

NGS data analysis service of LARC

Hoyoung Ghang

1. Introduction

차세대게놈시퀀싱(NGS, Next Generation Sequencing)법이 많은 생명과학자들에게 도입되면서, 실험동물 마우스에서도 이를 이용한 기법들이 최근 빠르게 사용되고 있습니다. 이러한 NGS는 다음과 같은 Genome(DNA), Exome(DNA), Transcriptome(RNA), Epigenome (후성유전체)등의 다방면으로 연구에 활용되고 있습니다 (McFadden, D. G. et al, 2014; Stephen E Gould et al, 2015).

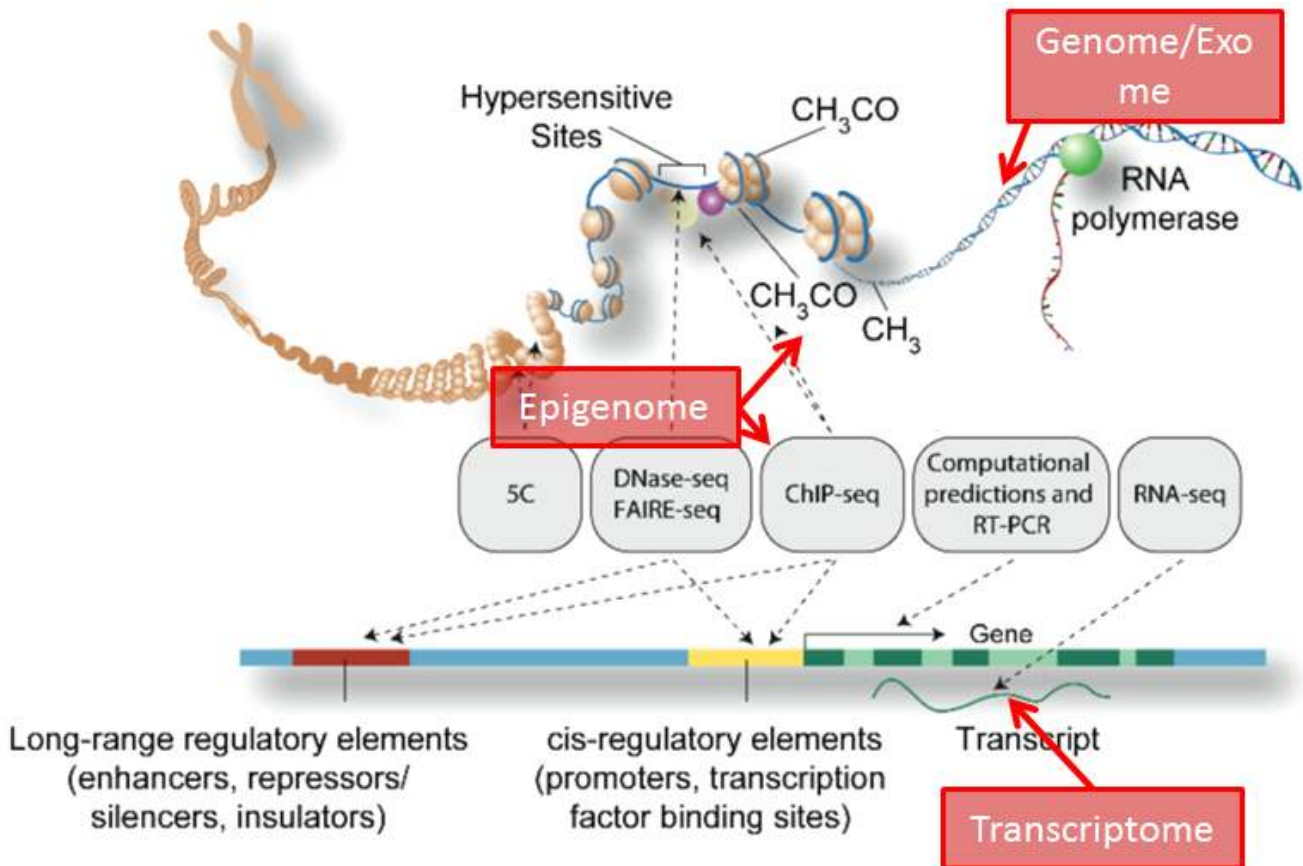


Fig from A User's Guide to the Encyclopedia of DNA Elements (ENCODE) in PLoS Biology.

- Whole genome sequencing (Re-sequencing and De-novo sequencing):** DNA상의 유전적 변이들(SNV, InDel, SV, CNV등)을 확인하기 위한 시퀀싱으로 whole genome이 대상
- Exome sequencing:** DNA상의 유전자부위(gene)을 대상으로 하는 선택적 시퀀싱. 비용절감 가능
- Transcriptome sequencing (mRNA and small RNAs):** RNA를 대상으로 하며, mRNA의 발현 패턴 분석 (Gene expression profile), non-coding RNA(piRNA, tRNA, snRNA, snoRNA, rRNA, microRNA등) 발현 분석등을 위한 시퀀싱
- Epigenome sequencing:** 유전자 발현을 조절하는 DNA-methylation profile 분석을 위한 시퀀싱

이러한 NGS는 본질적으로 대용량 데이터를 생산하며, 이를 정렬하여 원하는 결과를 볼 수 있는 일련의 프로그램 집합체(파이프라인, pipeline)가 필요합니다. 많은 Bioinformatics 연구자들을 중심으로 이를 개발하고, 배포하고 있지만, 실무적으로 쓰이기에는 미흡한 점이 많습니다. 동시에 일반 실험실에서는 전산장비의 부족과 전문적 운용 기술자의 부족으로 이를 분석하기가 쉽지 않은 경우가 많습니다.

실험동물자원센터 (LARC, Laboratory Animal Resource Center, KRIBB)는 다양한 NGS 시퀀싱 플랫폼들을

지원하는 “NGS 데이터 분석지원서비스”를 제공합니다. 이러한 서비스는 wild type을 포함한 genetically engineered mouse, syngeneic mouse, xenograft mouse, aged mouse등을 이용한, 약물, 노령화, 스트레스 등의 특정 제한 환경에서의 DNA, RNA, Epigenome 데이터에 대한 정보분석을 지원하는 것이 주요 목적입니다.

2. Analysis pipelines

정보분석지원서비스는 실험자의 특정 목적을 달성하기 위한 일련의 프로그램들로 이루어져 있습니다. 이러한 일련의 프로그램 집합을 “파이프라인”이라 부르며, 공개된 소프트웨어들과 자체 개발 소프트웨어들의 집합체입니다.

2.1. 정보분석지원서비스 파이프라인의 종류

현재 구현된 파이프라인은 아래와 같습니다. DNA 레벨에서의 게놈 변이 (SNV, InDel, SV, CNV등)를 관측하는 Whole genome sequencing을 기본으로, 시퀀싱의 비용적 문제를 고려하여 유전자 부위(exon)만을 선택적으로 시퀀싱하는 Exome 시퀀싱의 정보분석을 지원합니다. RNA 정보분석은 다양한 파이프라인들로 구성하였으며, 실험 디자인에 따라 재구성이 가능합니다. DNA methylation에 대한 정보를 분석하는 파이프라인은 현재 Medip-Seq 기반으로 구성되어 있으나, 다른 종류의 Chip-Seq도 가능합니다.

☐ DNA applications

- Exome analysis pipeline
- (Whole genome sequencing: resequencing and de-novo sequencing)

☐ RNA applications

- Transcriptome: gene expression analysis
- Transcriptome: variation calling
- Transcriptome: small-RNA analysis
- Transcriptome: alternative splicing analysis

☐ Epigenetic applications

- Epigenome: Medip-Seq analysis

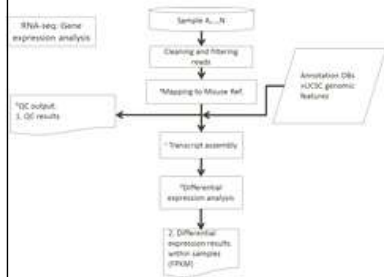


Fig. Workflow used in the RNA-seq gene expression analysis. Each number 1, 2 indicates an output result file. The used tools in each step are as follows: a-bowtie2 and TopHat, b-Quasimapper, c-cuffmerge/cuffdiff, d-cuffdiff, and the pipeline was tuned with perl.

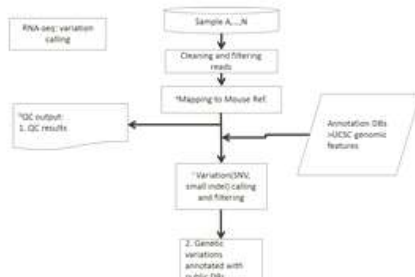


Fig. Workflow used in the variation calling analysis or CNV-seq. Each number 1 and 2 indicates an output result file. The used tools in each step are as follows: a-bowtie2 and TopHat, b-Quasimapper, c-cnvkit/cnvkit, and the pipeline was tuned with perl.

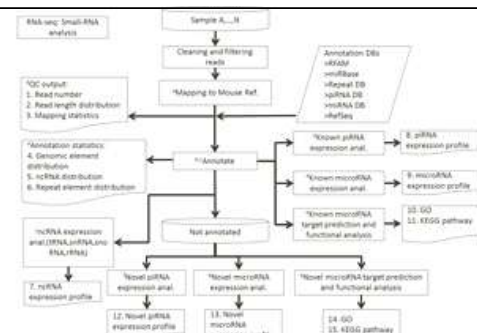


Fig. Workflow used in the small RNA analysis. Each number 1, 2 indicates an output result file. The used tools in each step are as follows: a-bowtie2, b-bowtie2, c-bowtie2, d-bowtie2, e-bowtie2, f-bowtie2, g-bowtie2, and the pipeline was tuned with perl.

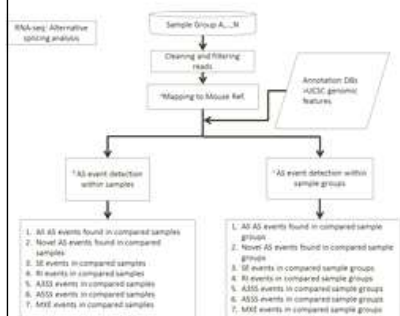


Fig. Workflow used in the alternative splicing analysis. Each number 1, 2 indicates an output result file. The used tools in each step are as follows: a-bowtie2 and TopHat, b-cuffdiff, c-cuffdiff, d-cuffdiff, and the pipeline was tuned with perl and python. The types of AS events - SE, RI, A3SS, A5SS, and MDE - are as follows:

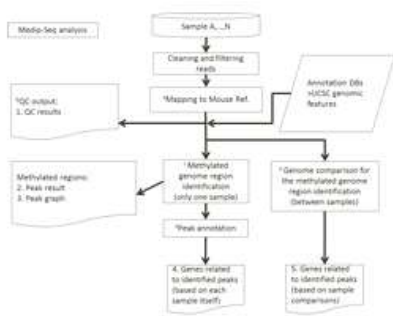
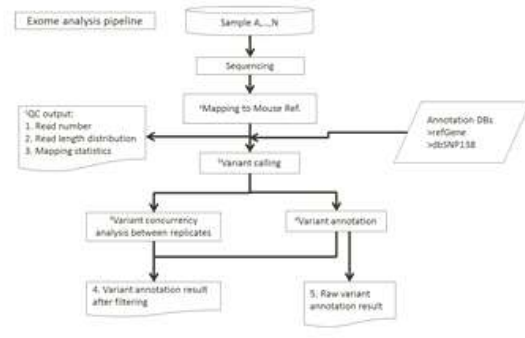


Fig. Workflow used in the Medico-Seq analysis. Each number 1, 2 indicates an output result file. The used tools in each step are as follows: a-bowtie2, b-bowtie2, c-bowtie2, d-bowtie2, e-bowtie2, f-bowtie2, g-bowtie2, and the pipeline was tuned with perl.



Pipeline was controlled with Perl.

- a. Done with TMAT (Ternary Mapping Alignment Program)
- b. Done with TCC (Ternary Variant Caller)
- c. Done with Quasimapper2.0
- d. Done with Hmmer (hidden Markov models)

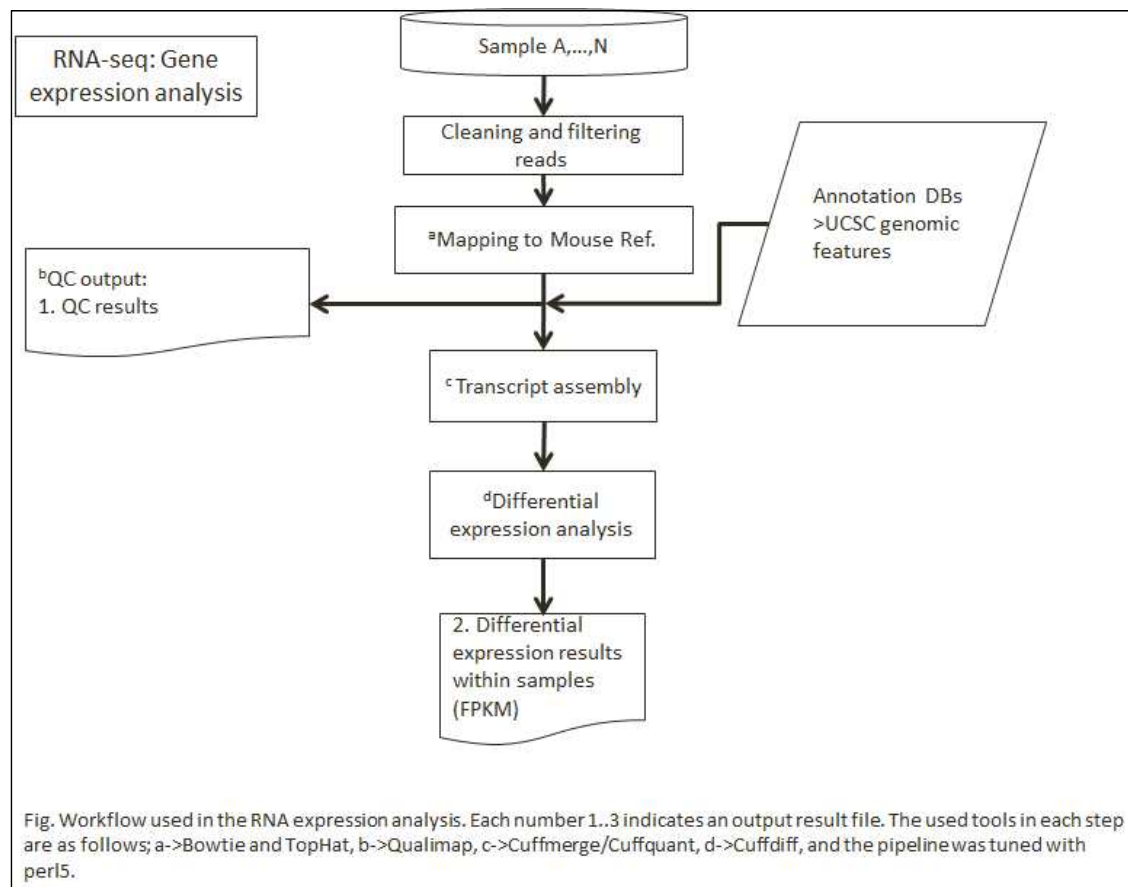
2.2. Transcriptome: gene expression analysis

2.2.1. 분석 목표

- 이 파이프라인은 생체내의 mRNA 발현 패턴을 분석하는 것이 목표입니다. 따라서 특정 상황에서 채취된 샘플의 mRNA 구성이 대조군의 mRNA 구성과 차이나는 부분을 확인해서, 세포수준에서의 환경 적응을 위한 유전자 발현 패턴의 변화를 확인할 수 있습니다. Gene, CDS, primary transcript, isoform 수준에서 조사됩니다.

2.2.2. 구성

- 분석의 과정은 “RNA-seq: Gene expression analysis(그림)”의 과정을 따르는 파이프라인으로 구성되어 있습니다.
- Raw read의 reference genome에의 매핑은 Bowtie와 TopHat이 이용되며, 이후 일련의 과정을 거쳐 RNA 발현량의 차이점이 각 transcript 또는 gene별로 출력됩니다.



2.2.3. 결과의 예제

a. QC results

Mouse reference sequence에의 서열 mapping 결과를 보여줍니다. 기본적인 품은 아래와 같으며, 첨부서류 (Qualimap_report_BAMQC.pdf)를 참고하시면 됩니다.

1. Summary (inside of regions)

1.1. Warnings

Some regions are not loaded	40847 regions were skipped because chromosome name was not found in the BAM file.
-----------------------------	---

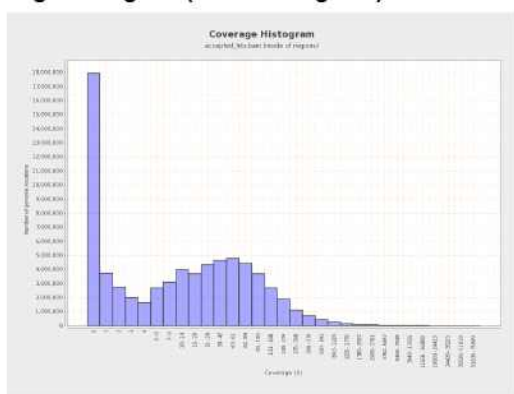
1.2. Globals

Reference size	3,095,693,983
Number of reads	68,044,634
Mapped reads	68,044,634 / 100%
Unmapped reads	0 / 0%
Paired reads	68,044,634 / 100%
Mapped reads, only first in pair	33,951,091 / 49.9%
Mapped reads, only second in pair	34,093,543 / 50.1%
Mapped reads, both in pair	65,993,600 / 96.99%
Mapped reads, singletons	2,051,034 / 3.01%
Read min/max/mean length	90 / 90 / 90
Clipped reads	0 / 0%
Duplication rate	48.54%

1.3. Globals (inside of regions)

Regions size/percentage of reference	141,459,010 / 4.57%
Mapped reads	54,426,965 / 79.99%
Mapped reads, only first in pair	27,159,671 / 39.91%
Mapped reads, only second in pair	27,267,294 / 40.07%

4. Coverage Histogram (inside of regions)



b. Differential expression results

- 기본적으로 다음과 같이 gene, CDS, primary transcript, isoform들 수준에서의 differential expression 결과 값을 제공합니다. 이는 'Cufflinks'의 결과입니다.

isoform_exp.diff	Transcript-level differential expression.
gene_exp.diff	Gene-level differential expression. Tests differences in the summed FPKM of transcripts sharing each gene_id
tss_group_exp.diff	Primary transcript differential expression. Tests differences in the summed FPKM of transcripts sharing each tss_id
cds_exp.diff	Coding sequence differential expression. Tests differences in the summed FPKM of transcripts sharing each p_id independent of tss_id

- 이들은 기본적으로 다음 표와 같은 양식을 따릅니다.

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
XL0C_001847	XL0C_001847	ZZZ3	chr1:78029730-78149222	Liver	Kidney	OK	3.22646	7.62958	1.24165	1.24008	0.00925	0.037508	yes
XL0C_026013	XL0C_026013	ZZEF1	chr17:3907512-4046253	Liver	Lung	OK	20.225	5.75667	-1.81284	-1.91732	0.00025	0.005309	yes
XL0C_026013	XL0C_026013	ZZEF1	chr17:3907512-4046253	Liver	Lung	OK	20.225	7.97655	-1.34231	-1.52567	0.00135	0.021638	yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...

- 위 표의 각 칼럼은 다음의 의미를 가집니다.

Column number	Column name	Example	Description
1	Tested id	XL0C_000001	A unique identifier describing the transcript, gene, primary transcript, or CDS being tested
2	gene	Lypla1	The gene_name(s) or gene_id(s) being tested
3	locus	chr1:4797771-4835363	Genomic coordinates for easy browsing to the genes or transcripts being tested.
4	sample 1	Liver	Label (or number if no labels provided) of the first sample being tested
5	sample 2	Brain	Label (or number if no labels provided) of the second sample being tested
6	Test status	NOTEST	Can be one of OK (test successful), NOTEST (not enough alignments for testing), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents testing.
7	FPKMx	8.01089	FPKM of the gene in sample x
8	FPKMy	8.551545	FPKM of the gene in sample y
9	log2(FPKMy/FPKMx)	0.06531	The (base 2) log of the fold change y/x
10	test stat	0.860902	The value of the test statistic used to compute significance of the observed change in FPKM
11	p	value 0.389292	The uncorrected p-value of the test statistic
12	q	value 0.985216	The FDR-adjusted p-value of the test statistic
13	significant	no	Can be either "yes" or "no", depending on whether p is greater than the FDR after Benjamini-Hochberg correction for multiple-testing

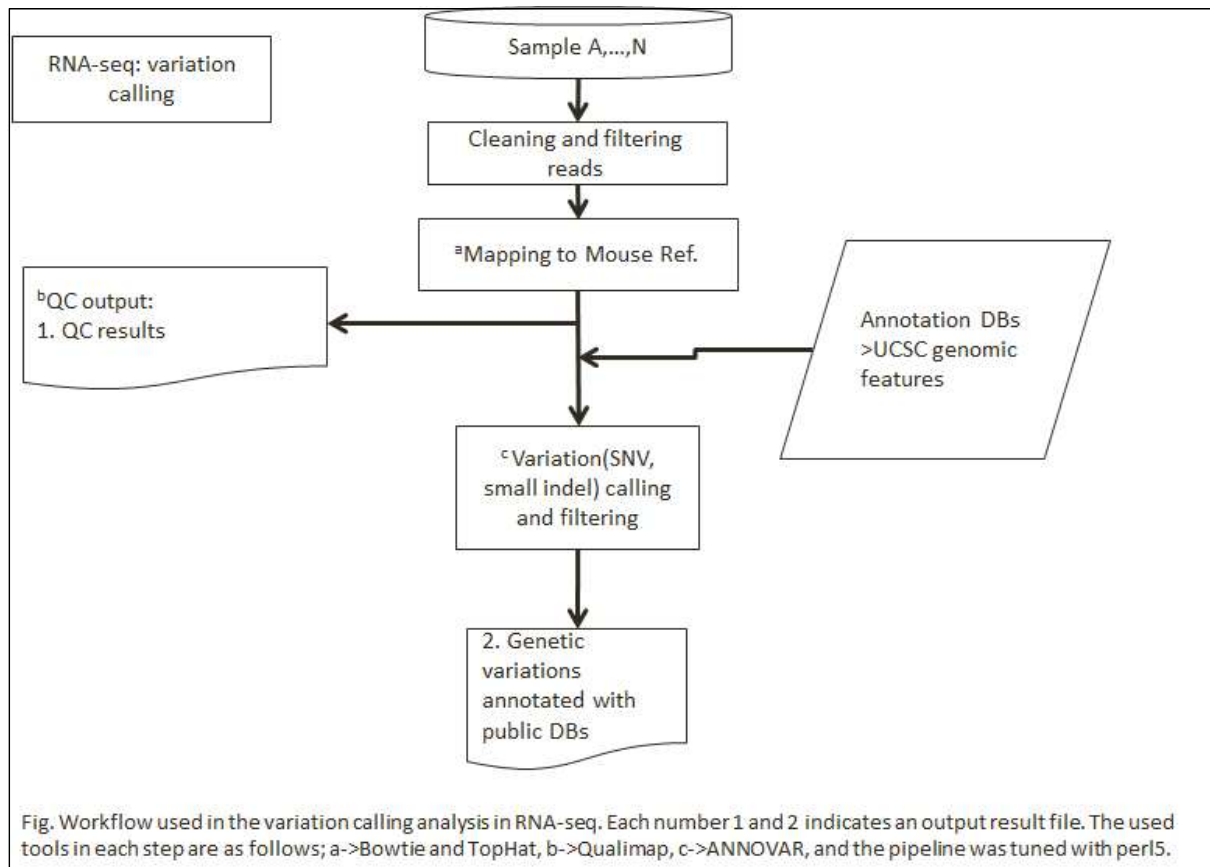
2.3. Transcriptome: variation calling

2.3.1. 분석 목표

- RNA는 DNA를 주형(template)로 합성되므로 RNA-seq을 통하여 SNV, Indel, SV등의 DNA상의 유전변이(genetic variation)을 확인하고, 이 변이의 기능 및 보고 여부를 확인해 보는 것이 이 분석 파이프라인의 목적입니다.
- RNA-seq이기에 유전변이는 다시 DNA seq을 통하여 재확인 하는 것을 추천합니다.

2.3.2. 구성

- NGS의 raw read들은 Bowtie 및 TopHat을 통해 reference genome에 mapping됩니다(아래 그림 참고). Mapping의 Quality는 Qualimap을 통해 제공되며, 이후의 variation calling은 samtools를 통해 이루어지며, 이들은 ANNOVAR를 통해 filtering 및 annotation을 거치게 됩니다.



2.3.3. 결과의 예제

a. QC results

- 2.2.2의 QC results 참고하시기 바랍니다.

b. Genetic variations annotated with public DBs

- 샘플 N1과 N2에 대하여 다음 표와 같은 결과가 제공됩니다. 단 아래 표는 human sample을 분석할때의 결과 포맷이며, 붉은색 부분은 제공되지 않습니다. 단 마우스와 연관된 human 샘플을 분석할때는 제공될것입니다.
- 추가로 각 변이의 mapping quality, depth등의 정보 가공도 가능합니다.
- KeggPathway, genomic duplication, cpg island, cytoband등의 정보 제공도 가능합니다.

[illegible]

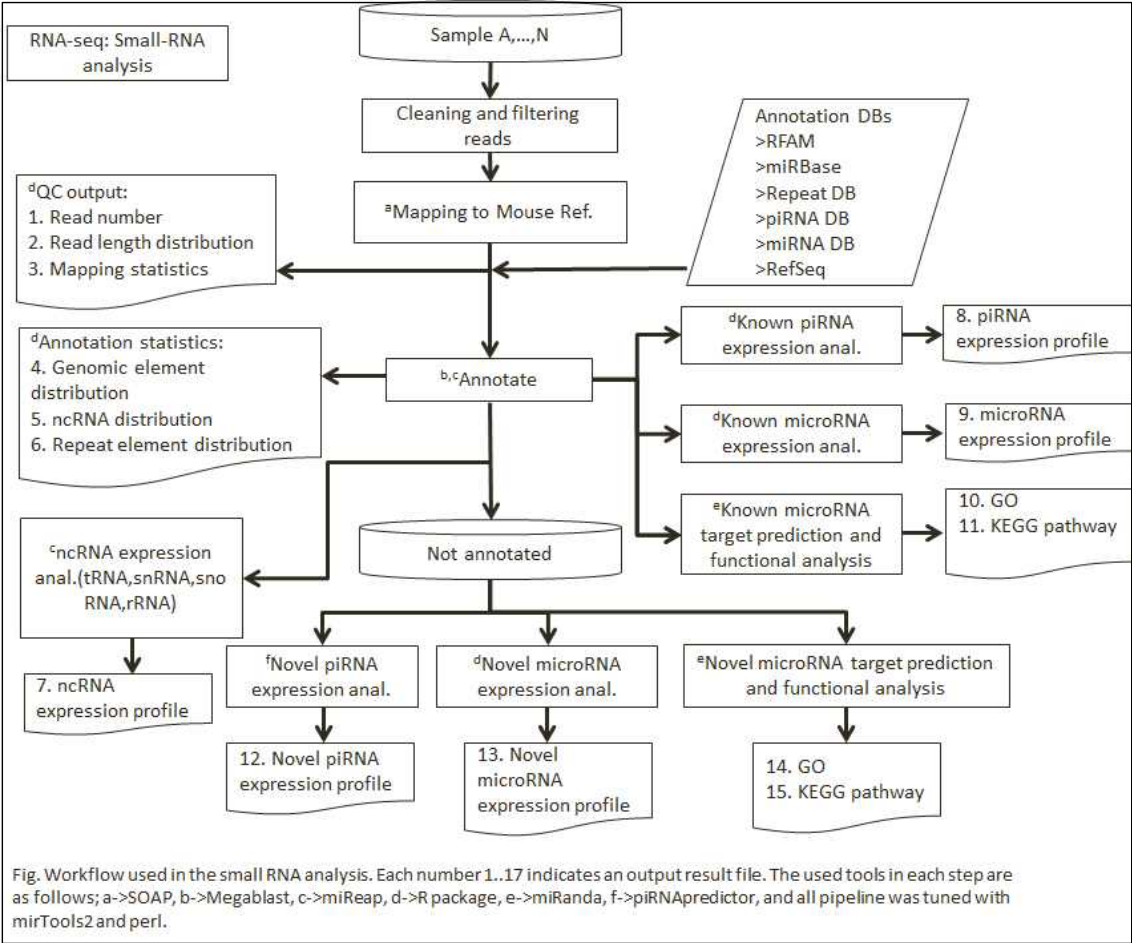
2.4. Transcriptome: small-RNA analysis

2.4.1. 분석 목표

- DNA를 주형으로 cell내에서 합성되는 RNA는 다양한 역할을 수행하는 것으로 알려져 있습니다. 대표적으로 단백질 번역의 주형이 되는 mRNA, 아미노산을 나르는 tRNA, 단백질을 합성하는 rRNA 외에도 다양한 기능을 수행하는 RNA가 보고되고 있습니다. 이 파이프라인은 mRNA를 제외한 모든 non-coding RNA(small RNA)를 분석하기 위해 구성된 것입니다.

2.4.2. 구성

- 현재 분석 가능한 항목은 tRNA, snRNA, snoRNA, rRNA, piRNA, microRNA등이며 기존에 알려진 것을 확인하는 것과 알려지지 않은 것들에 대한 예측이 가능합니다(miReap, miRandda, piRNApredictor 등을 이용함). 그 외에 KEGG pathway 분석이나 expression profile 분석 결과도 제공됩니다.



2.4.3. 결과의 예제

a. Read number

-포맷: 아래 표는 human 샘플을 분석한 자료입니다. 동일 포맷으로 마우스 결과도 제공됩니다.

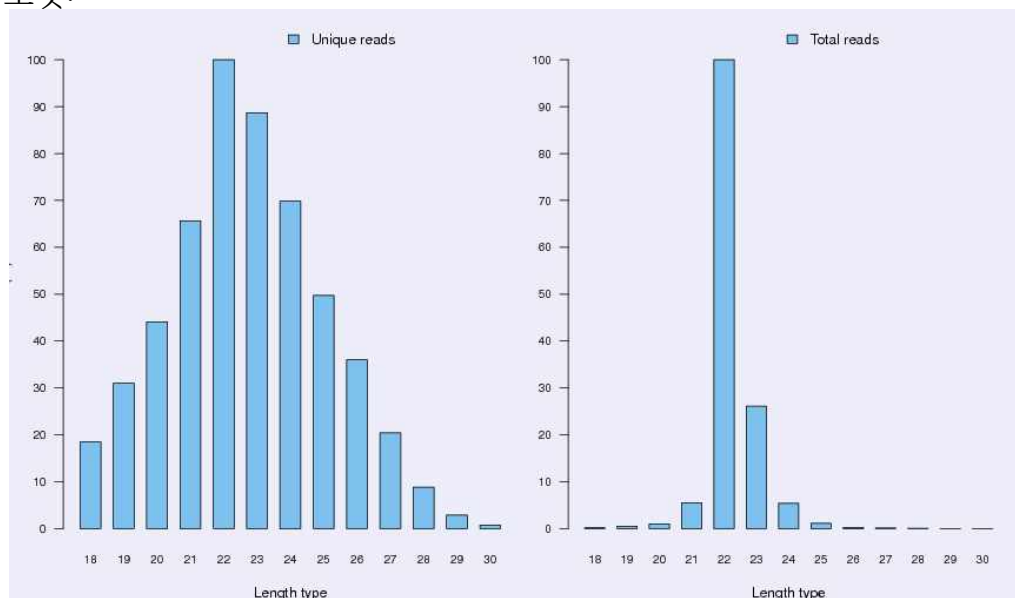
unique_query	157442														
total_query	18046672														
reference genome	hsa														
original query															
	18	19	20	21	22	23	24	25	26	27	28	29	30	SUM	
	5414	9094	1294 4	1926 2	2933 4	2601 3	2048 4	1458 6	1055 8	5999	2594	845	220	1573 47	*Tag number

	29101	6661 0	1279 04	7079 09	1285 4385	3355 316	6974 71	1447 07	3739 0	1577 1	8334	1281	372	1804 6551	**Sum of read number (= sum of (TagID * TagDepth))
	1048	2020	2793	3110	3131	1804	1097	850	645	403	156	50	15	1712 2	genome_uniq ue_num
	9782	1110 6	1318 2	2022 6	5371 5	3523 7	5689	3077	1383	1336	566	69	44	1554 12	genome_tota l_num
	86	180	250	361	466	337	246	202	186	174	87	25	11	2611	rfam_unique _num
	3158	4090	5798	4190	2878 0	3331	1318	730	689	1055	492	44	35	5371 0	rfam_total_ _num
	16	46	94	149	129	57	13	6	1	1	1	1	0	514	miRNA_uniqu e_num
	1405	451	1415	3554	5430	2642 2	1318	32	1	3	1	1	0	4003 3	miRNA_total _num
	304	615	902	988	991	564	358	306	225	124	37	16	3	5433	repeat_uniq ue_num
	1622	1576	2674	8551	1062 3	2274	552	359	288	140	37	16	8	2872 0	repeat_tota l_num
	44	83	107	125	144	112	87	74	63	31	4	0	0	874	mRNA_unique _num
	50	115	139	196	694	286	200	191	146	33	4	0	0	2054	mRNA_total_ _num
	598	1096	1440	1487	1401	734	393	262	170	73	27	8	1	7690	unannotatio n_unique_nu m
	3547	4874	3156	3735	8188	2924	2301	1765	259	105	32	8	1	3089 5	unannotatio n_total_num
	novel_num	35													

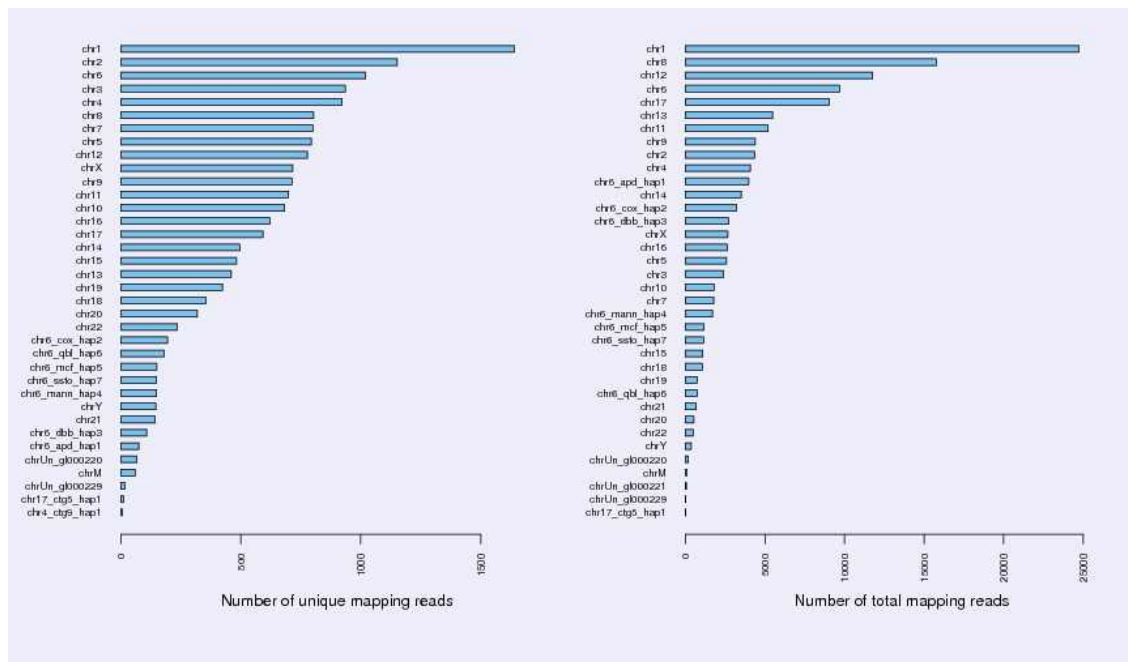
표 Small-RNA seq의 결과는 ">SampleName_TagXXX_xNN TCGTTTCCGGCTCG..."로 이루어져 있음. 위의 결과는 샘플 SampleName으로부터 TagXXX라는 서열이 NN번 검출되었다는 의미임. 따라서 *Tag number의 의미는 "18mer에서 30mer의 대표 시퀀스들이 몇 번 검출되었는지를 담고 있음. **Sum of read number는 각각의 Tag에다가 NN을 곱해서 산출한 것임. 'genome_unique_num'은 reference genome에 mapping되는 수이며, 'genome_total_num'은 이에 NN을 곱해서 합친 수임. 나머지 이 하 동일.

b. Read length distribution

-포맷:

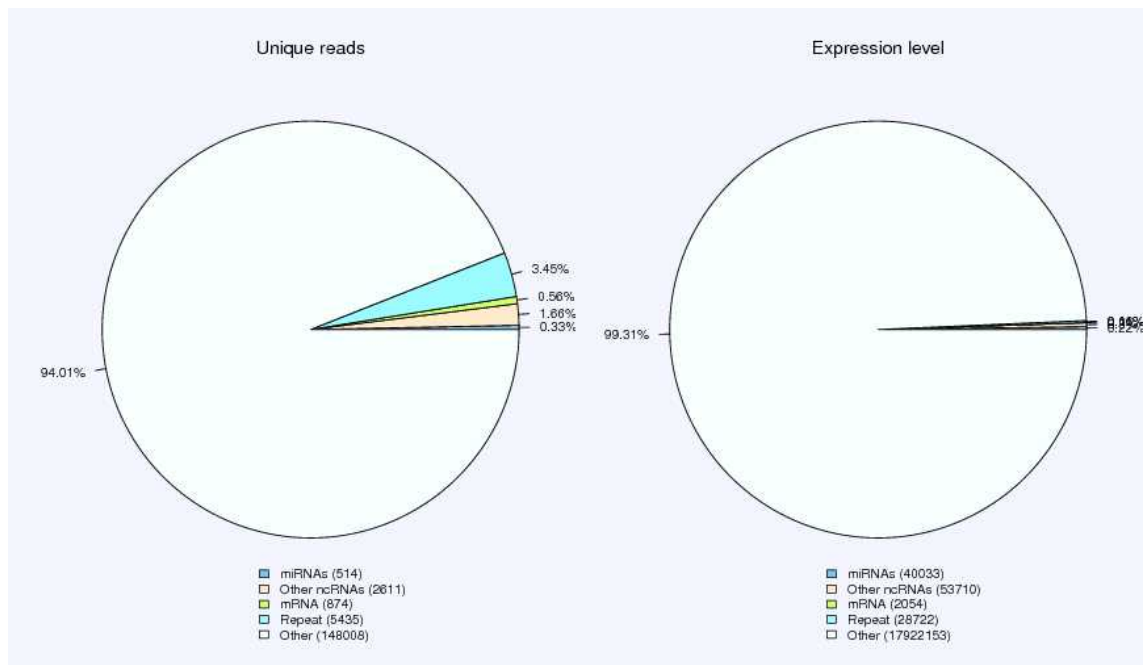


c. Mapping statistics

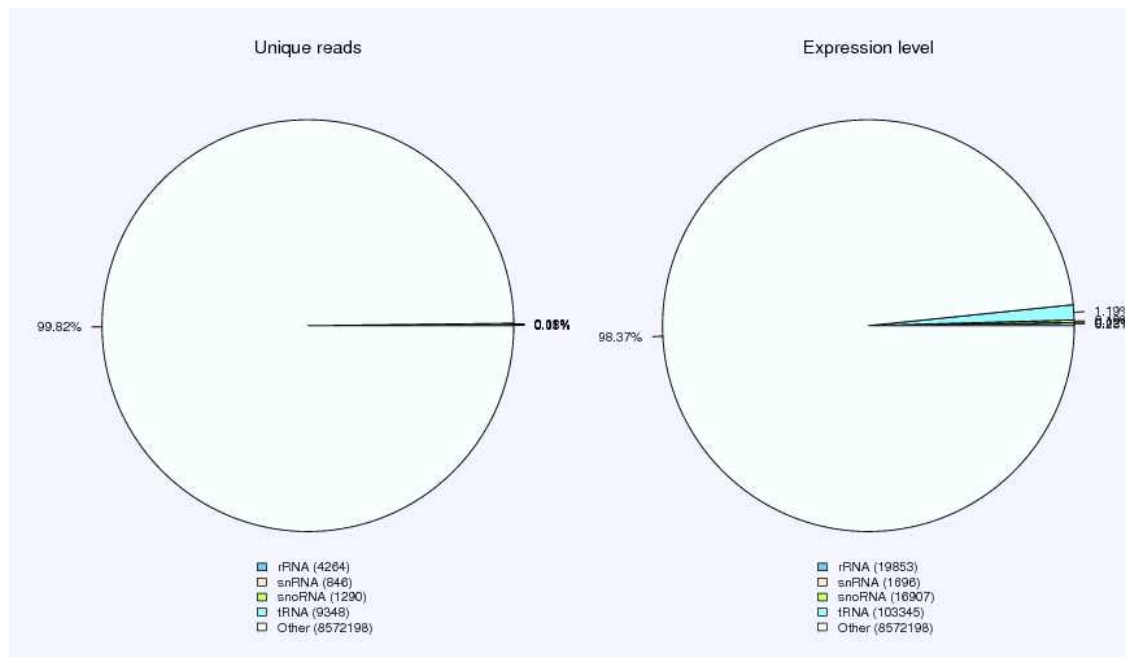


d. Genome element distribution

- RNA_stat.png: genome element들에 관한 통계량임.

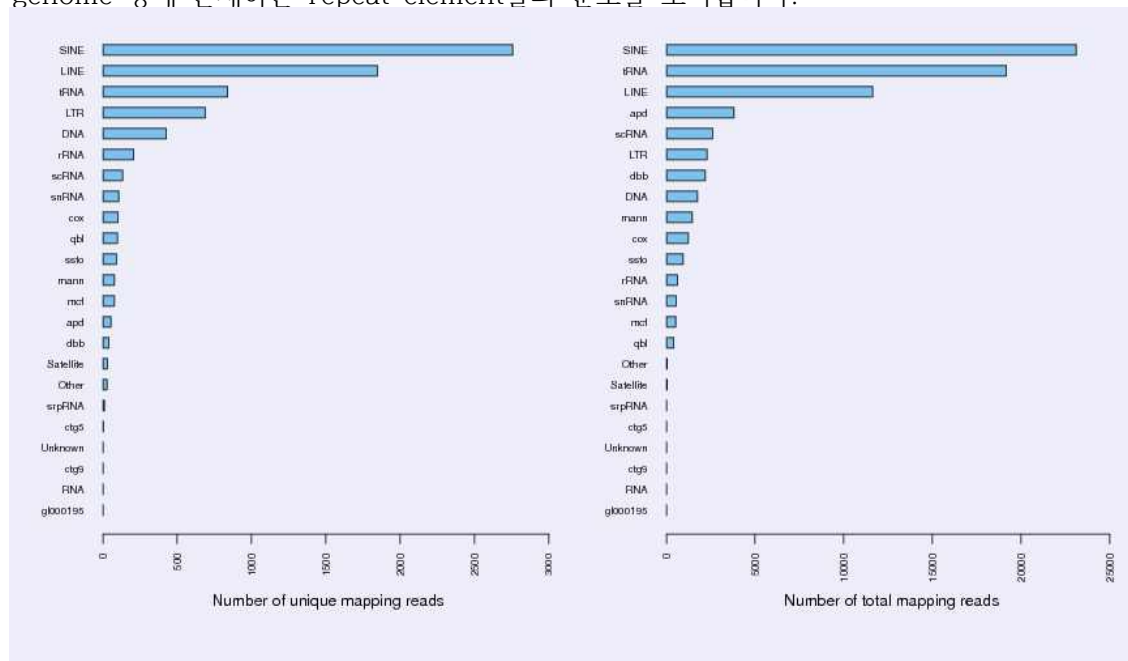


e. ncRNA distribution



f. Repeat element distribution

- genome 상에 존재하는 repeat element들의 분포를 보여줍니다.



g. ncRNA expression profile (miRNA, snRNA, snoRNA, tRNA, rRNA 포함)

g.1. snRNA
-포맷:

SN	*snRNA	Absolute count	Relative count	Sequence	Most abundant tag			
					Tag ID	*Absolute count	****Relative count	Tag sequence
1	AANU01245863.1/5270-5163_RF00026:U6;	220	12971.7	GUGCUCACUUCAGCAGCACAGACAUCCAAUUGGAAAGAUACAGAAAGAUUAGCAUGGCUAU AAGAAAAGAAAAAGAACAGCAGUCUGAGCAACAAAGUGAGACC	NY-D1_t0003528_x40	40	23584.91	TCTGGGCAACAAAGT GAGACCTAT
2	AAIY01493423.1/1716-1526_RF00004:U2;	163	96108.49	AUUGCUUCUCGGCCUUUUGGCUAAGAUCAAGUGUAGUUAUCUGUUCUUAUCAGUUUAUUAUCUGA UACGUCCUCUUAUCUGAGGACAUAUAUUAUAAUGGAUUUUUGGAGCAGGAGAUUGGAACAGGAGC UUGCUCCGUCCACUCCACACAACUGAUUAUACAGUACUUCAGAAUUGGUGCACCACAAA	NY-D1_t0003231_x45	45	26533.02	AAATGGATTTTGGG GCAGGGAGA

표 *snRNA: 서치된 Rfam sequence accession ID/Start-Stop_RfamFailiyID;RfamClanName, **Tag ID: 시퀀싱 된 read의 고유 ID, ***Absoute count: snRNA의 read수, ****Relative count: Normalized snRNA read 수, (snRNA의 read수/시퀀싱된 total ncRNA) * 10⁶

- g.2. snoRNA
 - 포맷: “g.1. snRNA”와 동일
- g.3. tRNA
 - 포맷: “g.1. snRNA”와 동일
- g.4. rRNA
 - 포맷: “g.1. snRNA”와 동일

- h. piRNA
 - RNAdb (http://research.imb.uq.edu.au/rnadb/rnadb2_archive.htm) 검색 결과입니다.
 - 포맷: “g.1. snRNA”와 동일

- i. miRNA
 - 포맷:
 - *miRNA: 서치된 mirBase sequence accession ID
 - **Tag ID: 시퀀싱된 read의 고유 ID
 - ***Absolute count: miRNA의 read 수
 - ****Relative count: Normalized miRNA read 수, (miRNA의 read수/시퀀싱된 total ncRNA) * 10⁶

SN	*miRNA	Absolute count	Relative count	Sequence	Most abundant tag			
					Tag ID	*Absolute count	****Relative count	Tag sequence
1	hsa-miR-222-3p	24888	621687.1	AGCUACAUCUGGCUA CUGGGU	NY-D1_t0000048_x23619	23619	589988.3	AGCTACATCTGGCTAC TGGGTCT
2	hsa-miR-29b-3p	2265	56578.32	UAGCACCAUUUGAAA UCAGUGUU	NY-D1_t0000269_x1892	1892	47261.01	TAGCACCATTTGAAAT CAGTGTT

- j. GO functional analysis of known miRNAs
 - 이 분석은 (1)miRNA가 간섭할 것으로 여겨지는(또는 알려진) 타겟 유전자를 밝히는 것으로 시작해서, (2)이 유전자들이 속한 Gene Ontoloty(GO)의 기능적 분석으로 이어진다.
 - j.1. Target prediction result of known miRNAs (miRanda의 결과)
 - 포맷:
 - *Targets: miRNA와 상호작용(binding)할것으로 예측되는 gene name
 - **Score: mircrorna.org에서 제공되는 mirSVR score입니다. 가장 최신의 miRanda prediction rules (seed-site pairing, site context, free-energy, and conservation)을 이용한 예측 기법입니다. 최소값은 140입니다.
 - ***Minimum free energy (MFE): Free energy refers to the minimum free energy and shows how strong the binding of a miRNA with its target is. Normally free energy is a negative real value and its unit is kcal/mol. The lower the free energy, the firmer the binding structure is and the more likely it suggests the true binding.

miRNAs	*Targets	**Score	***Minimum free energy	Target prediction tool	Supported by other tools
hsa-miR-320a	CDC14B	158	-19.3	miRanda	microT
hsa-miR-320a	IGF1R	283	-21.9	miRanda	microT
hsa-miR-29b-3p	C3orf72	149	-18.42	miRanda	-

- j.2. GO report of known miRNA targets
 - 포맷:

GO cluster	GO term	*N	**n	***M	****m	Gene description	Related genes	#Enrichment fold	##P value
biological_process	GO:0006397	99638	7184	196	98	mRNA processing	A1CF ADARB1 ADARB2 ADAR AFF2 AICDA APOBEC2 APOBEC4 BCAS2 CASC3 CDK12 CELF1 CELF2 CELF5 CELF6 ...생략	6.93	0
biological_process	GO:0010467	99638	7184	638	290	gene expression	A1CF AARS2 ADARB1 ADAR AIMP1 AIMP2 AR BDP1 BRF2 CARS2 CASC3 CCNT1 CCNT2 CD2BP2 CDC40 CDK9 CLNS1A CNOT1 CNOT4 ...생략	6.3	0
biological_process	GO:0016554	99638	7184	2	1	cytidine to uridine editing	A1FF	6.93	0.19

*N: total number of genes annotated by GO/KEGG in whole genome.

**n: total number of genes annotated by a specific GO/KEGG term in whole genome.

***M: total number of genes annotated by GO/KEGG in predicted miRNA targets.

****m: total number of genes annotated by a specific GO/KEGG term in predicted miRNA targets.

#Enrichment fold: 샘플내의 특정 GO term에 속하는 유전자들과 상호작용하는 miRNA들의 발현량 증감 정도

##P-value: hypergeometric test를 통한 유의도

k. KEGG functional analysis of known miRNAs

-이 분석은 miRNA가 간섭할 것으로 여겨지는(또는 알려진) 타겟 유전자를 밝히는 것으로 시작해서, 이 유전자들이 속한 KEGG의 기능적 분석으로 이어집니다.

→포맷:

KEGG pathway ID	*N	**n	***M	****m	Related genes	#Enrichment fold	##P value
hsa00650	15840	10039	30	14	AACS ABAT ACADS ACAT1 ALDH5A1 BDH1 GAD1 HADH HMGCS1 HMGCS2 L2HGDH OXCT2 PDHA1 PDHB	0.74	0.44
hsa00380	15840	10039	42	17	AADAT ACAT1 ALDH7A1 CBL1 CYP1A2 CYP1B1 HADH IL4I1 INMT KMO KYNU MAOA MAOB OGDHL TPH1 WARS WARS2	0.64	0.14

*N: total number of genes annotated by GO/KEGG in whole genome.

**n: total number of genes annotated by a specific GO/KEGG term in whole genome.

***M: total number of genes annotated by GO/KEGG in predicted miRNA targets.

****m: total number of genes annotated by a specific GO/KEGG term in predicted miRNA targets.

#Enrichment fold: 샘플내의 특정 GO term에 속하는 유전자들과 상호작용하는 miRNA들의 발현량 증감 정도

##P-value: hypergeometric test를 통한 유의도

l. Novel piRNA expression profile

-RNAdb (http://research.imb.uq.edu.au/rnadb/rnadb2_archive.htm)-search-result

-포맷:

SN	novel piRNA	Absolute count	Relative count	Sequence	Most abundant tag			
					Tag ID	Absolute count	Relative count	Tag sequence
1	novel_piRNA_281	7	58333.33	AAACTCAAAGGACCTGGCGGTGCTTC	NY-D1_t0069370_x1	1	8333.33	AACTCAAAGGACCTGGCGG
2	novel_piRNA_129	6	50000	TGAAAAATTTGAGAAGGAGGCTGCTG	NY-D1_t0034255_x2	2	16666.67	AAAAATTTGAGAAGGAGGCTGCTG

m. Novel microRNA expression profile

-포맷:

*Sequence: 이 miRNA의 Hairpin structure의 이미지가 제공됨

*Precursor sequence: 이 서열은 게놈상에서 이 miRNA가 발현되는 위치를 표시함.

SN	Novel miRNA	Arm	Absolute count	Relative count	pre-miRNA number	*Sequence	Most abundant tag				**Precursor sequence
							Tag ID	Absolute count	Relative count	Tag sequence	
1	xxx-m0034-3p	3p	217	2511 57.4	1	AGGCCGGGA GTGGTGGCT CAT	NY-D1_t 0001308 _x178	178	20601 8.5	CGGGCGGG AGTGGTGG CTTTT	GGATTACAGCGT GAGACACCGC ACCCAGCCACCC CTTTCTTTTTTT AAAAAACTTTTG AGGCCGGGAGT GGTGGCTCATGC CTGTAATC
2	xxx-m0004-3p	3p	182	2106 48.2	1	GAAGTTGGG GCTCTTATT TT	NY-D1_t 0001744 _x114	114	13194 4.4	GAAGCGGG GGCTCTTA TTTT	GAATTCAGGGA ACTATGGGAGC TCTCACAAGCAT GAAAAATCCAAG TATTCAGATGGA AGTTGGGGCTCT TATTTTCCAGCT TTGGG

m. GO report of novel miRNA targets

- 이 분석은 (1)miRNA가 간섭할 것으로 여겨지는(또는 알려진) 타겟 유전자를 밝히는 것으로 시작해서, (2)이 유전자들이 속한 Gene Ontology(GO)의 기능적 분석으로 이어집니다.

m.1. Target prediction result of novel miRNAs (miRanda의 결과)

- 포맷

miRNAs	*Targets	**Score	***Minimum free energy	Target prediction tool	Supported by other tools
xxx-m0034-3p	TGFB3	149	-24.03	miRanda	-
xxx-m0034-3p	TGFB3	149	-24.03	miRanda	-

m.2. GO report of novel miRNA targets

- 포맷:

*N: total number of genes annotated by GO/KEGG in whole genome.
 **n: total number of genes annotated by a specific GO/KEGG term in whole genome.
 ***M: total number of genes annotated by GO/KEGG in predicted miRNA targets.
 ****m: total number of genes annotated by a specific GO/KEGG term in predicted miRNA targets.
 #Enrichment fold: 샘플내의 특정 GO term에 속하는 유전자들과 상호작용하는 miRNA들의 발현량 증감 정도
 ##P-value: hypergeometric test를 통한 유의도

GO cluster	GO term	*N	**n	***M	****m	Gene description	Related genes	#Enrichment fold	##P value
biological_process	GO:0010466	99638	6287	9	2	negative regulation of peptidase activity	A2ML1 SERPINE2	3.52	0.14
biological_process	GO:0052548	99638	6287	3	2	regulation of endopeptidase activity	A2ML1FUURIN	10.57	0.03

n. KEGG report of novel miRNA targets

- 이 분석은 miRNA가 간섭할 것으로 여겨지는(또는 알려진) 타겟 유전자를 밝히는 것으로 시작해서 이 유전자들이 속한 KEGG의 기능적 분석으로 이어집니다.

- 포맷:

*N: total number of genes annotated by GO/KEGG in whole genome.
 **n: total number of genes annotated by a specific GO/KEGG term in whole genome.
 ***M: total number of genes annotated by GO/KEGG in predicted miRNA targets.
 ****m: total number of genes annotated by a specific GO/KEGG term in predicted miRNA targets.
 #Enrichment fold: 샘플내의 특정 GO term에 속하는 유전자들과 상호작용하는 miRNA들의 발현량 증감 정도
 ##P-value: hypergeometric test를 통한 유의도

KEGG pathway ID	*N	**n	***M	****m	Related genes	#Enrichment fold	##P value
hsa00650	15840	8895	30	9	AACS ABAT ACADS ACAT1 ALDH5A1 BDH1 BDH2 HADH L2HGDH	0.53	0.1
hsa00380	15840	8895	42	17	AADAT ACAT1 AFMID ALDH1B1 ALDH3A2 ALDH7A1 CCBL1 CYP1A1 CYP1A2 HAAO HADH IL4I1 KMO MAOA MAOB OGDH WARS2	0.72	0.28

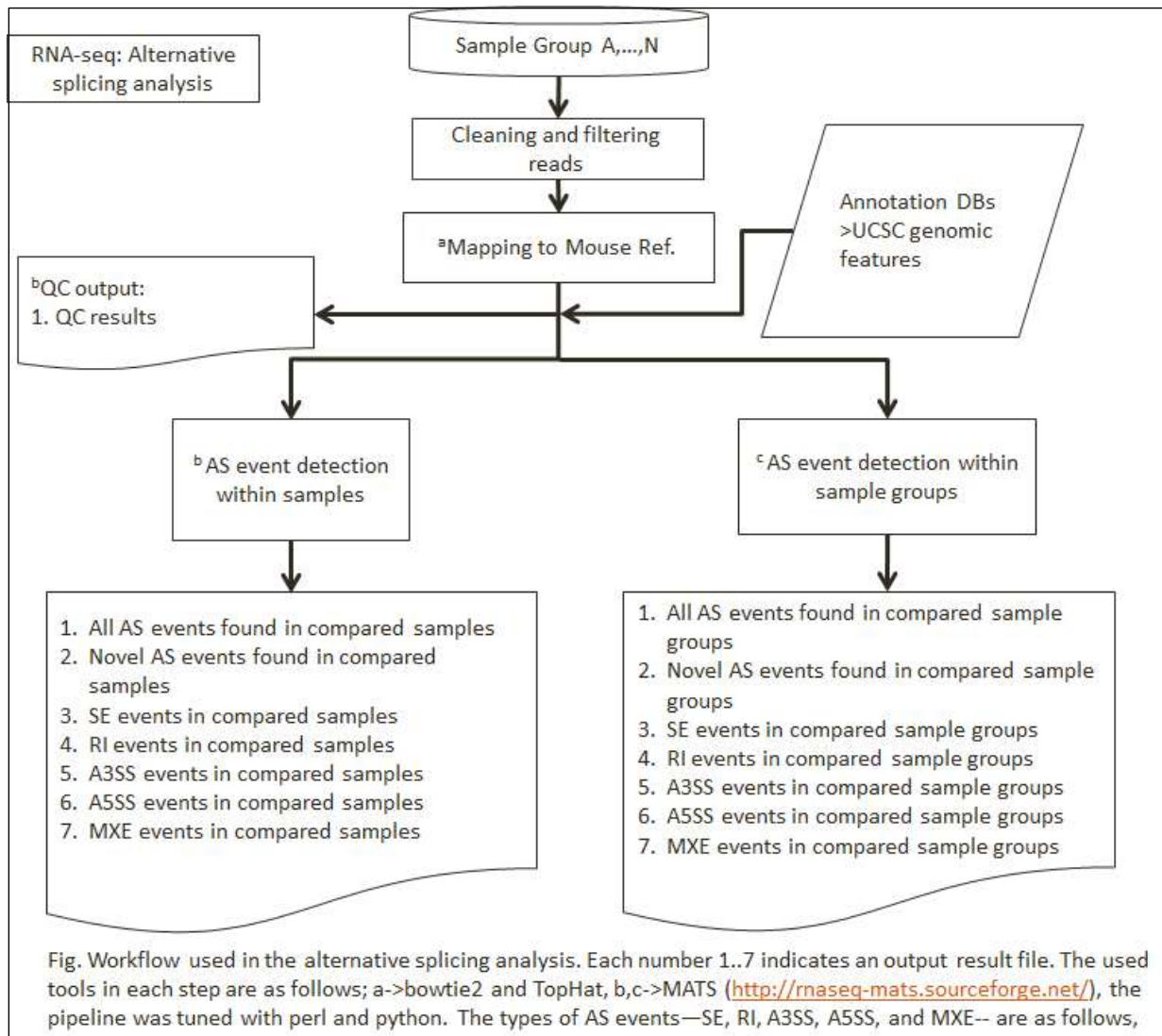
2.5. Transcriptome: alternative splicing analysis

2.5.1. 분석 목표

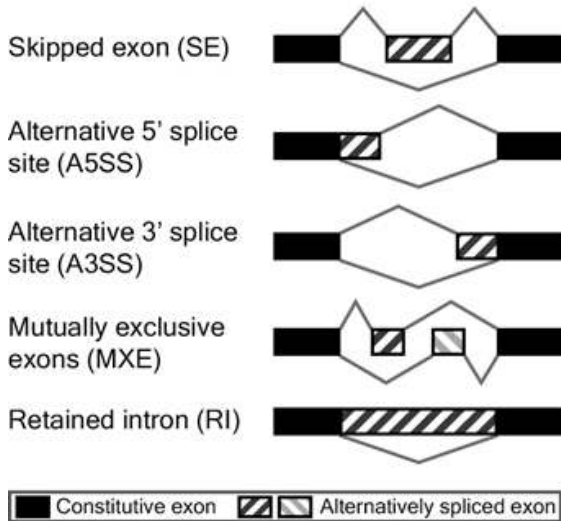
- 인간을 비롯한 고등생명체의 복잡한 생화학적 특성을 유지하기 위한 유전자의 개수를 2000년대 초까지만 해도 3만여개 이상으로 학자들은 추정했었습니다. 하지만 연구결과 이보다 적은 2만여개의 유전자만이 확인되었고, 나머지는 유전자 발현의 환경에 따른 다형성(isoform)을 통해 해결하는 것으로 최근 보고되었습니다.
- 이 파이프라인의 목표는 이러한 유전자의 발현 다형성이 유지되는 기작인 Alternative splicing을 확인하고, 샘플 간 비교를 수행하는 것을 목적으로 합니다.

2.5.2. 구성

- 유전자 발현의 다양한 형태를 확인하는 방법으로 기본적으로 MATS(<http://rnaseq-mats.sourceforge.net/>)을 사용합니다. 결과로는 5가지 Alternative splicing event들에 대한 정보를 제공합니다 (아래 그림 참고).



Alternative Splicing Events



2.5.3. 결과의 예제

a. QC results

- 2.2.2의 QC results 참고하시기 바랍니다.

b. All AS events

-포맷: SE

ID	GeneID	geneSymbol	chr	strand	exonStart_Obase	exonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE
1	FCAMR	FCAMR	chr 1	-	2.07E+08	2.07E+08	2.07E+08	2.07E+08	2.07E+08	2.07E+08
2	NFU1	NFU1	chr2	-	69627495	69627670	69623244	69623422	69633153	69633214

-포맷: A3SS, A5SS

ID	GeneID	geneSymbol	chr	strand	longExonStart_Obase	longExonEnd	shortTES	shortTEE	flankingES	flankingEE
1	MECR	MECR	chr 1	-	29543099	29543324	29543099	29543197	29557242	29557454
2	FOXM1	FOXM1	chr 12	-	2973848	2973921	2973848	2973918	2975558	2975687

-포맷: MXE

ID	GeneID	geneSymbol	chr	strand	1stExonStart_Obase	1stExonEnd	2ndExonStart_Obase	2ndExonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE
1	AARSD1	AARSD1	chr 17	-	41103816	41103911	41105740	41105795	41102542	41102745	41106892	41106984
2	AARSD1	AARSD1	chr 17	-	41103816	41103911	41106892	41106984	41102542	41102745	41107174	41107241

-포맷: RI

ID	GeneID	geneSymbol	chr	strand	riExonStart_Obase	riExonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE
1	FCAMR	FCAMR	chr 1	-	2.07E+08	2.07E+08	2.07E+08	2.07E+08	2.07E+08	2.07E+08
2	DERL3	DERL3	chr22	-	24176689	24179341	24176689	24179149	24179250	24179341

c. Novel AS events

-포맷: 'b'와 동일 포맷입니다. 단 기존의 "UCSC에 확인된 것" 외의 splicing event만 기록됩니다.

d. SE events

- 두 가지 결과가 있음

-1번결과: A file that evaluates differential AS using only reads that span splice junctions.

ID	GeneID	geneSymbol	chr	strand	exonStartObase	exonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE	ID
1033	AKR1A1	AKR1A1	chr1	+	46034156	46034356	46033653	46033849	46034598	46034671	1033
10350	RPN2	RPN2	chr20	+	35866804	35866852	35864982	35865112	35869705	35870025	10350

*IJC_SAMPLE_1	**SJC_SAMPLE_1	IJC_SAMPLE_2	SJC_SAMPLE_2	IncFormLen	SkipFormLen	PValue	FDR	IncLevel1	IncLevel2	IncLevelDifference
836	14	181	47	141	66	0	0	0.965	0.643	0.322
123	1127	8	1373	82	75	0	0	0.091	0.005	0.086

*IJC_SAMPLE_N column is for the inclusion junction counts of SAMPLE_N and **SJC_SAMPLE_N column is for the skipping junction counts of SAMPLE_N.

주의) What are IJC and SJC?

일단 샘플을 RNA-seq 하고, gene A의 exon 2번이 Skipped exon이라 할 때,

RNA seq 결과가 "exon1(3read count) - exon2(1) - exon3(3)"이면,

이는 아래와 같은 결과를 말함.

```

chr1      5'=====3'
gene      A      ----->
exon              e1-----e2-----e3-----
RNA              -----
read count              3          1          3

*IJC(inclusion junction count) 1
*SJC(skipped junction count)  2

```

-2번결과: A file that evaluates differential AS with both reads that span splice junctions and reads that map to target exons.

ID	GeneID	geneSymbol	chr	strand	exonStartObase	exonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE	ID
1033	AKR1A1	AKR1A1	chr1	+	46034156	46034356	46033653	46033849	46034598	46034671	1033
10350	RPN2	RPN2	chr20	+	35866804	35866852	35864982	35865112	35869705	35870025	10350

*IC_SAMPLE_1	**SC_SAMPLE_1	IC_SAMPLE_2	SC_SAMPLE_2	IncFormLen	SkipFormLen	PValue	FDR	IncLevel1	IncLevel2	IncLevelDifference
1326	14	302	47	252	66	0	0	0.961	0.627	0.334
123	1127	8	1373	82	75	0	0	0.091	0.005	0.086

*IC_SAMPLE_N is for the inclusion counts of SAMPLE_N

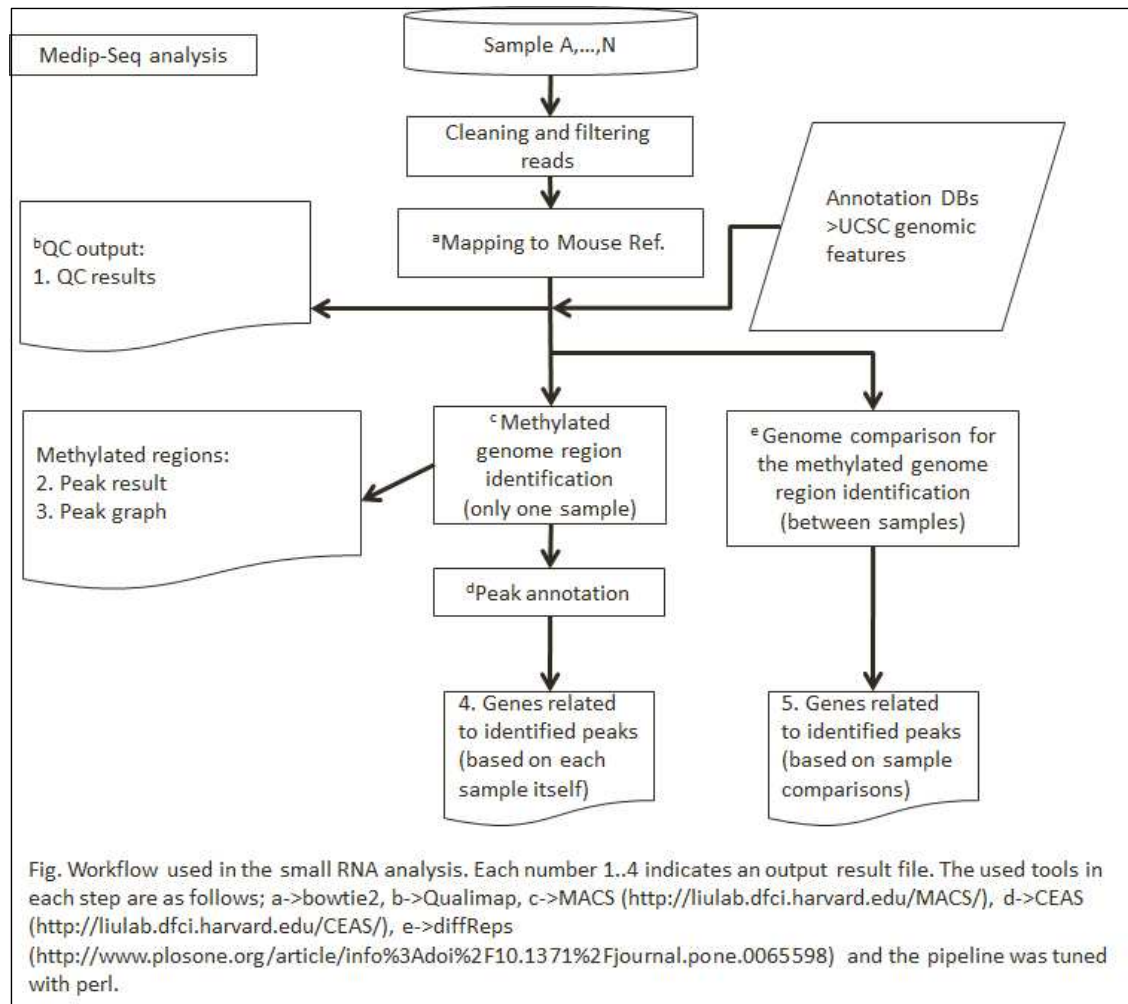
**JC_SAMPLE_N is for the skipping counts of SAMPLE_N

2.6. Epigenome: Medip-Seq analysis

2.6.1. 분석 목표

- 최근 후성유전체(epigenome) 연구가 활발하게 진행되고 있습니다. 후성유전은 후천적으로 획득된 형질의 유전을 말하며, DNA 발현을 조절하는 DNA 메틸화를 통해서 이루어집니다. 일반적으로 획득형질이 되는 DNA 메틸화 정보는 정자와 난자를 만드는 감수분열(meiosis)시 지워지나, 이들 중에서 일부는 후손에 그대로 전달되는 것으로 알려졌습니다. 최근 NGS는 이러한 DNA 메틸화 정도를 검사하기 위해 사용되며, 이 파이프라인은 이를 분석하기 위한 것입니다.

2.6.2. 구성



2.6.2. 결과의 예제

a. QC results

- 2.2.2의 QC results 참고하시기 바랍니다.

b. Peak result

- 포맷:

e. Genes related to identified peaks (based on sample comparisons)

e.1. Peak detect and annotation

- 샘플간 비교를 통해 확인된 peak들의 정보와 관련 유전자를 표시함.
- P-value는 chi-square test의 결과임
- 포맷:

Chrom	Start	End	Length	Treatment.cnt	Control.cnt	Treatment.avg	Control.avg	Treatment.enr	Control.enr	Event	log2FC
chr1	9101	10100	1000	233.63	331.45	233.63	331.45	NA	NA	Down	-0.5

pval	padj	winSta	winEnd	winFC	winP	winQ	GName	TName	Strand	TSS	TES	Feature	D2TSS
3.87E-05	0.000308	9101	10100	-0.5	3.87E-05	0.000273	DDX11L1	NR_046018	+	11873	14408	Promoter3k	-2272.5

Column Name	Explanation
Chrom	Chromosome name
Start	1-based start coordinate
End	1-based end coordinate (inclusive)
Length	Length of the differential site in bps
Treatment.cnt	Normalized read counts of the treatment group (separated by semi-colon)
Control.cnt	Normalized read counts of the control group (separated by semi-colon)
Treatment.avg	Avg. count (normalized) of the treatment group
Control.avg	Avg. count (normalized) of the control group
Treatment.enr	Fold enrichment vs. input based on avg. count (normalized) for treatment group
Control.enr	Fold enrichment vs. input based on avg. count (normalized) for control group
Event	Direction of enrichment change using the control group as reference
log2FC	Log2 fold change
pval	P-value
padj	BH-adjusted p-value (FDR)
winSta	Start coordinate of the core window which is the window with most significant p-value among the site
winEnd	End coordinate of the core window
winFC	Log2 fold change of the core window
winP	P-value of the core window
winQ	Adjusted p-value of the core window
GName	Associated gene name
TName	Associated transcript name
Strand	Associated transcription strand
TSS	Associated transcript start site
TES	Associated transcript end site
Feature	Classification of the site's location
D2TSS	Distance from the site center to the associated transcript's TSS

e.2. chromatin modification hotspot

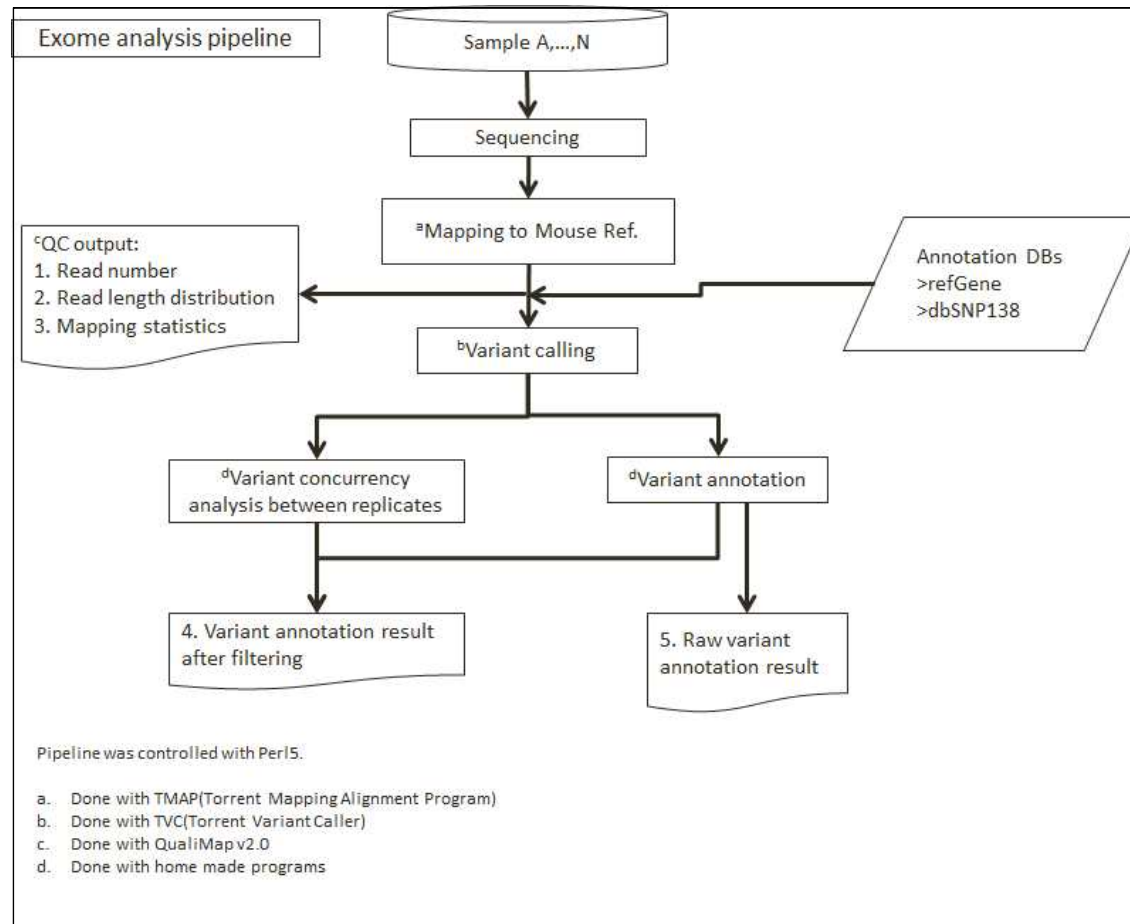
- 샘플간 비교를 통해 확인된 peak의 군집 정보(hotspot)

-포맷:

Column Name	Explanation
Chrom	Chromosome name
Start	1-based start coordinate
End	1-based end coordinate (inclusive)
Length	Length of the hotspot in bps
enrich	Enrichment ratio (compared with background)
pval	P-value of enrichment
padj	Adjusted p-value
nsite	Number of differential sites in the hotspot
Sites	One-string description of the hotspots. The differential sites are separated by semi-colon. Each differential site is described by filename:(info columns separated by comma):line_number.
ntype	Number of different types of chromatin marks in the hotspot
MarkType	Unique chromatin marks in the hotspot. In order for find Hotspots to recognize mark names, you must make the diffReps output file names contain "standard" chromatin mark names. Such as, diffout.h3k4me3.txt, diffout.h3k9ac.txt, diffout.poli1.txt. If it cannot recognize the marks, it will put the whole file names here.

2.7. Exome analysis pipeline

-구축중입니다.



3. Hardware

- LARC는 다음과 같은 서버를 가지고 있습니다. 이들은 CentOS 6.6(64bit)를 OS로 하여 움직입니다.

Machine	CPU number	Memory size	Main storage size
M1	16	47G	RAID10 5.4T
M2	16	47G	RAID10 5.4T
M3	8	47G	RAID10 5.4T
M4	16	36G	RAID10 3T
M5	16	36G	RAID10 3T
Total	72		

4. Publications (NGS related)

- Ji-Youn Han, Yeon-Su Lee, Eun Soon Shin, Jung-Ah Hwang, Seungyoon Nam, Seung-Hyun Hong, **Ho Young Ghang**, Sung Jin Yoon, and Jin Soo Lee (2012), **A genome-wide association study of survival in small cell lung cancer patients treated with irinotecan plus cisplatin chemotherapy**, *Pharmacogenomics J.*
- Han JY, Shin ES, Lee YS, **Ghang HY**, Kim SY, Hwang JA, Kim JY, Lee JS (2012), **A genome-wide association study for irinotecan-related severe toxicities in patients with advanced non-small-cell lung cancer**. *Pharmacogenomics J.*
<http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201224a.html>
- Ghang H**, Han Y, Jeong S, Bhak J, Lee S, Kim TH, Kim C, Kim S, Al-Mulla F, Youn CH, Yoo HS, The HUGO Pan-Asian SNP Consortium (2011), **How Many SNPs Should Be Used for the Human Phylogeny of Highly Related Ethnicities? A Case of Pan Asian 63 Ethnicities**. *Genomics Informatics*, 9(4):181-188, [http://koreamed.org/SearchBasic.php?RID=0117GNI%2F2011.9.4.181&DT=1&QY=%22Genomics+Inform%22+\[JTI\]++AND+2011+\[DPY\]+AND+Dec+\[DPM\]](http://koreamed.org/SearchBasic.php?RID=0117GNI%2F2011.9.4.181&DT=1&QY=%22Genomics+Inform%22+[JTI]++AND+2011+[DPY]+AND+Dec+[DPM])
- Chumpol Ngamphiw, Anunchai Assawamakin, Shuhua Xu, Philip J. Shaw, Jin Ok Yang, **Ho Ghang**, Jong Bhak, Edison Liu, Sissades Tongsim, and the HUGO Pan-Asian SNP Consortium (2011), **PanSNPdb: The Pan-Asian SNP Genotyping Databases**, *PLoS One*.
<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021451>
- So-Jung Choi, Sung-Hyun Kim, **Ho Y. Kang**, Jinseon Lee, Jong H. Bhak, Insuk Sohn, Sin-Ho Jung, Yong Soo Choi, Hong Kwan Kim, Jungho Han, Nam Huh, Gyusang Lee, Byung C. Kim and Jhingook Kim (2011), **Mutational hotspots in the mitochondrial genome of lung cancer**, *Biochemical and Biophysical Research Communications*, doi:10.1016/j.bbrc.02.078,
<http://www.sciencedirect.com/science/article/pii/S0006291X1100283X>
- Kang H**, Jung J, Cho Y S, Oh J H, Ryu M H, et al. (2010), **Gene Flow between the Korean Peninsula and Its Neighboring Countries**. *PLoS ONE* 5(7):e11855.
<http://www.plosone.org/article/info:doi/10.1371/journal.pone.0011855>
- The HUGO Pan-Asian SNP Consortium (2009), **Mapping Human Genetic Diversity in Asia**, *SCIENCE*, 2009, 326, 1541, <http://www.sciencemag.org/content/326/5959/1541.abstract>
- Sungwoong Jho, Byoung-Chul Kim, **Ho Ghang**, Ji-Han Kim, Daeui Park, Hak-Min Kim, Soo-young Jung, Ki-young Yoo, Hee-Jin Kim, Sunghoon Lee and Jong Bhak (2009), **COMUS: Clinician-Oriented locus-specific MUTation detection and deposition System**, *BMC Genomics*, 10(Suppl 3):S35,
<http://www.biomedcentral.com/1471-2164/10/S3/S35/>
- Kaput J, Cotton RG, Hardman L, Watson M, Al Aqeel AI, Al-Aama JY, Al-Mulla F, Alonso S, Aretz S, Auerbach AD, Bapat B, Bernstein IT, Bhak J, Bleoo SL, Blöcker H, Brenner SE, Burn J, Bustamante M, Calzone R, Cambon-Thomsen A, Cargill M, Carrera P, Cavedon L, Cho YS, Chung YJ, Claustres M, Cutting G, Dalgleish R, den Dunnen JT, Diaz C, Dobrowolski S, dos Santos MR, Ekong R, Flanagan SB, Flicek P, Furukawa Y, Genuardi M, **Ghang H**, Golubenkov MV, Greenblatt MS, Hamosh A, Hancock JM, Hardison R, Harrison TM, Hoffmann R, Horaitis R, Howard HJ, Barash CI, Izagirre N, Jung J, Kojima T, Laradi S, Lee YS, Lee JY, Gil-da-Silva-Lopes VL, Macrae FA, Maglott D, Marafie MJ, Marsh SG, Matsubara Y, Messiaen LM, Möslein G, Netea MG, Norton ML, Oefner PJ, Oetting WS, O'Leary JC, de Ramirez AM, Paalman MH, Parboosingh J, Patrinos GP, Perozzi G, Phillips IR, Povey S, Prasad S, Qi M, Quin DJ, Ramesar RS, Richards CS, Savige J, Scheible DG, Scott RJ, Seminara D, Shephard EA, Sijmons RH, Smith TD, Sobrido MJ, Tanaka T, Tavtigian SV, Taylor GR, Teague J, Töpel T, Ullman-Cullere M, Utsunomiya J, van Kranen HJ, Vihinen M, Webb E, Weber TK, Yeager M, Yeom YI, Yim SH, Yoo HS (2009), **Planning the human variome project: the Spain report**. *Hum Mutat.*, 30(4):496-510,
<http://onlinelibrary.wiley.com/doi/10.1002/humu.20972/abstract;jsessionid=EA3B528A650328B7E50C42A3536AD465.d03t03>
- Woo-Yeon Kim, Sang-Yoon Kim, Tae-Hyung Kim, Sung-Min Ahn, Ha Na Byun, Deokhoon Kim, Dae-Soo Kim, Yong Seok Lee, **Ho Ghang**, Daeui Park, Byoung-Chul Kim, Chulhong Kim, Sunghoon Lee, Seong-Jin Kim and Jong Bhak (2009), **Gevab: a prototype genome variation analysis browsing server**, *BMC Bioinformatics*, 10:S3, <http://www.biomedcentral.com/1471-2105/10/S15/S3/>
- Sung-Min Ahn, Tae-Hyung Kim, Sunghoon Lee, Deokhoon Kim, **Ho Ghang**, Daesoo Kim, Byoung-Chul Kim, Sang-Yoon Kim, Woo-Yeon Kim, Chulhong Kim, Daeui Park, Yong Seok Lee, Sangsoo Kim, Rohit Reja, Sungwoong Jho, Chang Geun Kim, Ji-Young Cha, Kyung-Hee Kim, Bonghee Lee, Jong Bhak and Seong-Jin Kim (2009), **The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group**. *Genome Res.*, 19(9):1622-9,
<http://genome.cshlp.org/content/early/2009/05/26/gr.092197.109>
- Jong Bhak, **Ho Ghang**, Rohit Reja and Sangsoo Kim (2008). **Personal Genomics, Bioinformatics, and Variomics**, *Genomics & Informatics*, 6(4) 161-165,
[http://koreamed.org/SearchBasic.php?RID=0117GNI%2F2008.6.4.161&DT=1&QY=%22Genomics+Inform%22+\[JTI\]++AND+2008+\[DPY\]+AND+Dec+\[DPM\]](http://koreamed.org/SearchBasic.php?RID=0117GNI%2F2008.6.4.161&DT=1&QY=%22Genomics+Inform%22+[JTI]++AND+2008+[DPY]+AND+Dec+[DPM])

5. References

McFadden, D. G. et al. Cell 156, 1298–1311 (2014).

Stephen E Gould, Melissa R Junttila & Frederic J de Sauvage. Nature Medicine 21, 431–439 (2015)

García-Alcalde, et al. "Qualimap: evaluating next generation sequencing alignment data." Bioinformatics(2012) 28 (20): 2678–2679