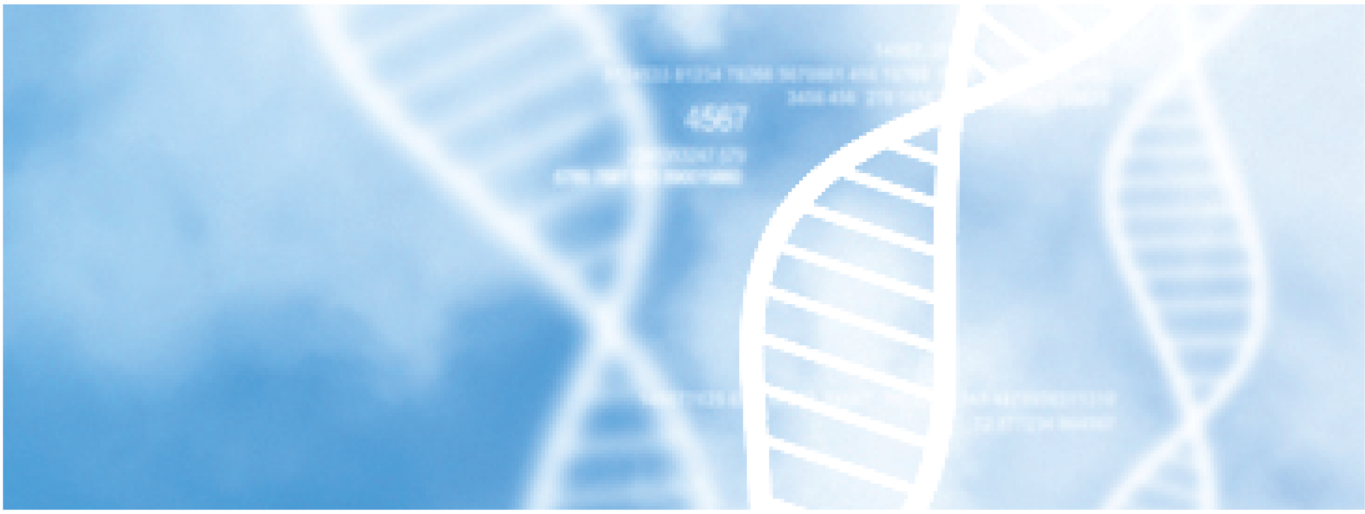
Humanizing Genomics
macrogen



Mus musculus
Transcriptome Sequencing Report

April 2015

Research use only

Project Background Information

Customer	
Company/Institution	
Order Number	Mouse
Species	*Mus musculus*
Reference	
Sample Type	
Library Type	
Types of Read	
Read Length	
Number of Samples	
Type of Analysis	

Project Results Summary

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

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

Analyses were successfully performed on all 12 paired-end samples as requested. Figure 1 below shows the amount throughput between raw data and trimmed data. Figure 2 shows the % Q30 score (% of bases with quality over phred score 30) per sample between raw and trimmed data.

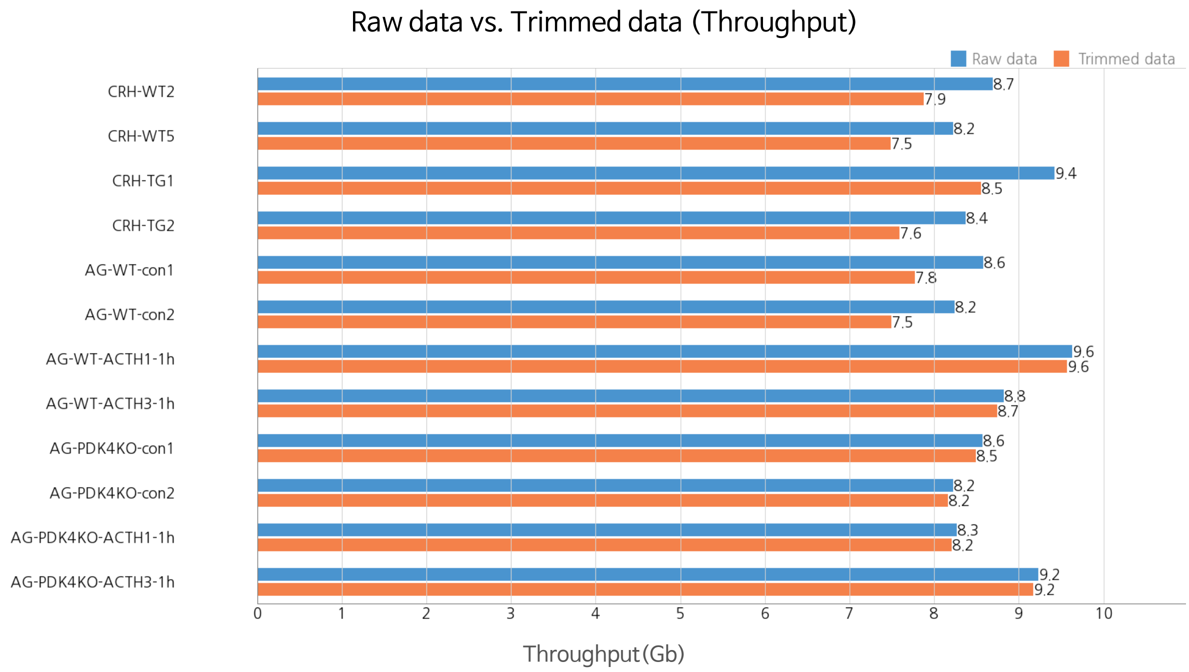


Figure 1. Throughput output between Raw and Trimmed data

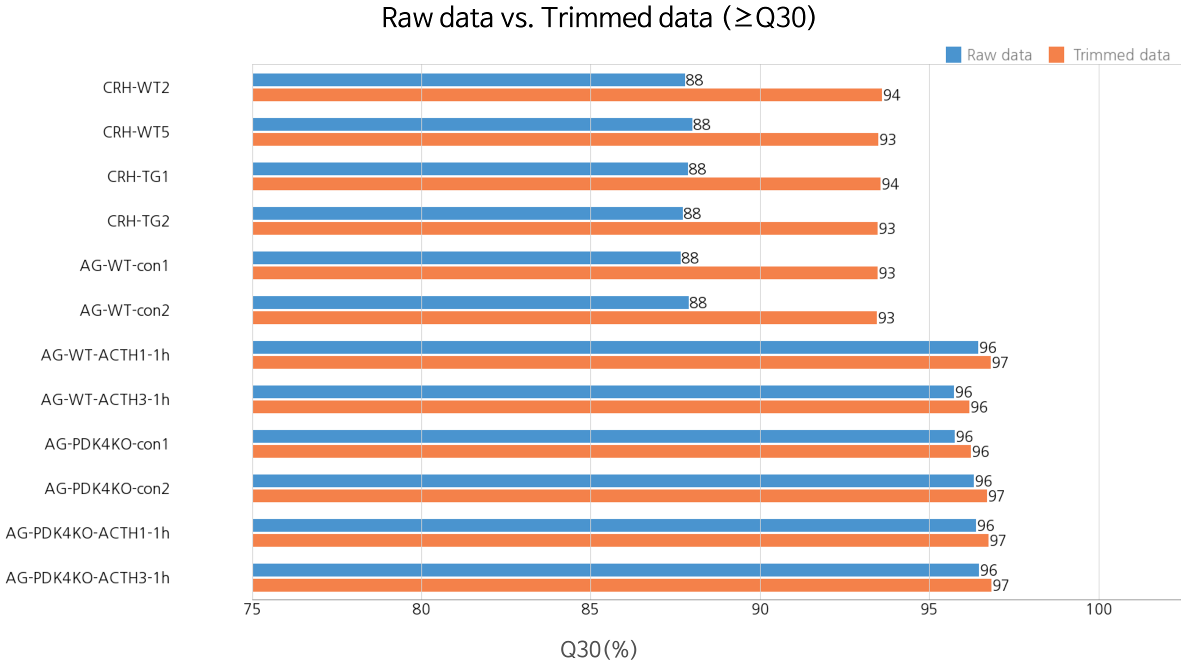


Figure 2. Q30 score between Raw and Trimmed data

TopHat was used to map trimmed reads with the reference genome. Figure 3 shows the overall read mapping ratio between trimmed read with the reference genome per sample.

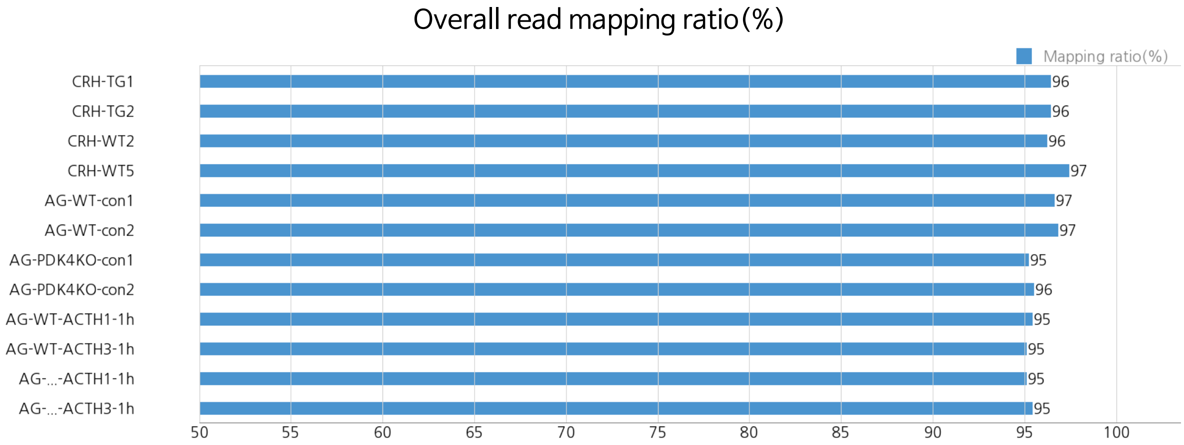


Figure 3. Overall read mapping ratio(%)

After the read mapping process, cufflink was used for transcript assembly process. Using these assembled transcripts, each expression profile was analyzed per sample, per transcript, and per FPKM (Fragment per Kilobase of transcript per Million mapped reads).

These values were used for comparison as 5 requested and were used for DEG (differentially Expressed Genes) analysis. The results showed total of 1,555 transcripts which satisfied |fc|≥2 & LPE test raw p-value〈0.05 conditions in at least one comparison.

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 per sample and per gene.

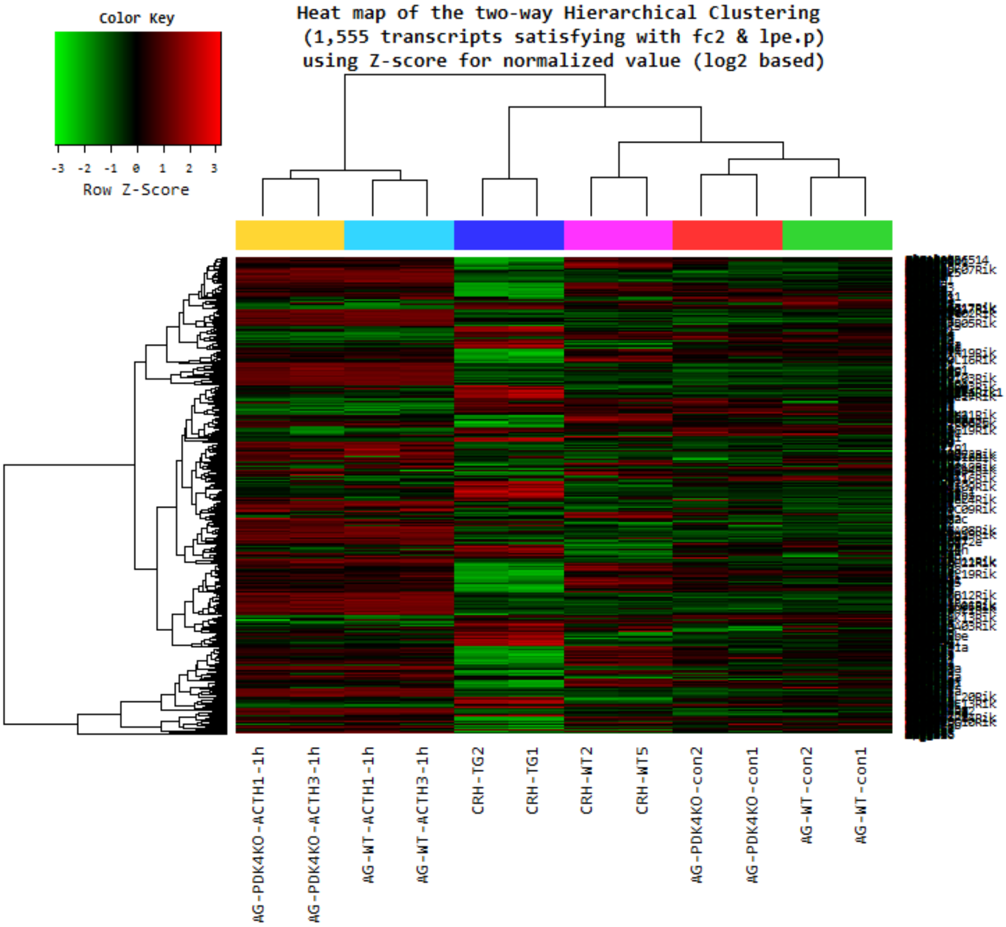


Figure4. Heatmap for DEG list

DEG list was further analyzed by DAVID tool(http://david.abcc.ncifcrf.gov/) for gene set enrichment analysis per biological process (BP), cellular component (CC), molecular function (MF). The Figure 5, 6 and 7 below show the gene set by each category.

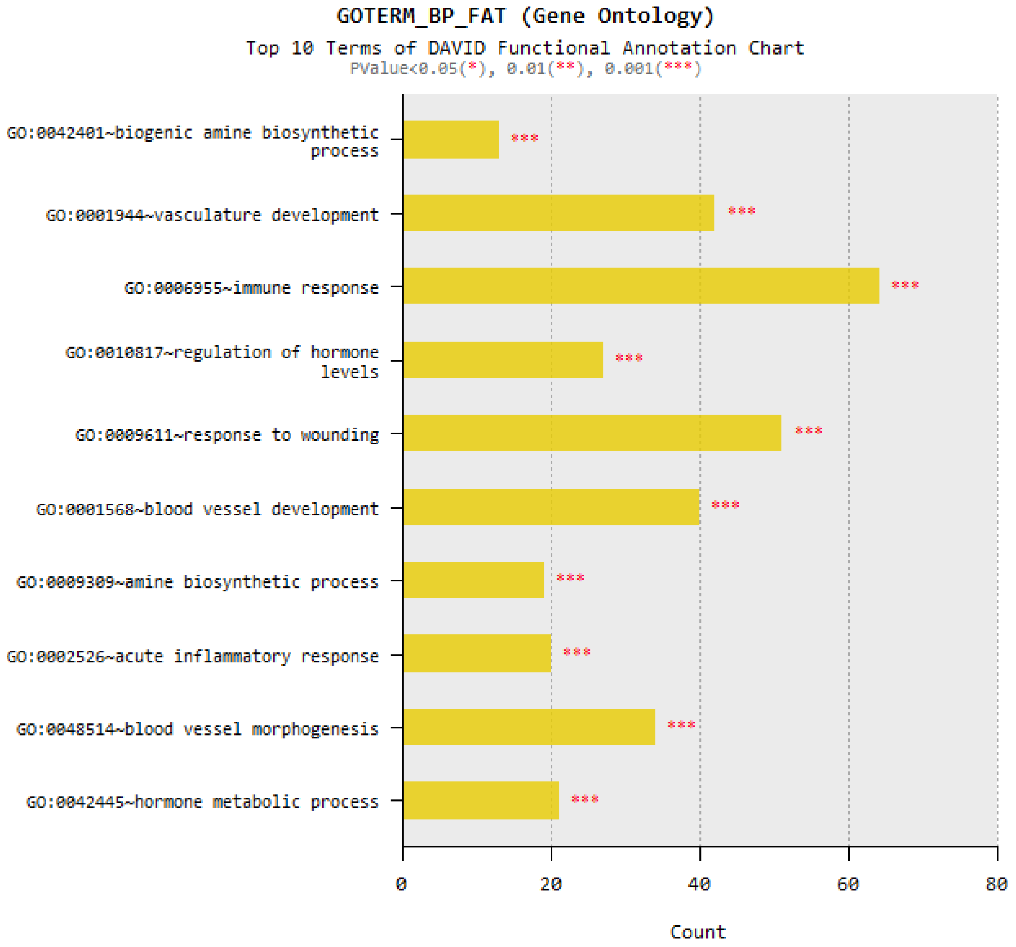


Figure 5. Gene Ontology terms related to Biological Process

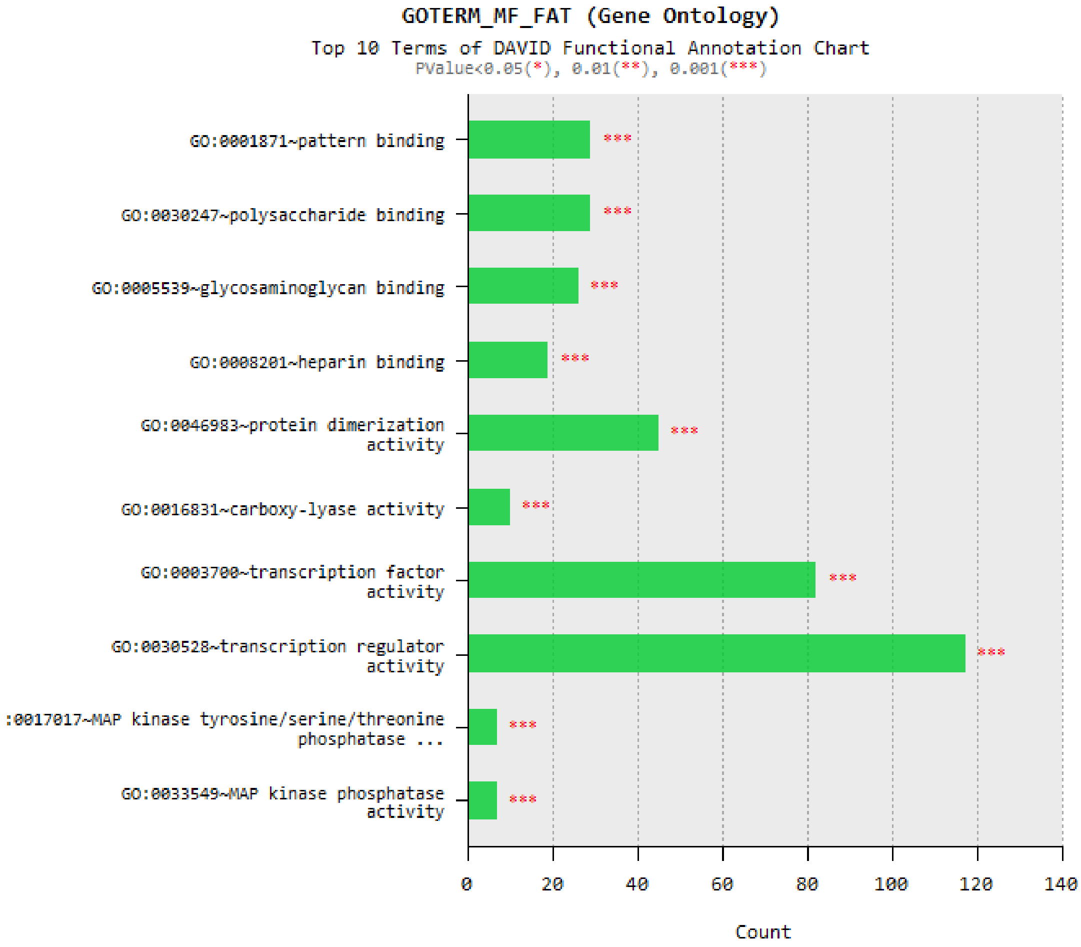


Figure 6. Gene Ontology Terms related to Molecular Function

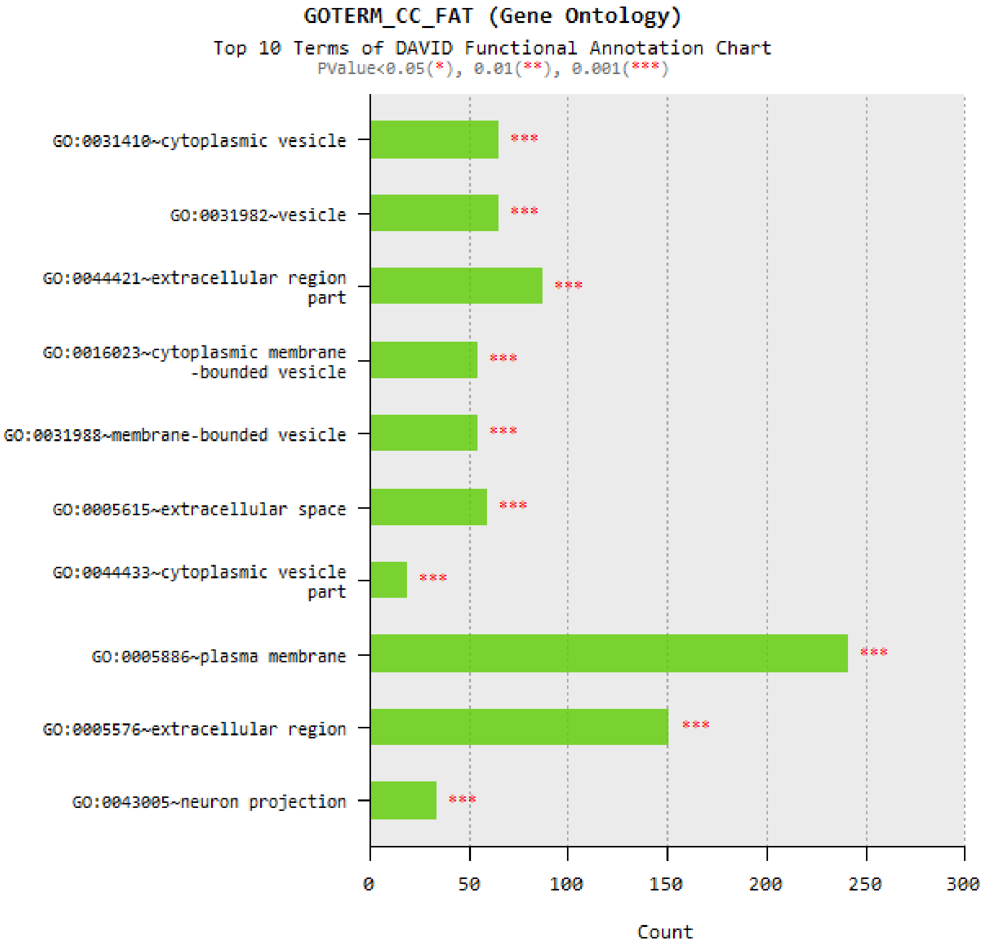
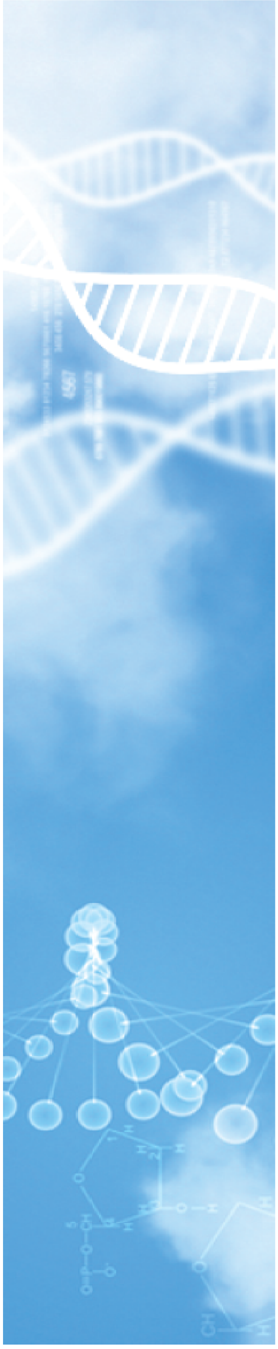


Figure 7. Gene Ontology Terms related to Cellular Component

In addition, novel transcript and novel alternative splicing transcripts were found per sample and SNV calling, variant annotation and fusion gene detection through defuse results were summarized (please refer to the main body of this report for detailed explanations).

Table of Contents



1. Experimental Methods and Workflow

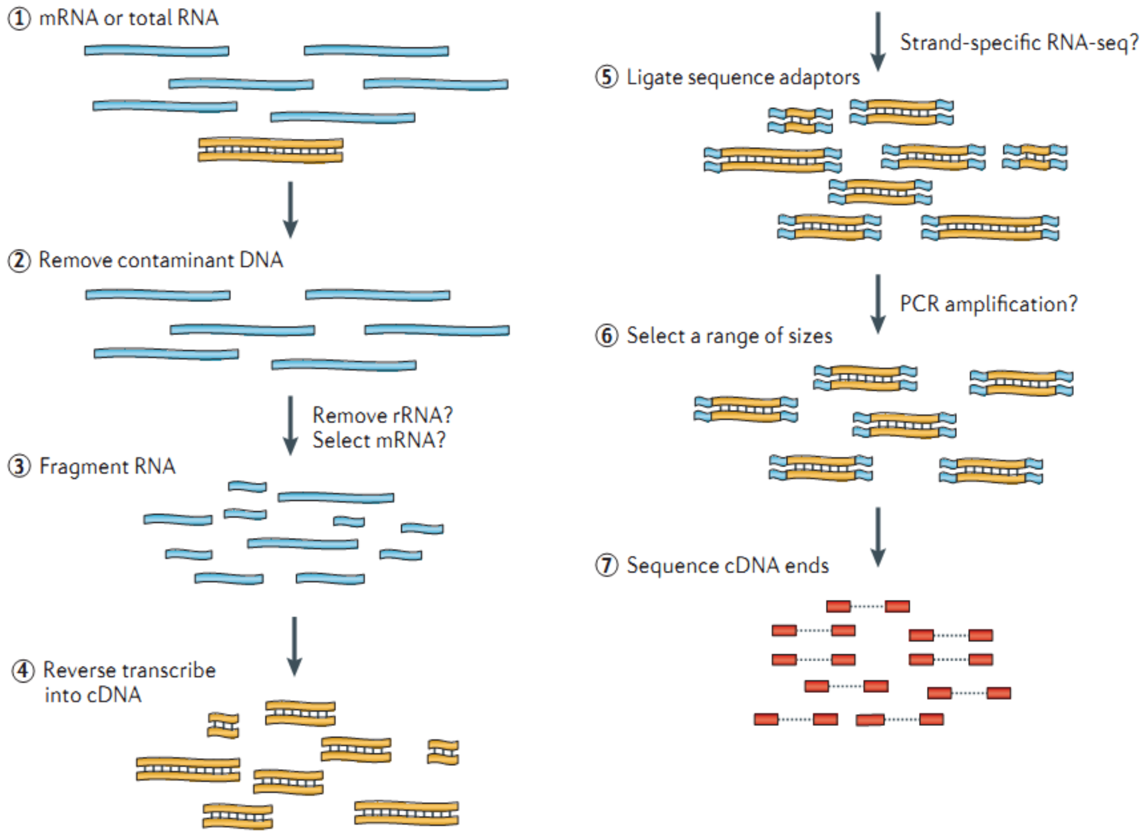


Figure 1. RNA Sequencing Experiment Workflow

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) Depending on the types of RNA, choose an appropriate kit for library prep process. For mRNA with poly-A tail, use mRNA purification kit; for noncoding 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.

8) For paired-end sequencing, both ends of the cDNA is sequenced by the read length.

2. Analysis Methods and Workflow

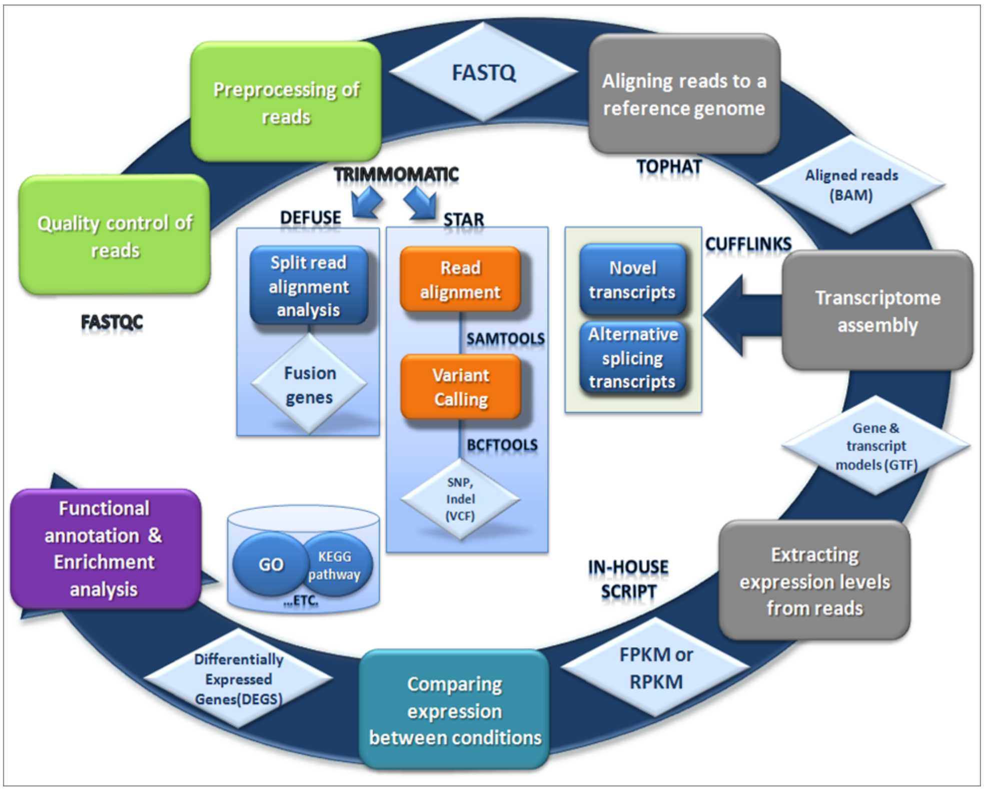

FASTQ
Preprocessing of reads
Aligning reads to a reference genome
TRIMMOMATIC
TOPHAT
Aligned reads (BAM)
Quality control of reads
DEFUSE
STAR
CUFFLINKS
Split read alignment analysis
Read alignment
Novel transcripts
Transcriptome assembly
FASTQC
SAMTOOLS
Alternative splicing transcripts
Fusion genes
Variant Calling
BCFTOOLS
Gene & transcript models (GTF)
SNP, Indel (VCF)
Functional annotation & Enrichment analysis
GO
KEGG pathway
...ETC.
IN-HOUSE SCRIPT
Extracting expression levels from reads
FPKM or RPKM
Differentially Expressed Genes(DEGS)
Comparing expression between conditions

Figure 2. 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 biased in analysis, artifacts such as low quality reads, adaptor sequence, contaminant DNA, or PCR duplicates are removed.

3) Aligned reads are generated using TopHat to align reads against the reference genome.

4) Transcript assembly of aligned reads is generated using Cufflinks. This process provides information on known transcripts, novel transcripts, and alternative splicing transcripts.

5) Mapped transcripts per sample allow calculation of differentially expressed profiles. Expression profiles between samples are compared through normalization of transcript length and depth of coverage. For paired-end sequencing FPKM (Fragments Per Kilobase of transcript per Million Mapped reads), for single end sequencing RPKM (Reads Per Kilobase of transcript per Million mapped reads), values are used within normalization for obtaining expression profile.

6) For groups of two or more with different conditions, genes or transcripts that express differentially are chosen through hypothesis verification.

7) Functional annotation and gene-set enrichment analysis was performed using GO and KEGG database on differentially expressed genes.

8) If SNV calling is done on RNA seq data, reads are mapped on genomic DNA reference using Star. Afterwards, the variant calling on the reads are executed using SAMTOOLS and BCFTOOLS.
LINK http://samtools.sourceforge.net/
LINK https://samtools.github.io/bcftools/bcftools.html

9) deFuse program is used to predict fusion genes.

3. Data Production Summary

3. 1. Raw Data Basic Statistics

(Refer to Path: 0.Stats 〉 rawData 〉 raw_throughput.stats)

The transcriptome raw data total read bases, number of reads, GC (%), Q20(%), Q30(%) of the 12 samples are calculated. For example, the CRH-WT2 sample produced 85,951,308 reads, and total length combined was 8.7Gbp. The GC content (%) was 50.31% and percentage of reads with over Q30 was 87.77%.

Table 1: Raw data stats

Index	Sample id	Total read bases*	Total reads	GC(%)	Q20(%)	Q30(%)
1	CRH-TG1	9,416,804,690	93,235,690	49.94	93.72	87.85
2	CRH-TG2	8,363,304,394	82,804,994	50.24	93.65	87.70
3	CRH-WT2	8,681,082,108	85,951,308	50.31	93.58	87.77
4	CRH-WT5	8,213,938,322	81,326,122	49.61	93.90	87.99
5	AG-WT-con1	8,575,900,304	84,909,904	50.28	93.59	87.65
6	AG-WT-con2	8,229,402,230	81,479,230	49.83	93.84	87.89
7	AG-PDK4KO-con1	8,561,801,714	84,770,314	50.46	97.42	95.74
8	AG-PDK4KO-con2	8,216,771,574	81,354,174	50.69	97.77	96.29
9	AG-WT-ACTH1-1h	9,623,999,524	95,287,124	50.33	97.87	96.44
10	AG-WT-ACTH3-1h	8,814,578,656	87,273,056	50.39	97.40	95.71
11	AG-PDK4KO-ACTH1-1h	8,261,083,708	81,792,908	50.39	97.82	96.36
12	AG-PDK4KO-ACTH3-1h	9,227,188,098	91,358,298	50.28	97.89	96.46

(* 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 reads that have phred quality score over 20
- Q30(%) : Ratio of reads that have phred quality score over 30

3. 2. Average Base Quality at Each Cycle

(Refer to path: 0.Stats 〉 rawData 〉 A_fastqc)

The quality of produced data is determined by the phred quality score of each reads. FastQC can be used to produce the box plot containing the average read quality. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

The x-axis shows number of cycles; y-axis shows phred quality score. Phred quality score 20 means 99% accuracy and reads over score 20 can be accepted as good quality reads.

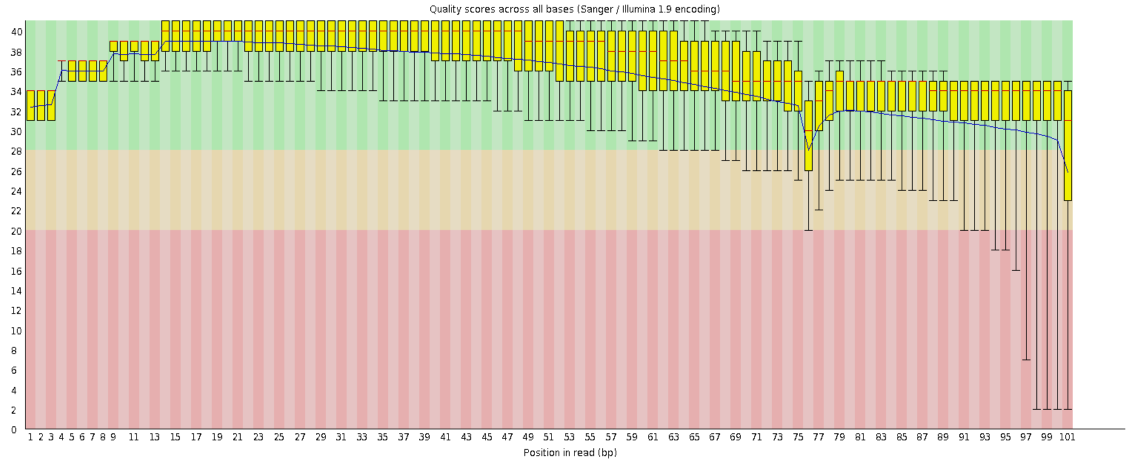


Figure 3. Read quality per cycle of CRH-TG1 (read1)

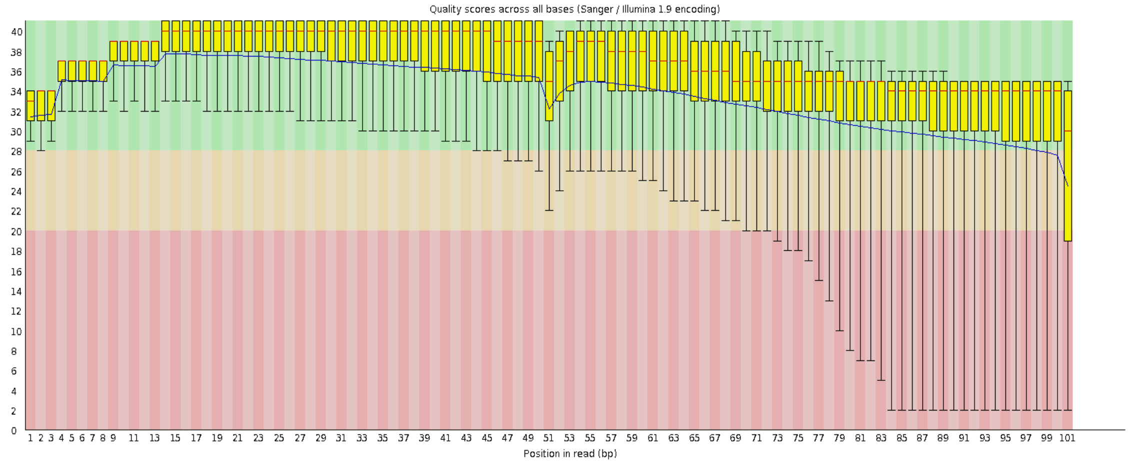


Figure 4. Read quality per cycle of CRH-TG1 (read2)

- Yellow box : Interquartile range (25-75%) of phred score per cycle
- Red line : Median of phred score per cycle
- Blue line : Average of phred score per cycle
- Green background : Good quality
- Orange background : Acceptable quality
- Red background : Bad quality

3. 3. Trimming Data Basic Statistics

(Refer to Path: 0.Stats 〉 trimmedData 〉 trim_throughput.stats)

Before starting analysis, Trimmomatic program is used to remove adapter sequences and remove reads with base quality lower than three from the ends. Also using sliding window trim method, reads that does not qualify for window size=4, and mean quality=15 are removed. Afterwards, reads with minimum length of 36bp are removed to produce cleaned data.

Table 2. Trimmed Data Stats

Index	Sample id	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
1	CRH-TG1	8,545,231,955	87,534,134	49.66	98.79	93.54
2	CRH-TG2	7,581,019,888	77,683,680	49.97	98.77	93.45
3	CRH-WT2	7,865,437,812	80,649,260	49.98	98.79	93.59
4	CRH-WT5	7,477,496,332	76,537,836	49.42	98.77	93.48
5	AG-WT-con1	7,767,133,826	79,601,946	50.01	98.77	93.45
6	AG-WT-con2	7,486,139,905	76,686,066	49.58	98.76	93.43
7	AG-PDK4KO-con1	8,487,985,707	84,314,174	50.42	97.80	96.21
8	AG-PDK4KO-con2	8,155,114,686	80,971,348	50.66	98.09	96.68
9	AG-WT-ACTH1-1h	9,560,866,500	94,912,352	50.30	98.16	96.79
10	AG-WT-ACTH3-1h	8,740,094,056	86,817,038	50.35	97.77	96.16
11	AG-PDK4KO-ACTH1-1h	8,201,787,556	81,430,414	50.36	98.12	96.73
12	AG-PDK4KO-ACTH3-1h	9,164,289,490	90,976,062	50.25	98.18	96.82

- Total read bases : Total number of reads bases after Trimming
- Total reads : Total number of reads after Trimming
- GC(%) : GC Content
- Q20(%) : Ratio of reads that have phred quality score over 20
- Q30(%) : Ratio of reads that have phred quality score over 30

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

(Refer to Path: 0.Stats 〉 trimmedData 〉 A_fastqc)

Figure 5 and 6 shows average base quality at each cycle after trimming.

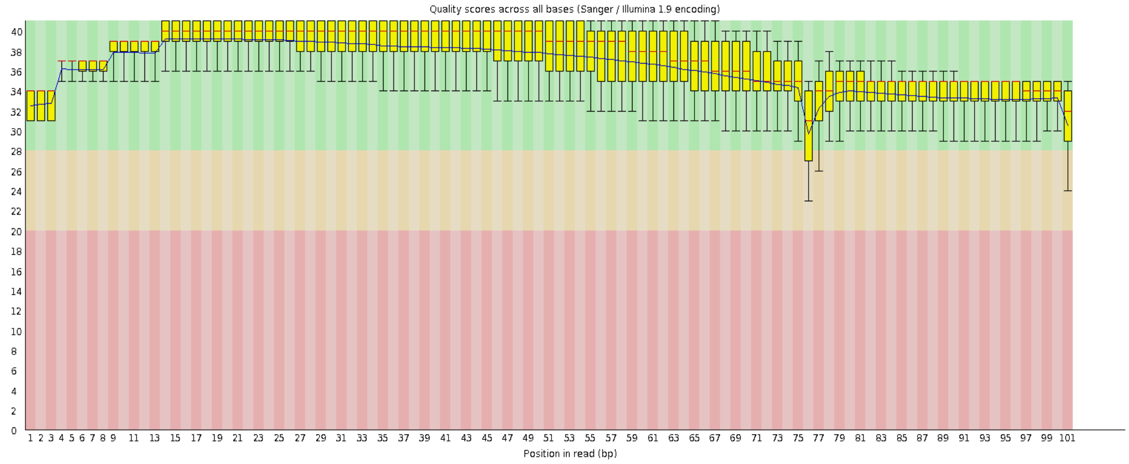


Figure 5. Average base quality of CRH-TG1 (read1) at each cycle after Trimming

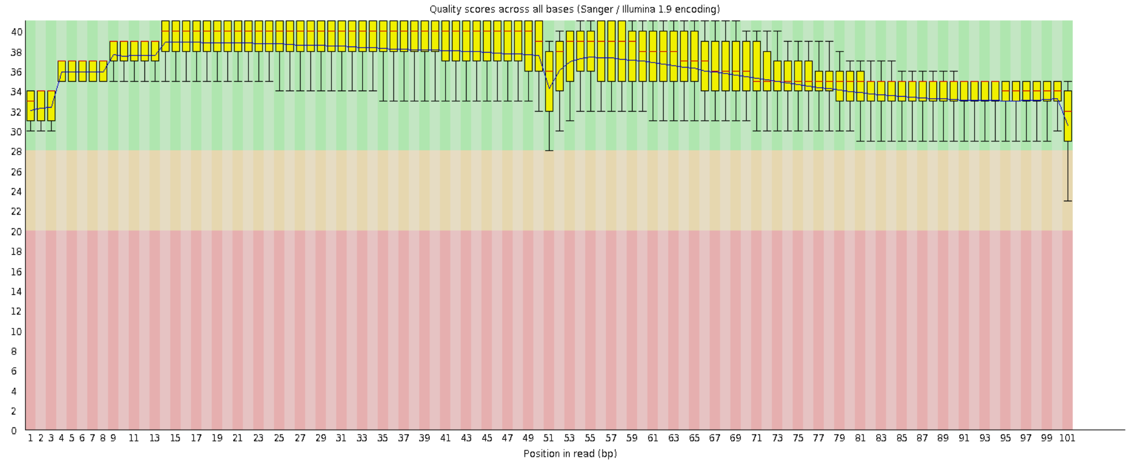


Figure 6. Average base quality of CRH-TG1 (read2) at each cycle after Trimming

- Yellow box : Interquartile range (25-75%) of phred score per cycle
- Red line : Median of phred score per cycle
- Blue line : Average of phred score per cycle
- Green background : Good quality
- Orange background : Acceptable quality
- Red background : Bad quality

4. Reference Mapping and Assembly Results

4. 1. Mapping Data Stats

(Refer to Path: 0.Stats 〉 mapping.stats)

In order to map cDNA fragments obtained from RNA seq process, genome DNA reference of was used. Below shows the statistic obtained from Tophat, which is obtained from spliced read mapping through Bowtie aligner. You can check number of processed reads, number of mapped reads, number of reads removed by multiple mapping, and overall mapping ratio.

Figure 3. Mapped Data Stats

Sample id	read type	# of processed reads	# of mapped reads	# of suppressed reads by multiple mapping	overall read mapping ratio
CRH-WT2	1	40,324,630	38,729,285 (96.0%)	2,272,540 (5.9%)	96.2%
CRH-WT2	2	40,324,630	38,869,909 (96.4%)	2,283,696 (5.9%)	
CRH-WT5	1	38,268,918	37,224,219 (97.3%)	2,229,995 (6.0%)	97.4%
CRH-WT5	2	38,268,918	37,289,452 (97.4%)	2,234,676 (6.0%)	
CRH-TG1	1	43,767,067	42,159,001 (96.3%)	3,382,148 (8.0%)	96.4%
CRH-TG1	2	43,767,067	42,232,132 (96.5%)	3,390,123 (8.0%)	
CRH-TG2	1	38,841,840	37,425,186 (96.4%)	2,716,788 (7.3%)	96.4%
CRH-TG2	2	38,841,840	37,486,149 (96.5%)	2,722,444 (7.3%)	
AG-WT-con1	1	39,800,973	38,421,717 (96.5%)	2,184,310 (5.7%)	96.6%
AG-WT-con1	2	39,800,973	38,490,718 (96.7%)	2,190,769 (5.7%)	
AG-WT-con2	1	38,343,033	37,102,219 (96.8%)	2,373,663 (6.4%)	96.8%
AG-WT-con2	2	38,343,033	37,164,682 (96.9%)	2,379,314 (6.4%)	

AG-WT-ACTH1-1h	1	47,456,176	45,460,666 (95.8%)	3,304,866 (7.3%)	95.4%
AG-WT-ACTH1-1h	2	47,456,176	45,111,113 (95.1%)	3,281,071 (7.3%)	
AG-WT-ACTH3-1h	1	43,408,519	41,558,038 (95.7%)	3,128,100 (7.5%)	95.1%
AG-WT-ACTH3-1h	2	43,408,519	40,997,828 (94.4%)	3,088,461 (7.5%)	
AG-PDK4KO-con1	1	42,157,087	40,384,677 (95.8%)	2,351,800 (5.8%)	95.2%
AG-PDK4KO-con1	2	42,157,087	39,855,228 (94.5%)	2,322,379 (5.8%)	
AG-PDK4KO-con2	1	40,485,674	38,830,570 (95.9%)	2,204,747 (5.7%)	95.5%
AG-PDK4KO-con2	2	40,485,674	38,494,941 (95.1%)	2,186,793 (5.7%)	
AG-...-ACTH1-1h	1	40,715,207	38,864,608 (95.5%)	2,632,567 (6.8%)	95.1%
AG-...-ACTH1-1h	2	40,715,207	38,568,816 (94.7%)	2,613,133 (6.8%)	
AG-...-ACTH3-1h	1	45,488,031	43,580,603 (95.8%)	3,102,594 (7.1%)	95.4%
AG-...-ACTH3-1h	2	45,488,031	43,234,938 (95.0%)	3,078,957 (7.1%)	

- # of processed reads : Number of cleaned reads after trimming
- # of mapped reads : Number of reads mapped against the reference
- # of suppressed reads by multiple mapping : Number of reads removed due to multiple mapping
- overall read mapping ratio : # of total mapped reads / # of total processed reads

4. 2. Transcriptome Assembly and Expression Level

Cufflinks with the reference gene model can be used to assemble novel transcripts, alternative splicing transcripts and known transcripts.

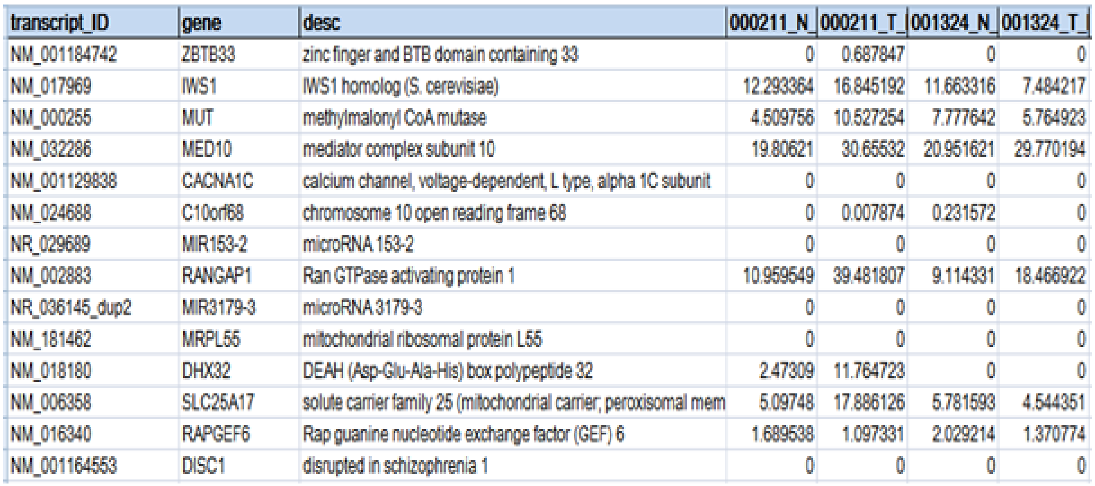
After assembly, the abundance of transcripts is shown in within sample normalized value. In the case of paired-end sequencing, FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and in the case of single-end sequencing, RPKM (Reads Per Kilobase of Transcript per Million Mapped reads) can be calculated.

4. 2. 1. Known transcripts expression level

(Refer to Path: 1.Expression_profile_G 〉 AnnoOnly_FPKM_from_all_samples_in_mm10.addDesc.xlsx)

Table 4 is an example of known transcript expression level per sample in FPKM value. This result is obtained by Reference Annotation Based Transcript (RABT) method using -G option of Cufflinks without novel transcript assembly.

Table 4. Known transcripts Expression Level (example)



transcript_ID	gene	desc	000211_N_	000211_T_	001324_N_	001324_T_
NM_001184742	ZBTB33	zinc finger and BTB domain containing 33	0	0.687847	0	0
NM_017969	IWS1	IWS1 homolog (S. cerevisiae)	12.293364	16.845192	11.663316	7.484217
NM_000255	MUT	methylmalonyl CoA mutase	4.509756	10.527254	7.777642	5.764923
NM_032286	MED10	mediator complex subunit 10	19.80621	30.65532	20.951621	29.770194
NM_001129838	CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit	0	0	0	0
NM_024688	C10orf68	chromosome 10 open reading frame 68	0	0.007874	0.231572	0
NR_029689	MIR153-2	microRNA 153-2	0	0	0	0
NM_002883	RANGAP1	Ran GTPase activating protein 1	10.959549	39.481807	9.114331	18.466922
NR_036145_dup2	MIR3179-3	microRNA 3179-3	0	0	0	0
NM_181462	MRPL55	mitochondrial ribosomal protein L55	0	0	0	0
NM_018180	DHX32	DEAH (Asp-Glu-Ala-His) box polypeptide 32	2.47309	11.764723	0	0
NM_006358	SLC25A17	solute carrier family 25 (mitochondrial carrier; peroxisomal mem	5.09748	17.886126	5.781593	4.544351
NM_016340	RAPGEF6	Rap guanine nucleotide exchange factor (GEF) 6	1.689538	1.097331	2.029214	1.370774
NM_001164553	DISC1	disrupted in schizophrenia 1	0	0	0	0

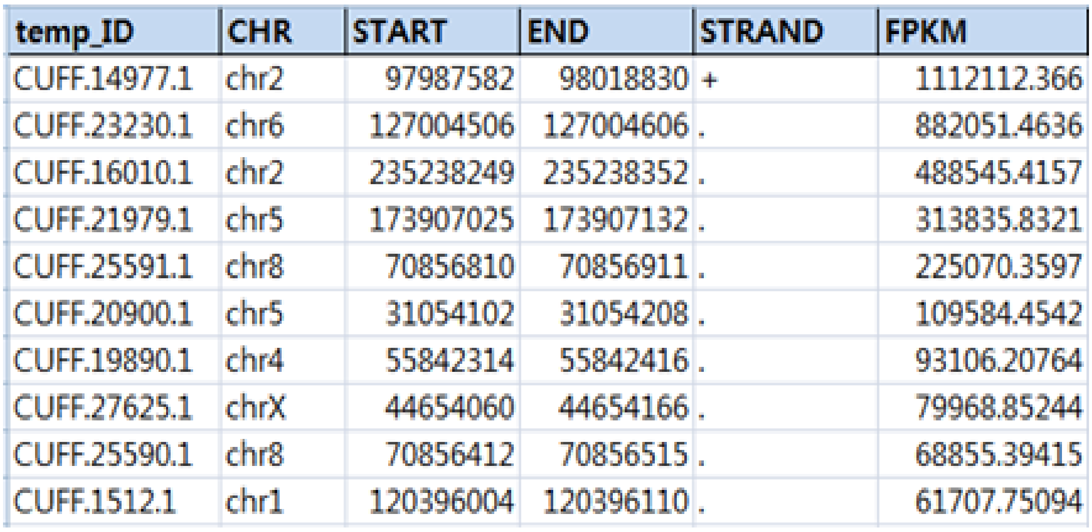
- Transcript_ID : splicing variant (isoform/transcript)
- Gene : Name of the gene
- Description : Description of the gene
- [Sample Name]_FPKM : FPKM normalized value per sample

4. 2. 2. Novel Transcripts

(Refer to Path: 2.Expression_profile_g 〉 novel_in_*.xlsx)

Novel transcripts are produced by reads that are mapped against novel exons or genes. Table 5 is an example of results obtained by cufflinks Reference Annotation Based Transcript Assembly (RABT) method, allowing discovery of reference transcripts and novel transcripts using –g option.

Table 5. Novel transcript List (Example)



temp_ID	CHR	START	END	STRAND	FPKM
CUFF.14977.1	chr2	97987582	98018830	+	1112112.366
CUFF.23230.1	chr6	127004506	127004606	.	882051.4636
CUFF.16010.1	chr2	235238249	235238352	.	488545.4157
CUFF.21979.1	chr5	173907025	173907132	.	313835.8321
CUFF.25591.1	chr8	70856810	70856911	.	225070.3597
CUFF.20900.1	chr5	31054102	31054208	.	109584.4542
CUFF.19890.1	chr4	55842314	55842416	.	93106.20764
CUFF.27625.1	chrX	44654060	44654166	.	79968.85244
CUFF.25590.1	chr8	70856412	70856515	.	68855.39415
CUFF.1512.1	chr1	120396004	120396110	.	61707.75094

- Temp_ID : If there are several transcripts within the same gene region, cufflinks assign an temporary “CUFF.xxxx.y” ID. Here xxxx specifies the gene region’s locus ID, and y specifies the specific number of transcript occurring in that region.

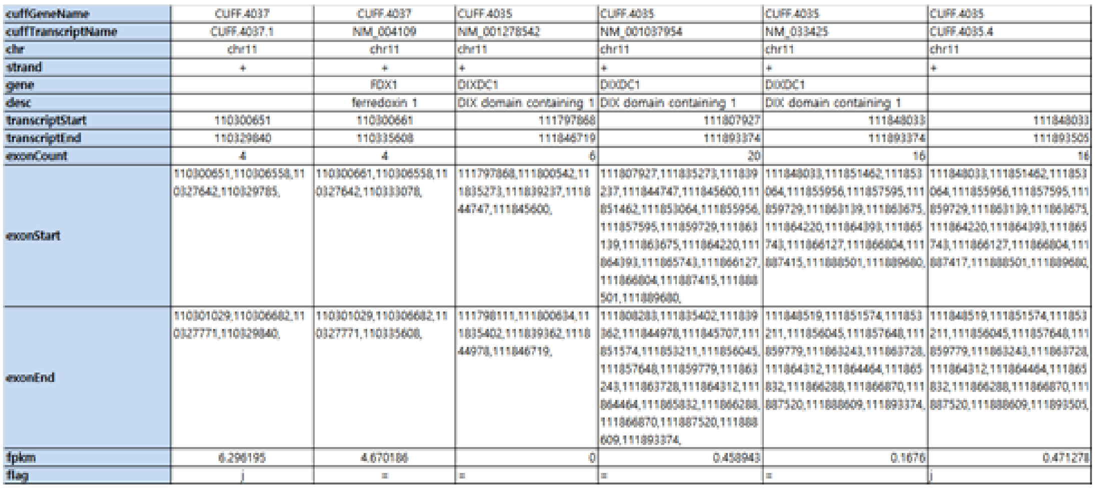
4. 2. 3. Novel Alternative splicing transcript

(Refer to Path: 2.Expression_profile_g 〉 novelSplicingVariant_*.addDesc.xlsx)

This refers to transcripts that did not map on known exon but mapped on a novel exon or transcripts that show different structure from usual isoforms. Table 6 shows an example of results obtained from cufflink using the –g option.

If novel alternative splicing transcript exists, GeneName and transcriptName is numbered using prefix “CUFF”. If TranscriptName is a known transcript, it is identified as RefSeq number, however if it is a novel splicing variant, it is identified as CUFF ID. Transcript start, transcript end, exon count, exon start, exon end position, FPKM flag value is provided for each transcript.

Table 6. Alternative splicing transcript list (Example)



cuffGeneName	CUFF.4037	CUFF.4037	CUFF.4035	CUFF.4035	CUFF.4035	CUFF.4035
cuffTranscriptName	CUFF.4037.1	NM_004109	NM_001278542	NM_001037954	NM_033425	CUFF.4035.4
chr	chr11	chr11	chr11	chr11	chr11	chr11
strand	+	+	+	+	+	+
gene		FDX1	DIXDC1	DIXDC1	DIXDC1	
desc		ferredoxin 1	DIX domain containing 1	DIX domain containing 1	DIX domain containing 1	
transcriptStart	110300651	110300661	111797868	111807927	111848033	111848033
transcriptEnd	110329840	110335608	111846719	111893374	111893374	111893505
exonCount	4	4	6	20	16	16
exonStart	110300651,110306558,110327642,110329785,	110300661,110306558,110327642,110333078,	111797868,111800542,111835273,111839237,111844747,111845600,	111807927,111835273,111839237,111844747,111845600,111851462,111853064,111855956,111857595,111859729,111863139,111863675,111864220,111864393,111865743,111866127,111866804,111887415,111888501,111889680,	111848033,111851462,111853064,111855956,111857595,111859729,111863139,111863675,111864220,111864393,111865743,111866127,111866804,111887415,111888501,111889680,	111848033,111851462,111853064,111855956,111857595,111859729,111863139,111863675,111864220,111864393,111865743,111866127,111866804,111887417,111888501,111889680,
exonEnd	110301029,110306682,110327771,110329840,	110301029,110306682,110327771,110335608,	111798111,111800634,111835402,111839362,111844978,111846719,	111808283,111835402,111839362,111844978,111845707,111851574,111853211,111856045,111857648,111859779,111863243,111863728,111864312,111864464,111865832,111866288,111866870,111887520,111888609,111893374,	111848519,111851574,111853211,111856045,111857648,111859779,111863243,111863728,111864312,111864464,111865832,111866288,111866870,111887520,111888609,111893374,	111848519,111851574,111853211,111856045,111857648,111859779,111863243,111863728,111864312,111864464,111865832,111866288,111866870,111887520,111888609,111893505,
fpkm	6.298195	4.670186	0	0.458943	0.1676	0.471278
flag	j	=	=	=	=	j

- Flag : “j” identifies novel splicing alternative transcript, “=” identifies known transcript.

5. Differentially Expressed Gene Analysis Results

5. 1. Data Analysis Quality Check and Workflow

After transcriptome assembly, the FPKM value of known transcripts and differentially expressed genes are selected. Before further analysis, data quality check, normalization between samples, and if biological replicates are present, the similarity between samples is checked and the data quality is verified.

(Refer to Path: 1.Expression_profile_G 〉 DEG_result)

5. 1. 1. Sample information and analysis design

Total of 12 samples were used for analysis.

Index	Sample.ID	Sample.Group
1	AG-PDK4KO-con1	AG-PDK4KO
2	AG-PDK4KO-con2	AG-PDK4KO
3	AG-PDK4KO-ACTH1-1h	AG-PDK4KO-ACTH
4	AG-PDK4KO-ACTH3-1h	AG-PDK4KO-ACTH
5	AG-WT-con1	AG-WT
6	AG-WT-con2	AG-WT
7	AG-WT-ACTH1-1h	AG-WT-ACTH
8	AG-WT-ACTH3-1h	AG-WT-ACTH
9	CRH-TG1	CRH-TG
10	CRH-TG2	CRH-TG
11	CRH-WT2	CRH-WT
12	CRH-WT5	CRH-WT

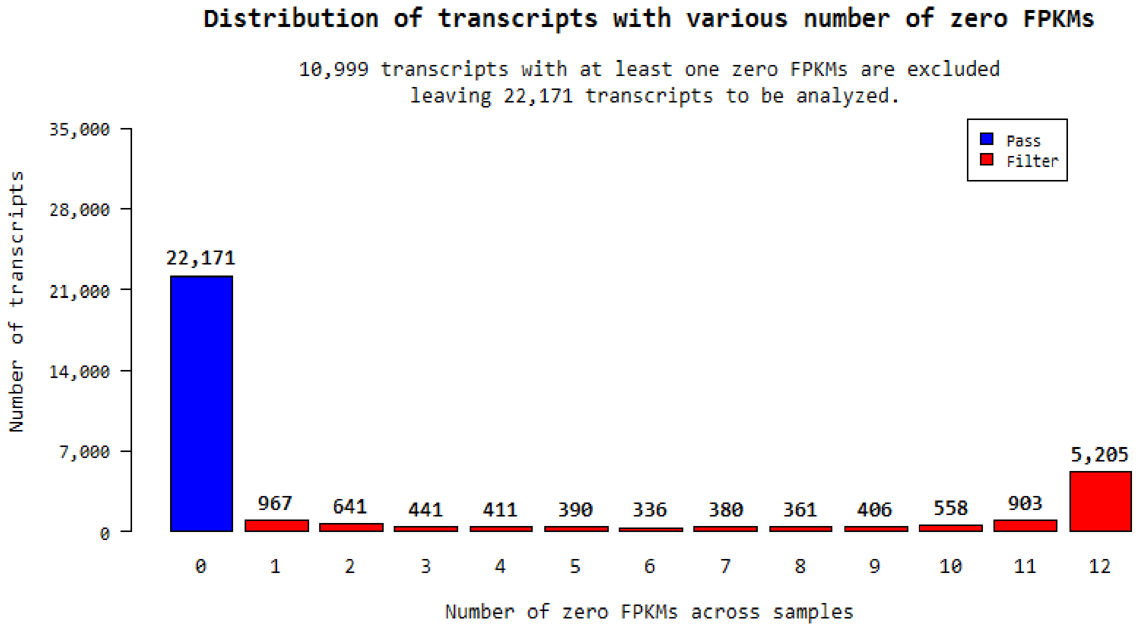
Comparison pair and the results statistics method is as follows.

Index	Test vs. Control	Statistical Method
1	CRH-TG vs. CRH-WT	Fold Change, LPE Test, Hierarchical Clustering
2	AG-WT-ACTH vs. AG-WT	Fold Change, LPE Test, Hierarchical Clustering
3	AG-PDK4KO vs. AG-WT	Fold Change, LPE Test, Hierarchical Clustering
4	AG-PDK4KO-ACTH vs. AG-WT-ACTH	Fold Change, LPE Test, Hierarchical Clustering
5	AG-PDK4KO-ACTH vs. AG-PDK4KO	Fold Change, LPE Test, Hierarchical Clustering

5. 1. 2. DATA Quality Check

(Refer to Path: 1.Expression_profile_G 〉 DEG_result 〉 Data Quality Check)

각 transcript 별, 전체 12개 샘플에서 적어도 한 샘플 이상에서 0인 FPKM값을 가지는 transcript는 분석에서 제외하였습니다. 따라서, 총 33,170개 transcript 중에서 10,999개를 제외한 22,171개 transcript을 대상으로 통계분석을 진행하였습니다.

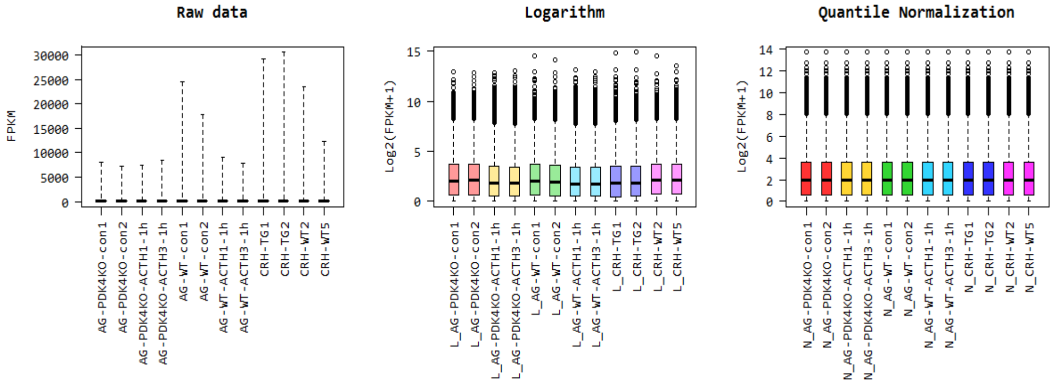


5. 1. 3. Data Alteration and Normalization

The Raw signal(FPKM)+1 is selected and simplified and processed with log2 based transformation. The reason for this is because raw signals are scattered along wide range and most signals are concentrated on the low signal value, so log transformation reduces the range of the signals and produces more even data distribution. After log transformation, in order to reduce systematic bias, quantile normalization is used to normalize data between samples. ('preprocessCore' R library used).

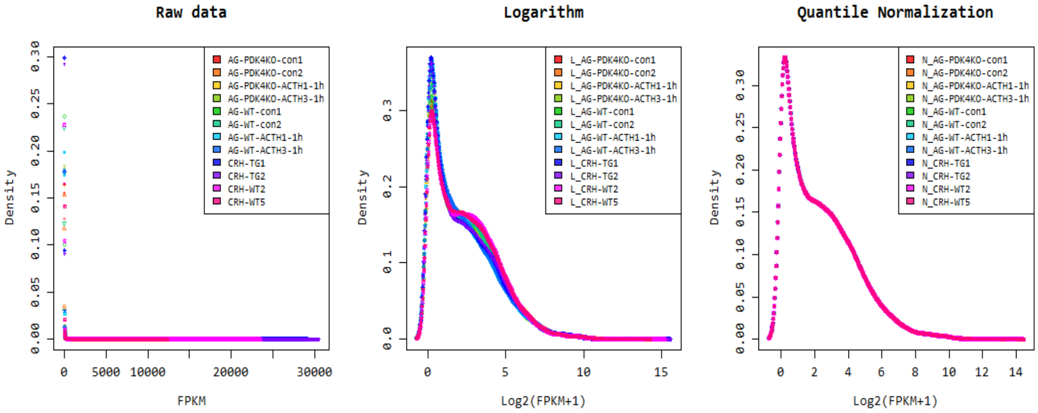
5. 1. 3. 1. Boxplot of expression difference between samples.

Below boxplots show before and after of raw signal (FPKM)+1 Log2 transformation, before after of Quintile Normalization and corresponding sample's expression scatter based on percentile, median, 50 percentile, 75 percentile, maximum and minimum.



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

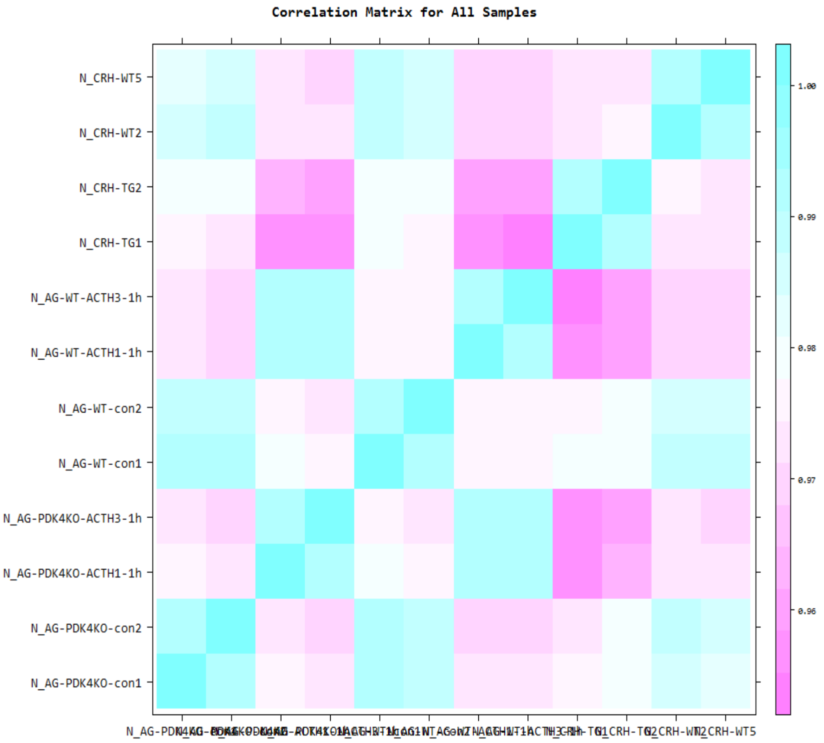
Below boxplots show before and after of raw signal (FPKM)+1 Log2 transformation, before after of Quintile Normalization and corresponding sample's expression scatter as a density plot.



5. 1. 4. Correlation Analysis between samples

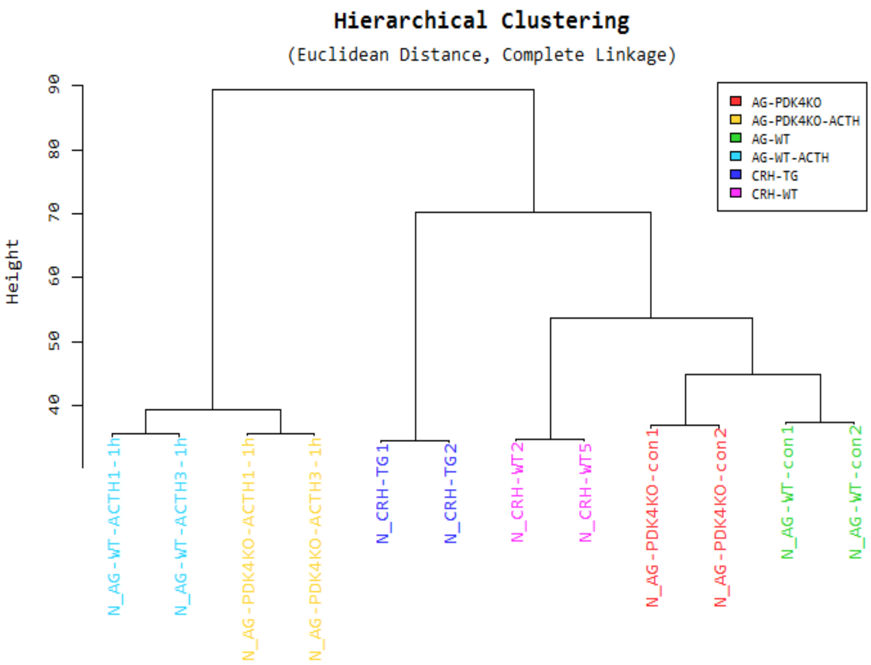
The similarity between samples are obtained through Pearson's coefficient of the Log2(FPKM+1) value. For range: -1≤ r ≤ 1, value closer to 1 means close correlation between samples.

Correlation matrix of all samples is as follows.



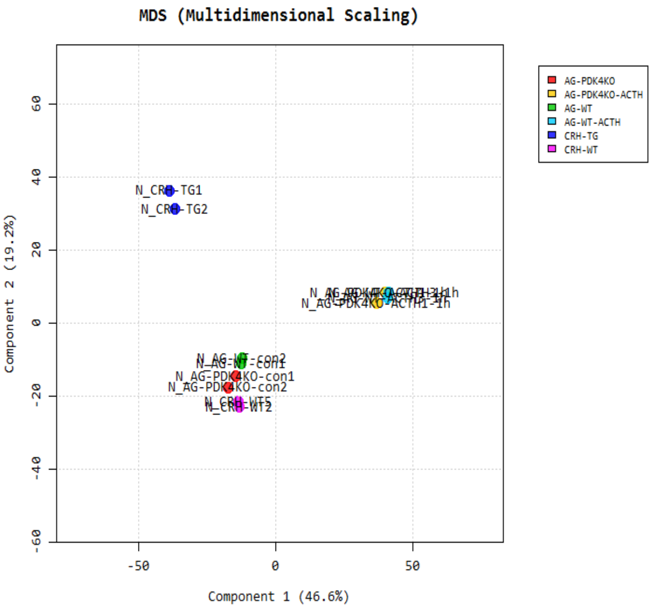
5. 1. 5. Hierarchical clustering Analysis

Using each sample's Log2(FPKM+1) value, the expression similarities were grouped together. (Distance metric = Euclidean distance, Linkage method= Complete Linkage)



5. 1. 6. MDS, Multidimensional Scaling

Using each sample's Log2(FPKM+1) value, the similarity between samples is graphically shown in a 2D plot to show the variability of the total data. This allows identification any outlier samples, or similar expression patterns between sample groups.



5. 2. Differentially Expressed Gene Analysis Workflow

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

1) the FPKM value of known transcriptions obtained through - G option of the Cufflinks were used as the original raw data.

- Raw data

(Refer to Path: 1.Expression_profile_G 〉 AnnoOnly_FPKM_from_all_samples_in_mm10.addDesc.xlsx)

: 33,170 transcripts, 12 samples

2) During data processing and QC process, low quality transcripts were filtered and log(FPKM+1) was performed. Afterwards, quantile normalization was performed.

- Processed data

(Refer to Path: 1.Expression_profile_G 〉 DEG_result 〉 data2.xlsx)

: 22,171 transcripts, 12 samples

3) Statistics Analysis was performed using Fold Change, LPE Test per comparison pair and results were selected on conditions of |fc|≥2 & LPE test raw p-value〈0.05. data3_*.xlsx was saved significant transcripts which satisfied |fc|≥2 & LPE test raw p-value〈0.05 conditions at least one comparison.

(Refer to Path: 1.Expression_profile_G 〉 DEG_result)

- Significant data (data3_fc2 & lpe.p.xlsx)

: 1,555 transcripts

- Significant data (data3-CRH-TG_vs_CRH-WT_fc2 & lpe.p.xlsx)

: 808 transcripts

- Significant data (data3-AG-WT-ACTH_vs_AG-WT_fc2 & lpe.p.xlsx)

: 585 transcripts

- Significant data (data3-AG-PDK4KO_vs_AG-WT_fc2 & lpe.p.xlsx)

: 95 transcripts

- Significant data (data3-AG-PDK4KO-ACTH_vs_AG-WT-ACTH_fc2 & lpe.p.xlsx)

: 58 transcripts

- Significant data (data3-AG-PDK4KO-ACTH_vs_AG-PDK4KO_fc2 & lpe.p.xlsx)

: 600 transcripts

4) For significant gene list, hierarchical clustering analysis was performed to determine and group the similarities between samples and genes. These results were graphically depicted using heatmap and dendogram.

- Hierarchical Clustering (Euclidean Distance, Complete Linkage)

(Refer to Path: 1.Expression_profile_G 〉 DEG_result 〉 Cluster image)

5) For similar gene lists, gene ontology(http://geneontology.org/), KEGG(http://www.genome.jp/kegg/) etc., based gene-set enrichment analysis was performed using DAVID tool (http://david.abcc.ncifcrf.gov/).

Please refer to the second sheet (DAVID_cluster) of data3 file and the third sheet (DAVID_chart).

Following reports are provided.

- Functional annotation chart report
- Functional annotation clustering report

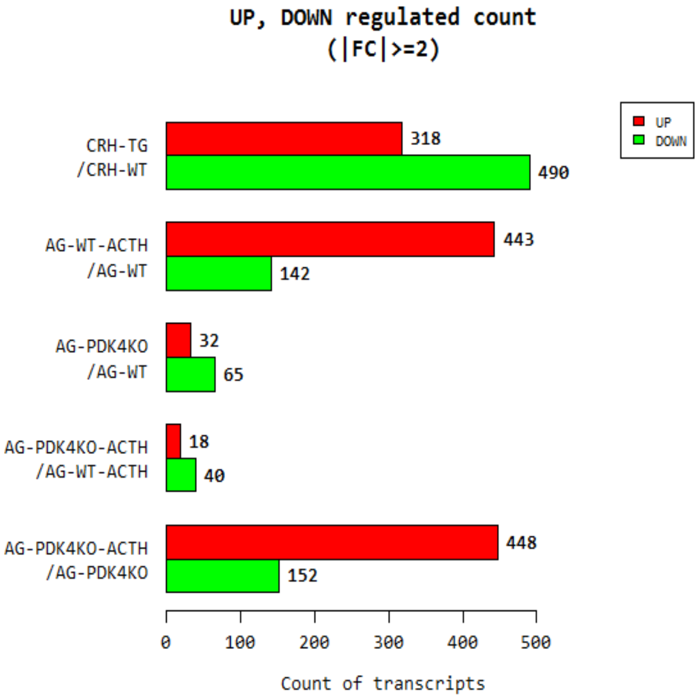
(Refer to Path: 1.Expression_profile_G 〉 DEG_result 〉 DAVID)

5. 3. Differentially expressed compare union statistics

(Refer to Path: 1.Expression_profile_G 〉DEG_result 〉Plots)

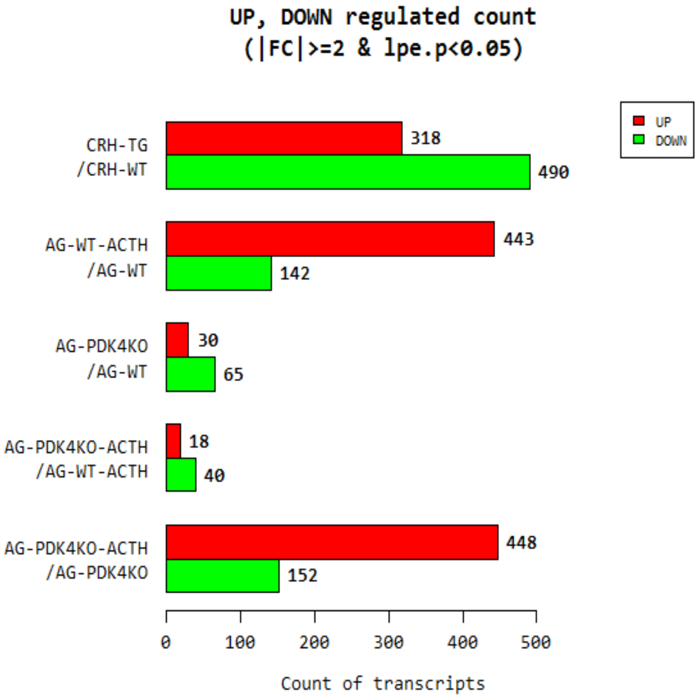
5. 3. 1. Number of transcripts per up and down based on fold change

Shows number of transcripts per up and down based on comparison pair fold change.



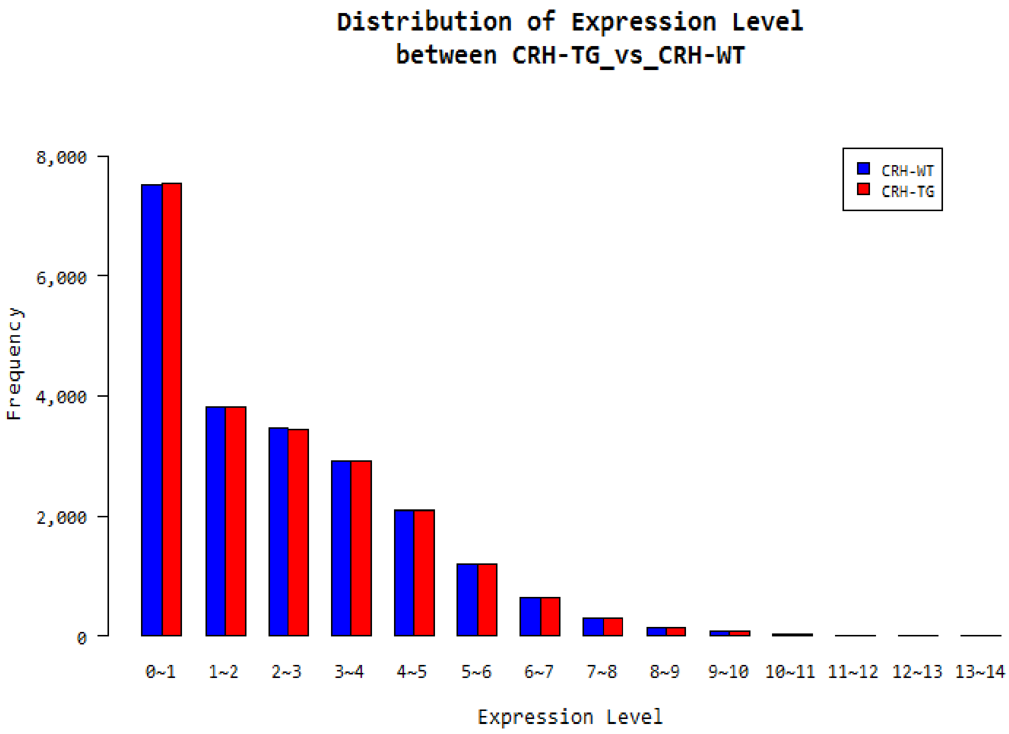
5. 3. 2. Number of transcripts per up and down based on fold change and p-values

Shows number of transcripts per up and down based on fold change and p-values.



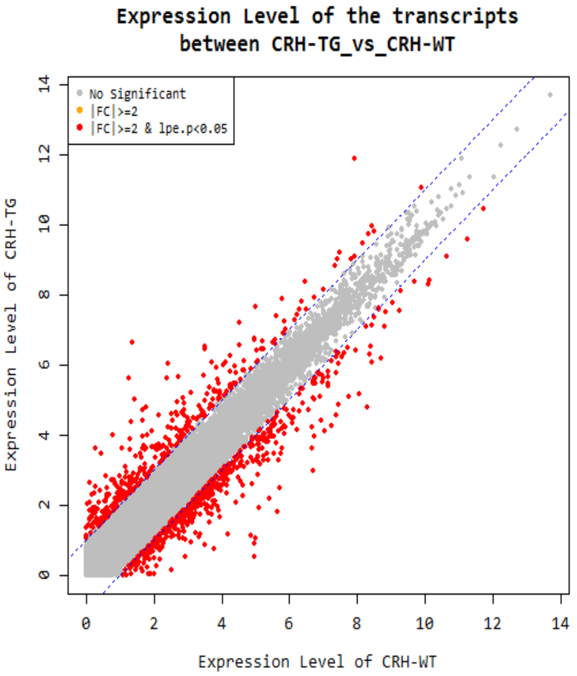
5. 3. 3. Distribution of expression level between two groups

Shows distribution of Normalized Log2(FPKM+1) per group for comparison pair.



5. 3. 4. Scatter plot of expression level between two groups

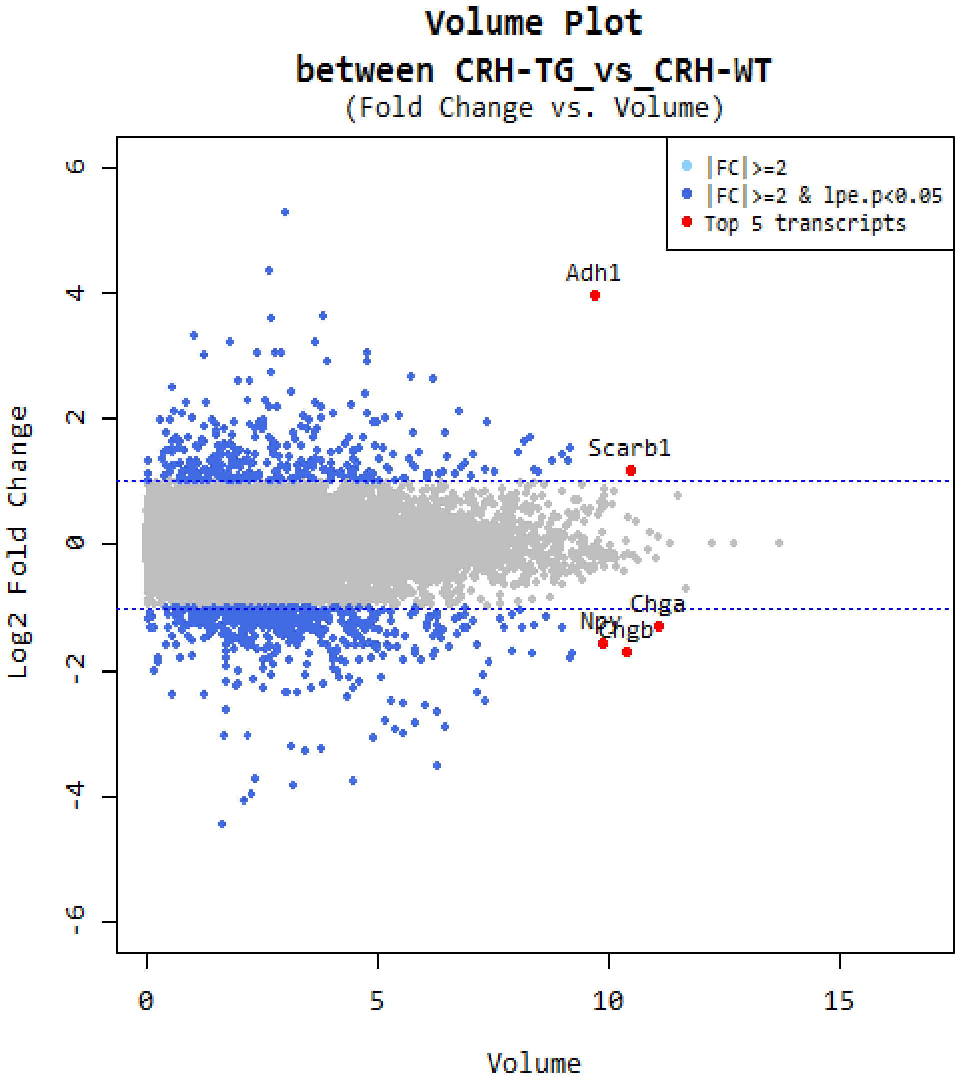
Shows expression levels between comparison pair as a scatter plot. X-axis as control and Y-axis as test group’s normalized value average.



5. 3. 5. Volume plot of different genes depending on expression volume

Expression volume was defined as the geometric average of two group's expression level In order to confirm the transcripts that showed higher expression volume compared to the control, volume plot was drawn. (X-axis: Volume, Y-axis: log2 Fold change).

For example, even though fold change might be different by two-fold, the transcripts with higher volume may be more credible.

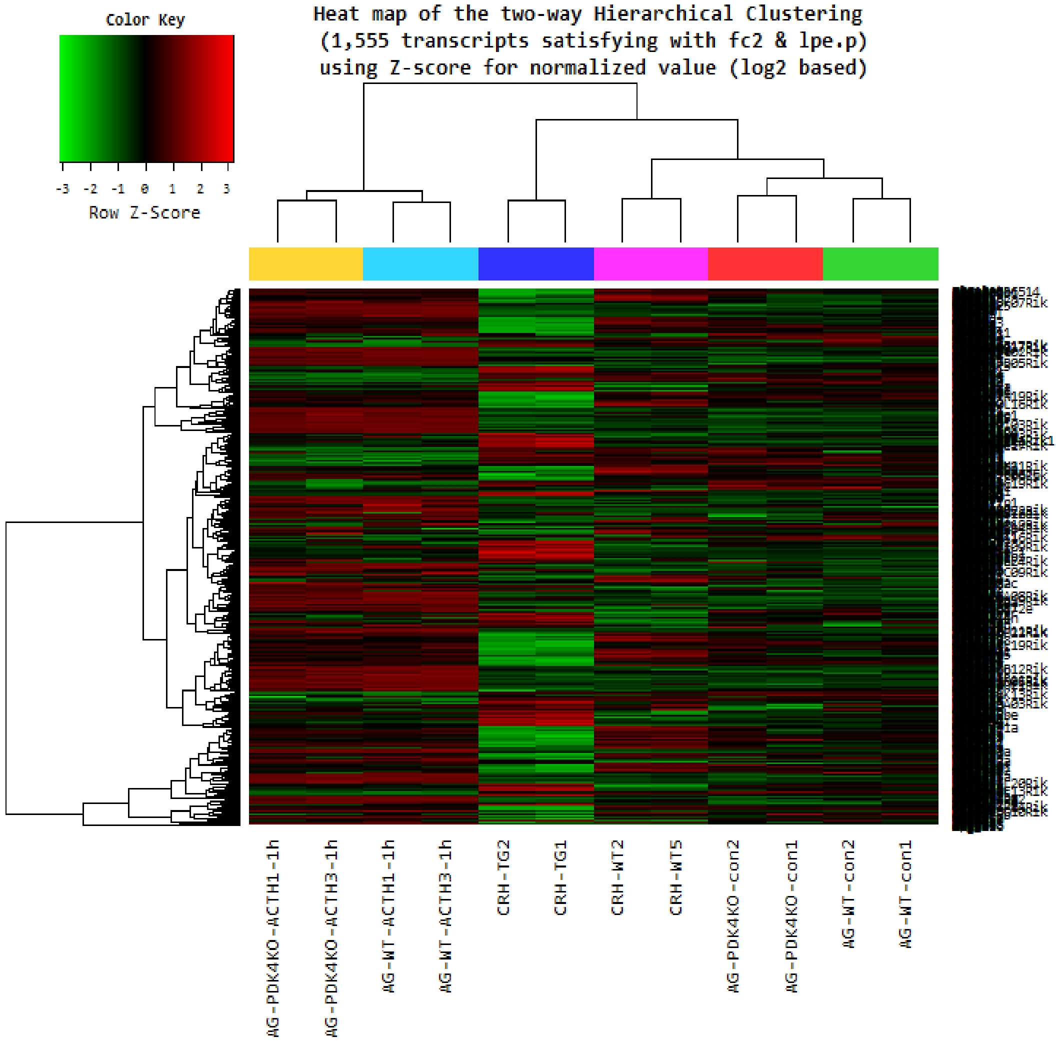


- red dot : Top five transcripts by volume which satisfies, |fc|≥2 & LPE test raw p-value<0.05

5. 3. 6. Hierarchical Clustering Analysis

(Refer to Path: 1.Expression_profile_G 〉 DEG_result 〉 Cluster image)

Heatmap shows results of hierarchical clustering analysis (Euclidean Method, Complete Linkage) of transcript groups of similar expression level (normalized value) from the DEG list at least one comparison.



5. 4. Function Classification and Gene-set enrichment Analysis

(Refer to Path: 1.Expression_profile_G 〉 DEG_result 〉 DAVID)

(Please refer to data3 file's second sheet (DAVID_cluster) and third sheet (DAVID_chart))

For DEG list, gene ontology(http://geneontology.org/),KEGG(http://www.genome.jp/kegg/) and other functional annotation database based gene-set enrichment analysis was performed using DAVID tool((http://david.abcc.ncifcrf.gov/).

Two reports are provided for Enrichment analysis.

- Functional annotation chart report
- Functional annotation clustering report

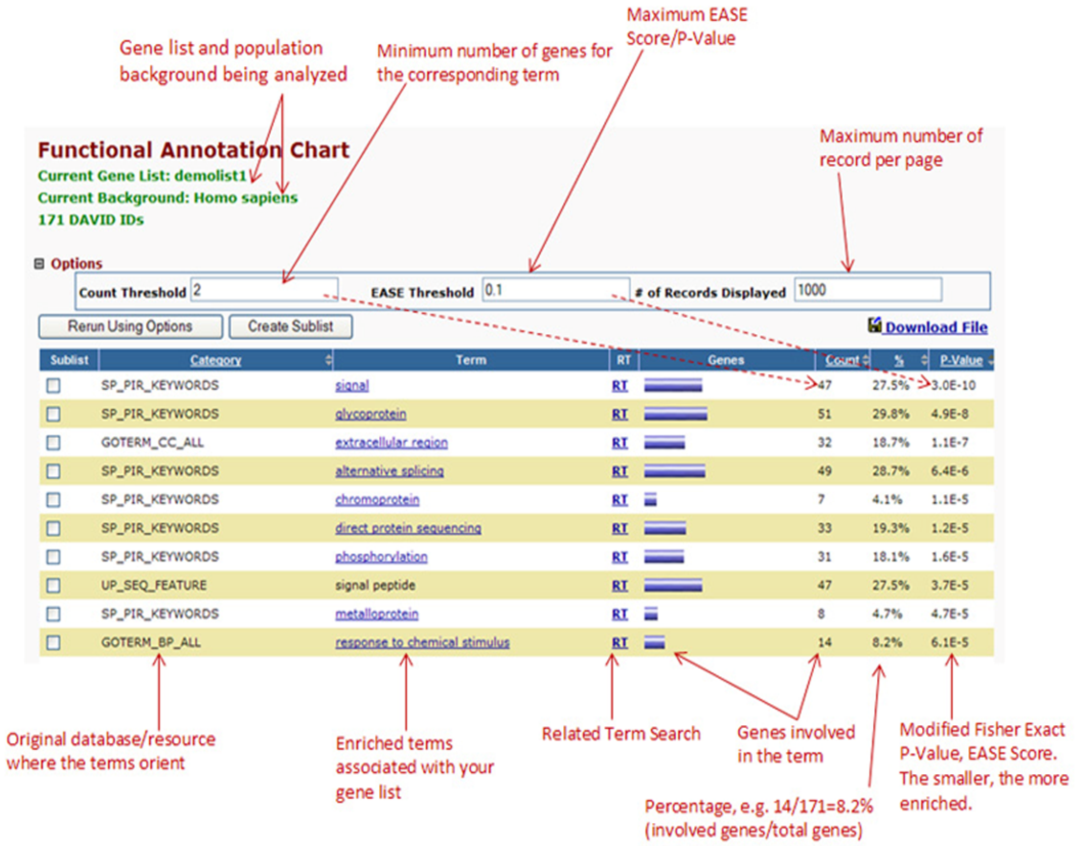
Chart below shows gene set databases that are used for DAVID tool.

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

5. 4. 1. Functional annotation chart report

Figure below shows example results of Functional annotation chart report.

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



Functional Annotation Chart

Current Gene List: demolist1

Current Background: Homo sapiens

171 DAVID IDs

⊟ Options

Count Threshold 2 EASE Threshold 0.1 # of Records Displayed 1000

Rerun Using Options | Create Sublist | Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value
□	SP_PIR_KEYWORDS	signal	RT		47	27.5%	3.0E-10
□	SP_PIR_KEYWORDS	glycoprotein	RT		51	29.8%	4.9E-8
□	GOTERM_CC_ALL	extracellular region	RT		32	18.7%	1.1E-7
□	SP_PIR_KEYWORDS	alternative splicing	RT		49	28.7%	6.4E-6
□	SP_PIR_KEYWORDS	chromoprotein	RT		7	4.1%	1.1E-5
□	SP_PIR_KEYWORDS	direct protein sequencing	RT		33	19.3%	1.2E-5
□	SP_PIR_KEYWORDS	phosphorylation	RT		31	18.1%	1.6E-5
□	UP_SEQ_FEATURE	signal peptide	RT		47	27.5%	3.7E-5
□	SP_PIR_KEYWORDS	metalloprotein	RT		8	4.7%	4.7E-5
□	GOTERM_BP_ALL	response to chemical stimulus	RT		14	8.2%	6.1E-5

Gene list and population background being analyzed

Minimum number of genes for the corresponding term

Maximum EASE Score/P-Value

Maximum number of record per page

Original database/resource where the terms orient

Enriched terms associated with your gene list

Related Term Search

Genes involved in the term

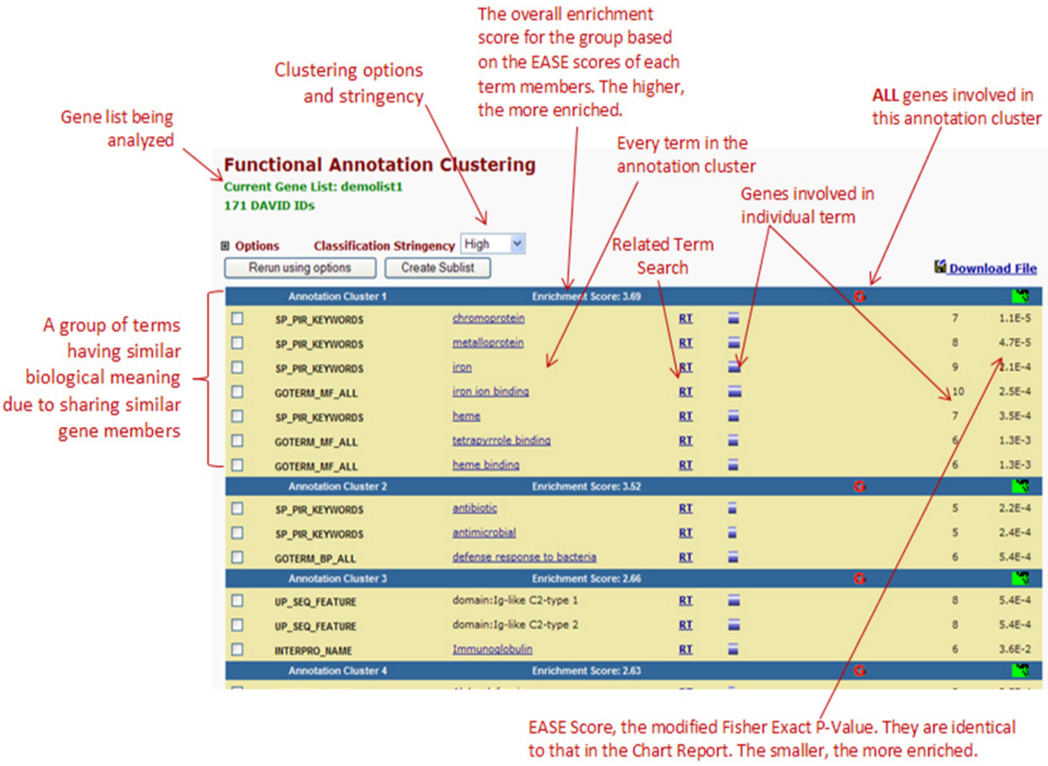
Percentage, e.g. 14/171=8.2% (involved genes/total genes)

Modified Fisher Exact P-Value, EASE Score. The smaller, the more enriched.

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

5. 4. 2. Functional annotation clustering report

Functional annotation clustering report groups similar gene members and gene set terms into “annotation clusters”, which undergoes the enrichment analysis. Below figure shows an example of the functional annotation clustering report.



Functional Annotation Clustering

Current Gene List: demolist1

171 DAVID IDs

⊞ Options Classification Stringency High

Rerun using options | Create Sublist

Download File

Annotation Cluster 1		Enrichment Score: 3.69				
□	SP_PIR_KEYWORDS	chromoprotein	RT	▬	7	1.1E-5
□	SP_PIR_KEYWORDS	metalloprotein	RT	▬	8	4.7E-5
□	SP_PIR_KEYWORDS	iron	RT	▬	9	2.1E-4
□	GOTERM_MF_ALL	iron ion binding	RT	▬	10	2.5E-4
□	SP_PIR_KEYWORDS	heme	RT	▬	7	3.5E-4
□	GOTERM_MF_ALL	tetrapyrrole binding	RT	▬	6	1.3E-3
□	GOTERM_MF_ALL	heme binding	RT	▬	6	1.3E-3
Annotation Cluster 2		Enrichment Score: 3.52				
□	SP_PIR_KEYWORDS	antibiotic	RT	▬	5	2.2E-4
□	SP_PIR_KEYWORDS	antimicrobial	RT	▬	5	2.4E-4
□	GOTERM_BP_ALL	defense response to bacteria	RT	▬	6	5.4E-4
Annotation Cluster 3		Enrichment Score: 2.66				
□	UP_SEQ_FEATURE	domain:Ig-like C2-type 1	RT	▬	8	5.4E-4
□	UP_SEQ_FEATURE	domain:Ig-like C2-type 2	RT	▬	8	5.4E-4
□	INTERPRO_NAME	Immunoglobulin	RT	▬	6	3.6E-2
Annotation Cluster 4		Enrichment Score: 2.63				

Gene list being analyzed

Clustering options and stringency

The overall enrichment score for the group based on the EASE scores of each term members. The higher, the more enriched.

Every term in the annotation cluster

ALL genes involved in this annotation cluster

Genes involved in individual term

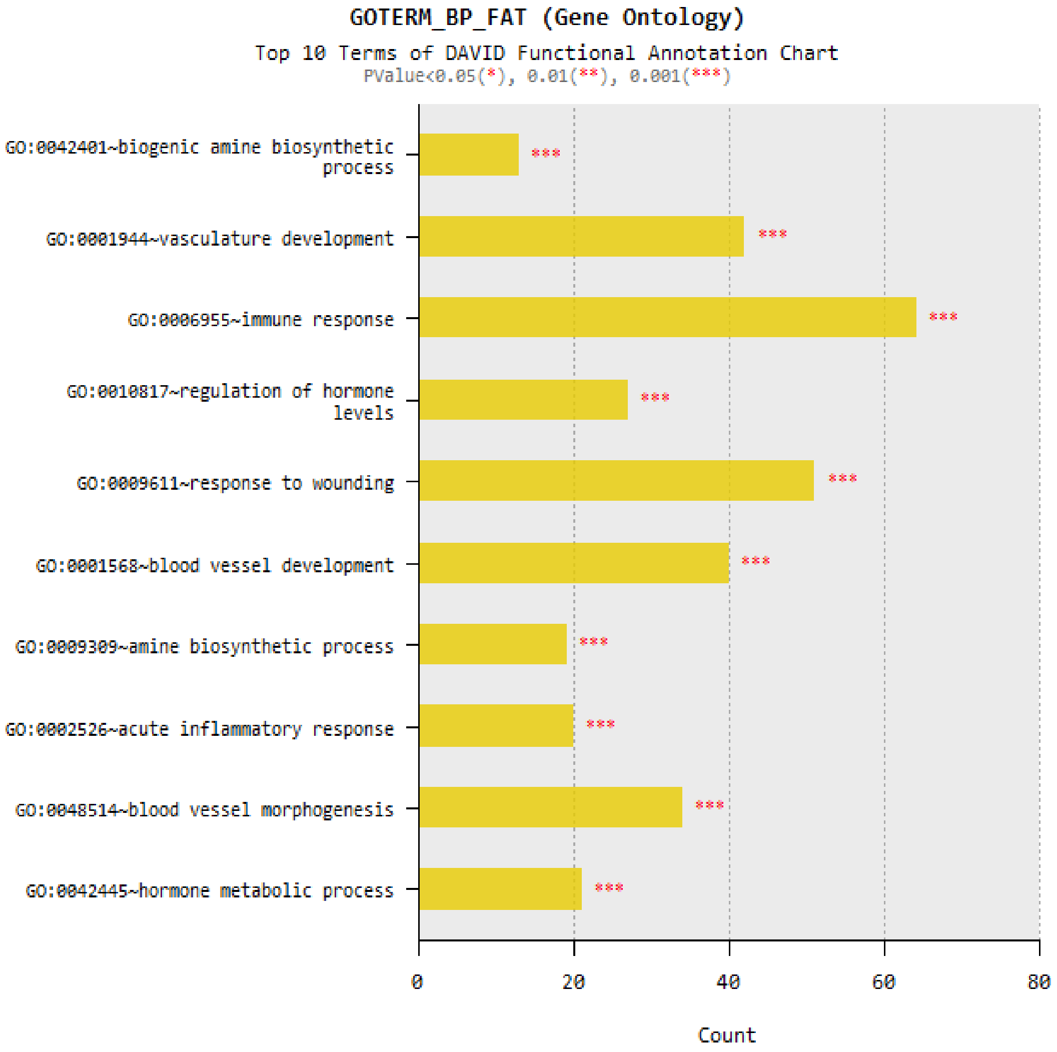
Related Term Search

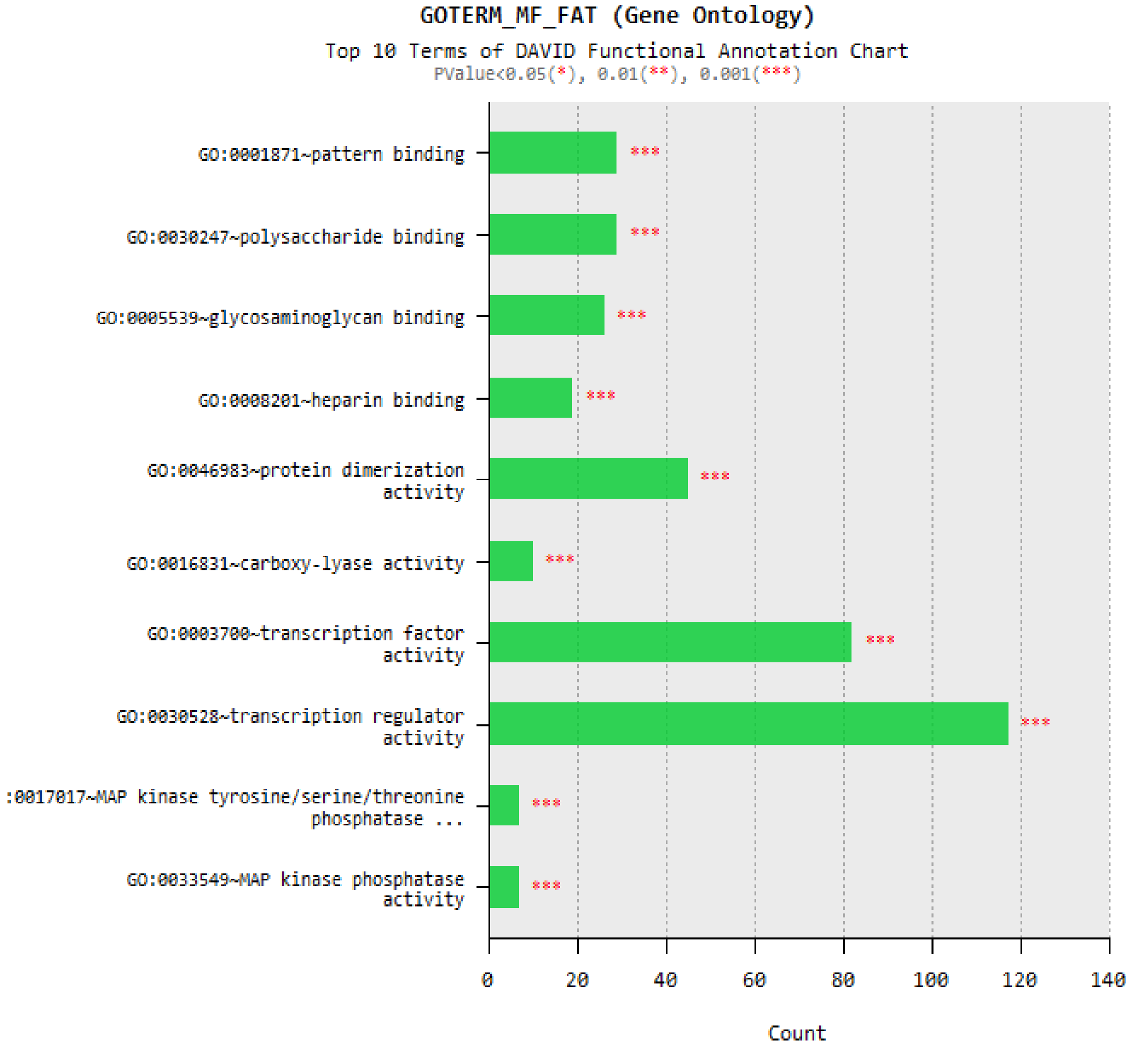
A group of terms having similar biological meaning due to sharing similar gene members

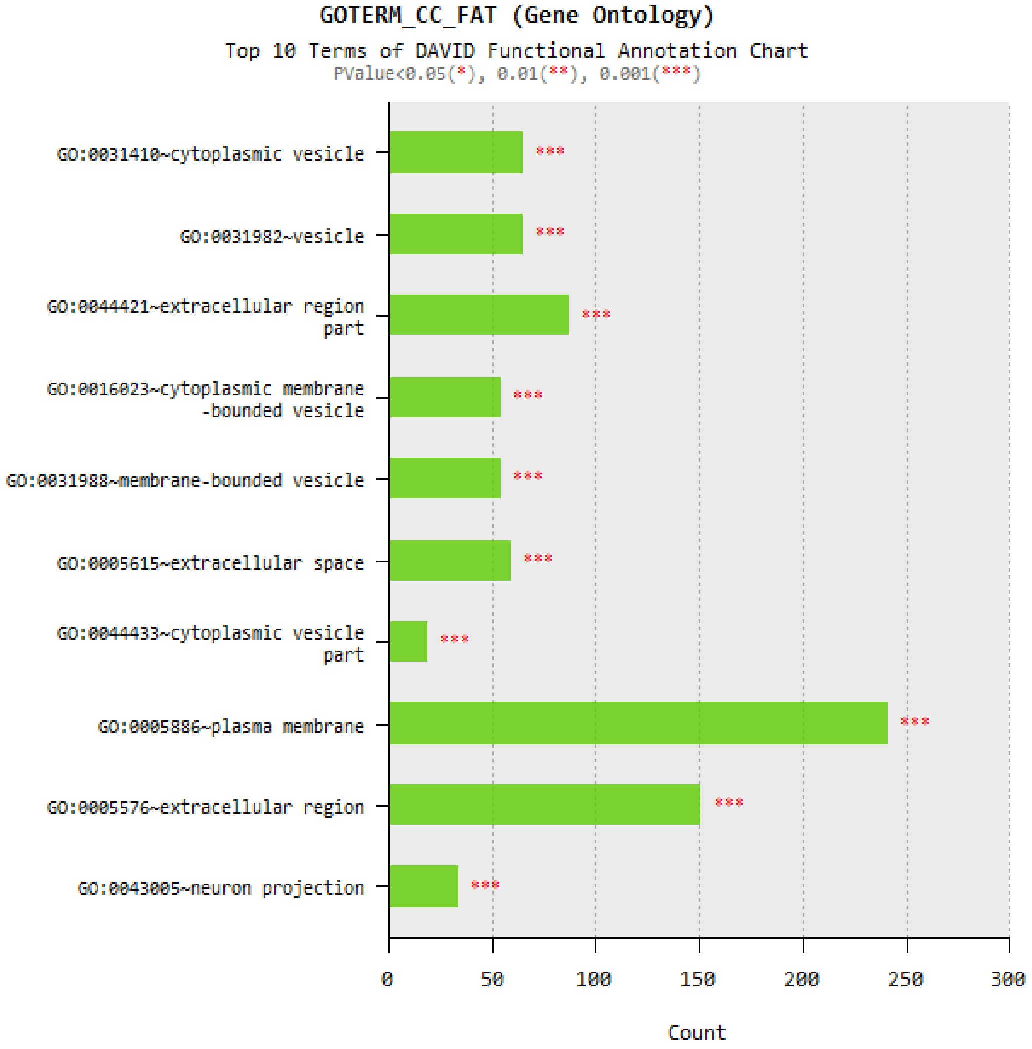
EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.

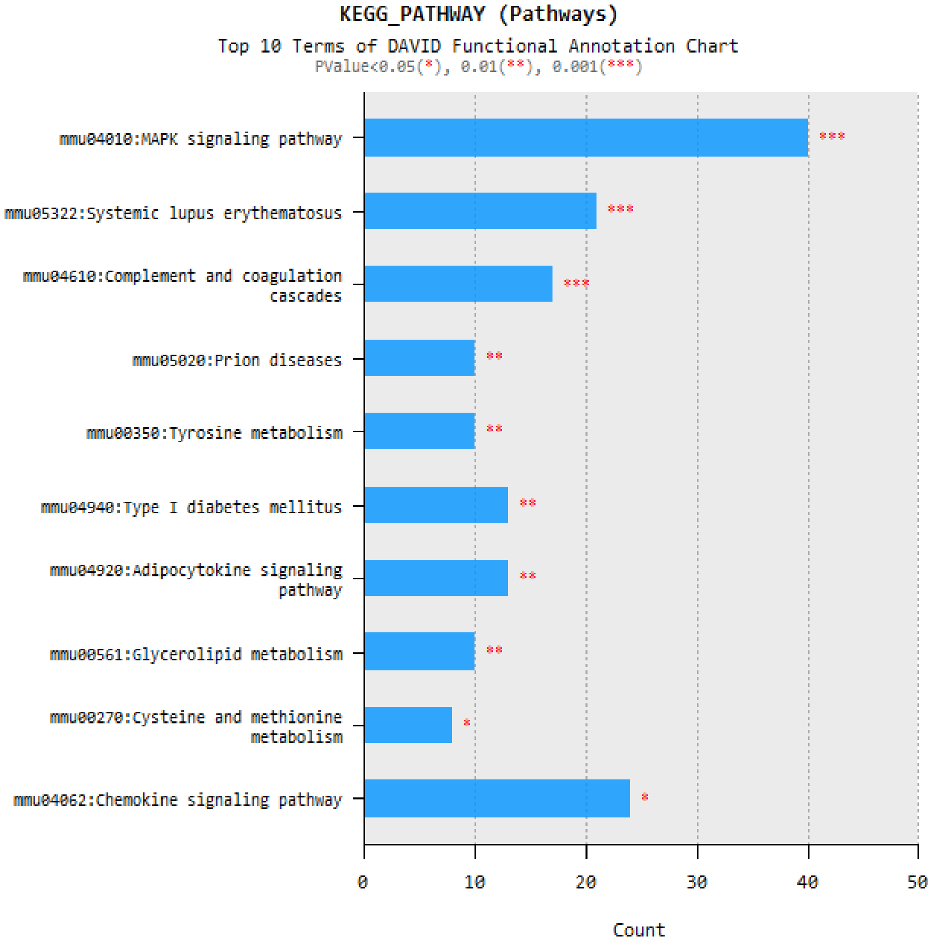
- Annotation cluster : Cluster of gene sets that have similar gene members and similar biological meanings.
- Enrichment Score : Refers to the enrichment score of each clusters. It is the －logP of average of EASE scores of each cluster’s gene-set term members. Higher value means that the cluster has been enriched.
- Category : Database which defines the gene set
- Term : Description of gene set
- Genes : List of gene that are included in the gene set term
- Percentage, % : Ratio of number of similar genes in the gene set term with the total number of genes
- P value : Also known as EASE score, the p-value from the Modified Fisher exact test to determine the enrichment of the gene from the gene set. If this value is lower than 0.05, it is classified as enrichment.
- Bonferroni, Benjamin, FDR : Due to multiple testing issue and to reduce the false positive value, p value corrected by (Bonferroni/ Benjamin/ FDR) method.

The bar plot below shows the results of the enrichment analysis through Gene Ontology, KEEG, and DAVID's functional annotation on the total of 1,555 similar transcripts
at least one comparison. (These plots were made based on functional annotation chart report.)




GOTERM_MF_FAT (Gene Ontology)
Top 10 Terms of DAVID Functional Annotation Chart
PValue<0.05(*), 0.01(**), 0.001(***)
GO:0001871~pattern binding
GO:0030247~polysaccharide binding
GO:0005539~glycosaminoglycan binding
GO:0008201~heparin binding
GO:0046983~protein dimerization activity
GO:0016831~carboxy-lyase activity
GO:0003700~transcription factor activity
GO:0030528~transcription regulator activity
:0017017~MAP kinase tyrosine/serine/threonine phosphatase ...
GO:0033549~MAP kinase phosphatase activity
***
0
20
40
60
80
100
120
140
Count


GOTERM_CC_FAT (Gene Ontology)
Top 10 Terms of DAVID Functional Annotation Chart
PValue<0.05(*), 0.01(**), 0.001(***)
GO:0031410~cytoplasmic vesicle
GO:0031982~vesicle
GO:0044421~extracellular region part
GO:0016023~cytoplasmic membrane -bounded vesicle
GO:0031988~membrane-bounded vesicle
GO:0005615~extracellular space
GO:0044433~cytoplasmic vesicle part
GO:0005886~plasma membrane
GO:0005576~extracellular region
GO:0043005~neuron projection
***
0
50
100
150
200
250
300
Count


KEGG_PATHWAY (Pathways)
Top 10 Terms of DAVID Functional Annotation Chart
PValue<0.05(*), 0.01(**), 0.001(***)
mmu04010:MAPK signaling pathway
***
mmu05322:Systemic lupus erythematosus
***
mmu04610:Complement and coagulation cascades
***
mmu05020:Prion diseases
**
mmu00350:Tyrosine metabolism
**
mmu04940:Type I diabetes mellitus
**
mmu04920:Adipocytokine signaling pathway
**
mmu00561:Glycerolipid metabolism
**
mmu00270:Cysteine and methionine metabolism
*
mmu04062:Chemokine signaling pathway
*
0
10
20
30
40
50
Count

6. SNP and Indel Discovery

(Refer to Path: 3.SNV_calling_result 〉 SNV_Call_*.xlsx)

SNV calling was performed on each sample, and the variant annotation based on the refGene Database, was performed as well.

For SNV calling, STAR program was used. This process maps the cDNA sequences reads to the genomic DNA reference. The reads that are obtains are processed for SAMTOOLS and BCFTOOLS for variant calling.

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

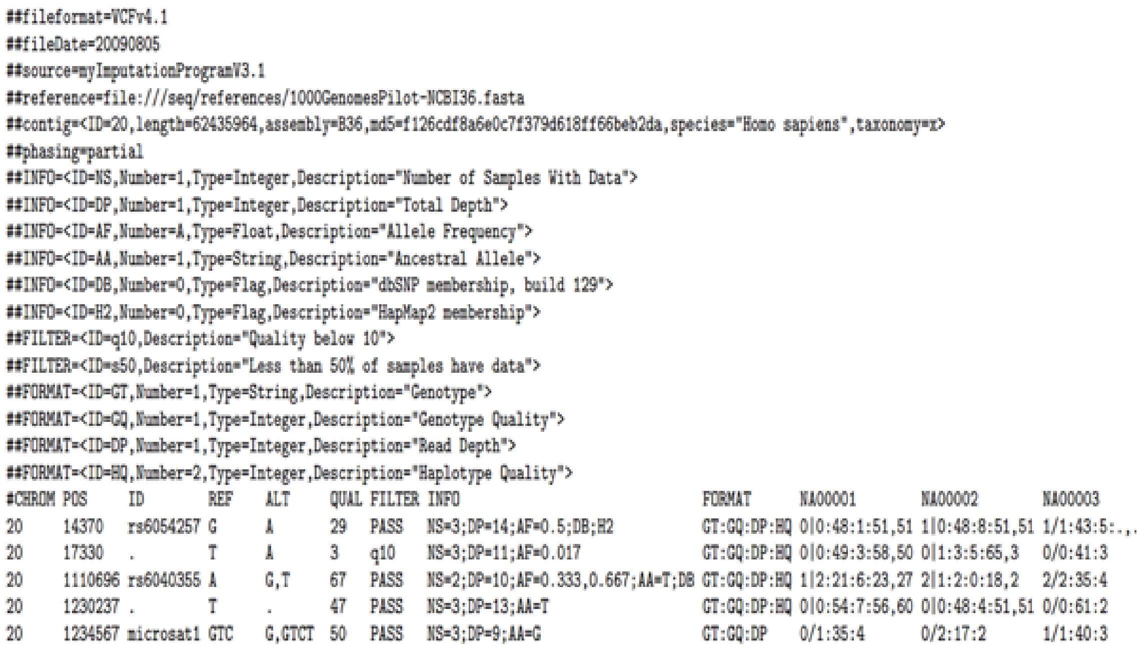
Below summarizes the results for 12 samples' SNV analysis.

Table 7. Summary of SNV Frequencies

Sample_ID	Number of SNPs	Number of coding SNPs	Number of indels	Number of coding indels	Ratio of hom variants (hom/(hom+het))
CRH-TG1	40,742	2,108	8,407	458	23.90%
CRH-TG2	38,237	2,079	7,713	414	23.35%
CRH-WT2	55,143	3,329	12,989	676	21.30%
CRH-WT5	47,082	2,272	11,216	638	22.24%
AG-WT-con1	41,889	1,898	9,766	538	22.04%
AG-WT-con2	41,629	2,052	9,578	550	23.77%
AG-PDK4KO-con1	59,998	2,905	11,406	501	23.27%
AG-PDK4KO-con2	54,610	2,878	10,051	498	22.49%
AG-WT-ACTH1-1h	52,231	2,971	9,998	568	23.61%
AG-WT-ACTH3-1h	48,966	3,142	8,570	489	24.33%
AG-PDK4KO-ACTH1-1h	50,664	3,074	8,941	501	24.39%
AG-PDK4KO-ACTH3-1h	60,292	3,091	10,964	595	22.54%

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

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

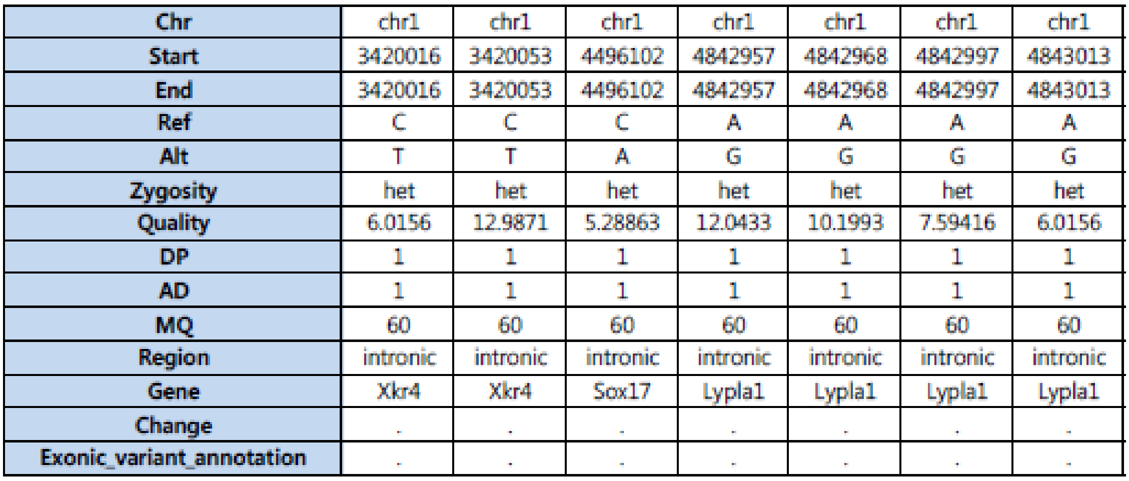


```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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
- POS : Reference position (1 based)
- 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 (q10: Quality <10, s50: less than 50% of samples are called, filter out). If it does not satisfies the filter condition, it will show the condition that hat it did not pass.
- 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.

The discovered SNV results are not only saved as vcf but along with refGene data information as excel file.

Table8. An example of annotation of individually discovered SNV



Chr	chr1	chr1	chr1	chr1	chr1	chr1	chr1
Start	3420016	3420053	4496102	4842957	4842968	4842997	4843013
End	3420016	3420053	4496102	4842957	4842968	4842997	4843013
Ref	C	C	C	A	A	A	A
Alt	T	T	A	G	G	G	G
Zygosity	het	het	het	het	het	het	het
Quality	6.0156	12.9871	5.28863	12.0433	10.1993	7.59416	6.0156
DP	1	1	1	1	1	1	1
AD	1	1	1	1	1	1	1
MQ	60	60	60	60	60	60	60
Region	intronic	intronic	intronic	intronic	intronic	intronic	intronic
Gene	Xkr4	Xkr4	Sox17	Lypla1	Lypla1	Lypla1	Lypla1
Change	.	.	.	.	.	.	.
Exonic_variant_annotation	.	.	.	.	.	.	.

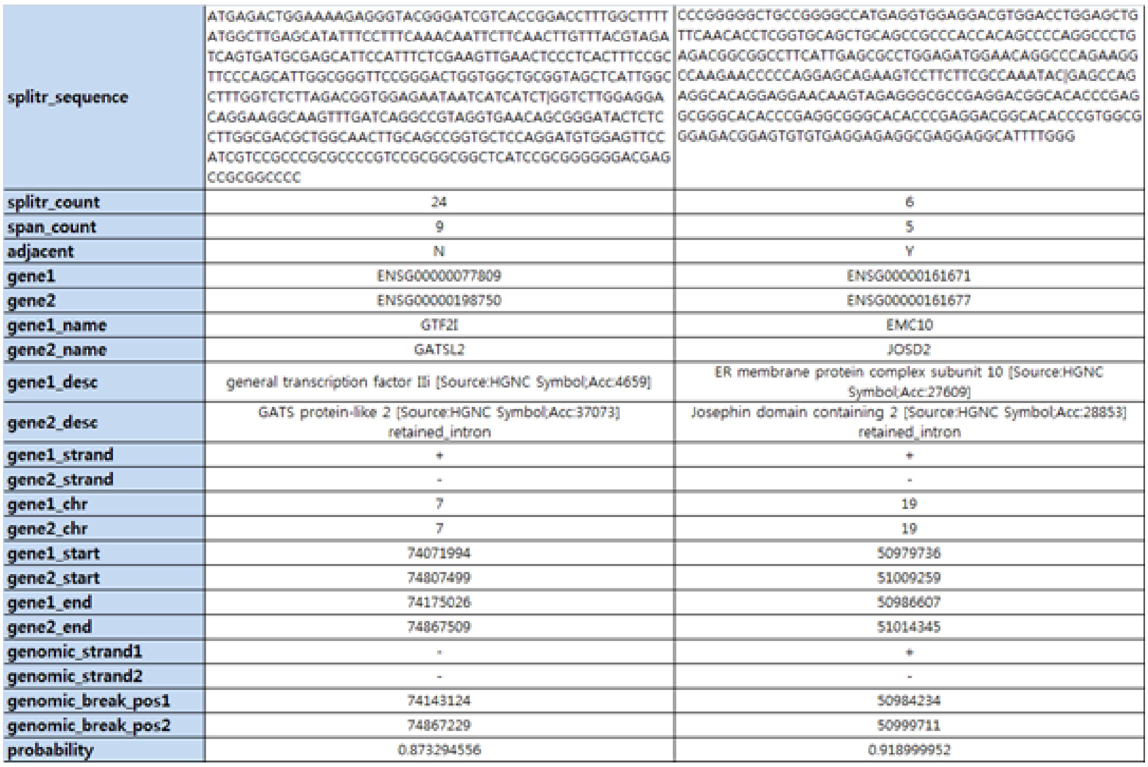
- Chr : chromosome
- Start, End : SNV position information
- Ref : Reference sequence regarding specific position
- Alt : Non-reference sequence
- Zygosity : Shows genotype, "hom" means non-reference homozygote, “het” means heterozygote
- Quality : Genotype quality.
- DP : Position’s read depth
- AD : Position’s alt read depth
- MQ : Mapping quality.
- Region : Functional region (exonic, intronic, 5’UTR, 3’UTR etc.)
- Gene : Gene symbol
- Change : If amino acid change exists, marked as nonsynoymous_SNV, if amino acid change does not exist, marked as synonymous_SNV.
- Exonic_variant_annotation : If amino acid change exists, detailed position information is shown. For example, if position is A2M:NM_000014:exon16:c.1915A>G:p.N639D, A2M gene, mRNA sequence of NM_000014, 16th exon’s 1915th position’s A changed to G, so protein change of 63th position’s N to D occurred.

7. Fusion Gene Prediction Results

(Refer to Path: 4.Fusion_gene_result)

Defuse program was used to predict fusion genes. Defuse predicts fusion genes region by clustering non concordant paired-end alignments (both spanning reads and split reads) and determines the possibility of real fusion gene through heuristic filter.

Table 9. Example of Fusion Gene Prediction Results

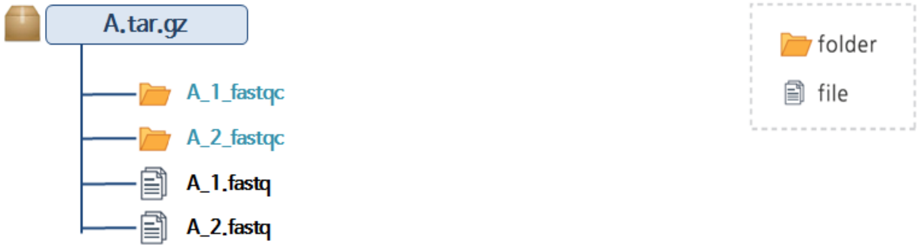


splitr_sequence	ATGAGACTGGAAAAGAGGGTACGGGATCGTCACCGGACCTTTGGCTTTT ATGGCTTGAGCATATTTCCTTTCAAACAATTCTTCAACTTGTTTACGTAGA TCAGTGATGCGAGCATTCCATTTCTCGAAGTTGAACTCCCTCACTTTCCGC TTCCCAGCATTGGCGGGTTCCGGGACTGGTGGCTGCGGTAGCTCATTGGC CTTTGGTCTCTTAGACGGTGGAGAATAATCATCATCT|GGTCTTGGAGGA CAGGAAGGCAAGTTTGATCAGGCCGTAGGTGAACAGCGGGATACTCTC CTTGGCGACGCTGGCAACTTGCAGCCGGTGCTCCAGGATGTGGAGTTCC ATCGTCCGCCCGCGCCCCGTCCGCGGCGGCTCATCCGCGGGGGGACGAG CCGCGGCCCC	CCCGGGGGCTGCCGGGGCCATGAGGTGGAGGACGTGGACCTGGAGCTG TTCAACACCTCGGTGCAGCTGCAGCCGCCCACCACAGCCCCAGGCCCTG AGACGGCGGCCTTCATTGAGCGCCTGGAGATGGAACAGGCCCAGAAGG CCAAGAACCCCCAGGAGCAGAAGTCCTTCTTCGCCAAATAC|GAGCCAG AGGCACAGGAGGAACAAGTAGAGGGCGCCGAGGACGGCACACCCGAG GCGGGCACACCCGAGGCGGGCACACCCGAGGACGGCACACCCGTGGCG GGAGACGGAGTGTGTGAGGAGAGGCGAGGAGGCATTTTGGG
splitr_count	24	6
span_count	9	5
adjacent	N	Y
gene1	ENSG00000077809	ENSG00000161671
gene2	ENSG00000198750	ENSG00000161677
gene1_name	GTF2I	EMC10
gene2_name	GATSL2	JOSD2
gene1_desc	general transcription factor IIi [Source:HGNC Symbol;Acc:4659]	ER membrane protein complex subunit 10 [Source:HGNC Symbol;Acc:27609]
gene2_desc	GATS protein-like 2 [Source:HGNC Symbol;Acc:37073] retained_intron	Josephin domain containing 2 [Source:HGNC Symbol;Acc:28853] retained_intron
gene1_strand	+	+
gene2_strand	-	-
gene1_chr	7	19
gene2_chr	7	19
gene1_start	74071994	50979736
gene2_start	74807499	51009259
gene1_end	74175026	50986607
gene2_end	74867509	51014345
genomic_strand1	-	+
genomic_strand2	-	-
genomic_break_pos1	74143124	50984234
genomic_break_pos2	74867229	50999711
probability	0.873294556	0.918999952

- split_sequence : Shows fusion sequences. The two sequences of the donor and acceptor are shown in separate columns.
- 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_desc, gene2_desc : 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_pos1, genomic_break_pos2 : 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.

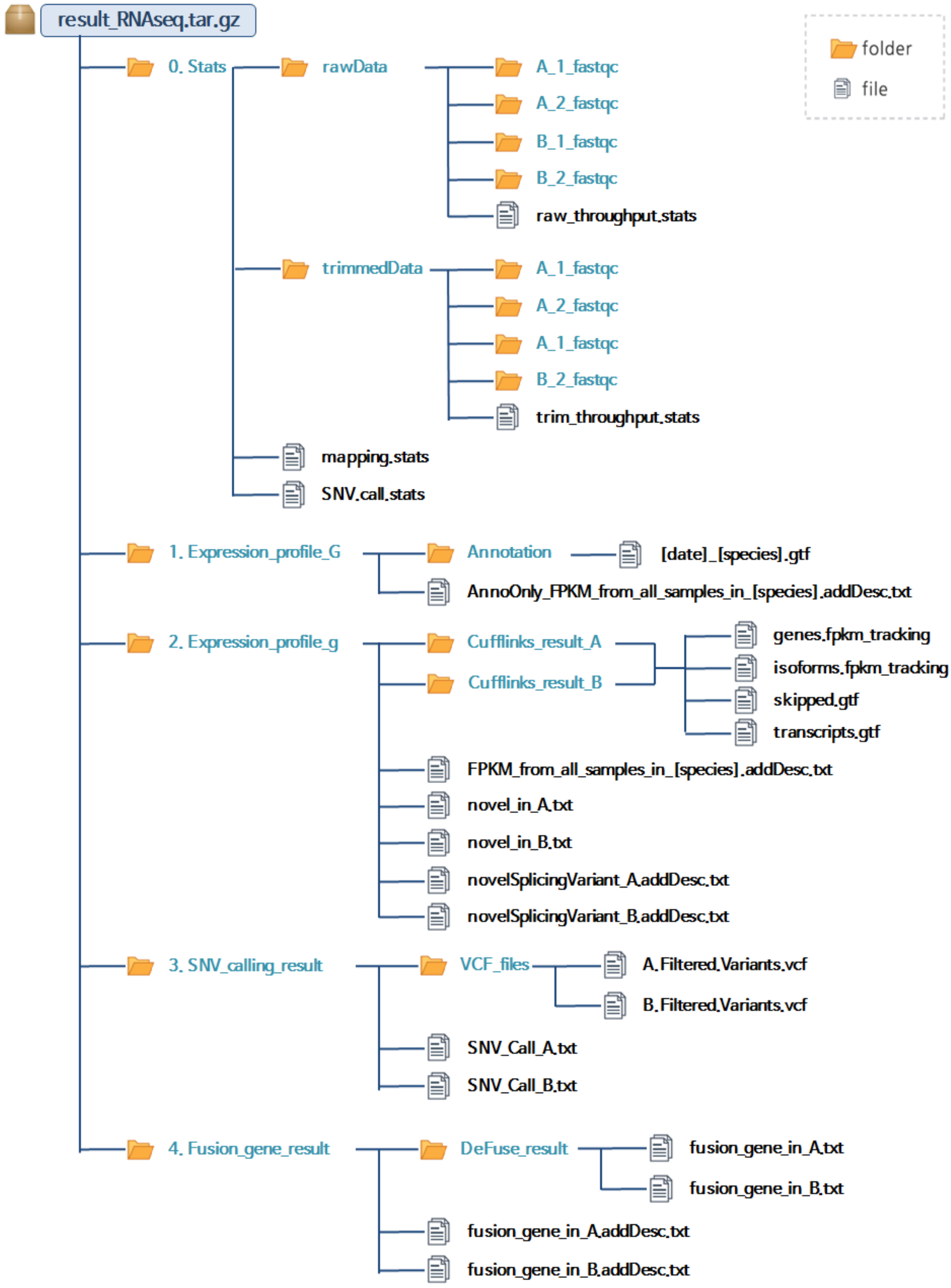
8. Data Download Information

8. 1. Raw Data



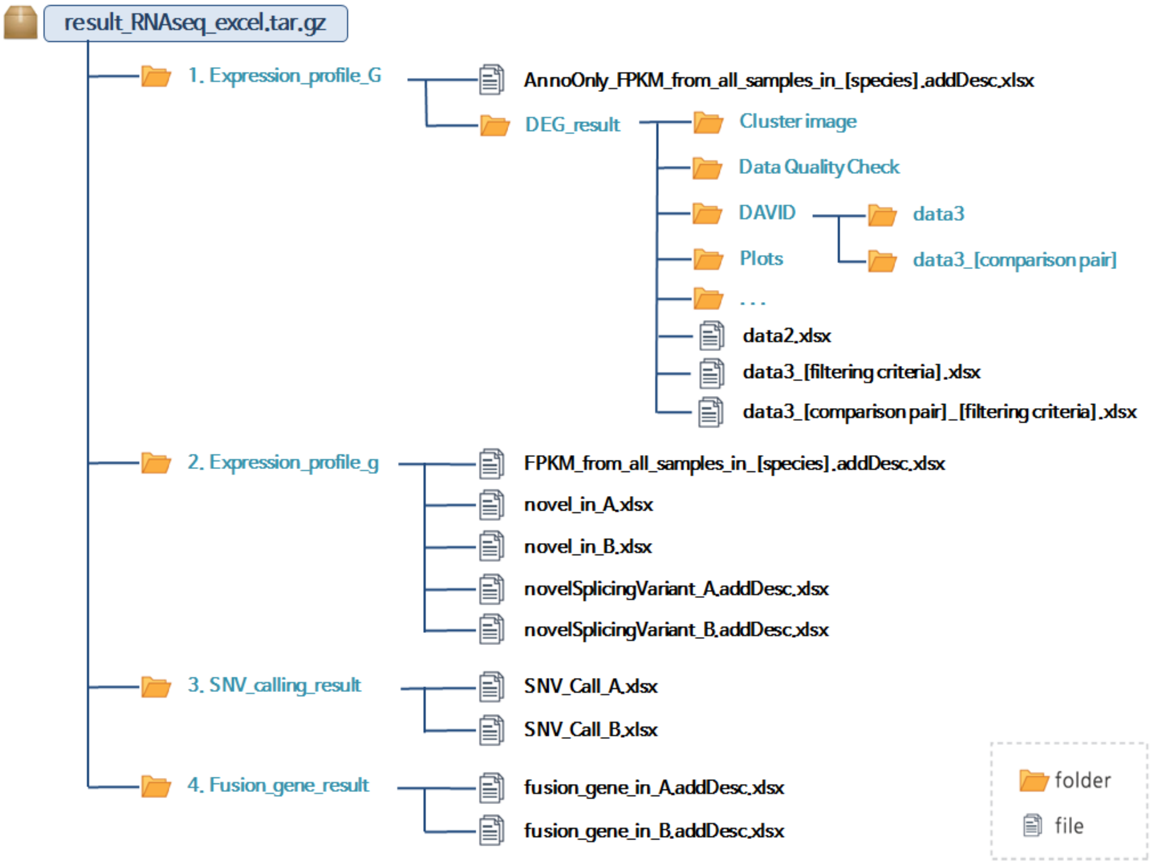
Index	Sample ID	Link
1	CRH-TG1	Download
2	CRH-TG2	Download
3	CRH-WT2	Download
4	CRH-WT5	Download
5	AG-WT-con1	Download
6	AG-WT-con2	Download
7	AG-PDK4KO-con1	Download
8	AG-PDK4KO-con2	Download
9	AG-WT-ACTH1-1h	Download
10	AG-WT-ACTH3-1h	Download
11	AG-PDK4KO-ACTH1-1h	Download
12	AG-PDK4KO-ACTH3-1h	Download

8. 2. Analysis Results



```
result_RNAseq.tar.gz
├── 0. Stats
│   ├── rawData
│   │   ├── A_1_fastqc
│   │   ├── A_2_fastqc
│   │   ├── B_1_fastqc
│   │   ├── B_2_fastqc
│   │   └── raw_throughput.stats
│   ├── trimmedData
│   │   ├── A_1_fastqc
│   │   ├── A_2_fastqc
│   │   ├── A_1_fastqc
│   │   ├── B_2_fastqc
│   │   └── trim_throughput.stats
│   ├── mapping.stats
│   └── SNV.call.stats
├── 1. Expression_profile_G
│   ├── Annotation
│   │   └── [date]_[species].gtf
│   └── AnnoOnly_FPKM_from_all_samples_in_[species].addDesc.txt
├── 2. Expression_profile_g
│   ├── Cufflinks_result_A
│   ├── Cufflinks_result_B
│   │   ├── genes.fpkm_tracking
│   │   ├── isoforms.fpkm_tracking
│   │   ├── skipped.gtf
│   │   └── transcripts.gtf
│   ├── FPKM_from_all_samples_in_[species].addDesc.txt
│   ├── novel_in_A.txt
│   ├── novel_in_B.txt
│   ├── novelSplicingVariant_A.addDesc.txt
│   └── novelSplicingVariant_B.addDesc.txt
├── 3. SNV_calling_result
│   ├── VCF_files
│   │   ├── A.Filtered.Variants.vcf
│   │   └── B.Filtered.Variants.vcf
│   ├── SNV_Call_A.txt
│   └── SNV_Call_B.txt
└── 4. Fusion_gene_result
    ├── DeFuse_result
    │   ├── fusion_gene_in_A.txt
    │   └── fusion_gene_in_B.txt
    ├── fusion_gene_in_A.addDesc.txt
    └── fusion_gene_in_B.addDesc.txt
```

LINK result_RNAseq.tar.gz : Download


result_RNAseq_excel.tar.gz
1. Expression_profile_G
AnnoOnly_FPKM_from_all_samples_in_[species].addDesc.xlsx
DEG_result
Cluster image
Data Quality Check
DAVID
data3
data3_[comparison pair]
Plots
. . .
data2.xlsx
data3_[filtering criteria].xlsx
data3_[comparison pair]_[filtering criteria].xlsx
2. Expression_profile_g
FPKM_from_all_samples_in_[species].addDesc.xlsx
novel_in_A.xlsx
novel_in_B.xlsx
novelSplicingVariant_A.addDesc.xlsx
novelSplicingVariant_B.addDesc.xlsx
3. SNV_calling_result
SNV_Call_A.xlsx
SNV_Call_B.xlsx
4. Fusion_gene_result
fusion_gene_in_A.addDesc.xlsx
fusion_gene_in_B.addDesc.xlsx
folder
file

LINK result_RNAseq_excel.tar.gz : Download

⚠ The data retention period is three months, please contact a representative e-mail (ngskr@macrogen.com) or representative if you need long-term storage.

9. Appendix

9. 1. Phred Quality Score Chart

Phred quality score numerically express the accuracy of each nucleotide. Higher Q number signifies higher accuracy. Q20 means the probability of wrong base is 1% and Q30 is probability of wrong base as 0.1%. Below is the Phred Quality Score chart.

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 by $-10\log_{10}P$, where P is probability of erroneous base call.

9. 2. Programs used in Analysis

9. 2. 1. FastQC v0.10.0

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 of 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 0.32

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 : Removes adapter and specific sequences from the reads
- SLIDINGWINDOW : Performs sliding window trimming. If quality is lower than the threshold within the window, the reads are trimmed.
- LEADING : If quality is lower than threshold, reads at the start are removed.
- TRAILING : If quality is lower than the threshold, reads at the ends are removed.
- CROP : Reads are removed at specific lengths.
- HEADCROP : Trim specific number of bases from the start.
- MINLEN: Drop reads under specific lengths.
- TOPHRED33 : Change quality score to phred-33.
- TOPHRED64 : Change quality score to phred-64.

9. 2. 3. TopHat version 2.0.13, bowtie2 2.2.3

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

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

9. 2. 4. Cufflinks version 2.2.1

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

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

9. 2. 5. deFuse 0.6.2

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 non-concordant 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. 3. References

1. BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, btu170.

2. TRAPNELL, Cole; PACHTER, Lior; SALZBERG, Steven L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25.9: 1105-1111.

3. KIM, Daehwan, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 2013, 14.4: R36.

4. LANGMEAD, Ben, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10.3: R25.

5. TRAPNELL, Cole, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 2010, 28.5: 511-515.

6. ROBERTS, Adam, et al. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology, 2011, 12.3: R22.

7. BI, Yong-Mei, et al. High throughput RNA sequencing of a hybrid maize and its parents shows different mechanisms responsive to nitrogen limitation. BMC genomics, 2014, 15.1: 77.

8. TRAPNELL, Cole, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology, 2013, 31.1: 46-53.

9. TRAPNELL, Cole, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 2012, 7.3: 562-578.

10. 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.

11. LI, Heng, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16: 2078-2079.

12. MCPHERSON, Andrew, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS computational biology, 2011, 7.5: e1001138.

Humanizing Genomics

macrogen

Thank you

Macrogen Korea

10F, 254 Beotkkot-ro, Geumcheon-gu, Seoul
Rep. of Korea
Phone : +82-2-2113-7100

Contact

Web : www.macrogen.com
Lims : http://dna.macrogen.com

FOR RESEARCH USE ONLY