Methods

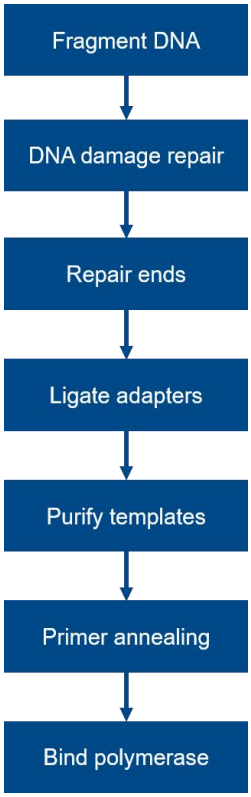
Sequencing

1. Sample Quality Control

Methods of sample quality control refer to QC report.

2. Library Construction, Quality Control and Sequencing

Amplified DNA fragments were end-repaired, and A-tailed. The sequencing adapters were ligated to the ends of the DNA fragments using DNA-binding enzyme, and the DNA fragments were purified using AMpure PB magnetic beads to construct a SMRTbell library. Finally, sequencing primer was annealed to the SMRTbell templates, followed by binding of the sequence polymerase to the annealed templates. The experimental procedures of DNA library preparation are shown as follows:



Workflow of library construction

The library was checked with Qubit for quantification. Quantified libraries will be pooled and sequenced on PacBio Sequel II/IIe systems, according to effective library concentration and data amount required.

Data analysis

1. Sequencing data processing

Export PacBio offline data to a bam format file. Use lima software to distinguish the data of each sample based on barcode sequences, and save all sample sequences in bam format. Then use CCS (SMRT Link v7.0) to correct the sequence, with a correction parameter of CCS=3 and a minimum accuracy of 0.99 [21]. Remove sequences with lengths less than 1340 and longest sequences with lengths greater than 1640, and store them in fastq and fastta; Subsequently, SSR filtration was performed and the primers were removed using cutadapt to filter out sequences containing consecutive identical base numbers>8. The Reads obtained after the above processing are the final valid data (Clean Reads).

2. ASVs Denoise and Species annotation

2.1 ASVsDenoise

For the Effective Tags obtained previously, denoise was performed with DADA2 or deblur module in the QIIME2 software (Version QIIME2-202006) to obtain initial ASVs (Amplicon Sequence Variants) (default: DADA2), and then ASVs with abundance less than 5 were filtered out[3].

2.2 Species Annotation

Species annotation was performed using QIIME2 software. For 16S/18S, the annotation database is Silva Database, while for ITS, it is Unite Database.

2.3 Phylogenetic Relationship Construction

In order to study phylogenetic relationship of each ASV and the differences of the dominant species among different samples(groups), multiple sequence alignment was performed using QIIME2 software.

2.4 Data Normalization

The absolute abundance of ASVs was 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 based on the output normalized data.

3.AlphaDiversity

In order to analyze the diversity, richness and uniformity of the communities in the sample,alpha diversity was calculated from 7 indices in QIIME2, including Observed_otus, Chao1, Shannon, Simpson, Dominance, Good's coverage and Pielou_e.

Three indices were selected to identify community richness:

Observed_otus - the number of observed species (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.observed_otus.html);

Chao- the Chao1 estimator (http://scikitbio.org/docs/latest/generated/skbio.diversity.alpha.chao1.html#skbio.diversity.alpha.chao1);

Dominance - the Dominance index (http://scikitbio.org/docs/latest/generated/skbio.diversity.alpha.dominance.html#skbio.diversity.alpha.dominance);

Two indices were used to identify community diversity:

Shannon - the Shannon index (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.shannon.html#skbio.diversity.alpha.shannon);

Simpson - the Simpson index (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.simpson.html#skbio.diversity.alpha.simpson);

One indice was used to calculate sequencing depth:

Coverage - the Good's coverage (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.goods_coverage.html#skbio.diversity.alpha.goods_coverage);

One indice was used to calculate species evenness:

Pielou_e - Pielou's evenness index (http://scikitbio.org/docs/latest/generated/skbio.diversity.alpha.pielou_e.html#skbio.diversity.alpha.pielou_e).

4. Beta Diversity

In order to evaluate the complexity of the community composition and compare the differences between samples(groups), beta diversity was calculated based on weighted and unweighted unifrac distances in QIIME2.

Cluster analysis was performed with 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 3.5.3).

Principal Coordinate Analysis (PCoA) was performed to obtain principal coordinates and visualize differences of samples in complex multi-dimensional data. A matrix of weighted or unweighted unifrac distances among samples obtained previously was transformed into a new set of orthogonal axes, where the maximum variation factor was demonstrated by the first principal coordinate, and the second maximum variation factor was demonstrated by the second principal coordinate, and so on. The three-dimensional PCoA results were displayed using QIIME2 package, while the two-dimensional PCoA results were displayed using ade4 package and ggplot2package in R software (Version 2.15.3).

To study the significance of the differences in community structure between groups, the adonis and anosim functions in the QIIME2 software were used to do analysis. To find out the significantly different species at each taxonomic level (Phylum, Class, Order, Family, Genus, Species), the R software (Version 3.5.3) was used to doMetaStat and T-test analysis. The LEfSe software (Version 1.0) was used to do LEfSe analysis (LDA score threshold: 4) so as to find out the biomarkers.

Further, to study the functions of the communities in the samples and find out the different functions of the communities in the different groups, the PICRUSt2 software (Version 2.1.2-b)was used for function annotation analysis.

Reference

[1]Magoč T,Salzberg S L. FLASH: fast length adjustment of short reads to improvegenome assemblies. Bioinformatics 27.21 (2011): 2957-2963.

[2]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.

[3]LiMinjuan,Shao Dantong,Zhou Jiachen et al. Signatures within esophageal microbiota with progression of esophageal squamous cell carcinoma.[J] .Chin J Cancer Res, 2020, 32: 755-767.