Small RNA Analysis Report

Demo Report

Overseas Department

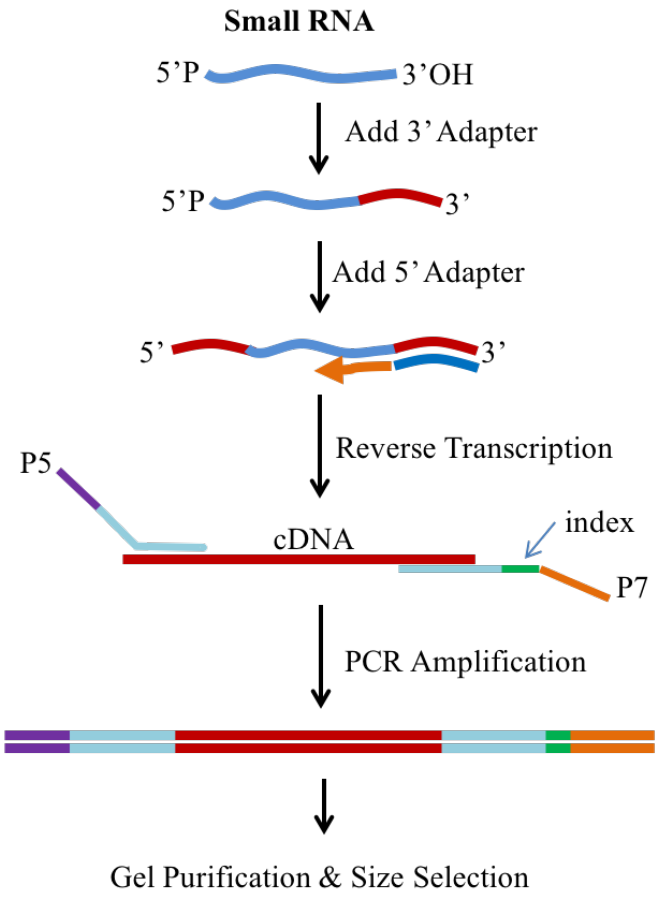
March 1, 2017

Contents

1 Library Preparation and Sequencing

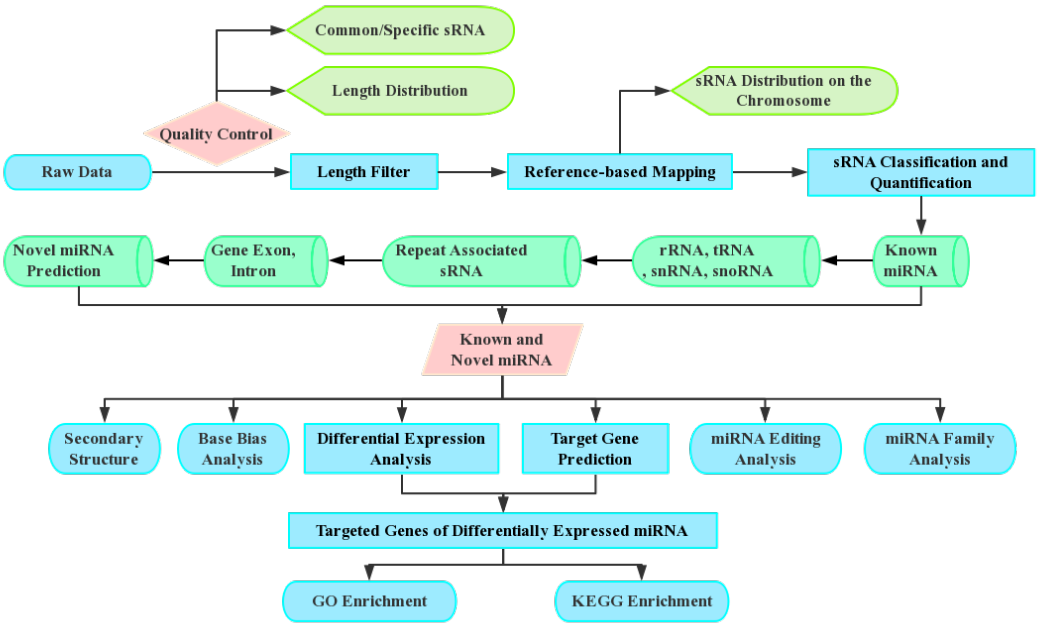
Small RNA is a special kind of molecule in organisms that induces gene silencing and plays an important role in the regulation of cell growth, gene transcription, and gene translation. Small RNA digitalization analysis, based on small RNA digitalization analysis, uses SBS (sequencing by synthesis), has the benefits of small sample requirements, high throughput, high accuracy, and a simple automatic platform.

HiSeq analysis can obtain millions of small RNA sequence tags in one shot, comprehensively identify the small RNA of certain species in certain conditions, predict novel miRNA, and construct a small RNA differential expression profile between samples. This expression can be a powerful tool for researching small RNA functions. Small RNA sequencing experiments are carried out as follows:



2 Data Analysis Process

Considering that the sample was obtained from an animal whose genome had been sequenced, the small RNA analysis process with a reference genome was used. This process is as follows:



3 Results

3.1 Raw Data

Raw picture data was obtained through the high throughput sequencing platform Illumina HiSeq 2500. Base calling was then used to transform these data to sequenced reads, which contain a read sequence and corresponding base quality information in FASTQ format. In this format, a read is stored as four lines.

Each line contains the following information (Cock et al. 2010):

- Line 1: the at sign (@) followed by Illumina sequence identifiers and optional description information
- Line 2: base sequence (A, G, C, and T)
- Line 3: the plus sign (+) optionally followed by the same Illumina sequence identifiers and description information
- Line 4: the quality of each base, corresponding to the data on line 2

The following is an example:

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458
1:N:0:CGATGT
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHHFAE0ECFFD/AEHH
```

The following table describes the Illumina sequence identifiers in the preceding read:

Identifier	Description
HWI-ST1276	Unique identifier of the sequencing instrument
71	Instrument run number
C1162ACXX	Flowcell ID
1	Lane number (positive integer)
1101	Tile number (positive integer)
1208	X coordinate of the spot (integer)
2458	Y coordinate of the spot (integer)
1	Read number (1 for single reads; 1 or 2 for p aired ends)
N	Filtering (Y if the read is filtered out and not in the delivered FASTQ file; N otherwise)
0	Control number (0 if none of the control bits are on; an even number otherwise)
CGATGT	Illumina index sequences

The ASCII value corresponding to each character in the fourth line, subtracted by 33, is equal to the sequencing quality value of the corresponding base in the second line. The relationship between sequencing error rate E and sequencing quality Q_{phred} is described by the following formula:

$$Q_{\text{phred}} = -10 \log_{10} E$$

The relationship between Phred quality score Q_{phred} and base calling error E was predicted using Illumina Casava version 1.8. It was found that Phred quality score was logarithmically linked to base calling error:

Phred score	Base calling error rate	Base calling accuracy	Q-score
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(Gbp)	Error rate	Q20	Q30	GC content
S_1	12047960	0.602	0.01%	97.20%	93.93%	49.36%
S_2	12174594	0.609	0.01%	97.12%	93.75%	49.95%
S_3	12748703	0.637	0.0001	0.9713	0.9384	0.4961

(1) Sample: Sample ID

(2) Reads: Statistics of the original sequence data

(3) Bases: Sequence number multiplied the length of the sequence, expressed in gigabase pairs (Gbp)

(4) Error rate: Sequencing error rate

(5) Q20: Percentage of bases whose Phred values exceed 20

(6) Q30: Percentage of bases whose Phred values exceed 30

(7) GC content: G and C bases as a percentage of all bases

3.2.2 Data Cleaning

The raw reads obtained included low quality and other problematic reads, such as those with missing insert tags, oversized inserts, poly(A) tags, and small tags. Data cleaning was therefore performed on the FASTQ file to obtain the final clean reads.

Data cleaning was performed as follows:

1. Reads in which more than 50% of bases had a Q_{phred} less than or equal to 5 were discarded.
2. Reads in which *N* accounted for more than 10% were discarded. (*N* indicates that base information was indeterminable.)
3. Reads with 5' primer contamination were discarded.
4. Reads lacking a 3' primer or insert tag were discarded.
5. The 3' primer sequence was trimmed.
6. Reads with poly(A), poly(T), poly(G), or poly(C) tails were discarded.

Small RNA adapter 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 Contamination	3' Adapter or Insert Missing	With Ploy(A)/(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 Contamination: Percentage of reads with 5' adapter contamination

(6) 3' Adapter or Insert Missing: Percentage of reads without a 3' adapter or insert

(7) With Poly(A)/(T)/(G)/(C) : Percentage of reads with poly(A), poly(T), poly(G), or poly(C) tails

(8) Clean Reads: Total clean reads followed by clean reads as a percentage of raw reads

3.2.3 Length Distribution

The length of sRNA typically ranges from 18 to 40 nucleotides (nt). Analyzing length distribution helps determine the composition of small RNA samples. For example, miRNA is normally 21 or 22 nt, siRNA is 24 nt, and piRNA is between 28 and 30 nt.

Length distribution varies between plants and animals. In plants, there are often peaks for 21 or 24 nt, whereas peaks for 22 nt are normal in animals.

Table 3.2.3 Type and quantity of sRNA

Sample	Total Reads	Total Bases (bp)	Unique Reads	Unique Bases (bp)
S_1	9742471	219606009	600206	14119632
S_2	9550931	220775560	1301461	31610866
S_3	10041867	231329567	966227	23095143

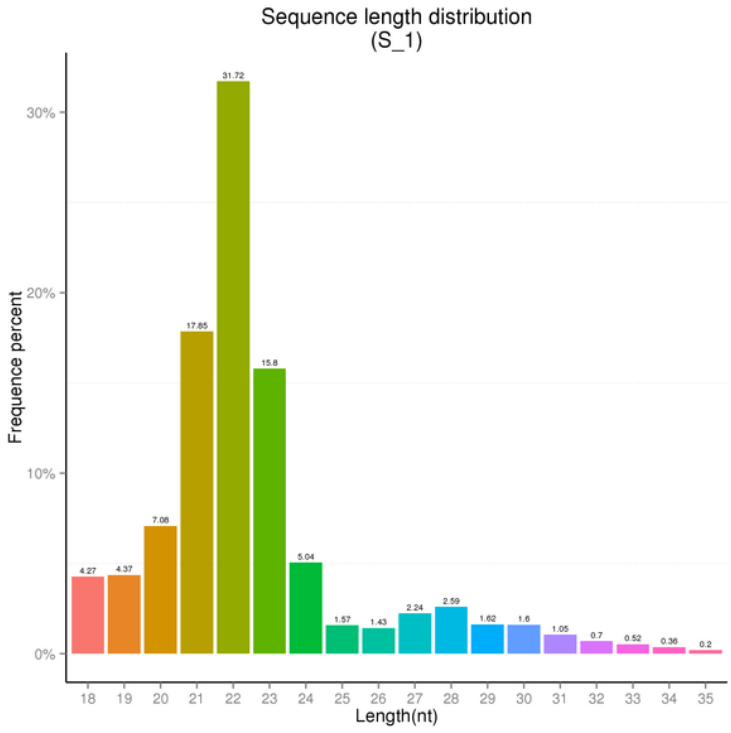
(1) Sample: Sample ID

(2) Total reads: Total number of sRNA reads

(3) Total bases (bp): Total reads multiplied by sequence length

(4) Unique reads: Types of sRNA

(5) Unique bases (bp): Unique reads multiplied by sequence length



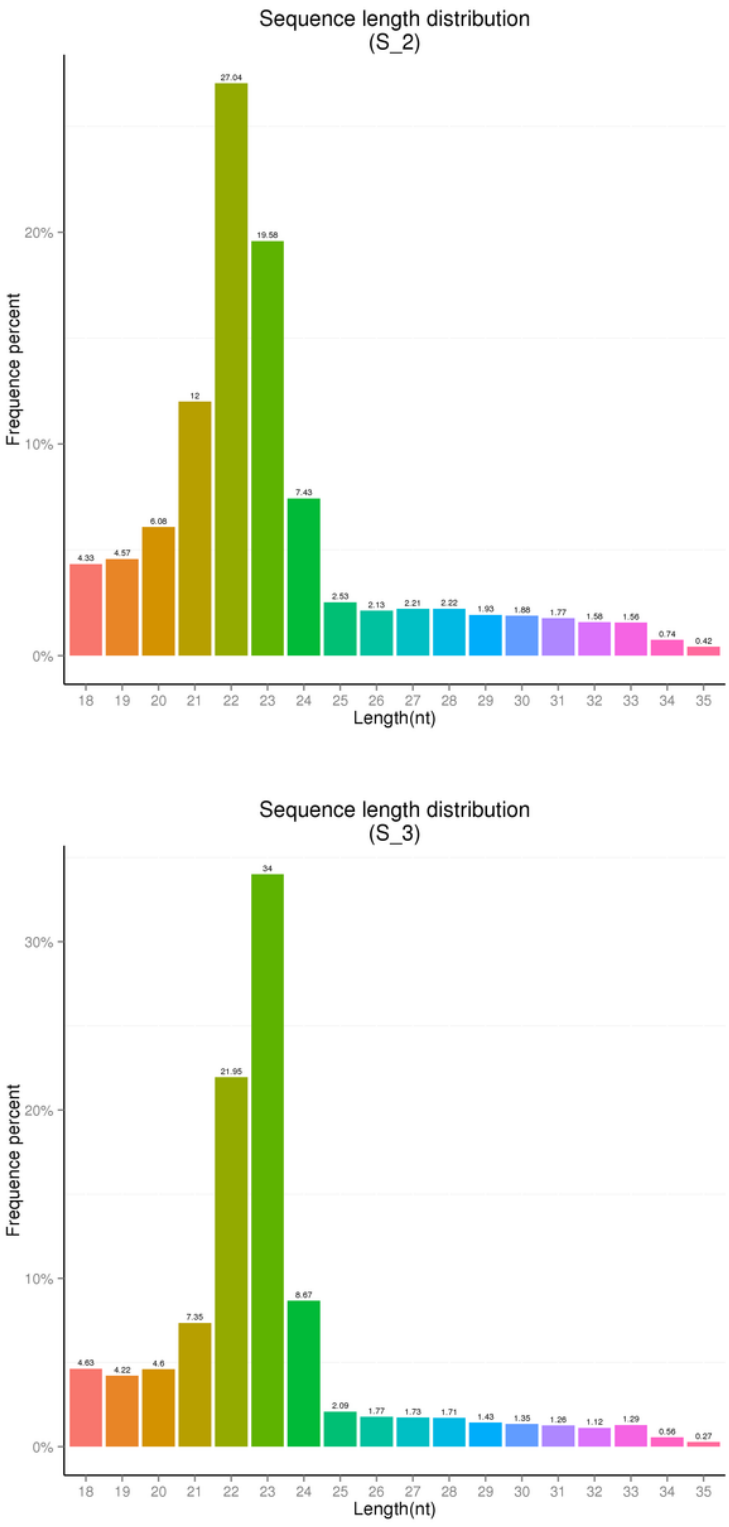
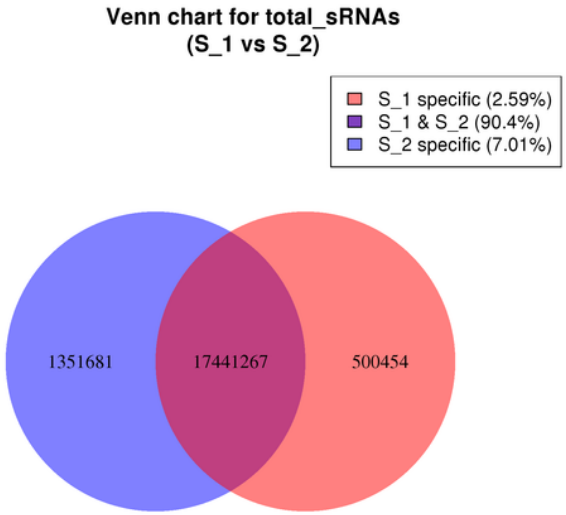


Figure 3.2.3 Length distribution of total sRNA

Note: The abscissa is the length of sRNA reads. The ordinate is one length read as a percentage of total sRNA.

3.2.4 Common and Specific Reads between Samples

The common and specific reads of two samples, including unique and total reads, were summarized. It was observed that a great difference in reads exists among different samples but that common reads are concentrated. This demonstrates that the overall sequencing uniformity of different samples is good.



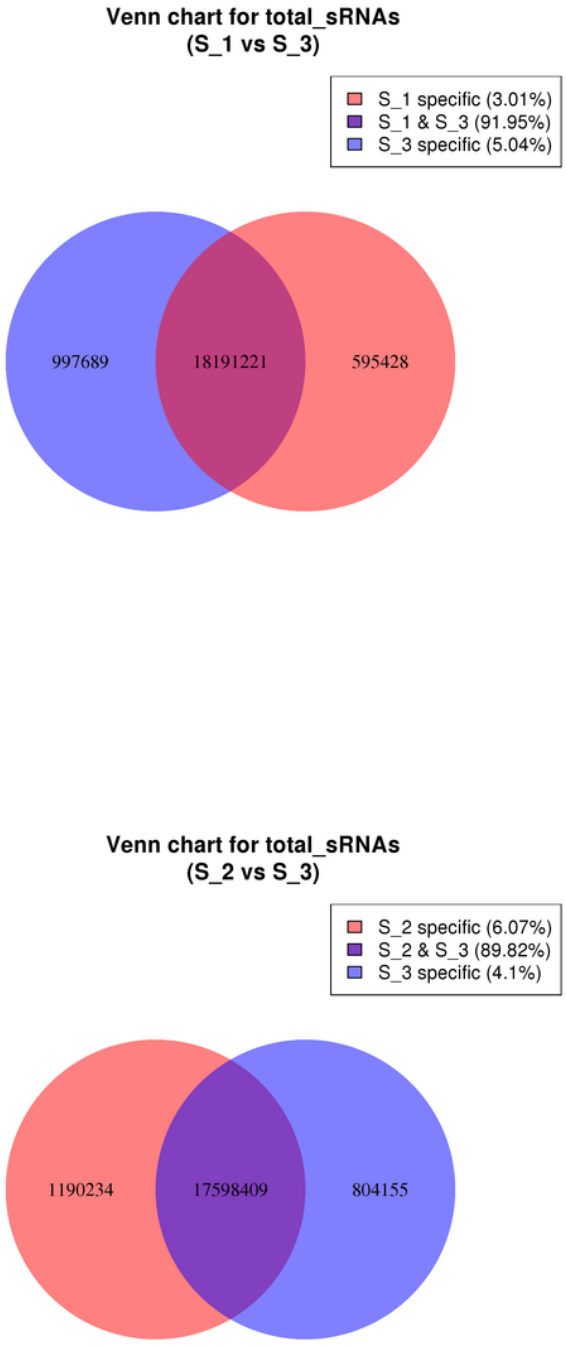
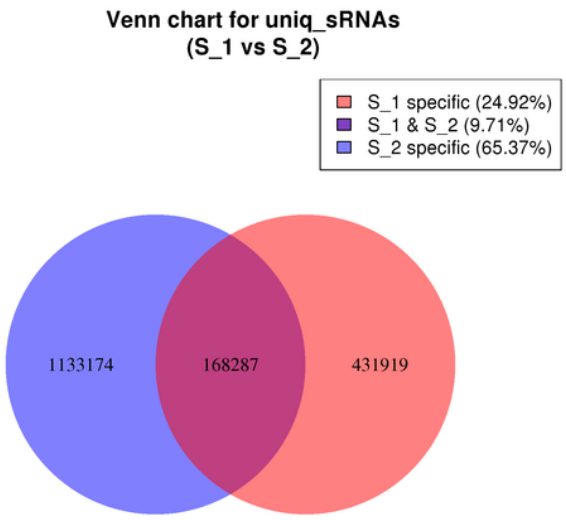
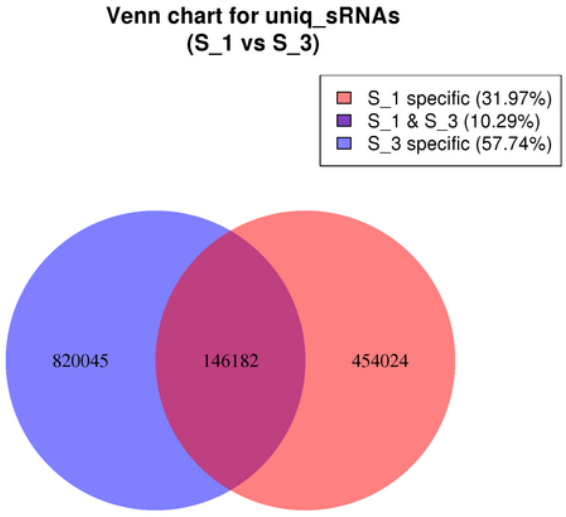


Figure 3.2.4.1 Common and specific reads between samples (total sRNA)


Venn chart for uniq_sRNAs
(S_1 vs S_2)
S_1 specific (24.92%)
S_1 & S_2 (9.71%)
S_2 specific (65.37%)
1133174
168287
431919


Venn chart for uniq_sRNAs
(S_1 vs S_3)
S_1 specific (31.97%)
S_1 & S_3 (10.29%)
S_3 specific (57.74%)
820045
146182
454024

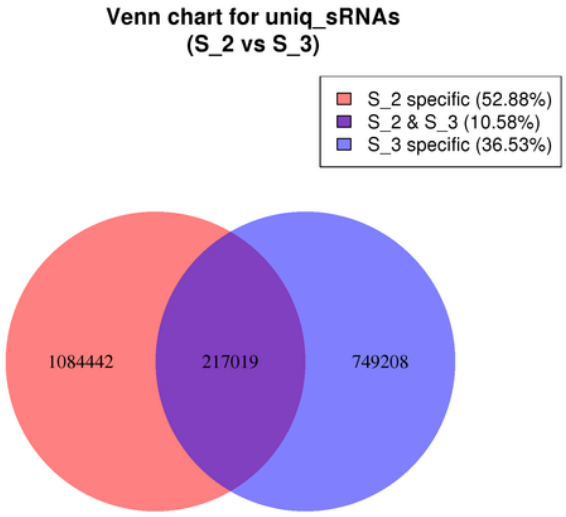


Figure 3.2.4.2 Common and specific reads between samples (Unique sRNA)

(1) Sample1 specific: Reads specific to sample 1

(2) Sample1 & Sample2: Reads common to sample 1 and sample 2

(3) Sample2 specific: Reads specific to sample 2

3.3 Mapping to Genome

The small RNA reads were mapped to the genome using Bowtie so as 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: Number of total sRNA after length filtering

(3) Mapped sRNA: Number and percentage of sRNA mapped to genome

(4) + Mapped sRNA: Number and percentage of mapped sRNA in the same direction as the genome

(5) – Mapped sRNA: Number and percentage of mapped sRNA in the opposite direction to the genome

The density of small RNA reads on each chromosome was determined for each sample. Circos was used to view the distribution of reads on each chromosome. The longest 10 contigs or scaffolds were chosen for analysis.

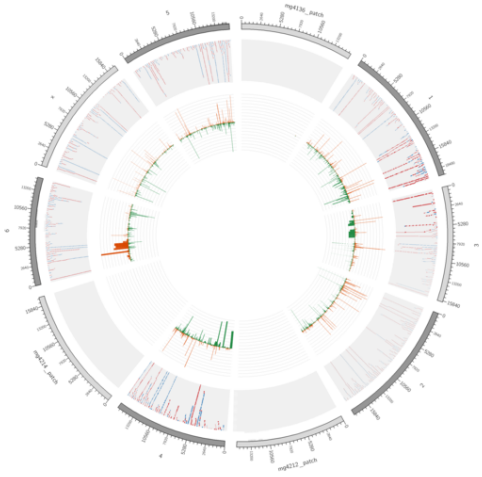


Figure 3.3 Reads distribution per chromosome

The chromosome is shown on the outer circle. The grey background in the middle area shows the distribution of 10,000 reads on the chromosome. Red represents the number of sRNAs on the sense strand of the chromosome, and blue represents the number of sRNAs on the antisense strand. All reads are shown in the center area of the circle. Yellow represents the number of sRNAs on the sense strand of the chromosome, and green represents the number of sRNAs on the antisense strand.

3.4 Analysis of Known miRNA

The above reads on the reference sequence were compared with the specified sequence in the miRBase to obtain details of the sRNAs on each sample match, including the secondary structure of the matched miRNAs, the sequence of miRNAs in each sample, length, the number of occurrences and other information. The specificity of the cleavage site makes the first base of the miRNA mature sequence highly biased, so the miRNAs with different length are also carried out. The first base site distribution, in addition to the base of each site of miRNA distribution statistics.

The results are shown in the tables and figures below.

Table 3.4.1 Summary of known miRNA in each sample

sRNAs	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

(2) Mapped hairpin: The number of sRNAs 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

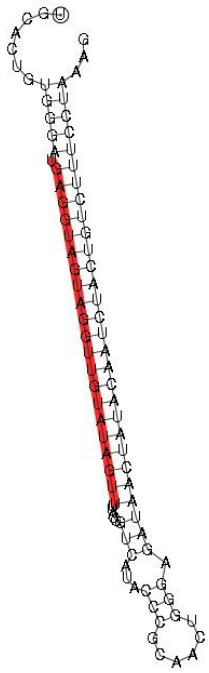


Figure 3.4 The secondary structure of the known miRNAs on partial schematic matches. The entire sequence is an miRNA precursor, the red section is the mature sequence.

Table 3.4 Known miRNA expression profiles

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

(1) miRNA: Mature miRNA id

(2) Columns 2-n+1 show the corresponding sample's read count

The details of mapped known miRNA:

>hsa-let-7a-1

total read count 1792398

hsa-let-7a-3p read count 3817

hsa-let-7a-5p read count 1788190

remaining read count 391

exp fffff5555555555555555555555ffffffffffffffffffffffffffffff3333333333333333333fff

pri_seq ugggaugagguaguagguuguauaguuuuagggucacacccaccacugggagauaacuauacaaucuacugucuuuccua

pri_struct ((((((.(((..((((((((((((((((...(((.....))).((....)).....))))))))))))))))..)))))))) #MM +/-

T_1_333432_x1 ....augagguaguagguuguauagu..... 0 +

T_2_323252_x2 ....augagguaguagguuguauagu..... 0 +

> 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

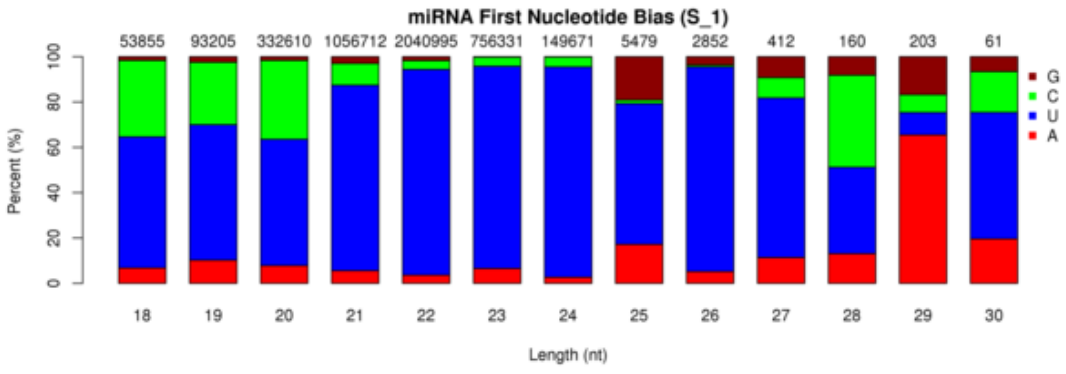


Figure 3.4.2 miRNA first nucleotide bias

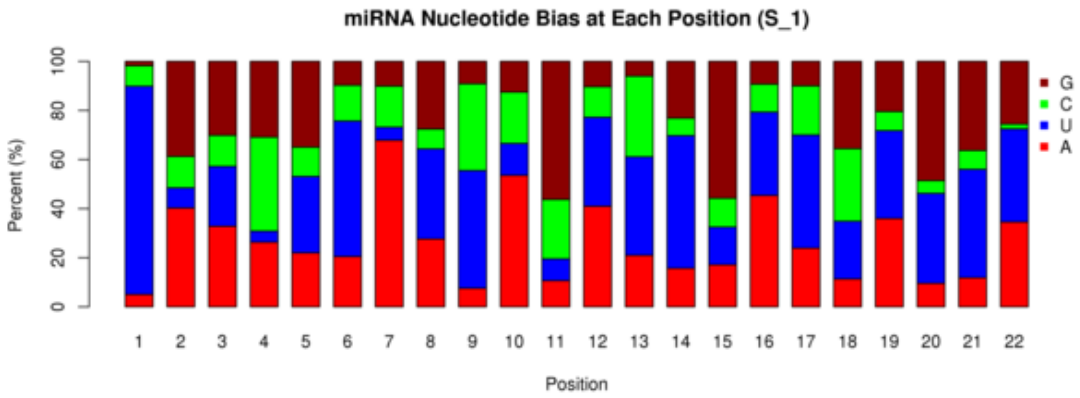


Figure 3.4.3 miRNA nucleotide bias at each position

3.5 ncRNA Analysis

Small RNA reads were annotated with sequences from the ncRNA sequence of the species, If there is a ncRNA annotation of the species. if not, small RNA reads were annotated with sequences from Rfam. Then the matched reads from rRNA, tRNA, snRNA, and snoRNA were removed.

The following table shows how many reads are mapped to different kinds of ncRNA.

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

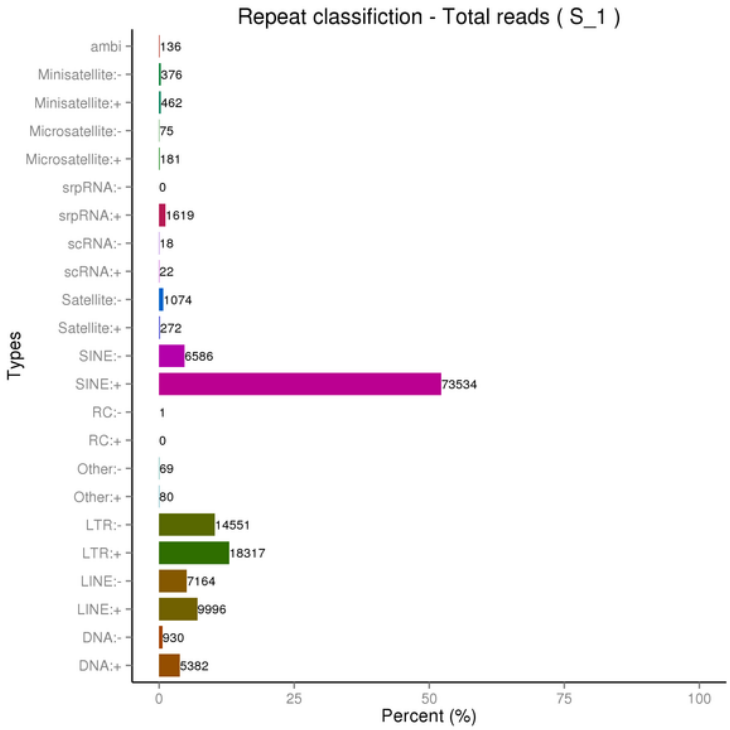
(1) Types: The kinds of ncRNA

(2) Columns 2-n+1 show the corresponding samples reads mapping to that kind of ncRNA

3.6 Repeat Associated RNA alignment

If the species repeat transponson information, the sequence information should be repeated to annotate the measured sRNA. Otherwise the ab initio repeats are performed based on reference sequence information and the sRNA is aligned with the repeats. Statistics on the various repeats repeat sRNAs (expressed in uniq) and the number of sRNAs.

Results of the repeat classifiction are shown in the following figures.



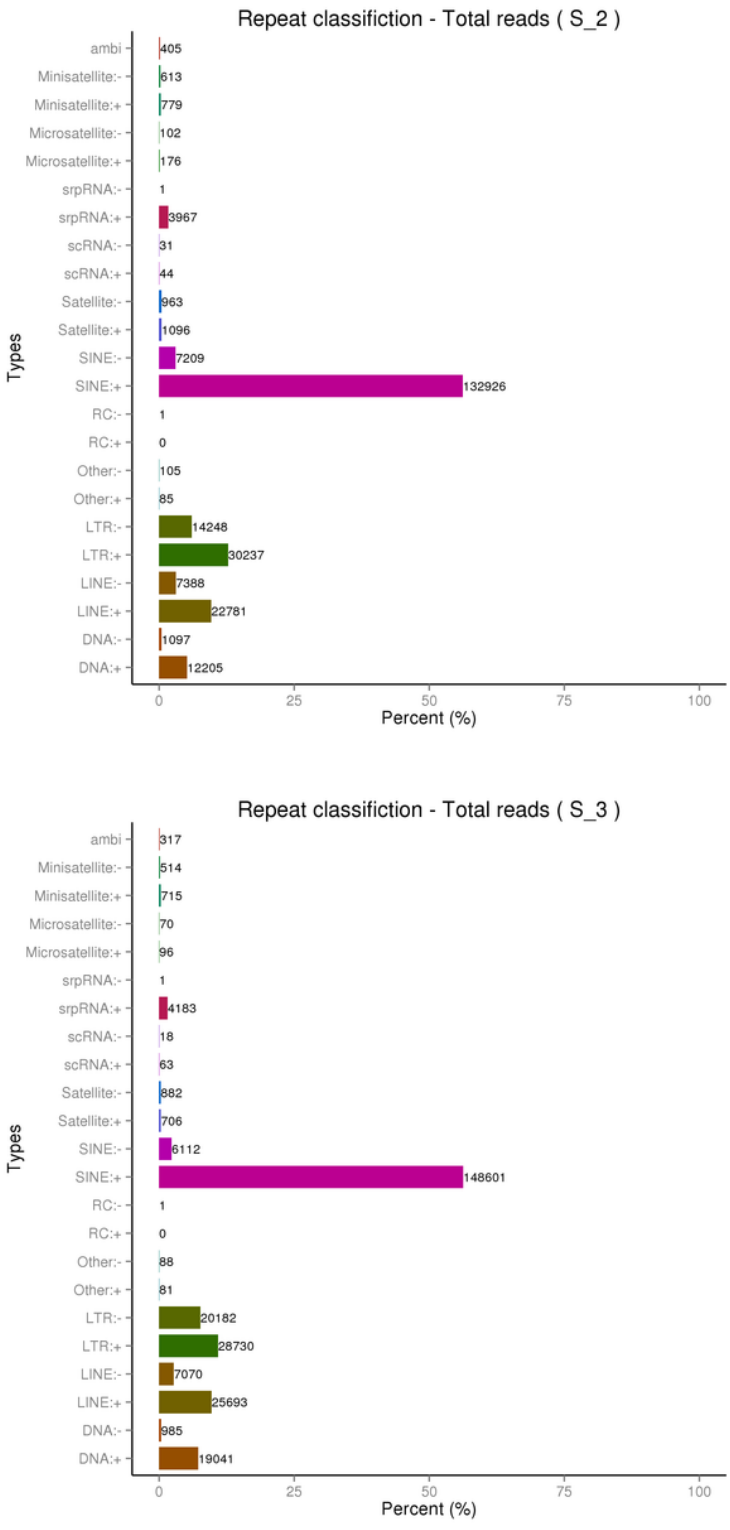
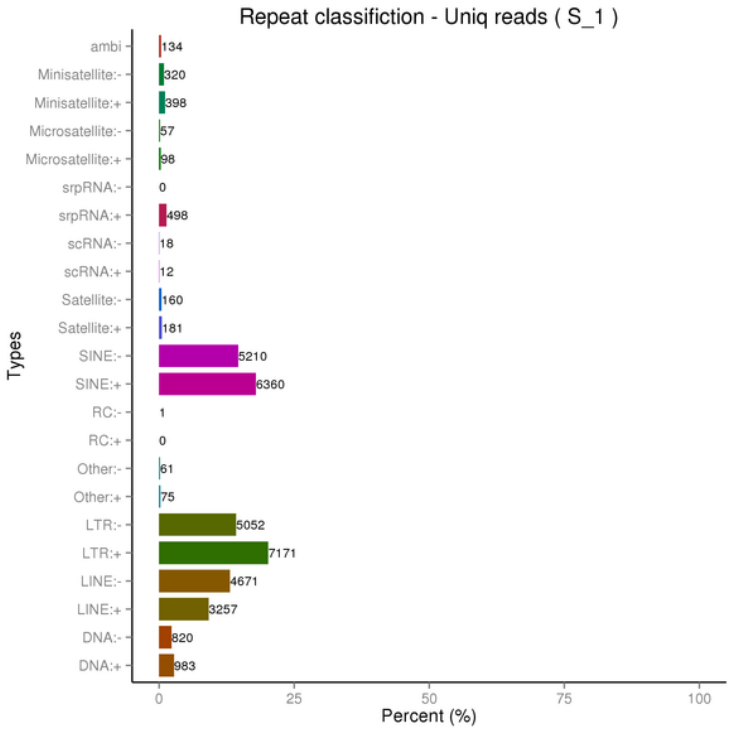
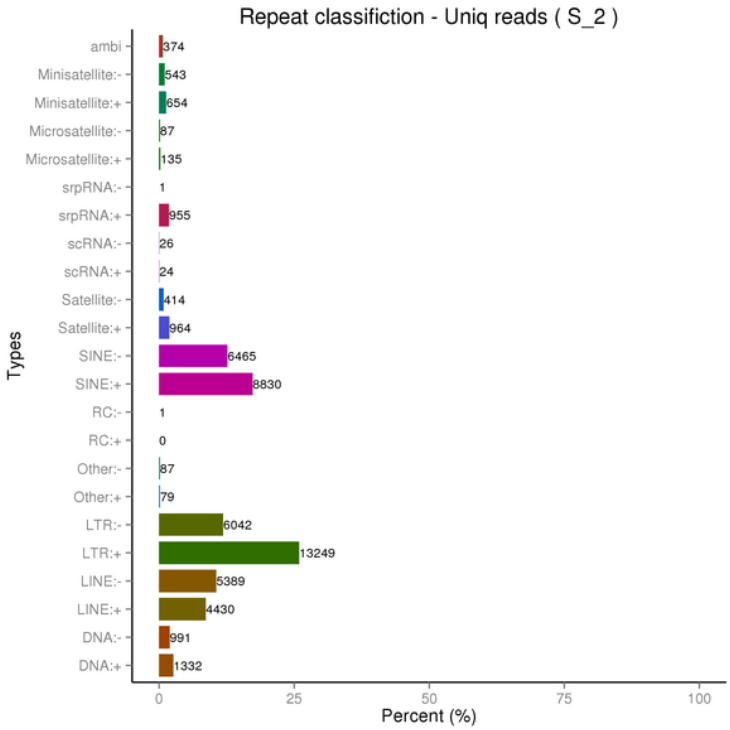


Figure 3.6.1 Total repeat reads


Repeat classifiction - Uniq reads (S_1)
ambi 134
Minisatellite:- 320
Minisatellite:+ 398
Microsatellite:- 57
Microsatellite:+ 98
srpRNA:- 0
srpRNA:+ 498
scRNA:- 18
scRNA:+ 12
Satellite:- 160
Satellite:+ 181
SINE:- 5210
SINE:+ 6360
RC:- 1
RC:+ 0
Other:- 61
Other:+ 75
LTR:- 5052
LTR:+ 7171
LINE:- 4671
LINE:+ 3257
DNA:- 820
DNA:+ 983
Types
0 25 50 75 100
Percent (%)


Repeat classifiction - Uniq reads (S_2)
ambi 374
Minisatellite:- 543
Minisatellite:+ 654
Microsatellite:- 87
Microsatellite:+ 135
srpRNA:- 1
srpRNA:+ 955
scRNA:- 26
scRNA:+ 24
Satellite:- 414
Satellite:+ 964
SINE:- 6465
SINE:+ 8830
RC:- 1
RC:+ 0
Other:- 87
Other:+ 79
LTR:- 6042
LTR:+ 13249
LINE:- 5389
LINE:+ 4430
DNA:- 991
DNA:+ 1332
Types
0 25 50 75 100
Percent (%)

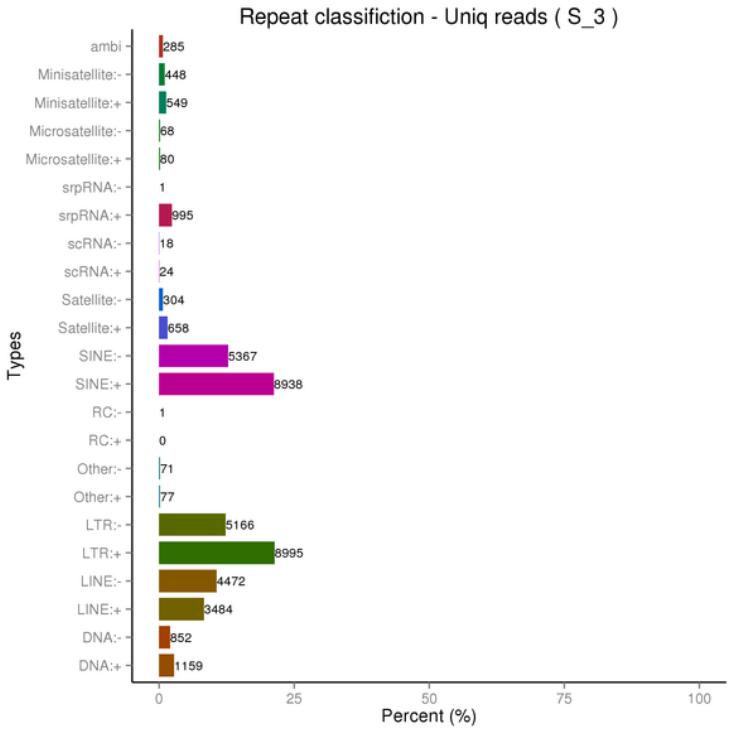


Figure 3.6.2 Unique repeat reads

3.7 Exon and Intron Alignment

Small RNA reads were aligned to the exons and introns of mRNA to find the degraded fragments of mRNA in the small RNA reads. The results are shown in the following table.

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

(1) Types: Types of exon and intron

(2) Columns 2-n+1 show each corresponding sample's results

3.8 Novel miRNA Prediction

The characteristic hairpin structure of miRNA precursors can be used to predict novel miRNA. We used miREvo (Wen et al., 2012) and mirdeep2 (Friedlander et al., 2011) to predict novel miRNA. The results are shown in the following tables and figures.

Table 3.8.1 Summary of novel miRNA

Type	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) Type: The type of sRNA mapping

(2) The total amount

(3) Columns 3-n+2 show the amount aligned to the predicted hairpin in each corresponding sample

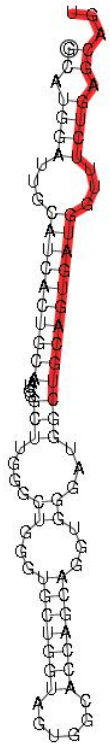


Figure 3.8.1 The secondary structure of the known miRNAs on partial schematic matches. The entire sequence is an miRNA precursor, the red section is the mature sequence

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

(1) miRNA: Mature miRNA id

(2) Columns 2-n+1 show the corresponding sample's read count

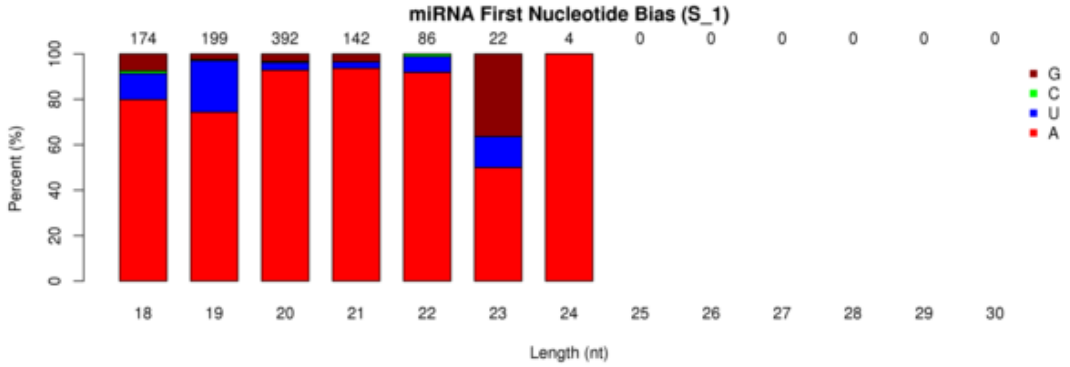


Figure 3.8.2 miRNA first nucleotide bias

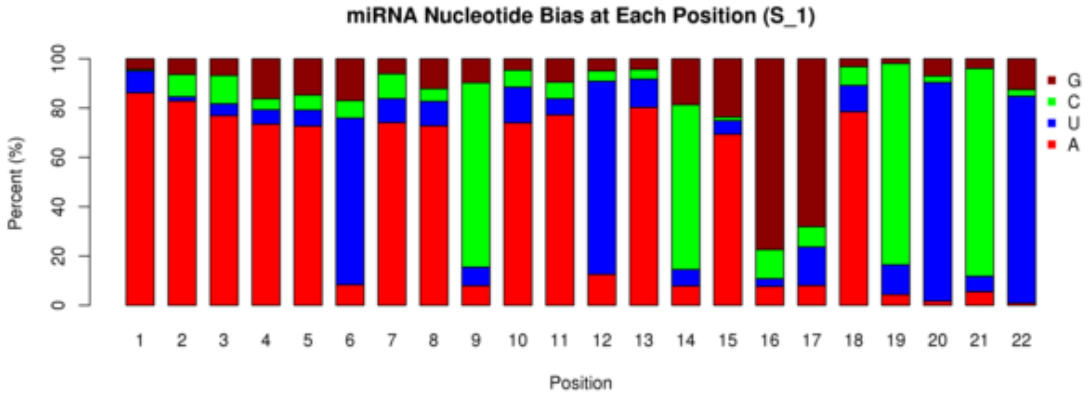


Figure 3.8.3 miRNA nucleotide bias at each position

3.9 Small RNA Annotation

All sRNAs were compared to all kinds of RNA and results were summarized in the comments. Because of the existence of a single sRNA, comparing several comparison results each unique sRNA must have a unique annotation. These annotations are:

Known miRNA > rRNA > tRNA > snRNA > snoRNA > repeat > gene > novel miRNA

This sequence traverses sRNA.

The results of this process are shown below.

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 mapped to known miRNA, ncRNA, repeat, novel miRNA, exon/intron.

3.10 miRNA Base Edit

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

Results:

>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%

Pre-miRNA: The number of reads mapped to precursor, the number and percentage of reads with base edits

Mature miRNA: The number of reads mapped to mature miRNA, and the number and percentage of reads with base edits

Site[1-n]: The details of each base of this mature miRNA. Each line represents the number and percentage of reads with base edits

3.11 miRNA Family Analysis

The following table explores the occurrence of known miRNA and novel miRNA families identified from samples of other species. A "+" means that the miRNA family exists in the species, and a "-" means the miRNA family does not exist in the 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

The expression of known and unique miRNAs in each sample were statistically analyzed and normalized by TPM (Zhou et al., 2010)

The normalized expression = (read count*1,000,000)/libsize. Libsize is the sample miRNA read count.

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) Column 1 shows the miRNA mature ID of each

(2) Columns 2-n+1 show the read count of the corresponding sample

(3) Columns n+2-2n+1 show the read count of the corresponding sample (TPM normalization)

3.12.2 miRNA TPM Distribution

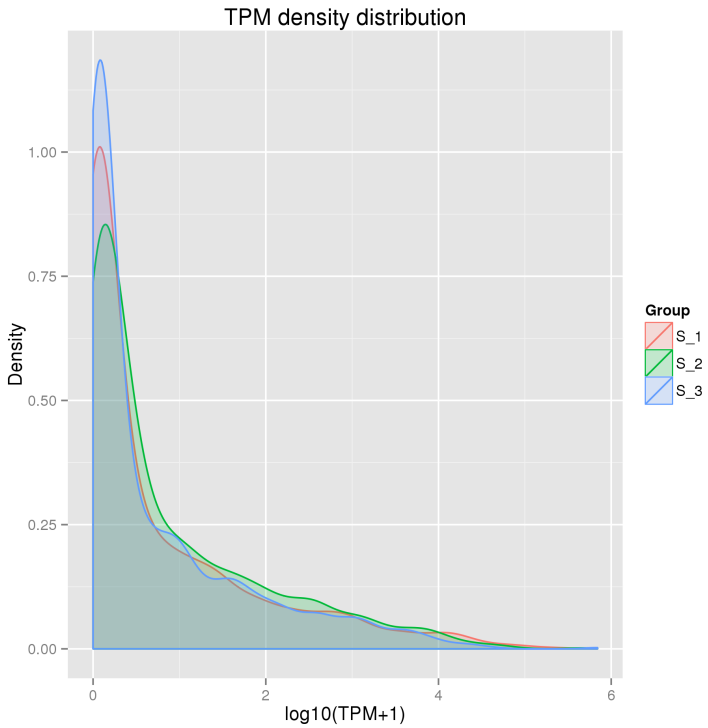


Figure 3.12.2 TPM distribution

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 serve two purposes. 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 an 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 than 0.8.

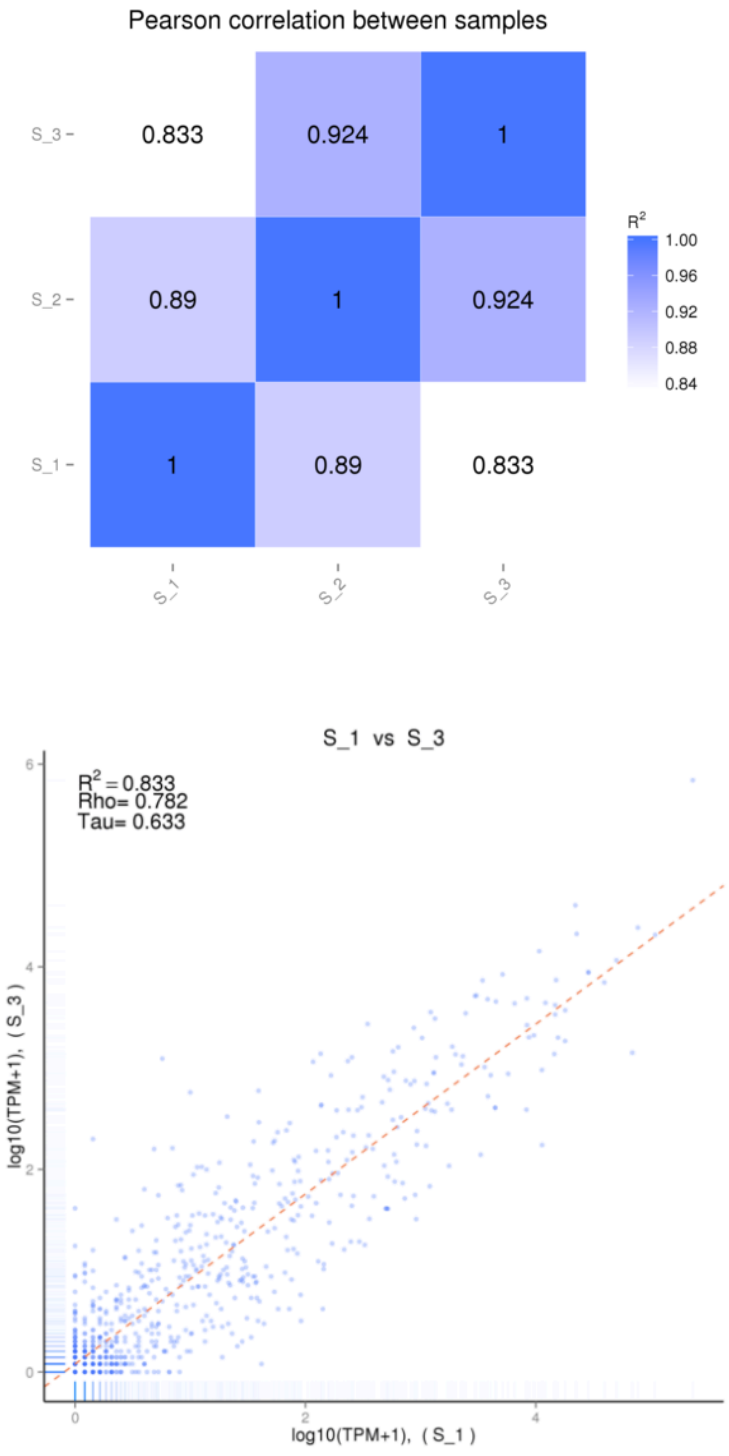


Figure 3.12.3 RNA-Seq Correlation

The x axis and y axis represent the log10(TPM+1)

R^2: Pearson RSQ

Rho: Spearman correlation coefficient

Tau: Kendall-tau correlation

3.12.4 Differential Expression

The read count value from the miRNA expression level analysis was used to create an miRNA expressions list. For samples with biological replicates, DESeq2 (Michael et al.，2014) was used to complete 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 complete the analysis. The different miRNA expressions list is 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: Read count values of Sample1 after normalization

(3) group2: Read count values of Sample2 after normalization

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

(4) p.value: The p value in a hypergenometric test

(5) q.value.Storey.et.al..2003.: p value after normalization

3.12.5 Filtering Different Expression miRNA

A volcano plot could be used to infer the overall distribution of different miRNA expressions. For an experiment with no biological replicate, the threshold is normally set as follows:

$$|\log2(\text{Fold Change})| > 1$$

$$\text{qvalue} < 0.01$$

For experiments with a biological replicate, as the DESeq2 has already eliminated the biological variation, our threshold is normally set to the following:

padj < 0.05

Figure 3.12.5 shows the results of a volcano plot.

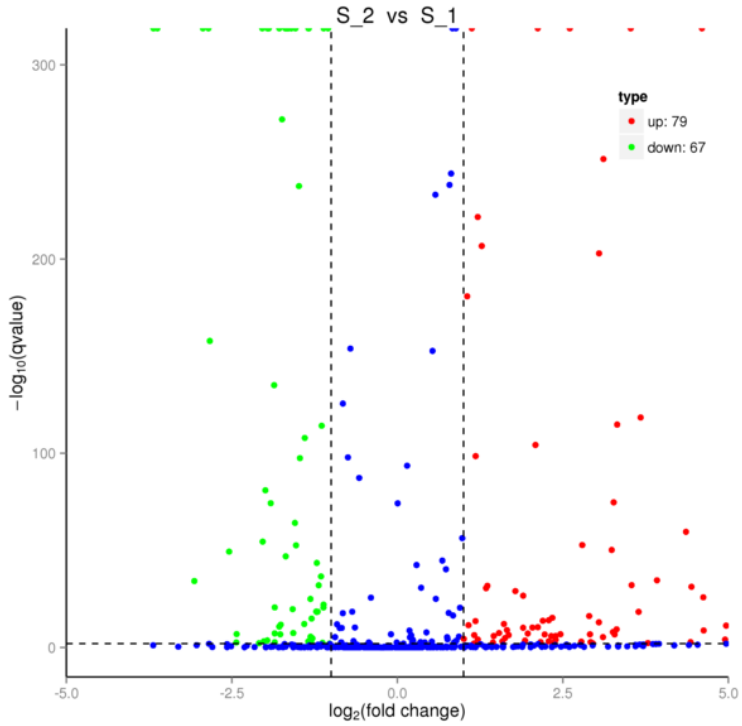


Figure 3.12.5 Volcano Plot

Statistically significant differences are are represented by red dots.

3.12.6 Cluster Analysis of the Differences between miRNAs Expressions

Cluster analysis is used to find miRNA expression patterns under a variety of experiment conditions. By clustering miRNAs with similar expression patterns, it is possible to recognize unknown functions of miRNAs and/or the function of unknown miRNAs.

In hierarchical clustering, differently-colored areas represent different groups of the cluster. miRNAs within each group may have similar functions or take part in the same biological process. In addition to the TPM cluster, K-means and SOM were also used to cluster the log2(ratios). miRNAs within the same cluster have the same changing trend in expression levels under different conditions.

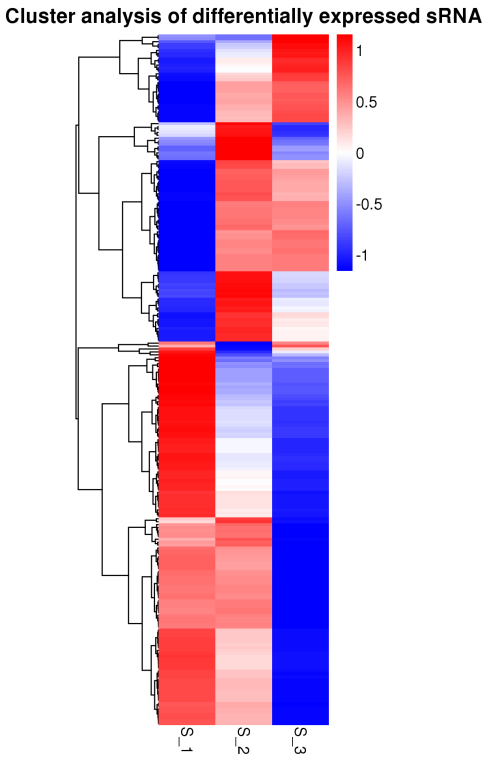


Figure 3.12.6 Cluster Analysis

- The three rows represent the overall TPM cluster analysis result, clustered by log10(TPM+1) value.
- The spectrum from red to blue represents the log10(TPM+1) value from large to small
- The lines near the bottom are a log2(ratios) line chart. Every grey line represents the relative expression value in different experiment conditions of a miRNA cluster and the blue line represents its mean value
- The x-axis represents the relative expression value and the y-axis represents the experiment conditions

3.12.7 Different Expression miRNA Venn Diagram

The Venn diagrams in figure 3.12.7 represent 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 (as shown in Figure 12.5).

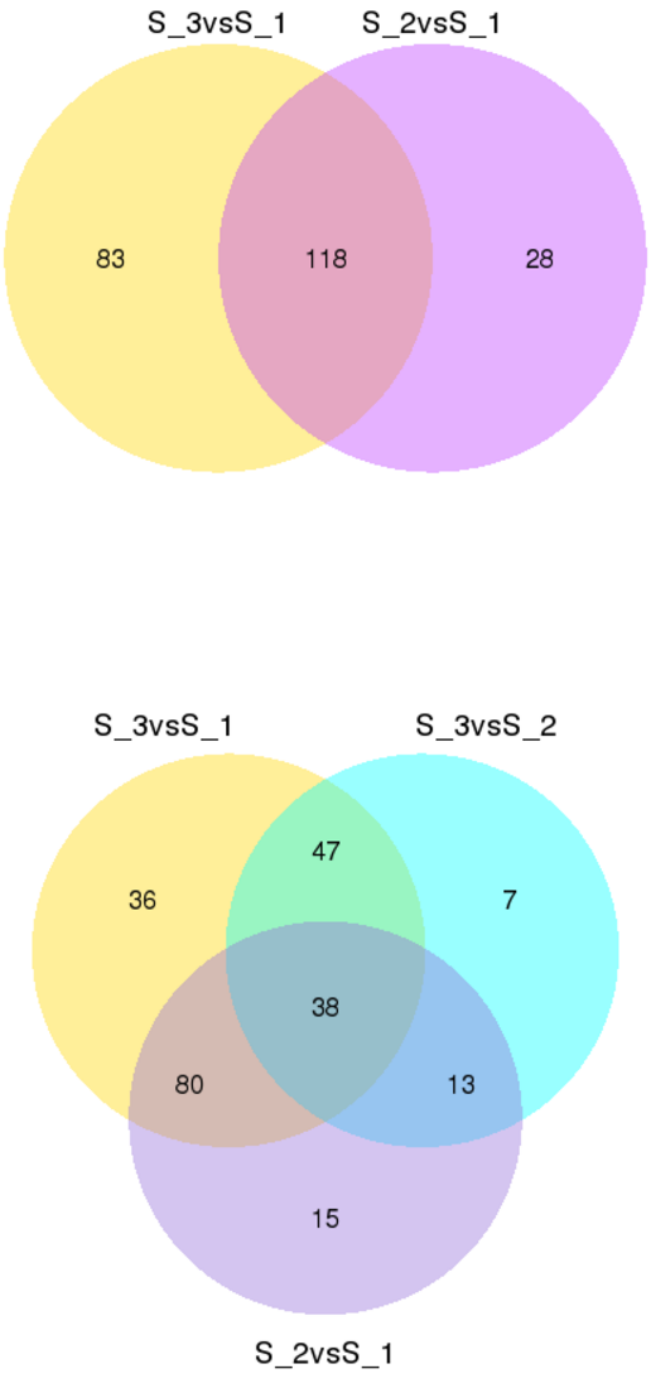


Figure 3.12.7 Different miRNA expressions Venn diagram

Each circle represents the total number of miRNAs in a combination. Each overlap represents the number of miRNA expressions in the corresponding combinations.

3.13 Target Gene Prediction for Known and Novel miRNA

The target gene of known and novel miRNAs were predicted, and the relationships between miRNAs and the corresponding target genes were found. The results are as follows:

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
- Biological process

The basic units of GO are GO-terms, each of which belong to one type of ontology. GO enrichment analysis is used on predicted target gene candidates of known and novel miRNAs. GO enrichment analysis will provide GO terms for predicted target gene candidates of known and novel miRNAs, which can refer to the genes reference background and biological functions. The results could reveal the functions related to the 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. It then uses Wallenius non-central hyper-geometric distribution to find significantly enriched GO terms in target gene candidates relative to the reference gene background.

Table 3.14.1 Results

GO Accession	Description	Term Type	Over Represented pValue	Corrected pValue	DEG Item	DEG List	Bg Item	Bg List
GO:0005488	binding	molecular_function	6.06E-82	2.46E-78	7635	12183	9099	16013
GO:0005515	protein binding	molecular_function	2.98E-54	6.06E-51	3723	12183	4294	16013
GO:0003824	catalytic activity	molecular_function	1.54E-43	2.08E-40	4845	12183	5697	16013
GO:0008152	metabolic process	biological_process	1.09E-39	1.11E-36	5705	12183	6856	16013
GO:0043167	ion binding	molecular_function	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 (cellular_component, biological_process, or 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 genes related to this term

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

Figure 3.14.1.1 shows enriched target gene candidates.

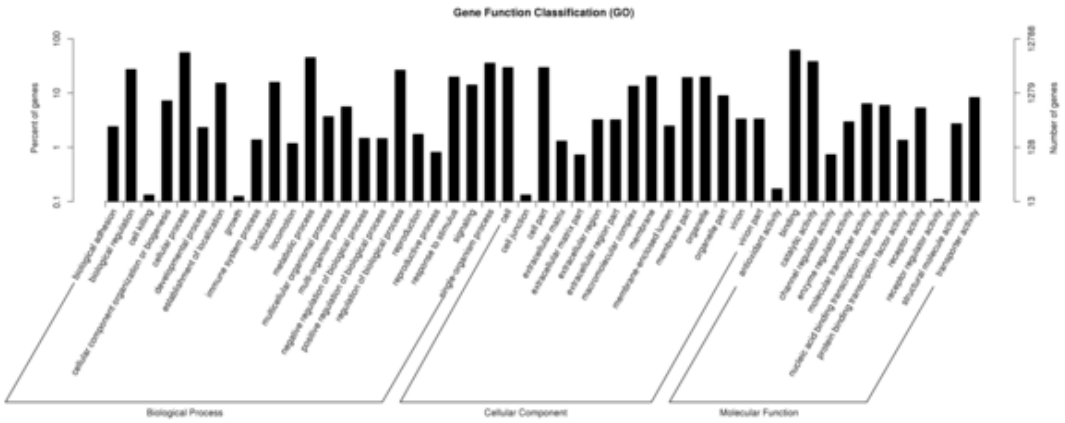


Figure 3.14.1.1 The histogram of target gene candidates

Figure 3.14.1.2 shows DAGs of GO enrichment. DAGs for biological process, molecular function and cellular component are shown respectively.

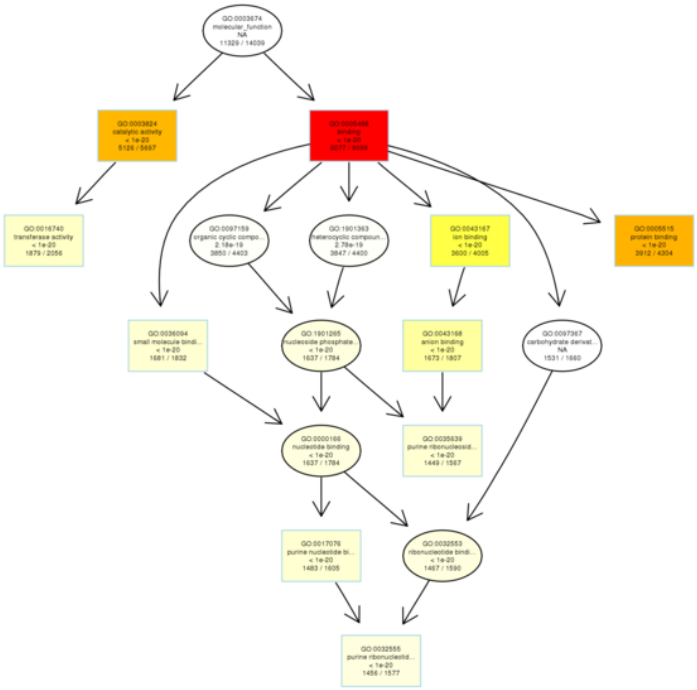


Figure 3.14.1.2 DAGs of GO enrichment

Each node represents a GO term, and the TOP10 GO terms are boxed. The name and p-value of each term are shown on the node.

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 graphs. Associated terms are also represented. Color depth indicates enrichment level.

3.14.2 KEGG Pathway Analysis

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of manually curated databases related to 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 related studies.

Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed miRNA target genes related to the whole genome background:

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

N = the number of all genes with KEGG annotation
n = the number of target gene candidates in N
M = the number of all genes annotated to a certain pathway
m = the number of target gene candidates in M

Genes with BH smaller than 0.05 are considered significantly enriched in target gene candidates. KEGG analysis could reveal the main pathways with 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

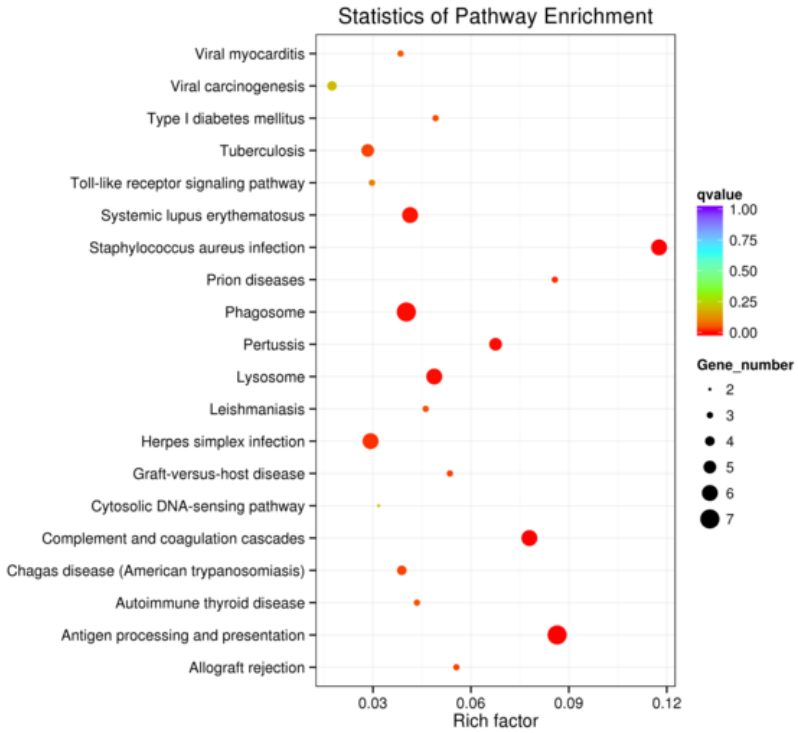


Figure 3.14.2.1 KEGG enrichment scatter plot of DEGs

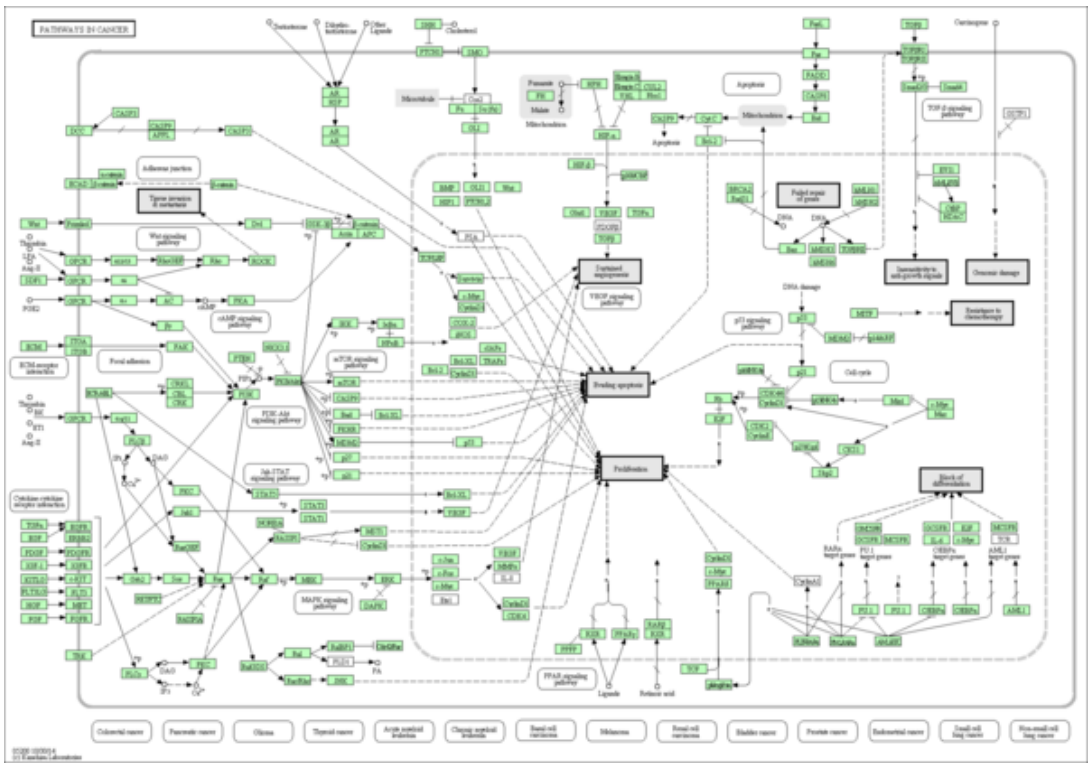


Figure 3.14.2.2 Metabolic map of target genes

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

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