CircRNA 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 was used as input material for the RNA sample preparations. rRNA was removed from total RNA by using specific probes and rRNA free residue was cleaned up by ethanol precipitation. Subsequently, the linear RNA was digested with RNase R. After fragmentation, the first strand cDNA was synthesized using random hexamer primers. Then the second strand cDNA was synthesized using dUTP, instead of dTTP. The directional library was ready after end repair, A-tailing, adapter ligation, size selection, amplification, and purification.

After the library construction, 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 were 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 sequencing 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 were 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 CircRNA Identification

Find circ (v1.2) (Memczak et al, 2013) and CIRI2 (v2.0.6) (Gao et al, 2015) are used to detect and identify circRNA to improve the accuracy of circRNA identification.

2.1.5 Quantitative Analysis of CircRNA

The expression levels of known and new circRNAs in each sample are statistically analyzed, and the expression levels are normalized by TPM (Zhou et al, 2010). Normalized expression level = (read count*1,000,000) / libsize (libsize: sum of sample circRNA read count).

2.1.6 CircRNA Difference Analysis

For samples with biological duplications, differential expression analysis between the two comparison combinations is performed by DESeq2 (v1.42.0) (Love MI et al., 2014), which provided a statistical procedure to determine differential expression in digital gene expression data using a model based on negative binomial distribution. For samples without biological duplications, edgeR (v4.0.16) (Robinson MD et al., 2010) TMM algorithm is used to standardize read count data for analysis.

2.1.7 Enrichment Analysis of Different CircRNA Source Genes

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.

2.1.8 Prediction of miRNA binding sites

circRNA can inhibit the function of miRNA by binding to miRNA (Hansen TB et al, 2013). Target miRNA sites of circRNA are predicted by miRanda software.

2.1.9 Prediction of CircRNA Coding Potential

IRESfinder software (Zhao J et al, 2018) is used to predict IRES scores, and then CPC, CNCI, and PAFM are used to identify whether circRNA has coding potential.

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.

Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495(7441):333-338. doi:10.1038/nature11928.

Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol. 2015;16(1):4. Published 2015 Jan 13.

doi:10.1186/s13059-014-0571-3.

Zhou L, Chen J, Li Z, et al. Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. PLoS One. 2010;5(12):e15224. Published 2010 Dec 30. doi:10.1371/journal.pone.0015224.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

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.

Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. Nature. 2013;495(7441):384-388. doi:10.1038/nature11993.

Zhao J, Wu J, Xu T, Yang Q, He J, Song X. IRESfinder: Identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features. J Genet Genomics. 2018;45(7):403-406. doi:10.1016/j.jgg.2018.07.006.