1.Sample collection and preparation

1.1 RNA quantification and qualification

RNA degradation and contamination, especially DNA contamination, was monitored on 1.5% agarose gels. RNA concentration and purity was measured using the NanoDrop 2000 Spectrophotometer (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 lncRNA sequencing

A total amount of 1.5 μg RNA per sample was used as input material for rRNA removal using the Ribo-Zero rRNA Removal Kit (Epicentre, Madison, WI, USA). Sequencing libraries were generated using NEBNext^{R} Ultra^{TM} Directional RNA Library Prep Kit for Illumina^{R} (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, 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 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 insert fragments of preferentially 150~200 bp in length, the library fragments were purified with AMPure XP Beads (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 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 and qPCR.

1.3 Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kitv3-cBot-HS(Illumia) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina platform and reads were generated.

2.Data analysis

2.1 Quality control

Raw data (raw reads) of fastq format were firstly processed through in-houseperl 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 and GC-content of the clean data were calculated. All the downstream analyses were based on clean data with high quality.

2.2 Gene analysis

2.2.1 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 were further analyzed and annotated based on the reference genome. HISAT2 soft were used to map with reference genome.

2.2.2 Gene functional annotation

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

2.2.3 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.

2.2.4 Quantification of gene expression levels

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

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

2.2.5 Differential expression analysis

1)For the samples with biological replicates:

Differential expression analysis of two conditions/groups was performed using the DESeq R package. DESeq provide statistical routines for determining differential expressionin digital gene expression data using a model based on the negative binomial distribution. The resulting Pvalues 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 (2010) 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.

2.2.6 GO enrichment analysis

Gene Ontology (GO) enrichment analysis of the differentially expressed genes (DEGs) was implemented by the clusterProfiler R packages.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.

2.2.7 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 clusterProfiler R packages to find KEGG pathway that are significantly enriched compared to the entire genome background.

2.2.8 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).

2.3 lncRNA analysis

2.3.1 lncRNA identification

The transcriptome was assembled using the StringTie based on the reads mapped to the reference genome. The assembled transcripts were annotated using the gffcompare program.The known lncRNAs were differentiated from the assembled transcripts if the sequencing species has known lncRNA annotations.

The remaining transcripts(unknown transcripts) were used to screen for putative lncRNAs. Three computational approaches include CPC/CNCI/Pfam/CPAT were combined to sort non-protein codingRNA candidates from putative protein-coding RNAs in the unknown transcripts. Putative protein-coding RNAs were filtered out using aminimum length and exon number threshold. Transcripts with lengths more than 200 nt and have more than two exons were selected as lncRNA candidates and further screened using CPC/CNCI/Pfam/CPAT that have the power to distinguish the protein-coding genes from the non-coding genes. As well as the different types of lncRNAs include lincRNA, intronic lncRNA, anti-sense lncRNA ,sense lncRNA were selected using gffcompare.

2.3.2 Quantification of lncRNA expression levels

StringTie (1.3.1) was used to calculate FPKMs of lncRNAs .

2.3.3 Differential expression analysis

As mentioned above.

2.3.4 Target gene functional annotation

Gene function was annotated based on the following databases:

NR(non-redundant protein sequence database);

Swiss-Prot(A manually annotated, non-redundant protein sequence database);GO(Gene Ontology database);

COG(The database of Clusters of Orthologous Groups of proteins);

KOG(The database of Clusters of Protein homology);

Pfam(The database of Homologous protein family);

KEGG(The database of Kyoto Encyclopedia of Genes and Genomes).

2.3.5 Target gene functional enrichment analysis

1) GO enrichment analysis

Gene Ontology(GO) enrichment analysis of the target gene of differentially expressed lncRNAs was implemented by the clusterProfiler R packages.

2) KEGG pathway enrichment analysis

We used clusterProfiler R packages to find KEGG pathway that are significantly enriched compared to the entire genome background.

2.3.6 Target gene PPI (Protein Protein Interaction)

As mentioned above.

2.4 CircRNA analysis

2.4.1 CircRNA identity

We can used CIRI (CircRNA Identifier) tools and find_circ software to identification circRNA.

CIRI scans SAM files twice and collects sufficient information to identify and characterize circRNAs. Briefly, during the first scanning of SAM alignment, CIRI detects junction reads with PCC signals that reflect a circRNA candidate.Preliminary filtering is implemented using paired-end mapping (PEM) and GT-AG splicing signals for the junctions. After clustering junction reads and recording each circRNA candidate,CIRI scans the SAM alignment again to detect additional junction reads and meanwhile performs further filtering to eliminate false positive candidates resulting from incorrectly mapped reads of homologous genes or repetitive sequences. Finally, identified circRNAs are output with annotation information.

The find_circ software will first be able to take 20 bp from both ends of the reads on the genomic alignment as anchor points, and then the anchor points as independent reads mapped to the reference genome and find the unique matching site. If the alignment position of the two anchor points is reversed in the linear direction, the reading of the anchor point is extended until the joint position of the circular RNA is found. When the signal is spliced for GT/AG, it is judged to be a circular RNA.

The intersect of the results of the two methods will be the final prediction result.

2.4.2 CircRNA target prediction

We can use miRanda(anmial), RNAhybrid (animal),targetscan(animal) and TargetFinder(plant) tools to predict target miRNA. The input files are miRNA and circRNA FASTA sequences files.

2.4.3 Quantification of circRNA expression levels

The expression of circRNA were determined by the number of junction reads identify by CIRI tools and find_circ software.

2.4.4 Differential expression analysis

As mentioned above.

2.4.5 CircRNA-host gene functional annotation

As mentioned above.

2.4.6 Target gene functional enrichment analysis

1) GO enrichment analysis

As mentioned above.

2) KEGG pathway enrichment analysis

As mentioned above.