GENEWIZ NGS DATA REPORT

1. Project Summary

Project ID: 1
Project Name: 1
Project Description: 1

2. Description of Workflow

2.1. Experimental Design

100 ng/ul DNA in volume is composed of concentrated and high-molecular integrity. Whereas concentrated signals are not significantly different across samples, high-molecular integrity varies significantly across samples. Samples with low-molecular integrity are expected to have degraded DNA. Samples with degraded DNA may not be suitable for some downstream applications. Samples with degraded DNA are still acceptable for library preparation and sequencing. Samples with degraded DNA are acceptable for sequencing but results are interpreted with caution. Samples with degraded DNA are acceptable for library preparation and sequencing but results are interpreted with caution.

Illumina Genome Analyzer Hiseq 2500 platform was used to generate 150 bp paired end reads. Adapter and quality trimmed reads from FASTQ files were used for downstream analysis. Quality trimmed reads were aligned to the reference genome using Bowtie2 (Langmead and Salzberg, 2012). Alignment files were processed using Samtools (Li et al., 2009). Gene hit counts were calculated using featureCounts (Liao et al., 2013). Transcripts per million (TPM) and reads per kilobase of transcript per million mapped reads (RPKM) were calculated in-house. Differential gene expression analysis was performed using DESeq2 (Love et al., 2014). Ontology analysis was done using GeneSCF (Gene Set Functional Enrichment in Current Frameworks) (Tian et al., 2010). Pathway analysis was done using KEGG (Kanehisa et al., 2012) and Fisher's exact test. Unless otherwise specified, default parameters were used for all analysis.

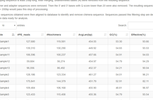


Figure 2.1 Illumina Hiseq 2500 platform workflow.

2.2. Bioinformatics Analysis Workflow

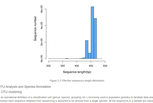


Figure 2.2 Bioinformatics Analysis workflow.

3. Analysis

3.1. Sequencing Quality Check

Sequencing reads length is a major factor that determines the information content of the sequence. Short reads have less of the sequence reads length than longer reads. Sequencing reads length is also a factor in mapping reads to the genome. Longer reads are easier to map than shorter reads. Sequencing reads length quality is also a factor.

Sequencing reads length and quality are measured in FASTQ format (Cock et al., 2009).

Plots are presented for samples:

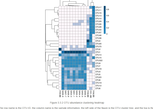
Sample1 Sample2 Sample3 Sample4 Sample5 Sample6 Sample7 Sample8

Sequencing reads length is a major factor that determines the information content of the sequence. Short reads have less of the sequence reads than longer reads. Sequencing reads length is also a factor in mapping reads to the genome. Longer reads are easier to map than shorter reads. Sequencing reads length quality is also a factor.

Sequencing reads length and quality are measured in FASTQ format (Cock et al., 2009).

Plots are presented for samples:

Sample1 Sample2 Sample3 Sample4 Sample5 Sample6 Sample7 Sample8



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 communities.

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.

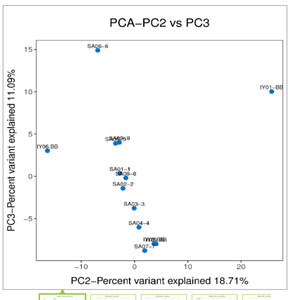


Figure 3.5.3 PCA graph

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.

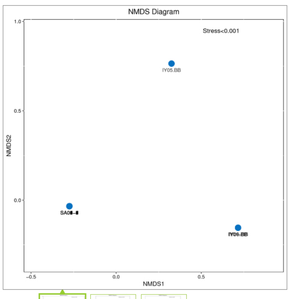


Figure 3.5.4 NMDS plot

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.

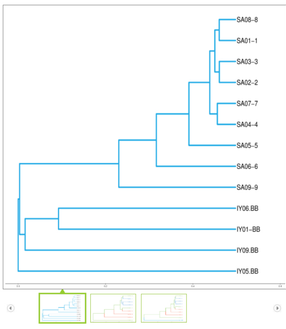


Figure 3.5.5 UPGMA tree

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

Show 10 entries Search:

Factor	R-Value	P-Value
GI	0.25	0.05
GS	0.00	0.46

Showing 1 to 2 of 2 entries

Copy CSV Excel PDF Print Previous 1 Next

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.

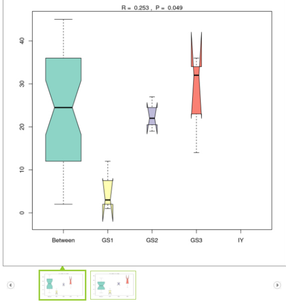


Figure 3.6.1 Group comparison by Anosim analysis

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, Glöckner 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