REPORT

weSEQ.iT

PROJECT INFORMATION

Client:
Institute:
Project: Library Preparation & small RNA Sequencing
Platform: Illumina
Bioinformatics Service: yes
Number of Samples: DEMO
Date:

Results

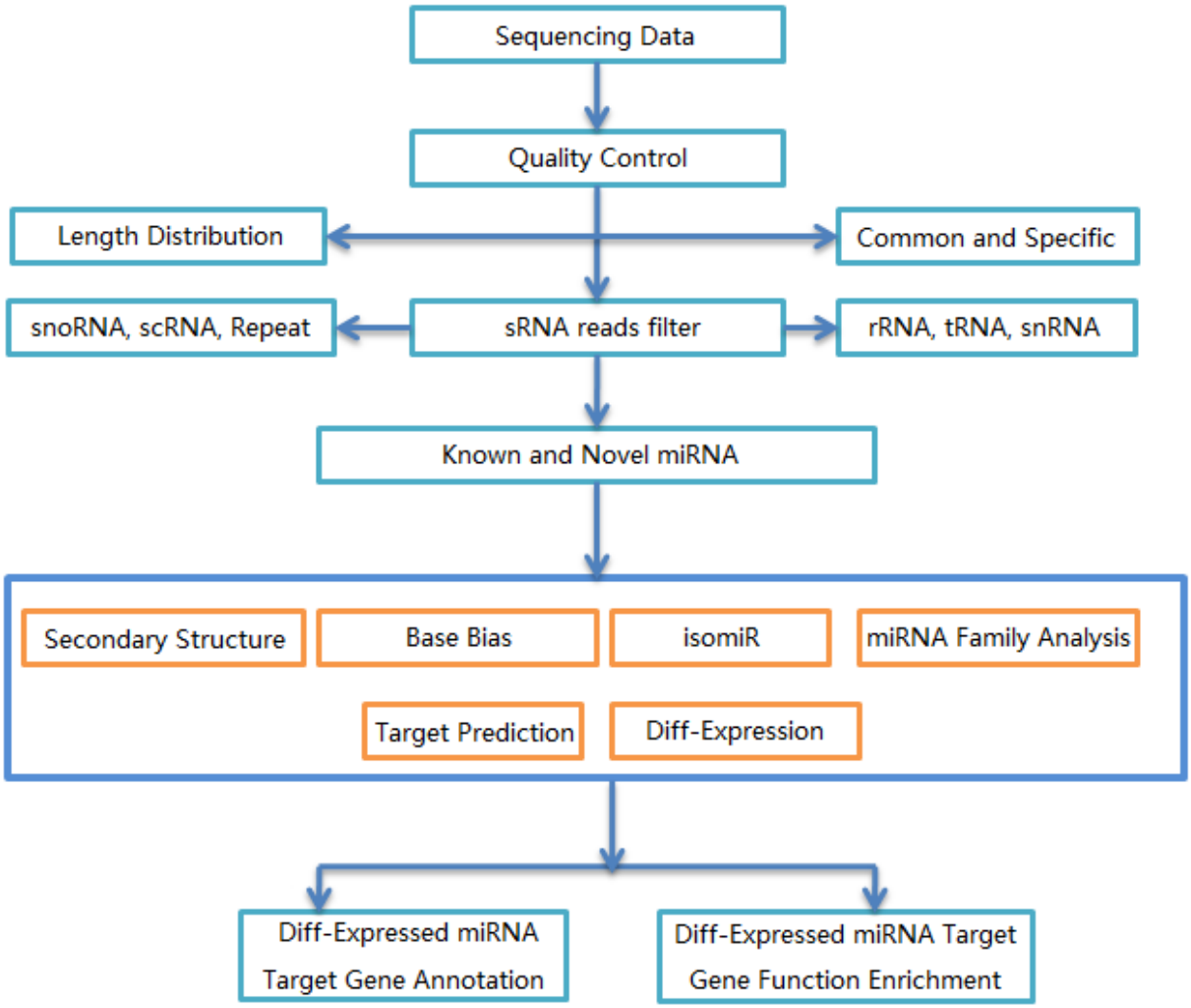
1. Experimental workflow

Items in contract:
(1) A total of 6 samples are to be processed for small RNA sequencing. Sequencing quality of all samples should reach Q30≥85%.
(2) Identification of known miRNA and prediction of novel miRNA.
(3) Quantification of miRNA expression and identification of differentially expressed miRNA.
(4) Prediction of miRNA target genes.
(5) Functional annotation and enrichment analysis on target genes of differentially expressed miRNA.
Summary of outputs:
(1) A total of 6 samples were processed for small RNA sequencing, generating 0.56 M Clean Reads (Minimum of 0.09 M Clean Reads in each sample).
(2) 346 miRNA were identified in total containing 99 known miRNAs and 247 novel miRNAs.
(3) Expression of miRNA in each samples were quantified and differentially expressed miRNAs between given groups were identified.
(4) A total of 8,148 miRNA target genes were identified. Those of differentially expressed miRNAs were annotated and processed for enrichment analysis.

2. Bioinformatics Analysis

2.1 Summary of Bioinformatics Analysis

Bioinformatic analysis scheme for small RNA sequencing is shown in the figure below. (For projects with single sample, analysis on common and unique sequences and differential expression are not included.)



2.2 Sequencing Data and Quality Control

In next generation sequencing, bases are inferred from light intensity signals generated by Illumina sequencing platform, which is known as base calling. Data generated directly from base calling is referred to as Raw data or Raw reads. Raw data was normally provided in FASTQ format, containing sequences and corresponding quality information. A demo FASTQ file is shown as below.

```
@HWI-D00621:162:HFMGFADXX:2:2215:14635:63583 1:N:0:GAGTGG
TTTTCCGTCTGATTTCCATATGAGATCGGAAGAGCACACGTCTGAACTCCA
+
??;B;ADDHHFFHIG<EHIH>H?@<CFDBA7CFDDEF>FEH=FAGIG>>B<
```

Figure. Demo Fastq format

Note: In FASTQ file, each sequence consists of 4 lines:
(1)The first line begins with @ and is followed by sequence ID and an optional description.
(2)The second line is a series of single letters representing sequence, i.e. reads.
(3)The third line begins with + and optional description.
(4)The last line is the corresponding quality value of the bases in the second line. The length of this line should be exactly the same as Line 2.

2.2.1 Sequencing bases quality score

Quality Score or Q-score represents the probability of an incorrect base. This Phred quality score is defined as following equation [1]:

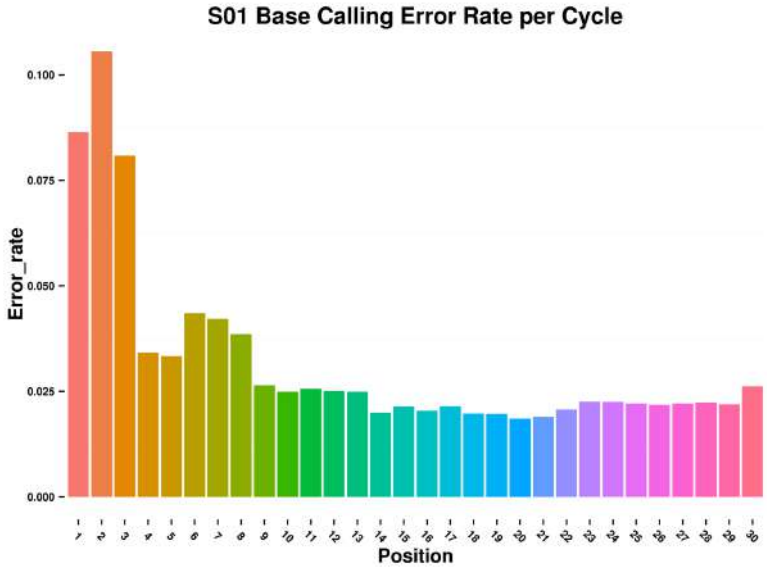
$$Q = -10 * log_{10}P$$

In the equation, P stands for the base calling error probabilities. Following table shows the relations between quality score and base calling accuracy:

Table. Quality score and base calling accuracy

Phred_Base_quality	Error_probability	Detection_accuracy
Q10	10%	90%
Q20	1%	99%
Q30	0.1%	99.9%
Q40	0.01%	99.99%

A high Q-score indicates lower error rate in base calling, i.e. higher accuracy. As shown in the table above, Q20 means that only one base calling in 100 is predicted to be incorrect. Q30 means that only one base calling in 1000 is predicted to be incorrect. Q40 means that only one base calling in 10,000 is predicted to be incorrect.
In order to describe the quality of sequences, base calling error is calculated for each sequencing reaction cycle. Base calling error rate of each sample were shown in the figures below.



Rate of error basing calling is influenced by the instrument, reagents, samples, etc. It is commonly found in RNA-seq that:
(1)The rate slowly climbs along the reading of sequence due to the consumption of reagents. It is commonly observed on Illumina sequencing platform.
(2)The error rate at first six bases of reads are normally higher, which is caused by inefficient binding between random hexamer primers and RNA templates.

2.2.2 Sequencing Data Assessment

Adapter and low-quality sequences in raw data need to be removed in order to ensure the reliability of downstream analysis. In prior to bioinformatic analysis, we processed very strict quality control to extract clean data from raw data. Detailed processes are listed below.
(1) Remove low quality sequences from each sample;
(2) Remove reads containing more than 10% N (unknown bases);
(3) Remove reads without 3' adapter sequences;
(4) Cut 3' adapter sequences from raw reads;
(5) Remove reads with length smaller than 18 or longer than 30 nt.
Summary of sequencing data was shown in the following table.

Table. Statistics on sequencing data

Samples	ID	Raw_reads	Length<18	Length>30	Low_quality	Containing'N'reads	Clean_reads	Q30(%)
S01	S01	114,359	1,241	16,882	0	0	96,236	90.60
S02	S02	106,104	1,186	12,983	0	0	91,935	90.18
S03	S03	105,942	2,350	14,458	0	0	89,134	92.79
S04	S04	103,792	1,096	7,609	0	0	95,087	95.37
S05	S05	102,384	1,207	6,636	0	0	94,541	95.53
S06	S06	104,365	1,322	7,314	0	0	95,729	95.45

Minimum of 0.09 M Clean reads were generated for each sample.

2.3 sRNA Classification

2.3.1 Annotation on ncRNA and Repeated Sequences

Bowtie [2] is a software designed for aligning high-throuput sequencing reads against sequences in database. Clean reads were mapped to Silva, GtRNAdb, Rfam and Repbase to remove ncRNAs including rRNA, tRNA, snRNA, snoRNA, etc. and repeated sequences. The rest unannotated reads were regarded as reads containing miRNAs. Statistics on sRNA classification was shown in the table below.

Table. Statistics on sRNA classification

S01.Data.stat.html
S02.Data.stat.html
S03.Data.stat.html
S04.Data.stat.html
S05.Data.stat.html
S06.Data.stat.html

2.3.2 Reference Genome Mapping

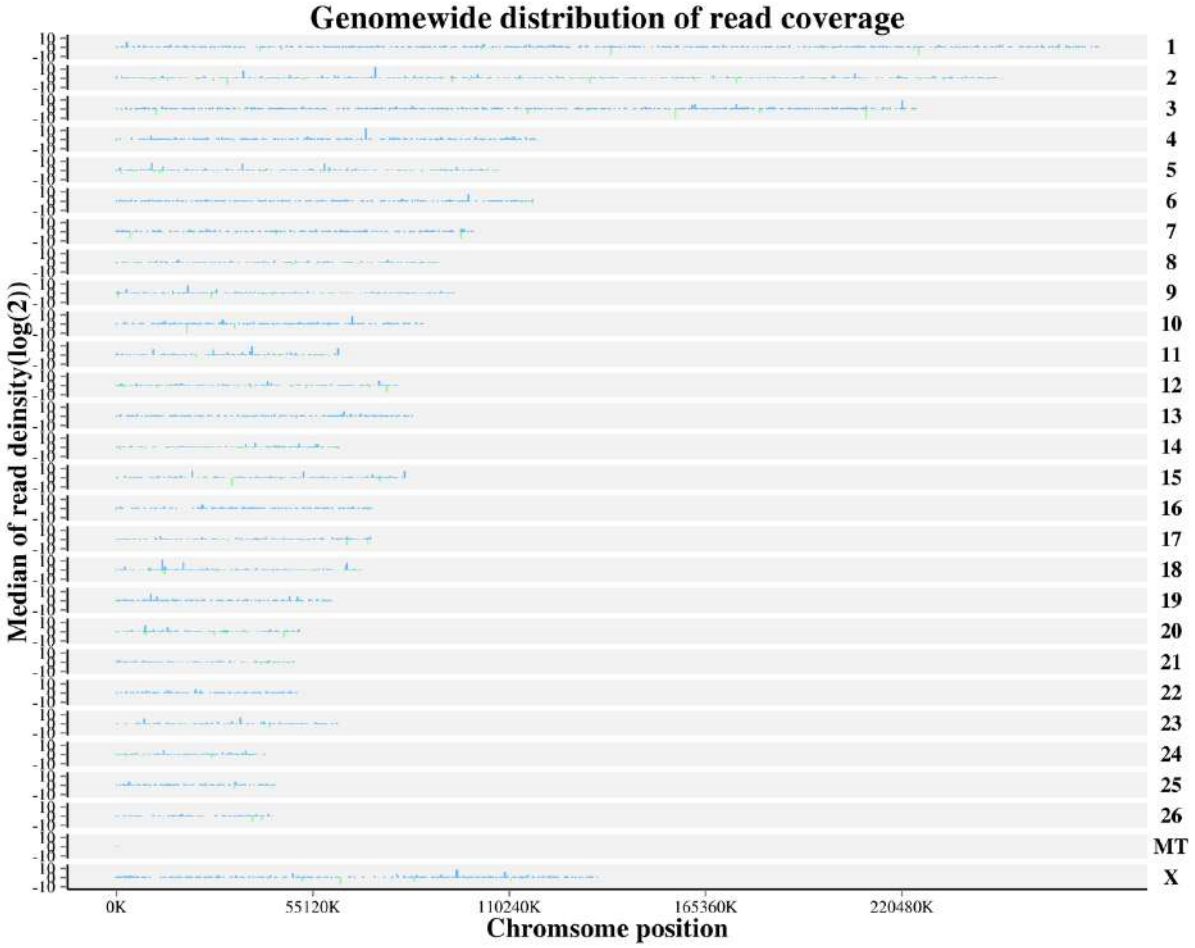
Reads were mapped to specified reference genome: Ovis_aries.Oar_v3.1.

Unannotated reads were mapped to reference genome with Bowtie to obtain their positions on reference genome. Reads with position are called mapped reads. Statistics on mapping was shown in the table below.

Table. Statistics on mapping against reference genome

#ID	Total_Reads	Mapped_Reads	Mapped_reads(+)	Mapped_reads(-)
S01	95,195	65,963(69.29%)	44,987(47.26%)	20,976(22.03%)
S02	90,973	63,925(70.27%)	41,139(45.22%)	22,786(25.05%)
S03	88,203	60,147(68.19%)	40,736(46.18%)	19,411(22.01%)
S04	94,158	72,907(77.43%)	55,464(58.91%)	17,443(18.53%)
S05	93,859	73,638(78.46%)	55,420(59.05%)	18,218(19.41%)
S06	95,065	73,603(77.42%)	56,240(59.16%)	17,363(18.26%)

Distribution of mapped reads on chromosome was calculated to generate a diagram showing overall coverage depth across chromosomes in reference genome. Distribution of reads on chromosomes were shown in the following figures.



2.4 miRNA Analysis

2.4.1 miRNA Identification

Known miRNAs were identified by comparing mapped reads with mature miRNA in miRBase(v22) database. Mature miRNA sequences with 2 nt up-stream and 5 nt down-stream were used in searching. Mapped reads with maximum 1 mis-match were regarded as matching to known miRNA.
MiRNA transcription start sites are more frequently found in intergenic regions, introns and reverse strand of coding regions. miRNA genes are firstly transcribed into primary miRNA (pri-miRNA) and processed into precursor miRNA (pre-miRNA), which is characterized by its hair-pin structure, and finally matured into miRNA with help of Dicer/DCL enzyme. The remaining reads were analyzed by miRDeep2 [3] to predict novel miRNAs based on specific species.
In miRDeep2 modules, potential miRNA precursors were extracted from reference genome based on reads mapping. A further selection of potential precursors counts on RNA secondary structure, where candidate precursors are expected to be able to partitioned into candidate mature, loop and star part

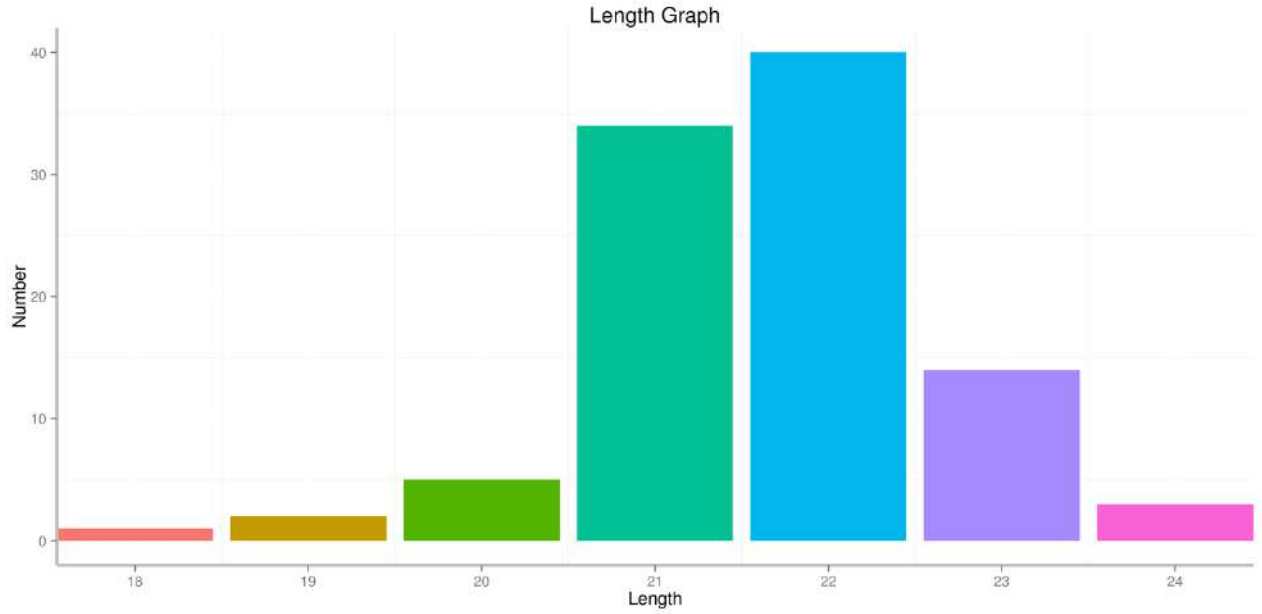
based on reads mapping. RNAfold randfold P-value will be given to a subset of potential precursors. Each precursor will be scored by Bayesian statistics to describe the fit of reads to the biological model of miRNA biogenesis. This software is mainly designed for animal miRNA prediction, however, by adjusting parameters and algorithm, plant miRNA prediction can also be achieved [4].
A total of 346 miRNA were identified in all samples, in which 99 were known miRNA and 247 were novel miRNA. Detailed summary of miRNA identification was shown in the following table.

Table 5. miRNA identification in samples

ID	Known-miRNAs	Novel-miRNAs	Total
S01	69	206	275
S02	67	198	265
S03	68	196	264
S04	82	176	258
S05	75	177	252
S06	79	175	254
Total	99	247	346

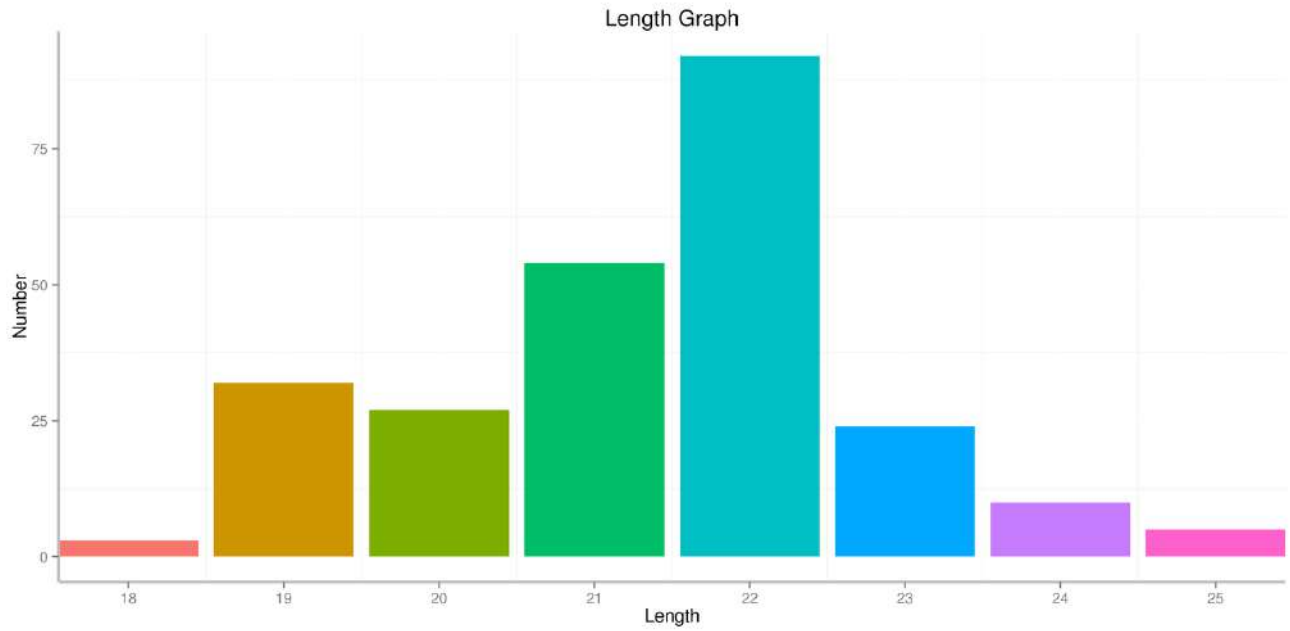
According to specificity of Dicer and DCL enzymes, length of mature miRNAs is mainly concentrated between 20 nt and 24 nt, in which plants miRNAs are normally 21 nt or 24 nt in length, while animal miRNAs are 22 nt in length. Length distribution of known miRNAs, novel miRNAs and total miRNAs were shown in the following figures.

Figure. Length distribution of known miRNAs



Note: X-axis: miRNA length; Y-axis: number of miRNA with corresponding length.

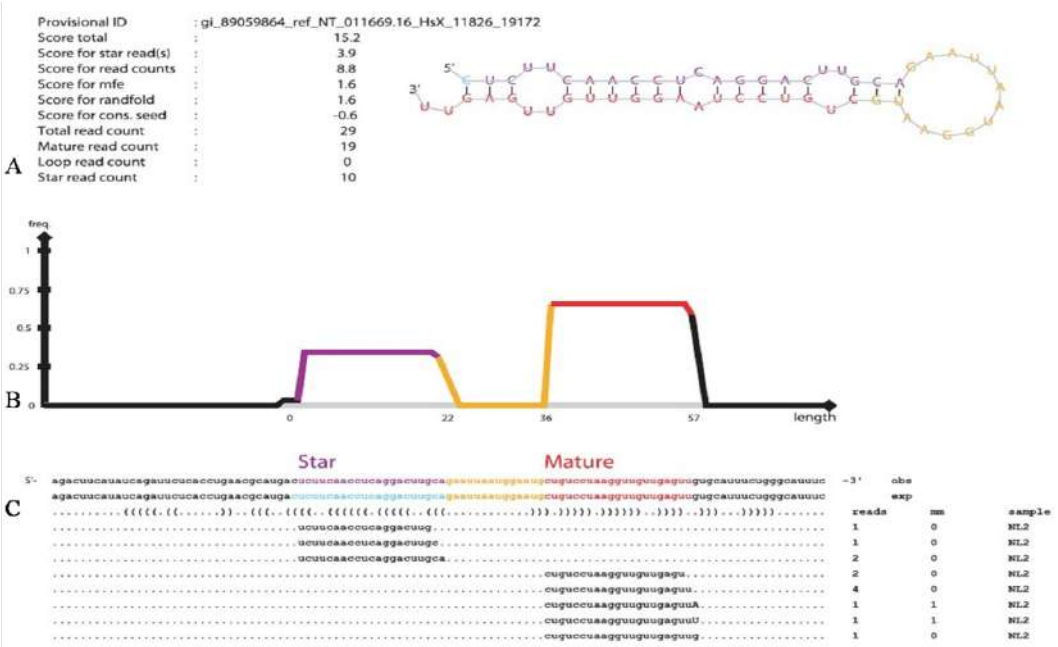
Figure. Length distribution of novel miRNAs



Note: X-axis: miRNA length; Y-axis: number of miRNA with corresponding length.

Each candidate precursor of novel miRNAs predicted by miRDeep2 has a pdf figure showing its structure and sequencing depth. A demo figure is shown below.

Figure. miRNA precursor structure and sequencing depth



3.4.2 miRNA Base Bias

Dicer and DCL enzymes are known to have strong sequence cleavage preference for 5'U. Analysis on miRNA base bias is used to compare that of identified miRNA with typical miRNA. First base preference of miRNA and base preference on all sites were shown in the following figures

Figure. First base preference of miRNA in different length-Known miRNAs

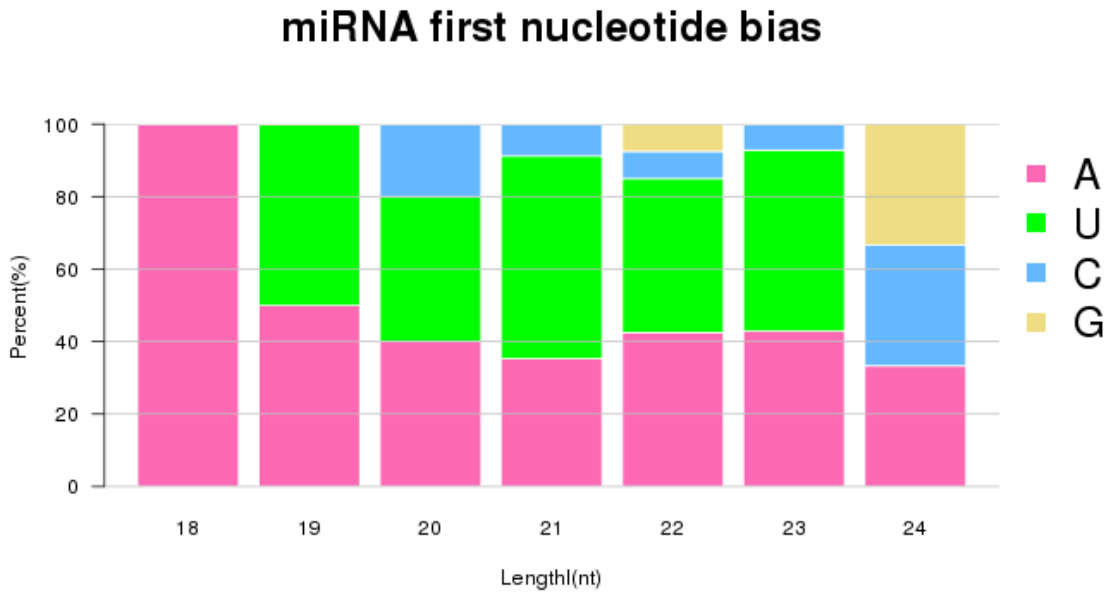


Figure. First base preference of miRNA in different length-Novel miRNAs

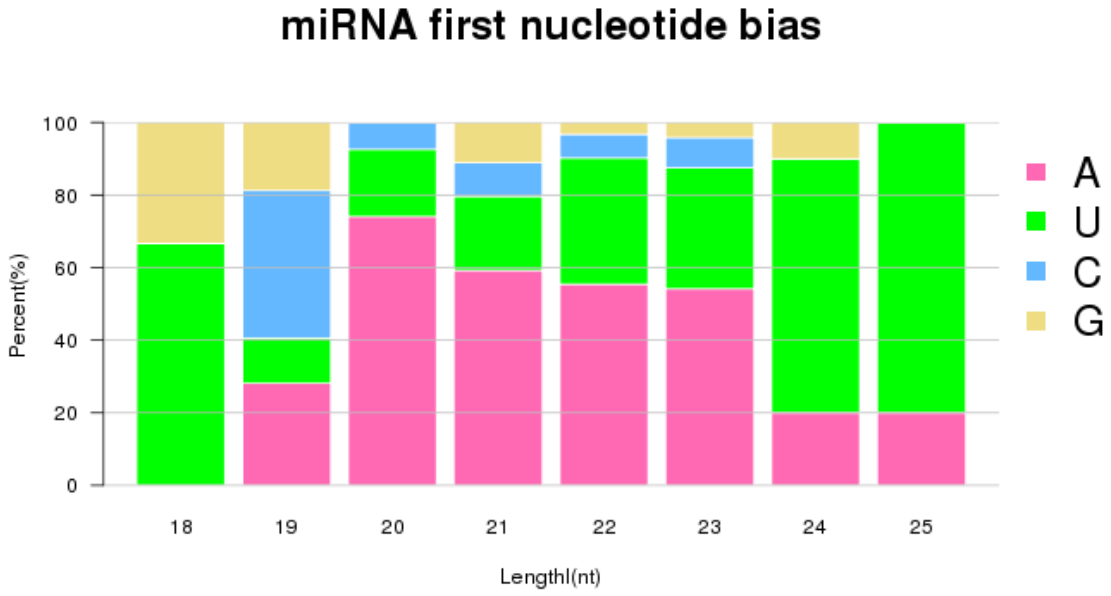


Figure. Base preference on miRNA-Known miRNAs

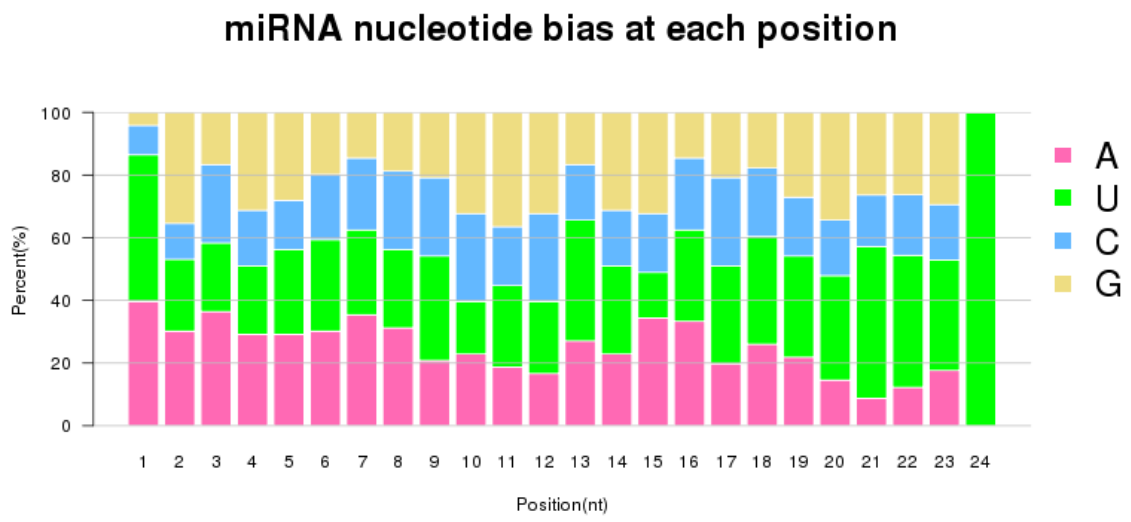
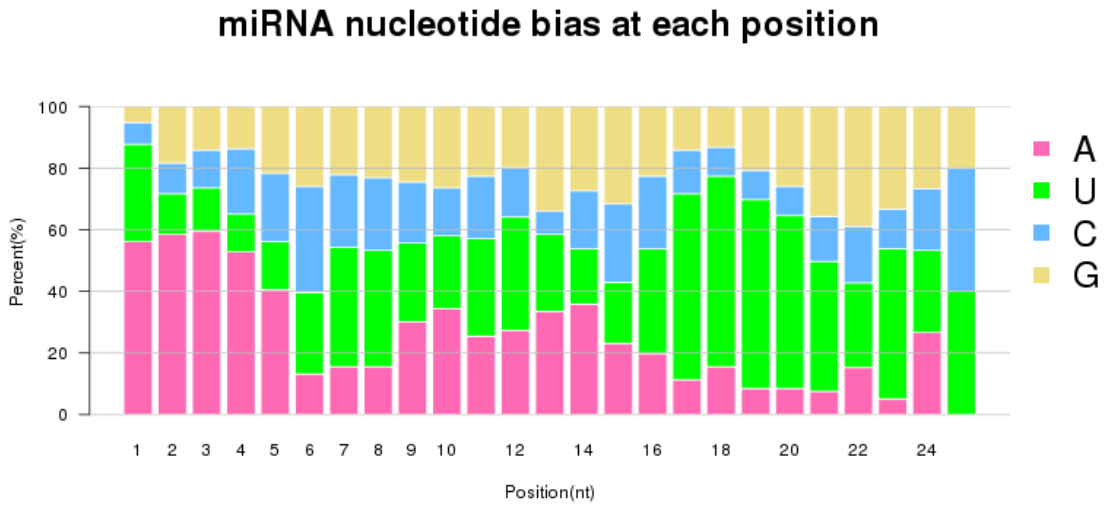


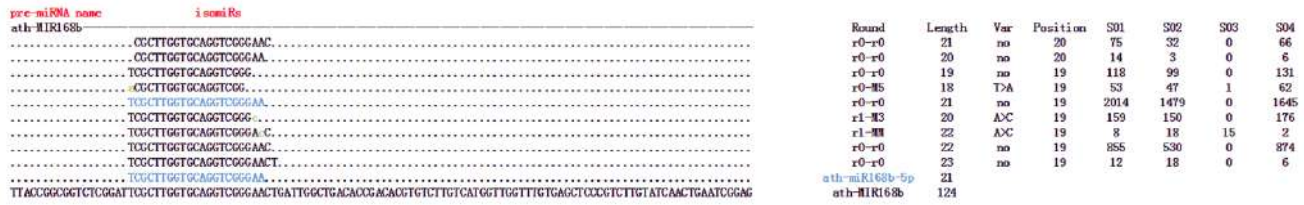
Figure. Base preference on miRNA-Novel miRNAs



3.4.3 miRNA Base Editing

MiRNA could be edited after transcription, resulting in changes in seed sequences. As a consequence, its target gene will change. Edited miRNA was identified by isomiRID. MiRNAs went through two rounds of mapping: Using bowtie to map against precursor sequences (r0), the perfectly matching sequences were regarded as reference sequence in next round; Mapping miRNAs with reference miRNA(r1, one mis-matched base maximum), mis-match on 3' end is marked as M3; mis-match on 5' end is marked as M5; mis-match in the middle is marked as MM. A demo figure on miRNA base editing was shown in the figure below.

Figure. Demo figure on miRNA base editing



pre-miRNA name	isomiRs
ath-MIR168b	
	CGCTTGGTGCAGGTCGGGAAC
	CGCTTGGTGCAGGTCGGGAA
	TCGCTTGGTGCAGGTCGGG
	CGCTTGGTGCAGGTCGG
	TCGCTTGGTGCAGGTCGGGAA
	TCGCTTGGTGCAGGTCGGG
	TCGCTTGGTGCAGGTCGGGAC
	TCGCTTGGTGCAGGTCGGGAAC
	TCGCTTGGTGCAGGTCGGGAACT
	TCGCTTGGTGCAGGTCGGGAA
TTACCGGCGGTCTCGGATTCGCTTGGTGCAGGTCGGGAACTGATTGGCTGACACCGACACGTGTCTTGTCATGGTTGGTTTGTGAGCTCCCGTCTTGTATCAACTGAATCGGAG	

Round	Length	Var	Position	S01	S02	S03	S04
r0-r0	21	no	20	75	32	0	66
r0-r0	20	no	20	14	3	0	6
r0-r0	19	no	19	118	99	0	131
r0-M5	18	T>A	19	53	47	1	62
r0-r0	21	no	19	2014	1479	0	1645
r1-M3	20	A>C	19	159	150	0	176
r1-M3	22	A>C	19	8	18	15	2
r0-r0	22	no	19	855	530	0	874
r0-r0	23	no	19	12	18	0	6
ath-miR168b-5p	21						
ath-MIR168b	124						

Note: Column 1: matching between small RNAs and pre-miRNA;
Column 2: Mapping round and editing type;
Column 3: Length of small RNA;
Column 4: Base variation compared to precursor sequence;
Column 5: Position of small RNA on precursor;
The rest columns are number of small RNA in each sample.

3.4.4 miRNA Family

miRNA is highly conserved within species. miRNA family classification and annotation on known and novel miRNAs is based on similarity in sequences. miRNA family annotation was shown in the table below.

Table. miRNA Family annotation

result.txt.html

3.5 miRNA Expression

3.5.1 miRNA Quantification

Expression of miRNAs in each sample was calculated and normalized by TPM algorithm [5]. Equation of TPM normalization is shown below.

$$TPM = \frac{Readcount * 1,000,000}{MappedReads}$$

In the equation, readcount stands for the number of reads mapped to a miRNA; Mapped Reads stands for the number of reads mapped to all miRNAs.
All miRNA expression in each sample was listed in the table below.

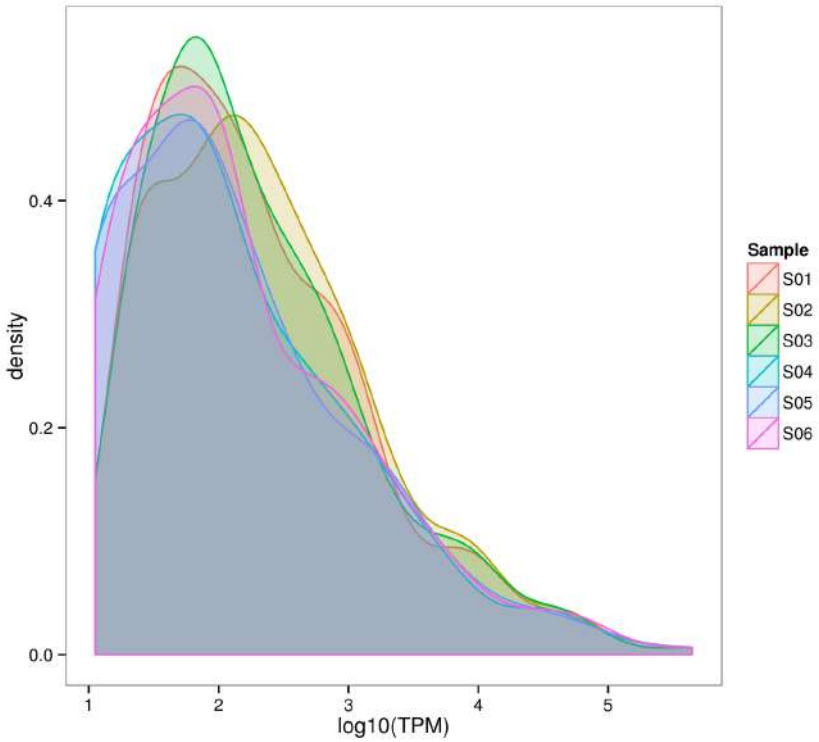
Table 7. All miRNA Expression

All_miRNA_expression.list.html

3.5.2 miRNA Expression Distribution

Distribution of miRNA expression describes overall miRNA expression pattern in each sample. Distribution of TPM density in each sample was shown in the figure below.

Figure. TPM density distribution



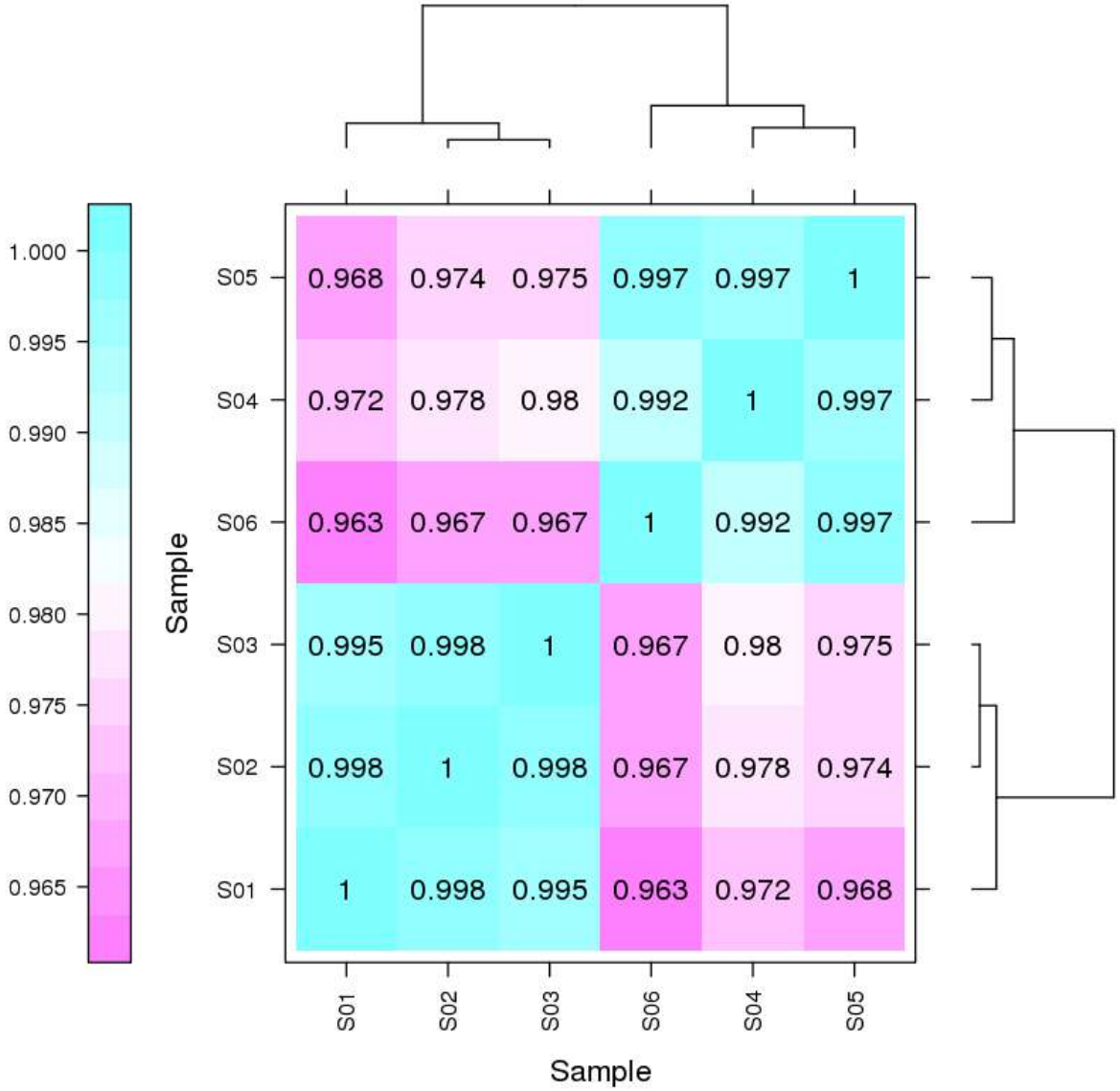
3.5.3 Correlation Assessment between Samples

Correlation coefficient(r) is a statistical measure of how well the relationships between two variables is. Currently, Pearson correlation coefficient and Spearman Correlation Coefficient are the most commonly used coefficient. Spearman's coefficient of two variables is equal to the Pearson correlation between the rank values of the two. In this project, Pearson's was used to calculate correlations between samples. Following equation describes the calculation of correlation.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(x)}\sqrt{E(Y^2) - E^2(Y)}}$$

In the equation, numerator is covariance and denominator is product of standard deviations of two variables. Standard deviation of X and Y cannot be 0. Correlation coefficient closer two 1 or -1 indicates the variables have stronger linear correlation("1": positive correlation; "-1": negative correlation). Correlation of 0 indicates that there is no tendency between two variables. Correlation(r^2) between samples were shown in the following figures.

Figure. Correlation between samples



Note: The color in the figure represents correlation coefficient. X and Y-axis: Samples.

3.6 miRNA Differential Expression Analysis

3.6.1 Differentially Expressed miRNA

Software applied for analyzing differentially expressed miRNAs(DE-miRNAs) should be selected based on practical situations. DESeq2 [6] is designed for differential expression analysis in experiments with biological replicates. edgeR [7] is designed for that without biological replicates. The differential analysis group is named as "A_vs_B". Normally, "A" represents control group, wild type or former time point. "B" normally represents corresponding treated group, mutant or later time point. The

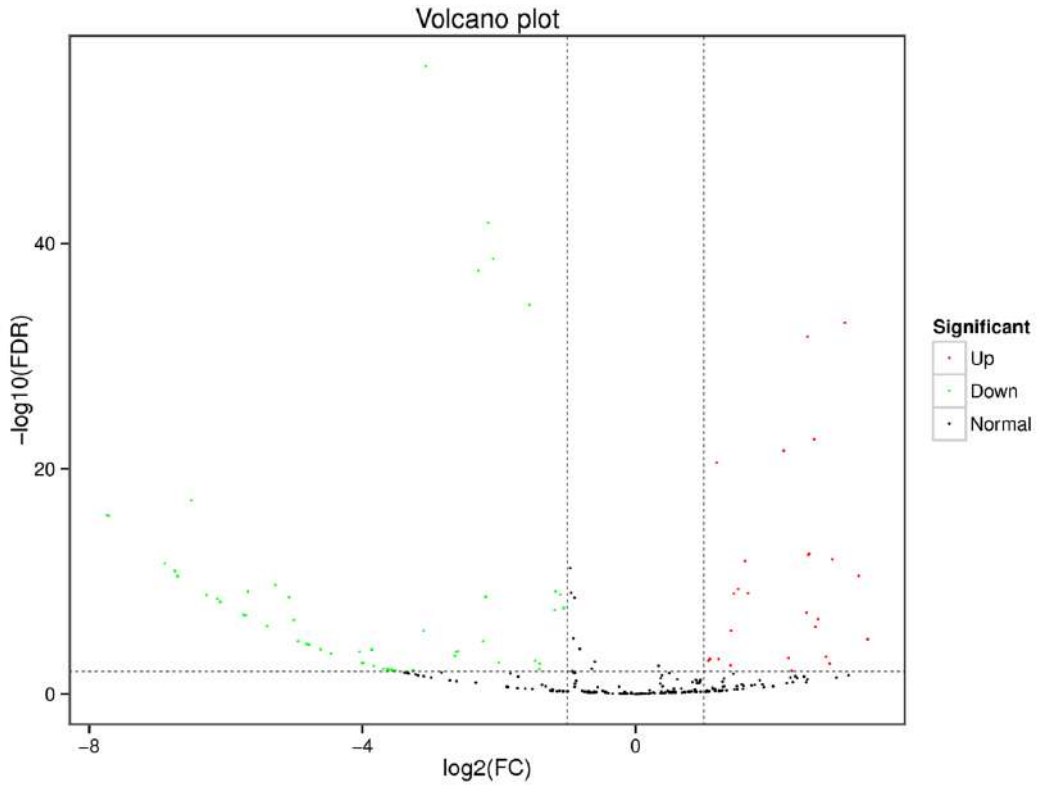
miRNAs with a higher expression level in B than A(B>A) are defined as up-regulated miRNAs. The ones with lower expression level in B(B<="" p="" style="box-sizing: border-box;">
In this project, threshold for defining DE-miRNAs was set as $|\log2(FC)|\geq1.00$;$FDR\leq0.01$. Fold change(FC) refers to ratio in expression between two samples(groups). P-value represents significancy of difference in expression. In order to minimize false-positive events in DE-miRNA identification, Benjamini-Hochberg procedure is required to correct the P-value of significancy test. The corrected P-value, known as False Discovery Rate (FDR) is applied as index for DE-miRNA screening.
In this report, differentially expressed miRNA set is named as "A_vs_B", e.g. DE-miRNAs between sample S01 and S02 is named as "S01_vs_S02".

Summary on DE-miRNAs between samples was shown in the table below.

Table. Summary on DE-miRNAs between samples

DEG Set	DEG Number	up-regulated	down-regulated
S01_S02_S03_vs_S04_S05_S06	91	26	65

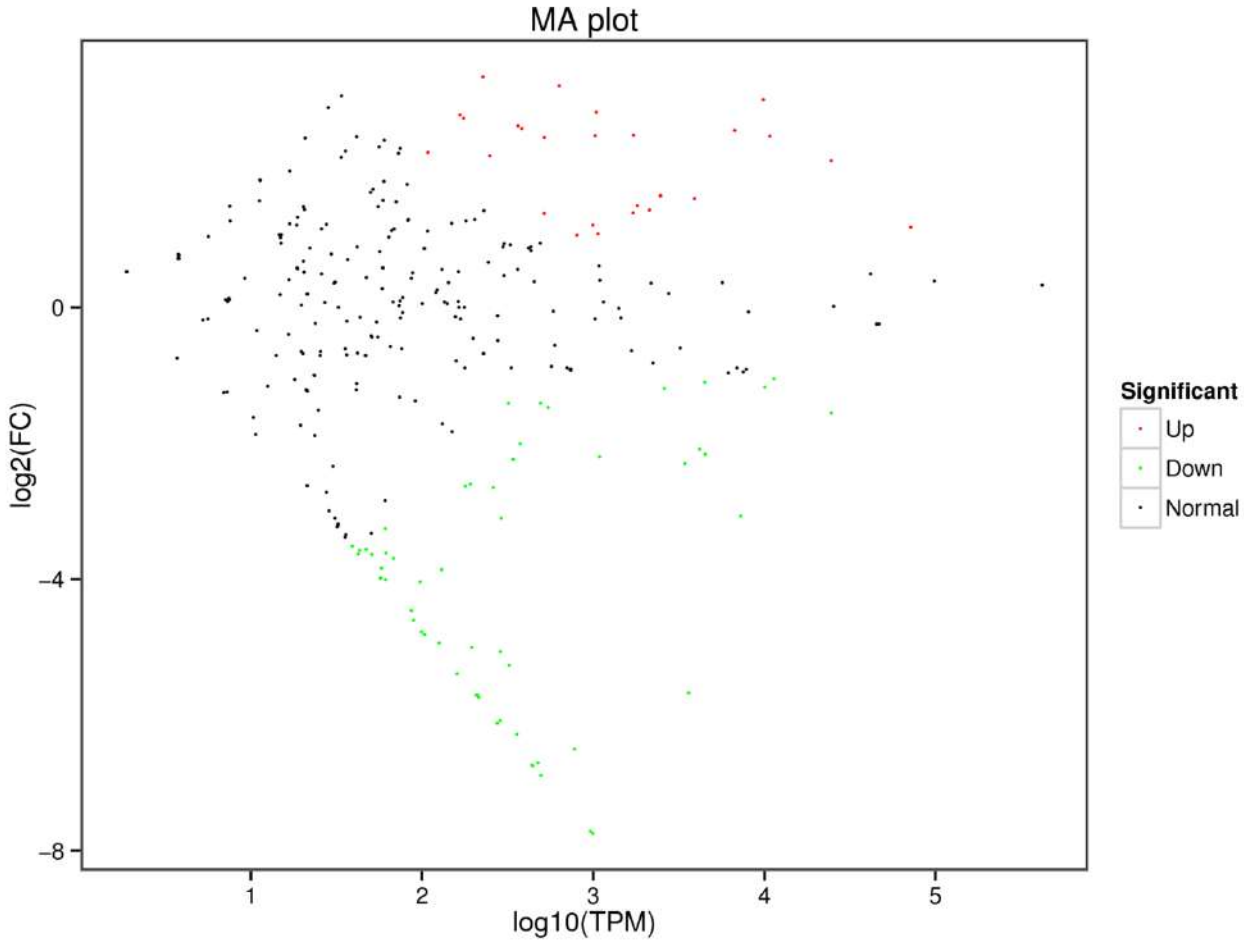
Volcano plot is a plot of log2(Fold change) against log10 FDR, which clearly shows the differences in miRNA expression between two samples and the corresponding significancy. Volcano plots between samples were shown in the following figures.



Note: In the plot, each dot represents a single miRNA. X-axis: log2(Fold change) between two samples; Y-axis: -log10 FDR. A larger absolute value on X-axis represents a larger difference in expression between the two sample; The larger the -log10FDR is, the more reliable the DE-miRNAs are. The dots colored in green are down-regulated miRNAs. Red dots are up-regulated miRNAs. Black dots stands for miRNAs without significant difference in expression between samples.

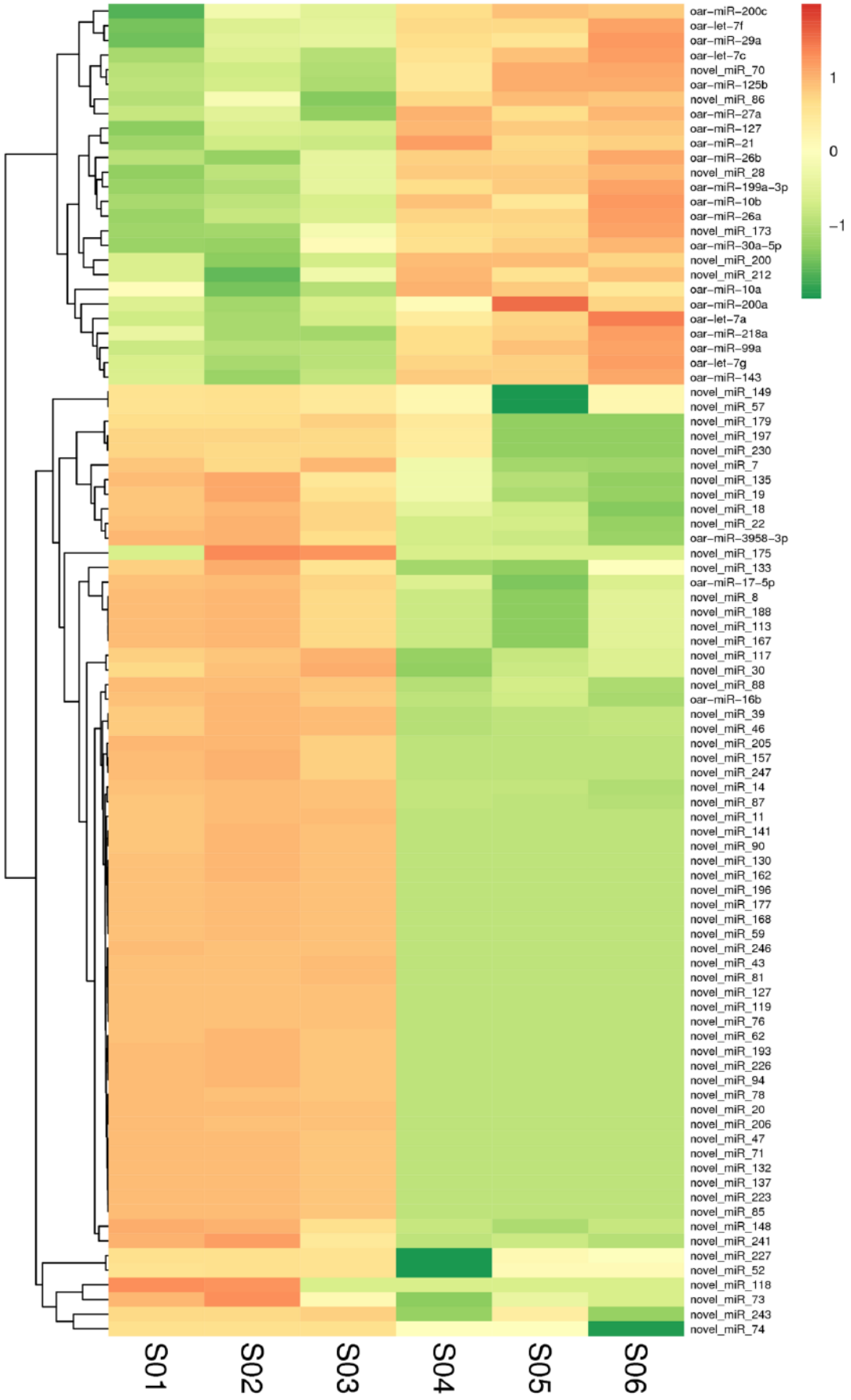
MA plots shows the overall distribution of miRNA expression and fold change of expression level between two samples. MA plot of differentially expressed miRNAs were shown in figures below.

Figure. MA plot on DE-miRNAs



3.6.2 Clustering of DE-miRNAs

Hierarchical clustering analysis was processed on differentially expressed miRNAs, i.e. miRNAs with same or similar expression mode were clustered together. Hierarchical clustering of DE-miRNAs between samples was shown in the figure below.


oar-miR-200c
oar-let-7f
oar-miR-29a
oar-let-7c
novel_miR_70
oar-miR-125b
novel_miR_86
oar-miR-27a
oar-miR-127
oar-miR-21
oar-miR-26b
novel_miR_28
oar-miR-199a-3p
oar-miR-10b
oar-miR-26a
novel_miR_173
oar-miR-30a-5p
novel_miR_200
novel_miR_212
oar-miR-10a
oar-miR-200a
oar-let-7a
oar-miR-218a
oar-miR-99a
oar-let-7g
oar-miR-143
novel_miR_149
novel_miR_57
novel_miR_179
novel_miR_197
novel_miR_230
novel_miR_7
novel_miR_135
novel_miR_19
novel_miR_18
novel_miR_22
oar-miR-3958-3p
novel_miR_175
novel_miR_133
oar-miR-17-5p
novel_miR_8
novel_miR_188
novel_miR_113
novel_miR_167
novel_miR_117
novel_miR_30
novel_miR_88
oar-miR-16b
novel_miR_39
novel_miR_46
novel_miR_205
novel_miR_157
novel_miR_247
novel_miR_14
novel_miR_87
novel_miR_11
novel_miR_141
novel_miR_90
novel_miR_130
novel_miR_162
novel_miR_196
novel_miR_177
novel_miR_168
novel_miR_59
novel_miR_246
novel_miR_43
novel_miR_81
novel_miR_127
novel_miR_119
novel_miR_76
novel_miR_62
novel_miR_193
novel_miR_226
novel_miR_94
novel_miR_78
novel_miR_20
novel_miR_206
novel_miR_47
novel_miR_71
novel_miR_132
novel_miR_137
novel_miR_223
novel_miR_85
novel_miR_148
novel_miR_241
novel_miR_227
novel_miR_52
novel_miR_118
novel_miR_73
novel_miR_243
novel_miR_74
1
0
−1
S01
S02
S03
S04
S05
S06

Note: In the heatmap, each column represents a sample. The expression level of miRNA is normalized to log10(TPM+1e-6) and presented as different colors based on scale bar (Red: higher expression; Green: lower expression).

3.7 miRNA Target Gene

miRNA target genes were predicted based on sequences of known miRNAs, novel miRNAs and gene sequences of corresponding species. TargetFinder [6] is employed for target gene prediction in plants. miRanda [7] and targetscan [8] are used for animals. Summary of miRNA target gene prediction was shown in the following table.

Table. Statistics on predicted miRNA target gene

Types	All_miRNA	miRNA_with_Target	Target_gene
Known_miRNA	99	55	4,479
Novel_miRNA	247	104	6,213
Total	346	159	8,148

Predicted miRNA target genes were listed in the following table.

Table. miRNA Target Genes

oar.mir2target.list.html

3.8 Annotation of miRNA Target Gene

Sequences of target genes were BLAST against NR [11], Swiss-Prot [12], GO [13], COG [14], KEGG [15], KOG [16] and Pfam [17]database to obtain their annotations. In 8,148 target gene, 8,134 of them were annotated. Summary on miRNA target gene annotation analysis was shown in the following table.

Table. Statistics on miRNA target gene annotation

# Anno_Database	Annotated_Number	300<=length<1000	length>=1000
COG_Annotation	2,642	341	2,299
GO_Annotation	3,374	667	2,689
KEGG_Annotation	5,613	1,011	4,575
KOG_Annotation	5,708	929	4,759
Pfam_Annotation	7,611	1,470	6,112
Swissprot_Annotation	6,350	1,277	5,041
eggNOG_Annotation	8,096	1,624	6,425

# Anno_Database	Annotated_Number	300<=length<1000	length>=1000
nr_Annotation	8,134	1,652	6,434
All_Annotated	8,134	1,652	6,434

Table. Statistics on miRNA target gene prediction in each sample

ID	All miRNA	miRNA with Target	Target gene
S01	275	128	5,386
S02	265	126	5,530
S03	264	131	4,988
S04	258	114	6,509
S05	252	112	7,095
S06	254	115	7,498

3.9 DE-miRNA Target Gene Annotation

Summary on annotation of DE-miRNAs target gene was shown in the table below.

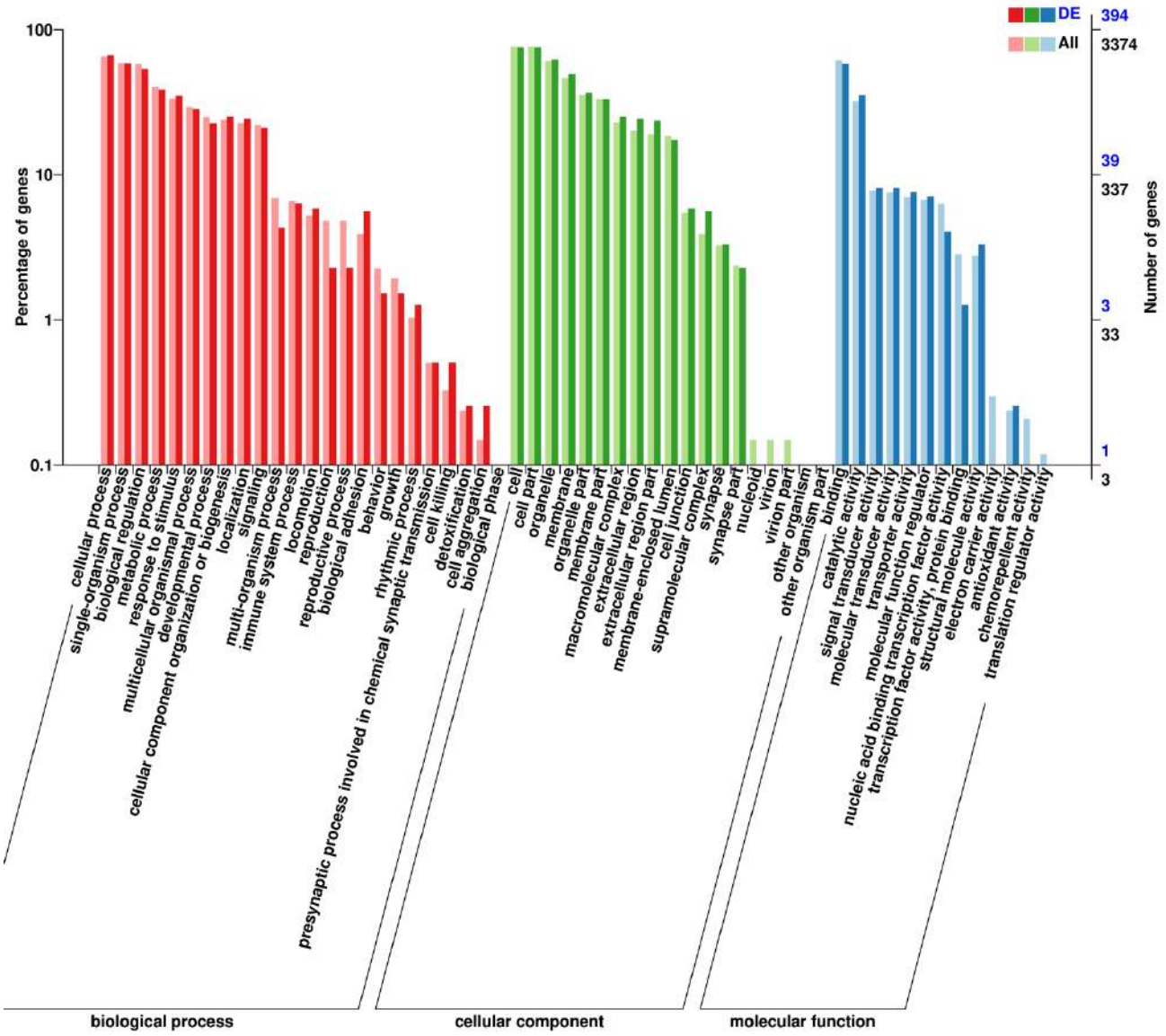
Table. Statistics on DE-miRNA target gene annotation

# DEG Set	Total	COG	GO	KEGG	KOG	NR	Pfam	Swiss-Prot	Egg NOG
S01_S02_S03_vs_S04_S05_S06	936	325	394	653	664	936	878	705	930

3.9.1 GO Classification on DE-miRNA Targeted Genes

GO (Gene Ontology) database is a structured biological annotation system established in 2000 containing a standard vocabulary of gene and gene products functions, which is applicable in all species.
GO classification of DE-miRNA targeted genes between samples was shown in the following figure.

Figure. GO classification of DE-miRNA targeted genes

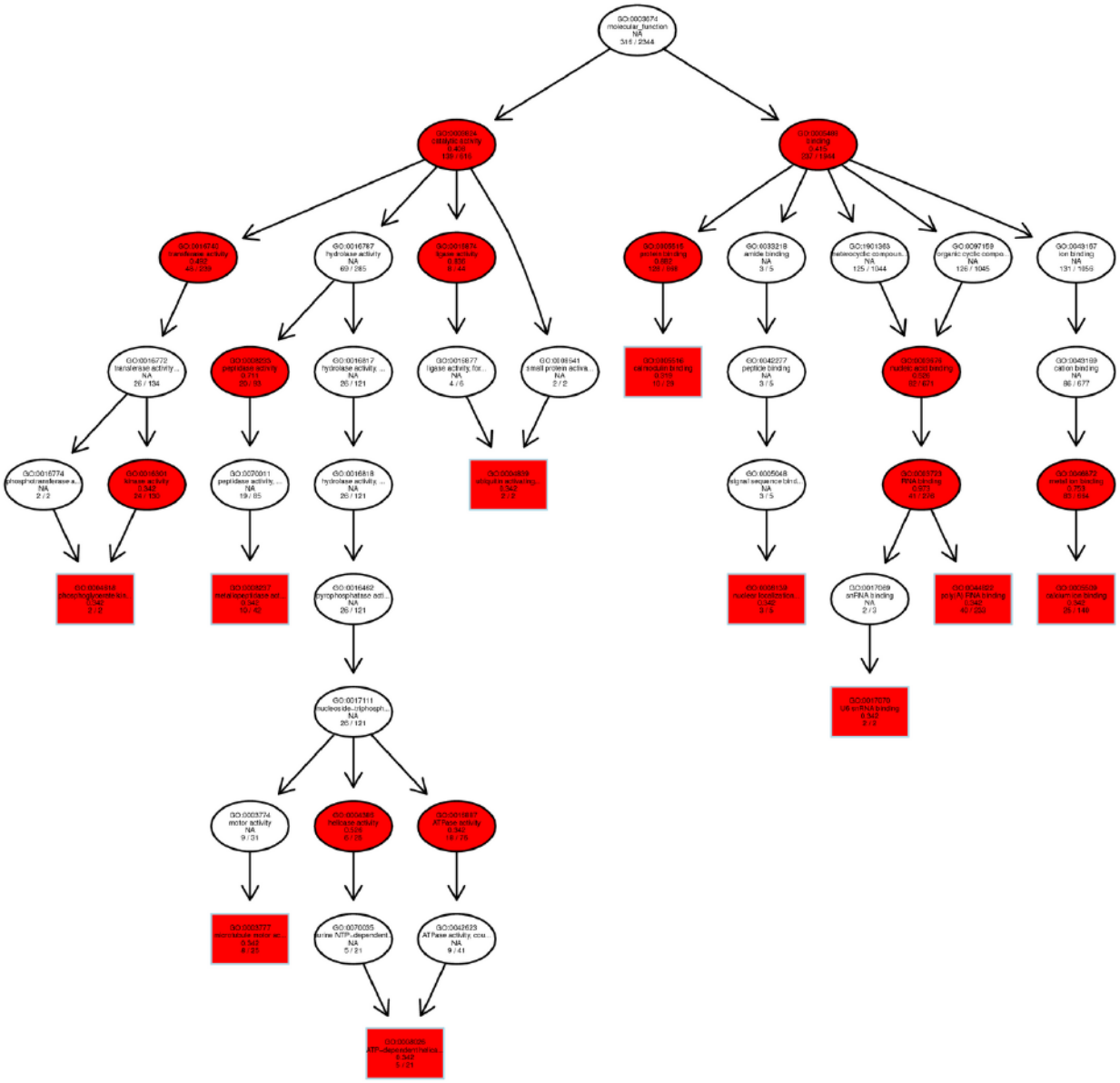


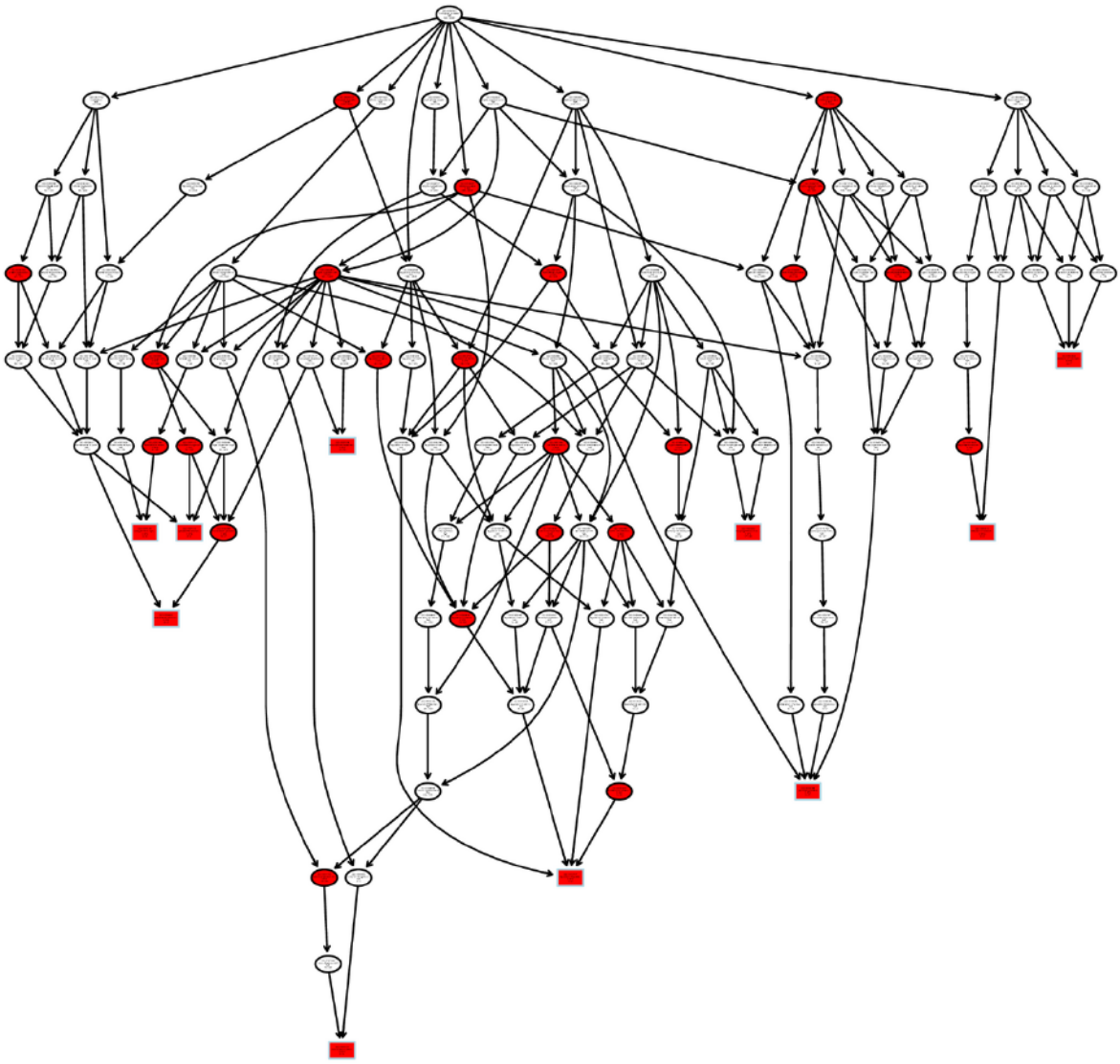
Note: X-axis: Go terms and classifications; Y-axis: Number of DEGs(genes) annotated to the term(right) and percentage of that in all DEGs(genes) (Left).This figure shows the GO enrichment in DEGs and in all genes, which indicates the importance of a specific GO term in DEGs and all genes respectively. The terms with two bars significantly different from each other can be picked up as potential targets for further analysis on functions, since these GO terms are enriched differently between DEGs-based and all-gene-based enrichment.

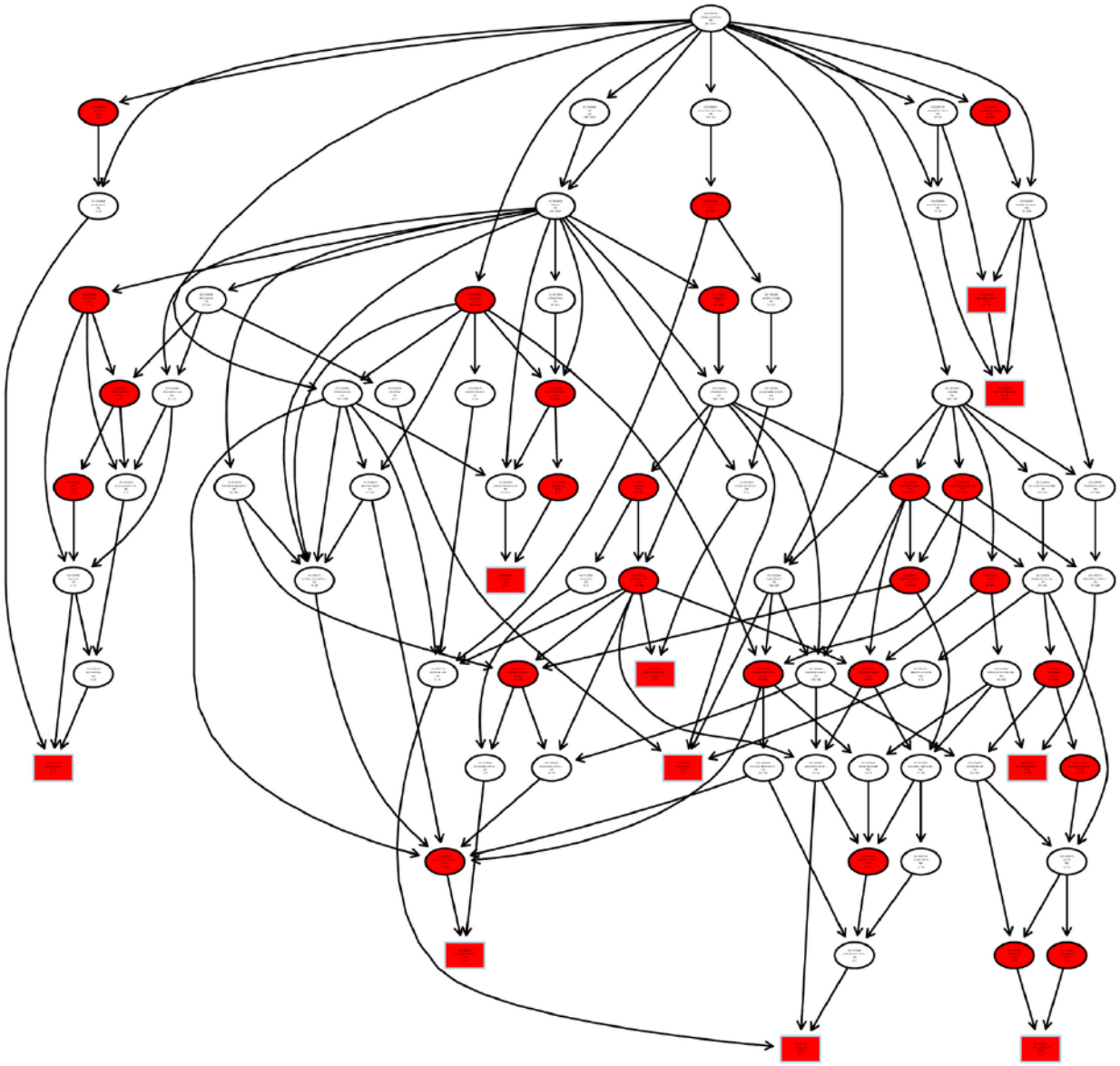
3.9.2 GO Enrichment Analysis on DE-miRNA Targeted Genes

clusterProfiler (See appendix) was employed in GO enrichment analysis on DE-miRNA targeted genes. Directed acyclic graph of the enriched terms were generated to show the hierarchical structure of the terms. In the figure, the direction of arrows represents inclusion relations between terms, i.e. the nodes are more specific than their upper nodes. Directed acylic graph of DE-miRNA targeted genes was shown in the figure below.

Figure. TopGO directed acyclic graph of DE-miRNA targeted genes







Note: The most significantly enriched 10 terms were shown in cubes, including their hierarchical structure. Each box or node contains a description of GO term and significancy value of enrichment. The color represents significancy, where a darker colour indicates a more significant enrichment.

Most significantly enriched functions in GO enrichment analysis is listed in the table below. Entire output of GO enrichment analysis can be checked in final result file.

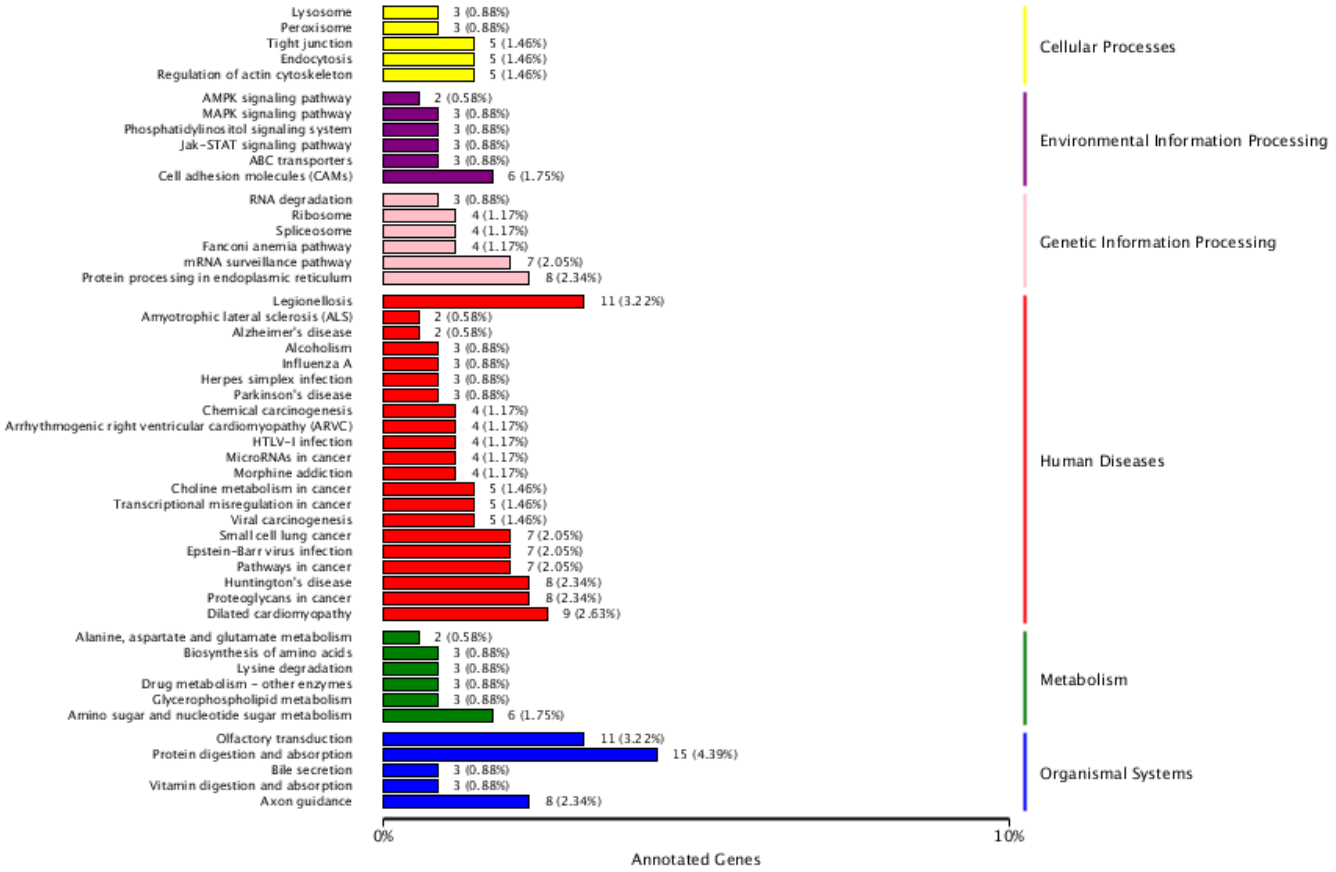
Table. topGO Enrichment

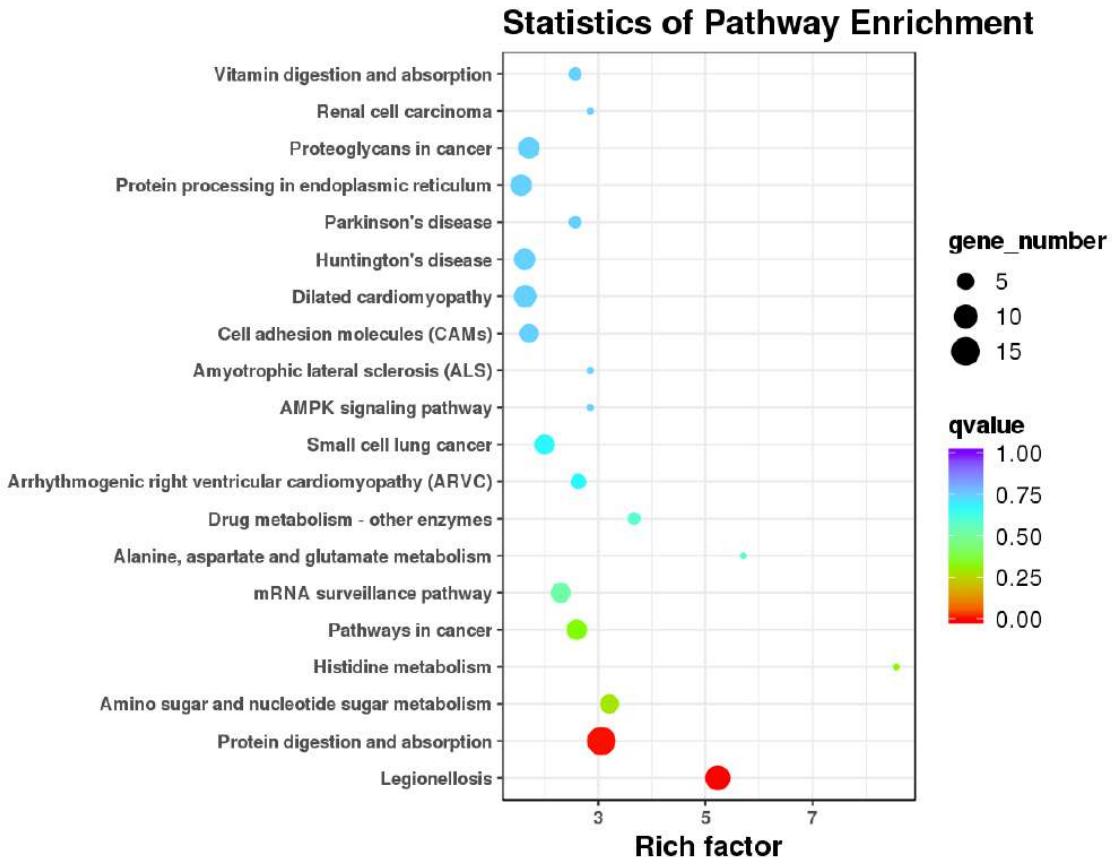
S01_S02_S03_vs_S04_S05_S06_Biological_Process_enrich.list.html
S01_S02_S03_vs_S04_S05_S06_Cellular_Component_enrich.list.html
S01_S02_S03_vs_S04_S05_S06_Molecular_Function_enrich.list.html

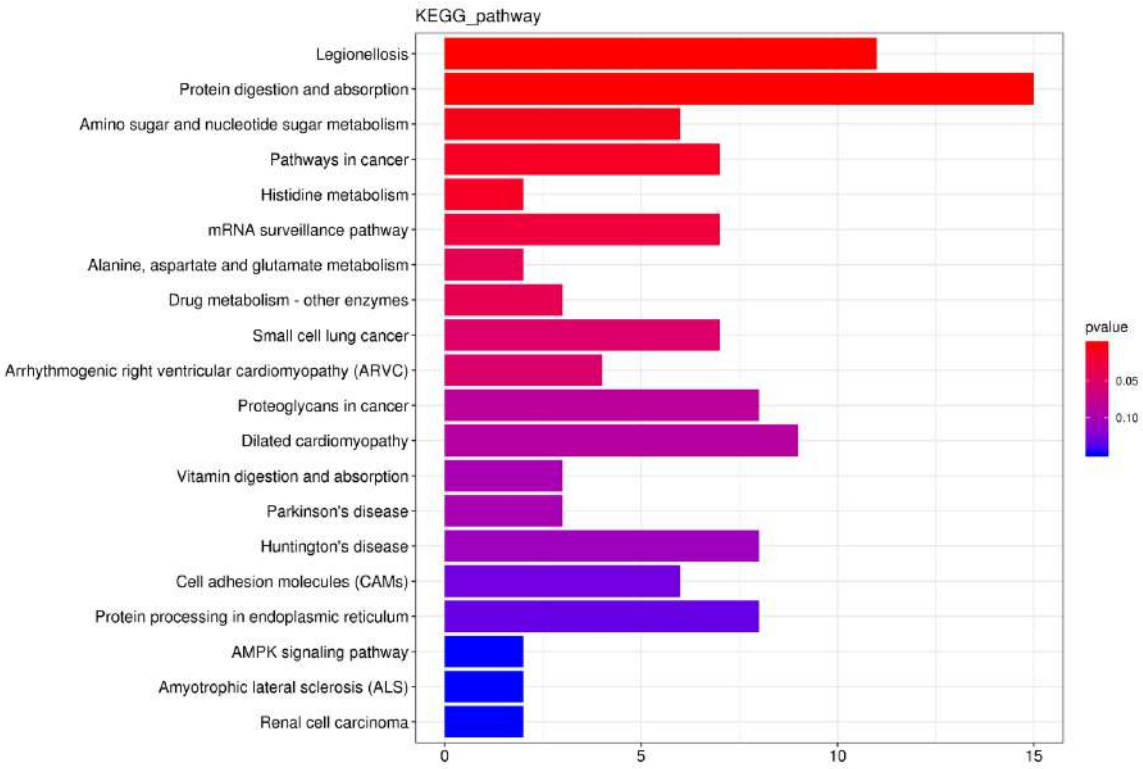
3.9.3 KEGG Annotation of DE-miRNA Targeted Genes

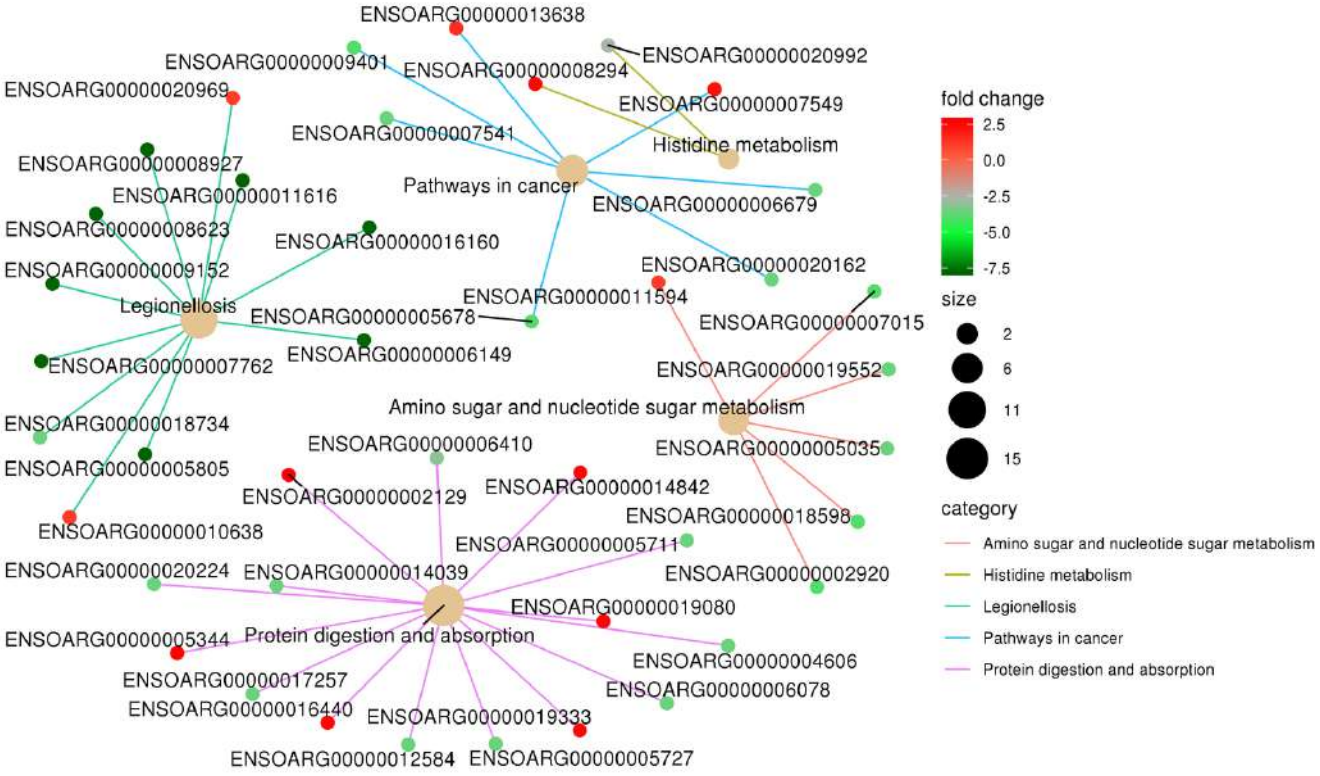
In biological organisms, series of gene products are working synergistically to perform biological functions, which is so called pathway. Annotating miRNA target genes within pathway networks could largely benefit further analysis on biological functions. KEGG (Kyoto Encyclopedia of Genes and Genomes) is one of the major databases on pathways.
A demo map of KEGG annotation on DE-miRNA target genes was shown in the figure below.
The KEGG annotations of DE-miRNA targeted genes were classified according to the type of pathways. Detailed classification was shown in the following figure.

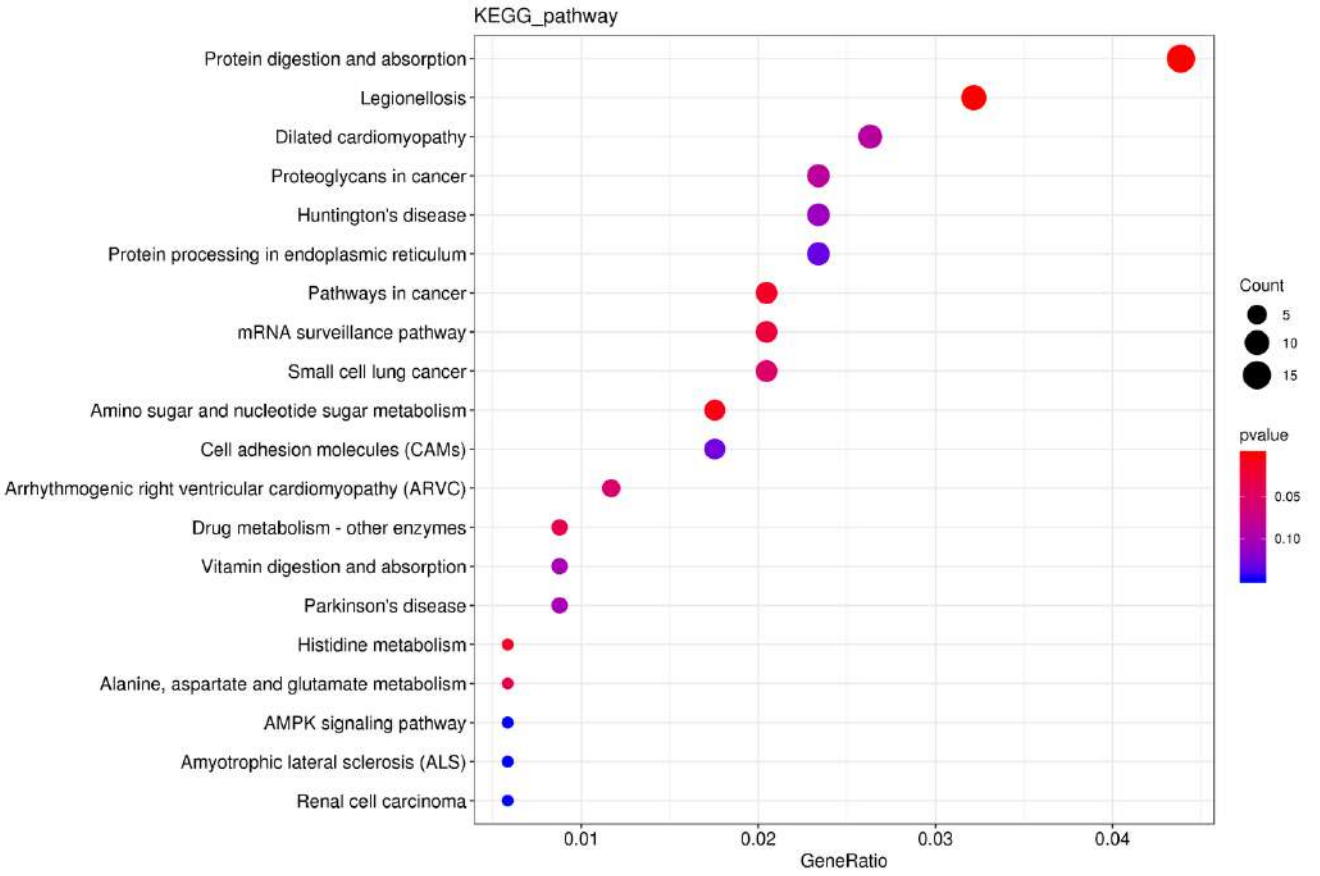
Figure. Classification of DE-miRNA targeted gene KEGG annotations




Statistics of Pathway Enrichment
Vitamin digestion and absorption
Renal cell carcinoma
Proteoglycans in cancer
Protein processing in endoplasmic reticulum
Parkinson's disease
Huntington's disease
Dilated cardiomyopathy
Cell adhesion molecules (CAMs)
Amyotrophic lateral sclerosis (ALS)
AMPK signaling pathway
Small cell lung cancer
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Drug metabolism - other enzymes
Alanine, aspartate and glutamate metabolism
mRNA surveillance pathway
Pathways in cancer
Histidine metabolism
Amino sugar and nucleotide sugar metabolism
Protein digestion and absorption
Legionellosis
3
5
7
Rich factor
gene_number
5
10
15
qvalue
1.00
0.75
0.50
0.25
0.00


KEGG_pathway
Legionellosis
Protein digestion and absorption
Amino sugar and nucleotide sugar metabolism
Pathways in cancer
Histidine metabolism
mRNA surveillance pathway
Alanine, aspartate and glutamate metabolism
Drug metabolism - other enzymes
Small cell lung cancer
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Proteoglycans in cancer
Dilated cardiomyopathy
Vitamin digestion and absorption
Parkinson's disease
Huntington's disease
Cell adhesion molecules (CAMs)
Protein processing in endoplasmic reticulum
AMPK signaling pathway
Amyotrophic lateral sclerosis (ALS)
Renal cell carcinoma
0
5
10
15
pvalue
0.05
0.10


ENSOARG00000013638
ENSOARG00000009401
ENSOARG00000008294
ENSOARG00000020992
ENSOARG00000020969
ENSOARG00000007549
ENSOARG00000007541
Histidine metabolism
ENSOARG00000008927
Pathways in cancer
ENSOARG00000011616
ENSOARG00000006679
ENSOARG00000008623
ENSOARG00000016160
ENSOARG00000009152
ENSOARG00000020162
Legionellosis
ENSOARG00000011594
ENSOARG00000005678
ENSOARG00000007015
ENSOARG00000006149
ENSOARG00000007762
ENSOARG00000019552
Amino sugar and nucleotide sugar metabolism
ENSOARG00000018734
ENSOARG00000006410
ENSOARG00000005035
ENSOARG00000005805
ENSOARG00000002129
ENSOARG00000014842
ENSOARG00000018598
ENSOARG00000010638
ENSOARG00000005711
ENSOARG00000020224
ENSOARG00000014039
ENSOARG00000002920
ENSOARG00000019080
ENSOARG00000005344
Protein digestion and absorption
ENSOARG00000004606
ENSOARG00000017257
ENSOARG00000006078
ENSOARG00000016440
ENSOARG00000019333
ENSOARG00000012584
ENSOARG00000005727
fold change
2.5
0.0
-2.5
-5.0
-7.5
size
2
6
11
15
category
Amino sugar and nucleotide sugar metabolism
Histidine metabolism
Legionellosis
Pathways in cancer
Protein digestion and absorption

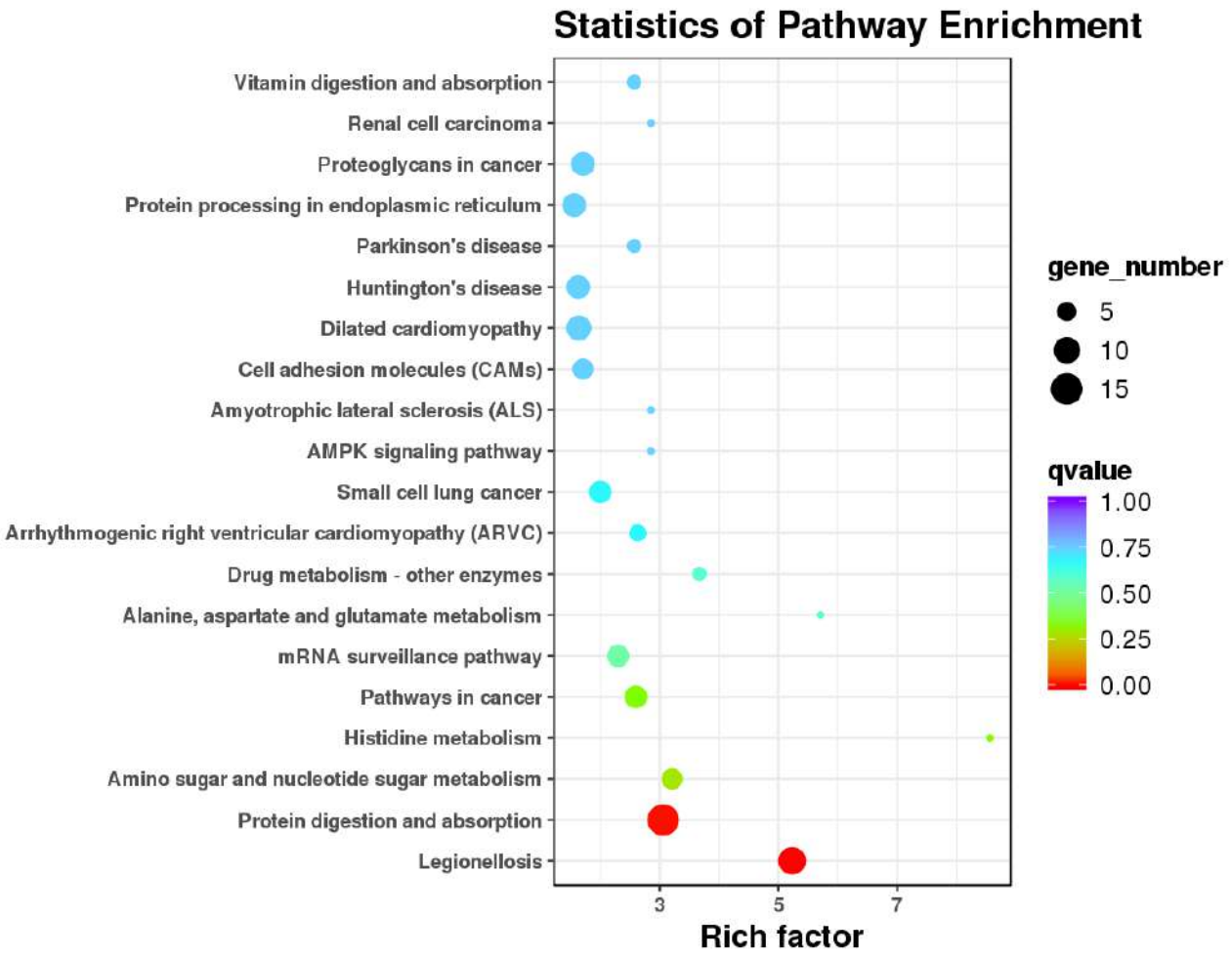

KEGG_pathway
Protein digestion and absorption
Legionellosis
Dilated cardiomyopathy
Proteoglycans in cancer
Huntington's disease
Protein processing in endoplasmic reticulum
Pathways in cancer
mRNA surveillance pathway
Small cell lung cancer
Amino sugar and nucleotide sugar metabolism
Cell adhesion molecules (CAMs)
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Drug metabolism - other enzymes
Vitamin digestion and absorption
Parkinson's disease
Histidine metabolism
Alanine, aspartate and glutamate metabolism
AMPK signaling pathway
Amyotrophic lateral sclerosis (ALS)
Renal cell carcinoma
0.01
0.02
0.03
0.04
GeneRatio
Count
5
10
15
pvalue
0.05
0.10

Note: Y-axis: KEGG pathway terms; X-axis: number and the percentage of genes annotated to the KEGG pathway.

3.9.4 KEGG Enrichment of DE-miRNA Targeted Genes

In this session, we examined if the pathways are over-presentation with DE-miRNA targeted genes. Enrichment factors and fisher test were applied in the determination of enrichment degree and significancy of the pathway.

Figure. KEGG pathway enrichment on DE-miRNA targeted genes



In this figure, the dots closer to lower right area are more reliable in differential analysis. Top enriched pathways (with smallest Q-value) were shown in the figure.

Table. KEGG Enrichment of DE-miRNA Targeted Genes

S01_S02_S03_vs_S04_S05_S06_KEGG_pathway_enrich.KEGG.list.html

Reference

1. Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 8, 186-194.
2. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10,R25.
3. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res., 40, 37-52.
4. Zhang Z, Jiang L, Wang J, et al. MTide: an integrated tool for the identification of miRNA-target interaction in plants[J]. Bioinformatics, 2014: btu633.
5. Li B, Ruotti V, Stewart R M, et al. RNA-Seq gene expression estimation with read mapping uncertainty[J]. Bioinformatics, 2009, 26(4): 493-500.
6. Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. Genome biology, 2014, 15(12): 550.
7. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26(1):139-40.
8. Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell 121, 207-221.
9. Betel D, Wilson M, Gabow A, et al. The microRNA. org resource: targets and expression[J]. Nucleic acids research, 2008, 36(suppl 1): D149-D153.
10. Lewis B P, Shih I H, Jonesrhoades M W, et al. Prediction of mammalian microRNA targets.[J]. Cell, 2003, 115(7):787-798.
11. Yangyang DENG, Jianqi LI, Songfeng WU, et al. Integrated nr Database in Protein Annotation System and Its Localization. Computer Engineering, 2006 32(5):71 74
12. Apweiler R, Bairoch A, Wu C H, et al. UniProt: the universal protein knowledgebase[J]. Nucleic acids research, 2004, 32(suppl_1): D115-D119.
13. Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, (25): 25¨C29
14. Roman L.Tatusov, Michael Y.Galperin, Darren A.Natale, et al. The COG database: a tool for genome scale analysis of protein functions and evolution. Nucleic Acids Res, 2000 7 1 28(1):33 6
15. Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, et al. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004, (32):D277 D280
16. Koonin EV, Fedorova ND, Jackson JD, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. [Genome Biology Italic], 2004, 5(2): R7.
17. Eddy S.R. Profile hidden Markov models (1998) [Bioinformatics Italic], 14 (9), pp. 755-763.

18. Rosenkranz D, Zischler H. proTRAC-a software for probabilistic piRNA cluster detection, visualization and analysis[J]. BMC bioinformatics, 2012, 13(1): 1.
19. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, Carrington JC. High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PLoS ONE 2, 2007, e219.
20. Moldovan D, Spriggs A, Dennis E S, et al. The hunt for hypoxia responsive natural antisense short interfering RNAs[J]. Plant signaling & behavior, 2010, 5(3): 247-251.
21. Katiyar-Agarwal S, Morgan R, Dahlbeck D, et al. A pathogen-inducible endogenous siRNA in plant immunity[J]. Proceedings of the National Academy of Sciences, 2006, 103(47): 18002-18007.
22. Borsani O, Zhu J, Verslues P E, et al. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis[J]. Cell, 2005, 123(7): 1279-1291.
23. Held M A, Penning B, Brandt A S, et al. Small-interfering RNAs from natural antisense transcripts derived from a cellulose synthase gene modulate cell wall biosynthesis in barley[J]. Proceedings of the National Academy of Sciences, 2008, 105(51): 20534-20539.
24. Zhang X, Xia J, Lii Y E, et al. Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function[J]. Genome biology, 2012, 13(3): R20.
25. Yu D, Meng Y, Zuo Z, et al. NATpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes[J]. Scientific reports, 2016, 6: 21666.
26. Qingli Guo, Xiongfei Qu, Weibo Jin; PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades, Bioinformatics, Volume 31, Issue 2, 15 January 2015, Pages 284–286
27. Yu G, Wang L, Han Y, He Q (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), 284-287

Appendix

Appendix 1: Methods and Data Analysis

Methods of sRNA extraction and detection

Sample collection and preparation

sRNA quantification and qualification

The RNA samples were extracted with Trizol.

The purity, concentration and integrity of RNA samples are tested using advanced molecular biology equipment to ensure the use of qualified samples for transcriptome sequencing.

Library preparation for sRNA sequencing

Briefly, First of all, ligated the 3′ SR and 5′ SR Adaptor. Then, reverse transcription synthetic first chain. Last, PCR amplification and Size Selection. PAGE gel was used to electrophoresis fragment screening purposes, rubber cutting recycling as the pieces get small RNA libraries. At last, PCR products were purified (AMPure XP system) and library quality was assessed.

Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina platform and single-end reads were generated.

Data analysis

Quality control

Raw data (raw reads) of fastq format were firstly processed through in-house perl 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. And reads were trimmed and cleaned by removing the sequences smaller than 18 nt or longer than 30 nt. At the same time, Q20, Q30, GC-content and sequence duplication level of the clean data were calculated. All the downstream analyses were based on clean data with high quality.

Comparative analysis

Use Bowtie tools soft, The Clean Reads respectively with Silva database, GtRNAdb database, Rfam database and Repbase database sequence alignment, filter ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA) and other ncRNA and repeats. The remaining reads were used to detect known miRNA and novel miRNA predicted by comparing

with Genome and known miRNAs from miRBase. Randfold tools soft was used for novel miRNA secondary structure prediction.

Target gene functional annotation

Gene function was annotated based on the following databases:

Nr (NCBI non-redundant protein sequences) ;

Pfam (Protein family) ;

KOG/COG (Clusters of Orthologous Groups of proteins) ;

Swiss-Prot (A manually annotated and reviewed protein sequence database) ;

KEGG (KEGG Ortholog database) ;

GO (Gene Ontology).

Quantification of miRNA expression levels

miRNA expression levels were estimated for each sample:

1. sRNA were mapped back onto the precursor sequence.

2. Readcount for each miRNA was obtained from the mapping results

Differential expression analysis

For the samples with biological replicates:

Differential expression analysis of two conditions/groups was performed using the DESeq2 R package (1.10.1). DESeq2 provide statistical routines for determining differential expression in digital miRNA expression data using a model based on the negative binomial distribution. The resulting P values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. miRNA with $|\log2(FC)| \geq 1.00$;FDR≤ 0.01 found by DESeq2 were assigned as differentially expressed.

For the samples without biological replicates:

Prior to differential gene expression analysis, for each sequenced library, differential expression analysis of two samples was performed using the edgeR. Pvalue was adjusted using q value (Storey et al, 2003). $|\log2(FC)| \geq 1.00$;FDR$\leq$0.01 was set as the threshold for significantly differential expression.

GO enrichment analysis

Gene Ontology (GO) enrichment analysis of the differentially expressed genes (DEGs) was implemented by the GOseq R packages based Wallenius non-central hyper-geometric distribution

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 KOBAS (Mao et al., 2005) software to test the statistical enrichment of differential expression genes in KEGG pathways.

Appendix 2: Software

Table. Software List

Tools	Version	Parameter
Bowtie	v1.0.0	-v 0
miRDeep2(animal)	v2.0.5	-g -1 -b 0
miRDeep2(plant)	v2.0.5	-g -1 -l 250 -b 0
DESeq	v1.18.0	default
IDEG6	--	default
edgeR	v3.8.6	bcv 0.1
RNAhybrid	v2.1.1	-d 1.9 0.28 -b 1 -e -25
miRanda	v3.3a	-sc 50.0 -en -20 -scale 4.0 -go -2.0 -ge -8.0

Tools	Version	Parameter
TargetFinder	v1.6	-c 3
randfold	v2.0	-s 99
RNAfold	v2.1.7	default
blast	v2.2.26	-b 100 -v 100 -e 1e-5 -m 7 -a 2
topGO	v2.18.0	nodeSize=6 firstSigNodes=10

Appendix 3. Database

Table. Database List

Database	Homepage
Silva	http://www.arb-silva.de/
GtRNAdb	http://lowelab.ucsc.edu/GtRNAdb/
Rfam	http://rfam.xfam.org/
Repbase	http://www.girinst.org/repbase/
miRbase	http://www.mirbase.org/
NR	ftp://ftp.ncbi.nih.gov/blast/db/
KOG	http://www.ncbi.nlm.nih.gov/KOG/
Pfam	http://pfam.xfam.org/
Swiss-Prot	http://www.uniprot.org/
GO	http://www.geneontology.org/
COG	http://www.ncbi.nlm.nih.gov/COG/
KEGG	http://www.genome.jp/kegg/
Ensembl	http://asia.ensembl.org/index.html