1. PROJECT SUMMARY

Project ID: 90-1130539759
Customer: RapolasJamontas
BI: GENEWIZ-NGS-BI

2. DESCRIPTION OF WORKFLOW

2.1. Experiment Workflow

16S/18S/ITS rRNA is composed of conserved and hypervariable regions. Whereas conserved regions are not significantly different across various microbial strains, the sequences of hypervariable regions are genus or species-specific, and differ in accordance to phylogenetic difference. Therefore, 16S/18S/ITS rDNA serve as identifiers of biological species, and are important for microbial phylogeny and taxonomic identification. 16S/18S/ITS rDNA amplicon sequencing has become an important tool for the study of the composition of microbial communities in environment.

16S rDNA amplicon sequencing includes the library construction using specific primers to amplify the variable region of prokaryotic 16S rDNA and data analysis of the 16S rDNA variable region sequence to identify the composition and abundance of prokaryotic microorganisms in the environment. The proprietary workflow at Azenta effectively amplifies the three variable regions of 16S rDNA (V3, V4, V5) and accurately identifies various species including archaea. For samples with eukaryotic contamination, only V3, V4 region are amplified. 18S / ITS rDNA amplicon sequencing includes the library construction using specific primers to amplify the variable region of eukaryotic 18S / ITS rDNA and data analysis to identify the composition and abundance of eukaryotic microorganisms in the environment. Illumina MiSeq sequencing platform is widely used for 16S / 18S / ITS rDNA amplicon sequencing because of its deep sequencing depth, high throughput, short run-time and high sequencing accuracy as well as reasonable cost. In recent years, pair-end chemistry has enabled MiSeq sequencing platform to read as long as 600bp, which further increased the accuracy of the results.

16S / 18S / ITS rDNA amplicon sequencing procedure includes genomic DNA extraction, quality control, rDNA variable region amplification, library construction, high-throughput sequencing and data analysis. All the steps are important for data quality and quantity, which in turn affects the subsequent data analysis. In order to ensure data accuracy and reliability, every step has to pass strict quality control before pooling the library by adjusting the volume of each library according to the target data volume for Illumina MiSeq sequencing. The workflow is as below:

Figure 2.1 Microflora diversity experimental workflow

2.2. Bioinformatics Analysis Workflow

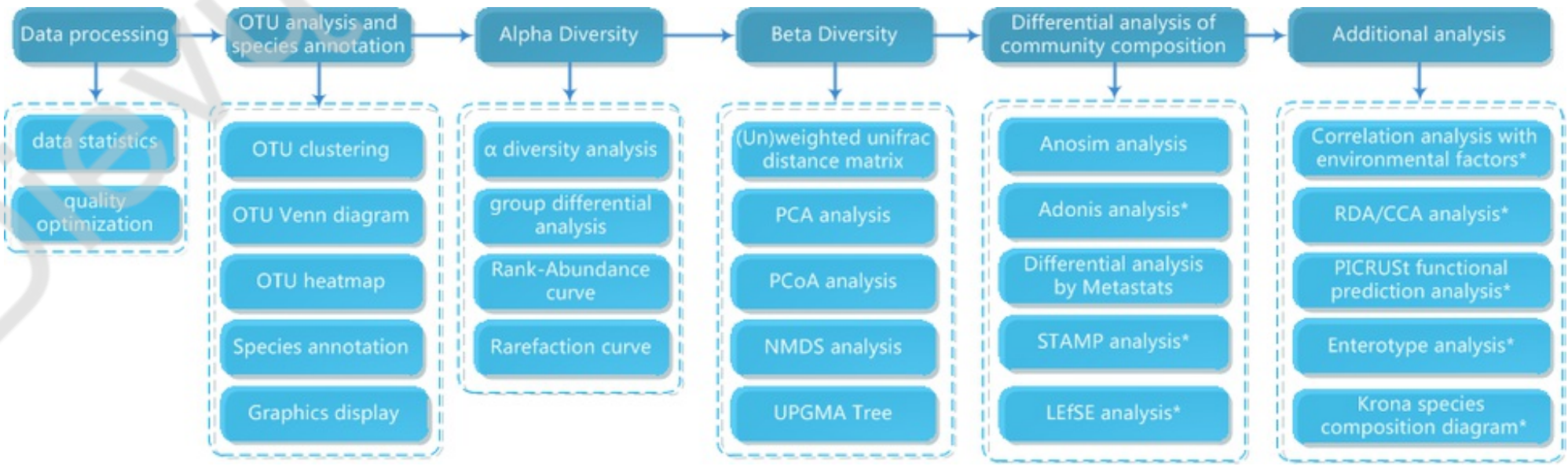

Figure 2.2 Microflora diversity data analysis workflow

Note:

(1) If the number of samples is less than 3, comparative analysis of diversity cannot be performed.

(2) The differential statistical analysis is meaningful only if there are at least three replicates in the biology group.

(3) Environmental factors are required for correlation analysis of community composition and environmental factors as well as for CCA / RDA analysis.

(4) Intestinal-type analysis only applies to animal or human intestinal or stool samples.

(5) Analysis with * are not included in the standard analysis and can be selected and analyzed according to individual sample and project.

3. ANALYSIS

3.1. Sample Summary Statistics

Show 10 entries

Search:

Sample	Length(bp)	#Reads	Bases(bp)	Q20(%)	Q30(%)	GC(%)	N(ppm)
DM1	249.33	562,694	140,294,457	96.80	91.41	52.05	13.73
DM2	249.20	507,848	126,557,334	96.52	90.88	52.30	17.47

Showing 1 to 2 of 2 entries

Copy CSV PDF Print

Previous 1 Next

3.2. Sequencing Data Quality Optimization

Sequencing errors might occur in high-throughput sequencing, and it is common that bases toward the end of the sequence reads have lower than average quality. In order to obtain higher quality and more accurate bioinformatic analysis results, it is necessary to optimize the raw data of the sequencing to obtain higher quality and more accurate bioinformatics analysis results.

Analysis software: Cutadapt(v1.9.1), Vsearch(1.9.6), Qiime(1.9.1)

Steps and parameters for optimization:

(1) The two sequences of each read pair were merged according to overlapping sequences. The read merge is deemed to be successful only if the overlapping sequence is least 20bp long. After merging, undetermined bases (N) were removed from the resulting sequence.

(2) Primer and adapter sequences were removed. Then the 5' and 3' bases with Q score lower than 20 were also removed. The resulting sequences with length > 200bp would pass this step of processing.

(3) The sequences obtained were then aligned to UCHIME 'Gold' database to identify and remove chimera sequence. Sequences passed this filtering step are deemed as clean data ready for analysis.

Show 10 entries

Search:

Sample	#PE_reads	#Nochimera	AvgLen(bp)	GC(%)	Effective(%)

DM1	281,347	247,991	458.84	51.99	88.14
DM2	253,924	205,016	458.27	52.40	80.74

Showing 1 to 2 of 2 entries

Copy CSV PDF Print

Previous 1 Next

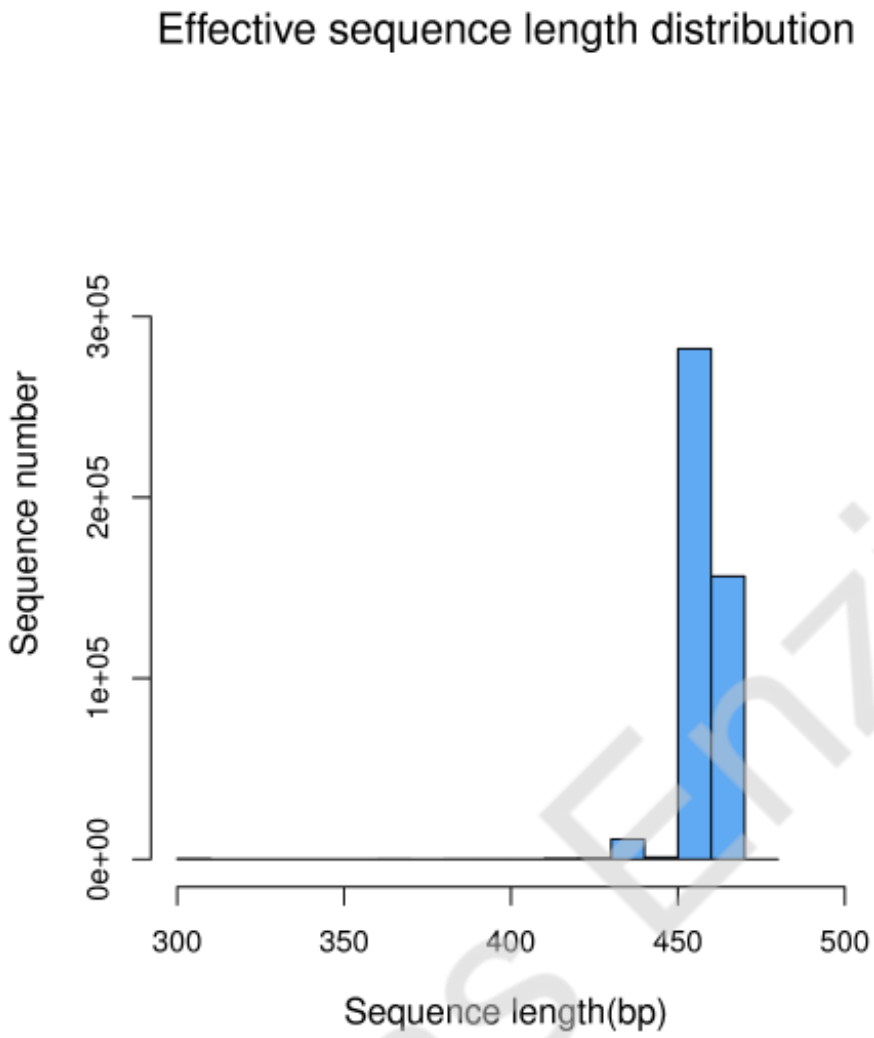

Figure 3.2 Effective sequences length distribution

3.3. OTU Analysis and Species Annotation

3.3.1. OTU clustering

OTU is an operational definition of a classification unit (genus, species, grouping, etc.) commonly used in population genetics to facilitate data analysis. In bioinformatics each sequence obtained from sequencing is assumed to be derived from a single species. All the sequences in a sample are classified to obtain information on species and genus. By classification, the sequences are grouped according to their similarity, and one group is an OTU. Typically, OTU clusters are defined by a 97% identity threshold for data statistics and analysis.

Analysis software:Qiime(1.9.1),Vsearch(1.9.6)

Analysis methods and steps:

(1) Unique sequences were extracted from the optimized sequences with the read count information.

(2) The unique sequences with 1 read count were removed.

(3) OTU clustering of unique sequences (read count > 1) was performed with similarity of 97%, and chimeric sequences were further removed to obtain the representative OTU sequences.

(4) All optimized sequences were compared with OTU representative sequences, and sequences of >97% similarity to a specific OTU representative sequence are considered to be of the same OTU. Finally, the OTU abundances were also summarized in the results table.

The following table shows the statistics of the sequence number in each sample's OTU:

Show 10 entries

Search:

OTU_ID	DM1	DM2	taxonomy
OTU1	170,373	11,403	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae
OTU10	6,619	14	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Leuconostocaceae;g__Oenococcus;s__oeni
OTU100	20	17	k__Bacteria;p__Bacteroidetes;c__[Saprospirae];o__[Saprospirales];f__Chitinophagaceae;g__;s__
OTU101	5	7	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae;g__;s__
OTU102	1	0	k__Bacteria
OTU103	4	1	k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__;g__;s__
OTU104	16	17	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Methylibium
OTU105	2	0	k__Bacteria;p__Chloroflexi;c__Anaerolineae;o__Caldilineales;f__Caldilineaceae;g__Caldilinea;s__
OTU106	6	4	k__Bacteria;p__Nitrospirae;c__Nitrospira;o__Nitrospirales;f__Nitrospiraceae;g__Nitrospira;s__
OTU107	50	54	k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae

Showing 1 to 10 of 220 entries

Copy CSV PDF Print

Previous 1 2 3 4 5 … 22 Next

3.3.2. OTU heatmap

The heatmap analysis shows the abundance information of selected OTU as well as the similarity and difference across OTUs and samples by similarity clustering. The figure below shows the top 30 OTUs with the highest abundance:

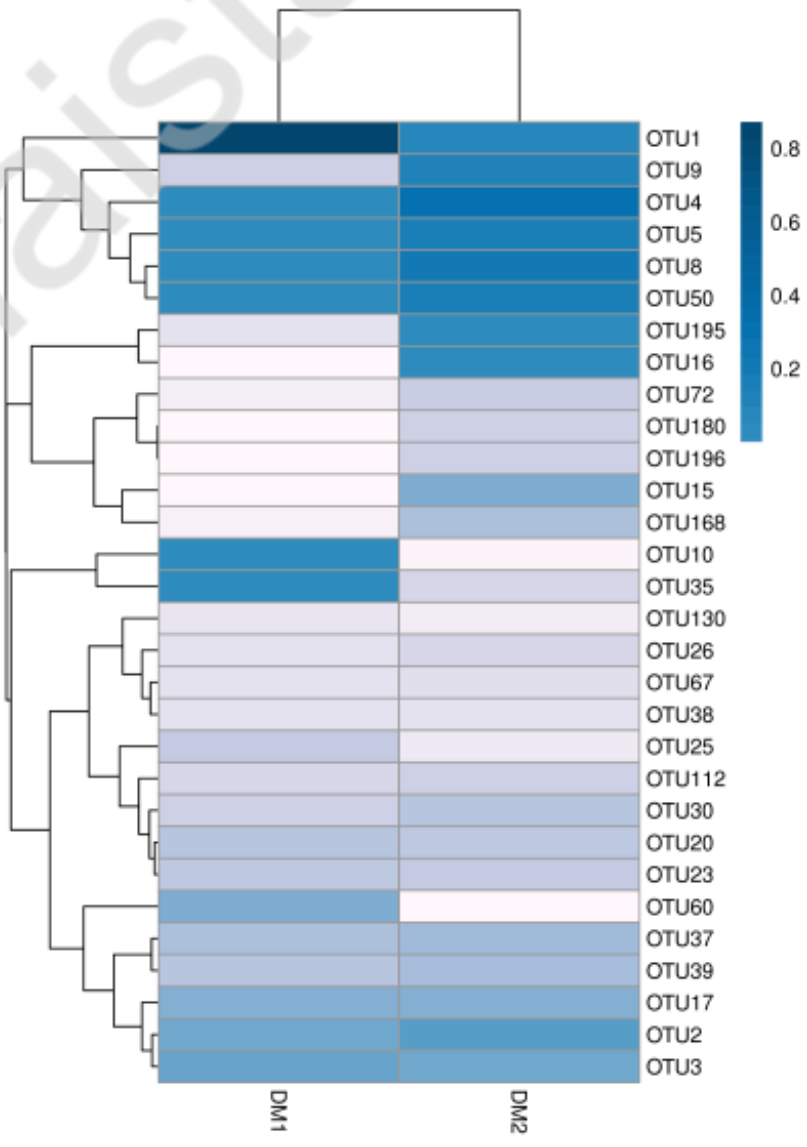

Figure 3.3.2 OTU abundance clustering heatmap

Note: The row name is the OTU ID, the column name is the sample information, the left side of the figure is the OTU cluster tree, and the top is the sample cluster tree. The value of each colored box is the relative abundance of each OTU after normalization.

3.3.3. OTU Venn diagram and petal diagram

According to the results of OTU cluster analysis, the common and unique OTUs of different samples/groups are analyzed. When the number of samples/groups is less than 5, the Venn diagram is drawn. When the sample/group is greater than 5, the petal diagram is drawn.

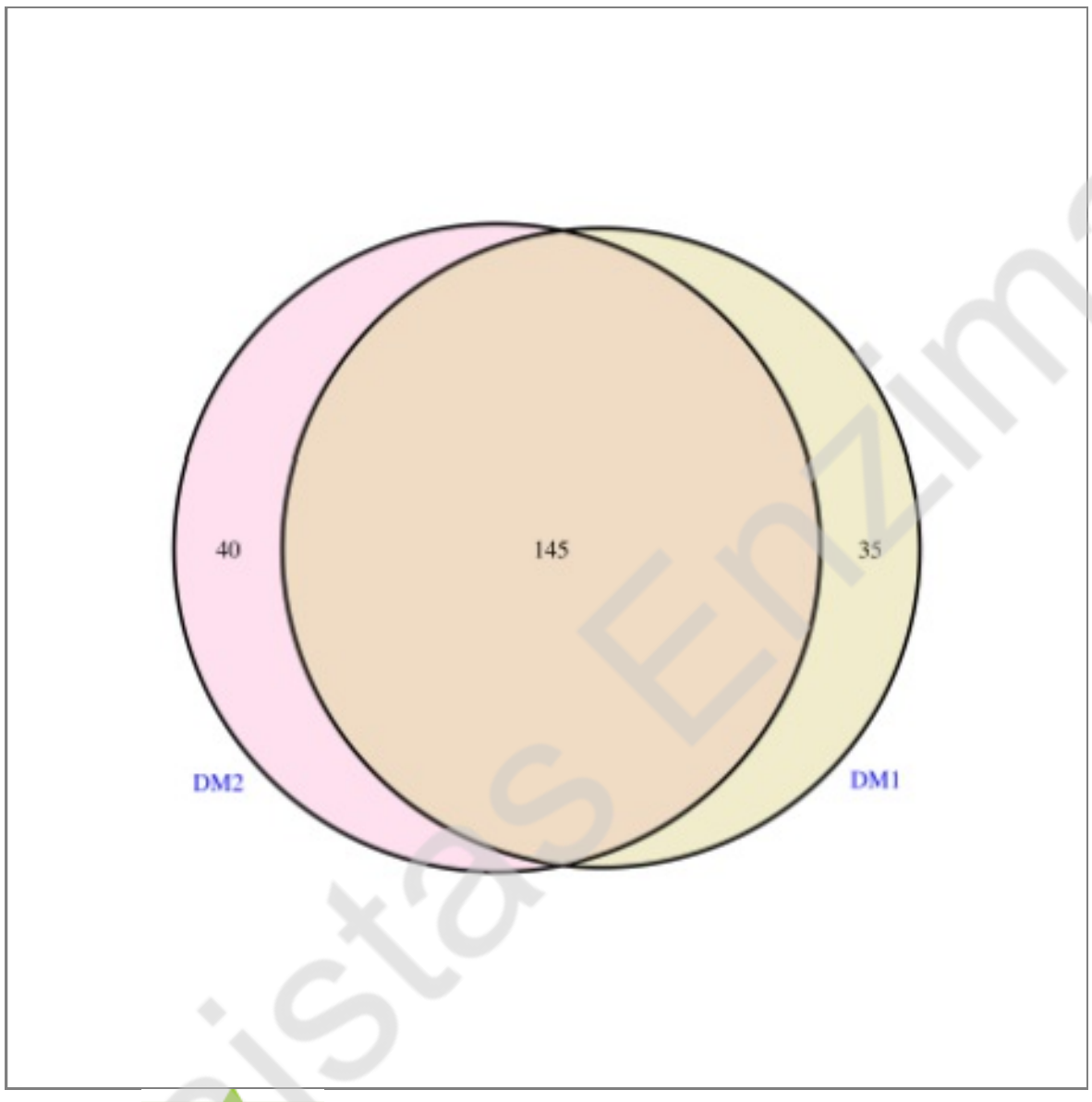

Figure 3.3.3 OTU Venn diagram or petal diagram

Note: the circles of different colors in the Venn diagram represent different samples or groups, and the numbers in the figure represent the numbers of OTUs unique or common to each sample or group. In the petal diagram, each petal represents a sample or group. The numbers on the petals represent the number of OTUs unique to the sample, and the white circle in the middle represents the number of OTUs shared by all samples and groups.

3.3.4. Species annotation statistics

In order to obtain the classification information of OTU, a representative sequence was selected for each OTU and annotated using the RDP classifier, thereby to obtain the community composition of each sample.

Analysis software: Qiime(1.9.1)

Analysis method: RDP classifier Bayesian algorithm was used to classify the OTU representative sequences of 97% similarity level, and the community composition of each sample was analyzed and summarized at all levels. The comparison database was Greengenes database

(http://qiime.org/home_static/dataFiles.html) / 18S rRNA database / ITS database.

For each sample, the percentage of each species at different taxonomic levels (Phylum, Class, Order, Families, Genus, Species) is shown in the table below:

Show 10 entries

Search:

Taxon	DM1	DM2
Acetobacter	267.00	269.00
Acidovorax	29.00	23.00
Acinetobacter	12.00	11.00
Bacillus	185.00	232.00
Balneimonas	2.00	0.00
Bifidobacterium	7.00	9.00
Caldilinea	64.00	53.00
Candidatus	24.00	26.00
Carica	0.00	9.00
Citrullus	1.00	20.00

Showing 1 to 10 of 48 entries

Copy CSV PDF Print

Previous 1 2 3 4 5 Next

Table 3.3.4.1 Taxa Statistics at Genus level

Show 10 entries

Search:

Samples	DM1	DM2
Class	25	21
Family	46	44
Genus	45	45
Kingdom	2	2
Order	42	35
Phylum	13	11
Species	15	14

Showing 1 to 7 of 7 entries

Copy CSV PDF Print

Previous 1 Next

Table 3.3.4.2 Statistics of Taxonomic Composition

3.3.5. Graphics display of species relative abundance

The distribution of the top 30 most abundant classifications in each sample or group at different taxonomic levels (Phylum, Class, Order, Families, Genus, Species) are shown as follows:

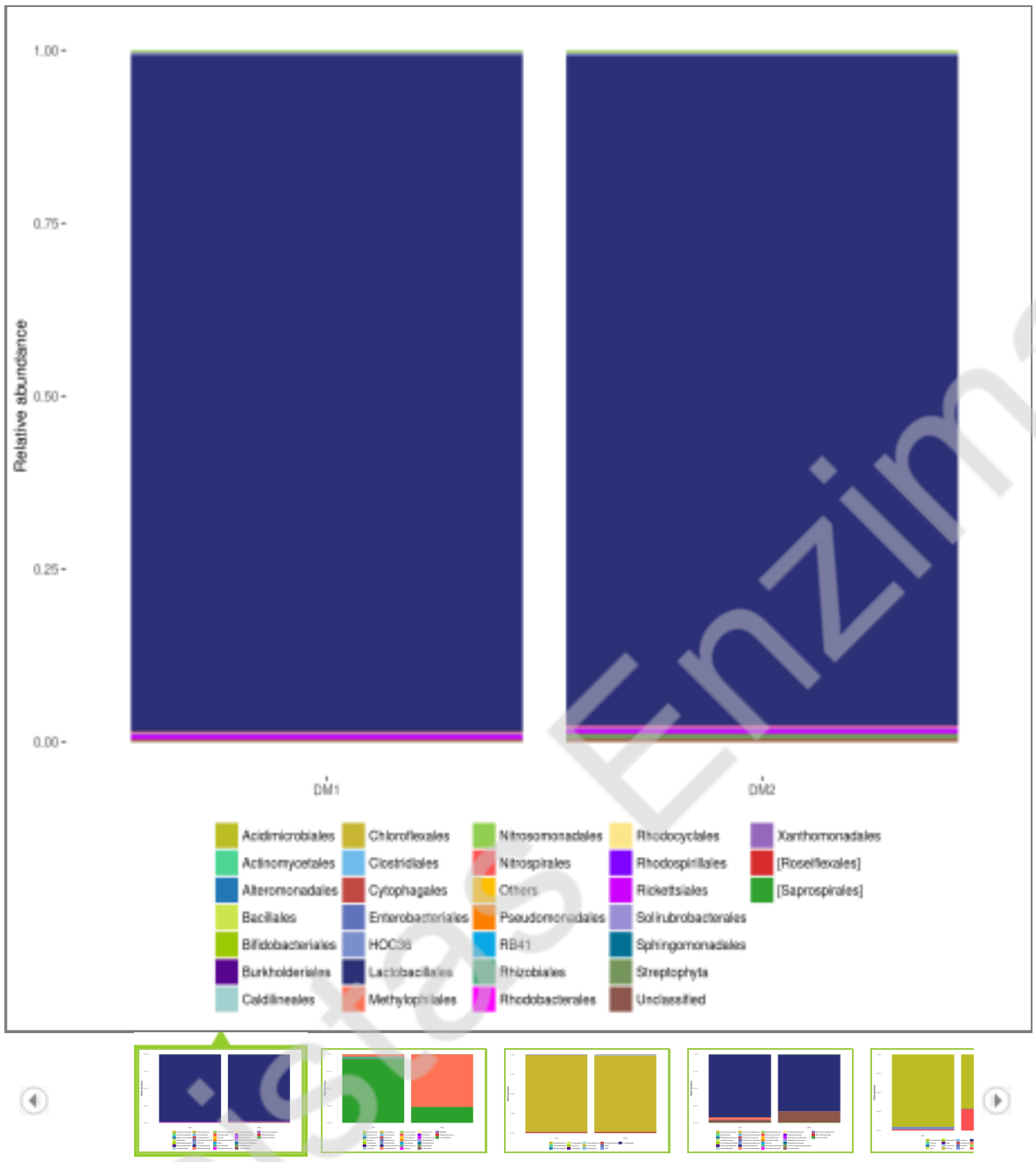

Figure 3.3.5.1 Stacked bar plot of species distribution

Note: X axis is the sample name or group name, and the Y axis is the relative abundance of different species. The legend is the name of the taxonomic classification of the species. 'Other' represents the relative abundance of all phylum level classifications other than the top 30.

The top 30 species distribution of each sample (or group) on different levels (Phylum, Class, Order, Families, Genus, Species) was clustered and plotted in a heatmap. The similarity and difference of each species is visualized by color scheme in the heat map. The heat maps plotted for the distribution of the species in each sample on different levels of classification are shown below:

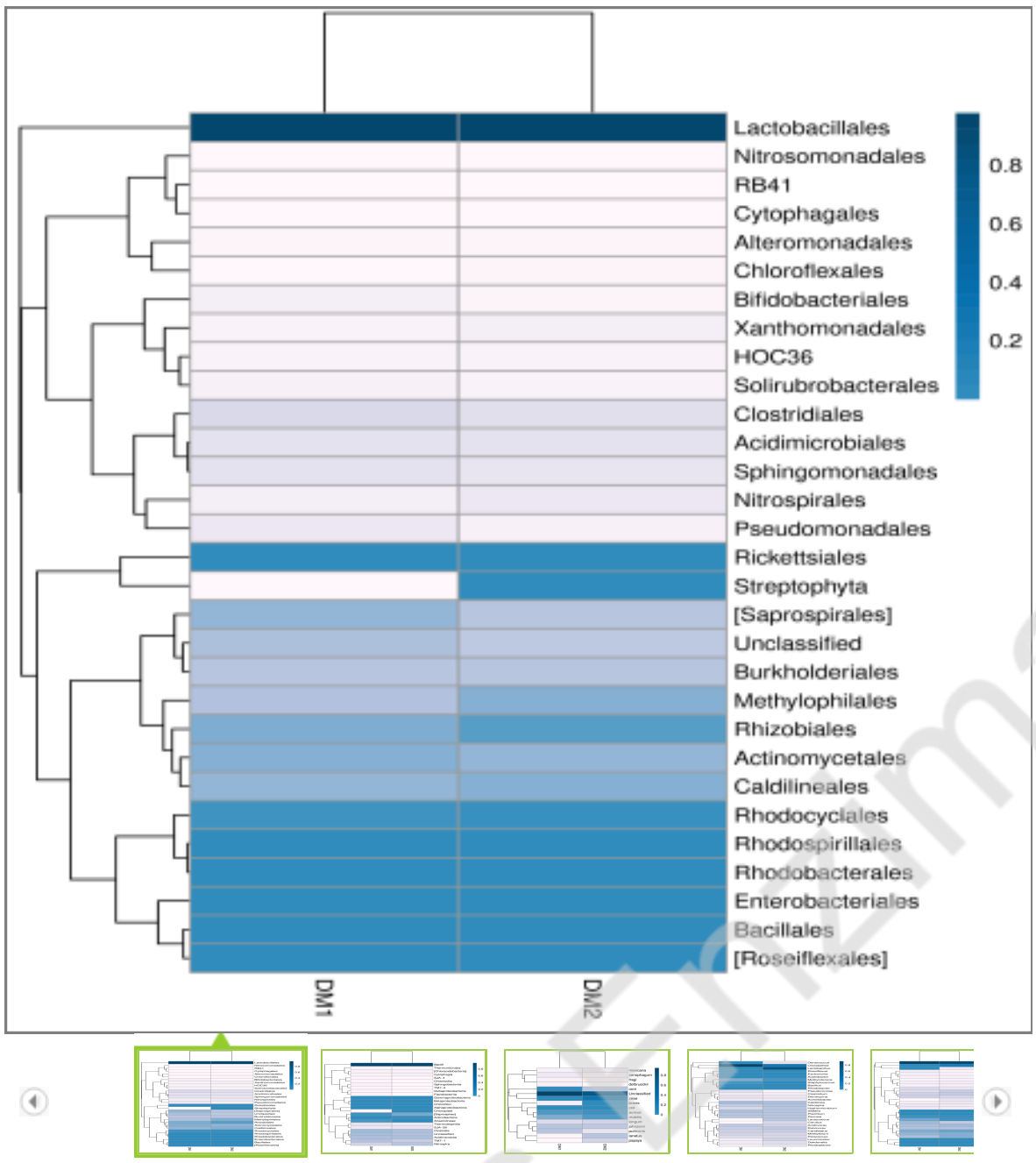

Figure 3.3.5.2 Species distribution heatmap

Note: The columns represent samples and/or groups and the rows represent species. The dendrogram above the heatmap is the cluster result of the samples and/or groups and the dendrogram to the left is the species cluster. The colors in the heat map represent the relative abundance of the corresponding species in the corresponding sample or group.

3.4. Sample complexity analysis

3.4.1. Alpha diversity analysis

In community ecology, alpha diversity is mainly used to reflect the diversity of each sample, which estimates the number of species in the microbial community as well as the abundance and diversity of species in environmental communities through a series of statistical indices.

The indices for community richness calculation include:

ACE: An index to estimate the number OTU in communities. It was proposed by the Chao and is commonly used to estimate the total number of species in ecology. (http://www.mothur.org/wiki/Ace)

Chao: It is an index that uses the Chao 1 algorithm to estimate the OTU number of samples. Chao is commonly used in ecology to assess the total number of microorganisms.(http://www.mothur.org/wiki/Chao)

The indices for community diversity calculation include:

Shannon: Commonly used to reflect the diversity index a for the estimation of microbial diversity. (http://www.mothur.org/wiki/Shannon)

Simpson: Simpson diversity index, proposed by Edward Hugh Simpson in 1949 and is commonly used in ecology to quantify biological diversity in a region. (http://www.mothur.org/wiki/Simpson)

The indices to measure sequencing depth (Coverage) include:

Goods Coverage: Refers to library coverage of each sample. The higher the value, the lower the probability that the sample did not cover the sequence.(http://www.mothur.org/wiki/Coverage)

Analysis software:Qiime (1.9.1)

Analysis method: sequences were randomly extracted and the valid sequences were subject to OTU analysis and a diversity index was calculated for each sample.

Alpha diversity results are summarized in the table below:

Show 10 entries

Search:

sample	ace	chao1	shannon	simpson	goods_coverage
DM1	183.66	182.25	1.09	0.25	1
DM2	187.07	188.50	2.92	0.82	1

Showing 1 to 2 of 2 entries

Copy CSV PDF Print

Previous 1 Next

Table 3.4.1 Collation of alpha diversity results

3.4.2. Between-group differential analysis using alpha-diversity indices

To perform between-group a diversity analysis, box plots were generated based on a diversity indices using R, which intuitively displays the maximum, minimum, median and outliers of the a diversity indices of samples in each group as well as the differences between groups.

Analysis method: Box plot was generated using R based on a-diversity index.

The box plot of the inter-group differential analysis based on chao1 and shannon indices is as below:

The alpha index boxplot can't be finished as don't define the group information.

Note: The left panel is the Chao1 index boxplot of each group. X axis indicates the names of the groups and Y axis indicates the Chao 1 index. Each box diagram shows the minimum, first quartile, medium, third quartile and maximum values of the chao1 index of the corresponding sample. The right graph is the Shannon index boxplot of each group.

3.4.3. Rank-abundance curve

Rank-abundance curve is used to analyze diversity. To generate a rank-abundance curve, the number of valid sequences in each OTU of a given sample was first calculated, and then all the OTUs were ranked in descending order based on their relative abundance (number of valid sequences), and finally the result was plotted with OTU ranking on the X axis and the number of sequences in the OTU on the Y axis. Y axis could also be OTU relative abundance in percentage.

Rank-abundance curve reflects both species abundance and species uniformity. The abundance of species is reflected by the length of the curve on the X axis. The more extended on the X axis, the more abundant the species is. Species uniformity is reflected by the shape of the curve. The smoother the curve, the higher the species uniformity.

Analysis method: R packages were used for graph generation based on the results of OTU analysis

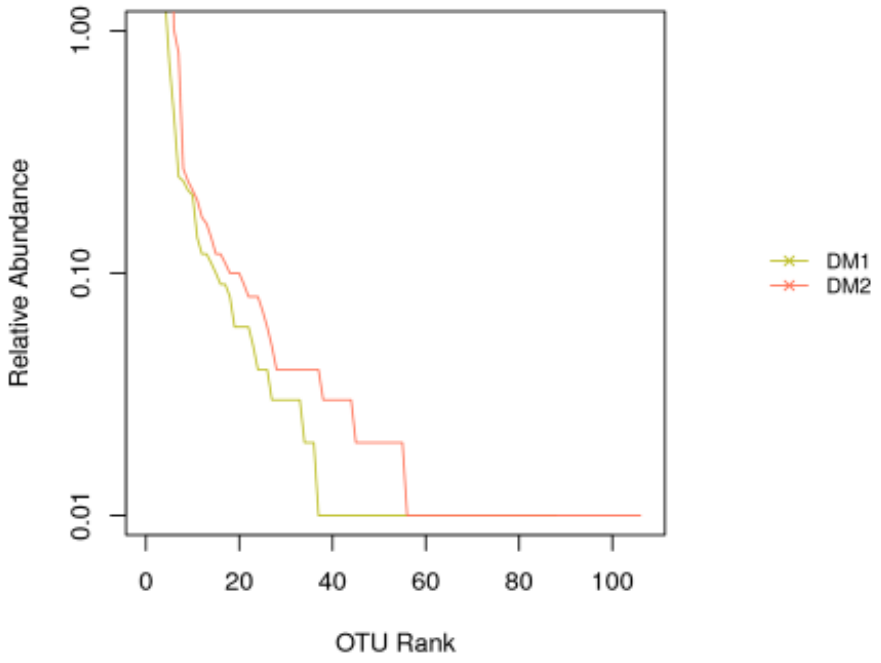

Figure 3.4.3 Rank-Abundance curve

Note: Each curve in the figure above corresponds to an individual sample. The X axis is the relative abundance of the OTU in descending order. The Y axis is the relative abundance of the OTU. '100' on the X axis indicates the OTU in the sample is ranked as the 100th abundant in descending order, and the corresponding value on the Y axis is the percentage of the sequence count in the OTU (the number of sequences of the OTU divided by the total number of sequences).

3.4.4. Rarefaction curve

The rarefaction curve is a useful tool to characterize the species composition of a sample and predicting the abundance of species in a sample. It efficiently deals with the increase of detected species due to the increase in sample size. It is widely used in biodiversity and community surveys to determine whether the sample size is sufficient and to estimate the species abundance. Therefore, the rarefaction curve can not only determine whether the sample size is sufficient, but also predict the species abundance when the sample size is sufficient.

Analysis software: Qiime (1.9.1)

Analysis method: The rarefaction curve was constructed by random sampling. The observed numbers of OTUs were plotted against the number of extracted sequences.

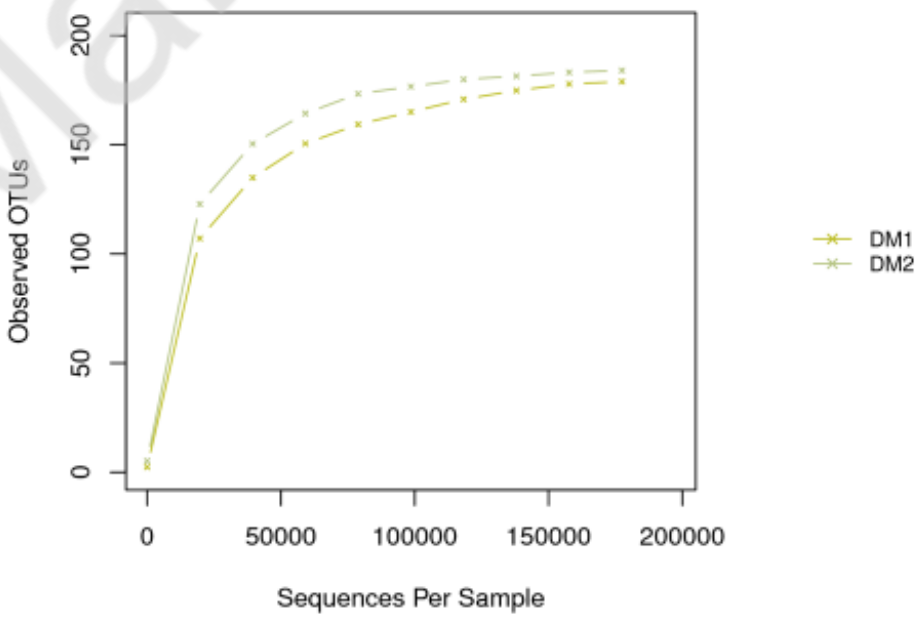

Figure 3.4.4 Observed OTUs rarefaction curves

Note: The X axis is the number of valid sequences extracted, and the Y axis is the number of OTUs (Observed OTUs). Each sample is represented by one curve with a unique color. The number of OTUs increases with the increase of extracted sequence count until reaching a plateau, which indicates the number of detected OTUs will not increase with the amount of extracted sequences and reflects the reasonable sequence depth.

3.5. Multi-sample comparative analysis

3.5.1. (Un)weighted unifrac distance matrix

Beta diversity reflects the diversity and the degrees of differences among samples. The distance between the samples can be calculated using the evolution and abundance information between the sample sequences to reflect whether there is significant difference in microbial community among the samples. This can be achieved by UniFrac analysis.

Analysis software: Qiime (1.9.1)

Analysis method: a phylogenetic tree was constructed using the OTU representative sequences from different environmental samples. The Unifrac metric was then used to measure the difference between two different environmental samples according to the length of the constructed evolutionary tree.

UniFrac analysis includes weighted unifrac and unweighted unifrac methods. The difference between the two is whether to include the relative abundance of sequences from different environmental samples. The weighted unifrac algorithm weights the sequence abundance information when calculating the length of the branch, so unweighted unifrac detects the change among the samples, and the weighted unifrac further quantifies the variation on different pedigrees.

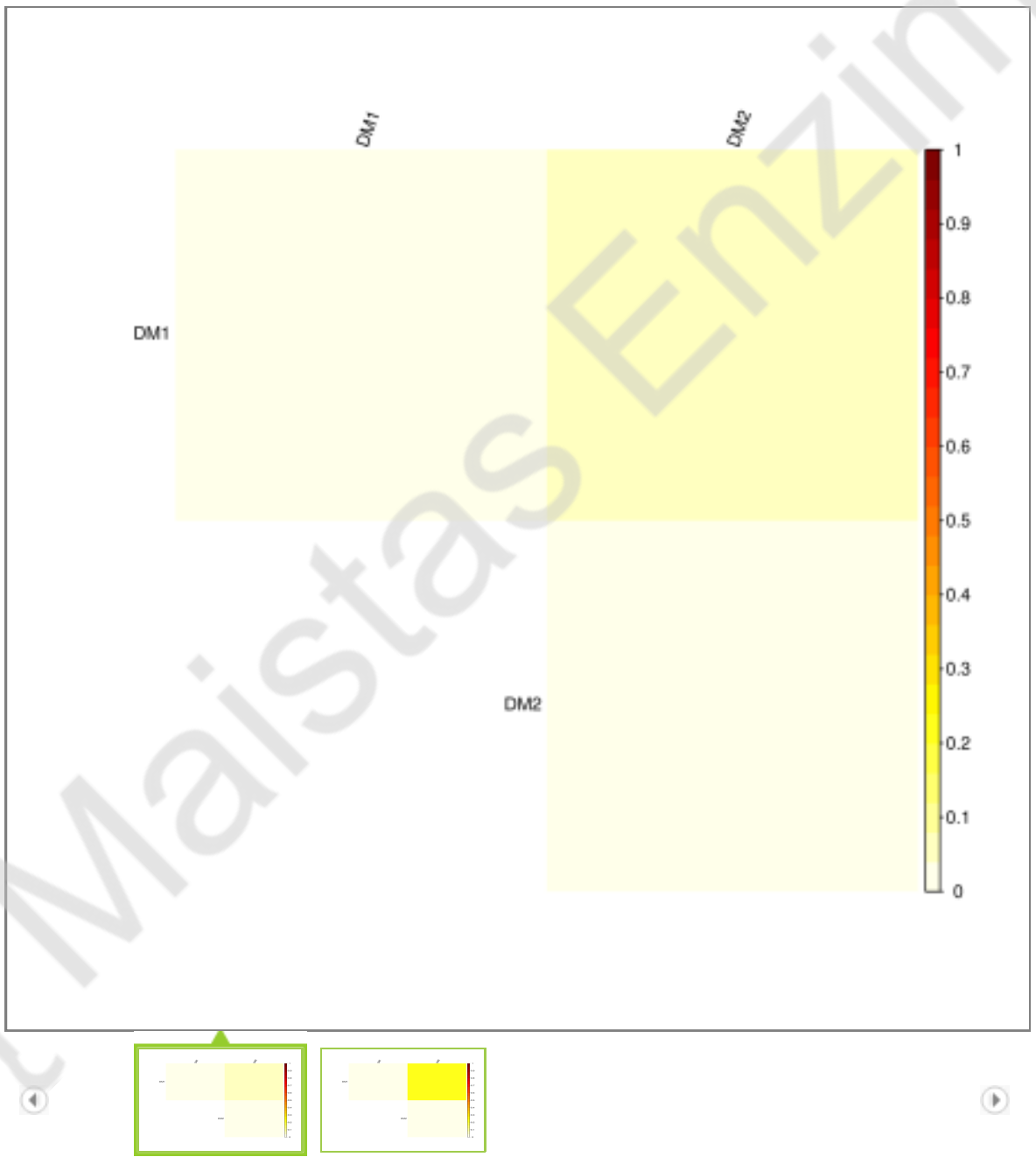

Figure 3.5.1 (Un)weighted unifrac distance matrix heatmap

Note: (Un)weighted unifrac distance matrix heat map. The color scheme in the heatmap represents the degree of difference between the two samples. The lighter the color, the smaller the coefficient between the two samples, and the smaller the difference of species diversity.

3.5.2. PCoA analysis

PCoA (Principal Co-ordinates Analysis) analysis characterizes and visualizes the similarities and differences of data. Similar to PCA, it sorts data based

on a series of eigenvalues and eigenvectors and uses the top ranked eigenvalues to determine the most important coordinates in the distance matrix. It also uses eigenanalysis to perform a rigid rotation of the original axes that only changes the coordinates but not the positional relationship of different sample points. The difference between the two is PCA determines the principal components based on sample similarity coefficient matrix whereas PCoA uses distance matrix to find the primary coordinates.

Analysis software: R

Analysis method: Analysis method: PCoA analysis was performed and plotted based on Brary-Curtis distance matrix

PCoA can't be finished as the number of samples doesn't meet the demand.

Figure 3.5.2 PCoA Plot

Note: Samples of the same group are represented in the same color and shape. PC1_vs_PC2 is the PCoA plot obtained for the first and second principal coordinates; the X and Y axes represent the first and second principal coordinates, respectively. The value in percentage in the axis label represents the contribution of the corresponding coordinate to the sample variance and measures how much this principal is extracted from the original information. The distance between the sample points indicates the similarity of the microbial community in the sample. The closer the points, the higher the similarity. Samples clustered together are composed of similar microbial compositions.

3.5.3. PCA analysis

PCA analysis (Principal Component Analysis) is a statistical technique for the determination of the key variables in a multidimensional data set that are most responsible for the differences in the observations, and thus is commonly used to simplify complex data analysis.

The difference and distance between samples can be reflected by the analysis of the gene functional distribution of different samples. The differences between multiple sets of data can be plotted on a two-dimensional chart using variance decomposition, with the axes representing two eigenvalues that reflects the largest variance values.

Figure 3.5.3 PCA graph

Note: PC1, PC2, PC3 represent the first, second and third principal components, respectively. The percentage after the principal component represents the contribution rate of this component to sample difference and measures how much information the principal component can extract from the original data. The distance between samples indicates the similarity of the distribution of functional classifications in the sample. The closer the distance, the higher the similarity.

3.5.4. NMDS analysis

The non-metric multidimensional scaling is a data analysis method that reduces multi-dimensional space to low-dimensional space to simplify the localization, analysis and classification of research objects. This method preserves the primitive relation among the objects. Its main feature is to position each object in multi-dimensional space based on its functional classification information and calculate the distances between different objects (points) as a measurement of their difference, which are used to obtain the spatial position map.

Analysis method: Graph was generated using vegan package in R based on the beta diversity distance matrix.

NMDS plot can't be finished as the number of samples doesn't meet the demand.

Figure 3.5.4 NMDS plot

Note: Each point represents a sample, and the distance between the points indicates the degree of difference. Samples of the same group are represented by the same color. Stress < 0.2 indicates NMDS can accurately reflect the difference between the samples.

3.5.5. UPGMA Tree

Cluster analysis uses evolutionary information derived from sample sequences to calculate whether samples in a specific environment is significantly different from a evolutionary lineage in microbial communities.

Analysis software: Qiime (1.9.1)

Analysis method:The UPGMA (Unweighted pair group method with arithmetic mean) clustering method was used to cluster the samples based on the Brary-Curtis distance matrices.

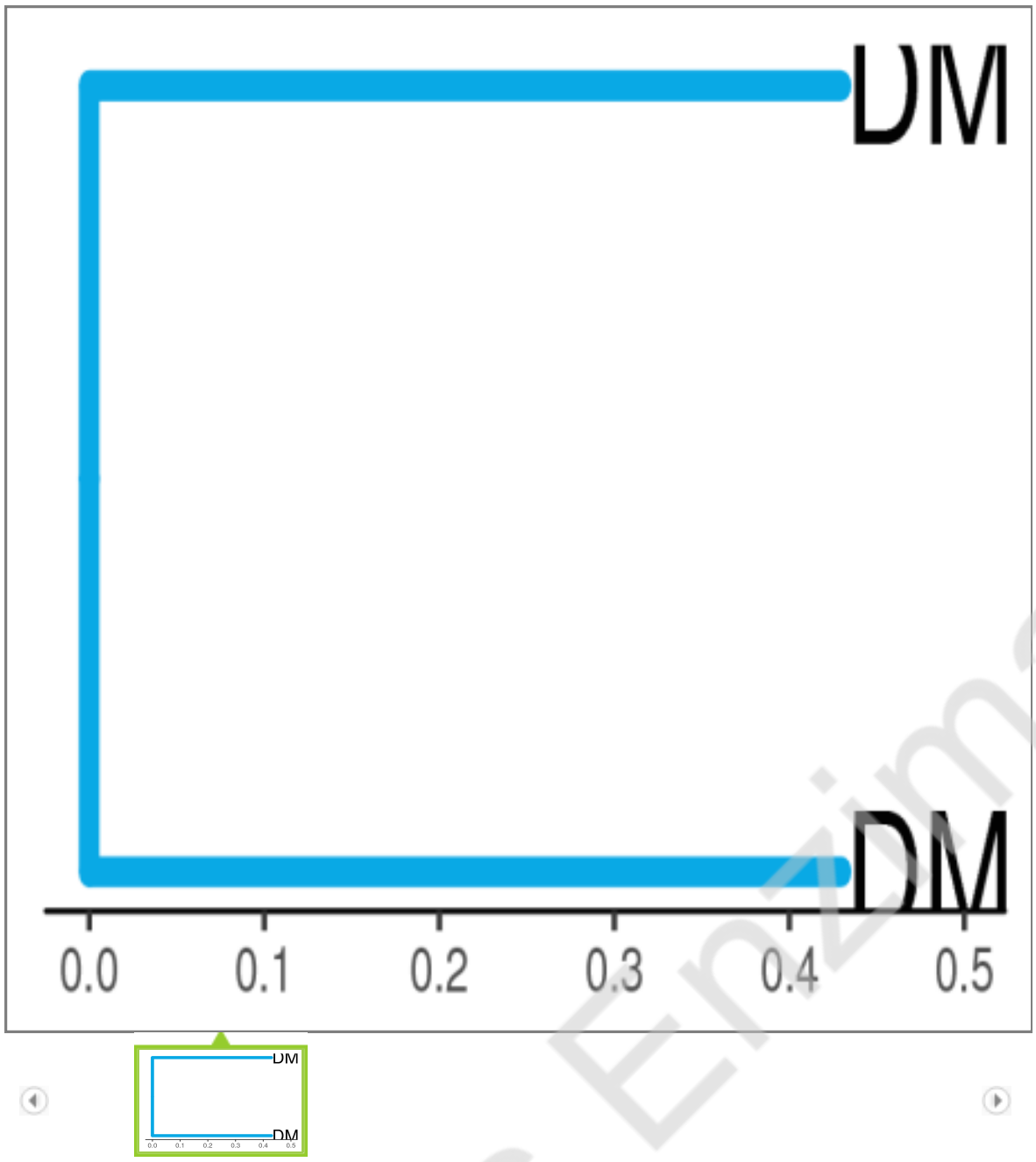

Figure 3.5.5 UPGMA tree

Note: Each branch in the figure represents a sample. Different colors representing different groups.

3.6. Differential analysis of community composition

3.6.1. Anosim analysis

ANOSIM is a nonparametric test used to examine whether the difference between two or more groups is statistically greater than the within-group difference. It was used to determine whether the grouping is meaningful. R value was obtained by analyzing the sample distance matrix.

Analysis software: vegan package in R

Anosim can't be finished as number of samples doesn't meet the demand.

Table 3.6.1 Group difference evaluation by Anosim

Column description:

(1) Factor:The factors used for grouping

(2) R value: The value of R ranges from 0 to 1. The closer it is to 0, the less significant the between-group difference is compared to within-group difference; the closer it is to 1, the more significance the between-group difference compared to within-group difference.

(3) P value: P-value indicates the statistical credibility, $P < 0.05$ indicates the result is statistically significant.

The results of the Anosim analysis were ranked by the value of the distances between the two samples. For each group pair, three set of data can be obtained: between-group distances and within-group distances of each of the two groups. The result is shown below in box plot.

Anosim plot can't be finished as the number of samples doesn't meet the demand.

Figure 3.6.1 Group comparison by Anosim analysis

Note: X axis is the ranking of the distances of samples. 'Between' represents the result between the two groups, the other two are the results of within-group distance calculation. R values close to 1 indicates the difference between the groups is significantly greater than the within-group difference. P < 0.05 indicates the result is statistically significant.

4. DELIVERABLES

00_Data: Raw data statistics

01_OTU_Taxa: OTU and Taxonomy results

02_Alpha_Diversity: Sample conplexity analysis results

03_Beta_Diversity: Multi-sample comparative analysis results

04_Diff_Comparison: Differential analysis results of community composition

5. REFERENCE

[1] JG, Kuczynski J. et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5): 335-336(2010).

[2] Crawford, P. A., Crowley, J. R., Sambandam, N., Muegge, B. D., Costello, E. K., Hamady, M., et al. (2009). Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. Proc Natl Acad Sci U S A, 106(27), 11276-11281.

[3] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069-5072 doi:10.1128/AEM.03006-05.

[4] Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Opens external link in new windowNucl. Acids Res. 41:e1

[5] Westram R, Bader K, Pruesse E, Kumar Y, Meier H, Glockner FO, Ludwig W (2011) ARB: a software environment for sequence data. In: de Bruijn FJ (ed) Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches.Opens external link in new window John Wiley & Sons, Inc., pp 399-406

[6] Yu Wang, Hua-Fang Sheng, et al. Comparison of the Levels of Bacterial Diversity in Freshwater, Intertidal Wetland, and Marine Sediments by Using Millions of Illumina Tags. Appl. Environ. Microbiol. 2012, 78(23):8264. DOI: 10.1128/AEM.01821-12.8

[7] Price MN, Dehal PS, Arkin AP (2010) FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5(3): e9490. doi:10.1371.journal.pone.0009490.

[8] Micah Hamady, Catherine Lozupone and Rob Knight. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. The ISME Journal (2010) 4, 17-27; doi:10.1038/ismej.2009.97