16S Amplicon Analysis

Demo Report

May 1st, 2016

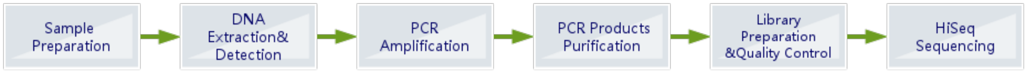
Contents

1. Overview

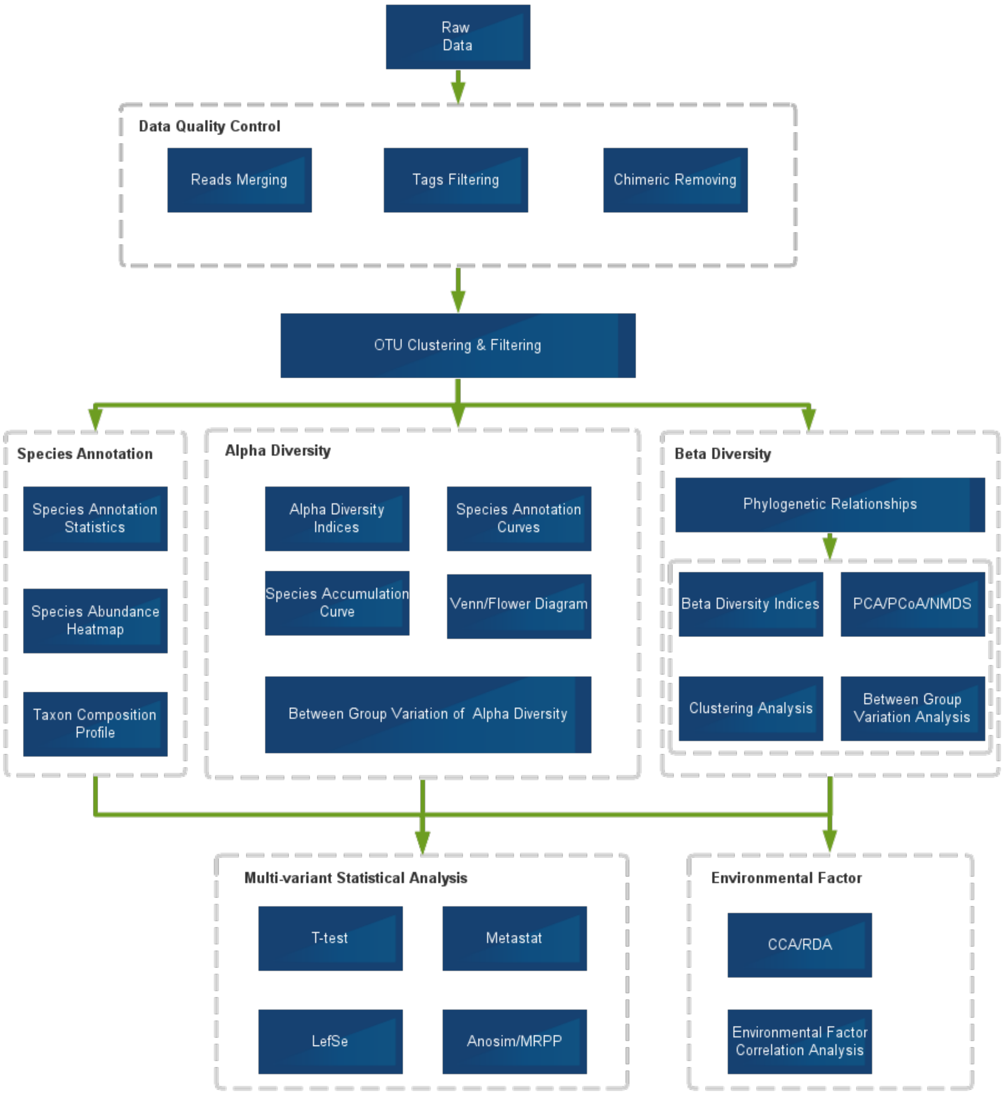
Sequence variation in the 16S ribosomal RNA (rRNA) gene is widely used to characterize taxonomic diversity presenting in microbial communities[1,2,3]. The 16S sequence is composed of nine hypervariable regions interspersed with conserved regions. The sequence of the 16S rRNA gene and its hypervariable regions have been determined for a large number of organisms, and are available to download from multiple databases such as Greengene[4] and the Ribosomal Database Project[5,6]. For taxonomic classification, it is sufficient to sequence individual hypervariable regions instead of the entire gene length.

2. Workflow

2.1 Experiment process and sequencing


Sample Preparation
DNA Extraction& Detection
PCR Amplification
PCR Products Purification
Library Preparation &Quality Control
HiSeq Sequencing

2.2 Information analysis process



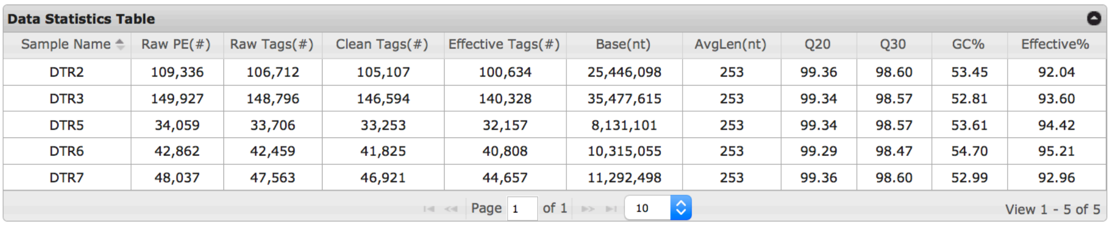
Raw Data
Data Quality Control
Reads Merging
Tags Filtering
Chimeric Removing
OTU Clustering & Filtering
Species Annotation
Species Annotation Statistics
Species Abundance Heatmap
Taxon Composition Profile
Alpha Diversity
Alpha Diversity Indices
Species Annotation Curves
Species Accumulation Curve
Venn/Flower Diagram
Between Group Variation of Alpha Diversity
Beta Diversity
Phylogenetic Relationships
Beta Diversity Indices
PCA/PCoA/NMDS
Clustering Analysis
Between Group Variation Analysis
Multi-variant Statistical Analysis
T-test
Metastat
LefSe
Anosim/MRPP
Environmental Factor
CCA/RDA
Environmental Factor Correlation Analysis

3. Results

3.1 Sequencing data processing

Amplification of certain region of 16S rRNA gene is performed on a paired-end Illumina HiSeq platform to generate 250bp paired-end raw reads (Raw PE).The raw reads are then merged and pretreated to obtain Clean Tags. The chimeric sequences on Clean Tags are detected and removed to obtain the Effective Tags subsequently. The data output of the above steps was shown in table 3.1.

Table 3.1 Statistical table for data pre-processing and quality control



Data Statistics Table

Sample Name	Raw PE(#)	Raw Tags(#)	Clean Tags(#)	Effective Tags(#)	Base(nt)	AvgLen(nt)	Q20	Q30	GC%	Effective%
DTR2	109,336	106,712	105,107	100,634	25,446,098	253	99.36	98.60	53.45	92.04
DTR3	149,927	148,796	146,594	140,328	35,477,615	253	99.34	98.57	52.81	93.60
DTR5	34,059	33,706	33,253	32,157	8,131,101	253	99.34	98.57	53.61	94.42
DTR6	42,862	42,459	41,825	40,808	10,315,055	253	99.29	98.47	54.70	95.21
DTR7	48,037	47,563	46,921	44,657	11,292,498	253	99.36	98.60	52.99	92.96

Results directory

Pair-end raw reads with primers and barcodes (PE reads):

result/00.RawData/Sample_Name/*.raw_1(2).fq.gz

Pair-end raw reads without primers and barcodes (PE reads):

result/00.RawData/Sample_Name/ *_1.fq.gz
result/00.RawData/Sample_Name/*_2.fq.gz

Merged raw tags (Raw Tags):

result/00.RawData/Sample_Name/*.extendedFrags.fastq

Tags with chimeric sequences and low quality sequences removal (Effective Tags):

result/01.CleanData/Sample_Name/*.fastq； result/01.CleanData/Sample_Name/*.fna

List of barcodesequences and primer sequences:

result/00.RawData/SampleSeq_info.xls

3.2 OTU analysis and species annotation

In order to analyze the species diversity within samples, we cluster all Effective Tags to OTUs (Operational Taxonomic Units) at 97% similarity. Then we perform species annotation based on the OTUs representative tags.

3.2.1 Statistics of OTU analysis and species annotation

During the process of OTUs construction, some basic information of different samples, such as Effective Tags number, low-frequency Tags number, Tags annotation info. etc. have all been collected below. The statistical dataset is showed as follows in Figure 3.2.1-1.

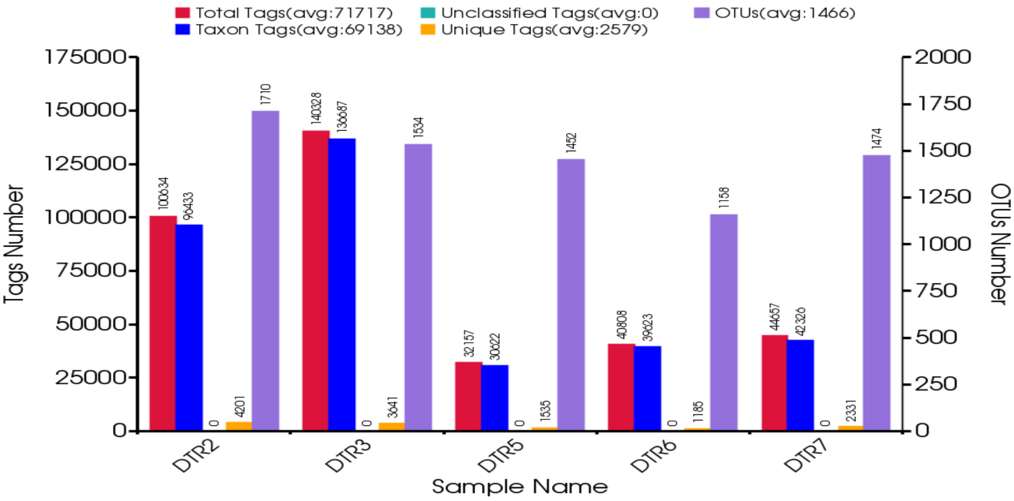


Figure 3.2.1-1 Statistic analysis of the tags and OTUs number of each samples.

The Y1-axis titled "Tags Number" means the numbers of tags: "Total tags"(Red bars) means the numbers of effective tags; "Taxon Tags" (Orange bars) means the numbers of annotated tags; "Unclassified Tags" (Orange bars) means the numbers of unannotated tags; "Unique Tags" means the numbers of tags with a frequency of 1 and only occurs in one sample. The Y2-axis titled "OTUs Numbers" means the numbers of OTUs displayed as "OTUs" (Purple bars) in the above picture to identify the numbers of OTUs in different samples.

According to the results of species annotation, the statistics of sequence number in different classification levels (Kingdom, Phylum, Class, Order, Family, Genus, Species) are calculated and displayed in Figure 3.2.1-2. Sample composition of each sample and differences among samples could be easily understood through the following picture.

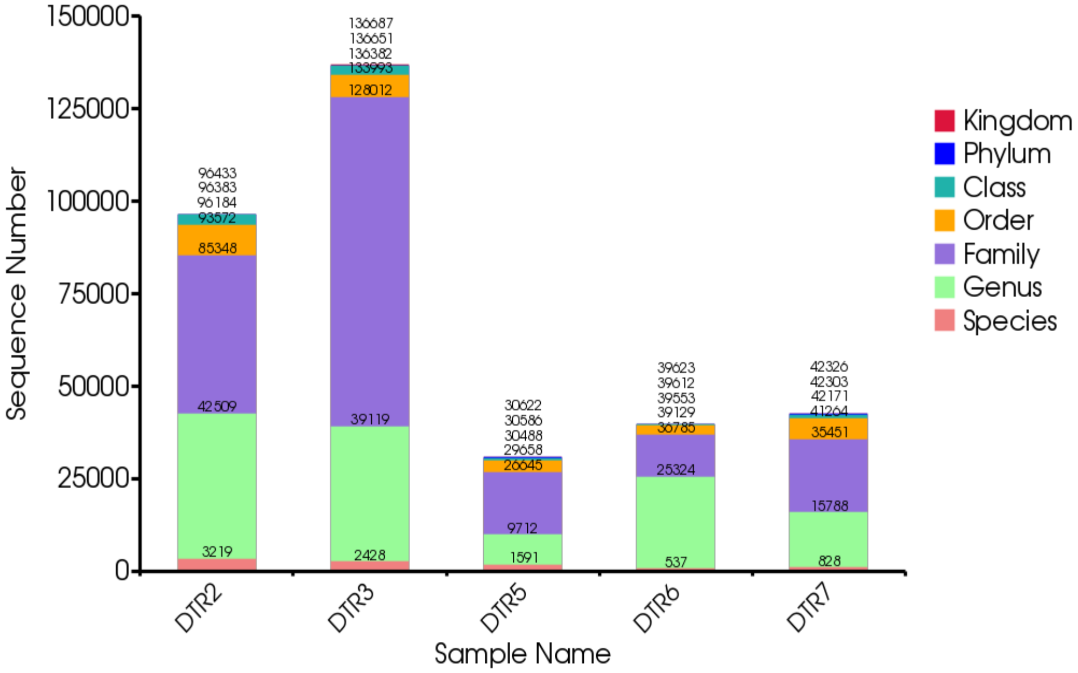


Figure 3.2.1-2 Tags abundance of different levels.

Plotted by the tags number of each classification level on the Y- axis and Samples Name on the X- axis.

Results directory

Tags distribution and OTU analysis layout: result/02.OTUanalysis/taxa_stat/Sample_Tags-OTUs_dis.{png,svg}

Tags and OTUs number statistics table: result/02.OTUanalysis/taxa_stat/Tags_stat.xls

Tags abundance layout in different levels: result/02.OTUanalysis/taxa_stat/Classified_stat.{png,svg}

Profiling taxonomy statistics table for each level: result/02.OTUanalysis/taxa_stat/classified_stat.xls

3.2.2 Species distribution

3.2.2.1 Species relative abundance layout

The top ten species in the classification level of phylum were selected, and the distribution histogram of relative abundance of species was formed as follows in Figure 3.2.2.1.

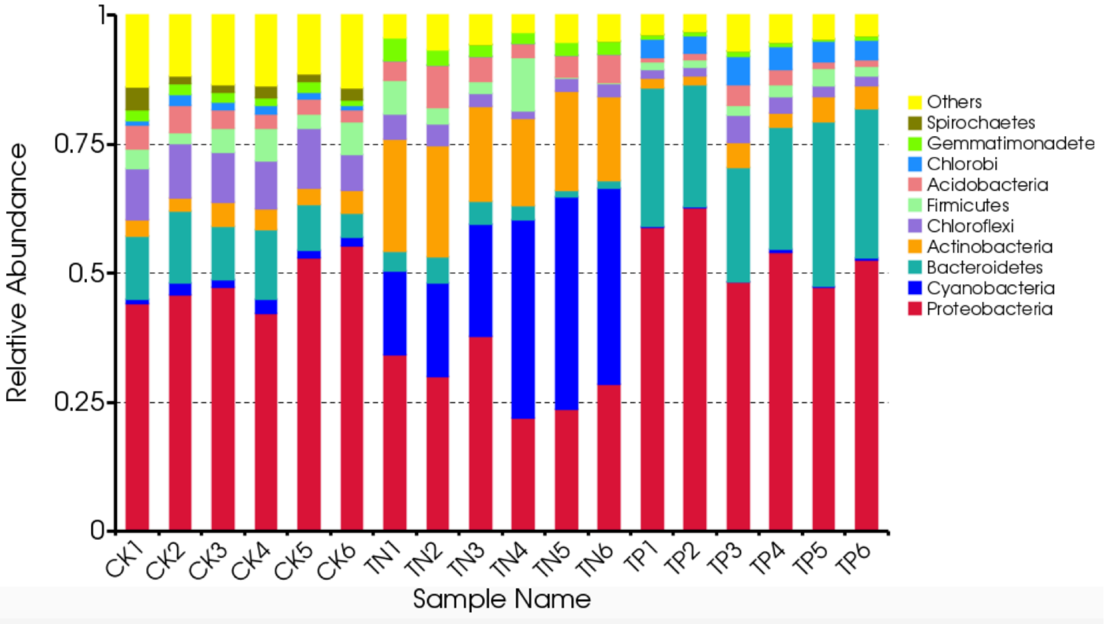


Figure 3.2.2.1 Relative abundance distribution of top 10 phyla.

Plotted by the "Relative Abundance" on the Y-axis and "Samples Name" on the X-axis. "Others" represents a total relative abundance of the rest phylum besides the top 10 phyla.

Results directory

Top 10 species abundance layout at each taxonomic level(phylum, class, order, family, genus): result/02.OTUanalysis/top10/.

3.2.2.2 Species abundance heatmap

The abundance distribution of dominant 35 genera among all samples was displayed in the Species abundance heatmap. Based on the information of clustering results of samples and taxa as well, we could check whether the samples with similar processing are clustered or not, and the similarity and difference of samples can also be observed. The result is shown in Figure 3.2.2.2.

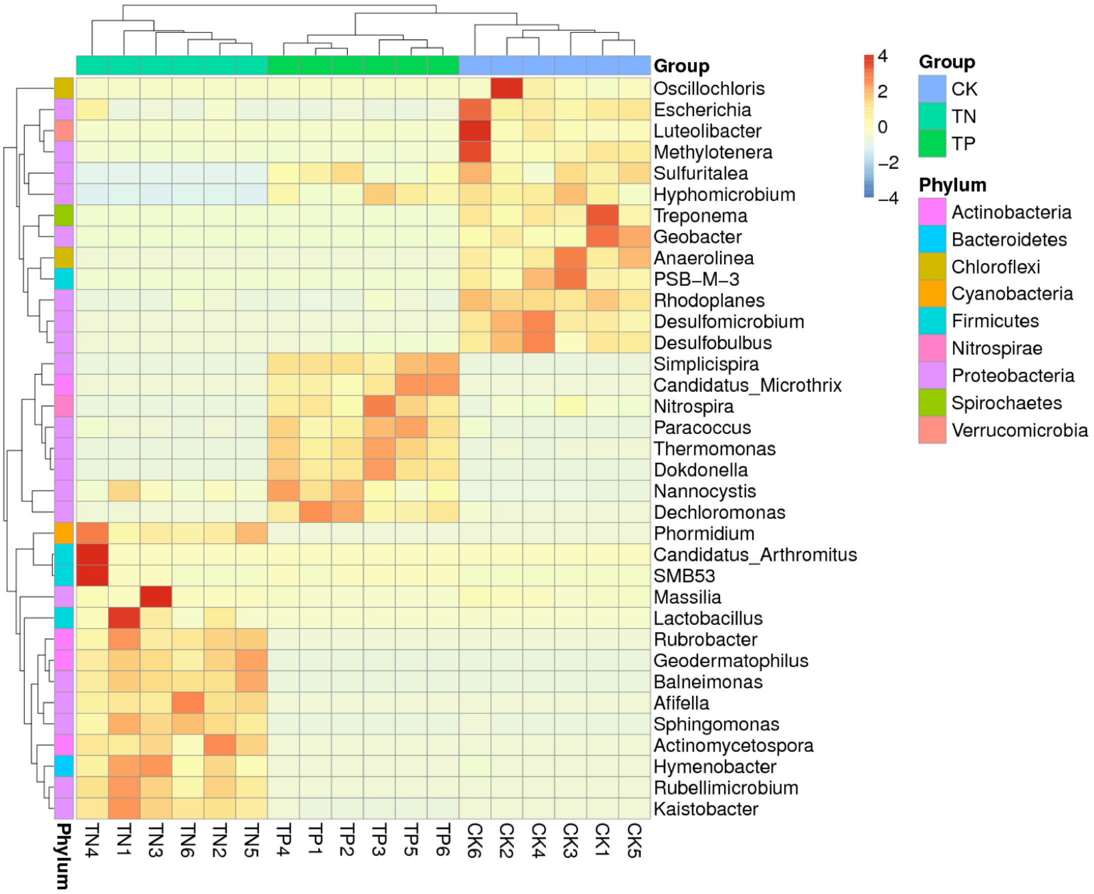


Figure 3.2.2.2 Species abundance heatmap.

Plotted by sample name on the X-axis and different genera on the Y-axis. The absolute value of 'z' represents the distance between the raw score and the mean population of the standard deviation. 'Z' is negative when the raw score is below the mean, and vice versa.

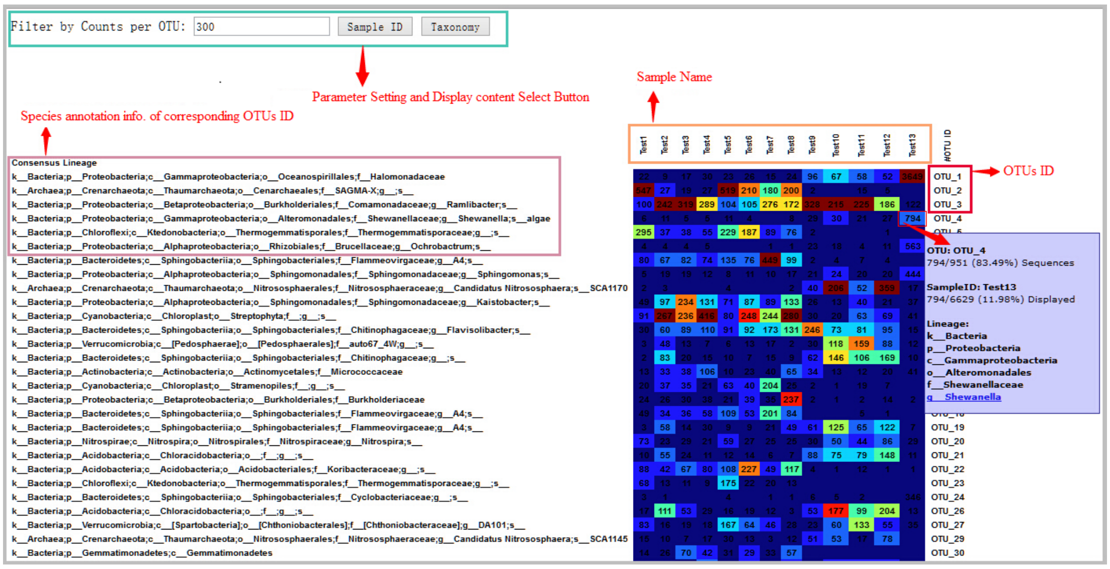
Results directory

Species abundance heatmap of different levels (p, c, o, f, g): result/02.OTUanalysis/phylo_tree/OTU.cluster.tree.{png,svg}

Profiling statistics table for each level: result/02.OTUanalysis/taxa_heatmap/cluster/*.txt

3.2.2.3 OTUs heatmap

We produce heatmap to achieve a interactive view of species composition and abundance among different samples by flexible web display . An example picture is as follows:



Consensus Lineage
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Halomonadaceae
k__Archaea;p__Crenarchaeota;c__Thaumarchaeota;o__Cenarchaeales;f__SAGMA-X;g__;s__
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Ramlibacter;s__
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Alteromonadales;f__Shewanellaceae;g__Shewanella;s__algae
k__Bacteria;p__Chloroflexi;c__Ktedonobacteria;o__Thermogemmatisporales;f__Thermogemmatisporaceae;g__;s__
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Brucellaceae;g__Ochrobactrum;s__
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Flammeovirgaceae;g__A4;s__
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__
k__Archaea;p__Crenarchaeota;c__Thaumarchaeota;o__Nitrososphaerales;f__Nitrososphaeraceae;g__Candidatus Nitrososphaera;s__SCA1170
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Kaistobacter;s__
k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__;g__;s__
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Chitinophagaceae;g__Flavisolibacter;s__
k__Bacteria;p__Verrucomicrobia;c__[Pedosphaerae];o__[Pedosphaerales];f__auto67_4W;g__;s__
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Chitinophagaceae;g__;s__
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae
k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Stramenopiles;f__;g__;s__
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Flammeovirgaceae;g__A4;s__
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Flammeovirgaceae;g__A4;s__
k__Bacteria;p__Nitrospirae;c__Nitrospira;o__Nitrospirales;f__Nitrospiraceae;g__Nitrospira;s__
k__Bacteria;p__Acidobacteria;c__Chloracidobacteria;o__;f__;g__;s__
k__Bacteria;p__Acidobacteria;c__Acidobacteria;o__Acidobacteriales;f__Koribacteraceae;g__;s__
k__Bacteria;p__Chloroflexi;c__Ktedonobacteria;o__Thermogemmatisporales;f__Thermogemmatisporaceae;g__;s__
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Cyclobacteriaceae;g__;s__
k__Bacteria;p__Acidobacteria;c__Chloracidobacteria;o__;f__;g__;s__
k__Bacteria;p__Verrucomicrobia;c__[Spartobacteria];o__[Chthoniobacterales];f__[Chthoniobacteraceae];g__DA101;s__
k__Archaea;p__Crenarchaeota;c__Thaumarchaeota;o__Nitrososphaerales;f__Nitrososphaeraceae;g__Candidatus Nitrososphaera;s__SCA1145
k__Bacteria;p__Gemmatimonadetes;c__Gemmatimonadetes

OTU: OTU_4
794/951 (83.49%) Sequences

SampleID: Test13
794/6629 (11.98%) Displayed

Lineage:
k__Bacteria
p__Proteobacteria
c__Gammaproteobacteria
o__Alteromonadales
f__Shewanellaceae
g__Shewanella

Figure 3.2.2.3 An example of OTU table heatmap.

The counts are colored based on the contribution percentage of each OTU to the total OTU count in one sample (blue: contributes low percentage of OTUs to sample; red: contributes high percentage of OTUs). Keeping the filter value unchanged, and click the "Sample ID" button, then a graphic will be generated as the example figure above.

Results directory

OTUs heatmap result:
result/02.OTUanalysis/taxa_stat/Sample_Tags-OTUs_dis.{png,svg}

3.2.3 Species classification analysis

3.2.3.1 Classification tree

Particular concerned species (top 10 genera for each sample, by default) were selected to draw the classification tree[8], and displayed by the independently developed software. The classification tree for single sample is shown in Figure 3.2.3.1-1. The classification tree for multiple samples is shown in Figure 3.2.3.1-2.

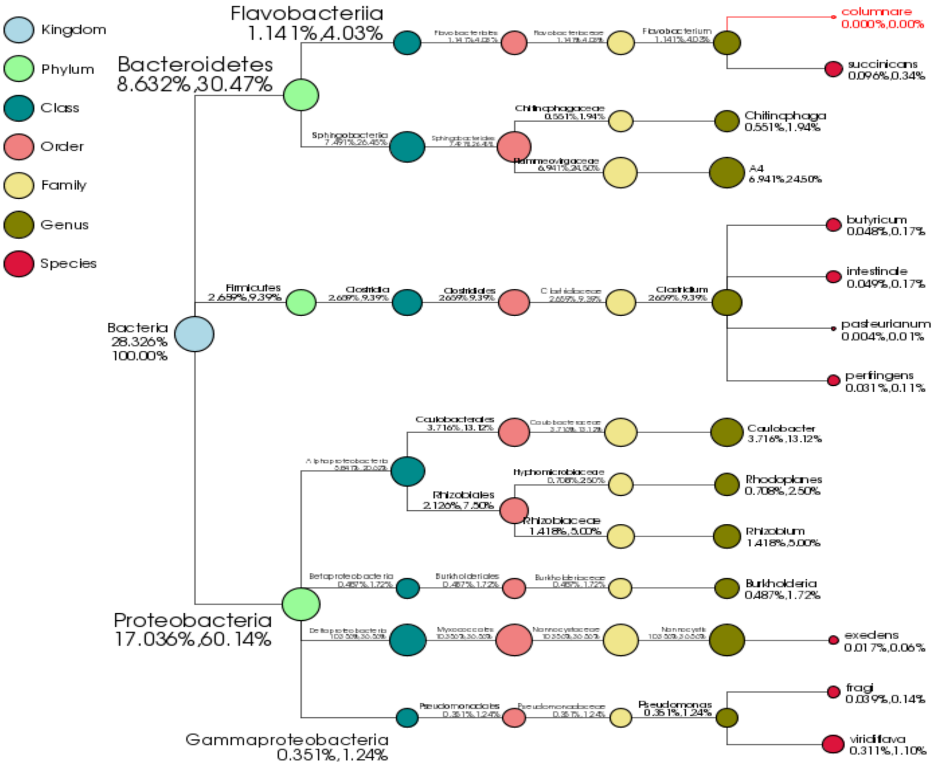


Figure 3.2.3.1-1 The classification tree for single sample

The number above (after the taxonomic ranks) represents the relative abundance of the whole corresponding taxon, while the second number represents the relative abundance of the selected corresponding taxon.

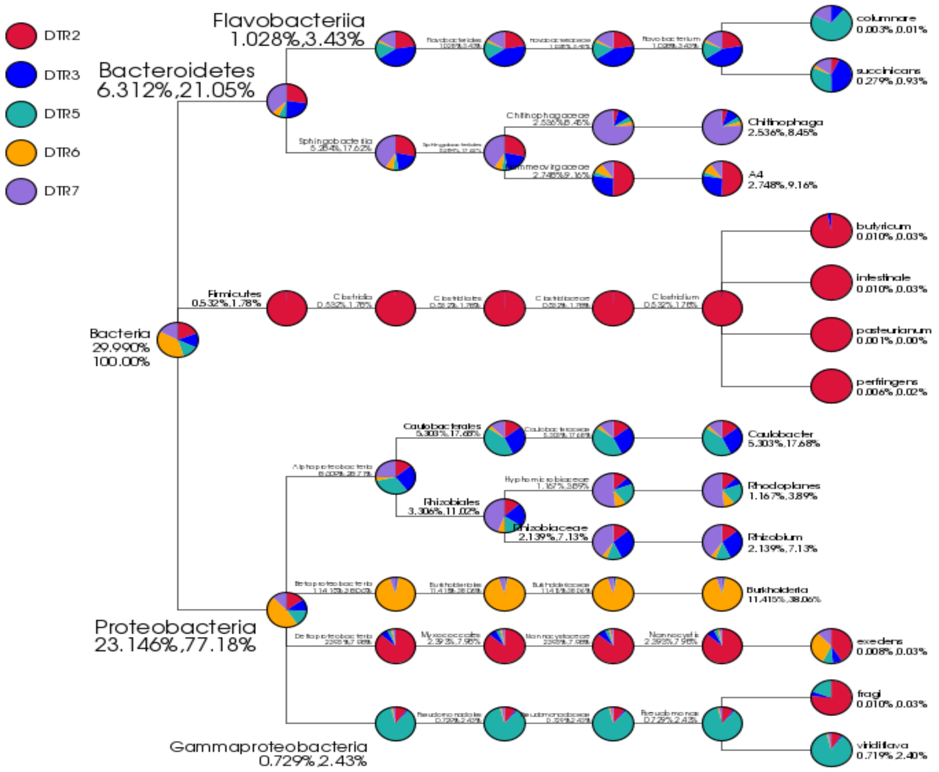


Figure 3.2.3.1-2 The classification tree for multiple samples

The number above (after the taxonomic ranks) represents the relative abundance of the whole corresponding taxon, while the second number represents the relative abundance of the selected of corresponding taxon.

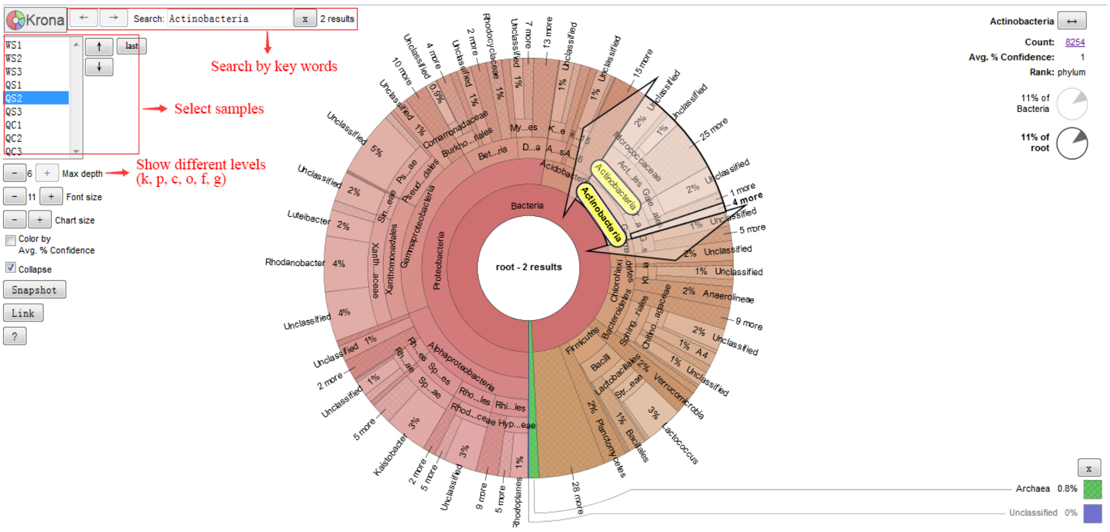
Results directory

Classification tree for multiple samples: result/02.OTUanalysis/taxa_tree/all.taxtree.{png,svg}

Classification tree for single samples: result/02.OTUanalysis/taxa_tree/*.taxtree.{png,svg}

3.2.3.2 Krona taxonomy visualization

The analysis result of species annotation is visually shown by KRONA[7] In the result display, circles from inside to outside stand for different classification levels, and the area of sector means respective proportion of different OTU annotation results, click for details. An example picture is as follows:



Actinobacteria
Count: 8254
Avg. % Confidence: 1
Rank: phylum
11% of Bacteria
11% of root

Figure 3.2.3.2 Krona taxonomy visualization

Results directory

Krona html visualization result: result/02.OTUanalysis/all_rep_set_tax_assignments.krona.html.

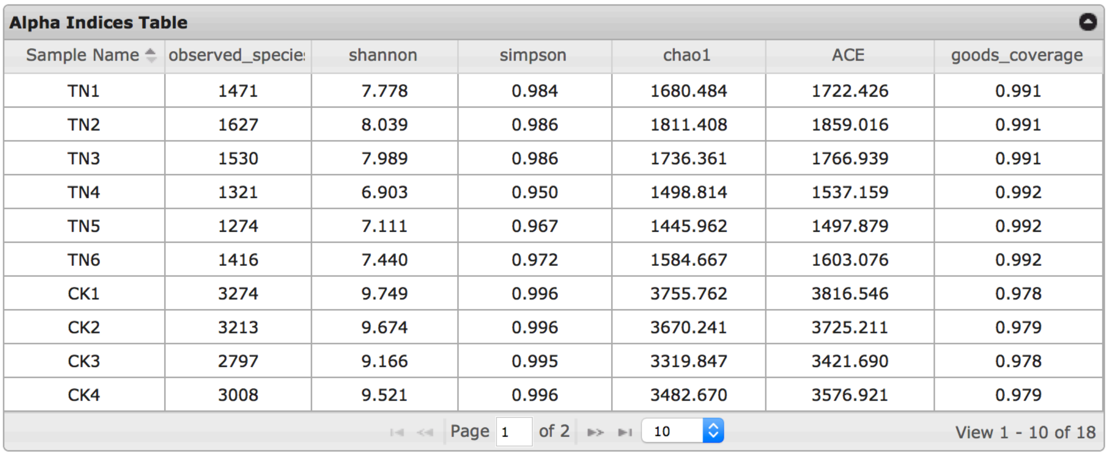
3.3 Alpha diversity analysis

Alpha diversity is widely used for the analysis of microbial community diversity[6]. It reflects the richness and diversity of microbial community by using a series of statistical indices, species accumulation curve and species richness curve.

3.3.1 Statistical indices for alpha diversity

Generally speaking, OTUs generated at 97% sequence identity are considered to be homologous on species level. Statistical indices of alpha diversity when the clustering threshold is 97% are summarized as below (Number of reads chosen for normalization: cutoff = 34778).

Table 3.3.1 Alpha indices table



Alpha Indices Table

Sample Name	observed_specie	shannon	simpson	chao1	ACE	goods_coverage
TN1	1471	7.778	0.984	1680.484	1722.426	0.991
TN2	1627	8.039	0.986	1811.408	1859.016	0.991
TN3	1530	7.989	0.986	1736.361	1766.939	0.991
TN4	1321	6.903	0.950	1498.814	1537.159	0.992
TN5	1274	7.111	0.967	1445.962	1497.879	0.992
TN6	1416	7.440	0.972	1584.667	1603.076	0.992
CK1	3274	9.749	0.996	3755.762	3816.546	0.978
CK2	3213	9.674	0.996	3670.241	3725.211	0.979
CK3	2797	9.166	0.995	3319.847	3421.690	0.978
CK4	3008	9.521	0.996	3482.670	3576.921	0.979

Page 1 of 2 10 View 1 - 10 of 18

Results directory

Alpha indices table: result/03.AlphaDiversity/alpha_diversity_index.xls

3.3.2 Species accumulation curve

Species accumulation curve describes the increase of species diversity with additional sample amount. It’s an effective tool for the investigation of species composition and prediction of species abundances. It’s also widely used to estimate whether sample amount is enough in the study of species diversity within community. When sample amount satisfies requirements, it can then be used to predict species richness(The threshold of sample is set higher than ten by default).

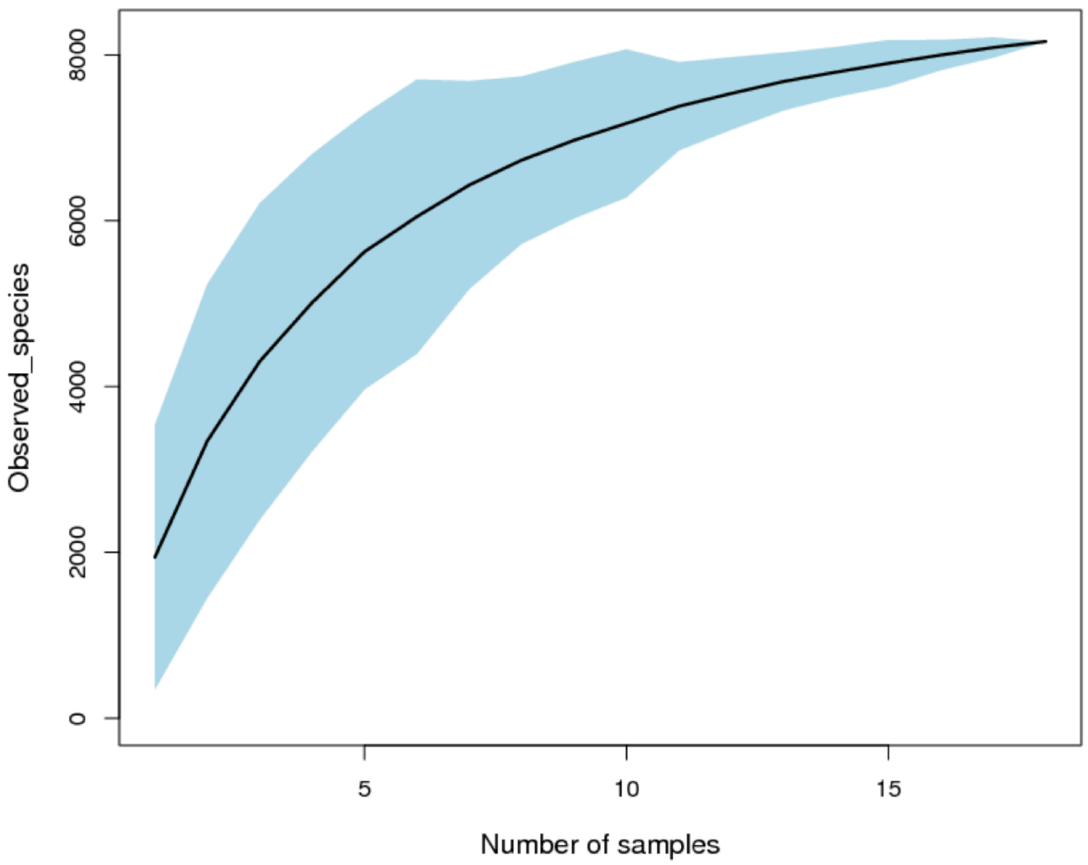


Figure 3.3.2 Species accumulation curve

X-axis represents sample amount, y-axis represents OTU number after sampling. The result reflects the occurrence rate of new OTUs (species) under continuous sampling. In a certain range, a sharp rising in the curve according to the increase of sample amount stands for a large number of species are discovered. A flat curve means that species in this environment won't increase much as the sample amount expanded. Species accumulation curve can be used to evaluate whether sample amount is enough. Sharping rising curve means lack of enough sample and more sampling are needed; instead, the sampling is enough and sequential data analysis is allowed.

Result directory

Species accumulation curve: result/05.StatTest/Specaccum/specaccum_test.{pdf,png}

3.3.3 Species richness curve

Rarefaction curve and Rank abundance curve are common methods to evaluate species richness. For rarefaction curve analysis, new OTUs (number of species) are generated by randomly resampling the complete set of OTUs versus the sampled number of reads.It reflects the reasonability of the number of sequencing reads used to be analyzed and can be used to infer species richness in the sample. A flat curve means that the number of sequencing reads is reasonable, and less new OTUs (new species) can be detected with increasing sequencing reads.

Rank abundance curves depict each pair of relative abundance and the corresponding abundance rank as a data point on the graph and then all the data points are linked to produce the curves. It directly reflects the species richness and species evenness in the sample. Species richness can be view as the range of the curve in the horizontal direction. The wider the curve range is, the higher the species richness is. Species evenness can be reflected by the steepness of the curve in the vertical direction. A shallow gradient indicates high evenness as the abundances of different species are similar[7]. (For high-quality picture of species richness curves please click)

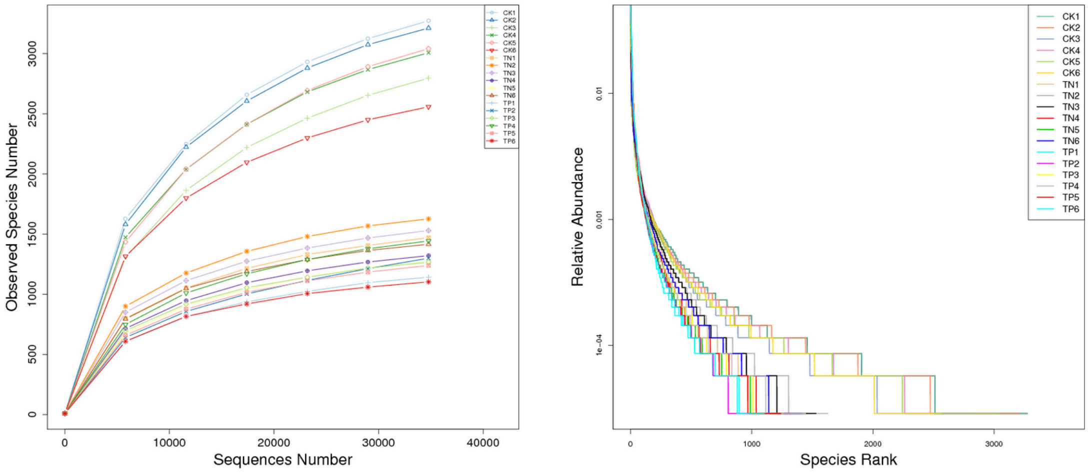


Figure 3.3.3 Rarefaction curves and rank abundance curves

In Rarefaction Curves plot, X-axis is number sequencing reads randomly chosen from a certain sample to obtain OTUs. Y-axis is corresponding OTUs. Curves for different samples are represented by different colors. In Rank Abundance Curves plot, X-axis is the abundance rank. The higher the abundance is, the smaller the rank is. Y-axis is the relative abundance. Curves for different samples are represented by different colors.

Result directory

Rarefaction curve: result/03.AlphaDiversity/observed_species.{pdf,png}

Rank abundance curve: result/03.AlphaDiversity/rank_abundance.{pdf,png}

Data for visualization: result/03.AlphaDiversity/plot_observed_species.txt

3.3.4 Venn diagram and flower diagram

According to the clustering analysis of OTUs and research requirements, common OTUs and unique OTUs belong to different samples or groups are counted. Venn diagram is provided with sample number or group number less than five, otherwise flower diagram would be provided. Both Venn diagram and flower diagram are plotted after data normalization for all samples.

3.3.4.1 Venn diagram based on OTUs

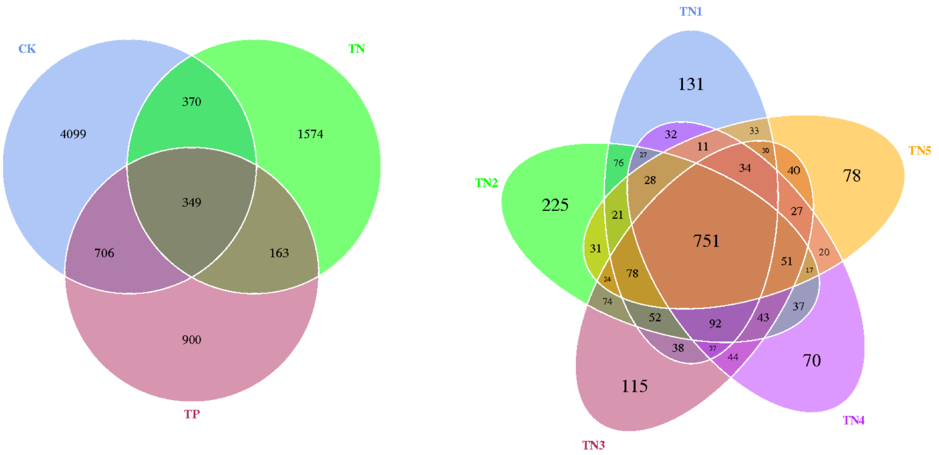


Figure 3.3.4.1 Each circle in the graph represents a sample or group

The number in the overlapping circles stands for common OTUs between different samples or groups, while the number in non-overlapping portion of the circle means the unique OTUs possessed by the corresponding sample or the group.

Result directory

Venn diagram: result/02.OTUanalysis/venn_figure/

Data for visualisation: result/02.OTUanalysis/venn_figure/venndata/

3.3.4.2 Flower diagram based on OTUs

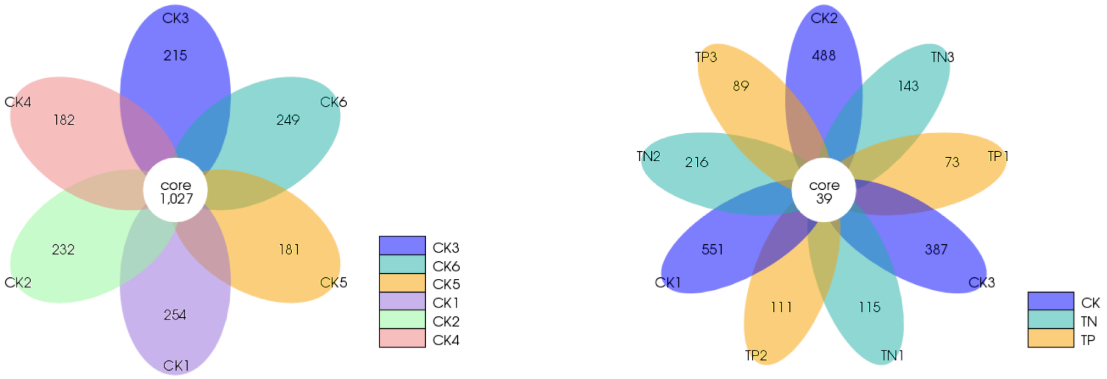


Figure 3.3.4.2 Flower diagram

Each petal in the flower diagram represents for a sample or group, with different colours for different samples or groups. The core number in the center is for the number of OTUs present in all samples, while number in the petal is for the unique OTUs only showing in each sample.

Result directory

Flower diagram: result/02.OTUanalysis/Flower_figure/

Data for visualization: result/02.OTUanalysis/flower_figure/flowerdata/

3.3.5 Between group variation analysis of alpha diversity indices

In the box plot of between group alpha diversity indices variation analysis, the mean value, the degree of dispersion, the maximum value, the minimum value, and the outliers of values of indices describing intra-group species diversity are displayed directly. It can also be used to analyze the significance of between group differences of species diversity(For the interpretation of box plot, please refer to box plot). The box plots of observed species and shannon index are displayed as below.

Box plot for between group variation analysis of alpha diversity indices

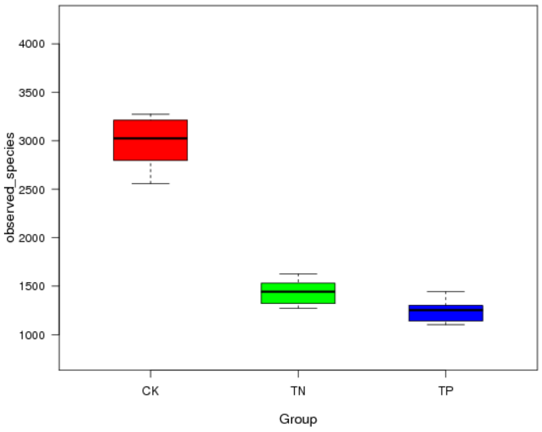


Figure 3.3.5-1 Box plot of observed species index

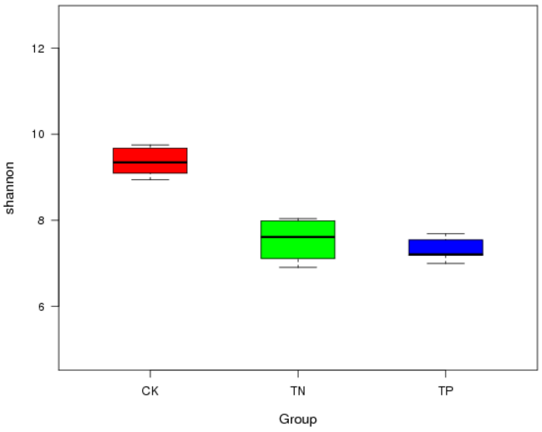


Figure 3.3.5-2 Box plot of Shannon index

Result Directory

Boxplot: result/05.StatTest/Alpha_div/*.{pdf,png}

ANOVA: result/05.StatTest/Alpha_div/*.txt

3.4 Beta diversity analysis

Beta diversity compares compositional heterogeneity among microbial communities. During beta diversity analysis, firstly, a Profiling Table is generated based on OTUs which are clustered into a single class according to their species annotation and abundance information. Then unweighted unifrac distance is calculated according to phylogenetic relationships of OTUs[8,9]. A matrix of unifrac distance is generated if there are more than two samples involved in the analysis. Weighted unifrac distance is calculated sequentially based on the unweighted unifrac distance by utilizing OTUs' abundance information[10]. At last, variations among samples or groups are determined by multi-variate statistical methods including Principle Component Analysis (PCA), Principal Co-ordinates Analysis (PCoA), Non-Metric Multi-Dimensional Scaling (NMDS), Unweighted Pair-group Method with Arithmetic Means (UPGMA).

3.4.1 Phylogenetic relationships

Phylogenetic relationships of the complete set of OTUs' representative sequences determined by multi sequence alignment is necessary and fundamental for further study of OTUs' phylogenetic relationships and indices of beta diversity. The phylogenetic relationships presented in the following are data chosen from the top 10 Genus ranked according to the maximum relative abundance of their corresponding OTUs combining with relative abundance of each OTU and confidence information of species annotation of each OTU's representative sequence. Phylogenetic Relationships of the top 10 Genus according to OTUs' information is as below. (For high-quality picture please click)

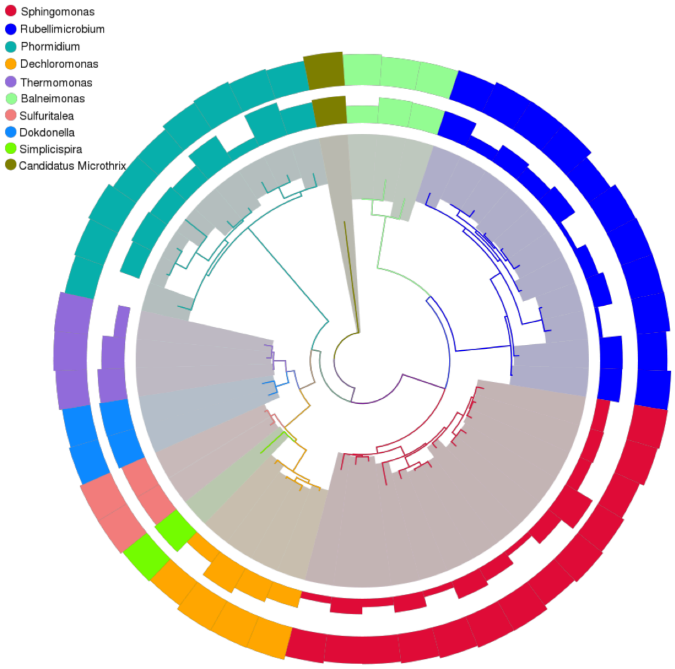


Figure 3.4.1 Phylogenetic Relationships of the top 10 Genus according to OTUs' information.

The inner-most layer is the phylogenetic tree constructed by representative sequence of OTUs. Each color corresponds to a unique genus. The second layer is the distribution of relative abundance of OTUs. The height of each bar stands for the numeric value of relative abundance of OTUs (The relative abundance is normalized according to the minimum value before presentation due to data may disperse across several scale). The outer- most layer is the distribution of the confidence of species annotation. The height of the bar equals to the confidence of the annotation.

Result directory
Phylogenetic relationship and species annotation:
result/02.OTUanalysis/phylo_tree/OTU.cluster.tree.{png,svg}
Phylogenetic tree: result/02.OTUanalysis/phylo_tree/OTUs.tre (viewed with MEGA)

3.4.2 Beta diversity indices

Unweighted unifrac and Weighted unifrac are chosen to estimate the variation coefficient of two samples. The smaller the number is, the less variation between the two samples in species diversity exists. Heatmap of unweighted unifrac distance and weighted unifrac distance is as below.

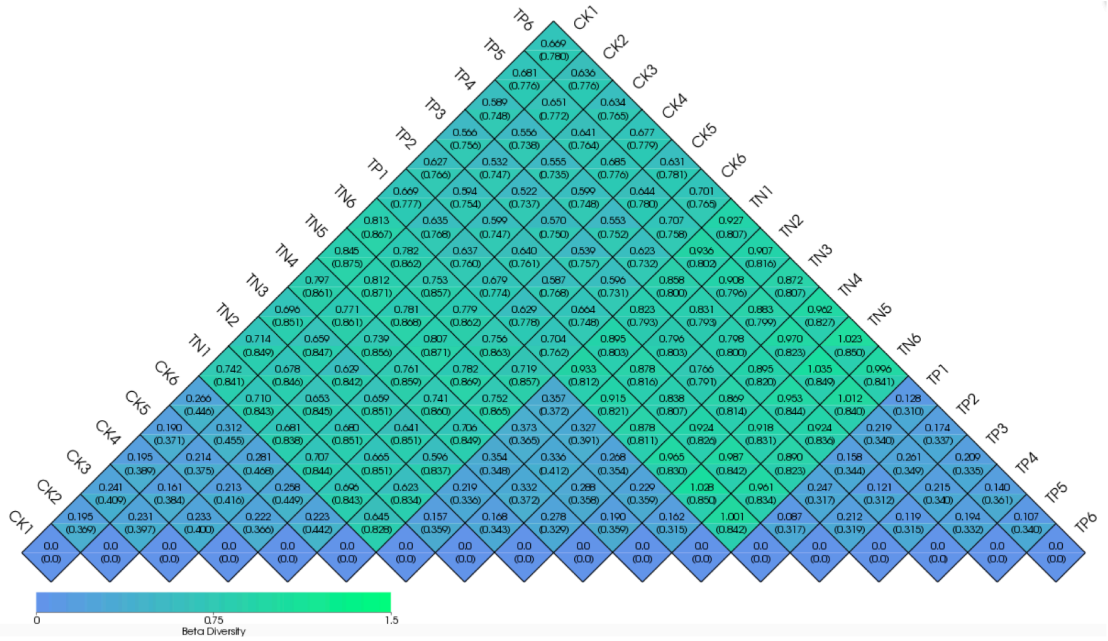


Figure 3.4.2 Heatmap of unweighted unifrac distance and weighted unifrac distance

The number in the square shows the variation of each pair of samples. The upper one stands for weighted unifrac distance, while the lower one stands for unweighted unifrac distance.

Result directory

Heatmap for indices of beta diversity:
result/04.BetaDiversity/beta_div_heatmap/beta_diversity.heatmap.*{png,svg}

Distance data file for visualization:
result/04.BetaDiversity/beta_div_heatmap/(un)weighted_unifrac_sorted_otu_table.txt

3.4.3 Principle component analysis(PCA)

Principal component analysis (PCA) is a statistical procedure to extract principle components and structures in data by using orthogonal transformation and reducing dimensionalities of data. It extracts the first two axises reflecting the variation of samples to the most extent thus can reflect high-dimensional data's variation in two-dimensional graph, which reveals the simple principle embedding in complex data.The more similar the composition of community among the samples are, the closer the distance of their corresponding data points on the PCA graph are. The result of PCA analysis based on OTUs is as below. (For high quality picture please click)

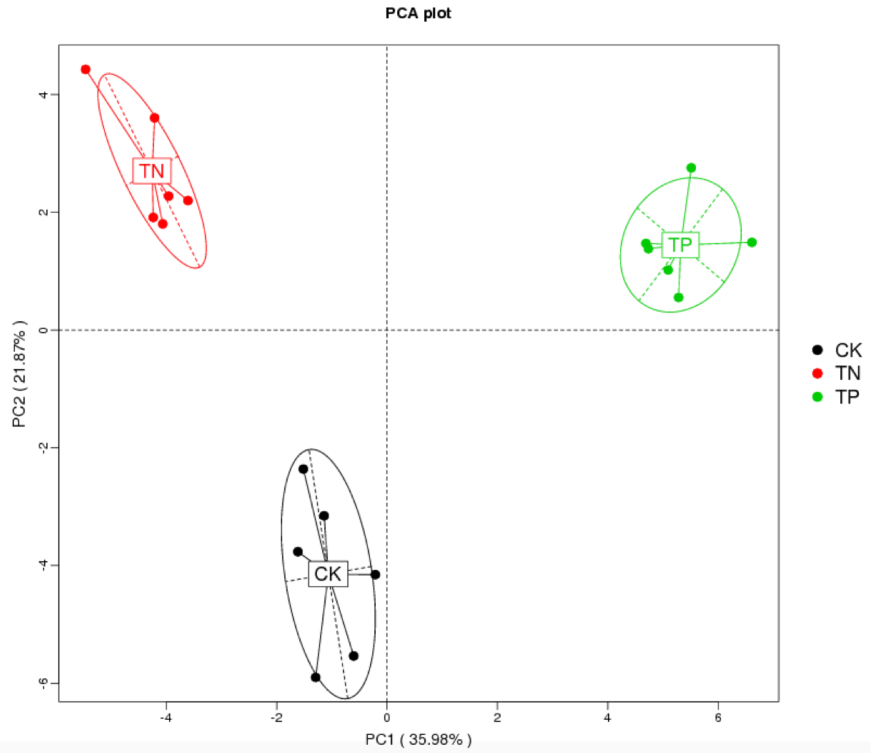


Figure 3.4.3 PCA analysis.

X-axis is the first principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Y-axis is the second principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Each data point in the graph stands for a sample. Samples belongs to the same group are in the same color. The clustering circle is added according to grouping information.

Result directory

Graph labeled with sample names: result/04.BetaDiversity/PCA/ PCA12.{png,pdf}

Graph unlabeled with sample names: result/04.BetaDiversity/PCA/PCA12_2.{png,pdf}

Graph labeled with sample names and clustering circle: result/04.BetaDiversity/PCA/PCA12_with_cluster.{png,pdf}

Graph unlabeled with sample names and clustering circle: result/04.BetaDiversity/PCA/PCA12_with_cluster_2.{png,pdf}

Result for PCA analysis: result/04.BetaDiversity/PCA/pca.csv

3.4.4 Principal co-ordinates analysis (PCoA)

Principal co-ordinates analysis is a similar method of reducing dimensions and ranking compared with PCA method, which extracts principle components and structures in multi-dimensional data via a series of eigenvalue and eigenvectors. The difference between PCoA and PCA is that PCoA searches the principle oordinate by distance matrix, while PCA does by similarity matrix. In our study, PCoA is based on unweighted unifrac distance and weighted unifrac distance. Principle coordinates combination that contributes most to variation in samples are chosen to be plotted. The closer the distance between different samples on the graph is, the more similar the species composition is. Samples with high similarity of community structure incline to be clustered together, while community with large variation will be separated remotely on the graph. The result of PCoA analysis is as below. (For high quality picture please click)

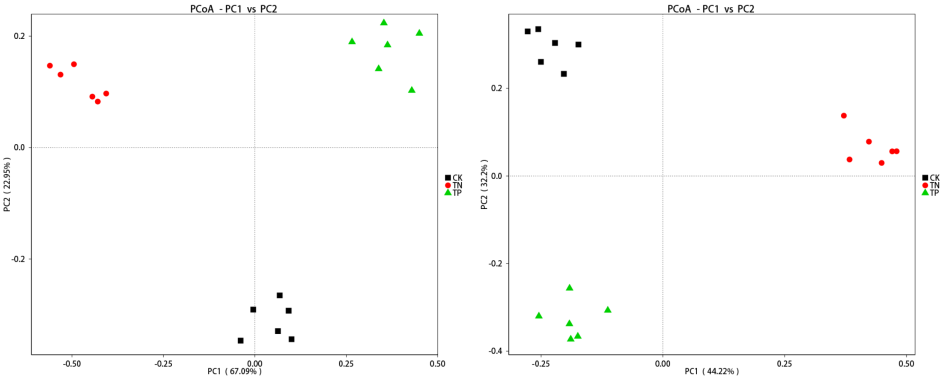


Figure 3.4.4 PCoA analysis

The left side picture is the result of PCoA based on weighted unifrac distance, while the right side picture is based on unweighted unifrac distance. X-axis represents the first principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Y-axis is the second principle component. Each data point in the graph stands for a sample. Samples belongs to the same group are in the same color.

Result directory

Result of PCoA: result/04.BetaDiversity/PCoA/(un)weighted_unifrac/

Matrix file for PCoA: result/04.BetaDiversity/PCoA/(un)weighted_unifrac/(un)weighted_unifrac_dm.txt

Principle component information for data visualisation: result/04.BetaDiversity/PCoA/(un)weighted_unifrac/(un)weighted_unifrac_pc.txt

3.4.5 Non-metric multi-dimensional scaling (NMDS)

Non-metric multi-dimensional scaling analysis is a ranking method applicable to ecological researches. It's a non-linear model designed for a better representation of non-linear biological data structure aiming at overcoming the flaws in methods based on linear model, including PCA and PCoA. The result of NMDS analysis based on OTUs is in Figure 3.4.5.

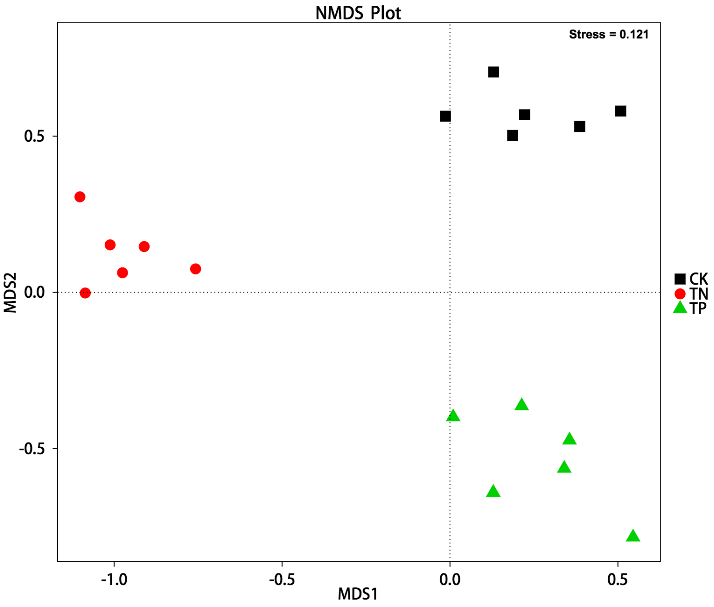


Figure 3.4.5 NMDS analysis

Each data point in the graph stands for a sample. The distance between data points reflects the extent of variation. Samples belongs to the same group are in the same color. When the value of Stress factor is less than 0.2, it's considered that NMDS is reliable to some extent.

Result directory

NMDS labeled with samples' names: result/04.BetaDiversity/NMDS/ NMDS.{png,pdf}

NMDS unlabeled with samples' names: result/04.BetaDiversity/NMDS/NMDS_2.{png,pdf}

NMDS Analysis Result: result/04.BetaDiversity/NMDS/NMDS_scores.txt

3.4.6 Clustering Analysis

To study the similarity among different samples, clustering analysis is applied and clustering tree can be constructed. Unweighted pair group method with arithmetic mean (UPGMA) is a type of hierarchical clustering methods widely used in ecology for the classification of samples. The basic ideas of UPGMA are as follows. First, samples with the closest distance are clustered together and a new node(as a new sample) is formed. It branching point is one half away from the original two samples. Then the average distance between the newly created "sample" and other samples is calculated and the nearest two samples could be found again to repeat above steps. A complete clustering tree could be obtained until all samples are clustered together.

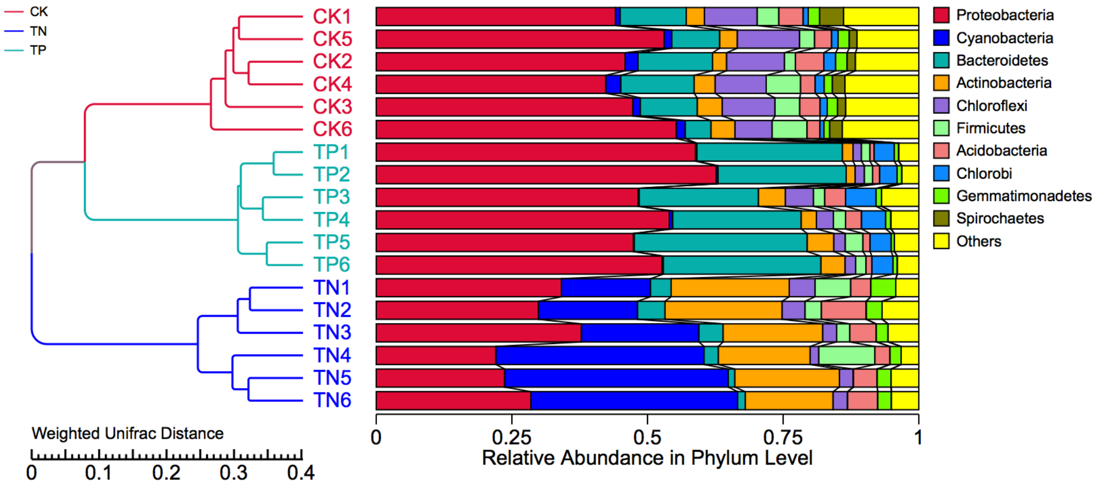


Figure 3.4.6-1 Clustering tree based on weighted unifrac distance

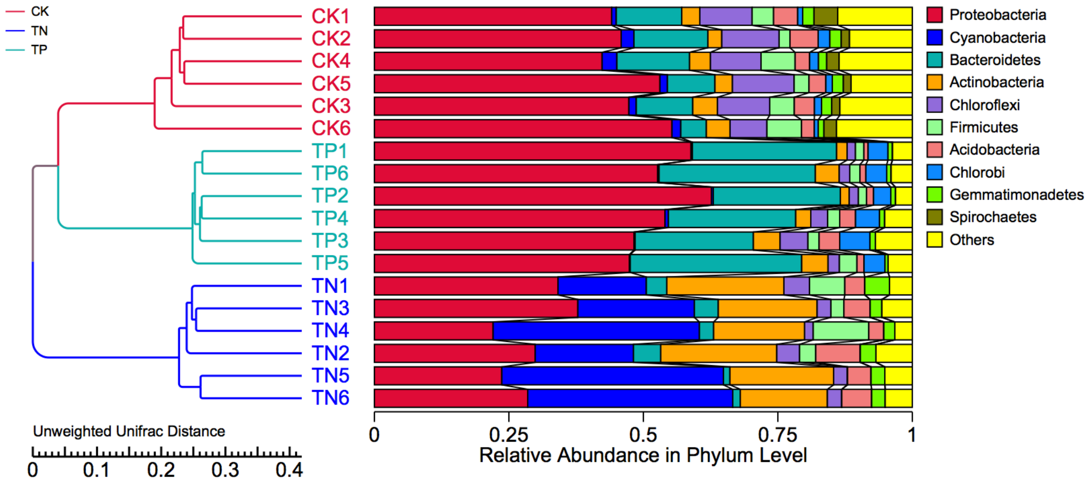


Figure 3.4.6-2 Clustering tree based on unweighted unifrac distance

Clustering results are displayed combined with each sample’s relative abundance on the level of phylum. The left side is the structure of clustering tree and the right side is the distribution of relative abundance.(For high-quality images please click weighted unifrac and unweighted unifrac).

Result directory

Sample clustering tree based on unweighted or weighted distances:
result/04.BetaDiversity/tree/(un)weighted_unifrac/(un)weighted_unifrac.{pdf,png}

Sample clustering tree combined with top 10 phyla distribution:
result/04.BetaDiversity/tree/(un)weighted_unifrac/UPGMA.W(UnW).tree.{png,svg}

3.4.7 Between group variation analysis of beta diversity indices

Beta diversity indices box plot directly reflect the mean value, the degree of dispersion, the maximum value, the minimum value, and the outliers of values of indices describing intra-group species diversity. It can also be used to analysis the significance of between group differences of species diversity(For the interpretation of box plot, please refer to box plot). The box plot for the analysis of between group variation of species diversity is as below.

Box plot for between group variation analysis of indices for alpha diversity. For high-quality picture please click.

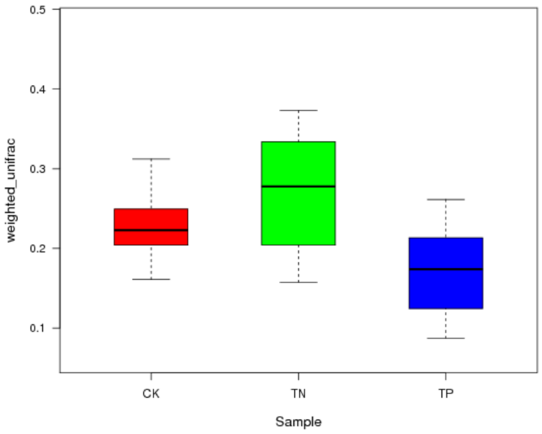


Figure 3.4.7-1 Box plot of weighted unifrac distance

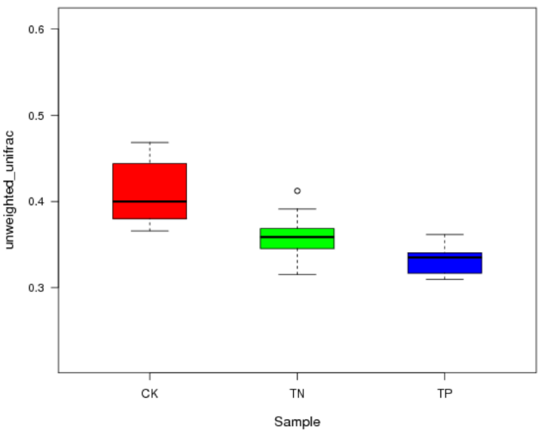


Figure 3.4.7-2 Box plot of unweighted unifrac distance

Result directory

Box plot for beta diversity:
result/05.StatTest/Beta_div/(un)weighted_unifrac.{pdf,png}

Significance of variation analysis:
result/05.StatTest/Beta_div/(un)weighted_unifrac_test.txt

3.5 Statistical analysis

Statistical analysis of different communities can be performed especially for those projects involving multiple groups. It captured those species whose abundance varies significantly among groups, meanwhile, the distribution of these variant species among the groups is also obtained. By comparing the within group variation and variation among groups, we can whether the variation of the community structure among different groups is significant can be determined.

3.5.1 Between groups t-test analysis

T-test is performed to determine species with significant variation between groups(p value < 0.05) at various taxon levels including phylum, class, order, family, genus, and species. For high-quality images please click.

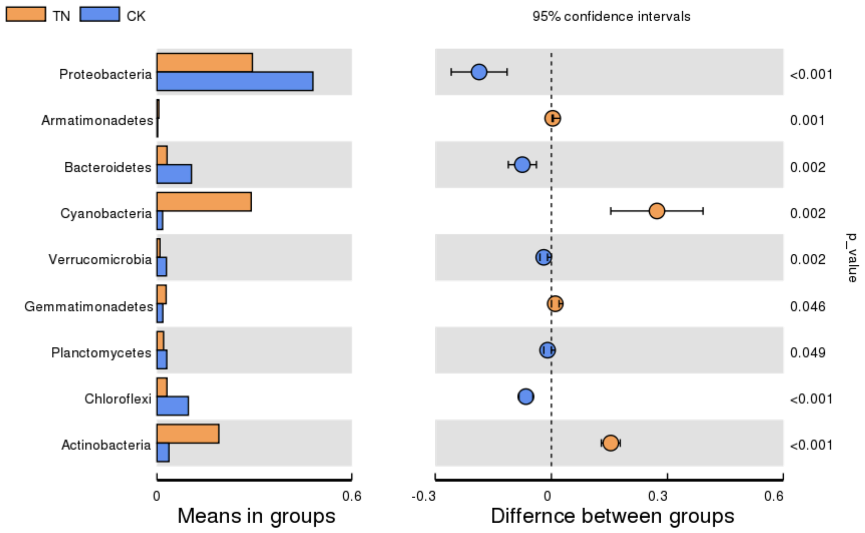


Figure 3.5.1 Between groups T-test analysis

The left panel is the abundance of species showing significant between group variation. Each bar represents the mean value of the abundance in each group of the specie showing significant between group variation. The right panel is the confidential interval of between group variation. The left-most part of each circle stands for the lower limit of 95% confidential interval, while the right-most part is the upper limit. The center of the circle stands for the difference of the mean value. The color of the circle is in agree with the group whose mean value is higher. The right-most value is the p-value of the significance test of between group variation.

Result directory

T-test of between group variation on various taxon level:
result/05.StatTest/t.test_bar_plot

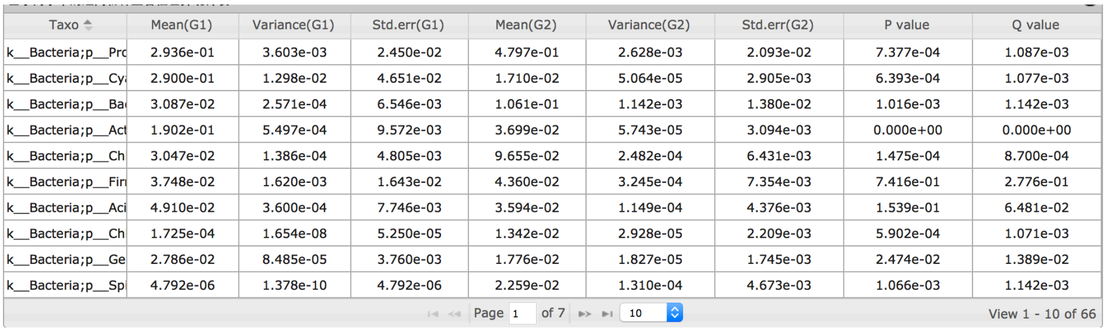
T-test of between group variation on the level of phylum: result/05.StatTest/t.test_bar_plot/phylum/*.(png,svg)

T-test result: result/05.StatTest/t.test_bar_plot/phylum/*.psig.xls

3.5.2 Metastats analysis

Species with significant intra-group variation are detected via metastats, a strict statistical methods, according to their abundance[23]. The significance of observed abundance's differences among groups is evaluated via multiple hypothesis-test for sparsely-sampled features and false discovery rate(FDR)(Table 5.2).

Table 3.5.2 Statistical results of intra-group variation of species abundance at phylum level



Taxo	Mean(G1)	Variance(G1)	Std.err(G1)	Mean(G2)	Variance(G2)	Std.err(G2)	P value	Q value
k__Bacteria;p__Pr	2.936e-01	3.603e-03	2.450e-02	4.797e-01	2.628e-03	2.093e-02	7.377e-04	1.087e-03
k__Bacteria;p__Cy	2.900e-01	1.298e-02	4.651e-02	1.710e-02	5.064e-05	2.905e-03	6.393e-04	1.077e-03
k__Bacteria;p__Ba	3.087e-02	2.571e-04	6.546e-03	1.061e-01	1.142e-03	1.380e-02	1.016e-03	1.142e-03
k__Bacteria;p__Act	1.902e-01	5.497e-04	9.572e-03	3.699e-02	5.743e-05	3.094e-03	0.000e+00	0.000e+00
k__Bacteria;p__Ch	3.047e-02	1.386e-04	4.805e-03	9.655e-02	2.482e-04	6.431e-03	1.475e-04	8.700e-04
k__Bacteria;p__Fir	3.748e-02	1.620e-03	1.643e-02	4.360e-02	3.245e-04	7.354e-03	7.416e-01	2.776e-01
k__Bacteria;p__Aci	4.910e-02	3.600e-04	7.746e-03	3.594e-02	1.149e-04	4.376e-03	1.539e-01	6.481e-02
k__Bacteria;p__Ch	1.725e-04	1.654e-08	5.250e-05	1.342e-02	2.928e-05	2.209e-03	5.902e-04	1.071e-03
k__Bacteria;p__Ge	2.786e-02	8.485e-05	3.760e-03	1.776e-02	1.827e-05	1.745e-03	2.474e-02	1.389e-02
k__Bacteria;p__Sp	4.792e-06	1.378e-10	4.792e-06	2.259e-02	1.310e-04	4.673e-03	1.066e-03	1.142e-03

Taxonomic information is shown in the column of Taxon Numbers listed in columns of Mean(G1), Variance(G1) and Std.err(G1) are the first group's the mean value, the variation and standard variation, separately, while Mean(G2), Variance(G2), Std.err(G2) are the second group's. P value is the p-value calculated from the hypothesis-test, Q value is the q-value corrected by the p-value.

Result directory

Results ofmetastats analysis at taxonomic level (phylum, class, order, family, genus, species): result/04.BetaDiversity/MetaStat

Results of metastats analysis at phylum level: result/04.BetaDiversity/MetaStat/*.test.xls

Statistical metrics obtained from metastats analysis at phylum level when p-value is less than 0.05: result/04.BetaDiversity/MetaStat/phylum /*.psig.xls

Statistical metrics obtained from metastats analysis at phylum level when q-value is less than 0.05: result/04.BetaDiversity/MetaStat/phylum /*.qsig.xls

3.5.3 LEfSe analysis

LEfSe (linear discriminant analysis (LDA) effect size) analysis detects biomarkers with statistical differences among groups, namely, species with significant intra-group variation. LEfSe is a software aiming at discovering high-dimensional biomarkers and revealing metagenomic features, including genes, metabolics, or taxa, thus can be used to distinguish two or more biological classes. It emphasizes statistical significance, biological consistency, and effect relevance and allows researchers to identify features of abundance and related classes. Its result is consisted of the histogram of LDA scores, the Cladogram and the histogram of statistically different biomarkers' relative abundance among groups (Figure 3.5.3 for high quality picture please click Histogram of LDA Scores and Cladogram).

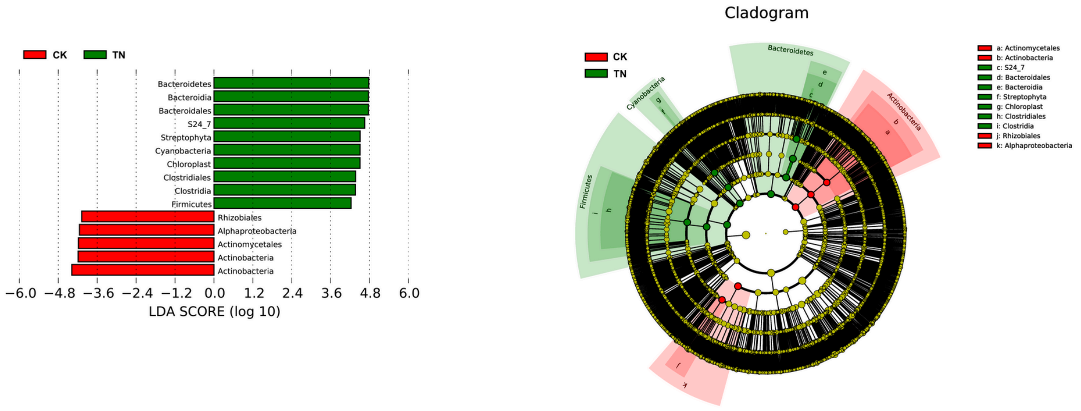


Figure 3.5.3 LEfSe analysis. Histogram of the LDA scores and Cladogram

Histogram of the LDA scores and Cladogram are shown as the results of LEfSe analysis for evaluating of biomarkers with statistically difference among groups. The histogram of the LDA scores presents species(biomarker) whose abundance shows significant differences among groups. The selecting criteria is that LDA scores are larger than the set threshold(4 set by default). The length of each bin, namely, the LDA score, represents the effect size (the extent to which a biomarker can explain the differentiating phenotypes among groups).

In Cladogram, circles radiating from inner side to outer side represents taxonomic level from phylum to genus(species). Each circle stands for a distinct taxon at corresponding taxonomic level. Each circle's diameter is proportional to the taxon's relative abundance. Coloring principles are as the followings. Yellow stands for species with non-significant differences. Species (biomarkers) with significant differences are colored according to corresponding group's color. Red nodes means these microbiota contributes a lot in the group denoted by red color, so do the green nodes. Letters above the circles and corresponding species are annotated on the right side.

Result directory

The histogram of LDA scores: result/04.BetaDiversity/LEfSe/*/LDA.*.(pdf,png)

The Cladogram: result/04.BetaDiversity/LEfSe/*/LDA.*.tree.(pdf,png)

The relative abundance of biomarker in each group: result/04.BetaDiversity/LEfSe/*/biomarkers_raw_images/

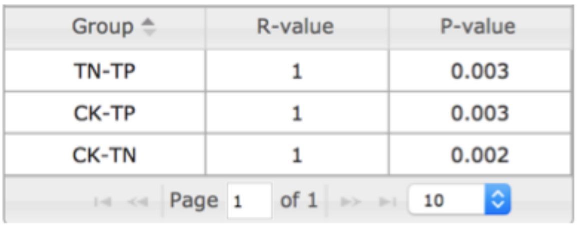
3.5.4 Anosim and MRPP

Anosim and MRPP analysis estimate the significance of differences among groups of community and compare the inner-group and inter-group variation.

3.5.4.1 Anosim

Anosim analysis is a nonparametric test to evaluate whether variation among groups is significantly larger than variation within groups, which helps to evaluate the reasonability of the division of groups. For detailed calculating steps please refer to Anosim.

Table 3.5.4.1 Anosim



Group	R-value	P-value
TN-TP	1	0.003
CK-TP	1	0.003
CK-TN	1	0.002

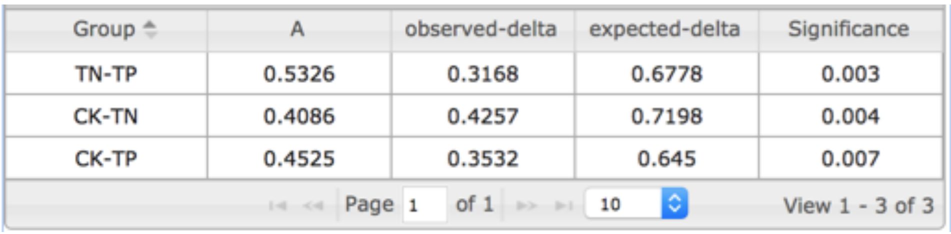
Page 1 of 1
10

R-value is a number between -1 and 1. A positive R value means that inter-group variation is considered significant, while a negative R-value suggests that inner-group variation is larger that inter-group variation, namely, no significant differences. The confidence degree is represented by P-value, whose value less than 0.05 suggests statistical significance.

3.5.4.2 MRPP

MRPP is similar with Anosim, which aims at determining whether the difference of microbial community structure among groups is significant. It’s usually applied with methods for dimension reduction like PCA, PCoA, and NMDS. For detailed calculating steps, please refer to MRPP.

Table 3.5.4.2 MRPP



Group	A	observed-delta	expected-delta	Significance
TN-TP	0.5326	0.3168	0.6778	0.003
CK-TN	0.4086	0.4257	0.7198	0.004
CK-TP	0.4525	0.3532	0.645	0.007

Page 1 of 1
10
View 1 - 3 of 3

A small value of the number in the column titled observe-delta indicates that the inner-group variation is small, while a large one in the column of expected-delta means that the inter-group variation is large. A positive A-value suggests that variation among groups is larger than variation within groups, while a negative one shows the opposite relationship. The difference among groups is significant if the number in the column of Significance is less than 0.05.

Result directory

Anosim: result/04.BetaDiversity/Anosim/stat_anosim.txt

MRPP: result/04.BetaDiversity/MRPP/stat_mrpp.txt

3.6 Data Mining

16S amplicon sequencing is widely used for microbial community comparison among samples from various natural or endozoic environments such as soil, water, host intestine etc. In order to achieve these objectives, several important results needed to be highly concerned.

Firstly, OTUs cluster and species annotation results are summarized in result/02.OTUanalysis/. Tags are clustered with 97% identity, all the represented tags for each OTU are list in result/02.OTUanalysis/OTUs.fasta. These OTUs are then annotated and collected in result/02.OTUanalysis/OTUs.tax_assignments.txt. Species abundance are displayed in two important directory: Absolute/ (containing absolute species composition of in different taxonomic levels), Evenabs/ (containing absolute species composition after normalization), and Relative/ (containing relative abundance for each sample after normalization, which are mainly summarized for the subsequent alpha diversity and beta diversity analysis). For instance, the directory Relative/ contains species relative abundance of each sample on different taxonomic levels (kingdom, phylum, class, order, family, genus, species). From these results, we can visualize species composition of various samples, and focus on some concerned species or vastly different species among samples (or groups) correlate with our certain research objectives.

Dominant species distribution among samples are visualized in the directory result/02.OTUanalysis/top10 (with bar chart and profiling table on p, c, o, f, g level) so that we could locate the notable predominant species fast and convenient, and then goes onto abundance analysis and difference tests.

Results about sample complexity are mainly included in the directory result/ 03.AlphaDiversity/ with six different alpha diversity indices (Observed_species, Goods_coverage, Chao1, ACE, Shannon, Simpson).

As for the difference comparison of microbial communities between samples, results are displayed in the directory result/04.BetaDiversity. Firstly, the Unifrac distance between pairwise samples are visualized as a heatmap to measure and view the dissimilarity extent, the result is represented in result/04.BetaDiversity/beta_div_heatmap. The dissimilarity are then calculated with gradient analysis and displayed with ordination plots (PCA, PCoA, etc.), samples with similar microbial community structure tend to be gathered, and vice versa. Samples could then clustered by UPGMA based on the acquired distance matrix, and visualized in result/04.BetaDiversity/Tree/. From these results, we can figure out the complexity differences between samples, and explain the differences between samples (or groups) combining with specific underlying biological problems. For instance, we can explain sample cluster results with UPGMA considering high-abundance taxa to achieve the underlying driving factors.

When there are more than 2 groups, more advanced analysis could be done.

For species differences, we can use Metastat to obtain the significance of all species between groups and select obvious different species between groups on various taxonomic levels (p, c, o, f, g, s) for further analysis, or choose LefSE analysis to figure out statistic significant different biomarkers among groups.

Anosim and MRPP analysis could be used to determine whether community structure significant differs between groups, or comparing the differences between groups and within groups.

For an exploratory analysis, if there are several environmental factors concerned, we could select CCA or RDA analysis to extracts environmental gradients from ecological datasets, and to find environmental driving factors which influence the development of certain microbial communities.

NMDS analysis could be selected as a supplementary method with unexpected results through PCA and PCoA, for it is based on nonlinear model (PCA and PCoA are both based on linear model), and may offer a better explanation of the nonlinear structure in ecological datasets.

4. Methods

4.1 Sequencing

4.1.1 Extraction of genome DNA

Total genome DNA from samples was extracted using CTAB/SDS method. DNA concentration and purity was monitored on 1% agarose gels. According to the concentration, DNA was diluted to 1ng/μL using sterile water.

4.1.2 Amplicon Generation

16S rRNA/18SrRNA/ITS genes of distinct regions(16SV4/16SV3/16SV3-V4/16SV4-V5, 18S V4/18S V9, ITS1/ITS2, Arc V4) were amplified used specific primer(e.g. 16S V4: 515F-806R, 18S V4: 528F-706R, 18S V9: 1380F-1510R, et. al) with the barcode. All PCR reactions were carried out with Phusion® High-Fidelity PCR Master Mix (New England Biolabs).

4.1.3 PCR Products quantification and qualification

Mix same volume of 1X loading buffer (contained SYB green) with PCR products and operate electrophoresis on 2% agarose gel for detection. Samples with bright main strip between 400-450bp were chosen for further experiments.

4.1.4 PCR Products Mixing and Purification

PCR products was mixed in equidensity ratios. Then, mixture PCR products was purified with Qiagen Gel Extraction Kit(Qiagen, Germany).

4.1.5 Library preparation and sequencing

Sequencing libraries were generated using TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, USA) following manufacturer's recommendations and index codes were added. The library quality was assessed on the Qubit@ 2.0 Fluorometer (Thermo Scientific) and Agilent Bioanalyzer 2100 system. At last, the library was sequenced on an IlluminaHiSeq2500 platform and 250 bp paired-end reads were generated.

4.2 Data analysis

4.2.1 Paired-end reads merging and quality control

1）Data split: Paired-end reads was assigned to samples based on their unique barcode and truncated by cutting off the barcode and primer sequence.

2）Reads merging: Paired-end reads were merged using FLASH (V1.2.7, http://ccb.jhu.edu/software/FLASH/) [15], a very fast and accurate analysis tool, which was designed to merge paired-end reads when at least some of the reads overlap the read generated from the opposite end of the same DNA fragment, and the splicing sequences were called raw tags.

3）Data Filtration: Quality filtering on the raw tags were performed under specific filtering conditions to obtain the high-quality clean tags [16] according to the Qiime(V1.7.0，http://qiime.org/scripts/split_libraries_fastq.html)[17] quality controlled process.

4） Chimera removal: The tags were compared with the reference database(Gold database，http://drive5.com/uchime/uchime_download.html)using UCHIME algorithm(UCHIME Algorithm, http://www.drive5.com/usearch/manual/uchime_algo.html)[18] to detect chimera sequences, and then the chimera sequences were removed [19]. Then the Effective Tags finally obtained.

4.2.2 OTU cluster and Species annotation

1) OTU Production: Sequences analysis were performed by Uparse software (Uparse v7.0.1001， http://drive5.com/uparse/) [20]Sequences with ≥97% similarity were assigned to the same OTUs. Representative sequence for each OTU was screened for further annotation.

2) Species annotation: For each representative sequence, the Greengene Database (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi)[21]was used based on RDP classifier(Version 2.2,http://sourceforge.net/projects/rdp-classifier/)[22]algorithm to annotate taxonomic information.

3) Phylogenetic relationship Construction: In order to study phylogenetic relationship of different OTUs, and the difference of the dominant species in different samples(groups), multiple sequence alignment were conducted using the PyNAST software(Version 1.2)[23] against the "Core Set" dataset in the Greengene database.

4) Data Normalization: OTUs abundance information were normalized using a standard of sequence number corresponding to the sample with the least sequences. Subsequent analysis of alpha diversity and beta diversity were all performed basing on this output normalized data.

4.2.3 Alpha Diversity

Alpha diversity is applied in analyzing complexity of species diversity for a sample through 6 indices, including Observed-species, Chao1, Shannon, Simpson, ACE, Good-coverage. All this indices in our samples were calculated with QIIME(Version 1.7.0) and displayed with R software(Version 2.15.3).

Two indices were selected to identify Community richness:

Chao - the Chao1 estimator(http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.chao1.html#skbio.diversity.alpha.chao1);

ACE - the ACE estimator (http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.ace.html#skbio.diversity.alpha.ace);

Two indices were used to identify Community diversity:

Shannon - the Shannon index (http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.shannon.html#skbio.diversity.alpha.shannon);

Simpson - the Simpson index (http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.simpson.html#skbio.diversity.alpha.simpson);

One indice to characterized Sequencing depth:

Coverage - the Good’s coverage (http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.goods_coverage.html#skbio.diversity.alpha.goods_coverage)

4.2.4 Beta Diversity

Beta diversity analysis was used to evaluate differences of samples in species complexity, Beta diversity on both weighted and unweighted unifrac were calculated by QIIME software (Version 1.7.0).

Cluster analysis was preceded by principal component analysis (PCA), which was applied to reduce the dimension of the original variables using the FactoMineR package and ggplot2 package in R software(Version 2.15.3).

Principal Coordinate Analysis (PCoA) was performed to get principal coordinates and visualize from complex, multidimensional data. A distance matrix of weighted or unweighted unifrac among samples obtained before was transformed to a new set of orthogonal axes, by which the maximum variation factor is demonstrated by first principal coordinate, and the second maximum one by the second principal coordinate, and so on. PCoA analysis was displayed by WGCNA package, stat packages and ggplot2 package in R software(Version 2.15.3).

Unweighted Pair-group Method with Arithmetic Means(UPGMA) Clustering was performed as a type of hierarchical clustering method to interpret the distance matrix using average linkage and was conducted by QIIME software (Version 1.7.0).

5. References

[1] Caporaso, J. Gregory, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences 108.Supplement 1 (2011): 4516-4522.
[2] Youssef, Noha, et al. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. Applied and environmental microbiology 75.16 (2009): 5227-5236.
[3] Hess, Matthias, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331.6016 (2011): 463-467.
[4] Luo, Chengwei, et al. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS one 7.2 (2012): e30087.
[5] Degnan, Patrick H., and Howard Ochman. Illumina-based analysis of microbial community diversity.The ISME journal 6.1 (2012): 183-194.
[6] Caporaso, J. Gregory, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal 6.8 (2012): 1621-1624.
[7] Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. Interactive metagenomic visualization in a Web browser.BMC bioinformatics 12.1 (2011): 385.
[8] Li, Bing, et al. Characterization of tetracycline resistant bacterial community in saline activated sludge using batch stress incubation with high-throughput sequencing analysis. Water research 47.13 (2013): 4207-4216.
[9] Whittaker, Robert H. Evolution and measurement of species diversity. Taxon (1972): 213-251.
[10] Lundberg, Derek S., et al. Practical innovations for high-throughput amplicon sequencing. Nature methods 10.10 (2013): 999-1002.
[11] Lozupone, Catherine, and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology 71.12 (2005): 8228-8235.
[12] Lozupone, Catherine, et al. UniFrac: an effective distance metric for microbial community comparison. The ISME journal 5.2 (2011): 169
[13] Lozupone, Catherine A., et al. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology 73.5 (2007): 1576-1585.
[14] Avershina, Ekaterina, Trine Frisli, and Knut Rudi. De novo Semi-alignment of 16S rRNA Gene Sequences for Deep Phylogenetic Characterization of Next Generation Sequencing Data. Microbes and Environments 28.2 (2013): 211-216.
[15] Magoč T, Salzberg S L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27.21 (2011): 2957-2963.

[16] Bokulich, Nicholas A., et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods 10.1 (2013): 57-59.
[17] Caporaso, J. Gregory, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7.5 (2010): 335-336.
[18] Edgar, Robert C., et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27.16 (2011): 2194-2200.
[19] Haas, Brian J., et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research 21.3 (2011): 494-504.
[20] Edgar, Robert C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10.10 (2013): 996-998.
[21] DeSantis, Todd Z., et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 72.7 (2006): 5069-5072.
[22] Wang, Qiong, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 73.16 (2007): 5261-5267.
[23] Caporaso, J. Gregory, et al. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26.2 (2010): 266-267.