LncRNA Method (Illumina)

1. Experimental Procedure

1.1 Sample Quality Control

Please refer to QC report for methods of sample quality control.

1.2 Library Construction, Quality Control and Sequencing

Total RNA is used as input material for the RNA sample preparations. rRNA is removed from total RNA by using specific probes. Fragmentation is carried out using divalent cations under elevated temperature in First Strand Synthesis Reaction Buffer. First strand cDNA is synthesized using random hexamer primer and reverse transcriptase. Then the second strand of cDNA is synthesized by adding buffer and dNTPs (dTTP in dNTP is replaced by dUTP). The synthesized double-stranded cDNA is treated by terminal repair, adding A, connecting adaptors, and PCR enrichment is performed after fragment selection. In order to select cDNA fragments with a length of 370-420bp, PCR products are purified with AMPure XP beads to obtain strand-specific library.

After the library construction is completed, preliminary quantification is carried out using Qubit. Subsequently, the inserted fragments of the library are detected. Once the inserted fragments meet expectations, the effective concentration of the library is accurately quantified using qRT-PCR to ensure the quality of the library.

The qualified libraries are pooled and sequenced on Illumina platforms, according to effective library concentration and data amount required.

2. Bioinformatics Analysis Pipeline

2.1 Data Quality Control

2.1.1 Raw Data

The original fluorescence image files obtained from Illumina platform are transformed to short reads (Raw data) by base calling and these short reads are recorded in FASTQ format (Cock P. et al, 2010), which contains sequence information and corresponding sequencing quality information.

2.1.2 Evaluation of Data (Data Quality Control)

Sequence artifacts, including reads containing adapter contamination, low-quality nucleotides and unrecognizable nucleotide (N), undoubtedly set the barrier for the subsequent reliable bioinformatics analysis. Hence quality control is an essential step and applied to guarantee the meaningful downstream analysis. We used Fastp (v0.23.1) (Chen S. et al, 2018) to perform basic statistics on the quality of the raw reads.

The steps of data processing are as follows:

(1) Discard a paired reads if either one read contains adapter contamination;

(2) Discard a paired reads if more than 10% of bases are uncertain in either one read;

(3) Discard a paired reads if the proportion of low quality (Phred quality <5) bases is over 50% in either one read.

2.1.3 Mapping to Reference Genome

We used Hisat2 (v2.2.1) to compare clean reads with the reference genome to obtain the location information of reads on the reference genome. The higher mapping rate, the higher accuracy rate of finding junction reads (Kim D et al., 2015). As a comparison tool, Hisat2 can generate a concatenation database based on gene

model annotation files, quickly process a large number of sequencing data, provide high-precision comparison results, and efficiently detect junction reads, which has a better effect than other non-concatenation comparison tools.

2.1.4 Transcript Assembly

Based on the comparison results, Stringtie (v2.2.3) is used to concatenate reads into transcripts and quantify them (Pertea M et al., 2015). Stringtie has specific parameters for different libraries, which can concatenate transcripts more accurately and achieve transcription quantification.

2.1.5 Transcript Identification

Stringtie is used to merge the transcripts obtained from the concatenated samples, and the transcripts whose chain direction is uncertain and length does not exceed 200nt are removed. Gffcompare (v0.12.6) is used to compare with the known database, filter the known transcripts in the database, and finally predict the coding potential of the selected new transcripts.

2.1.6 Quantification

The expression levels of mapped, spliced, screened transcripts and predicted transcripts need to be quantitatively analyzed. FPKM is the number of fragments per kilobase of transcript sequence sequenced per million base pairs. It is a simple and commonly used expression level normalization method that can standardize sequencing depth and genome size.

2.1.7 Differential Expression Analysis

We used edgeR (v4.0.16) (Robinson MD et al., 2010) software to analyze the

significance of expression differences, and p value or padj is used to determine the significance level. It can be divided into 3 steps:

(1) Standardize the original read count, mainly to correct the sequencing depth;

(2) Calculate p value of the statistical model;

(3) The FDR value is obtained by multiple hypothesis testing;

2.1.8 Enrichment Analysis

We used clusterProfiler (v4.8.1) (Yu G et al, 2012) to achieve functional enrichment analysis of differentially expressed genes in GO (Gene Ontology). Differentially expressed genes are significantly enriched, and padj < 0.05 as the threshold of significant enrichment. KEGG (Kyoto Encyclopedia of Genes and Genomes) is the main public database of pathway significant enrichment analysis, hypergeometric test is applied to identify the pathway of significant enrichment in candidate target genes, and padj < 0.05 as the threshold of significant enrichment. The differentially expressed genes in the KEGG pathway is analyzed by clusterProfiler.

2.1.9 Prediction and Functional Analysis of LncRNA Target Genes

LncRNA is rich in biological functions and widely participates in various important physiological processes of organisms. LncRNA can regulate the expression of target genes at the transcriptional and post-transcriptional levels, and identify the differential functional pathways generated by different treatments (Schmitt AM et al., 2016). Predict the target genes of lncRNAs, and functional enrichment analysis (GO/KEGG) is performed on target genes to predict the main functions of lncRNAs.

2.1.10 Alternative Splicing Analysis

Alternative splicing means that the same precursor mRNA has multiple splicing procedures to form different mRNAs. It is an important mechanism to regulate gene expression and produce protein diversity. Alternative splicing analysis is performed by rMATS (Shen S et al., 2014). rMATS (v4.3.0) can analyze SE (Exon Skip), RI (Retained Intron), ME (Mutually Exclusive Exon), A5SS (Alternative 5 ' splicing site), A3SS (Alternative 3 ' splicing site).

2.1.11 Variant Sites Analysis

SNP (Single Nucleotide Polymorphism) refers to a genetic marker formed by a single nucleotide mutation on the genome, which has a large number and abundant polymorphisms. InDel stands for insertion deletion, which refers to the insertion deletion of a small fragment in a sample relative to the reference genome, which may contain one or more bases. GATK (v4.6.0.0) (McKennaAetal.,2010) is used to analyze the variant sites of the sample, and SnpEff is used to annotate the variant sites.

3. References

Cock P.J.A. et al (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771.

Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu; fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i884–i890.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory

requirements. Nat Methods. 2015;12(4):357-360. doi:10.1038/nmeth.3317.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290-295. doi:10.1038/nbt.3122.

Ghosh S, Chan CK. Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol. 2016;1374:339-361. doi:10.1007/978-1-4939-3167-5_18.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616.

Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-287. doi:10.1089/omi.2011.0118.

Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. Cancer Cell. 2016;29(4):452-463. doi:10.1016/j.ccell.2016.03.010.

Shen S, Park JW, Lu ZX, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111(51):E5593-E5601. doi:10.1073/pnas.1419161111.

McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.