Methods for metatranscriptome RNA-seq

1. Sample qualification and quantification

Total RNA was qualified and quantified as follows: (1) RNA sample was firstly qualified using 1% agarose gel electrophoresis for possible contamination and degradation; (2) RNA purity and concentration were then examined using NanoPhotometer® spectrophotometer; (3) RNA sample was precise qualified by Qubit2.0 Fluorometer; (4) RNA integrity and quantity were finally measured using RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system.

2. Library preparation

RNA library for metatranscriptome-seq was prepared as rRNA depletion and stranded method. Briefly, the ribosomal RNA was depleted from total RNA using the rRNA Removal Kit following manufacturer's instruction. RNA was then fragmented into 250~300 bp fragments and reverse-transcribed into cDNA subsequently. Remaining overhangs of double-strand cDNA were converted into blunt ends via exonuclease/ polymerase activities. After adenylation of 3' ends of DNA fragments, sequencing adaptors were ligated to the cDNA. In order to select cDNA fragments of preferentially 250~300 bp in length, the library fragments were purified with AMPure XP system. Amplification of cDNA was performed using PCR.

After library construction, the concentration of library was measured by the Qubit® fluorometer and adjusted to 1ng/μL. Agilent 2100 Bioanalyzer was deployed to examine the insert size of the acquired library. At last, the accurate concentration of cDNA library was again examined using qPCR. Once the insert size and concentration of the library was identical, the samples can then be subjected for sequencing.

3. Sequencing

After library preparation and pooling of different samples, the samples were subjected for Illumina sequencing. Commonly, the metatranscriptome-seq use PE150 (paired-end 150nt) sequencing for 5/10G raw data.

4. Quality control for raw data

Raw data (raw reads) of FASTQ format were firstly processed through fastp software. In this

step, clean data (clean reads) were obtained by removing following reads: (1) reads with 5' adapter; (2) reads without 3' adapter or insert sequence; (3) reads with more than 10% N; (4) reads with more than 50% nucleotides with Qphred<=20; (5) reads with ploy A/T/G/C. Adapter trimming for the removal of adapter sequences from the 3' ends of reads was also performed. At the same time, Q20, Q30 and GC content of the clean data were calculated. All the downstream analyses were based on the clean data with high quality.

5. Assembly of transcripts

For the clean reads obtained by preprocessing each sample, de novo assembly was performed with Trinity, then the sequence of all samples was integrated and repeat (set sequence identity threshold 0.95) was deleted using CORSET to obtain unigene dataset.

6. Species annotation

Use DIAMOND software to compare unigenes with the sequences of bacteria, fungi, archaea and viruses extracted from NCBI's NR database (blastp), Evalue ≤ 1e-5); Comparison result filtering : For each sequence comparison result , select the evalue <= minimum evalue *10 for subsequent analysis ; After filtering , reads that match the sequence of some genes , were placed to the lowest common ancestor (LCA) node of those species in the taxonomy that were known to have that gene (LCA algorithm , applied to the system classification of MEGAN software). The classification level before the first branch would appear as the species annotation information of the sequence.

According to the LCA annotation results and the gene abundance table, the abundance information and the table of gene numbers of each sample at each taxonomic level (genus, phyla, family, genus, and species) were obtained.

According to the abundance tables on each taxonomic level (genus, family, genus, and species), statistical maps and abundance heat maps at different taxonomic levels were drawn.

7. Annotation of common function database

In order to obtain comprehensive gene function information, gene function annotations in four major databases were prepared including: GO, KEGG, CAZy and eggNOG. Using DIAMOND software, unigenes were mapped to each functional database (blastp, evalue ≤ 1e-5). Mapping result filtering: Since the mapping result of each sequence may be more than one, to ensure the biological

significance of the follow-up study, the comparison result of each sequence was screened with the coverage ratio BCR (The BLAST Coverage Ratio) of reference and query in each mapping record greater than 40% .

8. Quantification and differential expression analysis

The clean reads of each sample were mapped to reference sequence that was spliced by Trinity. In this process, RSEM software was applied where the bowtie2 parameter mismatch was set to 0 (bowtie2 default parameter).

Quantification of the transcripts and genes was performed with the Fragments Per Kilobase of transcript sequence per Millions base pairs (FPKM)-normalization method using RSEM . DESeq2 or edgeR was used for differential expression analysis . The resulting P-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate . Genes with |log2 (Fold Change) | > 1 & padj < 0.05 were assigned as differentially expressed.

9. GO and KEGG enrichment analysis

The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. The GO covers three domains: cellular component, molecular function and biological process. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. The KEGG PATHWAY database, the wiring diagram database, is the core of the KEGG resource. It is a collection of pathway maps integrating many entities including genes, proteins, RNAs, chemical compounds, glycans, and chemical reactions, as well as disease genes and drug targets, which are stored as individual entries in the other databases of KEGG.

GO and KEGG enrichment analysis of differentially expressed RNAs was implemented by the cluster Profiler R package, in which gene length bias was corrected. The enrichment was considered to be significant when the corrected p-values was less than 0.05.