October 2022

# Changes in Gene Expression From Long-Term Warming Revealed Using Metatranscriptome Mapping to FAC-Sorted Bacteria

Christopher A. Colvin
*University of Massachusetts Amherst*

# Changes in Gene Expression From Long-Term Warming Revealed Using Metatranscriptome Mapping to FAC-Sorted Bacteria

A Thesis Presented

By

CHRISTOPHER A. COLVIN

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

September 2022

Molecular and Cellular Biology

**Changes in Gene Expression From Long-Term Warming Revealed Using Metatranscriptome Mapping to FAC-Sorted Bacteria**

A Thesis Presented
By
Christopher Colvin

Approved as to style and content by:

_____

Jeffrey Blanchard , Chair

_____

Peter Chien, Member

_____

John Gibbons, Member

_____

Thomas Maresca

Graduate Program Director

Molecular and Cellular Biology

DEDICATION

To my family and friends. Without you, I would not have been able to have the success that I have throughout my years of school. Thank you. To my mentors, thank you for your guidance and helping me discover my love for science.

ACKNOWLEDGEMENTS

I would first like to thank Dr. Jeffrey Blanchard. I am incredibly grateful for your mentorship and guidance throughout the past couple of years. Despite the difficult times of 2020, you still managed to maintain a passionate learning environment. The bioinformatic skills I have learned while conducting research in your lab along with the soft-skills fostered in the lab will be valuable assets in my career as a scientist. I feel very fortunate and privileged to be a part of the Blanchard Lab during my time at UMass. I would also like to thank Dr. Peter Chien and Dr. John Gibbons for their contributions as members of my thesis committee. Your advice and time put into your roles on the committee is greatly appreciated. To Dr. Chien, I am grateful to have had the privilege as one of your students for a semester. To Dr. Gibbons, I am grateful for giving me my first hands on experience with important programs I used to complete my research.

This work was made possible by Department of Energy grants CSP 504322 and FICUS 49483/60006003.

ABSTRACT

CHANGES IN GENE EXPRESSION FROM LONG-TERM WARMING REVEALED USING METATRANSCRIPTOME MAPPING TO FAC-SORTED BACTERIA

SEPTEMBER 2022

CHRISTOPHER COLVIN, B.S., UNIVERSITY OF MASSACHUSETTS AMHERST

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jeffrey Blanchard, Ph.D.

Soil microbiomes play pivotal roles to the health of the environment by maintaining metabolic cycles. One question is how will climate change affect soil bacteria over time and what could the repercussions be. To answer these questions, the Harvard Forest Long-Term Warming Experiment was established to mimic predicted climate change by warming plots of land 5℃ above ambient conditions. In 2017, 14 soil core samples were collected from Barre Woods warming experiment to mark 15 years since the establishment of the soil warming in that location. These samples underwent traditional metatranscriptomics to generate an mRNA library as well as a process coined cell-sorted or mini-metagenomics involving the sorting of single bacterial cells from the environment using FACS. This was followed by pooling into groups of 100 cells for more cost efficient genome recovery. 200 high-quality genomes were compiled, 12 of which were taxonomically identified as *Acidobacteria*. *Acidobacteria* are an extremely abundant and diverse phylum of bacteria that were found to be very well represented in the soil samples. Due to their abundance in many different soil environments as well as their known importance in many metabolic cycles, they were chosen as the candidate phylum to further investigate. Using a reference-based read mapping approach with the 12 *Acidobacteria* genomes and metatranscriptomic data, we identified over 3,000 differentially expressed genes within these organisms as a result of soil warming. Due to the diversity within the phylum itself, many of the genomes indicated different patterns of expression making it difficult to identify phylum-wide differential expression trends. However, the sigma70 factor, an important housekeeping gene used as a transcription regulator, was found to be up-regulated in a majority of the genomes. Over 30 different glycoside hydrolase encoding genes and glycosyltransferases were also found to be differentially expressed across the *Acidobacteria* reference genomes as well as 23 chemotaxis-related genes. Despite identifying four different groups of genes that showed statistically significant differences in expression levels, there may be more changes occurring in these soil bacteria and the soil microbiome as a whole due to climate change than previously measured by read-based analyses of metatranscriptomic data.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS……2

ABSTRACT……3

LIST OF TABLES……...7

LIST OF FIGURES……8

CHAPTER

1. INTRODUCTION……...9

1.1 Diversity in Soil……9

1.2 The Soil Organic Layer……..9

1.3 Current Climate Change and Impact……10

1.4 Bacteria Role in Soil Respiration……10

1.5 The Harvard Forest Warming Plots……11

1.6 Initial Findings From Short Term Warming Experiments……...12

1.7 Barre Woods Sample Collection……..13

1.8 Mini-Metagenomics vs. Metagenomics……...14

1.9 Initial ReadMapping Reveals a Phylum of Interest……16

1.10 What are Acidobacteria……..20

RESEARCH QUESTION……..23

SIGNIFICANCE……24

2. MATERIALS AND METHODS……..25

2.1 Acidobacteria MAGs Used……..25

2.2 General Feature Format Files (GFF) and Gene Product Names……..26

2.3 Metatranscript (Read) Files Used……26

2.4 List of Programs/Tools Used for Computational Analysis……..27

2.5 Read Mapping the Metatranscripts to the 12 Acidobacteria MAGs using STAR…...29

2.6 Differential Expression Analysis…….32

2.7 CAZy and Gene Product Annotation……...34

2.8 GTDB-Tk Analysis……..34

2.9 iTOL Visualization……...35

2.10 OrthoFinder Analysis…….35

3. RESULTS……...36

3.1 Acidobacteria Phylogeny…….36

3.2 Expressional Differences in the Acidobacteria MAGs……38

3.3 Number of Differentially Expressed Genes…….40

3.4 An Up-Regulation is Found in the Sigma 70 Gene…….42

3.5 Exploration of Differentially Expressed Carbohydrate Active Enzymes……46

3.6 Preliminary Results for Differentially Expressed Chemotaxis Related Genes……51

4. DISCUSSION AND FUTURE DIRECTIONS…….53

4.1 A Reference Based Approach to Read Mapping in Acidobacteria is Possible……...53

4.2 A Number of Differentially Expressed Genes are Found in Acidobacteria Due to Soil Warming……55

4.3 Up-regulation of Sigma70 May Explain High Abundance of Acidobacteria in Our Samples……58

4.4 Carbohydrate Active Enzymes are Found to Differentially Expressed in Some Ways…….59

4.5 Changes in Chemotaxis Related Genes Suggest Drying of Soil…….61

4.6 Future Directions…….63

APPENDIX: SUPPLEMENTARY MATERIAL……...65

WORKS CITED……69

LIST OF TABLES

Table 1: **Table 1. MetaTranscript Files Used in Read Mapping Analysis** ……..……26

Table 2: **Table 2: Condition Table for Organic DESeq2 Analysis**……33

Supplemental Table 1**: Condition Table for Mineral Layer DESeq2 Analysis**……65

Supplemental Table 2: **List of All Differentially Expressed Sigma70 Genes**……67

Supplemental Table 3: **List of All Differentially Expressed Chemotaxis-Related Genes**……67

Supplemental Table 4: **List of All Differentially Expressed CAZy Genes**……68

LIST OF FIGURES

Figure 1: **Mini-Metagenomics vs. Bulk Metagenomics**……15

Figure 2a: **Number of Genomic and Transcript Reads to MAGs**……17

Figure 2b: **Total Percents of Genomic and Transcript Reads to MAGs**……..18

Figure 3: **Phylogeny of 26 Different Acidobacteria Subgroups**……21

Figure 4: **Phylogenetic Tree of Acidobacteria MAGs**……37

Figure 5: **Expressional Differences Between 12 Acidobacteria Reference MAGs**……..38

Figure 6: **Total Number of Differentially Expressed Genes From Organic Samples**…… 41

Figure 7: **Expressional Data of Sigma70 (ECF subfamily) in Acidobacteria MAGs**……..44

Figure 8: **Differentially Expressed Carbohydrate Metabolizing Enzymes in *Acidobacteria* MAGs**……46

Figure 9: **Differentially Expressed GT and GH Genes in Reference MAGs**……...48

Figure 10: **Differentially Expressed Chemotaxis-Related Genes in Reference MAGs**……...51

Supplementary 1: **Heatmap of Differentially Expressed Genes at Padjusted 0.1 Cutoff** …...66

# CHAPTER 1

# INTRODUCTION

## *1.1 Diversity in Soil*

The diversity found in soil is second to none in the world's ecosystems with a single teaspoon of soil containing anywhere from 100 million to one billion bacteria [2], the diversity and life seen below the surface is unmatched [1]. Of the microorganisms found in soil, bacteria are the most diverse. Bacteria make contributions to the soil ecosystem in many ways such as nutrient cycling and disease suppression, but the ways in which they do so varies greatly [2]. This can be attributed to the wide diversity found between bacteria themselves. As not one species reigns supreme in a given sample and many bacteria vary from one sample to another, the ways in which these processes occur also diverge. However, understanding the basics of these different pathways and slowly accruing a deeper knowledge can contribute to a healthy ecosystem. To first understand the roles of different bacteria, we must first further understand their environment, that is soil.

## *1.2 The Soil Organic Layer*

Soil is home to billions of organisms and is composed of many dynamic layers. Within each of these layers, certain microbes reside and perform different biogeochemical processes. One main component is the top organic layer which houses a majority of microbial diversity and dictates many of the chemical properties associated with the soil environment [3]. Due to the high concentration of terrestrial carbon, as well as serving as the nitrogen reservoir for the soil, this layer is often studied in relation to microbes behavior [4]. Due to this high microbial density, the organic layer is the most dynamic, in that there is constant decomcompostion and building of this layer from new leaf litter. Within this layer the respiration from soil organisms releases carbon

back into the atmosphere. Thus, any changes to the microbiome may have drastic effects in the composition of the layer itself [5] and lead to increase in respiration which in turn increase atmospheric carbon dioxide levels [6]. As the biogeochemistry role of the SOM is pivotal to the health of the system, changes may impact the role of life above and within the soil.

*1.3 Current Climate Change and Impact*

Global warming is defined by NASA as the long-term warming of the Earth's climate that has been observed since the industrial period [7]. Essentially the idea is that the climate, or the long-term regional/global temperature, is increasing due to the various pollutants emitted into the atmosphere. Effects of this phenomenon are clearly seen from the higher amounts of carbon in the atmosphere as well as a four inch increase of global sea levels. Although the increase has been gradual, the scale and the rate of the increase is becoming more alarming each year. As of today, we have seen an increase of over 1 ℃ since 1880 [7]. However, there are projections of an increase from 1.1 - 6.4 ℃ within the next 100 years alone. One of the largest contributors to this spike in the emission of greenhouse gasses. The main four greenhouse gasses are carbon dioxide, methane, nitrous oxide and water vapor [8].

*1.4 Bacteria Role is Soil Respiration*

Bacteria are the most abundant living organisms within soil, with a single gram containing 1000-1,000,000 different species [9]. They also inhabit the organic layer of the soil that contains the reserve of carbon for select environments. As the soil environments act as a reliable carbon sink by holding thousands of pentagrams (Pg) of carbon, studies on microbial processes in relation to carbon are becoming of more concern. Previous studies have shown an initial increase in $CO_2$ flux as a result of short term warming [10].There are many possibilities as to why

this occurs with one reason being that warming leads to an increase in microbial activity, particularly in regards to cellular soil respiration. Soil respiration is the process of CO2 releasing into the atmosphere via root respiration and the decomposition of the SOM [11]. Approximately 74% of total soil respiration can be attributed to microbes. Previous short term warming studies suggest an increase in microbial activity [10]. This study conducted by Jerry Mellilo intended to anticipate the possible impacts of climate change. Using his estimation of an increase of 5 degrees ℃ over the next 100 years, he suggested an observed increase in bacterial activity [10]. In combination with this increase in activity and $CO_2$ flux, one may question the impact on soil respiration rates climate change may have. However, in order to study the question that climate change may have on bacterial activity, there must be an environment suitable to conduct these studies.

*1.5 The Harvard Forest Warming Plots*

In order to study the possible changes to the soil environment as a result of climate change, the Harvard Forest Warming plots were established. As experts predict that soil warming may affect carbon storage and other cycles within the biosphere, a location in which the soil is constantly heated at a controlled temperature was created [10]. The Harvard Forest warming experiment was started in 1991 at Prospect Hill located in Massachusetts. As climate models predict an increase of temperature anywhere between 2-5 ℃, the field warming experiment heats the soil 5 ℃ above the ambient temperature [12]. This heating mechanism offers scientists information to study three main goals: Track the measurements of carbon and nitrogen stocks

and flux over long-term soil warming, observe warming induced feedbacks from short-term manipulation and study targeted processes that may have been affected [12]. Along with the site located at Prospect Hill, the experiment has expanded into two other sites. The Soil Warming x Nitrogen Study site was established in 2006 and the Barre Woods site was established in 2002. Although slightly differing in plot sizes, both sites follow the same principles of heating the soil 5 ℃ above the ambient temperature.

*1.6 Initial Findings From Short Term Warming Experiments*

As previously mentioned, Jerry Mellilo made initial short term observations at the 7 year warming mark at the Prospect Hill plots. Researchers specifically observed the changes in the carbon flux over a seven year warming period. Plant carbon storage was quantified during tree growth measurements and SOC was measured via soil respiration rates, fine-root respiration and fine-root biomass [10]. Findings included an initial spike in soil respiration rates after two years of heating that eventually leveled off and fell below the control. There was also an overall decrease in carbon flux over this seven year period. The decrease in carbon flux was attributed to woody tree growth resulting in a decrease of fine-root mass. Fine-root mass was estimated to contribute roughly 26% of the total respiration rates initially, with the remainder attributed to microbes, but only 18% after the 7 year warming mark [10]. As fine-root mass plays an important role in carbon turnover and other nutrient cycling, it was chosen as one indicator of total soil respiration levels [13]. These findings suggest that short-term soil warming leads to changes in both soil respiration rates and carbon flux due to changes in fine-root mass. However, how is microbial activity affected? Initial spikes in respiration rates may indicate changes in microbes that, although not sustainable, impact their own cellular processes to compensate for the decrease in root

respiration. Questions of long-term warming impacts also come to mind as these changes were seen after only a seven year period. Fortunately, the Harvard Forest Warming Project is still active and since the creation in 1991, long-term warming data is now available.

*1.7 Barre Woods Sample Collection*

Barre Woods differs from the plot located at Prospect Hill in that it is a single 30 x 30m plot that is split in half to heated and non-heated sectors. Heating cables are buried 10 cm deep to heat half the plots [12]. Although a combined 15 x 15m space, each of the two plots are divided into sub plots that are used to denote sample collection sites. In 2017, 14 forest core samples were taken from the Barre Woods plot to signify the 15 year warming point. These are the samples that have produced the data used in this study. Seven samples were taken from the heated plots and seven samples were taken from the control plots. Each of the samples taken from both conditions were then split into two respective layers: The organic layer and the mineral layer given a total of 28 different samples [9]. The samples were denoted by the condition of warming or control, a number signifying the sublot the sample was taken from , and the soil layer (O for organic or M for mineral). For all 28 samples, traditional bulk-metatranscriptomics was conducted to collect mRNA transcript reads for transcriptome analysis. Along with this, four of the soil core samples underwent bulk-metagenomics and four samples underwent a new process known as mini-metagenomics. With the goal of generating metagenome-assembled genomes (MAGs) as references for organisms found in our samples, one method was found to be more effective than the other.

*1.8 Mini-Metagenomics vs. Metagenomics*

Of the 28 samples, four were chosen at random to undergo both traditional metagenomics and mini-metagenomics [9]. The four samples consisted of a single sample from each condition: One heated organic sample, one non-heated (control) sample, one heated mineral sample and one non-heated mineral sample. Genomics is the study of the complete genetic makeup of an organism through high-throughput sequencing technology. Metagenomics differs in the fact that it specifically studies the genome sequences of those organisms found in a specific community, or soil in this case [14]. As working with soil is difficult due to the diversity within the environment, traditional metagenomic strategies often lead to metagenome-assembled genomes (MAGs) that have high contamination. High contamination in this case refers to assembled genomes that incorporate genomes from multiple different organisms. Due to this difficulty when working with soil, a new method coined mini-metagenomics was conducted to obtain higher quality samples. Figure 1 outlines the key differences between traditional metagenomics and the mini-metagenomic method used. The main differences consist of filtration through a 5 micron filter to filter out larger eukaryotic organisms and saturate the samples with bacteria and other smaller microorganisms. Sorting into pools of 100 cells through fluorescent activated cell sorting (FACs) helps generate samples with less genomic material with the intention of assembling scaffolds with lower contamination from other genomes [9]. This is accomplished by staining with SYBR green which stains the genetic material in the cells to quantify a cell. Traditional metagenomics consisted of DNA extraction via the DNeasy PowerSoil extraction kit [15]. All library preparations and sequencing were performed at the Joint Genome Institute (JGI) where Illumina HiSeq was used to generate all reads [9].Sets of reads from both methods underwent trimming and quality checking through computational pipelines.

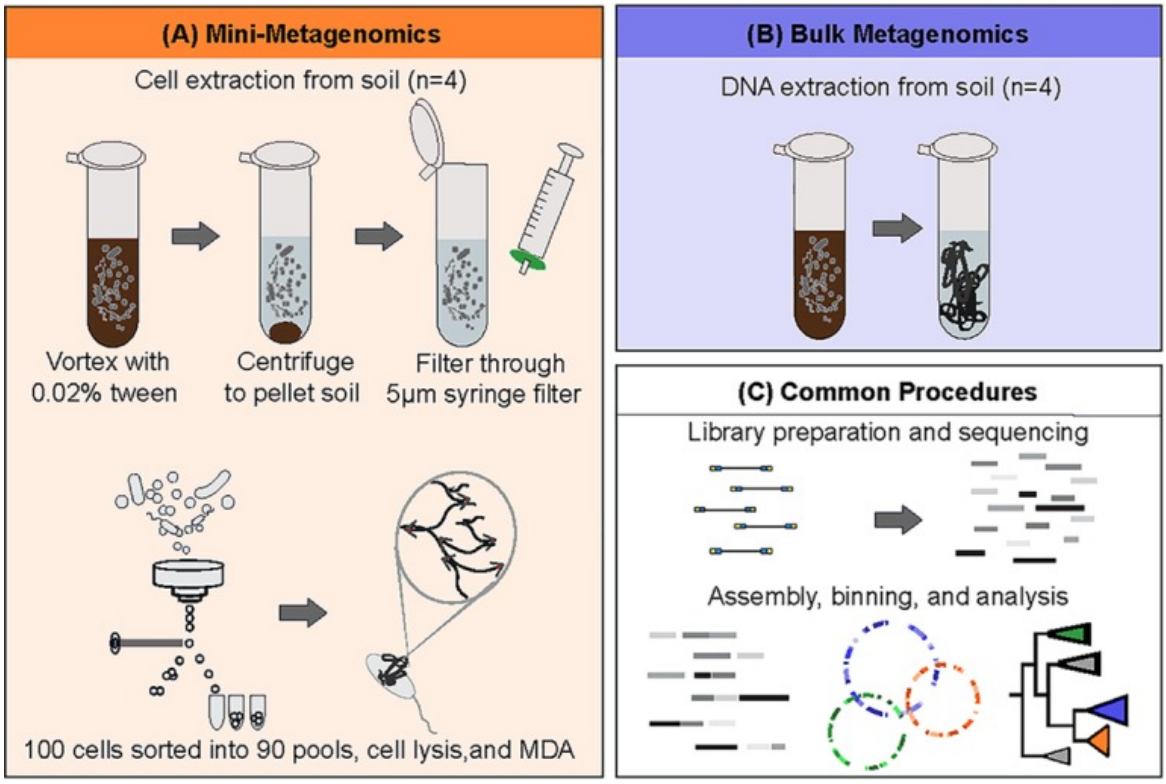



**Figure 1: Mini-Metagenomics vs. Bulk Metagenomics:** The Figure above compares the methods of Bulk-Metagenomics (Blue) to Mini-Metagenomics (Orange). 4 samples (n=4) underwent both processes. Mini-Metagenomics consists of treatment with 0.02% tween followed by cell extraction and filtration through a 5 micron filter. The samples are then stained with SYBR green to perform fluorescent activated cell sorting into pools containing 100 cells each. Traditional Bulk Metagenomics consists of standard DNA extraction and purification. In both methods, the samples are prepared for sequencing and computational pipelines are used for assembly and binning to obtain MAGs. Figure taken from *Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil* by Aleteio et Al [9].

The assembly of the reads into contigs and larger scaffolds was done through the program SPAdes [16]. Assembled contigs were then binned into the resulting MAGs using MetaBat2, a binning software that sorts contigs based on tetranucleotide frequency [17]. The final step included a quality assessment of the assembled MAGs to ensure that the MAGs consisted of >50% completeness of essential hallmark genes, <10% contamination from other genomes and <10% strain heterogeneity to be considered high-quality for downstream analysis [9]. This was

accomplished using the program CheckM [18]. After assembly, binning and quality filtration of the MAGs from both methods, 200 high-quality MAGs were extracted from the mini-metagenomic samples while only 29 high-quality MAGs were a result from the bulk-metagenomic samples.

*1.9 Initial ReadMapping Reveals a Phylum of Interest*

With metatranscriptome reads from the 28 samples and the assembled high-quality MAGs from mini-metagenomics, read mapping was performed using the Burrows-Wheeler aligner [19]. Figure 2 shows the resulting data identifying different phylum of bacteria in the samples. It is important to note that the total reads during this cycle did not filter based on heating vs non-heating. Figure 2a depicts the total reads as a result from looking at all 28 sets of transcript reads and the sets of genomic reads regardless of warming. Taxonomic classification for each of the 200 MAGs was performed and the phylum classification can be depicted by the key on the right.

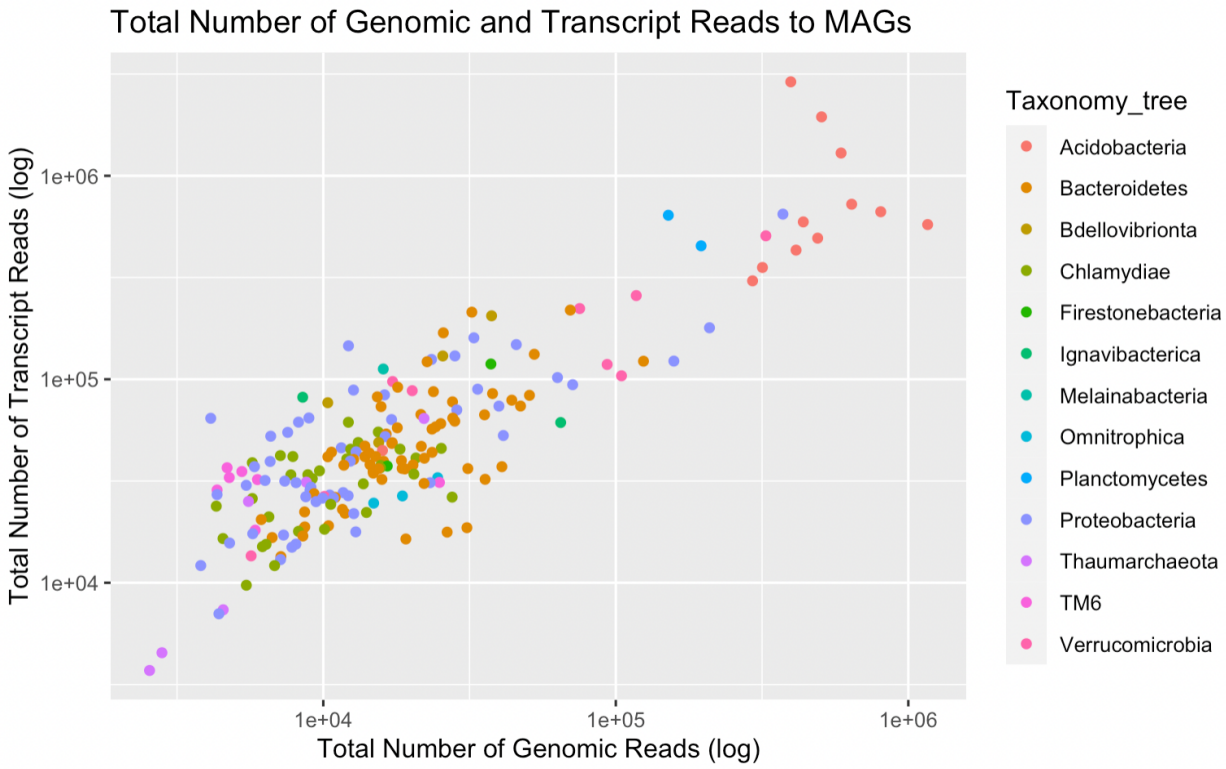



**Figure 2a: Number of Genomic and Transcript Reads to MAGs:** The plot above was created in RStudio using ggplot2. Depicted on the x-axis is the scaled total number of genomic reads that are mapped to high quality MAGs produced through the mini-metagenomics method. Depicted on the y-axis is the scaled total number of transcript reads that are mapped to those same MAGs. The key on the right indicates the phylum that each of the MAGs were classified as and is represented through color on the graph.

Results from Figure 2a are quite interesting. A variety of different bacterial phyla are represented in the samples and at very different levels. As the genomic and transcript samples are not from isolates and are obtained from the soil, one would expect to see a wide range of different organisms in the sample and with lower coverage on average than from an isolated sample. That is what we see in this case, but more interestingly, a cluster of one phylum in particular can be seen to have a higher number of reads mapped on average than the other phylum represented in

the sample. 11 Acidobacteria MAGs are seen to cluster at the top in both directions with 12 total being present in the sample. 11 of 12 Acidobacteria MAGs have a higher proportion of genomic reads mapped to them with the exception of two MAGs. A vast majority of MAGs also fall under the 1E5 point. Figure 2a also portrays that 11 of 12 Acidobacteria MAGs also had a higher number of transcript reads mapped when compared to the other phyla. Figure 2b shows another depiction of the disproportionate differences represented in the samples.

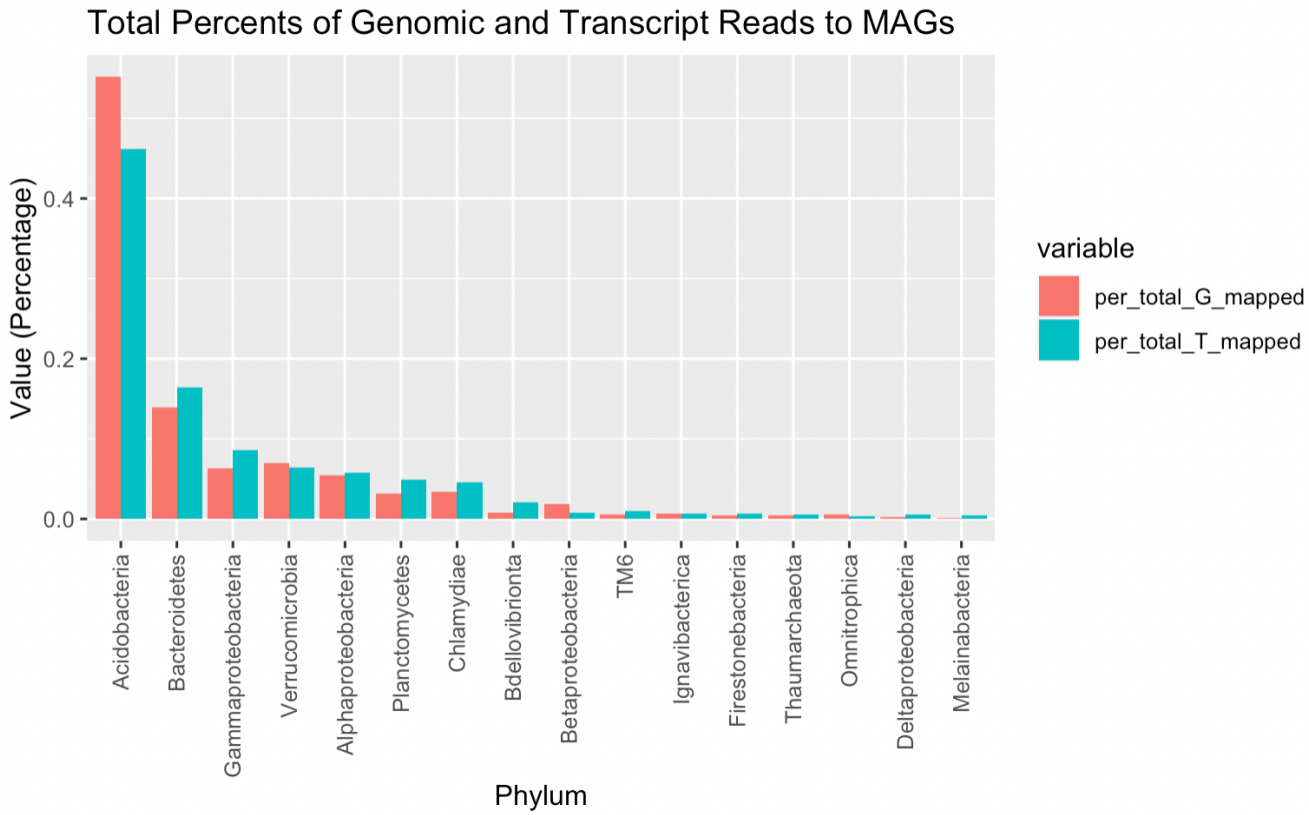



**Figure 2b: Total Percents of Genomic and Transcript Reads to MAGs:** The bar plot above was constructed in RStudio using ggplot2. The x-axis depicts the phyla represented in the samples. The y-axis depicts the values in percentages of the proportion of total number of reads mapped to each phyla against the total number of reads in total. The key on the right depicts the variable that each color represents; red represents the genomic reads mapped and blue represents the transcript reads mapped.

Figure 2b shows a barplot as opposed to the dot plot in Figure 2a that is filled by total genomic read percentages shown in red and total transcript read percentages in blue. Also, the values are not represented in relation to the MAG itself, but rather each phylum. Again, a disproportionate percentage of both the transcript and genomic reads map to the phylum *Acidobacteria*. Of the other phyla represented in the samples, none represent even half the proportion of the total reads in either set that the *Acidobacteria* population represents. This data tells us a few things. One is we know that *Acidobacteria* are well represented in the samples. As a high number of genomic reads mapped to the sample, one can conclude that not only are *Acidobacteria* found in the samples, but they appear to be relatively abundant based on the high number of genomic reads compared to the other phylum. It can also be seen that 12 different organisms of this phylum are represented in the sample as seen by the 12 individual points in Figure 2a. As each of these points represent a high-quality MAG, 12 total different *Acidobacteria* organisms are present in the samples. The second main takeaway is that *Acidobacteria* are also active in the samples as seen by the total number of transcript reads that mapped to the phylum specifically. Transcripts represent the precursors to proteins found in the samples. Large number of transcripts mapped indicate high levels of transcription from the *Acidobacteria*. This suggests that this phylum in particular is more active compared to the other bacteria identified. Also, as seen from Figure 2a, there are varying levels of abundance and activity within the 12 individual organisms themselves. This would suggest that there is diversity within the phylum itself. However, although this finding gives a phylum of interest to further study, it also raises more questions. As the mapping does not differentiate based on heating, we do not know how heating is affecting the organisms at the genomic and transcriptomic level. We also do not know what if any particular genes/pathways are impacted from this heating. It appears that further studying the

*Acidobacteria* found in the samples taken from Barre Woods may be used to infer impacts of long-term warming in our samples and bacteria in general. To answer these questions, more information regarding this phylum of bacteria is necessary.

*1.10 What are Acidobacteria*

*Acidobacteria* are gram-negative bacteria that have recently gained a lot of attention due to their role in soil ecosystems [20]. They are found predominantly in tropical agricultural climates, but are ubiquitous by nature due to their diversity [21]. Making up close to 50% of total bacterial population in known soil environments, this phylum has been found to compose up to 20% of the diversity in these populations [22]. First identified in 1991, the phylum was accordingly named after being found in acidophilic conditions in Japan [23]. However, despite being vastly abundant in soil environments, they are under-represented in science due to the difficulty isolating these bacteria. In lab environments, it has been very difficult to culture *Acidobacteria* which accounts for the lack of representation in past literature. Now with the genomic technology currently accessible, scientists have been able to identify this large cluster of bacteria as their own phylum.

*Acidobacteria* were previously grouped taxonomically with *Proteobacteria* with 87% of existing phylogenies placing the groups adjacent to each other [24]. The vast diversity found within the phylum has become of interest to scientists. Using phylogenetic analysis of 16S rRNA, the phylum is split into an astonishing 26 subdivisions, but this number is anticipated to climb [22]. *Acidobacteria* have often been described as ubiquitous due to their ability to thrive in a plethora of soil environments with a wide range of temperatures [25].

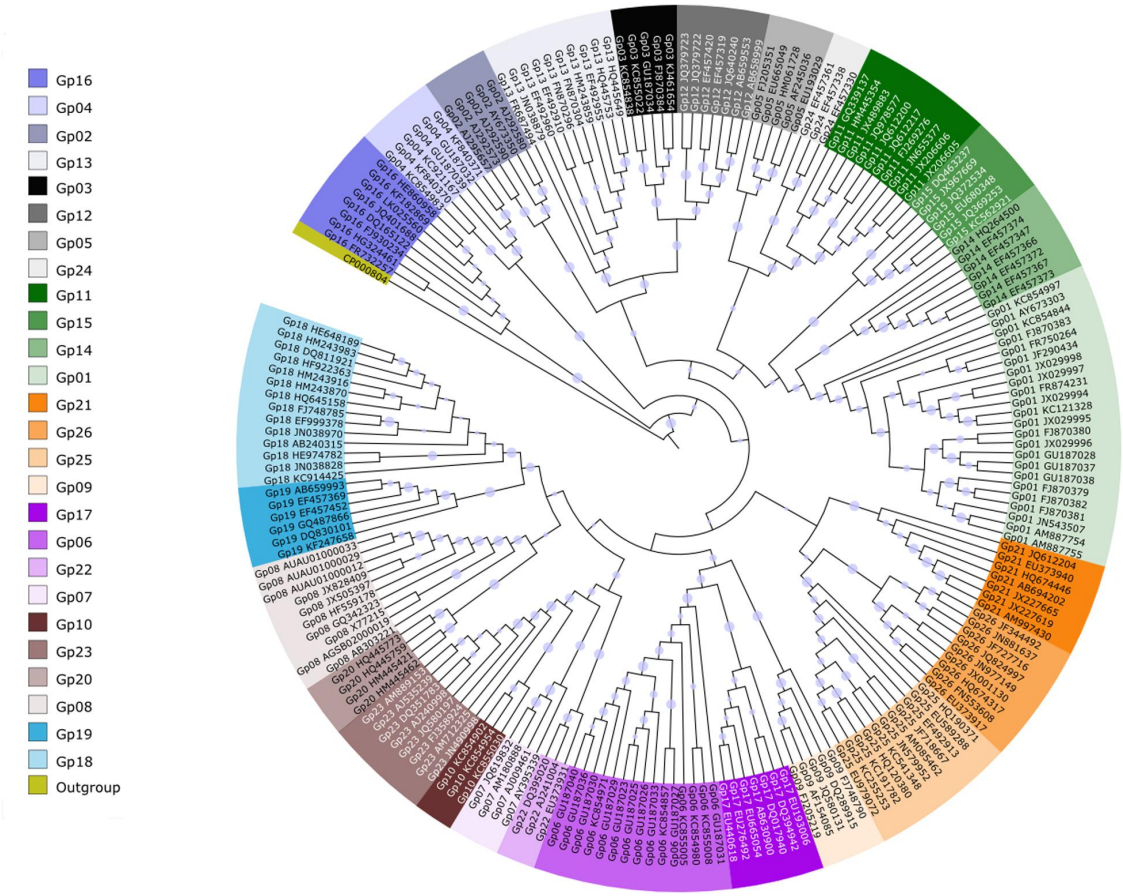



**Figure 3: Phylogeny of 26 Different Acidobacteria Subgroups**: The phylogenetic tree above illustrates the 26 different subgroups the Acidobacteria phylum is split into. The phylogeny consists of 220 different sequences taken from the Silva database. The circles found on some branches in the tree represent a bootstrap of more than 75%. Figure taken from *The Ecology of Acidobacteria: Moving beyond Genes and Genomes* by Kielak et al [22].

Figure 3 visually illustrates the diversity within the phylum itself through the analysis of 220 different *Acidobacteria* organisms. Subdivisions 1 and 3 are aerobic heterotrophs inhabiting areas of lower pH [24]. However, you also have subdivisions that prefer more extreme thermophilic conditions, such as subdivisions 4,8,10 and 23 [22]. With these vast differences in habitats, it is to be expected that the different subdivisions also readily express a variety of different genes. For example, subdivisions 1 and 3 possess a high-affinity for ABC-type transporters which generally couple the hydrolysis of ATP with translocation of solutes [24]. Subdivisions 1-6 also have an

affinity for bd-type oxidases which allow the bacteria to respire oxygen at atmospheric and sub-atmospheric conditions [25]. There are many more differences in expression within this large phylum, but perhaps one of the more interesting traits that differentiate this phylum is difference in carbohydrate metabolism.

Modern day sequencing of different *Acidobacteria* organisms have helped give information to the specific roles they play in their environment. One of significance from genomic studies is carbohydrate metabolism [22]. Compared to other groups of bacteria, *Acidobacteria* possess a disproportionate number of glycoside hydrolase-encoding genes (GHs). Specifically, the phylum is placed in the top 5% of bacterial genomes in the carbohydrate enzyme database [24]. Glycoside hydrolases are a diverse group of enzymes that are found in all domains and help break down different carbohydrates [26]. Specifically, the diversity within this gene family can be seen in *Acidobacteria* within the different subdivisions. For example, subdivision 1 metabolizes D-glucose, D-xylose and other different oligosaccharides. However, it lacks the ability to metabolize polysaccharides such as fructose or other disaccharides such as sucrose [22]. Subdivisions 3 and 4 do possess the ability to metabolize these two specific sugars, but at the cost of decreased carbon storage when compared to subgroup 1. Chitin is another known carbohydrate subdivisions 3 and 4 can utilize [22]. This wide range in carbohydrate metabolism genes may account for the diversity and abundance seen in *Acidobacteria* as it allows the group to outcompete other bacteria in soil.

With the wide diverse capabilities and survival strategies cited, *Acidobacteria* appear well equipped to survive and adapt to many different environments. Vast sources of carbon from the various metabolic genes they possess as well as the ability to encode their own auxiliary metabolic genes support why this phylum appears abundant in many soil environments. Because

of these factors among others such as the phylums role in nitrogen metabolism as well, *Acidobacteria* act as a great indicator for environmental changes. A detailed analysis into genomic changes to *Acidobacteria* specifically could possibly give an idea to changes occurring to many different bacteria in these soil environments as a result of climate change. This also could explain the disproportionate abundance of *Acidobacteria* in the samples collected from Barre Woods. For this reason, further research was performed to take a closer look at the expressional changes occurring in the 12 *Acidobacteria* organisms found in the samples taken after 15 years of warming.

## RESEARCH QUESTION

The primary research question investigated is does long-term soil warming affect the expression of given genes in *Acidobacteria*? In order to answer this question, the metatranscriptomic data will be used in combination with the 12 assembled *Acidobacteria* MAGs to perform a reference-based read mapping approach. All 12 MAGs have gene annotations as well giving the opportunity to observe changes in gene expression as a result of the long-term warming. This will be followed by differential expression analysis to discover statistically significant expressed genes between the heated and control samples.

## SIGNIFICANCE

Global warming is a reality that is becoming more apparent. Much is unknown the potential impacts climate change will have on all forms of life, including the soil microbiome. As soil lays the groundwork for a flourishing ecosystem, any changes in soil may also lead to a downstream effect. As *Acidobacteria* are found in many different soil climates and play pivotal roles in important metabolic processes, any changes to important genes seen in *Acidobacteria* may reflect changes occurring in other bacteria.

We are also attempting a reference-based read mapping approach within the construct of soil. Soil contains millions of organisms and it is extremely difficult to obtain isolates or look at changes within a particular group. However, this method of mini-metagenomics has generated reference genomes to map back to which have been taxonomically categorized. Typically a read based approach is used when working with soil samples as reference genomes are not typically available. By performing this initial step, a reference genome based approach can be explored and will serve to see how practical this method can be.

# CHAPTER 2

# MATERIALS AND METHODS

All raw metagenome and metatranscript sequences can be found on the Joint Genome Institute (JGI) genome portal along with all related annotation files [27]. Bacterial MAGs are also available in GenBank under Bioproject accession number PRJNA608716 [9].

## *2.1 Acidobacteria MAGs Used*

As previously mentioned, 12 different Acidobacteria MAGs were found to be of high-quality in the samples. These MAGs will serve as the reference genomes for which the transcript reads will map to and their IDs are listed below:

1. 3300021028.fa.3
2. 3300020651.fa.6
3. 3300020916.fa.5
4. 3300020743.fa.1
5. 3300020780.fa.4
6. 3300020904.fa.1
7. 3300021040.fa.2
8. 3300021013.fa.6
9. 3300020924.fa.1
10. 3300020985.fa.2
11. 3300021003.fa.3
12. 3300020924.fa.3

*2.2 General Feature Format Files (GFF) and Gene Product Names*

For each MAG, there is a General Feature Format (GFF) file with gene annotations. A GFF file contains information describing different genome features. In this case, each GFF contained a locus tag and a sequence ID representing a particular gene. All locus tags were identified and labeled with the gene product name along with its KEGG annotation. Using a python script, all GFF files used were subset to only contain gene sequences from the corresponding reference genome. This means that each *Acidobacteria* MAG has its own corresponding GFF file containing a locus tag and its corresponding sequence ID. Also, each MAG had its own corresponding file containing the product name for each locus tag and its KEGG annotation.

*2.3 Metatranscript (Read) Files Used*

14 soil core samples were taken from 14 different plots resulting in 28 sets of metatranscript files. 7 files represent samples from the soil organic layer that were not heated, 7 files represent samples from the soil organic layer that were heated, 7 files represent samples from the soil mineral layer that were not heated and 7 files represent samples from the soil mineral layer that were not heated. All read files underwent a quality check via FastQC with adapter sequences trimmed via Trimmomatic [28] [29]. Below is a table listing these files:

**Table 1. MetaTranscript Files Used in Read Mapping Analysis**

| Organic Control | Total # of Reads | Mineral Control | Total # of Reads | Organic Heated | Total # of Reads | Mineral Heated | Total # of Reads |
|---|---|---|---|---|---|---|---|
| NatBWC 12O | 154475132 | NatBWC 12M | 166053322 | NatBWH 11O | 166167809 | NatBWH 11M | 141756874 |
| NatBWC 14O | 199108669 | NatBWC 14M | 180897145 | NatBWH 17O | 150165463 | NatBWH 17M | 176294215 |
| NatBWC 19O | 174379752 | NatBWC 19M | 178043097 | NatBWH 26O | 189794985 | NatBWH 26M | 168886793 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| NatBWC 27O | 171495589 | NatBWC 27M | 170846067 | NatBWH 28O | 185624288 | NatBWH 28M | 176393643 |
| NatBWC 30O | 193252851 | NatBWC 30M | 187034288 | NatBWH 2O | 176084583 | NatBWH 2M | 192785046 |
| NatBWC 4O | 180562421 | NatBWC 4M | 137134632 | NatBWH 32O | 159187369 | NatBWH 32M | 136986557 |
| NatBWC 7O | 150682690 | NatBWC 7M | 172444343 | NatBWH 4O | 131990651 | NatBWH 4M | 181789077 |

*2.4 List of Programs/Tools Used for Computational Analysis*

As the soil core samples were previously collected and sequenced, the entirety of this analysis was computational. Below is a list of the programs used to conduct the analysis of the data along with a short description.

**STAR**- Spliced Transcript Alignment to a Reference (STAR) is a RNA-seq aligner used to map the transcript reads to reference genomes and obtain read count dataframes. Although typically used in eukaryotic mapping due to its splice site awareness, STAR can also be used within bacterial genomes with small adjustments in the parameters. STAR was chosen due to its higher mapping speeds and precision when compared to other aligners [30].

**DESeq2** - DESeq2 is a bioconductor package that tests for differential gene expression from RNA-seq count data. DESeq2 uses a negative binomial generalized linear model to calculate logarithmic fold changes and dispersions to help statistically identify differentially expressed genes [31].

**OrthoFinder** - OrthoFinder is a software that uses proteome inputs to infer orthogroups and orthologs of genes from input references. Other outputs include rooted species trees, gene duplication events as well as other comparative statistics [32].

**GTDB-Tk** - The Genome Taxonomy Database Toolkit (GTDB-Tk) is a software package that can be used to taxonomically classify bacterial and archeal genomes based on the Genome Taxonomy Database. GTDB-Tk is also specifically designed to work with MAGs making it a great tool for this dataset [33].

**iTOL**: The Interactive Tree of Life or iTOL is an online tool that allows for visualization of phylogenetic tree files. Other features include annotation capabilities of these constructed trees as well as figure export [34].

**RStudio** - RStudio is an application that creates a virtual environment for R, a programming language. R is specially designed to be used for statistical analysis and various graphics. RStudio allows users to download software packages for use. The packages used in RStudio are DESeq2 as mentioned above, ggplot2 which is used for graphic visualization and dplyr which is used for data manipulation [35].

**Unity HPC** - Unity is a High Performance Computing Cluster (HPC) composed of powerful computers. This allows users to run software that requires resources beyond the scope of an everyday laptop/desktop. As some of the programs used require a lot of computing power, jobs were scheduled on Unity to utilize the powerful resources available to obtain results. The following programs utilized the Unity HPC for computing power: STAR, GTDB-Tk and OrthoFinder [36].

**CAZy Database** - CAZy is an online database that has information regarding carbohydrate related enzymes and describes these gene families. Annotations of these gene and gene families are continuously added to the database to create a domain specific for carbohydrate enzymes. With a given reference sequence, CAZy can cross the reference sequence with the database and annotate any predicted carbohydrate enzymes [37].

*2.5 Read Mapping the Metatranscripts to the 12 Acidobacteria MAGs using STAR*

The mapping of reads with the STAR aligner is a two step process. Generating a genome index of the reference sequence is step one. To do so, the reference genome sequence in a fasta format is required as well as its corresponding GFF or GTF file. In this case, the *Acidobacteria* MAGs were used as the reference genome and its corresponding GFF was also used. The purpose of indexing a genome is to generate indices of where the aligner should map the reads to as the transcript reads are being mapped to the reference sequence. This also "places" the gene annotations along the reference sequence. This also allows the aligner to map at a faster rate opposed to strictly mapping to the reference sequence. Genome indexing with STAR was done on the Unity HPC. In order for the code to run successfully the following parameters were used:

–runMode - Direct STAR to run genome indexing

–genomeDir - Specifies path of output directory to store genome index

–genomeFastaFiles - Specifies path to reference genome that is being indexed

–sjdbGTFfile – Specifies path to GFF or GTF file used

–sjdbGTFfeatureExon – Defines element used to identify coding sequences in GFF file

–sjdbOverhang – Used to help create splice junction database (not important for bacterial genomes)

–genomeSAindexNbases - Scaler used for smaller genomes (STAR will suggest parameter if necessary based on genome length)

–outFileNamePrefix - Specifies prefix for output files

An example for the code to generate the genome index for Acidobacteria MAG 3300021040.fa.2 can be seen below:

STAR --runMode genomeGenerate --genomeDir 1040index/ --genomeFastaFiles RM330021040.fa.2.fa --sjdbGTFfile 1040gff.gff --sjdbGTFfeatureExon CDS --sjdbOverhang 129 --genomeSAindexNbases 9 --outFileNamePrefix 1040index

The resulting output is a directory that acts as a genome index for your reference genome. As there are 12 *Acidobacteria* genomes to use as references, 12 separate genome indexes were generated.

The second step following genome indexing is the read counting. This step takes considerably longer and requires more computing power in relation to the genome indexing step. However, the result is the data used for expressional analysis. For the second step, an index of the reference genome from step one is required, a file containing all the reads that will be mapped is required and the GFF or GTF file used for the indexing step is needed. Mapping of the reads to the reference index was all completed on the Unity HPC. In order for the mapping step to run successfully, the following parameters were used:

–runThreadN - Number of CPU cores utilized

–genomeDir - Name of directory the output will be stored in

–sjdbGTFfile - Path to GFF or GTF file used

–sjdbGTFfeatureExon - Defines element used to identify exons

–sjdbGTFtagExonParentGene - Defines tag name to use for the gene

–alignIntronMax - Defines maximum intron length (Set to 0 when working with bacterial genomes)

–readFilesIn - Path to file containing all transcript reads

–readFilesCommand - Defines if files are compressed or not so STAR can read them correctly

–outSAMtype - Defines output alignment type in binary format

–quantMode - Defines quantifying output. As reads per gene is the desired output, GeneCounts is used

–outFileNamePrefix - Specifies prefix for output files

–sjdbOverhang – Used to help create splice junction database (not important for bacterial genomes)

An example of the code used for the step can be seen below:

```
STAR --runThreadN 32 --genomeDir 330021040/1040index/ --sjdbGTFfile
330021040/1040gff.gff --sjdbGTFfeatureExon CDS --sjdbGTFtagExonParentGene locus_tag
–alignIntronMax 1 --readFilesIn ~/scratch/NatBWC12O_metat.trim.fastq.gz
--readFilesCommand zcat --outSAMtype BAM Unsorted --quantMode GeneCounts
--outFileNamePrefix Controlled/12O/ --sjdbOverhang 129
```

There are a number of results from the following step, but there are two main output files that will be used for further analysis. The first is a file labeled “ReadsPerGene.out.tab”. This file contains four columns: Column one is the gene ID, column two is the counts for unstranded RNA-seq, column three is for strand specific counts for the first read strand aligned with RNA, and column four is for strand specific counts for the second read strand aligned with RNA. Strandness for this dataset is not a concern so the second column of counts will be used for DESeq analysis. The second file of interest is labeled “Log.final.out”. This acts as an overall summary of the mapping that was performed giving overall statistics. Specifically, the total number of reads mapped and the percentage of reads that mapped will be used for further

analysis. If working with eukaryotes, this output also includes information regarding splicing events, but as this is a bacterial dataset, there is no information to log. The end result was 28 count matrices for each genome along with 28 summary outputs for each genome. Totaling 336 read count files and 336 final log reports.

*2.6 Differential Expression Analysis*

With the count matrices from the read mapping, the next step was to perform differential expression analysis to observe if any genes showed statistically significant differences in the control and heated plots. In order to do so, RStudio was used to subset the data specifically using the dplyr package [38]. For each of the 12 *Acidobacteria* genomes, two tables were generated using the read count data. One of the tables contained all the read count data for one of the *Acidobacteria* genomes from the 14 organic horizon samples. The second table contained all the read count data for one of the *Acidobacteria* genomes from the 14 mineral horizon samples. In total, 24 read count tables were constructed with 2 being from each genome, one table with all the read counts from the organic meta transcript samples and one table with all the read counts from the mineral meta transcript samples. For the final report files, a table was constructed that contained the total read counts and uniquely mapped reads from the different meta transcript samples (plots) to each genome. Soil layer was also defined in this table so that the organic and mineral samples were separated.

The package DESeq2 was used in RStudio to normalize the data to obtain the differential expression data. The inputs used were the newly created read count tables for each *Acidobacteria* genome, along with a table that defined the condition and type for each sample. This would tell the program which values to compare based on a condition. Table 2 shows what was used when

performing DESeq2 analysis on all the organic samples. Supplemental Table 1 shows what was used for the same analysis, but using the samples taken from the mineral layer. For each read count table, the DESeq2 pipeline was followed according to the bioconductor vignette [39] [40] [31]. Results are comparing conditions treated (warming) to untreated (no warming). One important output from DESeq2 is log2 fold change which indicates the up or down regulation of values based on a condition. As we are investigating gene expression as a result of warming, a positive log2 fold change will be indicative of an up-regulation of gene expression and a negative log2 fold change will be indicative of a down-regulation of expression.

**Table 2: Condition Table for Organic DESeq2 Analysis**

| Plot | Condition | Type |
|---|---|---|
| Control12O | untreated | organic |
| Control14O | untreated | organic |
| Control7O | untreated | organic |
| Control30O | untreated | organic |
| Control19O | untreated | organic |
| Control27O | untreated | organic |
| Control4O | untreated | organic |
| Heated2O | treated | organic |
| Heated4O | treated | organic |
| Heated11O | treated | organic |
| Heated17O | treated | organic |
| Heated26O | treated | organic |
| Heated28O | treated | organic |
| Heated32O | treated | organic |

*2.7 CAZy and Gene Product Annotation*

For each read count table, annotations from the CAZy database and gene product annotations were added. To obtain CAZy annotations, the fasta sequences for each *Acidobacteria* MAG was uploaded to the dbCAN meta server which takes nucleotide inputs of metagenomes and automatically annotates the sequences using the CAZy database [41]. The outputs include the fasta tags from the nucleotide sequences along with its corresponding HMMER annotation from the CAZy database if there is any. From here, a table join was conducted in RStudio to include CAZy annotations to the locus tags in the read count tables. Gene product annotations were also accomplished in the same manner by using a table join operator in RStudio. Annotations for the gene products of each locus tag was done by the Joint Genome Institute and the file containing the annotations for each reference MAG can be found on their genome portal. These annotations were downloaded for all 12 MAGs and added to the read count tables along with the KEGG annotations.

*2.8 GTDB-Tk Analysis*

To generate a phylogenetic tree of the 12 *Acidobacteria* MAGs in relation to other references, the GTDB-Tk was used. GTDB-Tk consists of three steps: Gene Calling (identification), aligning of the genomes with the identified markers from step one, and classifying the reference genomes to generate a reference tree from other genomes in the GTDB. The only necessary input files were the fasta sequences of the 12 *Acidobacteria* MAGs. The pipeline followed can be found on the github page for the GTDB-Tk [42]. GTDB-Tk was run on

the Unity HPC by using a bioconda install to generate a conda environment containing the GTDB-Tk packages necessary.

*2.9 iTOL Visualization*

Using the .tree output from GTDB-Tk, visualization of the phylogeny was accomplished using iTOL[34]. The original tree contained a total of 53,134 genome IDs from the GTDB. Using the pruning function, the phylogenetic tree was trimmed to contain only 62 genome identifiers, 12 of which were the *Acidobacteria* reference MAGs. Colored annotations were applied using the annotate feature found in iTOL. With the node ID provided for each leaf, the taxonomic classifications of each ID were searched using the GTDB and annotated according to its Order classification [43].

*2.10 OrthoFinder Analysis*

OrthoFinder was used to identify orthologous genes among the 12 *Acidobacteria* genomes. The pipeline used can be found on the github page for OrthoFinder [32]. Inputs for OrthoFinder include peptide sequences for the reference genomes which were previously translated from the *Acidobacteria* reference sequences. A bioconda install for OrthoFinder was completed on the Unity HPC to create a conda environment with all the required packages.

# CHAPTER 3

# RESULTS

## *3.1 Acidobacteria Phylogeny*

*Acidobacteria* is a highly diverse phylum with its 26 subdivisions. With that, it was assumed that there would be diversity among the 12 reference *Acidobacteria* used. Figure 4 shows a phylogenetic tree of the 12 Acidobacteria MAGs highlighted in yellow in comparison to other *Acidobacteria* genomes uploaded to GTDB. Of the nodes included in Figure 4, the *Acidobacteria* references fell into 3 different orders. These orders being *Acidobacteriales*, *Acidoferrales* and *Bryobacterales*. As our reference MAGs used were only classified by phylum, they were not categorized into any of these orders specifically and were instead annotated with a different color. 7 of the 12 *Acidobacteria* MAGs fell into clades with members classified as *Acidobacteriales*. 1 of the 12 fell into a clade with *Bryobacterales* and the remaining 4 were grouped into clades with *Acidoferrales*. As phylogenies display relationships between organisms based on evolutionary background, it can be assumed that organisms within the same clade share a more recent common ancestor [44]. Because of this, organisms that have a more recent common ancestor are often perceived as more genetically similar to organisms that branch off earlier or reside in different clades. This is why a separation can be seen in Figure 4 with the 3 different Orders occupying different sectors of the tree. Based on the tree, it appears that MAGs 33000780.fa.4 and 3300020743.fa.1 are the most evolutionarily similar.

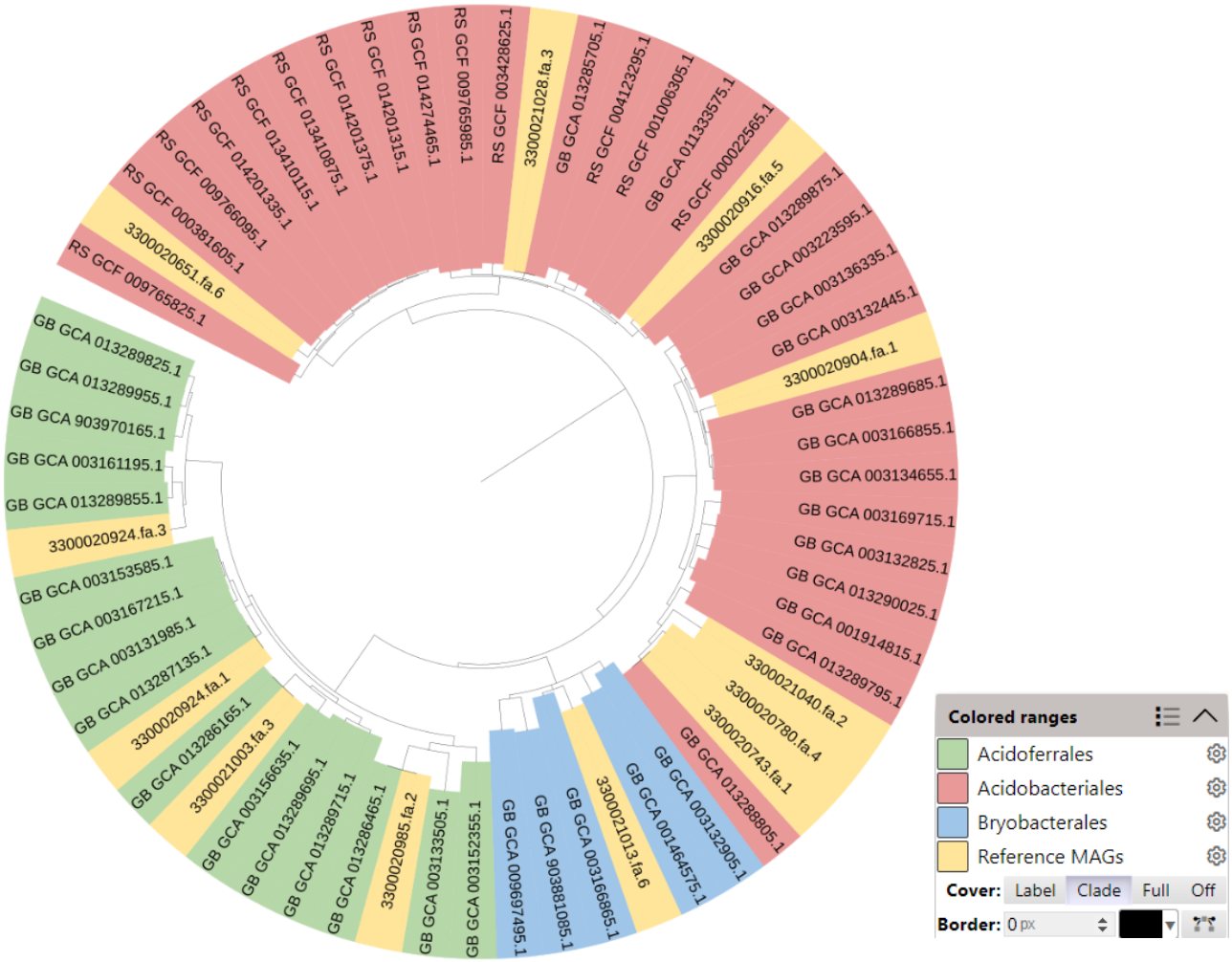



**Figure 4: Phylogenetic Tree of Acidobacteria MAGs:** The phylogenetic tree above was generated from the GTDB-Tk using the 12 *Acidobacteria* reference MAGs from Barre Woods. Originally containing 53,134 leaves, the tree was pruned to 62 leaves. All the organisms found on the tree are classified in the phylum *Acidobacteria*. Using the GTDB and the node IDs, each leaf was categorized based on order and color coded based on the key shown to the right of the phylogeny. *Acidoferrales* were labeled green, *Acidobacteriales* were labeled red, *Bryobacterales* were labeled blue and the Acidobacteria MAGs used in this study were labeled yellow. Visualization and annotation of the tree were accomplished using iTOL.

*3.2 Expressional Differences in the Acidobacteria MAGs*

Due to the diversity seen in the phylogeny, we wanted to observe expressional differences among the Acidobacteria MAGs as a whole. In order to do so, a graph was generated using the total percentage of uniquely mapped reads and number of reads from each soil sample to each *Acidobacteria* MAG. Figure 5 shows a boxplot of the findings.

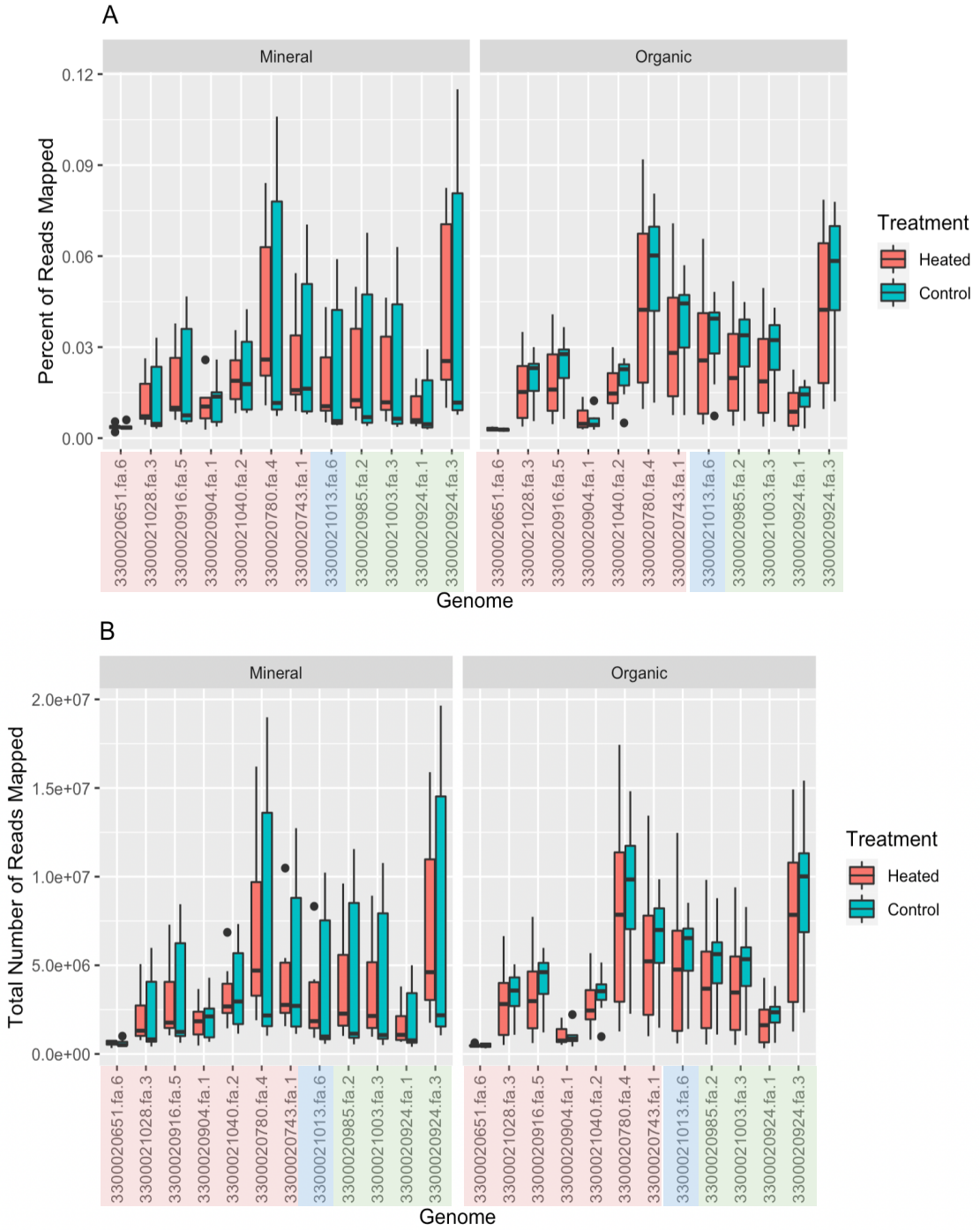


**Figure 5: Expressional Differences Between 12 Acidobacteria Reference MAGs**
(A)5a shows the percentage of uniquely mapped reads to each Acidobacteria MAG from each sample. The x-axis displays the genome ID for each Acidobacteria MAG and is colored according to the Order identified in Figure 4.. The y-axis displays the percentage of uniquely mapped reads from each sample to the reference MAG. Treatment refers to samples that were heated (red) and those that were not (blue). Samples were also split by layer with the percent of reads mapped from the mineral samples on the left and the percent of reads mapped from the organic samples shown on the right. Ggplot2 was used to generate the boxplot seen above using RStudio (B) 5b shows the same graphic as 5a, but the y-axis was changed to show the total number of reads that mapped.

The boxplot seen in 5a shows the percentage of uniquely mapped reads from each plot to each reference MAG. Only uniquely mapped read percentages were used as multimap reads can often lead to complications in downstream expressional analysis [45]. Split into Organic and Mineral layers as well as condition, the box itself represents percentages from the STAR mapping summary of each sample. This means that each box is composed of 7 different percentages as there were 7 samples for each condition (organic heated, organic control, mineral heated and minter control). Within both layers, a lower percentage of reads mapped to MAG 3300020651.fa.6 can be seen regardless of treatment. Both MAGs 3300020780.fa.4 and 3300020924.fa.3 had very similar expressional patterns with having a noticeable higher percentage of reads mapped despite being classified into different Orders. In the mineral layer, a higher mean in percentage of total reads mapped between the plots can be seen in the heated samples opposed to the control. This is indicated by the dash within the boxes. However in the organic layer, the opposite phenomenon can be seen with a higher average percentage in the control plots opposed to the heated. The average percentage of reads mapped for the genome in the mineral layer range from >0% to <3% in the heated samples and 0<3% in the control samples. Average percentage of reads mapped for the genome in the organic layer range from >0% to <4% in the heated samples and >0% to ≤~6% in the control samples. Figure 5b indicates

the same findings are 5a, however it used to help show the overall number of reads mapped. This is because the low percentages of reads mapped seen would indicate poor mapping percentages. In this case, it is known that there is contamination of other organisms in the transcript reads. Because of this, the read files contain over 100 million reads. The average number of reads mapped in the mineral layer range from >0 to < ~5million reads in the heated samples and >0 to < 3 million reads in the control samples. The average number of reads in the organic layer range from >0 to <~8 million reads in the heated samples and >0 to ≤~10 million reads in the control samples. Again, reference MAGs 3300020780.fa.4 and 3300020924.fa.3 had the highest averages and the control plots had a higher number of reads map to each genome on average opposed to the heated plots. Since a majority of bacteria reside in the organic layer and there appears to be less variance in this layer opposed to the mineral layer, we decided to focus the study on the differences of the gene expression in the organic layer specifically.

*3.3 Number of Differentially Expressed Genes*

With an idea of the overall expressional patterns of the genomes, we wanted to look at the genes that were differentially expressed. The results below only depict data from samples taken from the organic horizon. The R package DESeq2 was used to statistically test the difference in expressional data of specific genes based on the condition of soil warming. DESeq2 outputs from all reference MAGs were combined into a single table to observe any patterns across the genomes. Using three different statistical cut offs, Figure 5 shows the number of differentially expressed genes observed. The condition set was to test statistical significance differences between heated and controlled samples. All 12 reference MAGs indicated statistically significant differences in expression data at the gene level. Genome 3300020916.fa.5 indicated the highest number of differentially expressed genes at all three cut offs. 3300020916.fa.5 had 442 genes

with a p-value lower than 0.05, 45 genes with a p-adjusted value below 0.1 and 31 genes with a p-adjusted value below 0.05. The p-value comes from the Wald t-test equations and the p-adjusted

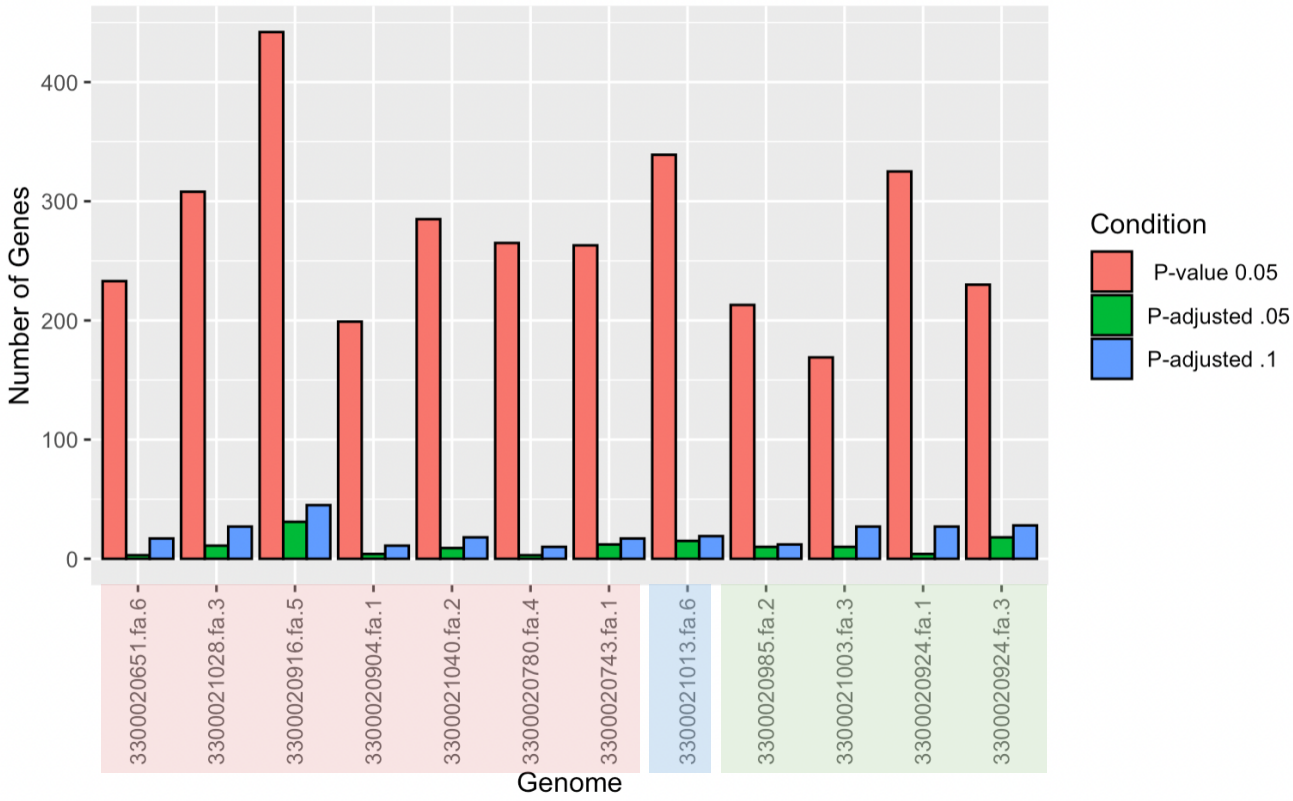



**Figure 6: Total Number of Differentially Expressed Genes From Organic Samples**
Figure 6 shows a grouped bar plot indicating the total number of differentially expressed genes at different cut offs. The x-axis shows the reference MAG genomes colored according to Order. The y-axis indicates the number of genes. The condition represents three different statistical cut offs. A Wald t-test p-value with a cut off at 0.05 is shown in red. A Benjamini and Hochberg method adjusted p-value with a cut off at .1 is shown in blue and a cut off at 0.05 is shown in green. DESeq2 was used to obtain p-values and R packages dplyr and tidyverse were used to combine data. Ggplot2 was used to generate the bar plot using RStudio

value comes from the Benjamini and Hochberg method. Based on the algorithm, it is to be expected that a decrease in the number of differentially expressed genes will be seen when adjusted. For the other 11 genomes, we see a similar pattern of the number of differentially expressed genes decreasing with a more strict cut off, but all showing at least some genes passing

each cut off. Again, this number varies from genome to genome, but most of the genomes show between ~200 to ~300 genes passing a cut off of 0.05, a statistical cut off that is typically used to discern significance in science. For an adjusted cutoff of 0.1, a range of 10-45 genes is seen passing this cut off and a range of 3-31 genes pass an adjusted cut off of 0.05. 3300021003.fa.3 had the lowest number of differentially expressed genes at the 0.05 p-value cut off with 169 total genes. 3300020780.fa.4 had the lowest number of genes differentially expressed at an adjusted cut off of 0.1 with 10 and the lowest with only 3 genes passing an adjusted cut off of 0.05 (3300020651.fa.6 also had 3 genes pass this threshold). Despite having a noticeable difference in reads mapped, 3300020651.fa.6 still showed a number of differentially expressed genes as well. Seeing a statistical difference in expression at the genic level gives evidence of changes occurring in all 12 *Acidobacteria* MAGs from long-term warming. However, the processes that are being affected cannot be discerned from looking at the number of genes itself. It is also important to note that the direction of expression cannot be discerned by only looking at the p-value. By looking at the product of the genes that are differentially expressed, some inferences can be made about the different cellular mechanisms affected and we can look to see if the genes that are differentially expressed are so at the phylum level. Using the log2 fold change value from DESeq2 will also indicate the directionality of the genes to determine whether a given gene was up or down regulated in the heated samples.

### *3.4 An Up-Regulation is Found in the Sigma 70 Gene*

We next wanted to look for any patterns in the expression of genes across the phylum as a whole. Specifically, we were looking for families of genes that were differentially expressed across the phylum or groups of genes that showed the same directionality due to the stress of

warming. To do so, all the genes with a p-value of less than 0.05 were put into a single table. Although an adjusted cut off of 0.1 is typically appropriate for larger genomic studies, there tends to be more variation in ecological research. Because of this, data will not always pass Bonferroni or Benjamini and Hochberg corrections. As this is a test to see if using MAGs for a reference based expressional analysis is viable in ecological research, we decided to include genes that pass the standard 0.05 p-value cut off. Supplemental Figure 1 is a heatmap that was used to identify genes of interest across the *Acidobacteria.* Due to difficulties visualizing all 3000+ plus genes with a 0.05 p-value cut off, the heatmap only portrays genes that passed an adjusted cut off of 0.1. If any hits were found, we then filtered out related genes within the table containing all the differentially expressed genes and continued from there. Unfortunately, we did see hits across the phylum for genes with unidentified product names. These were labeled as "hypothetical proteins". However, one interesting gene found to be differentially expressed across a majority of the MAGs was a Sigma 70 gene.

Sigma 70 falls under the category of sigma factor which are small subunits that aid RNA polymerases in bacterial organisms and help regulate transcription [46]. In total, 21 RNA polymerase sigma-70 factor genes (ECF subfamily) were found to pass a cutoff of 0.05. These genes were found to be differentially expressed in 11 of the 12 reference MAGs. Figure 7 shows a dot plot of these findings. The reference genomes can be found on the x-axis to signify what genome the gene was found to be differentially expressed in. In Figure 7, each dot represents a single gene. Although each dot represents a different gene, all the genes shown in the figure have predicted products of a Sigma 70 factor gene in the ECF subfamily with the same KEGG annotation. As condition set in DESeq2 analysis was generate to compare the warming condition to the control, a positive fold change corresponds to a higher number of reads mapped

(up-regulation) and a negative fold change corresponds to a lower number of reads mapped to that locus in the heated samples (down-regulation).

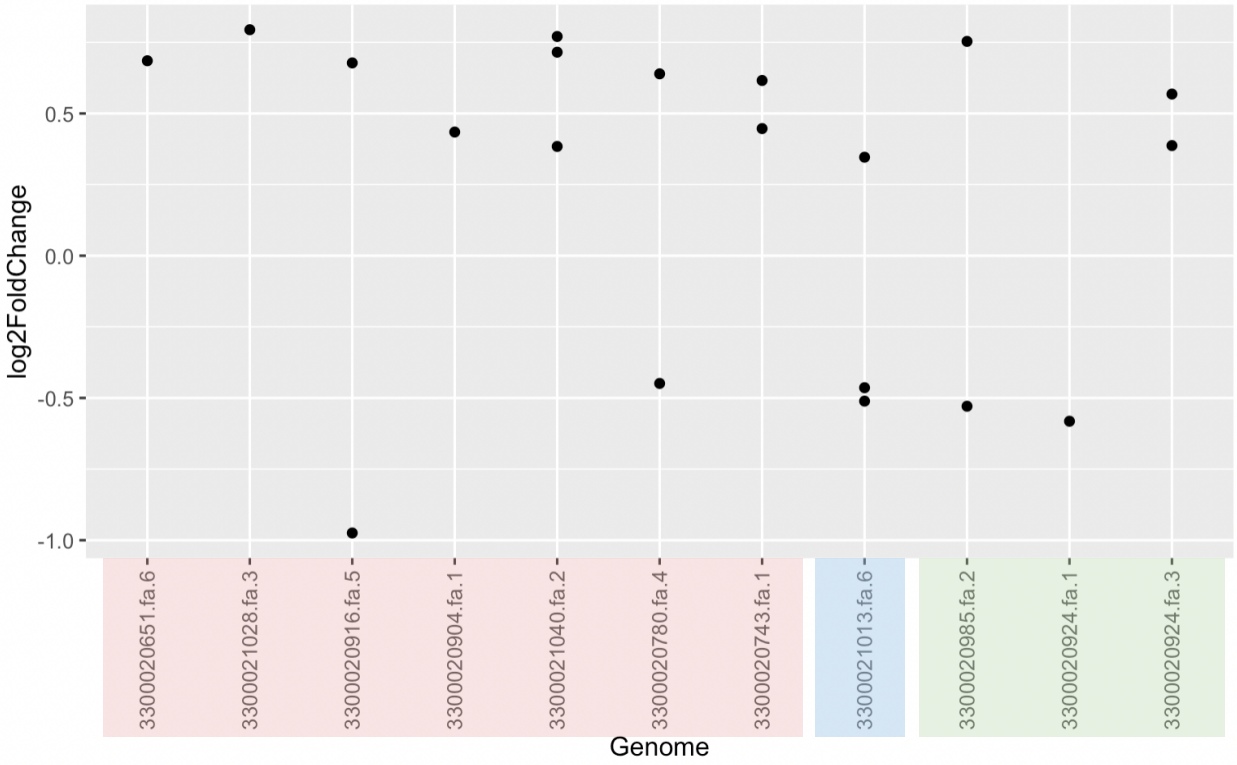



**Figure 7: Expressional Data of Sigma70 (ECF subfamily) in Acidobacteria MAGs**
Figure 7 illustrates the expression data of the Sigma70 gene found to be differentially expressed. Each dot on the graph represents a gene that passed a p-value cut off of 0.05. The x-axis represents the Acidobacteria genome that the gene was found to be differentially expressed in and is colored by Order. The y-axis represents the log2 Fold Change of the gene signifying the up or down regulation of the gene in the heated samples compared to the control.The figure above was created using ggplot2 with output data from the DESeq2 in RStudio.

Variation is seen among the *Acidobacteria* within this gene. In terms of numbers, 3300021040.fa.2 and 3300021013.fa.6 had three total sigma 70 genes found to be differentially expressed in the samples. All three genes had a positive fold change in the 3300021040.fa.2 MAG, while only one gene had a positive fold change in 3300021013.fa.6. 3300020904.fa.1 and 3300020924.fa.1 both only had one sigma 70 gene found to be differentially expressed.

3300020904.fa.1 only indicates an up-regulation in the heated conditions and 3300020924.fa.1 only indicates a down regulation. In general, the overall expressional pattern of the sigma 70 gene clusters to the top of the graph indicating a positive fold change. This is the general pattern seen among the reference genomes for this gene in particular. Besides 3300020924.fa.1 and 3300021003.fa.3, every reference shows indication of at least one sigma 70 gene being up-regulated. 6 of the 11 only show an upregulation of sigma 70, while 4 of the remaining 5 showed at least one sigma 70 gene being affected in both directions by the warming. The highest degree of log2 fold change is seen in 3300020916.fa.5 with ~-1. Besides this point, most of the other genes cluster close to 0.5 or -0.5 depending on directionality. As there could be multiple sigma 70 genes, we also wanted to include information gathered from OrthoFinder to see if we could further categorize these sigma70 genes. Supplemental Table 2 lists all the summary statistics and orthogroups from OrthoFinder. OrthoFinder sorted the 20 genes into 14 different orthogroups. Only two of the groups contained more than one of the genes being orthogroup OG0000008 and OG0001439. 5 sigma70 genes grouped into OG0000008 with three of genes up-regulated in the warming samples and two of them down-regulated. OG0001439 contained 2 genes, both of which were up-regulated in the warmed samples. In the two closely related genomes seen in Figure 4, both 3300020780.fa.4 and 3300020743.fa.1 exhibit an upregulated sigma 70 gene. However, 3300020780.fa.4 also shows one gene being down-regulated while 3300020743.fa.1 does not. Overall, there was variation found within the differentially expressed sigma 70 genes. However, seeing this one gene found in 11 of the 12 references with a general up-regulated trend is very interesting.

*3.5 Exploration of Differentially Expressed Carbohydrate Active Enzymes*

Carbohydrate metabolism was a category of interest when exploring these *Acidobacteria* genomes due to the high variation of metabolism among the phylum. Annotations from the CAZy database were used to classify the categories of these carbohydrate active enzymes as well as to identify the genes. If the genes had a CAZy annotation, they were filtered out and put into a separate table for further analysis. Again, only genes with a p-value below 0.05 were used for further analysis. Figure 8 shows a complete list of all the differentially expressed genes with CAZy annotations and there calculated log2 fold change. All information regarding

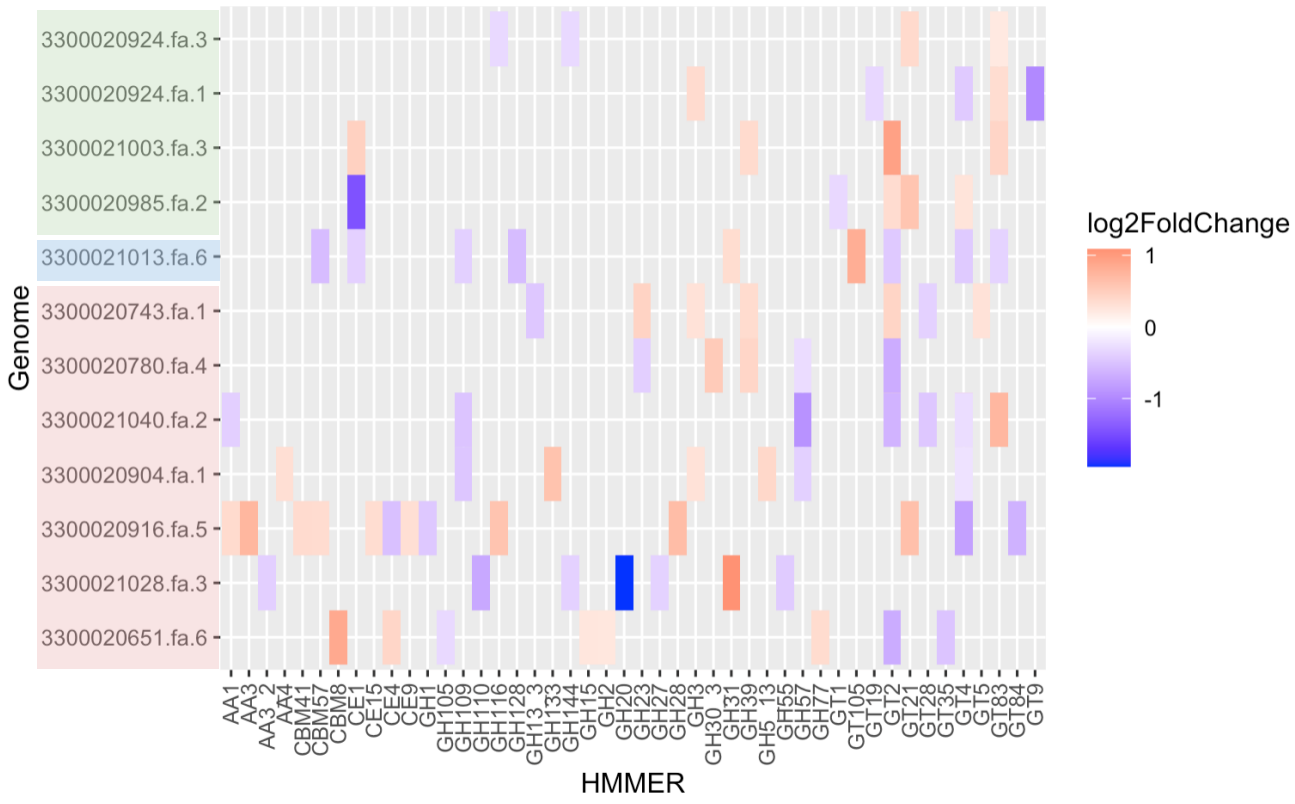


**Figure 8: Differentially Expressed Carbohydrate Metabolizing Enzymes in *Acidobacteria* MAGs**
Above is a heatmap showing a list of all differentially expressed genes with annotations in the CAZy database. On the x-axis is the HMMR annotation from the CAZy database that categorizes the type of enzyme each gene product falls under. "AA" refers to Auxiliary Activities, "CBM" refers to non-catalytic Carbohydrate Binding Molecules, "CE" refers to Carbohydrate Esterases, "GH" refers to Glycoside Hydrolases, and "GT" refers to Glycosyltranferases. The y-axis refers to the genome that gene was found to be differentially expressed in. The fill gradient corresponds to the log2 fold change with a darker shade of red corresponding to a positive fold change and a lighter shade of red/white color corresponding to a negative fold change. The heatmap was created using ggplot2 in RStudio with output data from DESeq2 and the CAZy database

Of the enzyme categories found on the CAZy database, 5 categories of carbohydrate-related metabolizing enzymes were found. 5 total genes fell into the Auxiliary Activity (AA) category, three of which had a positive fold change and two of which had a negative fold change. 4 total genes fell into the non-catalytic Carbohydrate Binding Molecule (CMB), three of which had a positive fold change. 7 total genes fell into the Carbohydrate Esterase (CE) category with 4 genes having a positive fold change, but 3 having a negative fold change. 36 total genes fell into the Glycoside Hydrolase category (GH), 18 of which show a positive fold change and 18 showing a negative fold change. Finally, 36 total genes fell into the Glycosyltransferase (GT) category, 16 of which have a positive fold change and 20 that have a negative fold change. In all, 81 total genes with CAZy annotations were found to be differentially expressed. Among the reference genomes, all 12 *Acidobacteria* had carbohydrate active genes differentially expressed. As to be expected, many glycoside hydrolase encoding genes were found across the *Acidobacteria*, but also surprising was the number of glycosyltransferases that were found to be differentially expressed. Among these categories, there was essentially an even split between the number of up and down regulated genes. However, given that *Acidobacteria* are known to have a high

percentage of GH genes and the fact that a high number of GT genes were differentially expressed, we decided to take a closer look at those categories specifically.

Figure 9a shows a barplot of the 36 Glycosyltransferase Genes. Specifically, the HMMER annotations were used as the fill to more easily identify the different glycosyltransferases with the fold change as the y-axis.

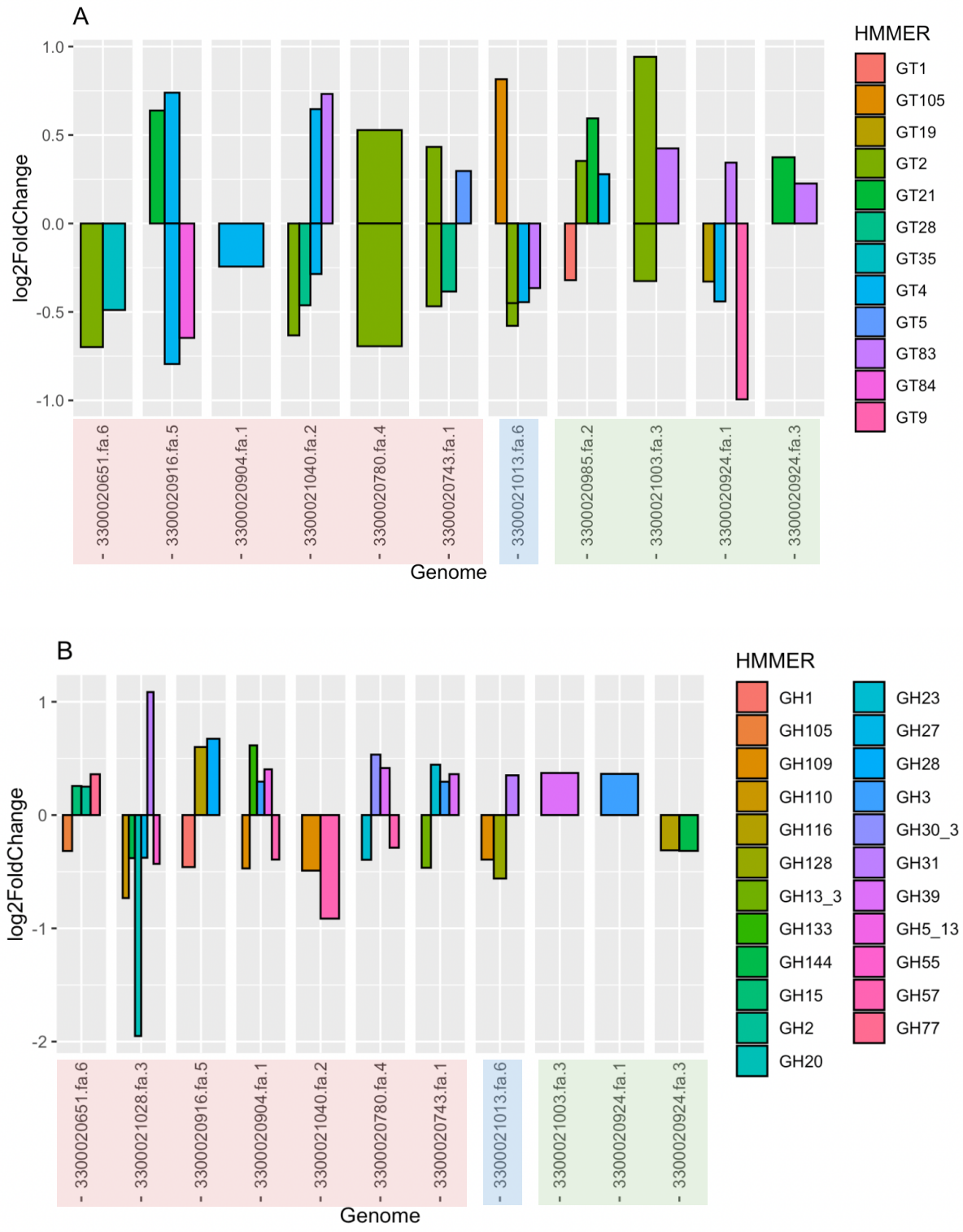


**Figure 9: Differentially Expressed GT and GH Genes in Reference MAGs**
(A) A grouped bar plot showing Glycosyltransferase genes that were found to be differentially expressed in the *Acidobacteria* MAGs. The x-axis represents the reference MAG with each MAG having its own boxed out section that is also colored by Order according to Figure 4. The y-axis represents the log2 fold change value from DESeq for each corresponding gene. Each bar represents a gene and the color corresponds to its HMMR annotation from the CAZy database. The plot was created using ggplot2 in RStudio with data from DESeq2.(B) A grouped bar plot showing Glycoside Hydrolase encoding genes that were found to be differentially expressed in the *Acidobacteria* MAGs. All the parameters are the same as in 9a other than the genes represented. The plot was created using ggplot2 in RStudio with data from DESeq2.

The diversity of expression can more easily be seen in GT genes from Figure 9a. There does not appear to be an overall pattern of expression as there appears to be an even distribution of up and down regulated genes. However, there are a few genes that mostly follow the same direction. GT83 is up-regulated in the warming plots in four different genomes: 3300021040.fa.3, 3300021003.fa.3, 3300020924.fa.1 and 3300020924.fa.3. According to the CAZy database and the gene product annotation from JGI, GT83 is a 4-amino-4 deoxy- L-arabinosyltransferase [47]. The enzyme is an integral membrane protein that modifies arabinose to generate a precursor for polymyxin resistance [48]. GT21 is another glycosyltransferase that is only found to be up-regulated in the heated samples in three different *Acidobacteria*. GT21 is a ceramide beta-glucosyltranferase that catalyzes the transfer of glucose to ceramide to produce GlcCer [49]. Only a few genes are found to only be down-regulated, such as GT28, but only 2 of the reference genomes exhibit down-regulation of this gene. Two specific glycosyltransferases also stand out, but are expressed bi-directional in the same genome and across genomes. These two genes are GT2 and GT4. 11 total GT2 genes were found to be differentially expressed with 7 being down-regulated. However, it also appears bi-directional in three different genomes. GT2 is a

cellulose synthase that is involved in cell wall synthesis [50]. 8 total GT4 genes were found to be differentially expressed with 5 being down-regulated. GT4 also was predicted to be involved in cell wall biosynthesis according to the product annotation, but is also classified as a sucrose synthase [51]. It also is bi-directional in two different reference genomes, both which are different from the two that exhibited this pattern with GT2.

Figure 9b shows a wide diversity in the expression of different glycoside hydrolase enzymes. Higher differences in expression between the two conditions are also seen within the GH genes. Genome 3300021028.fa.3 possess two different differentially expressed GH genes that are at either extreme: the most up-regulated hydrolase and the most down-regulated hydrolase. With a log2 fold change of over 1, GH31 is up-regulated in the heated samples and is also found to be up-regulated in one other genome. According to annotations, GH31 is alpha-glucosidase which essentially breaks down starches and other disaccharides to glucose [52]. The gene found in 3300021028.fa.3 particularly was further classified as a xylohydrolase. GH20 showed the opposite expressional pattern of being highly down-regulated in the heated sample in the same genome. With a log2 fold change of -1.95, GH20 is classified as a N-acetyl-beta-hexosaminidase and they act to cleave hexosamine residues into N-acetyl-beta-D-hexosaminides [53]. There are little to no consistencies found among the glycoside hydrolases within the reference genomes. 330021028.fa.3 which showed the highest number of differentially expressed GH genes indicates 5 of 6 glycoside hydrolases down-regulated. Only 11 of 12 *Acidobacteria* genomes contained any glycoside hydrolase encoding genes differentially expressed.

*3.6 Preliminary Results for Differentially Expressed Chemotaxis Related Genes*

Chemotaxis was one gene family of interest as previous results from the data set hypothesized expressional changes in the gene family. The Chemotaxis gene family encodes proteins vital for motility in bacteria. *Acidobacteria* use flagellar proteins for locomotion and members of the chemotaxis gene family are crucial for the regulation of many flagellar-like proteins [54]. Figure 10 illustrates a barplot of all differentially expressed chemotaxis-related genes discovered. More information regarding these genes can be found in Supplemental Table 3**.**

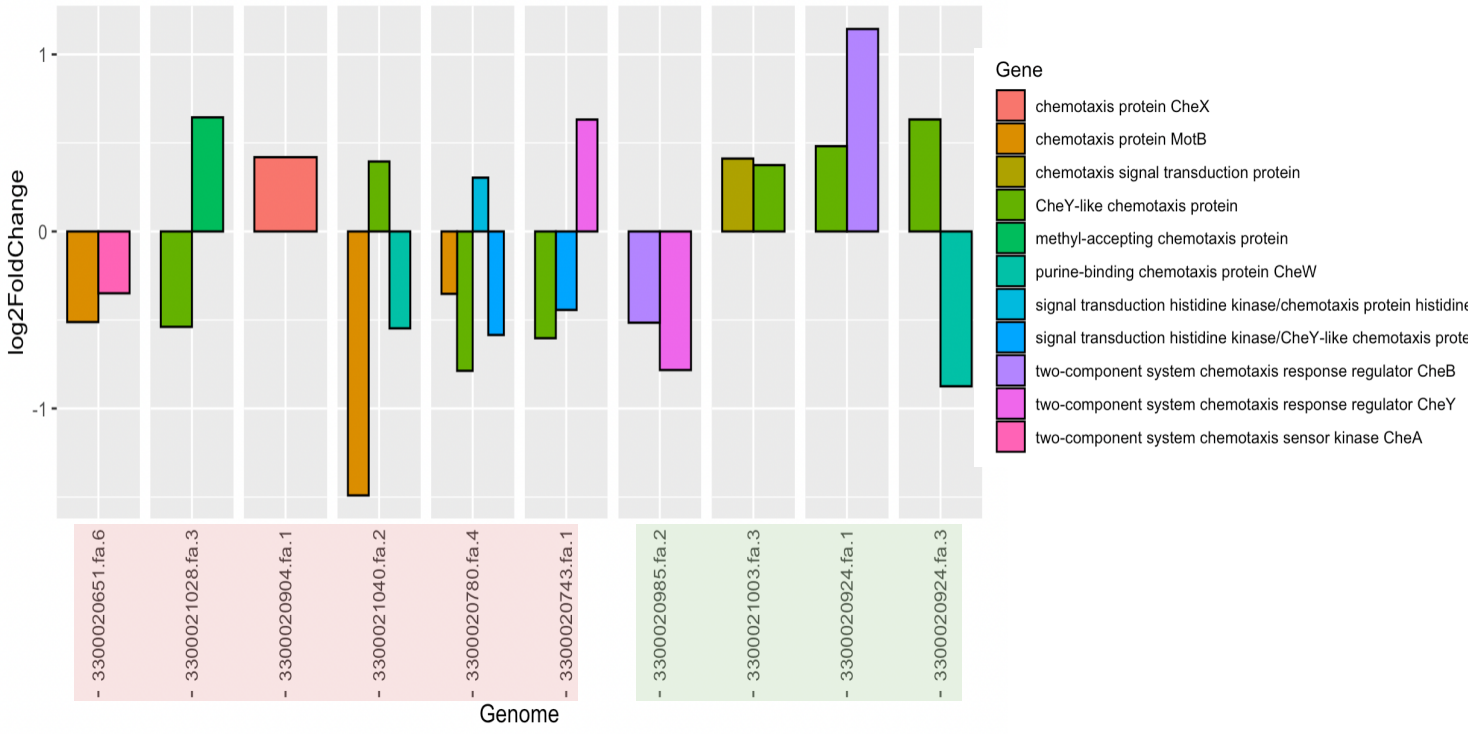



**Figure 10: Differentially Expressed Chemotaxis-Related Genes in Reference MAGs:** A grouped barplot is seen above displaying all chemotaxis-related genes found to be differentially expressed. The x-axis portrays the genome in which the genes were differentially expressed in. The y-axis represents the log2 fold change inferring the up or down-regulation under heated conditions. Each box represents a single gene and is filled as a color reflecting the predicted gene product. The figure was created using ggplot2 in RStudio in combination of data from DESeq2.

Figure 10 shows 23 total differentially expressed genes. 10 of 12 of the reference *Acidobacteria* contained at least one differentially expressed chemotaxis-related gene. The only two genomes that did not were 3300021013.fa.6 and 3300020916.fa.5. 3300020780.fa.4 possessed the highest number of genes found to be differentially expressed at 4 total and 3300020904.fa.1 had the fewest of the 10 with only one gene up-regulated in the heated environment. In total, 10 of the 23 genes were found to have a positive log2 fold change. The most common differentially expressed gene in this pool was the CheY-like chemotaxis protein which appears 7 times in Figure 10. Of the 7 times it appeared, 4 of these genes showed an up-regulation and 3 showed a down-regulation. Also related was a signal transduction histidine CheY-like gene that appeared 2 times and was down-regulated in two different genomes being 3300020780.fa.4 and 3300020743.fa.1. The gene with the predicted Chey-like chemotaxis protein was also found to be down-regulated in these two genomes. Purine binding chemotaxis protein CheW was found in only two genomes, but appeared in the same direction in both with a negative fold change. Finally, one last discernable pattern within the chemotaxis gene family involved Chemotaxis MotB. This protein was found in three different *Acidobacteria* genomes and was found to be down-regulated in all three. It not only has the highest difference in expression levels between the two conditions with a log2 fold change value of -1.49 in genome 3300021040.fa.2, but the MotB gene in 3300021040.fa.2 had the lowest p-value of all the chemotaxis related genes at 1.51E-7 making it one of three genes with an adjusted p-value below 0.1. Other chemotaxis protein coding genes are seen to be differentially expressed throughout, however it varies across the genomes.

# CHAPTER 4

# DISCUSSION AND FUTURE DIRECTIONS

## *4.1 A Reference Based Approach to Read Mapping in Acidobacteria is Possible*

As shown, read mapping and differential expression analysis does appear to be possible and viable in ecological research. Overcoming the challenges of working within a soil environment can be difficult. MAGs are becoming increasingly more common in the ecological sphere of genomics as pipelines are becoming increasingly accurate and advanced [55]. Although isolates are almost always preferred when possible and available, working with MAGs can provide information. With that being said, there are shortcomings when working with MAGs. One is the contamination of the reference sequences generated. Although programs such as CheckM are used to preliminary check the quality of the assembled genomes, it is expected that there will be contamination from other organisms in MAGs. This may lead to lower mapping percentages or the exclusion of data due to errors in the binning process. However, the method is viable and offers results to explore as can be seen by the research conducted.

*Acidobacteria* were chosen as the reference phylum to be used due to the high representation in the samples shown in Figure 2a and 2b. By choosing only mapping to those 12 high quality MAGs, it was to be expected that only a percentage of the total transcript reads would actually map to these 12 genomes. That is because the mRNA reads extracted came from all the organisms present in the samples. This explains why in Figure 5a, we see such a small percentage of total reads mapped. However, because the read files were so large containing 150+ million reads, the percentages can be deceiving. That is why 5b shows the total number of reads to indicate that read mapping was successful. From initial mapping data in Figure 2, it was to be assumed that different expressional levels would be observed as whole between the 12 reference

genomes. That is exactly what we see. *Acidobacteria* as a whole are a very diverse phylum with 26 different subgroups so we expected to see diversity within our small sample size as well. Figure 4 justifies that even at the Order level, there were evolutionary differences among the references. Figure 5 shows the difference in transcripts that mapped to each of the genomes. Genomes 3300020924.fa.3 and 3300020780.fa.4 had similar expression levels despite being classified in different Orders. Beyond that, most of the genomes displayed similar average expression levels to each other with each genome following the same trend depending on the condition and soil layer. There was one outlier being genome 3300020651.fa.6 that had a noticeably lower percentage of reads mapped. We did expect one of the 12 genomes to exhibit this behavior according to the preliminary results from Figure 2 and that is what we see. The overall differences seen may very well be attributed to diversity in the *Acidobacteria* phylum. It is known that the phylum is extremely diverse and ubiquitous so it is to be expected that different expressional patterns would be seen. However, it was a little surprising that those that were more evolutionarily similar did share more similar expression patterns as a whole. This may also be due to the environment. Although all samples taken are from the same large plot in Barre Woods, it is possible that there are ecological differences in the subplots. Some plots may be closer to larger trees or outside vegetation, while other plots may have a larger population of wildlife that changes the dynamic of the SOM.

Differences in the total number of reads mapped is seen based on the condition of soil warming. Perhaps more interesting is that the averages of the reads mapped in the heated and non-heated samples flips based on the sample layer. In the organic layer, the average number of reads mapped between all the samples is higher in the control plots compared to the heated plots. The general take away from this is that the stress of warming is causing changes in transcription

levels as a whole in the *Acidobacteria*. The result of this lower levels of transcription is that phylum of bacteria results in fewer reads mapped. However, as *Acidobacteria* often adapt to their environment, the stress of 5 ℃ is not too drastic so the changes in overall transcription, although lower, is not too different from the control. In the mineral layer, the opposite is true. On average, a higher number of reads mapped to the reference genomes in the long-term warming samples opposed to the control (Figure 4). Although surprising to see initially, this can be explained by increased mixing of the layers due to warming. Bacteria populations are known to be concentrated in the organic horizon as this is where a majority of the necessary nutrients are [56]. That is why the overall averages of reads mapped in each genome is higher in the organic samples opposed to the mineral samples, regardless of warming. Prior research in the lab has suggested that the organic and mineral layers tend to mix after warming. This mixing can be caused by a multitude of factors such as movement of larger eukaryotes causing a dispersion between the layer or other chemical weathering forcing soil mixing in the heated environment [57]. Regardless of how it happened, this theory of layers mixing in the warmed plots would explain the pattern seen in the mineral layer. If a lower percentage of bacteria live in the mineral layer, and this layer becomes mixed with the organic layer which harbors a higher population of bacteria, it would explain the increase of reads mapped. This would hypothetically cause an increase in the overall *Acidobacteria* population within the layer which would equate to a higher levels of transcription and mRNA in the mineral samples.

*4.2 A Number of Differentially Expressed Genes are Found in Acidobacteria Due to Soil Warming*

To gain a better understanding of changes brought about by long-term warming, we wanted to not only look at the overall expression, but also the number of differentially expressed

genes. Figure 6 displays the total number of differentially expressed genes. In all, 3271 genes were found to be differentially expressed if using a 0.05 Wald T-test p-value cut off. Although one cannot completely attribute these changes to long-term warming, it is reasonable to believe that the stressor of warming contributes to the differences seen. The Harvard Forest warming experiment was created to observe potential changes brought about by climate change. By being in the same general location, the overall climate and ecosystem between the plots is relatively similar. The only controlled difference is the 5 ℃ change in the soil temperature. Lab environments will always be more controlled than a natural field experiment by eliminating confounding variables, but considering the goal is to gain a better understanding of climate change, the Harvard Forest Warming plots are complemable. With that being said, we have determined that long-term warming is having an affect on the overall gene expression. Only looking at *Acidobacteria* extremely limits the scope of what is occurring in the entire microbial community. However, due to the abundance and cellular activity of this phylum in the samples, general conclusions about other bacterial members may be inferred.

When further investigating the gene products of the differentially expressed genes, we used the standard 0.05 p-value as the cut off. However, in biological multi omic studies, there are issues with this. First is that there are biases when using p-values from a standard t-test. Although DESeq2 uses an algorithm to normalize the read count and obtain p-values from these normalized and filtered count data, using arbitrary cut offs such as 0.05 can generate biases for false hits [58]. This is not to say that a 0.05 p-value cut off should not be used as it does suggest a strong correlation between the data and the condition, but it is important to recognize and address the biases. Another issue is quantifying the p-value for multiple testing scenarios. Although all variables except soil warming were controlled as best they could, it is reasonable to assume that

multiple variables can be accounting for the differences resulting in false positives. This is pertinent in multi omic studies with large biological datasets [58]. To account for this, corrections are made to generate an adjusted p-value such as Bonferroni corrections, or the Benjamini and Hochberg method which is a slightly adjusted algorithm from Bonferroni that DESeq2 uses. When analyzing the differential expression data, all these ideas were taken into consideration along with the high number of samples (genes) tested. To initially identify candidate gene families to further investigate, an adjusted p-value of 0.1 was used to look for any patterns seen across the phylum as whole. Although some related genes were identified with a 0.1 adjusted cut-off, it was difficult to discern consistencies across the phylum. One answer to this could be the vast diversity in *Acidobacteria.* As abundant as they were in the samples, each individual species identified is filling different niches in its environment and expressing different gene patterns. Just because they all fall into the same phylum does not mean a specific up or down regulation of a gene family will be seen, especially considering the diversity found in *Acidobacteria* specifically. Another reasonable explanation is the issues when working within an ecologic study. Field experiments and ecological studies do not always pass standard Bonferroni correction criteria due to the variability. In controlled lab environments, correcting for multiple independent variables makes sense as most variables are controlled. In this case, proximity was used to control for all variables except soil warming. Unfortunately, the weather conditions and nature can be extremely erratic, and the Barre Woods plots are no exception. Because of this, we made sure to take adjusted p-values into account when investigating the data set, but the cut off of 0.05 for the standard p-value was used. We believe that using this as a cut off was reasonable considering the source of the genomic and transcriptomic data. All p-values and adjusted p-values of genes identified in Figures 7-10 are also provided to maintain full transparency.

*4.3 Up-regulation of Sigma70 May Explain High Abundance of Acidobacteria in Our Samples*

One of the genes found to be differentially expressed across 11 of the 12 genomes was a sigma 70 ECF subfamily gene. Sigma 70 is an important housekeeping sigma factor that helps regulate transcription. Members of this family specifically direct RNA polymerase 10 to 35 base pairs upstream of the initiation site for a gene. In 10 of the 12 genomes, there was at least one instance of a sigma 70 gene being upregulated [46]. Due to the importance of sigma 70, it was interesting to see this gene differentially expressed. Of note is that there are many different sigma 70 genes in a genome that account for different housekeeping genes. Because of this, some may be more conserved than others. One method used to identify if there were differences among the sigma 70 genes was to organize them into orthogroups. The idea was that sigma 70 factors that regulated similar cellular processes would be orthologous to one another across genomes. Although five of the genes were found to be orthologous to one another, that particular gene was found to be both up and down regulated depending on the genome. Again, this can be attributed to the diversity seen between *Acidobacteria* as they may implore different survival strategies. Sigma factors are also known to be used as stress regulators in bacteria [59]. Due to their properties of controlling important pathways for survival, levels of sigma factors have been shown to change due to various stressors. For example, sigma 32 is a known heat shock protein found in *E. coli* that regulates a heat shock response to allow the organisms to acclimate to the stress of warming [60]. Studies identified that sigma 32 shares many similarities in binding affinity with sigma 70 [61]. With all this in mind, we hypothesize that this up-regulation of sigma 70 genes may explain the high representation of *Acidobacteria* in our samples. Sigma 70 does not only play important roles in the regulation of key cellular processes, but it shares similarities with other sigma factors that have shown to be up-regulated as a stress response to heat. Also, initial read

mapping data from Figure 2 indicates a high number of genomic reads mapping to the *Acidobacteria* phylum. Although this mapping did not separate heated and non-heated reads, it is reasonable to assume that these bacteria would have to be well represented in both sets to cluster at the top and appear to have such a disproportionately high number of genomic reads map. If sigma 70 is up-regulated in a majority of the genomes, then this suggests that *Acidobacteria* are adapting to the heated stress and able to continue reproducing and thrive despite warming. However, as this is only at the transcriptomic level, further proteomic studies would need to be conducted to confirm this hypothesis.

### *4.4 Carbohydrate Active Enzymes are Found to Differentially Expressed in Some Ways*

*Acidobacteria* have a large number of glycoside hydrolase encoding genes. They are also divided into many subgroups due to the diversity among the phylum in carbohydrate metabolism. As one major question regarding soil bacteria with climate change is differences among carbon recycling and thus impacts on metabolic pathways, we knew we wanted to explore these genes particularly. Using the CAZy database, a number of carbohydrate active genes were identified (Figure 8). As expected, we saw a number of genes both up and down regulated in the heated samples. In Figure 9a, an even split of 18 glycosyltransferase genes were found to be up and down regulated. As one of the driving characteristics that separate the phylum into subgroups is the substrates for carbohydrate metabolism, a lot of diversity was expected and that is what is seen. Based on the categories and directionality of the glycosyltransferases, a lot of different strategies in response to the soil warming can be seen. Ultimately, there were no specific glycosyltransferases that were differentially expressed across all 12 genomes. The most

frequently appearing glycosyltransferase was categorized as a GT2 and the synthesized gene product is a cellulose synthase. Three different genomes expressed two different GT2-categorized genes bi-directionality. Of the three genomes, two of them were 3300020743.fa.1 and 3300020780.fa.4 which are the most evolutionary similar according to the phylogeny in Figure 4. Due to this, one assumption made is that these two genomes may implore similar survival strategies at the carbohydrate metabolizing level. Unfortunately, only one similarity is not enough to make any conclusions regarding this. Generally, we do not see many other consistencies among the phylum. 4-amino-4 deoxy - L-arinonlytransferase encoding genes was found to be up-regulated in four different genomes. This is believed to be a general stress response to the warming as the protein product catalyzes reactions to form intermediates important for polymyxin resistance. Polymyxin is a polypeptide with antibiotic effects [62]. We hypothesize that the overall warming conditions are causing a general stress response in these bacteria. One defense mechanism is to up-regulate genes that increase resistance to certain harmful chemicals. Although polymyxin levels are not elevated in the soil samples, the *Acidobacteria* are responding to the stress of heating by increasing general resistance.

Figure 9b explores the opposite end of the carbohydrate metabolizing enzymes, specifically the glycoside hydrolases. From a general perspective, we do see that long-term warming has an impact on these types of genes. Again, due to the broad spectrum of the phylum and the diversity specifically in GH encoding genes in *Acidobacteria,* a lot of differences are seen among our small sample of references. Even among genomes 3300020743.fa.1 and 3300020780.fa.4, only one glycoside hydrolase was expressed in the same direction. Given that we know different subfamilies tend to be more involved in carbohydrate metabolism than others, it is not surprising that we see a discrepancy among the number of differentially expressed GH

genes as well. The takeaways from Figure 9b are that soil warming is affecting the transcription of these carbohydrate metabolizing enzymes to some degree. This may in turn affect the overall metabolizing of carbohydrates as whole, depending on the *Acidobacteria.* Unfortunately, without further classification into the subdivisions of the phylum, it is difficult to make any inferences on the data that is seen. For example if, it is known that genome 3300020924.fa.1 falls into a subdivision that can only metabolize specific beta-D-glucosides then it would make sense to see an up-regulation of a GH31 gene in order to outcompete for the carbon source. Further classifications of the Acidobacteria would provide more insight regarding these enzymes. One interesting observation is the pattern seen in genome 3300021028.fa.3. Specifically 6 different glycoside hydrolases were found to be differentially expressed, 5 of which are down-regulated. Also, it was the only reference genome not to show any differentially expressed glycosyltransferases. Another possibility as to no phylum wide associations can be seen is that these genes are highly conserved. Carbon is an essential component to life. Any genetic changes to carbon metabolizing enzymes may not be tolerated and thus we do not see a difference in expression patterning from warming. Ultimately, further research into these genes specifically is necessary.

*4.5 Changes in Chemotaxis Related Genes Suggest Drying of Soil*

Genes of the chemotaxis family are essential for proper locomotion in bacteria. We saw a number of chemotaxis related genes differentially expressed in the *Acidobacteria* after different degrees. However, these differences in expression levels were some of the most statistically significant seen. Although the gene encoding for chemotaxis protein MotB was only seen in 3 of the 12 genomes, it was vastly down regulated in the warmed samples and with high significance. The MotB protein is an outer membrane chemotaxis protein [63]. In *E. coli,* the MotB protein was

found to be one of eight torque generators in the flagellar motor and more specifically may help anchor the machinery to the cell wall of the organism [64]. Interestingly, in the three genomes that MotB was down-regulated in, genes of the GT4 family involved in cell wall synthesis were also down-regulated in those genomes. It is possible that the down regulation of the GT4 carbohydrate enzymes may have an impact on the production of the MotB protein due to mechanical discrepancies in cell wall synthesis and anchoring. A gene encoding for a CheY-like chemotaxis protein was also found to be differentially expressed a number of times. Although it appeared in both directions, CheY proteins are important signaling proteins for directional movement of flagellar proteins [65]. A number of these signaling proteins also are differentially expressed in the heated samples being CheX, CheA, CheW and ChB, but CheY has the highest frequency. Coordination of the flagellar proteins are important for these bacteria to move around in their environment. Preliminary results of the chemotaxis gene family do indicate changes occurring at the transcription level in response to warming temperatures. The Che-like signal proteins appearing in both directions does suggest that Acidobacteria are responding differently to the stressor, but changes are occurring. Genomes 3300020780.fa.4 and 3300020743.fa.1 were shown to be the most evolutionary similar of the reference genomes. Between these two genomes, we do see the CheY-like protein encoding gene down-regulated. Due to CheY's importance in the functionality of the flagella, it is possible that these two organisms are showing decreased levels of movement. Again, there are no trends other than MotB that are consistent across the phylum. However, one assumption made is that these two organisms are more closely related than the others so they would initiate similar responses. That is not to say that the other *Acidobacteria* found in our sample are not showing changes in their levels of movement, but further investigation is required. There was a fairly even split in terms of directionality among

the chemotaxis-related genes, but 13 of the 23 were down-regulated. Although the split is essentially evenly distributed, one possibility as to why we see a down-regulation of these genes is due to drying of the soil in the heated conditions. With soil warming, the temperature is not the only aspect being affected as the moisture of the soil may have been affected. Higher temperatures could lead to less moisture in the soil. This would make it more difficult for bacteria to propel through the environment. Due to this, one survival strategy may be to conserve resources and down-regulate important proteins for locomotion as the energy expenditure is too great. Again, this is a hypothesis and cannot be stated as fact by a transcriptomic study. Further proteomic and field work to test soil moisture would be required.

*4.6 Future Directions*

Ultimately, we can conclude that there are changes at the transcription level due to long-term warming. However, there is still much work to be done. Further classification of the 12 *Acidobacteria* MAGs into their subfamilies would allow for deeper analysis of the differentially expressed genes, especially those involved in carbohydrate metabolism. The study conducted was necessary to make initial observations. Due to constraints, only broad conclusions can be drawn without more knowledge. The phylum of *Acidobacteria* is very broad. Although the widespread nature makes them good candidates to study for generalizations, the diversity can make it difficult to come to specific conclusions. Classifications of the subfamilies should be the next taken for the continuation of the study.

Making phylum wide conclusions may also not be as effective. Although a standard in ecological prokaryotic research, phylum-wide generalizations may not be the best route. Acidobacteria could be an exception to this practice as the diversity is uncanny compared to other phyla, but it is something to consider. One possible solution would be to follow a similar

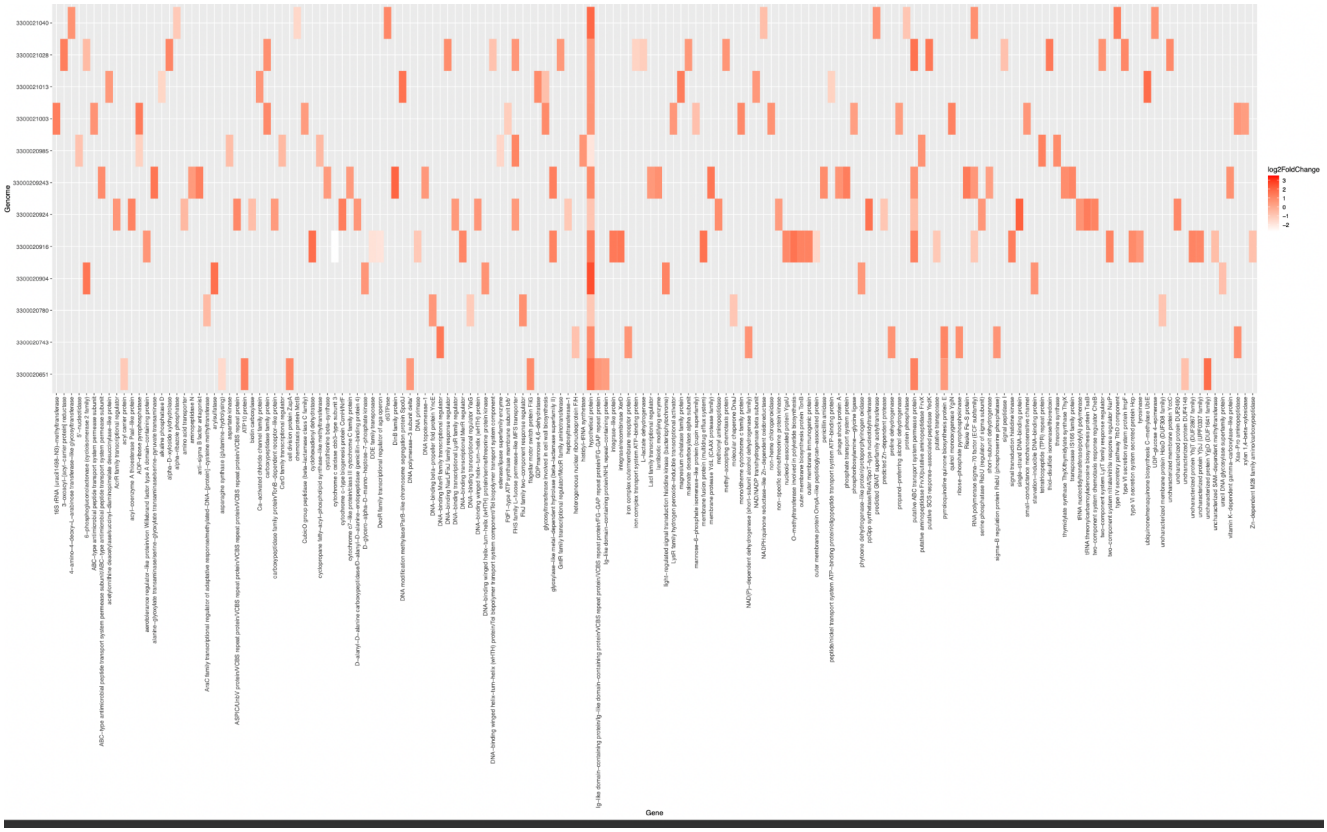
pipeline with a different phylum represented in the samples. For example, *Bacteroidetes* were the second most abundant phylum found within the samples. A reference based approach using *Bacteroidetes* would not only give insight to effectiveness of these methods in the ecological sphere of soil biology, but also help confirm conclusions drawn from *Acidobacteria*. Another possible avenue to take is to increase the sample size. Only 12 *Acidobacteria* MAGs were used as only 12 MAGs were generated from the mini-metagenomics process. However, bulk-metagenomics did produce a number of high-quality MAGs, abet far less than the mini-metagenomics pipeline. Some of these MAGs were identified as *Acidobacteria*. If these MAGs were incorporated into this study, more generalizations and conclusions may be apparent. Using isolates would be the ideal route to take as it would eliminate the possibility of contamination and incompletion when using MAGs. Unfortunately this is very difficult to accomplish when working with soil and even more difficult when working with *Acidobacteria*.

APPENDIX

SUPPLEMENTARY MATERIAL

**Supplemental Table 1: Condition Table for Mineral Layer DESeq2 Analysis**

| Plot | condition | type |
|---|---|---|
| Control12M | untreated | mineral |
| Control14M | untreated | mineral |
| Control7M | untreated | mineral |
| Control30M | untreated | mineral |
| Control19M | untreated | mineral |
| Control27M | untreated | mineral |
| Control4M | untreated | mineral |
| Heated2M | treated | mineral |
| Heated4M | treated | mineral |
| Heated11M | treated | mineral |
| Heated17M | treated | mineral |
| Heated26M | treated | mineral |
| Heated28M | treated | mineral |
| Heated32M | treated | mineral |




**Supplemental Figure 1: Heatmap of Differentially Expressed Genes at Padjusted 0.1 Cutoff**
The heatmap above displays all the genes found to be differentially expressed in the 12 reference MAGs that passed an adjusted cut-off of 0.1. The x-axis displays the predicted gene product for each gene. These annotations can be found on the JGI Genome Portal. The y-axis represents the *Acidobacteria* genome that each gene was found in. Each colored box represents a single gene and the gradient is representative of its log2 fold change. A positive fold change is indicated by a darker red and a negative value is indicated by a lighter white. The plot above was made using ggplot2 in RStudio.

## Supplemental Table 2: List of All Differentially Expressed Sigma70 Genes

| Genome | Locus_Tag | baseMean | log2FoldChang | lfcSE | stat | pvalue | padj | Gene | KEGG | HMMR | Orthogroup |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3300020651.fa | Ga0206911_13 | 44.83402 | 0.6853957 | 0.2499494 | 2.742138 | 0.006104063 | 0.31910066 | RNA polymeras | KO:K03088 | NA | OG0000722 |
| 3300020743.fa | Ga0207044_10 | 49.28705 | 0.6160554 | 0.2781822 | 2.214576 | 0.026789214 | 0.50223799 | RNA polymeras | KO:K03088 | NA | OG0000168 |
| 3300020743.fa | Ga0207044_10 | 41.5749 | 0.4470006 | 0.2272366 | 1.967115 | 0.049169909 | 0.60290947 | RNA polymeras | KO:K03088 | NA | OG0002253 |
| 3300020780.fa | Ga0206956_11 | 64.92007 | -0.4489006 | 0.2104061 | -2.133496 | 0.032884031 | 0.65479913 | RNA polymeras | KO:K03088 | NA | OG0003260 |
| 3300020780.fa | Ga0206956_11 | 48.49522 | 0.6394682 | 0.3074985 | 2.079582 | 0.037563908 | 0.65939461 | RNA polymeras | KO:K03088 | NA | OG0001807 |
| 3300020904.fa | Ga0207088_14 | 137.1845 | 0.4347811 | 0.1974738 | 2.201715 | 0.027685422 | 0.56080844 | RNA polymeras | KO:K03088 | NA | NA |
| 3300020916.fa | Ga0207021_10 | 210.16957 | -0.9745782 | 0.2327328 | -4.187542 | 2.82E-05 | 0.01825741 | RNA polymeras | KO:K03088 | NA | OG0000008 |
| 3300020916.fa | Ga0207021_10 | 52.48678 | 0.6778102 | 0.34021 | 1.992329 | 0.046335015 | 0.6716272 | RNA polymeras | KO:K03088 | NA | OG0001439 |
| 3300020924.fa | Ga0207015_10 | 30.28053 | -0.5817399 | 0.2210525 | -2.631683 | 0.008496316 | NA | RNA polymeras | KO:K03088 | NA | OG0001857 |
| 3300020924.fa | Ga0207015_11 | 64.7007 | 0.5682898 | 0.1609373 | 3.531125 | 0.000413796 | 0.04753781 | RNA polymeras | KO:K03088 | NA | OG0004513 |
| 3300020924.fa | Ga0207015_10 | 40.57755 | 0.3871949 | 0.1949673 | 1.985948 | 0.047039094 | NA | RNA polymeras | KO:K03088 | NA | OG0000009 |
| 3300020985.fa | Ga0206813_10 | 70.16876 | -0.5290554 | 0.1582303 | -3.343578 | 0.000827054 | 0.13525178 | RNA polymeras | KO:K03088 | NA | OG0000008 |
| 3300020985.fa | Ga0206813_10 | 53.5392 | 0.7535095 | 0.2704282 | 2.786357 | 0.005330413 | 0.29395006 | RNA polymeras | KO:K03088 | NA | OG0000008 |
| 3300021013.fa | Ga0206831_14 | 56.00661 | -0.5110628 | 0.1846712 | -2.76742 | 0.005650187 | 0.2951096 | RNA polymeras | KO:K03088 | NA | OG0000169 |
| 3300021013.fa | Ga0206831_10 | 64.48872 | -0.4639221 | 0.2242403 | -2.068861 | 0.038559104 | 0.56199605 | RNA polymeras | KO:K03088 | NA | OG0000468 |
| 3300021013.fa | Ga0206831_12 | 66.08124 | 0.3462988 | 0.1729853 | 2.001897 | 0.045295813 | 0.58958089 | RNA polymeras | KO:K03088 | NA | OG0000008 |
| 3300021028.fa | Ga0206863_10 | 43.23856 | 0.7945087 | 0.2587468 | 3.070603 | 0.002136268 | 0.18675728 | RNA polymeras | KO:K03088 | NA | OG0001915 |
| 3300021040.fa | Ga0206819_10 | 78.12383 | 0.7707349 | 0.2056513 | 3.747775 | 0.00017841 | 0.06597615 | RNA polymeras | KO:K03088 | NA | OG0001501 |
| 3300021040.fa | Ga0206819_10 | 49.49674 | 0.3841296 | 0.1694733 | 2.266609 | 0.023414137 | 0.53690679 | RNA polymeras | KO:K03088 | NA | OG0001439 |
| 3300021040.fa | Ga0206819_10 | 66.29 | 0.7154742 | 0.3416834 | 2.093968 | 0.036262788 | 0.57346368 | RNA polymeras | KO:K03088 | NA | OG0000008 |

## Supplemental Table 3: List of All Differentially Expressed Chemotaxis-Related Genes

| Genome | Locus_Tag | baseMean | log2FoldChang | lfcSE | stat | pvalue | padj | Gene |
|---|---|---|---|---|---|---|---|---|
| 3300021040.fa | Ga0206819_10 | 124.65767 | -1.4911356 | 0.283923 | -5.251901 | 1.51E-07 | 0.000556686 | chemotaxis protein MotB |
| 3300021028.fa | Ga0206863_10 | 92.26957 | 0.6443414 | 0.1776072 | 3.627902 | 2.86E-04 | 0.072937249 | methyl-accepting chemotaxis protein |
| 3300020924.fa | Ga0207015_13 | 133.64663 | 1.1436143 | 0.3496594 | 3.270652 | 1.07E-03 | 0.094570142 | two-component system chemotaxis response regulator CheB |
| 3300020780.fa | Ga0206956_12 | 36.47249 | -0.787305 | 0.2658971 | -2.960939 | 3.07E-03 | 0.273207663 | CheY-like chemotaxis protein |
| 3300020985.fa | Ga0206813_10 | 29.05863 | -0.7827674 | 0.2768519 | -2.827387 | 4.69E-03 | 0.288889706 | two-component system chemotaxis response regulator CheY |
| 3300020743.fa | Ga0207044_10 | 39.85465 | 0.6320509 | 0.2347294 | 2.692679 | 7.09E-03 | 0.304037988 | two-component system chemotaxis response regulator CheY |
| 3300021040.fa | Ga0206819_10 | 35.01451 | -0.5474555 | 0.2084956 | -2.625741 | 8.65E-03 | 0.373756721 | purine-binding chemotaxis protein CheW |
| 3300020780.fa | Ga0206956_10 | 302.04505 | -0.584793 | 0.2218694 | -2.635753 | 8.40E-03 | 0.383152965 | signal transduction histidine kinase/CheY-like chemotaxis protein/DNA-binding r |
| 3300020743.fa | Ga0207044_10 | 346.77788 | -0.4432887 | 0.1798838 | -2.464306 | 1.37E-02 | 0.394243697 | signal transduction histidine kinase/CheY-like chemotaxis protein/DNA-binding r |
| 3300021003.fa | Ga0206966_11 | 91.55202 | 0.3746239 | 0.1732176 | 2.162736 | 3.06E-02 | 0.458062509 | CheY-like chemotaxis protein |
| 3300021040.fa | Ga0206819_10 | 71.31832 | 0.3949947 | 0.1712537 | 2.306489 | 2.11E-02 | 0.523412497 | CheY-like chemotaxis protein |
| 3300020904.fa | Ga0207088_15 | 49.38546 | 0.4194366 | 0.1844185 | 2.274373 | 2.29E-02 | 0.533204499 | chemotaxis protein CheX |
| 3300021003.fa | Ga0206966_11 | 41.34524 | 0.4115965 | 0.2083876 | 1.975149 | 4.83E-02 | 0.54625215 | chemotaxis signal transduction protein |
| 3300020985.fa | Ga0206813_12 | 38.95439 | -0.5153258 | 0.259554 | -1.985428 | 4.71E-02 | 0.6032541 | two-component system chemotaxis response regulator CheB |
| 3300021028.fa | Ga0206863_10 | 23.22235 | -0.5389377 | 0.2739074 | -1.967591 | 4.91E-02 | 0.640702472 | CheY-like chemotaxis protein |
| 3300020651.fa | Ga0206911_11 | 105.92109 | -0.5119165 | 0.2460983 | -2.08013 | 3.75E-02 | 0.670066432 | chemotaxis protein MotB |
| 3300020780.fa | Ga0206956_10 | 57.18623 | -0.3525201 | 0.1736721 | -2.029803 | 4.24E-02 | 0.67030257 | chemotaxis protein MotB |
| 3300020651.fa | Ga0206911_10 | 119.02198 | -0.3492074 | 0.1734819 | -2.012933 | 4.41E-02 | 0.684695191 | two-component system chemotaxis sensor kinase CheA |
| 3300020780.fa | Ga0206956_10 | 205.07414 | 0.303675 | 0.1543804 | 1.967057 | 4.92E-02 | 0.686222781 | signal transduction histidine kinase/chemotaxis protein histidine kinase CheA |
| 3300020924.fa | Ga0207015_13 | 34.57761 | -0.8746483 | 0.2233475 | -3.916087 | 9.00E-05 | *NA* | purine-binding chemotaxis protein CheW |
| 3300020924.fa | Ga0207015_10 | 41.28845 | 0.632679 | 0.2374175 | 2.664837 | 7.70E-03 | *NA* | CheY-like chemotaxis protein |
| 3300020924.fa | Ga0207015_10 | 50.12058 | 0.4816927 | 0.2259182 | 2.132155 | 3.30E-02 | *NA* | CheY-like chemotaxis protein |
| 3300020743.fa | Ga0207044_10 | 25.14581 | -0.602941 | 0.3073058 | -1.962023 | 4.98E-02 | *NA* | CheY-like chemotaxis protein |

# Supplemental Table 4: List of All Differentially Expressed CAZy Genes

| Genome | Locus_Tag | baseMean | log2FoldChang | lfcSE | stat | pvalue | padj | Gene | KEG | HMMER |
|---|---|---|---|---|---|---|---|---|---|---|
| 3300021040.fa | Ga0206819_10 | 137.914891 | -0.3957849 | 0.17978242 | -2.201466 | 2.77E-02 | 0.5507843 | FtsP/CotA-like | COG2132 | AA1 |
| 3300020916.fa | Ga0207021_10 | 48.210168 | 0.3666728 | 0.17595394 | 2.083913 | 3.72E-02 | 0.63198662 | FtsP/CotA-like | COG2132 | AA1 |
| 3300021028.fa | Ga0206863_14 | 121.330325 | -0.4013043 | 0.17171918 | -2.33698 | 1.94E-02 | 0.52020992 | choline dehydr | COG2303,COG | AA3_2 |
| 3300020916.fa | Ga0207021_10 | 49.318569 | 0.7336857 | 0.32144021 | 2.282495 | 2.25E-02 | 0.519346 | choline dehydr | COG2303 | AA3 |
| 3300020904.fa | Ga0207088_11 | 170.285917 | 0.3253642 | 0.15529989 | 2.09507 | 3.62E-02 | 0.6064363 | glycolate oxida | KO:K00104 | AA4 |
| 3300020916.fa | Ga0207021_10 | 131.017012 | 0.3636983 | 0.14754804 | 2.464949 | 1.37E-02 | 0.41588048 | pullulanase | KO:K01200 | CBM41 |
| 3300020916.fa | Ga0207021_10 | 61.971771 | 0.3551735 | 0.15500911 | 2.291307 | 2.19E-02 | 0.51764536 | malectin (di-gl | pfam11721 | CBM57 |
| 3300021013.fa | Ga0206831_11 | 384.656441 | -0.5576787 | 0.19790608 | -2.817896 | 4.83E-03 | 0.27746186 | malectin (di-gl | pfam11721,pfa | CBM57 |
| 3300020651.fa | Ga0206911_12 | 529.464139 | 0.8645862 | 0.39316722 | 2.199029 | 2.79E-02 | 0.62742783 | glycosyl hydrol | pfam12891,pfa | CBM8 |
| 3300021013.fa | Ga0206831_10 | 147.133852 | -0.3871433 | 0.18834673 | -2.055482 | 3.98E-02 | 0.56894069 | dipeptidyl amir | COG1506 | CE1 |
| 3300021003.fa | Ga0206966_10 | 268.544633 | 0.4688378 | 0.18979801 | 2.470193 | 1.35E-02 | 0.33468034 | dipeptidyl amir | COG1506,COG | CE1 |
| 3300020985.fa | Ga0206813_11 | 120.372116 | -1.4558707 | 0.30157249 | -4.827598 | 1.38E-06 | 0.00365789 | esterase/lipase | COG4947 | CE1 |
| 3300020916.fa | Ga0207021_10 | 81.958741 | 0.3472349 | 0.15027392 | 2.31068 | 2.09E-02 | 0.50835187 | hypothetical pr | Hypo-rule appl | CE15 |
| 3300020651.fa | Ga0206911_10 | 84.163744 | 0.4155946 | 0.19354027 | 2.147329 | 3.18E-02 | 0.65828253 | peptidoglycan/ | COG0726 | CE4 |
| 3300020916.fa | Ga0207021_10 | 18.937189 | -0.5289919 | 0.26776412 | -1.975589 | 4.82E-02 | 0.68337994 | peptidoglycan/ | COG0726 | CE4 |
| 3300020916.fa | Ga0207021_10 | 40.830783 | 0.3245092 | 0.1619036 | 2.004336 | 4.50E-02 | 0.66433819 | N-acetylglucos | KO:K01443 | CE9 |
| 3300020916.fa | Ga0207021_13 | 68.657673 | -0.4591409 | 0.21521868 | -2.133369 | 3.29E-02 | 0.60275442 | beta-glucosida | KO:K05350 | GH1 |
| 3300020651.fa | Ga0206911_10 | 80.620157 | -0.3171103 | 0.15562026 | -2.037718 | 4.16E-02 | 0.67319707 | rhamnogalactu | COG4225 | GH105 |
| 3300021013.fa | Ga0206831_12 | 60.810868 | -0.3932721 | 0.17608828 | -2.23338 | 2.55E-02 | 0.48915622 | predicted dehy | COG0673 | GH109 |
| 3300021040.fa | Ga0206819_10 | 122.418064 | -0.490905 | 0.1845496 | -2.660017 | 7.81E-03 | 0.35777398 | predicted dehy | COG0673 | GH109 |
| 3300020904.fa | Ga0207088_15 | 296.84535 | -0.4710538 | 0.18117995 | -2.599922 | 9.32E-03 | 0.42577089 | myo-inositol 2- | KO:K00010 | GH109 |
| 3300021028.fa | Ga0206863_10 | 173.016463 | -0.7328097 | 0.34760777 | -2.108151 | 3.50E-02 | 0.61125243 | hypothetical pr | Hypo-rule appl | GH110 |
| 3300020916.fa | Ga0207021_10 | 44.106866 | 0.6005588 | 0.21031566 | 2.855511 | 4.30E-03 | 0.23620006 | uncharacterize | COG4354 | GH116 |
| 3300020985.fa | Ga0206813_10 | 122.909326 | -0.3204805 | 0.1494048 | -2.145048 | 3.19E-02 | 0.53713806 | UDP:flavonoid | COG1819 | GT1 |
| 3300021013.fa | Ga0206831_10 | 255.637934 | 0.8155364 | 0.25482749 | 3.200347 | 1.37E-03 | 0.16675292 | tetratricopepti | COG0457 | GT105 |
| 3300020924.fa | Ga0207015_10 | 60.386389 | -0.3284873 | 0.16435944 | -1.998591 | 4.57E-02 | 0.54727336 | lipid-A-disacch | KO:K00748 | GT19 |
| 3300021003.fa | Ga0206966_10 | 111.791993 | -0.3257924 | 0.16600245 | -1.962576 | 4.97E-02 | 0.54740674 | cellulose synth | COG1215 | GT2 |
| 3300021013.fa | Ga0206831_10 | 83.170031 | -0.578617 | 0.29271017 | -1.976758 | 4.81E-02 | 0.59190773 | glycosyltransfe | COG0463,COG | GT2 |
| 3300020985.fa | Ga0206813_10 | 64.738344 | 0.3533913 | 0.16826714 | 2.10018 | 3.57E-02 | 0.56149403 | hypothetical pr | KO:K00754 | GT2 |
| 3300021040.fa | Ga0206819_10 | 74.689851 | -0.632607 | 0.31969744 | -1.978768 | 4.78E-02 | 0.63412295 | glycosyltransfe | COG0463 | GT2 |
| 3300020743.fa | Ga0207044_10 | 159.830915 | -0.4679969 | 0.21343057 | -2.192736 | 2.83E-02 | 0.51312703 | glycosyltransfe | COG0463 | GT2 |
| 3300020743.fa | Ga0207044_10 | 92.778947 | 0.4327021 | 0.18790056 | 2.302825 | 2.13E-02 | 0.47072062 | glycosyltransfe | COG0463 | GT2 |
| 3300021003.fa | Ga0206966_10 | 200.789649 | 0.9418018 | 0.2843842 | 3.311723 | 9.27E-04 | 0.09447469 | glycosyltransfe | COG0463 | GT2 |
| 3300021013.fa | Ga0206831_11 | 43.750374 | -0.4501686 | 0.18884127 | -2.383846 | 1.71E-02 | 0.42506601 | hypothetical pr | KO:K07011 | GT2 |
| 3300020651.fa | Ga0206911_10 | 67.041426 | -0.6986842 | 0.32466853 | -2.151992 | 3.14E-02 | 0.65828253 | glycosyltransfe | COG0463 | GT2 |
| 3300020924.fa | Ga0207015_10 | 118.518667 | 0.3741534 | 0.1518556 | 2.463876 | 1.37E-02 | 0.3760884 | ceramide gluco | KO:K00720 | GT21 |
| 3300020916.fa | Ga0207021_10 | 66.230766 | 0.6383122 | 0.24320395 | 2.624596 | 8.68E-03 | 0.33256732 | ceramide gluco | KO:K00720 | GT21 |
| 3300020985.fa | Ga0206813_10 | 122.146631 | 0.5942013 | 0.2540388 | 2.339018 | 1.93E-02 | 0.43007095 | ceramide gluco | KO:K00720 | GT21 |
| 3300021040.fa | Ga0206819_10 | 94.752713 | -0.4625048 | 0.19200408 | -2.408828 | 1.60E-02 | 0.49312567 | 1,2-diacylglyce | KO:K03715 | GT28 |
| 3300020743.fa | Ga0207044_10 | 104.453844 | -0.3847814 | 0.13966601 | -2.755011 | 5.87E-03 | 0.27831498 | 1,2-diacylglyce | KO:K03715 | GT28 |
| 3300020651.fa | Ga0206911_10 | 410.066545 | -0.4890078 | 0.23066866 | -2.119958 | 3.40E-02 | 0.65828253 | starch phospho | KO:K00688 | GT35 |
| 3300021013.fa | Ga0206831_11 | 80.800555 | -0.4447977 | 0.18913753 | -2.351716 | 1.87E-02 | 0.43402005 | glycosyltransfe | COG0438 | GT4 |
| 3300021040.fa | Ga0206819_10 | 66.544311 | 0.6463159 | 0.20452388 | 3.1601 | 1.58E-03 | 0.1534817 | 1,2-diacylglyce | KO:K19002 | GT4 |
| 3300021040.fa | Ga0206819_10 | 90.168181 | -0.2856591 | 0.12504307 | -2.284486 | 2.23E-02 | 0.5234125 | glycosyltransfe | COG0438 | GT4 |
| 3300020924.fa | Ga0207015_10 | 56.746054 | -0.4405937 | 0.17383178 | -2.534598 | 1.13E-02 | 0.33073609 | glycosyltransfe | COG0438 | GT4 |
| 3300020916.fa | Ga0207021_10 | 248.173913 | 0.7393402 | 0.36746412 | 2.012006 | 4.42E-02 | 0.66122739 | glycosyltransfe | COG0438 | GT4 |
| 3300020904.fa | Ga0207088_10 | 136.695834 | -0.2436375 | 0.10906981 | -2.233776 | 2.55E-02 | 0.54459853 | glycosyltransfe | COG0438 | GT4 |
| 3300020916.fa | Ga0207021_12 | 43.862152 | -0.7946121 | 0.28322633 | -2.805573 | 5.02E-03 | 0.25081984 | glycosyltransfe | COG0438 | GT4 |
| 3300020985.fa | Ga0206813_10 | 208.035872 | 0.2784859 | 0.09780608 | 2.847327 | 4.41E-03 | 0.28888971 | glycosyltransfe | COG0438 | GT4 |
| 3300020743.fa | Ga0207044_10 | 90.711657 | 0.2967069 | 0.13651735 | 2.173401 | 2.98E-02 | 0.51312703 | starch synthase | KO:K00703 | GT5 |
| 3300021013.fa | Ga0206831_10 | 90.932271 | -0.3651039 | 0.15854641 | -2.30282 | 2.13E-02 | 0.47247286 | hypothetical pr | Hypo-rule appl | GT83 |
| 3300021003.fa | Ga0206966_10 | 806.965002 | 0.4242664 | 0.19355254 | 2.191996 | 2.84E-02 | 0.4579071 | 4-amino-4-deo | COG1807,COG | GT83 |
| 3300020924.fa | Ga0207015_10 | 177.748411 | 0.2257694 | 0.11202292 | 2.015386 | 4.39E-02 | 0.56314334 | 4-amino-4-deo | COG1807 | GT83 |
| 3300020924.fa | Ga0207015_10 | 87.976336 | 0.343895 | 0.14217395 | 2.418832 | 1.56E-02 | 0.40018774 | 4-amino-4-deo | COG1807 | GT83 |
| 3300021040.fa | Ga0206819_10 | 62.543945 | 0.7320442 | 0.18533029 | 3.949944 | 7.82E-05 | 0.03613388 | 4-amino-4-deo | COG1807 | GT83 |
| 3300020916.fa | Ga0207021_10 | 457.629401 | -0.6469806 | 0.25148693 | -2.572621 | 1.01E-02 | 0.34451915 | putative glucoa | pfam10091 | GT84 |
| 3300020924.fa | Ga0207015_10 | 125.314852 | -0.9942513 | 0.2708941 | -3.670258 | 2.42E-04 | 0.05220338 | heptosyltransfe | KO:K02841 | GT9 |
| 3300020780.fa | Ga0206956_10 | 55.079854 | -0.3948613 | 0.20025616 | -1.971781 | 4.86E-02 | 0.68622278 | NADPH:quinon | COG0604 | GH23 |
| 3300020780.fa | Ga0206956_14 | 35.631881 | 0.5336023 | 0.25785827 | 2.069363 | 3.85E-02 | 0.65939461 | glucosylcerami | KO:K01201 | GH30_3 |
| 3300020780.fa | Ga0206956_10 | 100.030401 | 0.2984699 | 0.14116419 | 2.114346 | 3.45E-02 | 0.65572728 | hypothetical pr | Hypo-rule appl | GH39 |
| 3300020780.fa | Ga0206956_10 | 170.242227 | 0.4150498 | 0.1725396 | 2.405534 | 1.61E-02 | 0.5196912 | hypothetical pr | Hypo-rule appl | GH39 |
| 3300020780.fa | Ga0206956_10 | 121.68225 | -0.2884993 | 0.13721705 | -2.102503 | 3.55E-02 | 0.65572728 | VCBS repeat pr | pfam13517,pfa | GH57 |
| 3300020780.fa | Ga0206956_11 | 80.362894 | 0.5277622 | 0.21299782 | 2.477782 | 1.32E-02 | 0.45962847 | peptidoglycan/ | COG1835 | GT2 |
| 3300020780.fa | Ga0206956_11 | 34.272406 | -0.6940023 | 0.26192352 | -2.649637 | 8.06E-03 | 0.38315297 | hypothetical pr | Hypo-rule appl | GT2 |

WORKS CITED


(1) *Understanding Soil Microbes and Nutrient Recycling*. https://ohioline.osu.edu/factsheet/SAG-16 (accessed 2022-07-30).

(2) *Soil Bacteria | NRCS Soils*. https://www.nrcs.usda.gov/wps/portal/nrcs/detailfull/soils/health/biology/?cid=nrcs142p2_053862 (accessed 2022-07-30).

(3) Gajda, A. M.; Czyż, E. A.; Dexter, A. R.; Furtak, K. M.; Grządziel, J.; Stanek-Tarkowska, J. Effects of Different Soil Management Practices on Soil Properties and Microbial Diversity. *Int. Agrophysics* **2018**, *32*, 81–91. https://doi.org/10.1515/intag-2016-0089.

(4) Craswell, E.; Lefroy, R. The Role and Function of Organic Matter in Tropical Soils. *Nutr. Cycl. Agroecosystems* **2001**, *61*, 7–18. https://doi.org/10.1023/A:1013656024633.

(5) Scharlemann, J. P.; Tanner, E. V.; Hiederer, R.; Kapos, V. Global Soil Carbon: Understanding and Managing the Largest Terrestrial Carbon Pool. *Carbon Manag.* **2014**, *5* (1), 81–91. https://doi.org/10.4155/cmt.13.77.

(6) Ondrasek, G.; Bakić Begić, H.; Zovko, M.; Filipović, L.; Meriño-Gergichevich, C.; Savić, R.; Rengel, Z. Biogeochemistry of Soil Organic Matter in Agroecosystems & Environmental Implications. *Sci. Total Environ.* **2019**, *658*, 1559–1573. https://doi.org/10.1016/j.scitotenv.2018.12.243.

(7) *Global Warming vs. Climate Change | Facts – Climate Change: Vital Signs of the Planet*. https://climate.nasa.gov/global-warming-vs-climate-change/ (accessed 2022-07-30).

(8) *FAQ: What is the greenhouse effect?*. Climate Change: Vital Signs of the Planet. https://climate.nasa.gov/faq/19/what-is-the-greenhouse-effect (accessed 2022-07-30).

(9) Alteio, L. V.; Schulz, F.; Seshadri, R.; Varghese, N.; Rodriguez-Reillo, W.; Ryan, E.; Goudeau, D.; Eichorst, S. A.; Malmstrom, R. R.; Bowers, R. M.; Katz, L. A.; Blanchard, J. L.; Woyke, T. Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *mSystems* **2020**, *5* (2). https://doi.org/10.1128/mSystems.00768-19.

(10) Melillo, J. M.; Butler, S.; Johnson, J.; Mohan, J.; Steudler, P.; Lux, H.; Burrows, E.; Bowles, F.; Smith, R.; Scott, L.; Vario, C.; Hill, T.; Burton, A.; Zhou, Y.-M.; Tang, J. Soil Warming, Carbon–Nitrogen Interactions, and Forest Carbon Budgets. *Proc. Natl. Acad. Sci.* **2011**, *108* (23), 9508–9512. https://doi.org/10.1073/pnas.1018189108.

(11) Fierer, N.; Colman, B. P.; Schimel, J. P.; Jackson, R. B. Predicting the Temperature Dependence of Microbial Respiration in Soil: A Continental-Scale Analysis. *Glob. Biogeochem. Cycles* **2006**, *20* (3). https://doi.org/10.1029/2005GB002644.

(12) *Harvard Forest*. https://harvardforest.fas.harvard.edu/ (accessed 2022-07-30).

(13) Pregitzer, K. S.; Zak, D. R.; Loya, W. M.; Karberg, N. J.; King, J. S.; Burton, A. J. CHAPTER 7 - The Contribution of Root – Rhizosphere Interactions to Biogeochemical Cycles in a Changing World. In *The Rhizosphere*; Cardon, Z. G., Whitbeck, J. L., Eds.; Academic Press: Burlington, 2007; pp 155–178. https://doi.org/10.1016/B978-012088775-0/50009-4.

(14) Hugenholtz, P.; Tyson, G. W. Metagenomics. *Nature* **2008**, *455* (7212), 481–483. https://doi.org/10.1038/455481a.

(15) *DNeasy PowerSoil Pro Kits*. https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/microbial-dna/dneasy-powersoil-pro-kit/ (accessed 2022-07-30).

(16) SPAdes – Center for Algorithmic Biotechnology.

(17) Kang, D. D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, *7*, e7359. https://doi.org/10.7717/peerj.7359.

(18) Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Res.* **2015**, *25* (7), 1043–1055. https://doi.org/10.1101/gr.186072.114.

(19) Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinforma. Oxf. Engl.* **2009**, *25* (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

(20) Dedysh, S. N.; Kulichevskaya, I. S.; Huber, K. J.; Overmann, J. Defining the Taxonomic Status of Described Subdivision 3 Acidobacteria: Proposal of Bryobacteraceae Fam. Nov. *Int. J. Syst. Evol. Microbiol.* **2017**, *67* (2), 498–501. https://doi.org/10.1099/ijsem.0.001687.

(21) de Chaves, M. G.; Silva, G. G. Z.; Rossetto, R.; Edwards, R. A.; Tsai, S. M.; Navarrete, A. A. Acidobacteria Subgroups and Their Metabolic Potential for Carbon Degradation in Sugarcane Soil Amended With Vinasse and Nitrogen Fertilizers. *Front. Microbiol.* **2019**, *10*.

(22) Kielak, A. M.; Barreto, C. C.; Kowalchuk, G. A.; van Veen, J. A.; Kuramae, E. E. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Front. Microbiol.* **2016**, *7*. https://doi.org/10.3389/fmicb.2016.00744.

(23) Kielak, A. M.; Cipriano, M. A. P.; Kuramae, E. E. Acidobacteria Strains from Subdivision 1 Act as Plant Growth-Promoting Bacteria. *Arch. Microbiol.* **2016**, *198* (10), 987–993. https://doi.org/10.1007/s00203-016-1260-2.

(24) 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.; Barabote, R. D.; Bradley, B.; Brettin, T. S.; Brinkac, L. M.; Bruce, D.; Creasy, T.; Daugherty, S. C.; Davidsen, T. M.; DeBoy, R. T.; Detter, J. C.; Dodson, R. J.; Durkin, A. S.; Ganapathy, A.; Gwinn-Giglio, M.; Han, C. S.; Khouri, H.; Kiss, H.; Kothari, S. P.; Madupu, R.; Nelson, K. E.; Nelson, W. C.; Paulsen, I.; Penn, K.; Ren, Q.; Rosovitz, M. J.; Selengut, J. D.; Shrivastava, S.; Sullivan, S. A.; Tapia, R.; Thompson, L. S.; Watkins, K. L.; Yang, Q.; Yu, C.; Zafar, N.; Zhou, L.; Kuske, C. R. Three Genomes from the Phylum Acidobacteria Provide Insight into the Lifestyles of These Microorganisms in Soils. *Appl. Environ. Microbiol.* **2009**, *75* (7), 2046–2056. https://doi.org/10.1128/AEM.02294-08.

(25) Eichorst, S. A.; Trojan, D.; Roux, S.; Herbold, C.; Rattei, T.; Woebken, D. Genomic Insights into the Acidobacteria Reveal Strategies for Their Success in Terrestrial Environments. *Environ. Microbiol.* **2018**, *20* (3), 1041–1063. https://doi.org/10.1111/1462-2920.14043.

(26) Hansen, L.; Husein, D. M.; Gericke, B.; Hansen, T.; Pedersen, O.; Tambe, M. A.; Freeze, H. H.; Naim, H. Y.; Henrissat, B.; Wandall, H. H.; Clausen, H.; Bennett, E. P. A Mutation Map for Human Glycoside Hydrolase Genes. *Glycobiology* **2020**, *30* (8), 500–515. https://doi.org/10.1093/glycob/cwaa010.

(27) *JGI IMG Integrated Microbial Genomes & Microbiomes*. https://img.jgi.doe.gov/ (accessed 2022-07-30).

(28) *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 2022-07-30).

(29) *USADELLAB.org - Trimmomatic: A flexible read trimming tool for Illumina NGS data*. http://www.usadellab.org/cms/?page=trimmomatic (accessed 2022-07-30).

(30) Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinforma. Oxf. Engl.* **2013**, *29* (1), 15–21. https://doi.org/10.1093/bioinformatics/bts635.

(31) Love, M. I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15* (12), 550. https://doi.org/10.1186/s13059-014-0550-8.

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

(33) *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database | Bioinformatics | Oxford Academic*. https://academic.oup.com/bioinformatics/article/36/6/1925/5626182 (accessed 2022-07-30).

(34) Letunic, I.; Bork, P. Interactive Tree Of Life (ITOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation. *Nucleic Acids Res.* **2021**, *49* (W1), W293–W296. https://doi.org/10.1093/nar/gkab301.

(35) *RStudio | Open source & professional software for data science teams*. https://www.rstudio.com/ (accessed 2022-07-30).

(36) *Unity Cluster Documentation*. https://docs.unity.rc.umass.edu/ (accessed 2022-07-30).

(37) Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics. *Nucleic Acids Res.* **2009**, *37* (Database issue), D233–D238. https://doi.org/10.1093/nar/gkn663.

(38) *Introducing dplyr*. https://www.rstudio.com/blog/introducing-dplyr/ (accessed 2022-07-30).

(39) Soneson, C.; Love, M. I.; Robinson, M. D. Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences. *F1000Research* **2015**, *4*, 1521. https://doi.org/10.12688/f1000research.7563.1.

(40) Trapnell, C.; Hendrickson, D. G.; Sauvageau, M.; Goff, L.; Rinn, J. L.; Pachter, L. Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq. *Nat. Biotechnol.* **2013**, *31* (1), 46–53. https://doi.org/10.1038/nbt.2450.

(41) Zhang, H.; Yohe, T.; Huang, L.; Entwistle, S.; Wu, P.; Yang, Z.; Busk, P. K.; Xu, Y.; Yin, Y. DbCAN2: A Meta Server for Automated Carbohydrate-Active Enzyme Annotation. *Nucleic Acids Res.* **2018**, *46* (W1), W95–W101. https://doi.org/10.1093/nar/gky418.

(42) *Example — GTDB-Tk 2.1.1 documentation*. https://ecogenomics.github.io/GTDBTk/examples/classify_wf.html (accessed 2022-07-30).

(43) Parks, D. H.; Chuvochina, M.; Rinke, C.; Mussig, A. J.; Chaumeil, P.-A.; Hugenholtz, P. GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy. *Nucleic Acids Res.* **2022**, *50* (D1), D785–D794. https://doi.org/10.1093/nar/gkab776.

(44) *Phylogenetic Trees and Monophyletic Groups | Learn Science at Scitable*. http://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956 (accessed 2022-07-30).

(45) Deschamps-Francoeur, G.; Simoneau, J.; Scott, M. S. Handling Multi-Mapped Reads in RNA-Seq. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1569–1576. https://doi.org/10.1016/j.csbj.2020.06.014.46.Burgess, R.R. Sigma Factors. {\i{}Refernce Module in Life Sciences}.1831 (2001).

(47) *CAZy - GT83*. http://www.cazy.org/GT83.html (accessed 2022-07-30).

(48) PubChem. *Undecaprenyl phosphate-alpha-4-amino-4-deoxy-L-arabinose arabinosyl transferase (Salmonella enterica subsp. enterica serovar Typhimurium str. LT2)*. https://pubchem.ncbi.nlm.nih.gov/protein/O52327 (accessed 2022-07-30).

(49) Liu, Y.-Y.; Hill, R. A.; Li, Y.-T. Chapter Three - Ceramide Glycosylation Catalyzed by Glucosylceramide Synthase and Cancer Drug Resistance. In *Advances in Cancer Research*; Norris, J. S., Ed.; The Role of Sphingolipids in Cancer Development and Therapy; Academic Press, 2013; Vol. 117, pp 59–89. https://doi.org/10.1016/B978-0-12-394274-6.00003-0.

(50) *CAZy - GT2*. http://www.cazy.org/GT2.html (accessed 2022-07-30).

(51) *CAZy - GT4*. http://www.cazy.org/GT4.html (accessed 2022-07-30).

(52) Tomasik, P.; Horton, D. Chapter 2 - Enzymatic Conversions of Starch. In *Advances in Carbohydrate Chemistry and Biochemistry*; Horton, D., Ed.; Academic Press, 2012; Vol. 68, pp 59–436. https://doi.org/10.1016/B978-0-12-396523-3.00001-4.

(53) Dasgupta, A. Chapter 7 - β-Hexosaminidase, Acetaldehyde–Protein Adducts, and Dolichol as Alcohol Biomarkers. In *Alcohol and its Biomarkers*; Dasgupta, A., Ed.; Clinical Aspects and Laboratory Determination; Elsevier: San Diego, 2015; pp 163–180. https://doi.org/10.1016/B978-0-12-800339-8.00007-9.

(54) Stock, J. B.; Baker, M. D. Chemotaxis. In *Encyclopedia of Microbiology (Third Edition)*; Schaechter, M., Ed.; Academic Press: Oxford, 2009; pp 71–78. https://doi.org/10.1016/B978-012373944-5.00068-7.

(55) Shakya, M.; Lo, C.-C.; Chain, P. S. G. Advances and Challenges in Metatranscriptomic Analysis. *Front. Genet.* **2019**, *10*.

(56) Raynaud, X.; Nunan, N. Spatial Ecology of Bacteria at the Microscale in Soil. *PLoS ONE* **2014**, *9* (1), e87217. https://doi.org/10.1371/journal.pone.0087217.

(57) Earle, S.; Earle, S.; Earle, S. 5.4 Weathering and the Formation of Soil. **2015**.

(58) Jafari, M.; Ansari-Pour, N. Why, When and How to Adjust Your P Values? *Cell J. Yakhteh* **2019**, *20* (4), 604–607. https://doi.org/10.22074/cellj.2019.5992.

(59) Typas, A.; Hengge, R. Differential Ability of Sigma(s) and Sigma70 of Escherichia Coli to Utilize Promoters Containing Half or Full UP-Element Sites. *Mol. Microbiol.* **2005**, *55* (1), 250–260. https://doi.org/10.1111/j.1365-2958.2004.04382.x.

(60) Grossman, A. D.; Straus, D. B.; Walter, W. A.; Gross, C. A. Sigma 32 Synthesis Can Regulate the Synthesis of Heat Shock Proteins in Escherichia Coli. *Genes Dev.* **1987**, *1* (2), 179–184. https://doi.org/10.1101/gad.1.2.179.

(61) Kourennaia, O. V.; Tsujikawa, L.; Dehaseth, P. L. Mutational Analysis of Escherichia Coli Heat Shock Transcription Factor Sigma 32 Reveals Similarities with Sigma 70 in Recognition of the -35 Promoter Element and Differences in Promoter DNA Melting and -10 Recognition. *J. Bacteriol.* **2005**, *187* (19), 6762–6769. https://doi.org/10.1128/JB.187.19.6762-6769.2005.

(62) Henkel, M.; Hausmann, R. Chapter 2 - Diversity and Classification of Microbial Surfactants. In *Biobased Surfactants (Second Edition)*; Hayes, D. G., Solaiman, D. K. Y., Ashby, R. D., Eds.; AOCS Press, 2019; pp 41–63. https://doi.org/10.1016/B978-0-12-812705-6.00002-2.

(63) *UniProt*. https://www.uniprot.org/uniprotkb/I0HPW1/entry (accessed 2022-07-30).

(64) Blair, D. F.; Kim, D. Y.; Berg, H. C. Mutant MotB Proteins in Escherichia Coli. *J. Bacteriol.* **1991**, *173* (13), 4049–4055.

(65) *Chemotaxis signaling protein CheY binds to the rotor protein FliN to control the direction of flagellar rotation in Escherichia coli | PNAS*. https://www.pnas.org/doi/10.1073/pnas.1000935107 (accessed 2022-07-30).