# Amplicon Project Report

**BGI Co., Ltd.**

Reported on Friday, 31st Aug., 2018

Table of Contents

## I Overview

### 1 Introduction of workflow

Once the DNA sample(s) was(were) received, a qulity test has been done first, then all the qualified DNA is used to construct a library(libraries). For PCR product, the jagged ends of DNA fragment would be converted into blunt ends by using T4 DNA polymerase, Klenow Fragment and T4 Polynucleotide Kinase. Then add an 'A' base to each 3' end to make it easier to add adapters. After all that, fragments too short would be removed by Ampure beads. For genomics DNA, we use fusion primer with dual index and adapters for PCR, fragments too short would be removed by Ampure beads too. In both cases, only the qualified library can be used for sequencing. The bioinformatics analysis will be carried on with sequencing data.

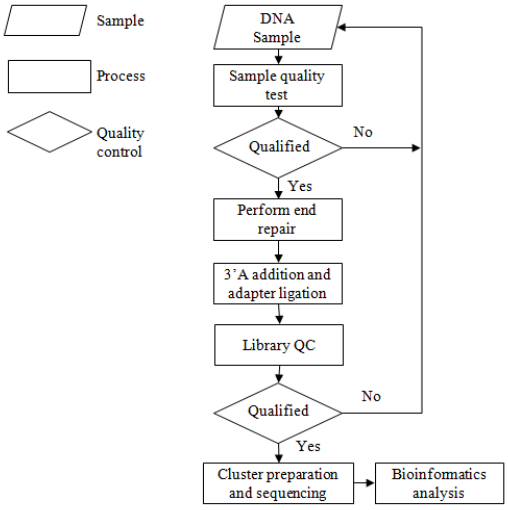



**Figure 1 Workflow**

## 2 Bioinformatics analysis workflow

The raw data were filtered to eliminate the adapter pollution and low quality to obtain clean reads, then paired-end reads with overlap were merged to tags. And tags were clustered to OTU at 97% sequence similarity. Taxonomic ranks were assigned to OTU representative sequence using Ribosomal Database Project (RDP) Naïve Bayesian Classifier v.2.2. At last, alpha diversity, beta diversity and the different species screening were analyzed based on OTU and taxonomic ranks.

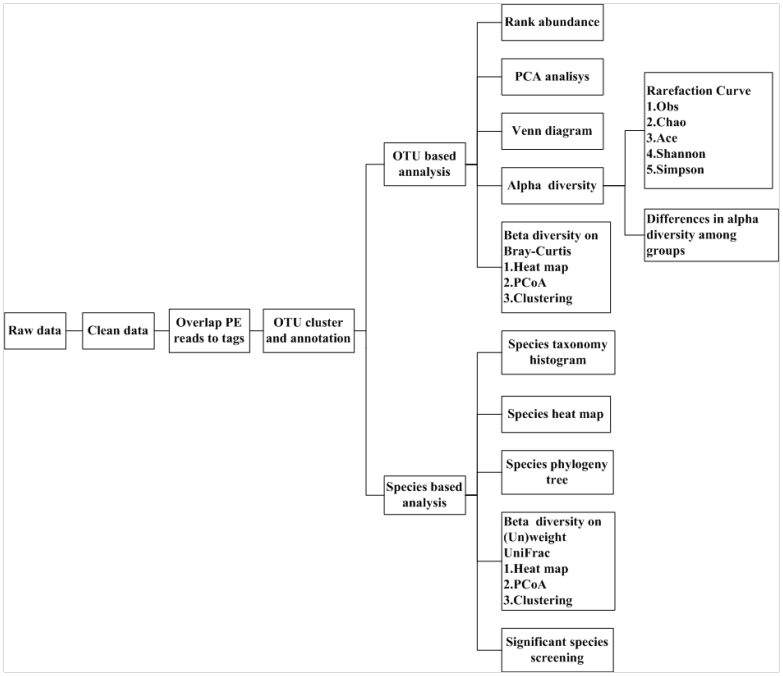



**Figure 2 Bioinformatics analysis pipeline of Amplicon Sequencing**

## II Methods and results

### 1 Data statistics

In order to obtain more accurate and reliable results in subsequent bioinformatics analysis[1], the raw data will be pre-processed to get clean data by in-house procedure as following:
1) Truncation of sequence reads not having an average quality of 20 over a 30 bp sliding window based on the phred algorithm, and trimmed reads having less than 75% of their original length, as well as its paired read, will be removed;
2) Removal of reads contaminated by adapter(default parameter: 15 bases overlapped by reads and adapter with maximal 3 bases mismatch allowed);
3) Removal of reads with ambiguous basa(N base), and its paired reads;
4) Removal of reads with low complexity(default: reads with 10 consecutive same base).
For pooling library with barcode samples mixed, the clean reads were assigned to corresponding samples by allowing 0 base mismatch to barcode sequences with in-house scripts.

Paired-end reads were generated with Illumina HiSeq/MiSeq platform, then the reads with sequencing adapters, N base, poly base, low quality etc were filtered out with default parameters(detailed in Method 1), and the data processing results was listed in Table 1-1.

**Table 1-1 Data Statistics**

| Sample Name | Reads Length (bp) | Raw Data (Mbp) | Adapter (%) | N base (%) | Ploy base (%) | Low Quality (%) | Clean Data (Mbp) | Data Utilization Ratio (%) | Raw Reads | Clean Reads | Read Utilization Ratio (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 611A | 250:250 | 33.65 | 0.000 | 0.076 | 0.003 | 5.285 | 30.32 | 90.11 | 67293*2 | 61194*2 | 90.94 |
| 612A | 250:250 | 34.01 | 0.000 | 0.075 | 0.003 | 4.963 | 30.65 | 90.12 | 68023*2 | 61948*2 | 91.07 |
| 621A | 250:250 | 32.29 | 0.000 | 0.094 | 0.001 | 3.215 | 30.18 | 93.47 | 64578*2 | 60787*2 | 94.13 |
| 622A | 250:250 | 32.64 | 0.000 | 0.087 | 0.005 | 3.359 | 30.35 | 92.99 | 65278*2 | 61233*2 | 93.80 |
| 631A | 250:250 | 33.19 | 0.000 | 0.065 | 0.001 | 4.002 | 30.57 | 92.10 | 66388*2 | 61636*2 | 92.84 |
| 632A | 250:250 | 32.79 | 0.000 | 0.081 | 0.004 | 4.020 | 30.25 | 92.25 | 65584*2 | 60997*2 | 93.01 |
| 641A | 250:250 | 32.00 | 0.000 | 0.095 | 0.002 | 2.823 | 30.26 | 94.58 | 63990*2 | 60834*2 | 95.07 |
| 642A | 250:250 | 32.44 | 0.000 | 0.082 | 0.001 | 3.165 | 30.42 | 93.78 | 64878*2 | 61228*2 | 94.37 |
| 711A | 250:250 | 33.31 | 0.000 | 0.089 | 0.002 | 4.643 | 30.17 | 90.58 | 66619*2 | 60945*2 | 91.48 |
| 712A | 250:250 | 33.36 | 0.000 | 0.094 | 0.000 | 3.899 | 30.70 | 92.02 | 66727*2 | 61833*2 | 92.67 |
| 721A | 250:250 | 33.85 | 0.000 | 0.081 | 0.002 | 4.394 | 30.78 | 90.92 | 67707*2 | 62069*2 | 91.67 |
| 722A | 250:250 | 32.78 | 0.000 | 0.068 | 0.002 | 3.812 | 30.26 | 92.32 | 65557*2 | 60940*2 | 92.96 |
| 731A | 250:250 | 31.86 | 0.000 | 0.094 | 0.001 | 2.126 | 30.57 | 95.96 | 63717*2 | 61378*2 | 96.33 |
| 732A | 250:250 | 32.44 | 0.000 | 0.093 | 0.000 | 2.990 | 30.45 | 93.87 | 64877*2 | 61360*2 | 94.58 |
| 741A | 250:250 | 32.78 | 0.000 | 0.077 | 0.001 | 3.066 | 30.80 | 93.95 | 65564*2 | 61994*2 | 94.55 |
| 742A | 250:250 | 31.69 | 0.000 | 0.087 | 0.002 | 1.938 | 30.46 | 96.12 | 63380*2 | 61204*2 | 96.57 |

**Results directory：BGI_results/Clean_Data/**

## 2 Paired end reads are merged to tags

If the two paired-end reads overlapped, the consensus sequence was generated by **FLASH**[2](Fast Length Adjustment of Short reads, v1.2.11), and the detailed method is as follows:
1) Minimal overlapping length: 15 bp;
2) Mismatching ratio of overlapped region: <= 0.1.
Removal of paired end reads without overlaps.

The high quality paired-end reads were combined to tags based on overlaps, 979025 tags were obtained in total with 61189 tags per sample on average, and the average length is 252 bp.

**Table 2-1 Tags statistics**

| Sample Name | Total Pairs Read Number | Connect Tag Number | Connect Ratio (%) | Average Length And SD | Tags Without Primer | Tag Utilization Ratio (%) | Average Length (bp) And SD |
|---|---|---|---|---|---|---|---|
| 611A | 61194 | 60978 | 99.65 | 252/0 | - | - | -/- |
| 612A | 61948 | 61770 | 99.71 | 252/0 | - | - | -/- |
| 621A | 60787 | 60636 | 99.75 | 252/0 | - | - | -/- |
| 622A | 61233 | 61108 | 99.80 | 252/0 | - | - | -/- |
| 631A | 61636 | 61466 | 99.72 | 252/0 | - | - | -/- |
| 632A | 60997 | 60845 | 99.75 | 252/0 | - | - | -/- |
| 641A | 60834 | 60704 | 99.79 | 252/0 | - | - | -/- |
| 642A | 61228 | 61098 | 99.79 | 252/0 | - | - | -/- |
| 711A | 60945 | 60774 | 99.72 | 252/0 | - | - | -/- |
| 712A | 61833 | 61681 | 99.75 | 252/0 | - | - | -/- |
| 721A | 62069 | 61928 | 99.77 | 252/0 | - | - | -/- |
| 722A | 60940 | 60791 | 99.76 | 252/0 | - | - | -/- |
| 731A | 61378 | 61205 | 99.72 | 252/0 | - | - | -/- |
| 732A | 61360 | 61159 | 99.67 | 252/0 | - | - | -/- |
| 741A | 61994 | 61821 | 99.72 | 252/0 | - | - | -/- |
| 742A | 61204 | 61061 | 99.77 | 252/0 | - | - | -/- |

Note: '-' in 'tags without primer' indicates there is no primer removal to the tags.

**Result directory: BGI_results/Connect_Tags/**

## 3 Analysis of community patterns

The tags were clustered to OTU(Operational Taxonomic Unit) by scripts of software USEARCH(v7.0.1090)[3], detailed as follows:
1) The tags were clustered into OTU with a 97% threshold by using UPARSE, and the OTU unique representative sequences were obtained;
2) Chimeras were filtered out by using UCHIME(v4.2.40);
The 16S rDNA and ITS sequences were screened for chimeras by mapping to gold database(v20110519 ) , UNITE(v20140703) separately, de novo chimera detection was done for 18S rDNA sequences;
3) All tags were mapped to each OTU representative sequences using USEARCH GLOBAL, then the tags number of each OTU in each sample will be summarized to OTU abundance table.

OTU representative sequences were taxonomically classified using Ribosomal Database Project (RDP) Classifier v.2.2 trained on the database Greengene_2013_5_99, using 0.6 confidence values as cutoff.

Databases used for species annotation:
16S rDNA is used for bacterial and archaea community: Greengene(default): V201305[8] ; RDP: Release11_5,20160930
18S rDNA is used for fungal community: Silva(default): Version_132,20180410
ITS is also used for fungal community: UNITE(default): Version_7_2,20171201

OTUs were Filtered as follow:
1) Unassigned OTUs were removed;
2) OTUs not assigned to the target specices were removed. For example, the OTUs assigned to archeae would be removed if the project is about 16S rDNA for bacterial community study.
The filtered OTUs were used to downstream processing。

### 3.1 OTU cluster and abundance

#### 3.1.1 OTU Statistics

Filtered tags are clustered into OTU (Operational Taxonomic Units) at 97% similarity and detailed information is summarized in Table 3-1, OTU number per sample primarily represents the degree of sample diversity.

**Table 3-1 OTU Statistics**

| Sample Name | Tag number | OTU number |
|---|---|---|
| 611A | 50909 | 409 |
| 612A | 51137 | 424 |
| 621A | 48911 | 439 |
| 622A | 49230 | 474 |
| 631A | 49455 | 446 |
| 632A | 49138 | 456 |
| 641A | 50208 | 427 |
| 642A | 50733 | 396 |
| 711A | 51404 | 393 |
| 712A | 50357 | 398 |
| 721A | 50267 | 427 |
| 722A | 47904 | 456 |
| 731A | 49744 | 491 |
| 732A | 48934 | 465 |
| 741A | 49158 | 497 |
| 742A | 50922 | 439 |

#### 3.1.1 OTU venn chart

Venn diagram could visually display the number of common/unique OTUs in multi-samples/groups. The core microbiomes of different environments could be obtained if combined with the OTU represent species.

Based on the OTU abundance, OTU of each group was listed, Venn diagram was drawn by VennDiagram of software R(v3.1.1), then the common and specific OTU ID were summarized.

Different color represents different samples or groups, The interior of each circle symbolically represents the number of observed OTUs in the certain sample/group. The overlapping area or intersection would represent the set of OTU commonly present in the counterpart samples/groups. Likewise, the single-layer zone represents the number of OTUs uniquely found in the certain sample/group. Venn diagram is suited to 2-5 samples or groups.

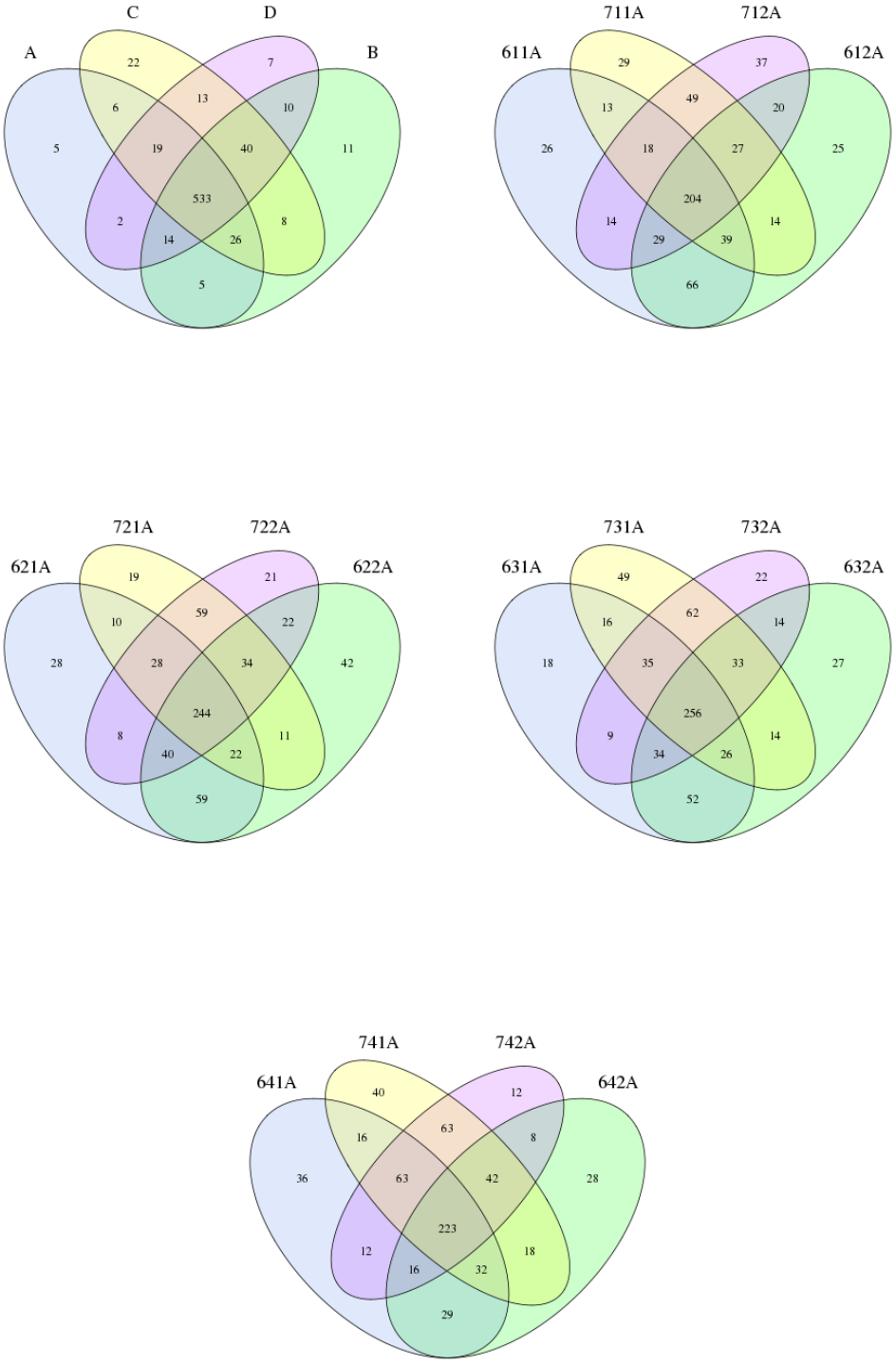



**Figure 3-1 Shared OTU across different samples or groups**

#### 3.1.3 Core-Pan OTU Analysis

Display the common and the uniq OTU of samples or groups.

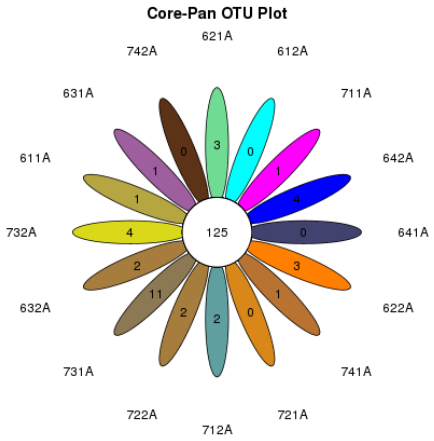



**Figure 3-2-1 Core-Pan OTU**

The middle circle indicates the number of shared OTUs in these samples o r groups, a nd the ellipse outside the middle circle indicates the number of OTUs that only this sample o r group has w hile others doesnt.

### 3.1.4 OTU PCA analysis

In order to display the differences of OTU composition in different samples, Principal component analysis(PCA) was used to construct 2-D graph to summarize factors mainly responsible for this difference, similarity is high if two samples are closely located.

Based on the OTU abundance information, the relative abundance of each OTU in each sample will be calculated, and the PCA of OTU was done with the relative abundance value. The software used in this step was package 'ade4' of software R(v3.1.1).

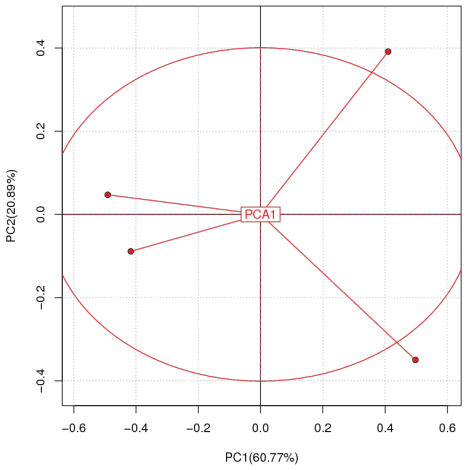



**Figure 3-3-1 PCA based on OTU abundance(Description)**

X-axis,1st principal component and Y-axis, 2nd principal component. Number in brackets represents contributions of principal components to differences among samples. A dot represents each sample, and different colors represent different groups.

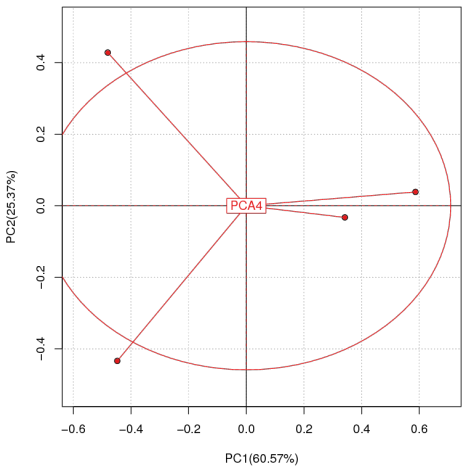



**Figure 3-3-2 PCA based on OTU abundance(Description)**

X-axis,1st principal component and Y-axis, 2nd principal component. Number in brackets represents contributions of principal components to differences among samples. A dot represents each sample, and different colors represent different groups.

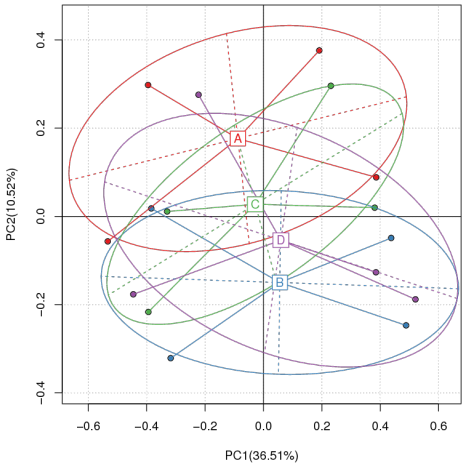



**Figure 3-3-3 PCA based on OTU abundance(Description)**

X-axis,1st principal component and Y-axis, 2nd principal component. Number in brackets represents contributions of principal components to differences among samples. A dot represents each sample, and different colors represent different groups.

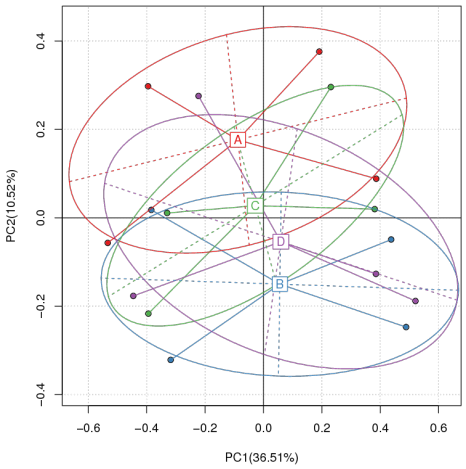


**Figure 3-3-4 PCA based on OTU abundance(Description)**

X-axis,1st principal component and Y-axis, 2nd principal component. Number in brackets represents contributions of principal components to differences among samples. A dot represents each sample, and different colors represent different groups.

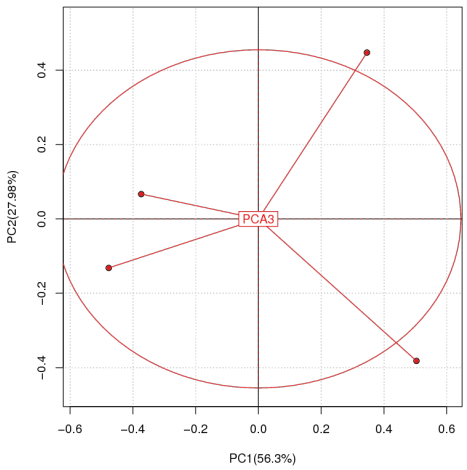



**Figure 3-3-5 PCA based on OTU abundance(Description)**

X-axis,1st principal component and Y-axis, 2nd principal component. Number in brackets represents contributions of principal components to differences among samples. A dot represents each sample, and different colors represent different groups.

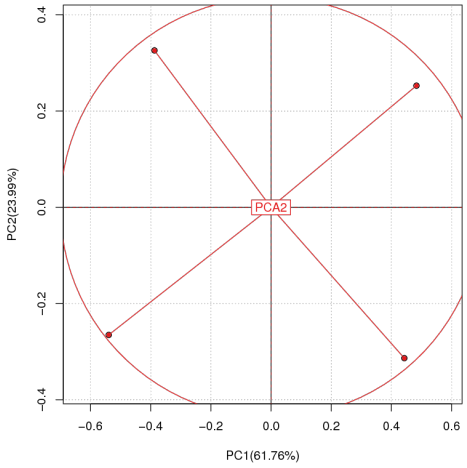



**Figure 3-3-6 PCA based on OTU abundance(Description)**

X-axis,1st principal component and Y-axis, 2nd principal component. Number in brackets represents contributions of principal components to differences among samples. A dot represents each sample, and different colors represent different groups.

#### 3.1.5 Species Accumulation analysis

Species Accumulation (SA) analysis. SA plots showing the increase in OTUs detected with the addition of each sample. Each bar represents 100 random draws (without replacement) of samples from the sample pool. The picture shows the curve obtained using all of the OTU data.

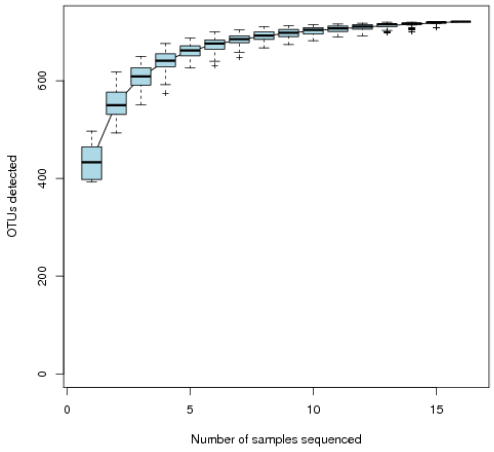



**图3-4-1 Species Accumulation analysis**

**3.1.6 PLS-DA analysis**

PLS-DA is performed in order to sharpen the separation between groups of observations, by hopefully rotating PCA components such that a maximum separation among classes is obtained, and to understand which variables carry the class separating information. Figure 3 is the graph of partial least squares discrimination analysis.

The software used in this step was package 'mixOmics' of software R.

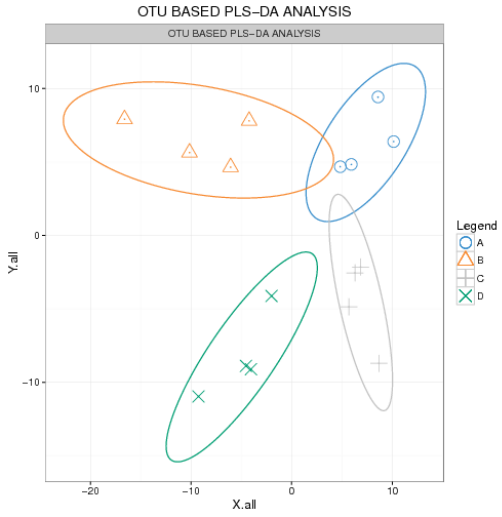



**Figure 3-5-1 PLS-DA based on OTU abundance(Description)**

The horizontal axis and the vertical axis indicate the top 2 components. Each dot represents one sample. Samples are colored and grouped by elipse according to their group information.

**3.1.7 OTU rank curve**

OTU rank abundance curve provides a means for visually representing species richness and species evenness. Species richness can be viewed as the number of different species on the chart (X-axis)i.e., how many species were ranked. Species evenness is derived from the slope of the line that fits the graph. A steep gradient indicates low evenness as the high ranking species have much higher abundances than the low ranking species. A shallow gradient indicates high evenness as the abundances of different species are similar.

OTU were ranked by the relative abundance value as x-axis, and the OTU relative abundance was as y-axis, then the rank curve were drawn by software R(v3.1.1).

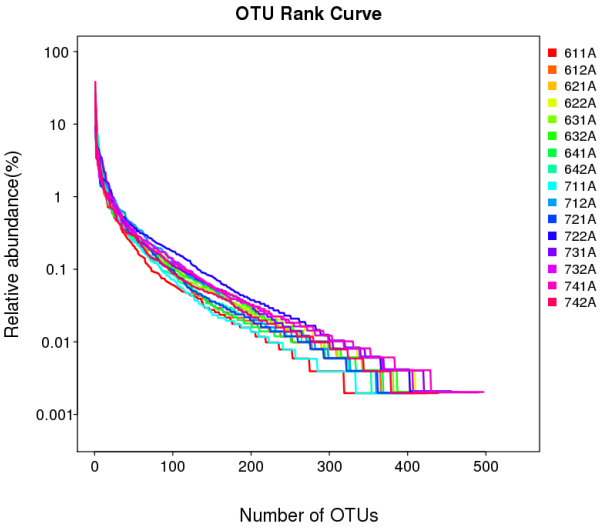



**Figure 3-6 OTU Rank Curve**

### 3.2 Species composition and abundance

#### 3.2.1 Species Annotation

The tags number of each taxonomic rank(Phylum, Class, Order, Family, Genus, Species) or OTU in different samples were summarized in a profiling table or histogram, and the histogram was drawn with the software R(v3.1.1).

**Figure 3-4-*** are the taxonomics composition distribution histograms of each sample were shown at Phylum, Order, Class, Family, Genus, Species level separately. The ratio of each species in certain sample is directly displayed. At Phylum, all species were used to draw the histogram. The species of which abundance is less than 0.5% in all samples were classified into 'others' in other ranks.

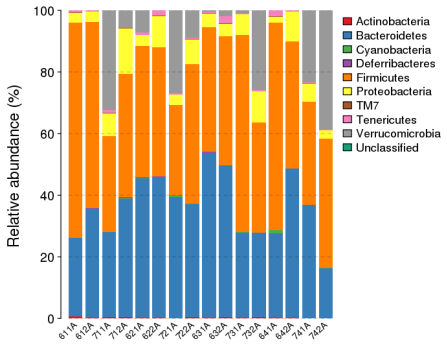



**Figure 3-4-1 The taxonomic composition distribution in samples of Phylum-level**

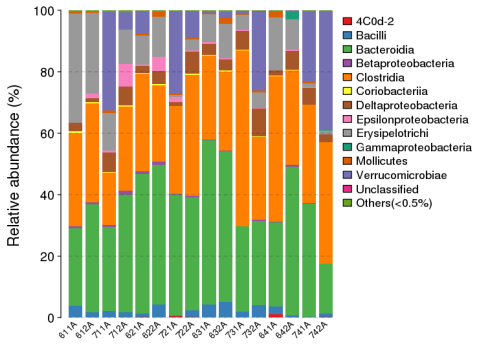



**Figure 3-4-2 The taxonomic composition distribution in samples of Class-level**

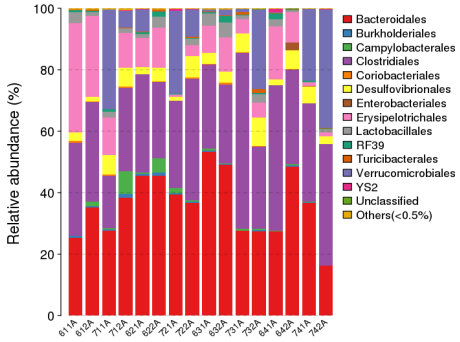



**Figure 3-4-3 The taxonomic composition distribution in samples of Order-level**

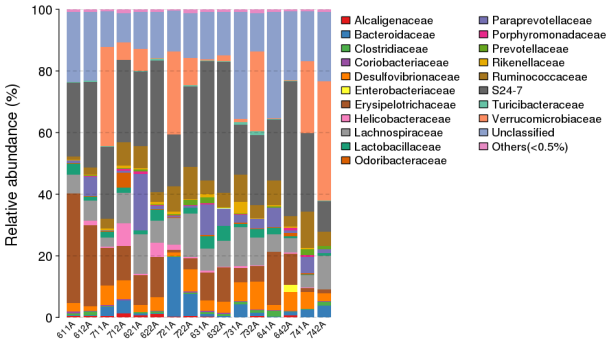



**Figure 3-4-4 The taxonomic composition distribution in samples of Family-level**

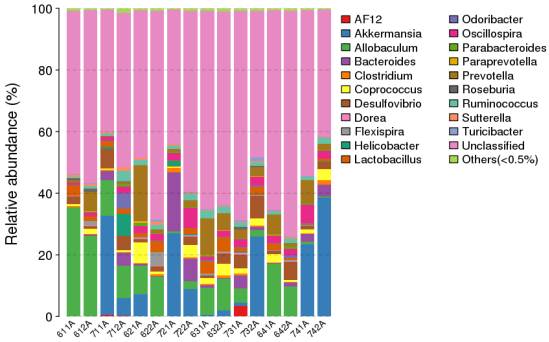



**Figure 3-4-5 The taxonomic composition distribution in samples of Genus-level**

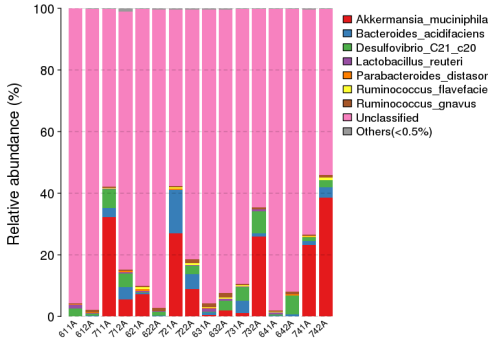



**Figure 3-4-6 The taxonomic composition distribution in samples of Species-level**

### 3.2.2 Specie heat map

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. Here species clustering based on the abundance of each species

was shown by heat map, longitudinal clustering indicates the similarity of all species among different samples, and the horizontal clustering indicates the similarity of certain species among different samples, the closer the distance is and the shorter the branch length is, the more similar the species composition is between the samples. At Phylum, all species were used to draw the heat map. The species of which abundance is less than 0.5% in all samples were classified into 'others' in other ranks. **Figure 3-5-*** shows the species clustering heat map at different taxonomic ranks.

Species heat map analysis was done based on the relative abundance of each species in each sample. To minimize the differences degree of the relative abundance value, the values were all log transformed. If the relative abundance of certain species is 0, the half of the minimum abundance value will substituted for it. Heatmaps were generated using the package 'gplots' of software R(v3.1.1) and the distance algorithm is 'euclidean', the clustering method is 'complete'.

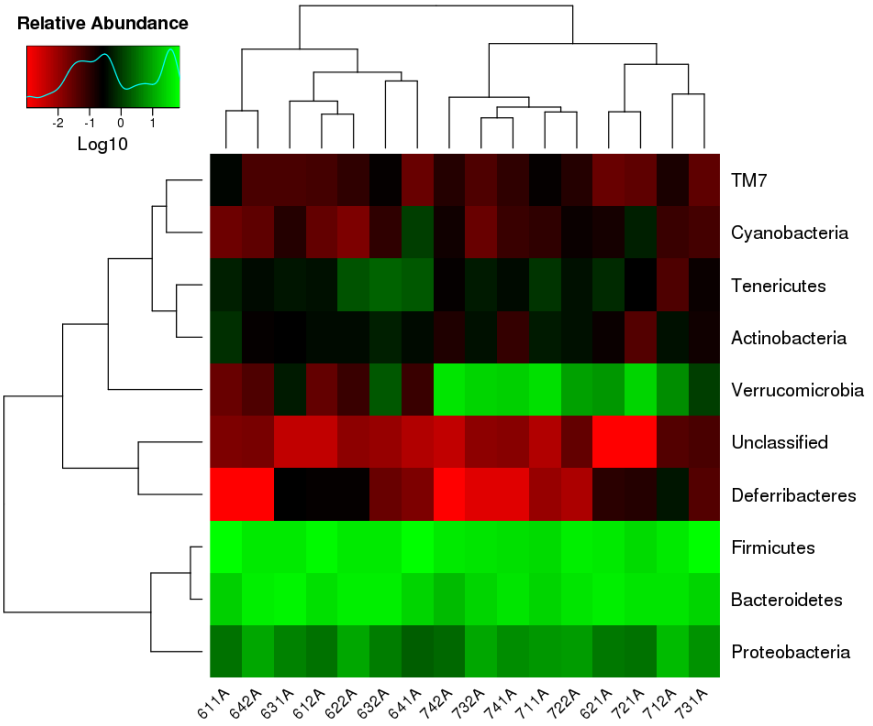



**Figure 3-5-1 Log-scaled percentage heat map of Phylum-level**

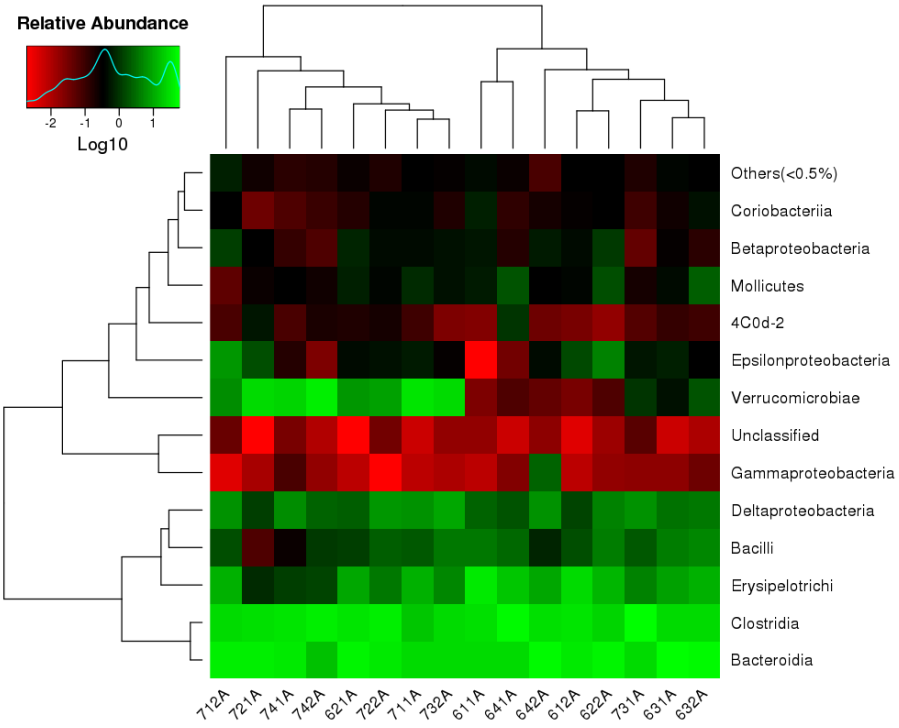



**Figure 3-5-2 Log-scaled percentage heat map of Class-level**

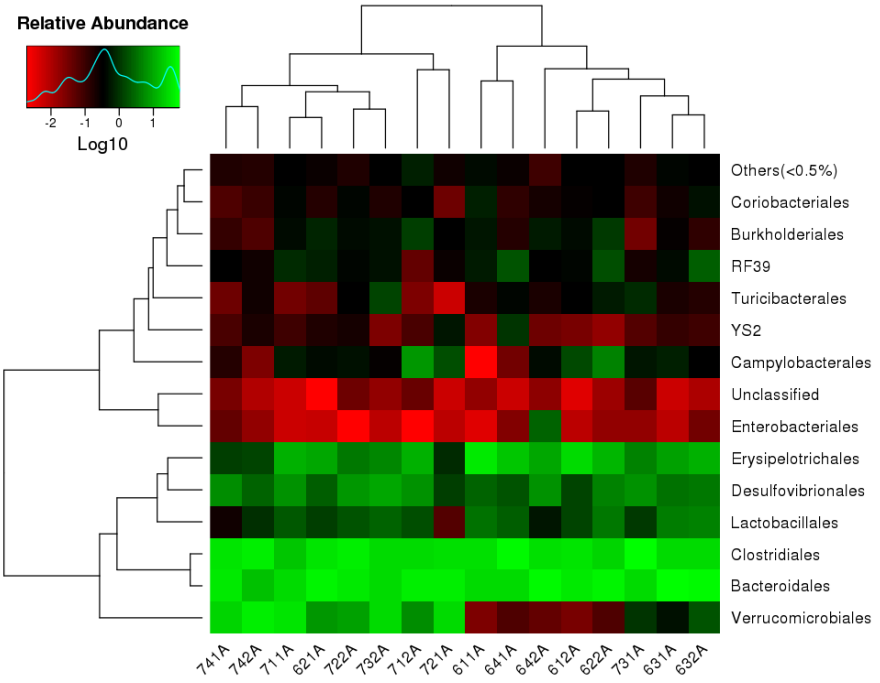



**Figure 3-5-3 Log-scaled percentage heat map of Order-level**

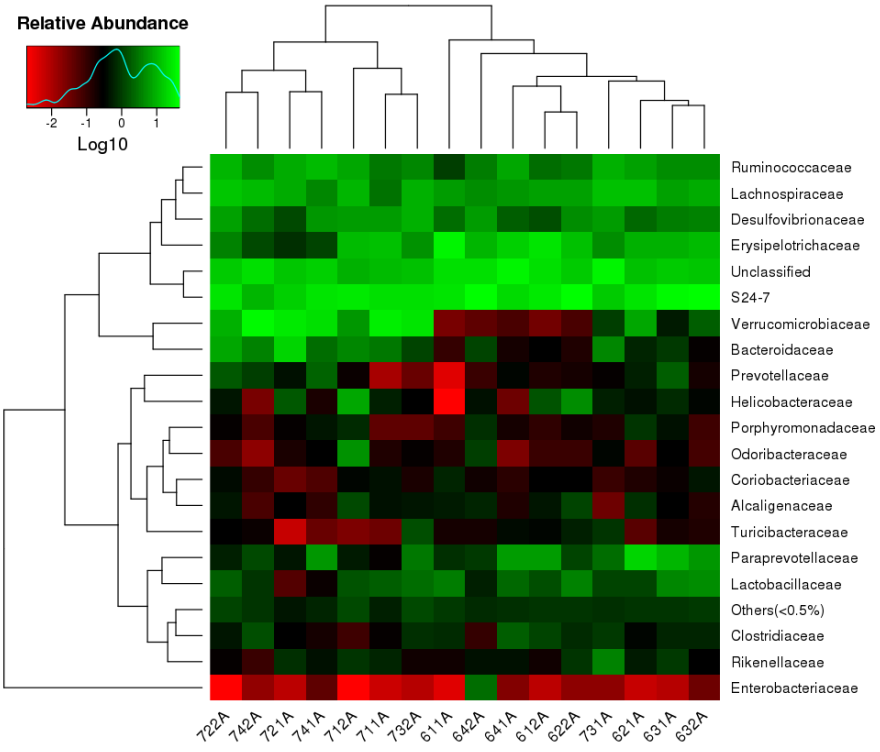



**Figure 3-5-4 Log-scaled percentage heat map of Family-level**

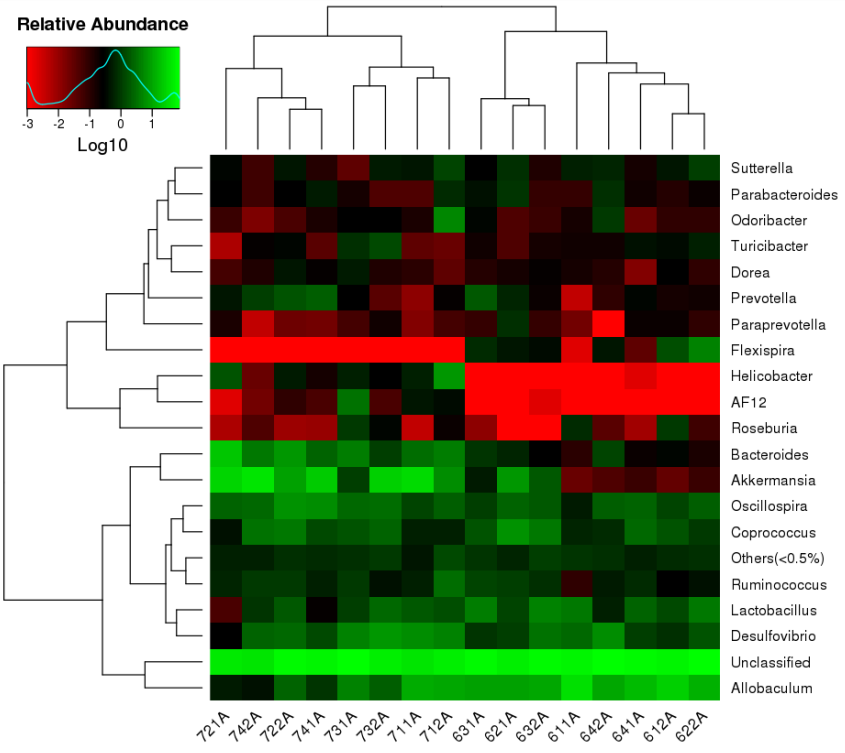



**Figure 3-5-5 Log-scaled percentage heat map of Genus-level**

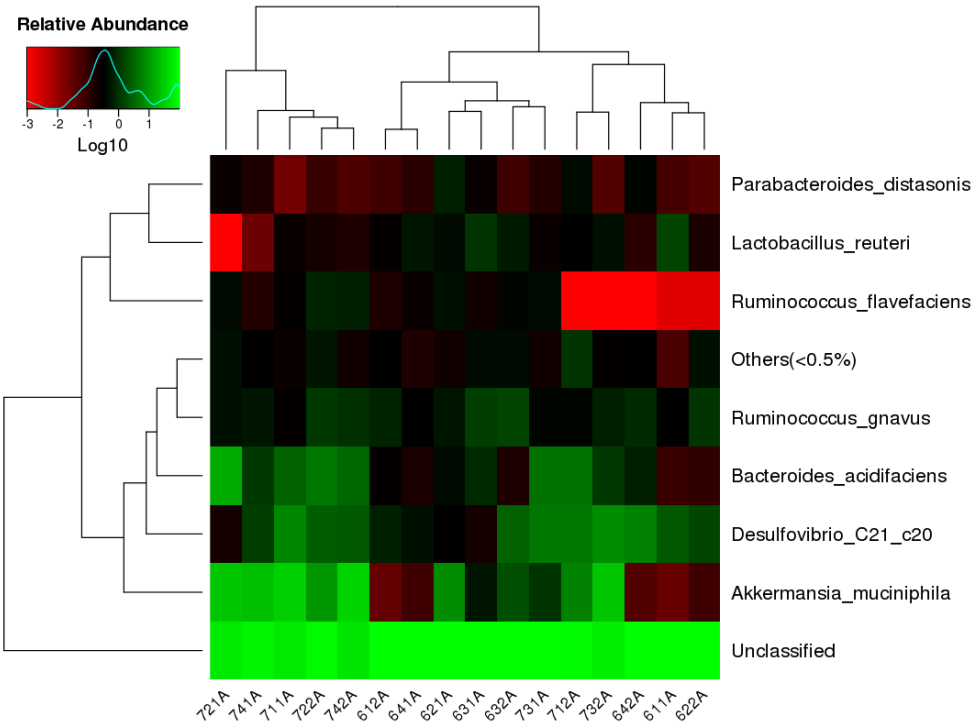



**Figure 3-5-6 Log-scaled percentage heat map of Species-level**

#### 3.2.3 Species phylogenetic analysis

A phylogenetic tree is a branching diagram showing the inferred evolutionary relationships among various biological species or other entities(their phylogeny) based upon similarities and differences in their physical or genetic characteristics. The evolution distance between species is closer if the branch length is shorter. Besides the species composition and abundance analysis above, phylogenetic tree could clarify the species evolution relationship further.

The representative sequences were aligned against the Silva core set(Silva_108_core_aligned_seqs) using PyNAST by 'align_seqs.py'. A representative OTU phylogenetic tree was constructed using the QIIME(v1.80) built-in scripts including the fasttree method for tree construction. The tags with highest abundance of each Genus was chosen as the corresponding Genus representative sequences, and Genus level phylogenetic tree was obtained by the same way of OTU phylogenetic tree. The phylogeny tree was imaged by software R(v3.1.1) at last.

**Genus species phylogeny tree**

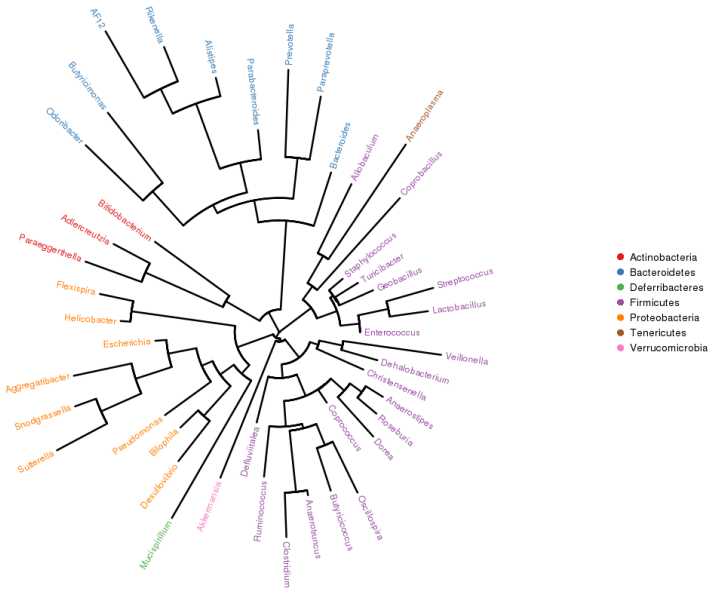



**Figure 3-6 Genus level phylogenetic tree**(The same Phylum is shown as the same color)

**Results directory: BGI_results/OTU_Cluster_Taxonomy/**

## 4 Diversity analysis

### 4.1 Diversity analysis with single sample

Alpha diversity is applied for analyzing complexity of species[11] diversity for a sample through several indices, including observed species, chao1, ace, shannon and simpson. The complexity of sample is proportional with the first four values, while with a negative correlation with simpson value.

Observed species value, chao1 value and ACE value can reflect the species richness of community, and the rarefaction curve based on the three values could also be used to evaluate if produced data is enough to cover all species in the community. When the curve tends to be smooth, it suggests the produced data is enough. Otherwise, when the curve continues to climb with increasing sequencing effort, it shows a high complexity in samples, and there still be species uncovered by the sequencing data.

Shannon value and simpson value can reflect the species diversity of the community, affected by both species richness and species evenness, that is the two values also consider the abundance of each species. With the same species richness, the greater the species evenness, the greater the community diversity.

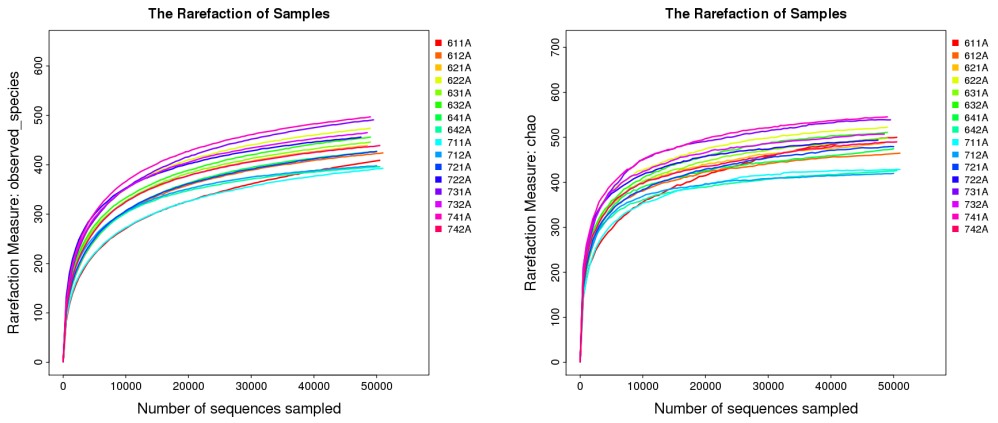
The indices are calculated by Mothur(v1.31.2), and the corresponding rarefaction curve are drawn by software R(v3.1.1). The calculation formula of each indice can refer to http://www.mothur.org/wiki/Calculators and the method of drawing rarefaction curve is as follows,
1) Calculating OTU numbers based on extracted tags (in multiples of 500);
2) Rarefraction curve was drawn using the indices calculated with extracted tags.

If the samples belong to different groups, and the sample number in per group is more than 3, differential analysis among groups could be done using the alpha diversity indices. And Wilcoxon Rank-Sum Test is used for two groups comparation, while Kruskal-Wallis Test is used for multi-groups comparation. And then plotbox of alpha diversity is drawn, the analysis above is done by software R(v3.1.1).

**Table 4-1 Alpha diversity statistics**

| Sample Name | sobs | chao | ace | shannon | simpson | coverage |
|---|---|---|---|---|---|---|
| 611A | 409.000000 | 500.000000 | 496.095952 | 3.305848 | 0.107189 | 0.998212 |
| 612A | 424.000000 | 464.738095 | 461.014661 | 3.606692 | 0.076133 | 0.998846 |
| 621A | 439.000000 | 487.157895 | 480.589014 | 3.989761 | 0.051323 | 0.998753 |
| 622A | 474.000000 | 522.372093 | 516.858605 | 4.208696 | 0.036014 | 0.998680 |
| 631A | 446.000000 | 499.333333 | 491.166736 | 3.955057 | 0.055367 | 0.998686 |
| 632A | 456.000000 | 510.886364 | 508.797702 | 4.130718 | 0.035861 | 0.998575 |
| 641A | 427.000000 | 474.527778 | 460.443019 | 3.731458 | 0.063911 | 0.998825 |
| 642A | 396.000000 | 426.100000 | 419.430355 | 3.777243 | 0.058066 | 0.999152 |
| 711A | 393.000000 | 428.400000 | 435.857603 | 3.398250 | 0.119816 | 0.998833 |
| 712A | 398.000000 | 419.794118 | 418.058471 | 4.311119 | 0.025888 | 0.999226 |
| 721A | 427.000000 | 479.317073 | 472.313024 | 3.614581 | 0.098802 | 0.998687 |
| 722A | 456.000000 | 493.657895 | 492.467038 | 4.469225 | 0.026079 | 0.998873 |
| 731A | 491.000000 | 538.788462 | 535.474223 | 4.262088 | 0.030033 | 0.998573 |
| 732A | 465.000000 | 507.775000 | 502.229190 | 3.998258 | 0.077648 | 0.998794 |
| 741A | 497.000000 | 545.468085 | 541.751610 | 3.982309 | 0.069966 | 0.998617 |
| 742A | 439.000000 | 489.833333 | 476.645198 | 3.447772 | 0.155148 | 0.998802 |

**Figure 4-1** shows the different curve based on observed species value, chao1 value and ACE value, shannon value, simpson value.



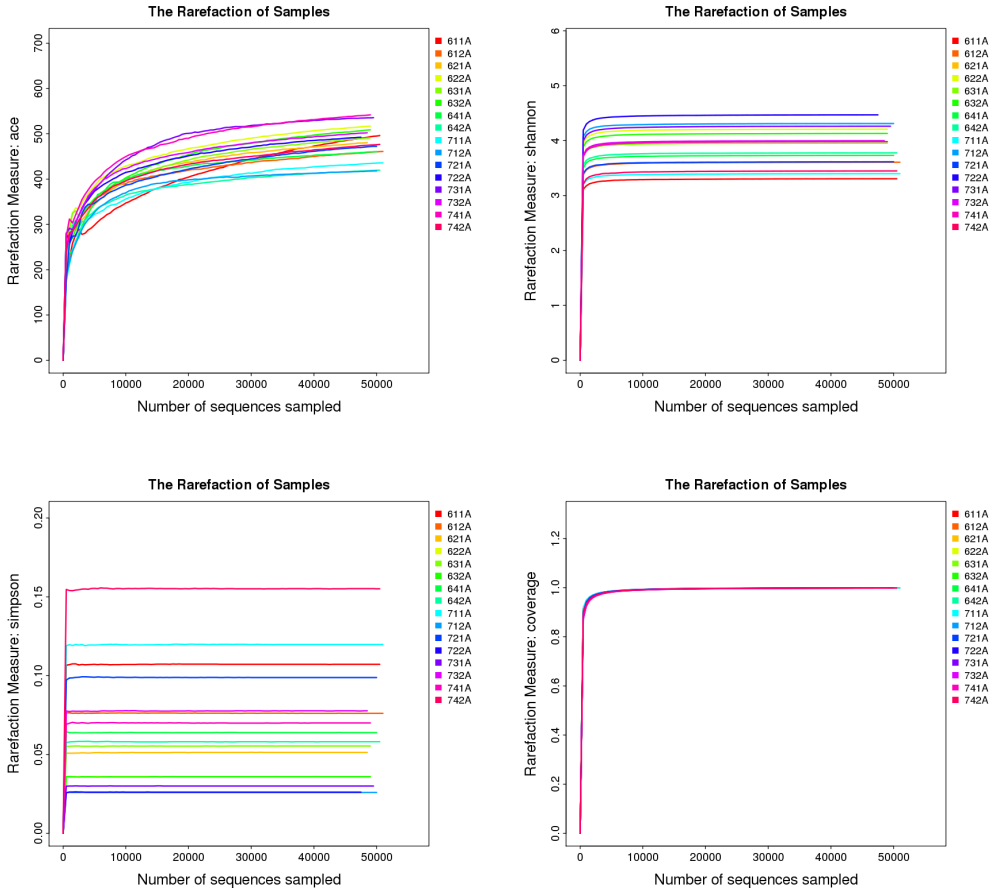


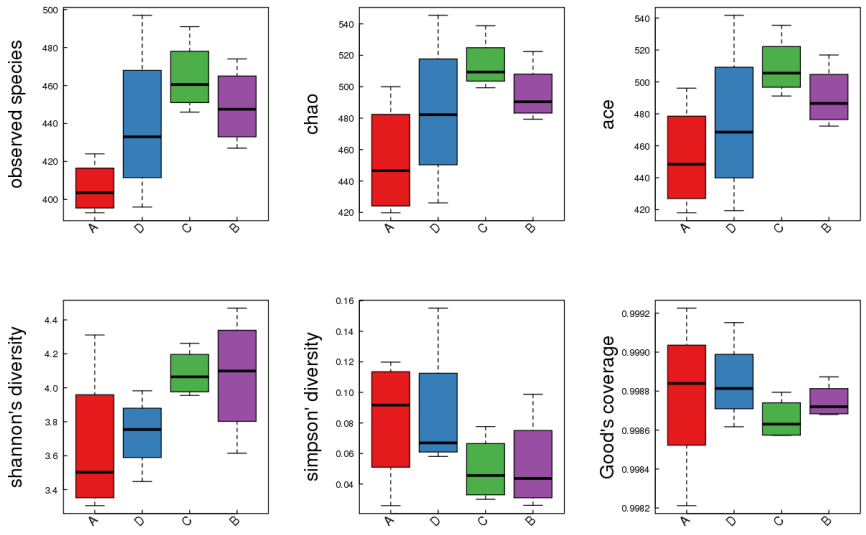

**图4-1 Sample-based rarefaction analysis**

The mean and standard deviation(SD) were calculated based on the alpha diversity values of all samples. If p value is less than 0.05, there is significant difference in alpha diversity among the groups. **Table 4-2-*** lists the comparison results and only shows the value of two groups, calculated value of other groups is detailed in the results file.

**Tabel 4-2-1 Alpha diversity comparison results among groups(Description)**

| #Alpha | mean(A) | SD(A) | mean(B) | SD(B) | ... | p-vaule |
|---|---|---|---|---|---|---|
| sobs | 406.00000 | 13.73560 | 439.75000 | 42.24827 | ... | 0.04547 |
| chao | 453.23305 | 36.76210 | 483.98230 | 49.17481 | ... | 0.14175 |
| ace | 452.75667 | 33.84293 | 474.56755 | 50.85151 | ... | 0.16493 |
| shannon | 3.65548 | 0.45485 | 3.73470 | 0.22020 | ... | 0.17782 |
| simpson | 0.08226 | 0.04182 | 0.08677 | 0.04584 | ... | 0.41605 |
| coverage | 0.99878 | 0.00042 | 0.99885 | 0.00022 | ... | 0.35259 |

Boxplot is used to visually display the differences of the alpha diversity among groups(**Figure 4-2-***). Five lines from bottom to top is the minimum value, the first quartile, median, the third quartile and the maximum value, and the abnormal value is shown as 'o'.




**Fighure 4-2-1 Alpha diversity indices boxplot among groups(Description)**

**Results directory: BGI_results/Alpha_Diversity/**

### 4.2 Diversity analysis among samples(n>=4)

Beta diversity analysis was used to evaluate differences of samples in species complexity. Beta diversity analysis was done by software QIIME(v1.80). There is differences in sequencing depth in different samples, normalization is introduced: Sequences is extracted randomly according to the minimum sequence number for all samples, the extracted sequences formed a new 'OTU table biom' file, then the beta diversity distance will be calculated base on the 'OTU table biom' file.

Various values, such as Bray-Curtis, weighted UniFrac, unweighted UniFrac, pearson, could be used to measure beta diversity, especially the first three values.

Bray-Curtis distance is a commonly used index to reflect the differences between two communities, and its value is between zero and one, zero Bray-Curtis represents exact similar community structure.

UniFrac uses the system evolution information to compare the composition of community species between samples. The results can be used as a measure of beta diversity. It takes into account the distance of evolution between the species, and the bigger index is the greater differences between samples. The UniFrac is divided into weighted UniFrac and unweighted UniFrac, and the weighted UniFrac considering the abundance of sequences while unweighted UniFrac do not.

**Beta diversity result:**

weighted_unifrac result：weighted_unifrac result

unweighted_unifrac result：unweighted_unifrac result

bray_curtis result：bray_curtis result

Heatmap of Beta diversity distance distribution is shown in **Figure 4-4-***. After clustering, the samples with similar beta diversity are clustered which reflect the similarity between the samples.

Beta diversity heat map was drawn by 'aheatmap' in package 'NMF' of software R(v3.1.1).

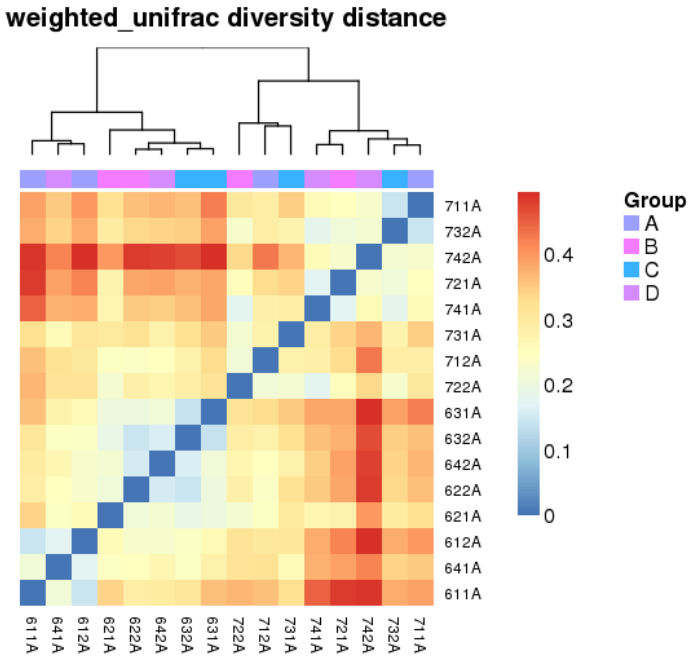



**Figure 4-3-1 Beta diversity heat map(Description, weighted_unifrac)**

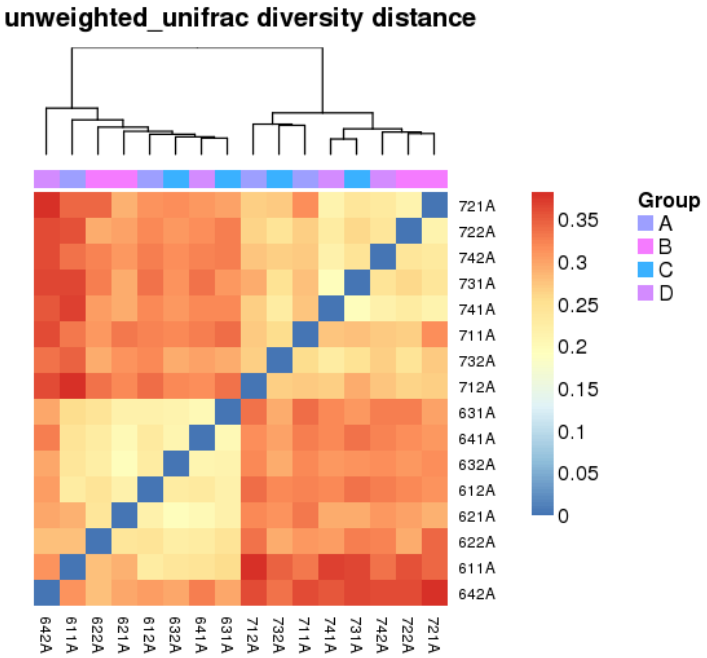



**Figure 4-3-2 Beta diversity heat map(Description, unweighted_unifrac)**

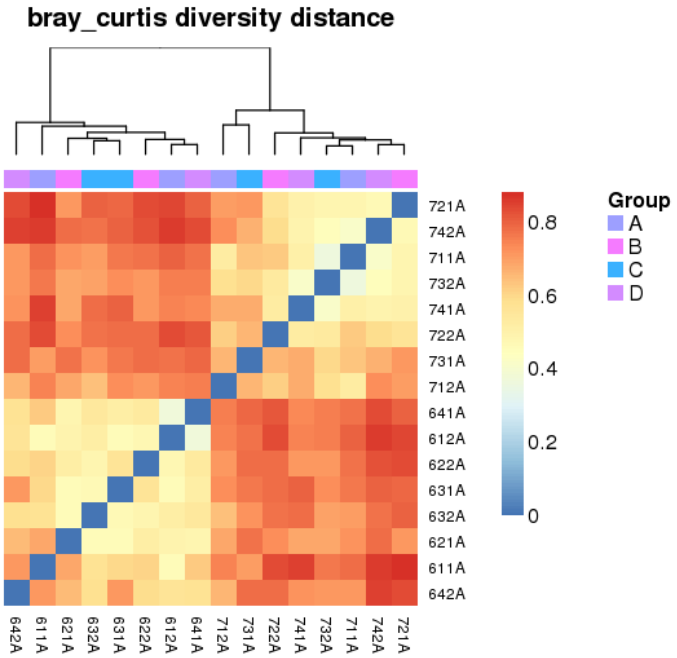



**Figure 4-3-3 Beta diversity heat map(Description, bray_curtis)**

Further, PCoA (Principal coordinate analysis) is used to exhibit the differences between the samples according to the matrix of beta diversity diatance. The closer distance stands for the similar species composition of the samples.

**PCoA result:**

weighted_unifrac PCoA 2D result: 2D PCoA result
weighted_unifrac PCoA 3D result:3D PCoA result
unweighted_unifrac PCoA 2D result: 2D PCoA result
unweighted_unifrac PCoA 3D result:3D PCoA result
bray_curtis PCoA 2D result: 2D PCoA result
bray_curtis PCoA 3D result:3D PCoA result

### 4.3 Clustering of Species Composition Among Samples (n>=4)

Similarity in species composition among samples was evaluated. The clustering results are shown in **Figure 4-4-*** and the same color represents the samples in the same group. Short distance between samples represents high similarity.

Unweighted Pair Group Method with Arithmetic mean (UPGMA) is a type of hierarchical clustering method using average linkage and was used to interpreting the distance matrix produced by beta diversity. To measure the robustness of this result to sequencing effort, we perform a jackknifing analysis, wherein 75% of the smallest sample sequences from each sample are chosen at random, and the resulting UPGMA tree from this subset of data is compared with the tree representing the entire available data set by QIIME(v1.80). This process is repeated with 100 random subsets of data, and the tree nodes which prove more consistent across jackknifed datasets are deemed more robust. And the figure is drawn by software R(v3.1.1).

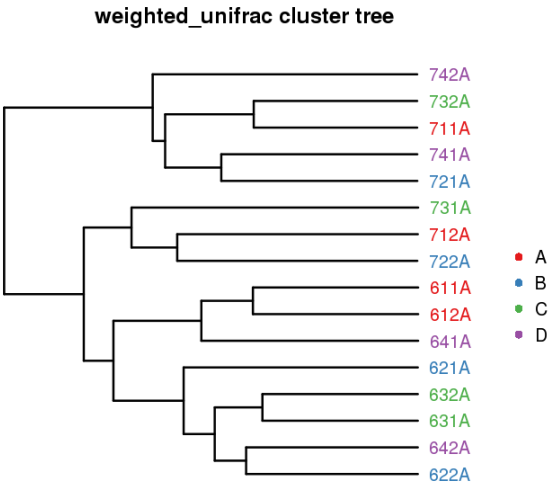



**图4-4-1 Samples Clustering result(Description, weighted_unifrac)**

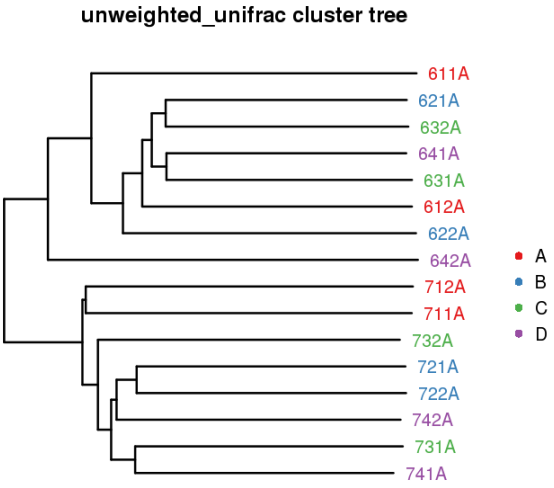



**图4-4-2 Samples Clustering result(Description, unweighted_unifrac)**

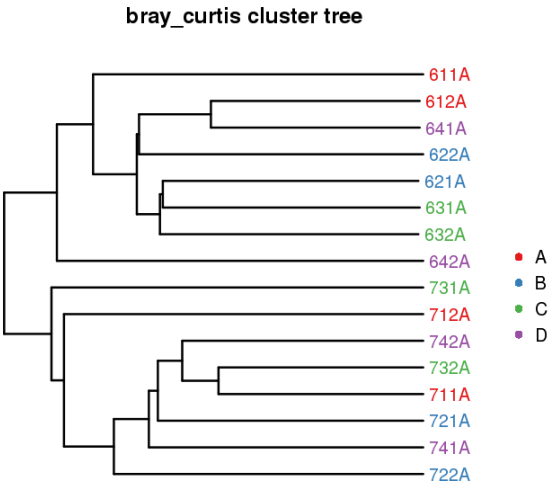



**图4-4-3 Samples Clustering result(Description, bray_curtis)**

**Results directory: BGI_results/Beta_Diversity/**

### 5 Significant differences analysis between groups of samples (groups>=2, samples per group>=3)

We use the method of statistical to get the abundance differences of microbial communities between samples, and FDR (false discovery rate) is adopted to assess the significance of differences. From the results, we can identify the samples that cause the species composition differences between two groups. The significant difference between the groups are analyzed at the level of Phylum, Class, Order, Family, Genus and Species.

Metastats(http://metastats.cbcb.umd.edu/) and R(v3.1.1) are used to determine which taxonomic groups were significantly different between groups of samples. We adjusted the obtained P-value by a Benjamini-Hochberg false discovery rate correction(function 'p.adjust' in the stats package of R(v3.1.1))[12].

**Method: kruskal.test**

**Results directory: BGI_results/Diff_Analysis/**

## 6 LDA Effect Size(LEFSE) Analysis)

LEfSe (Linear discriminant analysis Effect Size)[9] determines the features (organisms, clades, operational taxonomic units, genes, or functions) most likely to explain differences between classes by coupling standard tests for statistical significance with additional tests encoding biological consistency and effect relevance.

The analysis is done by software LEFSE.

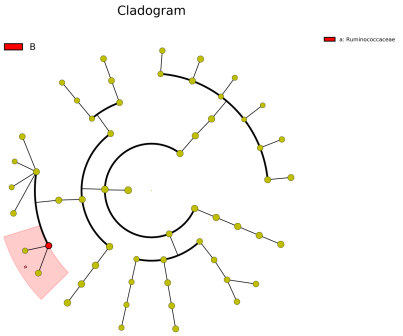



**Figure 6-1-1 LEFSE Analysis**

In the LEFse tree, different colors indicate different groups. Note colored in a group color shows an important microbe biomarker in the group and the biomark name will list in the upper right corner. The yellow notes represent the biomarker which do not show any importance in groups.

**Results directory: BGI_results/Diff_Analysis/LEFSe/**

### 7 PICRUST Analysis)

Using 16S information, PICRUSt[10] recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with quantifiable uncertainty. Here, we used PICRUSt to do functional classification scheme of KEGG Orthology (KOs) and Clusters of Orthologs Groups (COGs). Following table shows the significant differences analysis result of KOs/COGs.

The analysis is done by software PICRUST.

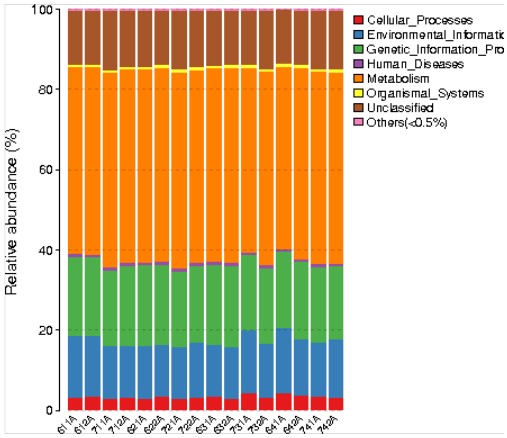



**Figure 7-1-1 PICRUST Analysis**

KOs prediction. In the figure, the x axis indicates different samples/groups, and the y axis indicates the relative abundance. Each color represent a KEGG pathway.

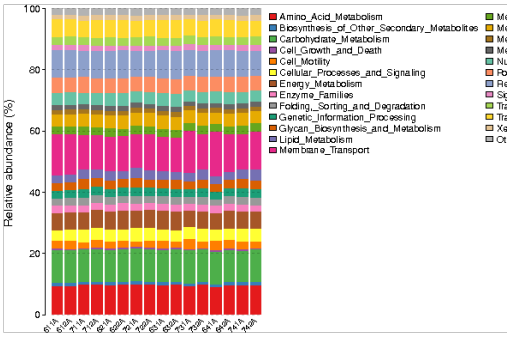



**Figure 7-1-1 PICRUST Analysis**

KOs prediction. In the figure, the x axis indicates different samples/groups, and the y axis indicates the relative abundance. Each color represent a KEGG pathway.

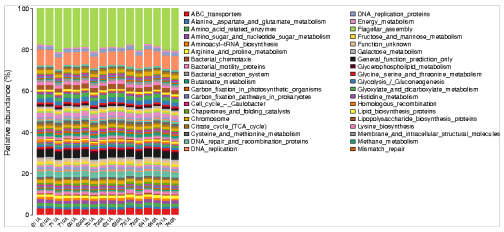



**Figure 7-1-1 PICRUST Analysis**

KOs prediction. In the figure, the x axis indicates different samples/groups, and the y axis indicates the relative abundance. Each color represent a KEGG pathway.

**Results directory: BGI_results/Functional_Analysis/**

## 8 Suggestions for Data Mining

16S/18S/ITS/Functional gene(FunGene, such as amoA, mcrA) analysis was mainly applied for host intestine, soil, water and so on. Analysis hotspots for this kind of research are mainly focused on differences among samples in species composition and corresponding abundance[13-17].

In analysis results, there are three important files in the directory **OTU_Cluster_Taxonomy**: **OTU_table_for_biom.txt** (containing species composition for each sample), **Species_heatmap**/ (visualized result of OTU taxonmy and abundance in each sample), **Taxa_summary** (including comparison among samples in different taxonomic levels. Please focus on the table in which different samples could be separated significantly, and then understand backgrounds of these species to correlate with environmental adaption). Results about sample diversity are mainly included in directory **Alpha_diversity**/ and **Beta_diversity**/ with files representing complexity in samples (Alpha diversity and rarefaction), complexity differences between samples(Beta diversity) etc.

For different groups, PCA on OTU abundance, heat map on species abundance and heat map on beta diversity all visually reflect the differences among groups. If the samples belonging to the same group were clustered together, the samples of the same group has the similar species composition. And **Diff_Analysis**/ could give the significant species responsible for the differences using statistics method.

## III References


[1] Douglas WF, Bing M, Pawel G, Naomi S, Sandra O, Rebecca MB. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:6.

[2] T. Magoc and S. Salzberg. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21): 2957-63.

[3] Edgar, R.C. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads,Nature Methods. 10(10):996-8.

[4] Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 27:2194-2200.

[5] J Gregory C, Justin K, Jesse S. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 7:335-336.

[6] Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl. Acids Res. 41:D633-D642.

[7] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Gl?ckner FO. 2013.: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.

[8] DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-72.

[9] Abarenkov, Kessy; Nilsson, R. Henrik; Larsson, Karl-Henrik; Alexander, Ian J.; Eberhardt, Ursula; Erland, Susanne; H?iland, Klaus; Kj?ller, Rasmus; Larsson, Ellen; Pennanen, Taina; Sen, Robin; Taylor, Andy F. S.; Tedersoo, Leho; Ursing, Bj?rn M.; Vr?lstad, Trude; Liimatainen, Kare; Peintner, Ursula; K?ljalg, Urmas. 2010. The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytologist. 186(2), 281-285.

[10] Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 73:5261-5267.

[11] Patrick DS, Sarah LW et al. (2009). Introducing mothur: Open-Source, Platform- Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75(23):7537-7541.

[12] James RW, Niranjan N, Mihai P. 2009. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS computational biology.

[13] McCafferty J, Mühlbauer M, Gharaibeh RZ, et al. (2013) Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J. 7(11):2116-25.

[14] Zhao L, Wang G, Siegel P, et al. (2013) Quantitative genetic background of the host influences gut microbiomes in chickens. Sci Rep. 3:1163.

[15] Rubin BE, Gibbons SM, Kennedy S, et al. (2013) Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS One. 8(7):1:6.

[16] Mao Y, Xia Y, Zhang T. (2013) Characterization of Thauera-dominated hydrogen-oxidizing autotrophic denitrifying microbial communities by using high-throughput sequencing. Bioresour Technol. 128:703-10.

[17] Peng X, Yu KQ, Deng GH, et al. (2013) Comparison of direct boiling method with commercial kits for extracting fecal microbiome DNA by Illumina sequencing of 16S rRNA tags. J Microbiol Methods. 95(3):455-62.

## IV Data format introdution

### 1 FASTQ format

FASTQ output format:
The first read in *1.fq:
@FC4290FAAXX:4:1:3:84#CAGATC/1
AGTTCCGGCGCACGGGTGAGTAACGCGTATCCAACCTTCCCCTTAGTAGGGCATAGCCCGGCGAAAGTCGGATTAATACTCTATGTTTTCCGTCGAGGACATCTGAAGTGGAACAAAGATT
+
CCCCFGCGGGFCCFGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGAFGFFGGFGGGGGGDGGGECBCFGDGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGEGGGGFGGGEFFEGFGFCF

The first read in *1.fq:
@FC4290FAAXX:4:1:3:84#CAGATC/2
CCGCTGCTGCTGGCACGGAATTAGCCGGTCCTTATTCATCAGGTACCTACAAAAAGGGACACGTCCCTCACTTTATCCCCTGATAAAAGCAGTTTACAACCCATAGGGCCGTCATCCTGCA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGEDAFGGGGGGGGFEGGGFGGGGGGGGCFCGFFGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

**Format explanation (from left to right).**

| Line | File Description |
|---|---|
| 1 | @Reads ID |
| 2 | Sequence |
| 3 | + Reads ID （Reads ID could be omitted） |
| 4 | Base quality |

### 2 FASTA format

Fasta format (also called Pearson format) which based on the format of text is used to record the sequences of DNA and protein. In the format, the sequence of DNA and protein are coded by single characters, and the name is allowed to be added as annotation at the beginning of each sequence. The first line of the sequence file is start by the symbol of '>' or ';' which is followed by annotation. The sequence is started from the second line, and only allowed characters can be used for coding. For DNA sequences, capitals or lowercases can be used for coding; for protein sequence, only capitals can be used. For example:
>Sample1_Tag1
TGAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGCAGGATGACGGTCCTATGGATTGTAAACTGCTTTTGTACGAGAAGAAAC
ACTCCTACGTGTAGGAGCTTGACGGTATCGTAAGAATAAGGATCGGCTAACTCC

The tags and OTU sequences are all FASTA format.

### 3 Illustration of OTU files

**OTU_taxonomy.xls** example:

| #OTUId | Abundance | Taxonomy |
|---|---|---|
| Otu1 | 1200 | Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;Acinetobacter_lwoffii |
| Otu2 | 900 | Bacteria;Firmicutes;Clostridia;Thermoanaerobacterales;Thermodesulfobiaceae;Coprothermobacter |
| Otu3 | 4100 | Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas |

From left to right, the column contains the information of OTU ID, The OTU Abundance, and The Annotated Species Information of OTU.

**OTU_table_for_biom.txt** example:

| #OTU ID | Sample1 | Sample2 | Sample3 | taxonomy |
|---|---|---|---|---|
| Otu1 | 100 | 300 | 800 | Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter_lwoffii |
| Otu2 | 500 | 200 | 200 | Bacteria; Firmicutes; Clostridia; Thermoanaerobacterales; Thermodesulfobiaceae; Coprothermobacter |
| Otu3 | 1000 | 1000 | 2100 | Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas |

The first column is OTU ID, column 2 to the second last column contains the tag number of each OTU in all samples, and the last column is the The Annotated Species Information of OTU.

**OTU_stat_detail.xls** example:
Num samples: 3
Num total OTUs: 121
Num Singletons: 0
Num Non-singletons: 121
Num total sequences: 6200

Sample all OTU summary:
Min: 64
Max: 83
Mean: 73
Standard deviation: 7

Sample detail:
Sample name Tag number OTU number
Sample1 1200 64
Sample2 900 72
Sample3 4100 79

Num samples: the number of samples; Num total OTUs: total OTU number; Num Singletons: Number of OTU with only 1 tag; Num Non-singletons: Number of OTU with 2 or more tags; Num total sequences: total number of tags.
Sample all OTU summary: The summarized information of OTU, including minimum number of OTU, the maximum number of OTU, the average number of OTU, the standard deviation.
Sample detail: the number of each sample's OTU. The first column is Sample ID, the second column is the total OTU number in each sample, the third column is the number of OTU without singletons, the fourth column is the ratio of OTU with singleton in all OTU.

***sharedOTU.venn.xls** example:

| group | shareOTU_num/uniuqeOTU_num | OTU_ID |
|---|---|---|
| Sample1-vs-Sample2 | 80 | Otu1,Otu10,Otu100,Otu101,Otu102,Otu103... |
| Sample1-vs-Sample3 | 85 | Otu1,Otu10,Otu100,Otu101,Otu102,Otu103... |
| Sample1 | 5 | Otu113,Otu116,Otu118,Otu60,Otu98 |

The corresponding excel sheet of Venn diagram with statistics of shared and unique OTU of samples or groups. The first column is sample or group ID, the second column is the number of shared OTU and unique OTU, and the third column is the shared or unique OTU ID.

**OTU_PCA.coordinate.xls** example:

| | Axis1 | Axis2 |
|---|---|---|
| Sample1 | 1.535 | -3.381 |
| Sample2 | 0.5531 | -3.0168 |
| Sample3 | 3.2705 | 3.4669 |

The first column is sample ID, the second column is the value of x-axis(PC1), and the third column is the value of y-axis(PC2).

**Genus.phylogeny.tree**: This file is in Newick format which is the usual format of phylogeny tree and could be identified by several software such as FigTree,Treebest, PHYLIP etc.

## 4 Alpha diversity analysis files

**Alpha_diversity.detail.xls** example:

| label | group | sobs | chao | chao_lci | chao_hci | ace | ace_lci | ace_hci | ... |
|---|---|---|---|---|---|---|---|---|---|
| 0.03 | Sample1 | 156.000000 | 521.500000 | 335.876875 | 898.676067 | 378.948085 | 283.807259 | 544.912564 | ... |
| 0.03 | Sample2 | 151.000000 | 297.176471 | 226.692999 | 433.292428 | 294.071360 | 231.482437 | 405.333926 | ... |

The first column is the differential level, the second column is the sample name, the third column is the obs value of alpha diversity, and from the fourth column to the last one, every 3 columns represent the alpha diversity value, the minimum value and the maximum value at 95% confidential intervals respectively.

***.Alpha.test.reslut.xls** example:

| #Alpha | mean(Group1) | SD(Group1) | mean(Group2) | SD(Group2) | ... | p-vaule |
|---|---|---|---|---|---|---|
| sobs | 191.00000 | 43.24350 | 682.50000 | 166.91016 | ... | 0.00075 |
| chao | 292.44654 | 16.15483 | 898.39669 | 154.11534 | ... | 0.00063 |
| ace | 382.62500 | 177.89841 | 854.80273 | 154.54333 | ... | 0.00114 |
| shannon | 2.33207 | 4.34743 | 3.04424 | 4.27114 | ... | 0.00066 |
| simpson | 0.18282 | 0.03691 | 0.15808 | 0.04751 | ... | 0.00167 |

The first column is the name of alpha diversity index, and from column 2, every two columns represent the mean value and SD value of each sample or group, and the last column is the p-value of the differential test for alpha diversity among groups.

## 5 Beta diversity analysis files

***Beta_diversity.xls** example:

| | Sample1 | Sample2 | Sample3 | Sample4 |
|---|---|---|---|---|
| Sample1 | 0.0 | 0.529648285613 | 0.690040650407 | 0.570607988689 |
| Sample2 | 0.529648285613 | 0.0 | 0.597119123365 | 0.449893955461 |
| Sample3 | 0.690040650407 | 0.597119123365 | 0.0 | 0.708907741251 |
| Sample4 | 0.570607988689 | 0.449893955461 | 0.708907741251 | 0.0 |

The document for the symmetric matrix values in the table for the first row and first column of the sample twenty-two calculated distance matrix.

## 6 Differences analysis files

***differentially_abundant.xls** example:

| Phylum | mean(Group1) | variance(Group1) | std.err(Group1) | mean(Group2) | variance(Group2) | std.err(Group2) | p-vlaue | FDR |
|---|---|---|---|---|---|---|---|---|
| Actinobacteria | 0.0490001 | 4.8e-05 | 0.039710 | 0.073479 | 0.000135 | 0.0601112 | 0.653410 | 0.682357 |
| Proteobacteria | 0.014310 | 2e-06 | 0.008470 | 0.018354 | 1e-06 | 0.0069 | 0.685421 | 0.676623 |

The first column is species names; the second, third and fourth column are average abundance , variance and standard deviation of Group1 respectively; the fifth, sixth and seventh column are average abundance, variance and standard deviation of Group2 respectively; the eighth column is the p-value of two groups; the ninth column is the false discovery rate values of two groups.