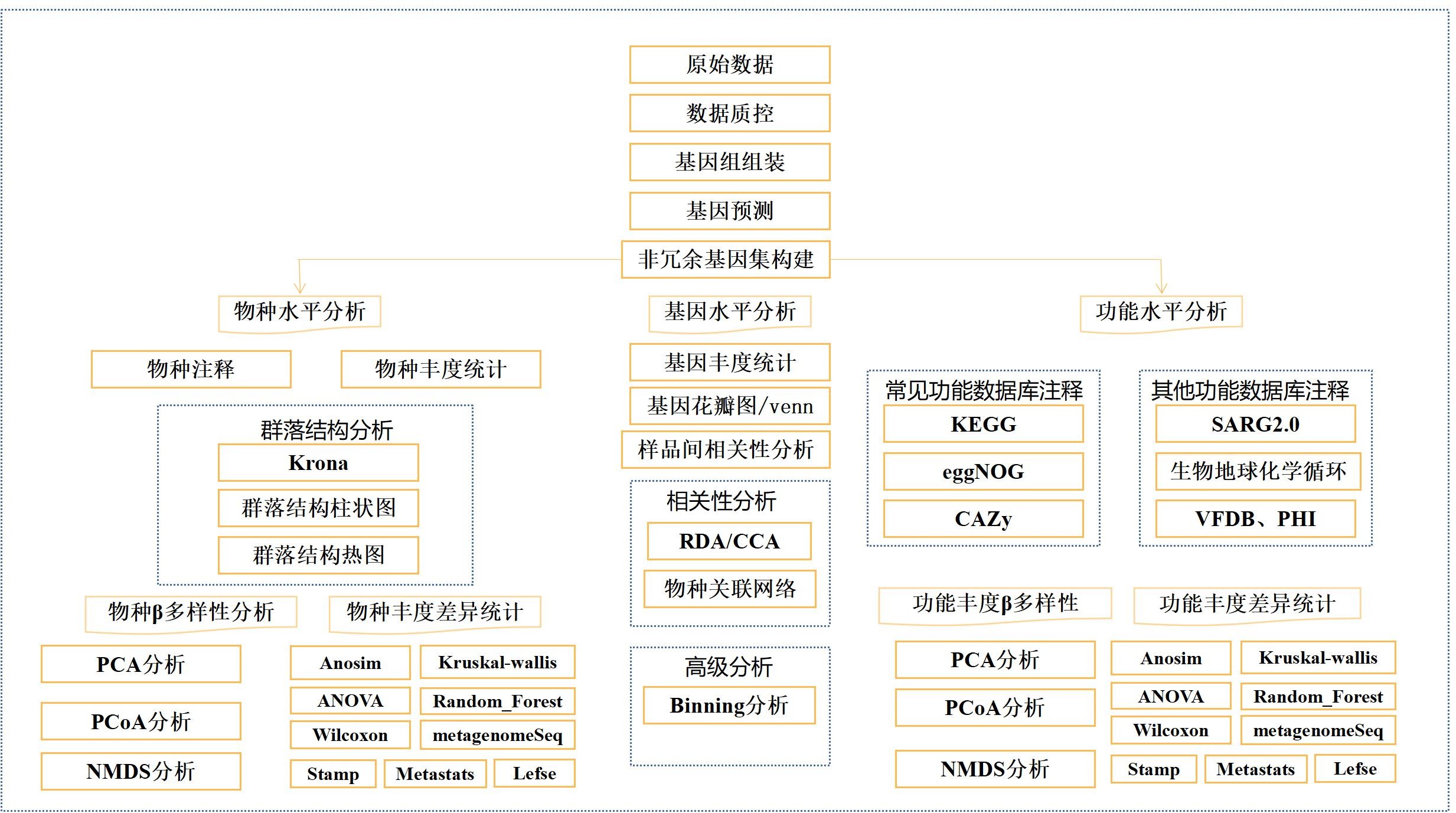
☑Genome Sequencing Project Final Report

**Table of contents**

1. [Bioinformatics Analysis](#_bookmark0)  [4](#_bookmark0)
   1. [Sequencing data statistics](#_bookmark1)  [4](#_bookmark1)
      1. [Sequencing base quality value](#_bookmark2)  [4](#_bookmark2)
      2. [Sequencing base content](#_bookmark3)  [distribution6](#_bookmark3)
      3. [Sequencing quality](#_bookmark4)  [control6](#_bookmark4)
      4. [Sequencing data output](#_bookmark5)  [statistics7](#_bookmark5)
   2. [Data assembly](#_bookmark6)  [8](#_bookmark6)
   3. [Gene prediction and gene set](#_bookmark7)  [construction9](#_bookmark7)
      1. [gene catalogue Basic information statistics](#_bookmark8)  [9](#_bookmark8)
      2. [analysis based on gene abundance11](#_bookmark9)
   4. [Species composition and abundance](#_bookmark10)  [analysis12](#_bookmark10)
      1. [Species abundance cluster heat map](#_bookmark11)  [14](#_bookmark11)
      2. [Species cluster](#_bookmark12)  [analysis15](#_bookmark12)
      3. [Dimensionality reduction analysis based on species](#_bookmark13)  [abundance16](#_bookmark13)
      4. [Association network](#_bookmark14)  [analysis17](#_bookmark14)
   5. [Analysis of intergroup differences based on species composition19](#_bookmark15)
      1. [Based on species abundance Anosim Analysis](#_bookmark16)  [19](#_bookmark16)
      2. [Based on species abundance Kruskal-Wallis Rank Sum Test](#_bookmark17)  [Analysis20](#_bookmark17)
      3. [Based on species abundance anova Analysis](#_bookmark18)  [20](#_bookmark18)
      4. [Based on species abundance Random\_Forest Analysis](#_bookmark19)  [21](#_bookmark19)
      5. [Based on species abundance Wilcoxon Analysis](#_bookmark20)  [21](#_bookmark20)
      6. [Based on species abundance metagenomeSeq Analysis](#_bookmark21)  [22](#_bookmark21)
      7. [Based on species abundance Stamp Analysis](#_bookmark22)  [22](#_bookmark22)
      8. [The differences among species Metastat Analysis](#_bookmark23)  [23](#_bookmark23)
      9. [The differences among species LqCy Analysis](#_bookmark24)  [23](#_bookmark24)
   6. [Functional level analysis](#_bookmark25)  [24](#_bookmark25)
      1. [Basic steps of functional annotation](#_bookmark26)  [25](#_bookmark26)
      2. [Annotation results of each database](#_bookmark27)  [26](#_bookmark27)
      3. [Functional relative abundance profile](#_bookmark28)  [27](#_bookmark28)
      4. [Dimensionality reduction analysis based on functional](#_bookmark29)  [abundance29](#_bookmark29)
   7. [Analysis of intergroup differences based on functional](#_bookmark30)  [abundance31](#_bookmark30)
      1. [Based on functional abundance Anosim Analysis](#_bookmark31)  [31](#_bookmark31)
      2. [Based on functional abundance anova Analysis](#_bookmark32)  [32](#_bookmark32)
      3. [Based on functional abundance Random\_Forest Analysis](#_bookmark33)  [32](#_bookmark33)
      4. [Based on functional abundance Wilcoxon Analysis](#_bookmark34)  [33](#_bookmark34)
      5. [Based on functional abundance metagenomeSeq Analysis](#_bookmark35)  [33](#_bookmark35)
      6. [Based on functional abundance Stamp Analysis](#_bookmark36)  [34](#_bookmark36)
      7. [The difference between the groups Metastat Analysis](#_bookmark37)  [34](#_bookmark37)
   8. [Analysis of biogeochemical cycle functions](#_bookmark38)  [35](#_bookmark38)
      1. [Schematic diagram of the circulation pathway](#_bookmark39)  [35](#_bookmark39)
      2. [Annotation results of each](#_bookmark40)  [database36](#_bookmark40)
      3. [Relative abundance profile of functional genes](#_bookmark41)  [36](#_bookmark41)
      4. [Dimensionality reduction analysis based on functional gene](#_bookmark42)  [abundance38](#_bookmark42)
      5. [Analysis of species contribution to](#_bookmark43)  [genes39](#_bookmark43)
      6. [Analysis of intergroup differences based on functional gene](#_bookmark44)  [abundance40](#_bookmark44)
   9. [Resistance gene annotation](#_bookmark45)  [44](#_bookmark45)
      1. [Relative abundance profile of resistance](#_bookmark46)  [genes44](#_bookmark46)
      2. [Analysis of species contribution to resistance](#_bookmark47)  [genes46](#_bookmark47)
      3. [Analysis of intergroup differences based on resistance gene](#_bookmark48)  [abundance46](#_bookmark48)
   10. [Virulence Factor Analysis](#_bookmark49)  [50](#_bookmark49)
       1. [Relative abundance profile of virulence factor genes](#_bookmark50)  [51](#_bookmark50)
       2. [Dimensionality reduction analysis based on functional](#_bookmark51)  [abundance52](#_bookmark51)
       3. [Analysis of species contribution to](#_bookmark52)  [genes54](#_bookmark52)
       4. [Analysis of intergroup differences based on virulence factor](#_bookmark53)  [abundance54](#_bookmark53)
   11. [Metagenome binning](#_bookmark54)  [60](#_bookmark54)
2. [References](#_bookmark55)  [61](#_bookmark55)

# Bioinformatics analysis



Metagenomic sequencing bioinformatics analysis flow chart

#### The original data is Illumina fastq Format, first of all, the raw data Perform splitting, quality shearing and other operations. If the microorganisms are inside or on the surface of the host, remove the host contamination and other operations at the same time. The obtained sequence is a high-quality sequence, that is, valid data , Use assembly software to assemble high-quality sequences and predict genes . Annotate and classify predicted genes by species and function, including NR , KEGG 、 eggNOG At the same time, the relative abundance of species in each sample was obtained, and on this basis, sample similarity clustering, ranking test, and statistical comparison of differences were performed.

## Sequencing data statistics

### Sequencing base quality value

#### quality value Score or Q-score ) is the base recognition ( Base The commonly used Phred base quality value formula [1] is:



Formula 1 Quality value calculation formula

#### Among them, P is the probability of base recognition error. The following table shows the corresponding relationship between base quality value and the probability of base recognition error:

surface 1 Correspondence table between base quality value and probability of base recognition error

|  |  |  |
| --- | --- | --- |
| **Base quality value** | **Probability of base calling error** | **Base calling accuracy** |
| Q10 | 1/10 | 90% |
| Q20 | 1/100 | 99% |
| Q30 | 1/1000 | 99.9% |
| Q40 | 1/10000 | 99.99% |

#### The higher the base quality value, the more reliable the base call and the higher the accuracy. For example, for a base call with a base quality value of Q20 , 1 in 100 bases will be called incorrectly, and so on.

#### The distribution of base quality values of the sample raw data is shown in the figure below :

Figure 2 Distribution of base sequencing error rates Note : The horizontal axis is the base position of the reads , and the vertical axis is the single-base error rate.

, and samples . Due to the consumption of chemical reagents during the sequencing process , the sequencing error rate will increase with the sequencing sequence . This characteristic is the increase in the length of the head . High-throughput sequencing platform.

### Sequencing base content distribution

#### Base type distribution test is used to detect the presence of AT and GC Separation phenomenon : Since the sequence being measured is a randomly interrupted DNA fragment, due to random interruption and the principle of base complementary pairing, theoretically, the content of G and C , A and T should be equal in each sequencing cycle, and the entire sequencing process is stable and unchanged, showing a horizontal line . The first few bases at the 5 ' end are random primer sequences with a certain preference, so there will be a large fluctuation at the front end in the base distribution map.

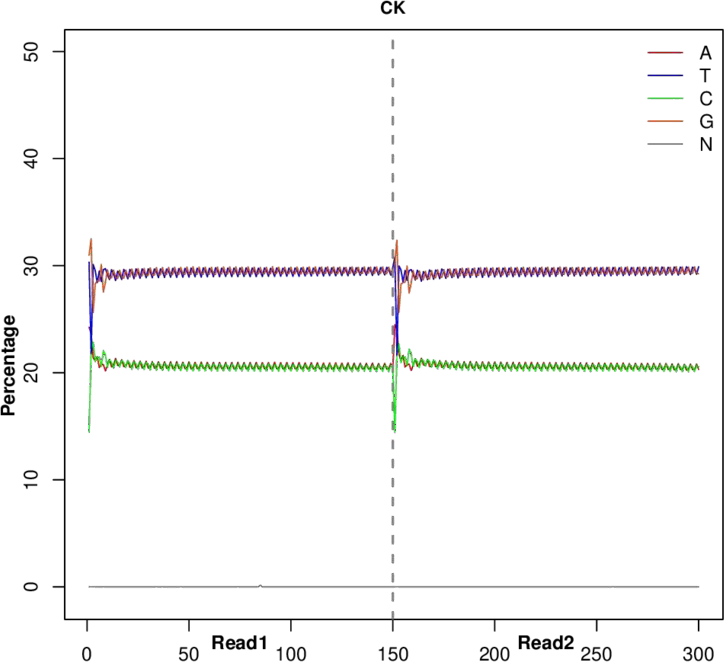


Figure 3 ATGC Content distribution diagram

Note: The horizontal axis is Reads The vertical axis represents the base position, and the vertical axis represents the proportion of single bases.

### Sequencing quality control

Before conducting data analysis, you first need to ensure that Reads The data is of high enough quality to ensure the accuracy of subsequent analysis . The data is strictly quality controlled and filtered as follows:

* + - 1. Remove the connector Reads ;
      2. Remove low-quality reads ( including removing N Reads with a ratio greater than 10% were removed; reads with a quality value of Q≤10 were removed Reads whose base number accounts for more than 50% of the entire read ).

After the above series of quality control, the high quality Clean Data , with FASTQ Format provided.

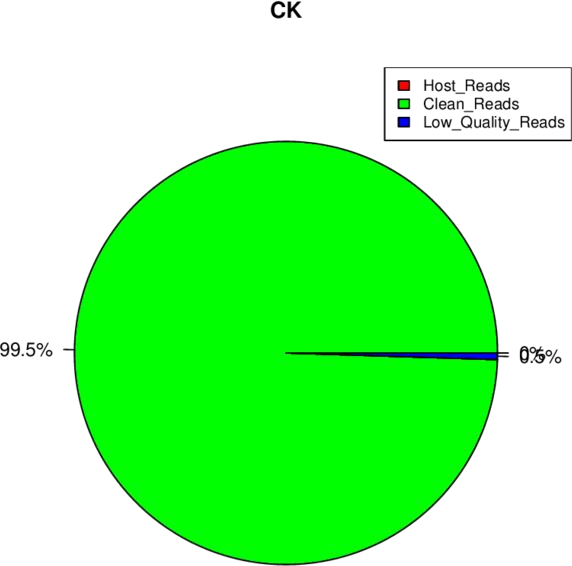


Figure 4 Raw data composition

Note: Adapter Related : Contains connector Reads Number of Raw Reads The ratio of the number . quality : Low- quality reads filtered out Number of Raw Reads The ratio of the number . Reads : Clean after filtering Reads Number of total Raw Reads The ratio of numbers.

### Sequencing data output statistics

#### The output statistics of each sample data of this project are shown in the following table:

Table 2 Sequencing data statistics table

##### Total

##### Clean

##### GC

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Samples Clean Reads bases( %>Q20** | | | | | | **%>Q30** | **Samples** |
|  | **Reads** |  | **GB)** | **Content** |  |  |  |
| **CK** | **37,529,490** | **37,342,294(99.5)** | **5.2078** | **58.74%** | **97.64%** | **93.15%** | **CK** |
| **eTmin** | **41,333,346** | **41,073,868(99.37)** | **5.7285** | **54.9%** | **97.97%** | **94.04%** | **eTmin** |
| **fTmin** | **37,089,744** | **36,873,118(99.42)** | **5.1424** | **57.76%** | **97.61%** | **93.1%** | **fTmin** |
| **tTmin** | **47,615,070** | **47,237,546(99.21)** | **6.5888** | **62.14%** | **98.01%** | **94.37%** | **tTmin** |

Note : Samples : Sample name in sample information sheet ; reads:Clean Data Middle pair-end Reads Total ; Clean bases : the total number of bases in Clean Data ; GC content : the GC content of Clean Data , that is, the percentage of G and C bases in the Clean Data in the total bases; ≥ Q30% : the percentage of bases with Clean Data quality values greater than or equal to 30 .

## Data assembly

#### After the sequence has passed quality control, in order to obtain more effective genomic information and perform more accurate functional annotation analysis on the sequence , it is necessary to perform metagenomic assembly analysis.

#### reads from all samples and use a De-Brujin graph- based The assembly software MEGAHIT ( or IDBA\_UD ) [2][3 ] based on the kmer The overlap relationship between them is constructed to construct the De-Brujin graph and obtain contigs , screening 800 bp Contigs above Statistics were performed and used for subsequent analysis.

#### Based on the assembly results, count the number of Contigs in each sample The distribution of length is plotted and the results are shown in the following figure:

#### Figure 5 Contigs Length distribution of contigs in the sample The length information is counted, the table is as follows: Contigs Length information 2-Assembly/contigs\_length.xls

#### Contigs​ The detailed information is summarized as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **sample num\_contigs total\_length(bp) min\_le** | | | | **max\_len** | **average\_le N50** | |
|  |  |  | **ngth** | **gth** | **ngth** |  |
| **CK** | **149,191** | **228,212,285** | **500** | **1,208,425** | **1,529.7** | **2,239** |
| **tTmin** | **127,050** | **185,547,793** | **500** | **1,033,918** | **1,460.4** | **1,943** |
| **fTmin** | **136,583** | **202,306,528** | **500** | **451,182** | **1,481.2** | **2,038** |
| **eT** | **155,280** | **223,585,856** | **500** | **322,070** | **1,439.9** | **1,807** |

## Gene prediction and gene set construction

#### Based on the genome assembly, gene prediction analysis is performed, and subsequent gene set construction and functional annotation analysis are performed on the predicted genes.

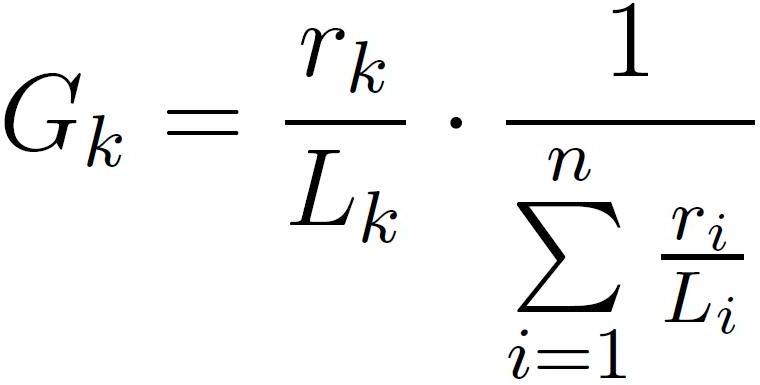
#### [1]. Use Prodigal [ 4 ] software to assemble the contigs ORF​ Predict and translate it into amino acid sequence.

#### [2] ORFs of each sample and mixed assembly Prediction results, using CD-HIT The software performs redundancy removal to obtain non-redundant initial genes catalogue (Here, the non-redundant continuous gene-encoded nucleic acid sequences are called genes ), by default, clustering is performed with identity 95 % and coverage 90 % , and the longest sequence is selected as the representative sequence.

#### [3]. Using bowtie2 Software, clean each sample separately reads Comparison with non-redundant gene sets

#### （ 95 % identity ) , count the abundance information of genes in the corresponding samples.

[4]. From the aligned reads Starting from the number and gene length, the abundance information of each gene in each sample is calculated . The calculation formula [5] [6] [7] [8] [9] is as follows:



Description : To align the reads of the gene Number, L is the length of the gene

### gene catalogue Basic information statistics

#### Using Python Gene​ catalogue The length of is counted, and the results are as follows:

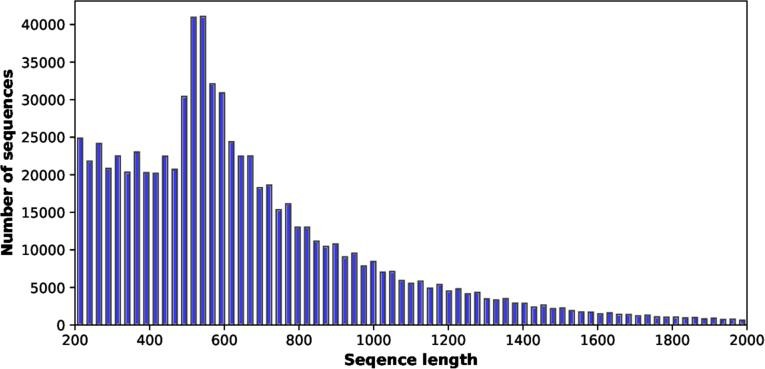


Figure 6 gene catalogue Length distribution

|  |  |  |
| --- | --- | --- |
| Table 3.1 gene catalogue | Length statistics table |  |
| **Length(>)** |  | **Count** |
| 200 |  | 91641 |
| 300 |  | 83613 |
| 400 |  | 85237 |
| 500 |  | 149577 |
| 600 |  | 89727 |
| 700 |  | 62701 |
| 800 |  | 46588 |
| 900 |  | 34746 |
| 1000 |  | 26468 |
| 1100 |  | 21147 |
| 1200 |  | 16836 |
| 1300 |  | 12790 |
| 1400 |  | 9847 |
| 1500 |  | 7148 |

1600 39972

#### Gene​ catalogue The basic information is statistically analyzed in the following table:

**ID num\_genes total\_length(bp) min\_length max\_length average\_length N50**

unique\_gene 770,642 537,962,910 201 25,533 698.1 780

### Correlation between samples based on gene abundance and Wayne analysis

#### The correlation of gene abundance between samples is an important indicator to test whether the selection of experimental samples is reasonable. The closer the correlation coefficient is to 1 , the higher the similarity of gene abundance patterns between samples. The clustering display results are as follows:

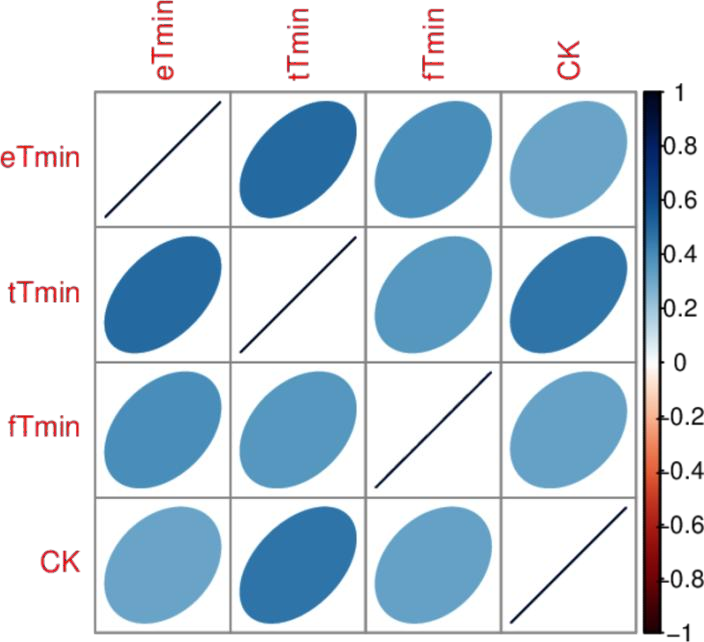


Figure 7 Heat map of correlation coefficients between samples

Image description: Different colors represent spearman The correlation coefficient is high or low; the relationship between the correlation coefficient and color is shown in the legend on the right; the darker the color, the greater the correlation coefficient between samples.

#### In order to examine the distribution of gene numbers between specified samples (groups) and analyze the common and unique information of genes between different samples (groups) , a Venn Graph or petal graph was drawn , and the results are shown as follows:

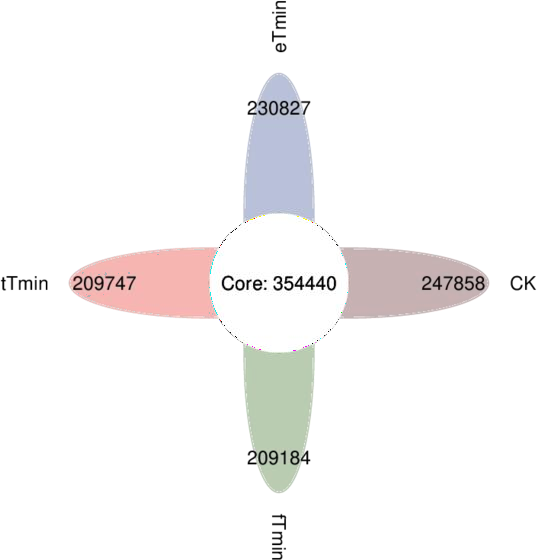


Figure 8 Gene number Venn diagram (petal diagram) analysis

## Species composition and abundance analysis

#### Using genes Compare with various functional databases . DIAMOND software [26 ] converts Unigenes NCBI​ Bacteria , fungi , archaea and viruses sequences extracted from the NR (Version: 2018.01) database were compared ( blastp , evalue ≤ 1e - 5 ) [6 ] ;

#### Alignment result filtering: For each sequence alignment result, select evalue <= minimum evalue\*10

#### The comparison results are then analyzed;

#### After filtering, since each sequence may have multiple alignment results and obtain multiple different species classification information , in order to ensure its biological significance, LCA is used The algorithm ( applied to the systematic classification of MEGAN [27 ] software ) takes the taxonomic level before the first branch as the species annotation information of the sequence;

#### From LCA Based on the annotation results and gene abundance table, we obtain the abundance information of each sample at each taxonomic level (kingdom , phylum, class, order, family , genus, species). The abundance of a species in a sample is equal to the sum of the abundances of genes annotated to that species [7,11,17] ;

#### From LCA Based on the annotation results and gene abundance table, we obtained a table of the number of genes at each taxonomic level (kingdom, phylum, class, order, family, genus, species) for each sample. The number of genes of a species in a sample is equal to the number of genes annotated as that species whose abundance is not 0. The number of genes;

#### Starting from the abundance tables at each taxonomic level (kingdom, phylum, class, order, family, genus, species) , Krona Analysis, relative abundance profile display, abundance clustering heat map display, PCA and NMDS Dimensionality reduction analysis, Anosim Analysis of differences between (and within) groups , Metastat and LEfSe multivariate statistical analysis of species with differences between groups.

* + 1. **Annotation analysis**

#### In order to comprehensively and intuitively display the relative abundance of species at different taxonomic levels in each sample, we used

#### Krona [ 11 ] visualizes the species annotation results . The example diagram is shown below:

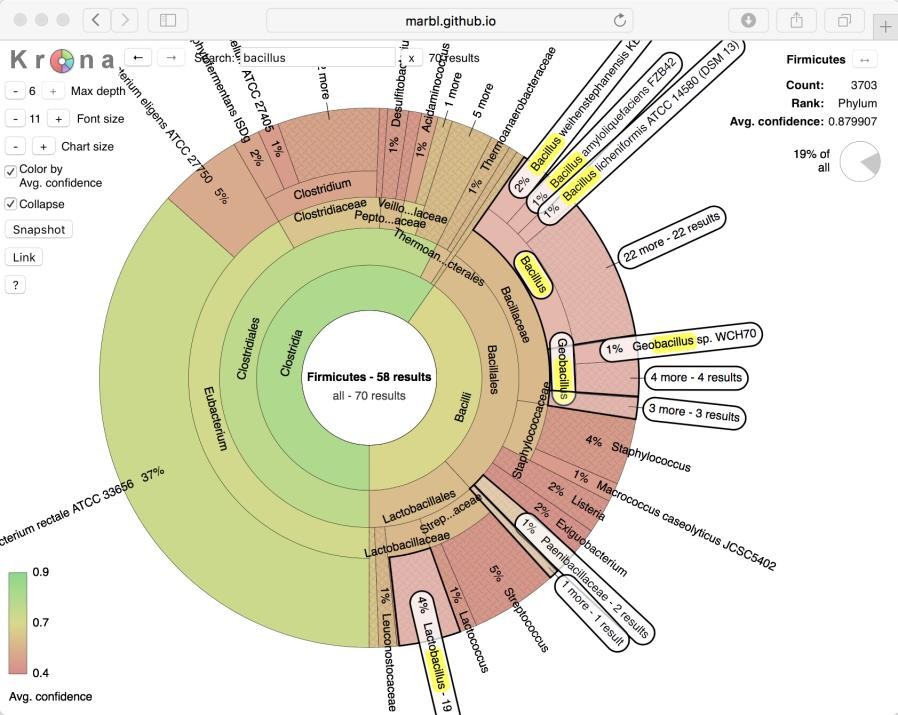


Figure 4.1 Using Krona Display of species annotation results (example image )

* + 1. **Species composition analysis**

According to the classification results, the classification results are classified into 4 Classes : Archaea , Bacteria , Fungi , and Virus . The table file is as follows :

The bar chart shows the species bar chart at the taxonomic level, from which we can intuitively see the species composition of each sample and the proportion of different species in each sample.

Using R Software, the dominant species in each sample ( here we take the top 20 in terms of overall abundance) The composition of species at each classification level is plotted in a bar graph, as shown below:

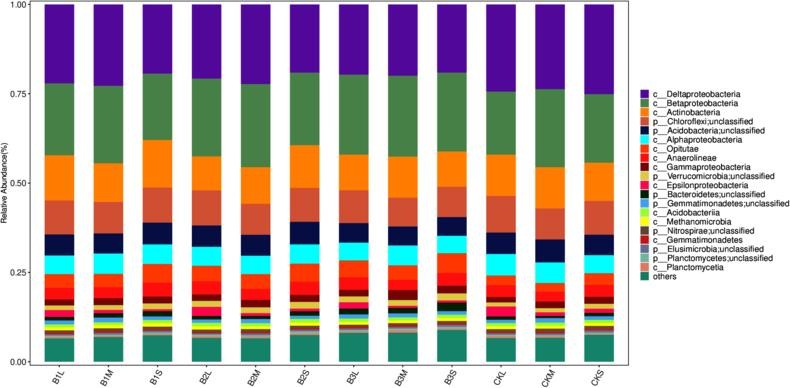


Figure 4.2.1 Species distribution histogram

Note: Kingdom , Phylum , Class , Order , Family , Genus , Species .

### Species abundance cluster heat map

Heat map​ It is a graphical display method that uses color gradients to represent the size of the values in the data matrix and clusters species or samples according to their abundance similarities . 5 0 The similarities and differences of the community composition of multiple samples are reflected by color gradients and similarity. Species heat map analysis is performed based on the species composition and relative abundance of each sample , and species at each taxonomic level are extracted. R The Euclidean distance between samples and species was calculated at the phylum, class, order, family, genus , and species levels , and then hierarchical clustering was performed . In the heat map clustering results, colors represent species abundance; vertical clustering indicates the similarity of abundance of different species among samples . The closer the distance between two species , the shorter the branch length, indicating that the abundance of the two species among samples is more similar; horizontal clustering indicates the similarity of abundance of each species in different samples. Like vertical clustering , the closer the distance between two samples, the shorter the branch length, indicating that the abundance of each species in the two samples is more similar.

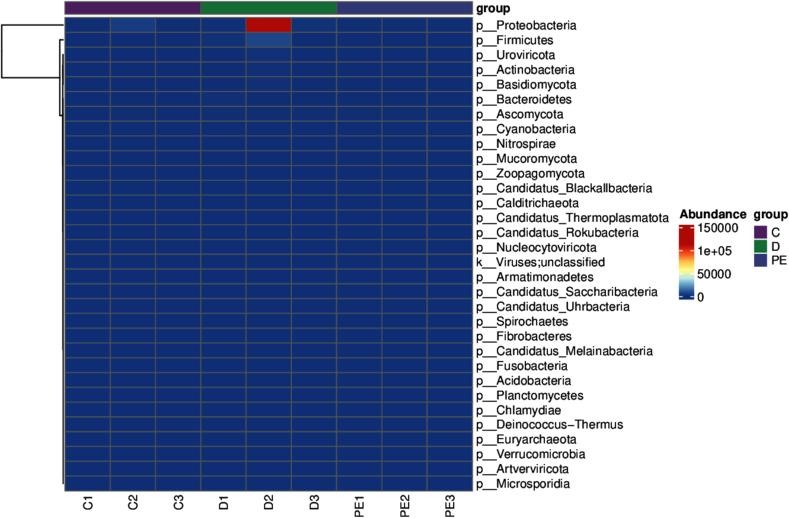
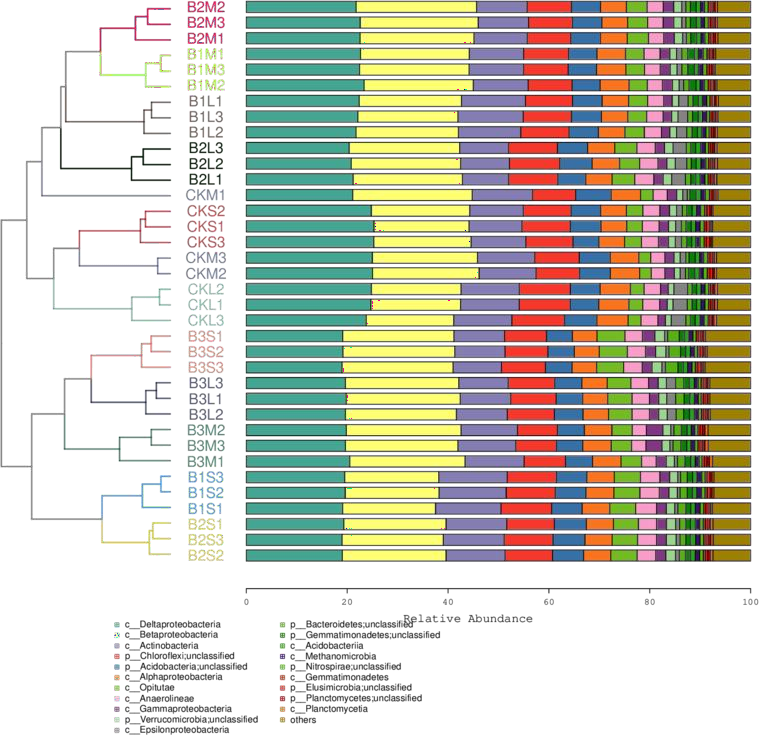


Figure 4.3.1 Species abundance cluster heat map

Note: Kingdom , Phylum , Class , Order , Family , Genus , Species .

### Species cluster analysis

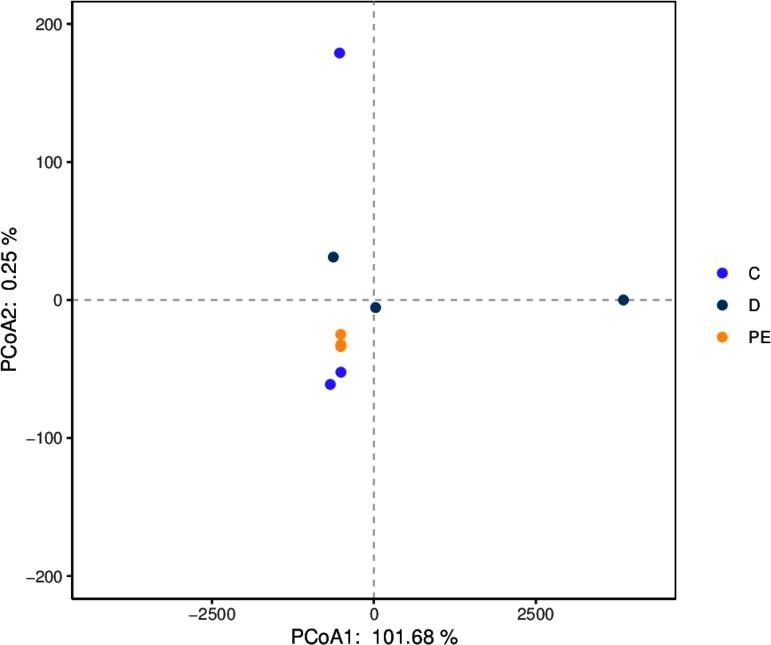


picture 4.4.1 Bray Cluster analysis and species composition stacking plot

Note: Kingdom , Phylum , Class , Order , Family , Genus , Species .

### Dimensionality reduction analysis based on species abundance

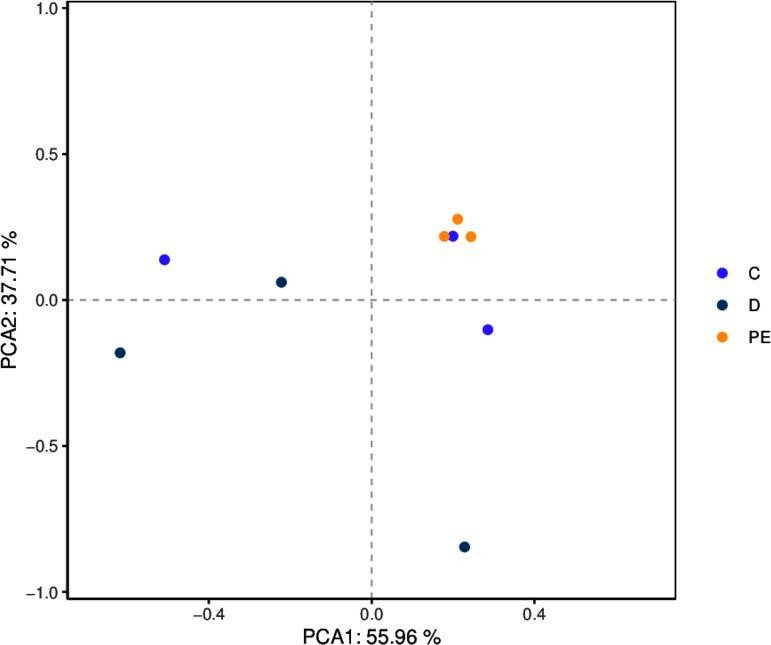
Principal coordinates analysis ( PCoA ) is a classic multidimensional Scaling , cMDScale) analysis method (Ramette, 2007) . It expands the sample distance matrix in a low-dimensional space after projection, and retains the distance relationship of the original sample to the maximum extent. PCoA considers the sample distance as a whole. Compared with principal component analysis ( PCA ) , it is more in line with the characteristics of ecological data. Therefore, it is more recommended as a sorting analysis method.



* + 1. PCoA Analysis chart

Note: The dots represent the species composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample differences; the vertical axis represents the second principal component and its contribution to the sample differences.

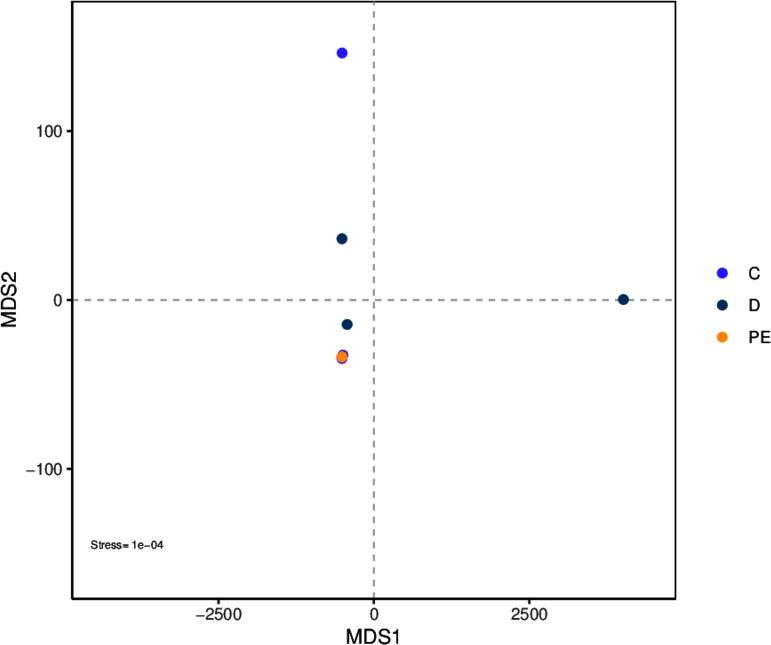
Principal component analysis​​​​​​ Component​​​​​​ An a l y s is ,​ PCA ) is a technique for analyzing and simplifying data sets. By decomposing the variance, the differences between multiple groups of data are reflected on a two-dimensional coordinate graph. The coordinate axes take the two eigenvalues that can best reflect the variance . By analyzing the composition of different samples ( 97 % similarity ), the differences and distances between samples can be reflected. The closer the distance between two samples on the graph, the more similar the species composition in the two samples is. R Language tools plot different classification levels separately P C A Analysis diagram, the sample diagram is as follows:



* + 1. PCA Analysis chart

Note: The dots represent the species composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample differences; the vertical axis represents the second principal component and its contribution to the sample differences.

Non-metric multidimensional scaling (N M D S ) and the above P C o A The analysis is similar to that of the previous one, which also reduces the dimension of the sample distance matrix and simplifies the data structure, thereby describing the distribution characteristics of the samples at a specific distance scale. P C o A Different analysis, N M D S The analysis does not rely on the calculation of eigenvalues and eigenvectors, but rather on ranking the sample distances so that the order of samples in the low-dimensional space is as close as possible to the distance between each other ( rather than the exact distance value ) . The analysis is not affected by the numerical value of the sample distance, and only considers the size relationship between each other. For data with complex structures, the sorting results may be more stable. N M D S The smaller the stress value ( S t ress ) of the result , the better. It is generally believed that when the value is less than 0.2​ When N M D S The results of the analysis are relatively reliable ( L e g e nd re, 199 8 ) .



* + 1. NMDS Analysis chart

Note: The dots represent the species composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample differences; the vertical axis represents the second principal component and its contribution to the sample differences.

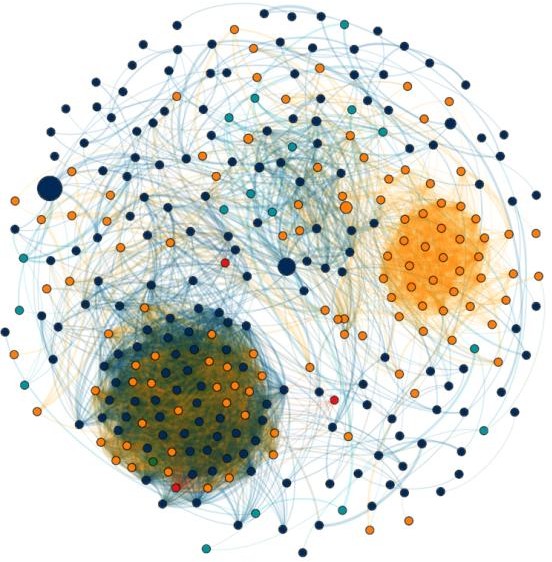
### Association network analysis

Network inference analysis based on the relationships between microbial members Analysis is also a common method for analyzing microbial communities (Faust and Raes, 2012) . The fundamental purpose of this type of analysis is to find specific

Co-occurrence or mutual exclusion of microbial communities driven by temporal and spatial changes and environmental processes

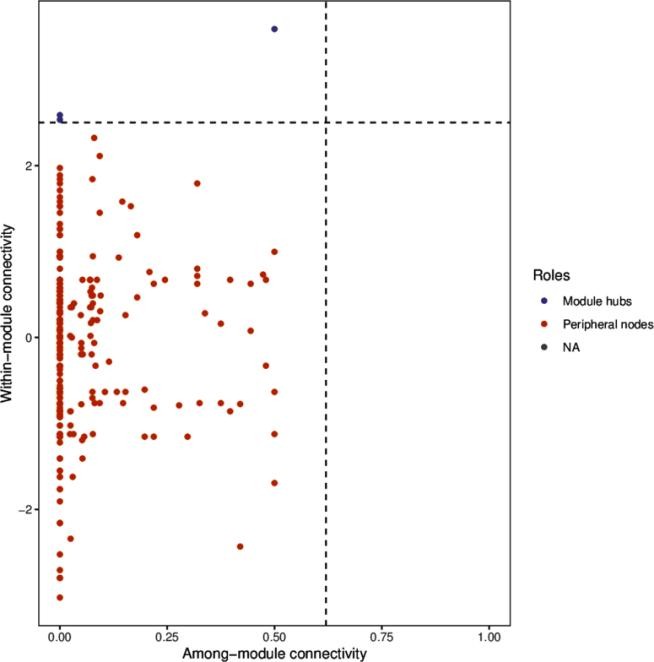
co-exclusion is explored to explore the following scientific questions: 1 ) Does the microbial community have specific modular units to complete specific ecological functions? 2 ) Are there keystone species whose changes are sufficient to leverage the composition changes of the entire community? 3 ) Do different environmental factors and environmental processes lead to differences in the species assembly of the community?

Use random matrix theory to determine the filtering threshold of correlation values, and then use igraph Build an association network.



* 1. network Analysis chart

Based on the network module, two important node characteristics are derived, the within-module connectivity (Within-module connectivity , Zi) and the between-module connectivity (Among-module connectivity , Pi) , which are calculated as follows (Guimerà and Nunes 2005) . According to the topological characteristics of the node, the node attributes can be divided into four types, including: Module hubs ( module center points, nodes with high connectivity within the module, Zi > 2.5 and Pi < 0.62 ), Connectors (connection nodes, nodes with high connectivity between two modules, Zi < 2.5 and Pi > 0.62 ), Network Hubs (network center, nodes with high connectivity in the entire network, Zi > 2.5 and Pi > 0.62 ) and Peripherals (peripheral nodes, nodes that do not have high connectivity within or between modules, Zi < 2.5 and Pi < 0.62 )



* 1. Node connectivity

Note: The horizontal axis is the connectivity between modules; the vertical axis is the connectivity within the module. Points of different colors represent nodes with different attributes .

## Analysis of intergroup differences based on species composition

### Based on species abundance Anosim analyze

A no s i m Analysis is a nonparametric test used to test whether the difference between groups is significantly greater than the difference within the group, so as to determine whether the grouping is meaningful. The detailed calculation process can be viewed A no s i m . Based on species level A no s i m The analysis results are as follows:

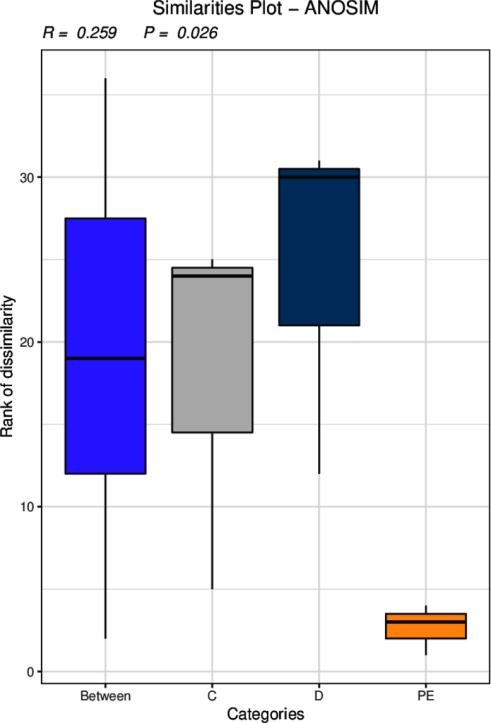


Figure 5.1 Anosim based on species level analyze

### Based on species abundance Kruskal-Wallis Rank Sum Test Analysis

Kruskal-Wallis The rank sum test is a nonparametric test for three or more groups of data that can calculate the existence of differences between the means of multiple independent groups of data. It only deals with the volatility of the data displayed on the graph, which is different from ANOVA. Different from the so-called non-parametric test, no assumptions are made about the data distribution. Through the Kruskal-Wallis rank sum test, we can get whether some OTUs at the OTU level are different, or whether some bacterial species are different in taxonomy, from multiple sets of data with multiple biological replicates. Here, we consider that p\_value less than 0.05 is considered to be different.

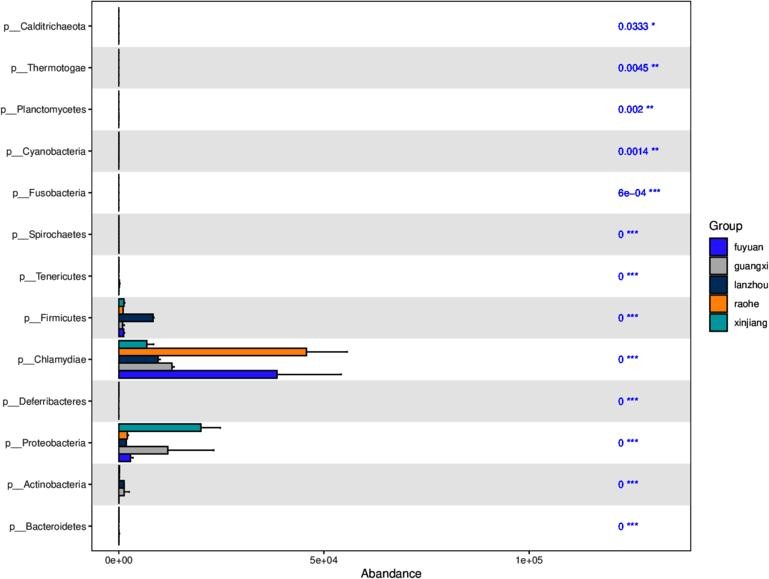


Figure 5.2 Kruskal-Wallis based on gate level analyze

Note: The horizontal axis of the bar graph represents the relative abundance of a species in different groups, the vertical axis represents the differential species, and different colors represent different groups.

### anova based on species abundance analyze

ANOVA (Analysis of variance) is abbreviated as analysis of variance or " coefficient of variation analysis " . It is used to test the significance of the difference between two or more sample means. It is mainly used to test whether the difference between two or more sample means is statistically significant or to test whether there is interaction between two or more factors.

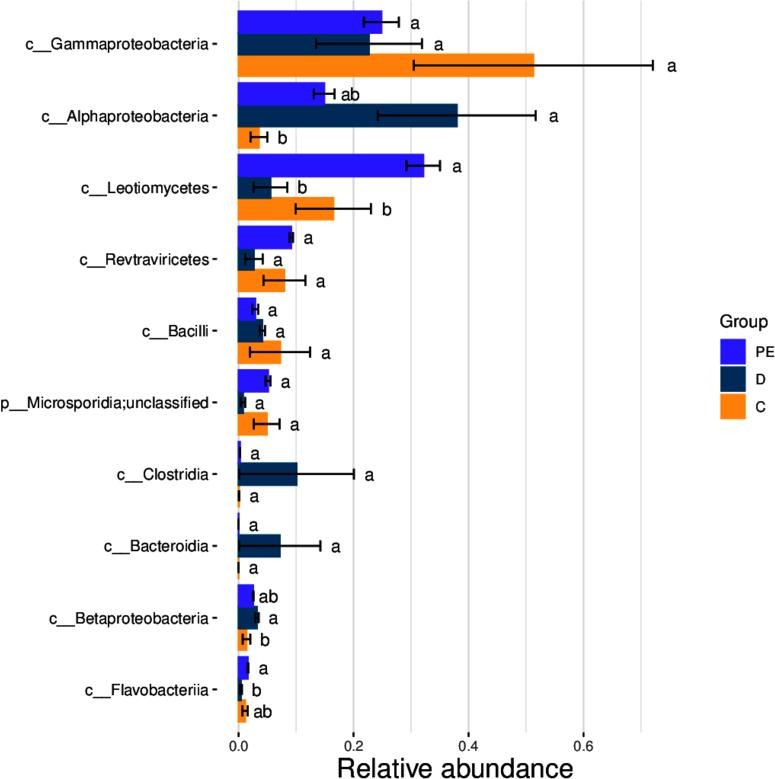


Figure 5.3 Anova based on class level analyze

### Random\_Forest based on species abundance analyze

Random forest is a highly flexible machine learning method. It is an algorithm that integrates multiple trees through the idea of ensemble learning . Its basic unit is the decision tree, and its essence belongs to a major branch of machine learning - ensemble learning method . Each decision tree is a classifier ( assuming that the problem is classification). For an input sample, N A tree will have N Random forest integrates all classification voting results and designates the category with the most votes as the final output. This is the simplest Baggin​​​​ Thought.

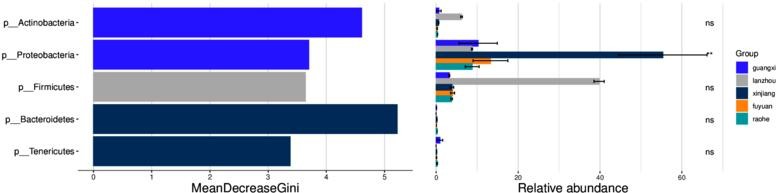


Figure 5.4 Random\_Forest based on gate level analyze

### Wilcoxon based on species abundance analyze

Will coxon​​​​​ Inspection (also known as M a nn -W it hn ey - W il c o x o n The χ2 test ) is a nonparametric test, meaning that it does not rely on the data belonging to a probability distribution family with any specific parameters. The goals of nonparametric tests are the same as those of parametric tests. However, they have one advantage over parametric tests: they do not require the assumption of normality of the distribution.

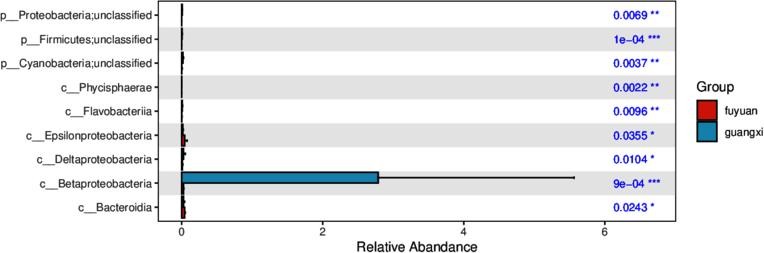


Figure 5.5 Bar chart of significantly different species

### MetagenomeSeq based on species abundance analyze

metagenomeSeq It is R A package developed by , the basic idea is to first standardize the data, then use zero - inflated Gaussian distribution to deal with the impact of sequencing depth, and finally find the difference based on the linear model . It is mainly used to compare the difference between two groups of samples, and whether the number of samples in each group is consistent has no effect on the result.

the ASV/OTU between the corresponding two groups Multiples​ change; log2 of FC)

Value, positive value represents B Group A Group upregulation (A The group refers to the group name in front of the folder name ) , and a negative value means a downward adjustment; the P value of the value is also given Pvalues and FDR - corrected P value (padj)

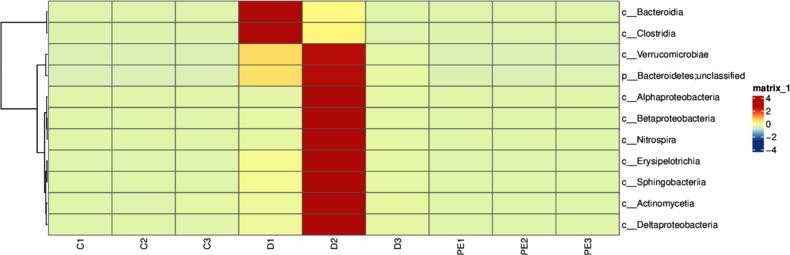


Figure 5.6 Heat map of significantly different species

### Stamp based on species abundance analyze

Stamp The software is used for inter-group difference analysis, which is suitable for various types of inter-group difference analysis, such as species distribution and abundance differences, genetic differences, functional differences, etc.

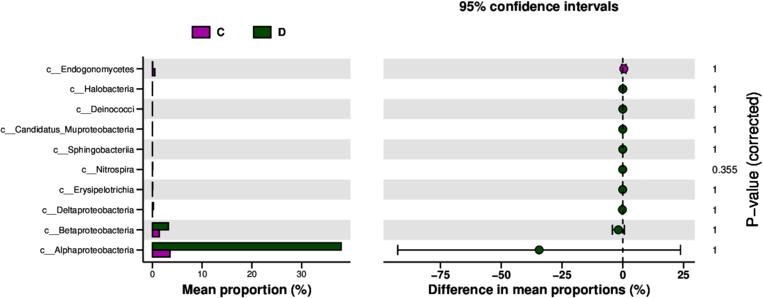


Figure 5.7 Difference test histogram

### Metastat of species with differences between groups analyze

In order to study the species with significant differences between groups, we used the Metastats[31 ] method to conduct hypothesis testing on the species abundance data between groups and obtained the p value. value , by The correction of the value is obtained value; finally according to q The values were used to screen species with significant differences and a bar chart of the difference species was drawn.

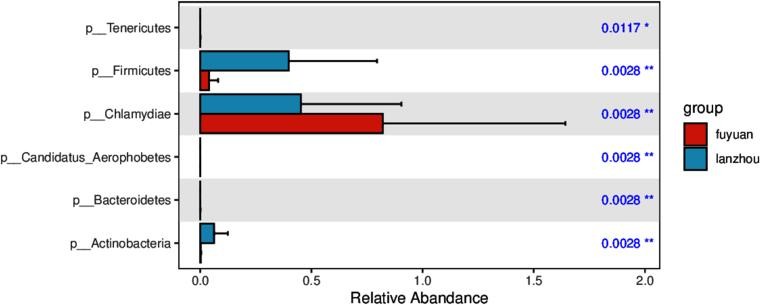


Figure 5.8 Bar chart of significantly different species

### LEfSe of species with differences between groups analyze

In order to screen the biomarkers with significant differences between groups , we first detected the different species between different groups by using the rank sum test method and used LDA (linear discriminant analysis) to reduce the dimension and evaluate the influence of the different species, that is, we obtained LDA score [32] ; Differences in species between groups LqCy The analysis results include three parts: LDA Value distribution histogram,

Evolutionary branch diagram (phylogenetic distribution) and biomarkers with statistical differences between groups The abundance comparison diagram in different groups. The LDA value distribution diagram of the differential species is as follows:

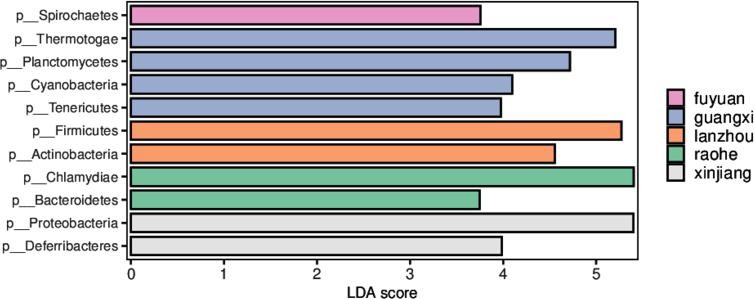


Figure 5.9 LDA of different species Value distribution graph

## Functional level analysis

The commonly used functional databases are:

Kyoto Encyclopedia of Genes and Genomes (KEGG) [16] [17] ;

Evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) [18] ;

Carbohydrate-Active enzymes Database (CAZy) [19] ;

KEGG Database in 1995 By Kanehisa Laboratories Release 0.1 It has now developed into a comprehensive database, the core of which is KEGG PATHWAY AND KEGG ORTHOLOGY Database . ORTHOLOGY In the database, genes with the same function are grouped together, called orthologs. Groups (KO entries) , each KO Contains information about multiple genes and is located in one or more pathways In KEGG PATHWAY In the database, biological metabolic pathways are divided into 7 The categories are: Cellular Processes , Environmental Information Processing , Genetic Information Processing , Human Diseases , Metabolism , Organismal Systems Systems ), Drug Development Development ) , each of which is systematically classified into the second, third, and fourth layers. The second layer currently includes 43 Seed pathway ; The third layer is its metabolic pathway diagram; the fourth layer is the specific annotation information for each metabolic pathway diagram.

eggNOG The database is based on the Smith-Waterman Orthologous gene clusters constructed by alignment algorithm Groups Functional annotation, eggNOG V4.1 Covering 2031 The genes of 19 species have been constructed . 10,000 Orthologous Groups .

CAY The database is a professional database for studying carbohydrate enzymes , mainly covering 6 Large functional class: Glycoside Hydrolases , GHs ), Glycosyl Transferases ( GTs ) , Polysaccharide Lyase ( Lyases , PLs , Carbohydrate esterase​ Esterases , CEs , Auxiliary Activities​​ , AAs) and Carbohydrate -Binding Modules ( CBMs ).

### Basic steps of functional annotation

[1] Using DIAMOND The software compares non-redundant genes with each functional database, taking e<1e-5 Based on the annotations, proteins with the highest sequence similarity are screened to obtain functional annotation information;

[2]. Alignment result filtering: For each sequence alignment result, select the score The highest alignment result ( one HSP > 60 bits ) for subsequent analysis [20 ] ;

[3]. Based on the comparison results, the relative abundance of different functional levels was calculated (the relative abundance of each functional level is equal to the sum of the relative abundance of genes annotated to that functional level [21 ] [ 22 ] [ 23 ] ), among which KEGG The database is divided into 5 Levels, eggNOG The database is divided into 3 Level, CAZy The database is divided into 3 The detailed division levels of each database are as follows:

|  |  |  |
| --- | --- | --- |
| **Database name** | **Divide into levels** | **Description of this level** |
| KEGG | level1 | KEGG Metabolic Pathway Level 17 macro metabolic pathways; |
| KEGG | level2 | KEGG Metabolic Pathway Level 243 Seed pathway ; |
| KEGG | level3 | KEGG pathway id (e.g. ko00010 ) ; |
| KEGG | ko | KEGG ortholog group ( e.g. K00010) ; |
| KEGG | ec | KEGG EC Number (Example: EC 3.4.1.1 ); |
| eggNOG | level1 | twenty four Large functional category; |
| eggNOG | level2 | ortholog group description ; |
| eggNOG | og | ortholog group ID ( e.g. ENOG410YU5S ) |
| CAY | level1 | 6 Large functional category; |

|  |  |  |
| --- | --- | --- |
| CAY | level2 | CAY family ( e.g. GT51 ); |
| CAY | level3 | EC number (e.g. murein polymerase ( EC 2.4.1.129) ); |

[4]. Based on the functional annotation results and gene abundance table, obtain the gene number table of each sample at each classification level . The number of genes with a certain function in a certain sample is equal to the number of genes with non- zero abundance among the genes annotated with this function .

[5]. Starting from the abundance table at each classification level, the number of annotated genes is counted, the relative abundance profile is displayed, the abundance clustering heat map is displayed, and PCA is performed. 、 PCoA and NMDS Dimensionality reduction analysis, Anosim based on functional abundance Inter-group (intra-group) difference analysis, comparative analysis of metabolic pathways, Metastat of inter-group functional differences and LEfSe analyze.

### Annotation results of each database

The genes annotated by each database are finally obtained in the following link

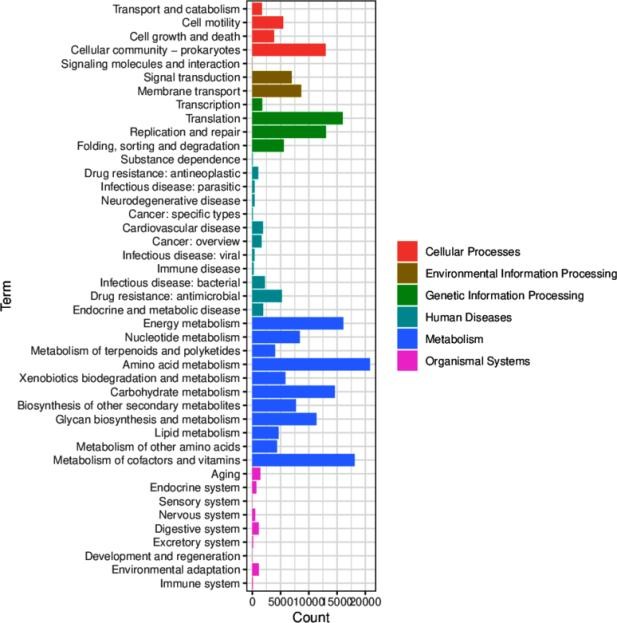


Figure 6.2 Statistics of the number of annotated genes in each database

Image caption: From left to right: KEGG , eggNOG , CAZy , GO The numbers on the bar graph represent the number of non -redundant genes in the annotation; the other axis is the level 1 The markings of each functional category and the annotations of the markings are shown in the corresponding legends.

GO Note: The horizontal axis represents GO Functional classification, the left vertical axis represents the gene annotated to this class The number of percentages, the vertical axis on the right represents the gene annotated to this class quantity.

### Functional relative abundance profile

Based on the classification information of genes in each database and the abundance table of genes in each sample, the relative abundance information of each database at different levels can be obtained . It is divided into three levels (the first level is the biological metabolic pathway, the second level is the sub-function, and the third level is the metabolic pathway diagram), eggNOG It is divided into three levels (the first level is the secondary functional classification, the second level is the functional description, and the third level is eggNOG Number), CAZy It is divided into two levels (the first level is the six major functional categories, and the second level is the sub-functions).

the relative abundance table of the first level ( level A ) of each database , a statistical graph of the first level corresponding to each sample in each database was drawn. Software, draw a bar chart of the functional composition of each sample, as shown below:

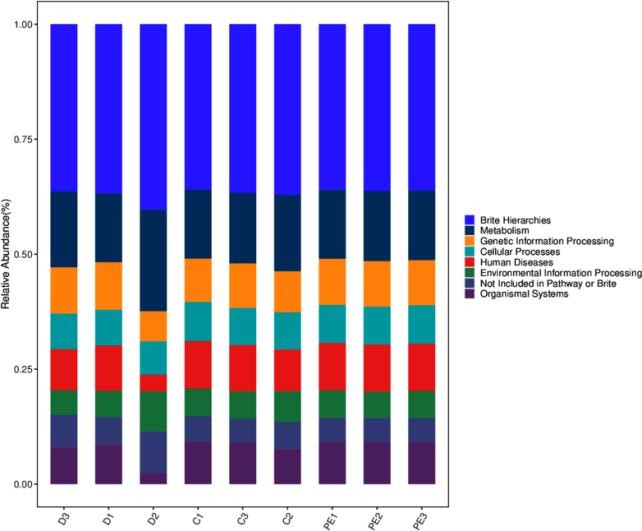


Figure 6.3.1 Functional abundance bar graph

Image caption: From left to right: KEGG eggNOG and CAZy The functional categories corresponding to each color block are shown in the legend below.

Heat maps were drawn based on the functional annotations and abundance information of all samples in various databases ( KEGG The third level is selected for display, eggNOG selects the third level for display, and CAZy selects the second level for display) , and clusters are performed at the function and sample levels.

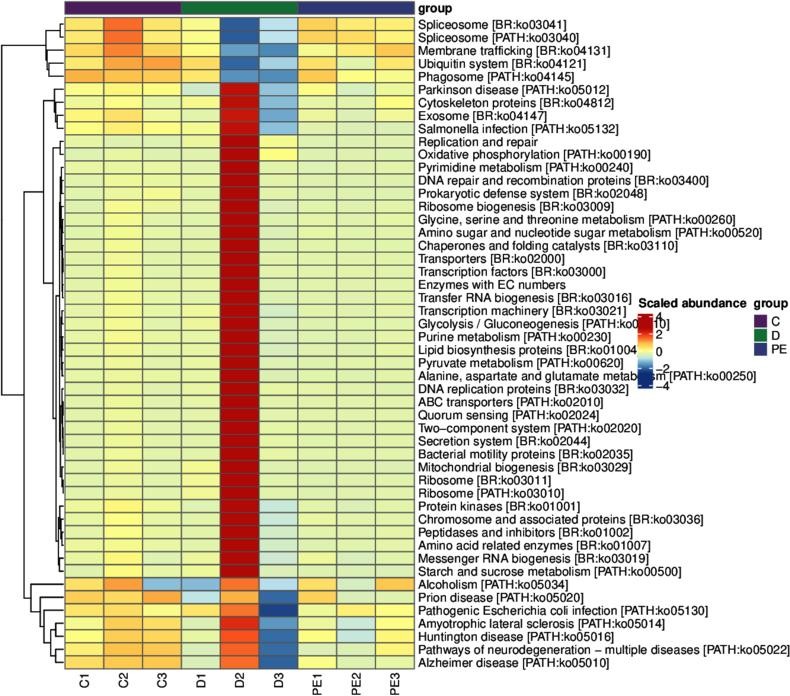


Figure 6.3.2 Function abundance heatmap

Image caption: From left to right: KEGG eggNOG and CAZy The functional categories corresponding to each color block are shown in the legend below.

According to all samples in each database A Based on the relative abundance information at the level of classification, we can calculate the Bray distance between samples and then perform hierarchical clustering based on the Bray distance. At the same time, for ease of understanding, the functional classification stacking diagram is combined with the clustering results of the Bray distance.

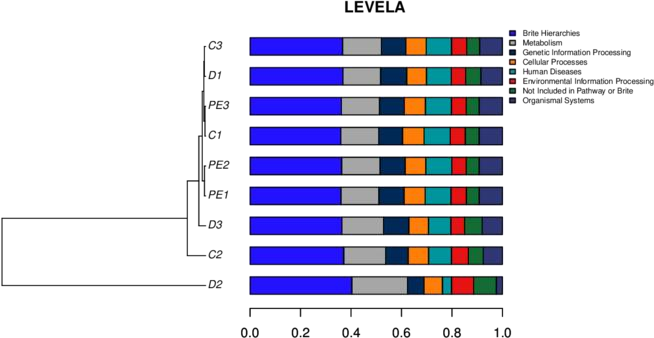
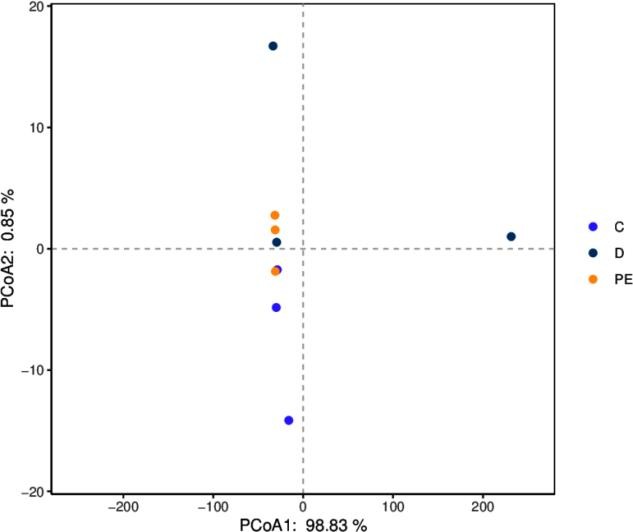


Figure 6.3.3 Functional clustering diagram

Image caption: From left to right: KEGG eggNOG and CAZy The functional categories corresponding to each color block are shown in the legend below.

### Dimensionality reduction analysis based on functional abundance

Principal coordinates analysis ( PCoA ) is a classic multidimensional Scaling , cMDScale) analysis method (Ramette, 2007) . It expands the sample distance matrix in a low-dimensional space after projection, and retains the distance relationship of the original sample to the maximum extent. PCoA considers the sample distance as a whole. Compared with principal component analysis ( PCA ) , it is more in line with the characteristics of ecological data. Therefore, it is more recommended as a sorting analysis method.

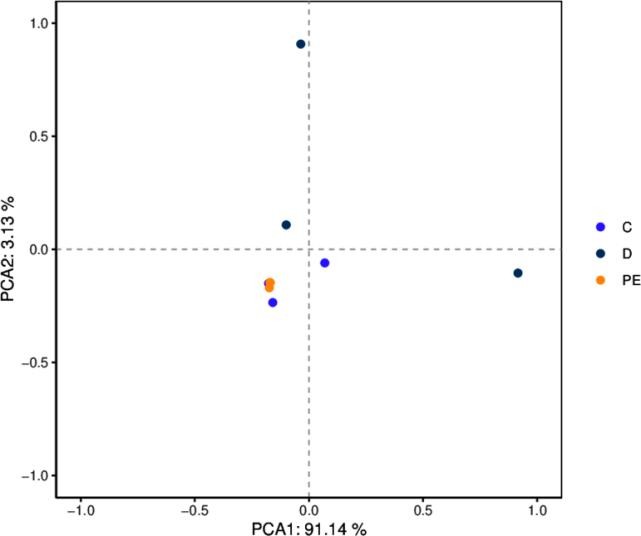


* + 1. PCoA Analysis chart

Note: The dots represent the functional composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample difference; the vertical axis represents the second principal component and its contribution to the sample difference.

Image caption: From left to right: KEGG eggNOG and CAZy The functional categories corresponding to each color block are shown in the legend below.

Principal component analysis​​​​​​ Component​​​​​​ An a l y s is ,​ PCA ) is a technique for analyzing and simplifying data sets. By decomposing the variance, the differences between multiple groups of data are reflected on a two-dimensional coordinate graph. The coordinate axes take the two eigenvalues that can best reflect the variance . By analyzing the composition of different samples ( 97 % similarity ), the differences and distances between samples can be reflected. The closer the distance between two samples on the graph, the more similar the functional composition of the two samples is. R Language tools plot different classification levels separately P C A Analysis diagram, the sample diagram is as follows:

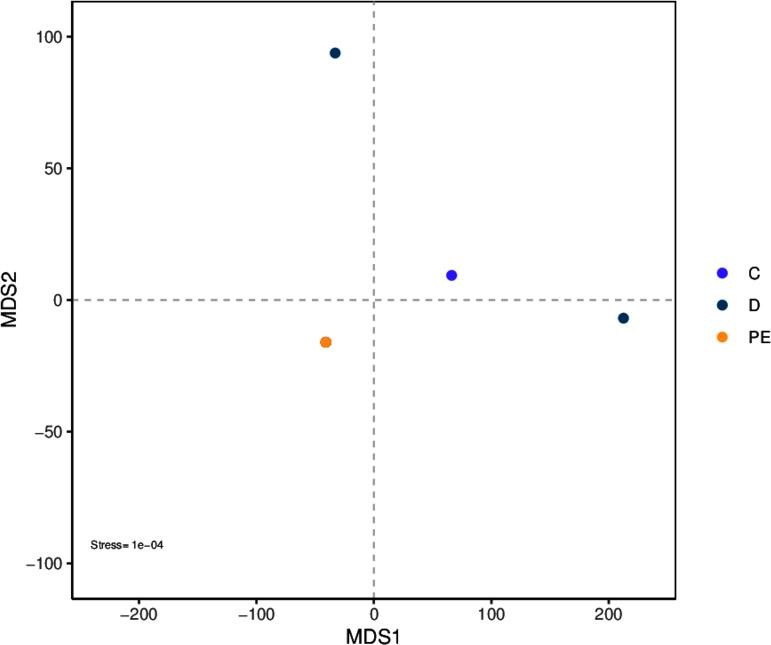


* + 1. PCA Analysis chart

Note: The dots represent the functional composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample difference; the vertical axis represents the second principal component and its contribution to the sample difference.

Image caption: From left to right: KEGG eggNOG and CAZy The functional categories corresponding to each color block are shown in the legend below.

Non-metric multidimensional scaling (N M D S ) and the above P C o A The analysis is similar to that of the previous one, which also reduces the dimension of the sample distance matrix and simplifies the data structure, thereby describing the distribution characteristics of the samples at a specific distance scale. P C o A Different analysis, N M D S The analysis does not rely on the calculation of eigenvalues and eigenvectors, but rather on ranking the sample distances so that the order of samples in the low-dimensional space is as close as possible to the distance between each other ( rather than the exact distance value ) . The analysis is not affected by the numerical value of the sample distance, and only considers the size relationship between each other. For data with complex structures, the sorting results may be more stable. N M D S The smaller the stress value ( S t ress ) of the result , the better. It is generally believed that when the value is less than 0.2​ When N M D S The results of the analysis are relatively reliable ( L e g e nd re, 199 8 ) .



* + 1. NMDS Analysis chart

Note: The dots represent the functional composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample difference; the vertical axis represents the second principal component and its contribution to the sample difference.

Image caption: From left to right: KEGG eggNOG and CAZy The functional categories corresponding to each color block are shown in the legend below.

## Analysis of intergroup differences based on functional abundance

### Based on functional abundance Anosim analyze

A no s i m Analysis is a nonparametric test used to test whether the difference between groups is significantly greater than the difference within the group, so as to determine whether the grouping is meaningful. The detailed calculation process can be viewed A no s i m . Based on functional abundance A no s i m The analysis results are as follows:

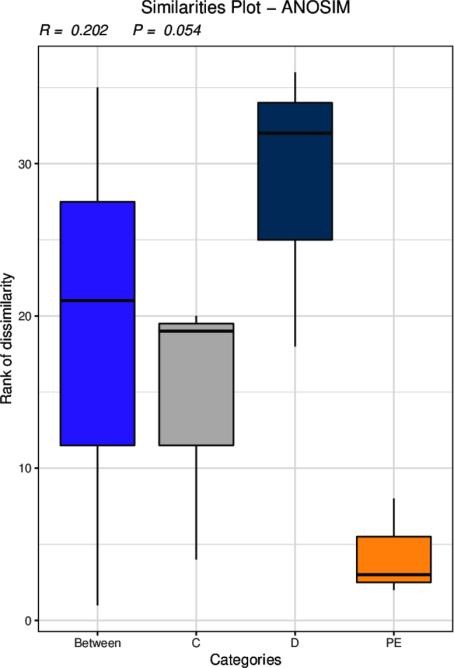


Figure 7.1 Anosim analyze

Note: R-value is between ( -1 , 1 ). If R-value is greater than 0 , it means that the difference between groups is significant. If R-value is less than 0 , it means that the difference within the group is greater than the difference between the groups. The credibility of statistical analysis is expressed by P-value. Indicates, P < 0.05 Indicates that the data are statistically significant.

* + 1. **Kruskal-Wallis based on functional abundance Rank Sum Test Analysis**

Kruskal-Wallis The rank sum test is a nonparametric test for three or more groups of data that can calculate the existence of differences between the means of multiple independent groups of data. It only deals with the volatility of the data displayed on the graph, which is different from ANOVA. In contrast, so-called nonparametric tests make no assumptions about the distribution of the data .

We can test whether there are differences in function from multiple sets of data from multiple biological replicates. p\_value

Less than 0.05 That is considered to be different.

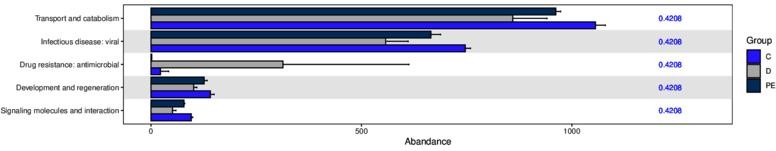


Figure 7.2 Kruskal-Wallis based on functional abundance analyze

Note: The horizontal axis of the bar graph represents the relative abundance of a function in different groups, the vertical axis represents the differential function, and different colors represent different groups.

### Based on functional abundance anova analyze

ANOVA (Analysis of variance) is abbreviated as analysis of variance or " coefficient of variation analysis " . It is used to test the significance of the difference between two or more sample means. It is mainly used to test whether the difference between two or more sample means is statistically significant or to test whether there is interaction between two or more factors.

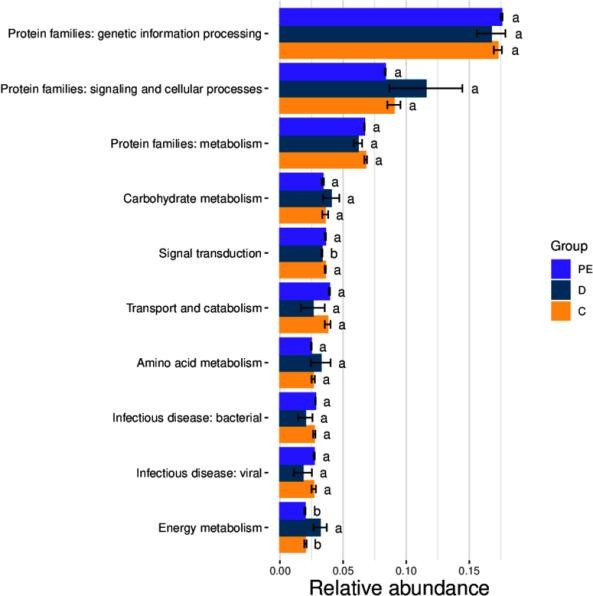


Figure 7.3 Anova based on functional abundance analyze

### Based on functional abundance Random\_Forest analyze

Random forest is a highly flexible machine learning method. It is an algorithm that integrates multiple trees through the idea of ensemble learning . Its basic unit is the decision tree, and its essence belongs to a major branch of machine learning - ensemble learning method . Each decision tree is a classifier ( assuming that the problem is classification). For an input sample, N A tree will have N Random forest integrates all classification voting results and designates the category with the most votes as the final output. This is the simplest Baggin​​​​ Thought.

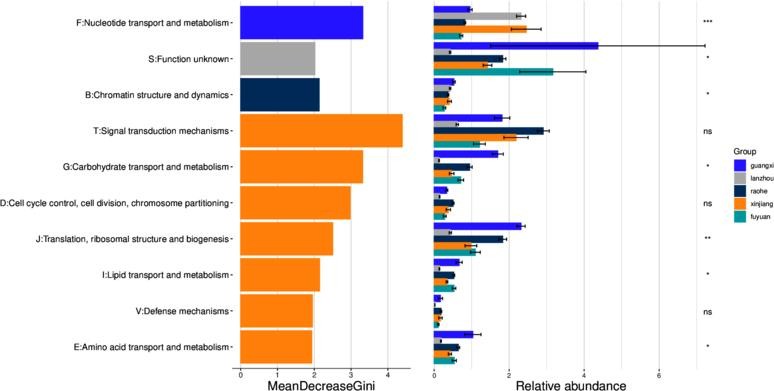


Figure 7.4 Random\_Forest based on functional abundance analyze

### Based on functional abundance Wilcoxon analyze

Will coxon​​​​​ Inspection (also known as M a nn -W it hn ey - W il c o x o n The χ2 test ) is a nonparametric test, meaning that it does not rely on the data belonging to a probability distribution family with any specific parameters. The goals of nonparametric tests are the same as those of parametric tests. However, they have one advantage over parametric tests: they do not require the assumption of normality of the distribution.

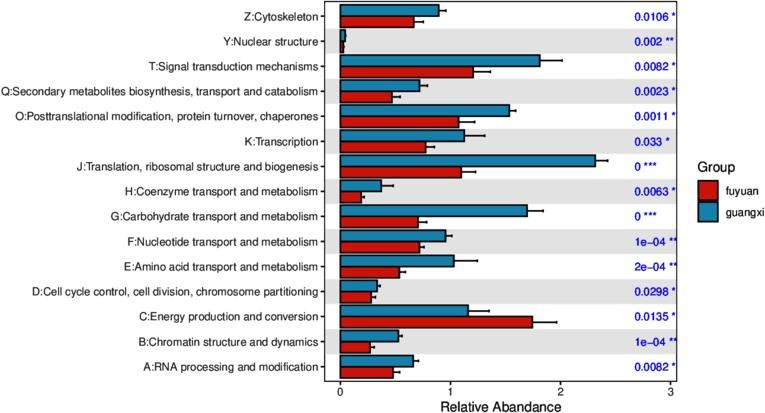


Figure 7.5 Significantly different functional bar chart

### MetagenomeSeq based on functional abundance analyze

metagenomeSeq It is R The basic idea is to first standardize the data, then use the zero -inflated Gaussian distribution to deal with the impact of sequencing depth, and finally find the optimal sequence based on the linear model. It is mainly used to compare the differences between two groups of samples , and whether the number of samples in each group is consistent has no effect on the results.

Each result file will give the ASV/OTU between the corresponding two groups Multiples​ change; FC) log2 value , positive value represents B Group A Group upregulation (A The group refers to the group name in front of the folder name ) , and a negative value means a downward adjustment; the P value of the value is also given Pvalues and FDR - corrected P Value (padj) .

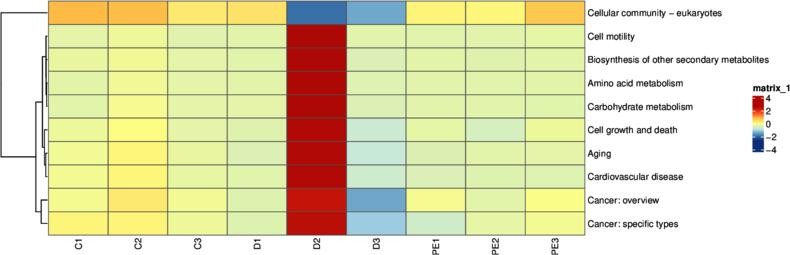


Figure 7.6 Significantly different function heatmap

### Stamp based on functional abundance analyze

Stamp The software is used for inter-group difference analysis, which is suitable for various types of inter-group difference analysis, such as species distribution and abundance differences, genetic differences, functional differences, etc.

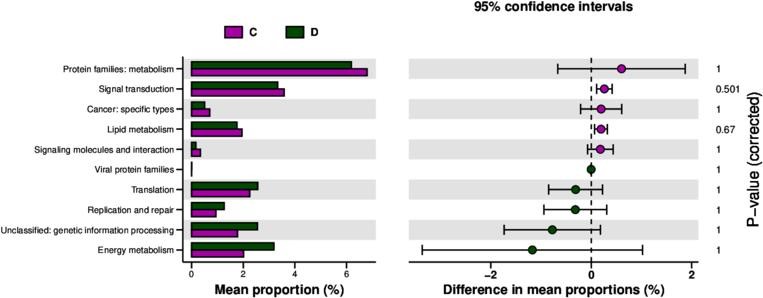


Figure 7.7 Difference test histogram

### Metastat for intergroup differences analyze

In order to investigate the functions with significant differences among the groups , Starting from the functional abundance tables at different levels , use

Metastats[31 ] method was used to perform hypothesis testing on the functional abundance data between groups to obtain p value , by Correction of the value,

to q value; finally according to q The functions with significant differences were screened and a bar chart of the differences was drawn.

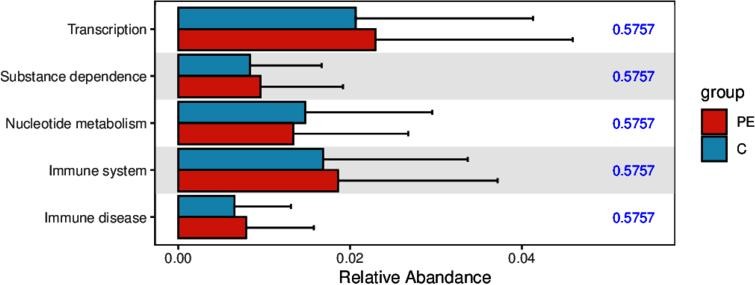


Figure 7.8 Significantly different functional bar chart

## Analysis of biogeochemical cycle functions

biogeochemical cycle​​​​​​​​​​​​ Cycle ) refers to the biosphere on the surface of the earth, where biological organisms , through life activities , Absorb elements and their compounds ( often called minerals ) from the medium of the living environment , and through biochemical reactions It is converted into life substances and excreted The decomposed matter returns to the environment and is decomposed into elements or chemicals after its death. minerals ) back to the surrounding medium. This cyclical process is called a biogeochemical cycle. Biogeochemical cycles also involve the transfer of nutrients from one organism (primary producer ) to another . Transfer of organisms (consumers ) or food chain transmission and effects. Biogeochemical cycle is one of the core research directions of Earth system science, and it studies carbon ( C ) , nitrogen ( N ), phosphorus ( P ) , Sulfur ( S ) and Describe, trace and predict the circulation process of heavy metals and other elements in the earth's sphere Biogeochemical The important content of academic cycle research.

Based on KEGG and published literature information , and constructed a carbon fixation ( 137 KOs ) , nitrogen cycle ( 32 KOs ) , phosphorus cycle ( 53 KOs ) and sulfur cycle ( 88 By extracting the above information from the database functional annotation results , based on TPM Abundance calculation method, functional composition analysis, difference analysis and correlation analysis.

### Schematic diagram of the circulation pathway

The schematic diagram of the carbon, nitrogen and sulfur cycle pathways is as follows:

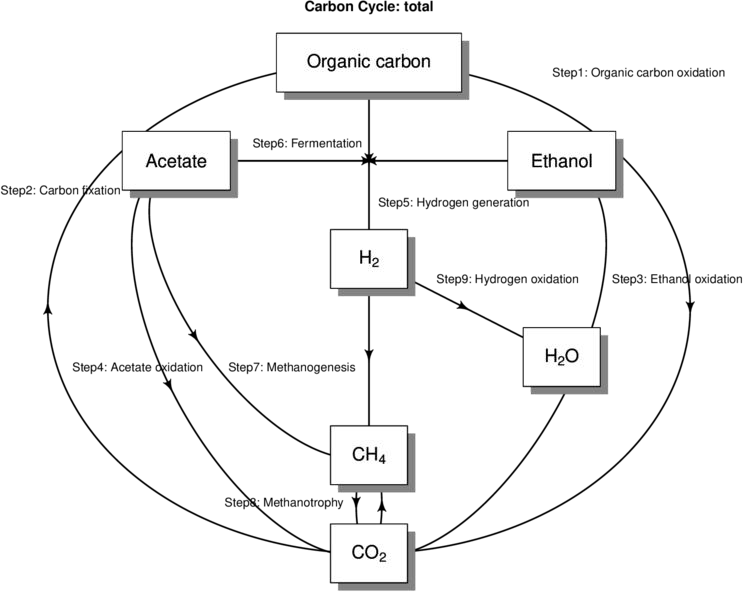


Figure 8.1.1 Schematic diagram of the circulation pathway

### Annotation results of each database

The final annotated genes for each cycle are shown in the following link :

### Overview of relative abundance of functional genes

The abundance ratio of each functional gene in the sample is counted, and the functional gene composition of each sample is intuitively displayed through the stacked column visualization method.

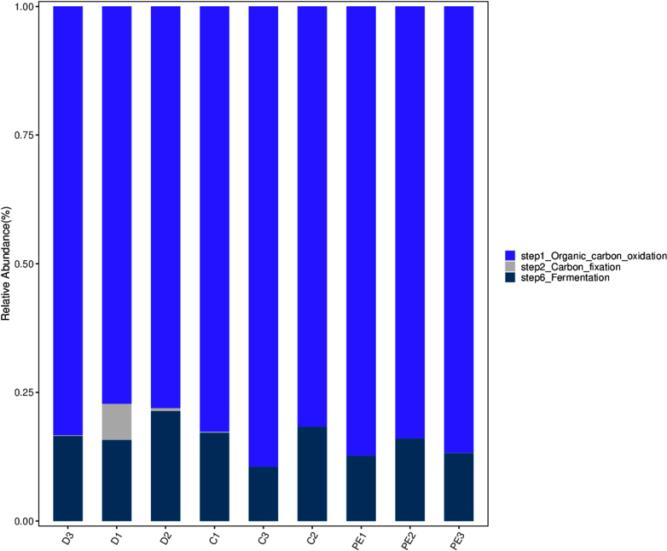


Figure 8.3.1 Functional gene abundance bar graph

the TPM of the functional gene in the sample Abundance value, columns of different colors represent different functional genes, and the length of the column represents the TPM of the functional gene in the sample Abundance size.

Heatmap Color changes can be used to reflect the data information in a two-dimensional matrix or table. It can intuitively represent the size of the data value with a defined color depth. Cluster the data based on the abundance similarity between functional genes or samples, and display the clustered data in a Heatmap. In the figure, high-abundance and low-abundance functional genes can be clustered into blocks, and the similarities and differences of functional genes of multiple samples can be reflected by color gradient and similarity.

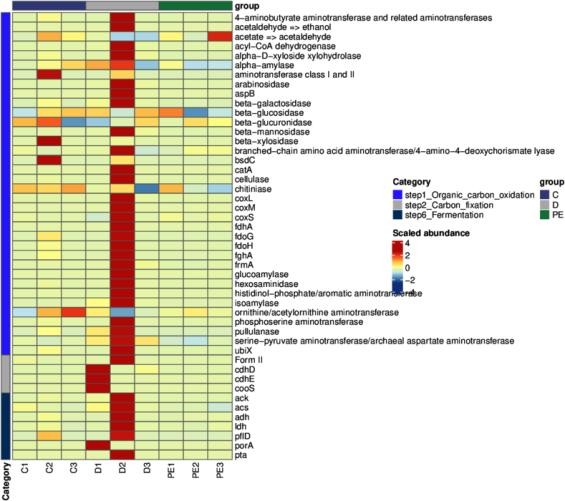


Figure 8.3.2 Functional gene abundance heat map

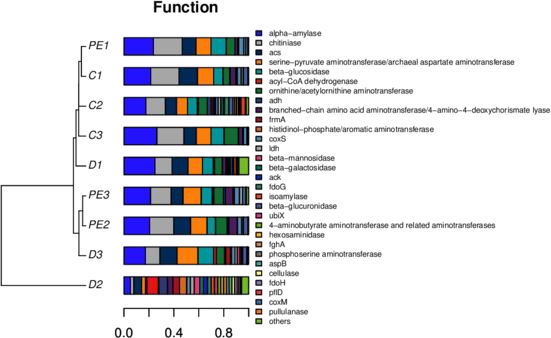
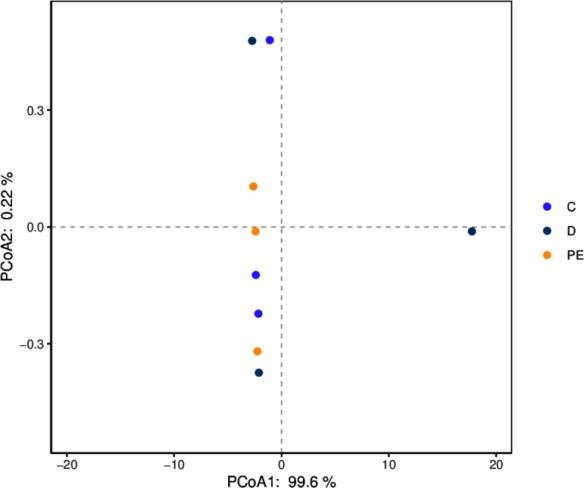
According to the relative abundance information of all sample functional genes (here we take the top 30 in terms of overall abundance) functional genes in the sample), we can calculate the Bray distance, then according to Bray Distance is used for hierarchical clustering, which can determine the similarity of samples. At the same time, for ease of understanding, the gene classification stacking diagram is compared with the Bray The clustering results of distances are displayed together.

Figure 8.3.3 Bray Cluster analysis and functional gene composition stacking diagram

### Dimensionality reduction analysis based on functional gene abundance

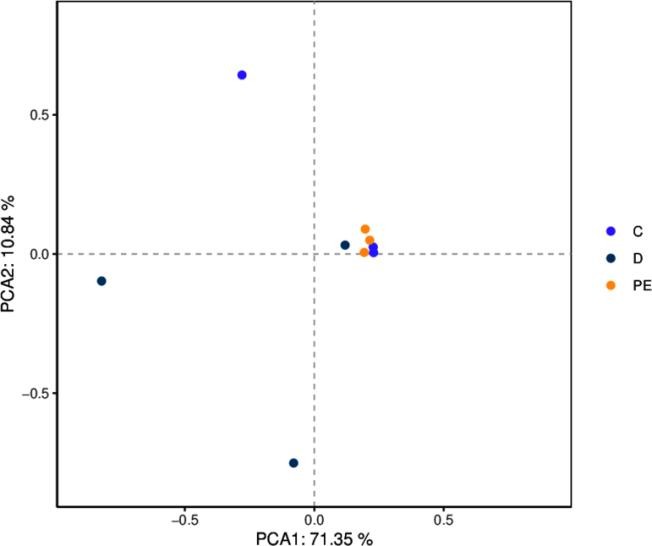
Principal coordinates analysis ( PCoA ) is a classic multidimensional Scaling , cMDScale) analysis method (Ramette, 2007) . It expands the sample distance matrix in a low-dimensional space after projection, and retains the distance relationship of the original sample to the maximum extent. PCoA considers the sample distance as a whole. Compared with principal component analysis ( PCA ) , it is more in line with the characteristics of ecological data. Therefore, it is more recommended as a sorting analysis method.



* + 1. PCoA Analysis chart

Note: The dots represent the functional gene composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample differences; the vertical axis represents the second principal component and its contribution to the sample differences.

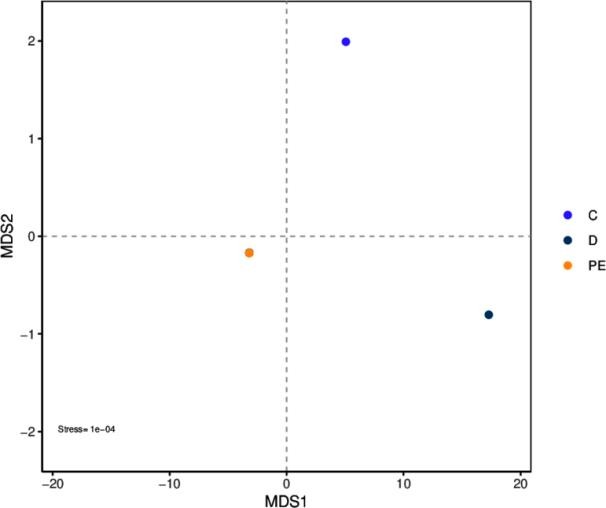
Principal component analysis​​​​​​ Component​​​​​​ An a l y s is ,​ PCA ) is a technique for analyzing and simplifying data sets. By decomposing the variance, the differences between multiple groups of data are reflected on a two-dimensional coordinate graph. The coordinate axes take the two eigenvalues that can best reflect the variance . By analyzing the composition of different samples ( 97 % similarity ), the differences and distances between samples can be reflected. The closer the distance between two samples on the graph, the more similar the composition of functional genes in the two samples is. R Language tools plot different classification levels separately P C A Analysis diagram, the sample diagram is as follows:



* + 1. PCA Analysis chart

Note: The dots represent the functional gene composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample differences; the vertical axis represents the second principal component and its contribution to the sample differences.

Non-metric multidimensional scaling (N M D S ) and the above P C o A The analysis is similar to that of the previous one, which also reduces the dimension of the sample distance matrix and simplifies the data structure, thereby describing the distribution characteristics of the samples at a specific distance scale. P C o A Different analysis, N M D S The analysis does not rely on the calculation of eigenvalues and eigenvectors, but rather on ranking the sample distances so that the order of samples in the low-dimensional space is as close as possible to the distance between each other ( rather than the exact distance value ) . The analysis is not affected by the numerical value of the sample distance, and only considers the size relationship between each other. For data with complex structures, the sorting results may be more stable. N M D S The smaller the stress value ( S t ress ) of the result , the better. It is generally believed that when the value is less than 0.2​ When N M D S The results of the analysis are relatively reliable ( L e g e nd re, 199 8 ) .



* + 1. NMDS Analysis chart

Note: The dots represent the functional gene composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample differences; the vertical axis represents the second principal component and its contribution to the sample differences.

### Analysis of species contribution to genes

Based on the gene sets screened from the metagenomic data of this project and the species information corresponding to the genes and sets, the contribution of species to specific functional genes is analyzed, and the results are visualized through stacked bar charts:

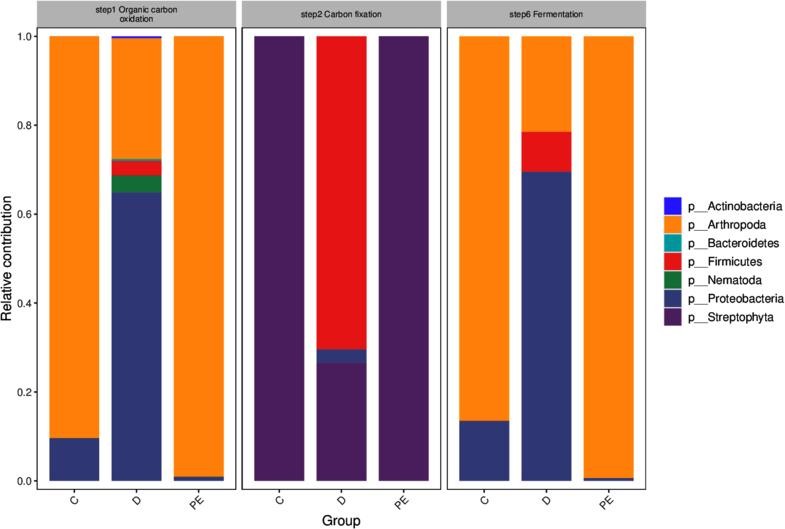


Figure 8.5.1 Species' contribution to genes

Note: The figure above shows the main species composition of specific functional genes (represented by the relative abundance of species ) , and the horizontal axes in the figure correspond to the sample groups.

### Analysis of intergroup differences based on functional gene abundance

Anosim Analysis is a non-parametric test used to test whether the difference between groups is significantly greater than the difference within the group, so as to determine whether the grouping is meaningful. The detailed calculation process can be viewed in Anosim . The results of Anosim analysis based on the abundance of functional genes are as follows:

Figure 8.6.1 Anosim based on functional gene abundance analyze

Note: R-value is between ( -1 , 1 ). If R-value is greater than 0 , it means that the difference between groups is significant. If R-value is less than 0 , it means that the difference within the group is greater than the difference between the groups. The credibility of statistical analysis is expressed by P-value. Indicates, P < 0.05 Indicates that the data are statistically significant.

Kruskal-Wallis The rank sum test is a nonparametric test for three or more groups of data that can calculate the existence of differences between the means of multiple independent groups of data. It only deals with the volatility of the data displayed on the graph, which is different from ANOVA. Different from the so-called non-parametric test, no assumptions are made about the data distribution. Through the Kruskal-Wallis rank sum test, we can determine whether there are differences in functional genes from multiple sets of data from multiple biological replicates. Here we consider that a p\_value less than 0.05 is considered to be different.

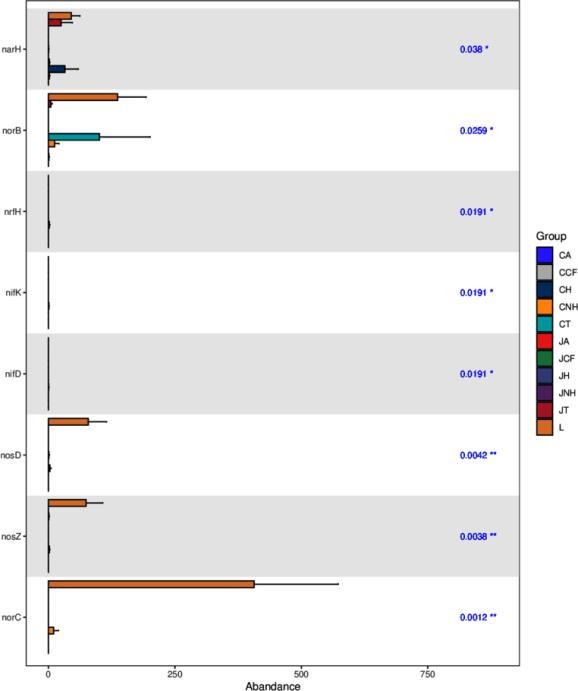


Figure 8.6.2 Kruskal-Wallis based on functional gene abundance analyze

Note: The horizontal axis of the bar graph represents the relative abundance of a functional gene in different groups, the vertical axis represents the differentially functional genes, and different colors represent different groups.

ANOVA ( A n a l y s i s​ o f ANOVA or " coefficient of variation analysis " is used to analyze the variance of two or more Significance test of the difference between the sample means. It is mainly used to test whether the difference between two or more sample means is statistically significant or to test whether there is interaction between two or more factors.

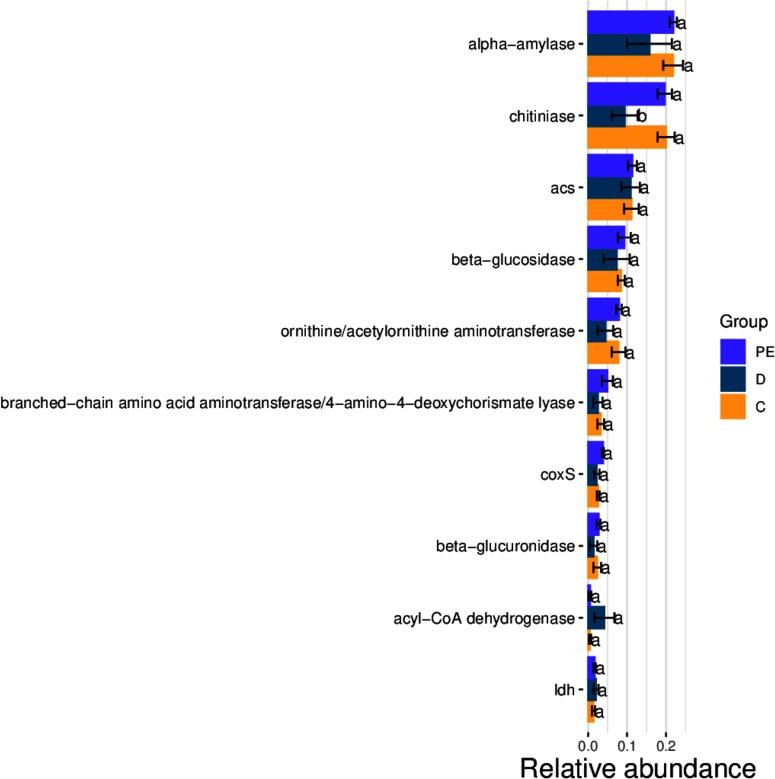


Figure 8.6.3 Anova based on functional gene abundance analyze

Random forest is a highly flexible machine learning method. It is an algorithm that integrates multiple trees through the idea of ensemble learning . Its basic unit is the decision tree, and its essence belongs to a major branch of machine learning - ensemble learning method . Each decision tree is a classifier ( assuming that the problem is classification). For an input sample, N A tree will have N Random forest integrates all classification voting results and designates the category with the most votes as the final output. This is the simplest Baggin​​​​ Thought.

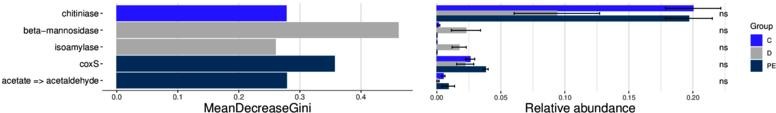


Figure 8.6.4 Random\_Forest based on functional gene abundance analyze

Will coxon​​​​​ Inspection (also known as M a nn -W it hn ey - W il c o x o n The χ2 test ) is a nonparametric test, meaning that it does not rely on the data belonging to a probability distribution family with any specific parameters. The goals of nonparametric tests are the same as those of parametric tests. However, they have one advantage over parametric tests: they do not require the assumption of normality of the distribution.

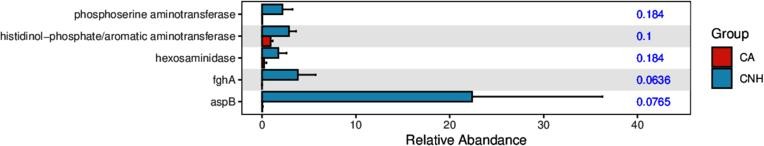


Figure 8.6.5 Bar graph of significantly differentially expressed functional genes

metagenomeSeq It is R A package developed by , the basic idea is to first standardize the data, then use zero - inflated Gaussian distribution to deal with the impact of sequencing depth, and finally find the difference based on the linear model . It is mainly used to compare the difference between two groups of samples, and whether the number of samples in each group is consistent has no effect on the result.

the ASV/OTU between the corresponding two groups Multiples​ change; log2 of FC)

Value, positive value represents B Group A Group upregulation (A The group refers to the group name in front of the folder name ) , and a negative value means a downward adjustment; the P value of the value is also given Pvalues and FDR - corrected P value (padj)

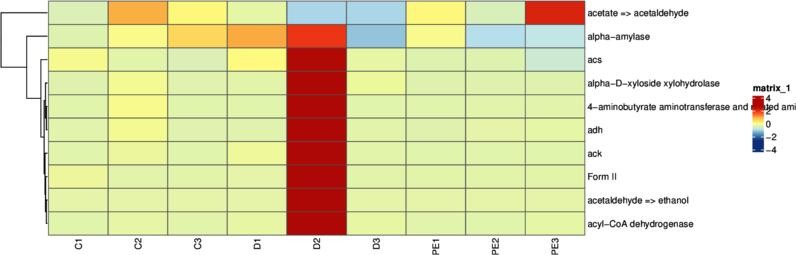


Figure 8.6.6 Heat map of significantly differentially expressed genes

Stamp The software is used for inter-group difference analysis, which is suitable for various types of inter-group difference analysis, such as species distribution and abundance differences, gene differences, functional gene differences, etc.

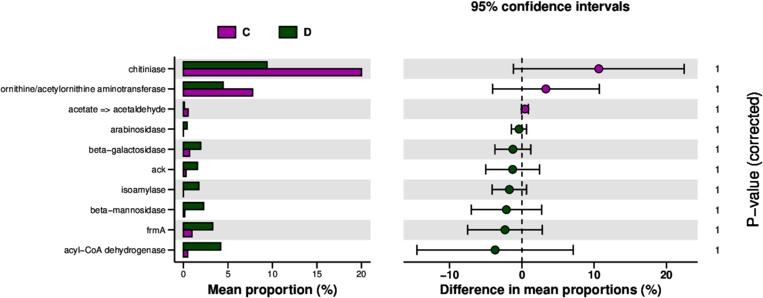


Figure 8.6.7 Difference test histogram

In order to study the functional genes with significant differences between groups, we used Metastats[31 ] to perform hypothesis testing on the functional gene abundance data between groups based on the functional gene abundance tables at different levels to obtain the p value , by The correction of the value is obtained value; finally according to q The functional genes with significant differences were screened and a bar graph of the differentially functional genes was drawn .

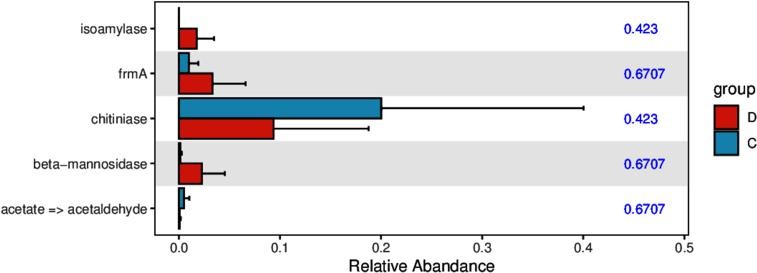


Figure 8.6.8 Bar graph of significantly differentially expressed functional genes

## Resistance gene annotation

Resistance genes are ubiquitous in both human intestinal microorganisms and other environmental microorganisms. The abuse of antibiotics leads to irreversible changes in the microbial communities in the human body and the environment, posing risks to human health and the ecological environment. Therefore, research on resistance genes has attracted widespread attention from researchers [ 40 ] . 2.0​ Version includes C A R D and A R D B database and incorporated N CB I- N R The latest sequence in the database is a new resistance gene database that has emerged in recent years. It has the advantages of comprehensive information, user-friendly, and timely update and maintenance. The core components of this database are A n ti b i o ti c R es i s t a n ce Ontology ( ARO ) , which integrates sequence , antibiotic resistance , mechanism of action , and provide online A R O and P D B 、 N C B I etc. database interfaces [ 41 ] .

### Overview of relative abundance of resistance genes

* + - 1. Using SARG2 Unigenes​ Compare with the database ( default evalue ≤ 1e-30[42] );
      2. according to SARG2 The comparison results, combined with Unigenes The abundance information of ARG The relative abundance of

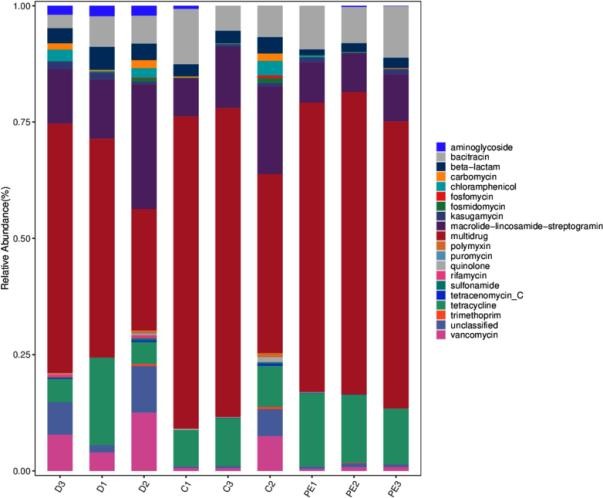


Figure 9.1.1 Resistance gene abundance bar graph

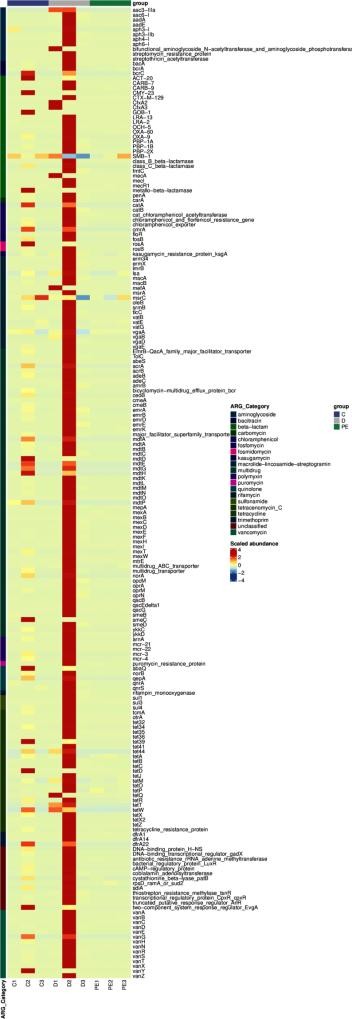
A heat map was drawn based on the resistance gene annotation and abundance information of all samples in the database , and clustering was performed at both the resistance gene and sample levels.

Figure 9.1.2 Heat map of resistance gene abundance

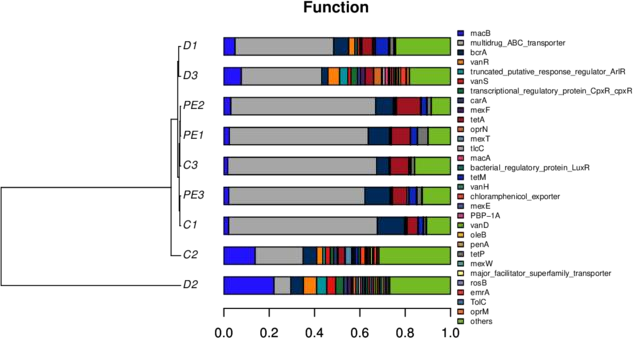
Based on the relative abundance information of all samples, we can calculate the Bray​ distance, and then according to Bray​ At the same time, for ease of understanding, the resistance gene classification stacking diagram is compared with Bray​ The clustering results of distances are displayed together.

Figure 9.1.3 Resistance gene clustering diagram

### Analysis of species contribution to resistance genes

Based on the sex gene sets screened from the metagenomic data of this project and the species information corresponding to the genes and sets, the contribution of species to specific functions is analyzed, and the results are visualized through stacked bar charts;

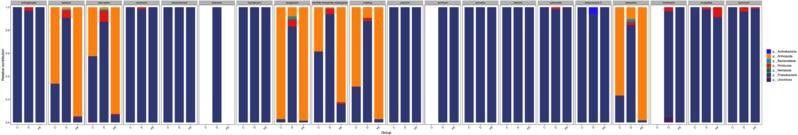


Figure 9.2.1 Species contribution to resistance genes

Note: The figure above shows the main species composition of specific functions (represented by the relative abundance of species ) , and the horizontal axes in the figure correspond to the sample groups.

### Analysis of intergroup differences based on resistance gene abundance

Anosim Analysis is a non-parametric test used to test whether the difference between groups is significantly greater than the difference within the group, so as to determine whether the grouping is meaningful. The detailed calculation process can be viewed in Anosim . The results of Anosim analysis based on the abundance of resistance genes are as follows:



Figure 9.3.1 Anosim based on resistance gene abundance analyze

Note: R-value is between ( -1 , 1 ). If R-value is greater than 0 , it means that the difference between groups is significant. If R-value is less than 0 , it means that the difference within the group is greater than the difference between the groups. The credibility of statistical analysis is expressed by P-value. Indicates, P < 0.05 Indicates that the data are statistically significant.

The Kruskal-Wallis rank sum test is a nonparametric test for three or more groups of data that can calculate the existence of differences between the means of multiple independent groups of data. It only deals with the volatility of the data displayed on the graph, which is different from ANOVA. In contrast, so-called nonparametric tests make no assumptions about the distribution of the data . Rank sum test We can use multiple sets of data from multiple biological replicates to determine whether the resistance genes are different. Here we consider that p\_value less than 0.05 is considered to be different.



Figure 9.3.2 Kruskal-Wallis analysis based on resistance gene abundance analyze

Note: The horizontal axis of the bar graph represents the relative abundance of a certain resistance gene in different groups, the vertical axis represents the differential resistance gene, and different colors represent different groups.

ANOVA (Analysis of variance) is abbreviated as analysis of variance or " coefficient of variation analysis " . It is used to test the significance of the difference between two or more sample means. It is mainly used to test whether the difference between two or more sample means is statistically significant or to test whether there is interaction between two or more factors.

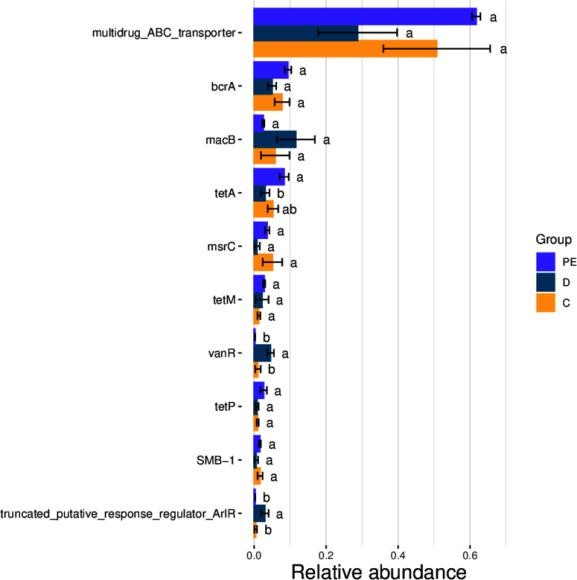


Figure 9.3.3 Anova based on resistance gene abundance analyze

Random forest is a highly flexible machine learning method. It is an algorithm that integrates multiple trees through the idea of ensemble learning . Its basic unit is the decision tree, and its essence belongs to a major branch of machine learning - ensemble learning method . Each decision tree is a classifier ( assuming that the problem is classification). For an input sample, N A tree will have N Random forest integrates all classification voting results and designates the category with the most votes as the final output. This is the simplest Baggin​​​​ Thought.

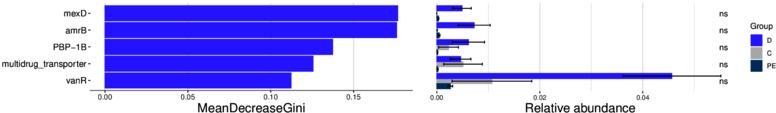


Figure 9.3.4 Random\_Forest based on resistance gene abundance analyze

Will Coxon​​​​​ Inspection (also known as M a nn - W ithn e y - W i l c o x o n The χ2 test ) is a nonparametric test, meaning that it does not rely on the data belonging to a probability distribution family with any specific parameters. The goals of nonparametric tests are the same as those of parametric tests. However, they have one advantage over parametric tests: they do not require the assumption of normality of the distribution.

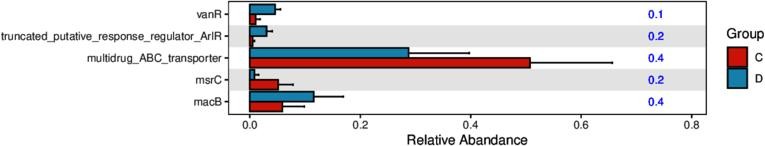


Figure 9.3.5 Bar graph of significantly different resistance genes

metagenomeSeq It is R A package developed by , the basic idea is to first standardize the data, then use zero - inflated Gaussian distribution to deal with the impact of sequencing depth, and finally find the difference based on the linear model . It is mainly used to compare the difference between two groups of samples, and whether the number of samples in each group is consistent has no effect on the result.

the ASV/OTU between the corresponding two groups Multiples​ change; log2 of FC)

Value, positive value represents B Group A Group upregulation (A The group refers to the group name in front of the folder name ) , and a negative value means a downward adjustment; the P value of the value is also given Pvalues and FDR - corrected P value (padj)



Figure 9.3.6 Heat map of significantly different resistance genes

Stamp The software is used for inter-group difference analysis, which is suitable for various types of inter-group difference analysis, such as species distribution and abundance differences, genetic differences, resistance gene differences, etc.

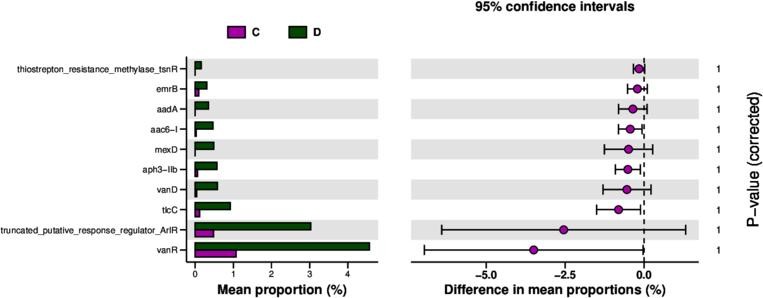
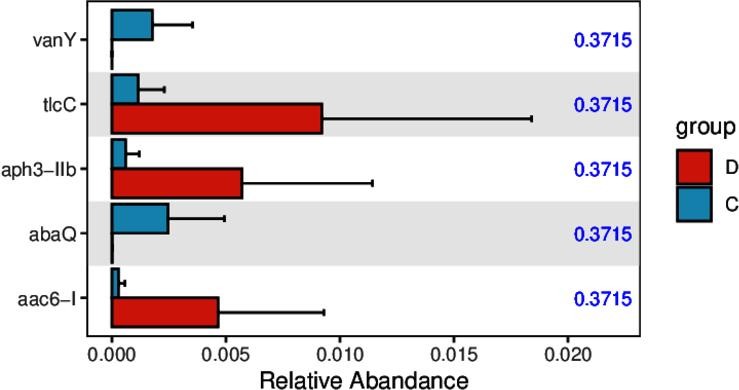


Figure 9.3.7 Difference test histogram

In order to study the resistance genes with significant differences between groups, we used the Metastats[31 ] method to perform hypothesis testing on the resistance gene abundance data between groups and obtained the p value. value , by The correction of the value is obtained value; finally according to q The resistance genes with significant differences were screened and a bar graph of the differential resistance genes was drawn .



9.3.8 Bar graph of significant difference resistance genes

## Virulence factor analysis

Substances derived from microorganisms that promote microbial infection and cause specific host diseases are called virulence factors . Virulence factor database mainly includes bacterial toxins, cell surface proteins that regulate bacterial adhesion, proteins that protect the bacteria themselves , cell surface carbohydrates , and hydrolases that cause bacterial pathogenicity. VFDB ( http://wwwc.ac.cn/VFS/ [) is a reliable , integrated network](http://www.mgc.ac.cn/VFs/) database [for the](http://www.mgc.ac.cn/VFs/) management [of pathogen virulence factor information .](http://www.mgc.ac.cn/VFs/)​

use DIAMOND [ 1,2 ] ( http://github.com/bbuchfink/diamond ) combines non - redundant gene sets with​​​​​​​​​​​​​​​​​​​​​​​​ V F D B core database comparison ( parameters: b l as t p ； E - v a lu e ≤ 1 e - 5 ), and obtain the functional annotation information of the virulence factor gene corresponding to the gene.

### Overview of relative abundance of virulence factor genes

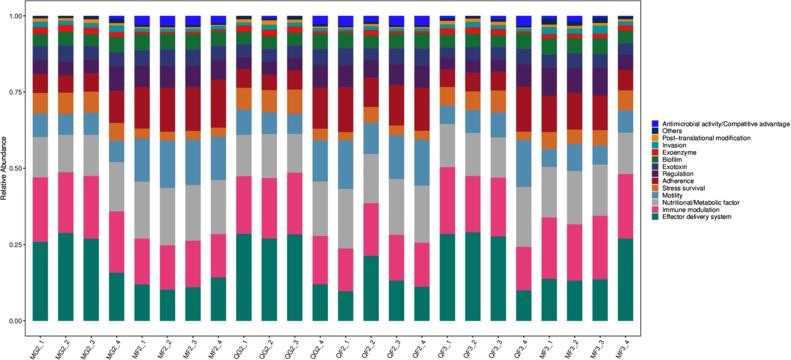


Figure 10.1.1 Virulence factor abundance bar graph

Heat map based on the virulence factor annotation and abundance information of all samples in the database



Figure 10.1.2 Virulence factor abundance heat map

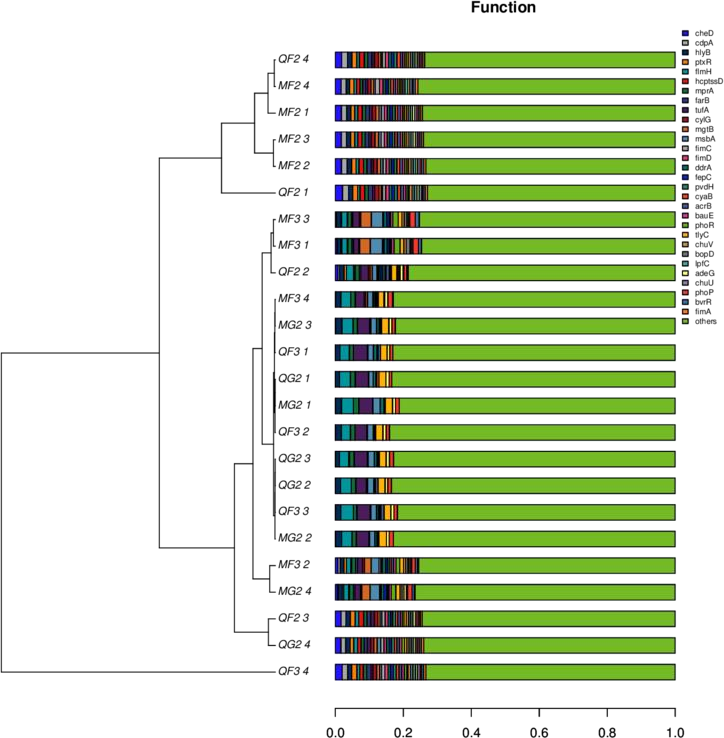
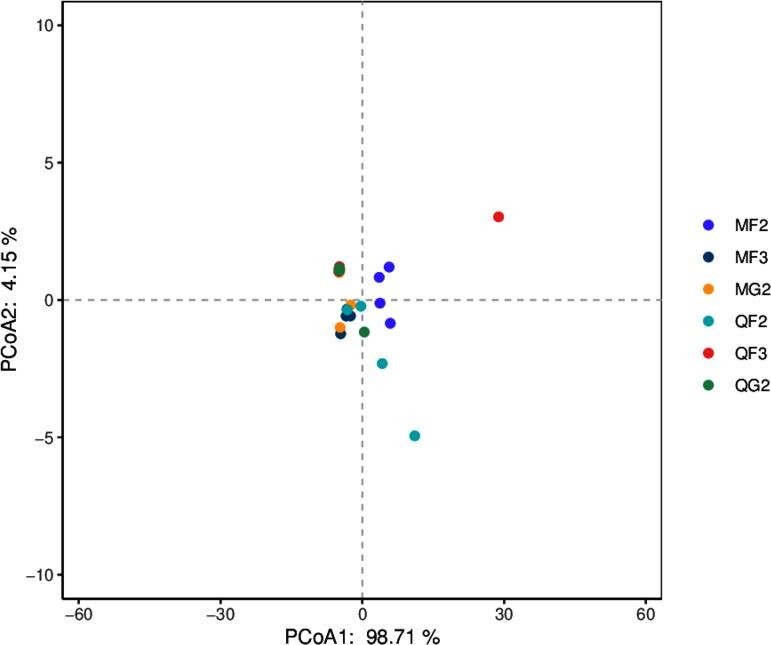
Based on the relative abundance information of all samples, we can calculate the Bray distance, then according to Bray To facilitate understanding, the virulence factor classification stacking diagram is compared with the Bray The clustering results of distances are displayed together .

Figure 10.1.3 Virulence factor clustering diagram

### Dimensionality reduction analysis based on functional abundance

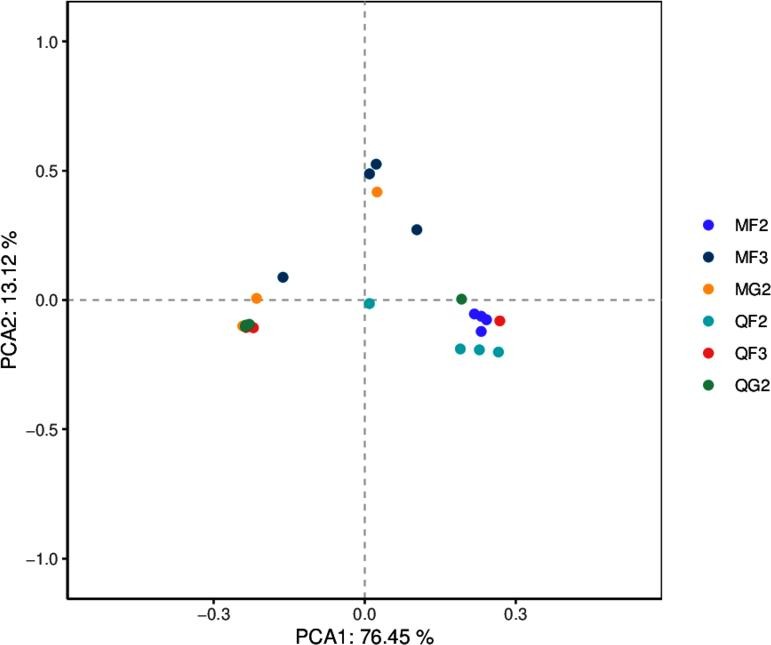
Principal coordinates analysis ( PCoA ) is a classic multidimensional Scaling , cMDScale) analysis method (Ramette, 2007) . It expands the sample distance matrix in a low-dimensional space after projection, and retains the distance relationship of the original sample to the maximum extent. PCoA considers the sample distance as a whole. Compared with principal component analysis ( PCA ) , it is more in line with the characteristics of ecological data. Therefore, it is more recommended as a sorting analysis method.



* + 1. PCoA Analysis chart

Note: The dots represent the functional composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample difference; the vertical axis represents the second principal component and its contribution to the sample difference.

Principal component analysis​​​​​​ Component​​​​​​ An a l y s is ,​ PCA ) is a technique for analyzing and simplifying data sets. By decomposing the variance, the differences between multiple groups of data are reflected on a two-dimensional coordinate graph. The coordinate axes take the two eigenvalues that can best reflect the variance . By analyzing the composition of different samples ( 97 % similarity ), the differences and distances between samples can be reflected. The closer the distance between two samples on the graph, the more similar the functional composition of the two samples is. R Language tools plot different classification levels separately P C A Analysis diagram, the sample diagram is as follows:



* + 1. PCA Analysis chart

Note: The dots represent the functional composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample difference; the vertical axis represents the second principal component and its contribution to the sample difference.

Non-metric multidimensional scaling (N M D S ) and the above P C o A The analysis is similar to that of the previous one, which also reduces the dimension of the sample distance matrix and simplifies the data structure, thereby describing the distribution characteristics of the samples at a specific distance scale. P C o A Different analysis, N M D S The analysis does not rely on the calculation of eigenvalues and eigenvectors, but rather on ranking the sample distances so that the order of samples in the low-dimensional space is as close as possible to the distance between each other ( rather than the exact distance value ) . The analysis is not affected by the numerical value of the sample distance, and only considers the size relationship between each other. For data with complex structures, the sorting results may be more stable. N M D S The smaller the stress value ( S t ress ) of the result , the better. It is generally believed that when the value is less than 0.2​ When N M D S The results of the analysis are relatively reliable ( L e g e nd re, 199 8 ) .



* + 1. NMDS Analysis chart

Note: The dots represent the functional composition of each sample; the horizontal axis represents the first principal component and its contribution to the sample difference; the vertical axis represents the second principal component and its contribution to the sample difference.

### Analysis of species contribution to genes

Based on the gene sets screened from the metagenomic data of this project and the species information corresponding to the genes and sets, the contribution of species to virulence factor genes is analyzed , and the results are visualized through stacked bar charts:

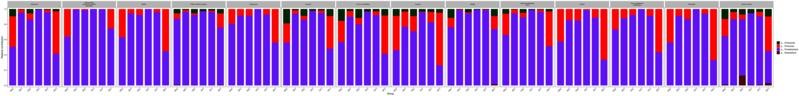


Figure 10.5.1 Species' contribution to genes

Note: The figure above shows the main species composition of specific functions (represented by the relative abundance of species ) , and the horizontal axes in the figure correspond to the sample groups.

### Analysis of intergroup differences based on abundance of virulence factors

Anosim Analysis is a non-parametric test used to test whether the difference between groups is significantly greater than the difference within the group, so as to determine whether the grouping is meaningful. The detailed calculation process can be viewed in Anosim . The results of Anosim analysis based on the abundance of resistance genes are as follows:

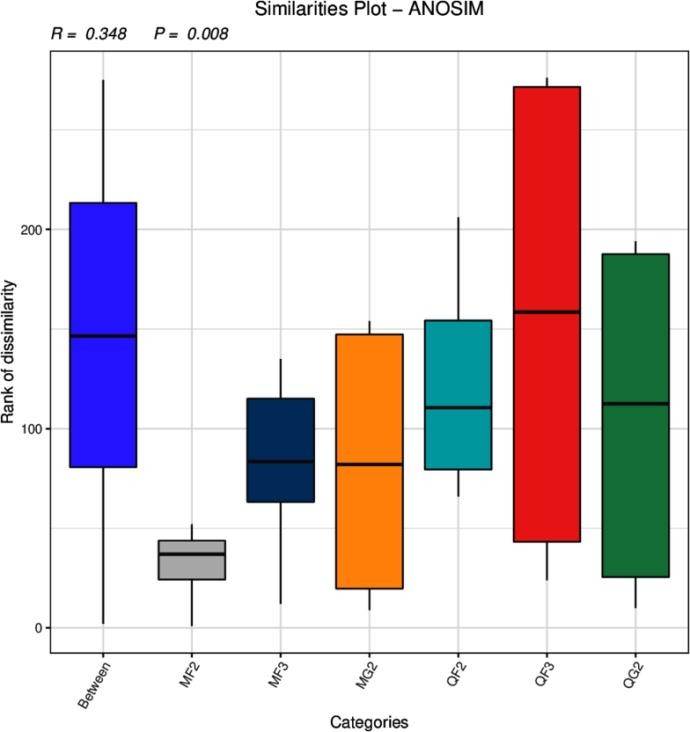


Figure 10.4.1 Anosim based on abundance of virulence factors analyze

Note: R-value is between ( -1 , 1 ). If R-value is greater than 0 , it means that the difference between groups is significant. If R-value is less than 0 , it means that the difference within the group is greater than the difference between the groups. The credibility of statistical analysis is expressed by P-value. Indicates, P < 0.05 Indicates that the data are statistically significant.

The Kruskal-Wallis rank sum test is a nonparametric test for three or more groups of data that can calculate the existence of differences between the means of multiple independent groups of data. It only deals with the volatility of the data displayed on the graph, which is different from ANOVA. In contrast, so-called nonparametric tests make no assumptions about the distribution of the data . Rank sum test We can use multiple sets of data from multiple biological replicates to determine whether the resistance genes are different. Here we consider that p\_value less than 0.05 is considered to be different.

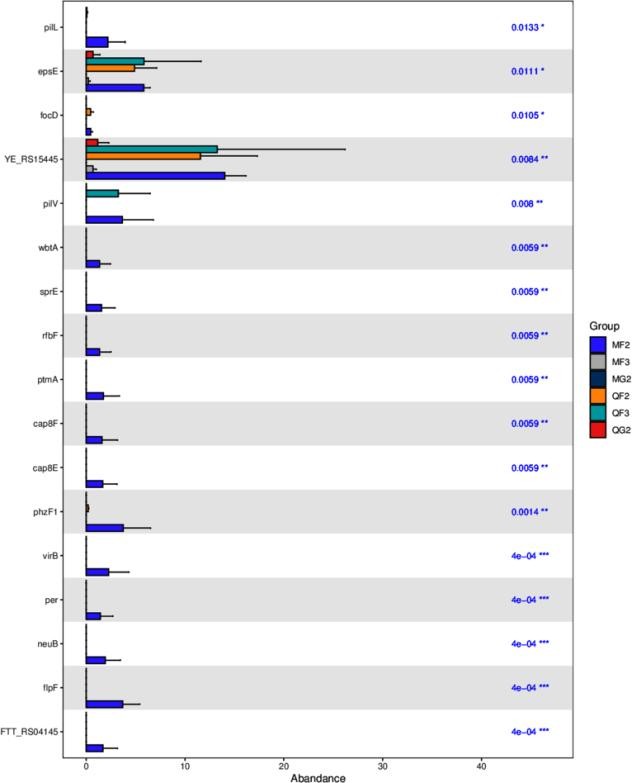


Figure 10.4.2 Kruskal-Wallis method based on abundance of virulence factors analyze

Note: The horizontal axis of the bar graph represents the relative abundance of a virulence factor in different groups, the vertical axis represents the differential virulence factor genes, and different colors represent different groups.

ANOVA (Analysis of variance) is abbreviated as analysis of variance or " coefficient of variation analysis " . It is used to test the significance of the difference between two or more sample means. It is mainly used to test whether the difference between two or more sample means is statistically significant or to test whether there is interaction between two or more factors.

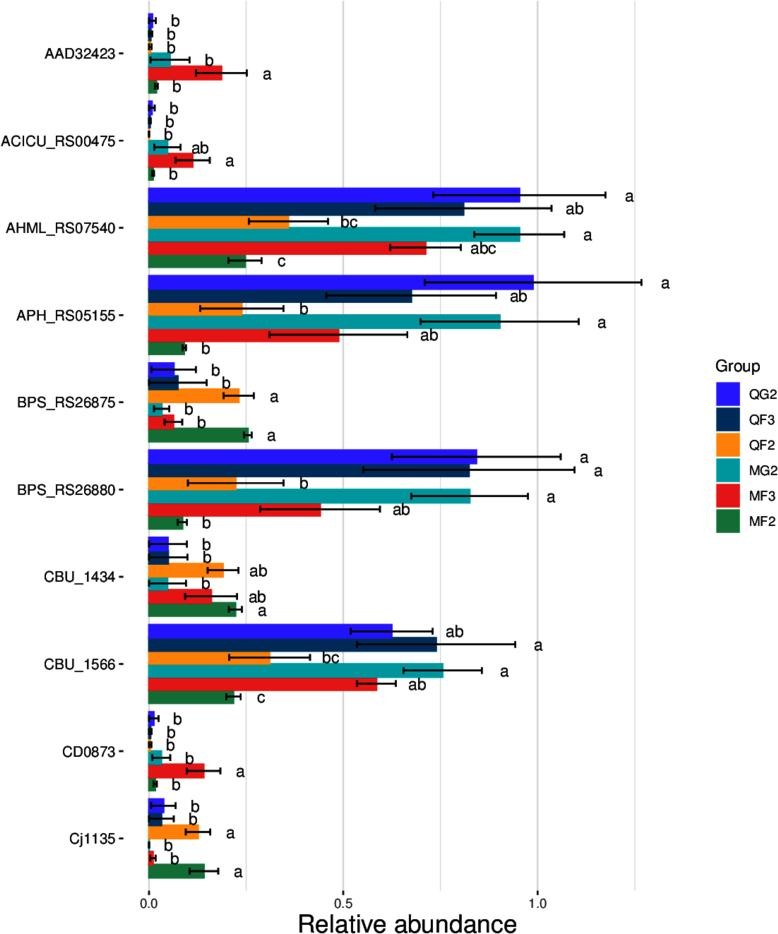


Figure 10.4.3 Anova based on virulence factor gene abundance analyze

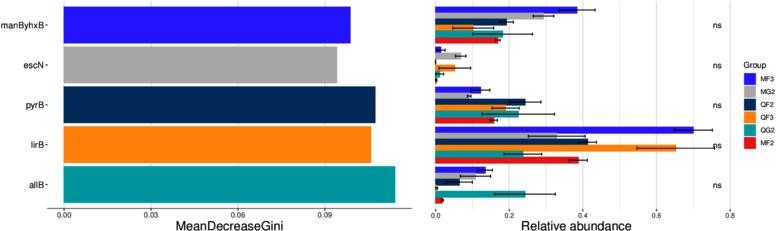
Random forest is a highly flexible machine learning method. It is an algorithm that integrates multiple trees through the idea of ensemble learning . Its basic unit is the decision tree, and its essence belongs to a major branch of machine learning - ensemble learning method . Each decision tree is a classifier ( assuming that the problem is classification). For an input sample, N A tree will have N Random forest integrates all classification voting results and designates the category with the most votes as the final output. This is the simplest Baggin​​​​ Thought.

图 10.6.4 基于毒力因子基因丰度的 Random\_Forest 分析

Will Coxon​​​​​ Inspection (also known as M a nn - W ithn e y - W i l c o x o n The χ2 test ) is a nonparametric test, meaning that it does not rely on the data belonging to a probability distribution family with any specific parameters. The goals of nonparametric tests are the same as those of parametric tests. However, they have one advantage over parametric tests: they do not require the assumption of normality of the distribution.

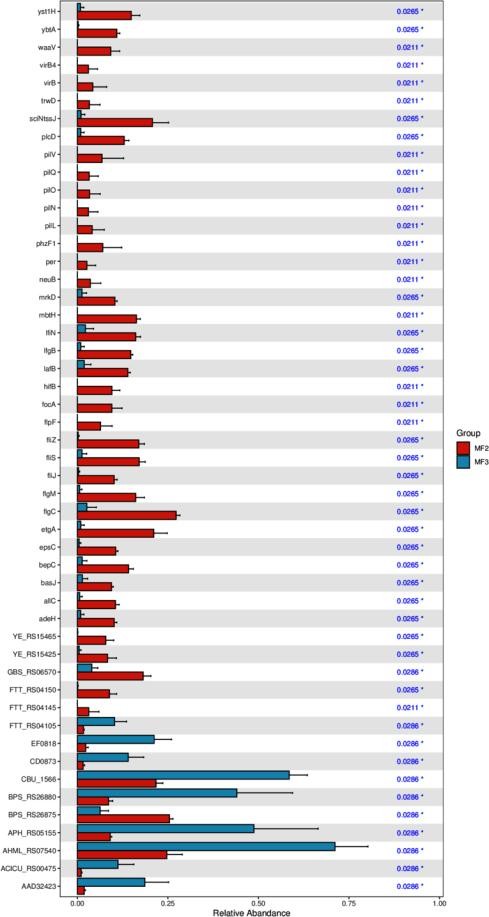


Figure 10.4.5 Column chart of significantly different virulence factor genes

metagenomeSeq It is R A package developed by , the basic idea is to first standardize the data, then use zero - inflated Gaussian distribution to deal with the impact of sequencing depth, and finally find the difference based on the linear model . It is mainly used to compare the difference between two groups of samples, and whether the number of samples in each group is consistent has no effect on the result.

the ASV/OTU between the corresponding two groups Multiples​ change; log2 of FC)

Value, positive value represents B Group A Group upregulation (A The group refers to the group name in front of the folder name ) , and a negative value means a downward adjustment; the P value of the value is also given Pvalues and FDR - corrected P value (padj)



Figure 10.4.6 Heat map of significantly different virulence factor genes

Stamp The software is used for inter-group difference analysis, which is suitable for various types of inter-group difference analysis, such as species distribution and abundance differences, genetic differences, resistance gene differences, etc.

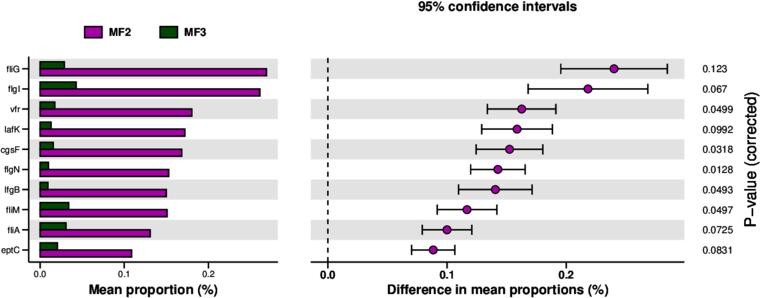
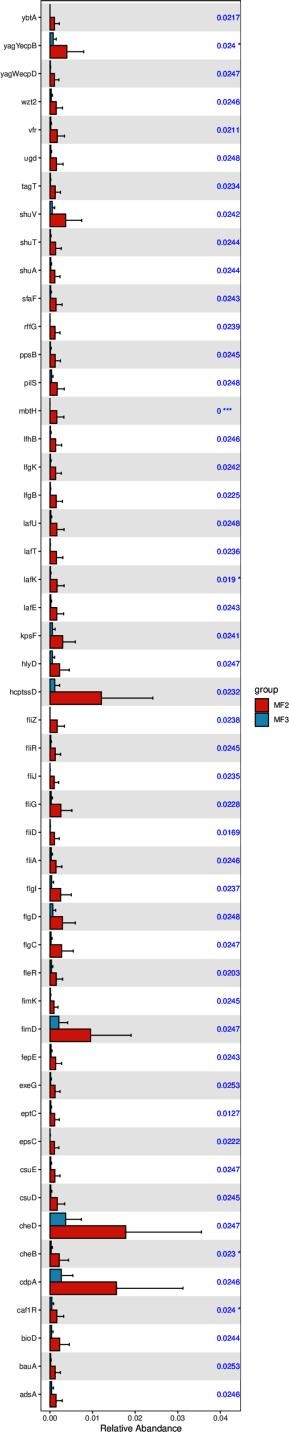


Figure 10.4.7 Difference test histogram

In order to study the virulence factor genes with significant differences between groups , we used the Metastats[31 ] method to perform hypothesis testing on the virulence factor gene abundance data between groups and obtained the p value. Value, through

P​ The correction of the value is obtained value; finally according to q The virulence factor genes with significant differences were screened and a bar graph of the differential virulence factor genes was drawn.



10.4.8 Column chart of significantly different virulence factor genes

## Metagenomic binning

Binning​​​​ The meaning of binning and clustering is to separate the sequences of different individuals from the sequences of microbial populations . or The process of separating the sequences from the same strain in the metagenomic data. That is , the sequences from the same strain in the metagenomic data are grouped together to obtain the genome of a strain . The technical principle is as follows:

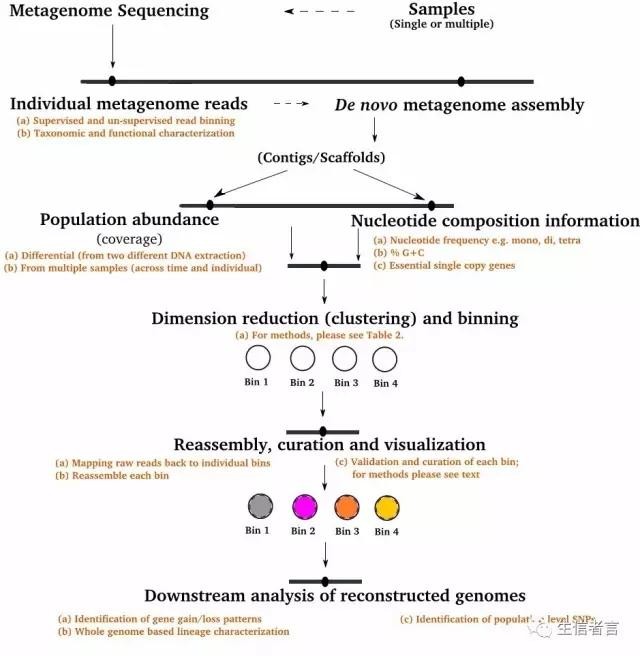


Figure 11.1 Binning Technical principle

Using vamb (2021, Nature Biotechnology, https://github.com/RasmussenLab/vamb) for Binning Analyze and screen the results. The screening criteria are as follows: genome integrity > 50%; contamination rate less than 10% ：

|  |  |  |  |
| --- | --- | --- | --- |
| **Bin\_Id** | **Marker\_Lineage** | **Completeness** | **Contamination** |

|  |  |  |  |
| --- | --- | --- | --- |
| bin\_S21C842 | o Lactobacillales | 98.6891385768 | 0.187265917603 |
| bin\_S24C812 | o Burkholderiales | 97.6616043614 | 0.49844236760099997 |

The genome integrity is >50%; the contamination rate is less than 10% of the standard screened bin The results are shown in the table below:

Note: Bin\_Id : bin Number; completeness : genome completeness; contamination : contamination rate; GC : GC content ; Marker\_Lineage : bin Classification status: Scaffolds : bin Scaffolds corresponding to the genome assembly size : the genome size corresponding to the bin , Predicted\_Genes : the number of genes predicted by the bin . Since the table file is too large, only part of the table content is shown here.

Using NCBI of nt The database is for bins with genome integrity > 50% and contamination rate less than 10% . Perform species annotation.

# References

[1]. Bolger A M, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data[J]. Bioinformatics, 2014, 30(15):2114-2120.

[2].Li D, Liu C M, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph[J]. Bioinformatics, 2015, 31(10): 1674-1676.

[3].Li D, Luo R, Liu C M, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices[J]. Methods, 2016, 102: 3-11.

[4].Land M L, Locascio PF, Chen G L, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification[J]. Bmc Bioinformatics, 2010, 11(1):119-119.

[5].Qin J J, Li R Q, Raes J J, et al. Qin J, Li R, Raes J, et al. A human gut microbial gene catalog established by metagenomic sequencing[J]. Nature, 2010, 464(7285):59-65.

[6]. Zeller G, Tap J, Voigt A Y, et al. Potential of fecal microbiota for early-stage detection of colorectalcancer[J]. Molecular Systems Biology, 2015, 10(11):766.

[7]. Villar E, Farrant G K, Follows M, et al. Ocean plankton. Environmental characteristics of Agulhas rings affect interocean plankton transport.[J]. Science, 2015, 348(6237):1261447.

[8]. Cotillard A, Kennedy S P, Ling C K, et al. Corrigendum: Dietary intervention impact on gut microbial gene richness[J]. Nature, 2013, 500(7464):585-8.

[9].Le C E, Nielsen T, Qin J, et al. Richness of human gut microbiome correlates with metabolic markers[J]. Nature, 2013, 500(7464):541-6.

[10]. Benjamin Buchfink, Chao Xie & Daniel H. Huson, Fast and Sensitive Protein Alignment using DIAMOND, Nature Methods, 12, 59 – 60 (2015) doi:10.1038/nmeth.3176.

[11]. Ondov B D, Bergman N H, Phillippy A M. Interactive metagenomic visualization in a Web browser.[J]. BMC Bioinformatics,12,1(2011-09-30), 2011, 12(1):385.

[12].Avershina E, Frisli T, Rudi K. De novo Semi-alignment of 16S rRNA Gene Sequences for Deep Phylogenetic Characterization of Next Generation Sequencing Data[J]. Microbes & Environments, 2013, 28(2):211-216.

[13].Noval RM, Burton OT, Wise P, et al. A microbiota signature associated with experimental food allergy promotes allergic Sensitization and Anaphylaxis[J]. Journal of Allergy & Clinical Immunology, 2013, 131(1):201-212.

[14].White JR, Nagarajan N, Pop M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples[J]. Plos Computational Biology, 2009, 5(4):e1000352.

[15]. Segata N, Izard J, Waldron L, et al. Metagenomic Biomarkers discovery and explanation.[J]. Genome Biology, 2011, 12(6):R60.

[16].Kanehisa M, Goto S, Hattori M, et al. From Genomics to chemical Genomics: new developments in KEGG.[J]. Nucleic Acids Research, 2006, 34(Database issue):FT454.

[17].Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG.[J]. Nucleic Acids Research, 2014, 42(Database issue):199-205.

[18].Powell S, Forslund K, Szklarczyk D, et al. eggNOG v4.0: nested orthology inference across 3686 organisms[J]. Nucleic Acids Research, 2014, 42(Database issue):231-9.

[19]. Cantarel B L, Coutinho P M, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics[J]. Nucleic Acids Research, 2009, 37(Database issue):FT233.

[20]. Backed F, Roswall J, Peng Y, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life.[J]. Cell Host & Microbe, 2015, 17(5):690-703.

[21].Qin J, Li Y, Cai Z, et al. A Metagenome-wide association study of gut microbiota in type 2 diabetes.[J]. Nature, 2012, 490(7418):55-60.

[22].Karlsson FH; Tremaroli V; Nookaew I; Bergström G; Behre CJ; Fagerberg B; Nielsen J; Bäckhed F. Gut metagenome in European women with normal, impaired and diabetic glucose control.[J]. Nature, 2013, 498(7452):99-103.

[23].Karlsson F H, Fåk F, Nookaew I, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome[J]. Nature Communications, 2012, 3(4):1245.

[24].Xie C, Mao X, Huang J, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases[J]. Nucleic Acids Research, 2011, 39(Web Server issue):316-22.

[25].Mistry J, Finn RD, Eddy SR, et al. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions.[J]. Nucleic Acids Research, 2013, 41(12):e121-e121.

[26].Altschul, SF, Gish, W., Miller, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[27].Chen L, Jian Y, Yu J, et al. VFDB: a reference database for bacterial Virulence factors[J].

Nucleic Acids Research, 2005, 33(Database issue):FT425- FT428.

[28].WinneMW1urg R, Baldwin T K, Urban M, et al. PHI-base: a new database for pathogen host interactions.[J]. Nucleic Acids Research, 2006, 34(Database issue):M1G59.