MetaGenomics Analysis Demo Report

May 1, 2016

Contents

1 Introduction

Microbes distribute ubiquitously in natural communities. Expanding from skin to gut, ranging from air in mountain area to mud in deep sea and dwelling across iced lake to volcano, the wildly spread microorganisms contribute a lot to the environment. Since the invention of microscope by Antoni van Leeuwenhoek several centuries ago, the traditional and powerful researching strategy in microbiology is culturing. Limited by the small percentage, 0.1%~1%, of all microorganisms in nature which can be successfully cultured in laboratory, these huge abundant resources remained unutilized.

Metagenomics is a strategy first proposed by Handelman to directly study the through genomic information contained in the samples. It was further defined by Kelvin that it is a discipline about studying the microbial community via genomics methods to circumvent the culturing step. It avoid culturing the microbe in the samples, provide a method to study microbes that can't be cultured, more truly react the component and interaction of microbes in the samples, and we can study it's metabolic pathway and gene function in molecule level.

With the rapid development of sequencing technology and informatics technology, metagenomics studies with Next Generation Sequencing (NGS) is a fundamental strategy to study the community diversity and characteristics being famous for its ability to get tremendous data and abundant information. More and more far-reaching projects have utilized NGS in their research, like the Human Microbiome Project (HMP, http://www.hmpdacc.org/) and Earth Microbiome Project (EMP, http://www.earthmicrobiome.org/).

2 Workflow

2.1 Library construct and sequencing

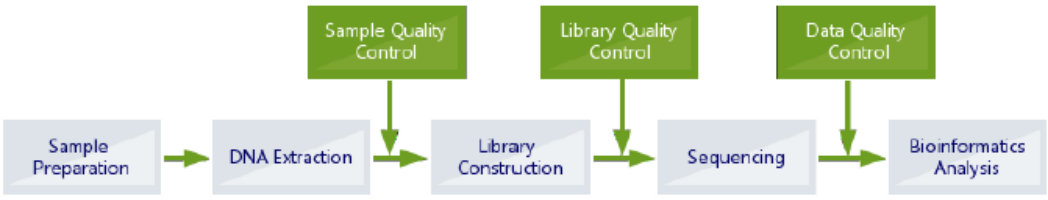


Figure 2.1 Experimental workflow of metagenomic sequencing

2.2 Bioinformatics analyze pipeline

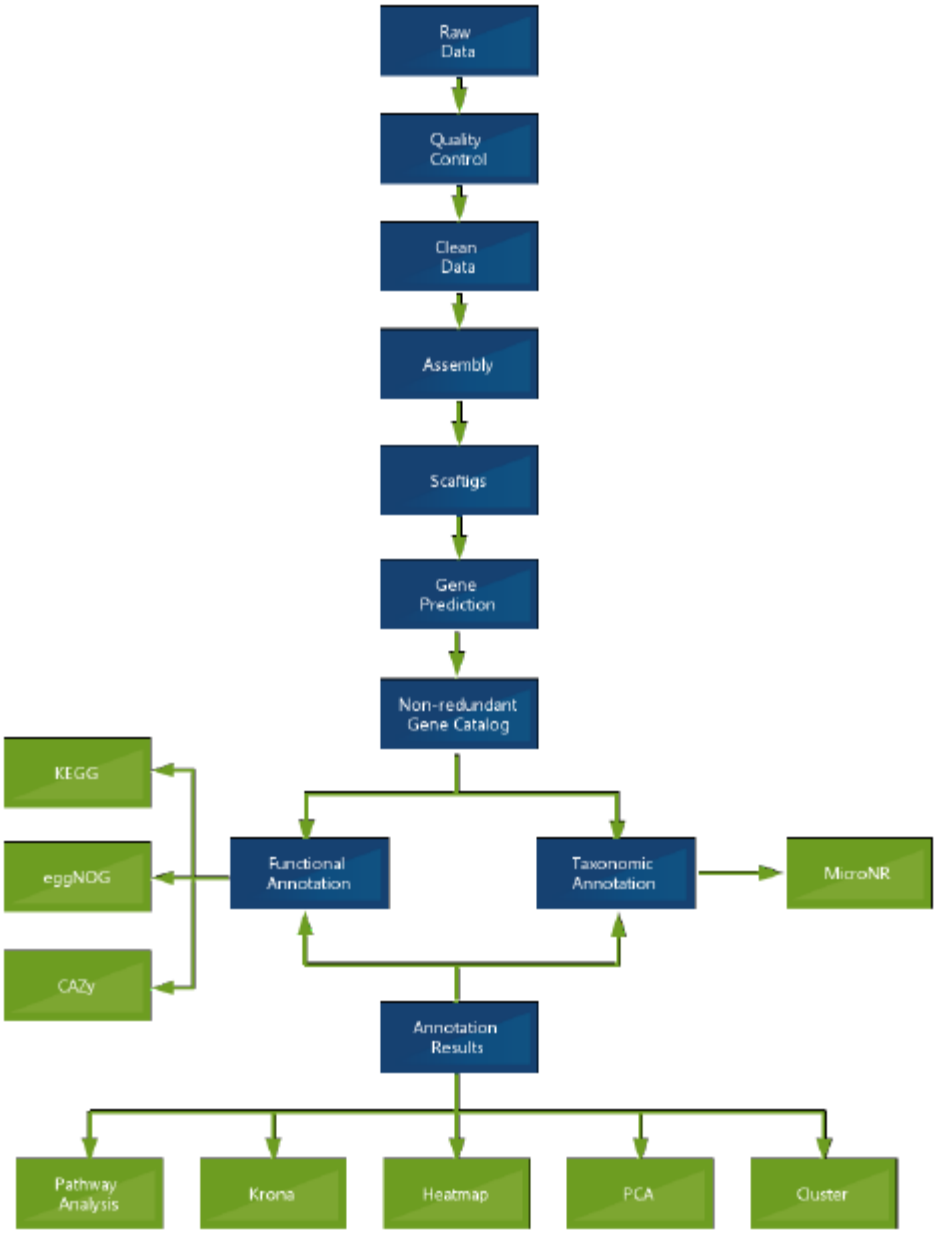


Figure 2.2 Workflow of Bioinformatics Analysis

3 Results

3.1 Data Pre-processing

Raw data from Illumina HiSeq platform sequencing exist a certain percentage of low quality data, to ensure the accuracy and the reliability of following information analysis, the initial step of metagenomic data analysis requires the execution of certain pre-filtering steps, and clean data is obtained. The protocols of data pre-processing are as the following:

1) Eliminate reads whose low-quality nucleotides (Q-value ≤ 38) exceed certain

threshold (40bp by default);
2) Eliminate reads which contain N nucleotides over certain threshold (10bp by default);
3) Eliminate reads which overlap with adapter over certain threshold (15bp by default);
4) If the sample exist contamination, they need to be blasted with host database, to filter reads which are probably come from the host (use SoapAligner by default, and the default arguments are: identity ≥ 90%, -l 30, -v 7, -M 4,-m 200,-x 400).

Steps above are all for read1 and read2 analysis. Data pre-processing statistics result is as table 3-1, for more detail information, please click QC_Report.

Table 3.1 Statistics for Data Pre-processing

#Sample	InsertSize(bp)	RawData	CleanData	Clean_Q20	Clean_Q30	Clean_GC(%)	Effective(%)
Monkey3	300	5,325.69	5,259.49	94.20	88.62	50.23	98.757
Monkey4	300	5,549.56	5,465.02	94.38	89.12	50.27	98.486
Monkey2	300	5,781.12	5,700.68	95.97	91.90	42.47	98.609
Human2	300	5,263.63	5,004.54	94.20	88.82	51.13	95.078
Human4	300	5,471.88	5,090.86	93.74	87.90	54.05	93.037
Human1	300	6,378.60	5,382.21	94.78	89.89	50.26	84.379
Monkey5	300	5,787.62	5,688.93	94.77	89.70	50.75	98.295
Monkey1	300	5,337.27	5,142.07	93.81	88.21	50.45	96.343
Human3	300	7,518.60	5,515.88	93.82	88.25	47.64	73.363
Human5	300	6,258.84	4,993.72	95.13	90.55	49.37	79.787

Page 1 of 1 10 View 1 - 10 of 10

Note:

The contents in the above table titled #sample, InsertSize, RawData, CeanData, Clean_Q20,Clean_Q30, Clean_GC(%), Effective(%) means samples' name, the fragments' length used in contructing sequencing library(bp), the data size of raw data, data volume remained after QC steps, the percentage of bases whose quality scores are greater than 20 or error rate is less than 0.01, the percentage of bases whose quality scores are greater than 30 or error rate is less than 0.001, the GC content, the ratio of the CleanData over the RawData.

3.2 Metagenome Assembly

1) Clean Data is assembled by SOAP denovo software;
2) Different K-mers are chosen for each individual sample for assembly (49,55,59 chosen as default). The result with the longest N50 is defined as the final result for assembly;
3) The Scaffolds are interrupted at N to get Scaftigs (continuous sequences within scaffolds);
4) Blast the cleandata from QC to Scaftigs from assembly of every sample using SoapAligner, obtain unutilized PE reads. Blast arguments: -d 1, -M 3, -R, -u, -F;
5) Put all the unutilized reads together, and conduct mixed assembly, while assembling, consider of computing cost and time cost, we choose only one kmer for assembly(-K 55 by default), other assemble arguments are the as single sample assembly;
6) The Scaffolds from mixed assembly are interrupted at N to get Scaftigs(continuous sequences within scaffolds);
7) Scaftigs whose length are less than 500bp are omitted and the remain is used for further statistical analysis and gene prediction.

Table3.2 Statistcs for assembled scaftigs

SampleID	Total len.(bp)	Num.	Average len.(bp)	N50 Len.(bp)	N90 Len.(bp)	Max len.(bp)
Human1	55,539,093	31,919	1,740.00	2,886	660	188,061
Human2	31,600,889	13,933	2,268.06	4,883	755	113,408
Human3	47,230,856	23,758	1,988.00	3,829	705	195,179
Human4	52,729,704	39,359	1,339.71	1,578	605	48,252
Human5	38,674,579	20,417	1,894.23	3,501	688	190,251
Monkey1	6,900,936	5,505	1,253.58	1,385	606	108,565
Monkey2	9,732,612	7,216	1,348.75	1,688	624	47,055
Monkey3	10,352,396	6,013	1,721.67	2,709	684	174,761
Monkey4	10,325,984	5,852	1,764.52	2,710	696	174,803
Monkey5	9,573,056	5,685	1,683.91	2,517	685	40,227

Page 1 of 2 | 10 | View 1 - 10 of 11

Note:

The contents in the above table titled Sample ID, Total len, Num, Average len, Scaftigs, N50, N90, Max Len means samples' name, the total length of all the assembled Scaftigs, the total number of all the assembled Scaftigs, the average length of all the Scaftigs, N50, N90, and the maximum length of all Scaftigs

According to the assembly result, do statistics to Scaftigs length distribution of each sample, and plot for it, the result is shown as the follow figure:

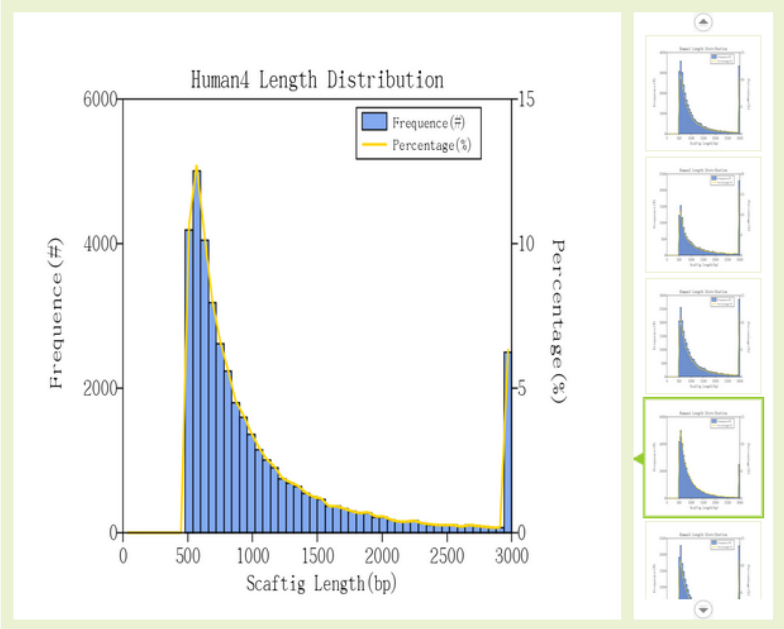


Figure 3.2 The length distribution of Scaftigs of each sample

The Y1-axis titled "Frequence (#)"means the numbers of Scaftigs of certain length; The Y2-axis titled "Percentage (%)" means the percentage of Scaftigs of certain length accounts for the total Scaftigs; The X-axis titled "Scaftigs Length (bp)" indicates the length of Scaftigs.

3.3 Gene prediction and abundance analysis

3.3.1 Gene prediction and abundance analysis workflow

1) Scaftigs are used for gene prediction by MetaGeneMark software.
2) The redundant predicted genes are omitted by CD-HIT software.
3) The reads from Clean Data that can be mapped to the representative non-redundant genes are calculated using SoapAligner;
4) The predicted genes with less than two reads supported are eliminated, and the remain is used for the sequential analysis of gene catalogue (Unigenes);
5) The gene abundance is calculated based on the total number of mapped reads and gene length. Computational formula is as following:

$$G_k = \frac{r_k}{L_k} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$$

R represent the total number of mapped reads, L represent gene length
6) Based on the abundance information of Unigenes in each sample, basic information is summarized, core-pan gene analysis and correlation analysis are performed, and venn figures are drawn.

3.3.2 Gene catalogue basic information statistics

Table 3.3.2 Gene catalogue basic information

ORFs NO.	203,305
integrity:end	35,581(17.5%)
integrity:none	20,013(9.84%)
integrity:start	40,501(19.92%)
integrity:all	107,210(52.73%)
Total Len.(Mbp)	157.92
Average Len.(bp)	776.75
GC percent	50.70

Note:

ORFs NO. indicates the gene number in gene catalogue; integrity:start means the number and percent of genes which only contain initiation codon; integrity:end means the number and percent of genes which only contain termination codon; integrity:none means the number and percent of genes which not only haven't initiation codon and also haven't termination codon; integrity:all means the number and percent of integrated genes(have both initiation codon and termination codon); Total Len.(Mbp) means the length of gene in gene catalogue, unit is million; Average Len. indicates the mean length of genes in catalogue; GC Percent indicates the total GC percent of genes in predicted gene catalogue.

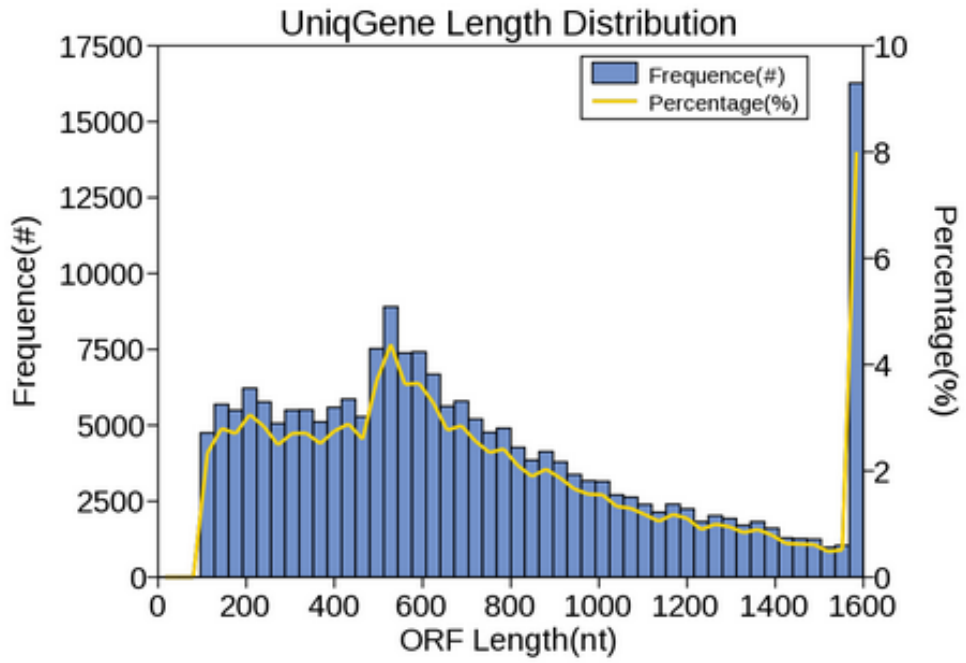


Figure 3.3.2 Gene catalogue length distribution

Plotted by the number of the genes along the Y1-axis, percentage of genes along the Y2-axis, and the length of genes along the X-axis.

3.3.3 Core-pan gene analysis

Based on the table of genes' abundance in every sample, we can get all samples' information about gene number, rarefaction curve is plotted for the number of core-gene and pan-gene separately by sampling different amount of samples, the figure is as following:

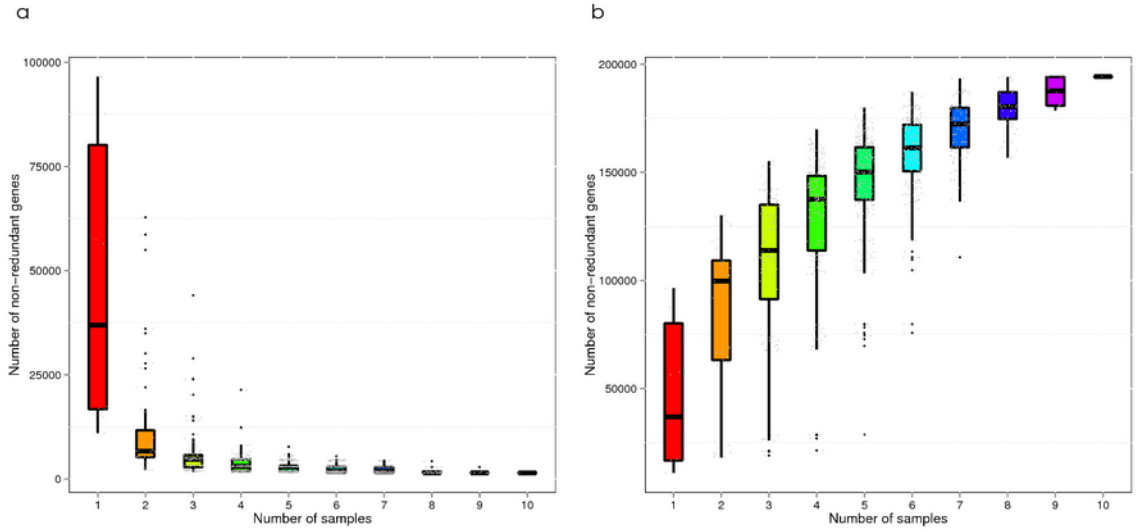


Figure 3.3.3 Core-pan gene rarefaction curve

a) Core gene rarefaction curve; b) Pan gene rarefaction curve. In the figure, X-axis means the number of samples that sampled; Y-axis means the gene number of sampled samples group

3.3.4 Gene number analysis among groups

To investigate the difference of gene number among groups, box-plot is figured for all groups' gene number, the result is as following:

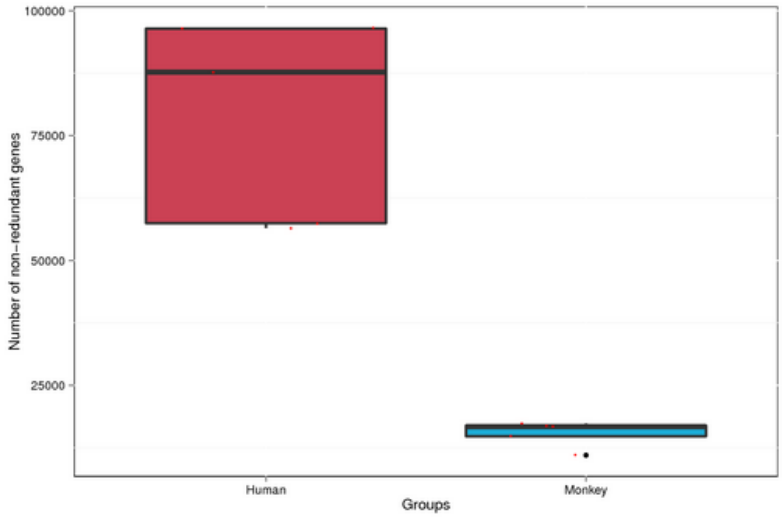


Figure 3.3.4 Box-plot for all groups' gene number

In the figure, X-axis indicates the group information; Y-axis indicates gene number

3.3.5 Venn figures analysis for gene numbers

To investigate the distribution of gene number among specified samples, and to analyze the common and peculiar information about genes, Venn figures are drawn, the result is as following:

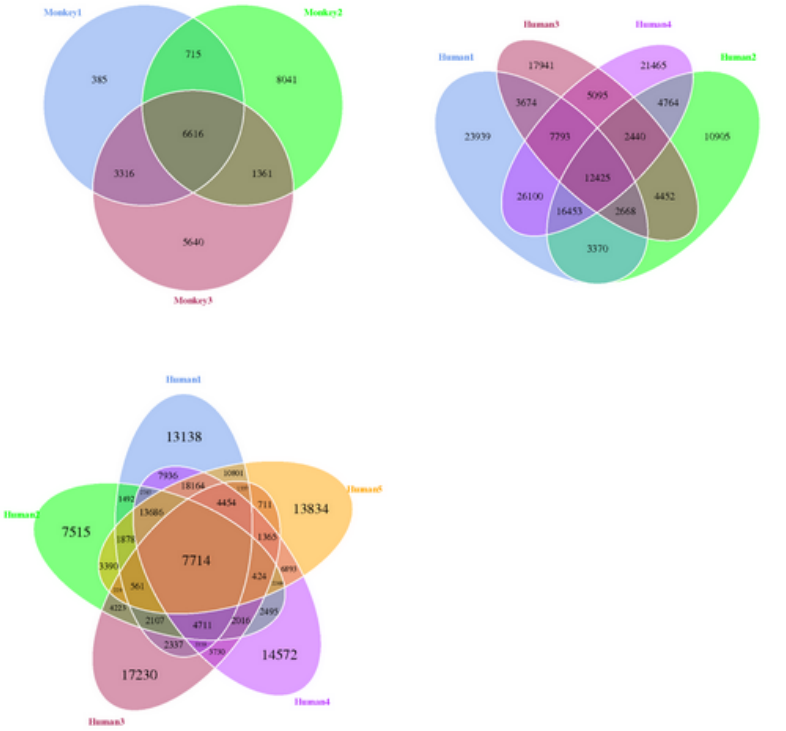


Figure 3.3.5 Venn figures for gene number among samples

In the figure, each circle reprent a sample; the overlapped part represent the number of common gene between samples; the part that don't overlap with any other circle represent the number of special gene of samples

3.3.6 The correlation analysis among samples

Repeat is essential in any biological experiment, High-throughput sequencing technology is also without exception. The correlation of gene abundance among samples is a critical index which reflects the reliability of experiment and the reasonability of the choice of samples. The closer the association coefficient is to 1, it indicate the more similar the abundance mode among samples are.

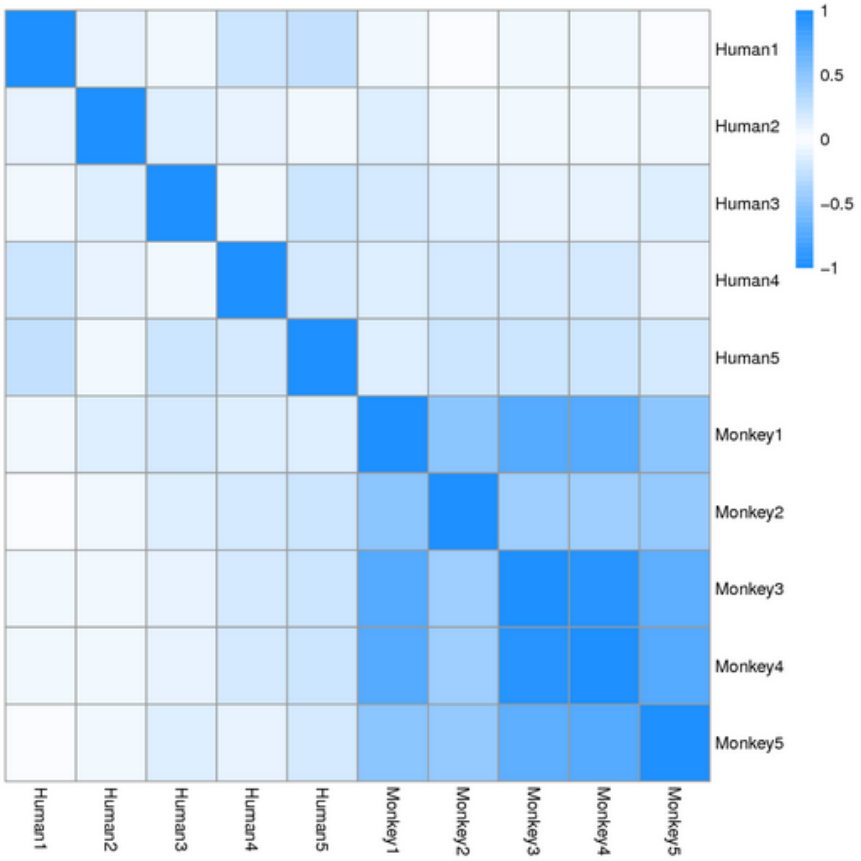


Figure 3.3.6 Heatmap for the relation among samples

In the figure, different colors represent different coefficients of association; the relationship between colors and coefficients of association is as the graphic symbol in the right; the deeper the color is the higher the coefficient of association is.

3.4 Taxonomy annotation

3.4.1 The workflow of taxonomy annotation

1) Align Unigenes with sequences of Bacteria, Fungi, Archaea, and Viruses extracted from NR database(NCBI: version 2014-10-19) using DIAMOND, a software which is 20,000 times than BLASTX, especially on short reads(E-value ≤ 1e-5).

2) Only those blast results whose E-value is less than 10 folds of the minimum E-value are selected for sequential analysis.

3) The taxonomic annotation for each Unigene is assigned using LCA(lowest common ancestor) algorithm.

4) Based on the taxonomic annotation for Unigenes and gene abundance table from the

above steps, abundance and gene numbers for each sample can be determined at each taxonomic level.

5) According to the abundance table on each taxonomic level, various analysis are performed including Krona, bar plot for abundant species, clustering heatmap according to abundance, PCA, and analysis of variation of significance (Heatmap and PCA are performed for abundant species and species with significant variation among groups, separately).

3.4.2 Krona analysis

To show the relative abundance of different level's species in different sample globally and visually, we adopt Krona to visually exhibit species' annotation result, the example of Krona is as following, for detail result please click here.

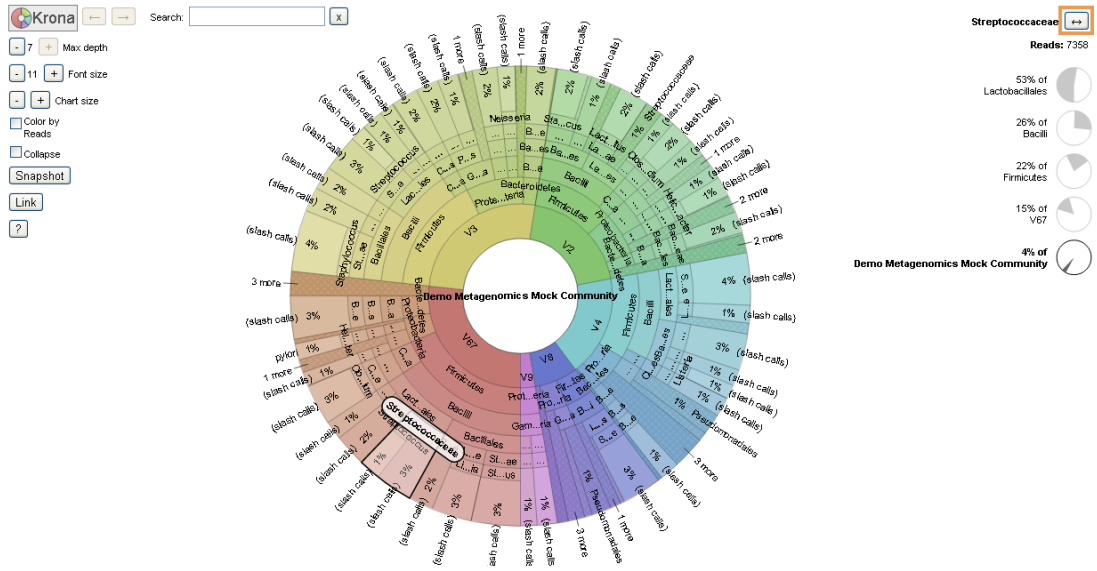


Figure 3.4.2 Exhibit species' annotation result using Krona

In the figure, circle from inside to outside represent different taxonomic level (kingdom, phylum, class, order, family, genus, species); the size of sector represent the relative proportion of different species; for more detail information please refer to detail explain of KRONA result.

3.4.3 Bar plot analysis for species' relative abundance

According to different level's relative abundance, pick out the 10 biggest relative abundance species in samples, set the other species as "Others", then draw bar plot that show the abundance in different level in species' annotation result of every sample.

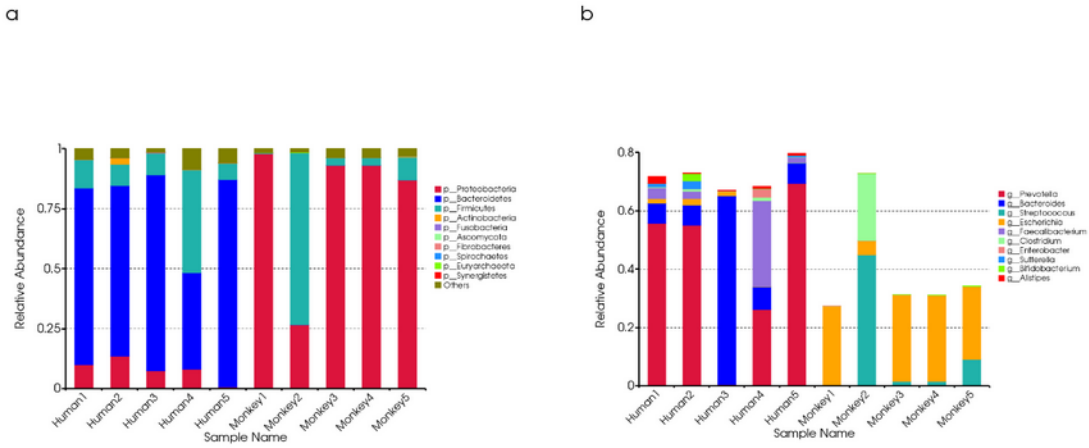


Figure 3.4.3 Species relative abundance bar plot in phylum level and genus level

a) Bar plot for relative abundance in phylum level; b) Bar plot for relative abundance in genus level; Y-axis represent relative proportion of species annotated to certain category; the relationship between colors and taxonomic level is as the graphic symbol in the right.

3.4.4 Gene number and relative abundance clustering analysis

Selecting the dominant 35 genera among all samples based on the results of relative abundance information in different taxonomic level. The abundance distribution of these dominant 35 genera is displayed in the Species abundance Heat-map in species level, to show the result clearly, and to find species that highly cluster in samples The result is shown in Figure 3.4.4.

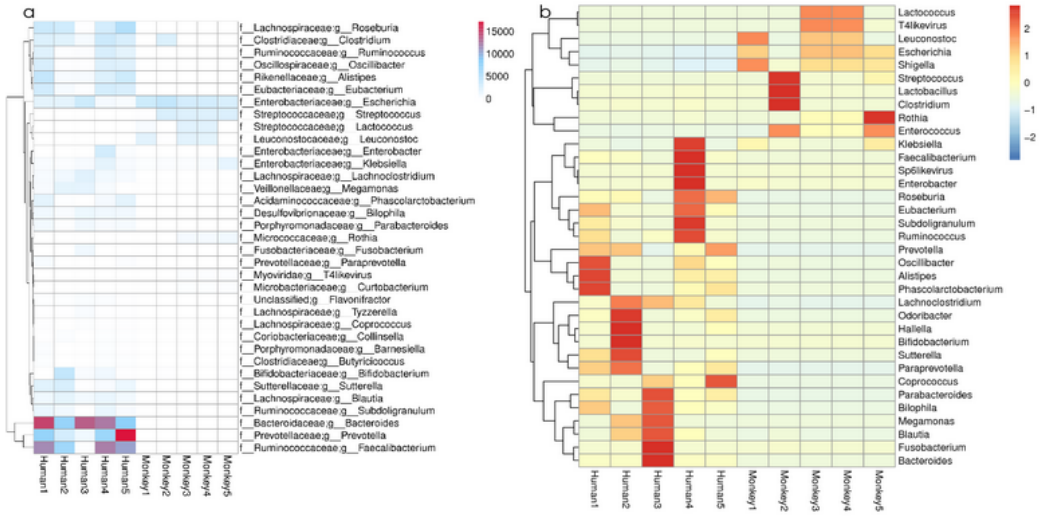


Figure 3.4.4 Gene number and abundance clustering heatmap in genus level

a) Heatmap for unigenes annotation number: X-axis indicates sample name; Y-axis indicates taxonomic information; different color represent different unigene number; b) abundance clustering heatmap in genus level: X-axis indicates sample name; Y-axis indicates taxonomic information; the clustering tree at the right of the figure is about species; he absolute value of "Z" represents the distance between the raw score and the population mean in units of the standard deviation. "Z" is negative when the raw score is below the mean, positive when above

3.4.5 Principal component analysis (PCA)

Principal component analysis (PCA) is a statistical procedure, using an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The first principal component accounts for the variability in the data as much as possible, and each succeeding component accounts for the remaining variability as much as possible. For the community composition of the samples, the more similar they are, the closer sample points in the PCA figure could be get. PCA result on the relative abundance of phylum level is displayed in Figure 3.4.5.

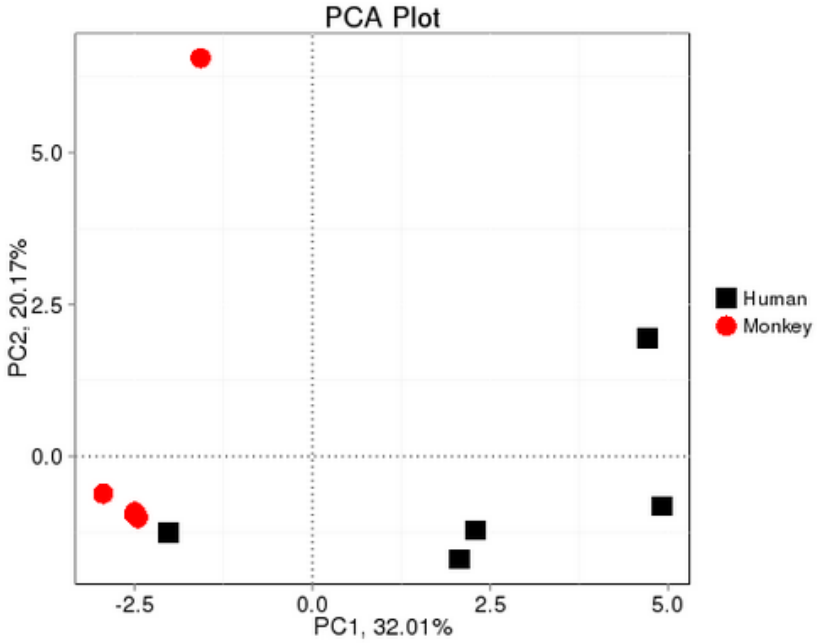


Figure 3.4.5 Species PCA result based on phylum level

In figure 3.4.5 X-axis is the first principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Y-axis is the second principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Each data point in the graph stands for a sample. Samples belong to the same group that are in the same color.

3.4.6 Sample clustering analysis

In order to study the similarity between different samples, we can also construct clustering tree by doing clustering analysis for samples. Bray-Curtis distance is the most widely used distance index in hierarchical clustering method, it mostly used to indicate the similarity between samples, and it is the major basis for sample clustering.

According to genes' relative abundance in samples, we conduct clustering analysis among samples based on Bray-Curtis distance, and combine the clustering result and relative abundance of different sample in different level to exhibit.

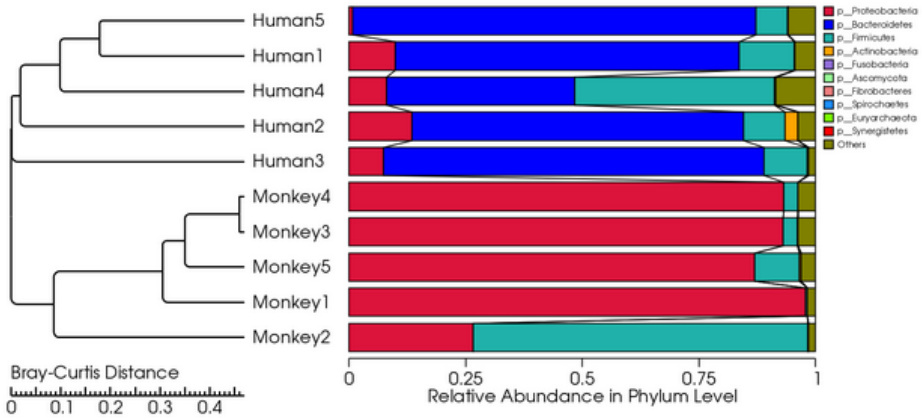


Figure 3.4.6 Clustering plot based on Bray-Curtis distance

In the figure, the left is clustering tree and the right is the relative phylum-level abundance map.

3.4.7 Significant variation analysis

Taxonomies whose abundance with significant variation among groups are detected via Metastat, a statistical method which performs hypothesis test on data of taxonomic abundance to calculate p-value. The p-value is further corrected as q-value to discover taxonomies with significant variation, and the box plot for abundance of species in groups is drawn, the result is as figure 3.4.7.1.

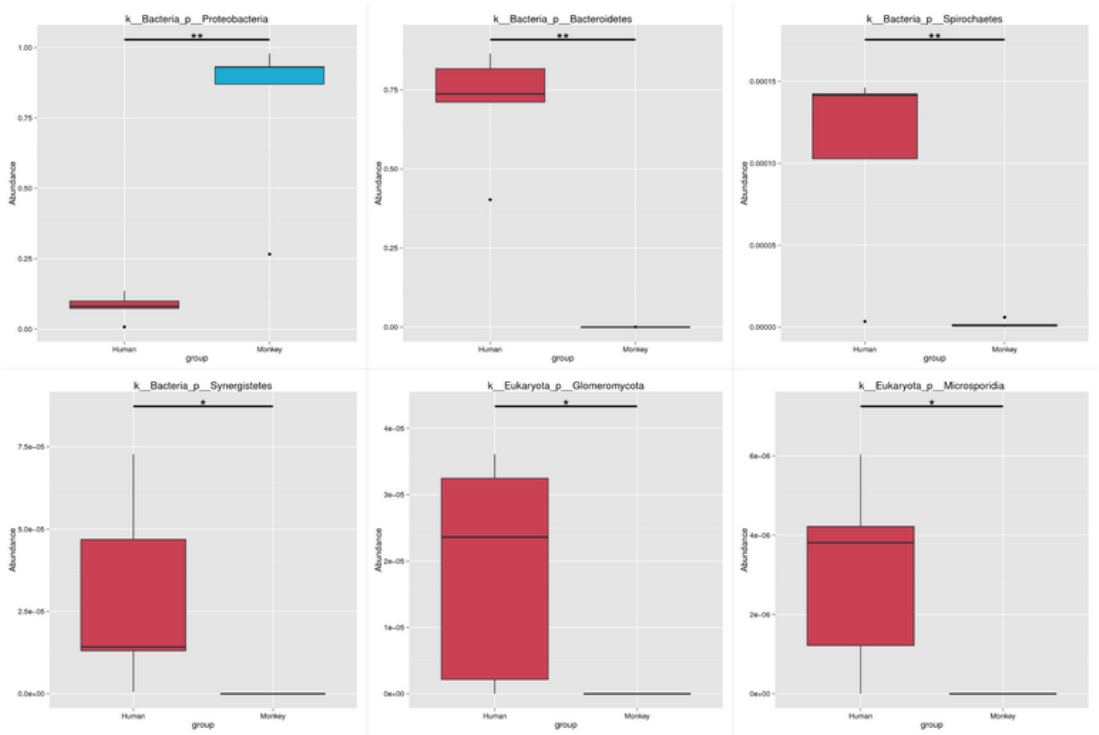


Figure 3.4.7.1 Box plot for relative abundance of significant variant species in phylum level

In the figure, X-axis indicates the group of samples; Y-axis indicates the relative abundance of corresponding species. X-axis have two significant variant groups, no group means this species having no significant variation in the two groups. "*" means the variation between the two groups is significant (q value <0.05), "**" means the variation between the two groups is very significant (q value <0.01).

Do PCA analysis and abundance clustering heatmap analysis for species which have variation between groups, the result is show in figure 3.4.7.2.

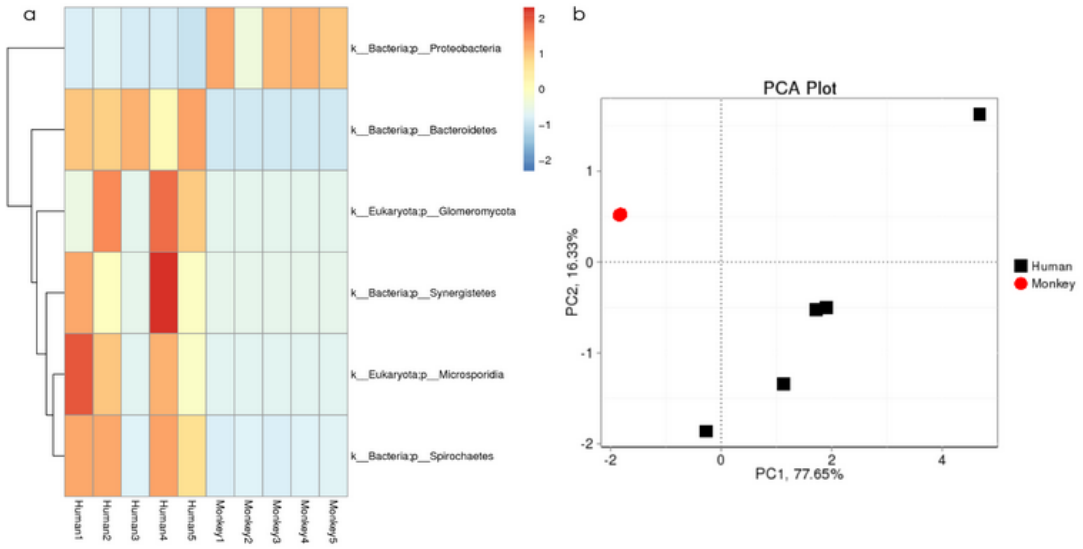


Figure 3.4.7.2 PCA analysis and abundance clustering heatmap based on significant variant species

a) The abundance clustering heatmap for significant variant species: X-axis represent the sample information; Y-axis represent species annotation information; the left of the figure is clustering tree of species; the value of the heatmap in the middle is Z value represent the relative abundance which is standardized; b) the PCA plot for significant variant species: each point represents a sample, plotted by the second principal component on the Y-axis and the first principal component on the X-axis, which was colored by group.

3.5 Function Annotation

The databases which provide annotation at present are mainly as following:
Kyoto Encyclopedia of Genes and Genomes (KEGG)[32,33];
Evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG)[34]; Version：4.1；
Carbohydrate-Active enzymes Database (CAZy)[35]; Version: 2014.11.25；

KEGG database is a comprehensive database，whose core are KEGG PATHWAY and KEGG ORTHOLOGY. KEGG PATHWAY sorts all pathways into 6 classes as: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism, Organismal Systems. Considering its important role in studies of functional genomics, KEGG database is indisapensable in Metagenomics analysis.

EggNOG database conduct function annotation by using the Orthologous Groups constructed by Smith-Waterman blast algorithm, eggnog V4.1 contains 2,031 species' gene, and has constructed 190,000 Orthologous Groups.

CAZy database is special database that study carbohydrase, mostly contain 6 kinds of functional carbohydrase: Glycoside Hydrolases(GHs), Glycosyl Transferases(GTs), Polysaccharide Lyases(PLs), Carbohydrate Esterases(CEs), Auxiliary Activities(AAs), Carbohydrate-Binding Modules(CBMs).

3.5.1 The workflow of function annotation

1) Searching Unigenes in each functional database(KEGG,eggNOG,CAZy)using DIAMOND software(blastp, E-value ≤ 1e-5)

2) Only the result with the highest score is used in sequential analysis;

3) Relative abundance and gene number are calculated at each taxonomic level based on above results;

4) According to the abundance table on each taxonomic level, various analysis are performed including annotated gene number analysis, bar plot of abundance, clustering heatmap based on abundance, PCA analysis, taxonomic information for orthologous groups, analysis of functional genes' variation of significance among groups, and pathway analysis

Database	Level	Level description
KEGG	level1	KEGG metabolic pathway level1, 6 main metabolic pathway;
KEGG	level2	KEGG metabolic pathway level2 43 seeded pathway;
KEGG	level3	KEGG pathway id (for example: ko00010);
KEGG	ko	KEGG ortholog group (for example: K00010);
KEGG	ec	KEGG EC Number (for example: EC 3.4.1.1);
eggNOG	level1	24 kinds of main functions;
eggNOG	level2	ortholog group description;
eggNOG	og	ortholog group ID (for example: ENOG410YU5S);
CAZy	level1	6 kinds of main functions;
CAZy	level2	CAZy family (for example: GT51);
CAZy	level3	EC number (for example: murein polymerase (EC 2.4.1.129));

3.5.2 Unigenes annotation number analysis

Based on Unigenes annotation results, draw summarization chart for the gene number annotated by every database, the result showed as following:

Figure 3.5.2.1 summarization chart for the gene number annotated by every database

Bar plot for annotation number of unigenes, top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database. Figure on bars are annotation numbers of unigenes, another axis stands for the codes of each function in level1 in all database, the explain of codes is show as the graphic symbol.

Based on annotated gene number of different level, draw heatmap for annotated gene number of different level from each database, the result showed as below:

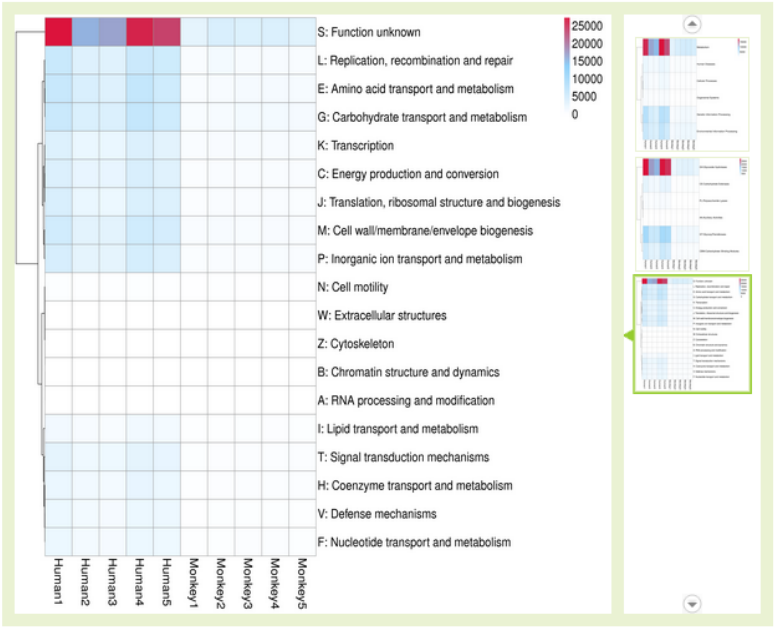


Figure 3.5.2.2 Heatmap for gene number in level1

Heatmap for annotation number of unigenes, top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database. X-axis stands for sample name, Y-axis stands for description of each level; different color represent different number of unigenes.

3.5.3 Function relative abundances Bar plot analysis

According to the relative abundances on level1 of each database, draw the abundance statistical chart in level1 for each sample.

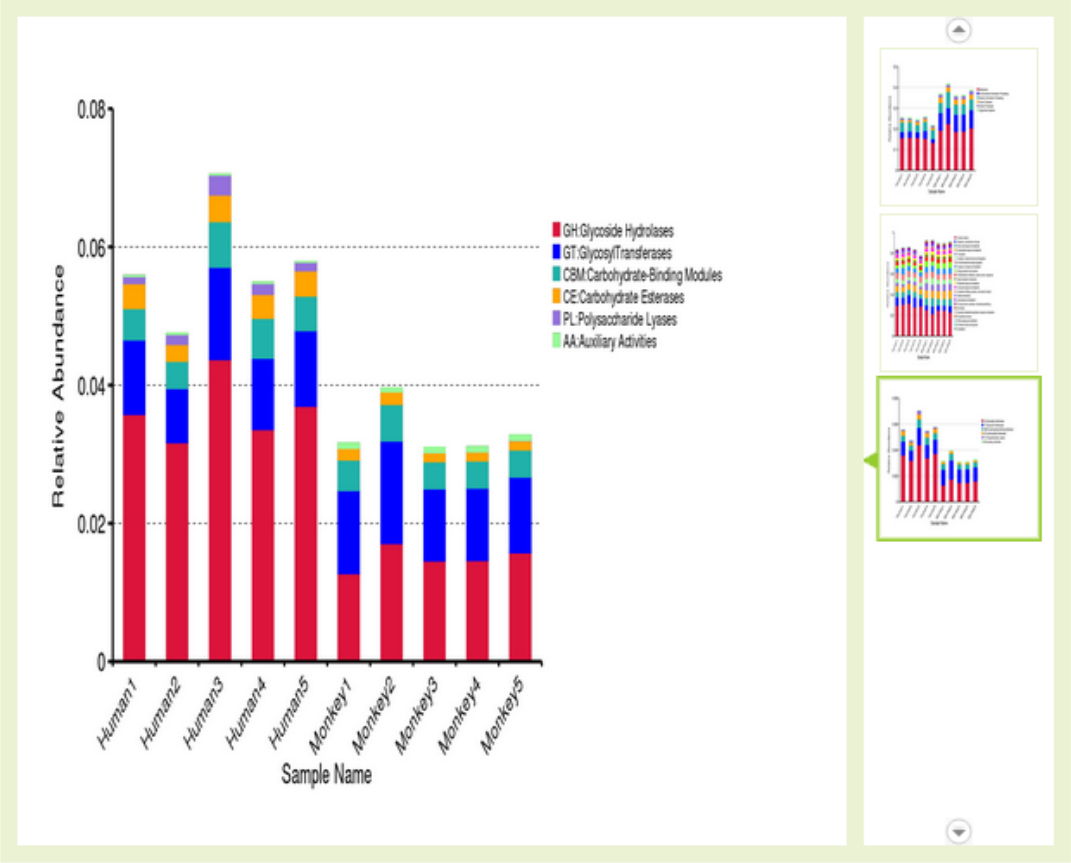
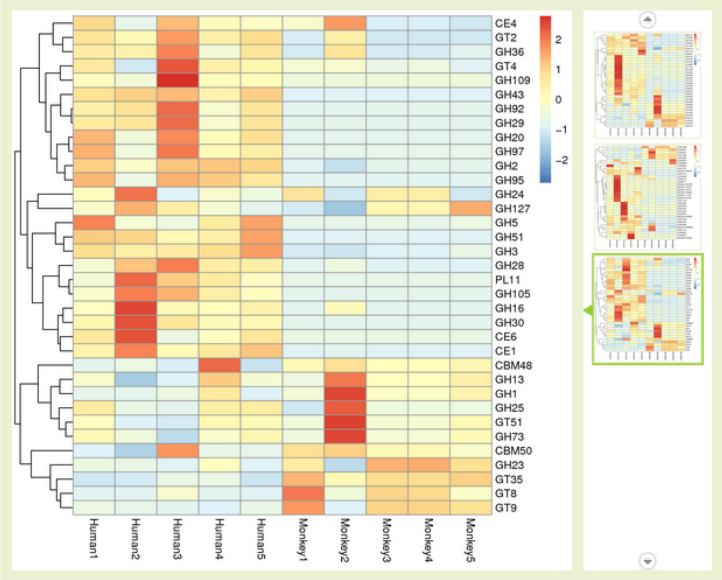


Figure 3.5.3 Abundance statistical chart in level1 in function annotation

Top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database. Y-axis stands for the relative percentage annotated to certain function; X-axis stands for samples; different color represents different function, which is show as the graphic symbol in right side.

3.5.4 Function abundance clustering analysis

According to all the samples' annotation information and abundance information in each database, select functions whose abundance is in the top 35 as well as every sample's abundance imformation to draw heatmap, and clustering from the aspect function variation.



Figue 3.5.4 Clustering heatmap for function abundance

Plotted by sample name on X-axis and genes on Y-axis. The absolute value of "Z" represents the distance between the raw score and the population mean in units of the standard deviation. "Z" is negative when the raw score is below the mean, positive when above.

3.5.5 Function principal component analysis(PCA)

Base on the function abundance table of different level, we conduct PCA analysis, the more similar the function compose of the sample are, the closer the their distance in PCA figure is.

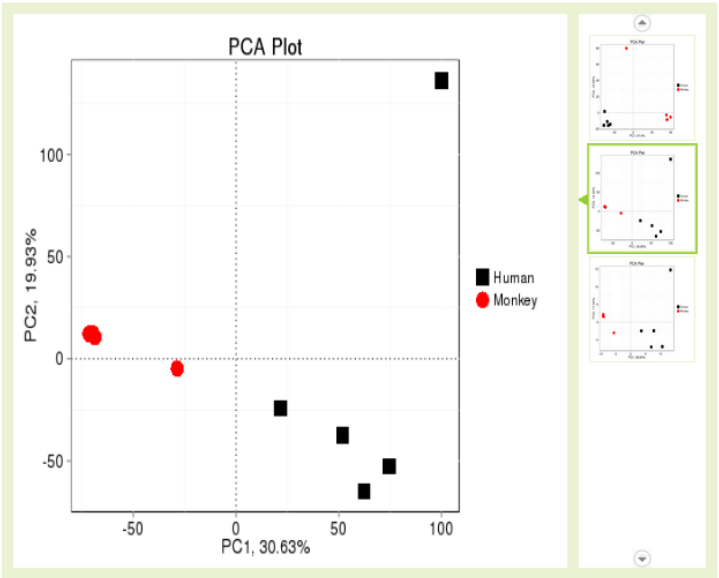


Figure 3.5.5 PCA analysis result based on function abundance

Top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database. X-axis is the first principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Y-axis is the second principle component, the percentage stands for the contribution of the first principle component to the variation in samples. Each data point in the graph stands for a sample. Samples belong to the same group that are in the same color.

3.5.6 Sample clustering analysis

In order to study the similarity of different sample, we can also construct clustering tree by making clustering analysis for samples. Bray-Curtis distance is the most widely used distance index in hierarchical clustering method, it mostly used to indicate the similarity between samples, and it is the major basis for sample clustering.

According to genes' relative abundance in samples, conduct clustering analysis among samples based on Bray-Curtis distance array, and combine the clustering result and relative abundance of different sample' function in level1 of each database to exhibit.

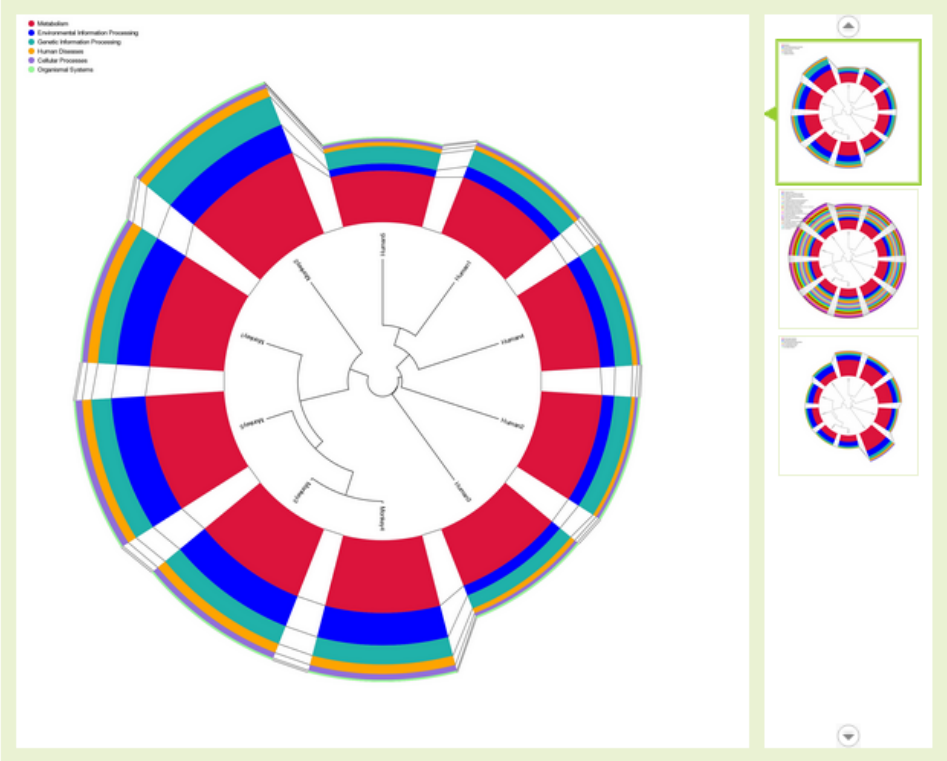


Figure 3.5.6 Clustering plot based on Bray-Curtis distance

Plotted with clustering tree in the center and the functional genes relative abundance from top level of three databases in the outer ring. Top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database.

3.5.7 Species affiliation analysis

In eggNOG database, every OG has it's detail species affiliation, according to the annotation result of level2(OG level), draw affiliation distribution circle chart of different OG species, the result is as following, for high definition chart please click here:

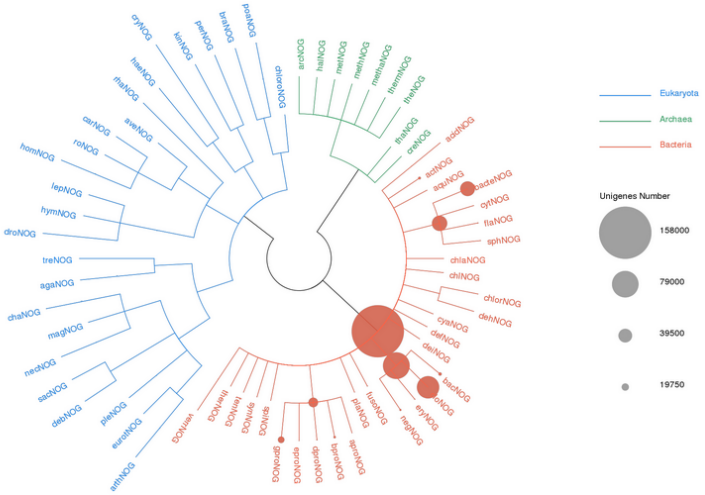


Figure 3.5.7 Species distribution circle chart of OG

Circle from inside to outside represent different taxonomic level (kingdom, phylum, class, order, family, genus, species); eucaryon, bacteria and archaea are marked with different color; every node represent the OG of a broad category of species; the size of the circle is positive correlate with gene number, the relation between the size of circle and the number of unigenes is show as the graphic symbol in the right side.

3.5.8 Function significant variation analysis

Function whose abundance with significant variation among groups are detected via Metastat, a statistical method which performs hypothesis test on data of function abundance to calculate p-value. The p-value is further corrected as q-value to discover function with significant variation, which are utilized for successive demonstration such as box plot.

Figure 3.5.8.1 Box plot for significant variant function

Top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database. X-axis indicates the group of samples; Y-axis indicates the relative abundance of corresponding function. X-axis have two significant variant groups, no group means this function having no significant variation in the two groups. "*" means the variation between the two groups is significant (q value <0.05), "**" means the variation between the two groups is very significant (q value <0.01).

Based on function with significant variation, conduct clustering heatmap and PCA:

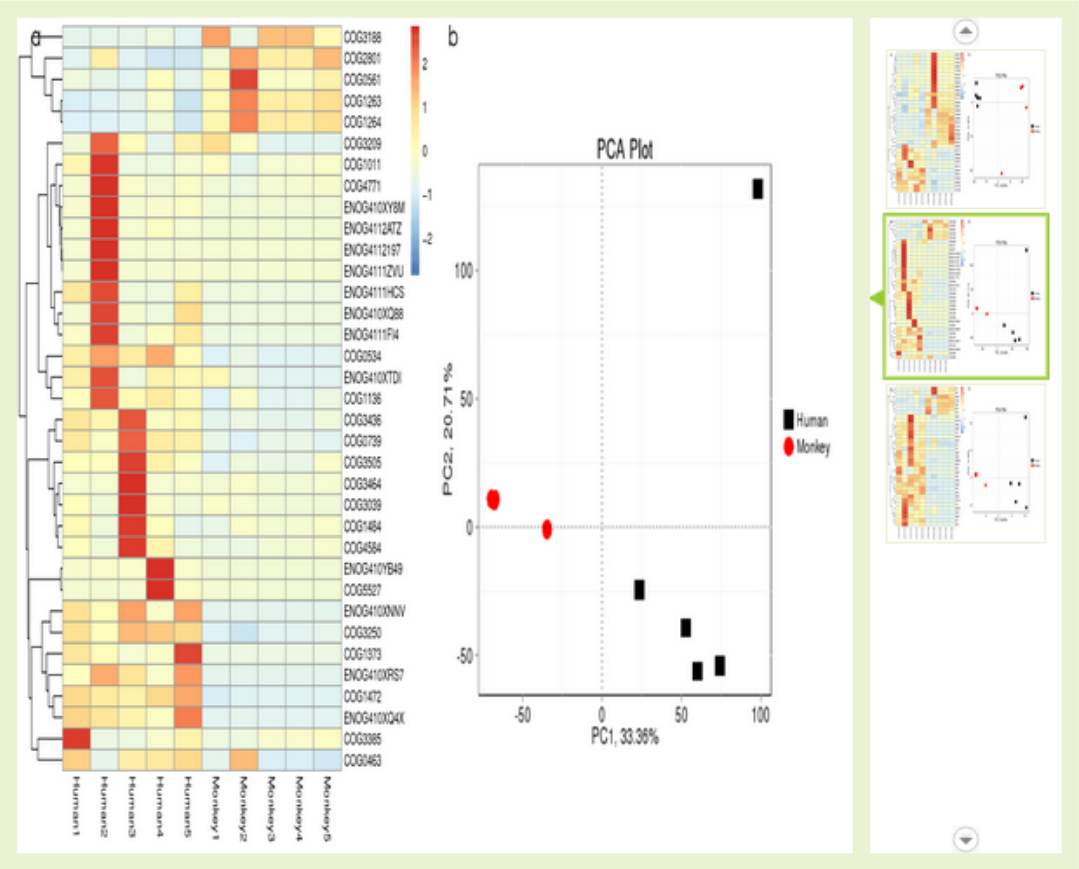


Figure 3.5.8.2 Clustering heatmap and PCA analysis based on function with significant variation

Top panel: KEGG database. Middle panel: eggNOG database. Low panel: CAZy database. a) The abundance clustering heatmap for significant variant function: X-axis represent the sample information; Y-axis represent function annotation information; the left of the figure is clustering tree of function; the value of the heatmap in the middle is Z value represent the function's relative abundance which is standardized; b) the PCA plot for significant variant function: each point represents a sample, plotted by the second principal component on the Y-axis and the first principal component on the X-axis, which was colored by group.

3.5.9 Evolution analysis on species of variant OG

To study the evolution of species with significant variant OG(ortholog group) in eggNOG database, for every significant variant OG, refer to the data from eggNOG database, we present it's species evolution tree.

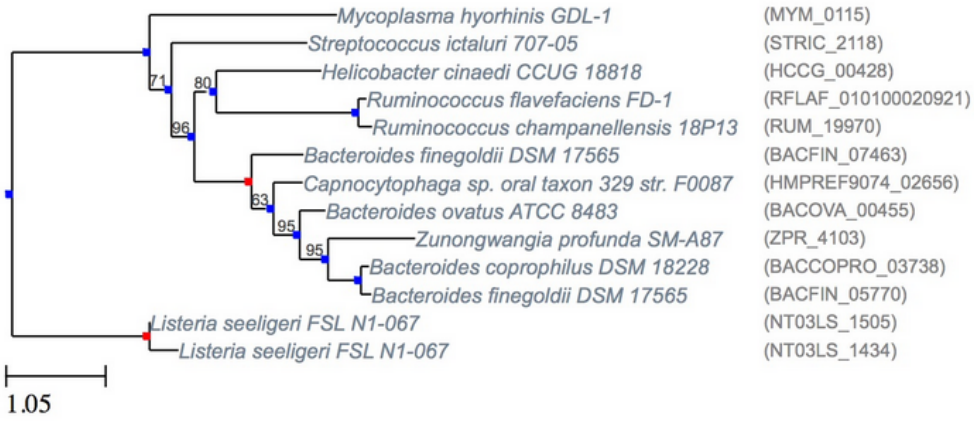


Figure 3.5.9 Plot species evolution tree

3.5.10 Metabolic Pathway Analysis

To study the variances of pathway patterns in different groups (samples), the web-version pathway figure was drawn. The whole report has two parts:

Part 1: pathway overview, showing shared and identical pathway information among groups or samples. In this metabolic pathway plot, nodes represent chemicals, and edges represent enzymatic reactions, of which the shared ones were marked in red, while the identical ones were marked in blue(group A/sample A) or green(group B/sample B);

Part 2: the annotated metabolic pathway plot. In this metabolic pathway plot, nodes represent chemicals, and frames represent enzymatic information (as default, black edges and white backspace). Different colors of the frames shows the Unigene number in this annotation， within which enzymes on yellow backspace are those significantly variant among groups(void without variance analysis). Touching your mouse on the enzymes, the abundance boxplot of variant enzymes will be displayed.

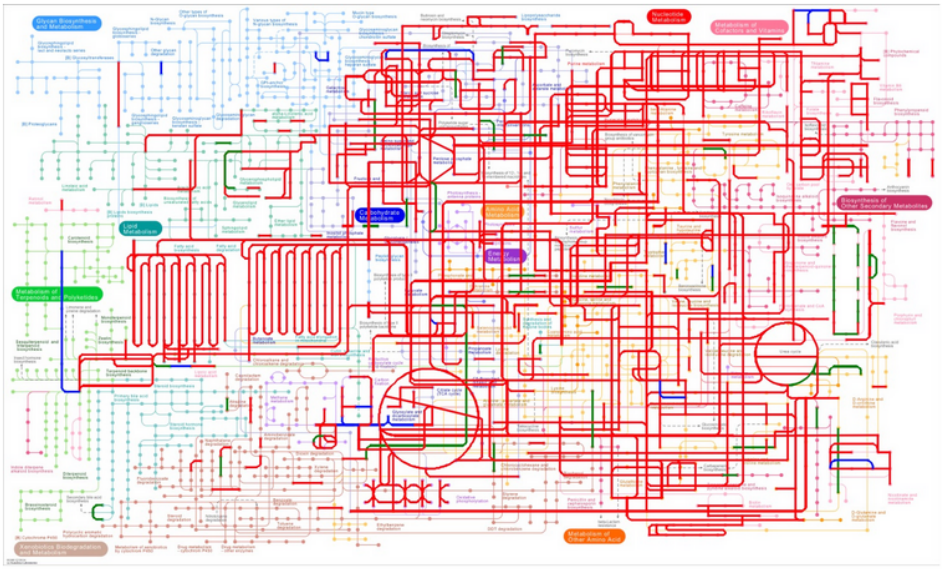


Figure 3.5.10 The compare analysis plot for Metabolic Pathway of multi-samples

4 Methods

4.1 Library construct and sequencing

First, samples are collected from the environment (like soil, ocean, fresh water, gut etc), and original samples or extracted DNA samples are delivered to our company under low temperature (below 0℃). Then we will conduct sample qualification.

For those qualified DNA samples, we will conduct library construction and qualification, and use those qualified libraries for sequencing with the Illumina HiSeq high-throughput sequencing platform. Then bioinformatics analysis will be carried on sequencing data. To ensure accuracy and reliability of the sequencing data form the source, 1st BASE control every aspect of sample testing, library construction and sequencing strictly, ensure high quality data from original, the experimental workflow is as following:

4.1.1 DNA sample quality control

Three key methods in DNA samples QC:
DNA integrity degree and purity are monitored by Agarose Gel Electrophoresis.
DNA purity (OD260/280) is checked using Nanodrop.
DNA concentration is accurately measured using Qubit.

4.1.2 Library Construction and Qualification

Qualified DNA samples are randomly fragmented to about 300bp by Covaris ultrasonic broken instrument, then fragments are end-polished, A-tailed, ligated with sequencing adaptor, and purified with further PCR amplification, sequencing libraries are generated through these procedures.

After constructed, those libraries will be primarily quantified by Qubit2.0, and diluted to 2ng/ul, then we will detect the library's insert size using Agilent 2100, if the insert size meet expectation, we will accurately quantify the effective concentration of the library using Q-PCR(effective concentration of the library>3nM) to ensure the library's quality.

4.1.3 Sequencing

After the library is qualified, pooling different libraries depend on effective concentration and aimed data size, then conduct Illumina HiSeq sequencing.

4.2 Bioinformatics analyze pipeline

a) Data Pre-processing: Raw data from sequencing exist a certain percentage of low quality data, to ensure the accuracy and the reliability of following information analysis, the initial step of metagenomic data analysis requires the execution of certain pre-

filtering, and the clean data is obtained;
b) Metagenomic assembly: Conduct metagenomic assembly based on clean data, and find information about low abundance species by putting in each sample's reads that have not been utilized for mixed assembly
c) Gene prediction: According to scaftigs from single sample assembly and mix assembly, predict genes using MetaGeneMark, then put all the samples and the predicted genes together to omit the redundant and construct gene catalogue, according to gene catalogue, together with every sample's clean data, we can obtain the abundance information about gene catalogue in every sample.
d) Species annotation: Based on gene catalogue, blast with MicroNR database, get the species annotation information of Unigenes, and combine with the gene abundance table, obtain species abundance table in different levels.
e) Function annotation: Based on gene catalogue, conduct Metabolic Pathway Analysis (KEGG), homologous gene group analysis (eggNOG), carbohydrase (CAZy) function annotation and abundance analysis;
f) Based on the species abundance table and the function abundance table, we can conduct abundance clustering analysis, PCA analysis, sample clustering analysis, significant variation analysis, metabolic pathway compare analysis, to discover the component difference of species and functional component difference between samples; at the same time, based on the standard analysis result, we can conduct a series of advanced analysis(like LEfSe analysis, significant variation analysis of community constitute between groups, CCA/RDA analysis, NMDS(Non-metric Multidimensional Scaling) analysis, ect), and combine with the environment factor, pathological indicators and special phenotype to study deep correlation, which can provide theoretical foundation for further study and use of samples' species and function.

Notes:
When the biological repetition in group is less than 3, statistical analyses like
PCA, LEfSe, CCA/RDA and so on have no statistical meaning, the result is just for reference.
When the number of samples is less than 3, in standard analysis, PCA analysis, clustering analysis, abundance clustering heatmap analysis can't be conducted.

5 References

[1] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology, 5(10), R245-R249.
[2] Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS computational biology, 1(2), e24.
[3] Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. Nature reviews genetics, 6(11), 805-814.
[4] Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H.

W., ... & Rubin, E. M. (2005). Comparative metagenomics of microbial communities. Science, 308(5721), 554-557.
[5] Raes, J., Foerstner, K. U., & Bork, P. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. Current opinion in microbiology, 10(5), 490-498.
[6] Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18.
[7] Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis[J]. Nature, 2014.
[8] Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence[J]. Nature communications, 2015, 6.
[9] Scher J U, Sczesnak A, Longman R S, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis[J]. Elife, 2013, 2: e01202.
[10] Brum J R, Ignacio-Espinoza J C, Roux S, et al. Patterns and ecological drivers of ocean viral communities[J]. Science, 2015, 348(6237): 1261498.
[11] Mende D R, Waller A S, Sunagawa S, et al. Assessment of metagenomic assembly using simulated next generation sequencing data[J]. PloS one, 2012, 7(2): e31386.
[12] Nielsen H B, Almeida M, Juncker A S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes[J]. Nature biotechnology, 2014, 32(8): 822-828.
[13] Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J, Backhed F: Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013, 498(7452):99-103.
[14] Karlsson F H, Fåk F, Nookaew I, et al. Symptomatic atherosclerosis is associated with an altered gut metagenome[J]. Nature communications, 2012, 3: 1245.
[15] Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing[J]. nature, 2010, 464(7285): 59-65.
[16] Zeller G, Tap J, Voigt A Y, et al. Potential of fecal microbiota for early - stage detection of colorectal cancer[J]. Molecular systems biology, 2014, 10(11): 766.
[17] Sunagawa S, Coelho L P, Chaffron S, et al. Structure and function of the global ocean microbiome[J]. Science, 2015, 348(6237): 1261359.
[18] Li J, Jia H, Cai X, et al. An integrated catalog of reference genes in the human gut microbiome[J]. Nature biotechnology, 2014, 32(8): 834-841.
[19] Oh J, Byrd A L, Deming C, et al. Biogeography and individuality shape function in the human skin metagenome[J]. Nature, 2014, 514(7520): 59-64.
[20] Zhu, Wenhan, Alexandre Lomsadze, and Mark Borodovsky. "Ab initio gene identification in metagenomic sequences." Nucleic acids research 38.12 (2010): e132-e132
[21] Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22(13):1658-1659.
[22] Fu L, Niu B, Zhu Z, Wu S, Li W: CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012, 28(23):3150-3152.
[23] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota

in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60.
[24] Villar E, Farrant G K, Follows M, et al. Environmental characteristics of Agulhas rings affect interocean plankton transport[J]. Science, 2015, 348(6237): 1261447.
[25] Cotillard A, Kennedy S P, Kong L C, et al. Dietary intervention impact on gut microbial gene richness[J]. Nature, 2013, 500(7464): 585-588.
[26] Le Chatelier E, Nielsen T, Qin J, et al. Richness of human gut microbiome correlates with metabolic markers[J]. Nature, 2013, 500(7464): 541-546.
[27] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12:59-60.
[28] Huson, Daniel H., et al. "Integrative analysis of environmental sequences using MEGAN4." Genome research 21.9 (2011): 1552-1560.
[29] Ondov B D, Bergman N H, Phillippy A M. Interactive metagenomic visualization in a Web browser[J]. BMC bioinformatics, 2011, 12(1): 385.
[30] Avershina, Ekaterina, Trine Frisli, and Knut Rudi. De novo Semi-alignment of 16S rRNA Gene Sequences for Deep Phylogenetic Characterization of Next Generation Sequencing Data. Microbes and Environments 28.2 (2013): 211-216.
[31] White J R, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples[J]. PLoS Comput Biol, 2009, 5(4): e1000352.
[32] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34(Database issue): D354–7.
[33] 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).
[34] Powell S, Forslund K, Szklarczyk D, et al. eggNOG v4. 0: nested orthology inference across 3686 organisms[J]. Nucleic acids research, 2013: gkt1253.
[35] 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.
[36] Bäckhed F, Roswall J, Peng Y, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life[J]. Cell host & microbe, 2015, 17(5): 690-703.

6 Documents

1 Links of related software:
SOAP denovo(Version 2.21): http://soap.genomics.org.cn/soapdenovo.html
SoapAligner(Version 2.21): http://soap.genomics.org.cn/soapaligner.html
MetaGeneMark(Version 2.10): http://exon.gatech.edu/GeneMark/metagenome/Prediction
CD-HIT(Version 4.5.8): http://www.bioinformatics.org/cd-hit/
2 Notes:

Files of result are recommended to open with Excel or specialized text editor like EditPlus.