Research use only

Mus musculus Transcriptome Sequencing

Report

May 2024

Humanizing Genomics macrogen

ver. MGTR3.4_HS_GRCm39

Project Information

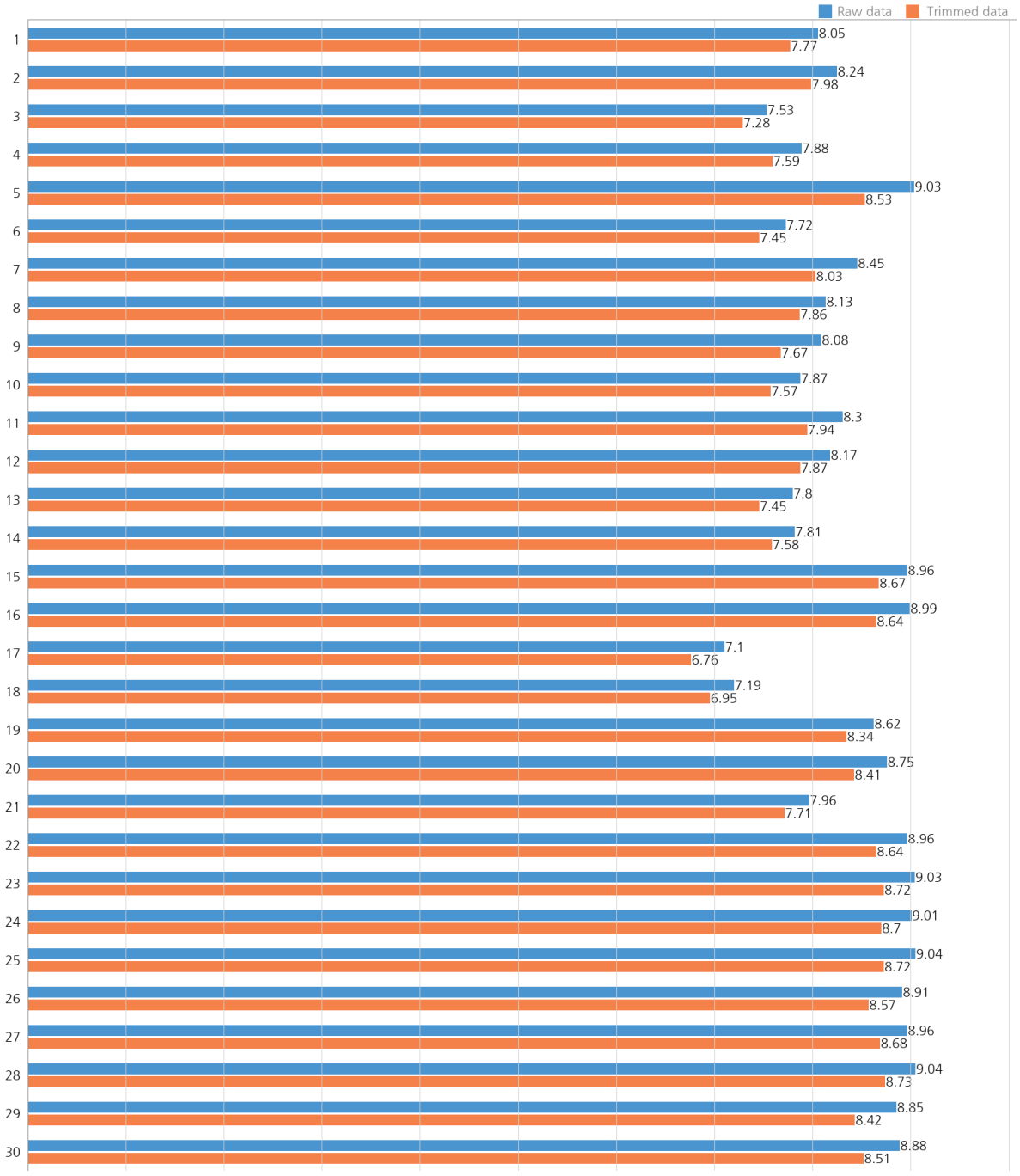
Client Name	TESTER
Company/Institution	
Order Number	
Species	*Mus musculus*
Reference	GRCm39
Annotation	NCBI_202304
Type of Read	Paired-ends
Read Length	151
Number of Samples	54
Library Kit	TruSeq Stranded mRNA Library Prep Kit
Library Protocol	TruSeq Stranded mRNA Reference Guide # 1000000040498 v00
Type of Sequencer	Illumina platform

Project Results Summary

In this study, *Mus musculus* whole transcriptome sequencing was performed in order to examine the gene expression profiles.

Analyses were successfully performed on all 54 paired-ends samples. Figure 1 shows the throughput of raw data and trimmed data. Figure 2 shows the Q30 percentage (% of bases with quality over phred score 30) of each sample's raw and trimmed data.

Raw data vs. Trimmed data (Throughput)



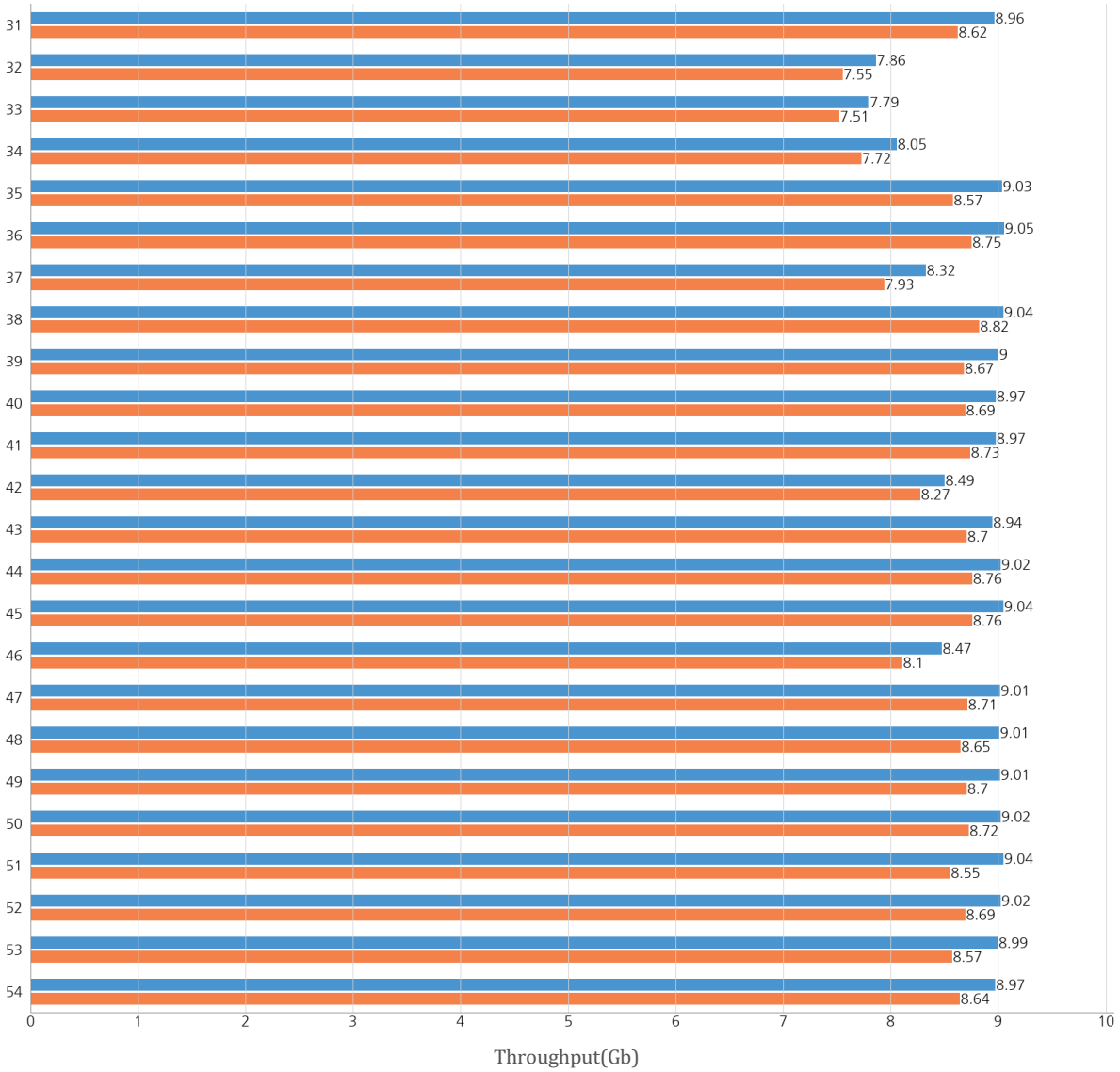
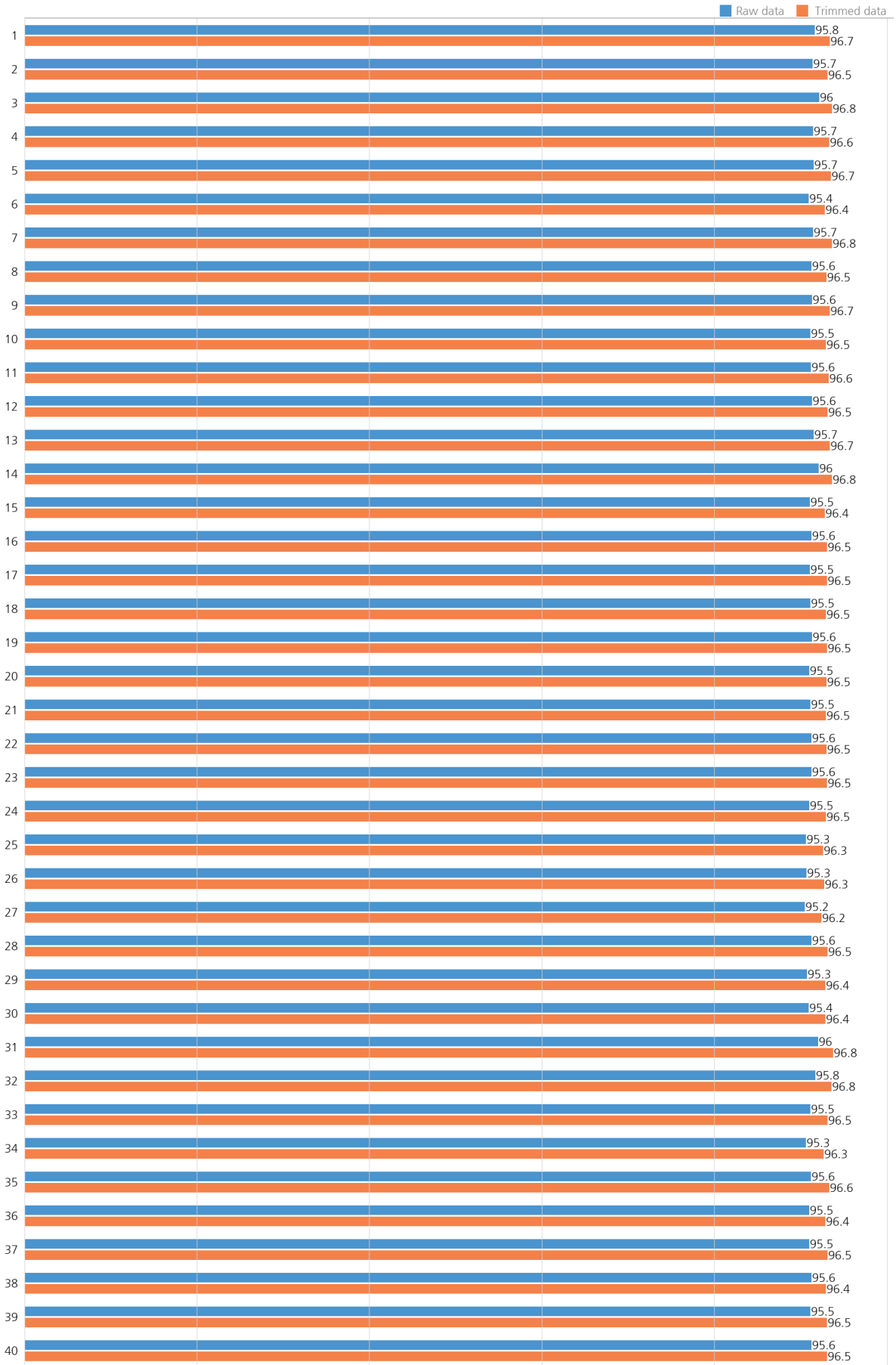


Figure 1. Throughput output of Raw and Trimmed data

Raw data vs. Trimmed data (≥Q30)



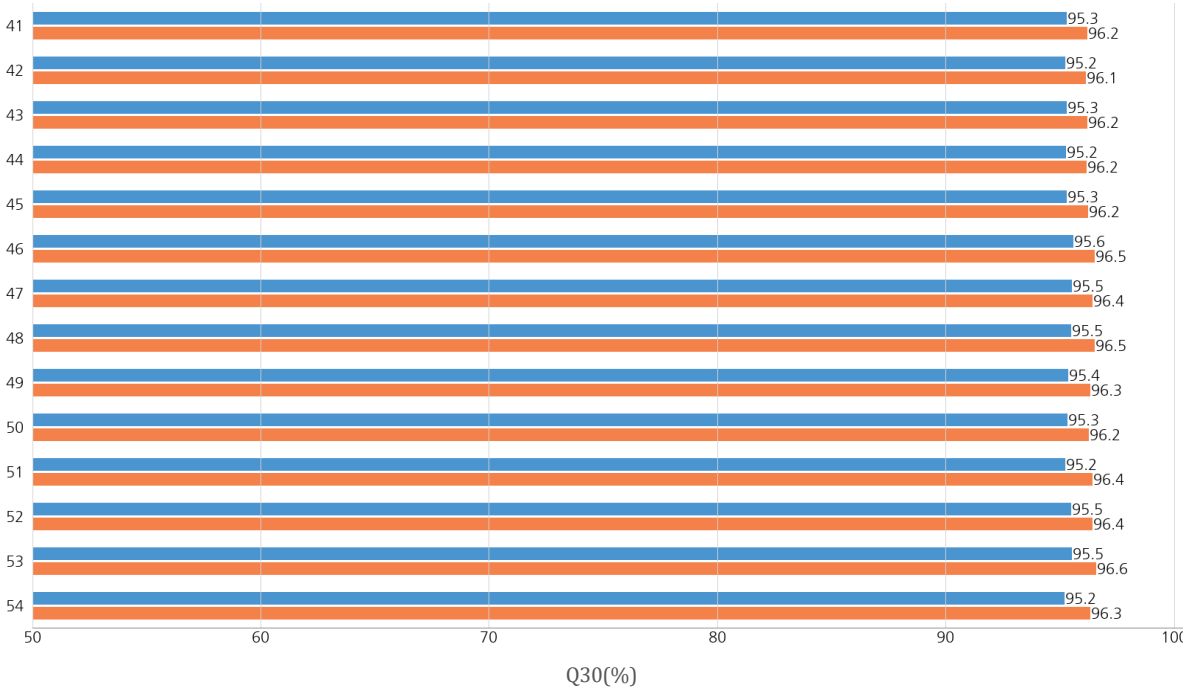


Figure 2. Q30 score of Raw and Trimmed data

Trimmed reads are mapped to reference genome with HISAT2. Figure 3 shows the overall read mapping ratio, the ratio of mapped reads to trimmed reads.

Overall read mapping ratio(%)

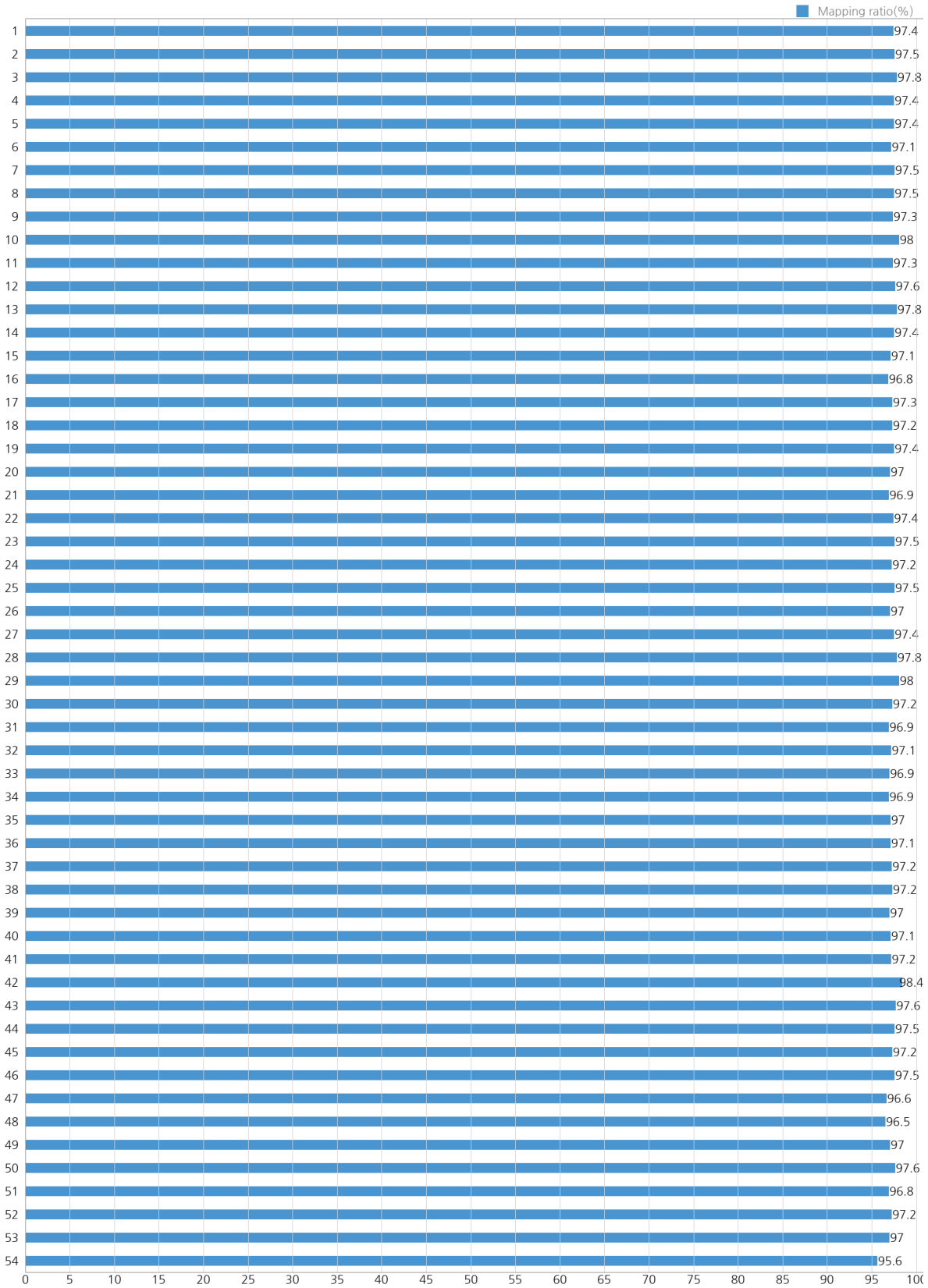


Figure 3. Overall read mapping ratio(%)

After the read mapping, Stringtie was used for transcript assembly. Expression profile was calculated for each sample and transcript/gene as read count, FPKM (Fragment per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Kilobase Million).

Table of Contents

1. Experimental Methods and Workflow

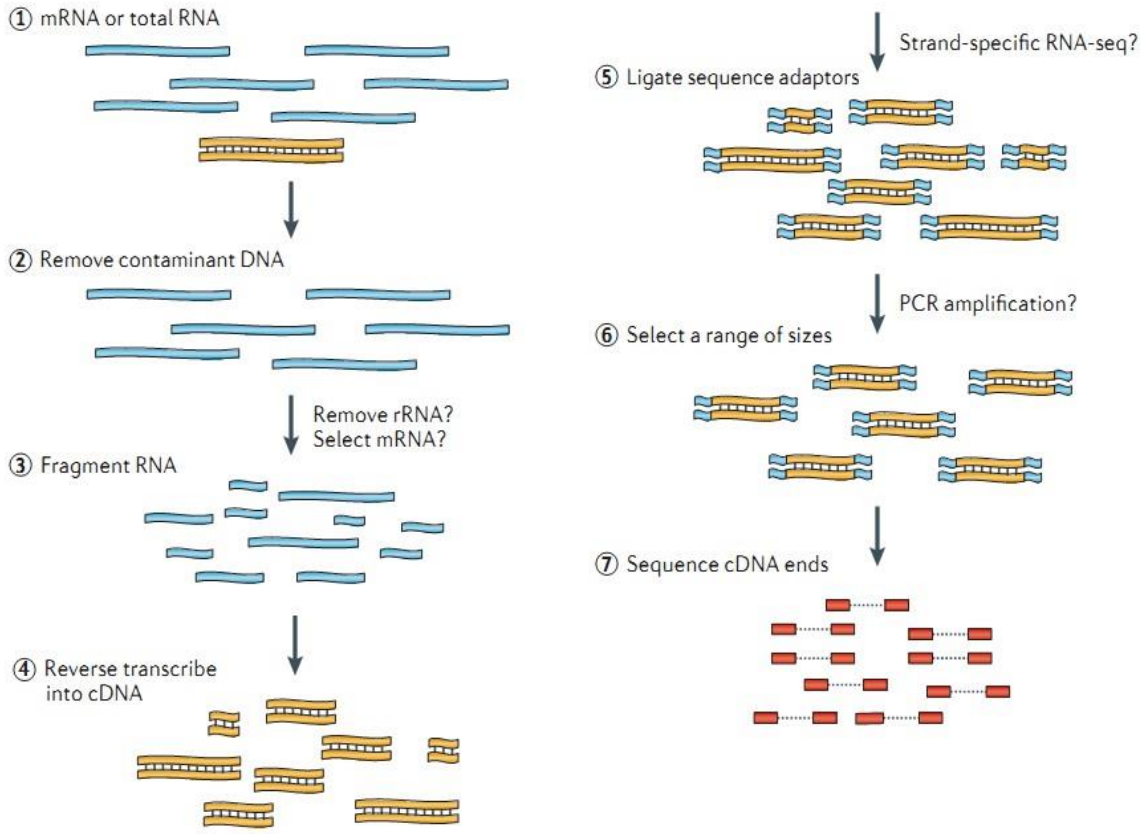


Figure 4. RNA Sequencing Experiment Workflow

REFERENCE Nat Rev Genet. 2011 Sep 7;12(10):671-82

1) Isolate the Total RNA from Sample of interest (Cell or Tissue).

2) Eliminate DNA contamination using DNase.

3) Choose an appropriate kit for library prep process depending on the types of RNA. For mRNA with poly-A tail, use mRNA purification kit; for non-coding RNAs, such as lincRNA, use ribo-zero RNA removal Kit to purify RNA of interest.

4) Randomly fragment purified RNA for short read sequencing.

5) Reverse transcribe fragmented RNA into cDNA.

6) Ligate adapters onto both ends of the cDNA fragments.

7) After amplifying fragments using PCR, select fragments with insert sizes between 200-400 bp. For paired-end sequencing, both ends of the cDNA is sequenced by the read length.

2. Analysis Methods and Workflow

Figure 5. Analysis Workflow

1) Analyze the quality control of the sequenced raw reads. Overall reads' quality, total bases, total reads, GC (%) and basic statistics are calculated.

2) In order to reduce biases in analysis, artifacts such as low quality reads, adaptor sequence, contaminant DNA, or PCR duplicates are removed.

3) Trimmed reads are mapped to reference genome with HISAT2, splice-aware aligner.

4) Transcript is assembled by StringTie with aligned reads.

5) Expression profiles are represented as read count and normalization values which are calculated based on transcript length and depth of coverage. Normalization values are provided as FPKM (Fragments Per Kilobase of transcript per Million Mapped reads) / RPKM (Reads Per Kilobase of transcript per Million mapped reads) and TPM(Transcripts Per Kilobase Million).

3. Summary of Data Production

3.1. Raw Data Statistics

(Refer to Path: result_RNAseq/Analysis_statistics/raw_throughput.txt)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 54 samples. For example, in 1, 53,334,504 reads are produced, and total read bases are 8.1Gbp. The GC content (%) is 50.29% and Q30 is 95.8%.

Table 1. Raw data stats

Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
1	8,053,510,104	53,334,504	50.29	98.59	95.8
2	8,244,713,854	54,600,754	49.79	98.55	95.66
3	7,530,857,428	49,873,228	51.32	98.69	96.04
4	7,882,857,152	52,204,352	50.78	98.55	95.7
5	9,028,230,506	59,789,606	51.5	98.54	95.72
6	7,722,462,838	51,142,138	49.96	98.47	95.43
7	8,450,130,328	55,961,128	50.75	98.5	95.69
8	8,125,643,408	53,812,208	49.17	98.55	95.61
9	8,082,878,094	53,528,994	50.53	98.51	95.63
10	7,869,804,108	52,117,908	48.73	98.51	95.55
11	8,303,264,406	54,988,506	50.0	98.49	95.58
12	8,172,111,544	54,119,944	49.99	98.53	95.63
13	7,795,439,930	51,625,430	50.16	98.55	95.72
14	7,811,772,694	51,733,594	50.04	98.68	96.01
15	8,957,830,682	59,323,382	49.63	98.5	95.49
16	8,986,611,886	59,513,986	50.2	98.53	95.59
17	7,100,274,284	47,021,684	50.93	98.45	95.52
18	7,191,397,952	47,625,152	49.83	98.48	95.55
19	8,621,108,232	57,093,432	50.45	98.53	95.62
20	8,754,740,212	57,978,412	49.82	98.45	95.46
21	7,959,415,964	52,711,364	49.57	98.51	95.55
22	8,961,334,788	59,346,588	49.89	98.52	95.6
23	9,034,910,746	59,833,846	48.91	98.53	95.59
24	9,005,959,214	59,642,114	48.68	98.45	95.47
25	9,042,469,806	59,883,906	48.72	98.4	95.28
26	8,906,843,418	58,985,718	49.2	98.39	95.3
27	8,961,179,560	59,345,560	49.25	98.4	95.23

28	9,042,166,900	59,881,900	49.41	98.5	95.59
29	8,849,577,272	58,606,472	49.38	98.4	95.33
30	8,883,419,090	58,830,590	49.6	98.46	95.44
31	8,956,281,422	59,313,122	49.01	98.64	95.99
32	7,858,836,374	52,045,274	48.6	98.59	95.84
33	7,792,035,182	51,602,882	48.88	98.49	95.55
34	8,048,290,034	53,299,934	48.36	98.38	95.27
35	9,031,260,170	59,809,670	50.06	98.48	95.58
36	9,046,396,712	59,909,912	49.37	98.47	95.47
37	8,323,620,716	55,123,316	49.18	98.43	95.48
38	9,043,205,780	59,888,780	48.79	98.55	95.61
39	8,996,136,362	59,577,062	49.28	98.48	95.53
40	8,974,618,560	59,434,560	48.65	98.52	95.6
41	8,973,989,494	59,430,394	48.85	98.43	95.27
42	8,492,754,910	56,243,410	48.35	98.42	95.23
43	8,942,043,330	59,218,830	48.73	98.44	95.27
44	9,018,099,010	59,722,510	48.55	98.43	95.25
45	9,041,543,874	59,877,774	48.53	98.41	95.28
46	8,471,472,970	56,102,470	49.36	98.5	95.56
47	9,013,147,418	59,689,718	48.71	98.49	95.5
48	9,006,907,192	59,648,392	48.7	98.45	95.47
49	9,008,241,428	59,657,228	48.84	98.44	95.36
50	9,015,693,278	59,706,578	49.08	98.44	95.32
51	9,042,642,248	59,885,048	49.83	98.32	95.21
52	9,016,244,126	59,710,226	49.37	98.48	95.47
53	8,994,213,226	59,564,326	49.61	98.44	95.5
54	8,968,957,872	59,397,072	49.35	98.33	95.18

(* 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: Analysis_statistics/rawData/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.

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

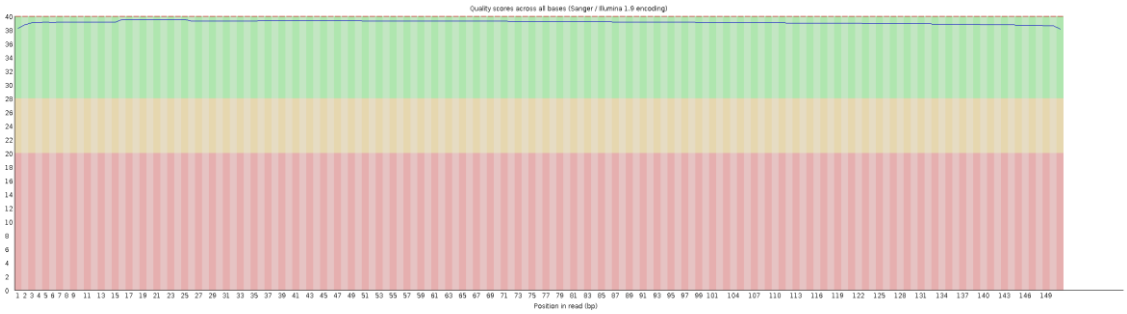


Figure 6. Read quality at each cycle of 1 (read1)

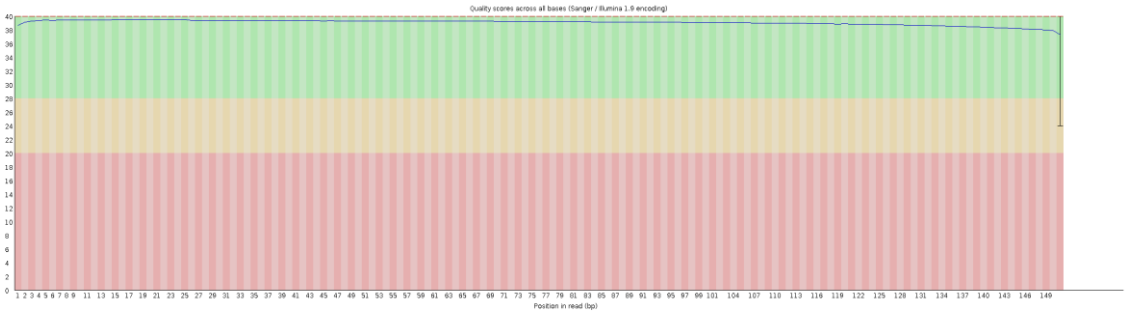


Figure 7. Read quality at each cycle of 1 (read2)

- 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_RNAseq/Analysis_statistics/trim_throughput.txt)

Trimmomatic program is used to remove adapter sequences and bases with base quality lower than three from the ends. Also using sliding window method, bases of reads that does not qualify for window size 4, and mean quality 15 are trimmed. Afterwards, reads with length shorter than 36bp are dropped to produce trimmed data.

Table 2. Trimming Data Stats

Sample id	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
1	7,768,511,481	53,065,742	50.18	99.07	96.66
2	7,981,309,942	54,288,416	49.67	99.02	96.53
3	7,284,037,645	49,670,884	51.26	99.11	96.81
4	7,589,085,641	51,937,398	50.68	99.06	96.63
5	8,527,382,944	59,535,654	51.34	99.09	96.74
6	7,451,550,275	50,770,686	49.77	98.99	96.39
7	8,025,668,722	55,672,502	50.65	99.1	96.78
8	7,862,761,206	53,511,666	49.02	99.01	96.46
9	7,667,876,479	53,270,250	50.32	99.07	96.65
10	7,568,827,930	51,837,752	48.56	99.0	96.45
11	7,941,243,032	54,678,264	49.81	99.05	96.6
12	7,870,809,757	53,830,858	49.87	99.03	96.55
13	7,452,483,545	51,388,952	50.04	99.08	96.68
14	7,579,147,153	51,492,688	49.95	99.11	96.78
15	8,666,711,846	58,977,202	49.49	98.98	96.36
16	8,640,538,780	59,159,310	50.05	99.02	96.51
17	6,758,296,946	46,806,570	50.76	98.99	96.49
18	6,951,434,217	47,312,164	49.69	98.97	96.45
19	8,341,507,451	56,794,880	50.34	99.02	96.51
20	8,414,970,124	57,621,286	49.68	99.0	96.46
21	7,709,658,165	52,382,340	49.44	99.0	96.45
22	8,644,773,862	59,050,882	49.77	99.01	96.48
23	8,719,553,806	59,522,136	48.74	99.02	96.49
24	8,697,691,280	59,252,870	48.5	99.0	96.45
25	8,719,352,716	59,432,412	48.52	98.95	96.29
26	8,568,720,994	58,590,348	49.03	98.97	96.33
27	8,679,673,971	58,916,152	49.09	98.92	96.18
28	8,733,795,722	59,480,508	49.25	99.03	96.53

29	8,424,892,420	58,165,704	49.14	98.99	96.41
30	8,513,066,412	58,500,986	49.42	98.99	96.41
31	8,619,774,779	59,026,554	48.85	99.12	96.85
32	7,548,679,296	51,790,508	48.43	99.1	96.76
33	7,514,510,193	51,237,440	48.67	99.02	96.53
34	7,721,995,200	52,908,658	48.13	98.96	96.31
35	8,573,773,778	59,498,022	49.84	99.05	96.62
36	8,748,757,188	59,535,916	49.23	98.99	96.4
37	7,932,766,270	54,815,696	48.97	99.03	96.54
38	8,820,741,312	59,506,700	48.66	99.0	96.43
39	8,674,528,443	59,151,170	49.11	99.02	96.51
40	8,690,209,304	59,089,874	48.5	99.01	96.5
41	8,731,125,465	58,991,970	48.69	98.92	96.17
42	8,267,513,301	55,812,584	48.21	98.91	96.13
43	8,699,006,236	58,746,162	48.57	98.93	96.19
44	8,756,353,942	59,313,952	48.38	98.92	96.16
45	8,755,446,459	59,444,662	48.36	98.94	96.23
46	8,104,315,994	55,811,466	49.2	99.01	96.5
47	8,710,744,435	59,317,162	48.53	98.99	96.42
48	8,646,452,540	59,242,764	48.48	99.02	96.49
49	8,701,549,161	59,249,598	48.68	98.96	96.31
50	8,723,783,433	59,247,002	48.87	98.95	96.25
51	8,547,968,036	59,485,752	49.67	98.99	96.42
52	8,687,675,822	59,283,300	49.18	99.0	96.42
53	8,565,176,183	59,191,968	49.39	99.04	96.58
54	8,639,289,869	58,879,748	49.19	98.95	96.3

- 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

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

(Refer to Path: result_RNAseq/Analysis_statistics/trimmedData/A_fastqc/)

Figure 8 and 9 show average base quality at each cycle after trimming.

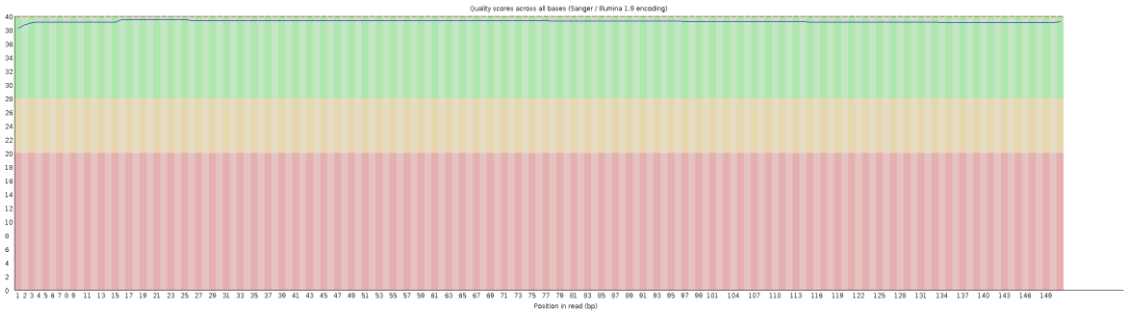


Figure 8. Average base quality of 1 (read1) at each cycle after trimming

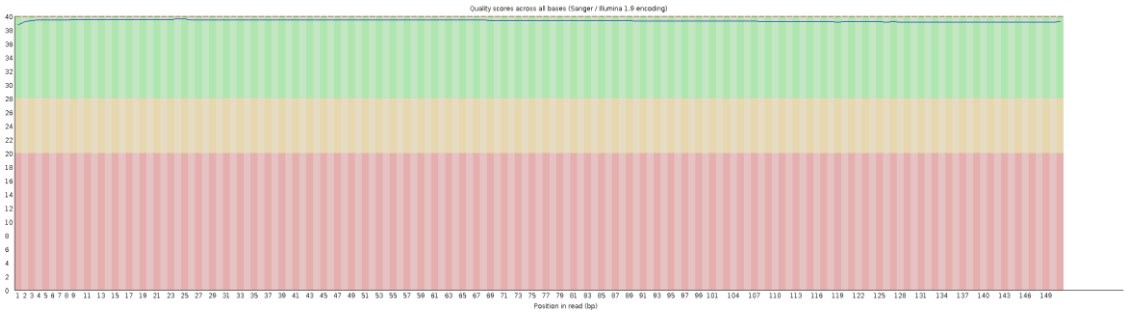


Figure 9. Average base quality of 1 (read2) at each cycle after trimming

- 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

4. Reference Mapping and Assembly Results

4.1. Mapping Data Statistics

(Refer to Path: result_RNAseq/Analysis_statistics/mapping.hisat.stats.txt)

In order to map cDNA fragments obtained from RNA sequencing, GRCm39 was used as a reference genome. Table 3 shows the statistic obtained from HISAT2, which is known to handle spliced read mapping. You can check number of processed reads, mapped reads.

Table 3. Mapped Data Stats

Sample ID	# of processed reads	# of mapped reads (%)	# of unmapped reads (%)
1	53,065,742	51,676,491 (97.38%)	1,389,251 (2.62%)
2	54,288,416	52,940,678 (97.52%)	1,347,738 (2.48%)
3	49,670,884	48,570,989 (97.79%)	1,099,895 (2.21%)
4	51,937,398	50,606,410 (97.44%)	1,330,988 (2.56%)
5	59,535,654	58,009,797 (97.44%)	1,525,857 (2.56%)
6	50,770,686	49,311,208 (97.13%)	1,459,478 (2.87%)
7	55,672,502	54,282,151 (97.5%)	1,390,351 (2.5%)
8	53,511,666	52,152,985 (97.46%)	1,358,681 (2.54%)
9	53,270,250	51,853,671 (97.34%)	1,416,579 (2.66%)
10	51,837,752	50,819,292 (98.04%)	1,018,460 (1.96%)
11	54,678,264	53,224,920 (97.34%)	1,453,344 (2.66%)
12	53,830,858	52,545,682 (97.61%)	1,285,176 (2.39%)
13	51,388,952	50,231,746 (97.75%)	1,157,206 (2.25%)

14	51,492,688	50,176,454 (97.44%)	1,316,234 (2.56%)
15	58,977,202	57,253,657 (97.08%)	1,723,545 (2.92%)
16	59,159,310	57,264,601 (96.8%)	1,894,709 (3.2%)
17	46,806,570	45,539,291 (97.29%)	1,267,279 (2.71%)
18	47,312,164	46,010,705 (97.25%)	1,301,459 (2.75%)
19	56,794,880	55,335,589 (97.43%)	1,459,291 (2.57%)
20	57,621,286	55,905,730 (97.02%)	1,715,556 (2.98%)
21	52,382,340	50,759,508 (96.9%)	1,622,832 (3.1%)
22	59,050,882	57,521,599 (97.41%)	1,529,283 (2.59%)
23	59,522,136	58,034,378 (97.5%)	1,487,758 (2.5%)
24	59,252,870	57,598,704 (97.21%)	1,654,166 (2.79%)
25	59,432,412	57,973,967 (97.55%)	1,458,445 (2.45%)
26	58,590,348	56,824,209 (96.99%)	1,766,139 (3.01%)
27	58,916,152	57,399,737 (97.43%)	1,516,415 (2.57%)
28	59,480,508	58,152,840 (97.77%)	1,327,668 (2.23%)
29	58,165,704	57,013,264 (98.02%)	1,152,440 (1.98%)
30	58,500,986	56,882,398 (97.23%)	1,618,588 (2.77%)
31	59,026,554	57,177,411 (96.87%)	1,849,143 (3.13%)
32	51,790,508	50,292,918 (97.11%)	1,497,590 (2.89%)

33	51,237,440	49,669,472 (96.94%)	1,567,968 (3.06%)
34	52,908,658	51,260,092 (96.88%)	1,648,566 (3.12%)
35	59,498,022	57,736,413 (97.04%)	1,761,609 (2.96%)
36	59,535,916	57,794,810 (97.08%)	1,741,106 (2.92%)
37	54,815,696	53,291,425 (97.22%)	1,524,271 (2.78%)
38	59,506,700	57,870,369 (97.25%)	1,636,331 (2.75%)
39	59,151,170	57,357,569 (96.97%)	1,793,601 (3.03%)
40	59,089,874	57,359,632 (97.07%)	1,730,242 (2.93%)
41	58,991,970	57,314,638 (97.16%)	1,677,332 (2.84%)
42	55,812,584	54,904,785 (98.37%)	907,799 (1.63%)
43	58,746,162	57,362,454 (97.64%)	1,383,708 (2.36%)
44	59,313,952	57,853,302 (97.54%)	1,460,650 (2.46%)
45	59,444,662	57,807,078 (97.25%)	1,637,584 (2.75%)
46	55,811,466	54,420,333 (97.51%)	1,391,133 (2.49%)
47	59,317,162	57,300,402 (96.6%)	2,016,760 (3.4%)
48	59,242,764	57,181,717 (96.52%)	2,061,047 (3.48%)
49	59,249,598	57,481,523 (97.02%)	1,768,075 (2.98%)
50	59,247,002	57,828,135 (97.61%)	1,418,867 (2.39%)
51	59,485,752	57,610,822 (96.85%)	1,874,930 (3.15%)

52	59,283,300	57,632,386 (97.22%)	1,650,914 (2.78%)
53	59,191,968	57,399,911 (96.97%)	1,792,057 (3.03%)
54	58,879,748	56,287,012 (95.6%)	2,592,736 (4.4%)

- Processed reads: Number of cleaned reads after trimming
- Mapped reads: Number of reads mapped to reference
- Unmapped reads: Number of reads that failed to align

4. 2. Expression Profiling

Known genes and transcripts are assembled with StringTie based on reference genome model.

After assembly, the abundance of gene/transcript is calculated in the read count and normalized values as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Kilobase Million) for a sample.

4. 2. 1. Known Transcripts Expression Level

(Refer to Path: result_RNAseq/Expression_profile/StringTie/ Expression_Profile.GRCm39.transcript.xlsx)

Table 4 is an example of known transcript expression level per sample in expression value. This result is obtained by -e option of StringTie does not consider novel transcript assembly.

Table 4. Known transcripts Expression Level (example)

Transcript_ID	Gene_ID	Gene Symbol	Description	Transcript_Locus	Trancript Length	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM	AM_TPM	BM_TPM
NM_130786	1	A1BG	alpha-1-B glycoprotein	chr19:58345183-58353492	3382	88	163	0.432396	0.678319	0.947053	1.504474
NR_040112	3	A2MP1	alpha-2-macroglobulin pseudog	chr12:9228533-9234207	1201	0	0	0	0	0	0
XM_017013947	9	NAT1	N-acetyltransferase 1, transcrip	chr8:18170419-18223689	2704	0	21	0	0.108737	0	0.241173
NM_001291962	9	NAT1	N-acetyltransferase 1, transcrip	chr8:18170467-18223689	2122	0	0	0	0	0	0
NM_000015	10	NAT2	N-acetyltransferase 2	chr8:18391282-18401218	1285	0	0	0	0	0	0
NM_001085	12	SERPINA3	serpin family A member 3	chr14:94612377-94624053	1590	8	75	0.084216	0.664787	0.184454	1.474461
XM_005247104	13	AADAC	arylacetamide deacetylase, trar	chr3:151814008-151828488	1620	0	12	0	0.102866	0	0.228152
NM_001086	13	AADAC	arylacetamide deacetylase	chr3:151814116-151828488	1563	108	108	1.152579	0.971041	2.524427	2.153715
XM_024452712	14	AAMP	angio associated migratory cell	chr2:218264127-218270181	2002	106	101	0.879142	0.710738	1.925533	1.576378
NM_001302545	14	AAMP	angio associated migratory cell	chr2:218264129-218270137	1763	1621	1797	15.408498	14.424821	33.74835	31.99344
NM_001087	14	AAMP	angio associated migratory cell	chr2:218264129-218270137	1760	9332	10212	88.854179	82.119453	194.6122	182.1363
NM_001166579	15	AANAT	aralkylamine N-acetyltransferas	chr17:76453351-76470117	1913	2	8	0.010678	0.052728	0.023387	0.116948
XM_017024259	15	AANAT	aralkylamine N-acetyltransferas	chr17:76465946-76470797	4252	4	11	0.013221	0.03452	0.028958	0.076564
NR_110548	15	AANAT	aralkylamine N-acetyltransferas	chr17:76467548-76470117	1082	0	0	0	0	0	0
NM_001088	15	AANAT	aralkylamine N-acetyltransferas	chr17:76467603-76470117	971	0	0	0	0	0	0
XR_933220	16	AARS	alanyl-tRNA synthetase, transc	chr16:70252295-70289509	3258	90	160	0.461517	0.694592	1.010834	1.540566
NM_001605	16	AARS	alanyl-tRNA synthetase	chr16:70252394-70289509	3344	22367	68204	112.089745	288.669189	245.5037	640.2521

- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_ID: Gene ID
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- Transcript_Locus: Transript locus
- Transcript_Length: Transcript length
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM normalized value of a sample
- [Sample Name]_TPM: TPM normalized value of a sample

4. 2. 2. Known Genes Expression Level

(Refer to Path: result_RNAseq/Expression_profile/StringTie/ Expression_Profile.GRCm39.gene.xlsx)

Table 5 is an example of known gene expression level per sample in expression value. This result is obtained by -e option of StringTie does not consider novel transcript assembly.

Table 5. Known genes Expression Level (example)

Gene_ID	Transcript_ID	Gene Symbol	Description	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM	AM_TPM	BM_TPM
1	NM_130786	A1BG	alpha-1-B glycoprotein	88	163	0.432396	0.678319	0.947053	1.504474
2	NM_000014,NM_001347423,	A2M	alpha-2-macroglobulin	0	0	0	0	0	0
3	NR_040112	A2MP1	alpha-2-macroglobulin pseudogene	0	0	0	0	0	0
9	NM_000662,NM_001160170,	NAT1	N-acetyltransferase 1	288	217	2.411185	1.490984	5.281078	3.306918
10	NM_000015,XM_017012938	NAT2	N-acetyltransferase 2	10	6	0.097138	0.050729	0.212756	0.112513
12	NM_001085	SERPINA3	serpin family A member 3	8	75	0.084216	0.664787	0.184454	1.474461
13	NM_001086,XM_005247104	AADAC	arylacetamide deacetylase	108	120	1.152579	1.073907	2.524427	2.381867
14	NM_001087,NM_001302545,	AAMP	angio associated migratory cell prot	11059	12110	105.141819	97.255012	230.2861	215.7062
15	NM_001088,NM_001166579,	AANAT	aralkylamine N-acetyltransferase	6	19	0.023899	0.087248	0.052345	0.193512
16	NM_001605,XR_933220	AARS	alanyl-tRNA synthetase	22457	68364	112.551262	289.363781	246.5145	641.7927
18	NM_000663,NM_001127448,	ABAT	4-aminobutyrate aminotransferase	327	175	1.143824	0.441216	2.505251	0.978593
19	NM_005502,XM_005251773,	ABCA1	ATP binding cassette subfamily A n	1496	2718	2.403716	3.695532	5.264719	8.196482
20	NM_001606,NM_212533,XM	ABCA2	ATP binding cassette subfamily A n	2500	3986	5.218521	6.986245	11.42982	15.4951
21	NM_001089	ABCA3	ATP binding cassette subfamily A n	2214	4876	5.619098	10.452255	12.30719	23.18251
22	NM_001271696,NM_001271	ABCB7	ATP binding cassette subfamily B n	2618	1974	9.550061	6.097788	20.91695	13.52455
23	NM_001025091,NM_001090	ABCF1	ATP binding cassette subfamily F n	11449	11921	56.366045	49.563715	123.4553	109.9295
24	NM_000350	ABCA4	ATP binding cassette subfamily A n	62	139	0.140036	0.267738	0.306712	0.593827

- Gene_ID: Gene ID
- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM normalized value of a sample

5. Data Download Information

5.1. Raw Data

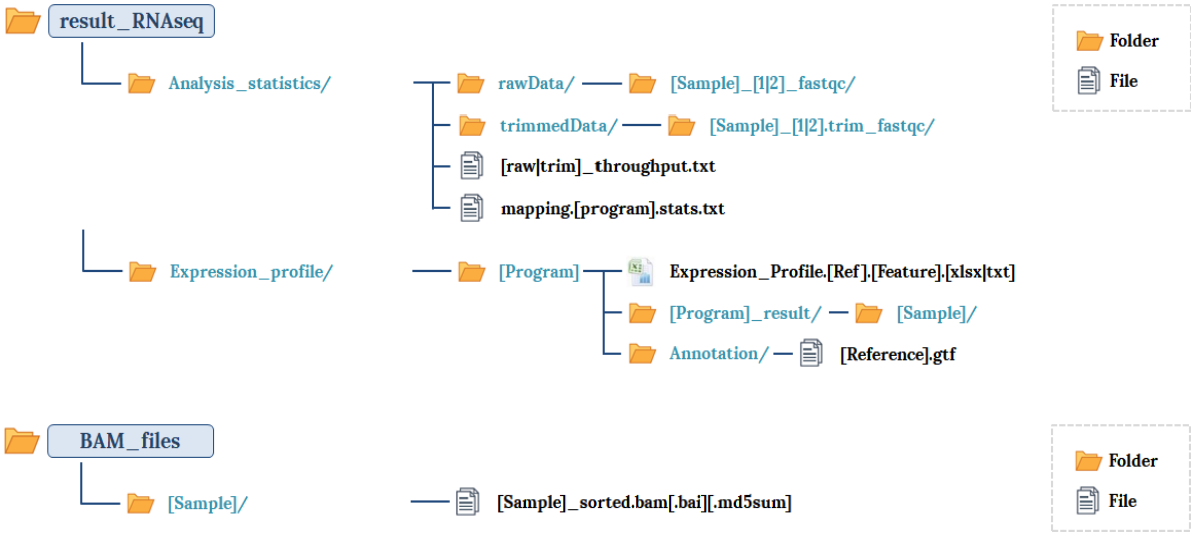
Raw data is the FASTQ file that isn't trimmed adapter sequence.

File name	File size	md5sum
1_1.fastq.gz	1.77G	
1_2.fastq.gz	1.77G	
2_1.fastq.gz	1.81G	
2_2.fastq.gz	1.81G	
3_1.fastq.gz	1.66G	
3_2.fastq.gz	1.65G	
4_1.fastq.gz	1.73G	
4_2.fastq.gz	1.73G	
5_1.fastq.gz	1.99G	
5_2.fastq.gz	1.98G	
6_1.fastq.gz	1.71G	
6_2.fastq.gz	1.72G	
7_1.fastq.gz	1.84G	
7_2.fastq.gz	1.85G	
8_1.fastq.gz	1.8G	
8_2.fastq.gz	1.8G	
9_1.fastq.gz	1.78G	
9_2.fastq.gz	1.79G	
10_1.fastq.gz	1.74G	
10_2.fastq.gz	1.73G	
11_1.fastq.gz	1.83G	
11_2.fastq.gz	1.83G	
12_1.fastq.gz	1.81G	
12_2.fastq.gz	1.8G	
13_1.fastq.gz	1.71G	
13_2.fastq.gz	1.72G	
14_1.fastq.gz	1.72G	
14_2.fastq.gz	1.71G	
15_1.fastq.gz	1.75G	
15_2.fastq.gz	1.75G	
16_1.fastq.gz	1.75G	
16_2.fastq.gz	1.75G	

17_1.fastq.gz	1.58G
17_2.fastq.gz	1.56G
18_1.fastq.gz	1.57G
18_2.fastq.gz	1.57G
19_1.fastq.gz	1.92G
19_2.fastq.gz	1.9G
20_1.fastq.gz	1.93G
20_2.fastq.gz	1.93G
21_1.fastq.gz	1.76G
21_2.fastq.gz	1.76G
22_1.fastq.gz	1.74G
22_2.fastq.gz	1.74G
23_1.fastq.gz	1.76G
23_2.fastq.gz	1.76G
24_1.fastq.gz	1.73G
24_2.fastq.gz	1.72G
25_1.fastq.gz	1.72G
25_2.fastq.gz	1.74G
26_1.fastq.gz	1.97G
26_2.fastq.gz	1.97G
27_1.fastq.gz	1.76G
27_2.fastq.gz	1.76G
28_1.fastq.gz	1.73G
28_2.fastq.gz	1.74G
29_1.fastq.gz	1.95G
29_2.fastq.gz	1.95G
30_1.fastq.gz	1.98G
30_2.fastq.gz	1.97G
31_1.fastq.gz	1.72G
31_2.fastq.gz	1.71G
32_1.fastq.gz	1.72G
32_2.fastq.gz	1.72G
33_1.fastq.gz	1.71G
33_2.fastq.gz	1.7G
34_1.fastq.gz	1.78G
34_2.fastq.gz	1.77G
35_1.fastq.gz	1.77G
35_2.fastq.gz	1.76G

36_1.fastq.gz	1.74G
36_2.fastq.gz	1.74G
37_1.fastq.gz	1.83G
37_2.fastq.gz	1.84G
38_1.fastq.gz	1.73G
38_2.fastq.gz	1.73G
39_1.fastq.gz	1.73G
39_2.fastq.gz	1.74G
40_1.fastq.gz	1.98G
40_2.fastq.gz	1.97G
41_1.fastq.gz	1.75G
41_2.fastq.gz	1.75G
42_1.fastq.gz	1.87G
42_2.fastq.gz	1.89G
43_1.fastq.gz	1.73G
43_2.fastq.gz	1.75G
44_1.fastq.gz	1.75G
44_2.fastq.gz	1.74G
45_1.fastq.gz	1.74G
45_2.fastq.gz	1.74G
46_1.fastq.gz	1.87G
46_2.fastq.gz	1.87G
47_1.fastq.gz	1.75G
47_2.fastq.gz	1.74G
48_1.fastq.gz	1.74G
48_2.fastq.gz	1.73G
49_1.fastq.gz	1.73G
49_2.fastq.gz	1.72G
50_1.fastq.gz	1.74G
50_2.fastq.gz	1.74G
51_1.fastq.gz	1.74G
51_2.fastq.gz	1.75G
52_1.fastq.gz	1.75G
52_2.fastq.gz	1.74G
53_1.fastq.gz	1.74G
53_2.fastq.gz	1.74G
54_1.fastq.gz	1.72G
54_2.fastq.gz	1.72G

- 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.



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

6. Appendix

6.1. 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.

6.2. Programs used in Analysis

6. 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.

6. 2. 2. Trimmomatic 0.38

LINK http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic is a program that performs trimming depending on various parameters on illumina paired-end or single-end.

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality.
- TRAILING: Cut bases off the end of a read, if below a threshold quality.
- CROP: Cut the read to a specified length.
- HEADCROP: Cut the specified number of bases from the start of the read.
- MINLEN: Drop the read if it is below a specified length.
- TOPHRED33: Change quality score to phred33.
- TOPHRED64: Change quality score to phred64.

6. 2. 3. HISAT2 version 2.1.0

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).

6. 2. 4. StringTie version 2.1.3b

LINK https://ccb.jhu.edu/software/stringtie/

StringTie is a fast and highly efficient assembler of RNA-Seq alignments into potential transcripts. It uses a novel network flow algorithm as well as an optional de novo assembly step to assemble and quantitate full-length transcripts representing multiple splice variants for each gene locus.

6. 3. References

1. BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, btu170.
2. KIM, Daehwan; LANGMEAD, Ben; SALZBERG, Steven L. HISAT: a fast spliced aligner with low memory requirements. Nature methods, 2015, 12.4: 357-360.
3. LI, Heng, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16: 2078-2079.
4. PERTEA, Mihaela, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology, 2015, 33.3: 290-295.
5. PERTEA, Mihaela, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols, 2016, 11.9: 1650-1667.

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 57, 1105 BA, 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

16F Time24 Building, 2-4-32 Aomi, Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@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