# Supplementary

**Sputum collection and disposition process**

1. Collection procedure: collect morning sputum, rinse your mouth with clean water before collection or clean your mouth and teeth with a toothbrush out of toothpaste. Dentures should be removed if patients have dentures. Cough up the sputum from the deep part of the respiratory tract, and the sputum should be spit directly into the phlegm cup of the yellow lid, and the amount of specimen should be more than 2g.

2. Disposition procedure: collect the sputum coughed by patients in a sterile dry petri dish, select the sputum suppository with aseptic tweezers and collect it in the aseptic drying test tube, weigh at least 0.3g, add 4 times the value of mL 0.1%DTT (i.e. 0.1g sputum with 0.4ml 0.1%DTT), blow and mix with a Pap straw at room temperature for 3-5 minutes to make it a uniform suspension. And separate the homogenized sputum samples (such as the sputum into 1.5ml EP tube, 0.5-1ml/ tube). Finally, it was stored at-80 ℃.

**The specific process of shotgun metagenomic sequencing**

1.DNA sample detection mainly includes two methods:

(1) the purity and integrity of DNA were analyzed by agarose gel electrophoresis (AGE).

(2) Quantify the concentration of DNA accurately with Qubit machine.

2. Library construction and library inspection.

The qualified DNA samples were randomly broken into fragments of about 350bp length by Covaris ultrasonic crusher, and the whole library was prepared by terminal repair, adding A tail, adding sequencing connector, purification, PCR amplification and so on.

After the construction of the library, the library was initially quantified by Qubit2.0 and diluted to 2ng/μl. Then, the insert size of the library was detected by Agilent 2100, and the effective concentration of the library was quantified accurately by Q-PCR method (the effective concentration of the library > 3nM) to ensure the quality of the library.

3. Computer sequencing.

After the library is qualified, different libraries are sequenced by Illumina PE150 after pooling according to the effective concentration and the target amount of data off the machine.

**Pretreatment and assembly process of sequencing results**

1. Pretreatment process.

1) Remove reads that contains low-quality bases (mass value < = 38) that exceeds a certain percentage (default is set to 40bp).

2) Remove reads with a certain percentage of N bases (default is 10bp).

3) Remove reads when overlap exceeds a certain threshold between Adapter and reads (default is set to 15bp).

4) If there is host contamination in the sample, it should be compared with the host sequence to filter out the reads that may come from the host[1-3] (default is Bowtie2 software, parameter setting:--end-to-end,--sensitive,-I 200,-X 400).

2. Assembly process.

1) After pre-processing, we get Clean Data, and use MEGAHIT assembly software to assemble and analyze (Assembly Analysis); assembly parameters: --presets meta-large.

2) Break the assembled Scaffolds from the N junction to get a sequence fragment without N, which is called Scaftigs[4, 5].

3) Compare all kinds of quality-controlled CleanData to the assembled Scaftigs of each sample by Bowtie2 software, and obtain the PE reads that has not been used. Comparison parameters: --end-to-end,--sensitive,-I 200,-X 400.

4) Put the unused reads of each sample together for mixed assembly[2, 4, 6], and the assembly parameters are the same as those of the single sample.

5) Break the mixed assembled Scaffolds from the N junction to get the N-free Scaftigs sequence.

6) For Scaftigs from single sample and mixed assembly, the fragments below 500bp were filtered out[1, 7, 8], and statistical analysis and subsequent gene prediction were carried out.

**Specific steps of gene prediction and abundance analysis**

1) Starting from the Scaftigs of each sample and mixed assembly (> = 500bp), using MetaGeneMark[2-4, 7, 9, 10]to predict ORF (Open Reading Frame), and from the prediction results, filtering out the information whose length is less than 100nt[5, 6, 11]; prediction parameters: using default parameters.

2) The ORF prediction results of each sample and mixed assembly are de-redundant by CD-HIT[12, 13] software to obtain non-redundant initial gene catalogue (In operation, the nucleic acid sequence encoded by non-redundant continuous gene is called genes[6]). By default, identity 95%, coverage 90% are clustered[7, 10], and the longest sequence is selected as the representative sequence. Adopt parameters: -c 0.95,-G 0,-aS 0.9,-g 1,-d 0.

3) Using Bowtie2, to compare the Clean Data of each sample to the initial gene catalogue, to calculate the reads number of gene alignment in each sample.Comparison parameters[5, 7]:--end-to-end,--sensitive,-I 200,-X 400.

4) Filter out the genes that support the number of reads < = 2[2] in each sample, and obtain the gene catalogue (Unigenes) that will eventually be used for follow-up analysis.

5) The abundance information of each gene in each sample is calculated from the number of reads and gene length in the comparison. The calculation formula can be found in the reference[14-16].

**Basic steps of species annotation**

1) Use genes to compare with each functional database. Unigenes was compared with Bacteria, Fungi, Archaea and Viruses sequences extracted from NCBI's NR (Version: 2018.01 database using DIAMOND software[17](blastp, evalue < = 1e-5)[6].

2) Filtering the alignment results: for the alignment results of each sequence, the alignment results of evalue < = minimum evalue\*10 are selected for subsequent analysis.

3) After filtering, because each sequence may have multiple alignment results, many different species classification information is obtained. In order to ensure its biological significance, the LCA algorithm (applied to the systematic classification of MEGAN[18] software) is adopted to take the classification level before the first branch as the species annotation information of the sequence.

4) Based on the results of LCA annotation and gene abundance table, the abundance information of each sample at each taxonomic level (genera, species, phylum and so on) is obtained. For a species in a sample, the abundance of a species is equal to the sum of the gene abundance annotated as that species[1, 5, 8].

5) Based on the results of LCA annotation and gene abundance table, the gene table of each sample at each taxonomic level (genera, species, phylum and so on) is obtained. For a species, the number of genes in a sample is equal to the number of genes whose abundance is not zero among the genes annotated as that species.

**Basic steps of function annotation**

1) Use DIAMOND software to compare Unigenes with each functional database (blastp, evalue <= 1e-5) [2, 7].

2) Filtering of alignment results: for the alignment results of each sequence, the highest alignment result of score (one HSP > 60 bits) is selected for subsequent analysis[7, 18].

3) Based on the comparison results, the relative abundance of different functional levels is calculated (the relative abundance of each functional level is equal to the sum of the relative abundance of genes annotated as this functional level[7]). Among them, the KEGG database is divided into 6 levels, the eggNOG database is divided into 3 levels, and the CAZy database is divided into 3 levels.

4) Based on the results of functional annotation and gene abundance table, the gene table of each sample at each classification level is obtained. For the number of genes with a certain function in a sample, it is equal to the number of genes annotated as that function whose abundance is not zero.

**References:**

[1]. FH, K., et al., Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature, 2013. 498(7452): p. 99-103.

[2]. FH, K., et al., Symptomatic atherosclerosis is associated with an altered gut metagenome. Nature communications, 2012. 3: p. 1245.

[3]. DR, M., et al., Assessment of metagenomic assembly using simulated next generation sequencing data. PloS one, 2012. 7(2): p. e31386.

[4]. HB, N., et al., Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature biotechnology, 2014. 32(8): p. 822-8.

[5]. J, Q., et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010. 464(7285): p. 59-65.

[6]. G, Z., et al., Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology, 2014. 10: p. 766.

[7]. J, L., et al., An integrated catalog of reference genes in the human gut microbiome. Nature biotechnology, 2014. 32(8): p. 834-41.

[8]. J, O., et al., Biogeography and individuality shape function in the human skin metagenome. Nature, 2014. 514(7520): p. 59-64.

[9]. W, Z., L. A and B. M, Ab initio gene identification in metagenomic sequences. Nucleic acids research, 2010. 38(12): p. e132.

[10]. N, Q., et al., Alterations of the human gut microbiome in liver cirrhosis. Nature, 2014. 513(7516): p. 59-64.

[11]. W, L. and G. A, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England), 2006. 22(13): p. 1658-9.

[12]. L, F., et al., CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England), 2012. 28(23): p. 3150-2.

[13]. J, Q., et al., A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 2012. 490(7418): p. 55-60.

[14]. S, S., et al., Ocean plankton. Structure and function of the global ocean microbiome. Science (New York, N.Y.), 2015. 348(6237): p. 1261359.

[15]. A, C., et al., Dietary intervention impact on gut microbial gene richness. Nature, 2013. 500(7464): p. 585-8.

[16]. Le Chatelier E, et al., Richness of human gut microbiome correlates with metabolic markers. Nature, 2013. 500(7464): p. 541-6.

[17]. B, B., X. C and H. DH, Fast and sensitive protein alignment using DIAMOND. Nature methods, 2015. 12(1): p. 59-60.

[18]. DH, H., et al., Integrative analysis of environmental sequences using MEGAN4. Genome research, 2011. 21(9): p. 1552-60.

Supplementary table 1. The table of relative abundance of the top ten species at phylum level among groups

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| phylum | p\_\_Firmicutes | p\_\_Proteobacteria | p\_\_Bacteroidetes | p\_\_Actinobacteria | p\_\_Fusobacteria | p\_\_Candidatus Saccharibacteria | p\_\_Chlamydiae | p\_\_Tenericutes | p\_\_Candidatus Gracilibacteria | p\_\_Chytridiomycota | Others |
| PeriodA | 0.134996 | 0.130529 | 0.04123 | 0.025691 | 0.005722 | 0.002149 | 0.002011 | 0.000246 | 0.00006 | 0.000207 | 0.657159 |
| PeriodB | 0.158 | 0.238272 | 0.047431 | 0.022255 | 0.005818 | 0.001507 | 0.001325 | 0.000241 | 0.000006 | 0.000137 | 0.525008 |
| PeriodC | 0.071417 | 0.044405 | 0.015885 | 0.009751 | 0.003538 | 0.001408 | 0.002913 | 0.000101 | 0.000017 | 0.000312 | 0.850253 |

Supplementary table 2. The table of relative abundance of the top ten species at genus level among groups

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| genus | g\_\_Streptococcus | g\_\_Neisseria | g\_\_Haemophilus | g\_\_Prevotella | g\_\_Moraxella | g\_\_Actinomyces | g\_\_Lautropia | g\_\_Veillonella | g\_\_Pseudomonas | g\_\_Porphyromonas | Others |
| PeriodA | 0.090228 | 0.052198 | 0.027154 | 0.025161 | 0.006587 | 0.008498 | 0.009389 | 0.015948 | 0.003492 | 0.00484 | 0.756505 |
| PeriodB | 0.117782 | 0.088008 | 0.103219 | 0.035077 | 0.008908 | 0.008121 | 0.005802 | 0.017956 | 0.002095 | 0.002031 | 0.611001 |
| PeriodC | 0.036944 | 0.007967 | 0.001997 | 0.0079 | 0.00003 | 0.002581 | 0.011583 | 0.00494 | 0.000375 | 0.001628 | 0.924055 |

Supplementary table 3. Relative abundance table of gene function between groups (KEGG ortholog group, KO)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| KO\_ID | K07481 | K01990 | K16087 | K01992 | K07052 | K02004 | K21572 | K21449 | K02003 | K02014 |
| PeriodA | 0.000704 | 0.000758 | 0.000459 | 0.000743 | 0.000432 | 0.000617 | 0.000331 | 0.000602 | 0.00048 | 0.000548 |
| PeriodB | 0.001193 | 0.000921 | 0.001348 | 0.000944 | 0.000524 | 0.000744 | 0.000403 | 0.000764 | 0.000572 | 0.000692 |
| PeriodC | 0.000179 | 0.000337 | 0.000061 | 0.000265 | 0.000198 | 0.000257 | 0.000123 | 0.000172 | 0.000199 | 0.000194 |