1. Pre-processing of sequencing results

Fastp software was used for Raw Data quality control, default software parameters were selected to preprocess Raw Data obtained by Illumina HiSeq sequencing platform, and Clean Data was obtained for subsequent analysis. If there is contamination in the sample, it should be compared with the host database to filter out reads that may be from the host. Select Bowtie2 and set the parameters to --sensitive, -I 200, -x 400.

2. Metagenome assembly

1. After pretreatment, Clean Data was obtained, and Assembly Analysis was performed using MEGAHIT Assembly software. Assembly parameters: -- K-min 35 -- K-max 95 -- K-Step 20 -- min-contig-Len 500;
2. CleanData of each sample after quality control was compared to contigs of each sample after assembly using Bowtie2 software to obtain PE reads that were not utilized. Comparison parameters: -I 200, -x 400;
3. Unutilized reads of each sample were put together for mixed assembly, and the assembly parameters were the same as those of single sample.

3. Gene prediction and abundance analysis

1. Based on contigs (>=500bp) of each sample and mixed assembly, MetaGeneMark was used to predict ORF (Open Reading Frame), and the default parameters were used.
2. Based on the predicted results, the predicted genes with length less than 100nt were filtered out.
3. Combine the ORF prediction results of all samples and mixed assembly, and use CD-HIT software to remove the redundancy, so as to obtain the non-redundant initial gene catalogue (here, operationally, the nucleic acid sequence encoded by non-redundant continuous genes is called genes). By default, identity 95% and Coverage 90% are used for clustering, and the longest sequence is selected as the representative sequence. Parameters: -c 0.95, -g 0, -AS 0.9, -g 1, -d 0;
4. Bowtie2 was used to compare Clean Data of each sample to the initial gene catalogue, and the number of reads that gene was compared in each sample was calculated. --end-to-end, --sensitive, -I 200, -x 400

5. Genes supporting reads <=2 in each sample were filtered out to obtain gene catalogue (Unigenes) for subsequent analysis.
6. Based on the number of reads and gene length compared, the abundance information of each gene in each sample was calculated, as shown in the following formula. R is the number of reads compared to the gene, and L is the length of the gene:

$$G_k = \frac{r_k}{L_k} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$$

4. Species Annotation

1. Comparison between genes and species database: DIAMOND software compared Unigenes sequences with those of Bacteria, Fungi, Archaea and Viruses extracted from NCBI NR database (BLastp, EVALue <= 1E-5).
2. LCA algorithm: Because each sequence may have multiple comparison results, multiple different species classification information can be obtained. In order to ensure its biological significance, LCA algorithm of MEGAN software is used to obtain the final species annotation information of the sequence.
3. Based on the results of LCA annotation and gene abundance table, the abundance information of each sample at each classification level (phylum and genus) was obtained. For the abundance of a species in a sample, the annotation was equal to the sum of the gene abundance of the species.
4. Based on the LCA annotation results and gene abundance table, the gene number table of each sample in each classification level (phylum compendium genus) was obtained. For the gene number of a species in a sample, it was equal to the gene number of the annotated species, gene abundance is not 0;
5. Krona analysis, relative abundance overview, abundance clustering thermal map, PCA and NMDS dimensional reduction analysis, Anosim inter-group (intra-group) difference analysis, and LEfSe multivariate statistical analysis of inter-group differential species were performed based on the abundance tables at each classification level (Genus and species of Phylum).

5. Common functions Database annotation

1. Sequence alignment: Unigenes was compared with each functional database using DIAMOND software (BLastP, EVALue <= 1E-5).
2. Comparison result filtering: For the comparison result of each sequence, the DIAMOND parameter --max-target-seqs 1 retains the unique comparison result;
3. Based on the comparison results, the relative abundance of different functional levels was calculated (the relative abundance of each functional level was equal to the sum of the relative abundance of genes annotated as this functional level). KEGG database was divided into 6 levels, eggNOG database into 3 levels, and CAZy database into 3 levels.
4. Based on the result of functional annotation and gene abundance table, the gene number table of each sample at each classification level was obtained. For the number of genes with a certain function in a sample, it was equal to the number of genes with a non-0 abundance among the genes annotated with this function.
5. Based on the abundance tables at each classification level, the number statistics of annotated genes, relative abundance overview, abundance clustering heat map, PCA and NMDS dimensionality reduction analysis, Anosim inter-group (within) difference analysis based on functional abundance, metabolic pathway comparative analysis were carried out. Metastat and LEfSe analysis of functional differences between groups.

6. Resistance gene annotation

1. Use the CARD database to hold Gene Identifier (RGI) software to compare the Unigenes and CARD database (https://card.mcmaster.ca/) (RGI built-in blastp, Default evalue < 1e-30);
2. The relative abundance of each ARO was calculated based on RGI and Unigenes abundance information.
3. Starting from the abundance of ARO, the abundance histogram, abundance clustering heat map, abundance distribution circle map and ARO difference between groups were displayed. Resistance genes (annotated to unigenes of ARO) and species attribution analysis of resistance mechanism,

etc. (For some aros with long names, the first three words and the form of underline abbreviations are displayed).