Microbiome Data Analysis with MicrobiomeAnalyst

User ID: guest5767366448178020058

February 14, 2020

1 Data Processing and Normalization

1.1 Reading and Processing the Raw Data

MicrobiomeAnalyst accepts count data in a variety of formats generated in microbiome studies including plain text table, biom format as well as output from mothur pipeline. User need to upload their data in one of three available formats and specify the taxonomic labels when uploading their data in order for MicrobiomeAnalyst to process the taxonomic information correctly. The hierarchical information for taxa can either be present within abundance data or uploaded as a separate taxonomy table file (.txt or .csv format). Also, uploading metadata file is required as plain text (.txt or.csv) with all three formats.

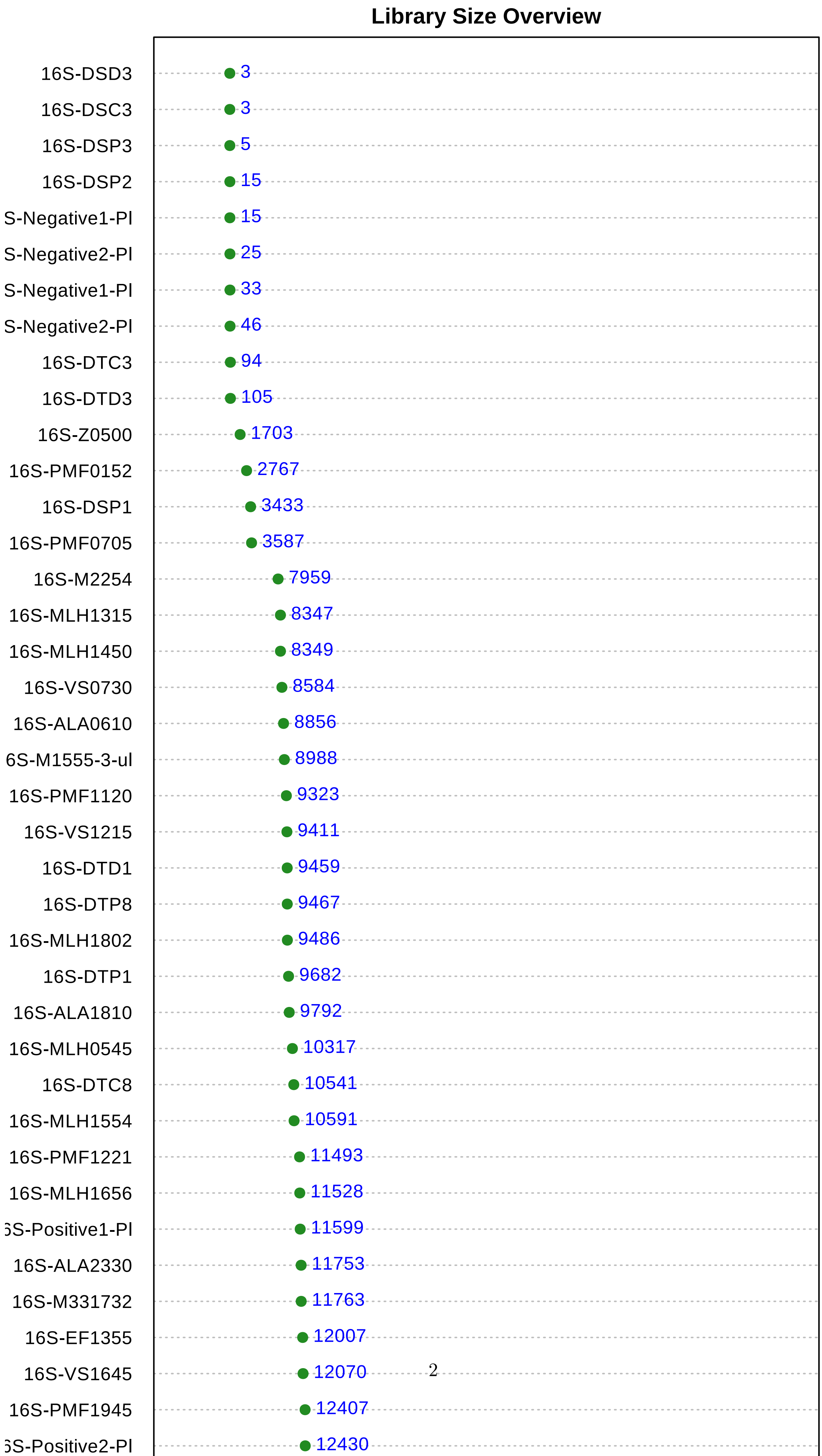
1.1.1 Reading abundance count data table

The abundance count data should be uploaded in tab-delimited text (.txt) or comma separated values (.csv) format. Samples are represented in columns, while rows contains the information about the features. Metadata file contains additional information about samples such as experimental factors or sample grouping.

A total of 112 samples and 4249 features or taxa are present. The sample data contains a total of 112 samples and 1 sample variables.The OTUs are annotated as Greengenes label.

1.1.2 Data Integrity Check

Before data analysis, a data integrity check is performed to make sure that all the necessary information has been collected. The sample variable should contain atleast two groups to perform most of the comparative analysis. *By default, sample variables which are found to be constant and continuous in nature will be removed from further analysis. Additionally, features just present in one sample will also be discarded from the data.* Figure 1 shows the library size for inspection of each sample.


Library Size Overview
16S-DSD3 3
16S-DSC3 3
16S-DSP3 5
16S-DSP2 15
S-Negative1-PI 15
S-Negative2-PI 25
S-Negative1-PI 33
S-Negative2-PI 46
16S-DTC3 94
16S-DTD3 105
16S-Z0500 1703
16S-PMF0152 2767
16S-DSP1 3433
16S-PMF0705 3587
16S-M2254 7959
16S-MLH1315 8347
16S-MLH1450 8349
16S-VS0730 8584
16S-ALA0610 8856
6S-M1555-3-ul 8988
16S-PMF1120 9323
16S-VS1215 9411
16S-DTD1 9459
16S-DTP8 9467
16S-MLH1802 9486
16S-DTP1 9682
16S-ALA1810 9792
16S-MLH0545 10317
16S-DTC8 10541
16S-MLH1554 10591
16S-PMF1221 11493
16S-MLH1656 11528
3S-Positive1-PI 11599
16S-ALA2330 11753
16S-M331732 11763
16S-EF1355 12007
16S-VS1645 12070
16S-PMF1945 12407
3S-Positive2-PI 12430

1.1.3 Data Filtering

The purpose of the data filtering is to identify and remove features that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step can usually improves the results. Features having low count and variance can be removed during the filtration step. Features having very few counts are filtered based on their abundance levels (minimum counts) across samples (prevalence). Other than sample prevalence, such features can also be detected using minimum count cutoff based on their mean and median values. Features or taxa with constant or less variable abundances are invaluable for comparative analysis. Such features are filtered based on their inter-quantile ranges, standard deviations or coefficient of variations. *By default, features having zero counts across all the samples, or only appears in one sample will be removed from further analysis.*

A total of 2111 low abundance features were removed based on prevalence. A total of 22 low variance features were removed based on iqr. The number of features remains after the data filtering step: 400

1.2 Data Normalization

The data is stored as a table with one sample per column and one variable (taxa or OTU) per row. The normalization procedures implemented below are grouped into three categories. Data rarefaction and scaling based methods deal with uneven sequencing depths by bringing samples to the same scale for comparison. While transformation based methods account for sparsity, compositionality, and large variations within the data. You can use one or combine all three to achieve better results. For more information about these methods, please refer to the paper by Weiss et al.[1] The normalization consists of the following options:

1. Data rarefying (with or without replacement)
2. Data scaling:
    - Total sum scaling (TSS)
    - Cumulative sum scaling (CSS)
    - Upper-quantile normalization (UQ)
3. Data transformation :
    - Relative log expression (RLE)
    - Trimmed mean of M-values (TMM)
    - Centered log ratio (CLR)

No data rarefaction was performed. No data scaling was performed. No data transformation was performed.

2 Marker Gene Analysis

MicrobiomeAnalyst offers a variety of methods commonly used in microbiome data analysis. They include:

1. Visual exploration:
    - Stacked bar/area plot
    - Rarefaction curve

[1] Weiss et al. *Normalization and microbial differential abundance strategies depend upon data characteristics*, Microbiome 2017

   - Phylogenetic tree
   - Heat tree

2. Community profiling:
   - Alpha diversity analysis
   - Beta Diversity analysis
   - Core microbiome analysis

3. Clustering analysis:
   - Heatmap
   - Dendrogram
   - Correlation analysis
   - Pattern Search

4. Differential abundance analysis:
   - Univariate analysis
   - metagenomeSeq
   - RNAseq methods

5. Biomarker analysis:
   - LEfSe
   - Random Forests

6. Predictive functional profiling:
   - PICRUSt
   - Tax4Fun

2.1 Visual Exploration

These methods are used to visualize the taxonomic composition of community through direct quantitative comparison of abundances. MicrobiomeAnalyst provides an option to view this composition at various taxonomic levels (phylum, class, order) using either stacked bar/stacked area plot or piechart. Viewing composition at higher-levels (phylum) provides a better picture than lower-levels (species) when the number of species in a community is large and diversified. Additionally, such taxonomic abundance or composition can be viewed at community-level (all samples), sample-group level (based on experimental factor) or at individual sample-level. Taxa with very low read counts can also be collapsed into `Others` category using a count cutoff based on either sum or median of their counts across all samples or all groups. Merging such minor taxa will help in better visualization of significant taxonomic patterns in data. Figure 2 shows the taxonomic composition using Stacked bar/area plot.

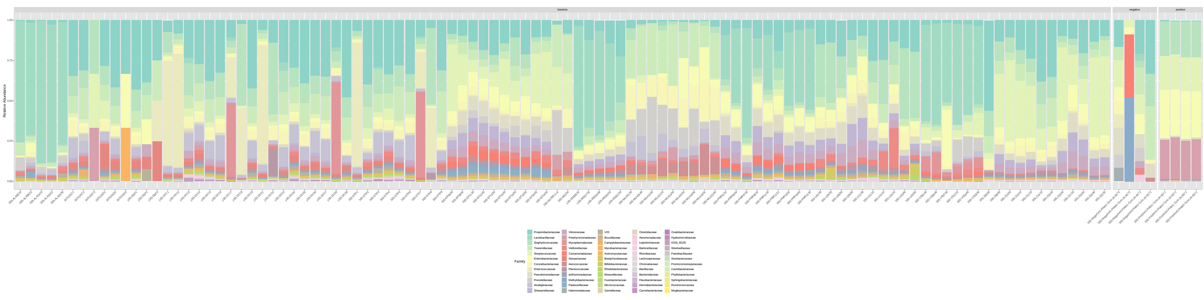


Figure 2: Taxonomic composition of community at `Family` level using `Stacked Bar` plot

2.2 Rarefaction curve analysis

This method is used to present relationship between number of OTUs and number of sequences It can infer if the reads of a sample is enough to reach plateau, which means that with increasing of sequences, the gain of newly discovered OTUs is limited. If sequence depth of some samples are not enough, you may consider to resequence these samples or removed from downstream analysis. User can choose different metadata variables as group, line colors and line types. Rarefaction curve analysis is performed using the modified function ggrare originated from `ranacapa` package[2]

Figure 3 shows the rarefaction curve.

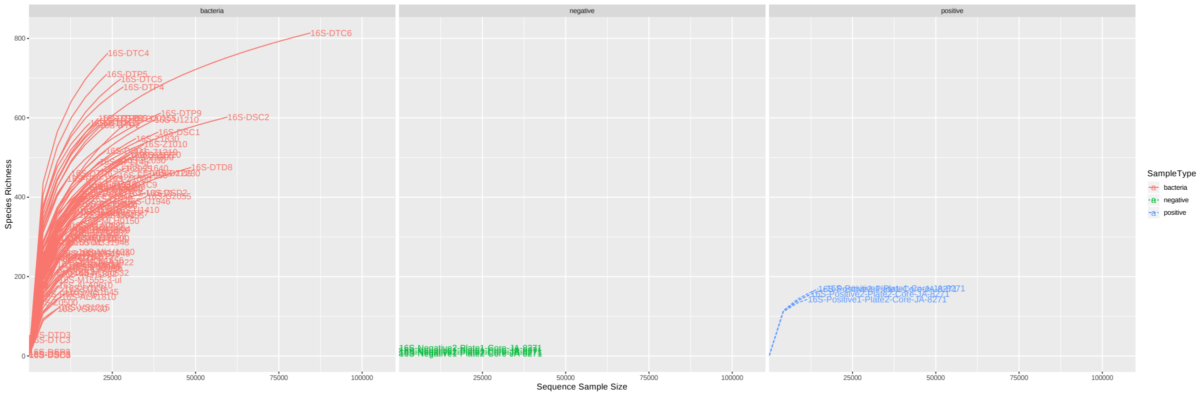


Figure 3: Rarefaction curve using `original` dataset

[2]Gaurav S. Kandlikar *ranacapa: An R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations.*, 2018.

2.3 Alpha diversity analysis

This method is used to measure the diversity present within a sample or community. Alpha diversity can be characterized via the total number of species (richness), the abundances of the species (evenness) or measures that considered both richness and evenness. How these measures estimates the diversity is need to be considered when performing alpha-diversity analysis. User can choose from richness based measure such as Observed index which calculates the actual number of unique taxa observed in each sample. While the Chao1 and ACE measures estimate the richness by inferring out the number of rare organisms that may have lost due to undersampling. Also,there are indices such as Shannon, Simpson and Fisher in which along with the number (richness), the abundance of organisms (evenness) is also measured to describe the actual diversity of a community.

Alpha diversity analysis is performed using the `phyloseq` package[3]. The results are plotted across samples and reviewed as box plots for each group or experimental factor. Further, the statistical significance of grouping based on experimental factor is also estimated using either parametric or non-parametric test. Figure 4 shows the alpha diversity measure across all the samples for given diversity index. Figure 5 shows the diversity distribution using box plot for a given group or experimental factor.

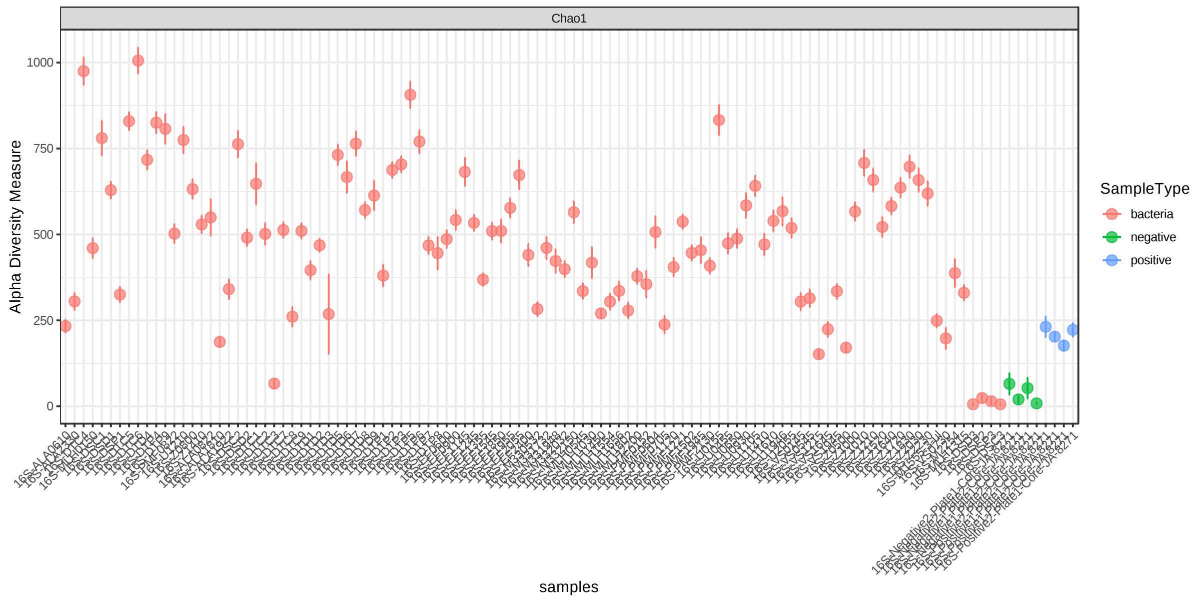


Figure 4: Alpha-diversity measure using `Chao1` at `OTU` level across all the samples. The samples are represented on X-axis and their estimated diversity on Y-axis. Each sample is colored based on `SampleType` class

[3]Paul J. McMurdie *phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data.*, 2013, R package version 1.19

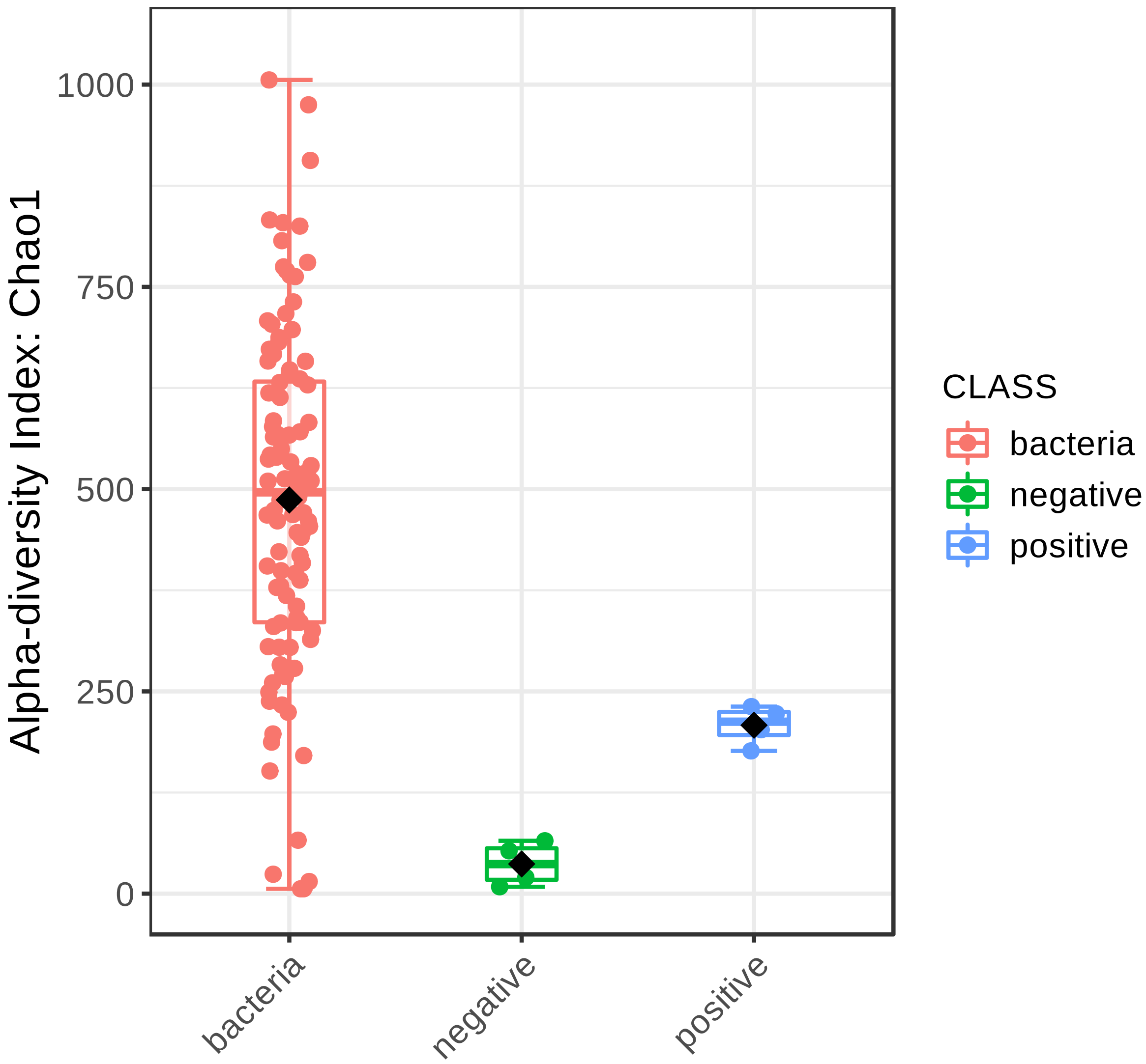


Figure 5: Alpha-diversity measure using `Chao1` at `OTU` level represented as boxplot. Each boxplot represents the diversity distribution of a group present within `SampleType` class [Statistical significance: `p-value: 1.2874e-05; [ANOVA] F-value: 12.508`]

2.4 Beta diversity Analysis

This method provides a way to compare the diversity or composition between two samples or microbial communities. These methods compare the changes in the presence/absence or abundance of thousands of taxa present in a dataset and summarize these into how 'similar' or 'dissimilar' two samples. Each sample gets compared to every other sample generating a distance or dissimilarity matrix. Two parameters need to be considered when performing beta diversity analysis. The first one is how similarity or distance between sample is measured which includes non-phylogenetic (Bray-Curtis distance, Shannon index, Jaccard index) and phylogenetic-based (weighted and unweighted UniFrac) distances. The other parameter is how to visualize such dissimilarity matrix in lower dimensions. Ordination-based methods such as Principle Coordinate Analysis (PCoA) and non-metric multidimensional scaling (NMDS) are used to visualize these matrix in 2 or 3-D plot where each point represents the entire microbiome of a single sample. Each axis reflects the percent of the variation between the samples with the X-axis representing the highest dimension of variation and the Y-axis representing the second highest dimension of variation. Further,each point or sample displayed on PCoA or NMDS plots is colored based on either sample group, features alpha diversity measures, or the abundance levels of a specific feature.

Also, the statistical significance of the clustering pattern in ordination plots can be evaluated using anyone among Permutational ANOVA (PERMANOVA), Analysis of group Similarities (ANOSIM) and Homogeneity of Group Dispersions (PERMDISP).

Beta diversity analysis is performed using the `phyloseq` package[4]. Figure 6 shows the ordination plot represented in 2-D; Statistical significance is found out using `[PERMANOVA] F-value: 4.8937; R-squared: 0.082394; p-value < 0.001` .

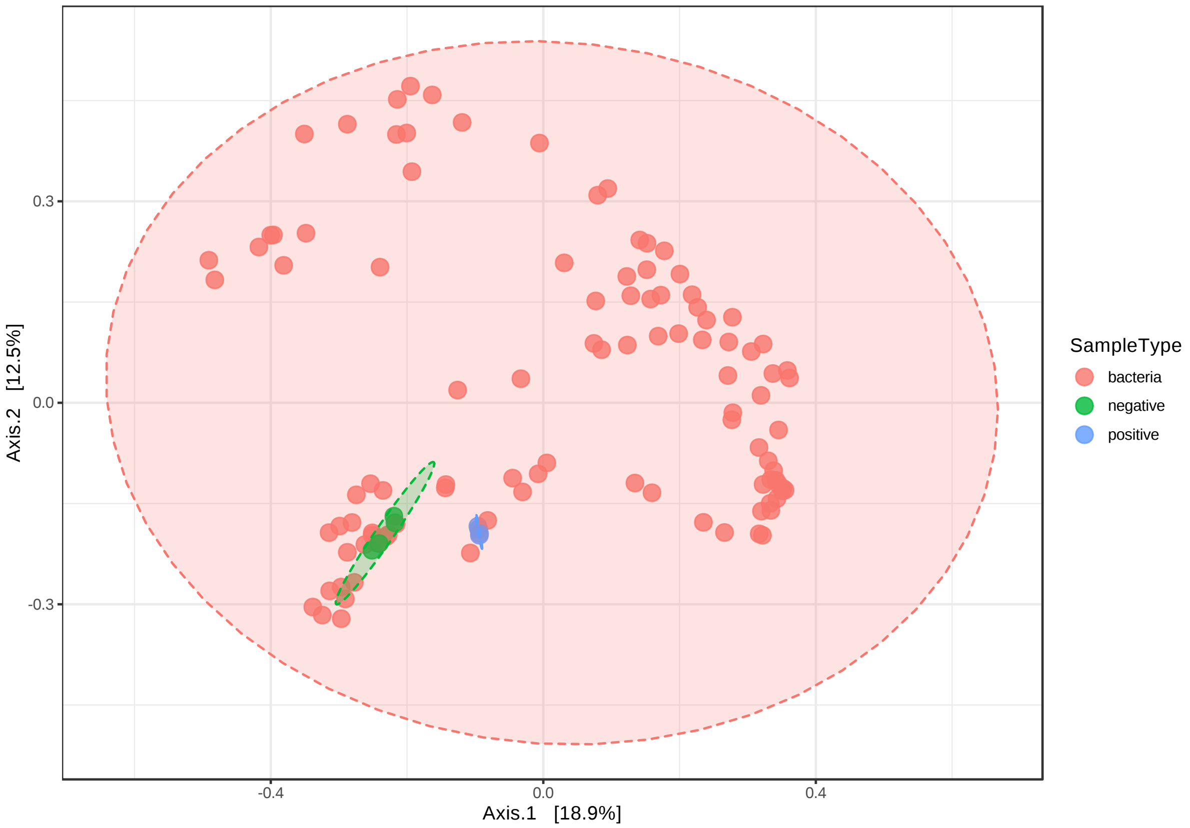


Figure 6: 2-D `PCoA` plot using `bray` distance. The explained variances are shown in brackets.

[4]Paul J. McMurdie *phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data.*, 2013, R package version 1.19

The report was generated on Fri Feb 14 18:47:46 2020 with R version 3.6.1 (2019-07-05).