Acknowledgements

Thank you to Dr. Jeffrey Blanchard for welcoming me into his lab and all his help and guidance over the past year. An additional thank you to all the members of the Blanchard Lab for their support and assistance.

Table of Contents:

List of Figures and Tables

List of Abbreviations

SOM: Soil Organic Matter
MAG: Metagenome Assembled Genome
BWA: Burrows–Wheeler Aligner
MEM: Maximal Exact Match
GTDB: Genome Taxonomy Database
SNP: Single Nucleotide Polymorphism

I. Introduction

With global temperatures predicted to rise anywhere from 2 to 5 degrees Celsius as a result of climate change, there will be great changes to the global ecosystem. One area that may be particularly affected is the soil ecosystem, which is the focus of soil warming experiments in the Barre Woods of the Harvard Forest in Petersham, Massachusetts. In this experiment, soil plots are consistently warmed 5 degrees above the surrounding temperature, replicating conditions under a warmer climate. Knowledge of how these temperatures affect microbial life will allow predictions to be made as to what changes are likely to occur in the future, having an impact on biodiversity and biogeochemistry.

In 2017, soil core samples from the Barre Woods warming plot within the Harvard Forest underwent cell sorting and sequencing. Early analysis of metagenomic and metatranscriptomic data from the bacterial population revealed that one phylum was disproportionately present in both areas. *Acidobacteria,* while not the most represented phylum identified in these samples, showed significantly higher expression than their more abundant counterparts. These bacteria, typically found in soil, have been found in a wide range of environments, able to withstand extreme weather, acidity, and metal-contamination (Barns et al., 2007). The phylum itself has been established only recently, yet members are frequently found to have a great deal of sequence diversity, resulting in the creation of 26 sub-groups within the phylum (Barns et al., 2007). *Acidobacteria* have been identified as microbial bioindicators, whose ubiquitous presence in soil ecosystems and sensitivity to temperature differences make them useful in a predictive capacity (Oliverio et al., 2017). Members of this phylum, however, have been difficult to culture

and therefore not much is known about what their role is in soil ecosystems, especially given the wide range of conditions in which members have been found.

Of the high-quality genomes acquired from the Barre Woods samples, twelve were in the *Acidobacteria* phylum, representing four different families. We aim to compare metagenomic data from these *Acidobacteria* under control and warmed conditions within the organic horizon of soil. We presume that there will be variants in these genomes as a result of adaptation to warmer temperatures. Previous studies suggest that *Acidobacteria* are involved in carbon and nitrogen metabolism, in addition to maintaining relationships with other soil bacteria such as *Proteobacteria* (Kielak et al., 2016). Any changes to these genes will likely have far-reaching effects. Therefore, understanding which genes are being affected and how can provide important insight into how global warming will affect other bacteria. Beyond the more traditional hypothesis-based approach to this research, we also hope to gain further knowledge about the function of *Acidobacteria* through this analysis. It is unclear why this phylum is so transcriptionally active within our data and in many other soil environments.

The Long-Term Ecological Research taking place in the Harvard Forest began in 1991, providing us with almost three decades of data related to changes in the microbial community within warmed soil. As such, there are numerous studies that have already revealed the effects of this warming. One estimate from these previous studies suggests that the microbial community is responsible for two-thirds of the carbon dioxide released from these soil plots thus-far, resulting in a 17% loss of soil carbon (Melillo et al., 2017). Clearly, understanding what is causing these changes at the microbial level will be key in both predictive and preventative efforts in response to a warming climate.

While previous research has already provided useful information, these studies have focused on microbial community responses as a whole. Given the enormous diversity harbored in the soil, it is difficult to tease out more subtle changes (Oliverio et al., 2017). Now, with metagenomic and metatranscriptomic data from the individual members of this community, we have the unique opportunity to analyze warming responses in finer detail. By focusing on *Acidobacteria,* a phylum which is present in a variety of environmental conditions, the responses we identify have the potential to be interpreted beyond the setting of the Harvard Forest, providing more specific estimations of change.

Additionally, the use of metagenomics in this study may allow us to make conclusions about the function of *Acidobacteria* that have not previously been possible. Due to the difficulty in culturing members of this phylum, studies of their physiological role have been limited. Samples from the Harvard Forest were subject to a mini-metagenomic approach. This method is termed 'mini' as microfluidics are used to reduce a large sample into smaller samples containing a defined number of cells. The cells are then sequenced, assembled, and binned within their sub-samples. Through the use of more individualized sequencing, the diversity within *Acidobacteria* and their subdivisions can be better understood, which may be a useful alternative while culturing is still unfeasible (Kielak et al., 2016).

II. Literature Review

Soils are predicted to contain 10^9 to 10^{10} microorganisms per dry gram (Eichorst et al., 2007). These microbes play an important role in the decomposition of soil organic matter (SOM), a carbon dioxide producing process that represents one of the greatest fluxes in the global carbon cycle (Schindlbacher et al., 2011). Microbial respiration, by some estimates,

results in up to 74% of total soil respiration (Melillo et al., 2011). Such soil microbial respiration releases around 60 pentagrams of carbon dioxide into the atmosphere (Karhu et al., 2014). With global temperatures predicted to rise 1.1°C to 6.4°C over the next 100 years, changes to microbial respiration rates are likely (Melillo et al., 2011). Increasing microbial respiration rates could have dramatic impacts on carbon stores within the soil. Soils and surface litter store 2 to 3 times as much carbon in the organic form as exists in the atmosphere (Scharlemann et al., 2014). This carbon comes from decaying vegetation, as well as microbial growth (Scharlemann et al., 2014). It is estimated that around 1500 pentagrams of carbon exist in the soil, two-thirds of which is organic while the remainder is inorganic (Scharlemann et al., 2014). With this in mind, understanding the effects of an increasing global temperature on soil microbes and the resultant indirect changes that follow, particularly on soil carbon pools, are necessary for predicting how the microbial world will react.

Previous studies of warming effects on soil have focused on a number of responses, one of which of course being carbon dioxide production. With temperature increases, decomposition of SOM is expected to increase, promoting a carbon dioxide efflux from the soil (Schindlbacher et al., 2011). One study found an increase in respiration within a plot heated for seven years of 8% (Melillo et al., 2011). While an increase in respiration has been seen in several cases, long-term studies have revealed that these initial increases diminish over time, likely due to adaptation by microbial community members (Melillo et al., 2017, Frey et al., 2008). Therefore, it is difficult to make definitive conclusions as to how respiration will be affected after many years.

There are differences in soil microbial response depending on the layer being studied. For the organic horizon, the top layer consisting of partially decomposed plant litter, there is a greater

response to environmental stressors than the next layer, the mineral soil, which holds less carbon (Pold et al., 2016). This response to warming in the organic horizon is made up of shifts in the functional and taxonomic makeup of the soil (Pold et al., 2016).

The effects of warming on microbial metabolism have also been observed. Several studies have identified an increase in bacterial metabolic activity as a result of rising temperatures, which in turn reduces the efficiency with which carbon is used (Schindlbacher et al., 2011, Melillo et al., 2017). The increase in microbial activity due to rising temperatures is likely to lead to a rapid turnover of SOM (Yang et al., 2017). Coupled with the soil drying that may accompany rising temperatures, substrate availability could be reduced as a result of the quick degradation of SOM and lower plant production, limiting microbial activity (Castro et al., 2010).

Genes encoding for carbohydrate degradation also undergo changes as a result of heating. Studies of three different plots in the Harvard Forest suggest that after 20 years of warming, expression of carbohydrate-active genes decreased, specifically in the organic horizon (Pold et al., 2016). Within the mineral layer, there was actually an increase in carbohydrate related genes (Pold et al., 2016). It is important to note that carbohydrate degrading enzymes are more abundant in the organic horizon, where there is greater decomposition, than in the mineral zone (Pold et al., 2016).

The relative abundance of soil bacteria has been demonstrated to change, with gram-negative bacteria decreasing with increasing temperature (Schindlbacher et al., 2011). Other studies have shown a similar decrease in abundance, but only under ambient carbon dioxide conditions (Castro et al., 2010). When temperature and carbon dioxide are both elevated,

bacterial abundance increases (Castro et al., 2010). In terms of total biomass, some studies identified a decrease in bacterial biomass (Melillo et al., 2017, Frey et al., 2008), while others saw no change in biomass (Schindlbacher et al., 2011, Melillo et al., 2017).

Other studies focused on changes in microbial gene expression as a result of soil warming. In one study of a nine-year warmed plot, there was an increase in genes encoding labile and recalcitrant C degradation, in addition to an increase in phosphorous and sulfur cycling genes (Xue et al., 2016). The shift towards increased recalcitrant carbon cycling suggests a decrease in the stability of soil carbon, causing further efflux (Xue et al., 2016). Nitrogen processing genes saw both an increase and a decrease in this study, potentially as a result of the focus on the microbial community as a whole rather than specific community members, which may have made it more difficult to make conclusions about warming effects (Xue et al., 2016).

In the majority of the above studies, the phylum *Acidobacteria* was frequently found to be one of the most abundant and responsive members of the soil microbial community. The first member of this phylum was identified in 1991 in an acid drain in Japan, as demonstrated by their name (Kielak et al., 2016). Currently split into 26 subdivisions, these bacteria have been further found in a diverse range of environments and can withstand extreme conditions including high acidity and metal-contamination (Barns et al., 2007). The most ubiquitous subdivisions are 1 and 3, which are incredibly versatile and typically found in soil environments (Eichrost et al., 2018). The more extremophilic subdivisions, 4, 8, 10 and 23, prefer thermophilic conditions (Eichrost et al., 2018). *Acidobacteria* are typically heterotrophs and are usually found in aerobic or microaerophilic environments, although some can survive anaerobically (Ward et al., 2009,

Kielak et al., 2016). Their genomes range in size from 4.9 to 6.7 megabases (Mb) with 68 to 76% protein coding sequences of predicted function (Eichrost et al., 2018).

Acidobacteria are second only to *Proteobacteria* in terms of abundance in soil environments, making up to 20% of the global microbial community (Janssen et al., 2006). Within these soil communities, subdivisions 1, 3, 4, and 6 are the most common (Eichorst et al., 2018). Due to their global presence, *Acidobacteria* likely make significant contributions to the carbon cycle (Ward et al., 2009). The specifics of their role, however, are still undefined. Much of the difficulty in studying this phylum is due to the challenges of culturing these bacteria (Jones et al., 2009). The few isolates that have been successfully cultured, which were only representative of 8 of the 26 subdivisions, did so under low pH and low carbon dioxide conditions (Jones et al., 2009, Fierer et al., 2007, Kielak et al., 2016), growing very slowly on low-nutrient media (Ward et al., 2009). The nutrients that are provided are non-traditional, with unique sources of carbon or complex polysaccharides (Kielak et al., 2016).

Given their slow growth, *Acidobacteria* are most abundant in older soil, appearing less in younger soils (Jones et al., 2009). They are assumed to be long-lived and to metabolize slowly, making them adaptable to changes in their environment (Ward et al., 2009). Due to their lower metabolism, they have been shown to favor lower C and lower nutrient environments (Castro et al., 2010), with their abundance negatively correlated with soil carbon access (Jones et al., 2009). As such, *Acidobacteria* are known as k-strategists, living near the carrying capacity of their environments when resources are limited (Fierer et al., 2007). Their abundance is also correlated with mild acidity, particularly for subdivision 1, explaining their prolific presence in soils, which tend to be mildly acidic (Eichorst et al., 2007).

Despite their slow growth, *Acidobacteria* are globally abundant, as noted earlier, raising the question as to how they are able to reach such abundance levels and outcompete faster growing species such as *Proteobacteria.* One possible explanation was noted in a strain of *Acidobacteria* that possesses a [NiFe]-hydrogenase, giving it the ability to consume hydrogen gas during starvation periods (Greening et al., 2015). In these persistence stages when carbon is at low concentrations, replication is halted and cells enter the stationary phase (Greening et al., 2015). This hydrogenase is upregulated, giving this *Acidobacteria* the ability to survive and adapt to carbon unavailability (Greening et al., 2015). The ability to enter a dormant state under environmental changes may explain why *Acidobacteria* are so common despite their slow growth.

Acidobacteria have also been shown to contain transposable elements, which have the ability to perform horizontal gene transfer between different strains of bacteria (Eichorst et al., 2018). By sharing metabolically important genes, members of this phylum have the ability to adapt and evolve in the long-term to changing environments and conditions (Eichorst et al., 2018). In addition to these mobile genetic elements, the presence of prophages was also found in acidobacterial genomes (Eichorst et al., 2018). Prophages and transposons have been demonstrated in *Escherichia coli* to encode their own metabolic genes which respond to environmental changes (Eichorst et al., 2018). The existence of both of these features within *Acidobacteria* could suggest that these bacteria also have this function, giving them better tools to adapt to unfavorable conditions.

Another possible explanation for their abundance is their ability to produce secondary metabolites. In two *Acidobacteria* studied, 12-14% of their genomes were found to be involved

in biosynthesis of these metabolites, primarily nonribosomal peptide synthetases and polyketide synthases which have a range of functions (Crits-Christoph et al., 2018). These enzymes produce antibiotics, antifungals, and immunosuppressants (Crits-Christoph et al., 2018). Their ability to produce chemicals and toxins, likely in defense against other microbes, suggests that the *Acidobacteria* studied here lead highly competitive lifestyles (Crits-Christoph et al., 2018). The transcriptional association that occurs between these metabolites, iron metabolism genes, and potential antimicrobial resistance genes also confirms their competitive lifestyle, particularly for iron sources and against competing microbes (Crits-Christoph et al., 2018). The presence of these metabolites may give the slow-growing *Acidobacteria* a better chance of surviving compared to their faster growing competitors.

Although there is still much to learn about the physiology and role of *Acidobacteria* in soils, prior studies have revealed some details. Due to the changing nature of soil environments, terrestrial *Acidobacteria* are more versatile than their extremophile counterparts, with more paralogous genes, suggesting the potential for adaptation to multiple conditions (Eichorst et al., 2018). In some, there are genes encoding for oxygen consumption at varying concentrations, as well as the ability to utilize inorganic and organic nitrogen sources (Eichorst et al., 2018). These genes may be ecoparalogues, which are paralogues that are expressed under different environmental conditions (Eichorst et al., 2018). A genomic and culture-based study of three *Acidobacteria* members identified that they make use of many carbon sources, ranging from simple sugars to complex molecules including hemicellulose, cellulose, and chitin (Ward et al., 2009). Their sugar transporters have low specificity and high affinity, making them ideal for low

nutrient environments and additionally allowing them to compete with species that cannot access these substrates at low concentrations (Ward et al., 2009).

Other physiological areas identified in Acidobacteria through genomic studies include iron metabolism and transporters beyond sugar (Kielak et al., 2016). While these areas have not yet been determined in culture, the genetic components of these functions have been identified (Kielak et al., 2016). In terms of iron metabolism, subdivisions 4 and 8 may potentially have the ability to uptake or scavenge for ferric iron (Kielak et al., 2016). In addition to iron uptake, transporters for amino acids, peptides, siderophores, and anions have been identified (Kielak et al., 2016).

Previous soil warming experiments have identified *Acidobacteria* to be very temperature responsive (Melillo et al., 2017). Within several prominent soil subdivisions (1, 4, 5, 6), their abundance has been shown to increase with elevated carbon dioxide and temperature, as well as with less precipitation (Castro et al., 2010). These results agree with previous work that demonstrated acidobacterial preference for oligotrophic niches, as dry environments and the related decline in plant growth results in decreased substrate availability (Castro et al., 2010). Their oligotrophic nature has also been supported by their low rRNA operon copy number, which typically indicates a preference for low-nutrient environments (DeAngelis et al., 2015). This is in contrast to copiotrophs, which possess many rRNA operon copies, indicative of a preference for a wide range of nutrients in high concentration (DeAngelis et al., 2015).

Members of *Acidobacteria* have also demonstrated an ability to break down recalcitrant C stores, which may, should their activity increase, cause the release of older recalcitrant carbon stores (DeAngelis et al., 2015). Within the organic horizon, polysaccharide associated genes

were unaffected by warming in one study, while there was a decrease in abundance of these genes in the mineral layer (Pold et al., 2016). They are also predicted to play an important role in rhizosphere carbon dynamics, near plant roots (Oliverio et al., 2017).

Due to their responsiveness to environmental changes, particularly in terms of abundance, *Acidobacteria* have been identified as an indicator species (DeAngelis et al., 2015, Oliverio et al., 2017). Their presence in most soil microbial communities and their identifiable changes in response to warming makes them useful for predicting the global microbial response to climate change (Oliverio et al., 2017). The family *Koribacteraceae,* within this phylum, holds 15 members which are temperature-responsive, 13 of which decrease in abundance with elevated temperature (Oliverio et al., 2017). *Acidobacteriaceae* and *Soilbacteres,* two other dominant families, have seven and six temperature responsive phylotypes, the majority of which are cold-responsive (Oliverio et al., 2017). Therefore, focusing on this clearly abundant indicator phylum on its own, as opposed to the microbial community overall, is likely to provide unique information as to the individual response of this phylum (DeAngelis et al., 2015). By understanding these individual responses, we can better predict how areas possessing members of this phylum will be affected.

III. Methods

Samples used in this analysis were prepared previously in the following manner. Soil cores were obtained from the Barre Woods warming plot within the Harvard Forest Long-Term Ecological Research Site in Petersham, Massachusetts in May 2017. The Barre Woods plot has been in operation since 2002. Warmed samples were exposed to heating cables within the soil

that brought the temperature of the soil to 5 degrees above the ambient temperature. Control samples were undisturbed.

Fourteen soil cores were taken, seven from subplots within the heated plots and seven from subplots within the control plots. The cores were separated into mineral zone samples and organic horizon samples, resulting in a total of 28 samples. These 28 samples underwent traditional bulk metagenomics and metatranscriptomics. Four samples of the 28 also underwent mini-metagenomics. Cells were prepared for fluorescence-activated cell sorting by undergoing sieving and filtering to remove soil contaminants. The samples were then fluorescently labeled with Sybr Green and sorted into groups of 100 cells for whole-genome amplification and sequencing. 359 mini-metagenomes were acquired and binned into metagenome assembled genomes (MAGs). Of the MAGs, those that were >50% complete, had <10% contamination, and <10% strain heterogeneity were classified as high-quality genomes, resulting in 200 MAGs. Within the high-quality MAGs, twelve were within the phylum *Acidobacteria,* eight of which were new species (Schulz et al., 2018, Alteio et al., 2020).

Our analysis of the metagenomic data began by mapping six of the organic zone samples, three control and three heated, to the 12 high-quality genomes acquired. The metagenomic data was previously filtered to remove adapter sequences and trim reads where quality was 0, in addition to removing reads with an average quality score under 3 and a minimum length less than or equal to 51. Common microbial contaminants were also filtered. For this analysis, reads were additionally using Trimmomatic v0.39 (Bolger et al., 2014), which scanned reads with a 4 base sliding window and cut when the average quality per base was under 20. Trimmed reads were then mapped to the 12 genomes using the Burrows–Wheeler Aligner (BWA) v0.7.17 (Li, 2013).

Specifically, the BWA maximal exact match (MEM) algorithm was used, which is better suited to longer reads than other aligners and has better performance. The default settings were used, which specify discarding a MEM with more than 10,000 occurrences in the genome and an error rate of 1.5%. BWA-MEM was completed by aligning each metagenomic sample to each reference genome. Therefore, a total of 72 variant files were created, with each of the 12 genomes having the 6 samples mapped to it. Variant calling was performed using bcftools v1.7, which detects single nucleotide polymorphisms (SNPs) (Li, 2011). Again, default settings were used and ploidy was set to 1 for bacterial data. The resulting SNPs were visualized using the Integrated Genomics Viewer (Robinson et al., 2011).

Using the Genome Taxonomy Database (GTDB), a tool which allows for the taxonomic definition of microbial isolates, a phylogenetic tree of the known Acidobacteria members was created (Parks et al., 2018). GTDB accesses genomes stored in RefSeq and GenBank and includes draft genomes of as of yet uncultured organisms. The GTDB toolkit accepts genome assemblies and assigns genomes to domains based on a set of 120 bacterial marker genes. After aligning marker genes, the genomes are placed into reference trees, resulting in the final phylogeny (Chaumeil et al., 2020). Visualization of the resulting phylogenetic tree was completed using the Interactive Tree of Life (iTOL) (Letunic and Bork, 2007).

Further analysis including statistical tests and quality assessment was primarily done in R (R Core Team, 2018) using the package vcfR (Knaus and Grünwald, 2017) and tidyverse (Wickham et al., 2019). Scripts developed in R were preserved in rmarkdown files, which include code chunks used to analyze and create visualizations of the obtained data in addition to comments about code function (Allaire et al., 2020). These files were then uploaded to the

Blanchard Lab's GitHub repository (https://github.com/OurMicrobiome). In the spirit of contributing to reproducible research, all code developed in this project will be shared on this repository for others to use.

IV. Results

The six metagenome files were first mapped to each of the 12 high-quality *Acidobacteria* genomes in order to obtain acidobacterial reads. Initially, the pipeline used created three separate files for each mapping. Optimizing this pipeline included reducing the number of produced files to preserve space. Following mapping and indexing, variant calling was performed by first calculating the coverage of reads against the reference genome before SNPs were called. The SNPs were then visualized at the whole genome and contig view (Figure 1). Each of the six metagenome samples were visualized against each reference genome, allowing visual comparison of the control and heated SNPs. Looking at just the largest contig, three genomes appear to show visible differences in SNP coverage between control and heated samples: 3300020924.fa.1, 3300020916.fa.5, and 3300020985.2. In 3300020924.fa.1, the last two heated tracks show less coverage in some areas. In 3300020916.fa.5, the first two heated tracks similarly have less coverage. For genome 3300020985.2, each of the heated tracks show much less coverage than the control tracks.

The quality of the resulting SNPs was determined in R using the vcfR package. The read depth, mapping quality, Phred scaled quality, and variants per site for the genome 3300020651.fa.6 under the two conditions are shown in Figure 2. For both the control and heated samples, read depth is heavily concentrated around 10 reads per SNP, with the highest read depth

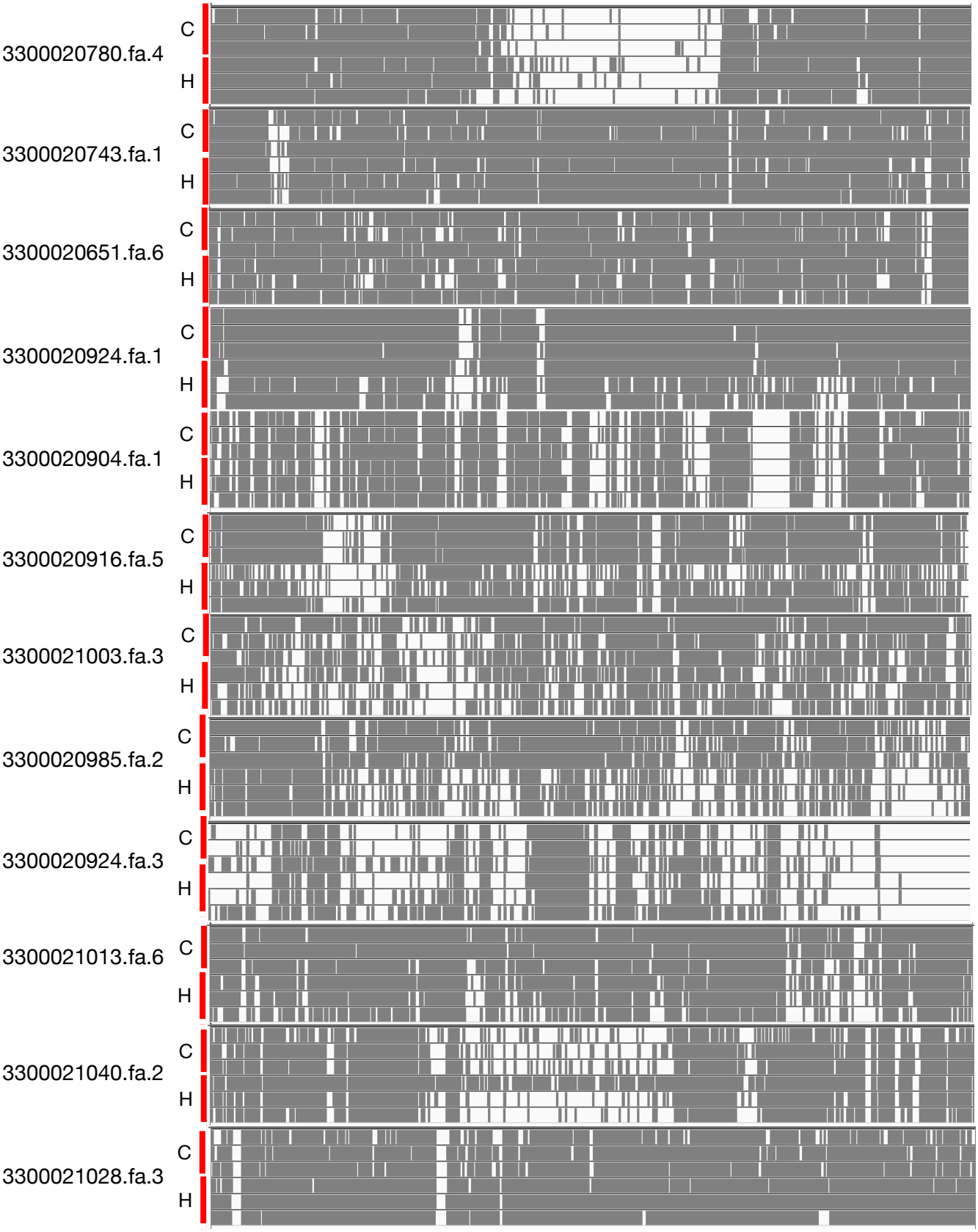

Figure 1. SNP Profiles under Control and Heated Conditions

For each genome studied, the SNP profile for each of the six metagenome samples was visualized with Integrated Genomics Viewer. The largest contig of each genome is shown here. The first 3 tracks (C) represent control samples and the last 3 tracks (H) represent heated samples. Gray areas represent sites with SNPs while white areas represent sites with no SNPs.

reaching almost 800. The mapping quality ranges from around 15 to the maximum level of 60 for both and Phred scaled quality ranges from 0 to around 240.

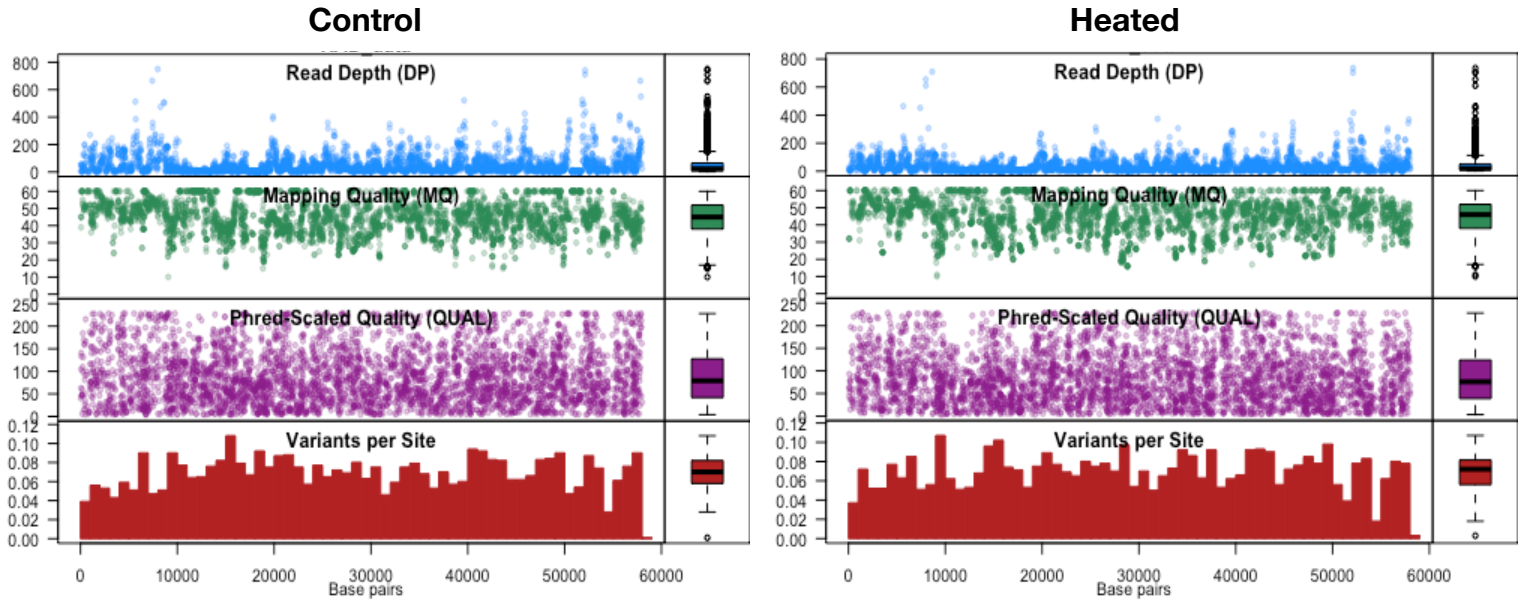

Figure 2. SNP Quality Report for 3300020651.fa.6 under Control and Heated Conditions
Using the vcfR package in R, the read depth, mapping quality, Phred scaled quality, and variants per site were reported for genome 3300020651.fa.6. Read depth states the number of times a read appeared in the sample, mapping quality states the confidence that a read is correctly aligned on a log scale, the Phred scaled quality is the probability of an alternative call existing at a site, and the variants per site gives the number of variants per SNP site.

Following variant calling, the SNPs per genome were recorded and the average for the control and heated samples computed (Table 1). A student t-test was performed to determine which samples were significantly different. Only one sample had a significant difference between control and heated, 3300020924.fa.1, which had a p-value of 0.02. Several other samples had p-values under 0.1, including 3300021028.fa.3, 3300020904.fa.1, and 3300021003.fa.3. There was a great deal of variance within the treatment samples, such as in samples mapped to genomes 3300020916.fa.5 and 3300020780.fa.4 (Figure 3).

A phylogenetic tree containing all known members of the phylum *Acidobacteria* was created using GTDBtk (Figure 4). The 12 highlighted genomes represent those used as reference

Table 1. SNP Averages under Control and Heated Conditions

Genome	Control Average	Heated Average	P-Value
3300021028.fa.3	92666	139857	0.095
3300020651.fa.6	115159	100837	0.203
3300020916.fa.5	226757	136115	0.156
3300020743.fa.1	75684	78715	0.913
3300020780.fa.4	101977	102534	0.986
3300020904.fa.1	95160	84769	0.088
3300021040.fa.2	90018	100709	0.653
3300021013.fa.6	127429	104285	0.156
3300020924.fa.1	208099	119014	0.002
3300020985.fa.2	47278	38377	0.204
3300021003.fa.3	58896	34855	0.067
3300020924.fa.3	63825	62971	0.757

Control and Heated n=3

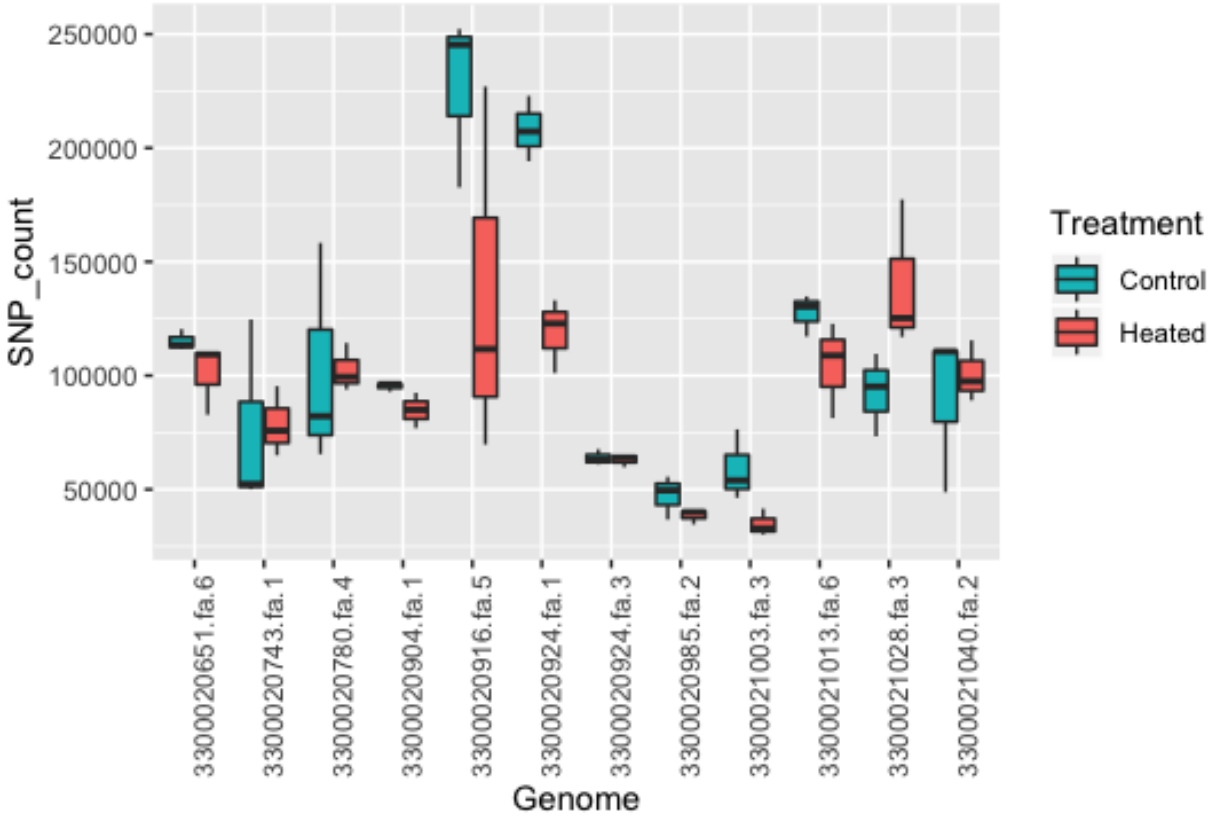

Figure 3. SNP Counts under Control and Heated Conditions

For each of the 12 high quality genomes, the SNP counts under control and heated conditions were recorded. Control samples are labeled in blue and heated samples are labeled in pink.

genomes in this study. 3300020924.fa.1, which demonstrated significant change between heated and controlled conditions, is labeled with a red asterisk. It grouped closely with 3300021003.fa.3. Other genomes which clustered together include 3300020743.fa.1, 3300020780.fa.4, and 3300021040.fa.2 as well as 3300021028.fa.3 and 3300020916.fa.5.

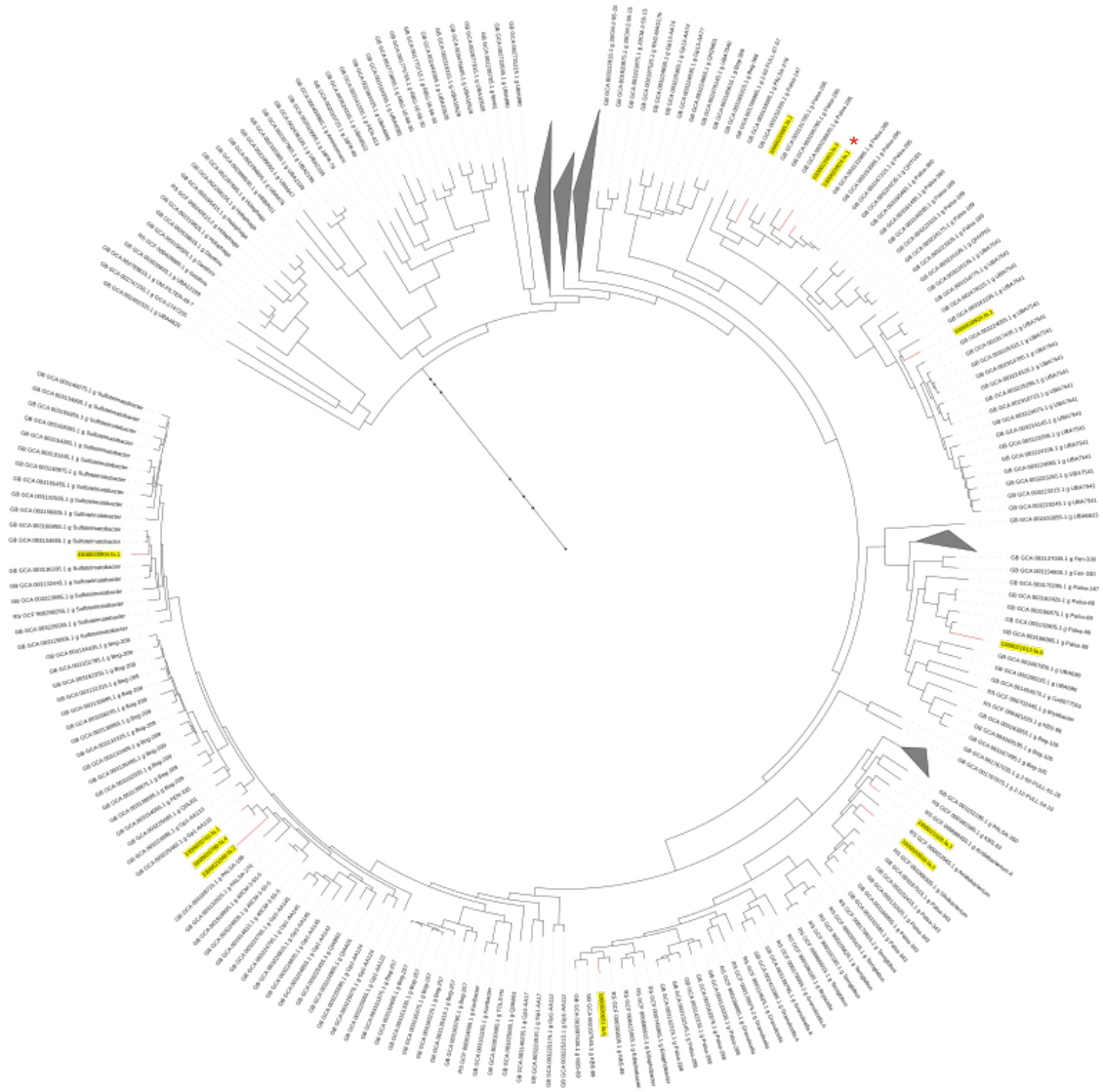

Figure 4. Phylogenetic Tree of all known *Acidobacteria* Genomes
After creation using GTDB toolkit, the resulting phylogenetic tree was visualized with iTOL. All known members of *Acidobacteria* are included with the high-quality reference genomes used in this study highlighted in yellow. The genome which experienced a significant change in SNP counts between control and heated treatment is identified with a red asterisk. Collapsed nodes are shown as gray triangles. (Image created by Alexander Truchon)

V. Discussion and Future Directions

While definitive conclusions cannot be made looking at the SNP profiles for each of the 12 genomes, it is possible to identify genomes that may have significant differences between control and heated samples (Figure 1). Of the three genomes that showed visually distinct profiles, 3300020924.fa.1, 3300020916.fa.5, and 3300020985.2, only 3300020924.fa.1 had a significant difference between control and heated samples (Table 1). Among all three, the overall SNP count differences ranged from almost 100,000 to under 10,000. This disparity is likely due to the fact that these profiles only represent the largest contig of each genome, so the visible changes, or lack therefore, may not appear on the other contigs. Therefore, it will be necessary to look at these changes both up close, to see in what areas the SNPs are occurring, and at the larger scale of the entire genome to compare overall SNP differences in order to make any definite conclusions. It is also important to mention that we are strictly comparing SNP counts and coverage, not location of SNPs and potential functional changes, which may account for more variance between treatments than simply the number of SNPs.

Looking at the quality assessment of one sample, 3300020651.fa.6, no reads had a mapping quality below 6 (Figure 2). Mapping quality is based on a log-based scale from 0 to 60, where 0 indicates a read is not uniquely mapped and 60 indicates the read is mapped in only one location. The score also represents the probability that a read is mapped to wrong position, based on the equation $10^{\frac{-Q}{10}}$ where Q indicates the mapping quality. The Phred scaled quality, which represents the probability that an alternative call exists, uses a similar equation, $10^{\frac{-(1-Q)}{10}}$, where Q represents Phred quality. Filtering reads below a specific read depth and selecting a probability cutoff for mapping and Phred scaled quality will likely assist with identifying significant SNPs.

Assessing the quality of the other samples may also reveal a natural cutoff in terms of read depth and quality that is consistent amongst all samples.

The SNP counts for each genome had a great deal of variation, with a maximum average SNP count reaching 226,757 and a minimum average count of 34,855 (Table 1). Only one genome showed a significant difference between control and heated SNP counts, 3300020924.fa.1, which had a p-value of 0.002. There is a marked difference between control and heated, with a decrease of about 100,000 SNPs, likely explaining its significance and giving us an acidobacterial strain that looks to be temperature responsive. This genome was not the only one with a drastic difference between control and heated counts; 3300020916.fa.5 also had almost a 100,000 SNP decrease. The p-value, however, reveals that there was clearly variance within the treatment groups (Figure 3), likely due to the small sample size used in this analysis. While we can only conclude here that there was a significant difference in SNP counts in one genome, several other genomes will benefit from further study, specifically 3300021028.fa.3, 3300020904.fa.1, and 3300021003.fa.3. Each of these genomes had p-values under 0.1 and show little to no overlap between their respective box-plots. These may represent species that are in fact adapting to increased temperature, once the remaining samples under each treatment are included in the analysis. There were several genomes with samples that clearly had very similar SNP counts and relatively less variation than other samples, such as 3300020924.fa.3, 3300020780.fa.4, 3300020743.fa.1, which could indicate species of *Acidobacteria* that are unaffected by warming. Including the remaining 22 samples in future analysis may reduce the observed variance, revealing other genomes that showed significant change between the treatments.

Observing the evolutionary relationships between these genomes, one of the nearest relatives to our genome with demonstrated significance (red asterisk), 3300021003.fa.3, had the second lowest p-value of all other genomes with 0.067 (Figure 4, Table 1). This may indicate that other closely related genomes will respond to the heated treatment and would make interesting candidates for future study. Both these genomes responded to heating with a decrease in SNP counts. On the other hand, the largest other grouping of genomes, 3300020743.fa.1, 3300020780.fa.4, and 3300021040.fa.2, make up some of the three largest p-values of all genomes (0.913, 0.985, and 0.653, respectively). These results may indicate that this clade is not temperature responsive, which is consistent with the minimal difference in SNP count between mean control and heated samples. The third visible clustering of genomes, 3300021028.fa.3 and 3300020916.fa.5, had very similar p-values of 0.095 and 0.156, respectively. Notably, the SNP count for 3300021028.fa.3 increased in response to heat treatment while the counts for 3300020916.fa.5 decreased due to treatment. If these genomes are in fact temperature responsive, they do so in opposite ways.

Beyond including the additional samples in the future and filtering based on quality, the SNPs should be annotated in order to specify where these variants are having an effect and determine the function of the genes in which SNPs are located. The number of SNPs per each gene will also aid our understanding of what changes are taking place in these genomes and perhaps identify which genes are reacting to the pressure resulting from increased temperature. Additionally, it is important to ensure that these results are reproducible. The mappings should be repeated in order to have confidence in the SNPs identified. Future mappings with BWA-MEM may also make use of stricter or different parameters, such as mismatch penalty and seed-length,

as the default settings were used here. It may also be beneficial to map each metagenomic sample to all reference genomes at once, rather than individually, in order to find the reads which best match each genome. The resulting SNPs will be specific to each genome, eliminating potential repetition of SNPs in each of the references. This may provide greater clarity as to which SNPs are most likely to be present in each of the genomes, allowing for more focused prediction of what genes are being affected by warming.

It may also be clarifying to calculate the average nucleotide identity between the reference genomes in order to understand how similar closely clustered genomes may be. For example, determining the average nucleotide identity between genomes 3300021028.fa.3 and 3300020916.fa.5 which are relatively close in our phylogeny but seem to have opposite responses to heating may help explain why this disparity in response is seen. Average nucleotide identity can also be used to conclude that similar responses, such as between 3300020924.fa.1 and 3300021003.fa.3 which both experienced an increase in SNP count, are as a result of their genetic closeness.

An assumption made in this study was that the 12 high-quality *Acidobacteria* genomes acquired from mini-metagenomics are the only members of this phylum sampled. It could be the case that other members of *Acidobacteria* were present and not detected. Including other environmental genomes from this phylum in our reference dataset could result in better read mapping and SNP detection. It would also allow us to compare the response of closely related genomes and determine if there are clade-specific responses to warming, such as an increase or decrease in SNP count or resistance to temperature changes. In the case of the closely related genomes with opposite responses to heat treatment, having information about the response of

their relatives may clarify the effect of heating on this greater clade. With this knowledge, we can develop a finer understanding of the differences in acidobacterial temperature sensitivity and further define the still not well-known subdivisions of *Acidobacteria*.

Identifying which subdivisions of *Acidobacteria* these 12 genomes belong to may also reveal more about their potential function and what their response will be to climate change. It may also serve us to explore the taxonomic relationships between the 12 genomes studied here. While this phylum is still not well understood, we do have some information about family-level responses to temperature change. Exploring whether these 12 genomes respond similarly to their own families or closely related families could reveal more about the diversity in function and physiology within *Acidobacteria*.

Pursuing a similar workflow with the metatranscriptomic data by mapping to the 12 high-quality genomes and identifying which genes are differentially expressed between the two conditions will add another layer of detail to our knowledge as to how *Acidobacteria* are affected by warming temperatures. While this and other related analysis will reveal how these bacteria are adapting over the long-term, gene expression provides more real-time information about how these microbes are responding in the moment and may specify the genes which are crucial to a heat response.

Overall, these results only indicate what occurs at the organic horizon of the soil. There may be differences in the SNP counts between the treatments, but more samples need to be included in order to confirm this, including samples from the mineral zone of the soil. The organic horizon, while more involved in degradation of SOM and more responsive to environmental changes than the mineral layer, can only reveal so much. Changes within the

mineral zone may indicate a more long-term adaptive response, as the mineral horizon is more involved in positive-feedback to climate, making it necessary to continue this analysis with such samples (Pold et al., 2016). Although it is difficult to make conclusions without a higher resolution understanding of which genes are being affected and what the roles of these genes are, there is some indication that *Acidobacterial* genomes are in fact adapting in response to warming. As we know from previous studies, bacteria and the greater microbial community will likely be affected by a rise in global temperatures, a change which may have a large impact on global carbon dioxide levels and the health of plants that rely on these soil microbes to survive.

VI. References

Allaire J., Xie Y., McPherson J., Luraschi J., Ushey K., Atkins A., Wickham H., Cheng J., Chang W., Iannone R. (2020). rmarkdown: Dynamic Documents for R. R package version 2.1. https://github.com/rstudio/rmarkdown

Barns, S.M., Cain, E.C., Sommerville, L., and Kuske, C.R. (2007). Acidobacteria Phylum Sequences in Uranium-Contaminated Subsurface Sediments Greatly Expand the Known Diversity within the Phylum. Appl. Environ. Microbiol. *73*, 3113–3116.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Castro, H.F., Classen, A.T., Austin, E.E., Norby, R.J., and Schadt, C.W. (2010). Soil Microbial Community Responses to Multiple Experimental Climate Change Drivers. Applied and Environmental Microbiology *76*, 999–1007.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics *36*, 1925–1927.

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., and Banfield, J.F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature *558*, 440–444.

DeAngelis, K.M., Pold, G., Topçuoğlu, B.D., van Diepen, L.T.A., Varney, R.M., Blanchard, J.L., Melillo, J., and Frey, S.D. (2015). Long-term forest soil warming alters microbial communities in temperate forest soils. Front. Microbiol. *6*.

Eichorst, S.A., Breznak, J.A., and Schmidt, T.M. (2007). Isolation and Characterization of Soil Bacteria That Define Terriglobus gen. nov., in the Phylum Acidobacteria. Applied and Environmental Microbiology *73*, 2708–2717.

Eichorst, S.A., Trojan, D., Roux, S., Herbold, C., Rattei, T., and Woebken, D. (2018). Genomic insights into the Acidobacteria reveal strategies for their success in terrestrial environments. Environmental Microbiology *20*, 1041–1063.

Fierer, N., Bradford, M.A., and Jackson, R.B. (2007). Toward an Ecological Classification of Soil Bacteria. Ecology *88*, 1354–1364.

Frey, S.D., Drijber, R., Smith, H., and Melillo, J. (2008). Microbial biomass, functional capacity, and community structure after 12 years of soil warming. Soil Biology and Biochemistry *40*, 2904–2907.

Greening, C., Carere, C.R., Rushton-Green, R., Harold, L.K., Hards, K., Taylor, M.C., Morales, S.E., Stott, M.B., and Cook, G.M. (2015). Persistence of the dominant soil phylum Acidobacteria by trace gas scavenging. Proc Natl Acad Sci U S A *112*, 10497–10502.

Janssen, P.H. (2006). Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes. Appl. Environ. Microbiol. *72*, 1719–1728.

Jones, R.T., Robeson, M.S., Lauber, C.L., Hamady, M., Knight, R., and Fierer, N. (2009). A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J *3*, 442–453.

Karhu, K., Auffret, M.D., Dungait, J.A.J., Hopkins, D.W., Prosser, J.I., Singh, B.K., Subke, J.-A., Wookey, P.A., Ågren, G.I., Sebastià, M.-T., et al. (2014). Temperature sensitivity of soil respiration rates enhanced by microbial community response. Nature *513*, 81–84.

Kielak, A.M., Barreto, C.C., Kowalchuk, G.A., van Veen, J.A., and Kuramae, E.E. (2016). The Ecology of Acidobacteria: Moving beyond Genes and Genomes. Front Microbiol *7*, 744.

Knaus, B.J., and Grünwald, N.J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour *17*, 44–53.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio].

Melillo, J.M., Butler, S., Johnson, J., Mohan, J., Steudler, P., Lux, H., Burrows, E., Bowles, F., Smith, R., Scott, L., et al. (2011). Soil warming, carbon–nitrogen interactions, and forest carbon budgets. PNAS *108*, 9508–9512.

Melillo, J.M., Frey, S.D., DeAngelis, K.M., Werner, W.J., Bernard, M.J., Bowles, F.P., Pold, G., Knorr, M.A., and Grandy, A.S. (2017). Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. Science *358*, 101–105.

Oliverio, A.M., Bradford, M.A., and Fierer, N. (2017). Identifying the microbial taxa that consistently respond to soil warming across time and space. Global Change Biology *23*, 2117–2129.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology *36*, 996–1004.

Pold, G., Billings, A.F., Blanchard, J.L., Burkhardt, D.B., Frey, S.D., Melillo, J.M., Schnabel, J., van Diepen, L.T.A., and DeAngelis, K.M. (2016). Long-Term Warming Alters Carbohydrate Degradation Potential in Temperate Forest Soils. Applied and Environmental Microbiology *82*, 6518–6530.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative Genomics Viewer. Nat Biotechnol *29*, 24–26.

Scharlemann, J.P., Tanner, E.V., Hiederer, R., and Kapos, V. (2014). Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Management *5*, 81–91.

Schindlbacher, A., Rodler, A., Kuffner, M., Kitzler, B., Sessitsch, A., and Zechmeister-Boltenstern, S. (2011). Experimental warming effects on the microbial community of a temperate mountain forest soil. Soil Biology and Biochemistry *43*, 1417–1425.

Schulz, F., Alteio, L., Goudeau, D., Ryan, E.M., Yu, F.B., Malmstrom, R.R., Blanchard, J., and Woyke, T. (2018). Hidden diversity of soil giant viruses. Nat Commun *9*.

Ward, N.L., Challacombe, J.F., Janssen, P.H., Henrissat, B., Coutinho, P.M., Wu, M., Xie, G., Haft, D.H., Sait, M., Badger, J., et al. (2009). Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. Appl. Environ. Microbiol. *75*, 2046–2056.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. Journal of Open Source Software *4*, 1686.

Xue, K., Xie, J., Zhou, A., Liu, F., Li, D., Wu, L., Deng, Y., He, Z., Van Nostrand, J.D., Luo, Y., et al. (2016). Warming Alters Expressions of Microbial Functional Genes Important to Ecosystem Functioning. Front Microbiol *7*.

Yang, Z., Yang, S., Van Nostrand, J.D., Zhou, J., Fang, W., Qi, Q., Liu, Y., Wullschleger, S.D., Liang, L., Graham, D.E., et al. (2017). Microbial Community and Functional Gene Changes in Arctic Tundra Soils in a Microcosm Warming Experiment. Front. Microbiol. *8*.