Materials for this workshop

- Concepts: this presentation
- Hands-on session:
  - handson.R
  - targets.txt
  - DE_resutls.txt
  - GSE60450_Lactation-GenewiseCounts.txt.gz

Please install the following library in Rstudio

```
Install.packages(magrittr)
Install.packages(edgeR)
Install.packages(org.Mm.eg.db)
Install.packages(tidyverse)
Install.packages(ggplot2)
Install.packages(statmod)
```

Intermediate RNA-seq data analysis

Gladstone Institutes

Michela Traglia and Min-Gyoung Shin
Bioinformatics Core, GIDB

22 May 2023

Assumed background

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

Introductions

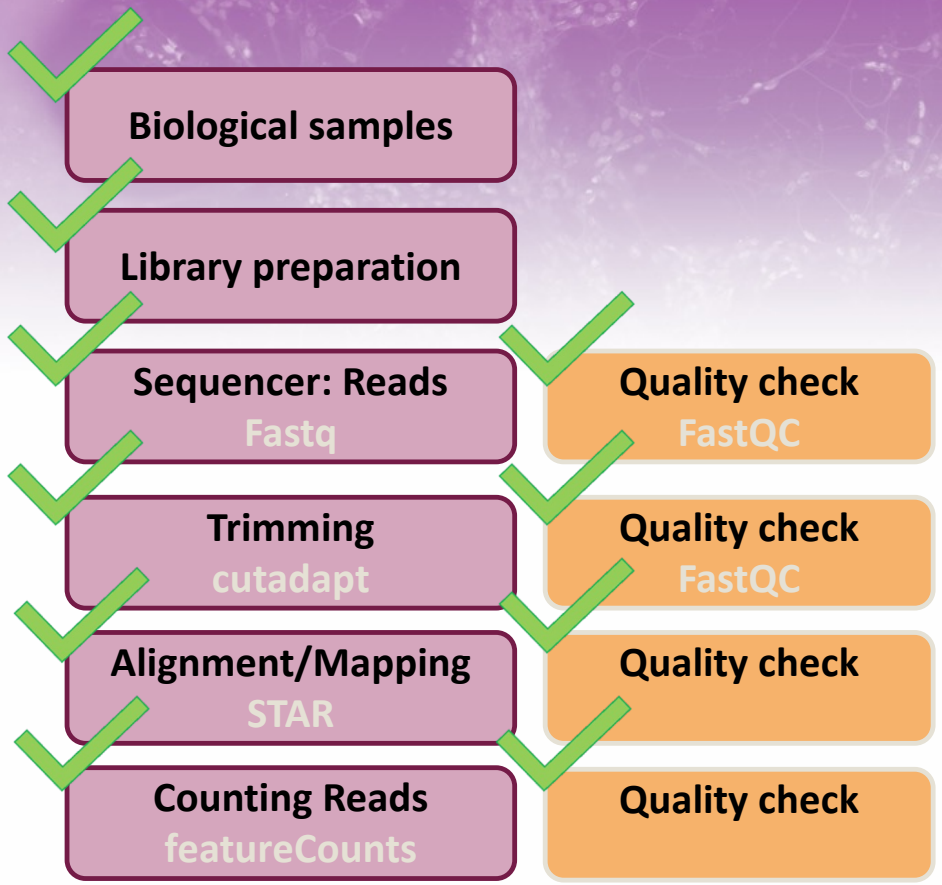
Min-Gyoung Shin

Bioinformatician II

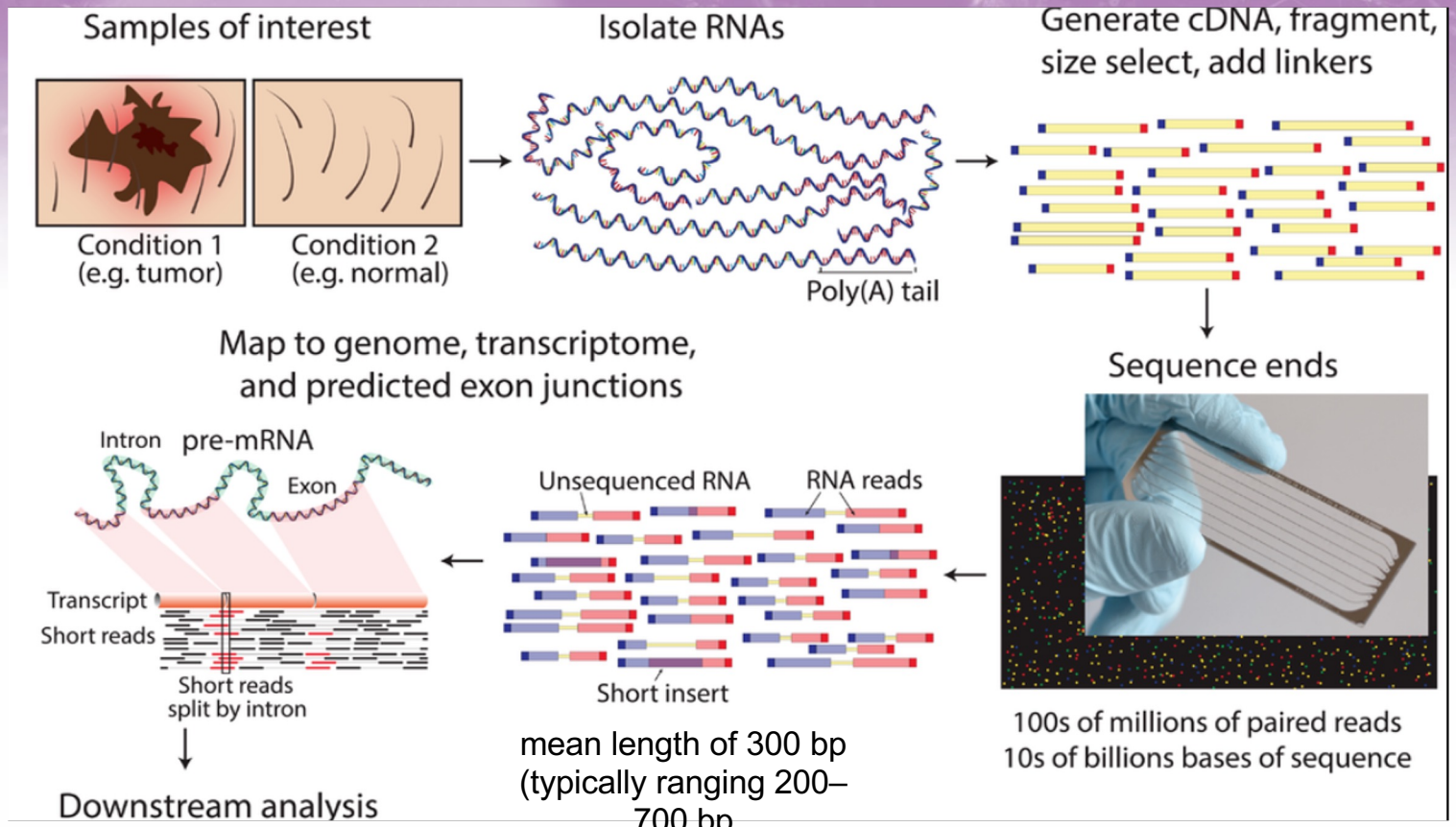
Michela Traglia

Statistician III

RNA-seq - analysis workflow



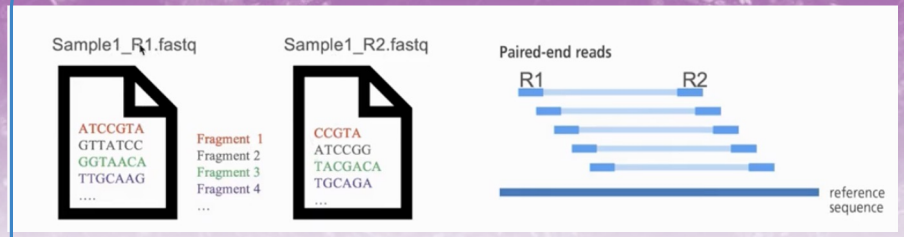
Typical RNA-seq protocol



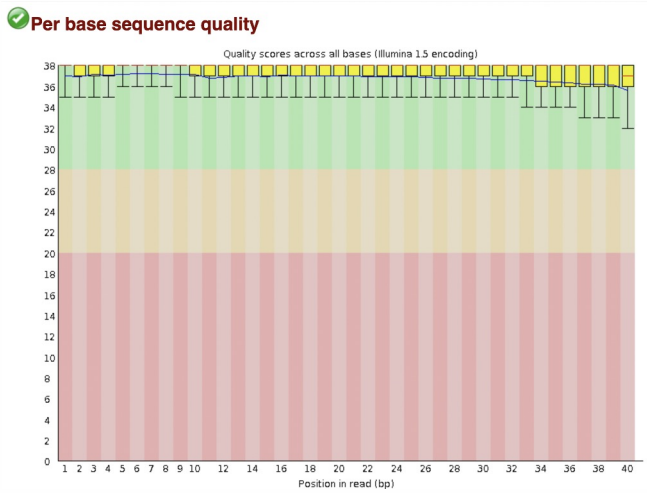
https://www.wikiwand.com/en/RNA-Seq

Bioinformatic pipeline - summary

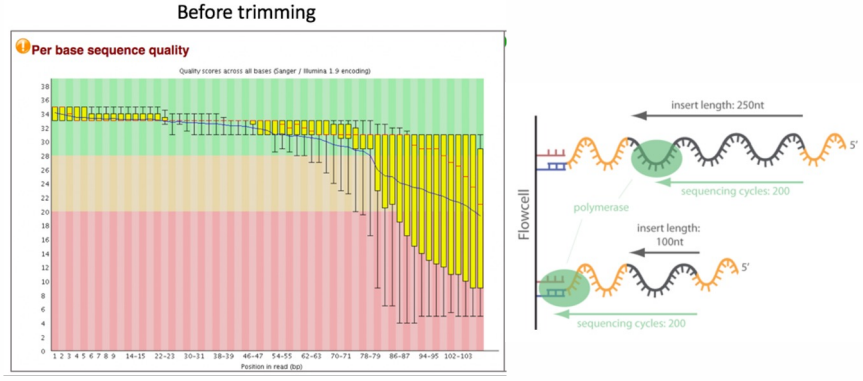
FASTQ format



FASTQ file

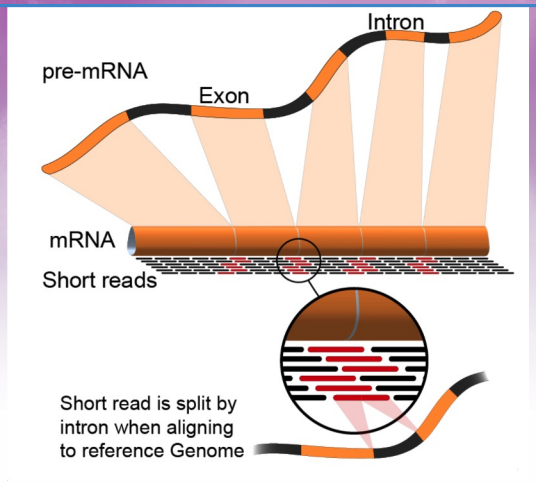


FastQC - good experiment

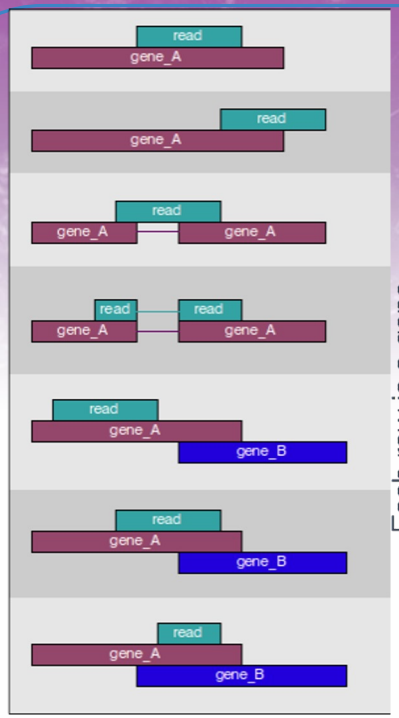


FastQC - bad experiment needs trimming - cutadapt

Bioinformatic pipeline - summary



Alignment and pseudo-alignment tools



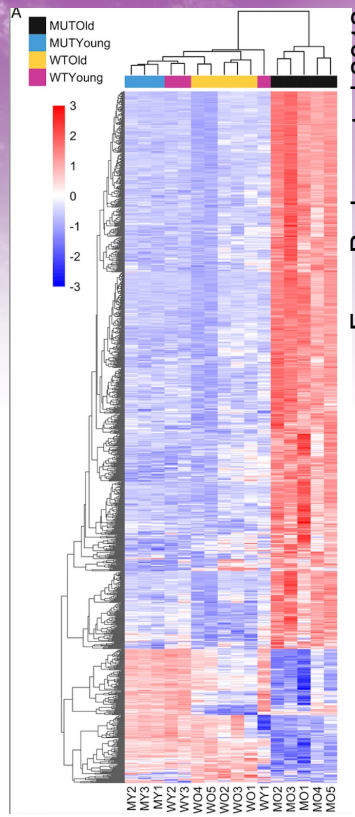
Each row is a gene

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

Feature counts

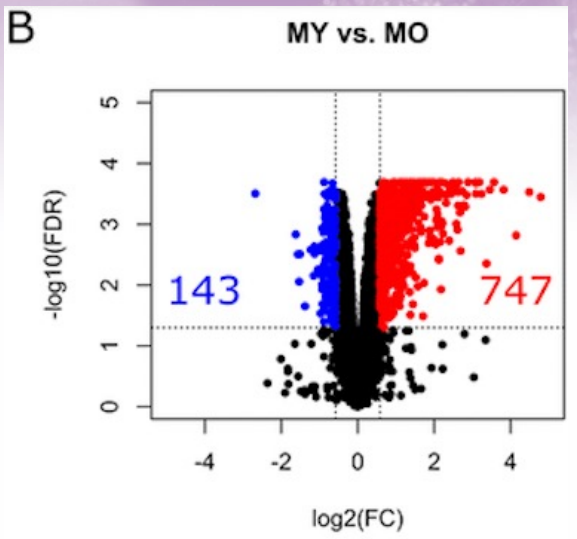
Goal of DEG analysis



From Barham et al 2013

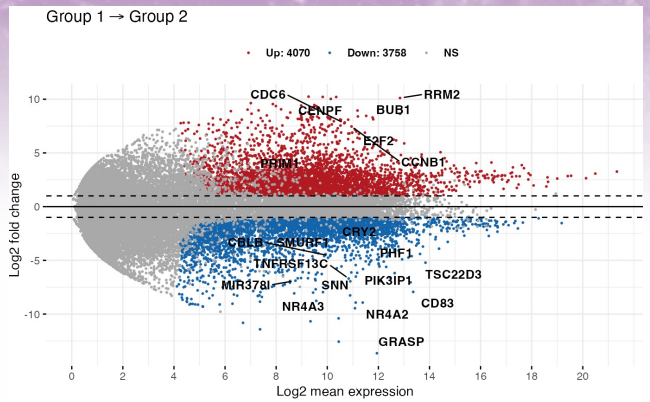
Heatmap

Volcano plot



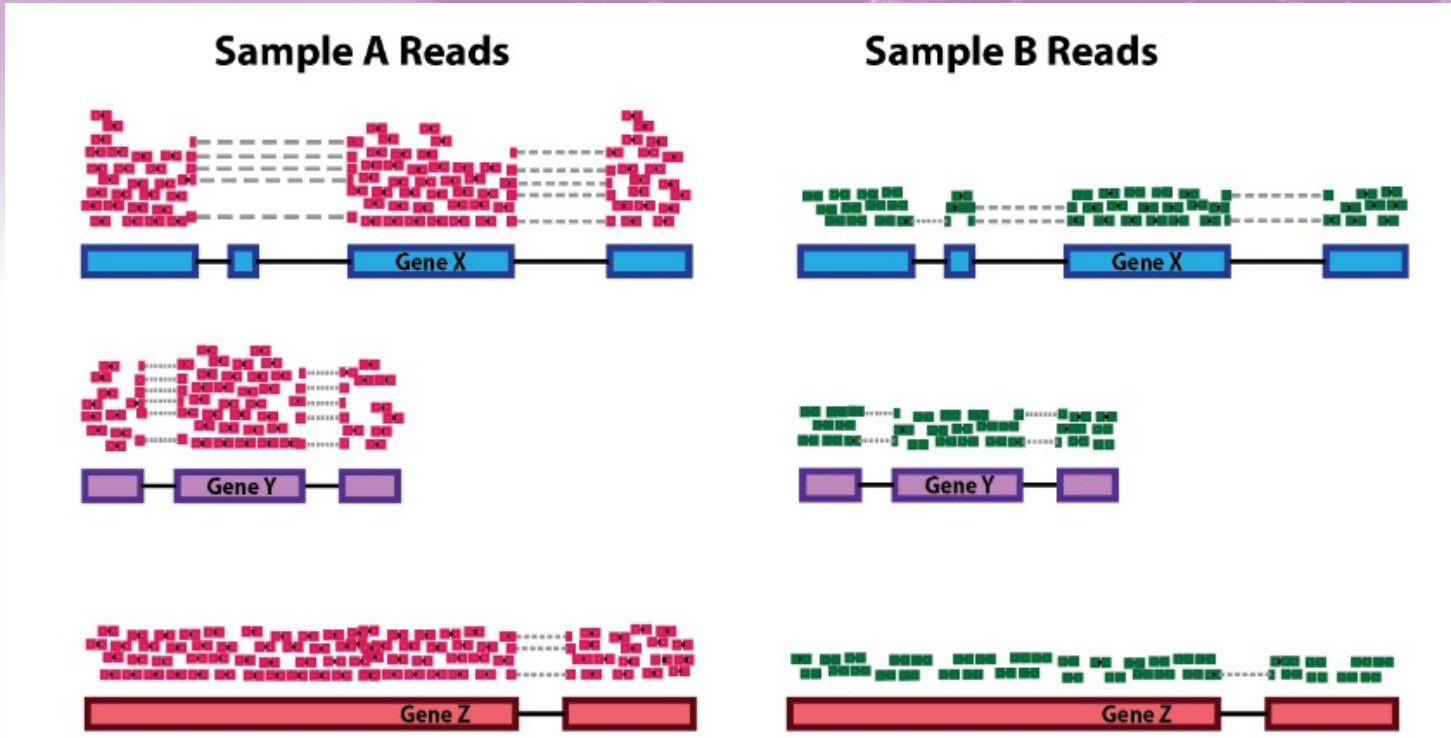
From Barham et al 2013

MA (mean-difference) plot



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

Many steps to calculate the Differentially Expressed Genes (DEG) between sample A and B



We need to make the counts **comparable** across samples

Workshop outline

- Intro to a real experiment – Demo
- Approach for Differentially Expressed Gene analysis: edgeR
- Filtering genes
- Normalization
- Exploratory visualization: MDS – PCA
- Fit the model for DEG
- Compare groups and visualize the DEG

Reference for the workshop

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

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[1,2], Aaron T. L. Lun[3], Gordon K. Smyth[1,4]

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia
[2]Department of Medical Biology, The University of Melbourne, Victoria, 3010, Australia
[3]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK
[4]Department of Mathematics and Statistics, The University of Melbourne, Victoria, 3010, Australia

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) x 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.

https://www.ncbi.nlm.nih.gov/geo/

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

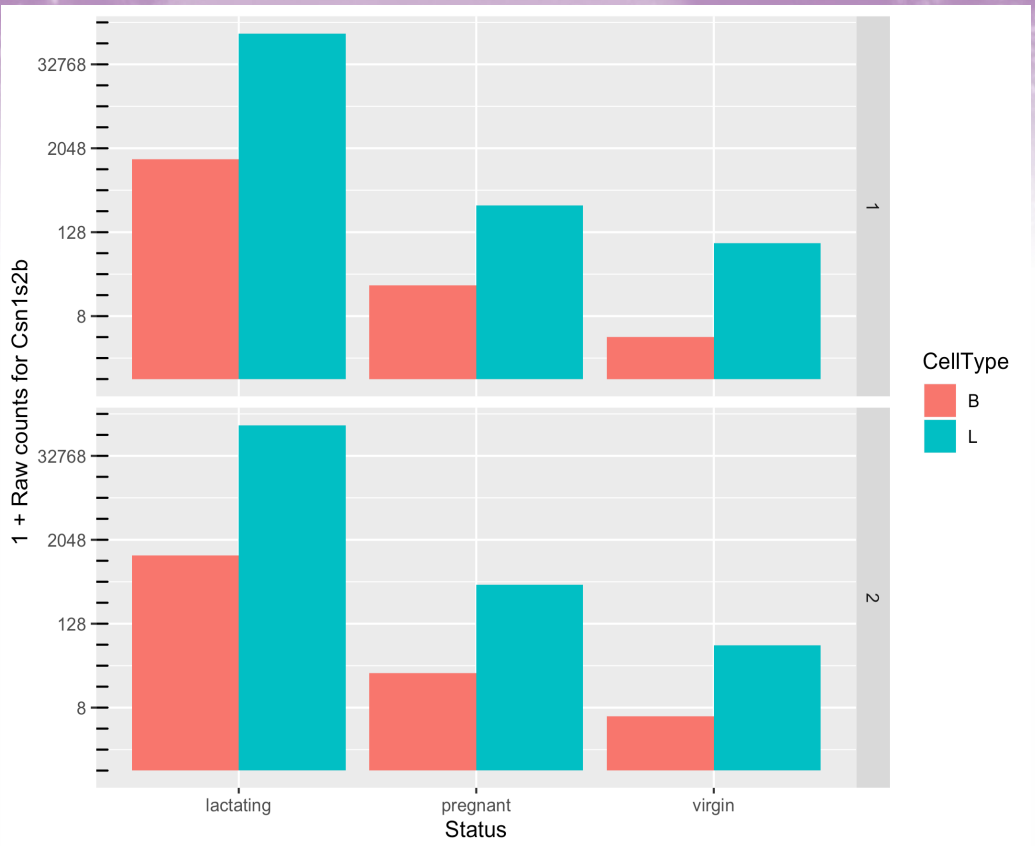
  Phenofile

- 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

Goal: Identify a set of genes that are differentially expressed across samples

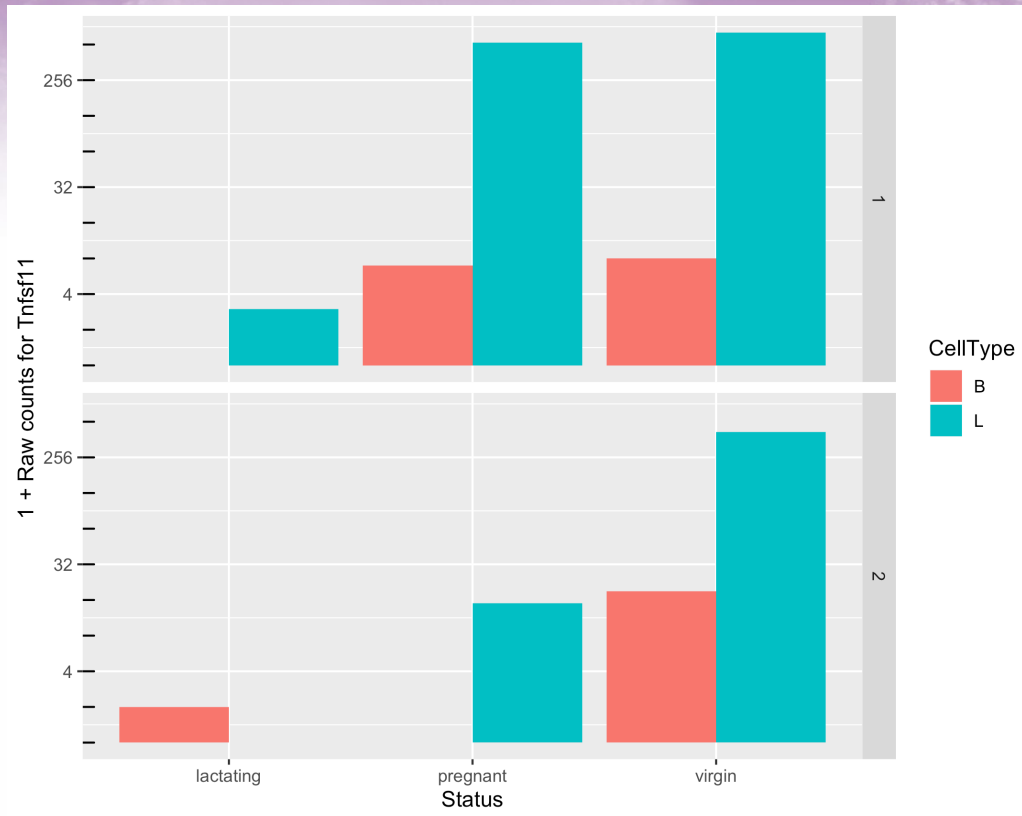


Which comparisons are we interested in?

Example:

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

Potential issues



How do we ensure high power for detection and high specificity when faced with noise?

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

Approach for DEG

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*)
- Assume that the data is generated according to this model.
- Given any observed level of difference in mean expression levels of a gene, compute the probability that the observation will result from the null model (p value)
- 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)

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.

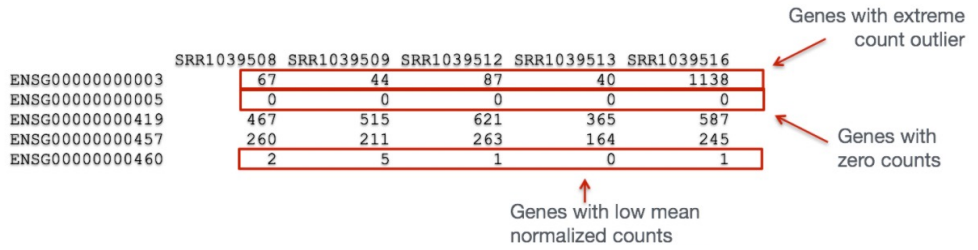
- The more genes we test, the more we must adjust.
- Reduce the number of tests by filtering out “uninteresting” genes.

“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



	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	67	44	87	40	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	2	5	1	0	1

Source of variability

Why we need to normalize the counts

Identify and correct technical biases removing the least possible biological signal

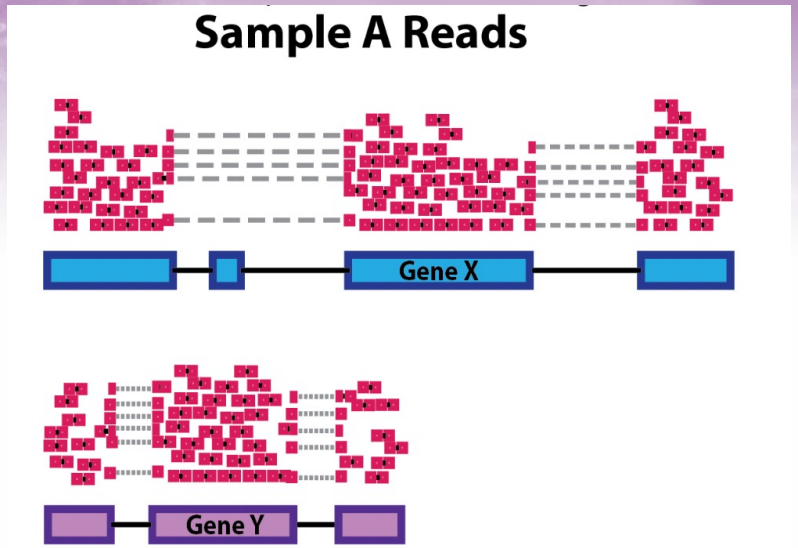
- library prep, sequencing technology, …

It is an essential step in the analysis of gene expression:

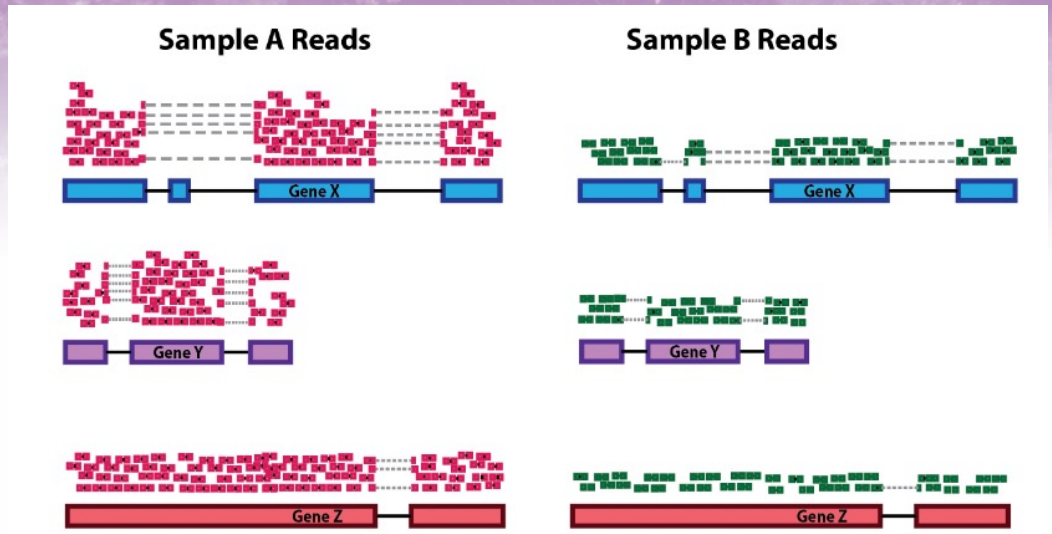
- to accurately estimate gene expressions in a sample
- to compare genes from different samples (differential analysis)

Why we need to normalize the counts

Within a sample



Between samples



Compare gene expressions across genes (features)

DEG: Compare gene expressions across samples

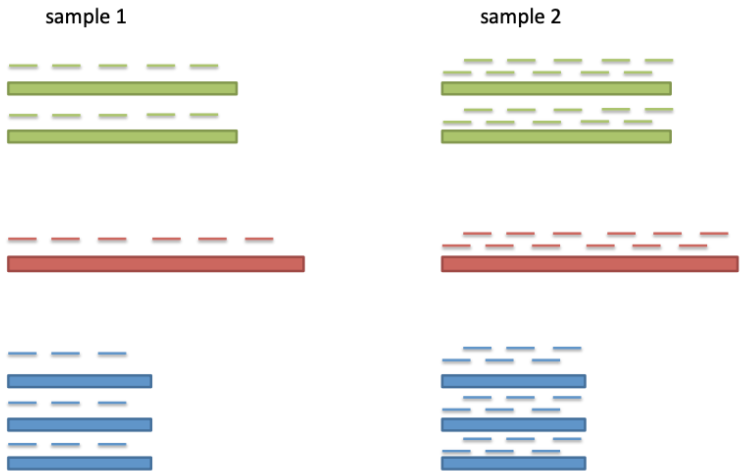
Source of technical variability to account for

- Gene length
- Library size or sequence depth (number of mapped reads)
- RNA sample composition
- Batch effects
- …

Gene-length and sequence depth

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

Higher sequencing depth, higher counts



Different sequence depths 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

Knowledge check 1

Is the gene length a technical bias for:

1. *Within same sample* gene expression quantification
2. *Between samples* differentially gene expression

Normalization methods

Several normalization methods to remove technical bias

- Various *gene expression units* such as RPM (or CPM), RPKM, FPKM, TPM, TMM, *DESeq*
- Measure of the abundance of gene or transcripts.
- Normalized expression units are necessary to remove technical biases
  - *Depth of sequencing* correction
  - *Depth of sequencing* and *gene length* correction
  - *RNA sample composition*
- To remove between samples batch effects

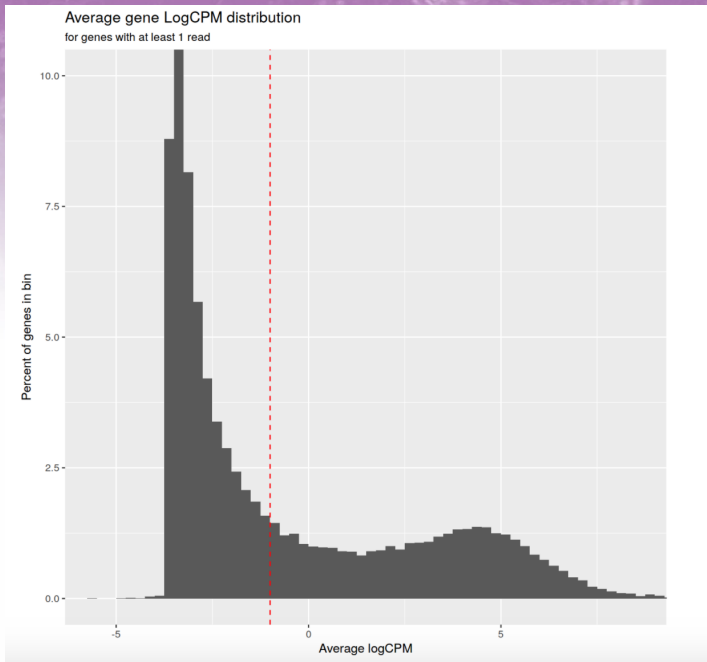
Sequence depth/Library size - Count Per Million mapped reads

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

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

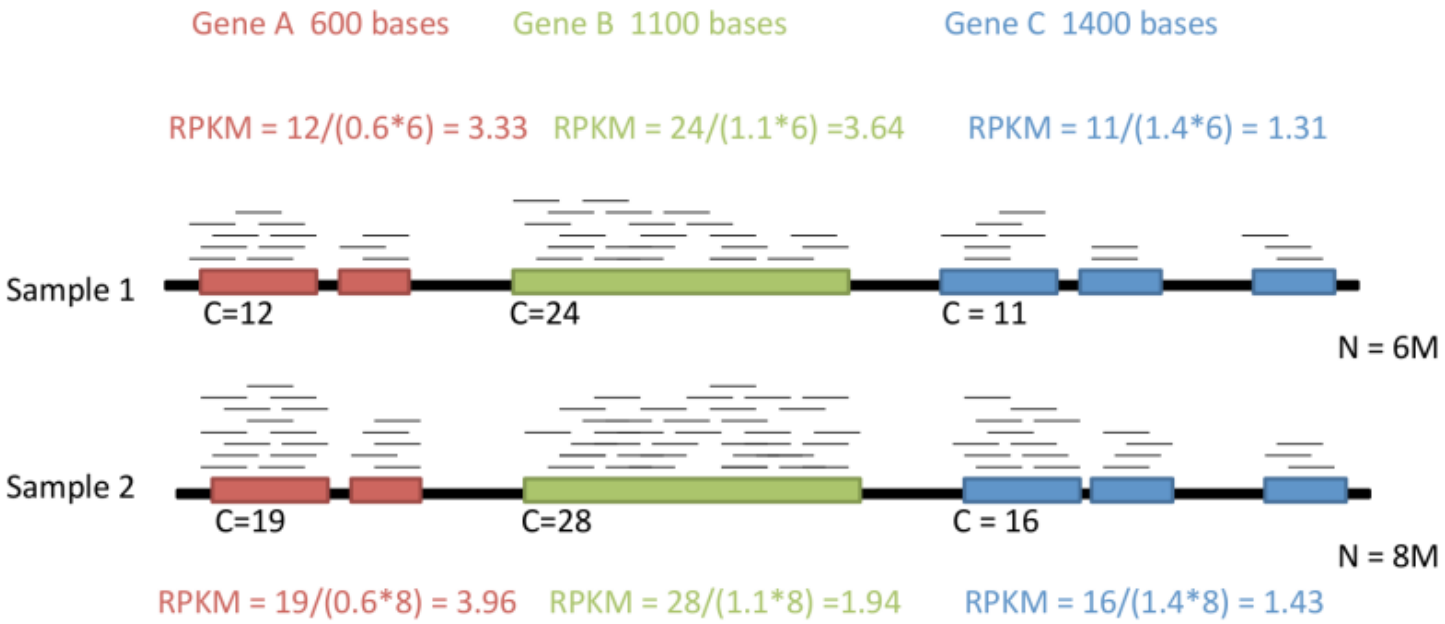
Common methods to normalize the counts considering gene length

Commonly used normalization method that includes gene length correction:

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

RPKM: seq depth and gene length bias

Reads/Fragments Per Kilobase per Million



One library with 5 M reads
Total 4 M matched to the genome sequence
5000 reads matched to a given gene *with a length of 2000 bp.*

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

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

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

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

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

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

TPM is equal to 10^6 (1 million) divided by the number of annotated transcripts in a given annotation so it is constant and proportional to the relative RNA molar concentration

Don’t use these methods for between condition/groups comparison!

When total RNA composition is similar across samples, these methods 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

Observed counts depend on total reads sequenced and RNA sample composition

Sample-to-sample variability in total RNA concentration

- Need to normalize for difference in total reads between samples.
  - Might be enough if total nucleic acid is the same in both samples.
  - Example: technical replicates

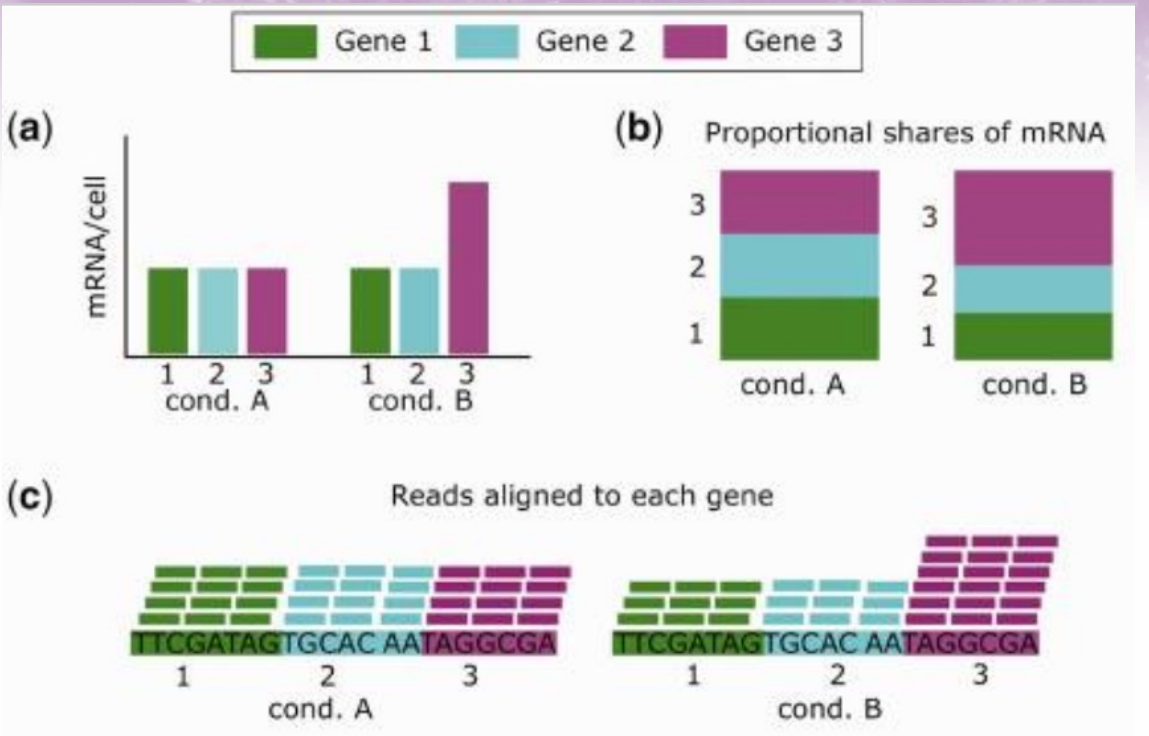
- Need to account for difference in RNA sample composition

Proportion of mRNA corresponding to a given gene may change across biological conditions

Total read counts for each sample is the same

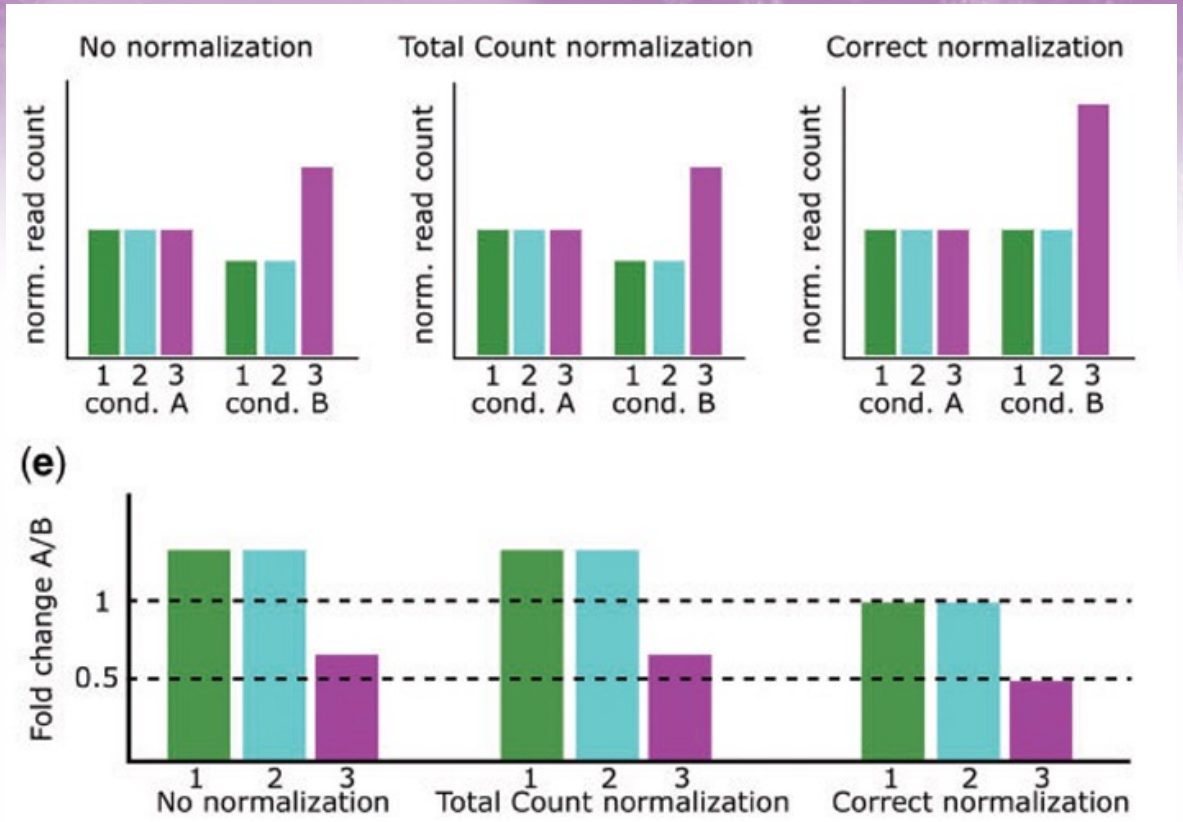
Few highly expressed genes make up a greater share of the total molecules (total library size) of samples B

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



doi: 10.1093/bib/bbx008

The goal is for differences in normalized read counts to represent differences in true expression



doi: 10.1093/bib/bbx008

Scaling factors

Normalize for RNA composition differences by scaling factor

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.

Between sample normalization

Normalize for RNA composition by a set of scaling factors that 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

Scaling factor: weighted mean of log ratios between the test and reference, from a gene set removing most/lowest expressed genes (avg read counts) and genes with highest/lowest log ratios (differences in expression)

Normalization with edgeR: Trimmed Mean of M-values

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.

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

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

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 results with real and simulated data
  - *R package DESeq*
- Upper quartile normalization by Bullard et al (2010)
  - Normalization factor: 75% quantile of the counts for each sample
  - Not recommended in general
- Control genes (housekeeping genes, spike-in) to estimate technical noise (RUVSeq - 2014)

Best practice to choose a normalization

An effective normalization should result in a stabilization of read counts across samples (eliminate composition biases between libraries)

- TC, RPKM, UQ - Adjustment of distributions, implies a similarity between RNA repertoires expressed
- DESeq, TMM - More robust ratio of counts using several samples, suppose that the majority of the genes are not DE
- RUVSeq - Powerful when a large set of control genes can be identified

Main normalization approaches summary

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** [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, RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**

https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/sample_level_QC.html

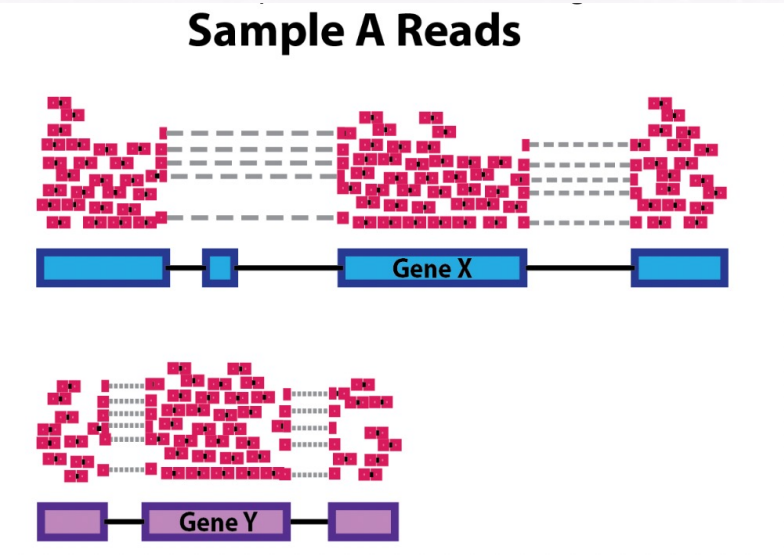
Break (10 min)

Knowledge check 2

Which normalization is suitable for within-sample gene expression quantification:

1. TPM
2. CPM
3. TMM

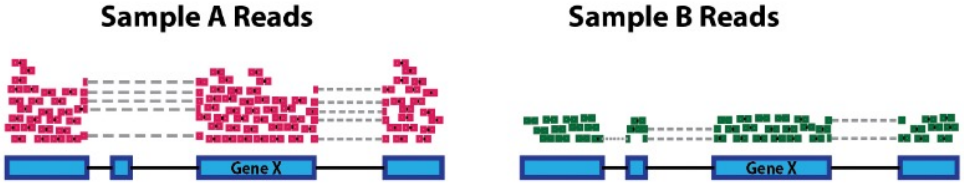
Within a sample



Knowledge check 3

Which normalization is suitable for comparing gene expression between disease samples vs control samples

1. RPKM
2. TMM
3. CPM

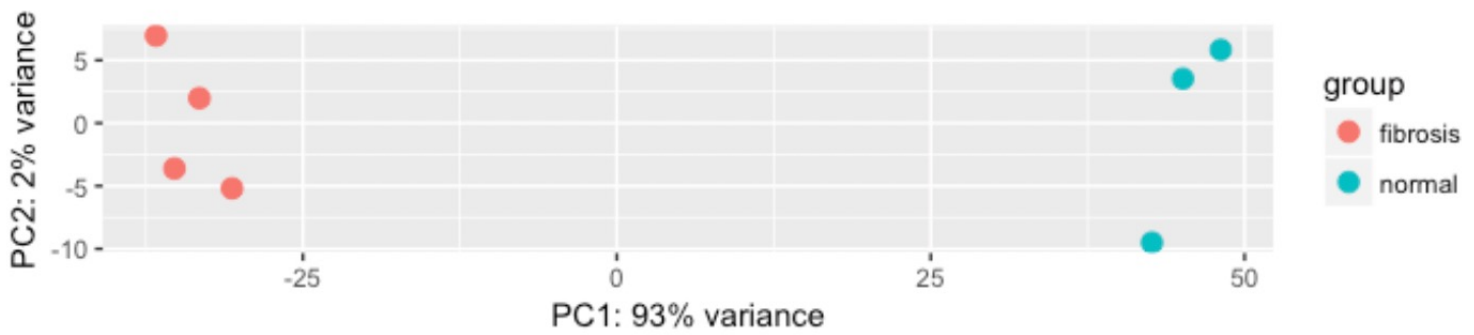


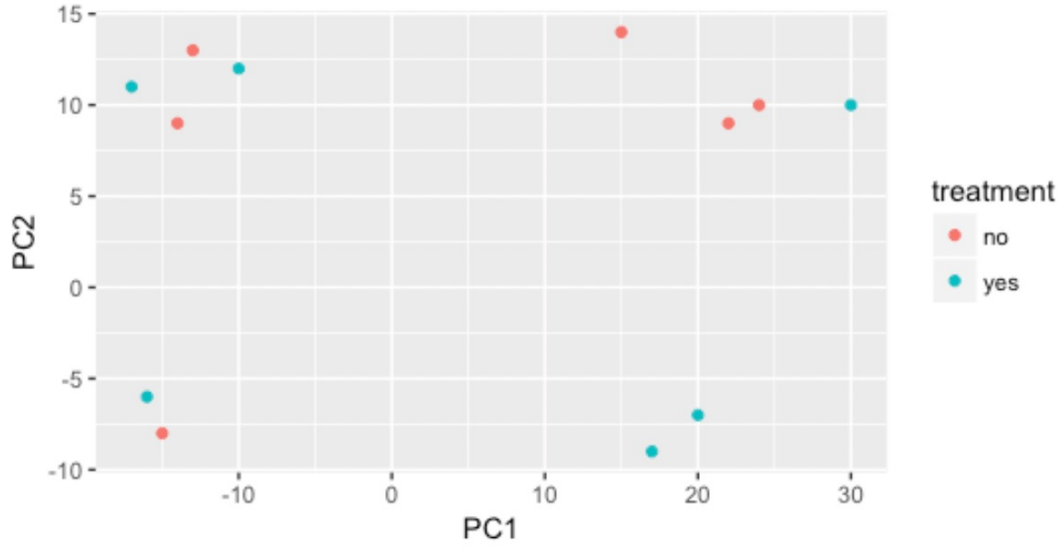
MDS and PCA plots

Assessing overall similarity across samples

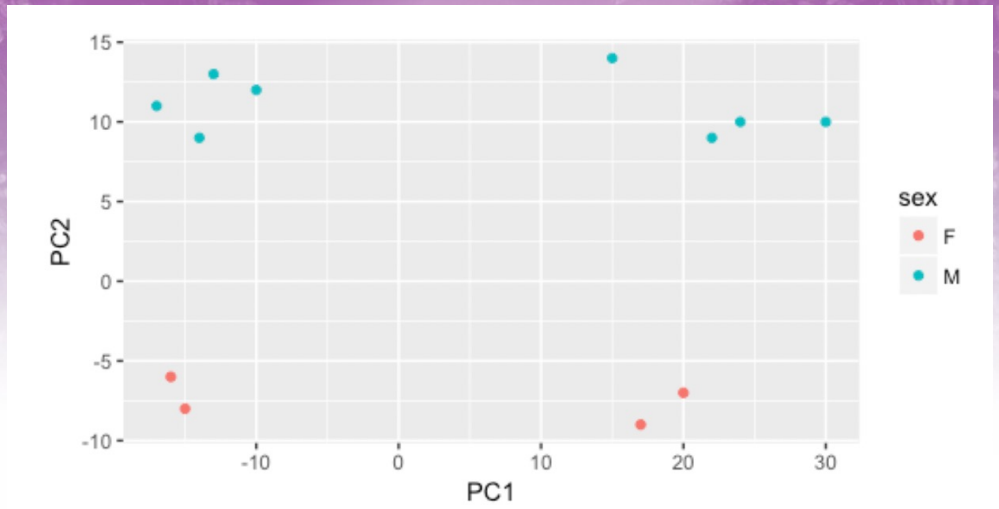
- Which samples are similar to each other, which are different?
- Does this fit to the expectation from the experiment's design?
- What are the major sources of variation in the dataset?

Ideal vs real experiment



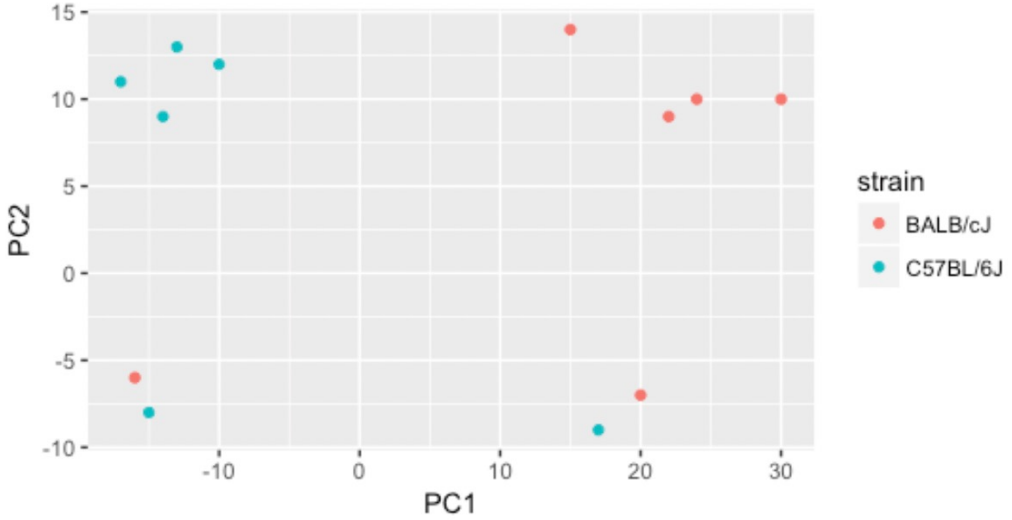


Identifying the source of variability



Variation in PC2

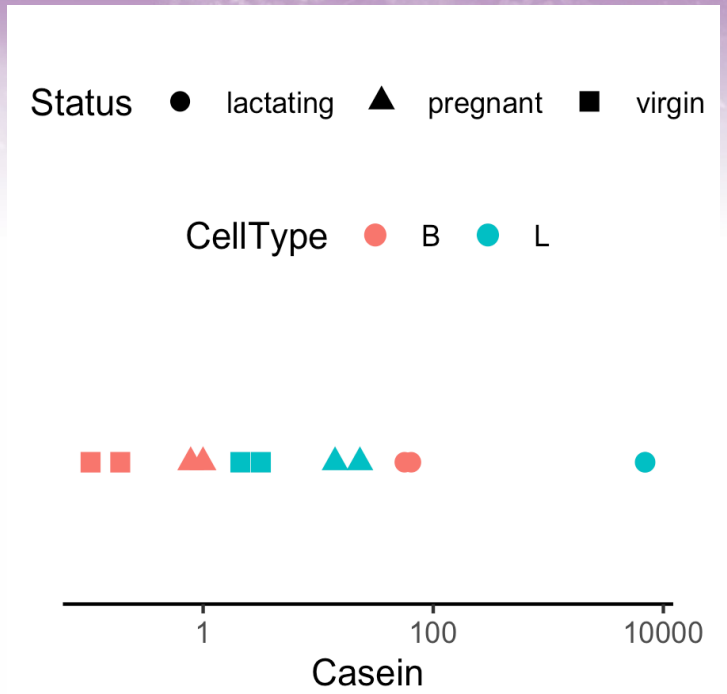
Variation in PC1



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?

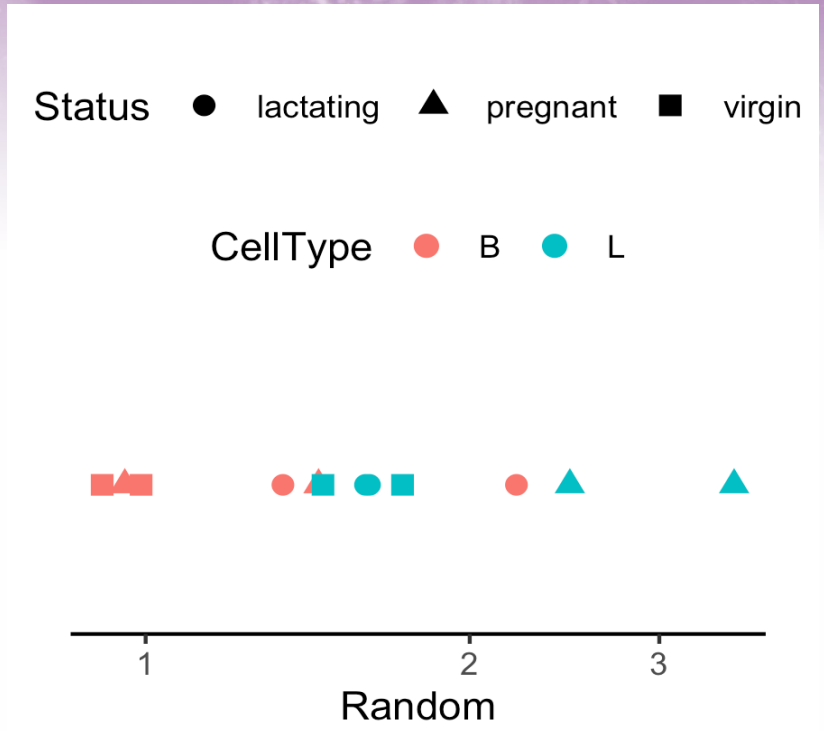
Could it be due to batch effects?



The distance between each pair of samples can be interpreted as the leading log-fold change between the samples for the genes that best distinguish that pair of samples.

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.



Fitting the model

How to model RNA-seq data

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.

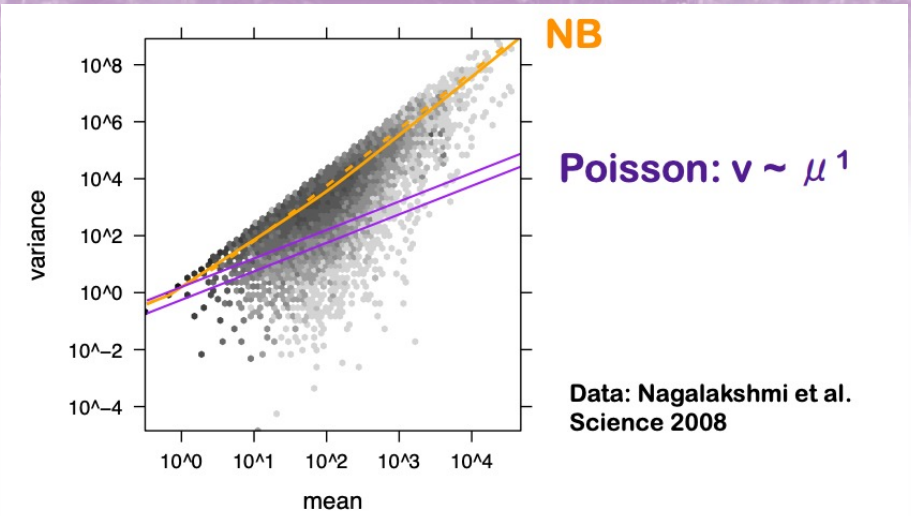
Discrete events sampled out of a large pool with low probability are usually modeled with Poisson distribution

RNA-seq data and overdispersion between samples

The number of reads that are mapped into a gene was first modeled using a Poisson distribution

Assumption: assumes that mean and variance are the same

The variance grows faster than the mean in RNAseq data.



Overdispersion in RNA-seq data

-> counts from biological replicates vary so tend to have variance exceeding the mean (highly expressed genes)

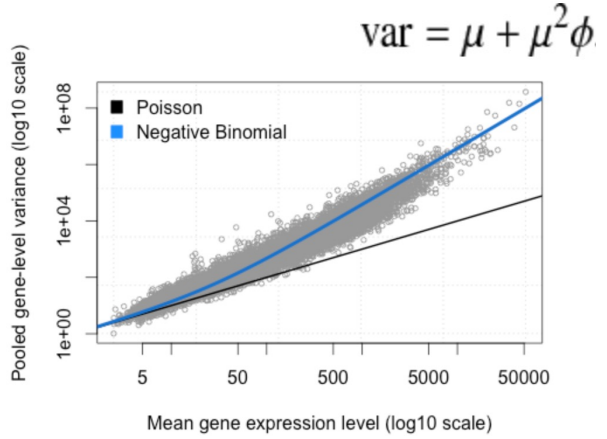
-> underestimation of the biological variance increased the probability to falsely declare a gene DE when it is non-DE (type I error rate)

Three dispersion estimates

The negative binomial (NB) distribution proposed as an alternative to model the read counts for each gene *in each sample*

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

$$\text{var} = \mu + \mu^2\phi$$



Many genes, few biological samples - difficult to estimate φ on a gene-by-gene basis
Using information *across all genes* for stable estimates of φ.
3 ways to estimate φ: common, trended, tagwise (moderated)

Empirical Bayes shrinkage moderated dispersion for each individual gene

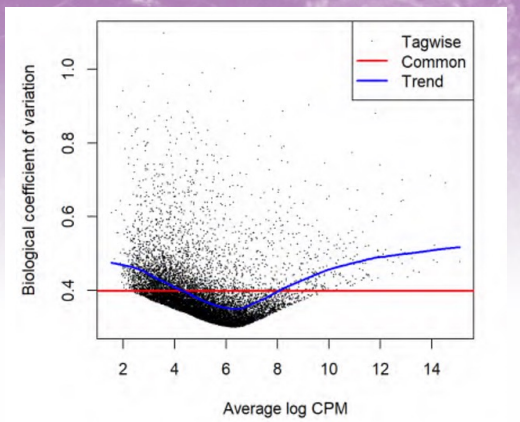
Empirical Bayes estimates of dispersion parameters: Learning from the experience of others

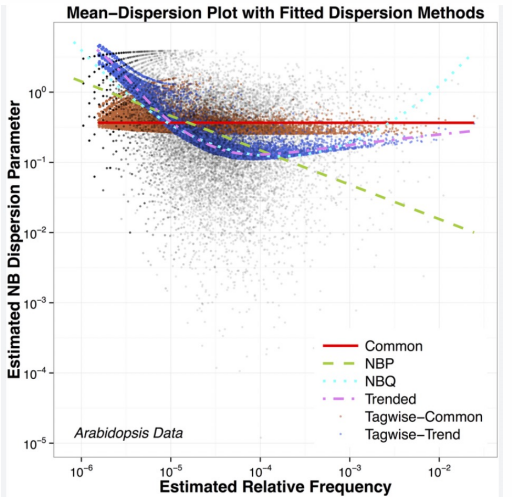
Dispersion accounts for variability between biological replicates

- 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

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

plotBCV (biological coeff variation)

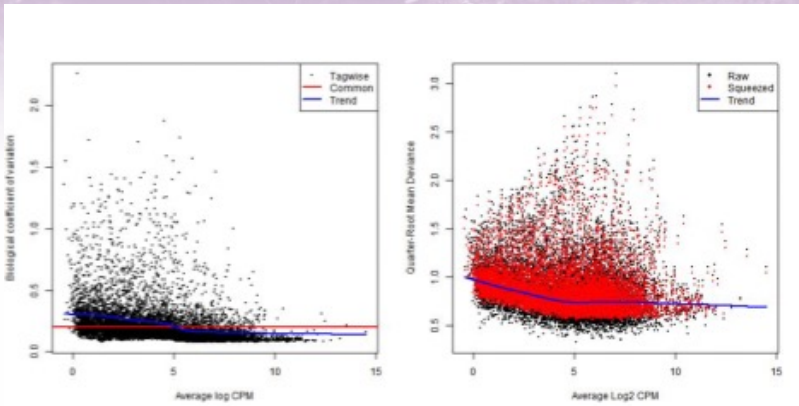




Fitting the model *: glmQLFit* to manage the variance of gene counts

- NB dispersions - higher for genes with very low counts and 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)
2) For each gene-specific variability above and below the overall level (deviance) is picked up by the QL dispersion

edgeR fitting models

- Classic (pairwise comparisons between two or more groups), glm and glmQL
- QL for bulk RNA-seq:
  - + stricter error rate control (more rigorous dispersion and uncertainty)
  - + speed improvement compared to other quasi-methods
  - + appropriate 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*

- 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 *decideTestsDGE() – set cutoff*

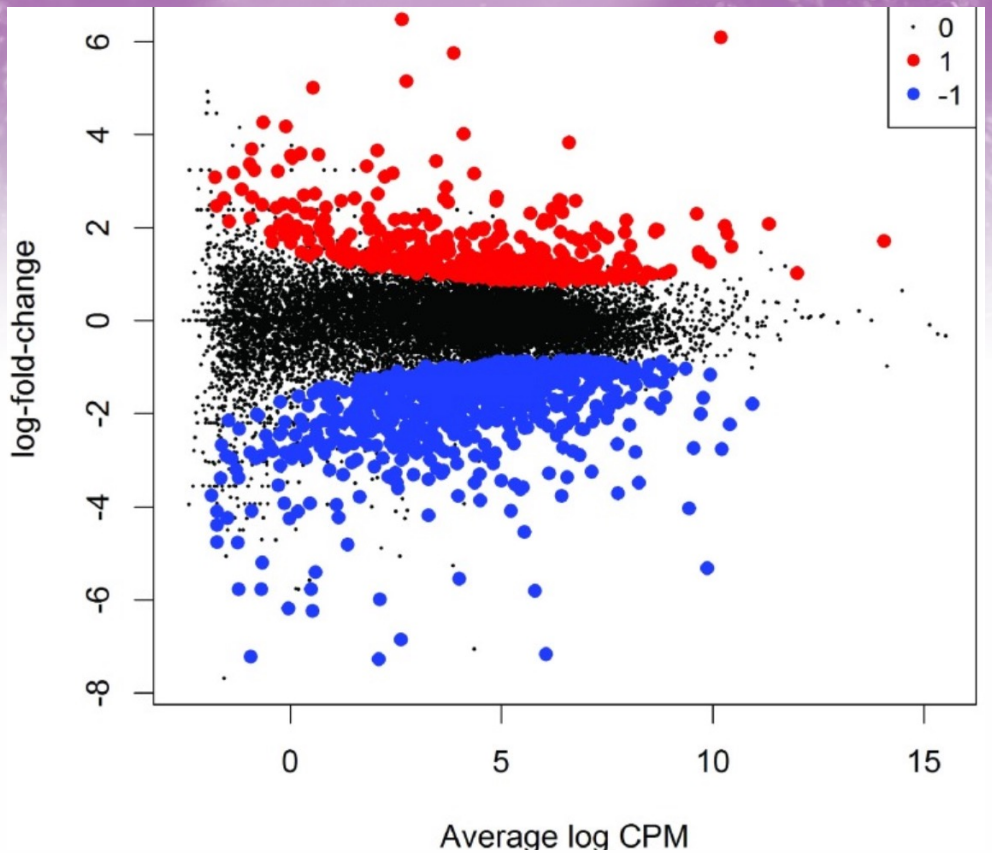
Get the DE genes - *glmQLFTest*

	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

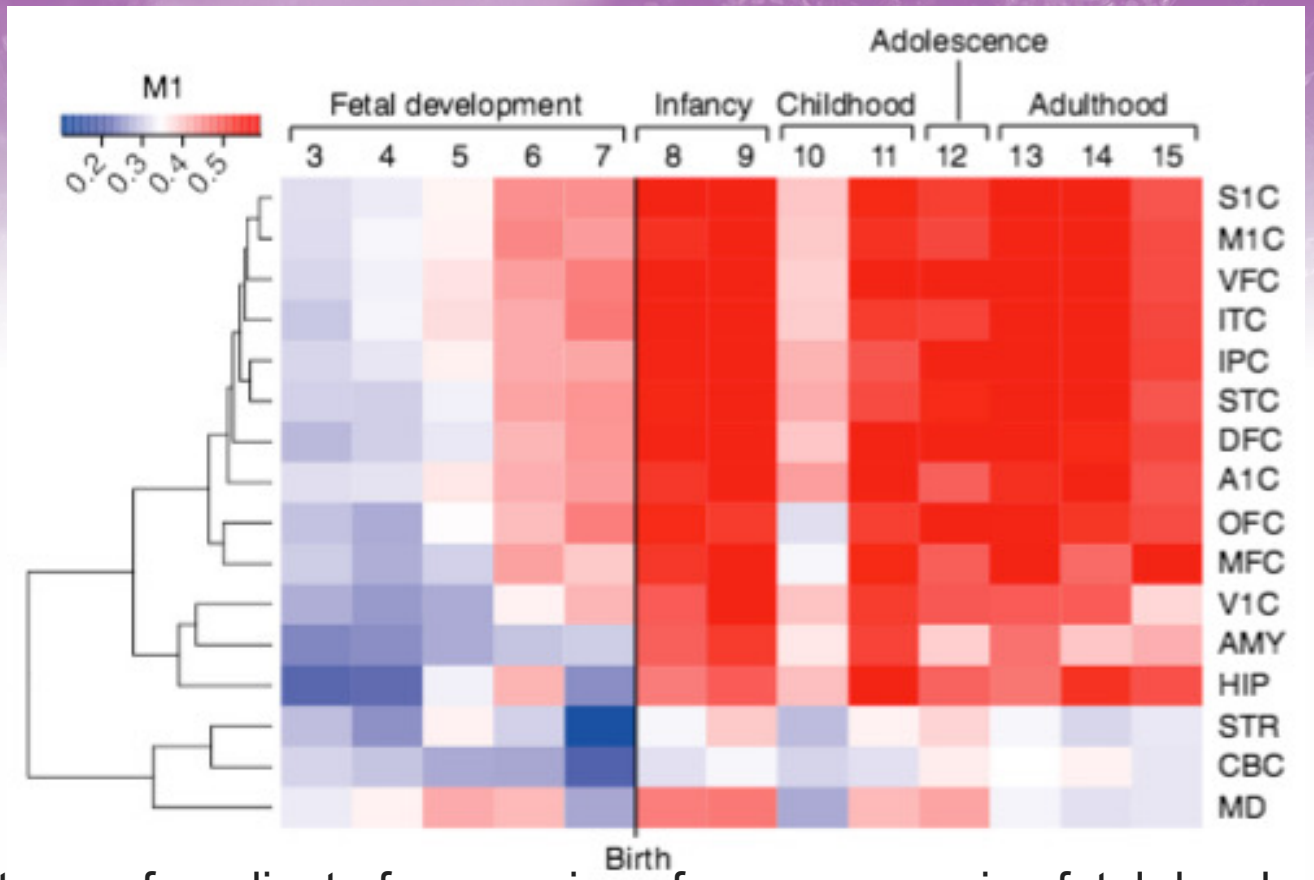
MD plot: Over and under expressed genes



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

Visualization - heatmap



Pathway analysis

Heatmap of gradient of expression of genes, spanning fetal development to late adulthood and expressed in cortical regions

In summary

- Raw counts are not comparable across samples/genes within a sample
- Several normalization methods: some more suitable for DEG
- Estimate the source of variability across samples
- Counts from biological replicates: variance exceeding the mean
- Estimate the dispersion, control for the technical variability
- Fit the model
- Make all the comparisons that you wish and visualize the DEG

Your feedback is important to us!

At the end of the hands-on session:

Please take the survey ~3 min:
https://www.surveymonkey.com/r/F75J6VZ

Real data might need additional analyses choices that need experience.

Consult with the Gladstone Bioinformatics core for such scenarios and data.

Demo

Materials for the demo

- Hands-on session:
  - handson.R
  - targets.txt
  - DE_resutls.txt
  - GSE60450_Lactation-GenewiseCounts.txt.gz

Please install the following library in Rstudio

Install.packages(magrittr)

Install.packages(edgeR)

Install.packages(org.Mm.eg.db)

Install.packages(tidyverse)

Install.packages(ggplot2)

Install.packages(statmod)

Hands-on session

- Load and reformat the data
- Exploratory visualization : MA plot
- Create DGEList object and retrieve gene symbols
- Filter genes with inadequate information
- 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


Thank you!


GLADSTONE
INSTITUTES