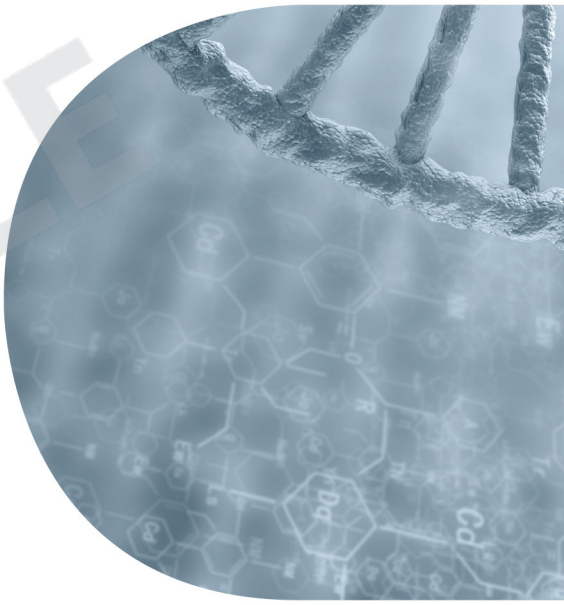
Research use only

Homo sapiens Transcriptome Sequencing

Report



ver. MGTR4.0_TC_GRCh38

Humanizing Genomics
macrogen

프로젝트 정보

고객명	TESTER
소속	Macrogen
수주번호	HN00000000
종 (학명)	*Homo sapiens*
레퍼런스	GRCh38
Annotation	NCBI_109.20200522
리드 종류	Paired-ends
리드 길이	101
샘플 수	6
Library Kit	TruSeq stranded mRNA
Type of Sequencer	Illumina platform

프로젝트 수행 결과 요약

본 연구에서는 *Homo sapiens* 의 전사체 서열분석(transcriptome sequencing)을 통해 유전자 발현값을 얻어 차별 발현 유전자를 분석하고, 유의한 유전자에 대하여 gene ontology 및 pathway 정보를 기반으로 기능분류 및 gene annotation을 진행하였습니다. 더불어 전사체 어셈블리(assembly) 과정에서 발견한 novel transcript 및 novel alternative splicing transcript를 정리하였습니다.

그 외 샘플 별 변이 발굴(SNV calling)과 variant annotation 및 융합유전자(fusion gene) 탐색 작업도 진행하였습니다.

의뢰하신 총 6 샘플에 대한 paired-ends 전사체 서열분석 결과, 모든 샘플이 정상적인 범위 내에 결과가 생산되었습니다. 아래 그림은 샘플 별 raw data와 전처리 과정을 거친 trimmed read를 총 데이터 량과 Q30(phred score, base quality 30이상) 값으로 각각을 비교한 데이터 입니다. (그림 1, 그림 2 참고. 전체 데이터의 샘플 별 구체적인 수치는 본문 참고)

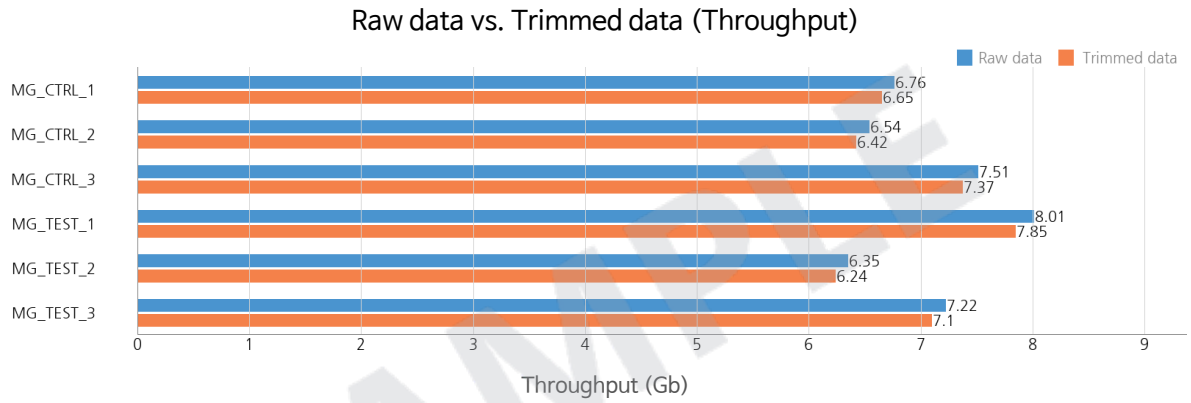


그림 1. Raw data와 Trimmed data의 데이터량 비교

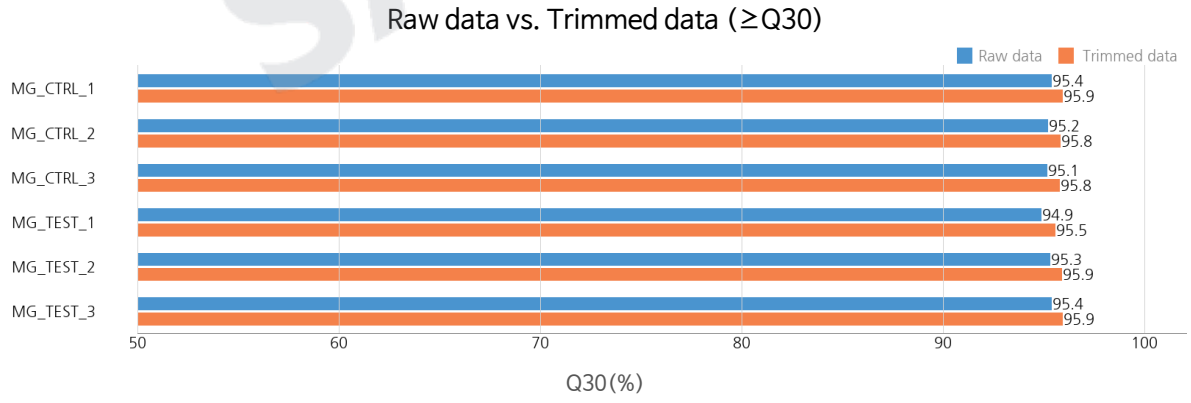


그림 2. Raw data와 Trimmed data의 Q30값 비교

전처리 과정을 거친 trimmed reads를 TopHat 프로그램을 이용하여 알려진 reference genome에 mapping합니다. 그림 3에서 각 샘플 별 trimmed reads 수 대비 mapped reads로 정의되는 mapping ratio를 확인하실 수 있습니다.

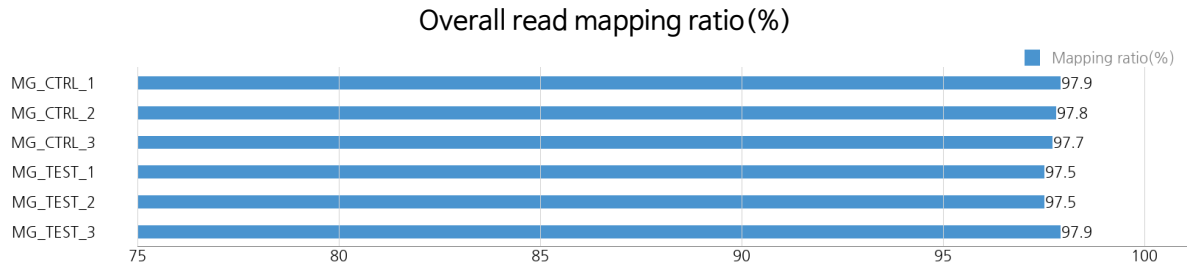


그림 3. Overall read mapping ratio (%)

Read mapping 후 Cufflinks 프로그램을 통해 transcript assembly 작업을 진행하였습니다. 그 결과로 알려진 transcript에 대해 각 샘플별로 expression profile 값을 얻었고, transcript/gene을 기준으로 read count, FPKM(Fragment per Kilobase of transcript per Million mapped reads), TPM(Transcripts Per Kilobase Million)값을 정리하였습니다.

이 값을 이용하여 요청하신 비교조합 (TEST_vs_CTRL)에 대해 DESeq2을 이용하여 DEG(Differentially Expressed Genes) 분석을 진행하였고, 두 그룹간 차별 발현하는 유전자를 선별하기 위하여 |fc|〉=2 & nbinomWaldTest raw p-value〈0.05 조건을 만족하는 gene 2,886개를 추출하였습니다.

그림 4는 유의한 gene을 대상으로 hierarchical clustering (distance metric = Euclidean distance, linkage method = complete) 분석을 통해 각 샘플의 gene별 발현 패턴의 유사성 정도를 이용하여 샘플 그룹별, gene별 그룹화된 정보를 시각화하여 나타낸 그림입니다.

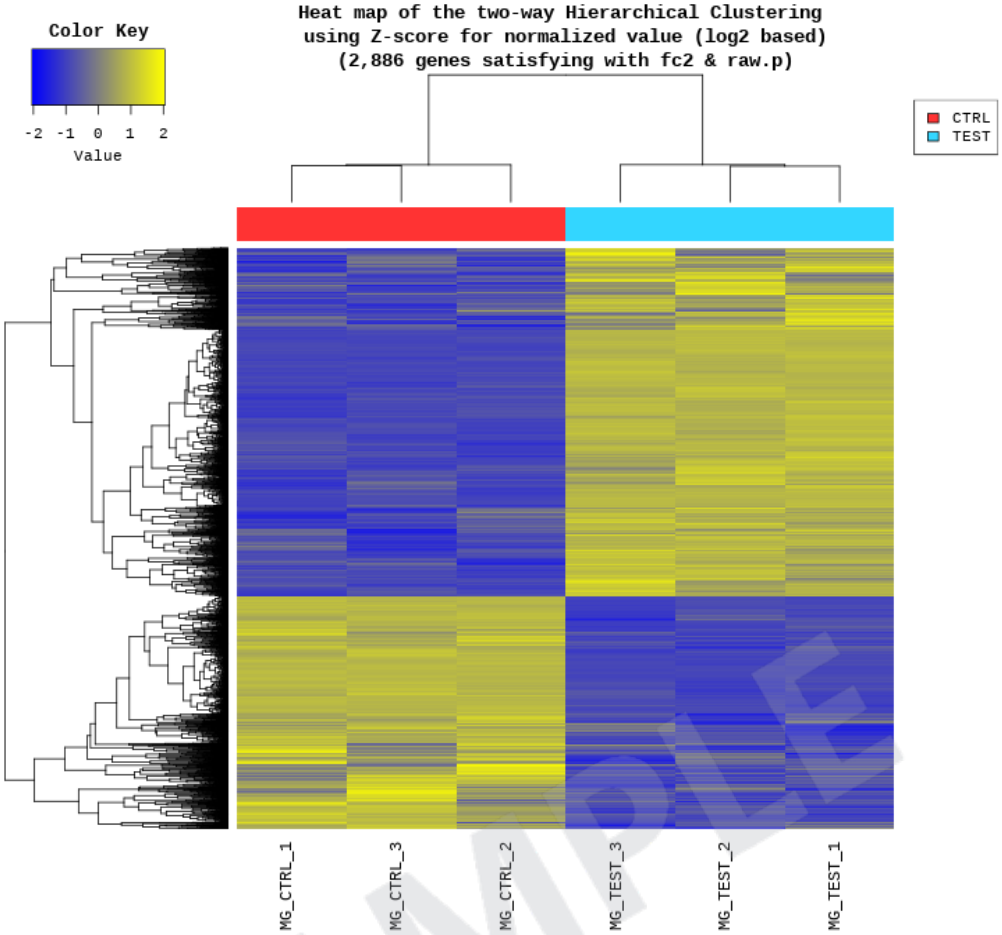


그림 4. DEG 리스트에 대한 Heatmap

유의한 DEG 리스트에 대해 gProfiler (https://biit.cs.ut.ee/gprofiler/orth)를 이용하여 gene ontology의 기능분류인 biological process(BP), molecular function(MF), cellular component(CC)별 gene set enrichment 분석을 진행하였습니다. 아래 그림 5, 그림 6, 그림 7은 각 카테고리별 유의한 gene set 리스트를 나타냅니다.

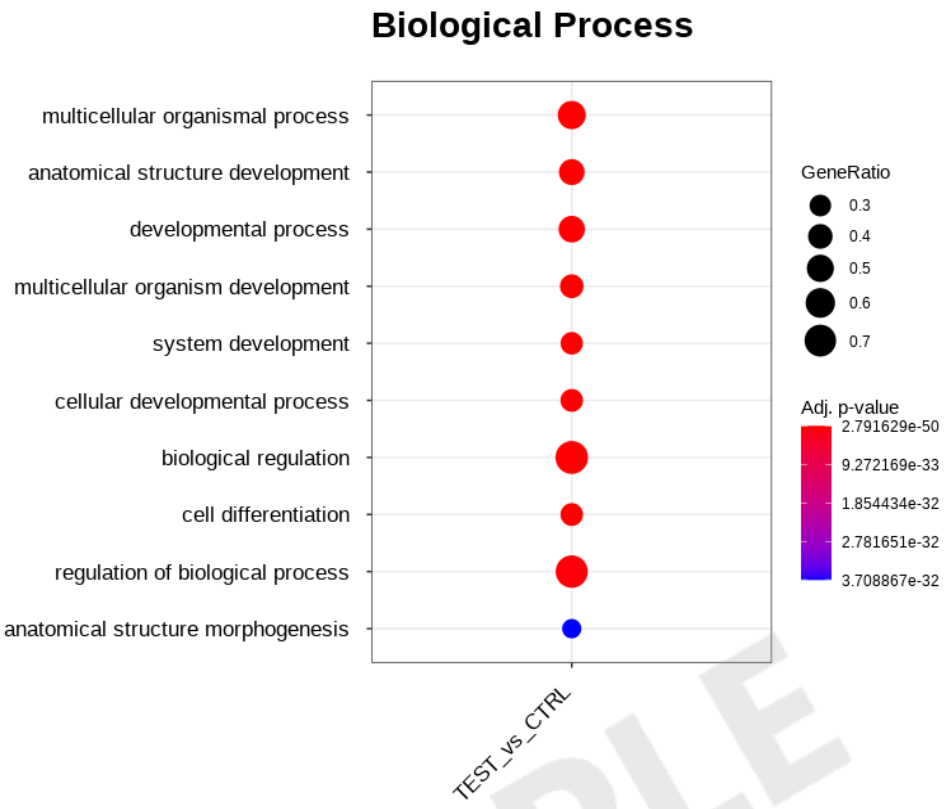


그림 5. Biological Process관련 GO term

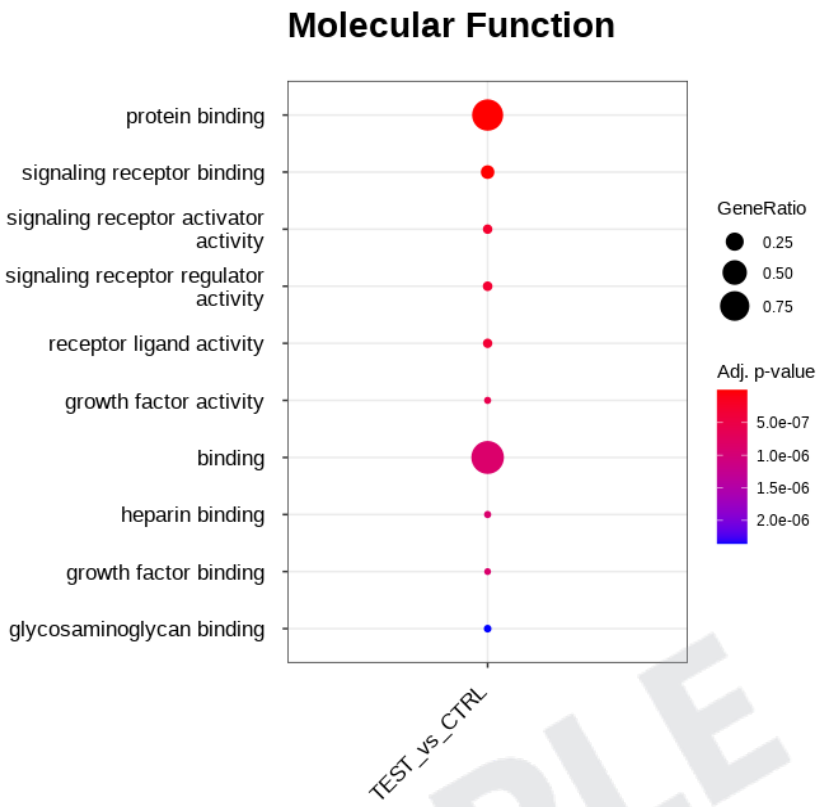


그림 6. Molecular Function관련 GO term

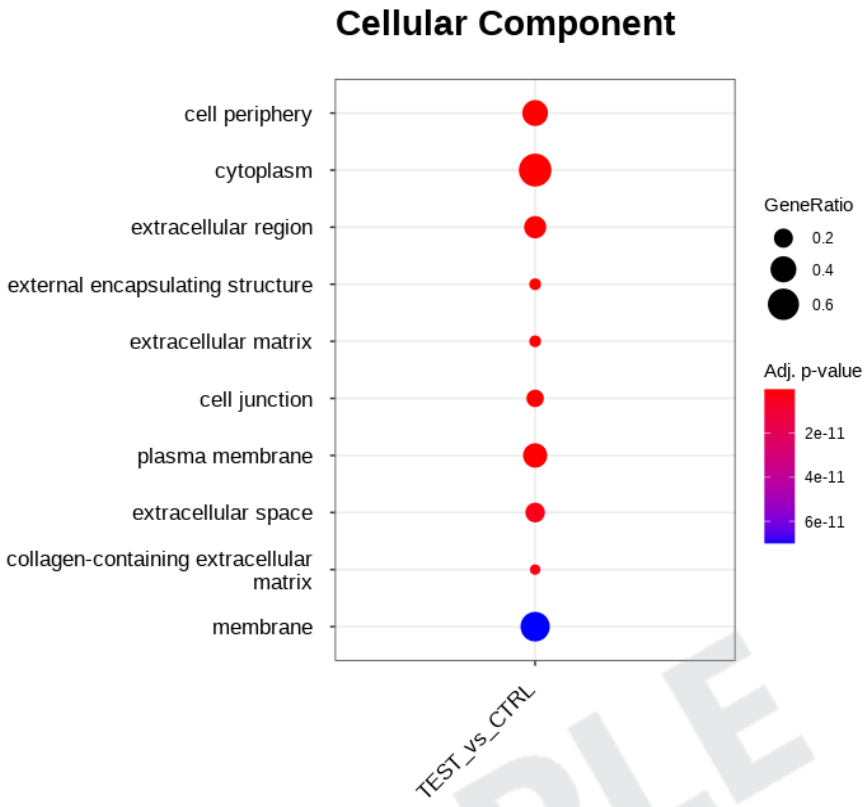


그림 7. Cellular Component관련 GO term

그 외 각 샘플별 novel transcript 및 novel alternative splicing transcript를 찾았으며, 변이 발굴(SNV calling) 및 주석 달기(variant annotation) 작업과 defuse, fusion catcher 프로그램을 통한 융합유전자(fusion gene)를 예측 작업을 진행하여 그 결과를 정리하였습니다. (세부 설명은 본문 참고)

목차

1. 실험 방법 및 순서도

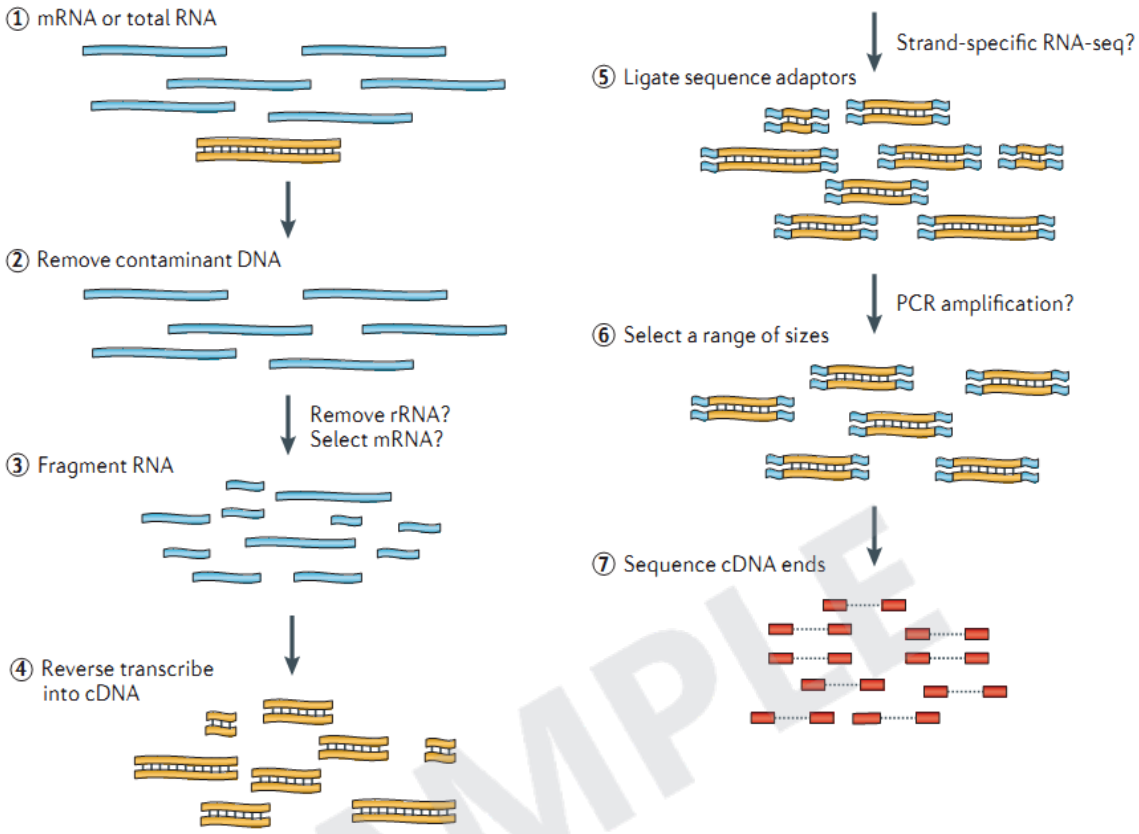


그림 8. RNA Sequencing Experiment Workflow

REFERENCE Nat Rev Genet. 2011 Sep 7;12(10):671-82

1) 샘플(세포 또는 조직)로 부터 Total RNA를 isolation 합니다.

2) DNase를 이용하여 DNA contamination을 제거합니다.

3) library제작 단계로 프로파일링 할 RNA 종류에 따라 kit를 선택 합니다. poly-A tail이 있는 mRNA를 대상으로 연구 할 경우 mRNA purification kit를, mRNA 뿐만 아니라 non-coding RNA(lincRNA 등)를 포함한 total RNA를 대상으로 연구 할 경우 ribo-zero rRNA removal kit를 이용하여 RNA를 정제합니다.

4) 정제된 RNA를 short read로 시퀀싱을 하기 위해 random하게 fragmentation 시킵니다.

5) 잘게 쪼개진 RNA fragment에 대해 reverse transcription 과정을 통해 cDNA로 만듭니다.

6) 만들어진 cDNA fragment 양 쪽 끝에 서로 다른 adapter를 붙이고 이를 ligation 합니다.

7) sequencing이 가능한 정도의 양으로 PCR 증폭 시킨 후 size selection과정을 통해 200-400 bp의 insert size 를 확보합니다. Paired-end sequencing일 경우, cDNA fragment의 양쪽 말단으로부터 read의 length만큼 sequencing이 됩니다.

2. 분석 방법 및 순서도

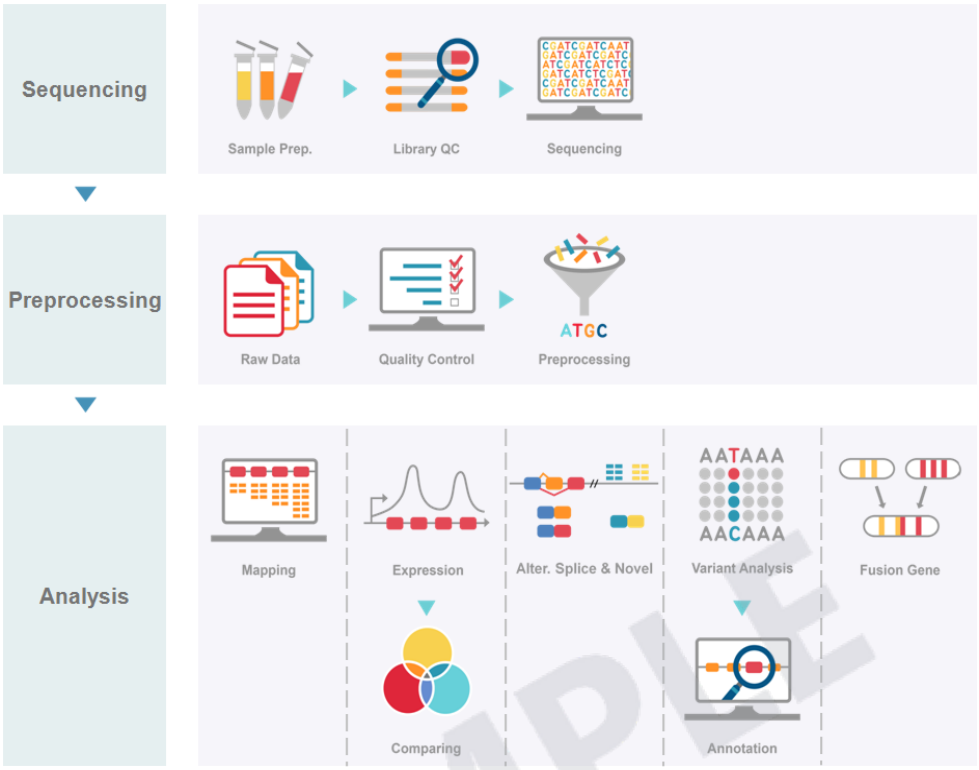


그림 9. Analysis Workflow

1) 시퀀싱을 통해 얻어진 raw reads의 quality control 분석을 진행합니다. 전체적인 read의 quality와 total bases, total reads, GC(%) 등 기본 통계치를 생산합니다.

2) 분석 결과의 bias를 줄이기 위해 low-quality를 가지거나 adaptor sequence, contaminant DNA, PCR duplicates와 같은 artifacts을 제거하는 전처리 과정을 거칩니다.

3) 전처리 과정을 거친 reads를 TopHat 프로그램을 이용하여 reference genome에 mapping한 후, aligned reads를 생성합니다.

4) reference 기반한 aligned reads의 정보를 이용하여 Cufflinks 프로그램을 통한 transcript 어셈블리를 진행합니다. 이 과정에서 알려진 transcripts 및 novel transcripts, alternative splicing transcripts에 대한 정보를 생산할 수 있습니다.

5) 각 샘플의 transcript quantification을 통해 얻은 발현량을 read count와 transcript length 및 depth of coverage를 고려한 normalization 값인 FPKM(Fragments Per Kilobase of transcript per Million mapped reads)/RPKM(Reads Per Kilobase of transcript per Million mapped reads)값과 TPM(Transcripts Per Kilobase Million)값으로 expression profile을 추출합니다.

6) 조건이 다른 두 그룹 이상의 발현값을 통계적인 가설검증을 통하여 차별 발현하는 유전자 또는 transcripts를 선별합니다.

7) 알려져 있는 유전자 정보가 있는 경우, 차별 발현 유전자를 대상으로 GO 및 KEGG database를 기반한 functional annotation 및 gene-set enrichment analysis를 진행합니다.

8) RNA-seq 데이터에서의 SNV calling은, STAR로 reads를 genomic DNA reference에 mapping한 후, Mark

duplication & sort 과정을 거칩니다. 그 이후 Split 'N' Trim 및 mapping quality reassignment, Indel realignment, base recalibration 과정을 통하여 분석 가능한 mapped reads를 만듭니다. 이렇게 data cleanup과정을 거친 reads들을 대상으로 haplotype caller를 이용하여 variant calling을 진행합니다.
LINK https://www.broadinstitute.org/gatk/guide/best-practices?bpm=RNAseq

9) Defuse, FusionCatcher, Arriba 프로그램을 사용하여 융합유전자(fusion gene)를 예측합니다.

3. 생산 데이터 요약

3. 1. Raw 데이터 기초 통계치

(경로: result_RNAseq/Analysis_statistics/raw_throughput.txt 참고)

의뢰하신 6 샘플에 대한 transcriptome raw data의 총 길이, read 수, GC(%), Q20(%), Q30(%) 입니다. 예를 들어, MG_CTRL_1 샘플은 66,947,992개 reads가 생성되었으며, 총 길이의 합은 6.8G bp로 생산되었습니다. GC content(%)가 47.97%였으며, 염기 품질 점수 30이상을 갖는 염기의 비율(Q30, %)은 95.36%로 나타났습니다.

표 1. Raw data 통계치

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 로 계산됨)

- Total read bases: 전체 시퀀싱 된 염기의 수
- Total reads: 전체 read의 수
- GC(%): GC 함량
- Q20(%): Phred quality score 20 이상의 품질을 갖는 염기의 비율
- Q30(%): Phred quality score 30 이상의 품질을 갖는 염기의 비율

3. 2. 사이클별 평균 염기품질

(경로: result_RNAseq/Analysis_statistics/rawData/A_fastqc/ 참고)

일반적으로 생산된 데이터의 품질은 각 염기의 프레드 품질 점수(phred quality score)로 판단됩니다. FastQC 를 사용하여 사이클별 평균 염기 품질을 박스 플롯으로 한눈에 볼 수 있습니다.

그래프의 x 축은 사이클을, y 축은 프레드 품질 점수를 보여줍니다. 프레드 품질 점수 20 은 99% 정확도를 나타내고, 일반적으로 20 이상의 프레드 품질 점수를 가지면 품질이 양호하다 할 수 있습니다.

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

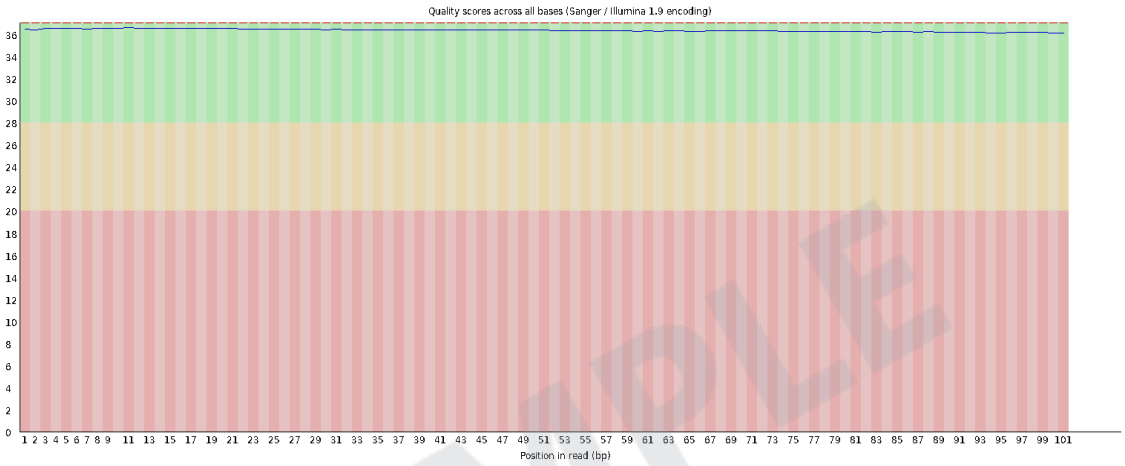


그림 10. MG_CTRL_1 (read1)의 사이클별 평균 염기 품질

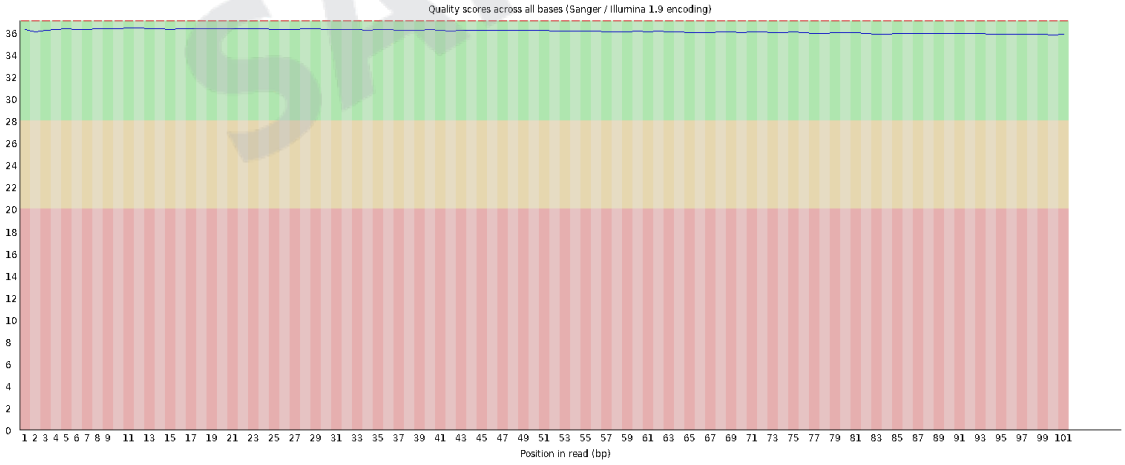


그림 11. MG_CTRL_1 (read2)의 사이클별 평균 염기 품질

- Yellow box: 사이클별 염기 품질 점수의 interquartile range (25-75%)를 나타냄.
- Red line: 사이클별 염기 품질 점수의 중앙 값
- Blue line: 사이클별 염기 품질 점수의 평균 값
- Green background: 우수한 품질을 나타냄.
- Orange background: 양호한 품질을 나타냄.
- Red background: 나쁜 품질을 나타냄.

3. 3. Trimming 데이터 기초 통계치

(경로: result_RNAseq/Analysis_statistics/trim_throughput.txt 참고)

분석에 들어가기 전, Trimmomatic 프로그램을 통하여 adapter sequence를 제거하고, reads의 ends로부터 base quality 3 미만인 bases와 슬라이딩 윈도우 trim기법으로 window size=4, mean quality=15를 만족하지 않으면 bases를 제거합니다. 그 후, min length=36 bp 보다 짧은 reads를 제거하는 단계를 거쳐 trimmed data를 생성하였습니다.

표 2. Trimmed Data 통계치

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: Trimming 후, 전체 시퀀싱 된 염기의 수
- Total reads: Trimming 후, 전체 read의 수
- GC(%): GC 함량
- Q20(%): Phred quality score 20 이상의 품질을 갖는 염기의 비율
- Q30(%): Phred quality score 30 이상의 품질을 갖는 염기의 비율

3. 4. Trimming 후 사이클별 평균 염기품질

(경로: result_RNAseq/Analysis_statistics/trimmedData/A_fastqc/ 참고)

그림 12, 그림 13은 trimming 단계를 거친 후 사이클별 평균 염기 품질을 나타냅니다.

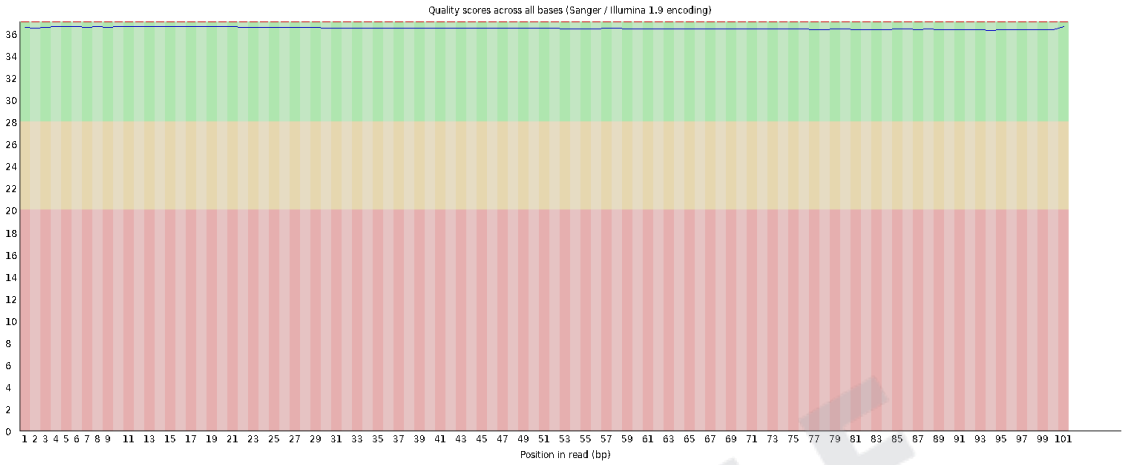


그림 12. Trimming 후 MG_CTRL_1 (read1)의 사이클별 평균 염기 품질

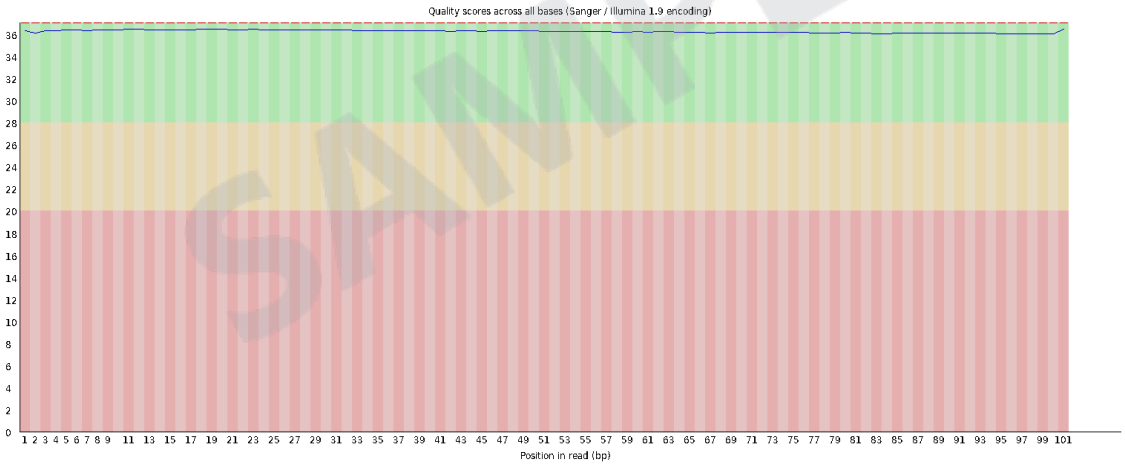


그림 13. Trimming 후 MG_CTRL_1 (read2)의 사이클별 평균 염기 품질

- Yellow box: 사이클별 염기 품질 점수의 interquartile range (25-75%)를 나타냄
- Red line: 사이클별 염기 품질 점수의 중앙 값
- Blue line: 사이클별 염기 품질 점수의 평균 값
- Green background: 우수한 품질을 나타냄.
- Orange background: 양호한 품질을 나타냄.
- Red background: 나쁜 품질을 나타냄.

4. 레퍼런스 기반 맵핑 및 전사체 어셈블리 결과

4. 1. 맵핑 데이터 통계치

(경로: result_RNAseq/Analysis_statistics/mapping.tophat.stats.txt 참고)

RNA-seq 실험을 통해서 얻은 cDNA fragment를 맵핑 하기 위해 genomic DNA reference (GRCh38)을 사용하였습니다. 아래는 Bowtie2 aligner를 통하여 spliced read mapping 처리가 가능한 TopHat 프로그램으로 mapping한 통계 결과입니다. 각 샘플별 read1, read2의 processed read 개수, mapped read 개수, multiple mapped read 개수 및 overall read mapping ratio를 확인 하실 수 있습니다.

표 3. Mapped Data 통계치

Sample ID	# of processed reads	# of mapped reads	# of failed to align reads	# of multiple mapped reads
MG_CTRL_1	66,167,454	64,779,936 (97.9%)	1,387,518 (2.1%)	2,003,840 (3.1%)
MG_CTRL_2	63,918,246	62,521,403 (97.8%)	1,396,843 (2.2%)	1,962,913 (3.1%)
MG_CTRL_3	73,393,434	71,740,691 (97.7%)	1,652,743 (2.3%)	2,365,013 (3.3%)
MG_TEST_1	78,155,316	76,171,999 (97.5%)	1,983,317 (2.5%)	2,163,176 (2.8%)
MG_TEST_2	62,086,894	60,551,614 (97.5%)	1,535,280 (2.5%)	1,762,387 (2.9%)
MG_TEST_3	70,627,636	69,138,148 (97.9%)	1,489,488 (2.1%)	1,888,886 (2.7%)

- # of processed reads: Trimming 후 cleaned read 수
- # of mapped reads: Reference에 mapping된 read 수
- # of multiple mapped reads: Multiple mapped read 수

4. 2. 레퍼런스 기반 전사체 어셈블리 및 발현값 추출

Cufflinks 프로그램을 통하여 reference gene model을 사용해 기존에 알려진 gene/transcript를 어셈블리 할 수 있습니다. 어셈블리 후, 해당 transcript의 abundance 양을 read count와 within sample normalized value인 FPKM (Fragments Per Kilobase of transcript per Million mapped reads) 값으로 발현량을 추정할 수 있습니다.

4. 2. 1. Known Transcript 발현값 추출

(경로: result_RNAseq/Expression_profile/Cufflinks/
Expression_Profile.GRCh38.transcript.xlsx 참고)

표 4는 기존에 알려진 transcript에 대한 샘플 별 발현값을 나타낸 예시 결과입니다. 이 결과는 Cufflinks의 -G 옵션을 이용해 novel transcript assembly를 고려하지 않고, reference annotation based transcript(RABT) 방식을 적용하여 기존에 알려진 transcript의 발현값을 추출한 결과를 나타냅니다.

표 4. Known transcripts 의 발현값 (예시)

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

- Transcript_ID: Splicing variant (isoform/transcript) 단위
- Gene_ID: 유전자의 이름
- Gene_Symbol: 유전자의 대표이름
- Gene_Description: 유전자의 설명
- Transcript_Locus: Genome에서 Trancript의 위치
- Transcript_Length: Transcript의 길이
- [Sample Name]_Read_Count: 샘플 별 read count
- [Sample Name]_FPKM: 샘플 별 FPKM (normalized value)

4. 2. 2. Known Gene 발현값 추출

(경로: result_RNAseq/Expression_profile/Cufflinks/
Expression_Profile.GRCh38.gene.xlsx 참고)

표 5는 유전자 단위로 샘플 별 발현값을 정리한 예시 결과 입니다. 기존에 알려진 transcript에 대한 샘플 별 발현값을 gene단위로 합한 결과이며, Cufflinks의 -G 옵션을 이용해 novel transcript assembly를 고려하지 않고, reference annotation based transcript (RABT) 방식을 적용하여 진행하였습니다.

표 5. Known gene 의 발현값 (예시)

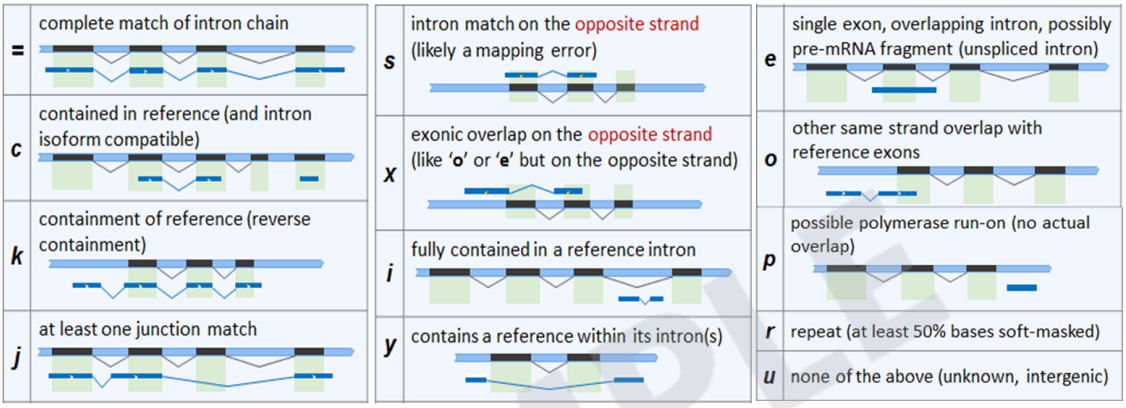
Gene_ID	Transcript_ID	Gene_Symbol	Description	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM
27	NM_001136000,NM_001136001,NM_001	ABL2	ABL proto-oncogene 2, non-receptor tyrosine	5145	6735	7.836987	9.9130971
28	NM_020469	ABO	ABO, alpha 1-3-N-acetylgalactosaminyltransfe	3	4	0.024627	0.0481101
37	NM_000018,NM_001033859,NM_001270	ACADVL	acyl-CoA dehydrogenase, very long chain	26094	26688	200.835542	195.690048
38	NM_000019	ACAT1	acetyl-CoA acetyltransferase 1	1042	3566	8.55795	27.8635
39	NM_001303253,NM_005891	ACAT2	acetyl-CoA acetyltransferase 2	933	1427	10.457377	15.259281
40	NM_001094,NM_183377	ASIC2	acid sensing ion channel subunit 2	4	0	0.013697848	0
172	NR_003226,NR_003227,NR_003228	AFG3L1P	AFG3 like matrix AAA peptidase subunit 1, pse	681	655	4.219935	3.9539
173	NM_001133	AFM	afamin	0	3	0	0.0183909
174	NM_001134	AFP	alpha fetoprotein	3	0	0.0185119	0
175	NM_000027,NM_001171988,NR_033655	AGA	aspartylglucosaminidase	251	284	2.082699679	2.2293071
176	NM_001135,NM_013227	ACAN	aggrecan	9	4	0.016414383	0.003922265
177	NM_001136,NM_001206929,NM_001206	AGER	advanced glycosylation end-product specific r	3120	2956	33.65818658	30.51683042
178	NM_000028,NM_000642,NM_000643,NM	AGL	amylo-alpha-1, 6-glucosidase, 4-alpha-glucan	4866	3634	11.58945114	8.23442506
181	NM_001138	AGRP	agouti related neuropeptide	0	0	0	0
182	NM_000214	JAG1	jagged 1	859	691	2.52453	1.94234
183	NM_000029	AGT	angiotensinogen	3	0	0.0143218	0
185	NM_000685,NM_004835,NM_009585,NM	AGTR1	angiotensin II receptor type 1	0	0	0	0

- Gene_ID : 유전자의 이름
- Transcript_ID: Splicing variant (isoform/transcript) 단위
- Gene_Symbol: 유전자의 대표이름
- Gene_Description: 유전자의 설명
- [Sample Name]_Read_Count: 샘플 별 read count
- [Sample Name]_FPKM: 샘플 별 FPKM (normalized value)

4. 3. Novel Transcript / Alternative Splicing Transcript 예측

Novel transcript와 novel alternative splicing transcript를 예측하기 위해 Cufflinks의 -g 옵션을 이용해 각 샘플의 mapping 결과로 transcript의 assemble을 추가로 진행하였습니다. 각 샘플의 assemble된 annotation 결과로 cuffmerge를 이용하여 annotation (GTF파일)을 합친 후 known / novel transcript에 대한 각 샘플별 발현값을 계산할 수 있습니다. 기존 annotation과의 비교 및 novel transcript의 형태를 구분하기 위해 cuffcompare 프로그램을 이용하였으며, 그 결과로 transcript를 아래 표 6과 같이 class code를 부여하여 기존 annotation 및 novel transcript의 타입을 분류하였습니다.

표 6. Novel splicing alternative transcript의 class code 설명



4. 3. 1. Known/Novel Transcript 예측 및 발현값 추출

(경로: result_RNAseq/Novel_transcript_analysis/Cufflinks/
Expression_Profile_with_Novel.GRCh38.transcript.xlsx 참고)

이 결과는 known transcript, novel transcript, novel alternative splicing transcript에 대하여 각 샘플별 발현값을 정리한 결과입니다.

(참고: 4.2의 레퍼런스 기반 전사체 어셈블리 및 발현값 추출은 novel transcript list가 포함되어 있지 않습니다.)

표 7는 Cufflinks를 이용하여 기존에 알려진 transcript 및 novel로 예측된 transcript에 대한 샘플 별 발현값을 나타낸 예시 결과입니다. 여기서 novel gene이 있을 경우 Cufflinks 프로그램에서는 이를 "XLOC_xxxxxx"로 임시 Gene ID를 부여하고, novel transcript나 alternative splicing transcript가 있을 경우에는 새로운 Transcript ID인 "TCONS_yyyyyyyy"로 임시 ID가 부여되며 각 transcript 마다 transcript locus, length, class code, read count, FPKM 값을 확인 할 수 있습니다.

(표 6의 class code를 참조)

표 7. Known/Novel transcript 발현값 (예시)

Transcript_ID	Gene_ID	Gene_Symbol	Description	Transcript Locus	Trancript Length	Class_Code	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM
NM_001297778	64802	NMNAT1	nicotinamide nucleotide adenylyltran	chr1:10002980-100455	3796	=	1	1	2.76598E-05	2.70542E-05
NM_001297779	64802	NMNAT1	nicotinamide nucleotide adenylyltran	chr1:10003485-100418	1094	=	48	66	0.799571	1.062
NM_022787	64802	NMNAT1	nicotinamide nucleotide adenylyltran	chr1:10003485-100455	3781	=	582	643	2.82687	3.00007
TCONS_00000309	64802	NMNAT1	nicotinamide nucleotide adenylyltran	chr1:10003603-100457	3986	j	96	48	0.442326	0.212003
TCONS_00000027	XLOC_0000		.	chr1:1000700-1002767	2067	u	342	438	2.97671	3.59819
NM_017734	54873	PALMD	palmdelphin	chr1:100111430-10016	2578	=	5	2	0.0270185	0.0138501
TCONS_00002121	54873	PALMD	palmdelphin	chr1:100111430-10017	1281	j	1	0	0.00360962	0
TCONS_00006694	391059	FRRS1	ferric chelate reductase 1	chr1:100136716-10023	4070	j	676	516	2.96286	2.13587
NR_037516	100500863	MIR548AA1	microRNA 548aa-1	chr1:100154610-10017	44	=	0	0	0	0
TCONS_00006695	391059	FRRS1	ferric chelate reductase 1	chr1:100168981-10023	7983	j	779	1182	1.74064	2.49467
NM_001013660	391059	FRRS1	ferric chelate reductase 1	chr1:100174258-10023	2678	=	322	157	2.14266	0.98449
NR_049882	100847055	MIR5697	microRNA 5697	chr1:10027438-100275	78	=	0	0	0	0
TCONS_00008992	XLOC_0026		.	chr1:1002858-1003976	1118	u	83	126	1.3243	1.92859
NM_000642	178	AGL	amylo-alpha-1, 6-glucosidase, 4-alp	chr1:100315639-10038	7371	=	4604	3525	10.9625	7.98643
TCONS_00002123	178	AGL	amylo-alpha-1, 6-glucosidase, 4-alp	chr1:100315639-10038	7997	j	56	48	0.121196	0.0986665
NM_000643	178	AGL	amylo-alpha-1, 6-glucosidase, 4-alp	chr1:100316044-10038	7169	=	1	1	1.44458E-05	8.71369E-06
NM_000028	178	AGL	amylo-alpha-1, 6-glucosidase, 4-alp	chr1:100316044-10038	7449	=	62	1	0.145269	4.02099E-05

- Transcript_ID: Splicing variant (isoform/transcript) 단위
- Gene_ID: 유전자의 이름
- Gene_Symbol: 유전자의 대표이름
- Gene_Description: 유전자의 설명
- Transcript_Locus: Genome에서 trancript의 위치
- Transcript_Length: Transcript의 길이
- Class_Code: Transcript ID에 따른 class code (표 6 참조)
- [Sample Name]_Read_Count: 샘플 별 read count
- [Sample Name]_FPKM: 샘플 별 FPKM (normalized value)

4. 3. 2. Known/Novel Gene 예측 및 발현값 추출

(경로: result_RNAseq/Novel_transcript_analysis/Cufflinks/
Expression_Profile_with_Novel.GRCh38.gene.xlsx 참고)

이 결과는 known transcript, novel transcript, novel alternative splicing transcript에 대하여 각 샘플별 발현값을 gene level로 정리한 결과입니다.

(참고: 4.2의 레퍼런스 기반 전사체 어셈블리 및 발현값 추출은 novel transcript list가 포함되어 있지 않습니다.)

표 8는 Cufflinks를 이용하여 기존에 알려진 gene 및 novel로 예측된 gene에 대한 샘플 별 발현값을 나타낸 예시 결과입니다. 여기서 novel gene이 있을 경우 Cufflinks 프로그램에서는 이를 "XLOC_xxxxxx"로 임시 Gene ID를 부여하고, novel transcript나 alternative splicing transcript가 있을 경우에는 새로운 Transcript ID인 "TCONS_yyyyyyyy"로 임시 ID가 부여되며 각 gene에 따른 Transcript ID와 gene_symbol, transcript별 class code, 샘플 별 read count와 FPKM 값을 확인 할 수 있습니다.

(표 6의 class code를 참조)

표 8. Known/Novel gene 발현값 (예시)

Gene_ID	Transcript_ID	Gene_Symbol	Description	Class_Code	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM
1	NM_130786,TCONS_00047533	A1BG	alpha-1-B glycoprotein	=,j	195	134	0.69357	0.4502962
2	NM_000014,NM_001347423,NM_001	A2M	alpha-2-macroglobulin	=,=,=,=,j,j	0	7	0	0.007631116
3	NR_040112	A2MP1	alpha-2-macroglobulin pseudoge	=	0	0	0	0
9	NM_000662,NM_001160170,NM_001	NAT1	N-acetyltransferase 1	=,=,=,=,=,=,=,=	58	121	0.396796035	0.982973461
10	NM_000015	NAT2	N-acetyltransferase 2	=	2	0	0.0303151	0
XLOC_000011	TCONS_00000024,TCONS_0000002	.		u,u,u,u	1065	1315	5.14250025	6.07279804
12	NM_001085	SERPINA3	serpin family A member 3	=	0	3	0	0.0238443
13	NM_001086	AADAC	arylacetamide deacetylase	=	0	5	0	0.0471377
14	NM_001087,NM_001302545	AAMP	angio associated migratory cell p	=,=	5650	7792	54.62937	71.39462
15	NM_001088,NM_001166579,NR_1105	AANAT	aralkylamine N-acetyltransferase	=,=,=,j	24	26	0.277158678	0.225565462
16	NM_001605	AARS	alanyl-tRNA synthetase	=	17643	26056	92.6457	130.121
XLOC_000017	TCONS_00000038	.	.	u	127	366	0.992825	2.82578
XLOC_000018	TCONS_00000039,TCONS_0000004	.	.	u,u,u	106	106	1.491445406	1.173963163
18	NM_000663,NM_001127448,NM_0206	ABAT	4-aminobutyrate aminotransferas	=,=,=,j,j	2161	1922	7.614496588	6.632409
19	NM_005502,TCONS_00088409	ABCA1	ATP binding cassette subfamily A	=,j	62	107	0.099729045	0.177826165
20	NM_001606,NM_212533,TCONS_000	ABCA2	ATP binding cassette subfamily A	=,=,j,j,j,j	13632	13426	28.499237	26.863917
21	NM_001089,TCONS_00033835,TCO	ABCA3	ATP binding cassette subfamily A	=,j,j,j	918	1087	2.653321	2.740275

- Gene_ID: 유전자의 이름
- Transcript_ID: Splicing variant (isoform/transcript) 단위
- Gene_Symbol: 유전자의 대표이름
- Gene_Description: 유전자의 설명
- Class_Code: Transcript ID에 따른 class code (표 6 참조)
- [Sample Name]_Read_Count: 샘플 별 read count
- [Sample Name]_FPKM: 샘플 별 FPKM (normalized value)

4. 3. 3. Novel Transcript 정리

(경로: result_RNAseq/Novel_transcript_analysis/Cufflinks/Novel_transcript_list.xlsx 참고)

이 결과는 기존의 알려진 exon이나 gene에 맵핑 되지 않고, intergenic region에 맵핑되어 novel transcript로 예측 된 결과입니다. 표 9은 known / novel transcript예측 결과에서 class code가 'u'인 list만을 선별하여 novel transcript list파일로 저장한 결과입니다.

표 9. Novel transcript 리스트 (예시)

Transcript_ID	TCONS_00000634	TCONS_00000957	TCONS_00002485	TCONS_00002500	TCONS_00002501
Gene_ID	XLOC_000199	XLOC_000284	XLOC_000722	XLOC_000728	XLOC_000728
Transcript_Locus	chr1:22482041-22484118	chr1:32001255-32002175	chr1:120150436-120158602	chr1:142804999-142810577	chr1:142805680-142827136
Trancript_Length	1902	920	8166	4875	2256
Strand	+	+	+	+	+
Exon_Count	2	1	1	2	8
Exon_Start	22482042,22483206	32001256	120150437	142805000,142806244	142805681,142810265,142812943, 142813101,142819266,142823389, 142824917,142826264
Exon_End	22483030,22484118	32002175	120158602	142805540,142810577	142806313,142810349,142812984, 142813293,142819336,142823552, 142825111,142827136
Class_Code	u	u	u	u	u
AM_Read_Count	28	11	183	133	48
BM_Read_Count	19	10	68	74	34
AM_FPKM	0.259766	0.204279	0.405756	0.555787	0.4318
BM_FPKM	0.155353	0.186772	0.139684	0.329609	0.322364

- Transcript_ID: 기존에 알려지지 않았던 새로운 exon이 포함된 transcript인 경우, Cufflinks 프로그램에서는 이를 "TCONS_yyyyyyyy"로 임시 ID를 부여
- Gene_ID: 기존에 알려지지 않았던 영역에 새롭게 탐색한 gene이 있을 경우, Cufflinks 프로그램에서는 이를 "XLOC_xxxxxx"로 임시 ID를 부여
- Transcript_Locus: Genome에서 Trancript의 위치
- Transcript_Length: Transcript의 길이
- Strand: Genomic region에서의 transcript의 방향
- Exon_Count: 해당 transcript의 exon 개수
- Exon_Start, End: 해당 transcript의 exon별 start와 end의 position
- Class_Code: Transcript ID에 따른 class code (표 6 참조)
- [Sample Name]_Read_Count: 샘플 별 read count
- [Sample Name]_FPKM: 샘플 별 FPKM값 (normalized value)

4. 3. 4. Novel Alternative Splicing Transcript 정리

(경로: result_RNAseq/Novel_transcript_analysis/Cufflinks/Novel_splicing_variant_list.xlsx 참고)

이 결과는 known / novel transcript 예측 결과에서 novel alternative splicing transcript (class code: 'j', 'c', 'k', 'e', 'i', 'o', 'p', 's', 'x')에 대해서만 추출하여 gene을 기준으로 저장한 결과입니다.

Novel alternative splicing transcript는 기존의 알려진 exon에 맵핑 되지 않고, 새로운 exon으로 예측되거나 기존 transcript와는 다르게 변형된 구조를 가진 transcript를 말합니다. 표 10은 샘플의 예시로 Cufflinks의 Reference Annotation Based Transcript assembly (RABT) 방식을 적용하여 reference와 다른 novel alternative splicing transcript를 탐색하여 추출한 결과를 나타냅니다. 가장 근접한 gene을 기준으로 해당되는 novel alternative splicing transcript의 start, transcript end, exon count, exon start, exon end position, read count, FPKM, class code 값을 확인 할 수 있습니다.
(표 6의 class code를 참조)

표 10. Novel alternative splicing transcript 리스트 (예시)

refGene_Name	43	48	49	81	54991
nearest_refTranscript_Name	NM_001302621	NM_001278352	NM_001097	NM_004924	NM_001330306
cuffGene_Name	43	48	49	81	54991
cuffTranscript_Name	TCONS_00081480	TCONS_00087654	TCONS_00058856	TCONS_00046202	TCONS_00004599
Gene_Symbol	ACHE	ACO1	ACR	ACTN4	C1orf159
Gene_Description	acetylcholinesterase (Cartwright blood group)	aconitase 1	acrosin	actinin alpha 4	chromosome 1 open reading frame 159
Transcript_Locus	chr7:100487424-100493754	chr9:32407455-32417830	chr22:51177917-51183858	chr19:39190811-39220132	chr1:1016862-1051736
Trancript_Length	3709	3338	1412	29238	3804
Strand	-	-	+	-	-
Exon_Count	4	3	3	2	9
Exon_Start	100487425,100488790, 100489955,100493423	32407456,32414501, 32416588	51177918,51182489, 51183081	39190812,39220085	1016863,1019861,1021258, 1022519,1022882,1025733, 1026852,1027371,1051440
Exon_End	100488709,100488959, 100491876,100493754	32408246,32415804, 32417830	51178405,51182634, 51183858	39220001,39220132	1019763,1019886,1021392, 1022584,1022977,1025808, 1026945,1027483,1051736
Class_Code	j	x	j	x	j
AM_Read_Count	1	117	23	13888	186
BM_Read_Count	1	110	1	20048	332
AM_FPKM	0.000819272	0.60924	0.287372	8.59587	0.86336
BM_FPKM	9.09894E-05	0.555086	1.99817E-06	11.8636	1.47506

- Gene_ID: Gene ID
- nearest_refGene_ID: Novel alternative splicing transcript가 예측된 부분에서 가장 근접한 유전자의 이름
- nearest_refTranscript_Name: Novel alternative splicing transcript가 예측된 부분에서 가장 근접한 transcript ID
- cuffGene_Name: Cufflinks 프로그램에서 부여된 gene의 ID (“XLOC_xxxxxx” 로 부여됨)
- cuffTranscript_Name: Cufflinks 프로그램에서 부여된 transcript의 ID (“TCONS_yyyyyyyy” 로 부여됨)
- Gene_Symbol: 근접한 유전자의 대표이름
- Gene_Description: 근접한 유전자의 설명
- Transcript_Locus: Genome에서 Transcript의 위치
- Transcript_Length: Transcript의 길이
- Strand: Genomic region에서의 transcript의 방향
- Exon_Count: 해당 transcript의 exon 개수
- Exon_Start, End: 해당 transcript의 exon별 start와 end의 position
- Class_Code: Transcript ID에 따른 class code (표 6 참조)

- [Sample Name]_Read_Count: 샘플 별 read count
- [Sample Name]_FPKM: 샘플 별 FPKM (normalized value)

5. 차별 발현 유전자 분석 결과

5. 1. 분석 데이터 품질 확인 및 전처리

전사체 어셈블리 결과로 얻은 known gene에 대한 read count 값을 가지고, 샘플간의 차별 발현 유전자를 선별하는 과정을 진행합니다. 전처리 과정에서는 분석에 들어가기 전 데이터의 quality check, 샘플간 normalization 과정을 진행하고, biological replicates이 존재할 때 샘플 간 유사성을 확인하여 데이터 신뢰성 여부를 파악합니다.

(경로: result_RNAseq/DEG_result/ [DataSet] /Analysis_Result.html 참고)

5. 1. 1. 샘플정보 및 분석 디자인

분석을 위해 사용한 샘플은 총 6샘플 입니다. 자세한 샘플 정보 및 분석 조합은 Sample.Info.txt 파일을 참고 바랍니다.

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

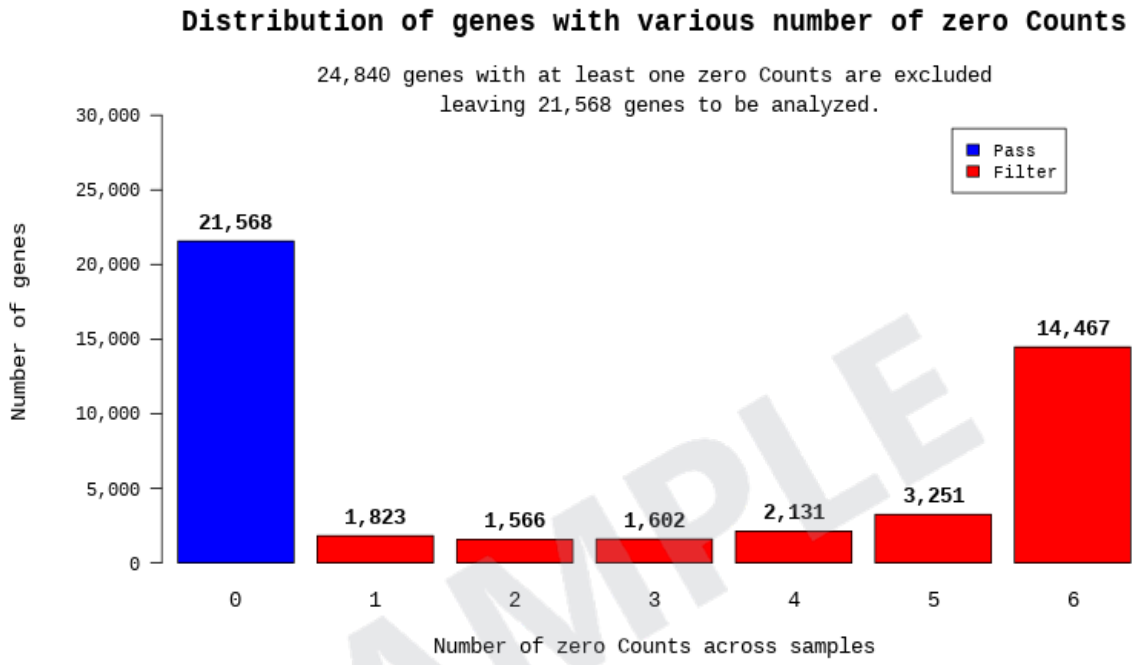
분석 비교조합과 해당 조합에 대한 통계분석 방법은 아래와 같습니다.

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

5. 1. 2. 데이터 Quality Check

(경로: result_RNAseq/DEG_result/ [DataSet] /Data Quality Check/ 참고)

각 gene 별, 전체 6개 샘플에서 적어도 한 샘플 이상에서 0인 Count값을 가지는 gene는 분석에서 제외함. 따라서, 총 46,408개 gene 중에서 24,840개를 제외한 21,568개 gene을 대상으로 통계분석을 진행함.



5. 1. 3. 데이터 변환 및 정규화

샘플간 비교에 있어 biological meaning에 영향을 줄 수 있는 systematic bias를 줄이기 위해 read count data를 이용하여 size factor를 추정하고, 이를 이용하여 Relative Log Expression (RLE) normalization 후, 통계분석을 진행했습니다. (DESeq2 R library 이용)

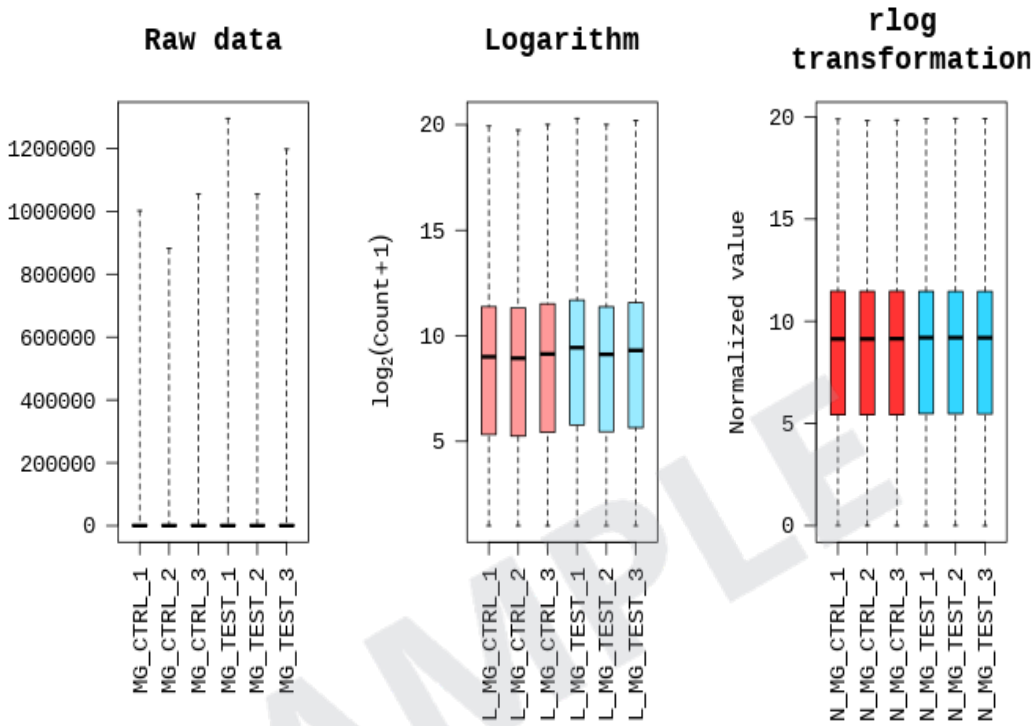
결과를 시각화 하기 위해 log2(read count+1) 값과 regularized log (rlog) transformation 값을 이용했습니다. rlog transformation은 발현값이 낮은 특정 gene/transcript에 대해서 샘플 간의 차이를 최소화 하는 방법입니다. 먼저 count data를 log2 scale로 변환한 뒤, library size factor로 normalization을 진행합니다. 샘플간 library size factor의 차이가 큰 경우에는 rlog 값을 이용하는 것이 데이터 시각화에 유리하다고 알려져 있습니다.

위 logarithm값들은 visualization을 위해서만 사용됩니다.

DESeq2를 이용한 통계 분석 시에는 RLE normalized count를 이용하여 negative binomial Wald Test(nbinomWaldTest)를 진행하였습니다.

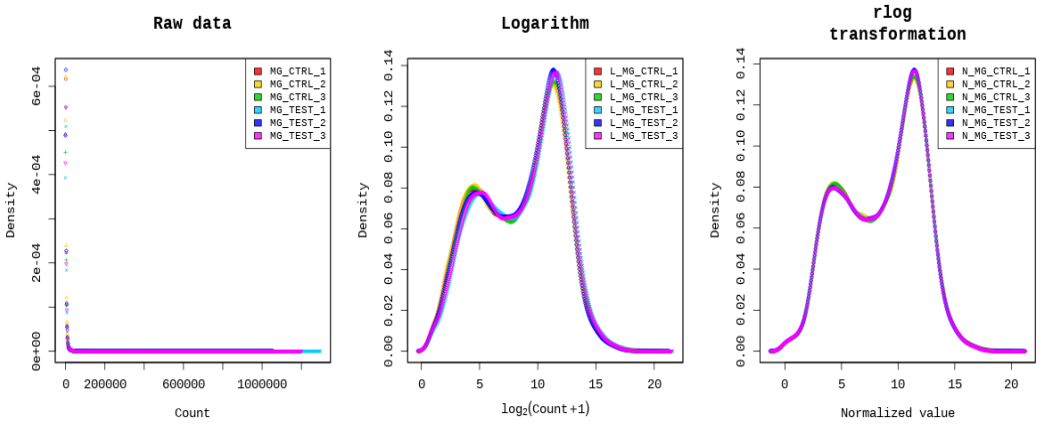
5. 1. 3. 1. 샘플별 발현값 분포의 상자그림

아래는 raw signal (read count)과 read count+1의 Logarithm (based 2), RLE Normalization 값에 대하여 해당 샘플별 발현값분포를 나타내기 위해 백분위수, 중앙값, 25백분위수, 75백분위수, 최대값, 최소값을 이용하여 시각적으로 표현한 상자 그림입니다.



5. 1. 3. 2. 샘플별 발현값 분포의 Density Plot

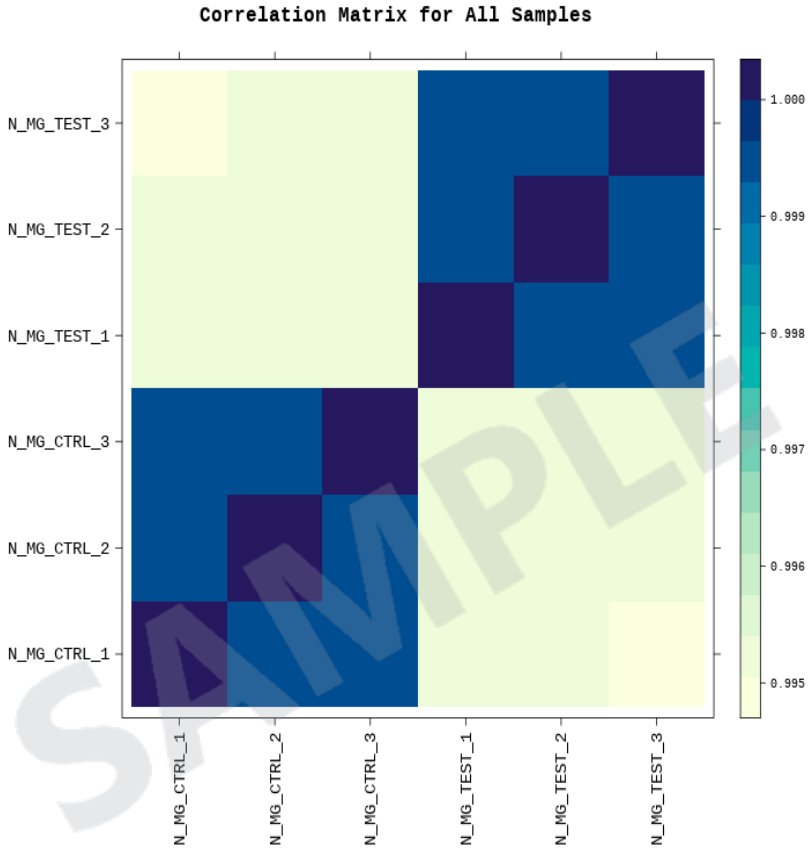
아래는 raw signal (read count)과 read count+1의 Logarithm (based 2), RLE Normalization 값에 대하여 각 샘플별 전체적인 발현값의 분포를 한 눈에 알아보기 위하여 density Plot으로 나타낸 그림입니다.



5. 1. 4. 샘플간 상관성 분석

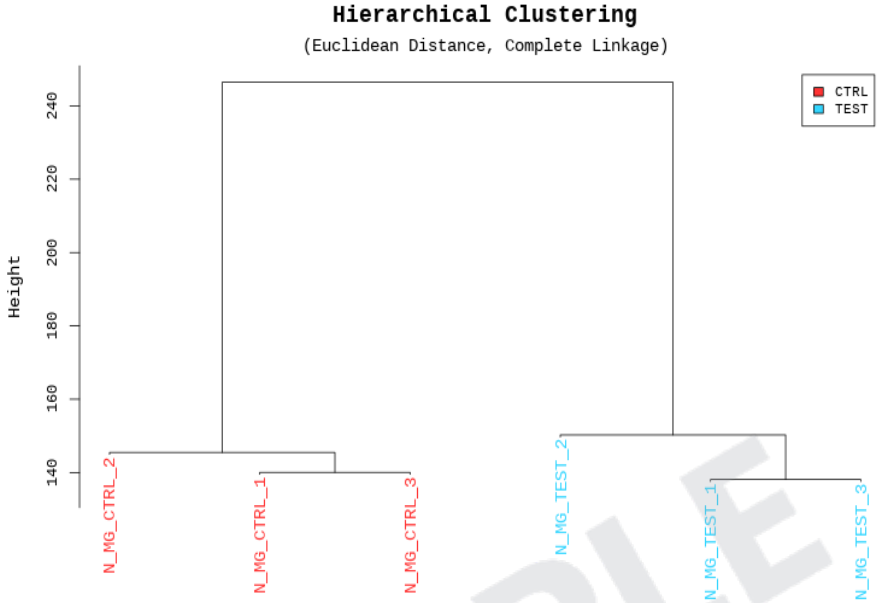
각 샘플별, rlog transformed value 를 사용하여 샘플간 유사성 정도(Pearson's coefficient, 피어슨 상관계수)를 살펴봄으로써 반복 샘플의 재현성 여부를 확인합니다. (Range: $-1 \leq r \leq 1$) 상관 계수값이 1에 가까울수록 샘플간 유사성이 높음을 의미합니다.

전체 샘플에 대한 correlation matrix는 아래와 같습니다.



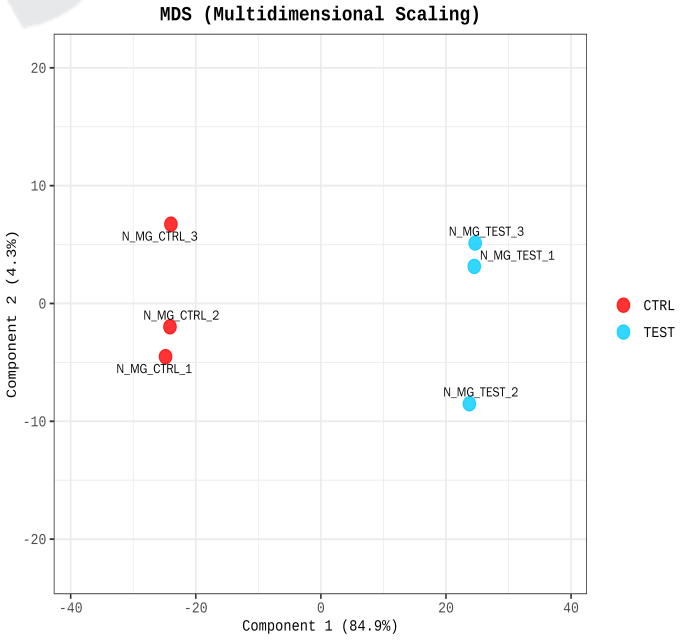
5. 1. 5. 계층형 군집(Hierarchical Clustering)분석

각 샘플별, rlog transformed value 를 사용하여 어떤 샘플끼리 발현 정도가 유사한지 그룹화 하였습니다. (Distance metric = Euclidean distance, Linkage method= Complete Linkage)



5. 1. 6. 다차원 척도법(MDS, Multidimensional Scaling)

각 샘플별, rlog transformed value 를 사용하여 샘플간 유사성 정도를 반영하는 2개의 Component를 이용하여 2차원 공간 상에 표현한 그림입니다. Outlier 샘플이 있는지, 샘플 그룹 간 유사한 발현 패턴을 가지는 군집이 있는지 여부를 확인 할 수 있습니다.



5. 2. 차별 발현 유전자 분석 절차

아래 설명은 DEG(Differentially Expressed Genes) 분석 순서를 나타냅니다.

1) Original Raw Data 는 Cufflinks -G option통해 얻은 기존의 알려진 genes에 대한 read count값을 대상으로 하였습니다.

- Raw data

(경로: result_RNAseq/Expression_profile/Cufflinks/ Expression_Profile.GRCh38.gene.xlsx 참고)
: 46,408 genes, 6 samples

2) 데이터 전처리 및 QC 과정에서 low quality를 가지는 genes를 filtering 후, RLE normalization을 진행하였습니다.

- Processed data

(경로: result_RNAseq/DEG_result/ [DataSet] /data2.xlsx 참고)
: 21,568 genes, 6 samples

3) 통계분석은 각 비교조합 별 Fold Change, nbinomWaldTest using DESeq2를 이용하였습니다.
유의한 결과는 |fc|>=2 & nbinomWaldTest raw p-value<0.05 조건으로 선별하였습니다.

- Significant data

(경로: result_RNAseq/DEG_result/DEG/data3_fc2_&_raw.p.xlsx 참고)
: 2,886 genes

4) 유의한 유전자 리스트에 대해 hierarchical clustering 분석을 통해 각 유전자별로 샘플별 유사성 정도를 그룹화하였고, 이에 대해 heatmap 및 dendrogram으로 시각화 하였습니다.

- Hierarchical Clustering (Euclidean Distance, Complete Linkage)

(경로: result_RNAseq/DEG_result/ [DataSet] /Cluster image/ 참고)

5) 유의한 유전자 리스트에 대해 g:Profiler tool (https://biit.cs.ut.ee/gprofiler/) 를 기반으로 Gene Ontology Enrichment 분석을 진행하였습니다.
data3 파일의 GO_stat, GO_genes sheet를 참고 바랍니다.

아래의 관련 결과가 제공됩니다.

- GO_stat
- GO_genes

6) 유의한 유전자 리스트에 대해 KEGG database (http://www.genome.jp/kegg/) 를 기반으로 gene-set enrichment 분석을 진행하였습니다.
data3 파일의 KEGG_stat, KEGG_genes sheet를 참고 바랍니다.

아래의 관련 결과가 제공됩니다.

- KEGG_stat
- KEGG_genes

또한 KEGG_pathway.html 파일로 KEGG enrichment analysis 결과를 확인하실 수 있습니다.

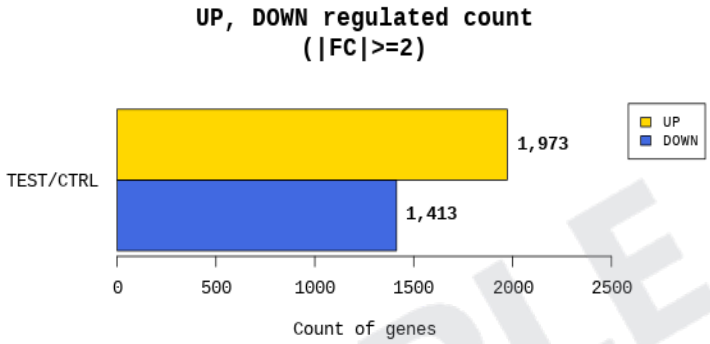
5. 3. 차별 발현 유전자 선별 결과

(경로: result_RNAseq/DEG_result/[DataSet]/Plots/ 참고)

이하 Result는 fc2_&_raw.p, TEST_vs_CTRL 에 대한 결과 예시입니다.

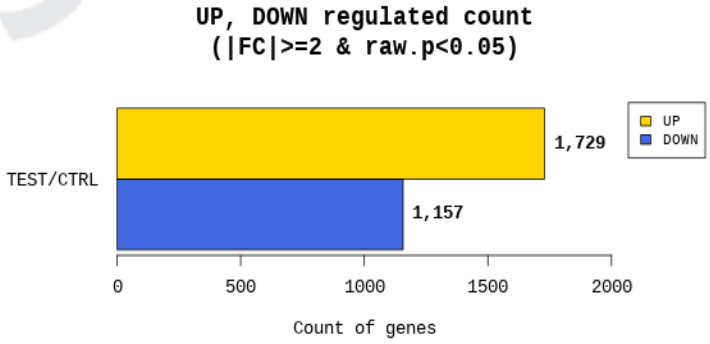
5. 3. 1. Fold change 기준 up, down별 Gene 개수

해당 비교조합별 fold change 기준으로 up, down별 gene 개수를 나타냅니다.



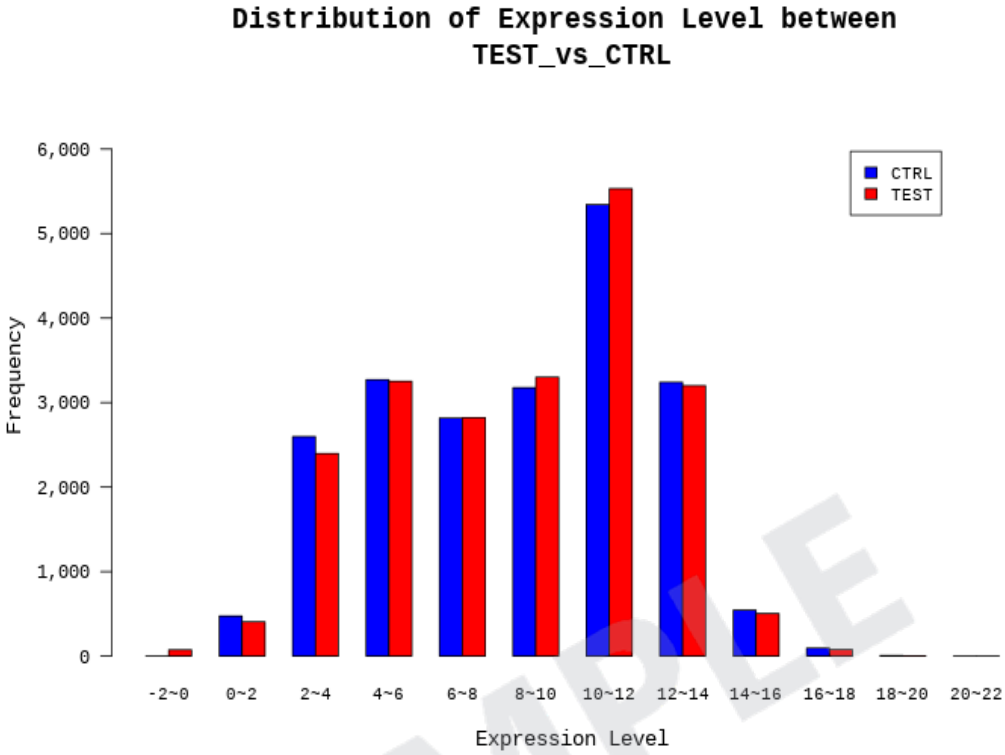
5. 3. 2. Fold Change 및 p-value 기준 Up, Down별 Gene 개수

해당 비교 조합별 fold change 및 p-value기준으로 유의한 gene의 개수를 나타냅니다.



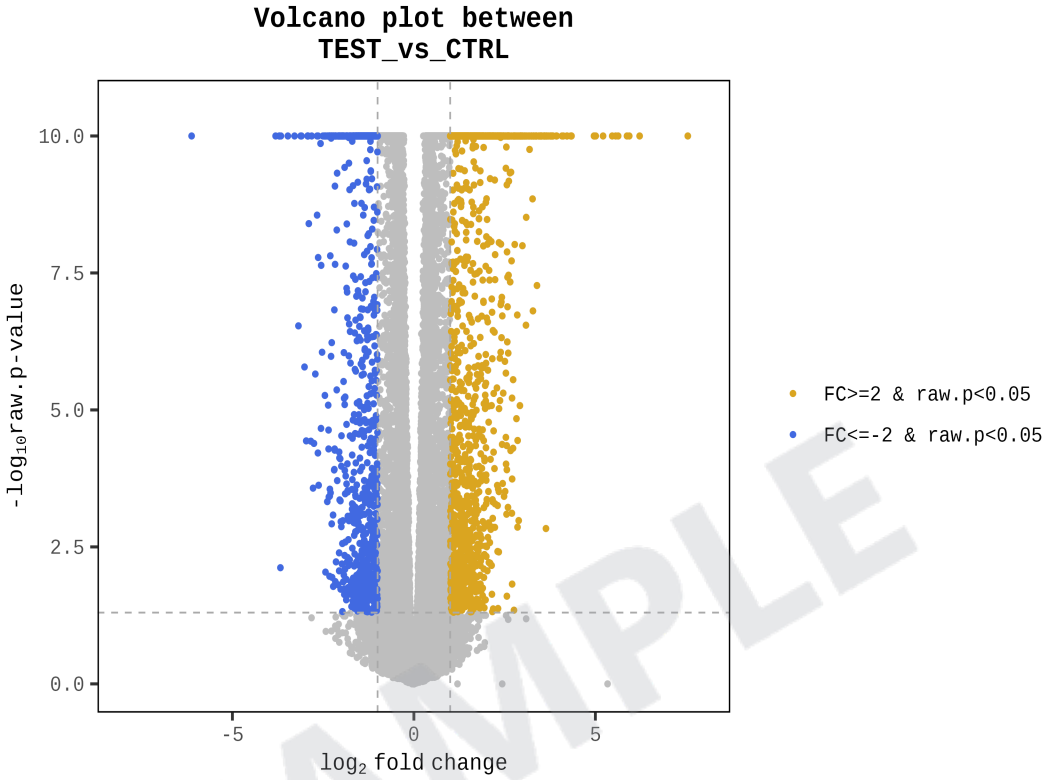
5. 3. 3. 두 그룹간 발현값 분포

해당 비교 조합에서 각 그룹별 normalized 값의 분포를 나타냅니다.



5. 3. 4. 두 그룹간 발현값의 Volcano Plot

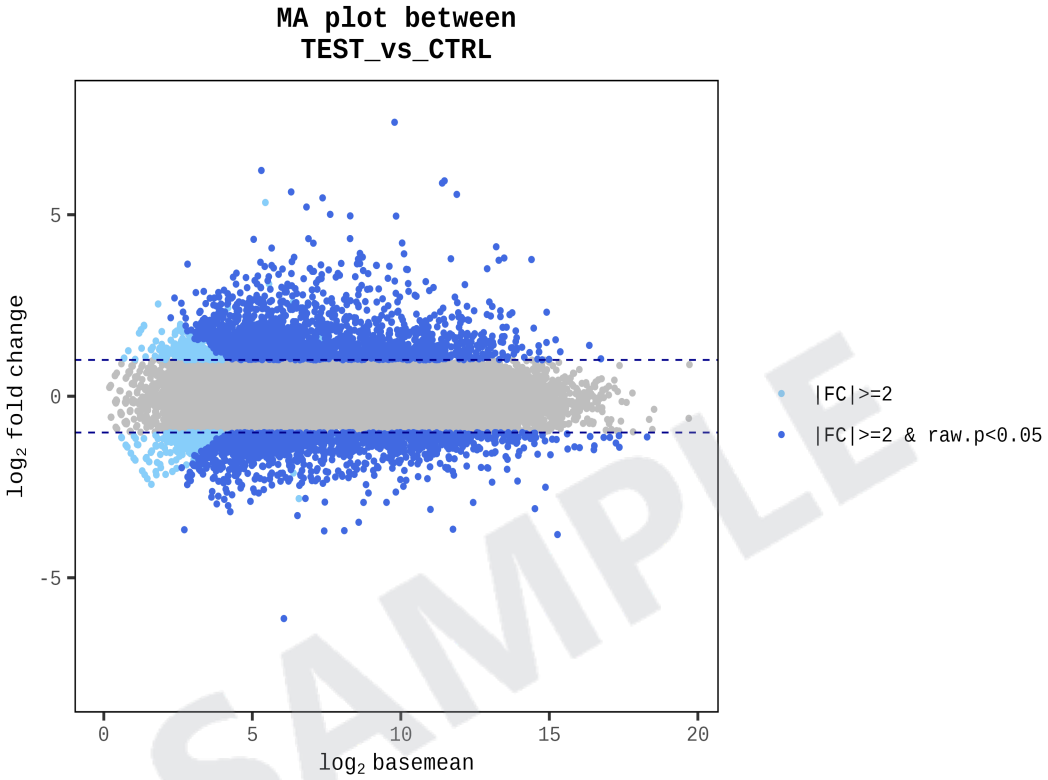
해당 비교조합간 발현값의 log2 fold change와 두 그룹간 평균 비교를 통해 도출된 p-value를 volcano plot으로 나타낸 그림입니다.
(X축: log2 Fold change, Y축: -log10 p-value)



5. 3. 5. 발현 강도에 따른 차별 발현 유전자 표현, MA Plot

두 그룹의 발현값의 평균이 높으면서 control 대비, test 그룹에서 차이가 있었던 gene을 확인하기 위해 MA plot(X축: mean of normalized counts, Y축: log2 Fold Change)으로 나타내었습니다.

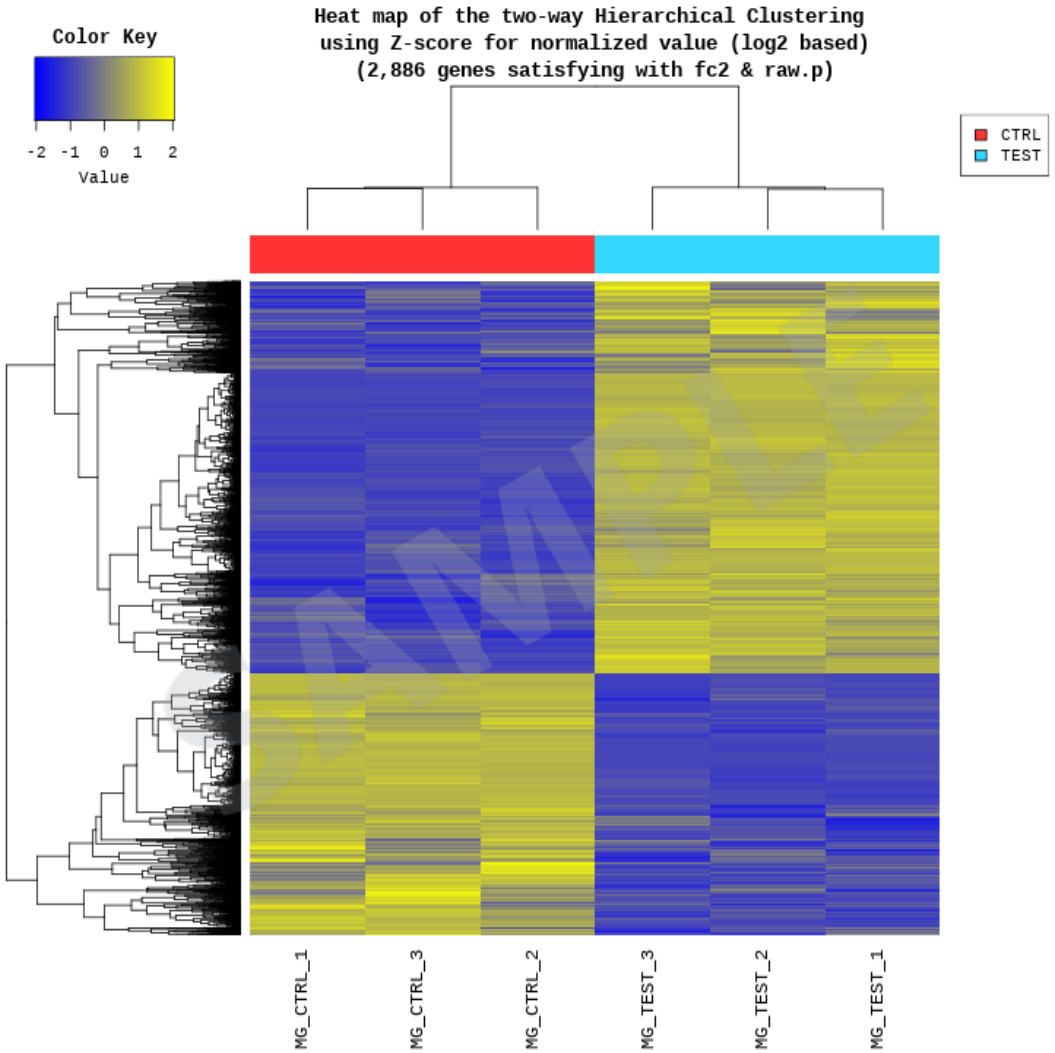
예를 들어, fold change가 동일하게 2배 이상 차이 나더라도 평균 발현값이 낮은 곳에서 2배 이상 차이 나는 것에 비해 높은 곳에서 차이가 나는 gene이 신뢰성이 높을 수 있습니다.



5. 3. 6. 계층형 군집(Hierarchical Clustering) 분석

(경로: result_RNAseq/DEG_result/ [DataSet] /Cluster image/ 참고)

유의한 DEG 리스트에 대하여 각 샘플의 gene별 발현값(rlog transformed value)을 이용하여 발현 정도가 유사한 샘플 및 genes를 hierarchical clustering analysis(Euclidean Distance, Complete Linkage)를 통하여 그룹화하여 나타내었습니다.



5. 4. GO Enrichment 분석

(경로: result_RNAseq/DEG_result/ [DataSet] /gprofiler/ 참고)

g:Profiler 는 Gene Ontology, biological pathways 등의 data source 에 기반한 Enrichment analysis를 수행하는 Tool 입니다.

GO 분석은 GO의 3가지 category에 대하여 진행되었습니다. GO의 graph 구조에 해당하는 Ontology file 및 Annotation file (GO consortium에 해당하는 각 종별 reference DB에서 제공하는 annotation, 혹은 Uniprot에서 제공하는 multispecies annotation)을 parsing하여 GO ID와 연관된 gene 혹은 gene product, molecule을 정리했습니다.

- 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 분석 결과는 아래의 2가지 형태로 DEG결과 (data3-*.xlsx 파일)의 각 시트에 정리되어 제공됩니다.

- GO_stat
- GO_genes

5. 4. 1. GO_stat Sheet

term_id를 기준으로, association되어있는 gene과 test stat을 정리한 결과입니다. 특정 term_id가 해당 DEG set을 이용한 enrichment 분석에서 얼마나 유의한지를 살펴볼 수 있습니다.

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

- source: Gene ontology의 3개 category 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: Hypergeometric test & multiple testing correction (FDR) 으로 도출된 보정된 p-value
- query_size: 해당 data source (the functional category)에 association된 unique DEG의 수
- intersection_size: 해당 term_id에 association된 unique DEG의 수
- term_size: 샘플 Species의 전체 Gene 중 해당 term_id에 association된 gene의 수
- effective_domain_size: 샘플 Species의 전체 Gene 중 해당 data source (the functional category)에 association된 gene의 수
- intersections: 해당 term_id에 association된 DEG (쉼표로 연결됨)

5. 4. 2. GO_genes Sheet

Gene을 기준으로, association된 term_id와 DEG분석 결과 stat를 정리한 결과입니다. 특정 Gene에 어떠한 term_id가 association되어있는지를 fold change, p-value, volume, normalized value와 같은 stat과 함께 살펴볼 수 있습니다.

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

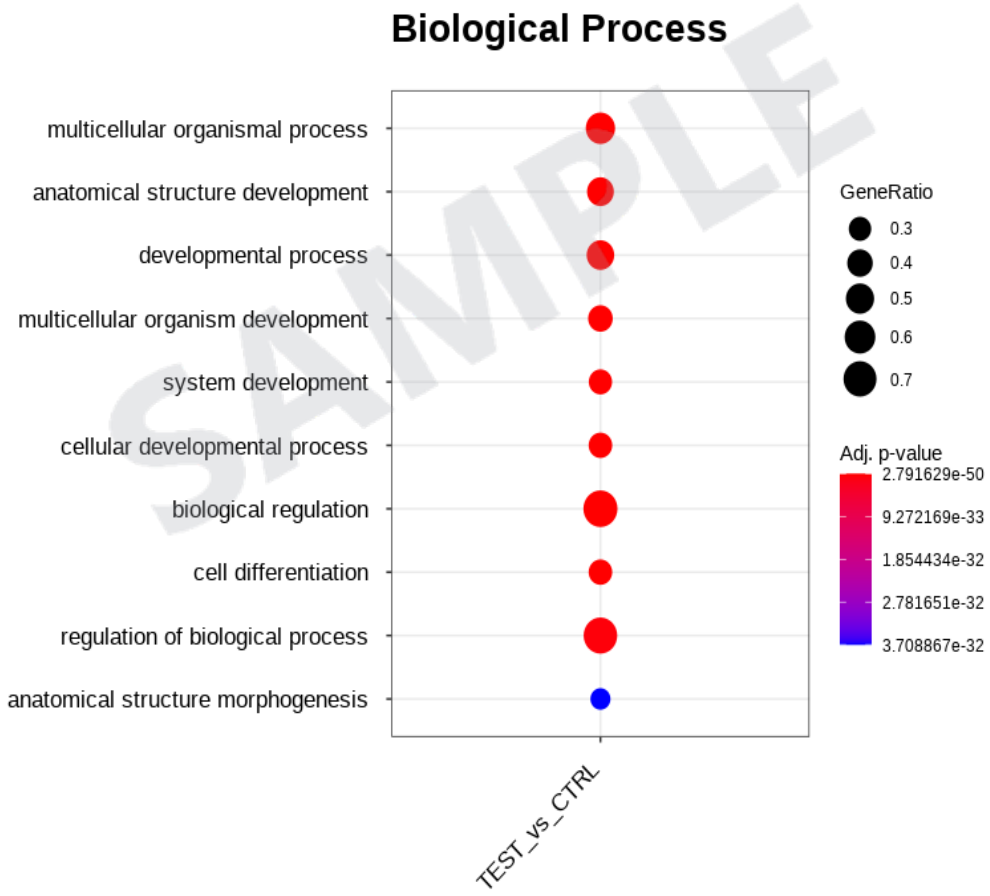
- source: Gene ontology의 3개 category Ex〉 GO:BP | GO:CC | GO:MF ...
- term_id: ID for the enriched term/functional category
- term_name: Readable name for the enriched term
- adjusted_p_value: Hypergeometric test & multiple testing correction (FDR) 으로 도출된 보정된 p-value
- intersection_size: 해당 term_id에 association된 unique DEG의 수

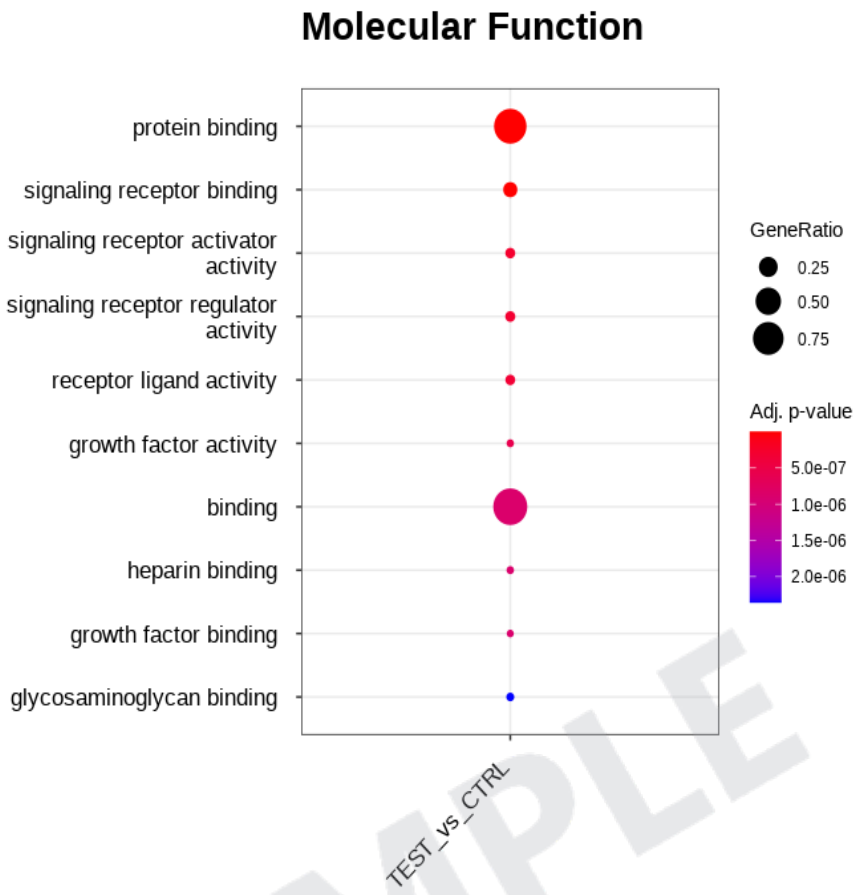
data3.GO_*.gprofiler.png: Gene Ontology Enrichment Analysis 결과 중 adjusted_p_value 기준 상위 20개의 term을 dot plot으로 나타냄.
(GO_stat 를 기반으로 작성됨)

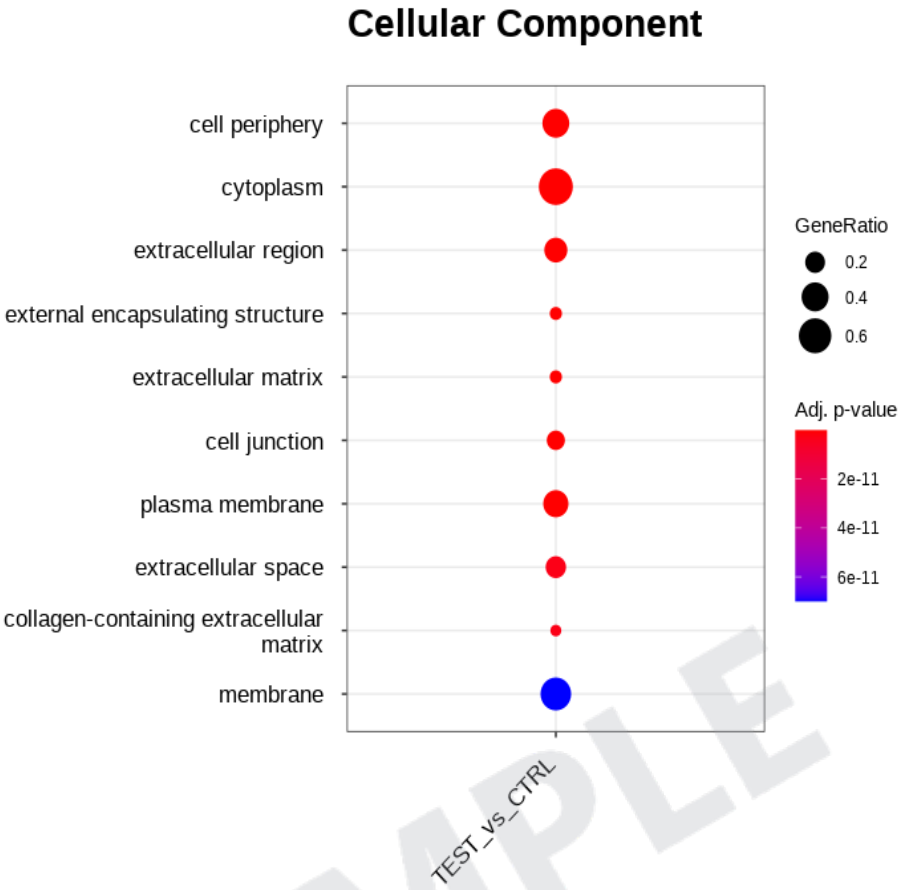
data3.GO_*.gprofiler.sizefilt.png: Gene Ontology Enrichment Analysis 결과 중 term_size filtering(min=10, max=500) 후 adjusted_p_value 기준 상위 20개의 term을 dot plot으로 나타냄.
(GO_stat 를 기반으로 작성됨. ./gprofiler/data3*.GO/폴더 참조)

- term_size filtering: Term size가 매우 적거나 큰 GO term의 경우, Hypergeometric test와 같은 검정에서 통계적 유의성이 과장될 수 있기 때문에 size filtering을 진행함.
- GeneRatio : 해당 term 에 관련된 DEG 개수 / 해당 functional category에 관련된 DEG 개수 (intersection_size / query_size)

아래는 유의한 genes를 대상으로 Gene ontology DB를 통해 enrichment 분석을 진행한 결과를 dot plot으로 나타내었습니다. (term size filtering 진행하지 않은 data3.GO_*.gprofiler.png 에 대한 예시)




Molecular Function
protein binding
signaling receptor binding
signaling receptor activator activity
signaling receptor regulator activity
receptor ligand activity
growth factor activity
binding
heparin binding
growth factor binding
glycosaminoglycan binding
TEST_vs_CTRL
GeneRatio
0.25
0.50
0.75
Adj. p-value
5.0e-07
1.0e-06
1.5e-06
2.0e-06


Cellular Component
cell periphery
cytoplasm
extracellular region
external encapsulating structure
extracellular matrix
cell junction
plasma membrane
extracellular space
collagen-containing extracellular matrix
membrane
TEST_vs_CTRL
GeneRatio
0.2
0.4
0.6
Adj. p-value
2e-11
4e-11
6e-11

5. 5. KEGG Enrichment 분석

(경로: result_RNAseq/DEG_result/ [DataSet] /KEGG_view/ 참고)

KEGG database는 molecular information (genome sequence, structure), chemical information (Metabolism, Glycans, Lipids etc.), molecular interaction information (physical interaction, co-expression)과 같은 다양한 종류의 omics 정보가 포함되어 있습니다.

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

KEGG pathway viewer는 해당 species에 관련된 pathway map 정보를 바탕으로 각 조합별 차별 발현된 유전자의 fold change 정보를 pathway map에 컬러로 표현하고, 각 pathway map별 유의한 DEG 리스트에서 매핑된 유전자 갯수와 해당 species에서 해당 pathway에 매핑되는 전체 유전자 갯수를 파악하여 gene-set enrichment test를 수행하고, 그 결과를 heatmap으로 제공합니다.

Enrichment 분석 결과는 아래의 2가지 형태로 DEG결과 (data3-*.xlsx 파일)의 각 시트에 정리되어 제공됩니다.

- KEGG_stat
- KEGG_genes

아래 heatmap은 각 pathway map별로 modified fisher's exact test를 이용하여 gene-set enrichment analysis를 수행한 결과 입니다. Legend는 enrichment p-value를 나타내며, p-value가 0.05 보다 낮을 수록 significantly enriched pathway term이라고 할 수 있습니다. 각 비교조합별 pathway map block을 클릭하면, 해당 pathway map에 매칭되는 유전자의 fold change coloring을 보실 수 있습니다.

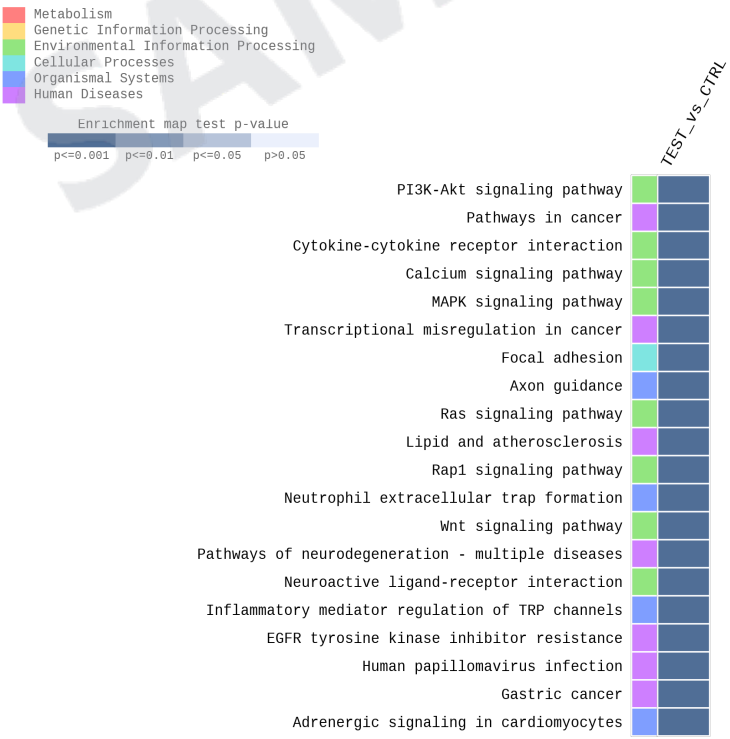


그림 14. Gene-set enrichment 분석결과 (p-value top 20)

5. 5. 1. KEGG HTML Viewer

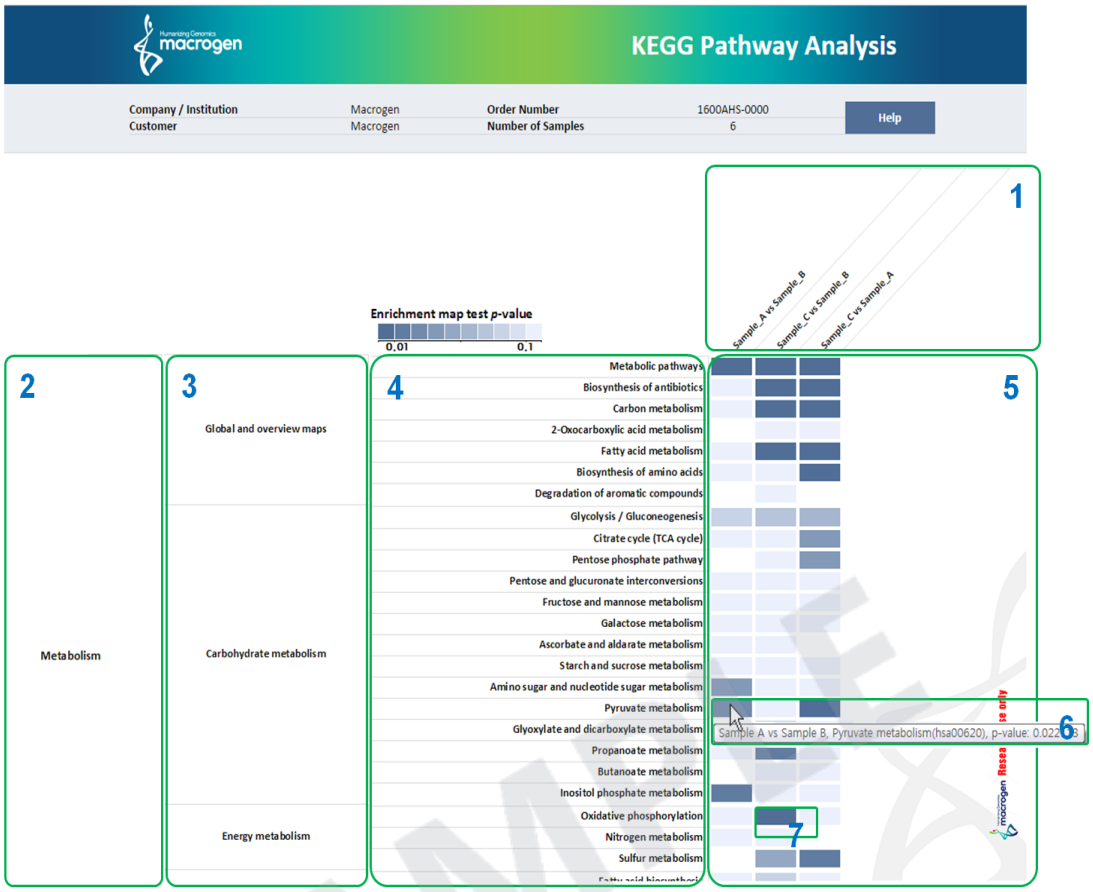


그림 15. KEGG Viewer 구조 설명

- 상자 1: DEG에 사용된 조합 목록
- 상자 2: KEGG pathway 최상위 category
- 상자 3: KEGG pathway 차상위 category
- 상자 4: Pathway map 이름
- 상자 5: KEGG enrichment map score (p-value) 에 대한 heatmap (empty = none mapped gene)
- 상자 6: Mouse를 올려두면 다음과 같은 정보들이 나타남 (조합 이름, pathway 이름, KEGG enrichment map score (p-value).
- 상자 7: Color 상자를 클릭하면 새로운 창에 해당 pathway map이 나타남

- "Global map and overview map"은 KEGG homepage에 자동으로 표시되기때문에 로딩 시간이 다소 걸릴수 있습니다.

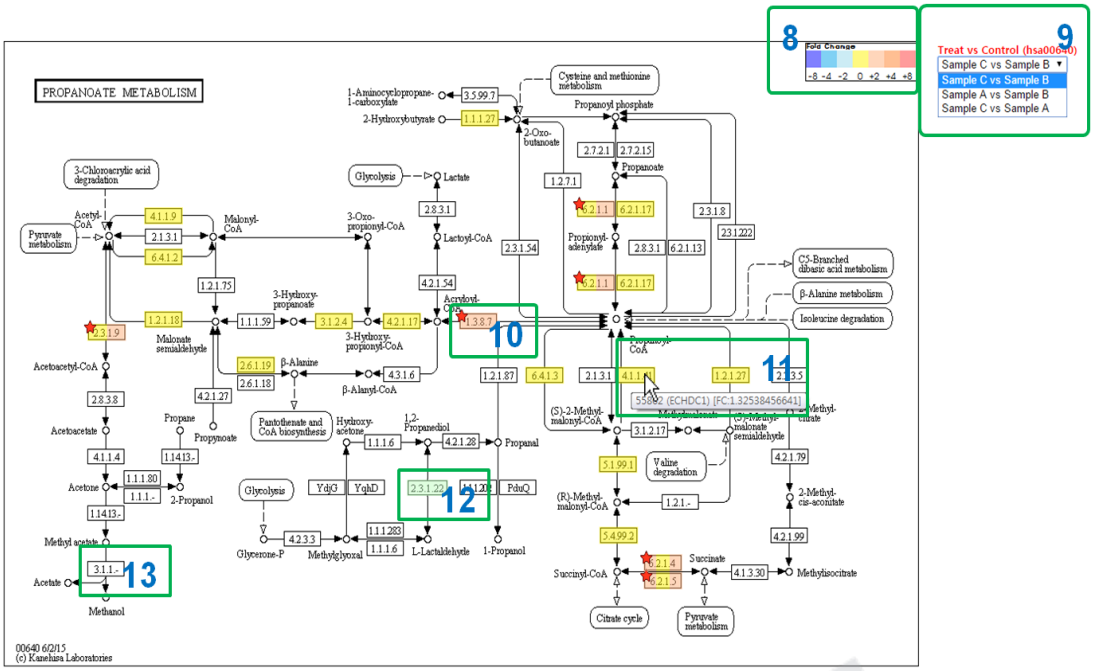


그림 16. KEGG pathway map 구조 설명

- 상자 8: 상자 9에 선택된 조합의 fold change값이 color key 기준으로 표현 (파란색으로 갈 수록 down regulated gene이 mapping되었다는 의미이며, 빨간색으로 갈수록 up regulated gene이 해당 module에 mapping 되었다는 것을 의미).
- 상자 9: 선택 상자를 이용하여 다른 비교조합의 동일한 pathway map으로 이동
- 상자 10: 해당 pathway map에 mapping 되면서, processed data (data2)에 속하는 경우는 상자 8의 color key를 기준으로 fold change가 각 module unit에 표현되며, 이 중, 유의하게 선별된 DEG (data3)를 포함한 module은 빨간색 별이 추가되어 표시 (하나의 module unit에 여러 유전자가 mapping되는 경우, 유전자 갯수 만큼 unit크기를 나눠 fold change color가 표시)
- 상자 11: 각 module에 마우스를 올리면, mapping되는 유전자 및 fold change정보가 나타남 (만약 gene id는 표시 되지만, fold change가 없는 경우는 processed data (data2)에는 존재하지 않은 유전자)
- 상자 12: 초록색으로 표시된 module은 species에 존재하는 유전자이기는 하나, 해당 조합에서는 발현을 확인할 수 없었던 유전자로 구성된 module인 경우를 의미
- 상자 13: 색이 없는 module은 해당 species에 존재하지 않는 유전자들로 구성된 module인 경우를 의미.

5. 5. 2. KEGG_stat Sheet

아래 table은 pathway map별로 modified fisher's exact test를 이용하여 gene-set enrichment analysis를 수행한 예시 결과 입니다. 본 분석 결과는 각 조합별 data3파일의 "KEGG_stat" sheet에서 확인 하실 수 있습니다.

Pathway map enrichment analysis 결과 예시

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 이름
- Number_of_SigGenes: data3 데이터 기준으로 해당 KEGG map에 매칭된 중복이 없는 유전자 갯수
- Genes: 해당 KEGG map에 매칭된 유전자 리스트, Number_of_SigGenes 갯수 만큼 존재함.
- Sig.NotIn.KEGG: data3 중 KEGG map에 매칭되지 않는 유전자 갯수
- Genome.In.KEGG: KEGG database에 존재하는 해당 species의 전체 유전자 갯수 중 KEGG map에 매칭된 유전자 갯수
- Genome.NotIn.KEGG: KEGG database에 존재하는 해당 species의 전체 유전자 갯수 중 KEGG map에 매칭되지 않은 유전자 갯수
- PValue: Modified fisher's exact test를 통하여 도출된 raw p-value
- Bonferroni: Bonferroni 방식으로 보정된 p-value
- FDR: FDR 방식으로 보정된 p-value

5. 5. 3. KEGG_genes Sheet

아래 table은 유전자별 pathway map enrichment analysis 결과를 정리한 예시 입니다. 본 분석 결과는 각 조합별 data3파일의 "KEGG_genes" sheet에서 확인 할 수 있습니다. 해당 sheet에서 엑셀 필터기능을 이용하여 관심 유전자의 enriched KEGG map을 filtering 하실 수 있습니다.

유전자 기준으로 정리된 KEGG pathway map enrichment analysis 결과 예시

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 ac	0.002975034	0.62475714	0.009054451	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	05410	Hypertrophic ca	9.33482E-05	0.01960313	0.000502644	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	05412	Arrhythmogenic	0.017901038	1	0.042238405	ITGA11	1.706859	11.100807	10.721833	11.493176
22801	05414	Dilated cardiom	0.002059901	0.432579199	0.006655065	ITGA11	1.706859	11.100807	10.721833	11.493176
3017	05034	Alcoholism	8.28056E-07	0.000173892	6.68814E-06	HIST1H2BD	1.647010	11.092905	10.738818	11.458667
3017	05203	Viral carcinoger	2.52581E-05	0.005304204	0.000156006	HIST1H2BD	1.647010	11.092905	10.738818	11.458667
3017	05322	Systemic lupus	2.5681E-06	0.0005393	1.85966E-05	HIST1H2BD	1.647010	11.092905	10.738818	11.458667
441024	00670	One carbon poo	1	1	1	MTHFD2L	1.747046	9.561974	9.167981	9.972899
441024	01100	Metabolic pathv	5.97272E-15	1.25427E-12	1.79181E-13	MTHFD2L	1.747046	9.561974	9.167981	9.972899
89853	04144	Endocytosis	0.033602909	1	0.075877535	FAM125B	1.677441	9.607461	9.241573	9.987835
7869	04360	Axon guidance	0.005283715	1	0.014994327	SEMA3B	-2.103133	8.787416	9.340035	8.267495
10135	00760	Nicotinate and	8.87463E-05	0.018636723	0.00049044	NAMPT	1.620452	10.752957	10.410395	11.106791
10135	01100	Metabolic pathv	5.97272E-15	1.25427E-12	1.79181E-13	NAMPT	1.620452	10.752957	10.410395	11.106791
534	00190	Oxidative phosp	1	1	1	ATP6V1G2	-1.647407	8.093609	8.461714	7.741517
534	01100	Metabolic pathv	5.97272E-15	1.25427E-12	1.79181E-13	ATP6V1G2	-1.647407	8.093609	8.461714	7.741517
534	04145	Phagosome	3.15039E-07	6.61582E-05	2.87644E-06	ATP6V1G2	-1.647407	8.093609	8.461714	7.741517

- InID: Matching key ID (ex. Entrez GeneID)
- MapID: KEGG map ID
- MapName: KEGG map name
- PValue: Modified fisher's exact test를 통하여 도출된 raw p-value
- Bonferroni: Bonferroni 방식으로 보정된 p-value
- FDR: FDR 방식으로 보정된 p-value

6. SNP 및 Indel 변이 분석

6. 1. SNP 및 Indel 변이 발굴

(경로: result_RNAseq/Variant_calling/STAR_GATK/VCF_files/*.rawVariants.vcf 참고)

변이 발굴 (SNV calling) 작업은 STAR 프로그램으로 cDNA sequence reads를 genomic DNA reference에 mapping 한 후, Mark duplication & sort 과정으로 진행됩니다. 이후 Split 'N' trim 및 base recalibration 과정을 통하여 분석 가능한 mapped reads를 만들고 HaplotypeCaller를 이용하여 variant calling을 수행합니다.

6. 2. SNP, Indel 변이 filtering 및 annotation 추가

(경로: result_RNAseq/Variant_calling/STAR_GATK/SNV_Call_*.xlsx 참고)

각 샘플 별 변이 발굴의 결과를 가지고 GATK의 VariantFilteration (Fisher Strand values (FS 〉 30.0) and Qual By Depth values (QD 〈 2.0) 및 Depth (DP 〉= 10)) 조건을 만족하는 variant를 선별하였습니다.

선별된 variant에 주석을 달기 위해 dbSNP, 1000 Genome Project database, ESP6500, SIFT database, CLINVAR 등 다양한 database를 참조하여 SnpEff, SnpSift를 수행했습니다.

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

아래는 6샘플의 SNV 요약 결과입니다.

표 11. SNV 빈도 요약

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%

개인별 SNV 결과는 vcf파일과 엑셀 파일로 제공 되며, 일반적인 vcf파일의 컬럼 예시는 아래와 같습니다.

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: 식별자 (dbSNP에 존재하는 variant인 경우는 rs#로 표시됨.)
- REF: 해당 position에 대한 reference sequence
- ALT: 적어도 한 샘플에서 call된 non-reference sequence
- QUAL: Phred scaled quality score, SNP quality 높은 QUAL 점수를 가지면 신뢰성이 높은 call이라 할 수 있음.
- FILTER: 'PASS'는 해당 position의 call이 filter 기준 (Fisher Strand values (FS 〉 30.0) and Qual By Depth values (QD 〈 2.0))을 만족했다는 의미임
- INFO: 세미콜론으로 추가적인 position에 대한 정보를 줄 수 있음 (vcf 생성에 따라 상이함)
  - NS: Number of Sample with Data
  - DP: Total depth
  - AF: Allele Frequency
  - AA: Ancestral Allele
  - DB: dbSNP 존재 유무
  - H2: HapMap2 존재 유무
- FORMAT: 샘플 column에 대한 데이터 포멧을 GT(Genotype):GQ(Genotype Quality):DP(Read Depth):HQ(Haplotype Quality) 순으로 나타냄.
- Sample Name: Sample에 대한 genotype 정보를 FORMAT column의 정보 순으로 나타냄.

개인별 발굴된 SNV 결과를 vcf파일뿐만 아니라 dbSNP, 1000 Genome Project database, ESP6500, SIFT database, CLINVAR 등 다양한 database의 정보를 추가하여 엑셀파일로 저장하였습니다.

예시는 전체적인 데이터를 축약하여 표현한 표입니다. 자세한 정보는 링크된 PDF 파일을 참조해주세요.

LINK AnnotDescription.pdf

표 12. 개인별 발굴된 SNV의 주석 달기 예시

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. 융합유전자(Fusion Gene)예측 결과

7. 1. Defuse 분석 결과

(경로: result_RNAseq/Fusion_gene_analysis/DEFUSE/ 참고)

Defuse프로그램을 사용하여 융합유전자(fusion gene)를 예측하였습니다. Defuse 프로그램은 서로 일치하지 않는 paired-end alignments(spanning reads와 split reads)를 clustering하는 방법으로 융합유전자의 결합 부위를 예측하고, 이 결과에 heuristic filter를 적용하여 real fusion gene의 여부를 분류합니다.

표 13. 예측된 융합유전자 리스트 예시

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	ENSG00000092820
Gene2	ENSG00000167193	ENSG00000058335	ENSG00000167193	ENSG00000058335
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: 해당 fusion gene이 발견된 샘플
- Split_Sequence: Fusion sequence를 나타내며 '|' 구분으로 융합유전자의 경계를 나타내며, 두 융합유전자의 sequence를 확인 할 수 있음.
- Split_Count: 한쪽 부분이 유전자의 경계에 align되고, 나머지 끝부분은 align이 안 되는 read 수
- Span_Count: Paired-ends가 서로 다른 유전자에 align이 되는 read 수
- Gene1, Gene2: Gene1, Gene2의 ensembl ID
- Gene1_Name, Gene2_Name: Gene1, Gene2의 유전자 이름
- Gene1_Description, Gene2_Description: Gene1, Gene2의 유전자 설명
- Gene1_Strand, Gene2_Strand: 유전자 방향
- Gene1_Chr, Gene2_Chr: Gene1, Gene2가 위치한 Chromosome
- Gene1_Start, Gene2_Start, Gene1_End, Gene2_End: 두 유전자의 start, end position
- Genomic_Strand1, Genomic_Stand2: Fusion splice/breakpoint의 각 유전자별 genomic strand
- Genomic_Break_Position1, Genomic_Break_Position2: Fusion splice/breakpoint의 각 유전자별 genomic position

- Probability: Fusion gene으로 분류될 확률을 나타냄. 이 값이 높을수록 fusion gene일 가능성이 높음.

7. 2. FusionCatcher 분석 결과

(경로: result_RNAseq/Fusion_gene_analysis/FusionCatcher/ 참고)

FusionCather는 RNA-seq data로부터 이미 알려져 있거나 새로운 somatic fusion genes, translocations, chimeras를 찾는 도구입니다. 기본적으로 read를 low quality filtering/trimming 한 후 정상적인 RNA fragment size distribution에서 벗어나는 read를 대상으로 이미 알려진 exon과 intron 정보를 이용하여 최소한 2개의 gene 모두 ENSEMBL Database에 존재하는 fusion junction을 찾습니다.

표 14. 예측된 융합유전자 리스트 예시

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	ATTTACAAACTGGCCA AGAAGGGCCTTACTCC TTCACAGATCG*CCATA ACCAGCAGACCACAG CATTCAGGCATCCTGT GACGGGA	GGGGATCGAGGATAC TCACCAGAAACCGAAA ATGCCGAAACCA*CAT TACTTCACGGTGAACTT CAGCCATGAGAACCA GAAAGCCT	ATTTACAAACTGGCCA AGAAGGGCCTTACTCC TTCACAGATCG*CCATA ACCAGCAGACCACAG CATTCAGGCATCCTGT GACGGGA	TGTTTTCGGGGATCGA GGATACTCACCAGAAA CCGAAAATGCCGAAA CCA*CATTACTTCACGG TGAACTTCAGCCATGA GAACCAGAAAGCCTTG GAGCT
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	MGRMHAPGKGLSQSA LPYRRSVPTWLKLTSD DVKEQIYKLAKKGLTPS QIAITSRPQHSGIL; ...	MPKPHYFTVNFSHENQ KALELRTEDAKDCDEW VAAIAHASYRTLA...DQ SFVMDEESLYESSLRIE PKLPT; ...	MGRMHAPGKGLSQSA LPYRRSVPTWLKLTSD DVKEQIYKLAKKGLTPS QIAITSRPQHSGIL; ...	MPKPHYFTVNFSHENQ KALELRTEDAKDCDEW VAAIAHASYRTLA...DQ SFVMDEESLYESSLRIE PKLPT; ...

- Sample: 해당 fusion gene이 발견된 샘플
- Gene_1_symbol, Gene_2_symbol: Fusion partner의 양쪽 Gene의 symbol
- Fusion_description: Fusion gene의 type
- Counts_of_common_mapping_reads: Fusion gene의 양쪽 gene 모두에 동시에 mapping 되는 read의 수
- Spanning_pairs: Fusion을 설명할 수 있는 pair-end reads의 수
- Spanning_unique_reads: Fusion junction에 unique하게 mapping된 read의 수
- Longest_anchor_found: Fusion junction에 unique하게 mapping된 read 중에 가장 긴 anchor
- Fusion_finding_method: Read를 align 할 때 사용했던 방법
- Fusion_point_for_gene_1, Fusion_point_for_gene_2: Fusion junction의 5' end 와 3' end 의 Chromosomal position; 1-based coordinate

- Gene_1_id, Gene_2_id: Fusion partner의 양쪽 Gene의 Ensembl gene id
- Gene_1_Description, Gene_2_Desciption: Fusion partner의 양쪽 Gene의 유전자 설명
- Exon_1_id, Exon_2_id: Fusion partner의 양쪽 Gene의 Ensembl exon id
- Fusion_sequence: Fusion junction으로 유추되는 sequence (asterisk 는 junction point를 의미함)
- Predicted_effect: Ensembl database로 부터 얻은 annotation 정보를 사용하여 예측된 fusion gene이 미치는 영향
- Predicted_fused_transcripts: 해당 gene들로 이루어진 모든 가능한 fused transcripts
- Predicted_fused_proteins: 모든 가능한 fused transcripts로 부터 예측되는 아미노산 서열

7. 3. Arriba 분석 결과

(경로: result_RNAseq/Fusion_gene_analysis/Arriba/ 참고)

Arriba 프로그램을 사용하여 융합유전자(fusion gene)를 예측하였습니다. Arriba 프로그램은 STAR 기반으로 alignment 진행 후, fusion gene 후보 내에서 Read-level filters, Event-level filters를 이용하여 fusion의 가능성 있는 후보를 제공합니다.

표 15. 예측된 융합유전자 리스트 예시

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	GCGGGAGCTGAGGCGGGAGCGGACAGG CTGGTGGGCGAGCGAGAGGCGGCGGAA TGGTGGACTACCACGCGGCGAACCAGTC GTACCAGTACGGCCCCAGCAGCGCGGG CAATGGCGCTGGCGGCGGGGGCAGCAT GGGCGACTACATGGCCCAGGAGGACGA CTGGGACCGGGACCTGCTGCTGGACCCG GCCTGGGAGAAGCAGCAGCGCAAG|TGT TACCTGTTTCACATGTACGTGGGTGTCCG GGCTGGCGGAGGCATTGGGGACGAGAT CGAGGACCCCGCGGGTGACGAATACGA GCTCTACAGGGTGGTCTTCGACATCACCT TCTTCTTCTTCGTCATCGTCATCCTGTTGG CCATCATCCAGG__GTCTGATCATCGACG CTTTTGGTGAGCTCCGAGACCAACAAGA GCAAGTGAAGGAGGATATGGAG__ACCA AGTGCTTCATCTGTGGAATCGGCAGTGAC TACTTTGATACGACACCGCATGGCTTCGA GACTCACACGCTGGAGGAGCACAACCTG GCC	CTCCCACTTTGGCCTTCCAAAATGTTGCG ATTATAGGCGTGAGCCACTGTGGCTGGC CTGAAATTTTCTAGTATCCACATTCATAAA GTAAAAAGAAAATAAAAAG|GGAAATAAA TGAAGGAAGACAAACATATATGCTTGGAT TAATGAGGAGTTTTCCCTTCCATCTTCCAT CAGCTTCGATTGTAATGAAAATTTTACTGT AGAGAATCTAGCAAGGAAGAAATGACAA TGATTCCCTCACTCAACAAGTATTTGGG	CAACAATAGAAACAGAACAAAAAGAGGA CAAAGGT__GAAGATTCAGAGCCTGTAAC TTCTAAGGCCTCTTCAACAAGTACACCTG_ _AACCAACAAGAAAACCTTGGGAAAGAA CAAATACAATGAATGGCAGCAAGTCACCT GTTATCTCCAG|CATTTGTCCCTGGAGGGT CTCTGAAAGTCCAGGTCAGCCCCTGGGC TGGGTCCCCACAGTAAGAAGAGAACTGT GATGGGCAACACCCCAGAAAAGAAGACT TGCAGCCTCACTTCAGGTCAATTTGCAAG AACTGACATCACACAGCAG	ACCTGGAAGAACCTGAAACAAATCCTCG CTTCTGAAAGGGCATTGCCGTGGCAACTG AACGATCCTAACT__ACTTCAGTATTGATG CTCCTCCATCCTTTAAGaCA...TCAGGTCTG CTT|GGTCTCACTCCATCACCCAGGCTGG AGTGCAGTGGTGCCATCTTGGCTCACTGC AACCTCCCTCTTCCAGGCTCAAGCAATCC TCTCACCTCAGCCTCCCTATAGCTGGACT ACAGaCACGCACCACCACACCTGGATAA T__GAACCACCAAACTGTTTTCCACAGAG GCTGCATCAATTGACATTCCCAC
reading_frame	in-frame	.	out-of-frame	out-of-frame
peptide_sequence	MVDYHAANQSYQYGPSSAGNGAGGGGS MGDYMAQEDDWDRDLLLDPAWEKQQRK| CYLFHMYVGVRAGGGIGDEIEDPAGDEYELY RVVFDITFFFFVIVILLAIIQGLIIDAFGELRDQQ EQVKEDMETKCFICGIGSDYFDTTPHGFETH TLEEHNL		TIETEQKEDKGEDSEPVTSKASSTSTPEPTRK PWERTNTMNGSKSPVISs|icpwrvsespgqplg wvptvrrel*	SGLL|gltpsprlecsgailahcnlplpgssnpltsasl*

- Sample ID: Sample 이름
- Gene1, Gene2: 해당 fusion gene에 대한 gene1/gene2의 gene symbol. Gene1은 fusion transcript의 5'쪽의 유전자를 나타내며, Gene2는 fusion transcript의 3'쪽의 유전자를 나타냄.
  - Intergenic region의 fusion인 경우, upstream과 downstream 내에 가장 가까이 있는 유전자를 표시하며, 괄호 안에 해당 fusion point와 해당 유전자 간의 떨어진 거리를 표시 함. 예를 들어, ZNF23(1396), ZNF19(10425)의 형태로 표시되면, 해당 fusion gene의 upstream 1,396bp 부근에 ZNF23 유전자가 존재하며, 해당 fusion gene의 downstream 10,425bp 부근에 ZNF19 유전자가 존재한다는 의미.
- Strand1, Strand2: 해당 fusion gene에 대한 gene1/gene2의 strand 정보.
  - (Gene/Fusion) 열에 (+/-)로 기재된 경우에는 기존의 gene이 genome 상에서는 (+) Strand에 존재하지만, fusion gene 서열은 (-) Strand로 전사됨을 의미함.
- Breakpoint1, Breakpoint2: 해당 fusion gene에 대한 gene1/gene2 기준 breakpoint.
- Site1, Site2: gene1/gene2 의 breakpoint가 있는 위치 정보(Splice-site, exon, intron, 5'UTR,

3'UTR, UTR, intergenic)를 나타냄.

- Type: Fusion을 설명하는 reads의 방향과 breakpoints 위치정보를 기반하여, 해당 fusion의 type (translocation, duplication, inversion, deletion)이 결정됨.
- Direction1, Direction2: gene1/gene2 기준 fusion partner의 방향 정보
  - Downstream: partner gene이 fusion breakpoint의 downstream 에서 융합되어 있음. 즉, partner gene이 fusion breakpoint 보다 높은 좌표 값에서 융합 됨.
  - Upstream: partner gene이 fusion breakpoint의 upstream에서 융합되어 있음. 즉, partner gene이 fusion breakpoint 보다 낮은 좌표 값에서 융합 됨.
- Split read1, Split read2: Fusion 을 설명하는 reads 수로, 각 fusion gene에 mapping된 split reads 수 (reads 수는 split reads의 더 긴 segment로 align되는 gene 기준으로 count됨.)
- Discordant mates: Fusion을 설명하는 discordant mate를 이루는 read의 수(spanning reads or bridge reads)를 나타냄.
- coverage1, coverage2: breakpoint1 및 breakpoint2 근처의 gene coverage 를 나타냄.
- confidence: 각 fusion에 대한 예측의 신뢰도로써 다음과 같이 고려하여 low, medium, high으로 구분됨.
  - supporting reads의 수(fusion을 설명하는 reads 수이며, split_reads1, split_reads2, discordant_mates를 뜻함.), split reads와 discordant mates의 균형, breakpoints사이의 거리, fusion의 event type, breakpoints의 intragenic여부, multiple isoform과 balanced translocations의 정보 등을 사용.

  (자세한 내용은 interpretation-of-results 참조)
- filters: 실제 fusion을 설명할 수 있는 reads 수는 split_reads1, split_reads2, discordant_mates, filters 항목의 reads수의 합으로 계산할 수 있는데, 여기서 제거된 reads의 사유.
- fusion_transcript: Fusion된 transcript sequence 를 나타내며 breakpoint는 파이프 기호 (|)로 표시됨.
  - 소문자: SNPs 나 SNVs를 나타냄.
  - "[", "]" 사이의 문자: Insertions
  - (-): Deleted bases
  - (___): Underscore 3개는 introns
  - (...): Coverage가 충분하지 않아 누락 된 정보
  - (?): reference서열과의 mismatch와 같은 모호한 위치
- reading_frame: Fusion gene의 3' end가 in-frame 으로 fusion이 되었는지, 또는 out-of-frame으로 fusion이 되었는지를 나타냄.
- peptide_sequence: peptide sequence를 나타내며 breakpoint는 파이프 기호("|")로 표시됨.

예측된 fusion에 대하여 chromosomal ideogram 상의 fusion 위치 및 annotation 정보를 아래의 그림 17과 같이 시각화합니다. 결과 폴더 내 PDF파일로 제공되며, 각 fusion 별로 한 페이지의 그림으로 생성됩니다.

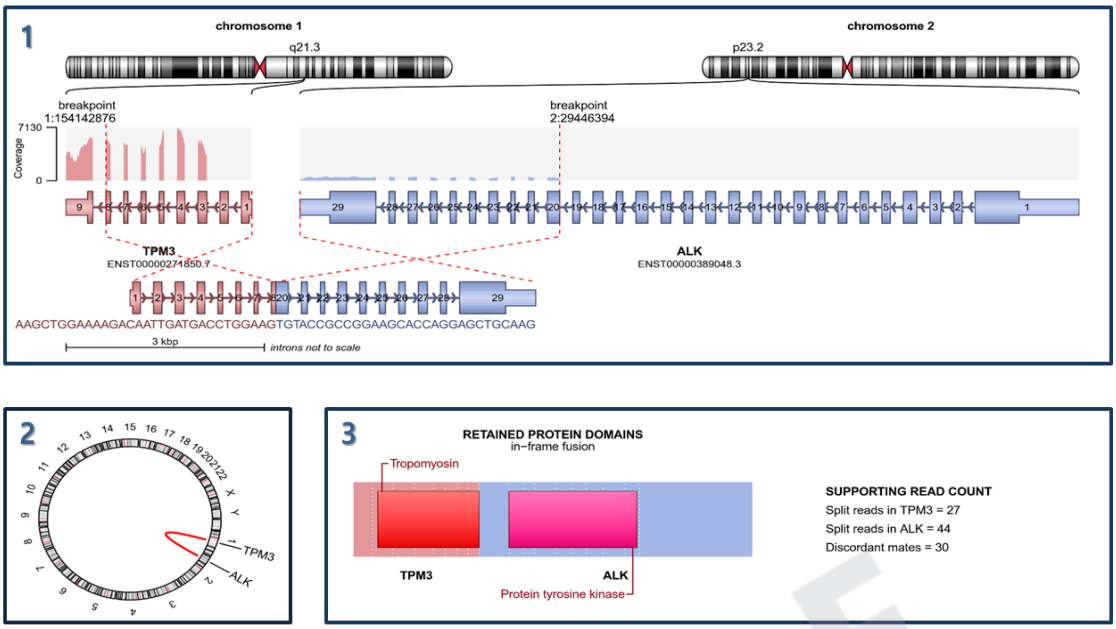


Figure 17. 예측된 융합유전자 예시

1. Fusion gene의 형태를 보여주며 fusion gene의 기본 정보(chromosome, transcript, coverage, sequence, break point)등을 나타냄.
2. CircosPlot으로 Fusion된 gene과 chromosome 위치 정보를 표시함.
3. 해당 fusion gene과 연관된 protein domain정보가 표시되며, fusion을 설명하는 supporting reads로 split reads와 discordant mates 별 read count가 표시됨. 연관된 protein domain이 없는 경우 공란으로 표시 됨.
   - Split reads in [geneA], Split reads in [geneB] : 각 유전자 상의 split 된 read의 수
   - Discordant mates: Discordant mate를 이루는 read의 수 (spanning reads or bridge reads)

8. 다운로드 안내

8. 1. Raw 데이터

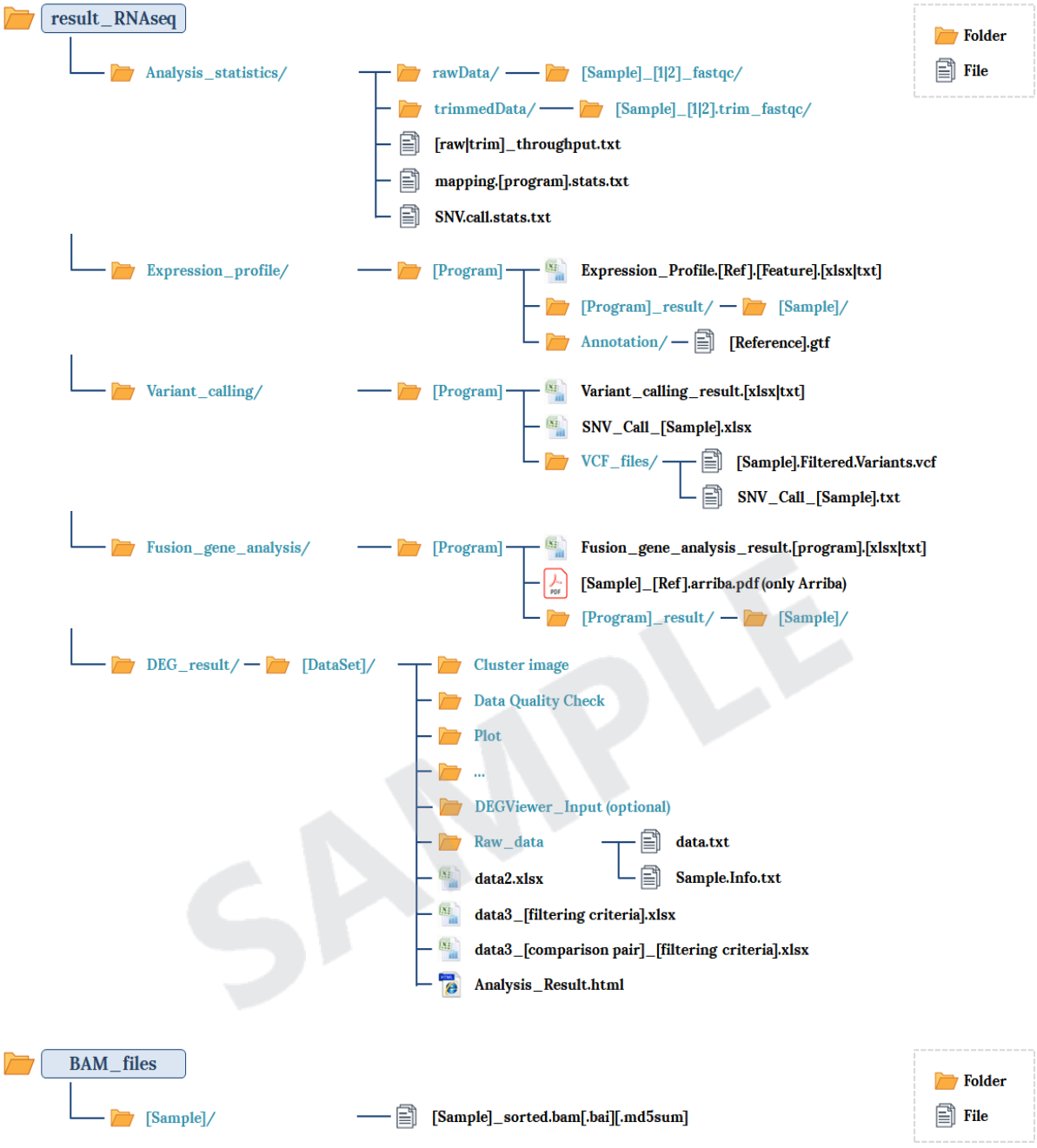
Raw 데이터는 adapter sequence가 제거되지 않은 FASTQ 파일입니다.

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 : 분석에 사용된 FASTQ가 압축된 파일입니다.
- md5sum : 파일의 온전성을 확인 하기 위하여 md5sum을 사용하였습니다. 만약 md5sum 값이 일치 한다면, 이 파일은 위조, 변경, 누락되지 않은 파일입니다.

8. 2. 분석 결과

Download link	File size
HN00000000_result_RNAseq.zip (md5sum: 1bc94df3a83b78ff34a0ce1ead0fb862)	1.18G
HN00000000_BAM_files.tar (md5sum: 77a63ec829467fab324c6d5cd21b8008)	43.6G



⚠ 데이터 보관기간은 3개월이며, 장기보관이 필요하신 경우
대표메일(ngskr@macrogen.com) 또는 담당자에게로 문의 바랍니다.

9. 부록

9. 1. 프레드 품질 점수표

프레드 품질 점수는 각각에 해당하는 염기가 얼마나 정확한지를 숫자로 나타낸 것으로 Q숫자가 클수록 해당 염기의 정확도가 높습니다. Q20은 wrong base probability가 1% 이고 Q30은 wrong base probability가 0.1% 입니다. 아래는 프레드 품질 점수표입니다.

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

프레드 품질 점수 Q는 $-10\log_{10}P$로 계산합니다. 여기서 P는 base call 오류 확률을 나타냅니다.

9. 2. 분석에 사용된 프로그램

9. 2. 1. FastQC

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

FastQC는 raw sequence에 대해 분석을 진행하기 전에 데이터에 문제가 없는지 quality check를 해 주는 프로그램입니다. 주된 기능들은 BAM, SAM, FastQ files을 import 하는 것과 어떤 곳에서 문제가 있었는지 quick overview를 제시해 주고, 요약된 그래프와 테이블 결과를 html 파일로 제공합니다.

9. 2. 2. Trimmomatic

LINK http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic은 illumina paired-end 또는 single-end에 대해 다양한 parameter로 trimming 작업을 수행하는 프로그램입니다. Trimming 단계에 제공되는 parameter들은 아래와 같습니다.

- ILLUMINACLIP: Read로부터 adapter 나 특정 sequence를 잘라냄
- SLIDINGWINDOW: Sliding window trimming을 수행합니다. Window 내 평균 quality가 threshold 미만인 경우 잘라냄
- LEADING: Threshold 미만인 경우, read의 start부분의 base를 잘라냄
- TRAILING: Threshold 미만인 경우, read의 end부분의 base를 잘라냄
- CROP: 특정 length로 read를 자름
- HEADCROP: Read의 start로부터 특정 base수 만큼 자름
- MINLEN: 특정 length 미만이면 해당 read를 drop
- TOPHRED33: Quality score를 phred33 score로 변환
- TOPHRED64: Quality score를 phred64 score로 변환

9. 2. 3. TopHat version v2.1.1, Bowtie2 2.5.1

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

Tophat은 전사체 시퀀싱 데이터를 bowtie를 이용하여 mammalian-sized genome에 read를 맵핑 시켜주는 도구입니다. 또한 맵핑 한 결과를 가지고, 잠정적인 엑손 위치와 엑손 결합부위(exon junction)를 찾아주는데, 엑손 결합부위 맵핑의 정확도 향상을 위해 splicing 위치에 있는 GT-AG 의 2개 염기 패턴 인식 정보를 고려합니다.

9. 2. 4. STAR 2.6.0c

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

STAR는 다량의 전사체 시퀀싱 데이터를 reference의 transcript 지역에 splicing 하여 맵핑하는 도구로써, uncompressed suffix arrays 를 만든 다음, seed clustering 과 stitching procedure를 이용하여 sequential maximum mappable seed search를 하는 RNA-seq alignment algorithm 이 사용되었습니다.

9. 2. 5. Cufflinks version v2.2.1

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

Cufflinks는 엑손 결합 부위에 대한 서열 정렬이 가능한 aligner 의 맵핑 결과를 입력으로 받아 단편 서열들을 이어 붙이는 서열 조립(sequence assembly) 프로그램으로, assemble된 전사체에 대한 발현 량 정도를 추정할 수 있으며 samples간의 발현 값의 차이를 확인하는 툴(cuffdiff) 등의 결과를 제공합니다.

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

GATK는 germline DNA와 RNA-seq data에서 SNP과 indels을 찾아내는 도구입니다. GATK best pratice에 따른 step-by-step으로 분석이 진행되며 STAR 2-pass mapping, Picard MarkDuplicate, Split 'N' Trim, Realignment, Base recalibration 의 과정이 포함되어 있습니다. Variant calling은 HaplotypeCaller로 수행됩니다.

9. 2. 7. SnpEff version 4.3t

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

SnpEff는 genetic variant를 annotation 해 주는 도구로, 유전자의 genetic variant로 인한 단백질 서열의 변화와 이로인한 functional effect를 예측합니다.

SnpEff는 아래와 같은 결과를 생성합니다.

- Variant에 의해 영향을 받은 gene과 transcript
- Variant의 위치
- Variant가 protein 합성에 주는 영향 (예: stop codon 생성)
- 여러가지 database에서 이미 알려져 있는 variant와의 비교

9. 2. 8. Defuse version 0.8.1

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

deFuse는 RNA-Seq data로 융합유전자를 발굴하는 프로그램으로 융합 유전자의 결합부위를 찾기 위해 서로 일치하지 않는 paired-end alignments (spanning reads와 split reads) 를 clustering 하는 방법을 이용하여 예측 가능한 fragment의 length distribution 비해 선택된 spanning reads와 split reads의 fragment길이가 true positive인지 여부를 heuristic filter를 적용하여 융합유전자 여부를 예측합니다.

9. 2. 9. FusionCatcher version 1.00

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

FusionCatcher는 RNA-seq data로부터 이미 알려져 있거나 새로운 somatic fusion genes, translocations, chimeras를 찾는 도구입니다. 기본적으로 read를 low quality filtering/trimming 한 후 정상적인 RNA fragment size distribution에서 벗어나는 read를 대상으로 이미 알려진 exon과 intron 정보를 이용하여 최소한 2개의 gene 모두 ENSEMBL Database에 존재하는 fusion junction을 찾습니다.

9. 2. 10. Arriba version 1.2.0

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

Arriba는 RNA-seq data로부터 fusion gene를 detection 하기위한 command-line tool 이며, 임상 연구 환경에서 사용하기 위해 개발되었습니다. 그렇기때문에 짧은 런타임과 높은 민감도를 기준으로 삼아 설계하였습니다. STAR aligner를 기반으로 하였고, post-alignment 런타임은 일반적으로 2분정도 입니다. Arriba는 STAR를 기반으로 한 다른 fusion detection tool과는 다르게 focal deletions으로 인해 발생하는 fusion을 detection 하기 위해 STAR의 alignIntronMax 매개 변수를 줄일 필요가 없습니다.

9. 3. 참고논문

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

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

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

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

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

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

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

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

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

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

11 MORTAZAVI, Ali, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 2008, 5.7: 621-628.

12 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.

13 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.

14 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.

15 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.

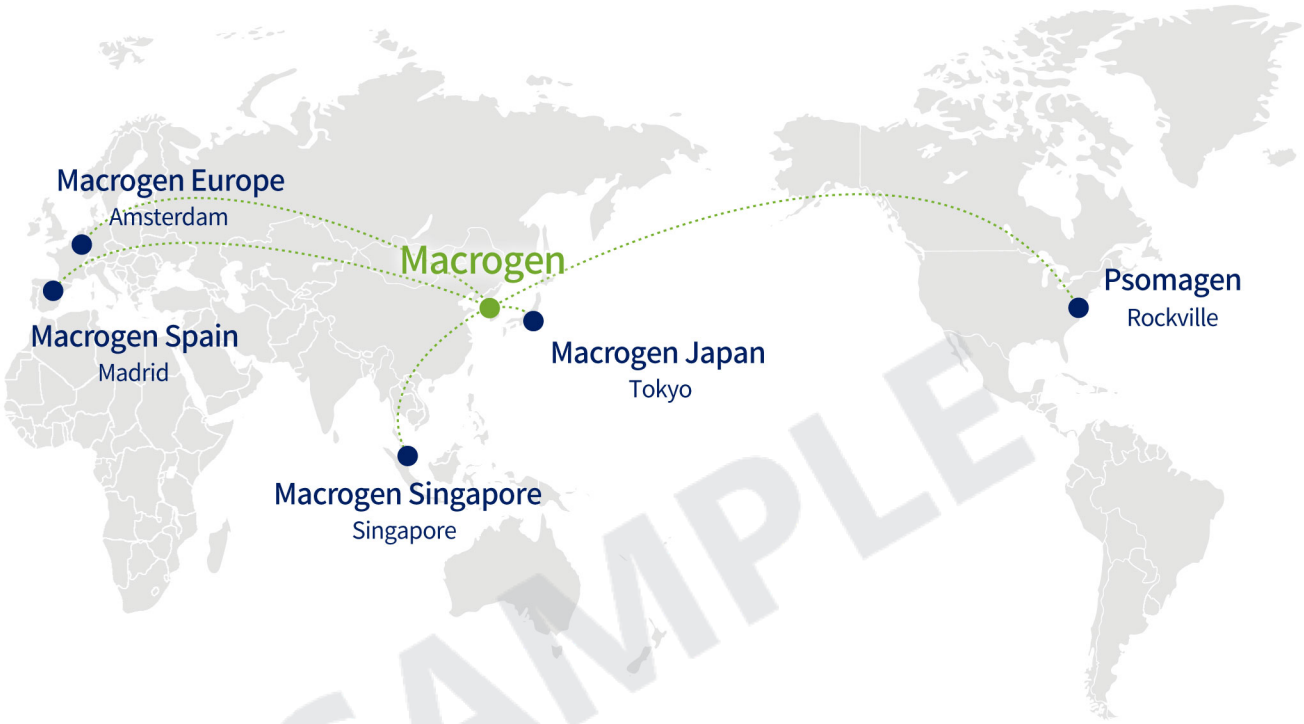
16 MCPHERSON, Andrew, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.

PLoS computational biology, 2011, 7.5: e1001138.

17 NICORICI, Daniel, et al. FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv, 2014, 011650.

18 RAUDVERE, Uku, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research, 2019.

Research use only



HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro, Geumcheon-gu, Seoul, Republic of Korea (Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com (Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 57, 1105 BA, Amsterdam, the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Macrogen Singapore

Laboratory, Business & Support Center

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

Psomagen (Macrogen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville, MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Japan

Laboratory, Business & Support Center

16F Time24 Building, 2-4-32 Aomi, Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas, 28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

Humanizing Genomics
macrogen