



# GLADSTONE INSTITUTES

# Requirements for the demo

**Please install the following library in Rstudio**

```
install.packages("magrittr")
install.packages("pheatmap")
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("BiocManager")
BiocManager::install("edgeR")
BiocManager::install("org.Mm.eg.db")
```

**Verify the installation:**

**Open Rstudio and type**

```
library(magrittr)
library(pheatmap)
library(tidyverse)
library(ggplot2)
library(edgeR)
```

We will be starting at 1.05p

# Intermediate RNA-seq data analysis using R

Michela Traglia  
Bioinformatics Core, GIDB

October 6th, 2025

**GLADSTONE**  
INSTITUTES

# Poll 1

How familiar you are with R/RNAseq concepts:

1. I attended the ‘Introduction to RNA-seq’ Gladstone workshop or similar courses
2. I attended ‘Introduction to R’ Gladstone workshop/experience with R
3. I have experience in analysing the RNASeq data
4. No experience with RNAseq
5. No experience with R

# What to expect from this workshop

- o Understanding why we need to normalize the reads
- o Overview of the normalization methods and edgeR pipeline
- o Understanding how to fit the reads in each gene to get the DGE and how to analyze complicated comparisons
- o Visualize the DEG and interpretation
- o Performing the analysis with a real dataset using R packages

# Assumed background

- o Familiarity with R and RStudio
- o Familiarity with RNA-seq protocol
- o Familiarity with basic concepts of statistics and hypothesis testing

# Introductions

Ayushi Agrawal

Bioinformatician III

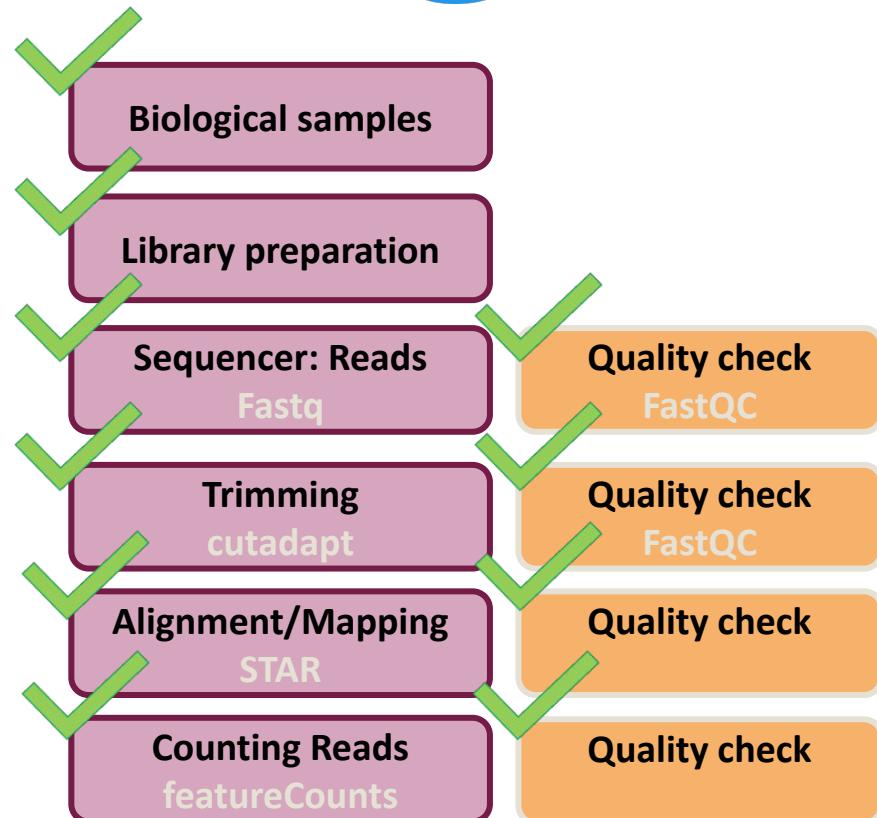
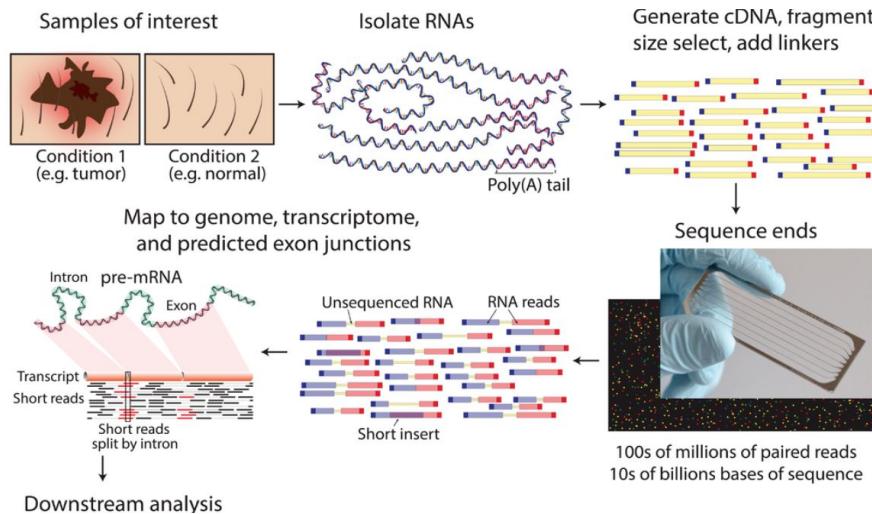
Michela Traglia

Senior Statistician

# Outline

- Intro to a real experiment, dataset and biological questions
- Source of variability
- Filtering genes
- Exploratory visualization: MDS – PCA
- Bias in Differentially Expressed Gene analysis (DEG)
- Approach for DEG analysis: normalization in edgeR
- Modeling the read in each gene to get the DEG
- Compare groups and visualize the DEG
  - Demo (from count matrix to DEG and plots)

# RNA-seq - analysis workflow

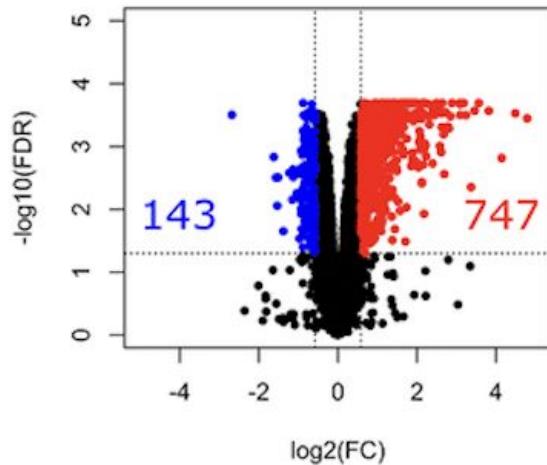


# From the count matrix to the DEG

Each row is a gene

Each column is a sample

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666
AAPC	1454	7777	7781	7781	1240	7781	7781	1257



# Other quantification from the count matrix

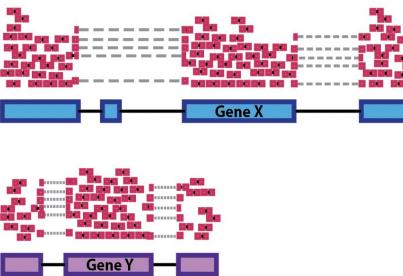
Each row is a gene

Each column is a sample

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666
AAZC	1451	7777	7704	7724	1240	7400	7074	1557

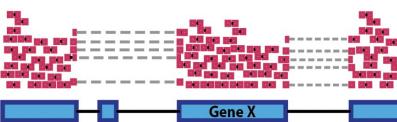
1. Get the expression for each gene within one sample (Sample A)

Sample A Reads

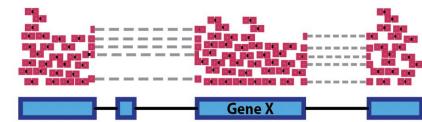


2. Compare the expression between technical replicates (Sample A1, A2)

Sample A.1 reads



Sample A.2 reads



# Reference for the workshop

F1000Research

F1000Research 2016, 5:1438 Last updated: 06 DEC 2018



Check for updates

SOFTWARE TOOL ARTICLE

**REVISED** From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; referees: 5 approved]

Yunshun Chen<sup>1,2</sup>, Aaron T. L. Lun <sup>3</sup>, Gordon K. Smyth <sup>1,4</sup>

<sup>1</sup>The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia

<sup>2</sup>Department of Medical Biology, The University of Melbourne, Victoria, 3010, Australia

<sup>3</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK

<sup>4</sup>Department of Mathematics and Statistics, The University of Melbourne, Victoria, 3010, Australia

# Materials for this workshop

Please download the compressed file **2025Oct\_intermediatRNAseq.zip**

Double click on the file, the unzipped folder includes:

## **Hands-on session files:**

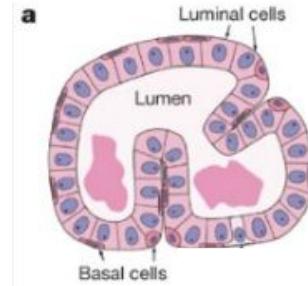
- 2025Oct\_IntermediateRNASeq\_handson.R #each step in R
- targets.txt #sample sheet
- DE\_results\_handson.txt #results table
- GSE60450\_Lactation-GenewiseCounts.txt.gz #count matrix

This **presentation** with concepts

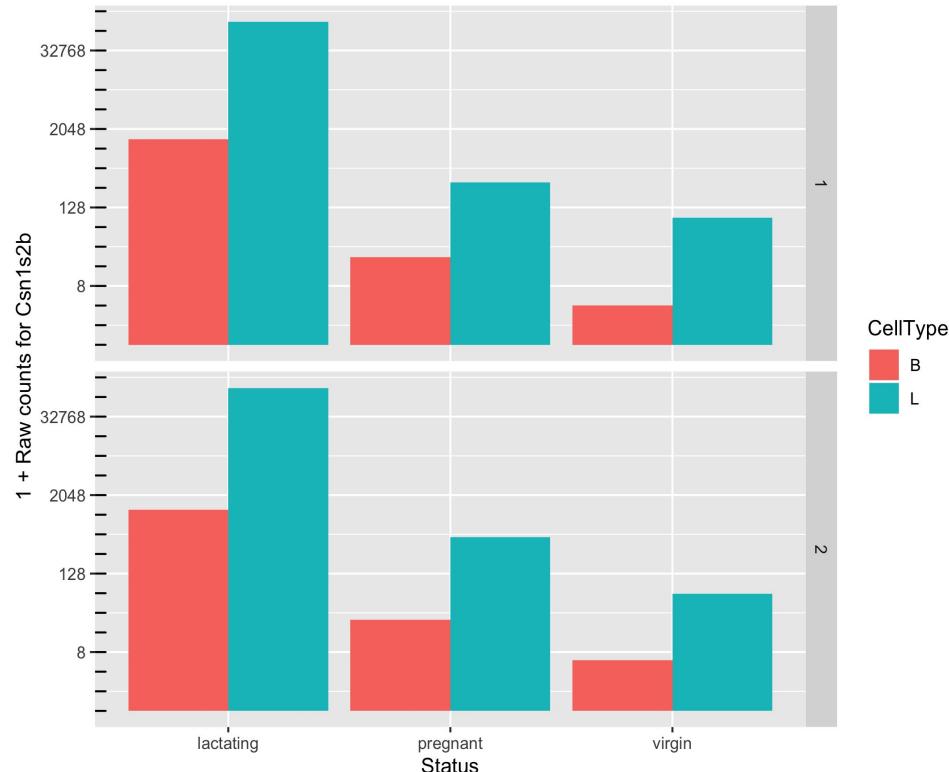
# Dataset

Transcriptome analysis of luminal and basal cell subpopulations in the lactating versus pregnant mammary gland

- GEO (gene expression omnibus) accession: **GSE60450**
- Tissue of origin: Mammary glands of mouse
- Cell types: Basal stem-cell enriched cells (B) and committed luminal cells (L)
- Biological conditions: Virgin, Lactating (2 day) and Pregnant (18.5 day)
- # of groups: 2 cell types (B/L) × 3 conditions (V/L/P) = 6 groups
- # of replicates: 2 of each group
- Illumina Hiseq sequencer - about 30 million 100bp single-end reads for each sample.



# Potential biological questions



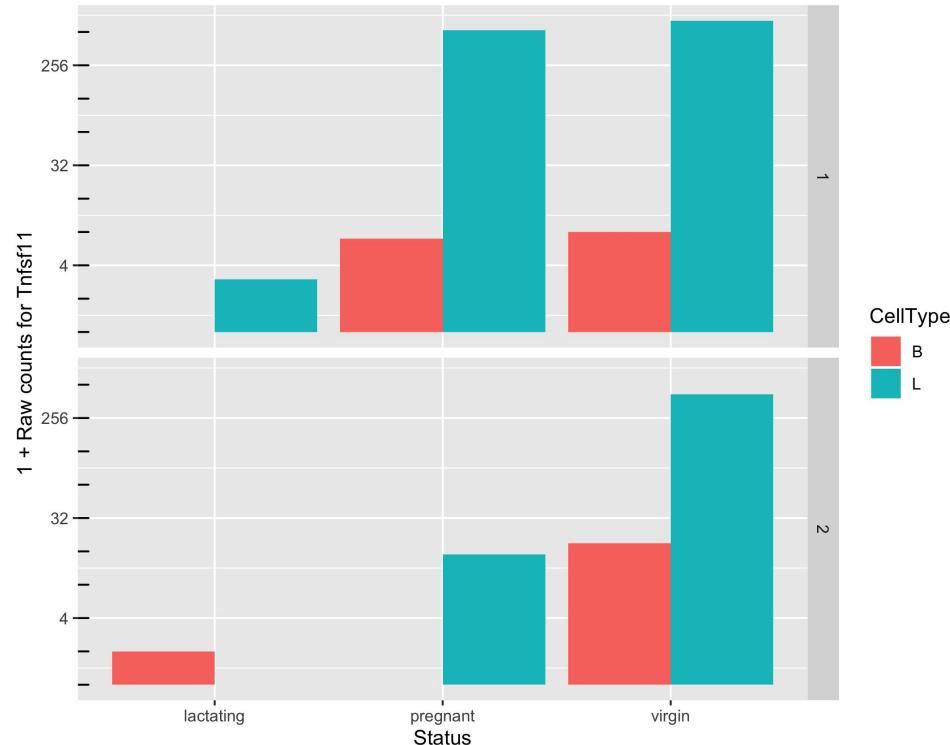
Which comparisons are we interested in?

Example:

1. B vs L,
2. B.lactating vs L.pregnant,
3. ...
4. All of them

Or not interested in comparisons but in gene expression of one sample

# Potential issues



Can we make reliable  
inferences for genes with very  
low counts? What should we  
consider “very low”?

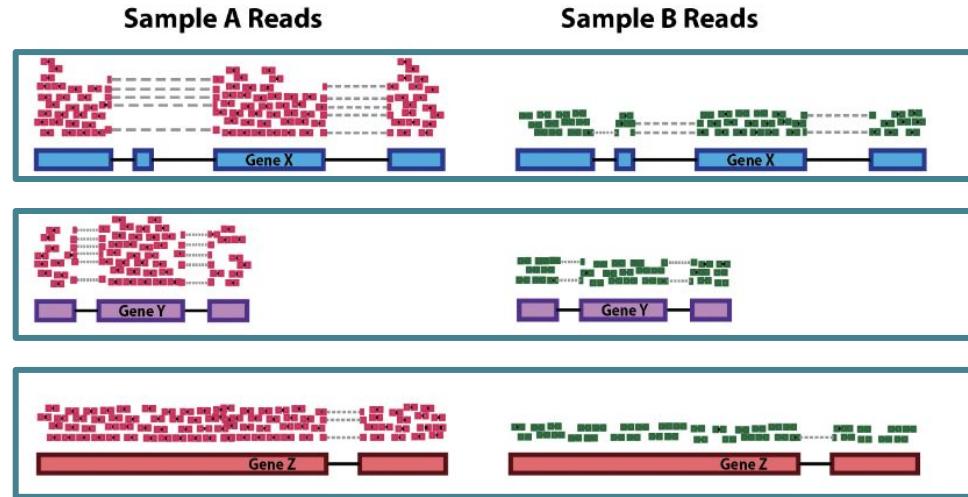
# Scenario 1

- **Tissue of origin:** Mammary glands of mouse
- **Cell types:** Basal stem-cell enriched cells (B) and committed luminal cells (L)
- **Biological conditions:** Virgin, Lactating (2 day) and Pregnant (18.5 day)
- **# of groups:** 2 cell types (B/L) × 3 conditions (V/L/P) = 6 groups
- **# of replicates:** 2 of each group

**Quantify if gene X and gene Y are differentially expressed in two sample groups (ex. DEGs Pregnant vs Lactating within B cell type, B.lactating vs L.pregnant)**

# Scenario 1

## DEG - between groups



Pairwise -  
Gene X group 1 compared  
to Gene Y group 2

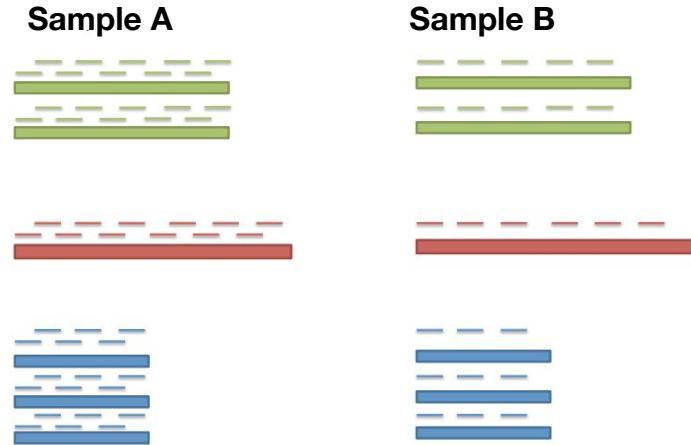
Gene Y group 1 compared  
to Gene Y group 2

**DEG:** Compare gene expressions across samples between conditions/genotype/groups

# Source of bias, gene filtering and PCA

# We need to detect and correct for the sources of bias

Sample A has more counts assigned to these 3 genes compared to sample B



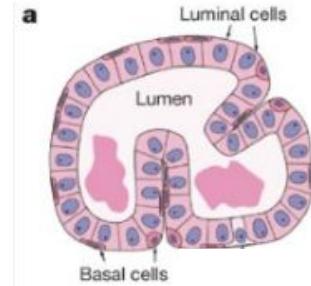
**Source of bias: Sequencing depth or library size or total number of reads of each sample**

# Different sequence depths may cause variation in counts

Variation in sequencing depths => Need to normalize counts

Group	Total counts
B.virgin	23085177
B.virgin	21628857
B.pregnant	23919152
B.pregnant	22490570
B.lactating	21382233
B.lactating	19884434

Group	Total counts
L.virgin	20213223
L.virgin	21509988
L.pregnant	22073815
L.pregnant	21837341
L.lactating	24638939
L.lactating	24581591



Library size about 20M

# Sequence depth/Library size - Count Per Million mapped reads

$$\text{RPM or CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

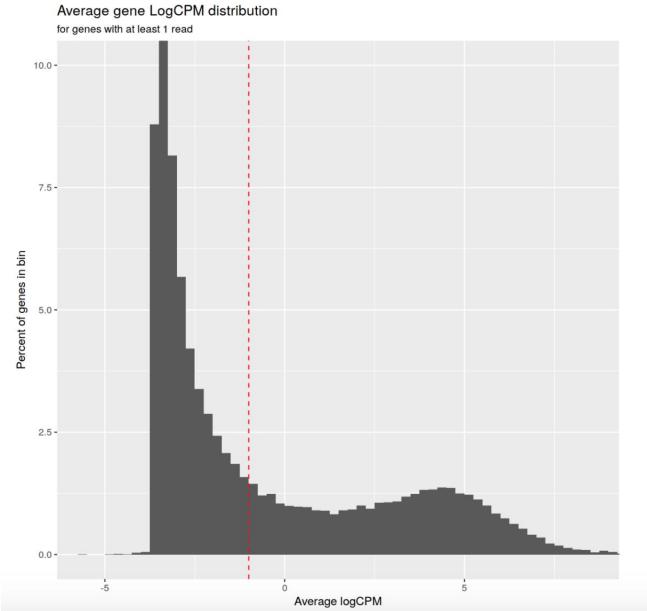
Sequenced one library with 5 million(M) reads.

Total 4 M matched to the genome sequence

5000 reads matched to a given gene

$$\text{RPM or CPM} = \frac{5000 \times 10^6}{4 \times 10^6} = 1250$$

Filter on count-per-million (CPM) values to avoid favoring genes that are expressed in larger libraries over those expressed in smaller libraries



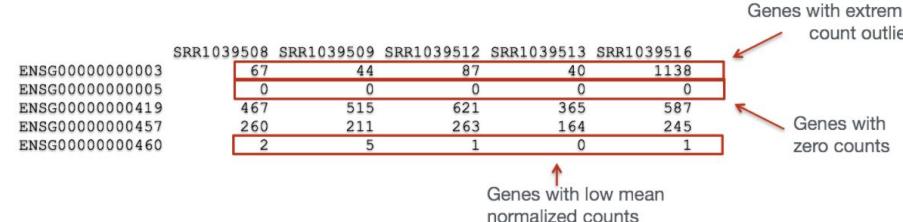
**A gene at least 10–15 counts in at least some libraries before it is considered to be expressed -> Identifying the CPM that corresponds to 10-15 counts**

# “Uninteresting” genes

## Filtering

- *Biological point of view*: minimal expression level of a gene -> translation into a protein -> biologically relevant
- *Statistical point of view*: low counts -> not enough statistical evidence.

Genes with consistently low counts are very unlikely be assessed as significantly DE



Using cpm to filter out genes

# Need to correct for multiple testing

P value represents the chance that we may be wrong in calling something significantly differential. Example:

- P = 0.01 means 1% chance that we may be wrong.
- P = 0.50 means 50% chance that we may be wrong.

More than 20k genes under consideration

=> if a certain difference in expression levels has only 1% chance of happening given the null model, it might be observed for 200 genes even if the null model were true for all the genes.

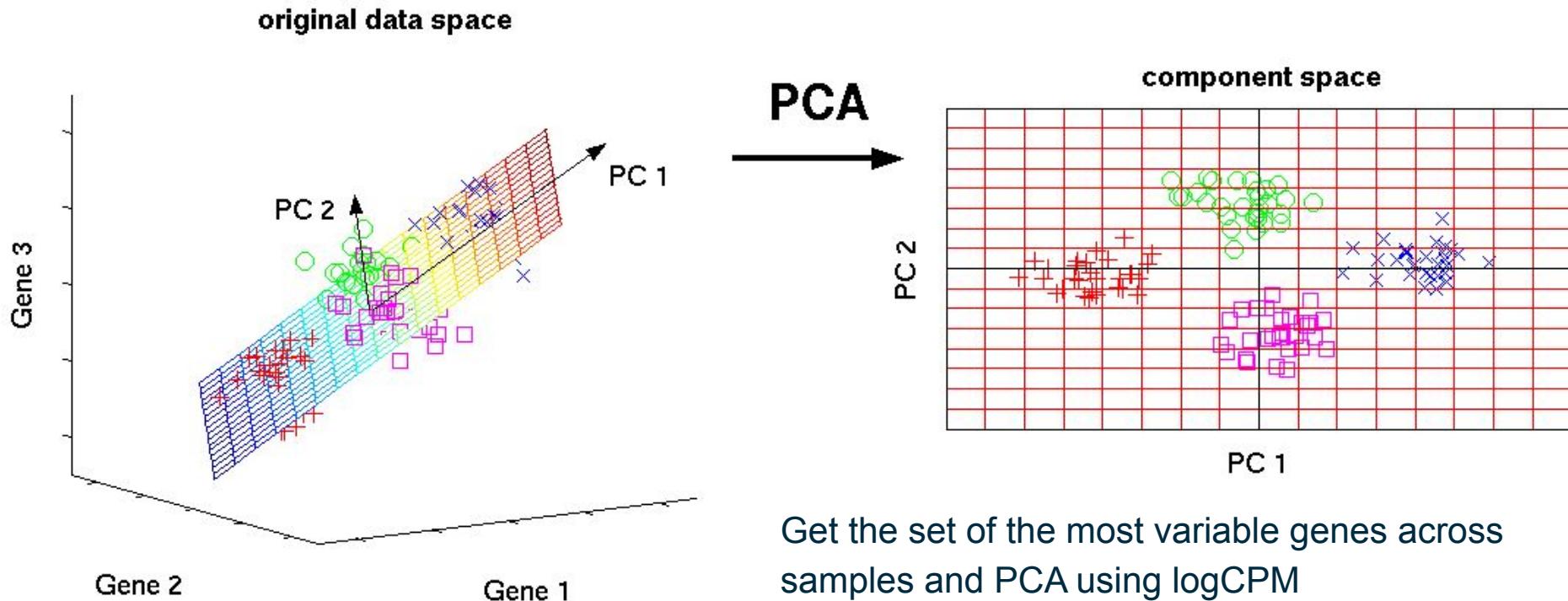
=> 200+ false positives

Hence, there is a need to adjust the p-values.

- The more genes we test, the more we must adjust.
- Reduce the number of tests by filtering out “uninteresting” genes.

# Assessing overall similarity across samples

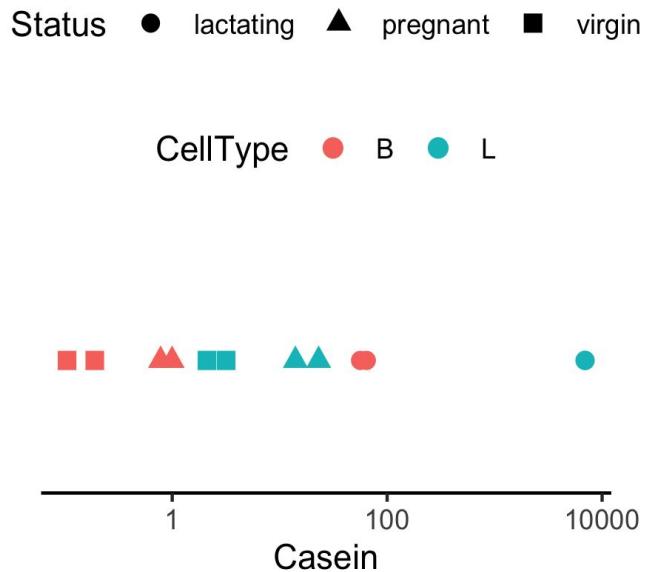
- Which samples are similar to each other, which are different?
- Does this fit to the expectation from the experiment's design?
- What are the major sources of variation in the dataset?



# Expression level of Casein varies in a way that is strongly indicative of the effect of CellType and Status.

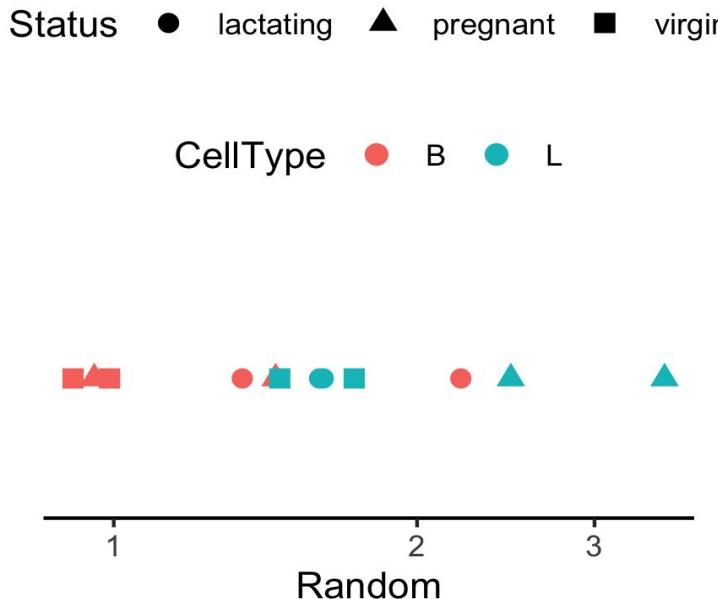
Why are the B.lactating samples not close to B.virgin and B.pregnant samples?

Could it be due to batch effects?



# Expression appears to vary across samples but...

In general, the way expression appears to vary across samples could be dominated by noise, batch effects, real signal, etc.



Identify the source for technical variability!

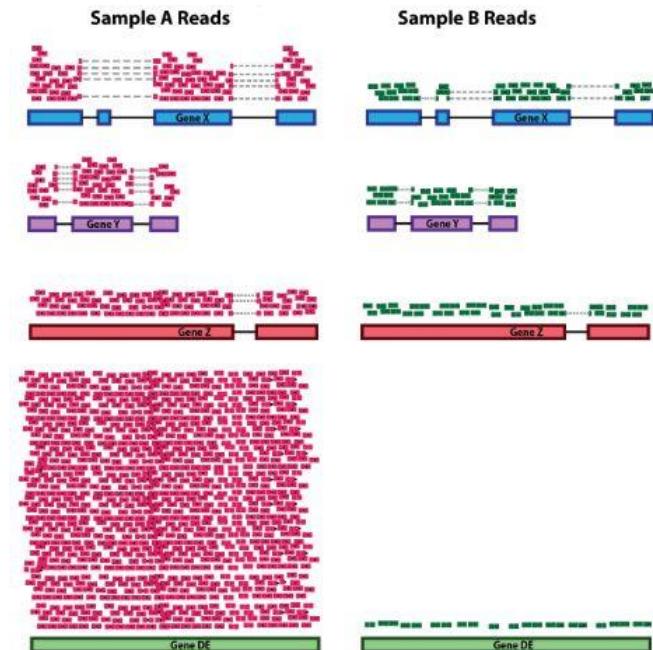
# Source of technical variability

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; <b>NOT for within sample comparisons or DE analysis</b>

logCPM used to filter genes with low counts and visualize the PCA but not enough for the DEG analysis or within sample comparison

# DEG other source of bias to account for

A gene is overexpressed in Sample A compared to Sample B and it seems that the other genes being lower expressed



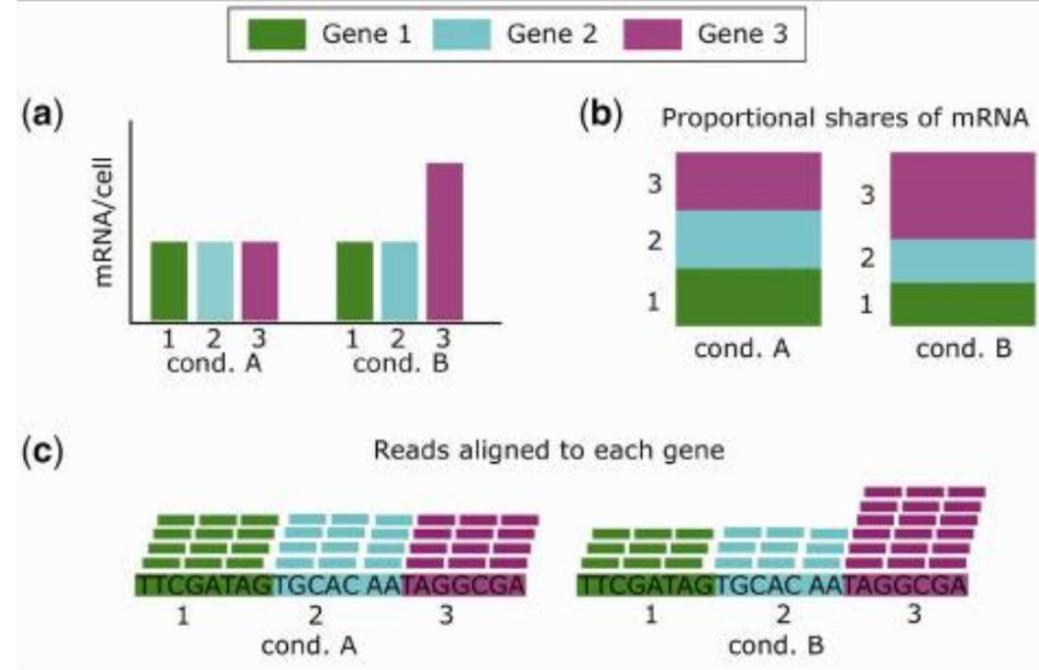
Different RNA sample composition

# Highly expressed genes bias the real DEG

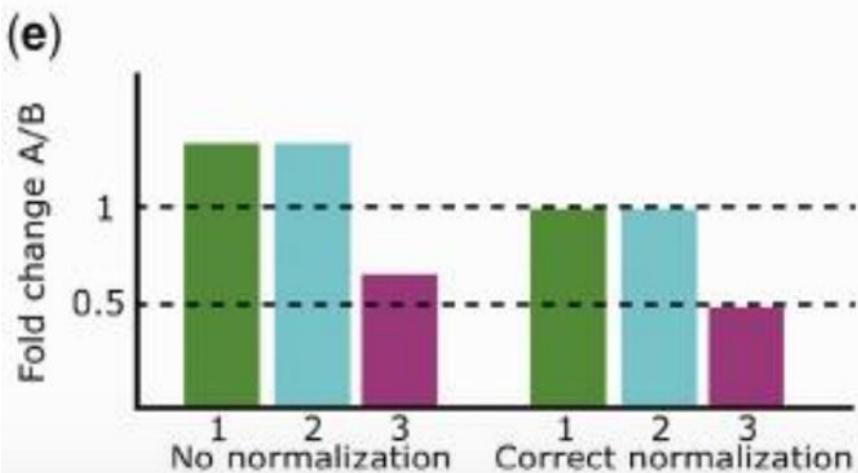
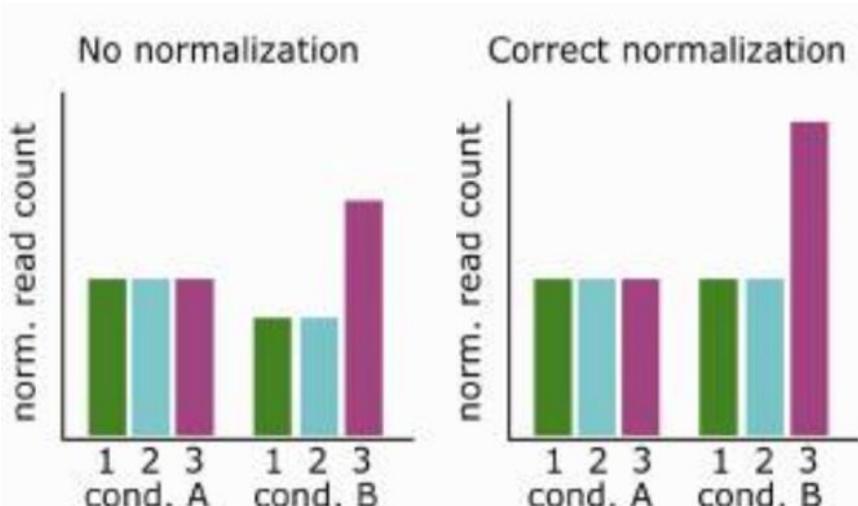
Total read count for each sample is the same.

Few highly expressed genes make up a greater share of the total molecules (total library size) of samples B.

Smaller fraction of the reads will be left for the other genes for that samples (undersampling).



doi: [10.1093/bib/bbx008](https://doi.org/10.1093/bib/bbx008)



Observed counts depend on total reads sequenced AND sample composition

**FC=1 noDE**

Gene 1 and 2 not-DE

Gene 3 is 2X DE

# Sample-to-sample variability in total RNA concentration

When **total RNA composition** is similar across samples, some normalization could be potentially used.

But you can't assume it!

Cells don't necessarily produce similar levels of RNA/cell between cell types, disease states or developmental stages is not always valid

$$\text{Number of reads for GENE1} = \frac{\text{Amount of nucleic acid from GENE1}}{\text{Total nucleic acid in sample}} \times \text{Total reads}$$

Normalization for DEG: TMM (edgeR), RLE (DESeq)

# Normalization for DEG analysis

**A few highly differentially expressed genes may have a strong influence on read counts -> minimizing effect of such genes**

Assumption: A majority of transcripts is not differentially expressed

Adjust counts such that for most genes, counts are not differential.

Normalize for RNA composition by a **set of scaling factors** that **minimize** the log-fold changes between the samples for most genes.

# Approach to identify DE genes: edgeR

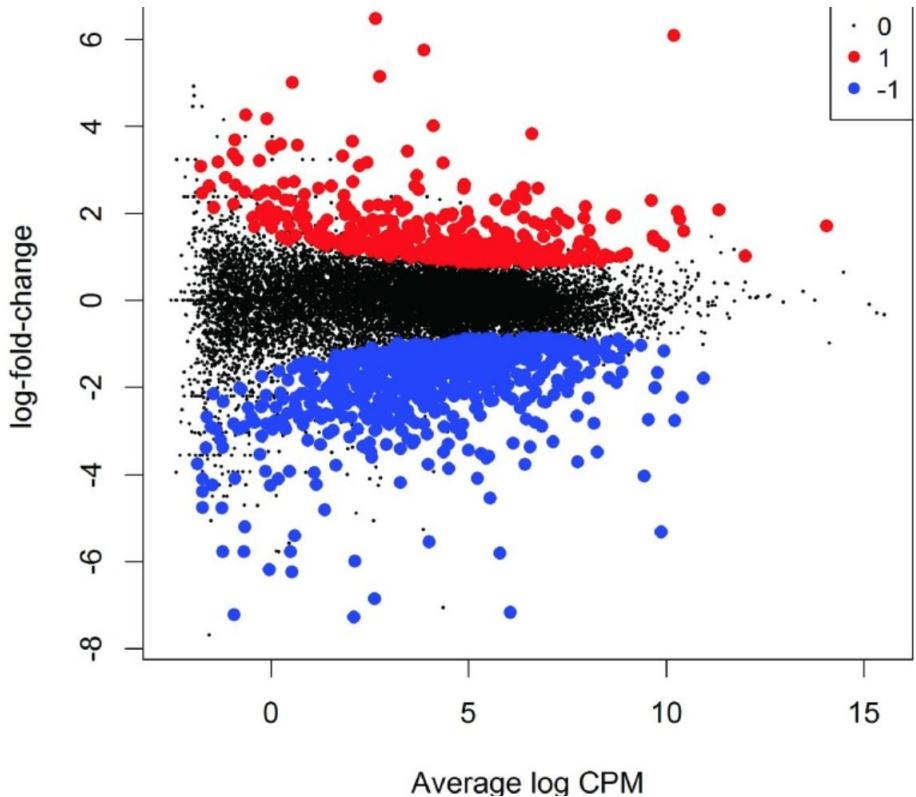
- Bioconductor package *edgeR* utilizes a *theoretical model* that captures some of the *known* processes leading to *noise in counts* data. (*null model*)
- If the probability is very low (e.g.,  $p < 0.05$ ), infer that something may be happening that we did not account for in the null model. (e.g., biological processes in L cells for milk production) - biological signal

## Scaling factors to minimize the log-fold changes between the samples for most genes

- *Reference sample*: have the closest average expressions to the mean of all samples
- *Test samples*: other samples

Generate a gene set removing most/lowest expressed genes ([avg read counts](#)) and genes with highest/lowest log ratios ([differences in expression](#))

# MD plot



Library size-adjusted log-fold change between two libraries (the difference) vs the average log-expression across those libraries (the mean)

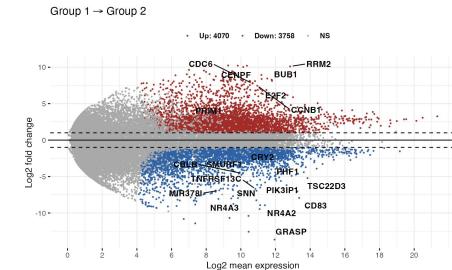
Log-fold change and average abundance of each gene

# Normalization with edgeR: Trimmed Mean of M-values

MA (mean-difference) plot

1. Choose a reference sample.
2. Compute the M and A values for all genes.

M=log FC between ref and test; A=avg count gene between ref and test



3. Filter genes that fall in the tails of M and A distributions.
4. Estimate variance of M values.
5. Estimate TMM --- the weighted average of trimmed M-values.
6. Size factor is  $2^{TMM}$ .
7. Adjust such that these multiply to 1.

TMM is implemented in edgeR and performs better for between-samples comparisons, when comparing the samples from different tissues or genotypes.

lib.size	norm.factors
23218026	1.2368993
21768136	1.2139485
24091588	1.1255640
22656713	1.0698261
21522033	1.0359212
20008326	1.0872153
20384562	1.3684449
21698793	1.3653200
22235847	1.0047431
21982745	0.9232822
24719697	0.5291015
24652963	0.5354877

`y <- calcNormFactors(y)`

# Other approaches to normalization

## Relative Log Expression (RLE) approach by Anders and Huber (2010)

- Reference: geometric mean of all samples
- Normalization factor: median ratio of each sample to the reference
- RLE and TMM give similar robust results with real and simulated data
- *R package DESeq - [tutorial](#)*

## Upper quartile normalization by Bullard et al (2010)

- Normalization factor: 75% quantile of the counts for each sample
- Not recommended in general - implies a similarity between RNA molecular repertoires expressed

## Control genes (housekeeping genes, spike-in) to estimate technical noise ([RUVSeq](#) – 2014 Remove Unwanted Variation)

- powerful when a large set of control genes can be identified

# Main normalization approaches summary

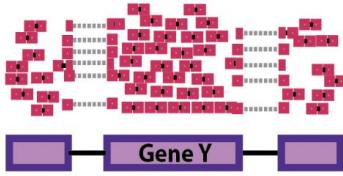
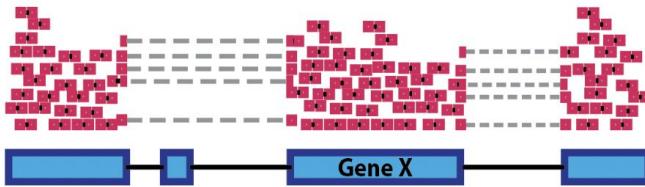
Normalization method	Description	Accounted factors	Recommendations for use
DESeq2's <b>median of ratios</b> [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for <b>DE analysis; NOT for within sample comparisons</b>
EdgeR's <b>trimmed mean of M values (TMM)</b> [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition	gene count comparisons between samples and for <b>DE analysis; NOT for within sample comparisons</b>

Why DESeq and TMM are not suitable for within sample comparison?

# Scenario 2

Which gene is more expressed in Sample A? Gene X or Gene Y?

Sample A Reads



What is the difference between Gene X and Gene Y?

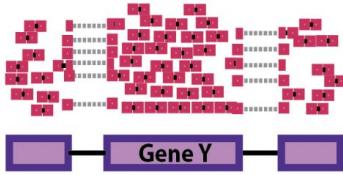
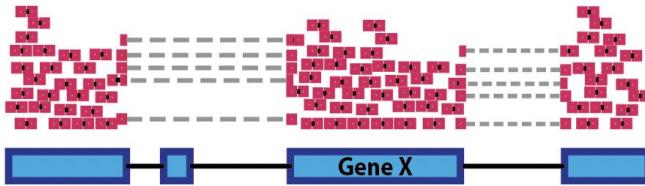


**Break (5 min)**

# Scenario 2

Which gene is more expressed in Sample A? Gene X or Gene Y?

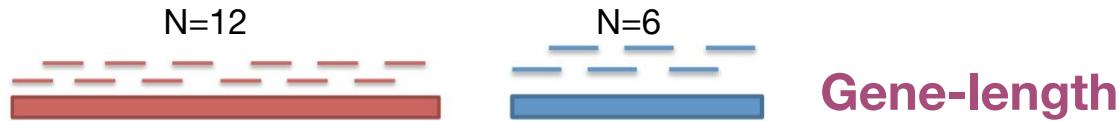
Sample A Reads



What is the difference between Gene X and Gene Y?

# Source of bias between genes

At the same expression level, a long gene will have more reads than a shorter gene



# Identify and correct technical biases removing the least possible biological signal

- Library size or sequence depth (number of mapped reads)
- Gene length

Commonly used normalization method that includes sequence depth and gene length correction:

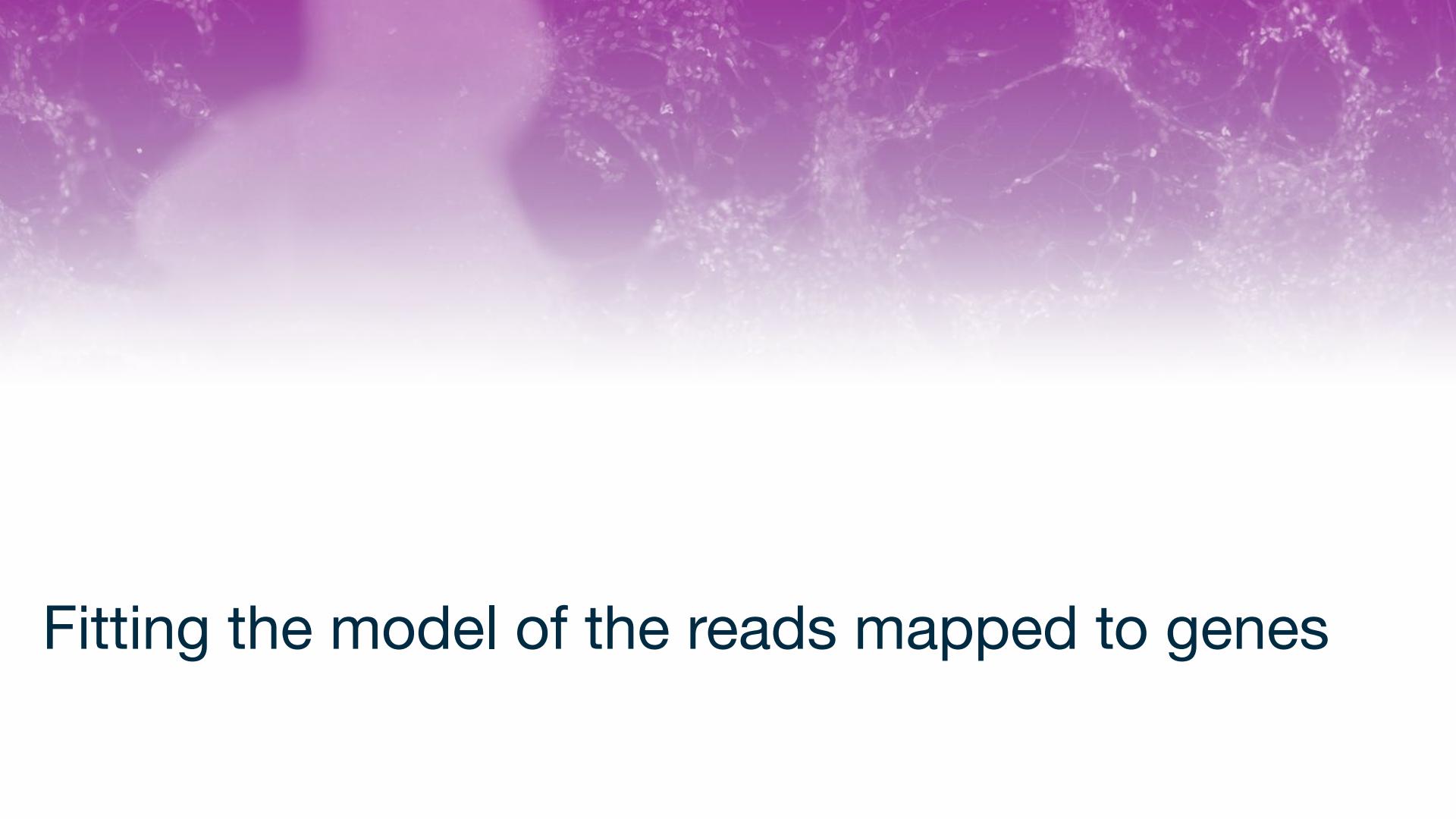
- RPKM /FPKM (Reads/Fragments Per Kilobase per Million)
- TPM (Transcripts Per kilobase Million)

## Poll 2

We want to detect whether Gene X is differentially expressed between Lactating and pregnant mice.

What are the sources of bias?

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



Fitting the model of the reads mapped to genes

# How to model the normalized RNA-seq read counts

Model the read mapped to a given gene and estimate the differences between the two groups/conditions.

$Y \sim \text{condition (B.Lactating vs B.Pregnant mice)}$

Total number of reads for a sample ~ millions

Counts per gene ~ tens /hundreds /thousands.

The chance of a *given read* to be mapped to any *specific gene* is rather small.

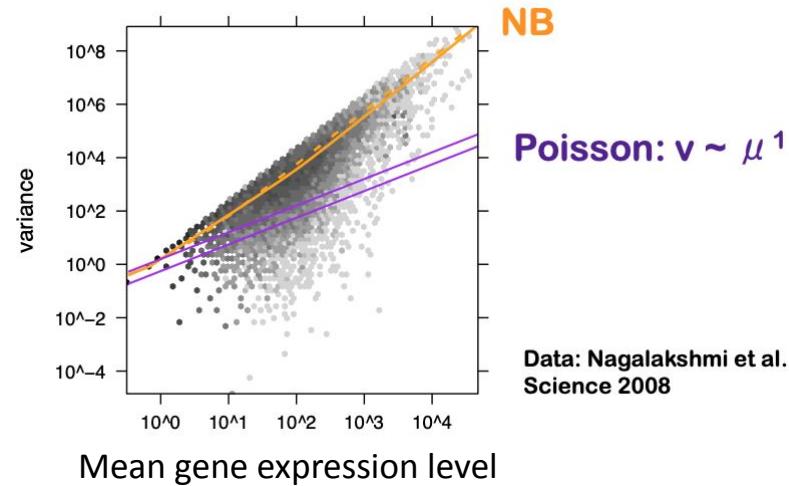
*Discrete events* sampled out of a large pool with low probability are usually modeled with Poisson distribution

# Modeling the read counts mapped to a gene - discrete and overdispersed

The number of reads mapped to a gene was first modeled using a **Poisson distribution** (Marioni *et al.* (2008))

**Assumption:** assumes that mean and variance are the same

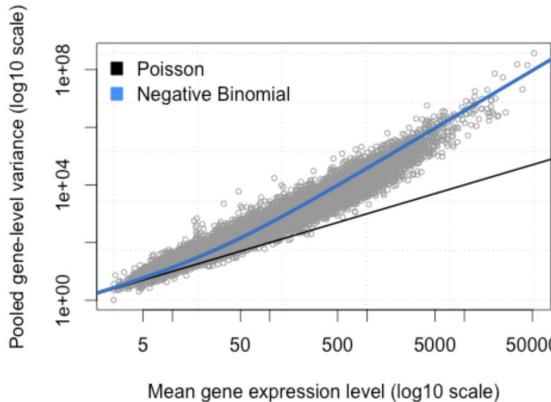
BUT the variance grows faster than the mean in RNAseq data.



Why?

# Overdispersion of read counts between samples

- > counts from biological replicates vary
- > variance is exceeding the mean for highly expressed genes
- > underestimation of the biological variance increased the probability to falsely declare a gene DE when it is non-DE (type I error rate)



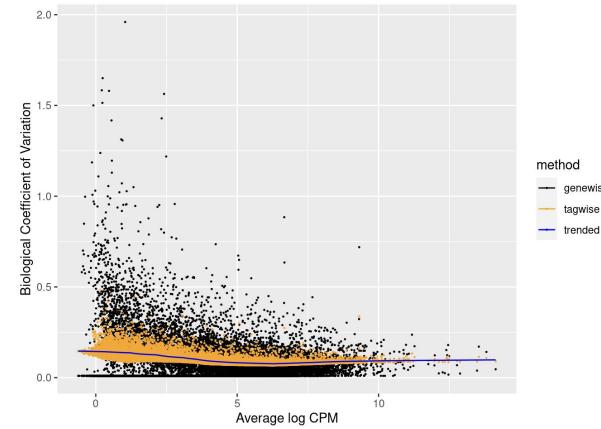
Negative binomial (NB) distribution  
-> alternative to model the read counts for each gene in each sample

The variance is always larger than the mean for the negative binomial  $\Rightarrow$  suitable for RNA-seq data

# Account for the relative deviation of the gene expression distribution relative to its mean

Dispersion  $\phi$  accounts for the variability between biological replicates

- Common dispersion: a global dispersion estimate averaged across the genes – not enough
- Trended dispersion: dispersion of a gene is predicted from its abundance – similar abundant genes
- Tagwise dispersion: measure of the degree of consistent inter-library variation for that tag - EB shrinkage to a common (trended) dispersion



The tag-wise dispersions (orange dots) are the result of shrinking the gene-wise dispersion (black dots) to the trend (blue line)

plotBCV (biological coeff variation)

Empirical Bayes estimates need to be controlled for the possibility of outlier genes with exceptionally large or small individual dispersions (robust=TRUE)

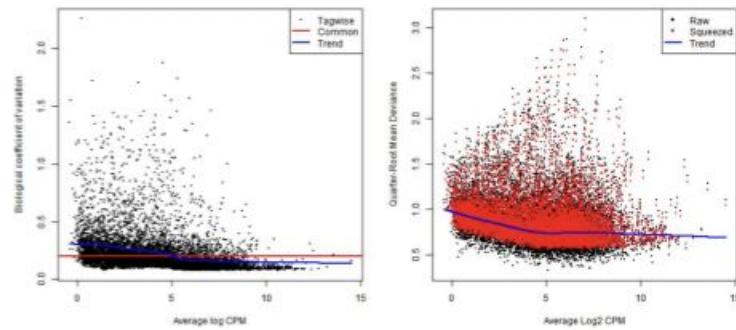
`y <- estimateDisp(y, design)`

# Fitting the model : *glmQLFit*

- NB dispersions - higher for genes with very low counts - decrease smoothly with abundance and asymptotically to a constant value for genes with larger counts.

- Extended NB model to account for gene-specific variability from both biological and technical sources (quasi-likelihood)

- 1) NB dispersion trend is used to describe the overall biological variability across all genes (fit GLM)
- 1) For each gene-specific variability above and below the overall level (deviance) is picked up by the QL dispersion



# edgeR fitting models

QL for bulk RNA-seq:

- + stricter error rate control (more rigorous dispersion and uncertainty)
- + speed improvement compared to other quasi-methods
- + for multiple treatment factors and with small # of biological replicates
- + relative changes in expression levels between conditions (not absolute)

Limma package for large scale datasets – high overlap across methods

# Get the DE genes - *glmQLFTest* and *glmTreat*

- Identifies differential expression based on statistical significance regardless of how small the difference might be -> 5000 DE genes between condition and control groups
- Interested only in genes with large expression changes -> subset of genes more biologically meaningful.
- Modify the statistical test to evaluate variability as well as the magnitude of change of expression values -> expression changes greater than a specified threshold
- Not equivalent to a simple fold change cutoff : “the fold-change below which we are definitely not interested in the gene”
- The total number of DE genes identified at an FDR of 5% can be shown with *decideTest()* – set cutoff

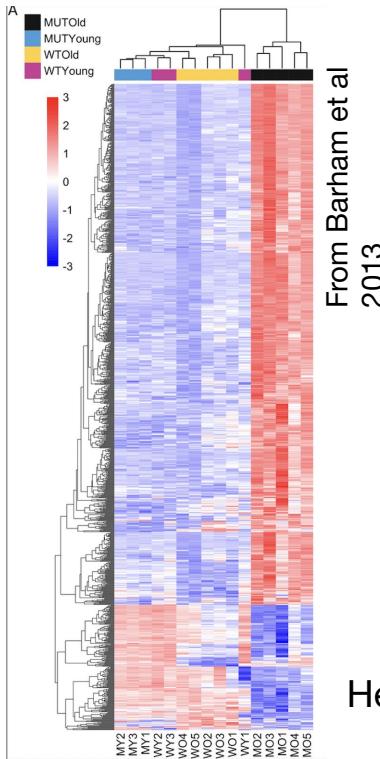
# Get the DE genes - *glmQLFTest* and *glmTreat*

	genes	logFC	logCPM	F	PValue	FDR
PCDHA10	PCDHA10	-3.602	5.676	499.9	8.164e-11	1.354e-06
CHGA	CHGA	2.923	5.976	185.4	1.972e-08	0.0001635
ARRB1	ARRB1	-3.914	5.015	158.3	4.627e-08	0.0002019
TSSC2	TSSC2	3.175	3.301	156.8	4.869e-08	0.0002019

## Complicated contrasts - *makeContrasts()*

- between lactating and pregnant mice is the same for basal cells as it is for luminal cells
- the interaction effect between mouse status and cell type

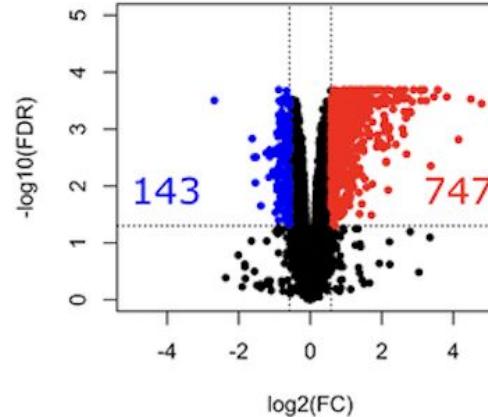
# Goal of DEG analysis



Volcano plot

B

MY vs. MO

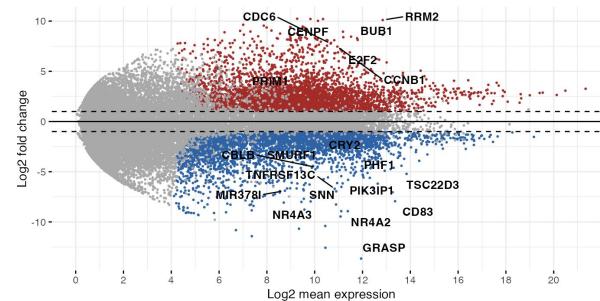


From Barham et al  
2013

MA (mean-difference) plot

Group 1 → Group 2

Up: 4070 · Down: 3758 · NS



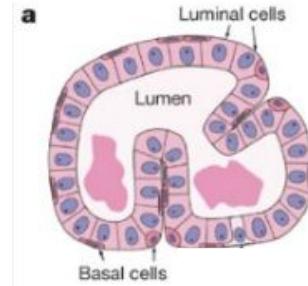
Identify genes (and molecular pathways) that are differentially expressed (DE) between two or more biological conditions

# Demo

# Dataset

Transcriptome analysis of luminal and basal cell subpopulations in the lactating versus pregnant mammary gland

- GEO (gene expression omnibus) accession: **GSE60450**
- Tissue of origin: Mammary glands of mouse
- Cell types: Basal stem-cell enriched cells (B) and committed luminal cells (L)
- Biological conditions: Virgin, Lactating (2 day) and Pregnant (18.5 day)
- # of groups: 2 cell types (B/L) × 3 conditions (V/L/P) = 6 groups
- # of replicates: 2 of each group
- Illumina Hiseq sequencer - about 30 million 100bp single-end reads for each sample.



# Files for the hands-on session

GEO	SRA	CellType	Status
MCL1.DG	GSM1480297	SRR1552450	B virgin
MCL1.DH	GSM1480298	SRR1552451	B virgin
MCL1.DI	GSM1480299	SRR1552452	B pregnant
MCL1.DJ	GSM1480300	SRR1552453	B pregnant
MCL1.DK	GSM1480301	SRR1552454	B lactating
MCL1.DL	GSM1480302	SRR1552455	B lactating
MCL1.LA	GSM1480291	SRR1552444	L virgin

- **targets.txt**

**Phenofile**

- **GSE60450\_Lactation-GenewiseCounts.txt.gz**

Counts for each sample for each gene (Entrez Gene Identifiers)

	Length	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA	MCL1.LB
497097	3634	438	300	65	237	354	287	0	0
100503874	3259	1	0	1	1	0	4	0	0
100038431	1634	0	0	0	0	0	0	0	0
19888	9747	1	1	0	0	0	0	10	3
20671	3130	106	182	82	105	43	82	16	25
27395	4203	309	234	337	300	290	270	560	464

# Hands-on session

- Load and reformat the data
- Exploratory visualization : MA plot
- Create DGElist object and retrieve gene symbols
- Filter genes with inadequate information
- Normalization
- Exploratory visualization : MDS and PCA plots
- Define and fit a model
- Hypothesis testing (four example hypotheses)
- Save results as a table and explore in Excel

# In summary

- Raw counts are not comparable across samples/genes within a sample
- Several normalization methods: some more suitable for DEG
- Estimate the dispersion, visualize the technical variability
- Fit the model: counts variance exceeding the mean
- Make all the comparisons that you wish using complicated contrasts
- Visualize the DEG and get a list for pathway analysis

# What if

- Q1. Experimental design with repeated measures - alternative methods
- Q2. Batch effects in bulk

Real data might need additional analyses choices that need experience.

Consult with the [Gladstone Bioinformatics core](#) for such scenarios and data.

Winter workshops schedule - [Data Science Training Program](#)

# Upcoming workshop

## Introduction to Pathway Analysis

October 16

## Statistics of Enrichment Analysis Methods

October 20–21

## Single Cell RNA-Seq Analysis

October 27–28

**Your feedback is important to us!**

Please take the survey ~3 min:

<https://www.surveymonkey.com/r/F75J6VZ>

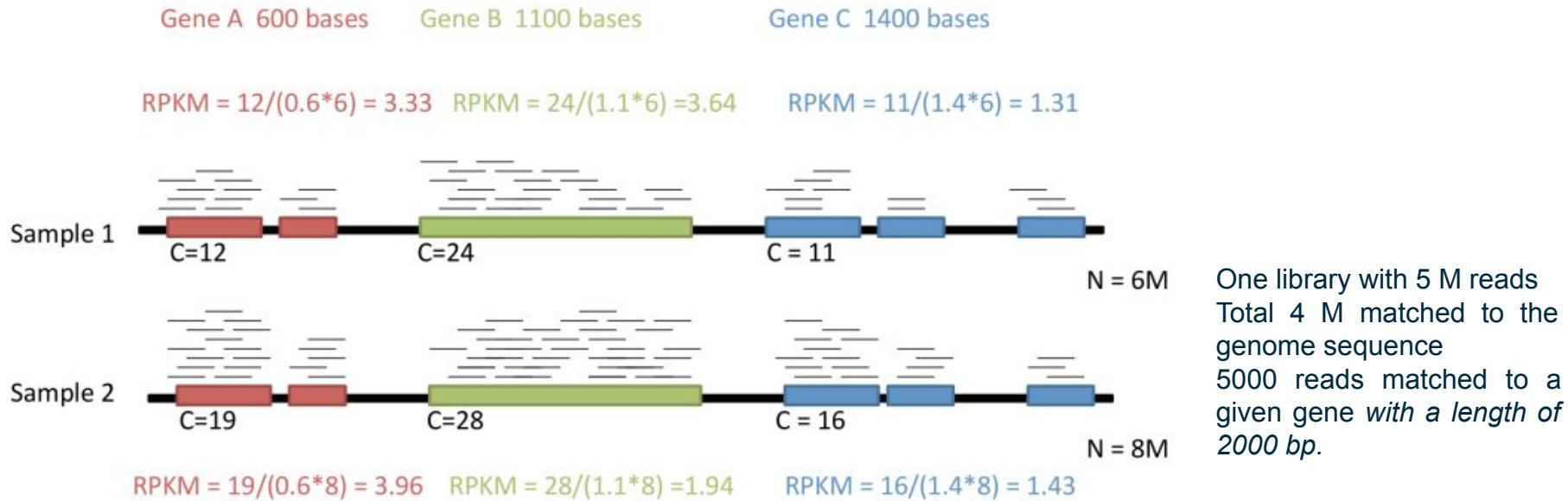


# GLADSTONE INSTITUTES

# **Appendix**

# RPKM: seq depth and gene length bias

Reads/Fragments Per Kilobase per Million



$$\text{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$

$$\text{RPKM} = \frac{5000 \times 10^3 \times 10^6}{4 \times 10^6 \times 2000} = 625$$

RPKM and FPKM -> *the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.*

# TPM: seq depth and gene length bias

TPM: Normalize for gene length first to get the Reads Per Kilobase, sum up all the RPK in a sample (across all genes), and then normalize for sequencing depth

$$\text{TPM} = 10^6 * \frac{\text{reads mapped to transcript / transcript length}}{\text{Sum(reads mapped to transcript / transcript length)}}$$

Represent the relative abundance of a transcript among a population of sequenced transcripts

*The sum of all TPMs in each samples are the same* -> to compare the proportion of reads that mapped to a gene in each sample.

Same denominator -> comparable - NOT for RPKM (different denominator)

Sample1 - gene1 TPM=2.5

Sample2 - gene1 TPM=2.5

# Main normalization approaches summary

Normalization method	Description	Accounted factors	Recommendations for use
<b>TPM</b> (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; <b>NOT for DE analysis</b>
<b>RPKM/FPKM</b> (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; <b>NOT for between sample comparisons or DE analysis</b>

Still not enough to perform DEG analysis



# GLADSTONE INSTITUTES