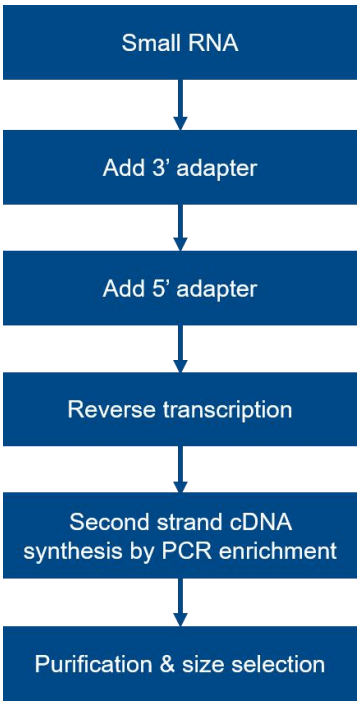
Methods

Sample Quality Control

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

Library Construction, Quality Control and Sequencing

Briefly, 3’ and 5’ adaptors were ligated to 3’ and 5’ end of small RNA, respectively. Then the first strand cDNA was synthesized after hybridazition with reverse transcription primer. The double-stranded cDNA library was generated through PCR enrichment. After purification and size selection, libraries with insertions between 18~40 bp were ready for sequencing on Illumina sequencing with SE50.



The library was checked with Qubit and real-time PCR for quantification and bioanalyzer for size distribution detection. Quantified libraries will be pooled and sequenced on Illumina platforms, according to effective library concentration and data amount required.

Data analysis

- **Quality control**
  Raw data (raw reads) of fastq format were firstly processed through custom perl and python scripts. In this step, clean data (clean reads) were obtained by removing reads containing ploy-N, with 5' adapter contaminants, without 3' adapter or the insert tag, containing ploy A or T or G or C and low quality reads from raw data. At the same time, Q20, Q30, and GC-content of the raw data were calculated. Then, chose a certain range of length from clean reads to do all the downstream analyses.

- **Reads mapping to the reference sequence**
  The small RNA tags were mapped to reference sequence by Bowtie (Langmead et al., 2009) without mismatch to analyze their expression and distribution on the reference.

- **Known miRNA alignment**
  Mapped small RNA tags were used to looking for known miRNA. miRBase20.0 was used as reference, modified software mirdeep2 (Friedlander et al., 2011) and srna-tools-cli were used to obtain the potential miRNA and draw the secondary structures. Custom scripts were used to obtain the miRNA counts as well as base bias on the first position of identified miRNA with certain length and on each position of all identified miRNA respectively.

- **Remove tags from these sources**
  To remove tags originating from protein-coding genes, repeat sequences, rRNA, tRNA, snRNA, and snoRNA, small RNA tags were mapped to RepeatMasker, Rfam database or those types of datas from the specified species itself.

- **Novel miRNA prediction**
  The characteristics of hairpin structure of miRNA precursor can be used to predict novel miRNA. The available software miREvo (Wen et al., 2012) and mirdeep2 (Friedlander et al., 2011) were integrated to predict novel miRNA through exploring the secondary structure, the Dicer cleavage site and the minimum free energy of the small RNA tags unannotated in the former steps. At the same time, custom scripts were used to obtain the identified miRNA counts as well as base bias on the first position with certain length and on each position of all identified miRNA respectively.

- **Small RNA annotation summary**
  Summarizing all alignments and annotations obtained before. In the alignment

and annotation before, some small RNA tags may be mapped to more than one category. To make every unique small RNA mapped to only one annotation, we follow the following priority rule: known miRNA > rRNA > tRNA > snRNA > snoRNA > repeat > gene > NAT-siRNA > gene > novel miRNA > ta-siRNA. The total rRNA proportion was used a marker as sample quality indicator. Usually it should be less than 60% in plant samples and 40% in animal samples as high quality.

- **miRNA editing analysis**

  Position 2~8 of a mature miRNA were called seed region which were highly conserved. The target of a miRNA might be different with the changing of nucleotides in this region. In our analysis pipeline, miRNA which might have base edit could be detected by aligning all the sRNA tags to mature miRNA, allowing one mismatch.

- **miRNA family analysis**

  Exploring the occurrence of miRNA families identified from the samples in other species. In our analysis pipeline, known miRNA used miFam.dat (http://www.mirbase.org/ftp.shtml) to look for families; novel miRNA precursor was submitted to Rfam (http://rfam.sanger.ac.uk/search/) to look for Rfam families.

- **Target gene prediction**

  Predicting the target gene of miRNA was performed by psRobot_tar in psRobot (Wu et al., 2012) for plants or miRanda (Enright et al, 2003) for animals.

- **Quantification of miRNA**

  miRNA expression levels were estimated by TPM (transcript per million) through the following criteria (Zhou et al., 2010):
  Normalization formula: Normalized expression = mapped readcount/Total reads*1000000

- **Differential expression of miRNA**

  For the samples with biological replicates:
  Differential expression analysis of two conditions/groups was performed using the DESeq R package (1.8.3). The P-values was adjusted using the Benjamini & Hochberg method. Corrected P-value of 0.05 was set as the threshold for significantly differential expression by default.

  For the samples without biological replicates:
  Differential expression analysis of two samples was performed using the DEGseq (2010) R package. P-value was adjusted using qvalue (Storey et al, 2003). qvalue < 0.01 and $|\log_2(\text{foldchange})| > 1$ was set as the threshold for significantly differential expression by default.

- **GO and KEGG enrichment analysis**

  Gene Ontology (GO) enrichment analysis was used on the target gene candidates of differentially expressed miRNAs ("target gene candidates" in the following). GOseq based Wallenius non-central hyper-geometric distribution（Young et al, 2010）, which could adjust for gene length bias, was implemented for GO 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 KOBAS (Mao et al., 2005) software to test the statistical enrichment of the target gene candidates in KEGG pathways.

References

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, 10(3), R25. (Bowtie)

Friedlander M.R., Mackowiak S.D., Li N., Chen W., Rajewsky N. (2011). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40:37-52. (miRDeep2)

Wen M., Shen Y., Shi S., and Tang T. (2010). miREvo: An Integrative microRNA Evolutionary Analysis Platform for Next-generation Sequencing Experiments. BMC Bioinformatics 13:140. (miREvo)

Wu HJ, Ma YK, Chen T, Wang M, Wang XJ (2012) PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res 40:W22–W28.(psRobot)

Enright AJ, John B, Gaul U, Tuschl T, Sander C, et al. (2003) MicroRNA targets in Drosophila. Genome Biol 5: R1.(miRanda)

Zhou L., Chen J., Li Z., Li X., Hu X., et al. (2010). Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. PLoS One 5: e15224. (TPM)

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value, Annals of Statistics. 31: 2013-2035.(qvalue)

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A(2010). goseq: Gene Ontology testing for RNA-seq datasets.(goseq)

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids research36:D480–484. (KEGG)

Mao, X., Cai, T., Olyarchuk, J.G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG orthology (KO) as a controlled vocabulary. Bioinformatics 21, 3787–3793.(KOBAS)

Wang L., Feng Z., Wang X., Wang X., Zhang X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-8.(DEGseq)

Anders, S., Huber, W. (2010).Differential expression analysis for sequence count data. Genome Biology,doi:10.1186/gb-2010-11-10-r106. (DESeq)