Small RNA Analysis Report

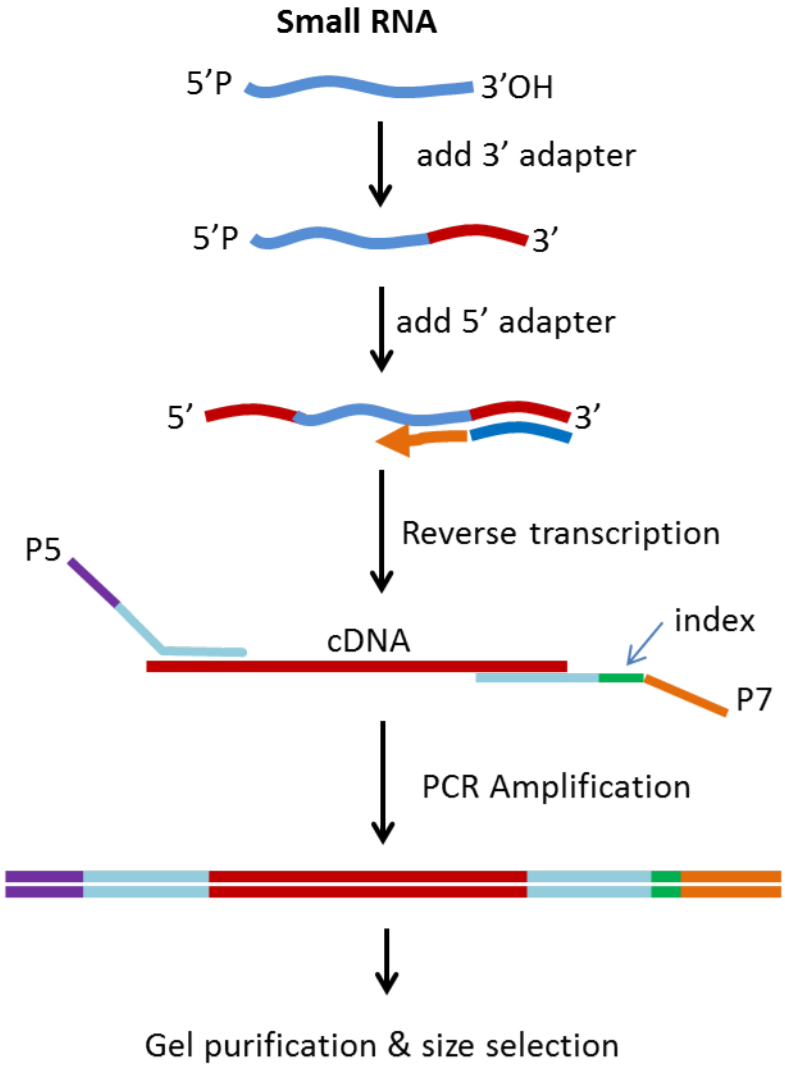
Demo Report

May 1, 2016

Contents

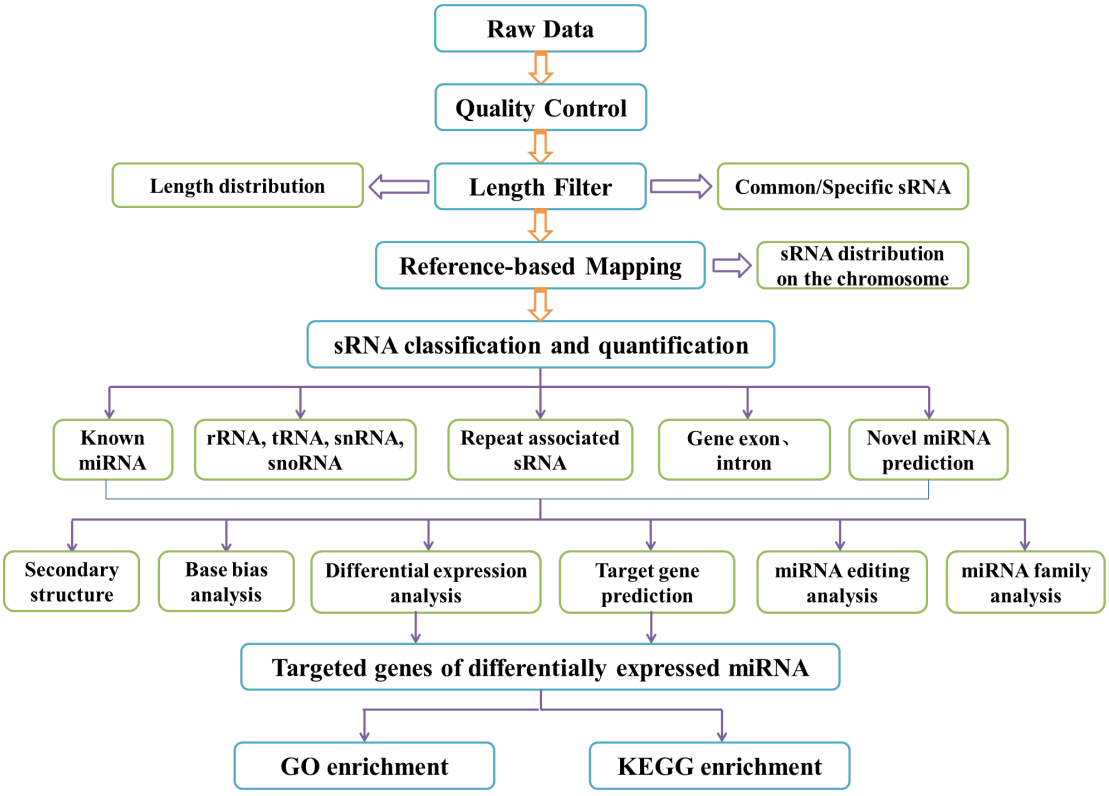
1 Library Preparation and Sequencing

Small RNA is a special kind of molecule in organisms which induces the gene silence and plays an important role in the regulation of cell growth, gene transcription and translation. The small RNA digitalization analysis based on Hi Seq high-throughput sequencing takes the SBS-sequencing by synthesis, which can decrease the loss of nucleotides caused by the secondary structure. It is also strong for its small requirement of sample quantity, high through-put, high accuracy with simply operated automatic platform. Such analysis can obtain millions of small RNA sequence tags in one shot, identify small RNA of certain species in certain condition comprehensively, predict novel miRNA and construct the small RNA differential expression profile between samples, which could be used as a powerful tool on small RNA function research. The experiment process of small RNA sequencing is as follows:



2 Data analysis process

Considering the sample was obtained from animal with a reference genome, we were using referenced animal small RNA analysis process, as follows:



3 Results

3.1 Raw Data

The original raw data from high throughput sequencing(illumina HiSeq) which consisted of raw pictures were first transformed to Sequenced Reads which contained reads sequence and corresponding base quality (in FASTQ format) through Base Calling.

Each read has four descriptive lines in as follow:

@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHFAE0ECFFD/AEHH

1st line: Illumina Sequence Identifiers and description information comes after "@" sign; 2nd line: bases sequence (A G C T) 3ed line: "+" sign, then Illumina Sequence

Identifiers(optional) 4th line: each base's quality corresponding to 2nd line (Cock et al.). Illumina Sequence Identifiers:

Identifier	Meaning
HWI-ST1276	Instrument – unique identifier of the sequencer
71	run number – Run number on instrument
C1162ACXX	FlowCell ID – ID of flowcell
1	LaneNumber – positive integer
1101	TileNumber – positive integer
1208	X – x coordinate of the spot. Integer which can be negative
2458	Y – y coordinate of the spot. Integer which can be negative
1	ReadNumber - 1 for single reads; 1 or 2 for paired ends
N	whether it is filtered - NB：Y if the read is filtered out, not in the delivered fastq file, N otherwise
0	control number - 0 when none of the control bits are on, otherwise it is an even number
CGATGT	Illumina index sequences

Each character in the fourth row of the corresponding ASCII value minus 33, which is corresponding to the second line base sequencing quality value. The relationship between sequencing error rate (E) and sequencing quality (s Q) is shown in the below formula:

$$\text{Formula one:} \quad Q_{phred} = -10\log10(e)$$

The relationship of Phred quality scores Q and base-calling error which were predicted by Illumina Casava 1.8 version Phred quality scores Q were logarithmically linked to base-calling errore:

Phred score	Base Calling error rate	Base Calling correct rate	Q-sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

3.2 Sequencing data quality evaluation

3.2.1 Data Quality summary

Table 3.2.1 Summary of the Data Production

Sample	Reads	Bases	Error rate	Q20	Q30	GC content
S_1	12047960	0.602G	0.01%	97.20%	93.93%	49.36%
S_2	12174594	0.609G	0.01%	97.12%	93.75%	49.95%
S_3	12748703	0.637G	0.0001	0.9713	0.9384	0.4961

(1) Sample: Sample id.

(2) Reads: Statistics of the original sequence data.

(3) Bases: Sequence multiplied the length of the sequence, and converted to the unit of G.

(4) Error rate: Sequencing error rate.

(5) Q20: Phred values greater than 20 base percentage accounted for the overall base.

(6) Q30: Phred values greater than 30 base percentage accounted for the overall base.

(7) GC content: The G base and C base accounted for the overall base.

3.2.2 Data cleaning

The unpurified raw data are 5' primer comtains, no insert tags, oversized insertion,low quality reads,poly A tags and small tags. We will get rid of some contaminant reads from the fq file and get the final clean reads.

The data is processed by the following steps:

(1) Get rid of reads which s Q <= 5 base percentage > 50% .

(2) Get rid of reads containing N > 10%.

(3) Get rid of reads with 5' primer contaminants.

(4) Get rid of reads without 3' primer and reads without the insert tag.

(5) Trim 3'primer sequenc.

(6) Get rid of reads with poly A/T/G/C.

Small RNA adapte sequences:

RNA 5’ Adapter (RA5), part:

5’-GTTCAGAGTTCTACAGTCCGACGATC-3’

RNA 3’ Adapter (RA3), part:

5’-AGATCGGAAGAGCACACGTCT-3’

Table 3.2.2 Data filtering summary

Sample	total reads	N% > 10%	low quality	5 adapter contamine	3 adapter null or insert null	with ployA/T/G/C	clean reads
S_1	12047960 (100.00%)	7 (0.00%)	3656 (0.03%)	667 (0.01%)	353829 (2.94%)	6160 (0.05%)	11683641 (96.98%)
S_2	12174594 (100.00%)	6 (0.00%)	4144 (0.03%)	875 (0.01%)	413564 (3.40%)	14126 (0.12%)	11741879 (96.45%)
S_3	12748703 (100.00%)	2 (0.00%)	6513 (0.05%)	936 (0.01%)	358624 (2.81%)	12360 (0.10%)	12370268 (97.03%)

(1) Sample : Sample id.

(2) total_reads : Total sequenced reads.

(3) N% > 10% : Percentage of reads with N > 10%.

(4) low quality : Percentage of low quality reads .

(5) 5'adapter contamine : Percentage of reads with 5'adapter contamine .

(6) 3'adapter null or insert null : Percentage of reads with 3'adapter null or insert null.

(7) with ploy A/T/G/C : Percentage of reads with ploy A/T/G/C.

(8) clean reads : Total clean reads and its percentage accounted for raw reads.

3.2.3 Length distribution

Generally speaking, the length of sRNA is between 18nt and 30nt. The length distribution analysis is helpful to see the compositions of small RNA sample. For example, miRNA is normally 21nt or 22nt, siRNA is 24nt, and piRNA is between 28nt and 30nt.The length distribution varies between plants and animals, in detail, the peak of plant often locates in 21nt or 24nt while animals is 22nt.The above data and information is helpful for us to do initial judgment.

Table 3.2.3 The type and quantity of sRNA

Sample	Total reads	Total bases (bp)	Uniq reads	Uniq bases (bp)
S_1	9742471	219606009	600206	14119632
S_2	9550931	220775560	1301461	31610866
S_3	10041867	231329567	966227	23095143

(1) Sample: Sample ID.

(2) Total reads: Total number of sRNA reads.

(3) Total bases (bp): Total reads multiplied the length of the sequence, and converted to the unit of G.

(4) Unique reads: The kinds of sRNA.

(5) Unique bases (bp): Unique reads multiplied the length of the sequence, and converted to the unit of G.

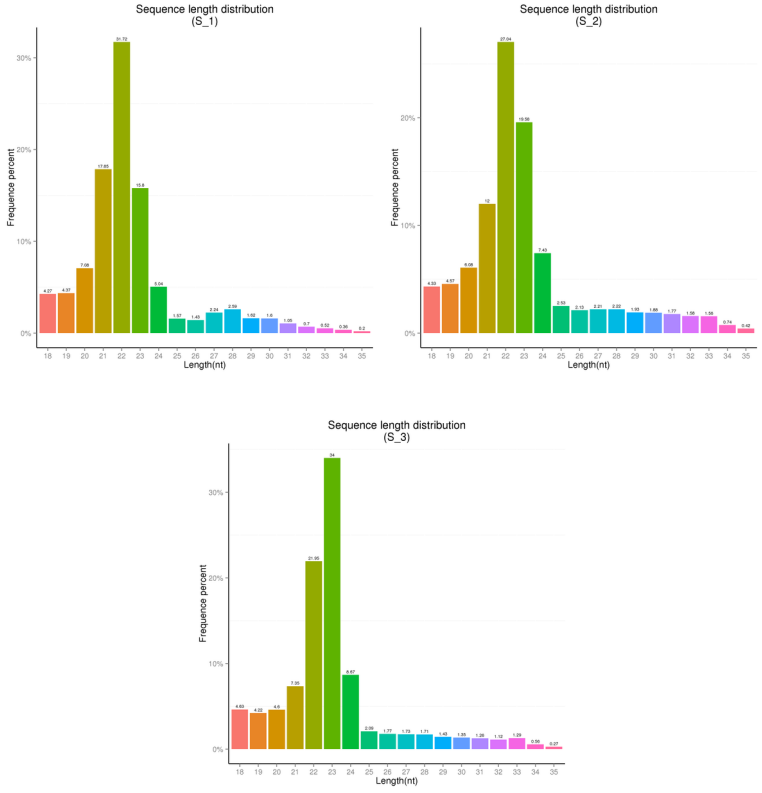


Figure 3.2.3 The length distribution of total sRNA

The abscissa is the length of sRNA reads, the ordinate is the percentage of one length read accounted for total sRNA.

3.2.4 Common and specific reads between samples

Summarize the common and specific reads of two samples, including the summary of unique reads and total reads. Generally speaking, a huge difference of reads exists among different samples but the common reads are concentrated, which demonstrates that uniformity of different samples on the whole of sequencing is good.

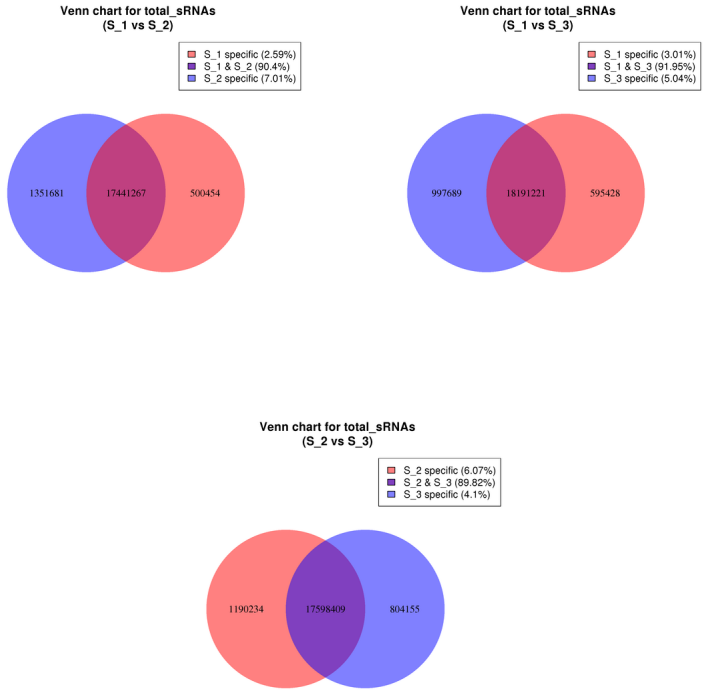


Figure 3.2.4.1 Common and specific reads between samples (Total sRNA)

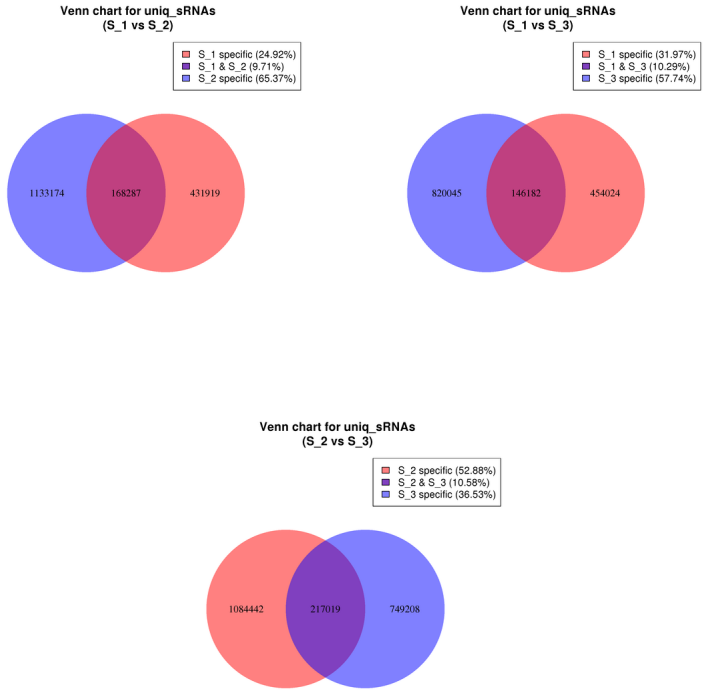


Figure 3.2.4.2 Common and specific reads between samples(Uniq sRNA)

(1) Sample1 specific: Specific reads in sample1.

(2) Sample1 & Sample2: Common reads between sample 1 and sample2.

(3) Sample2 specific: Specific reads in sample2.

3.3 Mapping to genome

Map the small RNA reads to genome by bowtie to analyze their expression and distribution on the genome.

Table 3.3 Statistics of mapping results

Sample	Total sRNA	Mapped sRNA	+ Mapped sRNA	- Mapped sRNA
S_1	9742471 (100.00%)	6236499 (64.01%)	3920295 (40.24%)	2316204 (23.77%)
S_2	9550931 (100.00%)	5875008 (61.51%)	2945268 (30.84%)	2929740 (30.67%)
S_3	10041867 (100.00%)	6787314 (67.59%)	2424492 (24.14%)	4362822 (43.45%)

(1) Sample: Sample id.

(2) Total sRNA: Quantily of total sRNAs after the length filter.

(3) Mapped sRNA: Quantily and percentage of sRNAs mapped to genome.

(4) "+" Mapped sRNA: Quantily and percentage of mapped sRNAs in the same direction as the genome.

(5) "-" Mapped sRNA: Quantily and percentage of mapped sRNAs in the oppsite direction as the genome. Count the number of each sample small RNA reads that locate on each chromosome, using Circos to view the distribution of reads on each chromosome. We choose the longest 10 contigs or scafords to analysis.

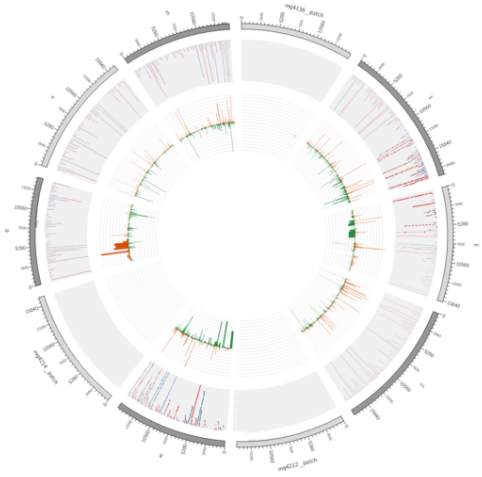


Figure 3.3 reads distribution on each chromosome

The chromosome is shown in the outer circle: grey background in the middle area shown 10000 reads distribution on the chromosome, red is the number of sRNAs on the sense strand of chromosome, blue is the number of sRNAs on the antisense strand of chromosome; in the center area of the circle shown all the reads, croci is the number of sRNAs on the sense strand of chromosome, green is the number of sRNAs on the antisense strand of chromosome.

3.4 Analysis of known miRNA

miRNA is produced by Dicer from pri-miRNA. Because of specificity of the cleavage site, miRNA has some preference on bases at different positions. For example, the first base from the 5' end has a strong preference of U, but resistant to G; bases from position 2 to 4 on the 5' end are usually resistant to U; bases from position10 (this position is the cleavage site when miRNA regulates mRNA) has a strong A preference. The mapped reads align to a specific special in the miRBase21, get the details of each sample's known miRNA information. Result table and figure:

Table 3.4.1 Summary of known miRNA in each sample

Types	Total	S_1	S_2	S_3
Mapped mature	977	702	823	761
Mapped hairpin	763	583	663	635
Mapped uniq sRNA	16921	5214	6257	5450
Mapped total sRNA	13292399	4492739	3855991	4943669

(1) Mapped mature: The number of sRNAs align to miRNA mature sequence. The second line is the number of all samples align to miRNA mature sequence, line(3) to line(n+1) is the number of each sample align to miRNA mature sequence.

(2) Mapped hairpin: The number of sRNAs align to miRNA hairpin sequence. The second line is the number of all samples align to miRNA hairpin sequence,line(3) to line(n+1) is the number of each sample align to miRNA hairpin sequence.

(3) Mapped uniq sRNA: The number of mapped unique sRNAs.

(4) Mapped total sRNA: The number of mapped total sRNAs.

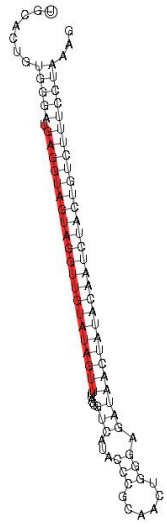


Figure 3.4 The stem loop structure of precursors of known miRNA red partial is mature sequence

Table 3.4 Known miRNA expression profile

miRNA	S_1	S_2	S_3

mmu-let-7a-1-3p	421	230	211
mmu-let-7a-5p	68708	26836	20976
mmu-let-7b-3p	26	7	9
mmu-let-7b-5p	15494	2125	697
mmu-let-7c-5p	38662	5003	1943

First line: mature miRNA id; the second line to n+2 line is each sample's readcount.

The details of mapped known miRNA:

- The details of mapped known miRNA:
- > miRNAs hairpin name
- total readcount : the number of reads mapped to this hairpin
- the mature miRNA name and the number of reads mapped to this mature
- remaining read count : reads mapped to the hairpin but not in the mature region
- exp: the annotation of miRNA hairpin each position,5 representing 5' region,3 representing 3' region,f representing others region of hairpin.
- pri_seq : hairpin sequences
- pri_struct : second structure of miRNA hairpin
- XXX_X_x_X : the first three character is sample name,next is "_",then is the id of this reads in each sample,next is also "_",at last is x and the frequency of this read

```
>rno-let-7a-1
total read count 61339
rno-let-7a-1-3p read count 1195
rno-let-7a-5p read count 60121
remaining read count 23
exp fffffffffff5555555555555555555555fffffffffffffffffffffffffffff333333333333333333333ffffffff
pri_seq uucacuguggugaugagguaguagguuguauaguuuuagggucacacccaccacugggagauaacuauacaaucuacugucuuuccuaaggugau
pri_struct .((((((.((((((.(((..(((((((((((((((((...(((.....))).((....)).....)))))))))))))))))..)))))))).)))))). #MM +/-
K_L_364221_x2 ...........augagguaguagguuguauagu.................................................................. 0 +
K_D_377491_x1 ...........augagguaguagguuguauagu.................................................................. 0 +
```

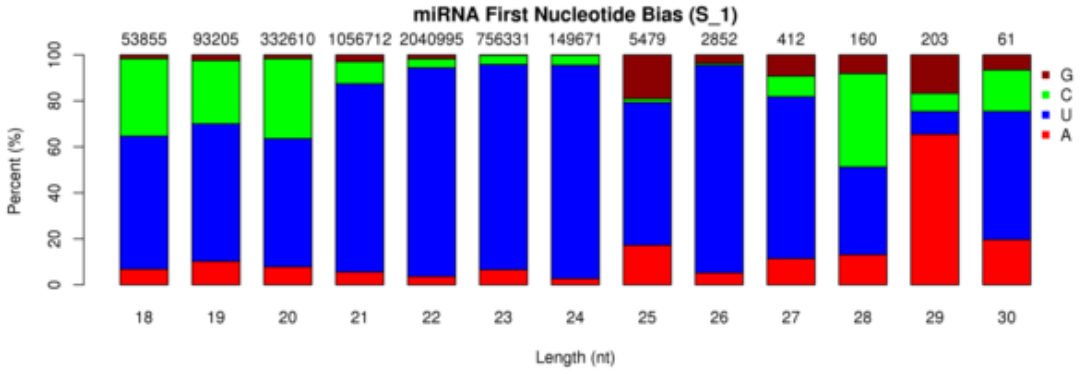


Figure 3.4.2 miRNA first nucleotide bias

The length of miRNAs is shown in the X axis, the Y axis is the percentage.

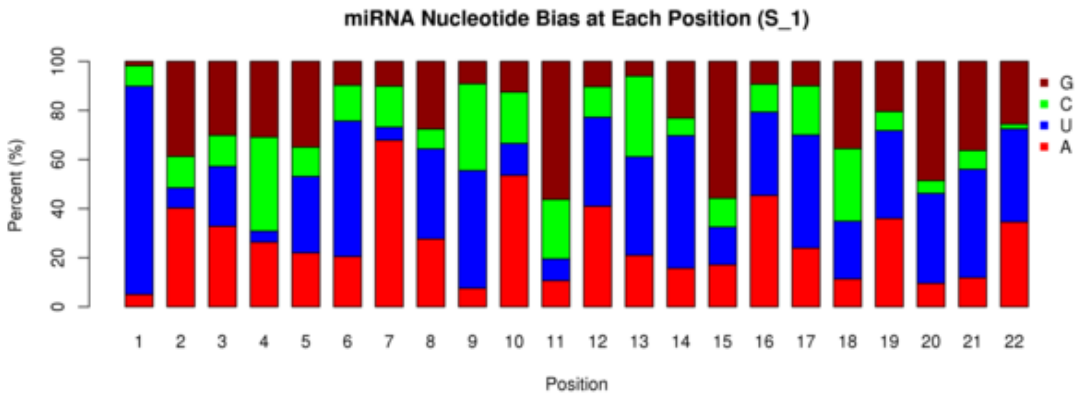


Figure 3.4.3 miRNA nucleotide bias at each position

The X axis shown each position of miRNA nucleotide, The Y axis shown the percentage

3.5 Alignment to Rfam

Annotate the small RNA reads with sequences from Rfam and get rid of matched reads from rRNA、tRNA、snRNA、snoRNA.

Table 3.5.1 Statistics of annotated ncRNA

Types	S_1	S_2	S_3
rRNA	50968	67947	48844
rRNA:+	50943	67920	48769
rRNA:-	25	27	75
tRNA	16970	21604	14592
tRNA:+	16904	21478	14416
tRNA:-	66	126	176
snRNA	36784	50590	51343
snRNA:+	36773	50581	51325
snRNA:-	11	9	18
snoRNA	866585	404552	262818
snoRNA:+	866576	404543	262813
snoRNA:-	9	9	5

First line: Types, the kinds of ncRNA; the second line to n+1 are sample(1) to sample(n+1)'s reads mapping to that kind of ncRNA.

3.6 Repeat associated RNA alignment

Align small RNA reads to repeat associated RNA to find matched tags in the sample. Repeat sequence was predicted by Repeat Masker.

Statistics the repeat type figure 3.6.1 and figure 3.6.2

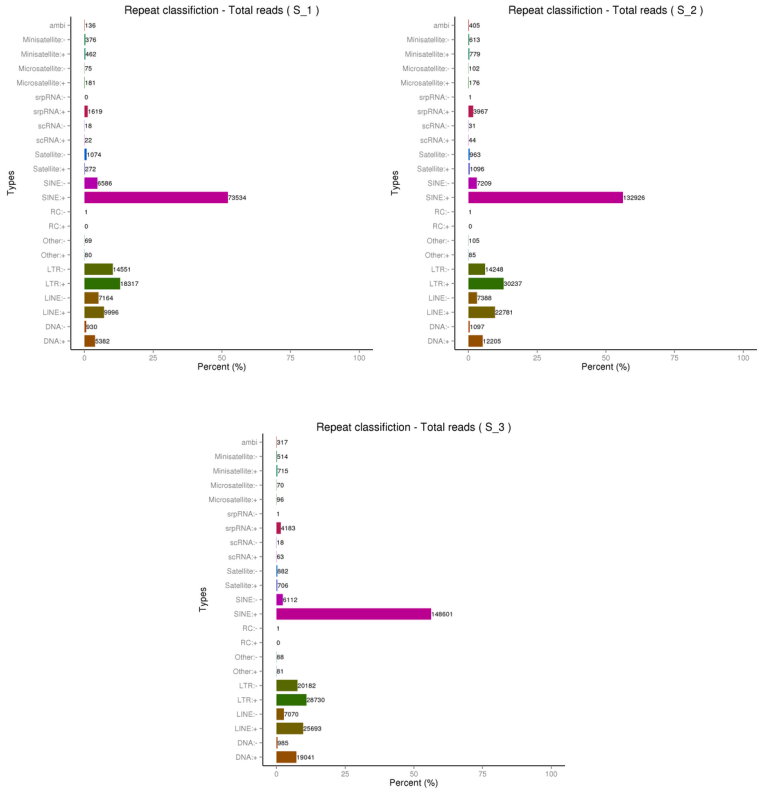
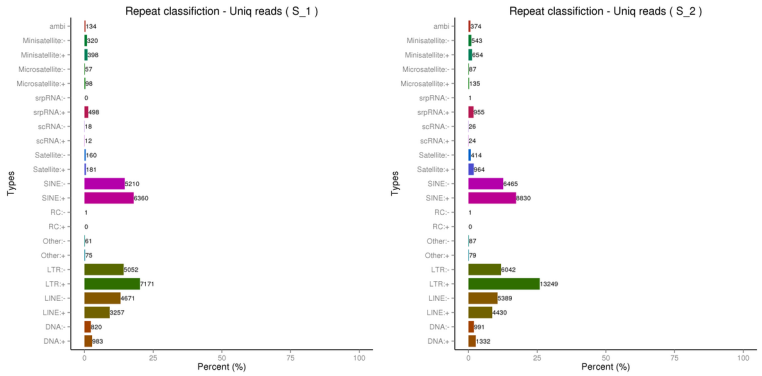


Figure 3.6.1 Total repeat reads



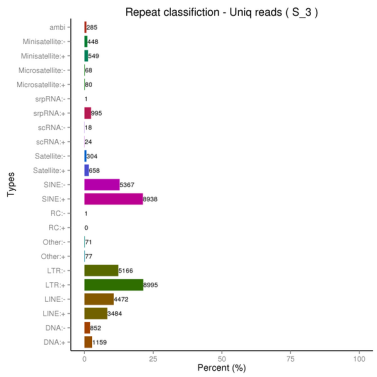


Figure 3.6.2 Uniq repeat reads

3.7 Exon and intron alignment

Align small RNA reads to exons and introns of mRNA to find the degraded fragments of mRNA in the small RNA reads.

Table 3.7 sRNAs mapped to exon and intron

Types	S_1	S_2	S_3
exon	194523	643392	538739
exon:+	160947	629132	509502
exon:-	33576	14260	29237
intron	312045	403123	452385
intron:+	210976	307047	330187
intron:-	101069	96076	122198

First line: our types. “exon : +” sRNAs mapped to sense exon, “exon: -” sRNAs mapping to antisense exon, “intron: +” sRNAs mapping to sense intron , “intron: -” sRNAs mapped to antisense intron. the second line to n+1 line is each sample's sRNAs..

3.8 Novel miRNA prediction

The characteristic hairpin structure of miRNA precursor can be used to predict novel miRNA. We use miREvo（Wen et al.，2012）and mirdeep2（Friedlander etal.，2011）to predict novel miRNA.

Table 3.8.1 Summary of novel miRNA

Types	Total	S_1	S_2	S_3
Mapped mature	14	13	13	13
Mapped star	2	1	1	2
Mapped hairpin	14	13	13	13
Mapped uniq sRNA	217	65	79	73
Mapped total sRNA	9477	1019	3808	4650

(1) Novel hairhin: Predicted hairpin. The second line is the number of all predicted hairpin,line(3) to line(n+2) is the number of each sample align to predicted hairpin .

(2) Mapped uniq sRNA: Mapped unique sRNAs.

(3) Mapped total sRNA: Mapped total sRNAs

Figure 3.8.1 The stem loop structure of precursors of novel miRNA

Red partial is mature sequence

Table 3.8.2 Novel miRNA expression profile

miRNA	S_1	S_2	S_3
novel_1	772	3249	4240
novel_10	0	0	5
novel_11	19	9	5
novel_12	9	23	29
novel_13	10	3	1

First line: mature miRNA id; the second line to n+2 line is each sample's readcount.

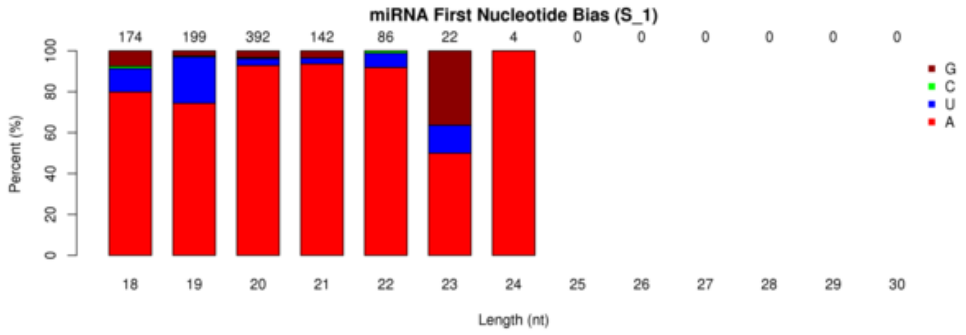


Figure 3.8.2 miRNA first nucleotide bias

The length of miRNAs is shown in the X axis, the Y axis is the percentage.

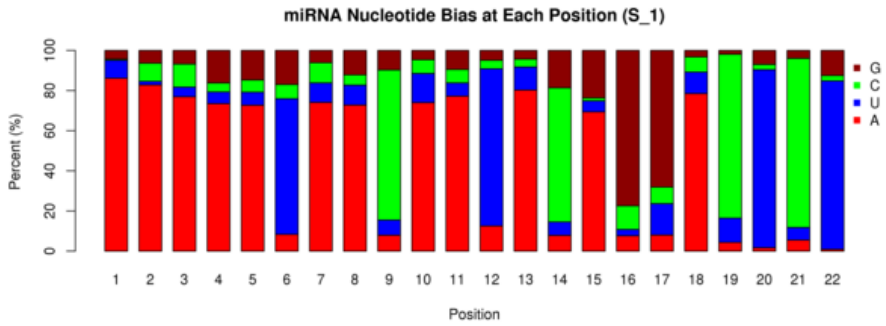


Figure 3.8.3 miRNA nucleotide bias at each position

The X axis shown each position of miRNA nucleotide, the Y axis shown the percentage

3.9 small RNA annotation

Summarize all alignments and annotation, some small RNA reads may be mapped to more than one categories. 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 > novel miRNA.

Table 3.9 Result table

Types	S_1	S_1(percent)	S_2	S_2(percent)	S_3	S_3(percent)
total	6236499	100.00%	5875008	100.00%	6787314	100.00%
known_miRNA	4492739	72.04%	3855991	65.63%	4943669	72.84%
rRNA	50968	0.82%	67947	1.16%	48844	0.72%
tRNA	16970	0.27%	21604	0.37%	14592	0.21%
snRNA	36784	0.59%	50590	0.86%	51343	0.76%
snoRNA	866585	13.90%	404552	6.89%	262818	3.87%
repeat	140845	2.26%	236459	4.02%	264149	3.89%
novel_miRNA	1019	0.02%	3808	0.06%	4650	0.07%
exon:+	160947	2.58%	629132	10.71%	509502	7.51%
exon:-	33576	0.54%	14260	0.24%	29237	0.43%
intron:+	210976	3.38%	307047	5.23%	330187	4.86%
intron:-	101069	1.62%	96076	1.64%	122198	1.80%
other	124021	0.0199	187542	0.0319	206125	0.0304

(1) total: The quantity of sRNA reads mapped to genome.

(2) known_miRNA: The number and percentage of sRNAs reads mapped to konwn miRNA.

(3) rRNA/tRNA/snRNA/snoRNA: The number and percentage of sRNAs reads mapped to rRNA/tRNA/snRNA/snoRNA.

(4) repeat: The number and percentage of sRNAs reads mapped to repeat region.

(5) novel_miRNA: The number and percentage of sRNAs reads mapped to novel miRNA.

(6) exon: +/exon : -/exon : +/intron : -/intron: The number and percentage of sRNAs reads mapped to exon (+/-) and intron(+/-).

(7) other: The number and percentage of sRNAs reads mapped genome but cound not map to known miRNA, ncRNA, repeat, novel miRNA, exon/intron.

3.10 miRNA base edit

Position 2~8 of a mature miRNA is called seed region which is highly conserved. The target of a miRNA might be different with the change of nucleotides in this region. In our analysis pipeline, miRNAs which might have base edit can be detected by aligning unannotated sRNA reads with mature miRNAs from miRBase.

Result:

pre-miRNA: miRNA precursor, the number of reads mapped to precursor, the number and percentage of reads with base edit.

matrue miRNA: mature miRNA, the number of reads mapped to mature miRNA, the number and percentage of reads with base edit.

site[1-n]: the details of each base of this mature miRNA, each line represent the number and percentage of reads with base edit.

>pre-miRNA: novel_106 1281 365 28.49%

mature miRNA: novel_106* 0 0.00%

mature miRNA: novel_106 1046 243 23.23%

site1: 10 0.96%

C->A: 3 0.29%

C->G: 1 0.1%

C->U: 6 0.57%

site2: 6 0.57%

C->A: 2 0.19%

C->G: 1 0.1%

3.11 miRNA family analysis

Explore the occurrence of known miRNA and novel miRNA families identified from the sample in other species. The first row is the species' name in mi RBase21, the first line is the name of known and novel miRNA precursor family name. "+" means that the miRNA family exists in a species and "-" means the inexistence of the miRNA family in a species.

Table 3.11 result

Species	mir-653	mir-328	mir-672	mir-128	mir-298	mir-21	mir-425	mir-188	mir-337
Macaca mulatta	+	-	-	+	+	+	+	+	+
Rattus norvegicus	+	+	+	+	+	+	+	+	+
Gorilla gorilla	-	+	-	+	-	+	-	+	+
Pongo pygmaeus	+	+	-	+	+	+	+	+	+
Mus musculus	+	+	+	+	+	+	+	+	+
Pan troglodytes	+	+	-	+	+	+	+	+	+
Canis familiaris	+	+	-	+	-	+	+	+	-
Cricetulus griseus	+	+	+	+	+	+	+	+	-
Sus scrofa	-	+	-	+	-	+	+	+	-

3.12 miRNA expression and differential expression

3.12.1 miRNA expression

Statistics the expression of known and novel miRNA in each samples, TPM (Zhou et al., 2010) method was used. The normalized expression=read Count*1,000,000)/libsize (libsize: sample miRNA readcount)

Table 3.12.1 result

sRNA.readcount	S_1.readcount	S_2.readcount	S_3.readcount	S_1.tpm	S_2.tpm	S_3.tpm
mmu-let-7a-1-3p	421	230	211	90.21559384	58.4011146	42.05058586
mmu-let-7a-5p	68708	26836	20976	14723.35635	6814.140484	4180.346393
mmu-let-7b-3p	26	7	9	5.571509358	1.777425227	1.793626885
mmu-let-7b-5p	15494	2125	697	3320.191	539.5755153	138.9064376
mmu-let-7c-5p	38662	5003	1943	8284.834416	1270.351201	387.2241153
mmu-let-7d-3p	1647	1174	1081	352.933689	298.0996023	215.4345181
mmu-let-7d-5p	10843	8725	6564	2323.533691	2215.433586	1308.151875

mmu-let-7e-3p	5	0	0	1.071444107	0	0
mmu-let-7e-5p	333	331	16	71.35817755	84.04682144	3.188670018
mmu-let-7f-1-3p	48	34	45	10.28586343	8.633208245	8.968134425

(1) First line-"miRNA", representing the miRNA mature id.

(2) Second line to n+1 line-"sample name", representing the sample 1 to n+1 readcount.

(3) n+2 line to 2n+1 line -"sample name(TPM)", represent the each samples readcount (TPM normalization)

3.12.2 miRNA TPM distribution

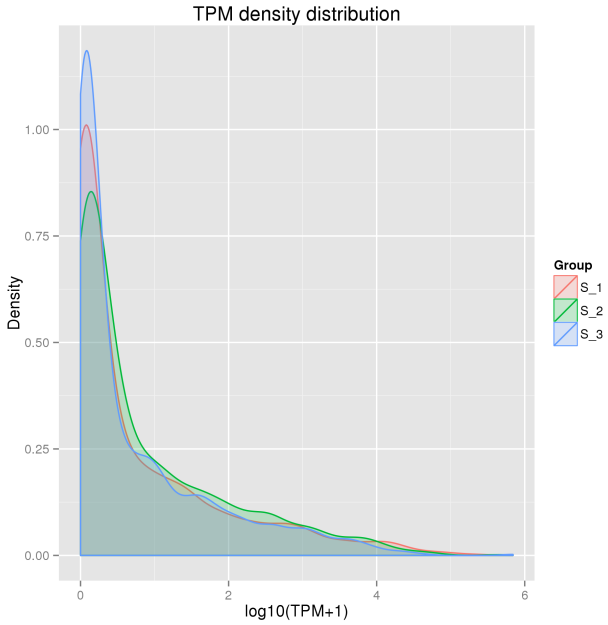


Figure 3.12.2 TPM distribution

The x axis is sample names and y axis is the price of miRNA log10(TPM+1).

3.12.3 RNA-Seq Correlation

Biological replicates are necessary for any biological experiment, including those involving RNA-seq technology (Hansen et al.). In RNA-seq, replicates have a two-fold purpose. First, they demonstrate whether the experiment is repeatable, and secondly, they can reveal differences in gene expression between samples. The correlation between samples is an important indicator for testing the reliability of the experiment. The closer the correlation coefficient is to 1, the greater the similarity of the samples. ENCODE suggests that the square of the Pearson correlation coefficient should be larger than 0.92, under ideal experimental conditions. In this project, the R2 should be larger that 0.8.

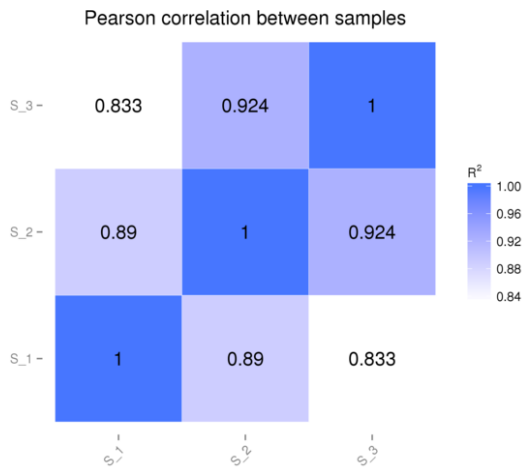


Figure 3.12.3 RNA-Seq Correlation

The x axis and y axis represented the log10(TPM+1)(R2: pearson RSQ; Rho: spearman coefficient of association; Tau: kendall-tau coefficient of association)

3.12.4 Differential expression

The input data is the readcount value from the miRNA expression level analysis. For samples with biological replicates, DESeq2 (Michael et al.，2014) was used to do the analysis. For the samples without biological replicates, TMM was first used to normalize the read count value, and the DEGseq (Wang et al.，2010) was used to do the analysis. The different expression miRNA list was as follows:

Table 3.12.4 result

sRNA	S_3	S_1	log2.Fold_change.	p.value	q.value.Storey.et.al..2003.
mmu-let-7a-5p	4423.124208	13915.21619	-1.6535	0	0
mmu-let-7b-5p	146.9735685	3137.951325	-4.4162	0	0
mmu-let-7c-5p	409.7125447	7830.093851	-4.2563	0	0
mmu-let-7f-5p	25710.67466	72640.10997	-1.4984	0	0
mmu-let-7g-5p	21825.04776	102597.6821	-2.2329	0	0

(1) sRNA: miRNA mature id.

(2) Group1: readcount values of sample1 after normalized.

(3) Group2: readcount values of sample2 after normalized.

(4) log2.Fold_change.: log2(Sample1/Sample2).

(5) p.value : the p.value in hypergenometric test.

(6) q.value.Storey.et.al..2003. : pvalue after normalized.

3.12.5 Filtering the Different Expression miRNA

Volcano plot could be used to infer the overall distribution of different expression miRNAs. For the experiment without biological replicate, the threshold is normally set as: |log2(Fold Change)| > 1 and qvalue < 0.01. For the experiment with biological replicate, as the DESeq2 has already eliminate the biological variation, our threshold is normally set as: padj < 0.05. result Figure 3.12.5:

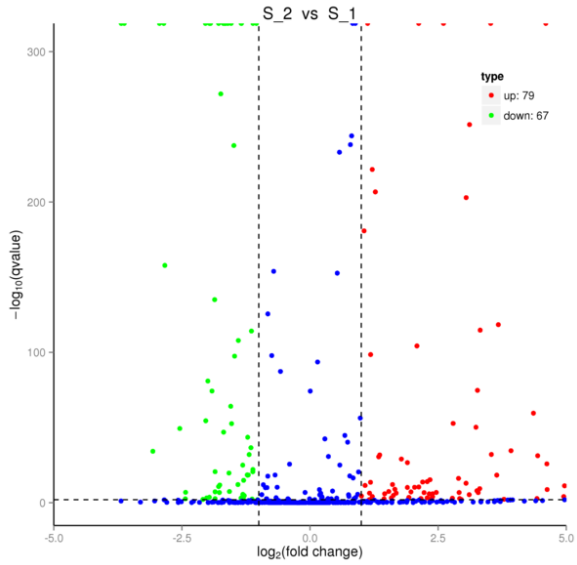


Figure 3.12.5 Volcano Plot

The x-axis shows the fold change in miRNA expression between different samples, and the y-axis shows the statistical significance of the differences. Statistically significant differences are represented by red dots..

3.12.6 Cluster Analysis of miRNAs Expression Difference

The cluster analysis is used to find miRNAs expression patterns under different experiment conditions. By clustering miRNAs with similar expression patterns, the unknown function of miRNAs or the function of unknown miRNAs could be recognized. In the hierarchical clustering, different area with different colours is presenting different groups of the cluster, and miRNAs within each group might have similar functions or take part in a same biological process. In addition to the TPM cluster, K-means and SOM was also be used to cluster the log2(ratios). miRNAs

within the same cluster have the same changing trend in expression levels under different conditions.

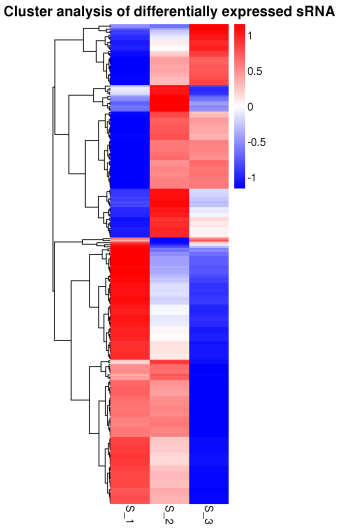


Figure 3.12.6 Cluster Analysis

Figure1: The overall TPM cluster analysis result, clustered by log10(TPM+1) value, red represents miRNAs with high expression level, blue represents miRNAs with low expression level. The colour from red to blue represents the log10(TPM+1) value from large to small. Figure2: log2(ratios) line chart. Every grey line in a sub line chart represents the relative expression value in different experiment conditions of a miRNA cluster, and the blue line represents the mean value of it. The x-axis represents the experiment condition and the y-axis represents the relative expression value.

3.12.7 Difference expression miRNA Venn diagram

The Venn diagram presents the number of miRNAs that are uniquely expressed within each group, with the overlapping regions showing the number of miRNAs that are expressed in two or more groups.（See Figure12.5）。

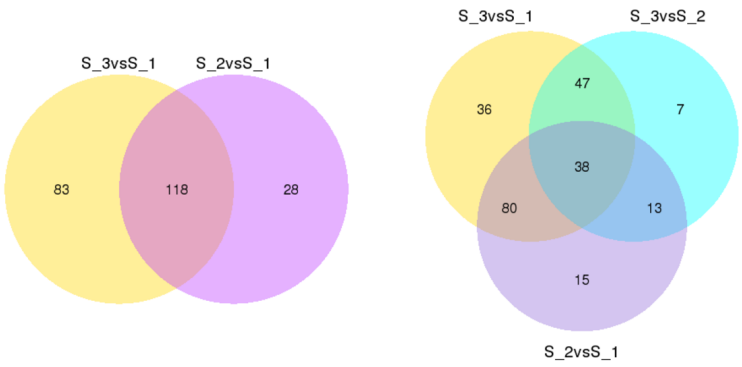


Figure 3.12.7 Difference expression miRNA Venn diagram

The large circle represents the total number of miRNA in a combination. Circle overlapping part is the number of miRNAs expression in all the overlapping combinations.

3.13 Target prediction for known and novel miRNA

Predicting the target gene of known and novel miRNA, and find the relationship between miRNA and target gene. The results are as follows: (the two line are IDs of miRNA and target gene respectively)

novel_106 ENSRNOG00000000091
novel_106 ENSRNOG00000000233
novel_106 ENSRNOG00000000246
novel_106 ENSRNOG00000000257
novel_106 ENSRNOG00000000264
novel_106 ENSRNOG00000000408
novel_106 ENSRNOG00000000415
novel_106 ENSRNOG00000000521
novel_106 ENSRNOG00000000549
novel_106 ENSRNOG00000000568

3.14 Enrichment analysis

3.14.1 GO enrichment analysis

Gene Ontology (GO) is an international standardized classification system for gene function, which supplies a set of controlled vocabulary to comprehensively describe the property of genes and gene products. There are 3 ontologies in GO: molecular function, cellular component and biological process. The basic unit of GO is GO-term, each of which belongs to one type of ontology. GO enrichment analysis is used on predicted target gene candidates of known and novel miRNAs ("target gene candidates" in the following). It will provide all GO terms significantly enriched in the predicted target gene candidates of known and novel miRNAs compared to the reference gene background, as well as the genes corresponding to certain biological function. The result could reveal the functions significantly related with predicted target gene candidates of known and novel miRNAs. This method(Young et al，2010) firstly maps all target gene candidates to GO terms in the database (http://www.geneontology.org/), calculating gene numbers for each term, then using Wallenius non-central hyper-geometric distribution to find significantly enriched GO terms in target gene candidates comparing to the reference gene background.

Table 3.14.1 Result

GO_accession	Description	Term_type	Over_represented_pValue	Corrected_pValue	DEG_item	DEG_list	Bg_item	Bg_list
GO:0005488	binding	nolecular_functio	6.06E-82	2.46E-78	7635	12183	9099	16013
GO:0005515	protein binding	nolecular_functio	2.98E-54	6.06E-51	3723	12183	4294	16013
GO:0003824	catalytic activity	nolecular_functio	1.54E-43	2.08E-40	4845	12183	5697	16013
GO:0008152	metabolic process	iological_process	1.09E-39	1.11E-36	5705	12183	6856	16013
GO:0043167	ion binding	nolecular_functio	5.0407E-33	4.0971E-30	3431	12183	4005	16013

(1) GO accession: Gene Ontology entry.

(2) Description: Detail description of Gene Ontology.

(3) Term type: GO types,including cellular_component,biological_process,molecular_function.

(4) Over represented pValue: P-value in hypergenometric test.

(5) Corrected pValue: Corrected P-value, GO with Corrected P-value < 0.05 are significantly enriched in DEGs.

(6) CAD item: The number of target gene candidates related to this term.

(7) CAD list: The number of target gene candidates with GO Annotation.

(8) Bg item: The number of reference gene related to this term.

(9) Bg list: The number of all gene in GO.

Using histogram to show enriched target gene candidates 3.14.1.1.

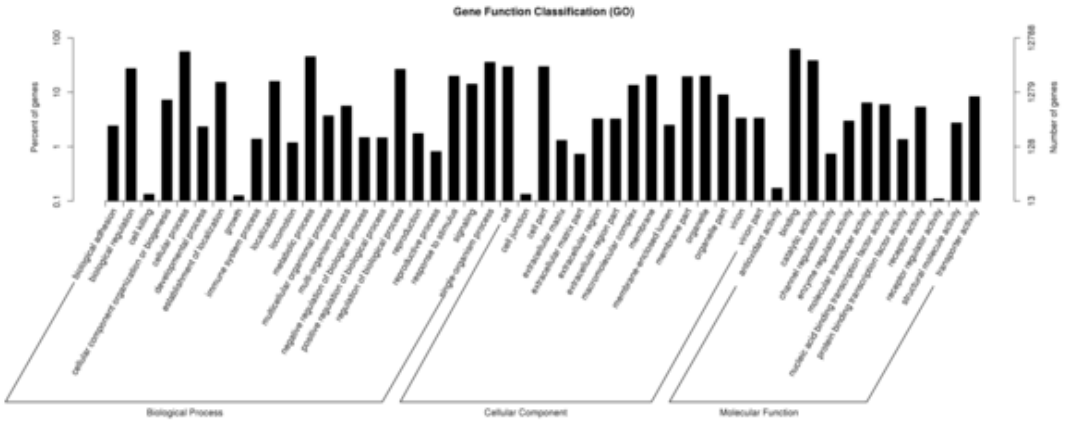


Figure 3.14.1.1 The histogram of target gene candidates

The x axis shown the 3 GO ontologies' next GO term, the y axis shown the number and percentage of target gene candidates annotated in this GO term.

The Directed Acyclic Graph (DAG) is used to visualize the GO enrichment, where branches represent inclusion of the two GO terms, and the scope of the term definitions becomes smaller and smaller from top to bottom. Normally, the top 10 results from GO enrichment are selected as main nodes in directed acyclic graph, where the associated terms are also represented and the depth of colors indicates enrichment level. DAGs for biological process, molecular function and cellular component are shown respectively.

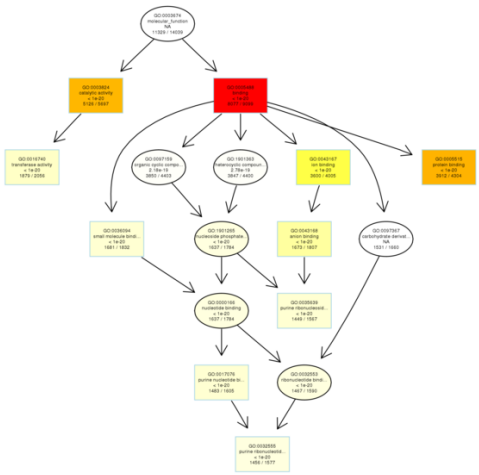


Figure 3.14.1.2 DAGs of GO enrichment

Each node represents a GO term, and TOP10 GO terms are boxed. The darker the color is, the higher is the enrichment level of the term. The name and p-value of each term are present on the node.

3.14.2 KEGG pathway analysis

The interactions of multiple genes may be involved in certain biological functions. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of manually curated databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. KEGG is utilized for bioinformatics research and education, including data analysis in genomics, metagenomics, metabolomics and other omics studies. Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed miRNAs target genes compared with the whole genome background:

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Here N is the number of all genes with KEGG annotation, n is the number of target gene candidates in N, M is the number of all genes annotated to a certain pathway, and m is the number of target gene candidates in M. Genes with BH smaller than 0.05 are considered as significantly enriched in target gene candidates. The KEGG analysis could reveal the main pathways which the target gene candidates are involved.

Table 3.14.2 Pathway annotation result

Term	ID	Sample number	Background number	P-Value	Corrected P-Value
Pathways in cancer	mmu05200	311	323	0.120353739	0.700794044
Metabolic pathways	mmu01100	1152	1256	0.12187924	0.700794044
MAPK signaling pathway	mmu04010	245	253	0.137186665	0.700794044
Oxytocin signaling pathway	mmu04921	157	158	0.140782639	0.700794044
Wnt signaling pathway	mmu04310	142	143	0.155602496	0.700794044

(1) Term: Description of this KEGG pathway.

(2) Id: Unique ID of this pathway in the KEGG database.

(3) Sample number: Number of target genes related to this pathway.

(4) Background number: Number of reference genes related to this pathway.

(5) P-value: P-value in hypergenometric test.

(6) Corrected P-value: Corrected P-value smaller than 0.05 are considered as significantly enriched in target gene candidates.

Scatter plot of KEGG enrichment of target genes:

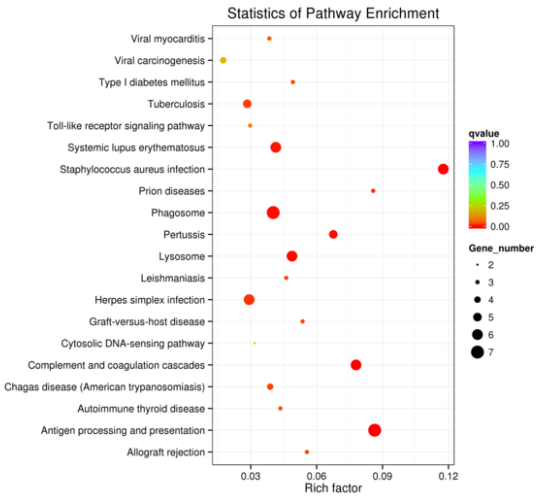
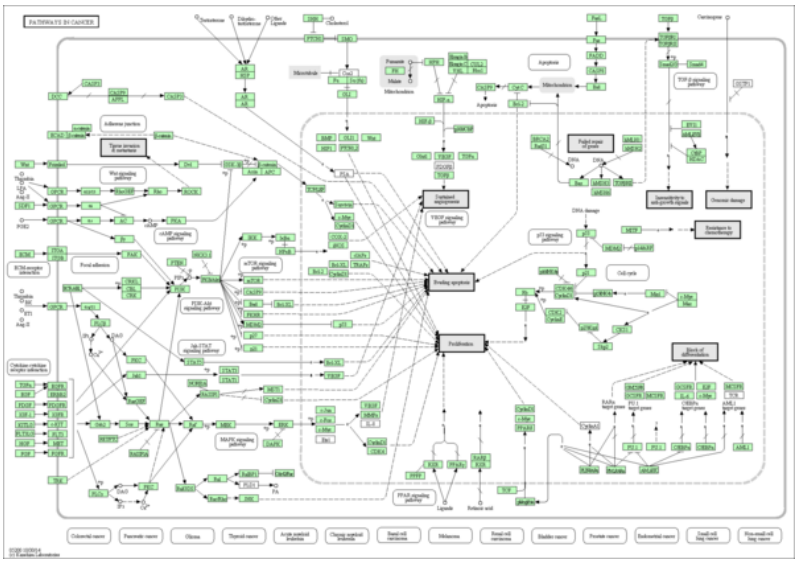


Figure 3.14.2.1 KEGG enrichment scatter plot of DEGs

The y-axis shows the name of the pathway and the x-axis shows the Rich factor. Dot size represents the number of target genes and the color indicates the q-value.

The metabolic map of target genes (**Figure 3.14.2.2**)



4 Reference

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)

Chen D., Yuan C., Zhang J., Zhang Z., Bai L., Meng Y., et al. (2011). Plant NATs DB: a comprehensive database of plant natural antisense transcripts. Nucleic Acids Research 40:D1:D1187–D1193. (Plant NATs DB)

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771.

Erlich, Y., and Mitra, P.P. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nature methods 5, 679-682.

Friedlander M.R., Mackowiak S.D., Li N., Chen W., Rajewsky N. (2011). mi RDeep2 accurately identifies known and hundreds of novel micro RNA genes in seven animal clades. Nucleic Acids Res 40:37-52. (mi RDeep2)

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B.(2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21, 1543-1551.

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)

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)

Moxon S., Schwach F., Mac Lean D., Dalmay T., J Studholme D., and Moulton V. (2008). A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics 24 (19): 2252-2253. (UEA sRNA tools)

Michael I Love,Wolfgang Huber,Simon Anders.(2014).Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.Genome Biology,DOI 10.1186/s13059-014-0550-8. (DESeq2)

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)

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

Wen M., Shen Y., Shi S., and Tang T. (2012). mi REvo: An Integrative micro RNA Evolutionary Analysis Platform for Next-generation Sequencing Experiments. BMC Bioinformatics13:140. (miREvo)

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

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

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

5 Notes

5.1 Result Directory Lists

Click to open the result directory.(Note: Please make sure the report directory and the result directory is under the same directory).

Result Directory Lists: html

```
../../../NHHWXXXXXX_species_results
├── 0.SuppFiles
├── 1.Example_data
│   ├── 1.1.RawData
│   └── 1.2.CleanData
├── 2.QualityControl
│   ├── 2.1.RawData_ErrorRate
│   ├── 2.2.RawData_Stat
│   ├── 2.3.ReadsClassification
│   ├── 2.4.Length_Filter
│   └── 2.5.Common_Specific_sRNA
├── 3.Mapping_Stat
├── 4.Known_miRNA
│   └── Structure_plot_example
├── 5. ncRNA: tRNA\rRNA\snRNA\snoRNA
├── 6.Repeat
├── 7. gene:    gene
├── 8.Novel_miRNA
│   └── Structure_plot_example
├── 9.Category
├── 10.    miRNA_editing: miRNA
├── 11.    miRNA_family:  miRNA
├── 12.DiffExprAnalysis
│   ├── 12.1.miRNAExp
│   ├── 12.2.miRNAExpdensity
│   ├── 12.3.CorAnalysis
│   ├── 12.4.DiffExprAnalysis
│   ├── 12.5.DEsFilter
│   ├── 12.6.DEcluster
│   └── 12.7.DEvenn
├── 13.miRNA_target
└── 14.Enrichment
```

5.2 Software List

Software and Parameter

Name	Version	Description	Main Parameter
Bowtie	bowtie-0.12.9	for mapping	-v 0 -k 1
miREvo	miREvo_v1.1	Modify mirdeep2 for known miRNA analysis ; Integration miREvo and mirdeep2 for novel miRNA prediction ; ViennaRNA for mirdeep2 internal call	-i -r -M -m -k -p 10 -g 50000
mirdeep2 ViennaRNA	mirdeep2_0_0_5 ViennaRNA-2.1.1		quantifier.pl -p -m -r -y -g 0 -T 10 default
srna-tools-cli	http://srna-tools.cmp.uea.ac.uk/	for plant TAS prediction	--tool phasing --abundance 3 --pval 0.001 --minsize 20 --maxsize 26 --trrna
RepeatMasker	open-4.0.3	for repeat analysis , based on RepBase18.07 , using trf and irf	-species –nolow -no_is –norna –pa 8
miRanda	miRanda-3.3a	animal target prediction	-sc 140 –en 10 –scale 4 –strict –out
psRobot	psRobot_v1.2	plant target prediction	-s –t –o –p 5
DEseq2	1.12.0	for Biological repeats analysis	padj<0.05
DEGSeq	1.2.2	for no Biological repeats analysis	qvalue<0.01 |log2foldchange|>1
EdgeR	3.2.4	for special circumstances analysis	padj<0.05 |log2foldchange|>1
GOSeq/topGO	Release 2.12	GO enrichment	enrichmentMethod: Wallenius; padjust:BH
KOBAS	v2.0	KEGG enrichment	blastx 1e-10; padjust:BH