1. Flow chart of Bioinformatics analysis

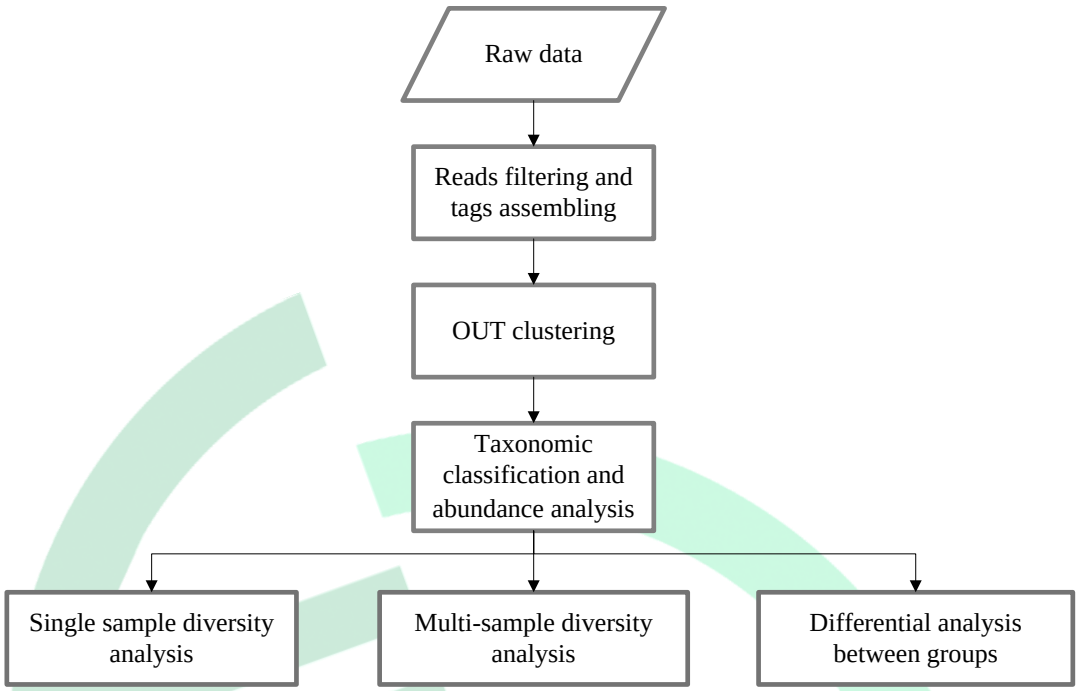

Fig.1 Flow chart of 16S rDNA bioinformatics analysis

2. Reads filtering

Raw data obtained after sequencing included dirty reads containing adapters or low quality bases which would affect the following assembly and analysis. Thus, to get high quality clean reads, raw reads were further filtered according to the following rules:

1) Removing reads containing more than 10% of unknown nucleotides (N);

2) Removing reads containing less than 80% of bases with quality (Q-value)＞20.

3. Tags assembling and abundance statistic

The filtered reads were then assembled into tags according to overlap between paired-end reads with more than 10bp overlap, and less than 2% mismatch. The software Mothur (v.1.34.0) [1] was used to remove the redundant tags to get unique tags. The obtained unique tags were then used to calculate the abundance.

4. Taxonomic classification of Tags

The software rdp classifier [2] was used to classify tags into different taxonomies against greengene database[3] (version 20101006) with Confidence Threshold of 0.5.

5. Advance analysis on Operational Taxonomic Unit (OTU) level

5.1 OTU clustering and abundance analysis

The software Mothur was used to cluster tags of more than 97% identity into OTUs, and then the abundances of OTUs were calculated.

5.2 Taxonomic classification of OTU

The taxonomic classification of OTUs was based on annotation result of contained tags according to the mode principle, that is, the taxonomic rank which contained more than 66% of tags was thought to be the taxonomic rank of this OTU, otherwise the higher rank would be considered.

The taxonomic ranks in descending order of size are: domain, phylum, class, order, family, genus, species.

5.3 OTU pathway annotation

The software PICRUSt[4] (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) was used to annotate pathways of OTUs against KEGG database.

5.4 OTU rarefaction curve

The rarefaction curve was used to evaluate whether the sequencing data amount was enough to cover all of the sample species and to reflect the species richness in samples. When the curve is becoming gentle or reaching a plateau, the sequencing data is enough to cover all of the sample species.

5.5 OTU alpha diversity

The alpha diversity represents species diversity in a single sample and was evaluated by several diversity indices such as chao1 value, ACE value, Shannon index and Simpson index [5], etc. The chao1 value and ACE value are used to predict microorganism species in a sample according to the tags numbers, OTU numbers and their relative proportion. The Shannon index is a diversity index considering OTU abundance and OTU evenness. The greater the Shannon index and npShannon index, the more diverse the species in a sample. The Simpson index is another diversity index to measure species diversity in a single sample. There are two ways to measure Simpson index[6]:

1)Simpson =$\Sigma(Pi)^2$

2) Simpson=1-$\Sigma(Pi)^2$

Here Pi is the proportional abundance of one species among the total species in a sample. The Simpson

index is constrained to between 0 and 1. The smaller the Simpson index, the more diverse the species in samples. The first method was used in our analysis.

5.6 OTU Shannon rarefaction curve

The Shannon rarefaction curve was used to evaluate whether the sequencing data amount was enough to cover all of the sample species and to reflect the species richness in samples. When the curve reaches a plateau, it means that the amount of sequencing data is enough to cover most of the species in a sample.

5.7 OTU Rank Abundance curve

The Rank Abundance curve was used to reflect the richness and evenness of species in a sample. The greater the span of curve on X axis, the more diverse the species in a sample. The smoother the curve on Y axis, the more even the species in a sample.

5.8 OTU heatmap analysis (samples≥2)

To show differentially expressed OTUs among samples, a heatmap was drawn using expression profile of OTU by R package pheatmap.

The heatmaps were also drawn at taxonomic level.

5.9 OTU PCA analysis (samples≥3)

The Principal Component Analysis (PCA) was used to reflect the microbial compositional differences between samples and the relationship of samples. PCA was analyzed based on the expression profile of OTUs using R package.

The PCA was also analyzed at taxonomic level.

5.10 OTU beta diversity(samples≥2)

Beta diversity is the ratio of all OTUs and common OTUs between two samples, which is often used to determine the differentiation among groups. Based on the OTU expression profile, the beta diversity of different samples at OTU level was calculated. The formula was shown as follows[7]:

$$\beta = \frac{S}{\bar{\alpha}} - 1$$

Here S is the total numbers of OTUs from all of samples, α is the common OTUs numbers among samples. Beta diversity was used to reflect the species diversity among samples. The closer to 0 the value of beta

diversity, the more similar the species composition of two samples.

The beta diversity was also analyzed at taxonomic level.

5.11 OTU cluster analysis (samples≥3)

According to the OTU expression profile, the distance between samples was calculated and then cluster analysis was carried out using R package pvclust[8] to predict the sample similarity on OTU level. Two kinds of P-value including AU (Approximately Unbiased) p-value and BP (Bootstrap Probability) p-value were provided to evaluate the reliability of cluster. The higher the value, the higher the reliability.

The cluster analysis was also performed at taxonomic level.

5.12 OTU bray-curtis distance coefficient analysis (samples≥2)

According to the OTU expression profile, the bray-curtis distance coefficient between samples was calculated and used to cluster samples to evaluate the sample similarity on OTU level.

The bray-curtis distance coefficient analysis was also performed at taxonomic level.

6. Advance analysis on taxonomic level

6.1 Taxonomic abundance profile

According to the taxonomic classification and abundance of OTU, the abundance of each samples on each taxonomic level of domain, phylum, class, order, family, genus and species were calculated. In this way, it is easier to compare the different abundance of one species in multiple samples and to find out the significant different species among samples.

6.2 Taxonomic classification tree

According to the expression of taxonomic units, those with high expression level were chosen to construct a taxonomic classification tree.

6.3 Species composition analysis

To display and compare the abundance of species in different samples, the composition of species of each sample at each taxonomic level was calculated and showed by stack graph. The species with low abundance (which counted less than 2% of all species abundance) were classified into “Other” group, and the tags without annotations at this taxonomic level were classified into “Unclassified” group.

6.4 Taxonomic evolutionary tree

To display and compare the predominant species between samples, the most abundant OTUs were used to construct the evolutionary tree at each taxonomic level.

6.5 Taxonomic heatmap analysis (samples≥2)

Same as OTU level.

6.6 Taxonomic PCA analysis (samples≥3)

Same as OTU level.

6.7 Taxonomic beta diversity analysis(samples≥2)

Same as OTU level.

6.8 Taxonomic cluster analysis(samples≥3)

Same as OTU level.

6.9 Taxonomic bray-curtis distance coefficient analysis (samples≥2)

Same as OTU level.

7 Differential analysis between groups

7.1 Differential analysis between groups

The software Metastats was used to detect the differentially abundant microbial community between two samples, and the FDR value was used to evaluate difference significance.

7.2 Differential pathway functional analysis between groups

Based on the result of Metastats analysis, the significant differential species were filtered by |Log2(FC)| ≥1 and P-value≤0.05 to do pathway enrichment analysis.

7.3 Differential LEfSe analysis between groups

Linear discriminant analysis (LDA) effect size (LEfSe) method was used to identify the most differentially abundant taxons between groups which would help to discover biomarkers. LEfSe used the Kruskal-Wallis rank sum test to detect features with significantly different abundances between all groups. Next Wilcoxon rank sum test was used to detect features with significantly different abundances between two groups

and then LDA was performed to estimate the effect size of each feature. In addition, a taxonomic cladogram representative of the structure of microbial community of each sample and their predominant bacteria was drawn to display the greatest differences in taxa between groups.

8. Reference

[1]. 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

[2]. http://rdp.cme.msu.edu/classifier/classifier.jsp

[3]. DeSantis, T. Z., P. Hugenholtz, et al. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-72.

[4] M. G.I.*; Zaneveld, J.* et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Langille, Nature Biotechnology, 1-10. 8 2013.

[5]. Paul FK, Josephine Y A. (2010). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. FEMS Microbiol.Ecol 47:161-177

[6]. Frosini BV. (2003). Descriptive measures of ecological diversity. In Environmetrics, A.H. El-Shaarawi and J. Jureckova (Eds.), Encyclopedia of Life Support Systems (EOLSS), EOLSS Publishers, Oxford (UK) (http://www.eolss.net)

[7]. Whittaker, R.H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. Ecological Monographs 30:279–338

[8]. http://www.is.titech.ac.jp/~shimo/prog/pvclust/