THE ROLE OF SPATIAL SCALES IN BIOGEOGRAPHY AND EVOLUTION OF SOIL-BORNE *STREPTOMYCES*

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Janani Hariharan

August 2022

THE ROLE OF SPATIAL SCALES IN BIOGEOGRAPHY AND EVOLUTION OF SOIL-BORNE *STREPTOMYCES*

Janani Hariharan, Ph. D.

Cornell University 2022

Soil microbes are critical to the sustenance of several nutrient cycles, communities of soil fauna, and ecosystem resilience and stability. Microbial biogeography allows us to understand how communities form and change over evolutionary time. My dissertation investigates the biogeography and population dynamics of soil bacteria using the genus *Streptomyces* as a model system. *Streptomyces* form long hyphae, have desiccation-resistant spores, and produce versatile biosynthetic products. Taken together, these traits aid in nutrient acquisition, growth, and competitive ability, and collectively improve their dispersal. In this dissertation, I assess the impact of environmental gradients and geographical separation on *Streptomyces* community assembly, horizontal gene transfer events, and population differentiation. A combination of amplicon sequencing and comparative genome analyses show that dispersal limitation occurs at local, regional, and continental scales, and is caused by both barriers to dissemination (such as elevation and distance) and inability to establish a population in a new habitat due to local biological conditions. Horizontal gene transfer (HGT) also emerges as an important evolutionary force shaping *Streptomyces* genomes. Bacterial populations that migrate to new locations show still show evidence of gene flow with their allopatric relatives. Recombination acts as a cohesive force that increases

genetic similarity within closely related groups of bacteria, and genes involved in amino acid production, vitamin utilization, and biosynthesis of secondary metabolites are frequently transferred within sympatric strains. Phylogroup-specific actinophage signatures were found in *Streptomyces* genomes, suggesting that transduction might be an important mechanism of HGT in the environment. The third chapter shows that migration to a new habitat can increase nucleotide-level diversity between conspecific strains even while high gene flow maintains their genetic cohesiveness well above the bacterial species boundary. Finally, the last chapter details the isolation and description of a novel species of soil *Streptomyces*, *Streptomyces apricus*. This research expands our understanding of soil microbial ecology and evolution by quantifying the impacts of space and time on bacterial communities. As climates and landscapes change, studying changes in gene content in microbial communities is crucial to understanding how soil ecosystems can be best managed for sustainability and productivity.

BIOGRAPHICAL SKETCH

Janani Hariharan was born on September 27, 1992 and raised in Chennai, India. Janani earned her B.Tech. in Biotechnology at Anna University in 2013 where she completed her first Microbiology lab course. Her undergraduate research at the Sálim Ali Centre for Ornithology and Natural History sparked an interest in ecology and a research career. In 2015, she earned an M.S. in Environmental Science at The Ohio State University, where she began working with soil bacteria under the guidance of Drs. Warren Dick and Parwinder Grewal. Her introduction to the world of bioinformatics led to the realization that she needed to build stronger foundational skills in programming, and she subsequently worked as a Junior Research Fellow in GANIT Labs from 2015-16. In 2017, she joined the School of Integrative Plant Science at Cornell University to work towards her Ph.D. under the supervision of Dr. Daniel Buckley. During her time at Cornell, she has enjoyed working with and admiring *Streptomyces*, teaching undergraduate microbiology courses, and working to build equity and justice into academic systems for graduate students.

To அம்மா and all the women who couldn't finish their Ph.Ds.

ACKNOWLEDGMENTS

I owe my intellectual and professional growth over the last five years to Dan Buckley's mentorship. Joining his research program was one of my best life decisions, and I will carry with me his curiosity and enthusiasm for microbes, commitment to his students' success, as well as his kindness and humility as a human being.

Many thanks to my committee members, Tory Hendry and Michael Stanhope, for the enjoyable conversations about microbial ecology, genomics, and life over the years. Thank you for taking a chance on me and agreeing to serve on my committee when I showed up at your office with a very not-well-defined plan. I'm grateful to the many professors I have interacted with and taught for, especially Sue Merkel and Janice Thies for encouraging my interests in pedagogy.

My Buckley lab officemates managed to make graduate school and even Bradfield seem like fun with their ability to offer help and silence in equal parts as needed. Roli, Cassi, Sam, Nicole, Sara, Steven, Sean – I've appreciated having you around for the R questions and presentation nerves. Kavya, Olenka, and Hannah – we met by random chance at our SIPS orientation and I'm so glad we've remained friends through everything since then.

I'd be remiss if I didn't thank the communities that kept me going when research was failing, and that nourished me as a human being. The OISE student leader community

helped me find a warm, nurturing space where I could be myself and hear my own voice for the first time. The IDP family (IDPhamily?) taught me so much about empathy, dialogue, and just being a good human being. I could not have asked for a better introduction to the compelling world of teaching. Special thanks to Sara Xayarath Hernández and Adi Grabiner Keinan for being the best role models a girl could ask for.

There are so many friends who made this dissertation possible with their love and enthusiasm. To PHWAGS for your unconditional support through the pandemic and general cheerleading through life. Manoj for being my longest and most supportive friend and mental health buddy. To Jazlin and Rachel for the million ways in which your friendship, creativity, and energy have changed my life.

Finally, to my mother and Aravind Natarajan. One of you inspired me to begin this journey, and the other ensured that I finished it.

TABLE OF CONTENTS

PREFACE

Microbial ecology has made deeper inroads into understanding species distributions across diverse environments thanks to high-throughput DNA sequencing. Biogeography is the discipline of deciphering the driving forces underlying those species distributions. Biogeographical patterns can help us understand how ecology and evolutionary history impact microbial interactions and influence how communities assemble in the environment. We have enough data about soil microbes to create a global soil atlas, but there is still a relative paucity of studies that integrate ecological theory with bacterial biogeography and community assembly, especially in free-living terrestrial bacteria. Designing and interpreting such studies in soil is not trivial, and there are several factors that increase the complexity of these experiments.

Soil can be an extremely heterogenous habitat at tiny scales, leading to micro-level environmental gradients even within a gram of soil. Ecological conditions influence genome evolution, but gene content dictates an organism's ability to rapidly adapt to a new habitat. Defining a bacterial population can be challenging in light of their propensity for horizontal gene transfer and long-distance dispersal. And while DNA is informative about a strain's evolutionary history, time is an important component of biogeography and thus, the appropriate phylogenetic markers should be used depending on the temporal scale under consideration. Thus, combining multiple disciplines is crucial to advancing our understanding of microbial ecology.

My dissertation investigates specific variables that can cause population differentiation under varying spatial and phylogenetic scales. Bacterial lifestyles are

highly variable even within the same environment, and it is often hard to make meaningful biological assumptions about ecology and evolution when questions are formulated at the community level. Thus, I chose to focus on a single, ubiquitous soil bacterial genus, *Streptomyces*. *Streptomyces* are excellent candidates for the study of biogeography broadly, and dispersal particularly, given their ability to form desiccation-resistant arthrospores as part of their developmental life cycle. Arthrospore production is likely to confer increased ability for dispersal in *Streptomyces* when compared to non-spore forming relatives or endospore-forming bacteria. *Streptomyces* are also likely critical members of soil ecosystems given their ability to degrade complex carbon substrates and produce an array of secondary metabolites.

Chapter 1 details the role of elevation in structuring *Streptomyces* community composition by limiting dispersal. Elevational gradients also vary strongly in other environmental variables such as temperature and vegetation type, making this a favorable set-up to distinguish the effects of environmental selection versus dissemination barriers in dispersal limitation. Soils were sampled from sites on a mountain that were separated by 100m elevation, and a watershed in the same region with similar horizontal separation was used as a contrast to ensure that we were measuring the effects of elevation and not just spatial distance.

We found that elevation increased beta diversity between *Streptomyces* communities, and that OTUs above 1000m were especially different from the ones below. Other measured variables like pH, soil organic matter content or temperature do not have as strong an effect on beta diversity. Horizontal distance at the flatter watershed also does not impact beta diversity; most of the turnover in the watershed

Streptomyces communities is a result of ecological drift. Dispersal is limited between the tops of the mountain and sites below, and homogenizing dispersal is common for the latter sites. Strain localization at the highest and lowest elevations resulted from a combination of selection and priority effects, leading me to the hypothesis that order of arrival at a newly opened site could play an important role in community assembly and ecology. The role of biotic processes like priority effects and density-dependent blocking is severely understudied in microbial ecology. These processes are especially likely to impact community assembly, structure, and function in dense but heterogenous habitats like soil and would be exciting new avenues for exploration with this model system.

Chapter 1 used the *rpoB* (RNA polymerase B) gene to quantify and identify *Streptomyces* OTUs as it provides finer taxonomic resolution for this genus than the 16S rRNA gene. While using a single phylogenetic marker allowed us to examine community composition along an elevational gradient, analysis of genome content from multiple individuals at the same site would offer a more comprehensive view of how the pangenome has evolved in response to geographical separation or gradients.

The other chapters of this dissertation focus on comparing sets of genomes from a geographically explicit culture collection to harness the power of comparative genomics in understanding the role of dispersal in *Streptomyces* population biology. Chapter 2 examines the role of horizontal gene transfer (HGT) on *Streptomyces* strains that are separated by over 2,000 km. The impacts of HGT on bacterial genome structure and composition have been studied previously, but the prevalence and

mechanism of HGT in environmental populations are still unknown, as is the functional importance of transfer events.

The strains compared in Chapter 2 share phylogenetic similarity and geographical proximity, allowing us to examine how gene sharing and frequency of HGT were impacted by these relationships. The majority of gene clusters were clearly vertically inherited from a common ancestor, including secondary metabolite gene clusters (SMGCs), but the presence of site-specific unique genes indicates that bacterial genomes may be constantly sampling the local gene pool with high deletion rates, until a beneficial gene undergoes a sweep to fixation in the population. The variation in genome sizes, even within the same species, highlights the importance of demographic processes like population bottlenecks and relaxed selection in creating conditions where gene acquisition could be favorable. *Streptomyces* strains typically have large genomes (8-10 Mb), and the phylogenetic reconstruction of this dataset suggests that HGT and gene deletions impact genome size under different conditions.

Recombination is seen to act as a net cohesive force that maintains the genetic similarity between strains of the same species with site-specific recombination between different species being rare but not absent. Host-phage specificity might play an important role in constraining HGT within species boundaries as clade-specific actinophage signatures were found in all *Streptomyces* genomes. Many horizontally transferred genes are annotated with putative adaptive functions including but not limited to antibiotic synthesis and resistance proteins, vitamin B synthesis enzymes, amino acid biosynthesis gene clusters, and metal transport and utilization proteins. Predicting function from a genome is a complicated and nuanced process however, and

observation of traits *in situ* is necessary to infer that these genes or clusters are advantageous to the strain.

Phylogenetic relatedness is clearly important in shaping genome evolution. However, dispersal to a new environment also leaves distinct signatures in a strain's gene content. Isolation-by-distance is known to create population structure in other bacterial groups including *Sulfolobus* and *Vibrio*, and Chapter 3 discusses the comparative genomics of 32 conspecific *Streptomyces* strains that were isolated from Washington (1 site) and Alaska (2 sites). The null hypothesis was that the Alaska strains would form a cohesive population with high rates of gene flow as compared to the Washington population, given that the sites are separated by distance and the Gulf of Alaska. While distance did impact genomic similarity, we observed higher population differentiation between the Alaska strains rather than the Alaska-Washington strains. In a similar vein, we observed a greater fraction of horizontally transferred genes shared between a site in Alaska and Washington, which is consistent with the hypothesis that these strains share recent gene flow despite the geographical distance but may be slowly differentiating. This chapter provides unique insight into microdiversity within a bacterial species and the processes and temporal scales that generate such diversity.

Finally, chapter 4 describes a novel *Streptomyces* strain that was isolated and characterized as part of the broader biogeography projects. Various challenges exist in naming and describing a bacterial species, including our inability to isolate pure cultures of the majority of soil bacteria. *Streptomyces* have unique, fungus-like morphology including the presence of substrate mycelia, aerial hyphae, and spore

formation. These properties enabled the isolation of several strains of *Streptomyces* from continental US soils that were spread plated on enrichment media containing antifungals. SUN51^{T} was isolated from Sun Prairie, Wisconsin was most closely related to *Streptomyces dioscori* but with an ANI of only 89.4%. Phylogenetic reconstruction using the 16S rRNA gene, concatenated 5-gene MLSA loci, and orthologous genes across the genome revealed this strain to be a novel, undescribed species. A polyphasic approach including biochemical characterization and morphological observations was used to describe the unique characteristics of this novel strain, and it is now validly published under the name *Streptomyces apricus*.

The data presented in these chapters offer insight into how biogeographical patterns emerge in soil bacteria with *Streptomyces* as a model system. Future work in this discipline would benefit from integrating and refining theory and methods from related fields such as population genetics and plant ecology to track the ancient and contemporary processes that structure soil bacterial communities.

CHAPTER 1

ELEVATIONAL GRADIENTS IMPOSE DISPERSAL LIMITATION ON *STREPTOMYCES*[*]

1.1 ABSTRACT

Dispersal governs microbial biogeography, but the rates and mechanisms of dispersal remain poorly characterized for most microbial taxa. Dispersal limitation is driven by limits on dissemination and establishment, respectively. Elevation gradients create striking patterns of biogeography because they produce steep environmental gradients at small spatial scales, and these gradients offer a powerful tool to examine mechanisms of dispersal limitation. We focus on *Streptomyces*, a bacterial genus common to soil, by using a taxon-specific phylogenetic marker, the RNA polymerase-encoding *rpoB* gene. By targeting *Streptomyces*, we assess dispersal limitation at finer phylogenetic resolution than is possible using whole community analyses. We characterized *Streptomyces* diversity at local spatial scales (100 m to 3000 m) in two temperate forest sites located in the Adirondacks region of New York State: Woods Lake (< 100 m elevation change), and Whiteface Mountain (> 1000 m elevation change). Beta diversity varied considerably at both locations, indicative of dispersal limitation acting at local spatial scales, but beta diversity was considerably higher at Whiteface Mountain. Beta diversity varied across elevation at Whiteface Mountain, being lowest at the mountain's base. We show that *Streptomyces* taxa exhibit

[*] Published as Hariharan J, Buckley DH. (2022) Elevational gradients impose dispersal limitation on *Streptomyces*. Frontiers in Microbiology 13: 856263. doi: 10.3389/fmicb.2022.856263

elevational preferences, and these preferences are phylogenetically conserved. These results indicate that habitat preferences influence *Streptomyces* biogeography and suggest that barriers to establishment structure *Streptomyces* communities at higher elevations. These data illustrate that *Streptomyces* biogeography is governed by dispersal limitation resulting from a complex mixture of stochastic and deterministic processes.

1.2 INTRODUCTION

For more than a century, elevational gradients have yielded unique insights into the ecological and evolutionary mechanisms that generate patterns of biogeography (Grinnell, 1917). Steep elevation gradients generate rapid shifts in habitat characteristics over short spatial distances, a property that is useful in determining the degree to which dispersal is driven by spatial distance or variation in habitat characteristics (Sundqvist et al., 2013). Elevation gradients have a strong influence on the biodiversity of plants and animals (Peters et al., 2016) with many taxa exhibiting mid-elevation peaks or "hump-shaped curves" in alpha diversity (Rahbek, 2004; Moradi et al., 2020), and similar patterns have been observed for microbes (Fierer et al., 2011; Singh et al., 2012; Liu et al., 2016; Siles and Margesin, 2016). Elevation can affect biodiversity by a range of mechanisms including ecological filtering by habitat preference (Wang et al., 2008; Shen et al., 2015, Fierer et al., 2011; Cho et al., 2018), variation in carrying capacity, and historical processes linked to patterns of climate change (Badgley and Fox, 2000; Rickart, 2001; de la Giroday et al., 2011; Schai-Braun et al., 2020). In addition, taxa that occupy mountain habitats are uniquely affected by

historical climate change as warming climates tend to push species distributions towards higher elevations (Flesch, 2019; Marshall et al., 2020; Neate-Clegg et al., 2021), minimizing dispersal opportunities for species of plants and animals found at high elevations (Sekercioglu et al., 2008).

Biogeographical patterns can be driven by ecological mechanisms (*e.g.,* assembly processes driven by ecological filtering and ecological drift), evolutionary mechanisms (*e.g.,* speciation due to selection and drift), and historical contingency (*e.g.* neutral processes linked to variation in geology and climate over time) (Hanson et al., 2012). Microbial biogeography is often thought to be constrained by ecological filtering, under the assumption that dispersal is largely unlimited and environmental gradients impose spatial structure on communities due to selection (Navarrete et al., 2015; Liu et al., 2018; Malard and Pearce, 2018; Malard et al., 2019). However, most evidence for unlimited microbial dispersal is obtained using highly conserved taxonomic markers (*e.g.*, rRNA genes) that have low taxonomic resolution and are insensitive to evolutionary processes that drive diversification (Hanson et al., 2012). Studies that use higher resolution taxonomic markers often find evidence for dispersal limitation with evidence that neutral processes can play a role in shaping patterns of microbial biogeography (Whitaker et al., 2003; Polz et al., 2013; Andam et al., 2016b; Choudoir et al., 2016; Choudoir and Buckley, 2018).

To explain the mechanisms that give rise to microbial biogeography we must first understand the forces that govern microbial dispersal. Dispersal is a two-part process comprised of dissemination, the movement from one place to another, and establishment, the successful colonization of a site characterized by the ongoing

production of viable offspring (Martiny et al., 2006). Dissemination can be either passive (as driven by wind, erosion, currents, and organismal vectors) or active (as driven by motility or hyphal growth) (Yang and van Elsas, 2018). It is likely that capability for dissemination varies considerably between microbial taxa. For example, windborne dissemination is likely to vary in relation to cell size (Wilkinson et al., 2012), and cell shape likely influences microbial dissemination and establishment (Young, 2006). In addition, dissemination is influenced by environmental states. For example, soil texture and temperature influence spore transport in *Phytophthora* fungi (MacDonald and Duniway, 1978), and weather patterns can affect aerial dissemination (de Groot et al., 2021). Successful dissemination, however, is insufficient for successful dispersal, as microbes must still establish a sustainable population in the new site. Establishment requires that the habitat be suitable for growth, and that competitive interactions (*e.g.*, antagonism, or density-dependent blocking) do not prevent ongoing reproduction (Woody et al., 2007; Cheong et al., 2021).

We performed analysis of *rpoB* amplicons to investigate community assembly in *Streptomyces* along an elevational and spatial gradient in the Adirondacks region in New York State. The use of *rpoB* as a taxonomic marker for this genus improves taxonomic resolution significantly (Rong and Huang, 2012; Higgins et al., 2021) as compared to analyses made using 16S rRNA genes. The use of high-resolution taxonomic markers is essential for investigating the mechanisms that govern microbial biogeography (Hanson, 2012; Chase et al., 2017).

Streptomyces are bacteria that form aerial hyphae and arthrospores (Flärdh, 2003), which facilitate dissemination. They are common in soil habitats worldwide

where they degrade a variety of common substrates derived from plant biomass (Yeager et al., 2017) and produce diverse antibiotics and secondary metabolites (Watve et al., 2005). Despite their high capacity for dissemination, and broad habitat tolerance, *Streptomyces* have been shown to exhibit endemism at regional scales, indicative of dispersal limitation (Andam et al., 2016a). Their wide distribution, ecological significance, theoretical capability for high dispersal, and their observed limited ranges make the *Streptomyces* genus an ideal group to understand dispersal and biogeography patterns in the soil.

We hypothesized that dispersal limitation would also occur at local scales, with dispersal limitation driven by barriers to establishment (*i.e.*, ecological filtering) rather than barriers to dissemination. To evaluate this hypothesis, we examined *Streptomyces* communities at two locations in the Adirondacks region of New York State. Sites at Whiteface Mountain varied greatly in elevation, while sites at Woods Lake varied little in elevation. All sites were broadly similar in habitat characteristics other than those linked to elevation. We predicted that high rates of local dissemination, coupled with ecological filtering as driven by elevation, would produce a strong gradient of beta diversity at Whiteface Mountain, and little variation in beta diversity at Woods Lake. We also predicted that ecological filtering would cause *Streptomyces* taxa to exhibit phylogenetic conservation with respect to elevation preference.

1.3 METHODS

1.3.1 Soil sampling for Whiteface Mountain and Woods Lake

Soil samples were collected from nine locations on Whiteface Mountain (WM)

in New York State (Fig. 1.1). The average elevation change and horizontal distance between sites at WM was 347 m and 1,361 m, respectively (metadata described in Table 1.1). The top of the mountain consists of shallow and well-drained loamy soil (Lythic Cryofolist) with moderately deep, well-drained Wallface-Skylight soils on gneiss bedrock at 1,200 m. The soil type changes to frigid Lithic Haplohumods characteristic of glaciated uplands below 800 m, and the base of Whiteface Mountain consists of deep, excessively drained sandy loam (Typic Haplorthod). To contrast the elevational gradient found at WM with the effects of spatial distance, we also collected samples from ten locations spanning two watersheds of Woods Lake (WL), which is situated in the Adirondacks region, 190 km from WM (Fig. 1.1). Details of soil collection from Woods Lake are described elsewhere (Melvin et al., 2013). The average elevation change and horizontal distance between sites at WL was 19 m and 353 m, respectively. WL contains sandy glacial till soils on top of hornblende granitic gneiss bedrock (Orthod Spodosols). Only non-limed soils from the Woods Lake watershed were included in this study.

Sample	Elevation (m)	pH	Temperature (ºC)	SOM (%)	Watershed
S01_R01	1200	3.93	15	16.63	NA
S01_R02	1200	4.1	15	31.93	NA
S01_R03	1200	3.74	15	29.19	NA
S02_R01	1100	3.86	16	22.14	NA
S02_R02	1100	3.61	17	53.96	NA
S02_R03	1100	3.86	16	35.99	NA
S03_R02	1000	3.67	16	26.33	NA
S03_R03	1000	4.05	17	27.41	NA
S04_R02	900	3.93	17	33.73	NA
S04_R03	900	4.69	17	27.67	NA
S05_R01	800	4.42	19	15.74	NA

S05_R02	800	4.52	19	14.04	NA
S05_R03	800	4.45	18	14.61	NA
S06_R01	700	4.96	20	17.92	NA
S06_R02	700	4.54	19	22.48	NA
S06_R03	700	4.95	18	13.23	NA
S07_R02	600	5.09	20	7.90	NA
S07_R03	600	4.7	19	8.87	NA
S08_R01	500	5.08	20	5.65	NA
S08_R02	500	4.81	20	5.04	NA
S08_R03	500	4.95	20	6.76	NA
S09_R01	400	4.81	19	11.48	NA
S09_R02	400	5.07	19	28.74	NA
S09_R03	400	5.11	19	9.08	NA
C1A1	610	4.14	NA	NA	C1
C1B1	615	4.1	NA	NA	C1
C1C1	657	4.28	NA	NA	C1
C1D1	620	4.16	NA	NA	C1
C1E1	638	4.59	NA	NA	C1
C2A1	609	4.45	NA	NA	C2
C2B1	622	4.62	NA	NA	C2
C2C1	630	4.28	NA	NA	C2
C2D1	643	4.21	NA	NA	C2
C2E1	643	4.88	NA	NA	C2

Table 1.1 Spatial and environmental metadata for each sample at Whiteface Mountain and Woods Lake. Sites that start with C (*e.g.*, C1A1) are in WL, and sites that start with S (*e.g.,* S01_R01) are in WM.

At each sampling site, soil cores were collected in triplicate using a soil probe (2.5 cm diameter, 5 cm depth). Soil temperature was measured at the time of sampling. Samples used to test soil properties were air dried for 24 hours and then sieved using a 2 mm mesh to remove plant debris and rocks, while soil used for DNA extraction was continuously stored at -20˚C. Soil pH was measured using the 1:1 soil:water method described elsewhere (Kalra, 1995), and soil organic matter (SOM) content was measured by the loss-on-ignition method described in the Kellogg Soil Survey

Laboratory Methods Manual (Burt, 2014).

1.3.2 DNA Extraction and Sequencing

DNA was extracted using the MoBio PowerSoil® DNA Isolation kit (Qiagen) and quantified using the PicoGreen fluorometric assay (ThermoFisher Scientific). A 406 bp region of the RNA polymerase gene (*rpoB*) was amplified by PCR (~25 ng DNA in a 25µl reaction) using *Streptomyces*-specific primers Smyces_rpoB1563F and Smyces_rpoB1968R as described elsewhere (Higgins et al., 2021).

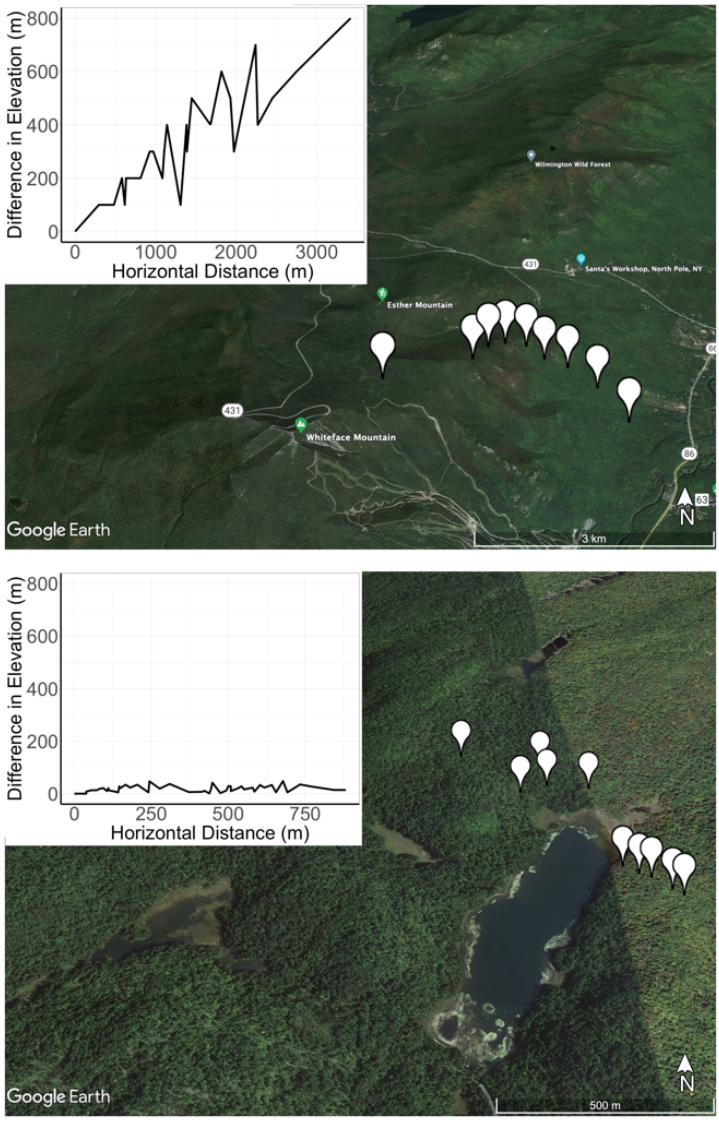

Figure 1.1 The sites at Whiteface Mountain (top) span more than 1000 m elevation while those at Woods Lake (bottom) span less than 100 m elevation. The topographical profile is provided in the inset. Horizontal distances (x-axis) are measured as geodesic distance or the shortest distance between the GPS coordinates of each site while the y-axis represents difference in elevation relevant to the base elevation at WM.

The PCR reactions consisted of 25 ng DNA, 12.5 µl of Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs, Ipswich, MA, USA), 0.625 µl of 4X Quant-iT PicoGreen dsDNA assay reagent (Thermo Fisher Scientific, Waltham, MA, USA), and 1.25 µl each of 10 µM dual-barcoded forward and reverse primers modified for

Illumina sequencing as described in (Kozich et al., 2013). PCR products from triplicate reactions were pooled and normalized using the SequelPrep Normalization Plate Kit (Thermo Fisher Scientific, Waltham, MA, USA). Fragments of 450 bp in length were size selected with a 1% agarose gel and subsequently extracted and purified from the gel band. Pooled samples were concentrated to 2 ng/µl using a vacuum concentrator and sequenced on an Illumina MiSeq instrument (2 x 300bp) at the Biotechnology Resource Center, Cornell University.

1.3.3 Additional datasets

In addition to the data generated from WM and WL, we also looked for evidence of elevational gradients in a larger dataset generated from soil samples obtained as part of the North American Soil Geochemical Landscapes Project (United States Geological Survey, 2012). This dataset consists of *Streptomyces rpoB* amplicons generated from 1,108 soil samples derived from sites across the USA and Mexico, spanning 7 – 3,483m in elevation (average elevation 674m). These *rpoB* amplicons were generated using the same primers and protocols described above (Steven Higgins, unpublished).

1.3.4 Data Analysis

Paired-end reads were joined using bbmerge (Bushnell et al., 2017) and trimmed using Trimmomatic-0.38 (Bolger et al., 2014). Sequences were dereplicated and size-sorted prior to OTU clustering at 99% identity using USEARCH (Edgar, 2010). The 0.99 similarity threshold corresponds to the species cut-off for

Streptomyces, and provides better resolution for classifying *Streptomyces* at the species level than the 16S rRNA gene (Rong and Huang, 2012; Andam et al., 2016b; Higgins et al., 2021). OTUs were classified with SINTAX (Edgar, 2016). Sequences were aligned using MAFFT v7.475 (Katoh and Standley, 2013) and phylogenetic trees were constructed using the maximum likelihood method with RAxML 8.2.12 (Stamatakis, 2014).

Samples that had fewer than 18 sequences (first quartile value) were discarded from further analyses. All other samples were normalized using the Cumulative Sum Scaling method (Paulson et al., 2013), wherein OTU relative abundances within each sample are divided by the sample's library size (total number of reads in the sample). Downstream analyses for beta diversity estimates, phylogenetic signal, and phylogenetic clustering were performed using the phyloseq and picante R packages (Kembel et al., 2010; McMurdie and Holmes, 2013). Distance-decay relationships were quantified using the mgram function in the ecodist package (Goslee and Urban, 2007). The relative contributions of ecological processes like drift, selection and dispersal to community assembly were assessed using methods and code described elsewhere (Anderson et al., 2011; Stegen et al., 2013). The indicspecies R package (Cáceres and Legendre, 2009) was used for indicator species analyses, and RAxML (Stamatakis, 2014) was used to reconstruct phylogenetic relationships. Phylogenetic trees were visualized using iTol (Letunic and Bork, 2021). All analyses were performed in R version 3.6.1.

1.4 RESULTS

1.4.1 Streptomyces Diversity at Whiteface Mountain and Woods Lake

Streptomyces exhibited greater richness at WM (74 ± 69 OTUs, average and standard deviation) than at WL (6 ± 3, ave. and s.d.), and this result was significant (Mann Whitney U-test, $p = 0.004$). *Streptomyces* richness at WM was maximal at mid elevations (500-900 m, Fig. 1.2), consistent with the classic hump-backed pattern of alpha diversity seen in previous studies (Singh et al., 2012; Liu et al., 2016; Kou et al., 2021). In contrast, sites at WL varied little in richness (Fig. 1.2). Although WL had fewer *Streptomyces* OTUs than WM, rarefaction curves indicate that both sites were adequately sampled (Fig. 1.3).

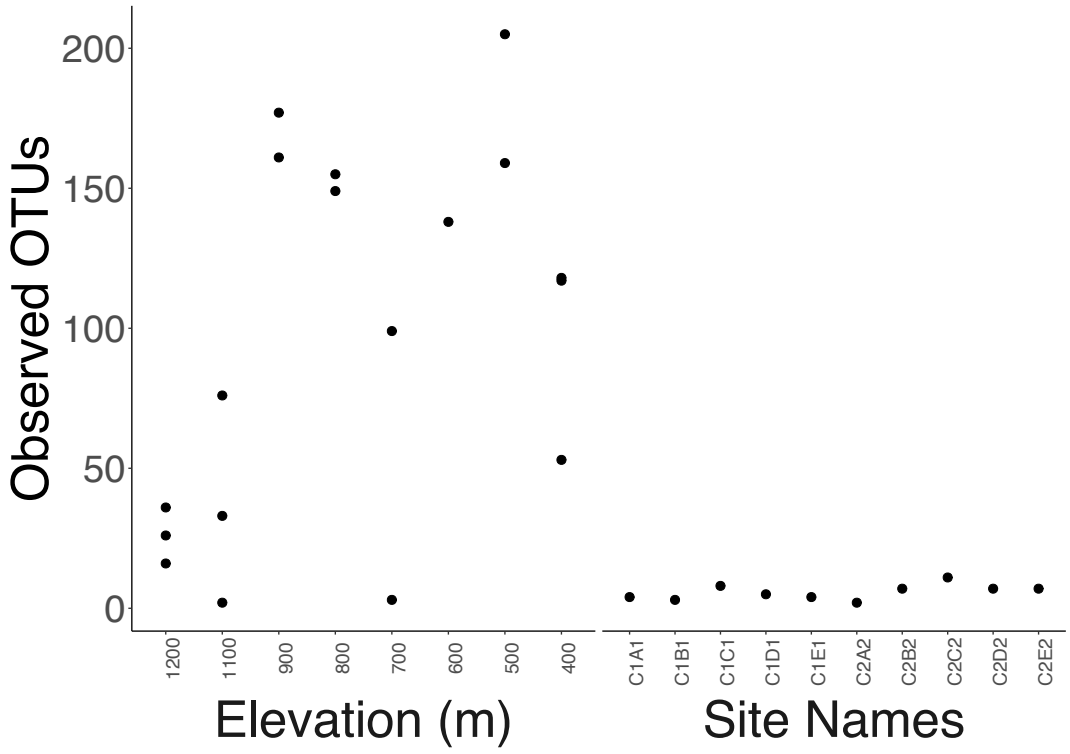

Figure 1.2 Alpha diversity is represented as number of observed OTUs in each sample. *Streptomyces* at Whiteface Mountain (left) exhibited maximal richness at mid elevations with decreasing alpha diversity at the highest and lowest sites. *Streptomyces* at Woods Lake (right) exhibited little change in alpha diversity with respect to

elevation change. Sample names are provided instead of elevations in the right panel because elevation does not vary across the WL watershed (largest elevation difference is 33m, Table 1.1).

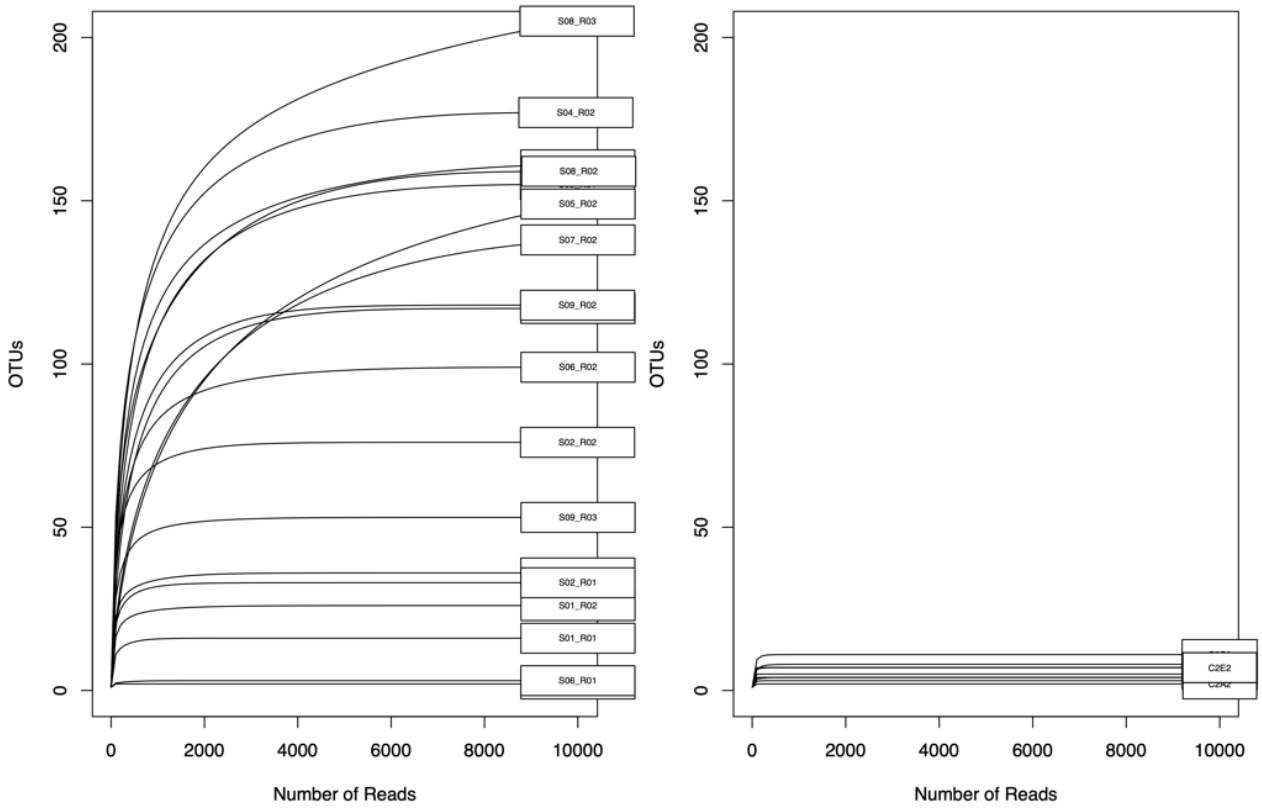

Figure 1.3 Rarefaction curves show that most sites within WM and all WL sites have been sampled to saturation.

Streptomyces exhibited greater beta diversity at WM (0.76 ± 0.28, unweighted UniFrac distance, ave. and s.d.) than at WL (0.35 ± 0.21, unweighted UniFrac distance, ave. and s.d.), and this difference was significant (Mann Whitney U-test, p <0.0001). In addition, *Streptomyces* communities at WM and WL were highly dissimilar (0.84 ± 0.164, unweighted UniFrac distance, ave. and s.d.). Analysis of similarities (ANOSIM) indicates that elevation is the strongest predictor of beta diversity at WM (R = 0.5284,

$p < 0.001$, 9,999 permutations). Additionally, beta diversity was partitioned into turnover and nestedness (Baselga, 2010). Species turnover underlies most of the beta diversity at both sites (90.91% at WM, 59.5% at WL), and turnover was significantly higher than nestedness at WM (paired t-test, $p < 0.0001$) (Table 1.2). Taken together, these results indicate that dispersal is limited between WM and WL, and it is limited across elevation at WM.

Site	Turnover	Nestedness
C1A1	0.599	0.301
C1B1	0.461	0.439
C1C1	0.40	0.500
C1D1	0.761	0.139
C1E1	0.505	0.395
C2A2	0.562	0.338
C2B2	0.849	0.051
C2C2	0.359	0.541
C2D2	0.739	0.161
C2E2	0.719	0.181
S01_R01	0.883	0.117
S01_R02	0.911	0.089
S01_R03	0.913	0.086
S02_R01	0.923	0.077
S02_R02	0.917	0.083
S02_R03	1	0
S04_R02	0.918	0.081
S04_R03	0.898	0.102
S05_R01	0.901	0.099
S05_R02	0.936	0.064
S06_R01	0.866	0.133
S06_R02	0.901	0.099
S07_R02	0.990	0.02
S08_R02	0.881	0.119
S08_R03	0.857	0.142
S09_R01	0.880	0.119

S09_R02	0.883	0.117
S09_R03	0.904	0.095

Table 1.2 Beta diversity partitioned into nestedness and species turnover components. Sites that start with C (*e.g.,* C1A1) are in WL, and sites that start with S (*e.g.,* S01_R01) are in WM.

A high amount of species turnover indicates that species are replaced from local pools rather than from the regional meta-community. Only 12 OTUs were shared between WM and WL. Ten of these shared OTUs have high relative abundance as compared to random expectations based on a random draw from the regional meta-community (paired t-test, 1000 permutations, $p = 0.002$), indicating that high-abundance OTUs are more likely to be shared at regional scales than we would expect due to chance.

1.4.2 Elevation Drives Community Structure on Whiteface Mountain

If elevation causes phylogenetic clustering, beta diversity should correlate with elevation change. Of the five variables tested, elevation was the only variable that was significantly correlated with beta diversity (Figs. 1.4, 1.5). The correlation coefficients indicate an intermediate-strength relationship, consistent with previous findings of dispersal limitation in other bacterial communities (Bell, 2010; Angermeyer et al., 2016). Beta diversity at WL was not significantly correlated with elevation, horizontal distance, pH, or any other measured variables.

Community assembly was subdivided into selection, dispersal, and ecological drift (as described by Stegen et al., 2012, 2013). Ecological drift was the dominant assembly process within WL. At WM, variable selection was the dominant assembly process at high elevation (above 1000 m), as expected if ecological filtering is driven

by elevation (Fig. 1.6(A)). However, at WM, the importance of homogenizing dispersal increases at elevations below 1000 m (Fig. 1.6(A)). This result suggests that dissemination at WM is driven by downward movement of soil and water from high to low elevation, with ecological filtering due to variable selection limiting the establishment of 'high elevation clades' at mid and low elevation sites, while homogenizing dispersal is more common between mid and low elevation sites (Fig. 1.6(B)).

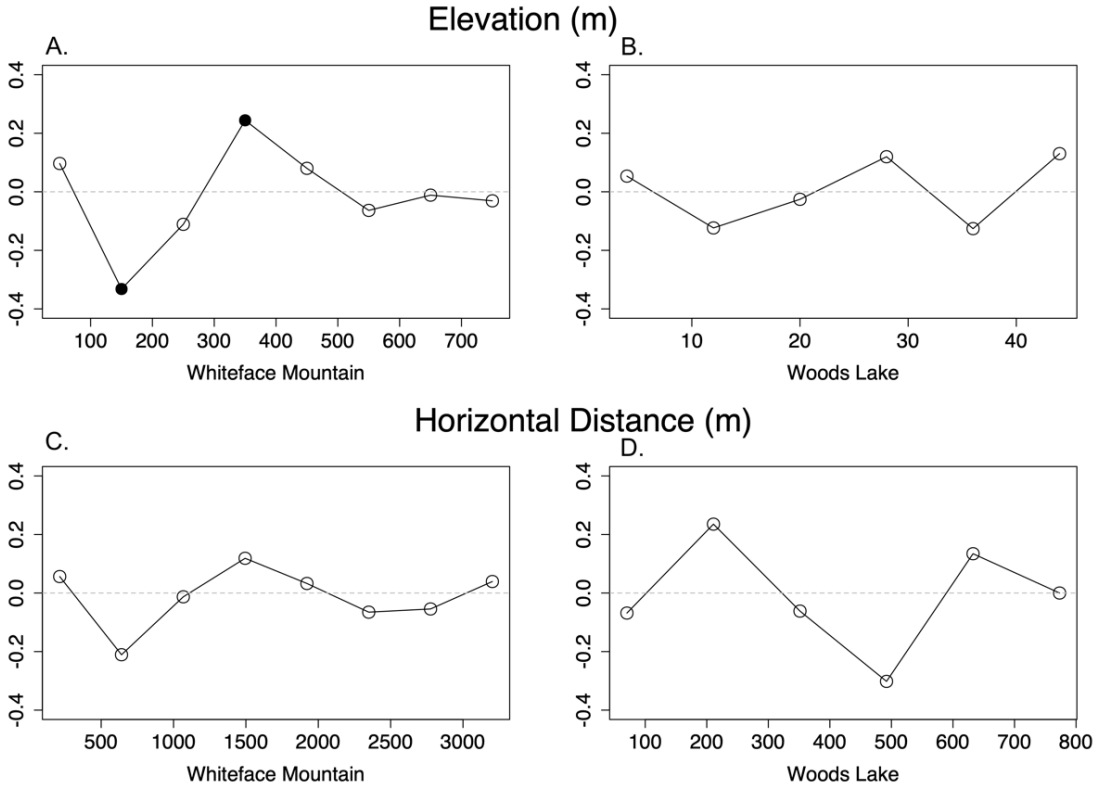

Figure 1.4 Partial Mantel tests indicate that beta diversity varies significantly with respect to elevation at Whiteface Mountain (A) but not at Woods Lake (B), and that beta diversity does not significantly correlate with horizontal distance at either site (C, D). Filled points indicate significant correlation ($p < 0.01$).

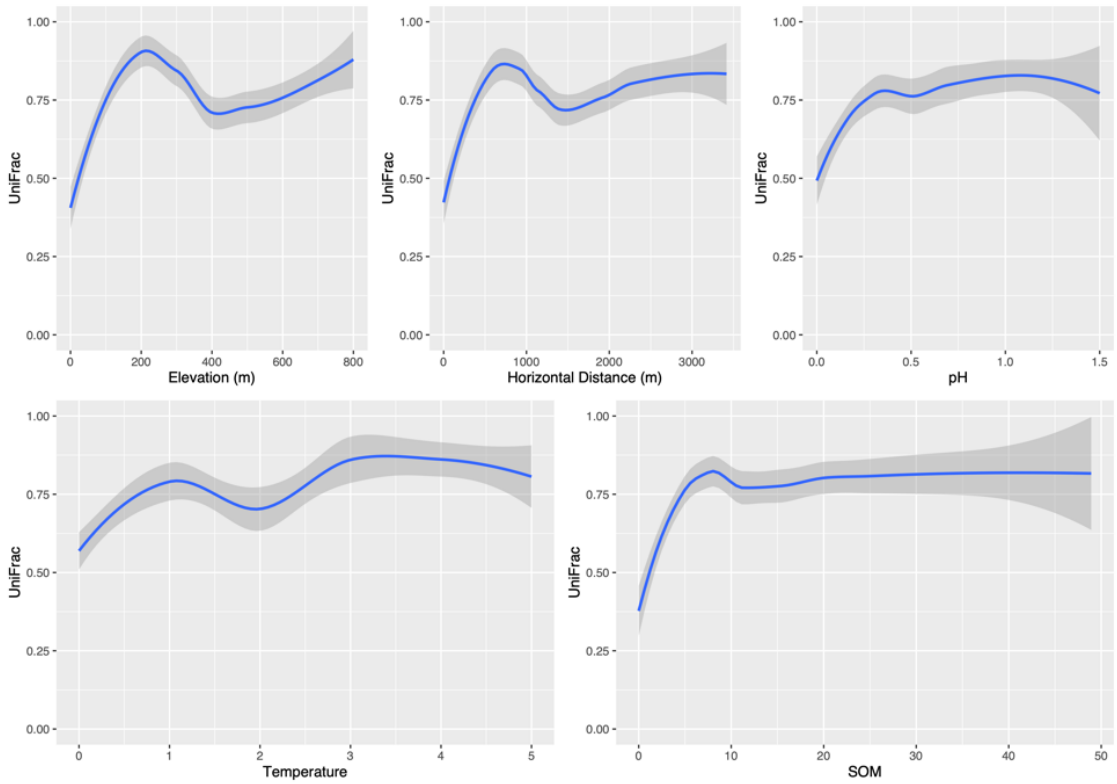

Figure 1.5 Elevation and horizontal distance are the best predictors of beta diversity at WM as measured by unweighted UniFrac distance along all measured environmental and spatial gradients.

1.4.3 Indicator Species Analysis

We conducted indicator species analysis to identify OTUs specific to elevation zones in WM (multipatt function, indicspecies R package). 16 OTUs were associated with elevations above 1000 m and 14 OTUs with elevations below 500 m. We evaluated pairwise phylogenetic distance of the indicator species with respect to elevation (Fig. 1.7). Low- elevation indicator species exhibited more phylogenetically similarity to each other than expected due to chance (*t*-test, $p = 0.006$), while high-elevation indicator OTUs were more diverse indicating the presence of multiple phylogenetic clusters (Fig. 1.7).

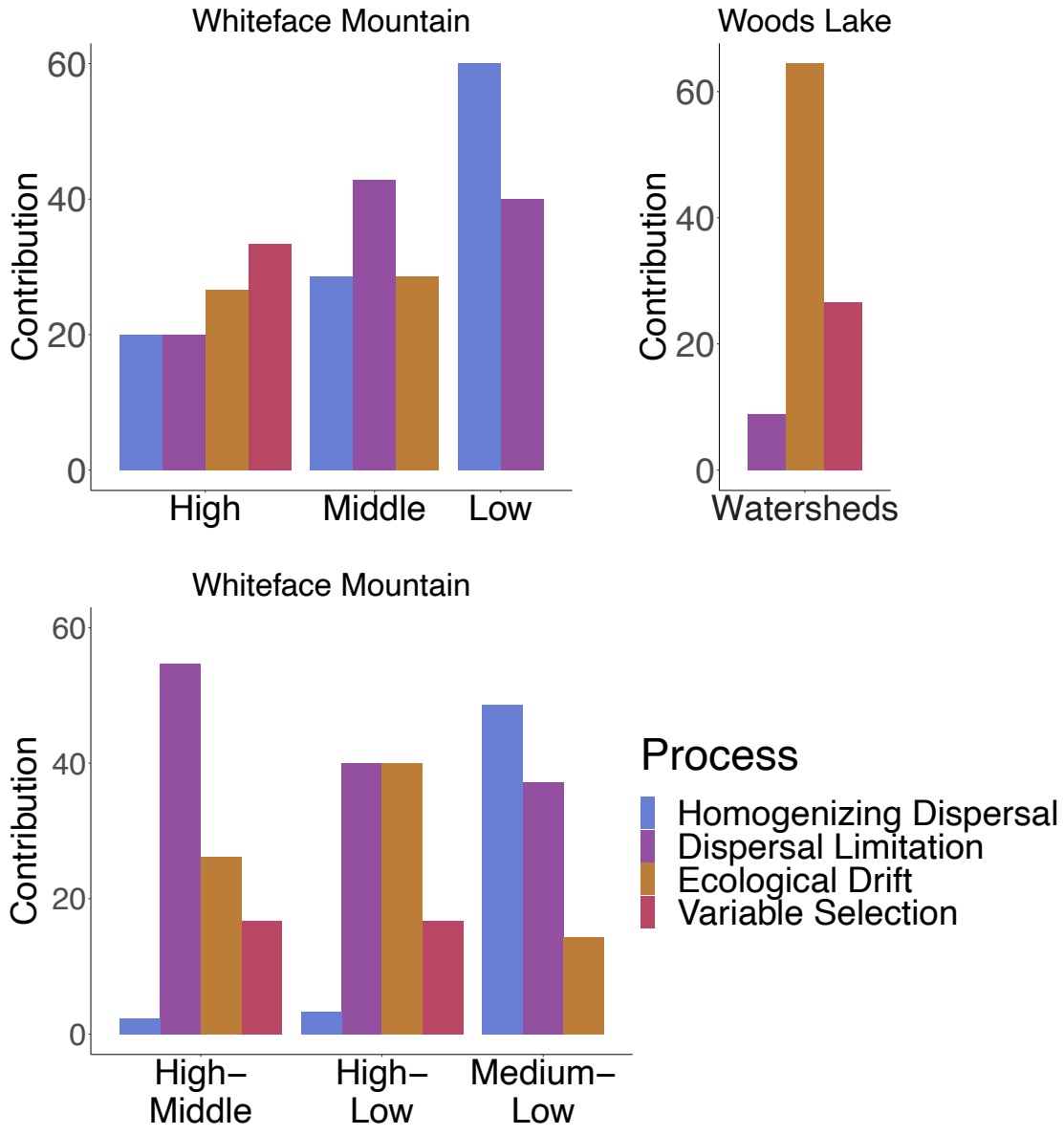

Figure 1.6 Relative contributions of selection, dispersal, and ecological drift to community assembly vary between sites and within elevation zones in Whiteface Mountain. For WM, High indicates elevations above 1000 m, Low indicates elevations below 500 m, and Middle indicates 500-1000m. (A) Variable selection drives beta diversity at sites above 1,000 m in WM while lower elevations have higher levels of homogenizing dispersal resulting in reduced beta diversity. (B) Dispersal limitation is highest between sites above 1,000 m and the rest of the mountain, while dispersal plays an important role in homogenizing *Streptomyces* communities between lower

elevations. (C) Community assembly at WL is dominated by ecological drift with some variable selection.

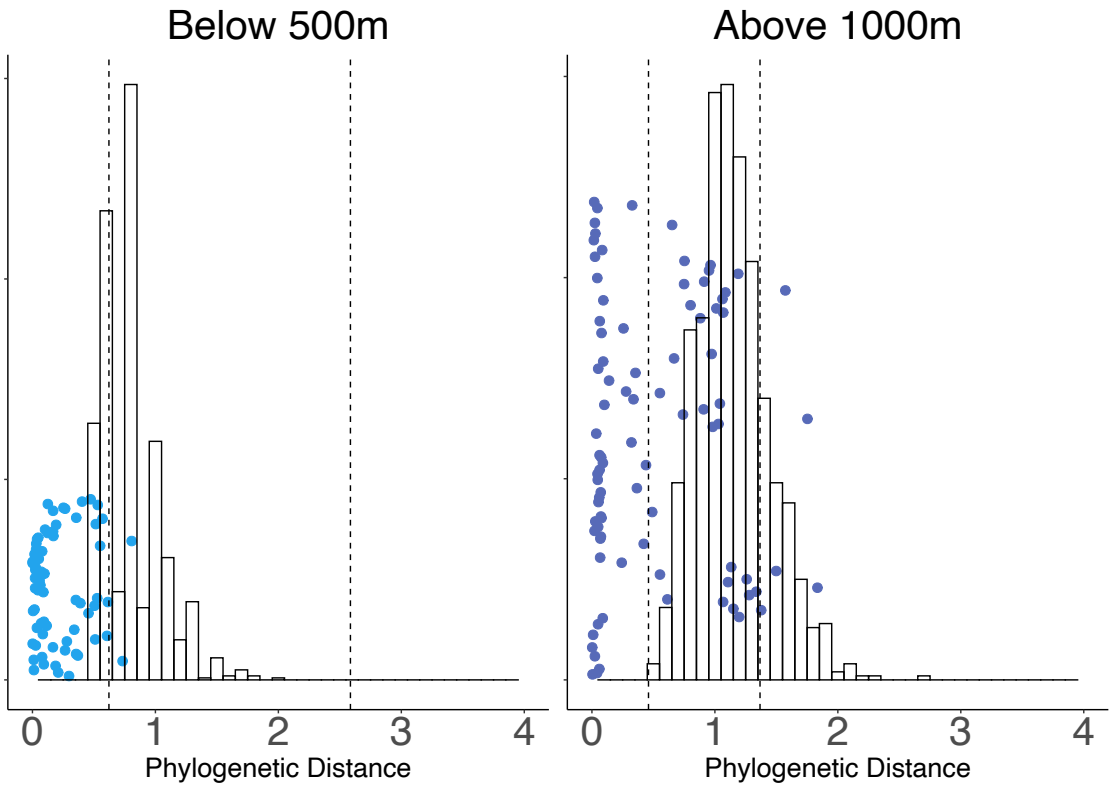

Figure 1.7 Scatter plots show the distribution of pairwise phylogenetic distances (A) between indicator OTUs below 500 m, and (B) between indicator OTUs above 1,00 0m. The histograms represent a bootstrapped distribution of 1,000 random draws from pairwise phylogenetic distances within the WM community, and 95% confidence intervals are indicated by vertical dashed lines. Each dot represents the phylogenetic distance between a pair of OTUs in that respective category.

1.4.4 Evaluating Indicator OTU Distribution in a Continental-scale Dataset

We further evaluated the elevational preferences of indicator OTUs by examining their distribution in a continental-scale dataset of *Streptomyces* biogeography. The continental-scale dataset contains *rpoB* sequences from *Streptomyces* communities across North America (see Methods). Briefly, indicator OTUs from WM were identified in the continental-scale dataset by clustering at 99%

identity. The preferred elevation for these OTUs was calculated as the normalized abundance-weighted average of elevations for all sites at which the OTU was detected. Elevational preference at WM had a strong influence on elevational distribution for conspecific OTUs in the continental survey (Cohen's d = 0.7, Fig. 1.8).

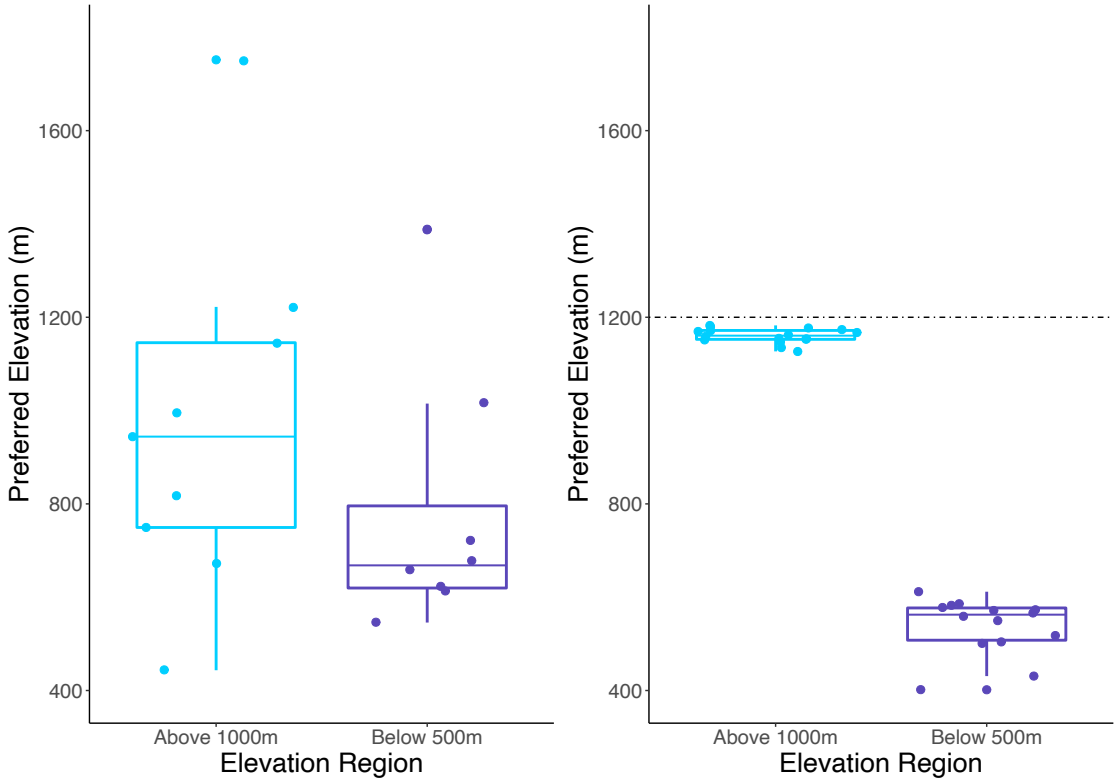

Figure 1.8 Abundance-weighted average elevation of each indicator OTU or its conspecific representative is represented as preferred elevation for (A) continental-scale distribution of *Streptomyces* and (B) the distribution of *Streptomyces* at WM. The dashed line indicates maximum elevation sampled at WM. Preferred elevations are significantly different between High and Low indicator OTUs at WM (Welch's t-test, $p < 0.001$).

Phylogenetic reconstruction of WM indicator OTUs and their continental relatives (Fig. 1.9) indicated that mixed clades (clades with both high- and low-indicator OTUs) were more evolutionarily divergent than clades with only one

indicator type as quantified by mean pairwise phylogenetic distances within each clade (*t*-test with 1,000 permutations, Bonferroni corrected $p = 0.002$). Additionally, clades with only high-elevation OTUs contained less divergence than those with only low-elevation OTUs (*t*-test with 1,000 permutations, Bonferroni corrected $p = 0.006$).

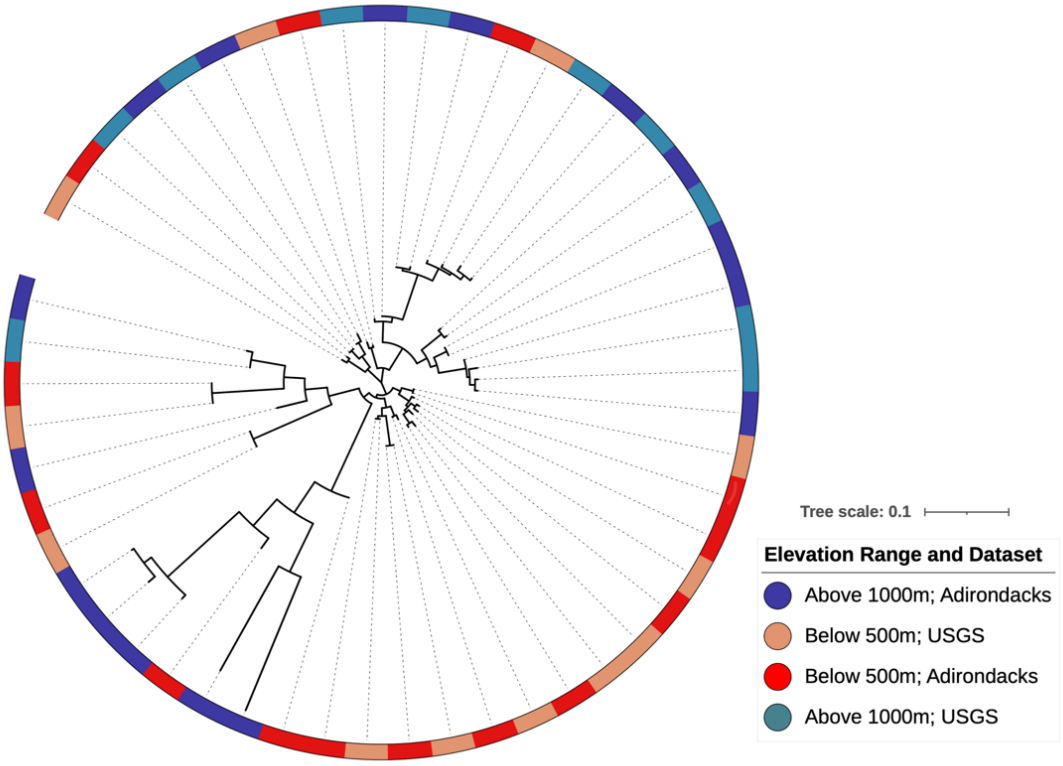

Figure 1.9 Phylogenetic reconstruction of indicator OTU lineages shows clade-level conservation of elevational preferences within *Streptomyces* found at WM. Some clades show mixed elevation preferences with branch lengths suggesting that indicator OTUs above 1,000 m have greater evolutionary divergence relative to those found at lower elevations. Colors are indicative of elevation and dataset for each OTU as indicated in the legend.

Patristic distance calculations suggest that low-elevation clades in Fig. 1.9 are younger than the mixed or high-elevation clades, as they have shorter root-to-tip

distances (0.09 ± 0.06 as compared to 0.16 ± 0.11 for high-elevation OTUs; Kruskal-Wallis test, $p = 0.004$). Hence, OTUs that are now localized above 1,000m exhibit greater evolutionary divergence than those present at lower elevations.

1.5 DISCUSSION

We show that *Streptomyces* communities are dispersal limited at local spatial scales and that dispersal limitation is governed by limits to both dissemination and establishment. *Streptomyces* communities at Woods Lake had low rates of dispersal, resulting in ecological drift at spatial scales spanning hundreds of meters. Low dispersal could result from low rates of dissemination, but the ability of *Streptomyces* to make aerial hyphae and desiccation resistant spores, and the fact that long range dispersal has been observed (Andam et al., 2016b; Choudoir et al., 2016; Higgins et al., 2021), suggest that limits to dissemination are unlikely to structure communities at local scales. Given the minimal habitat variation among sites at WL, and the broad habitat suitability expected for *Streptomyces*, it also seems unlikely that low dispersal at WL is driven by ecological filtering. Hence, we hypothesize that low dispersal and ecological drift at WL is likely driven by biotic interactions such as antagonism or density dependent blocking (Waters et al., 2013). *Streptomyces* are well known to produce diverse antimicrobial compounds, and other secondary metabolites, that alter biotic interactions (Kinkel et al., 2014; Schlatter and Kinkel, 2014; Vaz Jauri and Kinkel, 2014; Essarioui et al., 2016; Otto-Hanson and Kinkel, 2020). It seems likely that biotic interactions generate barriers to establishment that govern the structure of *Streptomyces* communities at WL.

We see strong evidence of *Streptomyces* dispersal limitation at Whiteface Mountain, with elevation having significant impacts on both dissemination and establishment. Certain clades preferentially occupy either high or low elevation habitats, indicating that barriers to establishment alter community structure across the mountain. The fact that these elevational habitat preferences are also observed in a continental-scale dataset suggests that ecological filtering by habitat preference constrains *Streptomyces* community structure across elevation. However, we also observe that homogenizing dispersal increases towards the base of the mountain (Fig. 1.6), and this suggests that the elevation gradient favors dissemination, likely due to movement of material down the mountain. Both alpha and beta diversity are significantly higher at WM than WL, consistent with the expectation that strong environmental gradients amplify patterns of microbial diversity despite high rates of dissemination. We expect that competitive interactions influence *Streptomyces* community composition at both WL and WM, but that the effect of elevation on dissemination and establishment is the main driver of community structure at WM.

Previous studies offer conflicting evidence for the effect of elevation on microbial biogeography. While research across a montane elevational gradient in Peru showed no effect of elevation on bacterial communities in soil (Fierer et al., 2011), a similar analysis on soils from the Andes Mountains found that bacterial and fungal diversity both varied with respect to elevation (Nottingham et al., 2018). Other studies have documented variation in bacterial community structure across elevation, attributing such variation to a range of factors including soil pH (Cho et al., 2018), aspect (Wu et al., 2017), soil carbon (King et al., 2008), and seasonality (Lazzaro et

al., 2015; Zhu et al., 2020).

Conflicting evidence on the relationship between bacterial biogeography and elevation could result from variation in spatial scales, habitat variability, and the phylogenetic resolution of taxonomic markers. Several environmental variables can co-vary with elevation (Sundqvist et al., 2013), making it difficult to disentangle the effect of elevation as opposed to other co-varying gradients. We also know that the spatial scale and taxonomic resolution at which diversity is measured can influence our ability to observe patterns of biogeography (Bent et al., 2003; Martiny et al., 2011; van de Guchte, 2017). Most prior studies of microbial diversity across elevation gradients have been performed using the 16S rRNA gene as a taxonomic marker. The low phylogenetic resolution of this marker makes it unsuitable for assessing mechanisms of dispersal limitation (Choudoir et al., 2012). For example, common taxonomic units defined on the basis of the 16S rRNA gene encompass strains whose ancestors may have diverged 50 - 150 million years ago (Ochman et al., 1999), and such taxonomic units lack the resolution needed to resolve the mechanisms that underlie extant patterns of microbial biogeography (Hanson et al., 2012). Our ability to identify the effect of environmental gradients on species distributions improves in proportion to the phylogenetic resolution at which diversity is characterized (Ramirez et al., 2018). Many phenotypic traits are conserved among closely related strains (Martiny et al., 2015; Barnett et al., 2021), and so experiments that use taxon-specific, fine-scale phylogenetic markers are vital to illustrate the processes driving microbial biogeography. Several examples of non-16S gene markers already exist in the literature; *dsrA*, *nirK*, *nirS*, and other MLST-based schemes have been used to detect

biogeographical patterns in environmental bacteria (Whitaker et al., 2003; Boucher et al., 2011; Angermeyer et al., 2016; Kou et al., 2021; Liao et al., 2021). In this study, the use of a *Streptomyces*-specific amplicon marker allows the exploration of phylogenetic patterns at a fine scale, sufficient for exploring the mechanisms that govern microbial dispersal.

Contemporary and historical climate variation is likely to influence patterns of *Streptomyces* biogeography. Elevation has a clear impact on temperature, as land temperatures decline 0.42˚C for every 100 m of elevation, such that a 200 m change in elevation approximates the temperature shift associated with a 1˚ change in latitude (Montgomery, 2006). Phylogenetic conservation of thermal traits has been shown to influence *Streptomyces* dispersal across latitude (Choudoir and Buckley, 2018), and such thermal adaptation likely contributes to the latitudinal diversity gradient observed for North American *Streptomyces* (Andam et al., 2016b). Whiteface Mountain is one of the highest peaks in the Adirondacks (1,484 m above sea level). Geological evidence indicates that the mountain was glaciated along with the entire Adirondacks region, until glacial retreat about 10,000 years ago (Franzi et al., 2000). This geological timeline means that *Streptomyces* have arrived fairly recently to WM and WL and hence the time for local diversification was limited. Prior to the period of glacial retreat, about 12,000 years ago, the climate in the Adirondacks region would have been approximately 2˚ C cooler than current conditions (Kaufman et al., 2020). Over time, as the climate warmed, species adapted to warmer climates would have dispersed into the wider Adirondacks region while cold adapted species would have found their habitat restricted to higher and higher positions on the mountain. In this

scenario, we hypothesize that ecosystem properties linked to climate variation influence microbial dispersal in soils by controlling the probability that species are able to establish at new sites. Changes in elevation influence a wide range of ecological variables both above and belowground (Sundqvist et al., 2013), and so it would be imprudent to conclude that temperature is the most important variable delimiting establishment, but it seems fair to conclude that ecological properties associated with climate variation can be expected to alter patterns of microbial establishment in soils.

Our findings indicate that a mixture of stochastic and deterministic processes govern *Streptomyces* dispersal. *Streptomyces* form aerial hyphae that produce desiccation-resistant, hydrophobic spores and their physiological traits should support broad habitat tolerance. As a result, we would generally assume that *Streptomyces* have a greater dispersal capacity than most other soil bacteria. However, we found high dissimilarity in communities that occupied similar habitats and similar elevations (350 m to 450 m elevation) at both WL and WM. This result suggests local limits on dispersal, likely driven by capacity for establishment as determined by competitive interactions between existing species and new immigrants. However, we did see evidence for homogenizing dispersal at the base of Whiteface Mountain suggesting that high rates of dissemination, likely driven by mass transport down the mountain, might overwhelm the ability of deterministic processes to constrain community assembly patterns. Our ability to identify dispersal limitation was enabled by the high phylogenetic resolution of the *rpoB* marker that we used, since 16S rRNA analyses provide little ability to resolve patterns of dispersal in *Streptomyces* (Higgins et al.,

2021). The existence and impact of dispersal limitation has now been documented for several microbial taxa across a range of ecosystems (Staley and Gosink, 2002; Whitaker et al., 2003; Bell, 2010; Eisenlord et al., 2012; Albright and Martiny, 2017; Bottos et al., 2018; Evans et al., 2019). Hence, it seems likely that microbial dispersal is finite and subject to change over time based on contemporary processes and historical contingencies (Hewitt, 2000; Mennicken et al., 2020; Liao et al., 2021). In the case of *Streptomyces*, we hypothesize that contemporary climate variation is a major determinant of establishment. We also hypothesize that historical variation in climate has contributed significantly to extant patterns of microbial biogeography in North America because rates of dispersal are low relative to rates of climate variation during the Quaternary Period.

1.6 Data Availability

Raw data for all amplicon sequences used in this study are available at NCBI under BioProject PRJNA790982 with BioSample accession numbers SAMN24252842 – SAMN24252878.

REFERENCES

Albright MBN, Martiny JBH. (2017) Dispersal alters bacterial diversity and composition in a natural community. *The ISME Journal* 12: 296–299. doi:10.1038/ismej.2017.161.

Andam CP, Choudoir MJ, Vinh Nguyen A, Sol Park H, Buckley, DH (2016a) Contributions of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces* lineages. *The ISME Journal* 10: 1731–1741. doi:10.1038/ismej.2015.230.

Andam CP, Doroghazi JR, Campbell AN, Kelly PJ, Choudoir MJ, Buckley DH. (2016b) A latitudinal diversity gradient in terrestrial bacteria of the genus *Streptomyces*. *mBio* 7: e02200-02215. doi:10.1128/mBio.02200-15.

Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, Sanders NJ, Cornell HV, Comita LS, Davies KF, Harrison SP, Kraft NJB, Stegen JC, Swenson NG. (2011) Navigating the multiple meanings of β diversity: A roadmap for the practicing ecologist. *Ecology Letters* 14: 19–28. doi:10.1111/j.1461-0248.2010.01552.x.

Angermeyer A, Crosby SC, Huber JA. (2016) Decoupled distance-decay patterns between *dsrA* and 16S rRNA genes among salt marsh sulfate-reducing bacteria: Biogeography of sulfate-reducing bacteria. *Environmental Microbiology* 18: 75–86. doi:10.1111/1462-2920.12821.

Badgley C, Fox DL. (2000) Ecological biogeography of North American mammals: Species density and ecological structure in relation to environmental gradients. *Journal of Biogeography* 27: 1437–1467. doi:10.1046/j.1365-2699.2000.00498.x.

Barnett SE, Youngblut ND, Koechli CN, Buckley DH. (2021) Multisubstrate DNA stable isotope probing reveals guild structure of bacteria that mediate soil carbon cycling. *Proceedings of the National Academy of Sciences* 118: e2115292118. doi:https://10.1073/pnas.2115292118.

Baselga A. (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography* 19: 134–143. doi:10.1111/j.1466-8238.2009.00490.x.

Bell T. (2010) Experimental tests of the bacterial distance–decay relationship. *The ISME Journal* 4: 1357–1365. doi:10.1038/ismej.2010.77.

Bent SJ, Gucker CL, Oda Y, Forney LJ. (2003) Spatial distribution of *Rhodopseudomonas palustris* ecotypes on a local scale. *Applied Environmental Microbiology* 69: 5192–5197. doi:10.1128/AEM.69.9.5192-5197.2003.

Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. doi:10.1093/bioinformatics/btu170.

Bottos EM, Kennedy DW, Romero EB, Fansler SJ, Brown JM, Bramer LM, Chu RK, Tfaily MM, Jansson JK, Stegen JC. (2018) Dispersal limitation and thermodynamic constraints govern spatial structure of permafrost microbial communities. *FEMS Microbiology Ecology* 94: 1–14. doi:10.1093/femsec/fiy110.

Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF. (2011) Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global *Vibrio cholerae* Populations. *mBio* 2: e00335-10.

Burt R. (2014) Kellogg Soil Survey Laboratory Methods Manual. U.S. Department of Agriculture, Natural Resources Conservation Service.

Bushnell B, Rood J, Singer E. (2017) BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE* 12: 1–15. doi:10.1371/journal.pone.0185056.

Cáceres MD, Legendre P. (2009) Associations between species and groups of sites: indices and statistical inference. *Ecology* 90: 3566–3574. doi:10.1890/08-1823.1.

Cheong JZA, Johnson CJ, Wan H, Liu A, Kernien JF, Gibson ALF, Nett JE, Kalan LR. (2021) Priority effects dictate community structure and alter virulence of fungal-bacterial biofilms. *The ISME Journal*, 2012–2027. doi:10.1038/s41396-021-00901-5.

Cho H, Tripathi BM, Moroenyane I, Takahashi K, Kerfahi D, Dong K, Adams JM. (2018) Soil pH rather than elevation determines bacterial phylogenetic community assembly on Mt. Norikura. *FEMS Microbiology Ecology* 95: 1–10. doi:10.1093/femsec/fiy216.

Choudoir M, Campbell A, Buckley DH. (2012) Grappling with Proteus: population level approaches to understanding microbial diversity. *Frontiers in Microbiology* 3: 336. doi:10.3389/fmicb.2012.00336.

Choudoir MJ, Buckley DH. (2018) Phylogenetic conservatism of thermal traits explains dispersal limitation and genomic differentiation of *Streptomyces* sister-taxa. *The ISME Journal* 12: 2176–2186. doi:10.1038/s41396-018-0180-3.

Choudoir MJ, Doroghazi JR, Buckley DH. (2016) Latitude delineates patterns of biogeography in terrestrial *Streptomyces*. *Environmental Microbiology* 18: 4931–4945. doi:10.1111/1462-2920.13420.

de Groot GA, Geisen S, Wubs ERJ, Meulenbroek L, Laros I, Snoek LB, Lammertsma DR, Hansen LH, Slim PA. (2021) The aerobiome uncovered: Multi-marker metabarcoding reveals potential drivers of turn-over in the full microbial community in the air. *Environment International* 154: 106551. doi:10.1016/j.envint.2021.106551.

de la Giroday HMC, Carroll AL, Lindgren BS, Aukema BH. (2011) Incoming! Association of landscape features with dispersing mountain pine beetle populations

during a range expansion event in western Canada. *Landscape Ecology* 26: 1097–1110. doi:10.1007/s10980-011-9628-9.

Edgar RC. (2016) SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161. doi:10.1101/074161.

Edgar RC. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461. doi:10.1093/bioinformatics/btq461.

Eisenlord SD, Zak DR, Upchurch RA. (2012) Dispersal limitation and the assembly of soil Actinobacteria communities in a long-term chronosequence. *Ecology and Evolution* 2: 538–549. doi:10.1002/ece3.210.

Essarioui A, Kistler HC, Kinkel LL. (2016) Nutrient use preferences among soil *Streptomyces* suggest greater resource competition in monoculture than polyculture plant communities. *Plant and Soil* 409: 329–343. doi:10.1007/s11104-016-2968-0.

Evans S, Bell-Dereske L, Dougherty K, Kittredge H. (2019) Dispersal alters soil microbial community response to drought. *Environmental Microbiology* doi:10.1111/1462-2920.14707.

Fierer N, Mccain CM, Meir P, Zimmermann M, Rapp JM, Silman MR, Knight R. (2011) Microbes do not follow the elevational diversity patterns of plants and animals. *Ecology* 92: 797–804. doi:10.1890/10-1170.1.

Flärdh K. (2003) Growth polarity and cell division in *Streptomyces*. *Current Opinion in Microbiology* 6: 564–571. doi:10.1016/j.mib.2003.10.011.

Flesch AD. (2019) Patterns and drivers of long-term changes in breeding bird communities in a global biodiversity hotspot in Mexico. *Diversity and Distributions* 25: 499–513. doi:10.1111/ddi.12862.

Franzi DA, Ridge JC, Pair DL, Desimone D, Rayburn JA, Barclay DJ. (2000) Post-Valley Heads Deglacation of the Adirondack Mountains and Adjacent Lowlands. *Adirondack Journal of Environmental Studies* 21: 119–146.

Goslee SC, Urban DL. (2007) The **ecodist** Package for Dissimilarity-based Analysis of Ecological Data. *Journal of Statistical Software* 22. doi:10.18637/jss.v022.i07.

Grinnell J. (1917) The Niche-Relationships of the California Thrasher. *The Auk* 34: 427–433. doi:10.2307/4072271.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology* 10: 497–506. doi:10.1038/nrmicro2795.

Hewitt G. (2000) The genetic legacy of the quaternary ice ages. *Nature* 405: 907–913. doi:10.1038/35016000.

Higgins SA, Panke-Buisse K, Buckley DH. (2021) The biogeography of *Streptomyces* in New Zealand enabled by high-throughput sequencing of genus-specific *rpoB* amplicons. *Environmental Microbiology* 23: 1452–1468. doi:10.1111/1462-2920.15350.

MacDonald JD, Duniway JM. (1978) Influence of Soil Texture and Temperature on the Motility of *Phytophthora cryptogea* and *P. megasperma* Zoospores. *Ecol. Epidemiology* 68: 1627–1630.

Kalra YP. (1995) Determination of pH of Soils by Different Methods: Collaborative Study. *Journal of AOAC International* 78: 310–324. doi:10.1093/jaoac/78.2.310.

Katoh K, Standley DM. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780. doi:10.1093/molbev/mst010.

Kaufman D, McKay N, Routson C, Erb M, Dätwyler C, Sommer PS, Heiri O, Davis B. (2020) Holocene global mean surface temperature, a multi-method reconstruction approach. *Scientific Data* 7: 1–13. doi:10.1038/s41597-020-0530-7.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26: 1463–1464. doi:10.1093/bioinformatics/btq166.

King AJ, Meyer AF, Schmidt SK. (2008) High levels of microbial biomass and activity in unvegetated tropical and temperate alpine soils. *Soil Biology and Biochemistry* 40: 2605–2610. doi:10.1016/j.soilbio.2008.06.026.

Kinkel LL, Schlatter DC, Xiao K, Baines AD. (2014) Sympatric inhibition and niche differentiation suggest alternative coevolutionary trajectories among *Streptomycetes*. *The ISME Journal* 8: 249–256. doi:10.1038/ismej.2013.175.

Kou Y, Liu Y, Li J, Li C, Tu B, Yao M, Li X (2021) Patterns and drivers of *nirK*-type and *nirS*-type denitrifier community assembly along an elevation gradient. *mSystems* 6: e00667-21. doi: 10.1128/mSystems.00667-21

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied Environmental Microbiology* 79: 5112–5129. doi: 10.1128/AEM.01043-13

Lazzaro A, Hilfiker D, Zeyer J. (2015) Structures of microbial communities in alpine soils: seasonal and elevational effects. *Frontiers in Microbiology* 6. doi:10.3389/fmicb.2015.01330.

Letunic I, Bork P. (2021) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* 49: W293–W296. doi:10.1093/nar/gkw290.

Liao J, Bergholz P, Wiedmann M. (2021) Adjacent terrestrial landscapes impact the biogeographical pattern of soil *Escherichia coli* strains in produce fields by modifying the importance of environmental selection and dispersal. *Applied Environmental Microbiology* 87: 1–17. doi:10.1128/AEM.02516-20.

Liu D, Wu X, Shi S, Liu H, Liu G. (2016) A hollow bacterial diversity pattern with elevation in Wolong Nature Reserve, Western Sichuan Plateau. *Journal of Soils and Sediments* 16: 2365–2374. doi:10.1007/s11368-016-1422-5.

Liu D, Yang Y, An S, Wang H, Wang Y. (2018) The biogeographical distribution of soil bacterial communities in the loess plateau as revealed by high-throughput sequencing. *Frontiers in Microbiology* 9. doi:10.3389/fmicb.2018.02456.

Malard LA, Anwar MZ, Jacobsen CS, Pearce DA. (2019) Biogeographical patterns in soil bacterial communities across the Arctic region. *FEMS Microbiology Ecology* 95: 1–13. doi:10.1093/femsec/fiz128.

Malard LA, Pearce DA. (2018) Microbial diversity and biogeography in Arctic soils. *Environmental Microbiology Reports* 10: 611–625. doi:10.1111/1758-2229.12680.

Marshall L, Perdijk F, Dendoncker N, Kunin W, Roberts S, Biesmeijer JC. (2020) Bumblebees moving up: shifts in elevation ranges in the Pyrenees over 115 years. *Proceedings of the Royal Society B* 287: 20202201. doi:10.1098/rspb.2020.2201.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Øvreås L, Reysenbach A, Smith VH, Staley JT. (2006) Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology* 4: 102–112. doi:10.1038/nrmicro1341.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC. (2011) Drivers of bacterial β-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences* 108: 7850–7854. doi:10.1073/pnas.1016308108.

Martiny JBH, Jones SE, Lennon JT, Martiny AC. (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* 350. doi:10.1126/science.aac9323.

McMurdie PJ, Holmes S. (2013) Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8. doi:10.1371/journal.pone.0061217.

Melvin AM, Lichstein JW, Goodale CL. (2013) Forest liming increases forest floor carbon and nitrogen stocks in a mixed hardwood forest. *Ecological Applications* 23: 1962–1975. doi:10.1890/13-0274.1.

Mennicken S, Kondratow F, Buralli F, Manzi S, Andrieu E, Roy M, Brin A. (2020) Effects of past and present-day landscape structure on forest soil microorganisms. *Frontiers in Ecology and Evolution* 8. doi:10.3389/fevo.2020.00118.

Montgomery K. (2006) Variation in temperature with altitude and latitude. *Journal of Geography* 105: 133–135. doi:10.1080/00221340608978675.

Moradi H, Fattorini S, Oldeland J. (2020) Influence of elevation on the species–area relationship. *Journal of Biogeography* 47: 2029–2041. doi:10.1111/jbi.13851.

Navarrete AA, Venturini AM, Meyer KM, Klein AM, Tiedje JM, Bohannan BJM, Nüsslein K, Tsai SM, Rodrigues JLM. (2015) Differential response of *Acidobacteria* subgroups to forest-to-pasture conversion and their biogeographic patterns in the western Brazilian Amazon. *Frontiers in Microbiology* 6. doi:10.3389/fmicb.2015.01443.

Neate-Clegg MHC, Jones SEI, Tobias JA, Newmark WD, Şekercioğlu ÇH. (2021) Ecological correlates of elevational range shifts in tropical birds. *Frontiers in Ecology and Evolution* 9: 215. doi:10.3389/fevo.2021.621749.

Nottingham AT, Fierer N, Turner BL, Whitaker J, Ostle NJ, McNamara NP, Bardgett RD, Leff JW, Salinas N, Silman MR, Kruuk LEB, Meir P. (2018) Microbes follow Humboldt: temperature drives plant and soil microbial diversity patterns from the Amazon to the Andes. *Ecology* 99: 2455–2466. doi:10.1002/ecy.2482.

Ochman H, Elwyn S, Moran NA. (1999) Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences* 96: 12638–12643. doi:10.1073/pnas.96.22.12638.

Otto-Hanson LK, Kinkel LL. (2020) Densities and inhibitory phenotypes among indigenous *Streptomyces* spp. vary across native and agricultural habitats. *Microbial Ecology* 79, 694–705. doi:10.1007/s00248-019-01443-2.

Paulson JN, Stine OC, Bravo HC, Pop M. (2013) Robust methods for differential abundance analysis in marker gene surveys. *Nature Methods* 10: 1200–1202. doi:10.1038/nmeth.2658.

Peters MK, Hemp A, Appelhans T, Behler C, Classen A, Detsch F, Ensslin A, Ferger SW, Frederiksen SB, Gebert F, Haas M, Helbig-Bonitz M, Hemp C, Kindeketa WJ, Mwangomo E, Ngereza C, Otte I, Röder J, Rutten G, Costa DS, Tardanico J, Zancolli G, Deckert J, Eardley CD, Peters RS, Rödel M, Schleuning M, Ssymank A, Kakengi V, Zhang J, Böhning-Gaese K, Brandl R, Kalko EKV, Kleyer M, Nauss T, Tschapka M, Fischer M, Steffan-Dewenter I. (2016) Predictors of elevational biodiversity

gradients change from single taxa to the multi-taxa community level. *Nature Communications* 7: 13736. doi:10.1038/ncomms13736.

Polz MF, Alm EJ, Hanage WP. (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics* 29: 170–175. doi:10.1016/j.tig.2012.12.006.

Rahbek C. (2004) The role of spatial scale and the perception of large-scale species-richness patterns: Scale and species-richness patterns. *Ecology Letters* 8: 224–239. doi:10.1111/j.1461-0248.2004.00701.x.

Ramirez KS, Knight CG, de Hollander M, Brearley FQ, Constantinides B, Cotton A, Creer S, Crowther TW, Davison J, Delgado-Baquerizo M, Dorrepaal E, Elliott DR, Fox G, Griffiths RI, Hale C, Hartman K, Houlden A, Jones DL, Krab EJ, Maestre FT, McGuire KL, Monteux S, Orr CH, van der Putten WH, Roberts IS, Robinson DA, Rocca JD, Rowntree J, Schlaeppi K, Shepherd M, Singh BK, Straathof AL, Bhatnagar JM, Thion C, van der Heijden MGA, de Vries FT. (2018) Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology* 3: 189–196. doi:10.1038/s41564-017-0062-x.

Rickart EA. (2001) Elevational diversity gradients, biogeography and the structure of montane mammal communities in the intermountain region of north America. *Global Ecology and Biogeography* 10: 77–100. doi:10.1046/j.1466-822x.2001.00223.x.

Rong X, Huang Y. (2012) Taxonomic evaluation of the *Streptomyces hygroscopicus* clade using multilocus sequence analysis and DNA-DNA hybridization, validating the MLSA scheme for systematics of the whole genus. *Systematic and Applied Microbiology* 35: 7–18. doi:10.1016/j.syapm.2011.10.004.

Schai-Braun SC, Lapin K, Bernhardt KG, Alves PC, Hackländer K. (2020) Effect of landscape type, elevation, vegetation period, and taxonomic plant identification level on diet preferences of Alpine mountain hares (*Lepus timidus varronis*). *European Journal of Wildlife Research* 66. doi:10.1007/s10344-020-01398-7.

Schlatter DC, Kinkel LL. (2014) Global biogeography of *Streptomyces* antibiotic inhibition, resistance, and resource use. *FEMS Microbiology Ecology* 88: 386–397. doi:10.1111/1574-6941.12307.

Sekercioglu CH, Schneider SH, Fay JP, Loarie SR. (2008) Climate change, elevational range shifts, and bird extinctions. *Conservation Biology* 22: 140–150. doi:10.1111/j.1523-1739.2007.00852.x.

Siles JA, Margesin R. (2016) Abundance and diversity of bacterial, archaeal, and fungal communities along an altitudinal gradient in alpine forest soils: what are the driving factors? *Microbial Ecology* 72: 207–220. doi:10.1007/s00248-016-0748-2.

Singh D, Takahashi K, Kim M, Chun J, Adams JM. (2012) A hump-backed trend in bacterial diversity with elevation on Mount Fuji, Japan. *Microbial Ecology* 63: 429–437. doi:10.1007/s00248-0.

Staley JT, Gosink JJ. (2002) Poles apart: biodiversity and biogeography of sea ice bacteria. *Annual Review of Microbiology* 53: 189–215. doi:10.1146/annurev.micro.53.1.189.

Stamatakis A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313. doi:10.1093/bioinformatics/btu033.

Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, Rockhold ML, Konopka A. (2013) Quantifying community assembly processes and identifying features that impose them. *The ISME Journal* 7: 2069–2079. doi:10.1038/ismej.2013.93.

Stegen JC, Lin X, Konopka AE, Fredrickson JK. (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *The ISME Journal* 6: 1653–1664. doi:10.1038/ismej.2012.22.

Sundqvist MK, Sanders NJ, Wardle DA. (2013) Community and ecosystem responses to elevational gradients: processes, mechanisms, and insights for global change. *Annual Review of Ecology, Evolution, and Systematics* 44: 261–280. doi:10.1146/annurev-ecolsys-110512-135750.

United States Geological Survey. (2012) History and progress of the North American Soil Geochemical Landscapes Project, 2001-2010. https://www.usgs.gov/publications/history-and-progress-north-american-soil-geochemical-landscapes-project-2001-2010 [Accessed January 14, 2022].

van de Guchte M. (2017) Horizontal gene transfer and ecosystem function dynamics. *Trends in Microbiology* 25: 699–700. doi:10.1016/j.tim.2017.07.002.

Vaz Jauri P, Kinkel LL. (2014) Nutrient overlap, genetic relatedness and spatial origin influence interaction-mediated shifts in inhibitory phenotype among *Streptomyces* spp. *FEMS Microbiology Ecology* 90: 264–275. doi:10.1111/1574-6941.12389.

Waters JM, Fraser CI, Hewitt GM. (2013) Founder takes all: density-dependent processes structure biodiversity. *Trends in Ecology and Evolution* 28: 78–85. doi:10.1016/j.tree.2012.08.024.

Watve MG, Tickoo R, Jog MM, Bhole BD. (2001) How many antibiotics are produced by the genus *Streptomyces*? *Archives of Microbiology* 176: 386–390.

Whitaker RJ, Grogan DW, Taylor JW. (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* 301: 976–978. doi:10.1126/science.1086909.

Wilkinson DM, Koumoutsaris S, Mitchell EAD, Bey I. (2012) Modelling the effect of size on the aerial dispersal of microorganisms. *Journal of Biogeography* 39: 89–97. doi:10.1111/j.1365-2699.2011.02569.x.

Woody ST, Ives AR, Nordheim EV, Andrews JH. (2007) Dispersal, density dependence, and population dynamics of a fungal microbe on leaf surfaces. *Ecology* 88: 1513–1524. doi:10.1890/05-2026.

Wu J, Anderson BJ, Buckley HL, Lewis G, Lear G. (2017) Aspect has a greater impact on alpine soil bacterial community structure than elevation. *FEMS Microbiology Ecology* 93: fix032. doi: 10.1093/femsec/fix032

Yang P, van Elsas JD. (2018) Mechanisms and ecological implications of the movement of bacteria in soil. *Applied Soil Ecology* 129: 112–120. doi:10.1016/j.apsoil.2018.04.014.

Yeager CM, Gallegos-Graves LV, Dunbar J, Hesse CN, Daligault H, Kuske CR. (2017) Polysaccharide degradation capability of actinomycetales soil isolates from a semiarid grassland of the Colorado Plateau. *Applied Environmental Microbiology* 83: e03020-16. doi:10.1128/AEM.03020-16.

Young KD. (2006) The selective value of bacterial shape. *Microbiology and Molecular Biology Reviews* 70: 660–703. doi:10.1128/MMBR.00001-06.

Zhu B, Li C, Wang J, Li J, Li X. (2020) Elevation rather than season determines the assembly and co-occurrence patterns of soil bacterial communities in forest ecosystems of Mount Gongga. *Applied Microbiology and Biotechnology* 104: 7589–7602. doi:10.1007/s00253-020-10783-w.

CHAPTER 2

HORIZONTAL GENE TRANSFER AND SITE-SPECIFIC GENE POOL CONTRIBUTE TO THE GEOGRAPHICAL SPREAD OF *STREPTOMYCES*

2.1 ABSTRACT

The role of horizontal gene transfer (HGT) in shaping bacterial genomes is well recognized, but its extent and impact is still debated. In this study, we sought to understand how HGT works in closely related *Streptomyces* genomes that are separated by geographical distance. To address this question, we sequenced the genomes of 17 strains of *Streptomyces griseus* or *Streptomyces pratensis,* closely related species co-isolated from two sites 2,500 km apart in North America, one in Wisconsin and one in Washington. Members of the genus *Streptomyces* are spore-forming filamentous bacteria that are ubiquitous in the soil and are a predominant source of many antibiotics. Phylogenomic analysis of core genome content in *S. griseus* and *S. pratensis* shows strong partitioning with respect to both phylogeny and geography. Core genes contained 557 recent and 457 ancient recombination events, and most were constrained by phylogeny and not geography. Accessory gene content was also defined largely by phylogeny but not by geography. The diversity of secondary metabolite gene clusters is highly variable between strains, and some gene clusters were conserved with respect to phylogeny while others were conserved with respect to geography. Strain specific (unique) gene content varied significantly with respect to geography (Wilcoxon test, $p = 0.033$). In particular, the Washington

genomes were missing genes associated with biosynthetic pathways such as ribonucleotide and nucleoside biosynthesis. We conclude that the most HGT in the core and accessory genome is phylogenetically constrained, while HGT of shell genes is more likely constrained by geography. This outcome likely reflects differences in evolutionary tempo, whereas changes in the shell genome respond rapidly to contemporary environmental conditions, and changes in the core genome integrate over much longer periods of evolutionary history.

2.2 INTRODUCTION

Horizontal gene transfer (HGT) is known to be a strong evolutionary force that shapes prokaryotic genomes. Microbiologists have debated the quantitative contribution of HGT to bacterial evolution, but it is thought to affect prokaryotic genomes by enriching functional gene transfers that have adaptive potential (González-Torres et al. 2019). For example, most metabolic innovations in the *E. coli* genome were acquired through HGT (Pang and Lercher 2019). Traits associated with human and ecosystem health such as antibiotic resistance and agricultural intensification (Aminov 2010; Lerminiaux and Cameron 2019; Kent et al. 2020; Woods et al. 2020) are mediated by HGT in natural and managed environments. The amount of homologous recombination within a species, however, appears to vary by lifestyle; free-living strains with large genomes exhibit the highest rates of homologous recombination and large accessory genomes (González-Torres et al. 2019).

The known diversity of genetically distinct bacterial species and lineages suggest that barriers to HGT must exist in nature. Allopatric speciation, wherein

spatially isolated lineages expand and give rise to local variants with local patterns of HGT, can limit opportunities for interactions and have been reported in some microbial lineages (Chase et al. 2019). However, multiple lineages of two closely related species can also be isolated from the same environment, indicating other barriers to gene flow. In such cases, fine-scale ecological differences may be more significant in the generation and co-existence of species and lineage clusters (Whitaker et al. 2003; Cordero and Hogeweg 2009; Wiedenbeck and Cohan 2011). Previous studies have shown that bacterial subpopulations and genetic diversity may emerge as a result of the existence of discrete microniches (Cordero et al. 2012) and selection driving frequencies of alleles (Rosen et al. 2015) within a broad, apparently homogenous environment. The presence of cryptic niches, which may be difficult to measure or observe in nature, can also drive the ecological differentiation and subsequent speciation of microbial species, as observed in *Campylobacter jejuni* (Sheppard et al. 2008). At the other end of the spectrum are microbes like the widespread marine genus *Procholoroccus*, which displays evidence of local microdiversity only at the 99.5% sequence similiarity cut-off, probably because cells evolve faster than ocean currents can mix them (Martiny et al. 2009).

The genus *Streptomyces* (phylum Actinobacteria) is a highly diverse group found predominantly in the soil and decaying vegetation. They play an important role in carbon cycling through degradation of complex carbon substrates like cellulose, hemicellulose, and chitin (Yeager et al. 2017). Unusually for bacteria, members of this genus are characterized by a complex secondary metabolism and a fungal-like morphological differentiation involving the formation of branching, filamentous

substrate mycelia, and aerial hyphae bearing long spores (Manteca and Sanchez 2009). They are also a major source of naturally derived antimicrobials and numerous bioactive, pharmaceutically relevant metabolites (Bérdy 2005). There are over 780 *Streptomyces* species (Labeda et al. 2017), including the genetic workhorse *Streptomyces coelicolor* and several phytopathogens such as *Streptomyces scabies*. While the clinical and agricultural importance of this genus has long been recognized, the factors driving its evolutionary history, ecology, and diversity remain poorly investigated.

Previous studies have reported the role of extensive horizontal gene transfer (HGT) in the evolution, diversity, and ecological adaptation of this genus (Li et al. 2019; Egan et al. 2001; Doroghazi and Buckley 2014; Choudoir and Buckley 2018; Cordero et al. 2012).This may in fact have contributed to the difficulties that have arisen in *Streptomyces* classification in the past (Anderson and Wellington 2001). The genetic coherence of *Streptomyces* species is therefore surprising, given the potential for widespread HGT within and between species, and even between *Streptomyces* and other genera. These observations suggest the existence of a mechanism for constraining inter-species HGT, giving rise to and maintaining the coherence of genetic clusters that correspond to *Streptomyces* species. The co-occurrence of distinct *Streptomyces* species in the natural environment despite widespread HGT presents a unique opportunity to investigate the dynamics of inter- and intra-species interactions at the genetic level. While genetic exchange in *Streptomyces* appears to be non-random, with inter-species HGT between species occurring at a much lower rate than within species (Doroghazi and Buckley 2014), it remains unclear what role the

environment and geography play in generating patterns of diversity and HGT between species.

Previous studies with symbiotic bacteria have found that phylogenetically and geographically proximal lineages will have the highest rates of gene exchange (Doroghazi and Buckley 2014; Greenlon et al. 2019). It is, however, unknown if the same trend applies to free-living soil dwellers given their potentially higher rates of dispersal and wider biogeographical distributions. Even with the *Streptomyces* genus, analyses of *S. pratensis* demonstrate high levels of HGT between strains from soils 740 km apart, indicative of either widespread contemporaneous dispersal or expansion of a population that was recently in equilibrium (Doroghazi and Buckley 2014). More recently, multi-locus sequence analyses of *Streptomyces* species suggest the existence of latitudinal barriers to dispersal and gene flow, which manifests as a latitudinal gradient of genetic diversity (Choudoir et al. 2016). Thus, the drivers and patterns of HGT are complex even within a single bacterial genus.

In this study, we investigate the contributions of phylogeny and geography towards HGT within and between two phylogroups of *Streptomyces, S. pratensis* and *S. griseus*. Strains in both groups are evenly distributed across two sites in the states of Washington and Wisconsin, USA (Fig. 2.1). This design allows us to contrast the frequency of gene transfer events and the magnitude of gene flux between co-localized and distant strains of the same. We hypothesize that (a) phylogenetic affinity rather than geography dictates gene flow and recombination patterns between *Streptomyces* strains given the species barrier to recombination, and (b) geography is likely to play a role in the acquisition of unique genes that may have adaptive significance to the local

habitat, but could also aid in the retention of non-adaptive genes under relaxed selection.

2.3 METHODS

2.3.1 Bacterial isolates and DNA extraction

A total of nine *S. pratensis* and eight *S. griseus* isolates (Table 2.1) were selected from an existing culture collection of more than 1,000 *Streptomyces* from 15 sites across the United States. Sites were selected to represent a narrow range of habitats dominated by perennial grasses and having neutral to slightly acidic pH (3.9-7.3). Soil was collected at 0-5 cm depth and air-dried, following which isolation of *Streptomyces* was carried out on glycerol-arginine media of pH 8.7 containing cycloheximide (300 mg/L) and rose bengal (35 mg/L) as previously described (Doroghazi and Buckley 2010). The genetic diversity of isolates in this collection was initially assessed using partial *rpoB* sequences. For this study, we selected isolates of *S. griseus* and *S. pratensis* collected from Rhinelander, Wisconsin and Bothell, Washington (straight line distance = 2,493 km) using previously described methods (Andam et al. 2016). Sub-groups were named as RH-pr for *S. pratensis* from Wisconsin, WA-pr for *S. pratensis* from Washington, RH-gr for *S. griseus* from Wisconsin and WA-gr for *S. griseus* from Washington. DNA was extracted from purified cultures, which were grown by shaking at 30°C in liquid yeast extract-malt extract medium (YEME) containing 0.5% glycine (Kieser et al. 2000), by using a standard phenol/chloroform/isoamyl alcohol protocol.

Name	Cluster	Sampling site	Latitude	Longitude	Size (Mbp)	N50	No. of contigs
RH45	RH-pr	Rhinelander, Wisconsin	45.63	-89.41	7.49	28462	475
RH79	RH-pr	Rhinelander, Wisconsin	45.63	-89.41	7.49	40914	365
RH155	RH-pr	Rhinelander, Wisconsin	45.63	-89.41	7.82	29432	538
RH185	RH-pr	Rhinelander, Wisconsin	45.63	-89.41	7.51	122364	173
RH254	RH-pr	Rhinelander, Wisconsin	45.63	-89.41	7.81	34852	455
WA1	WA-pr	Bothell, Washington	47.76	-122.21	8.35	194417	92
WA13	WA-pr	Bothell, Washington	47.76	-122.21	8.19	122969	148
WA22	WA-pr	Bothell, Washington	47.76	-122.21	8.17	172815	112
WA53	WA-pr	Bothell, Washington	47.76	-122.21	8.04	162578	137
RH34	RH-gr	Rhinelander, Wisconsin	45.63	-89.41	8.23	26146	630
RH195	RH-gr	Rhinelander, Wisconsin	45.63	-89.41	8.64	25829	621
RH206	RH-gr	Rhinelander, Wisconsin	45.63	-89.41	8.63	24347	671
RH207	RH-gr	Rhinelander, Wisconsin	45.63	-89.41	8.64	21115	786
WA1002	WA-gr	Bothell, Washington	47.76	-122.21	8.64	20023	893
WA1063	WA-gr	Bothell, Washington	47.76	-122.21	8.85	15367	1180
WA1064	WA-gr	Bothell, Washington	47.76	-122.21	8.7	136483	175
WA1071	WA-gr	Bothell, Washington	47.76	-122.21	8.75	11721	1401

Table 2.1 Assembly statistics of all draft genomes

2.3.2 Whole genome sequencing, annotation, and assembly

Genomic DNA over 10 kb was selected using gel purification (MO-BIO, Carlsbad, CA) and submitted to the Cornell Biotechnology Resource Center for DNA sequencing. DNA libraries were prepared using the Nextera library preparation kit and all libraries were run together in multiplex using GS FLX Titanium series reagents on a GS FLX instrument. The contiguous sequences were assembled de novo using the A5 genome assembly pipeline (Tritt et al. 2012). Annotations of the draft genomes were done using Prokka (Seemann 2014). Assembly statistics for each genome are found in Table 2.1. To measure the genomic relatedness between strain pairs, we

calculated the average nucleotide identity (ANI) values using fastANI (Jain et al. 2017) with default parameters. Pangenome gene content was analyzed using the Roary pipeline (Page et al. 2015).

2.3.3 Determination of orthologous gene families and secondary metabolite gene clusters

Orthologous groups of open reading frames (ORFs) from all strains were identified and clustered using OrthoMCL (Li et al. 2003) as implemented in GET_HOMOLOGUES REF using default settings of expectation value (E-value) for BLAST alignments of 1e-05 and a 75% minimum percentage of coverage required to call two sequences best hits.

2.3.4 Determination of HGT events

Draft genomes were aligned with each other using Mugsy (Angiuoli and Salzberg 2011), and a genome-based phylogeny was reconstructed using the maximum likelihood method as implemented in RAxML v8 (Stamatakis 2014). Effective recombination rates (r/m; recombination rate accounting for mutations) within populations were calculated using ClonalFrameML (Didelot and Wilson 2015).

fastGEAR (Mostowy et al. 2017) was used to identify ancient and recent recombination events as well as lineage probabilities for individual strains. All core genes identified by OrthoMCL were aligned with MUSCLE (Edgar 2004) and concatenated using FASconCAT-G (https://github.com/PatrickKueck/FASconCAT-G), and this alignment was used as the input to the fastGEAR algorithm. Selection

pressure on sites of recombination was assessed by calculating the ratio of non-synonymous substitutions to synonymous ones (dN/dS) using PAML (Yang 2007). The Panther database (Mi et al. 2021) was used to obtain gene ontologies for genes shared or lost between groups of interest. Only genes that have previously been identified in the *Streptomyces coelicolor* reference genome are reported in text unless otherwise specified. The tool HGTector (Zhu et al. 2014) was used to detect potential donors that were not any of the strains analyzed in this dataset.

2.3.5 Gene gain and loss

Orthogroups were detected using the OrthoFinder pipeline (Emms and Kelly 2019), and subsequently used for gene gain and loss analysis using Notung (Stolzer et al. 2012). The selected model allowed for HGT events, and so the Notung provided a count of genes gained, lost, and horizontally transferred at each node in the phylogeny. Genes gained or lost for specific groups of interest were functionally annotated using the COG database (Galperin et al. 2021) via the eggNOG mapper (Cantalapiedra et al. 2021). Selection pressure on flexible genes was measured using the dN/dS ratio from PAML as modified by GWideCodeML (Macías et al. 2020).

2.3.6 Identification of secondary metabolite gene clusters (SMGCs)

Secondary metabolite biosynthetic gene clusters were identified using AntiSMASH v3.0 (Weber et al. 2015). Codon-aware alignments were generated from the antiSMASH-identified nucleotide sequences using the pal2nal pipeline (Suyama et al. 2006). These alignments were subsequently processed through the PAML pipeline

to estimate the ratio of synonymous and non-synonymous substitutions within each cluster (ω) to identify whether neutral, positive, or negative selection was acting on secondary metabolite gene clusters.

2.3.7 Identification of mobile genetic elements

Mobile genetic elements were identified using sequence similarity searches with various databases: aclame for plasmid and phage elements (Leplae et al. 2010), ICEberg 2.0 for integrative and conjugative elements (Liu et al. 2019), and ISEScan for Insertion Sequence elements (Xie and Tang 2017). Sequenced phage genomes specific to the *Streptomyces* genus from the Actinobacteriophage Database (Russell and Hatfull 2017) were aligned with our draft genomes to identify actinophage signatures indicative of transductive gene transfer events.

All statistical tests and data visualizations were performed in R 3.6.1 (R Core Team 2021).

2.4 RESULTS

2.4.1 Strains cluster by phylogeny and geographic origin

A core genome-based phylogeny shows clades divided by both phylogeny and geography (Fig. 2.1). While the two major clades are divided based on phylotype, strains from different sites form sub-clades within each species clade. RH34 is an outlier to this clade structure such that the RH-GR and WA-GR clades appear to be more closely related than RH34 is to the other RH-GR strains. Nonetheless, we chose to include this strain in further analyses given that it is a representative of the *griseus*

species at Rhinelander, WI.

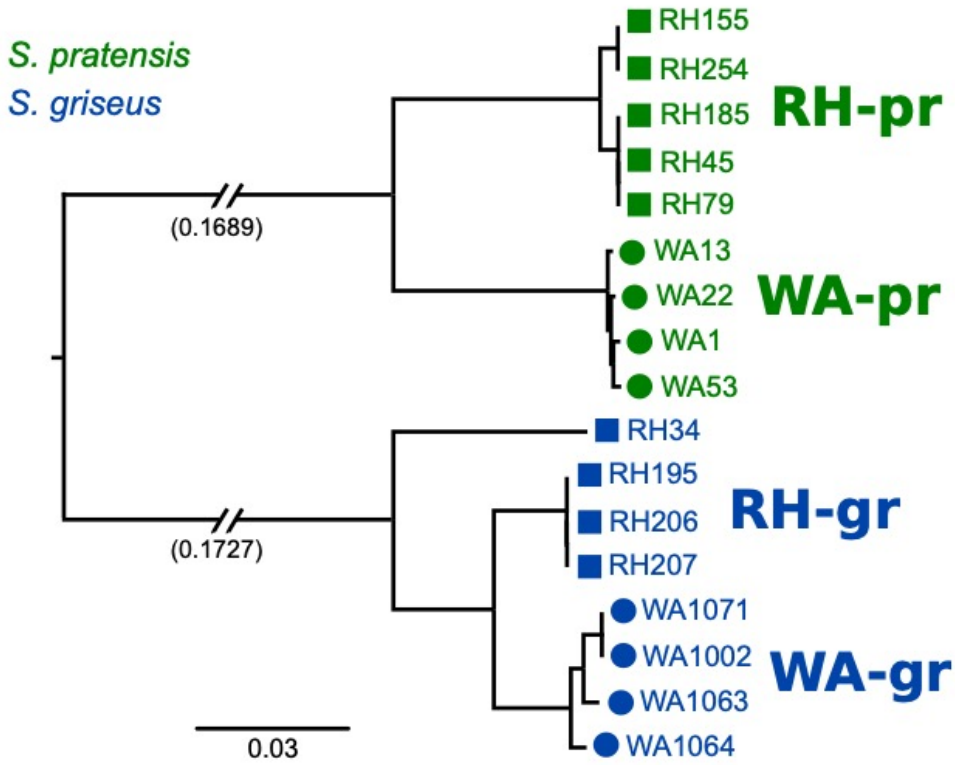

Figure 2.1 A maximum likelihood-based reconstruction of the core gene phylogeny shows strain differentiation based on their species grouping (*griseus* or *pratensis*) and site of isolation (Washington or Wisconsin).

All 17 strains were classified as either *pratensis* or *griseus* based on their rpoB sequences, and the ANI table in Table 2.2 confirms these species classifications. An analysis of homologous recombination in prokaryotes identified 95% ANI as the cut-off above which recombination rates outpaced background mutation rates in most species (González-Torres et al. 2019). Thus, we expect to see both intra-phylogroup and intra-site recombination events in this dataset.

2.4.2 Gene frequency distributions show sub-structure within species groups

Pratensis genomes share a higher number of core genes compared to the number of core genes shared within *griseus* or at the genus level. We estimated the pan-genome size of *griseus* at a lower bound of 35,660 protein families (or clusters) with an open pangenome (decay parameter alpha = 0.492 with 1,000 permutations) and the pangenome size of *pratensis* is closed (decay parameter alpha = 1.394 with 1,000 permutations) with a lower bound of 8,619 protein families using the R package micropan (Snipen and Liland 2015). The fluidity of the pan-genome, which determines how dissimilar genomes are at a gene level, was estimated for species within the genus *Streptomyces* (Kislyuk et al. 2011). The fluidity value was 0.37 +/− 0.31 for *griseus* and 0.13 +/− 0.06 for *pratensis*. This indicates that *griseus* genomes differ 37%, on average and the *pratensis* genomes differ 13% on average based on their gene cluster content.

Strains belonging to the same species typically show a characteristic U-shaped curve in a histogram of gene frequencies (Lobkovsky et al. 2013; Touchon et al. 2009) where number of genes shared are plotted against the number of genomes (Fig. 2.2). Based on our initial hypothesis, we expected these *Streptomyces* species to show a similar distribution, but the histograms in Fig. 2.2 indicate geographical structure in isolates within the same phylogroup.

	RH45	RH79	RH155	RH185	RH254	WA1	WA13	WA22	WA53	RH34	RH195	RH206	RH207	WA1002	WA1063	WA1064	WA1071
RH45																	
RH79	**99.96**																
RH155	**99.27**	**99.26**															
RH185	**99.92**	**99.95**	**99.26**														
RH254	**99.26**	**99.24**	**99.94**	**99.18**													
WA1	**93.85**	**93.87**	**93.88**	**93.88**	**93.84**												
WA13	**93.86**	**93.96**	**93.92**	**93.93**	**93.87**	**99.64**											
WA22	**93.83**	**93.89**	**93.86**	**93.88**	**93.84**	**99.72**	**99.62**										
WA53	**93.94**	**94**	**93.91**	**93.99**	**93.86**	**99.61**	**99.58**	**99.68**									
RH34	**84.76**	**84.76**	**84.85**	**84.96**	**93.74**	**84.63**	**84.52**	**84.52**	**84.54**								
RH195	84.84	84.79	84.86	84.87	84.72	84.83	84.74	84.68	84.77	***94.3***							
RH206	84.8	84.78	84.8	84.82	84.78	84.84	84.76	84.73	84.84	***94.27***	***99.94***						
RH207	84.84	84.83	84.84	84.84	84.76	84.81	84.7	84.7	84.85	***94.26***	***99.92***	***99.9***					
WA1002	84.91	84.82	84.76	84.81	84.85	84.6	84.62	84.64	84.71	***94.07***	***96.66***	***96.7***	***96.66***				
WA1063	84.93	84.87	84.86	84.88	84.91	84.68	84.69	84.73	84.83	***93.9***	***96.48***	***96.53***	***96.49***	***99.11***			
WA1064	84.89	84.82	84.74	84.83	84.81	84.72	84.69	84.62	84.68	***94.09***	***96.78***	***96.75***	***96.79***	***99.19***	***99.24***		
WA1071	84.94	84.88	84.78	84.9	84.82	84.79	84.72	84.74	84.77	***93.96***	***96.47***	***96.52***	***96.48***	***99.8***	***99***	***98.89***	

Table 2.2 Average Nucleotide Identity (ANI) between draft genomes analyzed in this dataset. ANI values between pairs of *griseus* strains are bolded and ANI values between *pratensis* strains are bolded and italicized

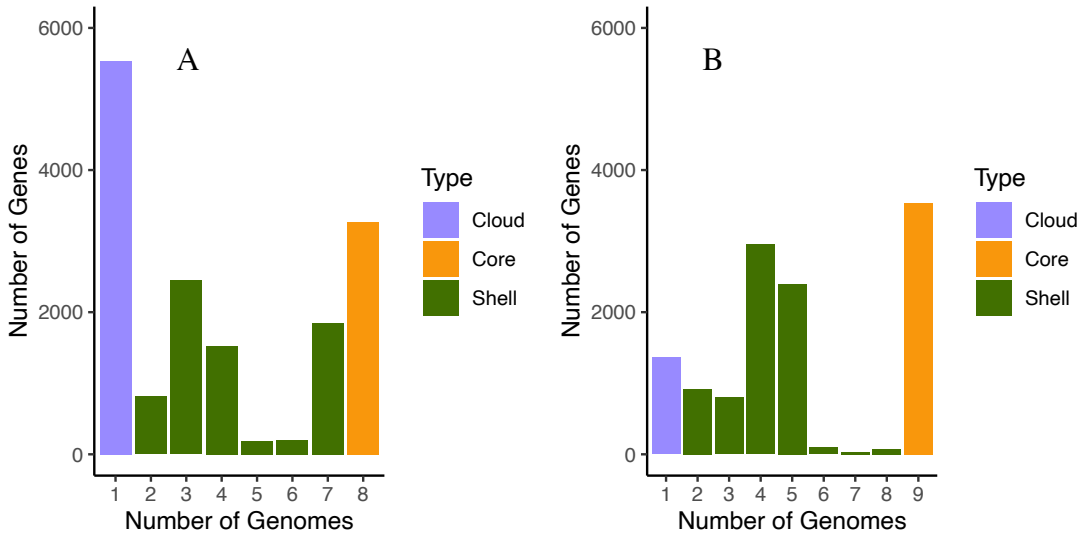

Figure 2.2 (A) Genes in the *griseus* strains show a higher number of shared core genes (3,269) when compared to the entire dataset as do the *pratensis* strains with each other (B; 3,535 genes).

Evidence for recombination acting as a cohesive force within species groups can be seen in the r/m values predicted with ClonalFrameML: 8.101 for *pratensis* and 6.272 for *griseus* strains. For comparison, r/m values of other soil bacteria have been reported to be in the order of 0.1-5.5 (Vos and Didelot 2009), and thus *Streptomyces* appears to have a higher-than-average r/m rate for terrestrial bacteria as seen in previous studies (Doroghazi and Buckley 2014). An r/m value over 0.25 is considered to represent the minimum threshold at which homologous recombination acts as a cohesive force on emerging population clusters (González-Torres et al. 2019).

2.4.3 Shell gene sharing patterns are driven by phylogeny

Phylogeny is the dominant factor in determining similarity of shell gene content, although gene sharing also occurs across phylogenetic and geographic

boundaries (Fig. 2.3) albeit to a limited extent. For instance, there is only 1 gene, *alsS* for acetolactate synthase, present exclusively in RH strains (with the exception of RH34), and one transcription factor (*tcrA*) shared by WA strains alone. In contrast, the *griseus* phylogroup shares 1601 exclusive genes and *pratensis* shares 1765 genes.

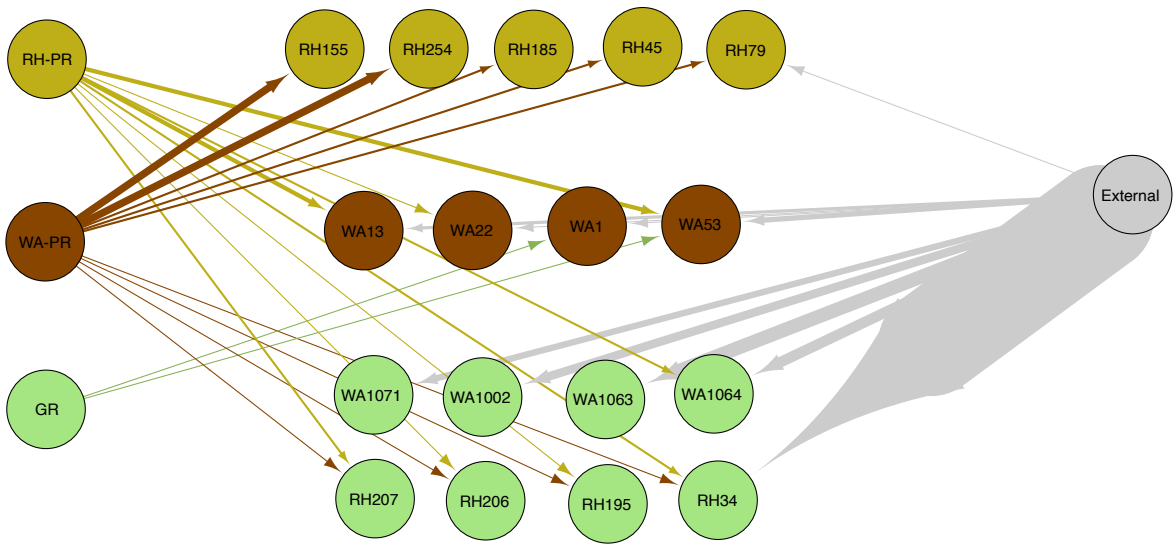

Figure 2.3 Strains were grouped into three lineages based on a concatenated core gene alignment. Membership in a lineage implies that the strain is more genetically similar to strains in the same lineage when compared to strains from other lineages. The *griseus* clade shows high genetic similarity despite being isolated from different sites while the *pratensis* strains are more similar to conspecific strains from the same site.

The number of shared genes increases when strains are considered on the basis of their clade grouping in Fig. 2.1. RH-GR strains share 517 genes without RH34 and 34 genes when RH34 is included, while RH-PR shares 924 genes. In Washington, WA-GR and WA-PR strains share 452 and 1118 genes respectively. These numbers indicate that even within a site, strains with higher genomic similarity share more genes (either due to shared ancestry or higher HGT). RH34, while consistent with the

species boundary that defines *griseus* strains, also has an unusually high number of unique genes (3611 versus an average of 201 genes across all other strains). When shell gene composition is analyzed without the RH34 genome, *pratensis* strains have a significantly higher number of shell genes than *griseus* regardless of site (Welch's t test; $p < 0.05$).

Common functions that seem to be specialized in clade-specific shell genes across environments and clades include biosynthesis of compounds needed for DNA replication (*de novo* purine and pyrimidine ribonucleotide biosynthesis), amino acids, peptidoglycan, and vitamin biosynthesis and vitamin metabolism (vitamins B6, B5, B7 biosynthesis, pyridoxal phosphate salvage and biosynthesis). While the entire list of shared genes and functional annotations is too long to report, we summarize clade-specific genes that may be of interest in adaptation to new environments in Table 2.3. Many such clade-specific flexible genes were also under positive selection pressure. Notably, within the WA-PR clade, genes encoding for formaldehyde dismutase, protein-L-isoaspartate O-methyltransferase, a putative prophage phiRv2 integrase, and trans-aconitate 2-methyltransferase had a $dN/dS > 1$, and the polyketide synthase PksM was under positive selection within all *pratensis* genomes.

RH-GR	**RH-PR**	**WA-GR**	**WA-PR**	**RH-GR without RH34**
Deferrochelatase/peroxidase EfeB	18 kDa heat shock protein	Aclacinomycin-T 2-deoxy-L-fucose transferase AknK	Aerobactin synthase IucC	55.5 kDa and 49.5 kDa sporulation proteins
Fructose import ATP-binding protein FruK	Aerobactin synthase IuC	Actinorhodin polyketide putative beta-ketoacyl synthase 1	Arginase Arg	Anthraniloyl-CoA anthraniloyltransferase PqsD
Hydroxydechloroatra	Agmatinase SpeB	Alkaline shock	Arsenate-mycothiol	Cobalamin

zine ethylaminohydrolase AtzB		protein 23	transferase ArsC2	biosynthesis protein CobD
Methionine synthase MetH	Alanine--anticapsin ligase BacD	Anaerobic magnesium-protoporphyrin IX monomethyl ester cyclase BchE	Arsenic resistance transcriptional regulator ArsR1	Erythromycin C-12 hydroxylase EryK
Nisin biosynthesis protein NisC	Antiseptic resistance protein QacA	Anhydrotetracycline monooxygenase OtcC	Arsenical-resistance protein Acr3	Ferrous iron permease EfeU
Ribose operon repressor RbSR	Antitoxin HigA1	Antiseptic resistance protein QacA	Bacillibactin exporter YmfD	Insertion sequence IS5376 putative ATP-binding protein
	Antitoxin YefM	Antitoxin MazE9	Beta-lactamase AmpC	Internalin-A InlA
	Arginase Arg	Antitoxin RelJ	Beta-lactamase inhibitory protein	Low-affinity putrescine importer PlaP
	Arginine biosynthesis bifunctional protein ArgJ	Arginine exporter protein ArgO	Cobalt import ATP-binding protein CbiO	Multidrug resistance protein 3 Bmr3
	Arsenate-mycothiol transferase ArsC2	Asparagine synthetase [glutamine-hydrolyzing] 1 AsnB	Cobalt transport protein CbiN	Multidrug resistance protein NorM
	Arsenic resistance transcriptional regulator ArsR1	Benzene 1,2-dioxygenase system ferredoxin--NAD(+) reductase subunit BedA	Colicin V secretion protein CvaA	Mycolysin Npr
	Arsenical-resistance protein Acr3	Cobyric acid synthase CpbQ	Denitrification regulatory protein NirQ	Nickel/cobalt efflux system RcnA
	Asparagine synthetase [glutamine-hydrolyzing] 3 AsnO	Dapdiamide A synthase DdaF	Dimodular nonribosomal peptide synthase DhbF	Penicillin acylase 2 proenzyme AcyII
	Beta-lactamase inhibitory protein	Daunorubicin/doxorubicin resistance ABC transporter permease protein DrrB	Enterobactin exporter EntS	Purine efflux pump PbuE
	Capsule biosynthesis protein CapA	DNA-invertase Hin	Ferredoxin-2 SubB	Purine ribonucleoside efflux pump NepI
	Chitin-binding protein CbpD	Ferric enterobactin transport system permease protein FepD	Germacradienol/geosmin synthase Cyc2	Rhodomycin D methylesterase DauP

	Flavohemoprotein Hmp	Ferric enterobactin transport system permease protein FepG	Large-conductance mechanosensitive channel MscL	Stress response kinase A SrkA
	Germacrene A synthase	Fused nickel transport protein NikMN	Multicopper oxidase Mco	Type I restriction enzyme EcoKI M protein HsdM
	Gramicidin S synthase 2 GsrB	Hydrogen peroxide-inducible genes activator OxyR	Multidrug resistance protein MdtH	Type-1 restriction enzyme R protein HsdR
	Haloacetate dehalogenase H-1 DehH1	Iron import ATP-binding/permease protein IrtA	Multiple antibiotic resistance protein MarR	Very short patch repair protein Vsr
	Insertion element IS6110 uncharacterized 12.0 kDa protein	Iron import ATP-binding/permease protein IrtB	Phthiocerol synthesis polyketide synthase type I PpsE	
	Insertion sequence IS5376 putative ATP-binding protein	L-cystine transport system permease protein YecS	Polyketide biosynthesis protein PksE	
	Iron-sulfur cluster carrier protein	L-ectoine synthase EctC	Purine efflux pump PbuE	
	L-aspartate oxidase NadB	Leucine efflux protein LeuE	Response regulator protein VraR	
	L-isoleucine-4-hydroxylase Ido	Metallo-beta-lactamase type 2 CphA	Rhodocoxin reductase ThcD	
	L-lysine N6-monooxygenase IucD	Nitrate/nitrite sensor protein NarX	Riboflavin biosynthesis protein RibD	
	L-methionine gamma-lyase Mgl	Plipastatin synthase subunit D PpsD	Septum site-determining protein MinD	
	L-proline cis-4-hydroxylase	Polyketide synthase PksL	Stage V sporulation protein D SpoVD	
	Lactose transport system permease protein LacG	Polyketide synthase PksN	Swarming motility protein SwrC	
	Lysostaphin Lss	Small-conductance mechanosensitive channel MscS	Tetracenomycin-F1 monooxygenase TcmH	
	Lysozyme M1 Acm	Streptothricin hydrolase SttH	Transposon Tn10 TetD protein TetD	
	Macrolide export ATP-binding/permease protein MacB	Subtilisin NAT AprN		
	Manganese-binding lipoprotein MntA	Surfactin synthase subunit 2 SrfAB		

	Mercuric reductase MerA	Toxin Doc		
	Molybdate-binding protein ModA	Toxin RelK		
	Molybdenum cofactor guanylyltransferase MobA	Tyrocidine synthase 3 TycC		
	Molybdenum import ATP-binding protein ModC	Tyrosine recombinase XerC		
	Multidrug efflux protein YfmO	Vancomycin B-type resistance protein VanW		
	Multiple sugar-binding protein MsmE			
	Oxygen regulatory protein NreC			
	Oxygen sensor histidine kinase NreB			
	Oxygen sensor histidine kinase response regulator DosT			
	Peptidoglycan-N-acetylmuramic acid deacetylase PdaA			
	Polyketide biosynthesis cytochrome P450 PksS			
	Polyketide synthase PksJ			
	Proline dehydrogenase 1 FadM			
	Proline iminopeptidase Pip			
	Resuscitation-promoting factor Rpf			
	Spore cortex-lytic enzyme SleB			
	Spore photoproduct lyase SplB			
	Sporulation-specific cell division protein SsgB			
	Subtilisin BL			
	Subtilisin DY			

	Surfactin synthase subunit 1 SrfAA			
	Tetracycline resistance protein, class B TetA			
	Threonine efflux protein RhTC			
	Threonine/homoserine exporter RhtA			
	Transposon Tn10 TetD protein			
	Valine--tRNA ligase ValS			
	Virulence sensor protein BvgS			

Table 2.3 Clade-specific shell genes that have the potential to aid in adaptation to new habitats. Genes specific to RH-GR are shown with and without RH34 given the unique nature of RH34's gene content and phylogenetic placement.

A significant geographical effect was seen on the number of unique genes with and without the RH34 genome, i.e., genes found in only one strain (WA > RH; Wilcoxon pairwise comparison test, $p < 0.05$). Given that these genes are only present in one genome, they are most likely genes that have been acquired from each strain's respective habitat and neighbors. Such low-frequency, strain-specific genes are common in bacterial pangenomes and are assumed to result from genetic drift (Choudoir et al. 2017).

2.4.4 Recombination events vary by species

Phylogenetic reconstruction partitions these strains into 4 sub-clades based on both phylogenetic affinity and geographic origin. Identifying recombination events on the basis of sequence similarity, however, indicates extensive gene transfer events that are not strictly constrained by either variable. Fig. 2.4 shows a clear effect of geography on the *pratensis* strains wherein they split into two distinct lineages. The

griseus strains cluster in the same lineage, presumably because their intra-lineage rates of recombination are higher than recombination rates with *pratensis* strains at the same sites. It is also possible that *griseus* strains have higher rates of gene exchange with other bacterial strains at the same location, as seen in the external recombination estimates for the *griseus* WA strains.

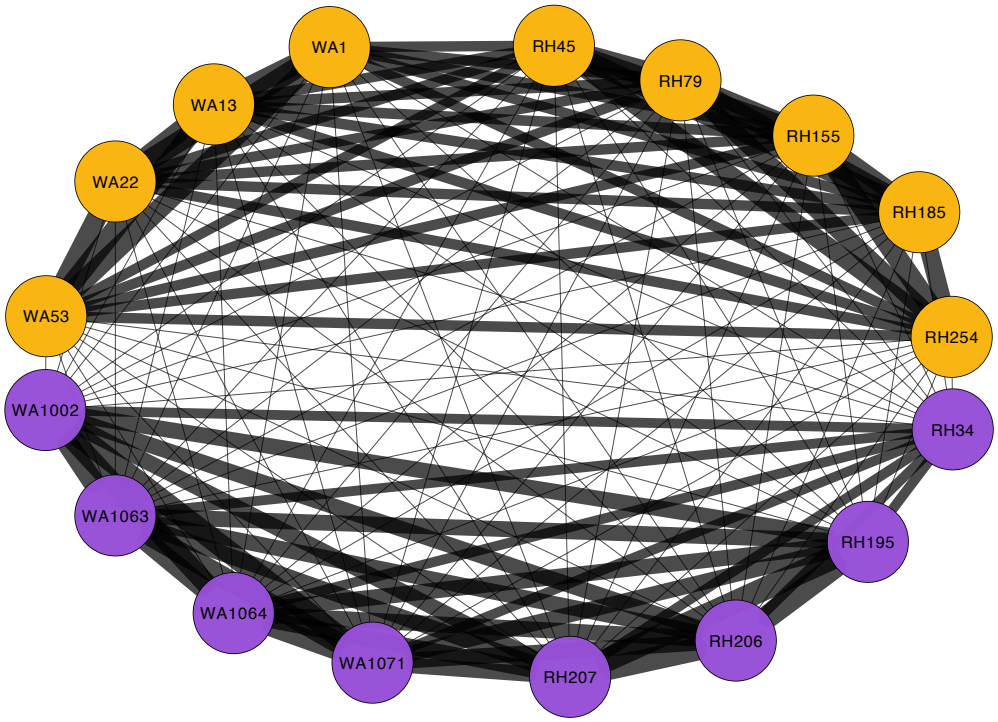

Figure 2.4 Strains with high genomic similarity share the most shell genes. The yellow and purple circles represent *pratensis* and *griseus* genomes respectively. Thickness of the connecting line is proportional to the number of shell genes shared between a pair of genomes.

557 recent and 457 ancient recombination events were identified in the core genome with fastGEAR. A recent recombination event is defined as a recombination event that affects one or multiple strains in a clade, but not all of them (which would

be an ancestral recombination event). The *pratensis* clade exhibited site-specific clustering when lineages were defined using recombination events, while the *griseus* strains clustered into a single lineage, indicating that recombination likely took place even after the migration event(s). Ancient recombination events support the idea that the *pratensis* clades once saw higher rates of recombination events, as evidenced by 457 recombination events detected between lineages 1 and 2 (WA-PR and RH-PR, respectively) as opposed to 20 recent events from WA-PR to RH-PR and 9 from RH-PR to WA-PR. Recombination with strains not analyzed in this dataset (labeled external recombination events in Fig. 2.3) also had a sizeable contribution to both genera; 37 for WA-GR, 10 for WA-PR, 1 for RH-PR and over 70% of all recent recombination events being between RH34 and external strains.

External recombination events usually involved other members of Actinobacteria, as detected by the HGTector algorithm. The most frequent recombination events were between donor strains in *Pseudonocardiaceae*, *Streptosporangiales*, *Actinomadura*, *Micromonospora*, *Actinoplanes, Amycolatopsis*, *Nocardiopsis*, and several other soil-dwelling bacteria in addition to very rare events that involve soil-dwelling Proteobacteria (eg. *Mesorhizobium*). We were especially interested in the RH34 genome because of its numerous external recombination events, and 15 of these fastGEAR-detected external recombination events were confirmed by HGTector, Most notable among these are antibiotic synthesis enzymes tetracenomycin-F1 monooxygenase from an unidentified bacterial donor, cephalosporin-C deacetylase from the *Micromonosporaceae*, as well as a membrane-associated carboxylesterase from the *Actinomadura*, an inner membrane protein and

melamine deaminase from unidentified Actinomycetia.

2.4.5 Gene acquisition and deletion over evolutionary time

Gene flux dynamics change as these 17 strains speciate and then form site-specific subclades. In particular, gene loss and transfers emerge as important factors that shape the pangenome at the tips of the tree, and these events are upto 9 times more frequent than gene loss or transfer along the internal nodes. Genome sizes are different between phylogroups (GR > PR, $p < 0.001$) and sites (WA > RH, $p < 0.05$), which could be a result of differing ancestral gene content within each species as well as demographic events local to strain habitat. In particular, RH-PR genomes have significantly smaller genomes compared to the other clades, and this size reduction seems to be driven by high gene loss.

Gene duplication events are lower by order of magnitude compared to loss and transfer events (Fig. 2.5), but duplications are of particular importance in identifying genes under positive selection or paralogs which are often associated with adaptive interactions with the environment (Bratlie et al. 2010). RH34 has an unusually high number of gene duplications compared to other strains, and it also has some of the highest copy numbers in individual gene duplication events, eg. one dimodular NRPS gene was duplicated 12 times. Some other genes were duplicated thrice; they encoded for DNA replication, transcription, and repair proteins (HTH-type transcriptional regulator *cdhR*, multifunctional non-homologous end-joining DNA repair gene *ligD*, replicative DNA helicase) and a gene encoding MbtH-like protein, which is known to mediate antibiotic and siderophore biosynthetic pathways in *Streptomyces coelicolor*

(Lautru et al. 2007). 35 of the duplicated genes in RH34 were identified as external recombination events by fastGEAR, including the multiple duplications described above. In addition, the amino acid carrier protein ALST, tetracenomycin-F1 monooxygenase, polyketide synthases PksM and PksL, melamine deaminase, linear gramicidin synthase subunit D, and several transport proteins (osmoprotectant import, oligopeptide transport, glutathione import, carnitine transport, and inner membrane transport proteins YdhP and YcjO) were duplicated and externally transferred.

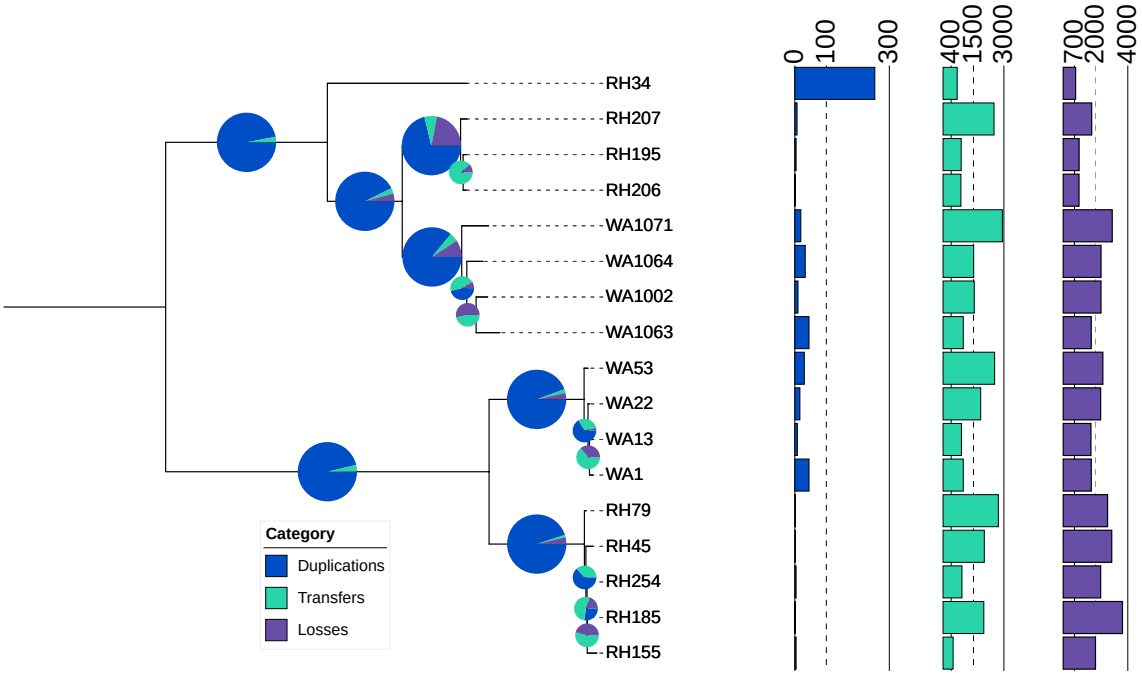

Figure 2.5 Gene duplications, losses, and transfers are represented as pie charts on the nodes and bar plots at the tips of the core gene tree. The number of gene loss and transfer events at each node are, on average, fewer than those seen at the tips while gene duplications are typically higher at the earlier nodes of the tree.

RH-PR strains experienced duplications in transposase domains, amidohydrolase family, haloacid dehalogenase-like hydrolase, and the putative peptidoglycan synthesis protein MviN; RH-GR in condensation domains (typically

found in enzymes that synthesize peptide antibiotic) in addition to the Lux regulon and GtrA-like protein (integral membrane components). Within WA, the PR group saw the highest number of annotated gene duplications across all clades including integrase elements, repeat domains in the Fasciclin I protein family, thermolysin propeptide motifs, a Tn3 transposase domain and another putative transposase of the IS4/5 family. WA-GR sees duplications in protein sequences annotated as permeases (amino acid, FtsX-like), periplasmic substrate-binding proteins, and a multiple antibiotic resistance protein.

A KEGG pathway analysis of genes lost in each clade revealed that WA-PR strains lost genes associated with ribosomal proteins *rplA*, *rplP*, *rpsR* (subunits L1, L16, S18), aminoacyl-tRNA biosynthesis (*glyS1* or glycyl-tRNA synthetase, *pheS* or phenylalanyl-tRNA synthetase), and the *aceE* gene for pyruvate dehydrogenase E1 (the entire gene cluster involved in pyruvate oxidation to acetyl-CoA was also lost, but other genes in this cluster were lost in other clades as well). The WA-GR clade also lost genes encoding several ribosomal proteins (*rplJ* for L10, *rplM* for L13), aminoacyl-tRNA biosynthesis genes *tyrS*, *lysK*, *lysS* (tyrosyl-tRNA synthetase, lysyl-tRNA synthetases, metabolic genes *leuB* and *pdhB* (3-isopropylmalate dehydrogenase in the leucine biosynthesis pathway, the beta subunit of the pyruvate dehydrogenase E1 component), and antibiotic resistance genes (streptomycin and chloramphenicol resistance). WA-GR also lost the genes required for pyridoxal phosphate biosynthesis (*pdxS* and *pdxT*), which is vital for vitamin B6 metabolism.

In Wisconsin, the RH-PR strains lost a gene involved in acyl-CoA synthesis (*fadD*), which is needed for beta oxidation in fatty acid metabolism. Ribosomal protein

genes *rplV*, *rpmG*, *rpsD* (L22, L33, S4) and aminoacyl-tRNA synthetase genes *trpS*, *alaS*, *metG*, *serS* (tryptophanyl-tRNA synthetase, alanyl-tRNA synthetase, methionyl-tRNA synthetase, seryl-tRNA synthetase) genes were also lost, similar to the WA clades. The RH-GR strains showed the least number of gene losses: genes *rplE* and *sucA* encoding ribosomal protein L5 and a component of 2-oxoglutarate dehydrogenase, an enzyme involved in the TCA cycle. All clades lost homologs of genes involved in biosynthesis of secondary metabolites (streptomycin, neomycin/kanamycin/gentamicin, prodigiosin, terpenoid backbone) and genes required for vitamin B (thiamine and biotin) metabolism.

WA-PR strains gained genes to encode ribosomal proteins L5 and S1, tryptophanyl-tRNA synthetase, and citrate synthase which is an essential part of the TCA cycle. Interestingly, WA-PR gained the gene cluster involved in acyl-CoA synthesis that was lost by RH-PR strains. WA-GR gained the *rplF* gene encoding ribosomal protein L6, a component of phenylalanyl-tRNA synthetase, and a dioxygenase that cleaves aromatic rings. All WA strains gained genes for ribosomal protein L36, S13, and S14 (*rpmJ*, *rpsM*, and *rpsN*), a fatty-acyl-CoA synthetase, a ligase involved in polyketide biosynthesis, various prenyltransferases that perform biosynthesis of secondary metabolites and cofactors, recombination-related genes (*dnaN*, *mutY*) as well as transporter proteins for taurine, phospholipids, iron-siderophores, vitamin B12, and ammonium ion channels.

RH-PR gained *rplQ* (L17) and *rplB* (L2) genes in addition to the gene encoding a lysine degradation enzyme (*sucB*). Finally, RH-GR gained genes *rplJ*, *rpmB*, *rpmG*, *rpmH*, *rpsE*, *rpsG* (L10, L28, L33, L34, S5, S7), *serS* (seryl-tRNA synthetase), *korB*,

pdhC, *leuC* (enzymes involved in leucine and isoleucine biosynthesis). All Wisconsin strains gained genes for asparagine synthesis (*asnB*), peptidoglycan flippase (*murJ*), pentalenic acid synthase for antibiotic production, *dnaX*, *ssb*, *sbcC*, *uvrA* genes for homologous recombination and repair, and transport proteins for osmoprotectants (*opuBD*), multiple sugars, nickel ions, a multiple antibiotic resistance protein (*marC*) and channels for urea and magnesium transport.

2.4.6 Secondary metabolite gene clusters are primarily vertically transmitted

Many SMGCs are found in all strains although the number of clusters within each type is highly variable even among closely related strains (Fig. 2.6). We also found that a number of biosynthetic gene clusters are restricted to a specific clade or site. Given that the production of certain secondary metabolites could provide an advantage in different environments, the differential distribution of secondary metabolites could be an indication of adaptation to ecological between phylogenetic clades.

Contrary to findings of positive selection pressure on some antibiotics in *Streptomyces* and *Salinispora* (Challis and Hopwood 2003; Penn et al. 2009), we found that all SMGCs shared by a phylogroup or all strains were under strong negative selection pressure (Table 2.4). Only sequences in siderophore biosynthetic gene clusters (BGCs) showed evidence of positive selection in the *griseus* clade: RH34 and WA1064 vs all other members of RH-GR (dN/dS = 2.79 and 1.32 respectively). Both sequences correspond to short ORFs upstream of IucA-IucC and aminotransferase domains that catalyze siderophore biosynthesis. Comparing the amino acid sequences

to known protein domains through InterProScan (Jones et al. 2014) revealed the closest hits to be non-cytoplasmic signal peptide domains.

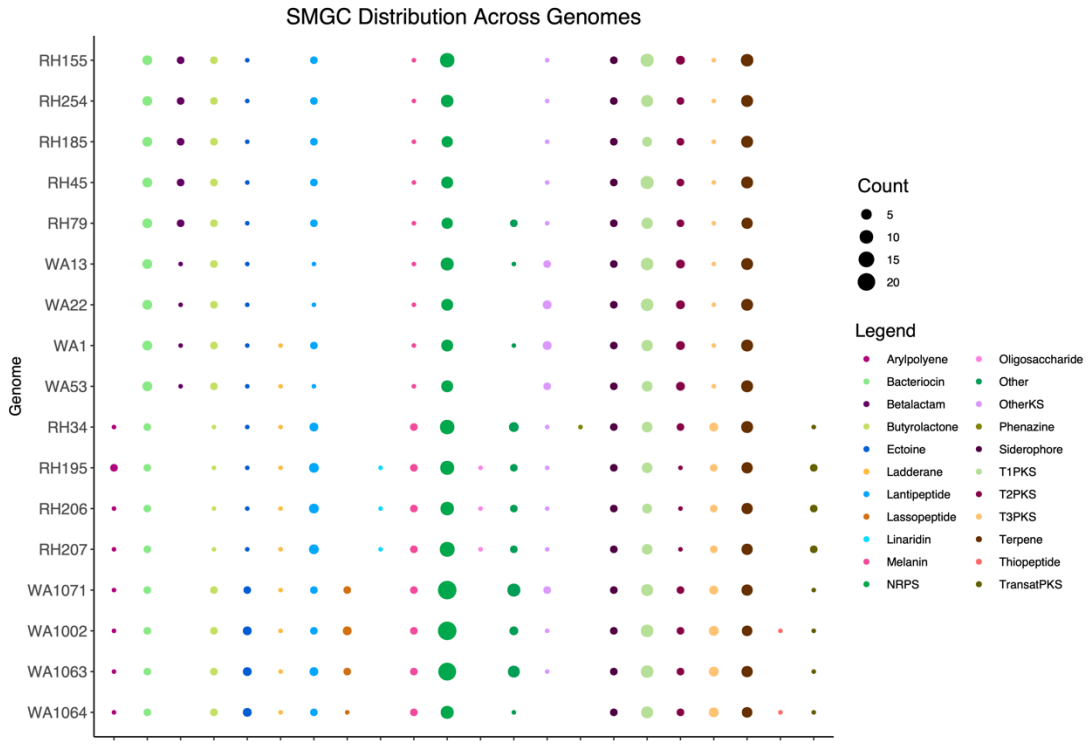

Figure 2.6 The bubble plot shows the number of gene clusters of each secondary metabolite category per genome, where size of the bubble is proportional to the number of clusters found in each strain. The y-axis labels are arranged in the same order as the tree in Fig. 2.1.

SMGC type	Clade	dN/dS range
Arylpolyene	GR	0.001 - 0.138
Bacteriocin	All	0.001 - 0.553
Betalactams	PR	0.001 - 0.072
Ectoine	All	0.001 - 0.495
Lantipeptides	All	0.001 - 0.551
Melanin	All	0.001 - 0.588
Siderophore	All	**0.035 - 2.795**
Terpene	All	0.001 - 0.797

Butyrolactone	All	0.001 - 0.164
Lassopeptides	WA-GR	0.001 - 0.252
Thiopeptides	WA-GR	0.001 - 0.22
NRPS	All	0.001 - 0.846
PKS	All	0.001 - 0.846

Table 2.4 Selection pressure (as measured by dN/dS ratio) for different SMGCs shared by all strains in a clade or the dataset. Presence of a specific clade in the second column indicates presence of a clade-specific SMGC, not that selection pressure exists between those strains.

SMGC trees were constructed through the maximum likelihood method and tree topologies were then analyzed for likelihood of HGT using the Shimodaira-Hasegawa test with 10,000 bootstrap replicates (Shimodaira and Hasegawa 1999) as implemented in the R package phangorn (Schliep 2011). First, homologs of genes present in all strains were used to reconstruct phylogeny for bacteriocin, lantipeptide, melanin, siderophore, and terpene gene clusters. Then, tree topologies were examined for individual, single-copy homologous genes or ORFs within each cluster. Two gene clusters (melanin and terpene) had topologies indicative of HGT ($p < 0.05$, 10000 bootstraps).

There was an increased likelihood of HGT even within SM gene clusters that were consistent with the core gene phylogeny. Sequences in the siderophore, terpene, and melanin gene clusters showed tree topologies inconsistent with vertical descent ($p < 0.05$, 10,000 bootstrap replicates, SH test). While two of these regions were short, unannotated ORFs, including a locus in the siderophore and melanin BGCs each, four adjoining protein sequences in a terpene BGC were identified as NRPS/PKS domains that could bind AMP, carrier proteins, and a thioesterase domain.

2.4.7 Genomes show signatures of conjugation and transduction

Identifying the mechanism of HGT (transformation, conjugation, or transduction) can be difficult in natural environments. The availability of draft genomes, however, allows us to identify and compare the genomic signatures of these processes. Core and accessory genes were scanned for evidence of mobile elements including plasmids, virulence factors, integrative and conjugative elements, insertion sequences as well as gene annotations that indicated the presence of transposons, phage elements, plasmids, conjugation elements, and integrons using keywords detailed elsewhere (Kent et al. 2020).

WA strains had more plasmids than those from RH ($p = 0.0005$; Wilcoxon rank-sum test) while RH strains had more insertion sequences ($p = 0.000059$; paired t-test). Virulence factors did not differ by site or phylogeny but the WA-GR clade had three unique virulence factors. Prophage elements, however, varied by phylogeny ($p = 0.0000228$; paired t-test); the GR strains had more prophage elements than PR.

Fig. 2.7 represents the number of phage genome fragments identified within each *Streptomyces* genome at 100% identity. While two actinophages (Beuffert and Sros11) are found at similar frequencies across all clades, clade membership has a clear influence on patterns of host-phage interactions. The WA-GR strains are particularly different compared to their phylogenetic and geographical neighbors – they lack 3 phage signatures that are found in all other genomes (OnionKnight, SPB78, VWB) and associate with a different phage that is unique to this clade (Bing). RH-GR strains also associate with two unique phages (JackieB and pZL12), thus independently validating the previous analysis that showed a higher number of annotated prophage

elements in *griseus* strains.

2.5 DISCUSSION

The strong geographical differentiation observed in these *Streptomyces* strains is in contrast to other bacteria in natural habitats, such as marine strains of *Vibrio cholerae*, which exhibit little biogeographic structure based on six representative core genes (Boucher et al. 2011).

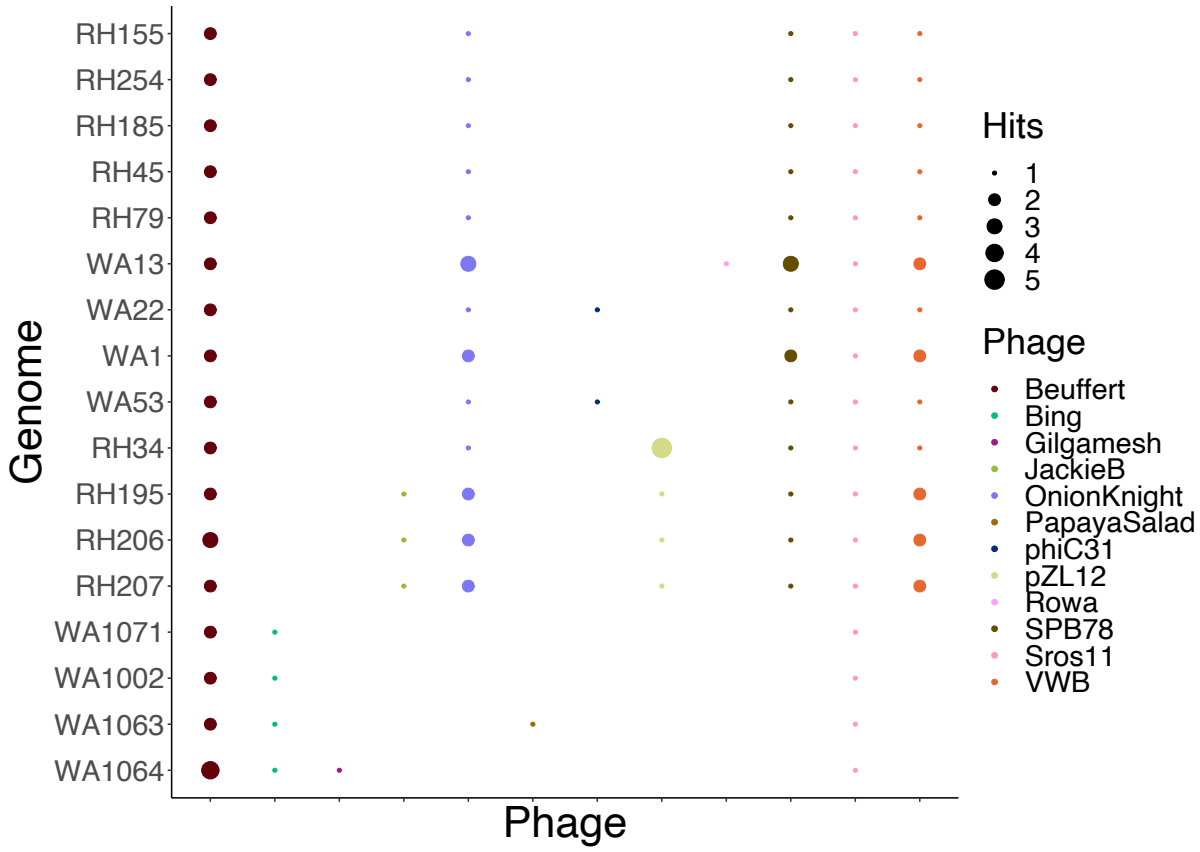

Figure 2.7 Actinophage sequences were distributed unevenly across all *Streptomyces* genomes. Bubble colors represent the phage sequence aligned to the *Streptomyces* strain on the y-axis, and the size of the bubble represents the number of unique hits from the phage. The WA-GR clade in particular showed a distinct phage element signature. Sequences from two phage genomes (Beuffert and Sros11) were present at 100% identity in all 17 strains.

Streptomyces have previously been identified to be dispersal limited at regional and local scales (Choudoir et al. 2016; Hariharan and Buckley 2022) and it is possible that the strains in this study experienced dispersal limitation following migration, which could contribute to reduced gene flow and the geographical structure seen in Fig. 2.1. Yet, the overall trends in gene content distribution suggest that recombination acts as a cohesive force that declines in strength with increasing sequence divergence (Fraser et al. 2007). Even within the same phylogroup, *griseus* strains are seen to have more gene flow, leading to the genetic cohesiveness seen in Fig. 2.3. This finding is supported by the ANI values in Table 2.2 where the average ANI between *griseus* strains is 97.82% ± 1.47 as compared to the average ANI between *pratensis* strains (96.41% ± 2.87).

A similar comparison within the species *S. flavogriseus* found that interspecies recombination rates were reduced by more than two orders of magnitude when compared with intraspecies recombination rates (Doroghazi and Buckley 2014). A similar trend is seen for the two species in this study; GR and PR strains share 13 recent recombination events as compared to 29 recent and 457 ancient events within PR. GR-GR recombination events are presumably more common (both ancient and recent) given that the strains cluster together as a lineage. The geographical split seen in the *pratensis* lineages could serve as an indicator of how isolation-by-distance might increase genetic diversity in bacterial populations and eventually lead to sister lineages diverging beyond the species boundary.

However, interspecies HGT events are not uncommon as seen in Fig. 2.3 and the potential external donors identified by HGTector. A previous study suggests that a reticulate network structure might be a better depiction of *Streptomyces* phylogenetic

relationships than a bifurcating tree (Doroghazi and Buckley 2014). Such reticulate evolutionary relationships could create fuzzy species boundaries within this genus, making gene exchange through homologous recombination favorable.

While shell gene content varied by phylogeny, strain-specific (unique) gene content varied by site. Given that these unique genes are only present in one genome and thus have not yet swept through the population, we propose that strain-specific genes are indicative of regular samplings of the local gene pool and thus representative of contemporary environmental conditions. Changes in more widely shared genes (i.e. core and shell) are likely integrated over much longer periods of evolutionary and geological history. Although accessory genes show localization by phylogeny and geography, several common functions are shared across clades. Annotations for clade-specific genes typically include amino acid biosynthesis or utilization, vitamin B metabolism, peptidoglycan biosynthesis, integrases and transposons, and enzymes involved in secondary metabolite production (particularly antibiotic synthesis and resistance). Several of these categories could improve a strain's ability for ecological differentiation and thus sweep through the local population. Future studies that investigate the adaptation of soil *Streptomyces* to new habitats should focus on how traits like switching amino acid synthesis pathways or finetuning the ability to use vitamin B could provide a competitive advantage.

Fig. 2.5 indicates differences in the tempo of gene gain and loss over evolutionary time as seen in the marked increase in gene losses at the tips. This pattern suggests the possibility of population expansion until more recent times, when gene loss becomes the dominant process that drives changes in genome size (eg. between

RH-GR and WA-GR or within RH-PR). Such an expansion would be consistent with the hypothesis that deglaciation drives differences in *Streptomyces* genome size (Choudoir et al. 2017).

Our genome size comparisons agree with previous observations that gene loss is more frequent than HGT, and that the phylogenetic distributions of the majority of protein families in a bacterial genome can be explained by vertical inheritance (Kunin and Ouzounis 2003). While the molecular signatures of genetic drift and selection can be especially hard to distinguish in free-living environmental strains, analysis of soil *Bradyrhizobium* genomes suggests that the strength of selection (both purifying and diversifying) is higher for accessory genes that have a high probability of being lost due to environmental stress like acidity or aridity (Simonsen 2021). Thus, signatures across multiple soil bacterial genomes suggest that gene loss can be used by these bacteria to reduce genome size when exposed to environmental stressors.

We expected that secondary metabolite gene clusters would follow site-specific distribution patterns as well, given their potential for adaptation to local environmental conditions and interactions with neighboring strains. Fig. 2.6 indicates that many detected SMGCs were present in all strains, and were hence derived from a common ancestor of both *S. griseus* and *pratensis*.

While there are some species-specific SMGCs and even some clade-specific ones, there are no SMGCs unique to either site. Some SM clusters are unique to specific strains, however, indicating the possibility of horizontal acquisition. Several NRPS- and PKS-linked clusters are unique to their genomes; for eg., RH155 and RH185 have a unique betalactam-T1PKS and betalactam-T1PKS-NRPS cluster each.

HGT has previously been identified as an important force shaping *Streptomyces* genomes, and is one of the proposed mechanisms to explain the diversity of biosynthetic pathways in this genus and the closely related *Salinispora* (Egan et al. 2001; Penn et al. 2009). An analysis of 122 *Streptomyces* genomes revealed that genes involved in secondary metabolism and xenobiotic metabolism were over-represented in HGT events while core biological functions like transcription and translation were under-represented, likely due to stronger selection against transferred core genes (McDonald and Currie 2017). Gene gain and loss are both major drivers of SMGC biodiversity, resulting in considerable variability within SMGCs inherited from a shared ancestor (Choudoir et al. 2018).

Finally, the possibility of phylogenetic bias in horizontal transfer cannot be ruled out. Following the tendency of similar genomes to exchange genes more frequently, SMGC distribution patterns could resemble vertical inheritance despite the presence of frequent HGT (McDonald and Currie 2017). Tidjani et al. (Tidjani et al. 2019) proposed that gene flux in *Streptomyces* can favor rapid acquisition or dissemination of BGCs within the population. In this framework, SMGCs are still transferred horizontally but within a genetically cohesive population cluster. Rapid turnover of BGCs would constantly modulate the interactions within the population and thus, favor the social cohesiveness of said population. In summary, SMGCs in *Streptomyces* strains are likely transferred both horizontally and vertically, as seen in *Streptomyces* with rugose-ornamented spores (Chung et al. 2021).

The mechanisms of HGT in environmental microbes are hard to identify. Across the prokaryotic world, large recombination events tend to be associated with

conjugation mechanisms and positive selection, both of which were observed in this dataset (González-Torres et al. 2019). Conjugation has been identified as the primary driver of indel flux and diversity in *Streptomyces* (Tidjani et al. 2019), with several hotspots located in the central region of the chromosome. The same study also identified an uneven indel distribution across the linear chromosome with highest variability towards the ends of the arms, which is typically made of species-specific genes (Kirby 2011; Choulet et al. 2006).

Examining these *Streptomyces* genomes confirms at the importance of conjugation and transduction for HGT in soil *Streptomyces*. Plasmids and actinophages in particular could be important drivers of HGT. While plasmids seem to be environmentally acquired, phage associations show species-level specificity. Many bacteriophages carry accessory genes that serve adaptive functions and are under purifying selection in bacterial genomes. In some enteric bacteria, genomes are thought to be in constant contact with closely related phages, with phage-derived genes frequently replaced by other incoming prophage-derived genes (Bobay et al. 2014). We are still discovering the host specificity and prevalence of actinophages in the soil, but such experiments will be essential in understanding how gene content is shared among *Streptomyces* and other soil bacteria.

Future work that focuses on detecting HGT across an environmental gradient will contribute to a deeper understanding of how HGT varies with different abiotic factors and across gradual shifts in microhabitats. In particular, such experimental designs should account for the different temporal scales at which phylogenetic and geographical barriers to gene exchange may operate. Different environmental gradients

(eg. a latitudinal temperature gradient versus a precipitation gradient) may result in varying rates of gene transfer and disparity in the transferred gene content. Analysis of accessory gene content and pangenome variation will illuminate the contribution of HGT to bacterial adaptation, niche partitioning, and ecotype differentiation in soil. Finally, experiments that identify the diversity of phages in the local bacterial community and plasmids or other conjugative elements in the local gene pool will improve our mechanistic understanding of HGT. Horizontal gene transfer clearly at multiple temporal and spatial scales and contributes to the richness of bacterial species and lifestyles in the environment. It is a vital ecological and evolutionary force that creates differentiation or cohesion between lineages. A better understanding of HGT dynamics in the environment will ultimately lead to a deeper understanding of bacterial speciation and taxonomic delineations as well.

REFERENCES

Aminov RI. (2010) A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Frontiers in Microbiology* 1. https://doi.org/10.3389/fmicb.2010.00134.

Andam CP, Doroghazi JR, Campbell AN, Kelly PJ, Choudoir MJ, Buckley DH. (2016) A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus *Streptomyces*. *mBio* 7 (2): e02200-02215. https://doi.org/10.1128/mBio.02200-15.

Anderson AS, Wellington EM. (2001) The Taxonomy of *Streptomyces* and Related Genera. *International Journal of Systematic and Evolutionary Microbiology* 51 (Pt 3): 797–814. https://doi.org/10.1099/00207713-51-3-797.

Angiuoli SV, Salzberg SL. (2011) Mugsy: Fast Multiple Alignment of Closely Related Whole Genomes. *Bioinformatics (Oxford, England)* 27 (3): 334–42. https://doi.org/10.1093/bioinformatics/btq665.

Bérdy J. (2005) Bioactive Microbial Metabolites: A Personal View. *The Journal of Antibiotics* 58 (1): 1–26. https://doi.org/10.1038/ja.2005.1.

Bobay L, Touchon M, Rocha EPC. (2014) Pervasive Domestication of Defective Prophages by Bacteria. *Proceedings of the National Academy of Sciences* 111 (33): 12127–32. https://doi.org/10.1073/pnas.1405336111.

Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF. (2011). Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global *Vibrio cholerae* Populations. *mBio* 2 (2): e00335-10. https://doi.org/10.1128/mBio.00335-10.

Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. (2010) Gene Duplications in Prokaryotes Can Be Associated with Environmental Adaptation. *BMC Genomics* 11 (1): 588. https://doi.org/10.1186/1471-2164-11-588.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. (2021) EggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. https://doi.org/10.1101/2021.06.03.446934.

Challis GL, Hopwood DA. (2003) Synergy and Contingency as Driving Forces for the Evolution of Multiple Secondary Metabolite Production by *Streptomyces* Species. *Proceedings of the National Academy of Sciences* 100 (suppl_2): 14555–61. https://doi.org/10.1073/pnas.1934677100.

Chase JM, McGill BJ, Thompson PL, Antão LH, Bates AE, Blowes SA, Dornelas M, Gonzalez A, Magurran AE, Supp SR, Winter M, Bjorkman AD, Bruelheide H, Byrnes JEK, Cabral JS, Elahi R, Gomez C, Guzman HM, Isbell F, Myers-Smith IH, Jones HP,

Hines J, Vellend M, Waldock C, O'Connor M. (2019) Species Richness Change across Spatial Scales. *Oikos* 128 (8): 1079–91. https://doi.org/10.1111/oik.05968.

Choudoir, MJ, Buckley DH. (2018) Phylogenetic Conservatism of Thermal Traits Explains Dispersal Limitation and Genomic Differentiation of *Streptomyces* Sister-Taxa. *The ISME Journal* 12 (9): 2176–86. https://doi.org/10.1038/s41396-018-0180-3.

Choudoir MJ, Doroghazi JR, Buckley DH. (2016) Latitude Delineates Patterns of Biogeography in Terrestrial *Streptomyces*. *Environmental Microbiology* 18 (12): 4931–45. https://doi.org/10.1111/1462-2920.13420.

Choudoir MJ, Panke-Buisse K, Andam CP, Buckley DH. (2017) Genome Surfing As Driver of Microbial Genomic Diversity. *Trends in Microbiology* 25 (8): 624–36. https://doi.org/10.1016/j.tim.2017.02.006.

Choudoir MJ, Pepe-Ranney C, Buckley DH. (2018) Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in *Streptomyces*. *Antibiotics* 7 (1): 12. https://doi.org/10.3390/antibiotics7010012.

Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou F, Fourrier C, Guérineau M, Decaris B, Barbe V, Pernodet J, Leblond P. (2006) Evolution of the Terminal Regions of the *Streptomyces* Linear Chromosome.

Chung Y, Kim H, Ji C, Je H, Lee D, Shim SH, Joo H, Kang H. (2021) Comparative Genomics Reveals a Remarkable Biosynthetic Potential of the *Streptomyces* Phylogenetic Lineage Associated with Rugose-Ornamented Spores. *mSystems* 6 (4): e00489-21. https://doi.org/10.1128/mSystems.00489-21.

Cordero OX, Hogeweg P. 2009. The Impact of Long-Distance Horizontal Gene Transfer on Prokaryotic Genome Size. *Proceedings of the National Academy of Sciences* 106 (51): 21748–53. https://doi.org/10.1073/pnas.0907584106.

Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF. (2012) Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. *Science (New York, N.Y.)* 337 (6099): 1228–31. https://doi.org/10.1126/science.1219385.

Didelot X, Wilson DJ. (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology* 11 (2): e1004041. https://doi.org/10.1371/journal.pcbi.1004041.

Doroghazi JR, Buckley DH. (2010) Widespread Homologous Recombination within and between *Streptomyces* Species. *The ISME Journal* 4 (9): 1136–43. https://doi.org/10.1038/ismej.2010.45.

Doroghazi JR, Buckley DH. (2014) Intraspecies Comparison of *Streptomyces oratensis* Genomes Reveals High Levels of Recombination and Gene Conservation

between Strains of Disparate Geographic Origin. *BMC Genomics* 15 (1): 1–14. https://doi.org/10.1186/1471-2164-15-970.

Edgar RC. (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Egan S, Wiener P, Kallifidas D, Wellington EMH. (2001). Phylogeny of Streptomyces Species and Evidence for Horizontal Transfer of Entire and Partial Antibiotic Gene Clusters. *Antonie van Leeuwenhoek* 79(2): 127-33. https://doi.org/10.1023/a:1010296220929

Emms DM, Kelly S. (2019) OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biology* 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the Nature of Bacterial Speciation. *Science (New York, N.Y.)* 315 (5811): 476–80. https://doi.org/10.1126/science.1127573.

Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D, Koonin EV. (2021) COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens. *Nucleic Acids Research* 49 (D1): D274–81. https://doi.org/10.1093/nar/gkaa1018.

González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. (2019) Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. *mBio* 10 (1). https://doi.org/10.1128/mBio.02494-18.

Greenlon A, Peter CL, Damtew ZM, Muleta A, Carrasquilla-Garcia N, Kim D, Nguyen HP, Suryawanshi V, Krieg CP, Yadav SK, Patel JS, Mukherjee A, Udupa S, Benjelloun I, Thami-Alami I, Yasin M, Patil B, Singh S, Sarma BK, von Wettberg EJB, Kahraman A, Bukun B, Assefa F, Tesfaye K, Fikre A, Cook DR. (2019) Global-Level Population Genomics Reveals Differential Effects of Geography and Phylogeny on Horizontal Gene Transfer in Soil Bacteria. *Proceedings of the National Academy of Sciences* 116 (30): 5200–15209. https://doi.org/10.1073/pnas.1900056116.

Hariharan J, Buckley DH. (2022) Elevational Gradients Impose Dispersal Limitation on *Streptomyces*. *Frontiers in Microbiology* 13 (May): 856263. https://doi.org/10.3389/fmicb.2022.856263.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. (2018) High-Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nature Communications* 9: 5114. https://doi.org/10.1101/225342.

Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M,

Yong S, Lopez R, Hunter S. (2014) InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Kent AG, Vill AC, Shi Q, Satlin MJ, Brito IL. (2020) Widespread Transfer of Mobile Antibiotic Resistance Genes within Individual Gut Microbiomes Revealed through Bacterial Hi-C. *Nature Communications* 11 (1): 4379. https://doi.org/10.1038/s41467-020-18164-7.

Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA. (2000) *Practical Streptomyces Genetics*. Norwich: John Innes Foundation.

Kirby R. (2011) Chromosome diversity and similarity within the *Actinomycetales*. *FEMS Microbiology Letters* 319 (1): 1–10. https://doi.org/10.1111/j.1574-6968.2011.02242.x

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. (2011) Genomic Fluidity: An Integrative View of Gene Diversity within Microbial Populations. *BMC Genomics* 12 (1): 32. https://doi.org/10.1186/1471-2164-12-32.

Kunin V, Ouzounis CA. (2003) The Balance of Driving Forces During Genome Evolution in Prokaryotes. *Genome Research* 13 (7): 1589–94. https://doi.org/10.1101/gr.1092603.

Labeda DP, Dunlap CA, Rong X, Huang Y, Doroghazi JR, Ju K, Metcalf WW. (2017) Phylogenetic Relationships in the Family *Streptomycetaceae* Using Multi-Locus Sequence Analysis. *Antonie van Leeuwenhoek* 110 (4): 563–83. https://doi.org/10.1007/s10482-016-0824-0.

Lautru S, Oves-Costales D, Pernodet J, Challis GL. (2007) MbtH-like Protein-Mediated Cross-Talk between Non-Ribosomal Peptide Antibiotic and Siderophore Biosynthetic Pathways in *Streptomyces coelicolor* M145. *Microbiology* 153 (5): 1405–12. https://doi.org/10.1099/mic.0.2006/003145-0.

Leplae R, Lima-Mendez G, Toussaint A. (2010) ACLAME: A CLAssification of Mobile Genetic Elements, Update 2010. *Nucleic Acids Research* 38 (Database issue): D57–61. https://doi.org/10.1093/nar/gkp938.

Lerminiaux NA, Cameron ADS. (2019) Horizontal Transfer of Antibiotic Resistance Genes in Clinical Environments. *Canadian Journal of Microbiology* 65 (1): 34–44. https://doi.org/10.1139/cjm-2018-0275.

Li L, Stoeckert CJ, Roos DS. (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* 13 (9): 2178–89. https://doi.org/10.1101/gr.1224503.

Li Y, Pinto-Tomás AA, Rong X, Cheng K, Liu M, Huang Y. (2019) Population Genomics Insights into Adaptive Evolution and Ecological Differentiation in *Streptomycetes*. *Applied and Environmental Microbiology* 85 (7). https://doi.org/10.1128/AEM.02555-18.

Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou H. (2019) ICEberg 2.0: An Updated Database of Bacterial Integrative and Conjugative Elements. *Nucleic Acids Research* 47 (D1): D660–65. https://doi.org/10.1093/nar/gky1123.

Lobkovsky AE, Wolf YI, Koonin EV. (2013) Gene Frequency Distributions Reject a Neutral Model of Genome Evolution. *Genome Biology and Evolution* 5 (1): 233–42. https://doi.org/10.1093/gbe/evt002.

Macías LG, Barrio E, Toft C. (2020) GWideCodeML: A Python Package for Testing Evolutionary Hypotheses at the Genome-Wide Level. *G3 Genes|Genomes|Genetics* 10 (12): 4369–72. https://doi.org/10.1534/g3.120.401874.

Manteca A, Sanchez J. (2009) *Streptomyces* Development in Colonies and Soils. *Applied and Environmental Microbiology* 75 (9): 2920–24. https://doi.org/10.1128/AEM.02288-08.

Martiny AC, Tai APK, Veneziano D, Primeau F, Chisholm SW. (2009) Taxonomic Resolution, Ecotypes and the Biogeography of *Prochlorococcus*. *Environmental Microbiology* 11 (4): 823–32. https://doi.org/10.1111/j.1462-2920.2008.01803.x.

McDonald BR, Currie CR. (2017) Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus *Streptomyces*. *mBio* 8 (3): e00644-17. https://doi.org/10.1128/mBio.00644-17.

Mi H, Ebert D, Muruganujan A, Mills C, Albou L, Mushayamaha T, Thomas PD. (2021) PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API. *Nucleic Acids Research* 49 (D1): D394–403. https://doi.org/10.1093/nar/gkaa1106.

Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. (2017) Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Molecular Biology and Evolution* 34 (5): 1167–82. https://doi.org/10.1093/molbev/msx066.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. (2015) Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* 31 (22): 3691–93. https://doi.org/10.1093/bioinformatics/btv421.

Pang TY, Lercher MJ. (2019) Each of 3,323 Metabolic Innovations in the Evolution of *E. coli* Arose through the Horizontal Transfer of a Single DNA Segment. *Proceedings*

of the National Academy of Sciences of the United States of America 116 (1): 187–92. https://doi.org/10.1073/pnas.1718997115.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR. (2009) Genomic Islands Link Secondary Metabolism to Functional Adaptation in Marine Actinobacteria. *The ISME Journal* 3 (10): 1193–1203. https://doi.org/10.1038/ismej.2009.58.

Rosen MJ, Davison M, Bhaya D, Fisher DS. (2015) Microbial Diversity. Fine-Scale Diversity and Extensive Recombination in a Quasisexual Bacterial Population Occupying a Broad Niche. *Science (New York, N.Y.)* 348 (6238): 1019–23. https://doi.org/10.1126/science.aaa4456.

Russell DA, Hatfull GF. (2017) PhagesDB: The Actinobacteriophage Database. *Bioinformatics* 33 (5): 784–86. https://doi.org/10.1093/bioinformatics/btw711.

Schliep KP. (2011) Phangorn: Phylogenetic Analysis in R. *Bioinformatics* 27 (4): 592–93. https://doi.org/10.1093/bioinformatics/btq706.

Seemann T. (2014) Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. (2008) Convergence of *Campylobacter* Species: Implications for Bacterial Evolution. *Science* 320 (5873): 237–39. https://doi.org/10.1126/science.1155532.

Shimodaira H, Hasegawa M. (1999) Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution* 16 (8): 1114–16. https://doi.org/10.1093/oxfordjournals.molbev.a026201.

Simonsen AK. (2021) Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *The ISME Journal* 16:423–434. https://doi.org/10.1038/s41396-021-01082-x

Snipen L, Liland KH. (2015) Micropan: An R-Package for Microbial Pan-Genomics. *BMC Bioinformatics* 16 (1): 79. https://doi.org/10.1186/s12859-015-0517-0.

Stamatakis A. (2014) RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. (2012) Inferring Duplications, Losses, Transfers and Incomplete Lineage Sorting with Nonbinary Species Trees. *Bioinformatics* 28 (18): i409–15. https://doi.org/10.1093/bioinformatics/bts386.

Suyama M, Torrents D, Bork P. (2006) PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments. *Nucleic Acids Research* 34: W609-612. https://doi.org/10.1093/nar/gkl315.

Tidjani A, Lorenzi J, Toussaint M, van Dijk E, Naquin D, Lespinet O, Bontemps C, Leblond P. (2019) Massive Gene Flux Drives Genome Diversity between Sympatric *Streptomyces* Conspecifics. *mBio* 10 (5): e01533-19. https://doi.org/10.1128/mBio.01533-19.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, El Karoui M, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EPC, Denamur E. (2009) Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLoS Genetics* 5 (1): e1000344. https://doi.org/10.1371/journal.pgen.1000344.

Tritt A, Eisen JA, Facciotti MT, Darling AE. (2012) An Integrated Pipeline for de Novo Assembly of Microbial Genomes. *PLOS ONE* 7 (9): e42304. https://doi.org/10.1371/journal.pone.0042304.

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH. (2015) antiSMASH 3.0–a Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters. *Nucleic Acids Research* 43 (W1): W237-243. https://doi.org/10.1093/nar/gkv437.

Whitaker RJ, Grogan DW, Taylor JW. (2003) Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. *Science (New York, N.Y.)* 301 (5635): 976–78. https://doi.org/10.1126/science.1086909.

Wiedenbeck J, Cohan FM. (2011) Origins of Bacterial Diversity through Horizontal Genetic Transfer and Adaptation to New Ecological Niches. *FEMS Microbiology Reviews* 35 (5): 957–76. https://doi.org/10.1111/j.1574-6976.2011.00292.x.

Woods LC, Gorrell RJ, Taylor F, Connallon T, Kwok T, McDonald MJ. (2020) Horizontal Gene Transfer Potentiates Adaptation by Reducing Selective Constraints on the Spread of Genetic Variation. *Proceedings of the National Academy of Sciences* 117 (43): 26868–75. https://doi.org/10.1073/pnas.2005331117.

Xie Z, Tang H. (2017) ISEScan: Automated Identification of Insertion Sequence Elements in Prokaryotic Genomes. *Bioinformatics* 33 (21): 3340–47. https://doi.org/10.1093/bioinformatics/btx433.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

Yeager CM, Gallegos-Graves LV, Dunbar J, Hesse CN, Daligault H, Kuske CR. (2017) Polysaccharide Degradation Capability of *Actinomycetales* Soil Isolates from a Semiarid Grassland of the Colorado Plateau. *Applied and Environmental Microbiology* 83 (6): e03020-16. https://doi.org/10.1128/AEM.03020-16.

Zhu Q, Kosoy M, Dittmar K. (2014) HGTector: An Automated Method Facilitating Genome-Wide Discovery of Putative Horizontal Gene Transfers. *BMC Genomics* 15 (1): 1–18. https://doi.org/10.1186/1471-2164-15-717.

CHAPTER 3

MIGRATION DIFFERENTIATES POPULATIONS WHILE MAINTAINING GENETIC COHESION IN CONSPECIFIC *STREPTOMYCES* STRAINS

3.1 INTRODUCTION

The effects of landscape structure on biogeography and dispersal have been examined for various plant, animal and insect species (Cristofoli et al., 2010; de la Giroday et al., 2011; Razgour et al., 2014) but only a handful of studies have been conducted on the effects of landscape structure and geographical barriers on the movement of environmental microbes (Whitaker et al., 2003; Ho et al., 2017). The soil microbiome is thought to be incredibly complex, diverse and a major force in mitigating climate change (Dubey et al. 2019) and it is thus essential that we understand the microorganisms that power this system. Distances of a few metres in a heterogenous soil environment could represent a significant dispersal barrier to some terrestrial bacteria (Oda et al., 2003) but the effects of spatial separation on closely related populations of soil bacteria are still unclear. Lack of specific information on range limits or dispersal mechanisms for bacteria (Wang and Soininen, 2017) often lead to difficulties in defining spatial scales for ecological and evolutionary processes or determining which landscape features could be impediments to gene flow between populations.

Continental-scale studies on the role of spatial scale in bacterial biogeography

indicate that bacteria have different biogeographical patterns than fungi or archaea (Ma et al., 2017) and edaphic variables like soil pH and land management influence bacterial communities (Fierer and Jackson, 2006; Karimi et al., 2018; Liu et al., 2018). However, existing literature only underscores that further research needs to be done to unravel the complexity of soil bacterial assemblages (Nemergut et al., 2011). Of particular interest is the fate of closely related strains that migrate to new habitats.

As in larger organisms, isolation-by-distance (IBD) has been reported in bacterial biogeography (i.e. genetic differentiation increases with geographic distance). Drift or selection could result in population structure, leading to reduced gene flow and allopatric speciation. Alternatively, ecological differentiation could result in barriers to gene flow and is known as isolation-by-environment (IBE) (Wang et al., 2022). Further studies in highly related strains collected from various habitats and geographic regions are necessary to explore how environmental selection and geographic isolation contribute to population differentiation.

The magnitude of gene flow between microbial populations is shaped predominantly by the genetic similarity and ecological overlap of the individual strains that make up those populations. The likelihood of transfer increases with greater physical contact between strains that occupy similar physical niches, but genetic relatedness also plays an important role in successful gene flow. (VanInsberghe et al., 2020).

Combinations of factors like dispersal limitation followed by ecological drift, or natural selection driven by differences across an environmental gradient could lead to changes in the genomes of soil bacteria. Examining the genomic signatures of these

processes can offer unprecedented insight into the interplay between biogeography and the evolution of the pangenome. For instance, both dispersal limitation and adaptation to local environments through their flexible genomes contribute to population structure in *Curtobacterium* strains (Chase et al., 2019).

The genus *Streptomyces* (phylum Actinobacteria) is a highly diverse group found predominantly in the soil and decaying vegetation. They play an important role in carbon cycling through degradation of complex carbon substrates like cellulose, hemicellulose, and chitin (Yeager et al., 2017). Unusually for bacteria, members of this genus are characterized by a complex secondary metabolism and a fungal-like morphological differentiation involving the formation of branching, filamentous substrate mycelia, and aerial hyphae bearing long spores (Manteca and Sanchez, 2009). They are also a major source of naturally derived antimicrobials and numerous bioactive, pharmaceutically relevant metabolites (Bérdy, 2005). Their cosmopolitan lifestyles, wide geographical distribution, and potential for adaptation to new habitats make *Streptomyces* ideal candidates to study isolation-by-distance in soil bacteria.

Previous work on *Streptomyces* ecology indicates the role of a latitudinal gradient in shaping diversification of gene content and phylogenetic relationships across *Streptomyces* genomes across the USA (Andam et al., 2016; Choudoir et al., 2017, 2018; Choudoir and Buckley, 2018). Briefly, these studies found that (a) dispersal limitation was responsible for structuring *Streptomyces* communities at regional scales, (b) historical deglaciation events could have contributed to diversification of gene content in the *Streptomyces* pangenome and (c) *Streptomyces* populations are differentiated on the basis of phylogenetically conserved traits

associated with thermal tolerance.

It has also been suggested that *Streptomyces* may not be ecologically stable in the soil and they undergo rapid adaptation and local selection (Traxler and Kolter, 2015). Large bacterial genomes are also thought to have a larger probability of being acted upon by purifying selection (Touchon and Rocha, 2016), leading to the intriguing possibility that *Streptomyces* are capable of rapid, long-range dissemination and adaptation to a variety of habitats. However, the effect of recent migration on the *Streptomyces* pangenome is still unknown, as is the potential for ecotype differentiation in new habitats. While sequence-discrete bacterial populations exist in nature, sampling from different ecological niches might confound the idea of clear species delineations in environmental populations (Caro-Quintero and Konstantinidis, 2012).

Population genetics theory and tools can be used to infer how dispersal shapes population structure across a landscape in *Streptomyces*. Quantifying genetic diversity and differentiation at individual sites is an effective way of understanding the process of effective dispersal across landscapes (Broquet and Petit, 2009). Analysis of pangenome content at these large scales will provide unique information on adaptation strategies and evidence of admixture and divergence between allopatric populations.

In this study, we analyze the pangenomes of 32 conspecific *Streptomyces* strains isolated from three sites in Alaska and Washington states. The Gulf of Alaska provides a natural barrier to gene flow between sites in Alaska and Washington, while a much smaller geographical distance separates the two sites in Alaska (DEN and MAN). Environmental conditions vary across sites, making this experimental design

effective in examining the effects of isolation-by-distance and isolation-by-environment on these bacterial genomes.

We hypothesize that geographically proximal *Streptomyces* strains in Alaska would display greater genetic cohesion and increased gene flow, with population differentiation increasing as spatial distance increases. Habitat similarity is expected to be the most important driver of the acquisition of accessory genes that aid in local adaptation (versus shared gene content from a common ancestor). Thus, *Streptomyces* strains occupying more similar habitats are likely to share a greater proportion of accessory genes and functional annotations.

3.2 METHODS

3.2.1 Bacterial isolates and DNA extraction

A total of 8 DEN, 11 MAN, and 13 WA strains were selected from an existing culture collection of more than 1,000 *Streptomyces* from sites across the United States. Inceptisol soils from perennial grass rhizosphere were collected at 0-5 cm depth and air-dried, following which isolation of *Streptomyces* was carried out on glycerol-arginine media of pH 8.7 containing cycloheximide (300 mg/L) and rose bengal (35 mg/L) as previously described (Doroghazi and Buckley 2010). The genetic diversity of isolates in this collection was initially assessed using partial *rpoB* sequences. For this study, we selected conspecific strains (96% < ANI < 100%) such that this group could be called a population capable of gene exchange. The strains were most closely related to *Streptomyces sp.* M41 (97.38% ANI; NCBI BioSample ID SAMN06311458) and the closest named species was *Streptomyces populi* (88.31% ANI). Geographical metadata for the three sites can be found in Table 3.1.

Site	Latitude	Longitude	pH	% Organic Matter	Average Temperature (ºC)	Annual Precipitation (cm)
Manley Hot Springs, AK (man)	63.87	-149.02	5	9.96	-3.58	58.22
Denali Highway, AK (den)	63.22	-147.68	6	1.89	-3.92	61.29
Bothell, WA (wa)	47.73	-122.24	5.31	11.34	11.36	83.1

Table 3.1 Geographical and environmental data for study sites in Alaska and Washington

DNA was extracted from purified cultures, which were grown by shaking at 30°C in liquid yeast extract-malt extract medium (YEME) containing 0.5% glycine (Kieser et al. 2000), by using a standard phenol/chloroform/isoamyl alcohol protocol.

3.2.2 Whole genome sequencing, annotation, and assembly

DNA libraries were prepared using the Nextera XT library preparation kit and all libraries were multiplexed and run together on an Illumina MiSeq instrument at the Cornell Biotechnology Resource Center. Demultiplexed and quality-filtered raw reads were assembled *de novo* using the SPADES assembly pipeline (Prjibelski et al. 2020). Annotations of the draft genomes were done using Prokka (Seemann, 2014). Assembly statistics for each genome are found in Table 3.2.

	Coverage	**# contigs**	**N50**	**Size (Mb)**	**Completeness (%)**
den1	20	4798	2735	8.91	95.1
den10	25	3205	4924	9.25	98.2
den12	23	4937	2685	8.79	95.5

den28	35	1447	11142	8.82	99.6
den30	7	1950	9001	9.58	99.4
den31	14	1794	18794	9.95	100
den35	22	3149	5236	9.37	99
den41	42	1824	9886	9.04	99.9
man116	39	2034	8711	9.56	99.4
man121	25	1634	9811	8.97	99.2
man143	21	2874	5019	9.04	97.8
man150	31	1885	8831	9.16	99.2
man153	24	3878	3803	9.03	98.5
man156	31	1617	11189	9.36	99.9
man163	31	1548	10174	8.99	99.7
man186	23	2221	6731	8.83	99.5
man25	22	4331	3976	9.75	97.6
man3	24	1984	7576	8.81	99.9
man34	23	2645	7140	9.23	99.9
wa1083	11	4040	3226	8.2	96.5
wa1098	33	1660	9816	8.98	99
wa18	20	3538	4083	8.78	98.3
wa19	6	3015	6126	9.61	98.8
wa21	20	2959	4784	8.77	98.7
wa30	23	2720	5632	8.86	98.6
wa31	46	1404	13926	9.26	99.8
wa32	15	2642	5921	8.87	98.6
wa34	27	2291	6194	8.79	95.8
wa44	31	1886	8347	8.99	98.7
wa46	81	1721	14723	9.6	99.8
wa60	32	1688	9741	9.4	99.8
wa9	19	2229	7376	8.64	98.8

Table 3.2 Genome assembly statistics

3.2.3 Variant calling and SNP filtering

Streptomyces avermitilis MA-4680 was chosen as a reference genome for

variant calling given that it was the closest validly named relative (average ANI 85%) with a complete genome. Quality-filtered and trimmed reads from each isolate were then aligned to this reference FASTA file (GCA_000009765; NCBI Genome) using the Burrows-Wheeler aligner (Li, 2013). Picard tools (http://broadinstitute.github.io/picard) were used to remove duplicate alignments, and subsequently sort, index, and validate the deduplicated BAM files. Variants (SNPS, MNPs, indels and complex variants) were called using the freeBayes algorithm (Garrison and Marth, 2012). Variants were filtered using VcfFilter (https://github.com/biopet/vcffilter) based on their quality scores, total read depths at a given locus, and number of samples without missing data (parameters QUAL > 100, DP > 50, NS > 4).

The ADMIXTURE software tool was used for maximum-likelihood ancestry estimation using the filtered SNPs (Alexander et al., 2009). The R package PopGenome was used to estimate nucleotide F_{ST} for pairs of populations (Pfeifer et al., 2014).

3.2.4 Comparative genomics

Pangenome content was analyzed using Roary (Page et al., 2015), and the gene presence-absence table generated by Roary was subsequently used to identify genes unique to individual populations using Scoary (Byrnildsrud et al., 2016) with 1000 permutations. Genes were identified as unique to a specific population if the empirical p-value from the permutations was lower than 0.01 and the odds ratio was greater than 2. PopCOGenT was used to identify recent population differentiation based on regions

of nucleotide similarity (Arevalo et al., 2019).

3.2.5 Horizontal gene transfer detection

Draft genomes were aligned using Mugsy (Angiuoli and Salzberg, 2011), and a genome-based phylogeny was reconstructed using the maximum likelihood method as implemented in RAxML v8 (Stamatakis, 2014). Effective recombination rates (r/m; recombination rate accounting for mutations) within populations were calculated using ClonalFrameML (Didelot and Wilson, 2015). fastGEAR (Mostowy et al., 2017) was used to identify ancient and recent recombination events as well as lineage probabilities based on core gene alignments, and HGTector (Zhu et al. 2014) was used to detect non-*Streptomyces* potential donors.

3.3 RESULTS AND DISCUSSION

3.3.1 Geographical proximity is not correlated with genetic cohesiveness

ADMIXTURE indicated that all isolates belonged to a single ancestral lineage, suggesting the possibility of a population whose genetic similarity is decaying slowly but wherein gene flow is recent enough to maintain genetic cohesion. ANI is higher between strains within MAN (98.6 ± 0.54) and WA (98.54 ± 0.32) when compared to MAN-WA (98.43 ± 0.32), DEN-WA (98.3 ± 0.34) or DEN-MAN (98.26 ± 0.34) ANI values (t-test; $p < 0.005$ for all comparisons). Strains isolated from Denali, however, had ANI values (98.4 ± 0.61) comparable to the ANI they shared with strains in MAN or WA ($p > 0.01$).

Weir-Cockham F_{ST} indicates that DEN is most distinct from other sites (Table 3.2) despite its shared latitude and environmental similarities with MAN. The low F_{ST} values, however, indicate that these populations are not highly differentiated and gene flow is high despite geographical separation.

	DEN	**MAN**
MAN	0.127	**-**
WA	0.106	0.078

Table 3.3 Weir-Cockham F_{ST} estimates of population differentiation between sites

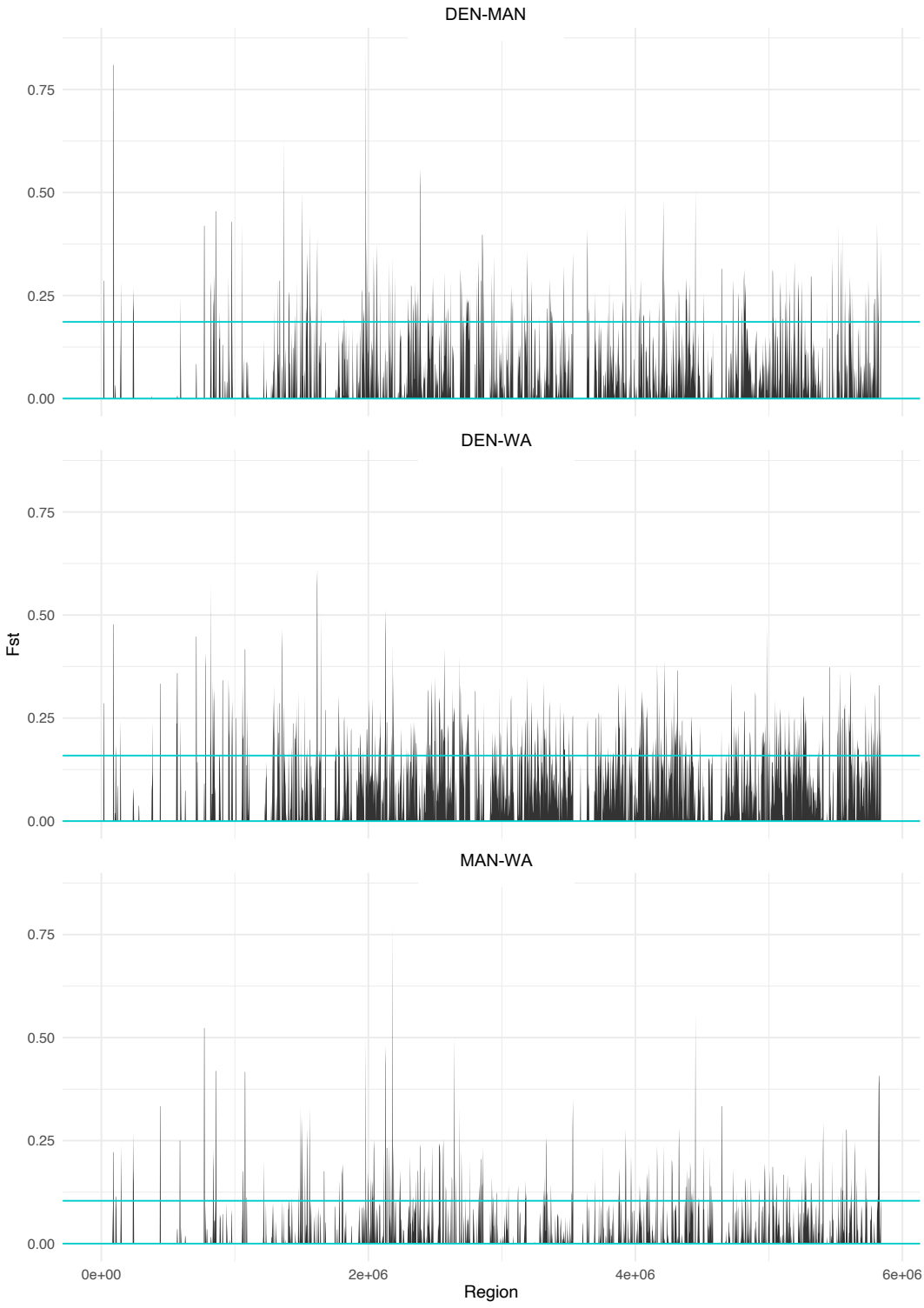

Figure 3.1 Each panel shows the F_{ST} (fixation index) between two populations per nucleotide over the whole genome region shared by all strains. The horizontal blue lines represent the mean ± standard deviation of F_{ST} for the populations in the panel.

Regions with large, consecutive increases in F_{ST} represent genomic regions that differentiate each set of populations.

3.3.2 Populations show evidence for recent substructure and unique gene content in their pangenomes

All but four strains (wa30, wa1098, man116, den1) clustered into the same population when analyzed with PopCOGenT but the DEN strains formed two distinct subpopulations within the main population: den10, den30 and den12 forming one group; and den28, den31, den35, den41 in the second subpopulation. This population sub-structuring is consistent with the F_{ST} values seen from nucleotide variants, and indicates that the DEN population may have undergone some recent gene sweeps that now differentiate them from their neighboring MAN relatives.

PopCOGenT identifies population sub-structure on the basis of shared similarity in genomic regions, and is based on the theory that recent gene flow results in nearly identical stretches of DNA since mutations accumulate over time. Once gene flow decreases due to speciation, genetic similarity decays slowly over time. Thus, even though strains may lie within the species boundary, recently speciated populations may not be recognizable by conventional detection methods because their genomes have not diverged enough (VanInsberghe et al., 2020).

Scoary found that the *lysX* gene (alpha-aminoadipate-LysW ligase for L-lysine biosynthesis) was preferentially enriched in DEN strains when compared to MAN ($p < 0.01$), while several genes were preferentially associated with all Alaska genomes including multidrug resistance proteins, enzymes involved in tryptophan and glycine biosynthesis, solute and oligopeptide transporters, copper homeostasis protein, and

vitamin B12 import protein among others. In contrast, the WA genomes were enriched in histidine, proline, and methionine biosynthesis enzymes and inner membrane protein-coding genes (cell and spore membranes). It is unclear whether these gene enrichments are a result of measured variables like temperature regimes or other variables (contemporary or historical) that were not measured.

3.3.3 Horizontal gene transfer patterns

BAPS (Corander et al., 2004) initially divided the genomes into two clusters; Cluster 1 contained man143, man150, man153, and man156 while the second cluster contained all other strains. Site-by-site HMM comparisons showed that the MAN strains were >50% similar to the other strains, thus collapsing all strains into a single fastGEAR lineage.

443 external recombination events were detected across sites. Strains in all three sites received the *rtcB* gene that codes for an RNA-splicing ligase while MAN and WA strains had common external transfers of chorismate dehydratase, cold shock like protein 7.0, hydroxyacylglutathione hydrolase GloC, peptidoglycan transpeptidase FtsI, and sortase SrtE2. DEN strains received aconitate hydratase, alkyl hydroperoxide reductase, beta-carbonic anhydrase, doxorubicin resistance transporter protein, diacetylchitobiose binding protein NgcE, ferric enterobactin transport protein, gas vescicle structural proteins, RNA polymerase sigma factor, universal stress protein, and zinc metalloprotease Rip1 while WA strains received 3-oxoacyl reductase FabG, alanine racemase, anthranilate synthase, ATP synthases, cell division protein CrgA, cobalt/magnesium transport protein CorA, global transcriptional regulator, ferredoxin,

homoserine kinase, peptidoglycan transpeptidase FtsI, Ser/The phosphatase, L-lysine N6-monooxygenase, rhomboid protease GluP, RNA-splicing ligase RtcB, transcriptional regulator protein AfsQ1, vitamin B12 import ATP-binding protein BtuD. The only unique transfer at MAN was transcriptional repressor NsrR that senses nitric oxide. MAN also had a significantly lower number of detected transfers compared to the WA strains (10.4 ± 5.8, DEN 14.9 ± 5.9, WA 16 ± 6.8, $p < 0.05$, Welch's t test).

HGTector detected external donors to be primarily other Actinomycetes from class *Actinomycetia* and orders *Streptosporangiales*, *Micromonosporales*, *Pseudonocardiales*, and *Streptomycetales*. Only rarely were transfers detected from external phyla (eg. one transfer from the order *Hyphomicrobiales* in phylum Proteobacteria), which is consistent with the idea that recombination is typically highest within phylogenetic groups even amongst free-living bacteria in soil.

The r/m value (effective recombination rate) for these genomes is 1.86, which is higher than the recombination rates observed in many soil bacteria (Vos et al., 2015), but lower than expected compared to other environmental *Streptomyces* populations (Doroghazi and Buckley, 2014). While the point mutation rates and frequency of recombination are comparable to other populations, the average recombined region length is lower than expected (82 bp compared to 348 bp for the genomes in Chapter 2).

The average gene length in the *Streptomyces avermitilis* genome is ~1,190 bp, and this threshold was used to differentiate short transfers (<1000 bp) versus large transfer events (>1000 bp). Only 400 of over 20,000 recombined regions reconstructed

by ClonalFrameML were large transfers (1.7%), which is in agreement with previous observations (González-Torres et al., 2019). Many strains contain only short fragment transfers with an average of 752 transfers per genome (compared to an average of 20 long fragment transfers per genome), but there is no site-specific bias.

The same study found that competence is important for short fragment exchange and that conjugation may be involved in the exchange of longer fragments. Thus, the length and frequency of gene transfer events might provide important clues about the mechanism of HGT in terrestrial bacteria. Homologous recombination (HR) leads to adaptive evolution as measured by amino acid substitutions in the core genome in soil *Rhizobium* strains, and increases the probability of fixation of advantageous mutations while lowering the probability of fixation for deleterious mutations (Cavassim et al., 2021). Thus, recombination and gene-specific selection are important modes of evolution among soil microbes and can structure populations at scales of meters in soils (Crits-Christoph et al., 2018).

3.3.4 SMGCs show patterns of vertical descent

Location does not affect SMGC distribution (Kruskal-Wallis test; $p > 0.3$), indicating that SMGCs may not be shared in the local gene pool as previously expected. There is evidence to suggest that individual genes in biosynthetic gene clusters may be horizontally transferred (see Chapter 2), but the broader sharing patterns still indicate the likelihood of vertical transfer or horizontal gene transfer within very closely related strains, thus mimicking vertical inheritance. The number of gene clusters within a particular SM category (eg. arylpolyene) do not vary by site

(Kruskal-Wallis test, $p = 0.34$), indicating that isolates across all three sites have similar diversity of SMGCs.

While the extent of HGT in these SMGCs is unknown, recombination can define population structure in the absence of dispersal limitation, meaning that allopatric populations can maintain a strong ancestral signature even in the presence of geographical separation as seen in some *Ferroplasma* populations (Tyson et al., 2004). In this particular study, three populations were identified with the presence of mosaic genome types, which could confer resilience to the species in the presence of environmental perturbations.

A previous study found habitat-specific genetic variations and fitness trade-offs in conspecific *Streptomyces olivaceus* isolates. Substrate utilization (sialic acid and glycogen) differentiated the two lineages in this study, indicating that adaptive processes drove ecological divergence among closely related streptomycetes, eventually leading to dispersal limitation, barriers to gene flow, and formation of a species complex consisting of two cryptic species (Wang et al., 2022). Further physiological tests will be necessary to identify traits that differentiate the sister lineages in this study, however.

Interestingly, although Fig. 3.2 shows the potential of these strains to produce different types of antibiotics, the genome annotations show that every genome has antibiotic resistance proteins, sometimes corresponding to the same antibiotic type that they produce. It has been shown that the probability of antagonism between *Vibrio* strains increases as the genetic distance between them increases (Cordero et al., 2012), and the same study identified the existence of a natural genetic boundary for

interference competition in *Vibrio*. A similar boundary for self-recognition and antagonism in terrestrial *Streptomyces* might indicate a new delineation for population clusters in the environment.

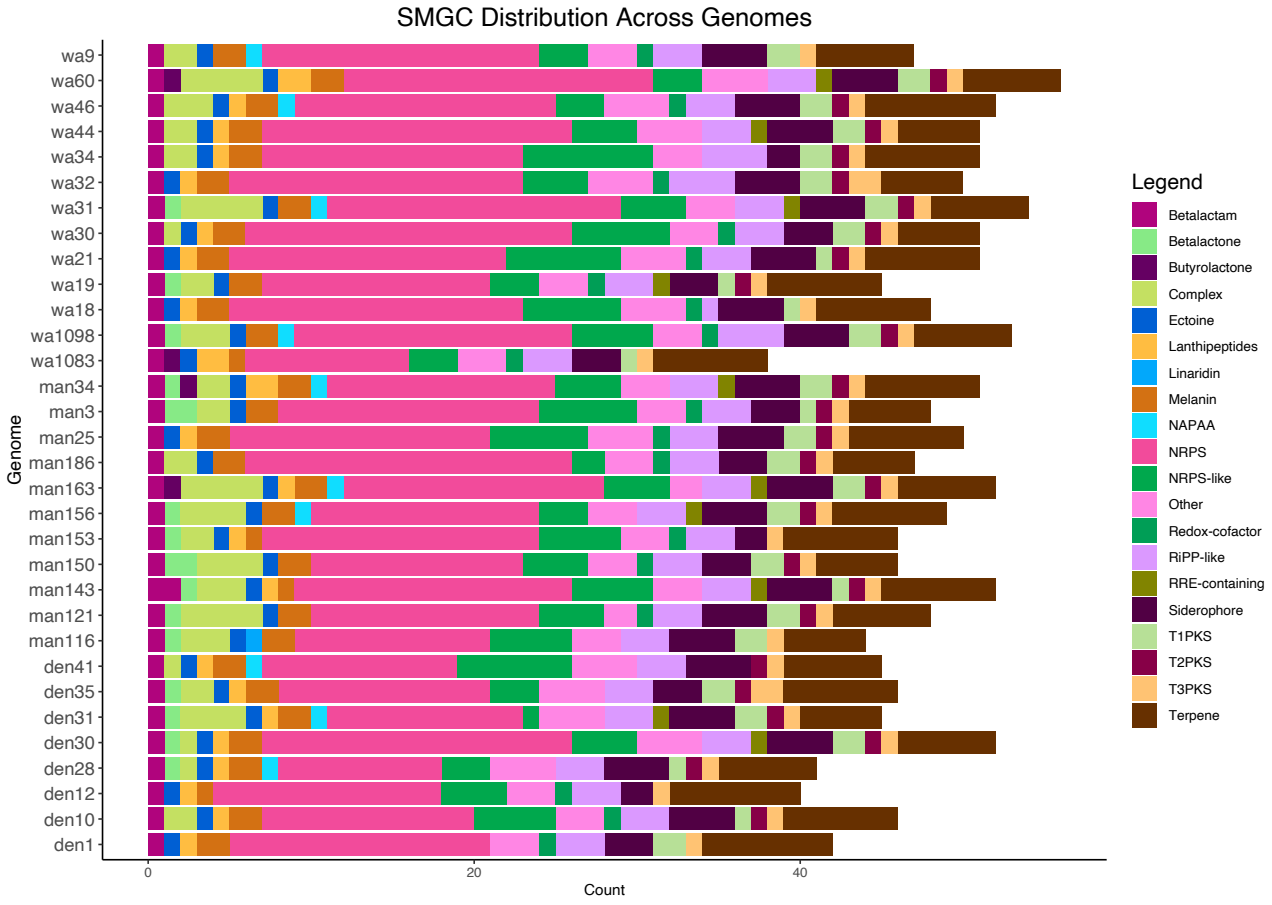

Figure 3.2 Most secondary metabolite gene clusters are shared by all strains, indicating that they were inherited from the ancestral lineage. While the numbers of SMGCs vary by genome, very few clusters are unique to a genome or site (linaridin is an exception only found in man116).

REFERENCES

Alexander DH, Novembre J, Lange K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109.

Andam CP, Choudoir MJ, Vinh Nguyen A, Sol Park H, Buckley DH. (2016). Contributions of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces* lineages. *ISME Journal* 10, 1731–1741. doi: 10.1038/ismej.2015.230.

Angiuoli SV, Salzberg SL. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342. doi: 10.1093/bioinformatics/btq665.

Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. (2019). A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* 178, 820-834.e14. doi: 10.1016/j.cell.2019.06.033.

Bérdy J. (2005). Bioactive Microbial Metabolites: A Personal View. *J Antibiot* 58, 1–26. doi: 10.1038/ja.2005.1.

Broquet T, Petit EJ. (2009). Molecular Estimation of Dispersal for Ecology and Population Genetics. *Annual Review of Ecology* 40, 193–216. doi: 10.1146/annurev.ecolsys.l.

Caro-Quintero A, Konstantinidis KT. (2012). Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* 14, 347–355. doi: 10.1111/j.1462-2920.2011.02668.x.

Cavassim MIA, Andersen SU, Bataillon T, Schierup MH. (2021). Recombination Facilitates Adaptive Evolution in Rhizobial Soil Bacteria. *Molecular Biology and Evolution* 38, 5480–5490. doi: 10.1093/molbev/msab247.

Chase AB, Arevalo P, Brodie EL, Polz MF, Karaoz U, Martiny JBH. (2019). Sympatric and allopatric differentiation delineates population structure in free-living terrestrial bacteria. *bioRxiv*, 644468. doi: 10.1101/644468.

Choudoir MJ, Buckley DH. (2018). Phylogenetic conservatism of thermal traits explains dispersal limitation and genomic differentiation of *Streptomyces* sister-taxa. *ISME Journal* 12, 2176–2186. doi: 10.1038/s41396-018-0180-3.

Choudoir MJ, Doroghazi JR, Buckley DH. (2016) Latitude delineates patterns of biogeography in terrestrial *Streptomyces*. doi: 10.1111/1462-2920.

Choudoir MJ, Pepe-Ranney C, Buckley DH. (2018). Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in *Streptomyces*. *Antibiotics* 7, 12. doi: 10.3390/antibiotics7010012.

Corander J, Waldmann P, Marttinen P, Sillanpaa MJ. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* 20, 2363–2369. doi: 10.1093/bioinformatics/bth250.

Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF. (2012). Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science* 337, 1228–1231. doi: 10.1126/science.1219385.

Cristofoli S, Monty A, Mahy G. (2010). Historical landscape structure affects plant species richness in wet heathlands with complex landscape dynamics. *Landscape and Urban Planning* 98, 92–98. doi: 10.1016/j.landurbplan.2010.07.014.

Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* 558, 440–444. doi: 10.1038/s41586-018-0207-y.

de la Giroday HMC, Carroll AL, Lindgren BS, Aukema BH. (2011). Incoming! Association of landscape features with dispersing mountain pine beetle populations during a range expansion event in western Canada. *Landscape Ecology* 26, 1097–1110. doi: 10.1007/s10980-011-9628-9.

Didelot X, Wilson DJ. (2015). ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology* 11, e1004041. doi: 10.1371/journal.pcbi.1004041.

Doroghazi JR, Buckley DH. (2014). Intraspecies comparison of *Streptomyces pratensis* genomes reveals high levels of recombination and gene conservation between strains of disparate geographic origin. *BMC Genomics* 15, 1–14. doi: 10.1186/1471-2164-15-970.

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *PNAS* 103(3), 626–631. doi: 10.1073/pnas.0507535103

Garrison E, Marth G. (2012). Haplotype-based variant detection from short-read sequencing. Available at: http://arxiv.org/abs/1207.3907 [Accessed June 1, 2022].

González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. (2019). Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. *mBio* 10. doi: 10.1128/mBio.02494-18.

Ho PT, Park E, Hong SG, Kim EH, Kim K, Jang SJ, Vrijenhoek RC, Won YJ. (2017). Geographical structure of endosymbiotic bacteria hosted by *Bathymodiolus* mussels at eastern Pacific hydrothermal vents. *BMC Evolutionary Biology* 17, 1–17. doi: 10.1186/s12862-017-0966-3.

Karimi B, Terrat S, Dequiedt S, Saby NPA, Horrigue W, Lelièvre M, Nowak V, Jolivet C, Arrquays D, Wincker P, Cruaud C, Bispo A, Maron PA, Bouré NCP, Ranjard L. (2018). Biogeography of soil bacteria and archaea across France. *Science Advances* 4. doi: 10.1126/sciadv.aat1808.

Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997 [Accessed June 1, 2022].

Liu D, Yang Y, An S, Wang H, Wang Y. (2018). The biogeographical distribution of soil bacterial communities in the loess plateau as revealed by high-throughput sequencing. *Frontiers in Microbiology* 9. doi: 10.3389/fmicb.2018.02456.

Ma B, Dai Z, Wang H, Dsouza M, Liu X, He Y, Wu J, Rodrigues JLM, Gilbert JA, Brookes PC, Xu J. (2017). Distinct Biogeographic Patterns for Archaea, Bacteria, and Fungi along the Vegetation Gradient at the Continental Scale in Eastern China. *mSystems* 2, 1–14. doi: 10.1128/msystems.00174-16.

Manteca A, Sanchez J. (2009). *Streptomyces* development in colonies and soils. *Appl Environ Microbiol* 75, 2920–2924. doi: 10.1128/AEM.02288-08.

Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. (2017). Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol* 34, 1167–1182. doi: 10.1093/molbev/msx066.

Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, Fierer N, Townsend AR, Cleveland CC, Stanish L, Knight R. (2011). Global patterns in the biogeography of bacterial taxa. *Environmental Microbiology* 13, 135–144. doi: 10.1111/j.1462-2920.2010.02315.x.

Oda Y, Star B, Huisman LA, Gottschal JC, Forney LJ. (2003). Biogeography of the Purple Nonsulfur Bacterium *Rhodopseudomonas palustris*. *Applied and Environmental Microbiology* 69, 5186–5191. doi: 10.1128/AEM.69.9.5186-5191.2003.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. (2014). PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution* 31, 1929–1936. doi: 10.1093/molbev/msu136.

Razgour O, Rebelo H, Puechmaille SJ, Juste J, Ibáñez C, Kiefer A, Burke T, Dawson DA, Jones G. (2014). Scale-dependent effects of landscape variables on gene flow and population structure in bats. *Diversity and Distributions* 20, 1173–1185. doi: 10.1111/ddi.12200.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033.

Touchon M, Rocha EPC. (2016). Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harbor Perspectives in Biology* 8, 1–18. doi: 10.1101/cshperspect.a018168.

Traxler MF, Kolter R. (2015). Natural products in soil microbe interactions and evolution. *Natural Product Reports* 32, 956–970. doi: 10.1039/c5np00013k.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43. doi: 10.1038/nature02340.

VanInsberghe D, Arevalo P, Chien D, Polz MF. How can microbial population genomics inform community ecology? *Philosophical Transactions of the Royal Society B: Biological Sciences* 375, 20190253. doi: 10.1098/rstb.2019.0253.

Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A. (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends in Microbiology* 23, 598–605. doi: 10.1016/j.tim.2015.07.006.

Wang J, Li Y, Pinto-Tomás AA, Cheng K, Huang Y. (2022). Habitat Adaptation Drives Speciation of a *Streptomyces* Species with Distinct Habitats and Disparate Geographic Origins. *mBio* 13, e02781-21. doi: 10.1128/mbio.02781-21.

Wang J, Soininen J. (2017). Thermal barriers constrain microbial elevational range size via climate variability. *Environmental Microbiology* 19, 3283–3296. doi: 10.1111/1462-2920.13823.

Whitaker RJ, Grogan DW, Taylor JW. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* 301, 976–978. doi: 10.1126/science.1086909.

Yeager CM, Gallegos-Graves LV, Dunbar J, Hesse CN, Daligault H, Kuske CR. (2017). Polysaccharide Degradation Capability of Actinomycetales Soil Isolates from a Semiarid Grassland of the Colorado Plateau. *Appl Environ Microbiol* 83, e03020-16. doi: 10.1128/AEM.03020-16.

Zhang B, Wu X, Zhang G, Li S, Zhang W, Chen X, Sun L, Zhang B, Liu G, Chen T. (2016). The diversity and biogeography of the communities of Actinobacteria in the forelands of glaciers at a continental scale. *Environmental Research Letters* 11. doi: 10.1088/1748-9326/11/5/054012.

CHAPTER 4

SPECIES DESCRIPTION OF *STREPTOMYCES APRICUS* SP. NOV., ISOLATED FROM SOIL*

4.1 ABSTRACT

A novel *Streptomyces* strain, SUN51^T, was isolated from soils in Wisconsin, USA, as part of a *Streptomyces* biogeography survey. Genome sequencing revealed that this strain had less than 90% average nucleotide identity (ANI) to type species of *Streptomyces*: SUN51^T was most closely related to *Streptomyces dioscori* A217^T (99.5% 16S rRNA gene identity, 89.4% ANI). Genome size was estimated at 8.81 Mb, and genome DNA G+C contents was 72%. The strain possessed the cellular fatty acids anteiso-$C_{15:0}$, iso-$C_{16:0}$, 16:1 ω7c, anteiso-$C_{17:0}$, iso-$C_{14:0}$ and $C_{16:0}$. The predominant menaquinones were MK9 and MK10. SUN51^T contained the polar lipids phosphatidic acid, phosphatidyl ethanolamine, phosphatidyl glycerol, and diphosphatidyl glycerol. The cell wall contained LL-diaminopimelic acid. The strain could grow on a broad range of carbon sources and tolerate temperatures of up to 40°C. The results of the polyphasic approach confirmed that this isolate represents a novel species of the genus *Streptomyces*, for which the name *Streptomyces apricus* is proposed. The type strain of this species is SUN51^T (=NRRL B-65543 =JCM 33736).

4.2 INTRODUCTION

* Published as Hariharan J, Choudoir MJ, Diebold P, Panke-Buisse K, Buckley DH. (2022) *Streptomyces apricus* sp. nov., isolated from soil. International Journal of Systematic and Evolutionary Microbiology 72:005178. doi: 10.1099/ijsem.0.005178

Species of the genus *Streptomyces* of the phylum *Actinobacteria* are found in a wide range of habitats and are common in soils worldwide. Filamentous actinomycetes are responsible for producing over two-thirds of known antibiotics, with *Streptomyces* producing a majority of these compounds (Challis and Hopwood, 2003). *Streptomyces* have been noted for their high rates of homologous recombination (Doroghazi et al., 2014), which can complicate species identification (Labeda et al., 2017; Cheng et al., 2016; Rong et al., 2013). In addition, the genetic diversity and secondary metabolite genes of *Streptomyces* species can diverge along geographical gradients (Choudoir et al., 2018; Andam et al., 2016[a]; Choudoir et al., 2016). These results suggest that while *Streptomyces* are capable of rapid dispersal, evolutionary and ecological processes can produce regionally defined species pools.

We describe a novel species of *Streptomyces* isolated from Wisconsin, USA, identified as part of a larger effort to characterize the biogeography of *Streptomyces* in soils across the continental United States (Andam et al., 2016[a]). The strain was identified from a culture collection comprising over 1,000 strains from 15 sites, as described in earlier studies (Andam et al., 2016[a]; Andam et al., 2016[b]; Choudoir et al., 2016). In order to identify unique species and clades, initial screening of the culture collection was performed by analyzing *rpoB* gene sequences (Andam et al., 2016[a]). SUN51^T (*Streptomyces apricus*) was isolated from a grass rhizosphere soil in Sun Prairie (43.17°N, 89.24°W) Wisconsin. The soil was a typic endoaquoll, 6.1% organic matter, pH 6.6, and the site has an average temperature of 8.4 ˚C. The strain was isolated using a standard plate dilution method and grown on glycerol-arginine agar supplemented with cycloheximide as described previously (Doroghazi et al., 2010).

After 7-14 days of aerobic incubation at 25°C, actinomycete colonies were transferred and purified on glycerol-arginine agar and maintained as spore stocks in 20% glycerol suspensions (v/v) at -80°C.

4.3 METHODS

Morphological characteristics were observed by light microscopy and scanning electron microscopy (MIRA3 TESCAN) using cultures grown on ISP (International Streptomyces Project) media (Shirling and Gottlieb, 1966) at 25°C for 14 days. Morphological characteristics were determined on ISP media 2-7 and R2A Nutrient Agar (Difco) after 14 days at 28°C. Temperature range and optima for growth were determined on glycerol-arginine agar after growth for 14 days. The pH range for growth (pH 4-10) was tested in Glucose Yeast Extract (GY) broth using the following buffers each at 50 mM (Ruzin, 1999; Ferguson et al., 1980; Good et al., 1966): acetic acid/acetate (pH 4-5), MES (pH 6), MOPS (pH 7), HEPPS (pH 8), AMPSO (pH 9), AMP (pH 10). Tolerance of NaCl (0-20% w/v) was also tested in GY broth. The utilization of sole carbon and nitrogen sources was examined using Biolog™ plates (Biolog, California, USA). In addition, Zym strips were used to identify enzyme production (BioMérieux, USA).

The antibiotic sensitivity of the strain was evaluated by measuring zones of inhibition in response to antibiotic discs. The discs were placed onto ISP2 plates freshly inoculated with cultures, which had been grown for 2-3 days in ISP2 liquid media with shaking. A portion of these cultures was homogenized using a sterile plastic mortar and pestle, and this was used to inoculate the plates. Plates were

incubated for 2 days before quantifying the zone of inhibition. We evaluated resistance to ampicillin (10 μg), minocycline (30 μg), penicillin (6 μg), novobiocin (30 μg), sulfamethoxazole (25 μg), clindamycin (2 μg), oxytetracycline (30 μg), erythromycin (15 μg), chloramphenicol (30 μg), vancomycin (5 μg), nitrofurantoin (300 μg), streptomycin (10 μg), tetracycline (30 μg), rifampicin (5 μg), nalidixic acid (30 μg) and kanamycin (30 μg).

Additionally, antibiotic production was evaluated by antagonism towards different bacterial species (referred to as test strains), namely *Bacillus subtilis*, *Arthrobacter globiformis*, *Pseudomonas aeruginosa*, *Serratia marcensens*, *Flavobacterium odoratum*, *Verrucomicrobium spinosum*, and *Vibrio vulnificus*. The *Streptomyces* strain was grown as a lawn on ISP2 agar for 30 days. Agar plugs were taken from this lawn using the tip of a 1 mL pipette. These plugs were placed culture side up onto a lawn of freshly spread test strain on 1.5% nutrient agar. The plates were incubated at 30°C, and the zones of inhibition were measured following 24 hours of growth.

Biomass for chemical studies was prepared by growing the organisms in GY broth in culture tubes (200 r.p.m) at 25°C for 7 days. Cells were harvested by centrifugation, washed with distilled water, and freeze-dried. Cellular phospholipids were extracted and identified by a previously described method (Gaspar et al., 2017). Specifically, cells were extracted with chloroform-methanol (2:1 v/v) and 0.6% NaCl and were not radioactively labeled. Fatty acid methyl esters were determined using the Sherlock MIDI system (MIDI Inc., DE, USA). Freeze-dried cells were sent to the DSMZ (Leibniz Institute, Germany) for analysis of diaminopimelic acid (DAP) and

menaquinones.

DNA was extracted from liquid cultures grown in Yeast Extract-Malt Extract (YEME) medium at 30°C for 7 days using the protocol outlined elsewhere (Nikodinovic et al., 2003). Extracted DNA was quantified using the PicoGreen assay (Thermo Fisher Scientific, MA, USA) prior to library preparation and barcoding with the Nextera XT kit (Illumina, Inc., CA, USA) according to the manufacturer's instructions. These libraries were then sequenced on a MiSeq machine using the v3 kit (2 x 300 bp reads) at the Biotechnology Resource Center, Cornell University.

Raw reads were assembled using the A5-miseq pipeline with default trimming and assembly parameters (Coil et al., 2015). Contigs generated at this step were assembled into scaffolds using MeDuSa (Bosi et al., 2015). Draft assemblies were evaluated for completeness and contamination with CheckM (Parks et al., 2015). Genome annotation of scaffolds was done using Prokka (Seemann, 2014).

Small subunit rRNA (16S rRNA) gene sequences were obtained from the annotated genomes using barrnap (https://github.com/tseemann/barrnap/). Genome ANI values were computed pairwise using fastANI (Jain et al., 2018). MLST profiles (*atpD*, *gyrB*, *recA*, *rpoB* and *trpB*) for reference strains were obtained from the ARS *Streptomycetaceae* MLSA database and compared with the concatenated sequences derived from the draft genome of SUN51^{T}. MEGA7 was used to align each pair of sequences using MUSCLE and to analyze the evolutionary distance between those sequences using the Kimura 2-parameter model (Kumar et al., 2016). Digital DNA-DNA hybridization (Digital DDH) values were calculated using GGDC 2.1 (Meier-Kolthoff et al., 2013). All phylogenetic analyses excluded strains that did not have

validly published names.

4.4 RESULTS AND DISCUSSION

Initial screening to identify unique isolates in our culture collection was performed by analyzing the *rpoB* gene sequences of strains relative to *Streptomyces rpoB* sequences available from the NCBI nucleotide database and from (Labeda et al., 2017) and (Andam et al., 2016[b]). Candidates for new species designation were identified using a 0.01 nucleotide dissimilarity cutoff for *rpoB*, as applied in (Andam et al., 2016[b]) and (Rong et al., 2012). Comparator strains were identified by searching for species having similar 16S rRNA genes, MLSA types, and genome composition (as described above). The genome sequences of candidate species were determined and used to calculate ANI with respect to all *Streptomyces* genome sequences present in the NCBI Genome and IMG (Markowitz et al., 2012) databases.

The complete 16S rRNA gene sequence of SUN51^T (1524 bp) was most closely related to *Streptomyces dioscori* A217^T (99.6% identity). Percent identity was determined using the BLAST 2.9.0+ algorithm. Phylogenetic analysis of 16S rRNA genes (Fig. 4.1) was performed using the neighbor-joining method as implemented in QuickTree with 1000 bootstraps (Howe et al., 2002) and verified using the maximum-likelihood method (Generalized Time Reversible Gamma model) as implemented in RAxML (Stamatakis 2014). This phylogenetic analysis includes all sequences in GenBank's 16S rRNA sequence database (Johnson et al., 2008) having > 98.7% sequence identity to SUN51^T. Unfortunately, 16S rRNA gene comparisons can be misleading in the identification of *Streptomyces* species because this locus has low

resolution for *Streptomyces* species (Labeda et al., 2017; Cheng et al., 2016; Rong et al., 2013; Rong et al., 2012). For example, *Streptomyces* species with highly similar, or even identical, 16S rRNA gene sequences often have highly divergent genomes (Labeda et al., 2017; Cheng et al., 2016). In addition, numerous instances of horizontal transfer have been documented for 16S rRNA genes within the genus *Streptomyces* (Rong et al., 2013; Doroghazi and Buckley 2010). As a result, phylogenetic incongruence between 16S rRNA gene sequence phylogeny and multi-locus or genomic phylogenies is not unusual in the genus *Streptomyces*, and multi-locus or orthology-based analyses are more reliable for species identification (Labeda et al., 2017; Guo et al., 2008). Since 16S rRNA gene sequence analyses are unable to reliably resolve closely related species of *Streptomyces*, we performed both MLSA and genomic analyses.

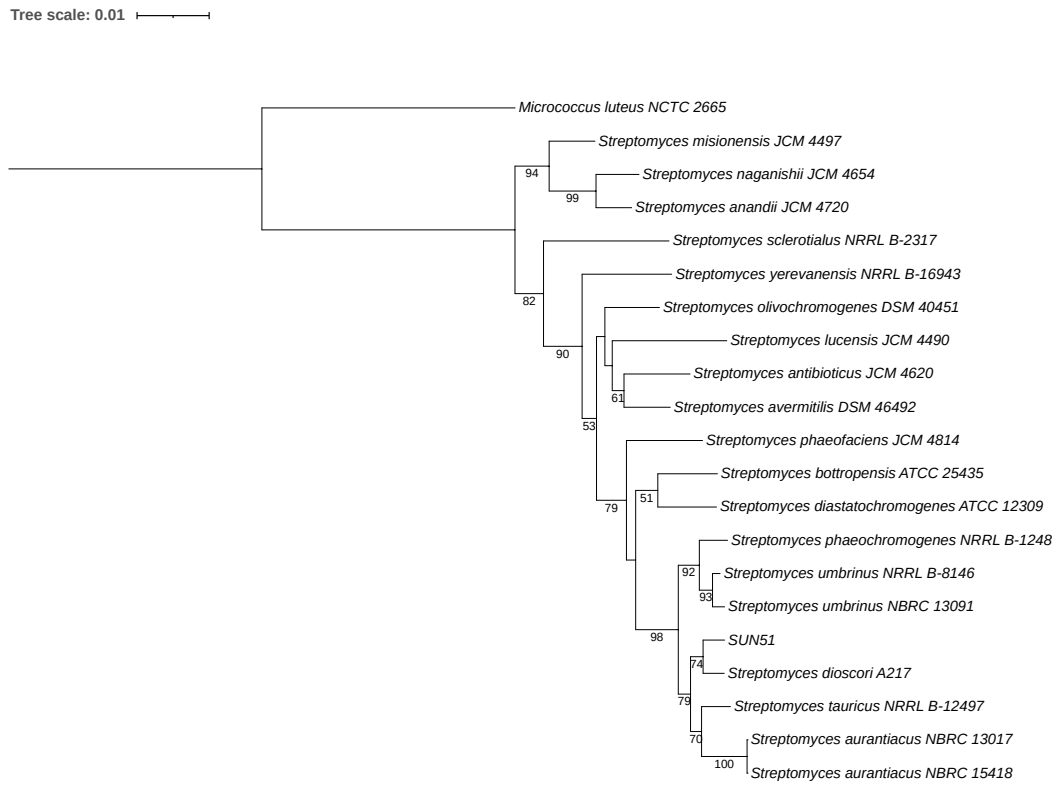

Figure 4.1 Neighbor-joining phylogenetic tree based on 16S rRNA gene sequence alignments. All type strains found in Genbank to have 16S rRNA gene sequence identity of > 98.7% to SUN51^T are included in the tree. The scale bar indicates substitutions per site and numbers indicate bootstrap values for all branches with > 50% support. Asterisks indicate branches also recovered in the maximum likelihood tree. Note: NRRL B-8146 was previously *S. ederensis* but has been reclassified as a later heterotypic synonym of *S. umbrinus*, and NBRC 15418 was previously *S. glomeroaurantiacus* but has been reclassified as a later heterotypic synonym of *S. aurantiacus* [35].

We searched for matching MLSA loci (*atpD*, *gyrB*, *recA*, *rpoB*, *trpB*) in the *Streptomyces* database on the ARS Microbial Genomic Sequence Database Server (USDA ARS, 2017). Rong and Huang (2012) established that an evolutionary distance of 0.007 between MLSA sequences is sufficient to delineate *Streptomyces* species [24]. SUN51^T has > 0.007 evolutionary distance to any validly named *Streptomyces*

species for which data is currently available (Table 4.1).

	1	2	3	4	5
MLSA distance	0.016	0.020	0.016	0.016	0.012
ANI (%)	89.1	89.4	83.9	87.6	89.6
16S rRNA (% identity)	97.7	99.6	98.5	98.9	98.8
GC content (%)	70.2	70.7	71.2	70.3	70.8
Digital DDH (%)	35	35	25	31	35

Table 4.1 MLSA-based distances, Average Nucleotide Identity (ANI), and 16S rRNA-based distances between SUN51^T and its closest phylogenetic relatives. The numbers identify comparator species as follows: (1) *S. aurantiacus* NBRC 13017, (2) *S. dioscori* A217, (3) *S. bottropensis* ATCC 25435, (4) *S. umbrinus* NBRC 13091, and (5) *S. tauricus* NRRL B-12497.

The draft genome of SUN51^T contains 1,098 contigs and 390 scaffold sequences with an N50 of 296,590 and genome size 8.81 Mb. CheckM estimated this draft genome to be 99.3% complete. The SUN51^T draft genome was evaluated against all *Streptomyces* genomes available in Genbank and IMG. All *Streptomyces* genomes that had >85% ANI to the candidate species were included in a phylogenetic analysis (Fig. 4.2) performed using OrthoFinder 2.3.8 (Emms and Kelly 2018; Emms and Kelly 2017; Emms and Kelly 2015). A total of 353 orthogroups were shared among these genomes, and these gene trees were used to reconstruct the species phylogeny. Whole genome ANI comparisons (Table 4.1) revealed the strain to have less than 90% ANI to any previously named or characterized species of *Streptomyces*. The closest genomic relatives to SUN51^T are *S. dioscori* A217^T (89.4% ANI) and *S. tauricus* NRRL B-12497 (89.6% ANI). The low ANI between SUN51^T and comparator strains supports the conclusion that SUN51^T represents a new species of *Streptomyces* (Goris et al.,

2007). This phylogenetic analysis also indicates that SUN51^T, for which we propose the name *Streptomyces apricus*, is a member of the *Streptomyces aurantiacus* phylogenetic clade identified previously (Saygin 2021).

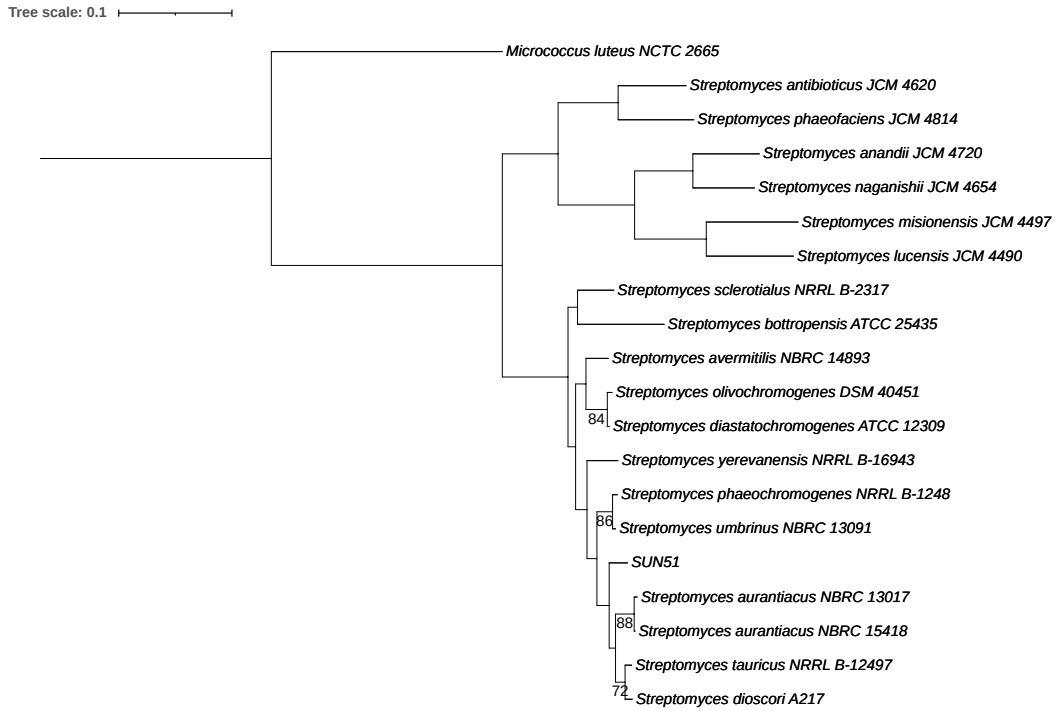

Figure 4.2 Genomic phylogeny generated using reconciliation of 353 orthogroup based gene trees. All type strains found to have either > 85% ANI or > 98.7% 16S rRNA similarity to SUN51^T are included in the tree. The scale bar indicates average number of substitutions per site and branch support values indicate STAG support values (*i.e.*, the fraction of trees that support each bipartition). Note: NBRC 15418 was previously *S. glomeroaurantiacus* but has been reclassified as a later heterotypic synonym of *S. aurantiacus* (Saygin 2021).

Strain SUN51^T grew between 10 – 40°C (optimum temperature 30°C), pH 6 – 9 (optimum pH 7), and NaCl concentration of 0 – 10% (w/v). Phenotypic properties for the novel species and comparator strains are described in detail in Table 4.2.

	1	2	3
Temperature (˚C)	10 to 40	4 to 37	10 to 37
pH	6 to 9	6 to 11	6 to 11
Salinity (%)	0 to 10	0 to 5	0 to 6
Citric Acid	+	-	+
D-Fructose	+	+	-
m-Inositol	+	-	+
Maltose	-	+	-
Sucrose	-	+	-
L-Arginine	-	+	+
L-Methionine	+	+	-
L-Serine	-	-	+
Tween 20	+	+	-
Fatty Acids			
iso-$C_{14:0}$	6.06	NA	7.9
$C_{15:0}$	2.08	NA	15.9
anteiso-$C_{15:0}$	12	NA	20.4
$C_{16:0}$	15.92	NA	36.4
iso-$C_{16:0}$	18.46	NA	NA
$C_{16:1\omega7c}$	12.67	NA	5.1
anteiso-$C_{17:0}$	5.28	NA	4.6
Polar Lipids			
Diphosphatidylglycerol	+	NA	+
Phosphatidylethanolamine	+	NA	+
Phosphatidylglycerol	+	NA	+
Phosphatidic Acid	+	NA	-

Table 4.2 Phenotypic characteristics that distinguish SUN51^T from its closest phylogenetic relatives. The numbers identify comparator species as follows: (1) SUN51^T, (2) *S. aurantiacus* NBRC 13017, (3) *S. dioscori* A217. Data for *S. aurantiacus* NBRC 13017 was obtained from (Saygin 2021; Lanoot et al., 2002) and data for *S. dioscori* A217 was obtained from (Wang et al., 2018).

SUN51^T contains phosphatidic acid, phosphatidyl ethanolamine, phosphatidyl glycerol, and diphosphatidyl glycerol in its cell membrane. The major fatty acids for this strain are $C_{15:0}$, anteiso- $C_{15:0}$, iso-$C_{16:0}$, 16:1ω7c, anteiso-$C_{17:0}$, iso-$C_{14:0}$, and $C_{16:0}$. The cell wall contained the menaquinones MK9 and MK10, and the LL-

diaminopimelic acid isomer.

Carbon and nitrogen source utilization were identified using Biolog™ plates. Discriminating substrates with respect to reference strains are indicated in Table 4.2. The following substrates were used by SUN51^T: L-Arabinose, N-Acetyl-D-Glucosamine, Succinic Acid, D-Galactose, L-Proline, D-Mannose, Glycerol, D-Gluconic Acid, L-Glutamic Acid, D,L-Malic Acid, D-Fructose, D-Xylose, Acetic Acid, α-D-Glucose, D-Glucosaminic Acid, Tween 40, α-Keto-Glutaric Acid, α-Keto-Butyric Acid, α-D-Lactose, Tween 80, α-Hydroxy-Butyric Acid, Maltotriose, Citric-Acid, Fumaric Acid, Bromo-SuccinicAcid, Propionic Acid, D-Cellobiose, L-Threonine, L-Alanine, Acetoacetic Acid, Methyl Pyruvate, L-Malic, Glycyl-L-Proline, Pyruvic Acid, Dextrin, Gelatin, Glycogen, Laminarin, Gentiobiose, D-Raffinose, Salicin, γ-Amino Butyric Acid, Butyric Acid, Caproic Acid, β-Hydroxy Butyric Acid, L-Alaninamide, and Hydroxy-L-Proline, L-Aspartic Acid, D-Sorbitol, D-Mannitol, D-Aspartic Acid, L-Arabitol, Lactitol, D-Melezitose, Sedoheptulosan, L-Sorbose, D-Tagatose, Melibionic Acid, L-Methionine, m-inositol, and L-Pyroglutamic Acid. SUN51^T produced the enzymes alkaline phosphatase, C4 esterase, Leucine arylamidase, Cystine arylamidase, acid phosphatase, napthol-AS-BI-phosphohydrolase, β-glucosidase, N-acetyl-β-glucosaminidase and α-mannosidase. A metabolic comparison of putative protein products capable of being synthesized by SUN51^T and its closest phylogenetic relative, *S. dioscori* A217, are presented in Table 4.3.

	1	2	3
Energy metabolism			
2,4-dienoyl-CoA reductase	-	+	+
Acetoin utilization protein AcuC	-	+	+
Aldo-keto reductase YhdN	+	-	-
Anaerobic sulfatase-maturating enzyme	-	+	-
Beta-glucanase	+	-	-
Cholesterol oxidase	-	+	-
Cytochrome c oxidase	-	+	+
Epoxide hydrolase B	+	-	+
Galactokinase	+	-	-
Gluconate 5-dehydrogenase	-	+	-
Glycogen synthase	-	+	-
Haloalkane dehalogenase	+	-	-
Nitrilase	+	-	-
Riboflavin biosynthesis protein RibBA	-	+	+
Zinc-type alcohol dehydrogenase-like protein	+	-	-
Amino acid metabolism			
Arginase	+	-	-
Arginine decarboxylase	+	-	+
Glutamate racemase	-	+	-
L-cystine binding protein TcyA	+	-	+
Inorganic nutrients			
Divalent metal cation transporter MntH	-	+	-
Ferrochelatase	-	+	-
Nickel import ATP-binding protein NikE	-	+	+
Motility			
Swarming motility protein SwrC	-	+	+
Fine tangled pili major subunit	-	+	+
Gas vesicle structural protein	+	-	-
Rodlin proteins RdlA, RdlB	+	-	-

Other			
Aerobactin synthase	+	-	-
Arsenical-resistance protein Acr3	-	+	+
Cytoskeleton protein RodZ	-	+	+
Erythromycin C-12 hydroxylase	+	-	-
Hypoxic response protein 1	-	+	-
Magnetosome protein MamB	+	-	+
Osmoprotectant-binding protein OsmX	-	+	+
Starvation-sensing protein RspA	-	+	+

Table 4.3 Comparison of predicted protein products between the genomes of SUN51T (1) and closest phylogenetic relatives *Streptomyces dioscori* A217 (2) and *Streptomyces tauricus* NRRL B-12497.

Antibiosis assays indicate that SUN51^T did not inhibit the growth of *Bacillus subtilis*, *Arthrobacter globiformis*, *Pseudomonas aeruginosa*, *Serratia marcensens*, *Flavobacterium odoratum*, *Verrucomicrobium spinosum*, and *Vibrio vulnificus.* SUN51^T was found to be resistant to monocycline, novobiocin, oxytetracycline, erythromycin, chloramphenicol, vancomycin, streptomycin, tetracycline, kanamycin, nalidixic acid, and rifampin.

The complete 16S rRNA gene sequence of *Streptomyces apricus* was obtained from the draft genome and have been deposited under accession number MN133488 in the GenBank database. The genome sequence for *Streptomyces apricus* sp. nov. has been deposited under the NCBI accession number VDFC00000000 (BioProject PRJNA496211).

4.5 DESCRIPTION OF *STREPTOMYCES APRICUS* SP. NOV.

Streptomyces apricus (aˈpri.cus. L. masc. adj. *apricus,* sunny, referring to the isolation of the organism from Sun Prairie, Wisconsin).

Aerobic actinomycete that forms well-developed substrate mycelium and aerial hyphae that differentiate into straight spore chains (*Rectiflexibiles* type) and produce non-motile spores with a smooth surface. Growth occurs at 10–40°C, pH 6.0–9.0 and in the presence of 0–10% (w/v) NaCl. The strain tested positive for hydrolysis of Tween 40, Tween 80, gelatin, and cellobiose. Optimal temperature and pH for growth are 30°C and pH 7.0, respectively. L-Arabinose, D-galactose, D-Mannose, D-Xylose, Glycerol, D-Glucose, m-inositol, and D-Lactose are some of the important sole carbon sources. L-Threonine, L-Alanine, L-Aspartic Acid, D-Aspartic Acid, and L-Methionine are some important sole nitrogen sources. The phospholipid profile consists of phosphatidylglycerol, diphosphatidylglycerol, phosphatidylethanolamine and phosphatidic acid. The major cellular fatty acids are iso-$C_{16:0}$, $C_{16:0}$, $C_{16:1\omega7c}$ and anteiso-$C_{15:0}$, the predominant menaquinones are MK9 and MK10, and the cell wall contains LL-diaminopimelic acid·

The type strain is SUN51^{T} (=NRRL B-65543 =JCM 33736), isolated from soil. The DNA G+C content of the type strain is 72%.

REFERENCES

Andam[a] CP, Choudoir MJ, Nguyen AV, Park HS, Buckley DH. (2016) Contributions of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces* lineages. *The ISME Journal* 10: 1731-1741. doi: 10.1038/ismej.2015.230

Andam[b] CP, Doroghazi JR, Campbell AN, Kelly PJ, Choudoir MJ, Buckley DH. (2016) A latitudinal diversity gradient in terrestrial bacteria of the genus *Streptomyces*. *mBio* 7(2):e02200-e02215. doi: 10.1128/mBio.02200-15

Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lio P, Crescenzi,P, Fani R, Fondi M. (2015) MeDuSa: a multi-draft based scaffolder. *Bioinformatics* 31(15): 2443-2451. doi: 10.1093/bioinformatics/btv171

Challis GL, Hopwood DA. (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *PNAS* 100 (2): 14555-14561. doi: 10.1073/pnas.1934677100

Cheng K, Rong X, Huang Y. (2016) Widespread interspecies homologous recombination reveals reticulate evolution within the genus *Streptomyces*. *Molecular Phylogenetics and Evolution* 102: 246 -254. doi: 10.1016/j.ympev.2016.06.004

Choudoir MJ, Doroghazi JR, Buckley DH. (2016) Latitude delineates patterns of biogeography in terrestrial *Streptomyces*. *Environmental Microbiology* 18(12): 4931-4945. doi: 10.1111/1462-2920.13420

Choudoir MJ, Pepe-Ranney CP, Buckley DH. (2018) Diversification of secondary metabolite biosynthetic gene clusters coincides with lineage divergence in *Streptomyces*. *Antibiotics* 7(1): 1-15. doi: 10.3390/antibiotics7010012

Coil D, Jospin G, Darling AE. (2015) A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* 31(4): 587-589. doi: 10.1093/bioinformatics/btu661

Doroghazi JR, Buckley DH. (2014) Intraspecies comparison of *Streptomyces pratensis* genomes reveals high levels of recombination and gene conservation between strains of disparate geographic origin. *BMC Genomics* 15: 970-984. doi: 10.1186/1471-2164-15-970

Doroghazi JR, Buckley DH. (2010) Widespread homologous recombination within and between *Streptomyces* species. *The ISME Journal* 4: 1136-1143. doi: 10.1038/ismej.2010.45

Emms DM, Kelly S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16 (157): 1-14. doi: 10.1186/s13059-015-0721-2

Emms DM, Kelly S. (2018) STAG: Species Tree inference from All Genes. *bioRxiv*. doi: 10.1101/267914

Emms DM, Kelly S. (2017) STRIDE: Species Tree Root Inference from gene Duplication Events. *Molecular Biology and Evolution* 34(12): 3267-3278. doi: 10.1093/molbev/msx259

Ferguson WJ, Braunschweiger KI, Braunschweiger WR, Smith JR, McCormick JM, Wasmann CC, Jarvis NP, Bell DH, and Good NE. (1980) Hydrogen ion buffers for biological research. *Analytical Biochemistry* 104: 300-310.

Gaspar ML, Chang Y, Jesch SA, Aregullin M, Henry SA. (2017) Interaction between repressor Opi1p and ER membrane protein Scs2p facilitates transit of phosphatidic acid from the ER to mitochondria and is essential for *INO1* gene expression in the presence of choline. *Journal of Biological Chemistry* 292 (45): 18713-18728. doi: 10.1074/jbc.M117.809970

Good NE, Winget GD, Winter W, Connolly TN, Izawa S, and Singh RMM. (1966) Hydrogen ion buffers for biological research. *Biochemistry* 5(2): 467-477. doi: 10.1021/bi00866a011

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57: 81-91. doi: 10.1099/ijs.0.64483-0

Guo Y, Zheng W, Rong X, Huang Y. (2008) A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *International Journal of Systematic and Evolutionary Microbiology* 58: 149-159. doi: 10.1099/ijs.0.65224-0

Howe K, Bateman A, Durbin R. (2002) QuickTree: building huge Neighbor-Joining trees of protein sequences. *Bioinformatics* 18(11): 1546-1547. doi: 10.1093/bioinformatics/18.11.1546

Jain C, Rodriguez-R LM, Philippy AM, Konstantinidis KT, Aluru S. (2018) High-throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*. doi: 10.1038/s41467-018-07641-9

Johnson MJ, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Research* 36(2): W5-W9. doi: 10.1093/nar/gkn201

Kumar S, Stecher G, Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33(7): 1870-4. doi: 10.1093/molbev/msw054

Labeda DP, Dunlap CA, Rong X, Huang Y, Doroghazi JR, Ju K, Metcalf WW. (2017) Phylogenetic relationships in the family *Streptomycetaceae* using multi-locus sequence analysis. *Antonie van Leeuwenhoek* 110: 563-583. doi: 10.1007/s10482-016-0824-0

Lanoot B, Vancanneyt M, Cleenwerck I, Wang L, Li W, Liu Z, Swings J. (2002) The search for synonyms among streptomycetes by using SDS-PAGE of whole-cell proteins. Emendation of the species *Streptomyces aurantiacus*, *Streptomyces cacaoi* subsp. *Cacaoi*, *Streptomyces caeruleus* and *Streptomyces violaceus*. *International Journal of Systematic and Evolutionary Microbiology* 52: 823-829. doi: 10.1099/ijs.0.02435-0

Markowitz VM, Chen IA, Palaniappan K, Chu K, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* 40: 115-122. doi: 10.1093/nar/gkr1044

Meier-Kolthoff JP, Auch AF, Klenk H, Göker M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14, 60: 1-14. doi: 10.1186/1471-2105-14-60

Nikodinovic J, Barrow KD, Chuck J. (2003) High yield preparation of genomic DNA from *Streptomyces*. *BioTechniques* 35:932-936. doi: 10.2144/03355bm05

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, and Tyson GW. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25(7): 1043-1055. doi: 10.1101/gr.186072.114

Rong X, Doroghazi JR, Cheng K, Zhang L, Buckley DH, Huang Y. (2013) Classification of *Streptomyces* phylogroup pratensis (Doroghazi and Buckley, 2010) based on genetic and phenotypic evidence, and proposal of *Streptomyces pratensis* sp. nov. *Systematic and Applied Microbiology* 36(2013): 401-407. doi: 10.1016/j.syapm.2013.03.010

Rong X, Huang Y. (2012) Taxonomic evaluation of the *Streptomyces hygroscopicus* clade using multilocus sequence analysis and DNA-DNA hybridization, validating the

MLSA scheme for systematics of the whole genus. *Systematic and Applied Microbiology* 35 (2012): 7-18. doi: 10.1016/j.syapm.2011.10.004

Ruzin SE. (1999) *Plant microtechnique and microscopy*. New York City, New York: Oxford University Press.

Saygin H. (2021) Genomic insight into the *Streptomyces aurantiacus* clade: reclassification of *Streptomyces ederensis* as a later heterotypic synonym of *Streptomyces umbrinus* and *Streptomyces glomeroaurantiacus* as a later heterotypic synonym of *Streptomyces aurantiacus*. *International Journal of Systematic and Evolutionary Microbiology* 71: 4797. doi: 10.1099/ijsem.0.004797

Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068-2069. doi: 10.1093/bioinformatics/btu153

Shirling EB, Gottlieb D. (1966) Methods for characterization of *Streptomyces* species. *International Journal of Systematic Bacteriology* 16(3): 313-340. doi: 10.1099/00207713-16-313

Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312-1313. doi: 10.1093/bioinformatics/btu033

(dataset) USDA Agricultural Research Service (2017) ARS Microbial Genomic Sequence Database Server. USDA Agricultural Research Service. https://data.nal.usda.gov/dataset/ars-microbial-genomic-sequence-database-server.

Wang Z, Tian J, Li X, Gan L, He L, Chu Y, Tian Y. (2018) *Streptomyces dioscori* sp. nov., a Novel Endophytic *Actinobacterium* Isolated from Bulbil of Dioscorea bulbifera L. *Current Microbiology* 75: 1384-1390. doi: 10.1007/s00284-018-1534-9