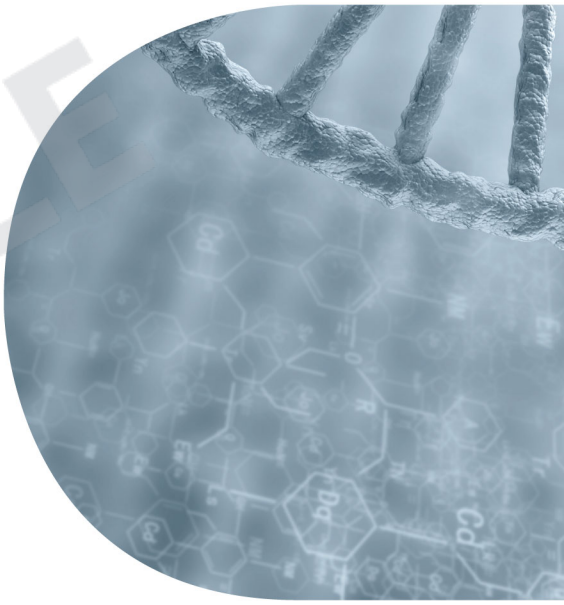
Research use only

Mus musculus Transcriptome Sequencing

Report



ver. MGTR4.0_TC_mm10

Humanizing Genomics
macrogen

Project Information

Client Name	TESTER
Company/Institution	Macrogen
Order Number	HN00000000
Species	*Mus musculus*
Reference	mm10
Annotation	NCBI_108
Type of Read	Paired-ends
Read Length	101
Number of Samples	6
Library Kit	TruSeq Stranded Total RNA Library Prep Gold Kit
Type of Sequencer	Illumina platform

Project Results Summary

In this study, *Mus musculus* 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 6 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.

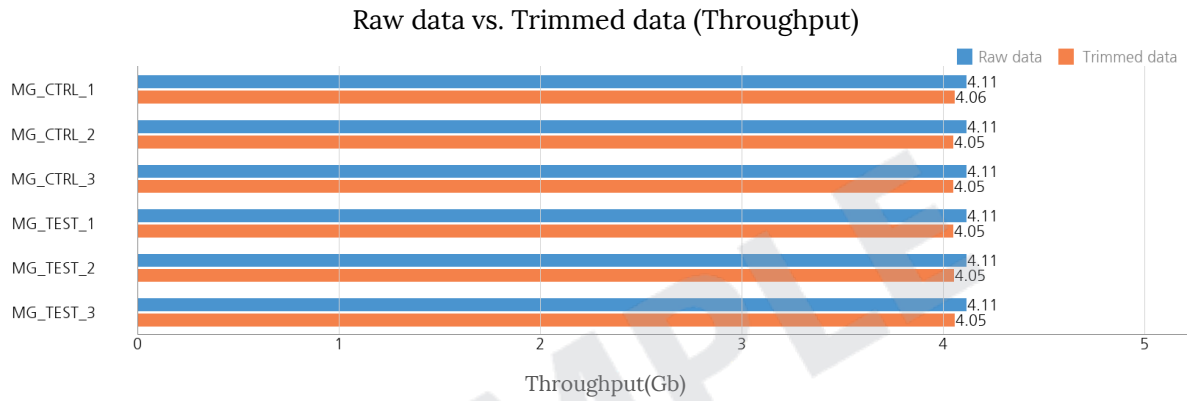


Figure 1. Throughput output of Raw and Trimmed data

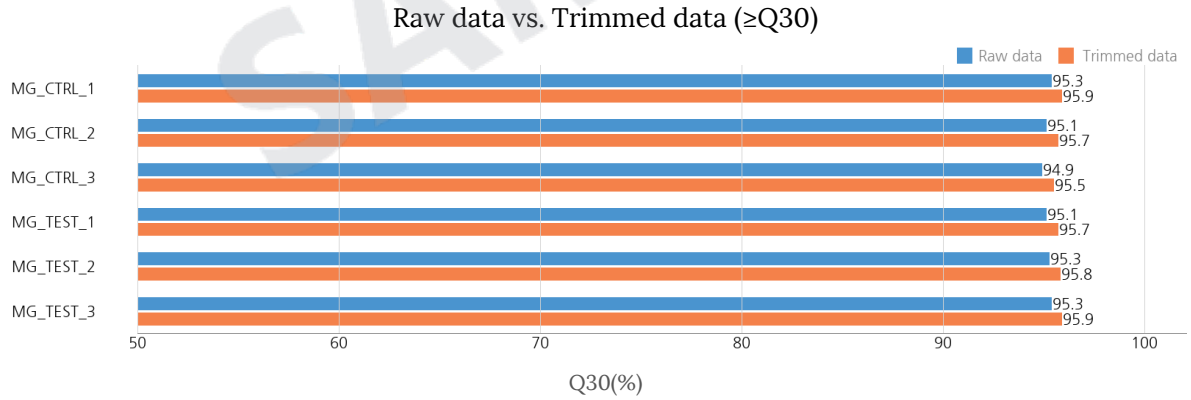


Figure 2. Q30 score of Raw and Trimmed data

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

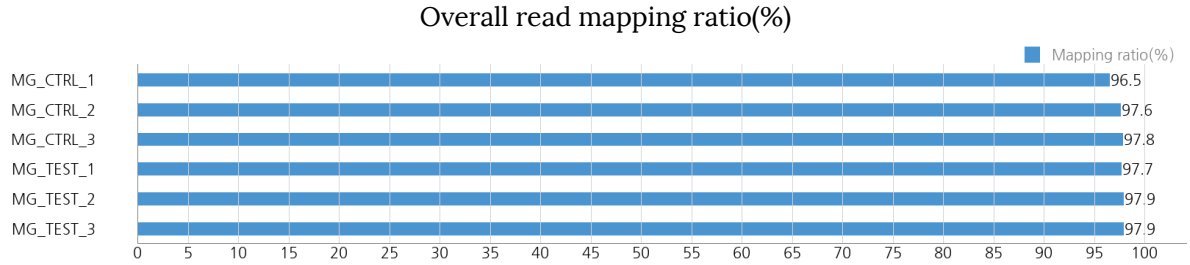


Figure 3. Overall read mapping ratio(%)

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

DEG (Differentially Expressed Genes) analysis was performed on a comparison pair (TEST_vs_CTRL) as requested using DESeq2. The results showed 244 genes which satisfied |fc|>=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.

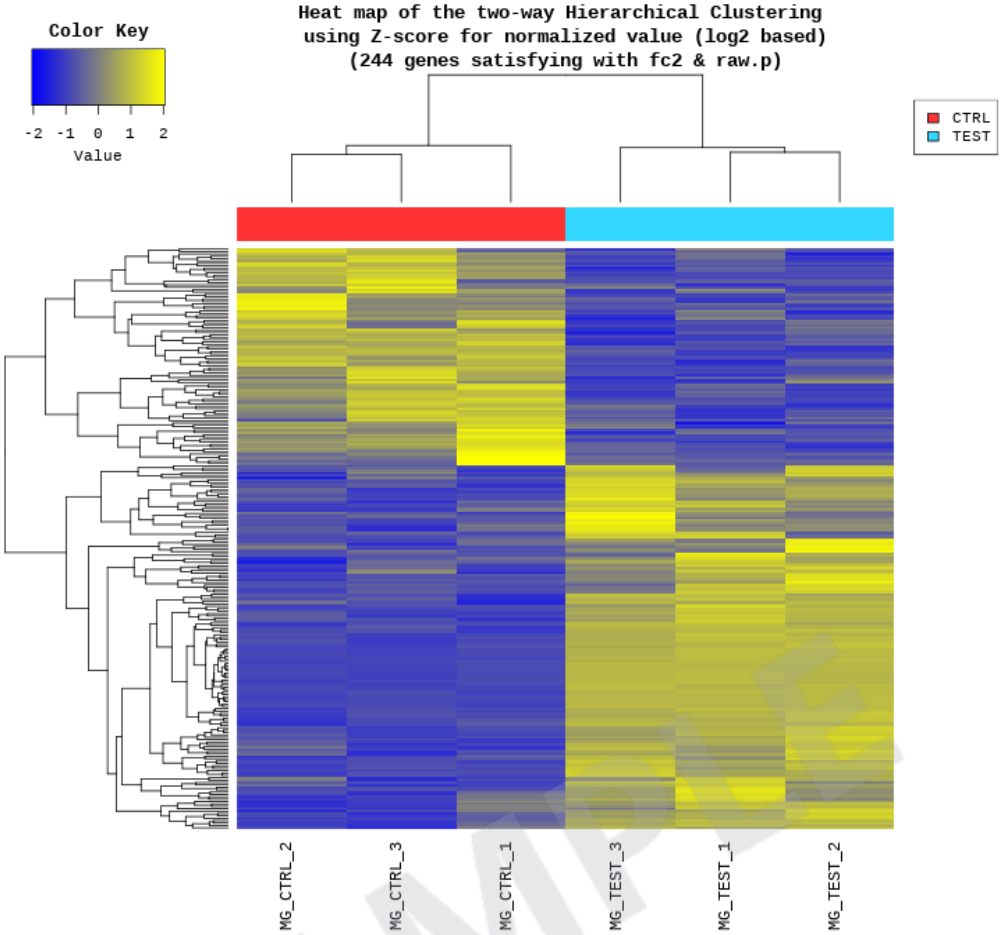


Figure 4. Heatmap for DEG list

DEG list was further analyzed with gProfiler (https://biit.cs.ut.ee/gprofiler/orth) for gene set enrichment analysis per biological process (BP), cellular component (CC) and molecular function (MF). The Figure 5, 6 and 7 show the significant gene set by each category.

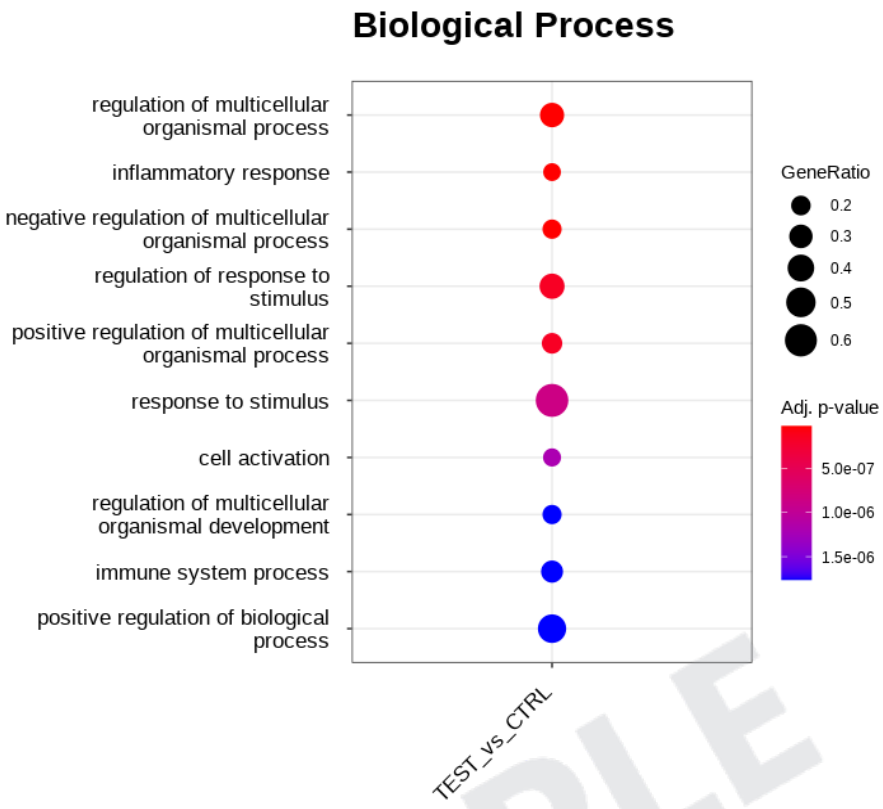


Figure 5. Gene Ontology terms related to Biological Process

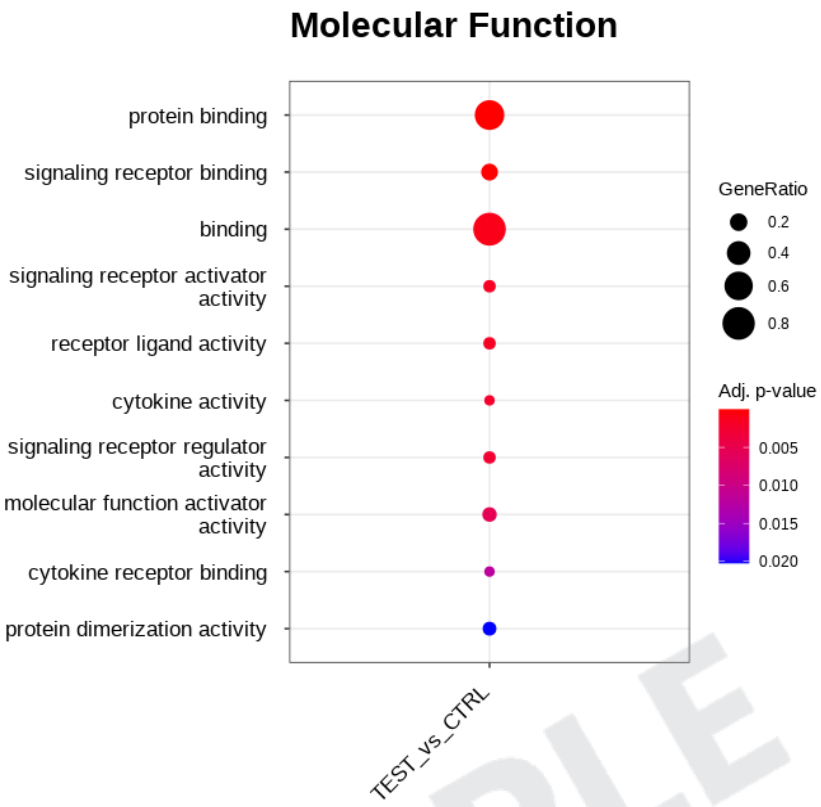


Figure 6. Gene Ontology Terms related to Molecular Function

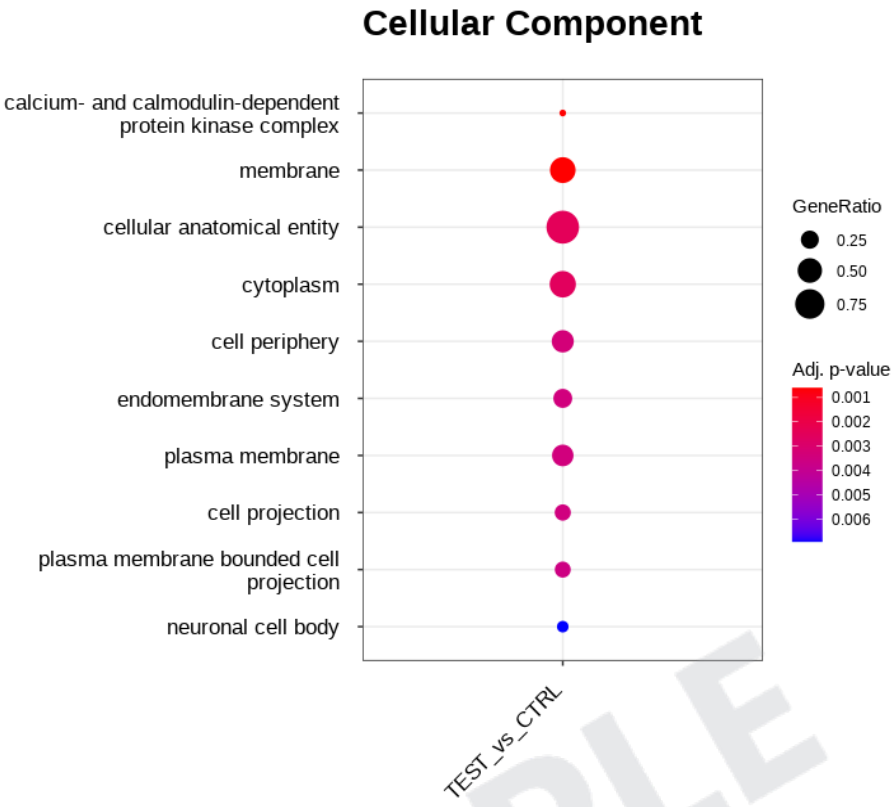


Figure 7. Gene Ontology Terms related to Cellular Component

Table of Contents

1. Experimental Methods and Workflow

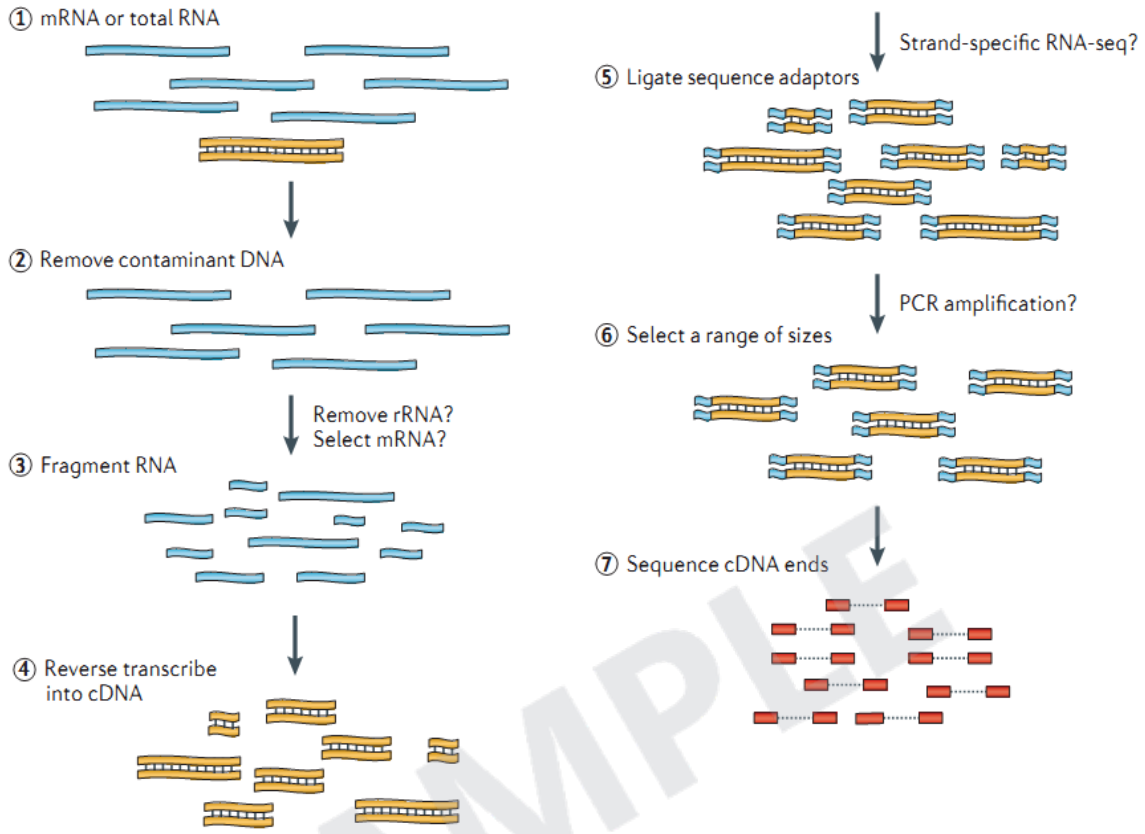


Figure 8. RNA Sequencing Experiment Workflow

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 9. 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 TopHat, splice-aware aligner.

4) Transcript is assembled by Cufflinks 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).

6) In groups with different conditions, genes or transcripts that express differentially are filtered out through statistical hypothesis testing.

7) In case of known gene annotation, functional annotation and gene-set enrichment analysis are performed using GO and KEGG database on differentially expressed genes.

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 6 samples. For example, in MG_CTRL_1, 40,718,992 reads are produced, and total read bases are 4.1Gbp. The GC content (%) is 47.81% and Q30 is 95.35%.

Table 1. Raw data stats

Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
MG_CTRL_1	4,112,618,192	40,718,992	47.81	98.52	95.35
MG_CTRL_2	4,112,811,910	40,720,910	48.15	98.41	95.11
MG_CTRL_3	4,112,524,464	40,718,064	48.13	98.34	94.9
MG_TEST_1	4,112,925,636	40,722,036	47.41	98.41	95.1
MG_TEST_2	4,112,882,408	40,721,608	47.89	98.48	95.27
MG_TEST_3	4,112,791,710	40,720,710	47.38	98.52	95.35

(* 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

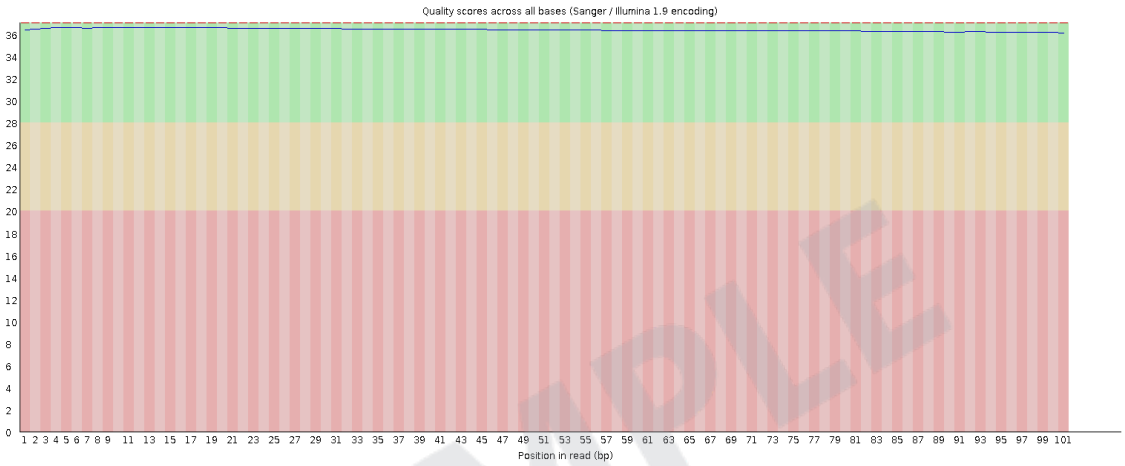


Figure 10. Read quality at each cycle of MG_CTRL_1 (read1)

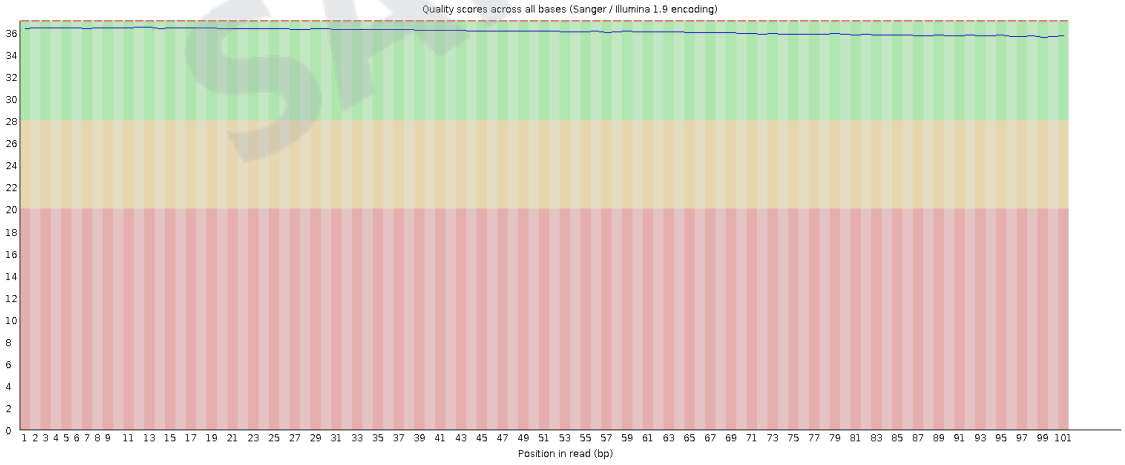


Figure 11. Read quality at each cycle of MG_CTRL_1 (read2)

- 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(%)
MG_CTRL_1	4,056,017,217	40,323,630	47.82	98.88	95.86
MG_CTRL_2	4,048,912,278	40,265,808	48.17	98.81	95.68
MG_CTRL_3	4,048,254,801	40,260,258	48.14	98.75	95.48
MG_TEST_1	4,048,521,949	40,263,096	47.42	98.81	95.68
MG_TEST_2	4,052,327,022	40,291,404	47.91	98.86	95.81
MG_TEST_3	4,054,156,420	40,310,810	47.39	98.89	95.87

- 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 12 and 13 show average base quality at each cycle after trimming.

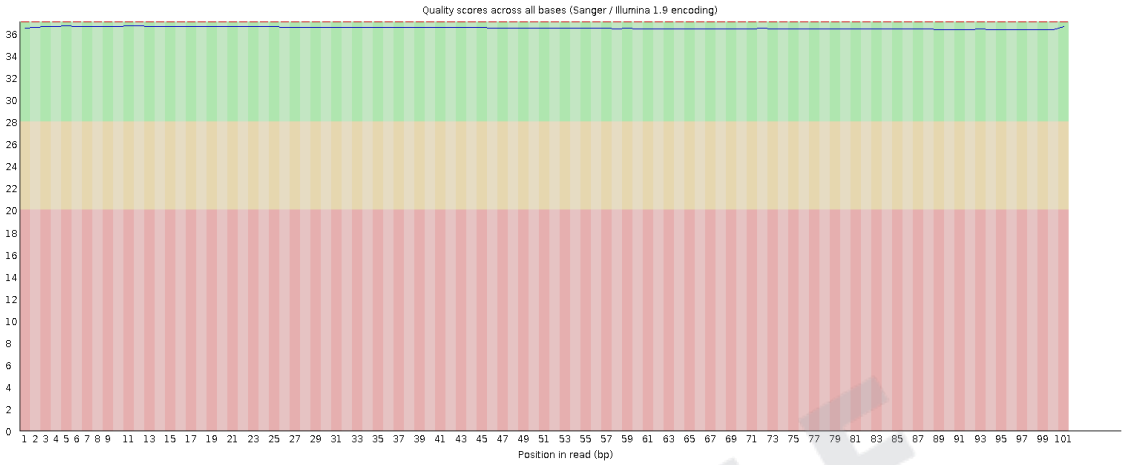


Figure 12. Average base quality of MG_CTRL_1 (read1) at each cycle after trimming

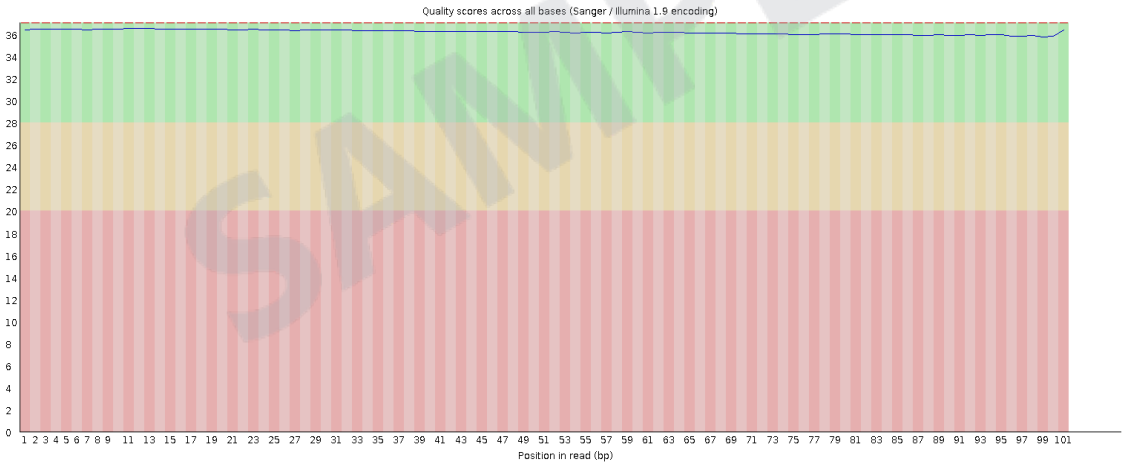


Figure 13. Average base quality of MG_CTRL_1 (read2) at each cycle after trimming

- 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.tophat.stats.txt)

In order to map cDNA fragments obtained from RNA sequencing, mm10 was used as a reference genome. Table 3 shows the statistic obtained from Tophat, which is known to handle spliced read mapping through Bowtie2 aligner. You can check number of processed reads, mapped reads, multiple mapped reads, and overall mapping ratio.

Table 3. Mapped Data Stats

Sample ID	# of processed reads	# of mapped reads	# of failed to align reads	# of multiple mapped reads
MG_CTRL_1	40,323,630	38,929,727 (96.5%)	1,393,903 (3.5%)	1,795,660 (4.6%)
MG_CTRL_2	40,265,808	39,313,717 (97.6%)	952,091 (2.4%)	1,731,515 (4.4%)
MG_CTRL_3	40,260,258	39,369,329 (97.8%)	890,929 (2.2%)	1,832,435 (4.7%)
MG_TEST_1	40,263,096	39,349,126 (97.7%)	913,970 (2.3%)	1,836,508 (4.7%)
MG_TEST_2	40,291,404	39,445,696 (97.9%)	845,708 (2.1%)	1,842,485 (4.7%)
MG_TEST_3	40,310,810	39,484,411 (97.9%)	826,399 (2.1%)	1,754,523 (4.4%)

- # of processed reads: Number of cleaned reads after trimming
- # of mapped reads: Number of reads mapped to reference
- # of multiple mapped reads: Number of multiple mapped reads

4. 2. Expression Profiling

Known transcripts, alternative splicing transcripts and novel transcripts are assembled with Cufflinks 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/Expression_profile/Cufflinks/ Expression_Profile.mm10.transcript.xlsx)

Table 4 is an example of known transcript expression level per sample in expression value. This result is obtained by Reference Annotation Based Transcript (RABT) method using -G option of Cufflinks 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
NM_001101	60	ACTB	actin beta	chr7:5566778-5570232	1812	101378	144745	1009.54	1362.35
NM_004301	86	ACTL6A	actin like 6A, transcript variant 1	chr3:179280667-1793061	1898	1125	2304	10.427	20.3277
NM_001130004	87	ACTN1	actinin alpha 1, transcript variant 1	chr14:69340839-6944608	3791	27	120	0.129273	0.548208
NM_001130005	87	ACTN1	actinin alpha 1, transcript variant 3	chr14:69340839-6944608	3710	75	49	0.368167	0.226404
NM_001102	87	ACTN1	actinin alpha 1, transcript variant 2	chr14:69340839-6944608	3725	19342	25769	94.5861	120.143
NM_001258371	89	ACTN3	actinin alpha 3 (gene/pseudogene), trans	chr11:66313865-6633080	3087	1	1	4.19515E-05	3.80965E-05
NR_047663	89	ACTN3	Homo sapiens actinin alpha 3 (gene/pse	chr11:66314311-6633080	2939	21	11	0.124547	0.0600351
NM_001105	90	ACVR1	activin A receptor type 1, transcript varia	chr2:158592957-1587316	3045	1	1	8.08978E-10	4.14268E-08
NM_001111067	90	ACVR1	activin A receptor type 1, transcript varia	chr2:158592957-1587323	2864	539	380	3.31363	2.31123
NM_000020	94	ACVRL1	activin A receptor like type 1, transcript v	chr12:52301201-5231714	4263	1	1	0.00007053	8.00872E-06
NM_001077401	94	ACVRL1	activin A receptor like type 1, transcript v	chr12:52306112-5231714	4126	28	34	0.113094	0.133785
NM_000666	95	ACY1	aminoacylase 1, transcript variant 1	chr3:52017299-52023218	1678	203	330	2.11537	3.32206
NM_001198897	95	ACY1	aminoacylase 1, transcript variant 4	chr3:52017299-52023218	1483	20	1	0.225968	1.21107E-11
NM_001198898	95	ACY1	aminoacylase 1, transcript variant 5	chr3:52017299-52023218	1573	50	63	0.555338	0.673857
NM_001198895	95	ACY1	aminoacylase 1, transcript variant 2	chr3:52017299-52023218	1673	1	1	2.53292E-05	8.57631E-05
NM_001198896	95	ACY1	aminoacylase 1, transcript variant 3	chr3:52017299-52023218	1462	1	1	4.79615E-12	8.87341E-23
NR_126393	97	ACYP1	acylphosphatase 1, transcript variant 2	chr14:75519927-7553075	702	1	11	1.70328E-19	0.250103

- 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

4. 2. 2. Known Genes Expression Level

(Refer to Path: result_RNAseq/Expression_profile/Cufflinks/ Expression_Profile.mm10.gene.xlsx)

Table 5 is an example of known gene expression level per sample in expression value. This result is obtained by Reference Annotation Based Transcript (RABT) method using -G option of Cufflinks 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
27	NM_001136000,NM_001136001,NM_001	ABL2	ABL proto-oncogene 2, non-receptor tyrosine	5145	6735	7.836987	9.9130971
28	NM_020469	ABO	ABO, alpha 1-3-N-acetylgalactosaminyltransfe	3	4	0.024627	0.0481101
37	NM_000018,NM_001033859,NM_001270	ACADVL	acyl-CoA dehydrogenase, very long chain	26094	26688	200.835542	195.690048
38	NM_000019	ACAT1	acetyl-CoA acetyltransferase 1	1042	3566	8.55795	27.8635
39	NM_001303253,NM_005891	ACAT2	acetyl-CoA acetyltransferase 2	933	1427	10.457377	15.259281
40	NM_001094,NM_183377	ASIC2	acid sensing ion channel subunit 2	4	0	0.013697848	0
172	NR_003226,NR_003227,NR_003228	AFG3L1P	AFG3 like matrix AAA peptidase subunit 1, pse	681	655	4.219935	3.9539
173	NM_001133	AFM	afamin	0	3	0	0.0183909
174	NM_001134	AFP	alpha fetoprotein	3	0	0.0185119	0
175	NM_000027,NM_001171988,NR_033655	AGA	aspartylglucosaminidase	251	284	2.082699679	2.2293071
176	NM_001135,NM_013227	ACAN	aggrecan	9	4	0.016414383	0.003922265
177	NM_001136,NM_001206929,NM_001206	AGER	advanced glycosylation end-product specific r	3120	2956	33.65818658	30.51683042
178	NM_000028,NM_000642,NM_000643,NM	AGL	amylo-alpha-1, 6-glucosidase, 4-alpha-glucan	4866	3634	11.58945114	8.23442506
181	NM_001138	AGRP	agouti related neuropeptide	0	0	0	0
182	NM_000214	JAG1	jagged 1	859	691	2.52453	1.94234
183	NM_000029	AGT	angiotensinogen	3	0	0.0143218	0
185	NM_000685,NM_004835,NM_009585,NM	AGTR1	angiotensin II receptor type 1	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/DEG_result/[DataSet]/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

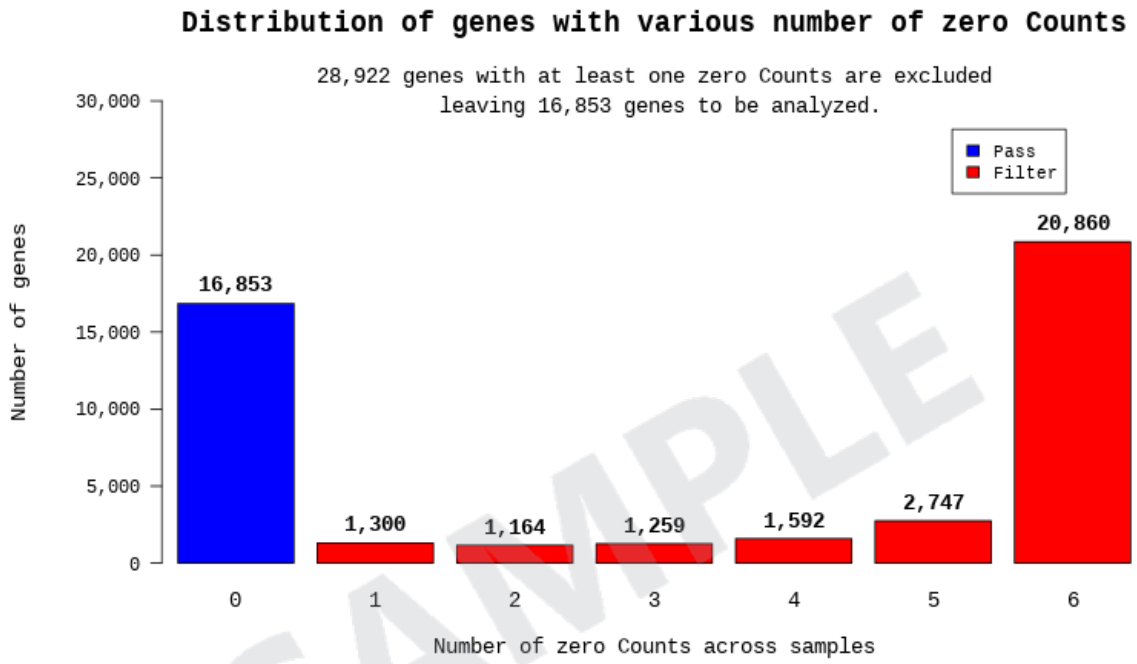
Comparison pair and statistical method for each pair are shown below.

Index	Test vs. Control	Statistical Method
1	TEST vs. CTRL	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering

5. 1. 2. DATA Quality Check

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/Data Quality Check/)

For 6 samples, if more than one read count value was 0, it was not included in the analysis. Therefore, from total of 45,775 genes, 28,922 were excluded and only 16,853 genes were used for statistic analysis.



5. 1. 3. Data Transformation and Normalization

In order to reduce systematic bias, size factors were estimated from the read count data (estimateSizeFactors method).

Using them, the read count data was normalized with Relative Log Expression (RLE) method in DESeq2 R library.

Then, statistical test was performed with the normalized data.

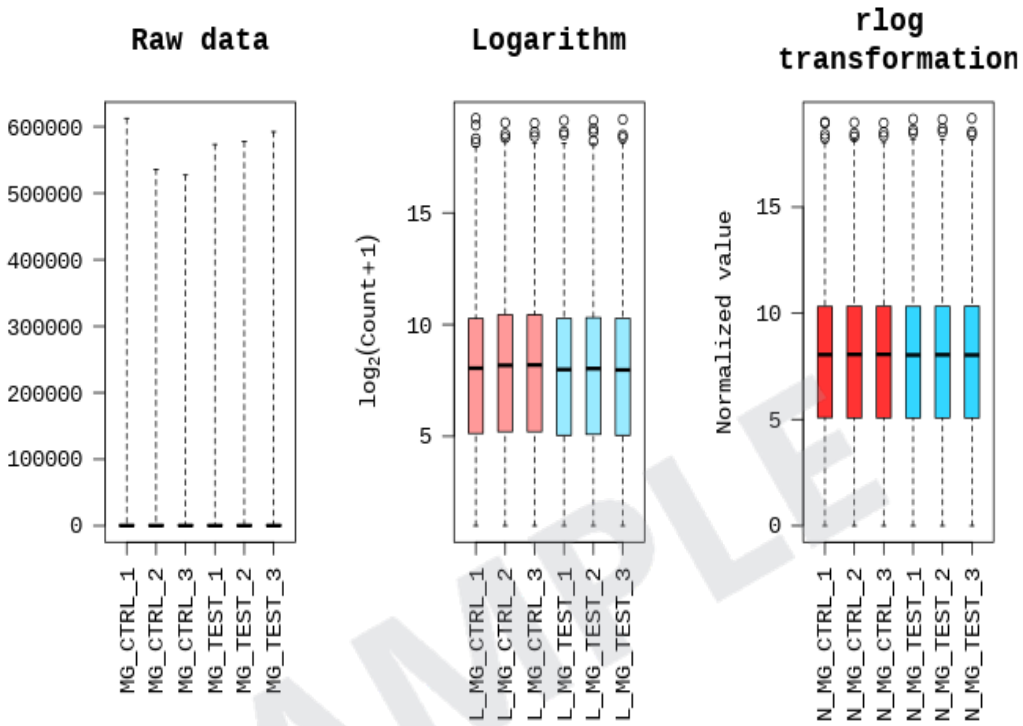
log2(read count+1) and regularized log (rlog) transformed values were used for data visualization. rlog transformation is a method to minimize differences between samples for genes/transcripts in low expression. It transforms count data into log2 scale and normalizes them with a library size factor. rlog is robust in the case when the size factors vary widely.

These logarithm figures were used only for visualization.

To proceed a statistical test, RLE normalized count was adopted for negative binomial Wald Test(nbinomWaldTest) in DESeq2.

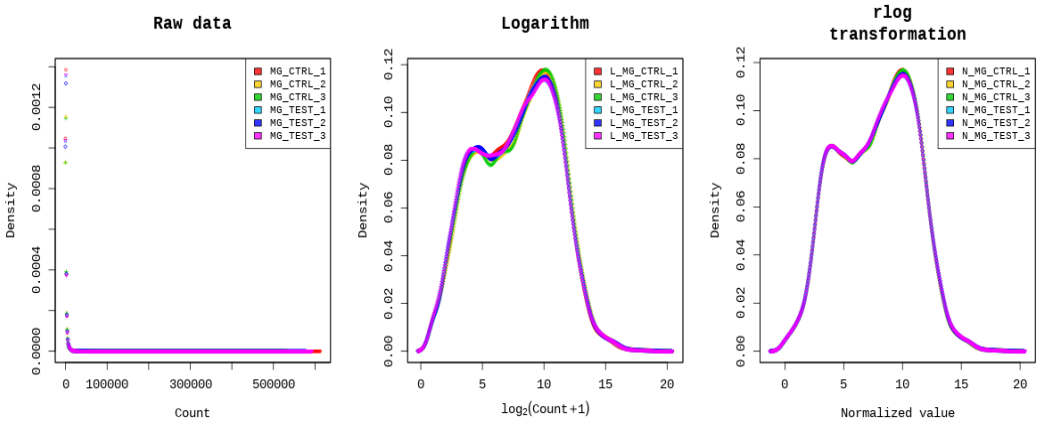
5. 1. 3. 1. Boxplot of Expression Difference between samples.

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

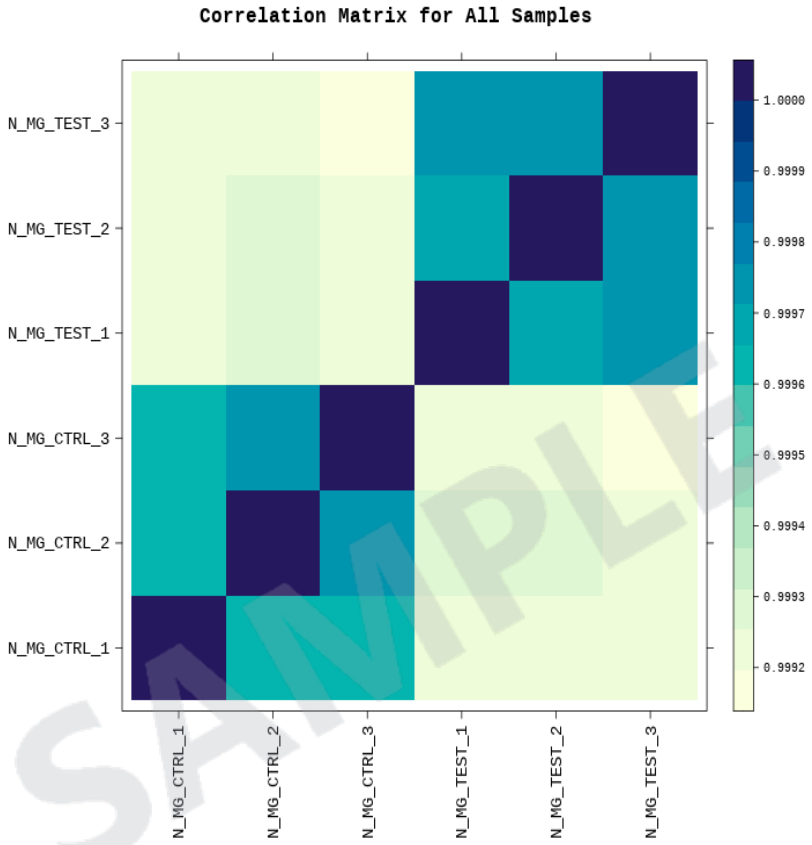
Below density plots show the corresponding samples expression distribution before and after of raw signal (read count), Log2 transformation of read count+1 and RLE Normalization.



5. 1. 4. Correlation Analysis between samples

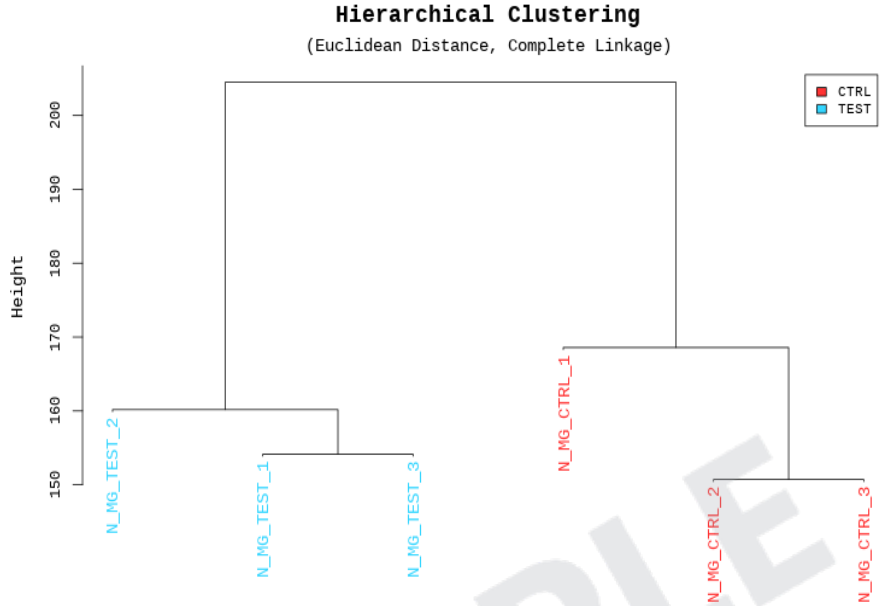
The similarity between samples are obtained through Pearson's coefficient of the rlog transformed value. For range: -1≤ r ≤ 1,the closer the value is to 1, the more similar the samples are.

Correlation matrix of all samples is as follows.



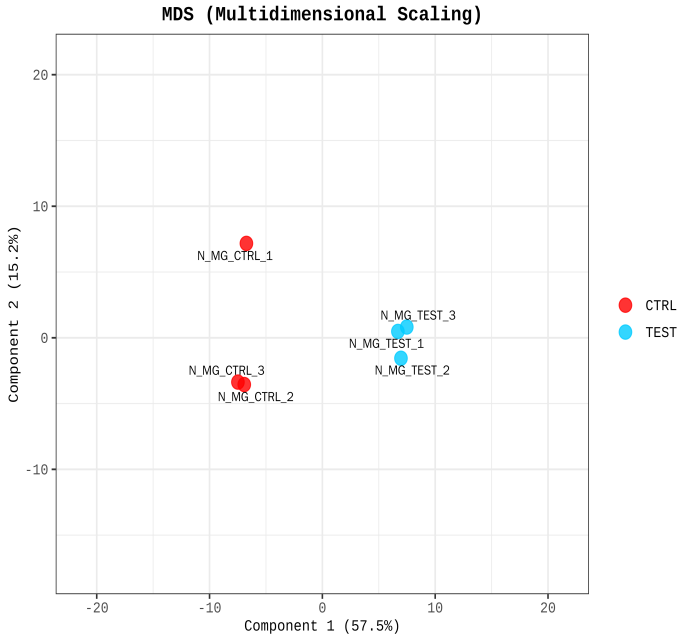
5. 1. 5. Hierarchical Clustering Analysis

Using each sample's rlog transformed 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 rlog transformed value, the similarity between samples is graphically shown in a 2D plot. It employs two components that well preserve thr degree of similarity between samples. 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 -G option of the Cufflinks were used as the original raw data.

- Raw data

(Refer to Path: result_RNAseq/Expression_profile/Cufflinks/Expression_Profile.mm10.gene.xlsx)

: 45,775 genes, 6 samples

2) During data preprocessing, low quality transcripts are filtered. Afterwards, RLE Normalization are performed.

- Processed data

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/data2.xlsx)

: 16,853 genes, 6 samples

3) Statistical analysis is performed using Fold Change, nbinomWaldTest using DESeq2 per comparison pair.
The significant results are selected on conditions of |fc|>=2 & nbinomWaldTest raw p-value<0.05.

- Significant data

(Refer to Path: result_RNAseq/DEG_result/DEG/data3_fc2_&_raw.p.xlsx)

: 244 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 dendogram.

- Hierarchical Clustering (Euclidean Distance, Complete Linkage)

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/Cluster image/)

5) For significant lists, gene-set enrichment analysis was performed based on gene ontology(https://biit.cs.ut.ee/gprofiler/).
Please refer to the GO_stat sheet and the GO_genes sheet of data3 file.

Following result are provided.

- GO_stat
- GO_genes

6) For significant lists, gene-set enrichment analysis was performed based on KEGG database(http://www.genome.jp/kegg/).
Please refer to the KEGG_stat sheet and KEGG_genes sheet of data3 file.

Following result are provided.

- KEGG_stat
- KEGG_genes

You can also see the KEGG enrichment result on the KEGG_pathway.html.

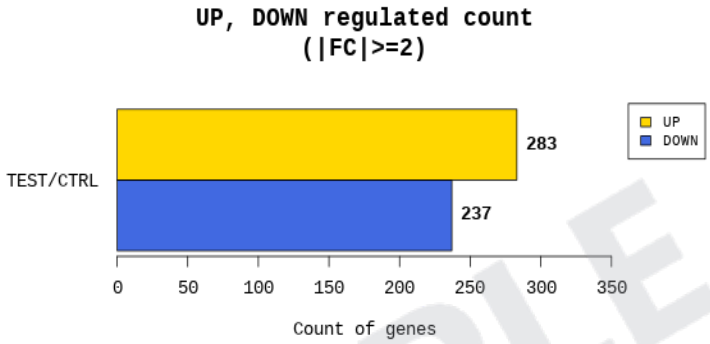
5. 3. Significant Gene Results

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/Plots/)

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

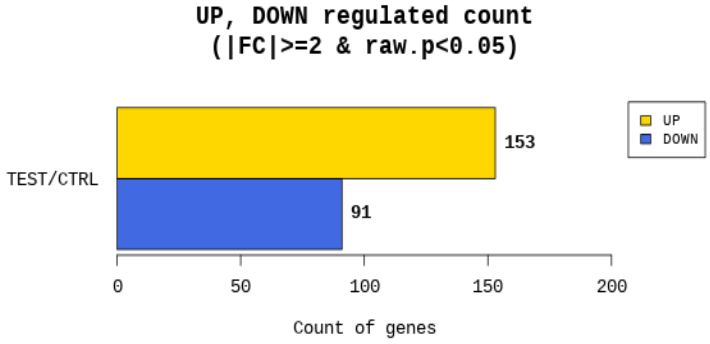
5. 3. 1. Up, Down Regulated Count by Fold Change

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



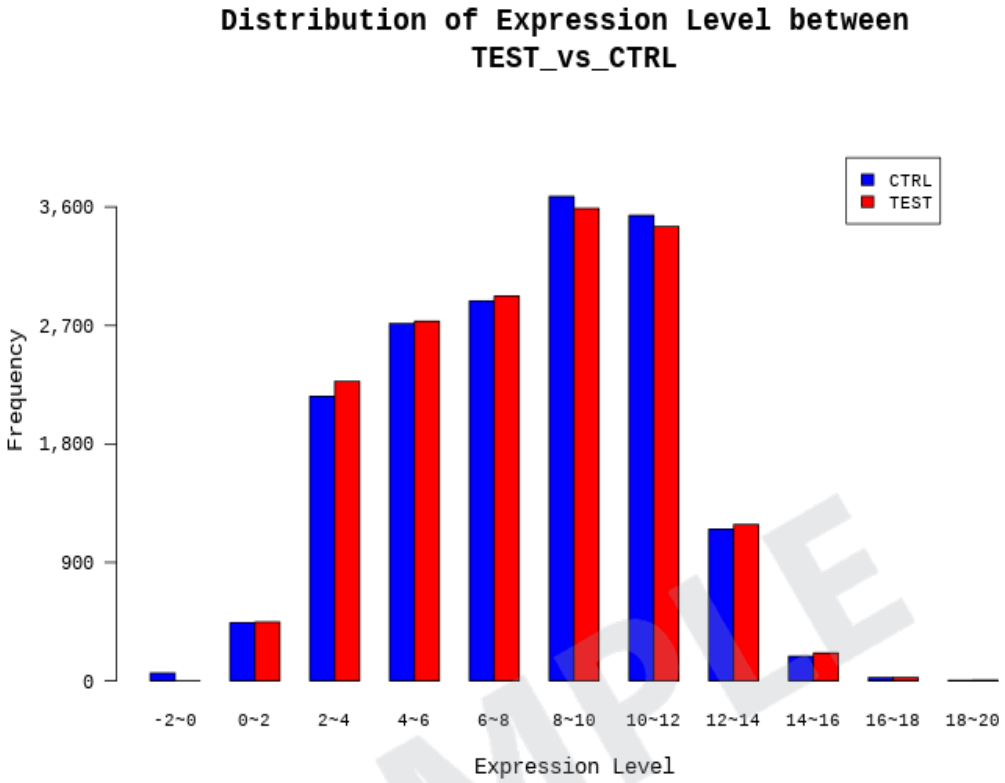
5. 3. 2. Up, Down Regulated Count by Fold Change and p-value

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

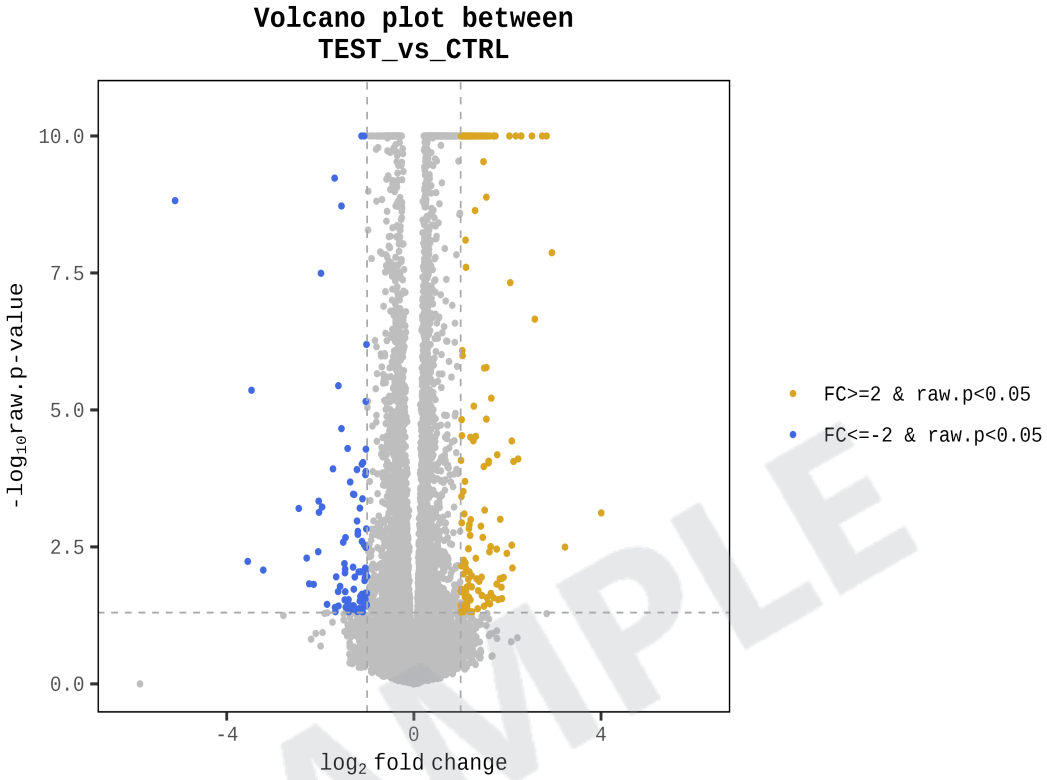
Shows distribution of normalized value of each group for comparison pair.



5. 3. 4. 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.

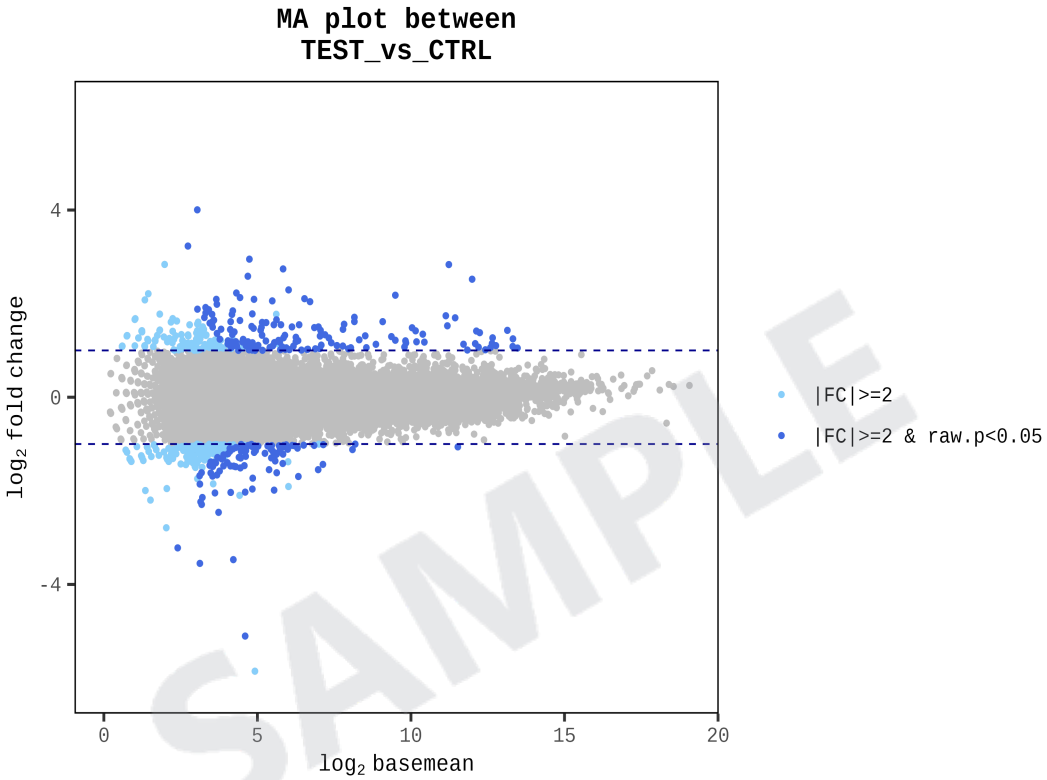
(X-axis: log2 Fold Change, Y-axis: -log10 p-value)



5. 3. 5. MA Plot

In order to confirm the transcripts that show higher expression difference compared to the control according to overall average expression level, MA plot is drawn. (X-axis: mean of normalized counts, Y-axis: log2 Fold Change).

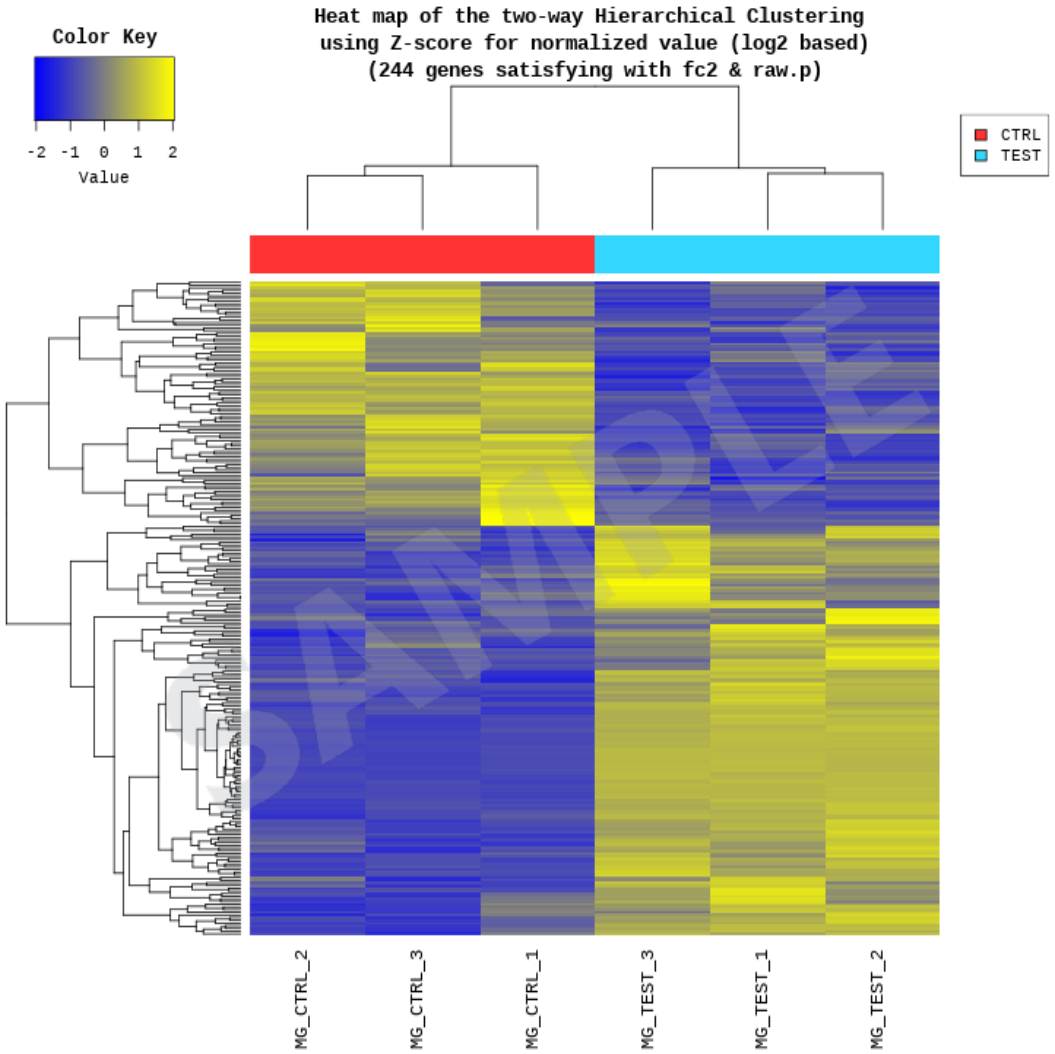
For example, even though fold change might be different by two-fold, the gene with higher mean of normalized counts may be more credible.



5. 3. 6. Hierarchical Clustering Analysis

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/Cluster image/)

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



5. 4. GO Enrichment Analysis

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/gprofiler)

For Enrichment test which based on Gene Ontology (http://geneontology.org/) DB was conducted with significant gene list using g:Profiler tool (https://biit.cs.ut.ee/gprofiler/).

The g:Profiler tool performs statistical enrichment analysis to find over-representation of information from Gene Ontology terms, biological pathways, regulatory DNA elements, human disease gene annotations, and protein-protein interaction networks.

Progressing about 3 categories of GO. The gene or gene product, molecule associated with GO ID was summarized by parsing the ontology file and the annotation file (multispecies annotation provided by Uniprot, or the annotation provided by each type reference DB for the GO consortium) for the GO graph structure.

- Link for the ontology documentation: http://geneontology.org/page/ontology-documentation
- Link for the ontology files: http://geneontology.org/page/download-ontology
- Link for the annotation files: http://geneontology.org/page/download-annotations

Enrichment test result was summarized at each sheet of DEG result(data3-*.xlsx file) by 2 forms below.

- GO_stat
- GO_genes

5. 4. 1. GO_stat Sheet

The result of associated gene and test stat was summarized by term_id. The significane of specific term_id in enrichment test with DEG set was summarized.

source	term_id	term_name	adjusted_p_value	term_size	query_size	intersection_size	effective_domain_size	intersections
GO:CC	GO:0022626	cytosolic ribosome	2.72198E-17	115	1921	50	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	3.60328E-15	96	1824	44	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:MF	GO:0003735	structural constituent of ribosome	1.32911E-14	170	1860	59	18098	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0006613	cotranslational protein targeting to membrane	2.03613E-14	101	1824	44	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:MF	GO:0005198	structural molecule activity	4.45523E-14	739	1860	151	18098	6134, 6206, 127294, 4586, 301, 3887, 6
GO:BP	GO:0045047	protein targeting to ER	7.18306E-14	109	1824	45	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:CC	GO:0044391	ribosomal subunit	2.36014E-13	195	1921	61	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0072599	establishment of protein localization to endoplasmic reticulum	2.82077E-13	113	1824	45	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:BP	GO:0070972	protein localization to endoplasmic reticulum	4.06119E-11	137	1824	47	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:CC	GO:0005840	ribosome	1.34069E-10	246	1921	65	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:CC	GO:0022625	cytosolic large ribosomal subunit	1.69728E-10	64	1921	29	18797	6134, 6155, 6168, 200916, 6167, 6161,
GO:BP	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	7.11348E-10	122	1824	42	17816	6134, 6206, 6155, 6204, 6168, 6167, 61
GO:CC	GO:0044459	plasma membrane part	1.34094E-09	2879	1921	400	18797	165829, 10326, 6405, 4283, 8322, 5743
GO:CC	GO:0071944	cell periphery	1.8891E-09	5662	1921	709	18797	829, 165829, 10326, 23256, 6405, 4283
GO:CC	GO:0005886	plasma membrane	5.37824E-09	5539	1921	692	18797	165829, 10326, 23256, 6405, 4283, 505
GO:CC	GO:0044444	cytoplasmic part	5.37824E-09	9685	1921	1125	18797	6134, 829, 84532, 10326, 5332, 23256,
GO:CC	GO:0005737	cytoplasm	5.47219E-09	11534	1921	1309	18797	6134, 829, 84532, 10326, 5332, 23256,
GO:BP	GO:0009888	tissue development	5.79564E-09	2068	1824	305	17816	6405, 5054, 8322, 5743, 144165, 12729
GO:BP	GO:0006612	protein targeting to membrane	5.95069E-09	195	1824	54	17816	6134, 6206, 6155, 6204, 6168, 6747, 51
GO:BP	GO:0051179	localization	1.23607E-08	6751	1824	824	17816	6134, 829, 10326, 10734, 23256, 6405,
GO:CC	GO:1903561	extracellular vesicle	1.65132E-08	2165	1921	309	18797	829, 5054, 10103, 2098, 9518, 4151, 43
GO:CC	GO:0043230	extracellular organelle	1.66899E-08	2167	1921	309	18797	829, 5054, 10103, 2098, 9518, 4151, 43
GO:CC	GO:0044445	cytosolic part	2.71585E-08	252	1921	60	18797	6134, 6206, 6155, 6204, 6168, 338321,
GO:BP	GO:0032501	multicellular organismal process	3.22915E-08	7718	1824	922	17816	6134, 829, 6405, 5670, 5054, 7079, 832

- source: Code for the data source. Ex> GO:BP | GO:CC | GO:MF ...
- term_id: ID for the enriched term/functional category
- term_name: Readable name for the enriched term
- adjusted_p_value: Adjusted p-value by FDR
- query_size: The number of unique DEG that are annotated to the data source (the functional category).
- intersection_size: The number of unique DEG that are annotated to the term_id
- term_size: The number of genes of species that are annotated to the term_id.
- effective_domain_size: The number of genes of species that are annotated to the data source (the functional category).
- intersections: list of unique DEG that are annotated to the term_id

5. 4. 2. GO_genes Sheet

The result of associated term_id and DEG analysis result was summarized based on Gene. term_id which associated with specific gene was summarized with stat such as fold change, p-value, volume, normalized value.

source	term_id	term_name	adjusted_p_value	intersection_size	Gene_ID	Transcript_ID	Gene_Symbol	test/control.fc	test/control.logCPM	test/control.raw.pval	test/control.bh.pval	N_control_1	N_control_2	N_test_1	N_test_2
GO:CC	GO:0044444	cytoplasmic part	5.37824E-09	1125	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:CC	GO:0005737	cytoplasm	5.47219E-09	1309	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:BP	GO:0070887	cellular response to ch	6.97255E-05	417	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:BP	GO:0050896	response to stimulus	0.000405505	1045	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:CC	GO:0005829	cytosol	0.078450245	563	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:CC	GO:0005622	intracellular	0.110987379	1522	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:MF	GO:0004060	arylamine N-acetyltrar	0.573292063	1	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:CC	GO:0005575	cellular_component	1	1921	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:BP	GO:0008150	biological_process	1	1824	9	NM_000662,NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:CC	GO:0044459	plasma membrane pa	1.34094E-09	400	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0071944	cell periphery	1.8891E-09	709	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0016020	membrane	0.000132828	1085	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0097458	neuron part	0.000244106	234	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0042995	cell projection	0.000353388	283	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0044425	membrane part	0.000390502	813	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:BP	GO:0050896	response to stimulus	0.000405505	1045	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:BP	GO:0051606	detection of stimulus	1	35	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:BP	GO:0008150	biological_process	1	1824	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0044444	cytoplasmic part	5.37824E-09	1125	34	NM_000016,NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040
GO:CC	GO:0005737	cytoplasm	5.47219E-09	1309	34	NM_000016,NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040
GO:BP	GO:0009888	tissue development	5.79564E-09	305	34	NM_000016,NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040
GO:BP	GO:0032501	multicellular organism	3.22915E-08	922	34	NM_000016,NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040
GO:BP	GO:0048731	system development	3.55854E-08	626	34	NM_000016,NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040
GO:BP	GO:0048513	animal organ develop	3.78565E-08	478	34	NM_000016,NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040

- source: Code for the data source. Ex> GO:BP | GO:CC | GO:MF ...
- term_id: ID for the enriched term/functional category
- term_name: Readable name for the enriched term
- adjusted_p_value: Adjusted p-value by FDR
- intersection_size: The number of unique DEG that are annotated to the term_id

data3.GO_*.gprofiler.png: Top 20 terms of Gene Ontology Enrichment Analysis result were described by dot plot.

(Plotting based on GO_stat)

data3.GO_*.gprofiler.sizefilt.png: After term_size filtering (min=10, max=500), top 20 terms of Gene Ontology Enrichment Analysis result were described by dot plot.

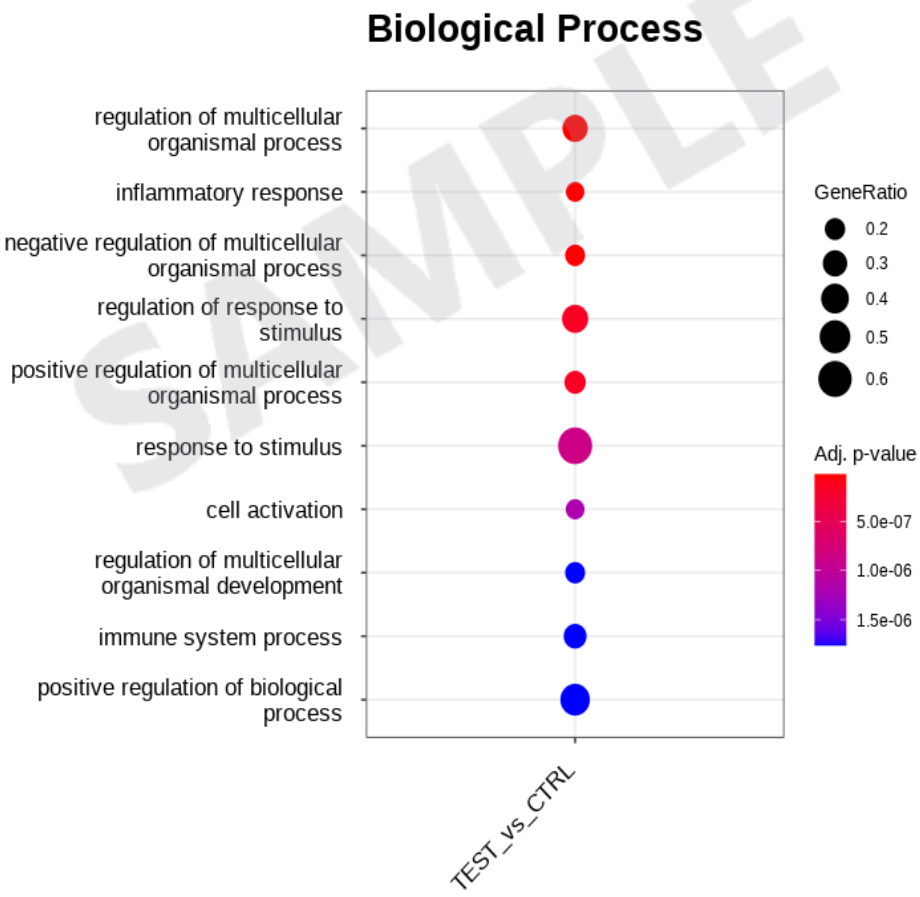
(Plotting based on GO_stat. Please refer to ./gprofiler/data3*.GO/folder.)

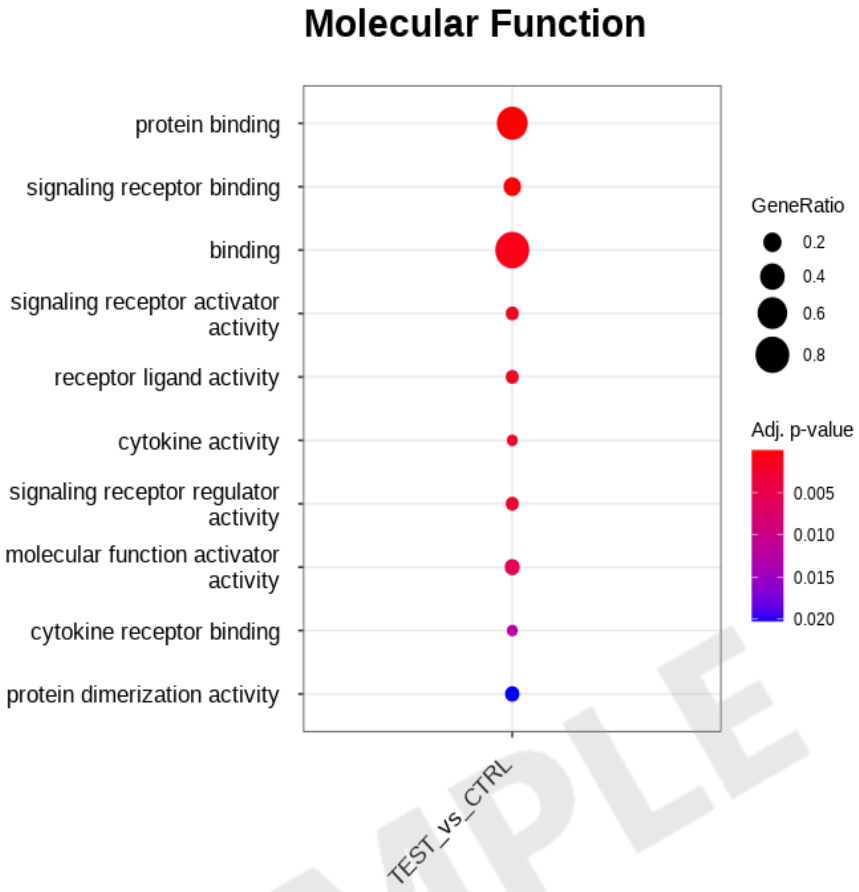
- term_size filtering: The GO Terms that are very large or small do not contribute to interpretability of results, and their statistical significance can be inflated when using certain statistical enrichment methods (e.g., Hypergeometric test).

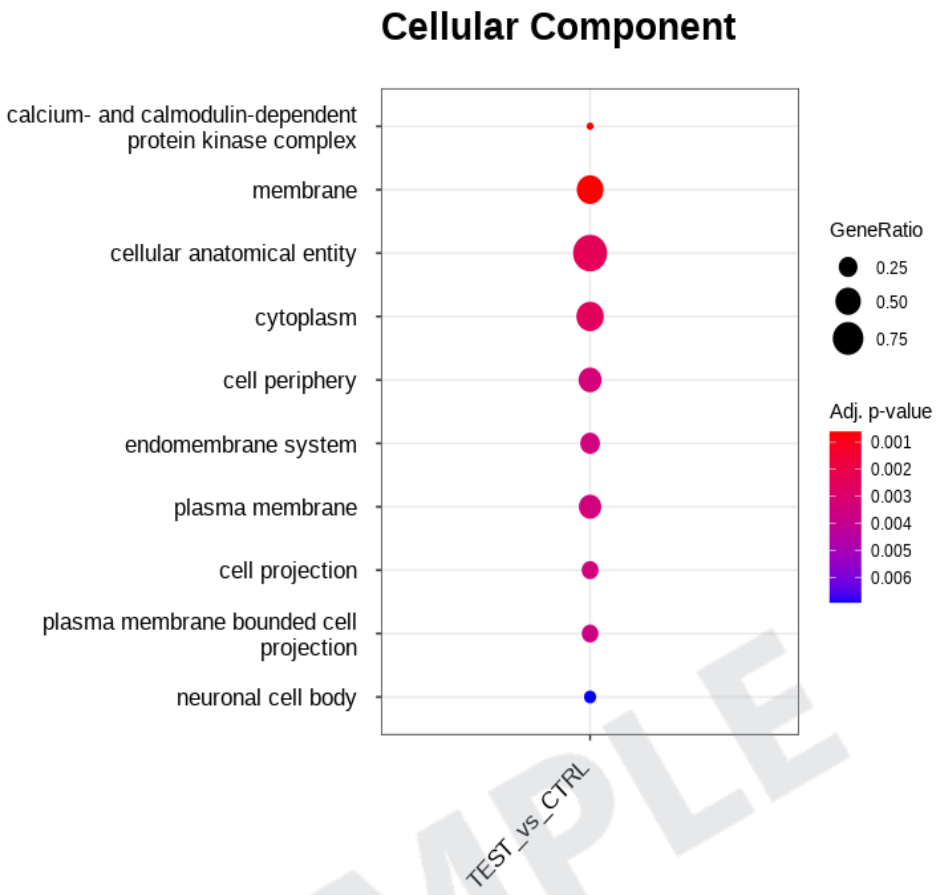
- GeneRatio: GeneRatio is calculated as the ratio of intersection_size and query_size.

The dot plot below shows the results of the enrichment analysis based on Gene Ontology DB for significant genes.

These dot plots are examples for data3.GO_*.gprofiler.png (without term_size filtering).




Molecular Function
protein binding
signaling receptor binding
binding
signaling receptor activator activity
receptor ligand activity
cytokine activity
signaling receptor regulator activity
molecular function activator activity
cytokine receptor binding
protein dimerization activity
TEST_vs_CTRL
GeneRatio
0.2
0.4
0.6
0.8
Adj. p-value
0.005
0.010
0.015
0.020


Cellular Component
calcium- and calmodulin-dependent protein kinase complex
membrane
cellular anatomical entity
cytoplasm
cell periphery
endomembrane system
plasma membrane
cell projection
plasma membrane bounded cell projection
neuronal cell body
TEST_vs_CTRL
GeneRatio
0.25
0.50
0.75
Adj. p-value
0.001
0.002
0.003
0.004
0.005
0.006

5. 5. KEGG Enrichment Analysis

(Refer to Path: result_RNAseq/DEG_result/[DataSet]/KEGG_view)

KEGG database contains various types of omics data such as molecular information (genome sequence, structure), chemical information (Metabolism, Glycans, Lipids etc.), molecular interaction information(physical interaction, co-expression).

KEGG pathway homepage: http://www.kegg.jp/kegg/pathway.html

KEGG pathway viewer provides the pathway map colored by fold change for significantly expressed genes by each comparison pair using pathway map information of given species. And it also gives you the enrichment test result and the heatmap of that on the main page. When clicking the KEGG_pathway.html, you can see the heatmap of enrichment test result for each pathway term. The detailed results for enrichment analysis are provided in the following sheets of data3.

Enrichment test result was summarized at each sheet of DEG result(data3-*.xlsx file) by 2 forms below.

- KEGG_stat
- KEGG_genes

The following heatmap shows the results of the enrichment analysis for each pathway term. The gradient legend shows the level of enrichment raw p-value from the modified fisher's exact test to determine the enrichment of each gene from the gene set. The raw p-value lower than 0.05 means that the pathway has been significantly enriched. By clicking the block of each pathway of pairs for comparison on the table, it would display the colored pathway in html format.

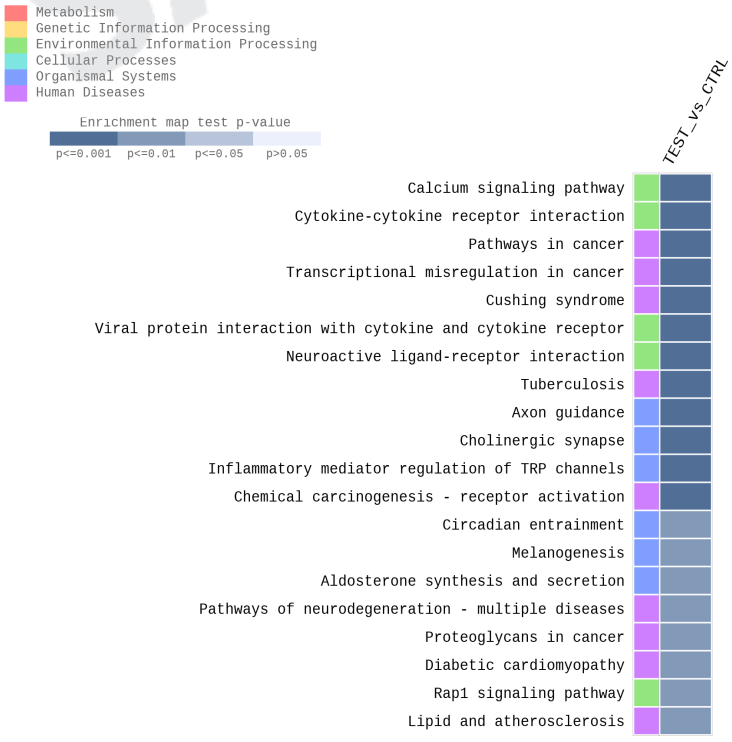


Figure 14. Result of gene-set enrichment analysis (p-value top 20)

5. 5. 1. KEGG HTML Viewer

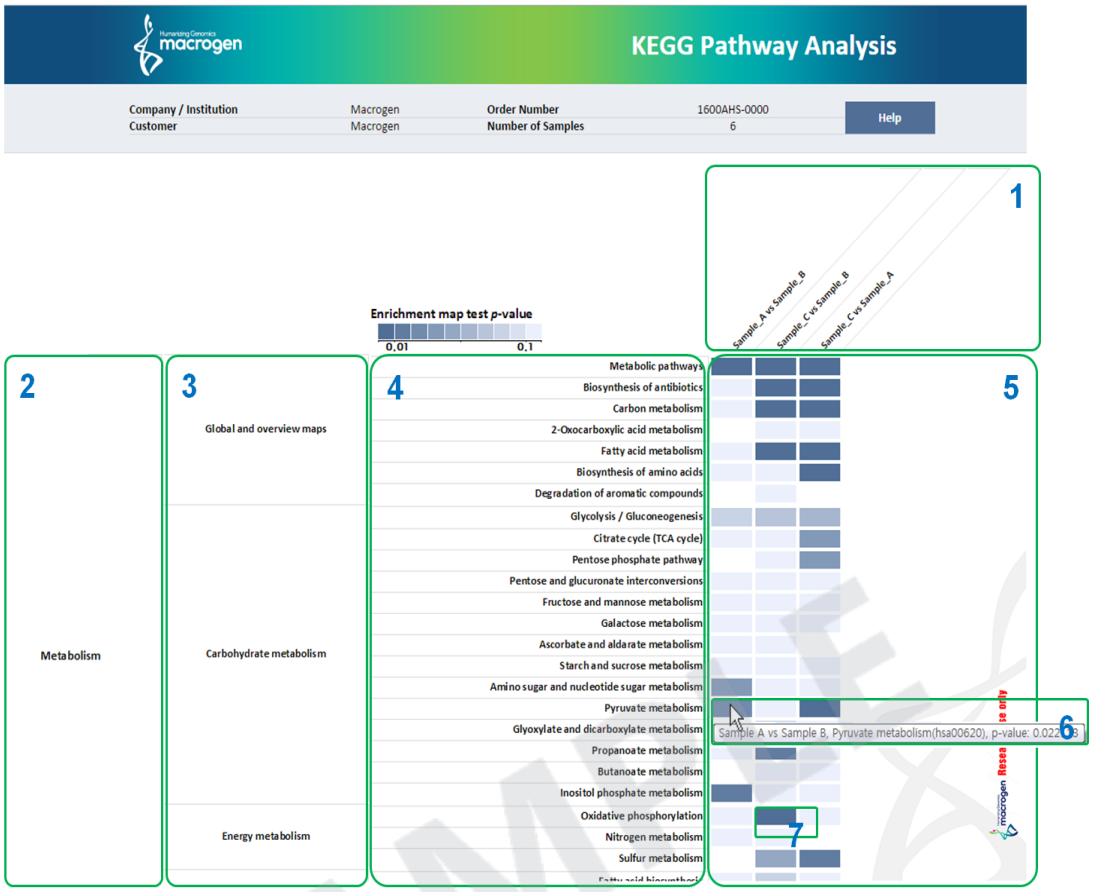


Figure 15. Description of KEGG Viewer frame

- Block 1: Differential expression gene combinations.
- Block 2: Metabolism, Cellular process, Environmental information processing, Genetic information processing, Organismal system
- Block 3: Categorized pathway map
- Block 4: Pathway map name
- Block 5: Heatmap of KEGG enrichment map score (p-value). (empty box means that there is not matched gene)
- Block 6: Following information are separated with comma and can be checked by putting mouse over. (Combination information , Pathway name , KEGG enrichment map score (p-value))
- Block 7: New window pops up when color box is clicked.

- "Global and overview maps" is not directly drawing the data saved from HTML. It directly shows genes from KEGG homepage. This may slow down the loading time.

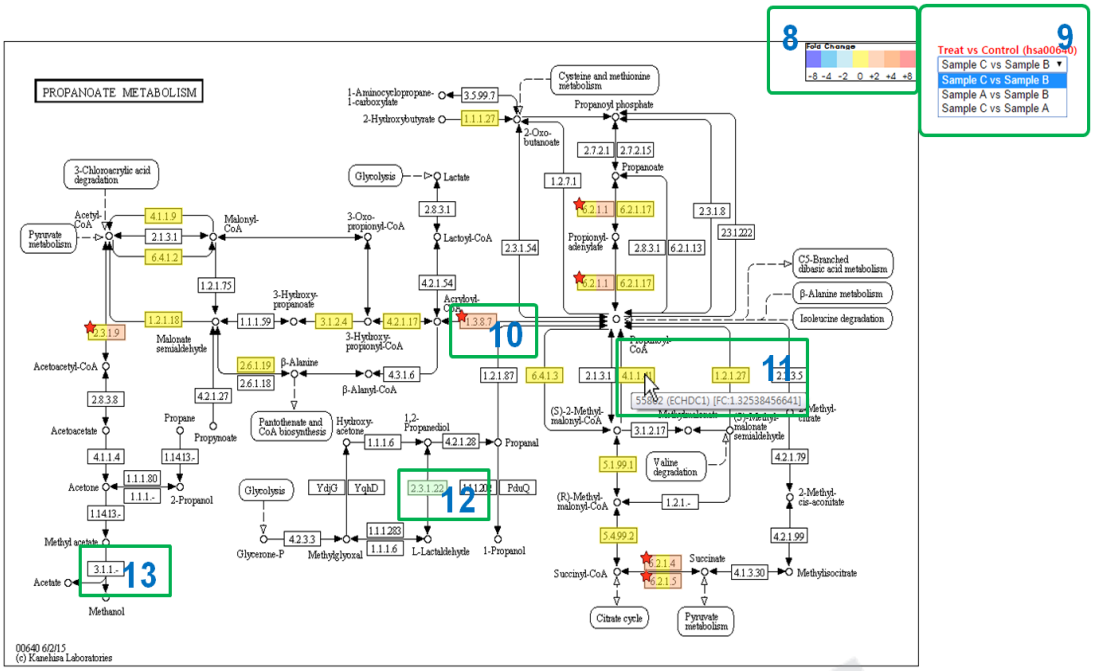


Figure 16. Description of KEGG pathway map frame

- Block 8: Fold change values of DEG are shown in colors.
- Block 9: You can change to different combination within the current KEGG pathway page. The combination shown in the box is currently shown combination.
- Block 10: Significant pathway module is marked with red star (based on data3 file of significance).
- Block 11: The name and fold change value of the gene are shown when mouse is over. (genes are separated with comma). If the gene id exists but there is no FC value on the title of module, then the gene does not exist in data2 file that is processed QC filtering step.
- Block 12: Green color box of pathway map is modules that are not mapped. Gene is in the pathway map but the expression is not shown.
- Block 13: White box of pathway map is module that is not relevant to the species.

5. 5. 2. KEGG_stat Sheet

This table shows the enrichment analysis result for each pathway term. You can find this table in the KEGG stat sheet of data3 file.

Example of KEGG pathway enrichment analysis result

MapID	MapName	Number_of_SigGenes	Genes	Sig.NotIn.KEGG	Genome.In.KEGG	Genome.NotIn.KEGG	PValue	Bonferroni	FDR
01100	Metabolic pathways	86	10229,10622,10797,10998,1109	281	1220	58263	8.6357E-61	2.29709E-58	2.29709E-58
01130	Biosynthesis of antibiotics	25	113675,1491,2026,2027,22934,	342	214	59269	5.67107E-22	1.5085E-19	7.54252E-20
05203	Viral carcinogenesis	22	1021,1026,1030,3017,3106,313	345	206	59277	1.32494E-18	3.52434E-16	1.17478E-16
04151	PI3K-Akt signaling pathway	25	10110,1021,1026,1280,2057,22	342	347	59136	1.79176E-17	4.76608E-15	1.19152E-15
04142	Lysosome	18	10577,138050,1514,175,1777,2	349	123	59360	2.54025E-17	6.75707E-15	1.35141E-15
05200	Pathways in cancer	26	1021,1026,1030,11211,2034,22	341	398	59085	3.16913E-17	8.42988E-15	1.40498E-15
05205	Proteoglycans in cancer	20	1026,11211,1514,1839,3678,40	347	204	59279	2.73765E-16	7.28215E-14	1.04031E-14
01230	Biosynthesis of amino acids	14	113675,1491,2026,2027,22934,	353	74	59409	9.20432E-15	2.44835E-12	3.06044E-13
05166	HTLV-I infection	20	1026,1030,11211,1958,2114,23	347	261	59222	1.77887E-14	4.7318E-12	5.25756E-13
01200	Carbon metabolism	15	113675,2026,2027,22934,230,2	352	113	59370	6.6255E-14	1.76238E-11	1.76238E-12
04010	MAPK signaling pathway	19	1649,1847,2248,2261,2264,235	348	257	59226	1.62278E-13	4.3166E-11	3.92418E-12
04390	Hippo signaling pathway	16	11211,126374,1490,166824,271	351	154	59329	2.11892E-13	5.63633E-11	4.69694E-12
04115	p53 signaling pathway	12	1021,1026,27113,5054,51246,5	355	68	59415	2.40037E-12	6.38498E-10	4.91153E-11
04145	Phagosome	14	10381,11151,1514,155066,3106	353	155	59328	4.8863E-11	1.29976E-08	9.28397E-10
05206	MicroRNAs in cancer	17	1021,1026,2261,3162,3371,367	350	297	59186	1.46683E-10	3.90177E-08	2.60118E-09
04550	Signaling pathways regulating pluripotency	13	11211,2261,2264,3625,5600,56	354	142	59341	2.51263E-10	6.6836E-08	4.17725E-09
04668	TNF signaling pathway	12	1051,1906,2353,3726,4323,468,	355	110	59373	2.6984E-10	7.17774E-08	4.2222E-09
05168	Herpes simplex infection	14	2353,3106,3133,3665,406,4938,	353	186	59297	4.01978E-10	1.06926E-07	5.94034E-09
00260	Glycine, serine and threonine metabolism	9	113675,1491,211,23464,2593,2	358	40	59443	5.52529E-10	1.46973E-07	7.73541E-09
04110	Cell cycle	12	1021,1026,10274,1028,1030,53	355	124	59359	8.7649E-10	2.33146E-07	1.16573E-08
04015	Rap1 signaling pathway	14	2248,2261,2264,2770,5600,560	353	211	59272	1.70866E-09	4.54503E-07	2.1643E-08
04068	FoxO signaling pathway	12	10110,1026,1030,10365,23710,	355	134	59349	1.87658E-09	4.9917E-07	2.26895E-08
04060	Cytokine-cytokine receptor interaction	15	2057,3576,3590,3625,51330,51	352	265	59218	2.64579E-09	7.03781E-07	3.05992E-08
05169	Epstein-Barr virus infection	13	1026,10622,3106,3133,3315,37	354	201	59282	1.01035E-08	2.68752E-06	1.1198E-07

- MapID: KEGG map ID
- MapName: KEGG map name
- Number_of_SigGenes: Number of (uniquely) differentially expressed genes that are included in the pathway
- Genes: List of gene that are included in the pathway (comma delimited)
- Sig.NotIn.KEGG: Number of (uniquely) differentially expressed genes that are not included in the pathway
- Genome.In.KEGG: Number of genes that are associated to this pathway among the genes in given species
- Genome.NotIn.KEGG: Number of genes that are not associated to this pathway among the genes in given species
- PValue: Raw p-value from the modified fisher's exact test
- Bonferroni: Corrected p-value by bonferroni method
- FDR: Corrected p-value by FDR method

5. 5. 3. KEGG_genes Sheet

This table shows the pathway enrichment analysis result according to gene. You can find this table in the KEGG genes sheet of data3 file.

Example of KEGG pathway enrichment analysis result sorted by gene

InID	MapID	MapName	PValue	Bonferroni	FDR	Gene	B/A.fc	B/A.volume	N_A	N_B
22801	04151	PI3K-Akt signali	5.34874E-08	1.12324E-05	5.34874E-07	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	04510	Focal adhesion	0.002603438	0.546721969	0.008040029	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	04512	ECM-receptor ir	0.001875844	0.393927235	0.006353665	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	04810	Regulation of a	0.002975034	0.62475714	0.009054451	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	05410	Hypertrophic ca	9.33482E-05	0.01960313	0.000502644	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	05412	Arrhythmogenic	0.017901038	1	0.042238405	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	05414	Dilated cardiom	0.002059901	0.432579199	0.006655065	ITGA11	1.706859	11.100807	10.721833	11.493176
3017	05034	Alcoholism	8.28056E-07	0.000173892	6.68814E-06	HIST1H2BD	1.647010	11.092905	10.738818	11.458667
3017	05203	Viral carcinoger	2.52581E-05	0.005304204	0.000156006	HIST1H2BD	1.647010	11.092905	10.738818	11.458667
3017	05322	Systemic lupus	2.5681E-06	0.0005393	1.85966E-05	HIST1H2BD	1.647010	11.092905	10.738818	11.458667
441024	00670	One carbon poo	1	1	1	MTHFD2L	1.747046	9.561974	9.167981	9.972899
441024	01100	Metabolic pathv	5.97272E-15	1.25427E-12	1.79181E-13	MTHFD2L	1.747046	9.561974	9.167981	9.972899
89853	04144	Endocytosis	0.033602909	1	0.075877535	FAM125B	1.677441	9.607461	9.241573	9.987835
7869	04360	Axon guidance	0.005283715	1	0.014994327	SEMA3B	-2.103133	8.787416	9.340035	8.267495
10135	00760	Nicotinate and	8.87463E-05	0.018636723	0.00049044	NAMPT	1.620452	10.752957	10.410395	11.106791
10135	01100	Metabolic pathv	5.97272E-15	1.25427E-12	1.79181E-13	NAMPT	1.620452	10.752957	10.410395	11.106791
534	00190	Oxidative phosp	1	1	1	ATP6V1G2	-1.647407	8.093609	8.461714	7.741517
534	01100	Metabolic pathv	5.97272E-15	1.25427E-12	1.79181E-13	ATP6V1G2	-1.647407	8.093609	8.461714	7.741517
534	04145	Phagosome	3.15039E-07	6.61582E-05	2.87644E-06	ATP6V1G2	-1.647407	8.093609	8.461714	7.741517

- InID: Matching key ID (ex. Entrez GeneID)
- MapID: KEGG map ID
- MapName: KEGG map name
- PValue: Raw p-value from the modified fisher's exact test
- Bonferroni: Corrected p-value by bonferroni method
- FDR: Corrected p-value by FDR method

6. Data Download Information

6. 1. Raw Data

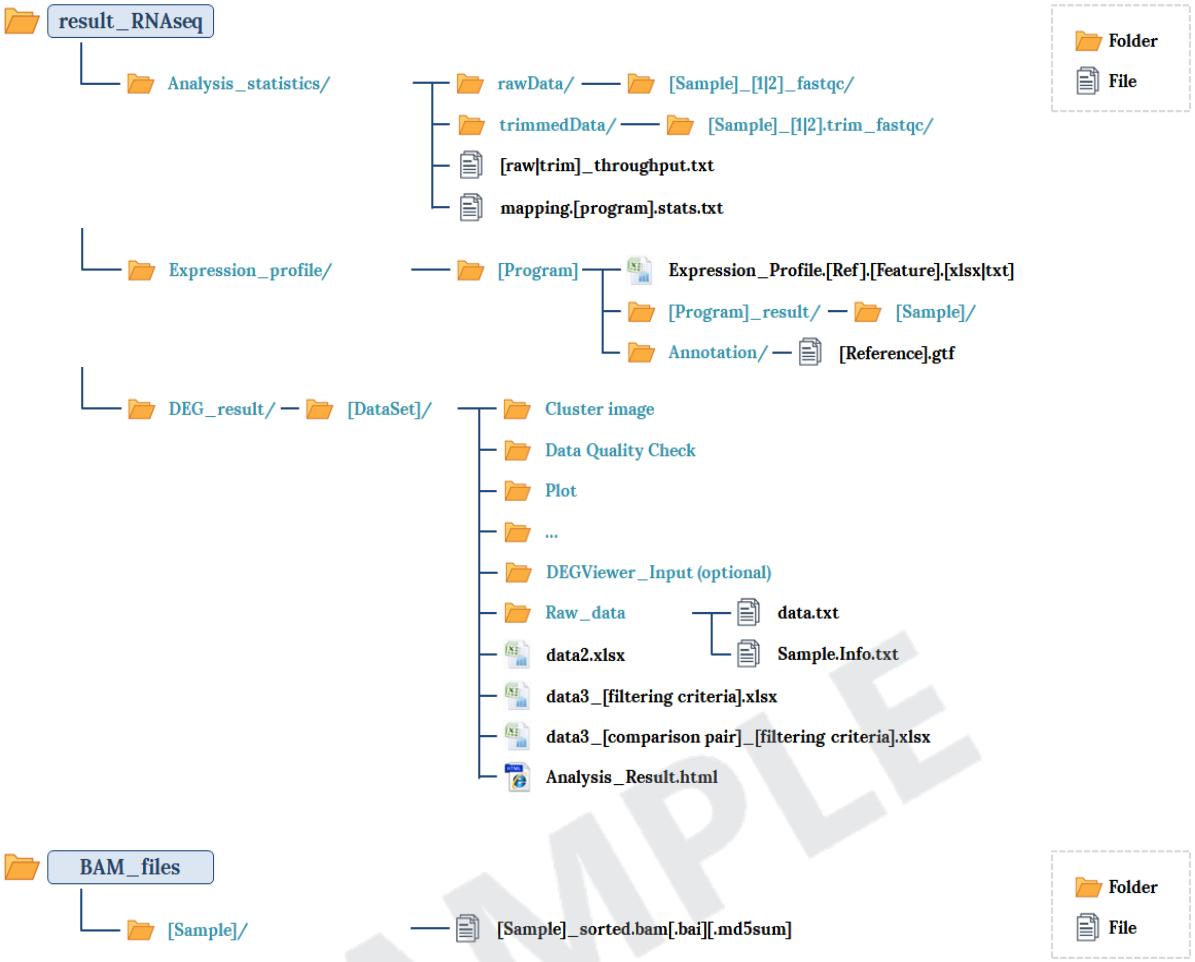
Raw data is the FASTQ file that isn't trimmed adapter sequence.

Download link	File size	md5sum
MG_CTRL_1_1.fastq.gz	915.53M	9caba6b113df1ad3c71d337f6aade342
MG_CTRL_1_2.fastq.gz	965.45M	84f806e22fcf94e26255e6363d775271
MG_CTRL_2_1.fastq.gz	918.55M	0014c4c56cc3aaef70a5d13e5e582682
MG_CTRL_2_2.fastq.gz	976.2M	31d2db72e2a8b9a1eca278ac46a98432
MG_CTRL_3_1.fastq.gz	926.58M	a87eb1f8fb8dacd137477ab0f57777d1
MG_CTRL_3_2.fastq.gz	981.42M	ccbdcbf25c80b4c79e543a66c39a6d1b
MG_TEST_1_1.fastq.gz	922.04M	c9ab662093cffd280bee590cf446b469
MG_TEST_1_2.fastq.gz	975.65M	3d928bf055ced6f22e1d23964ac221dd
MG_TEST_2_1.fastq.gz	917.4M	0ff808744648877e14a2c76bc3daab41
MG_TEST_2_2.fastq.gz	970.31M	e8404bc235277a1de1e16b5d1a63708e
MG_TEST_3_1.fastq.gz	914.96M	be9c425871e085591164949ebeba6fe0
MG_TEST_3_2.fastq.gz	966.77M	7ffdb61d18b462f8c3fb186f3574416f

- 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
HN00000000_result_RNAseq.zip (md5sum: 87a330c1f208e53e17a99e917557ceda)	205.36M
HN00000000_BAM_files.tar (md5sum: 1ed766a221ccf08fcfc15bef084fc238)	8.82G



⚠ 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

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

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. TopHat version v2.1.1, bowtie2 2.5.1

LINK http://ccb.jhu.edu/software/tophat/index.shtml

Tophat is a tool that maps transcriptome sequencing data on mammalian-sized genome using bowtie2. It uses this mapping results to provide provisional exon location and exon junctions. In order for increased mapping increase at exon binding site, it accounts for GT-AT's two nucleotide pattern information.

7. 2. 4. Cufflinks version v2.2.1

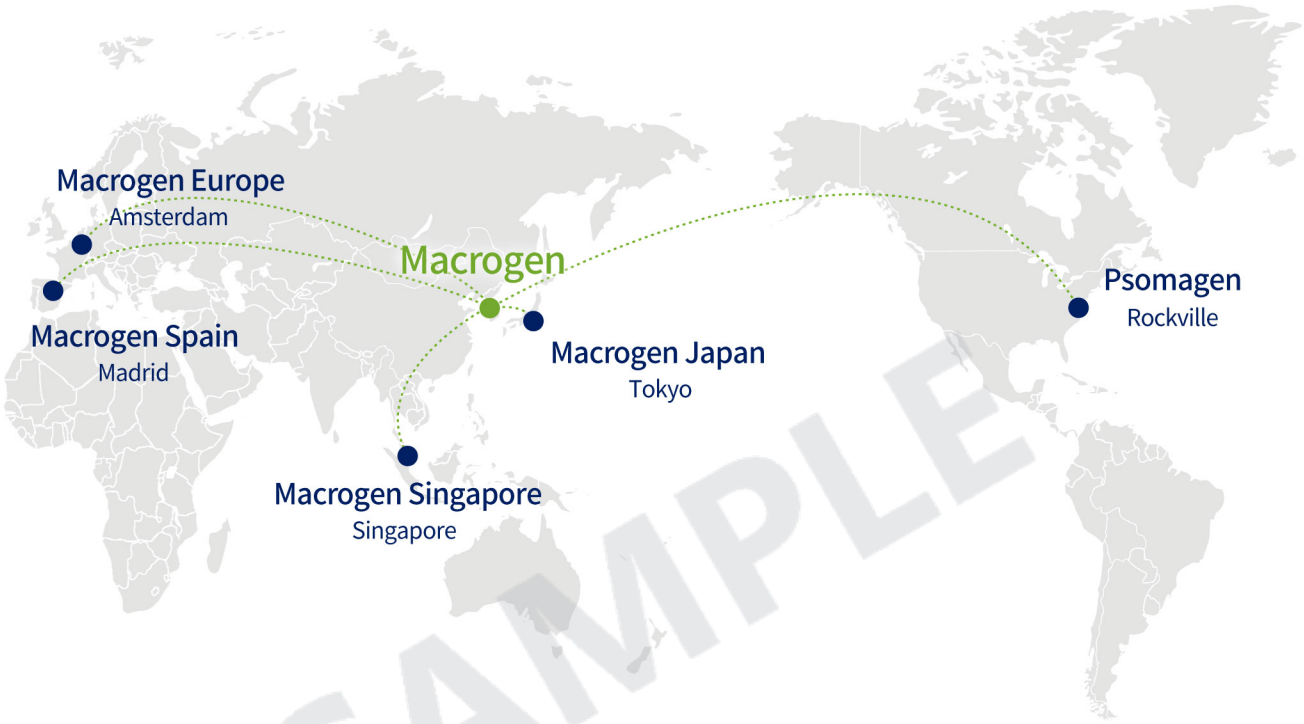
LINK http://cole-trapnell-lab.github.io/cufflinks/

Cufflink is a sequence assembly program that connects reads from the mapping results using the Tophat aligner. It can predict the expression level of the assembled transcriptomes and provides results for cuffdiff, which shows difference in expression between samples.

7. 3. References

1. BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, btu170.
2. TRAPNELL, Cole; PACHTER, Lior; SALZBERG, Steven L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25.9: 1105-1111.
3. KIM, Daehwan, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 2013, 14.4: R36.
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. LI, Heng, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16: 2078-2079.
6. TRAPNELL, Cole, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 2010, 28.5: 511-515
7. ROBERTS, Adam, et al. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology, 2011, 12.3: R22.
8. BI, Yong-Mei, et al. High throughput RNA sequencing of a hybrid maize and its parents shows different mechanisms responsive to nitrogen limitation. BMC genomics, 2014, 15.1: 77.
9. TRAPNELL, Cole, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology, 2013, 31.1: 46-53.
10. TRAPNELL, Cole, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 2012, 7.3: 562-578.
11. MORTAZAVI, Ali, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 2008, 5.7: 621-628.
12. RAUDVERE, Uku, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research, 2019.

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