CD Genomics
The Genomics Services Company

Metagenomic Shotgun Sequencing Report

Order#: XXXXXXXX

Date: XX/XX/XXXX

CD Genomics
The Genomics Services Company

Table of Contents

1. Experimental procedures of Metagenomic sequencing

1.1 Sample Testing

There are mainly three methods in QC for DNA samples:

(1) Analysis of DNA purity and integrity by agarose gel.

(2) DNA purity (OD260/OD280) was detected using the Nanodrop.

(3) DNA concentration was accurately quantified using the Qubit 2.0.

1.2 Library Construction

A total amount of 1μg DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using NEBNext® Ultra™ DNA Library Prep Kit for Illumina (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, the DNA sample was fragmented by sonication to a size of 350bp,then DNA fragments were end-polished, A-tailed, and ligated with the full-length adaptor for Illumina sequencing with further PCR amplification. At last, PCR products were purified (AMPure XP system) and libraries were analysed for size distribution by Agilent2100 Bioanalyzer and quantified using real-time PCR.

1.3 Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq platform and paired-end reads were generated.

2. Results

2.1 Basic Bioinformatics Analysis

2.1.1 Raw Data

The original data obtained from the high throughput sequencing platforms are transformed to sequenced reads by base calling. Raw data are recorded in a FASTQ file which contains sequenced reads and corresponding sequencing quality information. Every read in FASTQ format is stored in four lines like shown in follows:

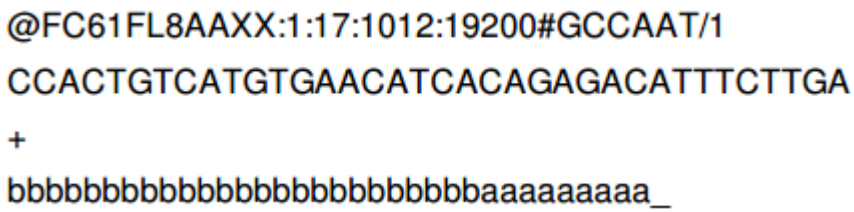


Figure 1. Schematic of FASTQ format file

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (such as a FASTA title line).

Line 2 is the sequence of the read.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the bases in Line 2.

2.1.2 Data Quality Control

The sequenced reads (raw reads) often contain low quality reads and adapters, which will affect the analysis quality. So it's necessary to filter the raw reads and get the clean reads. The filtering process is as follows:

(1) Remove reads containing adapters.

(2) Remove reads containing N > 10% (N represents the base that cannot be determined).

(3) Remove reads containing low quality (Qscore <= 5) base which is over 50% of the total base.

High-quality Clean Data, obtained after a series of quality controls described above, is available in the FASTQ format.

Table 1. Statistics of the data generated in the sequencing.

Sample	Raw Reads	Bases	GC (%)	Q20	Q30
H1	52237136	7826000861	63.56	99.13	94.60
H2	53729316	8052138314	61.08	99.14	94.60
H3	59643516	8939934551	61.28	99.09	94.40
H4	56844416	8521938164	62.16	99.12	94.45
H5	57096940	8558489324	63.11	99.10	94.38
T1	53064584	7953693124	64.19	99.05	94.15
T2	53052640	7948523002	62.78	99.09	94.40
T3	52104816	7807643134	62.61	99.19	94.94
T4	53349350	7996976406	62.93	99.08	94.29
T5	50861804	7620698212	62.75	99.18	94.88

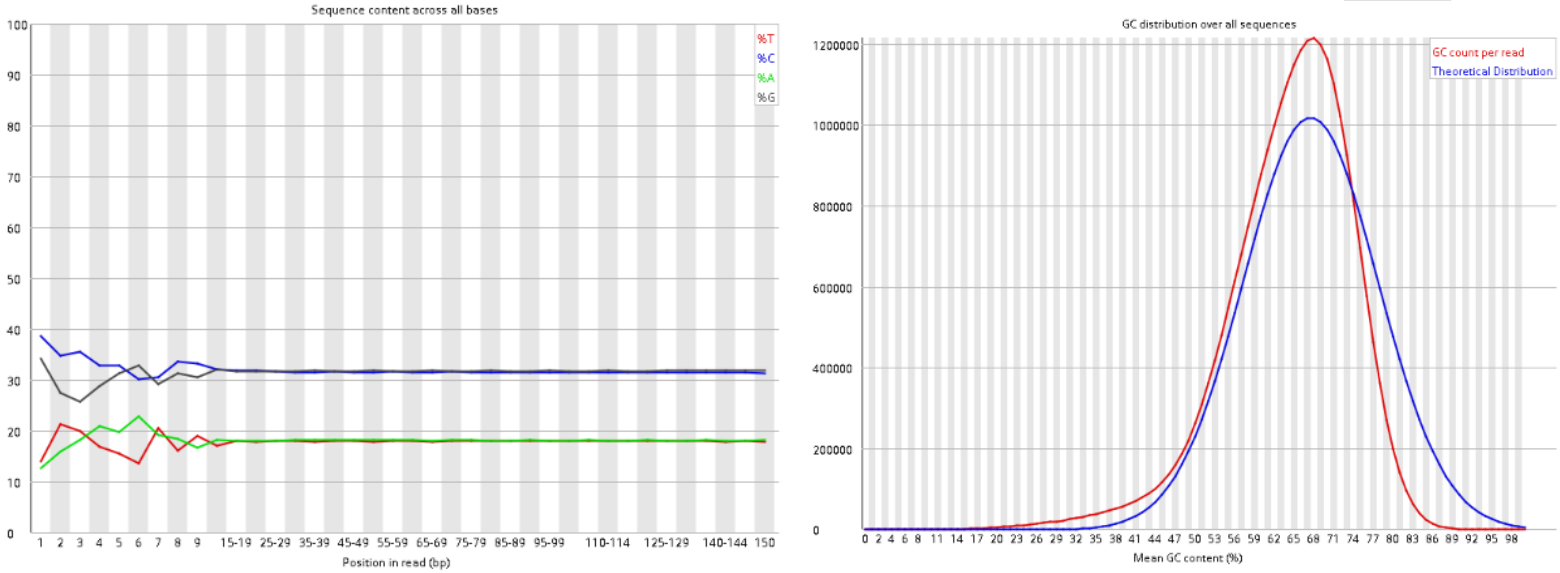


Figure 2. Per base sequence content (the left figure) and Per sequence GC content (the Right figure) for H1

2.2 Metagenome Assembly

1) Clean data is obtained after preprocessing, and assembly analysis is performed using megaHit [1] assembly software;

2) Combine all sample data. When megaHit is assembled, multiple parameters of K-mer = 21 ~ 141 are selected for assembly, and then the optimal result is selected to obtain the final assembly result;

3) Scaffold generated by mixed assembly, retains sequences longer than 500bp, and performs statistical analysis and subsequent gene prediction.

Table 2. Statistics of the assembly results.

Sample ID	# contigs	Largest contig	Total length	GC (%)	N50	N75	L50	L75
H1	194066	217096	1.91E+08	64.75	955	641	46579	109005
H2	177715	224837	1.77E+08	60.85	987	654	43448	99620
H3	204888	147959	2.02E+08	61.22	962	647	49866	115209
H4	174599	602973	2.11E+08	62.3	1315	703	25838	83737
H5	194463	277548	2.08E+08	63.48	1058	669	39484	103197
T1	28223	4252	18712551	63.52	630	550	11287	19274
T2	18188	15701	11672324	58.12	607	543	7461	12570
T3	29608	5139	21239969	60.18	667	559	10738	19520
T4	33765	8585	23335971	60.25	649	556	12880	22659
T5	29458	3891	19769917	61.39	637	552	11647	20022

Note: The “Sample ID” is a unique identifier assigned to each sample or dataset that has been subjected to metagenomic sequencing and subsequent assembly. The "# Contigs" column indicates the total number of contigs present in the assembled dataset. The "Largest Contig" refers to the length of the longest contig in the assembled dataset. The "Total Length" is the cumulative length of all the contigs in the assembly. “GC content” stands for Guanine-Cytosine content and is a measure of the proportion of nitrogenous bases in a DNA or RNA sequence that are either guanine (G) or cytosine (C). It's usually expressed as a percentage. N50 is a measure that represents the length at which the cumulative size of contigs (or sequences) reaches 50% of the total assembly length. Similar to N50, N75 represents the length at which

the cumulative size of contigs reaches 75% of the total assembly length. L50 is the number of contigs (or sequences) required to reach the N50 value. L75 is the number of contigs required to reach the N75 value, indicating how many contigs contribute to the first 75% of the assembly's total length.

2.3 Metaphlan Species Annotation

The analysis result of species annotation is visually shown by Metaphlan [2]. The result table is as follows:

Table 3. Statistics of the species annotation results.

Sample	Files
H1	05.Annotation\HPC1\metaphlan\metaphlan_anno.txt
H2	05.Annotation\HPC2\metaphlan\metaphlan_anno.txt
H3	05.Annotation\HPC3\metaphlan\metaphlan_anno.txt
H4	05.Annotation\HPC4\metaphlan\metaphlan_anno.txt
H5	05.Annotation\HPC5\metaphlan\metaphlan_anno.txt
T1	05.Annotation\TBFI1\metaphlan\metaphlan_anno.txt
T2	05.Annotation\TBFI2\metaphlan\metaphlan_anno.txt
T3	05.Annotation\TBFI3\metaphlan\metaphlan_anno.txt
T4	05.Annotation\TBFI4\metaphlan\metaphlan_anno.txt
T5	05.Annotation\TBFI5\metaphlan\metaphlan_anno.txt

Sample	Files
Taxonomy Annotation	05.Annotation\metaphlan

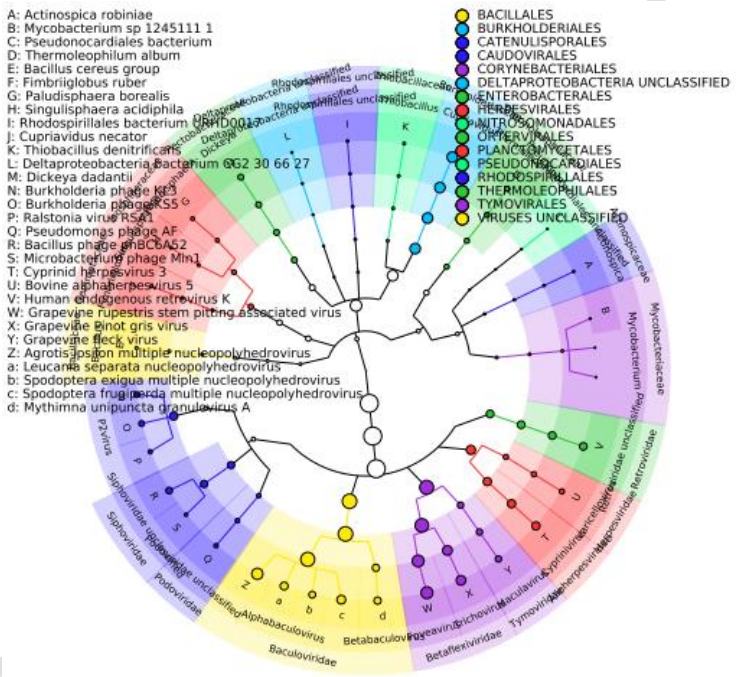


Figure 3. Merged_abundance. Different circles represent different taxonomic levels, from inside to outside, they are kingdom-phylum-class-order-family-genus-species. Each node represents a species, the more nodes larger means that the abundance of the species is higher, Colors mean different flora.

2.4 Taxonomy Distribution Histogram of All Samples

The taxonomy distributions histogram graph in each classification level (phylum, class, order, family, genus and species) are displayed in the below Figure. Each color represents a taxonomy, and the length of the color blocks indicates the proportion of the relative abundance of the taxonomy. In order to display the best view, the histogram shows only the abundance of the top ten taxonomy, and the other species are combined into 'Others' in the figure, 'Unknown' represents the taxonomy that has not been given annotations, the specific species information can be found in the corresponding species abundance table. The heatmap takes the family as an example, as shown in Figure 4.

Table 4. Statistics of the Taxonomy results

Level	Sample Abundance
Kingdom	\result\06.Taxonomy\level-1.csv
Phylum	\result\06.Taxonomy\level-2.csv
Class	\result\06.Taxonomy\level-3.csv
Order	\result\06.Taxonomy\level-4.csv
Family	\result\06.Taxonomy\level-5.csv
Genus	\result\06.Taxonomy\level-6.csv
Species	\result\06.Taxonomy\level-7.csv

Table 5. Krona charts of taxonomy results

Sample	Krona File
H1	\result\06.Taxonomy\krona\HPC1.krona.html
H2	\result\06.Taxonomy\krona\HPC2.krona.html
H3	\result\06.Taxonomy\krona\HPC3.krona.html
H4	\result\06.Taxonomy\krona\HPC4.krona.html

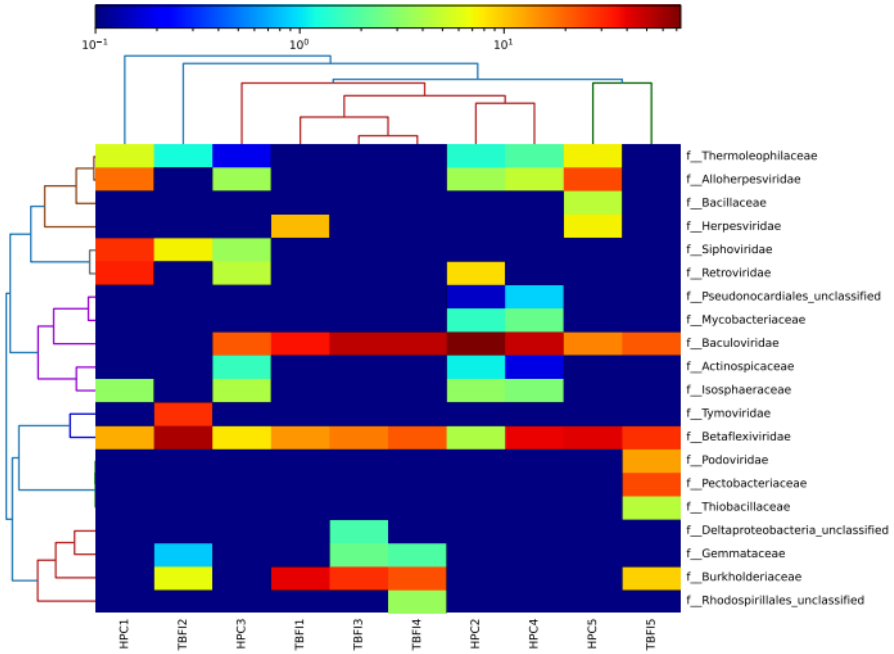


Figure 4. Abundance heatmap of family level

2.5 Functional Database Annotation

General databases for function annotations currently available include:

Kyoto Encyclopedia of Genes and Genomes (KEGG [3]); Version: 2018.01;

Evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG [4]); Version: 4.5;

Non-Redundant Protein Sequence Database (NR [5]);

Virulence Factors Database (VFDB [7]);

Transporter Classification Database (TCDB [8]);

Pathogen Host Interactions Database (PHI [9]);

Carbohydrate-Active enZYmes Database (CAZy [10]);

The Comprehensive Antibiotic Resistance Database (CARD [11]);

In the living body, different genes coordinate with each other to make their biological functions. The specific genes that involved in the major metabolic pathways and signal transduction pathways can be determined by Pathway significant enrichment analysis. KEGG is called Kyoto Encyclopedia of Genes and Genomes(http: / /www.genome.jp/ kegg/), it is the main public database of the pathways. A systematic analysis of the metabolic pathways of gene products and compounds in cells and the database of the function of these gene products. (KEGG PATHWAY), drug (KEGG DRUG), disease (KEGG DISEASE), functional model (KEGG MODULE), gene sequence (KEGG GENES) and the genome of the genome (KEGG GENOME) and so on. The KO (KEGG ORTHOLOG) system links the various KEGG annotation systems, and KEGG has developed a complete KO annotation system to annotate genomic or transcriptome functionalities of newly sequenced species.

EggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups, http: //eggnog.embl.de/) is a widely-recognized orthologous group of professional annotated databases, including COG/KOG functional classification and taxonomic functional annotations. Currently the database (v4.5.1) contains 1.9 million orthologous groups covering 2383 species (including 352 viruses). The sequence of the gene set was compared with the eggNOG database to obtain the Abbre corresponding to the gene, and then the Abbre abundance was calculated using the sum of Abbre corresponding gene abundances.

NR is called Non-Redundant Protein Database. It is a non-redundant protein database. It is created and maintained by NCBI. It is characterized by comprehensive content and species information in the annotation results, which can be used for species classification. But a lot of data in the database has not been verified, the reliability to be improved. The database is characterized by data is full, but not all of the features described are particularly accurate.

TCDB is a database for classifying Membrane Transport Protein. It has developed a Transporter Classification (TC System), which is similar to the EC system for classifying enzymes, except that the TC system provides both functional and evolutionary information.

PHI (Pathogen Host Interactions Database), whose contents have been experimentally verified, are mainly derived from fungal, oomycete, and bacterial pathogens. The infected hosts include animals, plants, fungi, and insects. The database plays an important role in the search for target genes for drug intervention, and the database also includes antifungal compounds and corresponding target genes. Each gene in the database contains nucleic acid and amino acid sequences, as well as detailed descriptions of protein functions predicted to infect the host.

CAZy (Carbohydrate-Active enZYmes) database describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. CAZy data are accessible either by browsing sequence-based families or by browsing the content of genomes in carbohydrate-active enzymes. New genomes are added regularly shortly after they appear in the daily releases of GenBank. New families are created based on published evidence for the activity of at least one member of the family and all families are regularly updated, both in content and in description.

CARD (Comprehensive Antibiotic Resistance Database, http://arpcard.Mcmaster.ca) contains a wide range of reference genes related to antibiotic resistance from various organisms, genomes, and plasmids, which can be used to guide the research of environmental, human, animal bacterial resistance groups and antibiotic resistance mechanisms.

2.5.1 Statistics of The Annotated Gene Numbers

Table 6. Statistics of the annotated gene numbers（H1）

DataBase	Annotated Number	Unannotated Number
CARD	257469	64478
CAZy	2916	319031

DataBase	Annotated Number	Unannotated Number
PHI	37354	284593
TCDB	26185	295762
VFDB	15156	306791
Total	266084	55863

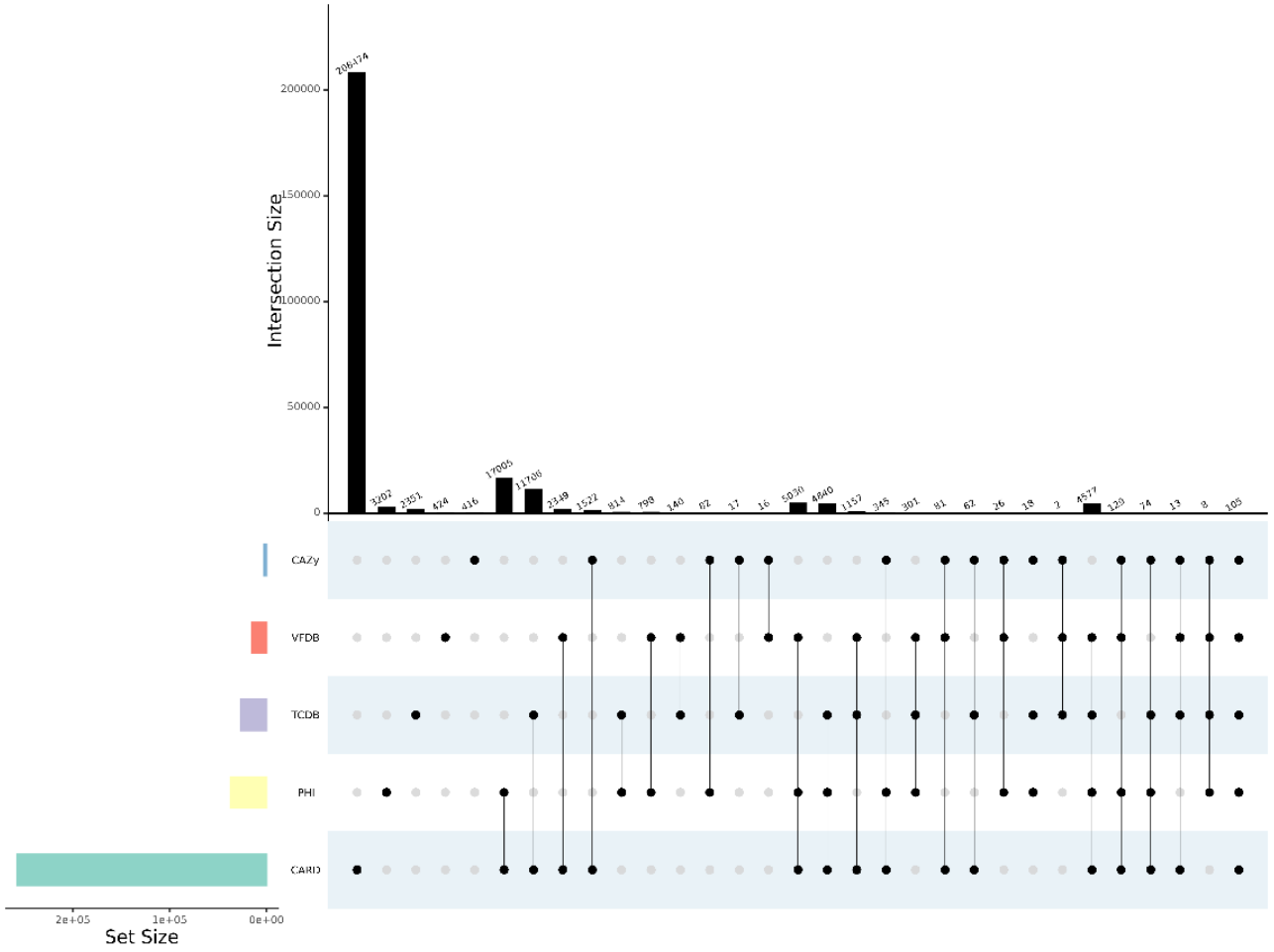


Figure 5. Statistical of specific function database common and unique annotation in H1.

2.5.2 KEGG

In the living body, different genes coordinate with each other to make their biological functions. The specific genes that involved in the major metabolic pathways and signal transduction pathways can be determined by Pathway significant enrichment analysis. KEGG is called Kyoto Encyclopedia of Genes and Genomes, it is the main public database of the pathways. A systematic analysis of the metabolic pathways of gene products and compounds in cells and the database of the function of these gene products. (KEGG PATHWAY), drug (KEGG DRUG), disease (KEGG DISEASE), functional model (KEGG MODULE), gene sequence (KEGG GENES) and the genome of the genome (KEGG GENOME) and so on. The KO (KEGG ORTHOLOG) system links the various KEGG annotation systems, and KEGG has developed a complete KO annotation system to annotate genomic or transcriptome functionalities of newly sequenced species. The annotation of differentially expressed genes is demonstrated in below figure:

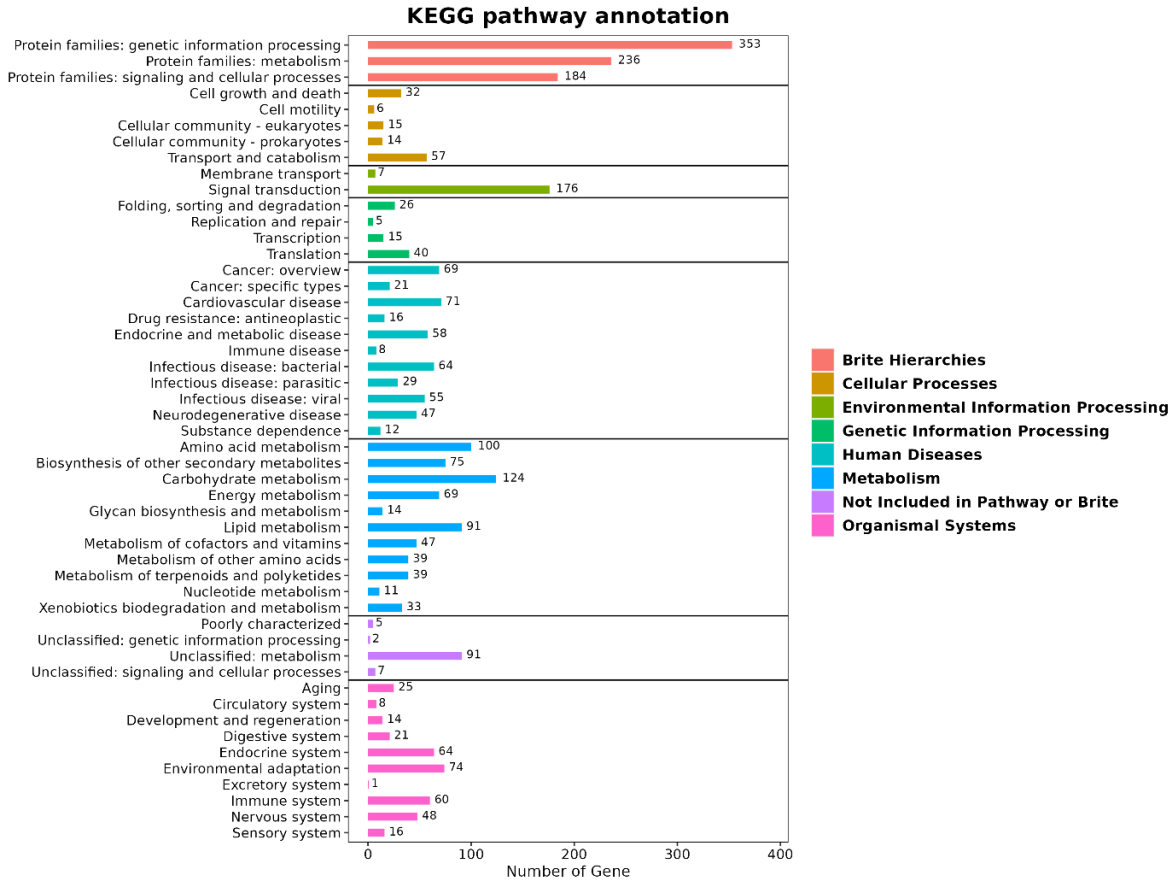


Figure 6. KEGG_classification. The height of the bars or data points on the x-axis indicates the number or proportion of genes or sequences that belong to each KEGG category. The y-axis labels represent different KEGG functional categories.

2.5.3 eggNOG

eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) is a widely used bioinformatics database and resource that provides functional annotation of genes based on orthology and evolutionary relationships. It is a part of the broader "Orthologous Groups" (OGs) databases family.

eggNOG classifies genes from various organisms into orthologous groups, which are sets of genes descended from a single common ancestral gene. By analyzing evolutionary relationships and similarities between genes, eggNOG assigns functional annotations to genes, helping researchers understand the functional properties and potential roles of genes in different biological processes.

File path: metagenome result\05.Annotation\T1\eggnog\eggnog.anno.xls

2.5.5 NR Annotation

The NR Database, also known as the "Non-Redundant Database" or "NCBI NR Database," is a comprehensive and fundamental resource in the field of bioinformatics and genomics. It is maintained by the National Center for Biotechnology Information (NCBI), a division of the National Institutes of Health (NIH) in the United States.

The NR Database contains a vast collection of non-redundant protein sequences compiled from various sources, including GenBank, Swiss-Prot, and Protein Information Resource (PIR). The content of this database provides researchers with a unique and diverse set of protein sequences, eliminating duplicates and representing each distinct protein only once.

File path: metagenome result\05.Annotation\T1\NR\nr_result.fmt6

Table 6. Statistics of NR database comparison (Top10)

Query Seq - id	Subject Seq - id	Percentage of identical matches	Alignment length	Expect value	unique Subject Taxonomy ID(s)	All Subject Title(s)
k95_17493:0-498	MSO67356.1	81.3	166	3.72E-98	--	XX
k95_122452:2-431	PYS70243.1	68.1	141	1.14E-41	1978231	XX
k95_244904:0-549	MBA2524687.1	88	183	2.17E-100	--	XX
k95_227411:2-644	RZL89876.1	84.1	214	2.08E-123	1871043	XX
k95_122453:1-166	MBR2119336.1	72.7	55	1.57E-16	--	XX
k95_157440:0-207	PYO92453.1	94	67	1.51E-34	2026742	XX
k95_157440:203-932	PYO72450.1	92.6	242	1.25E-148	2026742	XX
k95_157440:928-1066	PYP56854.1	84.8	46	1.10E-16	2026742	XX

Query Seq - id	Subject Seq - id	Percentage of identical matches	Alignment length	Expect value	unique Subject Taxonomy ID(s)	All Subject Title(s)
k95_262398:1-295	TMF08798.1	71.6	95	2.58E-37	2026724	XX
k95_262398:248-971	PYP16415.1	61.3	240	1.37E-101	2026742	XX

加 KEGG 和 ggNOG??

2.5.6 CARD Annotation

ARDB and CARD (The Comprehensive Antibiotic Resistance Database) are the two most widely used bacterial resistance gene databases. Although the ARDB database is comprehensive, it has stopped updating since 2009, and the CARD database includes the ARDB database. All resistance information is also updated monthly to ensure data validity. Therefore, we use the CARD database for drug resistance gene annotation. It is a rigorously curated collection of characterized, peer-reviewed resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO) and AMR gene detection models.

Use the blastp command of diamond (v0.9.12.113) tool to align the protein sequence of the predicted gene to the CARD database (v3.0.1), with an E-value<1e-5, and select the hit with the highest score as the final annotation result.

File path: metagenome result\05.Annotation\T1\card\card_result

2.5.7 CAZy Annotation

The CAZy database (Carbohydrate-Active enZYmes Database) describes the families of structurally related catalytic and carbohydrate- binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. It contains five main categories: Glycoside Hydrolases (GHs), GlycosylTransferases(GTs), Polysaccharide Lyases(PLs) and Carbohydrate Esterases(CEs), Auxiliary Activities(AAs).

We use dbCAN2 (web annotation tool for automated carbohydrate-related enzymes) dbCAN-HMMdb-V7 (dbCAN CAZyme domain HMM

database), annotation software HMMER (v3.1b2), parameter E- Value <= 1e-15, coverage >= 0.35 (refer to dbCAN2).

File path: metagenome result\05.Annotation\T1\dbCAN\CAZy_anno.xls

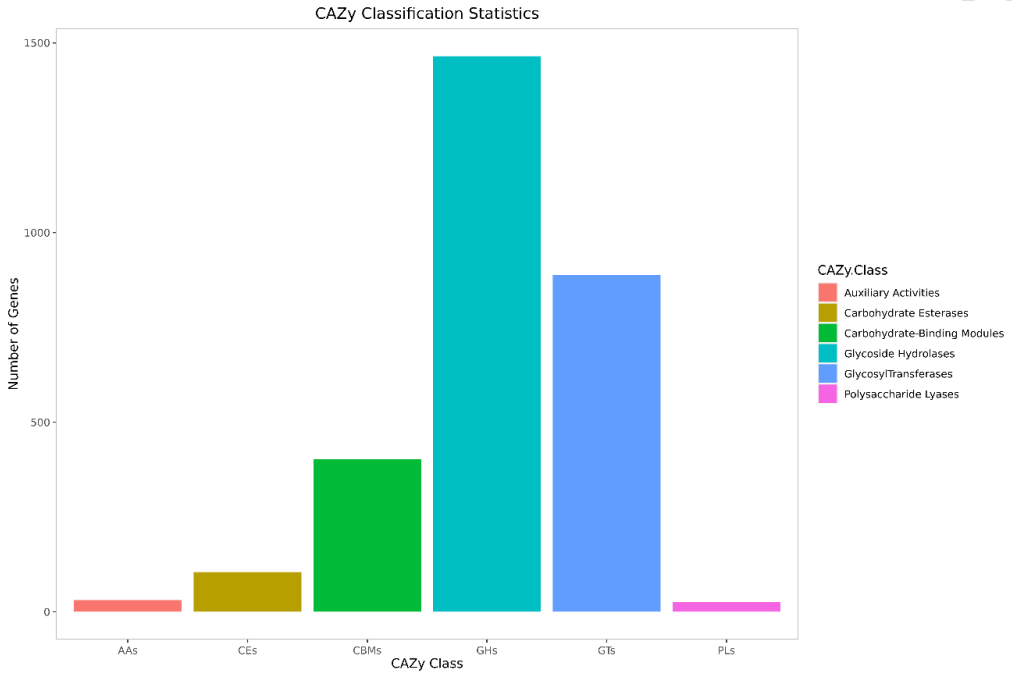


Figure 7. CAZy function classification for H1

2.5.8 PHI Annotation

PHI (Pathogen Host Interactions Database) is mainly derived from fungal, oomycete and bacterial pathogens. Infected hosts include animals, plants, fungi and insects. The database plays an important role in searching for target genes for drug intervention, and the database also includes antifungal compounds and corresponding target genes. Each gene in the database contains nucleic acid and amino acid sequences, as well as detailed descriptions of protein functions predicted during infection of the host.
Use the blastp command of diamond (v0.9.12.113) tool to align the protein sequence of the predicted gene to the PHI database, with an E- value<1e-5, and select the hit with the highest score as the final annotation result.

File path: metagenome result\05.Annotation\T1\phi\phi_anno.xls

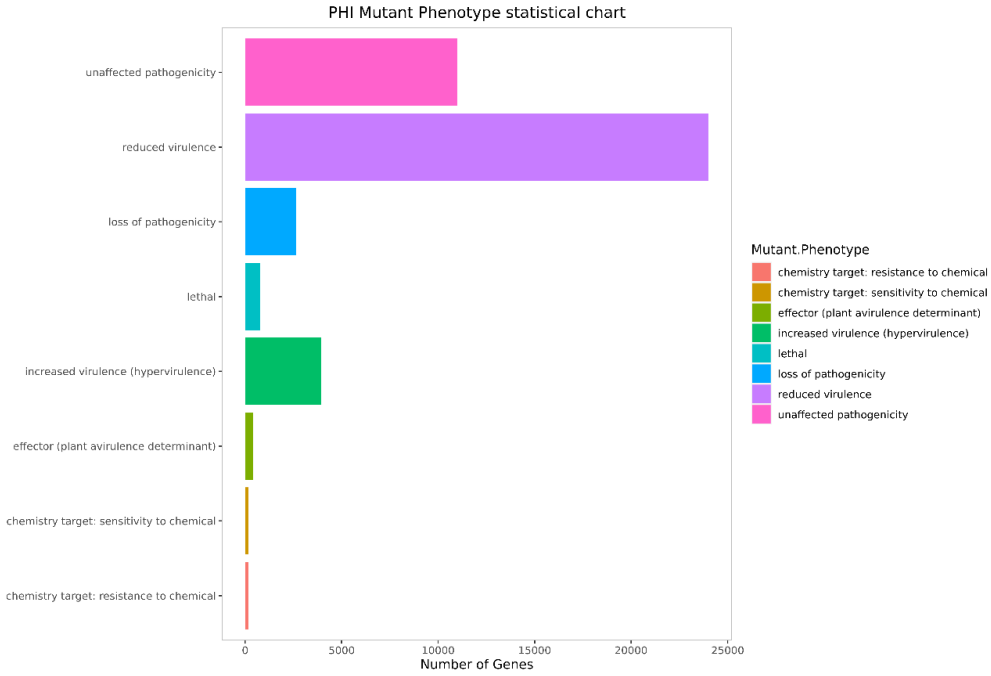


Figure 8. PHI phenotype classification for H1.

2.5.9 VFDB Annotation

The Virulence Factors of Pathogenic Bacteria database (VFDB) is an integrated and comprehensive online resource for curating information about virulence factors of bacterial pathogens.

The motivation for constructing VFDB was two fold:

First, to provide in-depth coverage major virulence factors of the best-characterized bacterial pathogens, with the structure features, functions and mechanisms used by these pathogens to allow them to conquer new niches and to circumvent host defense mechanisms, and cause disease.

Second, to provide current knowledge of the wide variety of mechanisms used by bacterial pathogens for researchers to elucidate pathogenic mechanisms in bacterial diseases that are not yet well characterized and to develop new rational approaches to the treatment and prevention of infectious diseases.

Use the blastp command of diamond (v0.9.12.113) tool to align the protein sequence of the predicted gene to the VFDB SetA library (Last updated: April 5, 2019), with an E-value<1e-5, and select the hit with the highest score as the final annotation result.

File path: metagenome result\05.Annotation\T1\vfdb\vfdb_result

2.5.10 TCDB Annotation

TCDB (Transporter Classification Database) details a comprehensive IUBMB approved classification system for membrane transport proteins known as the Transporter Classification (TC) system.

Use the blastp command of diamond (v0.9.12.113) tool to align the protein sequence of the predicted gene to the TCDB database, with an E-value<1e-5, and select the hit with the highest score as the final annotation result.

File path: metagenome result\05.Annotation\T1\tcdb\tcdb_anno.xls

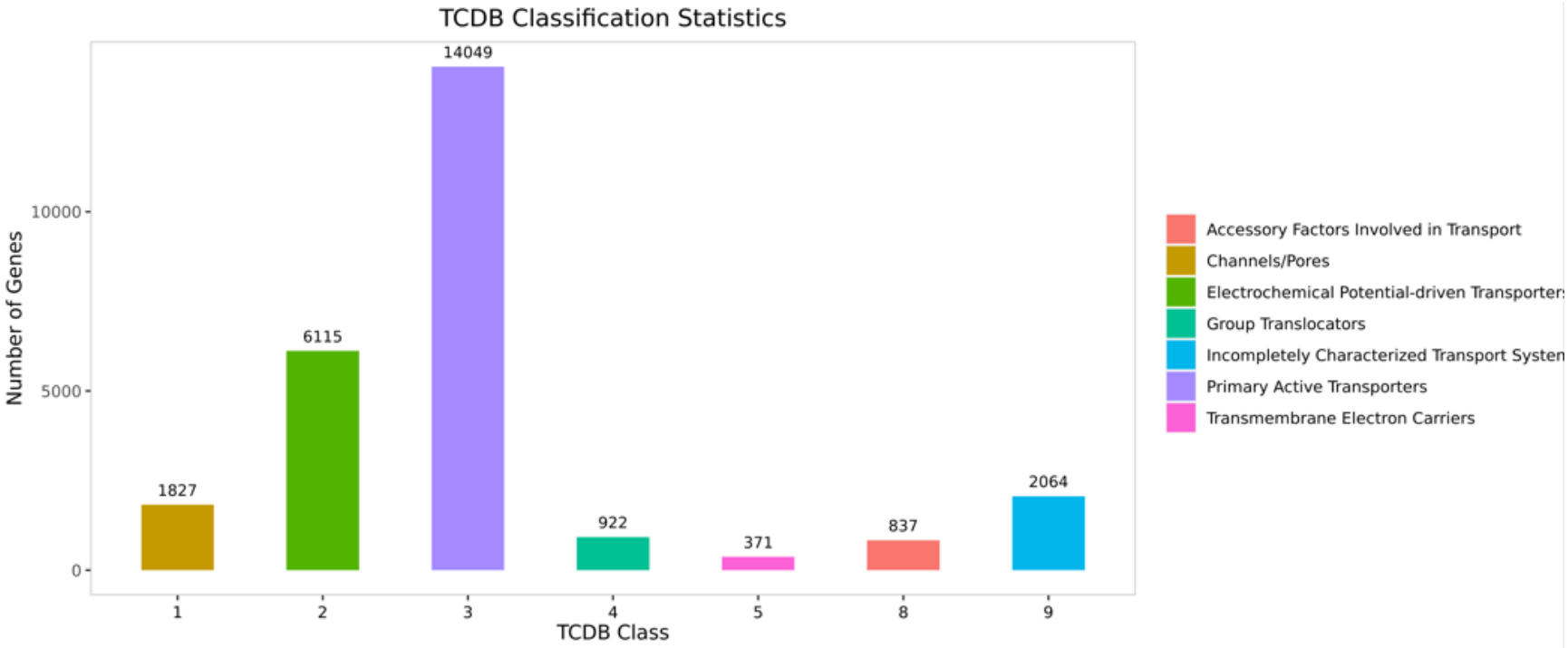


Figure 9. TCDB classification statistics for H1. The horizontal axis of the above figure (bar graph) is the transporter class, and the vertical axis is the number of genes annotated to the corresponding class. The legend corresponds to the definition of the transporter class; the below graphs correspond to the columns of the above bar graph one by one, and the fan-shaped area ratio represents the proportion of transporter subclasses in this transporter class.

2.6 Alpha Diversity Analysis

Microbial diversity can be assessed within a community (alpha diversity) or between the collections of samples (beta diversity). Four different metrics were calculated to assess the alpha diversity: Chao1 and Ace simply estimate the number of species in a community; Shannon and Simpson account for both richness and evenness of a community. Larger the Chao1, Ace and Shannon indices correspond to a smaller Simpson index value, indicating greater diversity of species [12]. In addition, the coverage of the sample library is reported. A higher value indicates a higher probability that the sequence is detected in the sample. The index reflects whether the results of this sequencing accurately represent the real population of microbes in the sample.

2.6.1 Statistical Data of Alpha Diversity

In order to compare the diversity indices between the samples, we have standardized the sequence number in each sample in the analysis process. At the level of 97% similarity, varied alpha metrics results were integrated and displayed on the following Table 7.

Table 7. Statistics of Alpha diversity indices (Top 10)

Sample	Observed species	ace	Chao 1	Simpson	Shannon
S145_1	910	910	910	0.654985035	3.914181389
S145_2	1891	1891	1891	0.830694297	5.83462709
S146_1	158	158	158	0.961684233	5.833334444
S146_2	87	87	87	0.55398131	1.616963214
S147_1	100	100	100	0.529850969	1.550806285
S147_2	802	802	802	0.594395889	2.948466552
S148_1	1141	1141	1141	0.789908199	4.70196164
S148_2	233	233	233	0.46312489	2.44111006
S149_1	3312	3312	3312	0.998912511	10.74202729
S149_2	199	199	199	0.538326826	1.745920168

2.6.2 Rarefaction Curve

Rarefaction curve [13] is created by random selection of a certain amount of sequencing data from the samples, then counting the number of the species these data represent. The left-side of the steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individual samples have been taken, suggesting that more intensive sampling is likely to yield only few additional species. The rarefaction curve can be used to judge the sequencing sufficiency of each sample. A sharp rise of the curve indicates that sequencing quantity is insufficient, and more reads are required.

2.7 Beta Diversity Analysis

Beta diversity represents the explicit comparison of microbial communities based on their composition. Beta diversity metrics therefore assess the differences between microbial communities. To compare microbial communities between every pair of community samples, a square matrix of distance was calculated, reflecting the dissimilarity between certain samples. The data in this distance matrix can be visualized with analyses such as Boxplot Analysis, Principal Coordinate Analysis (PCoA), hierarchical clustering, and so on.

Beta diversity analysis mainly uses four algorithms, binary jaccard, bray curtis, weighted unifrac (limited to bacteria), and unweighted unifrac (limited to bacteria), to calculate the distance between samples to obtain the β value between samples. These four algorithms can be divided into two categories: weighted (Bray-Curtis and Weighted Unifrac) and unweighted (Jaccard and Unweighted Unifrac) [14]. The use of unweighted methods is mainly to compare the presence or absence of species. A smaller β diversity between two groups indicates greater similarity in their relative species composition. Weighted methods consider both qualitative data (the presence or absence of species) and quantitative data about the relative abundance of species.

The metrics can be phylogeny based (the UniFrac metrics) or not (Bray-Curtis and Jaccard). The UniFrac distance take the phylogenetic relatedness of ASVs into account (only for bacteria), while the Bray-Curtis distance considers only the abundance.

Suggestion: In the microbial diversity analysis, the differences in microbial composition between different environments are tremendous, so the unweighted method is usually used for the analysis. However, if we want to study the relationship between the control and experimental treatment group using unweighted analysis, then no significant difference can be observed, and weighted method is recommended. Neither analytical method is inherently “better” or “worse”, but the appropriate method should be chosen for particular research purposes. Four types of Beta diversity analysis using a variety of algorithms have been included to provide you with a comprehensive analysis of the results, and you can choose the most suitable one to explain the biological issues of your project.

2.7.1 Boxplot Analysis

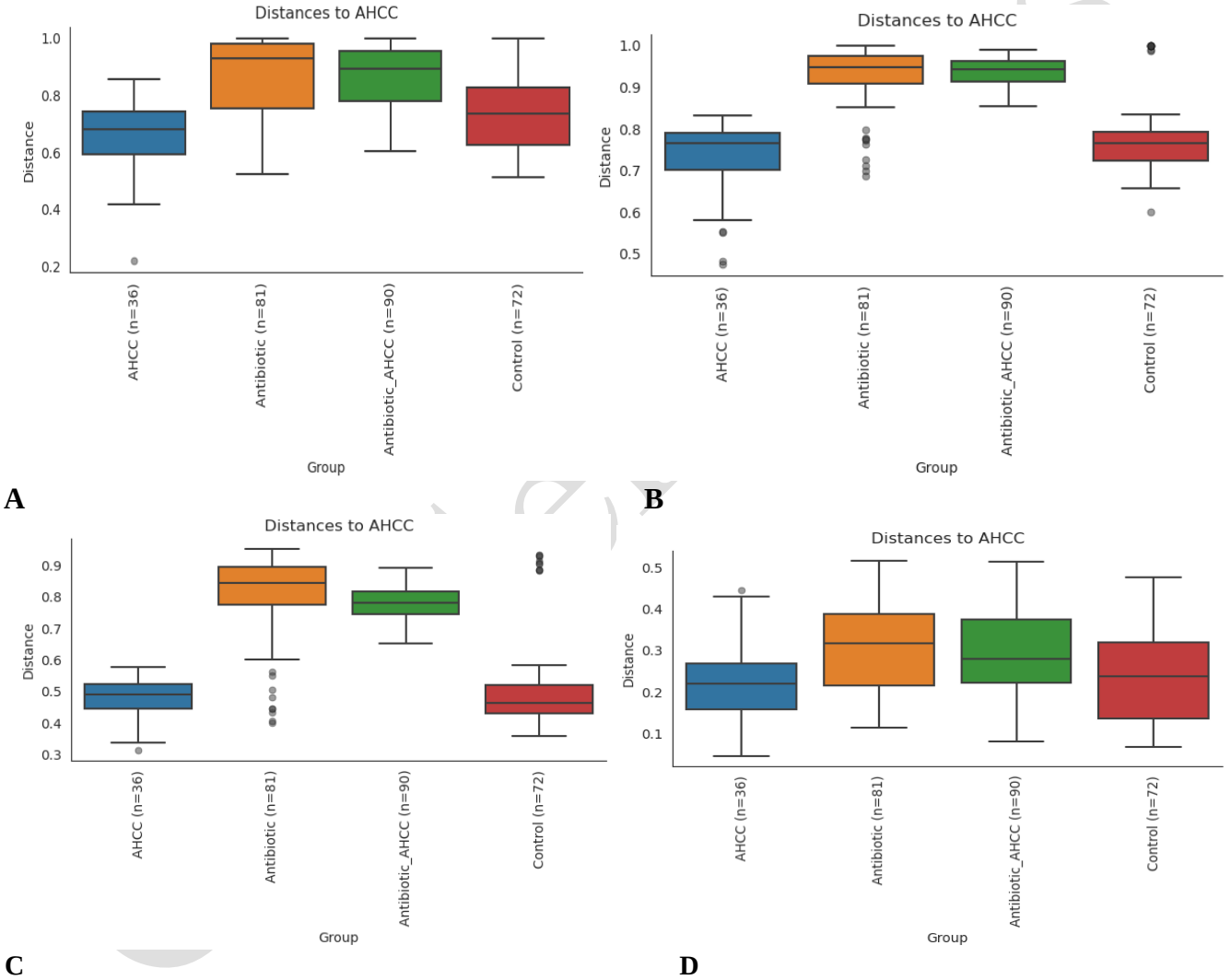
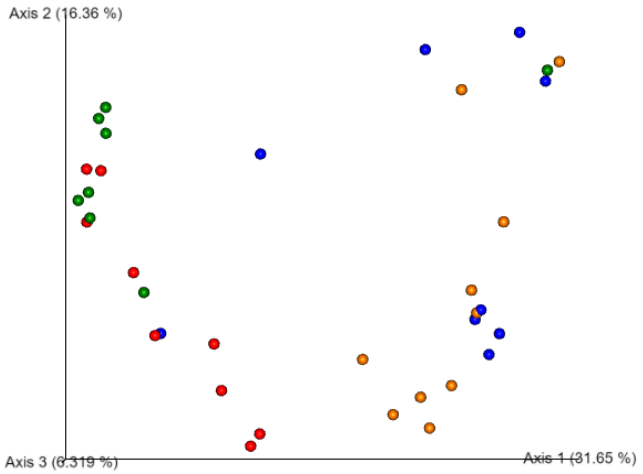


Figure 11. Boxplot analysis based on bray Curtis (A), binary jaccard (B), unweighted unifrac (C), and weighted unifrac (D). The boxplots represent the distribution of predicted functional profiles in the analyzed samples, with the box indicating the interquartile range and the median line inside. The whiskers extend to the minimum and maximum values within a specified range, providing insight into the variability and differences in functional potentials among the compared sample groups.

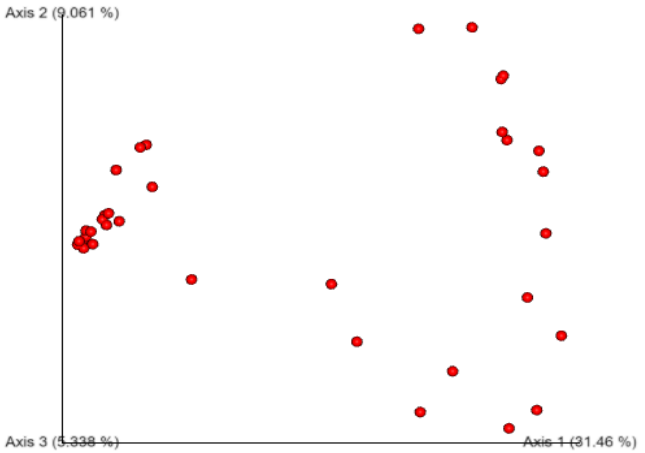
2.7.2 PCoA Analysis

Principal coordinates analysis (PCoA) [15] is an ordination technique similar to PCA, which picks up the main elements and structure from reduced multi-dimensional database series of eigenvalues and eigenvectors. It starts with a similarity matrix or dissimilarity matrix (distance matrix) and assigns for each item a location in a low-dimensional space. The technique has advantages over PCA in that each ecological distance can be investigated. PCA finds out the main coordinates based on the similarity coefficient matrix of all samples, while PCoA is based on the distance matrix. Weighted Unifrac and Unweighted Unifrac were calculated to assist the PCoA analysis. By using PCoA we can visualize individual and/or group differences, illustrating the microbial diversity between samples. Based on the four algorithms, principal coordinates analysis was calculated and displayed by QIIME 2 tool, you can view QIIME 2(QIIME 2 View) artifacts and visualizations at view.qiime2.org by uploading files. PcoA results is located at 04.Diversity\Beta\PcoA\ bray_curtis\index.html, and the PcoA result plots can be adjusted according to the link below.

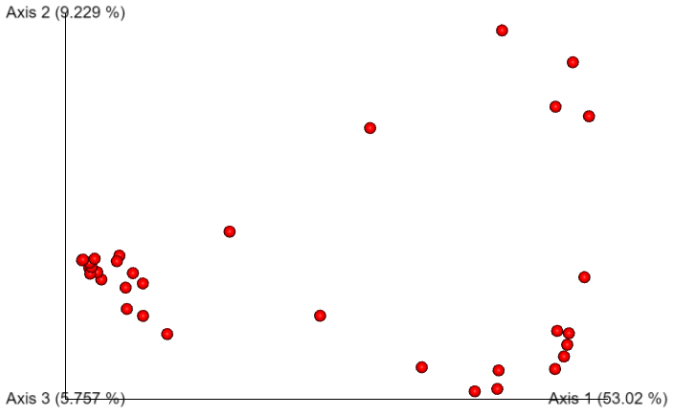
Result link: 07.Diversity\Beta\PCoA\bray_curtis\index.html


Axis 2 (16.36 %)
Axis 3 (6.319 %)
Axis 1 (31.65 %)

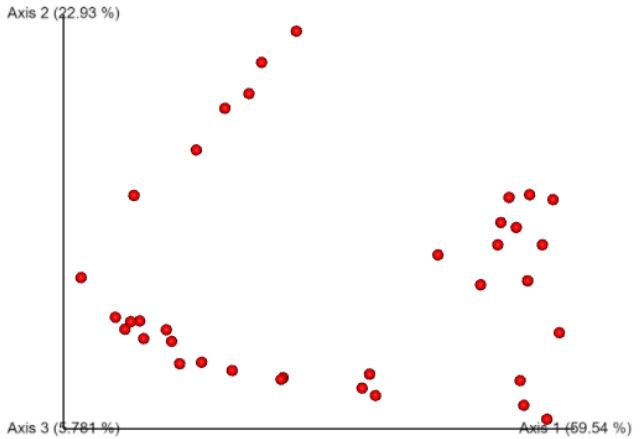
A


Axis 2 (9.061 %)
Axis 3 (5.338 %)
Axis 1 (31.46 %)

B


Axis 2 (9.229 %)
Axis 3 (5.757 %)
Axis 1 (53.02 %)

C


Axis 2 (22.93 %)
Axis 3 (5.781 %)
Axis 1 (59.54 %)

D

Figure 12. PCoA analysis based on bray Curtis (A), binary jaccard (B), unweighted unifrac (C), and weighted unifrac (D). Each point represents a sample, plotted by a principal component on the X- axis and another principal component on the Y- axis, which was colored by group. The percentage on each axis indicates the contribution value to discrepancy among samples.

2.7.3 UPGMA Analysis

Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is a type of hierarchical clustering method using average linkage. It is widely used in ecology for the classification of samples based on their pairwise similarities in relevant descriptor variables. The basic two ideas of UPGMA are as follows: First, it gathers two samples of the minimum distance together and forms a new node (a new sample), which is branched at the halfway point of the distance between the two samples. Second, it calculates the average distance between a new "sample" and the other samples and can find the minimum distance between two samples in order to cluster both. When all samples are gathered together, a complete clustering tree can be presented.

Based on the four algorithms, hierarchical clustering for samples using UPGMA was performed with the R language tool to assess the similarity of microbial composition between samples. The clustering results are displayed in Figure 12. A closer sample distance and a shorter branch, indicates more similarity in microbial composition between the samples.

The result is located at 04.Diversity\Beta\UPGMA\unweighted\tree.html and then click 'web-based ETE3 tree viewer', then click 'View tree!'. In this way, you can see the tree diagram of this result. The specific link is as follows.

Result link: 07.Diversity\Beta\UPGMA\unweighted\tree.html

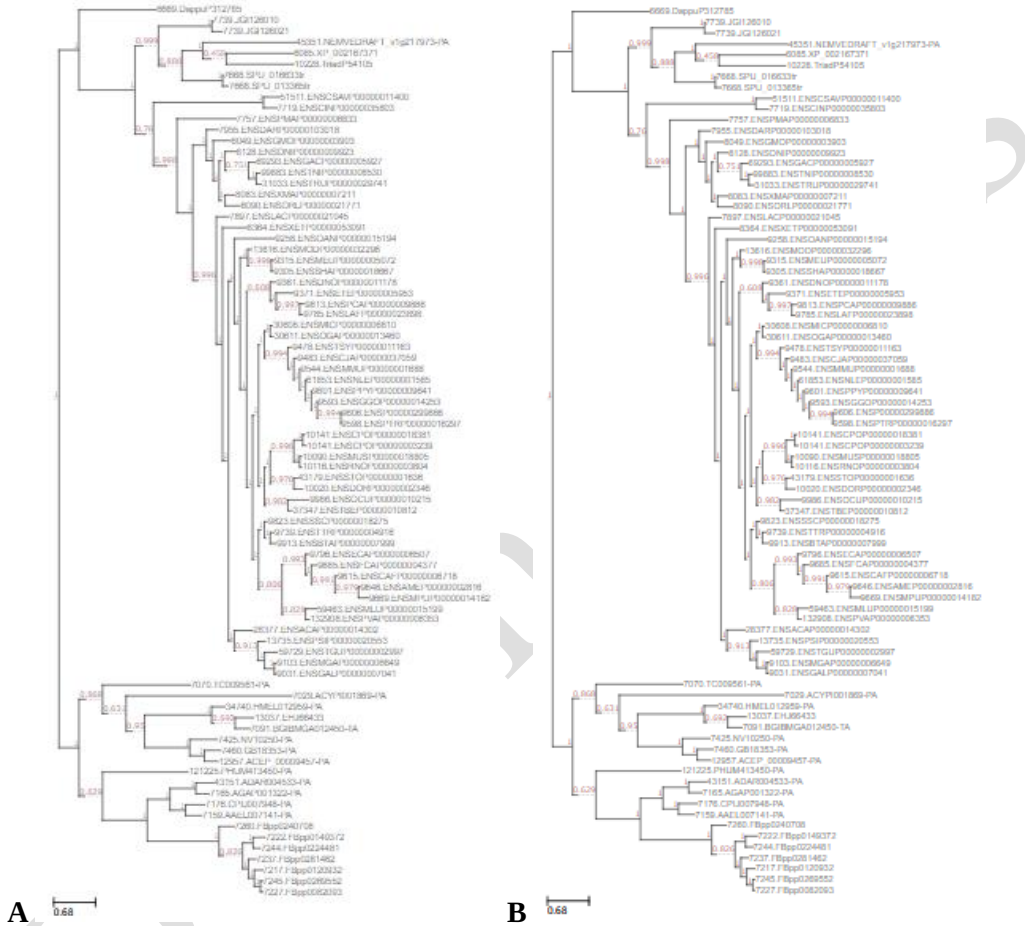


Figure 13. UPGMA clustering tree based on unweighted unifrac (A), and weighted unifrac (B). The different colors represent different grouping.

3. Analysis soft/database information:

Soft / Database	Source
Metaphlan	https://huttenhower.sph.harvard.edu/metaphlan/
KEGG	http://www.genome.jp/kegg/
NR	ftp://ftp.ncbi.nih.gov/blast/db/
CARD	https://card.mcmaster.ca/
CAZy	http://www.cazy.org/
PHI	http://www.phi-base.org/
VADB	http://www.mgc.ac.cn/VFs/
TCDB	https://www.tcdb.org/

4. Reference:

[1] Li D, Liu CM, Luo R, Sadakane K, Lam TW. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31(10):1674-6.

[2] Manghi P, Blanco-Míguez A, et al. (2023). MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. Cell Rep.42(5):112464.

[3] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–D205 (2014).

[4] Powell S, Forslund K, Szklarczyk D, et al. eggNOG v4. 0: nested orthology inference across 3686 organisms[J]. Nucleic acids research, 2013: gkt1253.

[5] Réau M, Lagarde N, Zagury JF, Montes M. (2019). Nuclear Receptors Database Including Negative Data (NR-DBIND): A Database Dedicated to Nuclear Receptors Binding Data Including Negative Data and Pharmacological Profile. J Med Chem. 62(6):2894-2904.

[7] Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. (2005). VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 33(Database issue):D325-8.

[8] Saier MH, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, Lam KJK, Tian N, Russum S, Wang J, Medrano-Soto A. (2021). The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 49(D1):D461-D467.

[9] Bertozzi Silva J, Storms Z, Sauvageau D. (2016). Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 363(4):fnw002.

[10] Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) .The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37:D233-238.

[11] Jia B, Raphenya A R, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database[J]. Nucleic Acids Research, 2017, 45(D1):D566.

[12] Grice EA, Kong HH, et al. (2009). Topographical and temporal diversity of the human skin microbiome. Science, 324(5931): 1190–1192.

[13] Wang Y, Sheng H-F, He Y, Wu J-Y, Jiang Y-X, Tam NF-Y, Zhou H-W: Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. Applied and environmental microbiology 2012, 78(23):8264-8271.

[14] Lozupone C, Knight R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol, 71 (12): 8228-8235.

[15] Sakaki T, Takeshima T, Tominaga M, Hashimoto H, Kawaguchi S: Recurrence of ICA-PCoA aneurysms after neck clipping. Journal of neurosurgery 1994, 80(1):58-63.