

Caenorhabditis elegans Transcriptome Sequencing

Report

November 2018



Project Information

Client Name	Tester
Company/Institution	Macrogen
Order Number	1800KHF-0000
Species	<i>Caenorhabditis elegans</i>
Reference	WBcel235
Annotation	WS264
Read Length	101
Number of Samples	6
Library Kit	TruSeq Stranded mRNA LT Sample Prep Kit
Library Protocol	TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E
Reagent	TruSeq rapid SBS kit or Truseq SBS Kit v4
Sequencing Protocol	HiSeq 2500 System User Guide Document # 15035786 v02 HCS 2.2.70
Type of Sequencer	HiSeq 2500
Sequencing Control Software	HCS 2.2.70
Comment	HiSeq 2500

Project Results Summary

In this study, *Caenorhabditis elegans* 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.

The novel transcripts and novel alternative splicing transcripts were discovered during the assembly.

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.

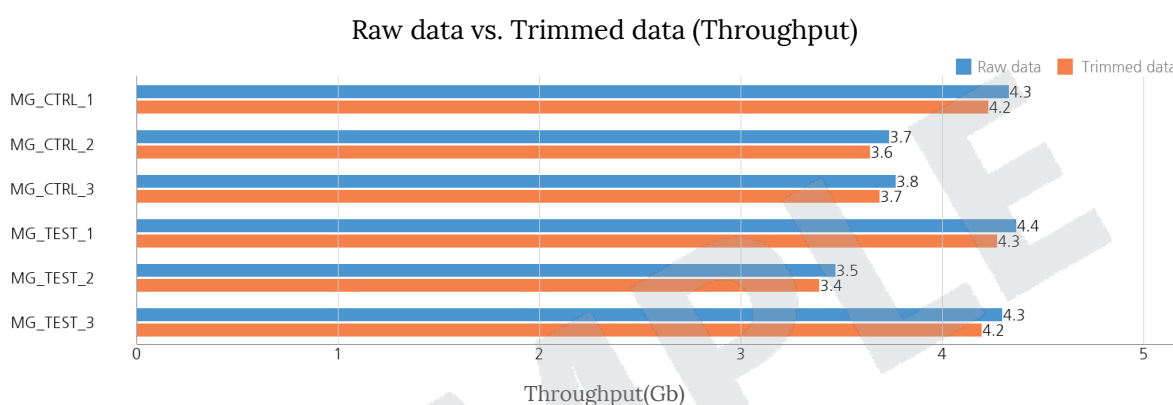


Figure 1. Throughput output of Raw and Trimmed data

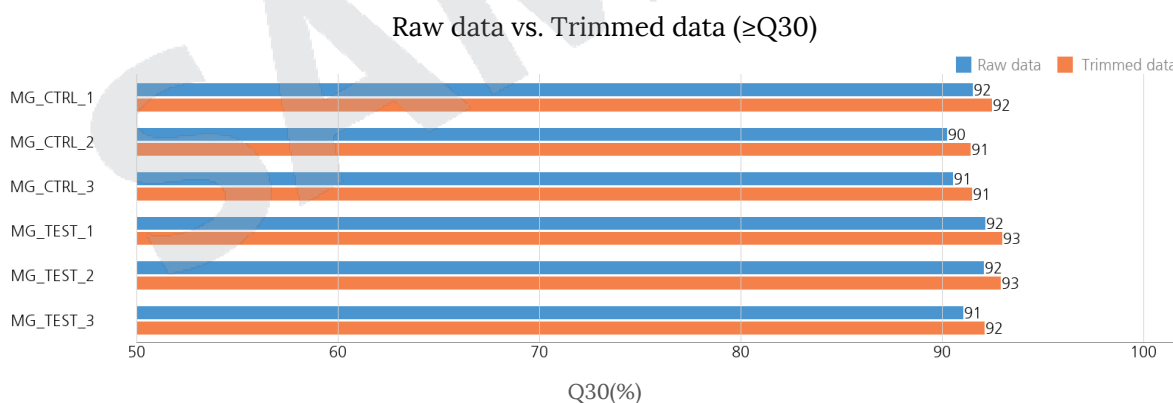


Figure 2. Q30 score of Raw and Trimmed data

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

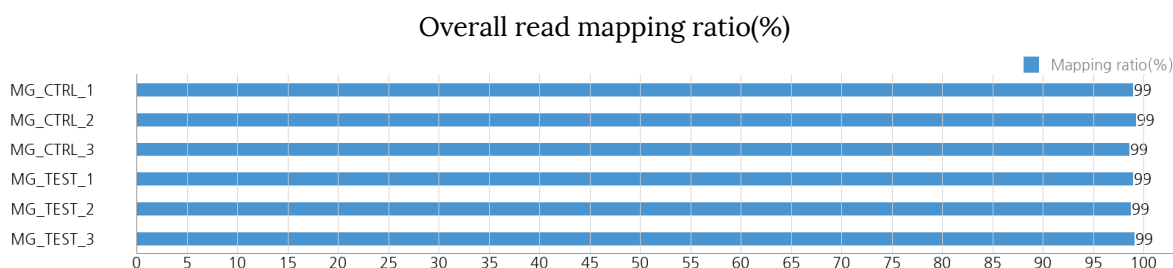


Figure 3. Overall read mapping ratio(%)

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

DEG (Differentially Expressed Genes) analysis was performed on a comparison pair (MG_TEST_vs_MG_CTRL) as requested using DESeq2. The results showed 564 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.

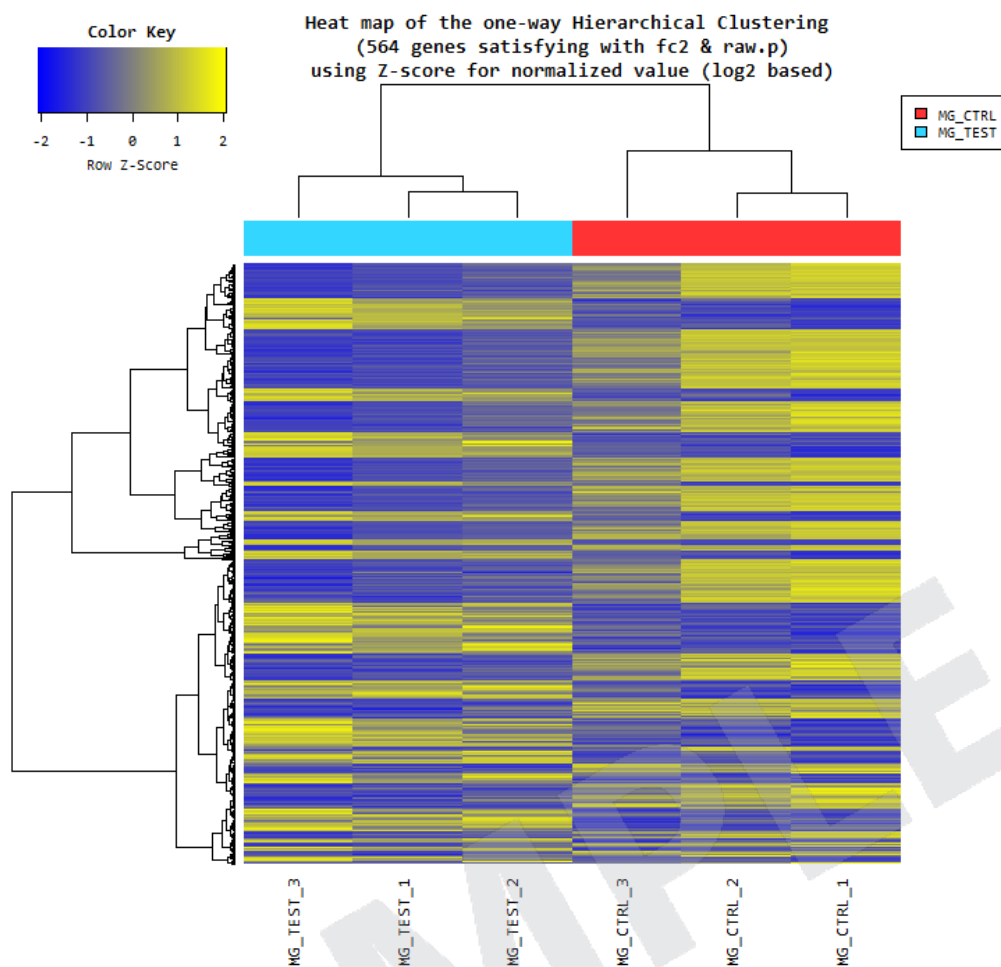


Figure 4. Heatmap for DEG list

DEG list was further analyzed with DAVID tool(<http://david.abcc.ncifcrf.gov/>) 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.

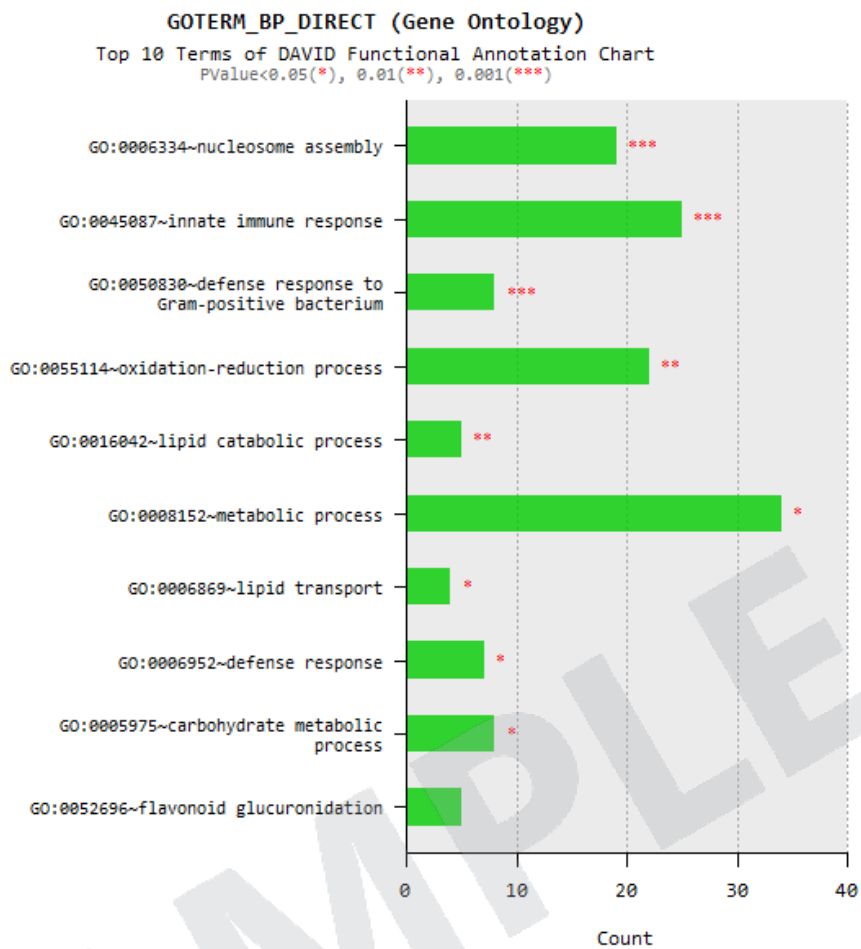


Figure 5. Gene Ontology terms related to Biological Process

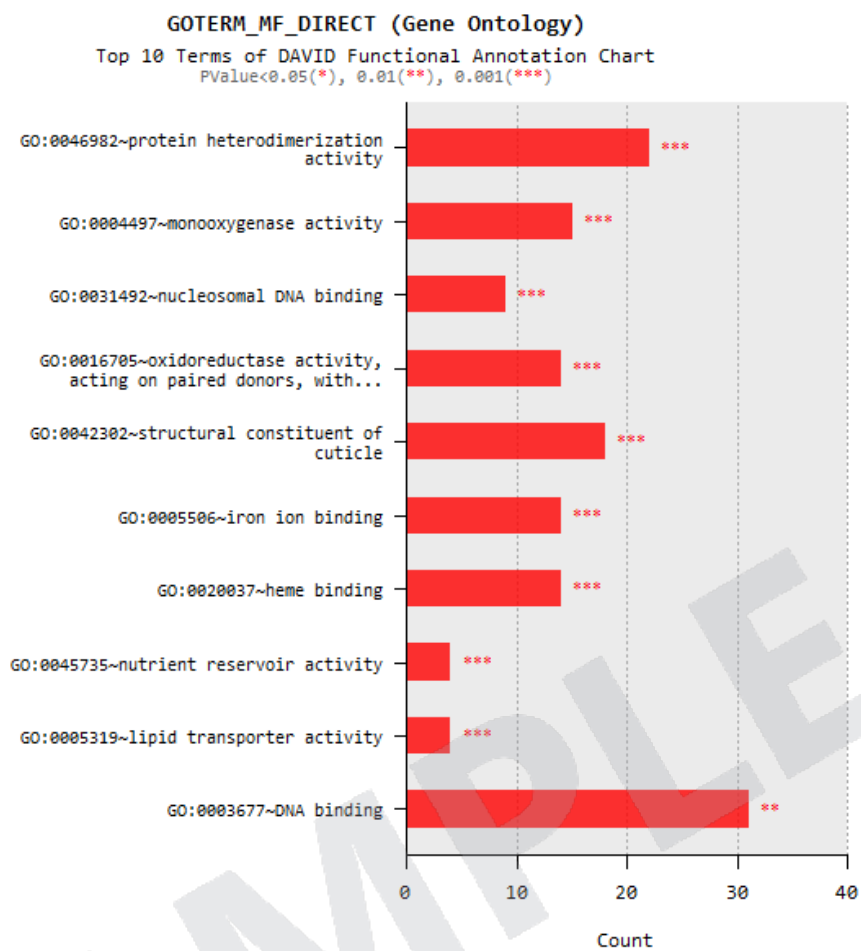


Figure 6. Gene Ontology Terms related to Molecular Function

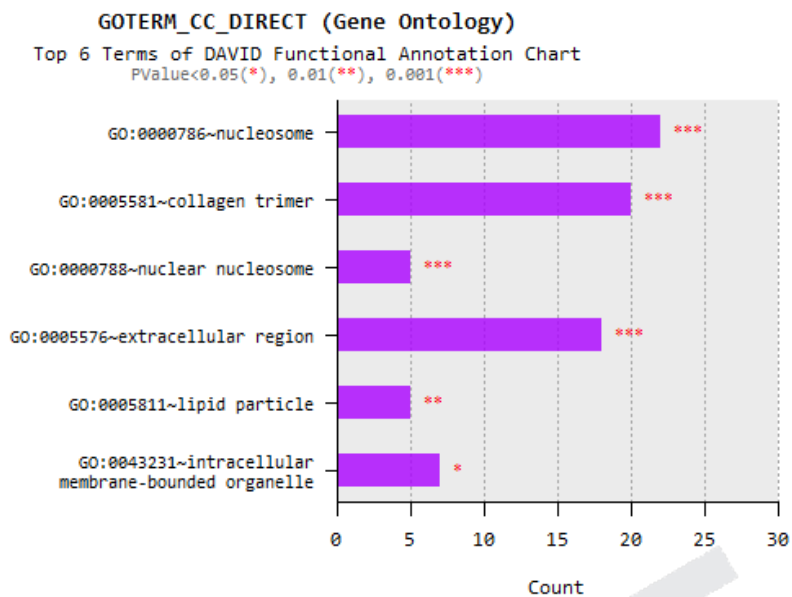


Figure 7. Gene Ontology Terms related to Cellular Component

In addition, novel transcript and novel alternative splicing transcripts were found each sample. (Please refer to the main body of this report for detailed explanations.)

SAMPLE

Table of Contents

Project Information	02
Project Results Summary	03
1. Experimental Methods and Workflow	10
2. Analysis Methods and Workflow	11
3. Summary of Data Production	12
3. 1. Raw Data Statistics	12
3. 2. Average Base Quality at Each Cycle	13
3. 3. Trimming Data Statistics	14
3. 4. Average Base Quality at Each Cycle after Trimming	15
4. Reference Mapping and Assembly Results	16
4. 1. Mapping Data Statistics	16
4. 2. Transcript Assembly and Expression Profiling based on Reference Genome	17
4. 3. Prediction of Novel Transcripts/Alternative Splicing Transcripts	19
5. Differentially Expressed Gene Analysis Results	25
5. 1. Data Analysis Quality Check and Preprocessing	25
5. 2. Differentially Expressed Gene Analysis Workflow	30
5. 3. Significant Gene Results	32
5. 4. DAVID Gene-Set Enrichment Analysis	36
5. 5. KEGG Enrichment Analysis	41
6. Data Download Information	46
6. 1. Raw Data	46
6. 2. Analysis Results	46
7. Appendix	49
7. 1. Phred Quality Score Chart	49
7. 2. Programs used in Analysis	50
7. 3. References	51

1. Experimental Methods and Workflow

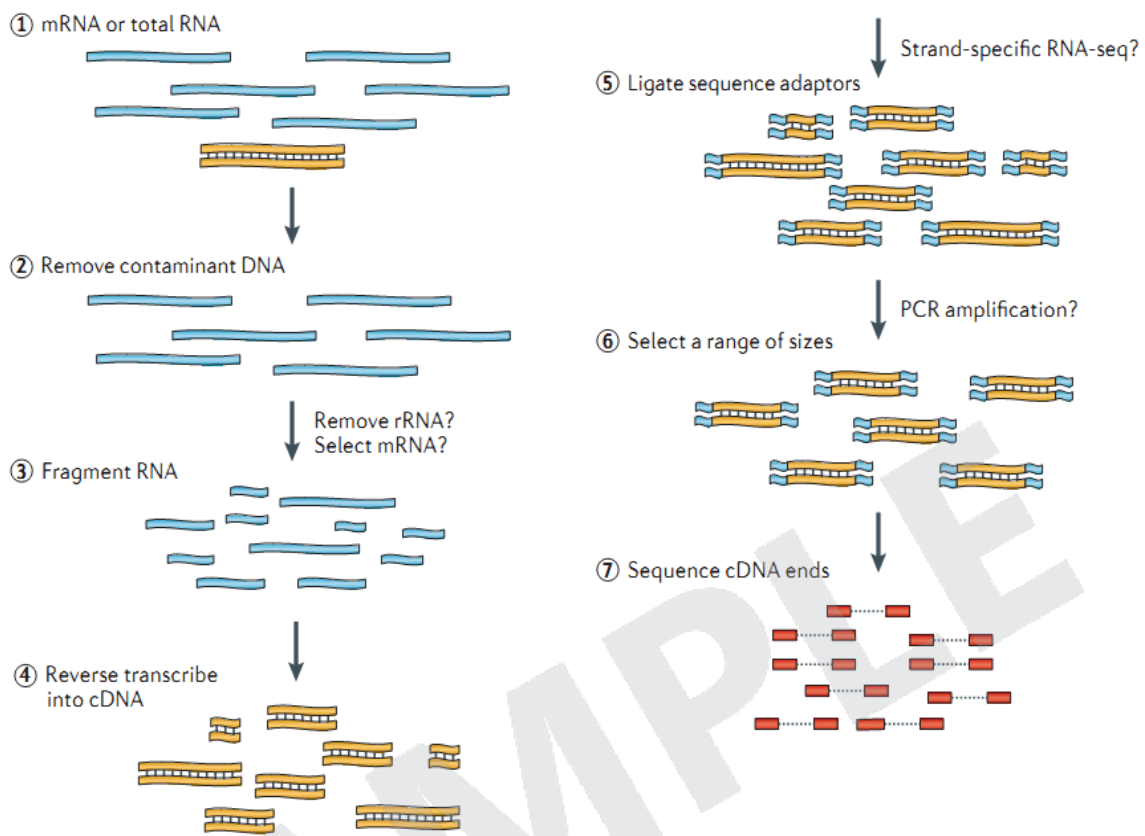


Figure 8. RNA Sequencing Experiment Workflow

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 adaptors 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 HISAT2, splice-aware aligner.
- 4) Transcript is assembled by StringTie with aligned reads. This process provides information of known transcripts, novel transcripts, and alternative splicing transcripts.
- 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.
- 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/rawData/raw_throughput.stats)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 6 samples. For example, in MG_CTRL_1, 42,854,214 reads are produced, and total read bases are 4.3Gbp. The GC content (%) is 47.55% and Q30 is 91.5%.

Table 1. Raw data stats

Index	Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
1	MG_CTRL_1	4,328,275,614	42,854,214	47.55	95.80	91.50
2	MG_CTRL_2	3,733,907,380	36,969,380	47.28	95.07	90.21
3	MG_CTRL_3	3,766,301,918	37,290,118	47.42	95.27	90.51
4	MG_TEST_1	4,365,852,462	43,226,262	47.22	96.12	92.12
5	MG_TEST_2	3,467,681,076	34,333,476	47.55	96.07	92.03
6	MG_TEST_3	4,296,524,042	42,539,842	47.03	95.50	91.04

(* 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>

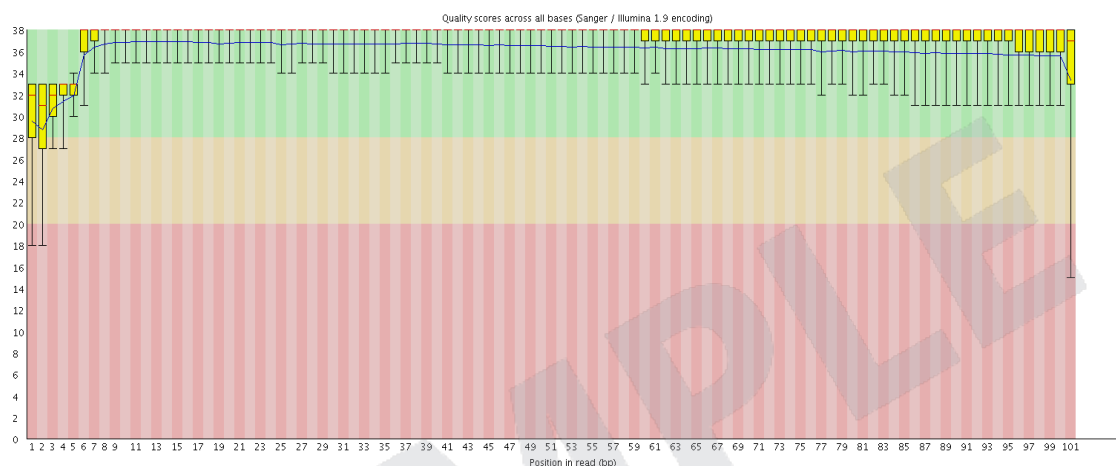


Figure 10. Read quality at each cycle of MG_CTRL_1 (read1)

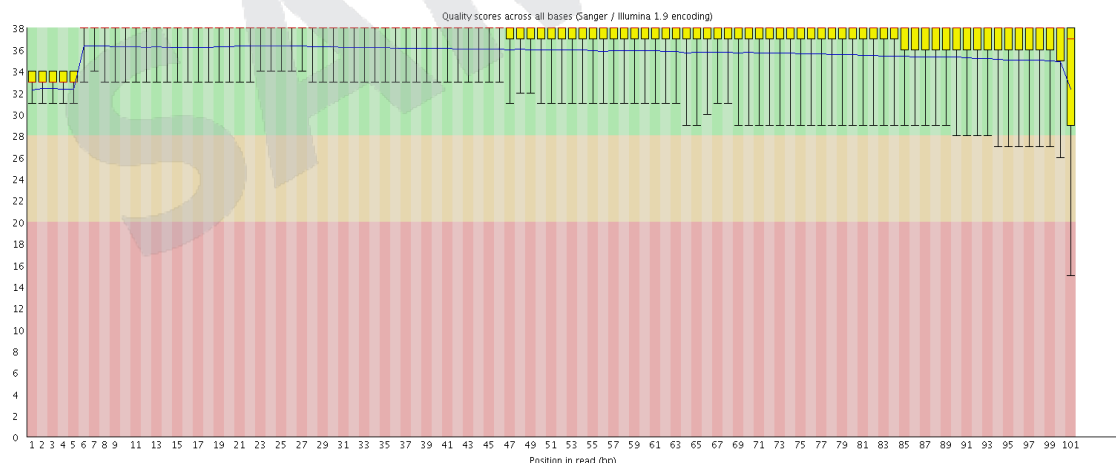


Figure 11. Read quality at each cycle of MG_CTRL_1 (read2)

- 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/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	MG_CTRL_1	4,227,003,626	42,288,760	47.52	96.47	92.44
2	MG_CTRL_2	3,637,085,474	36,338,480	47.25	95.90	91.39
3	MG_CTRL_3	3,685,541,399	36,789,176	47.38	95.94	91.45
4	MG_TEST_1	4,269,560,577	42,713,440	47.19	96.72	92.97
5	MG_TEST_2	3,387,328,325	33,930,482	47.53	96.67	92.89
6	MG_TEST_3	4,193,101,578	41,919,310	47.00	96.22	92.07

- 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.

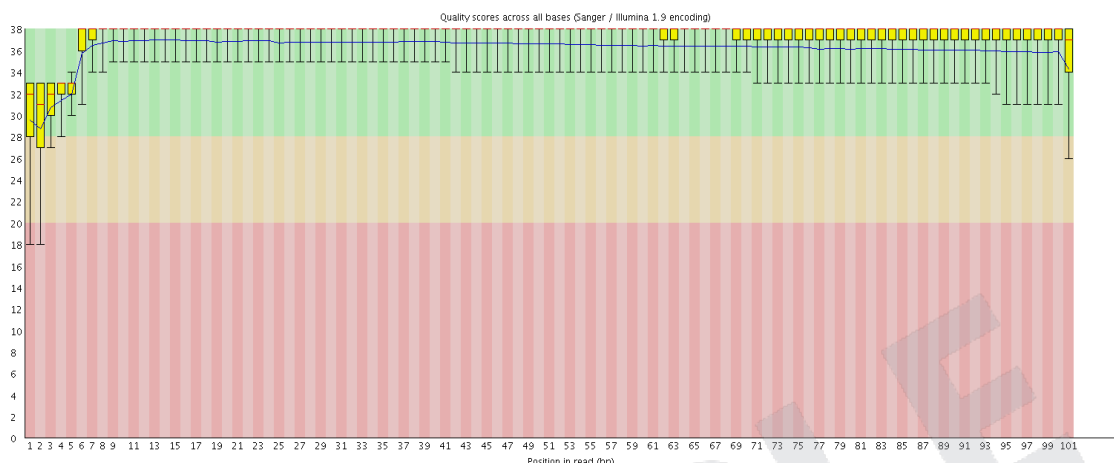


Figure 12. Average base quality of MG_CTRL_1 (read1) at each cycle after trimming

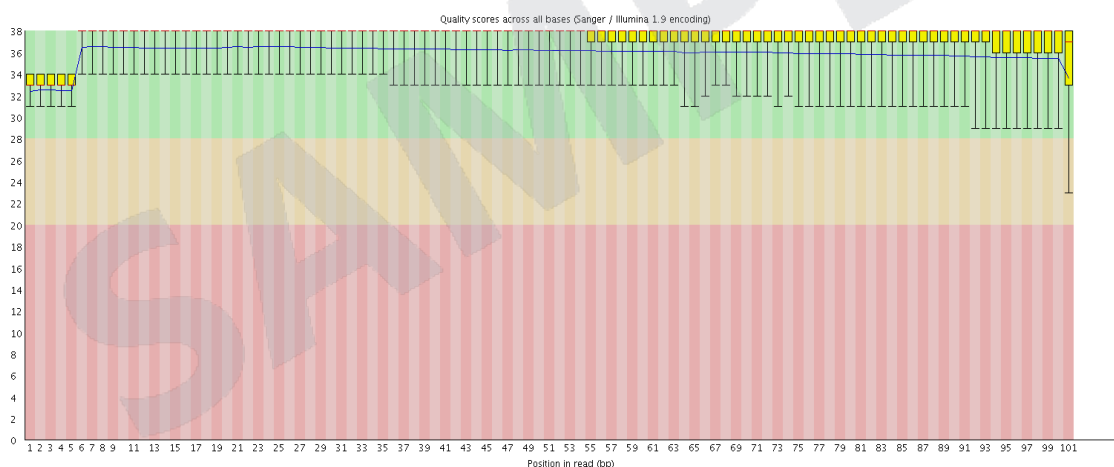


Figure 13. Average base quality of MG_CTRL_1 (read2) at each cycle after trimming

- 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.hisat.stats)

In order to map cDNA fragments obtained from RNA sequencing, WBcel235 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 (%)
MG_CTRL_1	42,288,760	42,087,125 (99.52%)	201,635 (0.48%)
MG_CTRL_2	36,338,480	36,114,446 (99.38%)	224,034 (0.62%)
MG_CTRL_3	36,789,176	36,593,363 (99.47%)	195,813 (0.53%)
MG_TEST_1	42,713,440	42,514,632 (99.53%)	198,808 (0.47%)
MG_TEST_2	33,930,482	33,762,426 (99.5%)	168,056 (0.5%)
MG_TEST_3	41,919,310	41,679,256 (99.43%)	240,054 (0.57%)

- 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. Transcript Assembly and Expression Profiling based on Reference Genome

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/RSEM/Expression_Profile.WBcel235.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		BM	
						Read_Count	FPKM	Read_Count	FPKM
NM_001302545	14	AAAMP	angio associated migratory cell prote	chr2:219128852-219134	1835	898	987	12.220251	12.415353
NM_001087	14	AAAMP	angio associated migratory cell prote	chr2:219128852-219134	1832	4678	6437	63.774289	81.140015
NM_001166679	15	ANAT	aralkylamine N-acetyltransferase, tra	chr17:74449433-744661	1913	46	30	0.599741	0.352587
NR_110548	15	ANAT	aralkylamine N-acetyltransferase, tra	chr17:74463830-744661	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	2485	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 pro	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	1.368953	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/HTseq/Expression_Profile.WBcel235.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	ALPL2	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


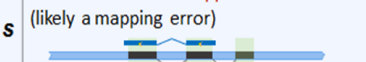
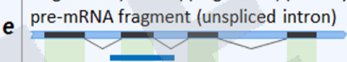
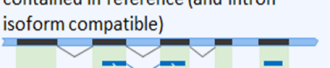
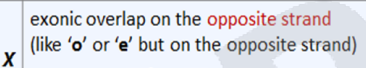
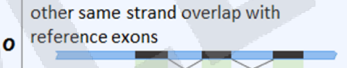
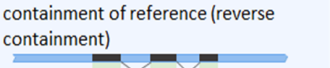
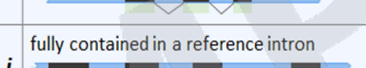
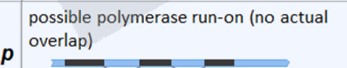
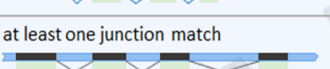
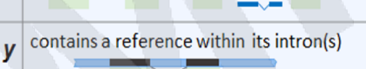
4. 3. Prediction of Novel Transcripts/Alternative Splicing

Transcripts

Transcripts are additionally assembled from the results of mapped reads to predict novel transcripts and novel alternative splicing transcripts without StringTie -e option.

Assembled annotation (GTF file) of samples is merged into one merged file with StringTie -merge option. After then, the abundances of samples are calculated for known and novel transcripts. The gffcompare program of GFF utilities is used to classify the types of known transcript and novel transcript, this resulted in known transcripts and novel transcripts are assigned the class code according to their alternative splicing type as the following the Table 6.

Table 6. Description of class code for various splicing alternative transcript type

= complete match of intron chain 	s intron match on the opposite strand (likely a mapping error) 	e single exon, overlapping intron, possibly pre-mRNA fragment (unspliced intron) 
c contained in reference (and intron isoform compatible) 	x exonic overlap on the opposite strand (like 'o' or 'e' but on the opposite strand) 	o other same strand overlap with reference exons 
k containment of reference (reverse containment) 	i fully contained in a reference intron 	p possible polymerase run-on (no actual overlap) 
j at least one junction match 	y contains a reference within its intron(s) 	r repeat (at least 50% bases soft-masked) u none of the above (unknown, intergenic)

4. 3. 1. Prediction of Known/Novel Transcripts and Estimation of Expression Levels

(Refer to Path: result_RNAseq_excel/Novel_transcript_analysis/StringTie/Expression_Profile_with_Novel.WBcel235.transcript.xlsx)

This result refers to expression level for each sample for each known transcript, novel transcript and novel alternative splicing transcript.

(Note: Expression profile on chapter 4.2 (Known transcript expression level based on Reference Genome Model) doesn't contain the expression profile of novel transcript.)

Table 7 shows an example result of the known/novel transcripts and their expression levels, which are predicted by StringTie for each sample. If novel gene exists, StringTie assigns the "MSTRG.xxxx" number as temporary gene ID. If novel transcript or alternative splicing transcript exists, it assigns "MSTRG.xxxx.yy" number for temporary transcript ID. The following Table 7 represents transcript locus, length, class code, read count, FPKM for each transcript.

(Refer to the class code of table 6)

Table 7. Known/novel transcript expression level (Example)

Transcript_ID	Gene_ID	Gene_Symbol	Description	Transcript_Locus	Transcript_Length	Class_Code	AM_Read_Count	BM_Read_Count	AM_FPKM	BM_FPKM
NM_130786	1	A1BG	alpha-1-B glycoprotein	chr19:58958172-58964	1766	=	43	17	0.502118	0.186489
NM_001086	13	AADAC	arylacetamide deacetylase	chr3:151531861-15154	1696	=	0	0	0	0
MSTRG_10251.2	14	AAMP	angio associated migratory cell protein	chr2:219128852-21913	1980	=	789	1359	8.01164	13.329059
NM_001302545	14	AAMP	angio associated migratory cell protein	chr2:219128852-21913	1835	=	781	781	8.7915	8.058047
MSTRG_10251.1	14	AAMP	angio associated migratory cell protein	chr2:219128852-21913	2275	=	659	874	5.986394	7.461948
NM_001097	14	AAMP	angio associated migratory cell protein	chr2:219128852-21913	1832	=	3812	4722	40.740643	50.081356
NM_001805	16	AARS	alanyl-tRNA synthetase	chr16:70286297-70323	3344	=	17776	26083	109.862747	151.563644
MSTRG_6156.1	18	ABAT	4-aminobutyrate aminotransferase	chr16:8768444-887843	8504	=	962	792	2.337547	1.808619
NM_020686	18	ABAT	4-aminobutyrate aminotransferase, trans	chr16:8768444-887843	4814	=	1353	1266	5.806647	5.10746
NM_000683	18	ABAT	4-aminobutyrate aminotransferase, trans	chr16:8806826-887843	5586	=	5	8	0.017256	0.025769
NM_001127448	18	ABAT	4-aminobutyrate aminotransferase, trans	chr16:8814573-887843	4908	=	0	0	0	0
MSTRG_19.1	MSTRG_19			chr11:945604-945843	240	u	2	1	0.119388	0.080965
NM_005502	19	ABCA1	ATP binding cassette subfamily A memb	chr9:107543284-10769	10502	=	57	103	0.112093	0.189561
MSTRG_16885.2	20	ABCA2	ATP binding cassette subfamily A memb	chr9:139901686-13992	8906]	4150	2834	9.628562	6.18288
NM_001606	20	ABCA2	ATP binding cassette subfamily A memb	chr9:139901686-13992	8154	=	3656	6078	9.26846	14.48372
MSTRG_16885.1	20	ABCA2	ATP binding cassette subfamily A memb	chr9:139901686-13992	8151]	6091	4932	15.444248	11.755709
NM_212533	20	ABCA2	ATP binding cassette subfamily A memb	chr9:139901686-13992	8146	=	573	525	1.451959	1.251235

- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_ID: Entrez gene ID
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- Transcript_Locus: Start and end position of transcript on genomic region
- Transcript_Length: Length of transcript
- Class_Code: Class code corresponding to transcript ID (Refer to Table 6)
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM value for each sample (normalized value)

4. 3. 2. Prediction of Known/Novel Genes and Estimation of Expression Levels

(Refer to Path: result_RNAseq_excel/Novel_transcript_analysis/StringTie/Expression_Profile_with_Novel.WBcel235.gene.xlsx)

This result refers to expression level for each sample for gene containing known transcripts and novel alternative splicing transcript or novel gene.

(Note: Expression profile on chapter 4.2 (Known transcript expression level based on Reference Genome Model) doesn't contain the expression profile of novel transcript.)

Table 8 shows an example result of the known/novel genes and their expression levels, which are predicted by StringTie for each sample. If novel gene exists, StringTie assigns the "MSTRG.xxxx" number as temporary gene ID. If novel transcript or alternative splicing transcript exists, it assigns "MSTRG.xxxx.yy" number for temporary transcript ID. The following Table 8 represents transcript ID, gene symbol, class code corresponding to transcript ID, read count, FPKM per sample for each gene.

(Refer to the class code of table 6)

Table 8. Known/novel gene expression level (Example)

Gene_ID	Transcript_ID	Gene_Symbol	Description	Class_Code	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM
1	NM_130786	A1BG	alpha-1-B glycoprotein	=	43	17	0.502118	0.186489
2	NM_000014.NM_001347423.NM_0013474	A2M	alpha-2-macroglobulin	=,=,=,=	3	5	0.011832	0.01613
27	MSTRG.1476.1.NM_001136000.NM_0011	ABL2	ABL proto-oncogene 2, non-receptor tyrosine kinase	j,=,=,=,=,=,=	5167	4054	9.00601	6.920619
28	NM_020469	ABO	ABO alpha 1-3-N-acetylgalactosaminyltransferase	=	2	4	0.026178	0.036919
29	MSTRG.6798.1.MSTRG.6798.10.MSTRG.6798.30	ABR	active BCR-related	j,j,j,j,j,o,=,=,=	6965	7608	26.611728	27.889476
30	MSTRG.11708.1.MSTRG.11708.2.MSTRG.11708.3	ACAA1	acetyl-CoA acyltransferase 1	u	121	371	0.948868	2.748202
31	MSTRG.7235.1.NM_198834.NM_198836	ACACA	acetyl-CoA carboxylase alpha	j,j,j,j,=,=,=	1656	1523	18.817823	16.274804
47	MSTRG.7321.3.NM_001096.NM_0013032	ACLY	ATP citrate lyase	j,=,=,=,=	11888	11450	25.661393	23.222174
48	MSTRG.16221.3.MSTRG.16221.4.MSTRG.16221.5	ACO1	aconitase 1	j,=,=,=,=	8052	10551	37.163913	45.821517
49	NM_001097	ACR	acrosin	j,j,j,j,j,j,x,=,=	12609	7106	14.989356	18.637128
50	MSTRG.11433.2.NM_001098	ACO2	aconitase 2	=	11	13	0.1535	0.180757
51	NM_001185039.NM_004035.NM_007292	ACOX1	acyl-CoA oxidase 1	j,=	3474	6082	21.17432	34.653712
52	MSTRG.9264.1.NM_001040649.NM_0043	ACP1	acid phosphatase 1, soluble	=,=,=	2785	5703	7.561765	14.583454
53	MSTRG.2844.1.NM_001302489.NM_0013	ACP2	acid phosphatase 2, lysosomal	j,=,=,=,=	2687	5420	26.545543	53.188392
54	NM_001111034.NM_001111035.NM_0011	ACP5	acid phosphatase 5, tartrate resistant	j,=,=,=,=	1214	1209	11.174793	10.68333
55	NM_001099.NM_001134194.NM_0012920	ACPP	acid phosphatase, prostate	=,=,=	12	26	0.143592	0.296069
					415	336	3.697107	3.063849

- Gene_ID: Entrez gene ID
- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- Class_Code: Class code corresponding to transcript ID (Refer to Table 6)
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM value for each sample (normalized value)

4. 3. 3. Filtering Novel transcripts

(Refer to Path: result_RNAseq/Novel_transcript_analysis/StringTie/Novel_transcript_list.xlsx)

Novel transcripts are predicted by reads that are not mapped to known exon or gene but to the intergenic region. Table 9 represents the list of novel transcripts filtered by transcripts with class code ‘u’ from the results of known and novel transcripts.

Table 9. Novel transcript list (Example)

Transcript_ID	MSTRG.3.3	MSTRG.3.4	MSTRG.1304.1	MSTRG.2029.1	MSTRG.2032.1
Gene_ID	MSTRG.3	MSTRG.3	MSTRG.1304	MSTRG.2029	MSTRG.2032
Transcript_Locus	chr1:148912-164783	chr1:158224-164789	chr1:155596570-155618518	chr10:38692080-38705692	chr10:42385511-42385720
Transcript_Length	747	978	1447	1065	210
Strand	-	-	-	+	-
Exon_Count	3	2	4	7	1
Exon_Start	148912,155767,164263	158224,164263	155596570,155604561,155607185,155618197	38692080,38696559,38697731,38701156,38702255,38704472,38705610	42385511
Exon_End	149072,155831,164783	158674,164789	155597518,155604637,155607283,155618518	38692478,38696668,38697862,38701371,38702320,38704530,38705692	42385720
Class_Code	u	u	u	u	u
AM_Read_Count	4	110	101	86	17
BM_Read_Count	6	106	157	82	6
AM_FPKM	0.110433	2.307731	1.439751	1.658545	1.647744
BM_FPKM	0.133012	2.093877	2.095667	1.488986	0.488918

- Transcript_ID: If there are detected novel transcripts with novel exons, StringTie assigns these transcripts to “MSTRG.xxxx.yy” of temporary transcript ID.
- Gene_ID: If there are detected novel genes in the intergenic region or unknown region, StringTie assigns these genes to “MSTRG.xxxx” of temporary gene ID.
- Transcript_Locus: Start and end position of transcript on genomic region
- Transcript_Length: Length of transcript
- Strand: Strand of transcript on genomic region
- Exon_Count: The number of exon in the transcript
- Exon_Start, End: The start and end position for each exon in the transcript
- Class_Code: Class code corresponding to transcript ID (Refer to Table 7)
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM value for each sample (normalized value)

4. 3. 4. Filtering Novel alternative splicing transcript

(Refer to Path: result_RNAseq/Novel_transcript_analysis/StringTie/Novel_splicing_variant_list.xlsx)

This result refers to the list of novel alternative splicing transcripts filtered by class code ('j', 'c', 'e', 'i', 'o', 'p', 'x') from the results of known and novel transcripts.

Novel alternative splicing transcript refers to the transcripts that are mapped to new exon or have different assembled structure from known transcript. Table 10 shows an example result of the known and novel transcripts obtained with novel network flow algorithm method of StringTie.

The result represents the list of novel alternative splicing transcripts on the basis of the nearest known transcript and known gene. You can find the information such as the start and end position of novel alternative splicing transcript, exon count of that, start and end position of each exon, read count, FPKM, class code assigned from StringTie.

(Refer to the class code of table 6)

Table 10. Novel alternative splicing transcript list (Example)

refGene_Name	30	166	224	377	409
nearest_refTranscript_Name	NM_001607	NM_198969	NM_001031806	NM_001659	NM_001257328
stringtieGene_Name	MSTRG.11708	MSTRG.8193	MSTRG.7080	MSTRG.3793	MSTRG.6859
stringtieTranscript_Name	MSTRG.11708.7	MSTRG.8193.1	MSTRG.7080.3	MSTRG.3793.1	MSTRG.6859.16
Gene_Symbol	ACAA1	AES	ALDH3A2	ARF3	ARRB2
Gene_Description	acetyl-CoA acyltransferase 1	amino-terminal enhancer of split	aldehyde dehydrogenase 3 family member A2	ADP ribosylation factor 3	arrestin beta 2
Transcript_Locus	chr3:38164213-38178588	chr19:3052908-3062964	chr17:19554700-19580877	chr12:49324912-49351252	chr17:4613940-4624792
Transcript_Length	1841	1651	7190	6947	4138
Strand	-	-	+	-	+
Exon_Count	9	7	8	6	9
Exon_Start	38164213,38167056,38167317,38168001,38169277,38170781,38173417,38178083,38178356	3052908,3054118,3055662,3056310,3057677,3061158,3062790	19554700,19555860,19561058,19564440,19566646,19568261,19575034,19578871	49324912,49327407,49333438,49333780,49334731,49351093	4613940,4618308,4620980,4621183,4622578,4623511,4623688,4623881,4624241
Exon_End	38164613,38167201,38167832,38168191,38169357,38170879,38173496,38178176,38178588	3054038,3054192,3055724,3056354,3057740,3061255,3062964	19555091,19559887,19561175,19564581,19566812,19568360,19575269,19580877	49325736,49332891,49333562,49333890,49334971,49351252	4614039,4620571,4621047,4621979,4622715,4623594,4623767,4623935,4624792
Class_Code	j	j	j	j	j
AM_Read_Count	804	158	393	846	707
BM_Read_Count	743	262	399	1134	627
AM_FPKM	9.024175	1.967786	1.12801	2.515832	3.530653
BM_FPKM	7.841952	3.074596	1.078702	3.170825	2.940787

- refGene_Name: The nearest Entrez gene ID from predicted novel alternative splicing transcript region
- nearest_refTranscript_Name: The nearest transcript ID form predicted novel alternative splicing transcript region
- stringtieGene_Name: Gene ID such as “MSTRG.xxxx” assigned as temporary gene ID in StringTie program
- stringtieTranscript_Name: Trnscript ID such as “MSTRG.xxxx.yy” assigned as temporary transcript ID in StringTie program.
- Gene_Symbol: Symbol of the nearest gene
- Gene_Description: Description of the nearest gene
- Transcript_Locus: Start and end position of transcript on genomic region
- Transcript_Length: Length of transcript
- Strand: Strand of transcript on genomic region
- Exon_Count: The number of exon in the transcript
- Exon_Start, End: The start and end position for each exon in the transcript

- Class_Code: Class code corresponding to transcript ID (Refer to Table 7)
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM value for each sample (normalized value)

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 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	MG_CTRL
2	MG_CTRL_2	MG_CTRL
3	MG_CTRL_3	MG_CTRL
4	MG_TEST_1	MG_TEST
5	MG_TEST_2	MG_TEST
6	MG_TEST_3	MG_TEST

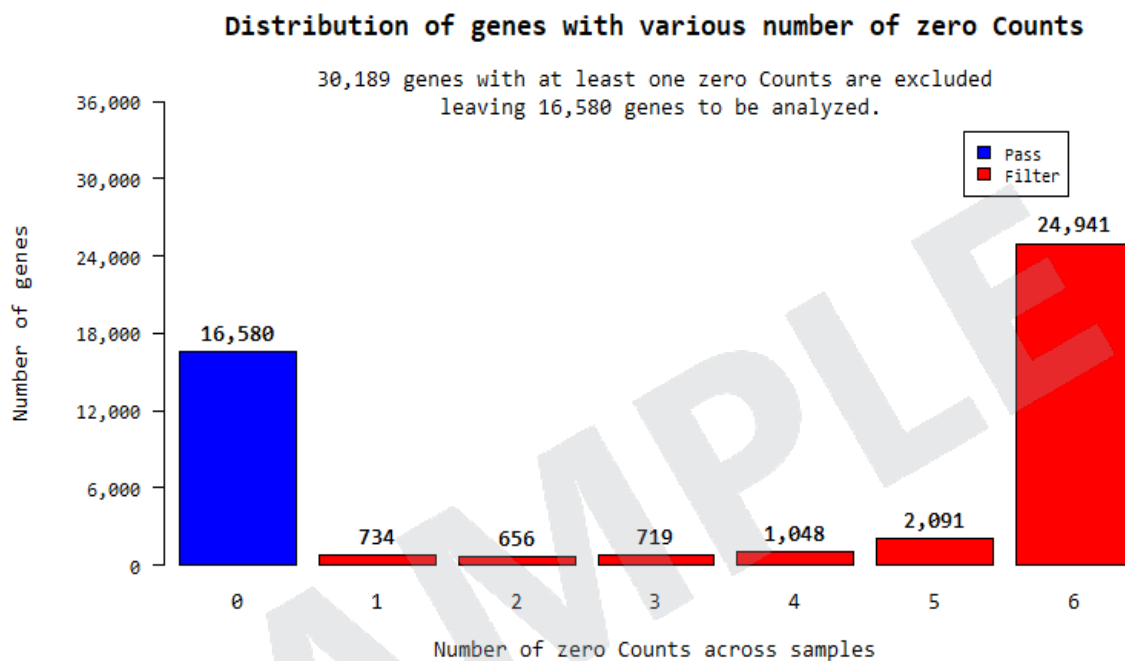
Comparison pair and statistical method for each pair are shown below.

Index	Test vs. Control	Statistical Method
1	MG_TEST vs. MG_CTRL	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 6 samples, if more than one read count value was 0, it was not included in the analysis. Therefore, from total of 46,769 genes, 30,189 were excluded and only 16,580 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.

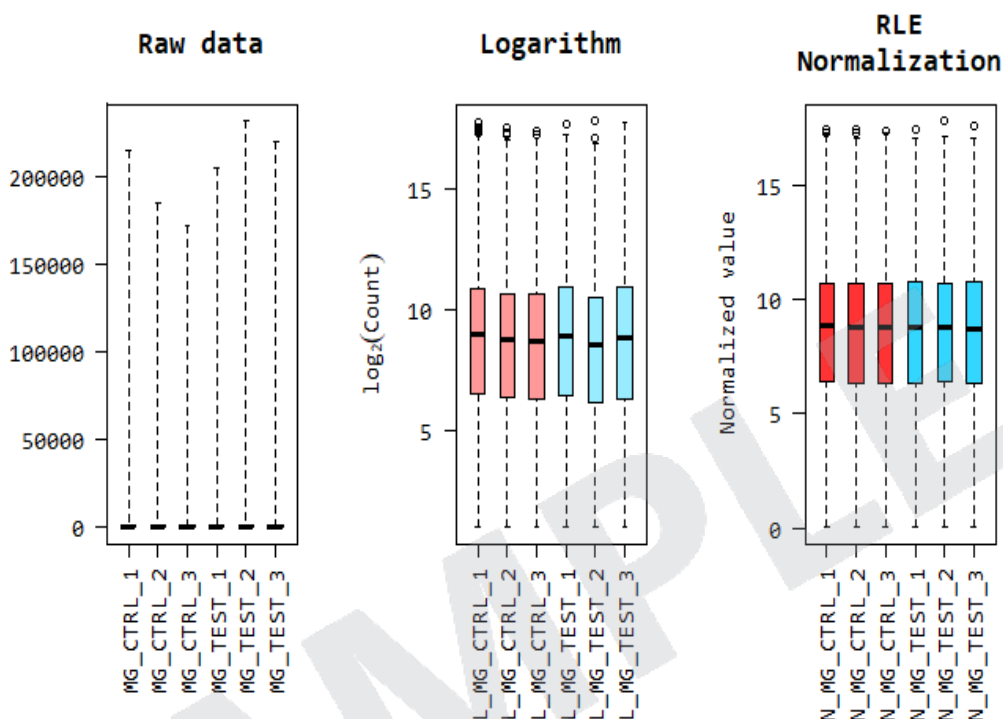
In case of DESeq2, read count+1 & Logarithm value is used to visualize the plots before normalization and regularized log (rlog) transformed value is used to visualize the plots after normalization.

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.

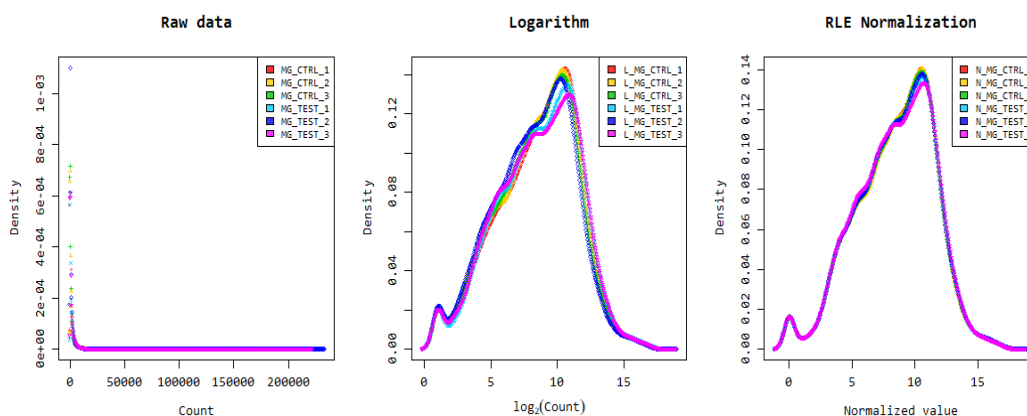
5. 1. 3. 1. Boxplot of Expression Difference between samples.

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

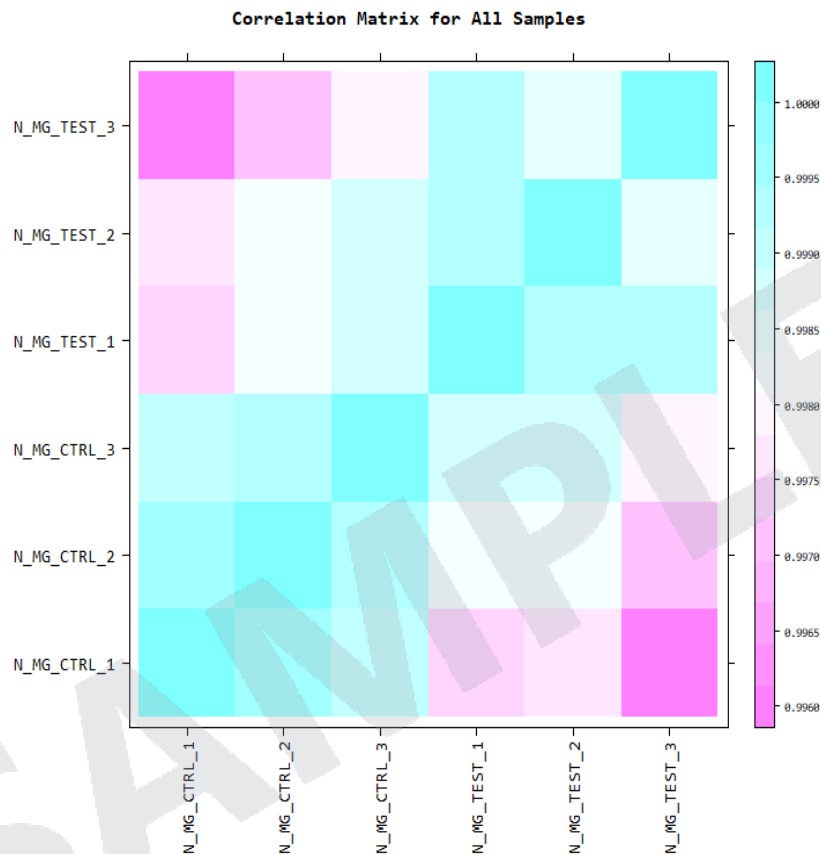
Below density plots show the corresponding samples expression distribution before and after of raw signal (read count), Log₂ transformation of read count+1 and RLE Normalization.



5. 1. 4. Correlation Analysis between samples

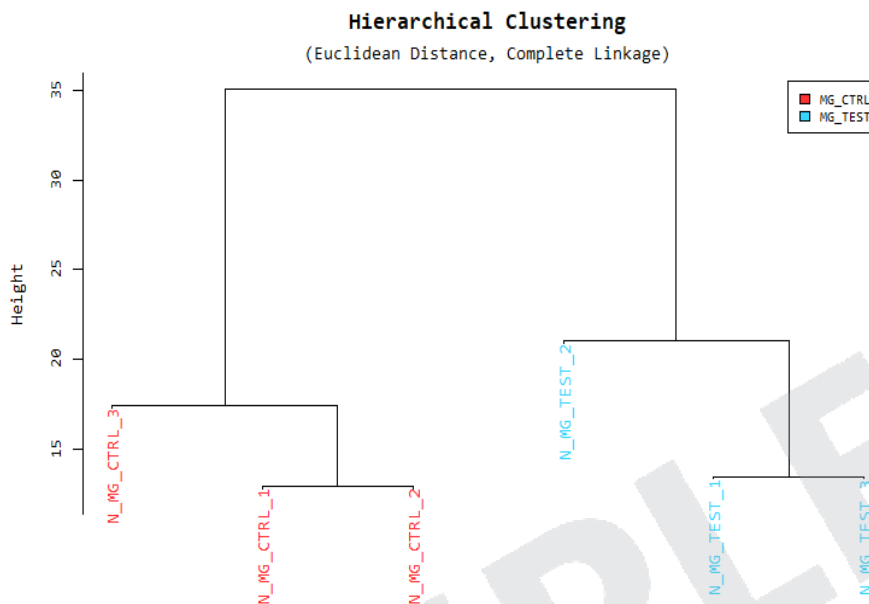
The similarity between samples are obtained through Pearson's coefficient of the normalized value. For range: $-1 \leq r \leq 1$, the closer the value is to 1, the more similar the samples are.

Correlation matrix of all samples is as follows.



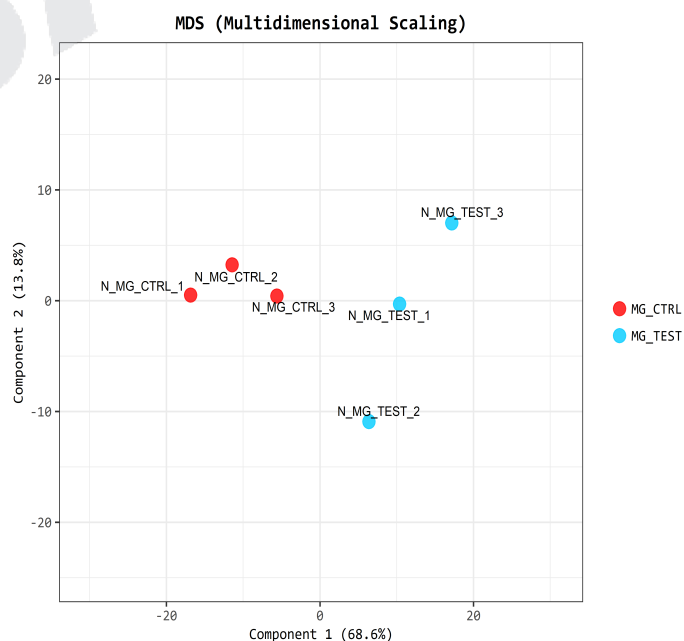
5. 1. 5. Hierarchical Clustering Analysis

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/HTseq/Expression_Profile.WBcel235.gene.xlsx)
: 46,769 genes, 6 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)
: 16,580 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 $|fc| \geq 2$ & nbinomWaldTest raw p-value < 0.05.
 - Significant data
(Refer to Path: result_RNAseq_excel/DEG_result/data3_fc2 & raw.p.xlsx)
: 564 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) For significant lists, gene-set enrichment analysis was performed using DAVID tool based on gene ontology, KEGG and many other functional annotation DBs.
<http://david.abcc.ncifcrf.gov/> Please refer to the DAVID_chart of data3 file.
Gene Ontology: <http://geneontology.org/>
KEGG: <http://www.genome.jp/kegg/>
DAVID: <http://david.abcc.ncifcrf.gov/>
A functional annotation chart report is provided for enrichment analysis.
(Refer to Path: result_RNAseq_excel/DEG_result/DAVID/)
- 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](#).

SAMPLE

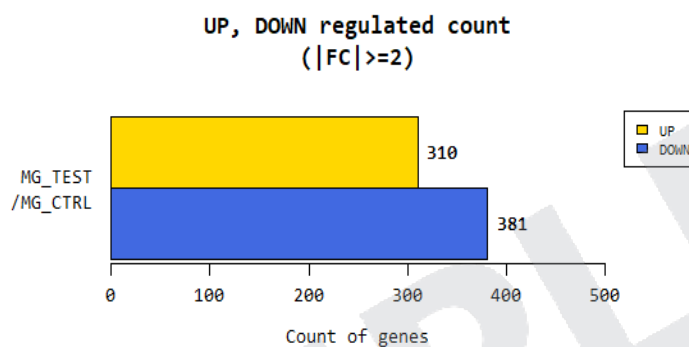
5. 3. Significant Gene Results

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

These are fc2 & raw.p, MG_TEST_vs_MG_CTRL results by example.

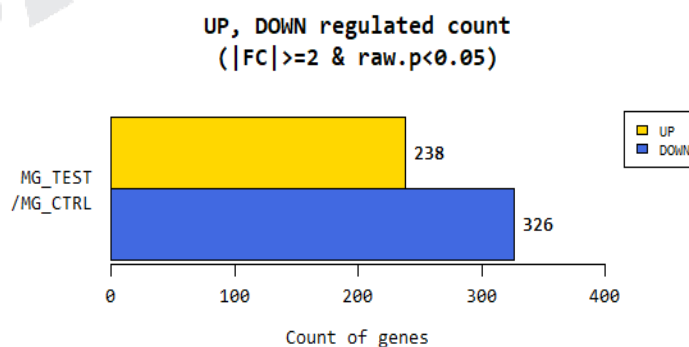
5. 3. 1. Up, Down Regulated Count by Fold Change

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



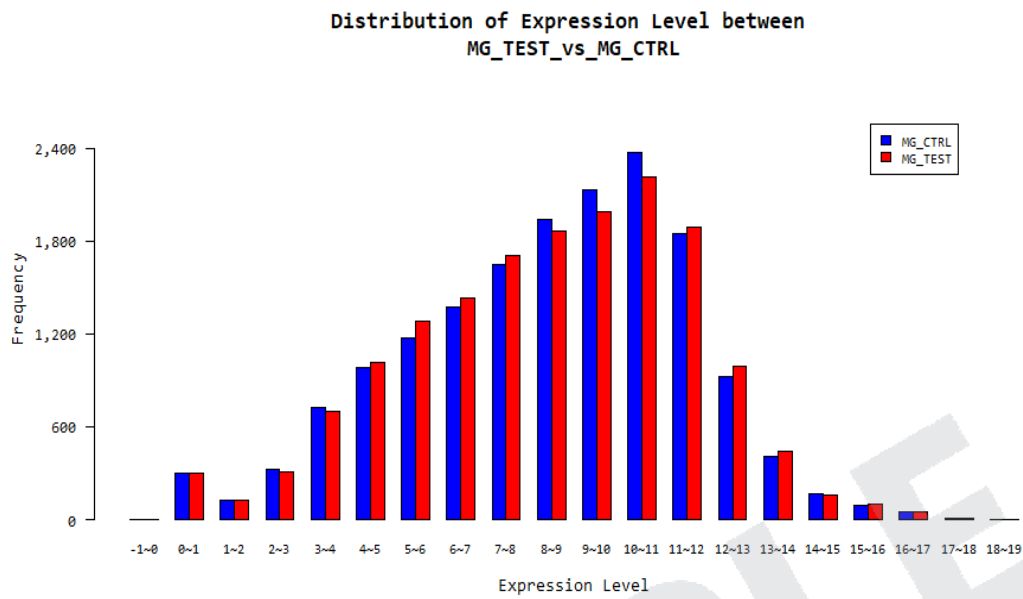
5. 3. 2. Up, Down Regulated Count by Fold Change and p-value

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.

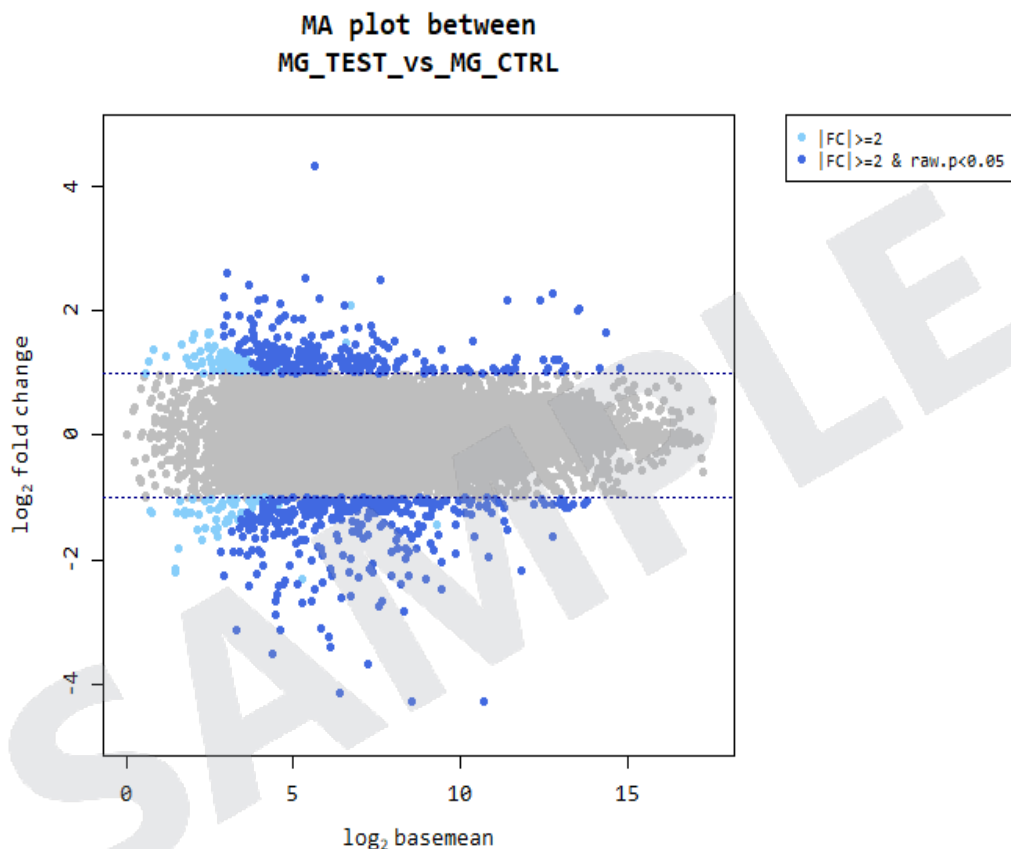


SAMPLE

5.3.4. 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).

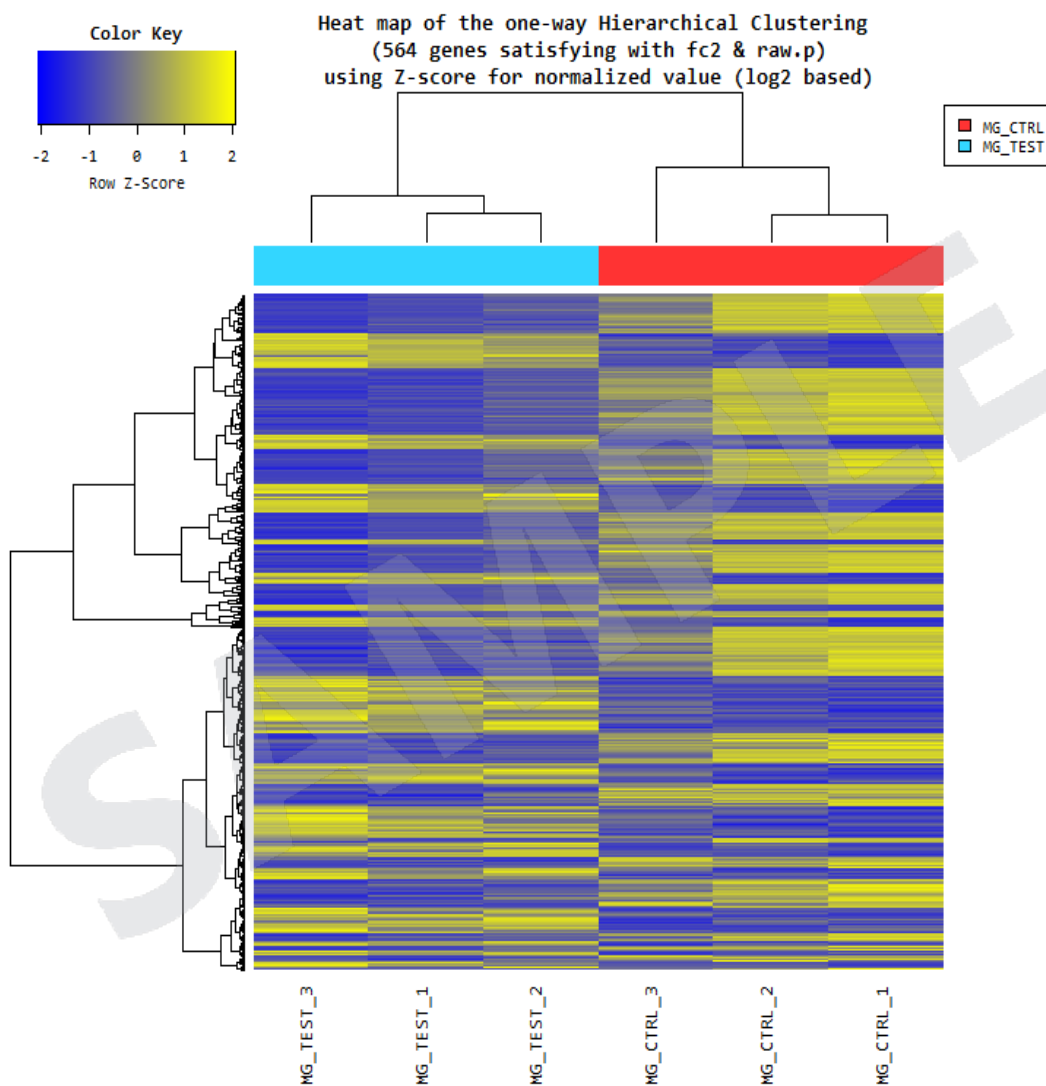
For example, even though fold change might be different by two-fold, the gene with higher mean of normalized counts may be more credible.



5. 3. 5. 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.



5. 4. DAVID Gene-Set Enrichment Analysis

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

For DEG list, gene-set enrichment analysis is performed with DAVID tool based on gene ontology, KEGG and other functional annotation databases.

A functional annotation chart report is provided for enrichment analysis.

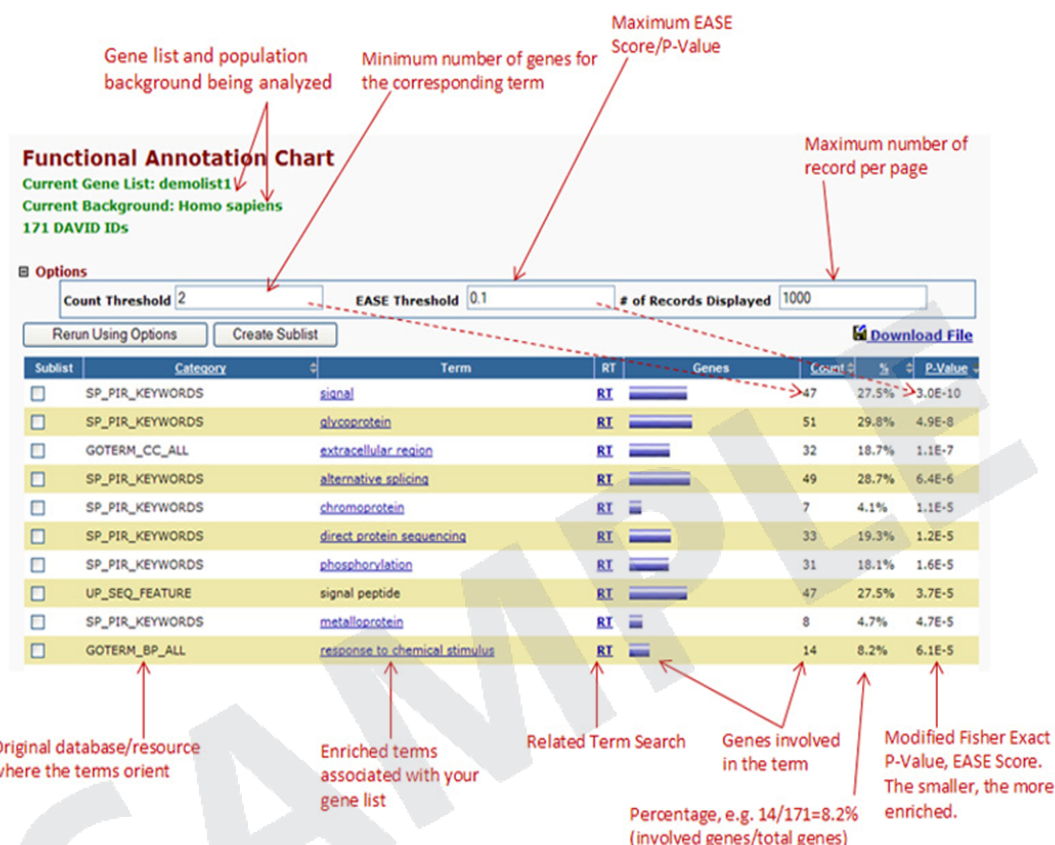
Chart below shows gene set databases that are used for DAVID tool.

Category	DB.class	URL
GOTERM_BP_FAT	Gene_Ontology	http://www.geneontology.org
GOTERM_CC_FAT	Gene_Ontology	http://www.geneontology.org
GOTERM_MF_FAT	Gene_Ontology	http://www.geneontology.org
INTERPRO	Protein_Domains	http://www.ebi.ac.uk/interpro
PIR_SUPERFAMILY	Protein_Domains	http://www.uniprot.org
SMART	Protein_Domains	http://smart.embl.de
BBID	Pathways	http://bbid.grc.nia.nih.gov
BIOCARTA	Pathways	http://www.biocarta.com/Default.aspx
KEGG_PATHWAY	Pathways	http://kegg.jp
COG_ONTOLOGY	Functional Categories	http://www.ncbi.nlm.nih.gov/COG
SP_PIR_KEYWORDS	Functional Categories	http://www.uniprot.org
UP_SEQ_FEATURE	Functional Categories	http://www.uniprot.org
OMIM_DISEASE	Disease	http://www.ncbi.nlm.nih.gov/omim

5. 4. 1. Functional Annotation Chart Report

Figure below shows an example of functional annotation chart report.

Caenorhabditis elegans is used as the background species. The enriched gene set results are extracted from the database used for the DAVID tool.



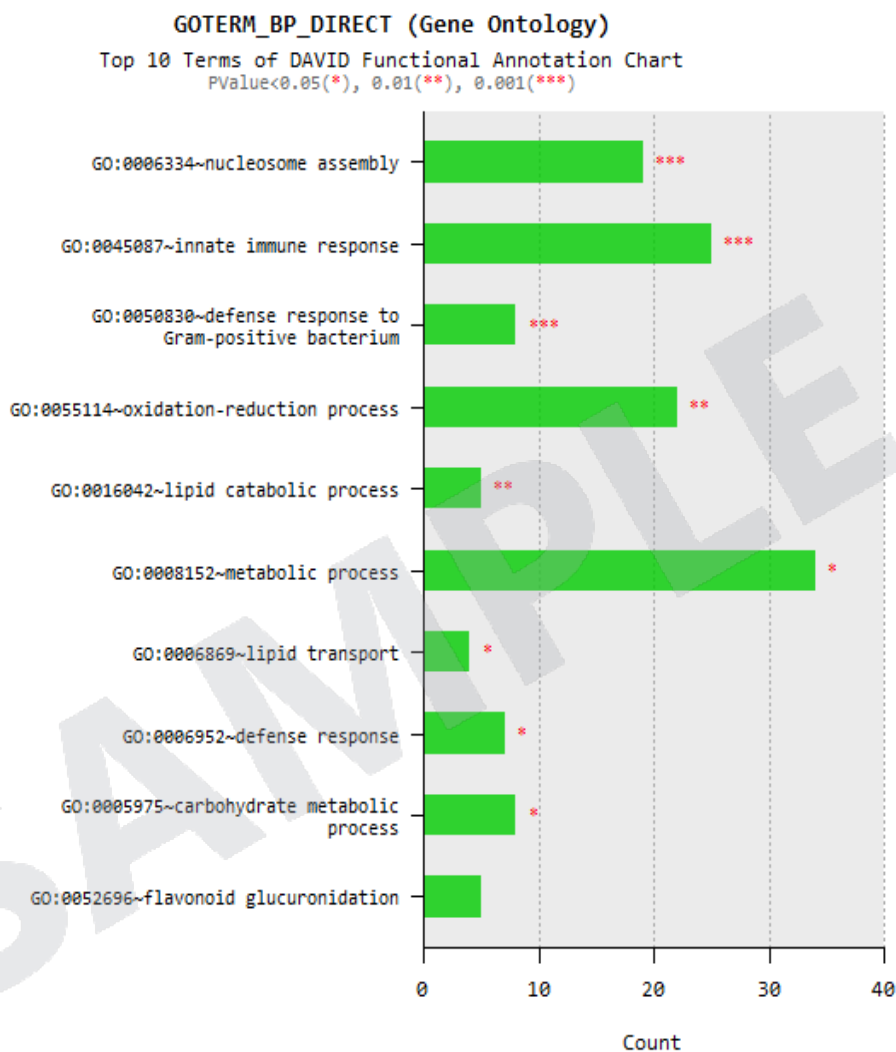
- Category: Database with defined gene set
- Term: Explanation on gene set
- Genes: Genes that are included in the gene set term
- Percentage, %: The ratio of genes that are included in the gene set term from input gene list (** Input gene list means the gene list matched with DAVID system from data3 gene list.)
- P-value: Also known as EASE score, the p-value from the Modified Fisher exact test to determine the enrichment of the gene from the gene set. If this value is lower than 0.05, it is classified as enrichment

(Additional columns in download file)

- List Total: Number of genes in the gene list
- Pop Hits: Number of genes in the total group of genes assayed that belong to the specific Gene Category
- Pop Total: Number of genes in the total group of genes assayed that belong to any Gene Category within the System

The bar plot below shows the results of the enrichment analysis through DAVID's functional annotation based on Gene Ontology, KEGG, and other functional annotation DBs for 564 significant transcripts.

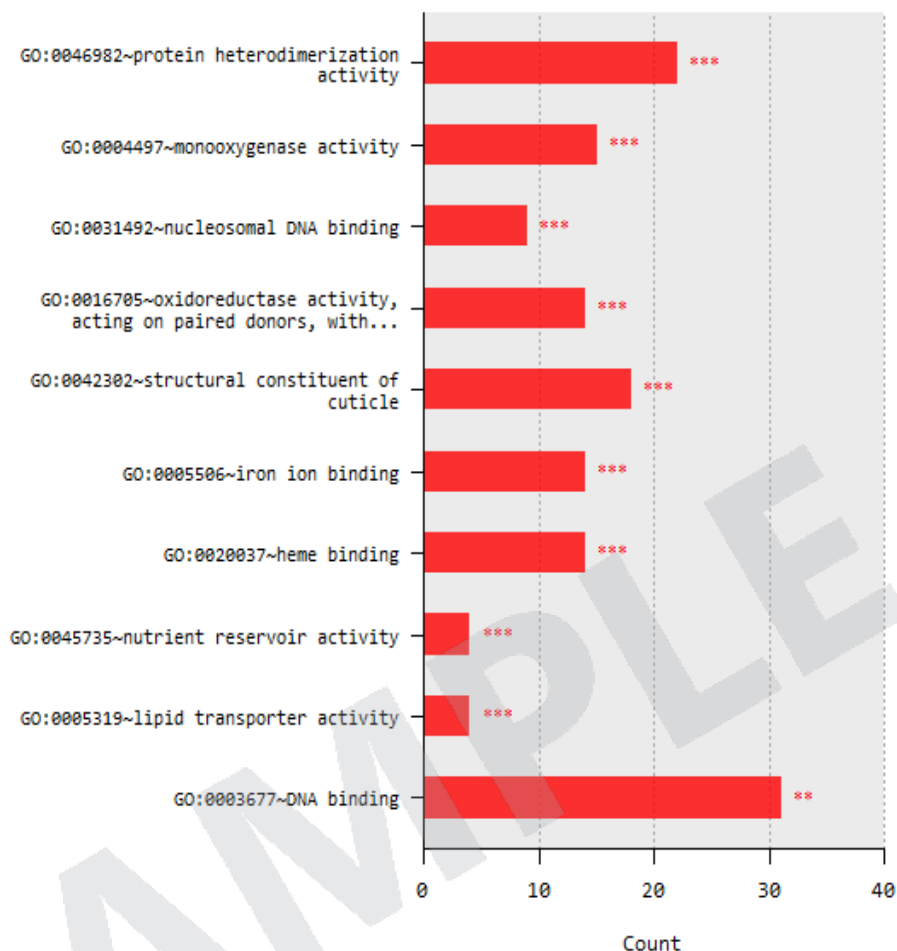
(These plots were made based on functional annotation chart report.)



GOTERM_MF_DIRECT (Gene Ontology)

Top 10 Terms of DAVID Functional Annotation Chart

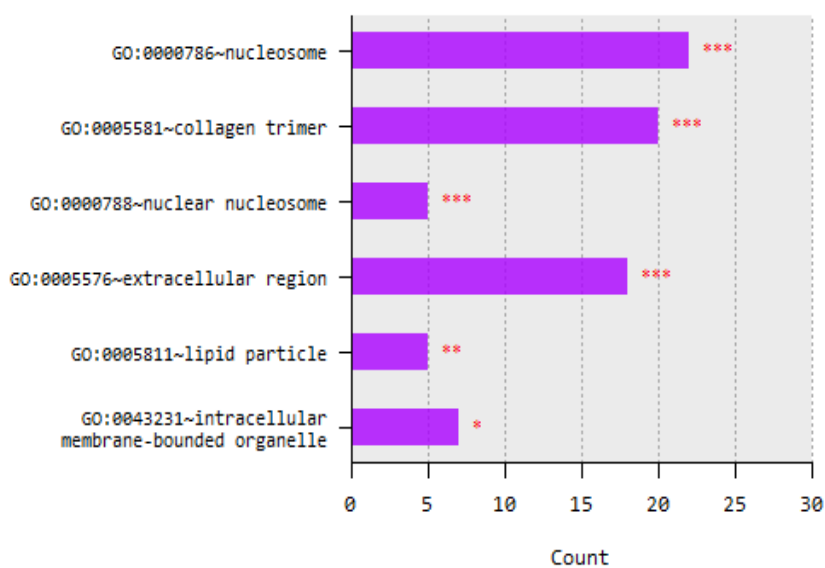
PValue<0.05(*), 0.01(**), 0.001(***)



GOTERM_CC_DIRECT (Gene Ontology)

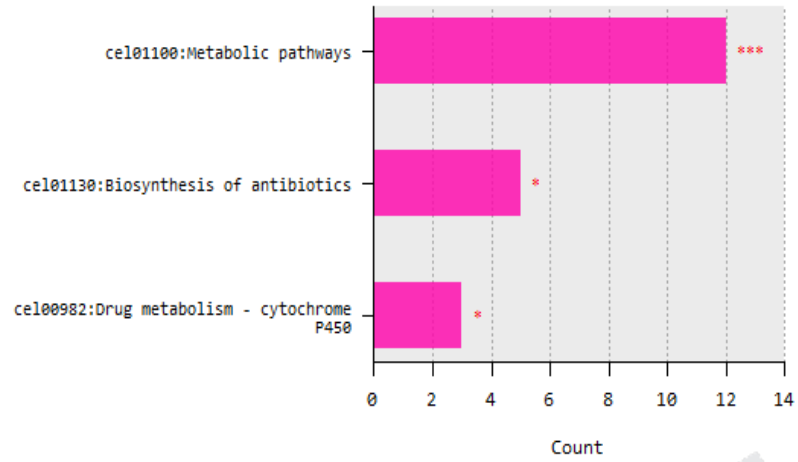
Top 6 Terms of DAVID Functional Annotation Chart

PValue<0.05(*), 0.01(**), 0.001(***)



KEGG_PATHWAY (Pathways)

Top 3 Terms of DAVID Functional Annotation Chart
PValue<0.05(*), 0.01(**), 0.001(***)



SAMPLE

5. 5. KEGG Enrichment Analysis

(Refer to Path: result_RNAseq_excel/DEG_result/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.

The two results are provided for enrichment analysis.

- 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.

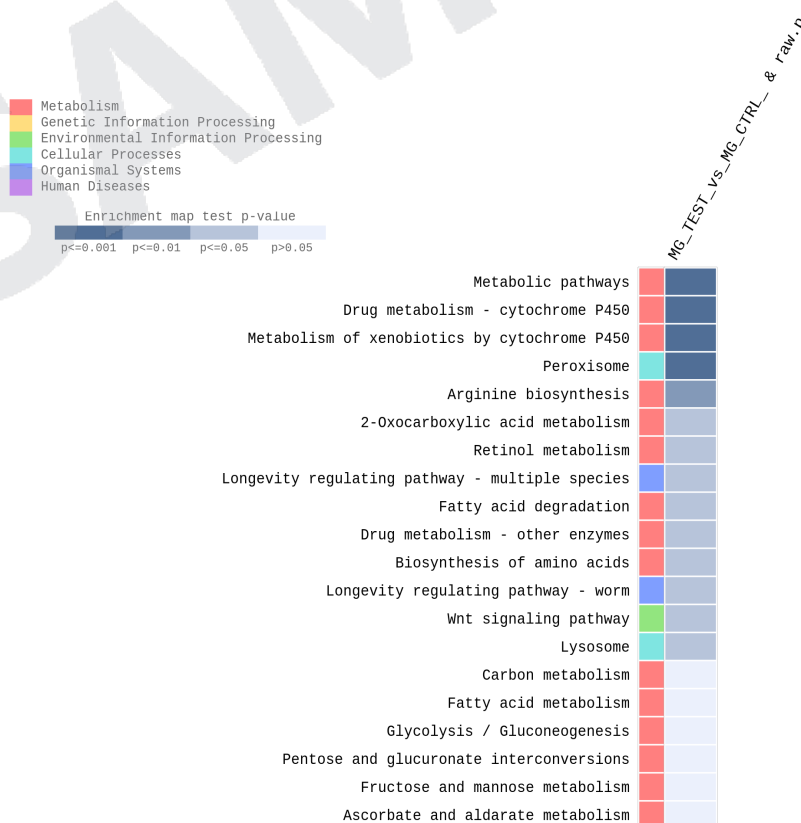


Figure 14. Result of gene-set enrichment analysis (p-value top 20)

5. 5. 1. KEGG HTML Viwer

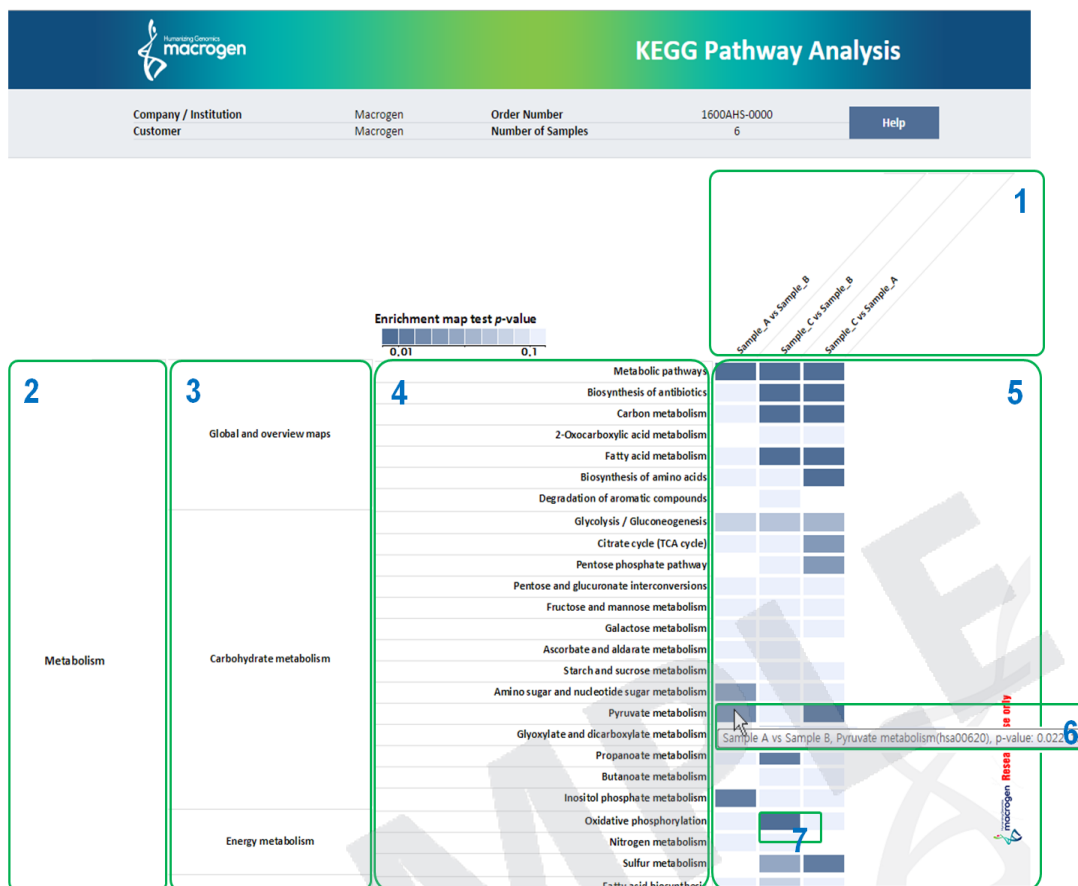


Figure 15. Description of KEGG Viewer frame

- 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.

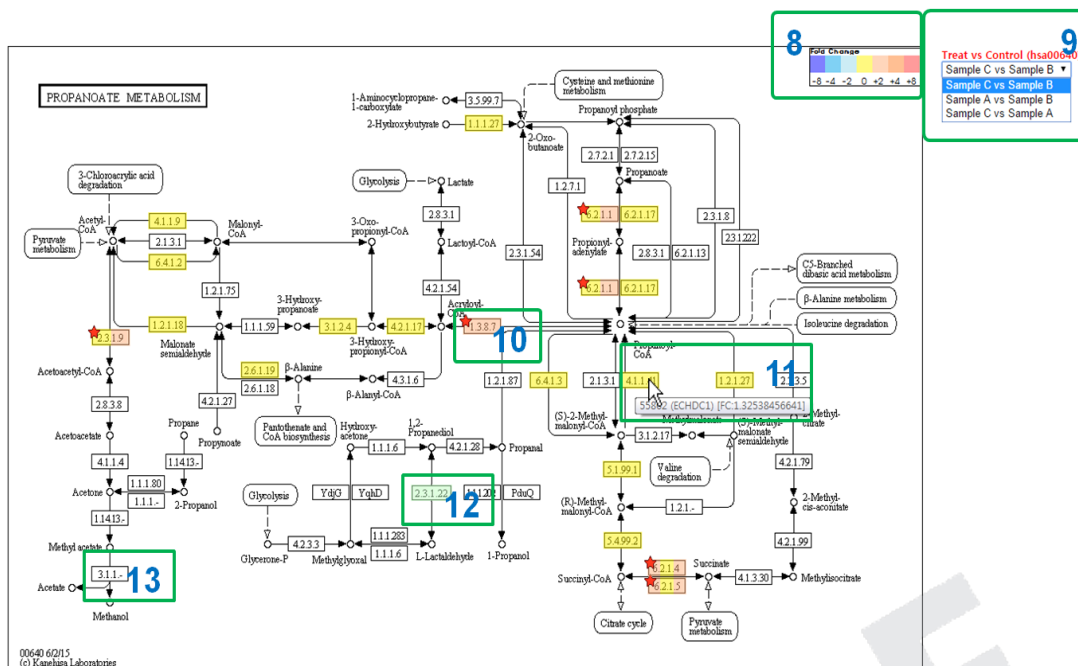


Figure 16. Description of KEGG pathway map frame

- 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 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	10229,10622,10797,10998,1106	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	7.54253E-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	4.88683E-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	2.6895E-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/Avolume	N_A	N_B
22801	04151	PI3K-Akt signal	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 in	0.001875844	0.393927235	0.006353665	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	04810	Regulation of ai	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 cardiomy	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 carcinogen	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 pool	1	1	1	MTHFD2L	1.747046	9.561974	9.167981	9.972899
441024	01100	Metabolic pathw	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 pathw	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 pathw	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

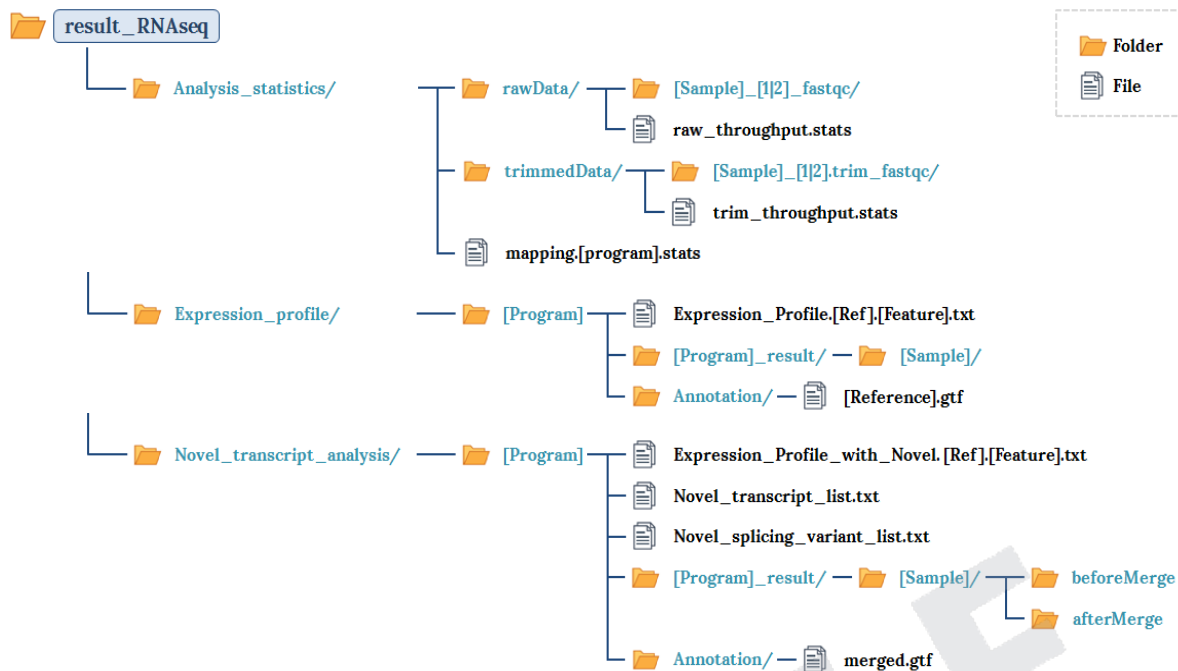
Raw data is the FASTQ file that isn't trimmed adapter sequence.

Download link	File size	md5sum
MG_CTRL_1_1.fastq.gz	1.42G	0bf458946ae8799f26c6e2a9da70ea8d
MG_CTRL_1_2.fastq.gz	1.42G	cf555c1ac94af4d648299eec5bcb294f
MG_CTRL_2_1.fastq.gz	1.23G	05a2b102cd7bee0e1b58ab56f01be73f
MG_CTRL_2_2.fastq.gz	1.29G	606c1d9494df064bba79b99589379d6f
MG_CTRL_3_1.fastq.gz	1.25G	0355f4b2a462c9cb2feb358b95e4a6e7
MG_CTRL_3_2.fastq.gz	1.27G	c60814b7c2b53aa2a25bd390869086d0
MG_TEST_1_1.fastq.gz	1.4G	af2e2695d823173d74d1617412ffec82
MG_TEST_1_2.fastq.gz	1.42G	f0f655780e437cb1a3de984d9835d253
MG_TEST_2_1.fastq.gz	1.12G	a12541b376988d65f75c6b72feea9fe9
MG_TEST_2_2.fastq.gz	1.12G	631b4c220ee972171d306ac26c02efa8
MG_TEST_3_1.fastq.gz	1.4G	72c4edbf6b5fb0290246fcbd3b2dd859
MG_TEST_3_2.fastq.gz	1.45G	f0844417b149ac959cd37691572da6c0

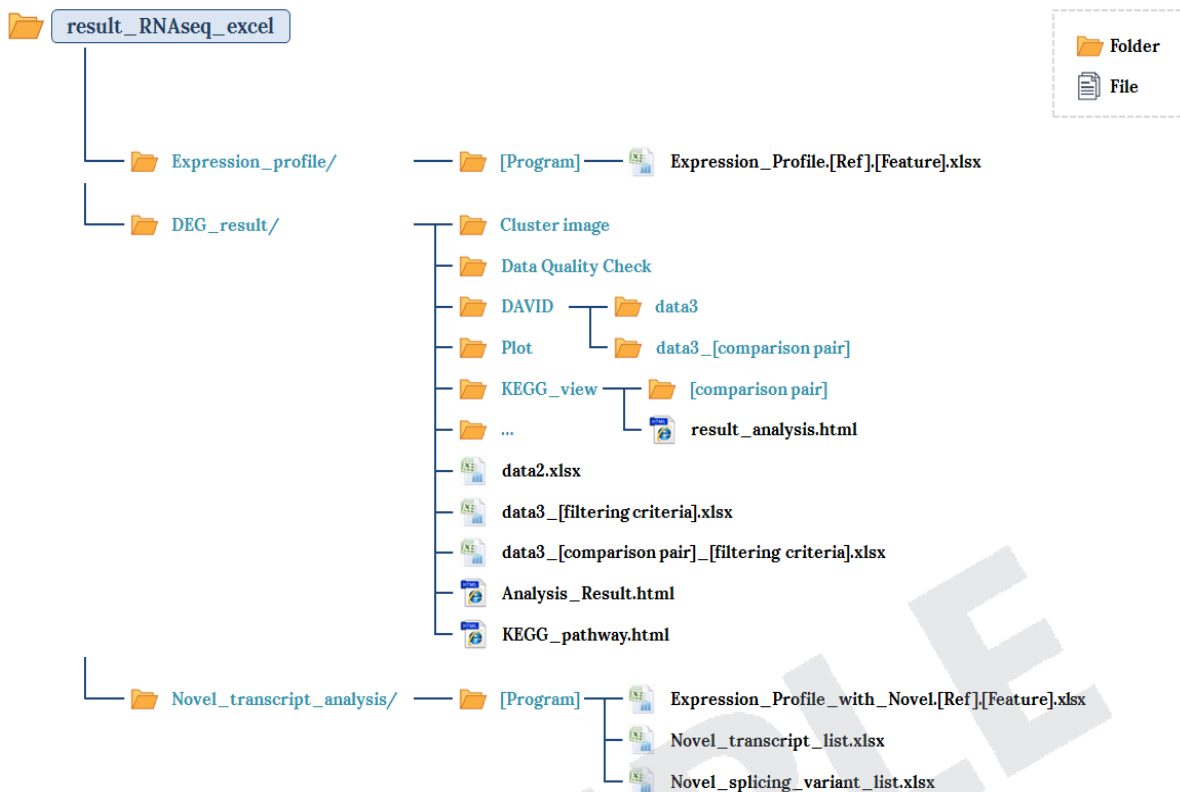
- fastq.gz : This is a zip file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.


6.2. Analysis Results

Download link	File size
SampleAnalysis_nonHuman_result_RNAseq.zip (md5sum: ba762e9aaa768a645a447b2fe8f7b13e)	281.39M
SampleAnalysis_nonHuman_result_RNAseq_excel.zip (md5sum: fe0c2eb74d38cd461d0d640f7873f57d)	54.58M



SAMPLE



 The data retention period is three months, please send an e-mail (ngs@macrogen.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?@ABCDEFGHIJ
20	1 in 100	99%	!"#\$%&'()*+,-./0123456789:;h=i?@ABCDEFGHIJ
30	1 in 1000	99.9%	!"#\$%&'()*+,-./0123456789:;h=i?@ABCDEFGHIJ
40	1 in 10000	99.99%	!"#\$%&'()*+,-./0123456789:;h=i?@ABCDEFGHIJ

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

7. 2. Programs used in Analysis

7. 2. 1. FastQC v0.11.7

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

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

7. 2. 2. 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.

SAMPLE



SAMPLE

Contact us

Tel: +82-2-2180-7016

Site: www.macrogen.com | <http://dna.macrogen.com>