Novogene

Amplicon Analysis Demo report

01-Aug-2017

Content

Ⅰ. Overview
Ⅱ. Workflow
1 Preparation and sequencing
2 Data analysis procedures
Ⅲ. Results
1 Sequencing data processing
2 OTU analysis and species annotation
2.1 OTU clustering and species annotation
2.2 Species distribution
3 Alpha Diversity analysis
3.1 Alpha Diversity Indices
3.2 Species diversity curves
3.3 Species accumulation boxplot
3.4 Venn and Flower diagram
3.5 Similarity analysis on alpha diversity indices
4 Beta Diversity Analysis
4.1 Beta Diversity Indices
4.2 Principal Coordinates Analysis(PCoA)
4.3 Principal component analysis (PCA)
4.4 Non-metric multidimensional scaling (NMDS)
4.5 Unweighted Pair Group Method with Arithmetic Mean(UPGMA) Clustering Analysis
4.6 Significance test of differences in community composition between groups
4.7 Variance analysis in species between groups
5 Data Mining
Ⅳ. Methods
Ⅴ. References
Ⅵ. Appendix

Ⅰ. Overview

The 16S ribosomal RNA (rRNA) sequence is composed of nine hypervariable regions interspersed with conserved regions. The bacterial 16S gene contains nine hypervariable regions (V1-V9) ranging from about 30-100 base pairs long that are involved in the secondary structure of the small ribosomal subunit. The degree of conservation varies widely between hypervariable regions, with more conserved regions correlating to higher-level taxonomy and less conserved regions to lower levels, such as genus and species. For taxonomic classification, it is sufficient to sequence individual hypervariable regions instead of the entire gene. Additionally, the 16S gene contains highly conserved sequences between hypervariable regions, enabling the design of universal primers. With the development of high-throughput sequencing platforms, sequence variation in the 16S gene is widely used to characterize diverse microbial communities[1,2,3].

Based on pair-end algorithms, amplicon sequencing is conducted on Illumina HiSeq platform.

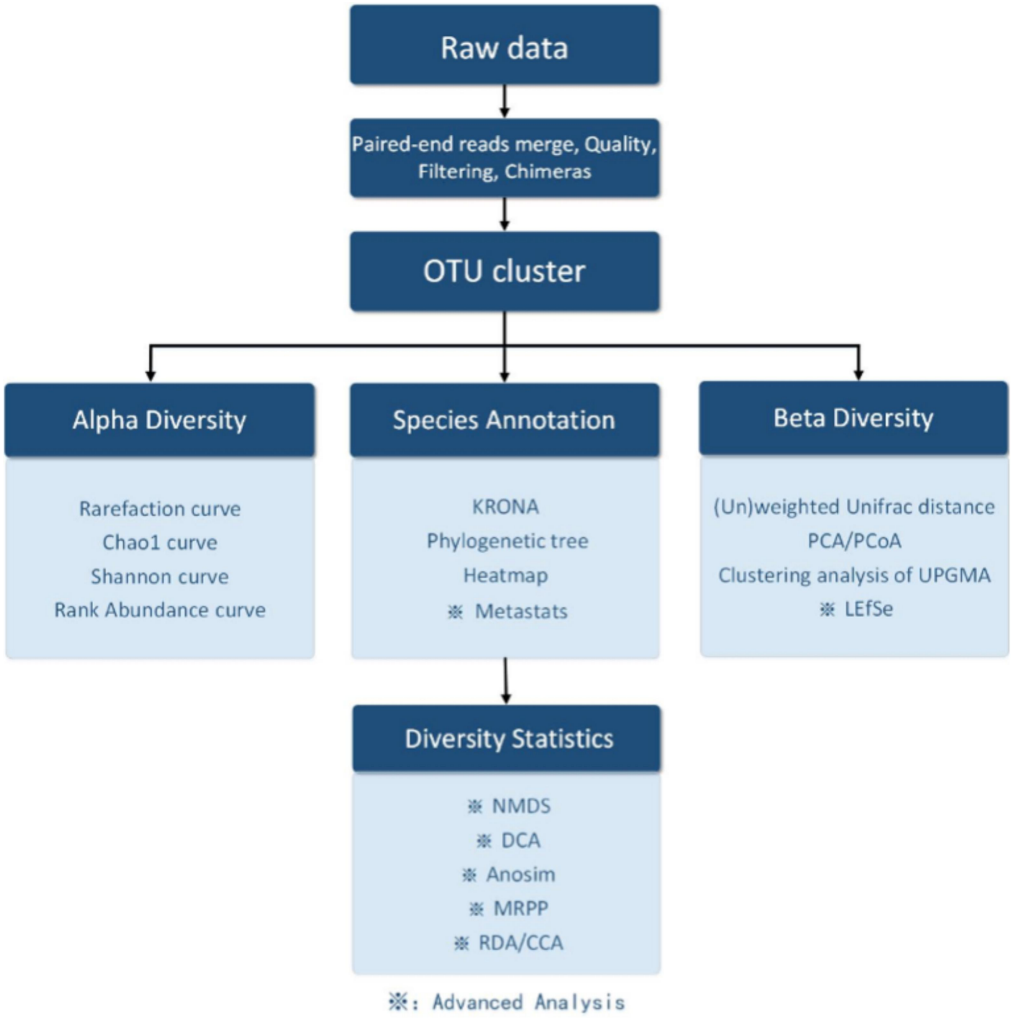
Ⅱ. Workflow

1 Preparation and sequencing

From the raw DNA samples to the final data, each step, including sample test, PCR, library preparation and sequencing, influences the quality of the data, and data quality directly impacts the analysis results. To keep the reliability of the data, quality control (QC) is performed at each step of the procedure. The workflow is as follows：

2 Data analysis procedures

For the veracity of sequencing data analysis, raw data would be merged and filtered to get clean data. The effective data is used to do OTU cluster and species annotation for the respective sequence of each OTU. Thus the relative species, evenness and abundance distribution can be analyzed with Alpha diversity, Beta diversity, venn or flower graph et al. Furthermore, OTU picking, taxonomic assignment construction of phylogenetic trees through downstream statistical analysis explain the community construction differences between samples or among groups via PCoA, PCA and NMDS. Statistic methods such as T-test、MetaStat、LEfSe、 Anosim and MRPP could test the significance of community composition and structure differences between groups. Meanwhile, the analysis result combines CCA/RDA to explore the major environmental factors. The workflow of data analysis shows as follow:



Notes: If samples number is less than 3, Beta diversity, significance test of community construction difference between groups, and environmental association analysis can not be proceeded. Significance test of community construction difference between groups and environmental association analysis can not proceed without grouping, or when repetition is less than 3. Environmental association analysis need the environmental factor data which clients offer.

Ⅲ. Results

1 Sequencing data processing

Amplicon was sequencing on Illumina HiSeq paired-end platform to generate 250bp paired-end raw reads (Raw PE), and then assembled and pretreated to obtain Clean Tags. The chimeric sequences in Clean Tags were detected and removed to obtain the Effective Tags finally. The data output has shown in table 1.

Table.1 QC stat

Data preprocessing and QC stat

Sample Name	Raw PE(#)	Raw Tags(#)	Clean Tags(#)	Effective Tags(#)	Base(nt)	AvgLen(nt)	Q20	Q30	GC%	Effective%
JG1	84,879	76,466	63,933	62,414	26,708,616	428	98.41	96.87	53.70	73.53
JG2	86,533	76,173	63,732	62,300	26,675,293	428	98.32	96.68	54.40	72.00
JG4	84,273	74,741	62,096	60,179	25,750,454	428	98.36	96.76	53.02	71.41
JG7	50,746	44,524	38,176	37,036	15,552,348	420	98.37	96.81	52.91	72.98
JPG7	92,620	81,079	69,761	68,017	28,791,674	423	98.40	96.88	50.65	73.44
JPG14	97,112	85,490	75,569	72,697	30,450,885	419	98.50	97.06	51.01	74.86
JPG21	83,269	74,583	66,106	62,856	26,463,582	421	98.57	97.20	51.66	75.49
JPG28	95,984	85,536	75,059	71,830	30,274,441	421	98.54	97.18	51.21	74.84
LG1	86,230	76,556	64,553	60,640	25,865,292	427	98.36	96.78	53.50	70.32
LG2	89,909	79,898	68,004	64,863	27,723,123	427	98.44	96.94	53.79	72.14

Notes: Raw PE means PE reads; Raw Tags means tags merged from PE reads; Clean Tags means tags after filtering; Effective Tags means tags after filtering chimera; Base means base number of Effective tags; AvgLen means average length of Effective Tags; Q20 and Q30 mean the precentage of base quantity that greater than 20 and 30; GC (%) means GC content in Effective Tags; Effective (%) means the percentage of Effective tags in Raw PE.

Results directory：

PE reads with barcode and primer sequence：result/00.RawData/Sample_Name/ *.raw_1(2).fq.gz; ;

PE reads without barcode and primer sequence：result/00.RawData/Sample_Name/ *_1.fq.gz ; result/00.RawData /Sample_Name/ *_2.fq.gz ;

Raw Tags：result/00.RawData/Sample_Name/ *.extendedFrags.fastq ;

Effective Tags：result/01.CleanData/Sample_Name/*.fastq ; result/01.CleanData/Sample_Name/*.fna ;

Barcodes and primers information：result/00.RawData/ SampleSeq_info.xls.

2 OTU analysis and species annotation

In order to analyze the species diversity in each sample, all Effective Tags were grouped by 97% DNA sequence similarity into OTUs(Operational Taxonomic Units)

2.1 OTU clustering and species annotation

2.1.1 Statistic analysis of annotation

During the construction of OTUs, basic information from different samples had been collected, such as Effective Tags data, low-frequency Tags data and annotation data of Tags. The statistical dataset is showed as followed in figure2.1.1 (To view full size picture please click**here**) 。

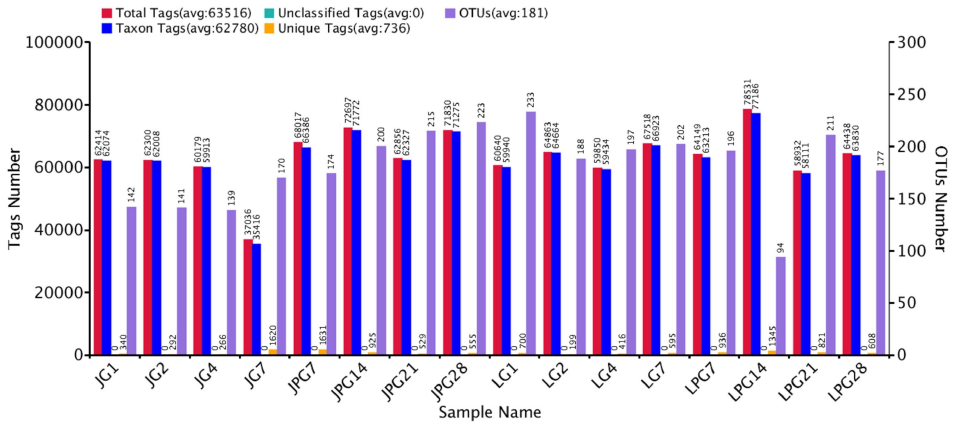


Figure2.1.1 Statistic analysis of the tags and OTUs number of each samples

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

2.1.2 Interactive view of species annotation

The heat-map shows a interactive view of species composition and abundance among different samples in the lived webpage**click here**. An example picture shows as follows:

Figure 2.1.2An example of OTU table heat-map, showing taxonomy assignment for each OTU.

Notes: 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. Click **here**for details

Result directory:

OTU heatmap：result/02.OTUanalysis/taxa_heatmap/otu_heatmap/sorted_otu_table.html.

2.1.3 GraPhlAn dispaly

Tree graph of species annotation for each group were construct by GraPhlAn[4]. The OTU tree of one group shows in figure2.1.3. (Click**here**)to view full size picture.

Figure 2.1.3 OTU annotation tree construct by GraPhlAn

Notes: Different taxonomic ranks range inside out. The size of circles stands for abundance of species. Different colors stand for different phylum. Solid circles stand for the top 40 species in high abundance.

Results directory:

Graphlan figures：result/02.OTUanalysis/GraPhlan/graphlan_*.{png,pdf}.

2.1.4 Taxonomy Tree

Specific species (Showing the top 10 genus in high relative abundance by default) (the top 10 genus in high relative abundance) were selected to make the taxonomy tree[5]by independently R&D software. Taxonomy tree in single sample is shown in Figure 2.1.4-1 (To view full size picture pleaseclick).Taxonomy tree in group is shown in figure 2.1.4-2 (To view full size picture pleaseclick) 。

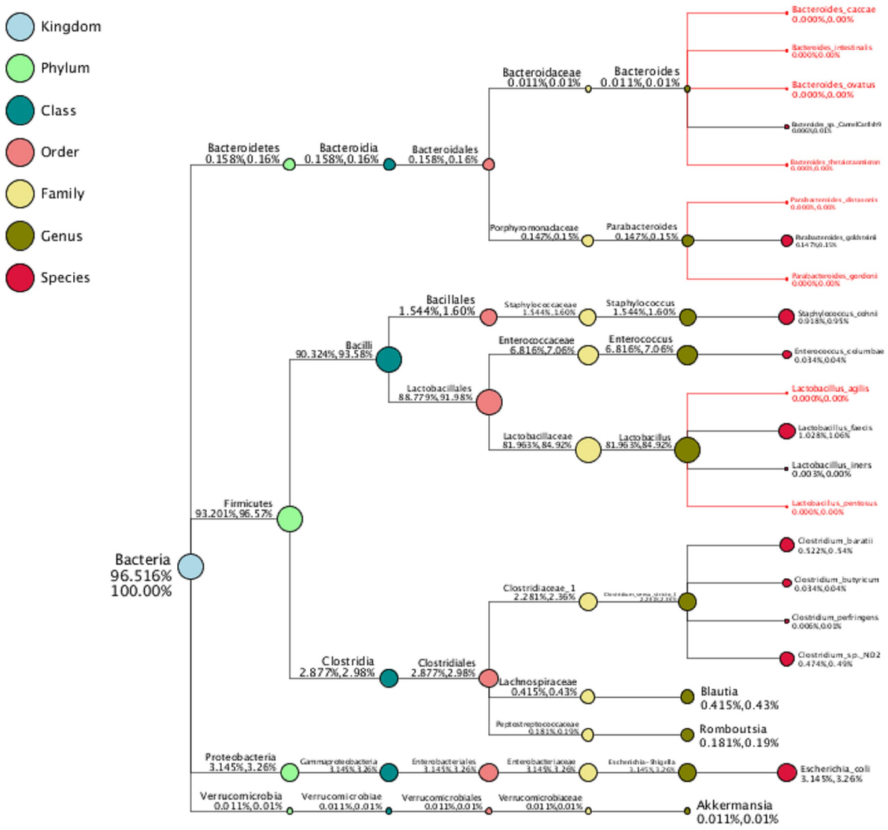


Figure 2.1.4-1 Taxonomy tree in single sample.

Notes: Different colours represent different taxonomic ranks. The size of circles stand for the relative abundance of species. The first number below the taxonomic name represents the percentage in the whole taxon, while the second number represents the percentage in the selected taxon.

Results directory: :

Taxonomy tree in samples：result/02.OTUanalysis/taxa_tree/*.{png,svg},all.taxtree.{png,svg} for all samples，*.taxtree.{png,svg} for single sample.

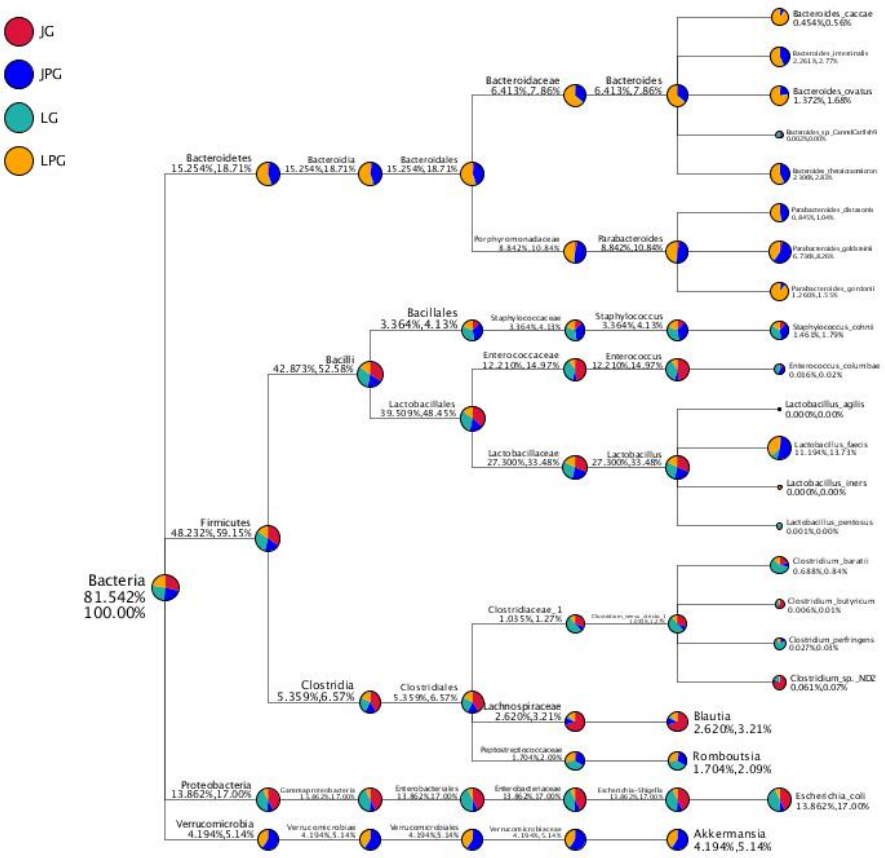


Figure 2.1.4-2 Taxonomy tree in groups.

Notes: Sectors with different colors represent different groups. The size of the sector represents the relative abundance. The first number below the taxonomic name represents the percentage in the whole taxon, while the second number represents the percentage in the selected taxon.

Results directory：

Taxonomy tree in groups：result/02.OTUanalysis/taxa_tree_group/*.{png,svg},all.taxtree.{png,svg} for all groups, *.taxtree.{png,svg} for single group.

2.1.5 Krona Display

KRONA[6]visually displays the analysis result of species annotation. Circles from inside to outside stand for different taxonomic ranks, and the area of sector means respective proportion of different OTU annotation results. More details please click**here**. An example picture shows below:

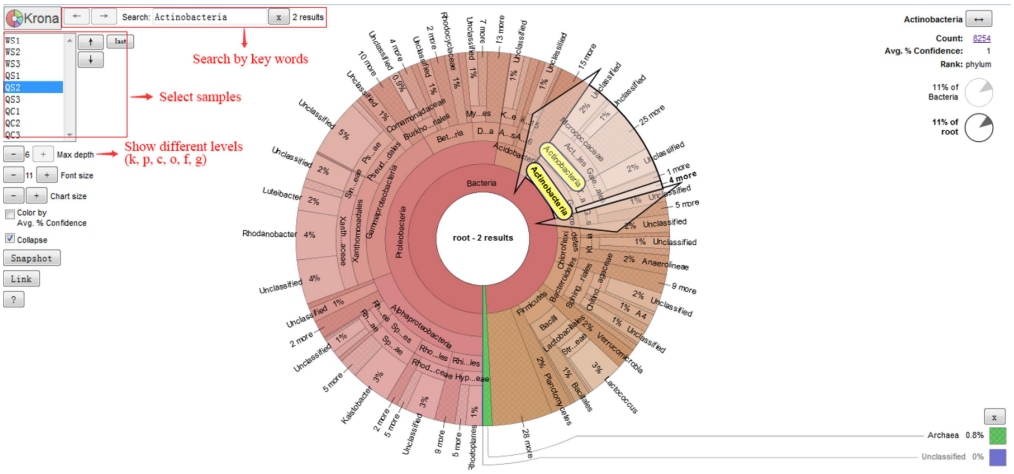


Figure 2.1.5 Krona display

Results directory：

Krona display：result/02.OTUanalysis/all_rep_set_tax_assignments.krona.html。

2.2 Species distribution

2.2.1 Species relative abundance layout

The top 10 species in the different taxonomic ranks were selected to form the distribution histogram of relative abundance. The distribution in phylum was shown as follow: (To view full size image please**click here**)

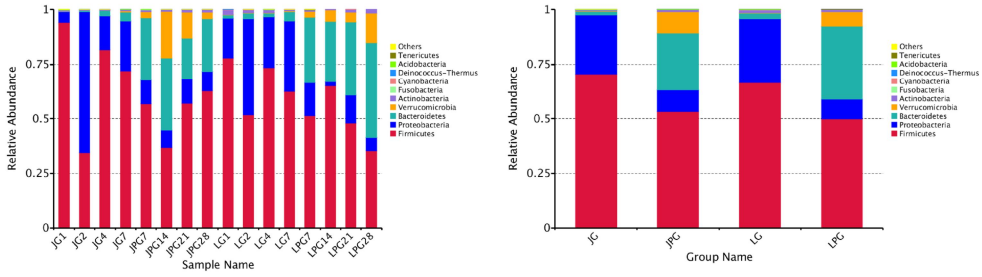


Figure 2.2.1 Species relative abundance in phylum

Notes: 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 phylum.

Results directory：

The top 10 species relative abundance in each taxonomic rank：result/02.OTUanalysis/top10(_group)/；Including Phylum, Class, Order, Family and Genus.

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 2.2.2 （To view full size image please**click**）。

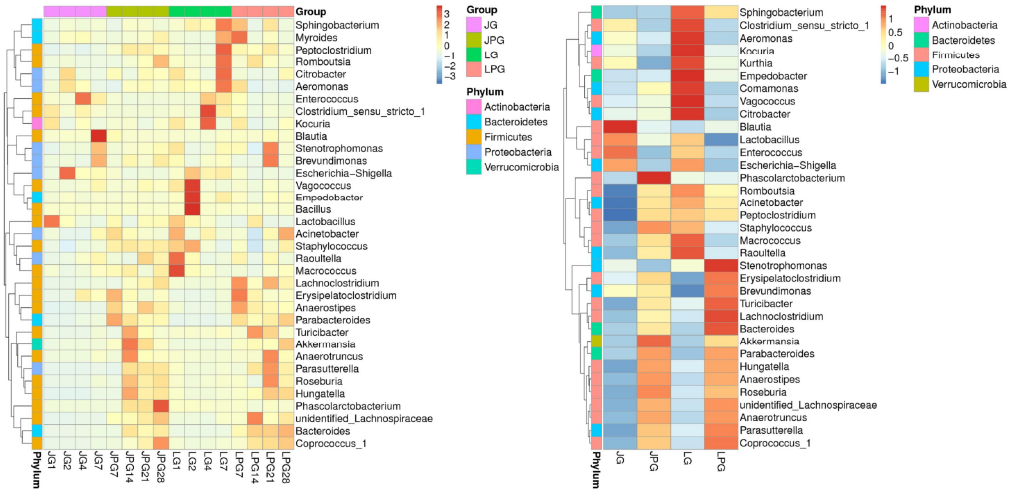


Figure 2.2.2 Species abundance heatmap

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

Results directory：

Species abundance heatmap in each taxonomic rank：result/02.OTUanalysis/taxa_heatmap/cluster/*.{png,pdf}；

Plot data for each heatmap：result/02.OTUanalysis/taxa_heatmap/cluster/*.txt。

2.2.3 Ternary plot

In order to outstand the difference of the dominant species among 3 samples in each taxonomic rank. The top 10 species were selected and the ternary plot [7]was drawn based on reletive abundance. The ternary plot in phylum shows as follow:（To view full size picture please **click**）：

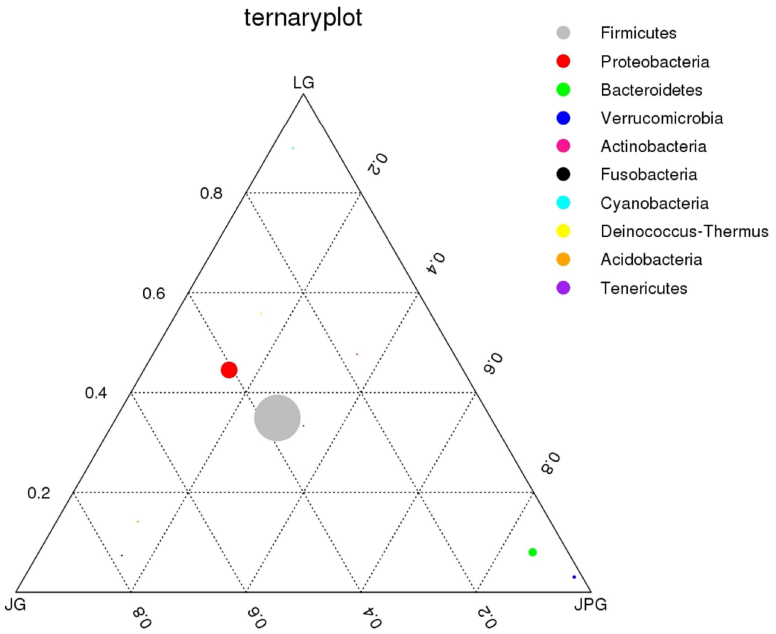


Figure 2.2.3 Ternary plot

Notes: Three vertexes represent three samples. Circles represent dominant species and the size represent the relative abundance. The group which the circle close to has higer abundance of this species.

Results directory：

ternary plot: result/02.OTUanalysis/ternaryplot2/*/*/ternary.{pdf,jpeg}

2.2.4 The evolutionary tree in genus

The top 100 genera were selected and the evolutionary tree were drawn using the aligned represent sequences. The relative abundance of each genus was also displayed beside the genus in figure 2.2.4. (To view full size picture please **click**) ：

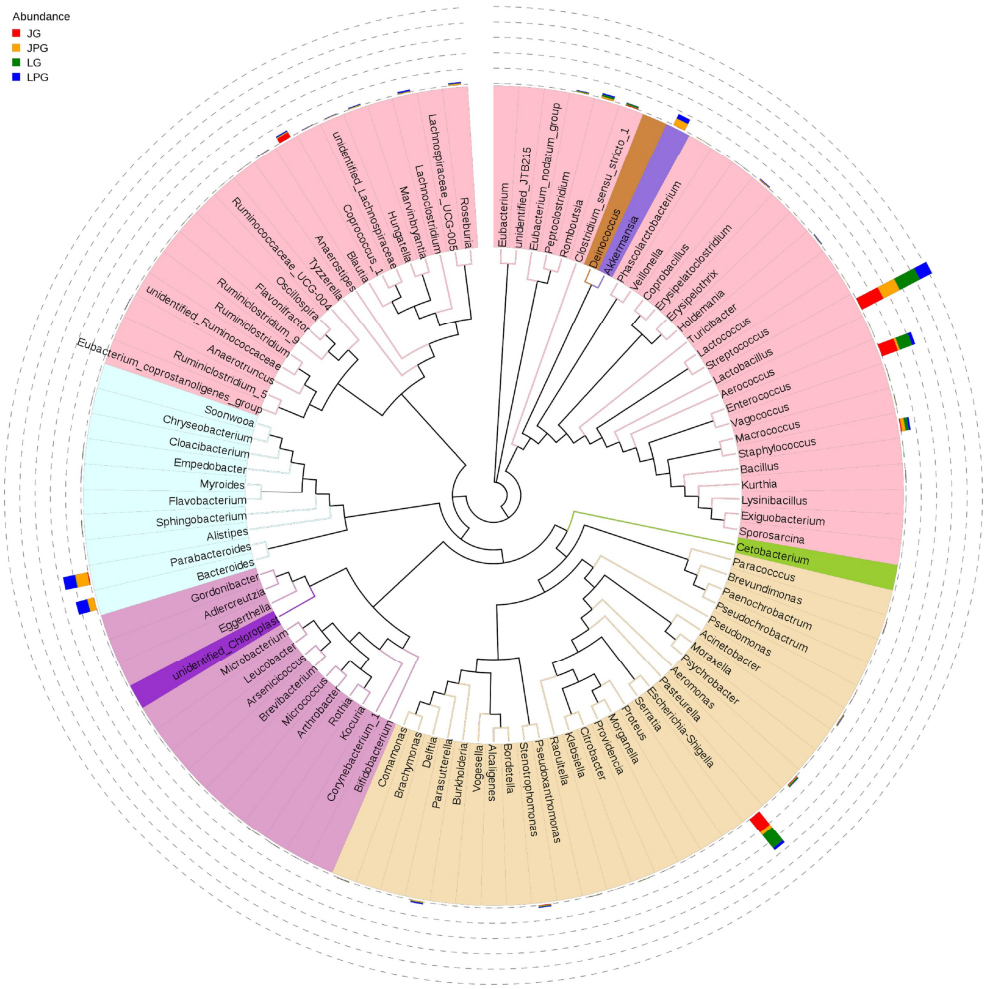


Figure 2.2.4 The evolutionary tree in genus

Notes: Different colors of the branchs represent different phyla. Relative abundance of each genus in each group was displayed outside the circle and different colours represent different groups.

Results directory：

The evolutionary tree in genus by groups: result/02.OTUanalysis/genus_evolutionary_tree_group/genus_group_100.tree.{svg,png}。

The evolutionary tree in genus by samples: result/02.OTUanalysis/genus_evolutionary_tree/genus_100.tree.{svg,png}。

3 Alpha Diversity Analysis

Alpha diversity is widely used to assess microbial diversity within community[8], including species accumulation boxplot, species diversity curves and statistical analysis indices.

3.1 Alpha Diversity Indices

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: Table 3.1 (Number of reads chosen for normalization：cutoff=. The meaning of each alpha diversity index is listed in Method-Information analysis-3 Alpha Diversity Analysis).

Table 3.1 Alpha Diversity Indices statistics

Alpha Indices Table

Sample Name	observed_species	shannon	simpson	chao1	ACE	goods_coverage	PD_whole_tree
JG1	128	1.434	0.339	157.333	162.282	0.999	13.950
JG2	124	1.939	0.568	143.500	148.614	0.999	11.435
JG4	124	2.317	0.616	140.235	146.472	0.999	11.054
JG7	170	3.518	0.834	190.300	194.262	0.999	14.064
JPG7	154	3.516	0.837	183.526	185.675	0.999	13.401
JPG14	173	4.444	0.903	176.143	178.424	1.000	13.996
JPG21	193	4.577	0.895	233.615	221.801	0.999	14.825
JPG28	202	4.660	0.893	213.000	215.024	0.999	16.295
LG1	214	3.429	0.711	226.037	229.800	0.999	16.851
LG2	175	3.083	0.746	203.333	204.871	0.999	15.352

Page 1 of 2 10 View 1 - 10 of 16

Results directory：

Alpha Diversity indices statistics：result/03.AlphaDiversity/alpha_diversity_index.xls

3.2 Species diversity curves

Rarefaction Curves and Rank abundance curves are widely used for indicating the biodiversity of the samples. Rarefaction Curve is created by selecting randomly certain amount of sequencing data from the samples, then counting the number of the species they represent. If the curve is steep, lots of the species remain to be discovered. If the curve becomes flatter, a credible number of samples have been taken, which means only the scarce species remain to be sampled.

Rank abundance curve is used to display relative species abundance. It also can be used to visualize species richness and evenness. It overcomes the shortcomings of biodiversity indices that cannot present the role the variables played in their assessment[9].

Species diversity curves (To view full size picture please **click**) :

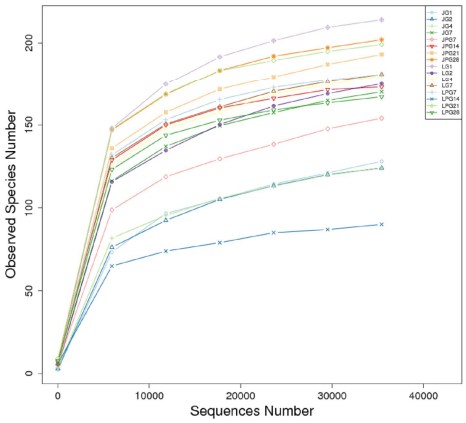


Figure 3.2-1 Rarefaction Curves

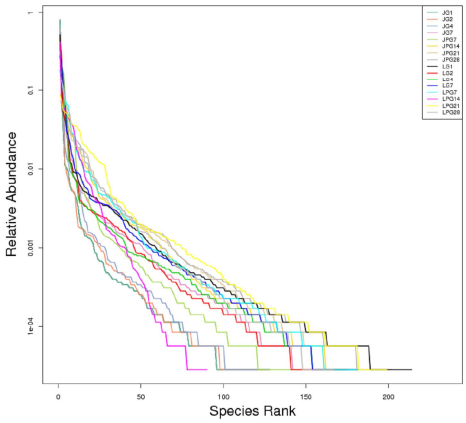


Figure 3.2-2 Rank Abundance curves

Notes: For the Rarefaction Curves, each curve represents a sample, and can be colored and shaped by each sample name supplied in the mapping file. The sequences number is on the X-axis and the observed OTUs number is on the Y-axis. For the Rank Abundance curves, each curve represents an single sample, plotted by OTU relative abundance on the Y-axis and the OTU abundance rank on the X-axis, which can be colored and shaped by each sample name supplied in the mapping file.

Species diversity curves by groups (To view full size picture please **click**) :

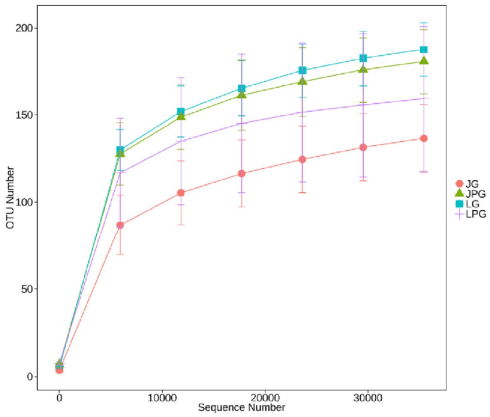


Figure 3.2-3 Rarefaction Curves

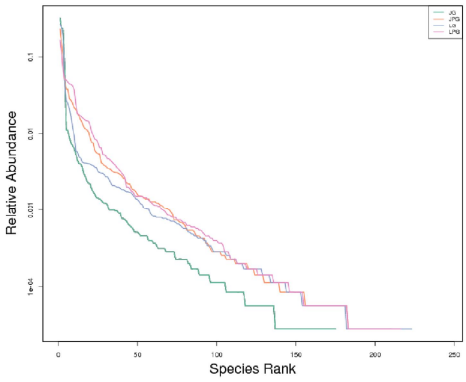


Figure 3.2-4 Rank Abundance curves

Results directory :

Rarefaction Curves：result/03.AlphaDiversity/(group_)observed_species.{pdf,png} ;

Plot data for Rarefaction Curves：result/03.AlphaDiversity/plot_observed_species.txt。

Rank Abundance curves：result/03.AlphaDiversity/(group_)rank_abundance.{pdf,png}。

3.3 Species accumulation boxplot

Species accumulation boxplot is used to display species diversity along with the increasing samples. It can be also be used for judging the adequacy of sample size and predicting the species richness. The result is shown in Figure 3.3 (To view full size picture please **click**) 。

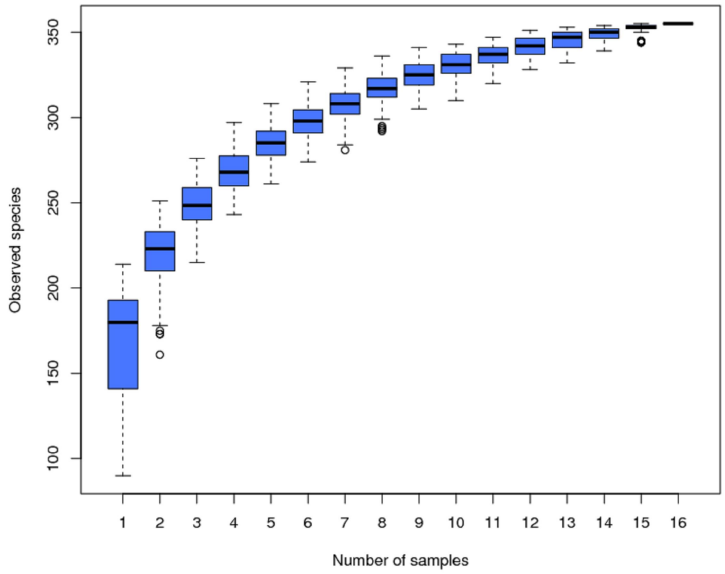


Figure 3.3 Species accumulation boxplot

Notes: Plotted by the "Sample num" on the X-axis and "OTU_num" on the Y-axis. With the increase of sample size, the boxplot showing a sharp rise indicates large amount of species found in samples while the boxplot turning to flat represents adequate sample numbers.

Results directory：

Species accumulation boxplot：result/03.AlphaDiversity/Specaccum/specaccum_test.{pdf,png}；

3.4 Venn and Flower diagram

According to the analysis result of OTUs clustering and the research requirements, we normalized the OTU table, analysed both the common and unique information for different samples (groups), and generate the Venn and Flower diagram.

3.4.1 Venn diagram based on OTUs

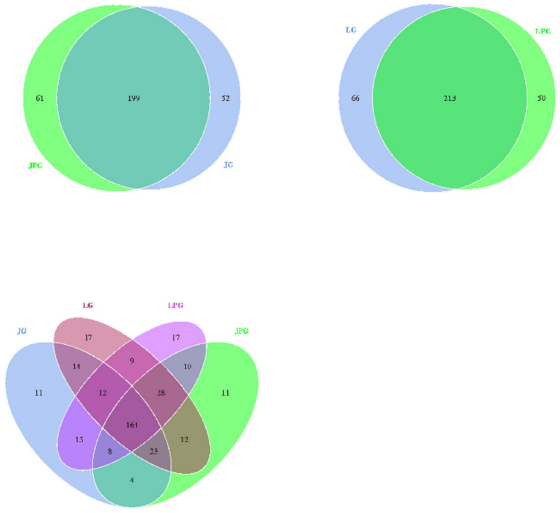


Figure 3.4.1 Venn diagram

Notes: Each circle represents one sample or group. Values in overlapping parts represent common OTUs. The Others are specific OTUs in each sample. To view full size picture please **click**。

Results directory：

Venn diagrams：result/03.AlphaDiversity/venn_figure/ ;

Plot data：result/03.AlphaDiversity/venn_figure/venndata/ ;

3.5 Variation analysis of alpha diversity indices between groups

Boxplots were formed to analyze difference of Alpha Diversity indices between groups. (Click **here** for introduction of boxplot). T-test, wilcox and Tukey tests (T-test and Wilcox test are for 2 groups while Wilcox and Tukey tests are for groups more than 2) are performed for analysis of significance of difference between groups. Boxplots based on observed_species and shannon indices are shown as follow. (To view full size picture please **click**) ：

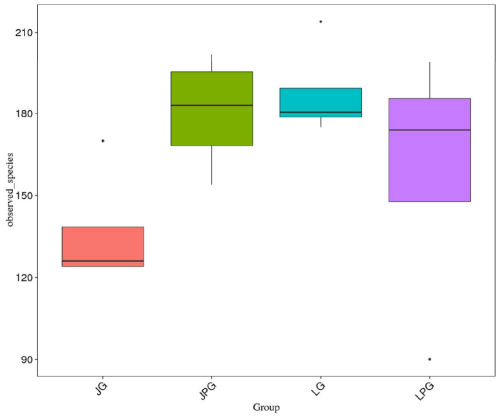


Figure 3.5-1 Box plot of difference of observed_species

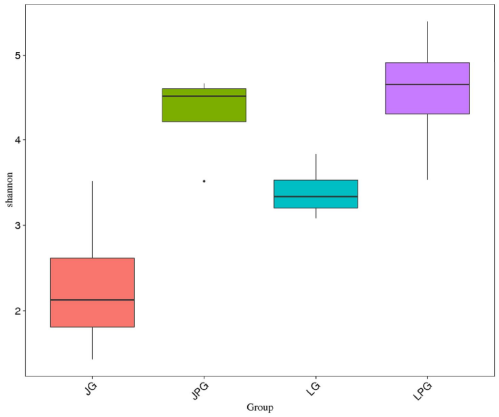


Figure 3.5-2 Boxplots for difference of shannon indices

Results directory：

Box plots with Alpha Diversity indices：result/03.AlphaDiversity/Alpha_div /{ACE，chao1，goods_coverge，observed_species，shannon，simpson，PD_whole_tree}/*.{pdf,png}；

Significance test results：result/03.AlphaDiversity/Alpha_div /{ACE，chao1，goods_coverge，observed_species，shannon，simpson，PD_whole_tree}/{*_wilcox.txt，*_t_test.txt 或*_Tukey.txt}。

4 Beta Diversity Analysis

Beta diversity represents the explicit comparison of microbial communities based on their composition. Beta-diversity metrics thus assess the differences between microbial communities. To compare microbial communities between every pair of community samples, a square matrix of "distance" or dissimilarity was calculated to reflecting the dissimilarity between certain samples, such as Unweighted Unifrac [10,11] and Weighted Unifrac distance[12]. The data in this distance matrix can be visualized with Principal Coordinate Analysis (PCoA), Principal Component Analysis (PCA), Non-Metric Multi-Dimensional Scaling (NMDS) and Unweighted Pair-group Method with Arithmetic Means (UPGMA).

4.1 Beta Diversity Indices

4.1.1 Beta diversity heatmap

Weighted Unifrac distance and Unweighted Unifrac distance were selected to measure the dissimilarity coefficient between pairwise samples, which are phylogenetic measures used extensively in recent microbial community sequencing projects. Heatmap based on Weighted Unifrac and Unweighted Unifrac distance is plotted in Figure 4.1.1 (To view full size picture please**click**) 。

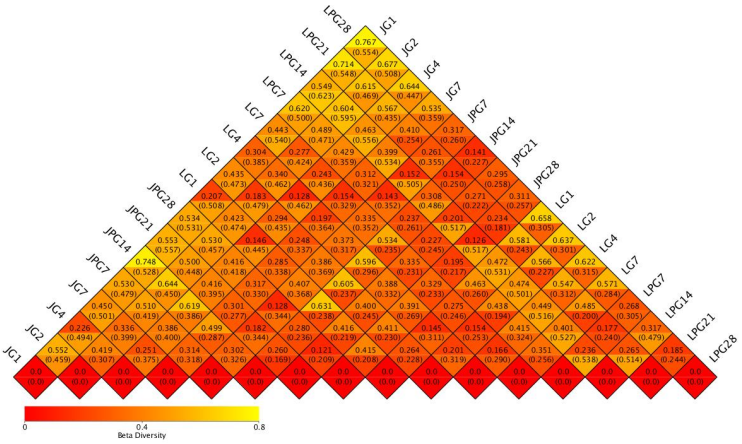


Figure 4.1.1 Beta diversity heatmap

Notes: Each grid represents pairwise dissimilarity coefficient between pairwise samples, in which Weighted Unifrac distance displayed above and Unweighted Unifrac distance conversely.

Results directory：

Beta diversity heatmap:result/04.BetaDiversity/beta_div_heatmap(_group)/beta_diversity.heatmap.*{png,svg}；

Plot data:result/04.BetaDiversity/beta_div_heatmap(_group)/(un)weighted_unifrac_sorted_otu_table.txt。

4.1.2 The difference of Beta Diversity indices between groups

Box plot was generated to show the difference of Beta Diversity indices between groups. (The details of box plot lists in **here**). T-test, wilcox and Tukey tests (T-test and wilcox test are for 2 groups while wilcox and Tukey tests are for groups more than 2) are performed for analysis of significance of difference between groups. (To view full size picture please **click**) ：

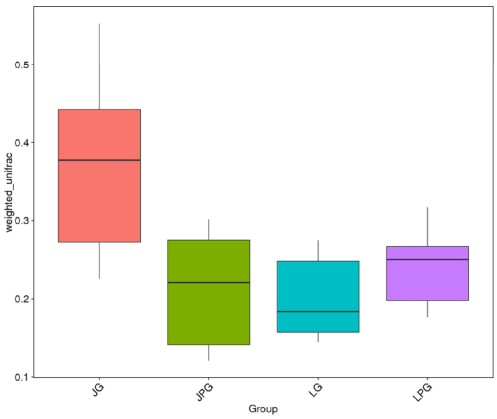


Figure 4.1.2-1 Boxplots based on Weighted Unifrac distance

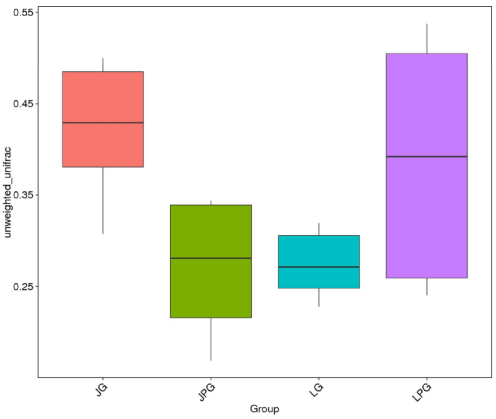


Figure 4.1.2-2 Boxplot based on Unweighted Unifrac distance

Results directory：

Beta diversity box plots：result/04.BetaDiversity/Beta_div/(un)weighted_unifrac.{pdf,png}；

Significance test results：result/04.BetaDiversity/Beta_div/{*_wilcox.txt , *_t_test.txt or *_TukeyHSD.txt}

4.2 Principal Coordinate Analysis (PCoA)

Principal coordinates analysis (PCoA) is an ordination technique, which picks up the main elements and structure from reduced multi-dimensional data series of eigenvalues and eigenvectors. The technique has the advantage over PCA that each ecological distance can be investigated. Weighted Unifrac and Unweighted Unifrac are calculated to assist the PCoA analysis. Gathered samples represent high species composition similarity than the separate ones. The result is shown as follow. (To view full size picture please **click**) :

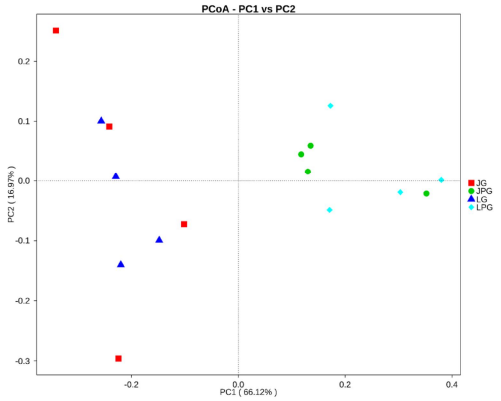


Figure 4.2-1 PCoA based on Weighted Unifrac distance

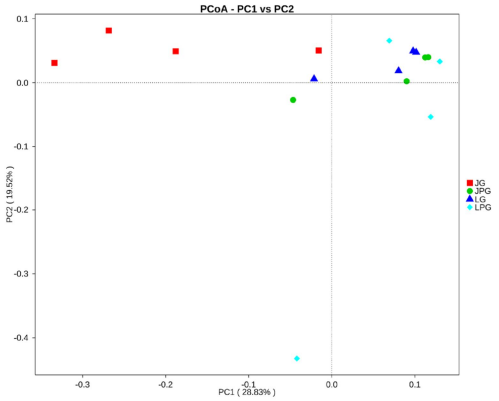


Figure 4.2-2 PCoA based on Unweighted Unifrac distance

Notes: Each point represents a sample, plotted by a principal component on the X-axis and another principal component on the Y-axis, which was colored by group. The percentage on each axis indicates the contribution value to discrepancy among samples.

Results directory :

PCoA results：result/04.BetaDiversity/PCoA/(un)weighted_unifrac/ ;

PCoA distance matrix files：result/04.BetaDiversity/PCoA/(un)weighted_unifrac/(un)weighted_unifrac_dm.txt ;

PCoA plot data：result/04.BetaDiversity/PCoA/(un)weighted_unifrac/(un)weighted_unifrac_pc.txt。

4.3 Principal 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[13]. 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 shown in Figure 4.3 (To view full size picture please **click**) 。

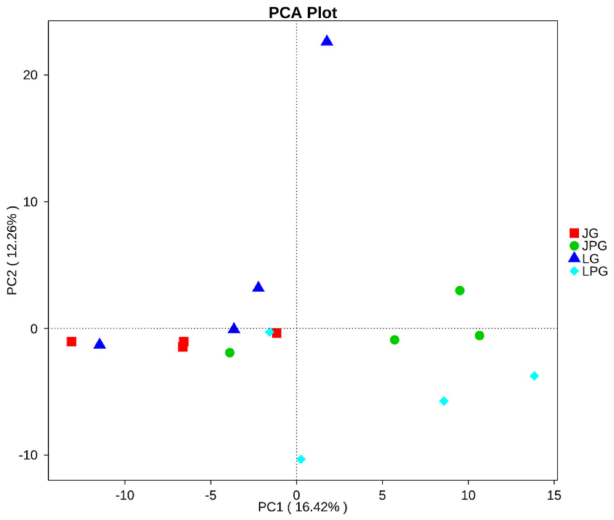


Figure 4.3 PCA

Notes: Each point represents a sample, plotted by the second principal component on the Y-axis and the first principal component on the X-axis, which was colored by group.

Results directory：

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

Graph without sample names：result/04.BetaDiversity/PCA/PCA12_2.{png,pdf}；

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

Graph without sample names and clustering circles：result/04.BetaDiversity/PCA/PCA12_with_cluster_2.{png,pdf}；

PCA results：result/04.BetaDiversity/PCA/pca.csv。

4.4 Non-Metric Multi-Dimensional Scaling (NMDS)

Non-metric multi-dimensional scaling analysis is a ranking method applicable to ecological researches[14]. 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 4.4 (To view full size picture please **click**) 。

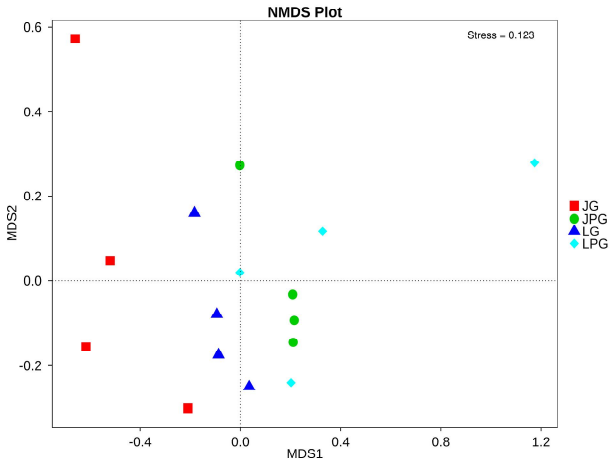


Figure 4.4 NMDS

Notes: 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.

Results directory：

Graph with sample names：result/04.BetaDiversity/NMDS/ NMDS.{png,pdf}；

Graph without sample names：result/04.BetaDiversity/NMDS/NMDS_2.{png,pdf}；

NMDS results：result/04.BetaDiversity/NMDS/NMDS_scores.txt。

4.5 Unweighted Pair-group Method with Arithmetic Mean (UPGMA)

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.

Weighted Unifrac distance matrix and Unweighted Unifrac distance matrix were calculated before used for UPGMA cluster analysis. They were displayed with the integration of clustering results and the relative abundance of each sample by phylum in Figure 4.5-1 and Figure 4.5-2. To view full size pictures please **Weighted Unifrac**click**Unweighted Unifrac** 。

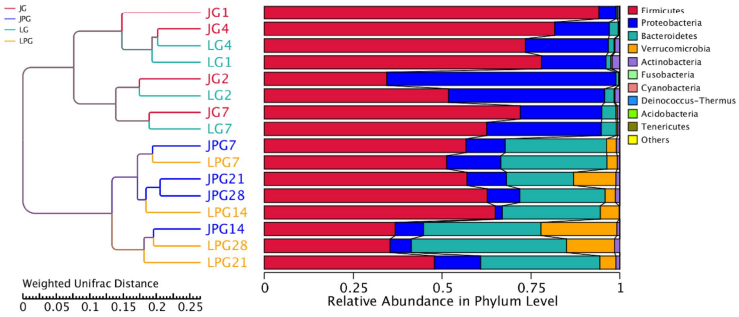


Figure 4.5-1 UPGMA cluster tree based on Weighted Unifrac distance

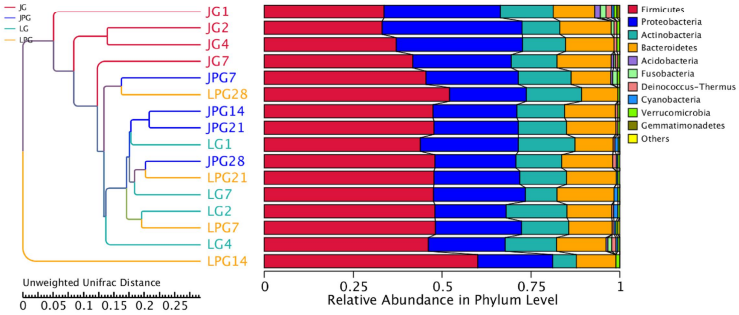


Figure 4.5-2 UPGMA cluster tree based on Unweighted Unifrac distance

Notes: Plotted with UPGMA tree on the left and the relative phylum-level abundance map on the right.

Results directory：

UPGMA results in (un)weighted_unifrac distance：result/04.BetaDiversity/tree/(un)weighted_unifrac /(un)weighted_unifrac.{pdf,png}；

UPGMA cluster tree combined with relative abundance of top10 phylum：result/04.BetaDiversity/tree /(un)weighted_unifrac/UPGMA.W(UnW).tree.{png,svg}。

4.6 Variation analysis of community structures between groups

4.6.1 Analysis of Similarity (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.Click **Anosim** for details. The results are shown in Table 4.6.1 and figure 4.6.1 (To view full size picture please **click**) :

Table 4.6.1 Anosim results

Anosim results

Group	R-value	P-value
JG-LG	-0.07292	0.653
LPG-LG	0.8958	0.026
LPG-JG	0.8229	0.018
JPG-LG	1	0.036
JPG-JG	0.9688	0.021
JPG-LPG	-0.03125	0.541

Page 1 of 1 10 View 1 - 6 of 6

Notes: 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.

According to Anosim results, rank was obtained from sorted distance between samples. Boxplots based on rank (Between group and Within group) are shown as follow :

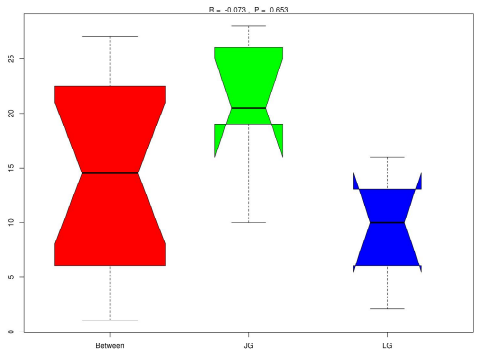


Figure 4.6.1 Anosim result

Notes: Plotted by the rank value on the Y-axis and the "Between group" and "Within group" on the X-axis.

Results directory :

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

Anosim gragh : result/04.BetaDiversity/Anosim/*.{pdf,png} ;

4.6.2 Multi-response permutation procedure (MRPP) analysis

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 more details of MRPP please click **MRPP**. The result is shown in Table 4.6.2 :

Table 4.6.2 MRPP result

MRPP test

Group	A	observed-delta	expected-delta	Significance
LG-LPG	0.2429	0.4695	0.6201	0.03
JPG-LG	0.2847	0.4194	0.5863	0.026
JPG-LPG	0.001812	0.4441	0.4449	0.453
JG-LG	-0.03439	0.5357	0.5179	0.709
JG-LPG	0.2056	0.5604	0.7054	0.037
JG-JPG	0.2505	0.5103	0.6809	0.04

Page 1 of 1 10 View 1 - 6 of 6

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.

Results directory :

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

4.6.3 Adonis

ADONIS is also called permutational MANOVA or nonparametric MANOVA, which is a method of nonparametric multivariate variance test according to distance matrix, e.g. Bray-Curtis, Euclidean, etc. This method can analysis the explanation of grouping factor on difference of samples and estimate the significance of grouping by permutation test[15,16,17,18]. For more details of ADONIS please click **ADONIS**. The result is in Table 4.6.3 :

Table 4.6.3 Adonis result

Adonis test

Vs_group	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
LPG-JPG	1(6)	0.09954(0.63367)	0.099542(0.105612)	0.94253	0.13576(0.86424)	0.549
LPG-LG	1(6)	0.75363(0.69821)	0.75363(0.11637)	6.4762	0.51909(0.48091)	0.029
LPG-JG	1(6)	0.86390(0.98764)	0.86390(0.16461)	5.2483	0.46659(0.53341)	0.022
JPG-LG	1(6)	0.73492(0.56030)	0.73492(0.09338)	7.87	0.56741(0.43259)	0.001
JPG-JG	1(6)	0.89246(0.84973)	0.89246(0.14162)	6.3017	0.51226(0.48774)	0.001
LG-JG	1(6)	0.09318(0.91426)	0.093184(0.152377)	0.61154	0.0925(0.9075)	0.684

Page 1 of 1 10 View 1 - 6 of 6

Notes: Df represents degree of freedom. SumsOfSqs represents sums of squares of deviations. MeanSqs represents SumsOfSqs/Df. F.Model represents F-test value. R2 represents the explanation of grouping factor on difference of samples, calculated from the ratio of grouping variance and total variance. Pr means P-value. Values in parentheses stands for Residual Error.

Results directory :

Adonis result：result/04.BetaDiversity/Adonis/bray_adonis.txt

4.6.4 Analysis of molecular variance (AMOVA)

AMOVA is similar with ADONIS, which is a kind of nonparametric method aiming at determining whether the difference of microbial community structure among groups is significant[19]. For more details of AMOVA please **click**. The result is shown in table 4.6.4 :

Table 4.6.4 Amova result

Amova test

vs_group	SS	df	MS	Fs	p-value
JG-JPG	0.355446(0.302681)	1(6)	0.355446(0.0504469)	7.04595	0.024
JG-LPG	0.474485(0.318249)	1(6)	0.474485(0.0530415)	8.94555	0.029
JPG-LG	0.326175(0.139889)	1(6)	0.326175(0.0233148)	13.99	0.036
JG-JPG-LG-LPG	0.826796(0.458138)	3(12)	0.275599(0.0381782)	7.21875	<0.001*
JG-LG	0.0181294(0.291938)	1(6)	0.0181294(0.0486564)	0.372601	0.736
JPG-LPG	0.0249505(0.166199)	1(6)	0.0249505(0.0276999)	0.900745	0.387
LG-LPG	0.454406(0.155457)	1(6)	0.454406(0.0259094)	17.5383	0.035

Page 1 of 1 10 View 1 - 7 of 7

Notes: SS means sums of squares of deviations. df represents degree of freedom. MS means SS/df. Fs means F-test value. Values in parentheses stands for Residual Error.

Results directory :

Amova result：result/04.BetaDiversity/Amova/stat_amova.txt

4.7 Between-group variation analysis of species

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.

4.7.1 T-test

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. The result is shown in Figure 4.7.1 (To view full size picture please **click here**).

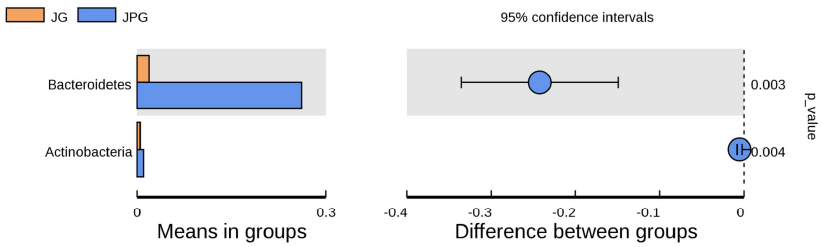


Figure 4.7.1Between-group T-test analysis

Notes: 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.

Results directory：

T-test of between group variation on various taxon ranks：result/04.BetaDiversity/t.test_bar_plot；

T-test of between group variation in phylum:result/04.BetaDiversity/t.test_bar_plot/phylum/*.(png,svg)；

T-test result:result/04.BetaDiversity/t.test_bar_plot/phylum/*.psig.xls。

4.7.2 MetaStat

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

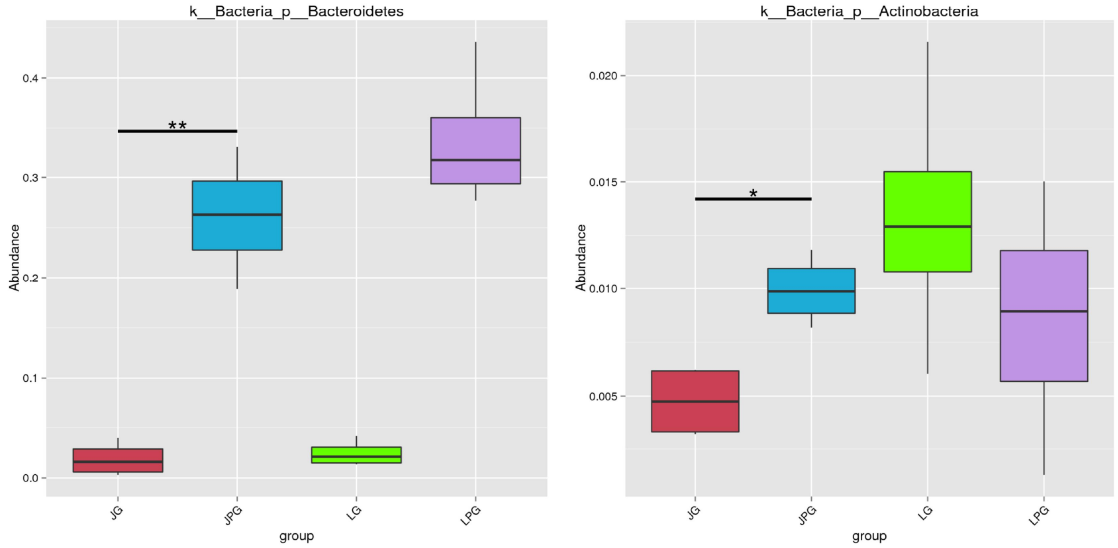


Figure 4.7.2 Between group metastat analysis

Notes: Plotted by relative abundance on the Y-axis and the group name on the X-axis. Horizontal line represents the two groups with significant variation. " * " represents significant variation (q value < 0.05) while " * * " represents very significant variation(q value < 0.01).

Results directory：

MetaStat result in each taxonomic rank：result/04.BetaDiversity/MetaStat；

MetaStat result in phylum：result/04.BetaDiversity/MetaStat/*.test.xls；

Statistical metrics obtained from metastats analysis in phylum with p-value less than 0.05：result/04.BetaDiversity /MetaStat/phylum /*.psig.xls；

Statistical metrics obtained from metastats analysis in phylum with q-value less than 0.05：result/04.BetaDiversity /MetaStat/phylum /*.qsig.xls。

4.7.3 LEfSe analysis

LEfSe(linear discriminant analysis(LDA) Effect Size) analysis detects biomarkers [20], with statistical differences among groups, 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. The results are as follow. (To view full size picture please click **the histogram of LDA score**and **Cladogram**) :

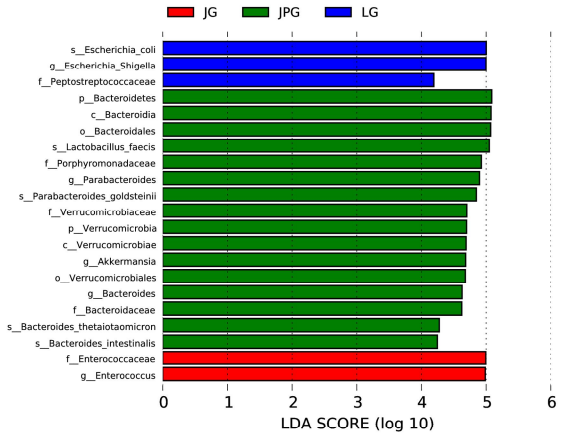


Figure 4.7.3-1 Histogram of LDA scores

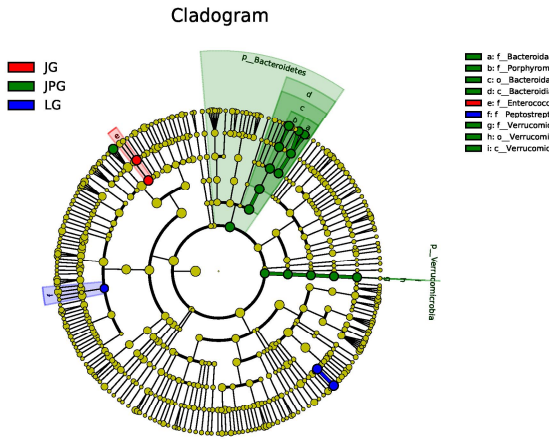


Figure 4.7.3-2 Cladogram

Notes: 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. The diameter of each circle represents proportionally the relative abundance of each taxon. 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 deNotesd by red color, so do the green nodes. Letters above the circles and corresponding species are annotated on the right side.

Results directory :

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

LEfSe cladogram:result/04.BetaDiversity/LEfSe/*/LDA.*.tree.(pdf,png);

Relative abundance plots of Biomarker in different samples: result/04.BetaDiversity/LEfSe/*/biomarkers_raw_images/。

5 Data mining

16S rDNA（18S rDNA，ITS）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.

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.

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.

IV. Methods

Sequencing preparation

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.

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).

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 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).

The libraries generated with TruSeq® DNA PCR-Free Sample Preparation Kit and quantified via Qubit and Q-PCR, would be analysed by HiSeq2500 PE250.

Information analysis

1 Sequencing data processing

Paired-end reads was assigned to samples based on their unique barcode and truncated by cutting off the barcode and primer sequence. Paired-end reads were merged using FLASH (V1.2.7, **http://ccb.jhu.edu/software/FLASH/**) [28], 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. Quality filtering on the raw tags were performed under specific filtering conditions to obtain the high-quality clean tags[29] according to the Qiime(V1.7.0, **http://qiime.org/scripts/split_libraries_fastq.html**) [30]quality controlled process. 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**)[31] to detect chimera sequences (**http://www.drive5.com/usearch/manual/chimera_formation.html**). And then the chimera sequences were removed [32]. Then the Effective Tags finally obtained.

2 OTU cluster and Species annotation

Sequences analysis were performed by Uparse software(Uparse v7.0.1001 **http://drive5.com/uparse/**) [33] using all the effective tags. Sequences with ≥97% similarity were assigned to the same OTUs. Representative sequence for each OTU was screened for further annotation. For each representative sequence, Mothur software was performed against the SSUrRNA database of SILVA Database (**http://www.arb-silva.de/**) [34]for species annotation at each taxonomic level(Threshold:0.8~1).[35] (kingdom, phylum, class, order, family, genus, species). To get the phylogenetic relationship of all OTUs representative sequences, the MUSCLE [36] (Version 3.8.31，**http://www.drive5.com/muscle/**) can compare multiple sequences rapidly. 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.

3 Alpha Diversity

Alpha diversity is applied in analyzing complexity of species diversity for a sample through6 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).

Alpha Diversity Indices：

Community richness indices:

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**);

Community diversity indices:

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**);

The index of 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**);

The index of phylogenetic diversity：

PD_whole_tree - PD_whole_tree index(**http://scikit-bio.org/docs/latest/generated/generated/skbio.diversity.alpha.faith_pd.html?highlight=pd#skbio.diversity.alpha.faith_pd**)

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).

LEfSe analysis was conducted by LEfSe software. Metastats was calculated by R software. P-value was calculated by method of permutation test while q-value was calculated by method of Benjamini and Hochberg False Discovery Rate[37]. Anosim, MRPP and Adonis were performed by R software(Vegan package: anosim function， mrpp function and adonis function). AMOVA was calculated by mothur using amova function. T_test and drawing were conducted by R software.

V. 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] Asnicar F, Weingart G, Tickle T L, et al. Compact graphical representation of phylogenetic data and metadata with GraPhlAn[J]. PeerJ, 2015.
[5] DeSantis, T. Z., et al. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic acids research 34.suppl 2 (2006): W394-W399.
[6] Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. Interactive metagenomic visualization in a Web browser. BMC bioinformatics 12.1 (2011): 385.
[7] Bulgarelli D, Garrido-Oter R, Münch P C, et al. Structure and function of the bacterial root microbiota in wild and domesticated barley[J]. Cell host & microbe, 2015, 17(3): 392-403.
[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] Lundberg, Derek S., et al. Practical innovations for high-throughput amplicon sequencing.Nature methods 10.10 (2013): 999-1002.
[10] Lozupone, Catherine, and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology 71.12 (2005): 8228-8235.
[11] Lozupone, Catherine, et al. UniFrac: an effective distance metric for microbial community comparison. The ISME journal 5.2 (2011): 169.
[12] 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.
[13] 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.
[14] Magali Noval Rivas, PhD, Oliver T. Burton, et al. A microbita signature associated with experimental food allergy promotes allergic senitization and anaphylaxis. The Journal of Allergy and Clinical Immunology.Volume 131, Issue 1, Pages 201-212, January 2013.
[15] Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology, 26: 32-46.
[16] McArdle, B.H. and M.J. Anderson. 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. Ecology, 82: 290-297.
[17] Warton, D.I., Wright, T.W., Wang, Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3, 89-101.
[18] Zapala, M.A. and N.J. Schork. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences, USA, 103:19430-19435.
[19] Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 131, 479-491.
[20] Segata, Nicola, et al. Metagenomic biomarker discovery and explanation. Genome Biol 12.6 (2011): R60.
[21] Algina, J., & Keselman, H. J. (1999). Comparing squared multiple correlation coefficients: Examination of a confidence interval and a test significance. Psychological Methods, 4(1), 76-83.
[22] Sheik CS, Mitchell TW, Rizvi FZ, Rehman Y, Faisal M, et al. (2012) Exposure of Soil Microbial Communities to Chromium and Arsenic Alters Their Diversity and Structure. PLoS ONE 7(6): e40059. doi:10.1371/journal.pone.0040059.
[23] Gross, J. (2003). Variance inflation factors. R News 3(1), 13–15.
[24] Clarke, K. R & Ainsworth, M. 1993. A method of linking multivariate community structure to environmental variables. Marine Ecology Progress Series, 92, 205–219.
[25] Yang S L, Zhang J, Xu X J. Influence of the Three Gorges Dam on downstream delivery of sediment and its environmental implications, Yangtze River[J]. Geophysical research letters, 2007, 34(10).
[26] Shuo Jiao, Zhenshan Liu,2016.Bacterial communities in oil contaminated soils: Biogeography and co- occurrence patterns.Soil Biology & Biochemistry 98 (2016) 64e73.
[27] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60.
[28] Magoč, Tanja, and Steven L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27.21 (2011): 2957-2963.
[29] Bokulich, Nicholas A., et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods 10.1 (2013): 57-59.
[30] Caporaso, J. Gregory, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7.5 (2010): 335-336.
[31] Edgar, Robert C., et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27.16 (2011): 2194-2200.
[32] 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.
[33] Edgar, Robert C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10.10 (2013): 996-998.
[34] 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.
[35] Quast C, Pruesse E, et al.The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. (2013) : D590-D596.
[36] MUSCLE: multiple sequence alignment with high accuracy and high throughputEdgar, 2004
[37] White, James Robert, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS computational biology 5.4 (2009): e1000352.

VI. Appendix

1. Result catalogue：**ReadMe.pdf**。

2. WebShow Instruction：**WebShow.pdf**。

3. Methods：**Methods.pdf**。

Notes: it is strongly advised to scan this report on Firefox box, for several browsers such as IE may disable for table display. More vectograms and statistic data are provided in delivery directory.