



JOHNS HOPKINS
SCHOOL *of* MEDICINE

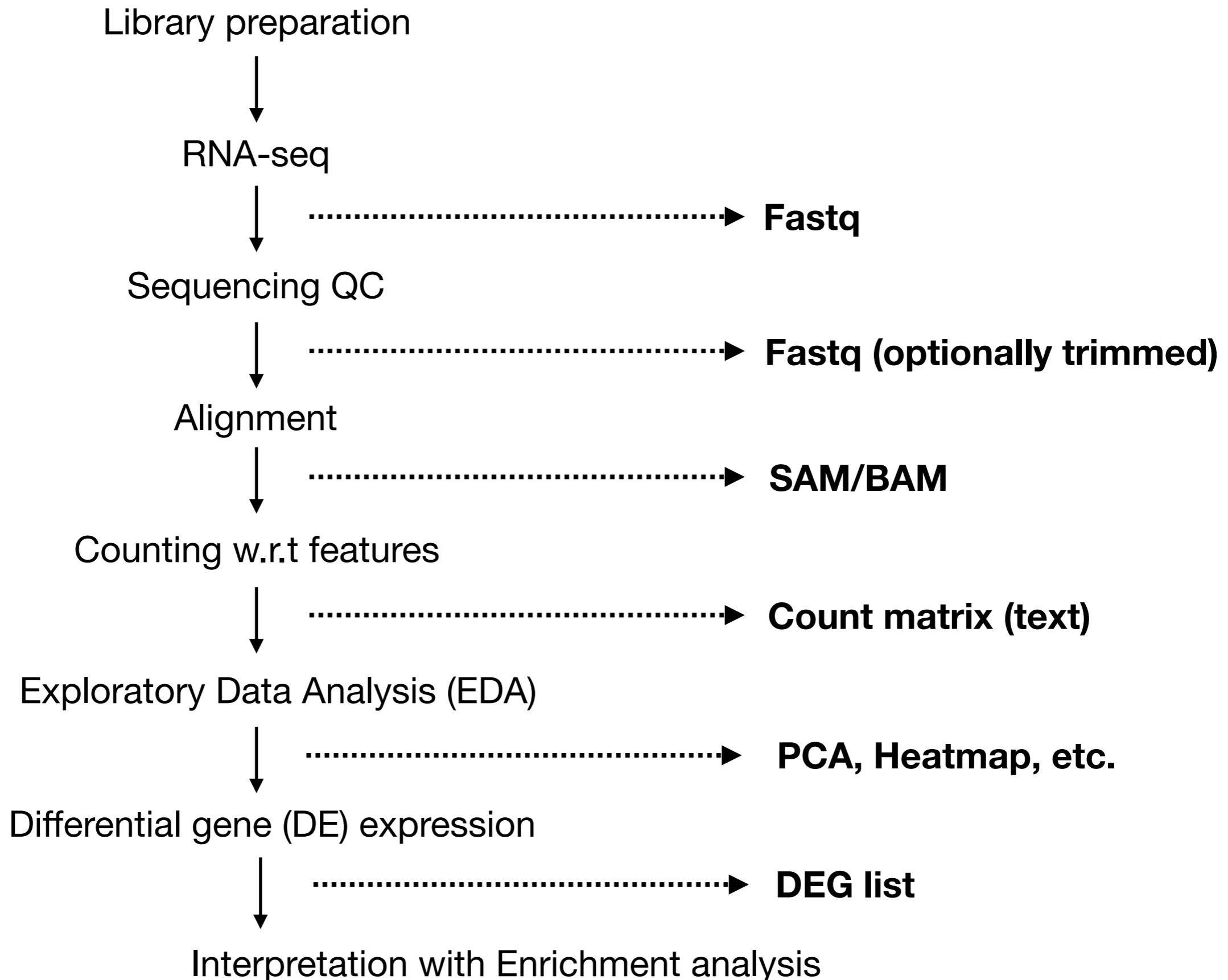
Quantitative Molecular Neurogenomics (ME.440.825)

Module 6 :
Practical RNA-seq workflow &
Enrichment analysis

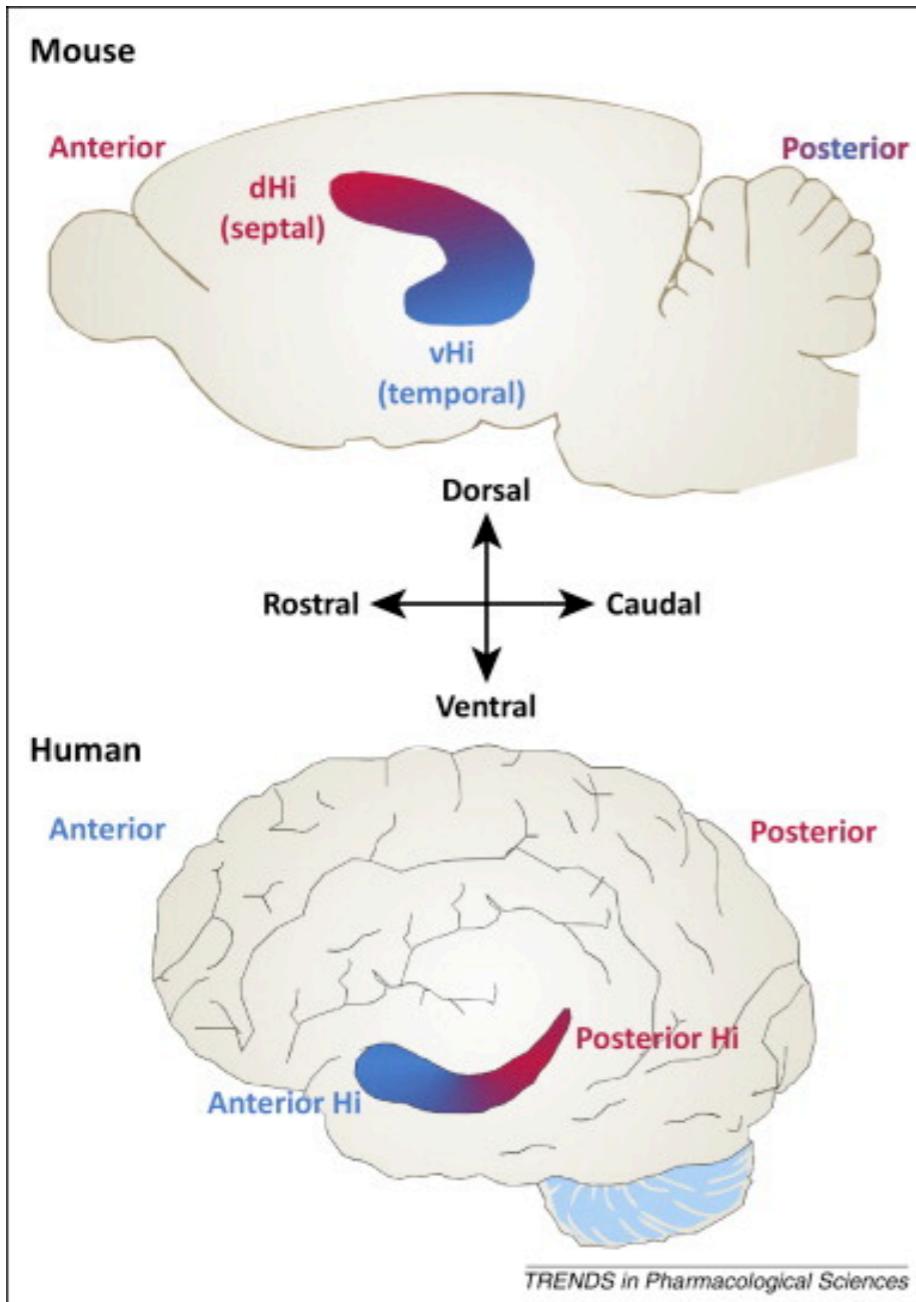
10/10/2022

Taeyoung Hwang, PhD
Lieber Institute for Brain Development,
Department of Neurology and Department of Neuroscience

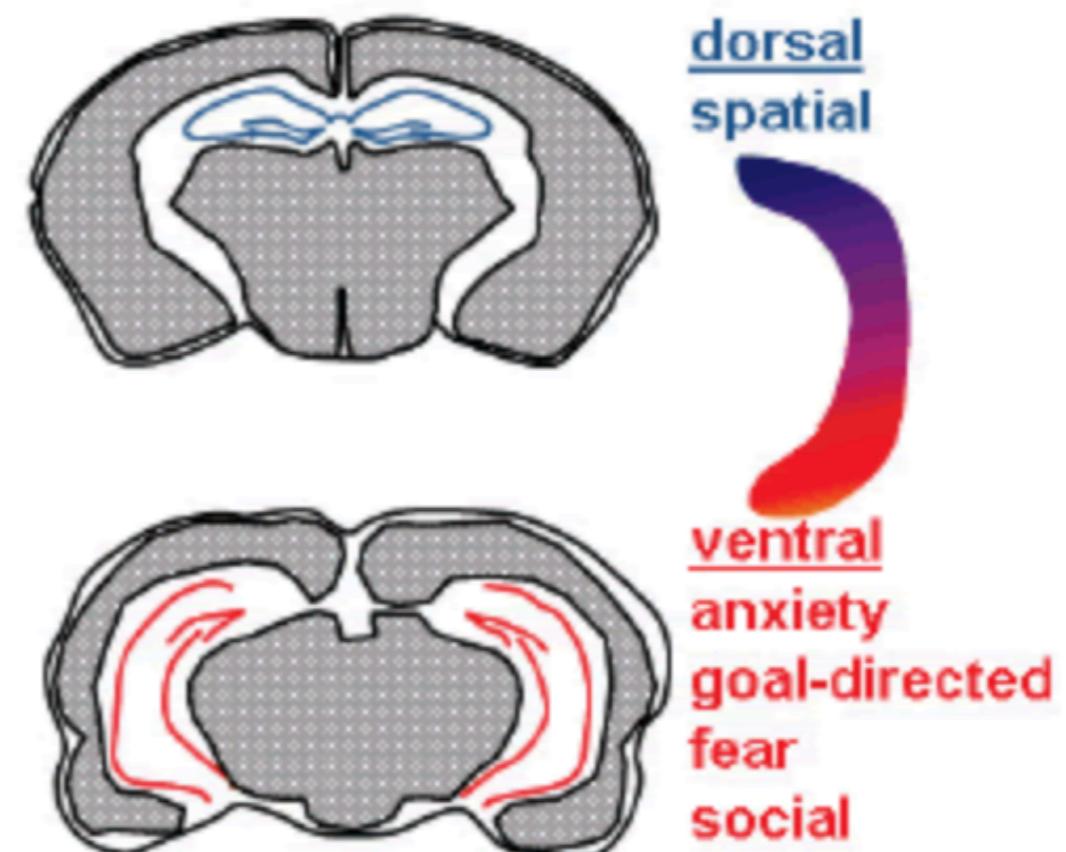
Typical RNA-seq workflow



Sample: mouse hippocampus tissue



The longitudinal axis of the hippocampus



Dataset: RNA-seq of mouse hippocampus

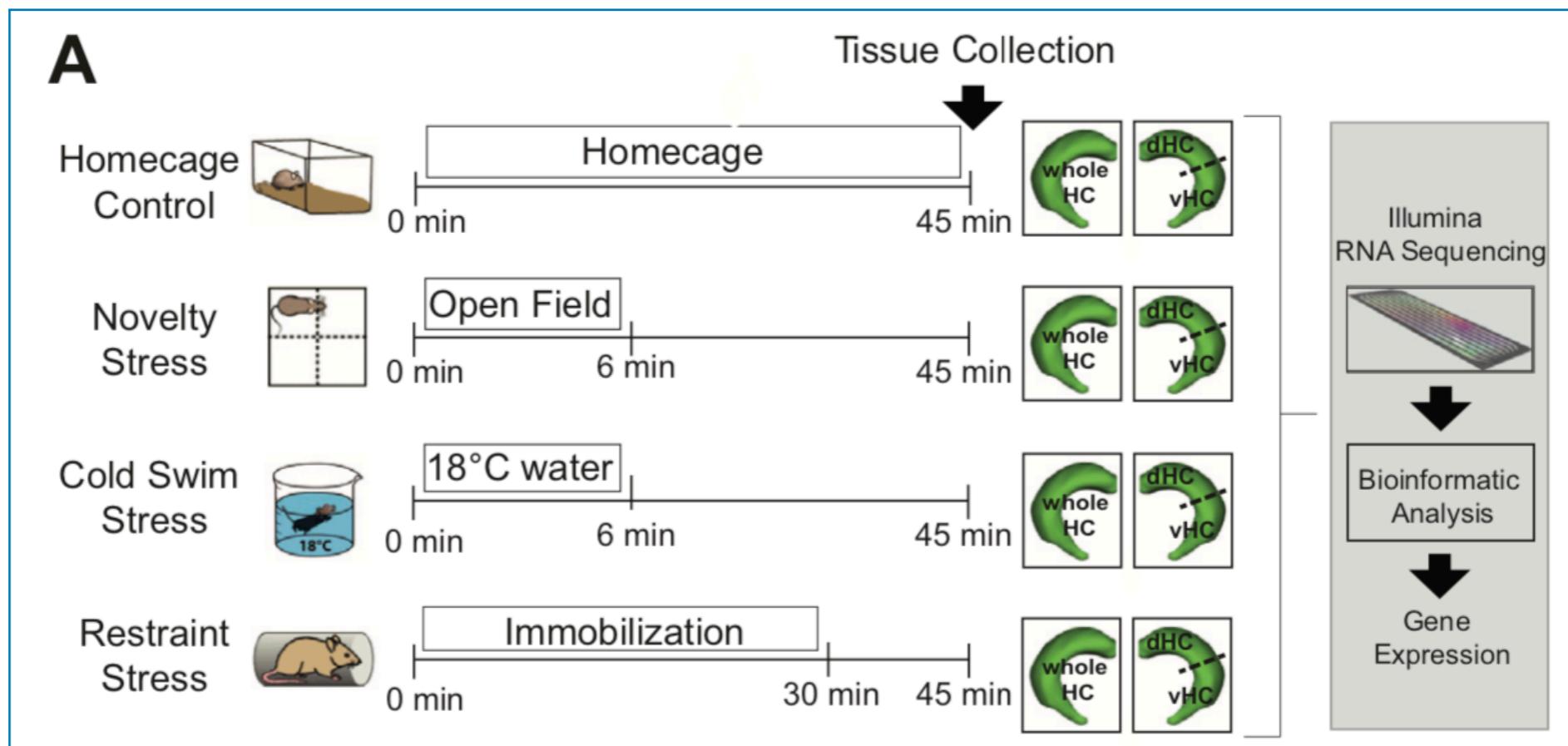
Biological Psychiatry October 1, 2018; 84:531–541

Archival Report

Distinct Proteomic, Transcriptomic, and Epigenetic Stress Responses in Dorsal and Ventral Hippocampus

Amalia Floriou-Servou, Lukas von Ziegler, Luzia Stalder, Oliver Sturman, Mattia Privitera, Anahita Rassi, Alessio Cremonesi, Beat Thöny, and Johannes Bohacek

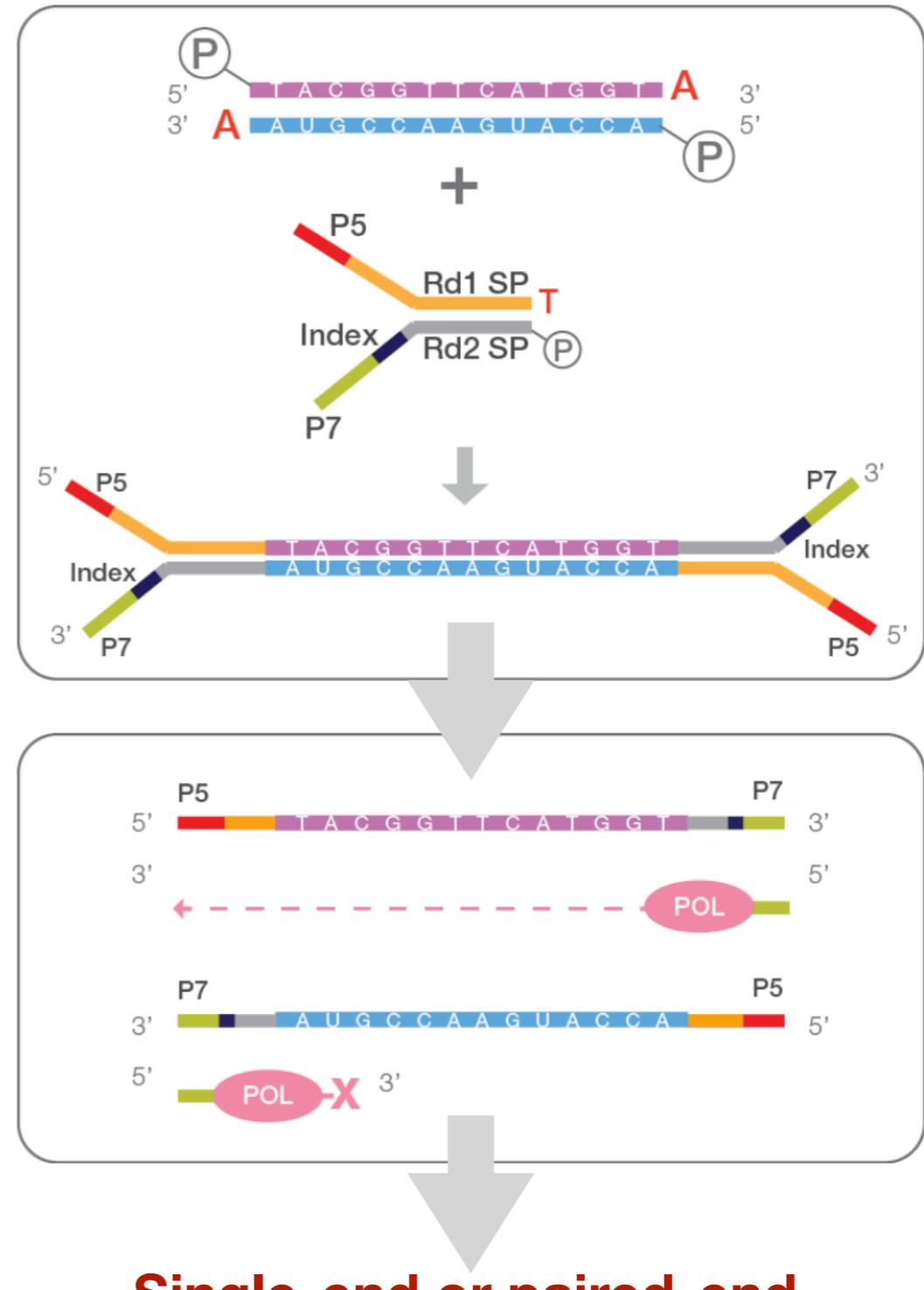
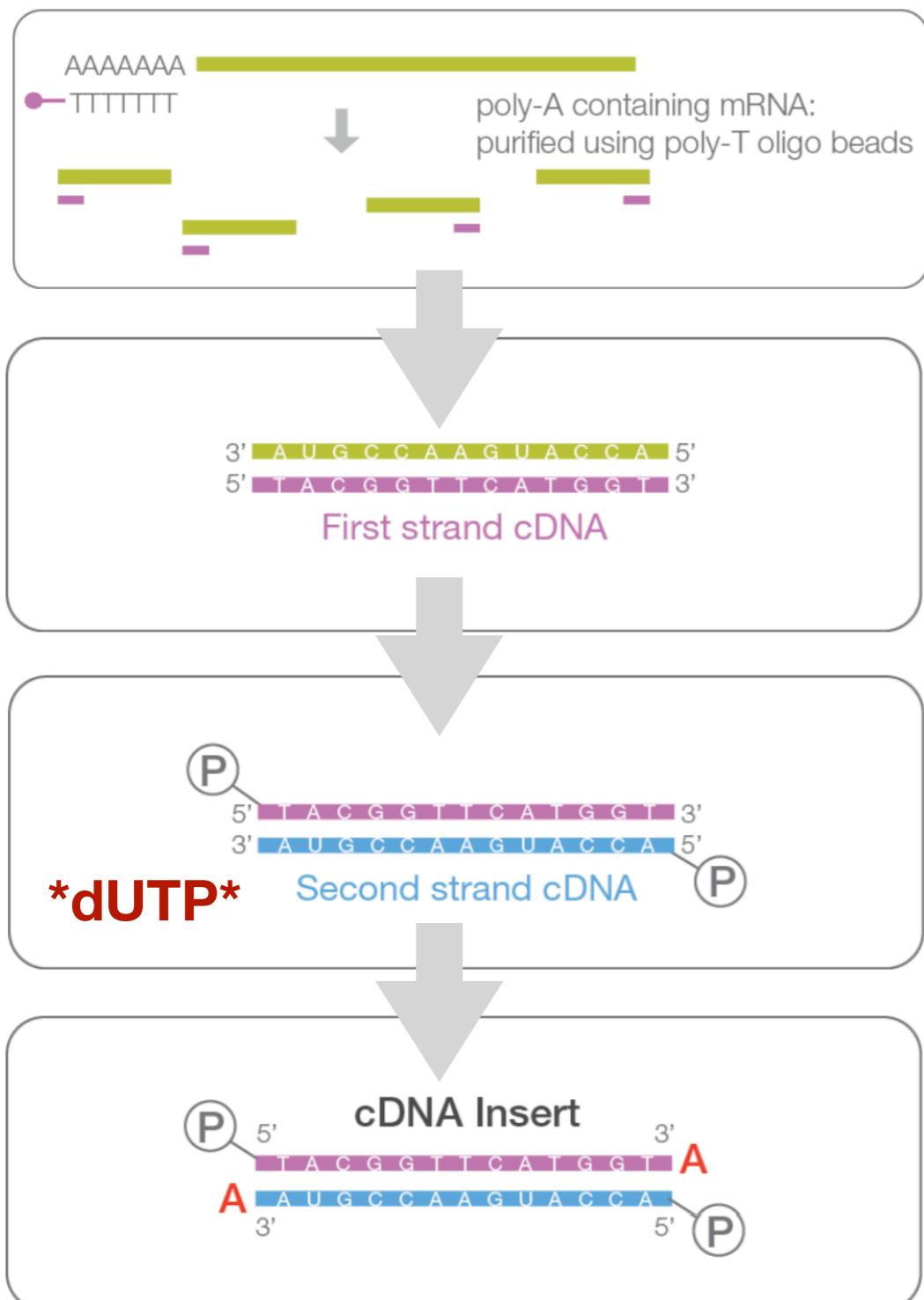
1	Sid	Region	Replicate
2	SRR5715043	Dorsal	Rep1
3	SRR5715044	Dorsal	Rep2
4	SRR5715045	Dorsal	Rep3
5	SRR5715063	Ventral	Rep1
6	SRR5715064	Ventral	Rep2
7	SRR5715065	Ventral	Rep3



RNA-seq library preparation

Supplementary Method

“For library preparation, the TruSeq stranded RNA kit (Illumina Inc.) was used according to the manufacturer’s protocol.”



Sequencing QC using FastQC

```
$ fastqc SRR5715043.fastq.gz
```

 **FastQC Report**

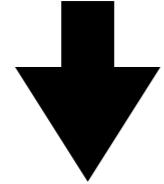
Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)

 **Basic Statistics**

Measure	Value
Filename	SRR5715043.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36027995
Sequences flagged as poor quality	0
Sequence length	126
%GC	49

 **Per base sequence quality**



****No worries****

Fastqc html report: an example

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)



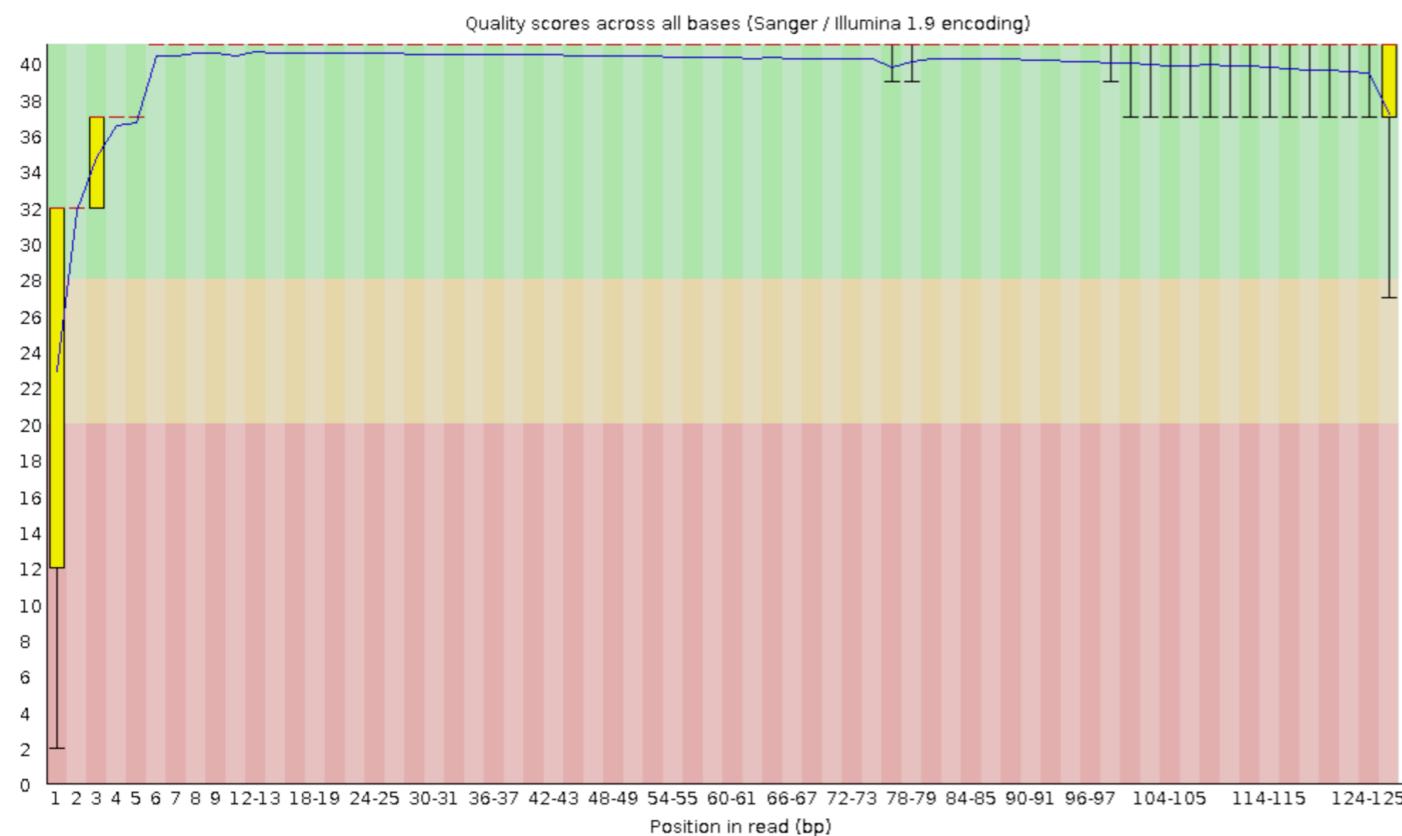
Basic Statistics

Measure	Value
Filename	SRR5715043.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36027995
Sequences flagged as poor quality	0
Sequence length	126
%GC	49

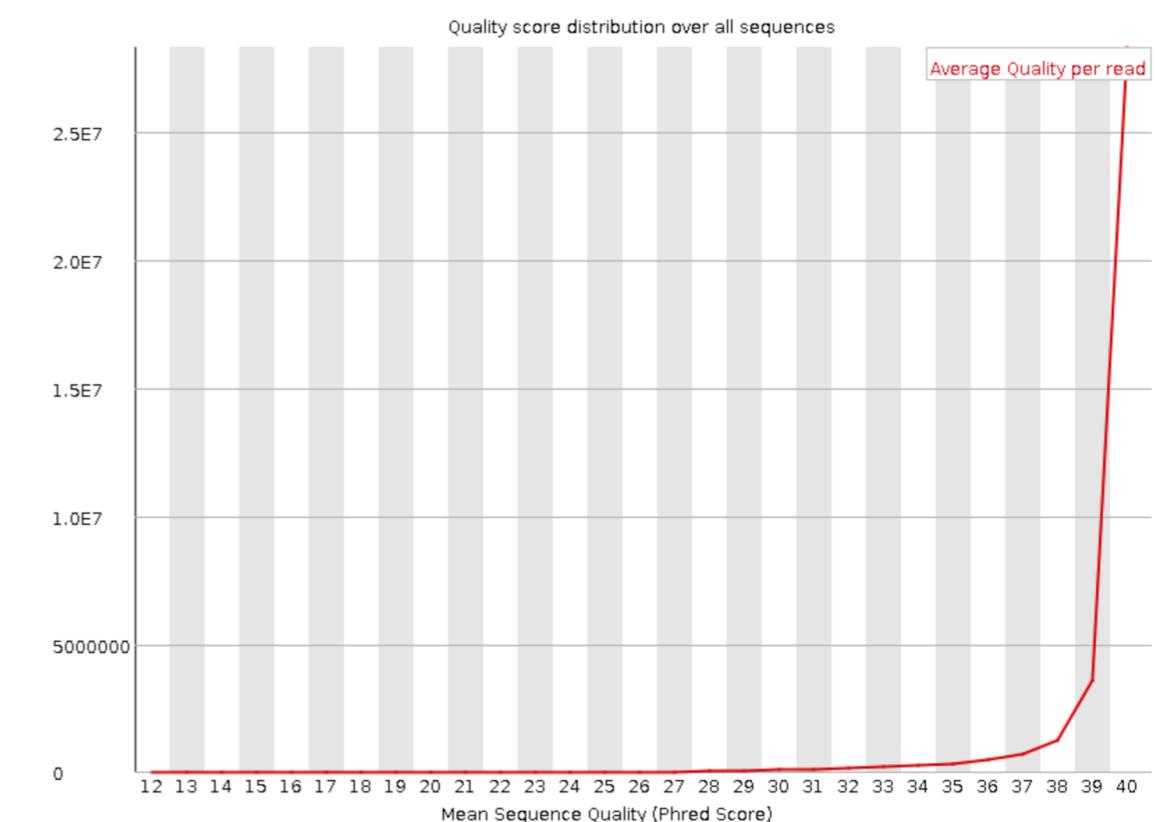
GC % in species

Species	Sequences used (Mb) ¹	CDS count ²	Exon count ³	Intron count	Average Exon length (bp)	Average Exon GC (%)	Average Intron length (bp)	Average Intron GC (%)
<i>A. thaliana</i>	119.2	23488	135697	118838	223.7	44.1	163.7	32.7
<i>B. taurus</i>	2434	16829	162223	151199	162.3	52.0	4516.4	46.9
<i>C. elegans</i>	100.3	27123	171102	149895	208.4	43.0	334.7	29.1
<i>C. familiaris</i>	2445	15960	161238	147118	157.2	50.6	3535.5	46.1
<i>D. melanogaster</i>	118.4	8119	43847	40732	370.0	52.7	1530.7	36.5
<i>D. rerio</i>	1547	20256	173438	156972	156.2	48.7	2276.4	34.5
<i>G. gallus</i>	1032	14367	142329	133690	152.1	48.2	2811.8	42.6
<i>H. sapiens</i>	3077	40430	381122	378089	162.7	50.8	5848.7	45.9
<i>M. musculus</i>	2644	30546	254605	243458	170.5	51.1	4683.6	46.0
<i>O. sativa</i>	372.1	41046	178106	144863	306.0	51.0	396.1	37.7
<i>P. troglodytes</i>	3176	30010	296830	287132	156.2	50.7	6003.7	45.3
<i>R. norvegicus</i>	2719	22243	195940	186864	172.9	51.1	4406.4	46.3
<i>T. nigroviridis</i>	217.3	15455	122253	108901	174.3	54.5	599.5	45.4

Per base sequence quality



Per sequence quality scores



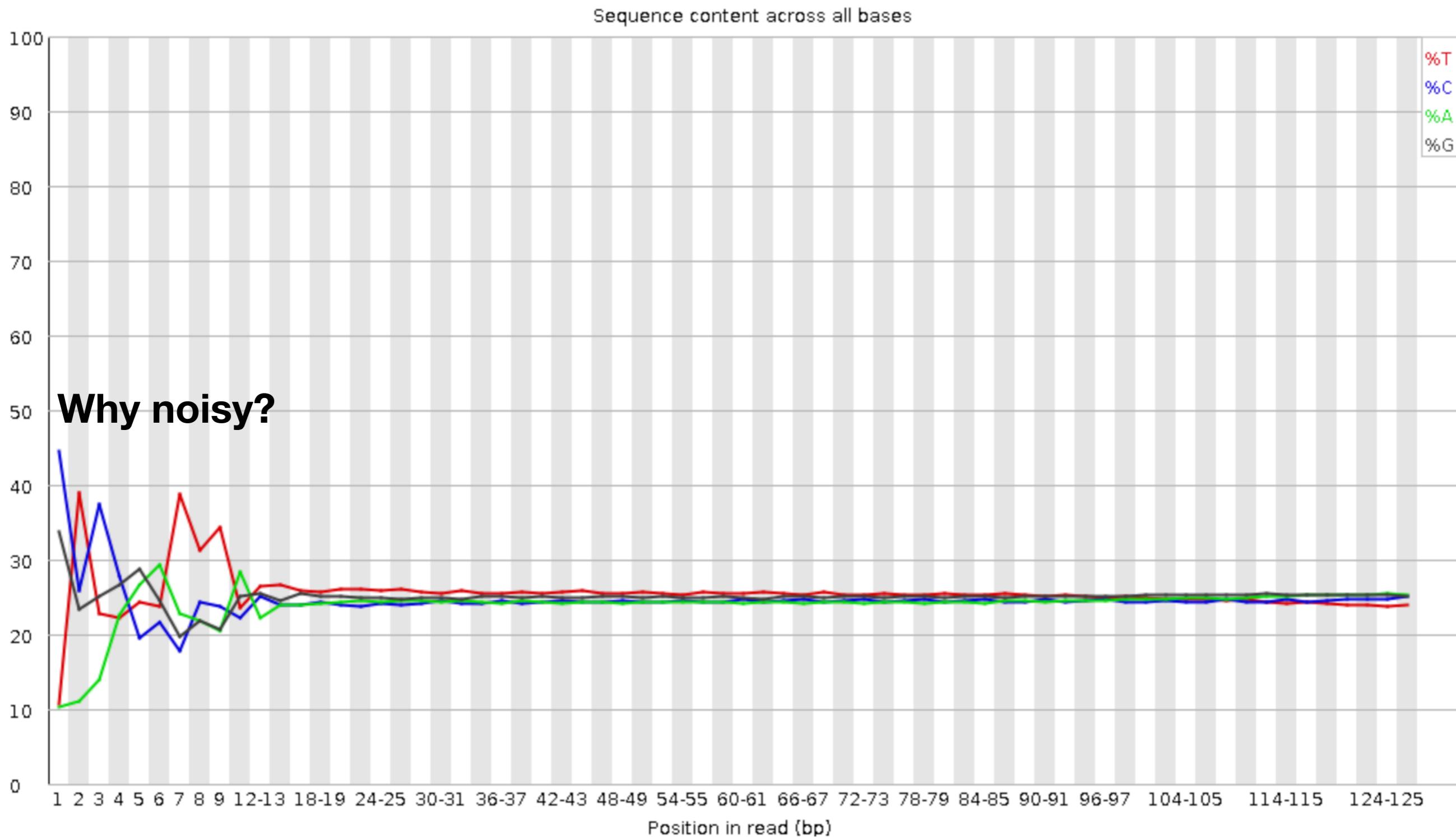
Phred Quality Score

10
20
30
40

Probability of incorrect base call

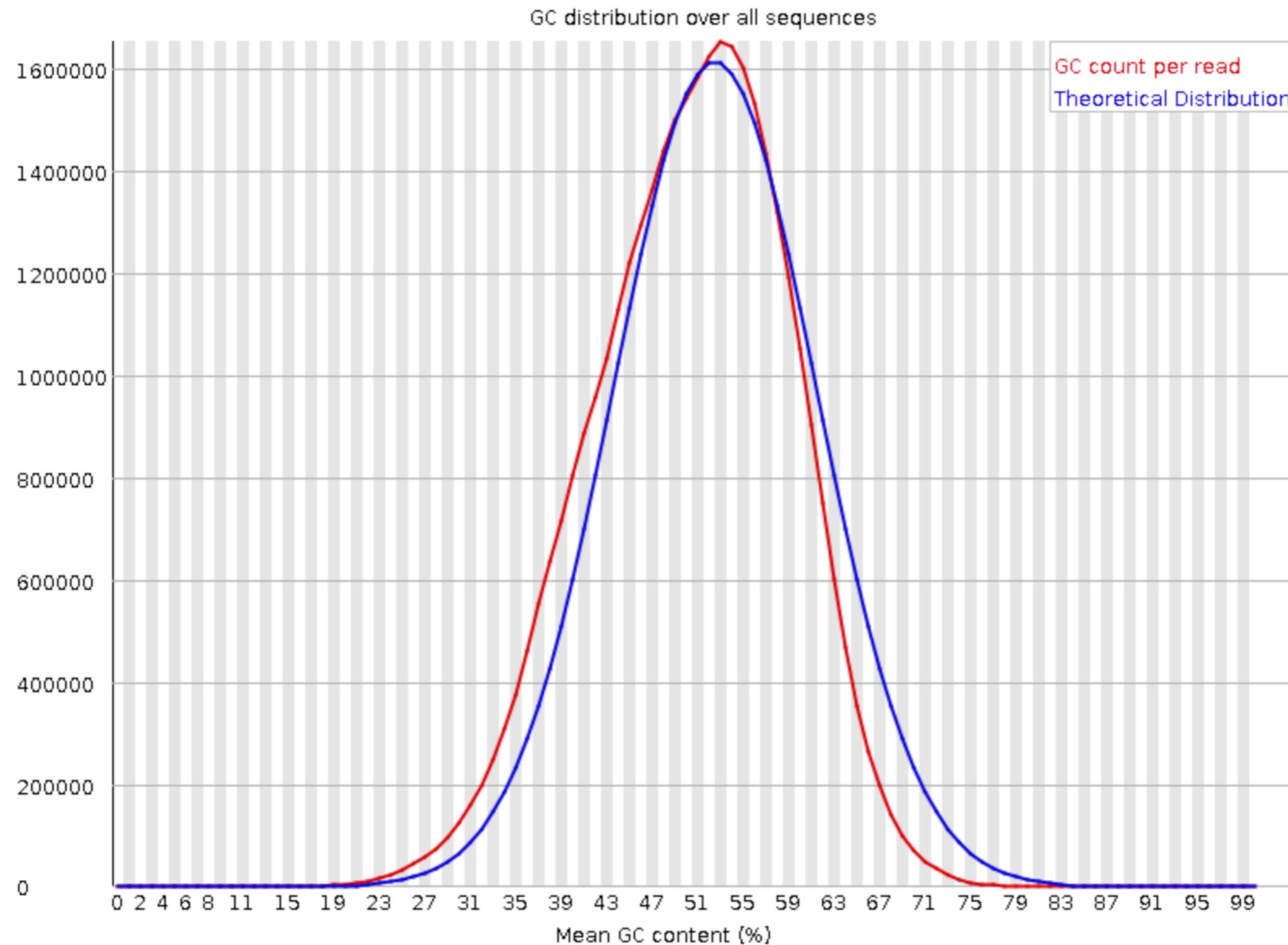
1 in 10
1 in 100
1 in 1000
1 in 10,000

✖ Per base sequence content

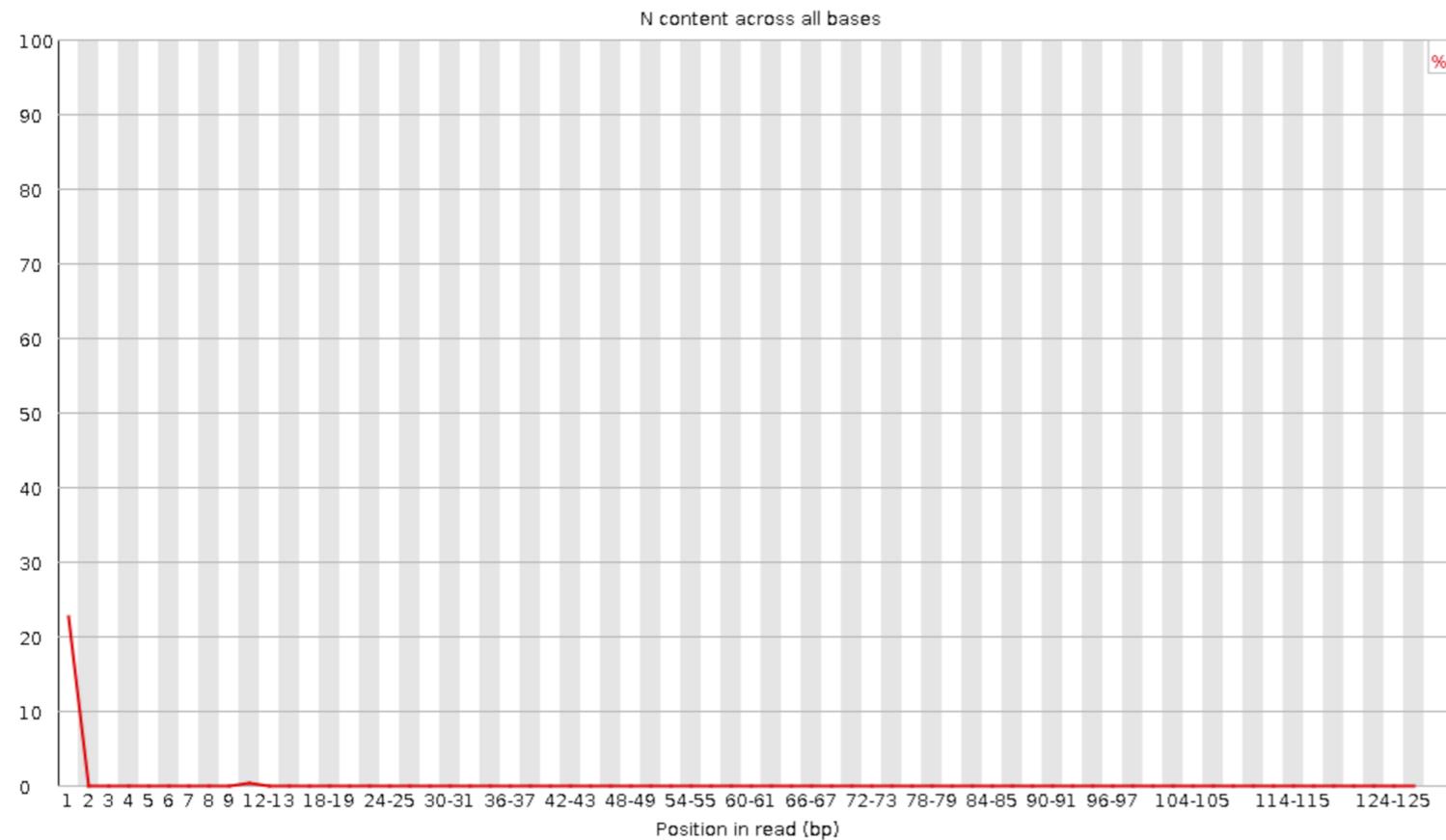




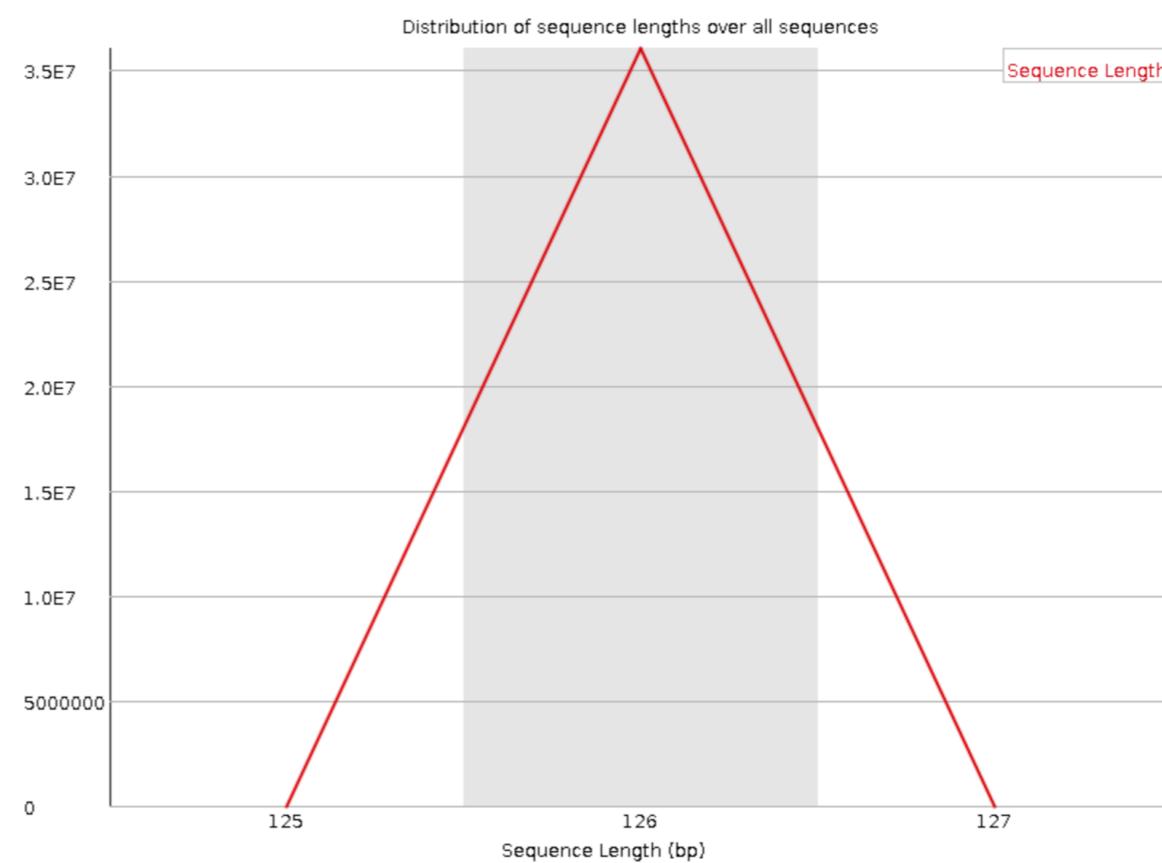
Per sequence GC content



Per base N content

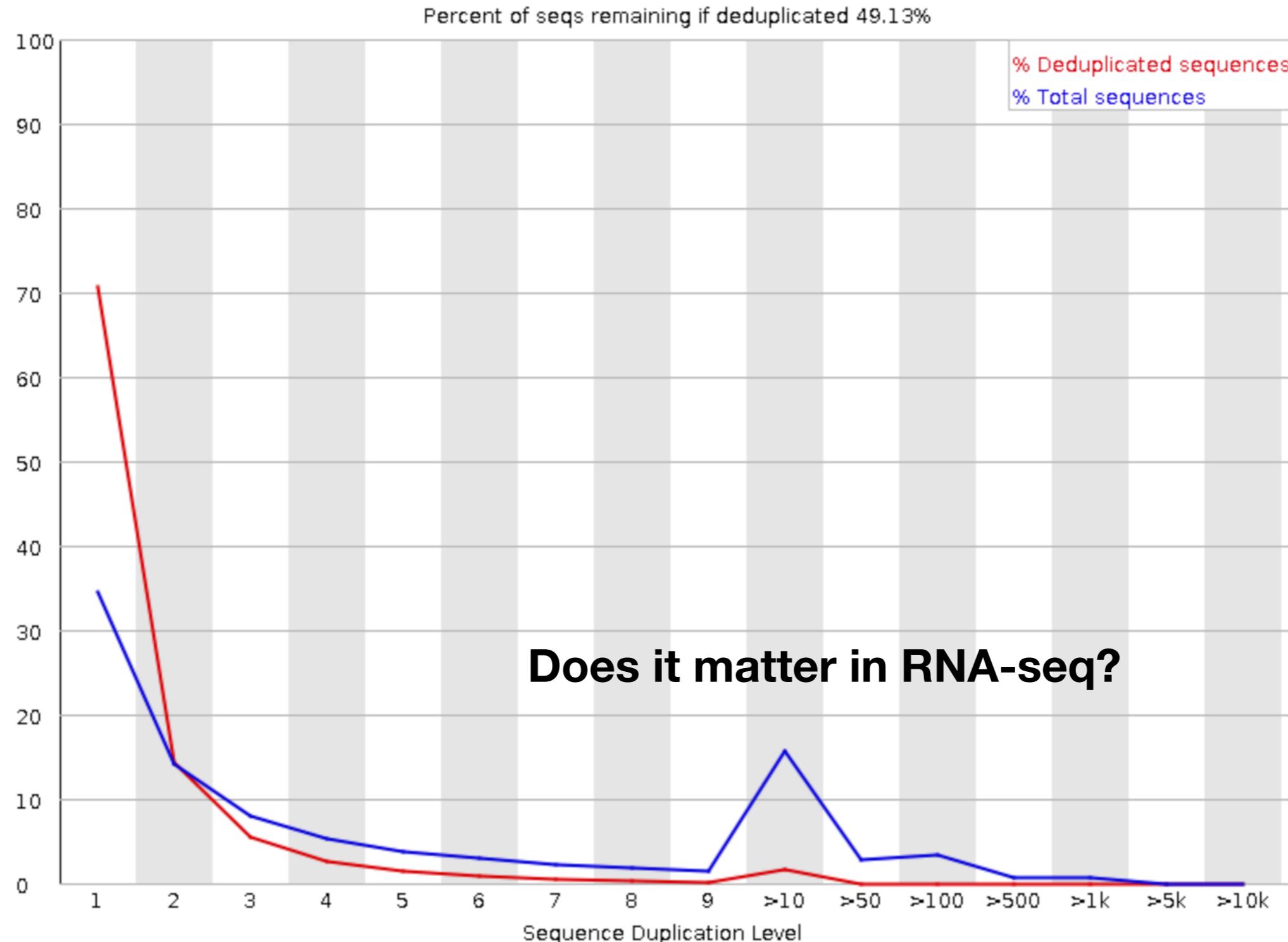


Sequence Length Distribution





Sequence Duplication Levels



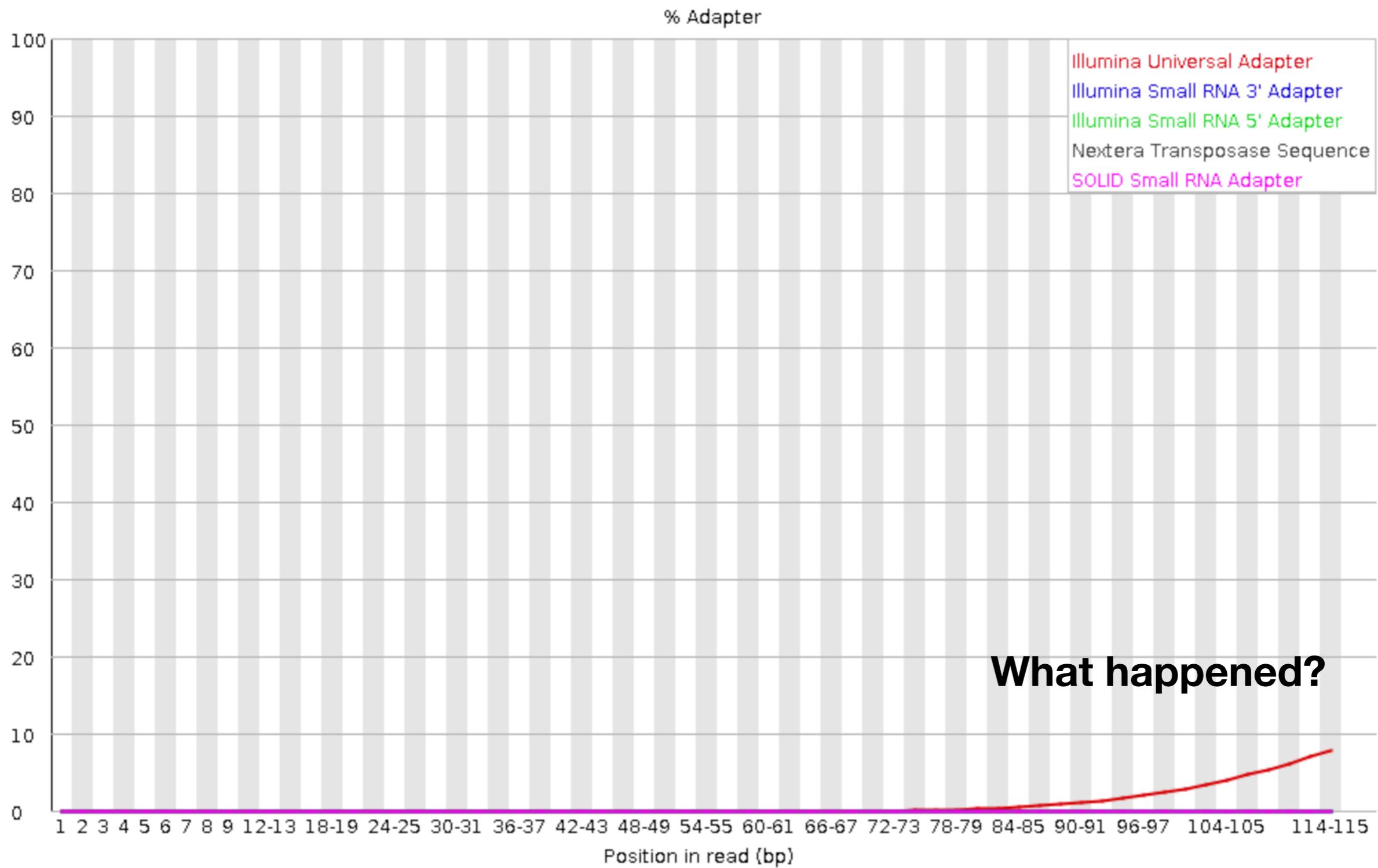


Overrepresented sequences

No overrepresented sequences



Adapter Content



Trimming FastQC using cutadapt

```
$ cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -m 20 -o SRR5715043.trimmed.fastq.gz SRR5715043.fastq.gz
```

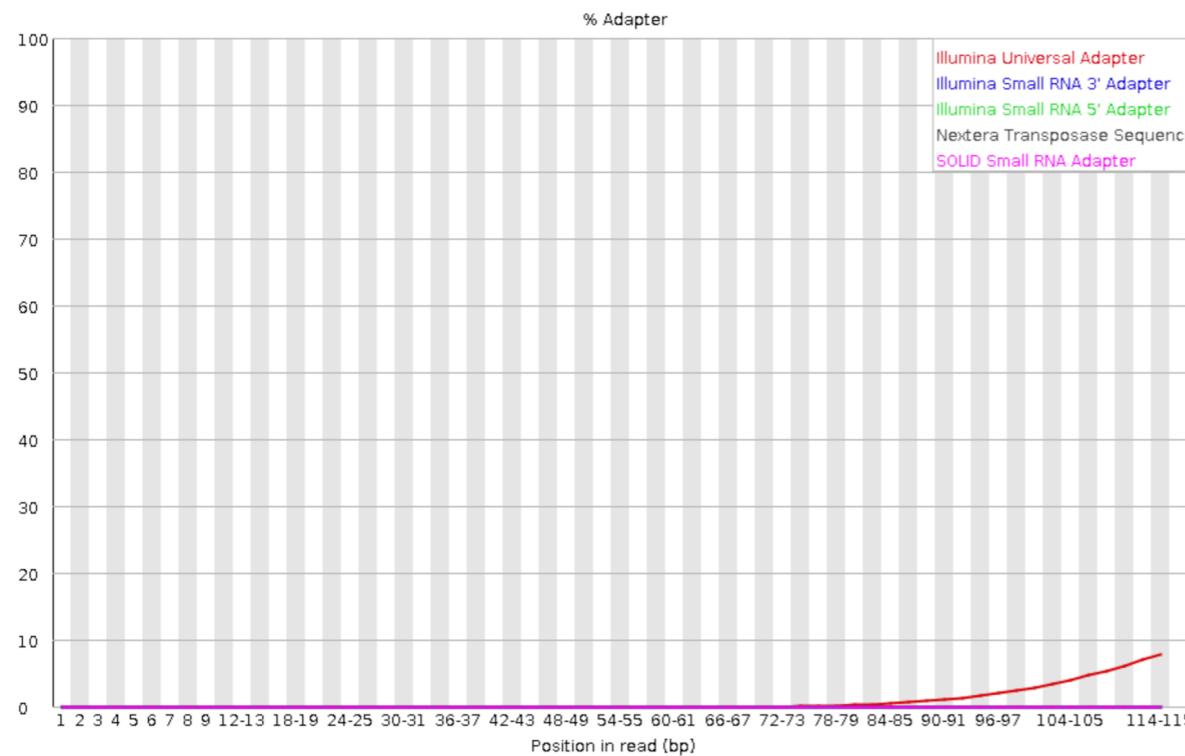
Before

Measure	Value
Filename	SRR5715043.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36027995
Sequences flagged as poor quality	0
Sequence length	126
%GC	49

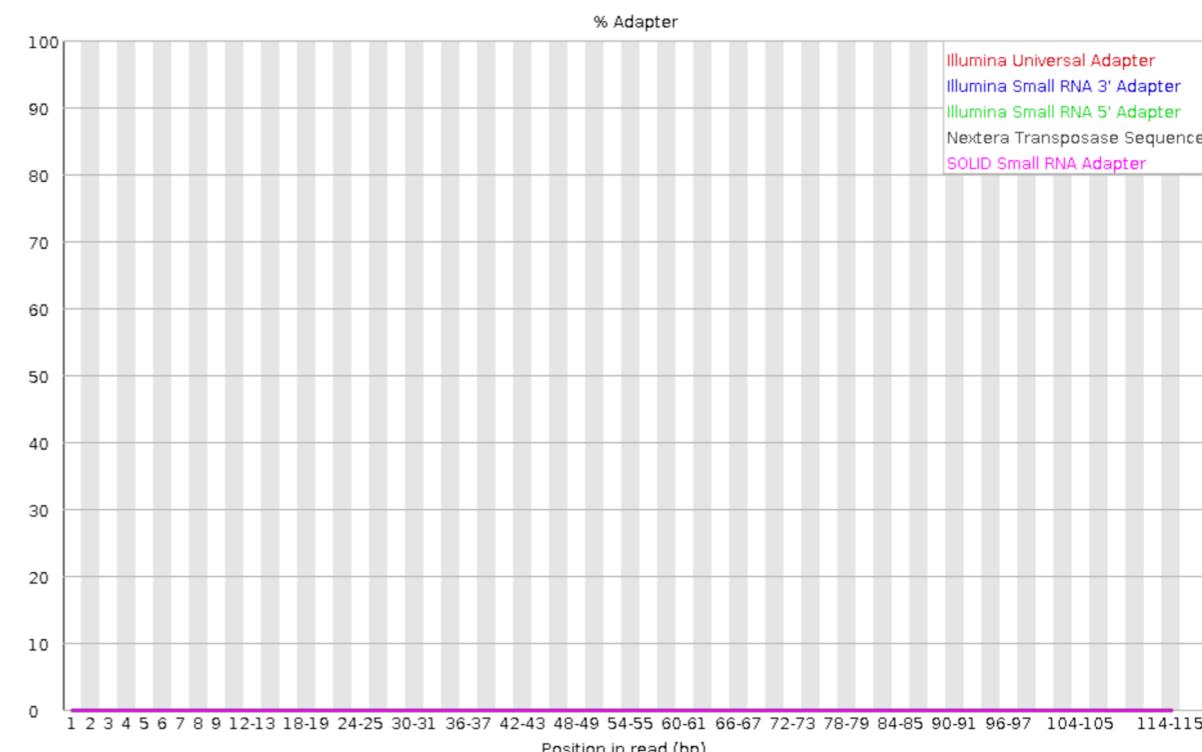
After

Measure	Value
Filename	SRR5715043.trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36022775
Sequences flagged as poor quality	0
Sequence length	20-126
%GC	49

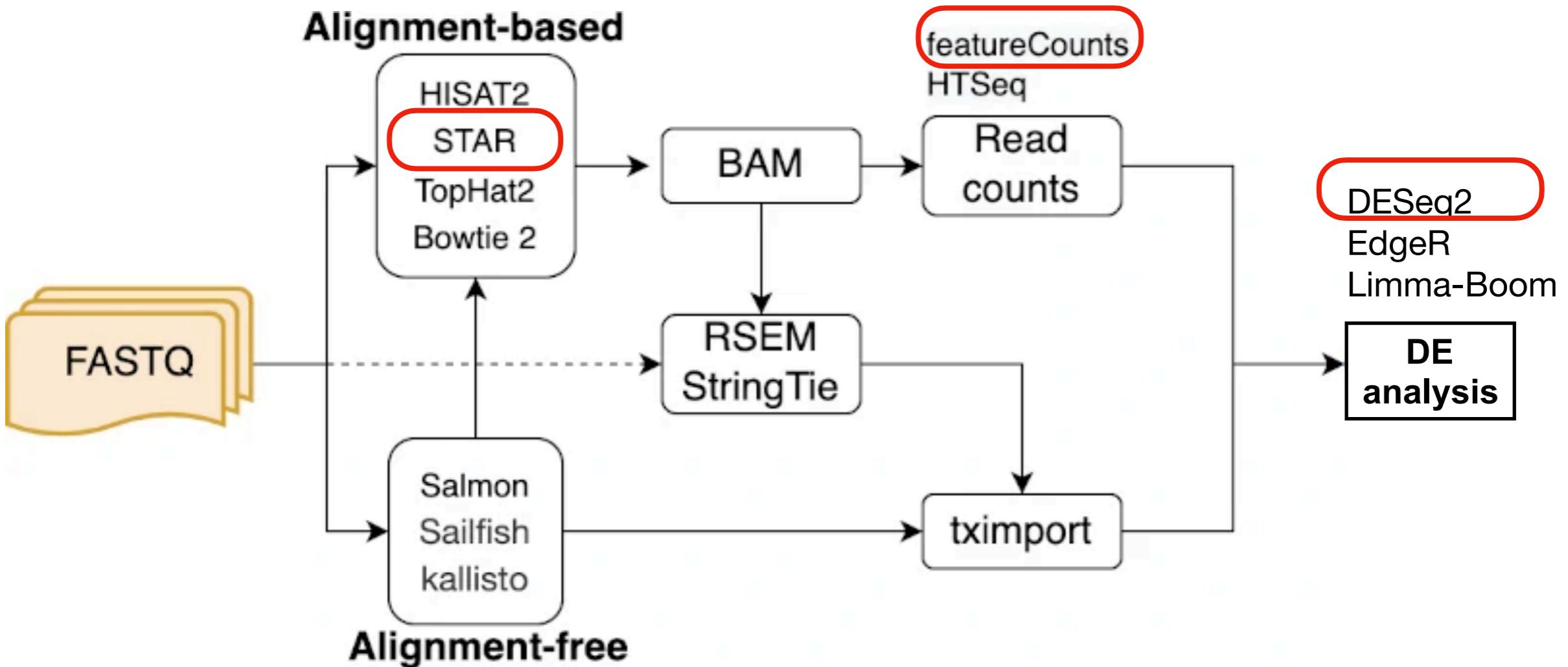
Adapter Content



Adapter Content



Our RNA-seq tools



Alignment using STAR

Spliced Transcripts Alignments to a Reference (STAR)

Indexing (only one time per genome+transcript annotation)

```
$STAR --runMode genomeGenerate \
--runThreadN 6 \
--genomeFastaFiles genome.fa \
--sjdbGTFfile annotation.gtf \
--sjdbOverhang 99 \
--genomeDir STAR_index
```

Mapping (multiple samples)

```
$STAR --genomeDir ./Genome/STAR_Index \
--runThreadN 4 \
--readFilesCommand zcat \
--readFilesIn SRR5715043.trimmed.fastq.gz \
--outSAMtype BAM SortedByCoordinate \
--outFileNamePrefix SRR5715043.
```

A review of module 5

Excerpted from Loyal Goff's slide

Alignment QC - What are we looking for?

- **Number** of reads aligned per sample
- **Percentage** of reads aligned per sample
- **Number/percentage** of reads uniquely mapping
 - % of reads with multiple mapping to reference (ambiguous)
- Paired End read **concordance**
 - # of PE reads uniquely mapping
 - # of PE reads mapping but not together (discordant)
 - # of PE reads where only one mapped [uniquely]
- **Consistent** scores/values across samples
- *Values are often reported after each sample is processed. Useful to try and capture this information in a log file.*

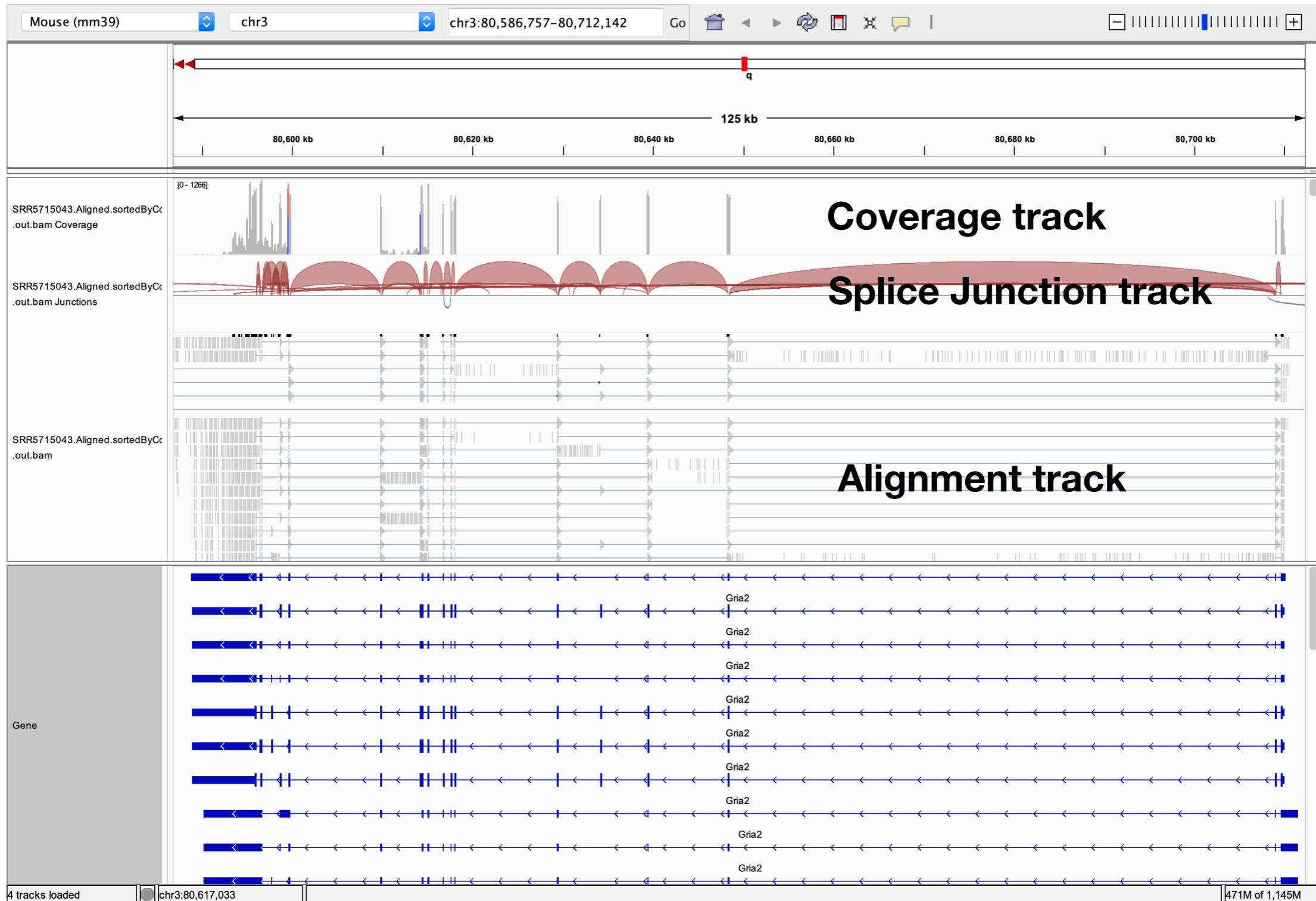
STAR Output example

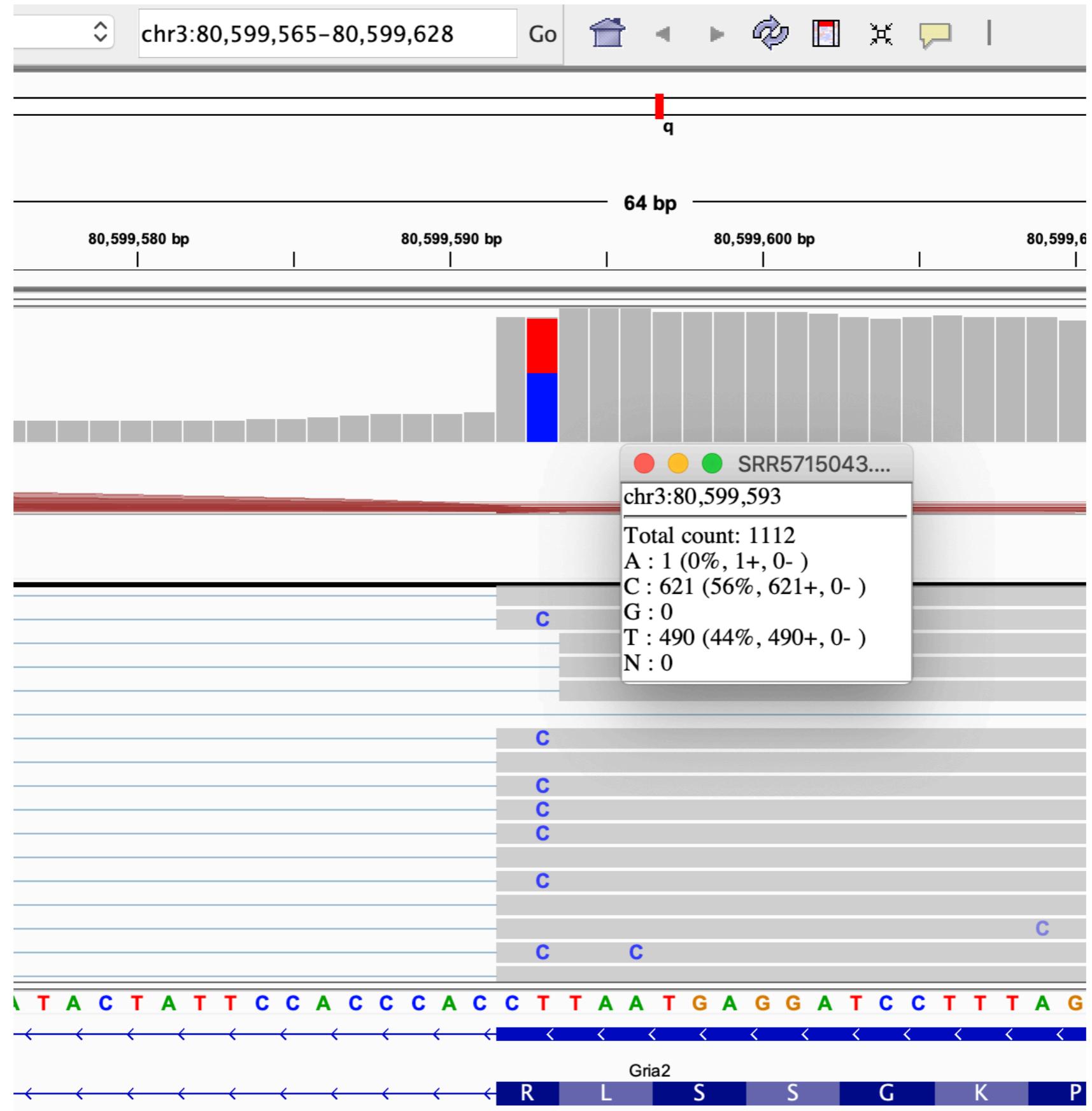
```
$ cat SRR5715043.Log.final.out
```

Started job on	Oct 03 17:10:04
Started mapping on	Oct 03 17:10:55
Finished on	Oct 03 17:16:23
Mapping speed, Million of reads per hour	395.37
Number of input reads	36022775
Average input read length	123
UNIQUE READS:	
Uniquely mapped reads number	33636271
Uniquely mapped reads %	93.38%
Average mapped length	123.01
Number of splices: Total	11521857
Number of splices: Annotated (sjdb)	11418927
Number of splices: GT/AG	11399495
Number of splices: GC/AG	98254
Number of splices: AT/AC	12071
Number of splices: Non-canonical	12037
Mismatch rate per base, %	0.11%
Deletion rate per base	0.01%
Deletion average length	1.96
Insertion rate per base	0.00%
Insertion average length	1.54
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	1724928
% of reads mapped to multiple loci	4.79%
Number of reads mapped to too many loci	74776
% of reads mapped to too many loci	0.21%
UNMAPPED READS:	
Number of reads unmapped: too many mismatches	0
% of reads unmapped: too many mismatches	0.00%
Number of reads unmapped: too short	536652
% of reads unmapped: too short	1.49%
Number of reads unmapped: other	50148
% of reads unmapped: other	0.14%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%

Visualization with IGV

```
$ samtools index  
SRR5715043.Aligned.sortedByCoord.out.bam
```

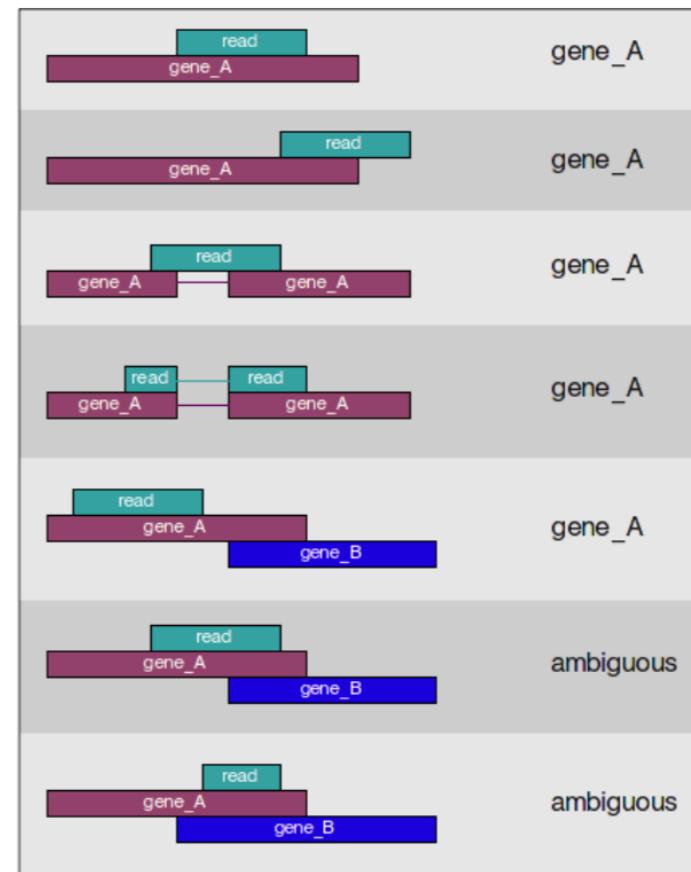




Counting reads using featureCounts

<https://subread.sourceforge.net/featureCounts.html>

1. Count the reads for every feature (by default, exon).
2. Sum the feature counts according to meta features (for example, gene).



```
$featureCounts SRR5715043.Aligned.sortedByCoord.out.bam \
-a gencode.vM30.annotation.gtf \
-o SRR5715043.fCounts \
-p # Paired-end (not in our case)
-s 2 \ # Strand of read 1 : 0 (unstranded), 1 (stranded), 2 (rc stranded)
-g gene_id # Meta feature ( a set of features) in GTF
-T 4 \ # Number of threads
```

FeatureCounts Output example

Standard output

```
=====
===== S|D|D|P|E|A|D|
===== v2.0.3

//===== featureCounts setting =====\\
Input files : 1 BAM file
              SRR5715043.Aligned.sortedByCoord.out.bam
Output file : SRR5715043.fCounts
Summary : SRR5715043.fCounts.summary
Paired-end : no
Count read pairs : no
Annotation : gencode.vM30.annotation.gtf (GTF)
Dir for temp files : ./3.Count

Threads : 4
Level : meta-feature level
Multimapping reads : not counted
Multi-overlapping reads : not counted
Min overlapping bases : 1
\\=====

//===== Running =====\\
Load annotation file gencode.vM30.annotation.gtf ...
Features : 868862
Meta-features : 56691
Chromosomes/contigs : 22

Process BAM file SRR5715043.Aligned.sortedByCoord.out.bam...
Strand specific : reversely stranded
Single-end reads are included.
Total alignments : 37797275
Successfully assigned alignments : 28832430 (76.3%)
Running time : 0.16 minutes

Write the final count table.
Write the read assignment summary.
8 Summary of counting results can be found in file "./3.Count/SRR5715043.fCounts.summary"
\\=====
```

SRR5715043.fCounts.summary

```
'Status ./2.Alignment/SRR5715043.Aligned.sortedByCoord
Assigned 28832430
Unassigned_Unmapped 0
Unassigned_Read_Type 0
Unassigned_Singleton 0
Unassigned_MappingQuality 0
Unassigned_Chimera 0
Unassigned_FragmentLength 0
Unassigned_Duplicate 0
Unassigned_MultiMapping 4161004
Unassigned_Secondary 0
Unassigned_NonSplit 0
Unassigned_NoFeatures 4381820
Unassigned_Overlapping_Length 0
Unassigned_Ambiguity 422021
```

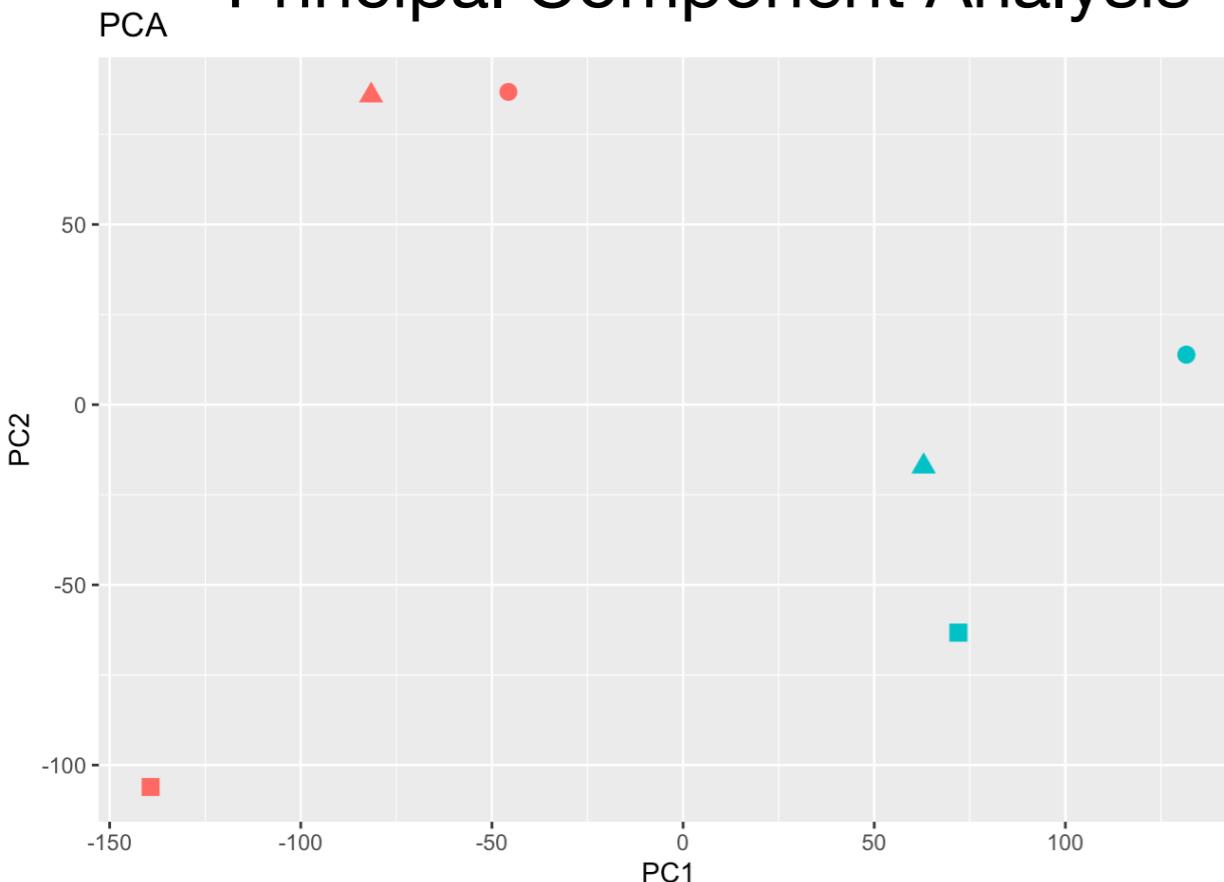
Exploratory Data Analysis (EDA)

Normalization should be done before any EDA.

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same sample group; NOT for within sample comparisons or DE analysis
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; NOT for DE analysis
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; NOT for between sample comparisons or DE analysis
DESeq2's median of ratios [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for DE analysis; NOT for within sample comparisons
EdgeR's trimmed mean of M values (TMM) [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth and RNA composition	gene count comparisons between samples and for DE analysis; NOT for within sample comparisons

Exploratory Data Analysis (EDA)

Principal Component Analysis



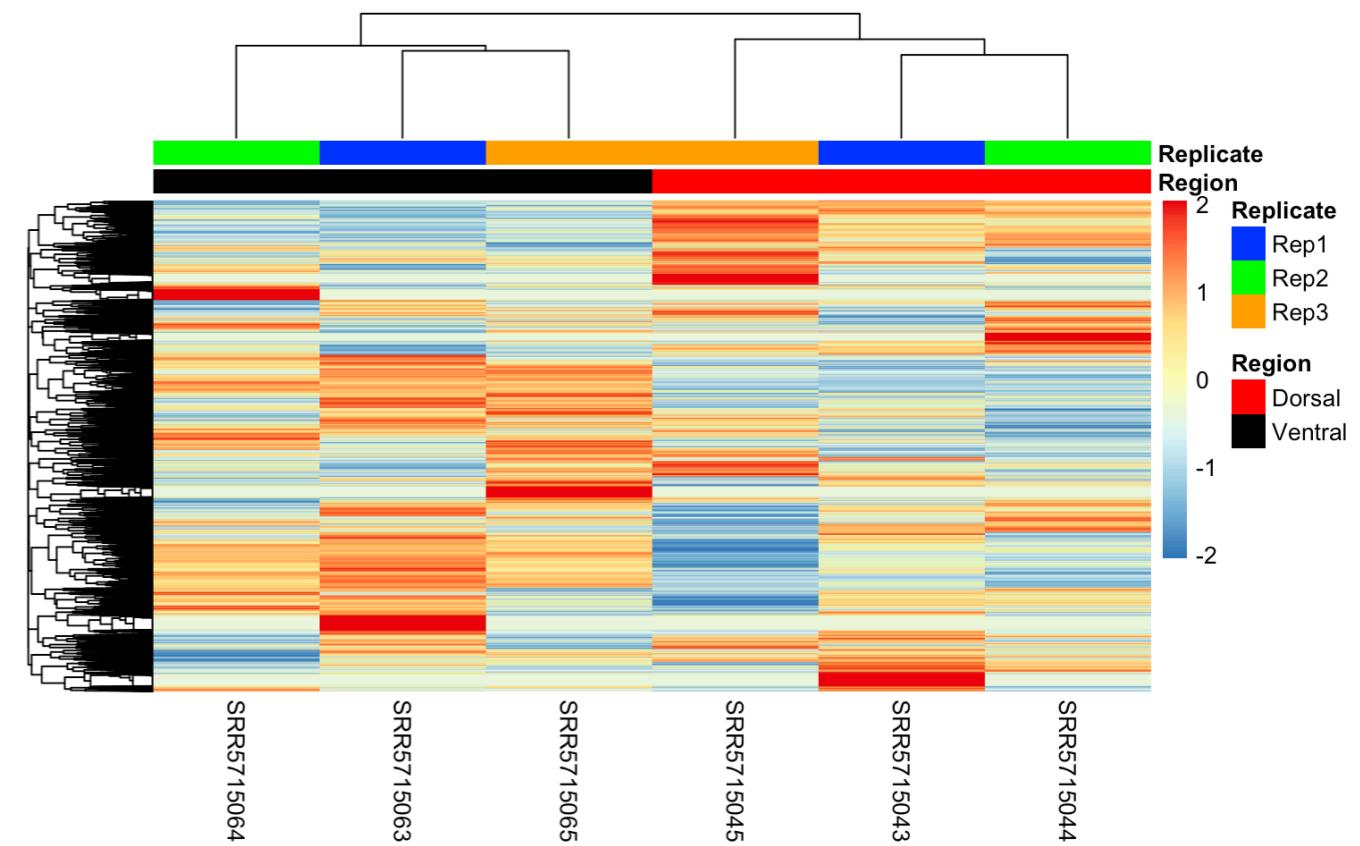
Replicate

- Rep1
- ▲ Rep2
- Rep3

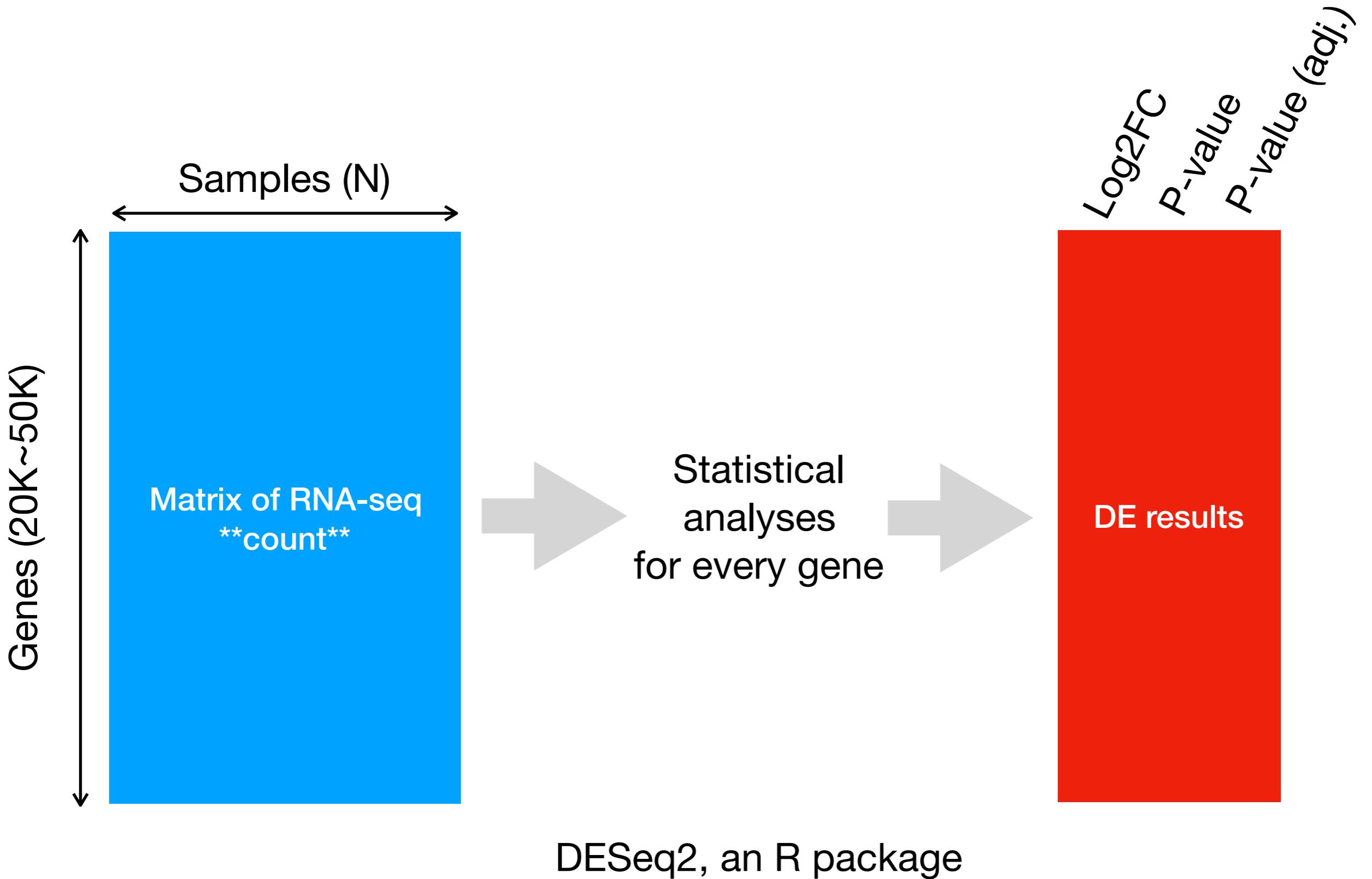
Region

- Dorsal
- Ventral

Heatmap



Differential Expression (DE)

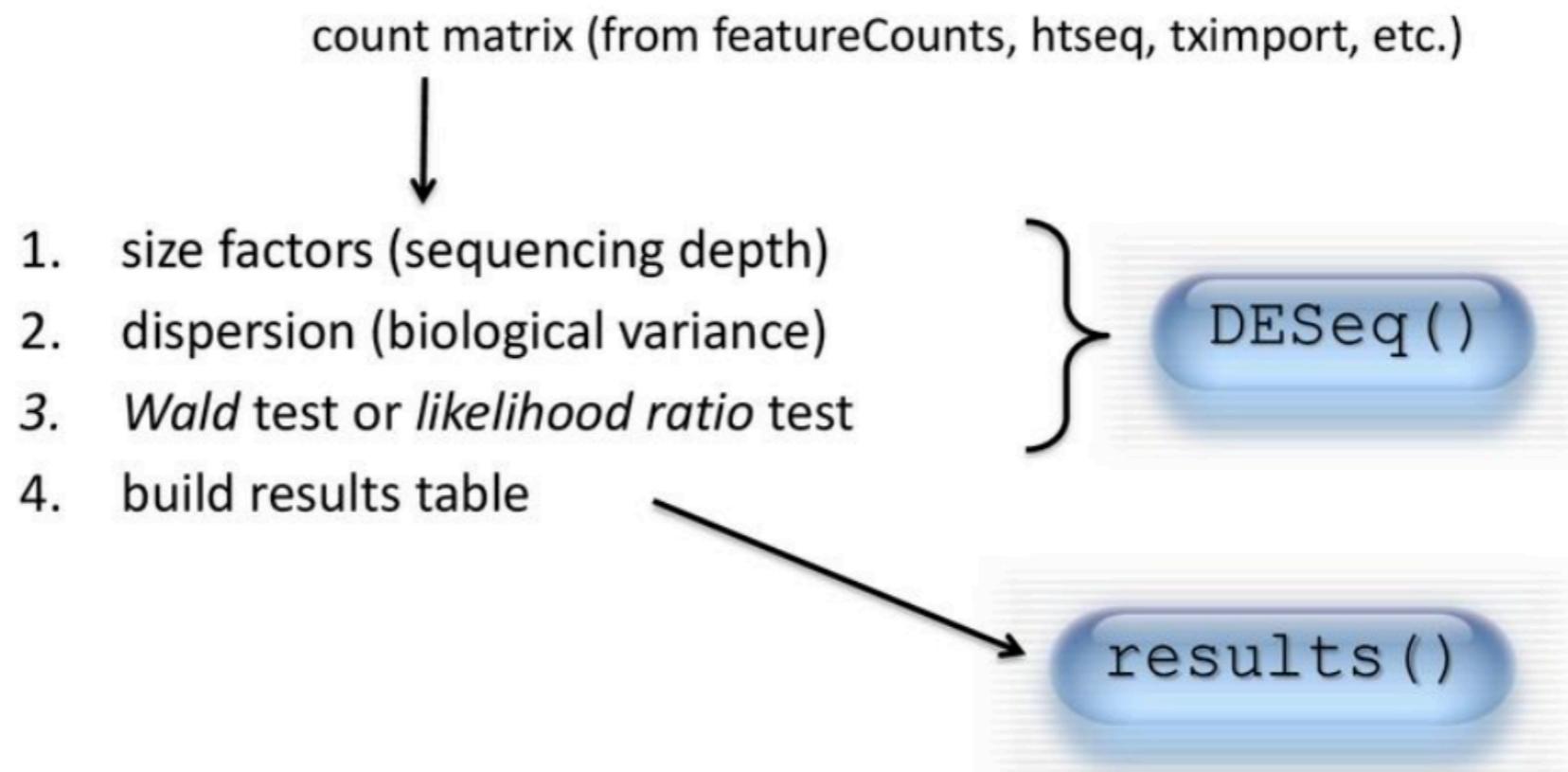


DE R packages

DE tools	Normalization	Count model	Statistical Test (Null: LFC==0)
EdgeR	TMM	Negative Binomial	Exact test or LRT
DESeq2	DESeq's median of ratio	Negative Binomial	Wald test or LRT
Limma-Voom	TMM	Log Normal	Moderated t-test

DESeq2

```
dds <- DESeqDataSetFromMatrix(countM, colData=sampleSheet,  
design= ~ Region)  
dds <- DESeq(dds)  
Res <- results(dds, contrast=c("Region", "Dorsal", "Ventral"))
```



Design formula and contrast

```
dds <- DESeqDataSetFromMatrix(countM, colData=sampleSheet,  
design= ~ Region)  
dds <- DESeq(dds)  
Res <- results(dds, contrast=c("Region", "Dorsal", "Ventral"))
```

sampleSheet

	Sid	Region	Replicate
1	SRR5715043	Dorsal	Rep1
2	SRR5715044	Dorsal	Rep2
3	SRR5715045	Dorsal	Rep3
4	SRR5715063	Ventral	Rep1
5	SRR5715064	Ventral	Rep2
6	SRR5715065	Ventral	Rep3
7			

- Simple comparison

design = ~ region

- Controlling covariates

design = ~ replicate + region

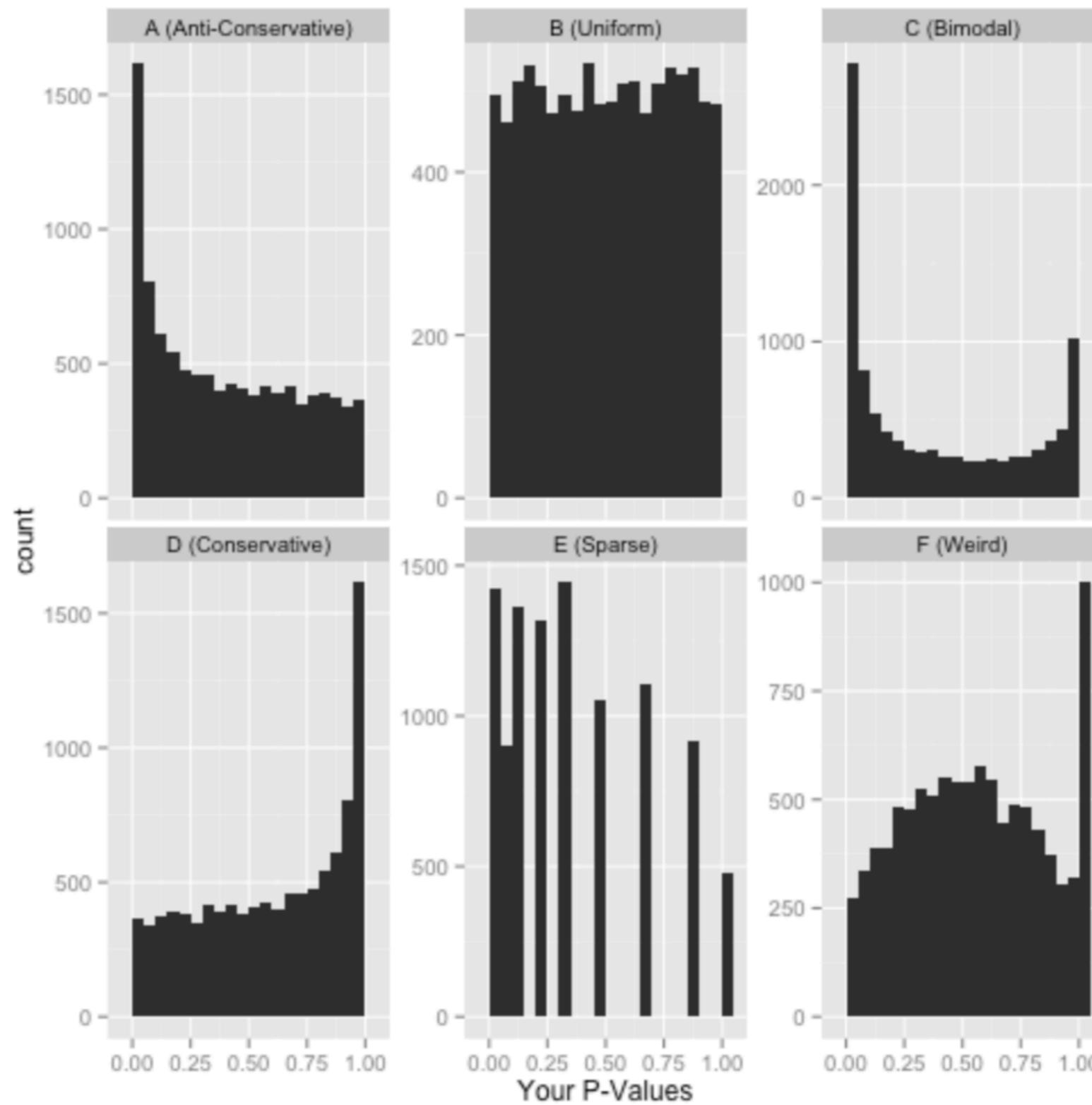
design = ~ replicate + sex + region

- Investigating interaction effects

design = ~ replicate + sex + region + sex : region

** Check the DESeq2 R package help page **

DE QC: p-value distribution

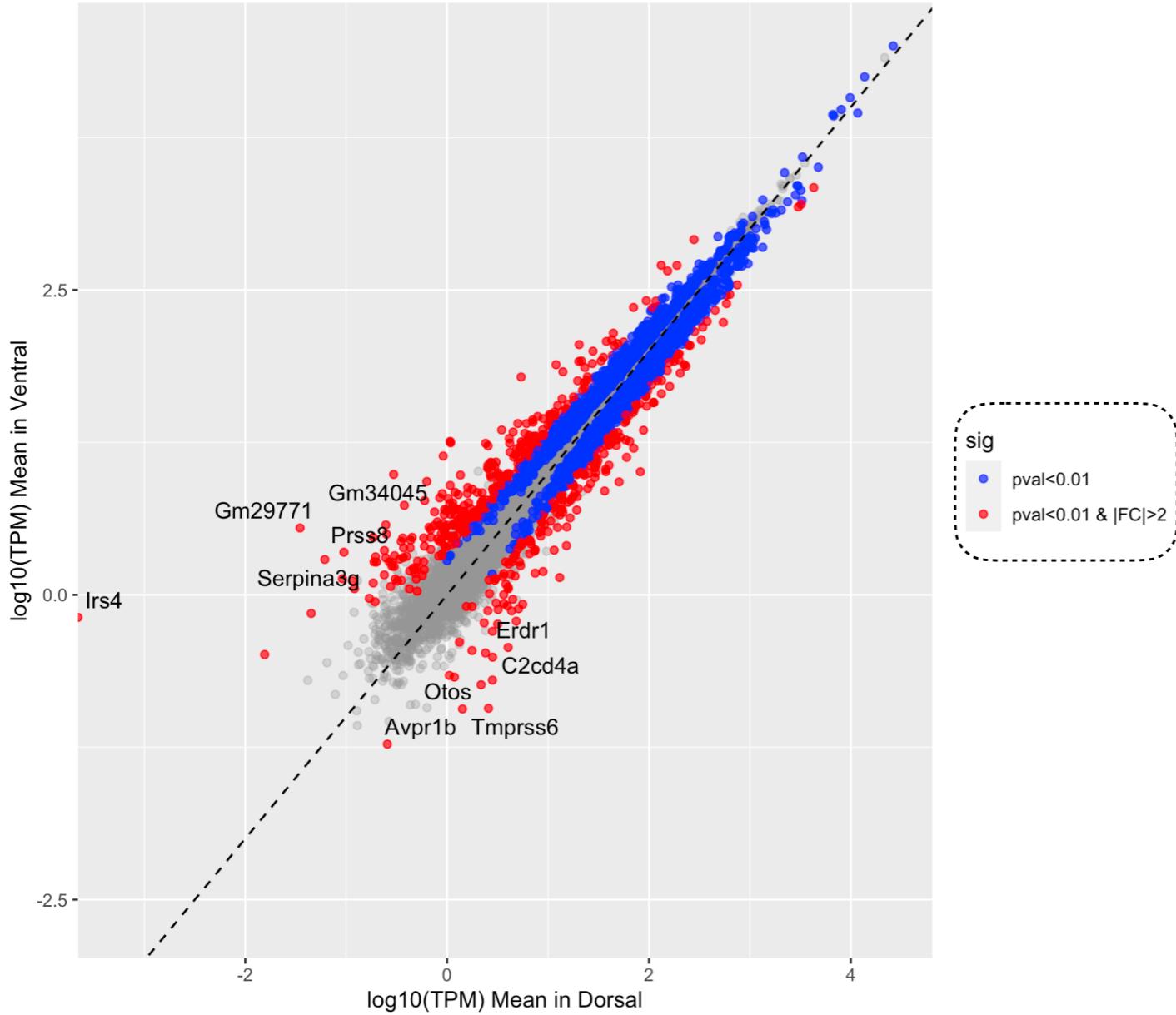


DE Results

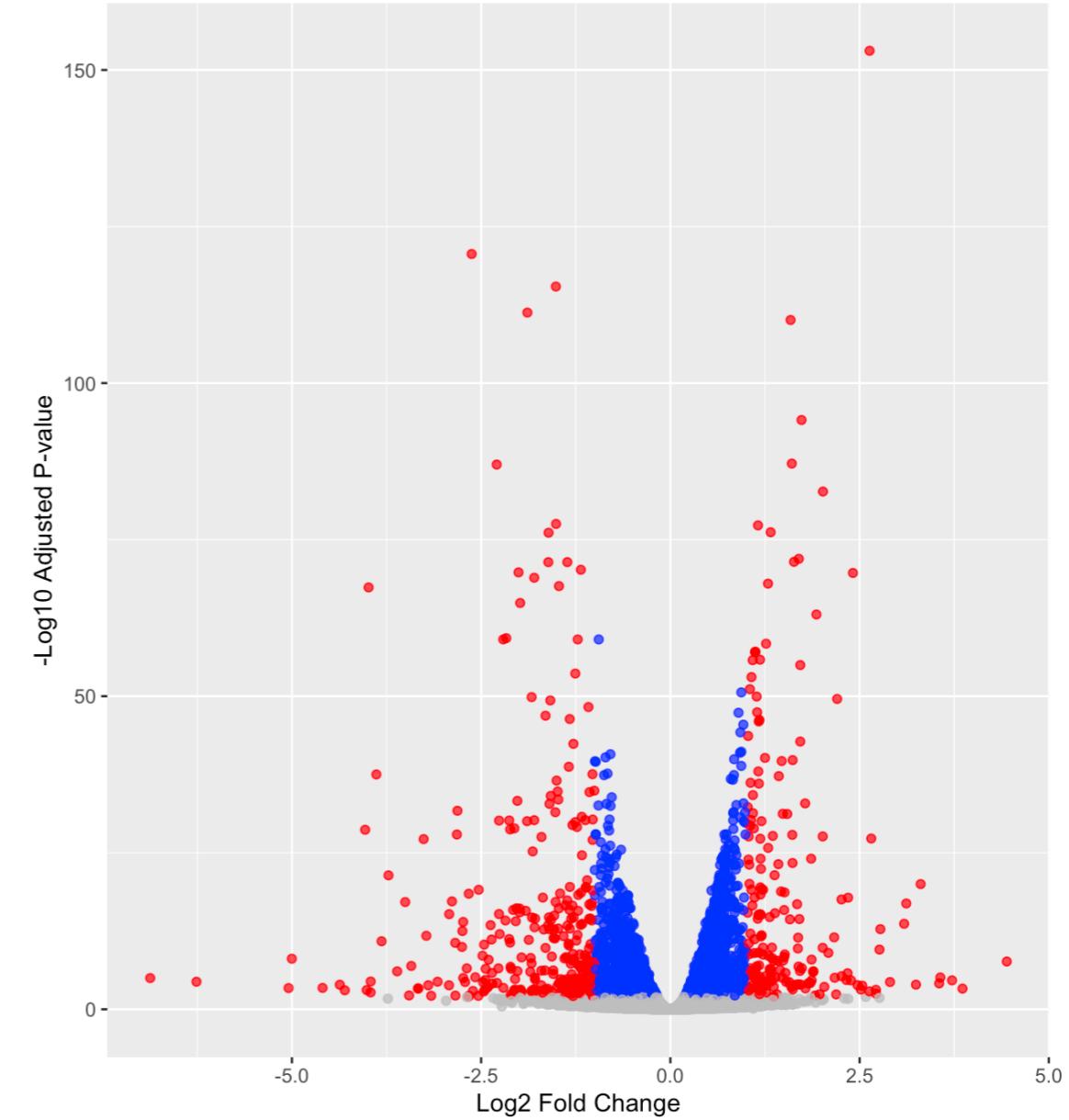
```
> res
log2 fold change (MLE): Region Dorsal vs Ventral
Wald test p-value: Region Dorsal vs Ventral
DataFrame with 56691 rows and 6 columns
  baseMean log2FoldChange lfcSE      stat      pvalue      padj
  <numeric>    <numeric> <numeric>    <numeric>    <numeric>    <numeric>
ENSMUSG00000102693.2     0          NA       NA        NA        NA        NA
ENSMUSG0000064842.3     0          NA       NA        NA        NA        NA
ENSMUSG0000051951.6 934.374636735581 -0.3552436281259 0.103883171365474 -3.41964558317256 0.000627027707598258 0.0042158466553528
ENSMUSG00000102851.2     0          NA       NA        NA        NA        NA
ENSMUSG00000103377.2 0.965337795076986 -0.10445910151227 2.64821564471098 -0.0394450888925515 0.968535532124425        NA
...
...          ...
ENSMUSG0000064368.1 554.832874720361 -0.268513399642424 0.121424238061778 -2.21136573659874 0.0270105205370121 0.0973450770345984
ENSMUSG0000064369.1 4.70528683513818 -0.229770388610795 1.14361158725448 -0.200916457275861 0.840763899060845        NA
ENSMUSG0000064370.1 153824.906100841 -0.188788470434184 0.0613087510925515 -3.07930706579212 0.00207482709210156 0.0118964348163462
ENSMUSG0000064371.1 14.6385304120105 0.333703279931054 0.656922522781579 0.507979660246796 0.611467607501022 0.808762996851356
ENSMUSG0000064372.1 29.7519207584107 -0.239430038273013 0.465265290440962 -0.51460971448373 0.606825775339794 0.806035794669969
```

DE result plots

- Scatter plot



- Volcano plot

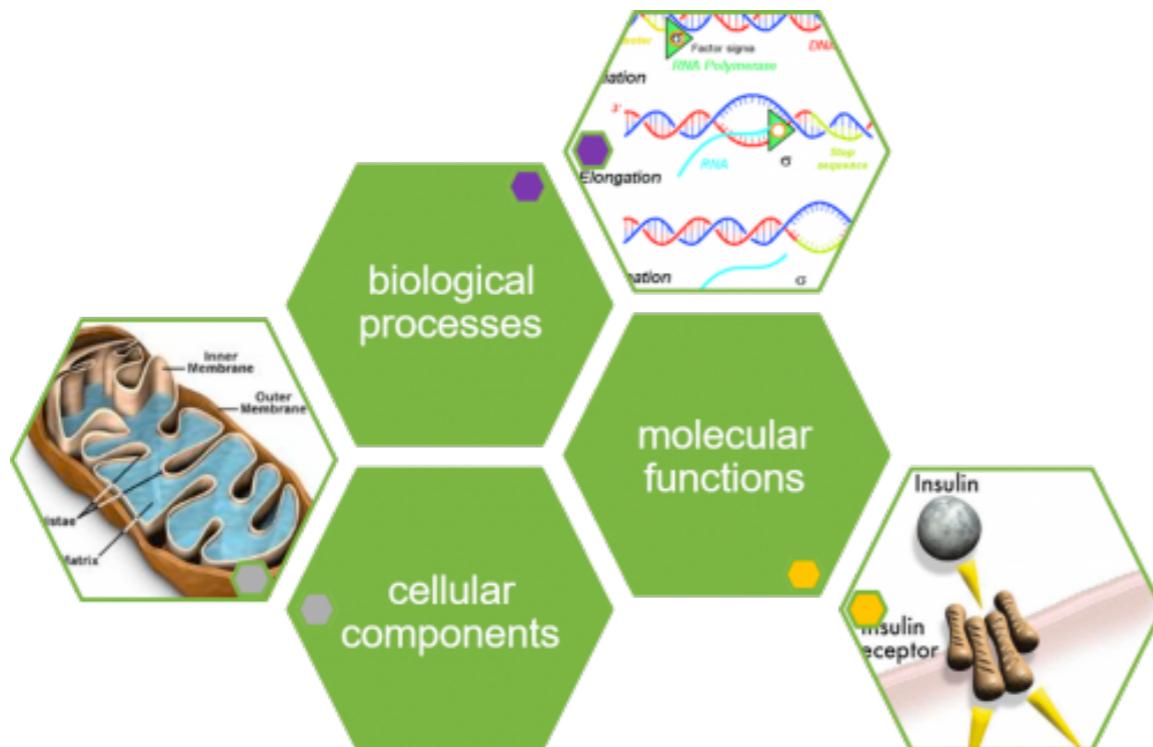


Gene Ontology database

GO:1903204

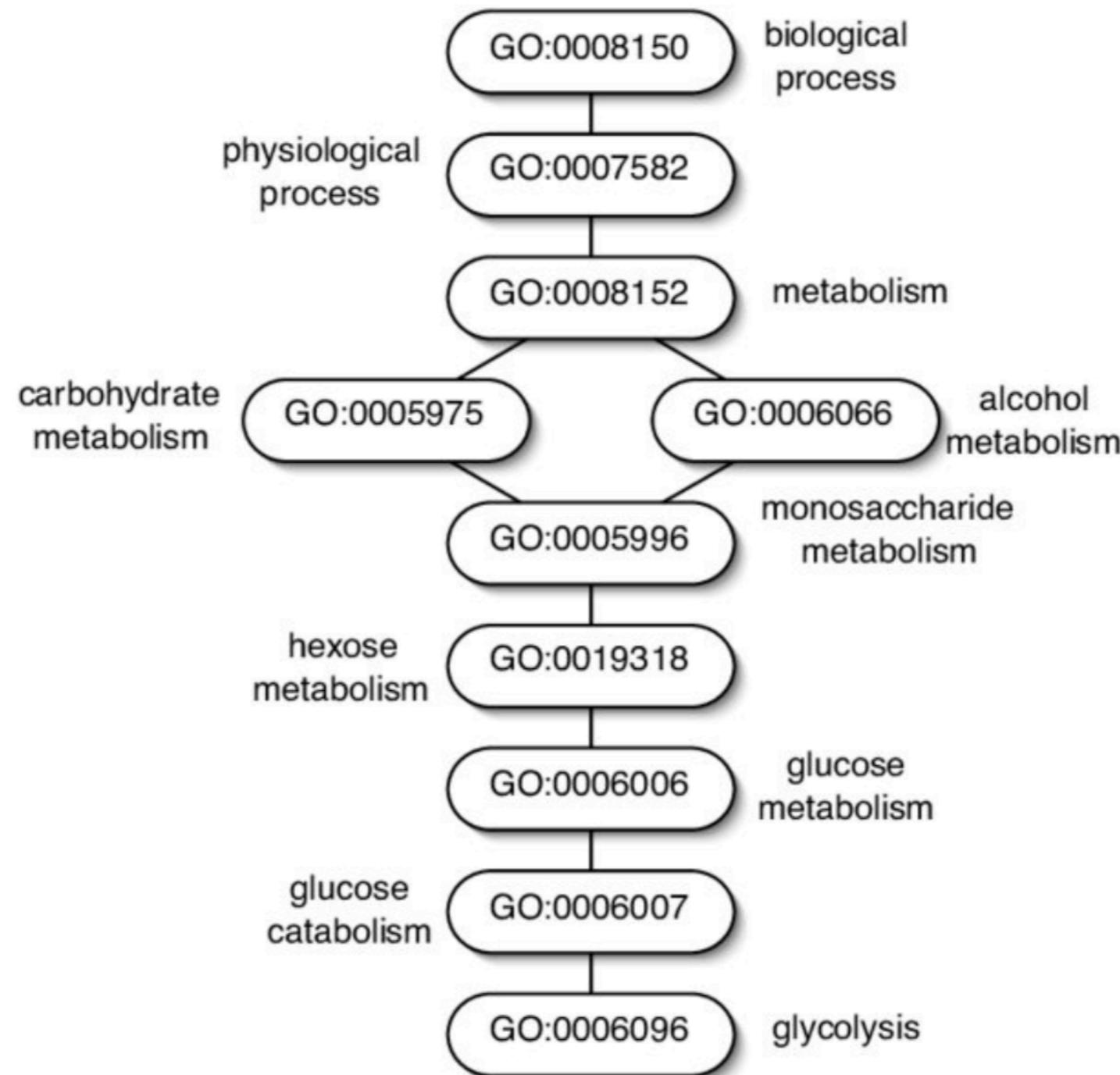
negative regulation of oxidative stress-induced neuron death

(#18) FBXO7, NONO, PARK7, RACK1, WNT1, PINK1, REST, IL10, NCOA7, MEAK7, HIF1A, SLC7A11, FZD1, GFER, CTNNB1, NR4A3, OXR1, ATF4.



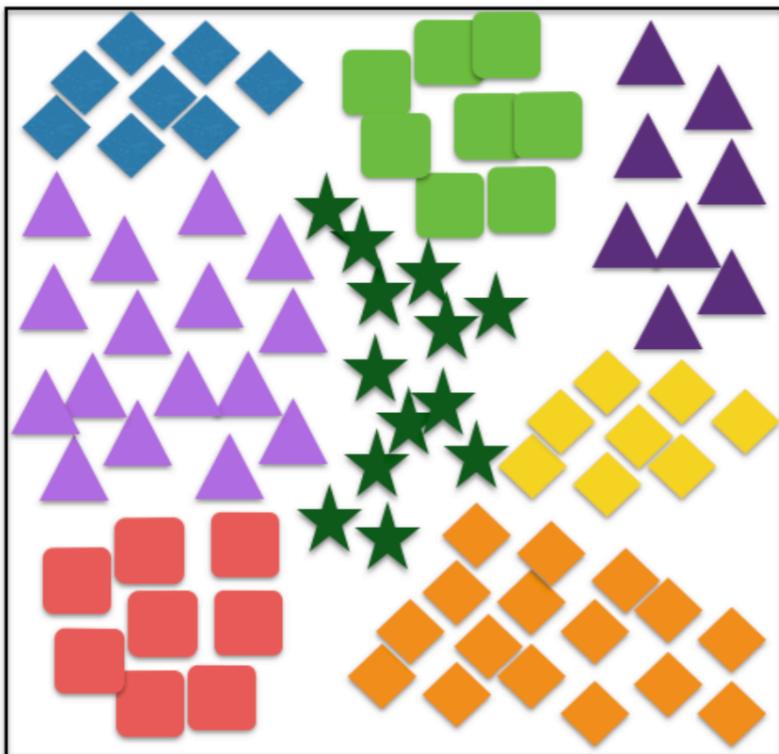
“cytochrome c” can be described by the **molecular function oxidoreductase activity**, the **biological process oxidative phosphorylation**, and the **cellular component mitochondrial matrix**.

GO: Hierarchical structure

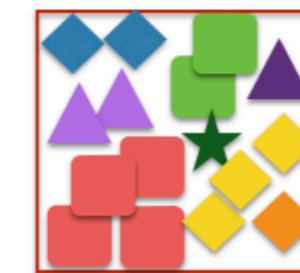


Enrichment analysis

All known genes in a species
(categorized into groups)



DE analysis



DEGs

Genes categories	Organism-specific Background	DE results	Over-represented?
Functional category 1	35/13000	25/1000	Likely
Functional category 2	56/13000	4/1000	Unlikely
Functional category 3	90/13000	8/1000	Unlikely
Functional category 4	15/13000	10/1000	Likely
...			

Hypergeometric test

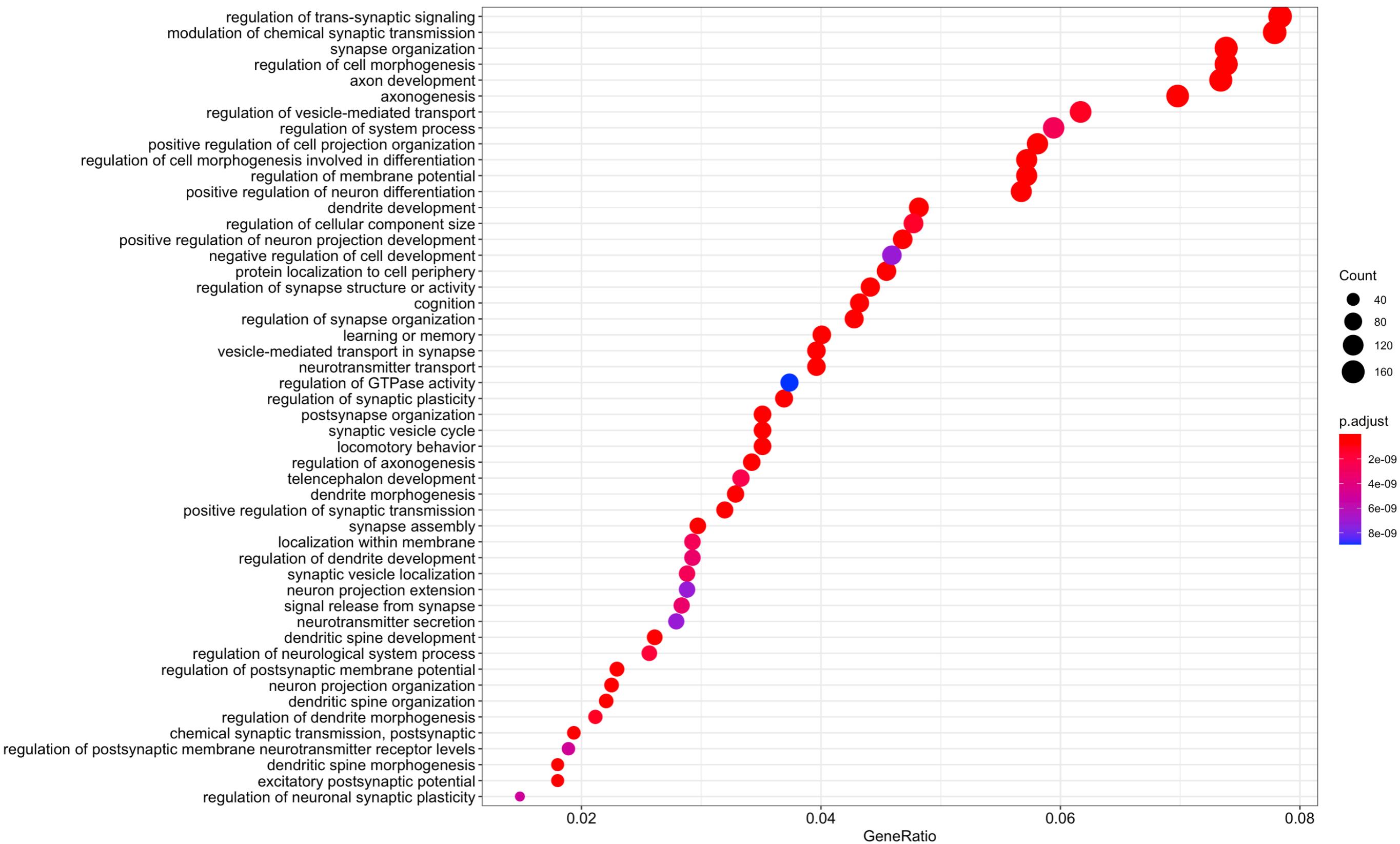
Genes categories	Organism-specific Background	DE results	Over-represented?
Functional category 1	35/13000	25/1000	Likely
Functional category 2	56/13000	4/1000	Unlikely
Functional category 3	90/13000	8/1000	Unlikely
Functional category 4	15/13000	10/1000	Likely
...			

Null distribution: hypergeometric distribution

$$\Pr(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

For category 1, N=13000, K=35, n=1000, k=0 to 35
For category 2, N=13000, K=56, n=1000, k=0 to 56

GO terms enriched in DE genes



KEGG pathway database

<https://www.genome.jp/kegg/pathway.html>

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

- 1. Metabolism**
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Organismal Systems**
- 6. Human Diseases**
- 7. Drug Development**

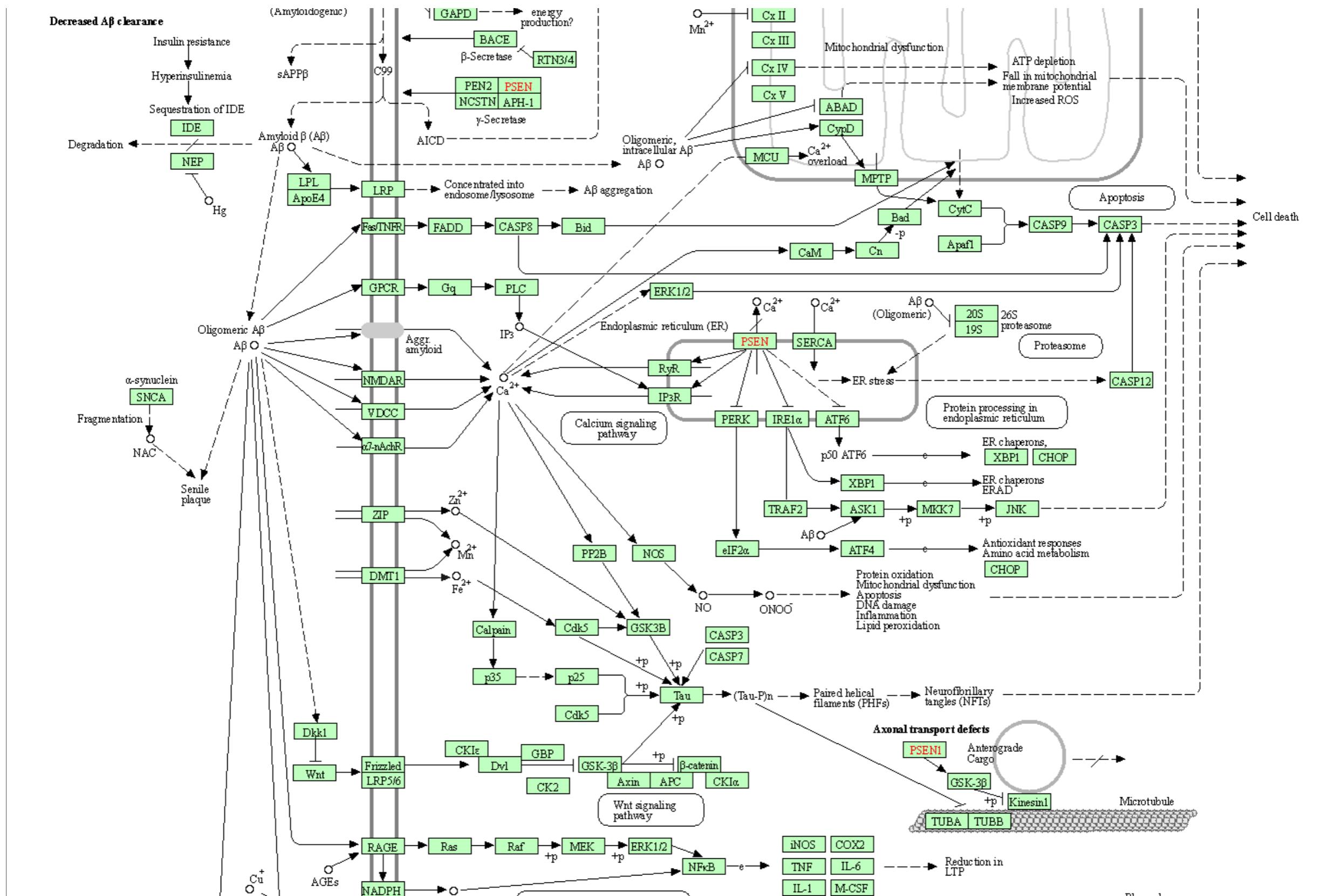
← → C 🔒 genome.jp/pathway/hsa05010



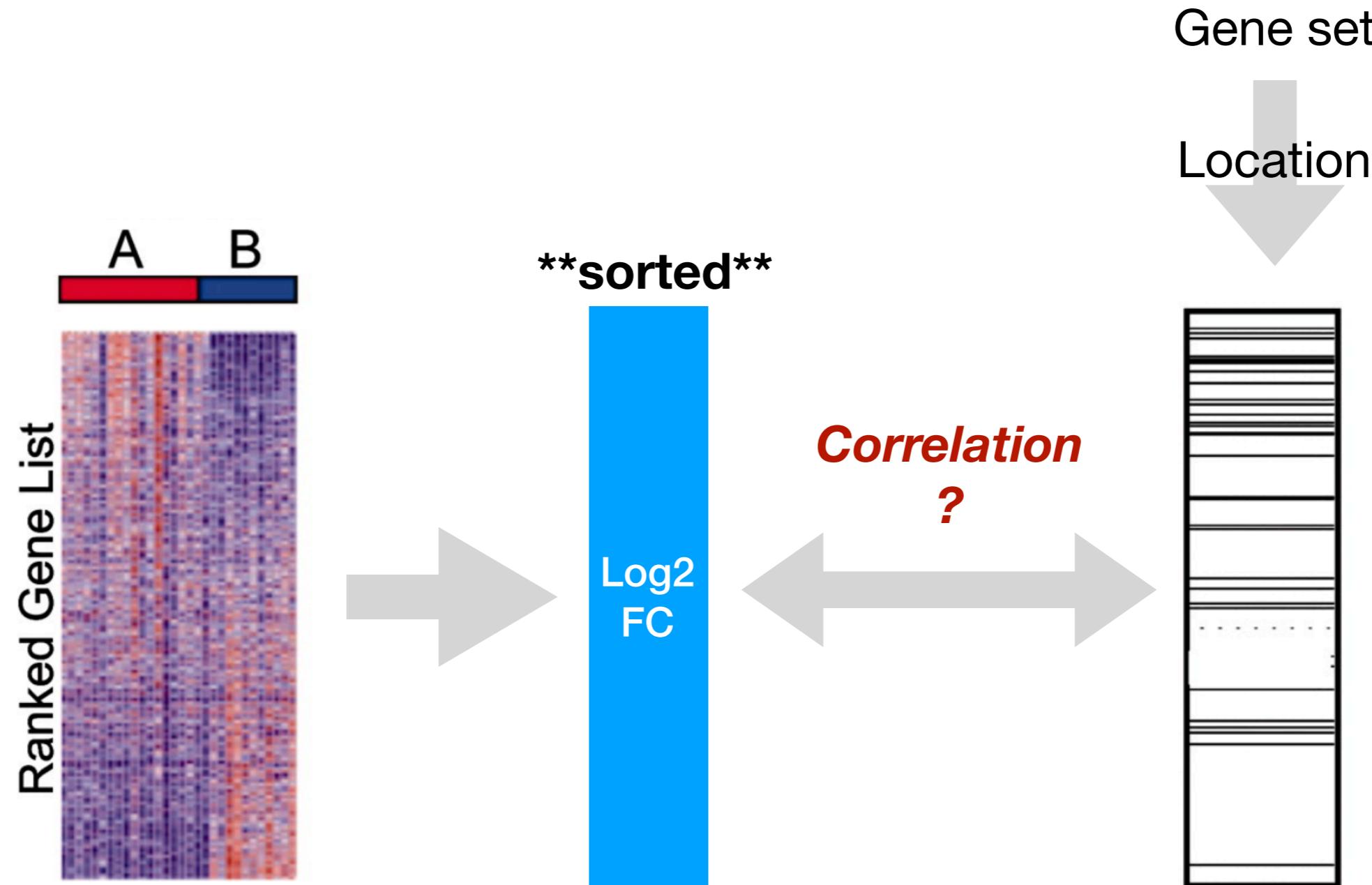
Alzheimer disease - Homo sapiens (human)

[[Pathway menu](#) | [Organism menu](#) | [Pathway entry](#) | [Download KGML](#) | [Hide description](#) | [Image file](#) | [Help](#)]

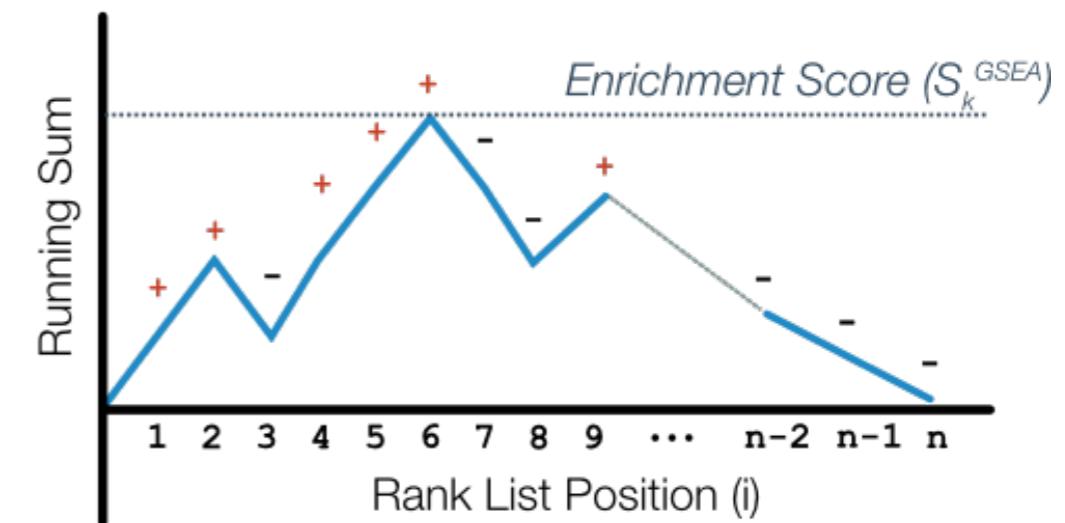
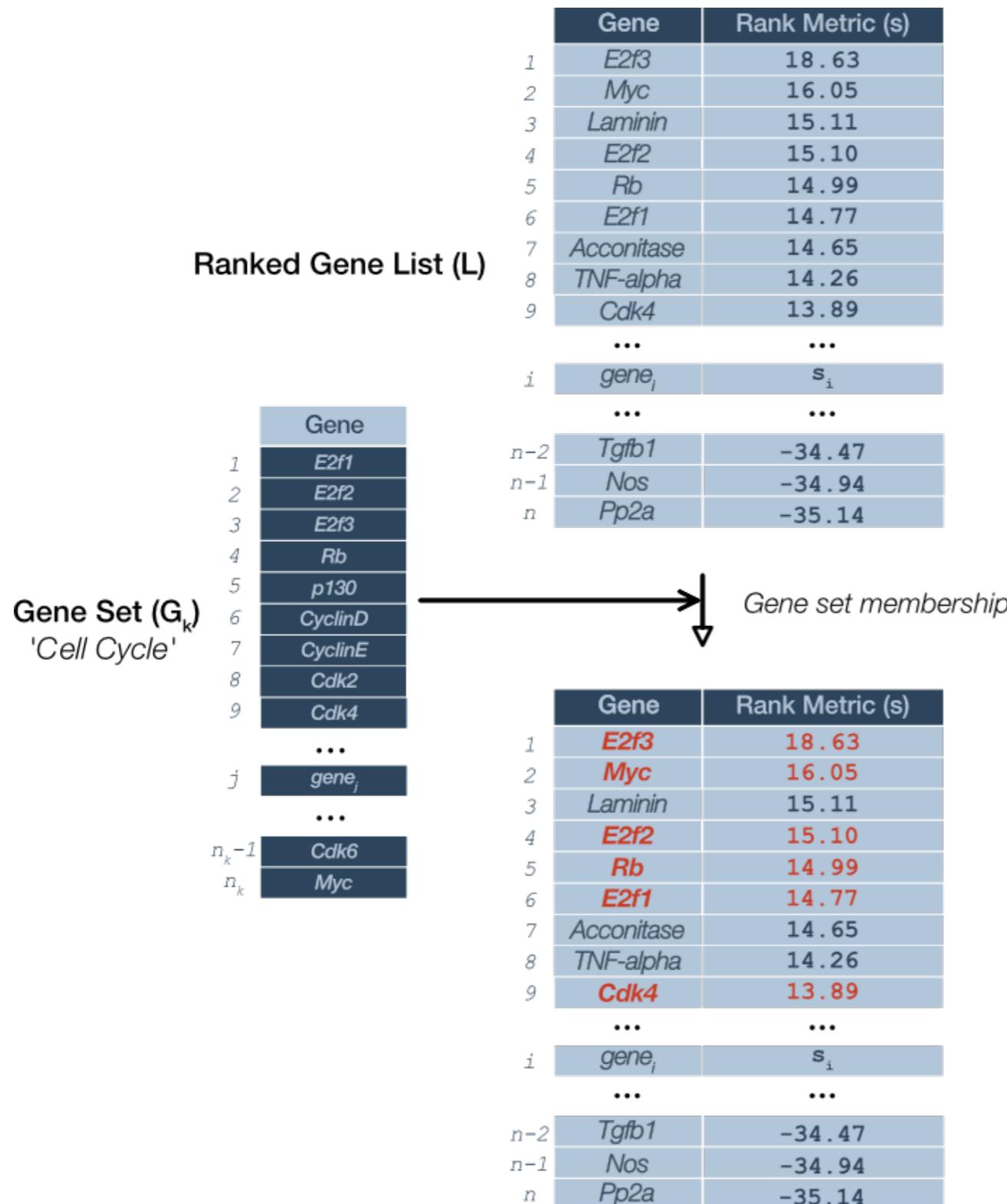
Alzheimer disease (AD) is a chronic disorder that slowly destroys neurons and causes serious cognitive disability. AD is associated with senile plaques and neurofibrillary tangles (NFTs). Amyloid-beta (Abeta), a major component of senile plaques, has various pathological effects on cell and organelle function. To date genetic studies have revealed four genes that may be linked to autosomal dominant or familial early onset AD (FAD). These four genes include: amyloid precursor protein (APP), presenilin 1 (PS1), presenilin 2 (PS2) and apolipoprotein E (ApoE). All mutations associated with APP and PS proteins can lead to an increase in the production of Abeta peptides, specifically the more amyloidogenic form, Abeta42. It was proposed that Abeta form Ca²⁺ permeable pores and bind to and modulate multiple synaptic proteins, including NMDAR, mGluR5 and VGCC, leading to the overfilling of neurons with calcium ions. Consequently, cellular Ca²⁺ disruptions will lead to neuronal apoptosis, autophagy deficits, mitochondrial abnormality, defective neurotransmission, impaired synaptic plasticity and neurodegeneration in AD. FAD-linked PS1 mutation downregulates the unfolded protein response and leads to vulnerability to ER stress.



Gene Set Enrichment Analysis (GSEA)



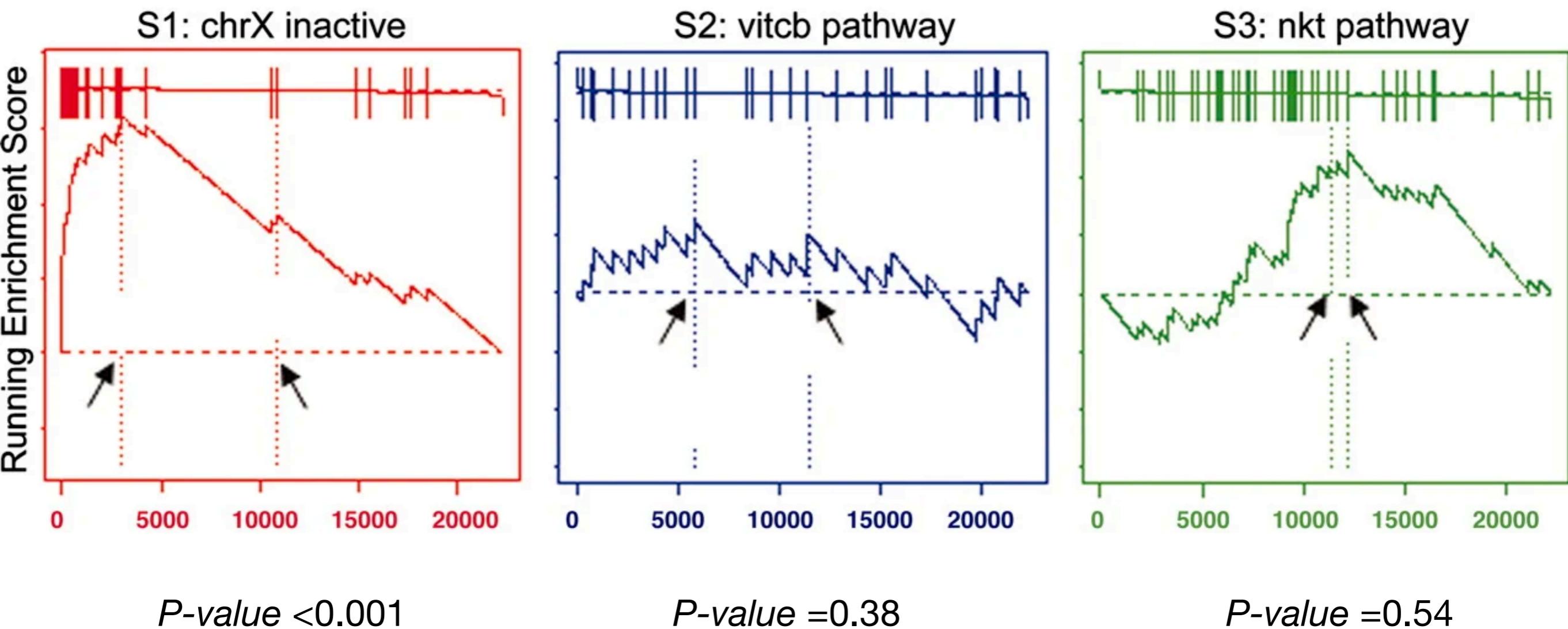
GSEA enrichment score



P-value computed by permutation test

GSEA enrichment examples

"the list of genes in the male vs female lymphoblastoid cell lines



Readings

- **Teaching materials at the Harvard Chan Bioinformatics Core (recommended)**
https://github.com/hbctraining/DGE_workshop_salmon_online/blob/master/schedule/links-to-lessons.md
- **UC-Davis Workshop material**
https://ucdavis-bioinformatics-training.github.io/2019_March_UCSF_mRNAseq_Workshop/data_reduction/alignment.html
- **Galaxy training**
<https://training.galaxyproject.org/archive/2019-02-07/topics/transcriptomics/tutorials/ref-based/tutorial.html>
- **Read each tool's manual (at least one time, strongly recommended) !!**