

This technical document is only made available to the clients who have purchased the corresponding service with BMKGENE. Please do NOT share with any third parties. Thank you for your understanding.

Bioinformatics Analysis Workflow for NGS-based Metagenomics

Contents

1 Background Introduction	3
2 Bioinformatics Analysis Workflow	3
3 Standard Analysis Results	3
3.1 Sequencing Data Quality Control	3
3.2 Metagenome Assembly	4
3.3 Metagenome Composition Analysis	4
3.3.1 Gene Prediction	4
3.4 Construction of Non-redundant Gene Set	4
3.5 Functional Annotation Analysis	4
3.5.1 Nr Annotation	4
3.5.2 GO Annotation	5
3.5.3 KEGG Annotation	5
3.5.4 eggNOG Annotation	6
3.5.5 Pfam Annotation	6
3.5.6 SwissProt Annotation	7
3.5.7 CAZy Annotation	7
3.5.8 CARD Annotation	8
3.5.9 PHI-base Annotation	8
3.5.10 CYPED Annotation	9
3.5.11 QS Annotation	9
3.5.12 BacMet Annotation	10
4 Functional Diversity Analysis	10
4.1 Functional Gene Composition and Abundance Analysis	10
4.1.1 Scatter Plot of Functional Composition	10
4.1.2 CAZy Functional Annotation	10
4.1.3 CARD Functional Annotation	11
4.2 Beta Diversity Analysis of Functional Genes	11
4.2.1 Principal Component Analysis	11
4.2.2 Principal Coordinates Analysis	11
4.2.3 NMDS Analysis	12
4.2.4 UPGMA Analysis	12
4.2.5 Analysis of Similaritie	12
4.2.6 PerMANOVA/Adonis	13
4.3 Inter-group Differences Analysis	13

4.3.1 Parametric Tests	13
4.3.2 Non-parametric Tests	13
4.3.3 MetagenomeSeq Analysis	14
4.3.4 Lefse	14
4.4 Functional Gene Correlation Analysis	14
4.4.1 Correlation Analysis	14
4.4.2 RDA/CCA	15
4.4.3 Functional Genes-Environmental Factors Correlation Heatmap	15
5 Taxonomic Diversity Analysis	16
5.1 Taxonomic Composition and Abundance Analysis	16
5.1.1 Taxonomic Annotation and Statistics	16
5.1.2 Taxonomic Composition Bar Plot	16
5.2 Alpha Diversity Analysis	16
5.3 Beta Diversity Analysis	17
5.3.1 Principal Component Analysis	17
5.3.2 Principal Coordinates Analysis	17
5.3.3 NMDS Analysis	17
5.3.4 UPGMA Analysis	18
5.3.5 Analysis of Similarities	18
5.3.6 PerMANOVA/Adonis	19
5.4 Inter-group Differences Analysis	19
5.4.1 Parametric Tests	19
5.4.2 Non-Parametric Tests	19
1) Paired samples - Signed-rank test	20
5.4.3 MetagenomeSeq Analysis	20
5.4.4 LEfSe	20
5.4.5 Random Forest Analysis	21
5.5 Correlation Analysis	21
5.5.1 Taxonomic Correlation Analysis	21
5.5.2 RDA/CCA	21
5.5.3 Taxonomy-Environment Factors Correlation Heatmap	22
6 Reference	22

1 Background Introduction

Second-generation metagenomic sequencing is based on Illumina high-throughput sequencing, using the paired-end sequencing method to construct short fragment libraries for sequencing. It aims to obtain gene functional information and species composition information from environmental samples through bioinformatics analysis. The functional diversity of environmental samples is analyzed from a gene functional perspective, and the species diversity contained in the environmental samples is analyzed from a species perspective.

2 Bioinformatics Analysis Workflow

The raw reads obtained from sequencing are subjected to quality control and filtering to obtain clean reads, which are used for subsequent bioinformatics analysis. The clean reads are assembled and predicted for coding genes, and the functional annotation is performed using both general and specialized databases. Taxonomic analysis is also conducted on the clean reads to determine the species composition and abundance information of the samples. The analysis and final report generation are performed using the BMKCloud platform.

3 Standard Analysis Results

3.1 Sequencing Data Quality Control

The raw sequences obtained from sequencing (raw reads) may contain low-quality sequences. To ensure the quality of information analysis, it is necessary to filter the raw reads and obtain clean reads for subsequent analysis. The main steps of data filtering are as follows:

- 1) The software fastp (version v0.23.1, parameters -5 -W 50 -M 20 -l 60 -n 0 -g -A) is used to filter the raw tags and obtain high-quality sequencing data (clean tags).
- 2) The software bowtie2 (version v2.2.4, parameters --seed 123456 -I 200 -X 1000 --un-conc-gz) is used to align the clean tags to the host genome sequence to remove host contamination (if a host reference genome is provided).

3.2 Metagenome Assembly

The software MEGAHIT [1] (version v1.1.2, default parameters) is used for metagenome assembly, with contig sequences shorter than 300 bp filtered out. The assembly results are evaluated using the software QUAST [2] (version v2.3, default parameters).

3.3 Metagenome Composition Analysis

3.3.1 Gene Prediction

The software MetaGeneMark [3] (version v3.26, default parameters) is used to predict coding regions in the genome. The default parameters (-A -D -fG) are used. The software can be accessed at http://exon.gatech.edu/meta_gmhmmmp.cgi, Version 3.26.

3.4 Construction of Non-redundant Gene Set

The MMseqs2 software (<https://github.com/soedinglab/mmseqs2>, Version 11-e1a1c) is used to remove redundancy, with a similarity threshold set at 95% and a coverage threshold set at 90%.

3.5 Functional Annotation Analysis

3.5.1 Nr Annotation

The Nr database [5], which stands for Non-Redundant Protein Database, is a comprehensive protein sequence and annotation database created and maintained by NCBI. It contains non-redundant protein sequences and annotation information, including corresponding species information. Compared to other databases, using the Nr database generally results in a higher annotation rate for genes in the genome. However, many protein sequences and annotation information in this database have not been validated and their reliability needs to be improved.

For specific annotation, the protein sequences of non-redundant genes are compared to the Nr database using BLAST alignment (diamond v0.9.29.130, alignment filtering threshold: E-value 1e-5). The most similar sequence in the Nr database corresponds to the annotation information of the corresponding gene in the sequenced genome.

3.5.2 GO Annotation

The GO database [6] (Gene Ontology) is a database established by the Gene Ontology Consortium, aiming to create a semantic vocabulary standard for limiting and describing gene and protein functions that is applicable to various species and can be updated as research progresses.

Currently, genes with the same function may be described using different terms in different functional databases, similar to different dialects. This leads to discrepancies in annotation when multiple databases are used, which hinders the long-term development and use of functional annotation. The Gene Ontology project was established to address this issue and ensure consistent functional descriptions of gene products across various databases.

The GO database represents biological functions in a hierarchical tree-like structure. The highest-level functional nodes include: (1) Cellular Component, which describes subcellular structures, locations, and macromolecular complexes; (2) Molecular Function, which describes the functions of genes and gene products; and (3) Biological Process, which describes the biological processes in which gene-encoded products are involved.

In specific annotation, based on the previous annotation results from the Pfam database, genes that align to Pfam are mapped to nodes in the GO database that contain functional annotation information. The annotation information corresponding to these nodes represents the annotation of genes in the sequenced genome.

3.5.3 KEGG Annotation

KEGG [8] (Kyoto Encyclopedia of Genes and Genomes) is a comprehensive database that collects information on genomes, pathways, and compounds in organisms. The database analyzes and clusters the included sequences to form orthologous protein groups, and assigns different KO (KEGG Orthology) sequences to different protein groups. Based on published literature, KEGG has manually drawn numerous biological pathway maps (such as metabolic pathways and signaling pathways), where specific rectangular boxes or lines correspond to protein groups with specific KO numbers that perform corresponding functions. These pathway maps provide a more intuitive representation of biological processes.

KEGG is a database that collects information on genomes, biological pathways, diseases, drugs, and chemical substances. Within KEGG, there is a "unique term" called KO (KEGG Orthology), which is a classification system for proteins (enzymes). It groups protein sequences that are highly similar and have similar functions in the same pathway, and assigns them a KO label. Each KO label is associated with corresponding functional annotation information. These KO labels correspond to many metabolic pathways or cellular processes in the genome within KEGG.

In specific annotation, the protein sequences of non-redundant genes are compared to the protein sequences included in the KEGG database using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value $1e-5$). The most similar sequence in the KEGG database corresponds to the annotation information of the sequence, including the associated KO number and its position in the KEGG pathway. This information represents the annotation of the corresponding gene in the sequenced genome, including the KO number and its position in the biological pathway.

3.5.4 eggNOG Annotation

eggNOG [9] is a database that catalogs clusters of orthologous genes in organisms and is continuously updated based on the COG database. Proteins that constitute each cluster of orthologous groups (COG) are assumed to originate from a common ancestral protein and have the same function. This database is generated by comparing the protein sequences of complete genomes. It is commonly used to classify and annotate genes in newly sequenced genomes.

In specific annotation, the protein sequences of non-redundant genes are compared to the eggNOG database using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value $1e-5$). The most similar sequence in the eggNOG database corresponds to the annotation and classification information of the corresponding gene in the sequenced genome.

3.5.5 Pfam Annotation

The Pfam [10] database is a protein family database that contains annotation and multiple sequence alignment information. The multiple sequence alignment information is generated using hidden Markov models. This database provides comprehensive and accurate classification information for

protein families and functional domains. Unlike typical BLAST searches, Pfam uses the software MMseqs2 (version 11-e1a1c, alignment filtering threshold: E-value $1e-5$) to compare unknown functional sequences with the Pfam database.

3.5.6 SwissProt Annotation

The SwissProt [11] database is a manually curated, non-redundant, and high-quality protein sequence database. Its distinguishing feature is that the annotation results have corresponding experimental validation, making them highly reliable.

During annotation, the protein sequences of non-redundant genes are separately compared to the protein sequences included in the SWISS-PROT and TrEMBL databases using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value $1e-5$). The most similar sequence in the two databases corresponds to the annotation information of the corresponding gene in the sequenced genome.

3.5.7 CAZy Annotation

The CAZy [12] (Carbohydrate-active enzymes) database is a collection of carbohydrate-active enzymes. It is formed by the collection and classification of published literature and related proteins, and is meticulously maintained by experts. Proteins in this database are classified into five major functional categories: glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and non-catalytic carbohydrate-binding modules (CBMs). Each of these major categories contains numerous different families. However, each family is composed of protein sequences and does not highlight the characteristic functional domains shared by the family. Carbohydrate-active enzymes in cells generally have multiple structural domains. The dbCAN tool specifically analyzes each family collected in the CAZy database and constructs hidden Markov models for the characteristic domains of each family. The software HMMER uses these hidden Markov models to identify conserved functional domains belonging to specific families.

During specific annotation, the HMMER software (version 3.0) is used to compare the protein sequences of non-redundant genes with the hidden Markov models of each family in the CAZy

database (alignment parameters: default; default filtering threshold: "if alignment > 80aa, use E-value < 1e-5, otherwise use E-value < 1e-3; covered fraction of HMM > 0.3"). This process identifies all families that meet the filtering threshold, allowing for the identification of carbohydrate-active enzymes in the genome and analysis of the number of conserved carbohydrate-related functional domains they contain.

3.5.8 CARD Annotation

The CARD [13] (Comprehensive Antibiotic Research Database) is a continuously updated database of antibiotic resistance-related information. CARD includes descriptions of antibiotics and their targets, as well as antibiotic resistance genes, related proteins, and information on antibiotic resistance mechanisms. At the core of CARD is a highly developed Antibiotic Resistance Ontology (ARO), which is used for the classification of antibiotic resistance gene data.

During annotation, the tool software rgi (version 4.2.2, default Perfect, Strict algorithm) from the CARD database is used to compare the protein sequences of non-redundant genes with the database. It identifies matching sequences in the database and obtains corresponding resistance genes and resistance-related information.

3.5.9 PHI-base Annotation

The PHI-base [14] (Pathogen Host Interactions Database) is a database of pathogen-host interactions. It includes experimentally validated genes primarily from fungal, oomycete, and bacterial pathogens, and the infected hosts include animals, plants, fungi, and insects. This database contains detailed descriptions of predicted proteins involved in the pathogen infection process.

During specific annotation, the protein sequences of non-redundant genes are compared to the protein sequences included in the PHI-base database using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value 1e-5). The most similar sequence in the database corresponds to the annotation information and classification information of the corresponding gene in the sequenced genome.

3.5.10 CYPED Annotation

Cytochrome P450 monooxygenases (CYPs) are heme-containing enzymes involved in the physiological metabolism of important substances in many microorganisms, plants, animals, and humans. CYPs catalyze the oxidation of a wide range of endogenous substances in biosynthetic and biodegradation pathways, including xenobiotics such as drugs and environmental pollutants. Therefore, understanding the specific substrates of human CYPs is crucial in drug development. The CYPED [15] (Cytochrome P450 Engineering Database) is a database of cytochrome P450 that contains information on sequences, sequence alignments, annotations, and structures. The corresponding data, including sequences, structures, and annotation information, are extracted from GenBank and PDB and manually confirmed.

During specific annotation, the protein sequences of non-redundant genes are compared to the protein sequences included in the Cytochrome P450 Engineering Database using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value $1e-5$). The most similar sequence in the database is identified, providing the corresponding classification information for the P450 family.

3.5.11 QS Annotation

Quorum sensing (QS) is a bacterial collective behavior regulation mechanism that has received increasing attention recently. It refers to the changes in the physiological and biochemical characteristics of bacterial populations as their density changes, exhibiting features and collective behaviors that individual or few bacteria do not possess in response to environmental changes. During quorum sensing, many bacteria can synthesize and secrete one or more autoinducers, and bacteria sense these autoinducers to assess the density of the bacterial population and changes in the surrounding environment. When the population reaches a certain threshold, a series of genes are activated to regulate the collective behavior of the bacteria.

QS genes are retrieved from the UniprotKB (<http://www.uniprot.org/>) database and the QS database is constructed. The QS genes are classified into nine categories: activator, autoinducer, autoinducer_producer, autoinducer_receptor, effector, regulator, repressor, transporter, and unclear (unclassified). During specific annotation, the protein sequences of non-redundant genes are

compared to the protein sequences included in the QS database using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value $1e-5$). The most similar sequence in the database is identified, providing the corresponding annotation information for the QS genes.

3.5.12 BacMet Annotation

In addition to antibiotics, antibacterial biocides and metal resistance can also affect the development of bacterial antibiotic resistance through co-selection. The BacMet (antibacterial biocide and metal resistance genes database) is a database that includes antibacterial biocide and metal resistance genes. BacMet integrates resistance gene information based on molecular function and induced resistance phenotypes. It consists of 753 experimentally validated resistance genes and 155,512 predicted resistance genes (with sequence similarity to experimentally validated resistance genes).

During specific annotation, the protein sequences of non-redundant genes are compared to the protein sequences included in the BacMet database using BLAST alignment (diamond v0.9.29, alignment filtering threshold: E-value $1e-5$). The most similar sequence in the database is identified, providing the corresponding information on antibacterial biocide and metal resistance genes.

4 Functional Diversity Analysis

4.1 Functional Gene Composition and Abundance Analysis

4.1.1 Scatter Plot of Functional Composition

In the KEGG and eggNOG functional annotations, the annotations of each sample were counted. In the KEGG functional annotation, the number of KO (KEGG Orthologs) and pathway annotations for each sample were calculated.

4.1.2 CAZy Functional Annotation

In the CAZy functional annotation, the number of annotations to each sample for Class and Family were calculated and counted.

4.1.3 CARD Functional Annotation

In the CARD functional annotation, the number of annotations to each sample for ARO and Resistance were calculated and counted.

4.2 Beta Diversity Analysis of Functional Genes

4.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a technique for analyzing and simplifying a dataset. By decomposing the variance and reflecting the differences between multiple sets of data on a two-dimensional coordinate graph, the two features with the largest variance are selected as the coordinate axes. Analyzing the composition of functional genes in different samples can reflect the differences and distances between samples. The closer two samples are on the PCA plot, the more similar the composition of functional genes in these two samples. PCA plots at different levels were generated using the R programming language tool.

4.2.2 Principal Coordinates Analysis

Principal Coordinates Analysis (PCoA) [16] is a dimensionality reduction and ordination method similar to PCA. The principle assumes that there is data measuring the differences or distances between N samples. This method finds a Cartesian coordinate system that represents N samples as points, where the squared Euclidean distances between points exactly correspond to the original dissimilarity data. This allows for the quantitative transformation of qualitative data and extraction of the main elements and structure from multidimensional data. PCoA can be used to classify multiple samples and further demonstrate the diversity of functional genes between samples.

Based on the distance matrices obtained from Beta diversity analysis, PCoA results were plotted using the R programming language tool. Samples that are closer on the coordinate plot indicate greater similarity in functional composition.

4.2.3 NMDS Analysis

Non-Metric Multi-Dimensional Scaling (NMDS) is an ordination method commonly used in ecological research. It simplifies the study objects (samples or variables) in multidimensional space to a lower-dimensional space for positioning, analysis, and classification, while preserving the original relationships between objects. Similar to PCA or PCoA, differences between groups or within groups can be observed from the distribution of samples. NMDS was originally designed to overcome the limitations of linear models, including PCA and PCoA, by employing a nonlinear model. It better reflects the nonlinear structure of ecological data, and some studies suggest that NMDS performs better than PCA/PCoA. It is recommended to have a minimum of 10 samples when not grouping them, and for grouped samples, each group should have a minimum of 5 samples. The analysis software version parameters are Python for analysis and R for plotting.

4.2.4 UPGMA Analysis

UPGMA (Unweighted Pair-group Method with Arithmetic Mean) is a commonly used clustering analysis method, also known as hierarchical clustering.

Based on the distance matrices obtained from Beta diversity analysis, UPGMA clustering was performed using the R programming language tool. UPGMA clustering is used to assess the similarity of functional gene composition between samples. In the sample hierarchical clustering tree, samples that are closer together have shorter branch lengths, indicating a greater similarity in the composition of functional genes between the two samples.

4.2.5 Analysis of Similarities

Anosim (Analysis of Similarities) is a statistical method used to analyze the similarity between groups in multidimensional data. In PCoA and NMDS analyses, there is no corresponding statistical test to determine the significance of differences between sample points in different groups. Anosim analysis is a non-parametric statistical test that can assess whether there are significant differences between samples in different groups.

4.2.6 PerMANOVA/Adonis

PerMANOVA (Permutational Multivariate Analysis of Variance), also known as Adonis analysis, is a permutation-based multivariate analysis of variance. Similar to Anosim analysis, it decomposes the total variance using a semi-metric (e.g., Bray-Curtis) or metric distance matrix (e.g., Euclidean) to analyze the explanatory power of different grouping factors on sample differences. Permutation tests are used to assess the statistical significance of the results.

4.3 Inter-group Differences Analysis

4.3.1 Parametric Tests

Common statistical methods can be divided into two types: parametric tests and non-parametric tests. Parametric tests are used when the population distribution is known. They involve inferring or hypothesis testing of population parameters using sample statistics. Non-parametric tests, on the other hand, do not rely on knowledge of the population distribution and compare the positions of the population distributions rather than population parameters.

Welch's t-test, also known as the t' test, is used to test differences between two groups of samples. ANOVA (Analysis of Variance), also known as the F test, is primarily used to test whether the means of multiple samples are equal, thereby determining the significance of differences in means between two or more samples.

When there are two groups, Welch's t-test is used for parametric tests of inter-group differences. If there are more than two groups, ANOVA analysis is used for parametric tests of inter-group differences.

4.3.2 Non-parametric Tests

The rank-sum test is a non-parametric test method. In this test, the "rank" refers to the order or position of the data, and the sum of the ranks is called the "rank sum." The rank-sum test uses the rank sum as a test statistic for hypothesis testing.

The rank-sum test is used in practical situations when the population distribution is unknown, when

the population distribution is known but does not follow a normal distribution, when certain variables may not be precisely measured, or when there is heterogeneity of variances.

Different analysis methods are used for different types of samples in this analysis:

- 1) Paired samples - Signed-rank test.
- 2) Two independent groups - For small samples, the Wilcoxon rank-sum test is used. For large samples, the Mann-Whitney U test is used. Small samples are defined as having a sample size less than or equal to 20, while large samples have a sample size greater than 20.
- 3) Multiple independent groups - Kruskal-Wallis rank-sum test. Pairwise comparisons can be conducted using the Nemenyi test. The Kruskal-Wallis rank-sum test requires that each group has a sample size of at least 5.

4.3.3 MetagenomeSeq Analysis

metagenomeSeq [17] (v1.22.0) is an R package used for differential abundance testing of genes or species in metagenomic data. This software applies a novel normalization method to account for the impact of uneven sequencing depths. It also utilizes a zero-inflated Gaussian distribution mixture model to address the effects of abundance differences caused by insufficient sampling.

4.3.4 Lefse

lelse [18] (Linear Discriminant Analysis Effect Size) is a tool implemented in Python that can identify statistically different biomarkers between different groups. It is commonly used for analyzing KEGG metabolic pathways.

4.4 Functional Gene Correlation Analysis

4.4.1 Correlation Analysis

A correlation network graph is a representation of correlation analysis. The top 80 abundant functional genes are selected, and Spearman's algorithm is used to analyze the correlation (both positive and negative) between the abundance and variation of each functional gene across samples.

Statistical tests are performed, and data groups with correlation greater than 0.5 and p-value less

than 0.05 are selected. A correlation network graph is then generated using Python.

4.4.2 RDA/CCA

RDA (Redundancy Analysis) / CCA (Canonical Correspondence Analysis) [18] is a sorting method based on correspondence analysis. RDA analysis is based on a linear model, while CCA analysis is based on a unimodal model. These methods are primarily used to reflect the relationship between functional genes or samples and environmental factors. The best analysis model is selected based on the changes in functional distribution.

Principles for selecting RDA or CCA models: First, perform DCA analysis on the function-sample data and examine the size of the first axis in the Lengths of gradient result. If it is greater than 4.0, CCA should be chosen. If it is between 3.0 and 4.0, both RDA and CCA can be selected. If it is less than 3.0, RDA results are better than CCA.

RDA or CCA analysis and plotting are performed using the `rda` or `cca` functions in the R package "vegan." In the RDA/CCA analysis results of functional diversity between samples, the relationships between points are represented by distances, where closer distances indicate similar functional profiles between samples. The relationships between rays are represented by angles, where obtuse angles indicate negative correlations and acute angles indicate positive correlations.

4.4.3 Functional Genes-Environmental Factors Correlation Heatmap

Using R language, correlation analysis can be performed between the abundance of functional genes and environmental factors. This analysis reveals the correlation relationship between environmental factors and functional genes. In the heatmap, asterisks (*) in the squares indicate significant differences ($p < 0.05$), and double asterisks (**) indicate highly significant differences ($p < 0.01$).

5 Taxonomic Diversity Analysis

5.1 Taxonomic Composition and Abundance Analysis

5.1.1 Taxonomic Annotation and Statistics

Based on the species information of non-redundant gene sequences aligned to Nr, the species composition and relative abundance information of the samples are obtained.

5.1.2 Taxonomic Composition Bar Plot

A bar plot of species composition is generated using Python, showing the composition of species at the levels of kingdom, phylum, class, order, family, genus, and species. This plot provides a visual representation of the species composition of each sample and the proportions of different species within each sample.

5.2 Alpha Diversity Analysis

Alpha diversity reflects the richness and diversity of species within an individual sample. There are several diversity indices used for measurement, including Ace, Chao1, Shannon, Simpson, goods_coverage, and pielou_evenness. Chao1 and Ace indices measure species richness, i.e., the number of species present. Shannon and Simpson indices are used to measure species diversity, taking into account both species richness and evenness in the sample community. In cases where species richness is the same, communities with greater evenness are considered to have higher diversity. Higher values of Shannon and Simpson indices indicate higher species diversity in the sample. goods_coverage is an index reflecting sequencing depth. A value close to 1 indicates reasonable sequencing depth, meaning that the depth has covered most of the species present in the sample. Conversely, a lower goods_coverage suggests that many low-abundance species may not have been sequenced. pielou_evenness, also known as Shannon's evenness, measures the relative abundance of species in a community. It represents the ratio of the actual Shannon index of a community to the maximum Shannon index that can be obtained in a community with the same

species richness. If all species have the same relative abundance, the value is 1.

5.3 Beta Diversity Analysis

5.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a technique used to analyze and simplify a dataset. By decomposing the variance, it reflects the differences between multiple sets of data on a two-dimensional coordinate graph. The two axes of the graph are selected to maximize the representation of the variance. Analyzing the species composition of different samples can reveal the differences and distances between samples. The closer two samples are on the PCA plot, the more similar their species compositions are. PCA plots at different taxonomic levels are generated using R language tools.

5.3.2 Principal Coordinates Analysis

Principal Coordinates Analysis (PCoA), similar to PCA, is a dimensionality reduction and ordination method. It assumes that there is data measuring the differences or distances between N samples. PCoA finds a Cartesian coordinate system that represents N samples as N points, such that the squared Euclidean distances between the points exactly match the original dissimilarity data. This method quantitatively transforms qualitative data and extracts the most important elements and structures from multidimensional data. PCoA can be used to classify multiple samples and further demonstrate the differences in species diversity between samples.

Based on the four distance matrices obtained from beta diversity analysis, PCoA plots are generated using R language tools. The closer the samples are on the plot, the greater their similarity in species composition.

5.3.3 NMDS Analysis

Non-Metric Multi-Dimensional Scaling (NMDS) is an ordination method commonly used in ecological research. It simplifies the study objects (samples or variables) in multidimensional space

to a lower-dimensional space for positioning, analysis, and classification while preserving the original relationships between objects. Similar to PCA or PCoA, the distribution of samples can reveal differences between groups or within groups. NMDS was designed to overcome the limitations of previous ordination methods, including PCA and PCoA, which are linear models. The model of NMDS is nonlinear and better reflects the nonlinear structure of ecological data. Some studies suggest that NMDS performs better than PCA/PCoA. It is recommended to have a minimum of 10 samples when not grouping them, and when grouping samples, each group should have a minimum of 5 samples.

5.3.4 UPGMA Analysis

UPGMA is a commonly used clustering analysis method, also known as unweighted pair-group method with arithmetic mean. It assumes that in the process of evolution, the number of divergences is equal for each lineage, meaning that the substitution rate of nucleotides or amino acids is equal and constant. The phylogenetic tree generated by UPGMA can be considered a simplified representation of a species tree, where the length of the branches from the common ancestor node to the two OTUs is the same after each divergence event. Therefore, this method is often used for species tree reconstruction.

Based on the four distance matrices obtained from beta diversity analysis, UPGMA clustering is performed using R language tools to assess the similarity of species composition between samples. The hierarchical clustering dendrogram shows that samples closer together have shorter branch lengths, indicating a higher similarity in species composition between the two samples.

5.3.5 Analysis of Similarities

ANOSIM (Analysis of Similarities), also known as similarity analysis, is a statistical method used to analyze the similarity between groups in multidimensional data. In PCoA and NMDS analyses, there is no corresponding statistical test to determine the significance of differences between sample points from different groups. ANOSIM analysis is a non-parametric statistical test that can assess the significance of differences between samples from different groups.

Based on the distance matrices obtained from beta diversity analysis using the binary Jaccard and Bray-Curtis algorithms, ANOSIM analysis is performed using the vegan package in R.

5.3.6 PerMANOVA/Adonis

PerMANOVA (Permutational multivariate analysis of variance), also known as Adonis analysis, is similar to ANOSIM. It decomposes the total variance using a semi-metric (such as Bray-Curtis) or a metric distance matrix (such as Euclidean) to analyze the explanatory power of different grouping factors on sample differences. It uses permutation tests to assess the statistical significance.

Based on the distance matrices obtained from beta diversity analysis using the binary Jaccard and Bray-Curtis algorithms, PerMANOVA analysis is performed using the vegan package in R.

5.4 Inter-group Differences Analysis

5.4.1 Parametric Tests

Common statistical methods can be divided into two types: parametric tests and non-parametric tests. Parametric tests are used when the distribution of the population is known, and they involve inferring or hypothesis testing of population parameters based on sample statistics. Non-parametric tests, on the other hand, do not rely on knowledge of the population distribution and compare the positions of the population distributions rather than the population parameters.

Welch's t-test, also known as the t'-test, is used to test differences between two groups of samples. ANOVA (Analysis of Variance), also known as the F-test, is primarily used to test whether the means of multiple samples are equal, thereby assessing the significance of differences in means between two or more samples.

When there are only two groups, Welch's t-test is used to test for between-group differences in species composition. If there are more than two groups, ANOVA analysis is used to test for between-group differences in species composition.

5.4.2 Non-Parametric Tests

The rank-sum test is a method used in non-parametric tests. The "rank" refers to the order or position

of the data, and the sum of ranks is referred to as the "rank sum." The rank-sum test uses the rank sum as a test statistic for hypothesis testing.

The rank-sum test is used in practical situations where the population distribution is unknown, or when the known population distribution does not follow a normal distribution, or when certain variables cannot be precisely measured, or when there is heterogeneity of variances.

Different analysis methods are used for different types of samples:

- 1) Paired samples - Signed-rank test
- 2) Two independent groups - For small samples, the Wilcoxon rank-sum test is used; for large samples, the Mann-Whitney U test is used. Small samples are defined as having a sample size less than or equal to 20, while large samples have a sample size greater than 20.
- 3) Multiple independent groups - Kruskal-Wallis rank-sum test. Pairwise comparisons can be performed using the Nemenyi test. The Kruskal-Wallis rank-sum test requires a minimum of 5 samples per group.

These analyses can be used to assess significant differences in species, genes, or functions between two or multiple groups of samples. The p-values can also be adjusted (e.g., using the false discovery rate correction).

5.4.3 MetagenomeSeq Analysis

metagenomeSeq [17] is an R package used for differential abundance testing of genes or species. This software applies a novel normalization method to mitigate the impact of uneven sequencing depth. It also utilizes a zero-inflated Gaussian mixture model to address the influence of abundance differences caused by undersampling.

5.4.4 LEfSe

LEfSe [19] (Linear Discriminant Analysis Effect Size) is a tool implemented in Python that identifies biomarkers with statistically significant differences between different groups.

5.4.5 Random Forest Analysis

Random Forest Analysis (implemented in R) is a machine learning algorithm that consists of multiple decision trees. It efficiently and rapidly identifies the most important species categories (biomarkers) for classification based on the original samples.

5.5 Correlation Analysis

5.5.1 Taxonomic Correlation Analysis

A correlation network graph is a representation of correlation analysis. The top 80 abundant species are selected based on their abundance and variation across samples. Spearman's correlation algorithm is used to analyze the correlations (both positive and negative) between species, and statistical tests are conducted. Data groups with correlation coefficients greater than 0.5 and p-values less than 0.05 are selected. A correlation network graph is then generated using Python.

5.5.2 RDA/CCA

Redundancy Analysis (RDA) and Canonical Correspondence Analysis (CCA) [18] are ordination methods based on correspondence analysis. RDA is based on a linear model, while CCA is based on a unimodal model. These methods are primarily used to investigate the relationship between microbial communities or samples and environmental factors. The choice of the analysis model depends on the distribution pattern of the species.

The selection criteria for RDA or CCA models are as follows: first, perform Detrended Correspondence Analysis (DCA) using species-sample data and examine the length of the first axis (Lengths of gradient). If it is greater than 4.0, CCA should be chosen. If it falls between 3.0 and 4.0, either RDA or CCA can be used. If it is less than 3.0, RDA is preferred over CCA.

R language, specifically the vegan package, is used for RDA or CCA analysis and visualization. In the RDA/CCA analysis results plot, the relationships between points are represented by distances, where closer distances indicate similar sample compositions. The relationships between rays are represented by angles, with obtuse angles indicating negative correlations and acute angles

indicating positive correlations.

5.5.3 Taxonomy-Environment Factors Correlation Heatmap

R can be used to analyze the correlation between species abundance and environmental factors, revealing the relationships between them. A heatmap can be generated based on these correlation relationships.

6 Reference

1. Li D, Liu C M, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph[J]. *Bioinformatics*, 2015, 31(10):1674-1676.
2. Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: quality assessment tool for genome assemblies[J]. *Bioinformatics*, 2013, 29(8):1072-1075.
3. Zhu W, Lomsadze A, Borodovsky M. Abinitio gene identification in metagenomic sequences. *Nucleic acids research*. 2010, 38(12):132-132.
4. Fu L., et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 2012, 28(23): 3150-3152.
5. Deng Y Y, Li J Q, Wu S F, et al. Integrated nr database in protein annotation system and its localization[J]. *Comput Eng*, 2006, 32(5): 71-74.
6. Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology[J]. *Nature genetics*, 2000, 25(1): 25-29.
7. Conesa A, Götz S, García-Gómez J M, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. *Bioinformatics*, 2005, 21(18), 3674-3676.
8. Kanehisa M, Goto S, Kawashima S, et al. The kegg resource for deciphering the genome[J]. *Nucleic acids research*, 2004, 32(suppl 1): D277-D280.
9. Powell S , Forslund K , Szklarczyk D , et al. eggNOG v4.0: nested orthology inference across 3686 organisms[J]. *Nucleic Acids Research*, 2014, 42(Database issue):231-9.

10. Finn R D , Bateman A , Clements J , et al. Pfam: the protein families database.[J]. Nucleic Acids Research, 2014, 42(Database issue):222-30.
11. Bairoch A , Apweiler R . The SWISS-PROT protein sequence data bank and its new supplement TREMBL[J]. Nucleic Acids Research, 1996, 24(1):21.
12. Cantarel B L, Coutinho P M, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics[J]. Nucleic acids research, 2009, 37(suppl 1): D233-D238.
13. Jia B, Raphenya A R, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database[J]. Nucleic acids research, 2016:gkw1004.
14. Winnenburg R, Baldwin T K, Urban M, et al. PHI-base: a new database for pathogen host interactions[J]. Nucleic acids research, 2006, 34(suppl 1): D459-D464.
15. Fischer M , Knoll M , Sirim D , et al. The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family[J]. Bioinformatics, 2007, 23(15):2015-2017.
16. Sakaki T, Takeshima T, Tominaga M, Hashimoto H, Kawaguchi S: Recurrence of ICA-PCoA aneurysms after neck clipping. Journal of neurosurgery, 1994, 80(1):58-63.
17. Paulson JN, Stine O C, Bravo H C, et al. Differential abundance analysis for microbial marker-gene surveys[J]. Nature Methods, 2013, 10(12):1200.
18. Lindström ES, Kamst-Van Agterveld MP, Zwart G: Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. Applied and environmental microbiology, 2005, 71(12):8201-8206.
19. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation[J]. Genome Biology, 2011, 12(6):R60.