仅供客户写英文文章时参考，分析内容和方法请以结题报告为准

——转录调控研究部

Methods

Sample collection and preparation

- RNA isolation

  Total RNA was isolated from …

- 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 concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA). RNA integrity was assessed using the RNA 6000 Pico Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

- Transcriptome sample preparation for sequencing

  A total amount of 3 g RNA per sample was used as input material for the RNA sample preparations. All four samples had RIN values above 8. Sequencing libraries were generated using Illumina TruSeq™ RNA Sample Preparation Kit (Illumia, San Diego, USA) following manufacturer's recommendations and four 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 Illumina proprietary fragmentation buffer. First strand cDNA was synthesized using random oligonucleotides and SuperScript II. 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 and enzymes were removed. After adenylation of 3' ends of DNA fragments, Illumina PE adapter oligonucleotides were ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 200 bp in length the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). DNA fragments with ligated adaptor molecules on both ends were selectively enriched using Illumina PCR Primer Cocktail in a 15 cycle PCR reaction. Products were purified (AMPure XP system) and quantified using the Agilent high sensitivity DNA assay 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 TruSeq PE Cluster Kit v3-cBot-HS (Illumia) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq 2000 platform and 100 bp paired-end reads were generated.

Data analysis

- 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 the clean data with high quality.

- Reads mapping to the reference genome

  Reference genome and gene model annotation files were downloaded from genome website (http://) directly. Index of the reference genome was built using Bowtie v2.0.6 and paired-end clean reads were aligned to the reference genome using TopHat v2.0.7. We selected TopHat as the mapping tool for that TopHat can generate a database of splice junctions based on the gene model annotation file and thus a better mapping resultthan other non-splice mapping tools.

- Quatification of gene expression level

  HTSeq v0.5.3 was used to count the reads numbers mapped to each gene. And then RPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. RPKM, Reads Per Kilobase of exon model per Million mapped reads, 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 (Mortazavi et al., 2008).

- Differential expression analysis

  (For DESeq with biological replicates) Differential expression analysis of two conditions/groups (two biological replicates per condition) 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 <0.05 found by

DESeq were assigned as differentially expressed.

(For DEGSeq 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 conditions was performed using the DEGSeq R package (1.12.0). the P values were adjusted using the Benjamini & Hochberg method. Corrected P-value of 0.005 and log2(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 GOseq 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-throughput experimental technologies (http://www.genome.jp/kegg/). We used KOBAS software to test the statistical enrichment of differential expression genes in KEGG pathways.

- Novel transcripts prediction and alternative splicing analysis

  The Cufflinks v2.0.2 Reference Annotation Based Transcript (RABT) assembly method was used to construct and identify both known and novel transcripts from TopHat alignment results. Alternative splicing events were classified to 12 basic types by the software Asprofile v1.0, the number of AS events in each sample was estimated, separately.

- SNP analysis

  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. GATK software was used to perform SNP calling. Raw vcf files 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.

References

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

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

Anders, S. and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level-the DESeq package.(DEseq)

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

Kim, D., G. Pertea, et al. (2012). TopHat2: Parallel mapping of transcriptomes to detect indels, gene fusions, and more.(TopHat2)

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.(Bowtie 2)

Mao, X., Cai, T., Olyarchuk, J.G., Wei, L. (1995). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics.(KOBAS)

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

McKenna, A, Hanna, M, Banks, E, Sivachenko, A, Cibulskis, K, Kernytsky, A, Garimella, K, Altshuler, D, Gabriel, S, Daly, M, DePristo, MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research.(GATK)

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

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.(edgeR)

Robinson, M. D. & Oshlack, A. (2010)A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol.(DEGSeq)

Trapnell, C. et al. (2010).Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol.(Cufflinks)

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics.(TopHat)

Trapnell, C., A. Roberts, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. nature protocols.(Tophat & Cufflinks)

Wang, L.Feng, Z.Wang, X.Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics.(DEGseq)

Wang, Z., M. Gerstein, et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology.(GOseq)