supplementary information

method of microbiome

1. Quality control

1.1 Data split

Raw sequences were initially processed through the PacBio SMRT portal. Sequences were filtered for a minimum of 3 passes, and a minimum predicted accuracy of 90% (min full pass = 3, min Predicted Accuacy = 0.9). The predicted accuracy of 90%, which is defined as the threshold below which a CCS is considered as noise. The files generated by the PacBio platform were then used for amplicon size trimming to remove sequences outside the expected amplicon size (min Length 1340 bp, max Length 1640 bp). The reads was assigned to samples based on their unique barcode and truncated by cutting off the barcode and primer sequence.

2. OTU cluster and Species annotation

2.1 OTU Production

Sequences analysis were performed by Uparse software (Uparse v7.0.1001, http://drive5.com/uparse/). Sequences with ≥97% similarity were assigned to the same OTUs. Representative sequence for each OTU was screened for further annotation.

2.2 Species annotation

For each representative sequence, the SSUrRNA Database of Silva Database (https://www.arbsilva.de/) was used based on Mothur algorithmto annotate taxonomic information.

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/).

2.4 Data Normalization

OTUs abundance information were normalized using a standard of sequence number corresponding to the sample with the most 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 through indices, including Observed-species, Chao1, Shannon, Simpson, ACE, Good-coverage. All this indices in our samples were calculated with QIIME (Version1.9.1) and displayed with R software (Version 2.15.3).

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 1.9.1). 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).

5. Random Forest

The Random Forest (RF) classification model was performed using the R package. The genus level count data derived from the OTU table were used for downstream analyses. The RF parameters used were as follows: the number of trees in the forest (ntree) was set to 501 and the number of features randomly sampled at each node in a tree (mtry) was 11.