RNA-seq data analysis: From counts to differentially expressed genes using edgeR-quasi

Gladstone Institutes

Krishna Choudhary
Bioinformatics Core, GIDB

2020

Assumed background

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

Typical protocol

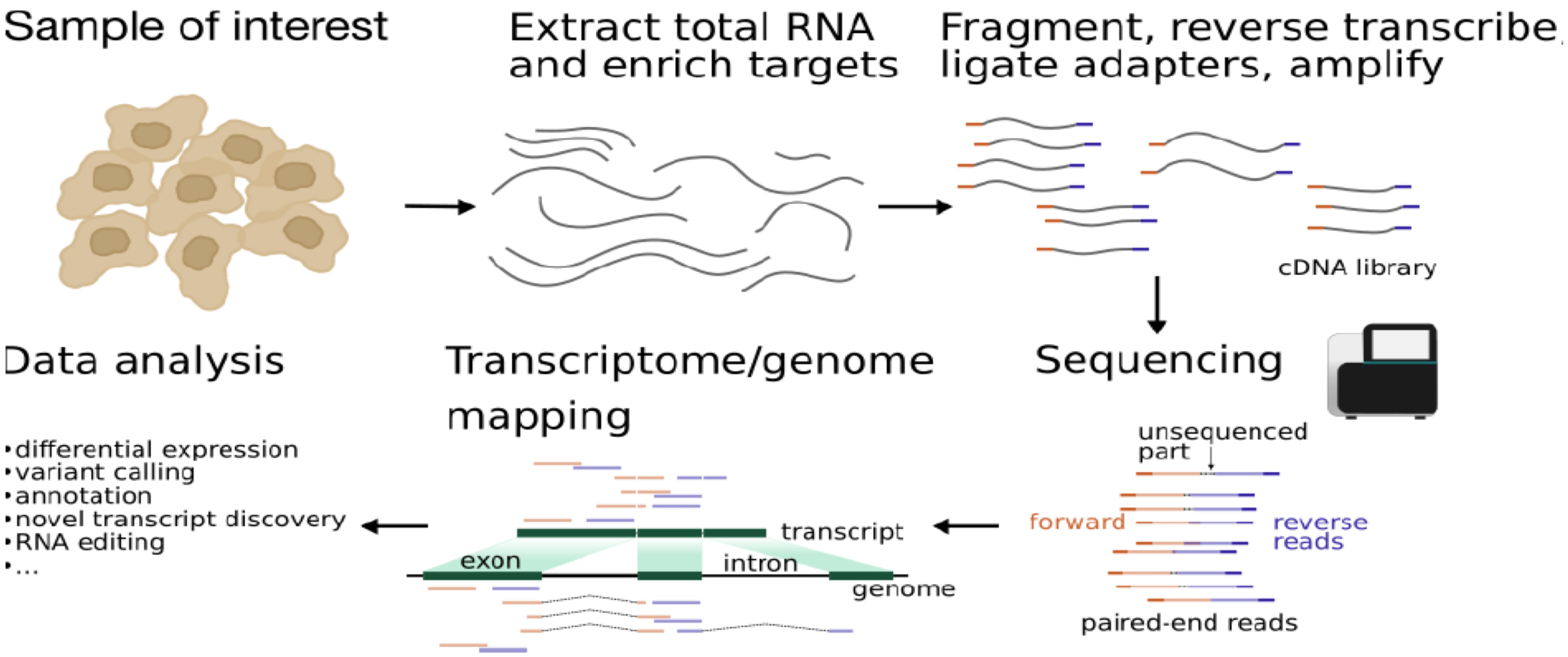


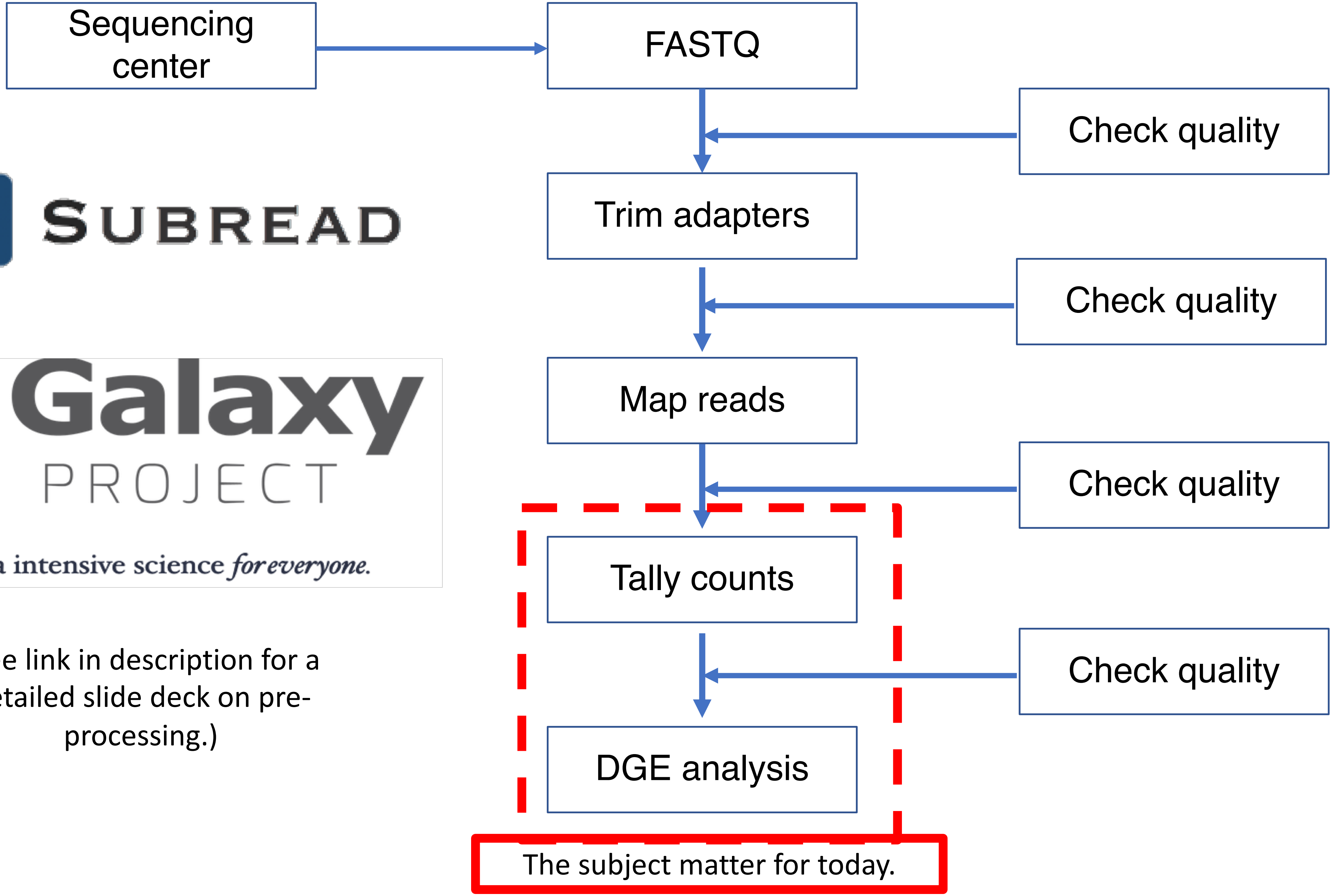
Figure: Berge et al., 2018, PeerJ Preprints.

Experiment design influences data analysis.
(should be planned to address relevant questions)

- What is the biological question that we seek to answer?
- How many tissue types and/or time points to compare?
- How deep should we sequence?
- Read length?
- Which sequencing platform?
- Single-end or paired-end?
- Pooling?
- Biological replicates?
- Technical replicates?
- Additional considerations?

Not the subject matter today!

- Workshop by Reuben Thomas: *Intro to statistics and experimental design*.
- Reading material: *RNA sequencing data : hitchhiker's guide to expression analysis* by Berge *et al.*, 2018



Sequencing center

FASTQ

Check quality

Trim adapters

Check quality

Map reads

Check quality

Tally counts

Check quality

DGE analysis

The subject matter for today.

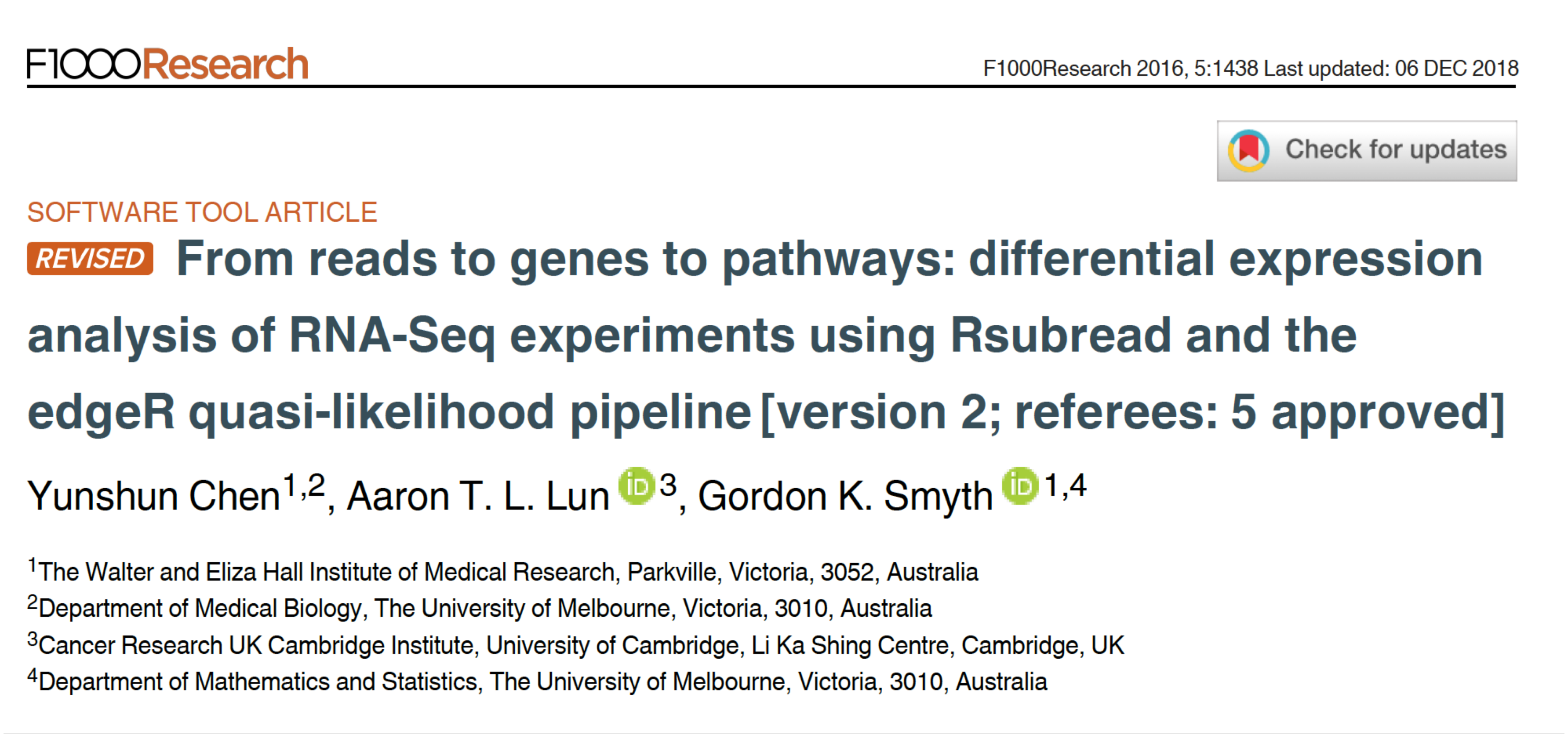
SUBREAD

Galaxy PROJECT

Data intensive science *for everyone.*

(See link in description for a detailed slide deck on pre-processing.)

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[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

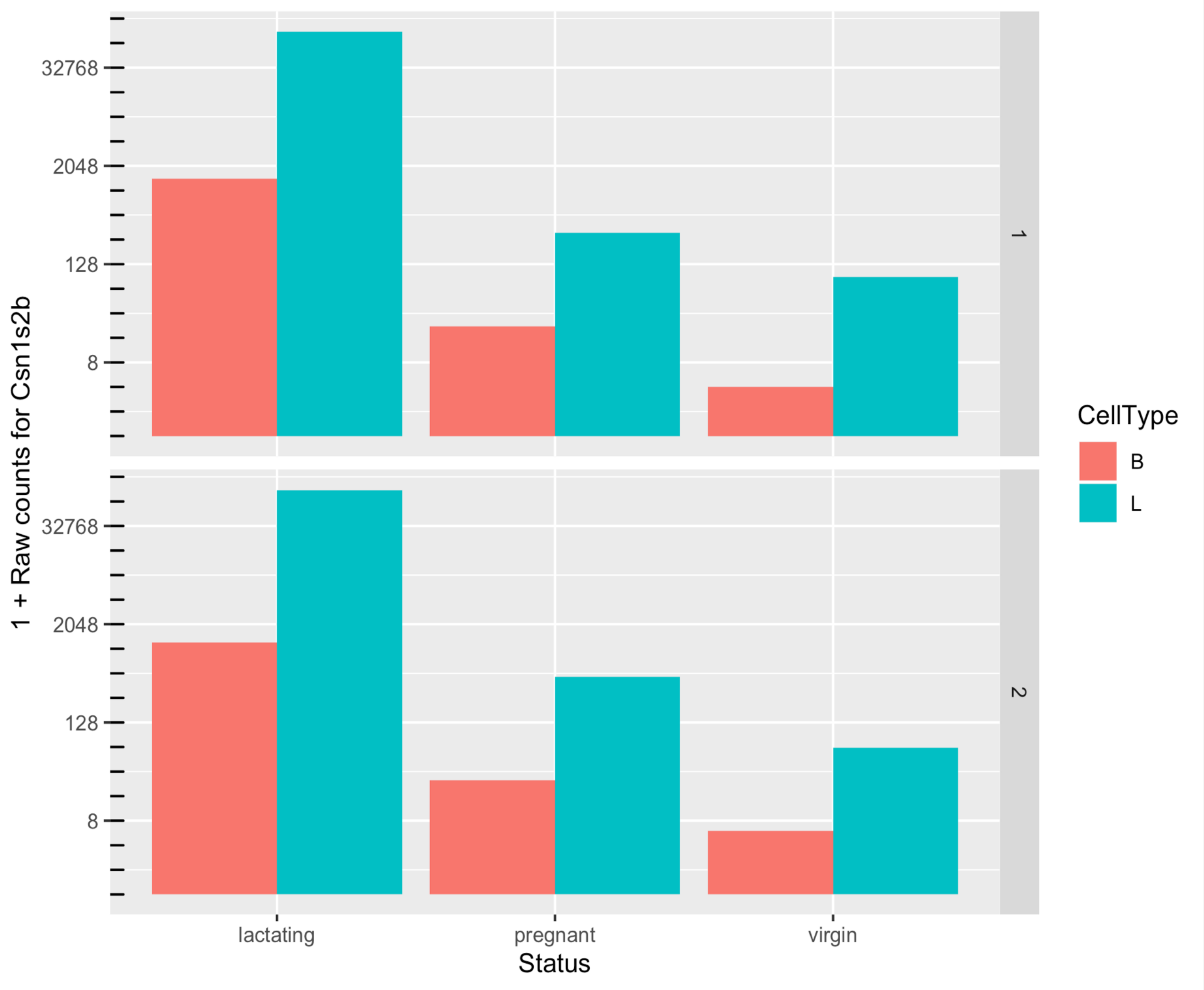
Outline

- Load and reformat the data
- Exploratory visualization : MA plot
- Create DGElist object and retrieve gene symbols
- Filter genes with inadequate information
- Normalize counts : *What's under the hood?*
- 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

Dataset

- GEO 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 and Pregnant
- # of groups: 2 cell types x 3 conditions = 6 groups
- # of replicates: 2 of each group

Goal: To identify a set of genes that are differentially expressed

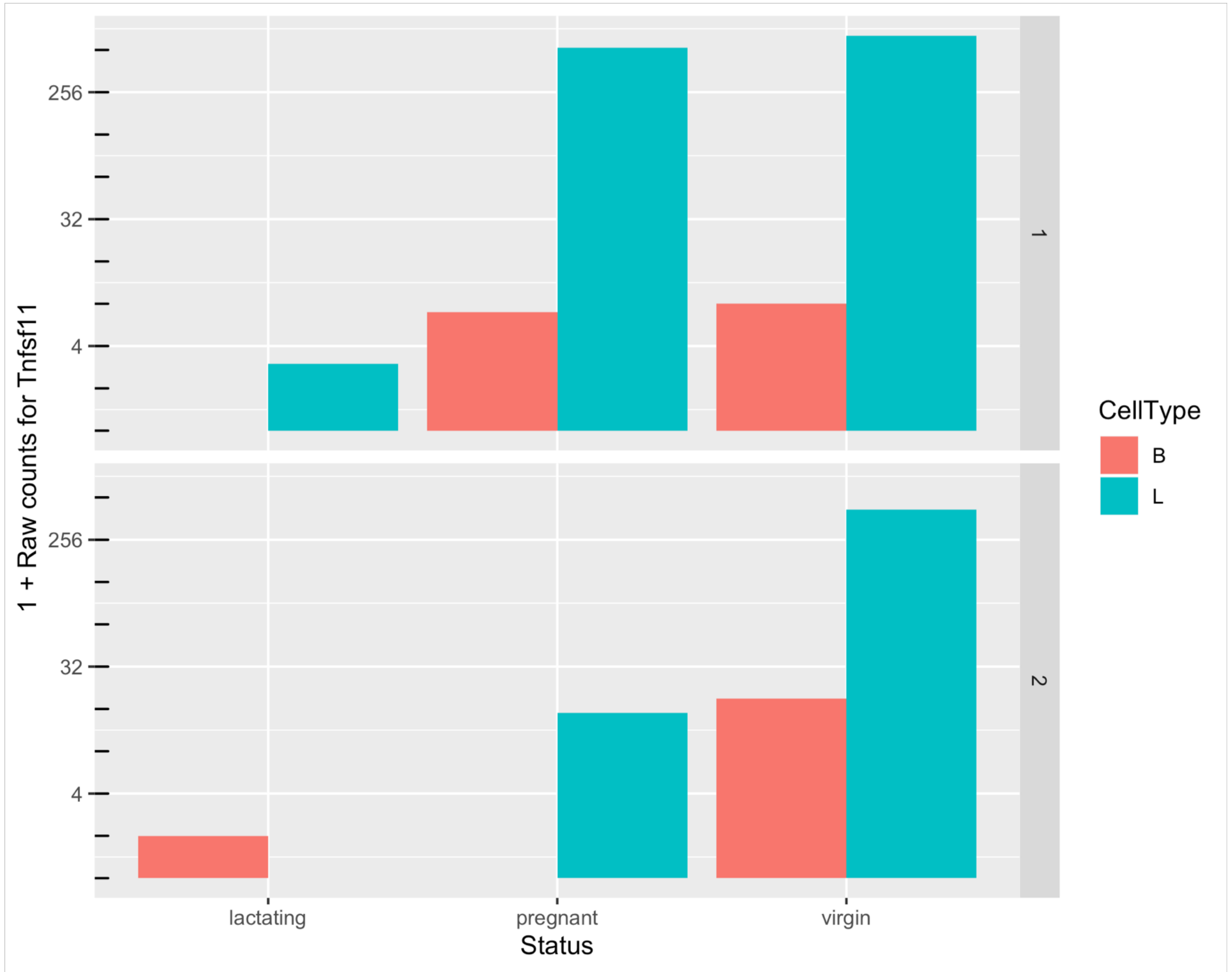


Which comparisons are we interested in?

Example:

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

Goal: To identify a set of genes that are differentially expressed



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 to identify DE genes:

- 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.

Factors other than differential gene expression that 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

Observed counts depend on total reads sequenced and sample composition

- # reads for YFG = $\frac{\text{Amount of nucleic acid from YFG}}{\text{Total nucleic acid in sample}}$ x Total reads
- 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 sample composition.
  - Assume that the large majority of genes are not differential.
  - Adjust counts such that for most genes, counts are not differential.

Normalization: Trimmed Mean of M-values

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

Other approaches to normalization

- RLE approach by Anders and Huber (2010)
  - Reference: geometric mean of all samples
  - Normalization factor: median ratio of each sample to the reference
  - Identical to TMM approach
  - See link in description for complete reference

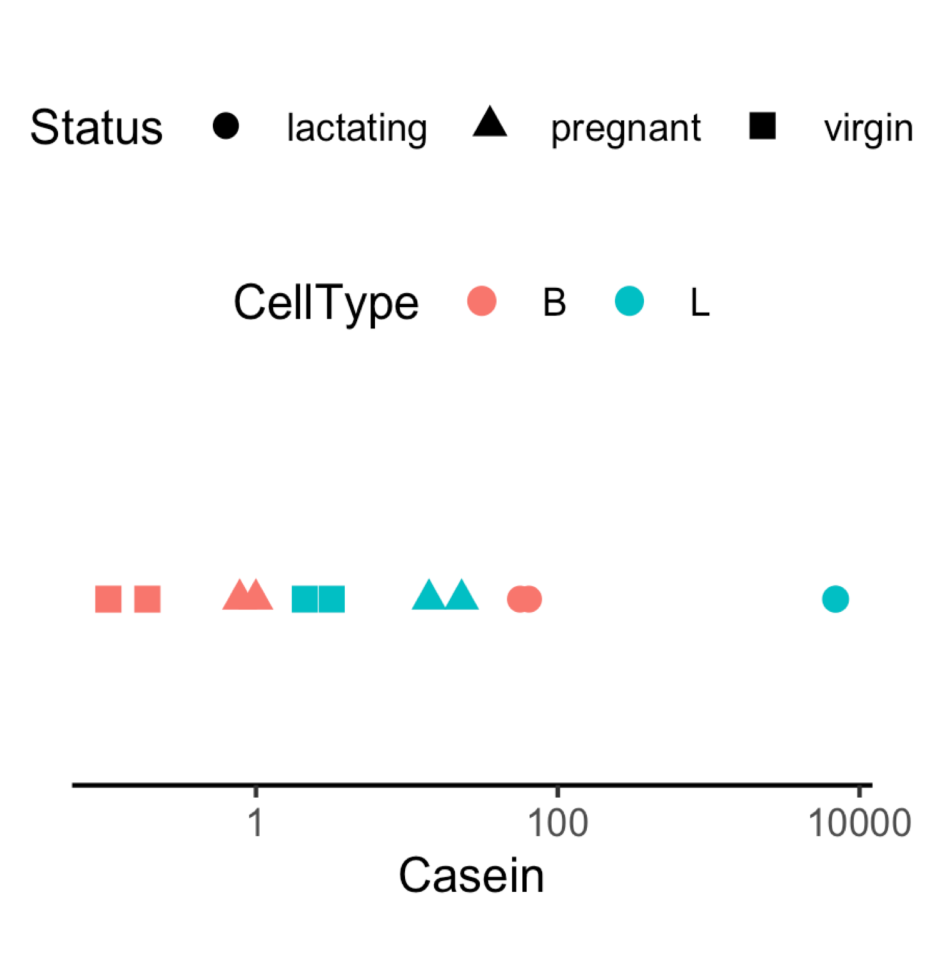
- Upper quartile normalization by Bullard et al (2010)
  - Normalization factor: 75% quantile of the counts for each sample
  - Not recommended in general
  - See link in description for complete reference

MDS and PCA plots

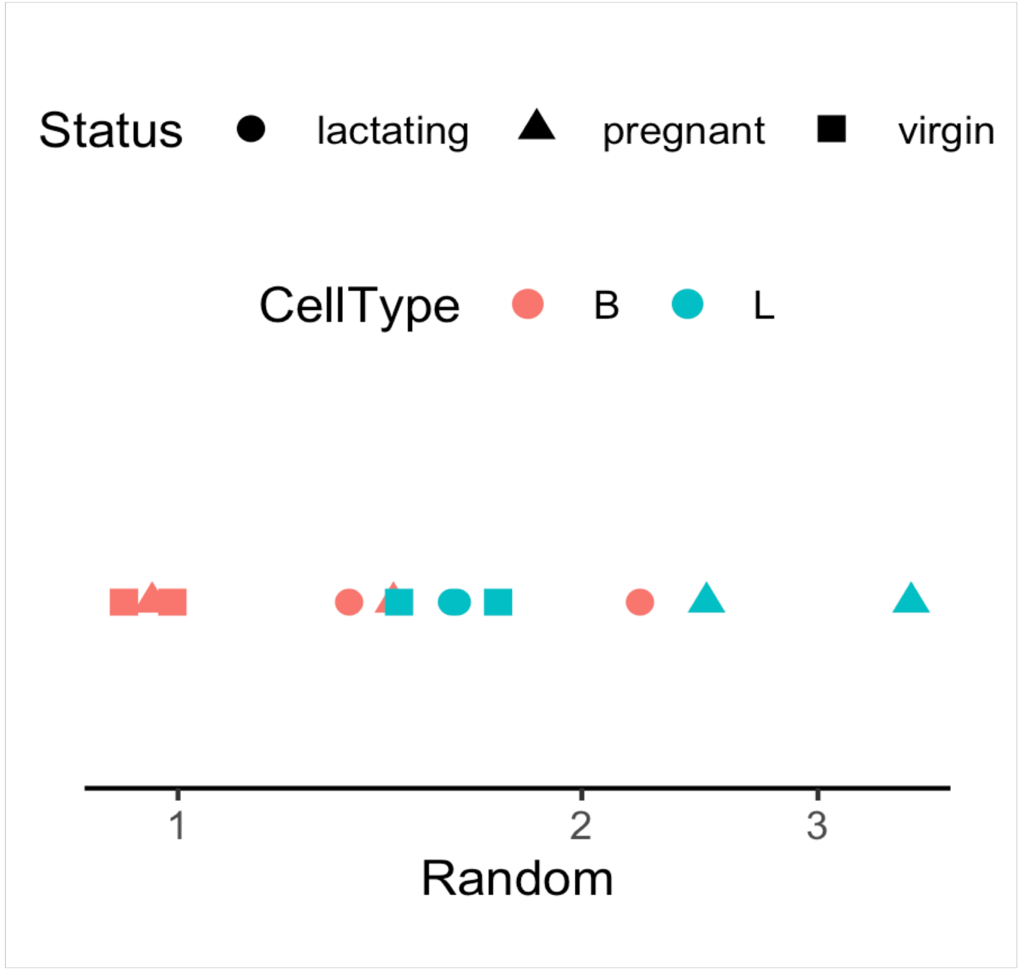
Expression level of Casein varies in a way that is strongly indicative of the effect of CellType and Status.

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

Could it be due to batch effects?

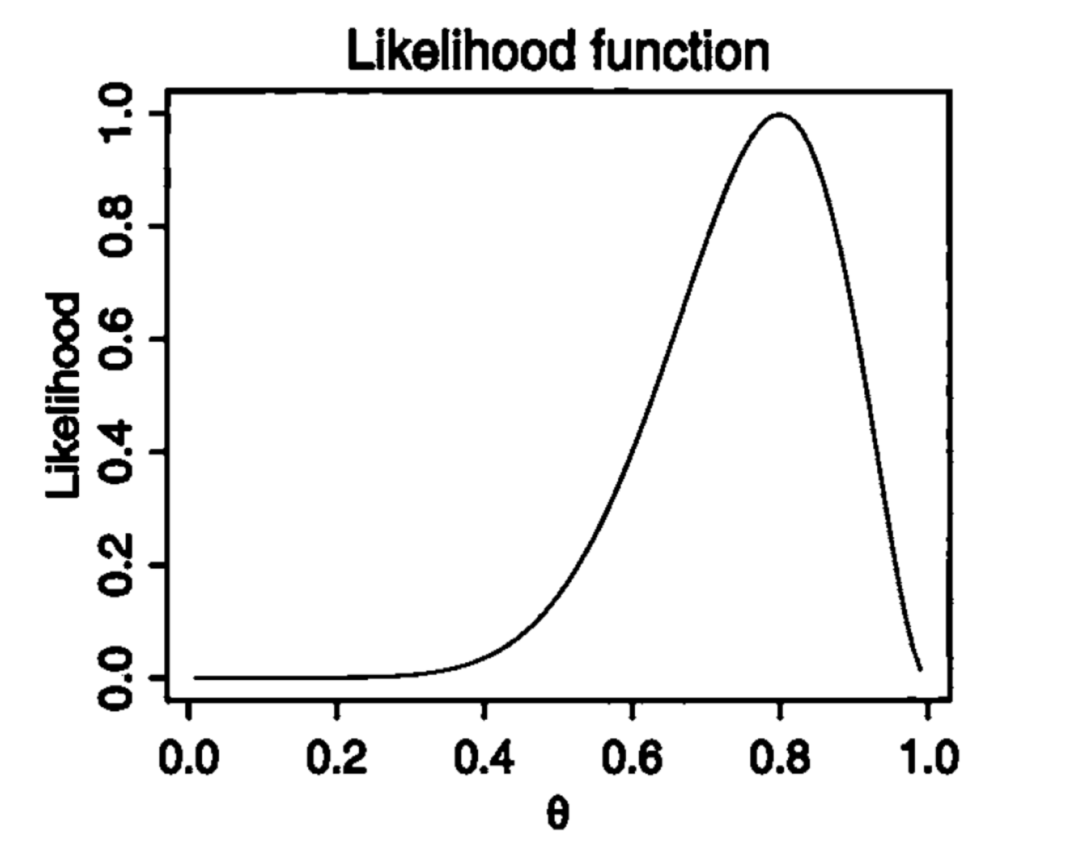


In general, the way expression appears to vary across samples could be dominated by noise, batch effects, real signal, etc.



Maximum Likelihood Estimates:

What’s the parameter value that makes the observed data the most likely observation?



Let’s run a “thought” experiment:

- Say, we randomly pick 10 mRNA molecules from a sample.
- We observe that 8 of them come from YFG.
- What is the proportion of YFG molecules out of all the mRNAs?

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

- For an intuitive explanation by Bradley Efron, see link in description

- Original paper: Robbins, Herbert. *An Empirical Bayes Approach to Statistics.* (see description for complete reference)

Upcoming workshops

- Intro to Pathway Modeling
- Whole Genome Sequence Analysis

Your feedback is important to us!

- https://bioinformatics-course-feedback.questionpro.com/
- ~3 min.

Thank you.