**Bioinformatics analysis**

**1. Sequencing data processing**

The original picture data was processed using Bcl2fastq software (v2.17.1.14) to identify the base sequence and perform preliminary quality analysis. In the sequencing process, Illumina's built-in software was used to assess the quality of each sequencing segment. The initial 25 bases were examined to determine if the read was kept or deleted. The findings were stored in the FASTQ file format, which included the sequencing sequence information (found in the second line of the FASTQ format) together with its accompanying sequencing quality information. During the sequencing process, Illumina's integrated software will acquire raw sequencing data (Pass Filter Data) by evaluating the quality of each sequencing fragment (read). The retention or discarding of a read is contingent upon the quality of the initial 25 bases. The outcomes will be saved in the FASTQ file format, which comprises the sequencing sequence information (the second line of the FASTQ format) and its matching sequencing quality information (the fourth line of the FASTQ format).

The sequencing bases' quality is influenced by various factors like as the sequencer, sequencing reagents, and samples. Typically, the first bases at the 5' end of the sequencing sequence have a high mistake rate, whereas the error rate of the bases at the 3' end increases as the sequencing sequence gets longer. This pattern is a result of the properties of high-throughput sequencing technology. The initial 6 bases exhibit a substantial mistake rate, and this length is also equivalent to the length of the random primer necessary for reverse transcription in the process of constructing an RNA-seq library. Hence, the substantial error rate observed in the first 6 bases of the sequencing is attributed to the incomplete binding of the random primers to the RNA template. The sequencing error distribution check is employed to identify any anomalous base positions with elevated error rates within the range of sequencing length. For instance, it detects if the sequencing error rate of bases in the middle position is notably greater compared to other positions. Typically, the error rate for sequencing each base location should be below 0.5%. The link between the sequencing error rate (e) and the quality value of bases in illumina HiSeq/MiSeq (Qphred) can be expressed as follows: .

**2. Optimisation of sequencing data quality**

High-throughput sequencing often exhibits common sequencing errors, such as point mutations. As a result, the latter part of the sequence is frequently of worse quality. To enhance the precision and quality of bioinformatics analyses, it is imperative to optimize the sequencing raw data.

**Analysis software:** Cutadapt (v1.9.1), Vsearch (1.9.6), and Qiime (1.9.1).

**Analysis methods and steps:**

(1) The two sequences should be compared and joined together based on the overlapping region at the end of the comparison, with a minimum overlap of 20 base pairs. The splicing result should exclude sequences that contain the nucleotide N.

(2) The primers and splice sequences should be eliminated, along with the bases that have quality values below 20 at both ends and the sequences that have lengths shorter than 200 bp.

(3) The spliced and filtered sequences mentioned above should be compared to the database. The chimera sequence should be excluded in order to acquire the final legitimate data.

**3. OTU analysis and species annotation**

3.1 OTU clustering

In the field of population genetics research, an operational taxonomic unit (OTU) is a standardized marker that is intentionally assigned to a specific taxonomic unit, such as a genus, species, or group. This marker is used to simplify the analysis process. Within the realm of bioinformatics analysis, every sequence obtained through the process of sequencing is derived from a distinct strain. To understand the variety of strains, genera, and other taxonomic units in the sequencing findings of a sample, it is necessary to cluster the sequences. Sequences can be grouped together based on their similarity to one another, enabling their classification. An operational taxonomic unit (OTU) is the term used to define a group. Generally, sequences are commonly categorized into Operational Taxonomic Units (OTUs) based on a similarity threshold of 97%. Subsequently, these OTUs are subjected to analysis using bioinformatic statistical methods.

**Analysis software:** Qiime (1.9.1), Vsearch (1.9.6).

**Analysis methods and steps:**

(1) Retrieve distinct sequences from the optimized valid sequences while preserving the count of occurrences for each sequence.

(2) Eliminate the distinct sequences that have a repetition count of 1.

(3) Conduct OTU clustering on the UNIQUE sequences (sequences that occur more than once) based on a 97% similarity threshold. Additionally, exclude any chimeric sequences during the clustering phase to obtain the representative sequences of the OTU.

(4) The optimized sequences were compared to the representative sequences of the OTUs. Sequences that had a similarity of 97% or higher to the typical sequences of the OTU were classified as part of the same OTU. A statistical analysis was performed to obtain the OTU abundance table.

3.2 Statistics of species annotation results

To acquire the species classification information associated with the OTUs, a representative sequence is chosen for each OTU. The representative sequence was annotated with species classification using the RDP classifier, allowing the determination of the community composition of each sample.

**Analysis software:** Qiime (1.9.1).

**Analysis methods and steps:** A taxonomic analysis was performed on the representative sequences of Operational Taxonomic Units (OTUs) at a similarity level of 97% using the RDP classifier Bayesian algorithm. The aim was to identify the composition of the community in each sample at every taxonomic level.

**Comparison database:** 16S – Silva\_132 16SrRNA database (<http://www.arb-silva.de/>)

**4. Alpha Diversity**

4.1 Analysis of the alpha diversity index

The Chao index is a statistical method used to estimate the number of operational taxonomic units (OTUs) present in a sample. It is calculated using the Chao1 algorithm, which is commonly employed in ecological studies to assess the total number of species present in a given community.

(<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.chao1.html>)

The indices employed to calculate community diversity are as follows:

(1) Shannon: This index is commonly used to reflect the alpha diversity index, which is used to estimate the microbial diversity in a sample.

(<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.shannon.html>)

(2) Simpson: The Simpson diversity index, proposed by Edward Hugh Simpson, is commonly used in ecology to quantitatively describe the biodiversity of a region.

(<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.simpson.html>)

Sequencing depth indices are defined as follows:

(1) Good's Coverage: This index represents the coverage of each sample library. A higher value indicates a lower probability that sequences within the sample have not been sequenced.

(<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.goods_coverage.html>)

**Analysis software:** Qiime (1.9.1).

**Analysis method:** The method of random sampling of sequences was employed to analyse the OTUs, with the number of valid sequences drawn. Each alpha diversity index was calculated separately.

4.2 Analysis of differences in α-diversity index between groups

The comparison of α-diversity index between different groups is conducted using the α-diversity index table, which was generated using the R programming language to create a box-and-whisker plot. This graphic depiction displays the upper and lower limits, median, and anomalous values of the α-diversity index for each sample group, providing a direct reflection of the extent of diversity disparities among the groups.

**Analysis method:** Construct box-and-line diagrams based on the results of the α-diversity index, employing the R language.

4.3 Rarefaction curve

The rarefaction curve is a valuable tool for examining the species makeup of a sample and forecasting the abundance of species within it. It is a reliable technique for assessing the sufficiency of a sample size and approximating the diversity of species. The curve in question is employed to depict the relationship between the increase in sample size and the corresponding increase in the number of detectable species. It has been extensively employed in biodiversity and community surveys to assess the sufficiency of the sample size and estimate the number of species present. Therefore, the dilution curve may determine both the adequacy of the sample size and estimate the species richness, assuming that the sample size is sufficient.

**Analysis software:** Qiime (1.9.1).

**Analysis method:** A random sampling of sequences was employed to construct a dilution curve, with the number of sequences sampled plotted against the number of operational taxonomic units (OTUs) they could represent.

**5. Beta diversity**

5.1 Unifrac Distance Matrix

The beta diversity value is the coefficient of variation between two samples, reflecting the difference in diversity between different samples. The distance between samples was calculated using the evolutionary and abundance information between the sequences of each sample. This reflects whether there are significant differences in microbial communities between the samples, which can be realised by UniFrac analysis.

**Analysis software:** Qiime (1.9.1).

**Analysis method:** Firstly, an evolutionary tree is constructed using representative sequences of OTUs from different environmental samples. The Unifrac metric then measures the difference between two different environmental samples based on the length of the constructed evolutionary tree branches.

5.2 PCoA Analysis

Principal Coordinates Analysis (PCoA) is a visualisation method used to study the similarity or difference of data. It is similar to Principal Component Analysis (PCA). After sorting through a series of eigenvalues and eigenvectors, the eigenvalues that are mainly in the top rank are selected to find the most dominant coordinates in the distance matrix. This results in a data matrix of A rotation, which does not change the mutual positional relationship between the sample points. However, it does change the coordinate system. The distinction between the two methods lies in their respective approaches to identifying principal components. PCA is based on the similarity coefficient matrix of the samples, whereas PCoA is based on the distance matrix.

**Analysis software:** R language.

**Analysis method:** Based on Brary-Curtis distance matrix for PCoA graphing analysis.

5.3 PCA Analysis

PCA (Principal Component Analysis) is a technique employed to simplify and analyse data, with the objective of identifying the most "major" elements and structures within the data set.

The distribution of strain communities in different samples can be analysed to reflect the similarity and difference between samples. This is achieved through variance decomposition, which reflects the differences between multiple data sets in a two-dimensional coordinate graph. The axes of this graph are able to maximally reflect the variance values of the two eigenvalues.

**Analysis software:** R language.

**Analysis method:** PCA graphing analysis based on strain community distribution.

5.4 NMDS analysis

The non-metric multidimensional scaling (NMDS) method is a data analysis method that simplifies the research objects in multidimensional space to low-dimensional space for positioning, analysis and categorisation, while retaining the original relationship between the objects. The method is characterised by the fact that the species information contained in the samples is reflected in the multidimensional space in the form of points. The degree of difference between different samples is reflected by the distance between points, and the spatial positioning point map of the samples is finally obtained.

**Analysis software:** Based on beta diversity distance matrix, graphing with R language vegan package.