Presence of the proteobacteria *Bradyrhizobium*, *Pseudomonas*, and *Burkholderia* indicate a healthy microbiome community at the UNDE site which could be strongly influence by acidic Sphagnum Bogs

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).

## UNDE Site

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. In the Great Lakes area NEON is particularly interested in land use and forest management. A balanced and well functioning soil microbiome is vital to the health of the forests found in the UNDE site and the Greak Lakes in general.

The UNDE site is managed by the University of Notre Dame Environmental Research Center (UNDERC). The site is at an elevation of 518 m, with a mean annual temperature of 3, and a moderate 854 mms of annual precipitation. 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*).

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 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. Although, some of the soil on the site is poorly drained, which may cause acidic sphagnum bogs. This is especially troublesome since climate warming can promote the growth of peat mosses and compound this issue (Breeuwer *et al.*, 2009). The presence of sphagnum is known to increase the amount of dissolved organic Nitrogen; reducing Nitrogen immobilization by soil microbes (Bragazza *et al.*, 2013). The competition for nutrients and changes caused by sphagnum bogs and climate warming, can shortcircuit the Nitrogen cycle. In years where sphagnum bogs are prevalent one could expect lower abundances of soil microbes. In addition, ecosystems with a high abundance of nitrogen-fixing bacteria may be more suited to withstand the disruption caused by sphagnum bogs.

## Study Objectives

The primary 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 potential for sphagnum bog growth at the UNDE site is a primary concern. To that end, I looked for signs to indicate plots at the site which may have issues with sphagnum growth. A secondary objective of this study is to carefully analyze matches against the BLAST database. Many of my colleagues who have collected data from various NEON sites had *Ralstonia solanacearum* as their top BLAST match, and often had *Staphylococcus aureus* among the top of their list.

The UNDE site has 1192 samples containing sequencing data from soil microbes. The data collected by NEON was be processed for quality control, and sub-sampled. At this point the sub-sampled files will be processed through two pipelines: DADA2 and BLAST. For the BLAST pipeline the samples were trimmed, converted to fasta format and compared to NCBI’s nucleotide database using the BLAST algorithm. After comparison to the nucleotide database, I curated the data in R based on percent identity, sequence match length, expected value, and bitscore. For the DADA2 pipeline, I discarded poor sequences and trimmed the higher quality sequences. Then, I genereated sample inferences using error models and compared the results to a training data set. I used the matches from both the nucleotide database and training data set to make inferences on the type of microbial organisms present.

## Major Findings

The subsampling of the dataset collected by NEON for the UNDE site contained a large number of quality sequences which could be used for downstream analysis. I found a variety of proteobacteria in both the BLAST matches and matches to the curated training dataset. Most notably, the samples contained *Bradyrhizobium*, *Pseudomonas*, and *Burkholderia*. The presence of *Bradyrhizobium* and *Pseudomonas* provide indications of a healthy soil microbiome community at the UNDE site. *Burkholderia* could either be benenificial or detrimental to the ecosystem especially with the potential threat of acidic sphagnum bogs.

The BLAST matches were dominated by low quality matches to *Ralstonia solanacearum* and *Staphylococcus aureus*. The high frequency of low quality matches is likely an artifact of the sequencing and sequence matching method. Interestingly, *Ralstonia solanacearum* was transferred from the *Burkholderia* genus to the *Ralstonia* genus. Thus, the matches may have been indicative of a strong presence of *Burkholderia* at 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 data set 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. I curated the BLAST results by discarding poor quality matches based on percent identity, length of sequence match, expected value, and bitscore. I compared the results from BLAST to cleaning the data into a dada2 pipeline. Metadata from the UNDE site to the dada2 data to create a phyloseq object. This object was used for data analysis and visualization on the change in the composition and diversity of Phyla over time.

### 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 occurred most frequently throughout the data set. 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 box plots 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 sub sampled 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 phyloseq(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.

## Taxonomic Results

I found a variety of proteobacteria, spingobacteriales, and firmicutes in the ten most frequent BLAST matches by the highest taxonomic level as seen in Table 1. They were *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 drop off to *Staphylococcus* and *Bradyrhizobium*. The number of raw matches to the NCBI nucleotide database may not be the most reliable method of species identification. So, I compared these top results to the matches from the curated training data set through the DADA2 pipeline and curated the BLAST data with ggplot.

A few of the matches were present in both lists, however a number of BLAST matches were not present in the curated database. The ten most frequent genus matches by abundance in the curated data set were *Bradyrhizobium*, *Pseudomonas*, *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. As a result, I am fairly confident the samples from UNDE site are dominated by proteobacteria. However, the lack of *Ralstonia* and *Staphylococcus* makes me question if the samples are present in the UNDE samples. A summary of the top genera and their abundances can be found in table 2.

Curation of the BLAST matches based on percent identity, sequence match length, expected value, and bitscore show low quality matches to *Ralstonia* and *Staphylococcus*. It is unlikely *Ralstonia* is present at the UNDE site because many of the matching sequences had short sequence matches with low percent identities and high expected values and bitscores. In fact, only 945 of the 68508 *Ralstonia* samples, or 1.3%, remained after curation. Similarly, I have low confidence in the presence of *Staphylococcus* because of low sequence match lengths and low expected values (expected value < 0.00010, and bitscore < 200). Less than 0.05% of the *Staphylococcus* samples remained after curation. In contrast, I have high confidence in the presence of *Bradyrhizobium*, *Pseudomonas*, and proteobacteria at large. All matches for proteobacteria met the specified criteria for curation. More than 90% of the *Bradyrhizobium* samples met the specified criteria while 86% of the *Pseudomonas* samples met the specified criteria. Figures 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. 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.

I found the top 3 species for the genera *Bradyrhizobium* and *Pseudomonas*. There was 1144 *Bradyrhizobium elkanii*, 849 *Bradyrhizobium sp.*, 89 *Bradyrhizobium erythrophlei*, 288 *Pseudomonas sp.*, 212 *Pseudomonas frederiksbergensis*, and 140 *Pseudomonas protegens*. The *Ralstonia* matches were predominantly *Ralstonia solanacearum* (68,505 of the 68,508 samples, or 99.99%). Similarly, *Staphylococcus* matches were predominantly *Staphylococcus aureus* (3,059 of the 3,068, or 99.71%). The species returned for each genus in the BLAST results can be seen in Figure 3 and Table 3.

Overall, I am highly confident the UNDE site contains a variety of proteobacteria including *Bradyrhizobium* and *Pseudomonas* for a number of reasons: the BLAST samples were predominantly proteobacteria including these two genera, the curated data set matched a large number of proteobacteria and these two genera, the BLAST matches for these samples were high quality (percent identity > 90, sequence match length > 200, expected value < 0.00010, bitscore > 200). However, the low quality BLAST matches and single species dominance makes me skeptical of the presence of *Ralstonia* and *Staphylococcus* in the samples 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. In total there was eighteen plots sampled from the UNDE site. Seven of the eighteen plots were sampled in both 2016 and 2017: 001, 006, 008, 037, 038, 043, 044. Eight of the plots were only sampled in 2016: 007, 013, 014, 017, 019, 027, 034, and 035. Three of the plots were only sampled in 2017: 002, 003, and 010. Both archaea and bacteria were present in 2016 and 2017 with only a small number of archaea. There was 22 distinct archaea samples and 1,757 distinct bacterial samples matched to the curated database. The data set contained 1,099 distinct samples in 2016. In contrast, the data set contained 710 samples in 2017. The relative number of samples by year and Kingdom is illustrated in Figure 10.

Next, I analyzed the abundance of different Phyla by plot, year, and month. For sixteen of the eighteen plots proteobacteria was the most abundant Phylum. In plot 002 acidobacteria was the most abundant Phylum. Verrucomicrobia was the most abundant Phylum in plot 014. In both cases this abundance was much lower than many of the plots with Proteobacteria as the most abundant Phylum. For both years, Proteobacteria, Acidobacteria, and Actinobacteria were the three most abundant Phyla. The ratio of Phyla remained constant from 2016 to 2017. However, the abundance for each Phyla significantly declined from 2016 to 2017. Unsurprisingly, Proteobacteria was the most abundant Phyla in each month collected for both years. The abundances for all of the Phyla were significantly lower in 2017. Interestingly, no Actinobacteria were found in may of 2017 while it was the third most abundant Phyla in 2016. No samples were collected in October of 2016 and July of 2017.

Since proteobacteria was the most abundant Phyla I looked into the abundance of different genera of Proteobacteria. The most abundant proteobacteria was *Bradyrhizobium* followed by *Pseudomonas*, *Roseiarcus*, and *Burkholderia*. Proteobacteria as a whole were more abundant during May and August. However, *Pseudomonas* and *Roseiarcus* were almost twice as abundant during May than August. *Gaiella* were the most abundant Actinobacteria, followed by *Myobacterium*. *Gaiella* was only present during May.

Finally, I assessed the diversity of samples by plot by month and by year. Diversity was much higher for each plot during 2016. However, there are a significantly fewer diversity points plotted for 2017. For the months May, July, and August there does not appear to be a clear pattern for diversity. In plots 001, 007, 027, 035, 037, and 038 diversity seems to be higher in August than May. In plot 019 diversity appears to be higher in May than July. The highest diversity was in plot 001 in 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 woody wetlands dominated by thicks of alder. However, the land has a known susceptibility to overgrowth of acidic sphagnum bogs which out compete other species impacting the soil microbiome.

The data collected by NEON from this site varied in quality, but contained a substantial amount of genetic information. The 1192 samples had 51,289,602 sequences. Half of the sub sampled sequences were high quality, passing the trimming step of the pipeline. The inspected fastqc samples were also of sufficient quality.

I found a variety of proteobacteria, acidobacteria, and actinobacteria among the matches to both the BLAST analysis and DADA2 pipeline. Most notably, I found a high abundance of proteobacteria especially *Bradyrhizbium* and *Pseudomonas*. Many of the BLAST matches for *Ralstonia* and *Staphylococcus* were low quality and were discarded after curation. It is likely these results were artifacts of sequencing or sequence matching.

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.

## Taxonomic Interpretation

I am not confident in the presence of *Staphylococcus aureus* and *Ralstonia solanacearum* at the UNDE site. The firmicute *Staphylococcus aureus* is frequently found on the skin of humans. Since the sample matches were relatively small there is a chance these samples matched to *Staphylococcus aureus* were contaminated from the time of collection to sequencing. *Ralstonia solanacearum* 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). Interestingly, *solanacearum* was transferred from the *Burkholderia* genus to the *Ralstonia* genus. Thus, there is a chance many of the samples matched to *Ralstonia solanacearum* were actually members of the *Burkholderia* genus. This would explain the large number of low quality matches to *Ralstonia solanacearum*. However, Workers at the UNDE site should still be aware of the slight possibility of this bacteria’s presence on the property. Since a goal of the site is tree management plants near important tree species should be inspected for infected roots and xylem.

I am very confident there is a large abundance of proteobacteria and most notably *Bradyrhizobium*, *Pseudomonas*, and *Burkholderia* at the UNDE site. *Bradyrhizobium* have a key role in nitrogen cycling, helping to maintain soil fertility (Delamuta *et al.*, 2015). Thus, their presence is a positive sign for the health of the UNDE ecosystem. They will be vital to the microbial response following a buildup of sphagnum moss. In addition, *Pseudomonas* is known to protect plants from soil-borne fungal pathogens (Ramette *et al.*, 2011). One of the predominant species in this genus *Pseudomonas frederikbergensis* has proved to provide a similar protection for plants by enhancing cold and soil salinity tolerance in plants (Chatterjee *et al.*, 2017). The combination of these two proteobacteria provide strong indications that the UNDE ecosystem is currently healthy from a soil microbiome perspective. However, the affinity for the UNDE site to develop peat overgrowth is still a primary concern. *Burkholderia* have strong acid tolerances with strains which promote plant growth and strains that are pathogens for humans, animals, and plants (Eberl and Vandamme, 2016). Thus, their presence at the UNDE site can be seen as a double-edged sword. Their ability to survive in acidic environments have important implications for how the soil microbiome of the UNDE site could respond to global warming. With an increase in acidic sphagnum bogs, the population of *Burkholderia* is likely to grow. If the *Burkholderia* present at the UNDE site is healthy they would be an ideal nitrogen fixing bacteria to combat the impact of sphagnum overgrowth. However, if the *Burkholderia* are pathogen causing an shift toward acidic sphagnum bogs at the UNDE site could have negative impacts to the surrounding environment.

## Changes over Time

Its difficult to make strong inferences on the changes of the composition and diversity of microbial samples collected from the UNDE site due to variance in the number of samples taken between months and years. In addition, the NEON site collection started in 2016 at the site so only two years of data was collected. Different approaches can be taken to improve the data collection to draw meaningful inferences and will be fully discussed in the future work section.

## Limitations

I had a number of limitations for this study of the UNDE site. The first limitation is computational tractability. The overwhelming number of samples collected were not possible to analyze in the time permitted. As a result, I had to sub sample 0.5% of the sequences collected. This could have introduced a number of bias to the data. For example, by random chance a larger number of samples than representative of the actual data set could have been sampled during a specific year (like 2016) or month (like May).

The sub sampled data I analyzed had uneven collections which could have resulted from the aforementioned sub sampling, or the way NEON sampled the site. In either case this made it difficult to make accurate inferences on the change in composition and diversity over time and by site. Finally, there are only two years of data currently collected at this site. As NEON continues work on the UNDE site, more data can be collected from different months and years to provide more context on the UNDE ecosystem.

## Future Work

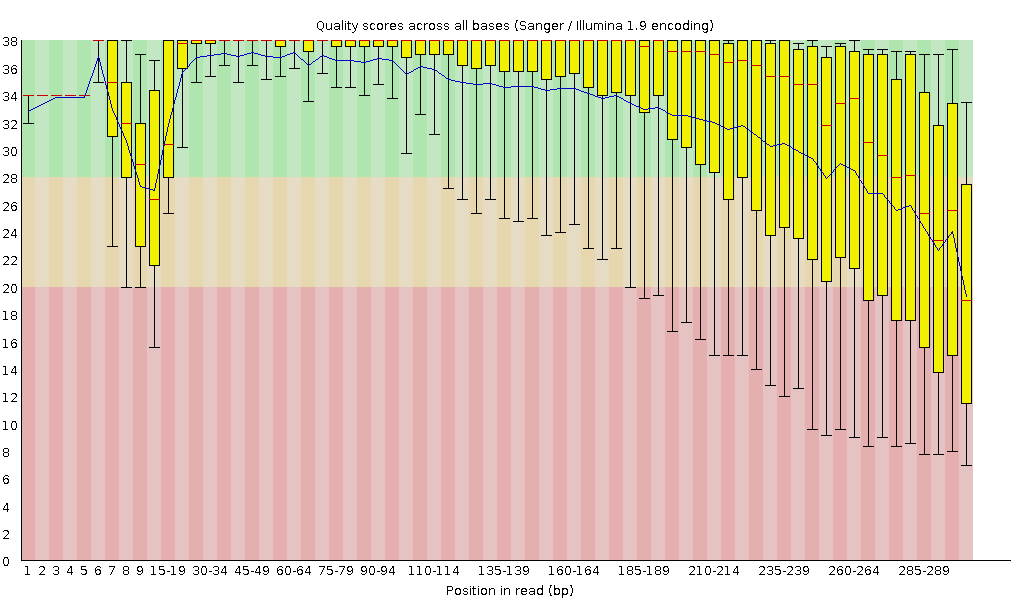
A large number of future projects can be analyzed from this data set. I would love to do more analysis without sub sampling data (or sub sampling to a lesser degree) or sub sample the data in a more delicate fashion. One way to go about sub sampling would be determine the different plots, months, and years NEON collected samples. Then, one could make sure to obtain an equal number of samples from each condition to make more even comparisons between conditions. Additionally, the metadata could be used to determine the relative percentage of the total samples in each condition. This could be used to ensure the sub sampled data matched the distribution of samples by condition. For instance, if 3% of the samples in the entire data set were collected at plot 001 in May of 2017, this could be reflected in the sub sampled data set. It was difficult to find information on what happened at the site during 2017. Further analysis into the metadata or an interview with staff at the UNDE site would be useful to determine if the uneven sampling was a result of decreased sampling or less organisms being present during 2017.

With more time I would break up the data by the type of plots present at the UNDE site: woody wetlands, mixed forest, deciduous forest to see what differences the soil microbiome samples have between site types. This information can be found on the NEON website and likely added to the metadata for the samples collected. I would also curate all of the BLAST matches based on quality score before determining the top 5 genus. This would provide context on the number of high quality matches against the BLAST database. In addition, more comparison matches would be useful to strengthen the inferences made. This could be done with additional training data sets.

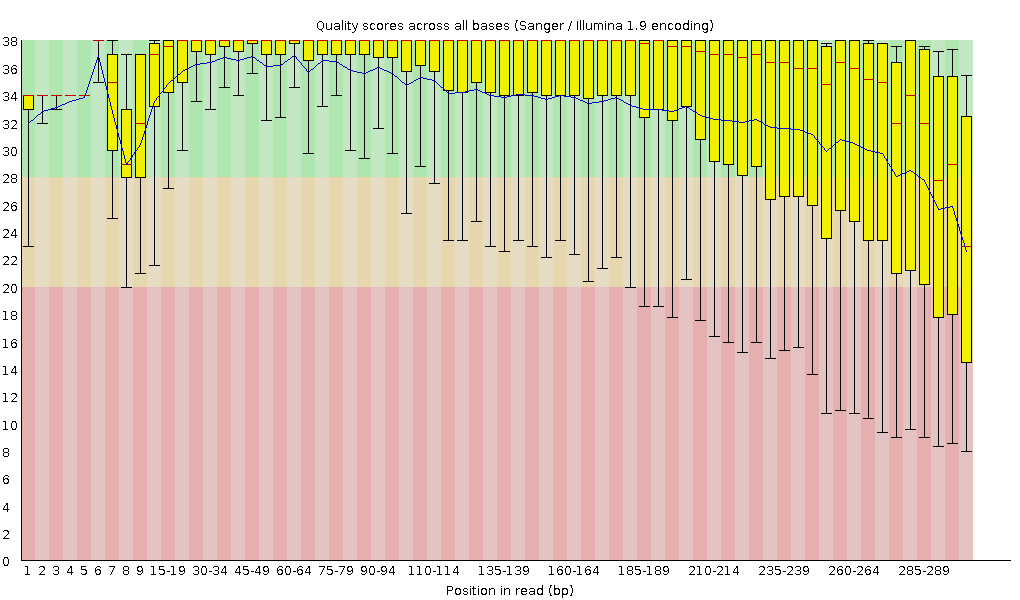
More analysis could be done on the samples which matched *Ralstonia solanacearum* in the NCBI nucleotide database. Since these samples could potentially be related to the *Burkholderia* genus the unique OTUs which matched *Ralstonia solanacearum* could be placed into phylogenetic context with different species of the *Burkholderia* genus and an out group. This would provide context into if these samples are a member of this genus or perhaps something else entirely. Additionally, as noted above the *Burkholderia* genus has both beneficial and detrimental strains. So, if the samples do fit into a larger *Burkholderia* phylogeny one could ascertain which strains of *Burkholderia* the mystery samples more closely resemble.

Finally, to target the primary objective of determining how the most abundant soil microbial taxa change over time continuing to track the site in the following years is vital. If sphagnum presence was added to the metadata one could track how the microbiome responds to and influences the changes in the ecosystem, especially as concerns over climate warming grow.

# 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.



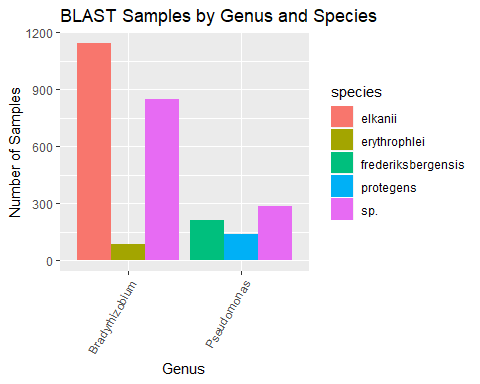
**Figure 2.** Fastqc per base sequence quality for BMI\_Tube52\_16S\_R1. This was one of the higher quality files. You can’t see a sharp decline until approximately 260 base pairs.

|  |  |
| --- | --- |
| 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 BLAST matches at the highest taxonomic level. Many of the genera listed are further divisions of proteobacteria including *Ralstonia*, *Bradyrhizobium*, *Pseudomonas*, and *Burkholderia*. The BLAST matches are predominantly *Ralstonia*.

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

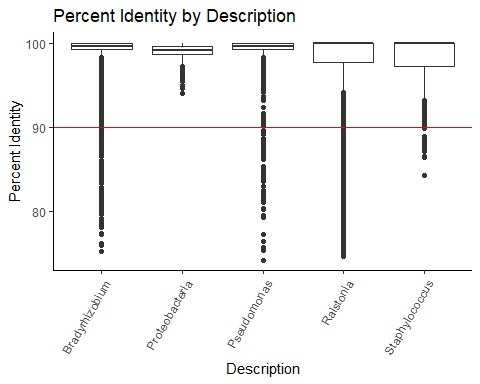
**Table 2.** Top 10 genus by abundance from curated database. *Bradyrhizobium*, *Pseudomonas*, and *Burkholderia* are in both the curated database and BLAST matches. *Rosiearcus*, *Mycobacterium*, *Thermosporothrix*, *Gaiella*, *Massilia*, *Pedomicrobium*, and *Gemmatimonas* are unique matches to the curated database.



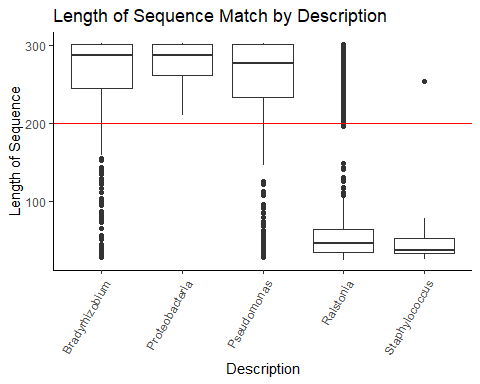
**Figure 3.** The top three species for *Bradyrhizobium* and *Pseudomonas* in BLAST matches. *Bradyrhizobium* is predominantly *B. elkanii* or *Bradyrhizobium sp.* while *Pseudomonas* is evenly distributed by *P. frederiksbergensis*, *P. protegens*, and *Pseudomonas sp.*. Both of these proteobacteria provide protection for a variety of plant species.

|  |  |
| --- | --- |
| Description | Count |
| *Ralstonia solanacearum* | 68505 |
| *Staphylococcus aureus* | 3059 |
| Proteobacteria | 897 |

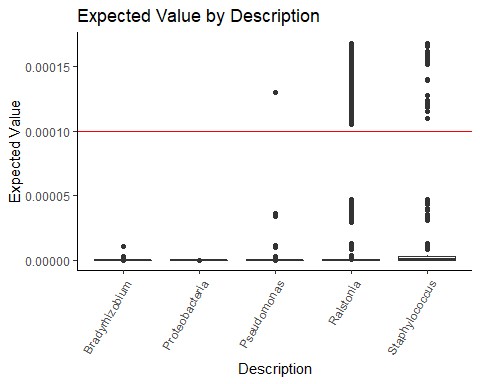
**Table 3.** Top 3 descriptions from BLAST matches. *Ralstonia* is dominated by *R. solanacearum* while *Staphylococcus* is predominantly *S. aureus*. Both of these counts are significantly higher than proteobacteria. Although proteobacteria is a broad category which also includes the genera in Figure 3.



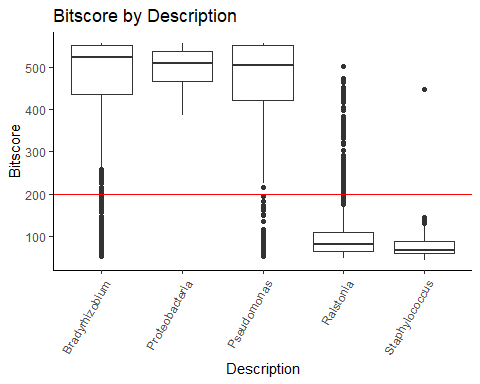
**Figure 4.** Boxplot of percent identity for the top 5 BLAST matches. The red line denotes the 90% cutoff used to curate the BLAST results. A majority of the samples have high percent identities, however proteobacteria is the only description with all matches above 90% identity.



**Figure 5.** Boxplot of length of sequence match for the top 5 BLAST matches. The red line denotes the 200 base pair cutoff used to curate the BLAST results. None of the proteobacteria matches are below the 200 base pair cutoff. A few *Bradyrhizobium* and *Pseudomonas* matches are below the cutoff. However, a majority of the *Ralstonia* and *Staphylococcus* matches have sequence matches of less than 100 base pairs.



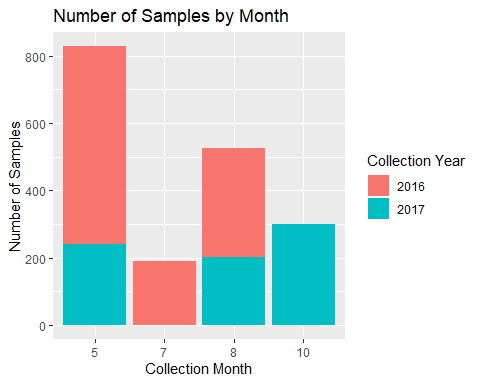
**Figure 6.** Boxplot of expected value for the top 5 BLAST matches. The red line denotes the 0.00010 cutoff used to curate the BLAST results. The figure was cutoff at the evalue of 0.0002 for visualization. However, a number of matches for *Ralstonia* and *Staphylococcus* were significantly higher. Many of the *Ralstonia* and *Staphylococcus* matches were above the cutoff and discarded.



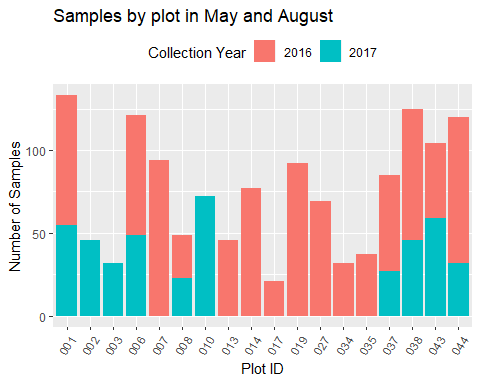
**Figure 7.** Boxplot of bitscore for the top 5 BLAST matches. The red line denotes the 200 bitscore cutoff used to curate the BLAST results. This also shows a large number of low quality matches for both *Ralstonia* and *Staphylococcus*.

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

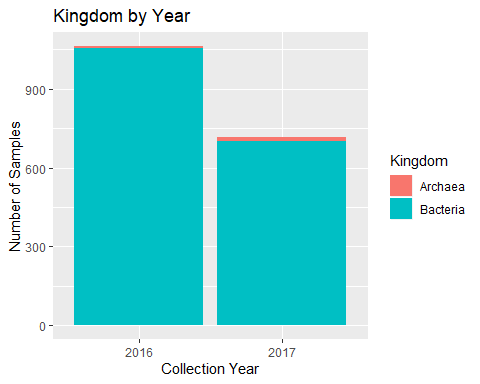
**Table 4.** Comparison of the initial counts for each BLAST match to their curated counts. *Ralstonia* and *Staphylococcus* drop off signficantly while *Bradyrhizobium*, *Pseudomonas*, and proteobacteria remain relatively constant. Thus, the matches to *Bradyrhizobium*, *Pseudomonas*, and proteobacteria had a higher quality. To see why samples were rejected see Figures 4-7.



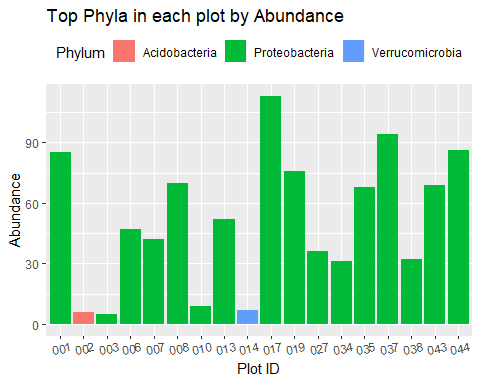
**Figure 8.** Number of samples by month. The number of samples varies from month to month influencing downstream analysis. July was only collected in 2016 while October was only collected in 2010. May had the most collections of any month by 300 samples.



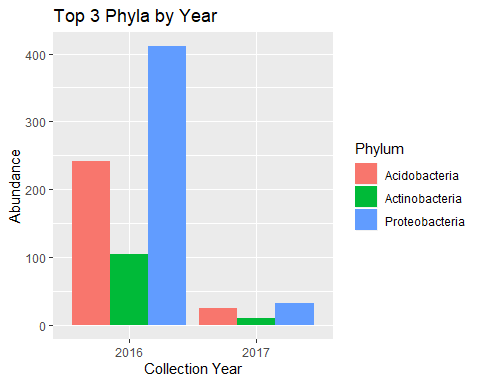
**Figure 9.** Number of samples by plot. There was variance in the number of samples taken in each plot. Plots 001, 006, 008, 037, 038, 043, 044 were collected in both 2016 and 2017. While plots 002, 003, and 010 were only collected in 2017 and plots 013, 014, 017, 019, 027, 034, 035 were only collected during 2016. Plot 001 had the highest number of samples collected while plot 017 had the least.



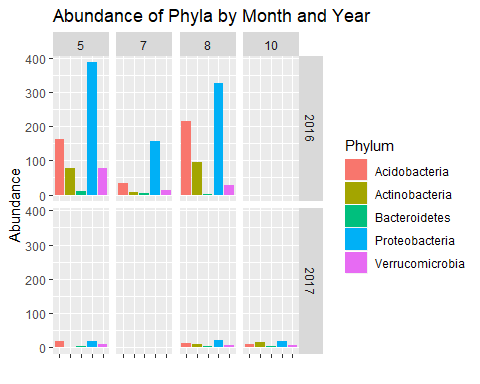
**Figure 10.** Summary of Kingdom by year. The samples were primarly bacterial, however a small number of the samples were archaeal. In addition, this graph shows that almost 300 more samples were collected in 2016.



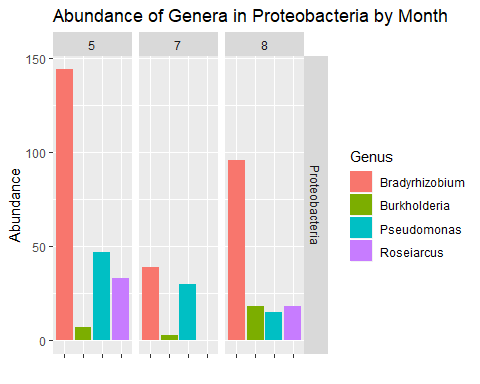
**Figure 11.** The top phyla in each plot. A majority of the plots are heavily dominated by proteobacteria. However, plot 002’s most abundant phyla was acidobacteria and plot 014’s most abundant phyla was verrucomicrobia. This could indicate a potentially acidic environment in plot 002, which could indicate a sphagnum bog.



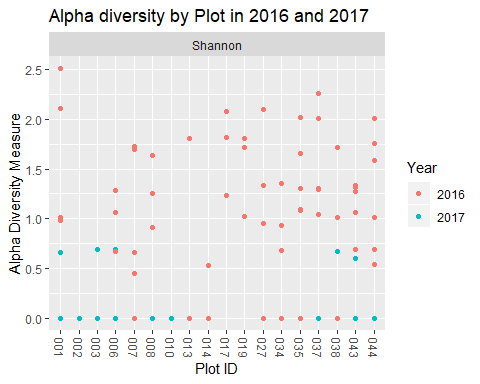
**Figure 12.** Top 3 Phyla by Year. Each Phyla is more abundant in 2016 than 2017. This could potentially be due to variance in the number of samples collected in 2016 and 2017. Both years have the same trend of proteobacteria, acidobacteria, and actinobacteria in similar proportions This indicates there may not be a significant change in the composition of Phyla from 2016 to 2017.



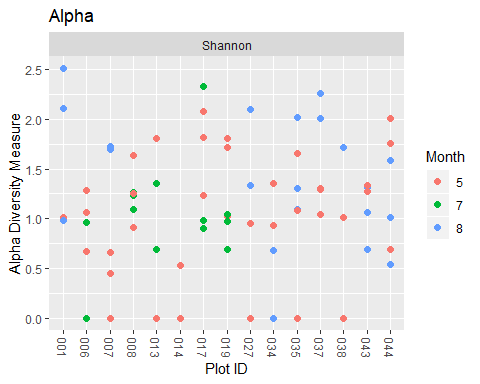
**Figure 13.** Abundance of Phyla by Month and Year. There is a significant difference between the abundance of samples in 2016 to 2017. In 2016 proteobacteria are the most abundant phyla for each month. Each month shows a similar trend. Except, May shows a higher relative abundance of verrucomicrobia.



**Figure 15.** Abundance of different genera in Proteobacteria by month. *Bradyrhizobium* is the most abundant genus each month samples were collected followed by *Pseudomonas*. *Burkholderia* abundance increases in August indicating the potential of a more acidic environment. In August the total abundance drops as well.



**Figure 15.** Alpha diversity by plot in 2016 and 2017. Alpha diversity was measured using the Shannon method. The diversity is consistently higher for each plot in 2016 over 2017. This could be due to a large number of samples being measured in 2016.



**Figure 15.** Alpha diversity by plot by month in 2016. Alpha diversity was measured using the Shannon method. The diversity doesn’t follow a consistent pattern,

# 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.

Bragazza,L. *et al.* (2013) Biogeochemical plant–soil microbe feedback in response to climate warming in peatlands. *Nature Climate Change*, **3**, 273.

Breeuwer,A. *et al.* (2009) Decreased summer water table depth affects peatland vegetation. *Basic and Applied Ecology*, **10**, 330–339.

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.

Chatterjee,P. *et al.* (2017) Beneficial soil bacterium pseudomonas frederiksbergensis os261 augments salt tolerance and promotes red pepper plant growth. *Frontiers in plant science*, **8**, 705.

Delamuta,J.R.M. *et al.* (2015) Bradyrhizobium tropiciagri sp. Nov. And bradyrhizobium embrapense sp. Nov., nitrogen-fixing symbionts of tropical forage legumes. *International Journal of Systematic and Evolutionary Microbiology*, **65**, 4424–4433.

Eberl,L. and Vandamme,P. (2016) Members of the genus burkholderia: Good and bad guys. *F1000Research*, **5**.

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*.

Ramette,A. *et al.* (2011) Pseudomonas protegens sp. Nov., widespread plant-protecting bacteria producing the biocontrol compounds 2,4-diacetylphloroglucinol and pyoluteorin. *Systematic and Applied Microbiology*, **34**, 180–188.

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.