1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

Novogene 16S Amplicon Analysis Report

Contract Information	Contract Content
Contract_No	H204SC00000000
Batch_ID	X204SC00000000-Z01-F001
Report_Time	2023-05-09

Novogene

Novogene Co., Ltd

1 Overview

16S ribosomal RNA (rRNA) is a component of the prokaryotic 30S ribosomal subunit. It contains 9 hypervariable regions (V1-V9) and the length of these hypervariable regions range from approximately 30 to 100 base pairs. The degree of conservation in prokaryotic 16s rRNA hypervariable region varies across bacterial species, where higher gene similarity was found between organism of closely related species. Currently, 16S rRNA gene was widely used in bacterial phylogenetic analysis and classification[1,2,3] .

The advancement of high-throughput sequencing technology has dramatically increased the number of microbial community studies based on 16S rRNA genes. In general, universal primers were designed using the sequence in conserved regions flanking at the upstream and downstream of the targeted hypervariable region(s). Although only selected 16S rRNA hypervariable region(s) was amplified and sequenced, studies have shown that the selected hypervariable region is sufficient to reveal the microbial profile in the microbial community of the environment studied. To date, microbial community profile of vast number of environment were characterized using 16S rRNA sequences generated by various sequencing platforms.

1.1 Experimental Workflow

The quality of TGS data was directly affected by the sample handling procedures including the DNA extraction, library preparation and sequencing processes. The data quality is then further impacting the analysis result. To ensure the accuracy and reliability of sequencing data, quality control (QC) is performed at each step of the procedure. The flowchart is as follows:

Fig 1.1 Experimental Workflow

1.2 Analysis Workflow

The PacBio BAM file is splited according to barcode and filtered to get clean data. Then, the effective data is used to do Amplicon Sequence Variants (ASVs) analysis and species annotation. Thus the relative species, evenness and abundance distribution can be analyzed with Alpha diversity, Beta diversity, venn or flower graph et al. Furthermore, ASV consructions, taxonomic assignment construction of phylogenetic trees and prediction of metagenome functions from PICRUSt2 through downstream statical analysis explain the community construction differences between samples or among groups via PCoA, PCA and NMDS[4]. Statistic methodss such as T-test, MetaStat, LEfSe, and ANCOM and could test the significance of community composition and structure differences between groups. Most of the analysis are finished by the Qiime2 platform and DADA2 package[5].

Notably, DADA2 method for ASV inference requires high-quality data and deep sequencing (recommend at least 1w CCS reads). For samples with high diversity, such as environmental samples, higher sequencing depth is required. If the project data situation cannot satisfy the DADA2 method, we will replace the ASV approach with the OTU approach, which is finished by clustering with 97% similarity (vsearch from Qiime2 platform).

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

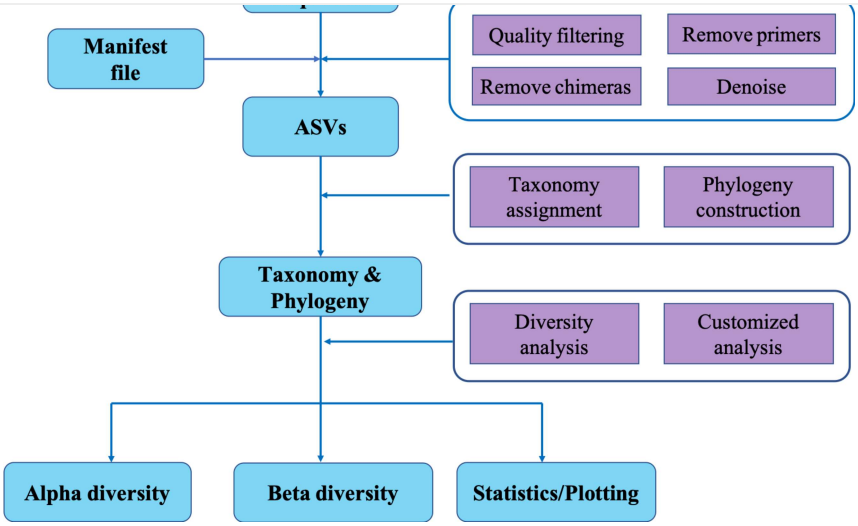


Fig 1.2 Analysis Workflow

Note: If there is no grouping information or the number of samples in the group is less than 3, then the Beta diversity analysis of differences between groups and the analysis of significant differences of species between groups cannot be performed. the ASV sequence would be further clustered by 100% identity before the downstream analysis for some identical ASVs might have variable lengths.

2 Data Processing

2.1 Data Filtering

Perform quality control on the qcCombined reads to obtain Clean reads. The statistical results obtained at each step in the data processing process are shown in the following table:

Table 2.1 Statistical results of data processing

Current: 1/1 page | Total items: 9 | First | Previous | Next | Last | Go to page Jump

Sample	NovoID	Clean reads	Clean data	AvgLen(nt)	Q20(%)	Q30(%)	GC(%)
Control1	FKDN230134872-1A	88295	129201185	1463	99.14	98.16	53.68
Control2	FKDN230134873-1A	81170	118379938	1458	99.23	98.31	53.47
Control4	FKDN230134875-1A	168873	246375574	1458	99.18	98.20	53.54
Control5	FKDN230134876-1A	71877	104680010	1456	99.17	98.19	53.81
Control6	FKDN230134877-1A	75336	109914442	1459	99.17	98.20	53.63
IonPoor1	FKDN230134878-1A	68963	100304522	1454	99.19	98.26	53.95
IonPoor2	FKDN230134879-1A	116765	165302704	1415	99.22	98.29	55.24
IonPoor4	FKDN230134881-1A	77340	109838646	1420	99.00	97.97	55.12
IonPoor5	FKDN230134882-1A	52018	75919257	1459	99.20	98.27	53.50

Note: Clean reads are the raw sequences filtered the low quality, that is Clean reads which are finally used for subsequent analysis; Clean data is the number of bases in the final Clean reads; AvgLen is the average length of Clean reads; Q20 and Q30 are the percentage of bases that have quality values in Clean reads greater than 20 (sequencing error rate less than 1%) and 30 (sequencing error rate less than 0.1%); GC (%) represents the content of GC bases in Clean reads; Effective (%) represents the percentage of the number of Clean reads to the number of Raw PE.

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

3 ASV Analysis

3.1 Denoise and Species Annotation

The DADA2 method[6] is mainly used for noise reduction. It no longer uses similarity clustering, but only performs dereplication or equivalent to 100% similarity clustering. Each de-duplicated sequence generated after noise reduction using DADA2 is called ASVs (Amplicon Sequence Variants), or feature sequence (corresponding to the OTU representative sequence), and the abundance table of these sequences in the sample is called the feature table (corresponds to the OTU table). The DADA2 method is more sensitive and specific than the traditional OTU method, and can detect the true biological mutations missed by the OTU method, while outputting fewer false sequences[7]. Compared with OTUs, ASVs improve the accuracy, comprehensiveness and repeatability of marker gene data analysis[8].

By applying QIIME2's classify-sklearn algorithm[9,10], a pre-trained Naive Bayes classifier is used for species annotation of each ASV.

According to the results of ASVs annotations and the feature tables of each sample, the species abundance tables at the level of kingdom, phyla, class, order, family, genus, and species are obtained. These abundance tables with annotation information are the core content of amplicon analysis. According to different experimental purposes, one or several species of key concern can be selected from the species abundance table of each classification level (usually focusing on the phylum and genus level), combined with the species composition and differential analysis of different samples (groups), and cluster analysis to conduct in-depth research.

3.2 Relative Abundance of Species

Based on the results of species annotations at different taxonomic levels, a histogram of species relative abundance is generated to view the species composition and proportions of each sample at different taxonomic levels. When the sample amount is large, the sample histogram will become crowded. It is recommended to display it in groups (the abundance is the average relative abundance of samples in the group) graphs.

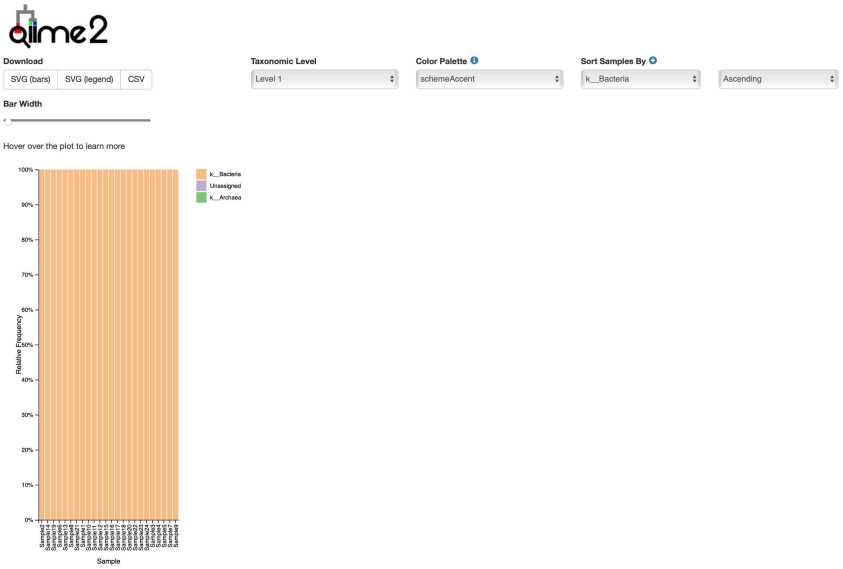


Fig 3.2 Histogram of Relative Abundance of Species

To view full size picture please click here

Results directory:

Representative sequences (ASVs) after denoise: result/02.ASVanalysis/ASV_table/rep_seqs_qza/ASV-dna-sequences.fasta

Frequency distribution of ASVs in each sample: result/02.ASVanalysis/ASV_table/asv_table_qza/ASV-feature-table.tsv

Visualization of annotation of ASVs: result/02.ASVanalysis/Seq_taxonomy/seq_taxonomy_qzv

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

3.3 Clustering of Species Abundance

According to the species annotations and abundance information of all samples at the genus level, select the top 35 genera in abundance, and cluster them from the species and sample levels according to their abundance information in each sample to draw the heatmap, to conveniently find the concentration of species in each sample. The results are shown below:

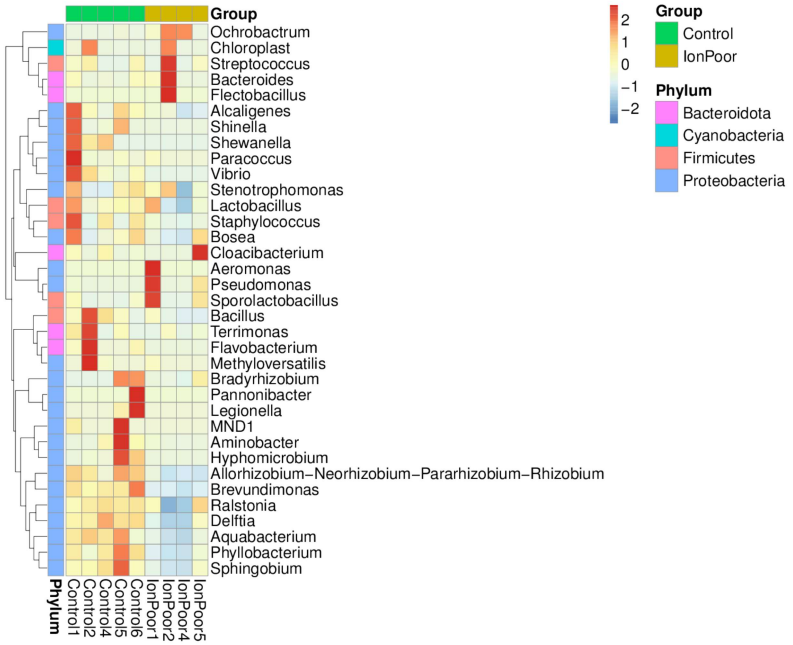


Fig 3.3 Heatmap of Clustering of Species Abundance

To view full size picture please click here

Note: x-axis represents the sample name and the y-axis represents the function annotation. The cluster tree on the left side of the figure is the species cluster tree; the corresponding value of the heatmap is the Z value of taxonomic relative abundance after standardization, that is, the Z value of a sample in a certain species is the differences between the relative abundance of the sample in the species and the average relative abundance of all samples in the species divided by the standard deviation of all samples in the species.

Results directory:

Cluster map of species abundance at different taxonomic levels: 03.AlphaDiversity/taxa_summary/cluster*.(png,pdf)

Plotting data: 03.AlphaDiversity/taxa_summary/cluster*.txt

3.4 Venn/Flower Diagram

According to the results of ASVs obtained from noise reduction and research needs, analyze the common and unique ASVs between different samples (groups). When the number of samples (groups) is less than or equal to 5, Venn diagram will be drawn. When the sample (group) number is greater than 5, Flower diagram will be drawn.

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

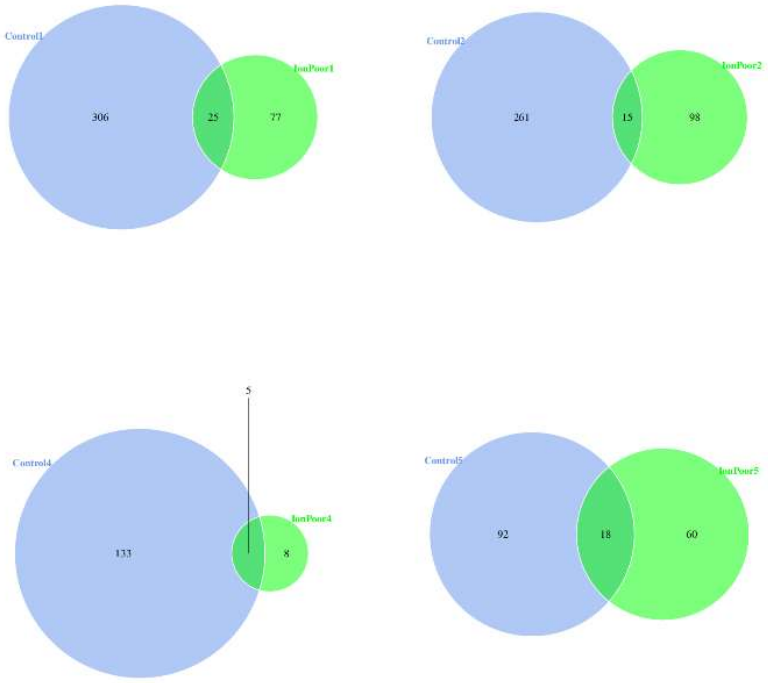


Fig 3.4.1 Venn Diagram

To view full size picture please click here

Note: each circle in the figure represents a sample (group), the number in the overlap of circles represents the number of common ASVs between samples (groups), and the number without overlapping parts represents the number of unique ASVs of the sample (group).

Results directory:

Analysis results: result/03.AlphaDiversity/Venn(_group)

Plotting data: result/03.AlphaDiversity/Venn(_group)/venndata

Novogene Co., Ltd

4 Alpha Diversity

Alpha Diversity is used to analyze the microbial community diversity in the sample (Within-community)[11]. Through the single-sample diversity analysis (Alpha Diversity), the richness and diversity of the microbial community in the sample can be reflected, including the use of a series of statistical analysis indexes, species diversity curves and species accumulation box plots, to evaluate the differences in species richness and diversity of microbial communities in each sample.

4.1 Alpha Diversity Indices

The Alpha Diversity analysis index (observed_otus, shannon, simpson, chao1, goods_coverage, dominance and pielou_e) of different samples are calculated, see the table below.

Table 4.1 Statistics of Alpha Diversity Indices

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

Control2	276.750	0.317	1.000	276	0.352	2.851	0.683
Control4	138.600	0.391	1.000	138	0.294	2.091	0.609
Control5	110.000	0.330	1.000	110	0.369	2.501	0.670
Control6	135.000	0.353	1.000	135	0.339	2.396	0.647
IonPoor1	102.000	0.251	1.000	102	0.408	2.720	0.749
IonPoor2	113.000	0.642	1.000	113	0.201	1.373	0.358
IonPoor4	13.000	0.604	1.000	13	0.303	1.121	0.396
IonPoor5	78.000	0.433	1.000	78	0.267	1.676	0.567

Note: chao1: Estimate the total number of species contained in the community sample. The more low-abundance species in the community, the greater the chao1 index; dominance: the probability when randomly select two sequences from the same sample, the better the community species uniformity, the larger the index; goods_coverage: coverage, the higher the sequencing coverage, the larger the index; observed_otus: the number of species observed directly, the larger the index, the more species are observed; pielou_e: the evenness index, the more even the species, the larger the pielou_e; Shannon: the total number of categories in the sample and their proportions. The higher the community diversity, the more uniform species distribution, and the greater the shannon index; simpson: characterizes the diversity and uniformity of species distribution in the community. The better the species uniformity, the greater the simpson index.

Results directory:

Analysis results: result/03.AlphaDiversity/Alpha_diversity_indices/*qza

Summary table of alpha diversity indices: result/03.AlphaDiversity/Alpha_diversity_indices/alpha_index.xls

4.2 Biodiversity Curves

Rarefaction curve is a common curve that describes the diversity of samples in a group. It randomly extracts a certain amount of sequencing data from the samples and counts their alpha diversity index values to determine the amount of sequencing data extracted and the corresponding index value to build the curve (cutoff=41457).

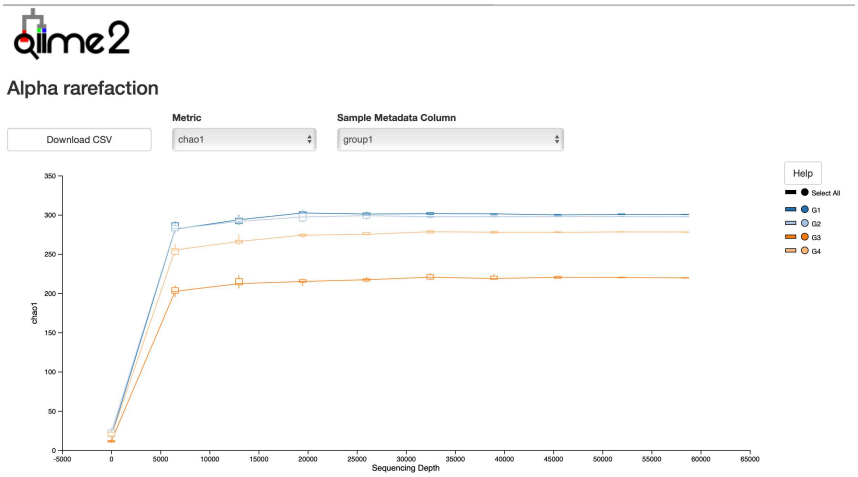


Fig 4.2.1 Rarefaction Curves

To view full size picture please click here

Note: The horizontal axis represents the amount of sequencing data, and the vertical axis represents the corresponding alpha diversity index. When the curve tends to be flat, it indicates that the amount of sequencing data is gradually reasonable, and more data will not have a significant impact on the alpha diversity index.

Species accumulation boxplot is an analysis that the species diversity increases as the sample size increases. It is an effective tool for investigating the species composition of the sample and predicting the species abundance in the sample. It is used in biodiversity and community surveys for judging whether the sample size is sufficient and estimating species richness. Therefore, the species accumulation boxplot can not only judge whether the sample size is sufficient, but also predict species richness under the premise of sufficient sample size (by default, analysis is performed when the sample size is greater than 10).

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

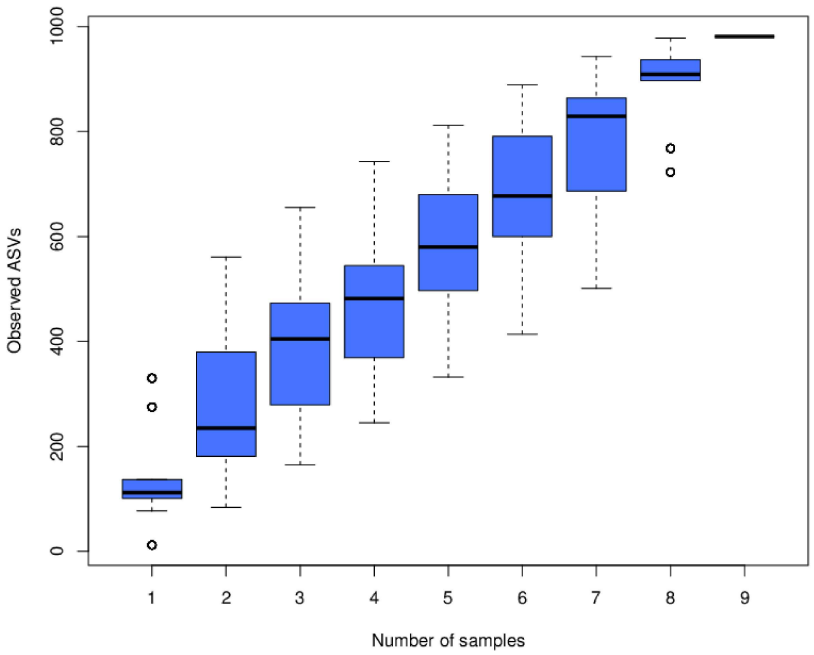


Fig 4.2.2 Species Accumulation Boxplot

To view full size picture please click here

Note: The horizontal axis is the sample size, and the vertical axis is the number of ASVs after sampling. The overall result reflects the rate of emergence of new ASVs under continuous sampling. Within a certain range, as the sample size increases, if the position of the box diagram shows a sharp rise, it means that a large number of species have been discovered in the community; when the position of the box diagram tends to be flat, it means that the species in the environment do not increase significantly as the sample size increases. The species accumulation boxplot can be used as a judgment of whether the sample size is sufficient. A sharp rise in the position of the boxplot indicates that the sample size is insufficient, and the sample size needs to be increased; otherwise, it indicates that the sampling is sufficient and data analysis can be carried out.

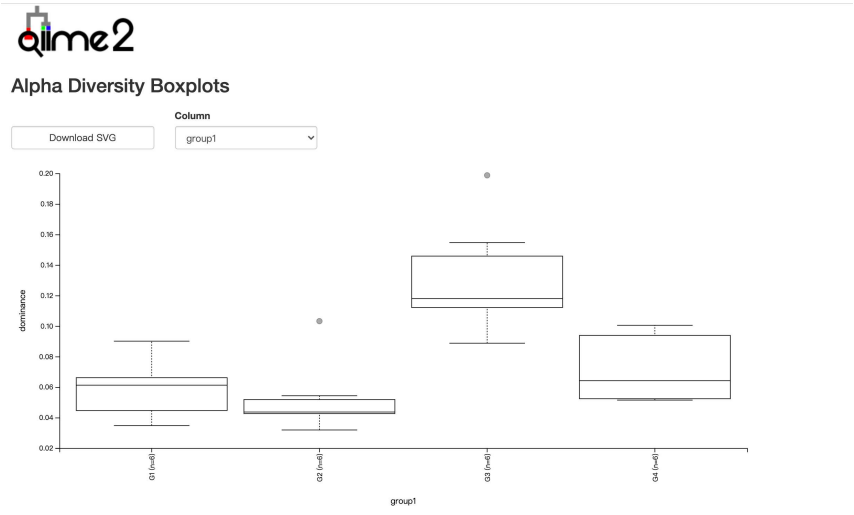
Results directory:

Results directory of rarefaction curves: result/03.AlphaDiversity/Alpha_rarefaction/*qzv

Analysis results of species accumulation boxplot: result/03.AlphaDiversity/Species_accum_box/*.(png,pdf)

4.3 Differential Analysis of Alpha Diversity Indices

In the analysis of differences between groups of the Alpha diversity indices, the boxplot can intuitively reflect the median, dispersion, maximum, minimum, and abnormal values of species diversity within the group. At the same time, Kruskal-Wallis was used to analyze whether the differences in species diversity between groups were significant. The results are as follows:



1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

Note: The horizontal axis of the boxplot represents the group, where n represents the number of samples in the group, and the vertical axis represents the corresponding alpha diversity index value. The table is the Kruskal-Wallis test results for all groups or between two groups.

To view detailed results of Dominance please click here

To view detailed results of Shannon please click here

To view detailed results of Chao1 please click here

To view detailed results of Simpson please click here

To view detailed results of Pielou_e please click here

To view detailed results of Observed_otus please click here

Results directory:

Analysis results: result/03.AlphaDiversity/Alpha_group_significance/*qzv

5 Beta Diversity

Beta diversity represents the explicit comparison of microbial communities based on their composition. Firstly, according to the species annotation results and the abundance information of ASVs of all samples, the species profiling table is obtained by merging the ASVs information of the same classification. The UniFrac distance (Unweighted Unifrac) is calculated by phylogenetic relationship of ASVs[12,13]. UniFrac distance is a method to calculate the distance between samples by using the evolutionary information of microbial sequences in each sample. For more than two samples, a distance matrix is obtained. Then, the Weighted UniFrac distance is obtained by modifying Unweighted Unifrac distance with abundance information of ASVs[14]. Then we analyze differences in samples (groups) with results on the differences of beta diversity index among samples (groups), Principal Component Analysis (PCA), Principal Coordinate Analysis (PCoA), and Non-Metric Multi-Dimensional Scaling (NMDS).

5.1 Beta Diversity Heatmap

In beta diversity study, Weighted UniFrac distance, Unweighted UniFrac distance, Jaccard distance and Bray distance were selected to measure the differences between samples. The smaller the value is, the smaller the differences in species diversity between the two samples are. The Heatmap by Unweighted UniFrac is as follow:

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

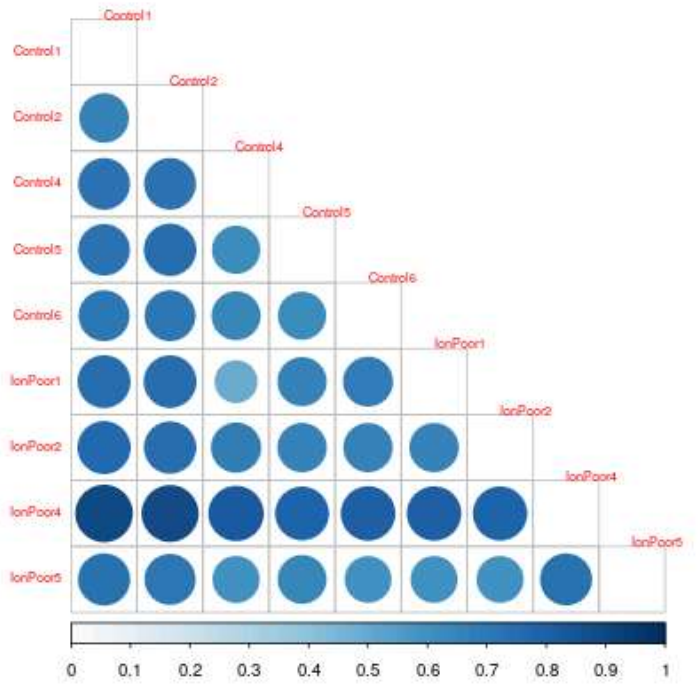


Fig 5.1 Beta Diversity Indices Heatmap of Unweighted Unifrac Distance Matrix

To view full size picture please click here

Note: The size and color of the circle in the square represent the differences coefficient between the two samples. The larger the circle is, the darker the corresponding color is, indicating that the differences between the two samples are greater; on the contrary, the smaller the circle is, the lighter the corresponding color is, indicating that the differences between the two samples are smaller.

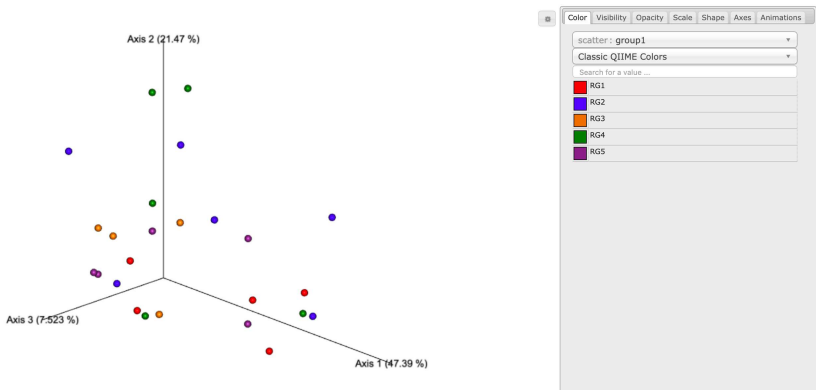
Results directory:

Analysis results: result/04.BetaDiversity/Core_metrics/*(qza,qzv)

5.2 Principal Coordinate Analysis (PCoA)

Principal coordinates analysis (PCoA)[15] is an ordination technique, which picks up the main elements and structures from reduced multi-dimensional data series of eigenvalues and eigenvectors. We use Weighted UniFrac distance and Unweighted UniFrac distance to analyze PCoA, and combine the principal coordinates with the largest contribution rate for mapping. The closer the sample distance is, the more similar the species composition structure is. Therefore, the samples with high similarity of community structure tend to gather together, while the samples with large differences of community structure tend to be far away.

For PCoA, two-dimensional and three-dimensional display forms are adopted. The first and second principal coordinates are selected for two-dimensional PCoA map, while three principal coordinates are selected for three-dimensional PCoA map for interactive web page display, and the coordinates can be flexibly adjusted. The results are as follows:



1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

To view full size picture please click here

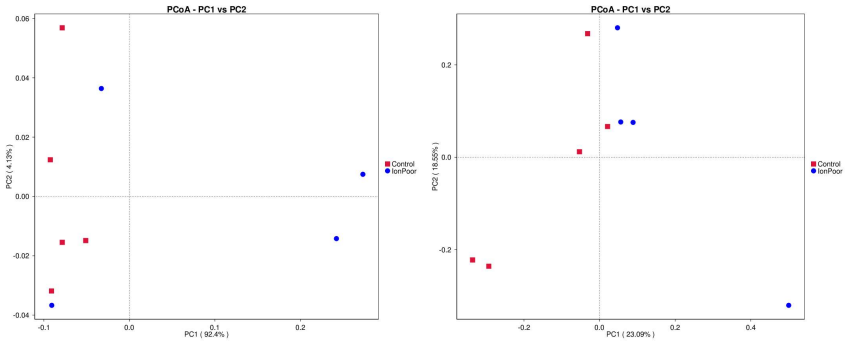


Fig 5.2.2 Two-dimensional PCoA Diagram

To view full size picture please click here

Note: the horizon represents one principal component, the ordinate represents another principal component, and the percentage represents the contribution value of the principal component to the sample differences; each point in the graph represents a sample, and the samples of the same group are represented by the same color.

Results directory:

Analysis results: result/04.BetaDiversity/PCoA/(weighted, unweighted)_unifrac

5.3 Principal Component Analysis (PCA)

Principal component analysis (PCA)[16], based on the relative abundance distribution of ASVs, is a statistical procedure to extract principle components and structures in data by using orthogonal transformation and reducing dimensionalities of data[17]. It extracts the first two axises reflecting the variation of samples to the most extent, which can reflect high-dimensional data's variation in two-dimensional graph, which reveals the simple principle embedding in complex data. The more similar the composition of community among the samples are, the closer the distance of their corresponding data points on the PCA graph are. The result of PCA analysis based on ASVs is shown in figure:

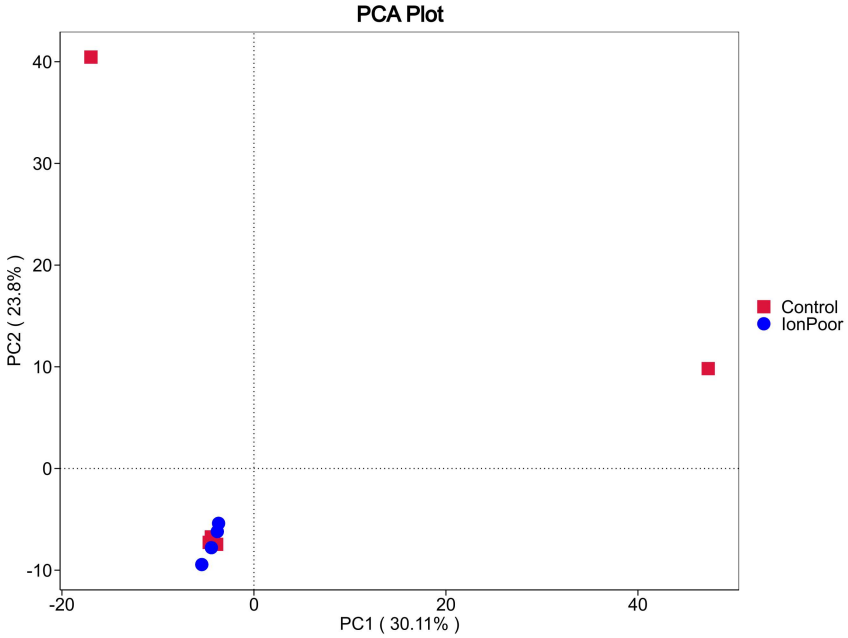


Fig 5.3 PCA

To view full size picture please click here

Notes: The X-axis represents the first principal component, and the percentage represents the contribution value of the first principal component to the sample differences; the Y-axis represents the second principal component, and the percentage represents the contribution value of the second principal component to the sample differences; each point in the graph represents a sample, and the

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

Analysis results: result/04.BetaDiversity/PCA/*(pdf,png)

5.4 Non-Metric Multi-Dimensional Scaling (NMDS)

Non-metric multi-dimensional scaling analysis (NMDS) is a ranking method applicable to ecological researches[18]. NMDS is the non-linear model based on Weighted UniFrac and Unweighted Unifrac, according to the species information contained in the samples, and is reflected on the two-dimensional plane in the form of points. The design overcomes the shortcomings of linear models (including PCA and PCoA) and better reflects the nonlinear structure of ecological data[19]. By using NMDS analysis, the species information contained in the sample is reflected in the multidimensional space in the form of points, while the degree of differences between different species is reflected by the distance between points, and also reflects the inter group and intra group differences of the samples.

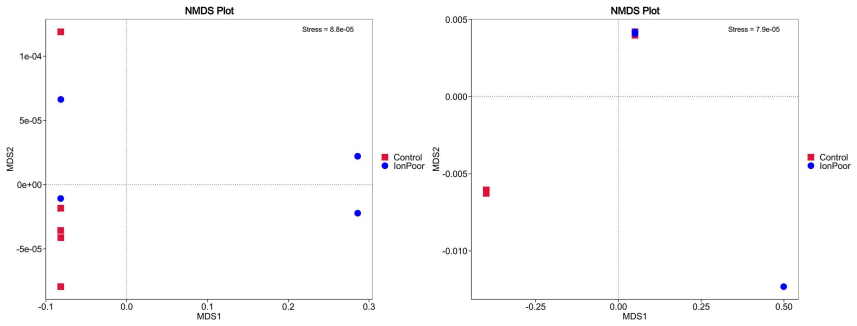


Fig 5.4 NMDS

To view full size picture please click here

Notes: each data point in the graph stands for a sample. The distance between data points reflects the extent of variation. Samples belongs to the same group are in the same color. When the value of Stress factor is less than 0.2, it's considered that NMDS is reliable to some extent.

Results directory:

Analysis results: result/04.BetaDiversity/NMDS/*/*.(png,pdf)

6 Significance Test of Community Structure Difference of Groups

The Adonis and Anosim methods are used for analyzing the differences of community diversity among groups based on the Weighted Unifrac and Unweighted Unifrac distances.

6.1 Adonis

ADONIS[20] is also called permutational MANOVA or nonparametric MANOVA, which is a method of nonparametric multivariate variance test according to distance matrix. This method can analyze the explanation of grouping factor on differences of samples and estimate the significance of grouping by permutation test[21,22,23,24]. The results could be obtained by adonis function in QIIME2 software. The results are as follows:

To view detailed results of weighted_unifrac please click here

To view detailed results of unweighted_unifrac please click here

Notes: Df represents degree of freedom; SumsOfSqs represents sums of squares of deviations; MeanSqs represents SumsOfSqs/Df; F.Model represents F-test value; R2 represents the explanation of grouping factor on differences of samples, calculated from the ratio of grouping variance and total variance; Pr means P-value, whose value less than 0.05 suggests statistical significance. Values in

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

Analysis results: 04.BetaDiversity/Beta_group_significance/adonis/*qzv

6.2 Anosim

Anosim[25] analysis, based on the Unifrac distance, is a nonparametric test to evaluate whether variation among groups is significantly larger than variation within groups, which helps to evaluate the reasonability of the division of groups. The results could be obtained by anosim function in QIIME2 software. The results are as follows:

To view detailed results of weighted_unifrac please click here

To view detailed results of unweighted_unifrac please click here

Notes: The box chart of beta diversity analysis can directly display the median, dispersion, maximum, minimum and abnormal values showing sample similarity within a group.

Results directory:

Analysis results: result/04.BetaDiversity/Beta_group_significance/anosim/*qzv

7 Inter-group Variation Analysis of Species

Statistical analysis of different communities can be performed especially for those projects involving multiple groups. Species with significant differences in abundance among groups were captured and their distribution among groups was obtained.

7.1 T-test

T-test is performed to determine species with significant variation between groups (p value < 0.05) at various taxon ranks including phylum, class, order, family, genus, and species. Result is displayed in phylum or genus by default:

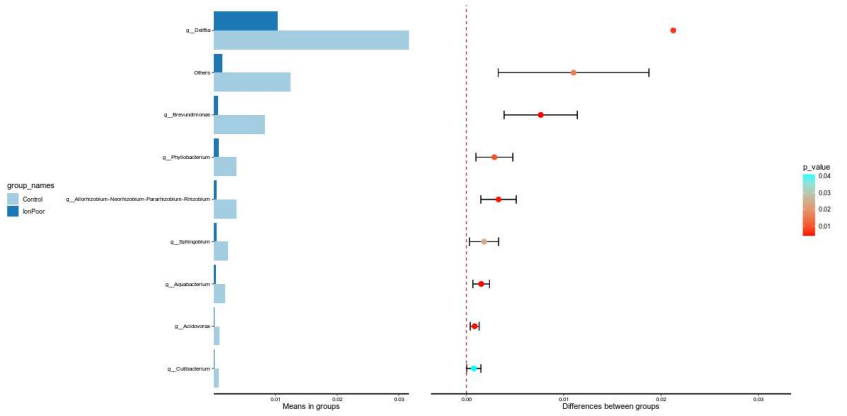


Fig 7.1.1 T-test Analysis of Species Differences between Groups

To view full size picture please click here

Note: the left figure shows the species abundance differences between groups, and each bar in the figure represents the average abuundance of the species in different groups; the right figure shows the confidence of inter-group differences. The leftmost point of each circle in the figure represents the lower limit of 95% confidence interval, and the rightmost point of the circle represents the upper limit of 95% confidence interval. The center of the circle represents the differences of the mean value, and the color of the circle represents the P value of differences significance test between groups of the corresponding species.

The following is the volcano map. By default, the results of the same classification level as above figure are shown:

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

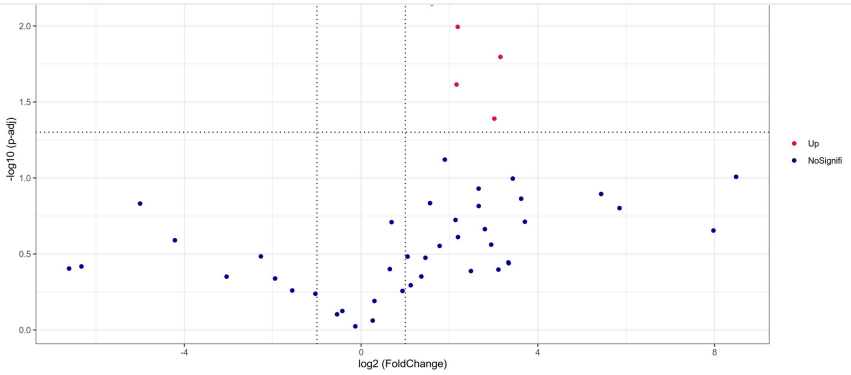


Fig 7.1.2 Volcano Map

To view full size picture please click here

Note: volcano map is a form of scatter diagram, which is usually composed of several parts: significantly up regulating species and significantly down regulating species. Generally, the X-axis is the multiple differences of the species with differences between groups, while the Y-axis is the p value of significance test. Each point in the figure represents a species with differences, where 'Up' represents that the abundance of the species with differences in the first comparison group is higher than that in the second, while 'Down' represents the opposite.

Results directory:

Analysis results: result/04.BetaDiversity/T_test/*/*.(xls,png,pdf)

Species with significant differences: result/04.BetaDiversity/T_test/*/boxplot

7.2 MetaStat

Taxa with significant inter-group variation are detected via metastats, a strict statistical method based on their abundance. The significance of observed abundance's differences among groups is evaluated via multiple hypothesis-test for sparsely-sampled features and false discovery rate (FDR). Result is displayed in phylum or genus by default:

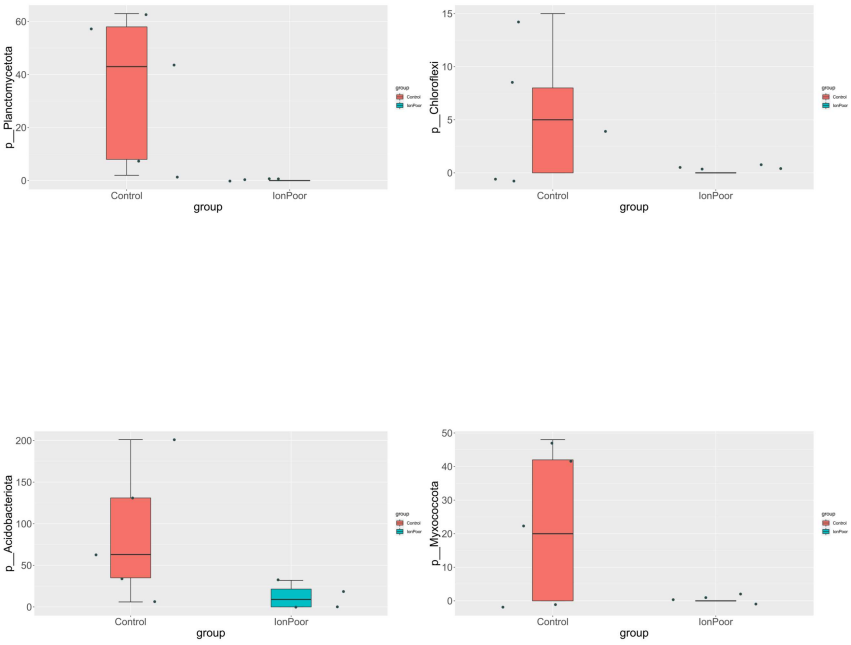


Fig 7.2.1 MetaStat Analysis of Species Differences between Groups

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

Note: The horizontal axis represents the group, the vertical axis represents the absolute abundance of the significant differences species in the sample, each point represents a sample, and the box chart of different colors represents different groups.

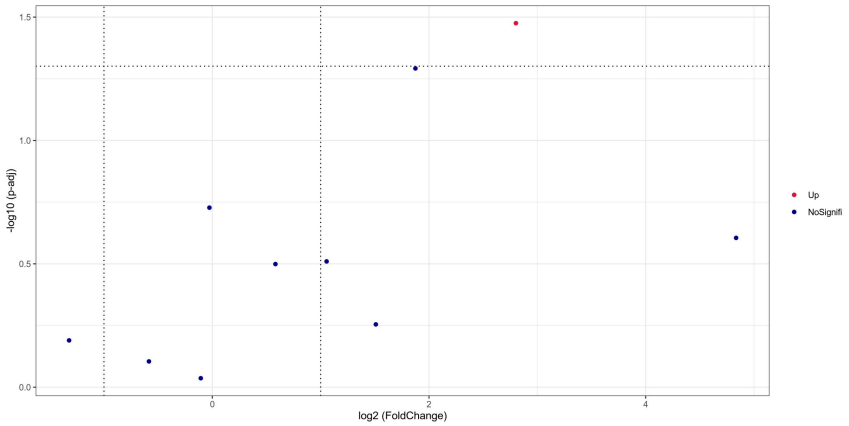


Fig 7.2.2 Volcano Map

To view full size picture please click here

Note: volcano map is a form of scatter diagram, which is usually composed of several parts: significantly up regulating species and significantly down regulating species. Generally, the X-axis is the multiple differences of the species with differences between groups, while the Y-axis is the p value of significance test. Each point in the figure represents a species with differences, where 'Up' represents that the abundance of the species with differences in the first comparison group is higher than that in the second, while 'Down' represents the opposite.

Results directory:

Analysis results: result/04.BetaDiversity/MetaStat/*/*.(xls,png,pdf)

Species with significant differences: result/04.BetaDiversity/MetaStat/*/boxplot

7.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) Effect Size)[26] is a software designed to discover high-dimensional biomarkers and reveal metagenomic characteristics (including genes, metabolites or taxa), so it can be used to distinguish two or more biological classes. It emphasizes statistical significance, biological consistency, and detect statistical significance biomarkers among groups, and allows researchers to identify characteristics of abundance and related classes. The result was composed of histogram of LDA score and the Cladogram among groups. The results are as follows:

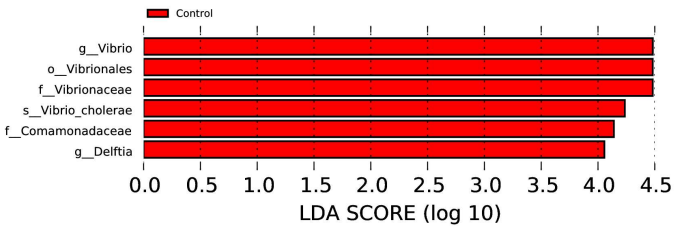


Fig 7.3.1 Histogram of LDA Scores

To view full size picture please click here

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

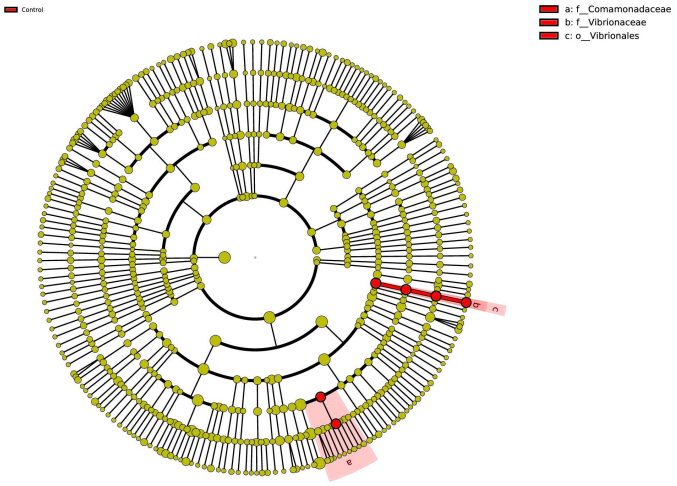


Fig 7.3.2 Cladogram

To view full size picture please click here

Note: The selected criteria is that LDA scores are larger than the set threshold (4 set by default). The length of each bin, namely, the LDA score, represents the effect size (the extent to which a biomarker can explain the differentiating phenotypes among groups). In Cladogram, circles radiating from inner side to outer side represents taxonomic rank from phylum to genus (species). Each circle stands for a distinct taxon at corresponding taxonomic rank. The diameter of each circle represents proportionally the relative abundance of each taxon. Coloring principle: Yellow stands for taxons with non-significant differences; Taxons (biomarkers) with significant differences are colored according to corresponding group's color; Red nodes means these microbiota contributes a lot in the group covered by red color, so do the green nodes. If one group in the figure is missing, it means that there is no species with significant differences in the group. The corresponding species of the letters above the circles are annotated on the right side.

Results directory:

Analysis results: result/04.BetaDiversity/LEfSe/*/*.(res,png,pdf)

Species with significant differences: result/04.BetaDiversity/LEfSe/*/biomarkers_raw_images

8 Function Prediction

PICRUSt2 is a bioinformatics software package for metagenomic function prediction based on marker gene (such as 16S rRNA). For detailed prediction process, please refer to the description on the webpage (see details: https://github.com/picrust/picrust2/wiki). At present, function prediction can be carried out according to 16S sequencing data based on KEGG, COG, PFAM, TIGRFAM, EC, KO databases.

8.1 PCA Analysis of Function Annotation

Principal component analysis (PCA) dimensionality reduction analysis was conducted based on the abundance statistical results of functional annotation for different databases. The more similar the composition of function among the samples are, the closer the distance of their corresponding data points on the PCA graph are. Here, only the prediction results of KO database are shown:

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

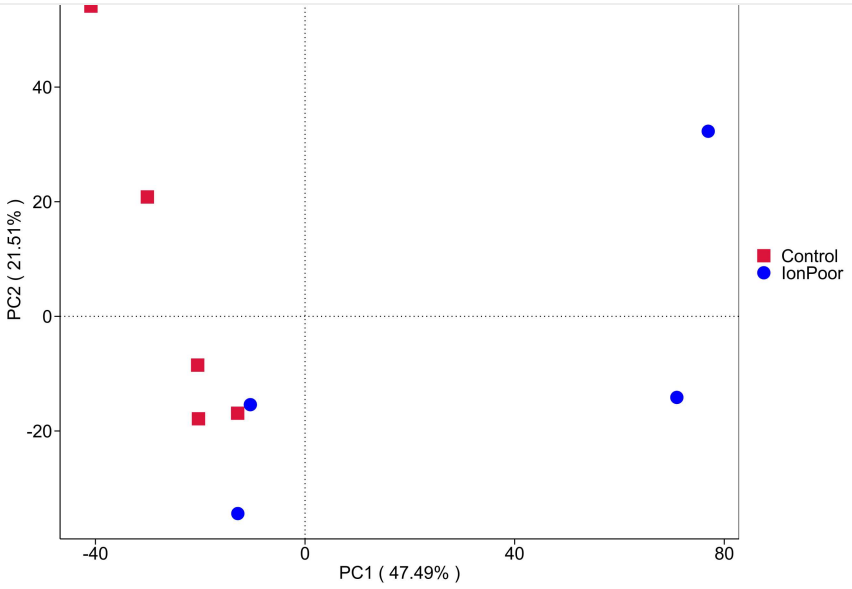


Fig 8.1 PCA Result of PICRUSt2 Function Annotation

To view full size picture please click here

Note: The X-axis represents the first principal component, and the percentage represents the contribution value of the first principal component to the function differences; the Y-axis represents the second principal component, and the percentage represents the contribution value of the second principal component to the function differences; each point in the graph represents a sample, and the samples in the same group are represented by the same color.

Results directory:

Analysis results: result/05.FunctionPrediction/PICRUSt2/picrust2_out_pipeline/*/PCA/*.(png,pdf)

8.2 Venn/Flower Diagram of Function Annotation

In order to investigate the distribution of function number among samples (groups), we analyze the common and unique functions among different samples (groups), the Venn/Flower diagram was drawn. Here, only the prediction results of KO database are shown:

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

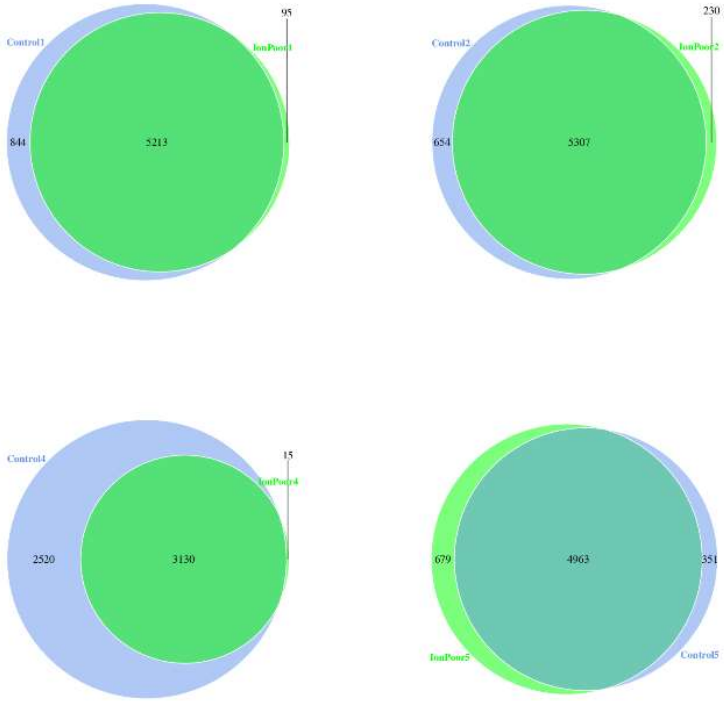


Fig 8.2.1 Venn Diagram of PICRUSt2 Results

To view full size picture please click here

Note: each circle in the figure represents a sample (group), the number in the overlap of circles represents the number of common functions between samples (groups), and the number without overlapping parts represents the number of unique functions of the sample (group).

Results directory:

Analysis results: result/05.FunctionPrediction/PICRUSt2/picrust2_out_pipeline/*/Venn(_group)

Plotting data: result/05.FunctionPrediction/PICRUSt2/picrust2_out_pipeline/*/Venn(_group)/venndata

8.3 T-test of Function Annotation

T-test was used to test the functional differences between different groups. Here, only the prediction results of KO database are shown:

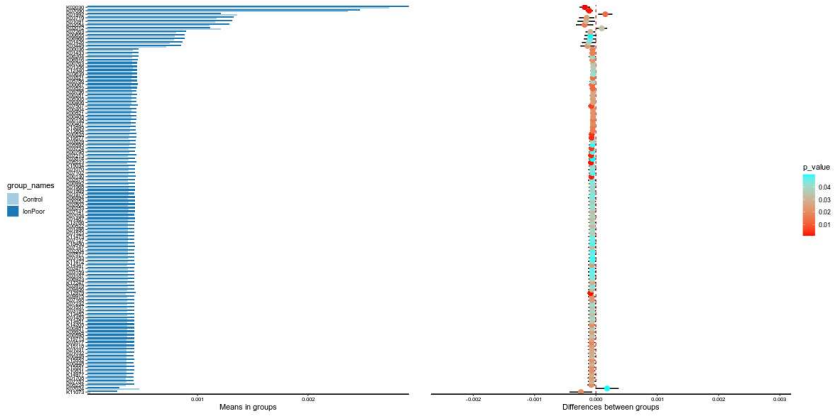


Fig 8.3.1 T-test Analysis of Functional Differences between Groups

To view full size picture please click here

upper limit of 95% confidence interval. The center of the circle represents the differences of the mean value, and the color of the circle represents the P value of differences significance test between groups of the corresponding functions.

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

The volcano diagram is as follows:

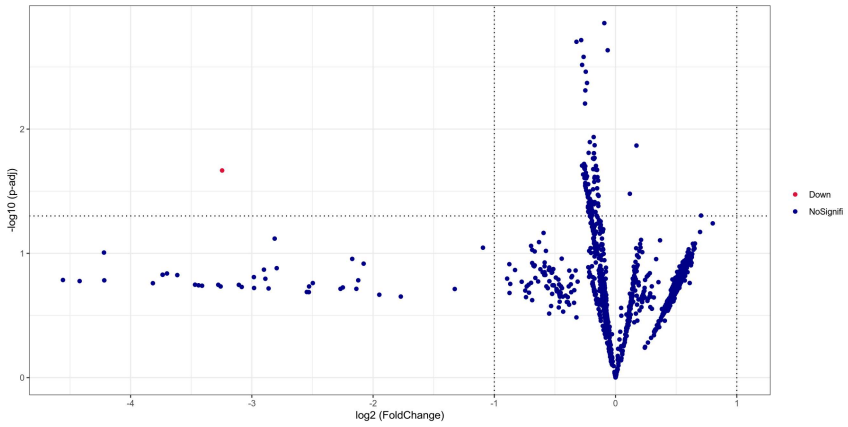


Fig 8.3.2 Volcano Map

To view full size picture please click here

Note: volcano map is a form of scatter diagram, which is usually composed of several parts: significantly up regulating function and significantly down regulating function. Generally, the X-axis is the multiple differences of the functions with differences between groups, while the Y-axis is the P value of significe test. Each point in the figure represents a function with differences, where 'Up' represents that the abundance of the function with differneces in the first comparison group is higher than that in the second, while 'Down' represents the opposite.

Results directory:

Analysis results: result/05.Function_prediction/PICRUSt2/picrust2_out_pipeline/*/T_test/*.(xls,pdf,png)

Species with significant differneces: result/05.Function_prediction/PICRUSt2/picrust2_out_pipeline/*/T_test/boxplot

9 Methods Introduction

9.1 Data Diming

16S rDNA (18S rDNA, ITS) amplicon sequencing is widely used for microbial community comparison among samples from various natural or endozoic environments such as soil, water, host intestine etc. In order to achieve these objectives, several important results need be highly concerned:

(1) ASVs denoise and species annotation results. The representative sequence after denoise process is shown in result/02.ASVanalysis/ASV_table/rep_seqs_qza/ASV-dna-sequences.fasta, and the results of species annotation are shown in 02.ASVanalysis/Seq_taxonomy/seq_taxonomy_qza/taxonomy.tsv.

(2) Species abundance. The species abundance of samples includes absolute abundance and relative abundance, and the downstream analysis of alpha and beta diversity is also based on the table of abundance after standardization. Through species abundance table, we can directly understand the species composition and distribution in the sample, and can also select significantly differences species or target species for further analysis based on their own project background. Taking the “relative” catalogue as an example, the relative abundance of species and the relative abundance of ASV in each sample are included in the catalogue, including seven taxonomic levels of Kingom, Phylum, Class, Order, Family, Genus and Species. For example, to see the relative abundance of Acinomycetaceae in each sample, you can search 'Actinomycetaceae' to query the corresponding relative abundance information in the file 02.ASVanalysis/Taxa_abundance/Relative/asv_table.*.relative.xls.

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

are observed_otus, Shannon, simpson, chao1, goods_ coverage, dominance and pielou_e. The diversity index value can reflect the complexity of the microbial community contained in the sample; in addition, through the alpha diversity index significant differences test results between groups (result/03.AlphaDiversity/Alpha_group_significance), we can quickly find groups with significantly increased or decreased species diversity for further analysis combined with biological treatment.

(4) Beta diversity. Beta diversity can reflect the differences of microbial community structure between samples, so we can further estimate whether the differences is consistent with the biological grouping and reasonably explain the differences combined with biological treatments. For example, the degree of community structure differences between two samples can be seen by the Unifrac distance between them; the differences of community structure can be displayed in two dimensions by PCA, PCoA, NMDS and other dimension reduction maps, so we can judge the community structure differences between groups (samples) according to the aggregation and dispersion between groups (samples).

Furthermore, if there are biological groups, the differences of microbial communities among groups can be further studied by statistical analysis. Through LEfSe analysis, we can find the Biomarker with statistical differences among groups. T-test and MetaStat analysis are carried out at six taxonomic levels of Phylum, Class, Order, Family, Genus and Species, and the species with significant differences between the two groups at different taxonomic levels were obtained. In addition, Adonis analysis can also help to determine whether there are significant differences in community structure among groups.

(5) Customized analysis. PICRUSt2 can be used to predict the function of microbial communities in ecological samples, which will be a great addition to the further understanding of community structure and the publication of articles.

9.2 Methods Description

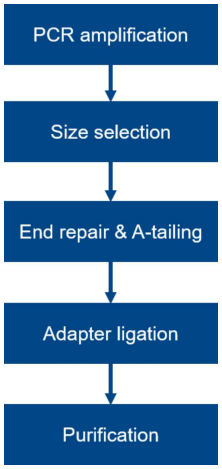
9.2.1 Sequencing

9.2.1.1 Sample Quality Control

Methods of sample quality control refer to QC report.

9.2.1.2 Library Construction, Quality Control and Sequencing

Amplified DNA fragments were end-repaired, and A-tailed. The sequencing adapters were ligated to the ends of the DNA fragments using DNA-binding enzyme, and the DNA fragments were purified using AMpure PB magnetic beads to construct a SMRTbell library. Finally, sequencing primer was annealed to the SMRTbell templates, followed by binding of the sequence polymerase to the annealed templates.The experimental procedures of DNA library preparation are shown as follows:



Workflow of library construction

The library was checked with Qubit for quantification. Quantified libraries will be pooled and sequenced on PacBio Sequel II/IIe systems, according to effective library concentration and data amount required.

9.2.2 Bioinfomatics Analysis

9.2.2.1 Sequencing Data Processing

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

analysis, Beta Diversity analysis, and ANCOM analysis are performed by the Qiime2(202203)

9.2.2.2 Denoise and Species Annotation of ASVs

For the effective tags, the DADA2 or deblur module in QIIME2 software was used to do denoise (DADA2 were used by default), and the sequences with less than 5 abundance were filtered out to obtain the final ASVs (Amplicon Sequence Variables) and feature table. Then, the Classify-sklearn moduler in QIIME2 software was used to compare ASVs with the database and to obtain the species annotation of each ASV.

9.2.2.3 Alpha Diversity

QIIME2 software was used to calculate alpha diversity indices including observed_ otus, shannon, simpson, chao1, goods_ coverage, dominance and pielou_e indices. The rarefaction curve and species accumulation boxplot were drawn. If there was grouping, the differences between groups of alpha diversity would be analyzed by default.

Alpha diversity indices:

Community richness indices:

Observed_otus - the number of observed species

Chao1 - the Chao1 estimator

Dominance - the Dominance index

Community diversity indices:

Shannon - the Shannon index

Simpson - the Simpson index

Index of sequencing depth:

Coverage - the Good's coverage

Index of sequencing uniformity:

Pielou_e - Pielou's evenness index

9.2.2.4 Beta Diversity

Firstly, the UniFrac distance was calculated by QIIME2 software, and the dimensionality reduction maps of PCA, PCoA and NMDS were drawn by R software. Among them, Ade4 and ggplot2 packages in R software were used to display PCA and PCoA. Then, Adonis and Anosim functions in QIIME2 software were used to analyze the significance of community structure differences among groups. Finally, LEfSe or R software was used to perform the species analysis of significant differences between groups. LEfSe analysis was performed by LEfSe software, and LDA score threshold was set to 4 by default. In MetaStat analysis, R software was used to test the differences between the two groups at the level of phylum, class, order, family, genus and species, and P value was obtained. The species with P value less than 0.05 were selected as the significant differences between the two groups; in T-test, R software was also used to analyze the significant differences of species at each taxonomic level.

9.2.2.5 Function Prediction

The full name of PICRUSt2 is Phylogenetic Investigation of Communities by Reconstruction of Unobserved Stats 2. Based on the ASVs tree and gene information of ASVs in Greengene database, the gene functional spectrum of their common ancestor was deduced. At the same time, the gene functional spectrum of other unknown species in Greengene database was deduced, and the gene functional prediction spectrum of the whole spectrum of Archaea and bacteria domain was constructed. Finally, the composition of bacteria group obtained by sequencing was 'mapped' to the database, so as to predict the metabolic function of bacteria group. See the analysis result file for details.

10 Reference

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

[1] Caporaso, J. Gregory, et al. 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 (2011): 4516-4522.

[2] Youssef, Noha, et al. 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 (2009): 5227-5236.

[3] Hess, Matthias, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331.6016 (2011): 463-467.

[4] Li Minjuan,Shao Dantong,Zhou Jiachen et al. Signatures within esophageal microbiota with progression of esophageal squamous cell carcinoma.[J] .Chin J Cancer Res, 2020, 32: 755-767.

[5] Callahan B J, McMurdie P J, Holmes S P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis[J]. The ISME journal, 2017, 11(12): 2639-2643

[6] Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. DADA2: high-resolution sample inference from Illumina amplicon data." Nature methods 13, no. 7 (2016): 581.

[7] Callahan B J, Wong J, Heiner C, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution[J]. Nucleic acids research, 2019, 47(18): e103-e103.

[8] Amir A, McDonald D, Navas-Molina J A, et al. Deblur rapidly resolves single-nucleotide community sequence patterns[J]. MSystems, 2017, 2(2).

[9] Bokulich NA, Kaehler BD, Rideout JR, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME2's q2-feature-classifier plugin. Microbiome. 2018a;6:90.

[10] Bolyen, E., Rideout, J.R., Dillon, M.R. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857 (2019).

[11] Li, Bing, et al. Characterization of tetracycline resistant bacterial community in saline activated sludge using batch stress incubation with high-throughput sequencing analysis. Water research 47.13 (2013): 4207-4216.

[12] Lozupone, Catherine, and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology 71.12 (2005): 8228-8235.

[13] Lozupone, Catherine, et al. UniFrac: an effective distance metric for microbial community comparison. The ISME journal 5.2 (2011): 169.

[14] Lozupone, Catherine A., et al. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology 73.5 (2007): 1576-1585.

[15] Minchin P R. An evaluation of the relative robustness of techniques for ecological ordination[J]. Vegetatio, 1987, 69(1/3):89-107.

[16] Jolliffe I T. Principal component analysis[J]. Journal of Marketing Research, 1986, 87(100):513.

[17] Avershina, Ekaterina, Trine Frisli, and Knut Rudi. De novo Semi-alignment of 16S rRNA Gene Sequences for Deep Phylogenetic Characterization of Next Generation Sequencing Data. Microbes and Environments 28.2 (2013): 211-216.

[18] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method[J]. Psychometrika, 1964, 29(2):115-129.

[19] Magali Noval Rivas, PhD, Oliver T. Burton, et al. A microbita signature associated with experimental food allergy promotes allergic senitization and anaphylaxis. The Journal of Allergy and Clinical Immunology.Volume 131, Issue 1, Pages 201-212, January 2013.

[20] Stat M, Pochon X, Franklin E C, et al. 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[J]. Ecology & Evolution, 2013, 3(5):1317-1329.

[21] Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology, 26: 32-46.

1 Overview
1.1 Experimental Workflow
1.2 Analysis Workflow
2 Data Processing
3 ASV Analysis
4 Alpha Diversity
5 Beta Diversity
6 Significance Test of Community Structure Difference of Groups
7 Inter-group Variation Analysis of Species
8 Function Prediction
9 Methods Introduction
10 Reference
11 Appendix

[23] Warton, D.I., Wright, T.W., Wang, Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3, 89-101.

[24] Zapala, M.A. and N.J. Schork. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences, USA, 103:19430-19435.

[25] M. G. Chapman, A. J. Underwood. Ecological patterns in multivariate assemblages: information and interpretation of negative values in ANOSIM tests[J]. Marine Ecology Progress, 1999, 180(3):257-265.

[26] Segata, Nicola, et al. Metagenomic biomarker discovery and explanation. Genome Biol 12.6 (2011): R60.

[27] Magoč, Tanja, and Steven L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27.21 (2011): 2957-2963.

11 Appendix

1. Result

2. Methods

Notes: Only partial results are shown in the report. For more results, please refer to released data.