1 Sample collection and preparation

1.1 RNA quantification and qualification

RNA concentration and purity was measured using NanoDrop 2000(Thermo Fisher Scientific, Wilmington, DE). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

1.2 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™ Directional 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. 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 240 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.

1.3 Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumia) according to the manufacturer’s instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq Xten platform and paired-end reads were generated.

2 Data analysis

2.1 Quality control

Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. In this step, clean data(clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30, GC-content and sequence duplication level of the clean data were calculated. All the downstream analyses were based on clean data with high quality.

2.2 Comparative analysis

The adaptor sequences and low-quality sequence reads were removed from the data sets. Raw sequences were transformed into clean reads after data processing. These clean reads were then mapped to the reference genome sequence. Only reads with a perfect match or one mismatch were further analyzed and annotated based on the reference genome. HISAT2 tools software were used to map to reference genome.

2.3 Gene functional annotation

Gene function was annotated based on the following databases: NR(NCBI non-redundant protein sequences database) ; Nt (NCBI non-redundant nucleotide sequences) ; Pfam (The database of Homologous protein family) ;COG (The database of Clusters of Protein homology) ; Swiss-Prot (A manually annotated non-redundant protein sequence database) ; KOG(The database of Clusters of protein homology); KEGG(The database of Kyoto Encyclopedia of Genes and Genomes) ; GO (Gene Ontology database).

2.4 SNP calling

Picard - tools v1.41 and samtools v0.1.18 were used to sort, remove duplicated reads and merge the bam alignment results of each sample. GATK2 or Samtools software was used to perform SNP calling. Raw vcffiles were filtered with GATK standard filter method and other parameters (clusterWindowSize: 10; MQ0 >= 4 and (MQ0/(1.0*DP)) > 0.1; QUAL < 10; QUAL < 30.0 or QD < 5.0 or HRun > 5), and only SNPs with distance > 5 were retained.

2.5 Quantification of gene expression levels

Gene expression levels were estimated by fragments per kilobase of transcript per million fragments mapped(FPKM). The formula is shown as follow:

$$\mathrm{FPKM} = \frac{\text{cDNA Fragments}}{\text{Mapped Fragments(Millions)} \times \text{Transcript Length(kb)}}$$

图 1 formula of FPKM

2.6 Differential expression analysis

(1) For the samples with biological replicates:
Differential expression analysis of two conditions/groups was performed using the DESeq R package (1.10.1). DESeq 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 < 0.01 and absolute value of log2(Fold change)>1 found by DESeq were assigned as differentially expressed.
(2) For the samples without biological replicates:
Prior to differential gene expression analysis, for each sequenced library, the read counts were adjusted by edgeR program package through one scaling normalized factor. Differential expression analysis of two samples was performed using the EBSeq R package. The resulting FDR (false discovery rate) were adjusted using the PPDE(posterior probability of being DE).The FDR < 0.05 & |log2

（foldchange）| ≥1 was set as the threshold for significantly differential expression.

2.7 GO enrichment analysis

Gene Ontology (GO) enrichment analysis of the differentially expressed genes (DEGs) was implemented by the clusterProfiler R package.Enrichment analysis uses hypergeometric testing to find GO entries that are significantly enriched compared to the entire genome background.GSEA(Gene Set Enrichment Analysis) can also be analysed by clusterProfiler..

2.8 KEGG pathway enrichment analysis

KEGG (Kanehisa et al., 2008) 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-throughput experimental technologies (http://www.genome.jp/kegg/). We used KOBAS (Mao et al., 2005) software to test the statistical enrichment of differential expression genes in KEGG pathways.We used clusterProfiler R packages to find KEGG pathway that are significantly enriched compared to the entire genome background.

2.9 PPI (Protein Protein Interaction)

The sequences of the DEGs was blast (blastx) to the genome of a related species (the protein protein interaction of which exists in the STRING database: http://string-db.org/) to get the predicted PPI of these DEGs. Then the PPI of these DEGs were visualized in Cytoscape (Shannon et al, 2003).

References

Altschul S F, Madden T L, Schaffer A A, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402. (BLAST)

Anders S, Huber W. (2010).Differential expression analysis for sequence count data. Genome Biology,doi:10.1186/gb-2010-11-10-r106. (DESeq)
Finn R D, Tate J, Mistry J, et al. (2008). The Pfam protein families database. Nucleic Acids Res 36, D281-D288. (Pfam)
Gotz S, Garcia-Gomez J M, Terol J, et al. (2008).High-throughput functional annotation and data mining with the Blast2GO suite.Nucleic Acids Research 36, 3420-3435. (BLAST2go)
Mao X, Cai T, Olyarchuk J G, et al. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.Bioinformatics 21, 3787-3793. (KOBAS)
McKenna A, Hanna M, Banks E, et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303. (GATK)
Li H, Handsaker B,Wysoker A. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009,25 (16): 2078-2079.(Samtools)
Kanehisa M, Araki M, Goto S, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids research 36:D480-D484. (KEGG)
Wang L, Feng Z, Wang X, et al. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-138. (DEGseq)
Robinson M D, McCarthy D J, Smyth G K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.(edgeR)
Shannon P, Markiel A, Ozier O, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498-2504. (Cytoscape)