NEED TO CHANGE TITLE

Kory Melton

December 7th, 2019

# Introduction

Microscopic organisms are among the oldest and most prevalent lifeforms on earth. Accordingly, they play a vital role in the health of ecosystems by affecting processes like carbon dioxide and nitrogen cycling (Martiny *et al.*, 2006). As the worlds climate changes over time understanding how microbes influence the dynamics of ecosystems will be necessary to preserve the environment. Currently, the impacts of climate change on the biogeochemical cycling of microbes are not well understood (Barnard *et al.*, 2005). However, some research suggests microbes may be able to mediate carbon-cycle feedbacks to climate warming (Zhou *et al.*, 2012).

Due to their importance in the function and stability of ecosystems, the biodiversity of microorganisms is vital to the overall diversity of macroscopic organisms. Typically the first objective of many ecological research projects is to determine the which organisms are present and their bio-geography (Zimmerman *et al.*, 2014). Previous research suggests the distribution of microbes is heterogeneous and spatially-structured (Maron *et al.*, 2011). Thus, accurate sampling of a wide variety of plots is necessary to understanding how microbes influence a large ecosystem.

For the microbial community, identifying distinct species is particularly difficult (Koeppel *et al.*, 2008). Recent advances in molecular biology, including high-throughput sequencing has allowed for genetic identification of microbial species without culturing them. Many researchers use amplicon sequencing of the 16S rRNA gene as a bio-marker for the microbial community (Tringe and Hugenholtz, 2008).

The National Science Foundation (NSF) funded National Ecological Observation Network (NEON) was designed to answer how the ecosystem and it’s components respond to changes in climate? (Keller *et al.*, 2008) To this aim NEON collects microbial samples in soil from a number of sites across the United States. I will investigate the UNDE site, near the Great Lakes in Michigan. The UNDE site is managed by the University of Notre Dame Environmental Research Center (UNDERC). The property has a second-growth Northern mesic forest and Evergreen forests. The dominant species are red and sugar maple (*Acer rubrum* and *A. saccharum*), aspen (*Populus tremuloides*), paper birch (*Betula papyrifera*), balsam fir (*Abies balsamea*), cedar (*Thuja occidentalis*), and black spruce (*Picea mariana*). In addition to the deciduous and mixed forests the site has woody wetlands dominated by thicks of alder (*Alnus incana*).

The climate is generally humid, cool, and wet with no true dry season and receives approximately 114 inches of snow annually. The soil families include coarse, loamy, mixed, superactive, frigid, argic fragiaquods.

Before the land was donated to the Notre Dame University in the 1930s, region-wide logging for pine cut most of the forested areas on the property. Timber harvest continued into the 1950s, but was followed by a forest regrowth. Recently, the site is used for recreational, educational, and research goals.

The objective of this study is to determine which microbes are present at the UNDE site, using the 16S barcode sequence, and how this changes over time. This information provides insight into how these ecosystems respond to climate change and their overall health and function.

The UNDE site has 1192 samples containing sequencing data from soil microbes. The data collected by NEON will be processed for quality control, sub-sampled, trimmed, converted to fasta format and compared to NCBI’s nucleotide database using the BLAST algorithm. The matches to the nucleotide database will be used to make inferences on the type of microbial organisms present.

The data set contained a large number of high quality sequences. I found *Ralstonia solanacearum* was the most frequent BLAST match with a few matches with a query cover of 100%. The finding of *Ralstonia solanacearum* is important because this could indicate a soil-borne plant pathogen on the UNDE site.

# Methods

NEON collected and sequenced the samples from the UNDE site, all data can be found in data product DP1.10108.001 (<https://data.neonscience.org/data-products/DP1.10108.001>) (NEON). I analyzed the samples which were formatted and uploaded by NEON to make inferences on the microbial sequences present in each sample over time.

## Sample Collection

Each site contains ten plots that are further sampled one to three times per year. NEON randomly selects three locations for each sample event, with each sampling location is not re-sampled. Soil sampling is collected to a maximum depth of thirty centimeters. After collection, samples are stores in sterile containers, frozen and shipped to a lab for downstream analysis. In the lab, DNA is extracted and the samples are prepared for high-throughput sequence analysis using primers for 16S sequences. The sequences are delivered to NEON for quality control and acceptance. After acceptance, NEON formats and uploads the files in repositories for public use.

## Analysis

I downloaded all unique 16S files for the UNDE site from the NEON database. Then, I checked each files quality using fastqc. I sub-sampled each file to make the dataset computationally tractable. I trimmed each file in order to capture high quality reads and converted the fastq files to fasta files. Next, I compared each fasta file to a nucleotide database to make inferences on which microorganisms were present in the samples from the UNDE site.

Finally, I used those BLAST results to BLANK!!! FILL THIS IN!!!!

### Downloading Data

I downloaded all unique files for the target gene 16S from the NEON database for the UNDE site. The script to download the data used the R library neonUtilities and the NEON API to download the data product. I used the data product ID DP1.10108.001. In addition, I downloaded the metadata for the raw data to include information on when the data was collected and where the data was collected from.

### Fastqc

I extracted the R1 files from their zipped folders. Then, I ran fastqc on each file to assess their quality and identify any problematic files (Andrews, 2010).

### Sub-sampling

I randomly sub-sampled 0.5% of the sequences in each file to make the data computationally tractable. For this, I used the srand function and the random seed 1234.

### Trimming

I trimmed the sub-sampled files using TrimmomaticSE (Bolger *et al.*, 2014). I used four threads, phred33 base quality encoding, a minimum quality threshold of 5 for both leading and trailing bases, and a minimum length of two-hundred base pairs. Additionally, I implemented a sliding window with window size eight and required quality of twenty-five.

### BLAST

I used bioawk (<https://github.com/lh3/bioawk>) to convert the sub-sampled, trimmed fastq files to fasta files for downstream BLAST analyses. I compared each fasta file to the GenBank’s nucleotide database (Benson *et al.*, 2000). I used the BLASTN algorithm, which uses a nucleotide query sequence to search the nucleotide database (Camacho *et al.*, 2009). I used four threads, the ‘10 sscinames std’ format, and returned one match for each sequence. I excluded 2018-09-19\_environmental\_sequence.gi to reduce the number of uncultured or environmental matches. The BLAST results are stored in a comma seperated file in the output directory.

### Analyzing Blast Data

I analyzed the BLAST matches found for each sequence using the seventh bash script in the pipeline. First, I curated the BLAST matches to exclude classifications that were unidentified or too generic. Then, I found the five BLAST matches which occured most frequently throughout the dataset. I analyzed the BLAST matches for the five most prevalent to determine my confidence in the results.

#### Curating

I curated the BLAST matches based on the description returned from the BLAST result. I removed all matches with the description uncultured or unidentified since these results would not provide any context on the UNDE site. In addition, I removed all files with the generic names of bacterium, Bacterium, or fungal.

#### Processing Curated File

After curation, I calculated a count for the frequency of each description. Then, I sorted out the top twenty most prevalent matches throughout all files. The count and description are saved in a curated summary comma separated file in the output directory. I also saved an output of the top thirty most prevalent BLAST matches without curation in the output directory.

I saved the data from the BLAST results for the top five most prevalent matches into a comma separated file for further analysis in R. I pulled the ten most prevalent matches based on the highest taxonomic level in the description. This was compared to the top ten genus by abundance from the curated database. Next, I found the most abundant species for each genus.

I used the percent identity, length of sequence match, expected value, and bitscore to assess the quality of the BLAST matches for the top five most prevalent matches by highest taxonomic level. The data were visualized with boxplots using ggplot2 (Wickham, 2016). I removed low quality matches based on cutoff values for percent identity, length of sequence match, expected value, and bitscore. I used a cutoff of 90% for percent identity. I sorted out all sequences with a match length below two-hundred base pairs. I used a upper bound of 0.00010 for expected value and two-hundred for bitscore.

### DADA2 Pipeline

I also processed the subsampled data with dada (Callahan *et al.*, 2016). The sequences were filtered and trimmed to ensure all analyzed sequences were high quality. I discarded any sequences with N’s. Only up to three expected errors were allowed. Reads were truncated at the first instance of a quality score below two. I also discarded reads which matched against the phiX genome. Error models were generated for each of the samples. I removed all duplicated sequences. I generated high resolution sample inferences from the dereplicated forward amplicon sequencing reads and the error models previously generated. I removed all chimeras using the consensus method. All sequences shorter than fifty were removed because there is no way to assign their taxonomy. I assigned taxonomy using a supplied training data set while checking the reverse complement of all sequences to see if it is a better match to the reference sequence. The results were saved to the output directory.

### Metadata

I downloaded the metadata for the UNDE site from the DP1.10108.001 data product by NEON. All duplicated rows were removed and matched to the data from the dada2 pipeline. I pruned the samples without any sequences.

### Analysis

I analyzed the data using ggplot (Wickham, 2016) and dplyr (Wickham *et al.*, 2019). First, I calculated the number of samples for each month, year, kingdom, and plot. Then, I found the top phyla in each plot by abundance. I separated the top three phyla by year to analyze the change of the samples from 2016 to 2017. For this analysis I only included the months of May and August since they were the only two months with collections in both 2016 and 2017. I analyzed the the abundance of Phyla by month and year to determine if the composition of samples changed over time. Then, I took the top six genus for each month and year separated by Phyla. Finally, I used (McMurdie and Holmes, 2013) to visualize the alpha diversity by month and year using the shannon method.

# Results

## Sequence Counts and Quality

The data set included 1192 fastq files with a range of sequences per file. The maximum number of sequences in a file was 413,851 sequences, while the minimum number of sequences was 118. The average number of sequences in a file was 43,028 sequences. There was a total of 51,289,602 sequences in the data set.

After sub-sampling there was 268,512 sequences total, or approximately .5% of the sequences. The maximum number of sequences after sub-sampling was 2,090, while the minimum was zero. Seven of the files did not have any sequences after sub-sampling. This was about 0.6% of the raw fastq files.

After trimming there was 154,922 sequences total, or approximately 58% of the sub-sampled sequences. The maximum number of sequences after trimming was 1,213, while the minimum was zero. Nineteen of the files did not have any sequences after sub-sampling. This was about 1.6% of the raw fastq files. files.

Overall, the sequence quality was good. None of the inspected fastqc reports contained sequences flagged as poor quality. Figure 1 shows the fastqc of a file which did not have any sequences after trimming, while Figure 2 shows the fastqc of the file with the most sequences after trimming. The decline in quality scores happens about 50 base pairs later in the better file.

## BLAST Results

The four most frequent BLAST matches I found in the data set were *Ralstonia solanacearum*, *Staphylococcus aureus*, *Geobacillus thermocatenulatus*, and *Sphingomonas melonis*.

The top matches I found in each file with a query cover of 100% were *Ralstonia solanacearum*, *Bradyrhizobium elkanii*, *Bacillus sp.*, *Actinomadura sp*, and *Acetobacteraceae bacterium*.

The ten most frequent BLAST matches by the highest taxonomic level can be seen in Table 1. They are *Ralstonia*, *Staphylococcus*, *Bradyrhizobium*, *Pseudomonas*, Proteobacteria, *Candidatus*, *Mucilaginibacter*, *Bacillus*, *Sphingomonas*, and *Burkholderia*. *Ralstonia* dominated the BLAST matches with 68,508 total matches in the data set. There was a large dropoff to *Staphylococcus* and *Bradyrhizobium*.

The ten most frequent genus matches by abundance were *Bradyrhizobium*, *Pseudomons*, *Roseiarcus*, *Burkholderia*, *Mycobacterium*, *Thermosporothrix*, *Gaiella*, *Massilia*, *Pedomicrobium*, and *Gemmatimonas*. *Bradyrhizobium*, *Pseudomonas*, and *Burkholderia* were prevalent in both lists while *Bradyrhizobium* and *Pseudomonas* appeared at near the top of the NCBI search and the curated search.

The species returned for each genus in the BLAST results can be seen in Figure 3 and Table 3. For *Ralstonia* and *Staphylococcus* one species dominated all of the BLAST matches for the genus. *Ralstonia solanacearum* was present in 68,505 of the 68,508 of the *Ralstonia* matches, or 99.99% of the matches. *Staphylococcus aureus* was present in 3,059 of the 3,068 *Staphylococcus* matches, or 99.71% of the matches. The BLAST matches for *Bradyrhizobium* and *Pseudomonas* had a wider variety of species. Figure 3 shows the the top three species for these two genera. *Bradyrhizobium* was dominated by *Bradyrhizobium elkanii* and other unidentified species with a small number of *Bradyrhizboium erythroplei*. *Pseudomonas* had a lot fewer samples overall, but they were equally spread among *Pseudomonas frederiksbergensis*, *Pseudomonas protegens*, and unidentified species.

Figues 4 through 7 show the distribution of values for the BLAST matches by their description. In each graph there is a red line depicting the cutoff value used for the corresponding value. As shown in Figure 4, a large majoriy of the percent identities for the BLAST matches fall close to 100%. However, all matches except for proteobacteria have matches with lower than 90% query cover. Figure 5 highlights the difference of the length of sequences by the different types of BLAST matches. A majoriy of the matches for *Ralstonia* and *Staphylococcus* are fairly short sequences less than 100 base pairs long. While proteobacteria, *Bradyrhizobium*, and *Pseudomonas* had a majority of their matches above 200 base pairs long. In addition, as seen in Figure 6, *Ralstonia* and *Staphylococcus* have higher expected values with many above the cutoff of 0.00010. *Bradyrhizobium*, proteobacteria, and *Pseudomonas* have low expected values which indicate strong confidence in the BLAST matches. A similar trend can be seen with the bitscore values returned from the BLAST matches. In figure 7, you can see *Ralstonia* and *Staphyloccocus* have low bitscores centering around 100 base pairs while *Bradyrhizobium*, proteobacteria, and *Pseudomonas* had bitscores centering around 500. Table 4 shows the BLAST results before and after curation to throw out lower quality matches. The counts for *Ralstonia* and *Staphyloccocus* drop drastically with curation while a majority of the matches to *Bradyrhizobium*, proteobacteria, and *Pseudomonas* remain after curation. With all of these metrics considered, I have high confidence of the presence of *Bradyrhizobium*, *Pseudomonas*, and proteobacteria in the samples from the UNDE site. In contrast, I am not confident there was actually *Ralstonia* and *Staphylococcus* present at the UNDE site.

## Samples over Time

The number of samples over varied quite a lot between the time of collection and the plot collected from. Figure 8 shows the variation of the number of samples collected by month and year. Only May and July had samples collected in both 2016 and 2017. In both of these months there was significantly more samples collected in 2016 than 2017. In addition, the most samples were collected in May; followed by August, October, and finally July. Figure 9 shows the samples by plot in 2016 and 2017 during the months of May and August.

# Discussion

## Overview

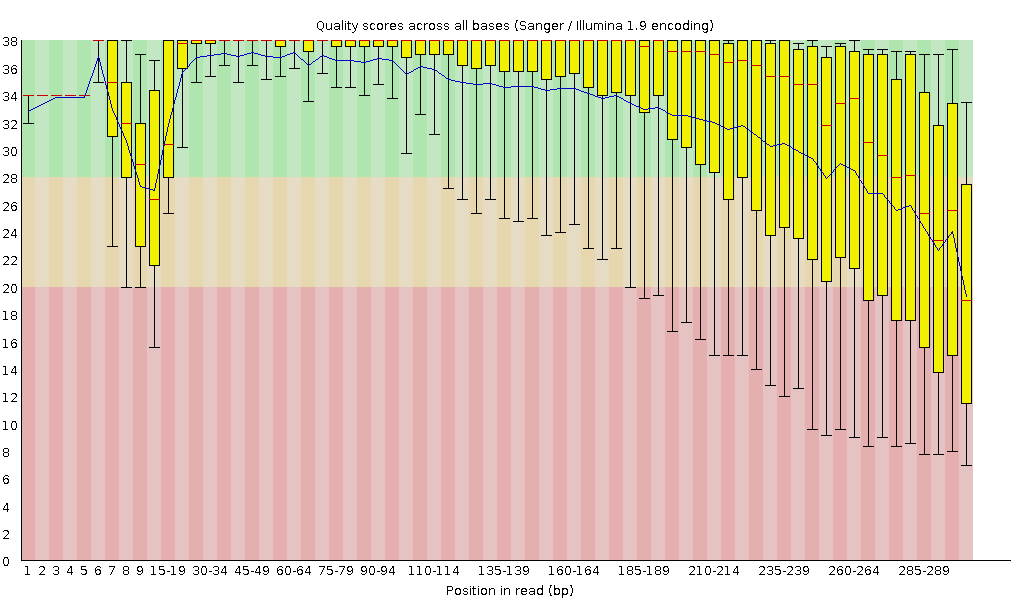
The UNDE site, in Michigan, near the great lakes has a second-growth Northern mesic forest and Evergreen forests with various trees including maple, aspen, birch, cedar, and black spruce. In addition, the site has a woody wetlands dominated by thicks of alder.

The data collected by NEON from this site was high-quality and contained a substantial amount of genetic information. The collected sequences matched a variety of bacteria widely associated to soil including *Ralstonia solanacearum* and *Acetobacteraceae bacterium*. These two species have very different impacts on plants. *Ralstonia solanacearum* is a plant pathogen while *Acetobacteraceae bacterium* can be found in symbiotic relationships with many different plants by colonizing their internal tissues.

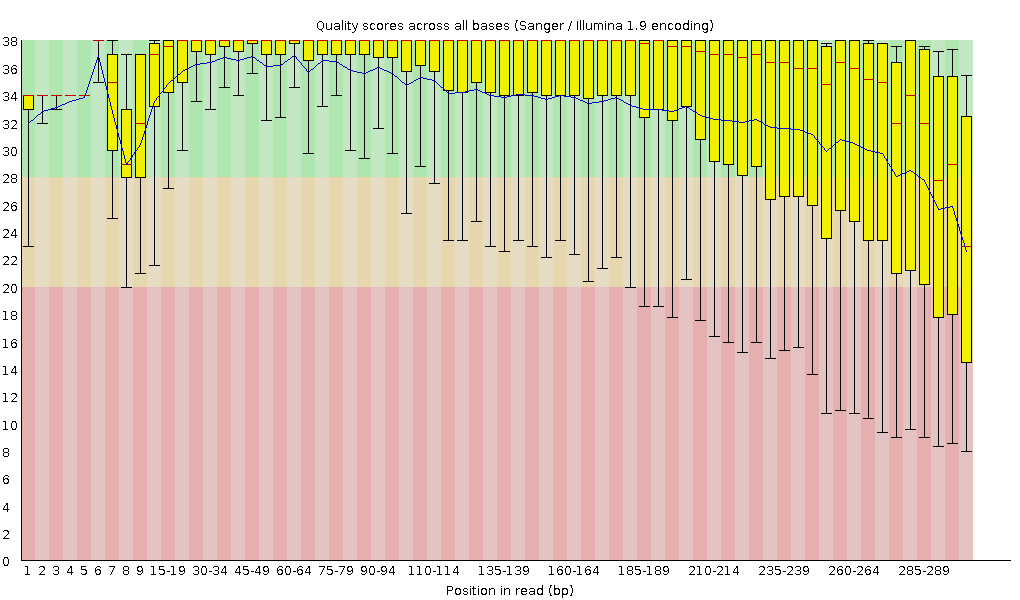
## *Ralstonia solanacearum*

*Ralstonia solanacearumm* is a soil-borne plant pathogen with unusually high prevalence worldwide (Salanoubat *et al.*, 2002). This pathogen naturally infects the roots of plants causing bacterial wilt in many plants and crops (Castillo and Greenberg, 2007). The DNA sequencing of the soil samples from the UNDE site indicate a high probability of *Ralstonia solanacearum*. The workers of UNDERC should look for indications of this pathogen in plants near the sampling sites and take preventative action to protect the biodiversity of the UNDE site.

# Tables and Graphs



**Figure 1.** Fastqc per base sequence quality for BMI\_Plate1WellA1\_16S\_R1. This was one of the lower quality files and did not have any acceptable sequences after sub-sampling and trimming. You can see a sharp decline on base sequence quality around 210 base pairs.



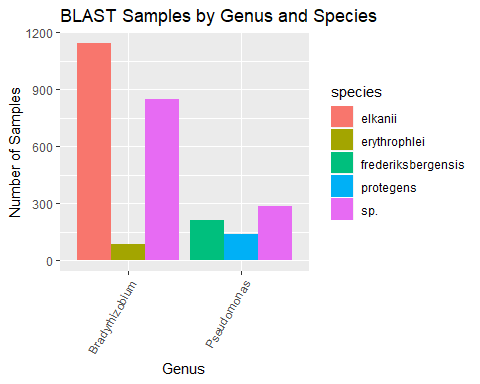
**Figure 2.** Fastqc per base sequence quality for BMI\_Tube52\_16S\_R1. This was one of the higher quality files. You can’t see a sharp decline until approximately 260 base pairs.

|  |  |
| --- | --- |
| Description | Count |
| Ralstonia | 68508 |
| Staphylococcus | 3068 |
| Bradyrhizobium | 2394 |
| Pseudomonas | 1095 |
| Proteobacteria | 897 |
| Candidatus | 731 |
| Mucilaginibacter | 691 |
| Bacillus | 605 |
| Sphingomonas | 565 |
| Burkholderia | 563 |

**Table 1.** Top 10 genus returned from blast results

|  |  |
| --- | --- |
| Genus | Abundance |
| Bradyrhizobium | 284 |
| Pseudomonas | 92 |
| Roseiarcus | 53 |
| Burkholderia | 28 |
| Mycobacterium | 18 |
| Thermosporothrix | 14 |
| Gaiella | 13 |
| Massilia | 9 |
| Pedomicrobium | 7 |
| Gemmatimonas | 6 |

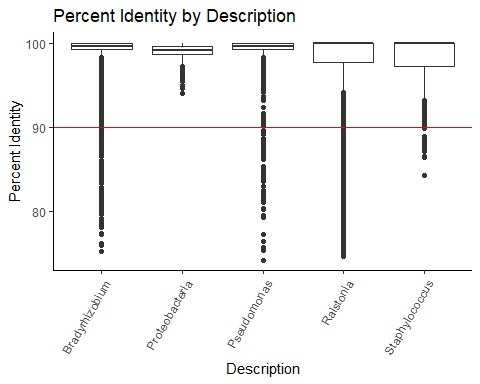
**Table 2.** Top 10 genus by abundance from curated database



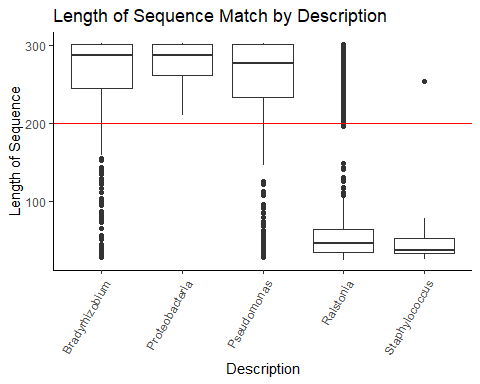
**Figure 3.**

|  |  |  |
| --- | --- | --- |
| genus | species | n |
| Ralstonia | solanacearum | 68505 |
| Staphylococcus | aureus | 3059 |
| Proteobacteria | Proteobacteria | 897 |

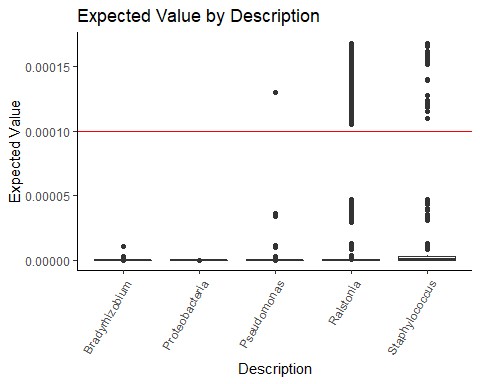
**Table 3.**



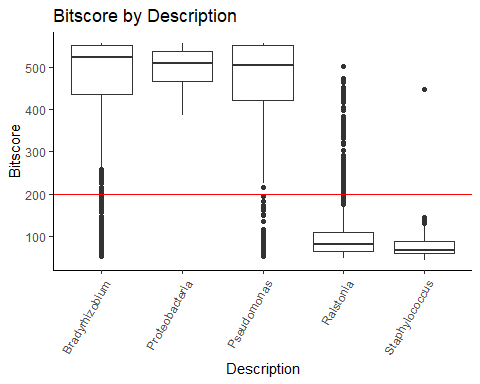
**Figure 4.**



**Figure 5.**



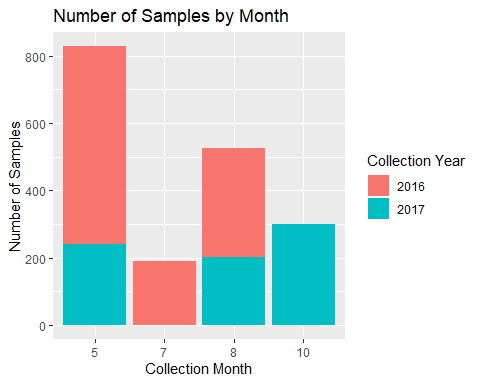
**Figure 6.**



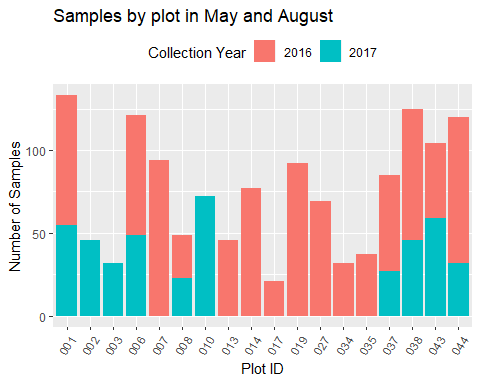
**Figure 7.**

|  |  |  |
| --- | --- | --- |
| Description | Count | Curated Count |
| Ralstonia | 68508 | 43 |
| Staphylococcus | 3068 | 1 |
| Bradyrhizobium | 2394 | 2224 |
| Pseudomonas | 1095 | 945 |
| Proteobacteria | 897 | 897 |

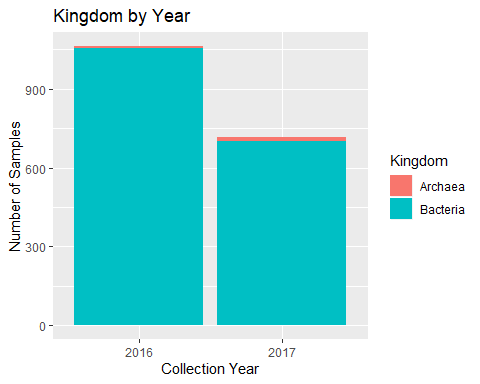
**Table 4.**



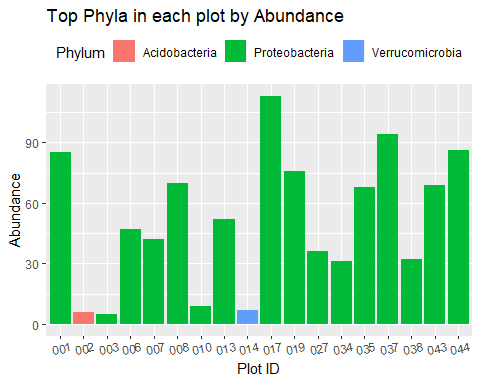
**Figure 8.** Number of samples by month



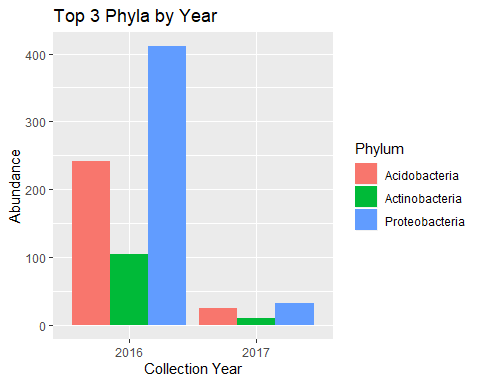
**Figure 9.** Number of samples by plot



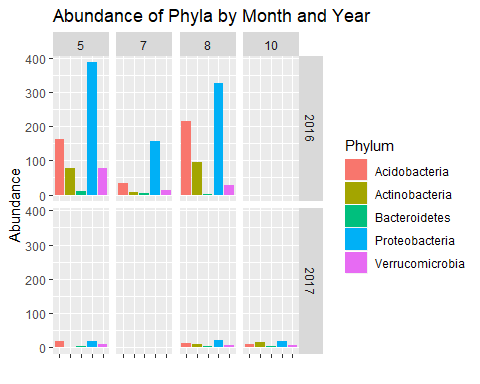
**Figure 10.** Summary of Kingdom found matched in phyloseq.



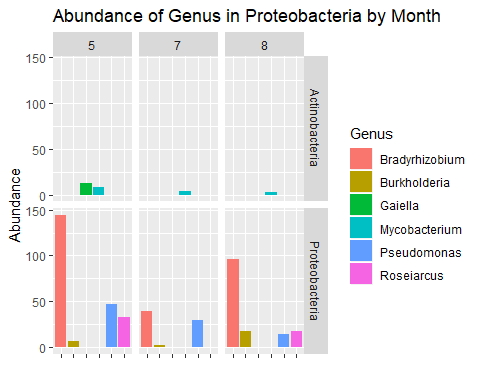
**Figure 11.**

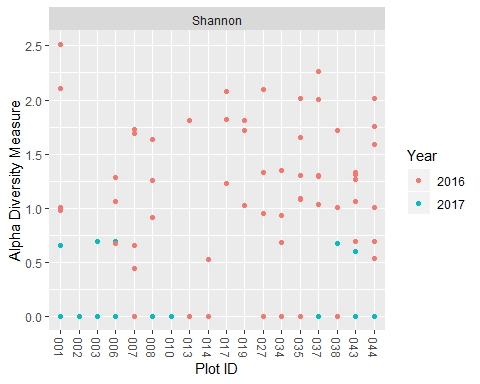


**Figure 12.**

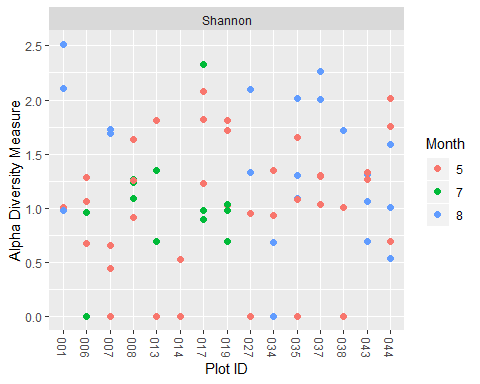


**Figure 13.**





**Figure 14.**



**Figure 15.**

# Sources Cited

Andrews,S. (2010) FastQC: A quality control tool for high throughput sequence data.

Barnard,R. *et al.* (2005) Global change, nitrification, and denitrification: A review. *Global biogeochemical cycles*, **19**.

Benson,D.A. *et al.* (2000) GenBank. *Nucleic acids research*, **28**, 15–18.

Bolger,A.M. *et al.* (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Callahan,B.J. *et al.* (2016) DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, **13**, 581–583.

Camacho,C. *et al.* (2009) BLAST+: Architecture and applications. *BMC bioinformatics*, **10**, 421.

Castillo,J.A. and Greenberg,J.T. (2007) Evolutionary dynamics of ralstonia solanacearum. *Appl. Environ. Microbiol.*, **73**, 1225–1238.

Keller,M. *et al.* (2008) A continental strategy for the national ecological observatory network. *Frontiers in Ecology and the Environment*, **6**, 282–284.

Koeppel,A. *et al.* (2008) Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences*, **105**, 2504–2509.

Maron,P.-A. *et al.* (2011) Soil microbial diversity: Methodological strategy, spatial overview and functional interest. *Comptes Rendus Biologies*, **334**, 403–411.

Martiny,J.B.H. *et al.* (2006) Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, **4**, 102.

McMurdie,P.J. and Holmes,S. (2013) Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, **8**, e61217.

NEON Data product dp1.10108.001, soil microbe marker gene sequences. *National Ecological Observatory Network, Battelle, Boulder, CO, USA NEON Downloaded on November 17, 2019*.

Salanoubat,M. *et al.* (2002) Genome sequence of the plant pathogen ralstonia solanacearum. *Nature*, **415**, 497.

Tringe,S.G. and Hugenholtz,P. (2008) A renaissance for the pioneering 16S rRNA gene. *Current opinion in microbiology*, **11**, 442–446.

Wickham,H. (2016) Ggplot2: Elegant graphics for data analysis Springer-Verlag New York.

Wickham,H. *et al.* (2019) Dplyr: A grammar of data manipulation.

Zhou,J. *et al.* (2012) Microbial mediation of carbon-cycle feedbacks to climate warming. *Nature Climate Change*, **2**, 106–110.

Zimmerman,N. *et al.* (2014) The unseen world: Environmental microbial sequencing and identification methods for ecologists. *Frontiers in Ecology and the Environment*, **12**, 224–231.