Innovative Metabolomics Insights for Better Health

Microbial 16S report

Metware Biotechnology Inc.

www.metwarebio.com

Address: 8A Henshaw St. , Woburn, MA 01801, USA Tell: +1（781）975-1541

MWXS-005 16S ASV demo report EN

1 Analysis Overview

The 16S rRNA is located on the small subunit of prokaryotic ribosomes and comprises 10 conserved regions and 9 hypervariable regions. The conserved regions exhibit little variation among bacteria, while the hypervariable regions are specific to genera or species, showing variations based on different phylogenetic relationships. Therefore, 16S rDNA serves as a characteristic nucleic acid sequence for revealing biological species and is considered the most suitable indicator for bacterial systematics, phylogeny, and classification. The 16S rDNA amplicon sequencing, typically selecting one or more variable regions, utilizes universal primers designed for the conserved regions to amplify through PCR. Subsequently, sequencing analysis and species identification are performed on the hypervariable regions. (Caporaso et al. 2011; Youssef et al. 2009; Hess et al. 2011).

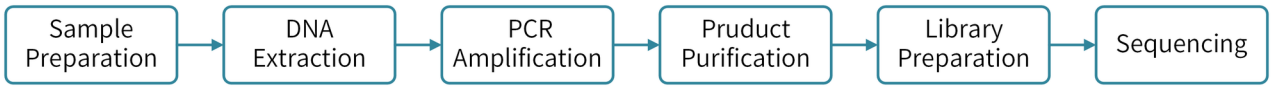
With the continuous development of high-throughput sequencing platforms, the upgraded NovaSeq sequencing platform now supports PE250 strategy for paired-end sequencing, achieving the same read length as the MiSeq platform but with significantly improved throughput and sequencing quality. This makes it a more suitable platform for 16S amplicon sequencing. The high throughput and depth of NovaSeq PE250 sequencing are particularly advantageous for identifying low-abundance microbial community species, thereby enhancing the comprehensiveness of microbial community studies. Hence, it is expected to become the preferred choice for studying microbial community diversity.

Based on the characteristics of the amplified 16S region, a small fragment library is constructed, and paired-end sequencing is performed on this library using the Illumina NovaSeq sequencing platform. After reads assembly and filtering, representative sequences are generated through clustering or denoising methods, enabling species annotation and abundance analysis. Through alpha diversity and beta diversity analyses, differences in species composition and community structure between samples can be revealed. Moreover, personalized analysis and in-depth data mining can be conducted according to project requirements.

2 Analysis Workflow

2.1 Experimental On-machine Process

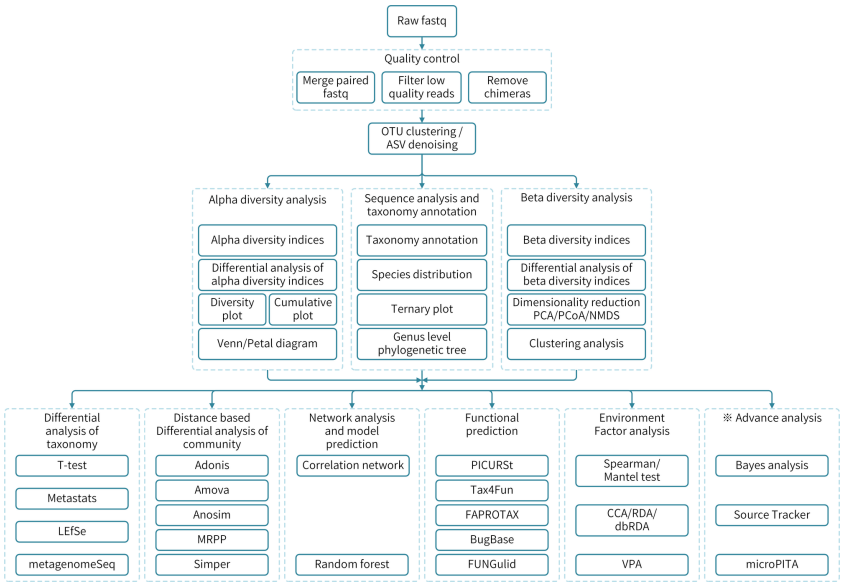
Throughout the process from DNA sample to final data acquisition, each step including sample detection, PCR, purification, library construction, and sequencing can impact both the quality and quantity of the data. The quality of the data, in turn, directly affects the results of subsequent information analysis. To ensure the accuracy and reliability of sequencing data from the source, Mevvy Metabolism rigorously controls every experimental step, including sample detection, library construction, and sequencing. This fundamental control ensures the production of high-quality data. The workflow diagram is presented below:



Flow Chart of Experimental Detection

2.2 Bioinformatics Analysis Pipeline

The sequenced raw data (Raw Data) contains a certain proportion of interference data (Dirty Data). To make the results of information analysis more accurate and reliable, the raw data is first filtered and spliced to obtain clean data (Clean Data). Then, denoising analysis is performed based on the effective data to generate Amplicon Sequence Variants (ASVs). According to the ASV analysis results, on the one hand, species annotation is performed for each ASV sequence to obtain corresponding species information and abundance distribution based on species. At the same time, abundance and Alpha diversity calculations, Venn diagrams, or petal diagrams are performed on ASVs to obtain information on species richness and evenness within samples, as well as shared and unique ASV information between different samples. On the other hand, multiple sequence alignments and phylogenetic trees are constructed for ASVs, and dimensional reduction analyses such as PCoA, PCA, NMDS, and sample clustering trees are displayed to explore differences in community structure between different samples or groups. To further explore the differences in community structure between grouped samples, statistical analysis methods such as T-test, Simper, Metastats, LEfSe, Anosim, and MRPP are used to test the significance of differences in species composition and community structure of grouped samples. The annotation results of amplicons can also be correlated with corresponding functional databases to predict the functional analysis of microbial communities in the samples. The information analysis process after obtaining the sequencing data is shown in the figure below:



Flow Chart of Bioinformatics Analysis

Description: When the number of samples is less than three, advanced analyses including Beta diversity analysis, significance analysis of differences in community structure between groups, significance analysis of differences in species between groups, and correlation analysis of environmental factors cannot be performed. If no grouping information is available or if there are less than three biological replicates, significance analysis of differences in community structure between groups and significance analysis of differences in species between groups cannot be performed. Correlation analysis of environmental factors requires environmental factor data from the client.

3 Species Annotation

3.1 Sequencing Data Preprocessing

Processing of Illumina NovaSeq sequencing raw data (Raw PE) involves quality control and splicing to obtain Clean Tags. Subsequently, chimeric filtering is applied to obtain Effective Tags, which are used for subsequent analysis. Statistical results at each step of the data processing are provided in the table below:

Table 1 Data pre-processing Statistics and Quality Control

Sample_name	Raw_tags	Clean_tags	Effective_tags	Effective_ratio(%)	Effective_bases(nt)	Min_length	Max_length	Mean_length	Q20(%)	Q30(%)	GC(%)
CC1	91,195	90,820	79,018	86.65	33,213,987	155	430	420	98.48	95.13	53.95
CC2	90,969	90,400	82,484	90.67	34,369,889	151	430	417	98.63	95.42	53.24
CC3	91,000	90,410	84,793	93.18	35,144,639	161	430	414	98.72	95.79	52.93
BB1	89,801	88,496	79,915	88.99	34,071,130	151	430	426	98.45	94.98	52.74
BB2	92,323	90,882	82,311	89.16	35,048,244	150	430	426	98.38	94.83	52.63
BB3	88,479	86,898	80,356	90.82	34,196,282	150	430	426	98.35	94.70	52.48
DD1	82,233	81,818	76,023	92.45	32,585,423	171	430	429	98.59	95.27	54.20
DD2	86,641	84,961	78,054	90.09	33,388,463	150	430	428	98.42	95.00	53.97
DD3	85,430	84,314	77,924	91.21	33,337,132	151	430	428	98.54	95.30	54.60
AA1	86,324	85,842	69,208	80.17	28,736,392	152	430	415	98.56	95.36	53.41
AA2	88,662	87,507	76,401	86.17	31,677,078	151	430	415	98.45	95.10	53.42
AA3	86,033	84,545	75,933	88.26	31,482,676	151	430	415	98.43	95.13	52.95

File path: 01.Quality_control/Clean_data

- Raw Tags: Paired-end reads from the original sequencing.
- Clean Tags: Sequences obtained after splicing and filtering for low quality and short length from Raw Tags.
- Effective Tags: Tags sequences used for subsequent analysis after filtering chimeras.
- Effective Ratio (%): The percentage of Effective Tags relative to Raw Tags.
- Effective Bases (nt): The number of bases in Effective Tags.
- Effective Base: The number of bases in the final Effective sequence.
- Min Length: The shortest length among Effective Tags.
- Max Length: The longest length among Effective Tags.
- Mean Length: The average length of Effective Tags.
- Q20 (%) and Q30 (%): The percentage of bases with quality values greater than 20 (sequencing error rate less than 1%) and 30 (sequencing error rate less than 0.1%) in Effective Tags.
- GC (%): Represents the GC content in Effective Tags.

3.2 ASV Analysis

In general, to study the species composition of each sample, effective data from each sample are clustered into Operational Taxonomic Units (OTUs) based on the 97% sequence similarity principle. Subsequently, the sequences of OTUs are annotated for species identification.

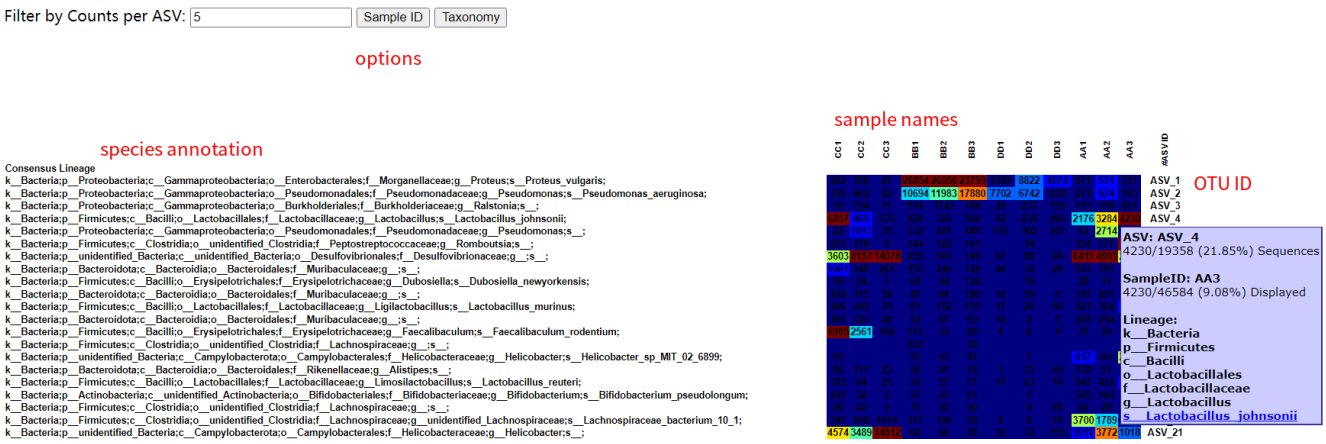
OTU clustering analysis addresses accuracy issues arising from sequencing errors but reduces phylogenetic resolution because sequences above the similarity threshold cannot be distinguished. To overcome

this problem, denoising methods attempt to achieve nucleotide resolution using statistical approaches. There are three main software tools for implementing denoising methods: DADA2, UNOISE3, and Deblur. While DADA2 and Deblur yield similar results, Deblur supports parallel processing, making it faster and more stable (producing identical sequences across different samples). Deblur is the default choice for denoising analysis, but if needed, DADA2 can also be employed. Both Deblur and DADA2 are implemented using QIIME 2.

Deblur compares Hamming distances within samples and between sequences, using upper error curves. It employs a greedy algorithm to obtain Amplicon Sequence Variants (ASVs). The DADA2 algorithm, on the other hand, constructs an error rate model to infer whether an amplicon sequence variant comes from the template, using its own data's error model as a parameter, independent of other parameter distribution models.

3.2.1 ASV Statistics

Statistical analysis of the number of ASVs in the samples and the generation of a heatmap are presented in the figure below:



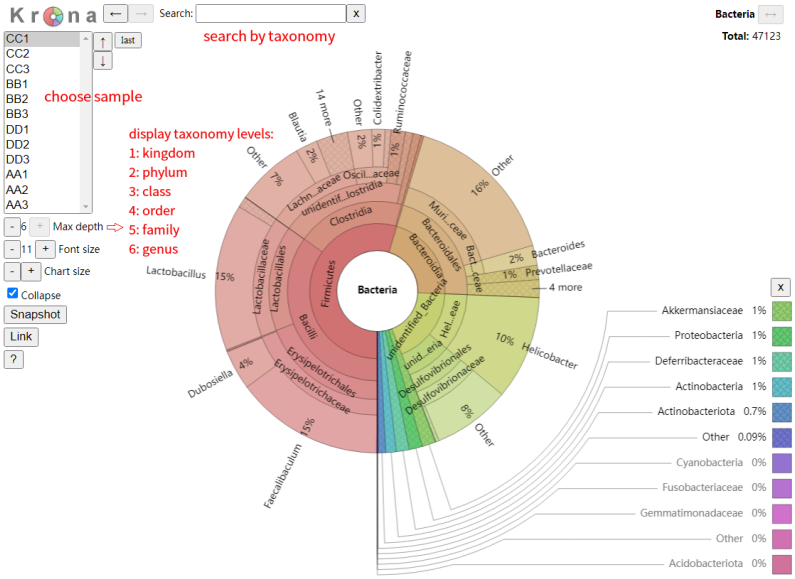
Heatmap of Sample ASV Sequence Numbers

Description: Horizontal represents ASVs, vertical represents samples, the number on the grid is the number of ASV sequences in the sample, the redder the color means the higher the number of ASVs, the bluer means the lower the number of ASVs.

File path: 03.ASV_visualization/ASV_heatmap/ASV.table.heatmap.html

Visualization of species annotation results using KRONA. In the displayed results, circles from the innermost to the outermost represent different taxonomic levels. The size of the sectors reflects the relative

proportions of different ASV annotation results. For more detailed information, please refer to the KRONA chart interpretation guide. An example is shown below:

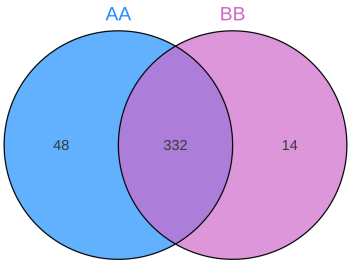


Visualization of Species Annotation Results

File path: 03.ASV_visualization/ASV_krona/ASV.krona.html

3.2.2 Venn Diagram or Petal Diagram Based on ASV

ASVs obtained through denoising methods were used to analyze shared and unique ASVs among different samples (groups). When the number of samples (groups) is less than or equal to 5, a Venn diagram is plotted. When the number of samples (groups) is greater than 5, a petal diagram is shown. Both the Venn diagram and petal diagram are normalized across all samples.



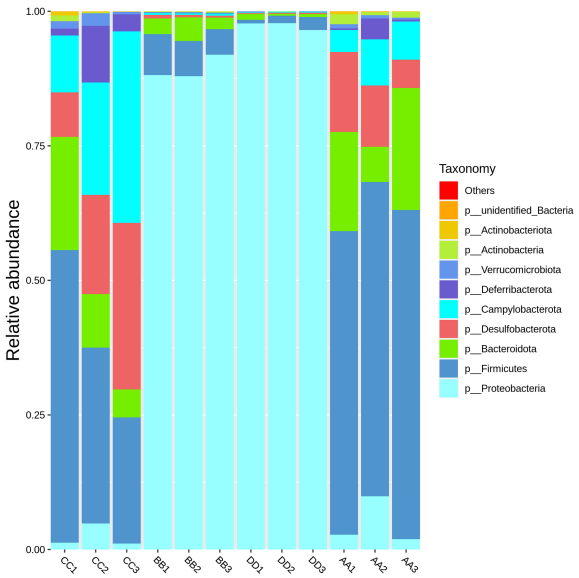
ASV-based Venn Diagram

Description: Each circle in the graph represents a sample (group). The number in the overlapping area between the circles represents the number of ASVs shared between these samples (groups), while the number in the non-overlapping area represents the number of ASVs specific to that sample (group).

File path: 03.ASV_visualization/venn_flower

3.3 Species Relative Abundance Display

Based on the species annotation results, the top 10 species with the highest abundance at each taxonomic level (Phylum, Class, Order, Family, Genus, Species) were selected for each sample or group. Stacked bar plots of species' relative abundance were generated to visually examine the species with higher relative abundance and their proportions in different taxonomic levels for each sample. An example of a stacked bar plot at the phylum level is shown below:

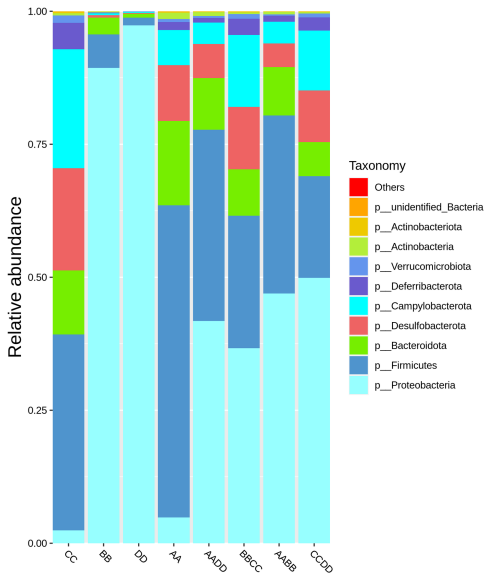


Stacked Bar Chart of ASV-based Relative Abundance of Species at the Phylum Level

Description: The horizontal coordinates represent the sample names; the vertical coordinates represent the relative abundances; "Others" represents the sum of the relative abundances of all the other phylums in the plot except for these 10 phylums.

File path: 03.ASV_visualization/top10_barplot/phylum

Stacked bar plots of species' relative abundance at the phylum level for different groups are shown below:



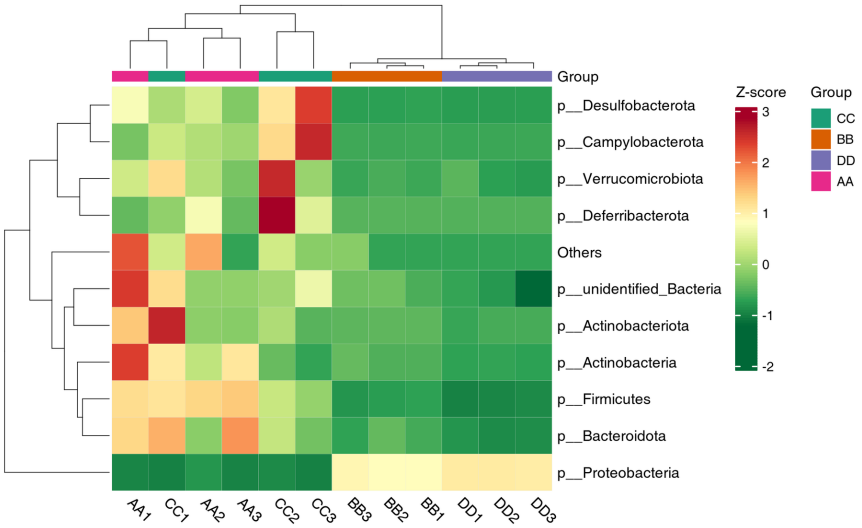
Stacked Bar Chart of Relative Abundance of Species at the Phylum Level for Different Groupings Based on ASV

Description: The horizontal coordinates represent the groupings; the vertical coordinates represent the relative abundances; "Others" represents the sum of the relative abundances of all the other phylums in the plot except for these 10 phylums.

File path: 03.ASV_visualization/top10_barplot_group/phylum

3.4 Species Abundance Cluster Heatmap

The top 35 microbial classifications based on the sum of quantitative values across all samples were selected. Clustering was performed based on the quantitative information of each species in each sample, and a heatmap was drawn. This allows for the identification of species that are more abundant or lower in content in specific samples and enables the assessment of clustering relationships between samples. The results are shown in the figure:

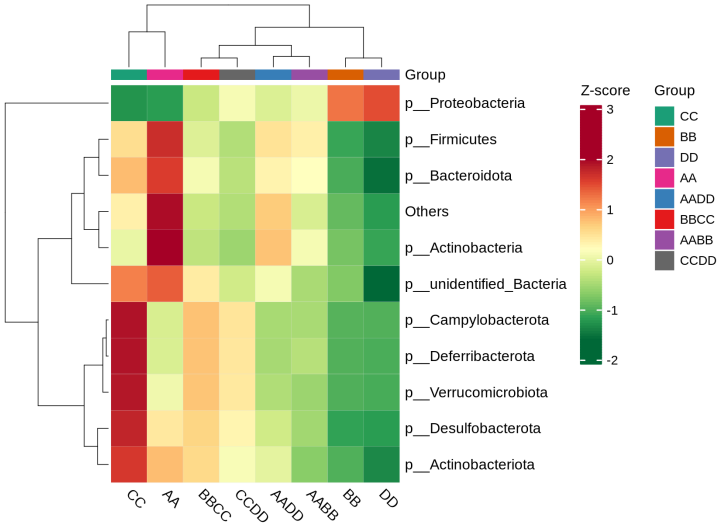


Heatmap of Relative Abundance of Species at the Phylum Level Based on ASV

Description: The sample information is shown horizontally, while the species classification information is shown vertically. The clustering tree in the figure represents the clustering structure of the species. The values displayed in the heatmap are the relative quantitative data standardized by Z-Score.

File path: 03.ASV_visualization/taxa_heatmap

Cluster heatmaps of species abundance at the phylum level for different groups are shown below:



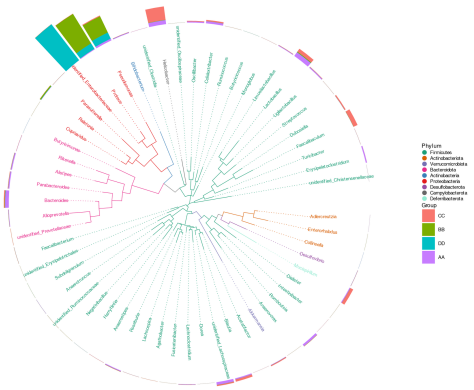
Heatmap of Species Abundance at the Phylum Level for Different Groupings Based on ASV

Description: The grouping information is shown horizontally, while the species classification information is shown vertically. The clustering tree in the figure represents the clustering structure of the species. The values displayed in the heatmap are the relative quantitative data standardized by Z-Score.

File path: 03.ASV_visualization/taxa_heatmap_group

3.5 Genus-Level Species Evolutionary Tree

To further study the systematic evolutionary relationships of species at the genus level, representative sequences of the top 100 genera are obtained through multiple sequence alignment and displayed in the tree below:



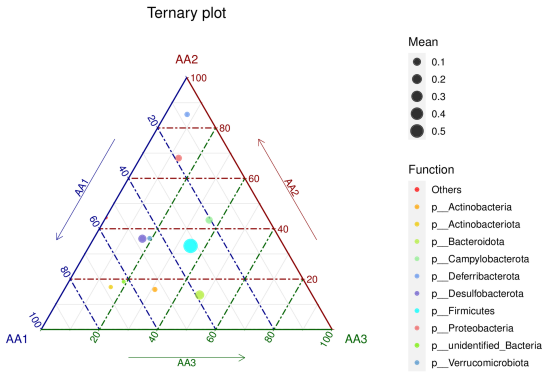
Phylogenetic Relationships between Species at the Genus Level Based on ASV

Description: Phylogenetic tree constructed from representative sequences of species at the genus level, with the colors of the branches and sectors indicating their corresponding phylum, and the stacked bars on the outside of the sector ring indicating information on the abundance distribution of the genus in different samples.

File path: 03.ASV_visualization/genus_evolution

3.6 Ternary Plot Analysis

To identify differences in dominant species between three sample groups at each taxonomic level (Phylum, Class, Order, Family, Genus, Species), the top 10 species in average abundance rankings for three sample groups at each taxonomic level were selected. A ternary plot was generated to visually examine the differences in dominant species at different taxonomic levels among the three sample groups. This analysis used the ternaryplot command from the R software vcd package. An example of a ternary plot at the phylum level is shown below:



ASV-based Ternary Phase Diagram

Description: The three vertices in the figure represent three groupings of samples, the circles represent species, the size of the circles is proportional to the relative abundance, the closer the circle is to the vertex, the higher the presence of this species in this grouping.

File path: 03.ASV_visualization/ternary_plot

4 Alpha Diversity Analysis

Alpha Diversity is used to analyze the microbial community diversity within samples. Alpha diversity analysis reflects the richness and diversity of microbial communities within individual samples. This includes assessing species richness and diversity in each sample using species accumulation boxplots, species diversity curves, and a series of statistical diversity indices.

4.1 Alpha Diversity Indices

Statistical analysis of different sample alpha diversity indices (Observed_ASV, Shannon, Simpson, Chao1, ACE, goods_coverage, PD_whole_tree) is shown in the table below:

Table 2 ASV-based Alpha Diversity Statistics Sheet

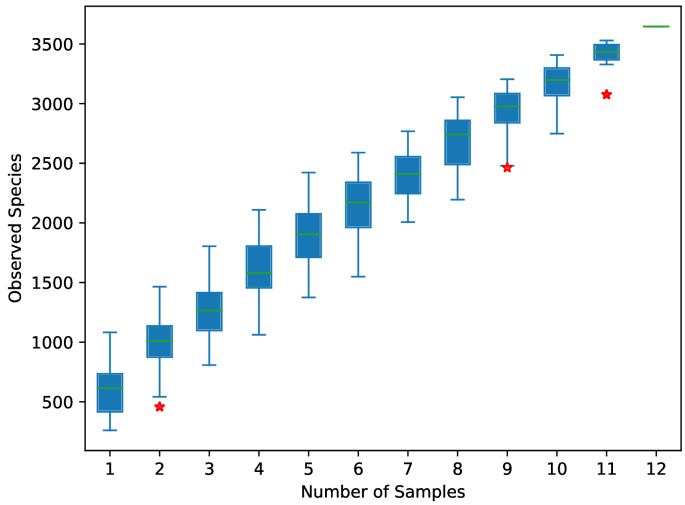
Sample	observed_ASV	Shannon	Simpson	Chao1	ACE	Goods_coverage	PD_whole_tree
CC1	342	3.710	0.931	343.000	344.226	1	23.533
CC2	308	3.700	0.934	309.500	309.432	1	22.267
CC3	298	2.657	0.801	300.786	302.581	1	21.744
BB1	271	1.560	0.560	273.528	276.405	1	21.188
BB2	300	1.653	0.590	303.886	306.018	1	22.750
BB3	215	1.385	0.594	221.333	223.399	1	19.231
DD1	120	0.807	0.366	121.552	124.261	1	13.821
DD2	114	1.115	0.543	115.037	116.872	1	13.203
DD3	152	0.745	0.284	153.875	155.588	1	14.124
AA1	348	4.349	0.962	348.857	348.891	1	24.299
AA2	305	4.073	0.962	305.273	305.792	1	22.240
AA3	294	4.021	0.962	294.625	295.415	1	21.507

File path: 04.Alpha_diversity

- Observed_ASV: The intuitively observed number of ASVs.
- Shannon: The total number of categories in the sample and their proportions. Higher community diversity and more even species distribution result in a larger Shannon index. When the sample number is 1, the base is the natural logarithm ‘e’; when greater than 1, the base is 2.
- Simpson: Characterizes the diversity and evenness of species distribution within the community.
- Chao1: Estimates the total number of species in the community sample.
- ACE: Estimates the number of ASVs in the community.
- Goods_coverage: Sequencing depth index.
- PD_whole_tree: Phylogenetic relationship of species within the community.
- Simpson has three display forms: Simpson’s Index (D), Simpson’s Index of Diversity (1 - D), and Simpson’s Reciprocal Index (1 / D). They are similar in reflecting community diversity, but the calculated results are presented differently. This analysis uses Simpson’s Index of Diversity (1 - D).

4.2 Species Accumulation Boxplot

The species accumulation boxplot is an analysis that describes the increase in species diversity as the sample size increases. It is an effective tool for investigating the species composition of samples and predicting species abundance in samples. In biodiversity and community surveys, it is widely used to determine whether the sample size is sufficient and to estimate species richness. Therefore, the species accumulation boxplot not only helps determine if the sample size is sufficient but, with a sufficient sample size, can also predict species richness (by default, analyzed when the sample size is greater than 10), as shown in the figure below:



Box plot of species accumulation based on ASV.

Description: The horizontal coordinates are sample volumes; the vertical coordinates are the post-sampling ASV numbers. The results reflect the rate of emergence of new ASVs (new species) under continuous sampling. Within a certain range, as the sample size increases, if the box plot position shows a sharp increase, it means that a large number of new species have been found in the community; when the box plot position tends to flatten, it means that the species in the environment does not increase significantly with the increase of sample volume. Species accumulation box plots can be used to determine the adequacy of the sample volume: a sharp rise in the box plot position indicates that the sample volume is insufficient and needs to be increased; if the other way around, it indicates that the sample volume is adequate for data analysis.

File path: 04.Alpha_diversity

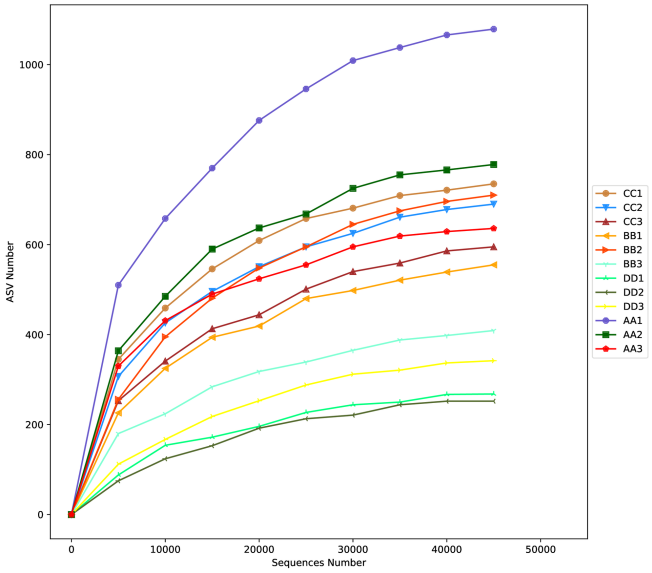
4.3 Species Diversity Curves

Rarefaction Curve and Rank Abundance Curve are common curves used to describe the diversity of samples within a group. The Rarefaction Curve involves randomly extracting a certain sequencing amount of

data from the sample, counting the number of species represented by these data (i.e., the number of ASVs), and constructing a curve with the extracted sequencing data amount and the corresponding number of species. The Rarefaction Curve directly reflects the reasonability of the sequencing data amount and indirectly reflects the richness of species in the sample. When the curve tends to flatten, it indicates that the sequencing data amount is progressively reasonable, and more data will only produce a small number of new species (ASVs).

The Rank Abundance Curve sorts ASVs in the sample in descending order by relative abundance (or the number of included sequences), obtains the corresponding sorting number, and then uses the sorting number of ASVs as the x-axis and the relative abundance of ASVs (or the relative percentage content of sequences in this ranked ASV) as the y-axis. Connecting these points with lines creates the Rank Abundance Curve, which intuitively reflects the richness and evenness of species in the sample. Horizontally, the width of the curve reflects the richness of species, with a wider span indicating higher species richness. Vertically, the smoothness of the curve reflects the evenness of species distribution, with a smoother curve indicating more even species distribution.

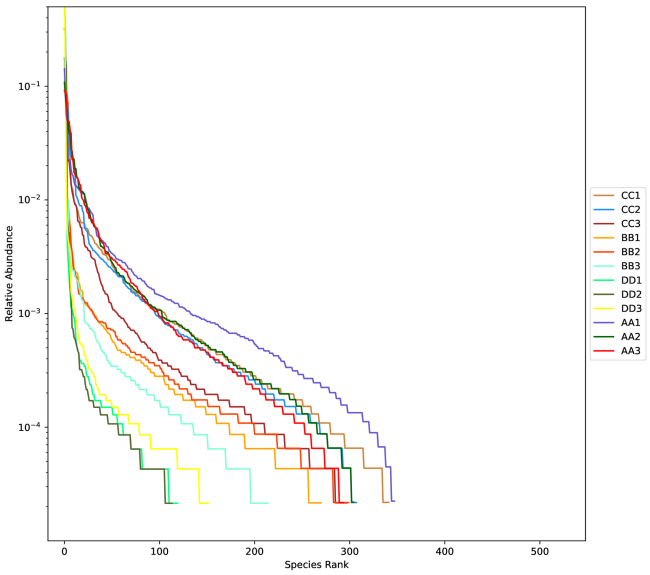
Species diversity curves are as follows:



Dilution curve of each sample based on ASV

Description: In the dilution curve, the horizontal coordinates represent the number of sequenced strips randomly selected from a sample, while the vertical coordinates represent the number of ASVs that can be constructed based on this number of sequenced strips to reflect the depth of sequencing. Different samples are represented by curves with different colors.

File path: 04.Alpha_diversity

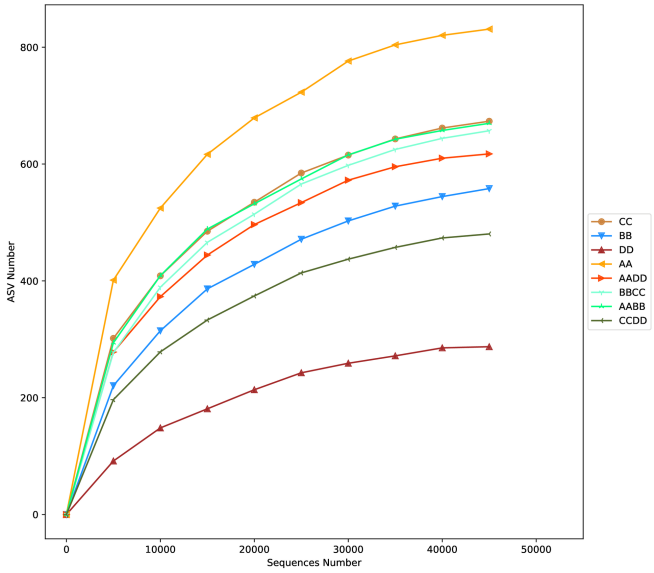


Rank Abundance curve of each sample based on ASV

Description: In the Rank Abundance curve, the horizontal coordinates are the ordinal numbers sorted by ASV abundance, and the vertical coordinates are the relative abundances of the corresponding ASVs. Different samples are represented by lines with different colors.

File path: 04.Alpha_diversity

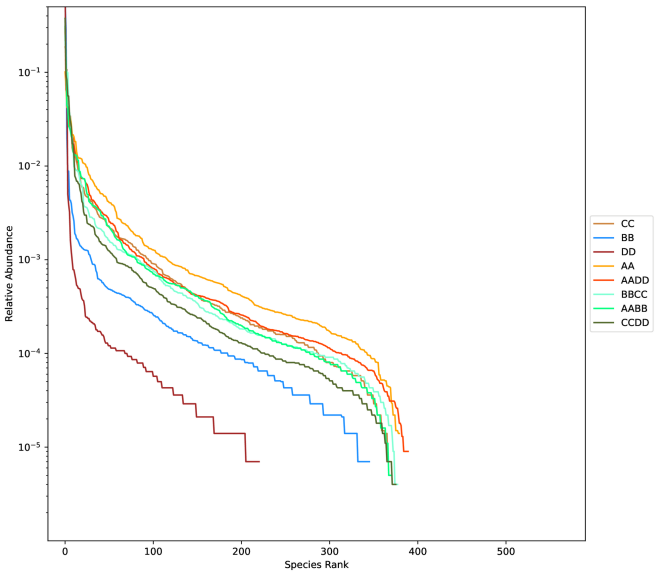
Species diversity curves analyzed by group are as follows:



Group dilution curve based on ASV

Description: In the dilution curve, the horizontal coordinates represent the number of sequenced strips randomly selected from each group, while the vertical coordinates represent the number of ASVs that can be constructed based on this number of sequenced strips to reflect the depth of sequencing. Different samples are represented by curves in different colors.

File path: 04.Alpha_diversity



Group Rank Abundance curve based on ASV

Description: In the Rank Abundance curve, the horizontal coordinates are the ordinal numbers sorted by ASV abundance, and the vertical coordinates are the relative abundances of the corresponding ASVs. Different groups are represented by lines with different colors.

File path: 04.Alpha_diversity

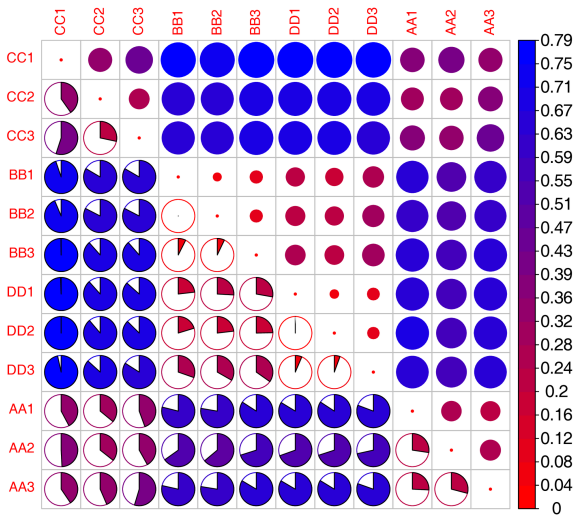
5 Beta Diversity Analysis

Beta Diversity compares the microbial community compositions of different samples. First, based on the species annotation results and ASV abundance information of all samples, the information of ASVs belonging to the same classification is merged to create a species abundance table (Profiling Table). Simultaneously, using the phylogenetic relationships between ASVs, Unifrac distance (Unweighted Unifrac) is calculated. Unifrac distance calculates the distance between samples using evolutionary information from microbial sequences in each sample, and for more than two samples, a distance matrix is obtained. Then, using the abundance information of ASVs, Weighted Unifrac distance is further constructed. Finally, through multivariate statistical methods such as Principal Component Analysis (PCA), Principal Co-ordinates Analysis (PCoA), Non-Metric Multi-Dimensional Scaling (NMDS), Unweighted Pair-group Method with Arithmetic Means (UPGMA), and analysis of differences in Beta diversity indices, differences between different samples

(groups) are discovered.

5.1 Distance Matrix Heatmap

In Beta diversity studies, Weighted Unifrac distance and Unweighted Unifrac distance are chosen to measure the dissimilarity between two samples. A smaller value indicates less difference in species diversity between these two samples. The Heatmap generated using Weighted Unifrac is shown in the figure:



Beta Diversity Index Heatmap based on ASV

Description: The circles in the grids within the upper triangle indicate the beta diversity between samples. The smaller the circle and the redder the color, means the smaller the beta diversity value and the smaller the diversity difference between samples. The meaning of the color and circle size in the lower triangle is the same as that in the upper triangle.

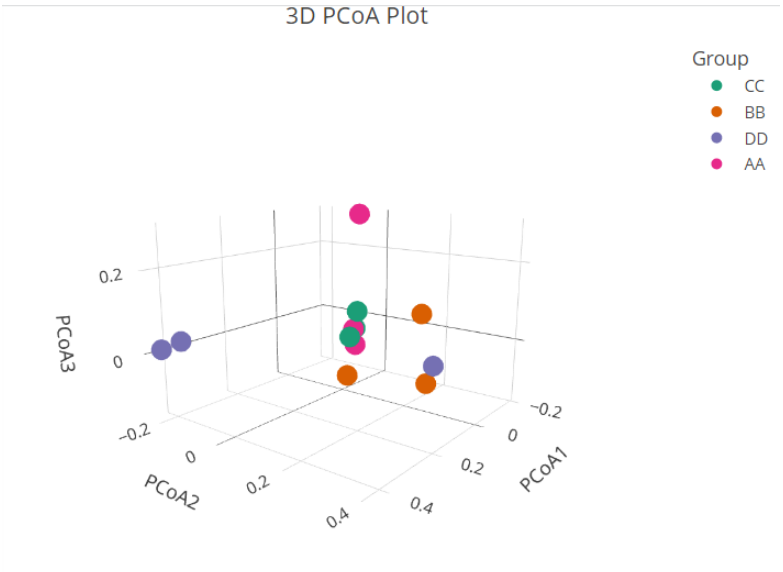
File path: 05.Beta_diversity/beta_heatmap

5.1.1 PCoA Analysis

Principal Co-ordinates Analysis (PCoA) extracts the most important elements and structures from multidimensional data through the sorting of eigenvalues and eigenvectors. We conducted PCoA analysis based

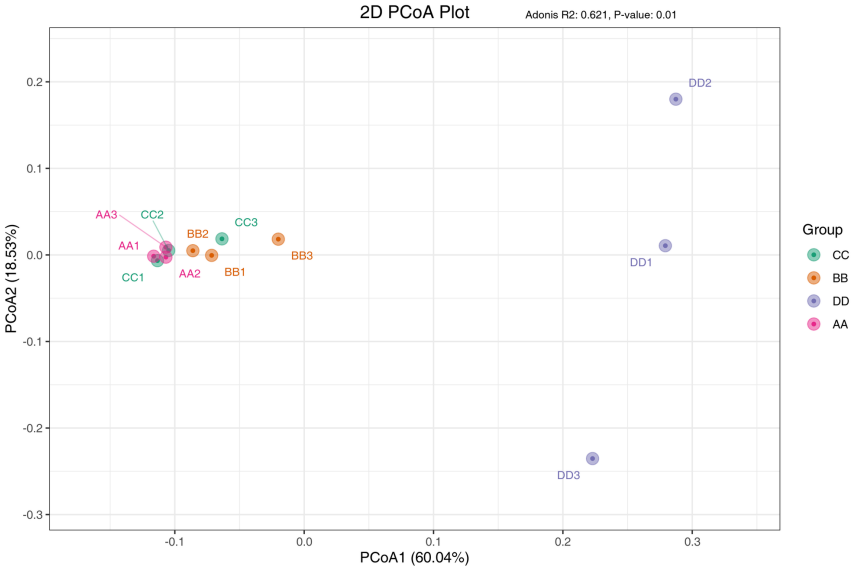
on Weighted Unifrac distance and Unweighted Unifrac distance, selecting the main coordinate combinations with the highest contribution rate for plotting. If sample distances are closer, it indicates a more similar species composition structure. Therefore, samples with high similarity in community structure tend to cluster together, while samples with significant differences in community structure are separated by a considerable distance.

PCoA is presented in two forms: two-dimensional and three-dimensional. The two-dimensional PCoA plot uses the first and second principal coordinates for display, while the three-dimensional PCoA plot uses three principal coordinates, and the coordinates can be flexibly adjusted. The results are presented in an interactive web page format and can be viewed in the webpage file.



Example of 3D PCoA

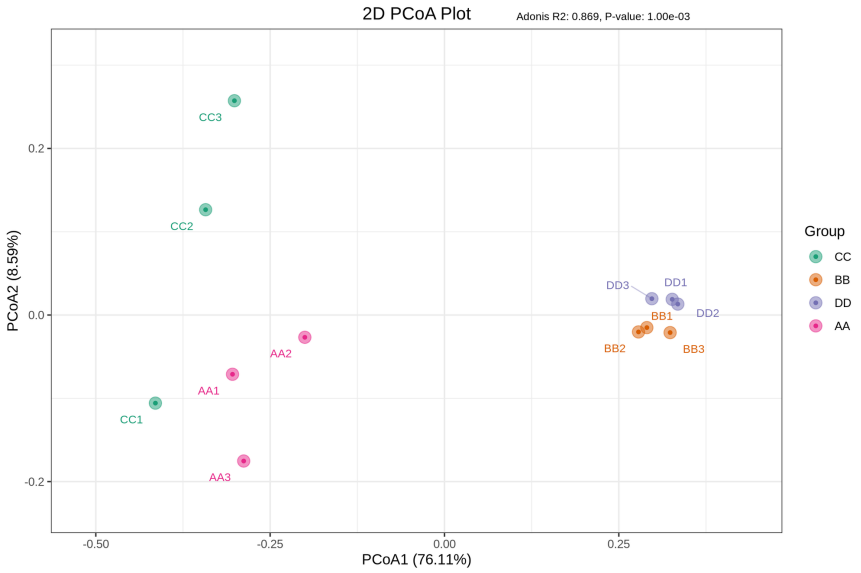
The two-dimensional PCoA results are displayed below:



Unweighted Unifrac Distance PCoA Based on ASV

Description: The horizontal coordinate indicates a principal component, the vertical coordinate indicates another principal component, and the percentage indicates the contribution of the principal component to the difference between samples; each point in the plot indicates a sample, and the samples from the same group are represented by the same color.

File path: 05.Beta_diversity/PCoA



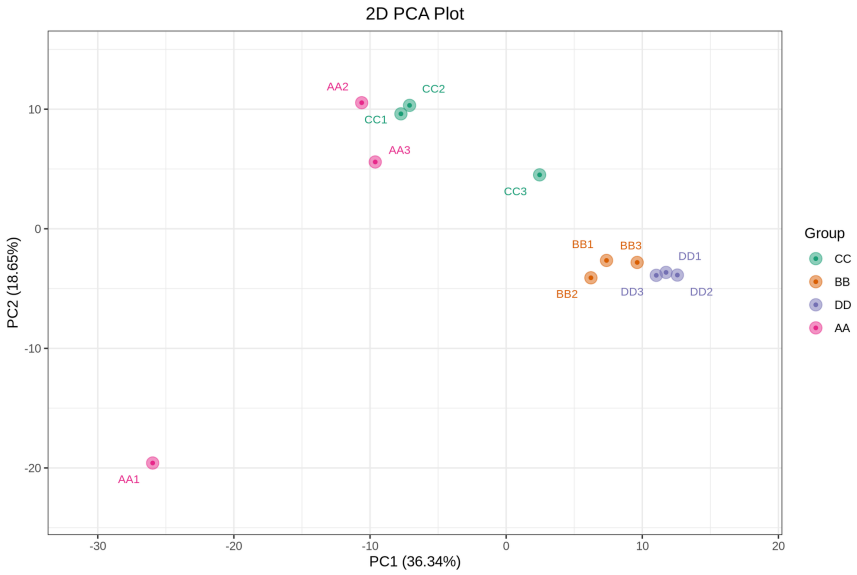
Weighted Unifrac Distance PCoA Based on ASV

Description: The horizontal coordinate indicates a principal component, the vertical coordinate indicates another principal component, and the percentage indicates the contribution of the principal component to the difference between samples; each point in the plot indicates a sample, and the samples from the same group are represented by the same color.

File path: 05.Beta_diversity/PCoA

5.1.2 PCA Analysis

Principal Component Analysis (PCA) is a method that decomposes variance based on Euclidean distances, reducing multidimensional data to extract the most significant elements and structures. PCA analysis extracts two coordinate axes that reflect the maximum differences between samples, portraying the differences in two-dimensional coordinate graphs and revealing simple patterns in complex data backgrounds. The closer the communities are in sample composition, the closer they are in the PCA plot. The PCA analysis results at the OTU level are shown in the figure:



PCA Based on ASV

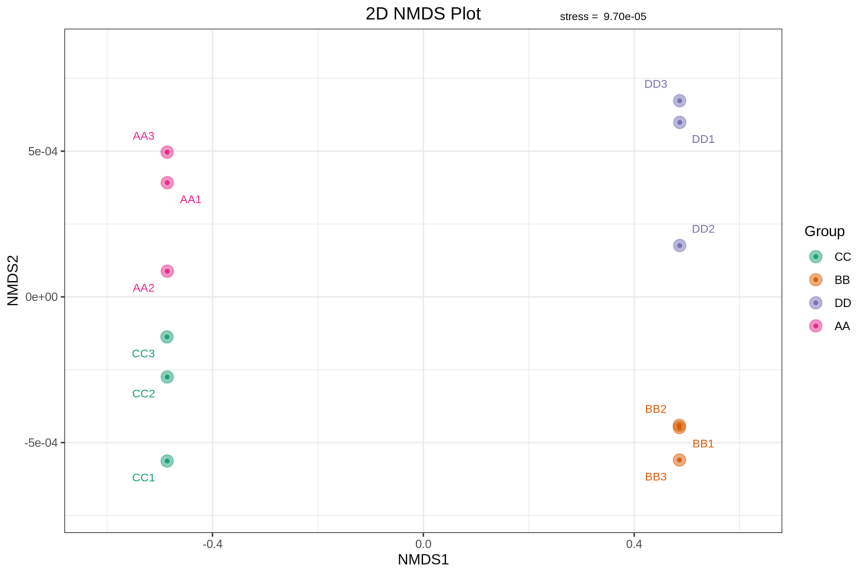
Description: the horizontal coordinate indicates the first principal component, while the percentage indicates the contribution of the first principal component to the sample variance; the vertical coordinate indicates the second principal component, while the percentage indicates the contribution of the second principal component to the sample variance; each dot in the plot indicates a sample, with samples in the same group represented by the same color; groups of more than 3 samples are allowed to insert ellipses to indicate confidence intervals, with the same color as the group.

File path: 05.Beta_diversity/PCA

5.1.3 NMDS Analysis

Non-Metric Multi-Dimensional Scaling (NMDS) is a statistical sorting method suitable for ecological research. NMDS is a non-linear model based on Bray-Curtis distance, reflecting the species information contained in samples in two-dimensional space. Its design aims to overcome the limitations of linear models (including PCA, PCoA) to better reflect the non-linear structure of ecological data. In NMDS analysis, points in multidimensional space reflect the species information contained in samples, and the degree of difference between different samples is represented by the distance between points. This method can effectively capture

inter-group and intra-group differences. The NMDS analysis results based on the OTU level are shown in the figure:



NMDS Based on ASV

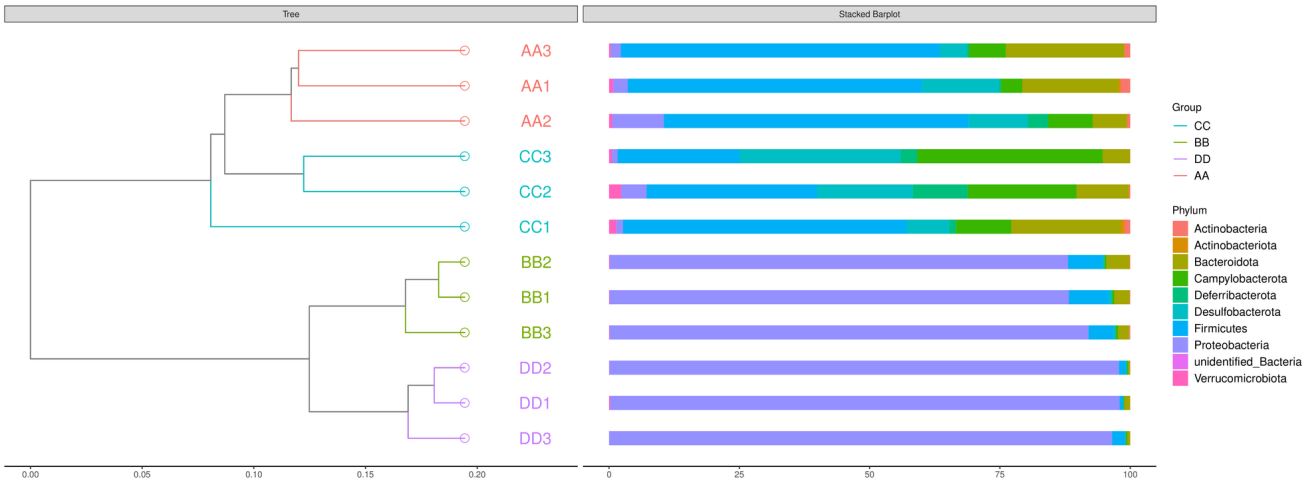
Description: Each dot in the plot represents a sample; the distance between the dots indicates the degree of variation; samples in the same group are represented with the same color. When Stress is less than 0.2, it means that NMDS can accurately reflect the degree of variation among samples.

File path: 05.Beta_diversity/NMDS

5.2 UPGMA Clustering Tree

To study the similarity between different samples, clustering analysis can be performed on the samples to construct a clustering tree. In environmental biology, Unweighted Pair-group Method with Arithmetic Mean (UPGMA) is a commonly used clustering analysis method, originally developed to solve classification problems. The basic idea of UPGMA is to first cluster the two samples with the smallest distance, forming a new node (a new sample) with the branching point at the midpoint of the distance between the two samples. Then, the average distance between the new “sample” and other samples is calculated, and the two samples with the smallest distance are clustered again. This process is repeated until all samples are clustered together, resulting in a complete clustering tree.

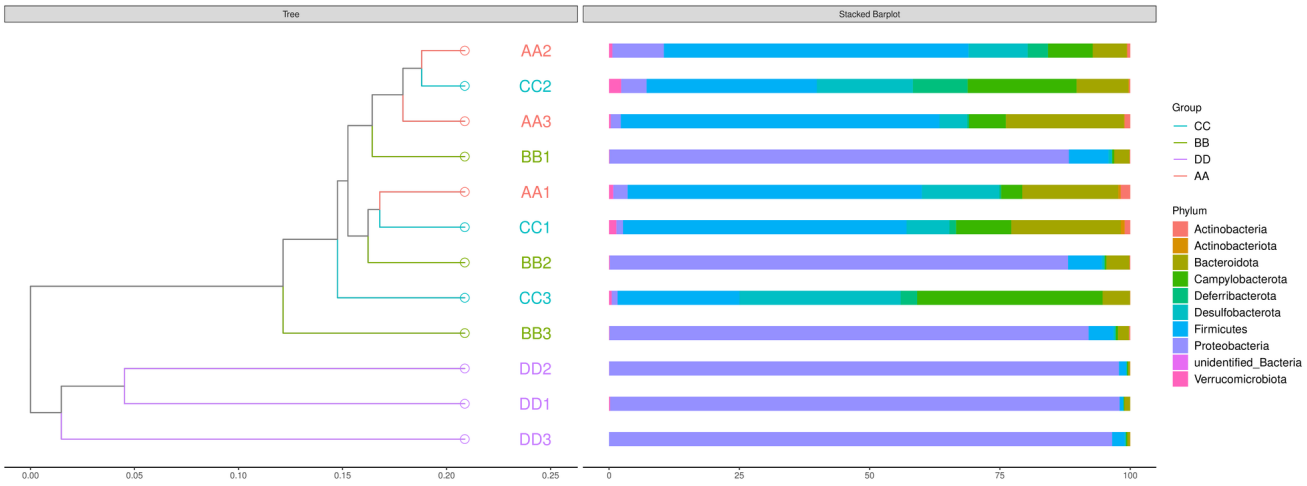
UPGMA clustering analysis was performed using Weighted Unifrac and Unweighted Unifrac distance matrices. The clustering results were integrated with the relative abundance of species at the phylum level for each sample, as shown in the figures below:



UPGMA Clustering Tree Based on Weighted Unifrac Distance of ASVs

Description: On the left is the UPGMA clustering tree structure, and on the right is the distribution of relative abundance of species at the phylum level for each sample.

File path: 05.Beta_diversity/Tree/wunifrac



UPGMA Clustering Tree Based on Unweighted Unifrac Distance of ASVs

Description: On the left is the UPGMA clustering tree structure, and on the right is the distribution of relative abundance of species at the phylum level for each sample.

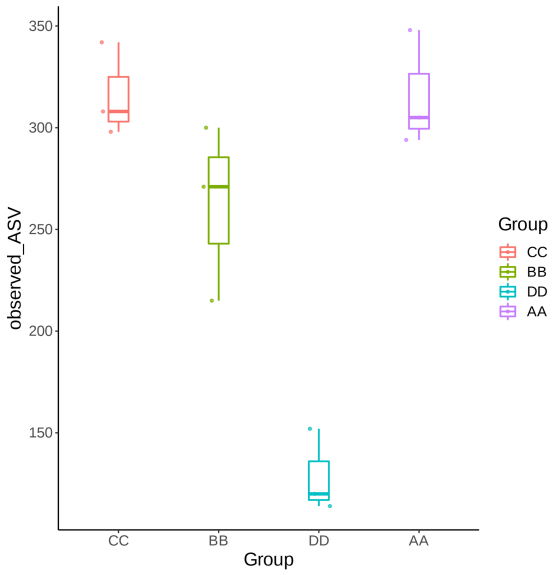
File path: 05.Beta_diversity/Tree/unifrac

6 Statistical Tests

6.1 Intergroup Difference Analysis

6.1.1 Intergroup Analysis of Alpha Diversity Indices

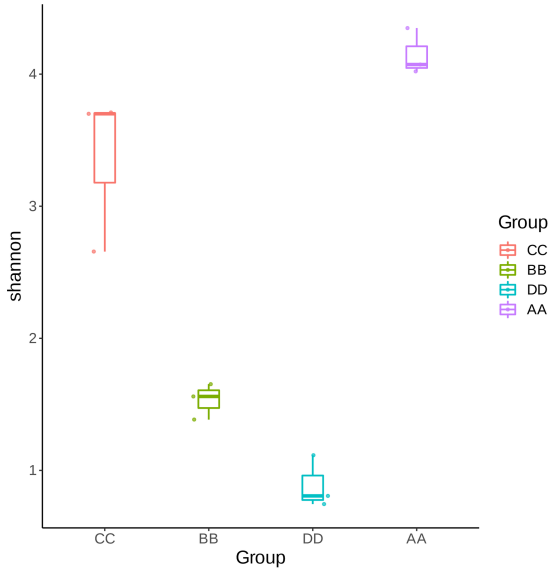
In the intergroup analysis of alpha diversity indices, boxplots visually reflect the median, dispersion, maximum, minimum, and outliers of species diversity within groups. Simultaneously, the significance of intergroup differences in species diversity is analyzed through T-test, Wilcox, Tukey, and Kruskal-Wallis tests (T-test and Wilcox rank-sum test are performed when there are only 2 groups, while Tukey and Kruskal-Wallis tests are performed when there are more than 2 groups). Taking observed_species and Shannon indices as examples, the boxplots for intergroup differences in species diversity are as follows:



Box Plot of Intergroup Variation in observed species Index Based on ASV

Description: The horizontal coordinate indicates the grouping name; the vertical coordinate indicates the observed species index; the horizontal line in the middle of the box plot indicates the median value.

File path: 06.Diff_analysis/alpha_stat



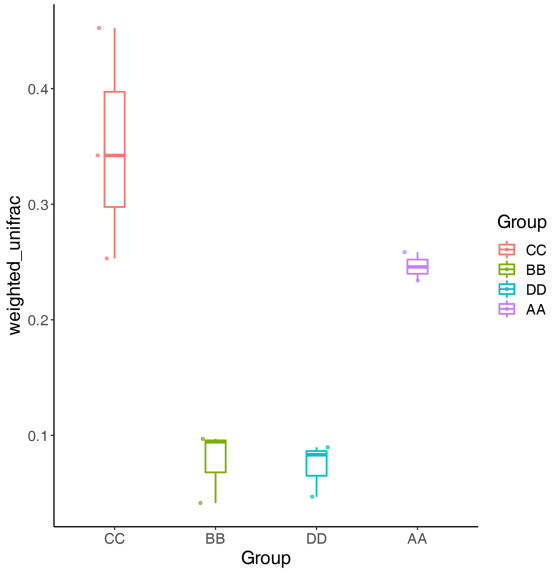
Box Plot of Intergroup Variation in Shannon Index Based on ASV

Description: The horizontal coordinate indicates the grouping name; the vertical coordinate indicates the shannon index; the horizontal line in the middle of the box plot indicates the median value.

File path: 06.Diff_analysis/alpha_stat

6.1.2 Intergroup Analysis of Beta Diversity Indices

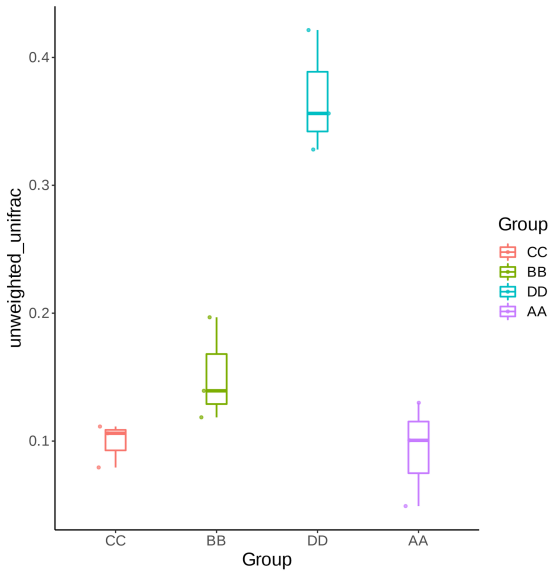
Boxplots for intergroup analysis of beta diversity visually depict the median, dispersion, maximum, minimum, and outliers of sample similarity within groups (For interpretation of boxplots, please refer to Box plot). Simultaneously, the significance of intergroup differences in beta diversity is analyzed through T-test, Wilcox, Tukey, and Kruskal-Wallis tests (T-test and Wilcox rank-sum test are performed when there are only 2 groups, while Tukey and Kruskal-Wallis tests are performed when there are more than 2 groups). The boxplots for intergroup analysis of beta diversity are shown below:



Box plot of Weighted Unifrac Beta Diversity Based on ASV

Description: The horizontal coordinate indicates the grouping name, the vertical coordinate indicates the intra-group weighted unifrac index, and the horizontal line in the middle of the box plot indicates the median value.

File path: 06.Diff_analysis/beta_stat/weighted_unifrac



Box Plot of Unweighted Unifrac Beta Diversity Based on ASV

Description: The horizontal coordinate indicates the grouping name; the vertical coordinate indicates the intra-group unweighted unifrac index; the horizontal line in the middle of the box plot indicates the median value.

File path: 06.Diff_analysis/beta_stat/unweighted_unifrac

6.2 Intergroup Significant Test of Community Structure Differences

6.2.1 Anosim

Anosim analysis is a non-parametric test used to determine if the differences between groups are significantly greater than the differences within groups, thereby assessing the meaningfulness of grouping. Anosim analysis uses the anosim function from the R vegan package, and it is based on the ranks of Bray-Curtis distance values for intergroup difference significance testing. Detailed calculation processes can be viewed here. The analysis results are presented in the table and figure below:

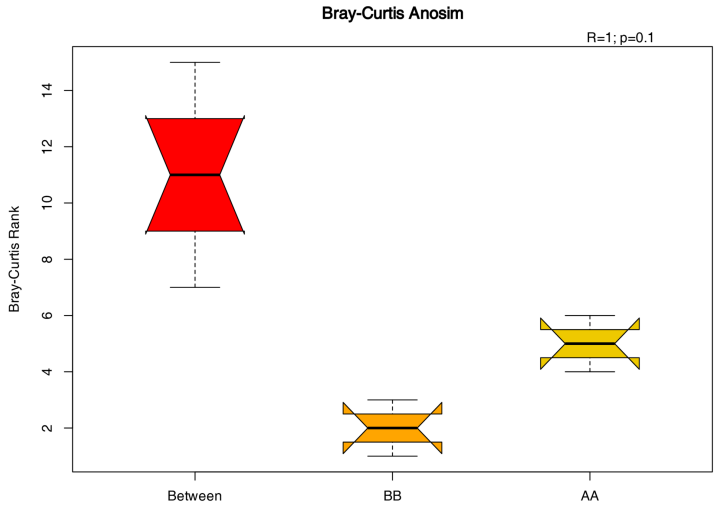
Table 3 Anosim Inter-group Variation Analysis Based on ASV

Group	R-value	P-value
CC_vs_BB	1.0000000	0.1
CC_vs_DD	1.0000000	0.1
CC_vs_AA	0.7407407	0.1
BB_vs_DD	1.0000000	0.1
BB_vs_AA	1.0000000	0.1
DD_vs_AA	1.0000000	0.1

File path: 06.Diff_analysis/Anosim

- Group: Grouping
- R-value: R-value ranges between (-1, 1), where an R-value greater than 0 indicates significant inter-group differences, and an R-value less than 0 indicates that within-group differences are greater than intergroup differences.
- P-value: $P < 0.05$ indicates statistical significance

For the Anosim analysis results, ranks based on the distance values between pairwise samples (between for intergroup, within for within-group) are obtained. This way, the comparison of any two groups can yield data for three categories, which are then presented in boxplots (if the notches of two boxes do not overlap, it indicates a significant difference in their medians), as shown below:



Anosim Inter-group Variation Analysis Based on ASV

Description: The vertical coordinate indicates the rank of the distance between the samples; the horizontal coordinate: Between shows the results between the two groups, and the other two are the results within their respective groups.

File path: 06.Diff_analysis/Anosim

6.2.2 MRPP

MRPP analysis is similar to Anosim, but MRPP is a parametric test based on Bray-Curtis distance, used to analyze whether the differences in microbial community structure between groups are significant. It is usually used in conjunction with dimensionality reduction plots such as PCA, PCoA, NMDS. MRPP analysis uses the mrpp function from the R vegan package, and detailed calculation processes can be viewed here. The analysis results are presented in the table:

Table 4 MRPP Inter-group Variation Analysis Based on ASV

Group	A	Observed_delta	Expected_delta	Significance
CC_vs_BB	0.5150544	0.3259808	0.6722007	0.1
CC_vs_DD	0.5060787	0.3551257	0.7189925	0.1
CC_vs_AA	0.1021204	0.5028547	0.5600469	0.1
BB_vs_DD	0.6883415	0.1613376	0.5176742	0.1
BB_vs_AA	0.5261285	0.3090665	0.6522159	0.1
DD_vs_AA	0.5208472	0.3382115	0.7058532	0.1

File path: 06.Diff_analysis/MRPP

- Group: Grouping
- A: A value greater than 0 indicates that intergroup differences are greater than within-group differences, and a value less than 0 indicates that within-group differences are greater than intergroup differences.
- Observed_delta: Smaller values indicate smaller within-group differences.
- Expected_delta: Larger values indicate larger between-group differences.
- Significance: Values less than 0.05 indicate significant differences.

6.2.3 Adonis

ADONIS, also known as permutational MANOVA or nonparametric MANOVA, is a nonparametric multivariate analysis of variance method based on Bray-Curtis distance. This method analyzes the explanatory power of different grouping factors on sample differences and uses permutation tests to analyze the statistical significance of grouping. ADONIS analysis uses the adonis function from the R vegan package, and detailed calculation processes can be viewed here. The analysis results are presented in the table:

Table 5 Adonis Inter-group Variation Analysis Based on ASV

Group	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
CC_vs_BB	1(4)	1.077(0.294)	1.077(0.0735)	14.666	0.786(0.214)	0.1
CC_vs_DD	1(4)	1.231(0.312)	1.231(0.078)	15.790	0.798(0.202)	0.1
CC_vs_AA	1(4)	0.288(0.511)	0.288(0.12775)	2.256	0.361(0.639)	0.1
BB_vs_DD	1(4)	0.84(0.056)	0.84(0.014)	60.113	0.938(0.062)	0.1
BB_vs_AA	1(4)	1.037(0.255)	1.037(0.06375)	16.235	0.802(0.198)	0.1

Table 5 Adonis Inter-group Variation Analysis Based on ASV Continued table

Group	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
DD_vs_AA	1(4)	1.22(0.274)	1.22(0.0685)	17.839	0.817(0.183)	0.1

File path: 06.Diff_analysis/Adonis

- Group: Grouping
- Df: Degrees of freedom. Values in parentheses correspond to residual items, and so on.
- SumsOfSqs: Total variance, also known as the sum of squared deviations.
- MeanSqs: Mean square (deviation), i.e., SumsOfSqs/Df.
- F.Model: F-test value.
- R2: Indicates the proportion of the sample difference explained by different grouping, i.e., the ratio of group variance to total variance. A higher R2 indicates a higher explanatory power of grouping for differences.
- Pr(>F): Indicates the P-value, with values less than 0.05 indicating high statistical confidence.

6.2.4 Amova

Analysis of Molecular Variance (AMOVA) is similar to ANOVA and is a nonparametric analysis method for testing the significance of differences between different groups based on weighted or unweighted Unifrac distance matrices. We conducted intergroup difference analysis using the mothur software's amova function based on Weighted Unifrac distance. Detailed calculation processes can be viewed here. The analysis results are presented in the table:

Table 6 Amova Inter-group Variation Analysis Based on ASV

Group	SS	df	MS	Fs	P_value
AA-BB-CC-DD	1.33374(0.201636)	3(8)	0.444578(0.0252046)	17.63880	<0.001*
AA-BB	0.520744(0.0673094)	1(4)	0.520744(0.0168274)	30.94630	0.096
AA-CC	0.106411(0.189239)	1(4)	0.106411(0.0473096)	2.24925	0.091
AA-DD	0.56307(0.0663512)	1(4)	0.56307(0.0165878)	33.94480	0.094
BB-CC	0.677793(0.135285)	1(4)	0.677793(0.0338213)	20.04040	0.119
BB-DD	0.084541(0.0123979)	1(4)	0.084541(0.00309947)	27.27600	0.112
CC-DD	0.714912(0.134327)	1(4)	0.714912(0.0335818)	21.28870	0.1

File path: 06.Diff_analysis/Amova

- Group: Grouping
- SS: Total variance, also known as the sum of squared deviations. Values in parentheses correspond to residual items, and so on.
- df: Degrees of freedom.
- MS: Mean square (deviation), i.e., SS/df.
- Fs: F-test value.
- P_value: Indicates the P-value, with values less than 0.05 indicating significant differences between groups.

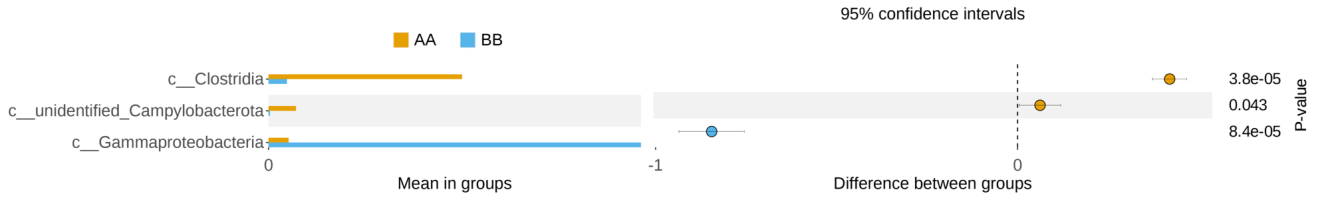
6.3 Intergroup Differential Species Analysis

For projects with grouping, in-depth studies can be conducted through statistical analysis of community structure differences. Through statistical analysis, species with significant abundance changes between groups can be identified, and the enrichment of differential species between different groups can be obtained. Additionally, the magnitude of intra-group and inter-group differences can be compared to assess whether the differences in community structure between different groups are statistically significant.

6.3.1 T-test

To find different species between groups at each taxonomic level (Phylum, Class, Order, Family, Genus, Species), a T-test is performed to identify species with significant differences (p-value < 0.05). Results are

initially displayed at the phylum level, and if there are no significantly different species at this level, the next taxonomic level is displayed, and so on. The results are shown in the figure:



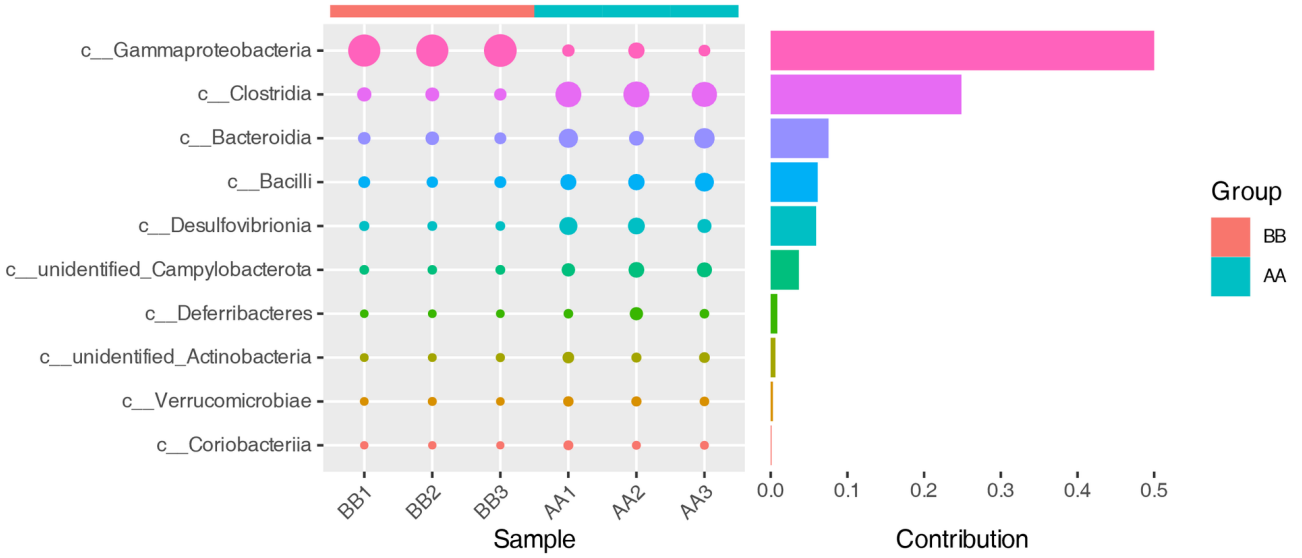
Plot of T-test Inter-group Species Variation Analysis Based on ASV

Description: The left panel shows the abundance of species that differ between groups, with each bar in the panel representing the mean value in each group for species that vary significantly in abundance between groups, respectively. The right panel shows the confidence level of the difference between groups, with the leftmost point of each circle in the panel indicating the lower limit of the 95% confidence interval for the difference in means, and the rightmost point of the circle indicating the upper limit of the 95% confidence interval for the difference in means. The center of the circle represents the difference in means. The group represented by the circle color is the group with the high mean. At the rightmost end of the displayed results are the p-values of the inter-group significance tests for the corresponding differential species.

File path: 06.Diff_analysis/T-test/phylum/

6.3.2 Simper

Simper (Similarity percentage) decomposes the Bray-Curtis dissimilarity index, quantifying the contribution of each species to the difference between two groups. The results show the top 10 species in terms of contribution to differences between two groups and their abundances. Simper analysis uses the simper function in the R vegan package, and the results are shown below:



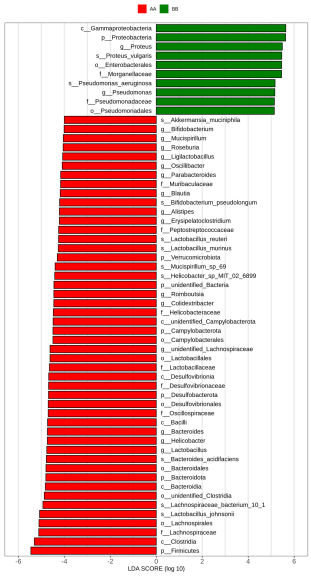
Simper Differential Contribution Based on ASV

Description: The top 10 contributing species are selected for plotting by default. The vertical axis represents the species, while the horizontal axis represents the samples. The size of the bubble represents the relative abundance of the species, and Contribution shows the contribution of the species to the variation between the two groups.

File path: 06.Diff_analysis/Simper/phylum/

6.3.3 LEfSe

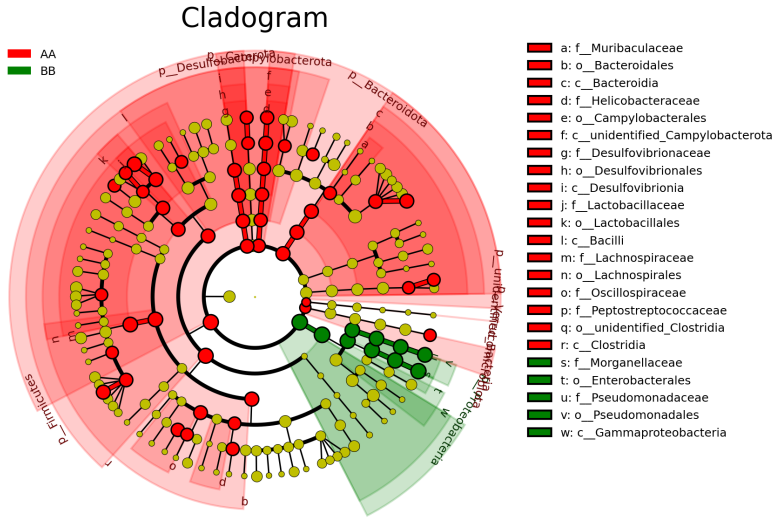
LEfSe (LDA Effect Size) is an analysis tool for discovering and interpreting biomarkers (genes, pathways, and taxonomic units) in high-dimensional biological data. It can be used to compare two or more groups, emphasizing statistical significance and biological relevance, helping researchers identify features of different abundances and associated categories between groups. LEfSe statistical results include a histogram of LDA values, an evolutionary branch diagram (phylogenetic distribution), and a comparison of the abundances of biomarkers with statistical differences between groups. The results are shown below:



Bar Chart of LDA Value Distribution Based on ASV

Description: The LDA value distribution bar chart shows species with an LDA Score greater than the set value (set to 4 by default), i.e. biomarkers that are statistically different between groups. It shows the species with significant differences in abundance in the different groups, with the length of the bars representing the magnitude of the contribution of the differential species (i.e., the LDA Score).

File path: 06.Diff_analysis/LEfSe



Phylogenetic Tree Based on ASV

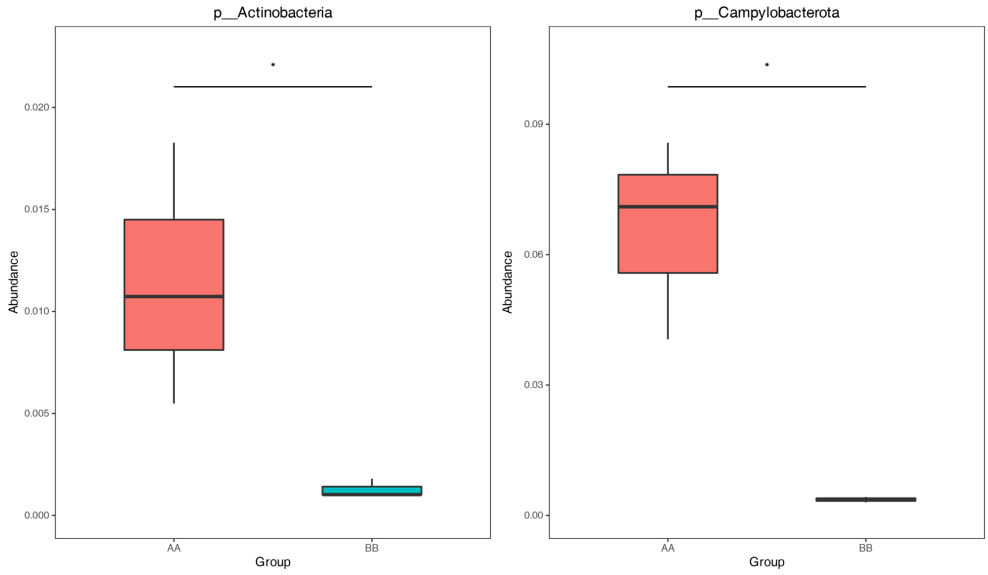
Description: In a phylogenetic tree , circles radiating from inside to outside represent taxonomic levels from phylum to genus (or species). Each small circle at a different taxonomic level represents a taxon at that level, and the size of the circle is proportional to the relative abundance. Coloring principle: species without significant differences are uniformly colored in yellow; biomarkers of differential species are colored following the group; red nodes indicate microbial taxa that play an important role in the red groups, while green nodes indicate microbial taxa that play an important role in the green groups; if a certain group in the plot is missing, it means that there are no species with significant differences in this group, hence this group is absent. The names of the species indicated by letters in the figure are shown in the legend to the right.

File path: 06.Diff_analysis/LEfSe

6.3.4 Metastats

To investigate species with significant differences between groups, Metastats is used to perform hypothesis testing on species abundance data between groups at different taxonomic levels. P-values are obtained

and corrected to q-values. Species with significant differences are filtered based on both p-values and q-values. Box plots of the abundance distribution of differentially abundant species between groups are then generated. The default is to show significantly different species at the phylum level, and if there are none, the next taxonomic level is displayed, and so on. The results are shown in the figure below:

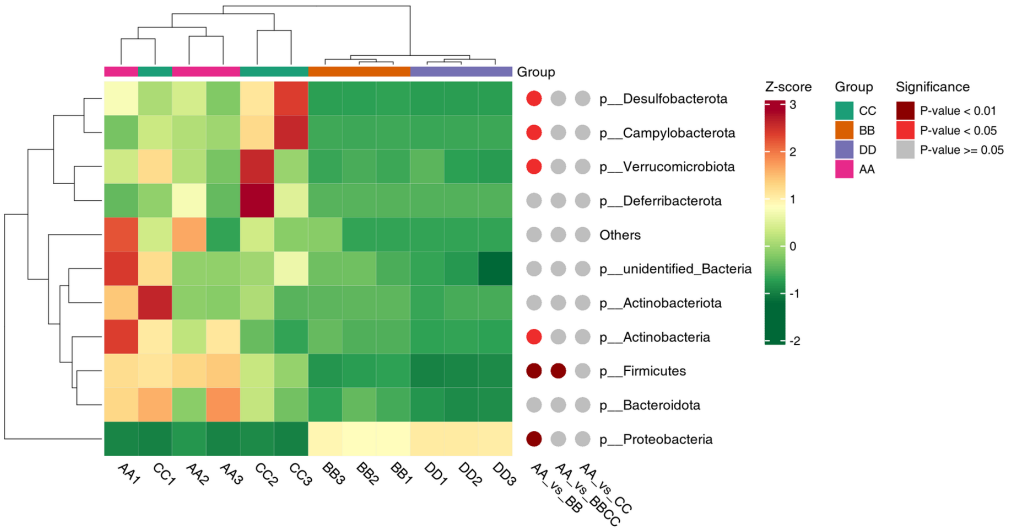


Metastats Significance Difference Statistical Plot Based on ASV for Inter-group Species

Description: In the figure, the horizontal axis shows the sample groups; the vertical axis shows the relative abundance of the corresponding species. The horizontal line represents the two groups with significant differences, while the absence of it indicates that this species does not differ between the two groups. '*' indicates a significant difference between the two groups (P-value < 0.05); '**' indicates a highly significant difference between the two groups (P-value < 0.01).

File path: 06.Diff_analysis/Metastats/phylum

The top 35 microbial taxa, ranked by the sum of quantitative values across all samples, are selected. Metastats differential significance labels are integrated with quantitative data heatmap (shown at the phylum level by default, and if not, the next taxonomic level is displayed). The results are shown in the figure below:



ASV-Based Heatmap of Quantitative Data and Metastats Significance Difference Annotation Plot.

Description: The heatmap is the same as the aforementioned species abundance heatmap, with the different colors on the right indicating the microbes' Metastats significance in the corresponding differential group.

File path: 06.Diff_analysis/Metastats/phylum

6.3.5 metagenomeSeq

Due to the sparsity of microbiome data and differences in sequencing depth between samples, there are limitations to the normalization methods used in LEfSe and Metastats differential analysis. MetagenomeSeq uses Cumulative Sum Scaling (CSS) and a zero-inflated log-normal mixture model to address these issues (Paulson et al. 2013).

LEfSe and Metastats use normalized data, where the total reads for each sample are equal. The results are differences at the phylum, class, order, family, genus, and species levels. In contrast, metagenomeSeq uses the original ASV/OTU abundance data, applies CSS for normalization, and then performs differential analysis between two groups, providing differences at the ASV/OTU level. The metagenomeSeq method is implemented using the R package metagenomeSeq (v1.38.0), and the results of the differential analysis are shown below:

Table 7 Differential analysis results based on ASV using metagenomeSeq

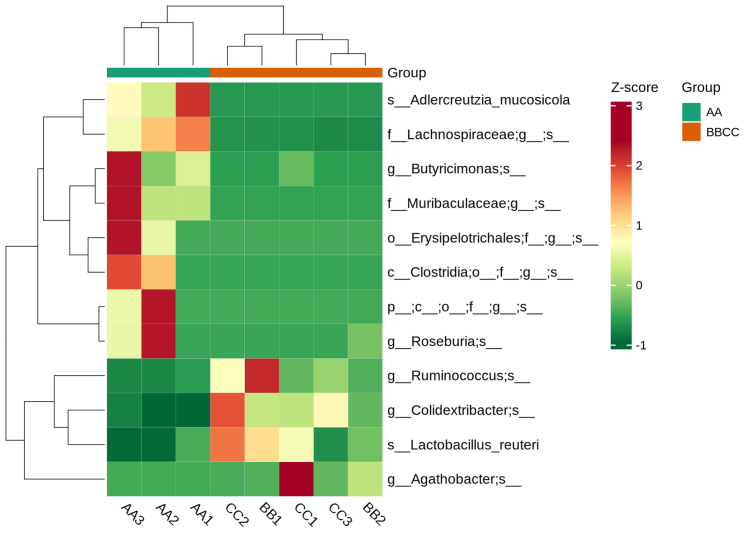
Index	Log2FC(BBCC/AA)	P-value	Adjusted P-value	AA1	AA2
ASV_732	-3.819741	0.0000000	0.0000000	36.446469	52.106430
ASV_1160	-2.357022	0.0000000	0.0000000	10.630220	7.760532
ASV_905	-2.797783	0.0000000	0.0000000	9.111617	27.716186
ASV_1157	-1.912067	0.0000000	0.0000014	3.796507	7.760532
ASV_152	2.363694	0.0000001	0.0000759	5.315110	0.000000
ASV_1215	-1.653191	0.0000004	0.0002182	9.870919	3.325942
ASV_539	3.440166	0.0000011	0.0004524	0.000000	0.000000
ASV_1146	-2.522591	0.0000012	0.0004524	0.000000	21.064302
ASV_1153	-1.723966	0.0000013	0.0004524	1.518603	9.977827
ASV_1483	-1.611203	0.0000030	0.0009352	2.277904	2.217295
ASV_1233	-1.935225	0.0000062	0.0016256	7.593014	3.325942
ASV_61	-2.669186	0.0000063	0.0016256	1873.955960	1463.414634
ASV_1214	-1.359196	0.0000184	0.0044077	1.518603	3.325942
ASV_1322	-1.344628	0.0000242	0.0053770	4.555809	2.217295
ASV_181	2.146247	0.0000281	0.0057828	0.000000	0.000000
ASV_707	-1.531163	0.0000298	0.0057828	0.000000	4.434590
ASV_1152	-1.611454	0.0000627	0.0114567	0.000000	9.977827
ASV_209	1.817841	0.0000799	0.0132557	1.518603	0.000000
ASV_193	1.752962	0.0000811	0.0132557	0.000000	0.000000
ASV_1295	-1.576436	0.0001048	0.0162732	1.518603	2.217295
ASV_177	1.540533	0.0001116	0.0165100	0.000000	0.000000
ASV_1142	-2.791322	0.0001829	0.0258277	24.297646	43.237251
ASV_1264	-1.327718	0.0002417	0.0326462	0.000000	2.217295
ASV_1148	-1.825669	0.0003547	0.0459141	0.000000	19.955654

File path: 06.Diff_analysis/metagenomeSeq

- Index: ASV/OTU ID
- Log2FC(Case/Control): Fold change, Log2FC = Case / Control, i.e., the log2 ratio of the means of the two groups
- P-value: Significance test P-value

- Adjusted P-value: Adjusted P-value
- Sample: CSS-normalized abundance. Note: The mean value of each group is not directly calculated from this abundance value to calculate Log2FC.
- Taxonomy: Taxonomic annotation of ASV/OTU

Abundance information of different ASVs/OTUs after normalization, classified and summed according to taxonomic annotation, and the heatmap is shown below:



Heat map of species for differential results based on ASV

Explanation: Vertical axis represents sample information, horizontal axis represents species classification information, and the clustered tree in the figure is the species clustering tree. The values corresponding to the heatmap are Z-Score standardized relative quantitative data.

File path: 06.Diff_analysis/metagenomeSeq

6.3.6 OPLS-DA

Partial Least Squares Discriminant Analysis (PLS-DA) is a supervised multivariate statistical analysis method that extracts components from the independent variable X and the dependent variable Y and calculates the correlation between components. PLS-DA maximizes differences between groups, facilitating

the identification of different features. Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) combines orthogonal signal correction (OSC) and PLS-DA methods. It decomposes the information in the X matrix into two types: relevant to Y and irrelevant to Y. This is done by removing unrelated differences to filter differential features. OPLS-DA is performed after log2 transformation and centralization of the original data. Here, X represents the matrix of sample quantitative information, and Y represents the matrix of sample grouping information.

Note that PLS-DA and OPLS-DA are common in metabolomics data analysis. Because microbiome data has characteristics such as zero inflation and compositional nature, these two algorithms may not be the optimal choice for microbiome data analysis. Therefore, the results of this method are for reference only.

6.3.6.1 Differential Microbes

Based on the OPLS-DA model (biological replicates >= 3), the Variable Importance in Projection (VIP) is calculated to preliminarily filter out differential features between different groups. The OPLS-DA algorithm is implemented using MetaboAnalystR (v1.0.1) in R (v3.5.1), with a default filtering criterion of VIP >= 1.

Table 8 Important microbes calculated by OPLS-DA

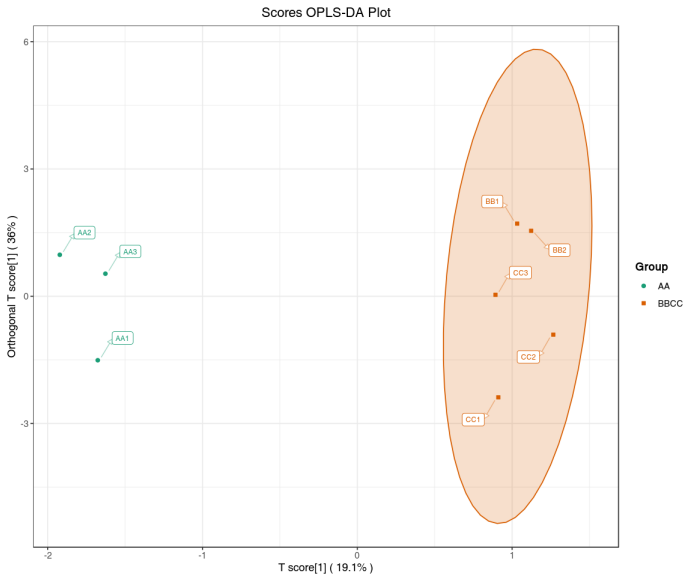
Index	VIP	AA1	AA2	AA3	CC1
k__Bacteria;p__Actinobacteria	1.014046	0.0182737	0.0054896	0.0107303	0.0105201
k__Bacteria;p__Verrucomicrobiota	1.130963	0.0077389	0.0062769	0.0033233	0.0138961
k__Bacteria;p__Campylobacterota	1.420164	0.0405511	0.0857555	0.0710283	0.1059418
k__Bacteria;p__Desulfobacterota	1.122250	0.1488067	0.1136190	0.0527825	0.0823968
k__Bacteria;p__Firmicutes	1.557716	0.5643830	0.5843667	0.6116035	0.5438230

File path: 06.Diff_analysis/OPLS-DA

- First column: Microbe name
- VIP: Variable Importance in Projection

6.3.6.2 OPLS-DA Model Overview

According to the analysis results of the OPLS-DA model, score plots are generated to visualize the differences between various groups.



OPLS-DA score plot

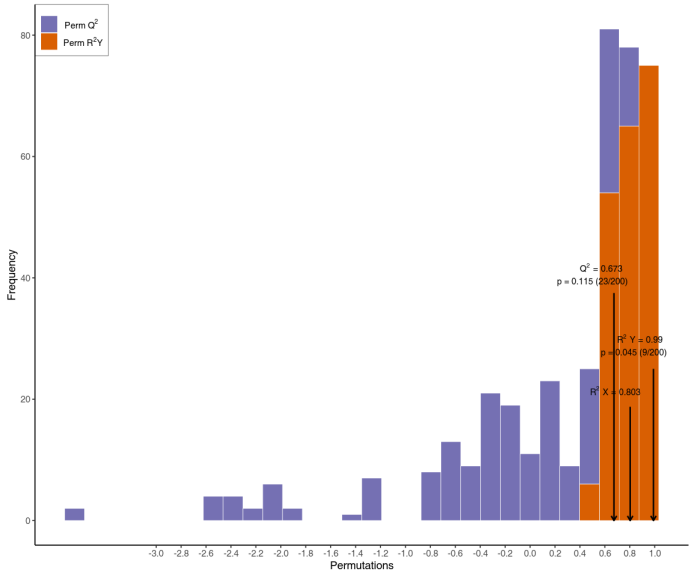
Explanation: The x-axis represents the predicted principal component, and the direction indicates the difference between groups; the y-axis represents the orthogonal principal component, and the direction indicates the difference within groups; the percentage indicates the explanatory rate of the component to the dataset. Each point in the figure represents a sample, and samples from the same group are represented in the same color.

File path: 06.Diff_analysis/OPLS-DA

6.3.6.3 OPLS-DA Model Validation

OPLS-DA evaluates the predictive parameters of the model, including R^2X, R^2Y, and Q^2. R^2X and R^2Y represent the explanatory rates of the model for the X and Y matrices, respectively. Q^2 indicates the predictive ability of the model. The closer these three indicators are to 1, the more stable and reliable the model. A model is considered effective when $Q^2 > 0.5$, and outstanding when $Q^2 > 0.9$. The validation plot below shows R^2Y and Q^2 values on the horizontal axis and the frequency of model classification effects on the vertical axis. The experiment involves 200 random permutations of the data. For example, if Q^2's $p = 0.02$, it means that in this permutation test, there are 4 random grouping models with predictive ability better than the current OPLS-DA model. Similarly, if R^2Y's $p = 0.545$, it indicates that in this permutation test, 109 random grouping

models have explanatory rates better than the current OPLS-DA model. In general, when $p < 0.05$, the model is considered optimal.



OPLS-DA model validation plot

Explanation: The x-axis represents the model R^2Y, Q^2 values, and the y-axis represents the frequency of occurrence of model classification effects in 200 randomly permuted experiments. In the figure, orange represents random grouping model R^2Y, purple represents random grouping model Q^2, and black arrows represent the values of the original model's R^2X, R^2Y, and Q^2 values.

File path: 06.Diff_analysis/OPLS-DA

7 Association Analysis and Model Prediction

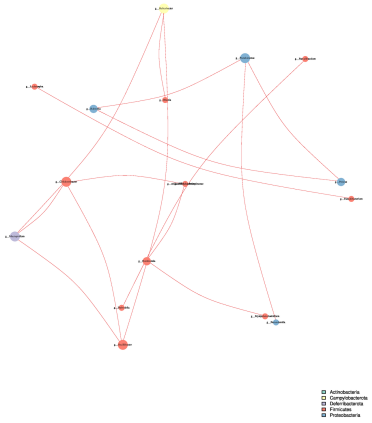
7.1 Network Analysis

The co-occurrence network provides a new perspective for studying the community structure and function of complex microbial environments. Due to the distinct co-occurrence relationships of microorganisms in different environments, the species co-occurrence network allows for a visual understanding of the impact

of different environmental factors on microbial adaptability. It also reveals dominant species and closely interacting species groups in a particular environment. These dominant species and species groups often play unique and crucial roles in maintaining the stability of microbial community structure and function in that environment.

Microbial data analysis is based on relative abundance. When conventional methods such as Pearson and Spearman correlation are used for compositional data analysis, there is a bias in correlation estimation. Since the sum of the scores must be 1, the scores are not independent. Karl Pearson warned in 1897 not to "attempt to explain the correlation between the ratio of the numerator and denominator containing common parts" (Pearson 1997). The SparCC algorithm (Friedman and Alm 2012) can reliably estimate correlations from compositional data. FastSpar (Watts et al. 2019), a C++ language rewrite of SparCC, performs more efficient calculations and supports parallel processing. Species correlation calculation is implemented using FastSpar.

Top 100 microbial genera were selected based on abundance for correlation analysis, with the following filtering conditions: (1) remove connections with an absolute correlation coefficient <= 0.8 (default value), (2) filter out self-connections, and (3) remove connections with a node abundance less than 0.005%. The resulting network graph is shown below:



ASV-based Network Plot

Description: Different nodes represent different genera; node size represents the average relative abundance of that genus; nodes of the same phylum are of the same color (as shown in the legend); the thickness of the connecting lines between nodes is positively correlated with the absolute value of the correlation coefficient of the species interactions; and the color of the connecting lines corresponds to the positive or negative correlation (positive correlation in red, negative correlation in blue).

File path: 03.ASV_visualization/genus_network

The topological parameters of the network graph, such as Network Diameter (ND), Average Degree (AD), Modularity (MD), Clustering Coefficient (CC), Graph Density (GD), and Average Path Length (APL), are displayed as follows:

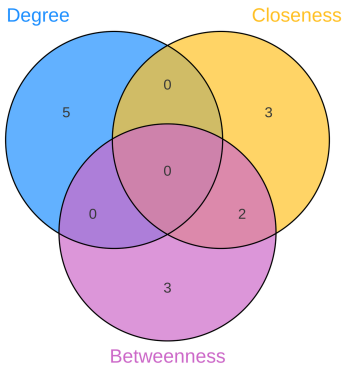
Table 9 ASV-based Network Parameters

ND(Network diameter)	MD(modularity)	CC(Clustering coefficient)	GD(graph density)	AD(Average degree)	APL(average.path.length)
4	0.6844444	0.5	0.1102941	1.764706	1.576923

File path: 03.ASV_visualization/genus_network

- ND (Network Diameter): The maximum measurement length of the network graph, i.e., the maximum of the shortest distances between any two points, among these shortest distances.
- AD (Average Degree): The average degree, i.e., the number of edges connected to a node. The average degree is the sum of all node degrees divided by the total number of nodes.
- MD (Modularity): The modularity measures the modularity of the network community structure. It describes the rationality of dividing the network into different modules or the distinctiveness between different modules.
- CC (Clustering Coefficient): The clustering coefficient represents the likelihood of a node's adjacent nodes being connected. The connectivity of the network graph is the average value of all node connectivities.
- GD (Graph Density): The graph density is the actual number of edges divided by the total possible number of edges.
- APL (Average Path Length/Mean Distance): The average path length is the sum of the shortest distances between all pairs of nodes divided by the number of node pairs.

Network graphs can be used to infer keystone species, defined as species whose absence would cause significant changes in the entire network. The selection criteria for keystone species include high average degree, high closeness centrality, and low betweenness centrality (Banerjee, Schlaeppi, and Heijden 2018).

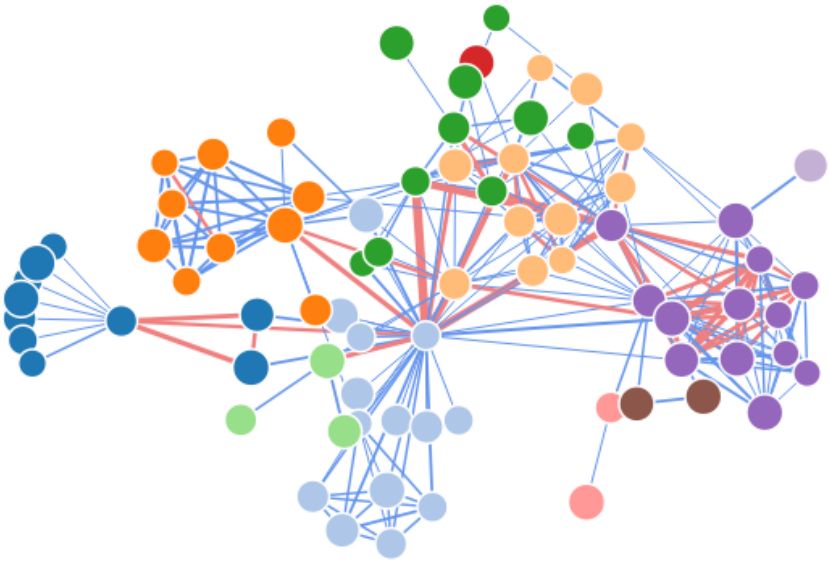


ASV-based Venn Diagram for Network Key Species

The Venn diagram of key species obtained by different screening criteria. The degree indicates the number of other nodes a node is connected to; Betweenness centrality: a node has high betweenness centrality if it is often located on the shortest paths between other nodes; Closeness centrality, a node has high closeness centrality if the shortest distances from this node to all other nodes are small. Closeness centrality is closer to geometrically centered positions than betweenness centrality.

File path: 03.ASV_visualization/genus_network

Dynamic network graphs provide a clearer view of the associations between a microbial genus and other microorganisms. An example is shown in the following figure:



Dynamic Network Graph Demonstration

Description: different nodes represent different genera, node size represents the degree of connectivity of the genus; the same color represents the same phylum level (as shown in the legend); the thickness of the connecting lines between the nodes is positively correlated with the absolute value of the correlation coefficient of the species interactions.

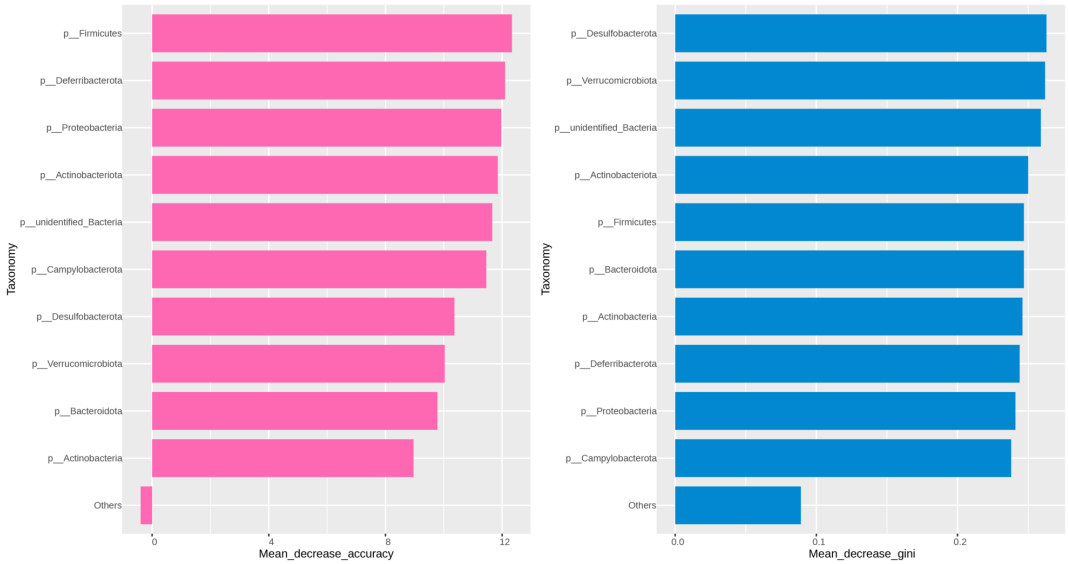
File path: 03.ASV_visualization/genus_network

7.2 RandomForest Analysis

Random Forest belongs to the ensemble type of machine learning algorithms. Random Forest uses the bootstrap aggregating resampling method to extract multiple samples with replacement from the original samples as the training set. It models decision trees on the training set and then combines the predictions of multiple decision trees through voting to obtain the final prediction. RF has a high prediction accuracy, good tolerance for outliers and noise, and is less prone to overfitting.

When extracting the training set using the bootstrap aggregating method, the probability that a sample in the original data is not selected is $(1–1/N)^N$, where N is the number of samples in the original data. When N is large enough, $(1–1/N)^N$ will converge to $1/e \approx 0.368$. This implies that close to 37% of the samples in the original data will not appear in the extracted training set, and these data are referred to as Out Of Bag

(OOB) data. OOB data will be used to estimate the model's performance. Random Forest is implemented using the R package varSelRF.

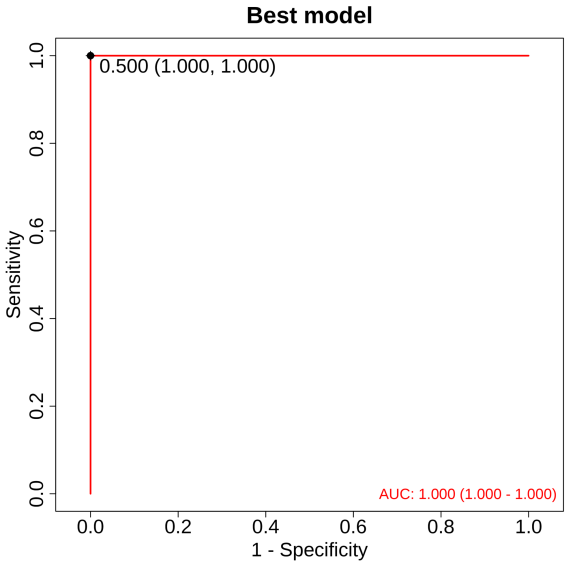


Importance Ranking Plot of Variables Based on ASV

Description: MeanDecreaseAccuracy measures how much the accuracy of a random forest prediction is reduced by changing the values of a variable to a random number. A larger value indicates a greater importance of the variable. MeanDecreaseGini compares the importance of variables by calculating the effect of each variable on the heterogeneity of observations at each node of the taxonomic tree through the Gini index. The larger this value indicates the greater importance of the variable. a) Horizontal coordinate: mean decrease in accuracy; vertical coordinate: top 50 important species; b) Horizontal coordinate: mean decrease in Gini index; vertical coordinate: top 50 important species.

File path: 06.Diff_analysis/random_forest/phylum

Based on the best model selected by the Random Forest method, ROC curves are plotted as shown in the figure below:



Optimal Model ROC Curve Based on ASV

Description: ROC curve for the training set, with the horizontal coordinate being 1 - Specificity, which indicates the rate of false positives, and the vertical coordinate being Sensitivity, which indicates the rate of true positives. Specificity = true negatives/(true negatives + false positives), Sensitivity = true positives/(true positives + false negatives). The ROC curve is shown in red. The dots on the curve indicate the optimal thresholds, and the three numbers indicate the optimal threshold, specificity, and sensitivity values, respectively. The AUC value is shown in the lower right corner of the graph, with the 95% confidence interval in parentheses.

File path: 06.Diff_analysis/random_forest/phylum

- Taking the phylum as an example, microbial importance ranking,
  - Sorted by mean decrease in accuracy:

    06.Diff_analysis / random_forest / phylum / *.mean_decrease_accuracy.pdf / png
  - Sorted by the average decrease in Gini coefficient:

    06.Diff_analysis / random_forest / phylum / *.mean_decrease_gini.pdf/png
  - Combined plot:

06.Diff_analysis / random_forest / phylum / *.accurary_gini.combine.pdf/png

- ROC curves of the best model selected by Random Forest: 06.Diff_analysis / random_forest / phylum / *.best_model.ROC.pdf
- ROC curves of individual microorganisms: single_variable_ROC.*

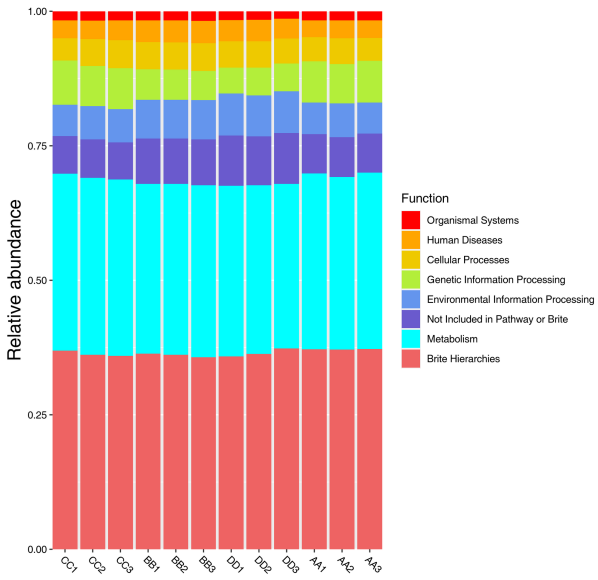
8 Functional Predictions

8.1 PICRUSt2

PICRUSt, which stands for Phylogenetic Investigation of Communities by Reconstruction of Unobserved States, is a bioinformatics software package for predicting metagenome functional content based on marker genes such as 16S rRNA. PICRUSt2(Douglas et al. 2020) is an extension of PICRUSt. Detailed prediction procedures can be found in the PICRUSt2 tutorial. Currently, it supports functional predictions based on 16S sequencing data using the KEGG database.

8.1.1 Display of Functional Annotation Relative Abundance

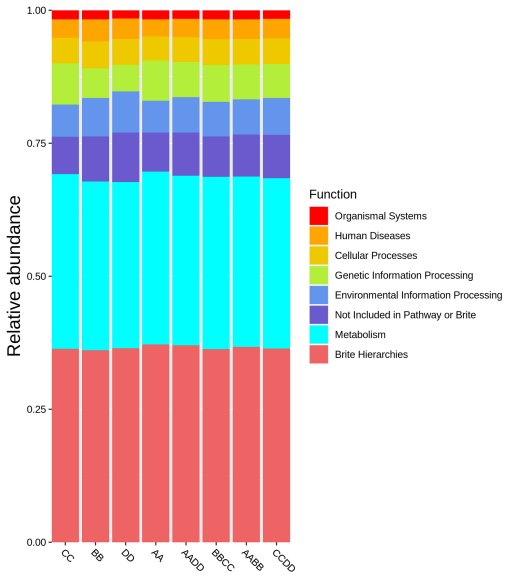
Based on the annotated database results, the top 10 functional categories with the highest abundance for each sample or group at each annotation level are selected. This information is used to generate a bar plot of the relative abundance, allowing an intuitive view of the functions with higher relative abundance and their proportions in different samples or groups. An example of the Level 1 relative abundance bar plot is shown below:



ASV-Based Relative Abundance Bar Chart of PICRUSt2 Functional Annotations for Each Sample

Description: The horizontal coordinates are sample names; the vertical coordinates indicate relative abundance.

File path: 07.Function_prediction/PICRUSt/03.top10_barplot



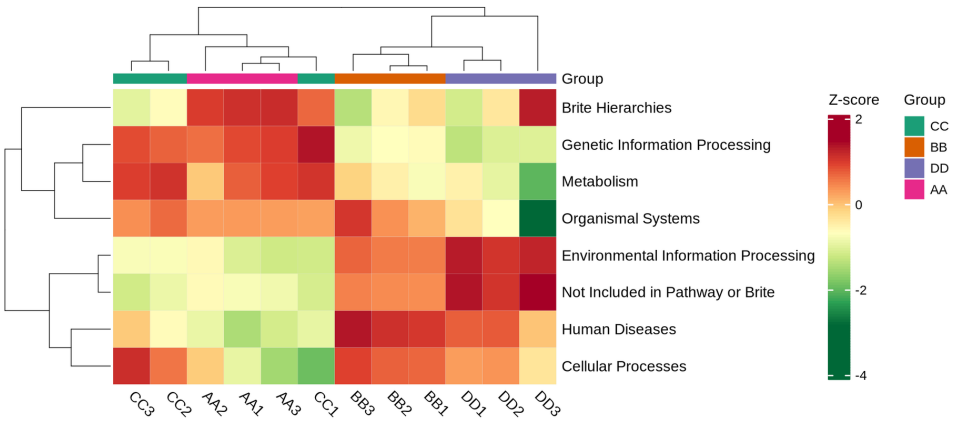
ASV-Based Relative Abundance Bar Chart of PICRUSt2 Functional Annotations for Each Group

Description: The horizontal coordinates are group names; the vertical coordinates indicate relative abundance.

File path: 07.Function_prediction/PICRUSt/03.top10_barplot/group

8.1.2 Clustering Analysis of Functional Relative Abundance

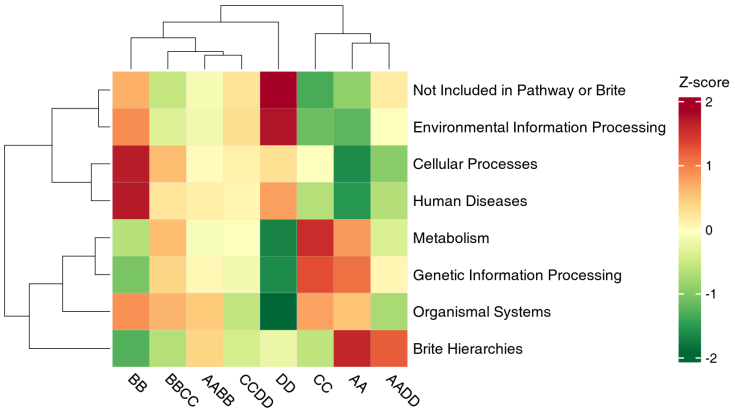
Based on the sum of the abundance of database-annotated functions across all samples, the top 35 functions are selected, and a heatmap is generated. This heatmap includes information on the abundance of these functions in each sample and performs clustering from the perspective of functional differences. Here, we show an example of a Level 1 hierarchical clustering heatmap:



ASV-Based Clustering Heatmap of PICRUSt2 Functional Annotations for Each Sample

Description: The horizontal axis represents functions, the vertical axis represents samples, and the cells represent relative abundance. The redder the color, the higher the relative abundance, while the bluer the color, the lower the relative abundance. Clustering is also made for functions and samples.

File path: 07.Function_prediction/PICRUSt/04.cluster_heatmap



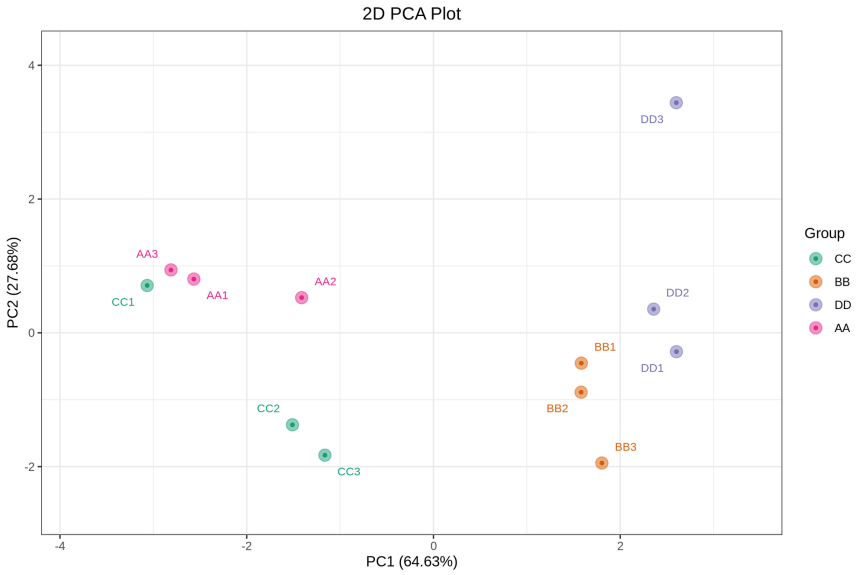
ASV-Based Clustering Heatmap of PICRUSt2 Functional Annotations for Each Group

Description: The horizontal axis represents functions, the vertical axis represents groups, and the cells represent relative abundance. The redder the color, the higher the relative abundance, while the bluer the color, the lower the relative abundance. Clustering is also made for functions and samples.

File path: 07.Function_prediction/PICRUSt/04.cluster_heatmap/group

8.1.3 Functional Annotation PCA Analysis

PCA dimensionality reduction analysis is conducted based on the abundance statistics of functionally annotated data. In this analysis, samples with more similar functional compositions exhibit closer distances in the reduced-dimensional plot:



Presentation of ASV-based PCA Results for PICRUSt2 Functional Annotations

Description: the horizontal coordinate indicates the first principal component, while the percentage indicates the contribution of the first principal component to the sample variance; the vertical coordinate indicates the second principal component, while the percentage indicates the contribution of the second principal component to the sample variance; each dot in the plot indicates a sample, with samples in the same group represented by the same color; groups of more than 3 samples are allowed to insert ellipses to indicate confidence intervals, with the same color as the group.

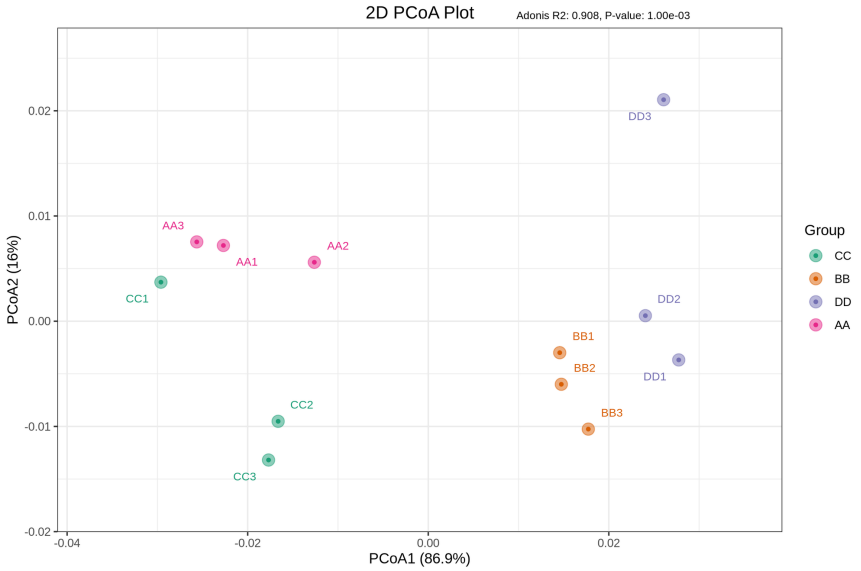
File path: 07.Function_prediction/PICRUSt/05.PCA/Level1

8.1.4 PCoA Analysis

Principal Coordinates Analysis (PCoA) involves extracting the most significant elements and structures from multidimensional data through a series of eigenvalue and eigenvector rankings. We conduct PCoA analysis based on Bray-Curtis distance, selecting the primary coordinate combinations with the highest contribution rates for plotting. In this analysis, samples with closer distances indicate more similar species composition structures, leading to the clustering of samples with similar community structures, while samples with significant community differences are distinctly separated.

Using the normalized abundance table of KEGG at each hierarchical level, we calculate the Bray-Curtis distance matrix between samples and perform PCoA analysis on these distance matrices. The results are as follows:

PCoA results for KEGG level 1 are displayed below:

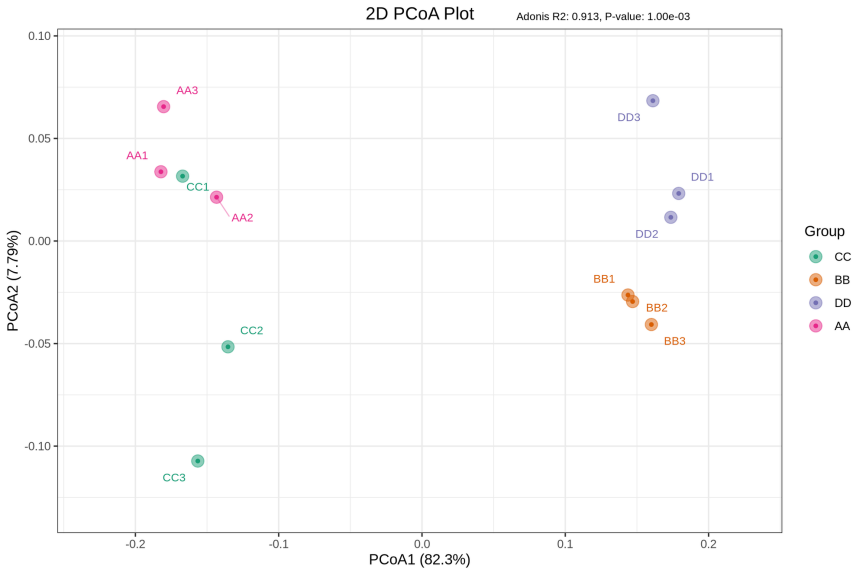


PcoA results display of PICRUSt2 functional annotation based on ASV

Explanation: The x-axis represents one principal component, the y-axis represents another principal component, and the percentage represents the contribution of the principal component to sample differences; each point in the figure represents a sample, and samples from the same group are represented in the same color.

File path: 07.Function_prediction/PICRUSt/06.PCoA/Level1

PCoA results for EC (Enzyme Commission) are displayed below:



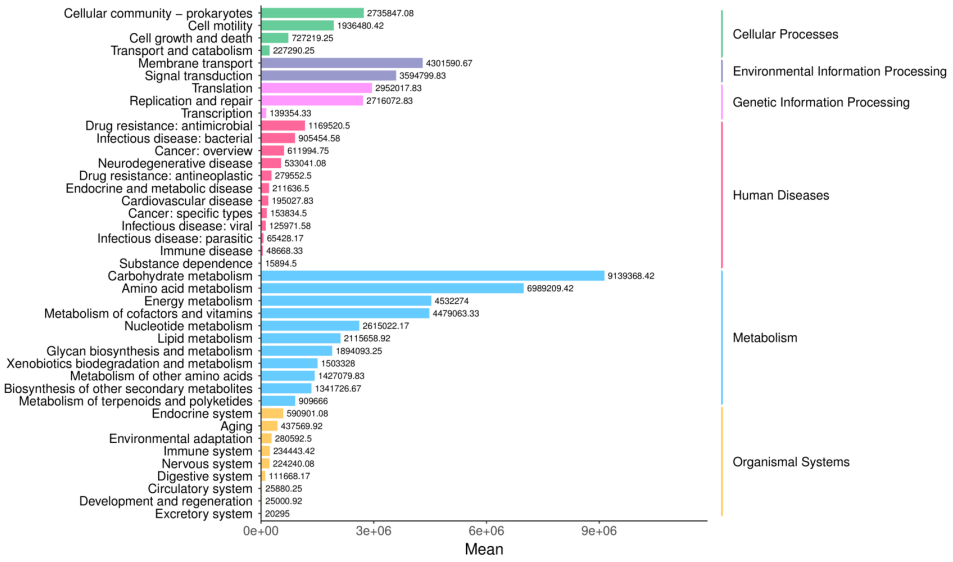
PcoA results display of PICRUSt2 EC annotation based on ASV

Explanation: The x-axis represents one principal component, the y-axis represents another principal component, and the percentage represents the contribution of the principal component to sample differences; each point in the figure represents a sample, and samples from the same group are represented in the same color.

File path: 07.Function_prediction/PICRUSt/06.PCoA/EC

8.1.5 Metabolic Pathway Statistics

The average abundance statistics for the secondary classification of the KEGG database are as follows:



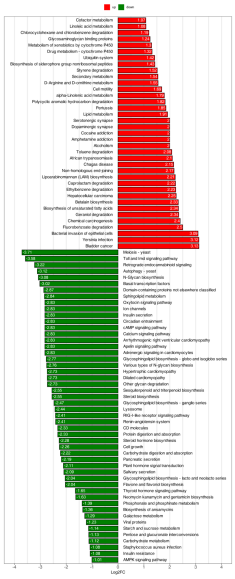
Bar chart of second-level classification abundance based on ASV in PICRUSt2

Explanation: The y-axis represents KEGG level 2 functional classification, the x-axis represents the average abundance of functional classification in all samples, and the right side indicates the corresponding level 1 classification.

File path: 07.Function_prediction/PICRUSt/09.pathway_stat

8.1.6 Differential Analysis of Metabolic Pathways

Differential metabolic pathway analysis is conducted using the metagenomeSeq method, implemented with the R package metagenomeSeq (v1.38.0). The bar plot showing differential pathways is displayed below:



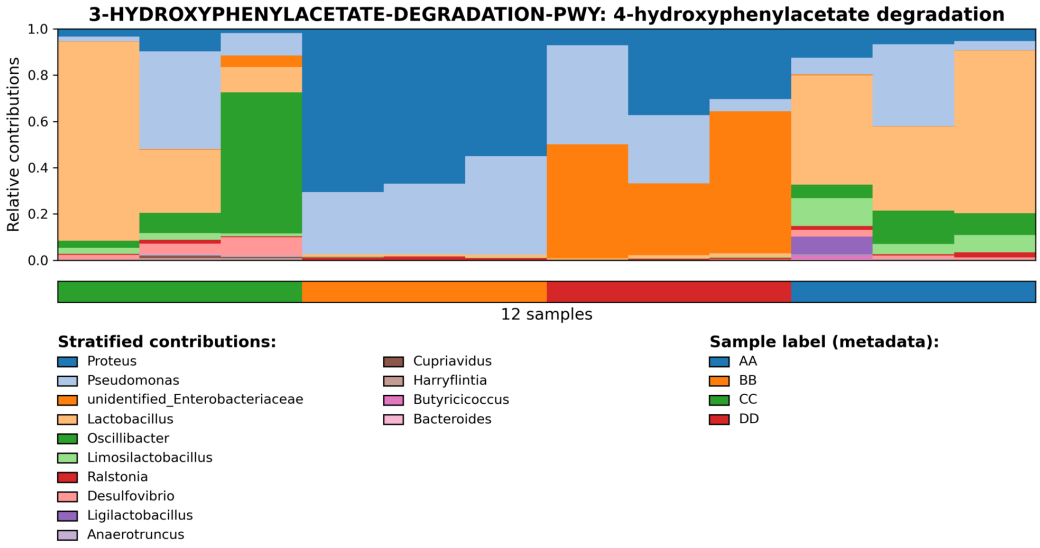
Bar chart of differential pathways based on ASV in PICRUSt2

Explanation: The y-axis represents KEGG pathway names, the x-axis represents the log2FC (Fold Change) of the pathway between 2 groups, and the log2FC of each pathway is displayed within the bar.

File path: 07.Function_prediction/PICRUSt/10.pathway_diff

8.1.7 Species Contribution to Pathways

The MetaCyc database contains various pathways, metabolites, biochemical reactions, enzymes, and genes involved in primary and secondary metabolism. It aims to classify the metabolic processes of all life by storing experimentally validated representative metabolic pathways. The contribution of species predicted by PICRUSt2 to MetaCyc pathways is visualized using tools provided with the humann program:



Bar chart of species contribution to MetaCyc pathways based on ASV in PICRUSt2

Explanation: The x-axis represents samples, with groups represented in different colors, the y-axis represents the relative abundance of metabolic pathways, and the graph represents the relative contribution of genus-level species to metabolic pathways.

File path: 07.Function_prediction/PICRUSt/11.pathway_taxon_contrib

8.2 Tax4Fun2

Tax4Fun2(Wemheuer et al. 2018) is an upgraded version of Tax4Fun, capable of rapidly predicting the functional profiles and functional redundancy of prokaryotes based on 16S rRNA gene sequences. By integrating user-defined, habitat-specific genomic information, the accuracy and robustness of predicting functional profiles can be significantly improved. Compared to the old version of Tax4Fun, Tax4Fun2 has the following advantages:

1) No longer limited to OTU/ASV abundance tables annotated with specific versions of SILVA. It allows direct input of OTU/ASV representative sequences and species annotation through alignment with a specified reference database. In addition to the pre-built reference sets provided by Tax4Fun2 (significantly expanded compared to before), it also allows us to provide custom reference sets, providing great flexibility.

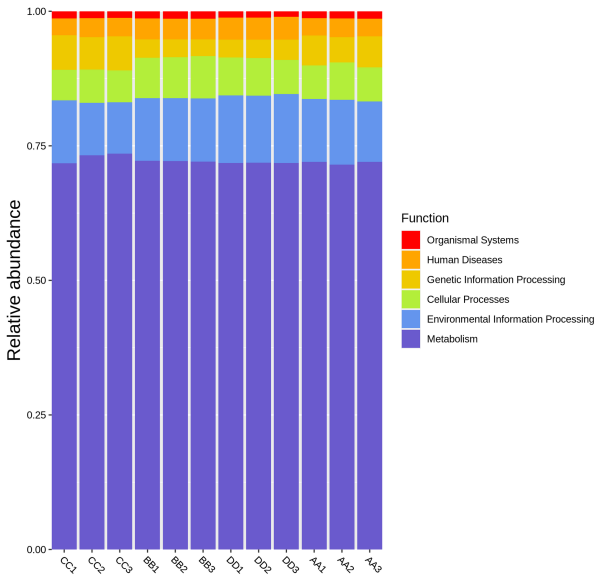
2) Focuses on prokaryotic data but can also incorporate eukaryotic data.
3) Provides methods for calculating specific functional redundancy, which is crucial for predicting the likelihood of losing specific functions during environmental disturbances.
4) Significant improvements in accuracy and stability.
5) Tax4Fun2 is currently in a continuous updating state.

Workflow of Tax4Fun2:

1) First, align the 16S rRNA gene sequences with reference sequences (which can be pre-built or user-defined) to identify the nearest neighbors.
2) Based on the results of the nearest neighbor search, summarize the OTU/ASV abundances for each sample.
3) An association matrix (AM) contains the reference functional profiles identified in the 16S rRNA search.
4) Integrate OTU/ASV abundances with the AM functional profiles to predict the metagenome of each sample.

8.2.1 Display of Functional Annotation Relative Abundance

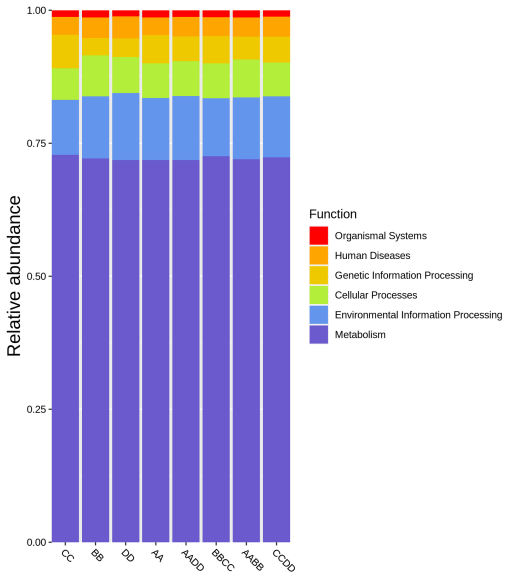
Based on the annotated results from the database, the top 10 functional categories with the highest relative abundance on each sample or group at each annotation level were selected. This generated a bar plot of functional relative abundance, allowing for a visual examination of the functions with higher relative abundance in different annotation levels for each sample. An example of a Level 1 relative abundance bar plot is shown below:



ASV-Based Relative Abundance Bar Chart of Tax4Fun2 Functional Annotations for Each Sample

Description: The horizontal coordinates are sample names; the vertical coordinates indicate relative abundance.

File path: 07.Function_prediction/Tax4Fun2/2.top10_barplot



ASV-Based Relative Abundance Bar Chart of Tax4Fun2 Functional Annotations for Each Group

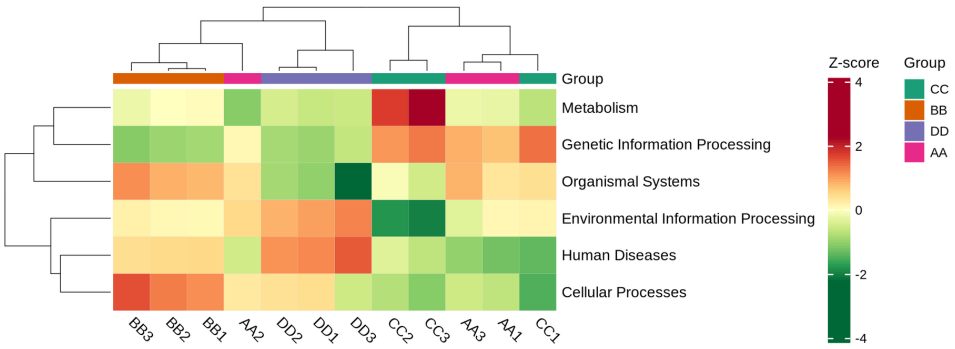
Description: The horizontal coordinates are group names; the vertical coordinates indicate relative abundance.

File path: 07.Function_prediction/Tax4Fun2/2.top10_barplot/group

Based on the annotated results from the database, T-test differential analysis was also conducted. Specific results can be found in the delivered documents.

8.2.2 Clustering Analysis of Functional Relative Abundance

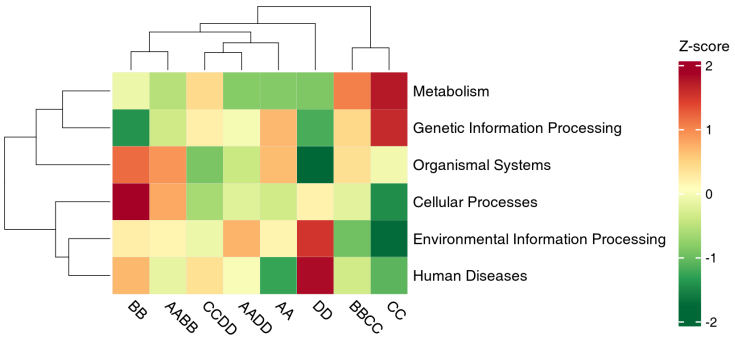
Based on the sum of the abundances of database-annotated functions in all samples, the top 35 functions were selected, and their abundance information in each sample was used to create a heatmap. Clustering was performed from the perspective of functional differences (only Level 1 hierarchical clustering heatmap is shown here):



ASV-Based Clustering Heatmap of Tax4Fun2 Functional Annotations for Each Sample

Description: The horizontal axis represents functions, the vertical axis represents samples, and the cells represent relative abundance. The redder the color, the higher the relative abundance, while the bluer the color, the lower the relative abundance. Clustering is also made for functions and samples.

File path: 07.Function_prediction/Tax4Fun2/3.cluster_heatmap



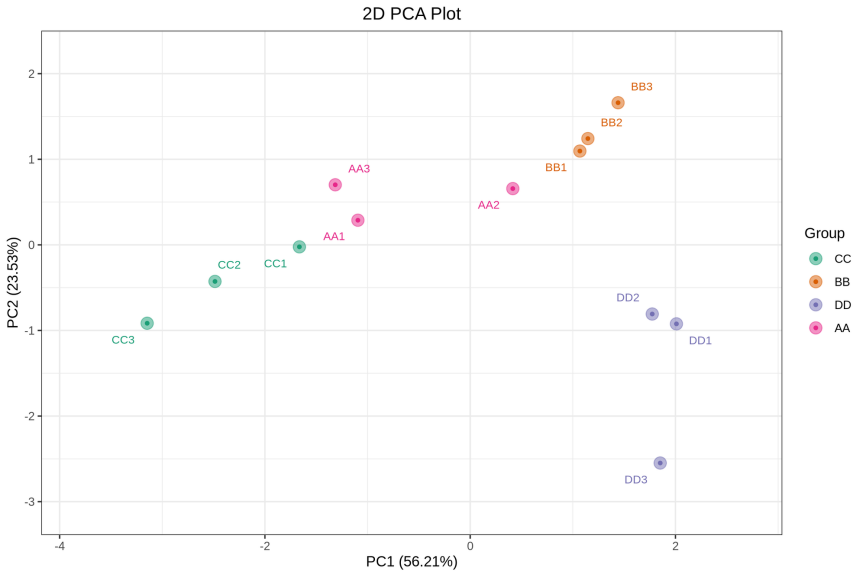
ASV-Based Clustering Heatmap of Tax4Fun2 Functional Annotations

Description: The horizontal axis represents functions, the vertical axis represents groups, and the cells represent relative abundance. The redder the color, the higher the relative abundance, while the bluer the color, the lower the relative abundance. Clustering is also made for functions and samples.

File path: 07.Function_prediction/Tax4Fun2/3.cluster_heatmap/group

8.2.3 Functional Annotation PCA Analysis

PCA dimensionality reduction analysis is performed based on the abundance statistics results of functional annotations from the database. If the functional composition of samples is more similar, the distances between them in the dimensionality reduction plot are closer:



Presentation of ASV-based PCA Results for Tax4Fun2 Functional Annotations

Description: the horizontal coordinate indicates the first principal component, while the percentage indicates the contribution of the first principal component to the sample variance; the vertical coordinate indicates the second principal component, while the percentage indicates the contribution of the second principal component to the sample variance; each dot in the plot indicates a sample, with samples in the same group represented by the same color; groups of more than 3 samples are allowed to insert ellipses to indicate confidence intervals, with the same color as the group.

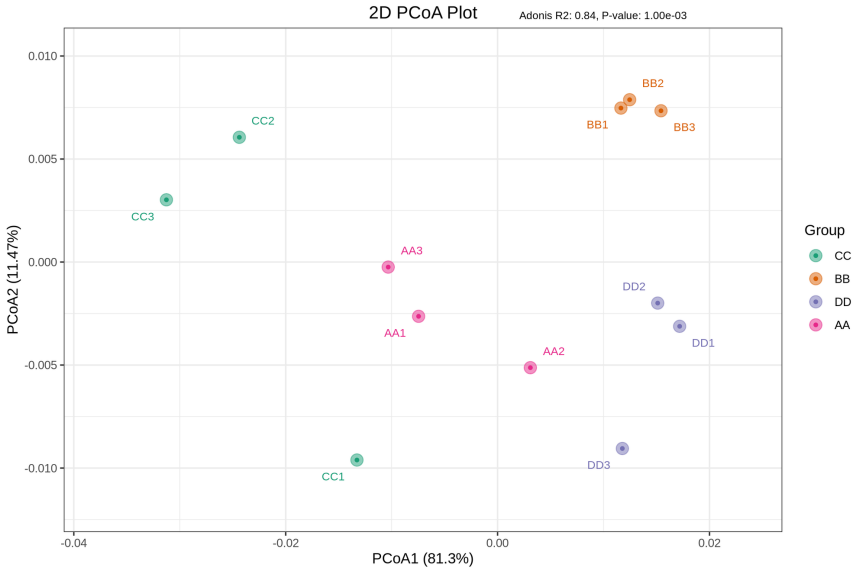
File path: 07.Function_prediction/Tax4Fun2/4.PCA/Level1

8.2.4 PCoA Analysis

Principal Coordinates Analysis (PCoA) is a method that extracts the most important elements and structures from multidimensional data through a series of sorted eigenvalues and eigenvectors. We conduct PCoA analysis based on Bray-Curtis distance and select the main coordinate combinations with the highest contribution rates for plotting. If samples are closer in distance, it indicates a more similar species composition structure. Therefore, samples with high similarity in community structure tend to cluster together, while samples with significant community differences are separated far apart.

Using the normalized abundance table for various levels of KEGG, we calculate the Bray-Curtis distance matrix between samples and perform PCoA analysis on these distance matrices. The results are as follows:

The PCoA results for KEGG level1 are shown below:



PcoA results display of Tax4Fun2 functional annotation based on ASV

Explanation: The x-axis represents one principal component, the y-axis represents another principal component, and the percentage represents the contribution of the principal component to sample differences; each point in the figure represents a sample, and samples from the same group are represented in the same color.

File path: 07.Function_prediction/Tax4Fun2/5.PCoA/Level1

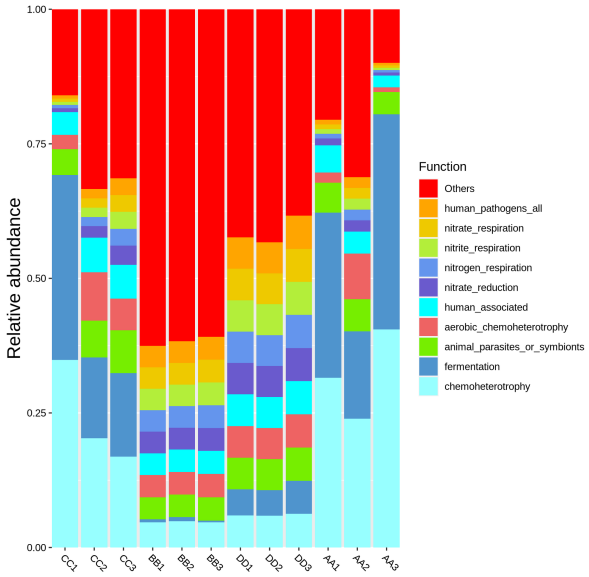
8.3 FAPROTAX

FAPROTAX (Functional Annotation of Prokaryotic Taxa) (Louca, Parfrey, and Doebeli 2016) is a manually curated database based on published literature. It includes the correspondences between taxonomic classifications (genus or species) of prokaryotic microorganisms and their associated metabolic or ecological functions. The database covers over 80 functional categories, including carbon, nitrogen, phosphorus, sulfur cycling, animal and plant pathogens, methane production, fermentation, and more. With over 7600 functional annotations, it spans more than 4600 prokaryotic species, making it suitable for functional annotation analysis of biochemical cycling processes in environmental samples.

FAPROTAX quantifies microbial abundance information into quantitative information for functional categories based on microbial taxonomic information identified in the samples and their functional annotation information in the database.

8.3.1 Functional Annotation Relative Abundance Display

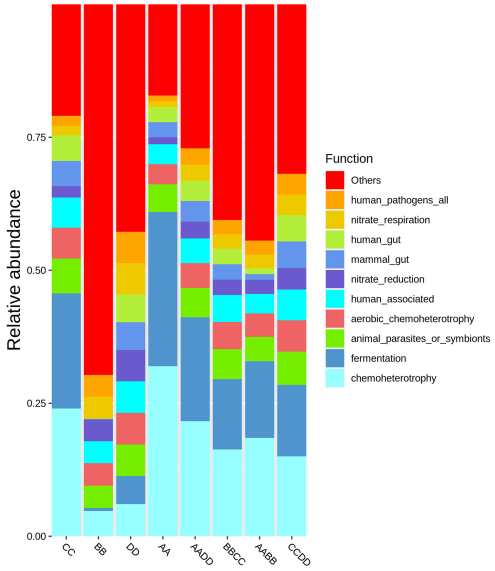
Based on the database annotation results, the abundance information for each sample and group is summarized. The results are displayed as follows:



ASV-Based Relative Abundance Bar Chart of FAPROTAX Functional Annotations for Each Sample

Description: The horizontal coordinates are sample names; the vertical coordinates indicate relative abundance.

File path: 07.Function_prediction/FAPROTAX/3.top10_barplot



ASV-Based Relative Abundance Bar Chart of FAPROTAX Functional Annotations for Each Group

Description: The horizontal coordinates are group names; the vertical coordinates indicate relative abundance.

File path: 07.Function_prediction/FAPROTAX/3.top10_barplot

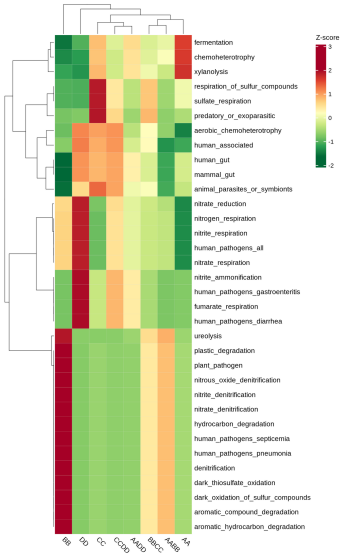
8.3.2 Functional Annotation Relative Abundance Clustering Analysis

A heatmap is generated based on the functional annotation and abundance information of samples in the database, and clustering is performed from the perspective of functional differences.

ASV-Based Relative Abundance Bar Chart of FAPROTAX Functional Annotations for Each Sample

Description: The horizontal axis represents functions, the vertical axis represents samples, and the cells represent relative abundance. The redder the color, the higher the relative abundance, while the greener the color, the lower the relative abundance. Clustering is also made for functions and samples.

File path: 07.Function_prediction/FAPROTAX/4.heatmap



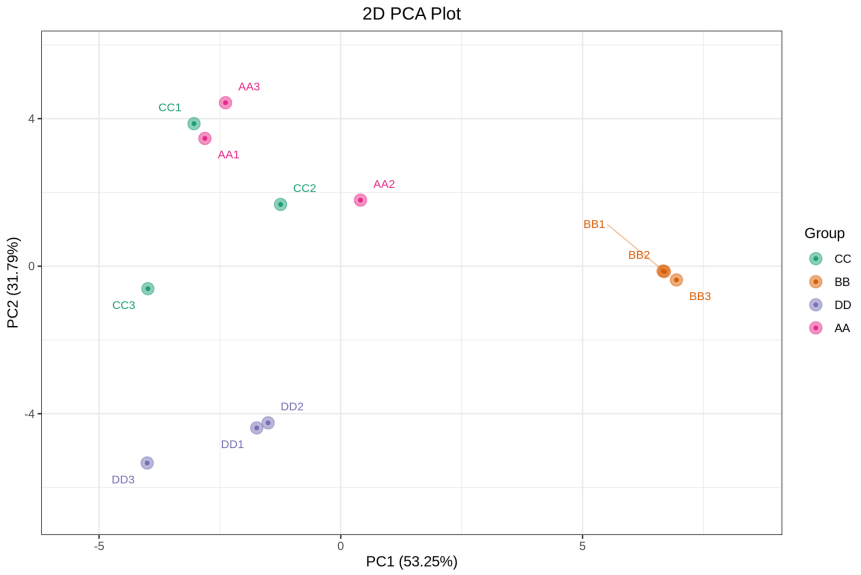
ASV-Based Relative Abundance Heatmap of FAPROTAX Functional Annotations for Each Group

Description: The horizontal axis represents functions, the vertical axis represents groups, and the cells represent relative abundance. The redder the color, the higher the relative abundance, while the greener the color, the lower the relative abundance. Clustering is also made for functions and samples.

File path: 07.Function_prediction/FAPROTAX/4.heatmap

8.3.3 Functional Annotation PCA Analysis

PCA dimensional analysis is performed on the abundance statistical results of functional annotation based on the database. If the functional compositions of samples are more similar, their distances in the dimensional plot are closer:



Presentation of ASV-based PCA Results for FAPROTAX Functional Annotations

Description: the horizontal coordinate indicates the first principal component, while the percentage indicates the contribution of the first principal component to the sample variance; the vertical coordinate indicates the second principal component, while the percentage indicates the contribution of the second principal component to the sample variance; each dot in the plot indicates a sample, with samples in the same group represented by the same color; groups of more than 3 samples are allowed to insert ellipses to indicate confidence intervals, with the same color as the group.

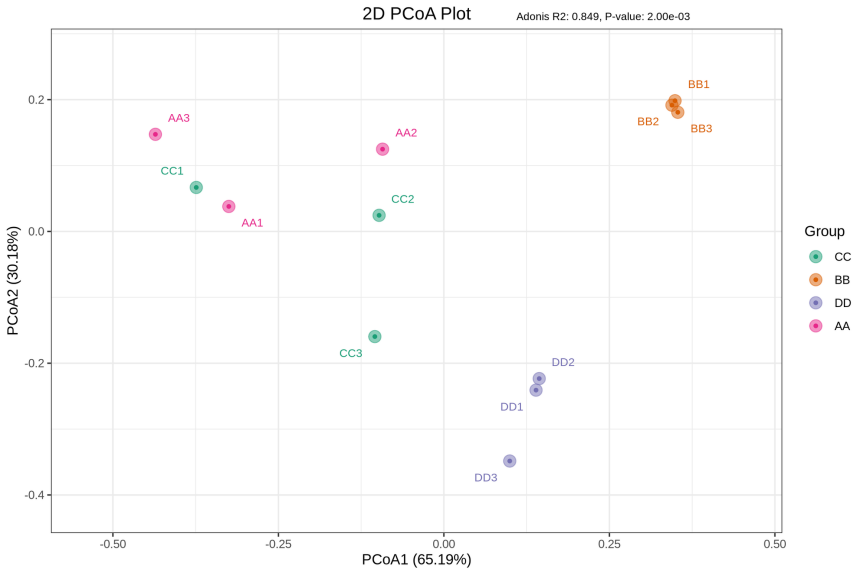
File path: 07.Function_prediction/FAPROTAX/5.PCA

8.3.4 Functional Annotation PCoA Analysis

Principal Coordinates Analysis (PCoA) is a method that extracts the most significant elements and structures from multidimensional data through the sorting of eigenvalues and eigenvectors. We conducted PCoA analysis based on Bray-Curtis distance and selected the main coordinate combinations with the highest contribution rate for plotting. If the samples are closer in distance, it indicates a more similar species composition structure. Therefore, samples with high similarity in community structure tend to cluster together, while samples with significant differences in community structure are separated by a considerable distance.

Using the normalized abundance table of functional categories, we calculated the Bray-Curtis distance matrix between samples and performed PCoA analysis on these distance matrices. The results are as follows:

The PCoA results for functional categories are displayed below:



PcoA results display of FAPROTAX functional annotation based on ASV

Explanation: The x-axis represents one principal component, the y-axis represents another principal component, and the percentage represents the contribution of the principal component to sample differences; each point in the figure represents a sample, and samples from the same group are represented in the same color.

File path: 07.Function_prediction/FAPROTAX/6.PCoA

9 Environmental Factor Correlation Analysis

9.1 Evaluation of Environmental Factors

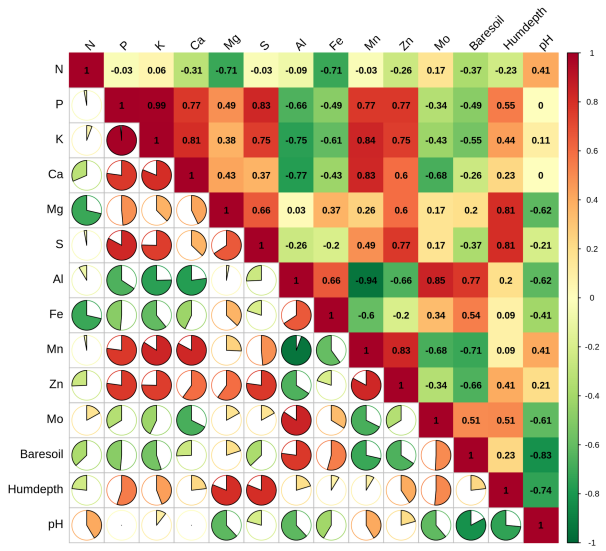
Before conducting correlation analysis with microbial species abundance, it is essential to assess the quality of environmental factor information. The project evaluates environmental factors from three aspects: the correlation between environmental factors, adonis analysis to test whether there are significant differences

between groups, and NMDS and envfit analysis to assess the significance of the impact of environmental factors on the community.

9.1.1 Correlation Between Environmental Factors

Correlation analysis is a statistical method for studying whether there is a dependency relationship between phenomena. It explores the direction and degree of dependence between specific phenomena and is a statistical method for studying the correlation between random variables.

There may be a correlation between environmental factors, indicating that different environmental factors change in a consistent trend. Therefore, it is necessary to calculate the correlation between environmental factors and visualize the results. Later in the text, the selection of environmental factors will eliminate collinear factors. The correlation results are as follows:



Correlation analysis between environmental factors

Explanation: Both horizontal and vertical lines represent environmental factors, the upper right triangular area shows the correlation coefficients in numerical form, the lower left triangular area shows the correlation coefficients using sector size to indicate magnitude (larger area indicates larger absolute value), and the right side is the color legend (red indicates positive correlation, green indicates negative correlation).

File path: 08.Environment_factor/environment_evaluation

9.1.2 Between-Group Adonis Analysis

To assess whether there are significant differences between groups, the project utilizes the previously introduced adonis analysis. The difference lies in the use of Euclidean distance for calculating the distance between environmental factors. The results are as follows:

Table 10 Inter-group adonis analysis of environmental factors

Group	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
CC_vs_BB	1(4)	34349.29(121442.673)	34349.29(30360.66825)	1.131	0.22(0.78)	0.4

File path: 08.Environment_factor/environment_evaluation

- Group: Grouping
- Df: Degrees of freedom. Values corresponding to residual terms are in parentheses, and the same applies below.
- SumsOfSqs: Total variance, also known as the sum of squares
- MeanSqs: Mean square (difference), i.e., SumsOfSqs/Df
- F.Model: F-test value
- R2: Indicates the explanatory power of different groups for sample differences, i.e., the ratio of group variance to total variance. A higher R2 indicates a higher explanatory power of groups for differences.
- Pr(>F): Represents the P-value, where values less than 0.05 indicate high confidence in the test.

9.1.3 Impact of Environmental Factors on the Community

The NMDS method, introduced earlier, can assess inter-group and intra-group differences. Combined with environmental factors, it allows the calculation of the significance of environmental factors' impact on the community.

Table 11 envfit analysis of NMDS based on ASV and environmental factors

Environment	NMDS1	NMDS2	R2	P-value
N	-0.6876526	0.7260398	0.8528714	0.0541667
P	-0.2840392	-0.9588127	0.0661035	0.8902778
K	-0.4444725	-0.8957925	0.0766932	0.8722222
Ca	0.2144874	0.9767268	0.2390917	0.7625000
Mg	0.5581300	0.8297535	0.5673265	0.2597222
S	0.5396877	0.8418653	0.1105540	0.8611111
Al	0.7108312	-0.7033626	0.1754274	0.7791667
Fe	0.7759130	-0.6308399	0.3053610	0.5986111
Mn	-0.4960844	-0.8682743	0.0249433	0.9597222
Zn	0.8813976	-0.4723751	0.0695287	0.8833333
Mo	0.1166998	0.9931672	0.1830592	0.7333333
Baresoil	0.6821466	0.7312155	0.4094918	0.4472222
Humdepth	0.3767463	0.9263165	0.4893654	0.3361111
pH	-0.6165744	-0.7872967	0.5781701	0.4000000

File path: 08.Environment_factor/environment_evaluation

- NMDS1: The cosine value of the angle between the environmental factor and the sorting axis, indicating the correlation between the environmental factor and the sorting axis.
- NMDS2: Same as NMDS1.
- R2: Represents the coefficient of determination of environmental factors on species distribution. A smaller value indicates a smaller impact of the environmental factor on species distribution.
- P-value: The P-value of the significance test. $P < 0.05$ indicates statistical significance.

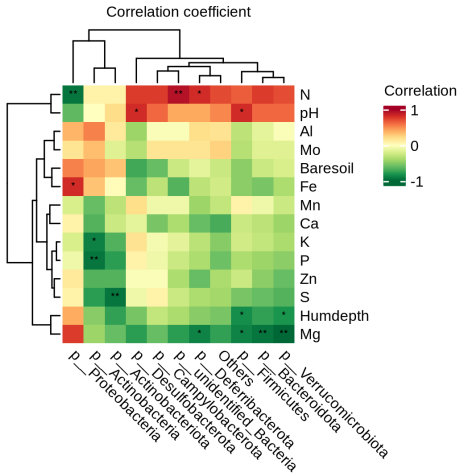
9.2 Spearman Correlation Analysis

When studying the correlation between environmental factors and species, as well as the correlation between environmental factors and species abundance (alpha diversity), commonly used methods include Spearman correlation analysis and Mantel test analysis. These methods examine the correlation between pairwise matrices and provide significance values.

Spearman correlation analysis uses the Spearman correlation coefficient as a measure, also known as the

rank correlation coefficient. It employs the rank order of two variables for linear correlation analysis, making no assumptions about the distribution of the original variables. It is a non-parametric statistical method with a wide range of applications.

Using Spearman rank correlation to study the relationship between environmental factors and microbial species abundance (alpha diversity), the analysis explores the mutual variation between environmental factors and species. The results include correlations and significant P-values between each pair. All environmental factors are used in the correlation analysis, and the results are presented below:



Heatmap of Spearman Correlations Between All Environmental Factors and Microbes Based on ASV

Environmental factors are shown vertically; species are shown horizontally; the values in the middle heatmap correspond to the Spearman correlation coefficient r, which is between -1 and 1, with r < 0 indicating a negative correlation and r > 0 indicating a positive correlation; the '*' denotes a significance test P-value < 0.05, and '**' denotes a P-value < 0.01.

File path: 08.Environment_factor/spearman/phylum

9.3 Environmental Factor Selection

There are often numerous environmental factors associated with microbial communities, but these factors frequently exhibit multicollinearity. Multicollinearity refers to a high correlation between environmental factors, leading to model distortion or difficulties in accurate estimation. Therefore, before environmental factor analysis, filtering is necessary to retain only those factors that have a significant impact on microbial communities.

The Variance Inflation Factor (VIF) is a measure that quantifies the severity of multicollinearity in a multiple linear regression model. A VIF value is calculated for each candidate environmental factor, and typically, a VIF value greater than 10 indicates collinearity. The calculation is performed iteratively, eliminating factors with VIF values exceeding 10 until all remaining environmental factors have VIF values below 10. After removing collinearity, the environmental factors are displayed as follows:

Table 12 Environmental Factors After Removing Collinearity

Environment	VIF
N	1.550707
Mg	1.817730
Al	2.663298
P	3.317414
Mn	3.425331

File path: 08.Environment_factor

- Environment: Represents environmental factors
- VIF: Variance Inflation Factor

The environmental factors selected by VIF filtering are then subjected to BioENV analysis. This analysis provides multiple combinations of environmental factors and calculates correlation values for each combination with the microbial community. Based on these correlation values, a combination with the highest correlation to the microbial community is identified, indicating the factors that have the greatest impact on the microbial community.

Table 13 Combinations of Environmental Factors

Environment	Size	Correlation
N + Mg	2	0.4321429
N + Mg + Al	3	0.3464286
N + Mg + Al + P	4	0.2750000
N + Mg + Al + P + Mn	5	0.0357143

File path: 08.Environment_factor

- Environment: Represents combinations of environmental factors
- Size: Number of environmental factors in the combination
- Correlation: Correlation between environmental factors and the microbial community

9.4 Mantel Test Analysis

The Mantel test is a test of the correlation between two matrices, commonly used in ecology to assess the correlation between environmental factors and microbial community data. The environmental factor combinations obtained from BioENV filtering will be used to calculate the overall correlation with community data. The Mantel test was performed using the vegan package (v2.6.2) of the R software, and the results are presented below:

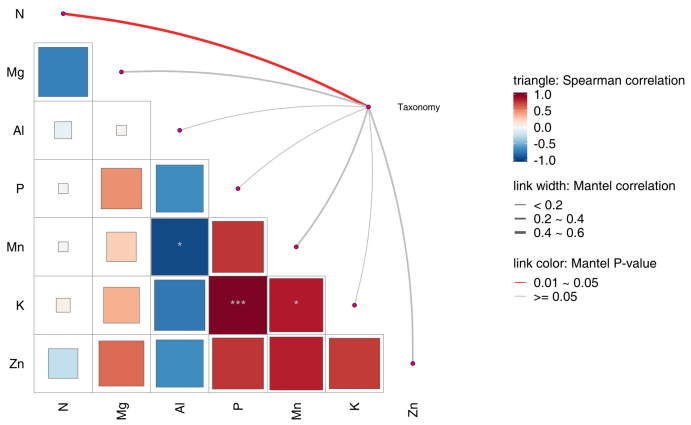
Table 14 Mantel Test Analysis of Environmental Factor Combinations and Microbes Based on ASV

Environment	Correlation	P-value
N + Mg	0.6892857	0.0166667
N + Mg + P	0.5321429	0.0263889
N + Mg + P + Mn	0.3321429	0.0875000
N + Mg + Al + P + Mn	0.3035714	0.1208333
N + P + K	0.1642857	0.2666667
Mn + Zn	-0.3750000	0.9486111
Zn + Al	-0.2678571	0.8597222

File path: 08.Environment_factor/Mantel-test

- Environment: Represents combinations of environmental factors
- Correlation: Represents the Spearman correlation coefficient between the environmental factor combination and the microbiota. The larger the absolute value, the greater the correlation between this set of environmental factors and species abundance.
- P-value: P-value of the significance test. $P < 0.05$ indicates statistical significance.

The results of the Mantel test analysis of the microbial community against individual metabolites are shown below:



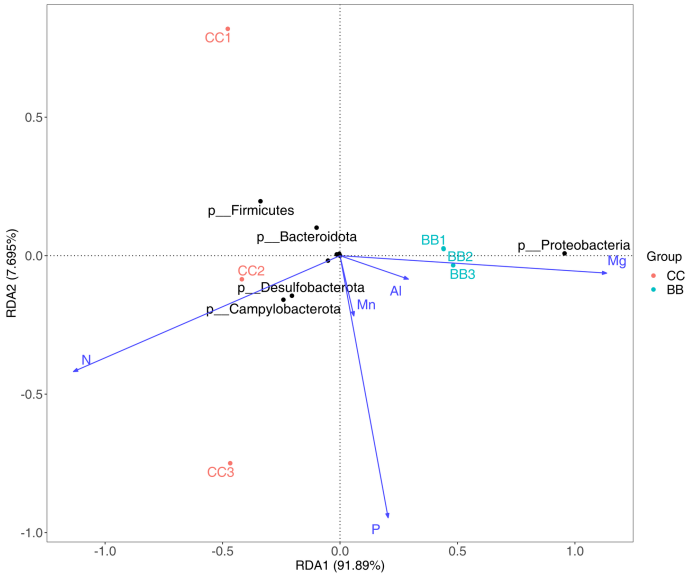
Mantel Test of Differential Microbes and Individual Environmental Factor

The lower triangle is the Spearman correlation between environmental factors. The size of the color block in the cell indicates the correlation coefficient, with red indicating a positive correlation and blue indicating a negative correlation. the significance test P-value is marked with '*'、'**' and '***', which indicate P-value < 0.05, P-value < 0.01, and P-value < 0.001, respectively, while the non-significant ones are without these markers. The connecting lines in the upper right area indicate the Mantel test results for communities and environmental factors. The length of the line indicates the overall correlation coefficient, and the color of the connecting line indicates the significance test result of the correlation coefficient.

File path: 08.Environment_factor/Mantel-test

9.5 CCA/RDA Analysis

CCA/RDA analysis is a sorting method developed based on correspondence analysis, combining correspondence analysis with multiple regression analysis. Each calculation step is regressed against environmental factors, also known as multivariate direct gradient analysis. CCA is based on a unimodal model, while RDA is based on a linear model. CCA/RDA analysis is mainly used to reflect the relationship between microbial communities and environmental factors. It can detect relationships between environmental factors, samples, and microbial communities or relationships between any two of them. It can identify important environmental driving factors influencing sample distribution. Detrended Correspondence Analysis (DCA) is used to assess whether CCA or RDA should be used. The program will automatically choose the appropriate method based on the DCA analysis results. The environmental factors filtered by VIF will be used for CCA/RDA analysis, and the results are shown below:



CCA/RDA Plot of Environmental Factor Combinations and Microbes Based on ASV

Black labels and dots indicate microbes; blue labels and arrows indicate environmental factors; colored labels indicate samples. The length of the arrow indicates the strength of the effect of the environmental factor on the microbial change. The longer the arrow, the greater the effect of the environmental factor on the microbial change. The vertical distance from the sample node to the environmental factor line segment and its extension line indicates the intensity of the environmental factor's impact on the sample - the closer the distance, the greater the impact of the environmental factor on the sample. Starting from the center origin, if the microbes are in the same direction as the arrows, it means that the environmental factors are positively correlated with the changes in the microbes, and vice versa indicates a negative correlation.

File path: 08.Environment_factor/CCA/phylum

Significance of each environmental factor is tested using the envfit function, as shown below:

Table 15 Regression Fitting Analysis of Environmental Factors and Microbes Based on ASV

Environment	RDA1	RDA2	R2	P-value
N	-0.9382255	-0.3460245	0.6886306	0.1375000
Mg	0.9984608	-0.0554626	0.6084834	0.2305556
pH	-0.9999306	0.0117832	0.5142674	0.2666667
P	0.2115651	-0.9773639	0.4412699	0.4000000
S	0.7934613	-0.6086207	0.4361771	0.4916667
Humdepth	0.9919594	-0.1265565	0.3626962	0.5388889
Baresoil	0.9803267	0.1973818	0.3234694	0.5750000
K	0.0638692	-0.9979583	0.3007229	0.5861111
Ca	0.9884792	0.1513567	0.2096358	0.6611111
Zn	0.8463833	-0.5325743	0.2102771	0.7180556
Fe	0.9825299	-0.1861046	0.1290215	0.7652778
Al	0.9601530	-0.2794748	0.0433018	0.8319444
Mo	0.5598667	-0.8285827	0.0021937	0.8666667
Mn	0.2612176	-0.9652799	0.0239437	0.9805556

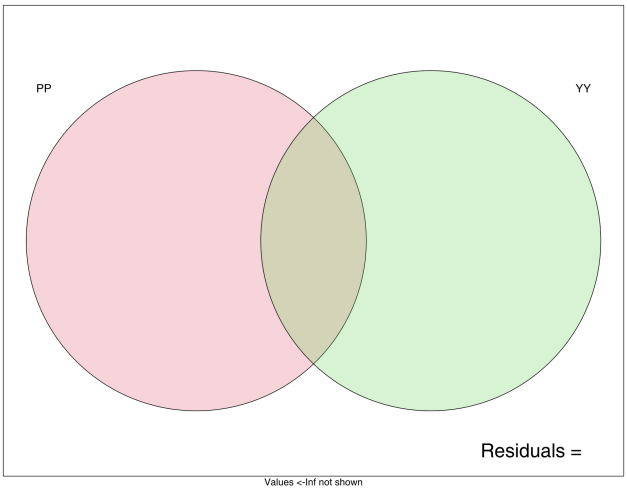
File path: 08.Environment_factor/CCA/phylum

- CCA1/RDA1: Represents the cosine value of the angle between the environmental factor and the ordination axis, indicating the correlation between the environmental factor and the ordination axis.
- CCA2/RDA2: Same as CCA1/RDA1.
- R2: Represents the coefficient of determination for the environmental factor's influence on species distribution. A smaller value indicates a smaller impact of the environmental factor on species distribution.
- P-value: P-value of the significance test. $P < 0.05$ indicates statistical significance.

9.6 VPA Analysis

CCA/RDA analysis is used to discover environmental factors that influence community structure. However, a drawback is the inability to intuitively and quantitatively show how a specific environmental factor affects the overall community change. When evaluating the contribution of environmental factors to community changes, Variance Partitioning Analysis (VPA) can be performed. In VPA analysis, environmental

factors need to be grouped first. Then, under the constraint of other categories of environmental factors, a sorted analysis of a specific category of environmental factors is conducted. This type of analysis is also known as partial CCA/RDA. After conducting partial analysis for each category of environmental factors, the contribution of each environmental factor individually and the interaction between different environmental factors to the changes in the biological community can be calculated. The specified combination of environmental factors will be used for VPA analysis, and the results are shown below:



VPA of Environmental Factor Groups and Microbes Based on ASV

The areas exclusively occupied by the two circles represent the contribution of that group of environmental factors to community change; the intersection of the circles represents the contribution of the interaction of the two groups of environmental factors to community change; and the out-of-circle residuals represent the community change that cannot be explained by either of the two groups of environmental factors.

File path: 08.Environment_factor/VPA

10 Advance Analysis

10.1 microPITA

microPITA is used to select suitable samples for metagenomic analysis. The software provides various methods to help choose representative and interesting samples.

- Unsupervised methods
  - diverse: Selects the sample with the highest α-diversity, defaulting to the Simpson method.
  - extreme: Selects the sample with the farthest β-diversity distance, defaulting to Bray-Curtis distance.
  - representative: Selects a representative sample that reflects overall differences, defaulting to Bray-Curtis distance.
  - features: Selects samples based on target species (OTU/ASV) with two methods,
    * rank: Selects the sample with the most target species.
    * abundance: Selects the sample with the highest abundance of target species.
- Supervised methods
  - distinct: Selects the sample with the largest β-diversity distance between groups based on phenotype/group features.
  - discriminant: Selects the sample closest to the group center based on phenotype/group features.

Table 16 Samples Screened by Different microPITA Methods

Method	Sample
diverse	DD2, DD1, DD3, BB3, BB1, AA3
extreme	DD2, AA3, DD1, AA1, DD3, BB3
representative	DD1, CC2, BB2, BB3, CC3, DD3
distinct	CC1, CC2, CC3, AA2, AA3, AA1, DD3, DD1, DD2, BB2, BB1, BB3
discriminant	CC1, CC2, CC3, AA2, AA3, AA1, DD3, DD1, DD2, BB2, BB1, BB3

File path: 09.Advance_analysis/microPITA

- Method: Selection method

- Sample: Selected samples

11 Method Description and Common After-sales

11.1 Data Mining Recommendations

16S rDNA (18S rDNA, ITS) amplicon sequencing is widely used for comparative analysis of microbial community structure differences in natural environments such as soil, water, and the gastrointestinal tracts of humans and animals. For studies with such objectives, the focus is generally on the following aspects of information analysis:

The primary focus is on OTU clustering and species annotation results. Default results include OTU clustering analysis with 97% similarity, where OTU clustering representative sequences are found in 02.OTU_analysis/otu.fasta, and annotation information for each OTU can be found in 02.OTU_analysis/otu.taxonomy_assignments.xlsx. The normalized absolute abundance information for species is available in the 02.OTU_analysis/taxa_abundance_absolute folder, and the normalized relative abundance can be found in 02.OTU_analysis/taxa_abundance_relative. Subsequent alpha and beta diversity analyses are based on the normalized OTU table. Taking taxa_abundance_relative as an example, it includes the relative abundance of species at six taxonomic levels (phylum, class, order, family, genus, species) and the relative abundance of each OTU in each sample. For instance, to view the relative abundance of Actinomycetaceae at the family level, you can open 02.OTU_analysis / taxa_abundance_relative / otu.table.relative.taxonomy.family.xlsx and search for Actinomycetaceae. Species composition and distribution in samples can be intuitively understood from the species annotation results, and significant or selected species can be further analyzed based on the project background.

If ASVs (Amplicon Sequence Variants) are generated using denoising methods, the corresponding files and directories are:

1) 02.ASV_analysis/ASV.fasta
2) 02.ASV_analysis/ASV.taxonomy_assignments.xlsx
3) 02.ASV_analysis/taxa_abundance_absolute
4) 02.ASV_analysis/taxa_abundance_relative
5) 02.ASV_analysis/taxa_abundance_absolute/ASV.table.absolute.taxonomy.family.xlsx

To view the distribution of major species in each sample, check the 03.OTU_visualization/top10_barplot or 03.ASV_visualization/top10_barplot folder. These folders include lists and bar plots of the distribution of the top 10 species with the highest abundance at the taxonomic levels of phylum, class, order, family, genus, and species. Understanding the distribution of major species in each sample helps identify dominant species between samples and differences in dominant species among samples.

For the assessment of species richness and evenness within samples, refer to 04.Alpha_diversity / alpha_diversity.xlsx, which includes results for seven different alpha diversity indices (ACE, chao1, goods_coverage, observed_otus/ASV, shannon, simpson, PD_whole_tree). The values of these indices reflect the complexity of the microbial community within the samples. By examining the significance test results of alpha diversity indices between groups (06.Diff_analysis/alpha_stat), species with significantly increased or decreased diversity can be quickly identified, facilitating further analysis in conjunction with biological treatments.

For beta diversity analysis, which involves comparing the microbial community structure differences between samples, there are various methods.

Firstly, unifrac distances between pairs of samples provide a visual representation of the degree of community structure differences between each pair of samples. Detailed results are available in 05.Beta_diversity/beta_heatmap.

PCA, PCoA, and NMDS plots provide a two-dimensional representation of these community structure differences, allowing for visual assessment of clustering and separation of grouped samples (or individual samples).

UPGMA clustering trees (05.Beta_diversity/Tree) enable hierarchical clustering of samples, providing a clearer view of similarity clustering between samples.

Through beta diversity analysis results, one can observe whether the differences in community structure between samples align with biological groupings. This clustering or differential situation can be interpreted in conjunction with biological questions. Additionally, in UPGMA, interpreting sample clustering can be combined with the distribution of high-abundance taxa.

For projects with groups, in-depth analysis can be conducted. LEfSe analysis can identify biomarkers with statistically significant differences between groups. T-test and Metastats analyses can identify significantly different species between different groups. Group-level species differential significance analysis is performed at six taxonomic levels (phylum, class, order, family, genus, species). Simper analysis quantifies the contribution of species to differences. Anosim and MRPP analyses determine whether differences

in community structure between groups are significant and allow for the comparison of the magnitude of within-group and between-group differences.

11.2 Method Description

11.2.1 Sequencing Part

1. Extraction and PCR Amplification of Genomic DNA

Genomic DNA of the samples is extracted using the CTAB or SDS method. Subsequently, the purity and concentration of DNA are assessed using agarose gel electrophoresis. An appropriate amount of sample DNA is taken in a centrifuge tube, and the sample is diluted with sterile water to a concentration of 1 ng/μl.

Using the diluted genomic DNA as a template and based on the selected sequencing region, specific primers with barcodes, New England Biolabs' Phusion High-Fidelity PCR Master Mix with GC Buffer, and a high-efficiency, high-fidelity enzyme are used for PCR to ensure amplification efficiency and accuracy.

Primer regions include:

- 16S V4 region primers (515F and 806R): Identify bacterial diversity.
- 18S V4 region primers (528F and 706R): Identify eukaryotic microbial diversity.
- ITS1 region primers (ITS5-1737F and ITS2-2043R): Identify fungal diversity.

Additionally, the amplified regions include: 16S V3-V4/16S V4-V5/16SV5-V7; Archaea 16S V4-V5/Archaea 16S V8; 18S V9 and ITS2 regions.

2. Mixing and Purification of PCR Products

PCR products are electrophoresed on a 2% agarose gel to check their concentration. Qualified PCR products are subjected to magnetic bead purification, quantified using enzyme labeling, mixed equimolarly based on PCR product concentration, thoroughly mixed, and electrophoresed again on a 2% agarose gel. The desired bands are recovered using the gel recovery kit provided by Qiagen.

3. Library Construction and Sequencing

Library construction is carried out using the TruSeq DNA PCR-Free Sample Preparation Kit. After quantification using Qubit and Q-PCR, qualified libraries undergo sequencing using NovaSeq6000.

11.2.2 Information Analysis Section

Processing of Sequencing Data

Barcode sequences and PCR amplification primer sequences are separated from the raw data for each sample. The Barcode and primer sequences are trimmed. The raw reads are filtered for high-quality reads using fastp (v0.22.0, https://github.com/OpenGene/fastp), employing the following filtering criteria: automatic detection and removal of adapter sequences; removal of reads with 1 or more N bases; removal of reads with over 40% low-quality bases (quality value < 15); deletion of bases with an average quality below 20 in a 4-base window; removal of polyG tails; and deletion of reads shorter than 150 bp. High-quality paired-end reads are then assembled using FLASH (v1.2.11, http://ccb.jhu.edu/software/FLASH/) to obtain high-quality Tags data (Clean Tags). Tags sequences are aligned with a species annotation database using vsearch (v2.22.1, https://github.com/torognes/vsearch/) to detect chimeric sequences. Chimeric sequences are then removed, resulting in the final effective data (Effective Tags).

OTU Clustering, ASV Denoising, and Species Annotation

For OTU clustering, the Uparse algorithm (from the USEARCH v7 software, http://www.drive5.com/uparse/) is applied to cluster all Effective Tags from all samples. Sequences are clustered into Operational Taxonomic Units (OTUs) with a default identity of 97%. Representative sequences are selected based on the highest frequency within the OTU, following the algorithm's principles. If ASV denoising is chosen, the Deblur (default, v1.1.1) or DADA2 (v1.26.0) method is employed. Both Deblur and DADA2 use QIIME 2 (v2023.2) (Bolyen et al. 2019).

The minimum total observation count of an OTU/ASV is set at 0.005% of the total observation (sequence) counts after rarefaction (Bokulich et al. 2013).

Species annotation for OTU/ASV sequences is performed using the Mothur (v1.48) method with the SILVA138.1 (http://www.arb-silva.de/) SSUrRNA database. Annotations are conducted with a threshold of 0.8 to 1. Taxonomic information is obtained, and the composition of communities at various taxonomic levels (phylum, class, order, family, genus, species) is statistically analyzed for each sample. Fast multiple sequence alignment is performed using MAFFT (v7.520, https://mafft.cbrc.jp/alignment/software/) to establish the phylogenetic relationships of all OTU/ASV representative sequences. Subsequently, data for each sample is normalized based on the sample with the least data, and subsequent Alpha and Beta diversity analyses are performed on the normalized data.

For ITS projects, the Mothur software is used to align with the UNITE (ver9, 29.11.2022, https://unite.ut.ee/) database for species annotation.

If the oral microbiome database is used for annotation, the eHOMD V15.22 (https://www.ehomd.org/) is employed, and subsequent analysis steps are the same as with the SILVA database.

Sample Complexity Analysis (Alpha Diversity)

R software (v4.2.0) with the phyloseq (v1.40.0) and vegan (v2.6.2) packages is used to calculate Observed_otus/ASV, Chao1, Shannon, Simpson, ACE, Goods-coverage, and PD_whole_tree indices. Dilution curves, Rank abundance curves, and species accumulation curves are plotted using R software, and Alpha diversity index inter-group difference analysis is conducted. Inter-group difference analysis includes both parametric and non-parametric tests. For two groups, T-test and Wilcoxon test are used, while for more than two groups, Tukey test and Kruskal-Wallis test are employed.

The specific description of Alpha diversity indices is as follows:

Calculation of the Community Richness index includes:

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

ACE - the ACE estimator (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.ace.html#skbio.diversity.alpha.ace);

Indices for calculating microbial community diversity include:

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);

Sequencing depth indices include:

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

Indices of phylogenetic diversity include:

PD_whole_tree - PD_whole_tree index(http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.faith_pd.html?highlight=pd#skbio.diversity.alpha.faith_pd)

Beta Diversity

Using the phyloseq package (v1.40.0) in R software (v4.2.0), Unifrac distances are computed, and a UPGMA (Unweighted Pair Group Method with Arithmetic Mean) sample clustering tree is constructed. PCA,

PCoA, and NMDS plots are generated using the stats and phyloseq packages in R software (v4.2.0). Beta diversity analysis is performed in R, employing both parametric and non-parametric tests. T-test and Wilcoxon test are chosen for two groups, while Tukey test and Kruskal-Wallis test are used for more than two groups.

LEfSe analysis is conducted using LEfSe (v1.1.2) software, with a default LDA Score filtering value of 4. Metastats analysis, performed in Mothur software, involves a permutation test at various taxonomic levels (Phylum, Class, Order, Family, Genus, Species) to obtain p-values. These p-values are then corrected using the Benjamini and Hochberg False Discovery Rate method to obtain q-values. Anosim, MRPP, and Adonis analyses utilize the anosim, mrpp, and adonis functions from the vegan package in R. AMOVA analysis is conducted using the amova function in Mothur. Species with significant inter-group differences are identified using T-test in R, and visualizations are generated.

Network Analysis

Microbial data analysis is based on relative abundance. When using conventional methods such as Pearson and Spearman correlation for compositional data, biases may arise due to the dependence of the correlation on the sum of fractions equaling 1. SparCC algorithm (Friedman and Alm 2012), and its more efficient C++ implementation, FastSpar (Watts et al. 2019), are employed to reliably estimate correlations in compositional data. FastSpar (v1.0.0) is used for species correlation calculations.

Top 100 microbial genera, selected based on abundance, undergo correlation analysis. Filtering criteria include: (1) removal of connections with absolute correlation coefficient <= 0.8 (default value), (2) filtering out self-connections of nodes, and (3) removal of connections with node abundance less than 0.005%.

Functional Annotation

The full name of PICRUSt is Phylogenetic Investigation of Communities by Reconstruction of Unobserved States. PICRUSt2 (Douglas et al. 2020) (v2.5.0) is a bioinformatics software package for predicting metagenomic functions based on marker genes, such as 16S rRNA. For detailed prediction processes, refer to the PICRUSt website. Currently, functional predictions based on 16S sequencing data can be made using the KEGG database.

Tax4Fun2 (Wemheuer et al. 2018) (v1.1.5) is an upgraded version of Tax4Fun, allowing rapid prediction of the functional profiles and redundancy of prokaryotes based on 16S rRNA gene sequences. By merging user-defined, habitat-specific genomic information, it significantly enhances the accuracy and robustness of predicted functional profiles. Compared to the old version Tax4Fun, Tax4Fun2 has the following advantages:

1) No longer limited to the OTU/ASV abundance table annotated with specific versions of SILVA. It allows direct input of OTU/ASV representative sequences for species annotation. In addition to the

built-in reference set provided by Tax4Fun2 (significantly expanded compared to previous versions), users can also provide custom reference sets, offering great flexibility.

2) Focuses on prokaryotic data but can also merge eukaryotic data.

3) Provides methods for calculating specific functional redundancy, which is crucial for predicting the potential loss of specific functions during environmental disturbances.

4) Substantial improvements in accuracy and stability.

FAPROTAX (Functional Annotation of Prokaryotic Taxa) (Louca, Parfrey, and Doebeli 2016) (v1.2.4) is a manually curated database constructed based on published literature. It includes associations between prokaryotic microbial taxonomy (genus or species) and functions related to metabolism or ecology. The database encompasses over 80 functional categories, including carbon, nitrogen, phosphorus, sulfur cycling, animal and plant pathogens, methane generation, fermentation, and more, with over 7600 functional annotations covering more than 4600 prokaryotic species. It is suitable for functional annotation analysis of biochemical cycling processes in environmental samples. FAPROTAX converts quantitative information about microbial abundance into quantitative information about functional classification based on the identified microbial taxonomy information and functional annotation information in the database.

BugBase (Ward et al. 2017) (v.0.1.0) is a database for predicting high-level phenotypes of microorganisms based on the GreenGenes (16S) (DeSantis et al. 2006) database. In addition to phenotype prediction, the database allows inter-group differential analysis and statistical chart display for different phenotypes. Currently, microbial communities can be classified based on seven phenotypes: Gram-positive, Gram-negative, biofilm-forming, pathogenic potential, mobile element-containing, oxygen utilizing (including aerobic, anaerobic, facultatively anaerobic), and oxidative stress-tolerant.

The functional prediction database FUNGuild (Nguyen et al. 2016) (v1.1) is used to associate ITS sequences with functional classifications of fungi. The database categorizes fungi into three major nutritional modes: 1) pathotroph, obtaining nutrients by damaging host cells and causing diseases; 2) saprotroph, obtaining nutrients by decomposing dead host cells; 3) symbiotroph, obtaining nutrients by exchanging resources with host cells. These three major categories are further divided into 12 functional groups (guilds).

Random Forest

Random Forest belongs to the ensemble type of machine learning algorithms. It utilizes the bootstrap aggregating (bagging) resampling method to extract multiple samples with replacement from the original dataset as training sets. Decision trees are then modeled on these training sets, and the predictions of multiple decision trees are combined to obtain the final prediction through voting. RF has high prediction accuracy,

good tolerance to outliers and noise, and is less prone to overfitting.

When extracting the training set with the bagging method, the probability that a sample is not selected from the original data is $(1–1/N)^N$, where N is the number of samples in the original data. When N is sufficiently large, $(1–1/N)^N$ converges to $1/e \approx 0.368$. This indicates that approximately 37% of the samples in the original data will not appear in the extracted training set, and these data are referred to as out-of-bag (OOB) data. OOB data will be used to estimate the model's performance. Random Forest is implemented using the R language package varSelRF (v0.7.8).

If conducting a Random Forest analysis, it is recommended that the number of samples in a single group be greater than or equal to 15.

microPITA

microPITA (v1.1.0) is used to select suitable samples for metagenomic analyses. The software provides various methods to help choose representative and interesting samples.

- Unsupervised Methods
  - diverse: Select samples with the highest α-diversity, defaulting to the Simpson method.
  - extreme: Select samples with the farthest β-diversity distance, defaulting to the Bray-Curtis distance.
  - representative: Choose representative samples that reflect overall differences, defaulting to the Bray-Curtis distance.
  - features: Select samples based on target species (OTU/ASV) using two methods,
    * rank: Choose samples with the highest number of target species.
    * abundance: Choose samples with the highest abundance of target species.
- Supervised Methods
  - distinct: Select samples with the maximum inter-group β-diversity distance based on phenotype/grouping characteristics.
  - discriminant: Select samples closest to the center of the grouping based on phenotype/grouping characteristics.

11.3 Appendix

1) Introduction to the Delivery Data Directory Structure: ReadMe.html

2) Instructions for Interactive Web Version: web_show.pdf

3) Methods in English: methods_english.pdf

4) After-Sales FAQ: FAQ.pdf

5) Introduction to Statistical Methods: statistics_method.pdf

6) Notes

The final report only presents part of the results. For the complete results, please refer to specific files. Each image in the delivered results will be provided not only in PNG format but also in PDF vector format.

Reference

Avershina, Ekaterina, Trine Frisli, and Knut Rudi. 2013. "*De Novo* Semi-Alignment of 16S rRNA Gene Sequences for Deep Phylogenetic Characterization of Next Generation Sequencing Data." *Microbes and Environments* 28 (2): 211–16. https://doi.org/10.1264/jsme2.ME12157.

Banerjee, Samiran, Klaus Schlaeppi, and Marcel G. A. van der Heijden. 2018. "Keystone Taxa as Drivers of Microbiome Structure and Functioning." *Nature Reviews. Microbiology* 16 (9): 567–76. https://doi.org/10.1038/s41579-018-0024-1.

Bokulich, Nicholas A, Sathish Subramanian, Jeremiah J Faith, Dirk Gevers, Jeffrey I Gordon, Rob Knight, David A Mills, and J Gregory Caporaso. 2013. "Quality-Filtering Vastly Improves Diversity Estimates from Illumina Amplicon Sequencing." *Nature Methods* 10 (1): 57–59. https://doi.org/10.1038/nmeth.2276.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Bulgarelli, Davide, Ruben Garrido-Oter, Philipp C. Münch, Aaron Weiman, Johannes Dröge, Yao Pan, Alice C. McHardy, and Paul Schulze-Lefert. 2015. "Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley." *Cell Host & Microbe* 17 (3): 392–403. https://doi.org/10.1016/j.chom.2015.01.011.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight. 2011. "Global Patterns of 16S rRNA Diversity at a Depth of Millions of Sequences

Per Sample." *Proceedings of the National Academy of Sciences* 108 (Supplement_1): 4516–22. https://doi.org/10.1073/pnas.1000080107.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5): 335–36. https://doi.org/10.1038/nmeth.f.303.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72. https://doi.org/10.1128/AEM.03006-05.

Douglas, Gavin M., Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. 2020. "PICRUSt2 for Prediction of Metagenome Functions." *Nature Biotechnology* 38 (6): 685–88. https://doi.org/10.1038/s41587-020-0548-6.

Edgar, R. C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–7. https://doi.org/10.1093/nar/gkh340.

Friedman, Jonathan, and Eric J. Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." *PLoS Computational Biology* 8 (9): e1002687. https://doi.org/10.1371/journal.pcbi.1002687.

Haas, B. J., D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, D. Ciulla, et al. 2011a. "Chimeric 16S rRNA Sequence Formation and Detection in Sanger and 454-Pyrosequenced PCR Amplicons." *Genome Research* 21 (3): 494–504. https://doi.org/10.1101/gr.112730.110.

———. 2011b. "Chimeric 16S rRNA Sequence Formation and Detection in Sanger and 454-Pyrosequenced PCR Amplicons." *Genome Research* 21 (3): 494–504. https://doi.org/10.1101/gr.112730.110.

Hess, Matthias, Alexander Sczyrba, Rob Egan, Tae-Wan Kim, Harshal Chokhawala, Gary Schroth, Shujun Luo, et al. 2011. "Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen." *Science* 331 (6016): 463–67. https://doi.org/10.1126/science.1200387.

Jiao, Shuo, Weimin Chen, and Gehong Wei. 2017. "Biogeography and Ecological Diversity Patterns of Rare and Abundant Bacteria in Oil-Contaminated Soils." *Molecular Ecology* 26 (19): 5305–17. https://doi.org/10.1111/mec.14218.

Jolliffe, Ian T., and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*

Sciences 374 (2065): 20150202. https://doi.org/10.1098/rsta.2015.0202.

Langille, Morgan G I, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, et al. 2013. "Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences." *Nature Biotechnology* 31 (9): 814–21. https://doi.org/10.1038/nbt.2676.

Li, Bing, Xuxiang Zhang, Feng Guo, Weimin Wu, and Tong Zhang. 2013. "Characterization of Tetracycline Resistant Bacterial Community in Saline Activated Sludge Using Batch Stress Incubation with High-Throughput Sequencing Analysis." *Water Research* 47 (13): 4207–16. https://doi.org/10.1016/j.watres.2013.04.021.

Louca, Stilianos, Laura Wegener Parfrey, and Michael Doebeli. 2016. "Decoupling Function and Taxonomy in the Global Ocean Microbiome." *Science (New York, N.Y.)* 353 (6305): 1272–7. https://doi.org/10.1126/science.aaf4507.

Lozupone, Catherine A., Micah Hamady, Scott T. Kelley, and Rob Knight. 2007. "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities." *Applied and Environmental Microbiology* 73 (5): 1576–85. https://doi.org/10.1128/AEM.01996-06.

Lozupone, Catherine, and Rob Knight. 2005. "UniFrac: A New Phylogenetic Method for Comparing Microbial Communities." *Applied and Environmental Microbiology* 71 (12): 8228–35. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.

Lozupone, Catherine, Manuel E Lladser, Dan Knights, Jesse Stombaugh, and Rob Knight. 2011. "UniFrac: An Effective Distance Metric for Microbial Community Comparison." *The ISME Journal* 5 (2): 169–72. https://doi.org/10.1038/ismej.2010.133.

Lundberg, Derek S, Scott Yourstone, Piotr Mieczkowski, Corbin D Jones, and Jeffery L Dangl. 2013. "Practical Innovations for High-Throughput Amplicon Sequencing." *Nature Methods* 10 (10): 999–1002. https://doi.org/10.1038/nmeth.2634.

Magoc, T., and S. L. Salzberg. 2011. "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies." *Bioinformatics* 27 (21): 2957–63. https://doi.org/10.1093/bioinformatics/btr507.

McGraw, Robert, and Renyi Zhang. 2008. "Multivariate Analysis of Homogeneous Nucleation Rate Measurements. Nucleation in the P-Toluic Acid/Sulfuric Acid/Water System." *The Journal of Chemical Physics* 128 (6): 064508. https://doi.org/10.1063/1.2830030.

Nguyen, Nhu H., Zewei Song, Scott T. Bates, Sara Branco, Leho Tedersoo, Jon Menke, Jonathan S. Schilling, and Peter G. Kennedy. 2016. "FUNGuild: An Open Annotation Tool for Parsing Fungal Community Datasets by Ecological Guild." *Fungal Ecology* 20 (April): 241–48. https://doi.org/10.1016/j.funeco.

2015.06.006.

Noval Rivas, Magali, Oliver T. Burton, Petra Wise, Yu-qian Zhang, Suejy A. Hobson, Maria Garcia Lloret, Christel Chehoud, et al. 2013. "A Microbiota Signature Associated with Experimental Food Allergy Promotes Allergic Sensitization and Anaphylaxis." *Journal of Allergy and Clinical Immunology* 131 (1): 201–12. https://doi.org/10.1016/j.jaci.2012.10.026.

Ondov, Brian D, Nicholas H Bergman, and Adam M Phillippy. 2011. "Interactive Metagenomic Visualization in a Web Browser." *BMC Bioinformatics* 12 (1): 385. https://doi.org/10.1186/1471-2105-12-385.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nature Methods* 10 (12): 1200–1202. https://doi.org/10.1038/nmeth.2658.

Pearson, Karl. 1997. "Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs." *Proceedings of the Royal Society of London* 60 (359): 489–98. https://doi.org/10.1098/rspl.1896.0076.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." *Nature* 490 (7418): 55–60. https://doi.org/10.1038/nature11450.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1): D590–D596. https://doi.org/10.1093/nar/gks1219.

Roewer, L. 1996. "Analysis of Molecular Variance (AMOVA) of Y-Chromosome-Specific Microsatellites in Two Closely Related Human Populations [Published Erratum Appears in Hum Mol Genet 1997 May;6(5):828]." *Human Molecular Genetics* 5 (7): 1029–33. https://doi.org/10.1093/hmg/5.7.1029.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. 2011. "Metagenomic Biomarker Discovery and Explanation." *Genome Biology* 12 (6): R60. https://doi.org/10.1186/gb-2011-12-6-r60.

Stat, Michael, Xavier Pochon, Erik C. Franklin, John F. Bruno, Kenneth S. Casey, Elizabeth R. Selig, and Ruth D. Gates. 2013. "The Distribution of the Thermally Tolerant Symbiont Lineage (*Symbiodinium*

Clade d) in Corals from Hawaii: Correlations with Host and the History of Ocean Thermal Stress." *Ecology and Evolution* 3 (5): 1317–29. https://doi.org/10.1002/ece3.556.

Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole. 2007. "Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73 (16): 5261–7. https://doi.org/10.1128/AEM.00062-07.

Ward, Tonya, Jake Larson, Jeremy Meulemans, Ben Hillmann, Joshua Lynch, Dimitri Sidiropoulos, John R. Spear, et al. 2017. "BugBase Predicts Organism-Level Microbiome Phenotypes." *bioRxiv*, January, 133462. https://doi.org/10.1101/133462.

Watts, Stephen C, Scott C Ritchie, Michael Inouye, and Kathryn E Holt. 2019. "FastSpar: Rapid and Scalable Correlation Estimation for Compositional Data." *Bioinformatics* 35 (6): 1064–6. https://doi.org/10.1093/bioinformatics/bty734.

Wemheuer, Franziska, Jessica A. Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer. 2018. "Tax4Fun2: A R-Based Tool for the Rapid Prediction of Habitat-Specific Functional Profiles and Functional Redundancy Based on 16S rRNA Gene Marker Gene Sequences." *bioRxiv*. https://doi.org/10.1101/490037.

White, James Robert, Niranjan Nagarajan, and Mihai Pop. 2009. "Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples." Edited by Christos A. Ouzounis. *PLoS Computational Biology* 5 (4): e1000352. https://doi.org/10.1371/journal.pcbi.1000352.

Youssef, Noha, Cody S. Sheik, Lee R. Krumholz, Fares Z. Najar, Bruce A. Roe, and Mostafa S. Elshahed. 2009. "Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys." *Applied and Environmental Microbiology* 75 (16): 5227–36. https://doi.org/10.1128/AEM.00592-09.