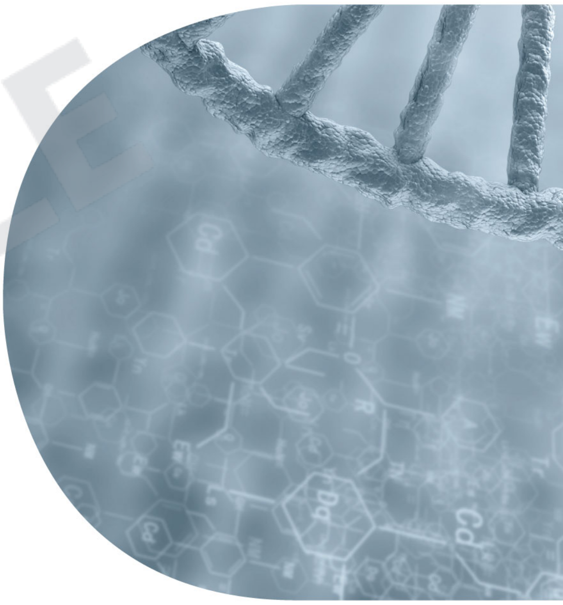
Research use only

Homo sapiens
Transcriptome Sequencing

Report



ver. MGTR4.0_HS_GRCh38

Humanizing Genomics
macrogen

Project Information

Client Name	TESTER
Company/Institution	Macrogen
Order Number	HN00000000
Species	*Homo sapiens*
Reference	GRCh38
Annotation	NCBI_109.20200522
Type of Read	Paired-ends
Read Length	101
Number of Samples	6
Library Kit	TruSeq stranded mRNA
Type of Sequencer	Illumina platform

Project Results Summary

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

The novel transcripts and novel alternative splicing transcripts were discovered during the assembly. In addition, SNV calling, variant annotation, and fusion gene detection were performed.

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.

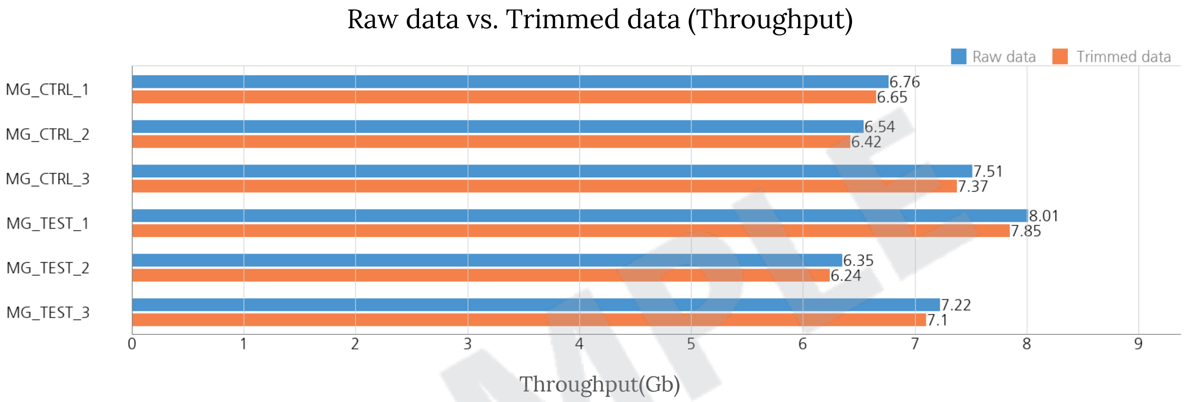


Figure 1. Throughput output of Raw and Trimmed data

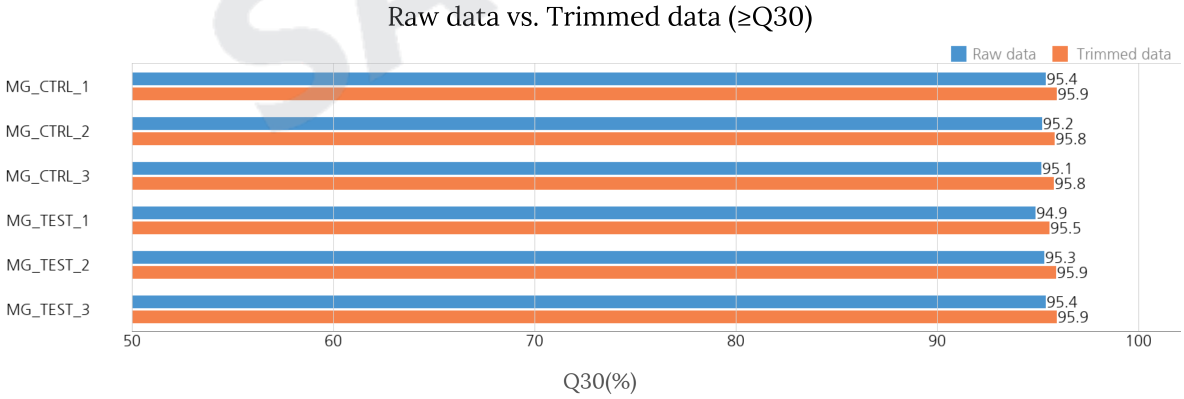


Figure 2. Q30 score of Raw and Trimmed data

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.

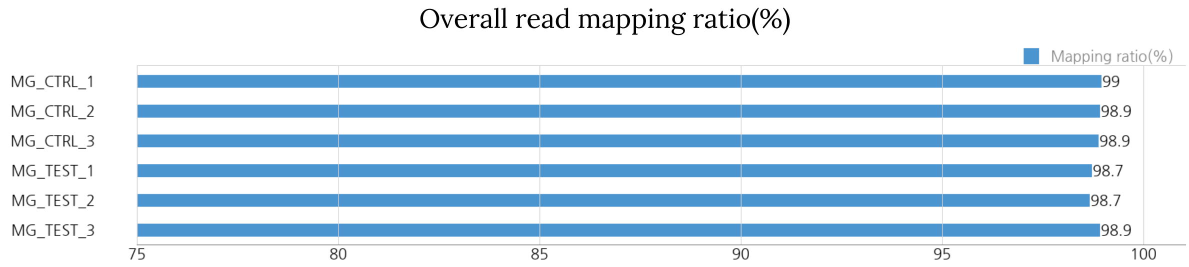


Figure 3. Overall read mapping ratio(%)

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

DEG (Differentially Expressed Genes) analysis was performed on a comparison pair (TEST_vs_CTRL) as requested using DESeq2. The results showed 2,700 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.

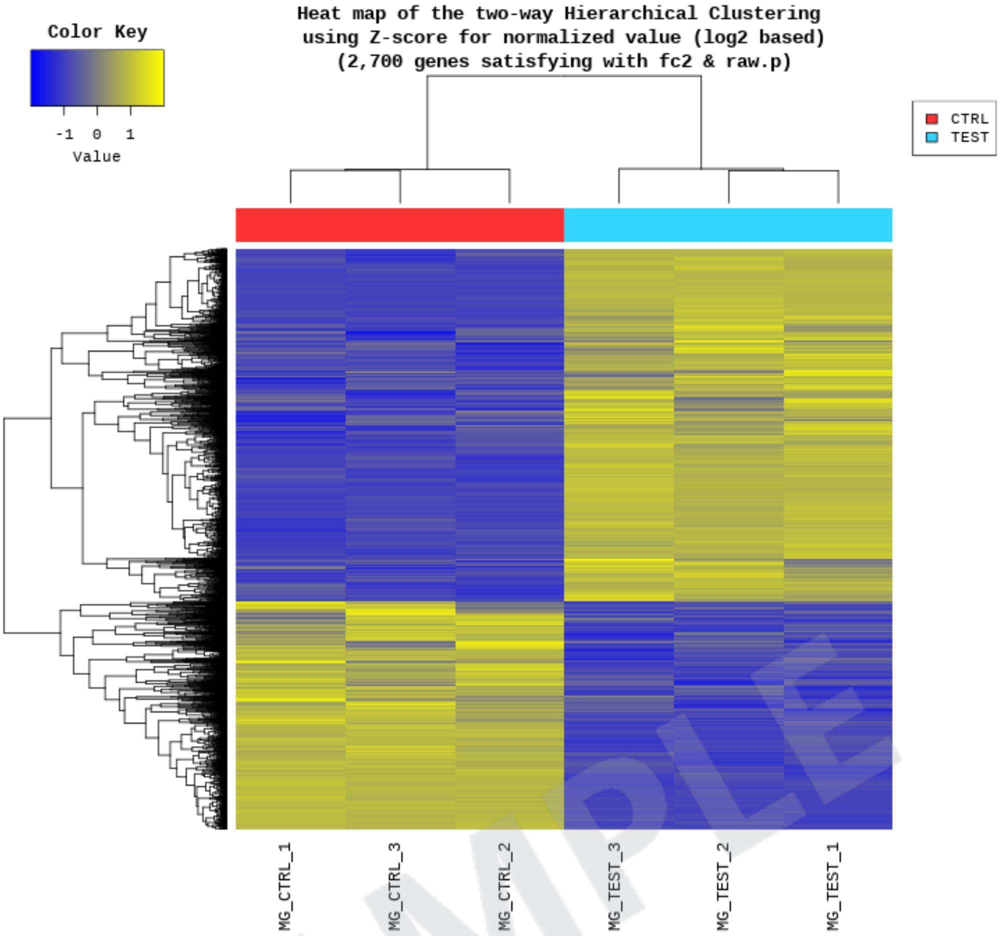


Figure 4. Heatmap for DEG list

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.

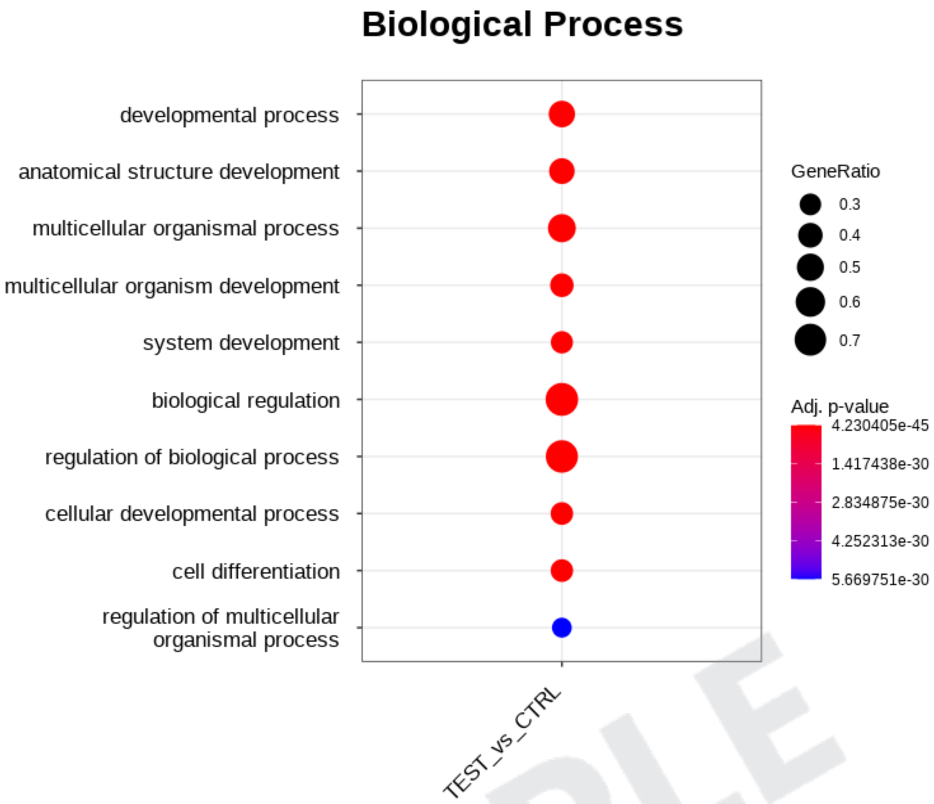


Figure 5. Gene Ontology terms related to Biological Process

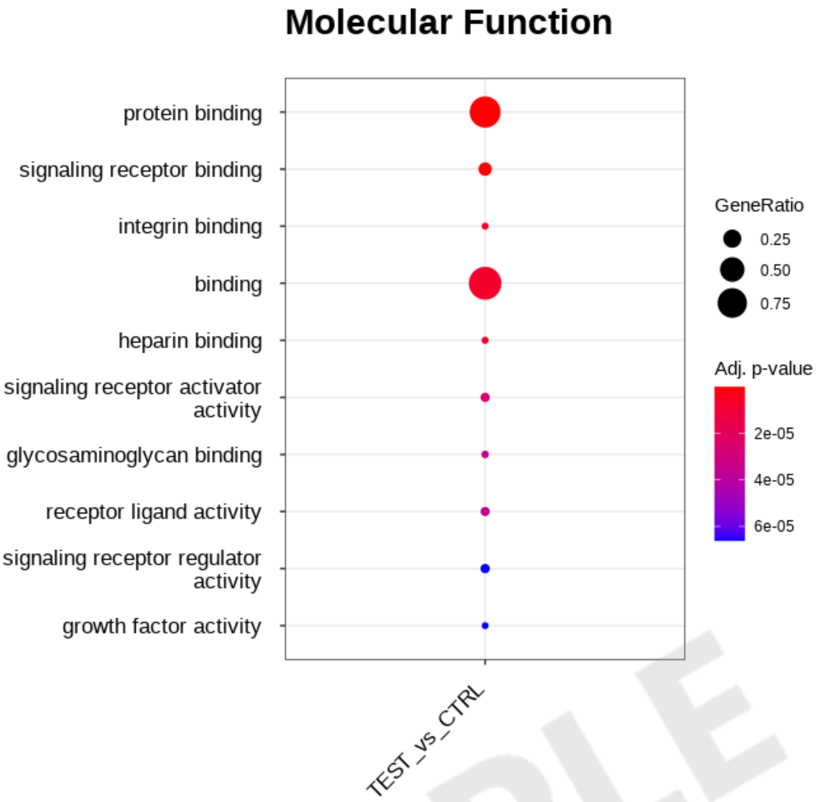


Figure 6. Gene Ontology Terms related to Molecular Function

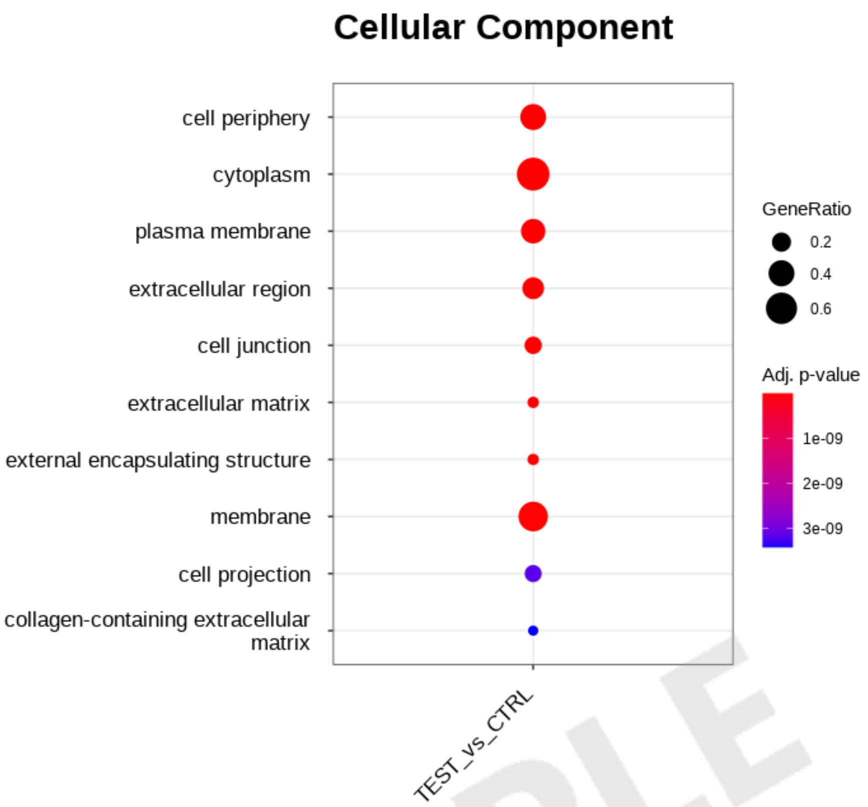


Figure 7. Gene Ontology Terms related to Cellular Component

In addition, novel transcript and novel alternative splicing transcripts were found each sample. Also SNV calling, variant annoation and fusion gene prediction results were summarized for each sample. (Please refer to the main body of this report for detailed explanations.)

Table of Contents

1. Experimental Methods and Workflow

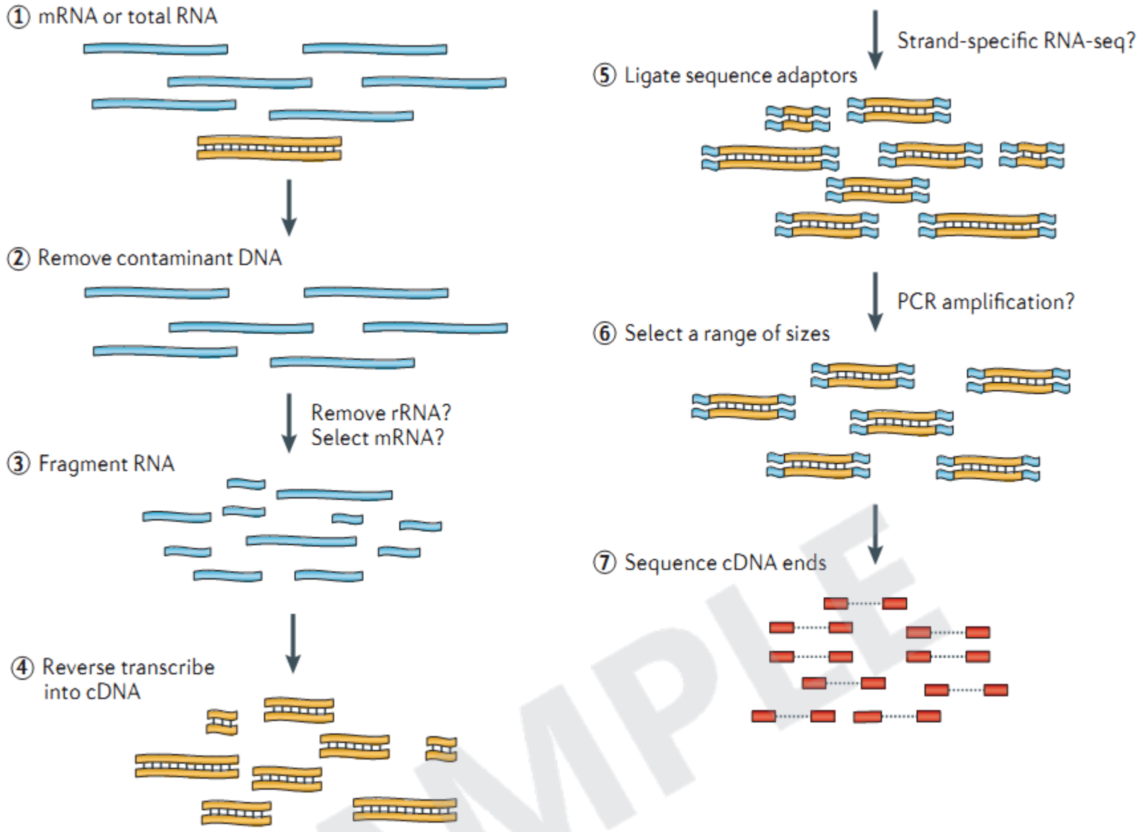


Figure 8. RNA Sequencing Experiment Workflow

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

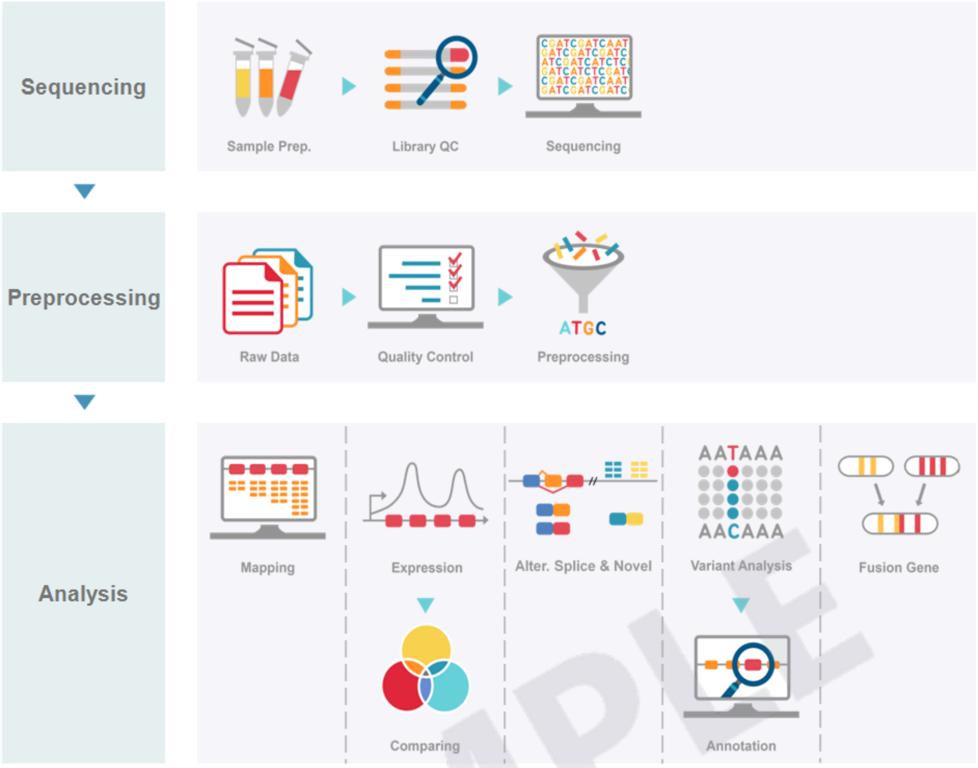


Figure 9. Analysis Workflow

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 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 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.

8) In SNV calling of RNA-seq data, reads are mapped to genomic DNA reference with STAR, then duplications are marked and sorted. Afterwards, mapped reads that can be used in analysis are created through Split 'N' Trim, mapping quality reassignment, indel realignment, and base

recalibration. The reads created in the previous step are used for variant calling with HaplotypeCaller
LINK https://www.broadinstitute.org/gatk/guide/best-practices?bpm=RNAseq

9) Fusion genes are predicted with Defuse, FusionCatcher and Arriba programs.

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, 66,947,992 reads are produced, and total read bases are 6.8Gbp. The GC content (%) is 47.97% and Q30 is 95.36%.

Table 1. Raw data stats

Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
MG_CTRL_1	6,761,747,192	66,947,992	47.97	98.49	95.36
MG_CTRL_2	6,538,936,142	64,741,942	48.13	98.41	95.2
MG_CTRL_3	7,510,790,462	74,364,262	48.43	98.38	95.13
MG_TEST_1	8,009,813,686	79,305,086	49.32	98.26	94.87
MG_TEST_2	6,347,729,608	62,848,808	48.91	98.45	95.31
MG_TEST_3	7,216,968,938	71,455,138	49.71	98.49	95.36

(* 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

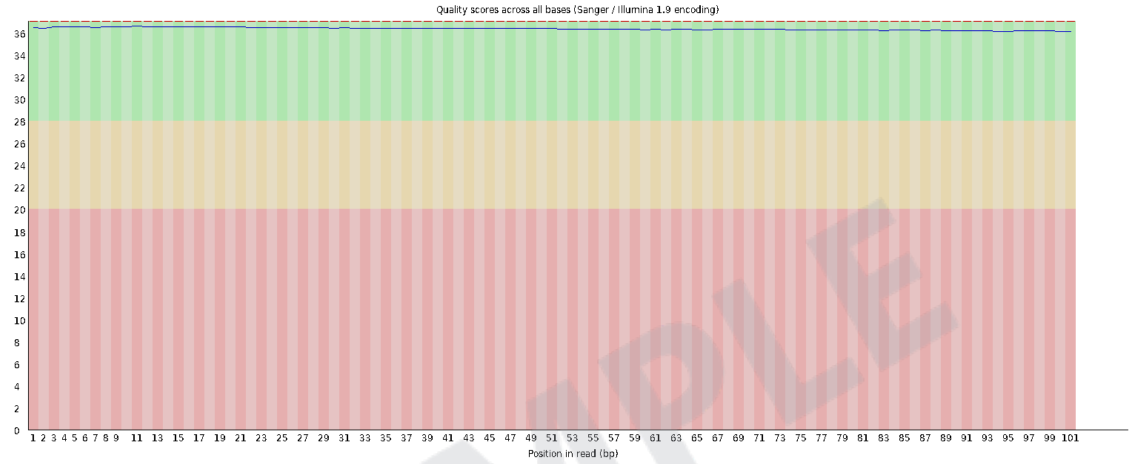


Figure 10. Read quality at each cycle of MG_CTRL_1 (read1)

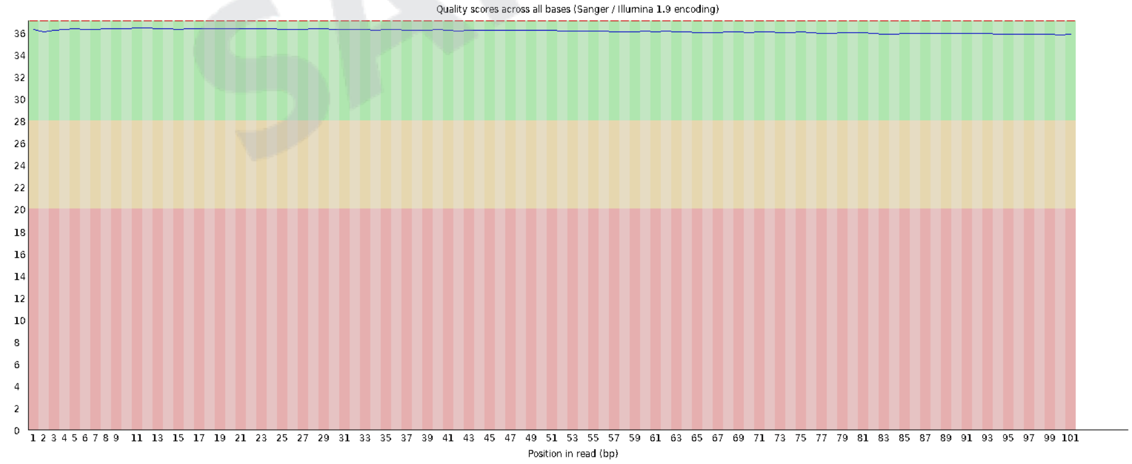


Figure 11. Read quality at each cycle of MG_CTRL_1 (read2)

- 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	6,649,641,185	66,167,454	47.98	98.88	95.92
MG_CTRL_2	6,420,603,246	63,918,246	48.15	98.83	95.8
MG_CTRL_3	7,372,930,453	73,393,434	48.44	98.82	95.75
MG_TEST_1	7,846,601,127	78,155,316	49.33	98.73	95.55
MG_TEST_2	6,237,415,963	62,086,894	48.93	98.86	95.89
MG_TEST_3	7,098,103,937	70,627,636	49.72	98.87	95.91

- 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.

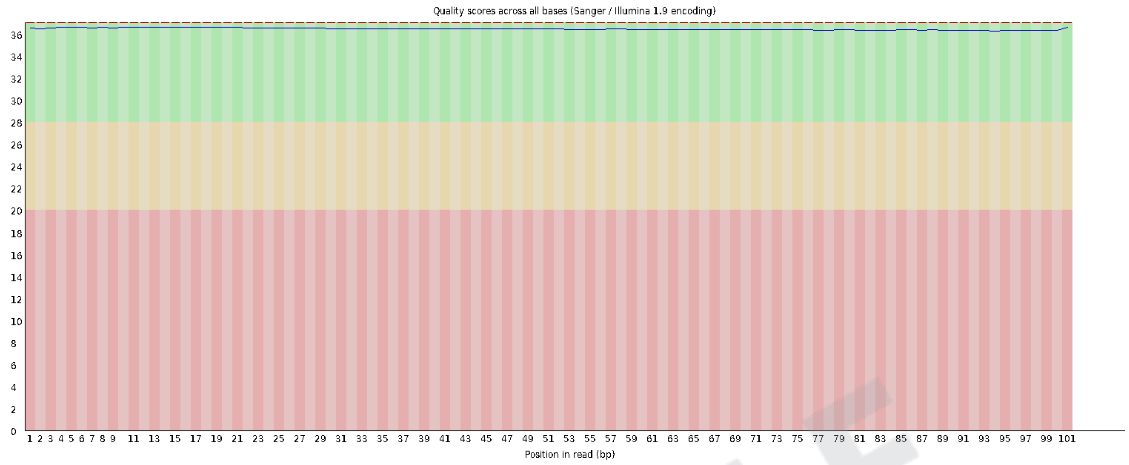


Figure 12. Average base quality of MG_CTRL_1 (read1) at each cycle after trimming

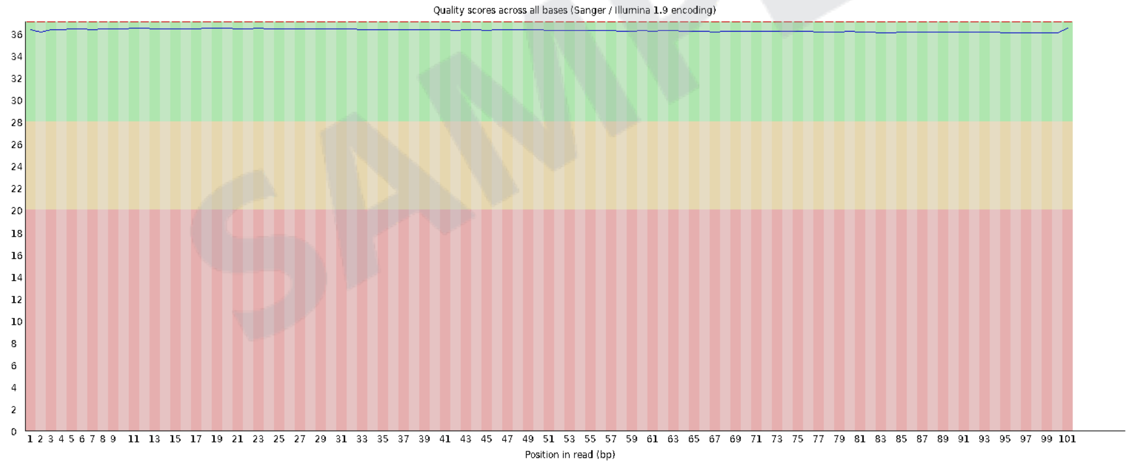


Figure 13. Average base quality of MG_CTRL_1 (read2) at each cycle after trimming

- 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.txt)

In order to map cDNA fragments obtained from RNA sequencing, GRCh38 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	66,167,454	65,473,700 (98.95%)	693,754 (1.05%)
MG_CTRL_2	63,918,246	63,213,512 (98.9%)	704,734 (1.1%)
MG_CTRL_3	73,393,434	72,562,152 (98.87%)	831,282 (1.13%)
MG_TEST_1	78,155,316	77,141,942 (98.7%)	1,013,374 (1.3%)
MG_TEST_2	62,086,894	61,249,232 (98.65%)	837,662 (1.35%)
MG_TEST_3	70,627,636	69,855,172 (98.91%)	772,464 (1.09%)

- 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.

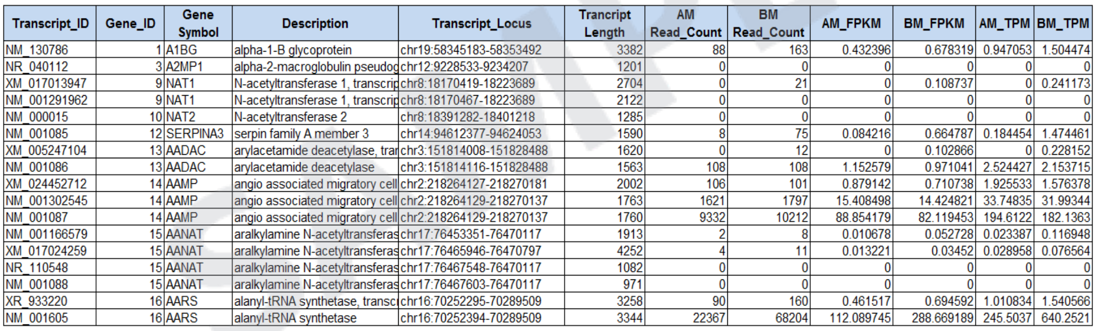
After assembly, the abundance of gene/transcript is calculated in the read count and normalized values as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Kilobase Million) for a sample.

4. 2. 1. Known Transcripts Expression Level

(Refer to Path: result_RNAseq/Expression_profile/StringTie/ Expression_Profile.GRCh38.transcript.xlsx)

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

Table 4. Known transcripts Expression Level (example)



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

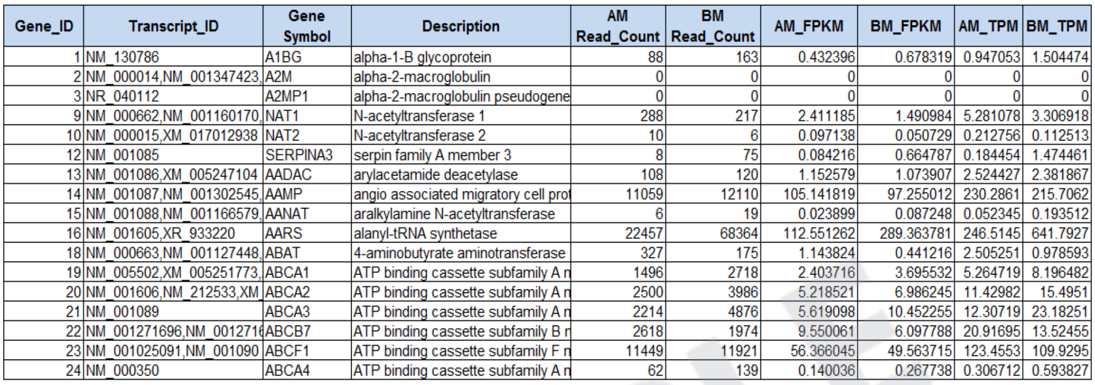
- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_ID: Gene ID
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- Transcript_Locus: Transript locus
- Transcript_Length: Transcript length
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM normalized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

4. 2. 2. Known Genes Expression Level

(Refer to Path: result_RNAseq/Expression_profile/StringTie/ Expression_Profile.GRCh38.gene.xlsx)

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

Table 5. Known genes Expression Level (example)



Gene_ID	Transcript_ID	Gene Symbol	Description	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM	AM_TPM	BM_TPM
1	NM_130786	A1BG	alpha-1-B glycoprotein	88	163	0.432396	0.678319	0.947053	1.504474
2	NM_000014,NM_001347423,	A2M	alpha-2-macroglobulin	0	0	0	0	0	0
3	NR_040112	A2MP1	alpha-2-macroglobulin pseudogene	0	0	0	0	0	0
9	NM_000662,NM_001160170,	NAT1	N-acetyltransferase 1	288	217	2.411185	1.490984	5.281078	3.306918
10	NM_000015,XM_017012938	NAT2	N-acetyltransferase 2	10	6	0.097138	0.050729	0.212756	0.112513
12	NM_001085	SERPINA3	serpin family A member 3	8	75	0.084216	0.664787	0.184454	1.474461
13	NM_001086,XM_005247104	AADAC	arylacetamide deacetylase	108	120	1.152579	1.073907	2.524427	2.381867
14	NM_001087,NM_001302545,	AAMP	angio associated migratory cell prot	11059	12110	105.141819	97.255012	230.2861	215.7062
15	NM_001088,NM_001166579,	AANAT	aralkylamine N-acetyltransferase	6	19	0.023899	0.087248	0.052345	0.193512
16	NM_001605,XR_933220	AARS	alanyl-tRNA synthetase	22457	68364	112.551262	289.363781	246.5145	641.7927
18	NM_000663,NM_001127448,	ABAT	4-aminobutyrate aminotransferase	327	175	1.143824	0.441216	2.505251	0.978593
19	NM_005502,XM_005251773,	ABCA1	ATP binding cassette subfamily A n	1496	2718	2.403716	3.695532	5.264719	8.196482
20	NM_001606,NM_212533,XM_	ABCA2	ATP binding cassette subfamily A n	2500	3986	5.218521	6.986245	11.42982	15.4951
21	NM_001089	ABCA3	ATP binding cassette subfamily A n	2214	4876	5.619098	10.452255	12.30719	23.18251
22	NM_001271696,NM_0012716	ABCB7	ATP binding cassette subfamily B n	2618	1974	9.550061	6.097788	20.91695	13.52455
23	NM_001025091,NM_001090	ABCF1	ATP binding cassette subfamily F n	11449	11921	56.366045	49.563715	123.4553	109.9295
24	NM_000350	ABCA4	ATP binding cassette subfamily A n	62	139	0.140036	0.267738	0.306712	0.593827

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

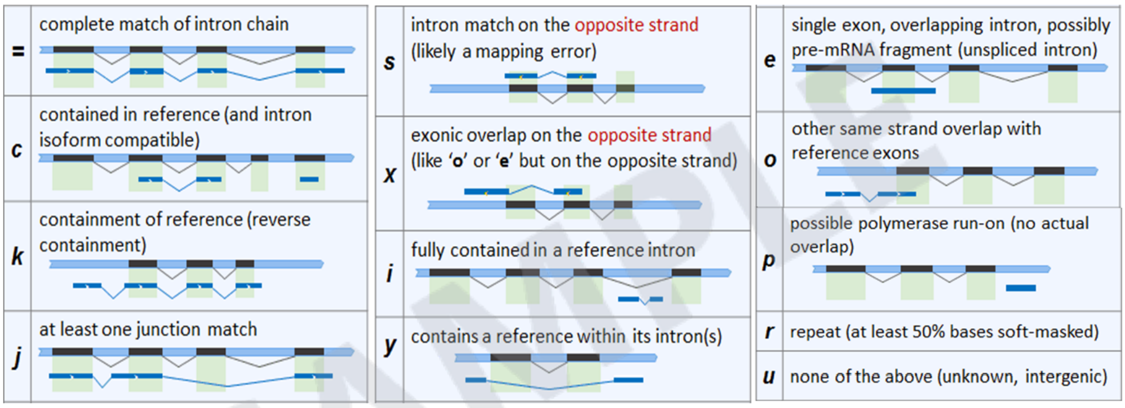
4. 3. Prediction of Novel Transcripts/Alternative Splicing

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
c	contained in reference (and intron isoform compatible)
k	containment of reference (reverse containment)
j	at least one junction match
s	intron match on the opposite strand (likely a mapping error)
x	exonic overlap on the opposite strand (like '**o**' or '**e**' but on the opposite strand)
i	fully contained in a reference intron
y	contains a reference within its intron(s)
e	single exon, overlapping intron, possibly pre-mRNA fragment (unspliced intron)
o	other same strand overlap with reference exons
p	possible polymerase run-on (no actual overlap)
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/Novel_transcript_analysis/StringTie/Expression_Profile_with_Novel.GRCh38.transcript.xlsx)

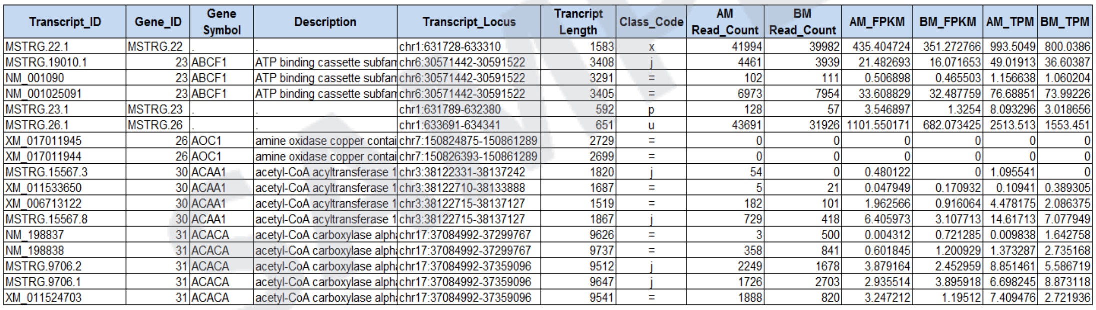
This result refers to expression level for each sample for each known transcript, novel transcript and novel alternative splicing transcript.

(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	Trancript Length	Class_Code	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM	AM_TPM	BM_TPM
MSTRG.22.1	MSTRG.22	.	.	chr1:631728-633310	1583	x	41994	39982	435.404724	351.272766	993.5049	800.0386
MSTRG.19010.1	23	ABCF1	ATP binding cassette subfan	chr6:30571442-30591522	3408	j	4461	3939	21.482693	16.071653	49.01913	36.60387
NM_001090	23	ABCF1	ATP binding cassette subfan	chr6:30571442-30591522	3291	=	102	111	0.506898	0.465503	1.156638	1.060204
NM_001025091	23	ABCF1	ATP binding cassette subfan	chr6:30571442-30591522	3405	=	6973	7954	33.608829	32.487759	76.68851	73.99226
MSTRG.23.1	MSTRG.23	.	.	chr1:631789-632380	592	p	128	57	3.546897	1.3254	8.093296	3.018656
MSTRG.26.1	MSTRG.26	.	.	chr1:633691-634341	651	u	43691	31926	1101.550171	682.073425	2513.513	1553.451
XM_017011945	26	AOC1	amine oxidase copper contai	chr7:150824875-150861289	2729	=	0	0	0	0	0	0
XM_017011944	26	AOC1	amine oxidase copper contai	chr7:150826393-150861289	2699	=	0	0	0	0	0	0
MSTRG.15567.3	30	ACAA1	acetyl-CoA acyltransferase 1	chr3:38122331-38137242	1820	j	54	0	0.480122	0	1.095541	0
XM_011533650	30	ACAA1	acetyl-CoA acyltransferase 1	chr3:38122710-38133888	1687	=	5	21	0.047949	0.170932	0.10941	0.389305
XM_006713122	30	ACAA1	acetyl-CoA acyltransferase 1	chr3:38122715-38137127	1519	=	182	101	1.962566	0.916064	4.478175	2.086375
MSTRG.15567.8	30	ACAA1	acetyl-CoA acyltransferase 1	chr3:38122715-38137127	1867	j	729	418	6.405973	3.107713	14.61713	7.077949
NM_198837	31	ACACA	acetyl-CoA carboxylase alph	chr17:37084992-37299767	9626	=	3	500	0.004312	0.721285	0.009838	1.642758
NM_198838	31	ACACA	acetyl-CoA carboxylase alph	chr17:37084992-37299767	9737	=	358	841	0.601845	1.200929	1.373287	2.735168
MSTRG.9706.2	31	ACACA	acetyl-CoA carboxylase alph	chr17:37084992-37359096	9512	j	2249	1678	3.879164	2.452959	8.851461	5.586719
MSTRG.9706.1	31	ACACA	acetyl-CoA carboxylase alph	chr17:37084992-37359096	9647	j	1726	2703	2.935514	3.895918	6.698245	8.873118
XM_011524703	31	ACACA	acetyl-CoA carboxylase alph	chr17:37084992-37359096	9541	=	1888	820	3.247212	1.19512	7.409476	2.721936

- 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)
- [Sample Name]_TPM: TPM normalized value of a sample

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

(Refer to Path: result_RNAseq/Novel_transcript_analysis/StringTie/ Expression_Profile_with_Novel.GRCh38.gene.xlsx)

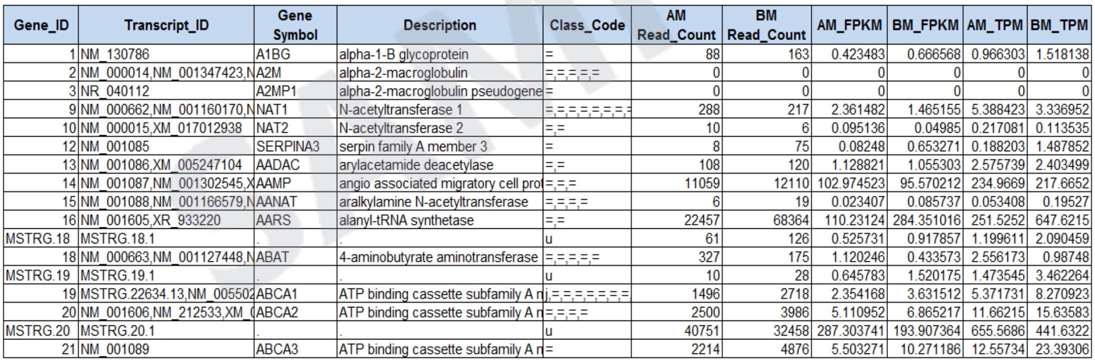
This result refers to expression level for each sample for gene containing known transcripts and novel alternative splicing transcript or novel gene.

(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	AM_TPM	BM_TPM
1	NM_130786	A1BG	alpha-1-B glycoprotein	=	88	163	0.423483	0.666568	0.966303	1.518138
2	NM_000014,NM_001347423,N	A2M	alpha-2-macroglobulin	=,=,=,=,=	0	0	0	0	0	0
3	NR_040112	A2MP1	alpha-2-macroglobulin pseudogene	=	0	0	0	0	0	0
9	NM_000662,NM_001160170,N	NAT1	N-acetyltransferase 1	=,=,=,=,=,=,=,=,=	288	217	2.361482	1.465155	5.388423	3.336952
10	NM_000015,XM_017012938	NAT2	N-acetyltransferase 2	=,=	10	6	0.095136	0.04985	0.217081	0.113535
12	NM_001085	SERPINA3	serpin family A member 3	=	8	75	0.08248	0.653271	0.188203	1.487852
13	NM_001086,XM_005247104	AADAC	arylacetamide deacetylase	=,=	108	120	1.128821	1.055303	2.575739	2.403499
14	NM_001087,NM_001302545,X	AAMP	angio associated migratory cell prot	=,=,=	11059	12110	102.974523	95.570212	234.9669	217.6652
15	NM_001088,NM_001166579,N	AANAT	aralkylamine N-acetyltransferase	=,=,=,=	6	19	0.023407	0.085737	0.053408	0.19527
16	NM_001605,XR_933220	AARS	alanyl-tRNA synthetase	=,=	22457	68364	110.23124	284.351016	251.5252	647.6215
MSTRG.18	MSTRG.18.1	.	.	u	61	126	0.525731	0.917857	1.199611	2.090459
18	NM_000663,NM_001127448,N	ABAT	4-aminobutyrate aminotransferase	=,=,=,=,=	327	175	1.120246	0.433573	2.556173	0.98748
MSTRG.19	MSTRG.19.1	.	.	u	10	28	0.645783	1.520175	1.473545	3.462264
19	MSTRG.22634.13,NM_005502	ABCA1	ATP binding cassette subfamily A n	j,=,=,=,=,=,=,=,	1496	2718	2.354168	3.631512	5.371731	8.270923
20	NM_001606,NM_212533,XM_(	ABCA2	ATP binding cassette subfamily A n	=,=,=,=	2500	3986	5.110952	6.865217	11.66215	15.63583
MSTRG.20	MSTRG.20.1	.	.	u	40751	32458	287.303741	193.907364	655.5686	441.6322
21	NM_001089	ABCA3	ATP binding cassette subfamily A n	=	2214	4876	5.503271	10.271186	12.55734	23.39306

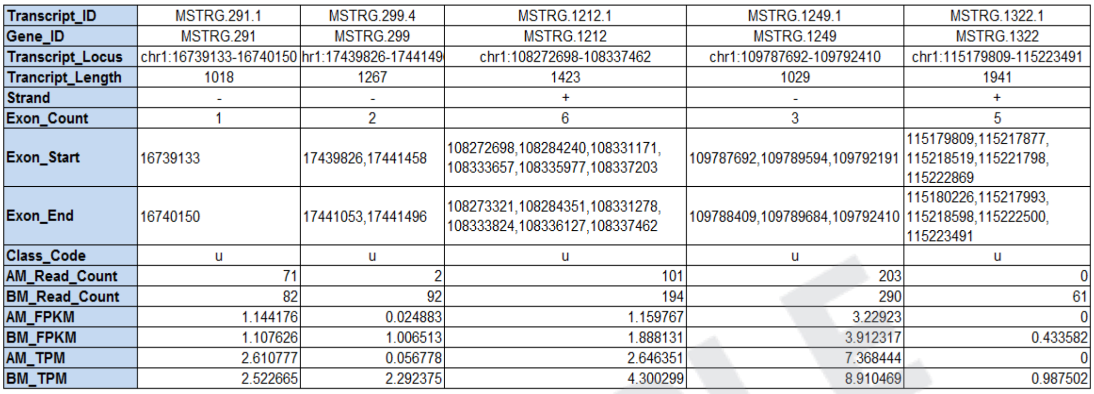
- Gene_ID: Entrez gene ID
- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- Class_Code: Class code corresponding to transcript ID (Refer to Table 6)
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM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

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.291.1	MSTRG.299.4	MSTRG.1212.1	MSTRG.1249.1	MSTRG.1322.1
Gene_ID	MSTRG.291	MSTRG.299	MSTRG.1212	MSTRG.1249	MSTRG.1322
Transcript_Locus	chr1:16739133-16740150	hr1:17439826-1744149	chr1:108272698-108337462	chr1:109787692-109792410	chr1:115179809-115223491
Trancript_Length	1018	1267	1423	1029	1941
Strand	-	-	+	-	+
Exon_Count	1	2	6	3	5
Exon_Start	16739133	17439826,17441458	108272698,108284240,108331171, 108333657,108335977,108337203	109787692,109789594,109792191	115179809,115217877, 115218519,115221798, 115222869
Exon_End	16740150	17441053,17441496	108273321,108284351,108331278, 108333824,108336127,108337462	109788409,109789684,109792410	115180226,115217993, 115218598,115222500, 115223491
Class_Code	u	u	u	u	u
AM_Read_Count	71	2	101	203	0
BM_Read_Count	82	92	194	290	61
AM_FPKM	1.144176	0.024883	1.159767	3.22923	0
BM_FPKM	1.107626	1.006513	1.888131	3.912317	0.433582
AM_TPM	2.610777	0.056778	2.646351	7.368444	0
BM_TPM	2.522665	2.292375	4.300299	8.910469	0.987502

- 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)
- [Sample Name]_TPM: TPM normalized value of a sample

4. 3. 4. Filtering Novel alternative splicing transcript

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

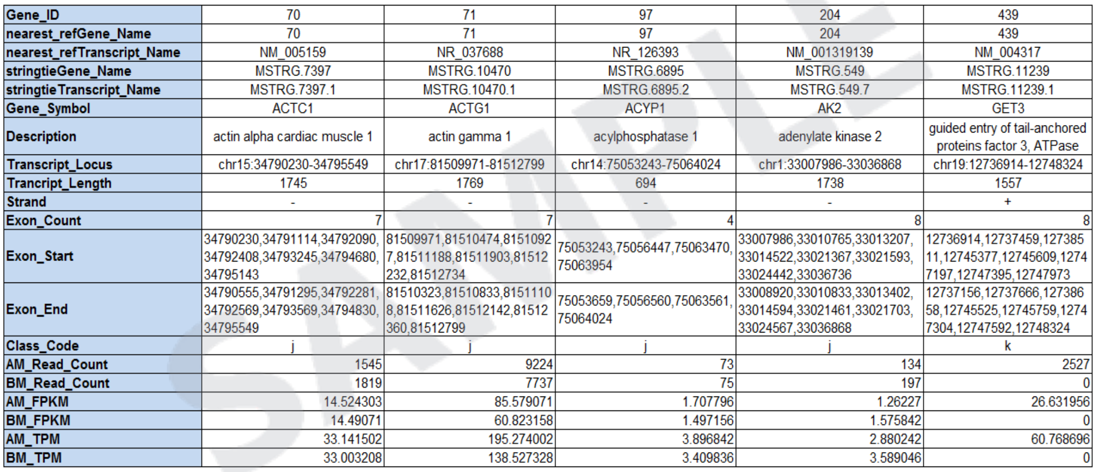
This result refers to the list of novel alternative splicing transcripts filtered by class code ('j', 'c', 'k', 'e', 'i', 'o', 'p', 's', 'x') from the results of known and novel transcripts.

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)



Gene_ID	70	71	97	204	439
nearest_refGene_Name	70	71	97	204	439
nearest_refTranscript_Name	NM_005159	NR_037688	NR_126393	NM_001319139	NM_004317
stringtieGene_Name	MSTRG.7397	MSTRG.10470	MSTRG.6895	MSTRG.549	MSTRG.11239
stringtieTranscript_Name	MSTRG.7397.1	MSTRG.10470.1	MSTRG.6895.2	MSTRG.549.7	MSTRG.11239.1
Gene_Symbol	ACTC1	ACTG1	ACYP1	AK2	GET3
Description	actin alpha cardiac muscle 1	actin gamma 1	acylphosphatase 1	adenylate kinase 2	guided entry of tail-anchored proteins factor 3, ATPase
Transcript_Locus	chr15:34790230-34795549	chr17:81509971-81512799	chr14:75053243-75064024	chr1:33007986-33036868	chr19:12736914-12748324
Trancript_Length	1745	1769	694	1738	1557
Strand	-	-	-	-	+
Exon_Count	7	7	4	8	8
Exon_Start	34790230,34791114,34792090,34792408,34793245,34794680,34795143	81509971,81510474,81510927,81511188,81511903,81512232,81512734	75053243,75056447,75063470,75063954	33007986,33010765,33013207,33014522,33021367,33021593,33024442,33036736	12736914,12737459,12738511,12745377,12745609,12747197,12747395,12747973
Exon_End	34790555,34791295,34792281,34792569,34793569,34794830,34795549	81510323,81510833,81511108,81511626,81512142,81512360,81512799	75053659,75056560,75063561,75064024	33008920,33010833,33013402,33014594,33021461,33021703,33024567,33036868	12737156,12737666,12738658,12745525,12745759,12747304,12747592,12748324
Class_Code	j	j	j	j	k
AM_Read_Count	1545	9224	73	134	2527
BM_Read_Count	1819	7737	75	197	0
AM_FPKM	14.524303	85.579071	1.707796	1.26227	26.631956
BM_FPKM	14.49071	60.823158	1.497156	1.575842	0
AM_TPM	33.141502	195.274002	3.896842	2.880242	60.768696
BM_TPM	33.003208	138.527328	3.409836	3.589046	0

- Gene_ID: Gene ID
- nearest_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: Trnascript 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 Name]_TPM: TPM 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

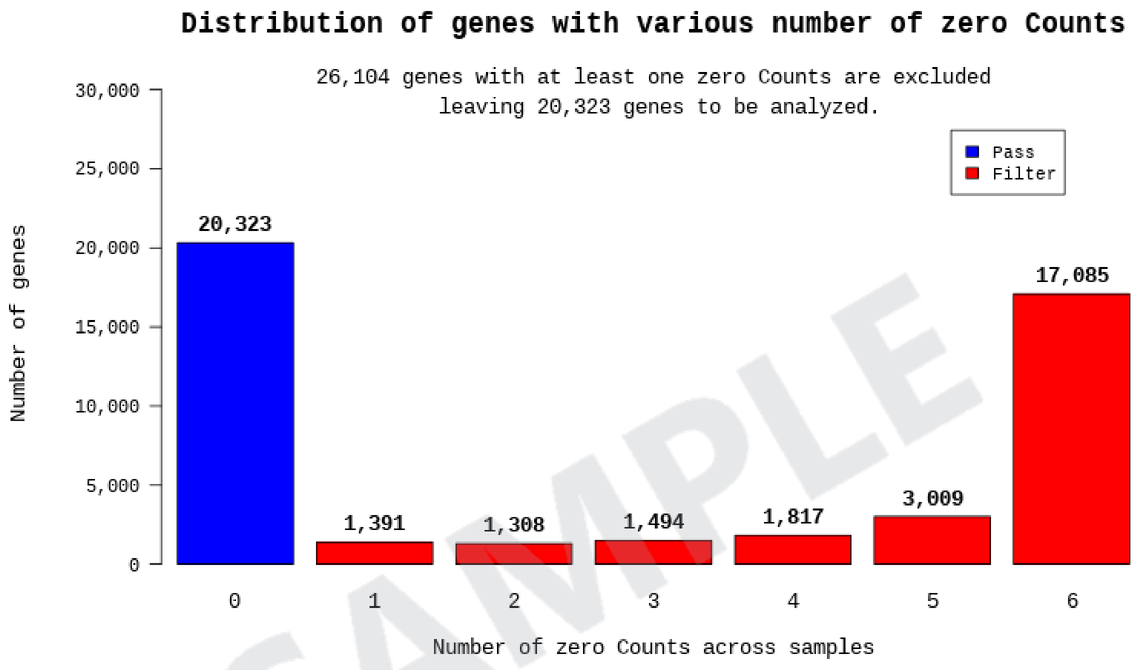
Comparison pair and statistical method for each pair are shown below.

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 46,427 genes, 26,104 were excluded and only 20,323 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.

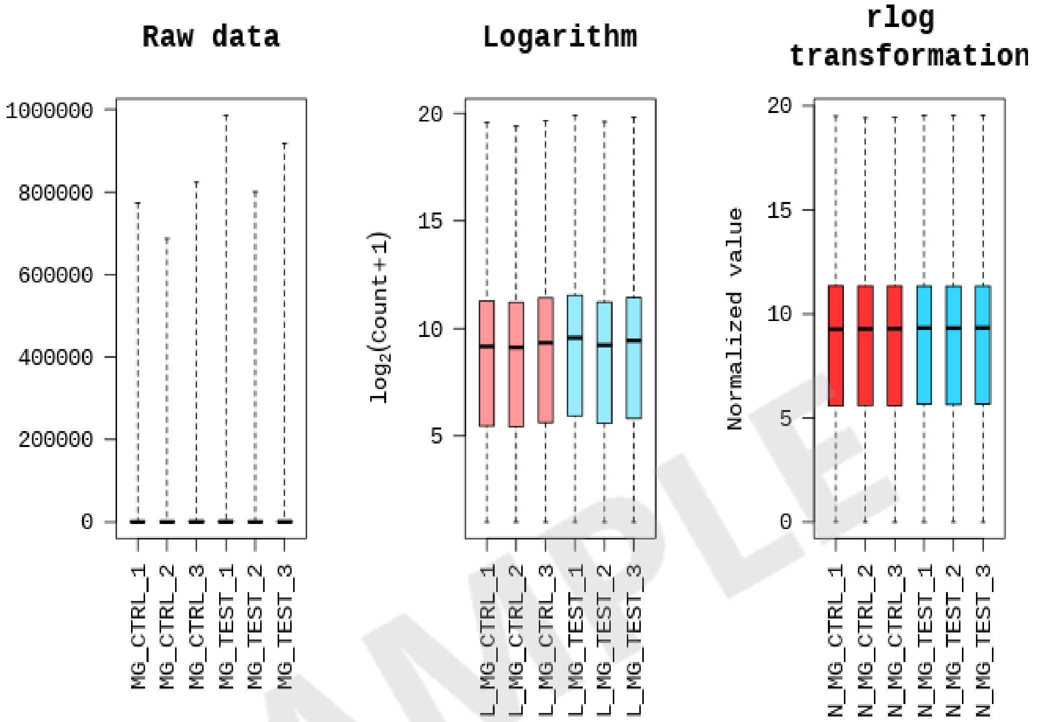
log2(read count+1) and regularized log (rlog) transformed values were used for data visualization. rlog transformation is a method to minimize differences between samples for genes/transcripts in low expression. It transforms count data into log2 scale and normalizes them with a library size factor.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

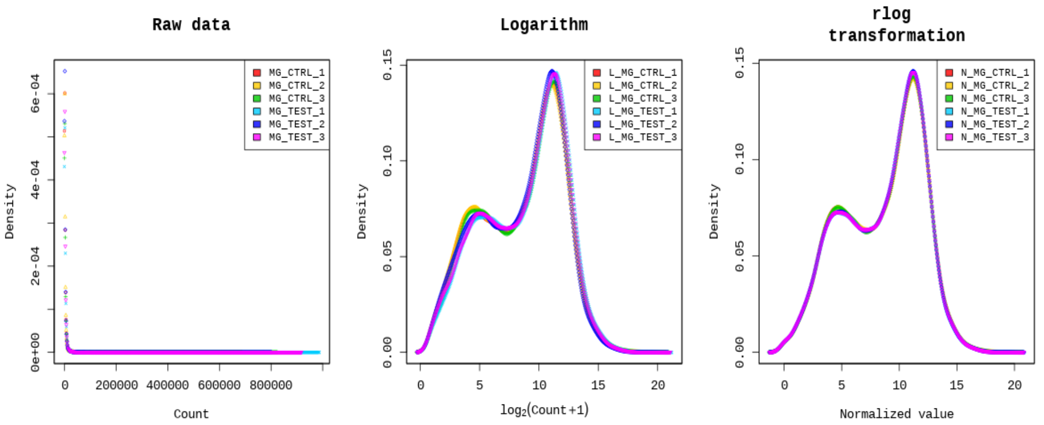
5. 1. 3. 1. Boxplot of Expression Difference between samples.

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

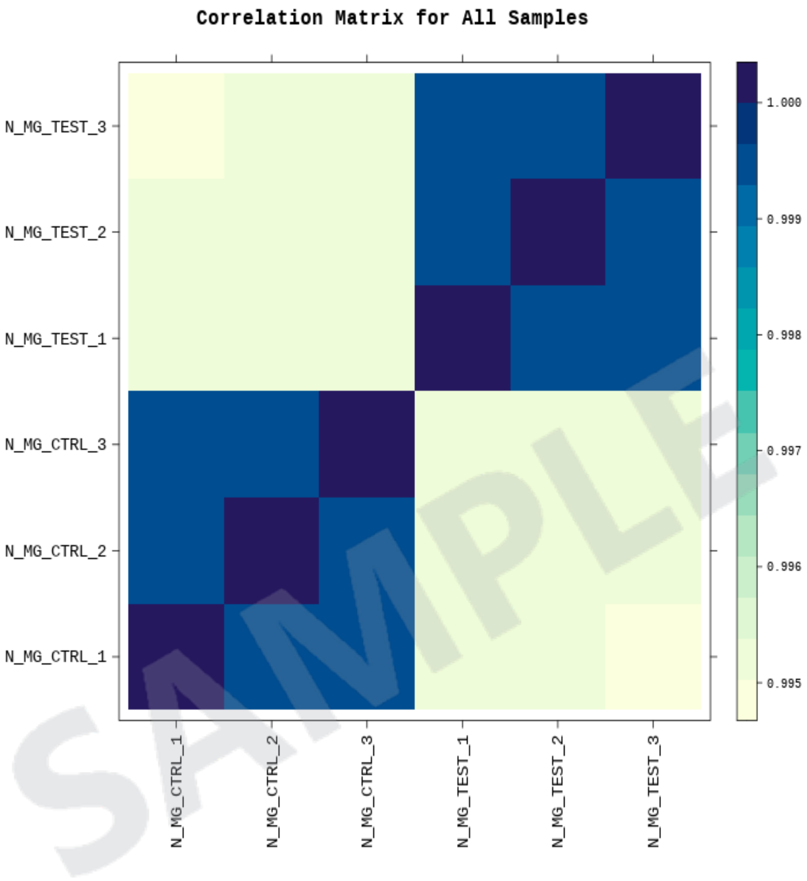
Below density plots show the corresponding samples expression distribution before and after of raw signal (read count), Log2 transformation of read count+1 and RLE Normalization.



5. 1. 4. Correlation Analysis between samples

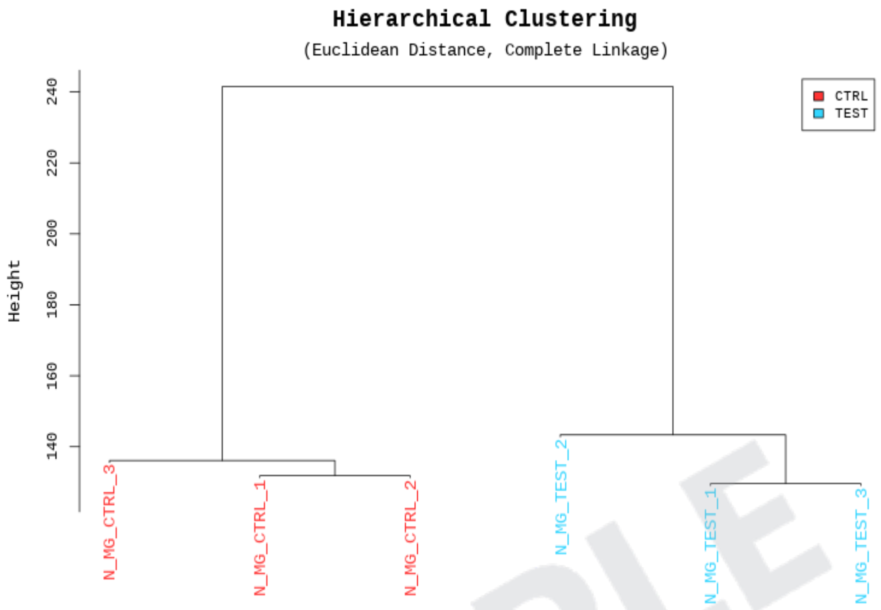
The similarity between samples are obtained through Pearson's coefficient of the rlog transformed value. For range: -1≤ r ≤ 1,the closer the value is to 1, the more similar the samples are.

Correlation matrix of all samples is as follows.



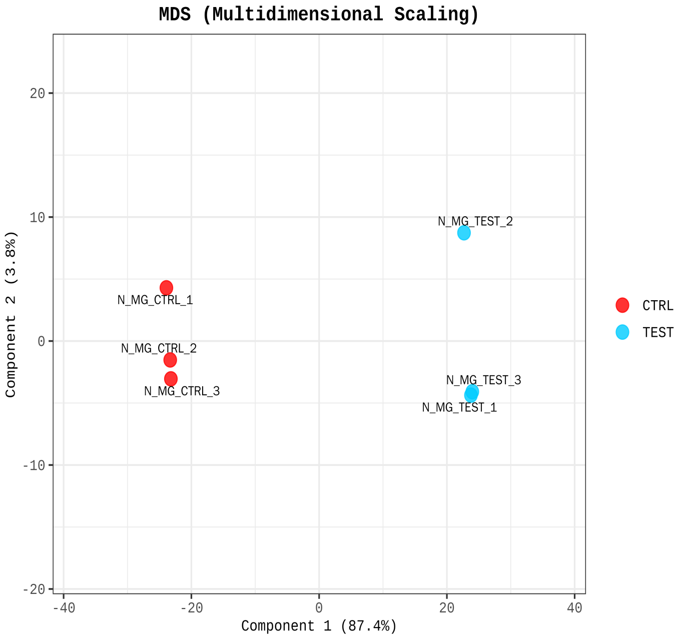
5. 1. 5. Hierarchical Clustering Analysis

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 -e option of the StringTie were used as the original raw data.

- Raw data

(Refer to Path: result_RNAseq/Expression_profile/StringTie/
Expression_Profile.GRCh38.gene.xlsx)

: 46,427 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)

: 20,323 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)

: 2,700 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.

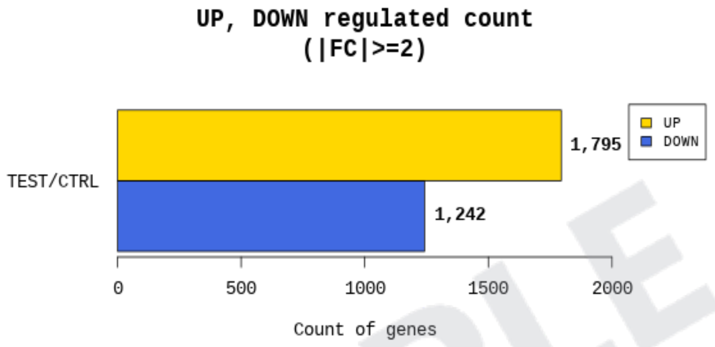
5. 3. Significant Gene Results

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

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

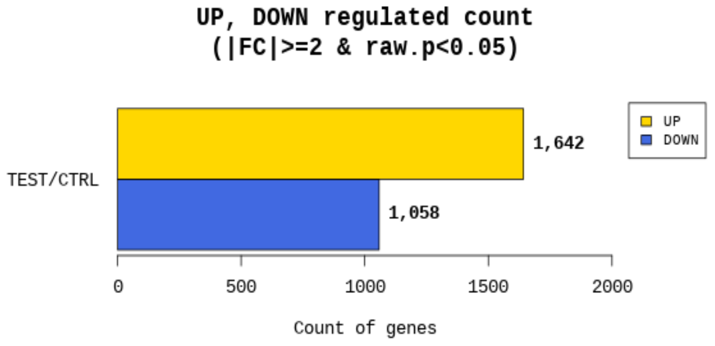
5. 3. 1. Up, Down Regulated Count by Fold Change

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



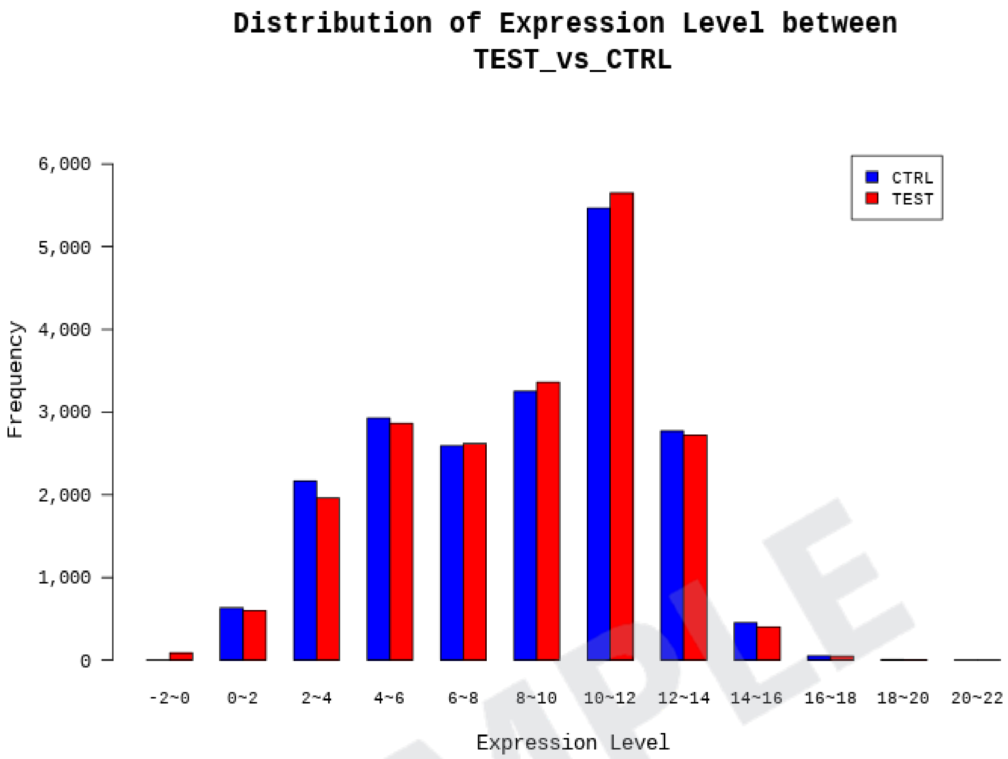
5. 3. 2. Up, Down Regulated Count by Fold Change and p-value

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



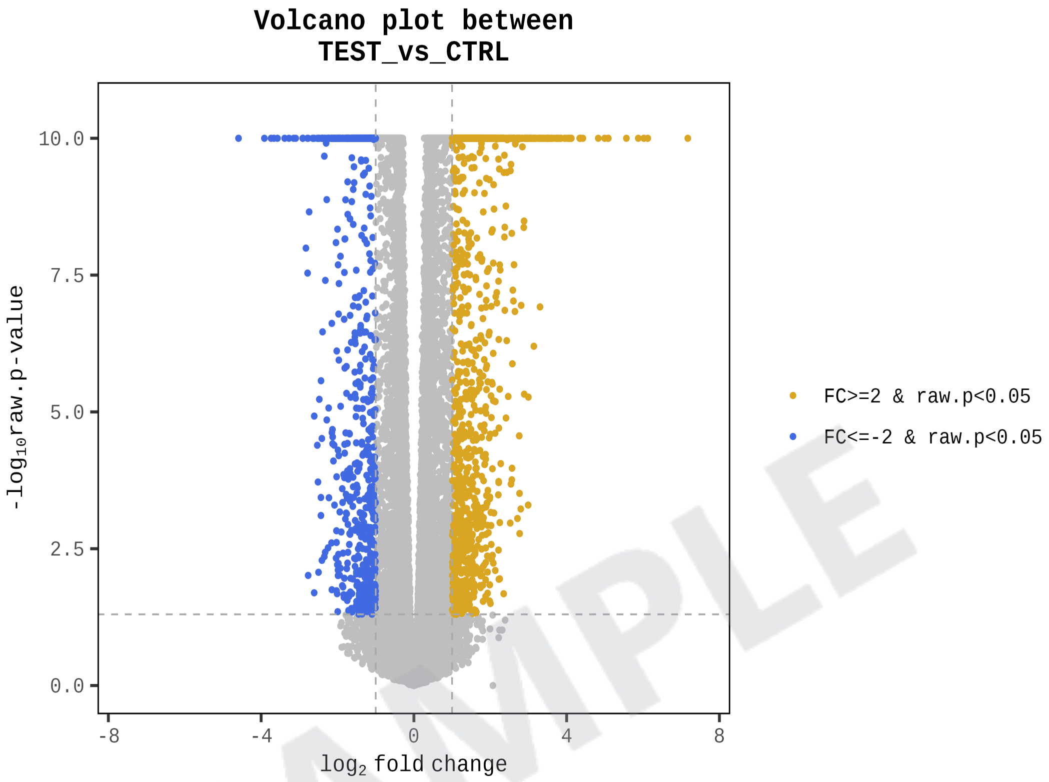
5. 3. 3. Distribution of Expression Level between two groups

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



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

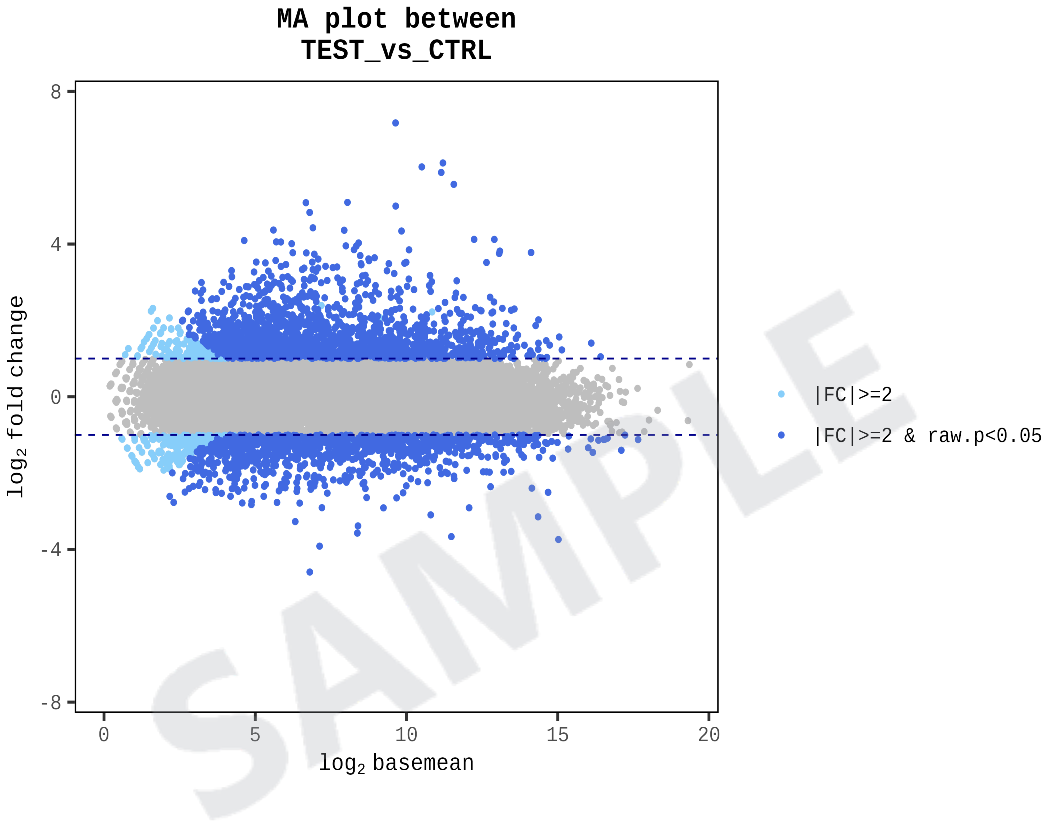
Log2 fold change and p-value obtained from the comparison between two groups plotted as volcano plot.
(X-axis: log2 Fold Change, Y-axis: -log10 p-value)



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).

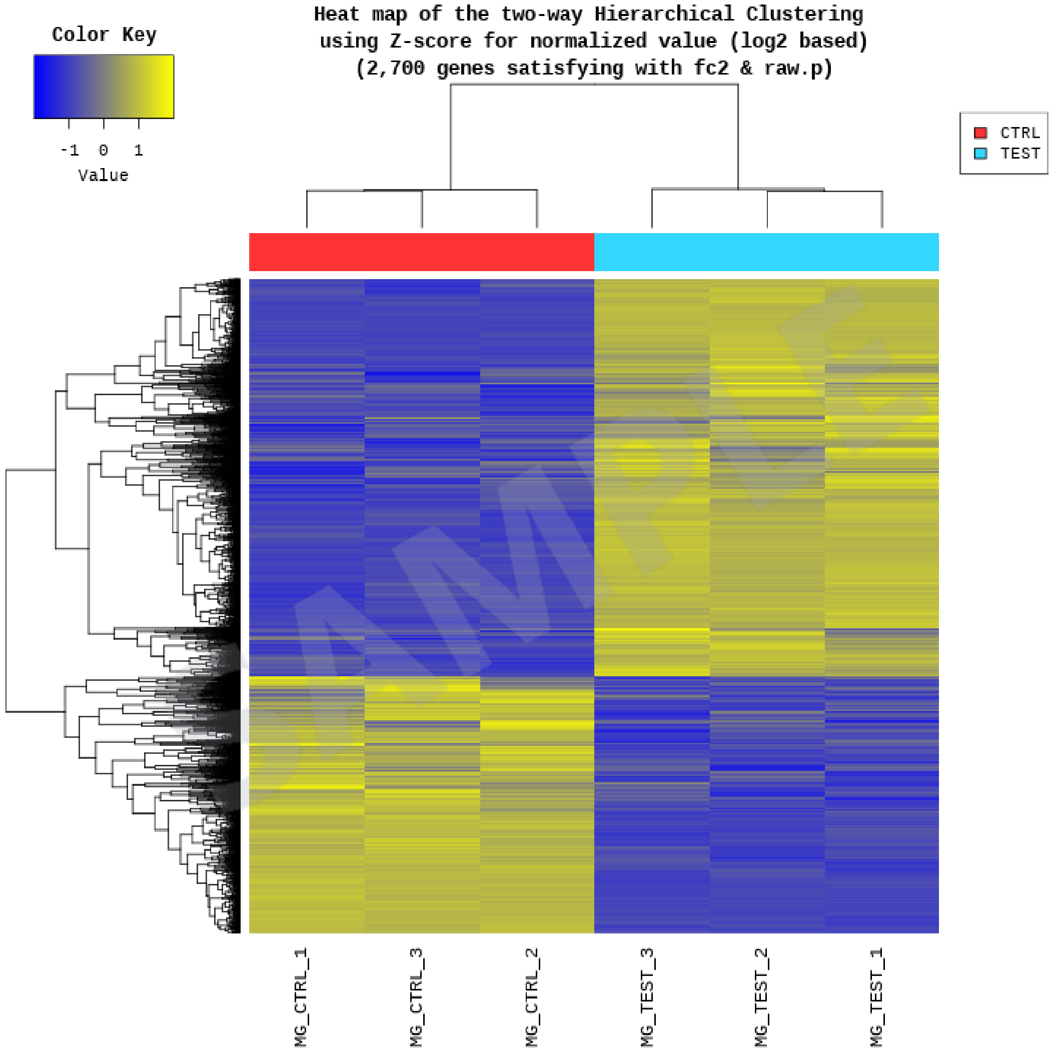
For example, even though fold change might be different by two-fold, the gene with higher mean of normalized counts may be more credible.



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.

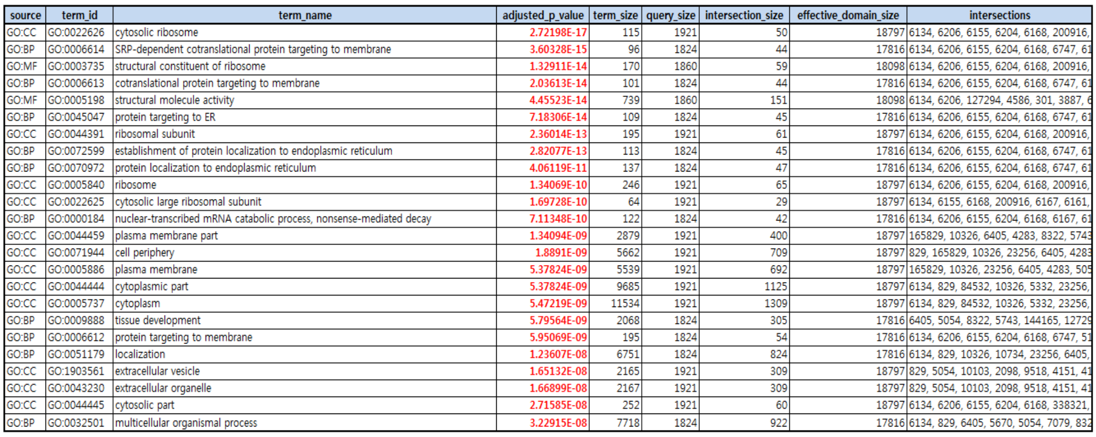
- Link for the ontology documentation: http://geneontology.org/page/ontology-documentation
- Link for the ontology files: http://geneontology.org/page/download-ontology
- Link for the annotation files: http://geneontology.org/page/download-annotations

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.

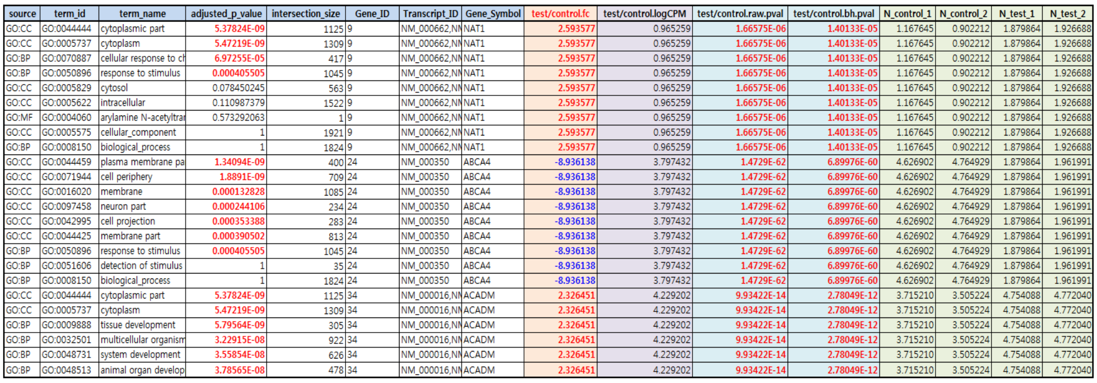


source	term_id	term_name	adjusted_p_value	term_size	query_size	intersection_size	effective_domain_size	intersections
GO:CC	GO:0022626	cytosolic ribosome	2.72198E-17	115	1921	50	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	3.60328E-15	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, 41
GO:CC	GO:0043230	extracellular organelle	1.66899E-08	2167	1921	309	18797	829, 5054, 10103, 2098, 9518, 4151, 41
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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	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,N	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040

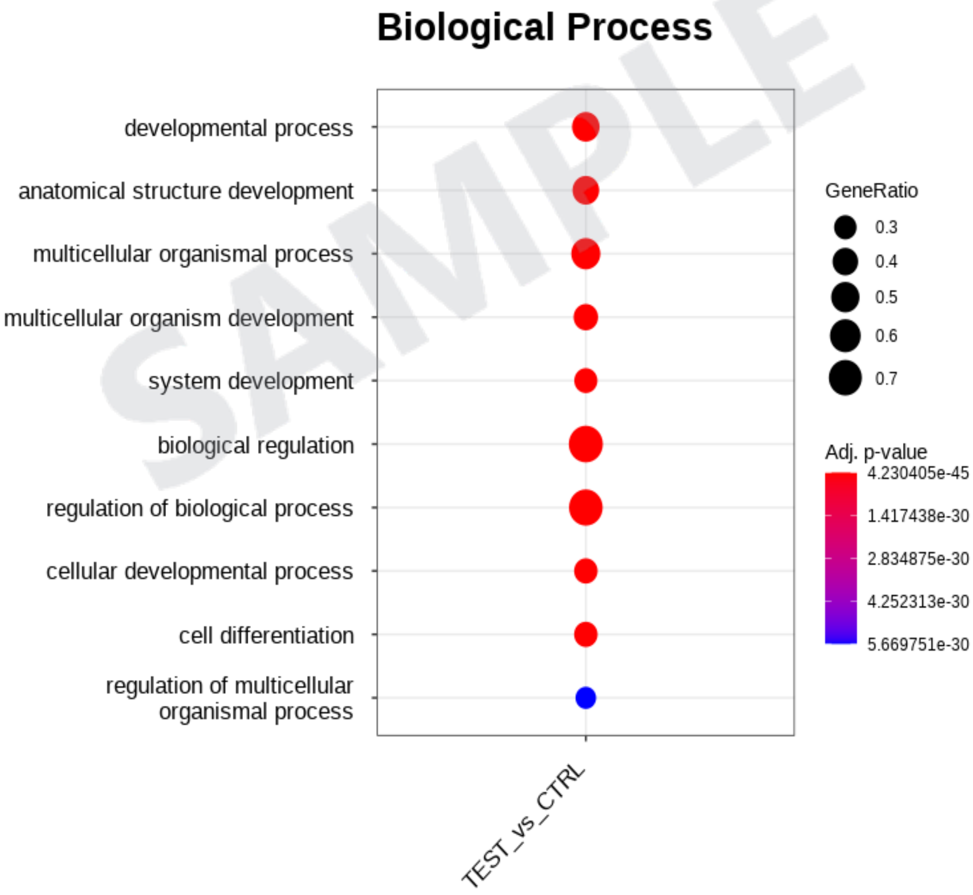
- source: Code for the data source. Ex> GO:BP | GO:CC | GO:MF ...
- term_id: ID for the enriched term/functional category
- term_name: Readable name for the enriched term
- adjusted_p_value: Adjusted p-value by FDR
- intersection_size: The number of unique DEG that are annotated to the term_id

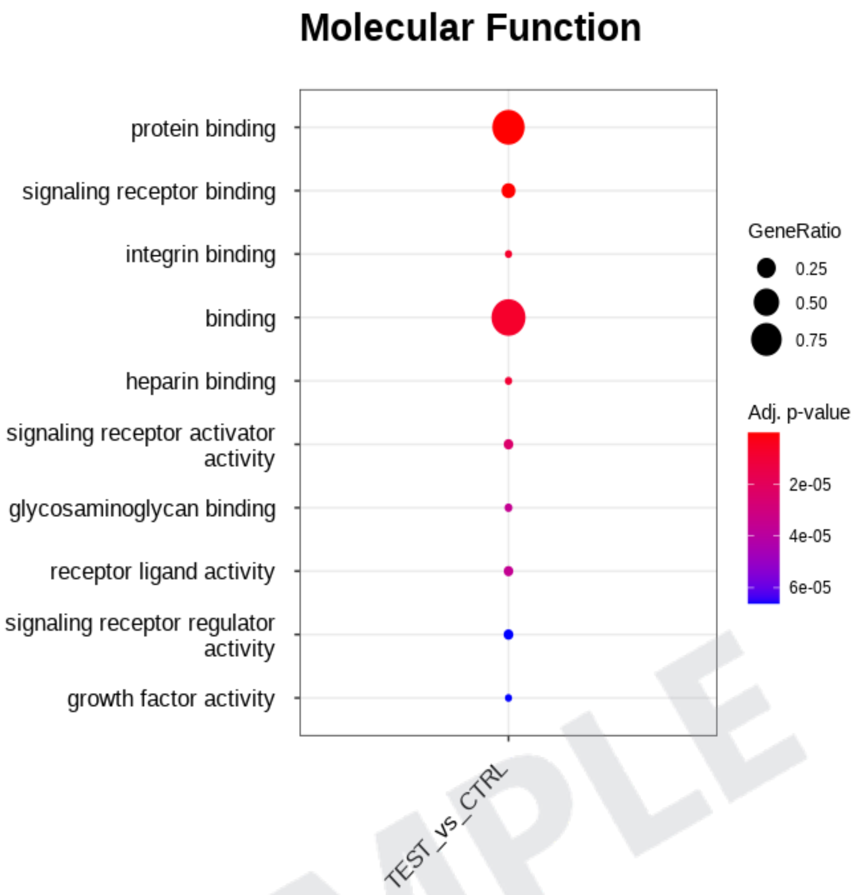
data3.GO_*.gprofiler.png: Top 20 terms of Gene Ontology Enrichment Analysis result were described by dot plot.
(Plotting based on GO_stat)

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.)

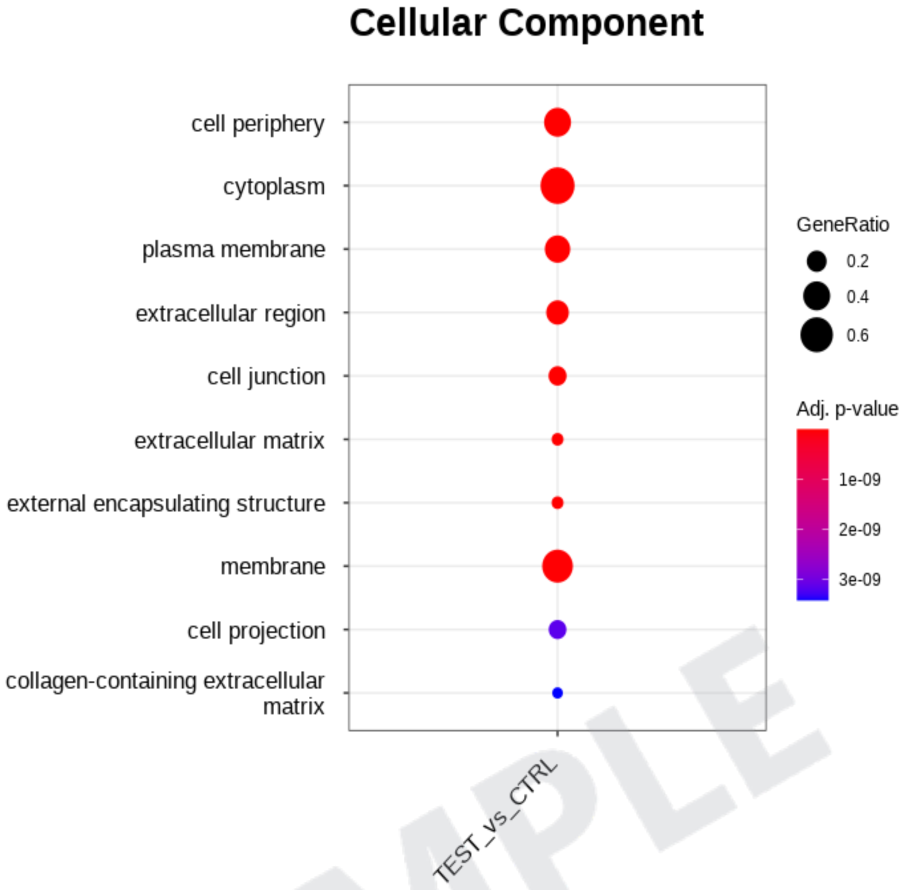
- term_size filtering: The GO Terms that are very large or small do not contribute to interpretability of results, and their statistical significance can be inflated when using certain statistical enrichment methods (e.g., Hypergeometric test).
- GeneRatio: GeneRatio is calculated as the ratio of intersection_size 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
integrin binding
binding
heparin binding
signaling receptor activator activity
glycosaminoglycan binding
receptor ligand activity
signaling receptor regulator activity
growth factor activity
TEST_vs_CTRL
GeneRatio
0.25
0.50
0.75
Adj. p-value
2e-05
4e-05
6e-05


Cellular Component
cell periphery
cytoplasm
plasma membrane
extracellular region
cell junction
extracellular matrix
external encapsulating structure
membrane
cell projection
collagen-containing extracellular matrix
TEST_vs_CTRL
GeneRatio
0.2
0.4
0.6
Adj. p-value
1e-09
2e-09
3e-09

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.

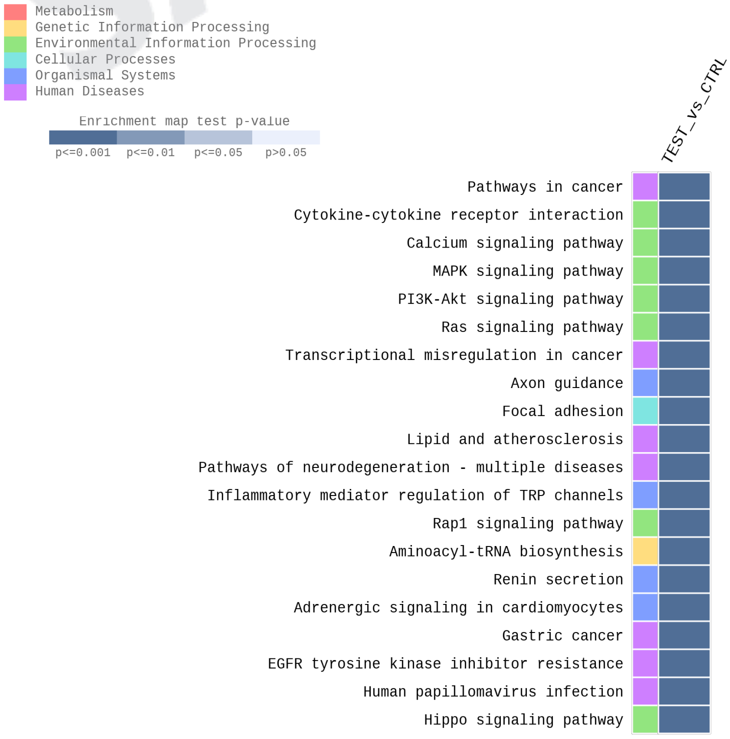
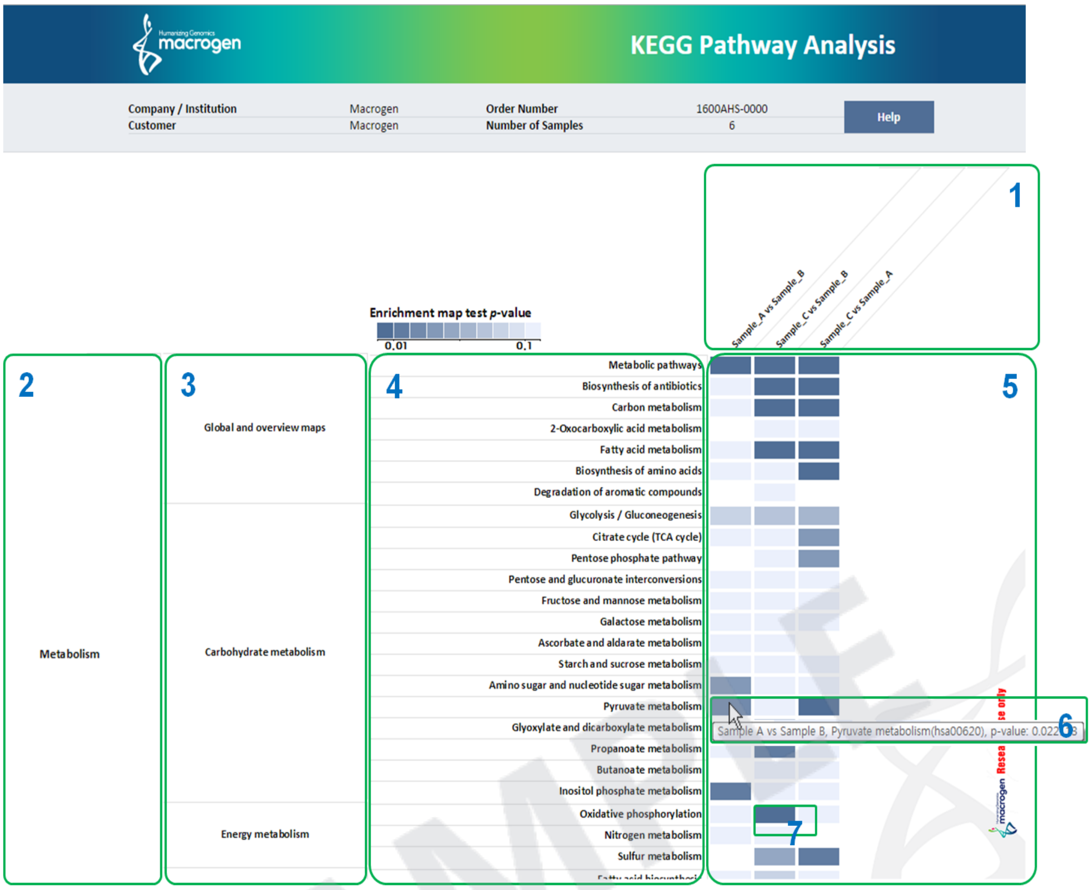


Figure 14. Result of gene-set enrichment analysis (p-value top 20)

5. 5. 1. KEGG HTML Viewer



KEGG Pathway Analysis

Company / Institution	Macrogen	Order Number	1600AHS-0000	Help
Customer	Macrogen	Number of Samples	6	

Sample A vs Sample B, Pyruvate metabolism(hsa00620), p-value: 0.022

macrogen
1
Sample_A vs Sample_B
Sample_C vs Sample_B
Sample_C vs Sample_A
Enrichment map test p-value
0.01
0.1
2
Metabolism
3
Global and overview maps
Carbohydrate metabolism
Energy metabolism
4
Metabolic pathways
Biosynthesis of antibiotics
Carbon metabolism
2-Oxocarboxylic acid metabolism
Fatty acid metabolism
Biosynthesis of amino acids
Degradation of aromatic compounds
Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Starch and sucrose metabolism
Amino sugar and nucleotide sugar metabolism
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Butanoate metabolism
Inositol phosphate metabolism
Oxidative phosphorylation
Nitrogen metabolism
Sulfur metabolism
5
6
7
macrogen Research use only

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.

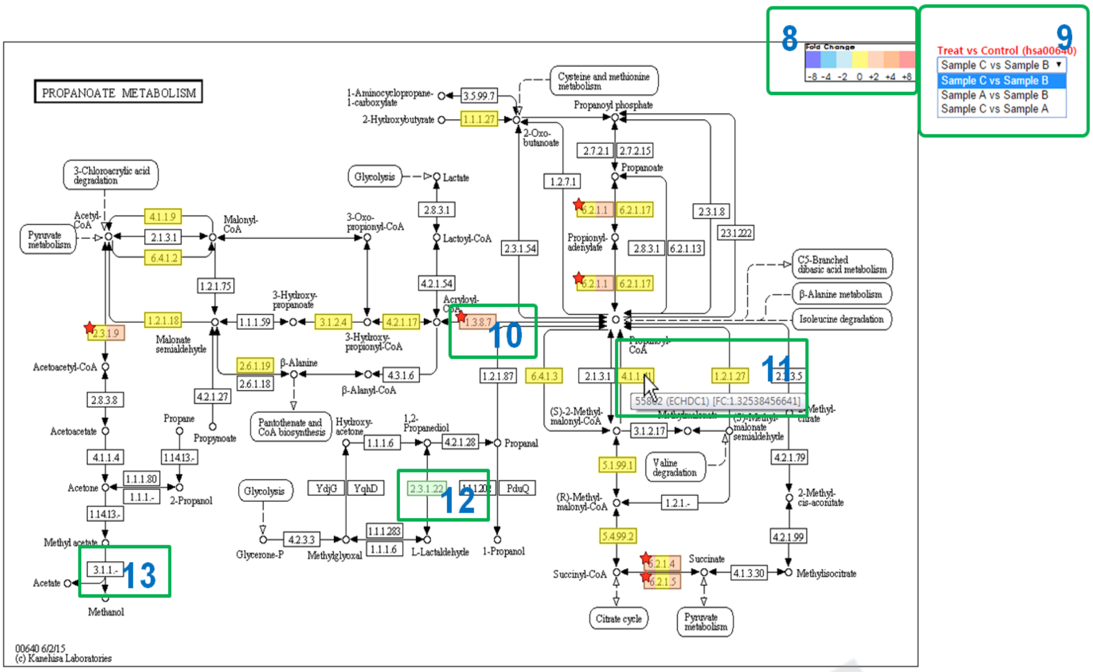


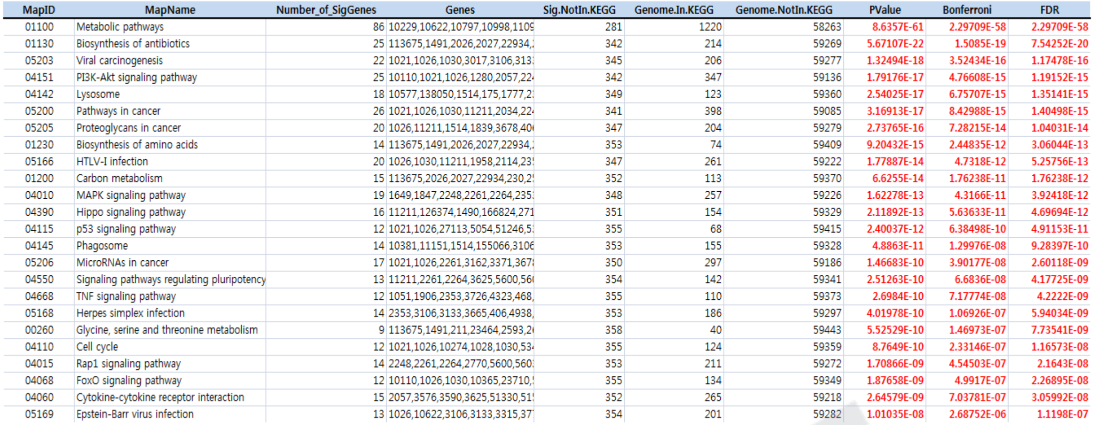
Figure 16. Description of KEGG pathway map frame

- 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



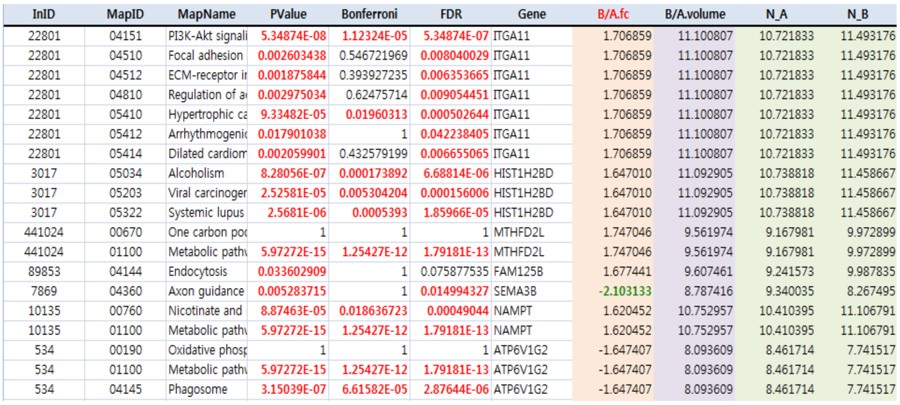
MapID	MapName	Number_of_SigGenes	Genes	Sig.NotIn.KEGG	Genome.In.KEGG	Genome.NotIn.KEGG	PValue	Bonferroni	FDR
01100	Metabolic pathways	86	10229,10622,10797,10998,1109	281	1220	58263	8.6357E-61	2.29709E-58	2.29709E-58
01130	Biosynthesis of antibiotics	25	113675,1491,2026,2027,22934,	342	214	59269	5.67107E-22	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 poc	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. SNP and Indel Analysis

6. 1. SNP and Indel Discovery

(Refer to Path: result_RNAseq/Variant_calling/STAR_GATK/VCF_files/*.rawVariants.vcf)

Identifying short variants (SNPs and Indels) is preformed according to the GATK's best practices workflow for RNA-Seq. SNV calling workflow is summarized in the following several steps. First, the trimmed reads are aligned to the reference genome using the STAR program. And pre-processing step such as mark duplication, sort, split 'N' trim, and base recalibration is performed. In the final step, HaplotypeCaller is used to call the SNP/Indel variants for each sample.

6. 2. Variant filtering and annotation

(Refer to Path: result_RNAseq/Variant_calling/STAR_GATK/SNV_Call_*.xlsx)

High quality variants are filtered by PASS filters (Fisher Strand values, FS > 30.0 and Quality By Depth values, QD < 2.0) in the VariantFilteration module and depth coverage higher than 10.

For the filtered variants, SNPEff and SNPSift are used to annotate them based on the various databases such as dbSNP, 1000 Genome Project database, ESP6500, SIFT database, and CLINVAR.

LINK https://www.broadinstitute.org/gatk/guide/best-practices?bpm=RNAseq

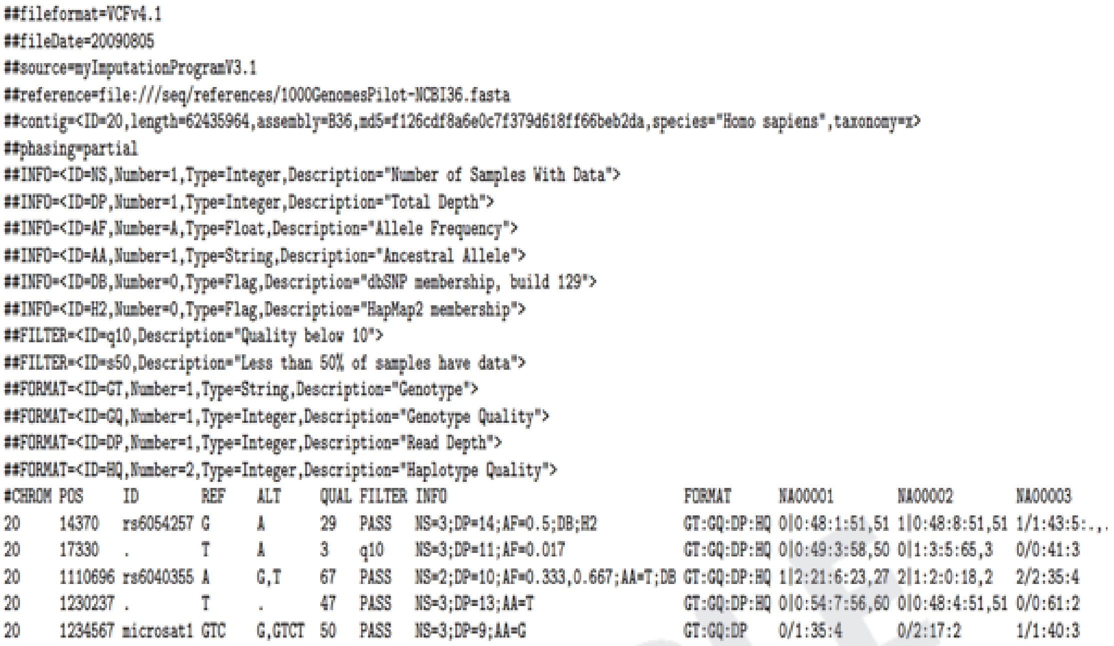
Below summarizes the results for 6 samples' SNV analysis.

Table 11. Summary of SNV Frequencies

Sample_ID	Number of SNPs	Number of coding SNPs	Number of synonymous SNPs	Number of nonsynonymous SNPs	Number of indels	Number of coding indels	Ratio of hom variants
MG_CTRL_1	40,787	36,881	5,860	4,310	8,444	7,768	22.50%
MG_CTRL_2	42,353	38,130	5,970	4,342	8,571	7,885	22.56%
MG_CTRL_3	43,815	39,345	6,025	4,419	8,818	8,086	22.51%
MG_TEST_1	48,222	42,534	6,309	4,623	9,193	8,289	23.20%
MG_TEST_2	42,603	38,081	6,070	4,460	8,041	7,348	22.93%
MG_TEST_3	45,498	40,366	6,429	4,761	8,011	7,284	23.36%

Individual SNV results are provided as vcf file and excel file. An example of vcf file is shown below.

LINK http://www.1000genomes.org/node/101



```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

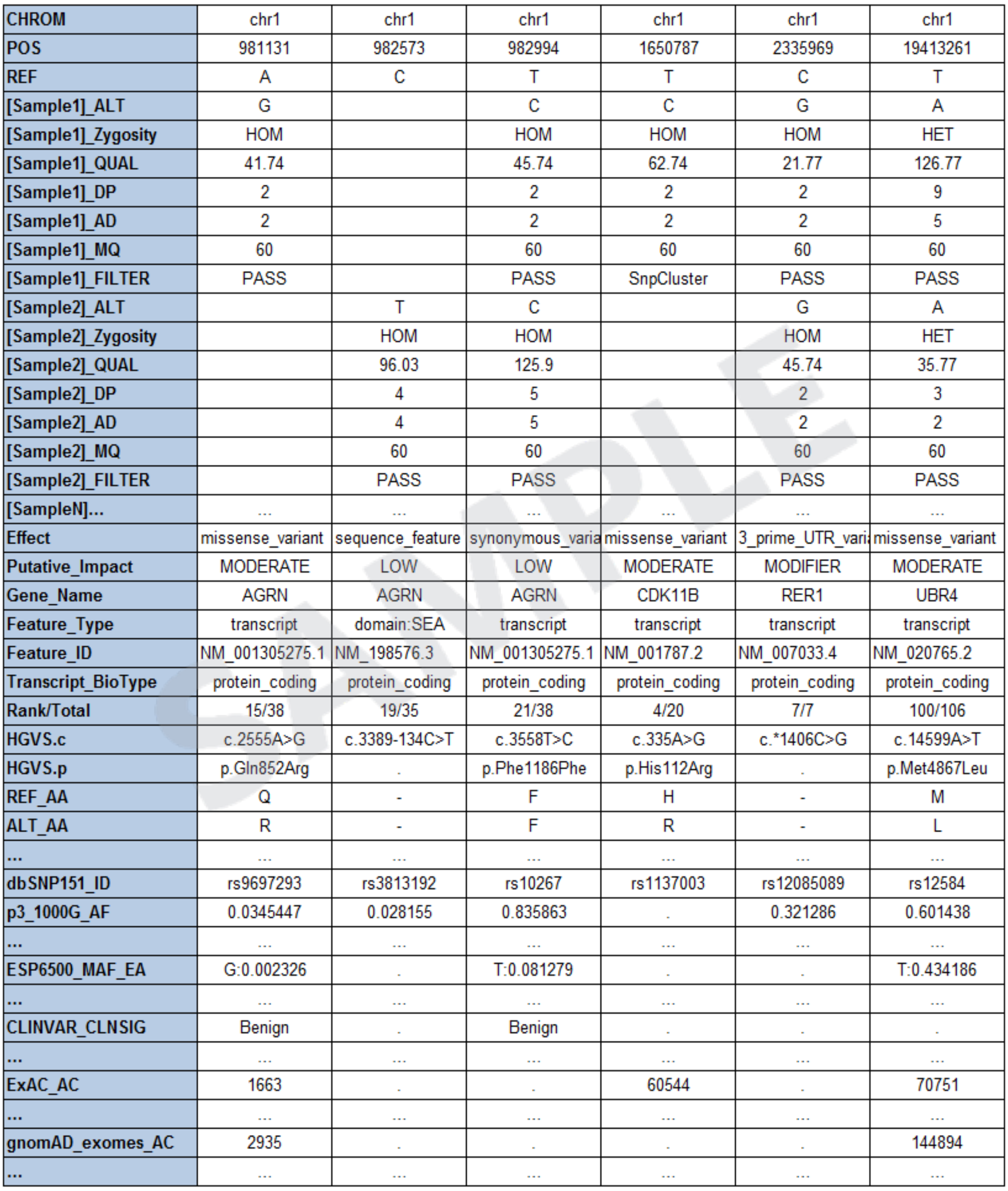
- CHROM: Chromosome name
- POS: Reference position (1-based coordinate)
- ID: Identifier (if it is a variant that exist in dbSNP, shown as rs#)
- REF: Reference Sequence regarding the position of interest
- ALT: Non-reference sequence
- QUAL: Phred scaled quality score. High QUAL score of SNP quality means credible call
- FILTER: 'PASS' if call at a specific position satisfies filter condition (Fisher Strand values, FS>30.0 and Quality By Depth values, QD <2.0).
- INFO: additional position information can be provided with semicolon (depending on the vcf production)
  - NS: Number of Sample with Data
  - DP: Total depth
  - AF: Allele Frequency
  - AA: Ancestral Allele
  - DB: Found in dbSNP or not
  - H2: Found in HapMap2 of not
- FORMAT: The data format is expressed in sample column in the order of GT(Genotype):GQ(Genotype Quality):DP(Read Depth):HQ(Haplotype Quality).
- Sample Name: Sample's genotype information is shows in FORMAT column in corresponding order.

You can find the discovered SNV results in Excel file format containing variant annotation information such as dbSNP, 1000 Genome Project database, ESP6500, SIFT database, CLINVAR etc.

The example is a table that summarizes the overall data. For more information, please refer to the PDF file linked below.

LINK AnnotDescription.pdf

Table 12. An example of annotation of individually discovered SNV



CHROM	chr1	chr1	chr1	chr1	chr1	chr1
POS	981131	982573	982994	1650787	2335969	19413261
REF	A	C	T	T	C	T
[Sample1]_ALT	G		C	C	G	A
[Sample1]_Zygosity	HOM		HOM	HOM	HOM	HET
[Sample1]_QUAL	41.74		45.74	62.74	21.77	126.77
[Sample1]_DP	2		2	2	2	9
[Sample1]_AD	2		2	2	2	5
[Sample1]_MQ	60		60	60	60	60
[Sample1]_FILTER	PASS		PASS	SnpCluster	PASS	PASS
[Sample2]_ALT		T	C		G	A
[Sample2]_Zygosity		HOM	HOM		HOM	HET
[Sample2]_QUAL		96.03	125.9		45.74	35.77
[Sample2]_DP		4	5		2	3
[Sample2]_AD		4	5		2	2
[Sample2]_MQ		60	60		60	60
[Sample2]_FILTER		PASS	PASS		PASS	PASS
[SampleN]...	...	...	...	...	...	...
Effect	missense_variant	sequence_feature	synonymous_varia	missense_variant	3_prime_UTR_vari	missense_variant
Putative_Impact	MODERATE	LOW	LOW	MODERATE	MODIFIER	MODERATE
Gene_Name	AGRN	AGRN	AGRN	CDK11B	RER1	UBR4
Feature_Type	transcript	domain:SEA	transcript	transcript	transcript	transcript
Feature_ID	NM_001305275.1	NM_198576.3	NM_001305275.1	NM_001787.2	NM_007033.4	NM_020765.2
Transcript_BioType	protein_coding	protein_coding	protein_coding	protein_coding	protein_coding	protein_coding
Rank/Total	15/38	19/35	21/38	4/20	7/7	100/106
HGVS.c	c.2555A>G	c.3389-134C>T	c.3558T>C	c.335A>G	c.*1406C>G	c.14599A>T
HGVS.p	p.Gln852Arg	.	p.Phe1186Phe	p.His112Arg	.	p.Met4867Leu
REF_AA	Q	-	F	H	-	M
ALT_AA	R	-	F	R	-	L
...	...	...	...	...	...	...
dbSNP151_ID	rs9697293	rs3813192	rs10267	rs1137003	rs12085089	rs12584
p3_1000G_AF	0.0345447	0.028155	0.835863	.	0.321286	0.601438
...	...	...	...	...	...	...
ESP6500_MAF_EA	G:0.002326	.	T:0.081279	.	.	T:0.434186
...	...	...	...	...	...	...
CLINVAR_CLNSIG	Benign	.	Benign	.	.	.
...	...	...	...	...	...	...
ExAC_AC	1663	.	.	60544	.	70751
...	...	...	...	...	...	...
gnomAD_exomes_AC	2935	.	.	.	.	144894
...	...	...	...	...	...	...

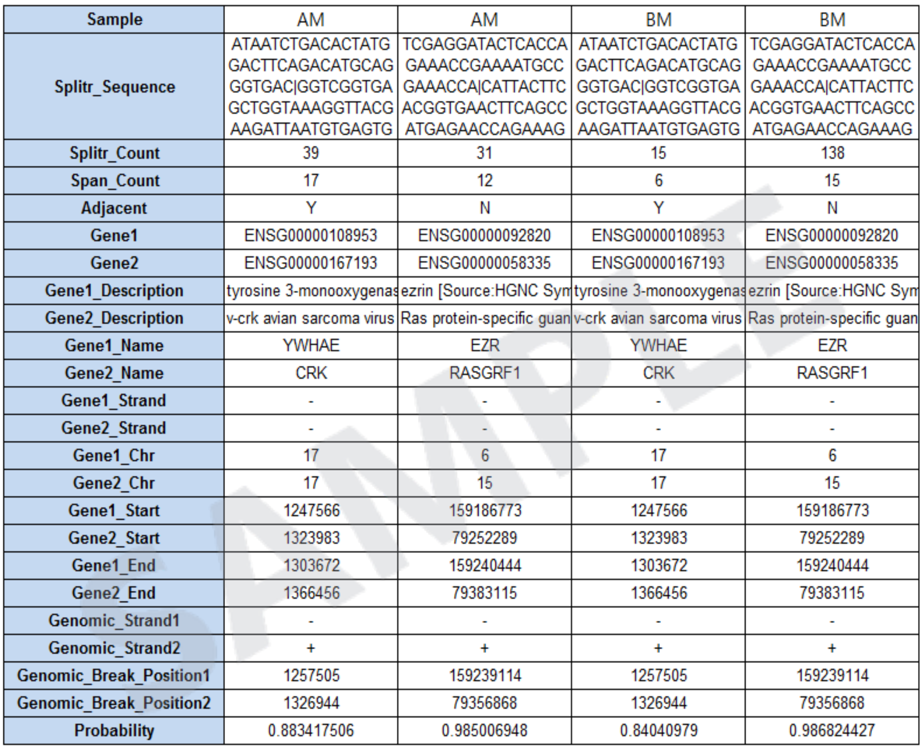
7. Fusion Gene Prediction Results

7. 1. Defuse Analysis Result

(Refer to Path: result_RNAseq/Fusion_gene_analysis/DEFUSE/)

Fusion genes were predicted with Defuse program. Defuse predicts fusion genes region by clustering discordant paired-end alignments (both spanning and split reads) and determines the probability of real fusion gene with heuristic filter.

Table 13. Example of Fusion Gene Prediction Results



Sample	AM	AM	BM	BM
Splitr_Sequence	ATAATCTGACACTATG GACTTCAGACATGCAG GGTGAC|GGTCGGTGA GCTGGTAAAGGTTACG AAGATTAATGTGAGTG	TCGAGGATACTCACCA GAAACCGAAAATGCC GAAACCA|CATTACTTC ACGGTGAACTTCAGCC ATGAGAACCAGAAAG	ATAATCTGACACTATG GACTTCAGACATGCAG GGTGAC|GGTCGGTGA GCTGGTAAAGGTTACG AAGATTAATGTGAGTG	TCGAGGATACTCACCA GAAACCGAAAATGCC GAAACCA|CATTACTTC ACGGTGAACTTCAGCC ATGAGAACCAGAAAG
Splitr_Count	39	31	15	138
Span_Count	17	12	6	15
Adjacent	Y	N	Y	N
Gene1	ENSG00000108953	ENSG00000092820	ENSG00000108953	ENSG0000092820
Gene2	ENSG00000167193	ENSG00000058335	ENSG00000167193	ENSG0000058335
Gene1_Description	tyrosine 3-monooxygenas	ezrin [Source:HGNC Sym	tyrosine 3-monooxygenas	ezrin [Source:HGNC Sym
Gene2_Description	v-crk avian sarcoma virus	Ras protein-specific guan	v-crk avian sarcoma virus	Ras protein-specific guan
Gene1_Name	YWHAE	EZR	YWHAE	EZR
Gene2_Name	CRK	RASGRF1	CRK	RASGRF1
Gene1_Strand	-	-	-	-
Gene2_Strand	-	-	-	-
Gene1_Chr	17	6	17	6
Gene2_Chr	17	15	17	15
Gene1_Start	1247566	159186773	1247566	159186773
Gene2_Start	1323983	79252289	1323983	79252289
Gene1_End	1303672	159240444	1303672	159240444
Gene2_End	1366456	79383115	1366456	79383115
Genomic_Strand1	-	-	-	-
Genomic_Strand2	+	+	+	+
Genomic_Break_Position1	1257505	159239114	1257505	159239114
Genomic_Break_Position2	1326944	79356868	1326944	79356868
Probability	0.883417506	0.985006948	0.84040979	0.986824427

- Sample: Sample name
- Split_Sequence: Shows fusion sequences. The two sequences of the donor and acceptor are separated by "|".
- Split_Count: Number of reads that align to the one end and does not align on the other end.
- Span_Count: Number of paired-ends reads that align at different genes
- Gene1, Gene2: Ensembl ID of gene1 and gene2
- Gene1_Name, Gene2_Name: Name of the gene1 and gene2
- Gene1_Description, Gene2_Description: Gene description
- Gene1_Strand, Gene2_Strand: Gene strand
- Gene1_Chr, Gene2_Chr: Chromosome
- Gene1_Start, Gene2_Start, Gene1_End, Gene2_End: Start, end position of two genes
- Genomic_Strand1, Genomic_Stand2: Genomic strand of each fusion splice/breakpoint
- Genomic_Break_Position1, Genomic_Break_Position2: Genomic position of of each gene's

fusion splice/breakpoint

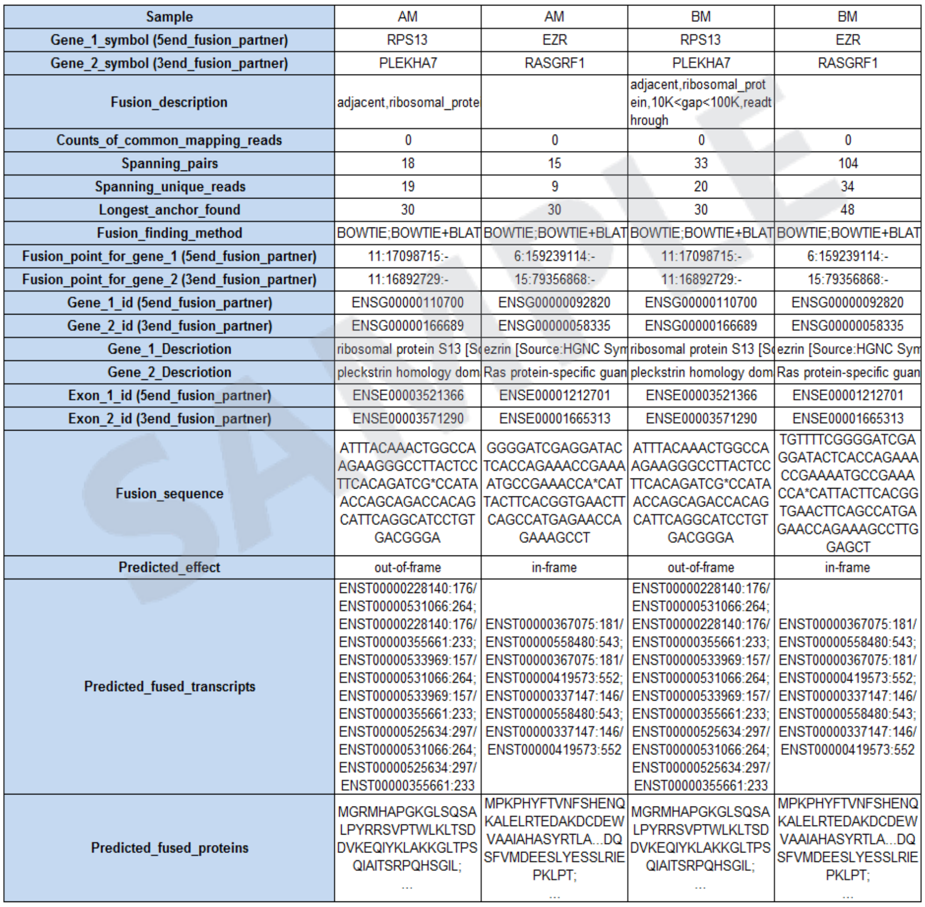
- Probability: Probability of sorted as fusion gene. Higher value means higher probability of being a fusion gene.

7. 2. FusionCatcher Analysis Result

(Refer to Path: result_RNAseq/Fusion_gene_analysis/FusionCatcher/)

Fusion genes were predicted with FusionCatcher program. FusionCatcher searches for novel/known somatic fusion genes, translocations, and chimeras in RNA-seq data. FusionCatcher is doing its own quality filtering/trimming of reads. This is needed because most a very important factor for finding fusion genes in RNA-seq experiment is the length of RNA fragments. Ideally the RNA fragment size for finding fusion genes should be over 300 bp. FusionCatcher is able to finding fusion genes even in cases where the fusion junction is within known exon or within known intron. The minimum condition for FusionCatcher to find a fusion gene is that both genes involved in the fusion are annotated in Ensembl database.

Table 14. Example of Fusion Gene Prediction Results



Sample	AM	AM	BM	BM
Gene_1_symbol (5end_fusion_partner)	RPS13	EZR	RPS13	EZR
Gene_2_symbol (3end_fusion_partner)	PLEKHA7	RASGRF1	PLEKHA7	RASGRF1
Fusion_description	adjacent,ribosomal_prote		adjacent,ribosomal_protein,10K<gap<100K,readthrough	
Counts_of_common_mapping_reads	0	0	0	0
Spanning_pairs	18	15	33	104
Spanning_unique_reads	19	9	20	34
Longest_anchor_found	30	30	30	48
Fusion_finding_method	BOWTIE;BOWTIE+BLAT	BOWTIE;BOWTIE+BLAT	BOWTIE;BOWTIE+BLAT	BOWTIE;BOWTIE+BLAT
Fusion_point_for_gene_1 (5end_fusion_partner)	11:17098715:-	6:159239114:-	11:17098715:-	6:159239114:-
Fusion_point_for_gene_2 (3end_fusion_partner)	11:16892729:-	15:79356868:-	11:16892729:-	15:79356868:-
Gene_1_id (5end_fusion_partner)	ENSG00000110700	ENSG00000092820	ENSG00000110700	ENSG00000092820
Gene_2_id (3end_fusion_partner)	ENSG00000166689	ENSG00000058335	ENSG00000166689	ENSG00000058335
Gene_1_Descriotion	ribosomal protein S13 [S	ezrin [Source:HGNC Sym	ribosomal protein S13 [S	ezrin [Source:HGNC Sym
Gene_2_Descriotion	pleckstrin homology dom	Ras protein-specific guan	pleckstrin homology dom	Ras protein-specific guan
Exon_1_id (5end_fusion_partner)	ENSE00003521366	ENSE00001212701	ENSE00003521366	ENSE00001212701
Exon_2_id (3end_fusion_partner)	ENSE00003571290	ENSE00001665313	ENSE00003571290	ENSE00001665313
Fusion_sequence	ATTTACAAACTGGCCAAGAAGGGCCTTACTCCTTCACAGATCG*CCATAACCAGCAGACCACAGCATTCAGGCATCCTGTGACGGGA	GGGGATCGAGGATACTCACCAGAAACCGAAAATGCCGAAACCA*CATTACTTCACGGTGAACTTCAGCCATGAGAACCAGAAAGCCT	ATTTACAAACTGGCCAAGAAGGGCCTTACTCCTTCACAGATCG*CCATAACCAGCAGACCACAGCATTCAGGCATCCTGTGACGGGA	TGTTTTCGGGGATCGAGGATACTCACCAGAAACCGAAAATGCCGAAACCA*CATTACTTCACGGTGAACTTCAGCCATGAGAACCAGAAAGCCTTGGAGCT
Predicted_effect	out-of-frame	in-frame	out-of-frame	in-frame
Predicted_fused_transcripts	ENST00000228140:176/ENST00000531066:264;ENST00000228140:176/ENST00000355661:233;ENST00000533969:157/ENST00000531066:264;ENST00000533969:157/ENST00000355661:233;ENST00000525634:297/ENST00000531066:264;ENST00000525634:297/ENST00000355661:233	ENST00000367075:181/ENST00000558480:543;ENST00000367075:181/ENST00000419573:552;ENST00000337147:146/ENST00000558480:543;ENST00000337147:146/ENST00000419573:552	ENST00000228140:176/ENST00000531066:264;ENST00000228140:176/ENST00000355661:233;ENST00000533969:157/ENST00000531066:264;ENST00000533969:157/ENST00000355661:233;ENST00000525634:297/ENST00000531066:264;ENST00000525634:297/ENST00000355661:233	ENST00000367075:181/ENST00000558480:543;ENST00000367075:181/ENST00000419573:552;ENST00000337147:146/ENST00000558480:543;ENST00000337147:146/ENST00000419573:552
Predicted_fused_proteins	MGRMHAPGKGLSQSALPYRRSVPTWLKLTSDDVKEQIYKLAKKGLTPSQIAITSRPQHSGIL; ...	MPKPHYFTVNFSHENQKALELRTEDAKDCDEWVAAIAHASYRTLA...DQSFVMDEESLYESSLRIEPKLPT; ...	MGRMHAPGKGLSQSALPYRRSVPTWLKLTSDDVKEQIYKLAKKGLTPSQIAITSRPQHSGIL; ...	MPKPHYFTVNFSHENQKALELRTEDAKDCDEWVAAIAHASYRTLA...DQSFVMDEESLYESSLRIEPKLPT; ...

- Sample: Sample name
- Gene_1_symbol, Gene_2_symbol: Gene symbol of the 5' end and 3' end fusion partner
- Fusion_description: Type of the fusion gene
- Counts_of_common_mapping_reads: Count of reads mapping simultaneously on both genes which form the fusion gene
- Spanning_pairs: Count of pair-end reads supporting the fusion
- Spanning_unique_reads: Count of unique read mapping on the fusion junction

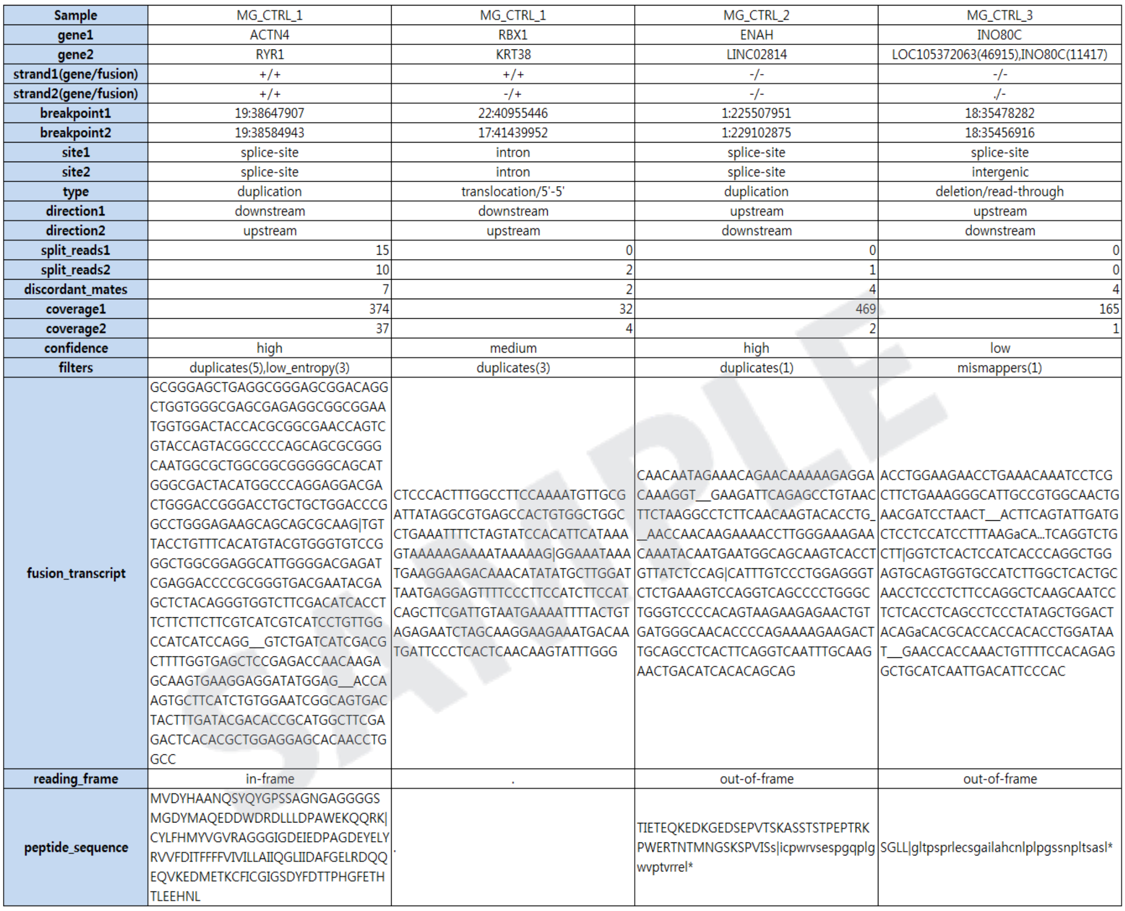
- Longest_anchor_found: Longest anchor (hangover) found among the unique reads mapping on the fusion junction
- Fusion_finding_method: Aligning method used for mapping the reads and finding the fusion genes.
- Fusion_point_for_gene_1, Fusion_point_for_gene_2: Chromosomal position of the 5' end and 3' end of fusion junction; 1-based coordinate
- Gene_1_id, Gene_2_id: Ensembl gene id of the 5' end and 3' end fusion partner
- Gene_1_Description, Gene_2_Description: Gene description of the 5' end and 3' end fusion partner
- Exon_1_id, Exon_2_id: Ensembl exon id of the 5' end and 3' end fusion exon-exon junction
- Fusion_sequence: The inferred fusion junction (the asterisk sign marks the junction point)
- Predicted_effect: Predicted effect of the candidate fusion gene using the annotation from Ensembl database
- Predicted_fused_transcripts: All possible known fused transcripts
- Predicted_fused_proteins: Predicted amino acid sequences of all possible fused proteins

7. 3. Arriba Analysis Result

(Refer to Path: result_RNAseq/Fusion_gene_analysis/Arriba)

Fusion genes were predicted by Arriba program. Based on the alignment result of STAR aligner, Arriba provides potential gene fusion candidates which pass all of its read-level filters and event-level filters.

Table 15. Example of Fusion Gene Prediction Results



Sample	MG_CTRL_1	MG_CTRL_1	MG_CTRL_2	MG_CTRL_3
gene1	ACTN4	RBX1	ENAH	INO80C
gene2	RYR1	KRT38	LINC02814	LOC105372063(46915),INO80C(11417)
strand1(gene/fusion)	+/+	+/+	-/-	-/-
strand2(gene/fusion)	+/+	-/+	-/-	./-
breakpoint1	19:38647907	22:40955446	1:225507951	18:35478282
breakpoint2	19:38584943	17:41439952	1:229102875	18:35456916
site1	splice-site	intron	splice-site	splice-site
site2	splice-site	intron	splice-site	intergenic
type	duplication	translocation/5'-5'	duplication	deletion/read-through
direction1	downstream	downstream	upstream	upstream
direction2	upstream	upstream	downstream	downstream
split_reads1	15	0	0	0
split_reads2	10	2	1	0
discordant_mates	7	2	4	4
coverage1	374	32	469	165
coverage2	37	4	2	1
confidence	high	medium	high	low
filters	duplicates(5),low_entropy(3)	duplicates(3)	duplicates(1)	mismappers(1)
fusion_transcript	GCGGGAGCTGAGGCGGGAGCGGACAGGCTGGTGGGCGAGCGAGAGGCGGCGGAATGGTGGACTACCACGCGGCGAACCAGTCGTACCAGTACGGCCCCAGCAGCGCGGGCAATGGCGCTGGCGGCGGGGGCAGCATGGGCGACTACATGGCCCAGGAGGACGACTGGGACCGGGACCTGCTGCTGGACCCGGCCTGGGAGAAGCAGCAGCGCAAG|TGTTACCTGTTTCACATGTACGTGGGTGTCCGGGCTGGCGGAGGCATTGGGGACGAGATCGAGGACCCCGCGGGTGACGAATACGAGCTCTACAGGGTGGTCTTCGACATCACCTTCTTCTTCTTCGTCATCGTCATCCTGTTGGCCATCATCCAGG__GTCTGATCATCGACGCTTTTGGTGAGCTCCGAGACCAACAAGAGCAAGTGAAGGAGGATATGGAG__ACCAAGTGCTTCATCTGTGGAATCGGCAGTGACTACTTTGATACGACACCGCATGGCTTCGAGACTCACACGCTGGAGGAGCACAACCTGGCC	CTCCCACTTTGGCCTTCCAAAATGTTGCGATTATAGGCGTGAGCCACTGTGGCTGGCCTGAAATTTTCTAGTATCCACATTCATAAAGTAAAAAGAAAATAAAAAG|GGAAATAAATGAAGGAAGACAAACATATATGCTTGGATTAATGAGGAGTTTTCCCTTCCATCTTCCATCAGCTTCGATTGTAATGAAAATTTTACTGTAGAGAATCTAGCAAGGAAGAAATGACAATGATTCCCTCACTCAACAAGTATTTGGG	CAACAATAGAAACAGAACAAAAAGAGGACAAAGGT__GAAGATTCAGAGCCTGTAACTTCTAAGGCCTCTTCAACAAGTACACCTG___AACCAACAAGAAAACCTTGGGAAAGAACAAATACAATGAATGGCAGCAAGTCACCTGTTATCTCCAG|CATTTGTCCCTGGAGGGTCTCTGAAAGTCCAGGTCAGCCCCTGGGCTGGGTCCCCACAGTAAGAAGAGAACTGTGATGGGCAACACCCCAGAAAAGAAGACTTGCAGCCTCACTTCAGGTCAATTTGCAAGAACTGACATCACACAGCAG	ACCTGGAAGAACCTGAAACAAATCCTCGCTTCTGAAAGGGCATTGCCGTGGCAACTGAACGATCCTAACT__ACTTCAGTATTGATGCTCCTCCATCCTTTAAGaCA...TCAGGTCTGCTT|GGTCTCACTCCATCACCCAGGCTGGAGTGCAGTGGTGCCATCTTGGCTCACTGCAACCTCCCTCTTCCAGGCTCAAGCAATCCTCTCACCTCAGCCTCCCTATAGCTGGACTACAGaCACGCACCACCACACCTGGATAAT__GAACCACCAAACTGTTTTCCACAGAGGCTGCATCAATTGACATTCCCAC
reading_frame	in-frame	.	out-of-frame	out-of-frame
peptide_sequence	MVDYHAANQSYQYGPSSAGNGAGGGGSMGDYMAQEDDWDRDLLLDPAWEKQQRK|CYLFHMYVGVRAGGGIGDEIEDPAGDEYELYRVVFDITFFFFVIVILLAIIQGLIIDAFGELRDQQEQVKEDMETKCFICGIGSDYFDTTPHGFETHTLEEHNL	.	TIETEQKEDKGEDSEPVTSKASSTSTPEPTRKPWERTNTMNGSKSPVISs|icpwrvsespgqplgwvptvrrel*	SGLL|gltpsprlecsgailahcnlplpgssnpltsasl*

- Sample ID: Sample name
- Gene1, Gene2: Gene symbols in gene1 and gene2 respectively. Gene1 contains the gene which makes up the 5' end of the transcript and gene2 the gene which makes up the 3' end.
  - If a breakpoint is in an intergenic region, Arriba lists the closest genes upstream and downstream from the breakpoint, separated by a comma. The numbers in parentheses after the closest genes indicate the distance to the genes.
  - For example, "ZNF23 (1396), ZNF19 (10425)" means that ZNF23 exists 1,396 bp upstream and ZNF19 exists 10,425 bp downstream from the breakpoint.
- Strand1, Strand2: Strand information in gene1 and gene2 respectively.
  - The strand before parenthesis indicates strand of the gene according to the gene annotation and the value after parenthesis indicates the strand that is transcribed, respectively.

- Breakpoint1, Breakpoint2: Coordinates of the breakpoints in gene1 and gene2, respectively.
- Site1, Site2: Information about the location of gene1/gene2 breakpoints (Possible values are splice-site, exon, intron, 5'UTR, 3'UTR, UTR and intergenic)
- Type: Based on the orientation of the supporting reads and the coordinates of breakpoints, the type of event can be inferred. Possible values are translocation, duplication, inversion and deletion.
- Direction1, Direction2: Orientation information of fusion partner based on gene1/gene2
  - Downstream: This means that the partner is fused downstream of the breakpoint, i.e. at a coordinate higher than the breakpoint.
  - Upstream: This means that the partner is fused upstream of the breakpoint, i.e. at a coordinate lower than the breakpoint.
- Split read1, Split read2: The number of supporting split fragments with an anchor in gene1 or gene2, respectively. The gene to which the longer segment of the split read aligns is defined as the anchor. i.e. the number of reads that mapped to both fusion partner genes. Reads are assigned to the gene containing the longer segment of them.
- Discordant mates: The number of pairs of discordant mates (spanning reads or bridge reads) supporting the fusion.
- coverage1, coverage2: The number of fragments (coverage) near breakpoint1 and breakpoint2 respectively.
- confidence: Each prediction is assigned one of the confidences low, medium, or high. Several characteristics are taken into account, including: the number of supporting reads(i.e. the number of reads describing fusion such as split_read1, split_reads2, discordant_mates), the balance of split reads and discordant mates, the distance between the breakpoints, the type of event, whether the breakpoints are intragenic or not, and whether there are other events which corroborate the prediction, e.g. multiple isoforms or balanced translocations.

  (Refer to interpretation-of-results)
- filters: The reasons why reads were excluded in the supporting reads.
  - The total number of supporting reads is calculated by summing up the reads given in the columns split_reads1, split_reads2, discordant_mates, and filters.
- fusion_transcript: Fusion transcript sequence. Breakpoint is marked by a pipe symbol (|).
  - Lowercase letters: SNPs or SNVs
  - Characters between "[", "]": Insertions
  - (-): Deleted bases
  - (___): Three underscores are introns
  - (...): Missing information due to insufficient coverage
  - (?): Ambiguous position, such as positions with diverse reference mismatches
  - reading_frame: Information about whether the gene at the 3'end of the fusion is fused "in-frame" or "out-of-frame".
- peptide_sequence: peptide sequence which is translated from the fusion transcript. Breakpoint is marked via a pipe symbol (|).

For each predicted fusion, the visualization for fusion location on chromosomal ideogram and annotation information is generated as below (Figure 17). It is provided in the results folder as a PDF file with one page for each predicted fusion.

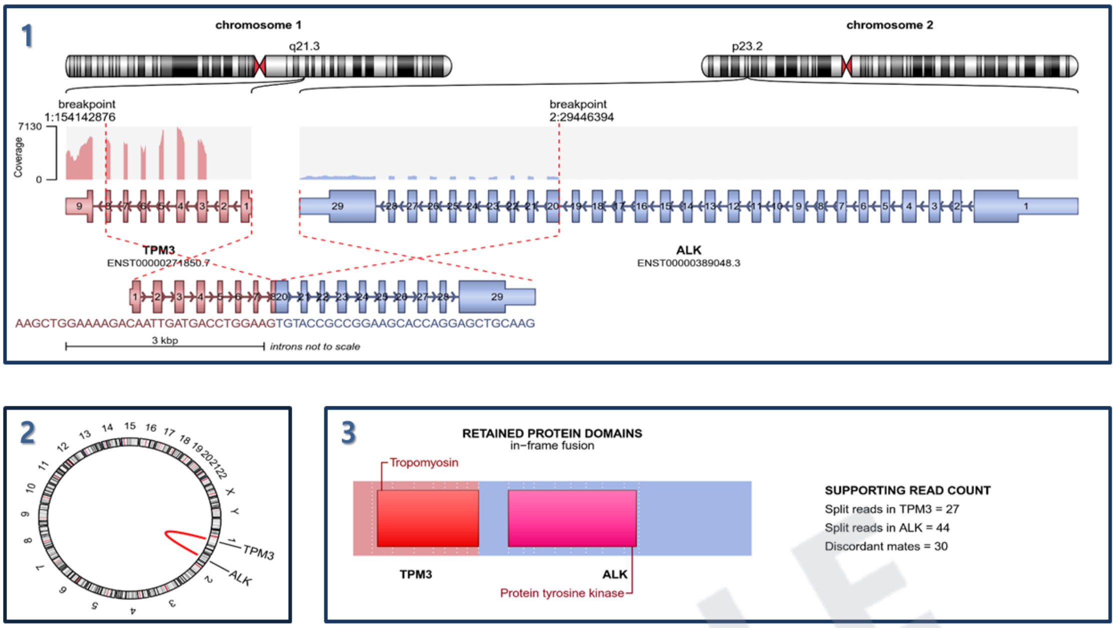


Figure 17. An example of a predicted fusion

1. The structure of fusion gene and its basic information (chromosome, transcript, coverage, sequence and break point).
2. CircosPlot of fusion gene containing its location on chromosome.
3. Retained protein domains and supporting read information associated with fusion gene.
   - If there is no associated protein domain, it is marked as blank.
   - Split reads in [geneA], Split reads in [geneB]: The number of split reads in each gene.
   - Discordant mates: The number of discordant mate reads (spanning reads or bridge reads)

8. Data Download Information

8. 1. Raw Data

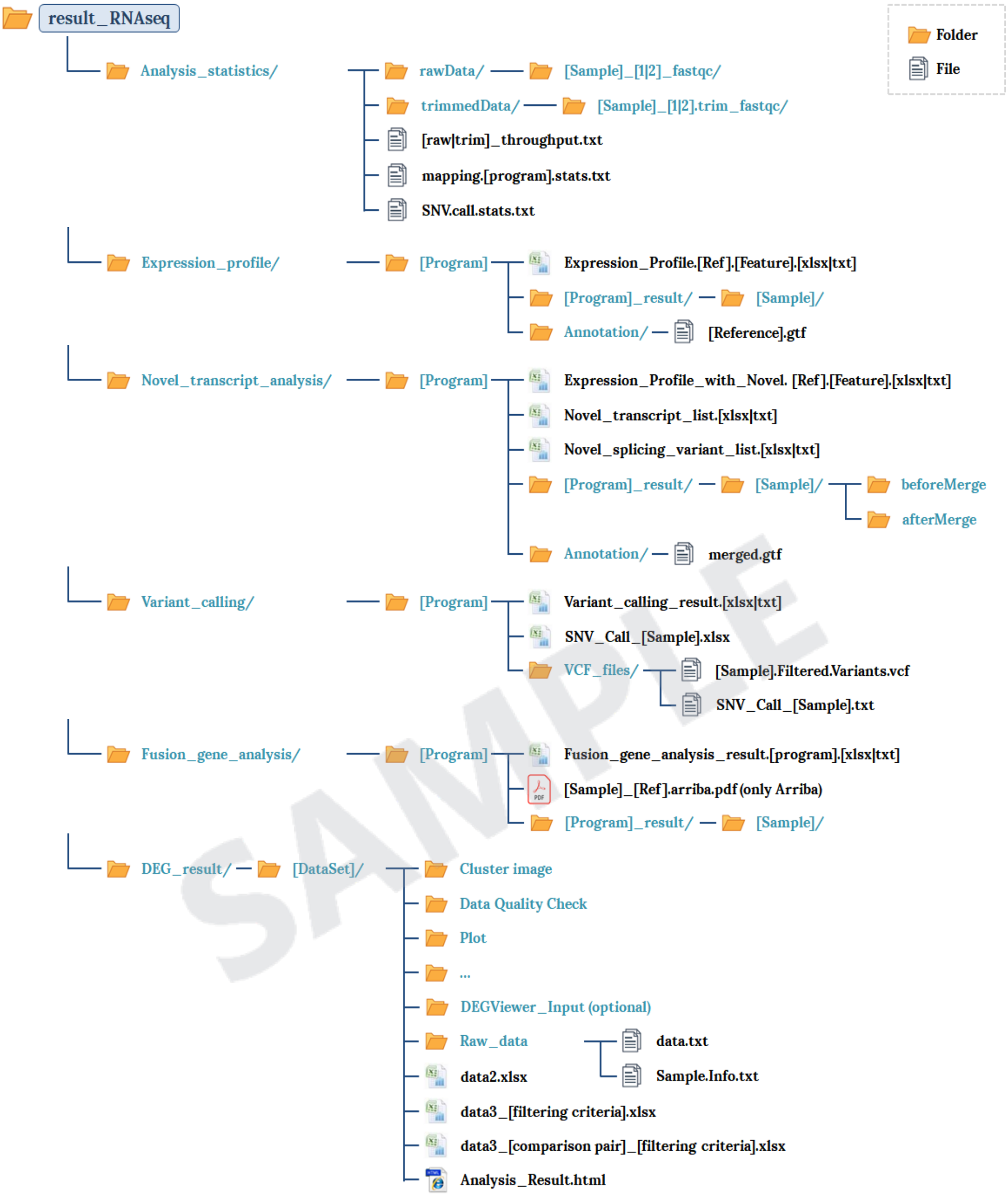
Raw data is the FASTQ file that isn't trimmed adapter sequence.

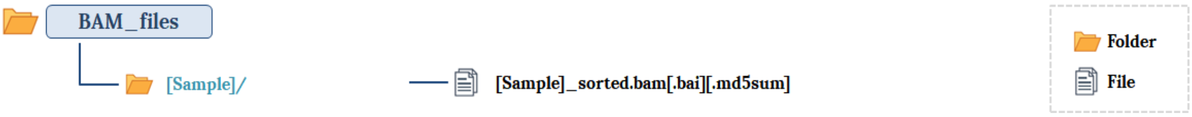
Download link	File size	md5sum
MG_CTRL_1_1.fastq.gz	8.3G	18cffa866442fe323d5612ce341f3d5c
MG_CTRL_1_2.fastq.gz	8.3G	1e82982a2fed892f4b27601d5100db1c
MG_CTRL_2_1.fastq.gz	8.02G	91637e3fbb20f714bea9323591b2ddcb
MG_CTRL_2_2.fastq.gz	8.02G	72233a9ec85c1a768eeac4dc63f719c8
MG_CTRL_3_1.fastq.gz	9.21G	306c928c518538973df1a463a293f1ce
MG_CTRL_3_2.fastq.gz	9.2G	dd1ae2c969fc66b0e111bf92dc9ce179
MG_TEST_1_1.fastq.gz	9.81G	24287dabbaa7c183200348debd493955
MG_TEST_1_2.fastq.gz	9.79G	96fe6a1645ff682b5297375f607f091a
MG_TEST_2_1.fastq.gz	7.79G	55066b44058fdb78c3aa175d371c9a80
MG_TEST_2_2.fastq.gz	7.79G	b65ef88c8e70c67a06f1328fdaf6031a
MG_TEST_3_1.fastq.gz	8.86G	977614bd41252cf399d5f56ab5af41db
MG_TEST_3_2.fastq.gz	8.86G	898ba04594825bc2800348b5bebb6cc5

- 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.

8. 2. Analysis Results

Download link	File size
HN00000000_result_RNAseq.zip (md5sum: a5b0c9ad4cc93c57447b927659e80f4f)	1.08G
HN00000000_BAM_files.tar (md5sum: a407fe964cb2b978a15c89c7559fdb4f)	44.55G


result_RNAseq
Folder
File
Analysis_statistics/
rawData/
[Sample]_[1|2]_fastqc/
trimmedData/
[Sample]_[1|2].trim_fastqc/
[raw|trim]_throughput.txt
mapping.[program].stats.txt
SNV.call.stats.txt
Expression_profile/
[Program]
Expression_Profile.[Ref].[Feature].[xlsx|txt]
[Program]_result/
[Sample]/
Annotation/
[Reference].gtf
Novel_transcript_analysis/
[Program]
Expression_Profile_with_Novel. [Ref].[Feature].[xlsx|txt]
Novel_transcript_list.[xlsx|txt]
Novel_splicing_variant_list.[xlsx|txt]
[Program]_result/
[Sample]/
beforeMerge
afterMerge
Annotation/
merged.gtf
Variant_calling/
[Program]
Variant_calling_result.[xlsx|txt]
SNV_Call_[Sample].xlsx
VCF_files/
[Sample].Filtered.Variants.vcf
SNV_Call_[Sample].txt
Fusion_gene_analysis/
[Program]
Fusion_gene_analysis_result.[program].[xlsx|txt]
[Sample]_[Ref].arriba.pdf (only Arriba)
[Program]_result/
[Sample]/
DEG_result/
[DataSet]/
Cluster image
Data Quality Check
Plot
...
DEGViewer_Input (optional)
Raw_data
data.txt
Sample.Info.txt
data2.xlsx
data3_[filtering criteria].xlsx
data3_[comparison pair]_[filtering criteria].xlsx
Analysis_Result.html



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

9. Appendix

9. 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.

9. 2. Programs used in Analysis

9. 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.

9. 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.

9. 2. 3. HISAT2 version 2.1.0

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

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

9. 2. 4. STAR 2.6.0c

LINK http://code.google.com/p/rna-star/

Spliced Transcripts Alignment to a Reference (STAR) software based on RNA-seq alignment algorithm which utilizes sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure.

9. 2. 5. StringTie version 2.1.3b

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

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

9. 2. 6. GATK version v4.2.0.0

LINK https://software.broadinstitute.org/gatk/
LINK https://www.broadinstitute.org/gatk/guide/best-practices?bpm=RNAseq

The GATK is the industry standard for identifying SNPs and indels in germline DNA and RNAseq data. The GATK Best Practices provide step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing (HTS) data. This analysis using STAR 2-pass mapping, Picard MarkDuplicate, Split 'N' Trim, Realignment, Base recalibration. Variant calling is performed on these reads using GATK haplotype caller.

9. 2. 7. SnpEff version 4.3t

LINK http://snpeff.sourceforge.net/SnpEff.html
LINK AnnotDescription.pdf

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes).

SnpEff can generate the following results :

- Genes and transcripts affected by the variant
- Location of the variants
- How the variant affects the protein synthesis (e.g. generating a stop codon)
- Comparison with other databases to find equal known variants

9. 2. 8. Defuse version 0.8.1

LINK https://bitbucket.org/dranew/defuse
LINK http://compbio.bccrc.ca/software/defuse/

Defuse is a discovers fusion genes from the RNA-Seq data. It clusters discordant paired-end alignments (spanning reads and split reads) to predict the correlation between fragment's length distribution and split reads and its arrangement lengths. Heuristic filter is applied to analyze the correlation and predict the existence of fusion genes.

9. 2. 9. FusionCatcher version 1.00

LINK https://github.com/ndaniel/fusioncatcher

FusionCatcher searches for novel/known somatic fusion genes, translocations, and chimeras in RNA-seq data. FusionCatcher is doing its own quality filtering/trimming of reads. This is needed because most a very important factor for finding fusion genes in RNA-seq experiment is the length of RNA fragments. Ideally the RNA fragment size for finding fusion genes should be over 300 bp. FusionCatcher is able to finding fusion genes even in cases where the fusion junction is within known exon or within known intron. The minimum condition for FusionCatcher to find a fusion gene is that both genes involved in the fusion are annotated in Ensembl database.

9. 2. 10. Arriba version 1.2.0

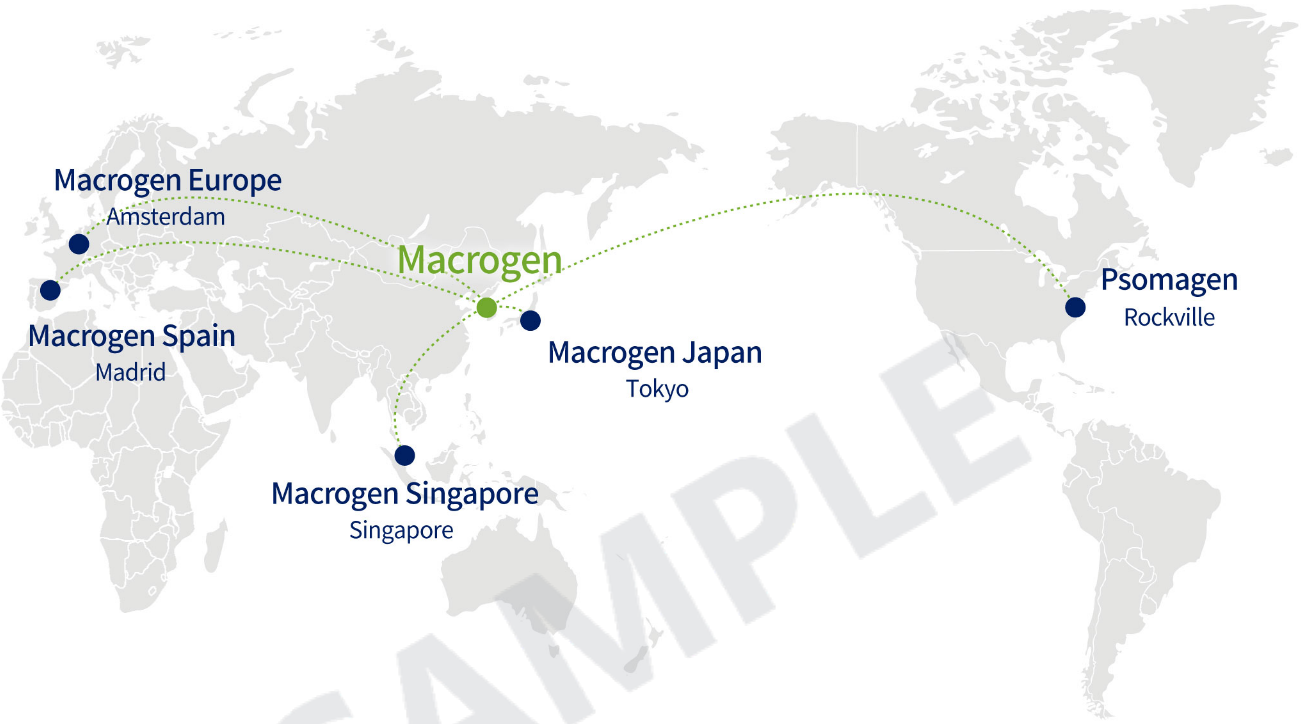
LINK https://arriba.readthedocs.io/en/latest/

Arriba is a command-line tool for the detection of gene fusions from RNA-Seq data. It was developed for the use in a clinical research setting. Therefore, short runtimes and high sensitivity were important design criteria. It is based on the ultrafast STAR aligner and the post-alignment runtime is typically just about 2 minutes. In contrast to many other fusion detection tools which build on STAR, Arriba does not require to reduce the alignIntronMax parameter of STAR to detect fusions arising from focal deletions.

9. 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.
6. AUWERA, Geraldine A., et al. From FastQ Data to HighConfidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics, 2013, 11.10.1-11.10. 33.
7. DEPRISTO, Mark A., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 2011, 43.5: 491-498.
8. MCKENNA, Aaron, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 2010, 20.9: 1297-1303.
9. CINGOLANI, Pablo, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 2012, 6.2: 80-92.
10. MCPHERSON, Andrew, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS computational biology, 2011, 7.5: e1001138.
11. NICORICI, Daniel, et al. FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv, 2014, 011650.
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

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 Singapore

**Laboratory,
Business & Support Center**

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