GENEWIZ®
Solid science. Superior service.

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

GENEWIZ MetaVx™ 2.0 Report

Client:

Quotation:

Date:

Email:

Table of Contents

1 Experimental Process

16S rRNA metagenomics is an important tool to determine the type and relative abundance of bacterial and archaeal species in heterogeneous samples, such as soil, water, and the gut microbiome. GENEWIZ has developed 16S MetaVx™ Sequencing, a proprietary assay that provides increased sensitivity and specificity in comparison to current 16S metagenomics assays. This improved performance is accomplished using a unique primer design shown to increase hybridization across a broad range of species and decrease taxonomy bias. Furthermore, primers are also designed to increase diversity within the amplicon, bypassing the need for control PhiX in the sequencing run, allocating more data to your research. 16S MetaVx™ Environmental analyzes the V3, V4, and V5 hypervariable regions of the 16S gene, whereas 16S MetaVx™ Mammalian analyzes the V3 and V4 regions.

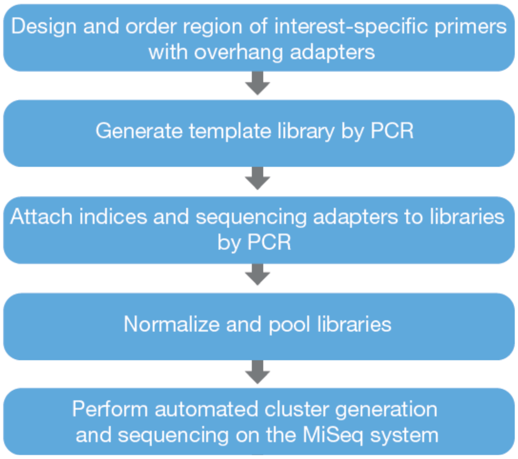

Design and order region of interest-specific primers with overhang adapters
Generate template library by PCR
Attach indices and sequencing adapters to libraries by PCR
Normalize and pool libraries
Perform automated cluster generation and sequencing on the MiSeq system

Figure 1.1 16S Metagenomics Sequencing Workflow

2 Bioinformatics Pipeline

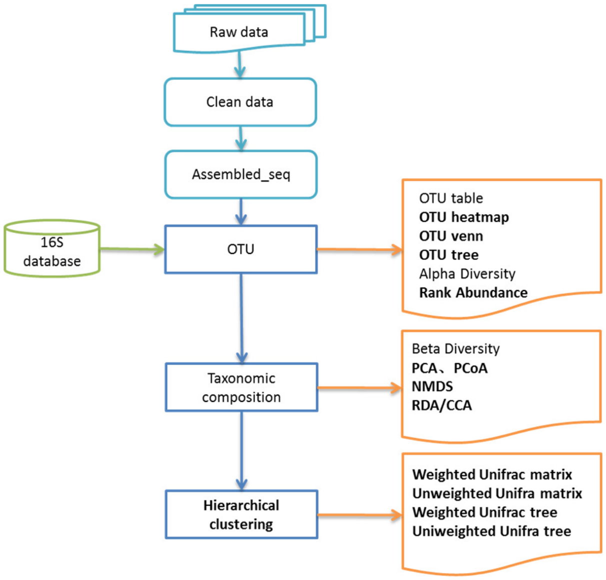


Figure 2.1 Pipeline of bioinformatics analysis.

3 Data Analysis

3.1 OTU analysis

Sequences were grouped into operational taxonomic units (OTUs) using the clustering program UCLUST, pre-clustered at 97% sequence identity, to produce an OTU table and OTU representative sequences.

Software: QIIME v1.7 (http://QIIME.org/tutorials/otu_picking.html)
Analysis methods: UCLUST method for OTU clustering, OTU of the sequence similarity is set to 97% to get the OTU list and OTU representative sequence.

Table 3.1 OTU table

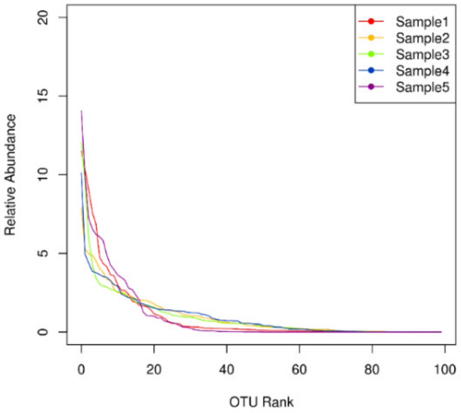
#OTU ID	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample N
OTU0	1	1	0	0	0	
OTU2	0	0	1	3	2	
OTU3	0	0	2	0	1	
OTU5	1	0	0	1	2	
OTU7	0	0	2	0	0	
OTU11	0	0	0	0	0	
OTU12	255	81	2	1	0	
OTUn						

Column name interpretation:

Column name	Description
#OTU ID	OTU number
Sample1	The abundance of OTU in sample 1 was obtained.
Sample2	The abundance of OTU in sample 2 was obtained.
…	…
SampleN	The abundance of OTU in sample N was obtained.

3.2 Rank-Abundance curve

A rank abundance curve or Whittaker plot is a chart used by ecologists to display relative species abundance, a component of biodiversity. It can also be used to visualize species richness and species evenness.



X-axis: The abundance rank. The most abundant species is given rank 1, the second most abundant is 2 and so on

Y-axis: The relative abundance. Usually measured on a log scale, this is a measure of a species abundance (e.g., the number of individuals) relative to the abundance of other species.

Figure 3.2 Rank abundance curve

3.3 Species taxonomy

The Ribosomal Database Program (RDP) classifier was used to assign taxonomic category to all OTUs at confidence threshold of 0.97. The RDP classifier uses Silva_111 16S rRNA database (http://www.arb-silva.de/) which has taxonomic categories predicted to the genus level.

Software: QIIME (http://QIIME.org/tutorials/otu_picking.html）

Table 3.3.1 Taxonomy tree file

Taxon level	rankID	Taxon	Sample1	Sample 2	Sample 3	Sample 4	Sample N	Total
Kingdom	0.1	k__Bacteria	101526	128445	108314	103809		1653735
Phylum	0.1.1	p__Deinococcus	3	768	136	1797		23257
Class	0.1.1.1	c__Deinococci	3	768	136	1797		23257
Order	0.1.1.1.1	o__Thermales	3	768	136	1797		23257
Family	0.1.1.1.1.1	f__Thermaceae	3	768	136	1797		23257
Genus	0.1.1.1.1.1.1	g__Thermus	3	768	136	1797		23257
Phylum	0.1.2	p__Nitrospirae	2674	5687	1683	3634		62539
Class	0.1.2.1	c__Nitrospira	0	0	1	0		2

Table 3.3.2 Taxa Statistics at Phylum level

Taxon	Sample 1	Sample 2	Sample 3	Sample 4	Sample N
Firmicutes	56.52	61.33	59.24	57.65	
Bacteroidetes	34.79	31.08	32.12	37.97	
Proteobacteria	4.49	4.74	4.12	2.19	
Deferribacteres	1.78	0.40	3.62	1.16	
Actinobacteria	2.19	2.22	0.72	0.84	
Verrucomicrobia	0.08	0.12	0.04	0.00	
Tenericutes	0.01	0.02	0.05	0.05	

Table 3.3.3 Statistics of Taxonomic Composition

Samples	Phylum	Class	Order	Family	Genus
Sample1	9	15	21	41	91
Sample2	9	14	21	41	87
Sample3	8	13	18	37	76
Sample4	8	15	20	38	73
SampleN					

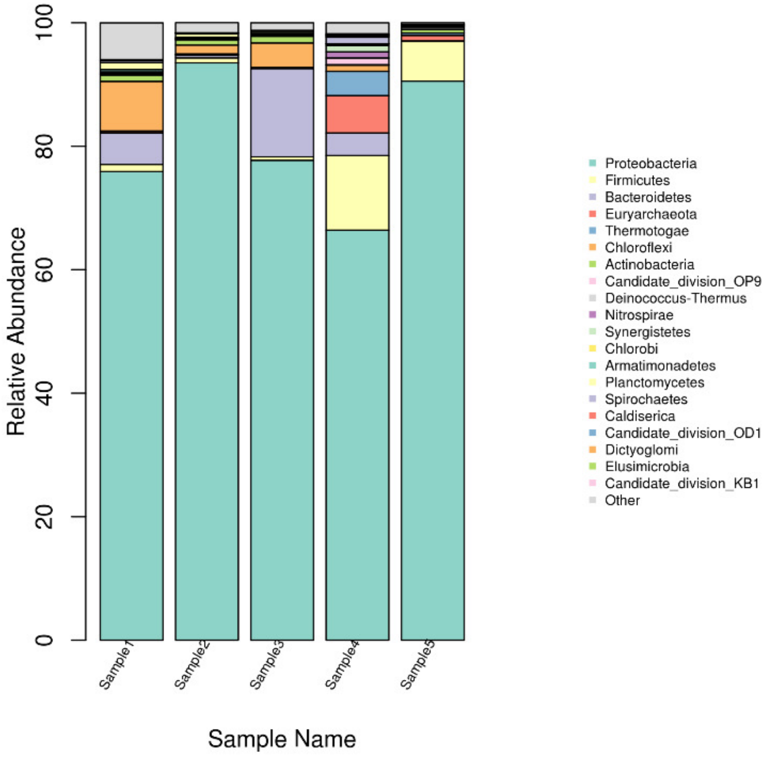


Figure 3.3 Taxa assignments at Phylum level

3.4 Rarefaction curve

Rarefaction allows the calculation of species richness for a given number of individual samples, based on the construction of so-called rarefaction curves. This curve is a plot of the number of species as a function of the number of samples.

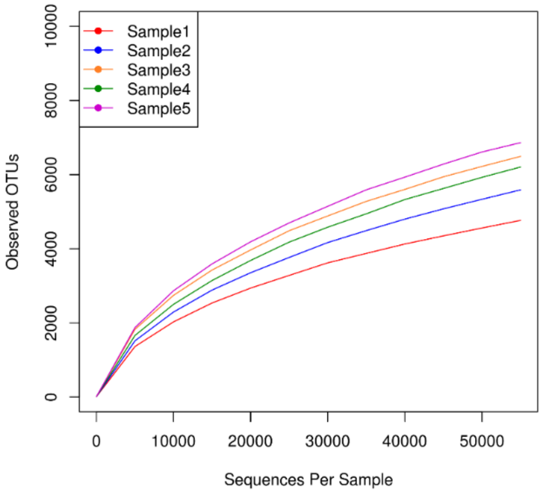


Figure 3.4 Observed OTUs rarefaction curves

3.5 Alpha diversity

Sequences were rarefied prior to calculation of alpha and beta diversity statistics. Alpha diversity indexes were calculated in QIIME from rarefied samples using for diversity the Shannon index, for richness the Chao1 index.

Software：QIIME（http://QIIME.org/tutorials/otu_picking.html）

Table 3.5 Collation of alpha diversity results

Sample	ACE	Chao1	Shannon	Simpson	Good's_coverage
Sample1	6057.815	5700.788	6.758925	0.95554	0.984373
Sample2	5868.596	5804.006	7.238077	0.968738	0.988376
…					
Sample N					

3.6 Beta diversity

Beta-diversity metrics assess the differences between microbial communities. The fundamental output of these comparisons is a square matrix where a “distance” or dissimilarity is calculated between every pair of community samples, reflecting the dissimilarity between those samples. The weighted and unweighted UniFrac matrix can be performed by Principal Coordinate Analysis (PCoA) and hierarchical clustering. Like alpha diversity, there are many possible metrics which can be calculated with the QIIME pipeline.

Software：QIIME（http://QIIME.org/tutorials/otu_picking.html）

Table 3.6.1 Weighted unifrac distance

	Sample1	Sample2	Sample3	Sample 4	Sample N
Sample1	0	0.34952	0.284133	0.45525	
Sample2	0.34952	0	0.261572	0.23688	
Sample 3	0.28413	0.26157	0	0.27705	
Sample 4	0.45525	0.23688	0.277046	0	
Sample N					

Table 3.6.2 Unweighted unifrac distance

	Sample1	Sample2	Sample3	Sample 4	Sample N
Sample1	0	0.53299	0.749778	0.73715	
Sample2	0.53299	0	0.73835	0.7125	
Sample 3	0.74978	0.73835	0	0.49223	
Sample 4	0.73715	0.7125	0.492227	0	
Sample N					

3.7 PCoA analysis

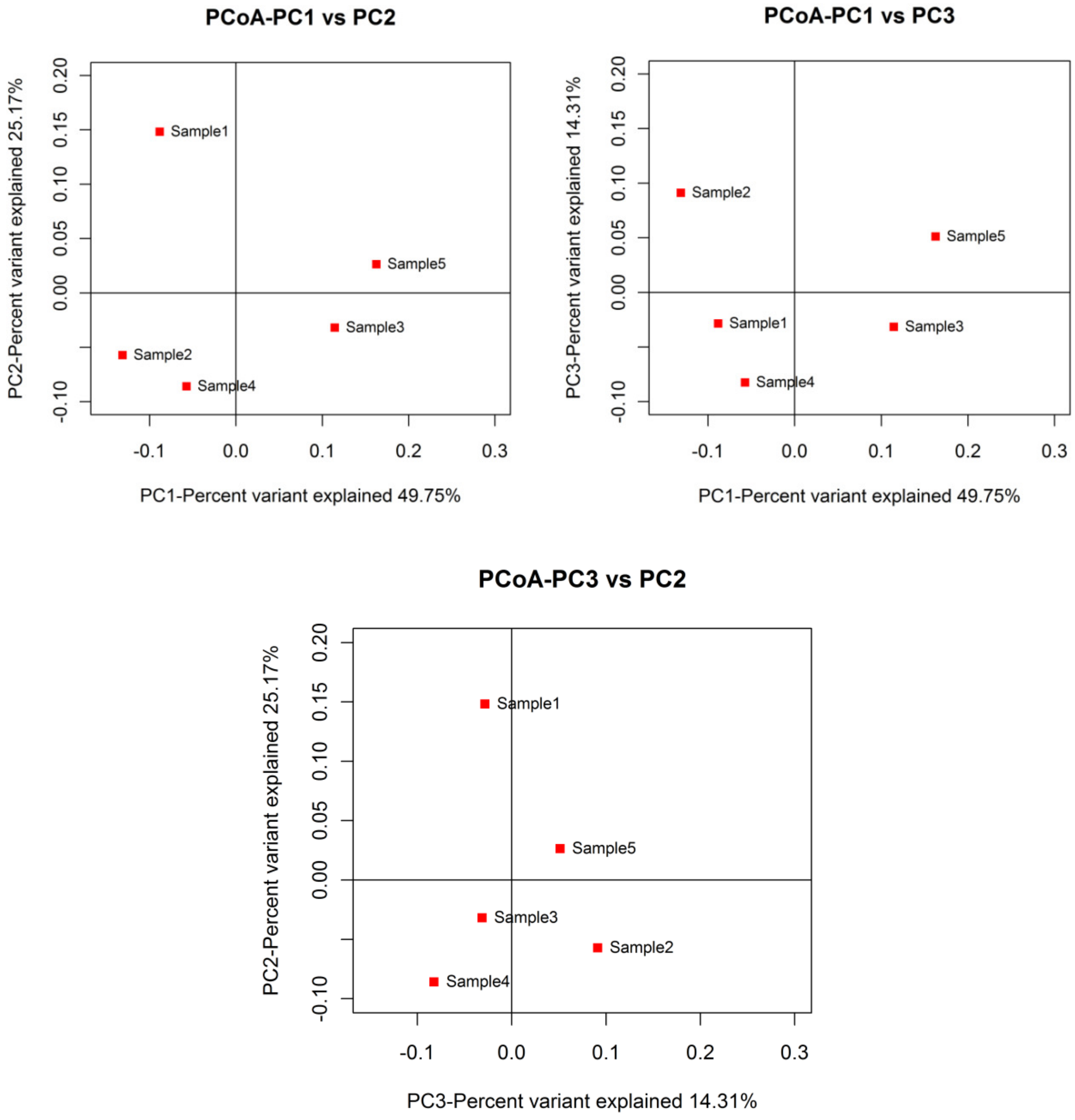


Figure 3.7.1 2D weighted unifrac PCoA Plot

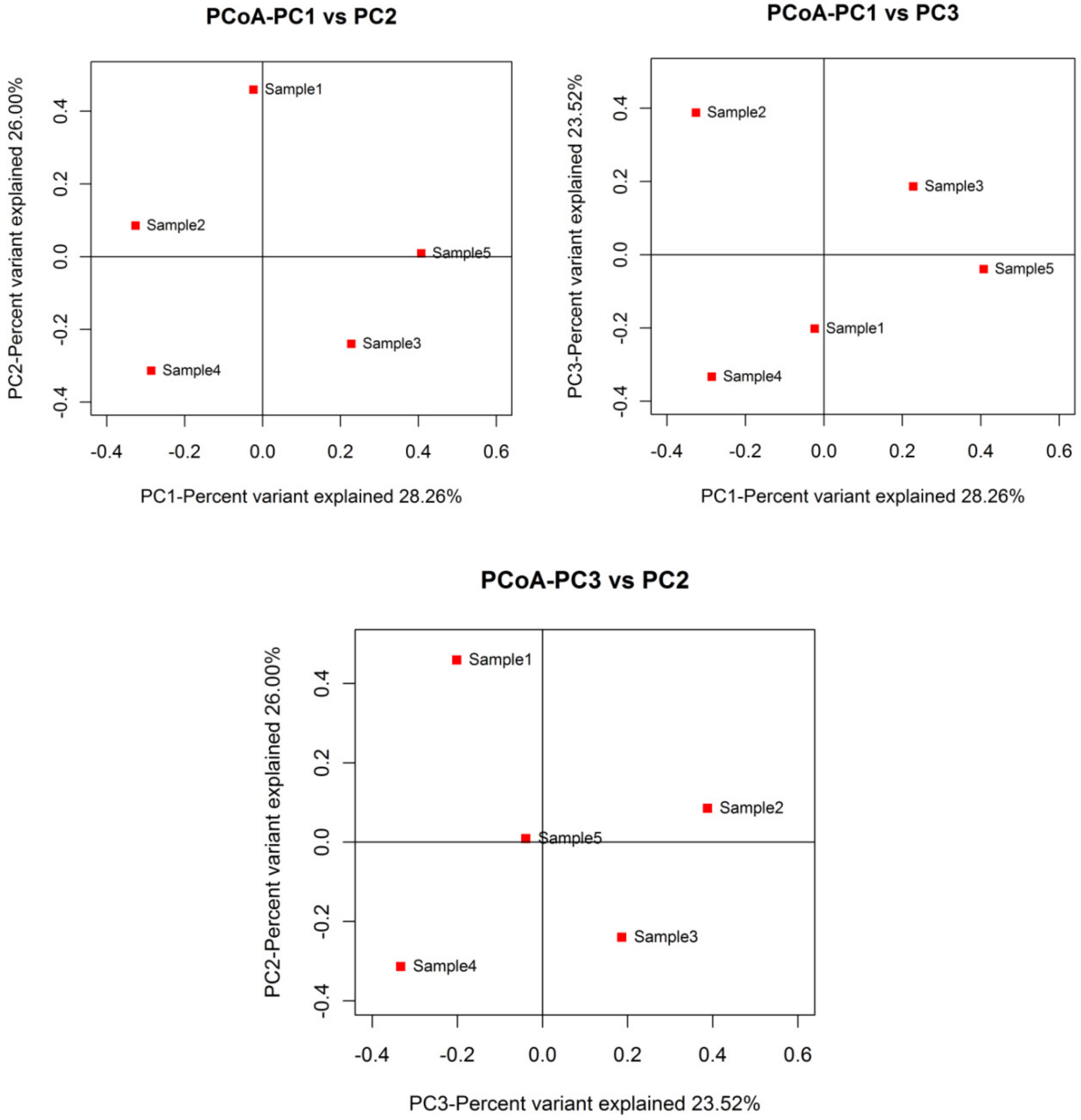


Figure 3.7.2 2D unweighted unifrac PCoA Plot

3.8 UPGMA Tree

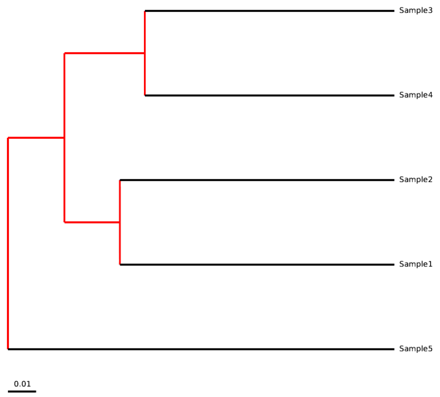


Figure 3.8.1 Weighted unifrac UPGMA tree

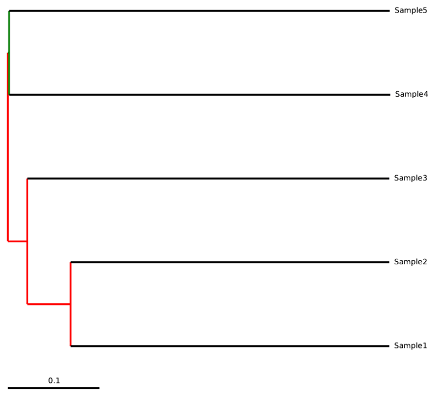
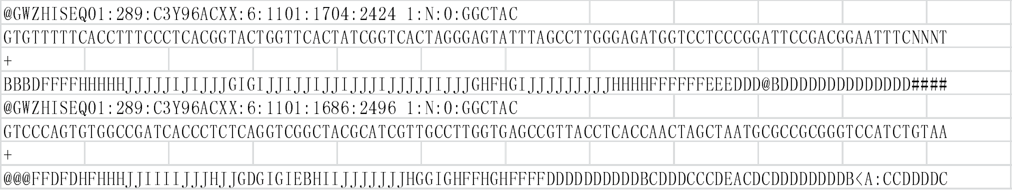


Figure 3.8.2 Unweighted unifrac UPGMA tree

4 Data Statistics

4.1 Data quality analysis

Image data generated by Miseq is transferred into raw reads through base calling software (BCL2FASTQ v2.17). These raw reads are stored in fastq format, which includes both a biological sequence (the second row in FASTQ) and its corresponding quality scores (the fourth row in FASTQ).



```
@GWZHISEQ01:289:C3Y96ACXX:6:1101:1704:2424 1:N:0:GGCTAC
GTGTTTTTCACCTTTCCCTCACGGTACTGGTTCACTATCGGTCACTAGGGAGTATTTAGCCTTGGGAGATGGTCCTCCCGGATTCCGACGGAATTTCNNNT
+
BBBDFFFFHHHHHJJJJJIJIJJJGIGIJJIJJIJJIJJJIJJJJJIJJJGHFHGIJJJJJJJJHHHHFFFFFFEEEDDD@BDDDDDDDDDDDDDD####
@GWZHISEQ01:289:C3Y96ACXX:6:1101:1686:2496 1:N:0:GGCTAC
GTCCCAGTGTGGCCGATCACCCTCTCAGGTCGGCTACGCATCGTTGCCTTGGTGAGCCGTTACCTCACCAACTAGCTAATGCGCCGCGGGTCCATCTGTAA
+
@@@FFDFDHFHHHJJIIIIJJJHJJGDGIGIEBHIIJJJJJJJHGGIGHFFHGHFFFFDDDDDDDDDDBCDDDCCCDEACDCDDDDDDDDB<A:CCDDDDC
```

Figure 4.1 FASTQ data

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description.
- Line 2 is the raw nucleotide sequence.
- 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 sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

GWZHISEQ01	Unique instrument name
321	Run ID
C5AL1ACXX	Flowcell ID
1	Flowcell lane
1101	Tile number within the flowcell lane
1184	'x'-coordinate of the cluster within the tile
2119	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
AGTCAA	Index sequence

4.2 Data statistics

The index sequences contained in the first 8 bp of each paired-end read were extracted and concatenated to form a 16 bp dual-index barcode specific for each paired read and sample.

Table 4.2 Statistics of Raw Data

Sample	length	# Reads	# Bases	Q20(%)	Q30(%)	GC(%)	N(ppm)
Sample 1	250.46	166246	41637410	94.41	91.85	54.34	60.11
Sample 2	250.38	236784	59285305	94.12	91.51	54.32	65.43
Sample 3	250.55	220638	55281245	96.34	94.50	54.31	73.55
Sample 4	250.23	264728	66243225	95.39	93.29	53.59	185.48
Sample N							

Format Description:

Column Number	Column Name	Description
1	Sample	Sample name
2	length	Average reads length
3	Reads	reads numbers
4	Bases	Bases numbers
5	Q20(%)	% of bases with <1% sequence error
6	Q30(%)	% of bases with <0.1% sequence error
7	GC(%)	% of Bases C+G content
8	N(ppm)	% of Undetermined bases per million bases

4.3 Data processing

Quality criteria:

1) The forward and reverse reads were joined using pandaseq (https://github.com/neufeld/pandaseq), truncation of sequence with “N” removal of sequence length less than 400.

2) Data filtering using Trimmomatic v0.30 (http://www.usadellab.org/cms/?page=trimmomatic), removal of primer and adaptor sequence, truncation of sequence reads with both pair end quality < 25, truncation of

sequence reads not having an average quality of 25 over a 4bp sliding window based on the phred algorithm.

3) Mapping clean reads using usearch (v8.0)

Table 4.3. Statistics of effective data

Sample	#PE_reads	#Nochimera	AvgLen(nt)	GC(%)	Effective(%)
Sample 1	83123	80773	454.78	54.29	97.17
Sample 2	118392	114610	457.50	54.32	96.81
Sample 3	110319	107055	447.86	54.26	97.04
Sample 4	132364	128951	452.64	53.66	97.42
Sample N					

Column name interpretation:

Colume name	description
Sample	Sample name
#PE_reads	Raw reads number
#Nochimera	Effective sequence number after removal of the chimeric
AvgLen(nt)	Average length of effective sequence
GC(%)	GC percentage content of effective sequence
Effective(%)	Nochimera/PE_reads

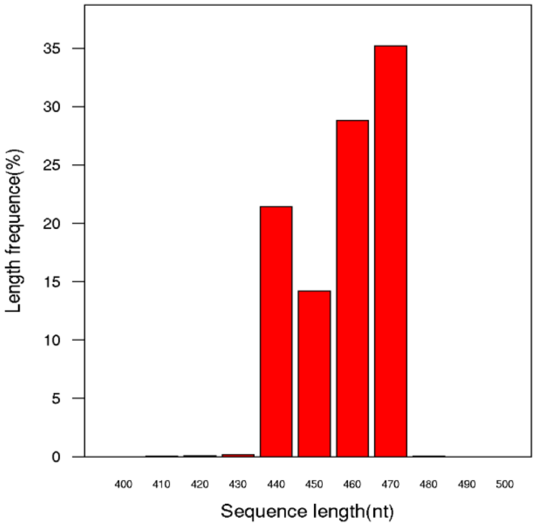


Figure 4.3 Sequence length distribution

5 Results Files

5.1 Results catalogue

```
00_Data
├── PFdata_stat.txt
├── final_len_distribution.tiff
└── effective_stat.txt
01_OTU
├── otu_table_mc2_w_tax.biom
├── otu_table.xls
├── otu_venn.tif
├── rep_set.fna
└── rep_set.tre
02_Rank_Abundance
└── rank_abundance.tif
03_Taxonomy
├── taxonomy_treefile.xls
├── taxa_summary_by_sample
└── taxa_summary_by_group
04_ Rarefaction_curve
└── Observed_OTUs_rarefaction_curves.tif
05_Alpha_Diversity
└── alpha_rarefaction.xls
06_Beta_Diversity
├── unweighted_unifrac.txt
└──weighted_unifrac.txt
07_PCoA
├── weighted_unifrac
|    ├── PC1_vs_PC2_plot.tif
|    ├── PC1_vs_PC3_plot.tif
|    └── PC3_vs_PC2_plot.tif
└── unweighted_unifrac
    ├── PC1_vs_PC2_plot.tif
    ├── PC1_vs_PC3_plot.tif
    └── PC3_vs_PC2_plot.tif

08_UPGMA_tree
├── weighted_unifrac.tif
└── unweighted_unifrac.tif
```

5.2 Documents browser

1．Documents includes sequence data and analysis results.

2．Documents Uncompress：

Unix/Linux/Mac system: tar –zcvf *.tar.gz “*.tar.gz”
gunzip *.tar.gz
Windows system：   WinRAR

3．Fastq format data: for Unix/Linux ,using ‘more’ or ‘less’ command； for Windows , text.

6 References

[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] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Opens external link in new windowNucl. Acids Res. 41 (D1): D590-D596.

[4] Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Opens external link in new windowNucl. Acids Res. 42:D643-D648

[5] Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucl. Acids Res. 35:7188-7196

[6] Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner 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

[7] 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

[8] 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

[9] 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.

[10] 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