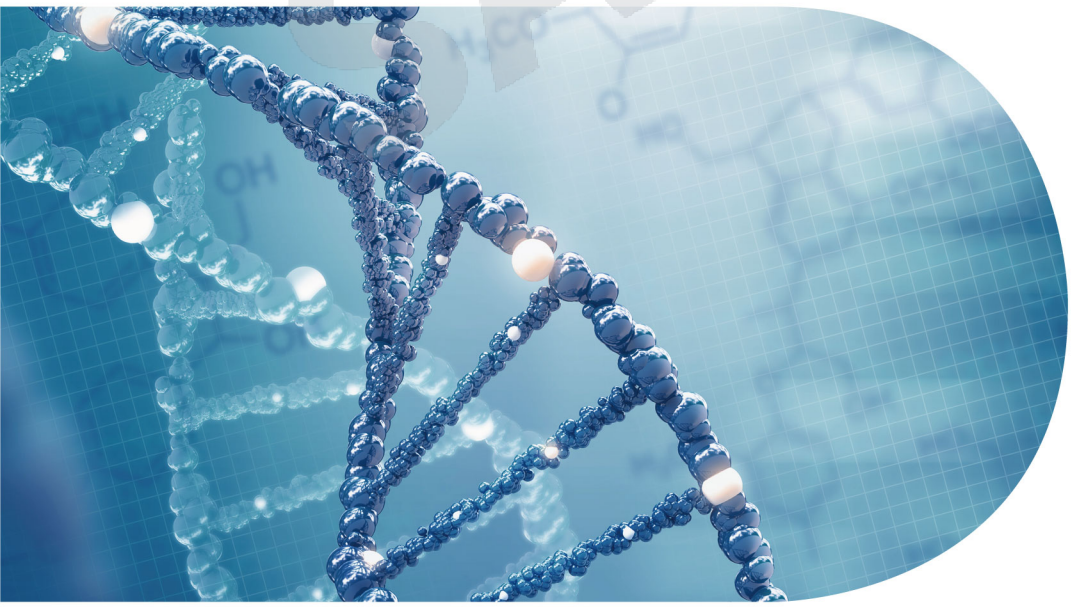
Research use only

Homo sapiens Small RNA Sequencing

Report

April 2020



Humanizing Genomics macrogen

ver. MGSR3.0_GRCh38

Project Information

Client Name	TESTER
Company/Institution	Macrogen
Order Number	HN00000000
Species	*Homo sapiens*
Reference	hg19
miRBase / RNAcentral version	miRBase v22.1 / RNAcentral 14.0
Read Length	51
Number of Samples	6
Library Kit	SMARTer smRNA-Seq Kit
Library Protocol	SMARTer smRNA-Seq Kit for Illumina User Manual
Reagent	TruSeq rapid SBS kit or Truseq SBS Kit v4
Sequencing Protocol	HiSeq 2500 System User Guide Document # 15035786 v02 HCS 2.2.70
Type of Sequencer	HiSeq 2500
Sequencing Control Software	HCS 2.2.70

Table of Contents

Project Information 02

1. Experimental Methods and Workflow 04

2. Analysis Methods and Workflow 05

3. Summary of Data Production 06
- 3. 1. Raw Data Statistics 06
- 3. 2. Average Base Quality at Each Cycle 07
- 3. 3. Trimming Data Statistics 08
- 3. 4. Average Base Quality at Each Cycle after Trimming 11
- 3. 5. Final processed reads after removing unwanted reads 12
- 3. 6. Read length distribution of each sample 13

4. smRNA Analysis Result 14
- 4. 1. Summary of Small RNA Composition 14
- 4. 2. Quantification of Mature miRNA Abundance 15
- 4. 3. Prediction of known/novel miRNA 20
- 4. 4. Small RNA Composition and Profiling 30

5. Differentially Expressed miRNA Analysis Results 33
- 5. 1. Data Analysis Quality Check and Preprocessing 33
- 5. 2. Differentially Expressed miRNA Analysis Workflow 38
- 5. 3. Significant Mature miRNA Results 39

6. Data Download Information 44
- 6. 1. Raw Data 44
- 6. 2. Analysis Results 44

7. Appendix 47
- 7. 1. Phred Quality Score Chart 47
- 7. 2. Programs used in Analysis 48
- 7. 3. References 50

1. Experimental Methods and Workflow

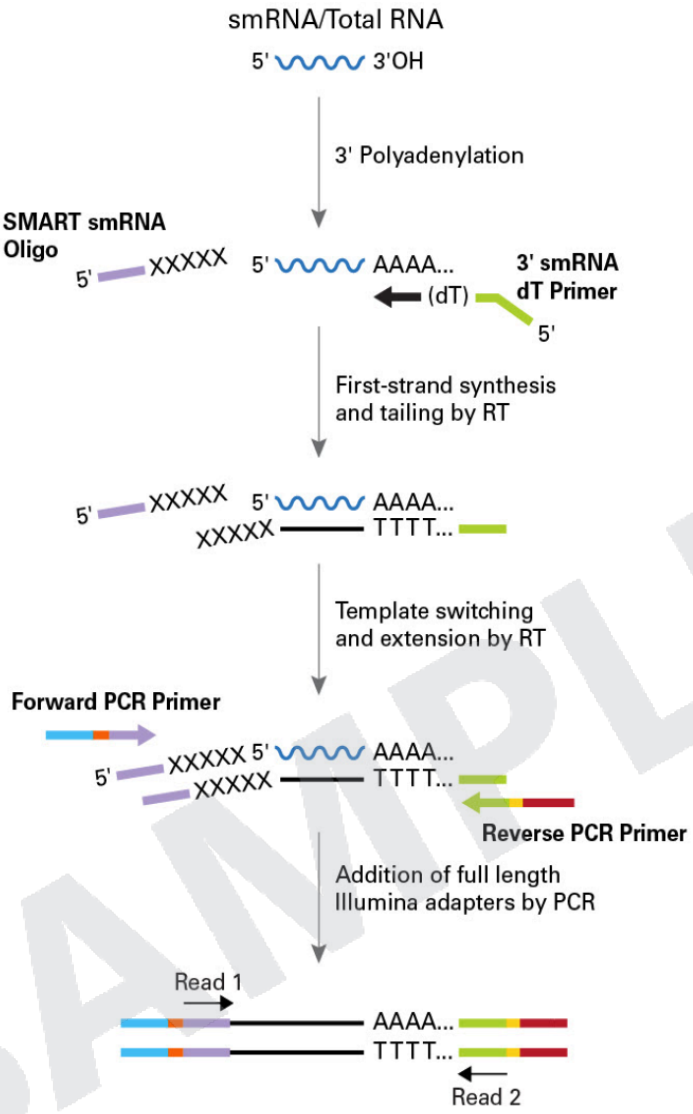


Figure 1. RNA Sequencing Experiment Workflow

REFERENCE SMARTer smRNA-Seq Kit for Illumina User Manual_040816.pdf

1) Input RNA is first polyadenylated in order to provide a priming sequence for an oligo(dT) primer.
2) cDNA synthesis is primed by the 3' smRNA dT Primer, which incorporates an adapter sequence (green) at the 5' end of each first-strand cDNA molecule.
3) When the MMLV-derived PrimeScript Reverse Transcriptase (RT) reaches the 5' end of each RNA template, it adds non-templated nucleotides which are bound by the SMART smRNA Oligo-enhanced with locked nucleic acid (LNA) technology for greater sensitivity.
4) PrimeScript RT uses the SMART smRNA Oligo as a template for the addition of a second adapter sequence (purple) to the 3' end of each first-strand cDNA molecule.
5) Full-length Illumina adapters are added during PCR amplification.
6) cDNA fragments are sequenced by the read length using sequence by synthesis method on the illumina platform.

2. Analysis Methods and Workflow

Figure 2. Analysis Workflow

The objectives of smRNA analysis are divided into four following contents.

- Quantification of smRNA expression
- Filtering differentially expressed smRNA
- Prediction of known / novel miRNAs
- Classification by categories of smRNA

1) After sequencing, the raw sequence reads are filtered based on quality. The adapter sequences are also trimmed off the raw sequence reads.
2) Both the trimmed reads and non-adapter reads as processed reads are used, to do analyzing long target (>= 50bp).
3) The processed reads are gathered forming a unique cluster. This cluster contains reads that are 100% match to the sequence identity as well as read length. The cluster is given its temporary cluster ID and the number of reads it holds.
4) In order to eliminate the effect of large amounts of rRNA from this study, the read was aligned to the rRNA sequence and have them removed.
5) rRNA removed reads are sequentially aligned to reference genome, miRBase v22.1 and non-coding RNA database, RNAcentral 14.0 to classify known miRNAs and other type of RNA such as tRNA, snRNA, snoRNA etc. Novel miRNA prediction is performed by miRDeep2.
6) The read counts for each smRNA are extracted from mapped smRNAs to report the abundance of each smRNA.
7) (Optional) Differentially expressed smRNAs are determined through comparing across conditions each smRNA using statistical methods.

3. Summary of Data Production

3. 1. Raw Data Statistics

(Refer to Path: result_smRNA/Analysis_statistics/FASTQ_stats.txt)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 6 samples. For example, in MG_CTRL_1, 48,516,689 reads are produced, and total read bases are 2.5Gbp. The GC content (%) is 42.06% and Q30 is 84.83%.

Table 1. Raw data stats

Sample ID	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
MG_CTRL_1	2,474,351,139	48,516,689	42.06	90.86	84.83
MG_CTRL_2	2,329,455,090	45,675,590	41.99	90.49	84.29
MG_CTRL_3	2,611,439,241	51,204,691	41.80	90.89	84.87
MG_TEST_1	2,559,360,795	50,183,545	41.58	92.16	86.69
MG_TEST_2	3,130,081,242	61,374,142	40.72	90.88	84.74
MG_TEST_3	3,159,045,009	61,942,059	40.43	90.84	84.83

(* Total read bases = Total reads x Read length)

- Total read bases: Total number of bases sequenced
- Total reads: Total number of reads
- GC (%): GC content
- Q20 (%): Ratio of bases that have phred quality score greater than or equal to 20
- Q30 (%): Ratio of bases that have phred quality score greater than or equal to 30

3. 2. Average Base Quality at Each Cycle

(Refer to Path: result_smRNA/Analysis_statistics/rawData_FASTQC/A_fastqc/)

The quality of produced data is determined by the phred quality score at each cycle. Box plot containing the average quality at each cycle is created with FastQC.

The x-axis shows number of cycles and y-axis shows phred quality score. Phred quality score 20 means 99% accuracy and reads over score of 20 are accepted as good quality.

Figure 3 shows average base quality at each cycle.

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc

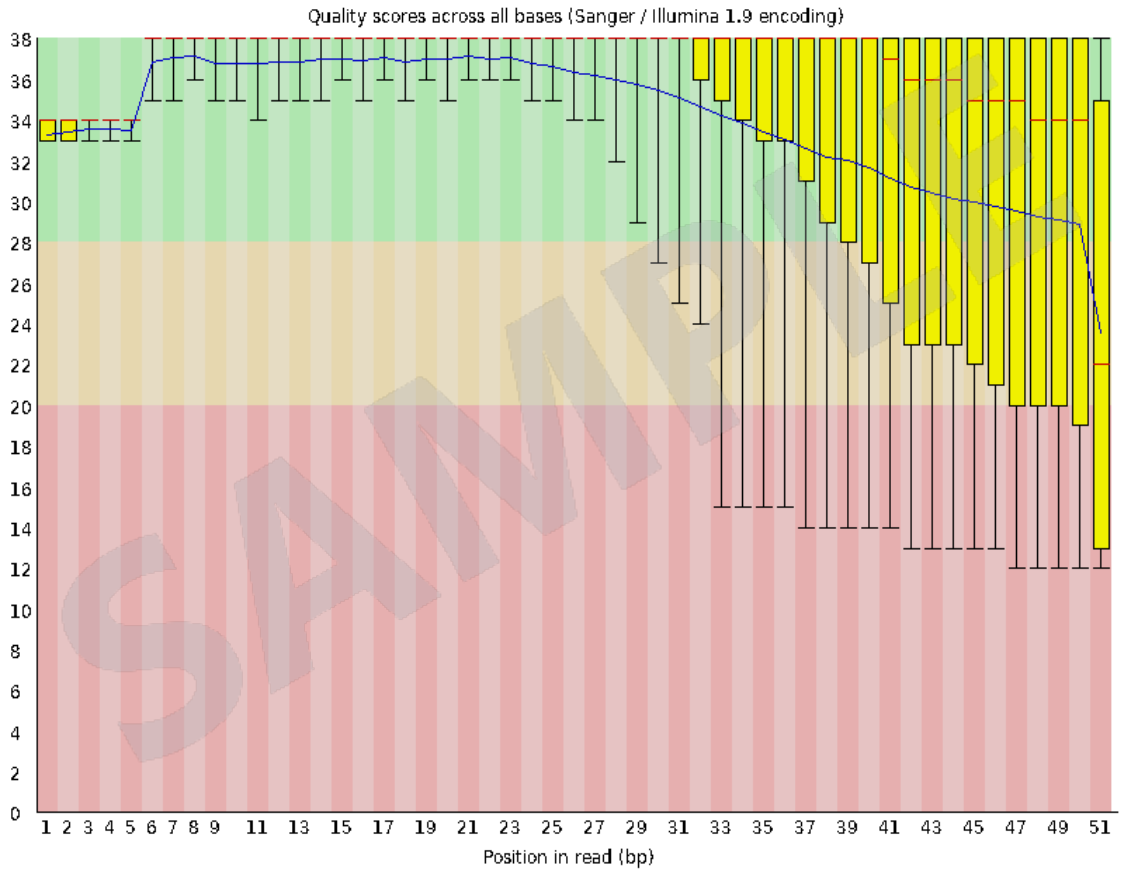


Figure 3. Read quality at each cycle of MG_CTRL_1 (read1)

- Yellow box: Interquartile range (25-75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

3. 3. Trimming Data Statistics

(Refer to Path: result_smRNA/Analysis_statistics/FASTQ_stats.txt)

The reads start at the first base after the 5' sequencing adapter and in Illumina sequencing typically end after 51 bp (in case of sequencing 51 bp SE). As mature miRNAs are normally up to 25 bp in length, the reads will contain part of 3' adapter sequence that has to be removed.

Sequenced reads are classified as trimmed reads, non-adapter reads, short reads and low quality reads according to these criteria. Adapter trimming process is done to eliminate the adapter sequences that exist in the read using Cutadapt. If a read matches at least first 5 bp of 3' adapter sequence, it is regarded as an adapter sequence, and then trimmed from the read. Trimmed reads should be at the minimum of 18 bp in order to be considered reliable for analysis. Afterwards, 3' end of each read is quality trimmed with -q 10 in Cutadapt. In addition, trimmed or nonAdapter read with 'N' base is regarded as low quality read and filtered.

Here is the definition of four read types described above.

- Trimmed Read: Read that is removed adapter sequences
- nonAdapter Read: Read that has not adapter sequences
- Short Read: Read with below 17bp in read length after adapter trimming
- Low Quality Read: Read with one or more 'N' base in trimmed or nonAdapter read

In case of analyzing short target (<50bp), the trimmed reads as processed reads are used. Otherwise, in case of analyzing long target (>= 50bp), both the trimmed reads and non-adapter reads as processed reads are used.

In this analysis, we analyzed both short target and long target to see comprehensive smRNA profiles.

For example, a total of 48,516,689 reads were produced from the MG_CTRL_1 sample, of which 21,722,746 reads included the adapter and the read length after trim was 18bp or more. 4,086,503 reads did not include the adapter, and 22,575,211 reads were shorter than 18bp after adapter trim.

For the analysis of long targets, 25,809,249 readings are used in the subsequent analysis, including trimmed reads and nonAdapter reads.

Table 2 - 4 show statistics of trimmed, nonAdapter and short FASTQ.

Table 2. Trimming Data Stats

Sample ID	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
MG_CTRL_1	529,787,140	21,722,746	53.09	98.29	96.08
MG_CTRL_2	489,528,561	20,510,349	53.09	98.37	96.26
MG_CTRL_3	510,908,370	21,376,564	52.48	98.39	96.32
MG_TEST_1	582,154,530	22,628,662	51.51	98.11	95.68
MG_TEST_2	576,774,529	23,878,445	51.47	98.41	96.32
MG_TEST_3	553,246,105	23,110,454	51.46	98.35	96.23

Table 3. nonAdapter Data Stats

Sample ID	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
MG_CTRL_1	196,151,790	4,086,503	48.00	96.58	93.33
MG_CTRL_2	146,265,569	3,047,205	47.94	96.23	93.00
MG_CTRL_3	228,197,922	4,754,132	48.04	96.82	93.72
MG_TEST_1	385,869,857	8,038,969	47.31	97.37	94.18
MG_TEST_2	248,006,488	5,166,811	48.88	96.80	93.67
MG_TEST_3	249,602,670	5,200,065	49.36	96.72	93.60

Table 4. Short Data Stats

Sample ID	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
MG_CTRL_1	269,587,536	22,575,211	66.02	98.16	96.01
MG_CTRL_2	264,314,948	21,997,473	66.40	98.18	96.03
MG_CTRL_3	297,540,014	24,939,233	66.54	98.20	96.09
MG_TEST_1	215,178,281	19,358,420	67.81	98.13	95.93
MG_TEST_2	356,650,740	32,177,953	68.31	98.28	96.22
MG_TEST_3	371,273,070	33,484,464	69.53	98.21	96.15

- Total read bases : Total number of read bases after trimming
- Total reads : Total number of reads after trimming
- GC(%) : GC Content
- Q20(%) : Ratio of bases that have phred quality score greater than or equal to 20
- Q30(%) : Ratio of bases that have phred quality score greater than or equal to 30

Table 5 and Figure 4 represents counts for trimmed reads, non-adapter reads, short reads and low quality reads by each sample.

Table 5. Summary of read preprocessing

Sample	Total Read Count	Trimmed Read Count	nonAdapter Read Count	Short Read Count	Low Quality Read Count
MG_CTRL_1	48,516,689	21,722,746 (44.77%)	4,086,503 (8.42%)	22,575,211 (46.53%)	132,229 (0.27%)
MG_CTRL_2	45,675,590	20,510,349 (44.9%)	3,047,205 (6.67%)	21,997,473 (48.16%)	120,563 (0.26%)
MG_CTRL_3	51,204,691	21,376,564 (41.75%)	4,754,132 (9.28%)	24,939,233 (48.7%)	134,762 (0.26%)
MG_TEST_1	50,183,545	22,628,662 (45.09%)	8,038,969 (16.02%)	19,358,420 (38.58%)	157,494 (0.31%)
MG_TEST_2	61,374,142	23,878,445 (38.91%)	5,166,811 (8.42%)	32,177,953 (52.43%)	150,933 (0.25%)
MG_TEST_3	61,942,059	23,110,454 (37.31%)	5,200,065 (8.4%)	33,484,464 (54.06%)	147,076 (0.24%)

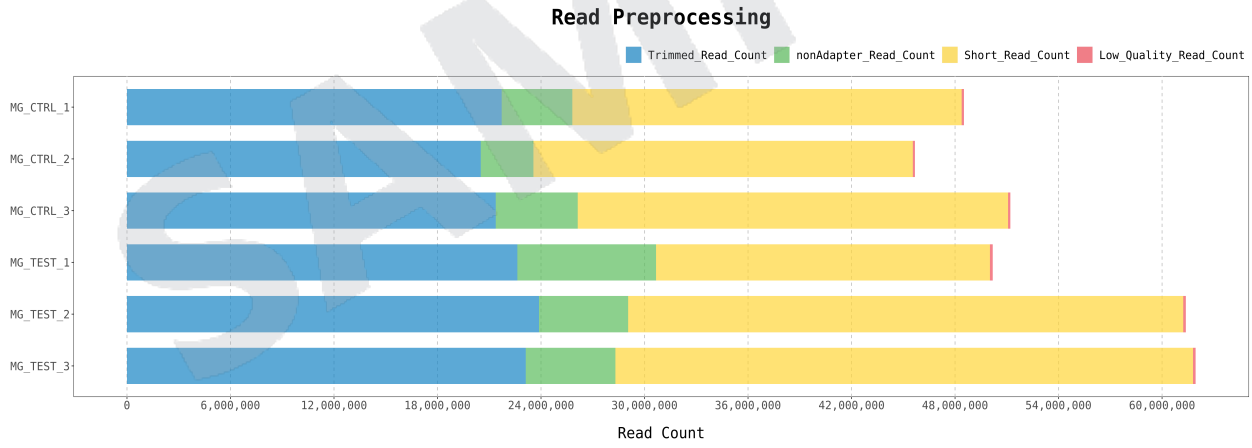


Figure 4. Summary of read preprocessing

3. 4. Average Base Quality at Each Cycle after Trimming

(Refer to Path: result_smRNA/Analysis_statistics/processedData_FASTQC/A_fastqc/)

Figure 5 shows average base quality at each cycle after trimming.

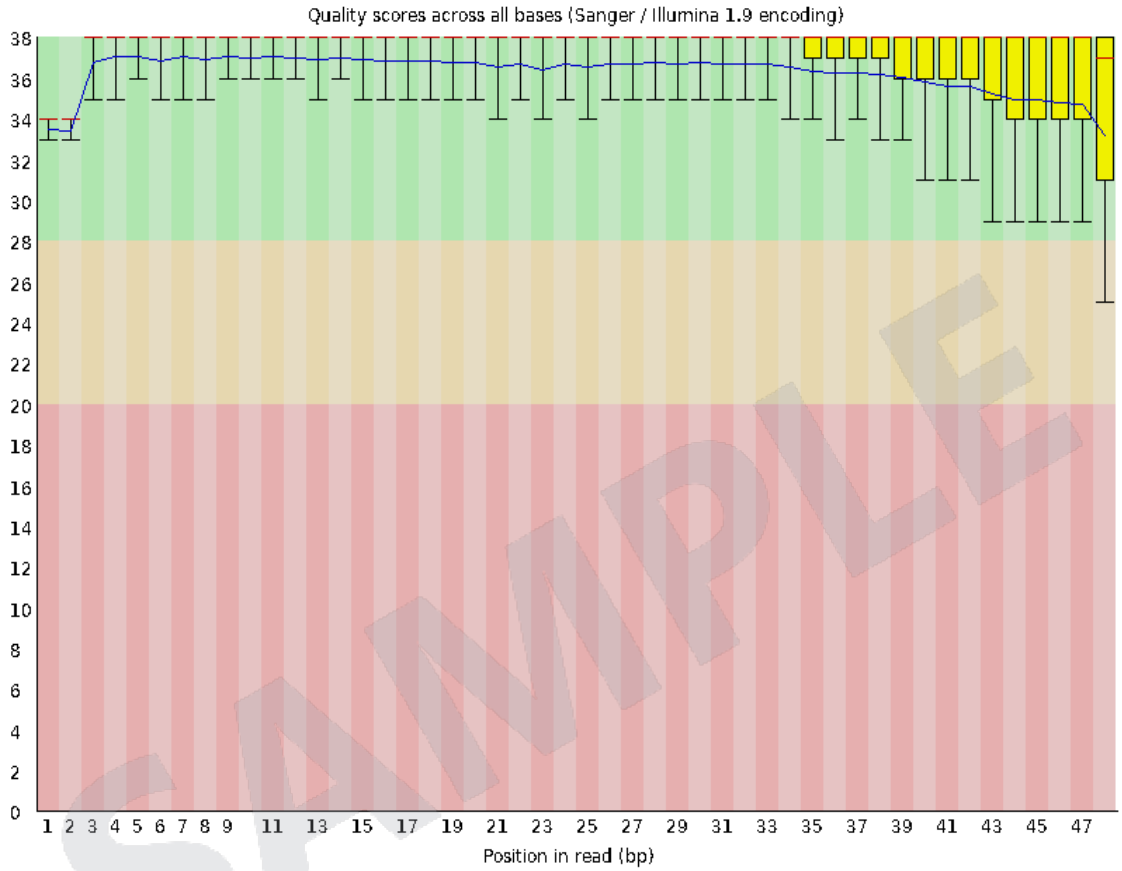


Figure 5. Average base quality of MG_CTRL_1 (read1) at each cycle after trimming

- Yellow box: Interquartile range (25-75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

3. 5. Final processed reads after removing unwanted reads

Most of the RNA composition is usually known as rRNA. In order to eliminate the effect of large amounts of rRNA from this study, reads were aligned to the 45S pre-rRNA and mitochondrial rRNA of *Homo sapiens* and have them removed.

Figure 6 represents the remaining reads (blue) after removing rRNA (red).

For example a total of 25,809,249 reads left after adapter trimming step from the MG_CTRL_1 sample, of which 8,195,759 reads did aligned to rRNA sequences. 17,613,490 reads finally remained in this filtering step. These reads were used in the following analysis step.

Table 6. Remaining reads after removing rRNA

Sample	Total Read Count	Remain Read Count	Filtered Read Count
MG_CTRL_1	25,809,249	17,613,490 (68.24%)	8,195,759 (31.76%)
MG_CTRL_2	23,557,554	16,588,880 (70.42%)	6,968,674 (29.58%)
MG_CTRL_3	26,130,696	19,344,740 (74.03%)	6,785,956 (25.97%)
MG_TEST_1	30,667,631	23,443,799 (76.44%)	7,223,832 (23.56%)
MG_TEST_2	29,045,256	24,430,051 (84.11%)	4,615,205 (15.89%)
MG_TEST_3	28,310,519	23,941,475 (84.57%)	4,369,044 (15.43%)

- Sample: Sample ID
- Total Read Count: Total read count of each sample
- Remain Read Count: Remaining reads count after removing rRNA
- Filtered Read Count: Filter out read count after removing rRNA

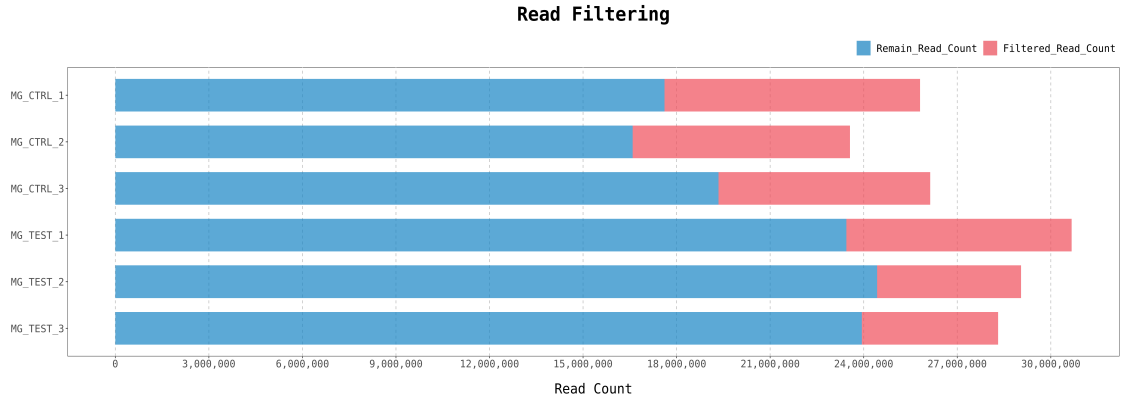


Figure 6. Summary of reads after removing rRNA

3. 6. Read length distribution of each sample

Figure 7 shows the read length distribution of each sample. Generally tRNA is 70-90 nt, snoRNA is about 90 nt, snRNA is between 100 and 300nt, miRNA is about 22 nt and piRNA is about 27 nt in length.

Figure 7. Length distribution of each sample

4. smRNA Analysis Result

4. 1. Summary of Small RNA Composition

For each sample, final processed reads are sequentially aligned to reference genome, miRBase v22.1 and non-coding RNA database (RNAcentral release 14.0) to classify known miRNAs and other types of RNA such as piRNA, tRNA, snRNA, snoRNA etc.

Genome mapping is processed by Bowtie and STAR using RSEM. Bowtie was subsequently used for miRDeep2 analysis using a genomic sequence. If it\'s necessary, the RSEM program can be used to align readings to transcript sequences using STAR.

Known/novel miRNA predicted by miRDeep2 and other smRNAs matching RNAcentral were aligned using Bowtie (target smRNA, <50 nt) and Bowtie2 (target smRNA, >=50nt).

Figure 8 represents the smRNA composition of each sample, which means the ratio of smRNA class type (such as known miRNA, candidate miRNA, rRNA, tRNA, snRNA, snoRNA etc.) classified from processed reads.

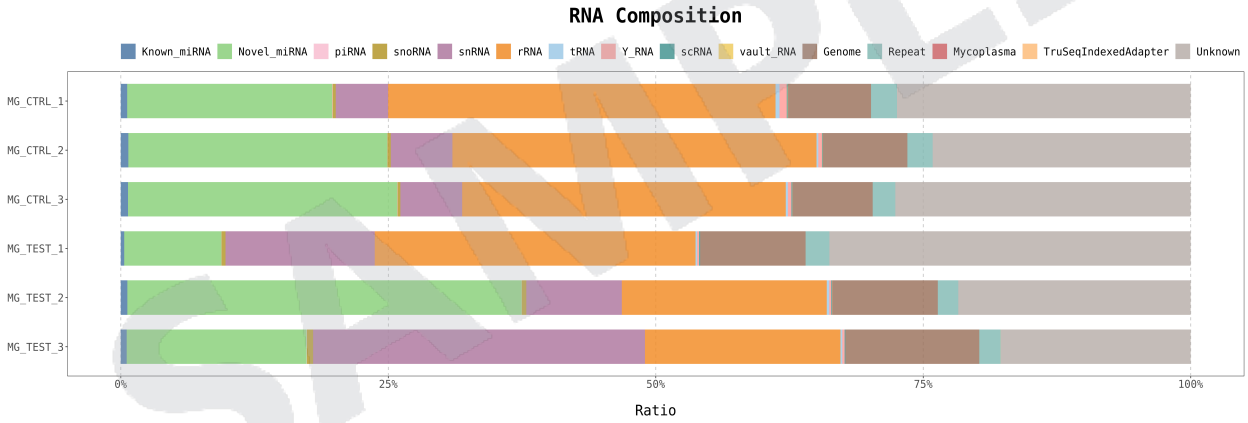


Figure 8. Summary of Small RNA Composition

4. 2. Quantification of Mature miRNA Abundance

Quantification step is divided into three procedures. At First, mature miRNA sequences of relevant species are aligned to precursor miRNA sequences from miRBase v21 (bowtie option : -f -v 0 -a --best --strata --norc). In the second step, unique clustered reads are aligned to precursor sequence (bowtie option : -f -v 1 -a --best --strata --norc). Bowtie aligner is used in the mapping steps. In the third step, we determine overlapping mature miRNA regions between two mapping results. Afterwards, the read count of mature miRNAs are extracted from the overlapping regions in the miRDeep2 Quantifier module. If a read is mapped to multiple mature miRNAs, the read counts are assigned equally to the abundance level of each mature miRNA. Table 7 shows the number of mapped reads to miRbase percursor, and the number of known mature miRNA with read count more than 1 for each sample.

Table 7. Mapped reads to miRBase precursor

Sample	Processed reads	Mapped reads	Known miRNA in Sample	Known miRNA in Species (miRBase v22.1)
MG_CTRL_1	17,613,490	155,910 (0.89%)	380	2,656
MG_CTRL_2	16,588,880	165,978 (1.0%)	349	2,656
MG_CTRL_3	19,344,740	176,423 (0.91%)	376	2,656
MG_TEST_1	23,443,799	96,734 (0.41%)	360	2,656
MG_TEST_2	24,430,051	177,994 (0.73%)	429	2,656
MG_TEST_3	23,941,475	155,117 (0.65%)	407	2,656

- Processed reads: Reads which were trimmed and removed unwanted sources from them
- Known miRNA in Sample: The number of known mature miRNA with read count more than 1 for each sample
- Known miRNA in Species: The number of known mature miRNA in miRBase database

Table 8 is an example of the expression profile of miRBase precursor for each sample in read count.

Table 9 shows the expression profile of known mature miRNA across samples. The expression profile of known mature miRNA is used to analyze differentially expressed miRNA (DE miRNA) later.

Table 8. Expression profile according to miRBase precursor (Example)

miRBase_Precursor_ID	hsa-mir-1246	hsa-mir-23a	hsa-mir-25	hsa-mir-191
Total_Read_Count	15191	2365	2176	2038
5p_miRNA_ID	hsa-miR-1246	hsa-miR-23a-5p	hsa-miR-25-5p	hsa-miR-191-5p
5p_Read_Counts	15191	1209	52	2036
3p_miRNA_ID		hsa-miR-23a-3p	hsa-miR-25-3p	hsa-miR-191-3p
3p_Read_Counts		1156	2121	2
Remaining_Reads	0	0	3	0
miRBase_5p_Sequence(s)	aauggauuuuuggagcagg	ugagguaguagguuguauaguu	gggguuccugggggaugggauuu	aggcggagacuugggcaauug
miRBase_3p_Sequence(s)		cuguacagccuccuagcuuucc	aucacauugccagggauuucc	cauugcacuugucucggucuga
miRBase_Precursor_Sequence	uguauccuugaauggauuuuuggagcaggaguggacaccugacccaaaggaaaucaauccauaggcuagcaau	agguugagguaguagguuguauaguuagaauuacaucaaggagauaacuguacagccuccuagcuuuccu	ggccggcugggguuccuggggaugggauuugcuuccugucacaaaucacauugccaggguuuccaaccgacc	ggccaguguugagaggcggagacuugggcaaauugcuggacgcugcccugggcauugcacuugucucggucugacagugccggcc

- miRBase Precursor ID: Clicking this field will display a pdf of the structure and read signature of the miRNA.
- Total Read Count: The sum of read counts for the 5p and 3p miRNA
- 5p miRNA ID
- 5p Read Counts: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the 5p sequence, including 2 nts upstream and 5 nts downstream.
- 3p miRNA ID
- 3p Read Counts: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the 3p sequence, including 2 nts upstream and 5 nts downstream.
- Remaining Reads: The number of reads that did not map to any of the 5p or 3p miRNA
- miRBase 5p sequence(s): The 5p miRNA sequence(s)
- miRBase 3p sequence(s): The 3p miRNA sequence(s)
- miRBase Precursor sequence: The precursor miRNA sequence

[PDF description based on Mature miRNA]

[Red Box]: Information box

- miRBase precursor: Precursor sequence ID of miRBase
- Total read count: Total read count of precursor
- 5p read counts: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the 5p sequence, including 2 nts upstream and 5 nts downstream
- 3p read counts: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the 3p sequence, including 2 nts upstream and 5 nts downstream
- remaining reads: The number of reads that did not map to any of the 5p or 3p sequences

[Blue Box]: Secondary structure box; (Figure)

- Red character: Mature miRNA region

[Orange Box]: Read frequency distribution

- Red line: Mature miRNA region

[Green Box]: Secondary structure box; (Text)

- First line: Genomic sequence (Red character: mature miRNA region)
- Second line: Secondary structure ("..."; Wobble, have not bond, "(", ")"; have bond)

[Purple Box]: Read Box

- Column 1: Read sequence (Upper case: mismatch sequence)
- Column 2: Read Count
- Column 3: Mismatch count

Table 9. Expression profile according to known mature miRNA (Example)

Mature_ID	hsa-let-7a-2-3p	hsa-let-7a-3p	hsa-let-7a-5p	hsa-let-7b-3p	hsa-let-7c-5p	hsa-let-7d-5p
Mature_Accession	MIMAT0010195	MIMAT0004481	MIMAT0000062	MIMAT0004482	MIMAT0000064	MIMAT0000065
Hairpin_ID	hsa-let-7a-2	hsa-let-7a-1, hsa-let-7a-3	hsa-let-7a-1, hsa-let-7a-2, hsa-let-7a-3	hsa-let-7b	hsa-let-7c	hsa-let-7d
Hairpin_Accession	MI0000061	MI0000060, MI0000062	MI0000060, MI0000061, MI0000062	MI0000063	MI0000064	MI0000065
miFAM_ID	let-7	let-7	let-7	let-7	let-7	let-7
miFAM_Accession	MIPF0000002	MIPF0000002	MIPF0000002	MIPF0000002	MIPF0000002	MIPF0000002
miRBase_Link	hsa-let-7a-2-3p	hsa-let-7a-3p	hsa-let-7a-5p	hsa-let-7b-3p	hsa-let-7c-5p	hsa-let-7d-5p
ENTREZGENE_Link	406882	406881	406881	406884	406885	406886
RFAM_Link	RF00027	RF00027	RF00027	RF00027	RF00027	RF00027
HGNC_Link	31477	31476	31476	31479	31480	31481
AM_RPM	0	35.197015	51669.21845	123.189554	11738.2046	6300.265737
BM_RPM	0	330.654552	67309.76581	244.396843	26754.26616	4600.411162
AM_Read_Count	0	2	2936	7	667	358
BM_Read_Count	0	23	4682	17	1861	320

- Mature ID: Mature miRNA ID
- Mature Accession: Mature miRNA accession ID of miRBase
- Hairpin ID: Hairpin (Precursor) ID
- Hairpin Accession: Hairpin accession ID of miRBase
- miFAM ID: miRNA gene family ID
- miRBase Link: miRBase ID (click to miRBase)
- ENTREZGENE Link: ENTREZGENE accession (click to NCBI)
- HGNC Link: HGNC accession (click to HUGO Gene Nomenclature Committee) if species is Homo sapiens, this column displays.
- MGI Link: MGI accession (click to Mouse Genome Information) if species is Mus musculus, this column displays.
- RFAM Link: RFAM accession (click to RNA families database)
- [Sample Name] RPM : Read per million of sample called "Sample Name"
- [Sample Name] Read Count: Read count of the sample called "Sample Name"

4. 3. Prediction of known/novel miRNA

To predict the known and novel miRNA, unique clustered reads are aligned against reference genome and precursor miRNAs separately. Novel microRNAs are predicted from mature, star and loop sequence according to the RNAfold algorithm using miRDeep2. The RNAfold function uses the nearest-neighbor thermodynamic model to predict the minimum free-energy secondary structure of an RNA sequence. RNAfold-generated graphic contains the actual in silico-folded hairpin, with the number of reads for each part of the hairpin, score for minimum free energy, score for randfold, and score for conserved seed sequence. In addition to detecting known and novel miRNAs, miRDeep2 estimates their abundance. Table 10 shows the mapping statistics to GRCh38 reference genome with Bowtie.

Table 10. Mapped reads to reference genome

Sample	Total reads	Mapped reads	Unmapped reads
MG_CTRL_1	17,613,490	2,407,310 (13.67%)	15,206,180 (86.33%)
MG_CTRL_2	16,588,880	2,266,399 (13.66%)	14,322,481 (86.34%)
MG_CTRL_3	19,344,740	2,467,926 (12.76%)	16,876,814 (87.24%)
MG_TEST_1	23,443,799	3,049,949 (13.01%)	20,393,850 (86.99%)
MG_TEST_2	24,430,051	3,478,218 (14.24%)	20,951,833 (85.76%)
MG_TEST_3	23,941,475	3,977,532 (16.61%)	19,963,943 (83.39%)

Table 11 - 14 and Genome-based PDF refer to result_smRNA_excel/miRDeep2_Result/[Sample_name]/[Sample_Name]_GenomeBased_result.xlsx

Table 11 shows an example of the summary result of predicted known/novel miRNAs by miRDeep2 at score cut-offs from -10 to 10.

Table 11. Survey of miRDeep2 performance at score cut-offs from -10 to 10

miRDeep2 score	10	9	8
novel miRNAs reported by miRDeep2	10	10	10
novel miRNAs, estimated false positives	8 +/- 3	8 +/- 3	8 +/- 3
novel miRNAs, estimated true positives	2 +/- 2 (24 +/- 21%)	2 +/- 2 (23 +/- 21%)	2 +/- 2 (21 +/- 20%)
known miRNAs in species	2588	2588	2588
known miRNAs in data	230	230	230
known miRNAs detected by miRDeep2	32 (14%)	34 (15%)	35 (15%)
estimated signal-to-noise	3.3	3.3	3.3
excision gearing	2	2	2

- miRDeep2 score: For details on how the log-odds score is calculated, see Friedlander et al., Nature Biotechnology, 2008.
- predicted by miRDeep2: Novel miRNA hairpins are defined by not having any of the reference mature miRNAs mapping perfectly (full length, no mismatches). The numbers show how many novel miRNA hairpins have a score equal to or exceeding the cut-off.
- estimated false positives: Number of false positive miRNA hairpins predicted at this cut-off, as estimated by the miRDeep2 controls (see Friedlander et al., Nature Biotechnology, 2008). Mean and standard deviation are estimated from 100 rounds of permuted controls.
- estimated true positives: The number of true positive miRNA hairpins are estimated as t = total novel miRNAs - false positive novel miRNAs. The percentage of the predicted novel miRNAs that is estimated to be true positives is calculated as p = t/total novel miRNAs. The number of false positives are estimated from 100 rounds of permuted controls. In each of the 100 rounds, t and p are calculated, generating mean and standard deviation of t and p. The variable p can be used as an estimation of miRDeep2 positive predictive value at the score cut-off.
- in species: Number of reference mature miRNAs for that species given as input to miRDeep2.
- in data: Number of reference mature miRNAs for that species that map perfectly (full length, no mismatches) to one or more of precursor candidates that have been excised from the genome by miRDeep2.
- detected by miRDeep2 : Number of reference mature miRNAs for that species that map perfectly (full length, no mismatches) to one or more of predicted miRNA hairpins that have a score equal to or exceeding the cut-off. The percentage of reference mature miRNAs in data that is detected by miRDeep2 is calculated as s = reference mature miRNAs detected / reference mature miRNAs in data. s can be used as an estimation of miRDeep2 sensitivity at the score cut-off.
- estimated signal-to-noise: For the given score cut-off, the signal-to-noise ratio is estimated as r = total miRNA hairpins reported/mean estimated false positive miRNA hairpins over 100 rounds of permuted controls.
- excision gearing: This is the minimum read stack height required for excising a potential miRNA precursor from the genome in this analysis.

There are three prediction results by miRDee2:

- novel miRNAs predicted by miRDeep2
- mature miRBase miRNAs detected by miRDeep2
- mature miRBase miRNAs not detected by miRDeep2

Table 12. novel miRNAs predicted by miRDeep2 (Example)

provisional id	chrX_33645	chr6_11385	chr13_22777
miRDeep2 score	258.3	258.2	156.4
estimated probability that the miRNA candidate is a true positive	24 +/- 21%	24 +/- 21%	24 +/- 21%
total read count	507	506	306
mature read count	504	504	278
loop read count	0	0	0
star read count	3	2	28
significant randfold p-value	yes	yes	yes
UCSC browser	blat	blat	blat
consensus mature sequence	ggaggcggagguugcagugagc	ggaggcggagguugcagugagc	ggaggcggagguugcagugagc
consensus star sequence	ugccauugcacuccagccugggcg	gccauugcacuccagccug	cacugcacuccagccugggc
consensus precursor sequence	ggaggcggagguugcagugagccaagauugugccauugcacuccagccugggcg	ggaggcggagguugcagugagccaagauugcgccauugcacuccagccug	ggaggcggagguugcagugagccaagauugcaccacugcacuccagccugggc
precursor coordinate	chrX:119035594..119035648:+	chr6:35484229..35484279:+	chr13:31066230..31066283:+

- provisional id: A provisional miRNA name assigned by miRDeep2. The first part of the id designates the chromosome or genome contig on which the miRNA gene is located. The second part is a running number that is added to avoid identical ids. The running number is incremented by one for each potential miRNA precursor that is excised from the genome. Clicking this field will display a pdf of the structure, read signature and score breakdown of the reported miRNA.
- miRDeep2 score: The log-odds score assigned to the hairpin.
- estimated probability that the miRNA candidate is a true positive: The estimated probability that a predicted novel miRNA with a score of this or higher is a true positive. To see exactly how this probability is estimated, mouse over the 'novel miRNAs, true positives' in the table at the top of the webpage.
- total read count: The sum of read counts for the predicted mature, loop and star miRNAs.
- mature read count: The number of reads that mapped to the predicted miRNA hairpin and are contained in the sequence covered by the predicted mature miRNA, including 2 nts upstream and 5 nts downstream.
- loop read count: The number of reads that mapped to the predicted miRNA hairpin and are contained in the sequence covered by the predicted miRNA loop, including 2 nts upstream and 5 nts downstream.
- star read count: The number of reads that mapped to the predicted miRNA hairpin and are contained in the sequence covered by the predicted star miRNA, including 2 nts upstream and 5 nts downstream.
- significant randfold p-value: If the estimated randfold p-value of the excised potential miRNA hairpin is lower or equal to than 0.05 (see Bonnet et al., Bioinformatics, 2004).
- UCSC browser: If a species name was put into miRDeep2, then clicking this field will initiate

a UCSC blat search of the consensus precursor sequence against the reference genome.

- consensus mature sequence: The consensus mature miRNA sequence as inferred from the deep sequencing reads.
- consensus star sequence: The consensus star miRNA sequence as inferred from the deep sequencing reads.
- consensus precursor sequence: The consensus precursor miRNA sequence as inferred from the deep sequencing reads. Note that this is the inferred Drosha hairpin product, and therefore does not include substantial flanking genomic sequence as does most miRBase precursors.
- precursor coordinate: The given precursor coordinates refer to absolute position in the mapped reference sequence.

Table 13. mature miRBase miRNAs detected by miRDeep2 (Example)

tag id	chr17_27374	chr7_14462	chr22_32774
miRDeep2 score	1580.6	174.5	122.3
estimated probability that the miRNA is a true positive	24 +/- 21%	24 +/- 21%	24 +/- 21%
predicted mature seq. in accordance with miRBase mature seq.	TRUE	STAR	TRUE
total read count	3097	347	236
mature read count	3080	346	233
loop read count	0	0	0
star read count	17	1	3
significant randfold p-value	yes	no	yes
mature miRBase miRNA	hsa-miR-423-5p	hsa-miR-25-5p	hsa-let-7b-5p
UCSC browser	blat	blat	blat
consensus mature sequence	ugaggggcagagagcgagacuuu	cauugcacuugucucggucug	ugagguaguagguuguguggguu
consensus star sequence	agcucggucugaggccccucagu	aggcggagacuugggcaauug	cuauacaaccuacugccuuccc
consensus precursor sequence	ugaggggcagagagcgagacuuu ucuauuuuccaaaagcucggucug aggccccucagu	aggcggagacuugggcaauugcug gacgcugcccugggcauugcacuu gucucggucug	ugagguaguagguuguguggguu ucagggcagugauguugccccucg gaagauaacuauacaaccuacugcc uuccc
precursor coordinate	chr17:28444112..28444171:+	chr7:99691194..99691253:-	chr22:46509570..46509646:+

- tag id: A tag id assigned by miRDeep2. The first part of the id designates the chromosome or genome contig on which the miRNA gene is located. The second part is a running number that is added to avoid identical ids. The running number is incremented by one for each potential miRNA precursor that is excised from the genome. Clicking this field will display a pdf of the structure, read signature and score breakdown of the miRNA.
- miRDeep2 score: The log-odds score assigned to the hairpin by miRDeep2
- estimated probability that the miRNA is a true positive: The estimated probability that a predicted miRNA with a score of this or higher is a true positive.
- total read count: The sum of read counts for the mature, loop and star miRNAs.
- mature read count: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the consensus mature miRNA, including 2 nts upstream and 5 nts downstream.
- loop read count: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the consensus miRNA loop, including 2 nts upstream and 5 nts downstream.
- star read count: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the consensus star miRNA, including 2 nts upstream and 5 nts downstream.
- significant randfold p-value: If the estimated randfold p-value of the miRNA hairpin is lower or equal to than 0.05 (see Bonnet et al., Bioinformatics, 2004).
- mature miRBase miRNA: The ids of any reference mature miRNAs for the species that mapped perfectly (full length, no mismatches) to the reported miRNA hairpin. If this is the case, the reported miRNA hairpin is assigned as a known miRNA. If not, it is assigned as a novel miRNA. If more than one reference mature miRNA maps to the miRNA hairpin, then only the id of the reference miRBase miRNA that matches the predicted mature sequence is output.

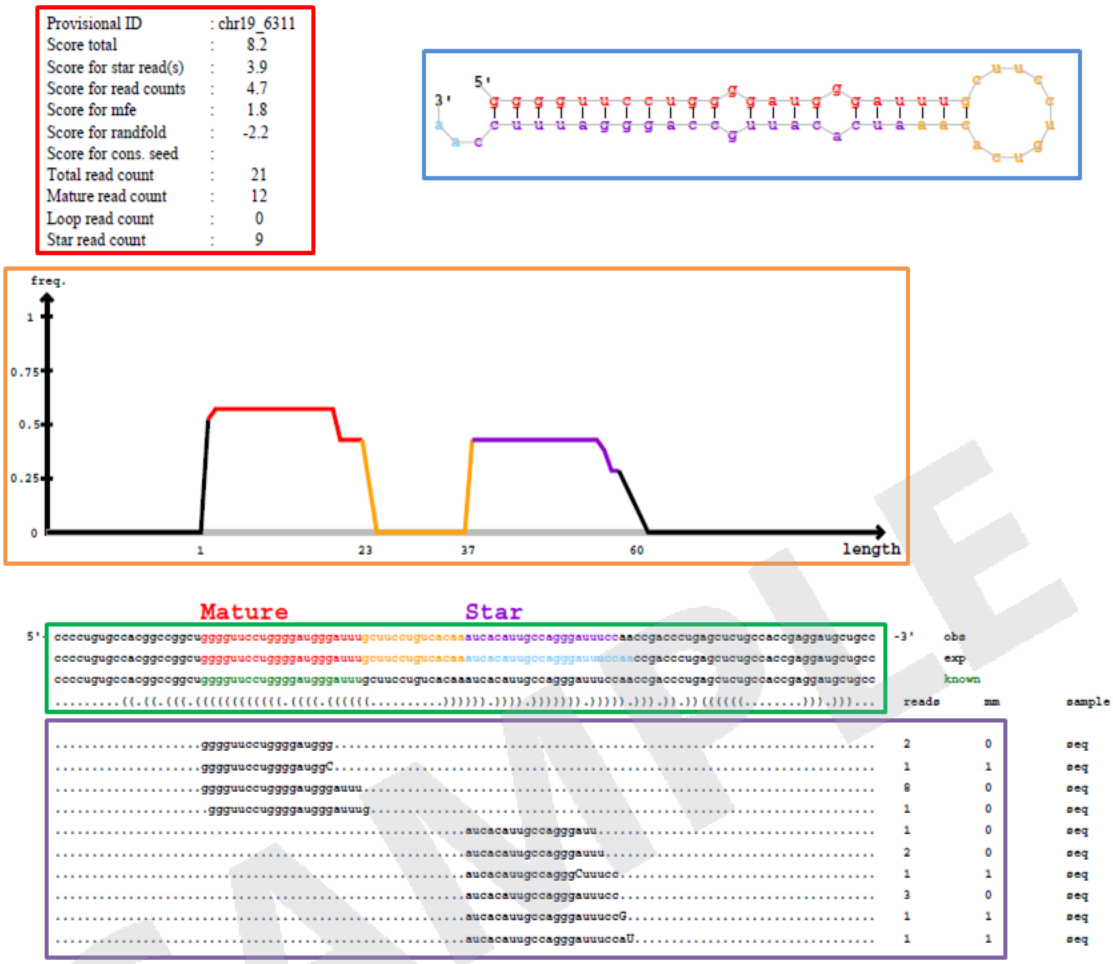
- UCSC browser: If a species name was put into miRDeep2, then clicking this field will initiate a UCSC blat search of the consensus precursor sequence against the reference genome.
- consensus mature sequence: The consensus mature miRNA sequence as inferred from the deep sequencing reads.
- consensus star sequence: The consensus star miRNA sequence as inferred from the deep sequencing reads.
- consensus precursor sequence: The consensus precursor miRNA sequence as inferred from the deep sequencing reads. Note that this is the inferred Drosha hairpin product, and therefore does not include substantial flanking genomic sequence as does most miRBase precursors.
- precursor coordinate: The given precursor coordinates refer to absolute position in the mapped reference sequence

Table 14. mature miRBase miRNAs not detected by miRDeep2 (Example)

miRBase_Precursor_ID	hsa-mir-1290	hsa-mir-6126	hsa-mir-619
Total_Read_Count	4475	2078	401
5p_Read_Counts	0	2078	222
3p_Read_Counts	4475	0	0
Remaining_Reads	0	0	179
miRBase_5p_Sequence(s)	-	gugaaggcccggcggaga	gcugggauuacaggcaugagcc
miRBase_3p_Sequence(s)	uggauuuuuggaucaggga	-	gaccuggacauguuugugcccagu
miRBase_Precursor_Sequence	gagcgucacguugacacucaaaaaguuucagauuuuggaacauuucggauuuuggauuuuuggaucagggaugcucaa	agccugugggaaagagaagagcagggcagggugaaggcccggcggagacacucugcccaccccacacccugccuaugggccacacagcu	cgcccaccucagccucccaaaaugcugggauuacaggcaugagccacugcggucgaccaugaccuggacauguuugugcccaguacugucaguuugcag

- miRBase precursor id: Clicking this field will display a pdf of the structure and read signature of the miRNA.
- total read count: The sum of read counts for the mature and star miRNAs.
- 5p read counts: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the 5p sequence, including 2 nts upstream and 5 nts downstream.
- 3p read counts: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the 3p sequence, including 2 nts upstream and 5 nts downstream.
- remaining reads: The number of reads that did not map to any of the mature and star sequences
- miRBase 5p sequence(s): The 5p miRNA sequence(s)
- miRBase 3p sequence(s): The 3p miRNA sequence(s)
- miRBase precursor sequence: This is the precursor miRNA sequence

[PDF description based on predicted precursor miRNA]



[Red Box]: Information box (score algorithm in miRDeep2_core_algorithm.pl)

- Provisional ID: A provisional miRNA name assigned by miRDeep2.
- Score total: Sum of the scores below
- Score for star read(s): Scores for the majority of potential star reads fall as expected from Dicer processing (true: 3, false: -1.3)
- Score for read counts: Scores for the count of reads that map to the potential precursor
- Score for mfe: Scores for the minimal free energy of the potential precursor
- Score for randfold: Scores for the Randfold folding test (true: 1.6, false: -2.2)
- Score for cons. Seed: scoring for conserved seed (true: +3, false: -0.6)
- Total read count: The sum of read counts for the mature, loop and star miRNAs.
- Mature read count: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the consensus mature miRNA, including 2 nts upstream and 5 nts downstream.
- Loop read count: The number of reads that mapped to the miRNA hairpin and are contained in the sequence covered by the consensus miRNA loop, including 2 nts upstream and 5 nts downstream.
- Star read count: The number of reads that mapped to the miRNA hairpin and are contained

in the sequence covered by the consensus star miRNA, including 2 nts upstream and 5 nts downstream.

[Blue Box]: Secondary structure box; (Figure)

- Red character: Mature miRNA region
- Purple character: Star region
- Yellow character: Loop region

[Orange Box]: Read frequency distribution

- Red line: Mature miRNA region
- Purple character: Star region
- Yellow character: Loop region

[Green Box]: Secondary structure box; (Text)

- "exp" line: mature, loop, and star region expected by miRDeep2 algorithm (Red Character: mature miRNA region, Purple character: star region, Yellow character: loop region)
- "obs" line: mature, loop, and star region observed by read align result (Red Character: mature miRNA region, Purple character: star region, Yellow character: loop region)
- "known" line: The region that matches the known mature miRNA of an already miRBase (Green character)
- Last line: Secondary structure ("..."; Wobble, have not bond, "(", ")"; have bond)

[Purple Box]: Read Box

- Column 1: Read sequence (Upper case: mismatch sequence)
- Column 2: Read Count
- Column 3: Mismatch count

In order to compare the predicted miRNA information for each sample, novel miRNAs overlapping 70% or more in the same strand were assumed to be the same miRNA.

Table 15, 16 shows an example of the summary of merged novel precursor miRNAs. Refer to Expression_Profile.hsa.9606.RNAcentral.xlsx - Novel.precursor.Exp, Novel.precursor.Info sheet.

Table 17, 18 shows an example of the summary of merged novel mature miRNAs. Refer to Expression_Profile.hsa.9606.RNAcentral.xlsx - Novel.mature.Exp, Novel.mature.Info sheet.

Table 15. Expression profile according to novel precursor miRNA (Example)

Precursor_ID	Chr	Start	End	Strand	Length	AM_Read_Count	BM_Read_Count
Precursor_1	chr1	248074635	248074715	-	81	5517	3238
Precursor_2	chr1	249115811	249115882	-	72	24	0
Precursor_3	chr2	3721463	3721549	+	87	78	0
Precursor_4	chr2	6985560	6985610	-	51	0	18
Precursor_5	chr2	7539768	7539831	+	64	24	0
Precursor_6	chr2	25564019	25564087	-	69	18	0
Precursor_7	chr2	26278302	26278348	-	47	415	0
Precursor_8	chr2	28508623	28508683	-	61	0	103
Precursor_9	chr2	38816200	38816241	+	42	277	0
Precursor_10	chr2	38830014	38830074	-	61	0	15

- Precursor ID: merged precursor miRNA ID
- Chr, Start, End, Strand: Genomic position of merged precursor miRNA
- Length: Length of merged precursor miRNA
- [Sample] Read Count: Read count of the sample called "Sample Name"

Table 16. Information according to novel precursor miRNA (Example)

Precursor_ID	Precursor_1	Precursor_2	Precursor_3	Precursor_4	Precursor_5	Precursor_6
Chr	chr1	chr1	chr2	chr2	chr2	chr2
Start	248074635	249115811	3721463	6985560	7539768	25564019
End	248074715	249115882	3721549	6985610	7539831	25564087
Strand	-	-	+	-	+	-
Length	81	72	87	51	64	69
Total_Count	8755	24	78	18	24	65
Total_Quality	4.5	2	0	2.3	0.2	7
AM_ID	chr1_4580|chr1:248074635-248074715:-	chr1_4581|chr1:249115811-249115882:-	chr2_4608|chr2:3721463-3721549:+	.	chr2_4626|chr2:7539768-7539831:+	chr2_6589|chr2:25564019-25564087:-
AM_Read_Count	5517	24	78	.	24	18
AM_Quality	2.2	2	0	.	0.2	3.5
BM_ID	chr1_4932|chr1:248074637-248074713:-	.	.	chr2_6887|chr2:6985560-6985610:-	.	.
BM_Read_Count	3238	.	.	18	.	.
BM_Quality	2.3	.	.	2.3	.	.

- Precursor ID: merged precursor miRNA ID
- Chr, Start, End, Strand: Genomic position of merged precursor miRNA
- Length: Length of merged precursor miRNA
- Total Count: Sum of read count for each sample

- Total Quality: Sum of quality for each sample
- [Sample] ID, Read Count, Quality: Novel precursor miRNA ID, Read Count, Quality of each sample

Table 17. Expression profile according to novel mature miRNA (Example)

Mature_ID	Chr	Start	End	Strand	Length	AM_Read_Count	BM_Read_Count
Mature_1	chr1	11423039	11423058	+	20	348	423
Mature_2	chr1	23150271	23150293	-	23	47	10
Mature_3	chr1	27320065	27320083	+	19	0	18453
Mature_4	chr1	27458337	27458359	+	23	0	425
Mature_5	chr1	31508166	31508187	+	22	208	0
Mature_6	chr1	33788507	33788526	-	20	12	8
Mature_7	chr1	38325303	38325321	+	19	0	38
Mature_8	chr1	39080606	39080624	-	19	514	0
Mature_9	chr1	42217030	42217048	+	19	26	19
Mature_10	chr1	248074635	248074653	-	19	5517	3238

- Precursor ID: merged mature miRNA ID
- Chr, Start, End, Strand: Genomic position of merged mature miRNA
- Length: Length of merged mature miRNA
- [Sample] Read Count: Read count of the sample called "Sample Name"

Table 18. Information according to novel mature miRNA (Example)

Mature_ID	Mature_1	Mature_2	Mature_3	Mature_4	Mature_5	Mature_6
Chr	chr1	chr1	chr1	chr1	chr1	chr1
Start	11423039	23150271	27320065	27458337	31508166	33788507
End	11423058	23150293	27320083	27458359	31508187	33788526
Strand	+	-	+	+	+	-
Length	20	23	19	23	22	20
Total_Count	771	57	18451	425	208	20
Total_Quality	3	2.1	2.9	0.8	0.4	1.9
AM_ID	chr1_559_mature|chr1:11423039-11423058:+	chr1_2858_mature|chr1:23150271-23150293:-	.	.	chr1_861_mature|chr1:31508166-31508187:+	chr1_3011_mature|chr1:33788507-33788526:-
AM_Read_Count	348	47	.	.	208	12
AM_Quality	1.5	1.5	.	.	0.4	1.9
BM_ID	chr1_599_mature|chr1:11423039-11423058:+	chr1_3050_mature|chr1:23150271-23150293:-	chr1_840_mature|chr1:27320065-27320083:+	chr1_843_mature|chr1:27458337-27458359:+	.	chr1_3211_mature|chr1:33788508-33788526:-
BM_Read_Count	423	10	18451	425	.	8
BM_Quality	1.5	0.6	2.9	0.8	.	0

- Precursor ID: merged mature miRNA ID
- Chr, Start, End, Strand: Genomic position of merged mature miRNA
- Length: Length of merged mature miRNA
- Total Count: Sum of read count for each sample
- Total Quality: Sum of quality for each sample
- [Sample] ID, Read Count, Quality: Novel mature miRNA ID, Read Count, Quality of each sample

4. 4. Small RNA Composition and Profiling

There are various kinds of small RNAs in cells. In this analysis, we used RNAcentral 14.0 DB to quantify the expression of small RNAs for each sample.

Final processed reads were aligned to small RNAs (50 nt or less; piRNA) of the DB using bowtie (option: -f -l 15 --norc -v 1 -a) and to small RNAs (50nt or greater; tRNA, snoRNA, etc.) of the DB using bowtie2 (option: -f -L 15 --norc -a) which assigned a result of 90% or more of the coverage as a corresponding RNA.

We also used HISAT2 (option: -f --dta --rna-strandness F) to identify sequences that did spliced RNA, and STAR (option: --estimate-rspd --append-names --seed-length 15 --strandedness forward) to align the transcript sequence.

In addition, bowtie (option: -f -l 15 --norc -v 1 -k 100) was used to match the repeat sequence in the genome.

Table 19 shows the categories of smRNAs present in *Homo sapiens* in RNAcentral.

Table 20 shows the priority of the small RNA category in this study.

Table 19. RNA categoties in RNAcentral Database for species

Taxon ID	Species	Small RNA category
9606	Homo sapiens	rRNA, tRNA, piRNA, snoRNA, snRNA, Y_RNA, scRNA, vault_RNA

Table 20. Priority of small RNA category

Priority	Small RNA category
1	Known miRNA
2	Novel miRNA
3	piRNA
4	snoRNA
5	snRNA
6	rRNA
7	tRNA
8	Y RNA
9	scRNA
10	vault RNA
11	Genome
12	Repeat
13	Mycoplasma
14	TruSeqIndexedAdapter

Figure 9 is the bar plot of RNA composition without priority for MG_CTRL_1.
Figure 10 is the pie chart of RNA composition with priority for MG_CTRL_1.
The plots of other samples are located in result_smRNA_excel/RNA_Composition_result/

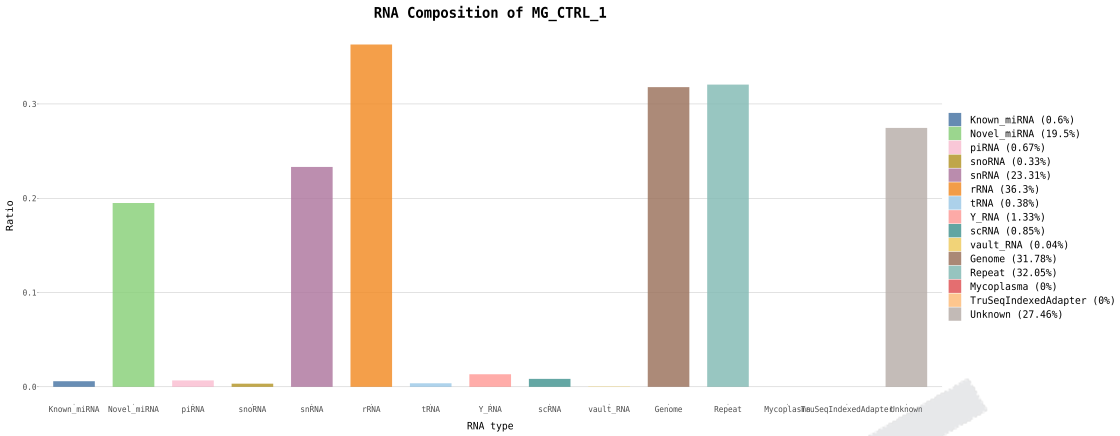


Figure 9. RNA composition without priority for MG_CTRL_1

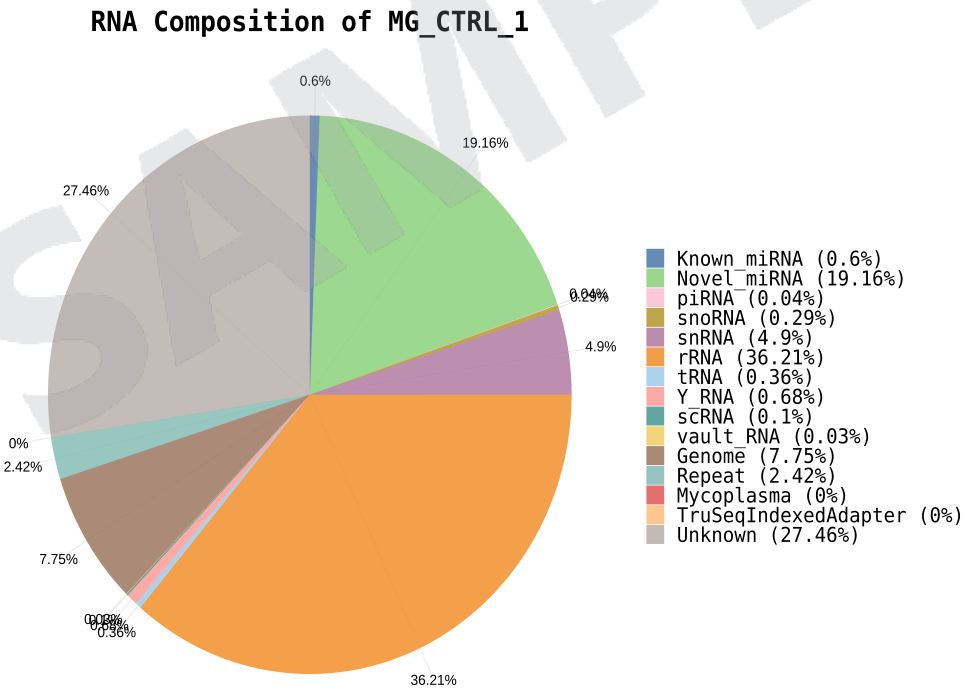


Figure 10. RNA composition with priority for MG_CTRL_1

Table 21 summarizes read counts for each small RNA ID, except for miRNA and rRNA. Expression_Profile.hsa.9606.RNAcentral.xlsx contains sheets of each smRNA category. You can find the expression profile of samples for each smRNA type from each sheet in the file.

Table 21. Expression profile of small RNA (example)

SeqID	Length	Description	AM_RPM	BM_RPM	AM_Read_Count	BM_Read_Count
URS000000022A	27	Homo sapiens (human) piR-61547	0	0	0	0
URS000000023A	30	Homo sapiens (human) piR-45735	0	0	0	0
URS0000002D9E	32	Homo sapiens (human) piR-36743	123683.0312	43332.98121	82246	7794
URS0000009A89	28	Homo sapiens (human) piR-34358	129.328365	928.484458	86	167
URS000000AA0C	26	Homo sapiens (human) piR-49039	7.519091	183.472977	5	33
URS000000E79F	31	Homo sapiens (human) piR-33081	1171.474373	1923.686361	779	346
URS000000EED0	28	Homo sapiens (human) piR-57559	7.519091	11.119574	5	2

- SeqID: Sequence ID of each RNA by RNAcentral DB
- Length: Sequence length
- Description: Description of sequence
- [Sample]_RPM: Read per million of each sample
- [Sample]_Read_Count: Read count of each sample

5. Differentially Expressed miRNA Analysis Results

5. 1. Data Analysis Quality Check and Preprocessing

There is a process that sorts differentially expressed miRNA among samples by read count value of mature miRNAs. In preprocessing, there are data quality and similarity checks among samples in case of biological replicates exist.

(Refer to Path: result_smRNA_excel/DEmiRNA_result/Analysis_Result.html)

5. 1. 1. Sample Information and Analysis Design

Total of 6 samples was used for analysis. For more information of samples and comparison pair, please refer to Sample.Info.txt file.

Index	Sample.ID	Sample.Group
1	MG_CTRL_1	CTRL
2	MG_CTRL_2	CTRL
3	MG_CTRL_3	CTRL
4	MG_TEST_1	TEST
5	MG_TEST_2	TEST
6	MG_TEST_3	TEST

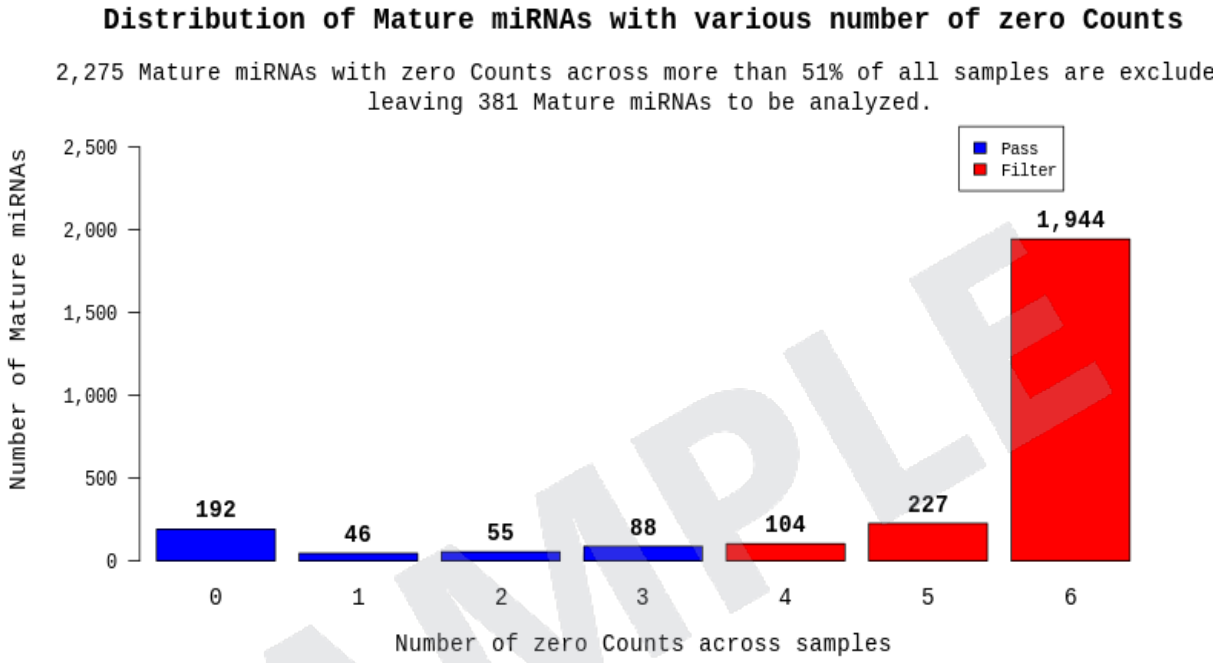
Comparison pair and statistical method for each pair are shown below.

Index	Test vs. Control	Statistical Method
1	TEST vs. CTRL	Fold Change, exactTest using edgeR, Hierarchical Clustering

5. 1. 2. DATA Quality Check

(Refer to Path: result_smRNA_excel/DEmiRNA_result/Data Quality Check/)

For each Mature miRNA, 2,275 Mature miRNAs with zero Counts across more than 51% of all samples are excluded leaving 381 Mature miRNAs to be analyzed.

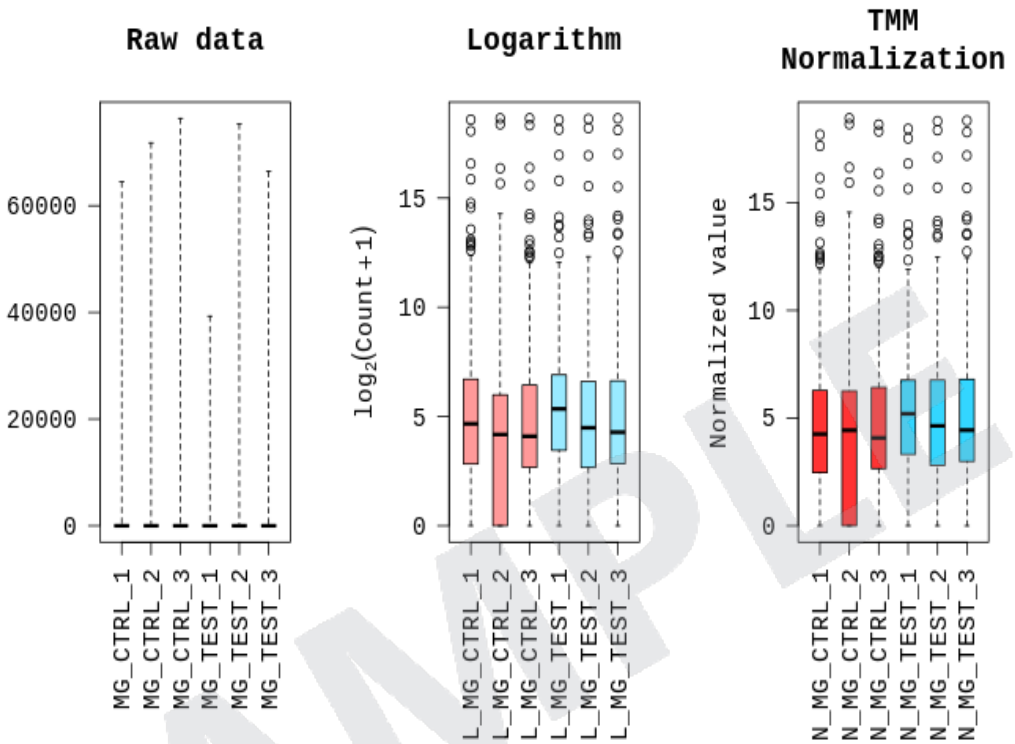


5. 1. 3. Data Transformation and Normalization

In order to reduce systematic bias, estimates the size factors from the count data and applies Trimmed Mean of M-values (TMM) normalization with edgeR R library.

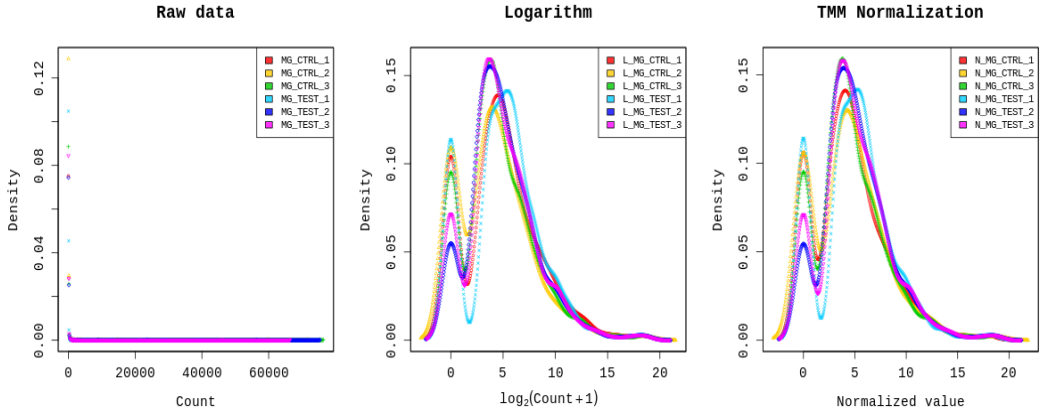
5. 1. 3. 1. Boxplot of Expression Difference between samples.

Below boxplots show the corresponding sample's expression distribution based on percentile (median, 50 percentile, 75 percentile, maximum and minimum) based on raw signal (read count), Log2 transformation of read count+1 and TMM Normalization.



5. 1. 3. 2. Expression Density Plot per sample

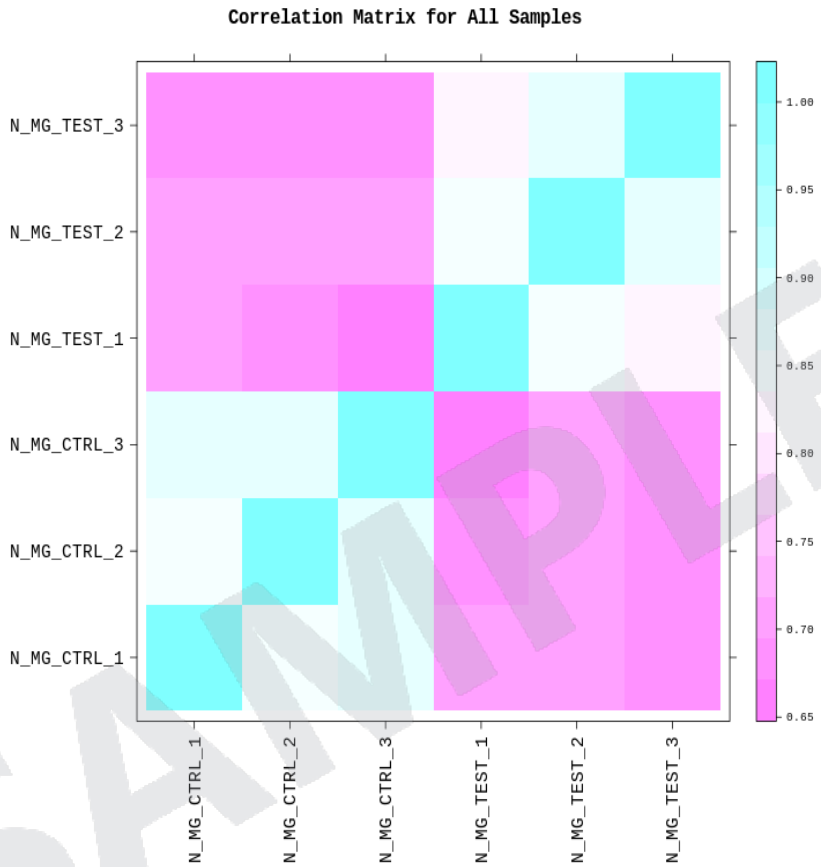
Below density plots show the corresponding samples expression distribution before and after of raw signal (read count), Log2 transformation of read count+1 and TMM Normalization.



5. 1. 4. Correlation Analysis between samples

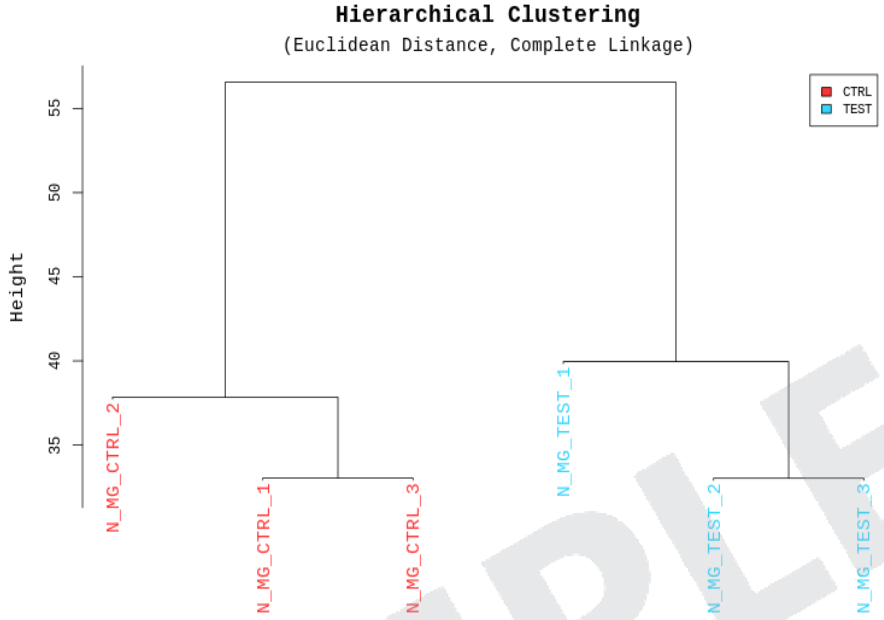
The similarity between samples are obtained through Pearson's coefficient of the normalized value. For range: -1≤ r ≤ 1,the closer the value is to 1, the more similar the samples are.

Correlation matrix of all samples is as follows.



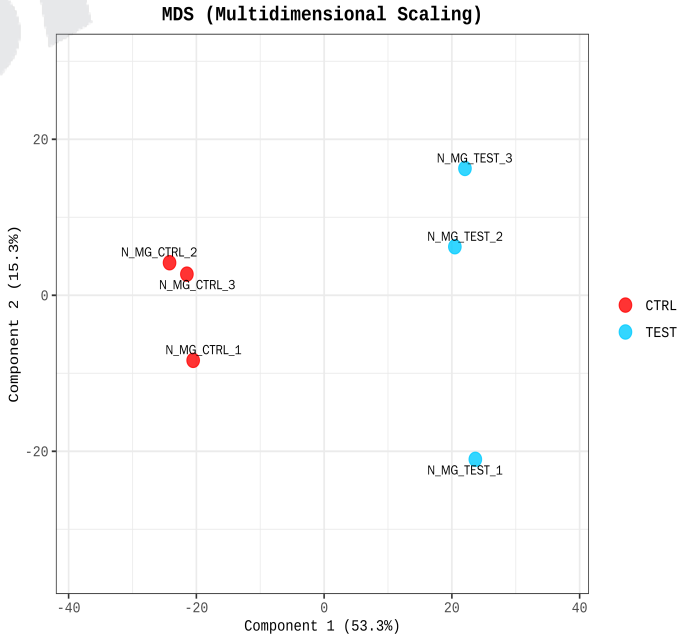
5. 1. 5. Hierarchical Clustering Analysis

Using each sample's normalized value, the high expression similarities were grouped together. (Distance metric = Euclidean distance, Linkage method= Complete Linkage)



5. 1. 6. Multidimensional Scaling Analysis

Using each sample's normalized value, the similarity between samples is graphically shown in a 2D plot to show the variability of the total data. This allows identification any outlier samples, or similar expression patterns between sample groups.



5. 2. Differentially Expressed miRNA Analysis Workflow

Below shows the orders of DEG (Differentially Expressed Genes) analysis.

1) The read counts of mature miRNA obtained from miRDeep2 Quantifier module are used as the original raw data.

- Raw data

(Refer to Path: result_smRNA_excel/Expression_Profile/Expression_Profile.hsa.GRCh38.miRNA.xlsx)

: 2,656 Mature miRNAs, 6 samples

2) During data preprocessing, low quality miRNAs are filtered. Afterwards, TMM Normalization are performed.

- Processed data

(Refer to Path: result_smRNA_excel/DEmiRNA_result/data2.xlsx)

: 381 Mature miRNAs, 6 samples

3) Statistical analysis is performed using Fold Change, exactTest using edgeR per comparison pair. The significant results are selected on conditions of |fc|>=2 & exactTest raw p-value<0.05.

- Significant data

(Refer to Path: result_smRNA_excel/DEmiRNA_result/data3_fc2_&_raw.p.xlsx)

: 85 Mature miRNAs

4) For significant lists, hierarchical clustering analysis is performed to group the similar samples and Mature miRNAs. These results are graphically depicted using heatmap and dendogram.

- Hierarchical Clustering (Euclidean Distance, Complete Linkage)

(Refer to Path: result_smRNA_excel/DEmiRNA_result/Cluster image/)

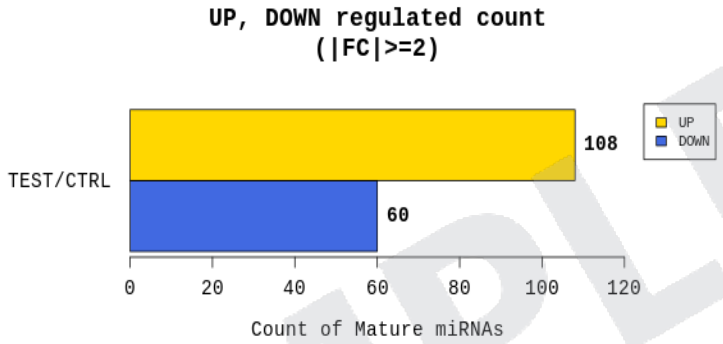
5. 3. Significant Mature miRNA Results

(Refer to Path: result_smRNA_excel/DEmiRNA_result/Plots/)

These are fc2_&_raw.p, TEST_vs_CTRL results by example.

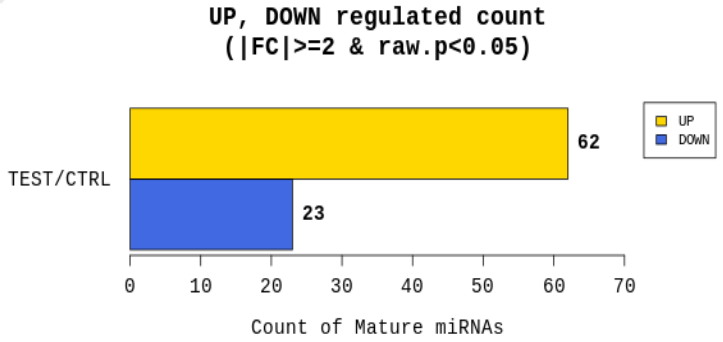
5. 3. 1. Up, Down Regulated Count by Fold Change

Shows number of up and down regulated mature miRNAs based on fold change of comparison pair.



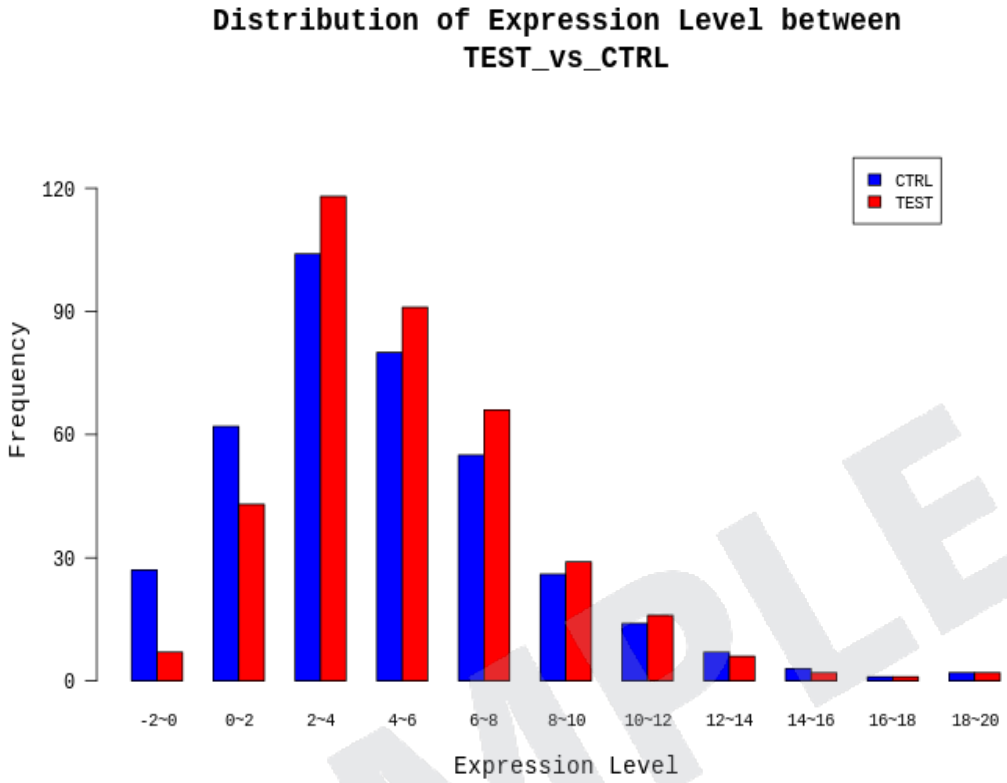
5. 3. 2. Up, Down Regulated Count by Fold Change and p-value

Shows number of up and down regulated mature miRNAs based on fold change and p-value of comparison pair.



5. 3. 3. Distribution of Expression Level between two groups

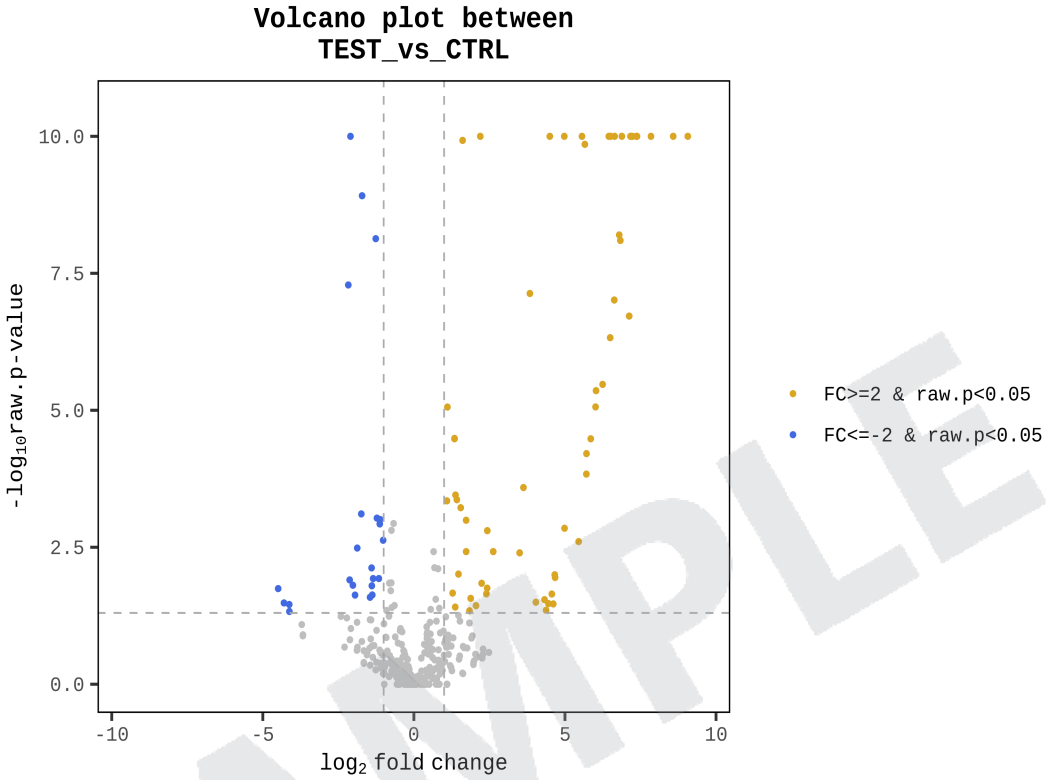
Shows distribution of normalized value of each group for comparison pair.



5. 3. 4. Volcano Plot of Expression Level of two groups.

Log2 fold change and p-value obtained from the comparison between two groups plotted as volcano plot.

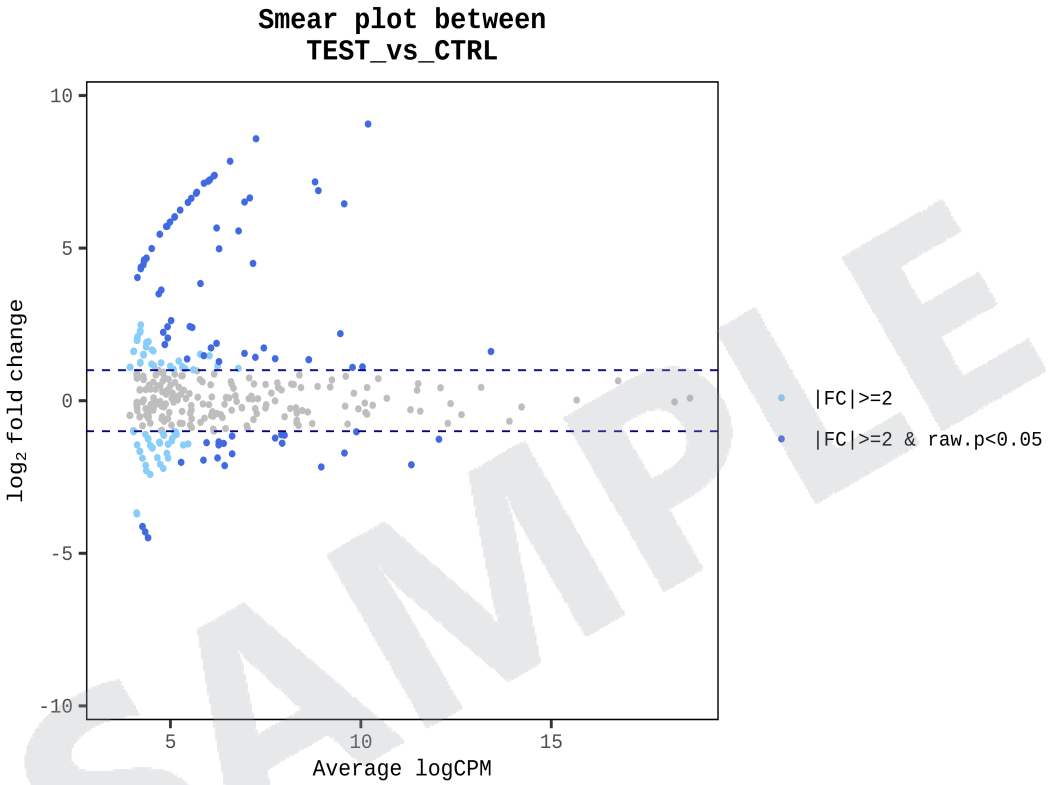
(X-axis: log2 Fold Change, Y-axis: -log10 p-value)



5. 3. 5. Smear Plot

In order to confirm the transcripts that show higher expression difference compared to the control according to overall average expression level (Smear), smear plot is drawn. (X-axis: Average logCPM, Y-axis: log2 Fold Change).

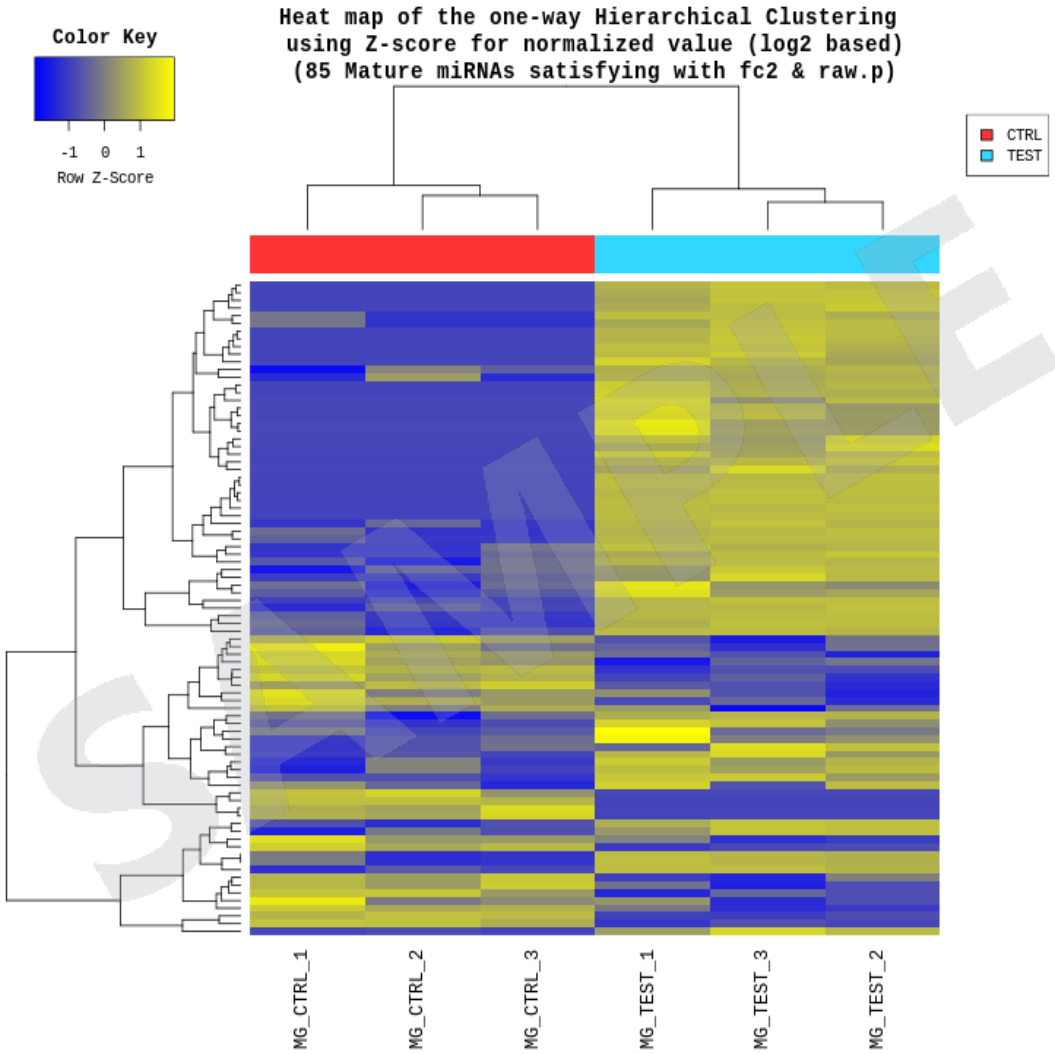
For example, even though fold change might be different by two-fold, the mature miRNA with higher average logCPM may be more credible.



5. 3. 6. Hierarchical Clustering Analysis

(Refer to Path: result_smRNA_excel/DEmiRNA_result/Cluster image/)

Heatmap shows result of hierarchical clustering analysis (Euclidean Method, Complete Linkage) which clusters the similarity of mature miRNAs and samples by expression level (normalized value) from significant list.



6. Data Download Information

6. 1. Raw Data

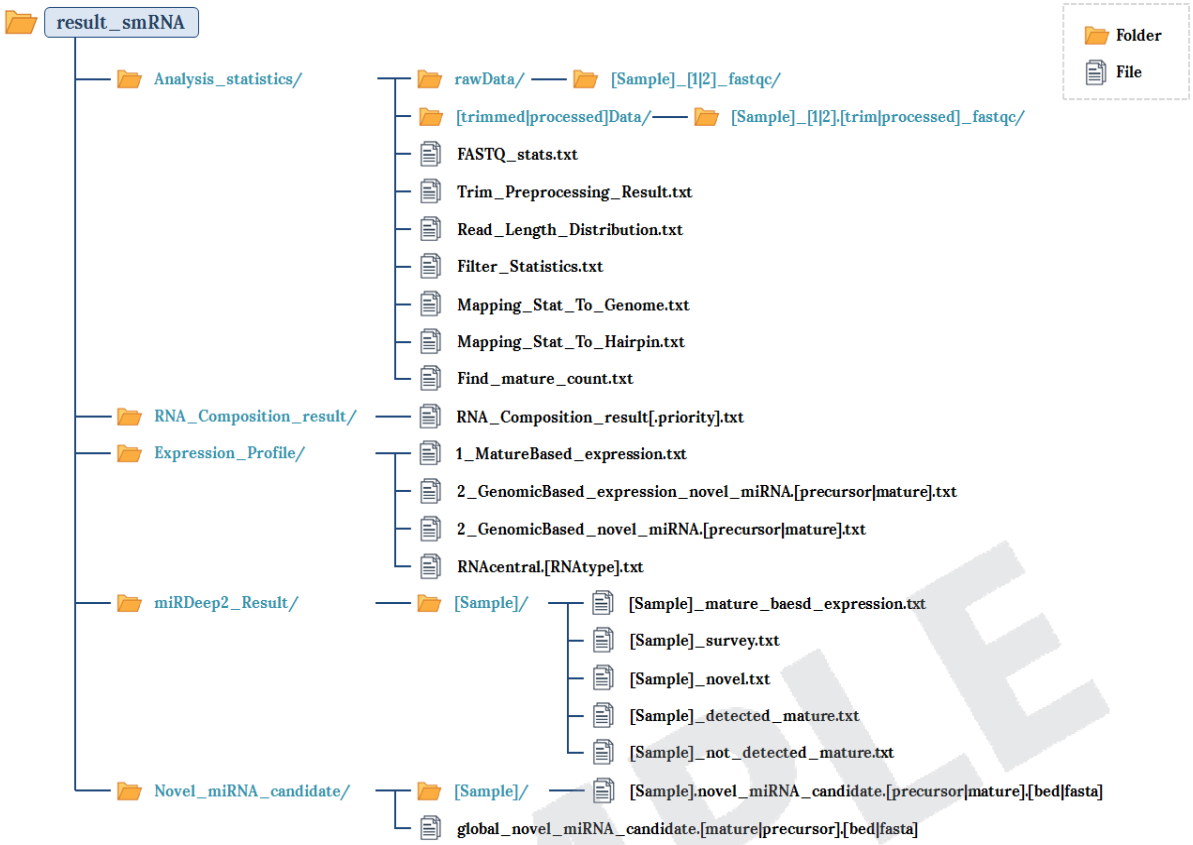
Raw data is the FASTQ file that isn't trimmed adapter sequence.

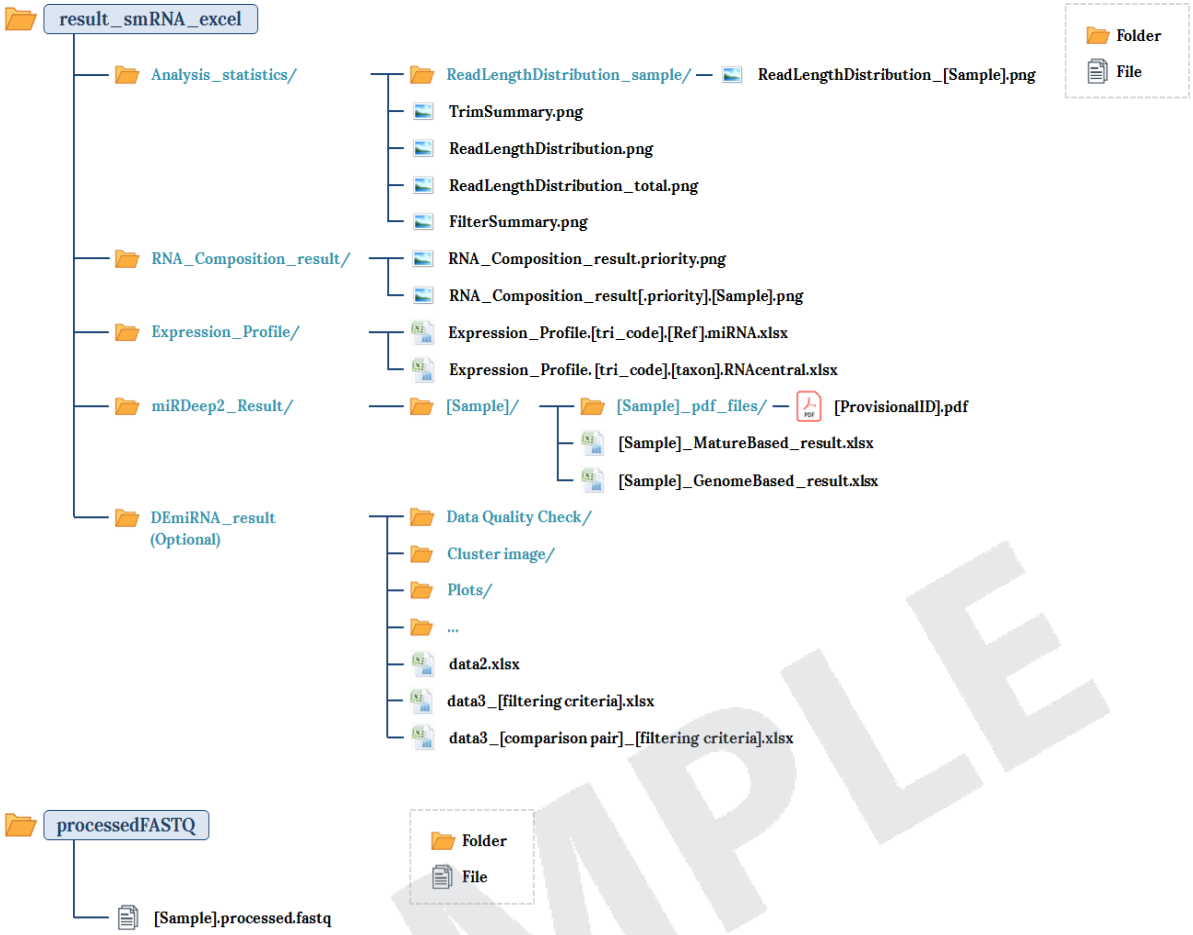
Download link	File size	md5sum
MG_CTRL_1_1.fastq.gz	1.46G	04ac48d7d91381a17e1f821afe685a16
MG_CTRL_2_1.fastq.gz	1.36G	77c40b4daa5dfc3e646a0087ac4c5e73
MG_CTRL_3_1.fastq.gz	1.53G	b4b0faf594ca8450c3c2c819f0ddfdb5
MG_TEST_1_1.fastq.gz	1.54G	2b3ea08f84818d67d8f55dbee8725b1c
MG_TEST_2_1.fastq.gz	1.81G	6a754673988dbfc4ce39751565ec8632
MG_TEST_3_1.fastq.gz	1.81G	007b05dadf9333d3e47872023d6a2273

- fastq.gz : This is a zip file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

6. 2. Analysis Results

Download link	File size
smRNA_result_smRNA.zip (md5sum: e76f7540427d040e265db1ac1ec33a84)	8.01M
smRNA_result_smRNA_excel.zip (md5sum: 98ae0be66cc9b003b8bfc3f79c4e8abf)	55.07M
smRNA_processedFASTQ.zip (md5sum: 4b2706255042d6a7e48831566be5f0ba)	3.21G


result_smRNA
Folder
File
Analysis_statistics/
rawData/
[Sample]_[1|2]_fastqc/
[trimmed|processed]Data/
[Sample]_[1|2].[trim|processed]_fastqc/
FASTQ_stats.txt
Trim_Preprocessing_Result.txt
Read_Length_Distribution.txt
Filter_Statistics.txt
Mapping_Stat_To_Genome.txt
Mapping_Stat_To_Hairpin.txt
Find_mature_count.txt
RNA_Composition_result/
RNA_Composition_result[.priority].txt
Expression_Profile/
1_MatureBased_expression.txt
2_GenomicBased_expression_novel_miRNA.[precursor|mature].txt
2_GenomicBased_novel_miRNA.[precursor|mature].txt
RNAcentral.[RNAtype].txt
miRDeep2_Result/
[Sample]/
[Sample]_mature_baesd_expression.txt
[Sample]_survey.txt
[Sample]_novel.txt
[Sample]_detected_mature.txt
[Sample]_not_detected_mature.txt
Novel_miRNA_candidate/
[Sample]/
[Sample].novel_miRNA_candidate.[precursor|mature].[bed|fasta]
global_novel_miRNA_candidate.[mature|precursor].[bed|fasta]



⚠ Your data will be retained in our server for 3 months.
Should you wish to extend the retention period, please contact us.

7. Appendix

7. 1. Phred Quality Score Chart

Phred quality score numerically express the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

7. 2. Programs used in Analysis

7. 2. 1. FastQC v0.11.7

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC is a program that performs quality check on the raw sequences before analysis to make sure data integrity. The main function is importing BAM, SAM, FastQ files and providing quick overview on which section has problems. It provides such results as graphs and tables in html files.

7. 2. 2. Cutadapt 1.16

LINK https://cutadapt.readthedocs.org/en/stable/

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

- -q: Try to trim low-quality ends from reads before adapter removal.
- -g: Sequence of an adapter that was ligated to the 5' end.
- -a: Sequence of an adapter that was ligated to the 3' end.
- -O: Minimum overlap length.
- --max-n: The max proportion of N's allowed in a read.
- -m: Discard trimmed reads that are shorter than LENGTH.
- -n: Try to remove adapters at most COUNT times.

7. 2. 3. miRDeep2 2.0.0.8, Bowtie 1.1.2

LINK https://www.mdc-berlin.de/content/mirdeep2-documentation

miRDeep2 is a completely overhauled tool which discovers microRNA genes by analyzing sequenced RNAs. The tool reports known and hundreds of novel microRNAs with high accuracy in seven species representing the major animal clades. The low consumption of time and memory combined with user-friendly interactive graphic output makes miRDeep2 accessible for straightforward application in current reasearch.

7. 2. 4. Bowtie 1.1.2

LINK http://bowtie-bio.sourceforge.net/index.shtml

Bowtie is an ultrafast, memory-efficient short read aligner. Bowtie's speed and small memory footprint are due chiefly to its use of the Burrows-Wheeler index in combination with the novel, quality-aware, backtracking algorithm.

7. 2. 5. HISAT2 version 2.1.0, Bowtie2 2.3.4.1

LINK https://ccb.jhu.edu/software/hisat2/index.shtml

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads to genomes. Its first implementation based on an extension of BWT for graphs, designed a graph FM index (GFM). In addition to using one global GFM index, HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover the human population). These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

7. 2. 6. RSEM verion v1.3.1, STAR 2.6.0c

LINK http://deweylab.github.io/RSEM/

RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data. The RSEM package provides an user-friendly interface, supports threads for parallel computation of the EM algorithm, single-end and paired-end read data, quality scores, variable-length reads and RSPD estimation. In addition, it provides posterior mean and 95% credibility interval estimates for expression levels.

7. 2. 7. miRBase release 22.1

LINK http://www.mirbase.org/

The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation.

7. 2. 8. RNAcentral version 14.0

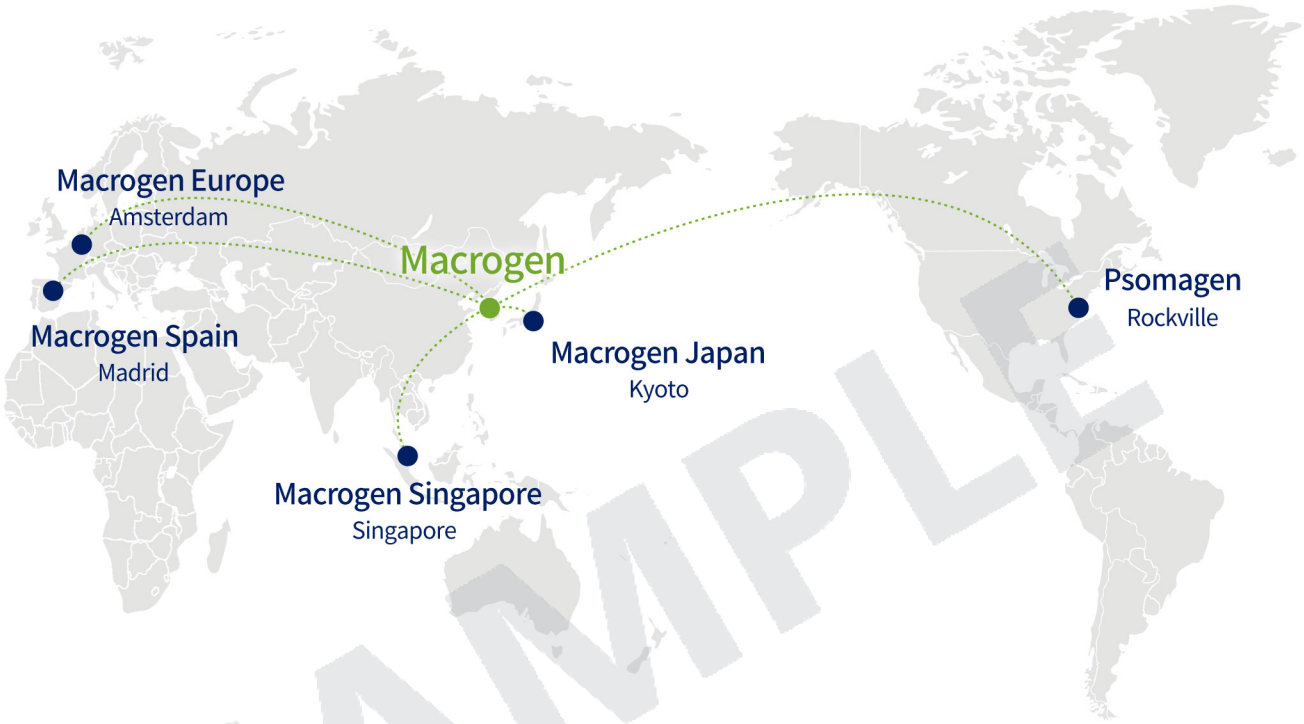
LINK https://rnacentral.org/

RNAcentral is a free, public resource that offers integrated access to a comprehensive and up-to-date set of non-coding RNA sequences provided by a collaborating group of Expert Databases representing a broad range of organisms and RNA types.

7. 3. References

1. MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 2011, 17.1: pp. 10-12.
2. FRIEDLAENDER, Marc R., et al. Discovering microRNAs from deep sequencing data using miRDeep. Nature biotechnology, 2008, 26.4: 407.
3. FRIEDLAENDER, Marc R., et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research, 2011, 40.1: 37-52.
4. LANGMEAD, Ben, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10.3: R25.
5. KIM, Daehwan; LANGMEAD, Ben; SALZBERG, Steven L. HISAT: a fast spliced aligner with low memory requirements. Nature methods, 2015, 12.4: 357-360.
6. LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. Nature methods, 2012, 9.4: 357.
7. LI, Bo; DEWEY, Colin N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics, 2011, 12.1: 1.
8. DOBIN, Alexander, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, 29.1: 15-21.
9. KOZOMARA, Ana; GRIFFITHS-JONES, Sam. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research, 2013, 42.D1: D68-D73.
10. RNACENTRAL CONSORTIUM. RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic acids research, 2016, gkw1008.
11. LI, Heng, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16: 2078-2079.
12. QUINLAN, Aaron R.; HALL, Ira M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010, 26.6: 841-842.

Research use only



HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro, Geumcheon-gu, Seoul, Republic of Korea (Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com (Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 31, 1105 AZ, Amsterdam, the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Macrogen Singapore

Laboratory, Business & Support Center

3 Biopolis Drive #05-18, Synapse, Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

Psomagen (Macrogen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville, MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Japan

Laboratory, Business & Support Center

3F Kyoto University International Science Innovation Bldg.
36-1 Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN
Tel: +81-75-746-2773
Email: customer@macrogen-japan.co.jp

BRANCH

Macrogen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas, 28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

Humanizing Genomics
macrogen