제 1회 KAIST genomics workshop
2일차

Quantification

Normalization

Advanced bash commands

(for, while, if, md5sum)

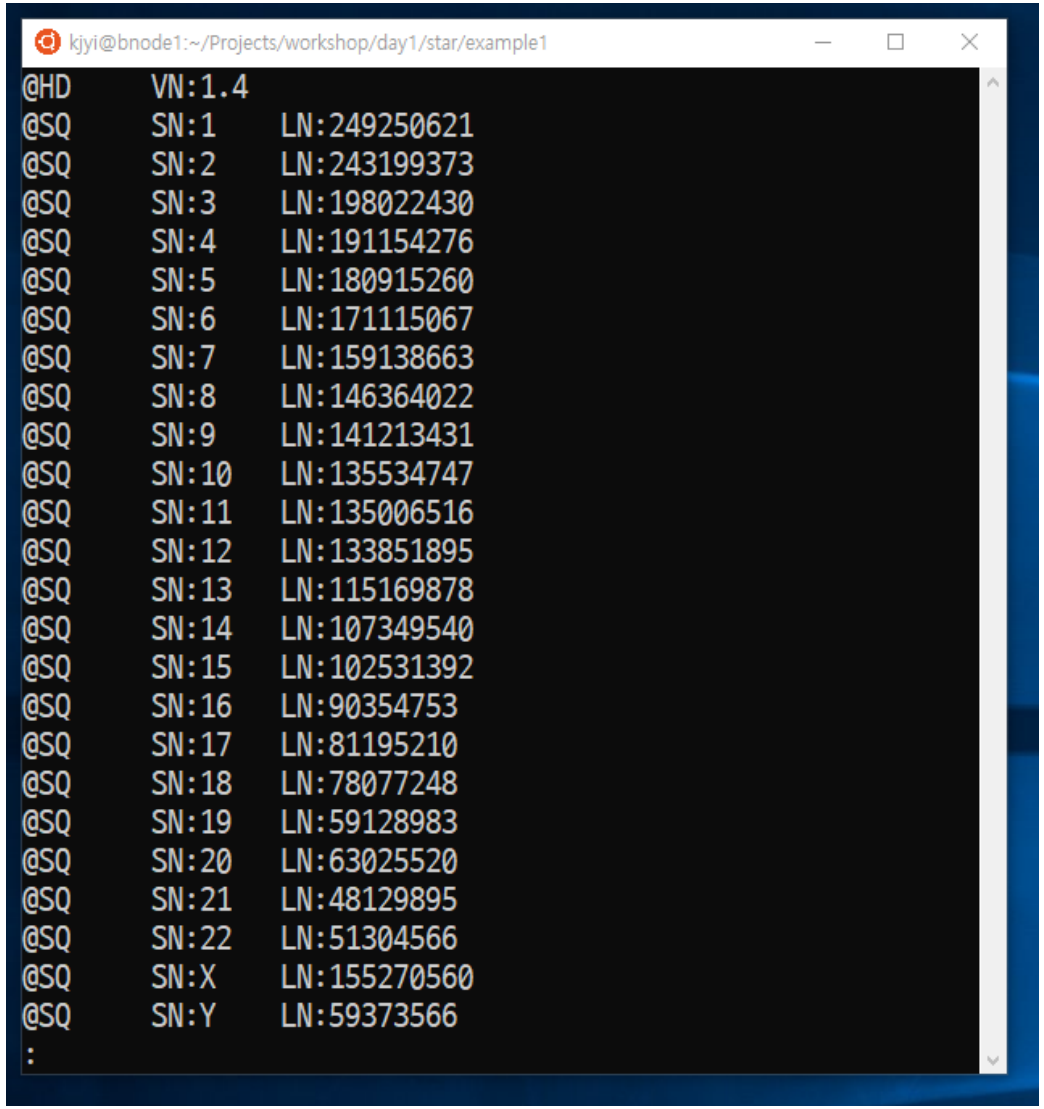
Introduction to R

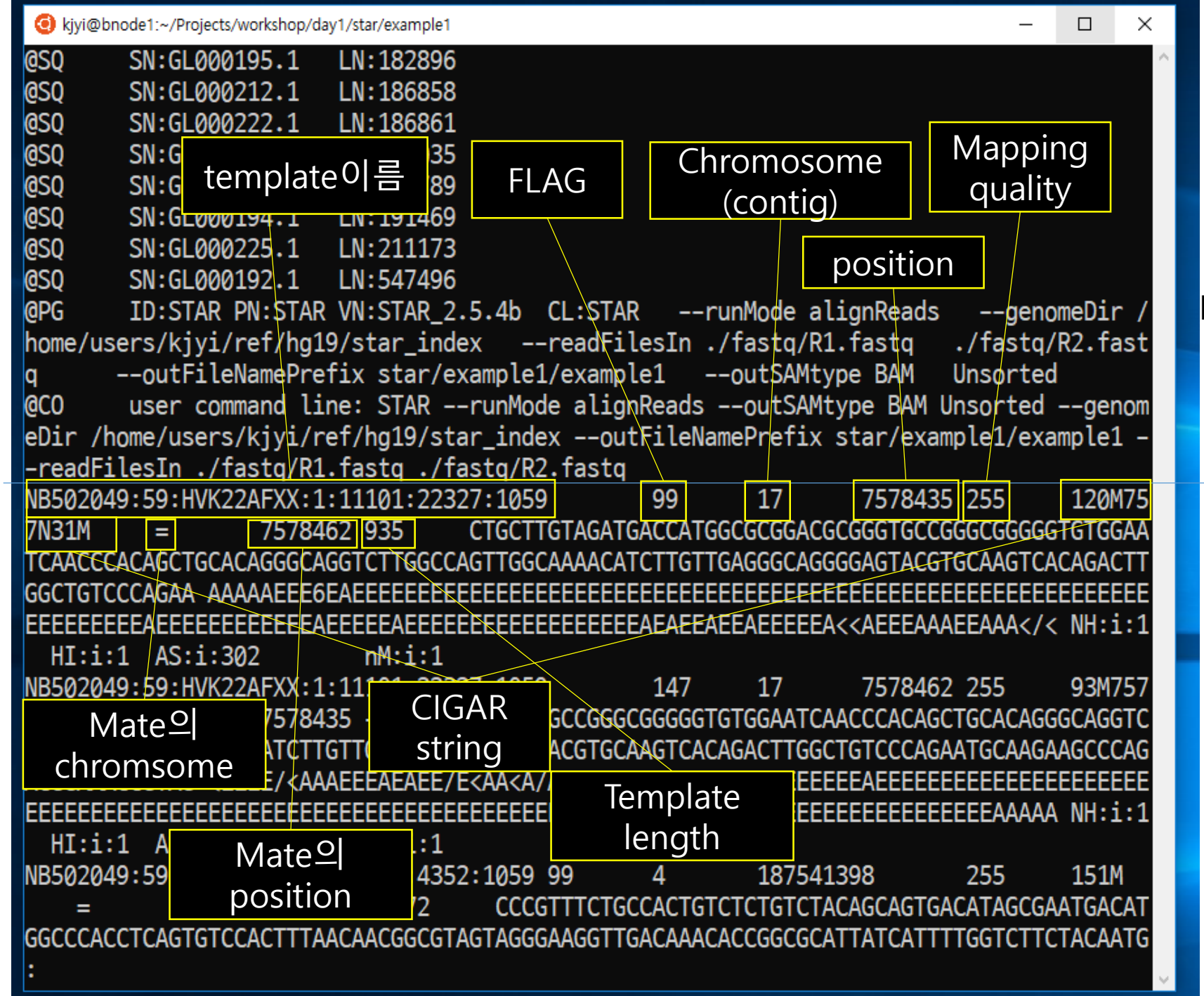
Mapping이 잘 되었는지 확인

```
$ cd ~/day1/star/example1
$ ls
example1Chimeric.out.sam
example1Chimeric.out.junction
example1Aligned.out.bam
example1SJ.out.tab
example1Log.progress.out
example1Log.out
example1Log.final.out
$ samtools -h view example1Aligned.out.bam ¦ less
```

HD = header, VN = format version

SQ = sequence dictionary, SN = reference sequence name, LN = Reference sequence length





PG = 프로그램 (bam 파일을 생성한)

CO = 코멘트 (STAR가 생성함)

여기부터 alignment section

https://samtools.github.io/hts-specs/SAMv1.pdf

FLAG: Combination of bitwise FLAGs.[7] Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Paired-end sequencin이면서 (0x1 = 1) (0x2 = 2)
두 pair가 proper mappin이 되고 (방향과 거리가 STAR에서 설정한 대로 적절함)
둘다 primary alignmen이면서, 이 read가 First in pair인 경우 FLAG는? 1+2+64=67
(0x40 = 4*16 = 64),

Mate의 FLAG는? 0x1 + 0x2 + 0x80 = 131

Sort, index

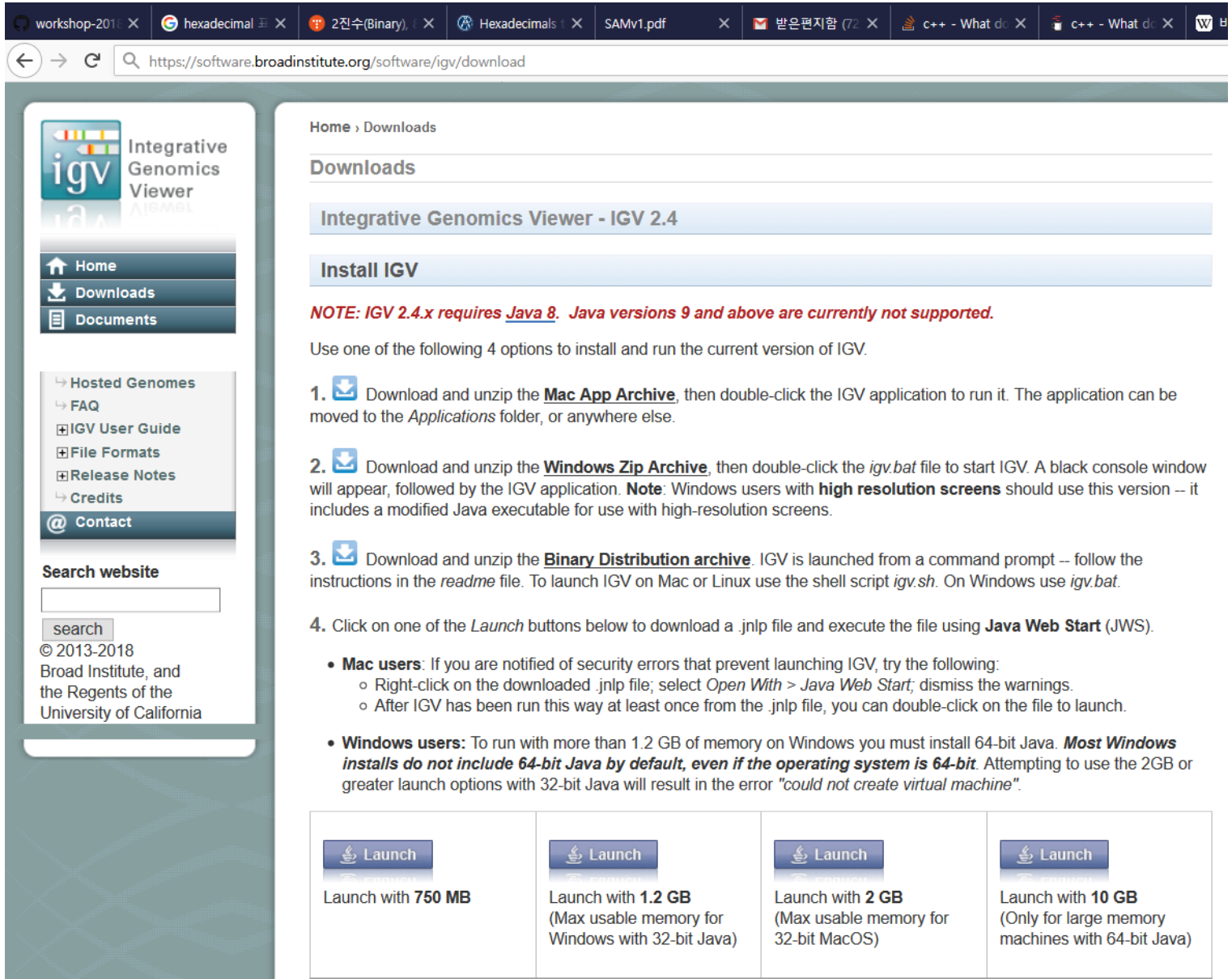
```
$ samtools sort example1Aligned.out.bam > example1Aligned.sort.bam
```

```
$ samtools index example1Aligned.sort.bam
```

(sort된 bam file 만 indexing할 수 있습니다.)

실습을 위해 example1Aligned.sort.bam file and example1Aligned.sort.bam.bai 파일을 다운로드하세요

https://software.broadinstitute.org/software/igv/download

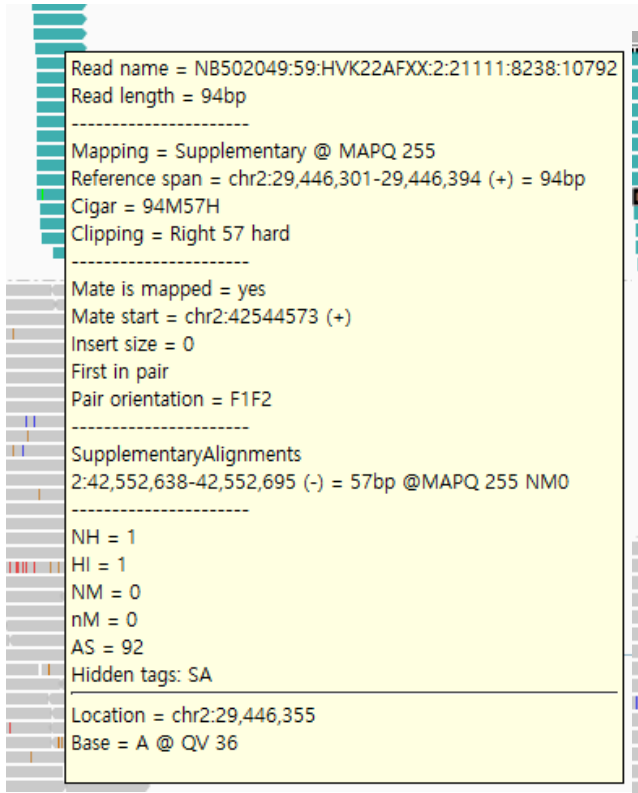


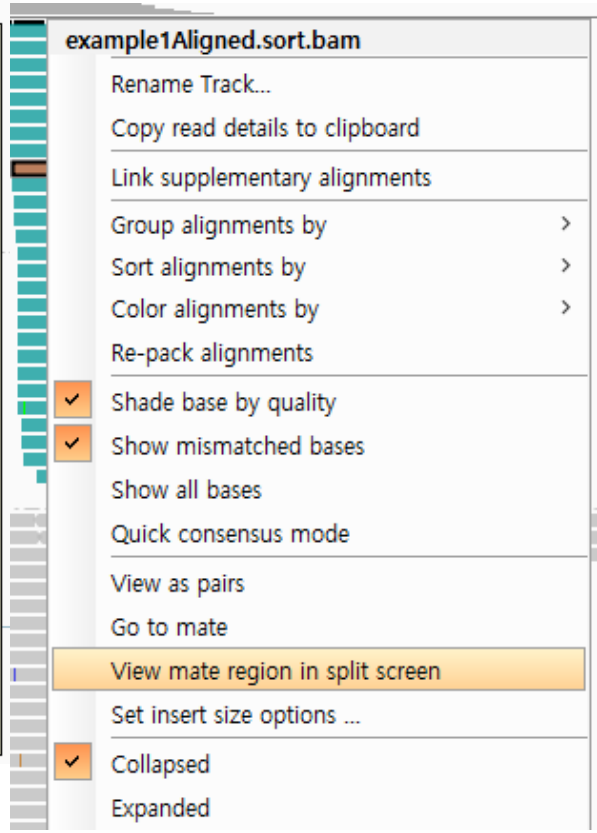
맥사용자

윈도우 사용자

리눅스, 기타 사용자

1. open bam file (sorted, indexed)
2. Go to the region
   chr2:29,445,324-29,447,615
3. right click alignment tract ->
   group by supplementary flag
4. Right click one of improperly mapped reads,
Select View mate region in split screen





제 1회 KAIST genomics workshop
2일차

Quantification

Normalization

Advanced bash commands

(for, while, if, md5sum)

Introduction to R

Quantification of Gene Expression

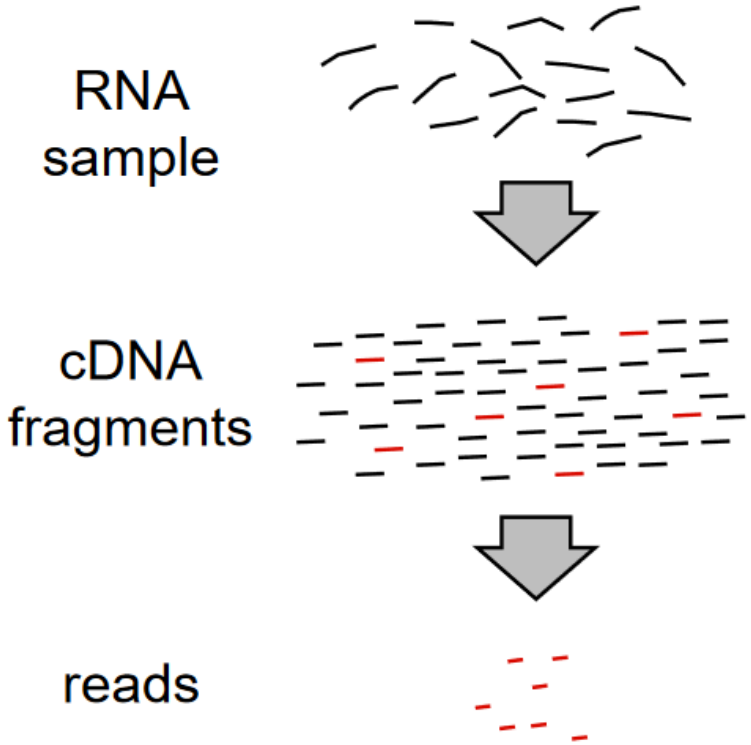
- Estimate relative abundance of transcripts
- Count reads, fetch depth/coverage

- Differential expression
  - 두 조건에서 얻은 gene expression profile을 비교하여, 어떤 transcript 가 두 조건 사이에서 발현에 차이를 보이는지를 찾는 것

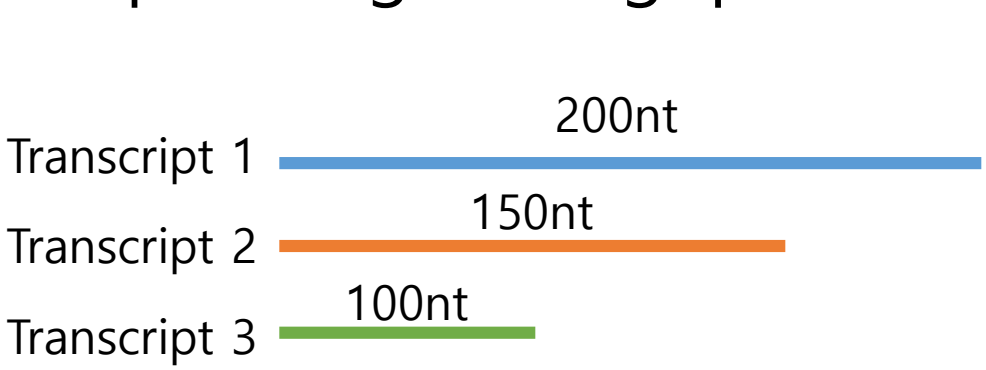
Quantification of Gene Expression (2)

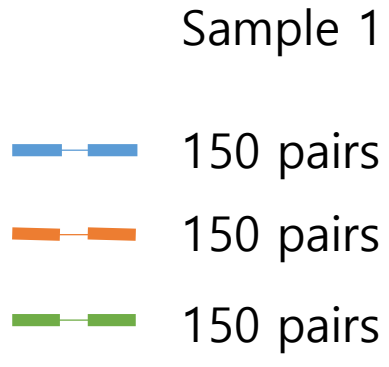
고려해야 할 것들

- Absolute quantification
- Sequencing throughput (depth)
- Gene length
- Transcript variant (different exon usage)



- Gene length
- Sequencing throughput





$$\hat{f} \propto \frac{\frac{150}{450}}{200} = \frac{1}{600}$$

150: Number of reads mapped in transcript 1

450: Total number of reads

200: Length of transcript 1

Relative abundance of Transcript 1 in sample 1

RPKM – Reads Per Kilobase per Million mapped reads

1. 샘플의 total read에 1,000,000을 나눈다. 이것이 per million scaling factor이다.
2. Read counts를 per million scaling factor로 나눈다. 이것이 reads per million (RPM)
3. RPM을 유전자(transcript)의 (평균)길이(kb 단위)로 나눈다. 이것이 RPKM이다.

FPKM은 read pair 수를 이용하면 된다. Pair 중 하나만 mapping이 된 경우, 1로 취급한다.

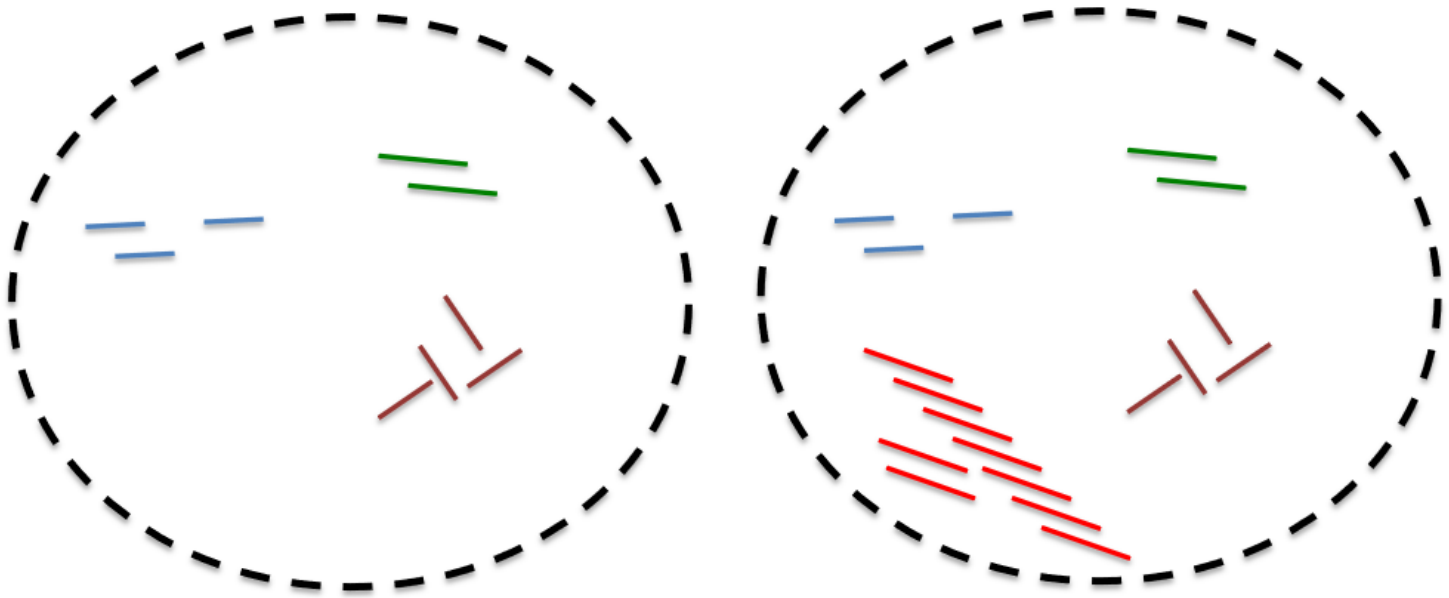
TPM – transcript per million

1. Read count를 각 유전자의 길이(kb 단위)로 나눈다. 이것이 reads per kilobase(RPK)
2. 한 샘플의 모든 transcript의 RPK를 다 더해서, 1,000,000로 나눈다.
   이것이 per million scaling factor이다.
3. RPK를 per million scaling factor로 나눈다. 이것이 TPM이다.

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

Normalization for comparing a gene across samples

1. Technical spike-in
2. Cross-sample normalization



Cell type I

Cell type II

$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}$$

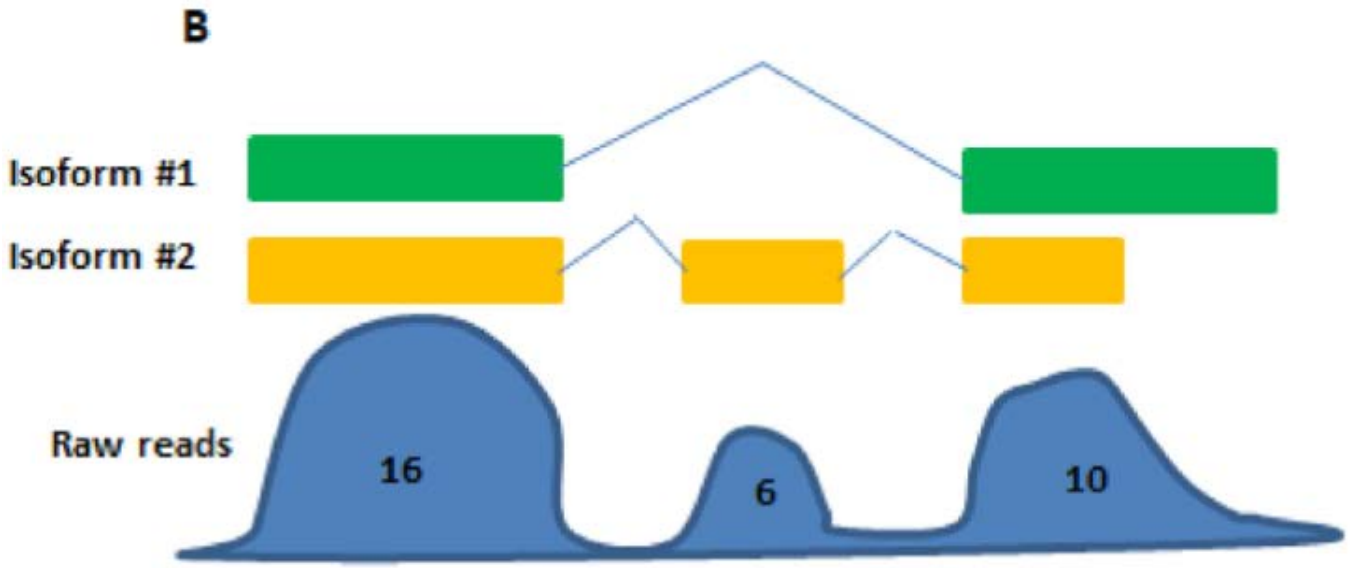
Gene i의 여러 샘플에서의 read count의 geometric mean

Size factor of sample j

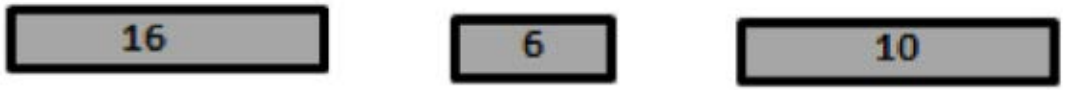
$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j}$$

http://bioinfo.umassmed.edu/content/pdf2014fall/Week4.pdf

Zhao, Shanrong, Li Xi, and Baohong Zhang. "Union exon based approach for RNA-seq gene quantification: To be or not to be?." *PLoS One* 10.11 (2015): e0141910.

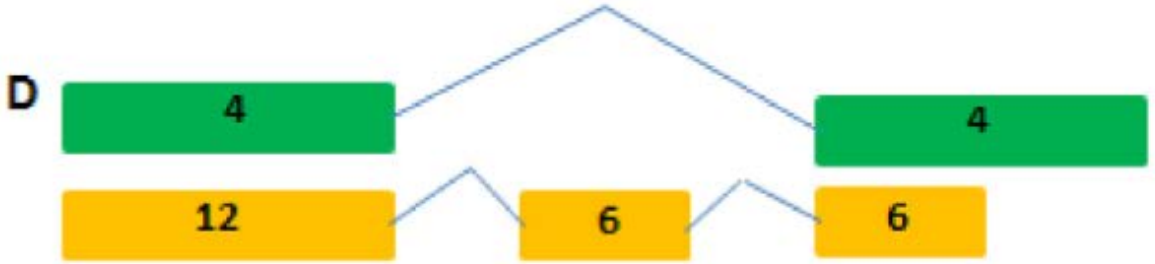


union-exon based approach



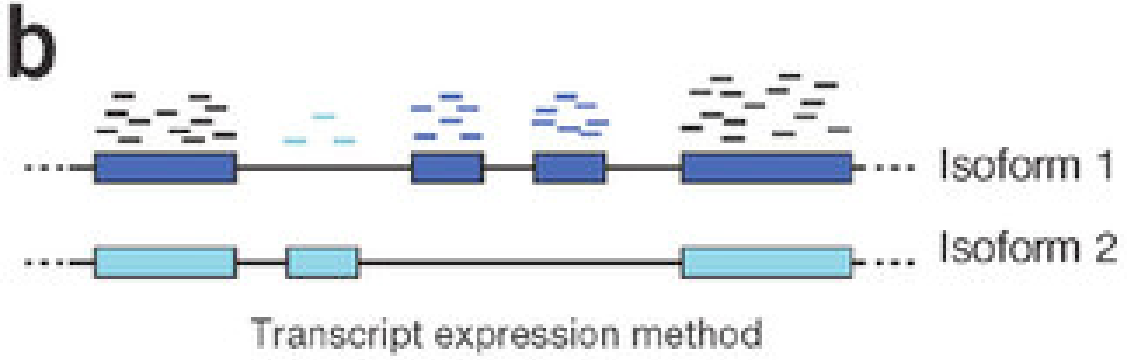
Total gene length after exon flattening: **5kb**
Total reads: **32**
RPKM for gene: **6.4 (=32/5)**

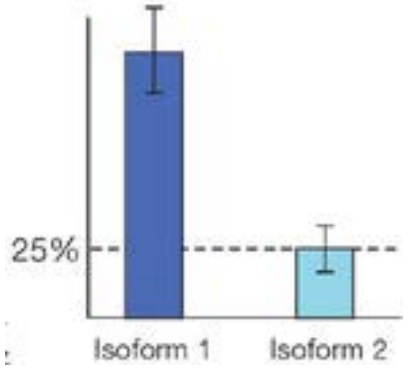
Transcript based approach

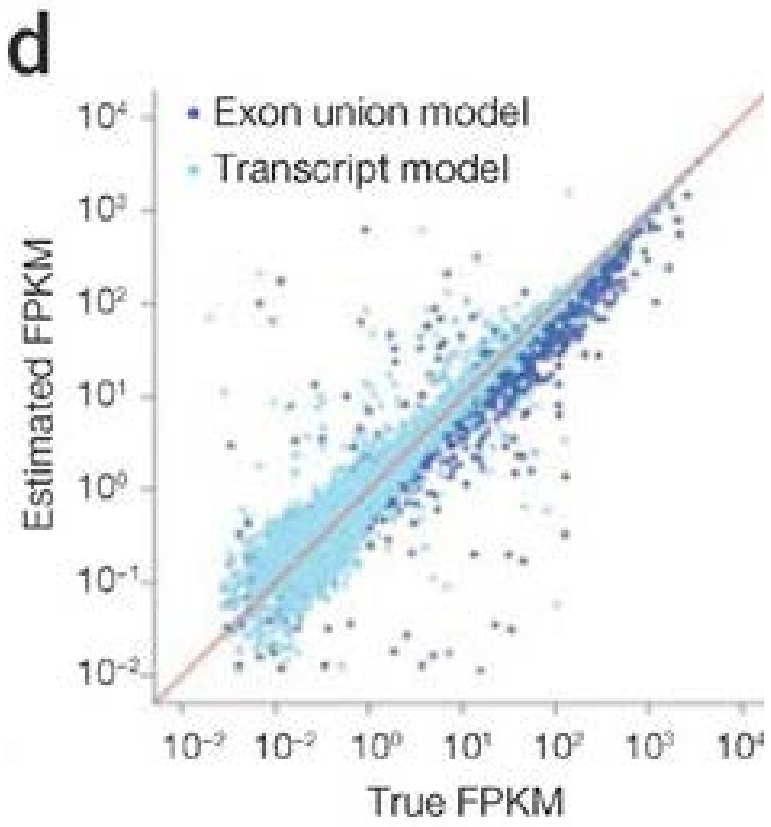


Relative isoform abundance (#1/#2): **25% / 75%**
RPKM for isoform #1 and #2: **2 and 6**
RPKM for gene (=sum of isoforms): **8 (=2+6)**

Garber et al. Nature Methods 8, 469–477 (2011)







Exon union model underestimate gene expression

Transcript model을 이용하는 counting을 해야 한다

Quantification

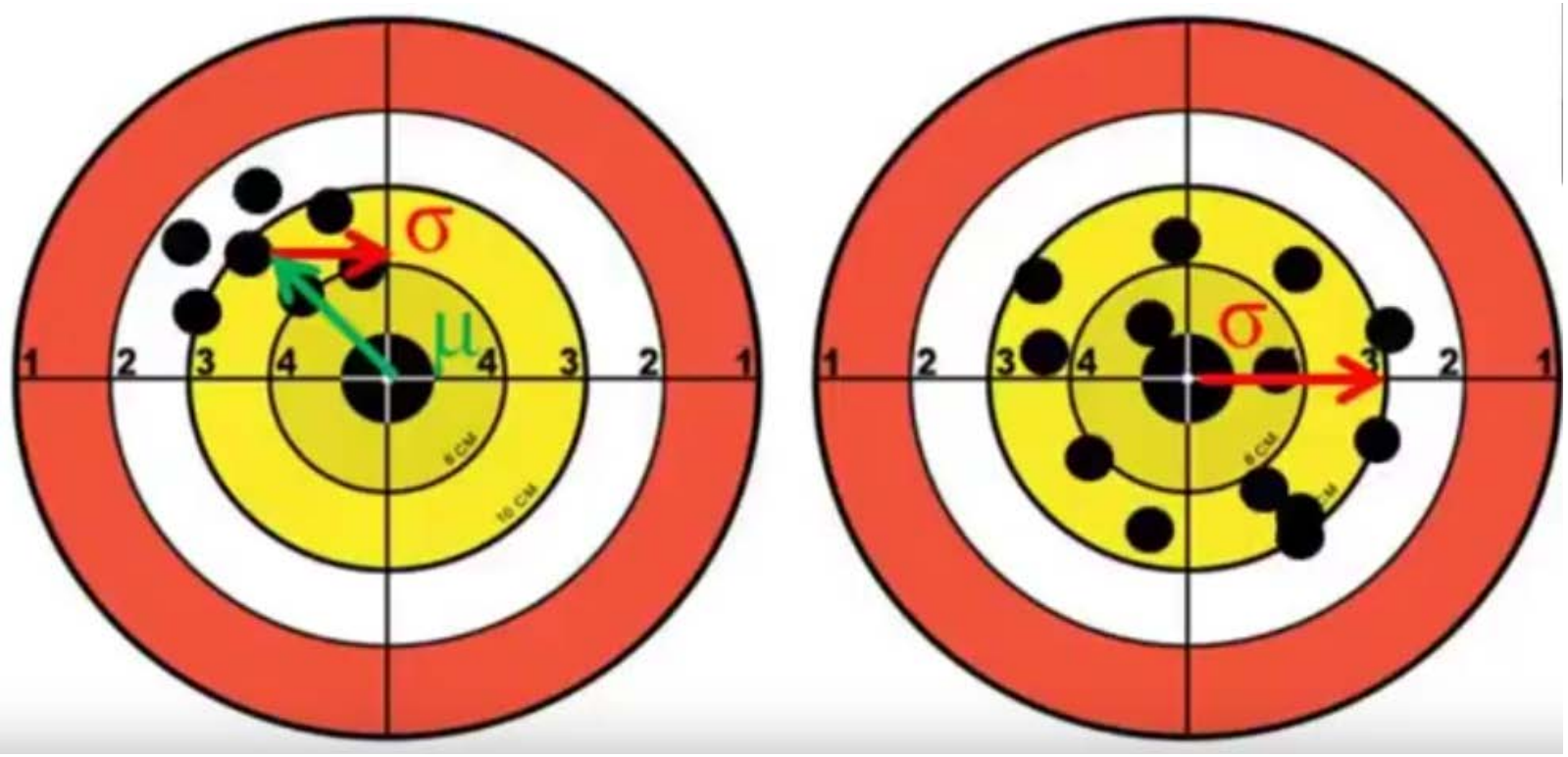
Normalization

Advanced bash commands

(for, while, if, md5sum)

Introduction to R

Normalization



Systematic measurement의 error를 보정하는 목적

Mean and Standard deviation
Z-score normalization

Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Z-score normalization

			mean	sd	(x-mean)/sd		
2	4	4	3.33	1.15	-1.2	0.6	0.6
5	4	14	7.67	5.51	-0.48	-0.67	1.15
4	6	8	6	2	-1	0	1
3	5	8	5.33	2.52	-0.93	-0.13	1.05
3	3	9	5	3.46	-0.58	-0.58	1.15

Quantile normalization example

Original

2	4	4
5	4	14
4	6	8
3	5	8
3	3	9

Ranked

2	3	4
3	4	8
3	4	8
4	5	9
5	6	14

Averaged

3	3	3
5	5	5
5	5	5
6	6	6
8	8	8

Re-ordered

3	5	3
8	5	8
6	8	5
5	6	5
5	3	6

https://www.youtube.com/watch?v=v0j4guy_z30

Quantile normalization 방법 자세히 소개 (쉬운 강의)

https://www.youtube.com/watch?v=v0j4guy_z30

다양한 normalization 방법을 자세히 소개 (어려움)

Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." *Genome biology* 11.10 (2010): R106.

Upper quartile normalization (TCGA method)

https://bioinformatics.stackexchange.com/questions/2586/how-to-apply-upperquartile-normalization-on-rsem-expected-counts