Method

The method is only used as a reference for publication purpose. Customers are responsible for the related risks of duplicate checking.

1 Experimental Procedure

1.1 Sample Quality Control

Please refer to QC report for methods of sample quality control.

1.2 Amplicon Generation

16S rRNA/18SrRNA/ITS genes of distinct regions (16SV4/16SV3/16SV3- V4/16SV4-V5, 18SV4/18SV9, ITS1/ITS2, ArcV4) were amplified used specific primer (e.g. 16SV4: 515F- 806R, 18SV4: 528F-706R, 18SV9: 1380F- 1510R, et. al) with the barcode. All PCR reactions were carried out with 15 μL of Phusion® High - Fidelity PCR Master Mix (New England Biolabs); 0.2 μM of forward and reverse primers, and about 10 ng template DNA. Thermal cycling consisted of initial denaturation at 98℃ for 1 min, followed by 30 cycles of denaturation at 98℃ for 10 s, annealing at 50℃ for 30 s, and elongation at 72℃ for 30 s and 72℃ for 5 min.

1.3 PCR Products Quantification and Qualification

The PCR products were purified using magnetic bead purification. Samples were mixed in equidensity ratios based on the concentration of PCR products. After thorough mixing, the PCR products were detected and target bands were recovered.

1.4 Library Preparation and Sequencing

Sequencing libraries were generated and indexes were added.The library was checked with Qubit and real-time PCR for quantification and bioanalyzer for size distribution detection. Quantified libraries were pooled and sequenced on Illumina platforms, according to effective library concentration and data amount required.

2 Bioinformatics Analysis Pipeline

2.1 Paired-end Reads Assembly and Quality Control

2.1.1 Data Split

Paired- end reads was assigned to samples based on their unique barcodes and truncated by cutting off the barcodes and primer sequences. The whole process was performed through Python (V3.6.13) and adaptors were removed through cutadapt (V3.3).

2.1.2 Sequence Assembly

Paired-end reads were merged using FLASH (V1.2.11, http://ccb.jhu.edu/software/FLASH/) (Magoc T et al.,2011), 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.

2.1.3 Data Filtration

Quality filtering on the raw tags were performed using the fastp (Version 0.23.1) software to obtain high-quality Clean Tags (Bokulich NA et al.,2012).

2.1.4 Chimera Removal

The tags were compared with the reference database (Silva database (16S/18S), https://www.arb-silva.de/; Unite Database (ITS), https://unite.ut.ee/) to detect chimera sequences. And the effective tags were obtained by removing the chimera sequences with the vsearch package (V2.16.0, https://github.com/torognes/vsearch) (Edgar RC et al.,2011).

2.2 OTU Cluster and Species Annotation

2.2.1 OTU Production

Sequences analysis were performed by Uparse software (Uparse v7.0. 1001 ， http://drive5.com/uparse/) (Edgar RC et al.,2013). Sequences with ≥ 97% similarity were assigned to the same OTUs. Representative sequence for each OTU was screened for further annotation.

2.2.2 Species Annotation

16S: For each representative sequence, the Silva Database (http://www.arb- silva.de/) (Quast C et al.,2012) was used based on Mothur algorithm to annotate taxonomic information.

18S: For each representative sequence, the Silva Database (http://www.arb- silva.de/) (Quast C et al.,2012) was used based on RDP classifier algorithm to annotate taxonomic information.

ITS: For each representative sequence, the Unite Database (https://unite.ut.ee/) (Herr JR et al.,2014) was used based on blast algorithm to annotate taxonomic information.

Unregular region: Micro_NT by default (a sub library obtained by extracting archaea, fungi, viruses, and bacteria from the NT) .

Note:

1.1. The fasta file (SILVA_138.1_SSURef_NR99_tax_silva.fasta) from the Silva offical website only provide species name rather than taxnonomic information, and thier taxonomic information file is incomplete. Therefore, we need use species name to get their taxonomic information from NCBI. At first, taxonomic information file from Silva offical is used to annotate, for those sequences that can not be annotated, a dmp file provided by NCBI would be use to supplement the taxonomic information.

2. Annotation principle of Micro_NT: perform sequence alignment (BLAST) between our sequence and sequence in Micro_NT. Then obtain the top20 result according to the score, screen by the maximum bit-score, use LCA algorithm to infer the Lowest Common Ancestor. Due to the large amount of unclassified information in the NT, in order to ensure the accuracy of the data, the annotation effect is greatly reduced, resulting in more unclassified results. So when we use LCA algorithm to obtain species information, we will ignore unclassified species to reduce their impact on annotation.

2.2.3 Phylogenetic Relationship Construction

In order to study phylogenetic relationship of different OTUs, and the difference of the dominant species in different samples (groups), multiple sequence alignment were conducted using the MUSCLE software (Version 3.8.31， http://www.drive5.com/muscle/) (Edgar RC et al.,2004).

2.2.4 Data Normalization

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

2.2.5 Relative abundance

Top 10 taxa of each samples at each taxonomic ranks(Phylum, Class, Order, Family, Genus, Species) were selected to plot the distribution histogram of relative abundance in Perl through SVG function.

2.2.6 Heatmap

The abundance information of top 35 taxa of each samples at each taxonomic ranks were used to drawn the heatmap, which visually display different abundance and taxa clustering. This was achieved in R through the pheatmap() function.

2.2.7 Ternaryplot

Ternary plot for top 10 taxa at each taxonomic ranks can be used to show the abundance difference among three samples. It was performed in R with vcd() function.

2.2.8 Venn and Flower diagram

Venn and Flower diagrams visually display the common and unique information between different samples or groups. Venn and Flower diagrams were produced in R with VennDiagram() function and in perl with SVG function, respectively.

2.2.9 Phylogenetic tree

Phylogenetic tree, also called evolutionary tree, can describe the evolutionary relationship between different species. One hundred genera with the highest abundance in the samples were selected and performed sequence alignment to draw the phylogenetic tree in perl with SVG function.

2.3 Alpha Diversity

2.3.1 Alpha Diversity

Alpha diversity is applied in analyzing complexity of species diversity for a sample through 6 indices, including Observed-species, Chao1, Shannon, Simpson, ACE, Good-coverage. All this indices in our samples were calculated with QIIME (Version 1.9.1) and displayed with R software (Version 4.0.3).

Two indices were selected to identify Community richness:

Chao -the Chao1 estimator (http://www.mothur.org/wiki/Chao); CE -the ACE estimator (http://www.mothur.org/wiki/Ace);

Two indices were used to identify Community diversity:

Shannon - the Shannon index (http://www.mothur.org/wiki/Shannon); Simpson - the Simpson index (http://www.mothur.org/wiki/Simpson); One indice to characterized Sequencing depth:

Coverage - the Good's coverage (http://www.mothur.org/wiki/Coverage)

2.3.2 Species accumulation boxplot

In order to evaluate the richness of microbial community and sample size. Species accumulation boxplot can be used to visualize, which performed with vegan package in R software.

2.3.3 Rank abundance curve

Rank abundance curve can reflect the richness and evenness of species with samples through observing the width and shape of the curves. It can be plotted by using the RColorBrewer package in R.

2.3.4 Rarefactioncurve

Insufficient sequencing number may lead to insufficient sample information, whereas excessive sequencing rounds can also result in unnecessary increase in cost. Therefore, the determination for the proper amount of sequencing is essential. Plotting rarefaction curves provides the ability to discover the viability of the sequencing numbers. This is achieved through R with the plyr library.

2.4 Beta Diversity

2.4.1 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 1.9.1).

Cluster analysis was preceded by principal component analysis (PCA), which was applied to reduce the dimension of the original variables using the ade4 package and ggplot2 package in R software(Version 4.0.3).

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 ade4 package and ggplot2 package in R software(Version 4.0.3).

2.4.2 Beta diversity heatmap

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. Then a heatmap was created to display the unifrac distance between samples, which was realized in Perl.

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

A cluster tree is constructed on UPGMA, which is based on the weighted unifrac distance matrix. This is widely used in ecology for evolutionary classification. UPGMA diagram was drawn through the upgma.tre function within Qiime.

2.4.4 Cluster analysis

Cluster analysis was preceded by principal component analysis (PCA), which was applied to reduce the dimension of the original variables using the ade4 package and ggplot2 package with R software(Version 4.0.3).

Principal Coordinate Analysis (PCoA) was performed to get principal coordinates and visualize from complex and multidimensional data. A distance matrix of weighted or unweighted unifrac among samples was obtained before transformation 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 ade4 package and ggplot2 package in R software(Version 4.0.3).

Non-metric multidimensional scaling (NMDS) was also implemented for data dimension reduction. Similar to the PCoA, NMDS also use the distance matrix but it emphasizes the numerical rank instead. The distance between sample points on the diagram can only reflect the rank information rather than the numerical differences. NMDS analysis was implemented through R software with ade4 package and ggplot2 package.

2.5 Communitydifference analysis

A series of statistical analyses which inlcude Anosim, Adonis, Multi-response permutation procedure (MRPP), Simper, T-test, MetagenomeSeq and LEfSe, were performed to reveal the community structure differentiation.

Anosim, Adonis and MRPP analysis are non-parametric tests that analyze the difference between high-dimensional data group. They can test whether the differences between groups are significant greater than the differences within the group, which can determine whether the grouping is meaningful. All of them were performed with vegan package and ggplot2 package within R.

Simper can reveal the contribution of each species to the differentiation between groups. Top 10 species were selected and presented on the graph. It was performed in R with Vegan package and ggplot2 package.

MetagenomeSeq can showcase the species that display significant differences between groups. It was performed in R with metagenomeSeq package.

LEfSe is widely used to discover biomarkers and it can reveal metagenomic characteristics. To achieve this, a exclusive package named lefse was utilized.

2.6 Functionprediction

PICRUSt (V1.1.4) is a package in R and is mainly used to predict the metagenomic functions based on marker genes. PICRUSt2 (V2.3.0) is the improved version of PICRUSt.

Tax4Fun (V0.3.1) is a R package that is widely used for intestinal and soil samples. In general, it can supply more accurate results when compared with PICRUSt, especially for soil samples.

BugBase is an excellent tool to discover the phenyotpe of microogranisms. It can classify the microbial communities according to seven phenotypes: Gram Positive, Gram Negative, Biofilm Forming, Pathogenic, Mobile Element Containing, Oxygen Utilization, including Aerobic, Anaerobic, and Cultivable Anaerobic, and Oxidative Stress Tolerance.

FunGuild is an excellent tool through python when working with fungi samples.

FAPROTAX through python can play a great role when elucidating the possible biochemical processes and elements in play.

2.7 Associationanalysis

To explore the symbiotic relationship between species and to reveal the environmental factor influence on community structures, 2D and 3D network diagrams were drew for visualization. Further analyses such as spearman correlation test, canonical correspondence analysis (CCA)/redundancy analysis (RDA) and dbRDA can be used to reflect the correlation between environmental factors and species abundance. All of these diagram and analysis were completed in R.

3 Reference

Bokulich NA, Subramanian S, Faith JJ, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.Nature Methods. 2012;10(1):57-59. doi:10.1038/nmeth.2276.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection.Bioinformatics. 2011;27(16):2194-2200. doi:10.1093/bioinformatics/btr381.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.Nucleic Acids Research. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340.

Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads.Nature Methods. 2013;10(10):996-998. doi:10.1038/nmeth.2604.

Herr JR, Öpik M, Hibbett DS. Towards the unification of sequence-based classification and sequence-based identification of host-associated microorganisms.New Phytologist. 2014;205(1):27-31. doi:10.1111/nph.13180.

Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies.Bioinformatics. 2011;27(21):2957-2963. doi:10.1093/bioinformatics/btr507.

Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.Nucleic Acids Research. 2012;41(D1):D590-D596. doi:10.1093/nar/gks1219.