Diverse Microbial Communities in Soil: A 16S rRNA Sequencing Study

Victoria Burke

Distributed Research Apprenticeships for Master's Program

December 2024

Abstract

Understanding the composition and diversity of soil microbiomes is critical for breaking down their roles in ecosystem functionality and resilience. This study employs 16S rRNA sequencing to analyze the microbial communities in 190 soil samples, focusing on beta diversity and phylogenetic relationships. Using QIIME2 for data processing, hundreds of thousands of gene sequences were analyzed to construct phylogenetic trees and explore microbial diversity patterns. Preliminary findings indicate distinct microbial community profiles influenced by environmental variables, offering insights into soil health and its implications for agriculture and sustainability. This research highlights the utility of 16S rRNA sequencing in advancing soil microbiome studies and sets the stage for future investigations into microbial ecosystems.

Introduction

The study of soil microbiomes has garnered increasing interest due to their vital role in ecosystem functioning, including nutrient cycling and plant health. One of the most effective approaches to analyzing microbial communities is through 16S ribosomal RNA (16s rRNA) sequencing, a method that provides insights to taxonomic composition and diversity of microbial populations. The 16S rRNA gene is part of the small subunit of the ribosome, a structure that

helps cells produces proteins. It is especially useful for studying microorganisms because of two unique characteristics:

1. Conserved Regions: These are parts of the gene that stay almost the same across different microbial species. Scientists use these conserved regions to design universal primers (short pieces of DNA used to start the process of copying the 16S rRNA gene during sequencing).
2. Variable Regions: Sections of the gene that vary from one species to another. These differences allow researchers to identify and classify different microbes based on their genetic information.

By sequencing the 16S rRNA gene, we can identify the types of microorganisms present in a sample and estimate their abundance. In this report, we will explore the soil microbiome of 190 samples by analyzing the 16S rRNA sequencing data using a bioinformatics tool called QIIME2 (pronounced chime). This tool helps process the sequencing data, compare microbial communities, and build phylogenetic trees to build a deeper understanding of the composition and relationships within the soil microbiome, contributing to our knowledge of microbial interactions and their environmental implications. According to Estaki et al., “QIIME 2 can also process other types of microbiome data, including amplicons of other markers such as 18S rRNA, internal transcribed spacers (ITS), and cytochrome oxidase I (COI), shotgun metagenomics, and untargeted metabolomics”. QIIME2, a comprehensive bioinformatics platform tailored for microbiome studies, uses an intuitive interface, detailed documentation, and a plugin-based architecture. Its focus on microbiome data and strong community support makes it a reliable and flexible choice for 16s rRNA analysis.

Related Work

There are extensive studies regarding 16S rRNA sequencing, surely becoming a cornerstone of microbial community analysis. Previous studies have leveraged 16S rRNA sequencing to investigate microbial diversity in contexts such as human health, agriculture, and environmental ecosystems. According to J. Michael Janda and Sharon L. Abbott from the Journal of Clinical Microbiology, 16S rRNA gene sequencing is the most common practice to study bacterial phylogeny and taxonomy.

Methods, Experiments, and Results

Soil samples were collected from consistent depths using sterile tools, stored at cool temperatures to preserve DNA integrity, and processed to extract microbial DNA. The DNA extraction process is intricate, and the first initial step is potentially the most crucial. The process of DNA extraction from soil can be done directly by lysing cells with soil particles present, or indirectly by separating cells before lysis (Young et al., 2014). These limitations could influence the accuracy of phylogenetic and diversity analyses, which are important to understanding microbial composition and health. Once DNA was extracted and purified, the DNA is sequenced and ready for analysis.

The first step in the microbiome analysis is to import the raw DNA sequencing data into QIIME2. QIIME2 uses a specific file format called .qza, which acts as a container for data like FASTQ files as well as metadata about the data type. FASTQ formats include either single-end or paired-end DNA reads. This data may be multiplexed, which means all the sequences are in a single file with barcodes identifying samples, or demultiplexed, where sequences are already

sorted by sample. For this analysis, paired-end demultiplexed FASTQ files were used, consisting of forward and reverse reads (commonly labeled as R1 and R2) along with a separate barcode file. After demultiplexing, I generated a summary of the results which allowed me to determine how many sequences were obtained per sample, as well as a summary of the distribution of sequence qualities at each position. In the analysis of the soil microbiome, sequencing data were successfully demultiplexed to yield a total of 6,404,955 paired-end reads, with equal contributions from forward and reverse reads. Sequence counts across 190 samples demonstrated consistent sequencing depth, with a mean of 33,710 reads per sample. Maximum reads per sample reached 50,476, while the minimum was notably lower at 78 reads.

While most samples exhibited sequencing depths within an acceptable range (35,000-47,000 reads), a small subset of samples presented with significantly lower read counts. A subset of samples, including sterile water controls (e.g., Sample ID 2158-096-SterileWater, 78 reads) and no-template controls (e.g., 2158-190-NTC4, 215 reads), displayed very low sequence counts.

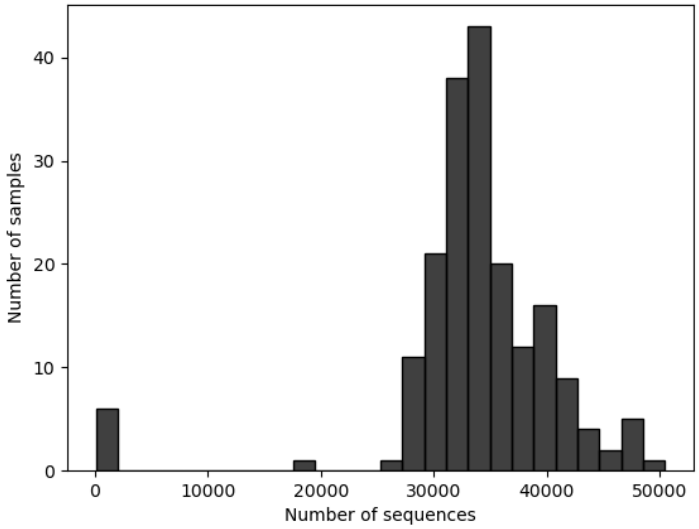

Figure 1. Forward and Reverse Reads Frequency Histogram

The even distribution of reads across samples, as indicated by comparable median and mean values, suggests minimal technical bias during sequencing.

Following the demultiplexing of sequencing data, which yielded a total of over 6.4 million paired-end reads across 190 samples, quality control measures were applies to ensure the integrity of the data. Using Deblur for denoising (correcting sequencing errors, removing artifacts, etc.), low-quality sequences were filtered out and reads were truncated at predefined thresholds. The results of this process, as summarized below, demonstrate the effectiveness of quality control and provide insights into the composition and integrity of the dataset.

	sample-id	reads-raw	fraction-artifact-with-minsize	fraction-artifact	fraction-missed-reference	unique-reads-derep	reads-derep	unique-reads-deblur	reads-deblur	unique-reads-hit-artifact	reads-hit-artifact	unique-reads-chimeric	reads-chimeric	unique-reads-hit-reference	reads-hit-reference	unique-reads-missed-reference	reads-missed-reference
0	2158-190-NTC4	215	0.958140	0.0	0.000000	4	9	4	9	0	0	0	0	3	7	0	0
1	2158-168-231021-FC	17954	0.587724	0.0	0.003617	1862	7402	1610	4947	0	0	292	523	1073	3934	5	16
2	2158-105-220509-HW	25642	0.523828	0.0	0.000000	2869	12210	2279	7338	0	0	404	701	1723	6381	0	0
3	2158-048-210723-WH	31603	0.507167	0.0	0.002592	3443	15575	2687	9072	0	0	485	971	1923	7589	7	21
4	2158-149-230527-JW	30328	0.505078	0.0	0.000000	3456	15010	2669	8886	0	0	443	885	1531	6474	0	0
5	2158-106-220509-Soy	32933	0.503993	0.0	0.000000	3658	16335	2791	9246	0	0	491	910	1664	7141	0	0
6	2158-096-SterileWater	78	0.500000	0.0	0.468750	5	39	5	32	0	0	0	0	0	0	1	15
7	2158-184-231021-HW	31345	0.496188	0.0	0.002336	3635	15792	2814	9059	0	0	470	925	2091	7681	5	19
8	2158-161-230527-HN	30896	0.489416	0.0	0.000000	3748	15775	3013	9198	0	0	570	1080	2129	7562	0	0
9	2158-086-220718-Soy	32735	0.488010	0.0	0.009126	3686	16760	2814	9600	0	0	520	1053	1859	7650	3	78
10	2158-152-230527-WH	35644	0.487151	0.0	0.016436	3999	18280	2936	10239	0	0	496	991	2075	8484	23	152

Figure 2. Deblur Summary (10 of 190 samples shown)

Most samples showed low levels of artifacts and retained ~7,000-10,000 unique reads, indicating robust recovery of biological sequences. Negative controls exhibited expected high artifact

fractions and low read counts, confirming minimal contamination. The dataset's high quality and consistency provide a solid basis for downstream analyses.

With the Deblur analysis providing a refined set of high-quality sequences, the next step involved organizing these sequences into a feature table. This table consolidates the data by quantifying the occurrence of each unique sequence across samples. The feature table illustrates that there are 15,216 unique features (ASVs or OTUs) identified across all samples, suggesting that there is a high level of biodiversity. ASVs (Amplicon Sequence Variants) represent unique DNA sequences found in the data after error correction and quality filtering which indicate that the analysis performed can achieve higher resolution in identifying and comparing microbial diversity across samples. Moreover, the feature table provides a median frequency of 1,851 reads per sample, with the first quartile at 8,081 and the third quartile at 9,660, showing that the majority of samples are evenly sequenced. This table helps affirm that the data is well-distributed across most samples, with a median close to the mean, indicating minimal skewness.

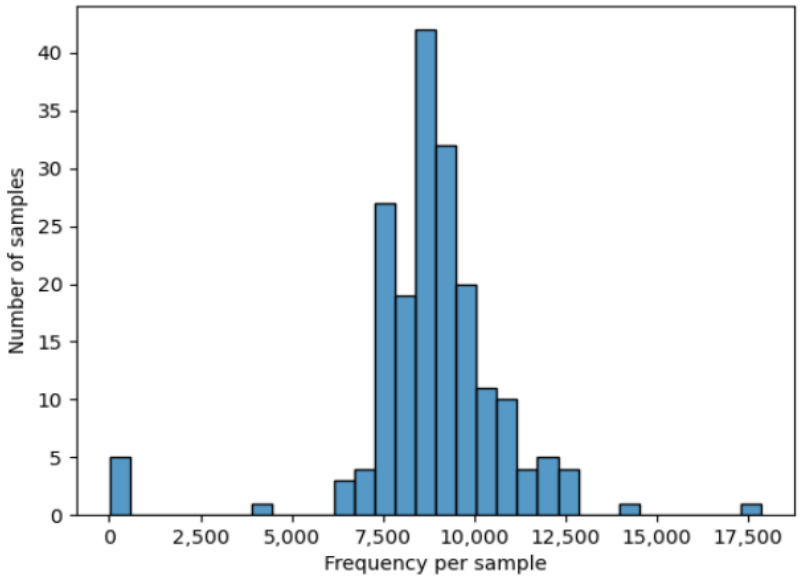

Figure 3. Feature Table (per sample)

The feature table also provides frequency per feature, which suggests that most features are relatively rare, while a small number of features are highly abundant. This is expected in microbial communities, where a few dominant taxa coexist with many rare ones.

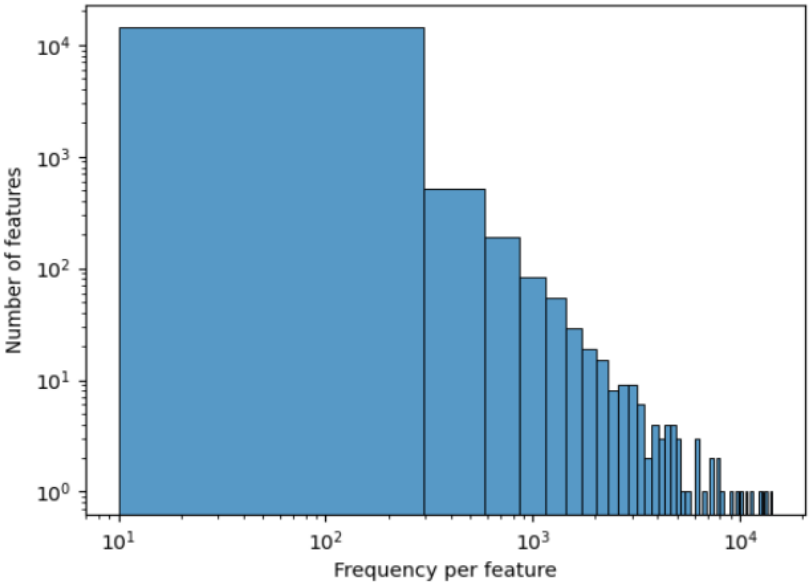

Figure 4. Feature Table (per feature)

Once I had composed the feature table, the next logical step was to generate a phylogenetic tree. A phylogenetic tree is a diagram that represents the evolutionary relationships among a group of organisms, genes, or sequences. There are four components of a phylogenetic tree: Nodes (common ancestors), Branches (evolutionary lineages), Root (most recent common ancestor of all entities in the tree), and Clades (groups of organisms or sequences that share a common ancestor). The phylogenetic tree helps to compute diversity metrics, which measure the variation in microbial communities within a sample (alpha diversity) or between various samples (beta diversity). To explore the compositional differences in microbial communities across samples, beta diversity analyses were performed using the Bray-Curtis dissimilarity and the Jaccard index.

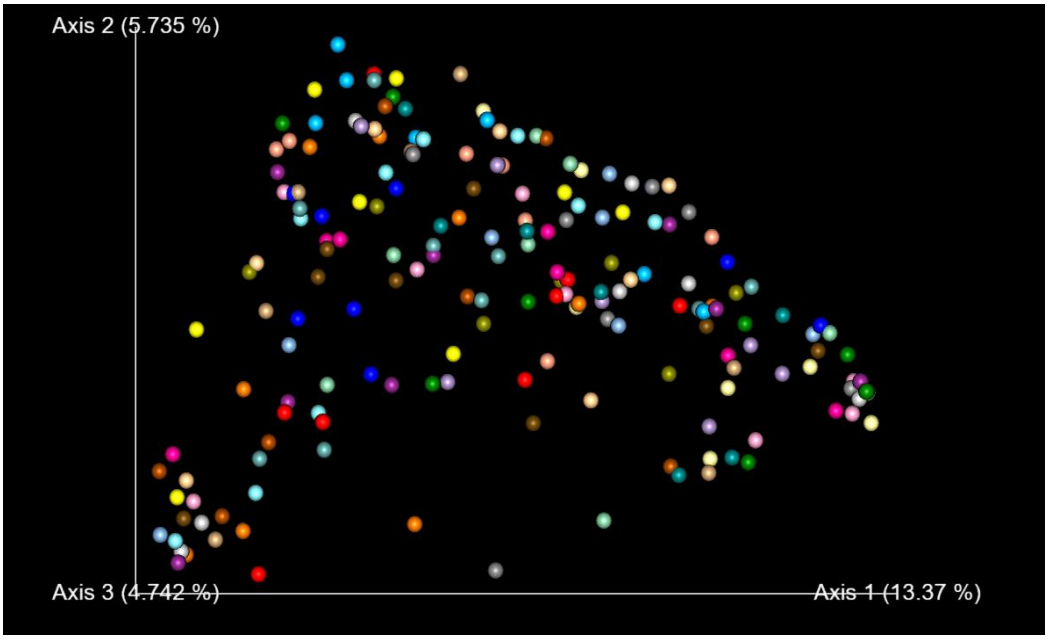

Figure 5. Bray-Curits Emperor Visualization

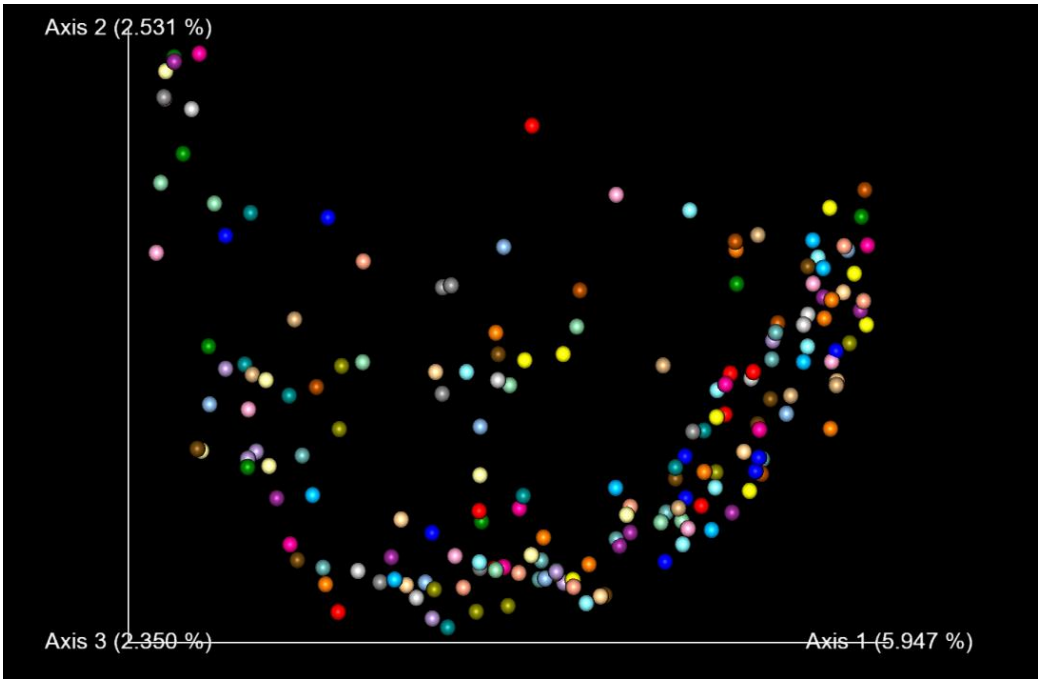

Figure 6. Jaccard Emperor Visualization

The Bray-Curtis plot highlights community differences based on the relative abundance of taxa. Axis 1 explains 13.37% of the variability, with Axes 2 and 3 accounting for 5.735% and

4.742% respectively. Samples cluster into distinct groups, suggesting that abundance differences drive community composition. This pattern indicates that dominant taxa play a significant role in shaping community structure, and samples with similar abundances cluster closer together. The Jaccard visualization focuses on the presence or absence of taxa. Axis 1 explains 5.947% of the variability, with Axes 2 and 3 contributing 2.531% and 2.350% respectively. While the clustering patters are similar, Jaccard emphasizes rare or unique taxa, which can result in different groupings. Both visualizations collectively highlight the diversity within and between microbial communities. Consistent clustering across both metrics suggests that shared taxa and their abundances equally contribute to community composition. However, discrepancies between the two indicate that abundance plays a critical role in observing patterns, particularly for dominant taxa.

Moreover, I used Faith's Phylogenetic Diversity Group Significance Analysis which is an alpha diversity metric that measures the total branch length of phylogenetic trees. The group significance analysis tests whether there are statistically significant differences in Faith PD values between predefined sample groups, which can help determine whether community diversity varies across experiments.

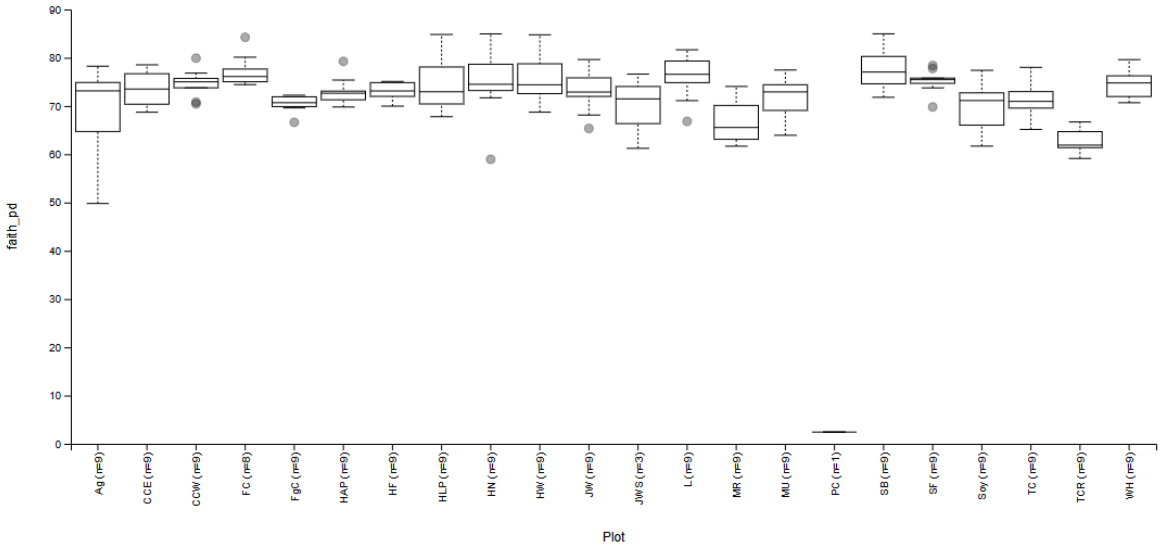

Figure 7. Alpha Diversity Boxplot

As the boxplot shows, the median values and range vary between groups, with some groups having a wider range compared to others. This means that certain groups exhibit more variability in their phylogenetic diversity while others are consistent. The alpha diversity analysis also included the Kruskal-Wallis test, which assesses if there are significant differences in Faith PD values. The Kruskal-Wallis test returned an H-statistic of 71.27 and a p-value of 2.19×10^−7, which reveals significant differences indicating that phylogenetic diversity varies across groups. The H-statistic measures the degree to which the distributions of a numeric variable differ between two groups or more, and in this case, a high H-statistic indicates greater difference between group median/distribution. The p-value is the probability of observing the data if the null hypothesis is true, and my tests returned an extremely small value meaning that there are statistically significant differences in Faith PD values.

Discussion and Conclusion

This study highlights the utility of 16S rRNA sequencing and advanced bioinformatics tools like QIIME2 for exploring microbiomes, successfully exploring the diversity and composition across 190 samples. The demultiplexing process yielded over 6 million paired-end reads, with a mean of 33,710 reads per sample, indicating an even distribution and minimal sequencing bias. Rigorous quality control using Deblur revealed low levels of artifacts, confirming minimal contamination. The phylogenetic tree constructed from the processed data revealed highly diverse microbiomes, highlighting the richness and variability of soil microbial communities. These findings align with broader ecological observations that emphasize the critical role of microbial communities in maintaining ecosystem functionality. The relationship between plant and microbial communities not only regulate productivity but also underscore the importance of conserving plant diversity to sustain ecological balance (Zak et al., 2003). By providing insight into the diversity and distribution of soil microbiomes, this study lays a foundation for future research into the interconnected roles of plants and microbes.

References

Amir, Amnon. “Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns | Msystems.” *Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns*, 17 Mar. 2017, journals.asm.org/doi/10.1128/mSystems.00191-16.

Bolyen, Evan, et al. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology*, U.S. National Library of Medicine, 9 Aug. 2019, pmc.ncbi.nlm.nih.gov/articles/PMC7015180/.

Chang, Hao-Xun, et al. “Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity.” *Frontiers*, Frontiers, 13 Mar. 2017, www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.00519/full.

Clarridge, Jill E. “IMPACT OF 16S Rrna Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases.” *Clinical Microbiology Reviews*, U.S. National Library of Medicine, Oct. 2004, pmc.ncbi.nlm.nih.gov/articles/PMC523561/.

Estaki, Mehrbod. “QIIME 2 Enables Comprehensive End-to-end Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data - ESTAKI - 2020 - Current Protocols in Bioinformatics - Wiley Online Library.” *QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data*, Current Protocols, 28 Apr. 2020, currentprotocols.onlinelibrary.wiley.com/doi/full/10.1002/cpbi.100.

Janda, J. Michael, and Sharon L Abbott. “16S Rrna Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls | Journal of Clinical Microbiology.” *Journal of Clinical Microbiology*, 1 Sept. 2007, journals.asm.org/doi/10.1128/JCM.01228-07.

Janssen, Peter H. “Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes.” *Applied and Environmental Microbiology*, 1 Mar. 2006, journals.asm.org/doi/full/10.1128/aem.02407-21.

Ricotta, C., and J. Podani. “On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning.” *Ecological Complexity*, Elsevier, 27 July 2017, www.sciencedirect.com/science/article/pii/S1476945X17300582.

Woese, Carl R. “Interpreting the Universal Phylogenetic Tree.” *PNAS*, 18 July 2000, www.pnas.org/doi/abs/10.1073/pnas.2301463120.

Yang, Dongyang, and Wei Xu. “Clustering on Human Microbiome Sequencing Data: A Distance-Based Unsupervised Learning Model.” *MDPI*, Multidisciplinary Digital Publishing Institute, 20 Oct. 2020, www.mdpi.com/2076-2607/8/10/1612.

Young, Jennifer M, et al. “Limitations and Recommendations for Successful DNA Extraction from Forensic Soil Samples: A Review.” *Science & Justice*, Elsevier, 11 Mar. 2014, www.sciencedirect.com/science/article/pii/S1355030614000082.

Zak, Donald R. “PLANT DIVERSITY, SOIL MICROBIAL COMMUNITIES, AND ECOSYSTEM FUNCTION: ARE THERE ANY LINKS?” *ESA Journals*, Wiley Library, 1 Aug. 2003, onlinelibrary.wiley.com/doi/abs/10.1890/02-0433.