CD Genomics
The Genomics Services Company

Metatranscriptome Sequencing Report

Order#: XXXXXXXX

Date: XX/XX/XXXX

CD Genomics
The Genomics Services Company

Table of Contents

1. Experimental procedures of Metatranscriptome 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:

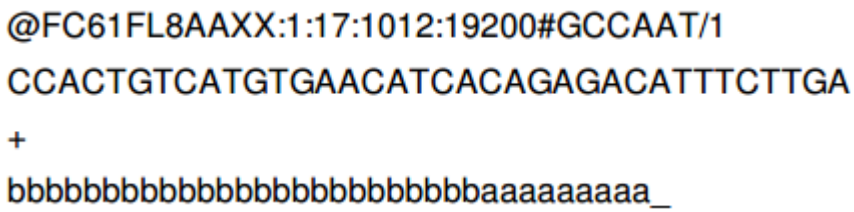


Figure 1. Schematic of FASTQ format file

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.(top10)

Sample	Raw Reads	Bases	Avg.Quality	GC (%)	Q20	Q30
H10	76267462	10715238194	145	43.40	98.36	97.10
H11	77236784	10877877466	146	43.50	98.39	97.14
H12	58219520	8210381988	146	43.80	98.40	97.15
H13	46975944	6525512488	144	44.18	98.28	97.01
H14	43700428	6058241762	144	44.22	98.25	96.95
H15	39808816	5570775712	145	43.60	98.29	97.01
H16	53709466	7435846010	144	44.35	98.24	96.96
H17	51129790	6970425524	142	44.43	98.19	96.87
H1	76088504	10706887154	146	42.82	98.34	97.07
H18	69511946	9537592646	143	43.65	98.16	96.80

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	Sequence Number	Largest contig	Total length	GC (%)	N50	N75
nucleotide	459514	19817	5.43E+08	39.77	1392	773

2.2.1 Metaphlan Species Annotation

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

Table 3. Statistics of the species annotation results(TOP10)

Sample	Files
H10	03.Taxonomy\metaphlan\H10\metaphlan\metaphlan_anno.txt
H11	03.Taxonomy\metaphlan\H11\metaphlan\metaphlan_anno.txt
H12	03.Taxonomy\metaphlan\H12\metaphlan\metaphlan_anno.txt
H13	03.Taxonomy\metaphlan\H13\metaphlan\metaphlan_anno.txt
H14	03.Taxonomy\metaphlan\H14\metaphlan\metaphlan_anno.txt
H15	03.Taxonomy\metaphlan\H15\metaphlan\metaphlan_anno.txt
H16	03.Taxonomy\metaphlan\H16\metaphlan\metaphlan_anno.txt
H17	03.Taxonomy\metaphlan\H17\metaphlan\metaphlan_anno.txt
H1	03.Taxonomy\metaphlan\H1\metaphlan\metaphlan_anno.txt
H18	03.Taxonomy\metaphlan\H18\metaphlan\metaphlan_anno.txt
Taxonomy Annotation	03.Taxonomy\metaphlan

2.2.2 Qiime2 Species Annotation

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.

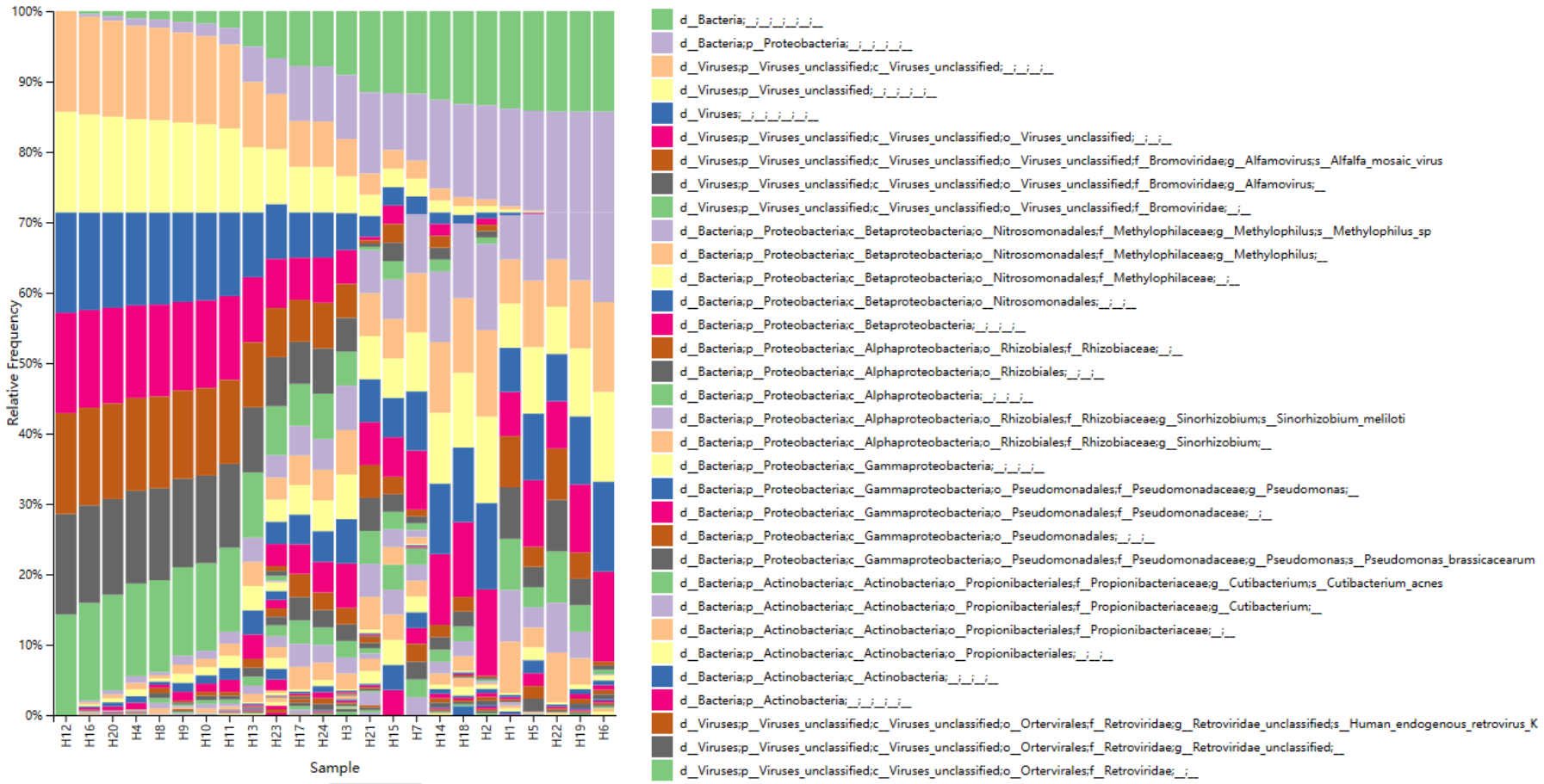


Figure 2. The taxonomy distribution of all sample in Phylum classification level. Other classification levels can be found in the taxonomy folder.

Table 4. Statistics of the Taxonomy results

Level	Sample Abundance
Kingdom	\result\03.Taxonomy\level-1.csv
Phylum	\result\03.Taxonomy\level-2.csv
Class	\result\03.Taxonomy\level-3.csv
Order	\result\03.Taxonomy\level-4.csv
Family	\result\03.Taxonomy\level-5.csv
Genus	\result\03.Taxonomy\level-6.csv
Species	\result\03.Taxonomy\level-7.csv

2.2.3 Species Abundance Heatmap

A Heatmap is a graphical representation of clustering using color gradients to represent the relative abundance of similar species in a sample. According to the taxonomic composition and relative abundance of each sample, heatmap analyses were carried out at each taxonomic level (phylum, class, order, family, genus and species respectively) and plotted using R language tools. In the heatmap clustering results, color represents the abundance of species, and vertical clustering indicates the similarity of the abundance between different species. A shorter distance between the two species and a shorter branch length indicates that the two species have a more similar abundance between the samples. The horizon clustering indicates the similarity of the abundance of different species between samples. As with the vertical clustering, the shorter distance and branch length between the samples indicates the more similarity of abundance. The heatmap at phylum level is illustrated in Figure 3.

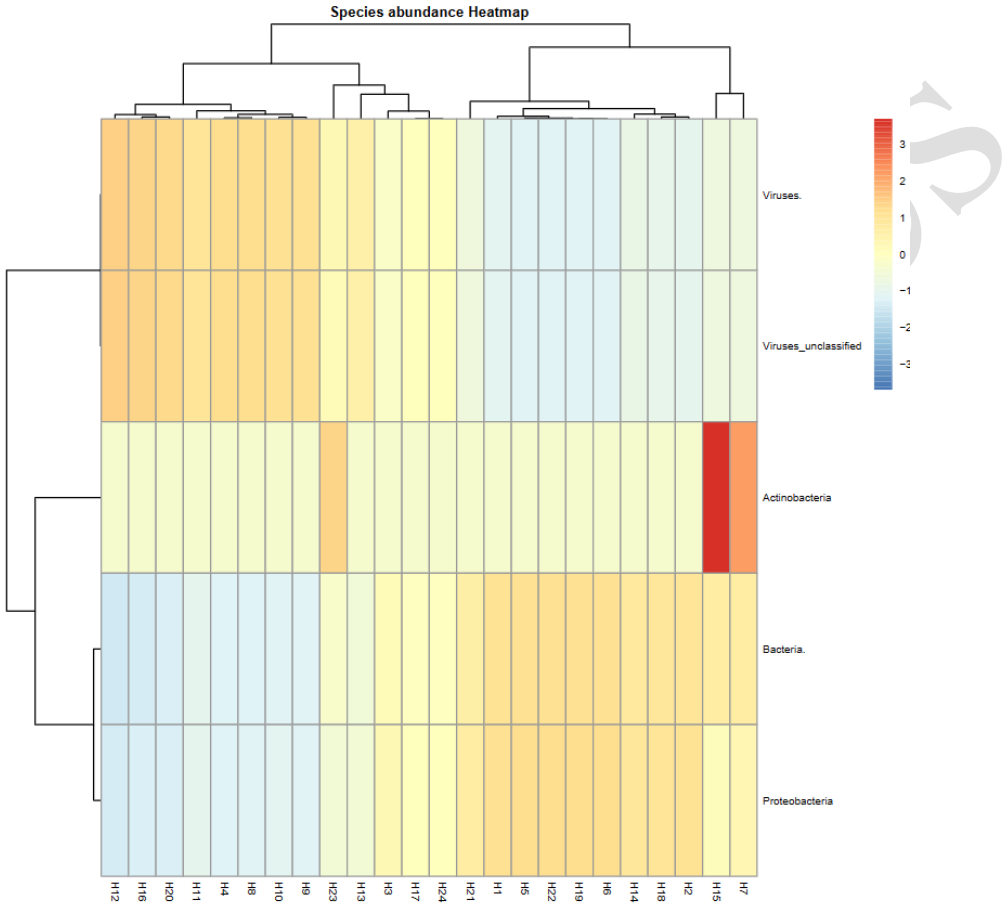


Figure 3. Species abundance Heatmap. Phylum. Plotted by sample name on the X-axis. The Y-axis represents the genus. The absolute value of the legend represents the distance between the raw score and the mean population of the standard deviation. The legend is negative when the raw score is below the mean.

2.3 Statistical Data of Alpha Diversity

2.3.1 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 5.

Table 5. Statistics of Alpha diversity indices

Sample ID	Observed species	Chao1	Simpson	Shannon
H1	25	24	0.912890465	3.780011627
H10	35	19	0.888399713	3.463859006
H11	31	19	0.893533523	3.593366156
H12	29	7	0.857142857	2.807616584
H13	31	34	0.926414128	4.055513167
H14	31	52	0.912123107	3.948416059
H15	25	25	0.948151256	4.47751725
H16	33	17.33333333	0.860846072	3.024437931
H17	35	38	0.945322108	4.381333703
H18	28	29.66666667	0.903555363	3.765482162

2.3.2 Rarefaction Curve

Rarefaction curve [2] 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.

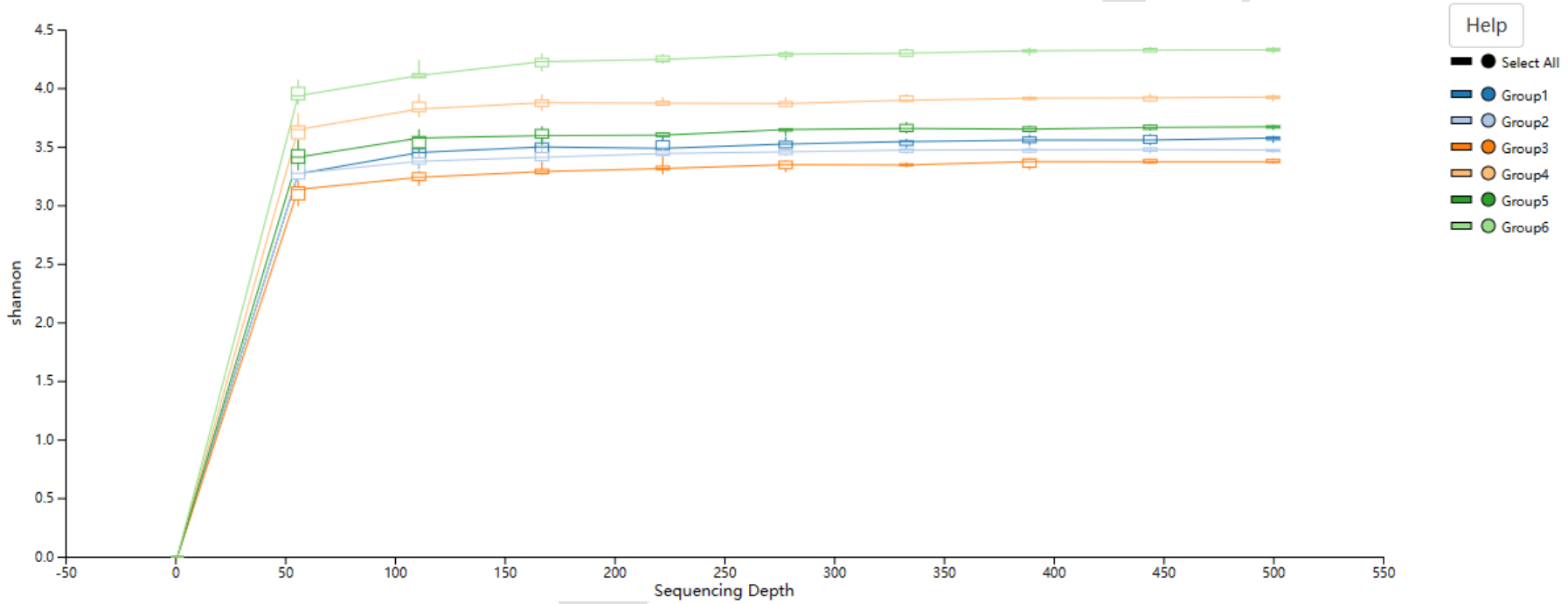


Figure 4. Rarefaction curve of the sequenced reads for all samples.

2.4 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 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). 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 OUTs 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.

2.4.1 Boxplot Analysis

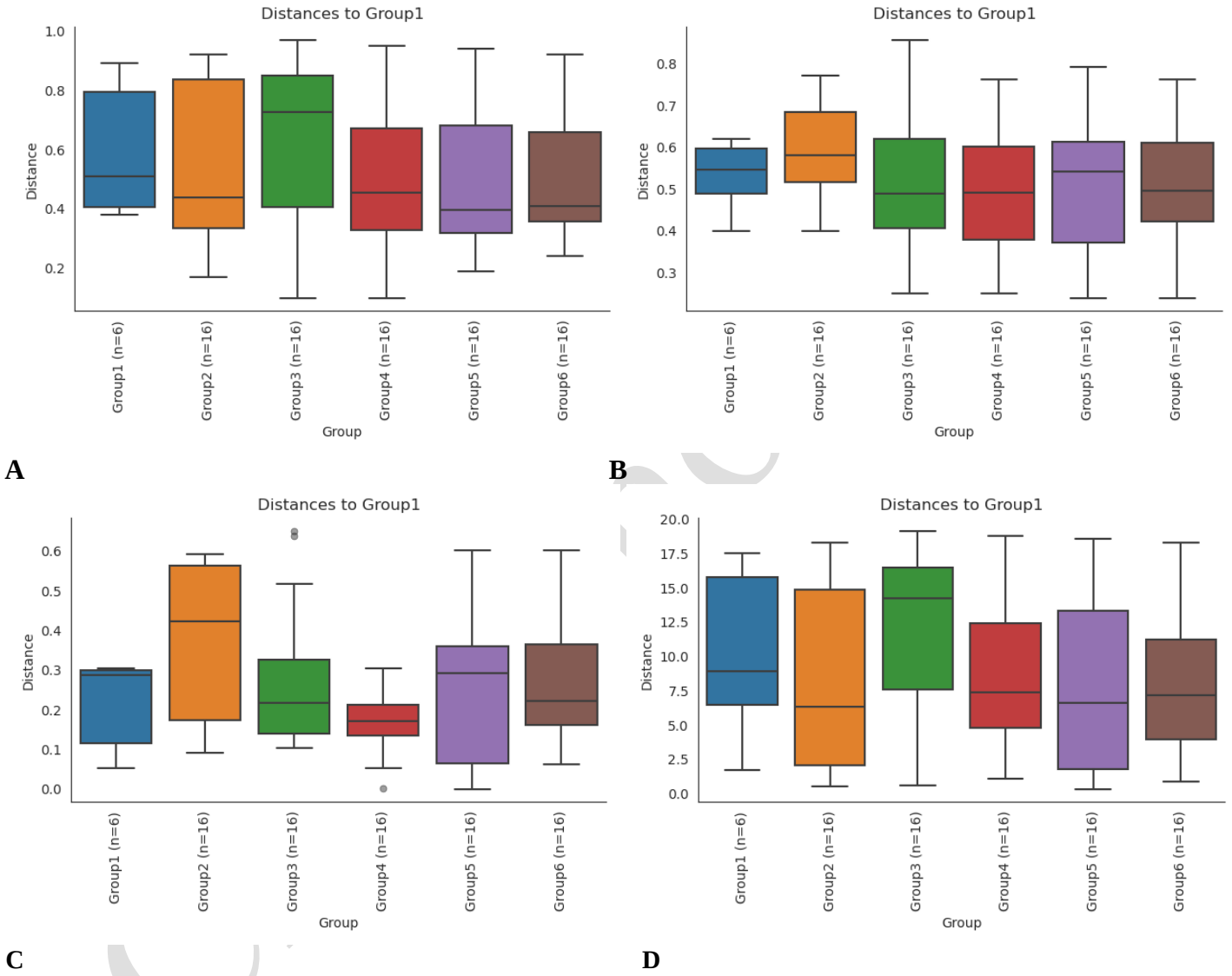
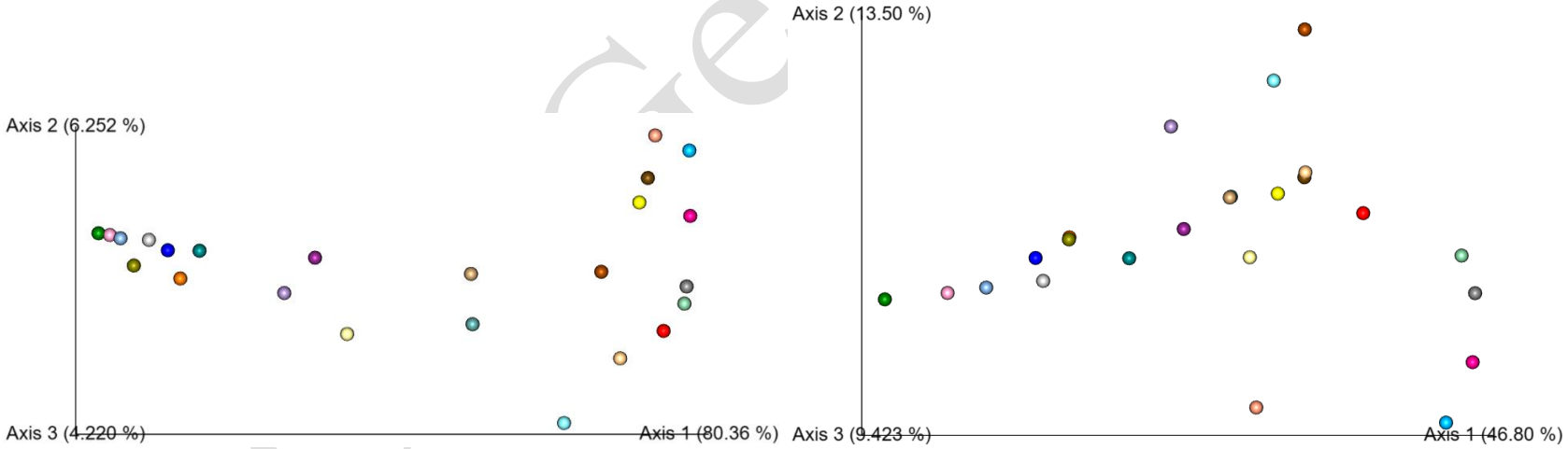


Figure 5. Boxplot analysis based on bray Curtis (A), binary jaccard (B), unweighted unifrac (C), and weighted unifrac (D). The boxplots represent the distribution of predicted functional profiles in the analyzed samples, with the box indicating the interquartile range and the median

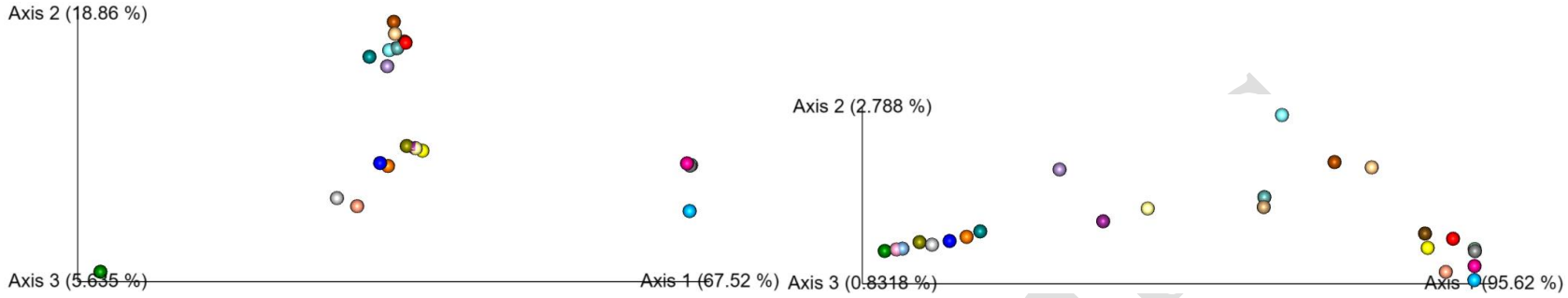
line inside. The whiskers extend to the minimum and maximum values within a specified range, providing insight into the variability and differences in functional potentials among the compared sample groups.

2.4.2 PCoA Analysis

Principal coordinates analysis (PCoA) [3] 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. Base 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.



A **B**



C **D**

Figure 6. 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.

2.4.3 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 7. A closer sample distance and a shorter branch, indicates more similarity in microbial composition between the samples.

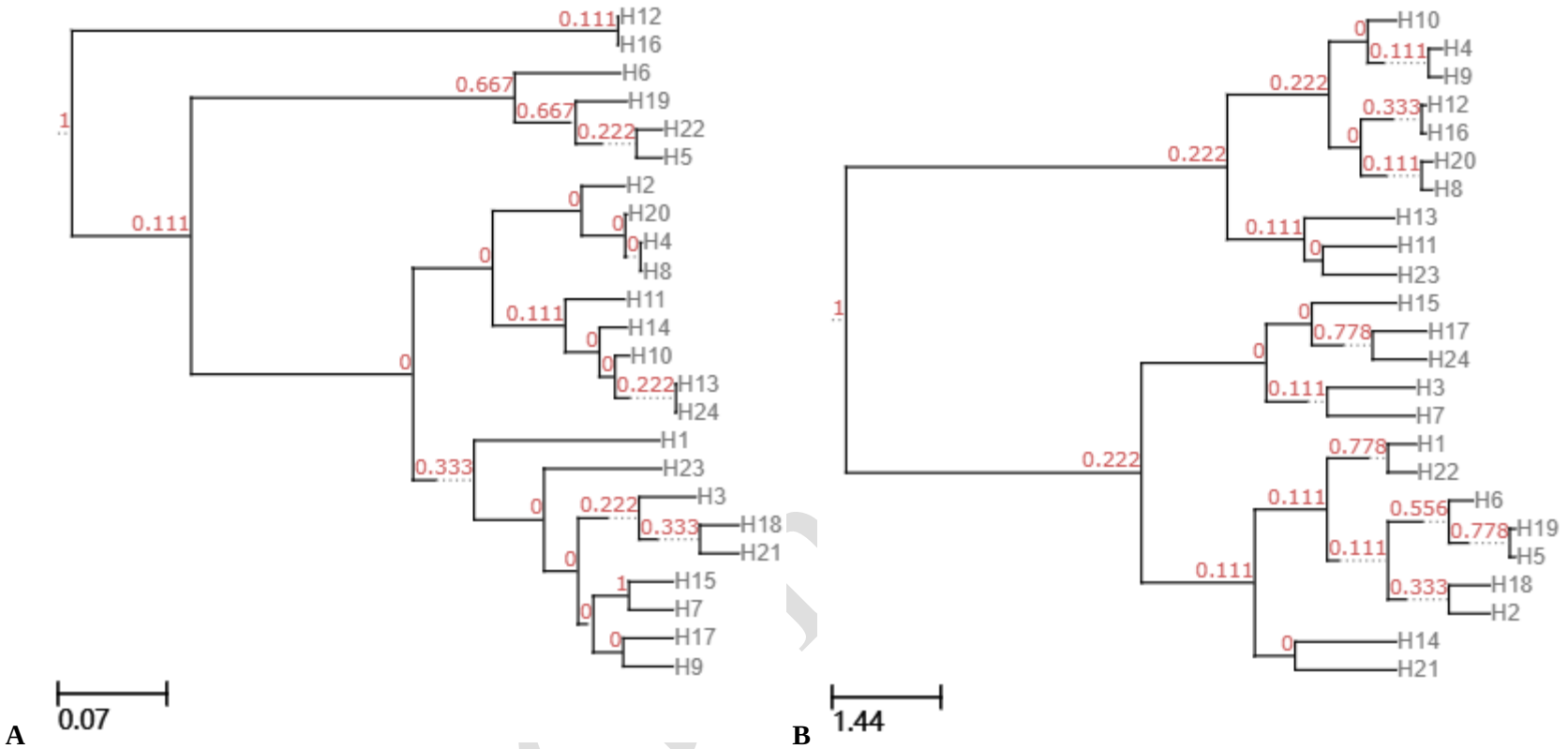


Figure 7. UPGMA clustering tree based on unweighted unifrac (A), and weighted unifrac (B). The different colors represent different grouping.

2.5 Gene Expression Analysis

2.5.1 Gene Expression Distribution

Using transcriptome data to detect gene expression has high sensitivity. Typically, TPM values for protein-coding gene expression that can be sequenced span six orders of magnitude from 10^{-2} to 10^{4} [4]. Boxplot and density plot of the TPMs of all transcripts are used to compare the expression of different samples.

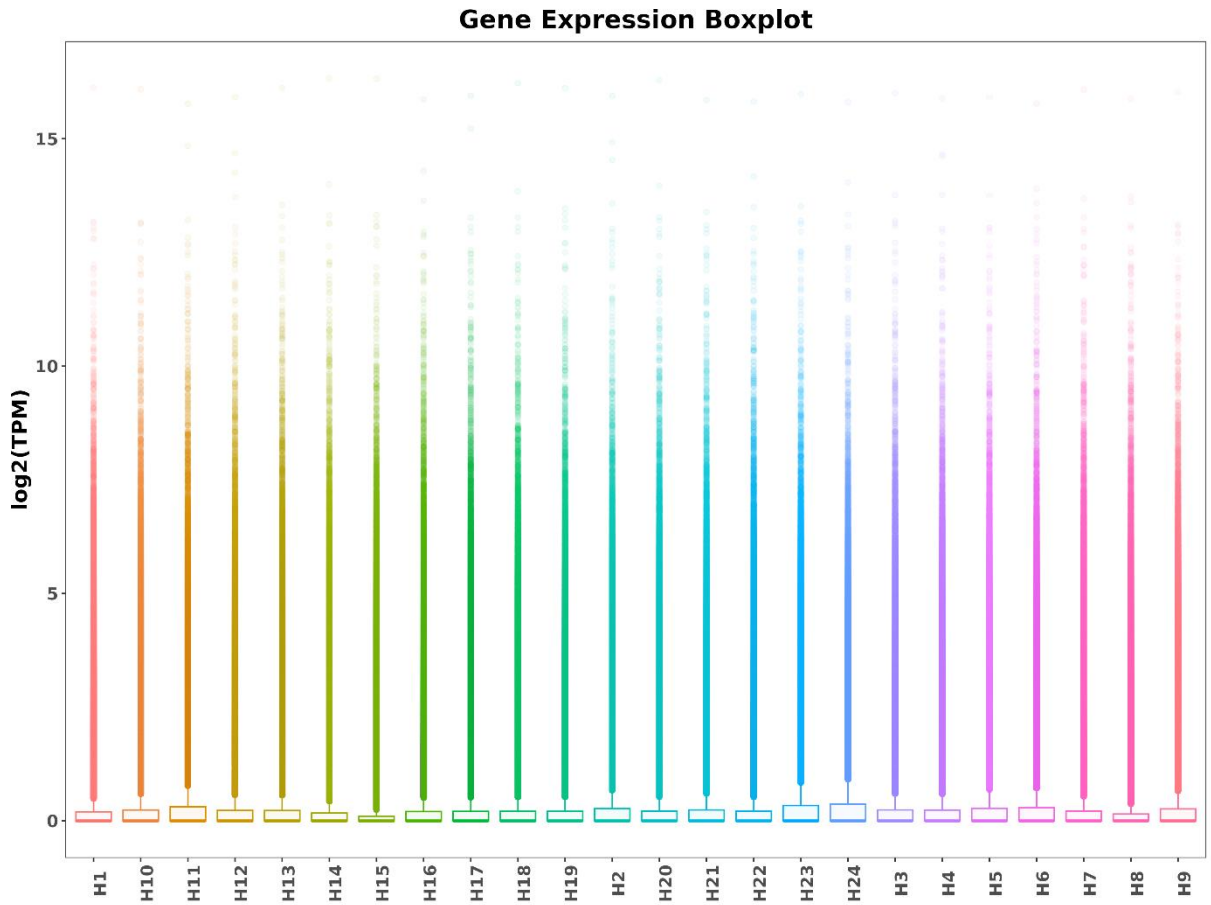


Figure 8. Boxplot of TPM for each sample. The x-axis shows the sample names and the y-axis shows the log10(TPM). Each box has five statistical magnitudes (max value, upper quartile, median, lower quartile and min value).

2.5.2 Correlation Bewteen samples

Biological repetition is necessary for any biological experiment, and high throughput sequencing technology is no exception. Gene abundance correlation between samples is an important index to test the reliability of experiments and the rationality of sample selection. The closer the correlation coefficient is to 1, the higher the similarity of gene abundance patterns between samples is.

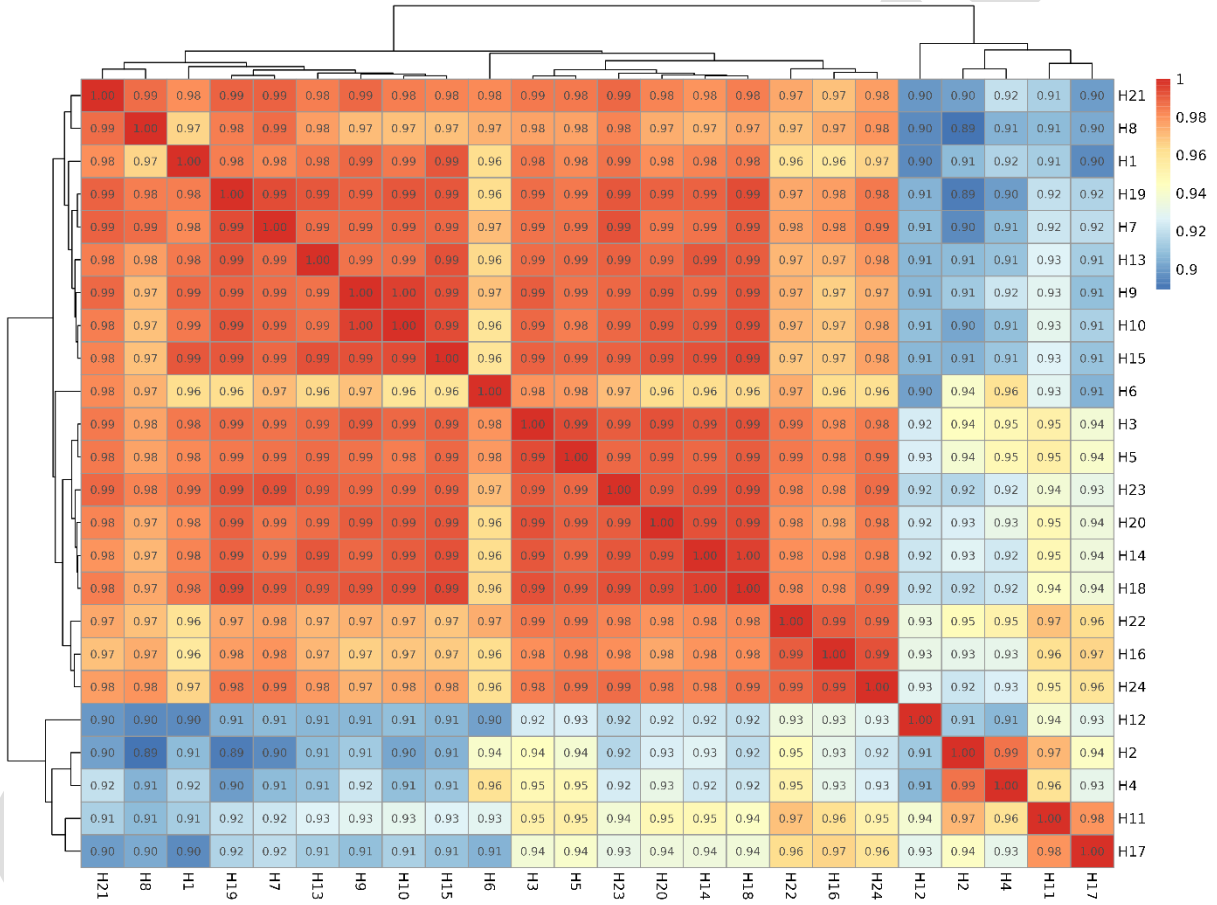


Figure 9. Correlation graph of gene number.

2.5.3 Gene Expression Difference Analysis

Read count value obtained from the gene expression analysis is used as the input data to do differential expression analysis.

For samples without biological replicates, TMM is first used to normalize the read count value, and DEGseq is used to do the analysis. The threshold is normally set as: [log2(Fold Change)] > 1 and q-value < 0.05.

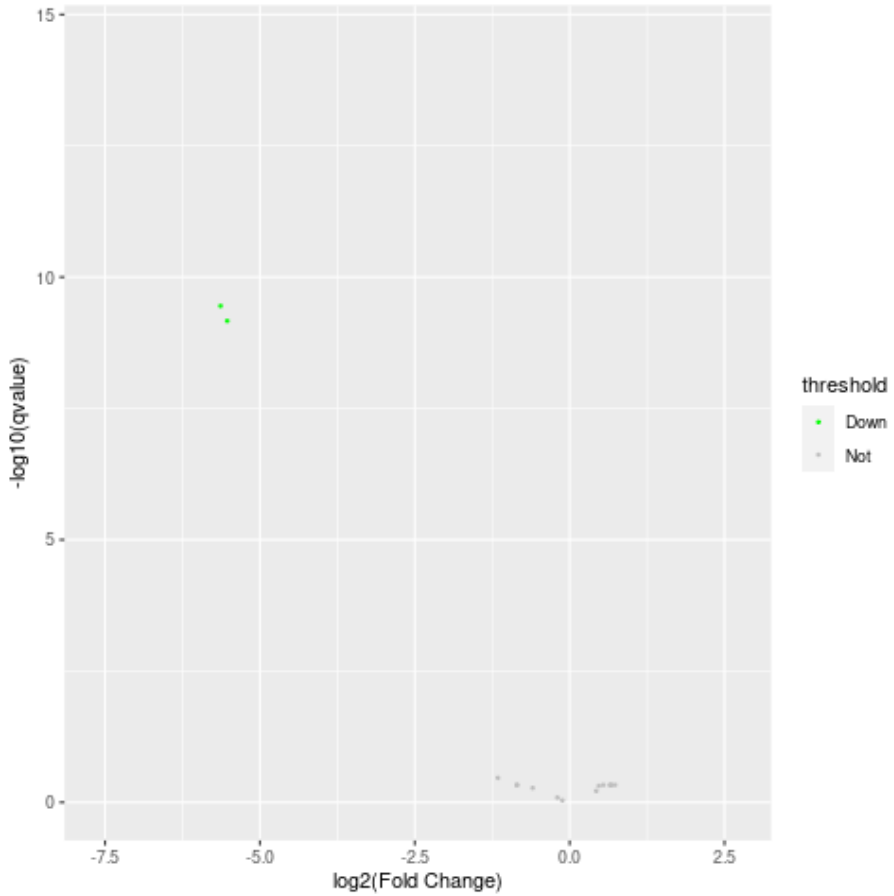


Figure 10. Volcano Plot.

2.5.4 Statistics of Differentially Expressed Genes

Table 6. Number of differentially expressed genes for each pairwise comparison.

DEG Set	All DEG	Up-regulated	Down-regulated
CLC_vs_SLC	2	0	2

2.6 Functional Annotation for Differentially Expressed Genes

The functional annotation of the database of differentially expressed genes, and the number of genes annotated by each differentially expressed gene are shown in the following table:

Table 7. Annotation of differentially expressed genes

DEG Set	Total	GO	KEGG	Pfam	Swiss-Prot	EggNOG	COG
CLC_vs_SLC	217	60	149	205	177	210	210

2.6.1 Classification of GO for DEGs

Gene ontology (GO - Gene Ontology Consortium, 2000) enrichment analysis is a set of the internationally standardized classification system of gene function description that attempts to identify GO terms that are significantly associated with differentially expressed protein coding genes. GO molecules are divided into three main categories: 1) Cellular Component: used to describe the subcellular structure, location and macromolecular complexes, such as nucleoli, telomere and recognition of the initial complex; 2) Molecular Function: used to describe the gene, gene products, individual functions, such as carbohydrate binding or ATP hydrolase activity; 3) Biological Process: used to describe the products encoded by genes involved in biological processes, such as mitosis or purine metabolism. The statistics for GO classification of DEGs is shown in the following figure:

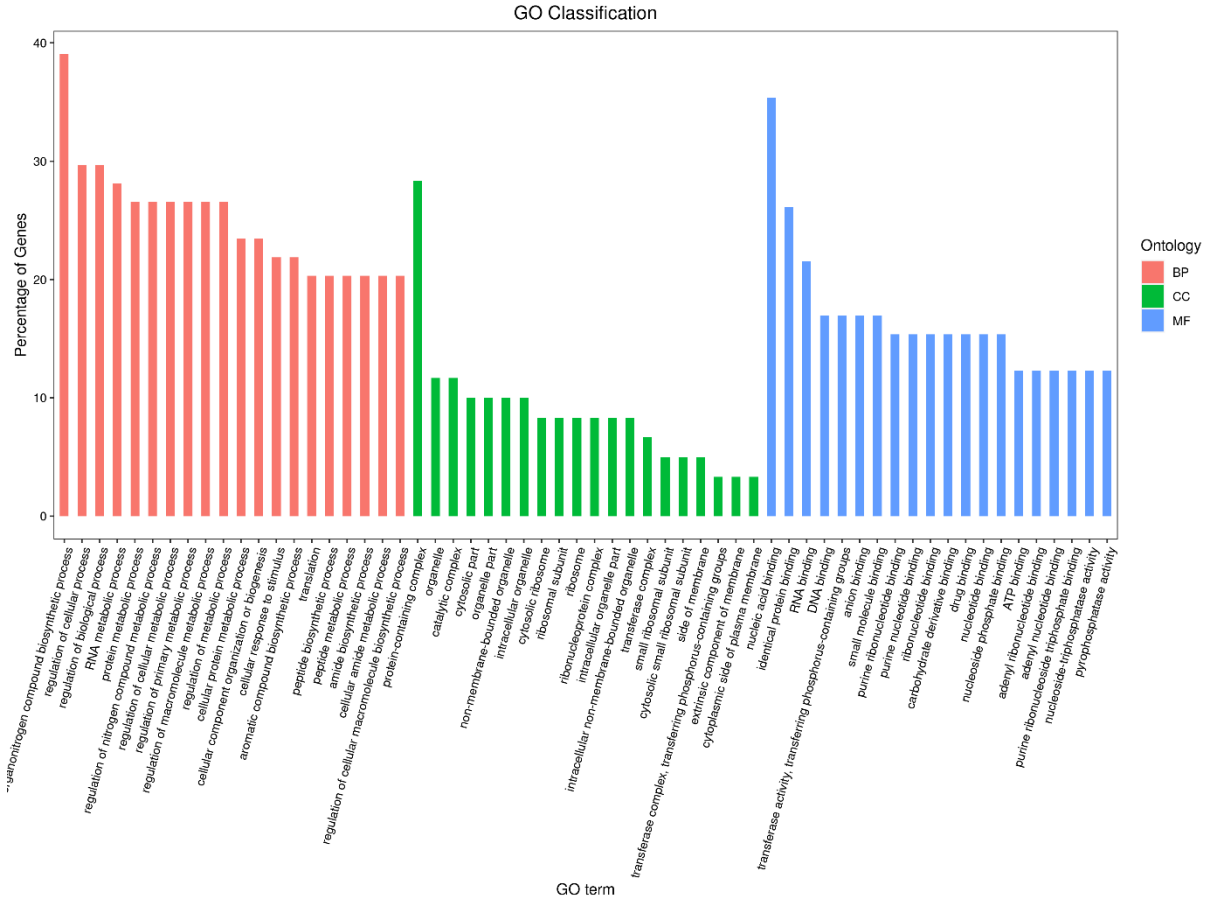


Figure 11. Statistics results of GO annotation for CLC_vs_SLC.

2.6.2 KEGG Analysis

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 [5]. (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 (CLC_vs_SLC) is demonstrated in below figure:

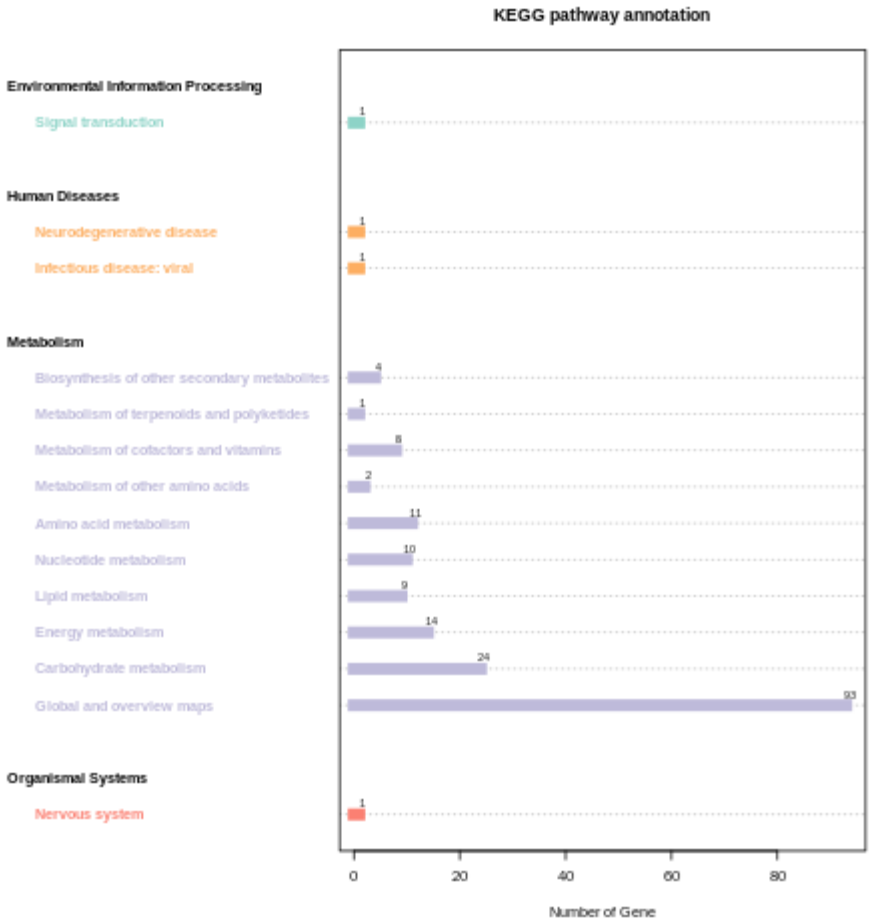


Figure 12. CLC_vs_SLC KEGG_classification

2.7 Functional Database Annotation

General databases for function annotations currently available include:

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

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

Non-Redundant Protein Sequence Database (NR);

UniProt Knowledgebase(UniProt);

Virulence Factors Database (VFDB);

Transporter Classification Database (TCDB);

Pathogen Host Interactions Database (PHI);

Carbohydrate-Active enZYmes Database (CAZy);

The Comprehensive Antibiotic Resistance Database (CARD);

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.

UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB).

It is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. Since 2002, it is maintained by the UniProt consortium and is accessible via the UniProt website.

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.

2.7.1 Statistics of The Annotated Gene Numbers

Table 8. Statistics of the annotated gene numbers（CLCW0_2）

DataBase	Annotated Number	Unannotated Number
CARD	28462	6784
CAZy	992	34254
PHI	7019	28227
TCDB	5603	29643
VFDB	2812	32434
Total	30261	4985

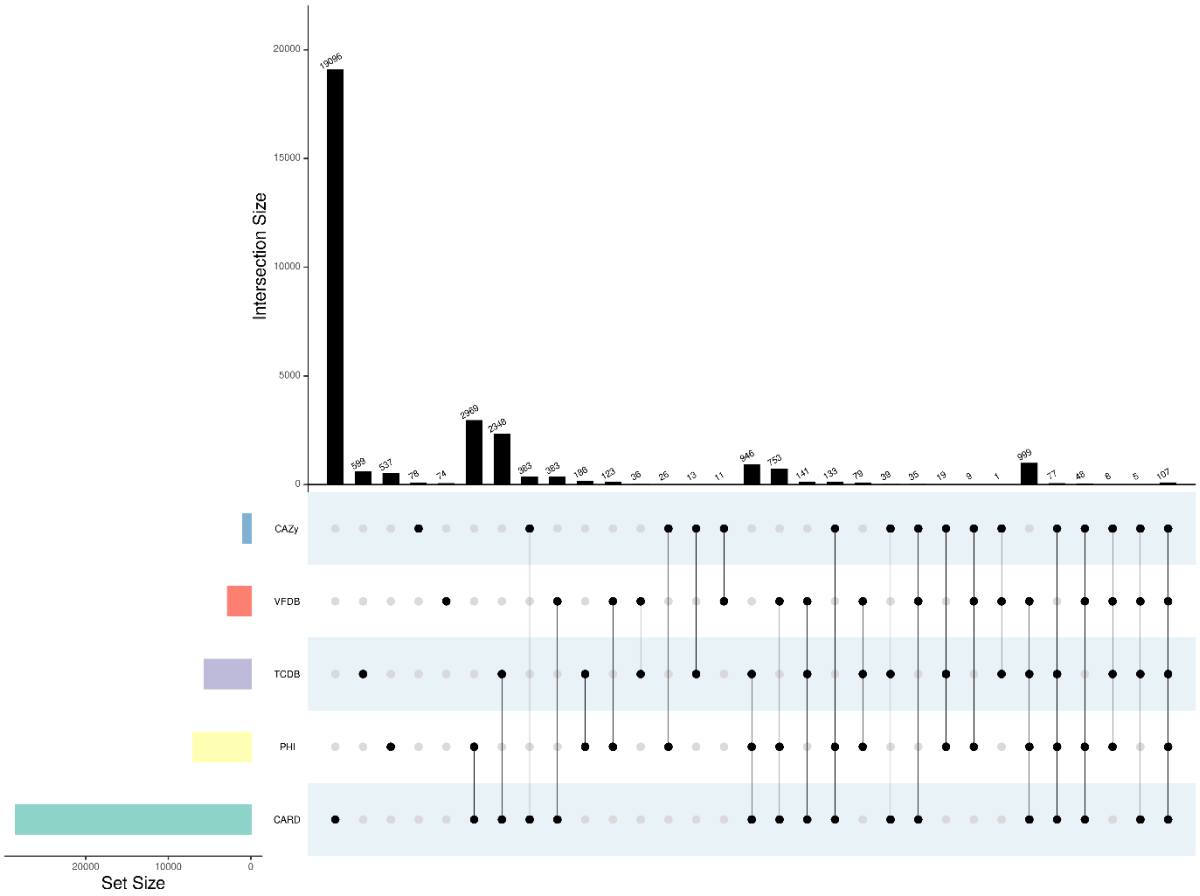


Figure 13. Statistical of specific function database common and unique annotation in CLCW0_2.

2.7.2 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.

2.7.3 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).

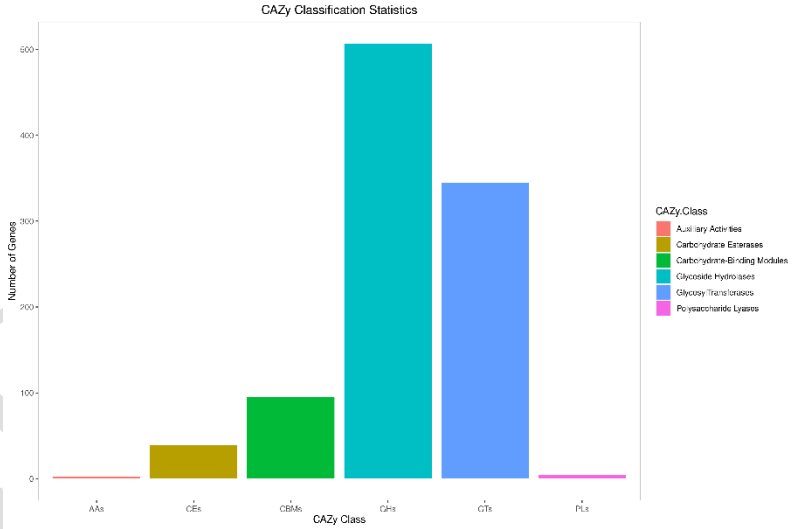


Figure 14. CAZy function classification for CLCW0_2

2.7.4 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.

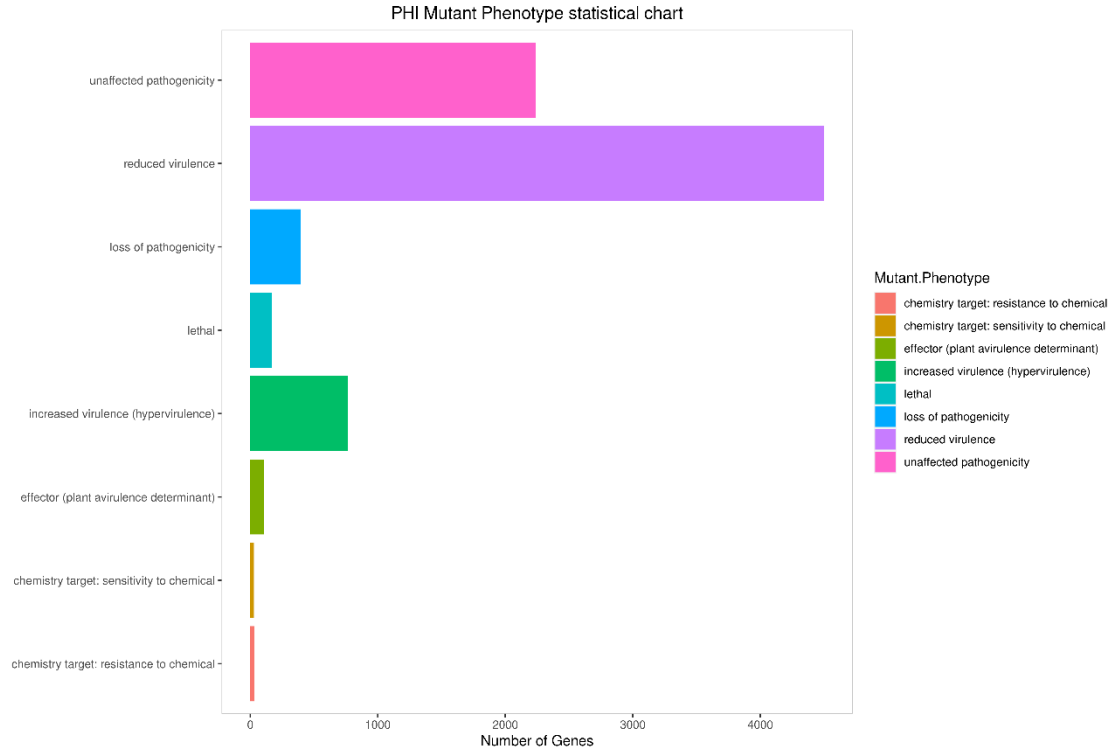


Figure 15. PHI phenotype classification for CLCW0_2.

2.7.5 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.

2.7.6 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.

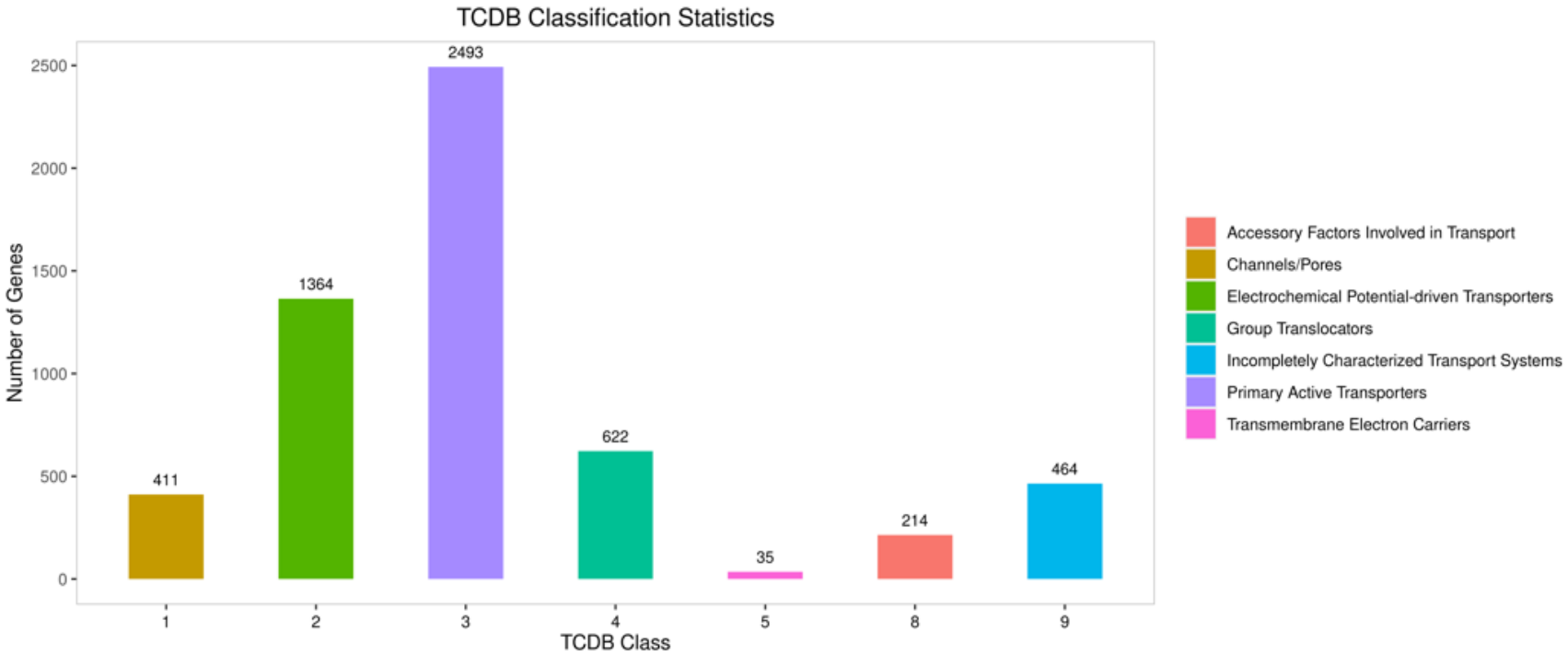


Figure 16. TCDB classification statistics for CLCW0_2. The horizontal axis of the above figure (bar graph) is the transporter class, and the vertical axis is the number of genes annotated to the corresponding class. The legend corresponds to the definition of the transporter class; the below graphs corresponds 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.

3. Soft List

Software	URL
megaHit	https://www.metagenomics.wiki/tools/assembly/megahit
Metaphlan	https://huttenhower.sph.harvard.edu/metaphlan/
Qiime2	https://qiime2.org/
Diamond	https://github.com/bbuchfink/diamond

4. Database List

Database	URL
KEGG	https://www.genome.jp/kegg/
eggNOG	http://eggnog5.embl.de/#/app/home
NR	https://arep.med.harvard.edu/seqanal/db.html
UniProt	https://www.uniprot.org/
VFDB	https://ngdc.cncb.ac.cn/databasecommons/database/id/516
TCDB	https://en.wikipedia.org/wiki/Transporter_Classification_Database
PHI	http://www.phi-base.org/
CAZy	http://www.cazy.org/
CARD	https://card.mcmaster.ca/

5. Reference:

[1] Li D, Liu CM, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6.

[2] Nikkeshi A, Hiraiwa MK, Ushimaru A, et al. Evaluation of sampling effort required to assess pollen species richness on pollinators using rarefaction. Appl Plant Sci. 2021 Feb 27;9(2):e11411.

[3] Wang Y, Sun F, Lin W, Zhang S. AC-PCoA: Adjustment for confounding factors using principal coordinate analysis. PLoS Comput Biol. 2022 Jul 13;18(7):e1010184.

[4] Zhao Y, Li MC, Konaté MM, et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med. 2021 Jun 22;19(1):269.

[5] Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(D1):D353-D361.