Microbial Community Composition Reflects Subduction Geochemistry in the Costa Rica Convergent Margin

Katherine Marie Fullerton

University of Tennessee, kfulle16@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_gradthes

Recommended Citation

https://trace.tennessee.edu/utk_gradthes/5439
Microbial Community Composition Reflects Subduction Geochemistry in the Costa Rica Convergent Margin

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Katherine Marie Fullerton
May 2019
Dedication

To my cat

Pokuro Kurosuke “Poe” Fullerton

My Yankee parents

Mary & Paul Robb

My Tennessee Family

Amber Roberts Shrewsberry

Keith & Janet Roberts
Acknowledgements

"Alone we can do so little, together we can do so much." --Helen Keller

First and foremost I must thank my major advisor, Dr. Karen G. Lloyd, for her unwavering support and encouragement through all the ups and down of my graduate career. She provided me with several once-in-a-lifetime opportunities to experience the world (literally) and develop myself as a field scientist.

I am forever indebted to those inhabiting the service elevator office of SERF, who were my steadfast friends through thick and thin: April Armes, Katie Sipes, and Dr. Joy Buongiorno-Altom.

The entire DCO Biology Meets Subduction Team, for cultivating my interdisciplinary scientific approach and exposing me to the wonders of the geological world.

The Deep Carbon Observatory and The Census of Deep Life at the Marine Biology Lab for making it possible to obtain terabytes of sequencing data for both Biology Meets Subduction field expeditions.

Veronica Brown of the UTK Next-Generation Sequencing Center for technical support with SAG generation and moral support.

Dr. Andrew Steen for encouraging reproducible data analysis using R, with significant help from Google and Stack Overflow for dependable debugging and coding help.

The Sloan Foundation, Deep Carbon Observatory, the Associate Alumnae of Douglass College, and the Graduate School for providing significant or supplemental funding for my research and post-grad education.
Abstract

Biogeochemical processes drive the cycling of nutrients on Earth, both in surface and subsurface environments, with subduction representing a critical link between shallow and deep geochemical cycles. In subduction zones, sediments and subseafloor basement basalts are either transported into the mantle or are “recycled” back to the surface through fluid fluxes, volatile degassing, or magmatic events. Transformations of these subducted sediments occur in response to a variety of abiotic, thermogenic, or biological processes on both geochemical and biological time scales. However, the biological component to biogeochemical cycling of volatiles in subduction zones has largely been overlooked in the past. Recent advancements in high throughput sequencing have opened the door for the systematic study of the diversity of these communities. When integrated with available geochemical data, one can start to gain a better understanding of the complex interactions between the biotic and abiotic processes driving these cycles. To investigate potential interactions between abundant taxa and their environment, ordination analysis was applied to a large, biogeochemical dataset from 24 geochemically diverse hydrogeological sites in the volcanic region of Costa Rica, including 16S rRNA gene libraries containing >56,000 and >27,000 unique bacterial and archaeal amplicon sequence variants (ASV) sequences, respectively. Fluids show low input of photosynthesis-related genes or carbon with photosynthetic isotope signals, indicating that fluid microbial communities largely reflect shallow subsurface geochemical processes. A pH gradient is the primary driver of across-arc variation between the Outer Forearc and Forearc/Arc, while changes in temperature corresponding to changes in offshore bathymetry define along-arc variation. Based on these two geochemical gradients, we propose a 4 region model of microbial composition: 1) Northern Forearc/Arc – acidic and thermophilic; 2) Central Forearc/Arc – acidic and thermophilic, but less so than its northern counterpart; 3) Northern Outer Forearc – alkaline; and 4) Central Outer Forearc – alkaline. Regional niche separation of primarily chemolithotrophic microbial taxa reflect local subduction geochemistry, such as the acidophilic Sulfurihydrogenibium sp. dominating high temperature acidic springs in the Forearc/Arc. This research establishes the microbial responses to regional-scale geochemistry in a geothermal system, and shows that the effects of tectonic-scale processes can be observed in microbial community compositions.
Table of Contents

Chapter 1 – Introduction ........................................................................................................... 1
Statement of Contribution ......................................................................................................... 1
The Terrestrial Subsurface Biosphere ...................................................................................... 1
Subduction: Linking Subsurface and Surface Processes .......................................................... 1
Biological Communities in Continental Geothermal Systems ................................................. 2
Costa Rica Geological Setting ................................................................................................. 3
Previous Microbiological Studies of Continental Costa Rica ................................................. 3
Biology Meets Subduction Field Expedition ........................................................................... 4
Carbon Isotope Composition of Geothermal Fluids in Costa Rica are Consistent with
Chemolithotrophy .................................................................................................................... 4
Moving Forward: Integration of ‘Omics and Geochemical Datasets ................................ ...... 5
Thesis Aims .............................................................................................................................. 5

Chapter 2 – Materials and Methods ....................................................................................... 7
Biological Sample Collection .................................................................................................. 7
Flow Cytometry ........................................................................................................................ 7
DNA Extraction and Sequencing ............................................................................................ 7
Analysis of 16S rRNA Gene Amplicon Data ....................................................................... 8

Chapter 3 – Results & Discussion ......................................................................................... 9
Site Description & Classification ............................................................................................. 9
Distribution of Biomass in Hot Spring Fluids .......................................................................... 9
Pre-Processing & Contamination Screening of 16S rRNA Gene Amplicon Libraries .......... 10
Hot Spring Fluids Show Low Photosynthetic Potential .......................................................... 10
Paired Fluid-Sediment Samples are Qualitatively Similar but Quantitatively Different ....... 11
Four Microbial Biogeographical Regions Defined by Across and Along Arc Gradients ......... 11
  Across Arc Variation in Microbial Communities Reflects Subduction Progression ............ 11
  Seafloor Bathymetry Drives Along Arc Variation in Microbial Communities .................. 11
  Four Region Model of Microbial Community Composition ............................................... 12
Putative Chemolithotrophic Bacteria Dominate in Geothermal Fluids ............................. 12
Sulfur and Iron Cycling Microbial Clades Show Niche Separation .................................... 13
  Iron Oxidizing Bacteria are Limited to the Central Costa Rica Region ............................ 13
  Diverse Sulfur Cycling Bacteria in the Costa Rica Volcanic Arc System ......................... 13
Arc Archaeal Hot Spring Communities are Dominated by Either Nitrososphaeria or
Bathyarchaeota ....................................................................................................................... 14

References ............................................................................................................................... 18

Appendices ............................................................................................................................... 26
Appendix I – R Code ................................................................................................................ 27
  Carbon Isotope Figure 3 Generation .................................................................................... 28
  Chloroplast & Photosynthetic Pigment Analysis & Figures 5 & 6 ..................................... 28
  Cell Abundance Analysis & Figures .................................................................................. 29
  ASV Classification ............................................................................................................... 32
    Brazelton Lab Python Scripts ......................................................................................... 32
    Mothur Code ................................................................................................................... 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Normalization</td>
<td>32</td>
</tr>
<tr>
<td>Bacteria</td>
<td>32</td>
</tr>
<tr>
<td>Archaea</td>
<td>37</td>
</tr>
<tr>
<td>Ordination Analysis &amp; Figures</td>
<td>40</td>
</tr>
<tr>
<td>Bacteria</td>
<td>40</td>
</tr>
<tr>
<td>Archaea</td>
<td>47</td>
</tr>
<tr>
<td>Archaea Heatmap Analysis &amp; Figure 15</td>
<td>52</td>
</tr>
<tr>
<td>Adonis Statistics</td>
<td>54</td>
</tr>
<tr>
<td>Bacteria</td>
<td>54</td>
</tr>
<tr>
<td>Archaea</td>
<td>56</td>
</tr>
<tr>
<td>Aqueous Geochemistry Analysis &amp; Figures</td>
<td>58</td>
</tr>
<tr>
<td>Iron</td>
<td>58</td>
</tr>
<tr>
<td>Ammonium</td>
<td>59</td>
</tr>
<tr>
<td>Appendix II – Single Cell Amplified Genome Workflow &amp; Results</td>
<td>61</td>
</tr>
<tr>
<td>Methods</td>
<td>62</td>
</tr>
<tr>
<td>Sample Collection &amp; Screening</td>
<td>62</td>
</tr>
<tr>
<td>Cell Sorting &amp; Whole Genome Amplification at Bigelow</td>
<td>62</td>
</tr>
<tr>
<td>Bioinformatic Workflow</td>
<td>62</td>
</tr>
<tr>
<td>Results</td>
<td>62</td>
</tr>
<tr>
<td>Appendix III – Tables and Figures</td>
<td>64</td>
</tr>
<tr>
<td>Vita</td>
<td>89</td>
</tr>
</tbody>
</table>
List of Tables

Table 1 Site Classification and Geochemical Characteristics ................................................................. 65
Table 2 Summary Statistics for Amplicon Libraries Post-Processing ......................................................... 67
Table 3 Genus Richness of Bacterial and Archaeal 16S rRNA Gene Amplicon Libraries ......................... 68
Table 4 Single Cell Genome Summary Statistics ...................................................................................... 69
Table 5 SILVA (v132) Classification of SAGs via 16S rRNA Gene Nearest Neighbor Analysis .............. 70
List of Figures

Figure 1 Cartoon Representation of a Terrestrial Geothermal Spring.................................71
Figure 2 Map of Costa Rica and Northern Panama. Google..................................................72
Figure 3 Carbon concentration and δ13C of CO2, DIC, DOC, and TOC Across Costa Rica Arc........73
Figure 4 Across and Along Arc Variation in Cell Abundance................................................74
Figure 5 Relative Abundance of Chloroplast Reads in Bacterial 16S Amplicon Libraries............75
Figure 6 Total Photosynthetic Pigments in Hot Spring Sediments........................................76
Figure 7 NMDS Ordination of UniFrac Distance Measure of Bacterial Sediments vs Fluids........77
Figure 8 NMDS Ordination of UniFrac Distance Measure of Archaeal Sediments vs Fluids........78
Figure 9 NMDS Ordination of UniFrac Distance Measures with Environmental Vectors..........79
Figure 10 NMDS Ordination of Archaeal UniFrac Distance Measures with Environmental Vectors...80
Figure 11 NMDS with Species Points of Top 10 Bacterial ASVs per Fluid Sample.................81
Figure 12 Iron Concentration Across the Costa Rica Volcanic Arc.......................................82
Figure 13 Percent of Total Iron Released Upon Acidification.............................................83
Figure 14 Abundance of Nitrosphaeria vs Ammonium Concentration..................................84
Figure 15 Order Level Heatmap of Archaeal Community Composition...............................85
Figure 16 Results of SAG Generation from Espabel (EP) Sediments....................................86
Figure 17 Results of SAG Generation from Quebrada Naranja (QN) Sediment Sample............87
Figure 18 Neighbor Joining Tree of SAG 16S rRNA Genes with SILVA Nearest-Neighbors........88
Chapter 1 – Introduction

Statement of Contribution

This thesis represents a portion of a larger collaborative project that generated of a large biogeochemical dataset from several collaborators. Within this thesis, aqueous geochemistry measurements, including ammonium and iron concentrations presented in Figure 12, Figure 13, and Figure 14, were performed by Dr. Mustafa Yucel at the Middle East Technical University (Mersin, Turkey). Additionally, concentrations of total photosynthetic pigments, included in Figure 6, were measured by Dr. Elena Manini at Institute of Marine Science – National Research Council (Ancona, Italy). All other data collection, processing, and analysis were performed by the author.

The Terrestrial Subsurface Biosphere

Early studies in the 1990’s estimated that that microbial abundance within the first 5km of the Earth’s continental subsurface to be 3.8-6x10^30 cells, but with an influx of studies of subsurface microbes in the past decade, estimates have been refined to a global subsurface biosphere estimate of 2-6x10^29 cells (1–3). Largely decoupled from photosynthetically-derived organic carbon, subsurface microbial communities are dependent on a productive autotrophic community, including sulfate reducers and methanogens, or detrital carbon from elsewhere, for bioavailable carbon (4, 5). Additionally, without the presence of highly energetic phototrophic and oxygen-dependent metabolisms, energy requirements of subsurface microbes becomes limited to chemolithotrophic or heterotrophic processes dependent on the geochemistry of the local environment (5, 6). However, subsurface rock ecosystems are highly heterogeneous and to date no reliable geochemical predictors of microbial community compositions in the continental subsurface have been identified, though community composition has been correlated with sample lithology (2). Due in large part to the logistical challenges of accessing samples, biogeochemical cycling in the vast terrestrial subsurface is poorly constrained, including our understanding of carbon biogeochemistry (2, 5).

Subduction: Linking Subsurface and Surface Processes

Biogeochemical processes drive the cycling of nutrients on Earth, both in surface and subsurface environments, with significant implications for the release and sequestration of volatiles (7). Subduction, the geological process where two tectonic plates converge forcing one plate to slide beneath the other, represents a critical link between shallow and deep geochemical cycles. During this process, sediments are transported into the mantle and are “recycled” back to the surface through fluid fluxes, volatile degassing, or magmatic events (8, 9). Regional heterogeneity in the “recycled” materials can be driven by two major trends. The first is how far the subduction has progressed across the arc, or the distance from the trench where the downgoing slab first begins to subduct under the overriding plate. The chemistry of the surface-expressing fluids should reflect the changing chemical constituents that evolve from the slab as it traverses different pressure and temperature regimes on its downward journey (10). The second factor that may dictate biogeochemistry of the recycled materials is the nature of the downgoing slab, since the presence of seamounts, or the dip angle of the subducting slab may change along the arc and will greatly affect hydration, chemical alteration, and physical properties of the slab (11).

Subduction zones can be broken up into several distinct regions based on their distance from the trench (Figure 1): the forearc (which may contain an Outer Forearc region), the volcanic arc, and the
backarc basin. In the forearc, the region between the oceanic trench and the associated volcanic arc, hydration of subducted ocean crust leads to the formation of serpentine (hydrated) minerals and hydrogen through a process known as serpentinization (12). Serpentinization drives the formation of molecular hydrogen causing a drastic increase in pH, resulting in the forearc’s characteristic alkaline conditions. This hydrogen production has also been shown to fuel microbial life in marine subsurface and shallow terrestrial serpentinizing systems (13–15). Additionally, serpentinizing areas are characterized by low dissolved inorganic carbon (DIC), due to the precipitation of carbonates at high pH, and enrichment with methane (16). Geochemistry in the arc region is tightly tied to the regional volcanic gas emissions and activity. While water vapor makes up the largest fraction of gases emissions within the arc, more acidic volatiles such as CO$_2$, SO$_2$, H$_2$S, and hydrogen halides increase in abundance moving from the trench to the volcanic arc itself (17).

Transformations of these subducted sediments will occur on both geochemical and biological time scales in response to a variety of abiotic, thermogenic, or biological processes. However, the biological component to biogeochemical cycling in subduction zones has largely been overlooked in the past. Recent advancements in high throughput sequencing have opened the door for the systematic study of the diversity and function of these communities. When integrated with the extensive pool of available geochemical data, one can start to gain a better understanding of the impact of these microbial communities on global biogeochemical cycling and the complex interactions between the biotic and abiotic processes driving these cycles.

Biological Communities in Continental Geothermal Systems

The geochemical composition of a hydrothermal system is directly related to the composition of the mantle and bedrock beneath it and the processes that occur as fluids travel through these layers (10). Much like hydrothermal vents on the seafloor act as a window to the marine subsurface biosphere, caves, boreholes, and hot springs can be used to learn about the continental subsurface biosphere (3). The physical and geochemical characteristics of geothermal fluids, including pH, temperature, and ionic composition, will have a significant impact on energy availability in the subsurface, which in turn will impart constraints on the microbial communities that can inhabit these environments (6). Limited amounts of photosynthetically derived carbon are available to subsurface microbes due to burial processes, but many microbial communities depend on autotrophic carbon fixation to provide bioavailable carbon (7). Geochemistry in subsurface fluids can be highly variable making it difficult to find ubiquitous trends applicable to all subsurface environment, but it is precisely this geochemical diversity that provides conditions that are favorable for chemolithoautotrophic microbes (18). Ascending fluids across the arc extract the essential nutrients required for life from host rocks, and mixing of these reduced subsurface derived nutrients with more oxidized species from the surface (Figure 1) provides the redox conditions necessary to provide energy for chemolithotrophic microbes (19).

A study of over 2,000 genomes from suboxic aquifer groundwater found that 26-36% of the genomes contained pathways for the cycling of carbon monoxide, hydrogen, or reduced sulfur species, implying a close link between these biochemical cycles and subsurface metabolism (20). Serpentinizing ecosystems, including ophiolites and serpentinizing springs, appear to share several universal phylotypes, including Clostridia, Bacteroidetes, Betaproteobacteria (specifically the genus Hydrogenphaga), and Methanobacterium, supporting the ideas of a core serpentinizing microbiome in the subsurface (21). Other representative organisms include the Nitrospirae, Chloroflexi, and Gallionellaceae, many of which utilize chemolithoautotrophic lifestyles (18). Bacterial OTU sequences belonging to the Chlorobi and Nitrospirae phyla in deep granitic groundwater, sedimentary aquifers, and hot springs further suggesting the presence of a ubiquitous deep subsurface biosphere (22).
and geothermal spring environments tend to be rich in sulfur and iron compounds that once oxidized generate acidic byproducts \((23, 24)\). Therefore, these environmental niches tend to be inhabited by acidophilic organisms involved in sulfur and iron cycling, such as the bacterial taxa *Acidithiobacillus*, *Leptospirillum*, *Sulfobacillus*, and *Aquificae* and the Archaeal *Ferroplasma* \((23, 25)\).

**Costa Rica Geological Setting**

Off the west coast of Central America along the Middle America Trench (MAT), the oceanic Cocos plate subducts under the Caribbean plate at a rate of 70-90mm/yr \((26)\). Shallow plate geometry \((35-65^\circ)\) of the subduction zone in Costa Rica helps promote dehydration of the slab prior to it reaching the magma generation zone, allowing for the release of large fluxes of organic carbon and reduced chemical species into the overlying plate, much of which is subaerial, and therefore easily accessible for sampling \((27, 28)\). Sediments in the subducted plate in CRVA contain a significant fraction of organic material, and as a result volcanic gases in the Costa Rica arc are particularly CO\(_2\)-rich, providing an inorganic source of carbon to microbial communities influenced by local geothermal fluids \((9)\). Costa Rica has a two well-defined Outer Forearc peninsulas: the Nicoya Peninsula in the north and the Osa Peninsula in the south.

There are three separate tectonic boundaries off Costa Rica (Figure 2) - the Triple Junction, the Fisher Seamount, and the Quepos Plateau - that roughly define four separate along-arc regions \((29)\). The first of these three boundaries, the Triple Junction, separates East Pacific Rise (EPR) lithology from Cocos-Nazca Spreading center (CNS) lithology within the Cocos Plate, which will furthermore be referred to as the EPR-CNS boundary. Shallow hydrothermal circulation within the EPR lithology, a result of high basement permeability, limits crustal heat loss at the seafloor, whereas the CNS lithology limits advective heat extraction and shows high heat flow \((30)\). The CNS lithology can further be divided into two distinct bathymetries, rough and smooth, by the Fisher Seamount, which will be referred to as the rough-smooth boundary. The older CNS-1 segment is proximal to the EPR crust and is characterized by smooth bathymetry, while the younger CNS-2 segment to the south is rough and riddled with seamounts. The presence of numerous seamounts in the CNS-2 segment facilitates exchange of fluid and heat between the crust and the ocean, further increasing the heat flow \((30, 31)\). Finally, the Quepos Plateau marks a transition to a much thicker upper crust that is characteristic of Cocos Ridge crust \((32)\).

There are two active volcanic mountain ranges in Costa Rica: the Cordillera de Guanacaste, located in Northern Costa Rica near Nicaragua, and the Cordillera Central, located in central Costa Rica. These two ranges are separated by an 80km gap with no active volcanic activity, that is only interrupted by Arenal Volcano \((33)\). This gap is the result of a Northward shift in volcanism that was a consequence of a shift in the subducting Cocos Plate to a thicker, more seamount-ridden bathymetry resulting in shallower subduction \((34, 35)\). The Cordillera de Guanacaste range is characterized by a shared, contagious geothermal reservoir which promotes lateral transport and mixing of deep, hydrothermal fluids and meteoric waters across the region \((36–38)\). This is in contrast to the Cordillera Central volcanic range, where the regional hydrothermal system is smaller and more isolated to individual volcanoes with little to no conductive heat flow between reservoirs, evidenced by a high magmatic heat flow signal \((39)\).

**Previous Microbiological Studies of Continental Costa Rica**

Extensive resources have been dedicated to studying the geological and geochemical processes taking place across CRVA. Several Ocean Drilling Program (ODP) and Integrated Ocean Drilling Program (IODP) expeditions have characterized the hydrogeology and geochemistry of the subducting marine sediments, the upper plate, and the offshore trench. Scientists from El Observatorio Vulcanológico y Sismológico de Costa Rica, Universidad Nacional (OVSICORI-UNA) located in Heredia, Costa Rica perform
active geochemical monitoring of several regional volcanoes (Poás, Turrialba, Irazú and Rincón de la Vieja) and their affiliated hot springs, with a focus on gas monitoring of carbon and sulfur volatiles. A small handful of studies have been undertaken to understand the microbial communities that are found across the Costa Rica convergent margin, mostly focused on the off-shore, marine sediment communities. Only two non-marine systems have been microbially characterized in Costa Rica: the Santa Elena Ophiolite in the Outer Forearc and Poas Volcano Crater Lake in the arc. Alkaline geothermal springs in the Santa Elena Ophiolite, the northernmost geotectonic complex in the Costa Rica forearc, support low microbial cell densities \((2.0 \times 10^4 - 1.51 \times 10^5 \text{ cells/mL})\) but contain evidence of microorganisms involved in hydrogen oxidation and methane cycling \((40)\). Metagenomic studies from these springs further reveal that these spring sustain a predominantly methanogenic ecosystem, though the isotopic signal of methanogenesis is potentially overshadowed by abiotic processes \((41)\).

Two known studies have been conducted to search for potential life in the acid crater lake of Poas Volcano, located in central Costa Rica. A 2002 culture-dependent study identified \(\textit{Thiobacillus}\) sp. in lake water and a recent 2018 amplicon based study identified a single OTU of \(\textit{Acidiphilium}\) dominating the microbial community \((42, 43)\). However, no large-scale survey of the effects of microbiology on degassing or fluid expulsion throughout a convergent margin have been performed. The few microbiological studies that have been performed in convergent margins have focused on sites across a smaller area, preventing any region-scale exploration of how microbes interact with these deep geological processes.

**Biology Meets Subduction Field Expedition**

In February 2017, a two-week field expedition across the Costa Rica volcanic arc system was conducted to gain insight into the transport of carbon between the subsurface and surface and the biological and chemical transformations that occur during this movement \((44)\). Twenty-six geochemically diverse sites \((\text{Table 1})\), including hot springs, mud pots, and volcanic lakes, were sampled for microbiology, geochemistry, and petrology in parallel. The goal of this expedition was to improve our understanding of carbon fluxes within the Costa Rica convergent margin, including the sources and potential influence of microbiological activity on carbon degassing at the interface of the subsurface and surface \((44)\). Metadata, including temperature, pH, and elevation, were collected for each sample site. Some sites were a clean clear stream of fluid gushing out of a rock face. Others had fluid gushing up through rock to make a small depression where hot spring fluids pooled at the surface and collected sediments at the bottom of the pool. Some of these were on the banks of rivers. Locations were in backyard creeks and rivers, tourist spas, cattle farms, or wild jungle. The common feature of all sites was that a central fluid source could be identified and was actively expelling fluids.

**Carbon Isotope Composition of Geothermal Fluids in Costa Rica are Consistent with Chemolithotrophy**

Stable isotopes can be used as a tracer of the source of carbon in environmental systems, specifically \(\delta^{13}\text{C}\). Biological metabolism of carbon, including photosynthetic carbon reduction pathways, will impart a highly depleted \(^{13}\text{C}\) signal, and this trend is characteristic of reduced organic carbon \((45)\). The isotopic composition of autotrophic and heterotrophic microorganisms will reflect their source of carbon, be it \(^{13}\text{C}\)-depleted organic matter or heavier oxidized carbon \((46)\). Therefore, we can compare the carbon isotopes of different environmental carbon fractions, including the dissolved inorganic carbon \((\text{DIC})\), dissolved organic carbon \((\text{DOC})\), and total organic carbon \((\text{TOC})\), to identify the source of carbon and metabolism utilized by microbes.
The $\delta^{13}$C of Costa Rica spring fluid DIC increased with distance from trench, and was more $^{13}$C-enriched than photosynthetically-derived carbon, suggesting that DIC in the gushing fluids was deeply-sourced (Figure 3). This is further supported by helium isotope measurements that indicate fluids are indeed subsurface influenced with a strong mantle signal in the Forearc/Arc, and therefore the DIC from these same fluids are representative of subsurface processes as well (47). TOC $\delta^{13}$C values in surface sediments did not vary across the arc, and fell within the range normally observed for photosynthetically-derived surface carbon (-20 and -30‰) (17). Depending on whether the microorganisms utilize the DIC in the fluids or the TOC in the sediments as their carbon source, the $\delta^{13}$C of the dissolved organic carbon (DOC) should parallel that of their carbon source (46). The $\delta^{13}$C of spring fluid DOC paralleled that of the DIC (Figure 3), consistent with either a chemolithoautotrophic source of DOC by DIC-utilizing microbes, or abiotic synthesis from DIC (47).

**Moving Forward: Integration of ‘Omics and Geochemical Datasets**

Due in large part to the logistical challenges of accessing the subsurface, biogeochemical cycling in the vast terrestrial subsurface is poorly constrained and the role of microbes in these processes remain cryptic (5). Until recently, biological studies usually focus on one or two hot springs at a time. A 2018 study represented the largest known study of a geothermal system in a regional scale (48). Current knowledge of biogeochemical cycling in continental convergent margins, which simultaneously represents a reservoir and flux of greenhouse gasses, is therefore largely limited to the abiotic and thermogenic processes influencing these volatile fluxes. Recent advancements in high throughput sequencing have opened the door for the systematic study of the diversity and function of these communities, but key questions remain regarding the biologically-mediated transformation of geothermal volatiles and in turn how these volatiles impact local microbial community dynamics. Additionally, while much work has been done to characterize the biogeochemistry of specific geothermally-impacted sites, such as individual springs in Yellowstone National Park or the Lost Chimney Hydrothermal Field, we lack a basic understanding of microbial diversity along an arc system at the regional scale. The coordination of geochemical and biological sampling leveraged in this study, resulting in the generation of a large integrated dataset spamming the Costa Rica Volcanic Arc System, provides a unique opportunity to begin to delineate the complex interactions between biotic and abiotic processes in a geothermally active subduction zone. Our goal is to integrate high-resolution microbial community dataset from subsurface-influenced fluids and their surface-associated sediments with geochemical measurements, to enhance our understanding of the spatial distribution and diversity of microbial communities across a geothermally active subduction zone in Costa Rica.

**Thesis Aims**

**Aim 1** Determine Whether Hot Spring Microbial Communities are Primarily Influenced by Subsurface or Surface Processes

- **Hypothesis 1.1**: Due to their greater isolation from the surface, hot spring fluid samples will be less influenced by photosynthesis than their corresponding sediments.
  - As evidenced by a low abundance of chloroplast sequences in their 16S rRNA gene amplicon libraries and low concentrations of total photosynthetic pigments.
- **Hypothesis 1.2**: Microbes with chemolithotrophic or mixotrophic metabolisms will be present in elevated abundances in samples with a strong subsurface influence.
  - This has been previously shown with carbon isotope data but has yet to be evaluated from a biological perspective (47)
Aim 2 Identify geochemical characteristics (pH, temperature, & ionic composition) that correlate with across- and along-arc changes in microbial abundance and community composition

- **Hypothesis 2.1**: Microbial cell abundance in fluid samples will negatively correlate with distance from trench in response to the elevated temperatures characteristic of the Forearc/Arc.
- **Hypothesis 2.2**: Microbial cell abundance in fluid samples will be lower in Northern Costa Rica than Central due to the elevated temperatures associated seen in those springs.
- **Hypothesis 2.3**: pH will be a primary driver of microbial community structure, reflecting the pH gradient between the Outer Forearc (alkaline) and Forearc/Arc (acidic-neutral).
- **Hypothesis 2.3**: There will be niche partitioning of chemolithotrophic microbes with similar metabolisms, i.e. sulfate reduction, that reflect the local pH and temperature.
Chapter 2 – Materials and Methods

Biological Sample Collection

At each sampling site, a ProDSS Multiparameter Water Quality Meter (YSI) was placed into the hot spring until it was fully submerged by the fluids to collect the following metadata: temperature, pH, and specific conductivity. Up to 1.5L of hydrothermal fluids, sampled as close to the source as possible, were filtered through Sterivex 0.2um filter cartridges (MilliporeSigma), and 15mL Falcon tubes were filled with sediments and immediately frozen at liquid nitrogen temperature in cryogenic dry shipper (ThermoFisher Scientific, Arctic Express 20) for transport back to the home laboratory. Less than 1.5L was filtered when precipitates clogged the filter. Samples for cell counts were taken as close to the source spring as possible, usually in an outflow from a rock outcrop or a small surface pool that was rapidly being refilled by the source. We placed 1 ml fluids into a 2 ml plastic tube with a rubber O-ring screwcap (to prevent evaporation) containing 500 µl 3% paraformaldehyde solution in phosphate-buffered-saline (PBS) pH?. Samples for single-cell amplified genome (SAG) analysis were collected in the field, preserved in a GlyTE solution (5% glycerol and 1x TE buffer) as recommended by Bigelow Single Cell Genomics Center, and stored at -80°C until analysis (49).

Flow Cytometry

Fluid samples (1mL) were preserved in 0.5mL of 3% paraformaldehyde in the field. Preserved water samples were diluted 1:5 with PBS and stained with 5X SybrGreen. Triplicate aliquots of each sample (200µL) were analyzed on a Millipore Guava Easy Cyte 6HT-2L flow cytometer. Gating strategy was optimized using stained, unstained, and filtered controls. Data were analyzed in R using the vegan package (50).

DNA Extraction and Sequencing

DNA extractions of Sterivex filters were performed using a modified phenol-chloroform extraction optimized for low biomass samples based on methods in Vetriani, et al. 1999, with additional modifications for use with Sterivex filters as described in Wright, et al. 2009. Briefly, extractions were performed via chemical lysis with lysozyme, Proteinase K, and SDS treatment, then purified with phenol-chloroform extractions and precipitation with sodium acetate and isopropyl alcohol. Initial extractions from sediment samples were performed at Rutgers University (New Brunswick, NJ) using the Qiagen DNeasy PowerSoil HTP 96 Kit, with additional extractions performed using the modified phenol-chloroform extraction described in Vetriani, et al. 1999, followed by concentration using the Zymo Genomic DNA Clean & Concentrator Kit. Extracted DNA was quantified using a NanoDrop 2000c (ThermoFischer Scientific) with additional PCR screening performed using universal bacterial primers (53, 54).

DNA was submitted to the Census of Deep Life (CoDL) at the Marine Biological Lab (MBL) at Woods Hole, MA, for amplicon sequencing with an Illumina MiSeq platform (55). The v4v5 hypervariable region of the 16S rRNA gene was amplified separately for bacteria and archaea (56, 57). Amplicon sequences were screened for quality, including chimera-checking with UCHIME, by the MBL as previously described and high-quality merged sequences were published on the Visualization and Analysis of Microbial Population Structures (VAMPS) website (58, 59).
**Analysis of 16S rRNA Gene Amplicon Data**

Taxonomic classification of amplicon sequence variants (ASVs) was performed using the mothur software package and the SILVA database v132 (60, 61). Analysis of ASVs was performed in R using the phyloseq and vegan packages (50, 62, 63). Any sequences classified as chloroplasts, mitochondria, or other eukarya were removed from the bacterial and archaeal libraries prior to further downstream analysis. Abundance filtering was performed to remove any ASVs with less than 5 reads within the entire dataset, and then read counts were normalized to a common-scale by transforming counts to relative abundance within a sample and then multiplying this proportion by the median library size across all samples (64, 65).

Ordination and multivariable analyses was performed using the vegan package on R (50, 63). Unifrac distances, weighted and unweighted, were calculated between all pairs of sites and then visualized using non-metric multidimensional scaling (NMDS) to examine the relationship between community composition across sites (66). Vector fitting of geochemical measurements was modelled using the envfit function and species scores were determined using the wascores function of the vegan package (50, 63). Additional statistical analysis, including Adonis analysis of variance, were performed using the vegan package.

See Appendix I for examples of all code run for the amplicon analysis.
Chapter 3 – Results & Discussion

Site Description & Classification

Twenty-four sites were sampled across northern and central Costa Rica (Figure 2), and were broadly categorized as one or a combination of the following types: spring, pool, farm, river, volcano, mud pot, rock, well (Table 1). With the exception of rock springs, which lacked any sediments, sites had a deep gushing source of fluid surrounded by surface sediments at the bottom of water (Figure 1). At sites where deep subsurface fluids interacted with surface-exposed sediments, we took whole sediment as a control for surface contamination. At sites where springs were feeding into a pool of standing water, we sampled as close to the source as possible. These included 8 sites from the Outer Forearc, 15 from the Forearc/Arc, and 1 from the backarc, with temperatures ranging from 23 to 89°C, pH ranging from < 1 to 10, and specific conductivity ranging from 197 to 91900 µS/cm². Two active volcanoes were sampled: a spring on the flank of Irazu Volcano and the crater lake of Poas Volcano.

Fluid and sediment samples were collected for molecular analysis in the field and frozen at -80°C until extraction. Only three sediment samples yielded no PCR amplifiable DNA (indicated with NA in Table 1). Samples yielding amplifiable DNA but no amplicon library are indicated with a minus sign (-), while those yielding a bacterial and/or archaeal library are indicated with a B or A respectively. 17 out of 21 sediment samples and 13 out of 24 fluid samples yielded at least one amplicon library, with 11 sites having a fluid-sediment pair. In total, 30 bacterial and 19 archaeal amplicon libraries were generated and analyzed.

Distribution of Biomass in Hot Spring Fluids

Cell abundance in geothermal spring fluids was evaluated using flow cytometry, measuring each sample in triplicate. The sample from Poas Lake was not included in downstream analyses as it represents a distinct type of geothermal system, in that it is a volcano crater lake, and due to the low cell abundance compared to other samples as measured in two separate samples (2.4x10³ cells/mL ± 1.42 x10³). There was slightly, but statistically significant, higher cell abundance in the Forearc/Arc as compared to the Outer Forearc (Wilcoxon test, p = 0.04129; Figure 4A). The average cell abundance in measured Forearc/Arc samples was 9.41x10⁵ cells/mL (± 8.21x10⁵) and 7.60x10⁵ cells/mL (± 1.13x10⁶) in the Outer Forearc. The geochemical boundary between the Outer Forearc and the Forearc/Arc can be defined primarily based on changes in pH and to some degree by variation in temperature. Springs in the Outer Forearc were characterized by high pH (>8.5) and cooler temperatures (<40°C), while springs in the Forearc/Arc were characterized by acidic pHs (<6) and warmer temperatures (>40°C). Given these geochemical characterizations, one would expect the Forearc/Arc to show a decrease in cell abundance as temperature increases, however cell abundance was positively correlated with temperature across the arc (Spearman = 0.468, p < 0.001). On the other hand, cell abundance was negatively correlated with pH (Spearman = -0.504, p < 0.00001), which suggests that the lower cell abundance in the Outer Forearc can likely be explained by its alkaline nature rather than by its lower temperatures. There is no statistical difference in along-arc cell abundance whether comparing sites split by the EPR-CNS boundary (Wilcoxon test, p = 0.2274; Figure 4B) or the rough-smooth boundary (Wilcoxon test, p = 0.3251, Figure 4C). This indicates that there is no along-arc variation in microbial cell abundance.
Pre-Processing & Contamination Screening of 16S rRNA Gene Amplicon Libraries

Before the amplicon data could be analyzed, the data needed to undergo pre-processing and normalization. Quality checking was performed at MBL as previously described with additional quality checking performed in mothur to confirm sequencing quality (58, 59). After this, the first step of pre-processing was to remove chloroplast and mitochondrial classified ASVs from both the bacterial and archaeal amplicon datasets (See Table 2), of which there were none found in the archaeal library.

Next, low quality libraries were removed from their respective datasets. From the bacterial dataset, two libraries of unknown identity were removed (designated S14 and S33), leaving 32 libraries for analysis. From the archaeal dataset, one unknown library was removed (S14) and the Qubrada Naranja fluid (QNF) and Poas Laguna Filter (PGF) samples were removed due to low read counts (<1000). The final pre-processing step was to remove low abundance ASVs where ASVs with less than 5 reads within the entire bacterial or archaeal datasets were removed. This abundance filtering removed between 14-46% of total reads from each sample library. At this point, the bacterial library was screened for putative contaminants based on a recent paper detailing common contaminants in CoDL datasets (67). Based on the low abundance of any individual identified contaminant ASV (<0.04% in the entire dataset and <0.01% in any individual library) and the fact that no ASV was present in all samples (possibly signifying a contaminant introduced during processing), no ASVs were removed as contaminants.

Hot Spring Fluids Show Low Photosynthetic Potential

In addition to the previously presented isotope data (see Chapter 2), the amount of surface contamination in fluid and sediment samples can be evaluated by looking at the relative abundance of chloroplast reads in amplicon libraries, as a proxy of plant input into biomass. Overall, all fluid samples, with the exception of Poas Laguna (PG), which is a lake, also called Bota Lake, had low abundance of chloroplast sequences in their 16S rRNA gene amplicon libraries, representing < 1% of all sequences (Figure 5). It should be noted that while bacteria 16S rRNA gene primers are known to amplify chloroplast 16S rRNA gene sequences, the efficiency of this amplification is highly dependent on the primer pair chosen, from 0.1-95% (68). Primer coverage rates for chloroplast sequences (Domain: Bacteria, Phyla = Cyanobacteria, Class = Oxyphotobacteria, Order = Chloroplast) were determined using the SILVA Test Probe, and shown to have up to 90% coverage for sequences within this taxon in the SILVA nr132 SSU database (69). Therefore, the relative abundance of chloroplast reads in individual sample libraries is a valid method by which to evaluate the presence of photosynthetic material in our samples.

Additionally, collaborators measured the concentrations of total photosynthetic pigments in the sediment samples collected (Figure 6), which can be used as a proxy of photosynthetic potential. Even though sediment samples were exposed to sunlight at the surface of the springs, they also had low concentrations of total photosynthetic pigments (< 5 µg/g) suggesting that despite being constantly exposed to the surface, sediment microbial communities potentially maintain some characteristics of the subsurface fluids that overwashed them. Exceptions were sediments from El Sitio with 46.4 µg/g total photosynthetic pigment, but less than 1% chloroplasts, and Pompilo’s Finca with 19.9 µg/g total photosynthetic pigments and 11.7% chloroplasts. Both samples represent agriculturally influenced sites used for cattle ranching. These sites, in addition to Poas Laguna, were also the only places with more than a few liters of water present in a long-term pool. Here, the residence time of the pooling water may have been long enough to establish a phototrophic community.
Paired Fluid-Sediment Samples are Qualitatively Similar but Quantitatively Different

Nonmetric multidimensional scaling (NMDS) ordinations of unweighted and weighted UniFrac distance matrices were generated to reveal differences in microbial community composition between sample types: fluids vs sediments. For the bacterial libraries, fluid and sediment samples from the same site ordinated close to each other within the unweighted UniFrac (Figure 7). This indicated that fluids and their corresponding sediments shared similar species richness, meaning the same species are present in both sample types. However, within the weighted UniFrac ordination, fluids and their corresponding sediments showed distinct separation in ordination space that was statistically significant (adonis \( r^2 = 0.08851, p = 0.002 \)). This suggests that while the presence/absence of a particular taxa is primarily driven by site, the relative abundances of these taxa are dependent on whether they are in the fluid or sediment phase. Separation of fluids and sediments in ordination space provided further evidence that the fluid 16S rRNA gene libraries were subsurface-derived and that their corresponding sediment 16S rRNA gene libraries, while containing similar taxa, likely represented a mixed community of subsurface and surface impacted microbes. These findings are consistent with recent work in Great Boiling Spring in Yellowstone National Park, which found that there were significant differences between the water and sediment communities found in pools (70).

The archaeal amplicon libraries do not show the same separation of fluids and sediments in ordination space as the bacterial amplicon libraries do (Figure 8, \( p > 0.7 \)). However, the archaeal amplicon libraries had significantly fewer ASVs than the bacterial libraries, 7992 ASVs vs 24021 ASVs respectively, in addition to less paired sediment-fluid samples from the same site (2 vs 12). Overall, this reduced our ability to delineate subsurface vs surface influenced samples to the same extent that could be seen in the bacterial dataset.

Four Microbial Biogeographical Regions Defined by Across and Along Arc Gradients

Across Arc Variation in Microbial Communities Reflects Subduction Progression

Within ordination space, samples from the Outer Forearc cluster separately from samples from the Forearc/Arc (Figure 9, Figure 10). Separation of the Outer Forearc samples is primarily driven by pH, which is inversely correlated with distance from trench. This grouping of sites by Forearc/Arc vs. Outer Forearc is statistically significant in the bacterial dataset whether you consider the unweighted (adonis \( p < 0.001 \)) or the weighted Unifrac (adonis \( p = 0.013 \)), and for the unweighted archaeal Unifrac ordination (adonis \( p = 0.041 \)). Based on geochemical modelling, clustering of samples in the Outer Forearc vs Forearc/Arc was primarily driven by differences in pH and temperature, which is consistent with the previously presented cell abundance data. Further analysis of specific taxa diagnostic of the Outer Forearc and Forearc/Arc will be presented in a later section.

Seafloor Bathymetry Drives Along Arc Variation in Microbial Communities

As previously discussed, there are two potential dividing boundaries along the arc: 1) the distinction between the EPR-CNS crust and 2) the rough-smooth boundary just south of this crustal change. Previously published research indicated that the carbon stable isotope composition of the incoming slab can be differentiated across the EPR-CNS boundary (Barry et al. 2019, in press). However, the EPR-CNS boundary did not significantly explain the clustering of microbial composition for the Unweighted Unifrac ordination for either the bacterial or the archaeal datasets (adonis \( p-values >0.18 \)). But it did explain 5-9% of the separation within the weighted Unifrac ordination (adonis \( p < 0.05 \)).
rough-smooth boundary was a better divider of along-arc biological changes within the bacterial datasets (unweighted $p < 0.001$, weighted $p < 0.005$), and the archaeal unweighted dataset ($p = 0.04$). Further analysis of specific taxa diagnostic of these along-arc changes will be presented in a later section.

**Four Region Model of Microbial Community Composition**

Upon further analysis of the unweighted bacterial NMDS (Figure 9), four distinct clusters of sites can be seen. These four clusters can be defined by both their across-arc and along-arc geographical positions. This two-dimensional separation of the bacterial data resulted in four clear clusters forming based on region: Smooth Outer Forearc (blue), Rough Outer Forearc (green), Smooth Forearc/Arc (red), and Rough Forearc/Arc (orange). This grouping of sites based on region was statistically significant when considering the entire bacterial dataset (unweighted & weighted $p < 0.001$). This regional cluster is also significant when considering only the fluid samples (unweighted $p < 0.001$, weighted $p = 0.012$). As these fluids are more representative of subsurface communities than sediment samples, we can infer that this regional clustering is at least in part driven by differences in geochemistry that are the result of subsurface geological processes.

Similar regional clustering is seen when considering the unweighted Unifrac NMDS of the archaeal community composition (Figure 10, $p < 0.001$), but it does not hold true when considering the weighted Unifrac. This indicates that while the presence or absence of particular archaeal ASVs is dependent on their geological region within the arc system, their abundance is dependent on other, currently unknown, factors. Due to a lack of fluid samples yielding archaeal libraries, ordinations of the fluid subset of archaeal samples is not informative about trends in subsurface archaeal communities. However, as the regional model holds true when considering either the entire bacterial dataset or just the fluid subset, broad conclusions on regional trends of particular archaeal taxa can be drawn from analysis of the entire archaeal dataset.

**Putative Chemolithotrophic Bacteria Dominate in Geothermal Fluids**

To identify patterns in bacterial composition within the four identified regions of the arc system, the weighted average scores of the most abundant ASVs were configured to the NMDS plots (Figure 11). Scores were calculated for the ten most abundance ASVs in each site, which represented between 35-91% of all reads in each normalized library. No ASVs were shared amongst all the bacterial libraries, nor within the fluid or sediment subsamples of the data. Most of the dominant bacterial ASVs are chemolithotrophic with diverse metabolic capabilities, including sulfur, hydrogen, and iron cycling bacteria. While no ASVs are shared, there were eight bacterial genera that were found in all fluid samples, including *Hydrogenophaga*, *Acinetobacter*, unclassified *Rhodocyclaceae*, unclassified *Burkholderiaceae*, uncultured *Anaerolineaceae*, uncultured *Pirellulaceae*, unfultured *Thermodefulfovibrionia*, and unclassified *Bacteria*. The high frequency of uncultured clades in these fluid samples is likely a reflection of the high frequency of uncultured phyla in geothermal systems (71). Additionally, as a whole the bacterial 16S rRNA gene amplicon libraries show higher genus richness than the archaeal libraries, with each sample containing between 86-697 bacterial genera and 23-71 archaeal genera (Table 3).

Many of these most abundance bacterial groups are have putative metabolisms dependent on the presence of oxidants to fuel their metabolism, which would imply these organisms are not from the deep subsurface or are dependent on periodic influx of surface water entrainment. They likely live at the interface where subsurface fluids ascend and mix with surface fluids, potentially in the sediments that are washing over with these mixing fluids. The chemical disequilibria caused by the mixing of reduced
hydrothermal fluids with oxidized surface materials can provide the energy necessary for microbial populations to thrive (72). Among the most abundance ASVs, each region contained at least one ASV representing an unclassified *Rhodocyclaceae*. This family is physiologically diverse with most organisms having a strictly respiratory metabolism that is dependent on the presence of oxidants, such as oxygen and nitrate, with select genera requiring sunlight for their metabolism (73). The presence of anoxygenic phototrophs, including the moderately acidophilic *Rhodoblastus*, the deeply branching *Chloroflexi*, and the putatively photoheterotrophic *Anaerolineaceae* further support this interface paradigm, as they require the reduced hydrogen from the subsurface fluids along with sunlight to survive (74–77). Very few of the most abundant bacterial taxa are known to survive anaerobically, but there are a few exceptions and these organisms may represent true, subsurface microbes. These clades include *Anoxybacillus* and *Ignavibacteriales* which have both been identified previously in thermophilic, geothermal systems (78, 79).

**Sulfur and Iron Cycling Microbial Clades Show Niche Separation**

**Iron Oxidizing Bacteria are Limited to the Central Costa Rica Region**

ASVs from putative iron-cycling bacterial, including *Gallionellaceae*, *Geothrix*, and *Geobacter*, were only present in relative abundances greater than 1% in fluid or sediment samples within the Central Forearc/Arc. The only archaeal ASV with iron-cycling potential was *Ferroplasma*, which is dominant in the Poas Background sediment sample but was absent from all other samples. Many of these sites had visible red-orange iron deposits in sediments and biofilms. The biological signal for the iron cycling is strong in these Central Forearc/Arc samples, but iron concentrations are not particularly high in this region (Figure 12). Additionally, when using geochemical models to identify characteristics that explain variation in the biological data, iron concentration is not included in the best geochemical models.

Iron concentrations were measured in both unacidified and acidified samples, measuring the soluble ferrous iron (II) and total iron respectively, which can provide insight into the fractionation of iron in these geothermal spring environments (80). The relative amount of ferric iron present in these samples was highest in samples from the Northern Forearc/Arc, ranging from 7-40% of the total iron concentration measured (Figure 13). This low content of ferrous iron (II) in the Northern Forearc/Arc, may be due to precipitation by sulfides present in the Guanacaste region hydrothermal system (36–38). The geothermal system in the Guanacaste region is also more developed than the Central hydrothermal system, allowing for increased mixing of geothermal fluids and magmas, allowing for any iron in the system to be fully reacted (de Moore, personal communication). This low abundance of soluble, and therefore bioavailable reduced iron, may explain the lack of iron oxidizing organisms in this region. However, the highly variable pH across the springs, particularly the acidic springs sampled in the Northern Forearc/Arc, likely contributes to inaccuracies in the measured iron concentrations. This inaccuracy is reflected in instances where the concentration of iron in acidified samples is lower than that measured in the unacidified samples.

**Diverse Sulfur Cycling Bacteria in the Costa Rica Volcanic Arc System**

Putative sulfur oxidizing bacteria were present in all four regions of the arc system. However, they showed a potential niche separation based on the geochemical characteristics of the region. More mesophilic, filamentous sulfur-oxidizing bacteria such as *Thiothrix* and *Thiovirga* were dominant in the Central Outer Forearc and could also be found farther north in the Nicoya Peninsula. The Northern Outer Forearc was also the only region with abundant representatives of sulfate reducing bacteria, the
Deltaproteobacteria Sva0485 and *Desulfatiirhabdium. Aquificales* of the genus *Sulfurihydrogenibium*, all cultured members of which are autotrophic oxidizers of sulfur species, were widespread in the Northern Forearc/Arc, with multiple ASVs associated with springs such as Mousetrap and Finca Ande (81). This is consistent with the prevalence of *Sulfurihydrogenibium* in weakly acidic hot springs in Yellowstone National Park and Mexico (25, 82, 83). The dominance of this clade in the Northern Forearc/Arc, and the lack of it elsewhere, is consistent with the temperature-dependent niche separation along-arc as members of the *Aquificales*, including *Sulfurihydrogenibium*, are in very low abundance in springs with temperatures below 50°C (48).

**Arc Archaeal Hot Spring Communities are Dominated by Either *Nitrososphaeria* or *Bathyarchaeota***

All archaeal amplicon libraries, with the exception of the Poas Background Sample sediment, were dominated by either *Nitrososphaeria* or *Bathyarchaeota* (Figure 15). The Poas Background sample was dominated by a single ASV belonging to the *Ferroplasma* within the *Thermoplasmatales* that is absent from all other archaeal libraries. This sample was the only one with no flowing or standing water associated with it, so this may explain why it has such a different archaeal community. The *Nitrososphaeria* are chemolithotrophic ammonia oxidizing archaea (AOA) that have been found in aquatic, terrestrial, and geothermal environments and contain mesophilic, acidophilic, and thermophilic representatives (84–87). Genomic evidence indicates that this clade likely assimilates carbon autotrophically utilizing a modified 3-hydroxypropionate/4-hydroxybutyrate pathway, which is absent in obligate heterotrophic archaea (88). *Bathyarchaeota* are postulated to play a key role in carbon biogeochemistry due to the diversity subgroups display in heterotrophic carbon assimilation pathways (89–92).

Since only three samples from the Outer Forearc generated successful archaeal amplicon libraries, across-arc trends within the Archaea cannot be accurately determined. However, these three samples indicated that there may be along-arc variation within the Outer Forearc. In the Central Outer Forearc, Quepos archaeal fluid communities were dominated by *Candidatus Nitrosopumilales* (>85%) and *Bathyarchaeota* (~10%). In contrast, the two Northern Outer Forearc samples were dominated by *Bathyarchaeota*. El Sitio sediments were dominated by *Bathyarchaeota* (>95%), with a small representation of Methanobacteriales and *Candidatus Nitrosopumilales*. Espabel sediments were the most diverse of the three Outer Forearc samples, with a dominance of *Bathyarchaeota* (~60%), *Methanobacteriales* (~25%), *Methanosarcinales* (~1%), and *Candidatus Nitrosopumilales* (~15%). In contrast to previous studies in serpentinizing systems, which are analogs to samples from the Outer Forearc in Costa Rica, there was little to no methane degassing present at the springs sampled. However, in samples where measurable methane was released, such as Espabel, methanogenic taxa were present in higher abundances.

While across-arc changes in archaeal diversity cannot be evaluated with the given dataset, along-arc variations in archaeal community composition within the Forearc/Arc can be seen. The smooth, Northern region of the Forearc/Arc had communities dominated by the *Nitrososphaeria*, while the Central Forearc/Arc are dominated by members of the *Bathyarchaeota*. In terms of the composition of *Nitrososphaeria*, the smooth, Northern Arc sees a preference for the thermophilic *Candidatus Nitrosocaldales*, the thermophilic branch of the *Nitrososphaeria*. This trend coincides with the analysis of the bacterial amplicon libraries where differences in composition between the smooth and rough regions along the arc could primarily be explained by variations in temperature. This is further supported by geological phenomena, where seamounts in rough subducting sediments result in increased heat loss and overall cooler subduction in Central Costa Rica (93). The dominance of *Candidatus Nitrosocaldales* in Poas Volcano Lake, which is located in Central Costa Rica, can be
explained by the fact that the lake is directly influenced by subsurface, magmatic processes, unlike the rest of the Central Forearc/Arc samples. Therefore, its archaeal composition was more reflective of the northern regions. The abundance of *Nitrososphaeria* was not correlated to ammonium concentrations in spring fluids (Figure 14), implying that other geophysical factors are driving their abundance in the northern region.

Finer classification of *Bathyarchaeota* ASVs could be achieved by aligning these short 16S rRNA gene sequences and inserting them into the primary *Bathyarchaeota* phylogenetic tree to assign subgroups to these ASVs (94, 95). Classification of these sequences to the sub-group level could allow for identification of metabolic niches of the *Bathyarchaeota* across hydrothermal systems, and represents a logical next step in the diversity analysis.
Chapter 4 — Conclusions

Previous work by collaborators showed that the plate boundary between EPR and CNS altered the carbon chemistry across the whole convergent margin transect, from trench to arc (47). However, the biological changes across and along this volcanic arc system had not been previously evaluated. Unlike the carbon geochemistry, the bacterial and archaeal microbial communities could not be differentiated across the EPR-CNS plate boundary, meaning that the changes in stable carbon isotopic ratio of the carbon coming off the slab caused by the plate boundary do not translate to other geochemical variations that greatly alter the microbial community. Although the CNS/EPR plate boundary delineated no changes in the microbial community composition, an apparent north-central regional divide within the Forearc/Arc and Outer Forearc aligns with a shift in oceanic bathymetry designed the rough-smooth boundary. While the difference in latitude between the EPR-CNS and rough-smooth boundaries is only about half a degree, changes in along-arc microbial composition can be better explained by the rough-smooth boundary when using more stringent p-value cutoffs. These along-arc biological changes are primarily driven by changes in temperature, with the Northern region of Costa Rica hosting thermophilic microbial taxa, including Candidatus Nitrosocylindrus and Sulfurihydrogenibium (Figure 11, Figure 15). This variation in temperature is consistent with the changes in seafloor bathymetry, where the rough, seaweed-laden crust in the CNS-2 segment results in increased heat loss, which corresponds with the cooler temperatures seen in terrestrial hot springs in Central Costa Rica. This is also consistent with the more well-developed geothermal system in the northern Guanacaste region providing a different array of substrates for chemolithoautotrophic communities.

We found that the major driver of variation in across-arc microbial community composition was primarily variations in subsurface fluid pH, with contributions of temperature, with increasing distance from the trench (Figure 9). Communities in the Outer Forearc contained clades known to be adapted to high pH and lower temperature, and those in the Forearc/Arc contained those adapted to low pH and higher temperature, reflecting local geochemistry. Geochemistry in the arc is tightly tied to the regional volcanic gas emissions and activity. The Outer Forearc’s characteristic alkalinity is the result of the release of consumption of protons into molecular hydrogen while the acidic nature of the Forearc/Arc is tied to acidic volatiles such as CO$_2$, SO$_2$, H$_2$S, and hydrogen halides[hydrohalic acids?] increase in abundance as you approach the volcanic arc itself (12, 17).

Taking into account both the along and across-arc geochemical gradients and changes in microbial community composition, we propose a 4 region model delineated by changes in pH and temperature: 1) a Northern Forearc/Arc – acidic and thermophilic; 2) a Central Forearc/Arc – acidic and thermophilic, but less so than its northern counterpart; 3) a Northern Outer Forearc – alkaline; and 4) a Central Outer Forearc – alkaline. As less sites were sampled in the Outer Forearc, specifically only a single site in the Central Outer Forearc, we cannot confidently analyze biogeochemical trends along this section of the arc system. However, ordinations (Figure 9, Figure 10) indicate that the few sites sampled from the Central Outer Forearc are distinct from their northern counterparts, highlighting a prime target for future sampling expeditions. Despite the stark differences in geochemistry between these four regions, we found common characteristics shared among the regions. Bacterial and archaeal chemolithoautotrophs were among the most abundant ASVs at each site, regardless of geochemistry. Variation in local geochemistry resulted in niche partitioning of certain microbial physiologies, such as microbial sulfur oxidation, with different taxa dominating in different regions. For example, the primary sulfur oxidizing bacteria in the Northern Forearc/Arc were Sulfurihydrogenibium and Hydrogenothermaceae, while the Outer Forearc contained diverse sulfur oxidizers including Thiothrix, Thiovirga, and Sulfitalea (Figure 11).
In the geothermal springs sampled, the local microbial communities were predominantly chemolithotrophic and their physiologies suggest that they likely take advantage of the redox interface generated by the mixing of reduced subsurface geothermal fluids and oxidized, photosynthetically derived materials on the surface. Similar studies of hydrothermal springs in Yellowstone National Park suggest that the moderately acidic pH values (those between 4-6) reflect the dilution of highly acidic volcanic gas phase-influenced fluids with meteoric water (83). Given the moderately acidic and alkaline nature of the springs sampled across the Costa Rica volcanic arc, a similar phenomenon may be occurring where ultra-alkaline or acidic geothermal fluids are mixing with surface waters, in the Outer Forearc and Forearc/Arc respectively. The lack of shared ASVs and limited number of shared genera suggest that niche selection drives the composition of these chemolithotrophic communities (Figure 11), such as is seen in the sulfur oxidizing bacteria in response to temperature and pH gradients (48).

The variability in geochemistry across the arc could be due to a single subsurface fluid source mixing with meteoric water to variable degrees. However, there is no evidence for correlation between chloride and bromide amongst our sites, implying that variability is due to different sources of subsurface input rather than a single subsurface endmember mixing with meteoric water (96). This heterogeneity has been well-characterized across the Costa Rica arc system. For example, thermal and non-thermal fluids discharged on the flanks of Rincon de la Vieja volcano either predominantly reflect meteoric input, others are consistent with steam heating of shallow groundwater, and others represent sulfur-deplete waters generated through rock-water interactions (97). Even river systems that are many tens of kilometers away from a volcanic system in the non-volcanically active Backarc will see localized heterogeneity in geochemistry as a result of upwelling geothermal fluids (98).

Of the geothermal sites sampled for this thesis, only one had been previously characterized microbiologically: Poas Volcano Lake (also know as Laguna Caliente). Our sampling of Poas Lake in February 2017, returned a community dominated by a single taxon: \textit{Sulfurihydrogenibium}. An early culture-dependent study of Poas Lake identified mat-associated bacteria classified as \textit{Thiobacillus} and \textit{Bacillus}, while a more recent amplicon-based study showed the microbial diversity was limited to a single species belonging to the \textit{Acidiphilium} (42, 43). All of these taxa are involved in sulfur cycling, so they are all logical taxa metabolically to be found in the highly sulfuric Poas Lake, but the changes in taxonomy over time are surprising. Alternatively, these differences may be a consequence of each study utilizing a different DNA extraction technique and amplifying a different region of the 16S rRNA gene, which are both known to introduce differences in sequencing data (99, 100). None of these studies, this one included, involved taking multiple samples, or if they did the extreme low biomass of the lake necessitated pooling samples, though this lack of replicates does not discount the results generated but limits their ability to be compared to each other (101).

This research provides a proof of concept foundation for studying regional-scale biogeochemistry in geothermal systems. Through the integration of high-throughput molecular data with a diverse geochemical dataset, we can begin to understand how subsurface processes influence surface biology along defined geochemical gradients. Our work complements the work of the Earth Microbiome Project and the Deep Carbon Observatory’s Census of Deep Life in characterizing microbial populations across the globe, but also highlights the need for further studies of the continental subsurface to understand spatial patterns in heterogeneity.
References


the deep biosphere of a fossil Lost City-type hydrothermal system at the Iberia Margin. Proc Natl Acad Sci U S A 112:12036–41.


Appendices
Appendix I – R Code
Carbon Isotope Figure 3 Generation

library(phyloseq)
library(ggplot2)
library(tidyverse)
library(ggrepel)
library(gridExtra)

bac_geo <- read.csv("L:/asv_analysis/bac/CORRECTED/bms2017_bac_geo.csv", stringsAsFactors = F)
%>%
column_to_rownames(var = "..sample")

bac_geo %>%
ggplot(aes(x = Distance_from_trench_km)) +
theme_bw() +
geom_point(aes(y = DIC_d13C), colour = "white", size = 2, shape = 17) +
#geom_point(aes(y = DOC_d13C), colour = "darkorchid2", size = 2, shape = 16) +
geom_point(aes(y = TOC_d13C), colour = "black", size = 2, shape = 18)+
labs(x = "Distance from Trench (km)", y = "d13C (per mil")

Chloroplast & Photosynthetic Pigment Analysis & Figures 5 & 6

library(ggplot2)
library(tidyverse)
library(vegan)

bac_chloro_data <- read.csv("190317_chloro.csv")
colnames(bac_chloro_data)[1] <- "sample"

#Figure 6
pig <- bac_chloro_data %>%
  filter(total_photo_pig >= 0) %>%
  filter(sample_type == "s") %>%
ggplot(aes(x = site)) +
geom_bar(aes(y = total_photo_pig), fill = "palegreen2", stat = "identity") +
theme(plot.subtitle = element_text(vjust = 1),
  plot.caption = element_text(vjust = 1),
  panel.background = element_rect(fill = NA)) +
labs(x = "Site", y = "Total Photosynthetic Pigments (ug/g")
+labs(title = "Total Photosynthetic Pigments in Hot Spring Sediments")

chloro_labels <- c(f = "F", s = "S")

#Figure 5
chloro <- bac_chloro_data %>%
  filter(site != "S14") %>%
  filter(site != "S33") %>%
ggplot(aes(x = sample_type)) +
  theme_bw() +
  geom_bar(aes(y = chloro_reads_rel, fill = sample_type), stat = "identity") +
  scale_fill_manual(values = c("lightskyblue3", "darkgreen")) +
  facet_grid(. ~ site, drop=TRUE, scale="free", space="free_x") +
  theme(plot.subtitle = element_text(vjust = 1),
        plot.caption = element_text(vjust = 1),
        panel.background = element_rect(fill = NA),
        legend.position = "bottom") +
  labs(x = "Site", y = "Relative Abundance of Chloroplast Reads")

**Cell Abundance Analysis & Figures**

fluid_counts <- read.csv("fluid_counts_1.csv")
colnames(fluid_counts)[1] <- "sample"
colnames(fluid_counts)[7] <- "fluid_cell_concentration"

fluid_counts$region = factor(fluid_counts$region, levels(fluid_counts$region) [c(3, 1, 4, 2)])

#REMOVE POAS
fluid_count_no_poas <- filter(fluid_counts, sample != "PL")

fluid_count_no_poas_mean <- fluid_count_no_poas %>%
  group_by(Province) %>%
  summarise(mean(fluid_cell_concentration))

wilcox.test(fluid_cell_concentration ~ epr_cns, data = fluid_count_no_poas, exact = F, paired = F)
wilcox.test(fluid_cell_concentration ~ rough_smooth, data = fluid_count_no_poas, exact = F, paired = F)
wilcox.test(fluid_cell_concentration ~ Province, data = fluid_count_no_poas, exact = F, paired = F)

across_count_mean <- fluid_count_no_poas %>%
  ungroup() %>%
  group_by(Province) %>%
  summarise(count = n(), across_mean = mean(fluid_cell_concentration), across_stdev =
            sd(fluid_cell_concentration))

#rough-smooth

trench_rough_smooth_no_poas_plot <- ggplot(fluid_count_no_poas, aes(x =
  Distance_from_trench_km, y = fluid_cell_concentration, color = rough_smooth)) +
  geom_point() +
  geom_smooth(method = "lm", fill = NA) +
  yscale("log10", .format = T)+
  labs(x = "Distance from Trench (km)", y = "Cell Concentration (cells/mL)",
       subtitle = "Fluid Microbial Cell Counts Along the Costa Rica Volcanic Arc: Rough-Smooth (No Poas)")

#outer forearc vs forearc/arc
trench_across_no_poas_plot <- ggplot(fluid_count_no_poas, aes(x = Distance_from_trench_km, y =
  fluid_cell_concentration, color = Province)) +
geom_point() +
geom_smooth(method = "lm", fill = NA) +
yscale("log10", .format = T)+
labs(x = "Distance from Trench (km)", y = "Cell Concentration (cells/mL)",
subtitle = "Fluid Microbial Cell Counts Across the Costa Rica Volcanic Arc (No Poas)"

#epr/cns
trench_epr_cns_no_poas_plot <- ggplot(fluid_count_no_poas, aes(x = Distance_from_trench_km, y = fluid_cell_concentration, color = epr_cns)) +
  geom_point() +
  geom_smooth(method = "lm", fill = NA) +
  yscale("log10", .format = T)+
labs(x = "Distance from Trench (km)", y = "Cell Concentration (cells/mL)",
subtitle = "Fluid Microbial Cell Counts Along the Costa Rica Volcanic Arc: EPR/CNS (No Poas)"

#regions
kruskal.test(fluid_cell_concentration ~ region, data = fluid_count_no_poas)

region_no_poas_plot <- ggplot(fluid_count_no_poas, aes(x = Distance_from_trench_km, y = fluid_cell_concentration, color = region)) +
  geom_point() +
  geom_smooth(method = "lm", fill = NA) +
  yscale("log10", .format = T)+
labs(x = "Distance from Trench (km)", y = "Cell Concentration (cells/mL)",
subtitle = "Fluid Microbial Cell Counts Along the Costa Rica Volcanic Arc: Region (No Poas)"

wilcox_region <- compare_means(fluid_cell_concentration ~ region, fluid_counts, method = "wilcox.test")

region_comparison_num <- list(c("1, 2"), c("1, 3"), c("1, 4"), c("2, 3"), c("2, 4"), c("3, 4"))

region_comparisons <- list(c("Smooth Northern Forearc/Arc", "Rough Central Forearc/Arc"),
epc("Smooth Northern Forearc/Arc", "Smooth Northern Outer Forearc"),
c("Smooth Northern Forearc/Arc", "Rough Central Outer Forearc"),
c("Rough Central Forearc/Arc", "Smooth Northern Outer Forearc"),
c("Rough Central Forearc/Arc", "Rough Central Outer Forearc"), c("Smooth Northern Outer Forearc", "Rough Central Outer Forearc"))

fluid_count_no_poas$region_num <- as.numeric(fluid_count_no_poas$region)

box_plot_region <- ggboxplot(fluid_count_no_poas, x = "region", y = 'fluid_cell_concentration', add = "jitter") +
  stat_compare_means(comparisons = region_comparisons, method = "wilcox.test", label = "p.signif") +
  stat_compare_means() +
labs(x = "Region", y = "Cell Concentration (cells/mL)",
subtitle = "Fluid Microbial Cell Counts Across the Costa Rica Volcanic Arc") +
yscale("log10", .format = T) #+
#stat_compare_means(aes(group = region), method = "wilcox.test")
#geom_signif(comparisons = region_comparisons, map_signif_level=TRUE)

fluid_count_no_poas %>%
  ggplot(aes(x = ph, y = fluid_cell_concentration)) +
  geom_point() +
  yscale("log10")

cor.test(x = fluid_count_no_poas$fluid_cell_concentration, y = fluid_count_no_poas$ph, use =
  "complete.obs", method = "spearman")

fluid_count_no_poas %>%
  ggplot(aes(x = temp, y = fluid_cell_concentration)) +
  geom_point() +
  yscale("log10")

cor.test(x = fluid_count_no_poas$fluid_cell_concentration, y = fluid_count_no_poas$temp, use =
  "complete.obs", method = "spearman")

fluid_count_no_poas %>%
  ggplot(aes(x = spc, y = fluid_cell_concentration)) +
  geom_point() +
  yscale("log10")

box_plot_across <- ggboxplot(fluid_count_no_poas, x = "Province", y = 'fluid_cell_concentration', add =
  "jitter") +
  stat_compare_means(comparisons = list(c("Outer Forearc", "Forearc/Arc")), method = "wilcox.test", label = "p.signif") +
  stat_compare_means() +
  labs(x = "Province", y = "Cell Concentration (cells/mL)",
       subtitle = "A") +
  yscale("log10", .format = T)

box_plot_epr_cns <- ggboxplot(fluid_count_no_poas, x = "epr_cns", y = 'fluid_cell_concentration', add =
  "jitter") +
  stat_compare_means(comparisons = list(c("EPR", "CNS")), method = "wilcox.test", label = "p.signif") +
  stat_compare_means() +
  labs(x = "EPR vs CNS", y = "Cell Concentration (cells/mL)",
       subtitle = "B") +
  yscale("log10", .format = T)

box_plot_rough_smooth <- ggboxplot(fluid_count_no_poas, x = "rough_smooth", y =
  'fluid_cell_concentration', add = "jitter") +
  stat_compare_means(comparisons = list(c("Rough Central", "Smooth Northern")), method =
  "wilcox.test", label = "p.signif") +
  stat_compare_means() +
  labs(x = "Rough-Smooth", y = "Cell Concentration (cells/mL)",
       subtitle = "C") +
  yscale("log10", .format = T)
ASV Classification

Brazelton Lab Python Scripts

**Python scripts were downloaded from https://github.com/Brazelton-Lab/lab_scripts/tree/master/16S, following the bioinformatics pipeline established in Brazelton et al (2017)**

**Parallel analysis was performed on the bacterial and archaeal 16S rRNA gene amplicon libraries**

# expand VAMPS fasta file according to abundance indicate at end of header
fasta-expander-vamps-2016.py <each_file_name>.fasta

# partition FASTA file into multiple files according to header
fasta-sort-by-header.py

# Creates a group file for a project by providing one or more fasta or fastq formatted files.
group_from_filename.py -separator . -position ! *.expanded.fa > <project>.group
# Concatenate all the separate fasta files into a single one for mothur input
Cat *expanded.fa > <project>.fa

Mothur Code

**Parallel analysis was performed on the bacterial and archaeal 16S rRNA gene amplicon libraries**

count.groups(group = <project>.group)
unique.seqs(fasta = <project>.fa)
count.seqs(name = current, group = current)
summary.seqs(count = current)
align.seqs(fasta = current, reference = silva.nr_v132_vamps.fasta)
summary.seqs(fasta = current, count = current)
filter.seqs(fasta = current, count = current)
summary.seqs(fasta = current, count = current)
unique.seqs(fasta = current, count = current)
summary.seqs(fasta = current, count = current)
classify.seqs(fasta = current, count = current, reference = silva.nr_v132_vamps.fasta, taxonomy = silva.nr_v132_vamps.tax)

Data Normalization

Bacteria

library(phyloseq)
library(ggplot2)
library(tidyverse)
library(vegan)

## generate OTU table to import
bac_abundance <- read.table("bms2017_bac_asv.expanded.unique.filter.count_table")
colnames(bac_abundance) <- as.character(unlist(bac_abundance[1,]))
bac_abundance = bac_abundance[-1, -2]

bac_abundance_rownames <- bac_abundance$Representative_Sequence

bac_otu <- bac_abundance %>%
  remove_rownames() %>%
  column_to_rownames(var = "Representative_Sequence") %>%
  select("DCO_LLO_Bv4v5--BRF1_BR170218_1"; "DCO_LLO_Bv4v5--VCS_VC170218")

# convert factors to numeric matrix
bac_otu <- sapply(bac_otu, function(x) as.numeric(as.character(x)))
# set rownames back to sample ID
row.names(bac_otu) <- bac_abundance_rownames

## generate taxa table to import
# expand taxon information columns and remove abundance
bac_taxa <- read.table("bms2017_bac_asv.expanded.unique.filter.unique.nr_v132.wang.taxonomy", stringsAsFactors = F)
colnames(bac_taxa) <- c("OTU", "taxonomy")

bac_taxa <- bac_taxa %>%
  separate(taxonomy, \c\(\'Domain\', \'Phyla\', \'Class\', \'Order\', \'Family\', \'Genus\', \'semi\'), ";"); extra = "merge") %>%
  select(Domain:Genus) %>%
  sapply(function(x) str_replace(x, \"\\(.*?\\)\", ""))

rownames(bac_taxa) <- bac_abundance_rownames

bac_name <- read.csv("L:/asv_analysis/bac/bac_name.csv", stringsAsFactors = F) %>%
column_to_rownames(var = "i..site")

# import data
OTU_b = otu_table(bac_otu, taxa_are_rows = TRUE)
TAX_b = tax_table(bac_taxa)
SAM_b = sample_data(bac_name)
physeq_bac = phyloseq(OTU_b, TAX_b, SAM_b)

# remove chloroplasts and mitochondria
physeq_bac_1 <- subset_taxa(physeq_bac,
  Domain == "Bacteria" &
  Family != "Mitochondria" &
  Order != "Chloroplast")
physeq_bac_1

#remove unknown sample S14 & s33
physeq_bac_2 <- subset_samples(physeq_bac_1, sample_name != "DCO_LLO_Bv4v5--S14_S_14" &
sample_name != "DCO_LLO_Bv4v5--S33_S_33")
physeq_bac_2

#remove low abundance reads from otu tables
physeq_bac_filter <- filter_taxa(physeq_bac_2, function(x) sum(x) > 4, TRUE)
physeq_bac_filter

#sanity check to make sure singletons removed
#check to see if taxa sums across all samples are 5+
bac_taxa_sum <- taxa_sums(physeq_bac_filter) %>%
data.frame()

bac_sum <- sample_sums(physeq_bac) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
bac_chloro_mito_remove_sum <- sample_sums(physeq_bac_1) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
bac_unknown_remove_sum <- sample_sums(physeq_bac_2) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
bac_filter_sum <- sample_sums(physeq_bac_filter) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
bac_contamination_sum <- bac_sum %>%
full_join(bac_chloro_mito_remove_sum, by = "Sample") %>%
full_join(bac_unknown_remove_sum, by = "Sample") %>%
full_join(bac_filter_sum, by = "Sample")

colnames(bac_contamination_sum) <- c("Sample", "total", "chloro_mito", "unknown", "filter")

bac_abundance_filter_sum <- bac_contamination_sum %>%
mutate(chloro_mito_remove = (total - chloro_mito)/total) %>%
mutate(total_remove = (total-filter)/total)
write.csv(bac_abundance_filter_sum, file =
"L:/asv_analysis/bac/CORRECTED/190314_bac_abundance_filter_library_sums.csv")

physeq_bac_3 <- transform_sample_counts(physeq_bac_filter, function(x) x/sum(x))

###CONTAMINATION SCREENING###

#list of potential contaminants
#contaminated genera

bms2017_bac_taxa <- data.frame(physeq_bac_filter@tax_table)

bms2017_asv_bac_shiek_contaminant <- bms2017_bac_taxa %>%
  rownames_to_column(var = "OTU") %>%
  #unite(taxon, c(Domain, Phyla, Class, Order, Family, Genus), remove = F) %>%
  filter(Genus %in% contamination_1) %>%
  select(OTU, Domain:Genus)
#5580 putative contaminants after abundance screening

bac_contamination_otu <- bms2017_asv_bac_shiek_contaminant$OTU

#filtered df as relative abundance
physeq_bac_filter_rel <- transform_sample_counts(physeq_bac_filter, function(x) x/sum(x))

#generate phyloseq object of putative contaminants from relative abundance physeq object
physeq_bac_contamination <- subset_taxa(physeq_bac_filter, Genus %in% contamination_1)
physeq_bac_contamination

physeq_bac_contamination_rel <- subset_taxa(physeq_bac_filter_rel, Genus %in% contamination_1)
physeq_bac_contamination_rel

#determine abundance of individual contaminant ASVs within each site
bac_contamination_rel <- veganotu(physeq_bac_contamination_rel) %>%
  t()%>%
data.frame()%>%
  rownames_to_column(var = "ASV") %>%
  #gather(-ASV, key = "Site", value = "rel") %>%
  #filter(rel > 0.001) %>%
  inner_join(bac_fasta_filter, by = c("ASV" = "seq_name"))
#write.csv(bac_contamination_rel, file = "L:/asv_analysis/bac/CORRECTED/bacterial_contaminants_within_site_abundance.csv")

#determine total abundance of contaminant ASVs in each sample
bac_contamination_site <- veganotu(physeq_bac_contamination_rel) %>%
  t() %>%
  data.frame() %>%
  rownames_to_column(var = "ASV") %>%
  gather(-ASV, key = "Site", value = "rel") %>%
  ungroup() %>%
  group_by(Site) %>%
  summarise(sum(rel))

#write.csv(bac_contamination_site, file = "L:/asv_analysis/bac/CORRECTED/bacterial_site_contaminantion_total_abundance.csv")

#determine total reads across all libraries
bac_total_reads <- physeq_bac_filter %>%
  sample_sums() %>%
  data.frame() %>%
  colSums()

bac_contamination_abun <- veganotu(physeq_bac_contamination) %>%
  t() %>%
  data.frame() %>%
  rowSums() %>%
  data.frame() %>%
  rownames_to_column(var = "ASV") %>%
  mutate(rel_abundance = ./bac_total_reads) %>%
  filter(ASV %in% bac_contamination_otu) %>%
  left_join(asv_shiek_blast, by = c("ASV"))

#contaminants abundance across all libraries
#write.csv(bac_contamination_abun, file = "L:/asv_analysis/bac/CORRECTED/bacterial_contaminants_within_dataset_abundance.csv")

median(bac_contamination_sum_1$filter) #84804 reads

#generate fasta file for export
bms2017_bac_taxa_filter <- data.frame(physeq_bac_filter@tax_table)

bac_otu <- rownames(bms2017_bac_taxa_filter)

library("Biostrings")

bac fasta <- readDNAStringSet("L:/asv_analysis/bac/fasta-1545400018724.fasta")
seq_name = names(bac fasta)
sequence = paste(bac.fasta)
bac.fasta.df <- data.frame(seq_name, sequence)

bac.fasta.df$seq.name <- gsub(':', '_', bac.fasta.df$seq.name)

#filter out singletons and chloroplasts
bac.fasta.filter <- bac.fasta.df %>%
  filter(seq.name %in% bac.otu)

#export fasta file to generate tree
library(seqRFLP)
dataframe2fas(bac.fasta.filter, file =
  "L:/asv_analysis/bac/CORRECTED/190214_bms2017_asv_bacteria_filtered.fasta")

#normalize read counts

total.b = median(sample.sums(physeq.bac.filter)) #74103
standf = function(x, t= total.b) round(t * (x / sum(x)))
physeq.bac_std = transform_sample_counts(physeq.bac.filter, standf)
physeq.bac_std

bms2017.bac.otu.filter_std <- data.frame(physeq.bac_std@otu_table)
bms2017.bac.taxa.filter_std <- data.frame(physeq.bac_std@tax_table)

#write.csv(bms2017.bac.taxa.filter_std, file =
# "L:/asv_analysis/bac/CORRECTED/190214_bms2017_bac_normalized_taxonomy.csv")
#write.csv(bms2017.bac.otu.filter_std, file =
# "L:/asv_analysis/bac/CORRECTED/190214_bms2017_bac_normalized_count.csv")

Archaea

library(phyloseq)
library(ggplot2)
library(tidyverse)

##generate OTU table to import
arc.abundance <- read.csv("bms2017.arc.count.csv")
colnames(arc.abundance)[1] <- "OTU"

arc.abundance.rownames <- arc.abundance$OTU

arc.otu <- arc.abundance %>%
  column_to_rownames(var = "OTU") %>%
  select(DCO_LLO.Av4v5..BRF1_BR170218_1:DCO_LLO.Av4v5..VCS_VC170218)

arc.otu <- arc.abundance[, -1]

#convert factors to numeric matrix
arc.otu <- sapply(arc.otu, function(x) as.numeric(as.character(x)))
# set rownames back to sample ID
row.names(arc_otu) <- arc_abundance_rownames

## generate taxa table to import
# expand taxon information columns and remove abundance
arc_taxa <- read.csv("bms2017_arc_tax.csv", stringsAsFactors = F)
colnames(arc_taxa) <- c("OTU", "taxonomy")

arc_taxa <- arc_taxa %>%
  separate(taxonomy, c('Domain', 'Phyla', 'Class', 'Order', 'Family', 'Genus', 'semi'), ';', extra = "merge")
%>%
  select(Domain:Genus) %>%
  sapply(function(x) str_replace(x, \"\(.*?\)\", ""))

rownames(arc_taxa) <- arc_abundance_rownames

arc_name <- read.csv("L:/asv_analysis/arc/arc_name.csv", stringsAsFactors = F) %>%
  column_to_rownames(var = "i..sample")

# import data
OTU_a = otu_table(arc_otu, taxa_are_rows = TRUE)
TAX_a = tax_table(arc_taxa)
GEO_a = sample_data(arc_name)
physeq_arc = phyloseq(OTU_a, TAX_a, GEO_a)
physeq_arc

# remove chloroplasts and mitochondria
physeq_arc_1 <- subset_taxa(physeq_arc, 
  Family != "Mitochondria" & 
  Order != "Chloroplast")

# remove QNF and PGF due to low read counts
# remove unknown sample S14
physeq_arc_2 <- subset_samples(physeq_arc_1, sample_name != "DCO_LLO_Av4v5..PGF_PG170224" & 
  sample_name != "DCO_LLO_Av4v5..QNF_QN170220" & sample_name != "DCO_LLO_Av4v5..S14_S_14")
physeq_arc_2

# remove low abundance ASVs from otu tables
physeq_arc_filter <- filter_taxa(physeq_arc_2, function(x) sum(x) > 4, TRUE)
physeq_arc_filter

# sanity check to make sure singletons removed
# check to see if taxa sums across all samples are 5+
arc_taxa_sum <- taxa_sums(physeq_arc_filter) %>%
data.frame()

arc_sum <- sample_sums(physeq_arc) %>%
data.frame()
rownames_to_column(var = "Sample")
arc_chloro_mito_remove_sum <- sample_sums(physeq_arc_1) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
arc_unknown_remove_sum <- sample_sums(physeq_arc_2) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
arc_filter_sum <- sample_sums(physeq_arc_filter) %>%
data.frame() %>%
rownames_to_column(var = "Sample")
arc_contamination_sum <- arc_sum %>%
  full_join(arc_chloro_mito_remove_sum, by = "Sample") %>%
  full_join(arc_unknown_remove_sum, by = "Sample") %>%
  full_join(arc_filter_sum, by = "Sample")
colnames(arc_contamination_sum) <- c("Sample", "total", "chloro_mito", "unknown", "filter")
arc_abundance_filter_sum <- arc_contamination_sum %>%
  mutate(chloro_mito_remove = (total - chloro_mito)/total) %>%
  mutate(total_remove = (total - filter)/total)
write.csv(arc_abundance_filter_sum, file =
  "L:/asv_analysis/arc/CORRECT/190314_arc_abundance_filter_library_sums.csv")

#generate fasta file for export
bms2017_arc_taxa_filter <- data.frame(physeq_arc_filter@tax_table)
arc_otu <- rownames(bms2017_arc_taxa_filter)

library("Biostrings")
arc_fasta <- readDNAStringSet("L:/asv_analysis/arc/fasta-1545402050342.fasta")
seq_name = names(arc_fasta)
sequence = paste(arc_fasta)
arc_fasta_df <- data.frame(seq_name, sequence)
arc_fasta_df$seq_name <- gsub(':', '_', arc_fasta_df$seq_name)
arc_fasta_filter <- arc_fasta_df %>%
  filter(seq_name %in% arc_otu)

#export fasta file to generate tree
library(seqRFLP)
dataframe2fas(arc_fasta_filter, file =
  "L:/asv_analysis/arc/CORRECT/190214_bms2017_asv_archaea_filtered.fasta")

#normalize read counts
total_a = median(sample_sums(physeq_arc_filter)) #96939.5
standf = function(x, t = total_a) round(t * (x / sum(x)))
physeq_arc_std = transform_sample_counts(physeq_arc_filter, standf)
physeq_arc_std

bms2017_arc_otu_filter_std <- data.frame(physeq_arc_std@otu_table)
bms2017_arc_taxa_filter_std <- data.frame(physeq_arc_std@tax_table)

write.csv(bms2017_arc_taxa_filter_std, file =
"L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_taxonomy.csv")
write.csv(bms2017_arc_otu_filter_std, file =
"L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_count.csv")

**Ordination Analysis & Figures**

*Bacteria*

library(phyloseq)
library(ggplot2)
library(tidyverse)
library(ggrepel)
library(gridExtra)

bac_otu <- read.csv("L:/asv_analysis/bac/CORRECTED/190214_bms2017_bac_normalized_count.csv",
stringsAsFactors = F) %>%
column_to_rownames(var = "X")

bac_tax <-
read.csv("L:/asv_analysis/bac/CORRECTED/190214_bms2017_bac_normalized_taxonomy.csv",
stringsAsFactors = F) %>%
column_to_rownames(var = "X") %>%
as.matrix()

bac_geo <- read.csv("L:/asv_analysis/bac/CORRECTED/bms2017_bac_geo.csv", stringsAsFactors = F)
%>%
column_to_rownames(var = "i..sample")

library(ape)
bac_tree <- read.tree(file = "L:/asv_analysis/bac/CORRECTED/bac_tree.tre")

#import data
OTU_b = otu_table(bac_otu, taxa_are_rows = TRUE)
TAX_b = tax_table(bac_tax)
GEO_b = sample_data(bac_geo)
physeq_bac = phyloseq(OTU_b, TAX_b, GEO_b, bac_tree)

bac_physeq_vegan <- veganotu(physeq_bac)

#generate weighted and unweighted unifrac matrices
bac_weighted_unifrac <- UniFrac(physeq_bac, weighted = TRUE)
bac_unweighted_unifrac <- UniFrac(physeq_bac, weighted = FALSE)
ordination and pull out datapoints for plots
bac_std_uni_ord <- ordinate(physeq_bac, method = "NMDS", distance = bac_unweighted_unifrac)
bac_std_uni_ord_points <- as.data.frame(bac_std_uni_ord$points)

bac_std_wuni_ord <- ordinate(physeq_bac, method = "NMDS", distance = bac_weighted_unifrac)
bac_std_wuni_ord_points <- as.data.frame(bac_std_wuni_ord$points)

###bioenv to select for best environmental parameters###
bac_geo_physeq <- data.frame(physeq_bac@sam_data)
bac_geo_num <- bac_geo_physeq[, sapply(bac_geo_physeq, class) == "numeric"]

##Unweighted##
#metas subset
bac_geo_meta <- bac_geo_num %>% select(1:7)
bac_meta_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_meta, use = "p")
bac_meta_bioenv

#aqueous isotopes
bac_geo_carbon <- bac_geo_num %>% select(8:11, 22, 23)
bac_carbon_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_carbon, use='p')
bac_carbon_bioenv

#sediment phase c and n and isotopes
bac_geo_total_cn <- bac_geo_num %>% select(12:21)
bac_total_cn_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_total_cn, use='p')
bac_total_cn_bioenv

#gas
bac_geo_gas <- bac_geo_num %>% select(24:32, 34, 35)
bac_gas_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_gas, use='p')
bac_gas_bioenv

#hydrocarbons subset
bac_geo_hydrocarbon <- bac_geo_num %>% select(55:57, 73)
bac_hydrocarbon_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_hydrocarbon, use='p')
bac_hydrocarbon_bioenv

#metals subset
bac_geo_metal <- bac_geo_num %>%
  select(80:88)

bac_metal_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_metal, use='p')
bac_metal_bioenv

#ions subset
bac_geo_ion <- bac_geo_num %>%
  select(89:99)

bac_ion_bioenv <- bioenv(bac_unweighted_unifrac, bac_geo_ion, use='p')
bac_ion_bioenv

##Weighted##

#metas subset
bac_meta_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_meta, use = "p")
bac_meta_bioenv_w

#aqueous isotopes
bac_carbon_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_carbon, use='p')
bac_carbon_bioenv_w

#sediment phase c and n and isotopes
bac_total_cn_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_total_cn, use='p')
bac_total_cn_bioenv_w

#gas
bac_gas_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_gas, use='p')
bac_gas_bioenv_w

#hydrocarbons subset
bac_hydrocarbon_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_hydrocarbon, use='p')
bac_hydrocarbon_bioenv_w

#metals subset
bac_metal_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_metal, use='p')
bac_metal_bioenv_w

#ions subset
bac_ion_bioenv_w <- bioenv(bac_weighted_unifrac, bac_geo_ion, use='p')
bac_ion_bioenv_w

#####EnvFit#####
##unweighted##

#select distance from trench, pH, temperature, and SPC
bac_geo_all_paper <- bac_geo_num[, c(3, 4, 6, 7)]
bac_all_uni_ord_envfit_paper <- envfit(bac_std_uni_ord, env = bac_geo_all_paper, na.rm = TRUE, permutations = 999, p.max = 0.0005)

#extracts relevant scores from envfit
env.scores.bac.all.paper <- as.data.frame(scores(bac_all_uni_ord_envfit_paper, display = "vectors"))
env.scores.bac.all.paper <- cbind(env.scores.bac.all.paper, env.variables = rownames(env.scores.bac.all.paper))

##weighted##

bac_all_wuni_ord_envfit_paper <- envfit(bac_std_wuni_ord, env = bac_geo_all_paper, na.rm = TRUE, permutations = 999, p.max = 0.0005)

env.scores.bac.w.all.paper <- as.data.frame(scores(bac_all_wuni_ord_envfit_paper, display = "vectors"))
env.scores.bac.w.all.paper <- cbind(env.scores.bac.w.all.paper, env.variables = rownames(env.scores.bac.w.all.paper))

###Bacterial Fluid Sample Top 10 ASVs###

#subset fluid samples
physeq_bac_fluid <- subset_samples(physeq_bac, sample_type == "F")

bac_fluid_vegan <- veganotu(physeq_bac_fluid)

#generate weighted and unweighted unifrac matrices
bac_fluid_weighted_unifrac <- UniFrac(physeq_bac_fluid, weighted = TRUE)
bac_fluid_unweighted_unifrac <- UniFrac(physeq_bac_fluid, weighted = FALSE)

#ordination
bac_fluid_uni_ord <- ordinate(physeq_bac_fluid, method = "NMDS", distance = bac_fluid_unweighted_unifrac)
bac_fluid_uni_ord_points <- as.data.frame(bac_fluid_uni_ord$points)

bac_fluid_wuni_ord <- ordinate(physeq_bac_fluid, method = "NMDS", distance = bac_fluid_weighted_unifrac)
bac_fluid_wuni_ord_points <- as.data.frame(bac_fluid_wuni_ord$points)

#calculate and extract species scores
bac_fluid_uni_wascores <- wascores(bac_fluid_uni_ord_points, bac_fluid_vegan)
bac_fluid_wuni_wascores <- wascores(bac_fluid_wuni_ord_points, bac_fluid_vegan)

#top 10 ASVs per fluid sample
#convert reads to relative abundance
physeq_bac_fluid_rel <- transform_sample_counts(physeq_bac_fluid, function(x) x/sum(x))
bac_fluid_vegan_rel <- veganotu(physeq_bac_fluid_rel)

bac_tax_1 <- bac_tax %>%
data.frame() %>%
rownames_to_column(var = 'ASV')

determine the top 10 ASVs per site
bac_fluid_top_10_site <- bac_fluid_vegan_rel %>%
t() %>%
data.frame() %>%
rownames_to_column(var = "ASV") %>%
gather(-ASV, key = "site", value = "rel_abund") %>%
group_by(site) %>%
top_n(10, rel_abund) %>%
ungroup() %>%
distinct(ASV)

calculate “coverage” of samples with the top 10 ASVs per site
bac_fluid_top_10_site_screened <- bac_fluid_vegan_rel %>%
t() %>%
data.frame() %>%
rownames_to_column(var = "ASV") %>%
inner_join(bac_fluid_top_10_site, by = "ASV") %>%
gather(-ASV, key = "site", value = "rel_abund") %>%
group_by(site) %>%
summarise(sum(rel_abund))

#extract top 10 ASVs/sample from unweighted species scores
bac_fluid_top10_scores <- bac_fluid_uni_wascores %>%
data.frame() %>%
rownames_to_column(var = "ASV") %>%
inner_join(bac_fluid_top_10_site, by = "ASV") %>%
inner_join(bac_tax_1, by = "ASV")

#extract top 10 ASVs/sample from weighted species scores
bac_fluid_top10_scores_w <- bac_fluid_wuni_wascores %>%
data.frame() %>%
rownames_to_column(var = "ASV") %>%
inner_join(bac_fluid_top_10_site, by = "ASV") %>%
inner_join(bac_tax_1, by = "ASV")

generate polygons generalizing area in ordination for each region
ordiplot(bac_fluid_uni_ord)
bac_fluid_ordihull <- ordihull(bac_fluid_uni_ord, bac_fluid_geo_physeq$region)
bac_fluid_ordi_north_arc <- bac_fluid_ordihull$Northern Arc %>%
```
data.frame()
bac_fluid_ordi_north_arc$region <- c("Northern Arc")
bac_fluid_ordi_north_forearc <- bac_fluid_ordihull$'Northern Forearc' %>%
data.frame()
bac_fluid_ordi_north_forearc$region <- c("Northern Forearc")
bac_fluid_ordi_south_forearc <- bac_fluid_ordihull$'Southern Forearc' %>%
data.frame()
bac_fluid_ordi_south_forearc$region <- c("Southern Forearc")
bac_fluid_ordi_south_arc <- bac_fluid_ordihull$'Southern Arc' %>%
data.frame()
bac_fluid_ordi_south_arc$region <- c("Southern Arc")
bac_fluid_ordi_south_volcano <- bac_fluid_ordihull$Volcano %>%
data.frame()
bac_fluid_ordi_points <- bac_fluid_ordi_north_arc %>%
  bind_rows(bac_fluid_ordi_north_forearc) %>%
  bind_rows(bac_fluid_ordi_south_arc) %>%
  bind_rows(bac_fluid_ordi_south_forearc) %>%
  bind_rows(bac_fluid_ordi_south_volcano)

#weighted fluid
ordiplot(bac_fluid_wuni_ord)
bac_fluid_ordihull_w <- ordihull(bac_fluid_wuni_ord, bac_fluid_geo_physeq$region)
bac_fluid_ordi_north_arc_w$region <- c("Northern Arc")
bac_fluid_ordi_north_forearc_w$region <- c("Northern Forearc")
bac_fluid_ordi_south_forearc_w$region <- c("Southern Forearc")
bac_fluid_ordi_south_arc_w$region <- c("Southern Arc")
bac_fluid_ordi_south_volcano_w$region <- c("Volcano")
bac_fluid_ordi_points_w <- bac_fluid_ordi_north_arc_w %>%
  bind_rows(bac_fluid_ordi_north_forearc_w) %>%
  bind_rows(bac_fluid_ordi_south_arc_w) %>%
  bind_rows(bac_fluid_ordi_south_forearc_w) %>%
  bind_rows(bac_fluid_ordi_south_volcano_w)

#### PLOTS####
```
#sample type, Figure 7
bac_unweighted_sample <- plot_ordination(physeq_bac, bac_std_uni_ord, shape = "sample_type", color = "sample_type", label = "Site_name", title = "Bacteria Unweighted Unifrac") +
  theme_bw() +
  theme(legend.position = "bottom") +
  geom_line(aes(group = Station), color = "black") +
  geom_point(size = 3) +
  scale_color_manual(name = "sample_type", values = c("navyblue", "red3"), labels = c("Filter", "Sediment"))

bac_weighted_sample <- plot_ordination(physeq_bac, bac_std_wuni_ord, color = "sample_type", shape = "sample_type", label = "Site_name", title = "Bacteria Weighted Unifrac") +
  theme_bw() +
  theme(legend.position = "bottom") +
  geom_line(aes(group = Station), color = 'black') +
  geom_point(size = 3) +
  scale_color_manual(name = "sample_type", values = c("navyblue", "red3"), labels = c("Filter", "Sediment"))

grid.arrange(bac_unweighted_sample, bac_weighted_sample, ncol = 2)

#region, Figure 9
bac_unweighted_region <- plot_ordination(physeq_bac, bac_std_uni_ord, shape = "sample_type", color = "region", label = "Site_name", title = "Bacteria Unweighted Unifrac") +
  theme_bw() +
  theme(legend.position = "bottom") +
  geom_line(aes(group = Station), color = "black") +
  geom_point(size = 3) +
  scale_color_manual(name = "region", values = c("red3","steelblue3", "darkerorange", "mediumseagreen", "black"), labels = c("Smooth, Northern Arc", "Smooth, Northern Outer Forearc", "Rough, Central Arc", "Rough, Central Outer Forearc", "Volcano")) +
  geom_segment(data=env.scores.bac.all.paper,aes(x=0,xend=NMDS1,y=0,yend=NMDS2, shape = NULL, color = NULL),
               color = "black", size = 0.5, linetype = 1, arrow = arrow(length = unit(0.1,"cm"))) +
  geom_text_repel(data = env.scores.bac.all.paper, aes(x = NMDS1, y = NMDS2, label=env.variables,
               shape = NULL, color = NULL),
               size = 4)

bac_weighted_region <- plot_ordination(physeq_bac, bac_std_wuni_ord, color = "region", shape = "sample_type", label = "Site_name", title = "Bacteria Weighted Unifrac") +
  theme_bw() +
  theme(legend.position = "bottom") +
  geom_line(aes(group = Station), color = 'black') +
  geom_point(size = 3) +
  scale_color_manual(name = "region", values = c("red3","steelblue3", "darkerorange", "mediumseagreen", "black"), labels = c("Smooth, Northern Arc", "Smooth, Northern Outer Forearc", "Rough, Central Arc", "Rough, Central Outer Forearc", "Volcano")) +
geom_segment(data=env.scores.bac.w.all.paper,aes(x=0,xend=NMDS1*0.5,y=0,yend=NMDS2*0.5, shape = NULL, color = NULL),
  color = "black", size = 0.5, linetype = 1, arrow = arrow(length = unit(0.1,"cm"))) +
geom_text_repel(data = env.scores.bac.w.all.paper, #labels the environmental variable arrows * "mult"
as for the arrows
  aes(x = NMDS1*0.5, y = NMDS2*0.5, label=env.variables, shape = NULL, color = NULL),
  size = 4)

grid.arrange(bac_unweighted_region, bac_weighted_region, ncol =2)

#species plots, Figure 11
bac_fluid_species_polygon_unweighted <- plot_ordination(physeq_bac_fluid, bac_fluid_uni_ord, color =
"region", title = "Fluid Unweighted Unifrac: Top 10 Bacterial ASV by Site") +
  theme_bw() +
  theme(legend.position = "bottom") +
  scale_color_manual(name = "region", values = c("red3","steelblue3", "darkorange",
  "mediumseagreen", "black"), labels = c("Smooth, Northern Forearc/Arc", "Smooth, Northern Outer
  Forearc", "Rough, Southern Forearc/Arc","Rough, Southern Outer Forearc", "Volcano")) +
  geom_polygon(inherit.aes = FALSE, data = bac_fluid_ordi_points, mapping = aes(x = NMDS1, y =
  NMDS2, fill = region)) +
  scale_fill_manual(name = "region", values = alpha(c("red3","steelblue3", "darkorange",
  "mediumseagreen", "black"), 0.35), labels = c("Smooth, Northern Forearc/Arc", "Smooth, Northern Outer
  Forearc", "Rough, Southern Forearc/Arc","Rough, Southern Outer Forearc", "Volcano")) +
  geom_point(inherit.aes = FALSE, data = bac_fluid_top10_scores, aes(x = MDS1, y = MDS2)) +#
  #geom_label_repel(inherit.aes = FALSE, data = bac_fluid_top10_scores, aes(x = MDS1, y = MDS2, label = Genus), size = 3)

bac_fluid_species_polygon_weighted <- plot_ordination(physeq_bac_fluid, bac_fluid_wuni_ord, color =
"region", title = "Bacterial Weighted Unifrac: Top 10 ASV by Site") +
  theme_bw() +
  theme(legend.position = "bottom") +
  scale_color_manual(name = "region", values = c("red3","steelblue3", "darkorange",
  "mediumseagreen", "black"), labels = c("Smooth, Northern Forearc/Arc", "Smooth, Northern Outer
  Forearc", "Rough, Southern Forearc/Arc","Rough, Southern Outer Forearc", "Volcano")) +
  geom_polygon(inherit.aes = FALSE, data = bac_fluid_ordi_points_w, mapping = aes(x = NMDS1, y =
  NMDS2, fill = region)) +
  scale_fill_manual(name = "region", values = alpha(c("red3","steelblue3", "darkorange",
  "mediumseagreen", "black"), 0.35), labels = c("Smooth, Northern Forearc/Arc", "Smooth, Northern Outer
  Forearc", "Rough, Southern Forearc/Arc","Rough, Southern Outer Forearc", "Volcano")) +
  geom_point(inherit.aes = FALSE, data = bac_fluid_top10_scores_w, aes(x = MDS1, y = MDS2)) +
  geom_label_repel(inherit.aes = FALSE, data = bac_fluid_top10_scores_w, aes(x = MDS1, y = MDS2, label = Genus), size = 3)

Archaea

library(phyloseq)
library(ggplot2)
library(tidyverse)
library(ggrepel)
library(gridExtra)

arc_otu <- read.csv("L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_count.csv", stringsAsFactors = F) %>%
  column_to_rownames(var = "X")

arc_tax <- read.csv("L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_taxonomy.csv", stringsAsFactors = F) %>%
  column_to_rownames(var = "X") %>%
  as.matrix()

arc_geo <- read.csv("L:/asv_analysis/arc/CORRECT/bms2017_arc_geo.csv", stringsAsFactors = F) %>%
  column_to_rownames(var = "i..")

# import tree
library(ape)
arc_tree <- read.tree(file = "L:/asv_analysis/arc/CORRECT/arc_tree.tre")

# import data
OTU_a = otu_table(arc_otu, taxa_are_rows = TRUE)
TAX_a = tax_table(arc_tax)
GEO_a = sample_data(arc_geo)
physeq_arc = phyloseq(OTU_a, TAX_a, GEO_a, arc_tree)

arc_physeq_vegan <- veganotu(physeq_arc)

# generate weighted and unweighted unifrac matrices
arc_weighted_unifrac <- UniFrac(physeq_arc, weighted = TRUE)
arc_unweighted_unifrac <- UniFrac(physeq_arc, weighted = FALSE)

# ordination
arc_std_uni_ord <- ordinate(physeq_arc, method = "NMDS", distance = arc_unweighted_unifrac)
arc_std_uni_ord_points <- as.data.frame(arc_std_uni_ord$points)

arc_std_wuni_ord <- ordinate(physeq_arc, method = "NMDS", distance = arc_weighted_unifrac)
arc_std_wuni_ord_points <- as.data.frame(arc_std_wuni_ord$points)

### bioenv to select for best environmental parameters###

arc_geo_physeq <- data.frame(physeq_arc@sam_data)
arc_geo_num <- arc_geo_physeq[, sapply(arc_geo_physeq, class) == "numeric"]

## Unweighted##

# metas subset
arc_geo_meta <- arc_geo_num %>%
  select(1:7)
arc_meta_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_meta, use = "p")
arc_meta_bioenv

#Best model has 2 parameters (max. 7 allowed):
# Longitude_E Distance_from_trench_km
#with correlation  0.4217006

#aqueous isotopes
arc_geo_carbon <- arc_geo_num %>%
  select(8:11, 22, 23)
arc_carbon_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_carbon, use='p')
arc_carbon_bioenv

#sediment phase c and n and isotopes
arc_geo_total_cn <- arc_geo_num %>%
  select(12:21)
arctotal_cn_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_total_cn, use='p')
arctotal_cn_bioenv

#gas
arc_geo_gas <- arc_geo_num %>%
  select(24:32, 34, 35)
arcgas_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_gas, use='p')
arcgas_bioenv

#hydrocarbons subset
arc_geo_hydrocarbon <- arc_geo_num %>%
  select(55:57, 73)
archydrocarbon_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_hydrocarbon, use='p')
archydrocarbon_bioenv

#metals subset
arc_geo_metal <- arc_geo_num %>%
  select(80:88)
arcmetal_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_metal, use='p')
arcmetal_bioenv

#ions subset
arc_geo_ion <- arc_geo_num %>%
  select(89:99)
arcion_bioenv <- bioenv(arc_unweighted_unifrac, arc_geo_ion, use='p')
arcion_bioenv
## Weighted##

### metas subset
```
arc_meta_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_meta, use = "p")
arc_meta_bioenv_w
```

### aqueous isotopes
```
arc_carbon_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_carbon, use='p')
arc_carbon_bioenv_w
```

### sediment phase c and n and isotopes
```
arc_total_cn_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_total_cn, use='p')
arc_total_cn_bioenv_w
```

### gas
```
arc_gas_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_gas, use='p')
arc_gas_bioenv_w
```

### hydrocarbons subset
```
ar arc_hydrocarbon_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_hydrocarbon, use='p')
ar c_hydrocarbon_bioenv_w
```

### metals subset
```
arc_metal_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_metal, use='p')
ar c_metal_bioenv_w
```

### ions subset
```
arc_ion_bioenv_w <- bioenv(arc_weighted_unifrac, arc_geo_ion, use='p')
ar c_ion_bioenv_w
```

### EnvFit###

### just distance from trench, temperature, pH, and SPC
```
ar c_geo_all_paper <- arc_geo_num[, c(3, 4, 6, 7)]
ar c_all Uni_ord_envfit_paper <- envfit(arc_std Uni_ord, env = arc_geo_all_paper, na.rm = TRUE, permutations = 999, p.max = 0.0005)
```

### extracts relevant scores from envfit
```
env.scores.arc.all.paper <- as.data.frame(scores(arc_all Uni_ord_envfit_paper, display = "vectors"))
env.scores.arc.all.paper <- cbind(env.scores.arc.all.paper, env.variables = rownames(env.scores.arc.all.paper))
```

```
ar c_all_wuni_ord_envfit_paper <- envfit(arc_std_wuni_ord, env = arc_geo_all_paper, na.rm = TRUE, permutations = 999, p.max = 0.0005)
env.scores.arc.w.all.paper <- as.data.frame(scores(arc_all_wuni_ord_envfit_paper, display = "vectors"))
```
env.scores.arc.w.all.paper <- cbind(env.scores.arc.w.all.paper, env.variables = rownames(env.scores.arc.w.all.paper))

####PLOTS####

#sample type, Figure 8
arc_unweighted_sample <- plot_ordination(physeq_arc, arc_std_uni_ord, shape = "sample_type", color = "sample_type", label = "Site_name", title = "Archaea Unweighted Unifrac") + theme_bw() + theme(legend.position = "bottom") + geom_line(aes(group = Station), color = "black") + geom_point(size = 3) + scale_color_manual(values = c("black", "red3"), labels = c("Filter", "Sediment")) + scale_shape_manual(values = c(19, 17))

arc_weighted_sample <- plot_ordination(physeq_arc, arc_std_wuni_ord, color = "sample_type", shape = "sample_type", label = "Site_name", title = "Archaea Weighted Unifrac") + theme_bw() + theme(legend.position = "bottom") + geom_line(aes(group = Station), color = "black") + geom_point(size = 3) + scale_color_manual(values = c("black", "red3"), labels = c("Filter", "Sediment")) + scale_shape_manual(values = c(19, 17))

grid.arrange(arc_unweighted_sample, arc_weighted_sample, ncol = 2)

#region, Figure 10
arc_unweighted_region <- plot_ordination(physeq_arc, arc_std_uni_ord, shape = "sample_type", color = "region", label = "Site_name", title = "Archaea Unweighted Unifrac") + theme_bw() + theme(legend.position = "bottom") + geom_line(aes(group = Station), color = "black") + geom_point(size = 3) + scale_color_manual(name = "region", values = c("red3","steelblue3","darkorange","mediumseagreen","black"), labels = c("Northern Arc","Northern Forearc","Southern Arc","Southern Forearc","Volcano")) + geom_segment(data = env.scores.arc.all.paper, aes(x=0,xend=NMDS1,y=0,yend=NMDS2, shape = NULL, color = NULL), color = "black", size = 0.5, linetype = 1, arrow = arrow(length = unit(0.1,"cm"))) + geom_text_repel(data = env.scores.arc.all.paper, aes(x = NMDS1, y = NMDS2, label = env.variables, shape = NULL, color = NULL), size = 4)

arc_weighted_region <- plot_ordination(physeq_arc, arc_std_wuni_ord, color = "region", shape = "sample_type", label = "Site_name", title = "Archaea Weighted Unifrac") + theme_bw() + theme(legend.position = "bottom") + geom_line(aes(group = Station), color = "black") +
geom_point(size = 3) +
scale_color_manual(name = "region", values = c("red3","steelblue3", "darkorange", "mediumseagreen", "black"), labels = c("Northern Arc", "Northern Forearc", "Southern Arc","Southern Forearc", "Volcano")) +
gemm_segment(data=env.scores.arc.w.all.paper,aes(x=0,xend=NMDS1*0.5,y=0,yend=NMDS2*0.5,
shape = NULL,
color = NULL),
color = "black", size = 0.5, linetype = 1, arrow = arrow(length = unit(0.1,"cm"))) +
gemm_text_repel(data = env.scores.arc.w.all.paper, aes(x = NMDS1*0.5, y = NMDS2*0.5,
label=env.variables, shape = NULL, color = NULL),
size = 4)
ggrid.arrange(arc_unweighted_region, arc_weighted_region, ncol =2)

Archaea Heatmap Analysis & Figure 15

library(phyloseq)
library(ggplot2)
library(tidyverse)

arc_otu <- read.csv("L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_count.csv",
stringsAsFactors = F) %>%
column_to_rownames(var = "X")

arc_tax <- read.csv("L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_taxonomy.csv",
stringsAsFactors = F) %>%
column_to_rownames(var = "X") %>%
as.matrix()

arc_geo <- read.csv("L:/asv_analysis/arc/CORRECT/bms2017_arc_geo.csv", stringsAsFactors = F) %>%
column_to_rownames(var = "X")

# import tree
#
library(ape)
#arc_tree <- read.tree(file = "arc_fasta_filter.tre")

# import data
OTU_a = otu_table(arc_otu, taxa_are_rows = TRUE)
TAX_a = tax_table(arc_tax)
GEO_a = sample_data(arc_geo)
physeq_arc = phyloseq(OTU_a, TAX_a, GEO_a)

arc_tax_1 <- arc_tax %>%
data.frame() %>%
rownames_to_column(var = "ASV")

# bar plot for paper
physeq_arc_rel <- transform_sample_counts(physeq_arc, function(x) x/sum(x))
arc_rel <- veganotu(physeq_arc_rel)

arc_rel_tidy <- arc_rel %>%
  t() %>%
  data.frame() %>%
  rownames_to_column(var = "ASV") %>%
  gather(-ASV, key = "Site", value = "Relative") %>%
  inner_join(arc_tax_1, by = "ASV")

arc_region <- data.frame(
  region = c("Northern Forearc", "Southern Forearc", "Northern Arc", "Southern Arc", "Volcano"),
  facet = c("Outer Forearc", "Outer Forearc", "Smooth, Northern Arc", "Rough, Central Arc", "Rough, Central Arc")
)

arc_geo_tidy <- arc_geo %>%
  rownames_to_column(var = "Site") %>%
  left_join(arc_region, by = 'region')

arc_rel_tidy_order <- arc_rel_tidy %>%
  group_by(Site, Order) %>%
  summarise(relative_order = sum(Relative)) %>%
  filter(relative_order > 0.01) %>%
  inner_join(arc_geo_tidy, by = "Site")

arc_order_1p <- arc_rel_tidy_order %>%
  ungroup() %>%
  distinct(Order) %>%
  lapply(as.character)

arc_rel_tidy_order_heat <- arc_rel_tidy %>%
  filter(Order %in% arc_order_1p$Order) %>%
  group_by(Site, Order) %>%
  summarise(relative_order = sum(Relative)) %>%
  #filter(relative_order > 0.01) %>%
  inner_join(arc_geo_tidy, by = "Site")

arc_order_heat <- arc_rel_tidy_order_heat %>%
  ggplot(aes(Site, Order)) +
  geom_tile(aes(fill = relative_order), colour = "white") +
  scale_fill_gradient(low = "white",high = "steelblue") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_discrete(expand = c(0, 0)) +
  theme(legend.position = "none",axis.ticks = element_blank(),axis.text.x = element_text(angle = 90, hjust = 1,size=8),axis.text.y = element_text(size=8)) +
  facet_grid(. ~ facet, drop=TRUE, scale="free",space="free") +
  theme(plot.subtitle = element_text(vjust = 1),
        plot.caption = element_text(vjust = 1),
axis.text.x = element_text(angle = 0),
plot.title = element_text(hjust = 0.5)) +
labs(title = "Relative Abundance of Archaea: Order Level",
  x = NULL) +
labs(y = "Order") +
theme(strip.text.x = element_text(size = 10)) + theme(axis.text.x = element_text(angle = 20))

### Adonis Statistics

**Bacteria**

```r
library(phyloseq)
library(ggplot2)
library(tidyverse)
library(ggrepel)
library(gridExtra)

bac_otu <- read.csv("L:/asv_analysis/bac/CORRECTED/190214_bms2017_bac_normalized_count.csv",
  stringsAsFactors = F) %>%
  column_to_rownames(var = "X")

bac_tax <-
read.csv("L:/asv_analysis/bac/CORRECTED/190214_bms2017_bac_normalized_taxonomy.csv",
  stringsAsFactors = F) %>%
  column_to_rownames(var = "X") %>%
  as.matrix()

bac_geo <- read.csv("L:/asv_analysis/bac/CORRECTED/bms2017_bac_geo.csv", stringsAsFactors = F)
%>
  column_to_rownames(var = "i..sample")

# #import tree
#
library(ape)

bac_tree <- read.tree(file = "L:/asv_analysis/bac/CORRECTED/bac_tree.tre")

#import data
OTU_b = otu_table(bac_otu, taxa_are_rows = TRUE)
TAX_b = tax_table(bac_tax)
GEO_b = sample_data(bac_geo)
physeq_bac = phyloseq(OTU_b, TAX_b, GEO_b, bac_tree)
physeq_bac

bac_physeq_vegan <- veganotu(physeq_bac)

#generate weighted and unweighted unifrac matrices
bac_weighted_unifrac <- UniFrac(physeq_bac, weighted = TRUE)
```
bac_unweighted_unifrac <- UniFrac(physeq_bac, weighted = FALSE)

#anosim
#analysis of similarity
#input dissimilarity matrix

#geochem table
bac_std_veg_geo <- data.frame(sample_data(physeq_bac))

#adonis call
bac_adonis_weighted_sample_type <- adonis(bac_weighted_unifrac ~ sample_type, data = bac_std_veg_geo)
bac_adonis_weighted_calcite_model <- adonis(bac_weighted_unifrac ~ calcite_model, data = bac_std_veg_geo)
bac_adonis_weighted_province <- adonis(bac_weighted_unifrac ~ Province, data = bac_std_veg_geo)
bac_adonis_weighted_arc_segment <- adonis(bac_weighted_unifrac ~ arc_segment, data = bac_std_veg_geo)
bac_adonis_weighted_region <- adonis(bac_weighted_unifrac ~ region, data = bac_std_veg_geo)
bac_adonis_weighted_trench <- adonis(bac_weighted_unifrac ~ Distance_from_trench_km, data = bac_std_veg_geo)
bac_adonis_weighted_temp <- adonis(bac_weighted_unifrac ~ temp_breaks, data = bac_std_veg_geo)
bac_adonis_weighted_ph <- adonis(bac_weighted_unifrac ~ pH_breaks, data = bac_std_veg_geo)

bac_adonis_unweighted_sample_type <- adonis(bac_unweighted_unifrac ~ sample_type, data = bac_std_veg_geo)
bac_adonis_unweighted_calcite_model <- adonis(bac_unweighted_unifrac ~ calcite_model, data = bac_std_veg_geo)
bac_adonis_unweighted_province <- adonis(bac_unweighted_unifrac ~ Province, data = bac_std_veg_geo)
bac_adonis_unweighted_arc_segment <- adonis(bac_unweighted_unifrac ~ arc_segment, data = bac_std_veg_geo)
bac_adonis_unweighted_region <- adonis(bac_unweighted_unifrac ~ region, data = bac_std_veg_geo)
bac_adonis_unweighted_trench <- adonis(bac_unweighted_unifrac ~ Distance_from_trench_km, data = bac_std_veg_geo)
bac_adonis_unweighted_temp <- adonis(bac_unweighted_unifrac ~ temp_breaks, data = bac_std_veg_geo)
bac_adonis_unweighted_ph <- adonis(bac_unweighted_unifrac ~ pH_breaks, data = bac_std_veg_geo)

bac_adonis_weighted_results <- bac_adonis_weighted_sample_type$aov.tab %>%
  rbind(bac_adonis_weighted_calcite_model$aov.tab) %>%
  rbind(bac_adonis_weighted_province$aov.tab) %>%
  rbind(bac_adonis_weighted_arc_segment$aov.tab) %>%
  rbind(bac_adonis_weighted_region$aov.tab) %>%
  rbind(bac_adonis_weighted_trench$aov.tab) %>%
  rbind(bac_adonis_weighted_temp$aov.tab) %>%
  rbind(bac_adonis_weighted_ph$aov.tab)
write.csv(bac_adonis_weighted_results, file =
"L:/asv_analysis/bac/CORRECTED/190328_bac_adonis_weighted_results.csv")

bac_adonis_unweighted_results <- bac_adonis_unweighted_sample_type$aov.tab %>%
  rbind(bac_adonis_unweighted_calcite_model$aov.tab) %>%
  rbind(bac_adonis_unweighted_province$aov.tab) %>%
  rbind(bac_adonis_unweighted_arc_segment$aov.tab) %>%
  rbind(bac_adonis_unweighted_region$aov.tab) %>%
  rbind(bac_adonis_unweighted_trench$aov.tab) %>%
  rbind(bac_adonis_unweighted_temp$aov.tab) %>%
  rbind(bac_adonis_unweighted_ph$aov.tab)
write.csv(bac_adonis_unweighted_results, file =
"L:/asv_analysis/bac/CORRECTED/190328_bac_adonis_unweighted_results.csv")

Archaean

library(phyloseq)
library(ggplot2)
library(tidyverse)
library(ggrepel)
library(gridExtra)

arc_otu <- read.csv("L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_count.csv",
  stringsAsFactors = F) %>%
  column_to_rownames(var = "X")

arc_tax <- read.csv("L:/asv_analysis/arc/CORRECT/190214_bms2017_arc_normalized_taxonomy.csv",
  stringsAsFactors = F) %>%
  column_to_rownames(var = "X") %>%
  as.matrix()

arc_geo <- read.csv("L:/asv_analysis/arc/CORRECT/bms2017_arc_geo.csv", stringsAsFactors = F) %>%
  column_to_rownames(var = "i..")

# import tree
library(ape)
arc_tree <- read.tree(file = "L:/asv_analysis/arc/CORRECT/arc_tree.tre")

# import data
OTU_a = otu_table(arc_otu, taxa_are_rows = TRUE)
TAX_a = tax_table(arc_tax)
GEO_a = sample_data(arc_geo)
physeq_arc = phyloseq(OTU_a, TAX_a, GEO_a, arc_tree)
physeq_arc

arc_physeq_vegan <- veganotu(physeq_arc)

# generate weighted and unweighted unifrac matrices
arc_weighted_unifrac <- UniFrac(physeq_arc, weighted = TRUE)
arc_unweighted_unifrac <- UniFrac(physeq_arc, weighted = FALSE)

# anosim
# analysis of similarity
# input dissimilarity matrix

# geochem table
arc_std_veg_geo <- data.frame(sample_data(physeq_arc))

# adonis call
arc_adonis_weighted_sample_type <- adonis(arc_weighted_unifrac ~ sample_type, data = arc_std_veg_geo)
arc_adonis_weighted_calcite_model <- adonis(arc_weighted_unifrac ~ calcite_model, data = arc_std_veg_geo)
arc_adonis_weighted_province <- adonis(arc_weighted_unifrac ~ Province, data = arc_std_veg_geo)
arc_adonis_weighted_arc_segment <- adonis(arc_weighted_unifrac ~ arc_segment, data = arc_std_veg_geo)
arc_adonis_weighted_region <- adonis(arc_weighted_unifrac ~ region, data = arc_std_veg_geo)
arc_adonis_weighted_trench <- adonis(arc_weighted_unifrac ~ Distance_from_trench_km, data = arc_std_veg_geo)

arc_adonis_unweighted_sample_type <- adonis(arc_unweighted_unifrac ~ sample_type, data = arc_std_veg_geo)
arc_adonis_unweighted_calcite_model <- adonis(arc_unweighted_unifrac ~ calcite_model, data = arc_std_veg_geo)
arc_adonis_unweighted_province <- adonis(arc_unweighted_unifrac ~ Province, data = arc_std_veg_geo)
arc_adonis_unweighted_arc_segment <- adonis(arc_unweighted_unifrac ~ arc_segment, data = arc_std_veg_geo)
arc_adonis_unweighted_region <- adonis(arc_unweighted_unifrac ~ region, data = arc_std_veg_geo)
arc_adonis_unweighted_trench <- adonis(arc_unweighted_unifrac ~ Distance_from_trench_km, data = arc_std_veg_geo)

arc_adonis_weighted_results <- arc_adonis_weighted_sample_type$aov.tab %>%
  rbind(arc_adonis_weighted_calcite_model$aov.tab) %>%
  rbind(arc_adonis_weighted_province$aov.tab) %>%
  rbind(arc_adonis_weighted_arc_segment$aov.tab) %>%
  rbind(arc_adonis_weighted_region$aov.tab) %>%
  rbind(arc_adonis_weighted_trench$aov.tab)
write.csv(arc_adonis_weighted_results, file = "L:/asv_analysis/arc/CORRECT/arc_adonis_weighted_results.csv")

arc_adonis_unweighted_results <- arc_adonis_unweighted_sample_type$aov.tab %>%
  rbind(arc_adonis_unweighted_calcite_model$aov.tab) %>%
  rbind(arc_adonis_unweighted_province$aov.tab) %>%
  rbind(arc_adonis_unweighted_arc_segment$aov.tab) %>%
  rbind(arc_adonis_unweighted_region$aov.tab) %>%
  rbind(arc_adonis_unweighted_trench$aov.tab)
rbind(arc_adonis_unweighted_trench$aov.tab)
write.csv(arc_adonis_unweighted_results, file =
"L:/asv_analysis/arc/CORRECT/arc_adonis_unweighted_results.csv")

Aqueous Geochemistry Analysis & Figures

Iron

library(phyloseq)
library(ggplot2)
library(tidyverse)

iron_data <- read.csv("L:/CELLFIES/190317_iron_analysis.csv") %>%
  mutate(iron_released = fe_acid - fe_unacid) %>%
  mutate(iron_released_per = iron_released/fe_acid) %>%
  mutate(iron_released_per = replace(iron_released_per, iron_released_per < 0, 0))

iron_data$region <- ordered(iron_data$region, levels = c("Northern Forearc/Arc", "Northern Outer Forearc", "Central Forearc/Arc", "Central Outer Forearc"))

#Figure 13
iron_data %>%
ggplot(aes(x = fe_acid, y = iron_released_per, color = region)) +
  theme_bw() +
  geom_point(size = 3) +
  scale_color_manual(name = "region", values = c("red3", "steelblue3", "darkorange", "mediumseagreen"), labels = c("Northern, Forearc/Arc", "Northern Outer Forearc", "Central Forearc/Arc", "Central Outer Forearc")) +
  xlab("Total Iron (umol/L)") +
  ylab("Ferric Iron (% of Total Iron)") +
  theme(legend.position = "bottom", legend.title = element_blank())

#Figure 12
iron_data %>%
ggplot(aes(x = distance, y = fe_acid, color = region)) +
  theme_bw() +
  geom_point(size = 3) +
  scale_color_manual(name = "region", values = c("red3", "steelblue3", "darkorange", "mediumseagreen"), labels = c("Northern, Forearc/Arc", "Northern Outer Forearc", "Central Forearc/Arc", "Central Outer Forearc")) +
  xlab("Distance from Trench (km)") +
  ylab("Total Iron (umol/L)") +
  theme(legend.position = "bottom", legend.title = element_blank())

cor.test(iron_data$pH, iron_data$fe_acid)
cor.test(iron_data$lat, iron_data$fe_acid, method = "pearson")
library(phyloseq)
library(ggplot2)
library(tidyverse)

arc_otu <- read.csv("L:/asv_analysis/arc/CORRECT/bms2017_arc_normalized_count.csv", stringsAsFactors = F) %>%
column_to_rownames(var = "X")

arc_tax <- read.csv("L:/asv_analysis/arc/CORRECT/bms2017_arc_normalized_taxonomy.csv", stringsAsFactors = F) %>%
column_to_rownames(var = "X") %>%
as.matrix()

arc_geo <- read.csv("L:/asv_analysis/arc/CORRECT/bms2017_arc_geo.csv", stringsAsFactors = F) %>%
column_to_rownames(var = "i..")

# import tree
library(ape)
arc_tree <- read.tree(file = "arc.fasta_filter.tre")

# import data
OTU_a = otu_table(arc_otu, taxa_are_rows = TRUE)
TAX_a = tax_table(arc_tax)
GEO_a = sample_data(arc_geo)
physeq_arc = phyloseq(OTU_a, TAX_a, GEO_a)
physeq_arc

arc_tax_1 <- arc_tax %>%
data.frame() %>%
rownames_to_column(var = "ASV")

physeq_arc_rel <- transform_sample_counts(physeq_arc, function(x) x/sum(x))

physeq_nitro <- subset_taxa(physeq_arc, Class == "Nitrososphaeria")
physeq_nitro_rel <- subset_taxa(physeq_arc_rel, Class == "Nitrososphaeria")

arc_geo_1 <- rownames_to_column(arc_geo, var = "sample")

nitro_rel <- sample_sums(physeq_nitro_rel) %>%
data.frame() %>%
rownames_to_column(var = "sample") %>%
inner_join(arc_geo_1, by = "sample")

#Figure 14
nitro_rel %>%
ggplot(aes(x = nh4, y = .)) +
theme_bw() +
geom_point() +
#geom_smooth(method = lm, se = FALSE) +
xlab("NH4 Concentration (mmol/L)") +
ylab("Relative Abundance of Nitrophaeria")

cor.test(nitro_rel$., nitro_rel$nh4, method = "spearman")
Appendix II – Single Cell Amplified Genome Workflow & Results
Methods

Sample Collection & Screening

Sediment preserved in glycerol TE for eventual preparation of single cell amplified genomes (SAGs) were thawed, diluted in PBS, and sonicated prior to immobilization on 0.2µm polycarbonate filters. Filters were stained with 5X SybrGold and mounted on glass slides with Vectashield Antifade Mounting Medium (Vector Laboratories) (102). All sediment samples were analyzed on an epifluorescence microscope to identify samples appropriate for cell-sorting.

Cell Sorting & Whole Genome Amplification at Bigelow

Samples from the Outer Forearc (Espabel) and Arc (Quebrada Naranja) were sent to Bigelow Single Cell Genomic Center for fluorescence activated cell sorting (FACS) and second-generation whole genome amplification (WGA-X) (49). Twenty-four WGA-X wells, fourteen from Espabel and 10 from Quebrada Naranja, were selected for library preparation and whole genome sequencing based on Cp values, which is the amount of time it took for detectable amplification-associated fluorescence. Cp values for the samples chosen for sequencing ranged from X to Y hours. Library preparation was performed using the Illumina Nextera XT DNA Library Preparation Kit. A magnetic bead clean-up was performed before sample quantification on a NanoDrop and an Agilent Technology 2100 Bioanalyzer. Paired-end reads with a length of 250bp were sequenced on an Illumina MiSeq at the UTK Genomics Core with a loading sample concentration of 10pM and 10% PhiX spike-in (56). All library preparation and sequencing was performed in the Next-Generation Sequencing Core at the University of Tennessee.

Bioinformatic Workflow

Downstream read processing was performing using the KBase platform (103). Illumina reads were trimmed using Trimmomatic and contigs assembled using SPADEs (104, 105). Contig quality was evaluated with CheckM and gene annotations performed with Prokka (106, 107). Taxonomy was evaluated via two different pipelines. Putative 16S rRNA gene sequences were input into the Basic Local Alignment Search Tool (BLAST) against the non-redundant nucleotide database to infer taxonomic identity of the SAGs (108). Additionally, putative 16S rRNA gene sequences were aligned using the SINA Aligner to identify nearest neighbor sequences from the SILVA database (60, 109). Multiple Sequence Analysis was performed using the MUSCLE algorithm in MEGA7 (110). The evolutionary history was inferred using the Neighbor-Joining method (111). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (112). The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site (113). Evolutionary analyses were conducted in MEGA7 (110).

Results

Based on microscopic screening of sediment samples, it was decided to send sediment samples from Quebrada Naranja (QN) and Espabel (EP) for SAG analysis at Bigelow because they contained a large amount of intact cells. The results of the WGA-X are shown in Figure 16 for Espabel and Figure 17 for Quebrada Naranja. Within each plate, the roughly 50% of the sorted cells were successfully amplified, as evidenced by a Cp value < 3:00. However, to maximize the chance of success with whole
genome sequencing, only the best samples were chosen based on those with the lowest Cp values. Samples for sequencing had Cp values of less than 1:40 minutes (Table 4).

Due to the untargeted approach utilized for cell-sorting, it was expected that SAGs would reflect the most abundant organisms at each site. Of the twenty-four WGA-X samples sequenced, only one sample (EP-15) failed to generate genomic assemblies, with twenty-two of the remaining genome being of bacterial lineage and one archaeal lineage, QN-7 (Table 4). Fifteen of the SAGs had either a full or partial 16S rRNA gene sequence present that was used to further taxonomically classify the SAG using the SILVA v132 database (Table 5, Figure 18). The diversity of intact microbial cells spans multiple microbial phyla and physiologies, including anaerobic, facultative lithotrophic, and alkaliphilic representatives. Sorting and sequencing of a putative acetoclastic methanogen SAG from a high methane site (BMS_QN7) may indicate the presence of a significant methanogen population.
Appendix III – Tables and Figures
Table 1 Site Classification and Geochemical Characteristics
For amplicon libraries, a B indicates a bacterial library was generated, A indicates an archaeal library was generated, NA indicates DNA was successfully extracted but a library could not be generated, and a minus (-) indicates no DNA could be extracted.

<table>
<thead>
<tr>
<th>Station</th>
<th>Site Name</th>
<th>Site Classification</th>
<th>Latitude (°N)</th>
<th>Longitude (°E)</th>
<th>Distance from trench [km]</th>
<th>Province</th>
<th>Temp (°C)</th>
<th>SPC (µS/cm)</th>
<th>pH</th>
<th>Fluid Amplicon Library</th>
<th>Sediment Amplicon Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES170215_1</td>
<td>Estrada River</td>
<td>9.899</td>
<td>-85.454</td>
<td>74.082</td>
<td>Outer Forearc</td>
<td>27.9</td>
<td>226.7</td>
<td>9.75</td>
<td>B</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>SR170216</td>
<td>Salitral el Rincon</td>
<td>10.254</td>
<td>-85.683</td>
<td>89.917</td>
<td>Outer Forearc</td>
<td>33</td>
<td>232</td>
<td>9.06</td>
<td>-</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>SM170216</td>
<td>Sabana Grande Spring</td>
<td>10.177</td>
<td>-85.480</td>
<td>97.884</td>
<td>Outer Forearc</td>
<td>31.8</td>
<td>310.5</td>
<td>9.25</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>QH170213_2</td>
<td>Quepos Hotsprings 2</td>
<td>9.562</td>
<td>-84.123</td>
<td>106.094</td>
<td>Outer Forearc</td>
<td>36.7</td>
<td>2692</td>
<td>8.69</td>
<td>BA</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>QH170213_1</td>
<td>Quepos Hotsprings 1</td>
<td>9.562</td>
<td>-84.123</td>
<td>106.118</td>
<td>Outer Forearc</td>
<td>48.7</td>
<td>4100</td>
<td>8.53</td>
<td>-</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>FA170219_1</td>
<td>Finca Ande Pool/Farm</td>
<td>10.337</td>
<td>-85.069</td>
<td>139.599</td>
<td>Forearc/Arc</td>
<td>55.2</td>
<td>5500</td>
<td>5.93</td>
<td>BA</td>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>SL170214</td>
<td>Santa Lucia Spring</td>
<td>10.291</td>
<td>-84.972</td>
<td>140.301</td>
<td>Forearc/Arc</td>
<td>57</td>
<td>2824</td>
<td>6.12</td>
<td>B</td>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>CY170214</td>
<td>Rio Cayuco Rock</td>
<td>10.287</td>
<td>-84.956</td>
<td>140.907</td>
<td>Forearc/Arc</td>
<td>72</td>
<td>5821</td>
<td>6.31</td>
<td>BA</td>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>MT170219</td>
<td>Mouse Trap Well</td>
<td>10.596</td>
<td>-85.238</td>
<td>152.554</td>
<td>Forearc/Arc</td>
<td>59.1</td>
<td>6150</td>
<td>6.32</td>
<td>BA</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>BQ170218</td>
<td>Borinquen Mud Pot</td>
<td>10.811</td>
<td>-85.414</td>
<td>159.488</td>
<td>Forearc/Arc</td>
<td>88.9</td>
<td>4980</td>
<td>2.11</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ST170223</td>
<td>Santa Theresa Pool</td>
<td>10.003</td>
<td>-83.828</td>
<td>164.110</td>
<td>Forearc/Arc</td>
<td>55.8</td>
<td>2725</td>
<td>4.51</td>
<td>-</td>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>PL170224</td>
<td>Poas Volcano lake</td>
<td>10.197</td>
<td>-84.230</td>
<td>164.522</td>
<td>Forearc/Arc</td>
<td>37.6</td>
<td>91900</td>
<td>0.85</td>
<td>-</td>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>HN170219</td>
<td>Las Hornillas Mud Pot</td>
<td>10.713</td>
<td>-85.177</td>
<td>167.413</td>
<td>Forearc/Arc</td>
<td>87.9</td>
<td>6051</td>
<td>1.82</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TC170221</td>
<td>El Tucano River</td>
<td>10.366</td>
<td>-84.381</td>
<td>172.669</td>
<td>Forearc/Arc</td>
<td>60</td>
<td>3500</td>
<td>6.24</td>
<td>B</td>
<td>BA</td>
<td></td>
</tr>
<tr>
<td>ET170220_1</td>
<td>Eco Thermales River</td>
<td>10.484</td>
<td>-84.676</td>
<td>173.293</td>
<td>Forearc/Arc</td>
<td>40</td>
<td>1911</td>
<td>6.06</td>
<td>BA</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>Station</td>
<td>Site Name</td>
<td>Site Classification</td>
<td>Latitude (°N)</td>
<td>Longitude (°E)</td>
<td>Distance from trench [km]</td>
<td>Province</td>
<td>Temp (°C)</td>
<td>SPC (µS/cm)</td>
<td>pH</td>
<td>Fluid Amplicon Library</td>
<td>Sediment Amplicon Library</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>---------------</td>
<td>----------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>-----------</td>
<td>-------------</td>
<td>-----</td>
<td>------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>QN170220</td>
<td>Quebrada Naranja</td>
<td>River</td>
<td>10.496</td>
<td>-84.697</td>
<td>173.682</td>
<td>Forearc/Arc</td>
<td>22.9</td>
<td>196.7</td>
<td>5.6</td>
<td>BA</td>
<td>B</td>
</tr>
<tr>
<td>BR170218_2</td>
<td>Blue River Spring 2</td>
<td>Spring</td>
<td>10.898</td>
<td>-85.329</td>
<td>173.684</td>
<td>Forearc/Arc</td>
<td>53.8</td>
<td>5755</td>
<td>5.87</td>
<td>B</td>
<td>BA</td>
</tr>
<tr>
<td>BR170218_1</td>
<td>Blue River Spring 1</td>
<td>Spring</td>
<td>10.898</td>
<td>-85.328</td>
<td>173.694</td>
<td>Forearc/Arc</td>
<td>59</td>
<td>5995</td>
<td>6.16</td>
<td>BA</td>
<td>BA</td>
</tr>
<tr>
<td>VC170218</td>
<td>Volcancito</td>
<td>Spring</td>
<td>10.898</td>
<td>-85.326</td>
<td>173.795</td>
<td>Forearc/Arc</td>
<td>59.8</td>
<td>6132</td>
<td>5</td>
<td>-</td>
<td>BA</td>
</tr>
<tr>
<td>RV170221</td>
<td>Recreo Verde</td>
<td>Well</td>
<td>10.322</td>
<td>-84.244</td>
<td>175.883</td>
<td>Forearc/Arc</td>
<td>42.7</td>
<td>45000</td>
<td>6.19</td>
<td>-</td>
<td>NA</td>
</tr>
<tr>
<td>PF170222</td>
<td>Pompilo’s Finca</td>
<td>Farm</td>
<td>10.518</td>
<td>-84.115</td>
<td>202.082</td>
<td>Backarc</td>
<td>28.7</td>
<td>3726</td>
<td>5.81</td>
<td>BA</td>
<td>B</td>
</tr>
</tbody>
</table>
Table 2 Summary Statistics for Amplicon Libraries Post-Processing
This table summarizes the number of ASVs (ie unique sequences) and total sequence reads in the entire dataset after each step of the ASV analysis pipeline. First, sequences classified as chloroplasts or mitochondria were removed, then low quality libraries were removed (ie those with <1000 reads). An abundance filter was applied to remove low abundance ASVs, which were defined as those with less than 5 reads across the entire dataset. Finally, the number of reads in each sample library was normalized to the median library size (indicated in the last row).

<table>
<thead>
<tr>
<th></th>
<th>Bacteria</th>
<th>Archaea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of ASVs</td>
<td>Total Number of Reads</td>
</tr>
<tr>
<td>Raw Data</td>
<td>1,217,085</td>
<td>4,315,644</td>
</tr>
<tr>
<td>Remove Chloroplasts/Mitochondria</td>
<td>1,201,947</td>
<td>4,280,851</td>
</tr>
<tr>
<td>Remove Low Quality Libraries</td>
<td>1,201,945</td>
<td>4,093,213</td>
</tr>
<tr>
<td>Remove Low Abundance ASVs</td>
<td>56,142</td>
<td>2,854,909</td>
</tr>
<tr>
<td>Median Library Size for Read Normalization</td>
<td>74,103</td>
<td></td>
</tr>
</tbody>
</table>
Table 3 Genus Richness of Bacterial and Archaeal 16S rRNA Gene Amplicon Libraries

The number of genera (i.e., genus richness) within each sample library was assessed as a crude measure of diversity. ASVs were agglomerated based on their taxonomic classification to the genera level.

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Bacteria</th>
<th></th>
<th>Archaea</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Filter</td>
<td>Sediment</td>
<td>Filter</td>
<td>Sediment</td>
</tr>
<tr>
<td>Blue River Spring 1</td>
<td>403</td>
<td>162</td>
<td>36</td>
<td>23</td>
</tr>
<tr>
<td>Blue River Spring 2</td>
<td>432</td>
<td>248</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Rio Cayuco</td>
<td>439</td>
<td>489</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>Espabel</td>
<td>469</td>
<td>406</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Estrada</td>
<td>437</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EcoThermales</td>
<td>378</td>
<td>412</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Finca Ande</td>
<td>419</td>
<td>210</td>
<td>44</td>
<td>34</td>
</tr>
<tr>
<td>Poas Background Sample</td>
<td></td>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>Mousetrap</td>
<td>303</td>
<td></td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Pompilo's Finca</td>
<td>600</td>
<td>455</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>Poas Laguna</td>
<td>420</td>
<td>340</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Poas Volcano Lake</td>
<td>207</td>
<td></td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Quepos Hotspring 1</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quepos Hotspring 2</td>
<td>137</td>
<td>135</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>697</td>
<td>418</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranchero el Salitral</td>
<td>405</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>El Sitio</td>
<td>324</td>
<td></td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Santa Lucia</td>
<td>564</td>
<td>266</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Santa Theresa</td>
<td>361</td>
<td></td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>El Tucano</td>
<td>566</td>
<td>184</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Volcancito</td>
<td>169</td>
<td></td>
<td>32</td>
<td></td>
</tr>
<tr>
<td><strong>Total Genera in Entire Dataset</strong></td>
<td><strong>1557</strong></td>
<td><strong>1557</strong></td>
<td><strong>121</strong></td>
<td><strong>121</strong></td>
</tr>
<tr>
<td>Sample</td>
<td>Plate</td>
<td>Well</td>
<td>Sample ID</td>
<td>Cp</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>------</td>
<td>-----------</td>
<td>-----</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>A02</td>
<td>QN-1</td>
<td>1:29</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>A03</td>
<td>QN-2</td>
<td>1:35</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>A09</td>
<td>QN-3</td>
<td>1:36</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>A11</td>
<td>QN-4</td>
<td>1:40</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>C04</td>
<td>QN-5</td>
<td>1:41</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>F23</td>
<td>QN-6</td>
<td>1:39</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>N22</td>
<td>QN-7</td>
<td>1:31</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>O21</td>
<td>QN-8</td>
<td>1:35</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>P11</td>
<td>QN-9</td>
<td>1:41</td>
</tr>
<tr>
<td>Quebrada Naranja</td>
<td>AH-250</td>
<td>P21</td>
<td>QN-10</td>
<td>1:37</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>A01</td>
<td>EP-11</td>
<td>1:24</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>A02</td>
<td>EP-12</td>
<td>1:32</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>B06</td>
<td>EP-14</td>
<td>1:34</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>B08</td>
<td>EP-15</td>
<td>1:31</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>C04</td>
<td>EP-16</td>
<td>1:33</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>C22</td>
<td>EP-17</td>
<td>1:33</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>I20</td>
<td>EP-18</td>
<td>1:34</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>L04</td>
<td>EP-19</td>
<td>1:34</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>M20</td>
<td>EP-20</td>
<td>1:26</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>N21</td>
<td>EP-21</td>
<td>1:30</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>O19</td>
<td>EP-22</td>
<td>1:27</td>
</tr>
<tr>
<td>Espabel</td>
<td>AH-253</td>
<td>P22</td>
<td>EP-24</td>
<td>1:31</td>
</tr>
<tr>
<td>Sample ID</td>
<td>Domain</td>
<td>Phylum</td>
<td>Class</td>
<td>Order</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>QN-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QN-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QN-3</td>
<td>Bacteria</td>
<td>Patescibacteria</td>
<td>Microgenomatia</td>
<td>Candidatus Levybacteria</td>
</tr>
<tr>
<td>QN-4</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Deltaproteobacteria</td>
<td>Syntrophobacterales</td>
</tr>
<tr>
<td>QN-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QN-6</td>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Ignavibacteria</td>
<td>Ignavibacteriales</td>
</tr>
<tr>
<td>QN-7</td>
<td>Archaea</td>
<td>Euryarchaeota</td>
<td>Methanomicrobia</td>
<td>Methanosarcinales</td>
</tr>
<tr>
<td>QN-8</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Deltaproteobacteria</td>
<td>Desulfuromonadales</td>
</tr>
<tr>
<td>QN-9</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Deltaproteobacteria</td>
<td>Syntrophobacterales</td>
</tr>
<tr>
<td>QN-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP-11</td>
<td>Bacteria</td>
<td>Nitrospirae</td>
<td>Thermodesulfovibrionia</td>
<td>uncultured</td>
</tr>
<tr>
<td>EP-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP-13</td>
<td>Bacteria</td>
<td>Nitrospirae</td>
<td>Thermodesulfovibrionia</td>
<td>uncultured</td>
</tr>
<tr>
<td>EP-14</td>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Sphingobacterales</td>
</tr>
<tr>
<td>EP-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP-16</td>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Sphingobacterales</td>
</tr>
<tr>
<td>EP-17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP-18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP-19</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Rhizobiales</td>
</tr>
<tr>
<td>EP-20</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Alphaproteobacteria</td>
<td>Sphingomonadales</td>
</tr>
<tr>
<td>EP-21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP-22</td>
<td>Bacteria</td>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Clostridiales</td>
</tr>
<tr>
<td>EP-23</td>
<td>Bacteria</td>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Chitinophagales</td>
</tr>
<tr>
<td>EP-24</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Deltaproteobacteria</td>
<td>Desulfobacterales</td>
</tr>
</tbody>
</table>
Figure 1 Cartoon Representation of a Terrestrial Geothermal Spring
Adapted from Schrenk, Brazelton, & Lang, 2013
Figure 2 Map of Costa Rica and Northern Panama. Google.
The EPR-CNS boundary is indicated by the orange box while the rough-smooth boundary, as separated by the Fisher Seamount, is denoted in yellow. Sites sampled during February 2017 field expedition are shown in green and red. Sites samples during a second field expedition in April 2018 are shown in blue.
Figure 3 Carbon concentration and $\delta^{13}$C of CO$_2$, DIC, DOC, and TOC Across Costa Rica Arc. Carbon concentrations (as indicated by circle size) and $\delta^{13}$C of CO$_2$ gas, dissolved inorganic carbon (DIC, purple), dissolved organic carbon (DOC, green), and total organic carbon in sediments surrounding the surface emanation of the springs (TOC, black) are shown vs. distance from trench. From Barry et al. 2019, in press.
Figure 4 Across and Along Arc Variation in Cell Abundance
A) Across arc, B) EPR-CNS boundary, and C) Rough-Smooth boundary
Figure 5 Relative Abundance of Chloroplast Reads in Bacterial 16S Amplicon Libraries
Figure 6 Total Photosynthetic Pigments in Hot Spring Sediments
Data generated by Dr. Elena Manini.
Figure 7 NMDS Ordination of UniFrac Distance Measure of Bacterial Sediments vs Fluids
Colors represent sample type: fluids in black and sediments in red.
Figure 8 NMDS Ordination of UniFrac Distance Measure of Archaeal Sediments vs Fluids
Colors represent sample type: fluids in black and sediments in red.
Figure 9 NMDS Ordination of UniFrac Distance Measures with Environmental Vectors
Colors represent four identified regional clusters: Smooth Outer Forearc (blue), Rough Outer Forearc (green), Smooth Forearc/Arc (red), and Rough Forearc/Arc (orange).
Figure 10 NMDS Ordination of Archaeal UniFrac Distance Measures with Environmental Vectors. Colors represent four identified regional clusters: Smooth Outer Forearc (blue), Rough Outer Forearc (green), Smooth Forearc/Arc (red), and Rough Forearc/Arc (orange).
Figure 11 NMDS with Species Points of Top 10 Bacterial ASVs per Fluid Sample
Colors represent four identified regional clusters: Smooth Outer Forearc (blue), Rough Outer Forearc (green), Smooth Forearc/Arc (red), and Rough Forearc/Arc (orange).
Figure 12 Iron Concentration Across the Costa Rica Volcanic Arc
Colors represent four identified regional clusters: Smooth Outer Forearc (blue), Rough Outer Forearc (green), Smooth Forearc/Arc (red), and Rough Forearc/Arc (orange). Iron data generated by Dr. Mustafa Yucel.
Figure 13 Percent of Total Iron Released Upon Acidification
Colors represent four identified regional clusters: Smooth Outer Forearc (blue), Rough Outer Forearc (green), Smooth Forearc/Arc (red), and Rough Forearc/Arc (orange). Iron data generated by Dr. Mustafa Yucel.
Figure 14 Abundance of *Nitrosphaeria* vs Ammonium Concentration

Ammonium concentration is not correlated with the abundance of *Nitrosphaeria* across the arc (Pearson cor = 0.324, p = 0.176; Spearman rho = 0.151, p = 0.538) [should read NH₄⁺?]). Ammonium data generated by Dr. Mustafa Yucel.
Figure 15 Order Level Heatmap of Archaeal Community Composition
Figure 16 Results of SAG Generation from Espabel (EP) Sediments

Diagram of 384-well plate generated from fluorescence-activated cell sorting (FACS) of Espabel sediment sample, followed by KOH lysis, and whole genome amplification via WGA-X. Color and time in each well correspond to the Cp value to reach detectable fluorescence, which is inversely correlated with quantity of DNA.
Figure 17 Results of SAG Generation from Quebrada Naranja (QN) Sediment Sample

Diagram of 384-well plate generated from fluorescence-activated cell sorting (FACS) of Quebrada Naranja sediment sample, followed by KOH lysis, and whole genome amplification via WGA-X. Color and time in each well correspond to the Cp value to reach detectable fluorescence, which is inversely correlated with quantity of DNA.
Figure 18 Neighbor Joining Tree of SAG 16S rRNA Genes with SILVA Nearest-Neighbors. 16S rRNA gene sequences from Espabel (◆ bms ep11-24) and Quebrada Naranja (◆ bms qn1-10) SAGs were aligned using the SINA Aligner to identify nearest neighbor sequences from the SILVA database. The evolutionary history was inferred using the Neighbor-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Evolutionary analyses were conducted in MEGA7.
Vita

Katherine Marie Fullerton was born and raised in the Garden State, New Jersey. Growing up near the Shore meant she was in the water before she learned to walk, and to this day is a self-proclaimed water-bug. Perfectly content to curl up in the bay window with a good book, it was no surprise to her parents that she loved to learn. She attended Biotechnology High School in Freehold, NJ, where her graduating class was the “last of the classics,” and she coined the school’s slogan: “Excellence is in our DNA.” Katherine paid her way through an undergraduate education at Rutgers University through a Presidential Scholarship and working as a recreational soccer referee and a residential networking consultant. As if her degree in microbial biotechnology was not enough to keep her busy, she was also an active member of the Douglass Residential College, a summer orientation leader, President of Designer Genes (the biotechnology student group), and a highly productive undergraduate researcher, spending 15+ hours a week in the lab. After graduating Summa cum Laude, she packed up her life into the back of a pickup truck and Cora, her Corolla, then made the 13 hour drive to Knoxville, TN where she would pursue graduate studies in microbiology. Little did she know that by the time she graduated with her Master’s degree three years later, she would spend 2 months of her time doing fieldwork in California, North Carolina, Costa Rica, and Panama and 2 years of her life screaming at a computer to “make the code work.” In her spare time, she moonlights as a cosplayer attending conventions along the East Coast to share her love to fandom through both her costumes and her panels on science in pop culture. While she may be the oddball in her family as the lone scientist amongst a group of mechanics, business owners, and bus drivers, Katherine’s family has always been supportive of her passion for the sciences. After graduation, Katherine hopes to find a job in either data science or undergraduate education, continuing to advocate for scientific literacy and communication.