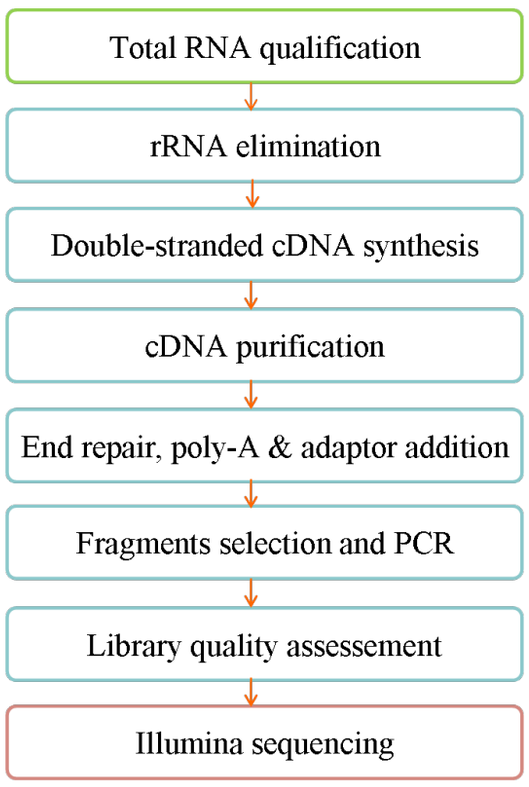
Sample collection and preparation

From the RNA sample to the final data, each step, including sample test, library preparation, and sequencing, influences the quality of the data, and data quality directly impacts the analysis results. To guarantee the reliability of the data, quality control (QC) is performed at each step of the procedure. The workflow is as follows:



RNA quantification and qualification

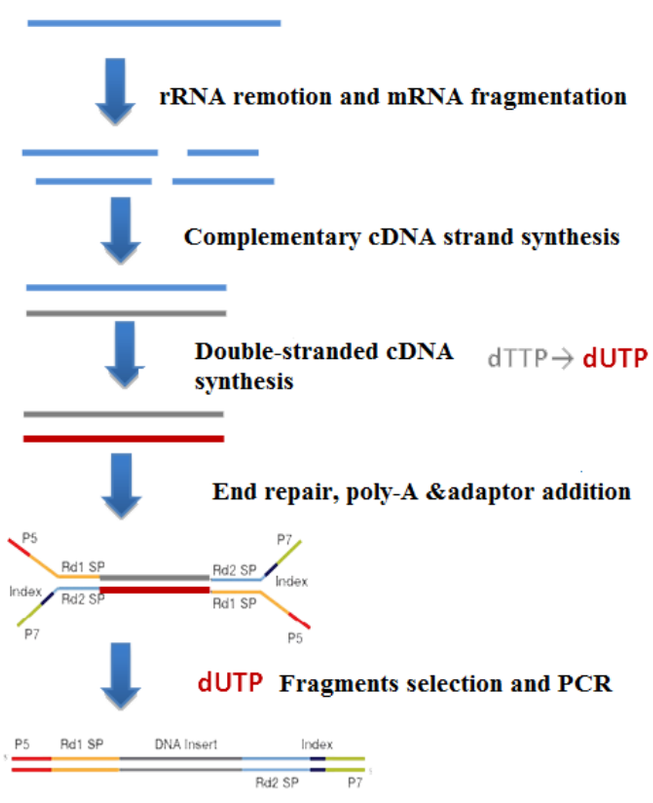
RNA degradation and contamination was monitored on 1% agarose gels.

RNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA).

RNA integrity and quantitation were assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

Library preparation for transcriptome sequencing

A total amount of 1 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 150~200 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 μl USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.



Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using PE Cluster Kit cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina platform and paired-end reads were generated.

Data Analysis

Quality control

Raw data (raw reads) of FASTQ format were firstly processed through fastp. In this step, clean data (clean reads) were obtained by triming reads containing adapter and removing poly-N sequences and reads with low quality from raw data. 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.

Reads mapping to the reference genome

Reference genome and gene model annotation files were downloaded from genome website directly. Both building index of reference genome and aligning clean reads to reference genome were used Bowtie2. (Langmead, B. and S.L. Salzberg, 2012)

Novel gene and gene structure analysing

Rockhopper was used to identify novel genes, operon and transcription start sites. It can be used for efficient and accurate analysis of bacterial RNA-seq data, and that it can aid with elucidation of bacterial transcriptomes (McClure, R., D.Balasubramanian, et al, 2013). Then, we extract upstream 700bp sequence of Transcription Start Site for predicting promoter using TDNN (Time-Delay Neural Network).

Quantification of gene expression level

FeatureCounts was used to count the reads numbers mapped to each gene. And then FPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. FPKM, expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced, considers the effect of sequencing depth and gene length for the reads count at the same time, and is currently the most commonly used method for estimating gene expression levels (Trapnell, Cole, et al., 2010).

Differential expression analysis

(For DESeq2 with biological replicates) Differential expression analysis of two conditions/groups (two biological replicates per condition) was performed using the DESeq2 R package. DESeq2 provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted P-value <0.05 found by DESeq were assigned as differentially expressed.

(For edgeR without biological replicates) Prior to differential gene expression analysis, for each sequenced library, the read counts were adjusted by Trimmed Mean of M-values (TMM) through one scaling normalized factor. Differential expression analysis of two conditions was performed using the edgeR R package. The *P* values were adjusted using the Benjamini and Hochberg methods. Corrected pvalue of 0.005 and $|\log_2^{(\text{Fold Change})}|$ of 1 were set as the threshold for significantly differential expression.

GO and KEGG enrichment analysis of differentially expressed genes

Gene Ontology (GO) enrichment analysis of differentially expressed genes was implemented by the clusterProfiler R package, in which gene length bias was corrected. GO terms with corrected Pvalue less than 0.05 were considered significantly enriched by differential expressed genes.

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular level information, especially large-scale molecular datasets generated by genome sequencing and other high-through put experimental technologies (http://www.genome.jp/kegg/). We used clusterProfiler R package to test the statistical enrichment of differential expression genes in KEGG pathways.

Predict UTR

According to the information of Transcription Start Site (Transcription terminal Site) and Translation start site (Translation terminal site), we extracted 5'UTR (3'UTR) sequences. Then, RBSfinder (Suzek, B, et al, 2001) and TransTermHP (Kingsford, C. L.et al, 2007) were used to predict SD sequence and terminator sequence respectively.

Analysis of ncRNA

IntaRNA was used to predict sRNA targets.And then we used RNAfold to predict RNA secondary structures (Busch, A.et al, 2008; Hofacker, I. L, et al, 2006).

Mutaion analysis

Firstly, Picard tools and Samtools were used to sort, mark duplicated reads and reorder the bam alignment results of each sample. Then the tool HaplotypeCaller in GATK software was used to perform variant discovery, including single nucleotide polymorphism (SNPs), insertions and deletions (INDELs). Raw VCF files were filtered with GATK standard filter method and other parameters (cluster: 3; WindowSize: 35; QD < 2.0 or FS > 60.0). Finally, SnpEff annotates variants based on their genomic locations and predicts coding effects.

Reference

Marioni, J. C., C. E. Mason, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research.

Mortazavi, A., B. A. Williams, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods.

Busch, A., A. S. Richter, et al. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics.(IntaRNA)

Hofacker, I. L. and P. F. Stadler (2006). Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics.(RNAfold)

Kingsford, C. L., K. Ayanbule, et al. (2007).Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome biology.(TransTermHP)

McClure, R., D. Balasubramanian, et al. (2013). Computational analysis of bacterial RNA-Seq data. Nucleic acids research.(Rockhopper)

Suzek, B., et al. (2001). A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics .(RBSfinder)

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol.(Bowtie)

Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. Nature methods.(Bowtie2)

Anders, S.(2010). HTSeq: Analysing high-throughput sequencing data with Python.(HTSeq)

McKenna, A, Hanna, M, et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research.(GATK)

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol.(DESeq)

Anders S, Huber W. (2010).Differential expression analysis for sequence count data.Genome Biology,doi:10.1186/gb-2010-11-10-r106. (DESeq2)

Robinson, M. D., et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.(edgeR)

Young, M. D., et al. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology.(GOseq)

Kanehisa, M., et al. (2008). KEGG for linking genomes to life and the environment. Nucleic acids research.(KEGG)

Mao, X., et al. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics.(KOBAS)

Waibel, A. H., et al. (1989).Phoneme Recognition Using Time-Delay Neural Networks IEEE Transactions on Acoustic, Speech, and Signal Processing Vol. 37 no.3,328-339.(TDNN)