CD Genomics
The Genomics Services Company

16S rRNA Sequencing Report

Customer: XXXXXXXXXXX

Date: XX/XX/XXXX

Table of Contents

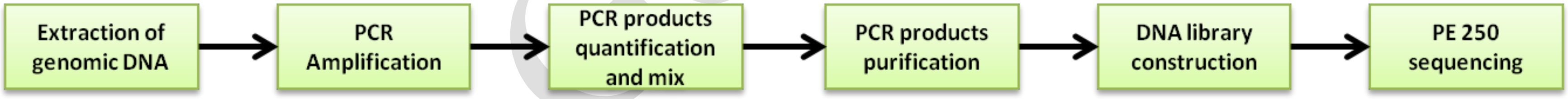
1. Background

Sequence variation in the 16S ribosomal RNA (rRNA) gene is widely used to characterize taxonomic diversity presenting in microbial communities. The 16S sequence is composed of nine hypervariable regions interspersed with conserved regions. The sequence of the 16S rRNA gene and its hypervariable regions have been determined for a large number of organisms, and are available to download from multiple databases such as Greengenes, Silva and the Ribosomal Database Project (RDP). For taxonomic classification, it is sufficient to sequence individual hypervariable regions instead of the entire gene length.

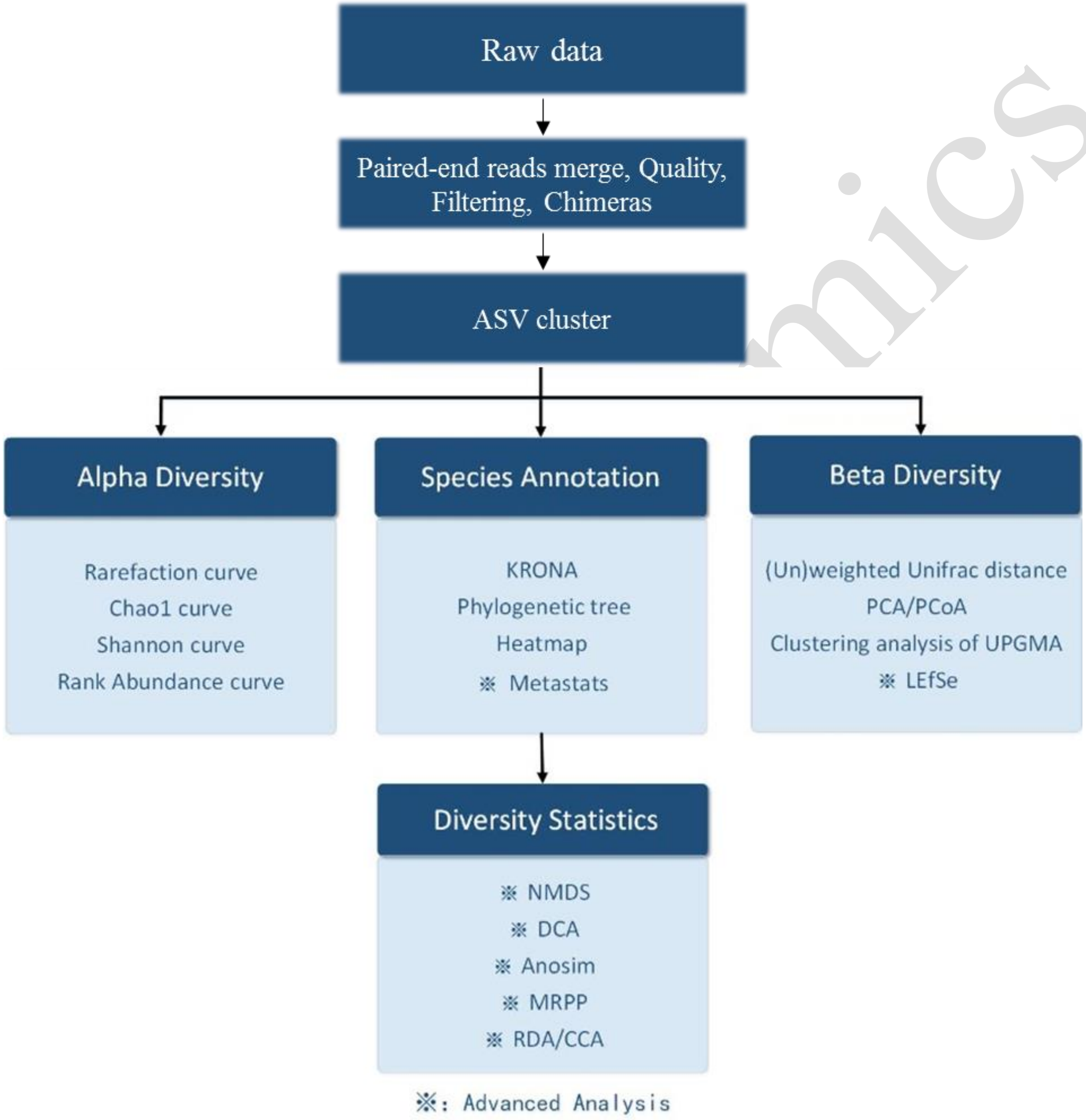
Microbial diversity is assessed by sequencing using a paired-end method on the Illumina platform to construct a small fragment library for sequencing. The microbial composition of a sample can be revealed by reads merging, ASV (Amplicon Sequence Variant), species annotation and abundance analysis. Furthermore, the Alpha diversity, Beta diversity analysis, and significant difference analysis can be used to explore the differences between samples.

2. Workflow

2.1 Experiment Process and Sequencing


Extraction of genomic DNA
PCR Amplification
PCR products quantification and mix
PCR products purification
DNA library construction
PE 250 sequencing

2.2 Bioinformation Analysis Process


Raw data
Paired-end reads merge, Quality, Filtering, Chimeras
ASV cluster
Alpha Diversity
Rarefaction curve
Chao1 curve
Shannon curve
Rank Abundance curve
Species Annotation
KRONA
Phylogenetic tree
Heatmap
※ Metastats
Beta Diversity
(Un)weighted Unifrac distance
PCA/PCoA
Clustering analysis of UPGMA
※ LEfSe
Diversity Statistics
※ NMDS
※ DCA
※ Anosim
※ MRPP
※ RDA/CCA
※: Advanced Analysis

2.3 Sample Information

The table below shows the samples ID and their corresponding information.

Table 1 Sample ID and the related information (Top 10)

Sample-id	group	C_vs_T	Name
S145_1	LST1	T	LST
S145_2	LST1	T	LST
S146_1	LET1	T	LET
S146_2	LET1	T	LET
S147_1	LSC1	C	LSC
S147_2	LSC1	C	LSC
S148_1	LEC1	C	LEC
S148_2	LEC1	C	LEC
S149_1	TST1	T	TST
S149_2	TST1	T	TST

3. Results

3.1 Data Processing and Statistics

Amplicons were performed on a paired-end Illumina MiSeq platform to generate 300 bp paired-end raw reads, and then pretreated. Specific processing steps are as follows：

1) Paired-end reads were assigned to a sample by unique barcode. The barcodes and primer sequences were then truncated.

2) Paired-end reads were merged using FLASH (V1.2.11,http://ccb.jhu.edu/software/FLASH/), a very fast and accurate analysis tool to merge pairs of reads when the original DNA fragments are shorter than twice the length of the read. The obtained splicing sequences are referred to as raw tags.

3) Quality filtering was then performed on the raw tags according to the Fastp quality control process. After filtering, high-quality clean tags were obtained. The data output of the above steps is shown in Table 2.

Table 2 Data statistics of the quality control (Top 10)

Sample Name	Raw_Reads(nt)	Clean Tags (nt)	Avglen (nt)	GC Conter (%)	Q20 (%)	Q30 (%)
S1	228272	98032	252	50.7	99.23	97.11
S2	210990	91908	252	50.64	99.26	97.18
S3	196918	85044	252	50.49	99.3	97.32
S4	182560	79198	252	51.21	99.24	97.11
S5	174606	75278	252	54.84	99.24	97.11
S6	168318	70309	252	55.61	99.22	97.03
S7	165110	70902	252	50.27	99.21	97.04
S8	159294	69376	252	49.1	99.18	96.96
S9	153876	64221	252	55.16	99.26	97.17
S10	148624	61667	252	56.79	99.27	97.12

PE Reads: the PE reads obtained from the sequencing platform. Clean Tags: tags following QC. AvgLen: the average length of the Clean Tags. GC (%): the percentage of G and C. Q20%&Q30%: the percentage of bases with a quality score equal to or higher than 20 (error rate <1%) and 30 (error rate <0.1%). Clean(%): the number of Clean Tags take up of the number of Raw PE.

The number of reads within a given length range was counted and displayed in length distribution graphs. The effective tags length distribution for C1_1_M is shown in below Figure 1. The distribution graphs for other samples can be found in the results folder.

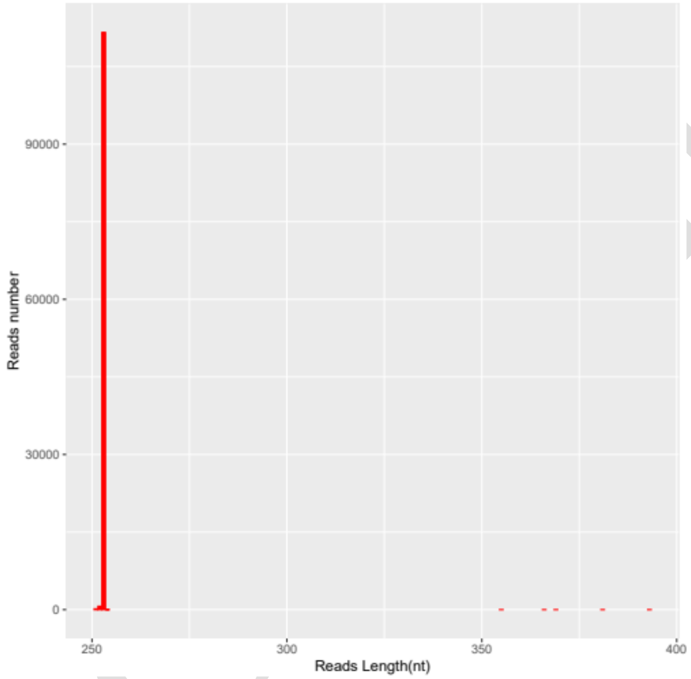


Figure 1. Effective tags length distribution of sample ID C1_1_M.

3.2 Feature table construction

QIIME 2[1] plugins are available for several quality control methods, including DADA 2[2] and Deblur. The feature table is the equivalent of the QIIME 1 ASV or BIOM table, and the QIIME 2 artifact is the equivalent of the QIIME 1 representative sequences file. Because the ASVs resulting from DADA2 and Deblur are created by grouping unique sequences, these are the equivalent of 100% ASVs from QIIME 1 and are generally referred to as sequence variants. In QIIME 2, these ASVs are higher resolution than the QIIME 1 default of 97% ASVs.

QIIME 2 ASVs also use more rigorous quality control steps than QIIME 1 ASVs, resulting in higher overall quality and more accurate estimates of both diversity and taxonomic composition than was achieved with QIIME.

File path： result\ 02.ASV_table\ ASV_table.tsv

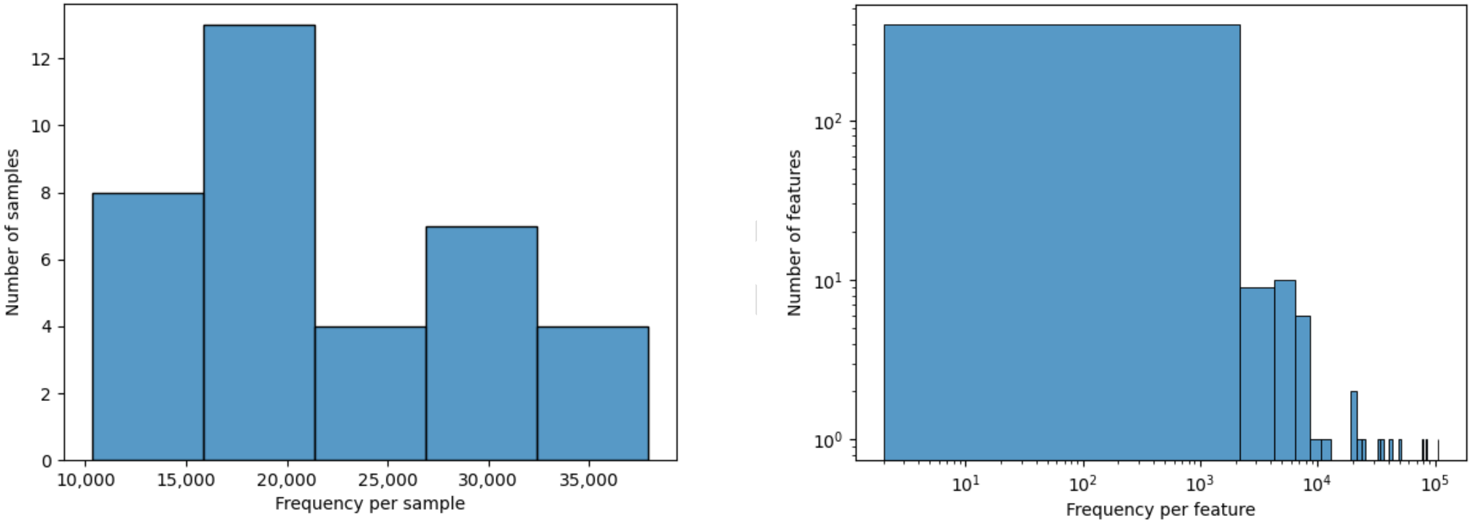


Figure 2. (A)The frequency of each sample and (B) the frequency of each feature.

3.3 Species Annotation and Taxonomic Analysis

In the next sections we will begin to explore the taxonomic composition of the samples and compare samples to the metadata. The first step in this process is to assign taxonomy to the sequences in our QIIME 2 artifact using a pre-trained Naive Bayes classifier and the plugin. This classifier was trained on the Silva 138 99% ASVs. We will apply this classifier to sample sequences and generate a visualization of the resulting mapping from sequence to taxonomy.

3.3.1 Taxonomy Distribution Histogram of All Samples

The taxonomy distributions histogram graph at the phylum classification level are displayed in the below Figure 3. Each color represents a taxonomy, and the length of the color blocks indicates relative abundance of the taxonomy. In order to display the best view, the histogram shows only the abundance of the top ten taxa, less abundant taxa are combined into 'Others' category. 'Unknown' indicates taxa that have not been annotated. Specific species information can be found in the corresponding species abundance table.

File path: result\ 03.Taxonomy\ 03.taxonomy.tsv

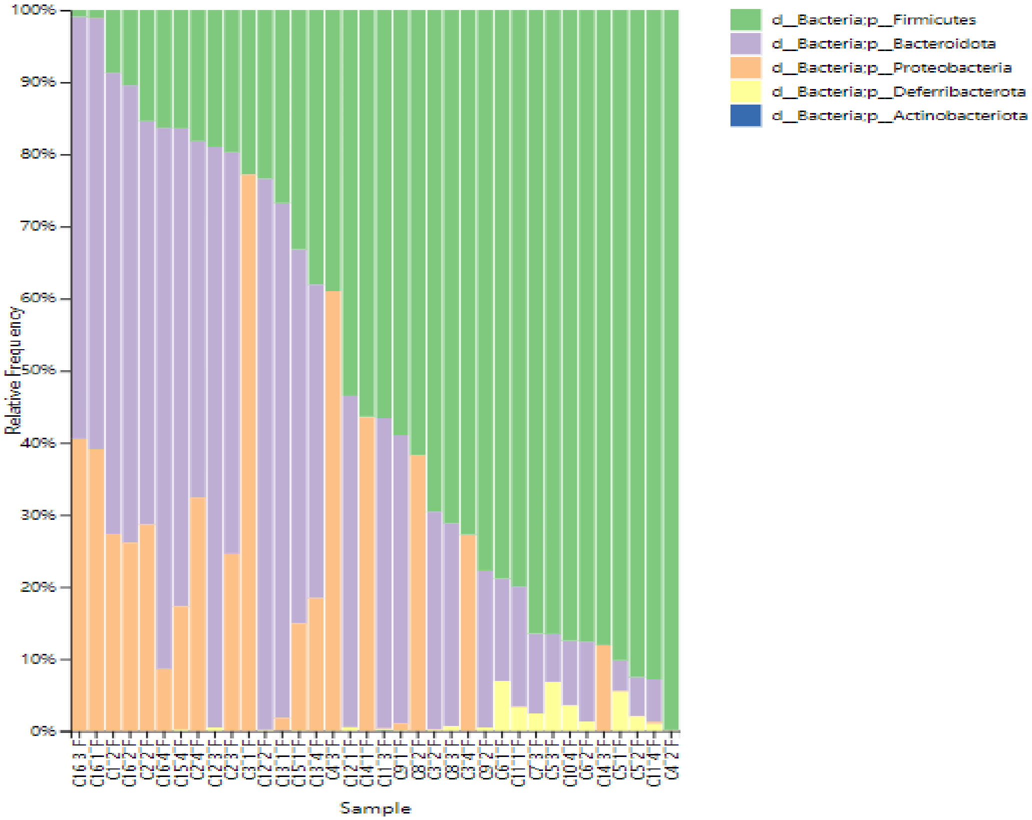
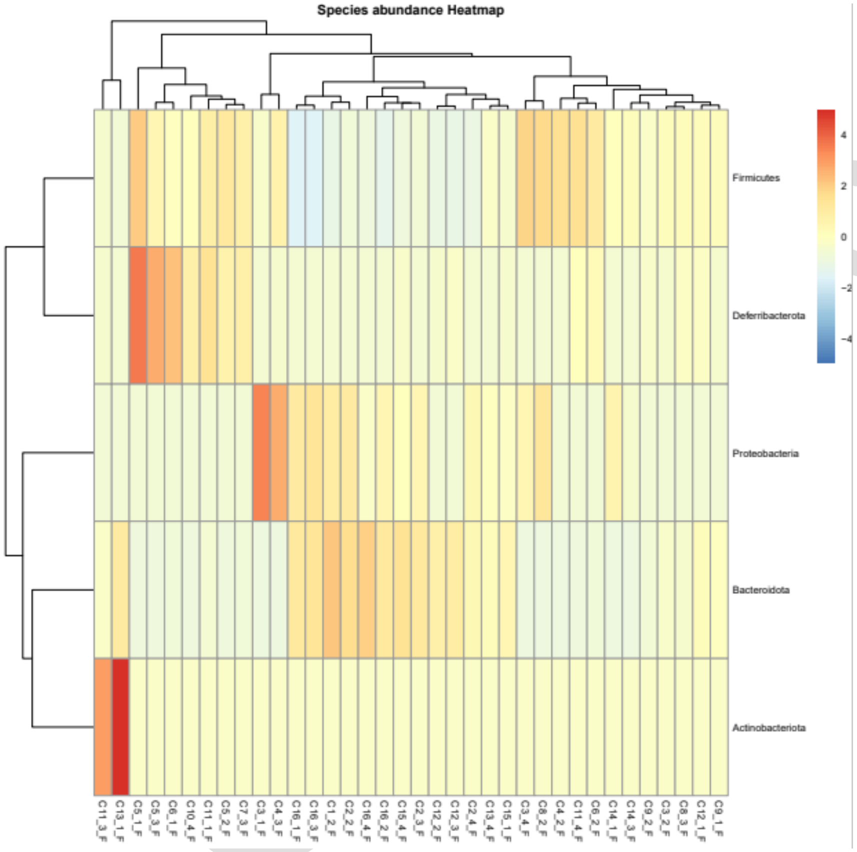


Figure 3. The taxonomy distribution of all sample in Phylum classification level. Other classification levels can be found in the taxonomy folder.

3.3.2 Species Abundance Heatmap

A Heatmap is a graphical representation of clustering using color gradients to represent the relative abundance of similar species in a sample. According to the taxonomic composition and relative abundance of each sample, heatmap analyses were carried out at each taxonomic level (phylum, class, order, family, genus and species respectively) and plotted using R language tools. In the heatmap clustering results, color represents the abundance of species, and vertical clustering indicates the similarity of the abundance between different species. A shorter distance between the two species and a shorter branch length indicates that the two species have a more similar abundance between the samples. The horizon clustering indicates the similarity of the abundance of different species between samples. As with the vertical clustering, the shorter distance and branch length between the samples indicates the more similarity of abundance. The heatmap at phylum level is illustrated in Figure 4.



Phylum

Figure 4. Species abundance Heatmap. Phylum. Plotted by sample name on the X-axis. The Y-axis represents the genus. The absolute value of the legend represents the distance between the raw score and the mean population of the standard deviation. The legend is negative when the raw score is below the mean.

3.3.3 Classification Tree

The Classification tree is a bifurcating tree that represents a hierarchical clustering of features. The hierarchical clustering uses Ward hierarchical clustering based on the degree of proportionality between features.

3.4 Alpha Diversity Analysis

Microbial diversity can be assessed within a community (alpha diversity) or between the collections of samples (beta diversity). Four different metrics were calculated to assess the alpha diversity: Chao1 and Ace simply estimate the number of species in a community; Shannon and Simpson account for both richness and evenness of a community. Larger the Chao1, Ace and Shannon indices correspond to a smaller Simpson index value, indicating greater diversity of species [3]. In addition, the coverage of the sample library is reported. A higher value indicates a higher probability that the sequence is detected in the sample. The index reflects whether the results of this sequencing accurately represent the real population of microbes in the sample.

3.4.1 Statistical Data of Alpha Diversity

In order to compare the diversity indices between the samples, we have standardized the sequence number in each sample in the analysis process. At the level of 97% similarity, varied alpha metrics results were integrated and displayed on the following Table 3.

Table 3. Statistics of Alpha diversity indices (Top 10)

Sample	Observed species	ace	Chao 1	Simpson	Shannon
A1	910	910	910	0.654985035	3.914181389
A2	1891	1891	1891	0.830694297	5.83462709
A3	158	158	158	0.961684233	5.833334444
A4	87	87	87	0.55398131	1.616963214
A5	100	100	100	0.529850969	1.550806285
A6	802	802	802	0.594395889	2.948466552
A7	1141	1141	1141	0.789908199	4.70196164
A8	233	233	233	0.46312489	2.44111006
A9	3312	3312	3312	0.998912511	10.74202729
A10	199	199	199	0.538326826	1.745920168

3.4.2 Rarefaction Curve

Rarefaction curve [4] is created by random selection of a certain amount of sequencing data from the samples, then counting the number of the species these data represent. The left-side of the steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individual samples have been taken, suggesting that more intensive sampling is likely to yield only few additional species. The rarefaction curve can be used to judge the sequencing sufficiency of each sample. A sharp rise of the curve indicates that sequencing quantity is insufficient, and more reads are required.

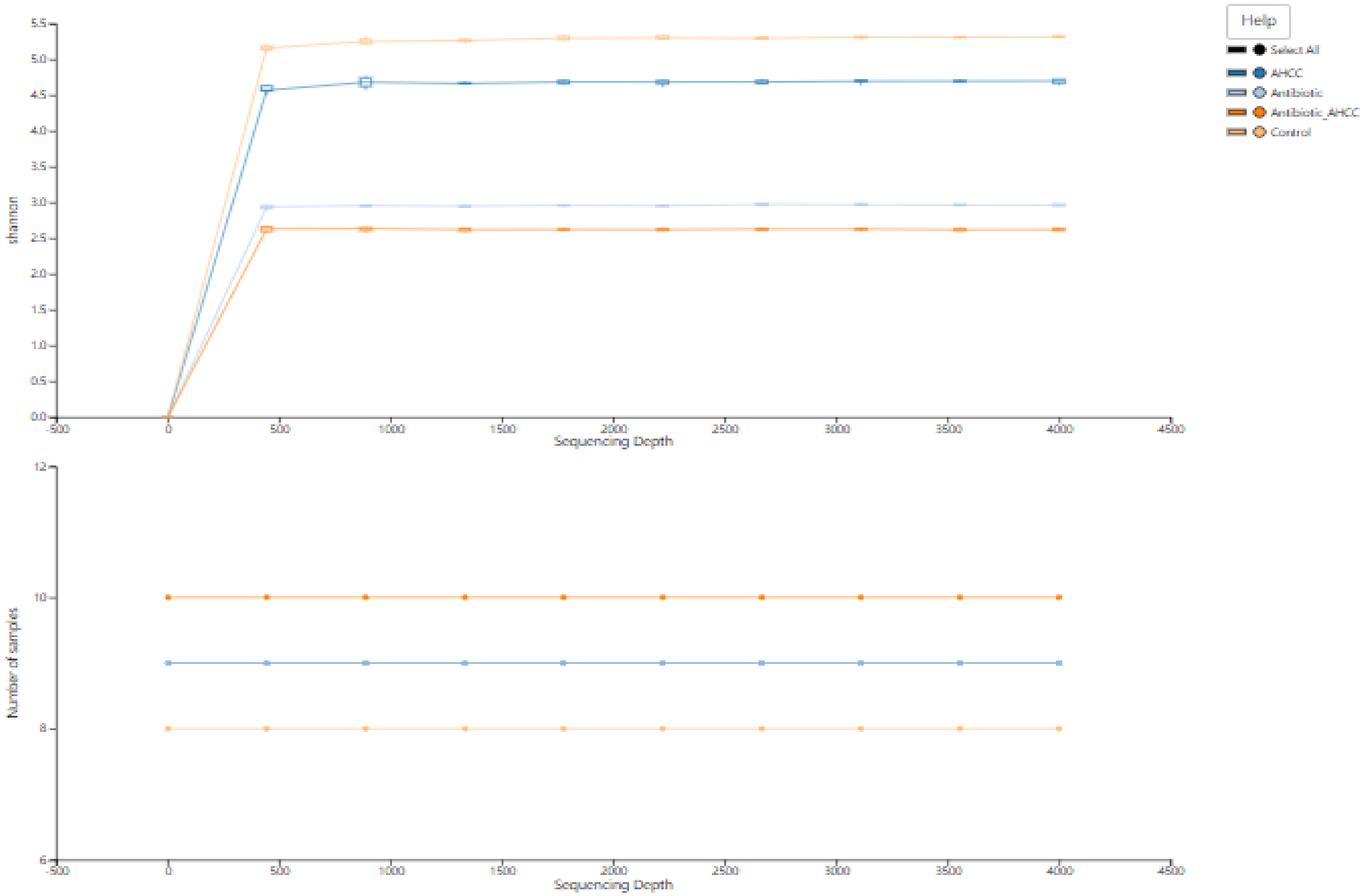


Figure 5. Rarefaction curve of the sequenced reads for all samples (The above figure) & The depth of the sequencing samples (The below figure).

3.5 Beta Diversity Analysis

Beta diversity represents the explicit comparison of microbial communities based on their composition. Beta diversity metrics therefore assess the differences between microbial communities. To compare microbial communities between every pair of community samples, a square matrix of distance was calculated, reflecting the dissimilarity between certain samples. The data in this distance matrix can be visualized with analyses such as Boxplot Analysis, Principal Coordinate Analysis (PCoA), hierarchical clustering, and so on.

Beta diversity analysis mainly uses four algorithms, binary jaccard, bray curtis, weighted unifrac (limited to bacteria), and unweighted unifrac (limited to bacteria), to calculate the distance between samples to obtain the β value between samples. These four algorithms can be divided into two categories: weighted (Bray-Curtis and Weighted Unifrac) and unweighted (Jaccard and Unweighted Unifrac) [5]. The use of unweighted methods is mainly to compare the presence or absence of species. A smaller β diversity between two groups indicates greater similarity in their relative species composition. Weighted methods consider both qualitative data (the presence or absence of species) and quantitative data about the relative abundance of species.

The metrics can be phylogeny based (the UniFrac metrics) or not (Bray-Curtis and Jaccard). The UniFrac distance take the phylogenetic relatedness of ASVs into account (only for bacteria), while the Bray-Curtis distance considers only the abundance.

Suggestion: In the microbial diversity analysis, the differences in microbial composition between different environments are tremendous, so the unweighted method is usually used for the analysis. However, if we want to study the relationship between the control and experimental treatment group using unweighted analysis, then no significant difference can be observed, and weighted method is recommended. Neither analytical method is inherently “better” or “worse”, but the appropriate method should be chosen for particular research purposes. Four types of Beta diversity analysis using a variety of algorithms have been included to provide you with a comprehensive analysis of the results, and you can choose the most suitable one to explain the biological issues of your project.

3.5.1 Boxplot Analysis

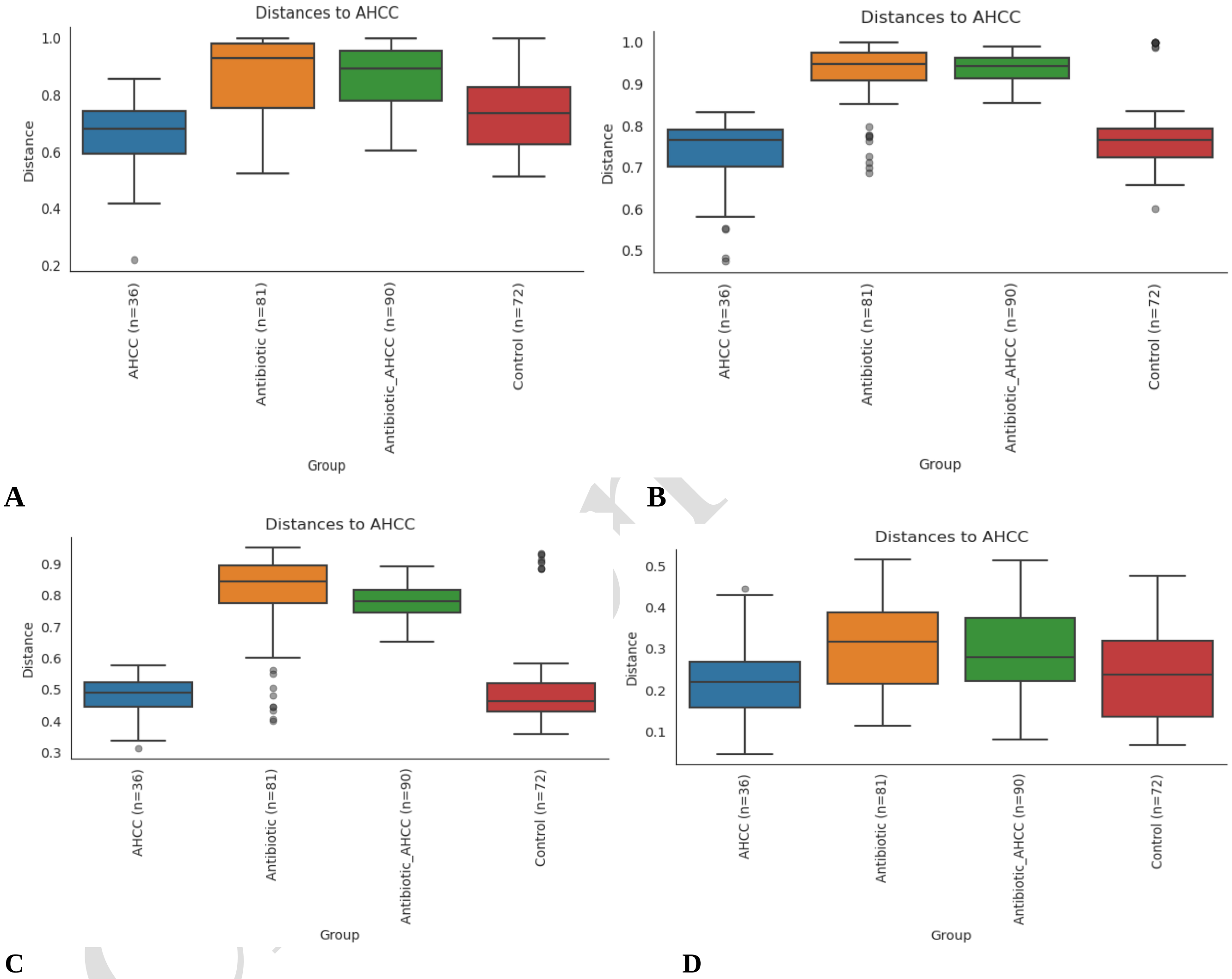


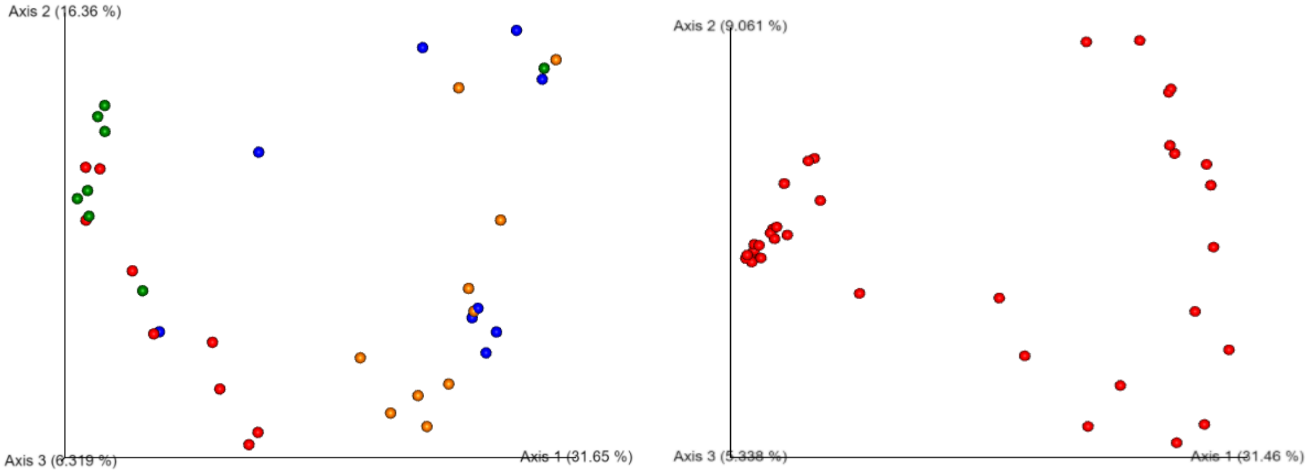
Figure 6. Boxplot analysis based on bray Curtis (A), binary jaccard (B), unweighted unifrac (C), and weighted unifrac (D). The boxplots represent the distribution of predicted functional profiles in the analyzed samples, with the box indicating the interquartile range

and the median line inside. The whiskers extend to the minimum and maximum values within a specified range, providing insight into the variability and differences in functional potentials among the compared sample groups.

3.5.2 PCoA Analysis

Principal coordinates analysis (PCoA) [6] is an ordination technique similar to PCA, which picks up the main elements and structure from reduced multi-dimensional database series of eigenvalues and eigenvectors. It starts with a similarity matrix or dissimilarity matrix (distance matrix) and assigns for each item a location in a low-dimensional space. The technique has advantages over PCA in that each ecological distance can be investigated. PCA finds out the main coordinates based on the similarity coefficient matrix of all samples, while PCoA is based on the distance matrix. Weighted Unifrac and Unweighted Unifrac were calculated to assist the PCoA analysis. By using PCoA we can visualize individual and/or group differences, illustrating the microbial diversity between samples. Based on the four algorithms, principal coordinates analysis was calculated and displayed by QIIME 2 tool, you can view QIIME 2(QIIME 2 View) artifacts and visualizations at view.qiime2.org by uploading files. PcoA results is located at 04.Diversity\Beta\PcoA\ bray_curtis\index.html, and the PcoA result plots can be adjusted according to the link below.

Result link: 04.Diversity\Beta\PCoA\bray_curtis\index.html



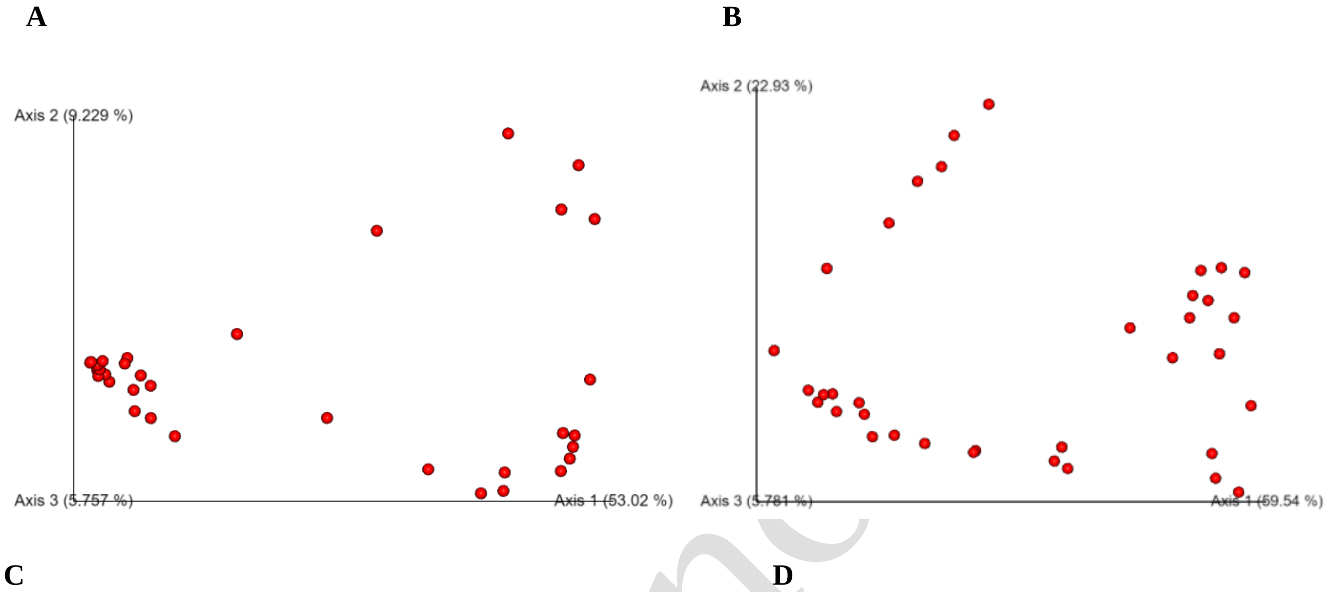


Figure 7. PCoA analysis based on bray Curtis (A), binary jaccard (B), unweighted unifrac (C), and weighted unifrac (D). Each point represents a sample, plotted by a principal component on the X- axis and another principal component on the Y- axis, which was colored by group. The percentage on each axis indicates the contribution value to discrepancy among samples.

3.5.3 UPGMA Analysis

Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is a type of hierarchical clustering method using average linkage. It is widely used in ecology for the classification of samples based on their pairwise similarities in relevant descriptor variables. The basic two ideas of UPGMA are as follows: First, it gathers two samples of the minimum distance together and forms a new node (a new sample), which is branched at the halfway point of the distance between the two samples. Second, it calculates the average distance between a new "sample" and the other samples and can find the minimum distance between two samples in order to cluster both. When all samples are gathered together, a complete clustering tree can be presented.

Based on the four algorithms, hierarchical clustering for samples using UPGMA was performed with the R language tool to assess the similarity of microbial composition between samples. The clustering results are displayed in Figure 8. A closer sample distance and a shorter branch, indicates more similarity in microbial composition between the samples.

The result is located at 04.Diversity\Beta\UPGMA\unweighted\tree.html and then click 'web-based ETE3 tree viewer', then click 'View tree!'. In this way, you can see the tree diagram of this result. The specific link is as follows.

Result link: 04.Diversity\Beta\UPGMA\unweighted\tree.html

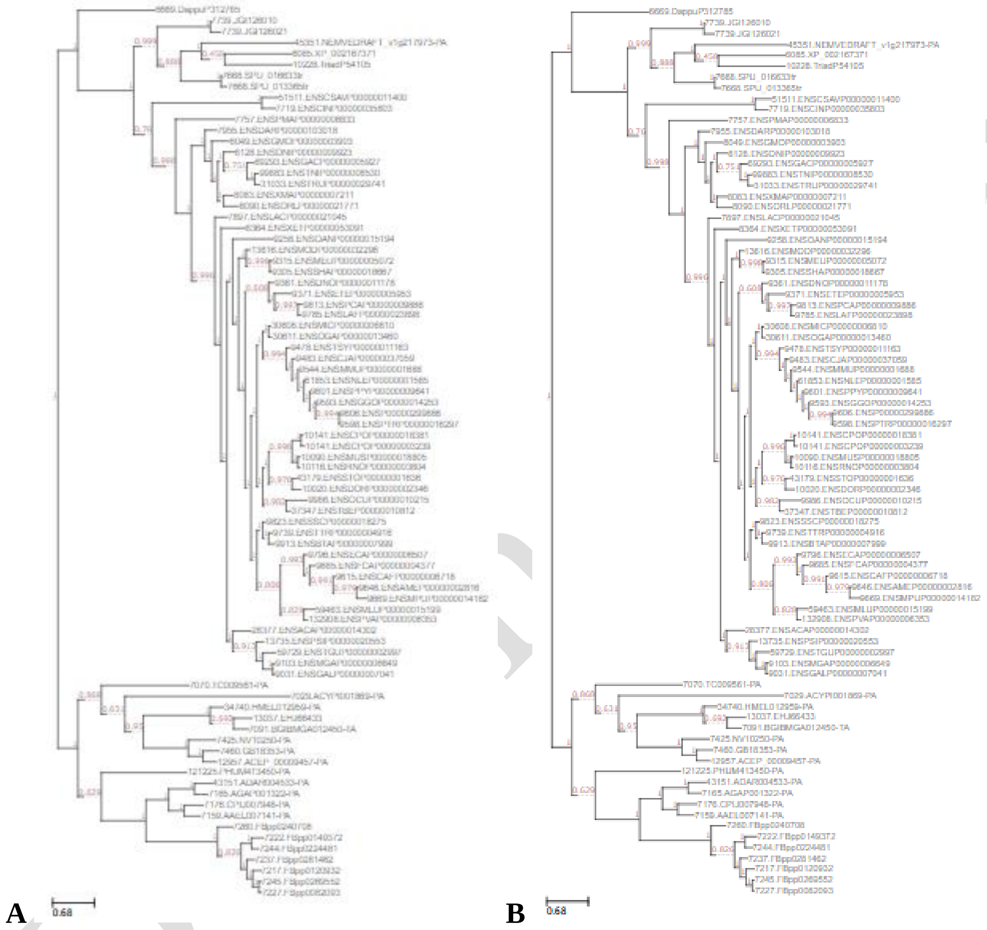


Figure 8. UPGMA clustering tree based on unweighted unifrac (A), and weighted unifrac (B). The different colors represent different grouping.

3.6 Significant Difference Analysis

Analysis of significant differences between groups is mainly used to detect biomarkers that have statistically significant differences between groups. The screening criteria for biomarkers is LDA score > 2. The most common analysis methods include Lefse that is used to screen biomarker and Metastats analysis to discover the significant difference between two groups in multiple taxonomy level through p and q value.

3.6.1 ANCOM Analysis

ANCOM can be applied to identify features that are differentially abundant (i.e., present in different abundances) across sample groups. (The following links take the phylum level as an example).

File path: 05.Composition\condition_1\level2\index.html

Table.4 ANCOM analysis partial result in phylum-level in GROUP

Percentile	0	25	50	75	100	0	25	50	75	100
Group	AHCC	AHCC	AHCC	AHCC	AHCC	Antibiotic	Antibiotic	Antibiotic	Antibiotic	Antibiotic
d__Bacteria;p__Bacteroidota	1007	2247	5634	9116	14790	1	16	4664	16765	24293
d__Bacteria;p__Proteobacteria	1	1	1	18	174	16	6645	7417	10599	25746
d__Bacteria;p__Firmicutes	3488	7413	9790	10722	18251	3285	5717	7584	13450	19714
d__Bacteria;p__Deferribacterota	19	32	91	190	591	1	1	1	1	27

3.6.2 **PicRust Analysis** (tool: PICRUSt2)

PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) is a computational tool developed to infer the functional potential of microbial communities based on 16S rRNA gene sequencing data. The analysis focuses on nderstanding the functional capabilities of the microbial community by predicting the presence and abundance of functional genes, pathways, and biological functions in the sampled environments.

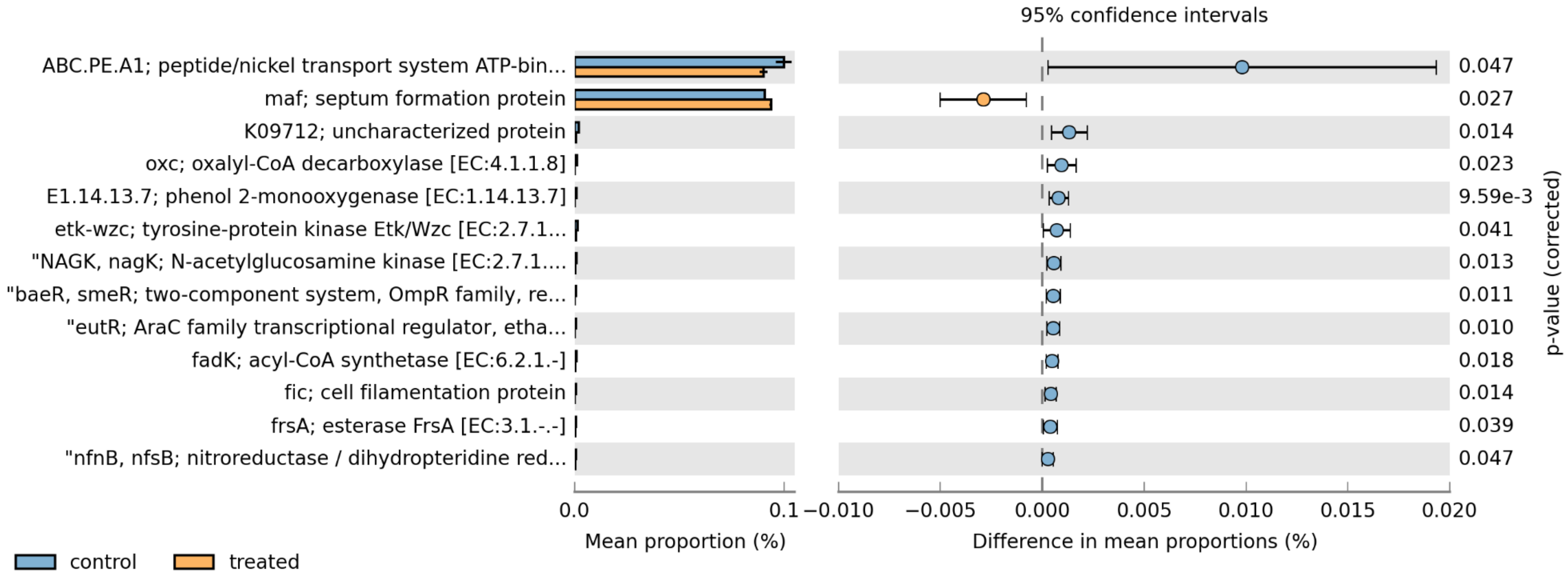


Figure 9. Mean proportion of treated and control group.

3.6.3 LEfSe Analysis

LEfSe (Linear discriminant analysis Effect Size) is a powerful statistical method used in microbiome research to identify differentially abundant features, such as microbial taxa or functional genes, that significantly differentiate between two or more biological conditions or sample groups. It is particularly valuable for exploring the associations between microbial communities and specific metadata, such as disease states, environmental conditions, or other experimental factors.

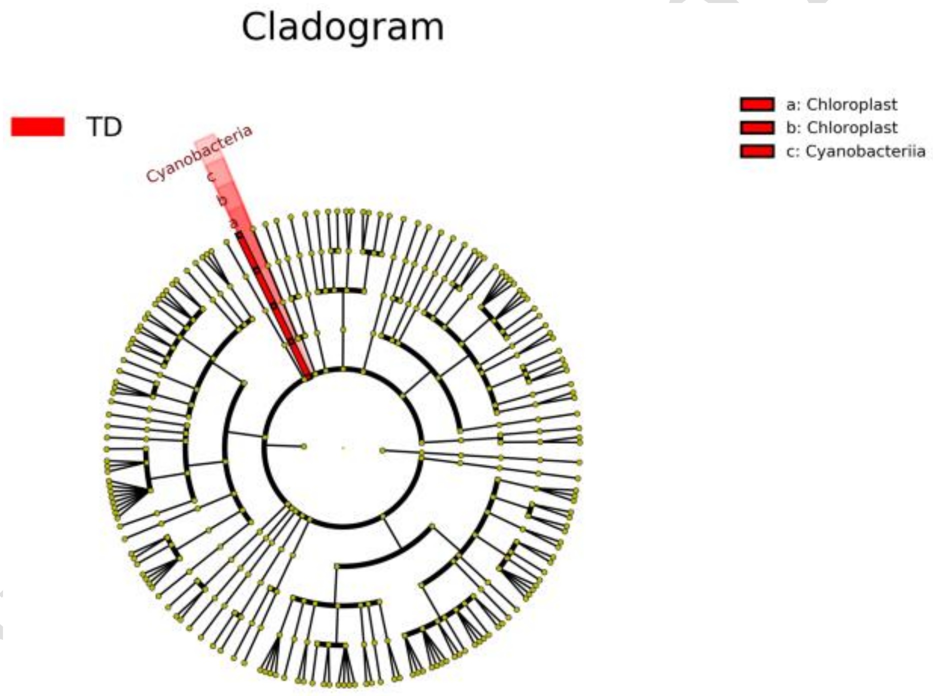


Figure 10. Cladogram. Different circles represent different taxonomic levels, from inside to outside, they are kingdom-phylum-class-order-family-genus-species. Each node represents a species, the more nodes larger means that the abundance of the species is higher, and yellow means that the species has no significant difference between the two groups.

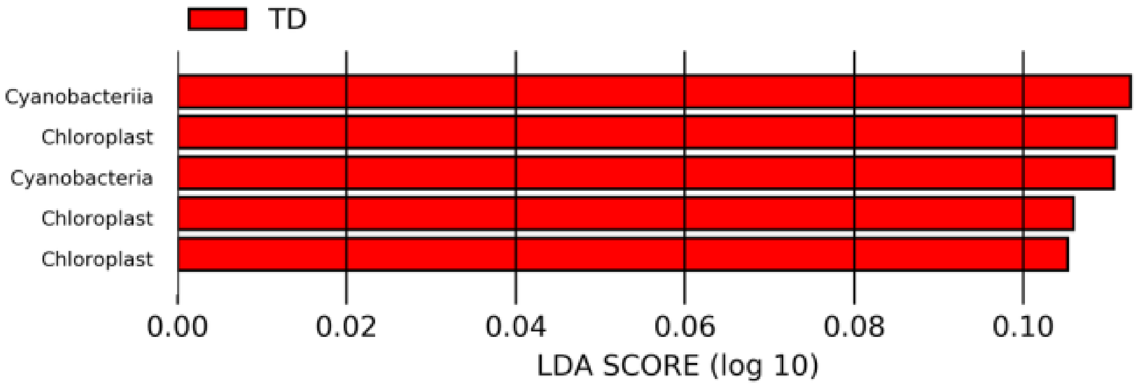


Figure 11. LDA SCORE. The distribution histogram mainly shows us the significantly different species whose LDA score is greater than the preset value, that is, the Biomaker with statistical difference. The default value is 2.0 (0.05 for this item); the color of the histogram represents the enriched group of each differential species, and the length of the column represents the size of the LDA score, that is, the degree of influence of the significantly different species between different groups.

4. Analysis software/ Database information:

Software / Database	Source
Silva	https://www.arb-silva.de/aligner/
UNITE	https://unite.ut.ee/
STAMP	https://github.com/donovan-h-parks/STAMP
PicRust	https://picrust.github.io/picrust/
LEfSe	https://huttenhower.sph.harvard.edu/lefse/
QIIME2	http://qiime.org/

5. Reference:

1. Hall M, Beiko RG. 16S rRNA Gene Analysis with QIIME2. Methods Mol Biol. 2018;1849:113-129. doi: 10.1007/978-1-4939-8728-3_8. PMID: 30298251.

2. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul;13(7):581-3. doi: 10.1038/nmeth.3869. Epub 2016 May 23. PMID: 27214047; PMCID: PMC4927377.

3. Grice EA, Kong HH, et al. (2009). Topographical and temporal diversity of the human skin microbiome. Science, 324(5931): 1190–1192.

4. Wang Y, Sheng H-F, He Y, Wu J-Y, Jiang Y-X, Tam NF-Y, Zhou H-W: Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. Applied and environmental microbiology 2012, 78(23):8264-8271.

5. Lozupone C, Knight R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol, 71 (12): 8228-8235.

6. Sakaki T, Takeshima T, Tominaga M, Hashimoto H, Kawaguchi S: Recurrence of ICA-PCoA aneurysms after neck clipping. Journal of neurosurgery 1994, 80(1):58-63.