MICROBIAL LIFE IN CHALLENGING ENVIRONMENTS

by

NICHOLAS BRADY DRAGONE

B.A., Middlebury College, 2014

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Ecology and Evolutionary Biology

2023

Committee Members:

Noah Fierer

Alisha Quandt

Steve Schmidt

Eve-Lyn Hinckley

Brian Hynek

Dragone, Nicholas Brady (Ph.D., Ecology and Evolutionary Biology)

Microbial Life in Challenging Environments

Thesis directed by Professor Noah Fierer

ABSTRACT

Microorganisms are nearly ubiquitous on Earth, but the identity and function of microbial communities are inherently dependent on the properties of the specific environment in question. Here, I have studied soils around the world to answer questions about how the functional attributes of microorganisms allow them to respond to challenging environmental conditions. First, I explore how microbial communities in soils change across environmental gradients in Antarctica. I show that microbes in Antarctic surface soils are most restricted by low temperatures, low water availability, and high concentrations of salt. Microbial communities near the polar plateau, the most challenging environment, are dominated by Actinobacteria and Chloroflexi, and are enriched in genes associated with the oxidation of hydrogen gas as an energy source. Second, I show that the earliest microbial colonizers of a newly-formed volcanic island in the Kingdom of Tonga are chemolithotrophs that appear to have come from nearby geothermal systems. While many of these microbes utilize sulfur as an energy source, the most abundant organisms have genes that indicate they can oxidize trace gases including carbon monoxide and hydrogen. Finally, I show that organisms associated with carbon limited subsurface soils tend to have smaller genomes, grow more slowly, and have more gene pathways associated with metabolism and the storage of carbon. Taken together, these studies shed light on microbial survival in challenging soil environments and show the varied ways in which microbial communities interact with and are affected by their surroundings.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

TABLES

FIGURES

CHAPTER I

INTRODUCTION AND OVERVIEW

Introduction

Microorganisms are nearly ubiquitous on Earth and can be found all around us. Not only are they in our oceans, soils, and homes, but viable organisms have been detected in even at the bottom of the deepest ocean trenches (Pathom-aree et al. 2006), in highly acidic geothermal systems (Hynek et al. 2018), and at the top of our tallest mountains (Liu et al. 2007). While it is possible that there are still environments on Earth that are kept effectively uninhabited by environmental conditions that restrict the ability of organisms to function and survive, current research suggests that microorganisms can adapt to a multitude of challenging conditions and that diverse communities of polyextremophilic taxa (organisms who can grow under one or more challenging environmental conditions) can be active in many of the most inhospitable environments on our planet (Merino et al. 2019).

For a microbial community to establish and persist in an environment, microbes must be able to first disperse and recruit to that environment (Merino et al. 2019). Once microbes arrive, they must be able to obtain energy and nutrients from the surrounding system. These challenges have a substantial impact on the identity and function of the microbial communities with certain organisms better adapted to the environmental pressures inherent to specific systems (Merino et al. 2019). To give one example, photosynthesizing cyanobacteria that can use light energy and fix carbon are often found to dominate nutrient-limited primary successional environments (Schmidt et al. 2008). However, generalizations about the structure of microbial communities, even within similar systems, should be approached with caution. The identity and function of microbial communities are inherently dependent on the properties of the specific environment in question.

While there have numerous studies focused on microbial adaptations to particular environmental conditions in selected environments (ex: thermophiles and acidophiles in hot springs and hydrothermal vents, (Power et al. 2018; Reysenbach et al. 2020)) there are many challenging environments around the world where microbial communities, and the adaptations of those community members, have not been well studied.

We know relatively little about how the microbial communities in soils are affected by and change in response to these challenging conditions. This knowledge gap persists, in part, because the soil microbiome is inherently difficult to study. Soils are highly complex, are very heterogeneous, and can contain high levels of microbial biomass (Fierer 2017). Even the most challenging soil environments on earth, like those in Antarctica (Cary et al. 2010) or on high elevation mountains (Vimercati et al. 2019), have been found to contain diverse communities of microorganisms, though there is considerable variation in community composition across these sites. Moreover, most of the microbial taxa found in soils are undescribed and generally are not amenable to cultivation-based study (Fierer 2017). Genomic information for most soil bacteria, even abundant and ubiquitous taxa, is not current available (Delgado-Baquerizo et al. 2018). This is even more relevant in soils exposed to extreme conditions, where many of the traditional techniques most often used in the field of microbial ecology, including culturing, are ineffective. However, it is important to study these soils as they are essential to understanding how terrestrial ecosystems will respond to ongoing environmental change and such soils can be among the first to show signs of impacts due to climate change (Turner et al. 2014).

In the subsequent chapters, I seek to answer questions about soil microorganisms: how they develop over time, how they are affected by and change in response to challenging environmental conditions. To do so, I used a suite of

culture-independent sequencing techniques to analyze soil samples collected from three distinct systems. First, I explore how microbial communities in soils change across environmental gradients in Antarctica (Chapters II and III). Second, I studied how microbial communities use a range of metabolic processes to survive after colonizing a new-formed volcanic island shortly after its formation (Chapter IV). Finally, I test hypotheses about how microorganisms adapt in response to resource limitation (Chapter V).

Thesis Overview

Chapter II: Exploring the boundaries of microbial habitability in soils.

Microorganisms are the most ubiquitous forms of life on Earth and can be found in even the most challenging environments. As a result, it is often assumed that microbes have come to inhabit every terrestrial surface on Earth. Previous work has hinted that soil environments, without any detectable microorganisms or microbial activity, might exist in Antarctica. To explore this potential limit of habitability, we used a range of approaches, including culturing, DNA sequencing, and metabolic assays to explore patterns of microbial communities in Antarctic surface soils. By testing >200 soils collected across the Shackleton Glacier Region, we sought to confirm whether microbial life in Antarctic surface soils is effectively undetectable under certain conditions. While we detected diverse microbial communities in many soils, we could not detect microbes in approximately 20% of the collected samples. Our results suggest that microbial habitability is limited by the unique combination of cold, dry, salty conditions experienced at inland, higher elevation sites throughout the Transantarctic Mountain. Additionally, the prevalence of fungi at many of the most challenging sites suggests that fungi may be better adapted to some of the most challenging soil environments on Earth than Bacteria and Archaea. Chapter II is adapted from Dragone et al. (2021).

Chapter III: Elevational constraints on the composition and genomic attributes of soil microbial communities in Antarctica. Antarctic soils represent an ideal system to study how environmental properties shape the taxonomic and functional diversity of microbial communities given the relatively low diversity of Antarctic soil microbial communities and the pronounced environmental gradients that occur across soils located in reasonable proximity to one another. Moreover, the challenging environmental conditions typical of most Antarctic soils present an opportunity to investigate the traits that allow soil microbes to persist in some of the most inhospitable habitats on Earth. We used cultivation-independent methods to study the bacterial and archaeal communities found in soils collected from across the Shackleton Glacier region of the Transantarctic Mountains. We show that those environmental characteristics associated with elevation have the greatest impact on the structure of these microbial communities, with the colder, drier, and saltier soils found in higher elevation soils sustaining less diverse communities that were distinct from those in more hospitable soils with respect to their composition, genomic attributes, and overall life history strategies. Notably, the harsher conditions found in higher elevation soils likely select for taxa with lower maximum potential growth rates and an increased reliance on trace gas metabolism to support growth. Chapter III is adapted from Dragone et al. (2022).

Chapter IV: The early microbial colonizers of a short-lived volcanic island in the Kingdom of Tonga. The volcanic island of Hunga Tonga Hunga Ha'apai (HTHH) in the Kingdom of Tonga represents a very rare example of new island formation and thus a unique opportunity to study how organisms colonize a new landmass. We found that the island was colonized by diverse microbial communities

shortly after its formation in 2015. The initial stages of primary succession in this system appears to be distinct from that typically observed in other terrestrial environments, with these microbes likely originating from nearby geothermal environments. Unlike other terrestrial primary succession environments, we found none of the photosynthetic cyanobacteria that typically dominate other early successional environments present in HTHH's sediments. Instead, the early microbial colonizers rely on the capacity to generate energy via sulfur, trace gas metabolism, and anoxygenic photosynthesis to survive on the surface of this newly formed island. While the near-complete destruction of the island of HTHH in January 2022 makes it impossible to revisit the site and conduct field measurements of metabolic activity to confirm these patterns, our results suggest that surtseyan islands represent a distinct form of microbial primary succession driven by microbes with unique metabolic capabilities. Chapter IV is adapted from Dragone et al. (2023).

Chapter V: Identifying the genomic attributes of oligotrophic soil bacteria.
Not all bacteria are fast growers. In soil, as in most other environments, bacteria exist along a continuum, from those that can grow rapidly under resource-rich conditions (copiotrophs) to those that are adapted to life in the 'slow lane' and can grow even when resources are limited (oligotrophs). However, the field of microbiology is built almost exclusively on the study of copiotrophs due, in part, to the ease of studying these faster growing taxa in vitro. Oligotrophs are far more difficult to study an remain more mysterious. We expect oligotrophs to be abundant in soil environments, but we know little about their traits and ecological strategies. To better understand the taxonomic and genomic attributes of oligotrophic bacteria, we analyze ~180 soil profile samples collected from across the U.S, treating soil depth gradients as resource gradients. By comparing the more copiotrophic surface

soils to the more oligotrophic subsurface soils, we were able to investigate the identity and functional attributes of those microorganisms that dominate in resource limited environments. We were able to identify certain taxa bacterial indicative of oligotrophic environments, most notably the phyla Dormibacteriota. In general, oligotrophs associated with the carbon limited subsurface had smaller genomes, were slower growers, and were generally underrepresented in culture collections. These bacteria were also enriched in gene pathways that allow them to metabolize a wide range of energy sources and store carbon, while genes associated with energy intensive functions like chemotaxis and motility were underrepresented.

CHAPTER II

EXPLORING THE BOUNDARIES OF MICROBIAL HABITABILITY IN SOILS

Abstract

Microbes are widely assumed to be capable of colonizing even the most

challenging terrestrial surface environments on Earth given enough time. We would

not expect to find surface soils uninhabited by microbes as soils typically harbor

diverse microbial communities and viable microbes have been detected in soils

exposed to even the most inhospitable conditions. However, if uninhabited soils do

exist, we might expect to find them in Antarctica. We analyzed 204 ice-free soils

collected from across a remote valley in the Transantarctic Mountains ($84 - 85^{o}$S,

$174 - 177^{o}$W) and were able to identify a potential limit of microbial habitability.

While most of the soils we tested contained diverse microbial communities, with

fungi being particularly ubiquitous, microbes could not be detected in many of the

driest, higher elevation soils - results that were confirmed using cultivation-

dependent, cultivation- independent, and metabolic assays. While we cannot

confirm that this subset of soils is completely sterile and devoid of microbial life, our

results suggest that microbial life is severely restricted in the coldest, driest, and

saltiest Antarctic soils. Constant exposure to these conditions for thousands of years

has limited microbial communities so that their presence and activity is below

detectable limits using a variety of standard methods. Such soils are unlikely to be

unique to the studied region with this work supporting previous hypotheses that

microbial habitability is constrained by near-continuous exposure to cold, dry, and

salty conditions, establishing the environmental conditions that limit microbial life in terrestrial surface soils.

Introduction

Microorganisms are nearly ubiquitous on Earth and can routinely be found in environments that include sediments of the deepest ocean trenches (Pathom-aree et al. 2006), highly acidic geothermal systems (Hynek et al. 2018), and subsurface aquifers > 1 km below the Earth's surface (Sahl et al. 2008). Viable microbes have been detected in even the most inhospitable environments (Merino et al. 2019) and it is widely assumed that all environments on Earth should contain detectable microorganisms. This assumption is likely incorrect. Lava flows and lava lakes, for example, are inhospitable to all microbial life but only remain so for relatively short periods of time until temperatures have cooled sufficiently to permit microbial colonization (Cockell et al. 2012; Cockell 2014). Few of Earth's stable terrestrial surface environments are expected to experience conditions challenging enough to restrict microbial habitability to the point where organisms cannot be detected.

In particular, we would expect that it is very unlikely to find surface soils without detectable microorganisms. Soils typically harbor diverse microbial communities of bacteria, fungi, and archaea (Fierer 2017). This holds true even for some of the most challenging soil environments found on (Goordial et al. 2016; Schmidt et al. 2018). For example, even soils in the driest regions of the Atacama Desert contain active microbial communities despite these sites receiving ~2 cm of precipitation annually (Schmidt et al. 2018). Soil microbes are impressively adept at tolerating 'extreme' conditions, including very high or very low temperatures, high levels of UV radiation, high salt concentrations, very low or high pH levels, and even a near-complete absence of liquid water (Fierer 2017; Merino et al. 2019; Morozkina et al. 2010; Schmidt et al. 2018). This wide range of microbial tolerances

to challenging environmental conditions suggests that, given sufficient time, all surface soils on Earth should harbor active microbial life.

Do surface soils exist where microbial life cannot be detected? If so, Antarctica would be the most likely location to find them. While we know that some Antarctic soils can harbor diverse and metabolically active microbial communities (Cary et al. 2010; Ji et al. 2017), other soils harbor some of the lowest recorded levels of microbial biomass on Earth (Goordial et al. 2016). Many ice-free surface soils in Antarctica are thought to have been exposed for millions of years and have remained largely unchanged over that time (Denton et al. 1989). These soils have generally been unaffected by direct anthropogenic impacts and are often very isolated (Archer et al. 2019; Bockheim 1997). Even if microbes can be dispersed to remote soil patches via aeolian transport (Archer et al. 2019), they often face a unique combination of cold temperatures, extremely low soil water potentials, and high salt concentrations that restrict the activity and survival of all but a few specifically adapted taxa (Goordial et al. 2016; Merino et al. 2019). It seems likely that some isolated Antarctic soils may harbor few, if any, living microorganisms due to these constraints.

To explore the limits of microbial habitability in terrestrial soil environments, we sampled Antarctic surface soils from 10 distinct features along the Shackleton Glacier in the Transantarctic Mountains (84 – 85°S, 174 – 177°W). We used a suite of microbiological approaches, including cultivation-dependent, cultivation- independent, and metabolic assays, to answer two questions: Are there soils in Antarctica with no detectable signs of microbial life? and, if so, What environmental conditions might be restricting the activity of microorganisms in these inland soils to the point where they cannot be detected using conventional methods?

Results

Microbial communities of the Shackleton Glacier Region

The 204 individual soils collected for this study represent a broad range of ice-free sites found across the Shackleton Glacier region. Sites ranged in elevation from ~150 m to 2221 m above sea level (m.a.s.l) and most soils are likely to have been exposed for prolonged periods of time, with the approximate time since last wetting (estimated from ClO_4^- concentrations, Jackson et al. (2015); A. Jackson et al. (2016)) ranging from <10 years to >2 million years (mean: ~20,000 years). Not surprisingly given the absence of plants in this region, the measured soil organic carbon concentrations were low, from ~3 to 60 mg·g soil^{-1} (mean = 13 mg·g soil^{-1}), and most of the collected soils have high soluble salt concentrations (Diaz et al. 2021). As expected, the soils contained almost no water at the time of collection (0.001 to 0.11 g H_2O·g dry soil^{-1}, mean of 0.02 g H_2O·g dry soil^{-1}). In general, soils located further inland at higher elevations were drier, saltier, and contained less organic carbon (Diaz et al. 2021). These soil characteristics are not unique to the Shackleton Glacier region as ice-free soils found in other regions of Antarctica have comparable edaphic characteristics (Bockheim 1997; Magalhães et al. 2012).

Of the 204 soil samples analyzed, 80% of the samples yielded sufficient amounts of PCR-amplifiable DNA to characterize the bacterial and/or fungal taxa found in these soils using cultivation-independent marker gene sequencing. As expected, the number of taxa identified per soil sample (microbial richness) was low with a mean of 205 bacterial phylotypes (2 – 729 exact sequence variants, ESVs) and only 5.5 fungal phylotypes (1 – 45 ESVs) detected per sample. In general, the soils from higher elevation features farther away from the Ross Ice Shelf had less diverse fungal and bacterial communities (Figure 2.1, Supplemental Figure S2.1).
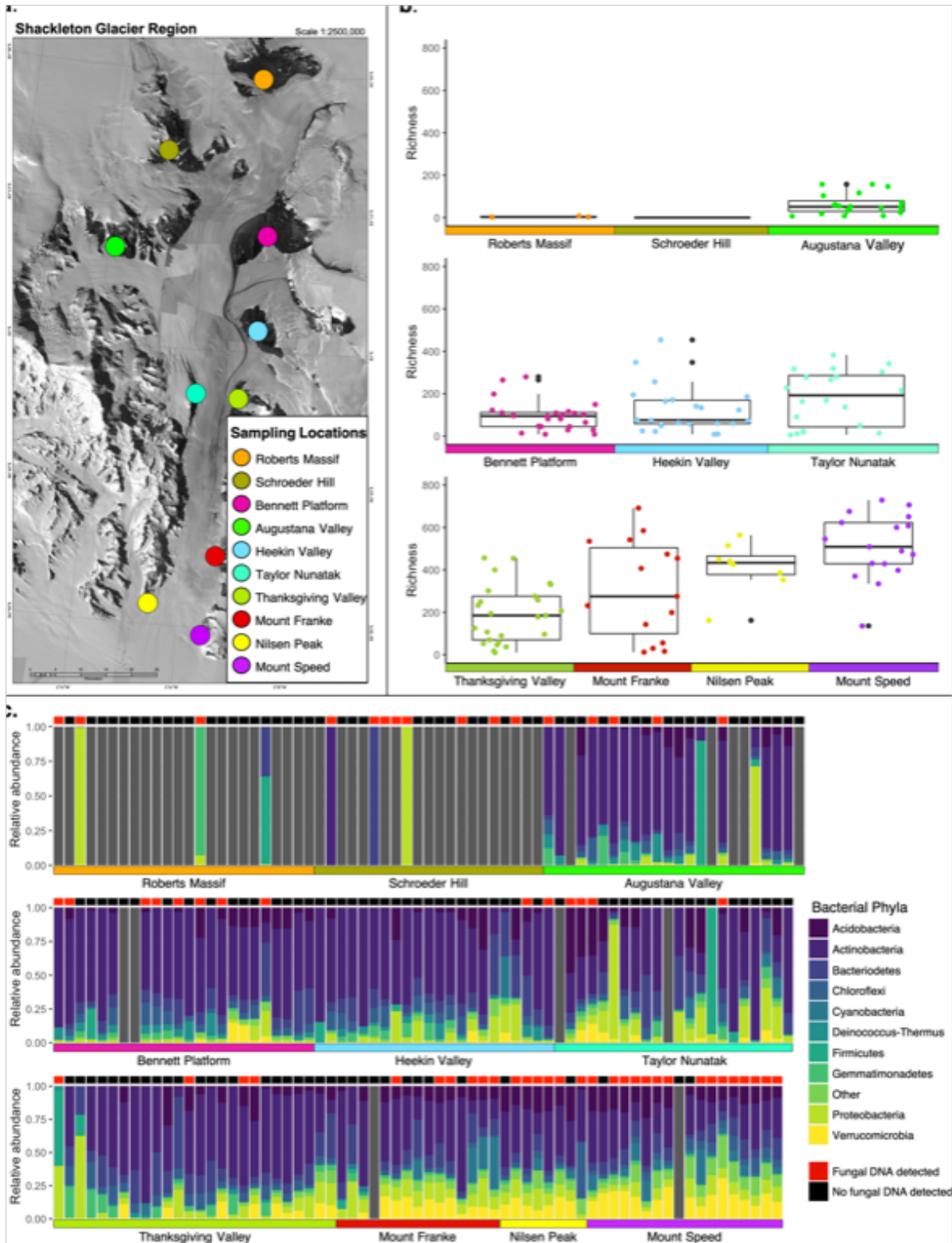
Figure 2.1: Richness and composition of the soil bacterial communities across the Shackleton Glacier region. **A)** A map of the Shackleton Glacier region (Lat: 84 -

85°S, Long: 174 – 177°W) with the locations of the 10 features where samples were collected indicated by the colored dots. **B)** The richness of the bacterial communities (number of bacterial phylotypes detected per sample) from each of the 10 sampled features. Box and whisker plots indicate the distribution of bacterial richness levels (number of phylotypes per sample) across all samples from each individual feature. **C)** Relative abundances of the dominant bacterial phyla from the 204 surface soil samples analyzed. Samples are grouped by feature and organized top to bottom from high elevation to low elevation (higher elevation sites are further south). Samples with no amplifiable bacterial DNA are represented by grey bars. The presence/absence of amplifiable fungal DNA is indicated by the red and black boxes over each bar. For more information about the fungal communities identified in each sample see Supplemental Figure S2.1.

The bacterial phyla with the highest relative abundances across all soils were: Bacteroidetes, Proteobacteria, Actinobacteria, and Acidobacteria, with the dominant taxa members of the families Solirubrobacteraceae, Blastocatellaceae, Chitinophagaceae, and Rubrobacteriaceae (Figure 2.1). The composition of the bacterial communities in these soils is consistent with results obtained using similar cultivation-independent analyses of other Antarctic soils, including those found in the McMurdo Dry Valley region (Aislabie et al. 2013; Cary et al. 2010; Lambrechts et al. 2019). The fungal communities in these soils were dominated by members of the Ascomycota and Basidiomycota phyla, including members of the families Herpotrichiellaceae, Trapeliaceae, Verrucariaceae, Filobasidiaceae, Mortierellaceae, Stereocaulaceae (Supplemental Figure S2.1). While there have been relatively few comparable studies of fungi in Antarctic soils, the dominant taxa identified are generally similar to those found in other Antarctic soils (Baeza et al. 2017; Gomes et al. 2019). Archaeal sequences associated with the phyla Thaumarchaeota, family Nitrososphaeraceae, were detected in 60 samples, but these archaeal sequences made up a maximum of 1.5% of all the 16S rRNA gene reads per sample (mean = 0.08% of reads per sample).

We found that approximately 20% of the soils (40 out of 204 soils) had no amplifiable microbial DNA as determined by our cultivation-independent

sequencing approach (see methods for more details). This is an unusual result for a culture-independent analysis of soil as the microbial communities found in a wide range of soils across the globe have been successfully characterized using similar techniques (Delgado-Baquerizo et al. 2018), and we have shown that the soils themselves do not appear to be inhibiting DNA extraction or PCR amplification (Supplemental Figure S2.2). We recognize that these results do not confirm that these soils are completely sterile, and we acknowledge that we could have further optimized our methods to try to detect microbial DNA (e.g. extracting DNA from larger sample volumes, testing a range of different extraction kits and PCR protocols). Instead, we note that while all of these soils have very low biomass, 20% of the soils failed to yield detectable amounts of amplifiable bacterial, archaeal, or fungal DNA using techniques that are routinely used to successfully characterize microbial communities in soils from different regions of Antarctica (Goordial et al. 2016; Lambrechts et al. 2019; Smith et al. 2006).

To complement the amplicon sequencing effort, we also conducted quantitative PCR (qPCR) assays to measure concentrations of bacterial and fungal DNA in each of the 204 soil samples. The qPCR results confirm the sequencing-based results presented in Figure 2.1 in that those samples found to have no detectable microbial DNA via sequencing were nearly always the same samples found to have no detectable microbial DNA via qPCR. Specifically, bacterial DNA concentrations varied from 10^{-1} - 10^4 genome equivalents $\cdot$ g soil^{-1} across the 204 samples, with 40 of the 50 samples found to have no detectable bacterial DNA via amplicon sequencing also having no detectable bacterial DNA as determined from the qPCR assays. In general, there were more samples that had qPCR DNA concentrations below detectable limits, but had detectable DNA as determined by amplicon sequencing, than vice versa. For fungi, measured fungal DNA concentrations varied from 10^{-1} to 106 genome $\cdot$ g soil^{-1}, but 96 of the 143 samples

found to have no detectable fungal DNA via amplicon sequencing also had fungal DNA concentrations near or below the level of qPCR detection.

Habitability Tests

To test whether we can detect microbial life and activity in soils where microbial DNA was near or below detectable levels (using cultivation-independent sequencing), we analyzed a subset of 35 samples with culture-based methods, metabolic assays to detect ^{13}C-glucose mineralization, and ATP quantification (in soils amended with and without glucose). This subset of 35 samples included both soils with and without microbial DNA (as determined by the cultivation-independent sequencing and qPCR, Figure 2.1, Figure 2.2, Supplemental Figure S2.3) with these soils collected from transects on each of the 10 features and spanning a range of edaphic properties (Supplemental Table S2.1).

The 35 soil samples were plated on 17 different types of solid culture media. All media targeted aerobic microorganisms (both heterotrophs and autotrophs), as obligate anaerobes are unlikely to be present in these extremely dry surface soils. This cultivation effort supports the results from the cultivation-independent analyses (Figure 2.2). Of the 35 samples tested; 4 samples did not grow any colonies on any of the different media types after three months. In general, those samples which did not yield any detectable microbial DNA also failed to yield any colonies, even when using such a broad range of aerobic media and relatively long incubation times. We acknowledge that our cultivation-based survey will not have captured all taxa potentially living in these samples as alternative cultivation strategies could always be employed, but the extensive cultivation-based analyses broadly agree with our cultivation-independent analyses in demonstrating that microbes are undetectable in an appreciable subset of the soil samples.
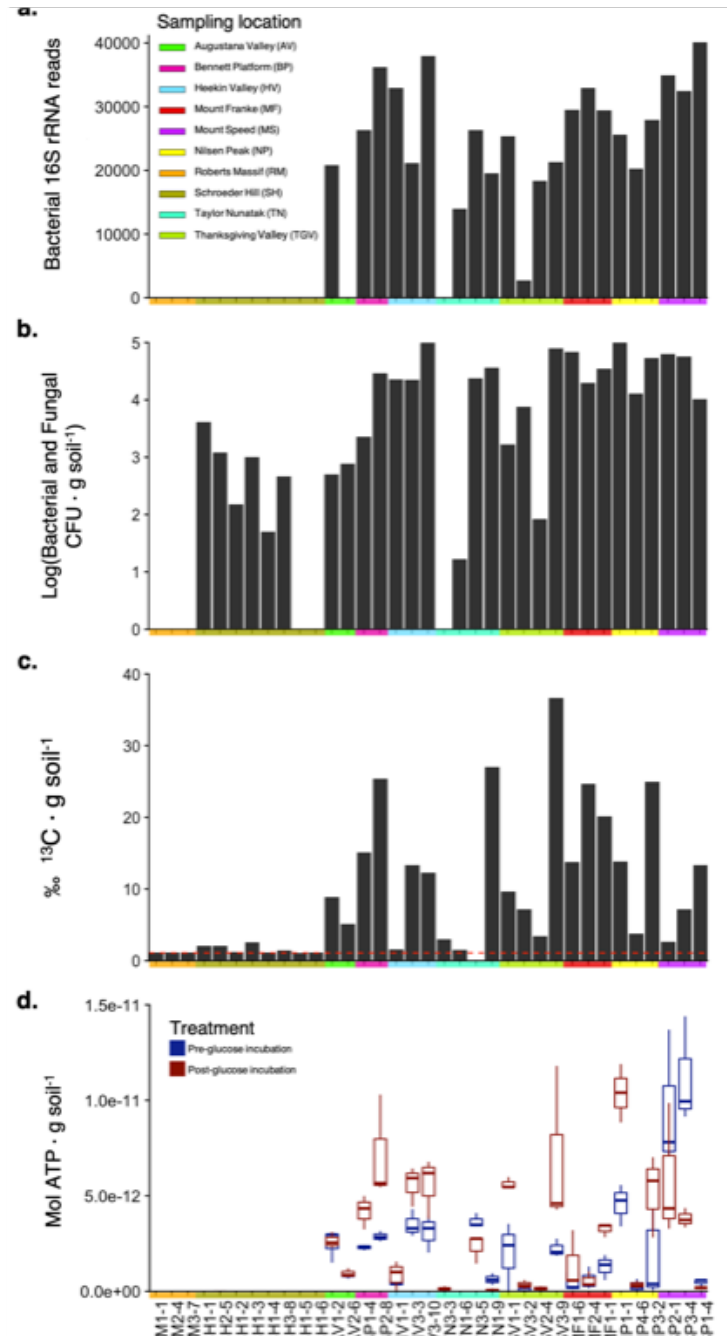
Figure 2.2: The four tests used to identify detect microorganisms and/or microbial activity in soils. **A)** The number of bacterial 16S rRNA reads that were classified to at least the phylum level of resolution for the subset of samples (n=35) used to verify the results of the culture-independent sequencing. Samples with a read number less than the threshold determined to identify and reliably detect prokaryotic DNA (see Methods) have been omitted. For all panels, samples are arranged on the x-axis by elevation (higher elevation features on the left). **B)** The log transformed total colony counts (Log(CFU · g soil⁻¹)) of microorganisms (both bacteria and fungi) grown from each of the 35 samples. The reported total cell

numbers are the cumulative number of colonies that grew on each plate over the three-month incubation period. Samples with no bar had no growth, i.e. no colonies detected. **C)** The ^{13}C ‰ released as CO_2 from the glucose-amended soils as calculated from the unautoclaved replicates of each of the 35 samples. The dashed red line indicates the ^{13}C ‰ measured in the blank samples. 3 samples (RM3-7, SH1-5, SH1-6) had ^{13}C ‰ from their unautoclaved replicates that was not significantly different from the values calculated from either their autoclaved replicated or the blank tubes and were below detection limits. Sample TN3-5, indicated by a star, was not successfully measured. **D)** ATP concentrations measured from the soils. Blue boxes were baseline readings performed on the soils before any glucose was added and before any incubation. Red boxes are from paired soil sub-samples amended with glucose and incubated for 24 h. Samples that are blank indicate that the ATP readings were below detection limits.

Of those soils (29 out of 35) which yielded visible bacterial or fungal colonies after 8 weeks, estimated cell numbers were highly variable, ranging from <20 to >130,000 CFU $\cdot$ g soil^{-1} (Figure 2.2). Even the soil with the highest cultivable cell numbers had values that were 50 times smaller than the 'positive control' soil collected from a lawn in Colorado (7.4 x 10^6 CFU $\cdot$ g soil^{-1}). These results are not surprising as we would expect ice-free Antarctic soils to have relatively low cultivable cell numbers (Lambrechts et al. 2019). However, it is worth noting that, while bacterial isolates dominated the culture collection, we were also able to culture and isolate viable fungi from 14 of these soils including some soils from which no bacteria could be cultured (Figure 2.2). These results extend the cultivation-independent fungal analyses and the qPCR results highlighted above in demonstrating that even viable fungi are often present and that bacteria are not the sole members of these Antarctic soil microbial communities. This prevalence of fungi in soils from higher elevation sites is consistent with the culture-independent results (Figure 2.1, Supplemental Figure S2.4) and suggests that certain fungi may in fact be better adapted than bacteria or archaea to some of the most challenging soil environments on Earth.

To further assess microbial activity, we performed an assay to detect microbial activity in these soils based on the catabolic mineralization of 13C-labeled glucose added to soils (Figure 2.2). We found that the 32 of the 35 samples tested had $^{13}C\text{-}CO_2$ values that were significantly higher than their autoclaved replicates and the corresponding control blanks. Three of the 35 samples had non-autoclaved $^{13}C\text{-}CO_2$ values that were indistinguishable from their paired autoclaved replicates or from the soil-free blanks. This suggests that we were unable to detect glucose mineralization resulting from microbial activity in these 3 soil samples (Supplemental Table S2.2). This same subset of soils also had no culturable bacteria or fungi (Figure 2.2), confirming that microbial cells were not detectable using the methods employed.

As a final test for microbial activity, we measured soil ATP concentrations both with and without glucose amendments (Figure 2.2). In theory, ATP assays should be able to detect very low levels of microbial activity that might not be detectable using other methods (Cowan et al. 2002; Schuerger et al. 2008). We found that 21 of 35 of the soils had measurable ATP concentrations with the addition of the PBS buffer only, while 23 of 35 had measurable ATP activity after amendment with glucose and incubated for 24 h. ATP production increased after glucose amendment in nearly all of the samples that had measurable ATP in the corresponding unamended soil. All reagent blanks and blank wells were below ATP detection limits of 10^{-15} mol ATP $\cdot$ g soil^{-1}. ATP concentrations in 12 of the 35 samples were below detection both with and without glucose amendment. These results suggest that there was no measurable microbial ATP in these soils, further evidence for a lack of actual or potential microbial activity in a subset of the samples collected.

Together these five distinct methods (cultivation-independent marker gene sequencing, qPCR, the extensive cultivation effort, the whole-soil $^{13}C\text{-}glucose$

metabolic assay, and the ATP assays, Figure 2.2) yielded similar results. We were unable to detect microorganisms or microbial activity in a subset of the samples (3 of 35) regardless of the methods employed: no colonies grew on any aerobic media type, no microbial metabolic activity was detected, and we could not detect any measurable ATP production. Although microscopy-based cell counting can be used to quantify microbial cell numbers in higher biomass soils, we did not use such an approach here given the problems associated with distinguishing between cells and soil particles in these soil types (Klauth et al. 2004).

We recognize that our tests do not prove the absence of living microbes and that there are always more strategies that could be employed. However, our use of multiple distinct methods suggests that any microbial life that may exist in these soils is below the limit of detection using methods that are routinely used in other low biomass microbial systems (Caporaso et al. 2012; Delgado-Baquerizo et al. 2018; Dineen et al. 2010; Hynek et al. 2018; Schmidt et al. 2018; Thompson et al. 2017; Vimercati et al. 2019). These soils may contain low levels of microbial biomass that we were unable to detect with these analytical approaches, which is why we do not conclude that the soils are sterile. Rather we emphasize that our results show that microbial life is restricted in certain soils to the point where we cannot detect microbes using the methods employed.

Why might microbial life be restricted in certain soils?

We next sought to identify patterns in microbial habitability and determine what soil or site characteristics differentiated those soils that had no detectable microorganisms from those in the same region that had detectable microorganisms. As previously mentioned, we were unable to perform all of our habitability tests on our full dataset of 204 samples. Thus, our random forest model was based on the results from the cultivation-independent marker gene sequencing, with those

samples that yielded no archaeal, bacterial, or fungal reads considered to represent soils with the potential to have no detectable microbial DNA. We found that elevation and soil chlorate concentrations were the best predictors of whether microbes could be detected in a soil (Figure 2.3, Supplemental Figure S2.5, Supplemental Table S2.3). Soils from lower elevations were more likely to have detectable microorganisms (37% of soils from sites >1190 m.a.s.l had detectable microbial DNA while only 3% of soils from sites <1190 m.a.s.l. had detectable DNA) and soils with lower chlorate concentrations were more likely to have detectable microbial DNA (soils with no detectable DNA or detectable DNA had average chlorate concentrations of 1675± 2595 µg·kg^{-1} and 161 ±1273 µg·kg^{-1}, respectively). Soils with no detectable microbial DNA also had higher median total salt and perchlorate concentrations (Supplemental Figure S2.5), but these differences were not significant (Supplemental Table S2.3).

Elevation and chlorate concentrations may be the best predictors of whether a soil contains detectable microorganisms, but it is unlikely that these factors, singly or in combination, are solely responsible for restricting microbial life in these soils. We know, for example, that there are soils at much higher elevations that have active microbial communities (Schmidt et al. 2018; Vimercati et al. 2019) and microorganisms have been found in soils with higher concentrations of deposited chlorate in the Atacama Desert and in the McMurdo Dry Valleys (Coates and Achenbach 2004; Rao et al. 2010). A more parsimonious explanation is that these variables are indicative of a suite of environmental properties that together restrict microbial life. Elevation likely correlates with a number of variables that increase in magnitude further inland and higher in elevation in the Shackleton Glacier region that may reduce microbial habitability: higher UV radiation, lower temperatures, lower water availability, and increased soluble salt concentrations (Diaz et al., 2021). The accumulation of high concentrations of chlorate salts, on the

other hand, may influence habitability due to its toxic properties but this seems unlikely due to the fact that we observed soils with detectable and undetectable microbial communities in close proximity across this region (Figure 2.1). Instead, high concentrations of chlorate, a highly soluble molecule, may indicate that these soils seldom, if ever, have had sufficient amounts of liquid water to sustain active microbial communities for longer than brief periods (Rao et al. 2010).
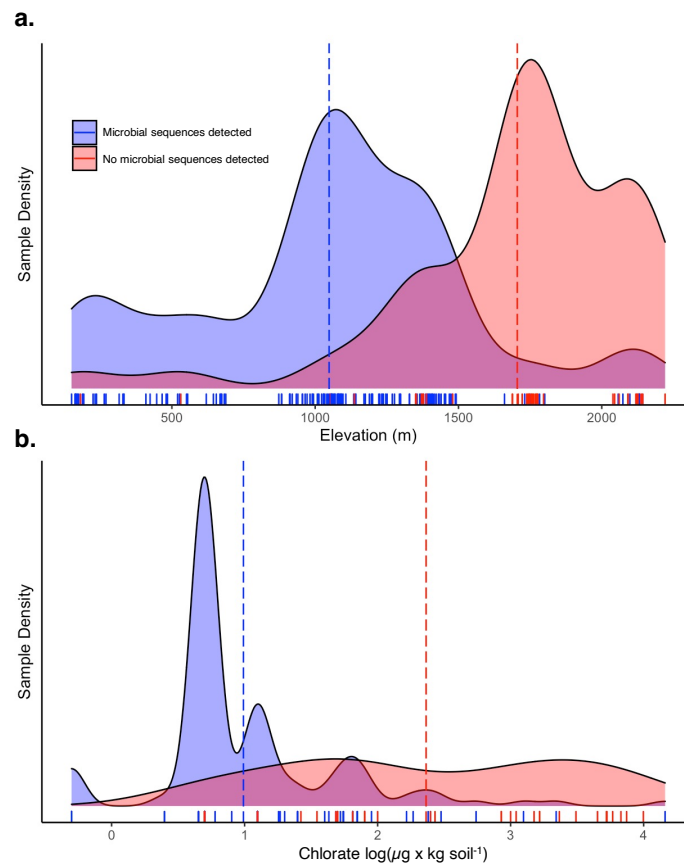
Figure 2.3: The best predictors of the habitability of a soil are elevation and chlorate concentration based on the random forest model. The distribution of samples with or without microbial DNA detected across the Shackleton Glacier Region in relationship to elevation (panel **A**) and soil chlorate concentrations (panel **B**). Y values show frequency density of each group with n=204 for elevation and n= 169 for chlorate. The area under each curve is equal to 1. Concentrations of chlorate ions (µg · kg soil^{-1}) have been log transformed. Dashed lines indicate the mean value for each group. To see the distribution of all other variables tested in the random forest model and the model results, see Supplemental Table S2.5 and Supplemental Table S2.3.

There is not likely a single environmental factor that restricts microbial life in Antarctic soils. Instead, it is likely that the constant exposure to deposited salts, low temperatures, and extremely low water potentials over thousands of years have created surface environmental conditions that repress the effective colonization, activity, and establishment of those microbes that may reach these sites via aeolian transport (Archer et al. 2019). The impact of these constraints likely change constantly, due to transient fluctuations in temperature, water availability, and salt concentrations. Clearly microbial activity in challenging soil environments is not ubiquitously distributed and local, landscape-level differences in environmental conditions can impose limits on microbial habitability in certain Antarctic soils.

Discussion

Our finding that some ice-free soils in the Shackleton Glacier region of Antarctica contain not detectable microbial life using standard methods is significant. These soils were neither recently formed, nor did they come from completely isolated areas. Instead, these soils are likely very old, and some were found on features in reasonably close proximity (within 50 m) to soils that contained detectable microbial life. This distinction is noteworthy as young soils may not have had a sufficient amount of time to accumulate microbial biomass. Instead it appears that a unique set of environmental conditions, those associated with higher elevation sites further inland and a near complete lack of liquid water over time, create conditions unfavorable to microorganisms. Soils typically contain large numbers of active microorganisms (often >1000 kg of microbial biomass carbon per hectare, Fierer (2017)) and even dry polar desert soils from the McMurdo Dry Valleys tend to have detectable, diverse, and active microbial communities (Cary et al. 2010; Goordial et al. 2016). To find soils with such restricted microbial habitability is unexpected.

Ice-free soils are not unique to the Shackleton Glacier region. More than 45,000 km2 of ice-free terrestrial surfaces can be found in Antarctica >5km from the coast (Brooks et al. 2019). The environmental and geochemical characteristics acting on these inland, high elevation, ice- free soils are similar to those we observed in the Shackleton Glacier region (Bockheim 1997; Magalhães et al. 2012) which suggests that the patterns in habitability we observed may be common elsewhere in Antarctica. This possibility is supported by the research of Goordial et al. (2016), whose work on soils in the University Valley also suggests that the unique combination of conditions found in Antarctica can severely limit microbial activity and survival. A similar constraint on habitability may also exist at higher elevation sites in the Altiplano of South America (Schmidt et al. 2018), but the absence of detectable microbial life in these soils remains hypothetical. Regardless, soils with such restricted microbial biomass and activity are likely very rare outside of Antarctica, as other ultra-arid regions, like the Atacama and Namib Deserts, harbor detectable microbial communities (Bull and Asenjo 2013; Schmidt et al. 2018).

We hypothesize that the microbial life in even the most challenging soils may not remain below detectable limits for long. Antarctic terrestrial systems have changed very little since the Last Glacial Maximum due to relatively stable climatic conditions (Thompson and Solomon 2002; Turner et al. 2014). Increases in temperature, precipitation and moisture availability are predicted to accelerate at high elevations in coming years due to ongoing climate change (Turner et al. 2014). These anticipated environmental changes would likely expand the range of habitable soil environments in Antarctica (Le Roux and McGeoch 2008) and decrease the likelihood of finding soils with no detectable microbial life in the future.

Conclusions

The existence of Antarctic soils lacking detectable microbial life, is significant to the study of terrestrial biology and astrobiology. We are not suggesting that we have found "lifeless" or "sterile" soils, nor have we identified the low temperature threshold for life (many locations on Earth maintain lower temperatures than the Shackleton Glacier Region and contain active microorganisms, Merino et al. (2019)). However, our inability to detect microbes in certain soils suggests that microbial activity and survival in Antarctic may be restricted by exposure to extremely cold, dry, and salty conditions. This hypothesis is supported by previous work performed elsewhere in Antarctica (Goordial et al. 2016) and is a different type of "limit to life" than what might be found in a hot, acidic environment (Belilla et al. 2019; Merino et al. 2019). Acknowledging that certain Antarctic soils may be more or less habitable will allow us to better understand the adaptations that allow organisms to survive and remain active in these unique, challenging environments and predict what other soils on Earth may be similarly restrictive to life. Finally, understanding patterns of terrestrial habitability on our planet will set the groundwork to better predict where microbes might, or might not, be found on other planets.

CHAPTER III

ELEVATIONAL CONSTRAINTS ON THE COMPOSITION AND GENOMIC
ATTRIBUTES OF MICROBIAL COMMUNITIES IN ANTARCTIC SOILS

Abstract

The inland soils found on the Antarctic continent represent one of the more
challenging environments for microbial life on Earth. Nevertheless, Antarctic soils
harbor unique bacterial and archaeal (prokaryotic) communities able to cope with
extremely cold and dry conditions. These communities are not homogeneous, and
the taxonomic composition and functional capabilities (genomic attributes) of these
communities across environmental gradients remain largely undetermined. We
analyzed the prokaryotic communities in soils collected from across the Shackleton
Glacier region of Antarctica by coupling quantitative PCR, marker gene amplicon
sequencing, and shotgun metagenomic sequencing. We found that elevation was the
dominant factor explaining differences in the structure of the soil prokaryotic
communities with the drier and saltier soils found at higher elevations harboring
less diverse communities and unique assemblages of co-occurring taxa. The higher
elevation soil communities also had lower maximum potential growth rates (as
inferred from metagenome-based estimates of codon usage bias) and an over-
representation of genes associated with trace gas metabolism. Together these
results highlight the utility of assessing community shifts across pronounced
environmental gradients to improve our understanding of the microbial diversity

found in Antarctic soils and the strategies used by soil microbes to persist at the limits of habitability.

Introduction

Not all of Antarctica is covered by ice. Ice-free surfaces in Antarctica represent >54,000 km^2 (~0.5%) of the total land area of the continent and most of these ice-free areas are located >5 km from the coast (Brooks et al. 2019). These inland soils can vary in age, from incipient soils that were recently covered in ice to soils that have been ice-free and developing in place for thousands of years or even longer (Diaz et al. 2021). The environmental conditions and geochemical characteristics of Antarctic soils can be highly variable (Cary et al. 2010; Goordial et al. 2016). However, nearly all of these soils developed in extremely cold and dry conditions - some of the coldest and driest conditions on Earth (Merino et al. 2019). Most Antarctic soils have extremely low organic carbon concentrations (Beilke and Bockheim 2013; Bockheim and Haus 2014) and the near-complete absence of liquid water and associated leaching, can lead many Antarctic soils (particularly those at higher elevations) to accumulate high concentrations of salts over time (Diaz et al. 2021). These salts include nitrate (NO_3^-), sulfate (SO_4^{2-}), perchlorate (ClO_4^-), and chlorate (ClO_3^-) salts derived from atmospheric deposition and chemical weathering (Diaz et al. 2021; Campbell and Claridge 1987; Ugolini and Bockheim 2008). Despite the extremely challenging conditions, Antarctic soils can harbor diverse and active microbial communities (Cary et al. 2010; Cowan et al. 2014).

Which microbes can persist in Antarctic soils and how they are able to tolerate the challenging environmental conditions have long been of interest to scientists (Cary et al. 2010; Goordial et al. 2016; Cowan et al. 2014; Aislabie et al. 2006; Chan et al. 2013). From this previous work, we know that soil microbial communities found in Antarctica are distinct from those in more temperate

ecosystems - distinct both with respect to their taxonomic composition and their genomic attributes (Aislabie et al. 2006; Chan et al. 2013; Cowan et al. 2015; Fierer et al. 2012). Antarctic soils are typically dominated by members of the bacterial phyla Actinobacteria, Firmicutes, Bacteroidetes, and Proteobacteria (Goordial et al. 2016; Cowan et al. 2014). Although these broader taxonomic groups also occur in more temperate soils (Delgado-Baquerizo et al. 2018), the specific bacterial taxa and lineages found in Antarctic soils are distinct and often only commonly found in Antarctic soils or other hyper-arid environments (Lambrechts et al. 2019).

Antarctic soil prokaryotes (bacteria and archaea) are not only taxonomically unique, they also have specific adaptations for life in Antarctica (Chan et al. 2013; Cowan et al. 2015). These adaptations can include those related to osmoregulation and psychrophily that allow microbes to maintain homeostasis and survive Antarctic conditions (Merino et al. 2019; Cowan et al. 2015). Microbial communities in Antarctica also use a variety of metabolic pathways to survive in the resource limited environments typical of most Antarctic soils (Lee et al. 2019; Ji et al. 2016; Ortiz et al. 2021). For example, there is accumulating evidence that the metabolism of atmospheric trace gases (including CO, H_2, and CO_2) is a key metabolic strategy used by microorganisms in Antarctica and other hyper-arid environments to fix carbon and generate energy (Ji et al. 2017; Leung et al. 2020).

Despite the distinct nature of Antarctic soil prokaryotic communities, they are not homogeneous. Previous studies, using both cultivation-dependent and cultivation-independent approaches, have documented a high degree of variance in the composition of Antarctic soil prokaryotic communities (Cary et al. 2010; Bottos et al. 2014). While this variability can be attributed to a range of soil and site factors, some of the more important factors shaping the composition of Antarctic soil prokaryotic communities can include temperature, water availability, soil pH, and soil salt concentrations, recognizing that many of these variables often co-vary

across Antarctic landscapes with drier soils at higher elevations often having higher pH and higher salt concentrations (Lee et al. 2018; Horn et al. 2013).

The relatively low diversity of most Antarctic soil microbial communities and the pronounced environmental gradients that can be found across sites in Antarctica (even sites located in close proximity), make this system well-suited for investigating how communities vary across environmental gradients, a core concept in both macroecology and microbial ecology. Likewise, the reduced diversity of Antarctic soil microbial communities, relative to the highly diverse soil communities typical of more temperate environments, makes it possible to relate the taxonomic composition of prokaryotic communities to differences in the functional attributes of these communities (Fierer 2017). Most soil microbes, including those found in Antarctica (Lambrechts et al. 2019), are difficult to cultivate and study in the lab. Fortunately, with advances in cultivation- independent approaches, including shotgun metagenomic analyses, it is now feasible to pair taxonomic and genomic-based investigations of Antarctic soil microbial communities (Wei et al. 2016; Baeza et al. 2017; Jindal 2020). Documenting how the genomic attributes of microbial communities vary across Antarctic soils can contribute important insights into the functional capabilities and adaptations of these unique microbial communities. Based on previous work, we predict that the microbial communities from soils collected further inland, i.e., those exposed to more challenging conditions, would be less diverse than those closer to the coast, both with regard to their taxonomic and their functional diversity (Lee et al. 2018; Horn et al. 2013; Dragone et al. 2021). We also expected that soil communities found further inland at higher elevations would contain more specialized taxa with unique metabolic capabilities, including an increased reliance on trace gas metabolism, that allow them to persist in more resource limited and challenging environmental conditions.

Here, we analyzed 204 soils collected from the Shackleton Glacier region of Antarctica. This region (~84.5 to 86.4°S, ~174.1 to 177.4°W) includes many ice-free features adjacent to a ~130 km long and ~10 km wide S-N outlet glacier of the East Antarctic Ice Sheet (EAIS). These soils are highly variable with respect to their age (amount of time ice-free), geochemistry, and other site conditions (including elevation, temperature, and moisture availability (Diaz et al. 2021)). We analyzed the prokaryotic communities in these soils by coupling a variety of cultivation-independent analyses including: quantitative PCR, marker gene amplicon sequencing, and shotgun metagenomic sequencing. Specifically, we used this collection of soils and the associated microbial analyses to address two questions: 1) What is the observed variation in the taxonomic composition and genomic attributes of soil prokaryotic communities across the Shackleton Glacier Region?; and, 2) What soil and site factors explain the observed changes in microbial communities across the Shackleton Glacier Region?

Results and Discussion

General characteristics of the soil microbial communities across the Shackleton Glacier region

The soils used for this study represent a wide range of conditions found across the Shackleton Glacier region. For example, the sampling locations ranged in elevation from ~100m to over 2000 m.a.s.l and these soils contained a wide range in the concentration of soluble salts (average: 5.7×10^3 mg/kg, range: $12.6 - 6.7 \times 10^4$ mg/kg). Due to the ambient temperatures being well below freezing for most of the year, water availability in these soils is low (Diaz et al. 2021; Cowan et al. 2014). In general, higher elevation soils were farther from the Ross Ice Shelf and were drier (estimate based on age of last wetting as estimated from nitrate concentrations by Diaz et al. (2021)), saltier, and contained less organic carbon (Diaz et al. 2021;

Dragone et al. 2021). We note that all 10 of the soil and site variables used for downstream analyses (see Methods) were positively correlated with elevation ($r > 0.5$, $p < 0.05$), though for NH_3, SiO_2, and Cl^- this correlation was weaker ($r < 0.25$, $p < 0.05$) (Supplemental Figure 3.1). More specific information on the environmental and geochemical properties of these soils are in Diaz et al. (2021).
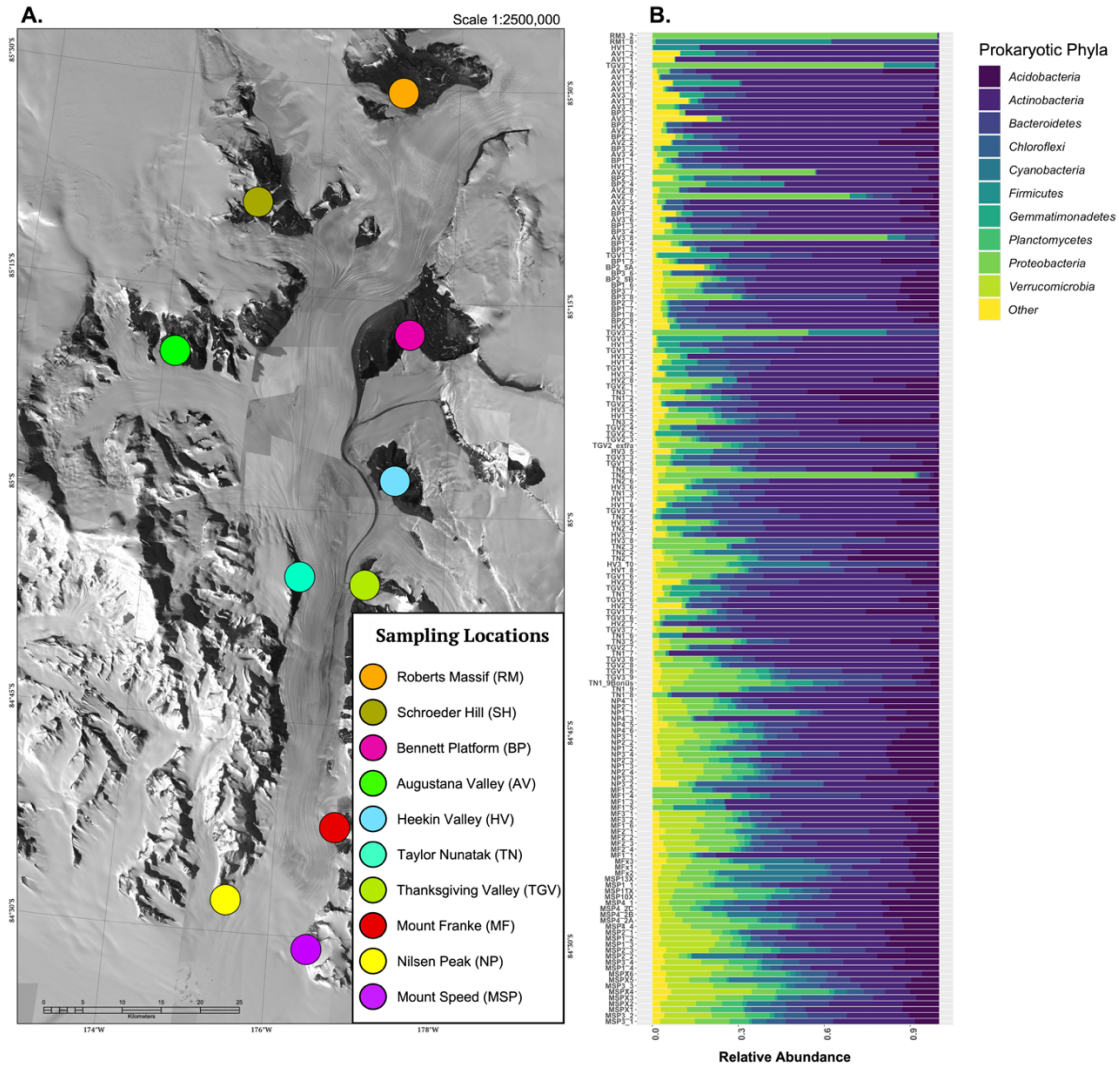
Figure 3.1: Overview of soil bacterial and fungal community composition across the Shackleton Glacier region. **A)** The Shackleton Glacier region (84 - 85°S, 174 – 177°W) with locations of the 10 features where samples were collected indicated by the colored dots (note that 14 - 26 samples were collected from transects at each of the 10 sampling locations). **B)** The relative abundances of the most abundant prokaryotic phyla for each of the 167 samples from which 16S rRNA marker gene sequencing were obtained. For panel B, samples are organized from highest elevation site at the top to the lowest elevation site at the bottom.

Only 167 of the 204 soils a yielded a sufficient number of prokaryotic 16S rRNA gene reads from the amplicon sequencing effort for inclusion in downstream analyses, see Methods and Dragone et al. (2021). The prokaryotic taxa with the highest relative abundances across these 167 soils included those assigned to the bacterial phyla Actinobacteria, Acidobacteria, Bacteriodetes, and Proteobacteria (Figure 3.1) which make up 46.0%, 11.6%, 10.0%, and 8.1% of total reads, respectively (Figure 3.1). Other phyla identified include Chloroflexi (6.2% of total reads), Verrucomicrobia (5.3%), Gemmatimonadetes (3.1%), Cyanobacteria (3.1%), Planctomycetes (2.8%), and Deinococcus-Thermus (1.5%). The most abundant ASVs across the region were assigned to the families Solirubrobacteraceae, Blastocatellaceae, Chitinophagaceae, and Rubrobacteriaceae (Figure 3.2). Archaeal sequences were found in 60 samples yet made up a maximum of 1.5% of all the 16S rRNA gene reads per sample (mean = 0.08% of reads per sample) and all were associated with the phylum Thaumarchaeota, family Nitrososphaeraceae. Prokaryotic richness, the number of distinct 16S rRNA gene phylotypes out of 2000 reads per sample, averaged 312 ASVs (21 – 853 ASVs) and prokaryotic genome equivalents measured with qPCR averaged $2.2\text{x}10^4$ genome equivalents $\cdot$ g soil^{-1} (Supplemental Figure S3.2).

The composition of the prokaryotic communities in these soils is consistent with results obtained using similar cultivation-independent analyses of other Antarctic soils, including those from the McMurdo Dry Valley region (Lee et al. 2018; Horn et al. 2013). However, we note that the composition of the microbial soil communities of the Shackleton Glacier region is highly variable, in particular with respect to the relative abundances of major taxonomic groups. For example, Actinobacteria and Chloroflexi were relatively more abundant in soils from the higher elevations (Figure 3.1). Most ASVs were detected only in a few soils; out of

the 8,641 prokaryotic ASVs identified through the 16S rRNA gene sequencing, 5,335 were found in <10 soils.

Figure 3.2: Phylogenetic tree of the taxa from the Shackleton Glacier region of Antarctica based on 16S rRNA gene sequencing (n = 167 samples). The ASVs represented in this tree include the top 100 most abundant ASVs that were identified from the 16S rRNA gene amplicon sequencing. The inset colors indicate the region of the tree associated with each bacterial phylum, while the grey bars around the ASV labels represent the bacterial order (internal dark grey ring) and family (external light grey ring). If no taxonomy is indicated, the ASV is not classified to that taxonomic level (taxonomy = "NA").

31

Key drivers of prokaryotic community composition

        Observed differences in the composition of the prokaryotic communities across the region were best described by a model that included elevation and total salt, perchlorate, and chlorate concentrations ($r = 0.62$, $p < 0.001$). Of these variables, elevation alone explained the majority of the dissimilarity ($r = 0.47$, $p < 0.001$), with the other variables contributing less to the overall correlation (total salt $r = 0.11$, perchlorate $r = 0.02$, chlorate $r = 0.03$, $p < 0.05$). The importance of elevation to the overall degree of dissimilarity in prokaryotic communities is further supported by the Mantel analyses which showed a reasonably strong correlation between elevation and Bray-Curtis distances across the 167 samples ($r = 0.45$, $p < 0.001$). These results suggest that elevation is the most important predictor of the degree of dissimilarity in prokaryotic communities across the Shackleton Glacier region of Antarctica. Elevation also had a reasonably strong influence on soil prokaryotic richness ($r = 0.56$, $p < 0.001$, Supplemental Figure S3.2). As elevation increased, we observed a steady decrease in prokaryotic richness up to ~2000 m. In general, the higher elevation soils also had the lowest concentrations of prokaryotic DNA estimated with qPCR, though this correlation was weak ($r = 0.09$, $p < 0.05$).

        Elevation is unlikely to be the sole factor driving observed differences in microbial community structure. Instead, these results support previous hypotheses that the soil environments found at higher elevations and further inland exert increasingly strong selective pressures on soil microbial communities (Cary et al. 2010; Goordial et al. 2016; Dragone et al. 2021) by virtue of these higher elevation soils being saltier, colder, and often drier. The differences in elevation may also be associated with other important variables we were unable to measure. For example, in Antarctic soils, elevation has been positively correlated with increases in UV radiation, decreases in temperature, and a decrease in available water (increased age of last wetting estimated from concentrations of water-soluble salts) (Diaz et al.

2021; Goordial et al. 2016). These variables are difficult to measure in this remote area where visits are short and infrequent, but all of these variables have been shown to have potentially important effects on Antarctic soil communities (Diaz et al. 2021; Merino et al. 2019; Dragone et al. 2021; Convey et al. 2014; Gilbert et al. 2010) (2, 6, 29 -31).

Certain taxa are associated with specific soil and site conditions

We were able to identify the environmental preferences of 28 of the 88 prokaryotic modules using network analyses and random forest analyses (Supplemental Table S3.1). The majority of these (16 of the 28 modules) were most strongly associated with elevation. For this reason, and the fact that nine of the remaining modules were predicted by variables strongly correlated with elevation (Supplemental Table S3.1, Supplemental Figure S3.3), we chose to focus the majority of our analyses and interpretation on the taxa assigned to modules associated with elevation. Out of the 16 modules best predicted by elevation, three were found to be associated with only the higher elevation sites (>800 m), 10 were found to be associated with only the lower elevation sites (<800 m) (Figure 3.3) and three modules were found to be associated with mid-elevation sites. There were more ASVs associated with low elevation modules (average = 38 ASVs, range = 2 - 154 ASVs) than with high elevation modules (average = 3 ASVs, range = 2 - 4 ASVs) and these 'low elevation' ASVs included representatives from 61 different families and 17 phyla. In comparison, all of the ASVs associated with the 'high elevation' modules consisted of taxa assigned to the phylum Actinobacteria (including taxa within the families Solirubrobacteraceae and Intrasporangiaceae) and Chloroflexi (Figure 3.3). Together, these results highlight that differences in elevation (or environmental variables associated with elevation) can explain a large portion of

the observed variation in overall community composition (Figure 3.1) and the distributions of particular prokaryotic taxa (Figure 3.3).
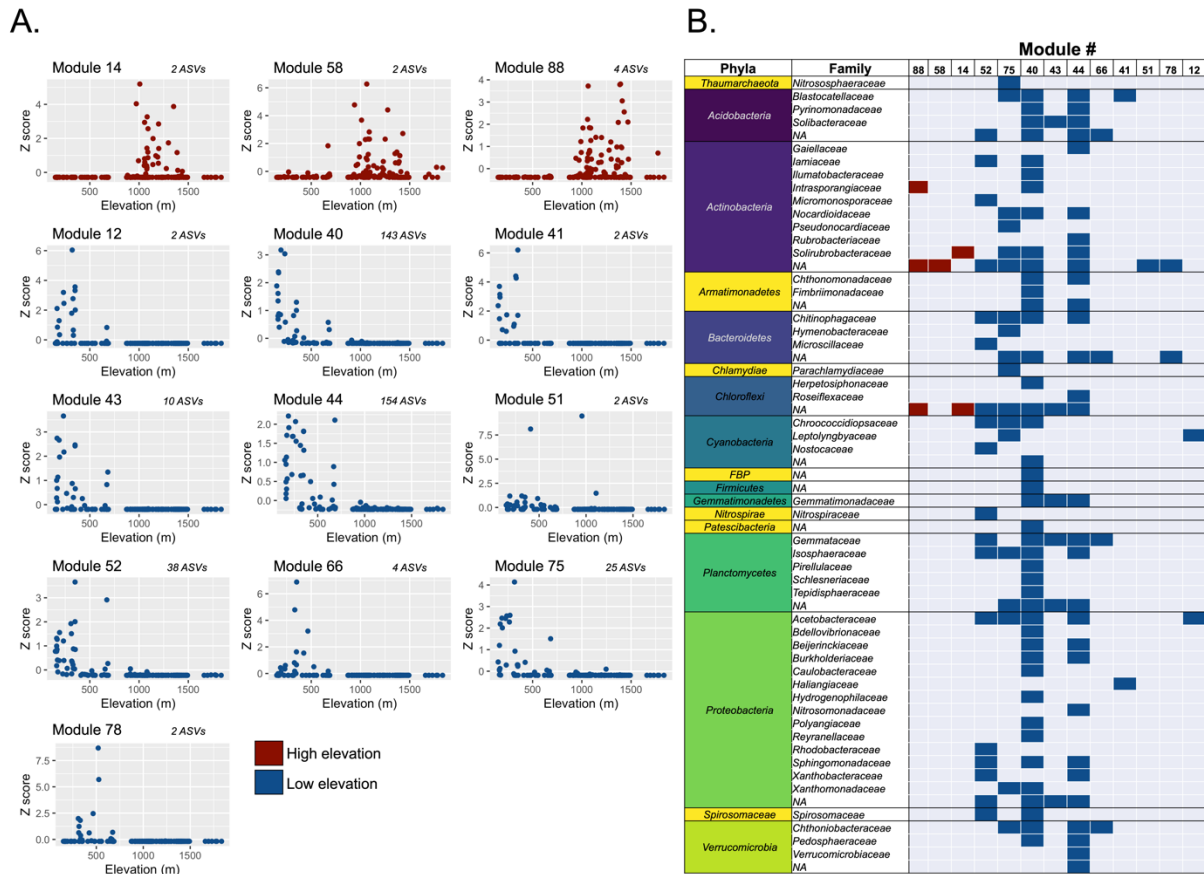

Figure 3.3: The taxonomic composition and elevational preferences of 13 prokaryotic modules identified from 167 samples. Modules are groups of ASVs (nearly all bacterial) that were identified as co-occurring based on the results of the network analysis. The modules displayed in this figure are modules with distributions best explained by site elevation, based on the results of the random forest analysis. **A)** The average standardized relative abundance (z score) plotted against elevation of the three high-elevation modules (colored in red) and the 10 low-elevation modules (colored in blue). The numbers of ASVs that are included in each module are listed next to the module number. **B)** The phyla and family-level taxonomic identities of those ASVs associated with each of the 13 modules (3 high-elevation, 10-low elevation).
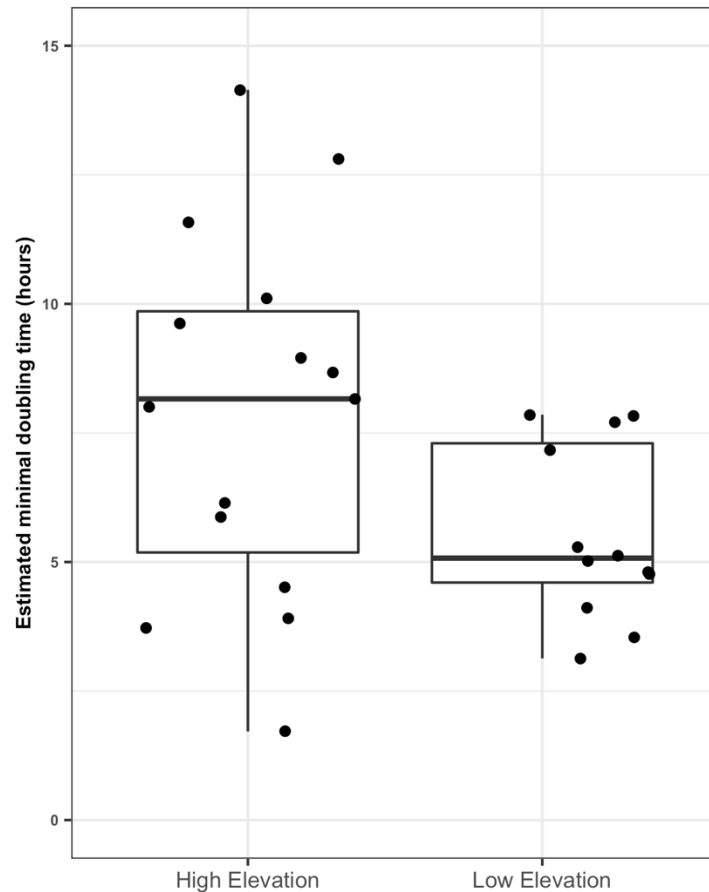
Figure 3.4: Patterns in estimated doubling time of samples (n = 27) across the Shackleton Glacier region. 15 samples make up the above 800 m a.s.l. "high elevation" group while 12 samples make up the below 800 m a.s.l. "low elevation" group. Estimated minimum doubling times, as inferred from gRodon (Weissman et al. 2021), were greater (lower maximum potential growth rates) in higher elevation samples than the lower elevation samples (Mann-Whitney U, p = 0.042). We note that, as gRodon was not designed for use in communities from these types of environments, the estimates of minimal doubling times are presented for comparative purposes only and the exact values should be considered with caution.

Genomic attributes of microbial communities

By analyzing shotgun metagenomic data obtained from 27 soils that were selected from the large sample set (see Methods), we found that, on average, the communities in the high elevation samples had longer estimated minimum doubling times (slower maximum potential growth rates) than the communities in the low

elevation samples (Mann-Whitney U, p = 0.042, Figure 3.4), though we caution that maximal potential growth rates should be considered estimates for comparative purposes only, not actual growth rates. A comparison of the dissimilarities in the functional gene profiles (relative abundance of functional genes) was positively correlated with the dissimilarities in the taxonomic composition of the prokaryotic communities (n = 27, r = 0.70, p < 0.001). This suggests that overall taxonomic dissimilarity in communities can be used to predict community-level differences in functional gene composition, a pattern consistent with observations from other metagenomic studies (Fierer et al. 2012; Gilbert et al. 2010). We also found that the number of distinct functional genes (richness) in the high elevation samples was significantly lower than in the low elevation samples (Mann-Whitney U, p < 0.001, Supplemental Figure S3.4). A larger number of genes were more than twice as abundant at low elevations than high elevations (6406 KEGGs) compared to those more abundant at higher elevations than low elevations (918 KEGGs). Functional gene richness was well-correlated with the observed patterns in taxonomic richness across these samples (n = 27, r = 0.77, p < 0.001).

The functional gene analyses to identify potential functional pathways of interest that may be over-represented in high elevation Antarctic soils identified many functional pathways associated with metabolism (49% of pathways containing high elevation-associated genes are related to metabolism) (Supplemental Figure S3.5). This identification of high elevation-associated metabolic genes supports previous work suggesting that microbial communities use a greater variety of genes coding for metabolic pathways in more challenging Antarctic environments (Lee et al. 2019; Ji et al. 2016; Ortiz et al. 2021). Of note, the metabolic pathway with the greatest number of genes over-represented in the higher elevation soils was the 'methane metabolism' pathway (KO00680). Trace gas metabolisms, including methanotrophy, are important metabolic strategies used by microorganisms in

Antarctica and other hyper-arid environments to generate energy and fix carbon (Ortiz et al. 2021; Ji et al. 2017; Ryan C. Lynch et al. 2014), yet this evidence suggests that the relative importance of these metabolic strategies may increase under more challenging conditions.

To complement the functional gene analyses described above, which only focused on broader gene categories, we performed more targeted analyses to compare abundances of genes associated with trace gas metabolism given that this category of functional genes was consistently over-represented in higher elevation soils (Supplemental Figure S3.5) and given the potential importance of trace gas metabolism as a strategy for microbial survival in hyper-arid systems (Bay et al. 2021; Ryan C. Lynch et al. 2014). We were able to identify genes associated with carbon monoxide oxidation, hydrogen oxidation, and methane oxidation in almost all the samples (Supplemental Figure S3.5). By comparing metagenomes from high elevation and low elevation soils, we found that five of the six genes involved in hydrogen oxidation were significantly more abundant in high elevation soils than low elevation soils, as was the gene coding for soluble methane monooxygenase (*MmoX*) (Figure 3.5, Supplemental Figure S3.6). Notably, the most abundant genes related to hydrogen oxidation were those in the recently identified group 1l [*NiFe*]-hydrogenases (*HylL*), which has been shown to be the primary catalysts of hydrogen oxidation in cold desert soils in other regions of Antarctica (Ortiz et al. 2021). The higher abundance of genes associated with hydrogen oxidation and methane oxidation suggests that trace gas metabolism may be particularly important for sustaining microbial life in higher elevation Antarctic soils. In contrast, the carbon monoxide dehydrogenase gene (*CoxL*) was found in almost all the samples and particulate methane monooxygenase (*PmoA*) was significantly more abundant in lower elevation soils (Figure 3.5, Supplemental Figure S3.6).
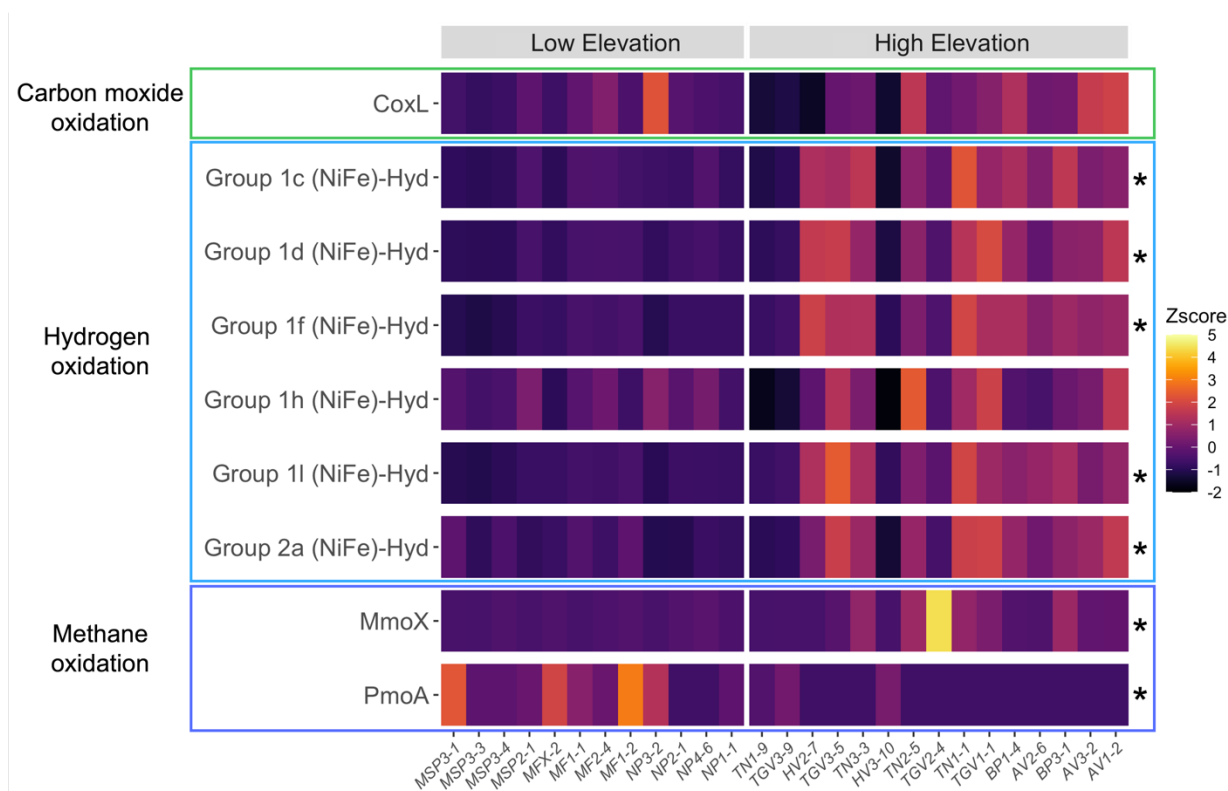
Figure 3.5: The abundance of nine genes associated with trace gas metabolism across the Shackleton Glacier Region. The 27 samples are grouped from lowest elevation to highest elevation and are grouped into "high elevation" and "low elevation" categories based on whether they were collected above or below 800 m a.s.l. Groups of genes associated with specific trace-gas oxidation pathways are outlined by colored boxes. Z scores were calculated based on the proportional gene abundances. Significant differences in gene abundances between the two groups are starred and the associated statistical information can be found in Supplemental Figure S3.6.

Our results build on those reported previously in highlighting that trace gas metabolism is likely an important metabolic strategy used by Antarctic soil microbial taxa (Ortiz et al. 2021; Ji et al. 2017) and supports previous work that atmospheric trace gas metabolism, and in particular hydrogen oxidation, is widespread in a variety of soil environments (Bay et al. 2021; Cordero et al. 2019; Lynch et al. 2014). These metabolic strategies may be particularly important in Antarctic soils, where biogeochemical studies suggest that the rate which soil

communities oxidize atmospheric trace gases may be sufficient to sustain their energy needs under certain conditions (Ji et al. 2017). The differential abundances of these genes across our dataset suggests that the importance of certain trace gas metabolisms by communities may vary depending on environmental or geochemical conditions. We may see more H_2 oxidation genes in higher elevation soil communities, for example, because H_2 occurs consistently in low abundance in the troposphere, with microbes using H_2 as an energy source to sustain metabolic activity in resource-limited environments (Ji et al. 2017; Ortiz et al. 2021; Leung et al. 2020; Lynch et al. 2014).

Conclusions

Microbial communities across the Shackleton Glacier region are highly variable in composition, and this variation is strongly associated with elevation (or environmental variables that strongly co-vary with elevation). Higher elevation soils typically had lower biomass, less diverse prokaryotic communities (Supplemental Figure S3.4), and communities with longer minimum doubling times (as estimated from metagenome-based analyses of codon usage bias, Figure 3.4). Likewise, we found that elevation was also associated with differences in the overall composition of the microbial communities, with the distributions of numerous specific prokaryotic taxa best predicted from site elevation (Supplemental Figure S3.3). Finally, the genomic attributes of the communities differed across the elevation gradient with a notable increase in the abundance of genes for trace gas metabolism in higher elevation soils and the prokaryotic communities found at higher elevations having lower estimated maximum potential growth rates (Supplemental Figure S3.4). The sampled elevation gradient captures a gradient in microbial community composition with the communities found in the more

challenging soil environments at the margins of habitability (Dragone et al. 2021) having distinct life history strategies and corresponding genomic attributes.

The patterns documented from our cultivation-independent methods would be difficult to infer from studying cultivated isolates, as most culturing studies are only able to isolate taxa that represented a fraction of the microbial diversity found in soil. Despite these limitations, detailed studies of cultivated isolates would make it possible to link the observed distribution patterns of particular taxa with experimental measurements of environmental preferences and tolerances (e.g. measuring growth responses across gradients in temperature and moisture in vitro). Likewise, while our analyses shed light on the potential functional attributes of high elevation and low elevation communities, the metabolic capabilities of specific taxa remain elusive. An important next step would be to identify which specific taxa harbor particular genes of interest, particularly those genes associated with metabolism of H_2, CO, and CH_4. This could be done by pairing community-level metagenomic analyses (as done here) with detailed analyses of particular metagenome-assembled genomes, as demonstrated previously (Ortiz et al. 2021). An integration of community-based and organismal-based ecological, genomic, and trait-based information will provide a more comprehensive understanding of microbial life in Antarctic soils and the adaptations that allow specific taxa to survive in one of the most challenging terrestrial environments on Earth.

CHAPTER IV

THE EARLY MICROBIAL COLONIZERS OF A SHORT LIVED VOLCANIC
ISLAND IN THE KINGDOM OF TONGA

Abstract

The island of Hunga Tonga Hunga Ha'apai (HTHH) in the Kingdom of

Tonga was formed by surtseyan eruptions and persisted for 7 years before being

obliterated by a massive volcanic eruption on January 15, 2022. Before it was

destroyed, HTHH was an unparalleled natural laboratory to study primary

succession on a newly-formed landmass. We characterized the microbial

communities found on the surface sediments of HTHH using a combination of

quantitative PCR, marker gene sequencing, and shotgun metagenomic analyses.

Contrary to expectations, photosynthetic cyanobacteria were not detected in these

sediments, even though they are typically dominant in the earliest stages of

primary succession in other terrestrial environments. Instead, our results suggest

that the early sediment communities were composed of a diverse array of bacterial

taxa, including trace gas oxidizers, anoxygenic photosynthesizers, and

chemolithotrophs capable of metabolizing inorganic sulfur, with these bacteria

likely sourced from nearby active geothermal environments. While the destruction

of HTHH makes it impossible to revisit the site to conduct in situ metabolic

measurements or observe how the microbial communities might have continued to

change over time, our results do suggest that the early microbial colonizers have

unique origins and metabolic capabilities.

Introduction

  Microorganisms are often the earliest colonizers of newly exposed or newly formed terrestrial surfaces (Fierer et al. 2010) with microorganisms having important roles in the development of ecosystems and the subsequent process of community succession (Connell and Slatyer 1977). In such environments, like a freshly cleaved rock face (Oosting and Anderson 1939; Woolhouse et al. 1985), recently deposited volcanic ash (del Moral and Wood 1993; Deligne et al. 2013), or sediments exposed following glacial retreat (Nemergut et al. 2007; Nicol et al. 2005), these early microbial colonizers tend to include oligotrophs, autotrophs, and other taxonomic groups that can survive in nutrient limited environments and/or fix carbon and nitrogen (Fierer et al. 2010; Brown and Jumpponen 2015; Ortiz-Álvarez et al. 2018). However, the identities of these earliest colonizers are likely dependent on the specific environment in question. For example, while photosynthetic cyanobacteria are often the earliest colonizers of sediments exposed after glacial retreat or newly formed sand dunes (Nemergut et al. 2007; Schmidt et al. 2008; Olff et al. 1993), the microbial colonizers of lava flows can also include many chemolithotrophs (King 2007).

  One of the most dramatic examples of microbial colonization occurs after the creation of new land. Most often this is a result of a volcanic eruption covering an existing surface with lava, ash, or tephra (del Moral and Wood 1993; Iimura et al. 2010). On occasion, volcanoes can also form completely new landmasses. 'Surtseyan' eruptions take place in shallow waters and rapidly push ejecta up through the water column (Kokelaar 1983; Schipper and White 2016), often leading to the formation of new islands (Wang et al. 2007; Thorarinsson 1967; Vaughan and Webley 2010). While many such surtseyan islands rapidly erode (within months to a few years) (Angus and Guest 1996; Ramalho et al. 2013), some may persist for long enough to be colonized by organisms (Brock 1973; Fattorini and Borges 2012). A

unique aspect of newly formed volcanic islands is the 'blank slate' they provide. Like other newly exposed terrestrial surfaces, there is initially little, if any, organic carbon on the sediments of new volcanic islands (Bishop et al. 2010; Göransson et al. 2011; Vitousek et al. 1993). Such volcanic substrates can also be challenging environments for organisms to recruit to and survive in because they typically contain high concentrations of heavy metals and are exposed to toxic volcanic gases, though the exact physical and chemical properties of these systems can vary depending on the geologic context (Deligne et al. 2013; King 2007; Fujimura et al. 2009; Iwasaki et al. 1963).

When the last two persisting surtseyan islands formed in the 1950s (Capelinhos, Azores) and 1960s (Surtsey, Iceland), research on these islands primarily focused on animal and plant colonizers, not microorganisms (Brock 1973; Fattorini and Borges 2012). For example, while qualitative observations of bacteria and algae were recorded as early as 1966 on Surtsey (Brock 1973; Brock 1966), the namesake of the eruption type, they were considered "extremely minor in importance"(Brock 1973). In fact, the first comprehensive survey of microorganisms on a surtseyan island did not occur until 2000, almost 40 years after the island of Surtsey had formed (Frederiksen et al. 2000). Thus, we have limited knowledge of the earliest microbial colonizers on newly formed volcanic islands. However, from work conducted in other volcanic systems, primary successional systems, and recent boreholes drilled on Surtsey (Bergsten et al. 2021; Jackson et al. 2019), we can hypothesize that the initial microbial colonizers likely include autotrophic taxa able to fix carbon by using light energy (phototrophs) or taxa that use inorganic energy sources to build biomass (chemolithotrophs) via trace gas oxidation, sulfur oxidation, and/or iron oxidation (G. M. King 2007; Fujimura et al. 2009; Magnússon et al. 2014; Dunfield and King 2004). The earliest microbial colonizers may also include oligotrophic heterotrophs able to survive on trace amounts of C and N in the

island's sediments or deposited by dust, sea spray, or from the atmosphere (Cockell et al. 2009; Ho et al. 2017).

On December 19, 2014, an underwater volcano in the Kingdom of Tonga in the southwestern Pacific Ocean, began a monthlong series of eruptions. A new island, referred to as Hunga Tonga Hunga Ha'apai (HTHH), had emerged by the end of January 2015 (Garvin et al. 2018; Hite et al. 2020). This new landmass formed between the two older islands of Hunga Tonga (HT) and Hunga Ha'apai (HH), connecting them with a ~120m tall, ~1.9 km2 tuff cone of tephra and ash. HTHH, the subaerial projection of the much larger Hunga volcano, was initially expected to erode after a few months, but instead persisted (Garvin et al. 2018). HTHH is only the third such landmass to have formed in the past 150 years that has persisted for longer than a year (Garvin et al. 2018). Unlike Capelinos and Surtsey which preceded it, HTHH formed in the tropics (latitude: 20.5^o S). While the tuff cone that formed in 2015 was destroyed during the explosive eruption of the HTHH volcano in 2022, which also stripped HT and HH of tens of meters of rock and sediment (Carr et al. 2022), samples of tephra collected from across the island three and four years after formation offered a rare opportunity to study the early microbial colonizers of the island's sediments. Using a suite of microbiological approaches, including cultivation independent marker gene sequencing, metagenomic "shotgun" sequencing, and quantitative PCR combined with a range of geochemical analyses, we addressed three questions: What taxa are the earliest microbial colonizers of sediments on HTHH?, From where did these microbial colonizers originate? and, What metabolic strategies are used by these microbes to persist in the challenging environmental conditions found on the recently formed landmass?
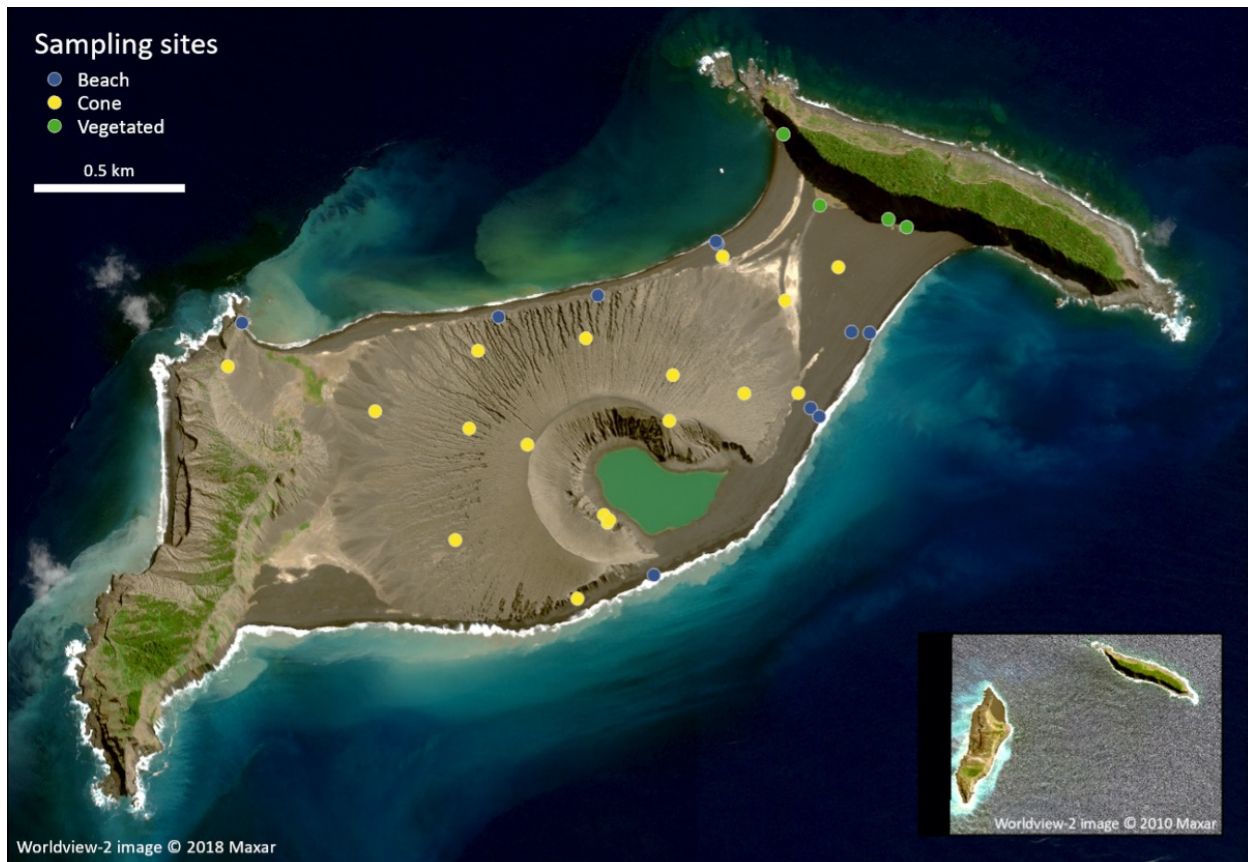
Figure 4.1. The island of Hunga Tonga Hunga Ha'apai, Kingdom of Tonga (Latitude: 20.536° S, Long: 175.382°W). The locations of the 32 surfaces where samples were collected. Background image is from August 19, 2018 and is orthorectified. The inset image displays the islands of Hunga Ha'apai (west) and Hunga Tonga (east) on September 11, 2010, prior to the 2014-2015 eruption. Imagery © Maxar.

Results and Discussion

We collected 32 samples across the 1.9km² volcanic cone of HTHH from surfaces that ranged from sea level to the summit of the crater ~120 m above sea level (m.a.s.l). While the focus of this study was on the unvegetated surfaces of the island cone, samples were also collected from sediments at the beach (marine-terrestrial interface) and from vegetated sediments around the island of HT that pre-date the 2014 - 2015 eruption cycle (Figure 4.1). While plants and animals could be found on the island at the time of collection, samples were collected from surfaces that were not visibly colonized by plants or animals unless otherwise noted (see

Figure 4.1, "vegetated" samples). For information on terrestrial and aquatic macrofauna and flora on HTHH see Hite et al. (2020) and Smallhorn-West et al. (2020).

The properties of the unvegetated, inland sediments collected from the tuff cone of HTHH suggest a distinct environment characterized by low concentrations of nutrients and organic carbon, high metal concentrations (including potentially toxic metals), and appreciable levels of sulfur (Supplemental Figure S4.1). Organic carbon concentrations in the cone sediments were low (avg: 0.32 mg · g^{-1}, range: 0.19 − 0.50 mg · g^{-1}) and typically 10 times lower than organic C concentrations in the 'vegetated' samples, which were collected from sediments where plants had recruited near the edge of HT (Figure 4.1, Supplemental Figure S4.1). The cone sediments had no detectable nitrogen but high concentrations of sulfur (avg: 2.1 mg · g^{-1}, range: 0.1 − 19.8 mg · g^{-1}), and iron (avg: 80 mg · g^{-1}, range: 74 − 86 mg · g^{-1}) (Supplemental Figure S4.1). The cone sediments also had high concentrations of other metals, including copper, vanadium, and cobalt concentrations that exceed those typically found in natural soils and are similar to the concentrations often found in metal-contaminated industrial sites and other volcanic systems (Supplemental Figure S4.1) (Flemming and Trevors 1989; Krishna and Govil 2007; Vukojević et al. 2019; Yang et al. 2017).

Bacterial and archaeal communities on HTHH

Bacteria and Archaea were detected in all the cone samples, but the amount of prokaryotic DNA, as determined via quantitative PCR, was two orders of magnitude lower in the cone samples than in the vegetated sediments collected around the island of Hunga Tonga (Supplemental Figure S4.2). The cone prokaryotic communities were also less diverse, with a mean of 108 amplicon sequence variants (ASVs) detected via targeted 16S rRNA gene sequencing as

compared to a mean of 473 ASVs in the vegetated samples (Supplemental Figure S4.3). The prokaryotic communities found in the unvegetated cone samples were also distinct in composition (Figure 4.2, Supplemental Figure S4.4) and dominated by members of the bacterial phylum Chloroflexi, which made up 26.4% of reads associated with these samples. Actinobacteria (18.1% of total reads), Firmicutes (15.7%), Proteobacteria (15.5%), and were also abundant. Other phyla that were identified in these samples included Bacteriodetes (6.2%), Planctomycetes (5.4%), Acidobacteria (3.3%), Cyanobacteria (2.8%), Gemmatimonadetes (1.5%), candidate phylum WPS-2 (Eremiobacteria, 1.5%). Archaea belonging to the phylum Thaumarchaeota were found in all samples, but they were relatively rare (<2% of reads across all samples). These taxonomic results obtained from targeted 16S rRNA gene sequencing mirrored results obtained by conducting shotgun metagenomic analyses on a subset of the same samples (Supplemental Figure S4.5). The dominant bacterial families identified in the cone samples included Acidiferrobacteraceae, Ktedonobacteraceae, and Sulfuricellaceae which include organisms that have been classified as autotrophic chemolithotrophs capable of oxidizing sulfur and iron based on studies conducted in other systems (Hu et al. 2018; Bennett et al. 2020). However, we note that many of the taxa come from groups that are poorly characterized. Of the top 100 taxa recovered with our targeted 16S rRNA gene sequencing, 40% could not be classified to a bacterial family (Figure 4.2). The early arrivals to the sediments of this new land mass are predominately bacterial taxa for which pre- existing information on their ecologies are limited, only 52% of the top 100 ASVs were from families that have previously been cultivated, thus necessitating the metagenomic-based analyses detailed below.
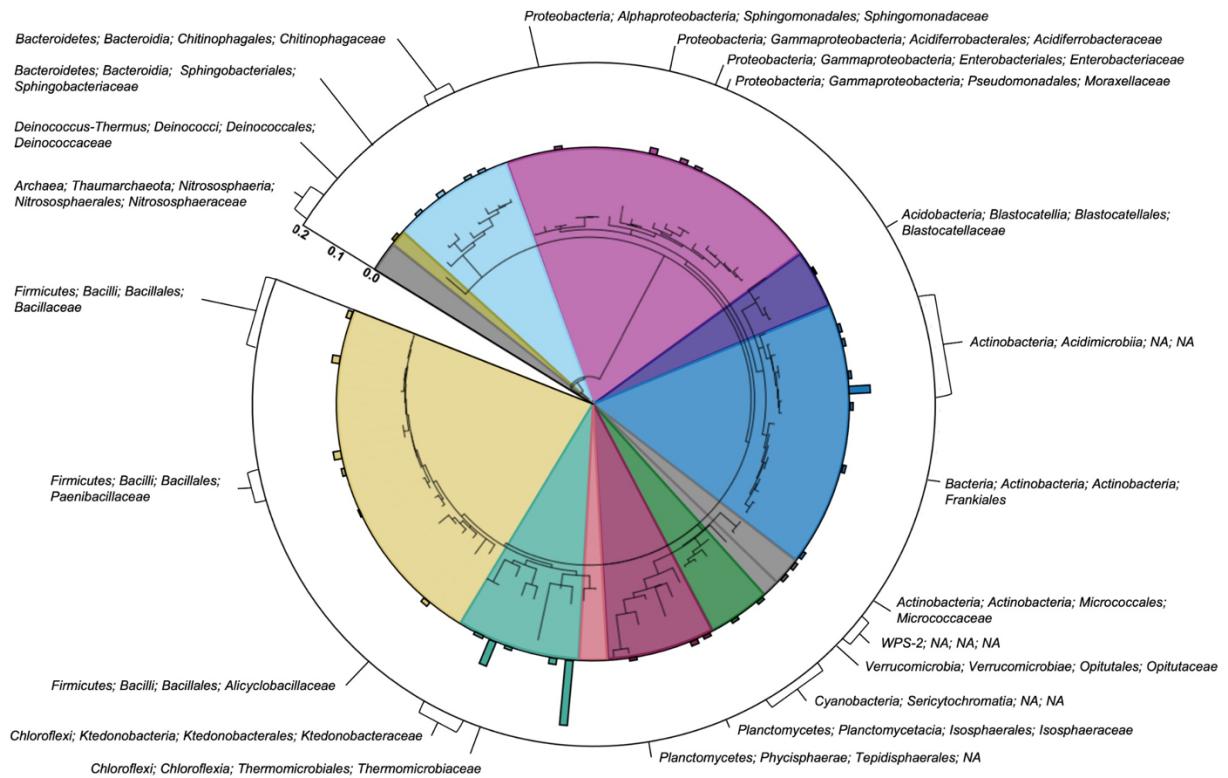
A.
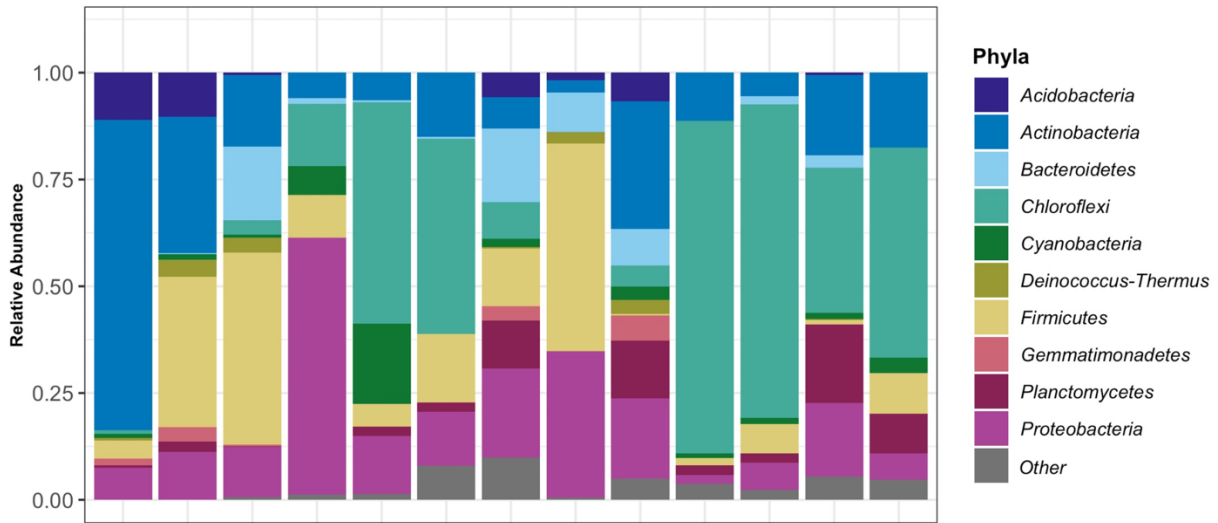
B.

Figure 4.2: Overview of the microbial community composition in the inland cone sediments. **A)** The proportional abundances of the dominant prokaryotic phyla found in each of the 13 inland cone samples from which 16S rRNA gene sequences were obtained. **B)** Phylogenetic tree of the most abundant amplicon sequence

variants (ASVs) identified in the inland cone samples of HTHH. The ASVs represented in this tree include the 100 most abundant prokaryotic ASVs that were identified from the 16S rRNA gene amplicon sequencing (2 archaeal sequences, 98 bacterial sequences). The inset colors indicate the region of the tree associated with each bacterial phylum and the proportional abundances of each ASV across the whole sample set are represented with the exterior bar plot. If an ASV made up less than 0.05% of reads, no bar is displayed. For those bacterial families that passed this abundance threshold, taxonomic information (phyla; class; order; family) is displayed outside of the bar plot. If an ASV could not be classified past the phylum level of resolution, no taxonomic label is included.

In contrast to the prokaryotic taxa commonly found in other newly exposed terrestrial surfaces, including surfaces exposed after glacial retreat (Fierer et al. 2010; Ortiz-Álvarez et al. 2018), the HTHH communities are distinct. For example, photosynthetic cyanobacteria, which are often considered to be characteristic of the earliest microbial communities (Nemergut et al. 2007), are completely absent, with the only cyanobacteria identified across all samples associated with the non-photosynthetic lineage of Sericytochromatia (Monchamp et al. 2019; Soo et al. 2017). We expect that the absence of cyanobacteria is due to the high concentrations of hydrogen sulfide (H_2S) typically released during volcanic activity as H_2S has been shown to restrict the growth of cyanobacteria and other oxygenic phototrophs (Cohen et al. 1986). However, we did find that members of the Chloroflexi phylum dominated the microbial communities in the HTHH sediments (Figure 4.2). While this phylum does not typically represent such large a proportion of the communities found in other early successional terrestrial environments (Schmidt et al. 2008), it is often found in volcanic environments where H_2S is present (King 2007; Fujimura et al. 2009; Iwasaki et al. 1963).

The cone samples are generally similar to those identified in other microbial communities described in studies of older volcanic deposits (sampled >10 years post-eruption) (King 2007; Fujimura et al. 2009; Dunfield and King 2004; Zeglin et al.

2016). To highlight one example, a study of basaltic lava deposits in Iceland found that the orders Planctomycetales, Rhizobiales, Rhodospirillales and Sphingomonadales were ubiquitous and abundant in all sampled basalts (Byloos et al. 2018). These orders were also dominant in the cone samples (Figure 4.2). Likewise, representatives of the phylum Chloroflexi, which were particularly abundant on HTHH, were also ubiquitous across Icelandic basalts (Byloos et al. 2018).

Source of bacterial and archaeal diversity

What are the likely sources of those microbes found to be dominant in the HTHH cone sediments? We might expect many of these microbes to be derived from surrounding ocean waters or from birds that deposit gut bacteria on the newly formed land mass. This does not seem to be the case. The bacterial taxa commonly associated with the bird microbiome and the taxa most often found in the marine environment were rare in the cone sediments (Supplemental Table S4.1) (Sunagawa et al. 2015; Capunitan et al. 2020), though we expect bird gut bacteria may be more abundant near nesting sites which were not sampled. Alternatively, the microbes we have classified on HTHH may have dispersed onto the cone from the soils found on the flanking islands of HT and HH that pre-dated the 2014 - 2015 eruption, either blown by the wind or carried by plants as they spread across the land bridge (Hite et al. 2020). While the microbial communities that dominate the vegetated sediments from HT (Figure 4.1) do share taxa with those in the HTHH cone samples (Supplemental Figure S4.6), our data do not suggest that these, or other, soils are the main source of microbes found in the cone sediments. Of the top 100 ASVs found in the cone samples, which make up on average 77% of the total reads in these samples ($50 - 94\%$ of reads), only 58 ASVs can be found in the vegetated samples, and they are generally rare, representing only 23% of the total reads in the

vegetated samples on average (6 – 57% of reads) (Supplemental Figure S4.7). Likewise, if we directly compare the ASVs from the cone samples to the 511 ASVs previously reported as being abundant and nearly ubiquitous in soils worldwide (Delgado-Baquerizo et al. 2018), we find that only 19 of the cone ASVs match with those in this reference database. In fact, only 19% of all reads from the cone sediments were assigned to bacterial classes that are the most common in soils (see Delgado-Baquerizo et al. (2018)), compared to 43% of all reads in the vegetated samples (Supplemental Table S4.1). While we cannot exclude the possibility that some of the cone microbes are derived from neighboring soils (or vice versa), the minimal overlap between the cone communities and those found on the neighboring islands or in a comprehensive database of global soils (Delgado-Baquerizo et al. 2018), suggests that most of the microbial colonizers on HTHH are unlikely to originate from soil.

Instead, we hypothesize that the microbial taxa found in the cone sediments may have been sourced from nearby volcanic systems and/or hydrothermal systems. While no detailed microbial information or equivalent sequence data is available from Tonga's active geothermal systems or from the seafloor and subsurface around HTHH pre-eruption, the communities we identified in our samples share similar taxa to those often found in these types of environments. Many of most abundant sequences we identified, including those assigned to Planctomycetales, Rhizobiales, Rhodospirillales and Sphingomonadales, are most similar to sequences in reference databases recovered in studies of volcanic environments in Iceland, Hawaii, New Zealand, and Alaska (Dunfield and King 2004; Zeglin et al. 2016; Byloos et al. 2018; Power et al. 2018). More specifically, we note that some of the more abundant taxa in the cone samples, including the uncultivated Chloroflexi which dominate our communities, are typically well represented in deep euphotic zone water (Orcutt et al. 2011), around hydrothermal vents (Zhou et al. 2020), and organic-poor seafloor

and subsurface sediments (Dick 2019). In fact, the most abundant 16S rRNA gene sequence found in the cone samples (ASV_1: Bacteria, Chloroflexi, AD3) is an exact sequence match to Chloroflexi recovered from sediments of the Brothers Volcano Complex in the Tonga- Kermadec Arc (Reysenbach et al. 2020) along with 12 of the other most abundant sequence variants. Finally, we also find that several of the more dominant taxa in the cone samples represented in our collection of metagenome-assembled genomes (MAGs) are most similar to MAGs obtained from Yellowstone hot springs (Bennett et al. 2020) and hydrothermal vent fields in the Atlantic and Pacific (Zhou et al. 2020; Kato et al. 2018).

The island of HTHH was simply the subaerial projection of the much larger submarine Hunga volcano, whose underwater caldera covers a total area of ~16km2. This volcano is extremely active, with large eruptions in 2009, 2014 - 2015, and 2022 following submarine and subaquatic venting and activity recorded since 1912 (Vaughan and Webley 2010; Garvin et al. 2018; Carr et al. 2022; Bryan et al. 1972). We see two potential mechanisms that could have brought organisms from subsea or subsurface to the subaerial cone. First, organisms found in the sediments around the active crater may have been transported up from the seafloor and subsurface sediments during the 2014-2015 eruption event that formed the island. Vertical microbial transport like this has been described on Surtsey where microbes found in the subsurface pore water and sediments, which includes Chloroflexi, were shown to be brought to the surface by fumarolic activity (Jackson et al. 2019). It is also possible that organisms may have been blown in from nearby subaerial volcanic systems. HTHH is just one of ~20 active volcanoes in Tonga. The nearest, Fonuafo'ou, although submarine, is only 25 km away and Tofua, which has frequent fumarolic activity is 100 km away (Bryan et al. 1972). Nearby explosive activity, like that of Late'iki in 2019 (Yeo et al. 2022), may have aerosolized and transported volcanic-associated microbes. From work conducted in New Zealand, we know that

such eruption-mediated microbial dispersal can transport viable microbial cells over distances exceeding 850 km (Van Eaton et al. 2013). However, due to the presence of the sediment-associated volcanic ASVs (described previously), we would expect that in the case of HTHH, microbial transport from the subsurface is a more likely explanation.

Inferred metabolic strategies of the microbial colonizers

Based on taxonomic information alone, we hypothesized that the microbial communities found on the newly formed land mass are dominated by chemolithotrophic bacteria and anoxygenic phototrophs. To infer the metabolic strategies used by these communities, we conducted shotgun metagenomic sequencing on a subset of samples, which included all the unvegetated, inland sediments collected from the cone (n = 13) with four vegetated samples included for comparison. We analyzed these metagenomic datasets to quantify the abundances of reads in each sample assigned to ~300 genes associated with metabolic strategies that we expect to be used by microbes to persist in this and other volcanic systems (King 2007; Dunfield and King 2004). We found that genes associated with pathways involving sulfur metabolism, CO oxidation, H_2 oxidation, and bacteriochlorophyll-mediated anoxygenic photosynthesis were enriched in the cone samples with the normalized abundance of genes associated with these pathways ~2 to 5 times higher than in the vegetated samples (Supplemental Table S4.2). For information on the relative abundances of genes associated with other functions, none of which were significantly different between the two categories of samples, see Supplemental Table S4.2.

Our finding that genes associated with sulfur metabolism are abundant in the HTHH sediment communities is in line with the high concentrations of sulfur measured in these samples. A closer investigation of the genes (see Yu et al. (2021)

for a full list), reveals that gene families coding for enzymes that are involved in the metabolism of thiosulfate were far more abundant in the cone sediment samples than in the nearby vegetated soils, specifically, thiosulfate reductase (*phsA, phsB, phsC*) and thiosulfate sulfurtransferase (*glpE*) (Figure 4.3). Other thiosulfate genes from the sulfur oxidation pathway, including *soxA, soxB, soxC, soxD, soxX, soxY, soxZ*, thiosulfate dehydrogenase (*tsdA, tsdB*), and thiosulfate sulfurtransferase (*sseA*), were found in all metagenomes though there were no significant differences between the two sample categories for any of these genes (Mann-Whitney U, p > 0.05, Figure 4.3). Thiosulfate acts as an intermediate in metabolic pathways that reduce, disproportionate, or oxidize inorganic sulfur compounds (Masuda et al. 2010) and thiosulfate is metabolized by almost all sulfur metabolizing chemolithotrophs, including those from deep sea hydrothermal vent systems (Nakagawa and Takai 2008) and volcanoes (King 2007). In the case of the surface sediments on HTHH, which we would expect to be well-oxygenated, these genes are not likely associated with anaerobic processes of sulfur reduction and disproportionation. Rather, our results suggest that the oxidation of inorganic sulfur compounds is likely an important strategy used by microbes to survive on HTHH.

Photosynthetic cyanobacteria were not found in the cone sediment samples and there was a corresponding absence of genes associated with oxygenic photosynthesis in the cone sediment metagenomes (Supplemental Table S4.2). However, we did observe that genes associated with other forms of phototrophy were enriched in the cone samples (Figure 4.3). In particular, the *pufM* and *pufL* genes were, on average, >3 times more abundant in the cone samples than in the vegetated soils (Figure 4.3). The *pufML* gene family codes for the type-II photosynthetic reaction center used by phototrophic organisms and anoxygenic photosynthesizers across a range of phyla (Imhoff et al. 2019), including some (like

Actinobacteria, Figure 4.3) that were abundant in the cone samples. However, we only found the *pufML* genes in two of our recovered MAGs, members of the families Beijerinckiaceae and Acetobaceraceae within the phyla Proteobacteria and Actinobacteria, respectively (Figure 4.3). While these putative phototrophs are not the dominant taxa in the cone samples, their presence suggests that anoxygenic photosynthesis, whereby sulfur is presumably used as an electron donor, may provide a competitive advantage in this environment.
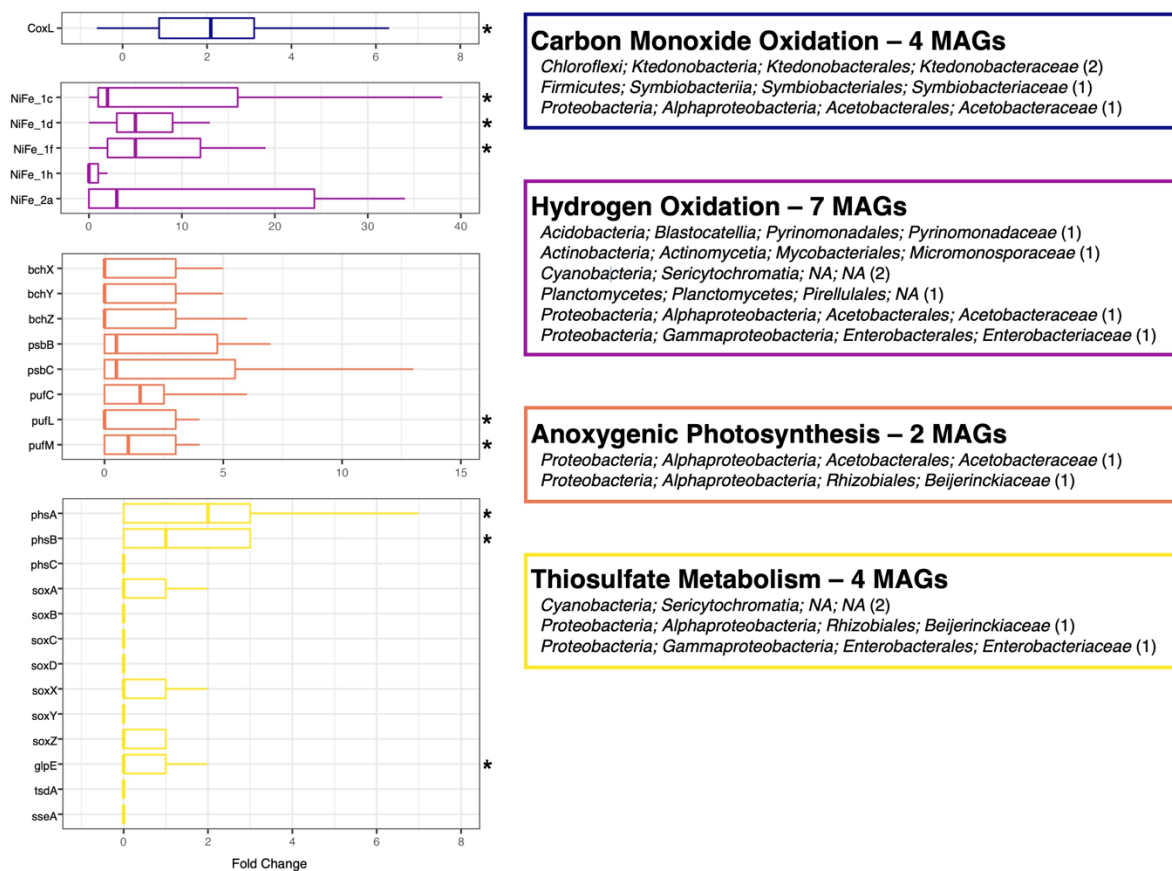

Figure 4.3: Differential abundances of genes between the inland cone sediments (n = 13) and the vegetated sediments (n = 4). The 27 genes are associated with the metabolic pathways of thiosulfate transformations in sulfur metabolism (13 genes), anoxygenic photosynthesis (8 genes), hydrogen oxidation (5 genes), and carbon monoxide oxidation (1 gene). Abundances are presented as the fold change in gene abundances (normalized reads per million) as compared to the average gene abundances in the vegetated samples. Genes with a significantly higher abundance in cone samples are indicated by a star (Mann-Whitney U, p < 0.05). The number of MAGs containing the genes associated with these four metabolic categories are

displayed in the boxes on the right with the taxonomy and number of MAGs within each taxonomic group indicated. If taxonomy could not be resolved below a certain level, it is listed as "NA."

We also found pronounced enrichment of genes associated trace gas metabolism, namely H_2 and CO oxidation, in the cone sediment metagenomes. Many of the MAGs associated with the most abundant taxa recovered from our 16S rRNA gene sequencing effort (including Chloroflexi; Ktedonobacteraceae) do have the genetic capacity for trace gas metabolism (Figure 4.3). More specifically, we identified genes for the catalytic subunits of group 1 [*NiFe*] hydrogenases, group 2 [*NiFe*] hydrogenases, and form 1 carbon monoxide (CO) dehydrogenases (*CoxL*) (King and Weber 2008). Atmospheric H_2 and CO oxidation are common strategies used by bacteria to sustain metabolic activity in resource-limited environments because these compounds are ubiquitous in the troposphere (Ortiz et al. 2021; Dragone et al. 2022). Organisms that use trace gas metabolism to sustain growth have been well described in Antarctic soils and other resource-limited systems (Cockell et al. 2009; Ortiz et al. 2021; Dragone et al. 2022; Bay et al. 2021). While we acknowledge that the presence of genes does not necessarily indicate that these metabolisms are occurring in situ, such metabolic processes have been confirmed in recent volcanic ejecta (Fierer et al. 2010; Ortiz-Álvarez et al. 2018). Likewise, trace gas oxidation has been shown to contribute significantly to microbial metabolism in older volcanic sediments, contributing to 6-10% of total respiratory activity 10 - 20 years after an eruption (King and Weber 2008). Our findings highlight the potential importance of CO and H2 oxidation for microbes inhabiting even more recently deposited sediments. We hypothesize that trace gas oxidation may play a similar role to oxygenic photosynthesis in other primary successional environments, promoting microbial colonization of newly formed landmasses (Schmidt et al. 2008).

Conclusions

Our analyses of the microbes found on the island of HTHH suggest that microbes arrive soon after the formation of new volcanic islands. These microbial colonizers are not oxygenic photosynthesizers, as has been observed in other newly formed or exposed surfaces (like glacial forefields) where cyanobacteria typically dominate the earliest microbial communities (Nemergut et al. 2007; Schmidt et al. 2008). Rather – the microbial colonizers on HTHH are more similar to taxa observed in terrestrial volcanic sediments and largely seem to be relying on trace gas and sulfur-based metabolisms to fuel microbial growth. Likewise, contrary to expectations, the initial colonizers are unlikely to be introduced to this newly formed landmass via dispersal from neighboring marine environments or vegetated soils. Rather, the microbes found on the surface of the young cone sediments appear to be derived from geothermally active environments, potentially including subsurface and/or subterranean sources. Volcanic activity may be seeding the earliest microbial colonizers, dispersing these organisms within and across systems, though it remains undetermined whether these early colonizing communities may have changed over the 3 years that elapsed between the formation of the island and when our samples were initially collected. We expect that if we had been able to continue to track the development of these communities over time, we would see a progression of microbial community change following models generated from other terrestrial volcanic systems, with chemolithotrophs decreasing in abundance as plants establish, organic matter accumulates, and soils develop over time (del Moral and Wood 1993; Deligne et al. 2013).

While we do not have direct field measurements of microbial growth and activity from these sediments, we assume that these taxa are indeed viable and active given that DNA from these organisms was recovered 3 - 4 years after the initial eruption. However, we do not know how long these organisms would persist

and how these communities might continue to change over time as succession progresses. Unfortunately, the near-complete destruction of the island of HTHH in January 2022 makes it impossible to revisit the site and conduct field measurements of metabolic activity on the subaerial cone, as has been done in previous studies (Bay et al. 2021; King and Weber 2008). These results need to be confirmed by performing in situ metabolic measurements in similar volcanic systems sampled soon after eruption, or the next time a surtseyan island emerges and persists.

CHAPTER V

USING SOIL DEPTH GRADIENTS TO DETERMINE THE TAXONOMIC AND
GENOMIC ATTRIBUTES OF OLIGOTROPHIC SOIL BACTERIA

Abstract

Not all bacteria are fast growers. In soil, as in most environments, bacteria exist along a continuum, from those that can grow rapidly under resource-rich conditions (copiotrophs) to those that are adapted to life in the 'slow lane' and can grow even under resource-limited conditions (oligotrophs). However, the field of microbiology is built almost exclusively on the study of copiotrophs due, in part, to the ease of studying them in vitro. Oligotrophic bacteria are expected to be abundant in soil, but they remain more mysterious, with their traits and ecological strategies remaining largely unresolved. To begin understanding the taxonomic and genomic attributes of more oligotrophic bacterial communities, we analyzed 185 soil samples collected from 20 soil profiles across the U.S., using both marker gene and shotgun metagenomic sequencing analyses. We used the soil depth to test hypotheses about the taxa and traits associated with more oligotrophic bacterial communities given that there is a strong decrease in organic carbon availability with depth and thus oligotrophic bacteria should be relatively more abundant in deeper soils compared to the corresponding organic carbon-rich surface soils. We were able to identify specific groups of likely oligotrophic taxa that were consistently more abundant in subsurface soils compared to the corresponding surface soils, including members of the Dormibacteriota and Chloroflexi phyla. In general, soil oligotrophs had smaller genomes, longer growing times, and were underrepresented in culture collections. These bacteria were also enriched in gene pathways that allow them to metabolize a wide range of energy sources and store carbon, while genes associated with energy intensive functions like chemotaxis and

motility were underrepresented. However, we saw a wide range of functional pathways in all subsurface soils, suggesting that oligotrophic taxa use a range of different metabolic strategies to thrive in resource limited environments.

Introduction

In 1991, A. Seminov defined oligotrophic microorganisms as those 'that are evolutionarily adapted to exploit ecological niches characterized by low substrate concentrations and low energy flow' (Semenov 1991). Compared with copiotrophs that can grow rapidly in carbon-rich environments, oligotrophs instead rely on efficient resource use to survive in environments where the supply of organic and inorganic substrates to fuel growth and metabolism are in limited supply (Poindexter 1981; Roller et al. 2016). However, unlike other microbial groups that can be defined more definitively based on their collective traits (e.g. photosynthesizers, obligate anaerobes), oligotrophs are more challenging to define. Rather than representing a discrete category, bacteria can be considered to span a continuous gradient from more copiotrophic to more oligotrophic lifestyles (Merchant and Helmann 2012; Harder and Dijkhuizen 1983) While it is likely there are numerous traits that characterize taxa that exist towards the more oligotrophic side of the spectrum, those traits are not necessarily exclusive to oligotrophs, nor would we expect them to be shared by all oligotrophs. Given the challenge of defining oligotrophy across all environments and microbial taxa, we restrict our discussion here to oligotrophic heterotrophs in soil who typically dominate in most soil environments (Delgado-Baquerizo et al. 2018). We do not focus on soil autotrophs as constraints on their growth are not generally directly linked to resource limitation (Bahram et al. 2018).

We expect to find soil oligotrophs dominating in environments with low concentrations of available organic C (Fierer et al. 2007). For example, we would

expect oligotrophs to be more abundant in bulk soil versus in rhizosphere soils (Liu et al. 2022), more abundant in deeper soils than in surface soils (Brewer et al. 2019), and generally more abundant in surface soils with low plant NPP than in surface soils where plant NPP is high, and inputs of fresh plant-derived organic C would be expected to be higher (Aragão et al. 2009). However, even soils with high concentrations of organic matter could favor oligotrophs if that organic matter is unavailable to fuel microbial metabolism, either due to chemical recalcitrance, physical protection, or other factors that make organic carbon resistant to microbial catabolism (Liang and Balser 2011; Anderson and Coleman 1985; Franzluebbers et al. 2001). Soil environments that we would expect to be dominated by more oligotrophic bacteria are likely common. As one line of evidence, consider that between 35 and 50% microbial biomass contained in soils is located in subsurface horizons that generally have lower levels of available organic carbon compared to more surface soils (Eilers et al. 2012; Brewer et al. 2019). Likewise, even at the scale of individual bacterial cells, most of available surface area in soil is not amenable to habitation - as evidenced by the extremely low percentage (<1%) of the available surface area is estimated to be colonized by microorganisms (Fierer 2017). Thus, environments that favor oligotrophic soil bacteria are likely the norm, not the exception.

We note that the amounts of available organic substrates are not the only factor limiting microbial growth in soil, there are abiotic stressors (e.g. low pH, moisture limitation, anaerobic conditions) and disturbances (e.g. predation, drying-rewetting and freezing-thawing events) that can also limit microbial growth even in soils where substrate concentrations are high (Merino et al. 2019; Fierer 2017). Thus, soils that favor oligotrophic bacteria due to reduced substrate availability can also be environments that might favor bacteria tolerant of other abiotic or biotic stressors or disturbances. To give one example, hyper-arid desert soils in

Antarctica and the Atacama Desert typically have low inputs of plant-derived organic C, but the microbes living in desert soils must also have to tolerate conditions of low moisture, high ultra-violet (UV) exposure, and high soluble salt concentrations (Schmidt et al. 2018; Dragone et al. 2022). Oligotrophic bacteria, by definition, must be able to tolerate environments where organic substrate availability is limited, but they may also have to tolerate other factors that could simultaneously act to constrain growth in such environments.

Genomic information for most soil bacterial taxa, even abundant and ubiquitous taxa, is not currently available (Delgado-Baquerizo et al. 2018). However, we would expect oligotrophs to be particularly under-represented in pre-existing genome databases given that they are likely difficult to cultivate using standard approaches which typically include reasonably high organic C concentrations in media (Poindexter 1981). While oligotrophs can be cultivated, most notably demonstrated through the cultivation of SAR11 from marine waters using extremely dilute media and long incubation periods (Henson et al. 2018; Rappé et al. 2002), doing so is rarely easy nor quick. This under-representation of soil oligotrophic bacteria in pre-existing culture collections has two important ramifications. First, it means that the physiological attributes of oligotrophs have not been as well-studied as those of more copiotrophic taxa which are more amenable to in vitro experimentation (Fierer et al. 2007). Second, it means that pre-existing genomic databases will be biased against soil oligotrophs as most soil bacterial genomes are obtained from the sequencing of cultured isolates. For example, 70% of the bacterial genomes in the Genome Taxonomy Database (GTDB, one of the largest curated genome database), are from isolates (Parks et al. 2022). For these reasons, it has remained difficult to identify the genes, or gene categories, that may be characteristic of soil oligotrophic bacteria and what those genomic attributes could tell us about the physiological adaptations of oligotrophic bacteria.

What are the expected life history characteristics of soil bacterial oligotrophs? Oligotrophic bacteria should be able to survive and grow under conditions where metabolizable organic substrates are infrequently supplied and/or supplied at consistently low concentrations (Semenov 1991; Poindexter 1981). We would expect that bacteria able to thrive under such soil conditions might share some ecological attributes. Those phenotypic traits that have long been thought to be characteristic of oligotrophic soil bacteria include (but are not limited to): long generation times, low maximal specific growth rates (umax), low maintenance energy requirements, high substrate uptake affinities, ability to accumulate intracellular storage polymers, smaller cell sizes (high surface area: volume ratios) and higher density of transport sites per unit cell surface area (and/or low specificity transporters) (Poindexter 1981; Semenov 1991; Lauro et al. 2009; Vieira-Silva and Rocha 2010; Roller et al. 2016; Noell and Giovannoni 2019). Most of these hypotheses regarding oligotrophy-associated traits are supported by limited evidence and there is considerable uncertainty regarding the validity of these hypotheses. For example, Noell and Giovannoni (2019) proposed that small genome size is associated with oligotrophs, while Vieira-Silva and Rocha (2010) have argued otherwise. A more detailed list of 18 hypothesized genes, gene categories, or other genomic features that might be associated with more oligotrophic soil bacteria, based on pre-existing work focused on soil bacteria and on heterotrophic bacteria found in environments (including aquatic environments), can be found in Table 1.

Here, we analyzed soil samples collected from 20 soil profiles representing distinct soil and ecosystem types (Brewer et al. 2019) to make predictions about which taxa tend to be consistently more abundant in deeper soils versus surface soils and about which genes tend to be more abundant in the deeper soils versus surface soils. For these soil profiles, we used depth as a proxy for carbon availability because most fresh carbon inputs are derived from plant litter and root exudates

with soil microbial biomass declining sharply with depth (Richter and Markewitz 1995; de Graaff et al. 2014; Trumbore 2000; Ajwa et al. 1998). Using a combination of marker gene amplicon sequencing, genomic analyses of representative genomes from taxa determined to be more copiotrophic versus more oligotrophic, and shotgun metagenomic sequencing. We used these three datasets to compare the taxonomic composition and genomic attributes of bacterial communities associated with the relatively organic carbon- rich surface soils (0 – 10 cm depth) to those associated with the more carbon-limited soils (>10 cm depth). Specifically, we used these sequence datasets to identify putatively oligotrophic bacterial taxa and to test the hypotheses outlined in Table 5.1, i.e., the genomic attributes hypothesized to be associated with oligotrophic bacterial heterotrophs.

Results and Discussion

Changes in organic carbon and microbial biomass with depth

As reported previously, we found that soil depth in our samples represents a gradient in organic C availability with conditions becoming more resource-limited in deeper soil depths (on average). Total organic C concentrations were significantly higher in surface soils than in subsurface soils (Supplemental Figure S5.1), likely because most fresh carbon inputs are coming from plant litters and root exudates. We note that while there was variation in the total organic C found in the subsurface soils, previous work has shown that organic C that remains at depth tends to be quite old and is likely unavailable to microbes (Ajwa et al. 1998; Trumbore 2000). We recognize that there are other abiotic stressors that vary with depth in soils, including but not limited to availability of other nutrients (including N, P), moisture, and temperature, but previous work with these samples has shown that other soil variables (including pH) exhibit minimal consistent changes with soil depth (Brewer et al. 2019; Dove et al. 2020). Notably, Dove et al. (2020) also found

that both extracellular enzyme activity and microbial biomass decrease with depth, further supporting the hypothesis that soil profiles represent a gradient in organic C availability, with deeper soils being more organic carbon limited and more likely to harbor oligotrophic bacteria as compared to surface soils.
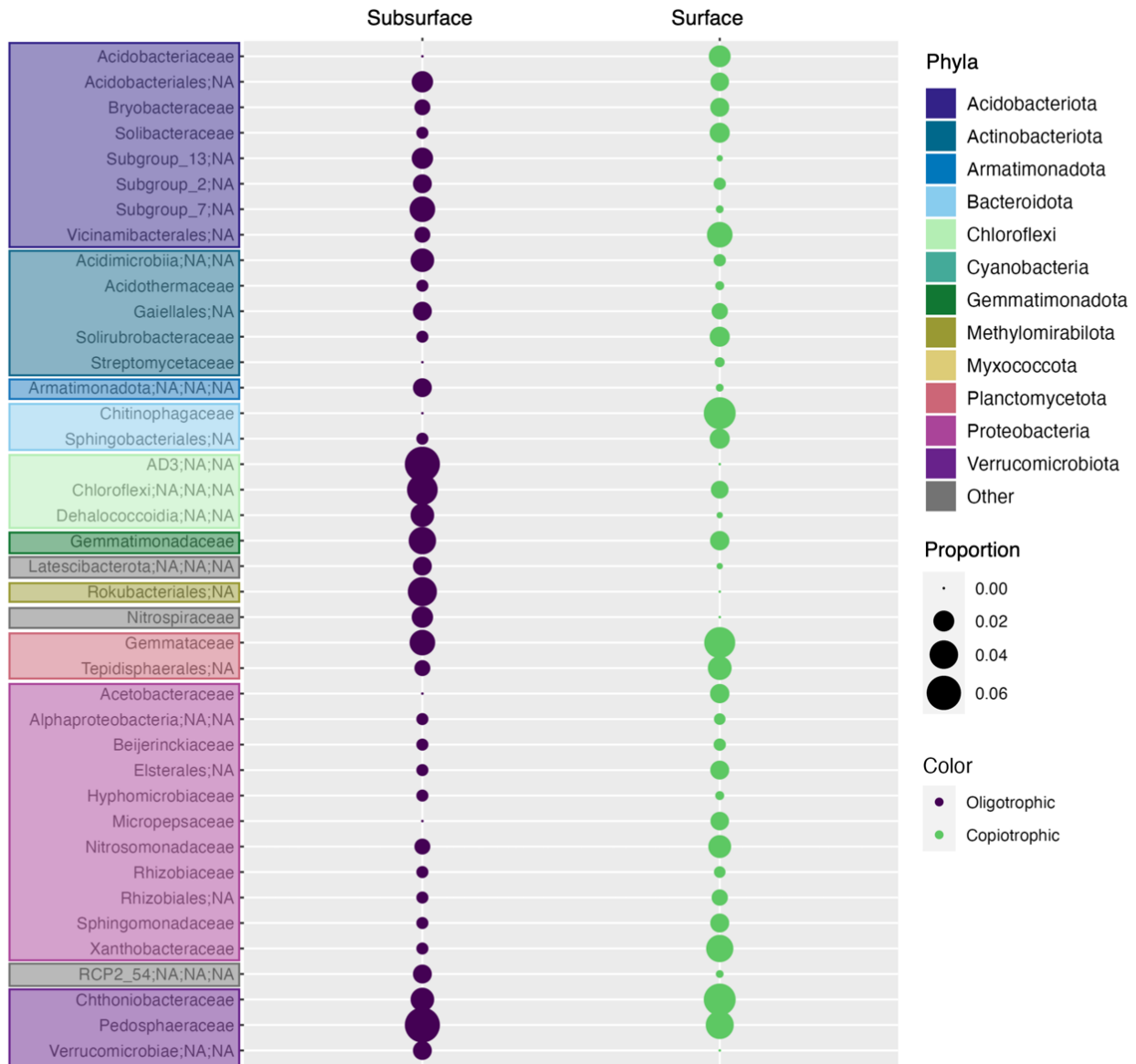

Figure 5.1: Proportion of the surface (n=1271) and subsurface (n=178) ASVs assigned to specific families. Families were only included if they made up at least 1.5% of the relative abundance of the surface or subsurface group. The size of the bubbles indicates the proportion of the total number of ASVs assigned to that taxonomic group.

Taxa that consistently increase or decrease with soil depth

By comparing the abundances of the 12075 ASVs recovered from the 16S rRNA gene sequencing across all 20 soil profiles, we identified 178 bacterial ASVs that were consistently more abundant in deeper soils and 1271 ASVs that were consistently more abundant in surface soils. The phyla that made up the largest proportion of the 178 taxa found to be consistently more abundant in deeper soils were Acidobacteriota (16.2% of oligotrophic ASVs), Chloroflexi (15.2%), Verrucomicrobiota (10.4%), Proteobacteria (9.4%), Actinobacteriota (9.4%), and Planctomycetota (6.2) (Figure S2). In contrast, the most abundant surface-associated phyla were Proteobacteria (27% of surface- associated ASVs), Acidobacteria (15.4%), Verrucomicrobiota (11.5%), Planctomycetota (11.4%), Bacteroidota (10.2%), and Actinobacteria (10.1%) (Supplemental Figure S5.2). At more specific taxonomic levels, there were more distinct families associated with the surface than the subsurface (159 and 82 families, respectfully), perhaps not surprising given we found nearly 7 times more surface-associated taxa than subsurface-associated taxa. In general, we found more taxa in the subsurface that could not be classified past the phylum or class levels of taxonomic resolution (Figure 5.1). The families that were most abundant out of the taxa consistently more abundant in deeper soils include Pedosphaeraceae, Gemmatimonadaceae, and Gemmataceae. The most represented families in the surface soils include Chthoniobacteraceae, Chitinophagaceae, Gemmataceae, Pedosphaeraceae, Xanthobacteraceae, Nitrosomonadaceae, Acidobacteriaceae, and Solibacteraceae.

Table 5.1: Genomic characteristics **A),** functional categories **B)**, and genes **C)** that have been hypothesized as more or less abundant in oligotrophs

A. Genomic Characteristic	Hypothesis	Reference
Delta ENC	codon usage bias in highly expressed genes should be lower for oligotrophs	Vieira-Silva and Rocha. 2010.
Estimated rRNA operon copy #	Oligotrophs, with lower maximum potential growth rates, should have fewer rRNA operons	Vieira-Silva and Rocha (2010), Roller et al. (2016).
Genome Size	Oligotrophs have smaller genomes	Lauro et al. (2009) Giovanni et al. (2014)
B. Functional Categories		
Amino acid transport and metabolism	Oligotrophs should have more genes associated with amino acid transport and metabolism to facilitate an enhanced utilization of proteinaceous substrates	Qin et al. mBio. (2019)
Chemotaxis and motility	sensing and moving is an energetically expensive foraging strategy and should be less abundant in oligotrophs	Roller et al. (2016), Lauro et al. (2009)
Lipid transport and metabolism	Oligotrophs are expected to be enriched in lipid transport and metabolism genes and may be using them as stored sources of C.	Lauro et al. (2009)
Secondary metabolite biosynthesis, transport, metabolism	Oligotrophs may have more genes associated with secondary metabolite metabolism	Lauro et al. (2009)
Defense mechanisms	Oligotrophs should have fewer genes allocating energy to defense	Dutta and Paul (2012)
Transcription	Oligotrophs should have fewer genes allocated to transcription	Dutta and Paul (2012)
Signal transduction	Oligotrophs should have fewer genes allocated to signal transduction	Dutta and Paul (2012)
Cellular replication, recombination, repair	Oligotrophs should have fewer genes and allocate less energy to with cellular replication, recombination, repair	Koch (2001)
C. Specific genes		
Glycine betaine ABC transporter (ProX)	glycine betaine ABC transporters are more abundant in oligotrophs	Noell and Giovannoni (2019)
RNA polymerase, extracytoplasmic E (rpoE)	transcription factor involved in environmental stress response. Should be more abundant in oligotrophs	Su et al. (2015)
Trehalose synthase and transporter	universal stress molecule and osmolyte that stabilizes proteins, associated with increased life span will be more abundant in oligotrophs	Bird et al. (2019)
Form 1 CO dehydrogenases (coxL)	Consumption of CO, even at low concentrations is beneficial for oligotrophs and genes associated with this metabolic pathway will be more abundant	Cordero et al. (2019)
[NiFe] hydrogenases	Genes involved in H2 metabolism, which can serve as an energy source in challenging environments, should be more abundant in oligotrophs	Greening et al. (2015), (2016),
Thiamine biosynthesis	Genes related to thiamine biosynthesis should be more abundant in oligotrophs	Roller et al. (2016)
Poly-B-hydroxybutyrate, polyhydroxyalkanoate	Oligotrophs should have more genes associated with poly-B-hydroxybutyrate, and polyhydroxyalkanoate synthesis, which can carbon and P to cope with periods of starvation.	Poindexter (1981)

We found that 30.7% of the taxa we identified can be found in both the surface and the subsurface (62 of the 202 families identified by this analysis included ASVs assigned to both the surface- and subsurface- associated groups). Unlike other ecological traits that are predictable from taxonomy alone, even at broad taxonomic levels (ex: oxygenic photosynthesis in Cyanobacteria, (Fischer et al. 2016)), bacteria span a continuous gradient from more copiotroph to more oligotrophic lifestyles (Merchant and Helmann 2012; Harder and Dijkhuizen 1983). It is not unexpected that many families could have both oligotrophic and copiotrophic representatives. However, there were some taxa with representatives more likely to be either oligotrophic or copiotrophic. For example, we found that the phylum Chloroflexi made up a much greater percentage of the subsurface-associated taxa compared to the surface (15% in the subsurface to 2% in the surface), while Bacteriodota made up a greater percent of the surface associated taxa (1% in the subsurface, 10% in the surface) (Supplemental Figure S5.2). Perhaps the most surprising pattern was the wide diversity of bacteria we found to be associated with the subsurface environment. In total 22 distinct phyla were represented in the group of 178 ASVs identified as being significantly more abundant in the carbon limited subsurface.

In general, the taxa we identified as more copiotrophic are organisms common to global topsoils (Delgado-Baquerizo et al. 2018). For example, Chitinophagaceae, the most abundant taxa at the surface, is a family that is often associated with rhizosphere communities (Madhaiyan et al. 2015). Members of the Chitinophagaceae were not found in the list of subsurface- associated taxa. Other taxa that were not found in the subsurface include Acidobacteriaceae, Streptomycetaceae, Chitinophagaceae, Acetobacteraceae, and Micropepsaceae. Many of these, including Streptomycetaceae and Acetobacteraceae are common soil bacteria (Reis and Teixeira 2015; Delgado-Baquerizo et al. 2018). On the other

hand, we found that the putatively oligotrophic taxa include groups that have been described to have functional adaptations to survive in challenging, low carbon environments. Dormibacteraeota, the most abundant taxa at depth, have been shown to survive in carbon limited environments, potentially via metabolic adaptations that allow them to use CO as a supplemental energy source, store and synthesize a diverse variety of carbohydrates, and potentially form spores (Brewer et al. 2019; Lennon 2020). Along with Dormibacteraeota, other taxa that were only present in the subsurface-associated group include Rokubacteriales, Nitrospiraceae, and Verrucomicrobiae.

Genomic characteristics of oligotrophic soil microbes

By matching the sequences of the subsurface and surface-associated ASVs described above and in Figure 1 against the Genome Taxonomy Database (GTDB) (Parks et al. 2018; Rinke et al. 2021), we compiled a dataset of 343 genomes to test hypotheses about the genomic characteristics of oligotrophic soil organisms (Supplemental Figure S5.3). 303 of these genomes matched to the surface-associated ASVs and were considered to be genomes representative of more-copiotrophic bacteria. Only 40 genomes matched to subsurface-associated ASVS that we considered to be more oligotrophic. The majority of the oligotrophic reference genomes were metagenome-assembled genomes (MAGs, 70.1%) while the majority of the copiotrophic genomes were derived from cultivated isolates (61.8%) (Figure 5.2). Given the challenges of cultivating oligotrophic taxa with standard techniques (Poindexter 1981), we were not surprised to find that the genomes of oligotrophs were more often generated through cultivation- independent methods than the genomes of copiotrophs.
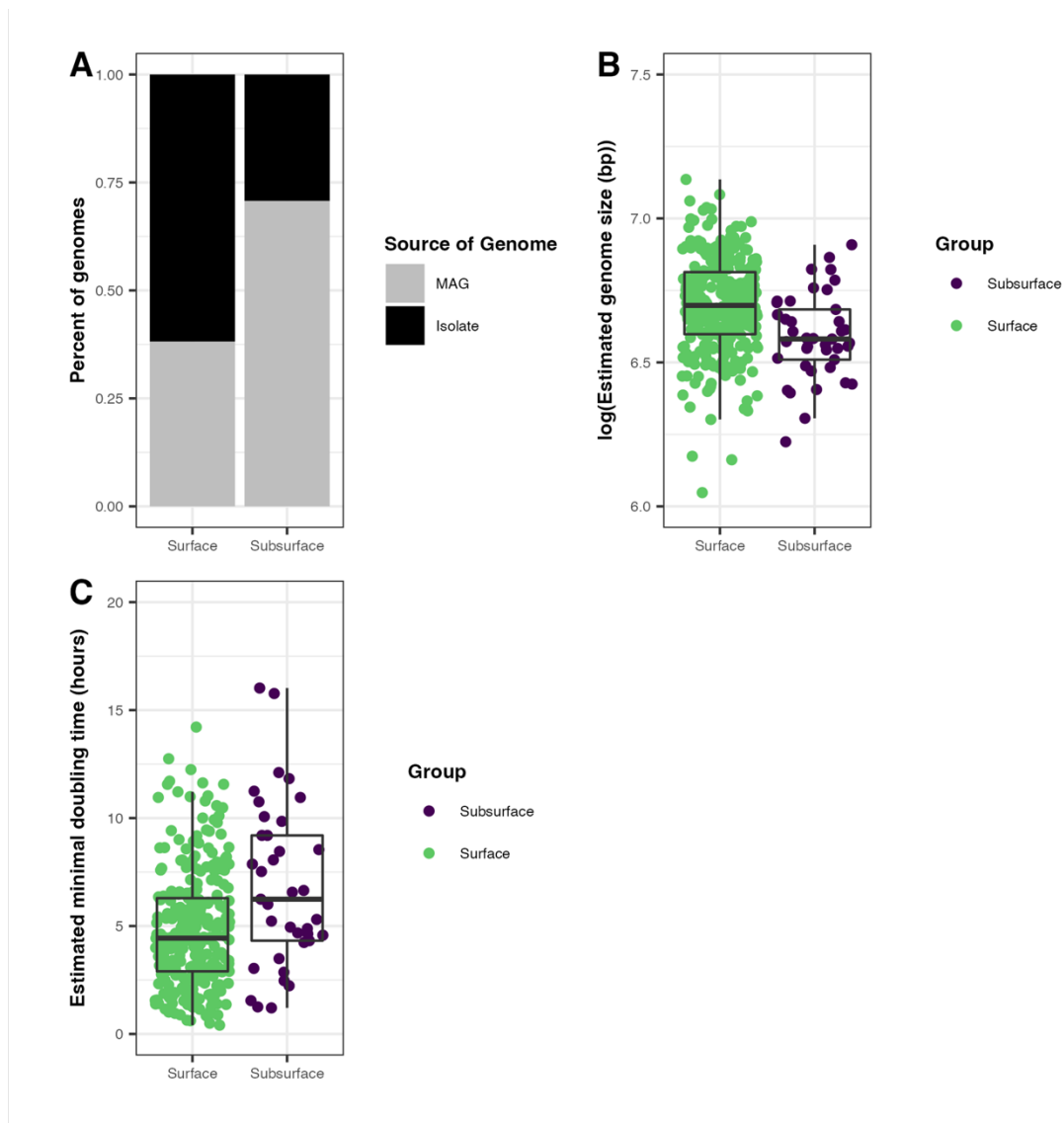
Figure 5.2: Genomic characteristics of the surface (n=303) and the subsurface (n=40) genomes. **A)** The majority of the surface genomes were generated from isolates (61.8%) compared to subsurface genomes which were mainly from MAGs (70.1%). **B)** Estimated genome size was significantly smaller in the subsurface genomes (Mann-Whitney U, p < 0.001). **C)** Estimated minimum doubling time, as calculated with grodon2 (Weissman et al. 2021) was significantly shorter in surface genomes (Mann-Whitney U, p < 0.001).

We found no significant difference in GC content between the copiotrophic and oligotrophic genomes (Supplemental Figure S5.4). We did find that estimated minimal doubling time was significantly longer in oligotrophic genomes than in copiotrophic genomes (Figure 5.2), which supports our hypothesis that oligotrophic

bacteria should have lower maximum potential growth rates to cope with carbon limitation (Table 5.1) (Weissman et al. 2021). Estimated genome size was significantly smaller (on average 1.2 times smaller) in oligotrophic genomes (Figure 5.2). A "streamlined" or reduced genome size has been hypothesized as one strategy used by oligotrophic bacteria, like the marine SAR11 clade, to reduce cellular metabolic costs (Swan et al. 2013; Giovannoni 2017; Henson et al. 2018). However, we note that genome reduction has also been linked to mutualist microbial taxa (Scherlach and Hertweck 2018), pathogens (Moran 2002), and challenging environmental conditions, including in response to extreme heat or cold (Merino et al. 2019) and small genomes may be associated with many ecological strategies besides oligotrophy.

Functional attributes of oligotrophs in soil

To test the hypotheses about the functional attributes of oligotrophs (see Table 5.1), we first compared the proportion of genes associated with the 25 COG categories between the copiotrophic genomes and the oligotrophic genomes. We found that the categories that were enriched in the oligotrophic genomes supported several of the hypotheses outlined in Table 5.1 (Figure 5.3). For example, the oligotrophic genomes had more genes associated with amino acid transport and metabolism. These genes facilitate an enhanced utilization of proteinaceous substrates, which is a metabolic process found to be beneficial in nutrient limited environments (Qin et al. 2013). We also found that oligotrophic genomes had a wider range of metabolic pathways related to lipid transport and metabolism, which are functions that have been linked to bacterial storage of C for later use (Lauro et al. 2009; Wang et al. 2022). On the other hand, copiotrophic genomes were enriched in genes associated with energetically expensive activities like chemotaxis and motility and cellular defense (Figure 5.3) (Lauro et al. 2009; Roller et al, 2016;

Dutta and Paul 2012). However, not all our findings supported the hypotheses outlined in Table 5.1. From previous work we would expect that oligotrophs have fewer genes associated with transcription and signal transduction (Dutta and Paul 2012) but these were enriched in the oligotrophic genomes. Similarly, we also found genes associated with secondary metabolite biosynthesis and coenzyme transport and metabolism to be enriched in copiotrophs, refuting the hypotheses outlined by Dutta and Paul (2012) and Lauro et al. (2009). Finally, we found a higher abundance of genes of unknown function in the surface (Figure 5.3). Given that we found higher relative abundance of uncultivated taxa in the subsurface compared to the surface (Supplemental Figure S5.2, Supplemental Figure S5.3), we would have expected the opposite, though we assume this is a result of the greater diversity or organisms represented in the copiotrophic genomes (see Supplemental Figure S5.3).

When we looked at more specific genes and functions that have previously been hypothesized to be associated with oligotrophs (Table 5.1), we found that only thiamine biosynthesis genes were significantly more abundant in oligotrophic genomes (Figure 3B). Thiamine (vitamin B1) is an essential coenzyme for life that helps organisms gain energy and carbon (Averianova et al. 2020). Organisms in carbon limited environments have been shown to be enriched in genes associated with thiamine biosynthesis to maintain growth rates and metabolic function required for survival (Roller et al. 2016). We saw no significant difference in the abundance of genes associated with other metabolic pathways that have been found to be important for bacterial survival in other in resource limited environments, including CO dehydrogenase genes and [*NiFe*] hydrogenase genes involved in trace gas metabolism (Greening and Grinter 2022; Greening et al. 2015; Cordero et al. 2019; Lynch et al. 2012) were not significantly enriched in the oligotrophic genomes. This might be due to non-penetration of gasses deep in soil. Nor did we find genes associated with the cellular storage of carbon to be enriched in the oligotrophic

genomes, including genes linked to the formation of poly-B-hydroxybutyrates and polyhydroxyalkanoates which are thought to be produced by certain microbes to cope with periods of starvation (Poindexter 1981; Wang et al. 2022).
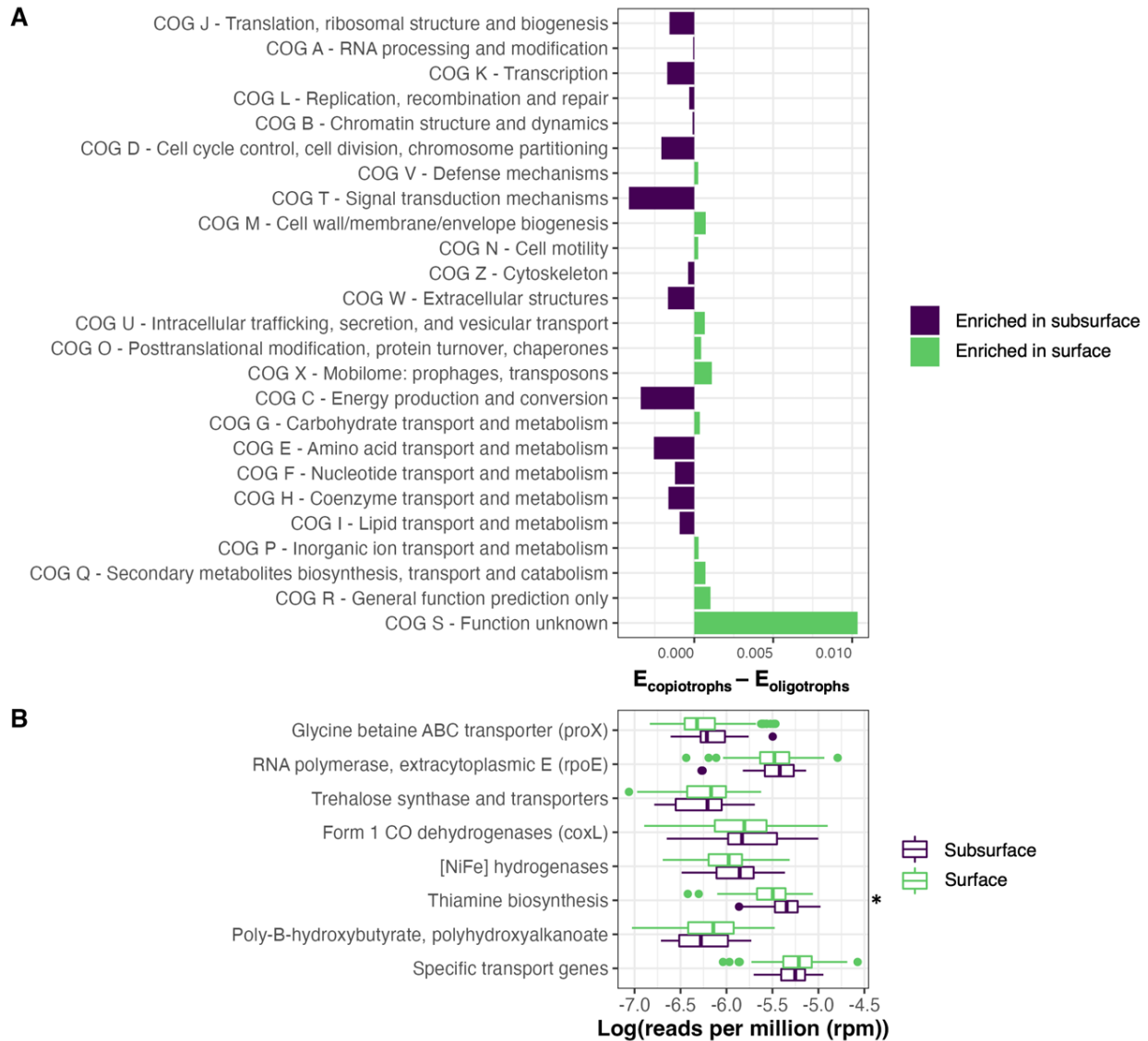

Figure 5.3: Functional comparison of 33 gene categories across the oligotrophic genomes (n=40) and copiotrophic metagenomes (n=303). A. The difference in the average proportion of genes in the surface metagenomes ($E_{surface}$) and the average proportion of genes in the subsurface ($E_{subsurface}$). Bars are colored based on which group has a higher proportion of those genes. B. Abundance of genes associated with the hypotheses outlined in Table 5.1. Significant differences in gene abundances between the two groups (Mann-Whitney U) are starred.

We did not see similar results when we compared the abundance of the same COGs and COG categories between the annotated metagenomes from the subsurface soils and the surface soils. Instead, we found that most of the COG categories were enriched in the copiotrophic bacteria (15 of 25), including lipid transport and metabolism, amino acid metabolism which we found to be enriched in the oligotrophic genomes (Supplemental Figure S5.5). However, we did find that COG category S (genes with unknown function) were more abundant in the subsurface metagenomes, Given that we saw a higher relative abundance of uncultivated taxa in the subsurface compared to the surface (Supplemental Figure S5.2, Supplemental Figure S5.3), this was expected though opposite the pattern we saw in the genomes. We also found that genes associated with form 1 carbon monoxide dehydrogenases (*coxL*) were significantly more abundant in the subsurface metagenomes (Supplemental Figure S5.5). Carbon monoxide oxidation, and other forms of trace gas metabolism, have been found to sustain bacterial energy requirements in challenging environments with limited carbon availability, like in Antarctica and new and old deposits of volcanic sediments (Greening and Grinter 2022; King and Weber 2008; Lynch et al. 2012). In general, we saw greater variation in the abundance of genes in the metagenomes than in the genomes, likely a result of the much more diverse communities represented in the metagenomes than the groups of genomes.

Conclusions

Our comparisons support several of the hypotheses about the functional attributes of oligotrophic organisms highlighted in previous research. We found that, in general, oligotrophs associated with the carbon limited subsurface had smaller genomes, were slower growers, and were generally underrepresented in culture collections. These bacteria are enriched in gene pathways that allow them to

metabolize a wide range of energy sources and store carbon. We also find that genes associated with complicated and energy intensive functions, like chemotaxis, are generally underrepresented in oligotrophic genomes. However, we also found that a wide range of taxa in the subsurface, with taxa assigned to 22 different phyla found to be associated with this carbon limited environment. We expect that this diversity explains the variation in functional gene abundance identified in both oligotrophic genomes and metagenomes. We expect that soil bacteria from distinct taxonomic groups might benefit from different strategies and adaptations to cope with carbon limitation.

CHAPTER VI

CONCLUSIONS

My thesis research has added to our understanding of soil microorganisms in three distinct environments – ice free surface soils in Antarctica, sediments of a newly formed volcanic island, and nutrient limited subsurface soils. In Chapter II I showed that microbial habitability is limited by the cold, dry, salty conditions experienced at inland, higher elevation sites throughout the Transantarctic Mountain. At the most challenging sites, microbial biomass and activity could not be detected using common microbial methods. In Chapter III I showed that the colder, drier, and saltier soils found in higher elevations in Antarctica sustain less diverse communities that were distinct from those in more hospitable soils at low elevations. The harsher conditions found in higher elevation soils likely select for smaller, less diverse communities made up of taxa with lower maximum potential growth rates and an increased reliance on trace gas metabolism to support growth. In Chapter IV I studied sediments from a newly formed surtseyan island in the Kingdom of Tonga. I described a distinct form of microbial primary succession driven by taxa sourced from nearby geothermal systems that generate energy via sulfur metabolism, trace gas metabolism, and anoxygenic photosynthesis. Finally, in Chapter V I showed that oligotrophs associated with the carbon limited subsurface had smaller genomes, were slower growers, and were generally underrepresented in culture collections. These bacteria were also enriched in gene pathways that allow them to metabolize a wide range of energy sources, while genes associated with energy intensive functions were underrepresented. This research has shown that microbial communities in challenging soil environments use a variety of functional adaptations to survive, with the specific functional adaptations dependent on the properties of their surroundings.

REFERENCES

Abarenkov K, Nilsson RH, Larsson KH, Alexander IJ, Eberhardt E, Erland S, Høiland K, *et al.* (2010). The UNITE database for molecular identification of Fungi – recent updates and future perspectives. *New Phytol* 186 (2): 281–285.

Aislabie JM, Chhour KL, Saul DJ, Miyauchi S, Ayton J, Paetzold RF, Balks MR. (2006). Dominant bacteria in soils of Marble Point and Wright Valley, Victoria Land, Antarctica. *Soil Biol Biochem* 38 (10): 3041–3056.

Aislabie JM., Lau A, Dsouza M, Shepherd C, Rhodes P, Turner SJ. (2013). Bacterial Composition of Soils of the Lake Wellman Area, Darwin Mountains, Antarctica. *Extremophiles* 17 (5): 775–786.

Ajwa HA, Rice CW, Sotomayor D. (1998). Carbon and nitrogen mineralization in tallgrass prairie and agricultural soil profiles. *Soil Sci Soc Am J* 62 (4): 942–951.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods* 11 (11): 1144–1146.

Anderson DW, Coleman DC. (1985). The dynamics of organic matter in grassland soils. *J. Soil Water Conserv* 40 (2): 211–216.

Angus P, Guest JCD. (1996). Capelinhos: the disappearing volcano. *Geol. Today* 12 (2): 68–72.

Aragão, LEOC, Malhi Y, Metcalfe DB, Silva-Espejo JE, Jiménez E, Navarrete D, Almeida S, *et al.* (2009). Above- and below-ground net primary productivity across ten Amazonian Forests on contrasting soils. *Biogeosciences* 6 (12): 2759–2778.

Archer SDJ, Lee KC, Caruso T, Maki T, Lee CK, Cary SC, Cowan DA, *et al.* (2019). Airborne microbial transport limitation to isolated Antarctic soil habitats. *Nat Microbiol* 4 (6): 925–932.

Atlas RM. (2004). Handbook of Microbiological Media, 3rd ed. CRC Press, Boca Raton, Fl, USA.

Averianova LA, Balabanova LA, Son OM, Podvolotskaya OB, Tekutyeva LA. (2020). Production of vitamin B2 (riboflavin) by microorganisms: an overview. *Front Bioeng Biotechnol* 8: 570828

Baeza M, Barahona S, Alcaíno J, Cifuentes V. (2017). Amplicon-metagenomic analysis of fungi from Antarctic terrestrial habitats. *Front Microbiol* 8: 2235.

Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PL, Bengtsson-Palme Johan, *et al.* (2018). Structure and function of the global topsoil microbiome. *Nature* 560 (7717): 233–237.

Bastida F, Torres IF, Hernández T, Bombach P, Richnow HH, García C. (2013). Can the labile carbon contribute to carbon immobilization in semiarid soils? Priming effects and microbial community dynamics. *Soil Biol Biochem* 57: 892–902.

Bay SK, Dong X, Bradley JA, Leung PM, Grinter R, Jirapanjawat T, Arndt SK, *et al.* (2021). Trace gas oxidizers are widespread and active members of soil microbial communities. *Nat Microbiol* 6:246-256.

Beilke AJ, Bockheim JG. (2013). Carbon and nitrogen trends in soil chronosequences of the Transantarctic Mountains. *Geoderma* 197:117-125.

Belilla J, Moreira JD, Jardillier B, Reboul G, Benzerara K, López-García JM, Bertolino P, *et al.* (2019). Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol Area. *Nat Ecol Evol* 3 (11): 1552–61.

Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H. (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiol* 10 (1): 189.

Bennett AC, Murugapiran SK, Hamilton TL. (2020). Temperature impacts community structure and function of phototrophic Chloroflexi and Cyanobacteria in two alkaline hot springs in Yellowstone National Park. *Environ Microbio. Rep* 12: 503–513.

Bergsten P, Vannier P, Klonowski AM, Knobloch S, Gudmundsson MT, Jackson MD, Marteinsson VT. (2021). Basalt-hosted microbial communities in the subsurface of the young volcanic island of Surtsey, Iceland. *Front Microbiol* 12: 728977.

Bishop JG, O'Hara NB, Titus JH, Apple JL, Gill RA, Wynn L. (2010). N-P co-limitation of primary production and response of arthropods to N and P in early primary succession on Mount St. Helens volcano. *PLoS One* 5: e13598

Bockheim JG. (1997). Properties and classification of cold desert soils from Antarctica. *Soil Sci Soc Am J* 61 (1): 224–231.

Bockheim JG, Haus NW (2014). Distribution of Organic Carbon in the Soils of Antarctica. In: Hartemink A, McSweeney K (eds). Soil Carbon: Progress in Soil Science. Springer, Cham, CH.

Bottos EM, Scarrow JW, Archer SDJ, McDonald IR, Cary SC. (2014). Bacterial Community Structures of Antarctic Soils. In: Cowan D (eds). Antarctic Terrestrial Microbiology. Springer, Berlin, DE.

Brewer TE, Aronson EL, Arogyaswamy K, Billings SA, Botthoff JK, Campbell AN, Dove NC, *et al.* (2019). Ecological and genomic attributes of novel bacterial taxa that thrive in subsurface soil horizons. *mBio* 10 (5): e01318-19.

Brock TD. (1966). Microbial life on Surtsey. *Surtsey Research Progress Report* 2: 10
Brock TD. (1973). Primary colonization of Surtsey, with special reference to the blue-green algae. *Oikos* 24: 239–243

Brooks ST, Tejedo P, O'Neill TA. (2019). Insights on the environmental impacts associated with visible disturbance of ice-free ground in Antarctica. *Antarctic Science* 31 (6): 304-314.

Brown SP, Jumpponen A. 2015 Phylogenetic diversity analyses reveal disparity between fungal and bacterial communities during microbial primary succession. *Soil Biol Biochem* 89: 52–60.

Bruinsma S, Burgess J, Schlingman D, Czyz A, Morrell N, Ballenger C, Meinholz H, *et al.* (2018). Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics* 19: 722

Bryan WB, Stice GD, Ewart A. (1972) Geology, petrography, and geochemistry of the volcanic islands of Tonga. *J Geophys Res* (1896-1977) 77: 1566–1585.

Buchfink B, Xie C, Huson DH. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12: 59-60.

Cole PD, Guest JE, Duncan AM, Pacheco JM. (2001). Capelinhos 1957–1958, Faial, Azores: deposits formed by an emergent surtseyan eruption. *Bull Volcanol* 63: 204.

Byloos B, Monsieurs P, Mysara M, Leys N, Boon N, Van Houdt R. (2018). Characterization of the bacterial communities on recent Icelandic volcanic deposits of different ages. *BMC Microbiol* 18: 112.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13: 581-583.

Campbell IB, Claridge GG. (1987). Antarctica: soils, weathering processes and environment, 1st ed. Elsevier, New York, NY, USA.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6: 1621-1624.

Capunitan DC, Johnson O, Terrill RS, Hird SM. (2020). Evolutionary signal in the gut microbiomes of 74 bird species from Equatorial Guinea. *Mol Ecol* 29: 829–847.

Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. (2016). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol* 2: 16242.

Carr JL, Horváth Á, Wu DL, Friberg MD. (2022). Stereo plume height and motion retrievals for the record-setting Hunga Tonga-Hunga Ha'apai eruption of 15 January 2022. *Geophys Res Lett* 49: e2022GL098131.

Cary SC, McDonald IR, Barrett JE, Cowan DA. (2010). On the rocks: the microbiology of Antarctic Dry Valley soils. *Nat Rev Microbiol* 8: 129-138.

Chan Y, Van Nostrand JD, Zhou J, Pointing SB, Farrell RL. (2013). Functional ecology of an Antarctic dry valley. *Proc Natl Acad Sci USA* 110: 8990-8995.

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36: 1925–1927.

Clarke KR, Ainsworth M. (1993). A method of linking multivariate community structure to environmental variables. *Mar Ecol Prog Ser* 92: 205-219.

Coates JD, Achenbach LB. (2004). Microbial perchlorate reduction: rocket-fueled metabolism. *Nat Rev Microbiol* 2 (7): 569–580.

Cockell CS. (2014). Types of habitat in the universe. *Int. J. Astrobiology* 13 (2): 158–64.

Cockell CS, Balme M, Bridges JC, Davila A, Schwenzer SP. (2012). Uninhabited Habitats on Mars. *Icarus* 217 (1): 184–93.

Cockell CS, Olsson K, Knowles F, Kelly L, Herrera A, Thorsteinsson T, Marteinsson V (2009). Bacteria in weathered basaltic glass, Iceland. *Geomicrobiol J* 26: 491–507.

Cohen Y, Jorgensen BB, Revsbech NP, Poplawski R. (1986). Adaptation to hydrogen sulfide of oxygenic and anoxygenic photosynthesis among Cyanobacteria. *Appl Environ Microbiol* 51: 398 – 407.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, *et al.* (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42: D633–D642.

Connell JH, Slatyer RO. (1977). Mechanisms of succession in natural communities and their role in community stability and organization. *Am Nat* 111: 1119–1144.

Convey P, Chown SL, Clarke A, Barnes DK, Bokhorst S, Cummings V, Ducklow HW, *et al.* (2014). The spatial structure of Antarctic biodiversity. *Ecol Monogr* 84:203-244.

Coplen TB, Brand WA, Gehre M, Gröning M, Meijer HAJ, Toman B, Verkouteren M. (2006). New guidelines for Δ13C measurements. *Anal Chem* 78 (7): 2439–2441.

Cordero PRF, Bayly K, Leung PM, Huang C, Islam ZF, Schittenhelm RB, King GM, Greening C. (2019). Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *ISME* 13: 2868-2881.

Cowan DA, Ramond JB, Makhalanyane TP, De Maayer P. (2015). Metagenomics of extreme environments. *Curr Opin Microbiol* 25: 97-102.

Cowan DA, Makhalanyane TP, Dennis PG, Hopkins DW. (2014). Microbial ecology and biogeochemistry of continental Antarctic soils. *Front Microbiol* 5: 154.

Cowan DA, Russell NJ, Mamais A, Sheppard DM. (2002). Antarctic Dry Valley mineral soils contain unexpectedly high levels of microbial biomass." *Extremophiles* 6 (5): 431–36.

Crowley SF. (2010). Mineralogical and chemical composition of international carbon and oxygen isotope calibration material NBS 19, and reference materials NBS 18, IAEA-CO-1 and IAEA-CO-8. *Geostand Geoanal Res* 34 (2): 193–206.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008.

Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, *et al.* (2018). A global atlas of the dominant bacteria found in soil. *Science* 359: 320-325.

Deligne NI, Cashman KV, Roering JJ. (2013). After the lava flow: the importance of external soil sources for plant colonization of recent lava flows in the central Oregon Cascades, USA. *Amst* 202: 15–32.

Denton GH, Bockheim JG, Wilson SC, Leide JE, Andersen BG. (1989). Late quaternary ice-surface fluctuations of Beardmore Glacier, Transantarctic Mountains. *Quat Res* 31 (2): 183–209.

Derrien D, Plain C, Courty CP, Gelhaye L, Moerdijk-Poortvliet TCW, Thomas F, Versini A, *et al.* (2014). Does the addition of labile substrate destabilise old soil organic matter? *Soil Biol Biochem* 76: 149–60.

Diaz MA, Gardner CB, Welch SA, Jackson WA, Adams BJ, Wall DH, Hogg ID, Fierer N, Lyons WB. (2021). Geochemical zones and environmental gradients for soils from the Central Transantarctic Mountains, Antarctica. *Biogeosciences* 18: 1629-1644.

Dick GJ. 2019. The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. *Nat Rev Microbiol* 17: 271–283.

Dineen SM, Aranda IVR, Anders DL, Robertson JM. (2010). An evaluation of commercial DNA extraction kits for the isolation of bacterial spore DNA from soil. *J App Microbiol* 109 (6): 1886–1896.

Dixon P. (2003). Vegan, a package of R functions for community ecology. *J Veg Sci* 6: 927-930.

Dove NC, Arogyaswamy K, Billings SA, Botthoff JK, Carey CJ, Cisco C, DeForest JL, *et al.* (2020). Continental-scale patterns of extracellular enzyme activity in the subsoil: an overlooked reservoir of microbial activity. *Environ Res Let* 15 (10): 1040a1.

Dragone NB, Diaz MA, Hogg ID, Lyons WB, Jackson WA, Wall DH, Adams BJ, Fierer N. (2021). Exploring the boundaries of microbial habitability in soils. *J Geophys Res Biogeosci* 126: e2020JG006052

Dragone NB, Henley JB, Holland-Moritz H, Diaz M, Hogg ID, Lyons WB, Wall DH, *et al.* (2022). Elevational constraints on the composition and genomic attributes of microbial communities in Antarctic soils. *mSystems* 7: e01330-21

Dragone NB, Whittaker K, Lord OM, Burke EA, Dufel H, Hite E, Miller F, *et al.* 2023. The early microbial colonizers of a short-lived volcanic island in the Kingdom of Tonga. *mBio* 0 (0): e03313-22.

Dunfield KE, King GM. (2004). Molecular analysis of carbon monoxide-oxidizing bacteria associated with recent Hawaiian volcanic deposits. *Appl Environ Microbiol* 70: 4242–4248.

Dutta C, Paul S. (2012). Microbial lifestyle and genome signatures. *Curr. Genomics* 13 (2): 153–162.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792-1797.

Egan S, Holmström C, Kjelleberg S. (2001). *Pseudoalteromonas ulvae sp. nov.*, a bacterium with antifouling activities isolated from the surface of a marine alga. *Int J Syst Evol Microbiol* 51(4): 1499-1504.

Eilers KG, Debenport S, Anderson S, Fierer N. (2012). Digging deeper to find unique microbial communities: the strong effect of depth on the structure of bacterial and archaeal communities in soil. *Soil Biol and Biochem* 50: 58-65.

Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3: e1319.

Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, *et al.* (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 6: 3–6.

Fattorini S, Borges PAV. (2012). Biogeographical kinetics on an island volcano (Capelinhos, Azores): fast colonization rates and dominance of arthropod exotic species. *Insect Conserv Divers* 5: 358–366.

Fierer N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* 15: 579-590.

Fierer N, Bradford MA, Jackson RB. (2007). Toward an ecological classification of soil bacteria. *Ecology* 88 (6): 1354–1364.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, *et al.* (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci U.S.A* 109 (52): 21390–21395.

Fierer N, Nemergut D, Knight R, Craine JM. (2010). Changes through time: integrating microorganisms into the study of succession. *Res Microbiol* 161: 635–642.

Figueras JB, Cox RP, Højrup P, Permentier HP, Miller M. (2002). Phylogeny of the PscB reaction center protein from green sulfur bacteria. *Photosynth Res* 71: 155–164.

Fischer WW, Hemp J, Johnson JE. (2016). Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci.* 44: 647-683.

Flemming CA, Trevors JT. (1989). Copper toxicity and chemistry in the environment: a review. *Wat Air Soil Pollut* 44: 143–158.

Franzluebbers AJ, Haney RL, Honeycutt CW, Arshad MA, Schomberg HH, Hons FM. (2001). Climatic influences on active fractions of soil organic matter. *Soil Biol Biochem* 33 (7): 1103–1111.

Frederiksen HB, Pedersen A, Christensen S. (2000). Substrate induced respiration and microbial growth in soil during the primary succession on Surtsey, Iceland. *Surtsey Research Progress Report* 11: 29–35.

Fujimura R, Sato Y, Nishizawa T, Nanba K, Oshima K, Hattori M, Kamijo T, Ohta H. (2012). Analysis of early bacterial communities on volcanic deposits on the Island of Miyake (Miyake-jima), Japan: a 6-year study at a fixed site. *Microb Environ* 27: 19–29.

Galperin MY, Makarova KS, Wolf YI, Koonin EV. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 43: D261-D269.

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 Res* 49: D274–D281.

Garber AI, Nealson KH, Okamoto A, McAllister SM, Chan CS, Barco RA, Merino N. (2020). FeGenie: A comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies. *Front Microbiol* 11: 37.

Garvin JB, Slayback DA, Ferrini V, Frawley J, Giguere C, Asrar GR, Andersen K. (2018). Monitoring and modeling the rapid evolution of Earth's newest volcanic island: Hunga Tonga Hunga Ha'apai (Tonga) using high spatial resolution satellite observations. *Geophys Res Lett* 45: 3445–3452.

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, *et al.* (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PloS One* 5: e15545.

Giovannoni SJ. (2017). SAR11 bacteria: the most abundant plankton in the oceans. *Ann Rev Mar Sci* 9 (1): 231–255.

Gomes EC, Figueredo HM, de Oliveira FS, Schaefer CE, Michel RF, Rosa CA, Rosa LH. (2019) Fungi present in soils of Antarctica. In: Rosa LH (ed). Fungi of Antarctica: Diversity, Ecology and Biotechnological Applications. Springer International Publishing, Cham, CH.

Goordial J, Davila A, Lacelle D, Pollard W, Marinova MM, Greer CW, DiRuggiero J, McKay CP, Whyte LG. (2016). Nearing the cold-arid limits of microbial life in permafrost of an upper dry valley, Antarctica. *ISME J* 10: 1613-1624

Göransson H, Venterink HO, Bååth E. (2011). Soil bacterial growth and nutrient limitation along a chronosequence from a glacier forefield. *Soil Biol Biochem* 43: 1333–1340.

Graaff ME, Jastrow JD, Gillette S, Johns A, Wullschleger SD. (2014). Differential priming of soil carbon driven by soil depth and root impacts on carbon availability." *Soil Biol Biochem* 69: 147–156.

Greening C, Carere CR, Rushton-Green R, Harold LK, Hards K, Taylor MC, Morales SE, *et al*. Persistence of the dominant soil phylum Acidobacteria by trace gas scavenging. *Proc Natl Acad Sci U.S.A.* 112 (33): 10497-10502.

Greening C, Grinter R. (2022). Microbial oxidation of atmospheric trace gases. *Nature Rev Microbiol* 12: 1-6.

Gruber-Vodicka HR, Seah BK, Pruesse E. (2020). PhyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. *Msystems* 5: e00920-20. Harder W, Dijkhuizen L. (1983). Physiological responses to nutrient limitation. *Ann Rev Microbiol* 37 (1): 1-23.

Henson MW, Lanclos VC, Faircloth BC, Thrash JC. (2018). Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *ISME J* 12(7): 1846-1860.

Hite EB, Garvin JB, Slayback DA, Burke EA, Callahan G, Joyce P, Whittaker K. (2020). Exploring the human-nature dynamics of Hunga Tonga Hunga Ha'apai, Earth's newest landmass. *J Volcanol Geotherm* 401: 106902.

Ho A, Di Lonardo DP, Bodelier PLE. (2017). Revisiting life strategy concepts in environmental microbial ecology. *FEMS Microbiol Ecol* 93: 3.

Horn DJV, Van Horn ML, Barrett JE, Gooseff MN, Altrichter AE, Geyer KM, Zeglin LH, Takacs-Vesbach CD. (2013). Factors controlling soil microbial biomass and bacterial diversity and community composition in a cold desert ecosystem: role of geographic scale. *PLoS One* 8: e66103.

Hu D, Cha G, Gao B. (2018). A phylogenomic and molecular markers based analysis of the class Acidimicrobiia. Front Microbiol 9: 987.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 11: 119

Hynek BM, Rogers KL, Antunovich M, Avard G, Alvarado GE. (2018). Lack of microbial diversity in an extreme Mars analog setting: Poás Volcano, Costa Rica. *Astrobiology* 18 (7): 923-933.

Iimura Y, Fujimoto M, Hirota M, Tamura K, Higashi T, Yonebayashi K, Fujitake N. (2010). Effects of ecological succession on surface mineral horizons in Japanese volcanic ash soil. *Geoderma* 159: 122–130.

Imhoff JF, Rahn T, Künzel S, Neulinger SC. (2019). Phylogeny of anoxygenic photosynthesis based on sequences of photosynthetic reaction center proteins and a key enzyme in bacteriochlorophyll biosynthesis, the chlorophyllide reductase. *Microorganisms* 7: 576.

Iwasaki I, Ozawa T, Yoshida M, Katsura T, Iwasaki B, Kamada M. (1963). Nature of volcanic gases and volcanic eruption. *Bull Volcanol* 26: 73–81.

Jackson A, Davila AF, Böhlke JK, Sturchio NC, Sevanthi R, Estrada N, Brundrett M, *et al.* (2016). Deposition, accumulation, and alteration of Cl−, NO3−, ClO4− and ClO3− salts in a hyper-arid polar environment: Mass balance and isotopic constraints. *Geochimica et Cosmochimica Acta* 182: 197-215.

Jackson MD, Gudmundsson MT, Weisenberger TB, Rhodes JM, Stefansson A, Kleine BI, Lippert PC, *et al.* (2019). SUSTAIN drilling at Surtsey volcano, Iceland, tracks hydrothermal and microbiological interactions in basalt 50 years after eruption. *Sci Drill* 25: 35–46.

Jackson WA, Böhlke JK, Andraski BJ, Fahlquist L, Bexfield L, Eckardt FD, Gates JB, *et al.* (2015). Global patterns and environmental controls of perchlorate and nitrate co-occurrence in arid and semi-arid environments. *Geochimica et Cosmochimica Acta* 164: 502–522.

Ji M, van Dorst J, Bissett A, Brown MV, Palmer AS, Snape I, Siciliano SD, Ferrari BC. (2016). Microbial diversity at Mitchell Peninsula, Eastern Antarctica: a potential biodiversity "hotspot". *Polar Biol* 39: 237-249.

Ji M, Greening C, Vanwonterghem I, Carere CR, Bay SK, Steen JA, Montgomery K, *et al.* (2017). Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* 552: 400-403.

Jindal S. (2020). Microbes in soil and their metagenomics. In Sharma SG, Sharma NR, Sharma M (eds). Microbial Diversity, Interventions and Scope, Springer, Gateway East, CA, USA.

Kanehisa M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Sci* 28: 1947-1951.

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 49: D545-D551.

Kanehisa M, Goto S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27-30.

Kang, DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7: e7359

Kato S, Shibuya T, Takaki Y, Hirai M, Nunoura T, Suzuki K. (2018). Genome-enabled metabolic reconstruction of dominant chemosynthetic colonizers in deep-sea massive sulfide deposits. *Environ Microbiol* 20: 862–877.

King AJ., Freeman KR, McCormick KF, Lynch RC, Lozupone C, Knight R, Schmidt SK. (2010). Biogeography and habitat modelling of high-alpine bacteria. *Nature Commun* 1 (1): 53.

King GM. (2007). Chemolithotrohic bacteria: distributions, functions and significance in volcanic environments. *Microbes Environ* 22: 309–319.

King GM, Weber CF. (2008). Interactions between bacterial carbon monoxide and hydrogen consumption and plant development on recent volcanic deposits. *ISME J* 2: 195–203.

Klauth P, Wilhelm R, Klumpp R, Poschen P, Groeneweg J. (2004). Enumeration of soil bacteria with the green fluorescent nucleic acid dye sytox green in the presence of soil particles. *J Microbiol Methods* 59 (2): 189–98.

Kokelaar BP. (1983) The mechanism of Surtseyan volcanism. *J Geol Soc London* 140: 939–944.

Krishna K, Govil P. (2007). Soil contamination due to heavy metals from an industrial area of Surat, Gujarat, Western India. *Environ Monit Assess* 124: 263–275.

Lambrechts S, Willems A, Tahon G. (2019). Uncovering the uncultivated majority in Antarctic soils: toward a synergistic approach. *Front Microbiol* 10: 242

Langmead B, Salzberg S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357-359.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106 (37): 15527–15533.

Le Roux PC, McGeoch MA. (2008). Rapid range expansion and community reorganization in response to warming. *Glob Chang Biol* 14 (12): 2950-2962.

Lee CK, Laughlin DC, Bottos EM, Caruso T, Joy K, Barrett JE, Brabyn L, *et al.* (2019). Biotic interactions are an unexpected yet critical control on the complexity of abiotically driven polar ecosystems. *Commun Biol* 2: 62.

Lee KC, Caruso T, Archer SD, Gillman LN, Lau MC, Cary SC, Lee CK, Pointing SB. (2018). Stochastic and deterministic effects of a moisture gradient on soil microbial communities in the McMurdo Dry Valleys of Antarctica. *Front Microbiol* 9: 2619.

Lennon JT. (2020). Microbial life deep underfoot. *MBio* 11 (1): e03201-19.
Letunic I, Bork P. (2007). Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. *J Bioinform* 23: 127-128.

Leung, PM, Bay SK, Meier DV, Chiri E, Cowan DA, Gillor O, Woebken D, Greening C. (2020). Energetic basis of microbial growth and persistence in desert ecosystems. *mSystems* 5: e00495-19.

Leung PM, Greening C. (2021). Compiled Greening lab metabolic marker gene databases. Monash University. Online resource. https://doi.org/10.26180/14431208.v1

Li D, Liu CM, Luo R, Sadakane K, Lam TW. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 10: 1674-1676.

Liang C, Balser TC. (2011). Microbial production of recalcitrant organic matter in global soils: implications for productivity and climate policy. *Nat Rev Microbiol* 9 (1): 75–75.

Liao Y, Smyth GK, Shi W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923-930.

Liu S, He F, Kuzyakov Y, Xiao H, Hoang DT, Pu S, Razavi BS. (2022). Nutrients in the rhizosphere: A meta-analysis of content, availability, and influencing factors. *Sci Total Environ*. 18: 153908.

Liu Y, Yao T, Kang S, Jiao N, Zeng Y, Huang S, Luo T. (2007). Microbial community structure in major habitats above 6000 m on Mount Everest. *Chin Sci Bull* 52 (17): 2350-2357.

Lynch RC, King AJ, Farías ME, Sowell P, Vitry C, Schmidt SK. (2012). The potential for microbial life in the highest-elevation (> 6000 masl) mineral soils of the Atacama region. *J Geophys Res Biogeosci* 117 (G2).

Lynch RC, Darcy JL, Kane NC, Nemergut DR, Schmidt SK. (2014). Metagenomic evidence for metabolism of trace atmospheric gases by high-elevation desert Actinobacteria. *Front Microbiol* 5: 698.

Madhaiyan M, Poonguzhali S, Senthilkumar M, Pragatheswari D, Lee JS, Lee KC. (2015). *Arachidicoccus rhizosphaerae* gen. nov., sp. nov., a plant-growth-promoting bacterium in the family Chitinophagaceae isolated from rhizosphere soil. *Int J Syst Evol Microbiol* 5 (2): 578-586.

Magalhaes C, Stevens MI, Cary SC, Ball BA, Storey BC, Wall DH, Türk R, Ruprecht U. At limits of life: multidisciplinary insights reveal environmental constraints on biotic diversity in continental Antarctica. *PLOS One* 7 (9*): e44578.*

Magnússon B, Magnússon SH, Ólafsson E, Sigurdsson BD. (2014). Plant colonization, succession and ecosystem development on Surtsey with reference to neighboring islands. *Biogeosciences* 11: 5521–5537.

Manor O, Borenstein E. (2015). MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol* 16: 53.

Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17: 10-12.

Masuda S, Eda S, Ikeda S, Mitsui H, Minamisawa K. (2010). Thiosulfate-dependent chemolithoautotrophic growth of *Bradyrhizobium japonicum*. *Appl Environ Microbiol* 76: 2402–2409.

Merchant SS, Helmann JD. (2012) Elemental economy: microbial strategies for optimizing growth in the face of nutrient limitation. *Adv Microb Physiol* 60: 91-210.

Merino N, Aronson HS, Bojanova DP, Feyhl-Buska J, Wong ML, Zhang S, Giovannelli D. (2019). Living at the extremes: extremophiles and the limits of life in a planetary context. *Front Microbiol* 10: 780.

Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, Harrison T, Paczian T, Trimble WL, Wilke A. (2019). MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief Bioinformatics* 20 (4): 1151-1159.

Monchamp ME, Spaak P, Pomati F. (2019). Long term diversity and distribution of non-photosynthetic Cyanobacteria in peri-alpine lakes. *Front Microbiol* 9: 3344.

Moral R, Wood DM. (1993). Early primary succession on the volcano Mount St. Helens. *J Veg Sci* 4: 223–234.

Moran NA. (2002). Microbial minimalism: genome reduction in bacterial pathogens." *Cell* 108 (5): 583–86.

Morozkina EV, Slutskaya ES, Fedorova TV, Tugay TI, Golubeva LI, Koroleva OV. (2010). Extremophilic microorganisms: biochemical adaptation and biotechnological application. *App Biochem Microbiol* 6 (1): 1-4.

Nakagawa S, Takai K. (2008). Deep-sea vent chemoautotrophs: diversity, biochemistry and ecological significance. *FEMS Microbiol Ecol* 65: 1–14.

Nemergut DR, Anderson SP, Cleveland CC, Martin AP, Miller AE, Seimon A, Schmidt SK. (2007). Microbial community succession in an unvegetated, recently deglaciated soil. *Microb Ecol* 53: 110–122.

Nicol GW, Tscherko D, Embley TM, Prosser JI. (2005). Primary succession of soil Crenarchaeota across a receding glacier foreland. *Environ Microbiol* 7: 337–347.

Nilsson RH, Larsson KH, Taylor AF, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, *et al.* (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 47 (D1): D259-D264.

Noell SE, Giovannoni SJ. (2019). SAR11 bacteria have a high affinity and multifunctional glycine betaine transporter. *Environ Microbiol* 21 (7): 2559–2575.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. (1993). metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27: 824–834.

Olff H, Huisman J, Van Tooren BF. (1993). Species dynamics and nutrient accumulation during early primary succession in coastal sand dunes. *J Ecol* 81: 693–706.

Oosting HE, Anderson LE. (1939). Plant succession on granite rock in eastern North Carolina. *Bot Gaz* 100: 750–768.

Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ. (2011). Microbial ecology of the dark ocean above, at, and below the seafloor. *Microbiol Mol Biol Rev* 75: 361–422.

Ortiz M, Leung PM, Shelly G, Jirapanjawat T, Nauer PA, Van Goethem MW, Bay SK, *et al.* (2021). Multiple energy sources and metabolic strategies sustain microbial diversity in Antarctic desert soils. *Proc Nat Acad Sci USA* 118: e2025322118.

Ortiz-Álvarez R, Fierer N, de los Ríos A, Casamayor EO, Barberán A. (2018). Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession. *ISME J* 12: 1658–1667.

Pachter L. (2011). Models for transcript quantification from RNA-Seq. arXiv:1104.3889v2

Parks DH, Chuvochina M, rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 50: D785–D794.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36 (10): 996-1004.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25: 1043–1055.

Pathom-Aree W, Stach JE, Ward AC, Horikoshi K, Bull AT, Goodfellow M. (2006). Diversity of actinomycetes isolated from Challenger Deep sediment (10,898 m) from the Mariana Trench. *Extremophiles* 10 (3): 181-189.

Poindexter JS. Oligotrophy. (1981). In Advances in microbial ecology. Springer, Boston, MA, USA.

Power JF, Carere CR, Lee CK, Wakerley GLJ, Evans DW, Button M, White D, *et al.* (2018). Microbial biogeography of 925 geothermal springs in New Zealand. *Nat Commun* 9: 2876.

Pulschen AA, Rodrigues F, Duarte RT, Araujo GG, Santiago IF, Paulino-Lima IG, Rosa CA, Kato MJ *et al.* (2015). UV-resistant yeasts isolated from a high-altitude volcanic area on the Atacama Desert as eukaryotic models for astrobiology. *Microbiologyopen* 4 (4): 574-588.

Qin S, Bian GK, Zhang YJ, Xing K, Cao CL, Liu CH, Dai CC, Li WJ, Jiang JH. (2013). *Modestobacter roseus* sp. nov., an endophytic actinomycete isolated from the coastal halophyte *Salicornia europaea Linn.*, and emended description of the genus *Modestobacter*. *Int J Syst Evol Microbiol* 63 (6): 2197-2202.

Quackenbush J. (2002). Microarray data normalization and transformation. *Nat Genetics* 32: 496-501.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41: D590-D5966.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, AT. https://www.R-project.org/

Ramalho RS, Quartau R, Trenhaile AS, Mitchell NC, Woodroffe CD, Avila SP. (2013). Coastal evolution on volcanic oceanic islands: A complex interplay between volcanism, erosion, sedimentation, sea-level change and biogenic production. *Earth Sci Revs* 127: 140–170.

Rao B, Hatzinger PB, Böhlke JK, Sturchio NC, Andraski BJ, Eckardt FD, Jackson WA. (2010). Natural chlorate in the environment: application of a new IC-ESI/MS/MS method with a Cl18O3-internal standard. *Environ Sci Technol* 44 (22): 8429-8434.

Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 18 (6898): 630-633.

Reasoner DJ, Geldreich E. (1985). A new medium for the enumeration and subculture of bacteria from potable water. *App Environ Microbiol.* 49 (1): 1-7.

Reis VM, Teixeira KR. (2015). Nitrogen fixing bacteria in the family Acetobacteraceae and their role in agriculture. *J Basic Microbiol* 55 (8): 931-949.

Reysenbach, AL, St. John E, Meneghin J, Flores GE, Podar M, Dombrowski N, Spang A, L'Haridon S, *et al.* (2020). Complex subsurface hydrothermal fluid mixing at a submarine arc volcano supports distinct and highly diverse microbial communities. *Proc Natl Acad Sci U.S.A.* 117: 32627–32638.

Richter DD, Markewitz D. (1995). How deep is soil? *BioScience* 45 (9): 600–609.

Rinke C, Chuvochina M, Mussig AJ, Chaumeil PA, Davín AA, Waite DW, Whitman WB, *et al.* (2021). A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* 6 (7): 946-959.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584.

Roller BRK, Stoddard SF, Schmidt TM. (2016). Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat Microbiol* 1 (11): 1–7.

Sahl JW, Schmidt R, Swanner ED, Mandernack KW, Templeton AS, Kieft TL, Smith RL, *et al.* (2008). Subsurface microbial diversity in deep-granitic-fracture water in Colorado. *App Environ Microbiol* 74 (1): 143–152.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, *et al.* (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12 (1): 1-2.

Scherlach K, Hertweck C. (2018). Mediators of mutualistic microbe–microbe interactions. *Nat Prod Rep* 35 (4): 303–308.

Schipper CI, White JDL. (2016). Magma-slurry interaction in Surtseyan eruptions. *Geology* 44: 195–198.

Schmidt SK, Gendron EM, Vincent K, Solon AJ, Sommers P, Schubert ZR, Vimercati L, *et al.* (2018). Life at extreme elevations on Atacama volcanoes: the closest thing to Mars on Earth? *Anton Leeuwen Int J G* 111 (8): 1389-1401.

Schmidt SK, Reed SV, Nemergut DR, Grandy AS, Cleveland CC, Weintraub MN, Hill AW, *et al.* (2008). The earliest stages of ecosystem succession in high-elevation (5000 metres above sea level), recently deglaciated soils. *Proc Royal Soc B* 275: 2793–2802.

Schnell IB, Bohmann K, Gilbert MT. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* 15 (6): 1289-1303.

Schuerger AC, Fajardo-Cavazos P, Clausen CA, Moores JE, Smith PH, Nicholson WL. (2008). Slow degradation of ATP in simulated martian environments suggests long residence times for the biosignature molecule on spacecraft surfaces on Mars. *Icarus* 194 (1): 86-100.

Semenov AM. (1991). Physiological bases of oligotrophy of microorganisms and the concept of microbial community. *Microb Ecol* 22 (1): 239–247.

Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 3: 836–843.

Singh D, Takahashi K, Kim M, Chun J, Adams JM. (2012). A hump-backed trend in bacterial diversity with elevation on Mount Fuji, Japan. *Microb Ecol* 63 (2): 429-437.

Smallhorn-West PF, Garvin JB, Slayback DA, DeCarlo TM, Gordon SE, Fitzgerald SH, Halafihi T, *et al.* (2020). Coral reef annihilation, persistence, and recovery at Earth's youngest volcanic island. *Coral Reefs* 39: 529–536.

Smith JJ, Tow LA, Stafford W, Cary C, Cowan DA. (2006). Bacterial diversity in three different Antarctic cold desert mineral soils. *Microb Ecol* 51 (4): 413-421.

Soo RM, Hemp J, Parks DH, Fischer WW, Hugenholtz P. (2017). On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. *Science* 355: 1436–1440.

Stamatakis A, Ludwig T, Meier H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456-463.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadi K, Salazar G, Djahanschiri B, *et al.* (2015). Structure and function of the global ocean microbiome. *Science* 348: 1261359.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, Luo H, *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U.S.A.* 110 (28): 11463-11468.

Tahon G, Willems A. (2017). Isolation and characterization of aerobic anoxygenic phototrophs from exposed soils from the Sør Rondane Mountains, East Antarctica. *Syst App Microbiol* 40 (6): 357-369.

Tamames J, Puente-Sánchez F. (2019). SqueezeMeta, a fully automatic metagenomic analysis pipeline from reads to bins. *Front Microbiol* 9: 3349.

Thompson DW, Solomon S. (2002). Interpretation of recent Southern Hemisphere climate change. *Science* 296 (5569): 895-599.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, *et al.* (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551 (7681): 457-463.

Thorarinsson S. (1967). The Surtsey eruption and related scientific work. *Polar Record* 13: 571–578.

Trumbore S. (2000). Age of soil organic matter and soil respiration: radiocarbon constraints on belowground C dynamics. Ecol Appl 10 (2): 399-411.

Turner J, Barrand NE, Bracegirdle TJ, Convey P, Hodgson DA, Jarvis M, Jenkins A, *et al.* (2014). Antarctic climate change and the environment: an update. *Polar Record* 50 (3): 237-259.

Ugolini FC, Bockheim JG. (2008) Antarctic soils and soil formation in a changing environment: a review. *Geoderma* 144: 1-8

Van Eaton AR, Harper MA, Wilson CJN. (2013). High-flying diatoms: Widespread dispersal of microorganisms in an explosive volcanic eruption. *Geology* 41: 1187–1190.

Vaughan RG, Webley PW. (2010). Satellite observations of a surtseyan eruption: Hunga Ha'apai, Tonga. *J Volcanol Geotherm* 198: 177–186.

Vieira-Silva S, Rocha EPC. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLOS Genetics* 6 (1): e1000808.

Vimercati L, Darcy JL, Schmidt SK. (2019). The disappearing periglacial ecosystem atop Mt. Kilimanjaro supports both cosmopolitan and endemic microbial communities. *Sci Rep* 9 (1): 1-4.

Vitousek PM, Walker LR, Whiteaker LD, Matson PL. (1993). Nutrient limitations to plant growth during primary succession in Hawaii Volcanoes National Park. *Biogeochemistry* 23: 197–215.

Vukojević V, Durdic S, Stefanovic V, Trifkovic J, Cakmak D, Perovic V, Mutic J. (2019). Scandium, yttrium, and lanthanide contents in soil from Serbia and their

accumulation in the mushroom *Macrolepiota procera* (Scop.) Singer. *Environ Sci Pollut Res* 26: 5422–5434.

Wagner GP, Kin K, Lynch VJ. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 131 (4): 281-285.

Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, Gilbert JA, *et al.* (2016). Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *Msystems* 1 (1): e00009-15.

Wang JL, Dragone NB, Avard G, Hynek BM. (2022). Microbial survival in an extreme martian analog ecosystem: Poás Volcano, Costa Rica. *Front Astron Space Sci* 9: 817900.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73: 5261-5267.

Warren-Rhodes KA, Lee KC, Archer SD, Cabrol N, Ng-Boyle L, Wettergreen D, Zacny K, *et al.* (2019). Subsurface microbial habitats in an extreme desert Mars-analog environment. *Front Microbiol* 10: 69.

Wei ST, Lacap-Bugler DC, Lau MC, Caruso T, Rao S, de Los Rios A, Archer SK, *et al.* (2016). Taxonomic and functional diversity of soil and hypolithic microbial communities in Miers Valley, McMurdo Dry Valleys, Antarctica. *Front Microbiol* 7: 1642.

Weissman JL, Hou H, Fuhrman JA. (2021). Estimating maximal growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc Natl Acad Sci U.S.A.* 118: e2016810118.

Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, Paczian T, *et al.* (2016). The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res* 44 (D1): D590-D594.

Woolhouse MEJ, Harmsen R, Fahrig L. (1985). On succession in a saxicolous lichen community. *Lichenologist* 17: 167–172.

Wu YW, Simmons BA, Singer SW. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32: 605–607.

Yang J, Teng Y, Wu J, Chen H, Wang G, Song L, Yue W, Zuo R, Zhai Y. (2017). Current status and associated human health risk of vanadium in soil in China. *Chemosphere* 171: 635–643.

Yeo IA, Mcintosh IM, Bryan SE, Tani K, Dunbabin M, Metz D, Collins PC, *et al.* (2022). The 2019–2020 volcanic eruption of Late'iki (Metis Shoal), Tonga. *Sci Rep* 12: 7468.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, *et al.* (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 42: D643- D648.

Yu X, Zhou J, Song W, Xu M, He Q, Peng Y, Tian Y, *et al.* (2021). SCycDB: A curated functional gene database for metagenomic profiling of sulphur cycling pathways. *Mol Ecol Resour* 21: 924–940.

Zeglin LH, Wang B, Waythomas C, Rainey F, Talbot SL. (2016). Organic matter quantity and source affects microbial community structure and function following volcanic eruption on Kasatochi Island, Alaska. *Environ Microbiol* 18: 146–158.

Zhou, Liu Y, Xu Wei, Pan J, Luo ZH, Li M. (2020). Genome- and community-level interaction Insights into carbon utilization and element cycling functions of Hydrothermarchaeota in hydrothermal sediment. *mSystems* 5: e00795-19.

CHAPTER II APPENDIX

EXPLORING THE BOUNDARIES OF MICROBIAL HABITABILITY IN SOILS

Materials and Methods

Sample Collection

Soil samples were collected from the Shackleton Glacier region from December 2017- January 2018. A total of 204 soils were collected from ten different features running the length of the valley. These features represent a range of elevations (150 - 2221 m) across a 120 km north-south distance spanning from the Ross Ice Shelf to the Polar Plateau (Figure 2.1). Between 14 and 26 soil samples were collected along elevational transects located on each of ten features to maximize variation in soil characteristics and soil exposure times (amount of time at the surface and uncovered by glacial ice) at each feature. Soils (0 - 5 cm depth) were collected in sterile polyethylene bags using hand trowels that were sterilized with ethanol before each collection. Approximately 2-5 kg of soil were collected from each location and, after thorough homogenization, a 50-100 g sub-sample was used for downstream microbial analyses (Diaz et al. 2021). GPS coordinates, photographs of the soil surface, elevation, and other metadata were collected at the time of soil sample collection. frozen immediately after collection and were kept frozen at -20°C during transport, with temperature monitored throughout the duration of transport to ensure that the samples never thawed. The soil samples remained frozen until being processed at the University of Colorado in Boulder, Colorado, USA.

DNA extractions

DNA was extracted from all of the collected samples in a laminar flow hood. After mixing 1 g of each soil in 1 mL of sterile PCR-grade water, DNA was extracted from a 500 µl aliquot of the soil slurry using the Qiagen DNeasy® Powersoil® HTP

96 Kit (Qiagen, Germantown, MD, USA) following the manufacturer's recommendations. A total of 6 extraction blanks (2 per 96-well plate) were included to test for any possible contamination introduced during DNA extraction.

Cultivation-independent microbial analyses via marker gene sequencing

The DNA aliquots extracted from each of the 204 soils and the 6 extraction blanks, were PCR-amplified using a primer pair that targets the hypervariable V4 region of the archaeal and bacterial 16S rRNA gene (515F: 5'-GTGCCAGCMGCCGCGGTAA-3' and 806-R: 5′- GGACTACHVGGGTWTCTAAT-3′). These primers are identical to those used by the Earth Microbiome Project and are routinely used to assess both archaeal and bacterial diversity in a wide range of environments (Walters et al. 2016). To assess the fungal communities that may be found in these soils, we also conducted a separate set of PCR amplifications using a primer pair (ITS1-F: 5'-CTTGGTCATTTAGAGGAAGTAA-3' and ITS2-R: 5′-GCTGCGTTCTTCATCGATGC-3′) that target the internal transcribed spacer of the fungal ribosomal RNA (rRNA) operon (Bellemain et al. 2010). Three no template PCR blanks (1 per 96 well plate) were run for each set of amplifications. Both primer sets included the appropriate Illumina adapters and unique 12 - bp barcode sequences to permit multiplexed sequencing (Caporaso et al. 2012). PCR was performed using GoTaq® Hot Start PCR Master Mix (Promega, Madison, WI, USA) in 25 µL reaction volumes. The reaction mixture included 10.5 µL master mix, 12.5 µL sterile PCR grade water, 1µL primers (0.5 each F and R), and 1 µL template DNA. Cycling parameters for both primer sets consisted of an initial denaturation step at 94 °C for 3 min, followed by 35 cycles of denaturation at 94 °C (45 s), annealing at 50 °C (60 s), extension at 70 °C (90 s), and a final extension step at 72 °C for 10 min. Both sets of amplified products were cleaned and normalized to equimolar concentrations using SequalPrep™ Normalization Plates (Thermo Fisher

Scientific, Carlsbad, CA, USA) with the 16S and ITS rRNA gene amplicons sequenced on separate Illumina MiSeq runs (Illumina, San Diego, CA, USA) using the V2 2 x 150 bp and 2 x 250 bp paired-end Illumina sequencing kits (for the 16S rRNA gene and ITS amplicons, respectively). All raw sequences can be found in the NCBI Sequence Read Archive, project accession number PRJNA699250.

16S rRNA gene sequences were processed using the DADA2 pipeline (Callahan et al. 2016). Sequences were quality filtered and clustered into exact sequence variants (ESVs). Taxonomic information was assigned to ESVs using a naive Bayesian classifier method (Q. Wang et al. 2007), which takes the set of ESVs generated and compares them to a training set of reference sequences from the 16S rRNA bacterial and archaeal SILVA database (Quast et al. 2013; Yilmaz et al. 2014). A minimum bootstrapping threshold required to return a taxonomic classification of 50% similarity was used for analysis. ESVs associated with chloroplast, mitochondria, eukaryotes, and those unassigned to the phylum level (477 ESVs) were removed prior to downstream analyses. Extraction blanks yielded an average of 420 reads (0 – 1831 reads) and these reads came from 14 phylotypes from 11 families Rubrobacteriaceae, Sphingomonadaceae, Sulfruspirillaceae, Xanthobacteraceae, Burkholderiaceae, Sporolactobacillus, Clostridiaceae, Bacillaceae, Planococcaceae, Deinococcaceae, Enterobacteriaceae, some of which have been classified as bacterial taxa commonly associated with reagent contamination (Salter et al. 2014). These 14 ESVs found in the extraction blanks represented 100% of the reads from these blanks, while these 14 ESVs only accounted for 0.31% of the reads from all extracted soil samples on average. No-template PCR blanks yielded 311 reads on average (126 - 586) and ESVs associated with these reads were common in other samples (including taxa within the families: Blastocatellaceae, Soilrubrobacteriaceae, Clostridiaceae, Paenibacillaceae, Chitinophagaceae, and Rubrobacteriaceae) and these taxa are not typical reagent

contaminants (Salter et al. 2014), but are instead most likely derived from the soil samples and represent 'tag switching' events (Schnell et al. 2015). For this reason, we used 586 reads per sample as a threshold for determining that we could reliably detect prokaryotic DNA using this DNA sequencing approach. All soils with <586 reads were considered to have no detectable PCR-amplifiable prokaryotic 16S rRNA genes. Across the 153 samples that had a sufficient number of bacterial or archaeal 16S rRNA gene reads for downstream analyses, the mean number of reads per sample was 25912 (629 - 58150) while the mean number of reads in the 51 samples that did not meet the threshold determined from the analysis of the 'blank' samples was only 125 reads per sample.

Fungal ITS sequences were also processed using the DADA2 pipeline (Callahan et al. 2016). Sequences were quality filtered and clustered into exact sequence variants (ESV). Taxonomic information was assigned to ESVs using the same naive Bayesian classifier method described previously (Q. Wang et al. 2007) but with the UNITE database (Nilsson et al. 2019). A minimum bootstrapping threshold required to return a taxonomic classification was set at 85% similarity (Nilsson et al. 2019; Abarenkov et al. 2010). All 6 of the extraction blanks and 3 no template PCR controls had zero reads after processing. We used all ESVs that could be classified to at least a fungal phylum for our analysis. After this filtering, 143 of the 204 samples had no remaining ITS reads. These 143 samples with no identifiable fungal ITS sequences were considered to have no amplifiable fungal DNA. The remaining 61 samples had a mean of 5886 reads per sample and were considered to have a sufficient number of fungal ITS reads for downstream analyses.

Quantitative PCR

We measured 16S rRNA gene copy number and ITS gene copy number using the same primers and DNA extracts used for sequencing. Reaction conditions and details followed methods described by Carini et al. (2016). Each of the 204 soil samples, along with corresponding DNA extraction blank, were run for both gene regions and no-template controls were included for each gene region (16 for 16S rRNA, 18 for ITS). Standard curves were calculated using purified genomic DNA from *Escherichia coli* for 16S rRNA copy number and *Aspergillus fumigatus* for ITS copy number. Ct values from the negative controls were used to identify a limit of DNA detection for each gene region. For 16S rRNA, samples with a Ct value greater than 31 were considered below detection limits. For ITS, samples with a Ct value greater than 35 were considered below detection limits. Calculated copy number measurements for each sample are reported as number of *A. fumigatus* or *E. coli genome* equivalents · g soil^{-1}.

DNA extraction and PCR inhibition test

From the cultivation independent ITS and 16S rRNA gene sequencing analyses described above, we identified ~ 20% of samples (40 out of the 204 individual soils) that had no amplifiable DNA from the targeted bacteria, archaea, or fungal marker genes. To confirm that the PCR amplification of microbial DNA from these 40 samples was not simply a product of DNA extraction problems or PCR inhibition, we performed an inhibition test. For this test, we used two 250 mg subsamples of three soils that did not PCR amplify and one soil that successfully amplified. Following the methods described by Warren-Rhodes et al. (2019), one sub-sample of each of the four soils was extracted after adding a 50 µL suspension of *E. coli* in trypticase soy broth at a concentration of 107 CFU. The other subsample was extracted with no addition of *E. coli*, with the DNA extractions on all samples conducted as described above. Extracted DNA was amplified using the

16S rRNA gene-targeting primers, reagents, and PCR conditions described above. The concentrations of both the extracted gDNA and amplified product from each soil sample replicate were measured with a Molecular Devices SpectraMax M2 microplate reader (Molecular Devices, LLC, San Jose, CA, USA) using an Invitrogen Quant- iT™ Picogreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Carlsbad, CA, USA). Amplified products were also visualized on a 2% agarose gel. No signs of inhibition were detected, i.e. the DNA from all sub-samples amended with *E. coli* cells were successfully PCR- amplified (Supplemental Figure S2.2). Thus, we can conclude that our failure to recover 16S and ITS rRNA gene reads from these 44 samples was not likely a result of PCR inhibition, but rather any microbial DNA in these samples was simply not detectable using our cultivation-independent sequencing approach.

Cultivation-dependent analyses

As an added test of the presence of microbial cells, we attempted to cultivate microbes from a subset of 35 soil samples (Supplemental Table 2.1). These 35 soils were selected to represent the range of environmental gradients encompassed by the larger dataset and included samples from all 10 features within the Shackleton Glacier region, including soils with detectable microbial communities and those with microbial communities below detection limits (as determined by the cultivation-independent sequencing). These 35 soil samples were plated on 17 different types of media (Supplemental Table 2.4) (Atlas 2004; Egan et al. 2015; Pulschen et al. 2015; Reasoner and Geldreich 1985; Singh et al. 2012; Tahon and Willems 2017) to capture a broad diversity of microbial taxa that could be living in these soils. One gram of each soil sample was homogenized with 1 mL of water and 60 µL of the resulting soil slurry was pipetted onto each of the 58 cm^2 plates and spread across the plates using flame sterilized cell spreaders. 'Blank'

plates inoculated with only 60 µL sterile water were prepared for each media type and handled in an identical manner to the 595 plates inoculated with the soil slurries. One 'positive control' soil sample was collected from a lawn on the University of Colorado Boulder's main campus and was plated on all media types using the same techniques. The TSA 24 plates, the PH, and PA plates were all left to incubate at 24°C while the rest of the plates were incubated at 4°C. All plates, besides the PA and PH plates, were left to incubate in the dark. Plated samples, and the uninoculated control plates, remained under these conditions for 8 weeks. Photos of the plates were taken each week. Every distinct colony on each of the 629 petri plates was counted weekly from these photos and microbial cell numbers (CFU · g soil^{-1}) were calculated from these colony counts. The reported total cell numbers are the cumulative number of colonies that grew on each plate over the 8-week incubation period. At the end of the 8-week incubation, those plates that had no visible colonies (307 of 629 plates) were incubated for an additional month. No colonies grew on any of the blank plates over the three-month incubation.

^{13}C glucose metabolism assay

To further verify whether there were, indeed, some soil samples with no detectable microbes and to complement the cultivation-independent and dependent approaches described above, we conducted a ^{13}C - glucose metabolic assay with the subset of 35 soil samples described above. We used ^{13}C-glucose as a substrate because glucose should be readily catabolized by most heterotrophic microbes. Two 1 g replicate sub-samples of each of the 35 soils were placed in sterile glass tubes. Three soil samples (including the one used in the culturing experiment) collected from the University of Colorado Boulder's main campus were prepared in the same way as the Shackleton Glacier samples to serve as 'positive' controls. One of the replicate sub-samples from each of the 38 soil samples, plus 6 soil-free blank tubes,

was autoclaved on a gravity cycle with a 60-minute sterilization and a 30-minute drying cycle so we could assess metabolic activity in paired autoclaved versus un-autoclaved sub-samples of each soil. A 250 µL solution ^{13}C - glucose (99 atom% U-^{13}C, Cambridge Isotope Laboratories, Tewksbury, MA, USA) dissolved in H2O was pipetted directly into each soil, an addition of approximately 250 µg glucose C per gram soil. This amount of glucose added to the soils is comparable to the glucose amendments used in other comparable soil metabolic assays (Bastida et al. 2013; Derrien et al. 2014). A 2 mL cryotube containing 1 mL of 1N sodium hydroxide solution was placed in each tube before all tubes were sealed with an airtight cap. We prepared a total of 76 soil incubations (35 Antarctic soils + 3 Colorado soils with and without soil autoclaving) and 6 blanks (3 tubes with NaOH traps without soil or glucose, and 3 tubes without NaOH traps without soil but with glucose). All tubes were incubated for one month at $4^{o}C$ after which an aliquot of the NaOH trap from each tube was transferred to pre-evacuated glass vials filled with helium gas. The NaOH traps were acidified with 1 mL of concentrated phosphoric acid to release CO_2 into the evacuated headspace. The quantity and isotopic composition of the released $^{13}CO_2$ was determined using a ThermoFisher Scientific Gasbench II (Thermo Fisher Scientific, Waltham, MA, USA) coupled to a Delta V Isotope Ratio Mass Spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). Data from the mass spectrometer was processed using the R package 'isoprocessCUBES'. Sample readings were corrected using the IAEA reference standard NBS18 (Crowley 2010). The delta ^{13}C values were converted to fraction of ^{13}C per mil, expressed as ‰, using the international standards of V-PDB (Vienna Pee Dee Belemnite, Coplen et al. (2006)). Positive glucose mineralization was detected in all measurements, including in control blanks who had an average reading of 1.095 ‰ ^{13}C · g soil^{-1}. Samples were identified as harboring active microbes based on comparisons between the replicate unautoclaved and autoclaved samples and positive controls.

Significance was calculated using Kruskal-Wallis tests as implemented in R with differences between groups determined by a Nemenyi test calculated using the kruskalmc function in the R package 'pgirmess.'

ATP assay

To further confirm the results of the ^{13}C glucose metabolism assay, we conducted an ATP assay with the same subset of 35 soils to test for any bacterial activity. One-gram subsamples of each of the soils were placed in sterile 2ml microcentrifuge tubes. Tubes were homogenized into a soil slurry with 1 mL of PBS. Three 100 µl aliquots of each homogenized soil slurry (35 Antarctic soils + 3 Colorado soils) were pipetted into black, 96-well Costar Assay plates (Corning Inc., Corning, NY, USA) and ATP production was analyzed using the Promega BacTiter-Glo™ Microbial Cell Viability Assay (Promega, Madison, WI, USA) following the manufacturer's instructions. This assay has been used in other studies to detect low levels of bacterial activity on simulated Martian environments (Schuerger et al. 2008). Luminescence readings were performed using a Biotek Synergy™ HT Plate Reader (Biotek Instruments, Inc., Winooski, VT, USA). Six reagent "blanks" (3 wells containing just the sterile PCR-grade water used to make the soil slurries, 3 wells with water and glucose), and 6 blank wells were measured using the same methods. After this "baseline" reading, a 250 µl filter-sterilized solution of glucose dissolved in H_2O was added to each of the soil slurries, an addition of approximately 250 µg glucose C per gram soil. Soils were mixed and were then left to incubate at 4^oC. After a 24- hour incubation the soil slurries were measured again following the previously described methods, with the same reagent blanks and blank wells.

ATP concentrations were calculated from luminescence readings based on a standard curve (R^2=0.99) generated from a 10-fold serial dilution (10^{-10} - 10^{-15} mols) of a Tris-buffered ATP standard (Thermo Fisher Scientific, Hampton, NH, USA).

Based on this standard curve, we determine that we were effectively able to detect ATP concentrations $>10^{-15}$ mol ATP $\cdot$ g soil^{-1}.

Soil Geochemical Analyses

To determine soil water content, 50 grams of soil were put into aluminum weigh dishes and dried at 105°C for 24 hrs. % H_2O was determined by the difference between the dried and initial weights of the soils. Since studies have shown that organic matter can and does degrade at this temperature, the water content analysis was done on a separate aliquot of samples.

Soils were leached at a 1:5, soil to DI water ratio and filtered through 0.45µm nucleopore filters. Anions were analyzed using a Dionex ICS-2100 ion chromatograph and an AS-DV automated sampler. Cations were analyzed by optical emission spectrum on a PerkinElmer Optima 8300 inductively coupled plasma mass spectrometer (ICP-OES) (Diaz et al. 2021). Relative soil exposure ages were estimated using perchlorate concentrations in 1:5 soil to water leaches and annual fluxes from the McMurdo Dry Valleys, Antarctica. Fluxes are estimated to range from 1 to 4 µg m^{-2} yr^{-1} and average 2 µg m^{-2} yr^{-1}. Perchlorate salts are highly soluble in water and the primary source is wet and dry atmospheric deposition (Jackson et al. 2015; Jackson et al. 2016). Therefore, surface soil concentrations can be used to calculate a rough estimate of the amount of time which has passed since the soil was last inundated or wetted.

Statistical Analyses

We performed a random forest analysis to determine if any of the measured environmental and geochemical variables could be used to predict whether a soil would have no detectable microorganisms or detectable microorganisms. The factors used in our models were chosen from a total of 37 different measurements. Highly

correlated variables, edaphic factors, and variables that were not measured on at least 80% of the 204 samples were not included in the analysis. The final variables used for analysis included elevation, the concentration of certain water-soluble salts (NO_3^-, Cl^-, ClO_3^-, ClO_4^-) leached at a 1:5 soil: water ratio, and a sum of total salt, total anions, and total cations calculated from these leached results. Cation and anion concentrations were log transformed before analysis. To determine what environmental variables might best predict which soils would have detectable microorganisms, we used the R package 'rfPermute' and performed a random forest analysis with 100 trees and 3 variables tried at each split to identify the most important predictors. As we were unable to perform all of our habitability tests on our full dataset of 204 samples, the random forest model was based on whether we could detect microbial DNA in each sample using the cultivation-independent marker gene sequencing approach (i.e. whether we could detect bacterial, archaeal, or fungal sequences).

Acknowledgments

Supplemental Figure S2.1: The relative abundances of fungal taxa identified using cultivation independent ITS gene sequencing in all 204 samples collected from across the Shackleton Glacier Valley. Samples are grouped by feature and organized from highest elevation site to lowest elevation site at each feature (left to right), with the higher elevation sites being further south. Grey bars represent samples with no amplifiable fungal DNA.

Sample	gDNA (ng/μl)	Amplified Product (ng/μl)	Well
SH1-5	0.00	11.13	B
SH1-5 + *E.coli*	5.29	426.99	C
SH1-6	0.00	15.91	D
SH1-6 + *E.coli*	8.49	407.58	E
RM3-7	0.00	14.50	F
RM3-7 + *E.coli*	5.01	381.96	G
TGV3-9	2.91	260.35	H
TGV3-9 + *E.coli*	17.28	370.79	I
E.coli	14.26	528.26	J
Blank (Extraction)	0.00	13.95	K
Blank (PCR)	n/a	15.95	L
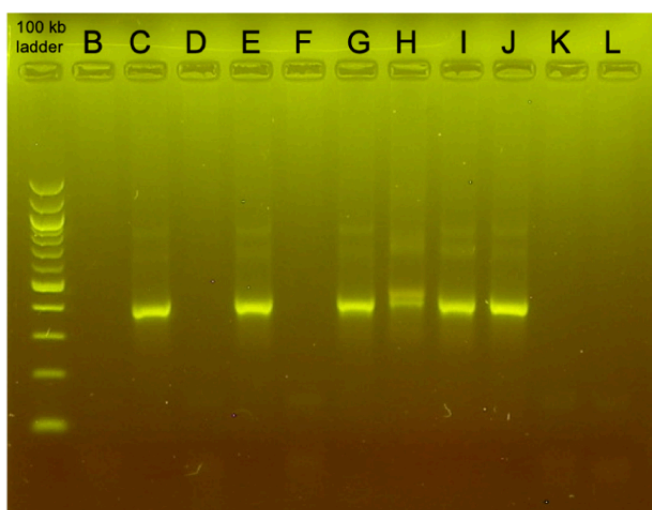
Supplemental Figure S2.2: The results of the PCR inhibition test. The PCR inhibition test used three samples that were below detectable limits across all tests (SH1-5, SH1-6, RM3-7) and one sample (TGV 3-9) that contained identifiable DNA based on the 16S rRNA sequencing. When the soils were spiked with E. coli cells, there was no evidence that DNA extractions or PCR amplifications were inhibited and thus the failure to detect amplifiable bacterial DNA in these 3 soils is due to the lack of sufficient bacterial DNA in those samples and not methodological issue.
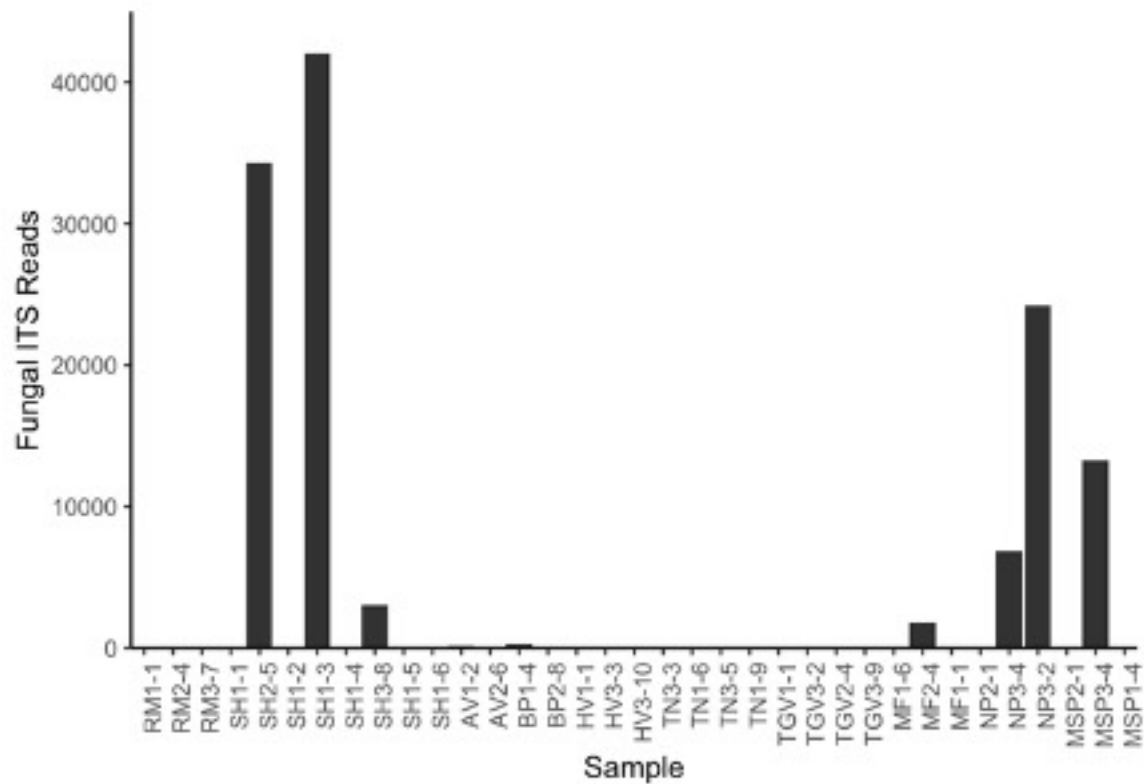
Supplemental Figure S2.3: The number of fungal ITS reads that were classified at least to a fungal phylum for the same subset of 35 samples analyzed using the cultivation-dependent, metabolic, and ATP assays (see Figure 2.1).
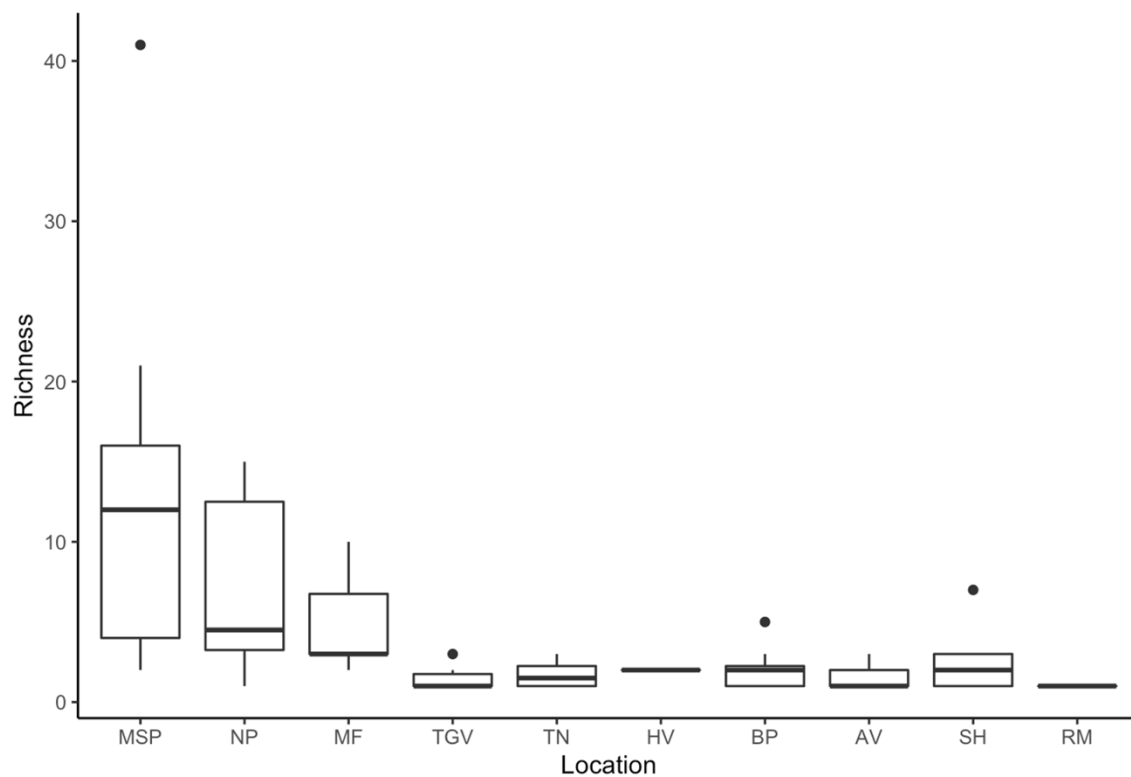
Supplemental Figure S2.4: The richness of fungal communities across the 10 features of the Shackleton Glacier Valley. Fungal richness, number of unique fungal phylotypes per sample, was determined based on the cultivation independent ITS gene sequencing results. Samples were included if they had at least one read identifiable to a fungal phylum.
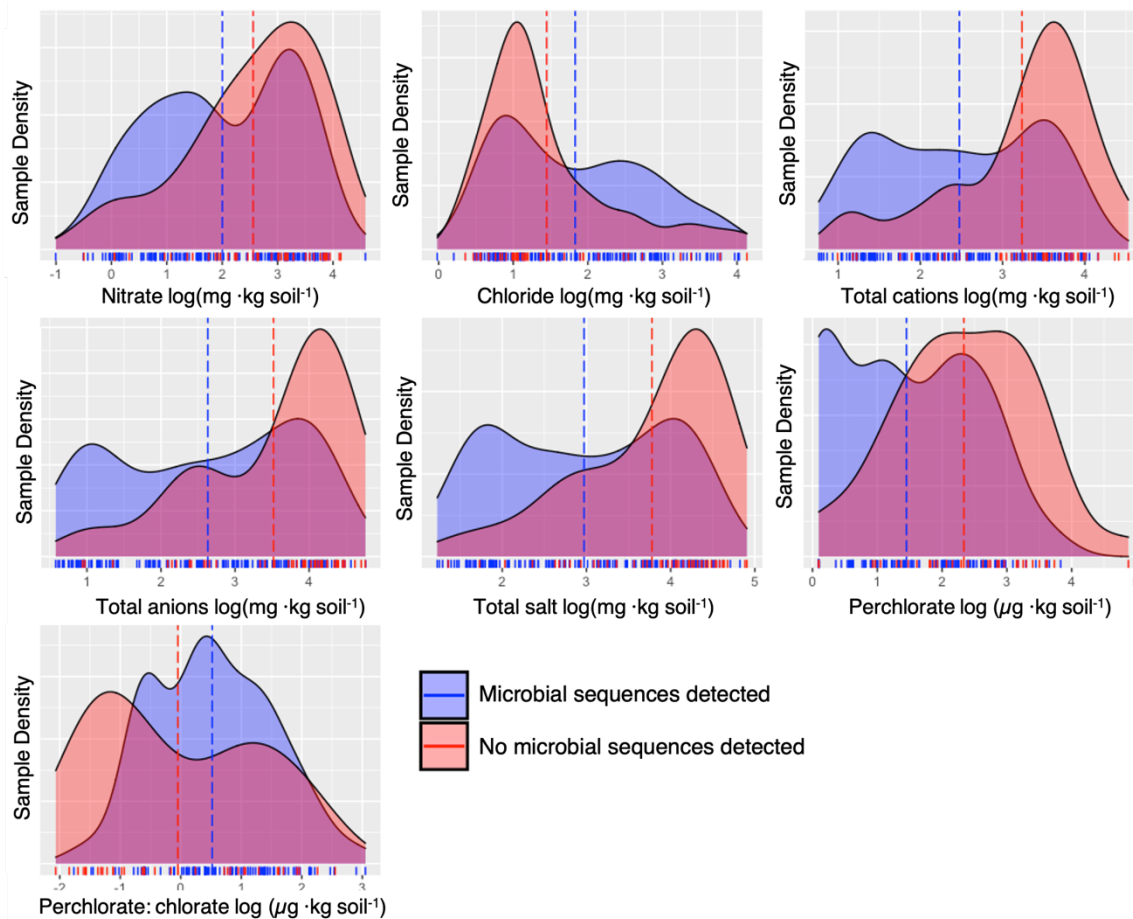
Supplemental Figure S2.5: The distribution of the geochemical and environmental variables used in the random forest model that were not significant predictors of a soil's habitability. Y axis shows sample density of each group (either with detectable microbial DNA or no detectable microbial DNA) with the number of samples (n) for each dataset ranging from 162-204. The area under each curve is equal to 1. Concentrations of all salt ions (µg·kg dry soil⁻¹, or mg·kg dry soil⁻¹) have been log transformed. Dashed lines indicate the average for each of the groups. The model results for each of the different variables are indicated in Supplemental Table S2.4.

Supplemental Table S2.1. The 35 samples used to confirm the results from the culture-independent genetic sequencing, their locations within the Shackleton Glacier Valley (see Methods).

Sample	Site	Elevation (m)	Distance from Coast (km)	Relative age of last wetting (yrs)
AV1-2	Augustana Valley	1492	72	3.61E+03
AV2-6	Augustana Valley	1376	72	4.06E+03
BP1-4	Bennett Platform	1329	82	4.47E+03
BP2-8	Bennett Platform	1222	82	3.47E+01
MF1-1	Mount Franke	409	9	6.43E+01
MF1-6	Mount Franke	484	9	3.47E+01
MF2-4	Mount Franke	424	9	9.72E+00
HV1-1	Heekin Valley	1660	63	8.68E+03
HV3-3	Heekin Valley	1140	63	4.75E+03
HV3-10	Heekin Valley	1030	63	5.86E+02
MSP1-4	Mount Speed	188	0	3.47E+01
MSP2-1	Mount Speed	270	0	3.47E+01
MSP3-4	Mount Speed	193	0	3.47E+01
NP1-1	Nielsen Peak		0	3.47E+01
NP3-2	Nielsen Peak	645	0	6.94E+00
NP4-6	Nielsen Peak		0	3.47E+01
RM1-1	Roberts Massif	1801	120	1.19E+04
RM2-4	Roberts Massif	1760	120	2.58E+02
RM3-7	Roberts Massif	1688	120	6.44E+02
SH1-1	Schroder Hill	2221	94	
SH1-2	Schroder Hill	2123	94	
SH1-3	Schroder Hill	2098	94	7.78E+04
SH1-4	Schroder Hill	2091	94	1.12E+05
SH1-5	Schroder Hill	2045	94	6.94E+04
SH1-6	Schroder Hill	2039	94	5.53E+04
SH2-5	Schroder Hill	2131	94	1.88E+05
SH3-8	Schroder Hill	2057	94	4.31E+03
TGV1-1	Thanksgiving Valley	1298	45	2.27E+04
TGV2-4	Thanksgiving Valley	1086	45	2.38E+04
TGV3-9	Thanksgiving Valley	911	45	
TN1-6	Taylor Nunatak	955	45	2.44E+04
TN1-9	Taylor Nunatak	883	45	1.72E+02
TN3-3	Taylor Nunatak	1023	45	2.65E+04

Supplemental Table S2.2: The results of the ^{13}C metabolism assay. The table shows the average %^{13}C (+/- 1 standard deviation) recorded from the ^{13}CO2 produced by the samples amended with glucose. 3 samples were classified as having no microbial activity detected, while 32 were classified as having microbial activity detected.

Treatment Group	Unautoclaved ^{13}C ‰	Autoclaved ^{13}C ‰
Microbial activity detected	9.90 +/- 9.65	1.15 +/- 0.09
No microbial activity detected	1.11 +/- 0.01	1.11 +/- 0.01
Positive control	15.91 +/- 10.01	1.10 +/- 0.01
Blanks	1.10 +/- 0.002	

Supplemental Table S2.3: The importance of the predictive variables of the random forest in describing whether a soil had detectable microbial DNA or not. Salt ion concentrations (μg·kg dry soil^{-1} or mg·kg dry soil^{-1}) were log-transformed before analysis. The importance of each variable is indicated by the % change of the mean standard error that is caused by the exclusion of that variable, and the significance (pval) of the estimation. This model explained 25.32% of the variance, with a mean of squared residuals of 0.130.

Variable	% of MSE explained	p Val
Elevation	7.85	0.02
Chlorate	7.66	0.01
ClO_4^-: ClO_3^- ratio	4.32	0.06
Total Salt	4.02	0.44
Total Cations	3.50	0.45
Perchlorate	3.23	0.29
Cl^-	2.40	0.45
NO_3^-	2.36	0.77
Total Anions	0.46	0.98

Supplemental Table S2.4: The growth media and growth conditions that were used for the culturing experiment.

Media	pH	Temperature (°C)	Source
Trypticase Soy Agar (TSA 24)	7.0	24	Atlas 1993
Trypticase Soy Agar (TSA 4)	7.0	4	Atlas 1993
M3 Acetate Agar (M3)	6.8	4	Atlas 1993
Antarctic Bacterial Medium (ABM)	7.0	4	Singh et al. 2016
Modified Nutrient Broth (MNB)	6.8	4	Pulschen et al. 2017
Luria Bertani Media (LB)	7.0	4	Atlas 1993
Malt Extract Agar (MEA)	5.5	4	Atlas 1993
Photoautotrophic media (PA)	7.0	24	Tahon and Willems 2017
Photoheterotrophic media (PH)	7.0	24	Tahon and Willems 2017
VNSS agar (VNSS)	7.0	4	Egan et al. 2015
V agar (V)	7.0	4	Egan et al. 2015
R2A (pH_5)	5.0	4	Reasoner and Geldreich 1985
R2A (pH_7)	7.0	4	Reasoner and Geldreich 1985
R2A (pH_9)	9.0	4	Reasoner and Geldreich 1985
R2A+5% NaOH (Salt_5)	7.0	4	Reasoner and Geldreich 1985
R2A+10% NaOH (Salt_10)	7.0	4	Reasoner and Geldreich 1985
Potato Glucose Agar (PGA)	5.6	4	Atlas 1993

ELEVATIONAL CONSTRAINTS ON THE COMPOSITION AND GENOMIC
ATTRIBUTES OF MICROBIAL COMMUNITIES IN ANTARCTIC SOILS

Materials and Methods

Sample collection and characterization

Soil samples were collected from the Shackleton Glacier region from December 2017- January 2018. The soil sampling process is described in detail in Diaz et al. (2021) and Dragone et al. (2021). In brief, soils were collected from 10 different features running the length of the valley including a range of elevations (150 - 2221 m above sea level (m.a.s.l.)) across a 120 km north- south distance spanning from the Ross Ice Shelf to the Polar Plateau (Figure 3.1). Between 14 and 26 soil samples were collected along elevational transects located on each of ten features to maximize variation in soil characteristics and soil exposure times (amount of time at the surface and uncovered by glacial ice) at each feature. Soils (0 - 5 cm depth) were collected in sterile polyethylene bags using ethanol cleaned hand trowels. GPS coordinates, photographs of the soil surface, elevation, and other environmental data were collected at the time of soil sample collection. All soils were transported to the field camp in insulated coolers where they were frozen at -20^oC and remained frozen until processed at the University of Colorado in Boulder, Colorado, USA.

Environmental and geochemical variables associated with each sample were measured as described in Diaz et al. (2021). For this study, we chose to focus on the following variables: elevation (m.a.s.l), nitrate (mg·kg soil^{-1}), chloride (mg·kg soil^{-1}), total cations (mg·kg soil^{-1}), total anions (mg·kg soil^{-1}), total salt (mg·kg soil^{-1}), perchlorate (µg·kg soil^{-1}), chlorate (µg·kg soil^{-1}), NH_3 (mg·kg soil^{-1}), SiO_2 (mg·kg soil-

1). We focus our analyses on these 10 variables as they are likely the most biologically relevant and were measured for at least 80% of the soil samples (Diaz et al. 2021; Dragone et al. 2021). All concentration measurements were log transformed prior to downstream analyses. Correlations between these environmental and biogeochemical data variables were calculated with the R function 'cor' (method = 'pearson') and correlation matrix plots were visualized using 'ggpairs' (R package 'GGally') (Supplemental Figure S3.1).

Cultivation-independent analysis

DNA was extracted from 204 samples in a laminar flow hood. After mixing 1 g of each soil in 1 mL of sterile PCR-grade water, DNA was extracted from a 500 µl aliquot of the soil slurry using the Qiagen DNeasy® Powersoil® HTP 96 Kit (Qiagen, Germantown, MD, USA) following the manufacturer's recommendations. A total of six extraction blanks (2 per 96-well plate) were included to test for any possible contamination introduced during DNA extraction.

The DNA aliquots extracted from each of the 204 soils, their associated six extraction blanks, and three no template controls, were PCR-amplified using a primer set that targets the hypervariable V4 region of the archaeal and bacterial 16S rRNA gene (515F: 5'- GTGCCAGCMGCCGCGGTAA-3' and 806-R: 5′-GGACTACHVGGGTWTCTAAT-3′) following the methods described in Dragone et al. (2021). These primers included the appropriate Illumina adapters and unique 12 - bp barcode sequences to permit multiplexed sequencing (Caporaso et al. 2012). The amplified products of all samples, blanks, and no template controls were cleaned and normalized to equimolar concentrations using SequalPrepTM Normalization Plates (Thermo Fisher Scientific, Carlsbad, CA, USA) and sequenced on an Illumina MiSeq run (Illumina, San Diego, CA, USA) using the V2 2 x 150 bp paired-end Illumina sequencing kits. The sequencing data generated from the soils can be

accessed in the NCBI Sequence Read Database, project accession number PRJNA699250.

The 16S rRNA gene sequences were processed using the DADA2 pipeline v.3.8 (Callahan et al. 2016). Sequences were quality filtered and clustered into exact sequence variants (ASVs), with taxonomy determined using a naïve Bayesian classifier method (Q. Wang et al. 2007) trained against the SILVA reference database v.132 (Quast et al. 2013; Yilmaz et al. 2014). A minimum bootstrapping threshold required to return a taxonomic classification of 50% similarity was used for analysis. For the soil DNA extracts, ASVs associated with chloroplast, mitochondria, eukaryotes, and those unassigned to the phylum level (717 ASVs) were removed prior to downstream analyses. We also excluded ASVs with fewer than 10 reads in total across the entire dataset (1567 ASVs). For our analysis, we used a cutoff of 1000 reads per sample as a threshold for inclusion in our analysis. This left 167 samples that had a sufficient number of prokaryotic 16S rRNA gene reads for downstream analyses with a mean number of reads per sample of 32048 (1086 - 73690). We note that our blanks and negative controls did not show any evidence of contamination during the extraction or amplification steps (see Dragone et al. (2021) for more detail).

Quantitative PCR

To estimate how prokaryotic DNA concentrations vary across the sample set, we used quantitative PCR (qPCR) to measure bacterial 16S rRNA gene copy numbers using the same primers and soil DNA extracts used for sequencing. Reaction conditions and details followed methods described previously (Carini et al. 2016). The 167 soil samples, corresponding extraction blanks, and 16 no-template controls were used for the 16S rRNA gene qPCR analyses. Standard curves were calculated using purified genomic DNA from *Escherichia coli* for 16S rRNA copy

number. Based on the data from the negative controls, samples with a cycle threshold (Ct) value greater than 31 were considered below detection limits. Calculated copy number measurements for each sample are reported as number of *E. coli* genome equivalents · g soil^{-1}.

Microbial community analyses via marker gene sequencing

Community analyses of the sequenced soils were performed in R v.4.0.5 (R Core Team 2017) Richness was calculated from the filtered 16S rRNA gene ASV tables using 'specnumber' (R package 'Vegan'). Plots of relative abundance were created using the R package 'mctoolsr' (https://github.com/leffj/mctoolsr/) as were the NMDS plots. To measure differences between communities across the Shackleton Glacier region, we calculated pairwise Bray-Curtis dissimilarities from the ASV tables using the 'calc_dm' function (R package 'mctoolsR'). To identify the best model that explains the differences in overall prokaryotic community composition across the soils, we used BIOENV (Clarke and Ainsworth 1993; Dixon 2003) to identify the subset of biologically relevant environmental and geochemical variables that maximizes the correlation to Bray Curtis- dissimilarities (method = spearman). For these analyses, we only included samples where every variable was measured (108 soils). We then confirmed the correlation of each variable identified to the pairwise Bray-Curtis dissimilarities from the full set of 167 samples with Mantel tests. For all Mantel tests, distance matrices were calculated with the R function 'dist' and Mantel statistics are based on the Pearson's product-moment correlation method.

Phylogenetic tree construction was performed with the 100 most abundant bacterial ASVs. Phylogenetic relatedness of the 100 ASVs were determined via maximum likelihood with RaxML v.8.0.0 (raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -number 100, (Stamatakis 2014a)) including *Gemmata obscuriglobus* as

the outgroup. Sequences were aligned using MUSCLE (Edgar 2004) and the tree was visualized and annotated using iTOL v.6.3.2 (Letunic and Bork 2016).

Network analysis and niche modeling of prokaryotic communities

To identify modules of co-occurring prokaryotic ASVs across the 167 soils, we performed network analyses on the filtered ASV table generated from the culture-independent sequencing of DNA extracted from soils. We included all ASVs that, after the filtering steps described above, were found in at least 10 samples (3710 ASVs). A correlation matrix was generated using the R function 'correlate' (method= "spearman"). This matrix was filtered so that only positive correlations >0.75 were kept. This left a final edge list of 4274 correlations from 885 nodes. Network analyses were conducted and visualized using the R package 'igraph.' Routes were generated from the node and edge lists with 'graph_from_data_frame' and community structure were found using 'cluster_louvain.' From these network analyses, we found 88 modules of co-occurring prokaryotic ASVs with each module containing between 2-154 ASVs.

We performed random forest analyses to determine which, if any, of the measured environmental and geochemical variables were the best predictors of where the 88 prokaryotic modules could be found across the Shackleton Glacier Region. For the purpose of our modeling, a module was reported as being present in a sample if reads associated with any of the ASVs assigned to that module was present. For our random forest models, we used the R package 'rfPermute' and performed a random forest analysis with 100 trees and three variables tested at each split to identify the most important predictors. Models were accepted if the percent variance explained was >10% and 38 modules had predictive models that passed this threshold. For these models, the variable that most increased the MSE was identified as the variable that was most predictive of where the taxa within

each module were most likely to be found, so long as that variable increased the MSE by at least 5% ($p < 0.05$). Predictive soil and site variables were identified for 28 of the prokaryotic modules. To visualize these relationships, the average standardized relative abundance (Z score) of each of the 28 modules was plotted against their respective predictive variable. Z scores of each ASV were calculated from the filtered table of read counts using the R function 'zscore.' The average standardized relative abundance of each module in each sample was calculated by averaging the Z scores of all ASVs assigned to that module.

Metagenomic sequencing and annotation

We chose 27 of the 167 samples for shotgun metagenomic sequencing (Table S2). This subset of samples was chosen to include at least two soils from each of 8 sampled features to span the range of edaphic properties found across our dataset. We chose not to include samples from Schroeder Hill (SH) and Roberts Massif (RM) because the results of our amplicon sequencing effort suggested that we would not be able to extract enough DNA from these soils. To obtain sufficient DNA for metagenomic sequencing of the 27 samples, we re-extracted DNA from these soils in triplicate using the Qiagen DNeasy® Powersoil® Kit (Qiagen, Germantown, MD, USA). The manufacturer's protocols were followed except that DNA from all three replicates were combined on the same spin filter at the final step. This DNA was used to generate metagenomic libraries with the Nextera DNA Flex library preparation kit (Illumina, San Diego, CA, USA). The manufacturer's protocol was followed except that the number of PCR cycles were increased for low biomass samples as suggested in Bruinsma et al. (2018). and by Illumina Tech Support. Libraries were sequenced on an Illumina NextSeq 500 run using a high output 300-cycle kit with paired-end chemistry at the University of Colorado Boulder's Next-

Generation Sequencing Facility. The raw metagenomic data can be also be accessed in the NCBI Sequence Read Database, project accession number PRJNA699250.

Prior to downstream analyses, we removed adapter sequences from the raw sequence data using Cutadapt v.2.1 with the recommended options for paired end Illumina reads (Martin 2011) and filtered reads based on sequence quality using Sickle v.1.33 (-q 20 -I 50) (https://github.com/najoshi/sickle). After this quality filtering, we obtained an average of 23.6 million quality filtered reads per sample (range 17.0 million to 28.4 million reads). The relative abundances and diversity of bacteria and archaea in the metagenomic samples were determined by extracting 16S rRNA gene reads from the metagenomic sequence data using phyloFlash v. 3.0 (Gruber-Vodicka et al. 2020). To verify that the 16S rRNA amplicon data was consistent with the taxonomic composition of the bacterial communities as inferred from the metagenomic data, we tested the correlation between the Bray-Curtis dissimilarity matrices of the amplicon and metagenomic datasets using Mantel tests as described above (r = 0.80, p <0.001).

Assembly-free analyses on the trimmed and quality filtered data were performed using SqueezeMeta v.0.1.0 with the alternative analysis mode sqm_reads.pl script (Tamames and Puente-Sánchez 2019) which uses DIAMOND v2.0.11 (Buchfink et al. 2015) to annotate reads with the KEGG ontology (Kanehisa and Goto 2000; Kanehisa 2019; Kanehisa et al. 2021). We obtained an average of 7.2 million annotated reads per sample across all 27 samples (range 5.2 million – 8.7 million reads). To control for differences associated with variation in the number of annotated reads per sample, we rarified each sample to 5,203,694 annotated reads per sample using the R package 'vegan'. This rarefied table was normalized using MUSiCC v.1.0.3 to obtain more robust measures of gene abundances normalized to abundances of universal single copy genes (Manor and Borenstein 2015).

Estimation of microbial growth rate

To estimate the maximal microbial growth rate, we used the tool gRodon which estimates maximal microbial growth rates from codon usage biases in highly expressed genes, an indicator of selection for rapid growth (Weissman et al. 2021; Vieira-Silva and Rocha 2010). Briefly, we assembled the sickle-filtered reads with MEGAHIT v.1.2.9 (preset: meta-large) (Li et al. 2015) and mapped the filtered reads back onto the reference using Bowtie2 v.2.4.4 (default parameters) (Langmead and Salzberg 2012). We then used Metaprokka v.1.14.6 (https://github.com/telatin/metaprokka) to annotate the assembled reads. After annotation, we used the tool featureCounts (Liao et al. 2014) to calculate the number of filtered reads mapping to each gene and then converted these mapping counts to transcripts per million (TPM) (Pachter 2011; Wagner et al. 2012) to normalize for differential sequencing depth across samples and differences in gene length. We then ran gRodon (Weissman et al. 2021) in metagenome mode to calculate codon usage biases between highly expressed ribosomal proteins and background codon usage. We also followed the authors' recommendations for extremophiles and used the temperature setting to set a growth temperature of 0°C for all samples. We note that gRodon is not calibrated for the extremely low temperature environments found in Antarctica and maximal microbial growth rates should be considered estimates. For this reason, we focus on the relative comparison in estimated maximal growth rates between categories of samples instead of the specific values obtained.

Analysis of metagenomic sequencing data

Functional diversity was determined using the rarefied and MUSiCC-normalized KEGG table. To identify which annotated genes were more abundant at different elevations, we grouped the 27 soils into two different categories. "High

elevation" samples were those collected above 800m (n = 15) while "low elevation" samples were those collected below 800 m (n = 12). We chose 800m as the separation between the two categories as no samples were collected from an elevation +/- 75 m of 800 m. This 150 m "gap" corresponds to the average elevation of this subset of 27 samples (853 m). Additionally, the group of 15 samples from above 800 m had significantly higher concentrations of nearly all measured geochemical variables than the group of 12 samples collected below 800 m (Supplemental Figure S3.7).

To identify differences in functional gene abundances between the high elevation and low elevation groups, we first compared KEGG richness (number of unique KEGGs in each sample) using a Mann-Whitney nonparametric test using the R function 'wilcox.test.' Then we calculated the log2 fold change in average gene abundance across the two elevation categories for each KEGG following methods described in (Quackenbush 2002). KEGGs were classified as being consistently more abundant at higher elevations if they were, on average, more than twice as abundant at higher elevations than lower elevations (log2 fold change < -1). KEGGs that were identified as being more abundant at higher elevation were annotated based on the KEGG Orthology database gene catalogs (Kanehisa 2019; Kanehisa and Goto 2000; Kanehisa et al. 2021).

To make predictions about potential functions that are more prevalent at higher elevation sites, annotated genes were categorized into larger functional categories based on their locations in the KEGG pathway database and/or the KEGG BRITE Database (Kanehisa 2019; Kanehisa and Goto 2000; Kanehisa et al. 2021). We did not consider pathways that are exclusively associated with the eukaryotic organisms (KEGG pathway categories: "organismal systems," "Human Diseases," "Drug Development," BRITE Categories: "Drugs," "Diseases," and others associated with eukaryotic organisms). For the purposes of assigning a potential

function, if a gene was associated with multiple pathways, it was included in both pathways. Pathways of interest were those identified as having at least 5 genes that were >2 times more abundant above 800m (Supplemental Figure S3.5). We note, none of the pathways of interest were complete, with every gene >2 times as abundant above 800m.

Targeted analysis of trace gas metabolism genes

For more detailed, targeted analyses of genes related to trace gas metabolism, we followed the approach described previously (Bay et al. 2021). To summarize, the quality filtered and trimmed paired end reads (see above for more details) were searched for the presence of 10 metabolic marker genes related to trace gas metabolism using the blastx function of DIAMOND v.2.0.11(Buchfink et al. 2015). These included CoxL, MmoX, PmoA, Group 1c (NiFe)-Hyd, Group 1d (NiFe)-Hyd, Group 1f (NiFe)-Hyd, Group 1h (NiFe)-Hyd, Group 1l (NiFe)-Hyd, Group 2a (NiFe)-Hyd, Group 3 (NiFe)-Hyd. More specifically, sequence reads were searched against protein sequences of these 10 genes downloaded from the Greening lab metabolic marker gene database v.1 (Leung and Greening 2021) using a query coverage of 80%. Following the methods of Bay et al. (2021), hits were kept if they had an identity threshold of 50% for the NiFe-Hyd genes or 60% for all others, and a maximum e-value threshold of 10^{-10}. Reads per gene were divided by the total number of trimmed and quality filtered reads and are reported in all downstream analysis as proportion of total reads. Differences between high elevation soils and low elevation soils were assessed using Mann- Whitney non-parametric tests as described previously. Z scores were calculated using the R function 'zscore.'

Acknoledgements

Supplemental Figure S3.1: Correlations between selected soil and site characteristics across the 153 samples used for this study for which all measured variables were available. Pearson correlation coefficients are inset into each scatterplot. While our analyses included soil NH_3 and SiO_2 concentrations, they have not been included in this figure as they were not strongly correlated with any other category.
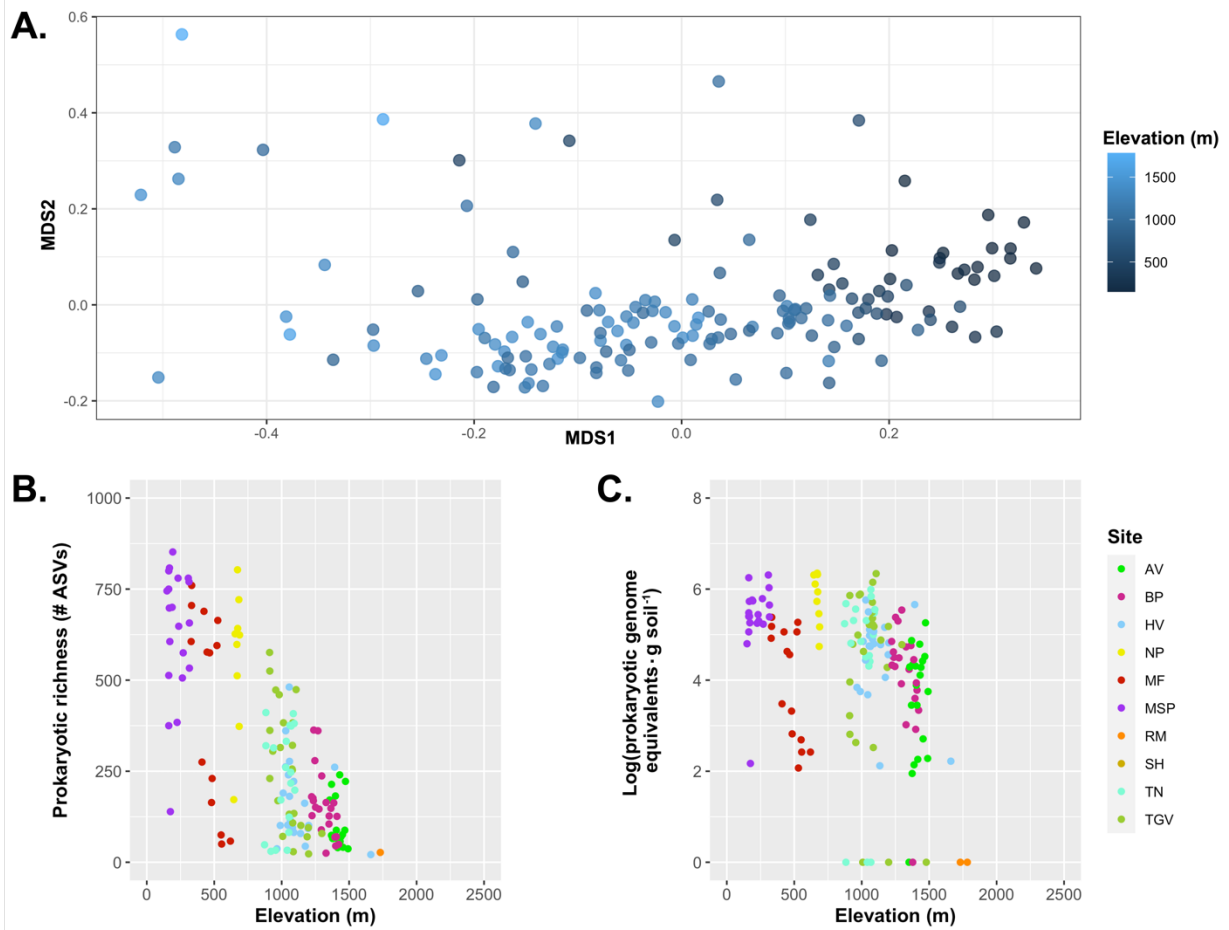
Supplemental Figure S3.2: Variation and richness of the microbial communities and amounts of microbial DNA in soils of the Shackleton Glacier region. **A)** NMDS showing dissimilarity in the taxonomic composition of the microbial communities in the 167 samples based on the 16S rRNA gene sequencing results. **B)** The relationship between bacterial richness (number of distinct prokaryotic ASVs out of 2000 reads per sample) and elevation (m.a.s.l). **C)** The relationship between the qPCR estimates of the prokaryotic DNA concentrations in each soil and elevation. Note that zero values indicate that DNA concentrations were below the limit of detection using our qPCR approach, not that these soils have a complete absence of microbial DNA. In all panels, soils are colored by the feature from which they were collected (see map, Figure 3.1, for location of sampled features within the Shackleton Glacier region).
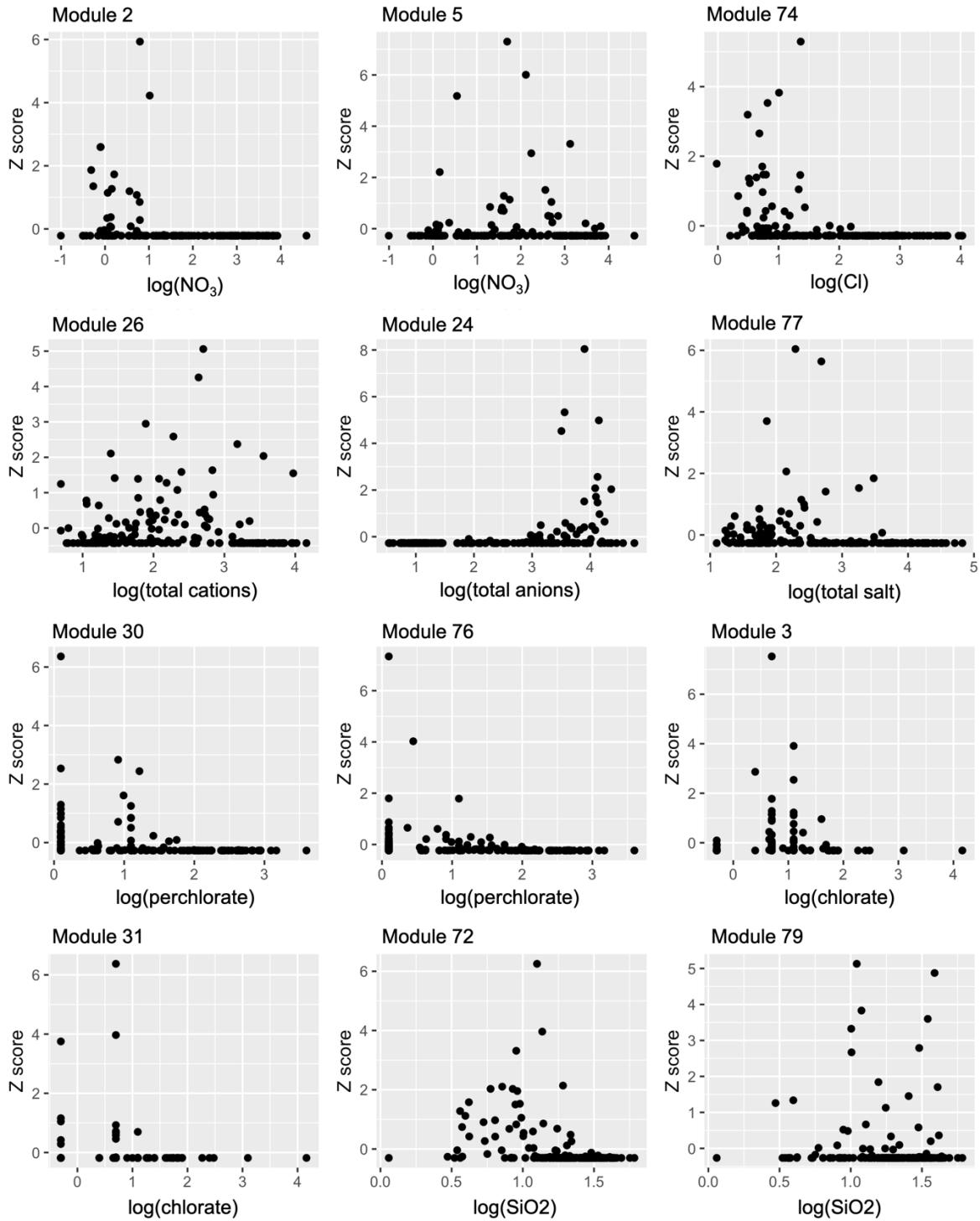
Supplemental Figure S3.3: The environmental preferences of 12 bacterial modules. The 12 bacterial modules that were not best predicted by elevation (see text for details). Z scores are the average standardized relative abundances of all of the ASVs associated with each module (range of 2 to 154 ASVs per module).
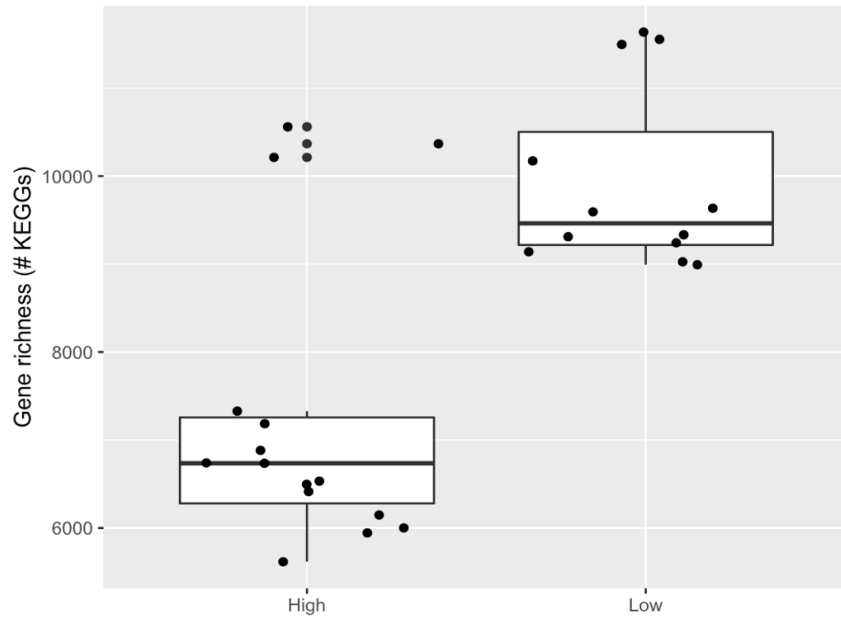
Supplemental Figure S3.4: KEGG gene richness in soils above and below 800m. 15 samples make up the above 800m "high elevation" group while 12 samples make up the below 800m "low elevation" group. KEGG gene richness samples was significantly lower in the higher elevation samples than in the lower elevation soil samples (Mann-Whitney U, p < 0.001).
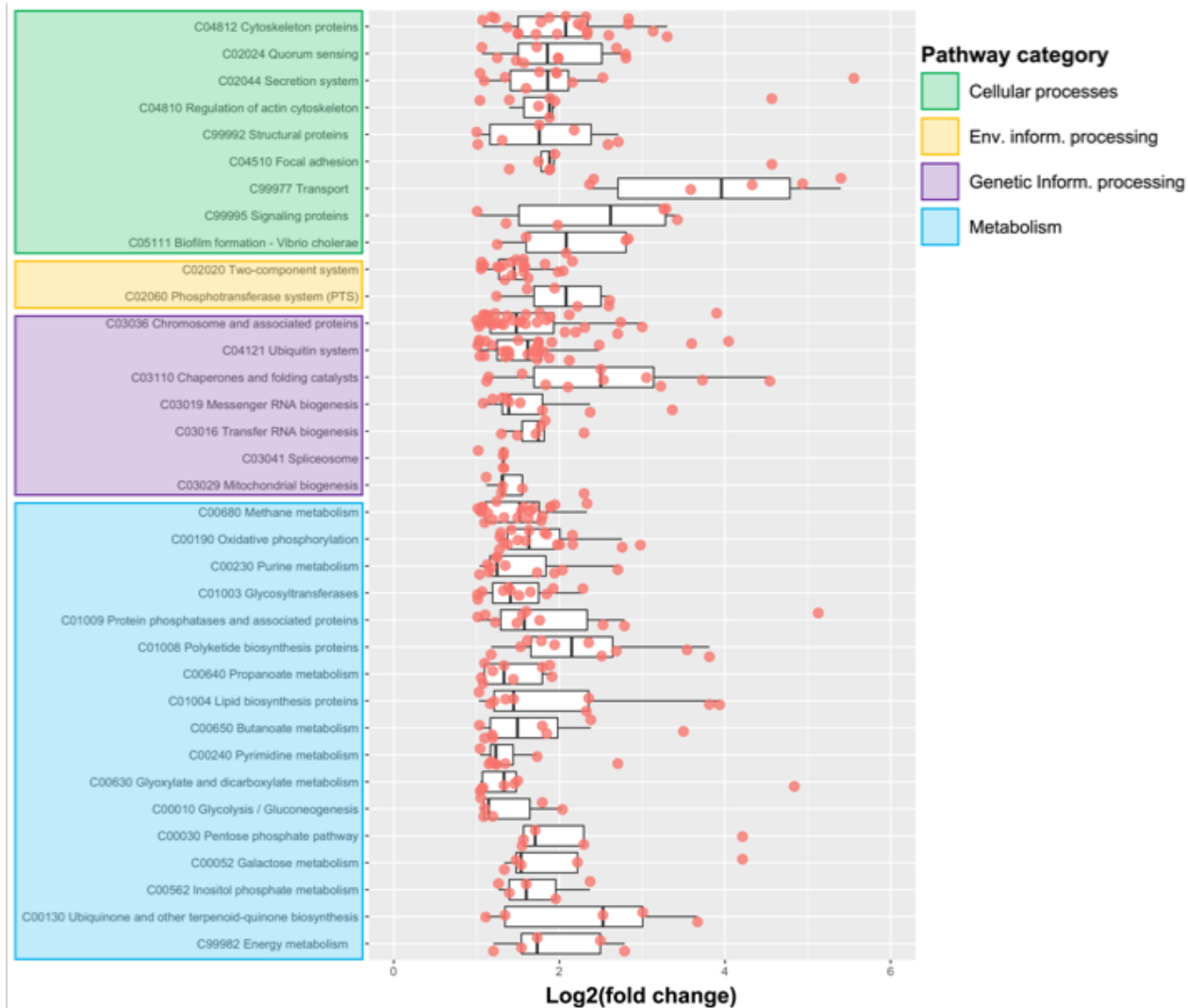
Supplemental Figure S3.5: Gene pathways that are relatively more abundant in higher elevation soils. Shown here are those gene categories with at least 5 genes that were identified as being >2 times more abundant in samples above 800 m across our sample set (comparing 15 soils collected from above 800 m and 12 soils below 800 m).
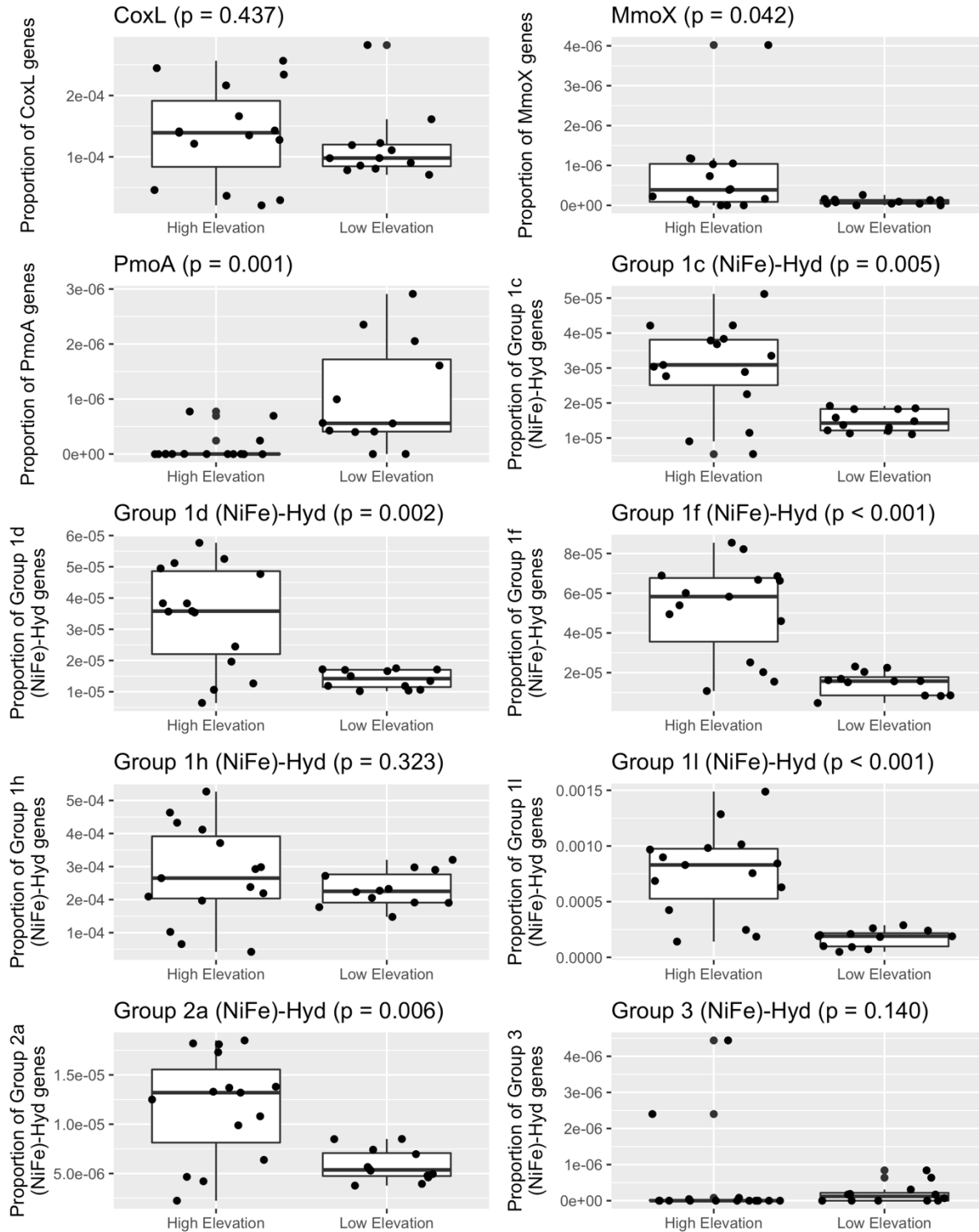
Supplemental Figure S3.6: A comparison of the abundance of 10 genes associated with trace gas metabolism. High elevation samples were collected from above 800 m (n=15) while low elevation samples were collected from below 800m (n=12). For all comparisons, the significance (p, as determined by a Mann-Whitney U test) is included with the gene name in the title of each plot.
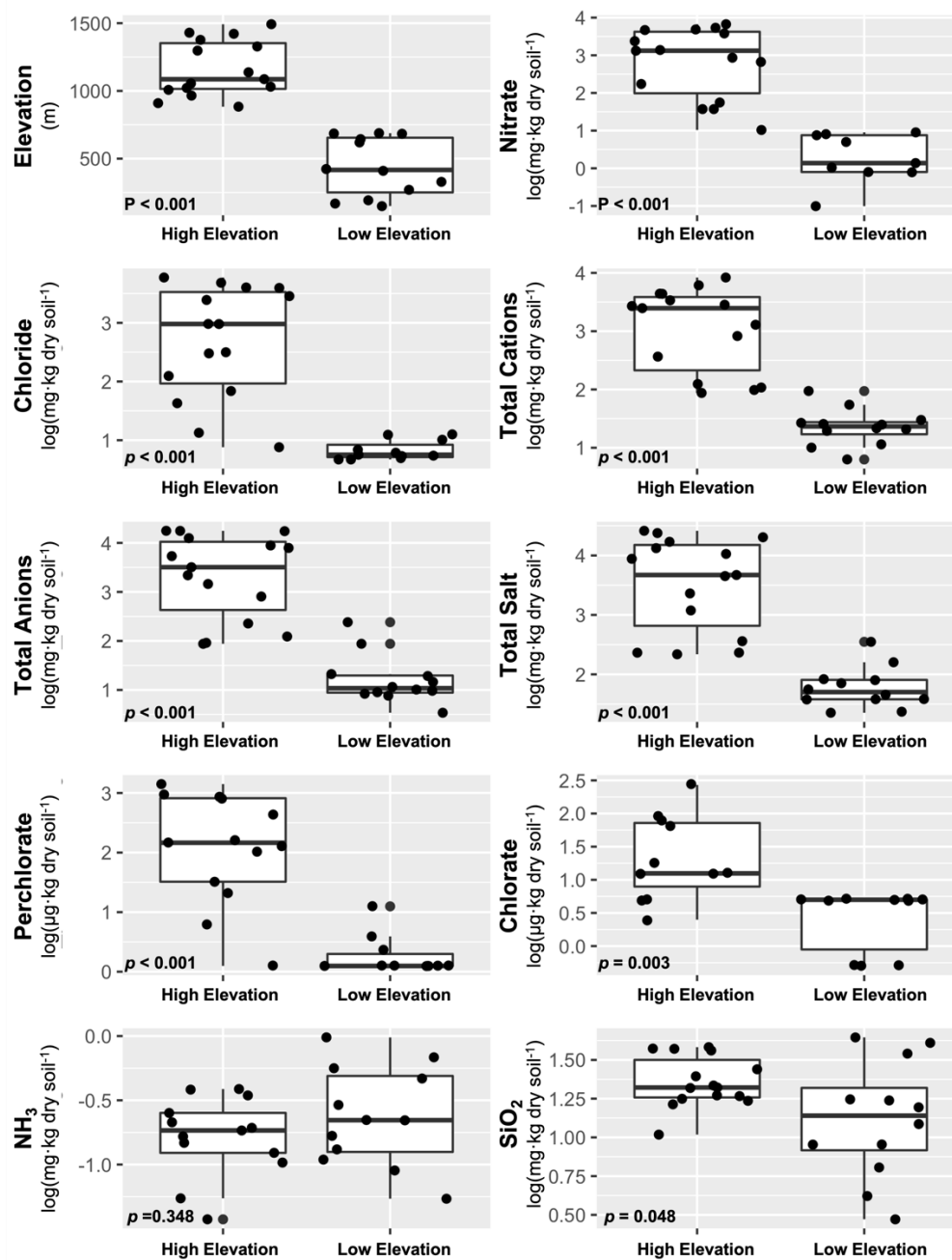
Supplemental Figure S3.7: A comparison of the properties of the high elevation and low elevation samples selected for shotgun metagenomic analyses. High elevation samples were collected from above 800 m (n=15) while low elevation samples were collected from below 800m (n=12). For all comparisons, the significance (p, as determined by a Mann-Whitney U test) is inset in the plot.

Supplementary Table S3.1: Summary of the results of the random forest analyses. A variable was determined to be the most predictive variable if the model explained at least 10% of the variance and if that variable increased the MSE by at least 5% ($p < 0.05$).

Most predictive variable	Prokaryotic Module
Elevation (m)	1,10,12,14,27,40,41,43,44,52,58, 66, 75,78,80,88
Nitrate (mg·kg soil^{-1})	2,5
Chloride (mg·kg soil^{-1})	74
Total cations (mg·kg soil^{-1})	26
Total anions (mg·kg soil^{-1}).	24
Total salt (mg·kg soil^{-1})	77
Perchlorate (µg·kg soil^{-1})	30,76
Chlorate (µg·kg soil^{-1})	3,31
NH$_3$ (mg·kg soil^{-1})	
SiO2 (mg·kg soil^{-1})	72,79

Supplementary Table S3.2: The 27 soil samples used for shotgun metagenomic sequencing.

Sample	Site	Elevation (m)	Group
AV1_2	Augustana Valley	1492	High Elevation
AV2-6	Augustana Valley	1376	High Elevation
AV3_2	Bennett Platform	1430.1	High Elevation
BP1_4	Bennett Platform	1329	High Elevation
BP3_1	Bennett Platform	1421.9	High Elevation
HV2_7	Heekin Valley	965	High Elevation
HV3_10	Heekin Valley	1030	High Elevation
MF1_1	Mount Franke	409.3	Low Elevation
MF1_2	Mount Franke	620	Low Elevation
MF2_4	Mount Franke	424	Low Elevation
MFx2	Mount Franke	329	Low Elevation
MSP2_1	Mount Speed	270	Low Elevation
MSP3_1	Mount Speed	149.9	Low Elevation
MSP3_3	Mount Speed	168.4	Low Elevation
MSP3_4	Mount Speed	192.6	Low Elevation
NP1_1	Nielsen Peak	688	Low Elevation
NP2_1	Nielsen Peak	682.7	Low Elevation
NP3_2	Nielsen Peak	644.9	Low Elevation
NP4_6	Nielsen Peak	685	Low Elevation
TGV1_1	Thanksgiving Valley	1298	High Elevation
TGV2_4	Thanksgiving Valley	1086	High Elevation
TGV3_5	Thanksgiving Valley	1007.6	High Elevation
TGV3_9	Thanksgiving Valley	910.5	High Elevation
TN1-1	Taylor Nunatak	1137	High Elevation
TN1_9	Taylor Nunatak	883	High Elevation
TN2_5	Taylor Nunatak	1056	High Elevation
TN3-3	Taylor Nunatak	1022.5	High Elevation

CHAPTER IV APPENDIX

THE EARLY MICROBIAL COLONIZERS OF A SHORT LIVED VOLCANIC
ISLAND IN THE KINGDOM OF TONGA

Materials and Methods

Sample Collection

Surface sediments (0 - 5cm depth) were collected from Hunga Tonga Hunga

Ha'apai (HTHH), by the scientists, students, and crew of the SSV Robert C.

Seamans on October 14, 2018, and from October 8-9, 2019. A total of 32 sites were

sampled (Figure 4.1). Sediment was collected in sterile polyethylene bags using

aseptic techniques. GPS coordinates, photographs of the soil surface, and other

metadata were taken at the time of sample collection. Samples remained frozen at -

$20^{\circ}C$ until they were processed at the University of Colorado in Boulder, Colorado,

USA.

Sediment characterization

Sediment pH was determined for all samples according to the method

described in (King et al. 2010). Specifically, 5g of soil and 5g of DI water were placed

in a 15ml conical tube and shaken for 2 hours at 200 rpm. Sediment pH was then

measured with an Orion Star™ A211 Benchtop pH meter (ThermoFisher Scientific,

Waltham, MA, USA). Other geochemical measurements were performed on freeze-

dried and crushed aliquots of each sample. Total nitrogen content (TN) and total

organic carbon (TOC) measurements for all samples were measured by the Arikaree

Laboratory at the University of Colorado Boulder using a Shimadzu

TOC-L/TNM-L TOC/TN analyzer. Additional geochemical analyses were performed

on 25 of the samples by Activation Laboratories Ltd. (Ancaster, Ontario, Canada)

with the composition of trace elements measured via lithium borate fusion

inductively coupled plasma mass spectrometry (ICP-MS) and instrumental neutron activation analysis (INAA). Statistical tests to assess differences in the environmental and geochemical properties across sample categories were conducted using Mann-Whitney U tests for two groups or Kruskal Wallis one way analysis of variance tests with the post hoc Dunn's multiple comparison test for three or more.

DNA Extractions of sediment

DNA was extracted from 0.5 g of each sample in a laminar flow hood using the Qiagen DNeasy® Powersoil® Kit (Qiagen, Germantown, MD, USA) following the manufacturer's recommendations. A total of 2 extraction blanks were included to test for any possible contamination introduced during the DNA extraction. Cultivation-independent microbial analyses via marker gene sequencing: The DNA aliquots extracted from each of the 32 samples and their associated 2 extraction blanks were PCR- amplified using a primer pair that targets the hypervariable V4 region of the archaeal and bacterial 16S ribosomal RNA (rRNA) gene (515F: 5'-GTGCCAGCMGCCGCGGTAA-3' and 806-R: 5′-GGACTACHVGGGTWTCTAAT-3′) (Walters et al. 2016). This primer sets included the appropriate Illumina adapters and unique 12 - bp barcode sequences to permit multiplexed sequencing (Caporaso et al. 2012). Two no-template PCR blanks were included with each set of PCR amplifications. Amplifications were performed on a SimpliAmp Thermal Cycler (Thermofisher Scientific, Waltham, MA, USA) using Platinum™ II Hot-Start PCR Master Mix (2X) (Invitrogen, Carlsbad, CA, USA) in 25 µL reaction volumes. Cycling parameters consisted of an initial denaturation step at 94 °C for 3 min, followed by 35 cycles of denaturation at 94 °C (45 s), annealing at 50 °C (60 s), extension at 70 °C (90 s), and a final extension step at 72 °C for 10 min. The amplified products were cleaned and normalized to equimolar concentrations using SequalPrep™ Normalization Plates (Thermo Fisher Scientific, Carlsbad, CA,

USA) and were sequenced on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) using the V2 2 x 150 bp paired-end Illumina sequencing kit at the University of Colorado Boulder's Next- Generation Sequencing Facility. Raw sequencing data can be accessed in the NCBI Sequence Read Archive, project accession number PRJNA914229.

The 16S rRNA gene sequences were processed using the DADA2 pipeline, version 3.8 (Callahan et al. 2016). Sequences were quality filtered and clustered into exact sequence variants (ASVs), with taxonomy determined using a naïve Bayesian classifier method (Q. Wang et al. 2007) trained against the SILVA v.132 reference database (Quast et al. 2013; Yilmaz et al. 2014). A minimum bootstrapping threshold required to return a taxonomic classification of 50% similarity was used for analysis. The two extraction blanks yielded a total of 593 reads. None of the 11 ASVs that these reads were assigned to were found in any of the samples. The no-template PCR controls had reads associated with 37 ASVs, 6 of which were also found in other samples. However, reads associated with these ASVs were not common and only accounted for 0.04% of the reads from all extracted samples and were found in a maximum of 2 samples. Prior to downstream analyses, ASVs associated with chloroplast, mitochondria, eukaryotes (137 ASVs total), and those unassigned to the phylum level (211 ASVs) were removed as were ASVs with fewer than 10 reads across all samples. After this filtering, samples were included in downstream analysis if they met a threshold of 10000 reads per sample. This left 27 samples that had sufficient prokaryotic 16S rRNA gene reads for downstream analyses with a mean number of reads per sample of 52000 (12207 - 86863).

Microbial community analyses

Community analyses of the sequenced soils were performed in R v.4.0.5 (R Core Team 2017). Richness, the number of distinct prokaryotic ASVs out of 10,000

reads per sample, was calculated from the filtered 16S rRNA gene ASV tables using 'specnumber' (R package 'Vegan') (Dixon 2003). Jaccard distances were also calculated with vegan, using the function 'vegdist(method= "jaccard")'. Plots of relative abundance were created using the R package 'mctoolsr' (https://github.com/leffj/mctoolsr/) as were the NMDS plots. Phylogenetic tree construction was performed with the 100 most abundant bacterial ASVs identified in the cone samples. Phylogenetic relatedness of the 100 ASVs were determined via maximum likelihood with RaxML v.8.0.0 (raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -number 100, (Stamatakis 2014b)) including *Cenarchaeum symbiosum* as the outgroup. Sequences were aligned using MUSCLE v3.8.1551 (Edgar 2004) and the tree was visualized and annotated using iTOL v.6.3.2 (Letunic and Bork 2016).

To determine the percentage of taxa identified in the inland cone sediments that were from families that had been previously cultured, we used the Ribosomal Database Project's (RDP) SeqMatch tool to search for the sequences of the top 100 ASVs recovered from the cone against the RDP v.11.5 database of 16S rRNA sequences filtered to include just those from isolated organisms (Cole et al. 2014).

We note that while samples were collected in two different years, we do not treat them as distinct sample categories. Comparisons between years could not be made because the specific locations sampled in 2018 were distinct from those sampled in 2019. However, identical methods were used to collect and process all samples and we found no significant differences in the composition of bacterial and archaeal communities across the two years (PermANOVA, r < 0.01).

Comparisons against reference databases of potential source environments

To shed light on the potential origin of the microbial diversity on the island, we examined the ASVs recovered from the inland cone sediments for bacterial taxa

characteristic of the potential source environments. First, we calculated the percentage of reads associated with the 6 bacterial families that were found to be the most ubiquitous in survey of the guts of 74 bird species (Capunitan et al. 2020). In the reference study of the 74 bird species, these 6 families were found in every single sample (Capunitan et al. 2020), yet we found they only make up a total of 1.31% of the total reads recovered from the inland cone samples. Next, we searched for the 10 most abundant bacterial orders found in seawater (Sunagawa et al. 2015). These 10 bacterial orders made up a total of 65% of all reads recovered in a global survey of the microorganisms in seawater (Sunagawa et al. 2015)but in total represented <0.05% of the reads in the inland samples. See Supplemental Table S4.1 for more information about the identity of these organisms.

To compare the microbial communities in the inland cone sediments to soil environments, we performed a similar search as described above for the most abundant bacterial phylotypes identified in a global survey of soil environments (see Supplemental Table S4.1 and Delgado-Baquerizo et al. (2018) for more detail). We then took the representative 16S rRNA gene sequences for each of the top 100 ASVs identified in the inland cone and vegetated samples and performed a blastn search at 100% identity against a reference database of 515 bacterial taxa that dominate the microbial communities in soil environments (Delgado-Baquerizo et al. 2018).

To compare the top ASV sequences against communities in hydrothermal vent environments, we reprocessed sequencing data downloaded from NCBI (BioProject accession: PRJNA546572) from samples collected from the Brothers volcano complex (Reysenbach et al. 2020) with the DADA2 pipeline following the methods described previously. The sequences of the top 100 ASVs identified in the cone sediments and the vegetated samples were then searched against this dataset following the same methods described previously for the soil reference database.

Quantitative PCR

To estimate how prokaryotic DNA concentrations vary across the sample set, we used quantitative PCR (qPCR) to measure bacterial 16S rRNA gene copy numbers in the 32 sediment samples, two corresponding extraction blanks, and two no-template controls. These analyses were performed on a BIO-RAD CFX Connect Real-Time System (Bio-Rad Laboratories, Hercules, CA, USA) using the same primers and soil DNA extracts used for sequencing. Reaction conditions and details followed methods described previously in (Carini et al. 2016). Reactions were performed in triplicate for each sample with the average of the readings being used for analysis. Standard curves were calculated using purified genomic DNA from *E. coli*. Calculated copy number measurements for each sample are reported as number of *E. coli* genome equivalents $\cdot$ g soil^{-1}.

Metagenomic Sequencing

A total of 19 samples were selected for shotgun sequencing. These samples included all the samples collected from the inland cone (>30 meters from the ocean) and from vegetated soils that predate the eruption from the bordering island of Hunga Tonga. With aliquots of the same DNA pools used for the amplicon sequencing, we generated metagenomic libraries with the Nextera DNA Flex library preparation kit (Illumina, San Diego, CA, USA). The manufacturer's protocol was followed except that the number of PCR cycles were increased for low biomass samples following suggestions by Illumina Tech Support based on previous studies (Bruinsma et al. 2018). Libraries were sequenced on an Illumina NextSeq 500 run using a high output 300- cycle kit with paired-end chemistry at the University of Colorado Boulder's Next-Generation Sequencing Facility. The raw metagenomic data have been deposited in the NCBI Sequence Read Archive, project accession number PRJNA914229.

The raw paired end sequencing reads were interleaved using BBmap v.38.94 (https://sourceforge.net/projects/bbmap/), Illumina adapters were removed using Cutadapt v.2.1 (Martin 2011) and were filtered based on sequence quality using Sickle v.1.33 (-q 20 -I 50) (https://github.com/najoshi/sickle). Two samples (HTHH_2019_5, HTHH_2019_15) had fewer than 100 reads remaining after this quality filtering and were not included in downstream analysis. For the remaining samples, we had an average of 46.8 million quality filtered reads pers sample (range 10.7 million to 64.7 million reads). The filtered FASTQ sequences for these remaining samples were reformatted to FASTA using BBmap.

We used phyloFlash v.3.0 (Gruber-Vodicka et al. 2020) on these filtered reads to verify that the metagenomic data were consistent with the taxonomic composition of the bacterial communities as inferred from the 16S rRNA gene amplicon data. We tested the strength and significance of this relationship with a Mantel test comparing the Bray-Curtis dissimilarity matrices of the amplicon and metagenomic datasets performed using the R package 'vegan' and found that they were well correlated (Mantel test, $r = 0.90$, $p < 0.001$, Figure S5) (Dixon 2003).

Targeted gene analysis

Targeted analyses of 294 genes related to photosynthesis, trace gas metabolism, sulfur metabolism, and iron metabolism were performed on the 23 trimmed and quality filtered metagenomes. Sequence reads were searched against compilations of protein sequences from existing databases and repositories using the blastx function of DIAMOND v.2.0.11 with a query coverage of 80% (Buchfink et al. 2015). The sequences of the trace gas genes were downloaded from the Greening lab metabolic marker gene database v.1 (Leung and Greening 2021), the sequences of the iron metabolism genes were from the FeGenie database (Garber et al. 2020), and the sequences of the sulfur metabolism genes were those from the National

Center for Biotechnology Information (NCBI) reference sequence database as compiled for SCycDB (Yu et al. 2021). The photosynthesis genes used were compiled from Imhoff et al. (2019), Leung and Greening (2021), Figueras et al. (2002). We note that some of the genes could be shared across multiple pathways (e.g. dsrA, dsrB, dsrC are all involved in both sulfur reduction and oxidation, (Yu et al. 2021)). Following the methods of (Bay et al. 2021), positive hits were considered those with an identity threshold >60% and a maximum e-value threshold of 10^{-10}. For each gene, abundances were normalized based on the abundance of 14 single copy genes calculated using the program SingleM v.13.2 (https://github.com/wwood/singlem) and are presented as normalized reads per million (RPM).

To make predictions about potential metabolic pathways that are enriched in the inland cone sediments as compared to the samples from the vegetated island, genes were grouped into larger functional categories based on (Yu et al. 2021; Imhoff et al. 2019; Leung and Greening 2021; Garber et al. 2020; Figueras et al. 2002). Abundances of these pathways in each of these samples were determined by summing the normalized RPM of all genes associated with a pathway or gene family. Pathways were considered enriched in the inland samples if the abundance was, on average, significantly higher in the inland samples than in the vegetated soils that pre-date the 2014 - 2015 eruption. Significance was determined using Mann-Whitney U tests as described above.

Recovery and Analysis of metagenomically assembled genomes (MAGs)

To determine the organisms that are responsible for the gene categories identified through the targeted gene analysis, metagenome-assembled genomes (MAGs) were recovered following methods described in Ortiz et al. (2021). In short, the sickle-trimmed sequences for 13 of the inland samples were assembled individually using metaSPAdes v.3.13.0 (Nurk et al. 2017) with default parameters.

Reads were mapped to assembled scaffolds using Bowtie2 v.2.4.5 (Langmead and Salzberg 2012), and abundance files were created and formatted with samtools v.1.14 (Danecek et al. 2021) and anvio v.7.1 (Eren et al. 2021; 2015). Scaffolds were then used to obtain genome-resolved bins with CONCOCT v.1.1.0 (Alneberg et al. 2014), MaxBin2 v.2.2.4 (Wu et al. 2016), and MetaBAT2 v.2.12.1 (Kang et al. 2019) and an optimized, dereplicated, non-redundant set of bins was identified and refined using The Dereplication, Aggregation, and Scoring Tool (DAS_Tool) v.1.1.2 (Sieber et al. 2018). The quality of the dereplicated bins was determined using CheckM v.1.1.3 1 (Parks et al. 2015). Following the suggestions outlined in Parks et al. (2015), bins that were greater than 70% complete and <10% contaminated were considered "high-quality" and were used for downstream analyses. Following these methods, we recovered 156 high-quality MAGs from the 13 samples. Taxonomy was assigned to each of these MAGs with the GTDBtk v.2.1.0 classify workflow (Chaumeil et al. 2020) based on the Genome Taxonomy Database (Parks et al. 2022), and open reading frames were predicted with Prodigal v.2.6.3 (Hyatt et al. 2010).

To identify which MAGs contain genes associated with the pathways we found to be elevated in abundance in samples from the inland cone (see above), the 48 genes associated with anoxygenic photosynthesis, carbon monoxide oxidation, hydrogen oxidation, sulfur disproportionation, and sulfur reduction were queried against the translated amino acid sequences of each MAG using the blastp function of DIAMOND v.2.0.11 with a query coverage of 80%, an identity threshold of 60%, and a maximum e-value threshold of 10^{-10} (Buchfink et al. 2015).

Supplemental Figure S4.1: Environmental and geochemical characteristics of the samples collected from across the island of HTHH. The pH of the inland cone sediments (n=13) was significantly lower than the pH of the beach sediments (n= 10) or the vegetated sediments (n=4) (Kruskal-Wallis one way analysis of variance, p <0.001). Total organic carbon in the vegetated samples was significantly higher than in the other sample categories (Kruskal-Wallis, p < 0.001) and sulfur content (%) was significantly higher in the cone samples than the vegetated samples (Mann-Whitney U, p = 0.033). Sediments from both cone and vegetated sediments had high concentrations of copper, vanadium, cobalt, and scandium, with these concentrations well above the range typically measured in 'typical', non-contaminated soils (indicated in red, taken from Flemming et al. (1989), Krishna and Govil (2007), Vukojević et al. (2019), Yang et al. (2017)).
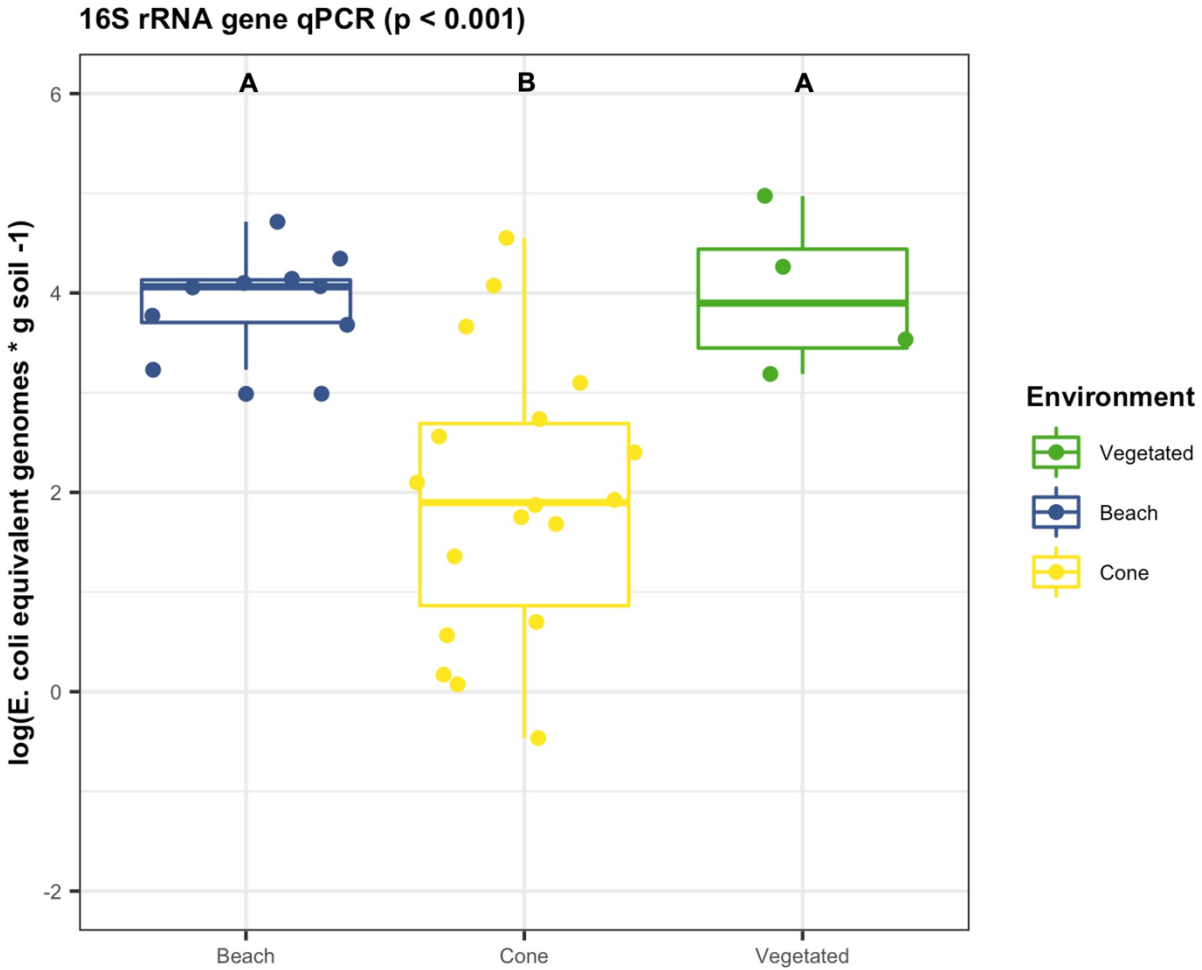
Supplemental Figure S4.2: Amount of prokaryotic (bacterial and archaeal) DNA found in the sediments of HTHH, as estimated by quantitative PCR (qPCR) analysis of the 16S rRNA gene. Cone sediments (n = 13) had significantly lower DNA concentrations than the beach sediments (n = 10) and vegetated sediments (n = 4) (Kruskal-Wallis one way analysis of variance, $p < 0.001$). Post hoc comparisons between groups, indicated with letters above each group, were performed using Dunn's Multiple Comparison Test.
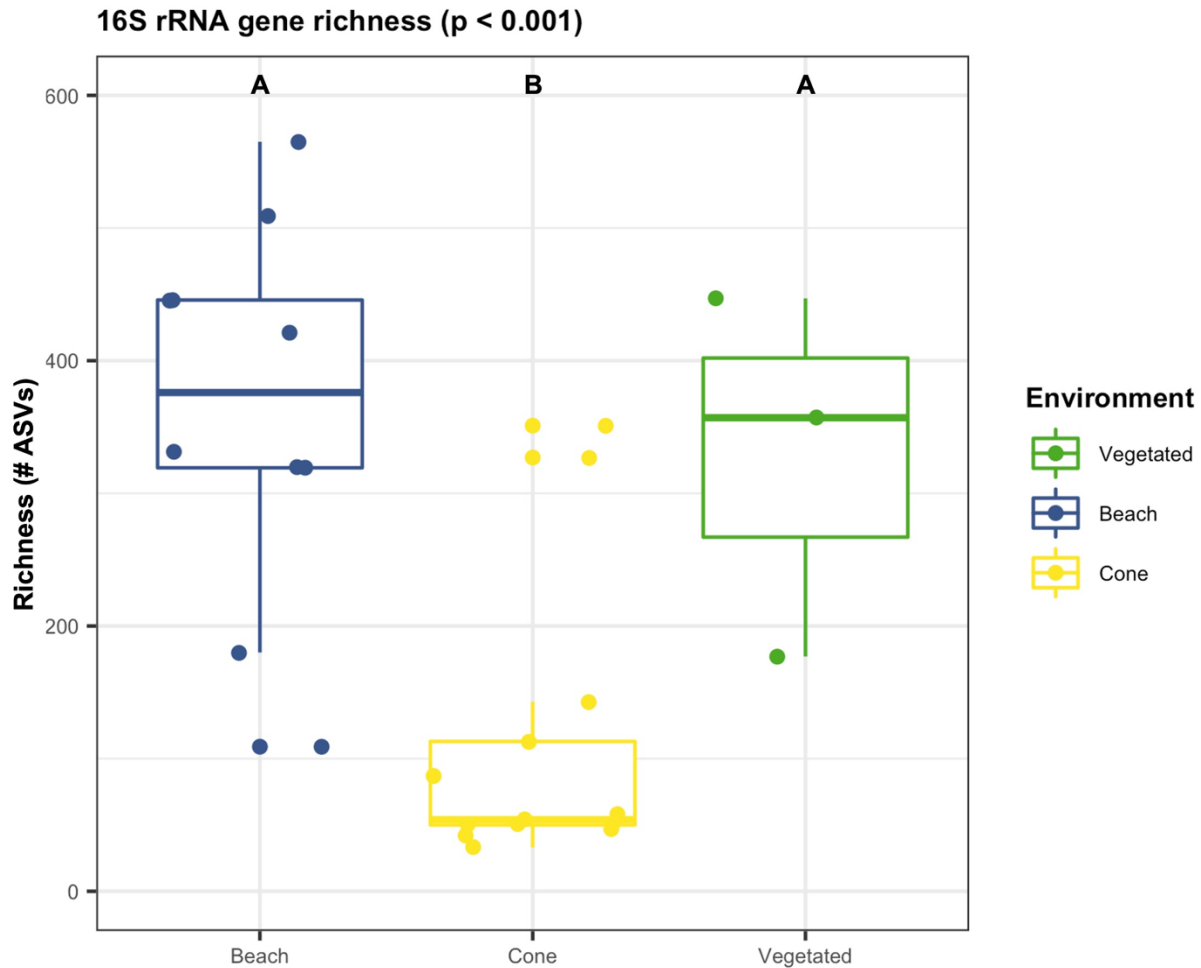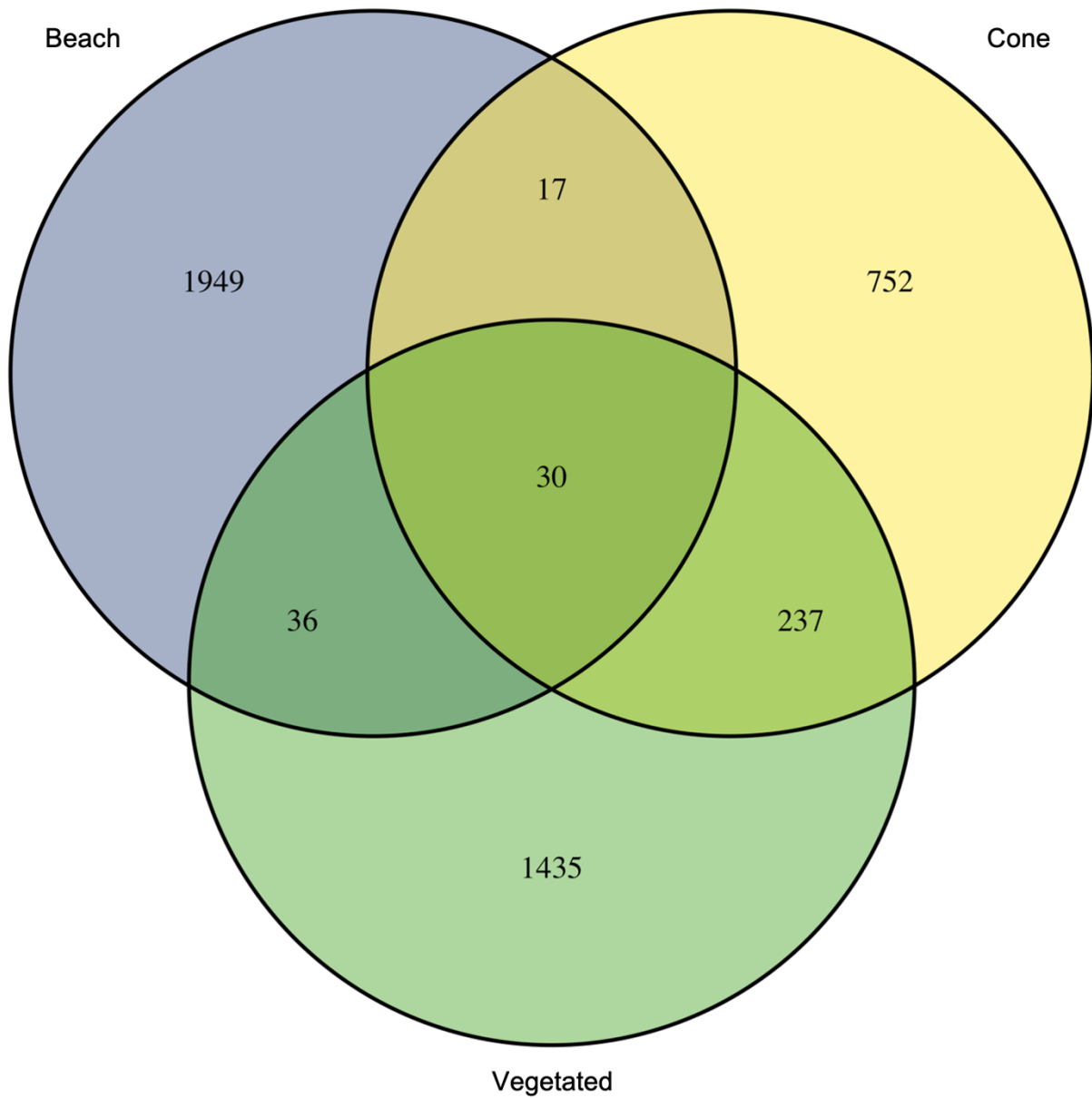
Supplemental Figure S4.3: Prokaryotic richness (number of distinct prokaryotic ASVs out of 10,000 reads per sample) of the beach, cone, and vegetated sediments. Cone sediments (n = 13) had significantly lower prokaryotic richness concentrations than the beach sediments (n = 10) and vegetated sediments (n = 4) (Kruskal-Wallis one way analysis of variance, $p < 0.001$). Post hoc comparisons between groups, indicated with letters above each group, were performed using Dunn's Multiple Comparison Test.
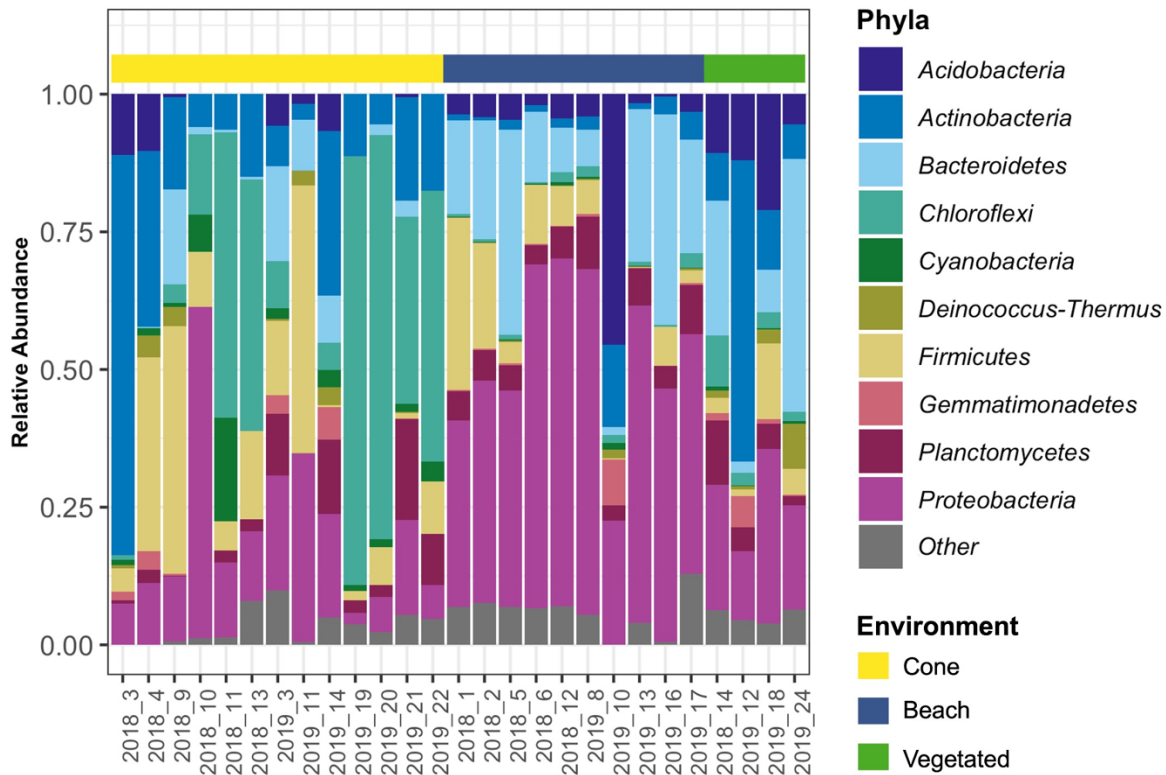
Supplemental Figure S4.4: Overview of bacterial and archaeal community composition across the island of HTHH. Shown here are the relative abundances of the top 10 most abundant phyla for each of the 27 samples for which 16S rRNA marker gene sequence data were obtained. Samples are grouped by environment type (Cone, Beach, Vegetated).
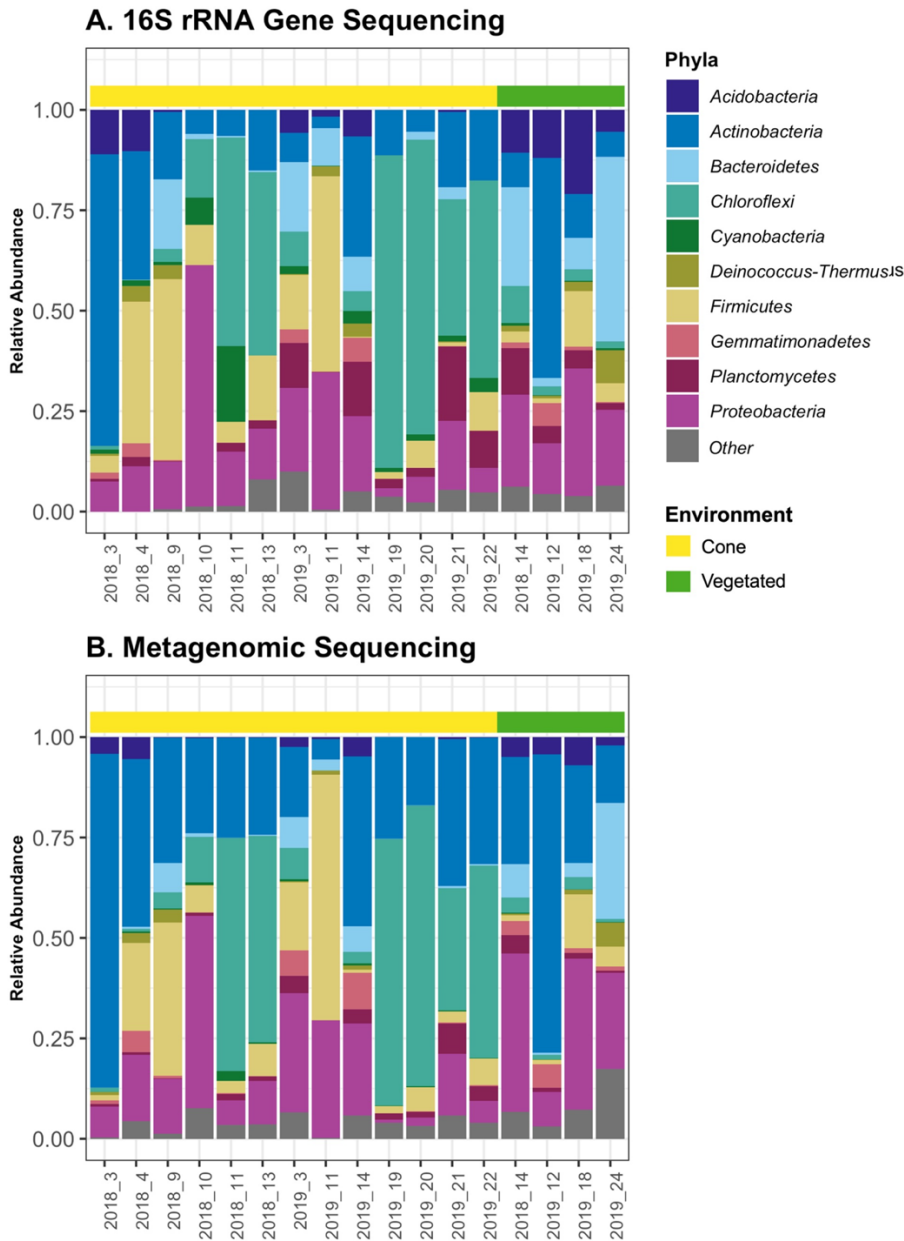
Supplemental Figure S4.5: Taxonomic information recovered from the amplicon and shotgun metagenomics sequencing efforts for the cone (n = 13) and vegetated sediments (n = 4). **A)** The relative abundances of the top 10 most abundant phyla as measured by amplicon sequencing of the 16S rRNA gene. **B)** Relative abundance of the top 10 most abundant phyla based on the 16S rRNA gene sequences recovered from the metagenomic sequencing data with phyloFlash (Gruber-Vodicka et al. 2020). We also note that the taxonomic composition of the prokaryotic communities, as inferred from the 16S rRNA amplicon data, was well-correlated with the taxonomic composition inferred from the metagenomic data (Mantel test of Bray-Curtis dissimilarity matrices, r = 0.90, p < 0.001).

Supplemental Figure S4.6: Shared prokaryotic taxa across the different sediment environments. Venn diagram was created with the ASV list compiled from the 16S rRNA gene sequencing effort of the beach (n = 10), cone (n = 13), and vegetated sediments (n = 4).
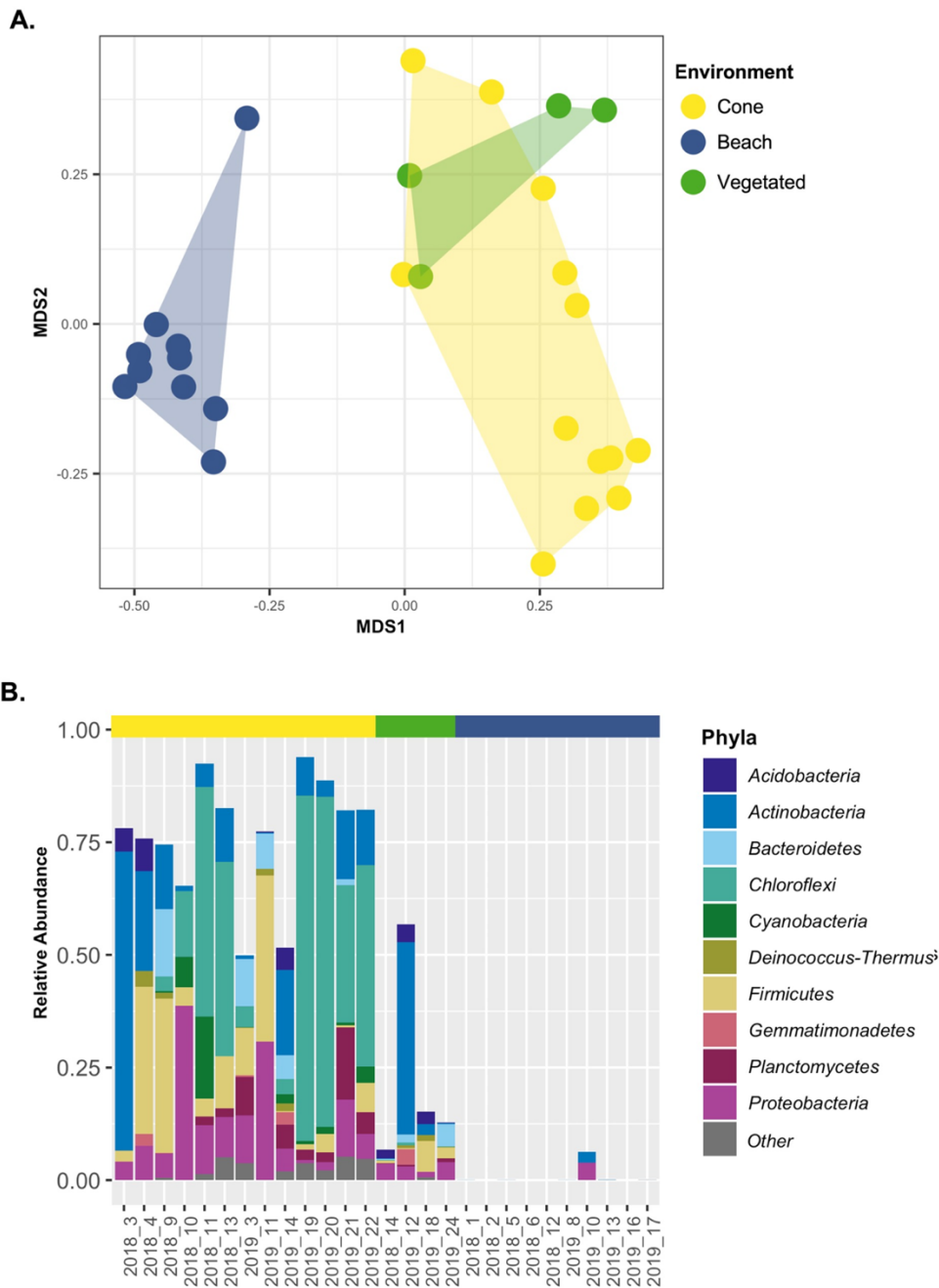
Supplemental Figure S4.7: Variation in microbial community composition across the island of HTHH. **A)** Non-metric multidimensional scaling (NMDS) plot based on the Jaccard distances in the 27 samples analyzed via 16S rRNA gene sequencing with colors indicating each of the three environment types sampled. **B)** The proportional abundances of the top 100 ASVs identified from the cone samples in each of the three other environments that were sampled. Samples are arranged by environment (indicated above the stacked bars) and ASVs are colored by phyla.

Supplemental Table S4.1: Potential sources of microbial colonizers. **A)** The abundance of the most ubiquitous bacterial families found in bird guts (as defined by Capunitan et al. (2020) in the inland cone samples. **B)** The abundance of the most common bacterial classes found in the marine environment (based on Sunagawa et al. (2015)) in the inland cone samples. **C)** Abundance of the dominant bacterial phylotypes found in soils. The percentage of total reads in the inland cone samples and the vegetated samples that are associated with the most common bacterial phylotypes found in soils, as defined by Delgado-Baquerizo et al. (2018).

A. Bird Gut/Fecal Microbiome

Bacterial family	% Reads
Enterobacteriaceae	1.21
Carnobacteriaceae	0
Enterococcaceae	0.01
Staphylococcaceae	0.09
Lactobacillaceae	0
Peptostreptococcaceae	0
Streptococcaceae	0

B. Marine

Bacterial order	% Reads
SAR11_clade	0
Oceanospirillales	0
Synechococcales	0
Flavobacteriales	0.02
Deferribacterales	0
Rhodospirillales	0
Alteromonadales	0.01
Rhodobacterales	<0.01
Acidimicrobiales	<0.01
Rickettsiales	<0.01

C. Vegetated

Phyla	Class	Cone (% of reads)	Vegetated (% of reads)
Proteobacteria	Alphaproteobacteria	5.40	18.96
	Betaproteobacteria	0	0
	Deltaproteobacteria	1.13	1.07
Actniobacteria	Actinobacteria	6.77	13.47
	Thermoleophilia	1.32	1.17
	Rubrobacteria	0.06	0.01
Acidobacteria	Chloracidobacteria	0	0
	Subgroup_6	0.22	1.53
	Soilibacteres	0	0
Planctomycetes	Phycisphaerae	1.15	1.49
	Planctomycetacia	3.47	5.25
Chloroflexi	Ellin6529	0	0
	Thermomicrobia	0	0
	Tk10	0.01	0.04
Verrucomicrobia	Spartobacteria	0	0
	Opitutae	0	0
	Pedosphaerae	0	0

Supplemental Table S4.2: Abundance of the genes related to 20 metabolic pathways in the cone (n = 13) and vegetated samples (n = 4). The abundances of genes are presented as the average normalized reads per million (RPM) of all samples in that group. The significance (p value) was determined by a Mann-Whitney U test. Genes are grouped based into pathways based on classifications given in Kato et al. (2018), Yu et al. (2021), Leung and Greening (2021), Garber et al. (2020), Figueras et al. (2002) with each gene appearing in just one group based on what function they have most often been described in. However, we note that many of these genes may be utilized by multiple pathways (ex: genes shared by sulfur reduction and sulfur oxidation).

Pathway	Cone Average normalized RPM	Vegetated Average normalized RPM	pval
Iron acquisition and transport	1.37 ± 0.4	1.11 ± 0.30	0.233
Iron oxidation	0.00 ± 0.00	0.00 ± 0.00	1.00
Iron reduction	0.00 ± 0.00	0.00 ± 0.00	1.00
Siderophore synthesis and transport	0.93 ± 0.9	1.97 ± 2.04	0.244
Magnetosome formation	0.05 ± 0.03	0.07 ± 0.01	0.104
Assimilatory sulfate reduction	4.18 ± 0.79	4.31 ± 0.96	0.549
Dissimilatory sulfur reduction and oxidation	0.47 ± 0.99	0.09 ± 0.04	0.079
Linkages between inorganic and organic sulfur transformation	6.91 ± 2.19	6.23 ± 1.16	0.233
Organic sulfur transformation	6.57 ± 2.67	6.14 ± 1.85	0.233
Sulfur metabolism (other)	6.37 ± 1.78	5.38 ± 1.74	0.233
SOX systems	0.36 ± 0.27	0.33 ± 0.17	0.412
Sulfur disproportionation	0.05 ± 0.06	0.01 ± 0.00	<0.001
Sulfur oxidation	0.69 ± 0.22	0.51 ± 0.22	0.36
Sulfur reduction	0.57 ± 0.21	0.28 ± 0.04	0.045
Carbon monoxide oxidation	0.27 ± 0.17	0.09 ± 0.06	0.003
Hydrogen oxidation	0.20 ± 0.15	0.05 ± 0.03	0.015
Hydrogenic fermentation	0.01 ± 0.02	0.02 ± 0.01	0.163
Methane oxidation	0.00 ± 0.00	0.00 ± 0.00	0.532
Bacteriochlorophyll-mediated anoxygenic photosynthesis	0.07 ± 0.06	0.02 ± 0.02	0.02
Chlorophyll photosynthesis	0.05 ± 0.06	0.01 ± 0.01	0.245

CHAPTER V APPENDIX

USING SOIL DEPTH GRADIENTS TO DETERMINE THE TAXONOMIC AND
GENOMIC ATTRIBUTES OF OLIGOTROPHIC SOIL BACTERIA

Materials and Methods

Sample collection and data acquisition

Details about the soil sampling process and soil characterization are
described in detail in Brewer et al. (2019). In brief, samples were collected at 10
different Critical Zone Observatories (CZO) across the United States of America.
Two soil profiles representative of distinct soil types found at each CZO site were
sampled in 10cm increments from the surface to 1m in depth or to saprolite when
excavation to 1m depth was difficult. In total, 185 samples were included for
downstream analysis used for downstream analysis. For all analysis described
subsequently, surface samples are any samples collected from 0-10 cm (39 samples)
while subsurface samples are samples collected from >10cm (129 samples).

DNA extraction and amplicon sequencing methods are also described in
Brewer et al. (2019). In summary, DNA was extracted from 0.25g of each sample
using the DNeasy PowerLyzer PowerSoil kit (Qiagen, Germantown, MD, USA).
Triplicate extractions were performed for each sample (0.75g total) and the
manufacturer's recommendations were followed until the stage when DNA was
eluted onto the spin filter; replicates were pooled at this point onto a single filter,
and extractions proceeded from this point as a single. Extracted DNA was amplified
in triplicate using the primer pair 515f/806r for sequencing the V4-V5 region of the
16S rRNA gene. Primers included the appropriate Illumina adapters and unique 12
- bp barcode sequences to permit multiplexed sequencing (Caporaso et al. 2012).
Amplicons were normalized using SequalPrep normalization plate kits (Thermo
Fisher Scientific, Waltham, MA), samples were pooled, and were sequencing at the

University of Colorado next-generation sequencing facility on an Illumina Miseq (Illumina, San Diego, CA, USA) using the V2 2 x 150 bp paired- end Illumina sequencing kit. Raw 16S rRNA gene sequencing data can be found on figshare at https://doi.org/10.6084/m9.figshare.4702711.

Taxonomic analysis via amplicon sequencing

The 16S rRNA gene sequences from the 185 samples were processed using the Dada2 pipeline v.1.26 with samples pooled (Callahan et al. 2016). Sequences were quality filtered and clustered into exact sequence variants (ASVs), with taxonomy determined using a naïve Bayesian classifier method (Q. Wang et al. 2007) trained against the SILVA reference database v.138 (Quast et al. 2013; Yilmaz et al. 2014). A minimum bootstrapping threshold required to return a taxonomic classification of 50% similarity was used for analysis. More details of the specific parameters used can be found at https://github.com/fiererlab/dada2_fiererlab. Prior to downstream analysis we removed samples that did not meet a threshold of having >10000 reads (7 samples) which left us with 178 samples. ASVs associated with chloroplast, mitochondria, and eukaryotes (785 ASVs) as well as those unassigned to the phylum level (613) were then removed. We then further paired down the ASV table based on abundance and ubiquity. ASVs with less than 50 reads across all the 178 samples were removed (21570 ASVs) and ASVs that were found in fewer than 5 profiles were also removed (6012 ASVs). A total of 12075 ASVs remained.

To determine which bacterial taxa are more likely to be found in the deeper soils versus the surface soils, we used Mann-Whitney nonparametric tests to compare the relative abundance of each ASV between the surface (39 samples) and subsurface (139 samples). The relative abundance for each ASV in each sample was calculated by dividing the number of reads assigned to that ASV by the total

number of reads for a sample remaining after the filtering described above. ASVs that were significantly more abundant in surface samples (1271 ASVs) are classified as surface-associated and therefore more-copiotrophic while those significantly more abundant in deeper samples (178 ASVs) are classified as subsurface-associated and considered to be more oligotrophic. 8658 ASVs showed no significant difference in abundance between the surface and subsurface environments. While we did find 13 archaeal ASVs to be classified as more oligotrophic, we focus just on bacteria.

The sequences of the 1462 ASVs that we identified as either being surface- or subsurface-associated (1271 ASVs and 178 ASVs respectfully) were searched against the Genome Taxonomy Database (GTDB) release 207 (Parks et al. 2022; Rinke et al. 2021) using VSEARCH v2.22.1 (--strand both --notrunclabels --iddef 0 -- id 0.97 --maxrejects 100 -- maxaccepts 100) (Rognes et al. 2016). If a single ASV matched to multiple GTDB genomes, the most complete genome with the lowest contamination was chosen as the reference. A total of 453 surface-associated ASVs and 66 subsurface-associated ASVs matched to reference genomes in GTDB. Before any additional analysis was performed reference genomes that matched to both categories (surface and subsurface) were removed (15 genomes) which gave us a total of 303 unique surface genomes and 40 unique subsurface genomes.

Genomic analyses of representative genomes

Relevant characteristics for the 343 genomes was gathered from the metadata associated with the GTDB reference database release 207 (Parks et al. 2022; Rinke et al. 2021). More specifically we used GTDB's information about the genome category (MAG vs isolate), predicted genome size, GC percentage, predicted small subunit (SSU) count, and taxonomy (based on SILVA reference database v.138 (Quast et al. 2013; Yilmaz et al. 2014)). For more information about how the

metadata were generated by GTDB, see details in Parks et al. (2018) and https://gtdb.ecogenomic.org/methods.

To estimate the predicted growth rate from the reference genomes, we used the tool gRodon2 on the genome scaffolds downloaded from GTDB following the authors recommendations for MAGs and genomes as detailed in Weissman et al. (2021) and https://github.com/jlw-ecoevo/gRodon. gRodon2 estimates maximal microbial growth rates from codon usage biases in highly expressed genes, an indicator of selection for rapid growth (Weissman et al. 2021; Vieira-Silva and Rocha 2010). We note that gRodon2 only provides a prediction of maximum potential growth rates (not actual growth rates) and the calculated values are just estimates useful for inferring broad patterns in maximum potential growth rates across genomes.

To determine the functional genes present in each reference genome, we used the blastp function of DIAMOND v.2.0.15 (-k 1 -e 10-10 --query-cover 90) (Buchfink et al. 2015) to annotate reads with the Database of Clusters of Orthologous Genes (COGs) ontology v.2020 (Galperin et al. 2021; 2015). For calculating the abundances of the COGs, COG categories, and COG groups listed in Table 5.1, we normalized the reads assigned to each of the 4877 individual COGs by dividing by the estimated genome size for each reference genomes and gene abundances are presented as reads per million base pairs. The abundances of COGs associated with different hypotheses were determined by summing the normalized gene abundances of each COG or COG category associated with that hypothesis (Supplemental Table S5.1). For the 25 COG categories, we also followed a method used by Weissman et al. (2021). Briefly, we calculated the proportion of COGs associated with each category by dividing the number of genes assigned to that category by the total number of COGs identified. If a COG was found to be assigned to multiple categories, it was counted in each. COG categories we identified as

enriched in one environment or the other had, on average, a greater proportion of COGs assigned to that category than the other environment.

To determine differences in the genomic characteristics, predicted growth rates, abundance of COGs, COG categories, and other functional properties we used Mann-Whitney nonparametric tests to compare the presence of these features between surface genomes (303 genomes) and subsurface genomes (40 genomes). Functional characteristics and categories that were significantly more abundant in surface soils are considered to be associated with more copiotroph-associated traits while those significantly more abundant in subsurface soils are considered to be oligotroph-associated characteristics.

Shotgun metagenomic analysis

Adapter trimmed and quality filtered shotgun metagenomic sequencing data from 73 of the samples described previously were gathered for this analysis. For more information about the generation of this data, which is publicly available on the MG-RAST metagenomics analysis server project ID: mgp80869 (Meyer et al. 2019; Wilke et al. 2016), see Brewer et al. (2019) for more information. The relative abundances and diversity of bacteria and archaea in the metagenomic samples were determined by extracting 16S rRNA gene reads from the metagenomic sequence data using phyloFlash v. 3.4 (Gruber-Vodicka et al. 2020). Read-based analyses on the trimmed and quality filtered data were performed using SqueezeMeta v.0.1.0 with the alternative analysis mode sqm_reads.pl script (Tamames and Puente-Sánchez 2019) which uses DIAMOND v.2.0.15 (Buchfink et al. 2015) to annotate reads with the COG ontology (Galperin et al. 2021; 2015). Three samples, which had fewer than <2000 reads assigned, were removed leaving a total of 70 samples with an average of 3.9 million annotated reads (range 496 thousand – 15 million reads). For each of these remaining samples, we removed assignments that were not to an

individual COG category or were assigned to ENOG categories. The reads assigned to each COG in each sample was normalized by dividing by the total number of remaining reads in that sample after this filtering, leaving COG abundances as reads per million base pairs. COG categories and COGs associated with specific functional hypotheses listed in Table 1 were summed as described previously for the annotated reference genomes.

To determine whether we see the same patterns in our metagenomic data that we found in our genome-based analysis described previously, comparisons were performed as described previously with the genomes, with significant differences between the surface metagenomes (18 metagenomes) and subsurface metagenomes (52 metagenomes) determined with Mann-Whitney nonparametric tests.

Plotting and additional analysis in R

Supporting analyses were performed in R v.4.2.2 (R Core Team, 2017). Statistical tests were performed using the base R functions 'wilcox.test' and the packages 'rstatix.' Plotting was performed using the R packages 'ggplot2' and 'cowplot.' ASV table filtering, stacked bar plots of relative abundance, and other ASV-based analysis was performed using the R packages 'mctoolsr' (https://github.com/leffj/mctoolsr/).
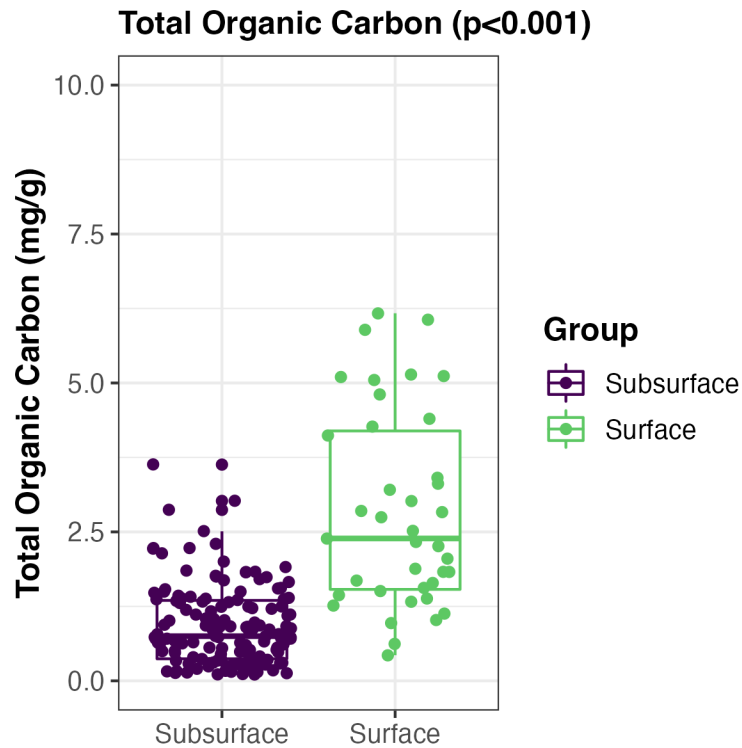
Supplemental Figure S5.1: Total organic carbon in the surface (n=39) and subsurface (139 ASVs) and sediments. Surface samples were collected from 0 – 10cm depth while subsurface samples were collected >10cm in depth. Subsurface samples had significantly less organic carbon than the surface samples (Mann-Whitney U, p <0.001).
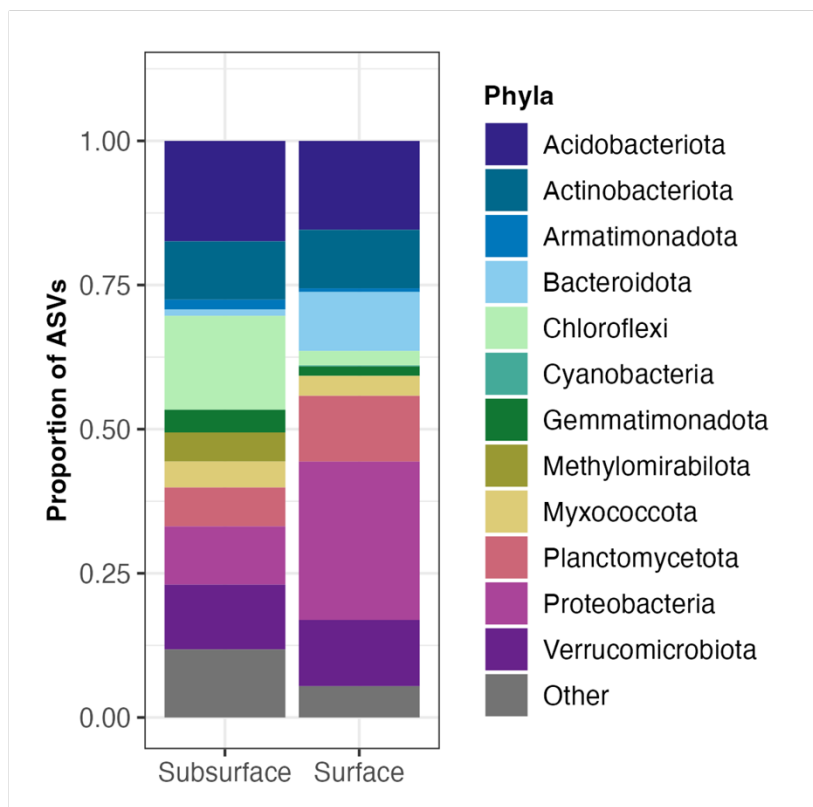
Supplemental Figure S5.2: Overview of the soil bacteria associated with the surface (1271 ASVs) and subsurface (191 ASVs) environments. The size of the bar represents the proportion of the total number of ASVs assigned to the phylum. More specific taxonomic information can be found in Figure 5.1.

	Subsurface	Surface
Acidobacteriaceae (Subgroup 1)	0	17
Acidobacteriales;NA	0	1
Pyrinomonadaceae	1	1
Solibacteraceae	0	2
Subgroup 13;NA	2	0
Subgroup 2;NA	2	2
Subgroup 22;NA	1	0
Subgroup 7;NA	2	0
Vicinamibacterales;NA	1	4
Frankiaceae	1	2
Geodermatophilaceae	0	3
Kineosporiaceae	1	1
Micromonosporaceae	1	5
Mycobacteriaceae	0	12
Nocardioidaceae	0	5
Pseudonocardiaceae	0	6
Solirubrobacteraceae	0	5
Streptomycetaceae	0	4
Streptosporangiaceae	0	1
Chitinophagaceae	0	23
Kryptoniales;NA	1	0
Sphingobacteriaceae	0	10
AD3;NA;NA	2	0
Chloroflexi;NA;NA;NA	3	0
Dehalococcoidia;NA	1	0
Desulfobacterota;NA;NA;NA	1	0
FCPU426;NA;NA;NA	1	0
Paenibacillaceae	1	1
Gemmatimonadaceae	3	1
Latescibacterota;NA;NA;NA	2	0
Rokubacteriales;NA	1	0
Anaeromyxobacteraceae	1	0
Nitrospiraceae	2	0
Acetobacteraceae	0	3
Beijerinckiaceae	1	3
Burkholderiaceae	0	9
Burkholderiales;NA	1	4
Caulobacteraceae	0	6
Comamonadaceae	2	12
Devosiaceae	0	2
Dongiaceae	0	1
Hyphomicrobiaceae	0	1
Micropepsaceae	0	4
Nitrosomonadaceae	1	2
Oxalobacteraceae	0	5
Reyranellaceae	0	5
Rhizobiaceae	1	5
Rhizobiales;NA	0	2
Sphingomonadaceae	1	14
Xanthobacteraceae	0	19
RCP2-54;NA;NA;NA	1	2
Chthoniobacteraceae	1	9
Opitutaceae	0	8
Pedosphaeraceae	1	5
Other	0	77

Supplemental Figure S5.3: Heatmap displaying the number of genomes identified as associated with either the surface (303) or subsurface (40). Genomes associated with the surface are considered more copiotrophic while those associated with the subsurface are considered more oligotrophic. Genomes are grouped by family and are ordered alphabetically by phyla. If no family is available, the most specific taxonomy is displayed.
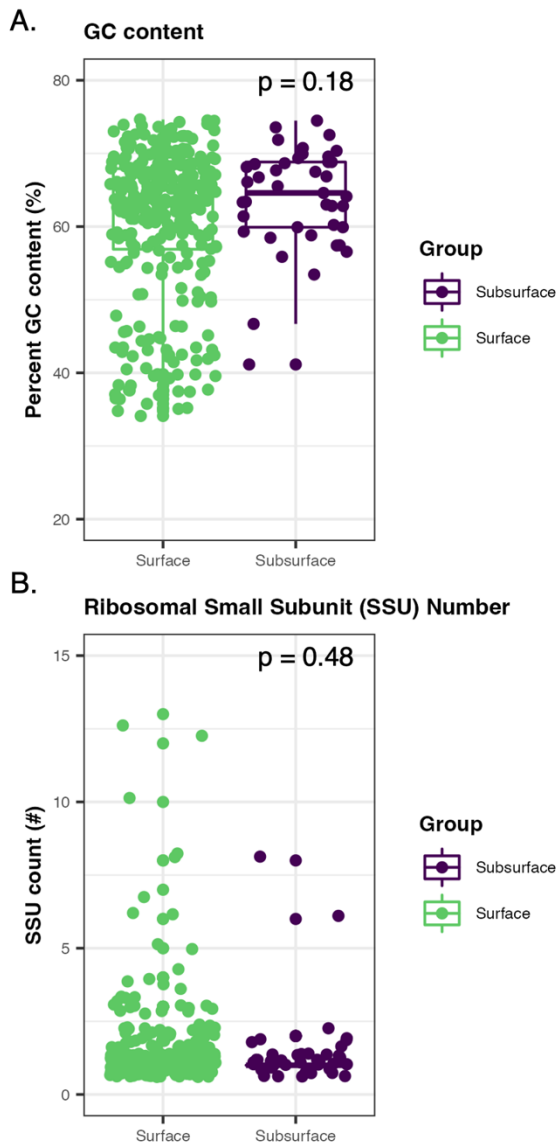
Supplemental Figure S5.4: Additional genomic characteristics of the surface (n=303) and subsurface (n=40) genomes. **A)** There was no significant difference in the percent GC content between the surface and subsurface (Mann-Whitney U, p = 0.18). **B)** There was no significant difference in the ribosomal small subunit (SSU) number between the surface and subsurface (Mann-Whitney U, p = 0.48).
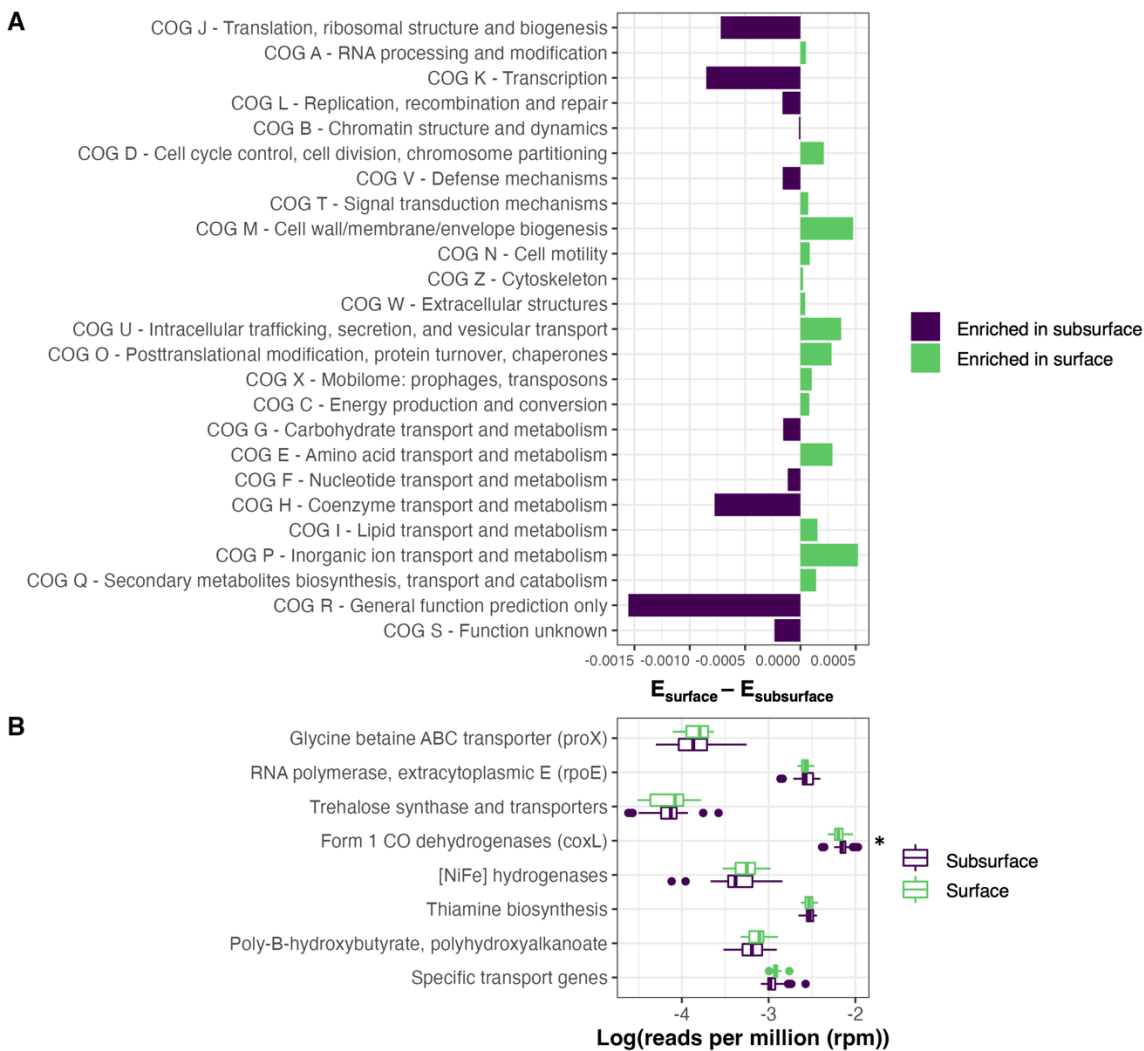
Supplemental Figure S5.5: Functional comparison of 33 gene categories across the subsurface metagenomes (n=52) and surface metagenomes (n=18). **A)** The difference in the average proportion of genes in the surface metagenomes ($E_{surface}$) and the average proportion of genes in the subsurface ($E_{subsurface}$). **B)** Abundance of genes associated with the hypotheses outlined in Table 5.1. Significant differences in gene abundances between the two groups (Mann-Whitney U) are starred.

Supplemental Table S5.1: The COGS and COG categories associated with each hypothesis outlined in Table 5.1.

Gene or Gene Category	Associated COGs	Reference
Amino acid transport and metabolism	COG category E	Qin et al. mBio. (2019)
Chemotaxis and motility	COG category N	Roller et al. (2016), Lauro et al. (2009)
Lipid transport and metabolism	COG category I	Lauro et al. (2009)
Secondary metabolite biosynthesis, transport, metabolism	COG category Q	Lauro et al. (2009)
Defense mechanisms	COG category V	Dutta and Paul (2012)
Transcription	COG category K	Dutta and Paul (2012)
Signal transduction	COG category T	Dutta and Paul (2012)
Cellular replication, recombination, repair	COG category L	Koch (2001)
Glycine betaine ABC transporter (ProX)	COG2113, COG3760, COG2113, COG3760	Noell and Giovannoni (2019)
RNA polymerase, extracytoplasmic E (rpoE)	COG1595, COG3343, COG5503	Su et al. (2015)
Trehalose synthase and transporter	COG3281, COG4813	Bird et al. (2019)
Form 1 CO dehydrogenases (coxL)	COG1529	Cordero et al. (2019)
[NiFe] hydrogenases	COG0374, COG0680, COG1740, COG1969, COG3260, COG3261, COG3262	Greening et al. (2015), (2016),
Thiamine biosynthesis	COG2145, COG0301, COG0351, COG0352, COG0422, COG0476, COG0611, COG1060, COG1564, COG2022, COG2104	Roller et al. (2016)
Poly-B-hydroxybutyrate, polyhydroxyalkanoate	COG3243, COG3937, COG5394, COG5490	Poindexter (1981)