UMI SmallRNA 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

A total amount of 100ng total RNA per sample was used as input material for the small RNA library. Sequencing libraries are generated using QIAseq miRNA Library Kit. The 3' adapter specifically recognized and connected to the mature small RNA 3' ends. After the 3' ends connection reaction, the 5' ends adapter is connected, and then a reverse transcription (RT) primer with UMI is used to reverse transcribe into cDNA. The cDNA is then purified using DNA Clean Beads and amplified by PCR. The PCR products are finally purified using DNA Clean Beads, eluted with nuclease-free water, and dissolved in 8ul elution buffer for subsequent QC and sequencing.

After the library construction is completed, quantification of the inserted fragments and the effective concentration 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 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.

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

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.

UMI sequences on each read are identified with UMI-tools (v1.1.4), and reads with UMIs are subjected to subsequent analysis.

2.1.3 Mapping to Reference Genome

To identify the duplicated reads, UMIs are firstly removed from the UMI reads, and

remained parts of reads are mapped to the reference genome by Bowtie (Langmead et al, 2009). Reads that mapped to the same location on the reference genome are identified as duplicated reads. Then UMIs on each read are recalled, and duplicated reads with the same UMI are identified as the non-natural duplication. The non-natural duplication identified as above is subsequently removed from the clean data. Bowtie is used to locate sRNA after length screening to the reference sequence, and analyzed the distribution of sRNA on the reference sequence.

2.1.4 Analysis of Known miRNA

Reads matched to reference sequences are compared with sequences in the specified range in miRBase. The sRNA information of each sample is obtained by mirdeep2 (Friedlander et al, 2011) and SRna-Tools-CLI. Including the known miRNA secondary structure, miRNA sequence, length, frequency of occurrence, etc.

2.1.5 Remove Reads from These Sources

Annotate sRNA using the ncRNA sequences of the species, or select rRNA, tRNA, snRNA, and snoRNA from RFAM to annotate sRNA, identifying and removing potential rRNA, tRNA, snRNA, and snoRNA. Using species-specific repeat sequence annotation information, or reference sequence information for de novo prediction of repeat sequences, align sRNA with repeat sequences to identify and remove potential repeat sequences, and statistically analyze the types and quantities of sRNA matching various repeat types.

2.1.6 Novel miRNA Prediction

The characteristic hairpin structure of miRNA precursors can be used to predict new

miRNAs. Using miREvo (Wen et al., 2012) and mirdeep2 (Friedlander et al., 2011), sRNA sequences of a certain length are extracted and aligned with the reference genome. These sequences are then analyzed for secondary structure, Dicer cleavage site information, energy characteristics, and other features to predict novel miRNAs. Statistically analyze the aligned sRNA sequences for their sequence, length, occurrence frequency, the distribution of the first nucleotide at different miRNA lengths, and the nucleotide distribution at each position for all miRNAs.

2.1.7 Small RNA Annotation

Summarize the alignment and annotation of all small RNAs with various types of RNAs. Since a single sRNA can match multiple different annotation categories, to ensure that each unique small RNA is assigned to only one annotation, follow the priority order: known miRNA > rRNA > tRNA > snRNA > snoRNA > repeats > genes > NAT-siRNA > novel miRNA.

2.1.8 miRNA Base Edit

MicroRNA may undergo nucleotide editing at certain positions, leading to changes in the seed sequence and subsequently altering target genes (Wei et al., 2009). By aligning the sRNA sequences from each sample with the detected known and novel mature miRNAs as well as their precursors, potentially mutated miRNAs can be identified.

2.1.9 miRNA Family Analysis

Conduct a family analysis of the detected known and novel miRNAs to explore the presence of their miRNA families in other species. Known miRNAs can be identified

using miFam.dat to determine their family origin, while novel miRNAs can be classified using RFAM to determine their RFAM family.

2.1.10 miRNA Expression and Differential Expression

Statistically analyze the expression levels of known and novel miRNAs in each sample, and normalize the expression levels using TPM (Zhou et al., 2010). TPM= (read count * 1,000,000) / libsize (libsize: total miRNA read count). For samples with biological replicates, use DESeq2 (Love MI et al., 2014) for differential expression analysis between two comparison groups. DESeq2 provides statistical routines to determine differential expression in digital gene expression data using a model based on the negative binomial distribution. For samples without biological replicates, use the edgeR (Robinson MD et al., 2010) TMM algorithm to normalize read count data for analysis.

2.1.11 Target Gene Prediction for Known and Novel miRNA

For animals, use miRanda and RNAhybrid to predict miRNA target genes, taking the intersection as the final targeting result. For plants, use psRobot to predict miRNA target genes. Perform target gene prediction for the analyzed known and novel miRNAs to obtain the relationships between miRNAs and their target genes.

2.1.12 Enrichment Analysis

We used clusterProfiler (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.

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.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25. (Bowtie)

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40(1):37-52. doi:10.1093/nar/gkr688. (miRDeep2)

Wen M, Shen Y, Shi S, Tang T. miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. BMC Bioinformatics. 2012;13:140. Published 2012 Jun 21. doi:10.1186/1471-2105-13-140.

Wei Y, Chen S, Yang P, Ma Z, Kang L. Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. Genome Biol. 2009;10(1):R6.

doi:10.1186/gb-2009-10-1-r6.

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. (DESeq2)

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.