Transcriptomics/ RNA-sequencing

for Medical Life Science

Neha Mishra

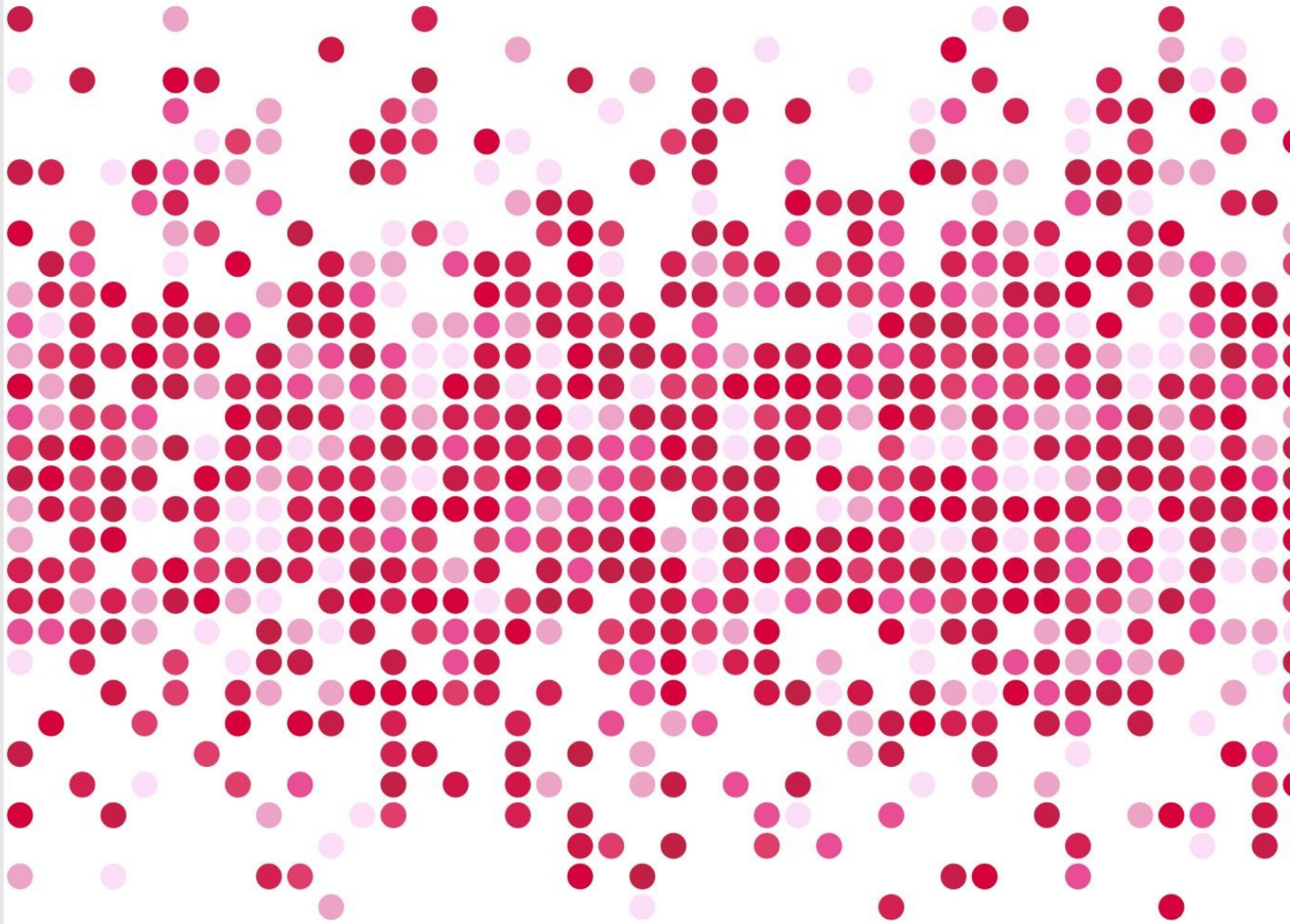
n.mishra@ikmb.uni-kiel.de

Florian Uellendahl-Werth

f.uellendahl-werth@ikmb.uni-kiel.de

Institute of Clinical Molecular Biology

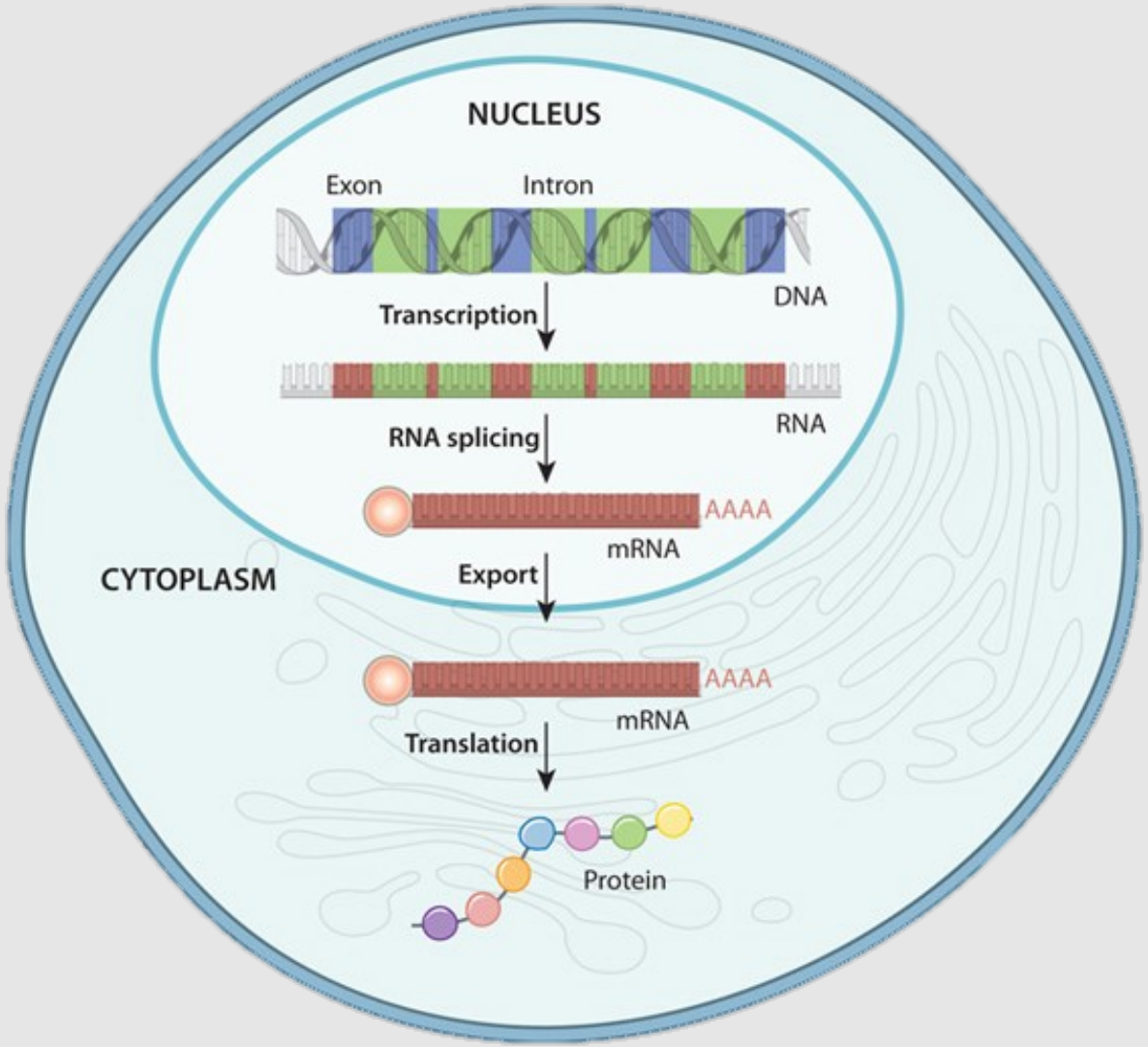
12.12.2023



Learning Objectives

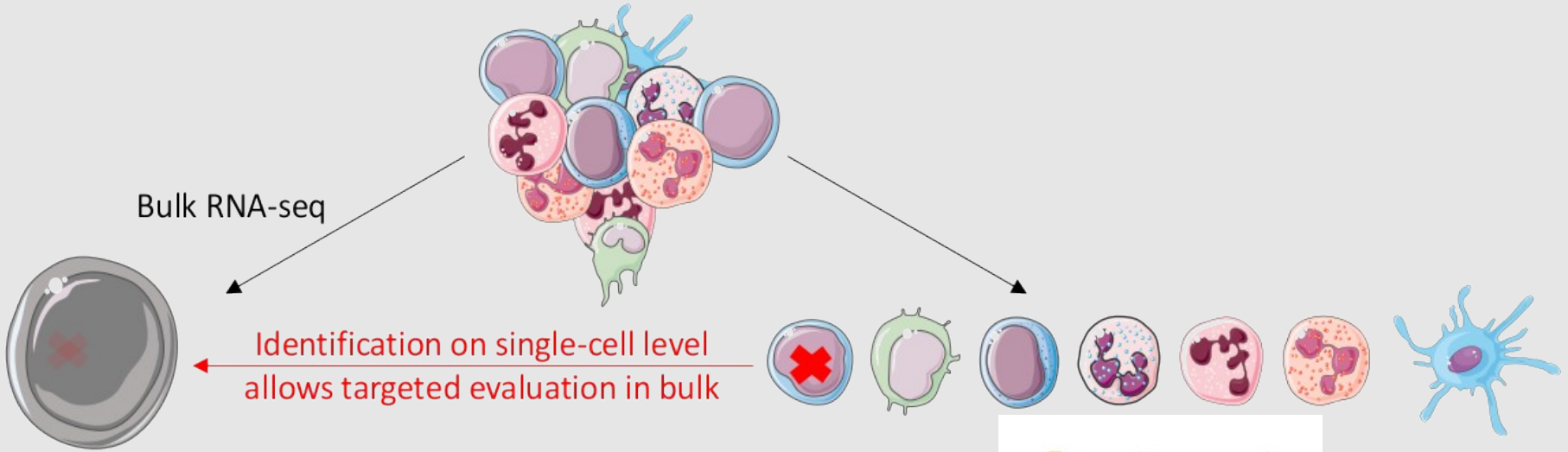
- Basics of RNA-sequencing
- Experimental design
- Basics of pre-processing of RNA-seq output (tools and pipelines)
- Differential expression analysis (tools and scripting)

Gene Expression



Courtesy 2010 Nature Education

RNA-sequencing



Modified from Servier Medical Art by Jonas Schulte-Schrepping
http://365-smoothie-rezepte.de/obst-beeren-und-gemuese-sorgfaeltig-auswaehlen/

Why do RNA-sequencing?

- Functional studies
  - Gene expression changes according to the experimental conditions
    - Eg. Treated vs. Untreated cells
    - Eg. Wild type vs. Knock out animals

Why do RNA-sequencing?

- Functional studies
  - Gene expression changes according to the experimental conditions
    - Eg. Treated vs. Untreated cells
    - Eg. Wild type vs. Knock out animals

- Whole genome coverage* (compared to qPCR and microarray)

- Prior information about the gene structure or sequence not required
  - Can also be used for gene annotation

Why do RNA-sequencing?

- Functional studies
  - Gene expression changes according to the experimental conditions
    - Eg. Treated vs. Untreated cells
    - Eg. Wild type vs. Knock out animals

- Whole genome coverage* (compared to qPCR and microarray)

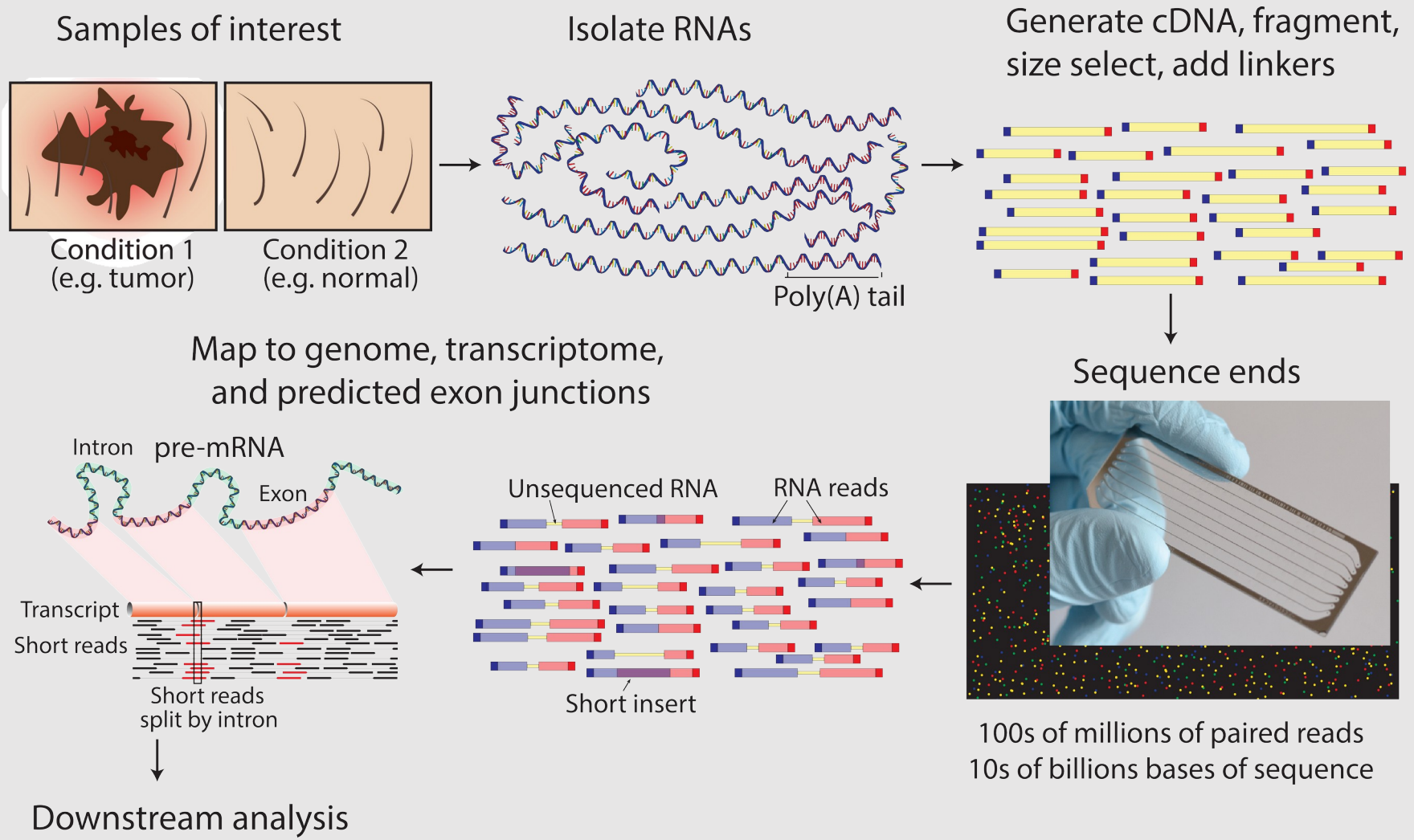
- Prior information about the gene structure or sequence not required
  - Can also be used for gene annotation

- Cheaper (than proteome profiling)

Why do RNA-sequencing?

- Functional studies
  - Gene expression changes according to the experimental conditions
    - Eg. Treated vs. Untreated cells
    - Eg. Wild type vs. Knock out animals
- Whole genome coverage* (compared to qPCR and microarray)
- Prior information about the gene structure or sequence not required
  - Can also be used for gene annotation
- Cheaper (than proteome profiling)
- RNA level features can be detected
  - Alternative isoforms, fusion transcripts

RNA-sequencing



Griffith et al Plos Computational Biology 2015

Considerations for RNA-sequencing

- RNA extraction kit
- RNA quality
- Library type

Considerations for RNA-sequencing

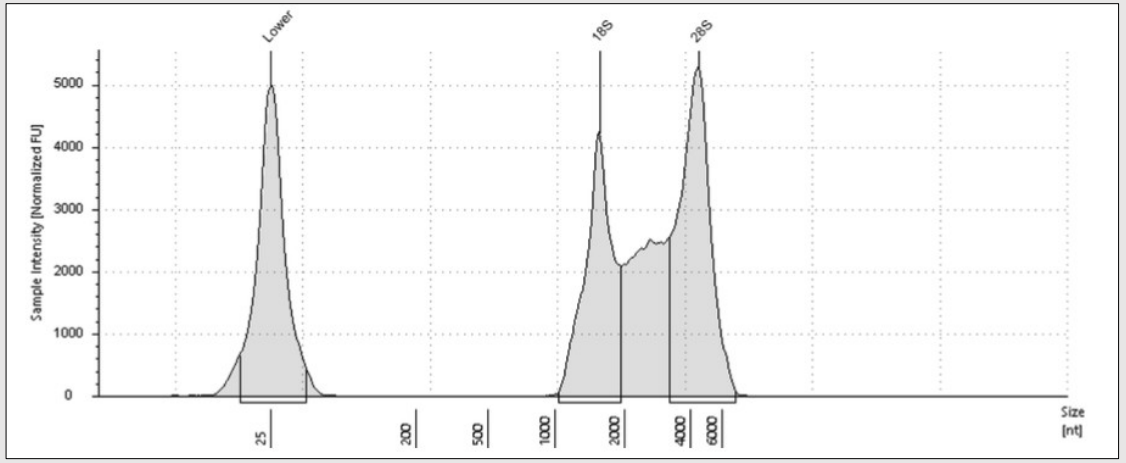
- RNA extraction kit
- RNA quality
- Library type

Choosing the right extraction kit

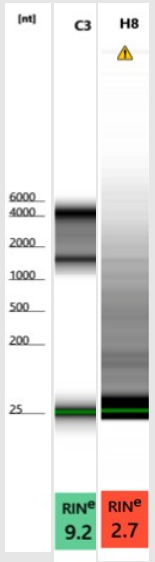
- Tissue/source
- Preservative

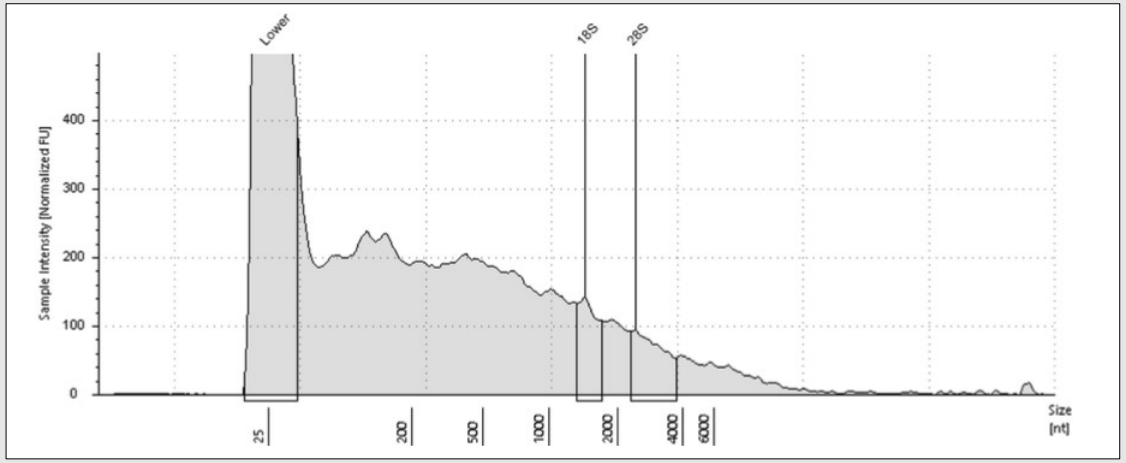
Considerations for RNA-sequencing

- RNA extraction kit
- RNA quality
- Library type



RIN = 9.2





RIN = 2.7

Considerations for RNA-sequencing

- RNA extraction kit
- RNA quality
- Library type

Choosing the right library type

- totalRNA
- mRNA
- 3´ tags
- smallRNA

Experimental Design

Replicates

Experimental Design

Replicates

Why do you need replicates?

- Reproducibility : The results of the experiments should be representative of the population and not the specific conditions you performed the experiment in.

- Statistical Power: The statistical tests will not make accurate inferences about the gene expression differences if the number of replicates is too low.

Experimental Design

Replicates

Why do you need replicates?

- Reproducibility : The results of the experiments should be representative of the population and not the specific conditions you performed the experiment in.
- Statistical Power: The statistical tests will not make accurate inferences about the gene expression differences if the number of replicates is too low.

How many replicates do you need?

Capture the breadth of variability and identify sources of noise.

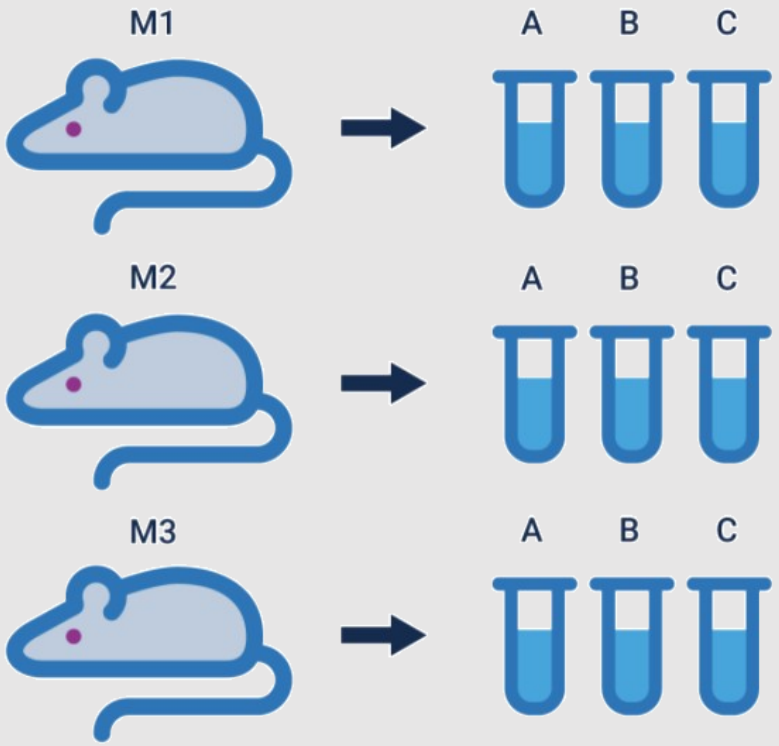
- Identify outlier samples
- *Remove outliers without losing too much information*

Experimental Design

Technical vs. Biological Replicates

Experimental Design

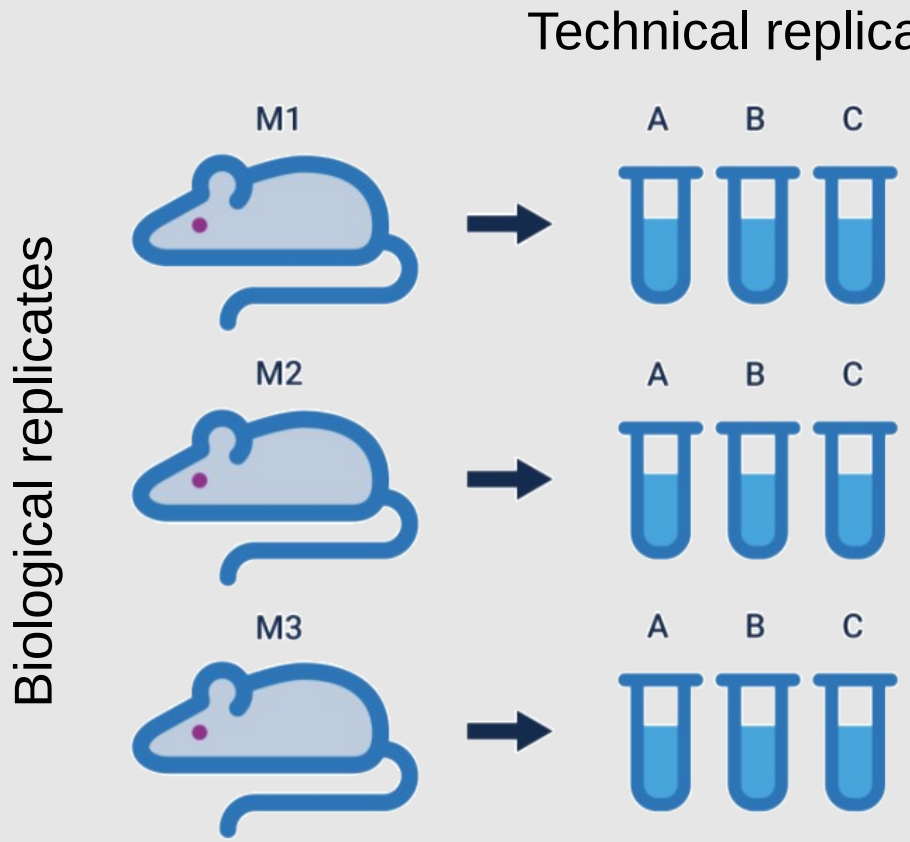
Technical vs. Biological Replicates



https://www.licor.com/bio/guide/westerns/replicates

Experimental Design

Technical vs. Biological Replicates



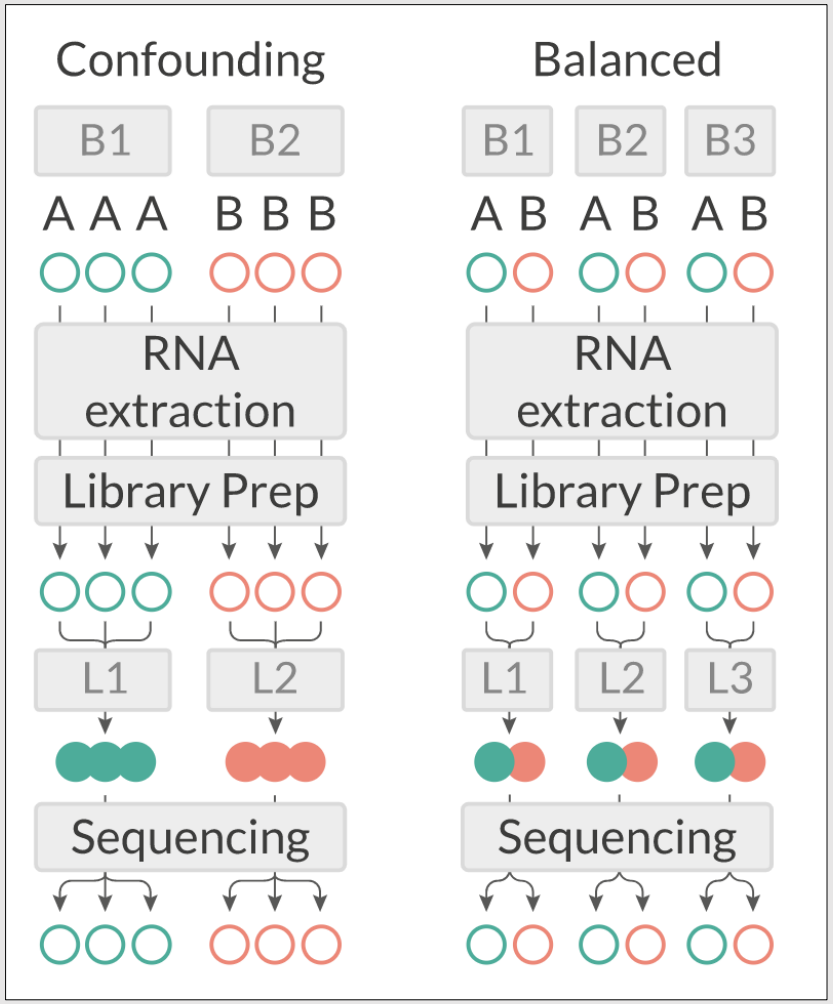
Technical replicates
- Repeated mesurements of the same sample
- Eg. Different library preparations

Biological replicates
- Parallel measurements of biologically distinct samples that capture random biological variation
- Eg. Different animals, cell cultures

https://www.licor.com/bio/guide/westerns/replicates

Experimental Design

Balanced experiment design



Raw Data (Reads)

After sequencing, we obtain short (50-150bp) reads in fastQ format

You can also download sequencing data from publically accessible databases such as the Sequence Read Archive (SRA)

Read ID + sequencing run info

Sequence bases

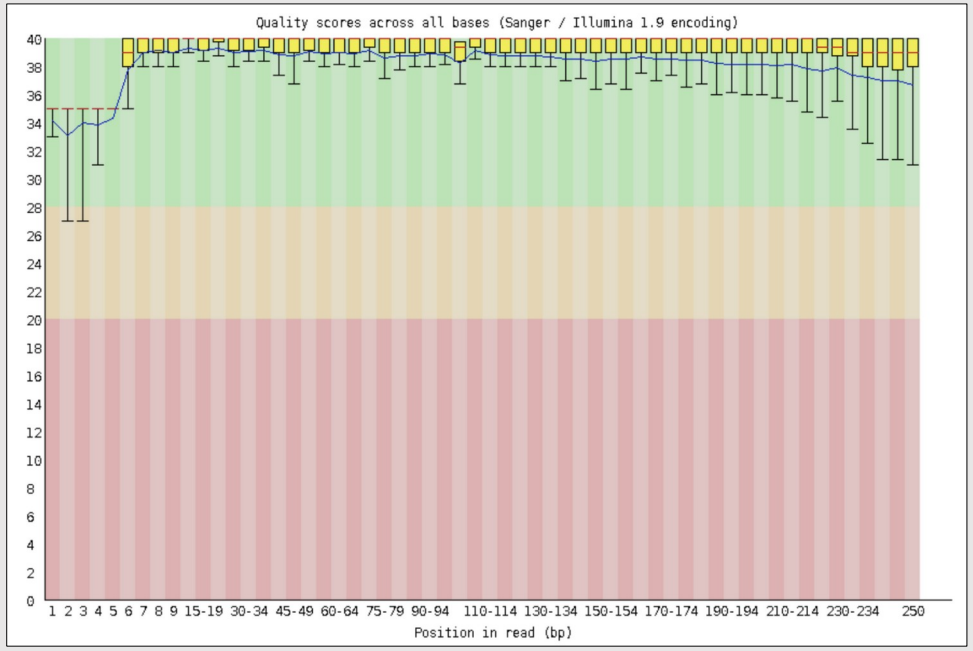
Quality scores for each base

```
$ zcat ERR459145.fastq.gz | head
@ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
+
@7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA'5?B@D
@ERR459145.2 DHKW5DQ1:219:D0PT7ACXX:2:1101:2652:2237/1
GCAGCATCGGCCTTTTGCTTCTCTTTGAAGGCAATGTCTTCAGGATCTAAG
+
@@;BDDEFGHHHHIIIGBHHEHCCHGCGIGGHIGHGIGIIGHIIAHIIIGI
@ERR459145.3 DHKW5DQ1:219:D0PT7ACXX:2:1101:3245:2163/1
TGCATCTGCATGATCTCAACCATGTCTAAATCCAAATTGTCAGCCTGCGCG
```

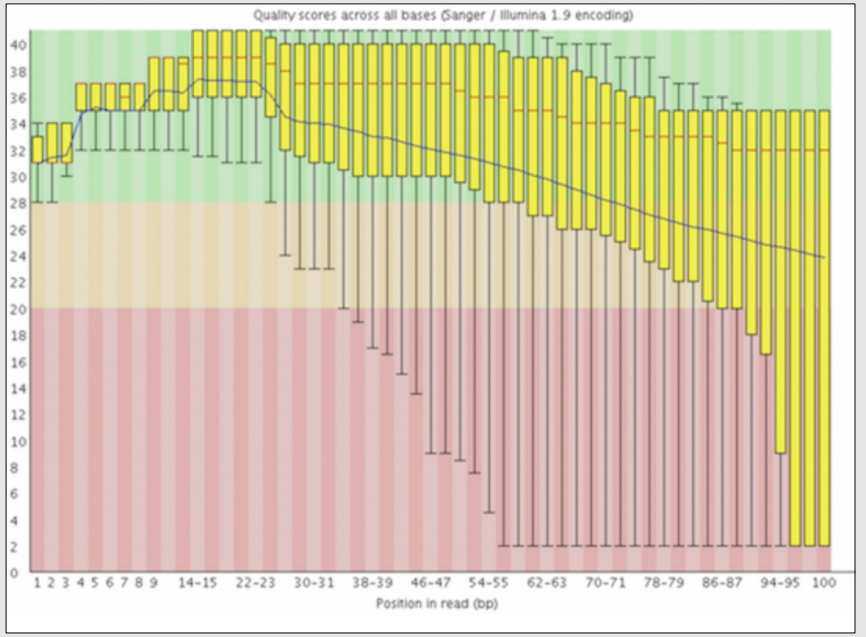
!Paired-end sequencing runs will have two FASTQ files – one for forward and one for reverse reads

Raw Data (Reads)

Quality control: Checking the quality of reads using FastQC

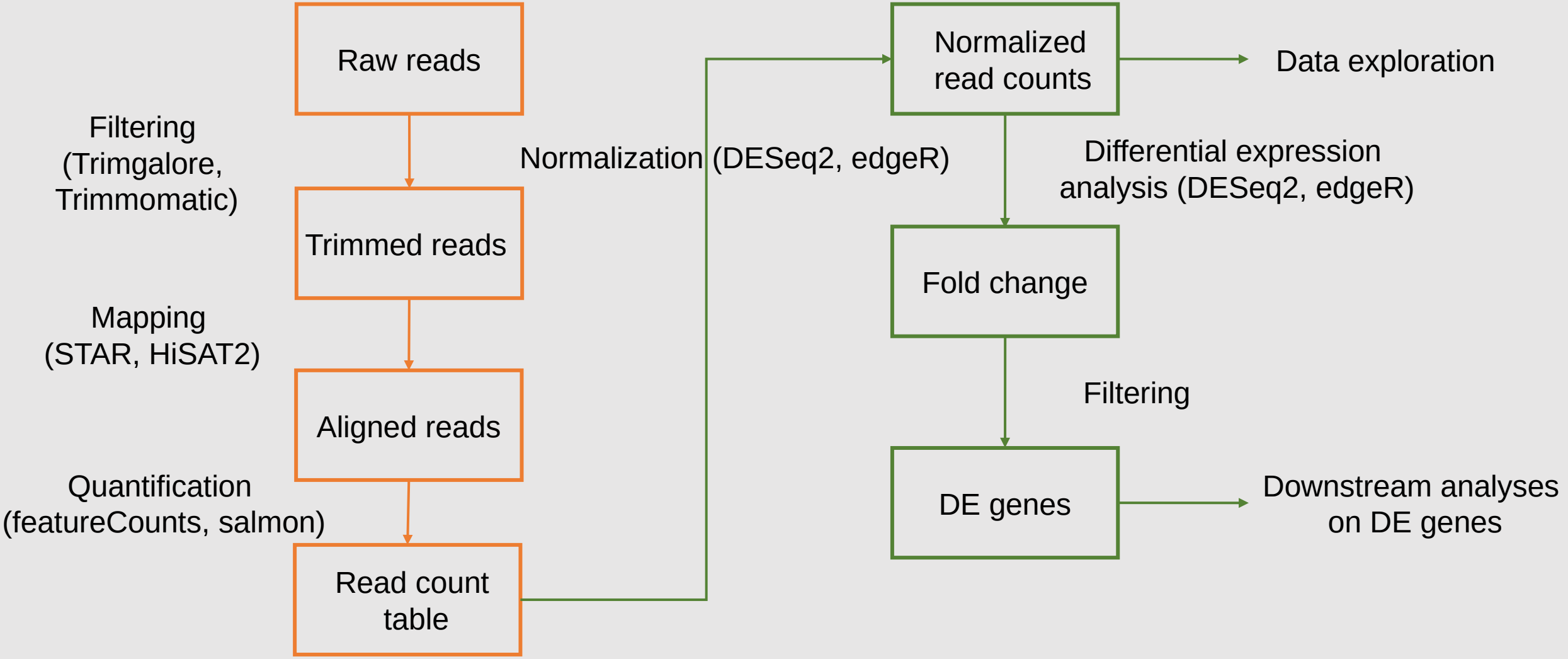


Good



Not Good

RNA-seq data analysis Pipeline



Read Mapping/Alignment

Reference ..GCTAGTTCGAATAGCTGCTGCAGTCGATGCATAAAGCCG..

Reads

GAATAGCTGCTGCAGTCGAT

ATAGCTCCTGCAGTCGATGCAT

TCGAATAGCTCCTG

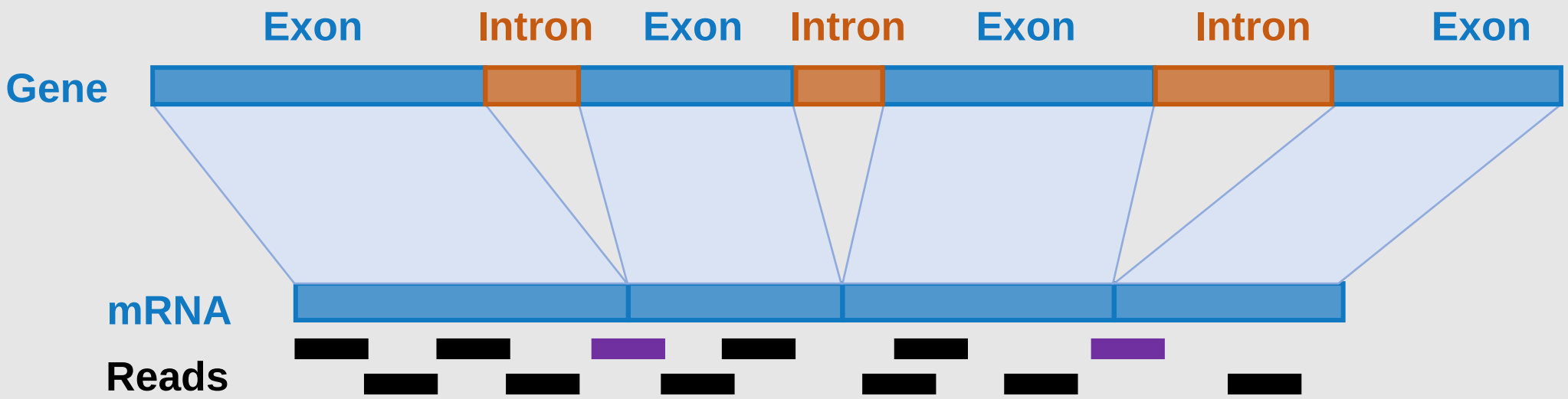
TGCAGTCGATGCATAAAGCCG

GCTGCTGCAGTCGATGCAT

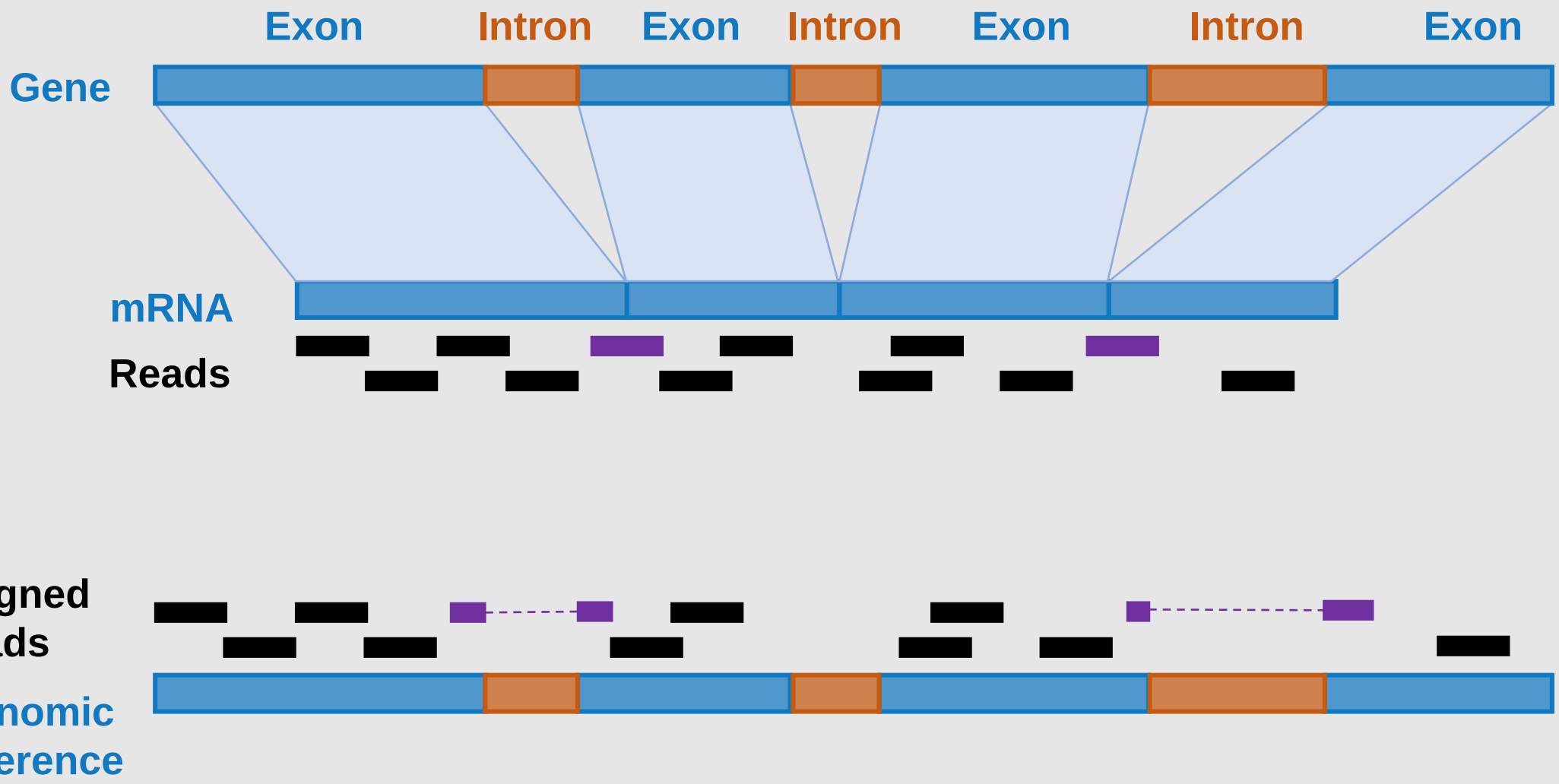
GCTGCTGCAGTCGATGCAT

Read Mapping/Alignment

Read Mapping/Alignment

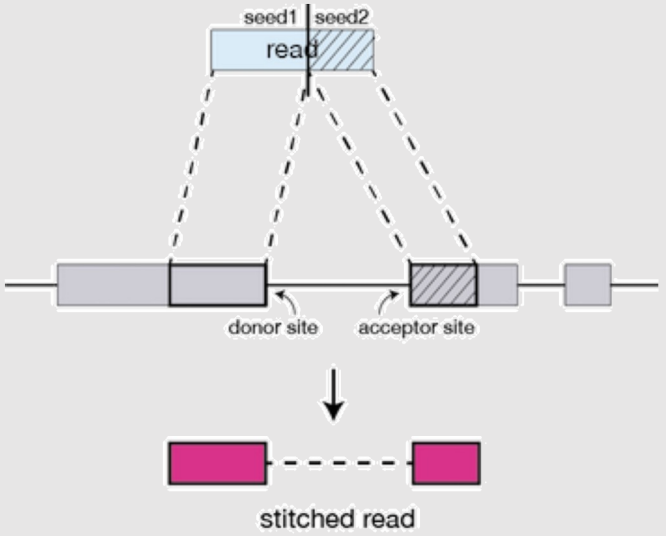


Read Mapping/Alignment



Read Mapping/Alignment

Alignment splice-aware algorithms (STAR, HISAT2)



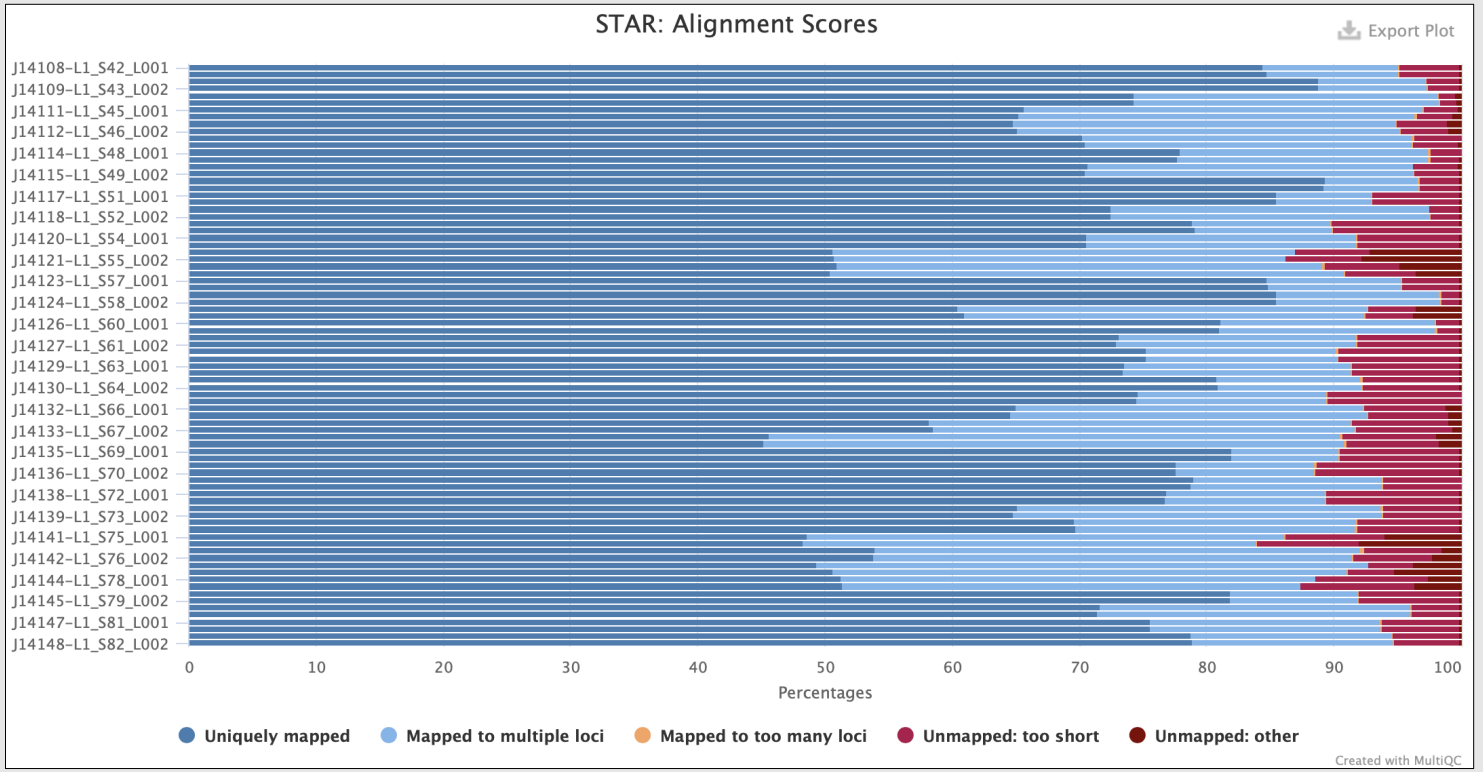
- Search for the longest sequence that exactly matches one or more locations on the reference genome.
- Then search again for only the unmapped portion of the read to find the next longest sequence that exactly matches the reference genome
- Stitch the read based on proximity and alignment score

Alignments are stored in SAM/BAM format

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

Read Mapping/Alignment

Alignment Scores



Read Mapping/Alignment

Visualisation using Integrated Genome Viewer (IGV)

Read Mapping/Alignment

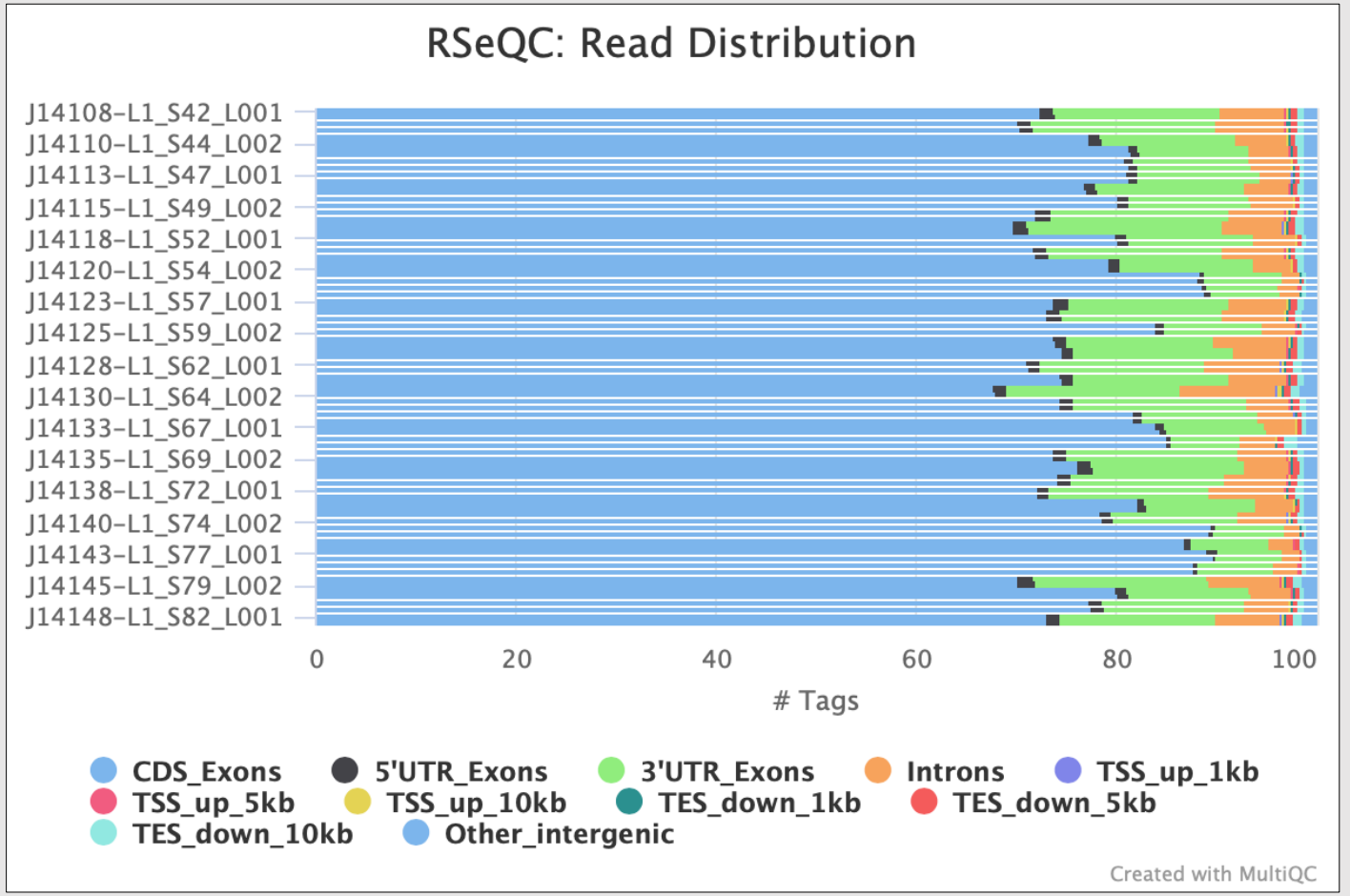
Quality control: Bias Identification

Potential biases

- High intron coverage: incomplete poly(A) enrichment or abundant presence of immature transcripts
- Intergenic reads: genomic DNA contamination (or abundant non-coding transcripts)
- 3’ bias: over-representation of 3’ portions of transcripts indicates RNA degradation

Read Mapping/Alignment

QC: Read distribution using RSeQC

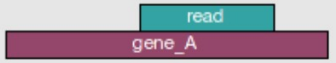
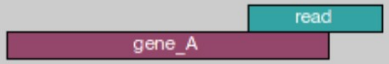
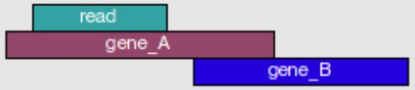
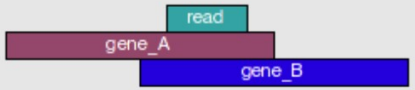


Read Quantification

When counting reads, make sure you know how the program handles the following

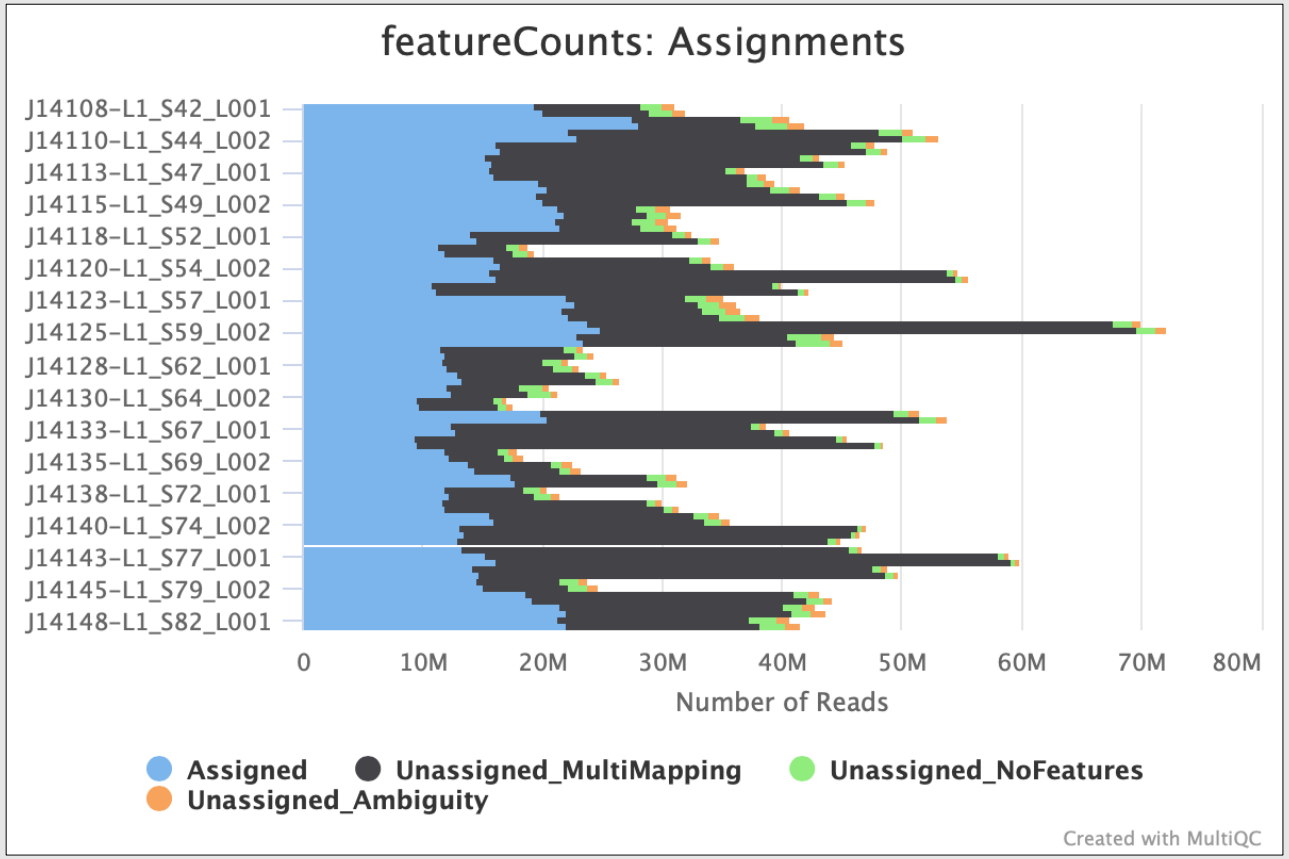
- overlap size (full read vs. partial overlap)
- multi-mapping reads
- reads overlapping multiple genomic features of the same kind
- reads overlapping introns

HTSeq offers three different modes
featureCounts default option is union

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

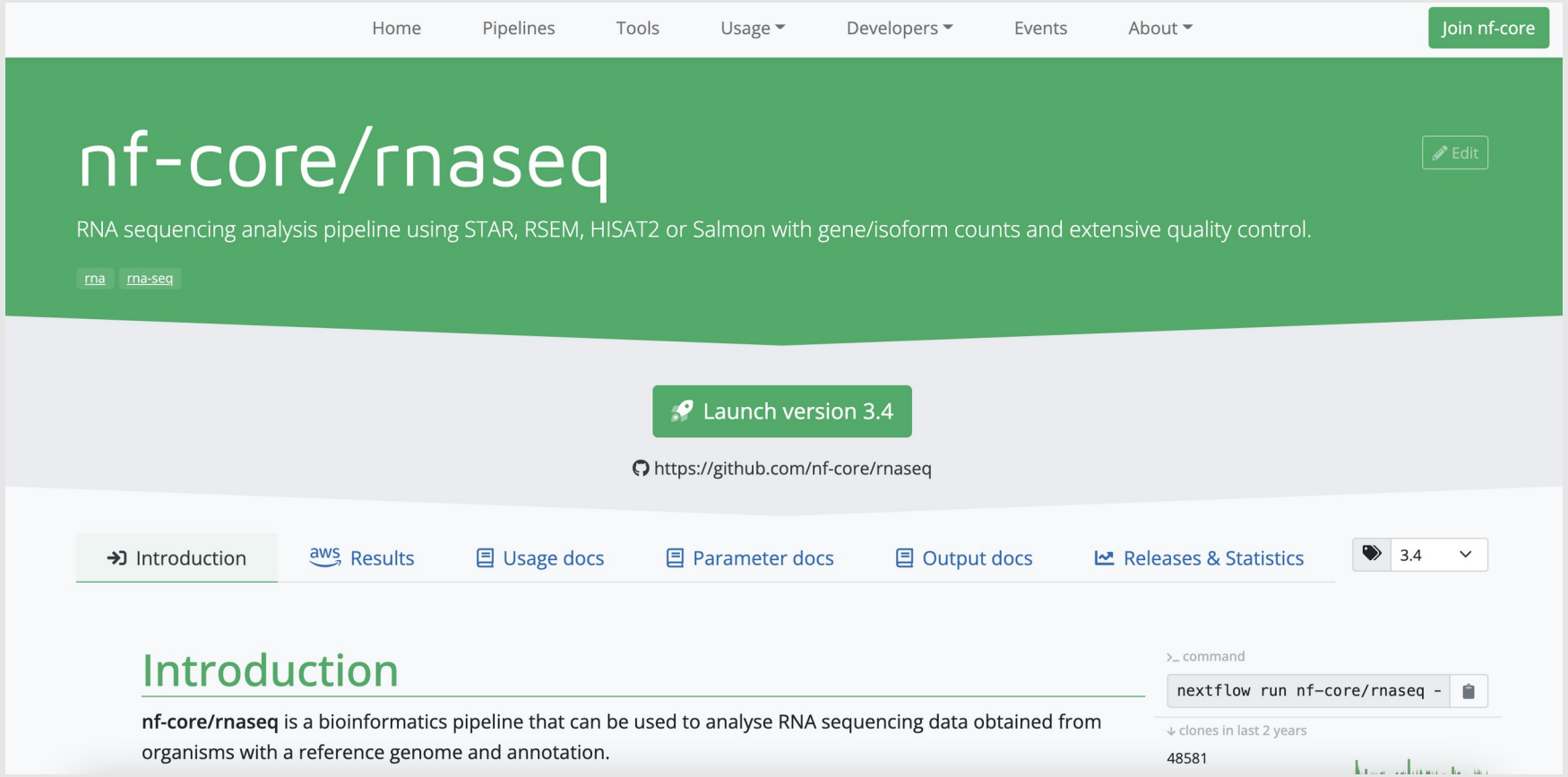
Read Quantification

QC: Features Assigned



RNA-seq preprocessing pipeline

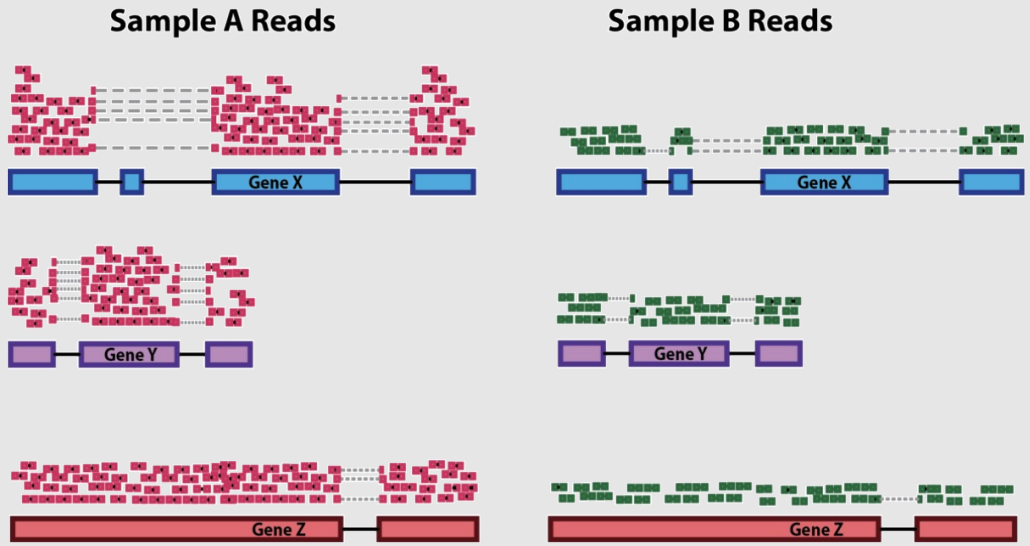
https://nf-co.re/rnaseq



Normalizing Read Counts

Why do we need normalization?

1. Sequencing depth

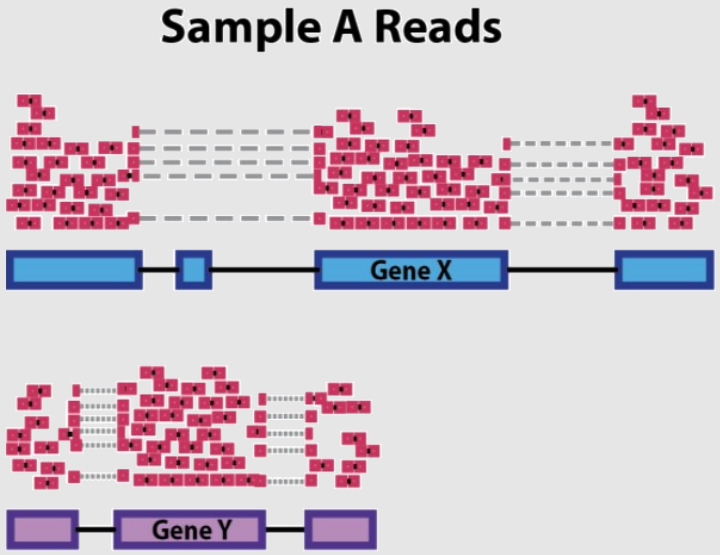


https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalizing Read Counts

Why do we need normalization?

1. Sequencing depth
2. Gene length

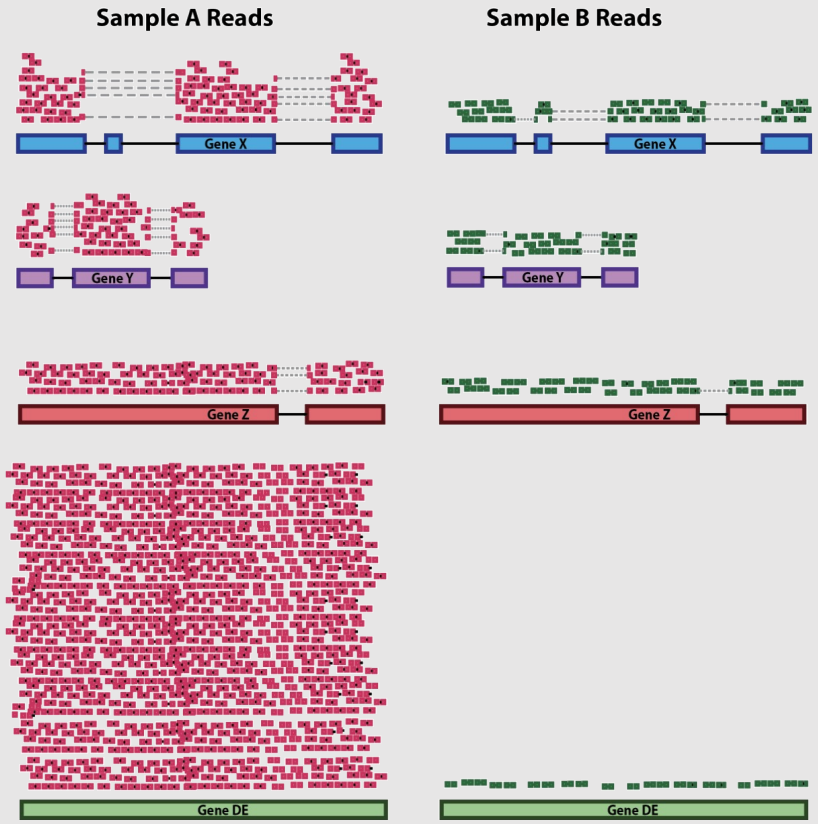


https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalizing Read Counts

Why do we need normalization?

1. Sequencing depth
2. Gene length
3. RNA composition



https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

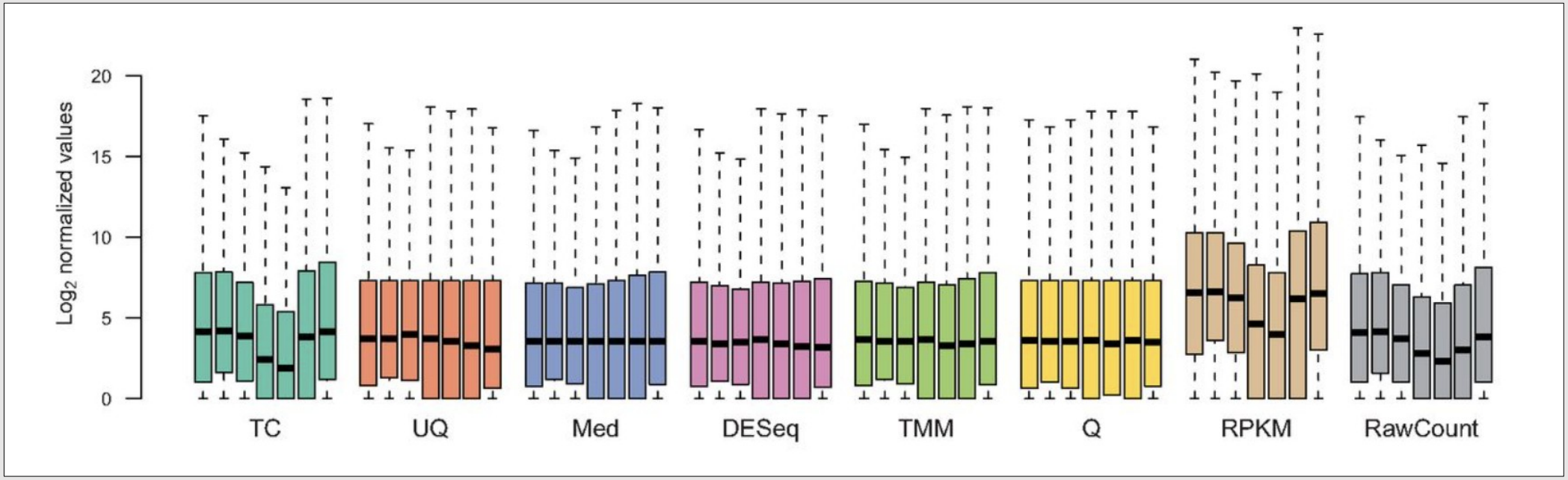
Normalizing Read Counts

Common normalization methods

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 samplegroup; **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**	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)**	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalizing Read Counts



Dillies et al Briefings in Bioinformatics 2013

Normalizing Read Counts

1. Create pseudo-reference sample (row-wise geometric mean)
2. Calculate ratio of each sample to pseudo-reference
3. Calculate normalization factor
4. Calculate normalized counts

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**

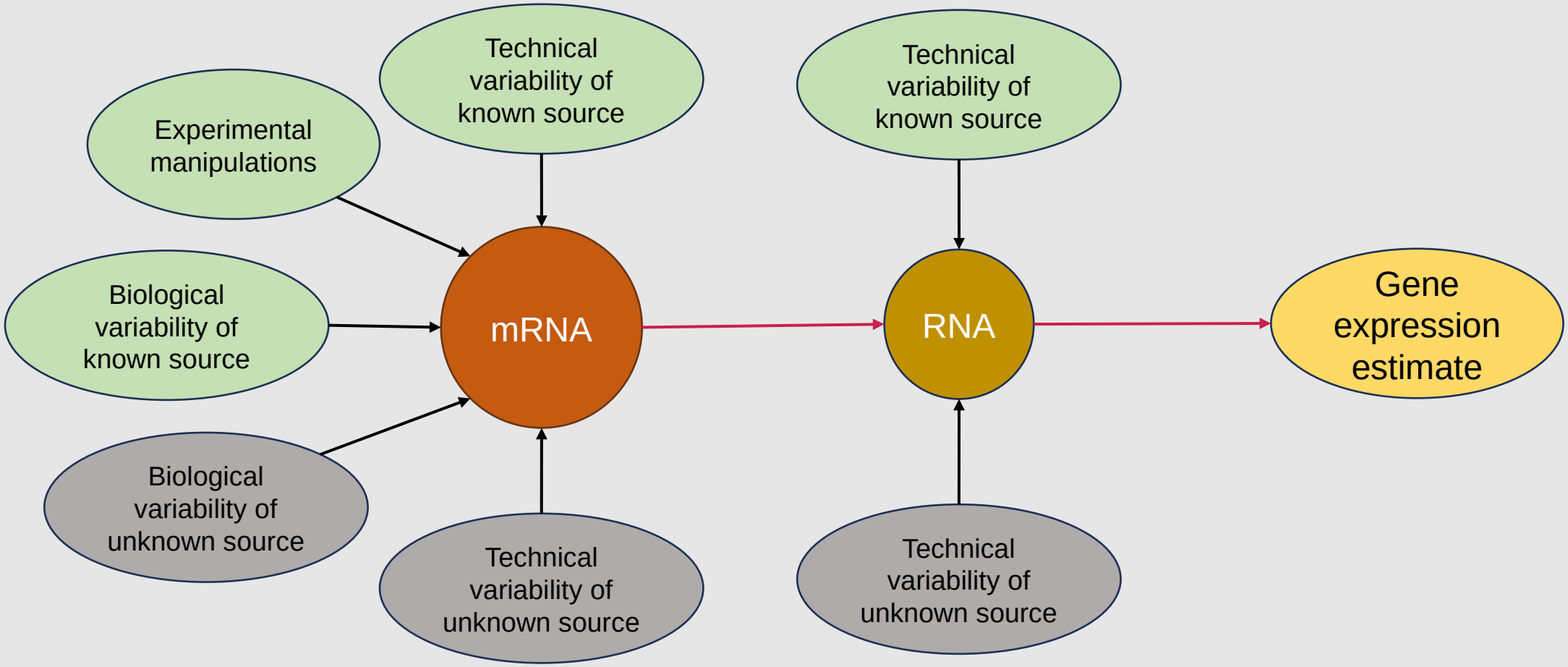
```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))

normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Data exploration

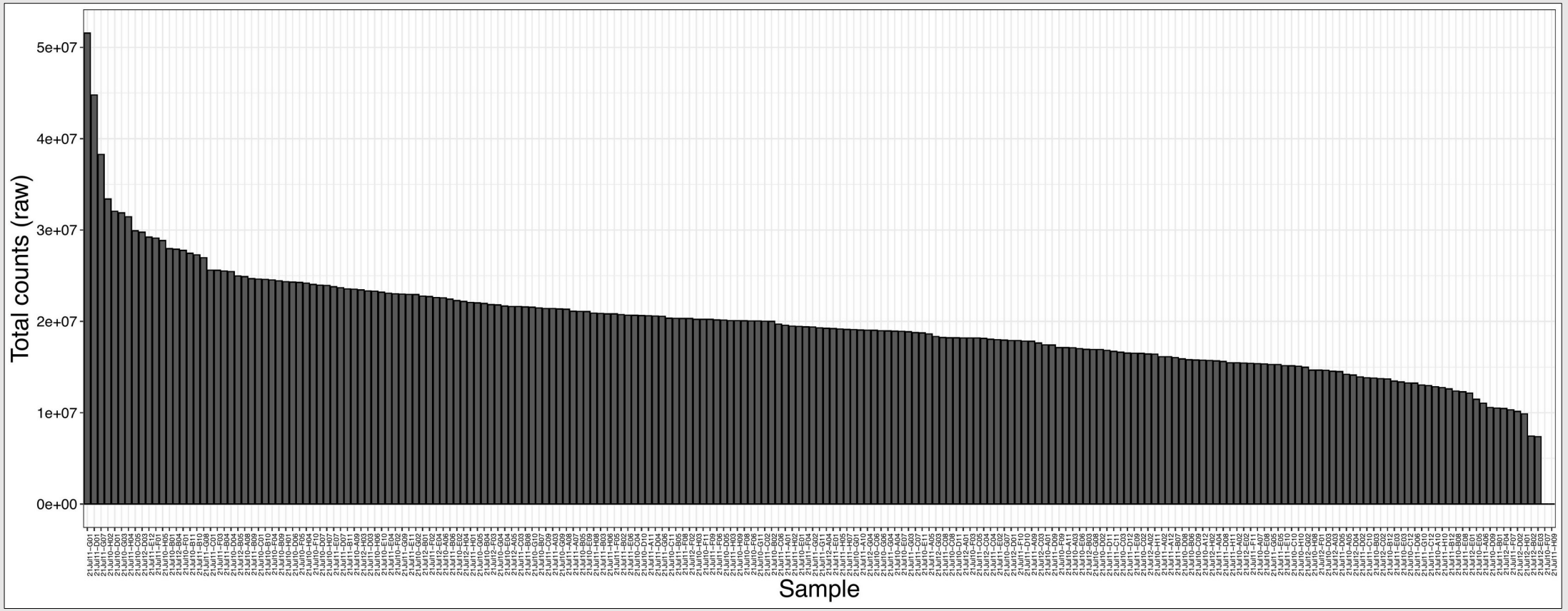


Adapted from DGE workshop by HBC training

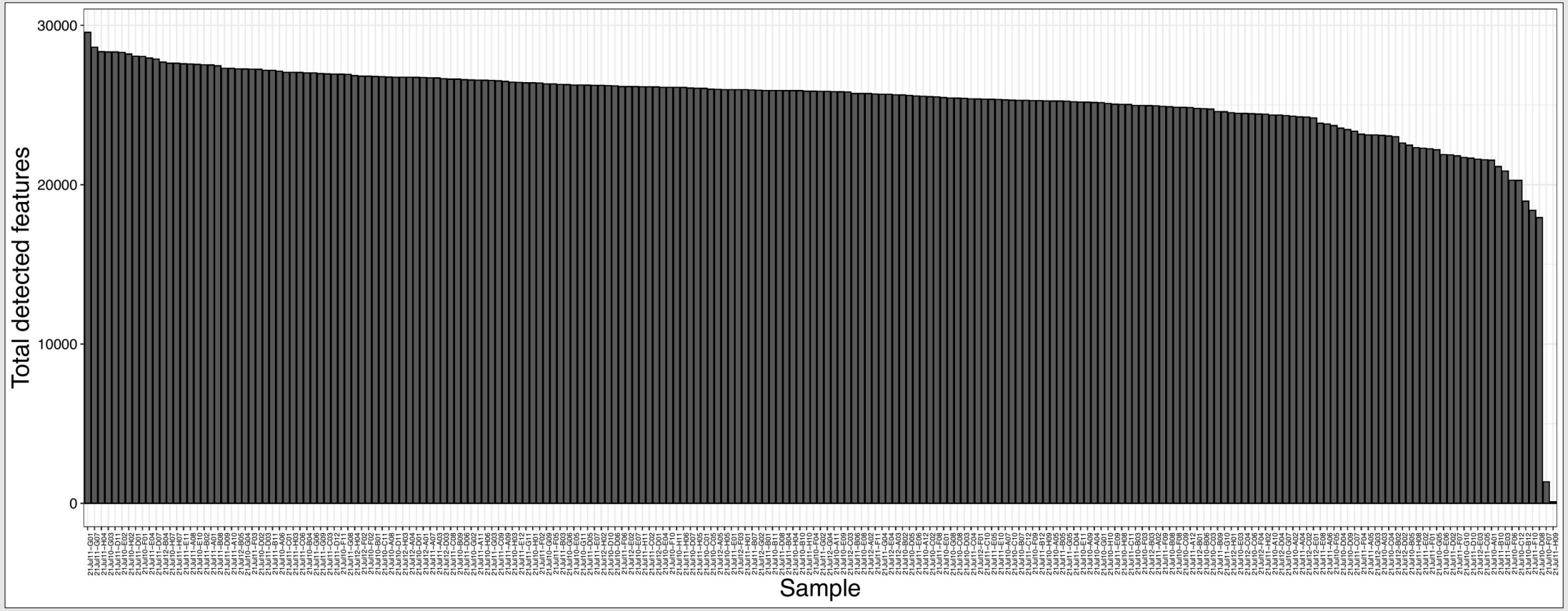
Data exploration

- Assessing the quality of the data
- Principal component analysis
- Detecting outliers

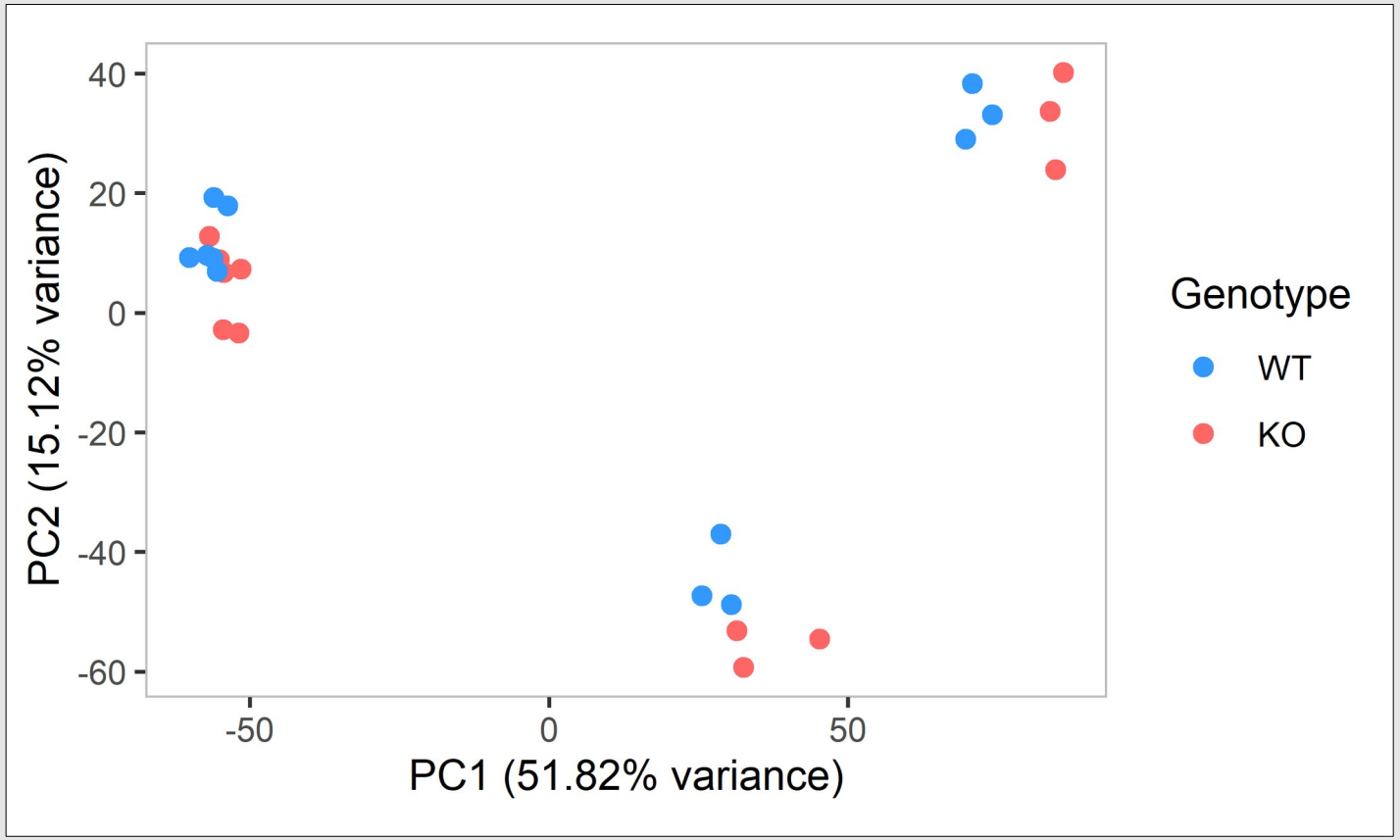
Distribution of counts



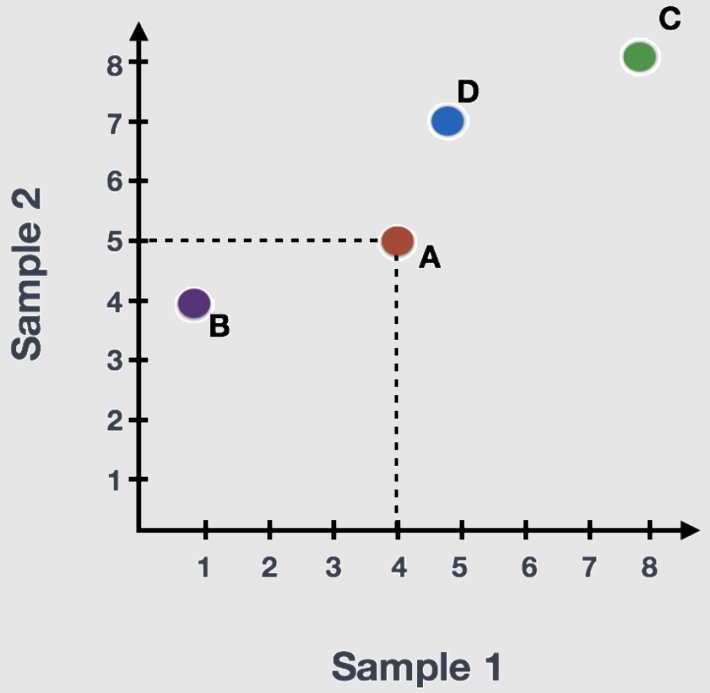
Distribution of genes



Principal Component Analysis



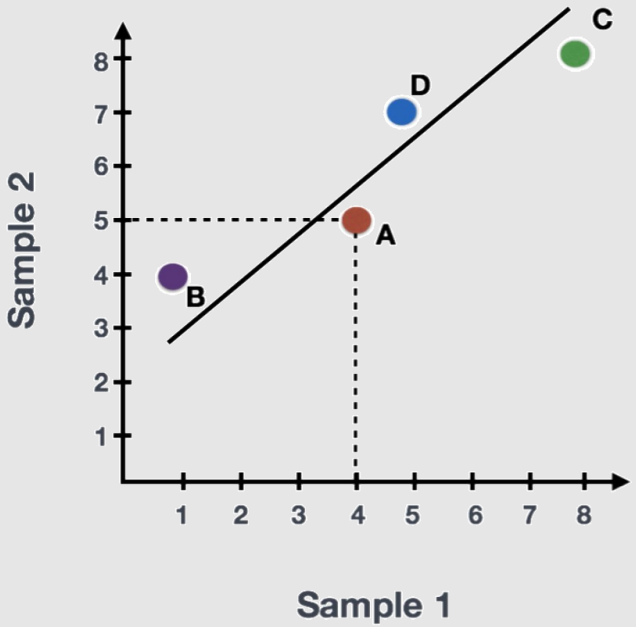
Principal Component Analysis



	Sample 1	Sample 2
Gene A	4	5
Gene B	1	4
Gene C	8	8
Gene D	5	7

https://hbctraining.github.io/DGE_workshop_salmon/lessons/03_DGE_QC_analysis.html

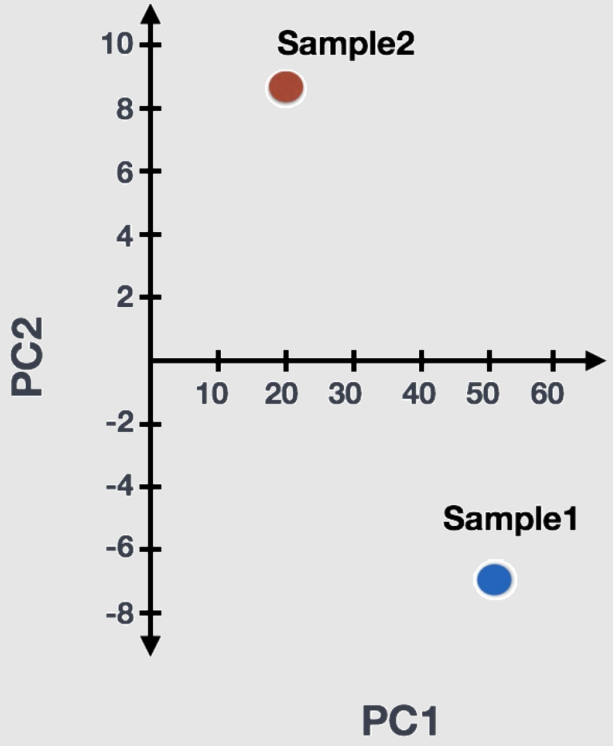
Principal Component Analysis



Sample PC1 score = (read count of Gene A * influence Gene A) +
(read count of Gene B * influence Gene B) +
(read count of Gene C * influence Gene C) +
(read count of Gene D * influence Gene D)

https://hbctraining.github.io/DGE_workshop_salmon/lessons/03_DGE_QC_analysis.html

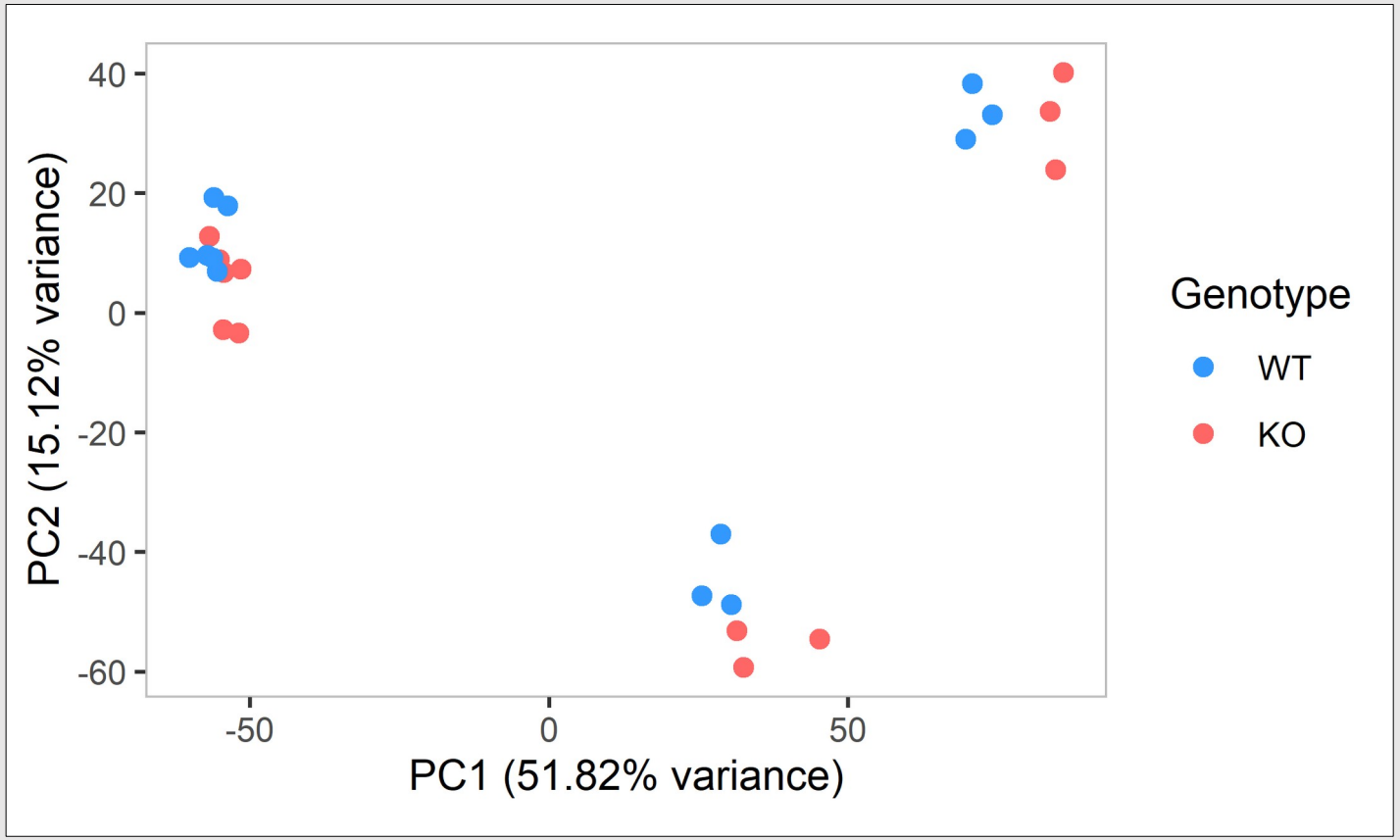
Principal Component Analysis



	PC1	PC2
Sample1	51	-7
Sample2	21	8.5

https://hbctraining.github.io/DGE_workshop_salmon/lessons/03_DGE_QC_analysis.html

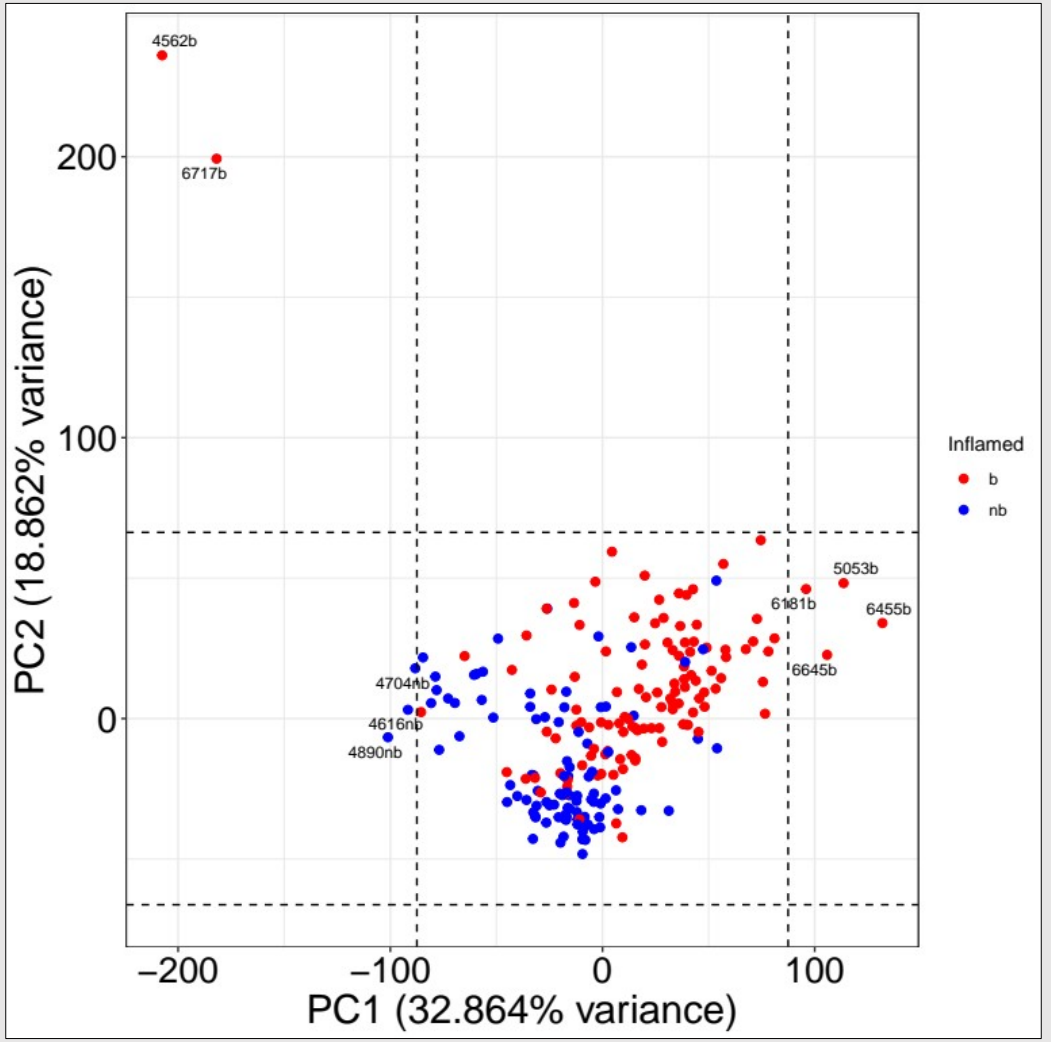
Principal Component Analysis



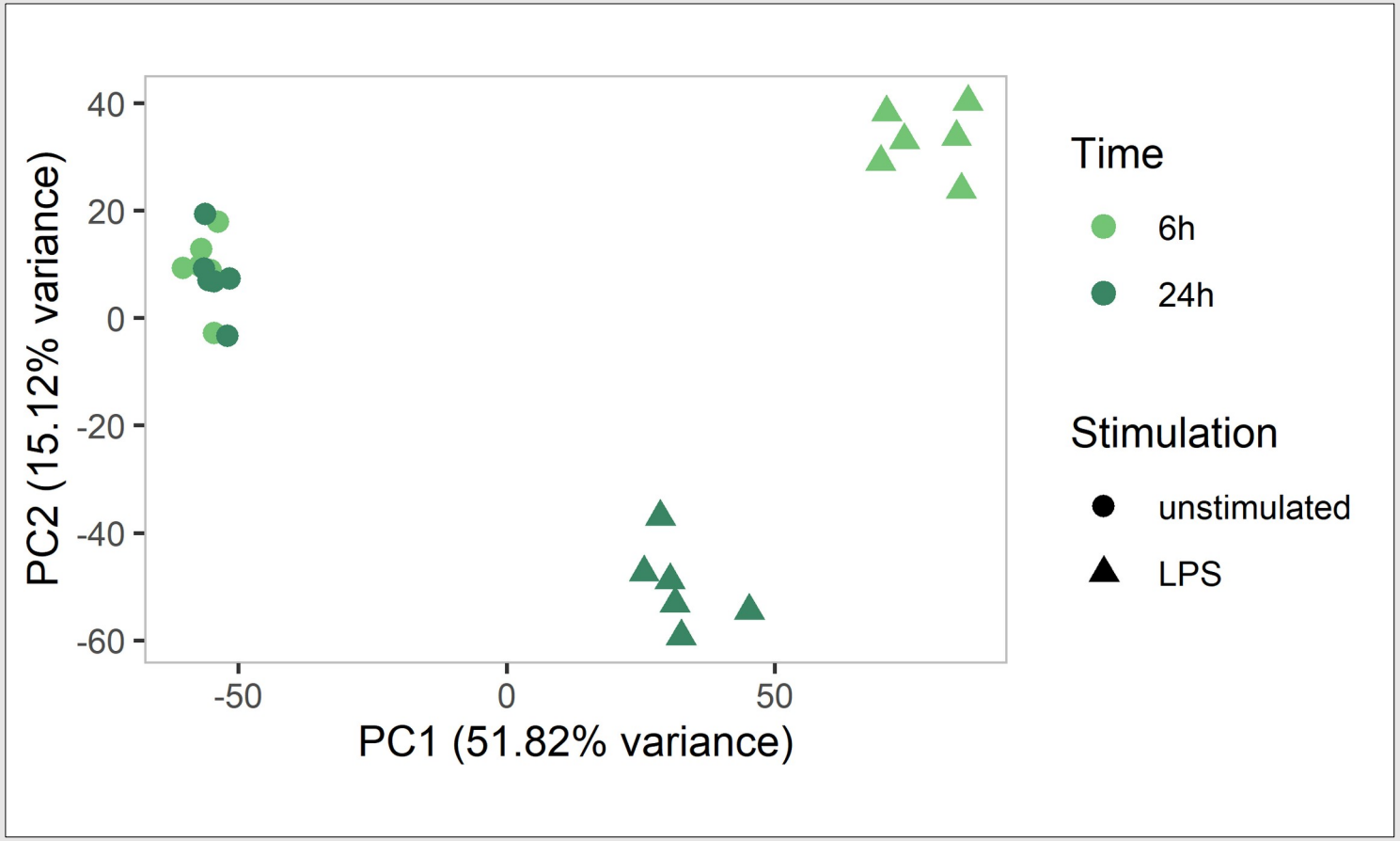
WT and KO cells

- Unstimulated
- Stimulated for 6h
- Stimulated for 12 h

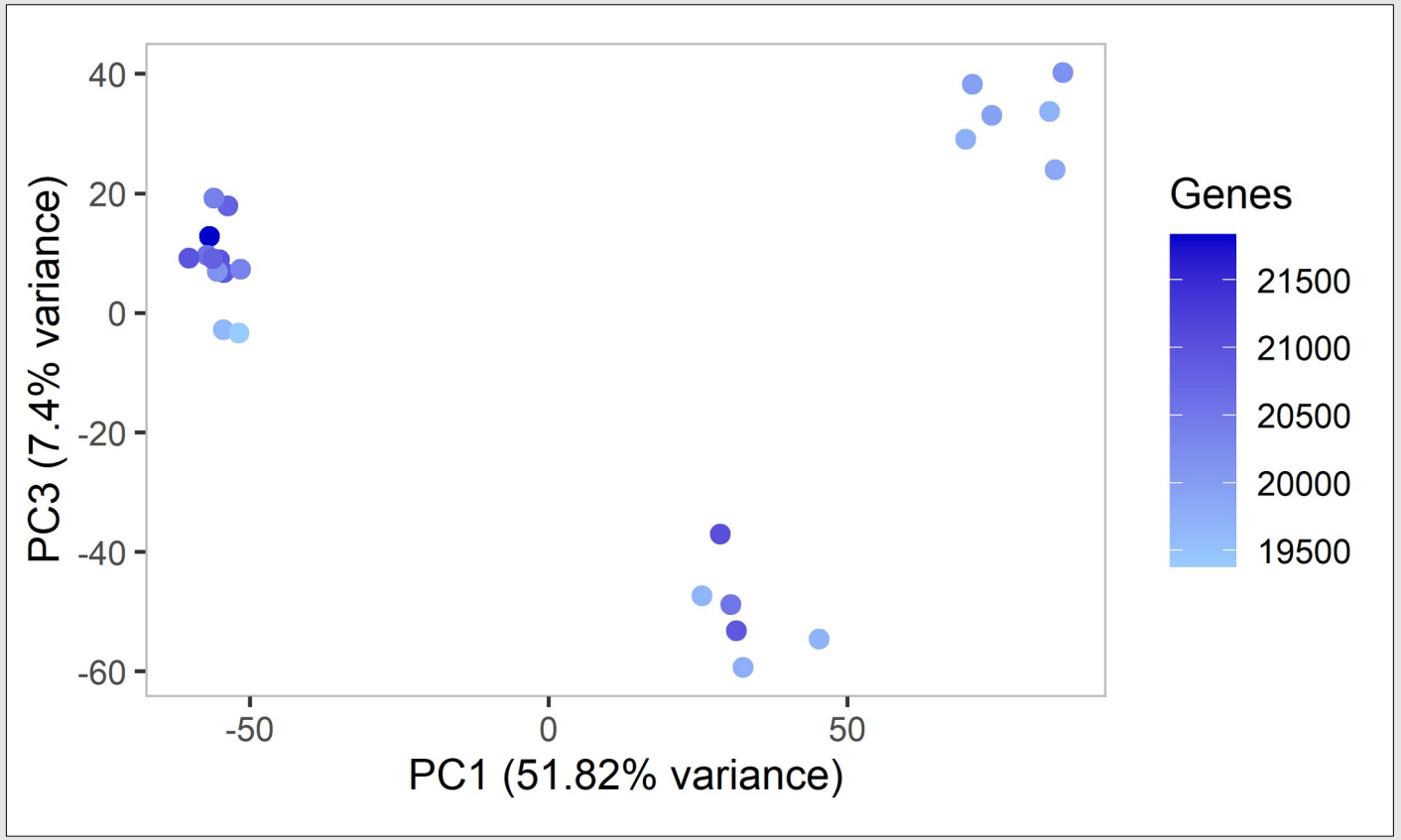
Detecting outliers



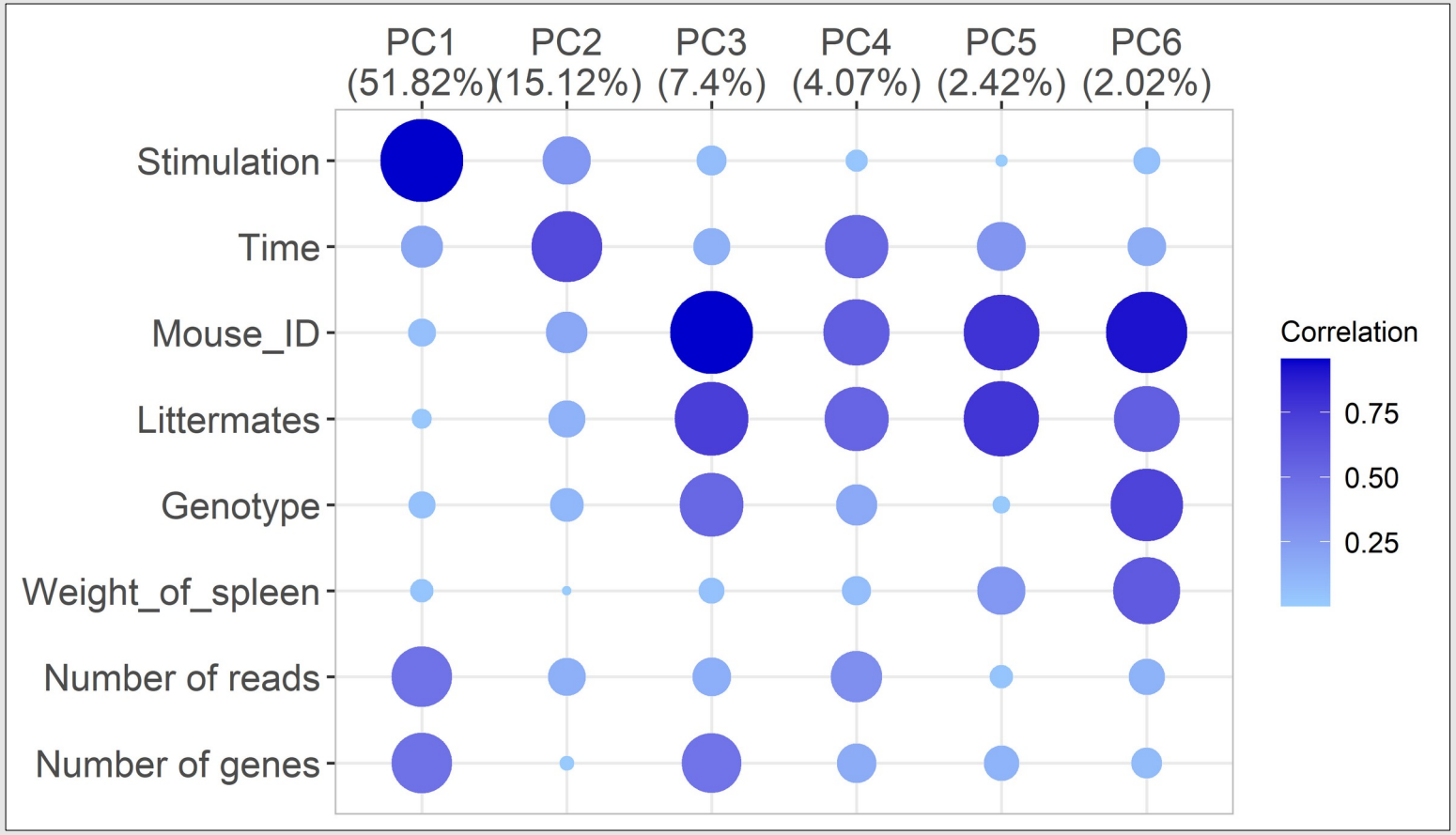
Principal Component Analysis



Principal Component Analysis

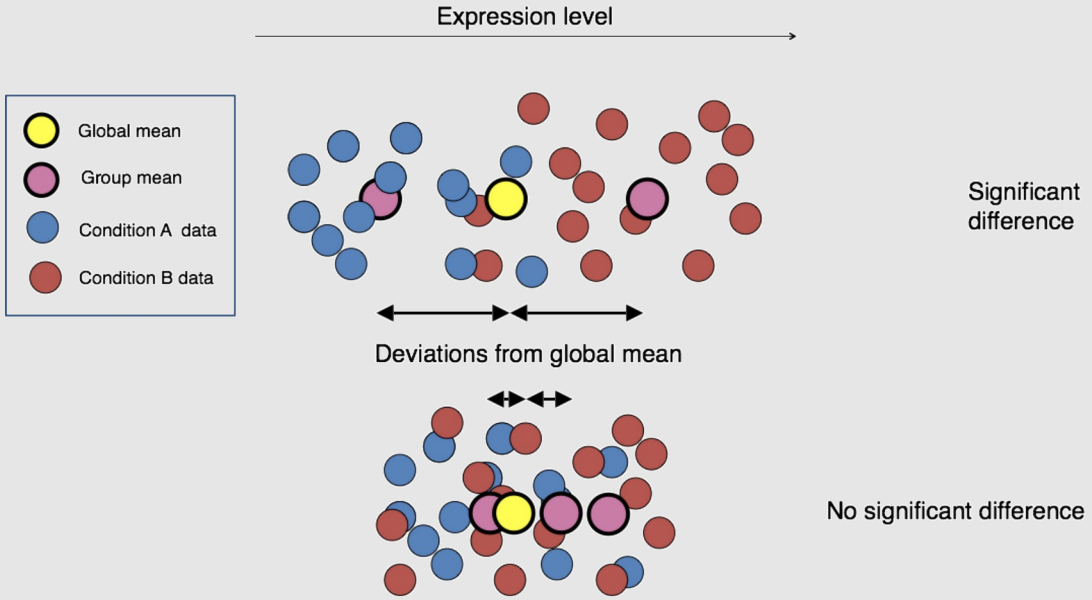


Principal Component Analysis



Differential Gene Expression

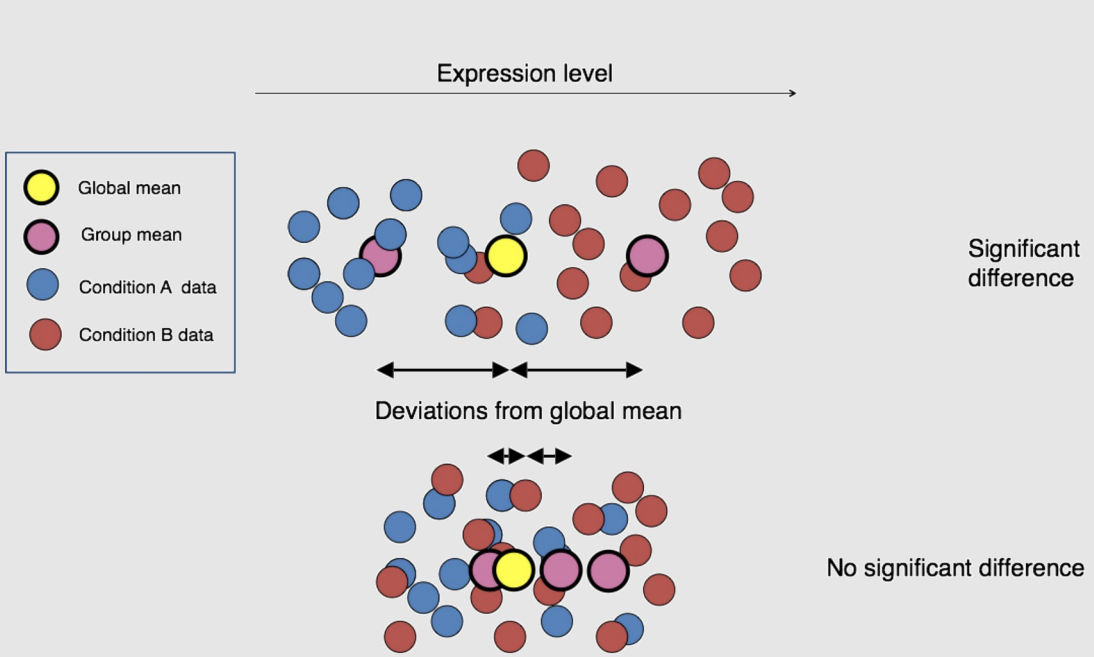
DESeq2 analysis workflow

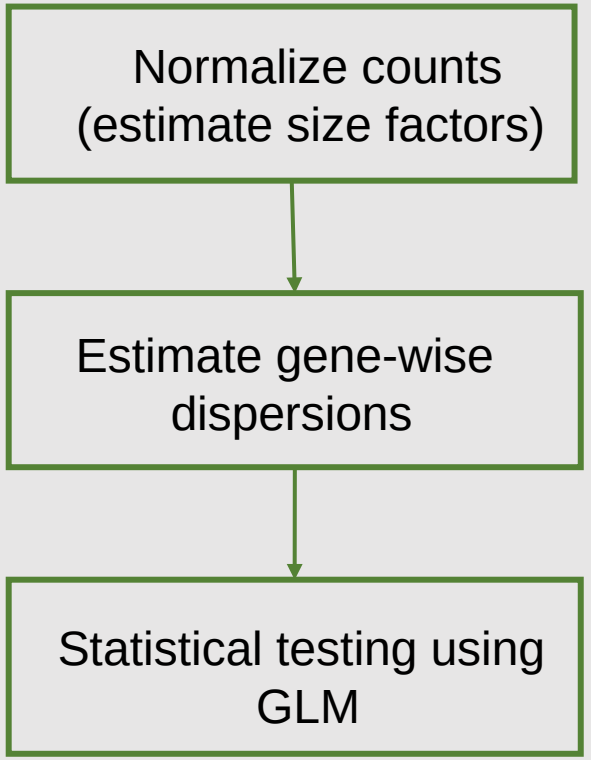


https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html
http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Differential Gene Expression

DESeq2 analysis workflow



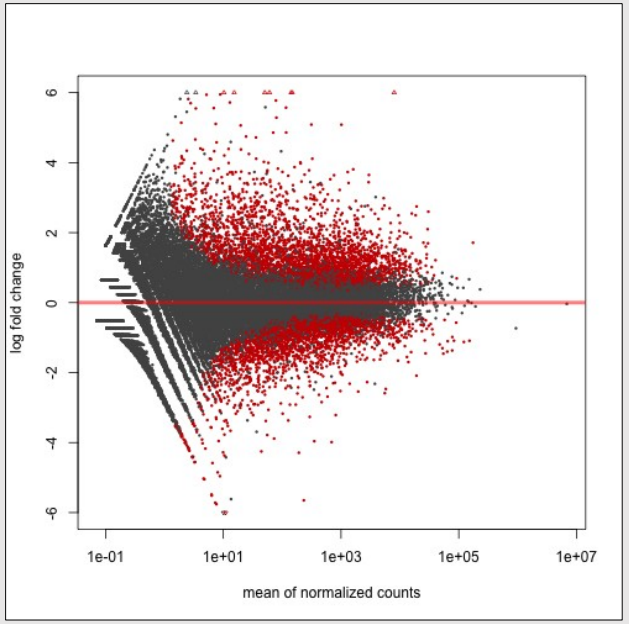


https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html
http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Differential Gene Expression

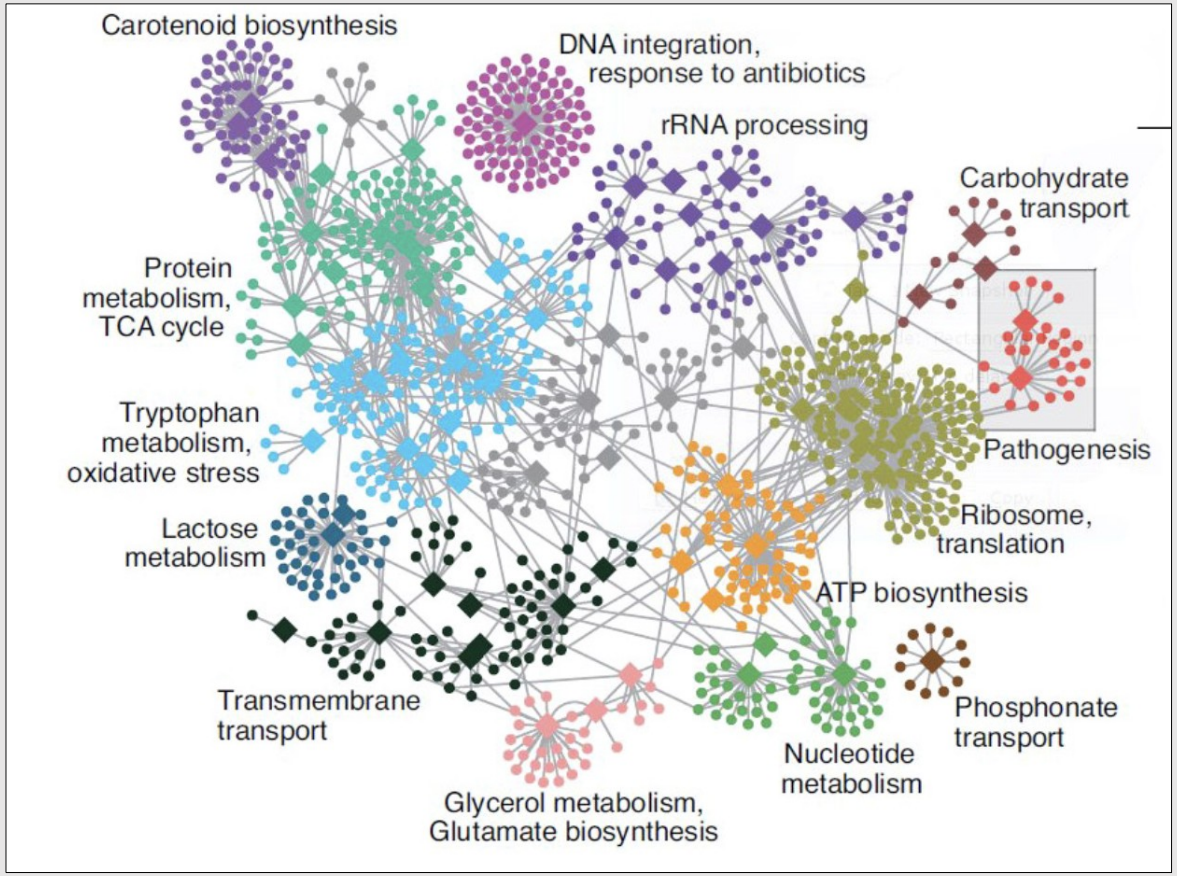
Result interpretation

	A	B	C	D	E	F	G	H	I
1		baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	gene	
2	ENSG00000135424	313.109892	5.087071073	0.239015731	21.2834153	1.62E-100	3.73E-96	ITGA7	
3	ENSG00000053108	70.3526611	4.857239409	0.409673254	11.85637423	1.99E-32	2.30E-28	FSTL4	
4	ENSG00000171241	68.4994869	2.94583187	0.26816503	10.98514551	4.51E-28	3.46E-24	SHCBP1	
5	ENSG00000157445	191.948513	-4.290148115	0.41983153	-10.21873731	1.63E-24	9.41E-21	CACNA2D3	
6	ENSG00000274383	351.587345	-3.960519808	0.391300552	-10.12142657	4.44E-24	2.05E-20	CTD-2017F17.2	
7	ENSG00000115425	231.512078	1.728793809	0.173157125	9.983959989	1.79E-23	6.88E-20	PECR	
8	ENSG00000172594	250.690737	2.352002487	0.24045012	9.78166486	1.35E-22	4.44E-19	SMPDL3A	
9	ENSG00000169679	116.550311	3.167021202	0.335901036	9.428435349	4.16E-21	1.20E-17	BUB1	
10	ENSG00000118113	1015.01076	5.083997701	0.544494027	9.337104628	9.90E-21	2.53E-17	MMP8	
11	ENSG00000189127	116.896432	4.860400364	0.541588933	8.974334712	2.85E-19	6.57E-16	ANKRD34B	
12	ENSG00000173825	589.655847	-2.688283635	0.300450782	-8.947500885	3.64E-19	7.61E-16	TIGD3	
13	ENSG00000115884	19.9554979	4.376934791	0.491659254	8.902374473	5.47E-19	1.05E-15	SDC1	
14	ENSG00000078053	58.278994	3.993929934	0.450573167	8.864109594	7.71E-19	1.37E-15	AMPH	
15	ENSG00000180535	15.4925082	2.726481506	0.311063524	8.765031252	1.87E-18	3.07E-15	BHLHA15	
16	ENSG00000138821	347.451712	2.378683313	0.276888502	8.590762319	8.64E-18	1.24E-14	SLC39A8	
17	ENSG00000164104	1765.50371	2.087084376	0.242739363	8.598046675	8.11E-18	1.24E-14	HMGB2	
18	ENSG00000276644	89.2865496	2.863486048	0.334498343	8.560538815	1.12E-17	1.52E-14	DACH1	
19	ENSG00000090376	3772.17918	2.538474177	0.303099183	8.375061086	5.52E-17	7.06E-14	IRAK3	
20	ENSG00000138190	625.645083	1.751181259	0.21066268	8.312726593	9.35E-17	1.13E-13	EXOC6	

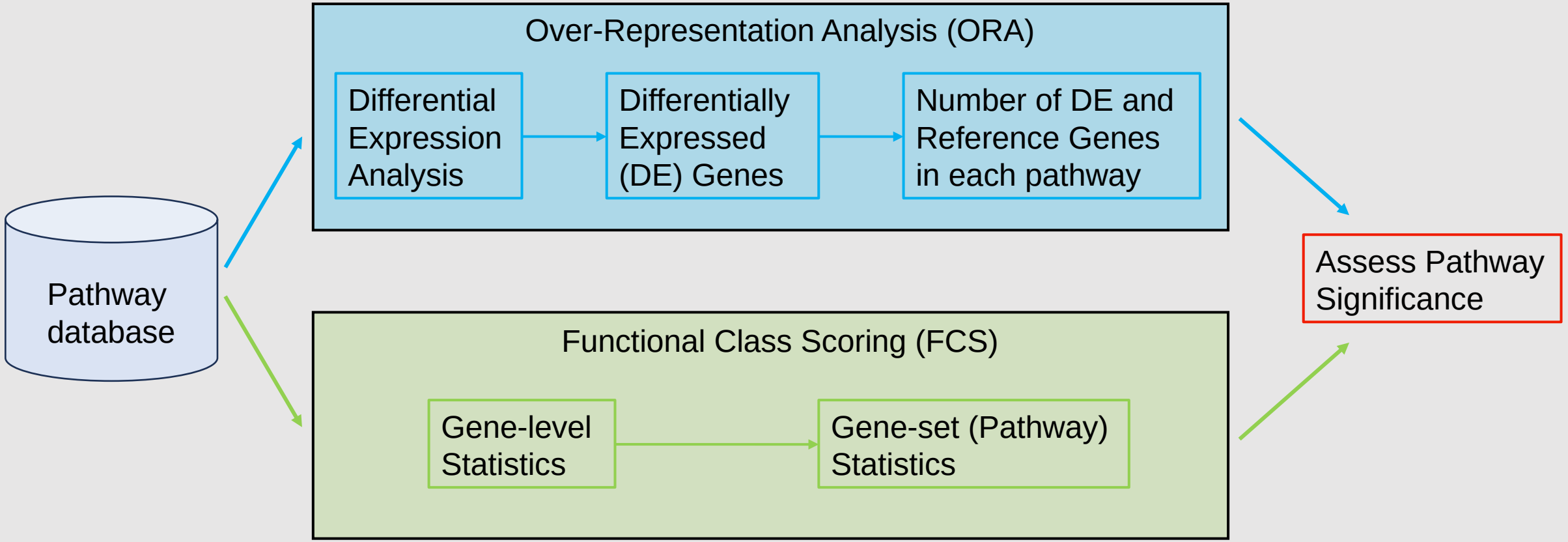


out of 34450 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up) : 3274, 9.5%
LFC < 0 (down) : 2756, 8%
outliers [1] : 60, 0.17%
low counts [2] : 11356, 33%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results

Functional Analysis

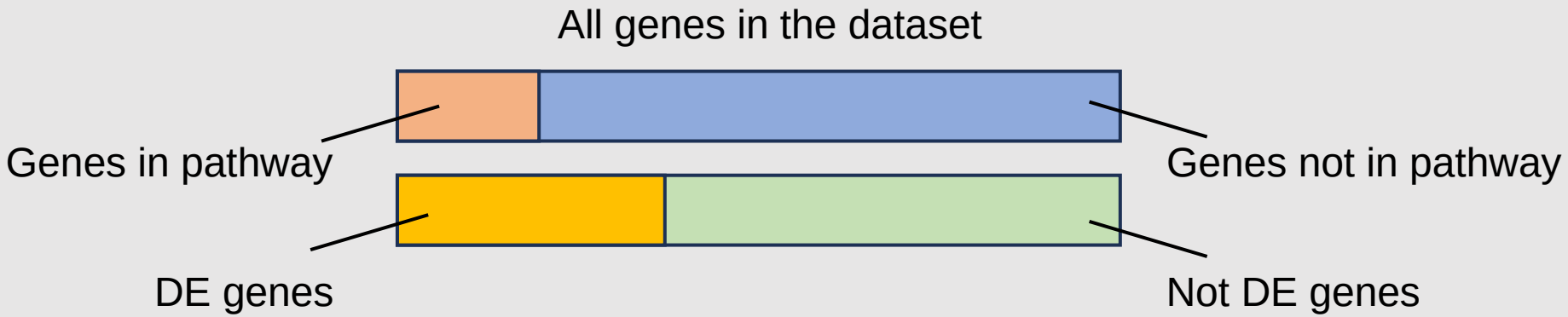


Functional Analysis



Adapted from DGE workshop by HBC training

Over-representation analysis

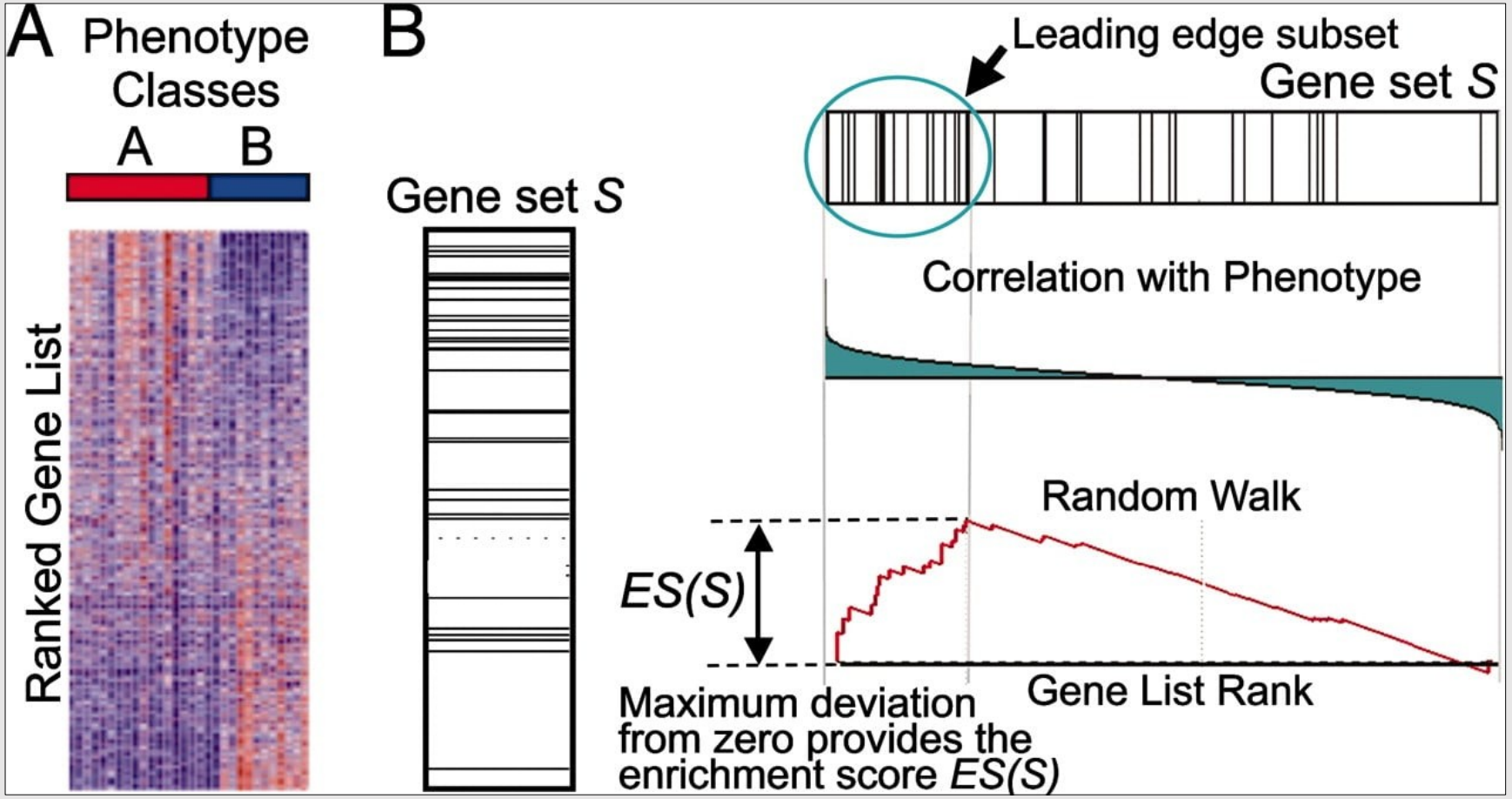


		DE genes		
		yes	no	
In pathway	yes	20	300	320
	no	80	19600	19680
		100	19900	20000

https://bioinformatics-core-shared-training.github.io/Bulk_RNAseq_Course_Apr22/

Functional Class Scoring (Advanced)

Gene Set Enrichment Analysis (GSEA)



https://doi.org/10.1073/pnas.0506580102

Resources

https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

https://github.com/hbctraining/DGE_workshop

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

https://lms.uni-kiel.de/url/Catalog/0/Search/0/Infos/5273846135
Password: biomedinf2022