

Homo Sapiens Transcriptome Sequencing Report

February 2019



Project Information

Client Name	Paola Vera-Licona
Company/Institution	University of Connecticut Health Center
Order Number	1901UQHS-0097
Species	<i>Homo Sapiens</i>
Reference	hg19
Annotation	RefSeq_2017_06_12
Read Length	50
Number of Samples	8

Project Results Summary

In this study, *Homo Sapiens* whole transcriptome sequencing was performed in order to examine the different gene expression profiles, and to perform gene annotation on set of useful genes based on gene ontology pathway information.

Analyses were successfully performed on all 8 single-end 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.

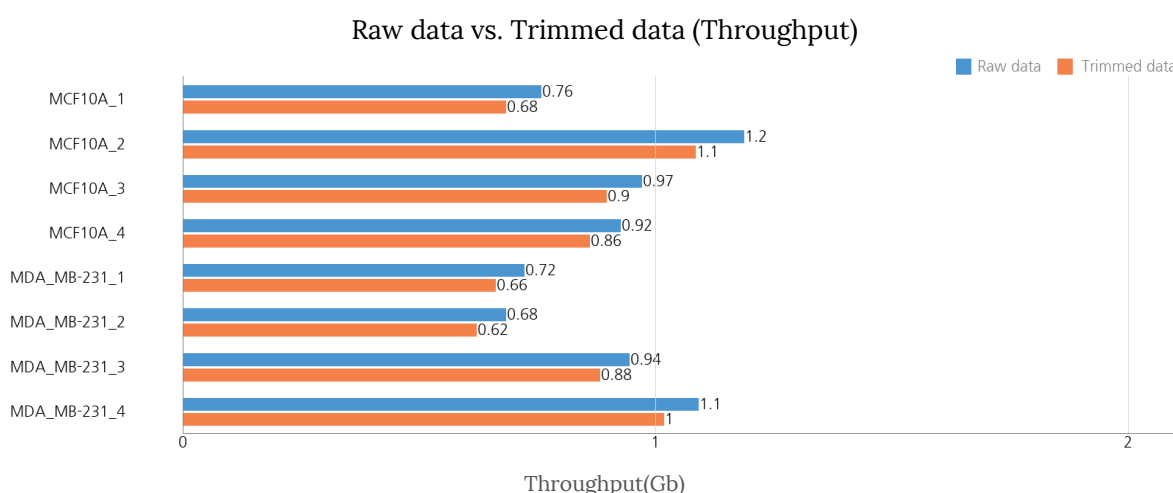


Figure 1. Throughput output of Raw and Trimmed data

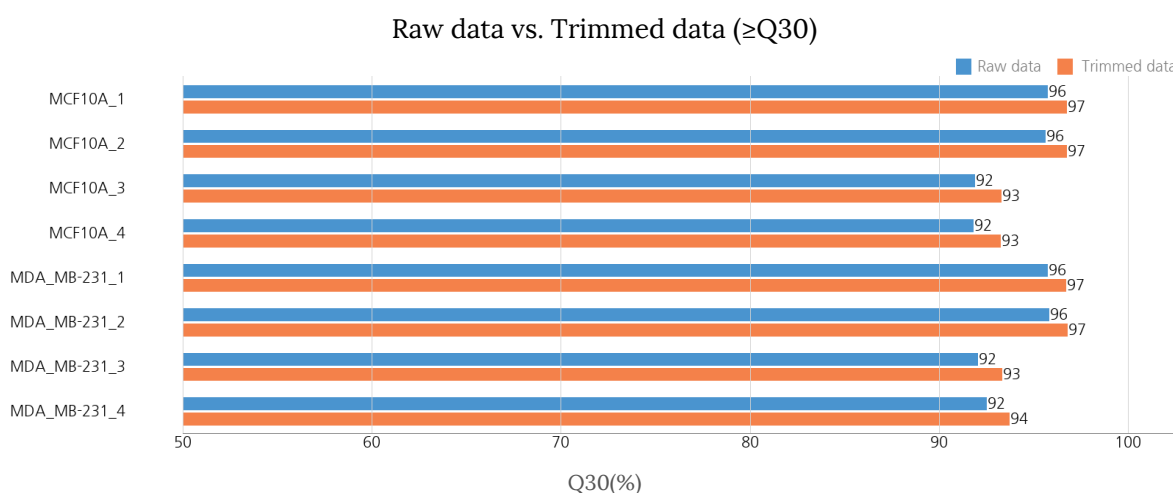


Figure 2. Q30 score of Raw and Trimmed data

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.

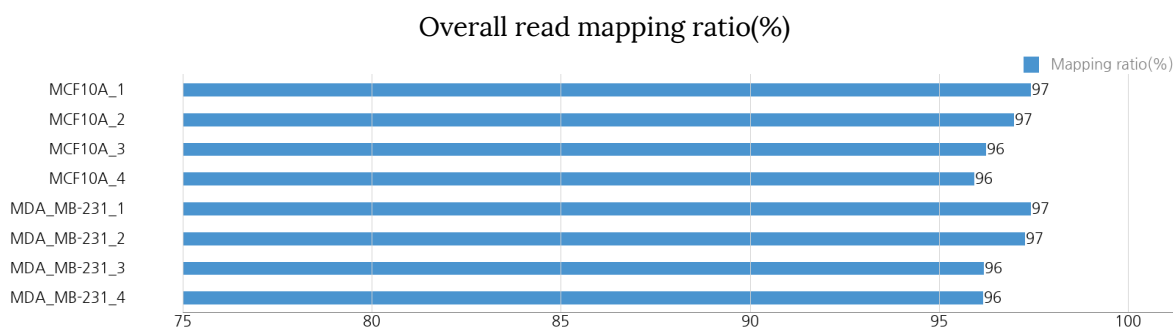


Figure 3. Overall read mapping ratio(%)

After the read mapping, Stringtie was used for transcript assembly. Expression profile was calculated for each sample and transcript/gene as read count and RPKM (Read per Kilobase per Million mapped reads).

DEG (Differentially Expressed Genes) analysis was performed on a comparison pair (MDA_MB-231_vs_MCF10A) as requested using DESeq2. The results showed 3,563 genes which satisfied $|fc| \geq 2$ & $nbinomWaldTest$ raw p -value < 0.05 conditions in comparison pair.

Figure 4 shows the result of hierarchical clustering (distance metric= Euclidean distance, linkage method= complete) analysis. It graphically represents the similarity of expression patterns between samples and genes.

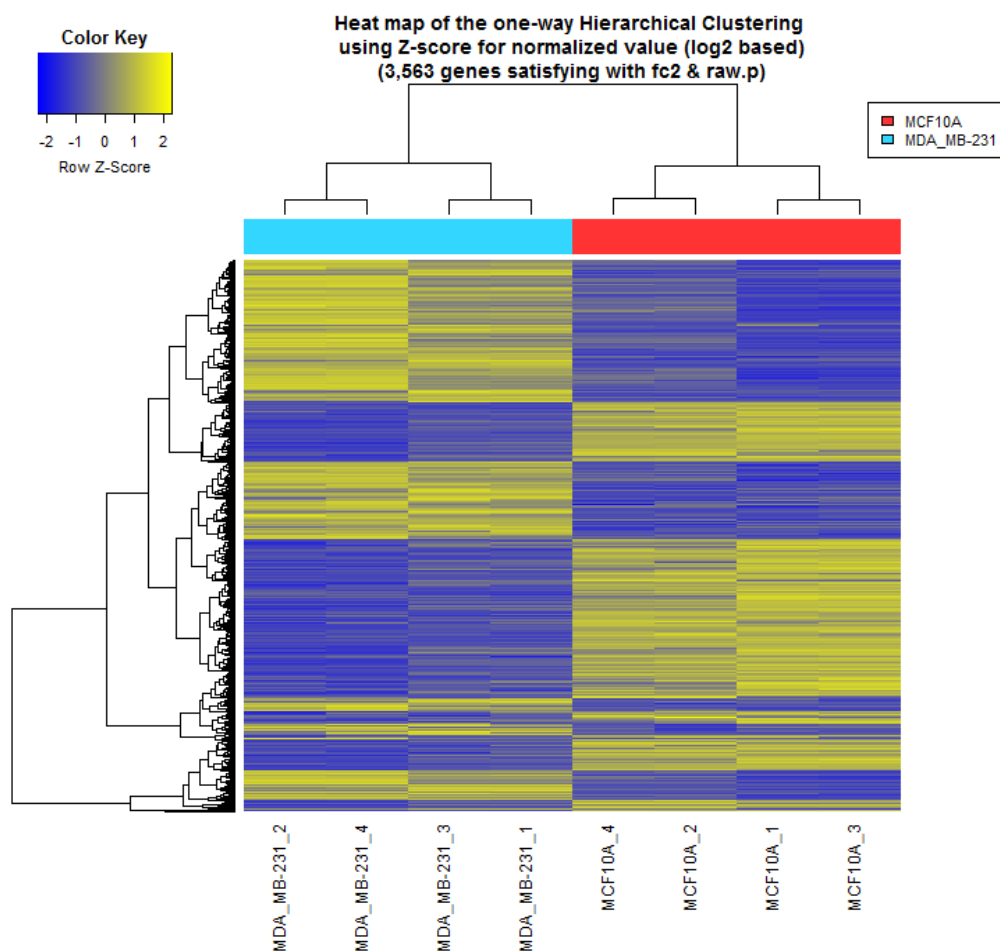


Figure 4. Heatmap for DEG list

Table of Contents

Project Information	02
Project Results Summary	03
1. Experimental Methods and Workflow	07
2. Analysis Methods and Workflow	08
3. Summary of Data Production	09
3. 1. Raw Data Statistics	09
3. 2. Average Base Quality at Each Cycle	10
3. 3. Trimming Data Statistics	11
3. 4. Average Base Quality at Each Cycle after Trimming	12
4. Reference Mapping and Assembly Results	13
4. 1. Mapping Data Statistics	13
4. 2. Expression Profiling	14
5. Differentially Expressed Gene Analysis Results	16
5. 1. Data Analysis Quality Check and Preprocessing	16
5. 2. Differentially Expressed Gene Analysis Workflow	21
5. 3. Significant Gene Results	22
6. Data Download Information	25
6. 1. Raw Data	25
6. 2. Analysis Results	25
7. Appendix	28
7. 1. Phred Quality Score Chart	28
7. 2. Programs used in Analysis	29
7. 3. References	30

1. Experimental Methods and Workflow

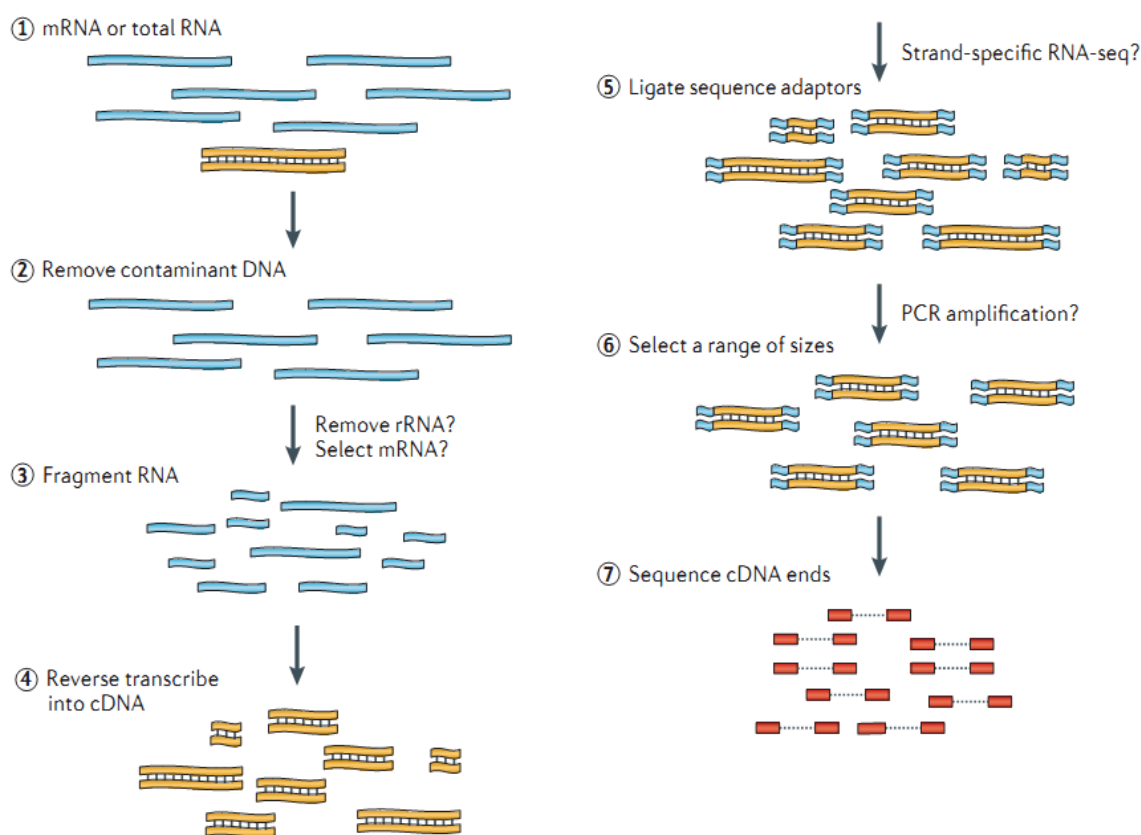


Figure 5. RNA Sequencing Experiment Workflow

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 6. 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 value which is based on transcript length and depth of coverage. The FPKM (Fragments Per Kilobase of transcript per Million Mapped reads) value or the RPKM (Reads Per Kilobase of transcript per Million mapped reads) is used as a normalization value.
- 6) In groups with different conditions, genes or transcripts that express differentially are filtered out through statistical hypothesis testing.

3. Summary of Data Production

3.1. Raw Data Statistics

(Refer to Path: result_RNAseq/Analysis_statistics/rawData/raw_throughput.stats)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 8 samples. For example, in MCF10A_1, 15,132,901 reads are produced, and total read bases are 756.6Mbp. The GC content (%) is 46.75% and Q30 is 95.71%.

Table 1. Raw data stats

Index	Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
1	MCF10A_1	756,645,050	15,132,901	46.75	98.35	95.71
2	MCF10A_2	1,185,887,300	23,717,746	46.44	98.22	95.61
3	MCF10A_3	969,849,100	19,396,982	46.89	96.77	91.87
4	MCF10A_4	924,635,400	18,492,708	46.64	96.66	91.78
5	MDA_MB-231_1	720,995,350	14,419,907	46.88	98.38	95.74
6	MDA_MB-231_2	683,251,800	13,665,036	45.97	98.38	95.80
7	MDA_MB-231_3	943,351,800	18,867,036	47.03	96.86	92.02
8	MDA_MB-231_4	1,090,284,700	21,805,694	46.15	97.03	92.49

(* 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>

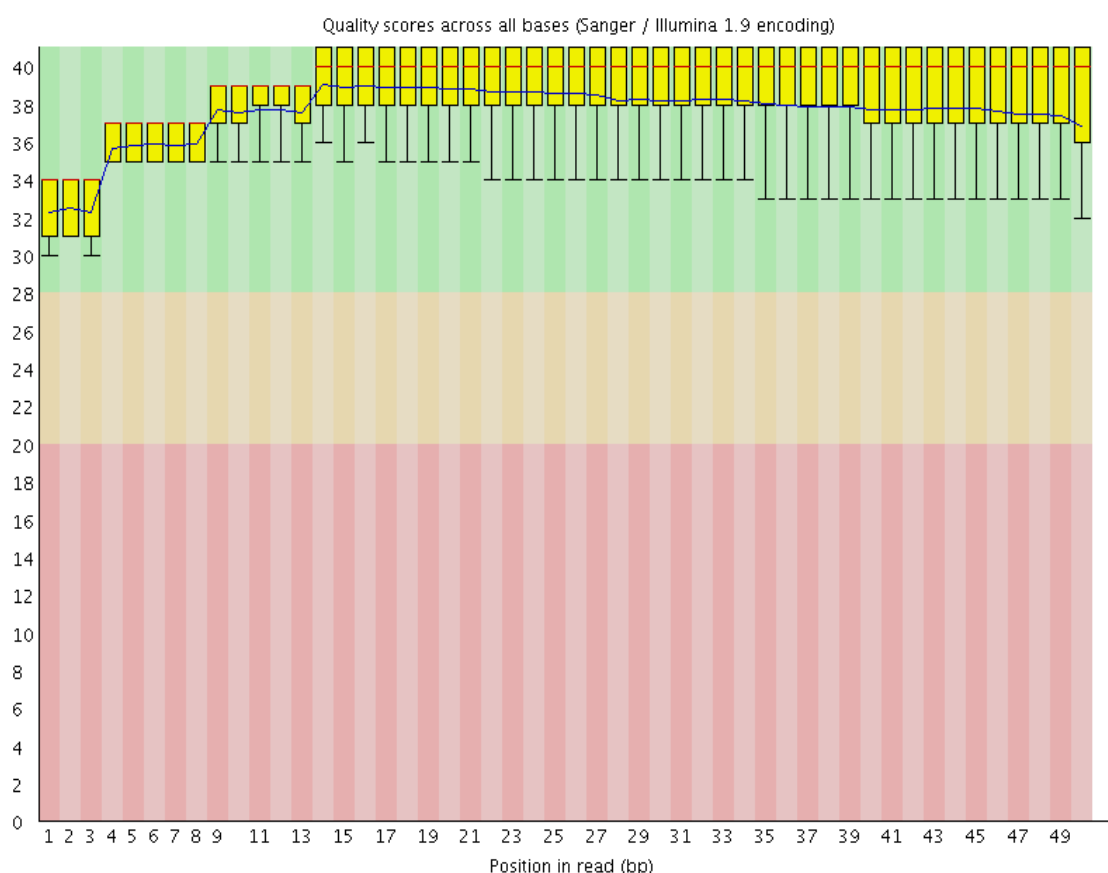


Figure 7. Read quality at each cycle of MCF10A_1 (read1)

- 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_RNAseq/Analysis_statistics/trimmedData/trim_throughput.stats)

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

Index	Sample id	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
1	MCF10A_1	683,322,164	13,936,472	47.38	99.16	96.74
2	MCF10A_2	1,084,330,105	22,115,734	47.37	99.15	96.74
3	MCF10A_3	895,112,653	17,974,082	46.76	97.86	93.27
4	MCF10A_4	860,717,290	17,286,171	46.76	97.86	93.25
5	MDA_MB-231_1	661,259,870	13,488,510	47.58	99.15	96.71
6	MDA_MB-231_2	620,838,689	12,661,780	46.70	99.17	96.79
7	MDA_MB-231_3	882,413,118	17,718,723	46.94	97.88	93.31
8	MDA_MB-231_4	1,016,690,938	20,408,169	46.14	98.00	93.71

- 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.

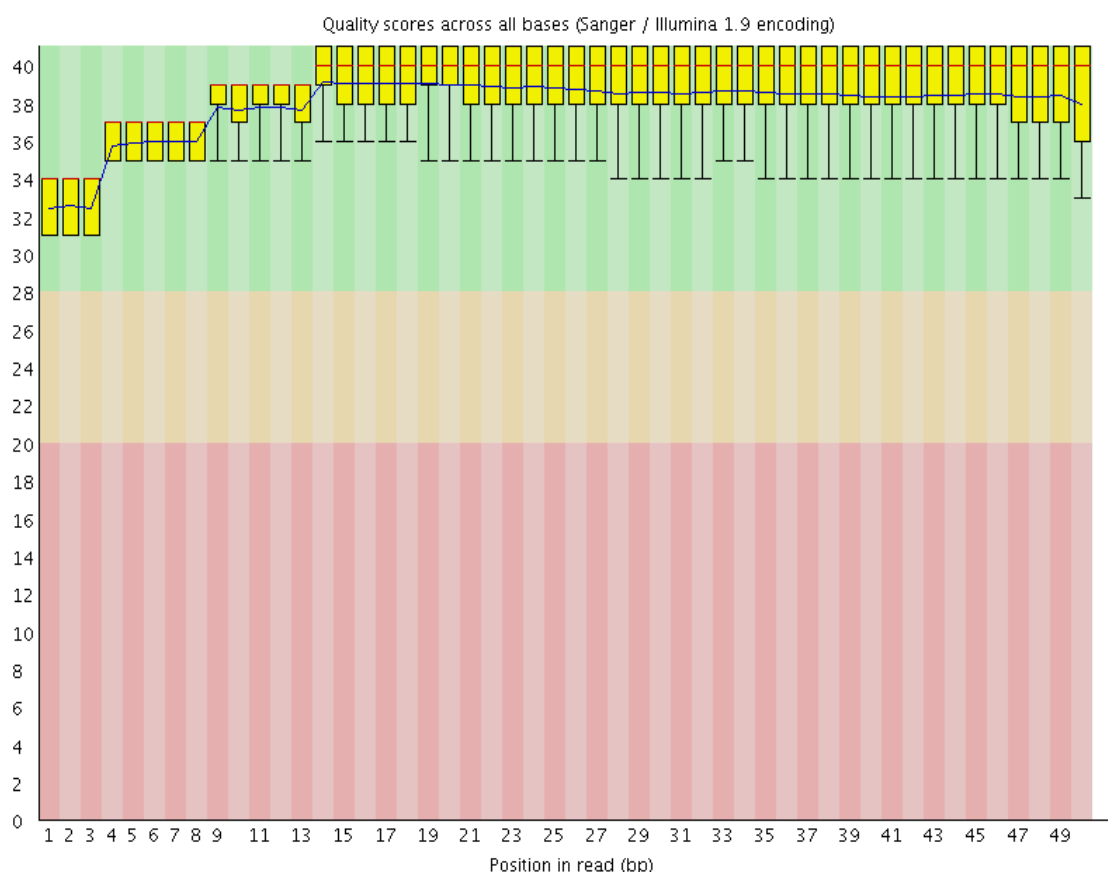


Figure 8. Average base quality of MCF10A_1 (read1) at each cycle after trimming

- 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

4. Reference Mapping and Assembly Results

4.1. Mapping Data Statistics

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

In order to map cDNA fragments obtained from RNA sequencing, hg19 was used as a reference genome. Table 3 shows the statistic obtained from HISAT2, which is known to handle spliced read mapping through Bowtie2 aligner. 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 (%)
MCF10A_1	13,936,472	13,575,777 (97.41%)	360,695 (2.59%)
MCF10A_2	22,115,734	21,444,952 (96.97%)	670,782 (3.03%)
MCF10A_3	17,974,082	17,297,085 (96.23%)	676,997 (3.77%)
MCF10A_4	17,286,171	16,580,659 (95.92%)	705,512 (4.08%)
MDA_MB-231_1	13,488,510	13,140,214 (97.42%)	348,296 (2.58%)
MDA_MB-231_2	12,661,780	12,313,048 (97.25%)	348,732 (2.75%)
MDA_MB-231_3	17,718,723	17,042,364 (96.18%)	676,359 (3.82%)
MDA_MB-231_4	20,408,169	19,622,226 (96.15%)	785,943 (3.85%)

- 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 value as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) for a sample.

4. 2. 1. Known Transcripts Expression Level

(Refer to Path: result_RNAseq_excel/Expression_profile/StringTie/Expression_Profile.hg19.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	Transcript_Length	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM
NM_001302545	14	AAMP	angio associated migratory cell protei	chr2:219128852-219134	1835	898	987	12.220251	12.415353
NM_001087	14	AAMP	angio associated migratory cell protei	chr2:219128852-219134	1832	4678	6437	63.774269	81.140015
NM_001166579	15	AANAT	aralkylamine N-acetyltransferase, tra	chr17:74449433-744661	1913	46	30	0.599741	0.352587
NR_110548	15	AANAT	aralkylamine N-acetyltransferase, tra	chr17:74463630-744661	1082	9	9	0.192813	0.186779
NM_001101	60	ACTB	actin beta	chr7:5566779-5570232	1812	93591	129901	1290.007935	1655.640503
NM_001161572	23764	MAFF	MAF bZIP transcription factor F, trans	chr22:38597939-386125	2465	1	150	0.002107	1.397431
NM_012323	23764	MAFF	MAF bZIP transcription factor F, trans	chr22:38597939-386125	2439	1682	2109	17.222849	19.96483
NM_001161574	23764	MAFF	MAF bZIP transcription factor F, trans	chr22:38597939-386125	2372	0	0	0	0
NM_001161573	23764	MAFF	MAF bZIP transcription factor F, trans	chr22:38599027-386125	2223	44	25	0.485203	0.252227
NM_001289905	23765	IL17RA	interleukin 17 receptor A, transcript v	chr22:17565849-175965	8506	1303	975	3.825815	2.644646
NM_014339	23765	IL17RA	interleukin 17 receptor A, transcript v	chr22:17565849-175965	8608	3241	1998	9.402107	5.359576
NR_028287	23766	GABARAPL3	GABA type A receptor associated pr	chr15:90889763-908926	1885	3	6	0.036076	0.073511
NM_001017526	23779	ARHGAP8	Rho GTPase activating protein 8, tra	chr22:45148438-452586	1725	460	641	6.645803	8.576918
NM_181335	23779	ARHGAP8	Rho GTPase activating protein 8, tra	chr22:45148438-452586	1632	1979	2405	30.27355	34.027134
NM_001198726	23779	ARHGAP8	Rho GTPase activating protein 8, tra	chr22:45148438-452586	1528	84	59	1.366953	0.889118
NM_030882	23780	APOL2	apolipoprotein L2, transcript variant a	chr22:36622255-366356	2545	559	1155	5.482551	10.474212
NM_145637	23780	APOL2	apolipoprotein L2, transcript variant b	chr22:36622255-366360	2686	1212	0	11.260728	0

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

4. 2. 2. Known Genes Expression Level

(Refer to Path: result_RNAseq_excel/Expression_profile/StringTie/
Expression_Profile.hg19.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
60	NM_001101	ACTB	actin beta	93591	129901	1290.007935	1655.640503
70	NM_005159	ACTC1	actin, alpha, cardiac muscle 1	20	6	0.1339	0.031949
175	NM_000027, NM_001171988, NR_001171988	AGA	aspartylglucosaminidase	252	279	2.995219	3.071083
176	NM_001135, NM_013227	ACAN	aggrecan	8	0	0.022519	0
177	NM_001136, NM_001206929, NM_001206929	AGER	advanced glycosylation end-product specific	3332	3124	51.224842	44.355004
178	NM_000028, NM_000642, NM_000642	AGL	amylase, alpha-1, 6-glucosidase, 4-alpha-gluc	4919	3679	16.662192	11.52329
191	NM_000687, NM_001161766, NM_001161766	AHCY	adenosylhomocysteinase	12053	13891	129.59984	138.005572
245	NR_002710, NR_120453	ALOX12P2	arachidonate 12-lipoxygenase pseudogene	8	5	0.070872	0.041258
246	NM_001140	ALOX15	arachidonate 15-lipoxygenase	785	710	7.302354	6.108678
247	NM_001039130, NM_001039131, NM_001039131	ALOX15B	arachidonate 15-lipoxygenase, type B	6	0	0.049592	0
248	NM_001631	ALPI	alkaline phosphatase, intestinal	13	3	0.098671	0.021092
249	NM_000478, NM_001127501, NM_001127501	ALPL	alkaline phosphatase, liver/bone/kidney	9	19	0.085416	0.164094
250	NM_001632	ALPP	alkaline phosphatase, placental	464	142	3.894943	1.098701
251	NM_031313	ALPPL2	alkaline phosphatase, placental like 2	88	12	0.876858	0.106491
257	NM_006492	ALX3	ALX homeobox 3	310	319	5.229297	4.975804
258	NM_016519	AMBN	ameloblastin	0	0	0	0
259	NM_001633	AMBP	alpha-1-microglobulin/bikunin precursor	0	0	0	0

- 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. Differentially Expressed Gene Analysis Results

5.1. Data Analysis Quality Check and Preprocessing

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

(Refer to Path: result_RNAseq_excel/DEG_result/Analysis_Result.html)

5.1.1. Sample Information and Analysis Design

Total of 8 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	MCF10A_1	MCF10A
2	MCF10A_2	MCF10A
3	MCF10A_3	MCF10A
4	MCF10A_4	MCF10A
5	MDA_MB-231_1	MDA_MB-231
6	MDA_MB-231_2	MDA_MB-231
7	MDA_MB-231_3	MDA_MB-231
8	MDA_MB-231_4	MDA_MB-231

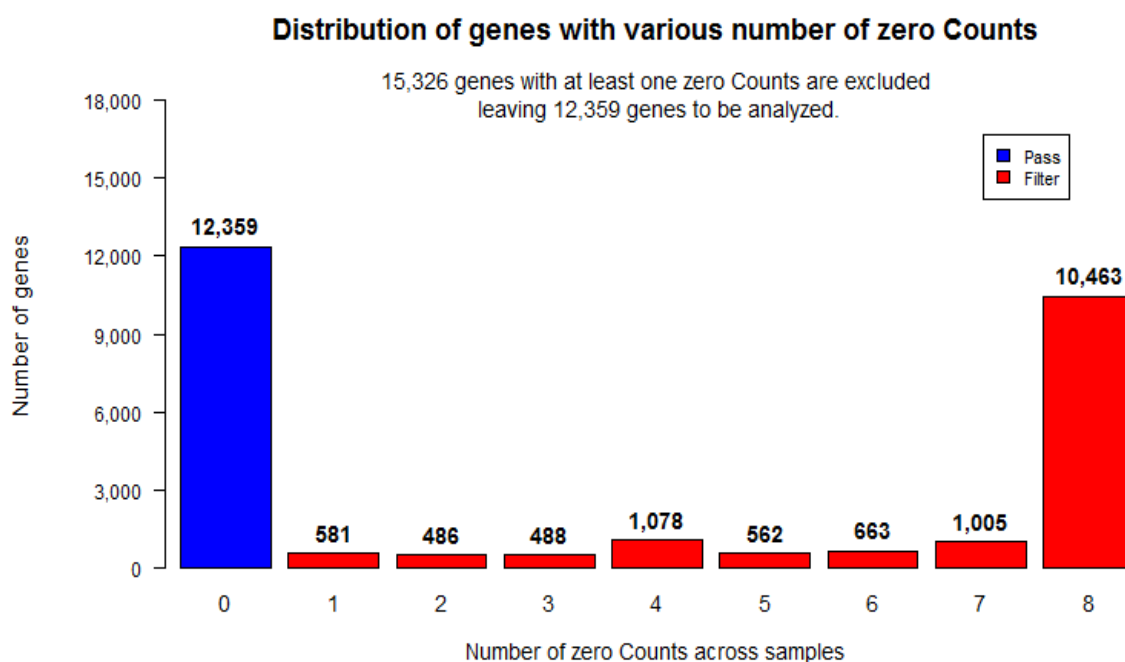
Comparison pair and statistical method for each pair are shown below.

Index	Test vs. Control	Statistical Method
1	MDA_MB-231 vs. MCF10A	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering

5. 1. 2. DATA Quality Check

(Refer to Path: result_RNAseq_excel/DEG_result/Data Quality Check/)

For 8 samples, if more than one read count value was 0, it was not included in the analysis. Therefore, from total of 27,685 genes, 15,326 were excluded and only 12,359 genes were used for statistic analysis.



5. 1. 3. Data Transformation and Normalization

In order to reduce systematic bias, estimates the size factors from the count data and applies Relative Log Expression (RLE) normalization with DESeq2 R library.

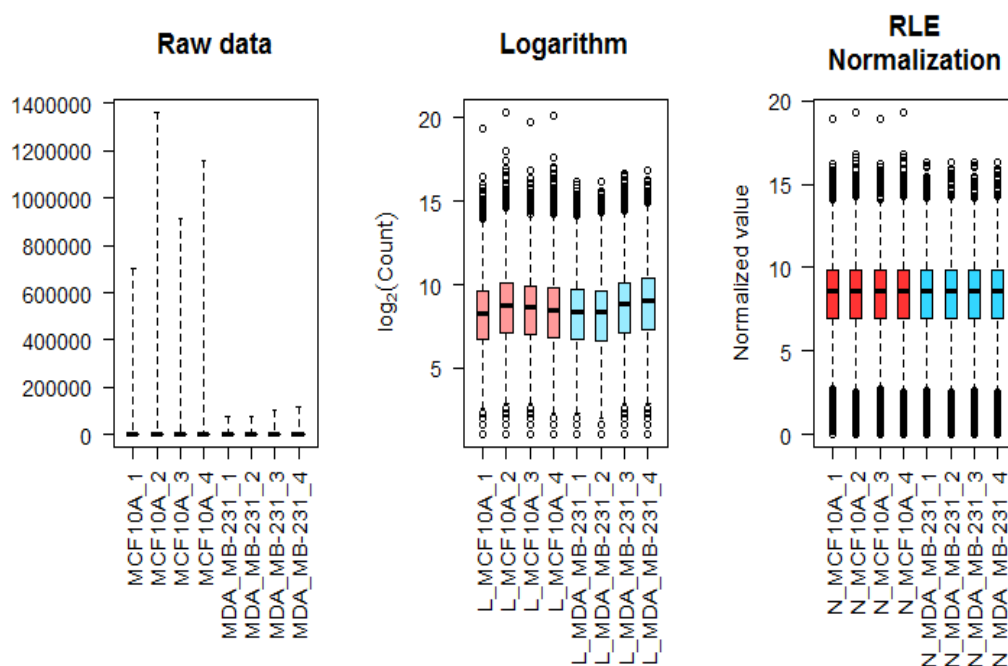
In case of DESeq2, read count+1 & Logarithm value is used to visualize the plots before normalization and regularized log (rlog) transformed value is used to visualize the plots after normalization.

Regularized log transforms the count data to the log₂ scale in a way which minimizes differences between samples for rows with small counts, and which normalizes with respect to library size.

The rlog transformation produces a similar variance stabilizing effect as Variance Stabilizing Transformation (VST), though rlog is more robust in the case when the size factors vary widely.

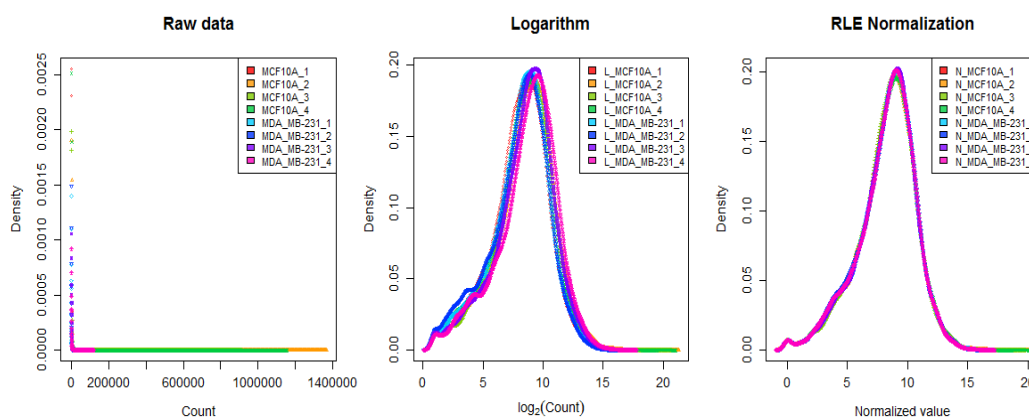
5. 1. 3. 1. Boxplot of Expression Difference between samples.

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 RLE Normalization.



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

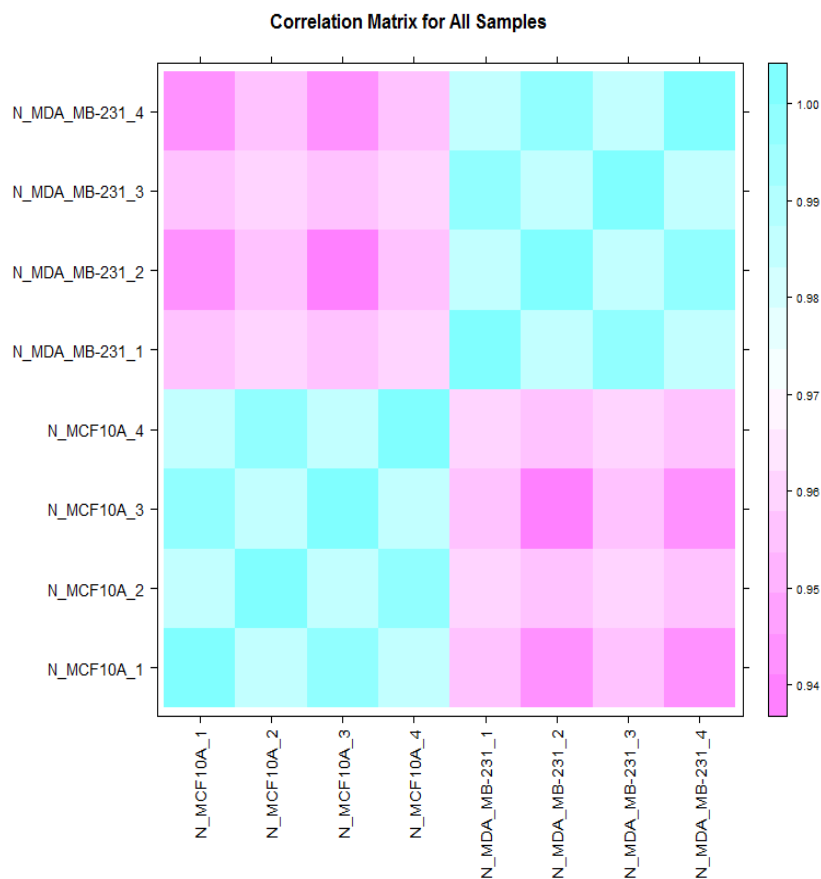
Below density plots show the corresponding samples expression distribution before and after of raw signal (read count), Log2 transformation of read count+1 and RLE Normalization.



5. 1. 4. Correlation Analysis between samples

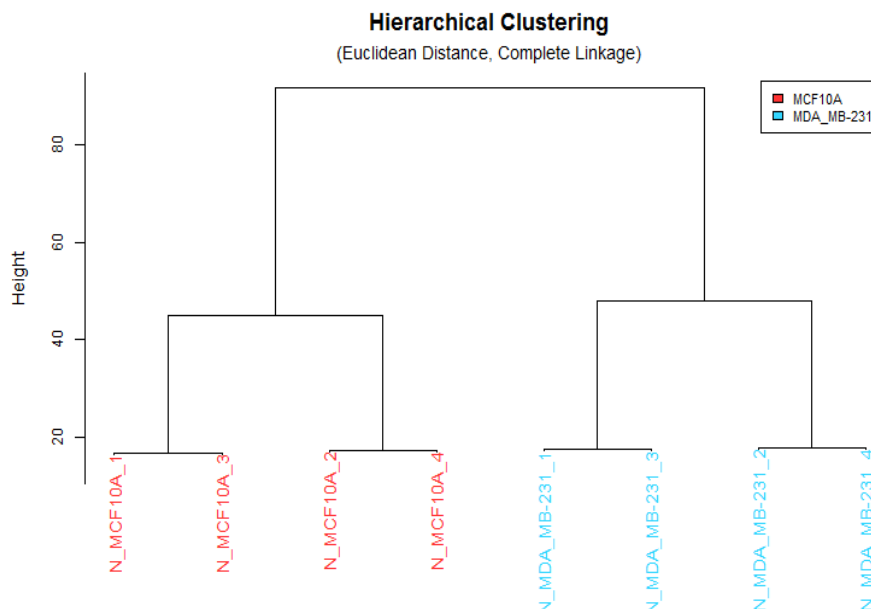
The similarity between samples are obtained through Pearson's coefficient of the normalized value. For range: $-1 \leq r \leq 1$, the closer the value is to 1, the more similar the samples are.

Correlation matrix of all samples is as follows.



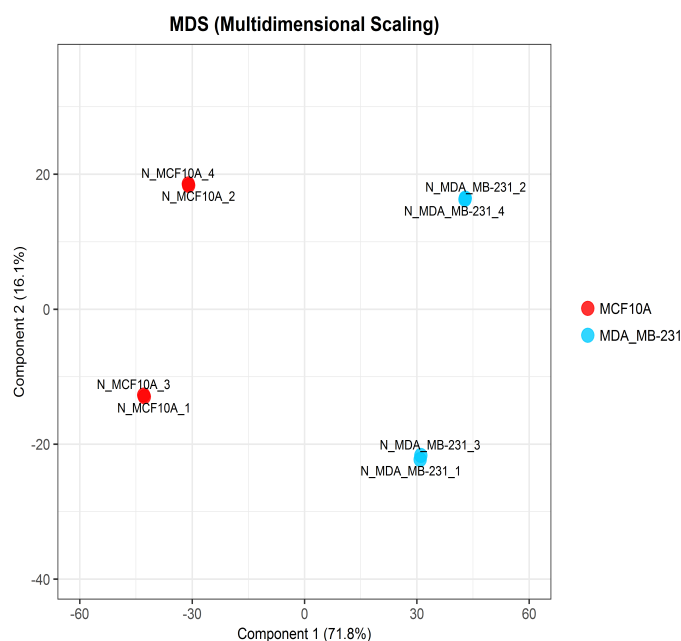
5. 1. 5. Hierarchical Clustering Analysis

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 Gene Analysis Workflow

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

- 1) the read count value of known genes obtained through -e option of the StringTie were used as the original raw data.
 - Raw data
(Refer to Path: result_RNAseq_excel/Expression_profile/StringTie/Expression_Profile.hg19.gene.xlsx)
: 27,685 genes, 8 samples
- 2) During data preprocessing, low quality transcripts are filtered. Afterwards, RLE Normalization are performed.
 - Processed data
(Refer to Path: result_RNAseq_excel/DEG_result/data2.xlsx)
: 12,359 genes, 8 samples
- 3) Statistical analysis is performed using Fold Change, nbinomWaldTest using DESeq2 per comparison pair.
The significant results are selected on conditions of $|fc| \geq 2$ & nbinomWaldTest raw p-value < 0.05 .
 - Significant data
(Refer to Path: result_RNAseq_excel/DEG_result/data3_fc2 & raw.p.xlsx)
: 3,563 genes
- 4) For significant lists, hierarchical clustering analysis is performed to group the similar samples and genes. These results are graphically depicted using heatmap and dendrogram.
 - Hierarchical Clustering (Euclidean Distance, Complete Linkage)
(Refer to Path: result_RNAseq_excel/DEG_result/Cluster image/)

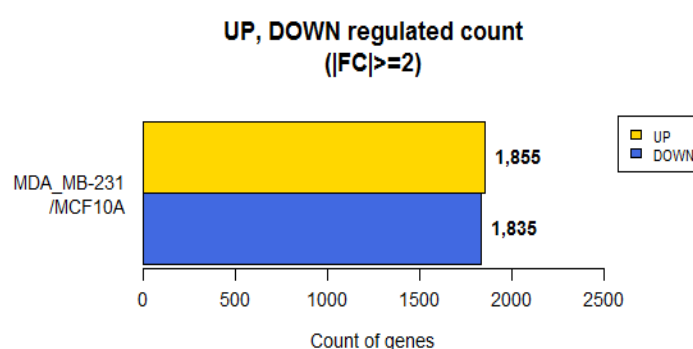
5. 3. Significant Gene Results

(Refer to Path: result_RNAseq_excel/DEG_result/Plots/)

These are DEG result of MDA_MB-231_vs_MCF10A meeting fc2 & raw.p by example.

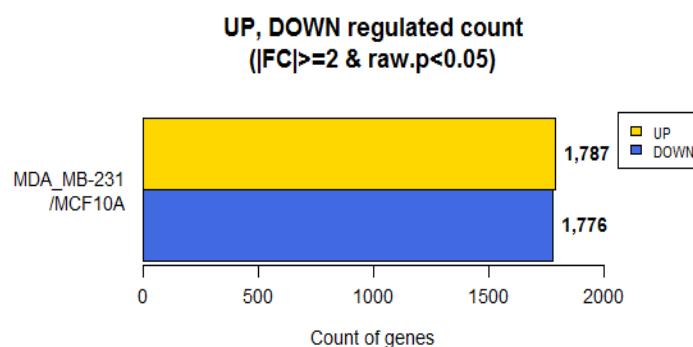
5. 3. 1. Up, Down Regulated Count by Fold Change

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



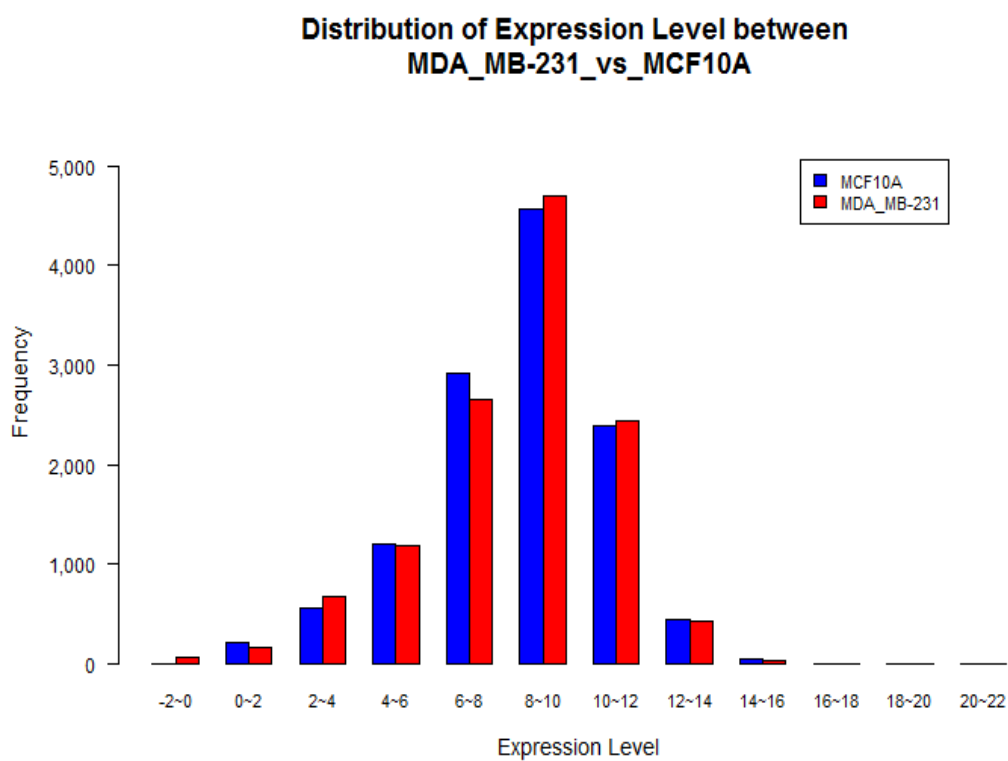
5. 3. 2. Up, Down Regulated Count by Fold Change and p-value

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



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

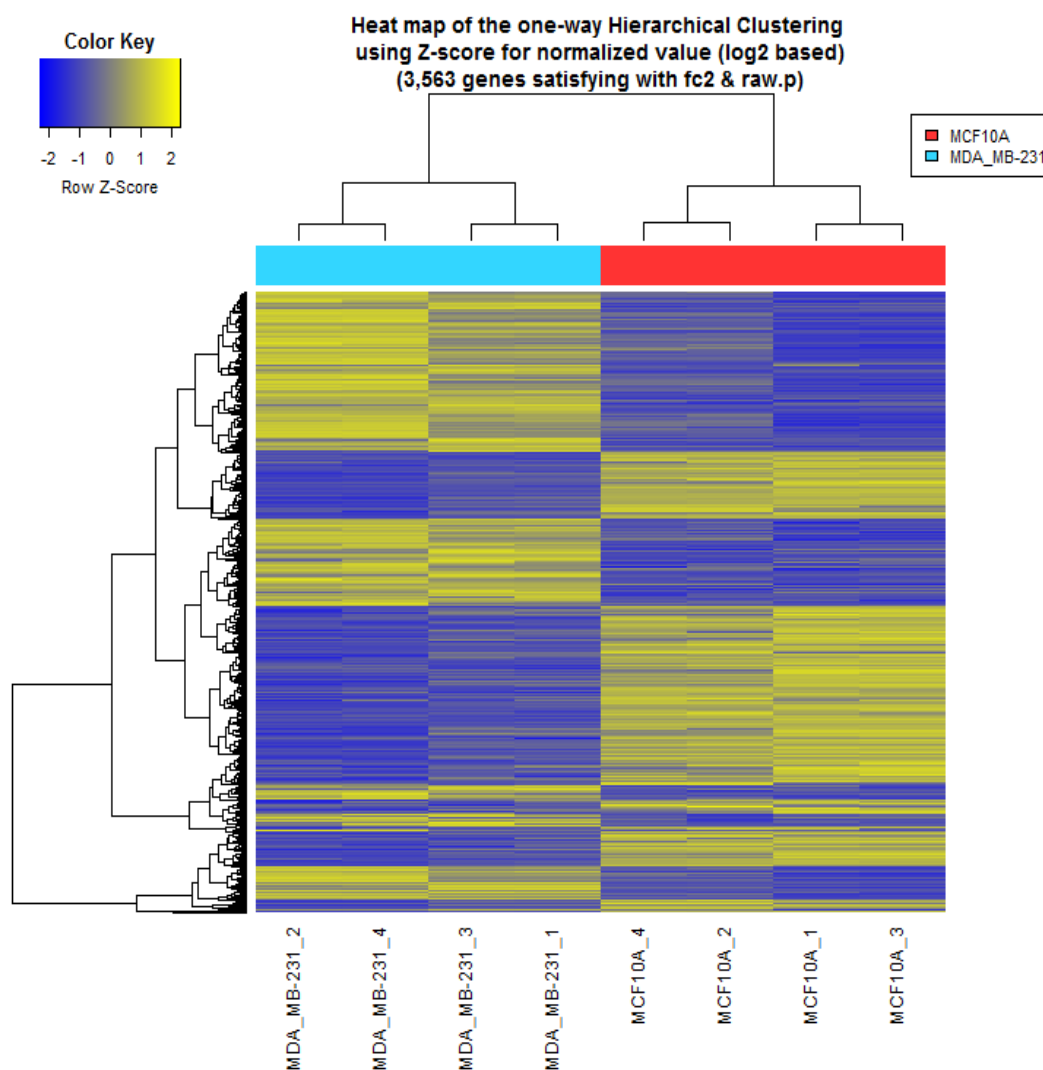
Shows distribution of normalized value of each group for comparison pair.



5. 3. 4. Hierarchical Clustering Analysis

(Refer to Path: result_RNAseq_excel/DEG_result/Cluster image/)

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



6. Data Download Information

6.1. Raw Data

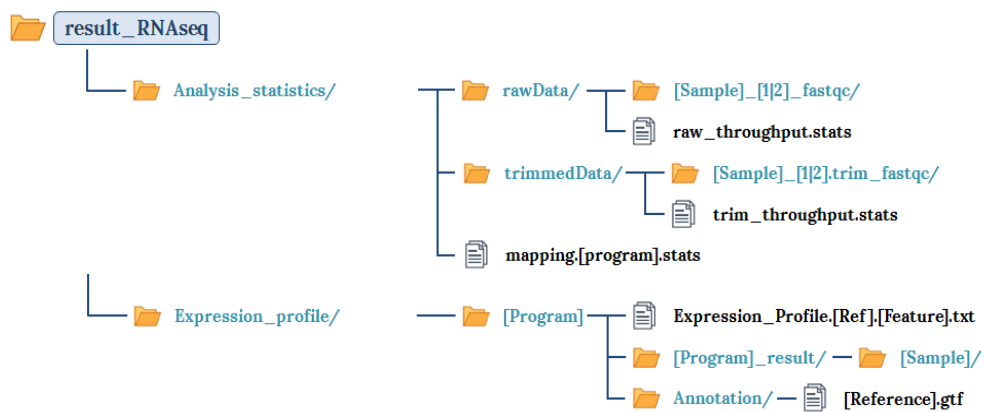
Raw data is the FASTQ file that isn't trimmed adapter sequence.

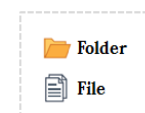
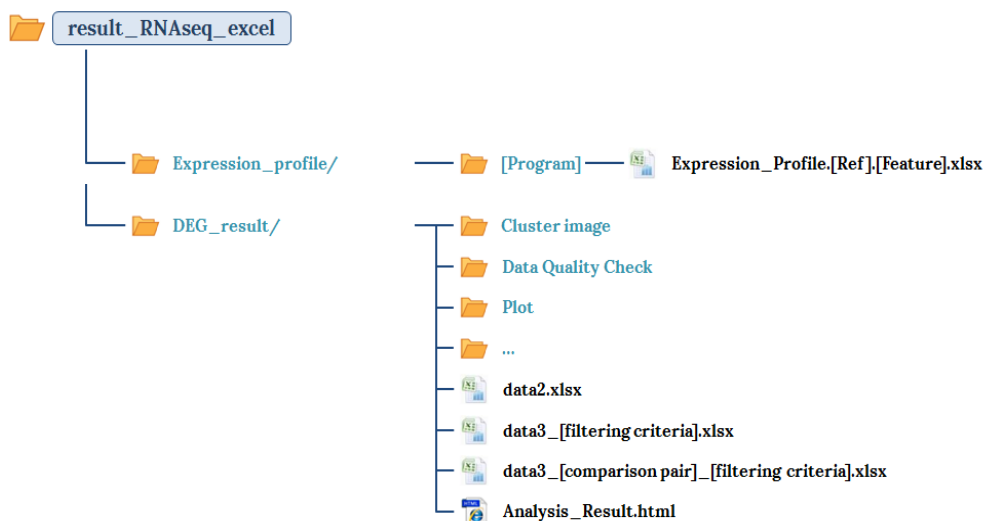
Download link	File size	md5sum
MCF10A_1_1.fastq.gz	747.17M	50165962b46325930a5be99d7db282b5
MCF10A_2_1.fastq.gz	1.14G	4bfb94b8eae33ff93c4e0bdcf03ff96a
MCF10A_3_1.fastq.gz	1.02G	3ed31a4b82dd352bf8a6b5301baf01d0
MCF10A_4_1.fastq.gz	992.42M	69169eb825fb42dc005f8aebcd2be078
MDA_MB-231_1_1.fastq.gz	720.31M	7d7ecfddf9ecc89c87b903e06b7123d5
MDA_MB-231_2_1.fastq.gz	677.59M	e6f27f809305a4d3b841acd3ba7a65c6
MDA_MB-231_3_1.fastq.gz	1022.27M	efa4efa6c7cd14851c49295dc983595b
MDA_MB-231_4_1.fastq.gz	1.14G	ca5d02e5fb3ce07c01d33964285939b2


- 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
1901UQHS-0097_result_RNAseq.zip (md5sum: bf0f071ddb24087e3e40c05eeb6207b5)	83.99M
1901UQHS-0097_result_RNAseq_excel.zip (md5sum: 9155bbb80c4618a5e871159a826f5631)	15.93M





 The data retention period is three months,
please send an e-mail (ngssales@macrogenlab.com)
or contact representative if you want longer retention period.

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%	!"#\$%&'()*+,-./0123456789:;h=i?
20	1 in 100	99%	6789:;h=i?
30	1 in 1000	99.9%	@ABCDEFGHIJ
40	1 in 10000	99.99%	

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. 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.

7. 2. 3. 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. 4. StringTie version 1.3.4d

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.

7. 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.

