

# Microflora Diversity Analysis Report

---

service@allbiolife.com

Contract No : NXXXXXX

Customer : XXXXXX

BI : test

Date : 2019-10-09

- ① **Experimental workflow**
- ② **Bioinformatics workflow**
- ③ **Bioinformatics analysis result**

### **1 Data processing**

- 1.1 Preliminary sequencing data statistics
- 1.2 Sequencing data quality optimization

### **2 OTU analysis and species annotation**

- 2.1 OTU clustering
- 2.2 OTU heatmap
- 2.3 OTU Venn diagram and petal diagram
- 2.4 Species annotation statistics
- 2.5 Graphics display of species relative abundance
- 2.6 Phylogenetic Tree of the Genus

### **3 Alpha Diversity**

- 3.1  $\alpha$  diversity analysis
- 3.2 Between-group differential analysis using  $\alpha$ -diversity indices
- 3.3 Rank-Abundance curve
- 3.4 Rarefaction curve
- 3.5 Species accumulation curve

### **4 Beta Diversity Analysis**

- 4.1 Unifrac distance matrix
- 4.2 PCoA analysis
- 4.3 PCA Analysis
- 4.4 NMDS Analysis
- 4.5 UPGMA-Tree Cluster Analysis
- 4.6 PCA\_Box Analysis
- 4.7 3D-PCoA Analysis

### **5 Variation analysis Between groups**

- 5.1 Group difference evaluation by Anosim
- 5.2 Adonis analysis
- 5.3 Differential analysis by Metastats
- 5.4 STAMP analysis
- 5.5 LEfSE analysis
- 5.6 Wilcoxon Rank sum test
- 5.7 ROC curve Analysis

## 6 Other analysis

6.1 Correlation analysis of community composition and environmental factors

6.2 RDA/CCA analysis

6.3 Enterotype analysis

6.4 Krona species composition diagram

6.5 GraPhlAn Analysis

6.6 Phylogenetic Tree

6.7 Collinearity between Samples and species

6.8 Network Analysis

## ④ Appendix

## ⑤ Reference



Experimental workflow

## ① Experimental workflow

16S/18S/ITS rRNA is composed of conserved and hypervariable regions. Whereas conserved regions are not significantly different across various microbial strains, the sequences of hypervariable regions are genus or species-specific, and differ in accordance to phylogenetic difference. Therefore, 16S/18S/ITS rDNA serve as identifiers of biological species, and are important for microbial phylogeny and taxonomic identification. 16S/18S/ITS rDNA amplicon sequencing has become an important tool for the study of the composition of microbial communities in environment.

16S rDNA amplicon sequencing includes the library construction using specific primers to amplify the variable region of prokaryotic 16S rDNA and data analysis of the 16S rDNA variable region sequence to identify the composition and abundance of prokaryotic microorganisms in the environment. The proprietary workflow at ALLBIO effectively amplifies the two variable regions of 16S rDNA (V3 and V4) and accurately identifies various species including archaea. 18S / ITS rDNA amplicon sequencing includes the library construction using specific primers to amplify the variable region of eukaryotic 18S / ITS rDNA and data analysis to identify the composition and abundance of eukaryotic microorganisms in the environment. Illumina MiSeq / NovaSeq sequencing platform is widely used for 16S / 18S / ITS rDNA amplicon sequencing because of its deep sequencing depth, high throughput, short run-time and high sequencing accuracy as well as reasonable cost. In recent years, pair-end chemistry has enabled sequencing platform to read longer, which further increased the accuracy of the results.

16S / 18S / ITS rDNA amplicon sequencing procedure includes genomic DNA extraction, quality control, rDNA variable region amplification, library construction, high-throughput sequencing and data analysis. All the steps are important for data quality and quantity, which in turn affects the subsequent data analysis. In order to ensure data accuracy and reliability, every step has to pass strict quality control before pooling the library by adjusting the volume of each library according to the target data volume for Illumina MiSeq / NovaSeq sequencing. The workflow is as below:



**Figure I.1** Microflora diversity experimental workflow

2

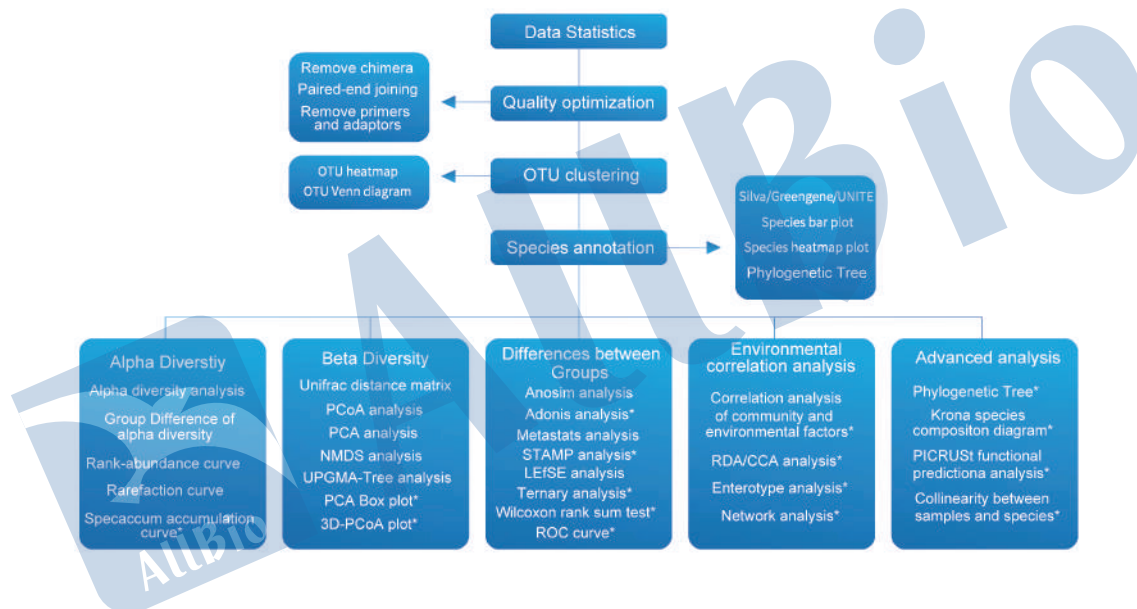
**Anaysis workflow**



## ① Anaysis workflow

First, adapters and low quality data were filtered out from the original data. Then the chimera sequences were removed to obtain the effective sequences for cluster analysis. Each cluster was called an OTU (Operational Taxonomic Unit). The taxonomy analysis of the representative sequence of each OTU was then performed to obtain species distribution information. Based on the results of OTU analysis,  $\alpha$ -diversity indices of each sample can be derived as well as the species richness and evenness. Based on taxonomic information, statistical analysis of community structure can be carried out at each classification level. UPGMA clustering tree and PCoA plots can be constructed based on Unifrac distance to illustrate the differences in community structure between different samples or groups.

Following the basic analysis above, a series of in-depth data mining can be carried out. For example, researchers can investigate the different community structure among different groups of samples using multiple statistical methods. This could be further combined with the environmental factors and species diversity to discover the environmental factors important for community structure.



**Figure II.1** Analysis workflow

Note:

- (1) If the number of samples is less than 3, comparative analysis of diversity cannot be performed.
- (2) The differential statistical analysis is meaningful only if there are at least three replicates in the biology group.
- (3) Environmental factors are required for correlation analysis of community composition and environmental factors as well as for CCA / RDA analysis.
- (4) Intestinal-type analysis only applies to animal or human intestinal or stool samples.
- (5) Analysis with "\*" are not included in the standard analysis and can be selected and analyzed according to individual sample and project.

3

Analysis result



## ① Analysis result

### 1 Data processing

The original image data were analyzed using Bcl2fastq (v2.17.1.14) for base calling and preliminary quality analysis. During the sequencing process, Illumina built-in software based on each sequencing segment, namely read, the first 25 The quality of the base determines whether the read is retained or discarded. The result is stored in the FASTQ file format and contains the sequencing sequence information (the second row in the FASTQ format) and the corresponding sequencing quality information FASTQ format fourth line).

FASTQ format has four lines of information for each sequence as shown below:

```
@ALLBIOHISEQ01:289:C3Y96ACXX:6:1101:1704:2425 1:N:0:GGCTAC
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTGAACTTCTCTGT
+
@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHGCCF
```

The first and third lines contain sequence identifier information produced by the sequencer (some fastq files omit name information and leaves it empty after the “+” sign on the third line to save space). The second line contains the sequence information. The fourth line depicts the quality information of each corresponding base on the second line. The fourth line contains sequence quality information, and the quality score is the ASCII value of the corresponding character minus 33. For example, the ASCII value of '@' is 64, and therefore the corresponding base quality score is 31 (64-33). Starting with Illumina GA Pipeline v1.8 (currently v1.9), base quality scores range from 0 to 41.

**Table 1.1** Explanation of the elements in illumina sequence identifiers.

Type	Description
GWZHISEQ01	Unique instrument name
289	Run ID
C3Y96ACXX	Flowcell ID
6	Flowcell lane
1101	Tile number within the flowcell lane
1704	'x'-coordinate of the cluster within the tile
2424	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
GGCTAC	Index sequence

Sequencing base quality is affected by sequencer, reagents, samples and other factors. The first few bases from the 5'-end are usually of higher error rate and the error rate drops afterward. With long read sequencing platforms, sequencing error rate might rise again close to the 3'-end. This is one of the inherent short-comings/features of high-throughput sequencing (Erlich and Mitra, 2008; Jiang et al.). The first six bases usually have a higher than average error rate. Since this is also the length of the random primer, it is suggested that the high error rate is due to the annealing between not perfectly matched primers and template (Jiang et al.). Statistics of sequencing error rate across all base positions can be used to spot the existence of abnormally high error rates. For example, it would raise a red flag if the base error rate in the

middle of the sequence is significantly higher than that of the positions close to the end. In general, the sequencing error rate for each base position is less than 0.5%. An error in the sequence is indicated by letter 'e'. The base quality scores of Illumina HiSeq (TM) /MiSeq platforms are expressed in Q Phred. The formula to calculate QPhred based on error rate is:

$$\text{Formulat: } Q_{\text{phred}} = -10\log_{10}(e)$$

**Table 1.2** The corresponding correlation between Illumina's Bcl2fastqbase calling and Qphred scores

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	100%

## 1.1 Preliminary sequencing data statistics

Information on data volume and sequencing quality of sequencing data (PF data) were summarized.

The statistics of raw data of each sample are in the following table:

**Table 1.1.1** Raw data statistics (shown is partial result. Details please see PFdata\_stat.xls)

Sample	Length(bp)	#Reads	Bases(bp)	Q20(%)	Q30(%)	GC(%)
A_1	300.00	126840	38052000	89.81	79.57	55.46
A_2	300.00	154500	46350000	90.18	80.05	55.13
A_3	300.00	169436	50830800	90.37	80.38	55.10
B_1	300.00	112418	33725400	90.10	79.95	55.12
B_2	300.00	144034	43210200	90.75	80.79	54.91

### Column description :

- (1) Sample: Sequencing sample name;
- (2) length(bp): Read average length;
- (3) #Reads: Read count;
- (4) Bases(bp): Base count;
- (5) Q20(%),Q30(%): The percentage of bases with Phred value greater than 20 or 30;
- (6) GC%: GC base percentage.

## 1.2 Sequencing data quality optimization

Sequencing errors such as point mutations might occur in high-throughput sequencing, and it's common that bases toward the end of the sequence reads have lower than average quality. In order to obtain higher quality and more accurate bioinformatic analysis results, it is necessary to optimize the raw data of the sequencing to obtain higher quality and more accurate bioinformatics analysis results.

Analysis software: Cutadapt (v1.9.1) ,Vsearch (1.9.6) ,Qiime (1.9.1)

#### Steps and parameters for optimization:

- (1) The two sequences of each read pair were merged according to overlapping sequences. The read merge is deemed to be successful only if the overlapping sequence is least 20bp long. After merging, undetermined bases (N) were removed from the resulting sequence.
- (2) Primer and adapter sequences were removed. Then the 5' and 3' bases with Q score lower than 20 were also removed. The resulting sequences with length > 200bp would pass this step of processing.
- (3) The sequences obtained were then aligned to database to identify and remove chimera sequence. Sequences passed this filtering step are deemed as clean data ready for analysis.

The statistical result of the data filtering is shown in the table below:

**Table 1.2.1** Statistics of filtered data (shown is partial result. Details please see effective\_stat.xls)

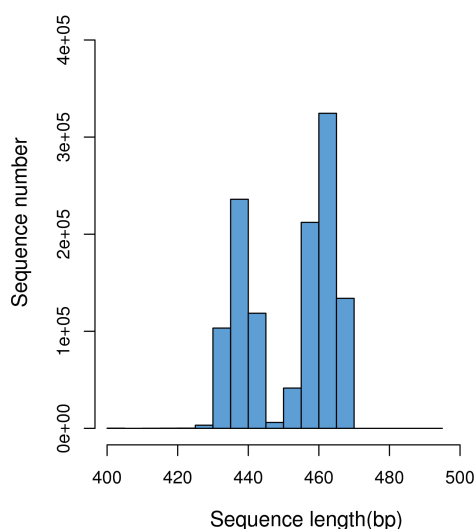
Sample	#PE_reads	#Nochimera	AvgLen(bp)	GC(%)
A_1	63420	53774	451.66	55.06
A_2	77250	66795	452.22	54.77
A_3	84718	73199	452.46	54.73
B_1	56209	49140	455.19	54.73
B_2	72017	84785	455.85	54.55

#### Column description :

- (1) Sample: Name of the sequencing sample
- (2) #PE\_reads: Number of the original PE reads
- (3) #Nochimera: Number of filtered sequences after chimera removal
- (4) AvgLen(bp): Average length of the filtered sequences
- (5) GC(%): GC percentage after filtering.

Effective sequence length distribution is shown as follows:

Effective sequence length distribution



**Figure 1.2.1** Effective sequence length distribution

Note: The X axis is the sequence length (bp), and the Y axis is the sequence count of the corresponding length.

## 2 OTU analysis and species annotation

### 2.1 OTU clustering

OTU is an operational definition of a classification unit (genus, species, grouping, etc.) commonly used in population genetics to facilitate data analysis. In bioinformatics each sequence obtained from sequencing is assumed to be derived from a single species. All the sequences in a sample are classified to obtain information on species and genus. By classification, the sequences are grouped according to their similarity, and one group is an OTU. Typically, OTU cluster are defined by a 97% identity threshold for data statistics and analysis.

Analysis software: Qiime (1.9.1), Vsearch (1.9.6)

#### Analysis methods and steps:

- (1) Unique sequences are extracted from the optimized sequences with the read count information.
- (2) Remove the unique sequence with 1 read count.
- (3) OTU clustering of unique sequences (read count > 1) was performed with similarity of 97%, and chimeric sequences were further removed to obtain the representative OTU sequences.
- (4) All optimized sequences are compared with OTU representative sequences, and sequences of >97% similarity to a specific OTU representative sequence are considered to be of the same OTU. Finally the OTU abundance were also summarized in the result table.

The following table shows the statistics of the sequence number in each sample's OTU

**Table 2.1.1** OTU Table (shown is partial result. Details please see otu\_taxa\_table.xls)

OTU_ID	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3	taxonomy
OTU2	4	0	3	4494	4237	4113	2507	2422	2464	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Hydrogenophilaes;f__Hydrogenophilaceae;g__Tepidiphilus;s__uncultured_bacterium
OTU1	1	0	4	2611	2738	2891	2405	2454	2732	k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;p__Sphingobacteriales;f__Lentimicrobiaceae;g__uncultured_bacterium;s__uncultured_bacterium
OTU3	2	1	0	1187	1319	1417	2140	2357	2523	k__Bacteria;p__Bacteroidetes;c__WCHB1-32
OTU4	1	0	0	1358	1551	1882	1568	1883	2147	k__Bacteria;p__Thermotogae;c__Thermotogae;o__Kosmotogales;f__Kosmotogaceae;g__Mesotoga;s__uncultured_bacterium
OTU5	2	2	2	1713	1516	1325	1761	1452	1738	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Ruminococcus_1

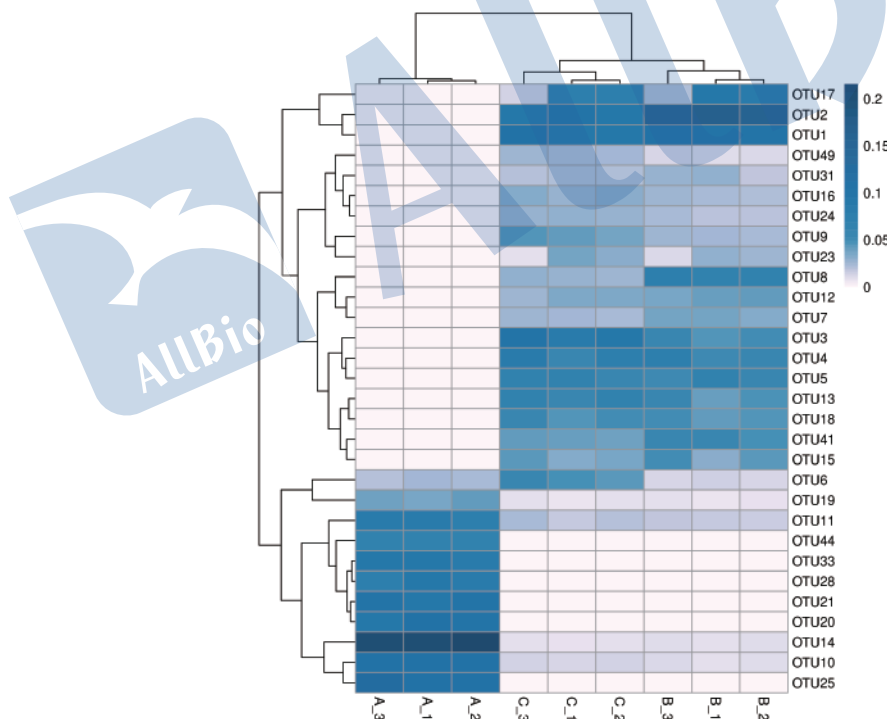
#### Column description :

- (1) OTU\_ID: OTU number;
- (2) Sample\_Name: The abundance of each OTU in the sample
- (3) taxonomy: Species annotation information of the corresponding OTU.

## 2.2 OTU heatmap

The heatmap analysis shows the abundance information of selected OTU as well as the similarity and difference across OTUs and samples by similarity clustering. The figure below shows the top 30 OTUs with the highest abundance:

Analysis software: Plot by R,basing on otu\_tables.xls



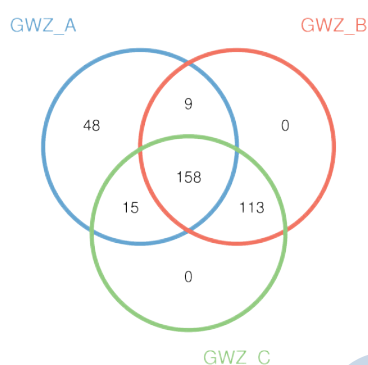
**Figure 2.2.1** OTU abundance clustering heatmap

Note: The row name is the OTU ID, the column name is the sample information, the left side of the figure is the OTU cluster tree, and the top is the sample cluster tree. The value of each colored box is the relative abundance of each OUT after normalization.

## 2.3 OTU Venn diagram and petal diagram

According to the results of OTU cluster analysis, the common and unique OTUs of different samples/groups are analyzed. When the number of samples/groups is less than 5, the Venn diagram is drawn. When the sample/group is greater than 5, the petal diagram is drawn.

Analysis software: Statistics and plot by R



**Figure 2.3.1** OTU Venn diagram or petal diagram

Note: the circles of different colors in the Venn diagram represent different samples or groups, and the numbers in the figure represent the numbers of OTUs unique or common to each sample or group. In the petal diagram, each petal represents a sample or group. The numbers on the petals represent the number of OTUs unique to the sample, and the white circle in the middle represents the number of OTUs shared by all samples and groups.

## 2.4 Species annotation statistics

In order to obtain the classification information of OTU, a representative sequence was selected for each OTU and annotated using the RDP classifier, thereby to obtain the community composition of each sample.

Analysis software: Qiime (1.9.1)

Analysis method: RDP classifier Bayesian algorithm was used to classify the OTU representative sequences of 97% similarity level, and the community composition of each sample was analyzed and summarized at all levels. The comparison database was Silva\_132 16S rRNA database (<http://www.arb-silva.de/>) / Silva\_132 18S rRNA database (<http://www.arb-silva.de/>) / ITS database (<https://unite.ut.ee/>)

For each sample, the percentage of each species at different taxonomic levels (Phylum, Class, Order, Families, Genus, Species) is shown in the table below:

**Table 2.4.1** Taxa Statistics at Genus level (shown is partial result. For details please see Genus\_abundance.xls)

Taxon	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3
Unclassified	57.35	58.72	58.25	37.43	38.72	40.17	43.99	44.15	46.00
Tepidiphilus	0.01	0.00	0.01	11.32	10.67	10.36	6.32	6.10	6.21
Ambiguous_taxa	2.71	2.43	2.36	5.61	5.72	5.75	7.45	8.13	7.81
Mesotoga	0.00	0.00	0.00	3.42	3.91	4.74	3.95	4.74	5.41
Blvii28_wastewater-sludge_group	0.00	0.00	0.00	6.03	5.73	1.45	4.88	5.02	0.92

#### Column description :

- (1) axonomic classification at the phylum level
- (2) Sample\_Name: The percentage of samples in different species classification

The statistics of the number of species on each taxonomic levels (Kingdom, Phylum, Class, Order, Families, Genus, Species) are as follows:

**Table 2.4.2** Statistics of Taxonomic Composition (shown is partial result. For details please see Sample\_tax\_stat.xls)

Samples	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3
Kingdom	2	1	1	2	2	2	2	2	2
Phylum	25	22	22	31	30	29	30	30	30
Class	42	39	37	45	48	44	47	44	46
Order	48	46	46	48	54	50	53	51	53
Family	72	71	72	74	79	72	77	77	77

#### Note:

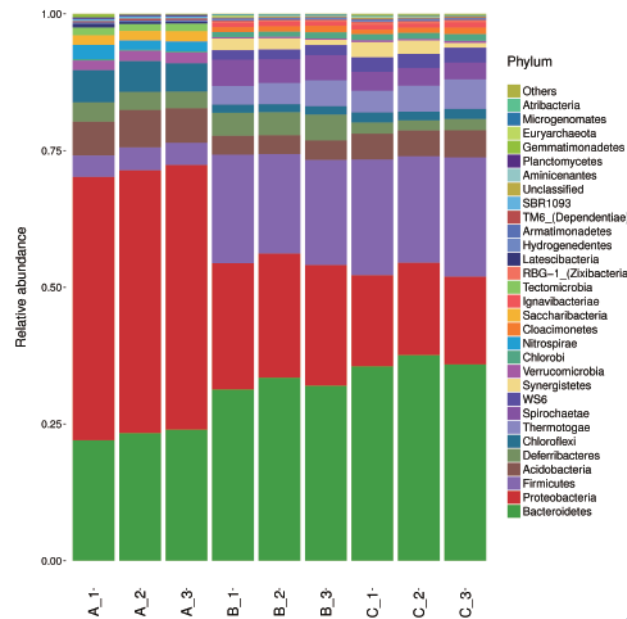
- (1) Samples: Sample name
- (2) Kindom: Species number of each sample on Kingdom level
- (3) Phylum: Species number of each sample on Phylum level
- (4) Class: Species number of each sample on Class level
- (5) Order: Species number of each sample on Order level
- (6) Family: Species number of each sample on Family level
- (7) Genus: Species number of each sample on Genus level
- (8) Species: Species number of each sample on Species level

## 2.5 Graphics display of species relative abundance

Analysis software: Statistics and plot by R

The distribution of the top 30 most abundant classifications in each sample or group at different taxonomic levels (Phylum, Class, Order, Families, Genus, Species) are shown as follows:

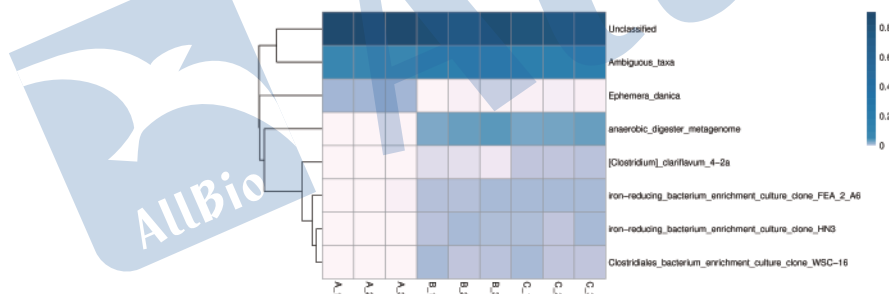




**Figure 2.5.1** Stacked bar plot of species distribution

Note: X axis is the sample name or group name, and the Y axis is the relative abundance of different species. The legend is the name of the taxonomic classification of the species. 'Other' represents the relative abundance of all phylum level classifications other than the top 30.

The top 30 species distribution of each sample (or group) on different levels (Phylum, Class, Order, Families, Genus, Species) was clustered and plotted in a heatmap. The similarity and difference of each species is visualized by color scheme in the heat map. The heat maps plotted for the distribution of the species in each sample on different levels of classification are shown below:



**Figure 2.5.2** Species distribution heatmap

Note: The columns represent samples and/or groups and the rows represent species. The dendrogram above the heatmap is the cluster result of the samples and/or groups and the dendrogram to the left is the species cluster. The colors in the heat map represent the relative for the distribution of the species in each sample on different levels of classification are shown below:

## 2.6 Phylogenetic Tree of the Genus

Phylogenetic Tree infers approximately-maximum-likelihood phylogenetic trees from alignments of the Top 30 OTU sequences.

Analysis software: Plot by R

Note: Phylogenetic Tree infers approximately-maximum-likelihood phylogenetic trees from alignments of the OTU sequences, the color of the branch said its corresponding Phylum, different colors represents different Phylum.

### 3.1 $\alpha$ diversity analysis

The indices for community richness calculation include:

( <http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.ace.html#skbio.diversity.alpha.ace> )

( <http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.chao1.html#skbio.diversity.alpha.chao1> )

The indices for community richness calculation include:

( <http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.shannon.html#skbio.diversity.alpha.shannon> )

( <http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.simpson.html#skbio.diversity.alpha.simpson> )

**Good's Coverage:** Refers to library coverage of each sample. The higher the value, the lower the probability that the sample did not cover the sequence.

( [http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.goods\\_coverage.html#skbio.diversity.alpha.goods\\_coverage](http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.goods_coverage.html#skbio.diversity.alpha.goods_coverage) )

Analysis software: Qiime ( 1.9.1 )

Analysis method: sequences were randomly extracted and the valid sequences were subject to OTU analysis and  $\alpha$  diversity index was calculated for each sample.

$\alpha$  diversity results are summarized in the table below:

**Table 3.1.1** Collation of alpha diversity results ( partial result is displayed, for details please see alpha\_rarefaction.xls )

Sample	ace	chao1	shannon	simpson	goods_coverage
A_1	226.988	224.5	6.553	0.982	1
A_2	226.676	227.429	6.507	0.981	1
A_3	225.355	224.125	6.489	0.981	1
B_1	270.857	269.5	5.97	0.966	1
B_2	275.492	285.333	5.982	0.967	0.999

#### Column description :

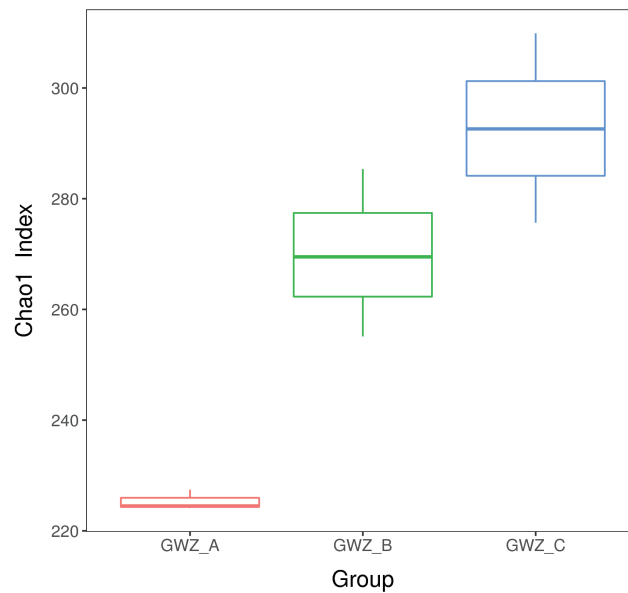
- (1) Sample: Sample Name
- (2) ace: ACE diversity index
- (3) chao1: Chao1 diversity index
- (4) shannon: Shannon diversity index
- (5) simpson: Simpson diversity index
- (6) good's\_coverage: Good's\_coverage diversity index

## 3.2 Between-group differential analysis using $\alpha$ -diversity indices

To perform between-group  $\alpha$  diversity analysis, box plots were generated based on  $\alpha$  diversity indices using R, which intuitively displays the maximum, minimum, median and outliers of the  $\alpha$  diversity indices of samples in each group as well as the differences between groups.

Analysis method: Box plot was generated using R based on  $\alpha$ -diversity index.

The box plot of the inter-group differential analysis based on chao1 and shannon indices is as below:



**Figure 3.2.1** Boxplot of between-group diversity comparison

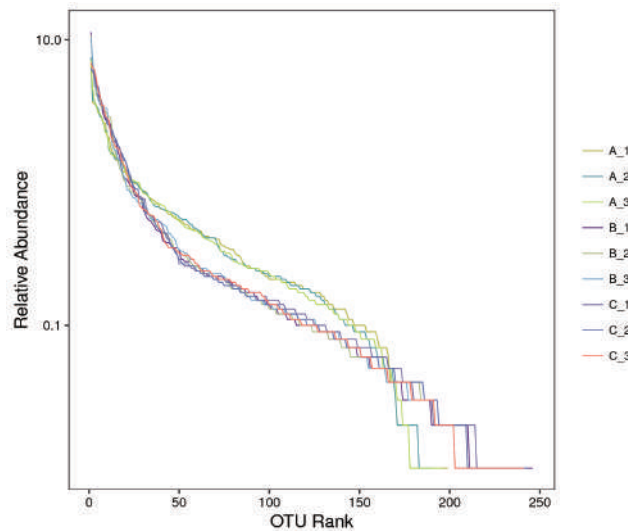
Note: The left panel is the Chao1 index boxplot of each group. X axis indicates the names of the groups and Y axis indicates the Chao 1 index. Each box diagram shows the minimum, first quartile, medium, third quartile and maximum values of the chao1 index of the corresponding sample. The right graph is the Shannon index boxplot of each group.

### 3.3 Rank-Abundance curve

Rank-abundance curve is used to analyze diversity. To generate a rank-abundance curve, the number of valid sequences in each OTU of a given sample was first calculated, and then all the OTUs were ranked in descending order based on their relative abundance (number of valid sequences), and finally the result was plotted with OTU ranking on the X axis and the number of sequences in the OTU on the Y axis. Y axis could also be OTU relative abundance in percentage.

Rank-abundance curve reflects both species abundance and species uniformity. The abundance of species is reflected by the length of the curve on the X axis. The more extended on the X axis, the more abundant the species is. Species uniformity is reflected by the shape of the curve. The smoother the curve, the higher the species uniformity.

Analysis method: R packages were used for graph generation based on the results of OTU analysis.



**Figure 3.3.1** Rank-Abundance curve

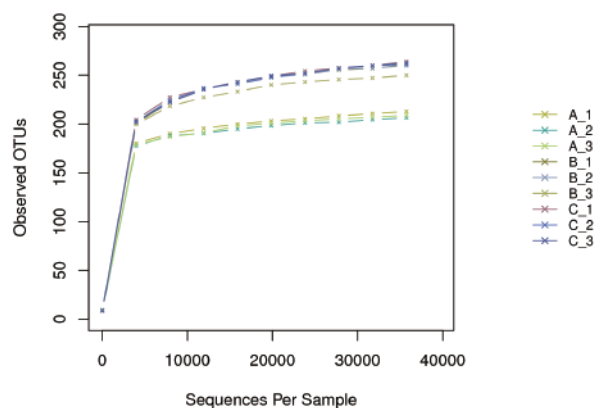
Note: Each curve in the figure above corresponds to an individual sample. The X axis is the relative abundance of the OTU in descending order. The Y axis is the relative abundance of the OTU. '100' on the X axis indicates the OTU in the sample is ranked as the 100th abundant in descending order, and the corresponding value on the Y axis is the percentage of the sequence count in the OTU (the number of sequences of the OTU divided by the total number of sequences).

### 3.4 Rarefaction curve

The rarefaction curve is a useful tool to characterize the species composition of a sample and predicting the abundance of species in a sample. It efficiently deals with the increase of detected species due to the increase in sample size. It is widely used in biodiversity and community surveys to determine whether the sample size is sufficient and to estimate the species abundance. Therefore, the rarefaction curve can not only determine whether the sample size is sufficient, but also predict the species abundance when the sample size is sufficient.

Analysis soft: Qiime (1.9.1) and Plot by R

Analysis method: The rarefaction curve was constructed by random sampling. The observed numbers of OTUs were plotted against the number of extracted sequences.



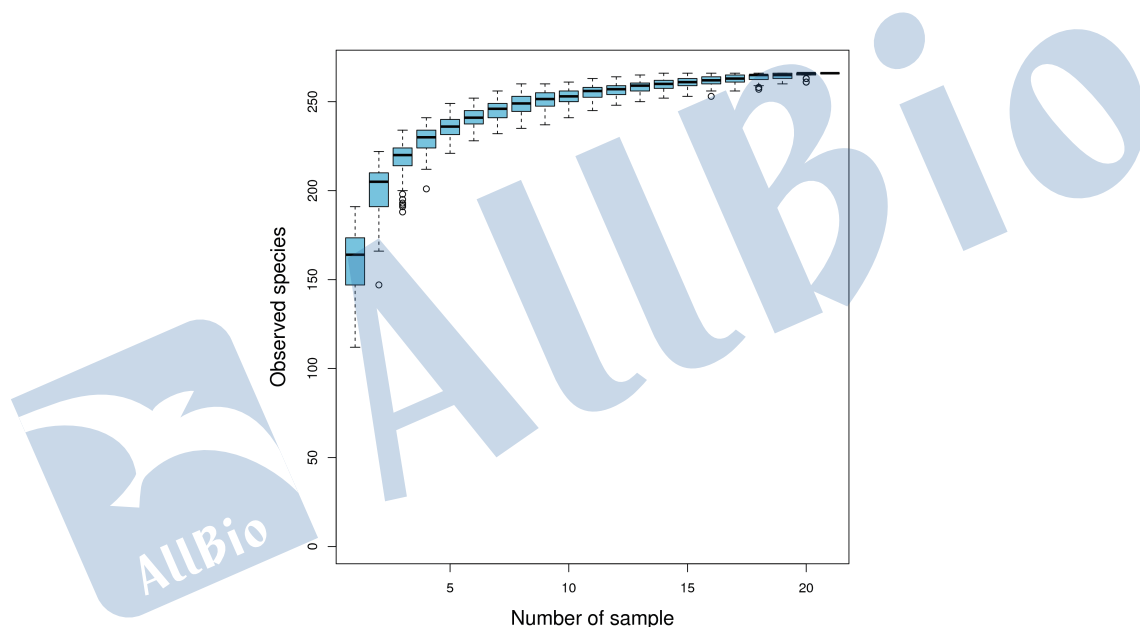
**Figure 3.4.1** Observed OTUs rarefaction curves

Note: The X axis is the number of valid sequences extracted, and the Y axis is the number of OTUs (Observed OTUs). Each sample is represented by one curve with a unique color. The number of OTUs increases with the increase of extracted sequence count until reaching a plateau, which indicates the number of detected OTUs will not increase with the amount of extracted sequences and reflects the reasonable sequence depth.

### 3.5 Species accumulation curve

Species accumulation curve is described as the sample size of the analysis of the increasing species diversity, is in the species composition of the sample and forecast sample species abundance effective tools, in biodiversity and community survey, is widely used in the sample size is sufficient judgment and species richness, species richness) estimates. Cumulative curve by species, therefore, not only can judge whether the sample size to fully, in the premise of enough sample size, the use of species accumulation curve can also forecast species richness (the default is analyzed when sample size is greater than 10).

Analysis software: Statistics and plot by R



**Figure 3.5.1** Species accumulation curve

Note: Abscissa for sample size; Ordinate: OTU number after sampling. The results reflect the continuous sampling is a new OTU rate (new species). Within a certain range, with the increase of sample size, if curve is characterized by a sharp rise in the said in the community has a lot of species have been found; When curve flattens, said the environment of species will not significantly increased with the increase of sample size. Species accumulation curve can be used as the sample size is sufficient judgment, curve has risen dramatically show that sample size is insufficient, need to increase the amount of sampling; On the other hand, suggests that the sampling is enough, can for data analysis.

## 4 Beta Diversity Analysis

### 4.1 Unifrac distance matrix

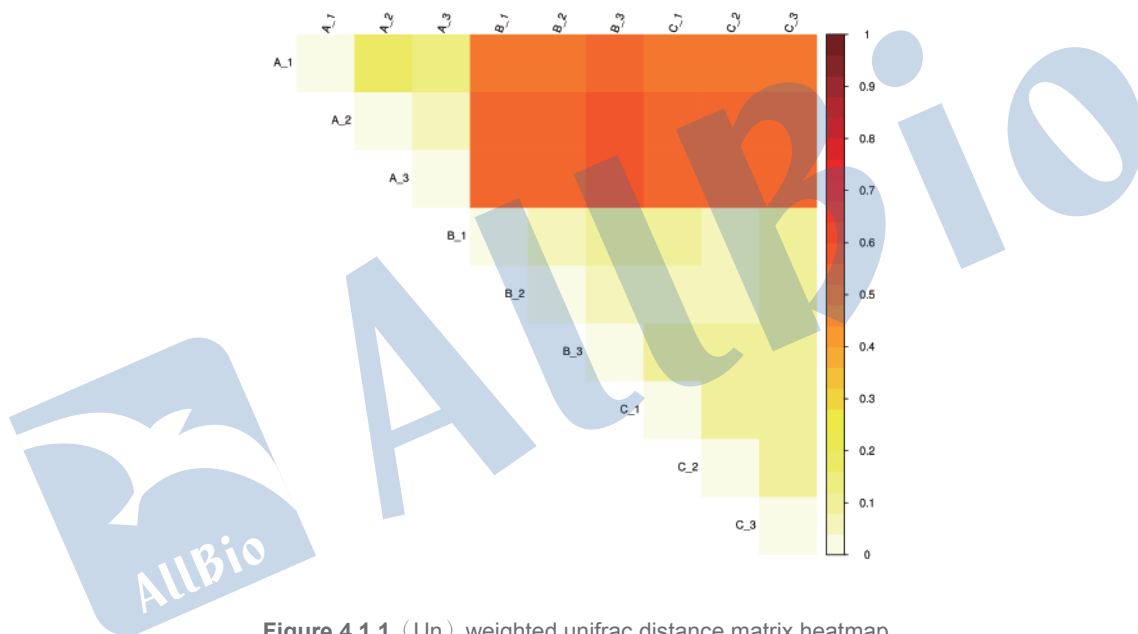
$\beta$  diversity reflects the diversity and the degrees of differences among samples. The distance between the samples can be calculated using

the evolution and abundance information between the sample sequences to reflect whether there is significant difference in microbial community among the samples. This can be achieved by UniFrac analysis.

Analysis software: Qiime (1.9.1) and Plot by R

Analysis method: a phylogenetic tree was constructed using the OTU representative sequences from different environmental samples. The Unifrac metric was then used to measure the difference between two different environmental samples according to the length of the constructed evolutionary tree.

UniFrac analysis includes weighted unifrac and unweighted unifrac methods. The difference between the two is whether to include the relative abundance of sequences from different environmental samples. The weighted unifrac algorithm weights the sequence abundance information when calculating the length of the branch, so unweighted unifrac detects the change among the samples, and the weighted unifrac further quantifies the variation on different pedigrees.



**Figure 4.1.1** (Un) weighted unifrac distance matrix heatmap

Note: (Un)weighted unifrac distance matrix heat map. The color scheme in the heatmap represents the degree of difference between the two samples. The lighter the color, the smaller the coefficient between the two samples, and the smaller the difference of species diversity.

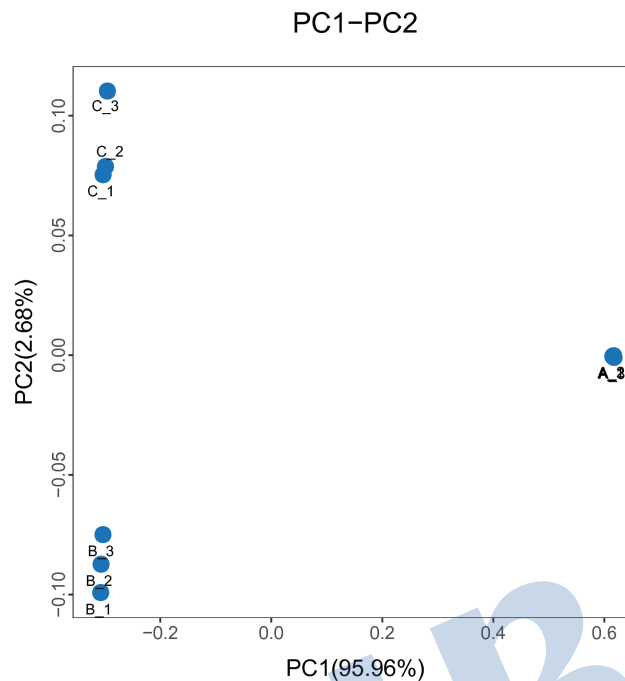
## 4.2 PCoA analysis

PCoA (Principal Co-ordinates Analysis) analysis characterizes and visualizes the similarity and difference of data. Similar to PCA, it sorts data based on a series of eigenvalues and eigenvectors and uses the top ranked eigenvalues to determine the most important coordinates in the distance matrix. It also uses eigenanalysis to perform a rigid rotation of the original axes that only changes the coordinates but not the positional relationship of different sample points. The difference between the two is PCA determines the principal components based on sample similarity coefficient matrix whereas PCoA uses distance matrix to find the primary coordinates.

Analysis software: R



Analysis method: PCoA analysis was performed and plotted based on Brary-Curtis distance matrix.



**Figure 4.2.1** PCoA plot

Note: Samples of the same group are represented in the same color and shape. PC1\_vs\_PC2 is the PCoA plot obtained for the first and second principal coordinates; the X and Y axes represent the first and second principal coordinates, respectively. The value in percentage in the axis label represents the contribution of the corresponding coordinate to the sample variance and measures how much this principal is extracted from the original information. The distance between the sample points indicates the similarity of the microbial community in the sample. The closer the points, the higher the similarity. Samples clustered together are composed of similar microbial compositions.

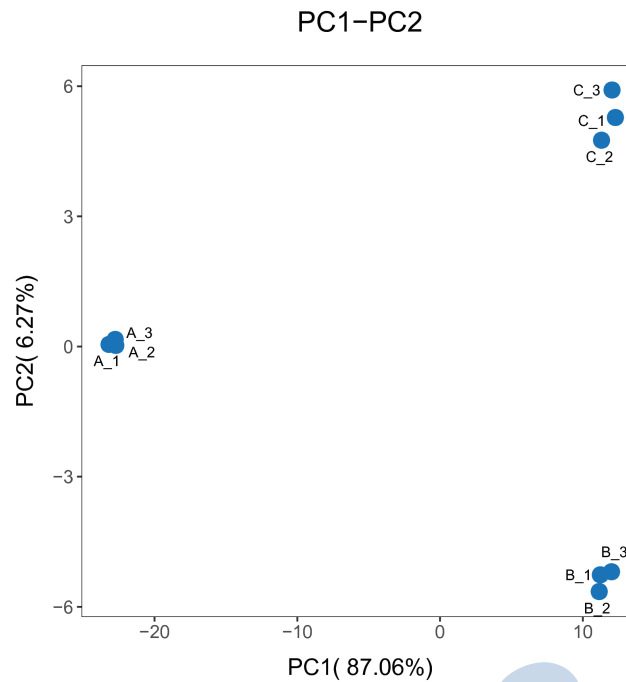
### 4.3 PCA Analysis

PCA analysis (Principal Component Analysis) is a statistical technique for the determination of the key variables in a multidimensional data set that are most responsible for the differences in the observations, and thus is commonly used to simplify complex data analysis.

The difference and distance between samples can be reflected by the analysis of the gene functional distribution of different samples. The differences between multiple sets of data can be plotted on a two-dimensional chart using variance decomposition, with the axes representing two eigenvalues that reflects the largest variance v

Analysis software: R

Analysis method: PCA analysis was performed and plotted based on Brary-Curtis distance matrix



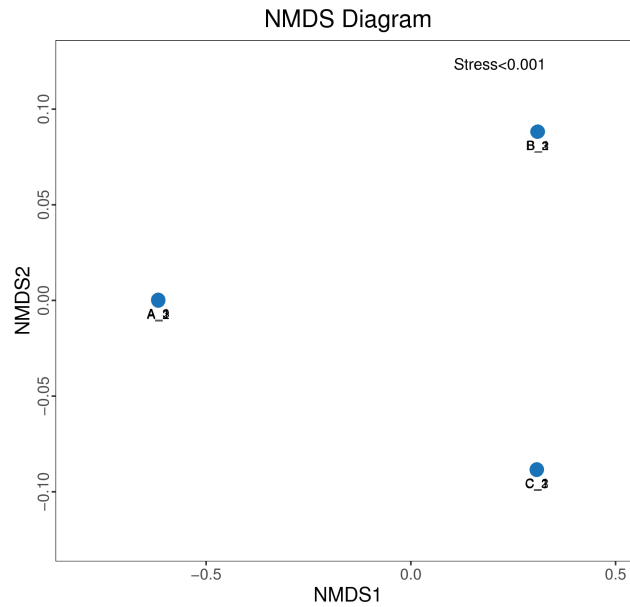
**Figure 4.3.1** PCA plot

Note: PC1, PC2, PC3 represent the first, second and third principal components, respectively. The percentage after the principal component represents the contribution rate of this component to sample difference and measures how much information the principal component can extract from the original data. The distance between samples indicates the similarity of the distribution of functional classifications in the sample. The closer the distance, the higher the similarity.

## 4.4 NMDS Analysis

The non-metric multidimensional scaling is a data analysis method that reduces multi-dimensional space to low-dimensional space to simplify the localization, analysis and classification of research objects. This method preserves the primitive relation among the objects. Its main feature is to position each object in multi-dimensional space based on its functional classification information and calculate the distances between different objects (points) as a measurement of their difference, which are used to obtain the spatial position map.

Analysis method: Graph was generated using vegan package in R based on the beta diversity distance matrix.



**Figure 4.4.1** NMDS plot

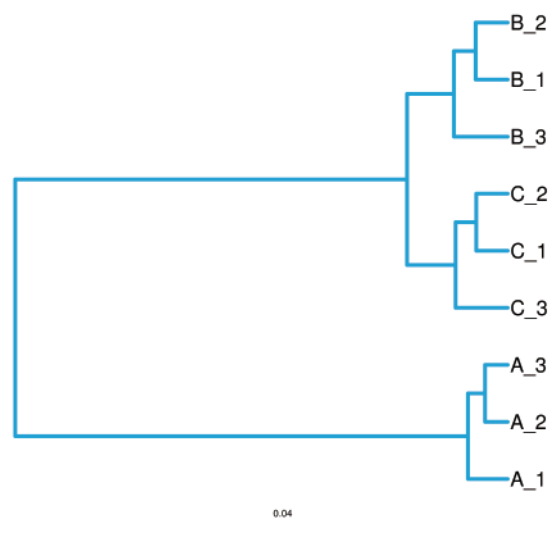
Note: Each point represents a sample, and the distance between the points indicates the degree of difference. Samples of the same group are represented by the same color. Stress < 0.2 indicates NMDS can accurately reflect the difference between the samples.

## 4.5 UPGMA-Tree Cluster Analysis

Cluster analysis uses evolutionary information derived from sample sequences to calculate whether samples in a specific environment is significantly different from a evolutionary lineage in microbial communities.

Analysis software: R

Analysis method: The UPGMA (Unweighted pair group method with arithmetic mean) clustering method was used to cluster the samples based on the Brary-Curtis distance matrices.



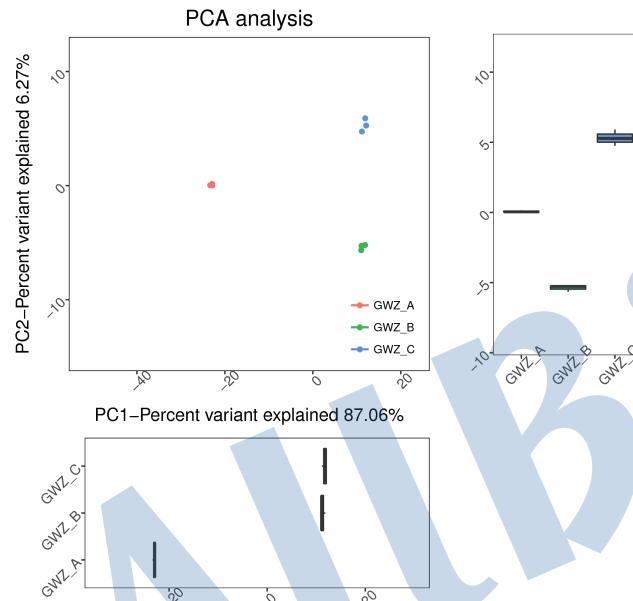
**Figure 4.5.1** UPGMA tree plot

Note: Each branch in the figure represents a sample. Different colors representing different groups.

## 4.6 PCA\_Box Analysis

Between different environmental samples will represent the distribution of dispersion and aggregation, PCA results explain sample difference degree is the highest in two or three components can be used to verify this hypothesis factors.

Analysis software: R



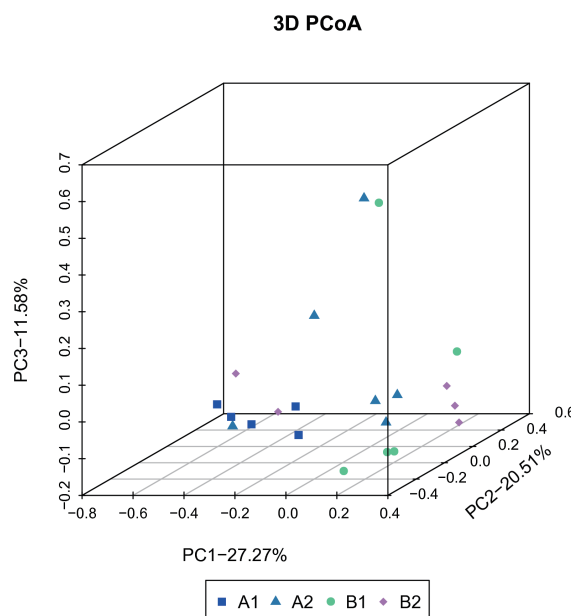
**Figure 4.6.1** PCA\_Box plot

Note: Differentiate between samples according to the group. Different groups are horizontal and vertical box figure is on the first and second axes value distribution.

## 4.7 3D-PCoA Analysis

PCoA (principal co-ordinates analysis) is a kind of data visualization method of the similarities or differences through a series of characteristic value or characteristic vector after sorting, selection in the top of several main characteristic values of PCoA distance matrix can be found the main, the coordinates of the results obtained by the rotation of the data matrix, only change the coordinate system, don't change position relationship between sample points.

Analysis software: R



**Figure 4.7.1** 3D-PCoA plot

Note: Diagram of different color or shape point representing different groups of sample group, scale of transverse and longitudinal axis is relative distance, no practical significance. PC1, PC2 respectively in two groups of samples of microbial components offset suspected the influence factors of the need to be combined with the feature of sample information.

## 5 Variation analysis Between groups

### 5.1 Group difference evaluation by Anosim

ANOSIM is a nonparametric test used to examine whether the difference between two or more groups is statistically greater than the within-group difference. It was used to determine whether the grouping is meaningful. R value was obtained by analyzing the sample distance matrix.

Analysis software: R

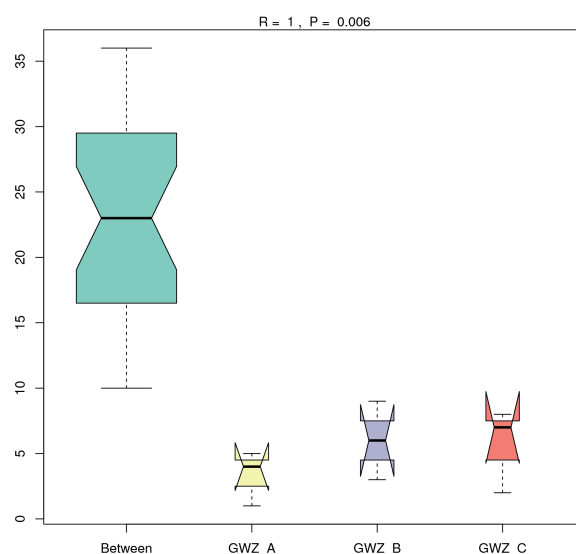
**Table 5.1.1** Group difference evaluation by Anosim

Factor	R-Value	P-Value
Group	1	0.006

#### Column description :

- (1) Factor: The factors used for grouping
- (2) R value: The value of R ranges from 0 to 1. The closer it is to 0, the less significant the between-group difference is compared to within-group difference; the closer it is to 1, the more significance the between-group difference compared to within-group difference.
- (3) P value: P-value indicates the statistical credibility,  $P < 0.05$  indicates the result is statistically significant.

The results of the Anosim analysis were ranked by the value of the distances between the two samples. For each group pair, three set of data can be obtained: between-group distances and within-group distances of each of the two groups. The result is shown below in box plot.



**Figure 5.1.1** Group comparison by Anosim analysis

Note: the distance Between the samples of ordinate rank, abscissa: Between results Between the two groups, the other two results for each group. If two non-overlapping cases of grooves, median suggests that they have significant differences. Box of top and bottom Angle represents the median rank value with more than threshold, the greater the Angle on behalf of outliers. Figure in the R value is close to 1 indicates the difference between groups is greater than the differences in the group,  $P < 0.05$  said statistics have significant.

## 5.2 Adonis analysis

Adonis is a function to perform permutational multivariate analysis of variance, and it is directly analogous to MANOVA (multivariate analysis of variance). It uses the distance matrix to partition sums of squared deviations and analyze the relevance of each grouping factor.

Significance tests are carried out using F-tests based on sequential sums of squares from permutations of the raw data.

Analysis software: R

**Table 5.2.1** Adonis between-group differential analysis

Factor	df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Group	2	1.73454605685859	0.867273028429295	259.681434087077	0.988579322286609	0.00497512437810945

### Column description :

- (1) Factor: The grouping factor in the test
- (2) df: Indicates the degree of freedom
- (3) SumsOfSqs: Total variance, also known from Ward
- (4) MeanSqs: Mean square (difference) that SumsOfSqs / Df
- (5) F.Model: F. test value
- (6) R2: Represent different interpretation of the sample group differences, namely grouping variance ratio of the total variance, R2 greater the higher the degree of differences in interpretation of the packet;
- (7) Pr(>F): Indicates P values less than 0.05 Description of the surveys can be high.

## 5.3 Differential analysis by Metastats

Differential analysis on species composition between groups can be performed based on differential abundance between different groups. Differences in abundance of microbial communities in two groups can be evaluated using strict statistical methods. The multiple hypothesis test and the false discovery rate (FDR) of the rare frequency data were performed to evaluate the statistical significance of the observed difference. This analysis can be carried out at different taxonomic levels, such as Kingdom, Phylum, Class, Order, and Family.

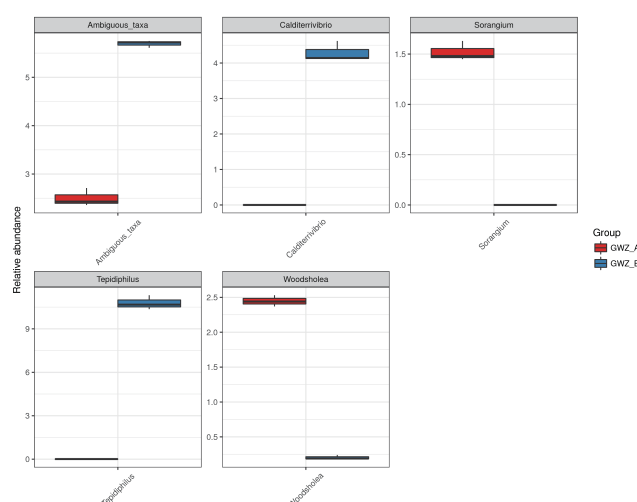
Analysis software: R

**Table 5.3.1** Differentially abundant features

Taxon	Group1_mean	Group1_variance	Group1_standard	Group2_mean	Group2_variance
Unclassified	0.580969587504357	4.80113180644139e-05	0.00400047155822135	0.387657562074362	0.000189778016652239
Tepidiphilus	6.66566683330334e-05	3.33233359993002e-09	3.33283342914605e-05	0.107810916379863	2.37045472408416e-05
Ambiguous_taxa	0.0249959040621966	3.43431901352253e-06	0.00106994065466619	0.0569221015906292	5.68410578841593e-07
Mesotoga	0	0	0	0.0402259452548503	4.45704761540994e-05
Blvii28_wastewater-sludge_group	0	0	0	0.0440241348557752	0.000656063450881351

### Column description :

- (1) Taxon: Strain classification
- (2) mean:Mean
- (3) variance:Variance;
- (4) standard: Standard deviation;
- (5) p value: False positive probability, usually p value <0.05 is considered statistically significant
- (6) q value: False-positive rate of the assessed value, also refers to the credibility of this calculation.



**Figure 5.3.1** Comparison of strain differential abundance



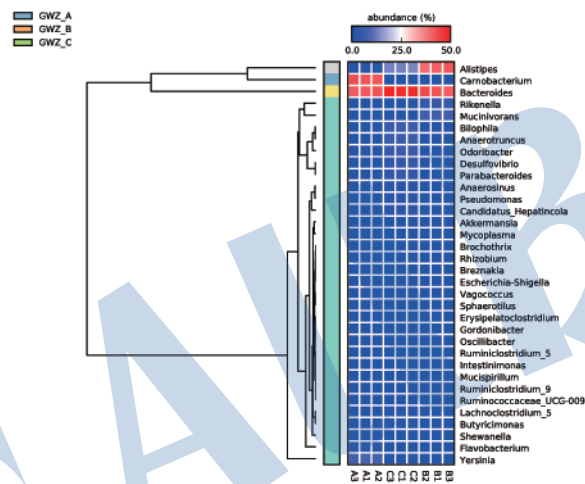
Note: Shown are the abundance distributions of the five strains with the largest between-group difference. The X axis indicates the names of the five strains and the Y axis represents the relative abundance of each.

## 5.4 STAMP analysis

The difference in strain abundance between different samples was analyzed using STAMP (The default setting for two-sample comparison is [Fisher's exact test](#), the default setting for two-group comparison is [Welch's t-test](#), the default setting for multi-sample comparison is [ANOVA](#), P value threshold of the significance test is 0.05 by default).

Differential abundance analysis was performed to compare bacteria abundances between two samples or between two groups of samples. It provides information on the strains of differential abundance between groups as well as their preferred environmental conditions.

Analysis software: STAMP



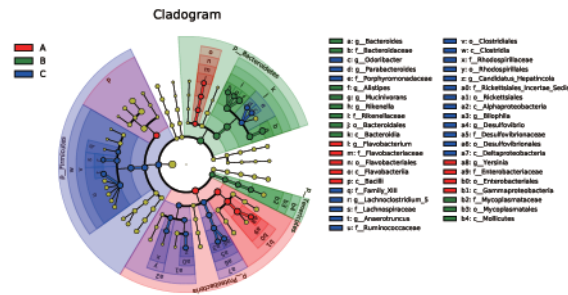
**Figure 5.4.1** Comparison of strain abundance between two samples or groups

The figure on the left shows the percent abundance of the strains in the two samples. The figure in the middle shows the proportion abundance of the strains within the 95% confidence interval. P values are on the right and  $p < 0.05$  indicates statistical significance. The figure on the left shows the percent abundance of the strains in the two samples. The figure in the middle shows the proportion abundance of the strains within the 95% confidence interval. P values are on the right and  $p < 0.05$  indicates statistical significance.

## 5.5 LEfSE analysis

LefSe (LDA Effect Size) is an analytical method to identify high-dimensional biomarkers and reveal genomic characteristics, including gene, metabolic and taxonomic categories. These biomarkers or genomic features can be used to characterize the differences between two or more biological conditions or groups.

Analysis software: LEfSE (online analysis) [http://huttenhower.sph.harvard.edu/galaxy/root?tool\\_id=lefse\\_upload](http://huttenhower.sph.harvard.edu/galaxy/root?tool_id=lefse_upload)



**Figure 5.5.1** LEfSE plot

Note: The figure on the left shows the categories of species that are significantly different between two groups as well as the LDA score from LDA analysis. The figure on the right is a cladogram that indicate the evolutionary relationships of different species. Nodes of different background colors (red or green) indicates different groups. The nodes in red represents microorganism groups that play an important role in the red group; the nodes in green represents microorganism groups that play an important role in the green group; the nodes in yellow represents the groups that do not play an important role in either group. The names of the species represented by the letters in the figure are shown on the right.

## 5.6 Wilcoxon Rank sum test

Wilcoxon rank-sum test , also is Mann-Whitney U test, is one of the two groups of independent samples nonparametric test method. The null hypothesis for two groups of independent samples from no significant difference between two population distribution, based on the research of the two groups of sample average rank to realize the differences determine whether two overall distribution, the analysis of the two groups of samples of species can be significant difference analysis, and the p value calculation false discovery rate (FDR) q value.

Analysis software: R

**Table 5.6.1** the results of Wilcoxon Rank sum test

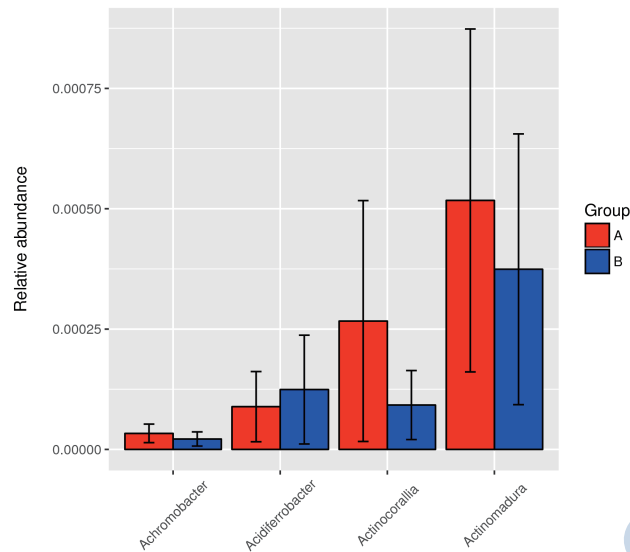
Taxon	Group1_mean	Group1_variance	Group1_standard	Group2_mean	Group2_variance
Unclassified	0.654875586245029	5.35140576090131e-05	0.00422350792623908	0.563311602684589	3.784318001553e-05
Arthrobacter	0.0240707740841431	1.0173481758658e-05	0.00184151040893592	0.0429173215923045	0.00010363661238876
Sphingomonas	0.0248374540728061	2.34283304652106e-06	0.000883710557162441	0.0399840620373353	1.5729063541286e-06
Gaiella	0.0148690804267464	3.44146239678926e-06	0.00107105281488033	0.01157172588705	1.21622546724092e-06
Steroidobacter	0.0112018636683966	2.98993729705554e-08	9.98321474118021e-05	0.012605456065471	2.80039274329776e-07

### Column description :

- (1) Taxon: Strain classification
- (2) mean: Mean
- (3) variance: Variance;
- (4) standard: Standard deviation;

(5) p value: False positive probability, usually p value <0.05 is considered statistically significant

(6) q value: False-positive rate of the assessed value, also refers to the credibility of this calculation.



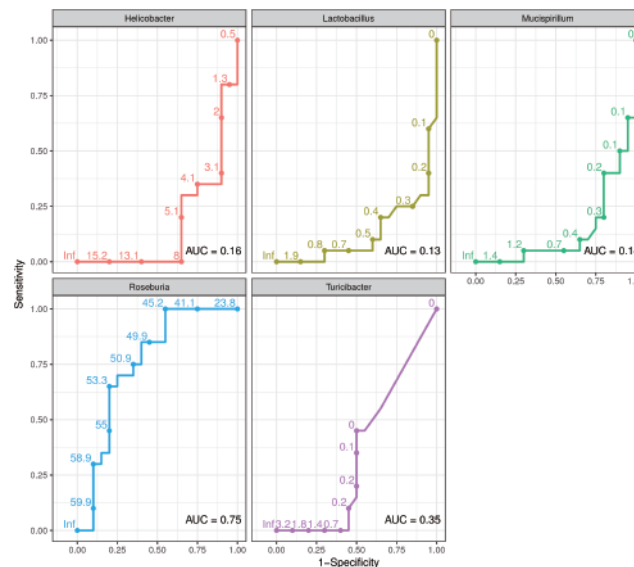
**Figure 5.6.1** Histogram between-group differential analysis

Note: Scheme respectively, two groups of sample average relative abundance of species, sd, respectively, two groups of sample standard deviation of relative abundance of species. P values for the two groups of testing the null hypothesis is true probability value,  $P < 0.05$  said differences,  $P < 0.01$  showed significant differences, q is a false discovery rate.

## 5.7 ROC curve Analysis

receive operating characteristic curve is reflecting the sensitivity and specificity of continuous variable comprehensive indexes, through the composition method to reveal the relationship of the sensitivity and specificity. ROC curve continuous variables set out a number of different threshold, and a series of sensitivity and specificity, is calculated and sensitivity as ordinate and abscissa (1 - specificity) to draw into a curve, the area under the curve, the greater the diagnosis accuracy is higher.

Analysis software: R



**Figure 5.7.1** Different genus of ROC diagram

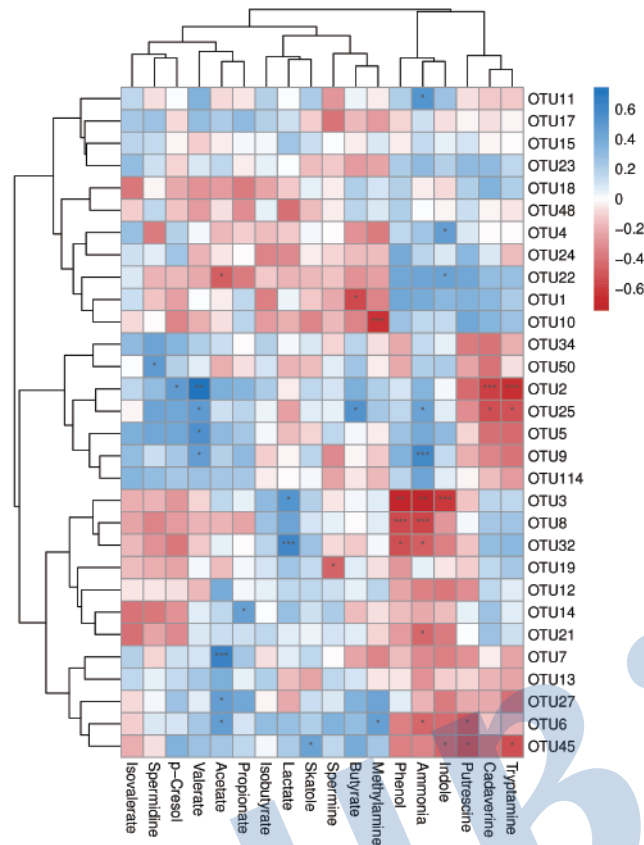
Note: The area under the ROC value between 1.0 and 0.5. In the case of  $AUC > 0.5$ , AUC is close to 1, to diagnose the better the results. AUC in 0.5 to 0.7 with low accuracy, AUC in 0.7-0.9 there is a certain accuracy, AUC in higher accuracy at above 0.9. The  $AUC = 0.5$ , the diagnostic method doesn't work completely, no diagnostic value.  $AUC < 0.5$  do not conform to the reality, seldom appears in practice. On the ROC curve, the point closest to the coordinate figure left for the sensitivity and specificity were higher threshold.

## 6 Other analysis

### 6.1 Correlation analysis of community composition and environmental factors

Correlation analysis uses statistical models to study the correlation between random variables. It investigates whether there is dependency between the phenomena as well as the nature and level of association. The relationship between community composition and environmental factors was analyzed using Spearman correlation coefficient between environmental factors and species or that between environmental factors and OTU. Spearman correlation coefficient is also known as the rank correlation coefficient. It is a linear correlation of the ranking of two variables. It is independent of the distribution of the original variable and is a non-parametric statistical methods with a wide range of applications.

Analysis method: Spearman correlation coefficient was performed using R based on the data of OTU abundance, species abundance and environmental factors. P values from significance test was also obtained. Heatmap was generated to illustrate the relationship between environmental factors and community composition.



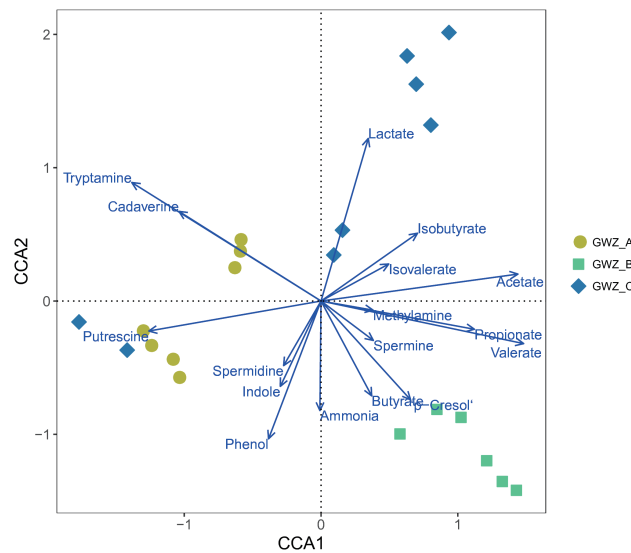
**Figure 6.1.1** Heatmap of Spearman correlation between OTU and environmental factors

Note: the column labels at the bottom represent environmental factors. The labels on the right represents OTU or species. The heatmap is plotted on the Spearman correlation coefficient ( $r$ ) between species or OTU and environmental factors.  $r$  ranges between -1 and 1.  $r > 0$  indicates positive correlation and  $r < 0$  negative correlation. (\* means  $p$  ranges between 0.01 and 0.05, \*\* means  $p$  ranges between 0.001 and 0.01, \*\*\* means  $p$  less than 0.001).

## 6.2 RDA/CCA analysis

RDA / CCA is a ordination method based on correspondence analysis. It couples correspondence analysis with multiple regression analysis, and performs regression analysis on environmental factors at each step of calculation, which is also known as multivariate direct gradient analysis. This analysis is mainly used to investigate the relationship between species distribution or functional classification and environmental factors. RDA is based on a linear model and CCA is based on a single-peak model. This analysis examines the relations across environmental factors, samples, and functional distributions.

Analysis method: package 'vegan' in R was used for analysis and figure generation based on beta diversity distance matrix and data on environmental factors.



**Figure 6.2.1** RDA/CCA plot

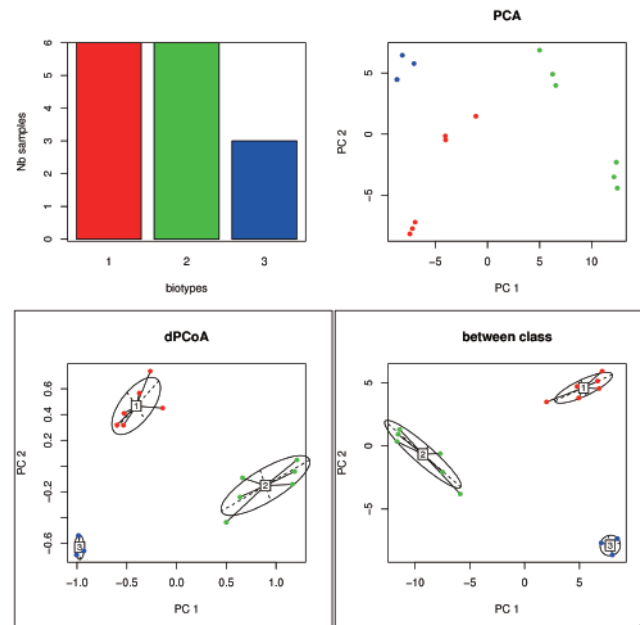
Note: Numbers in the figure indicate sample names, and different colors or shapes indicate groups of samples in different environments or conditions. Arrows represents environmental factors. The angles between the environmental factors represent positive and negative correlations among the environmental factors (Acute angle: positive correlation; obtuse angle: negative correlation; right angle: no correlation). Perpendicular lines were drawn from sample points to environmental factor arrows. The closer the projection points are, the more similar the environmental factor values are between the samples. That is, the environmental factors have similar levels of influence on the samples.

### 6.3 Enterotype analysis

The International Human Microbiome Consortium defined three enterotypes based on the differences in the number and types of bacteria in the human intestinal tract. Researchers named them Bacteroides-, Prevotella-, and Ruminococcus-enterotypes to reflect the dominant strain of bacteria in each ecosystem. Bacteroides obtain energy mainly from carbohydrates and proteins, whereas prevotella and ruminococcus efficiently digest intestinal glycoprotein.

Although the understanding of intestinal type is far than understanding of blood type, but scientists believe that intestinal type information can also provide reference for diagnosis and treatment of diseases. In the intestinal flora, type and quantity can reflect the digestive ability of different people, the immune ability and response to the drug.

Analysis method: Based on the relative abundance of the genus level data clustering analysis (using the JSD distance and divide around the center of the PAM clustering method), according to Calinski - Harabasz index to evaluate the results of cluster analysis and graphics display.

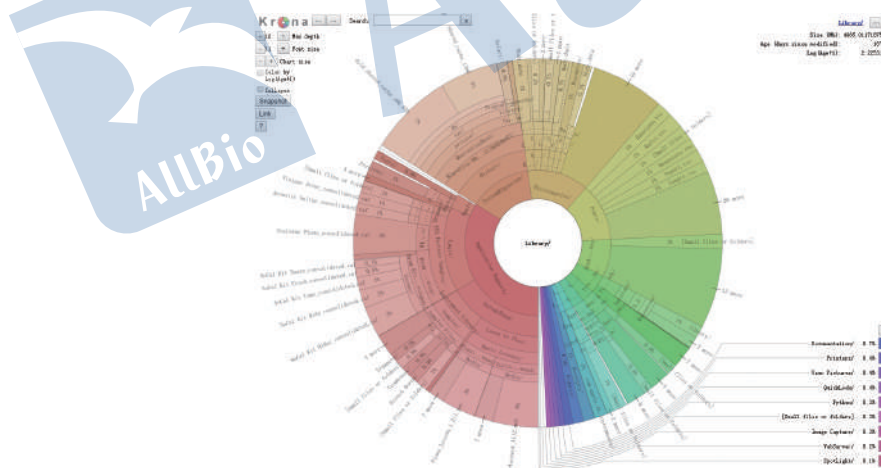


**Figure 6.3.1** Enterotype analysis plot

Note: The picture shows that different colors represent different intestinal types based on the results of intestinal type analysis, and the circle is the range PCAdiagram that conforms to the confidence interval.

## 6.4 Krona species composition diagram

KronaTools (v2.7) software was used to visualize the distribution of different species on different levels. The following is an example of the result generated by Krona. For detailed analysis, please see [krona.html](http://krona.html).



**Figure 6.4.1** Krona Krona species composition diagram (for illustration)

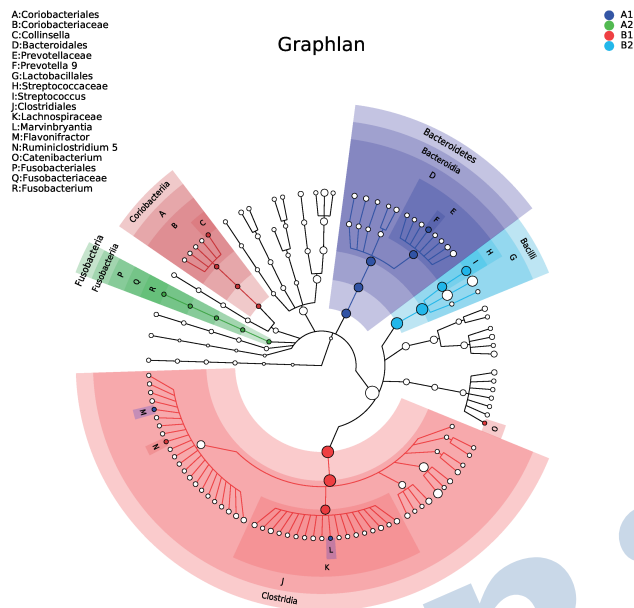
Note: Circles in the figure represent different taxonomic levels, from inside to outside: Kingdom, Phylum, Class, Order, Families, Genus, Species. The size of the area positively correlates with the abundance of the corresponding species composition.

## 6.5 GraPhlAn Analysis

Using GraPhlAn combining OTU Table of a grouping of all samples comments result overall display of OTU species, in order to see the advantage species.



Analysis software: GraPhlAn(online analysis)



**Figure 6.5.1** GraPhlAn plot

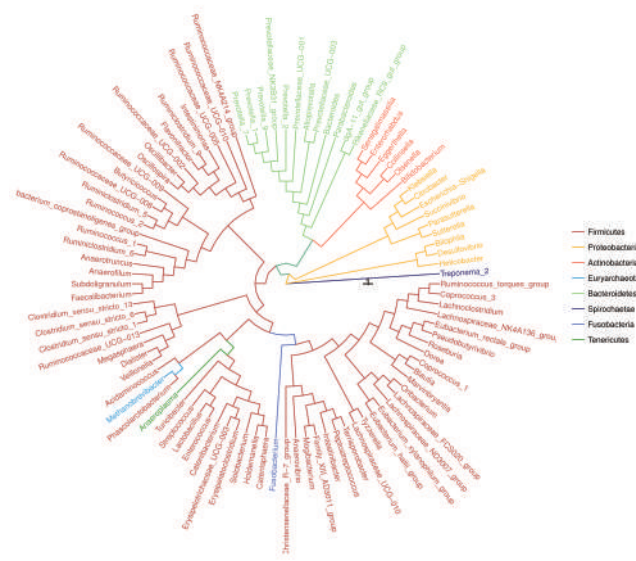
Note: The circle in the figure represents different species classification levels from inside to outside, the size of the circle is proportional to the abundance of species, and different colors represent dominant bacteria in different groups.

## 6.6 Phylogenetic Tree

In the study of molecular evolution. Phylogenetic inference to reveal the evolution process of sequence, understand the history and biological evolution mechanism, can be classified by a level sequence to build bases of the differences between the evolutionary tree.

By selecting the genera level (OTU or a species classification level) the corresponding OTU based on maximum likelihood method to construct the evolutionary tree, using R language map to ring form.

Analysis software: R



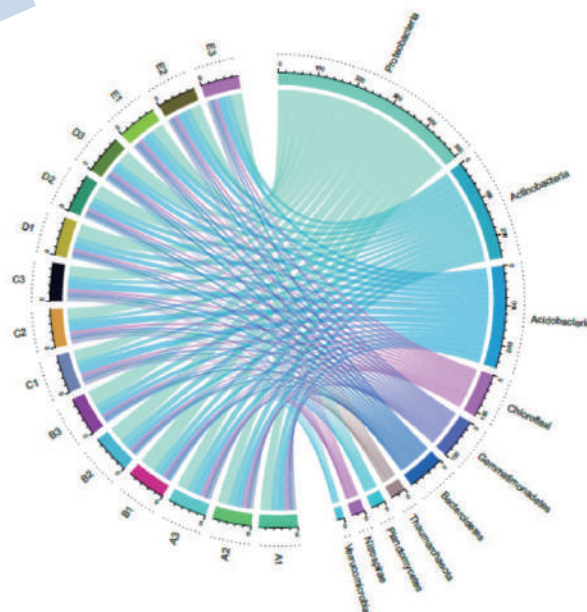
**Figure 6.6.1** Phylogenetic Tree plot

Note: The evolutionary tree of each branch represents a species, numerical length for the evolution of the distance between the two species, the species difference degree.

## 6.7 Collinearity between Samples and species

Samples and species of the chart is a linear relationship between samples and species, visual graph corresponding relationship which not only reflects the advantages of each sample species composition, but also reflects the ratio between the distribution of all the dominant species in different samples.

Analysis method: Circos(<http://circos.ca/software/download/>)



**Figure 6.7.1** plot of Collinearity between Samples and species

Note: the left semicircle (small circle) represents the species abundance composition of the sample in the colinear diagram of the sample and the right semicircle (large circle) represents the distribution proportion of phylum species in different samples. Circle from outside to inside: the first and second color circles: the left half circle represents the species composition corresponding to different samples, different colors represent different species, and the length represents the proportion of a species in the sample; The right half circle represents the distribution proportion of different samples in dominant species, different colors represent different samples, and the length represents the distribution proportion of this sample in a certain species (the percentage shown in the second circle). The third circle: color band within the circle, one end connects the sample (left semicircle), the end width of the band represents the species richness in the sample, the other end connects the species (right semicircle), the end width of the band represents the distribution proportion of the sample in the corresponding species, and the value outside the circle represents the abundance value of the corresponding species.

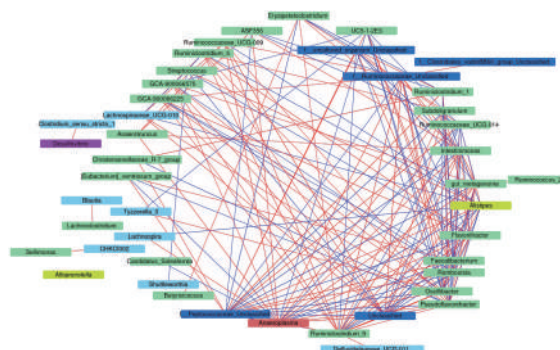
## 6.8 Network Analysis

Total network in the study of complex microbial community structure and function provides a new perspective of environment. Due to the relationship between microbial total different under different environment, total network diagram by species, can directly see the influence of different environmental factors on the microbial adaptation, and an environment of mutual advantage of the dominant species, close interactions of species, the dominant species and the species group tend to maintain the environment of the microbial community structure and function stability plays a unique and important role.

Through the study of the correlation index of all the samples (spearman correlation coefficient of SCC or Pearson correlation coefficient of the PCC) calculation, the correlation coefficient for species form, the absolute value of correlation coefficient in  $\text{cuoff} = 0.6$  filtering, using Cytoscape recombination species abundance to do.

Cytoscape is a molecular interaction network visualization analysis software, based on different species abundance information correlation analysis between the example that exist in the total species in environmental samples can be obtained, by interaction between the species in the same environment, further explain the formation mechanism of phenotypic differences between samples. Of metabolic pathways can also be abundance information of features such as correlation analysis, analysis of the relationship between the function, and the relationship with the connection between the environment, Cytoscape performance species from the visual Angle or metabolic function and the correlation between and among samples, sample classification, chart patterns or find important in information from the network.

Analysis software: Cytoscape(<http://www.cytoscape.org>)



**Figure 6.8.1** Network analysis

Note: Different nodes represent different genera, node size represents the average relative abundance, same color node door, attachment and the thickness of species interactions between nodes correlation coefficient absolute value is first close.



## Appendix

## ① Appendix

### 1 Documents

readme.pdf -- Analysis results directory description document.

method.pdf -- Documentation of experimental and analytical methods.

software.pdf -- Analyze software list documents

FAQ.pdf -- After-sales FAQ documents

### 2 Notes

The result file is recommended to be opened with a professional text editor such as Excel or EditPlus.



## References

## ① References

- [1] JG, Kuczynski J. et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5): 335-336(2010).
- [2] Crawford, P. A., Crowley, J. R., Sambandam, N., Muegge, B. D., Costello, E. K., Hamady, M., et al. (2009). Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. *Proc Natl Acad Sci U S A*, 106(27), 11276-11281.
- [3] Quast C, Priesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. [Opens external link in new window](#) *Nucl. Acids Res.* 41 (D1): D590-D596.
- [4] Yilmaz P, Parfrey LW, Yarza P, Gerken J, Priesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2014) The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. [Opens external link in new window](#) *Nucl. Acids Res.* 42:D643-D648
- [5] Priesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl. Acids Res.* 35:7188-7196
- [6] Klindworth A, Priesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. [Opens external link in new window](#) *Nucl. Acids Res.* 41:e1
- [7] Westram R, Bader K, Priesse E, Kumar Y, Meier H, Glöckner FO, Ludwig W (2011) ARB: a software environment for sequence data. In: de Bruijn FJ (ed) *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*. [Opens external link in new window](#) John Wiley & Sons, Inc., pp 399-406
- [8] Yu Wang, Hua-Fang Sheng, et al. Comparison of the Levels of Bacterial Diversity in Freshwater, Intertidal Wetland, and Marine Sediments by Using Millions of Illumina Tags. *Appl. Environ. Microbiol.* 2012, 78(23):8264. DOI: 10.1128/AEM.01821-12.8
- [9] Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5(3):] e9490. doi:10.1371.journal.pone.0009490.
- [10] Micah Hamady, Catherine Lozupone and Rob Knight. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal* (2010) 4, 17–27; doi:10.1038/ismej.2009.97