

CANCER  
RESEARCH  
UK

CAMBRIDGE  
CENTRE

# Introduction to Differential Gene Expression Analysis in R

CRUK Summer School 2021

Ashley Sawle

July 2021

# HTS Applications - Overview

## DNA Sequencing

- Genome Assembly
- SNPs/SVs/CNVs
- DNA methylation
- DNA-protein interactions (ChIPseq)
- Chromatin Modification (ATAC-seq/ChIPseq)

## RNA Sequencing

- Transcriptome Assembly
- **Differential Gene Expression**
- Fusion Genes
- Splice variants

## Single-Cell

- RNA/DNA
- Low-level RNA/DNA detection
- Cell-type classification
- Dissection of heterogenous cell populations

# RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

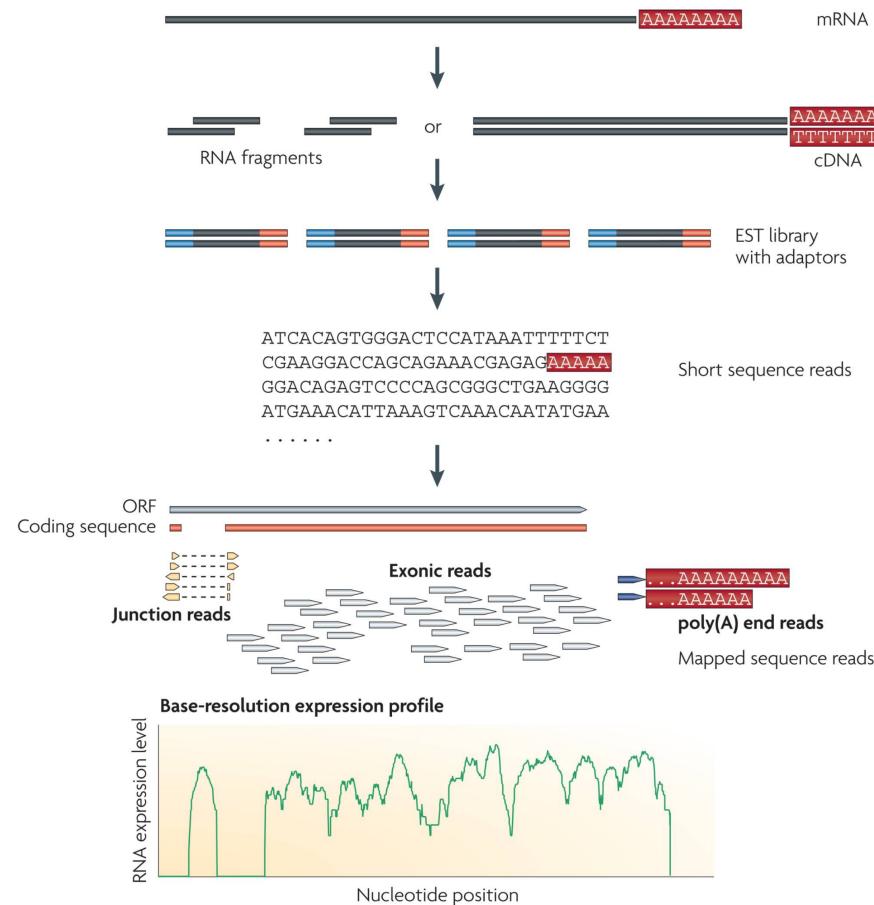


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

# Designing the right experiment

A good experiment should:

- Have clear objectives
- Have sufficient power
- Be amenable to statistical analysis
- Be reproducible
- More on experimental design later

# Designing the right experiment

## Practical considerations for RNAseq

- Coverage: how many reads?
- Read length & structure: Long or short reads? Paired or Single end?
- Library preparation method: Poly-A, Ribominus, other?
- Controlling for batch effects

# Designing the right experiment - How many reads do we need?

The coverage is defined as:

$$\frac{\text{Read Length} \times \text{Number of Reads}}{\text{Length of Target Sequence}}$$

The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample.

- For a general view of differential expression: 5–25 million reads per sample
- For alternative splicing and lowly expressed genes: 30–60 million reads per sample.
- In-depth view of the transcriptome/assemble new transcripts: 100–200 million reads
- Targeted RNA expression requires fewer reads.
- miRNA-Seq or Small RNA Analysis require even fewer reads.

# Designing the right experiment - Read length

## Long or short reads? Paired or Single end?

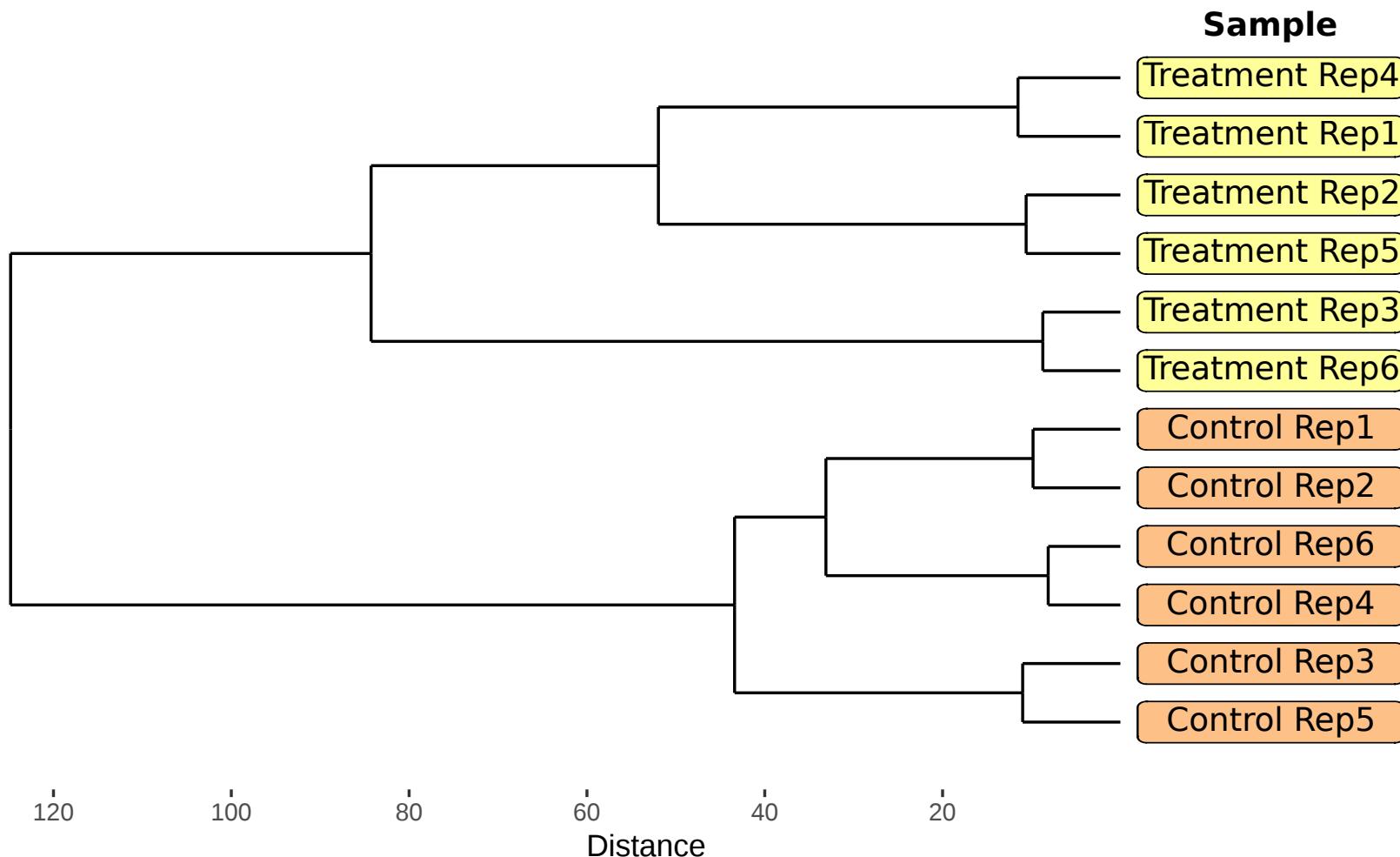
The answer depends on the experiment:

- Gene expression – typically just a short read e.g. 50/75 bp; SE or PE.
- kmer-based quantification of Gene Expression (Salmon etc.) - benefits from PE.
- Transcriptome Analysis – longer paired-end reads (such as 2 x 75 bp).
- Small RNA Analysis – short single read, e.g. SE50 - will need trimming.

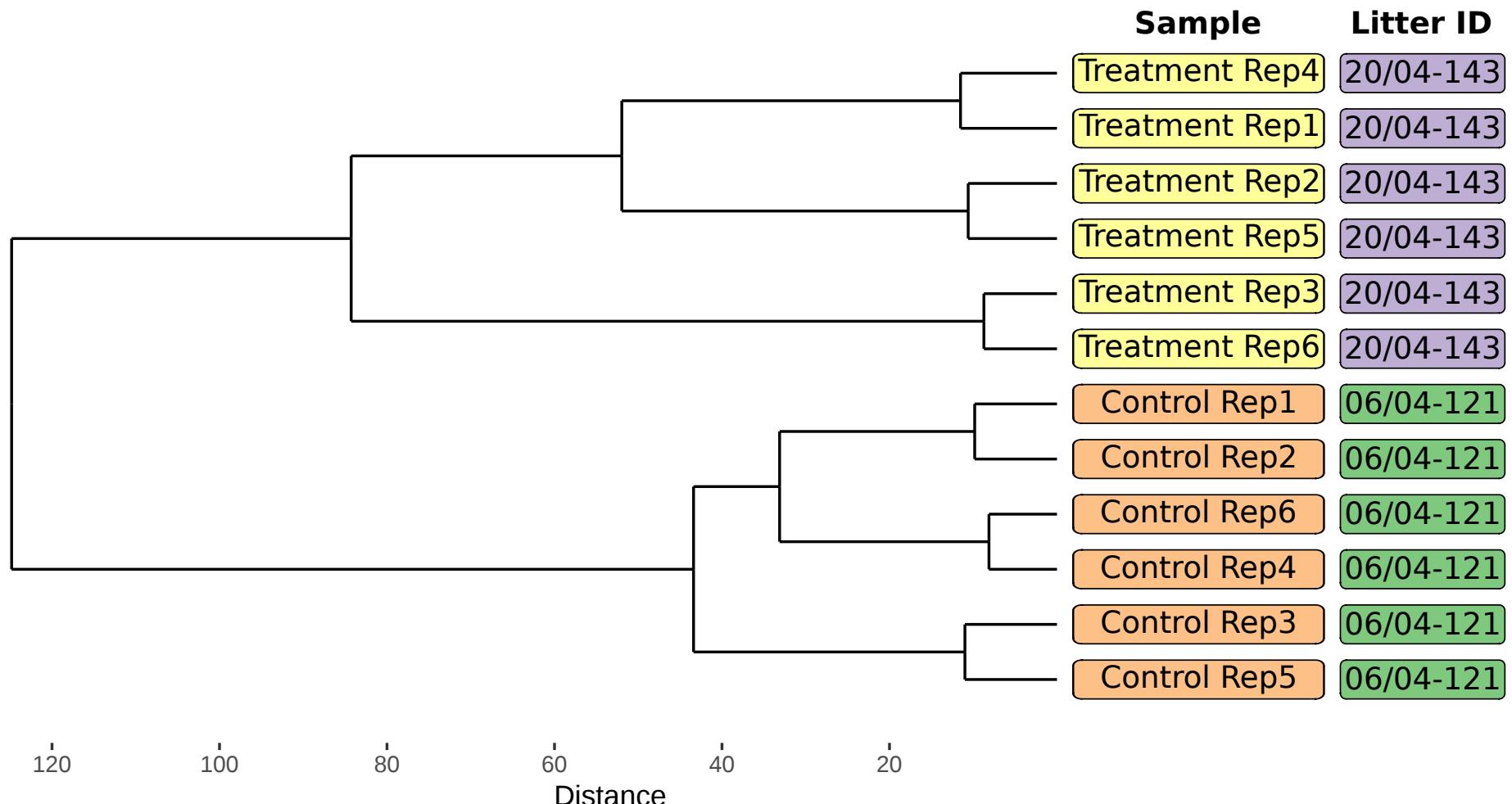
# Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.

# Designing the right experiment - Batch effects



# Designing the right experiment - Batch effects

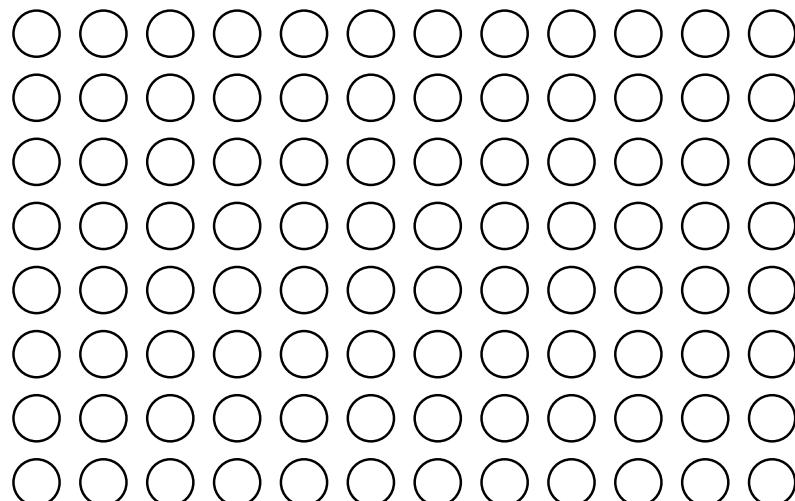


# Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.

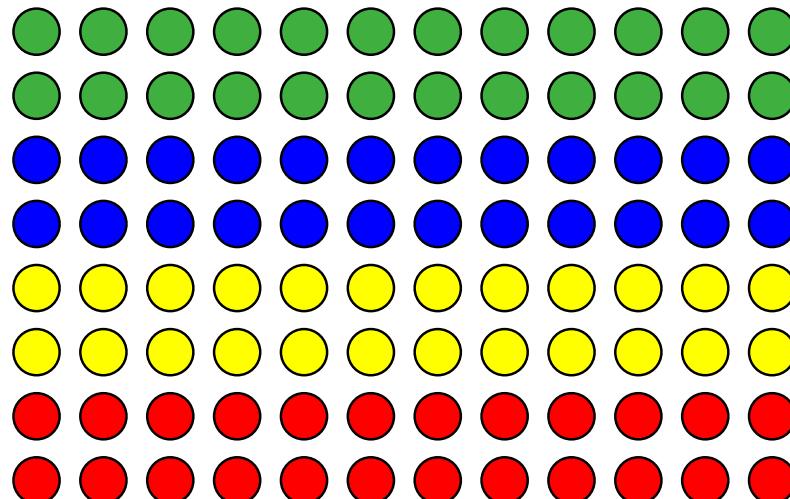
# Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



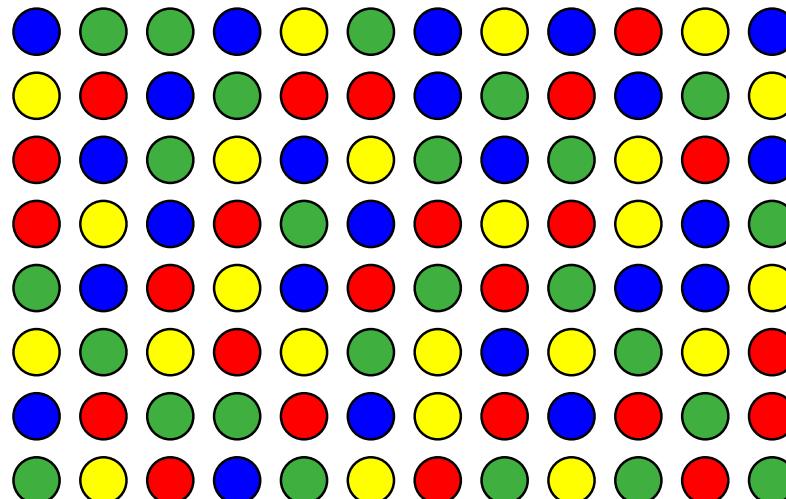
# Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



# Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



# Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects
- **Record everything:** Age, sex, litter, cell passage ..

# RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

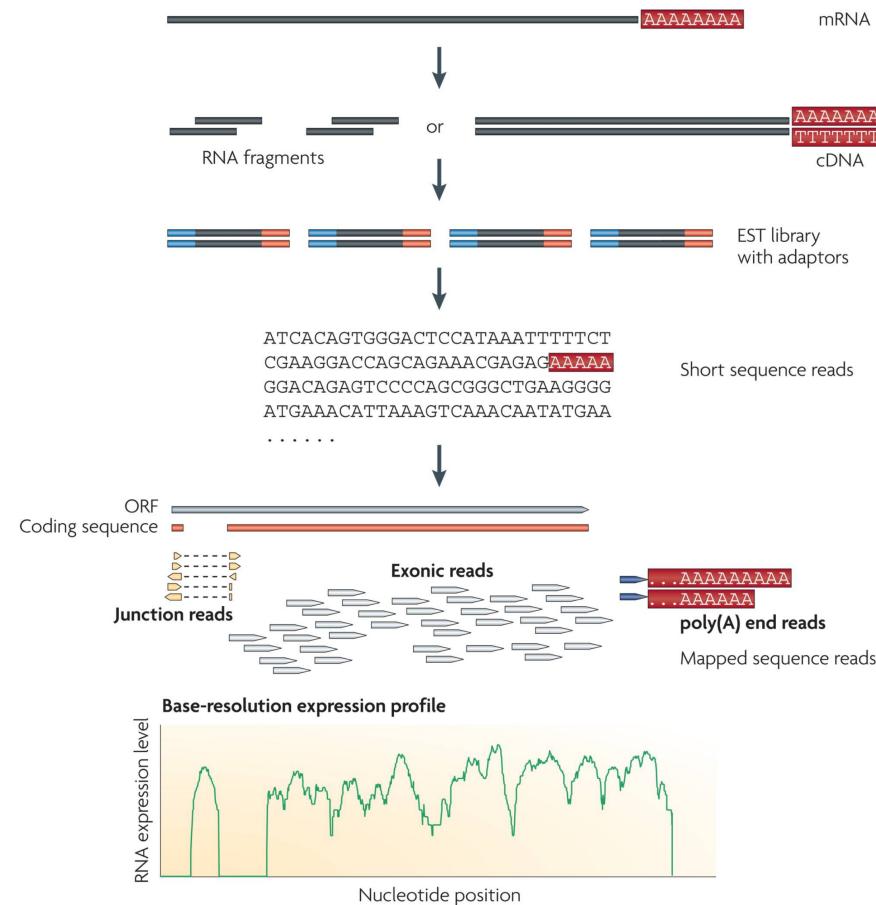
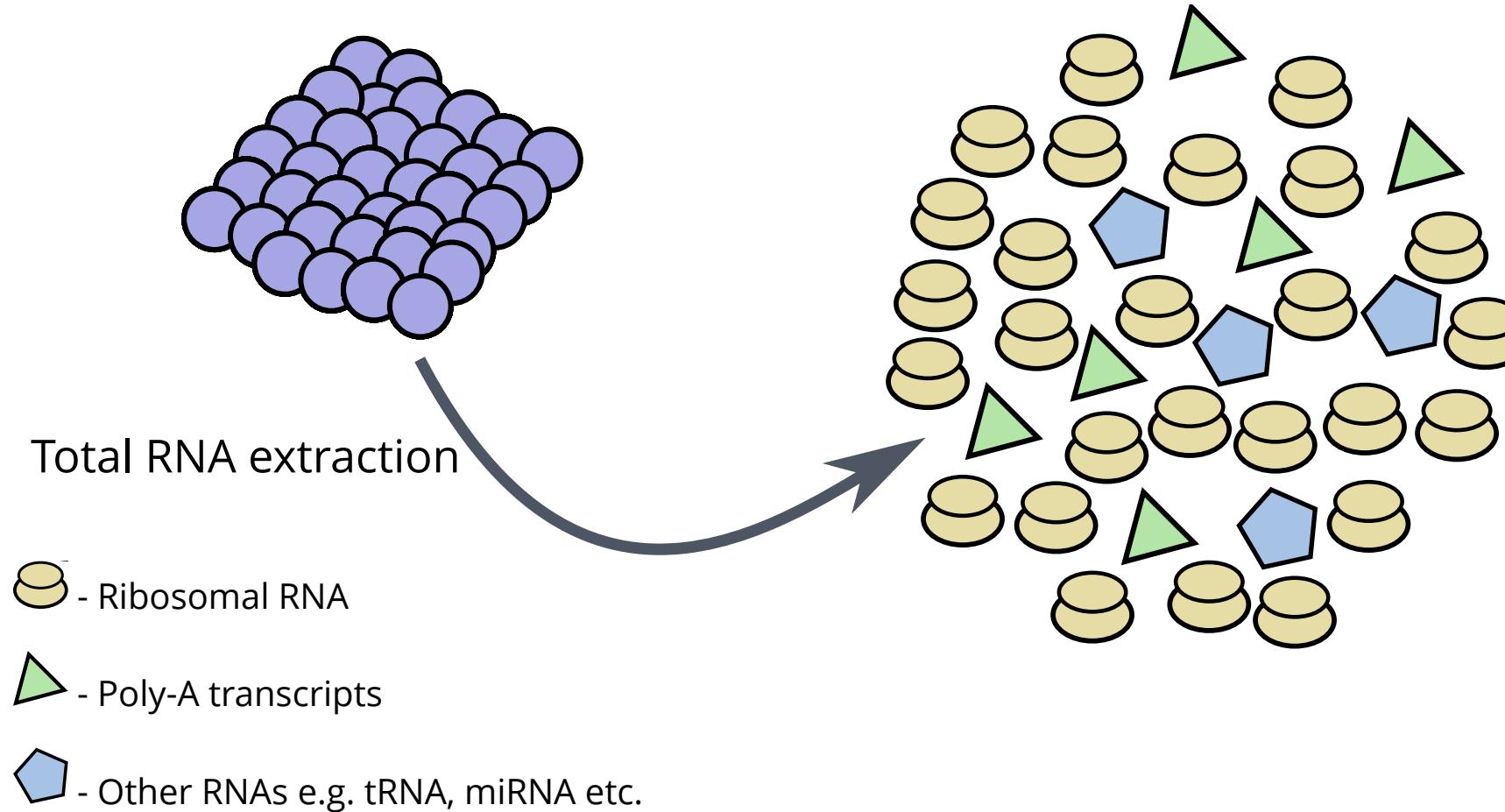


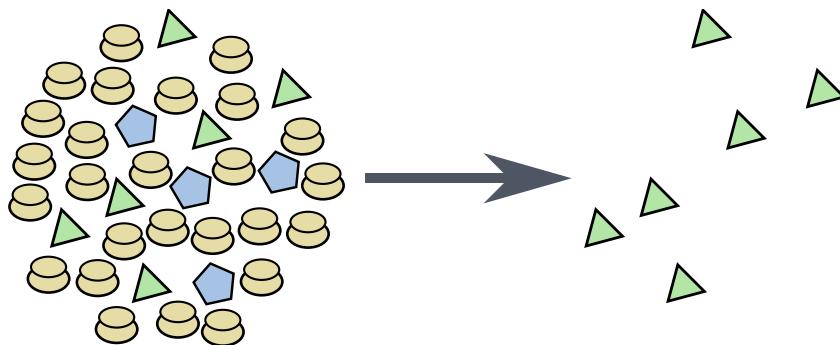
Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

# Library preparation



# Library preparation

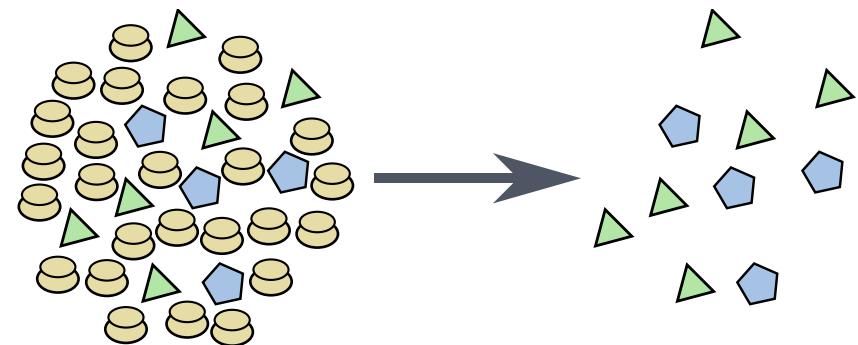
Poly-A Selection



Poly-A transcripts e.g.:

- mRNAs
- immature miRNAs
- snoRNA

Ribominus selection



Poly-A transcripts + Other mRNAs e.g.:

- tRNAs
- mature miRNAs
- piRNAs

# RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

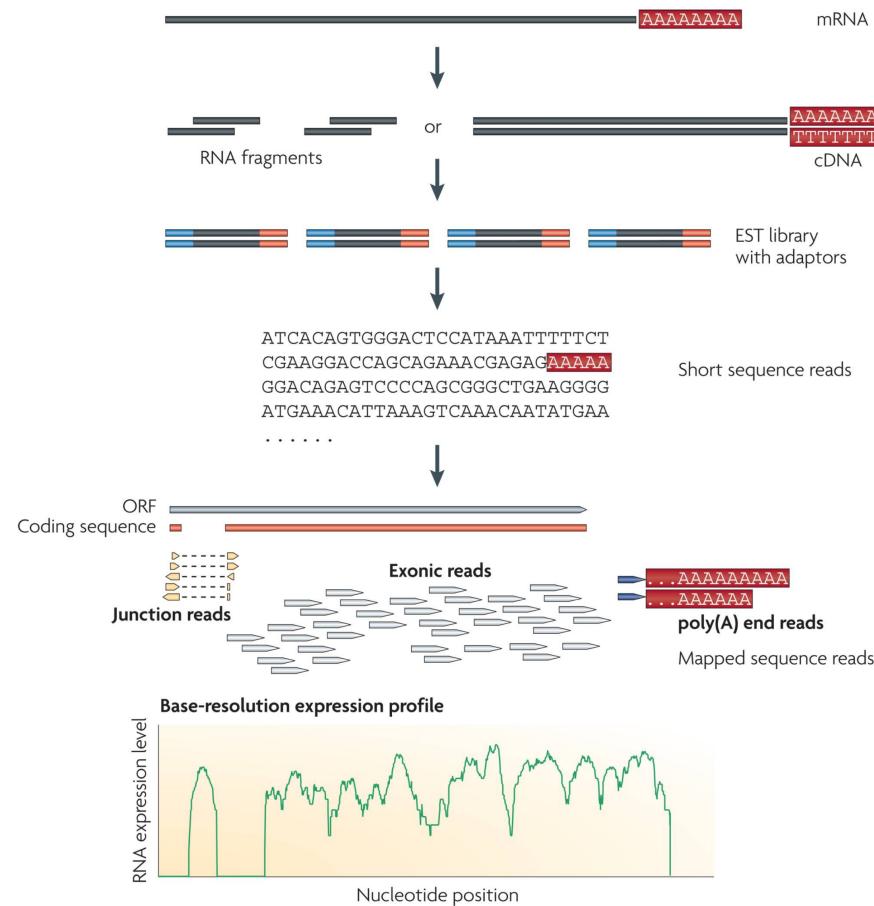


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

# RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

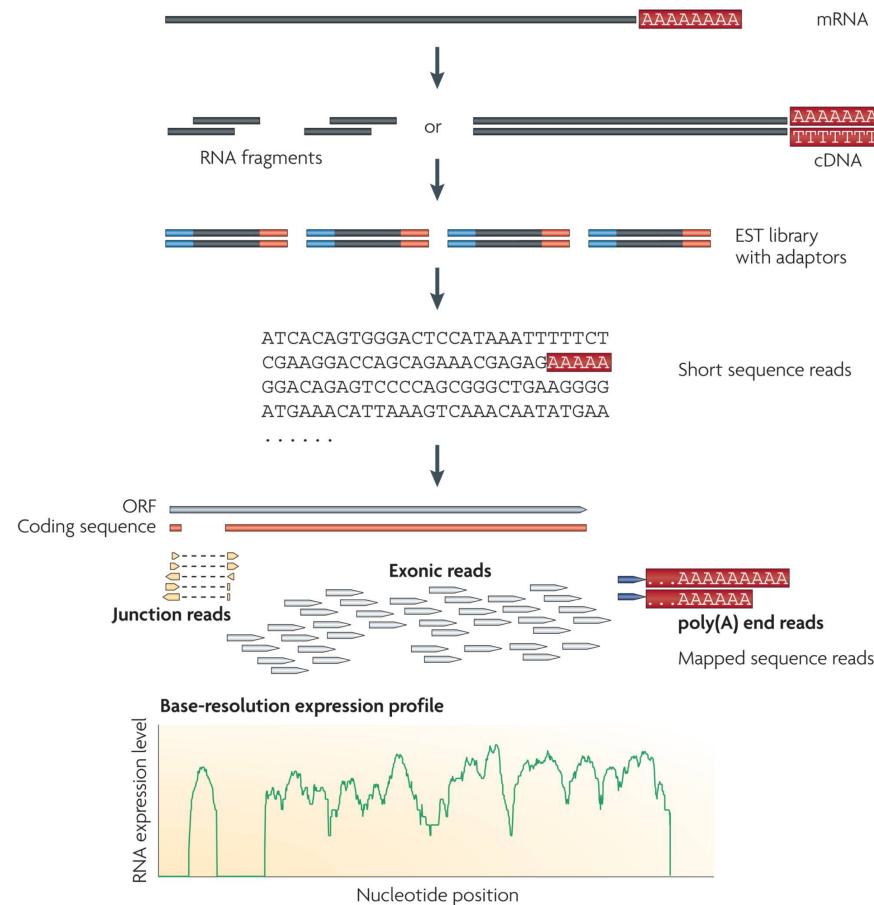
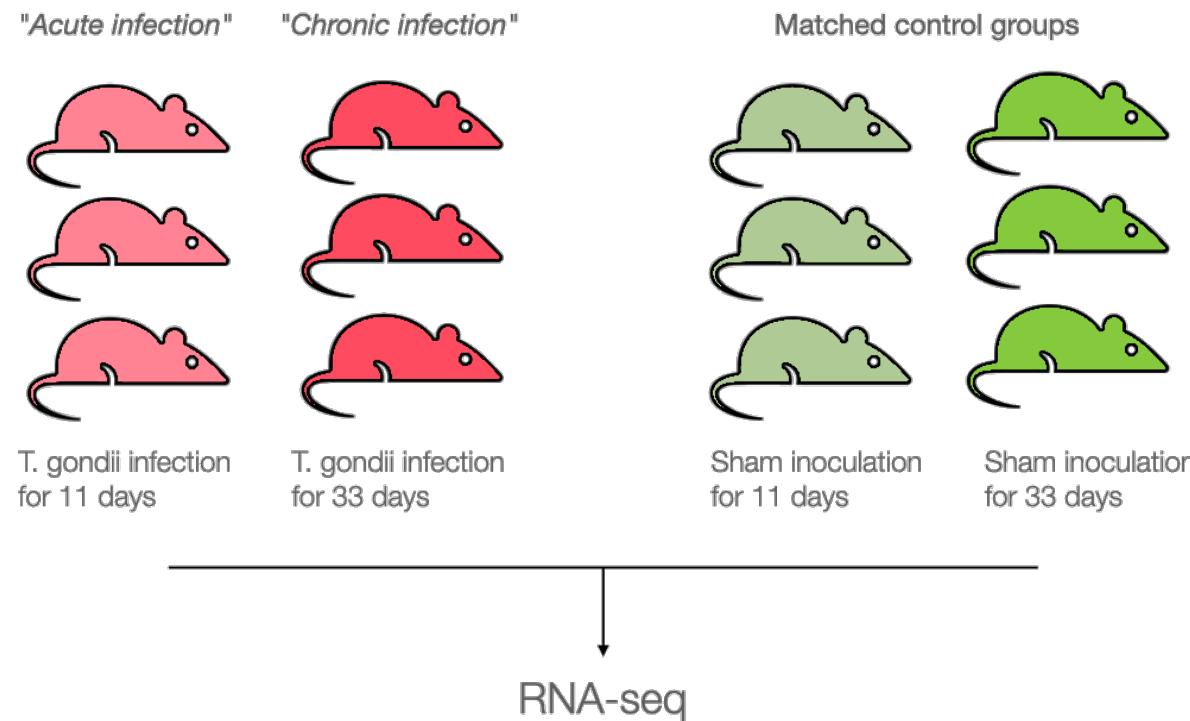


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

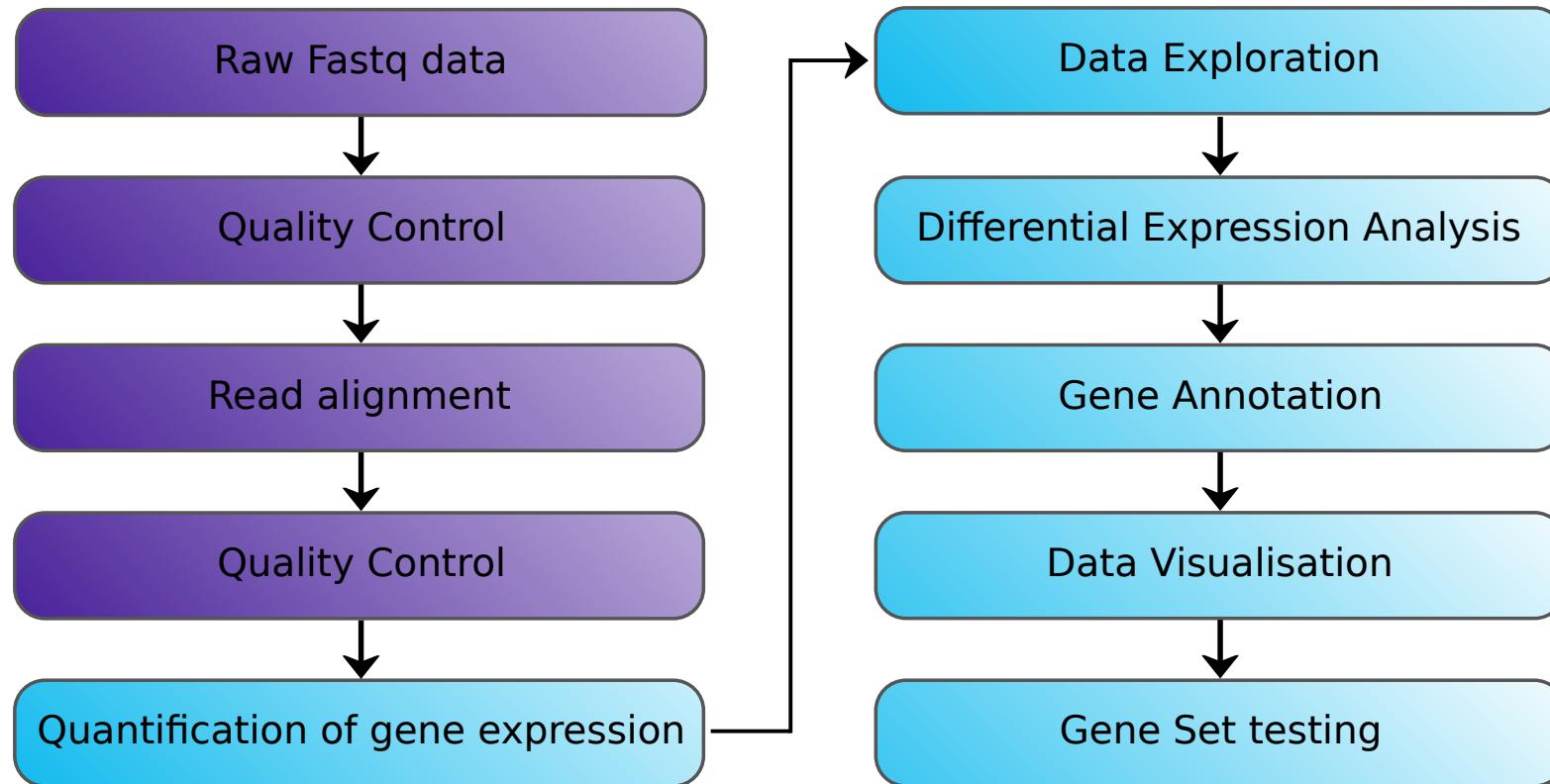
# Case Study

## Transcriptomic Profiling of Mouse Brain During Acute and Chronic Infections by *Toxoplasma gondii* Oocysts

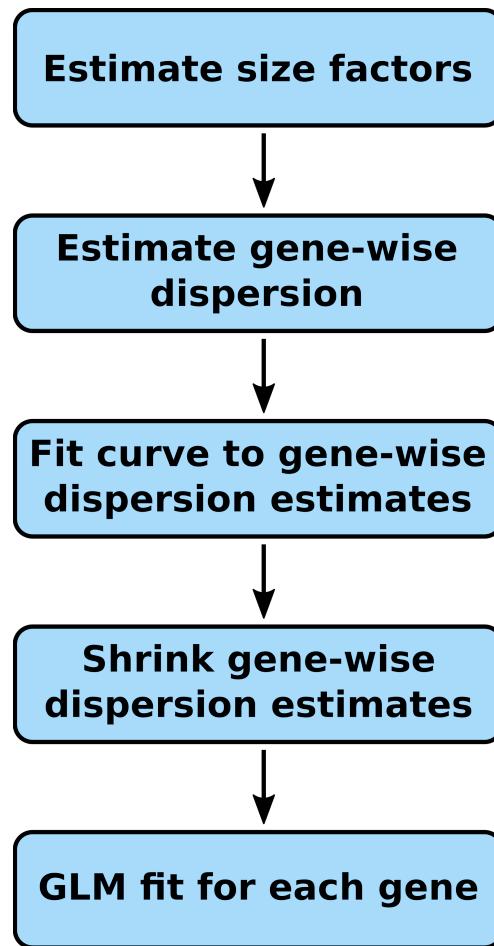
Rui-Si Hu<sup>1,2</sup>, Jun-Jun He<sup>1\*</sup>, Hany M. Elsheikha<sup>3</sup>, Yang Zou<sup>1</sup>, Muhammad Ehsan<sup>1</sup>, Qiao-Ni Ma<sup>1</sup>, Xing-Quan Zhu<sup>1,4</sup> and Wei Cong<sup>5\*</sup>



# Differential Gene Expression Analysis Workflow



# DESeq2 analysis workflow



# Normalisation

- Quantification estimates the *relative* read counts for each gene
- Does this **accurately** represent the original population of RNAs?
- The relationship between counts and RNA expression is not the same for all genes across all samples

## Library Size

Differing sequencing depth

## Gene properties

Length, GC content, sequence

## Library composition

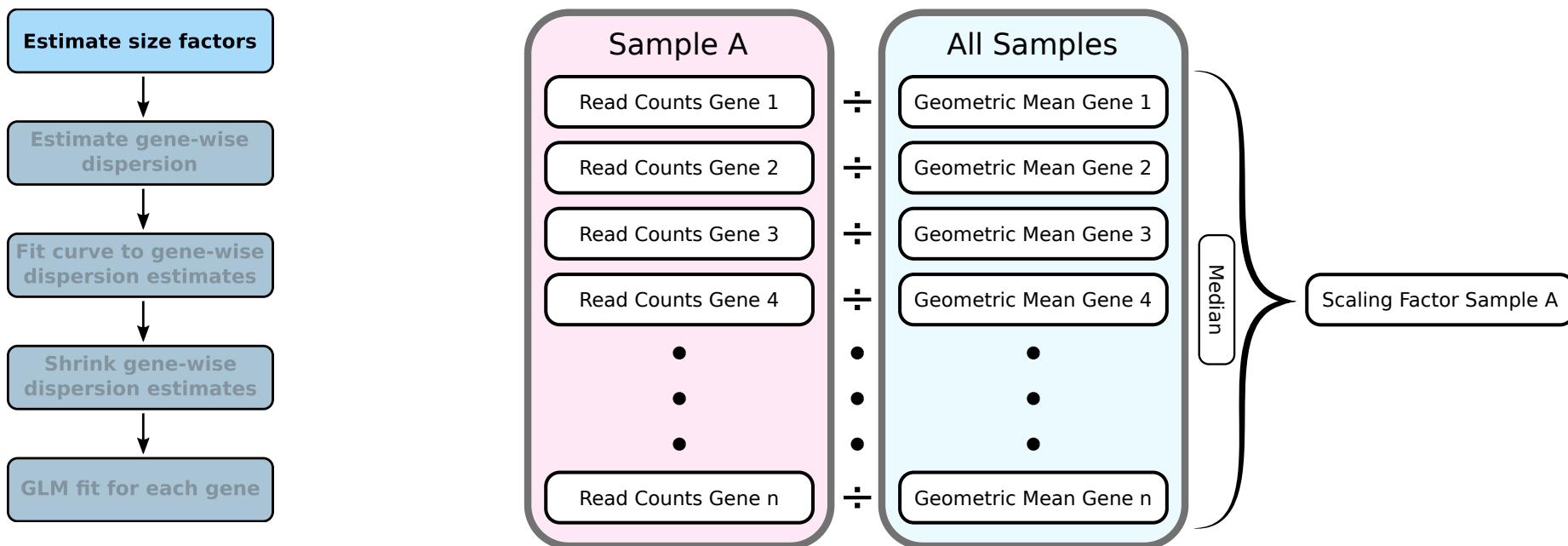
Highly expressed genes overrepresented at the cost of lowly expressed genes

“Composition Bias”

# Normalisation - Geometric mean scaling factor

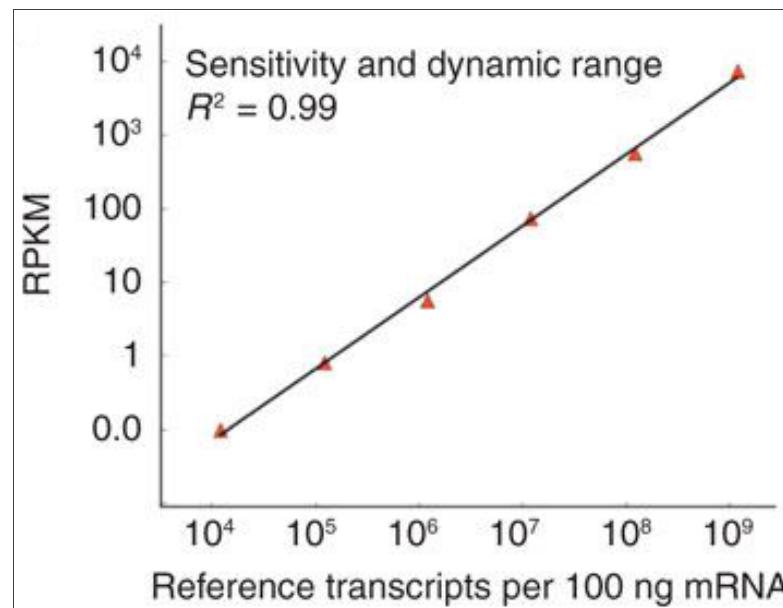
## ■ Used by DESeq2

1. For each gene calculate the geometric mean across all samples
2. For each gene in each sample, normalise by dividing by the geometric mean for that gene
3. For each sample calculate the scaling factor as the median of the normalised counts



# Differential Expression

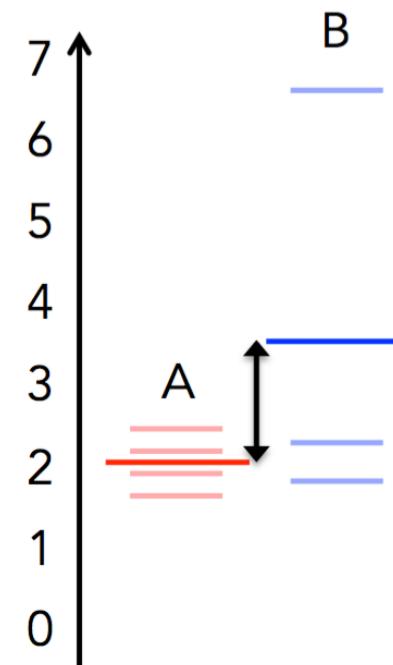
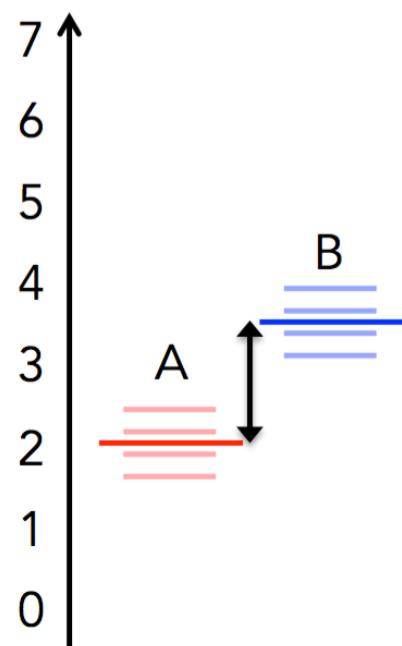
- Comparing feature abundance under different conditions
- Assumes linearity of signal
- When feature=gene, well-established pre- and post-analysis strategies exist



Mortazavi, A. et al (2008) Nature Methods

# Differential Expression

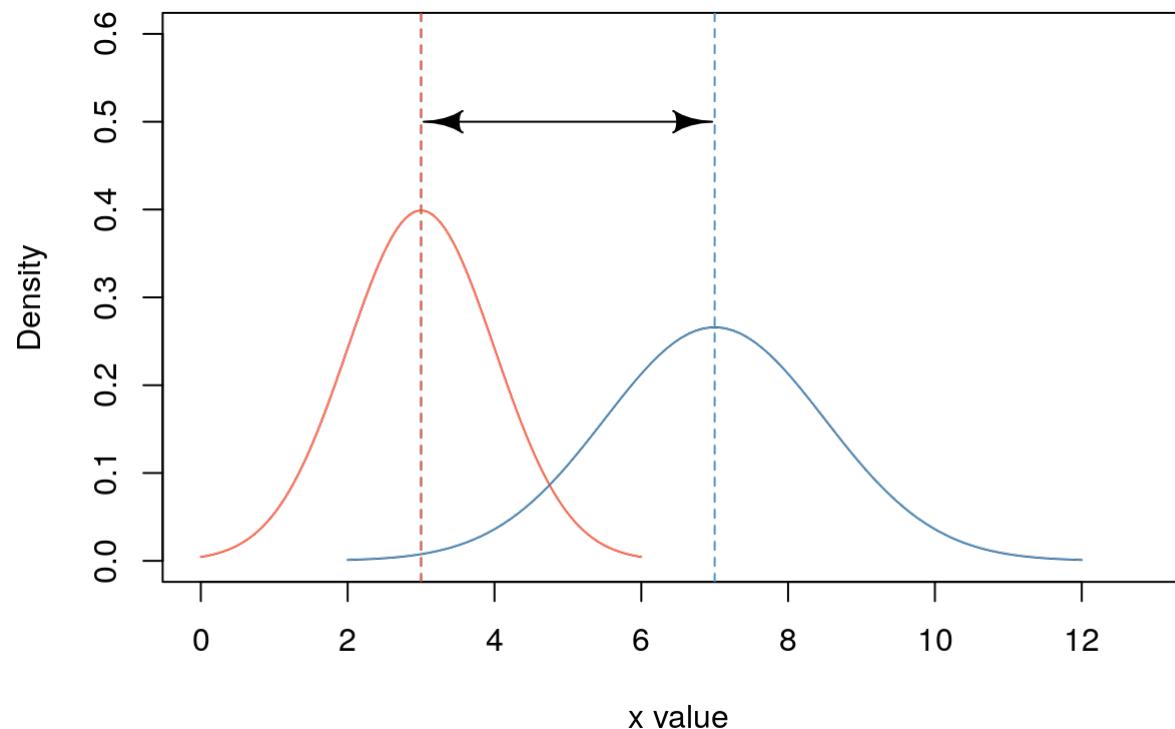
Simple difference in means



Replication introduces variation

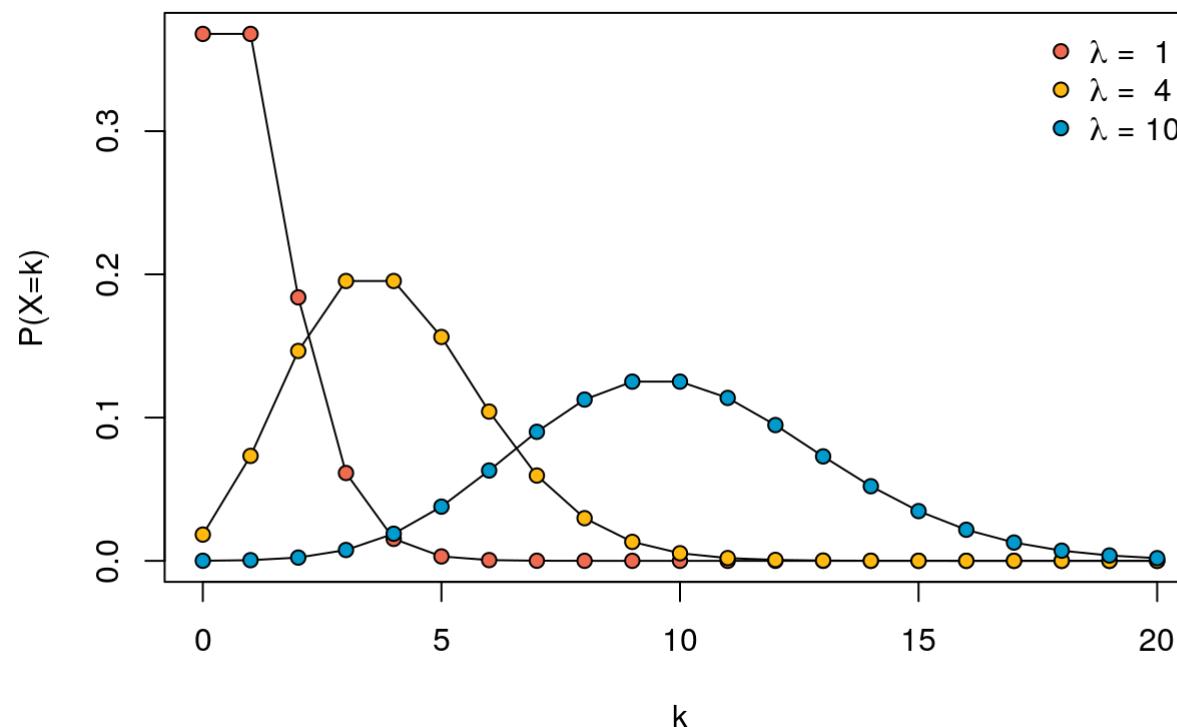
# Differential Expression - Modelling population distributions

- Normal (Gaussian) Distribution - t-test
- Two parameters - *mean* and *sd* ( $sd^2 = variance$ )
- Suitable for microarray data but not for RNAseq data

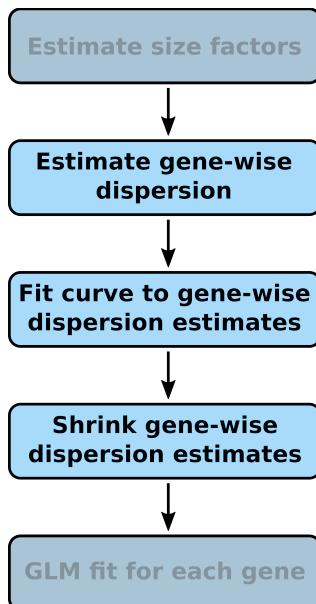


# Differential Expression - Modelling population distributions

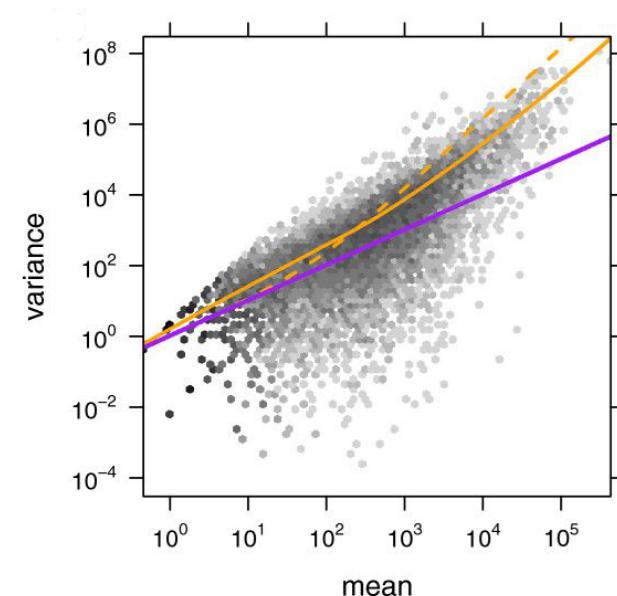
- Count data - Poisson distribution
- One parameter - *mean* ( $\lambda$ )
- $variance = mean$



# Differential Expression - Modelling population distributions



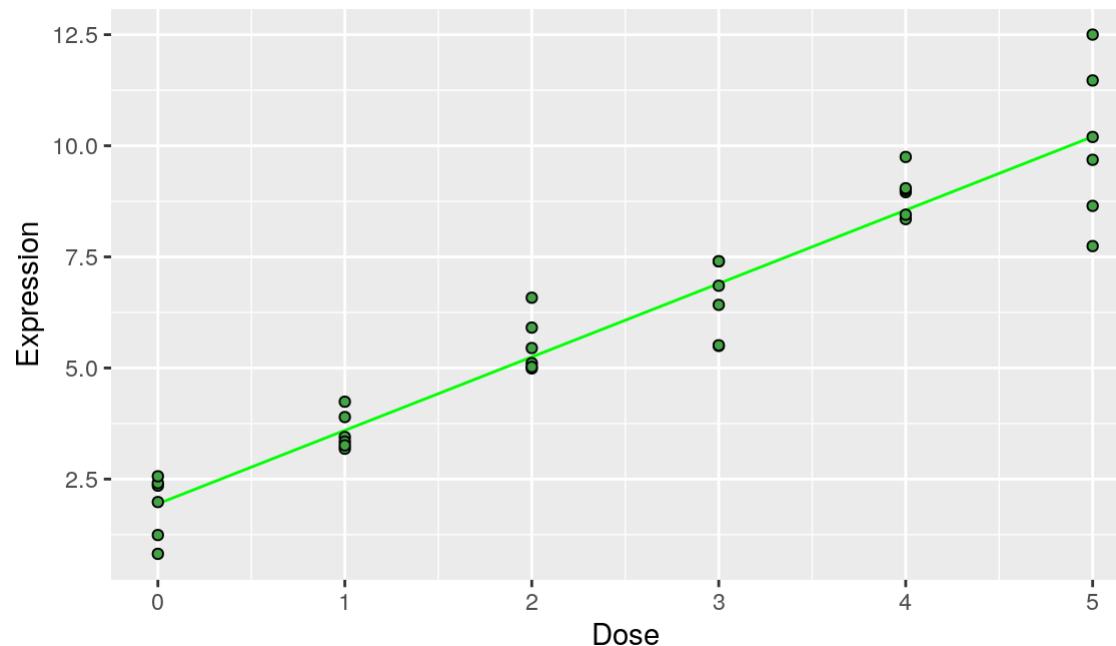
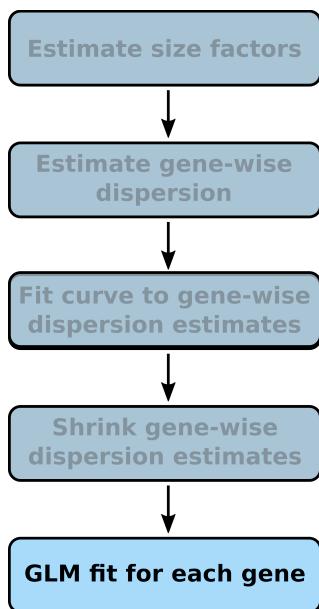
- Use the Negative Binomial distribution
- In the NB distribution *mean* not equal to *variance*
- Two parameters - *mean* and *dispersion*
- *dispersion* describes how *variance* changes with *mean*



Anders, S. & Huber, W. (2010) Genome Biology

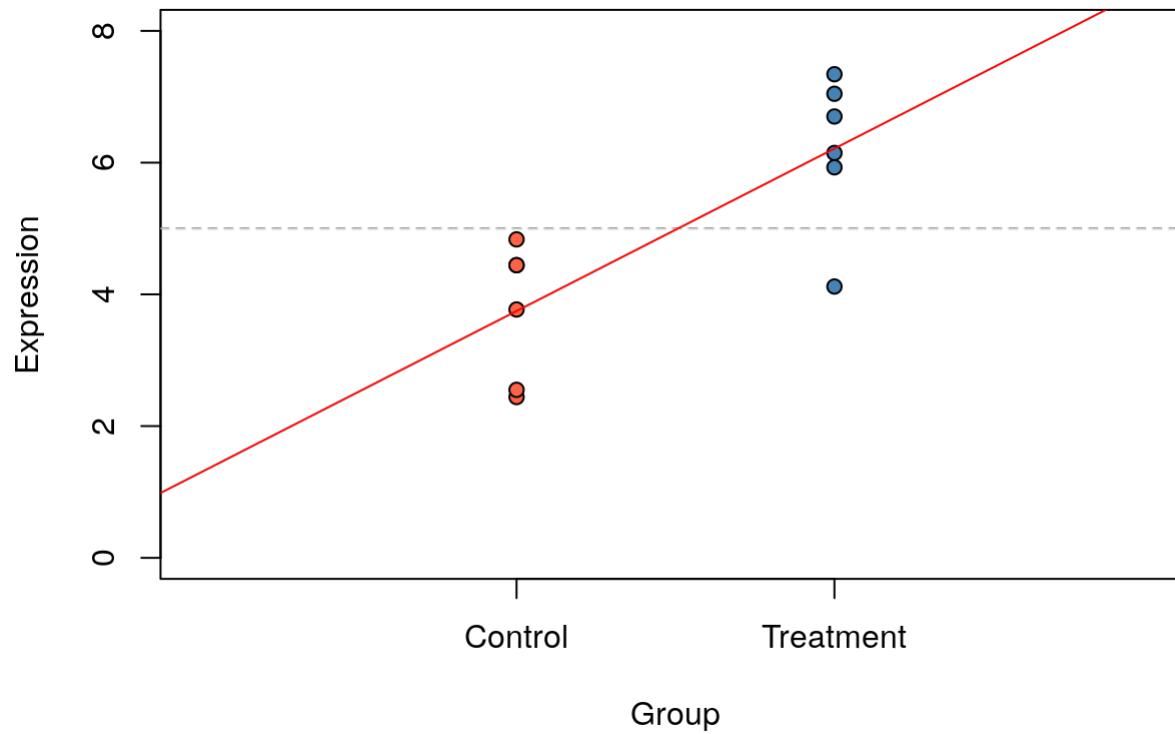
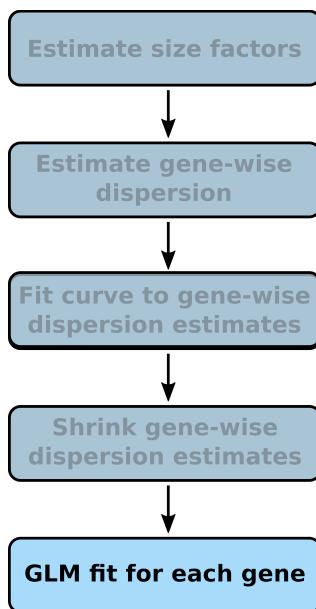
# Differential Expression - linear models

- Calculate coefficients describing change in gene expression
- Linear Model → Generalized Linear Model



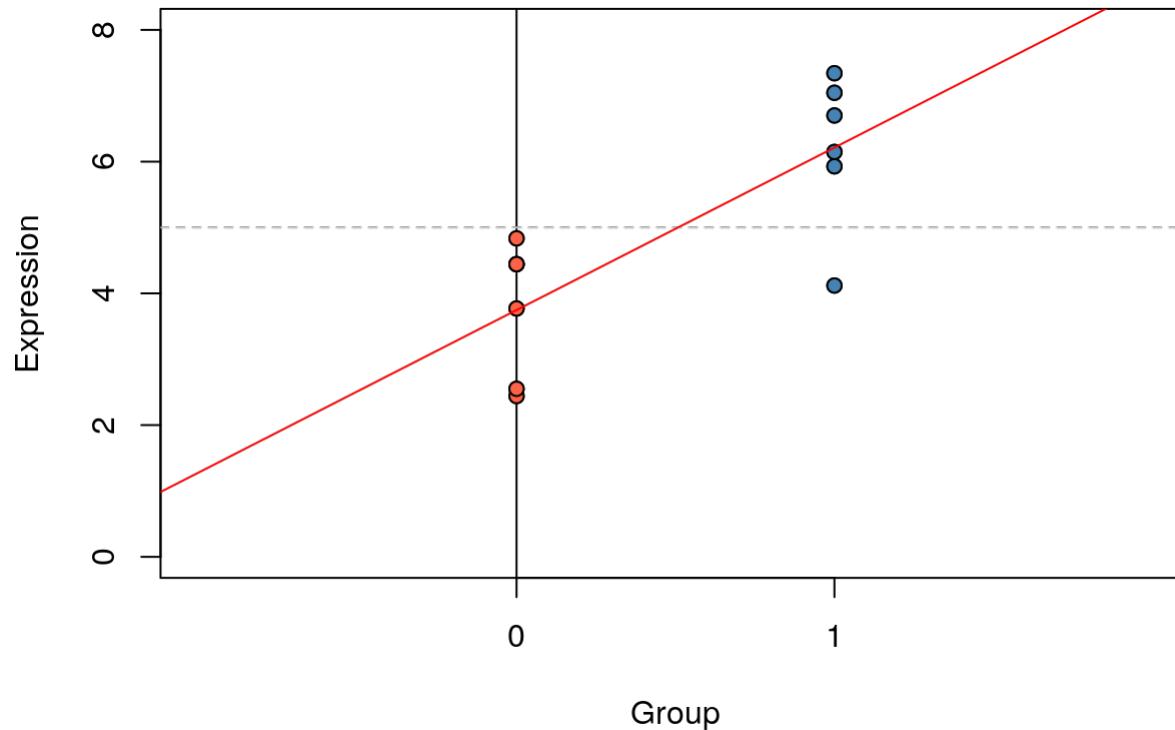
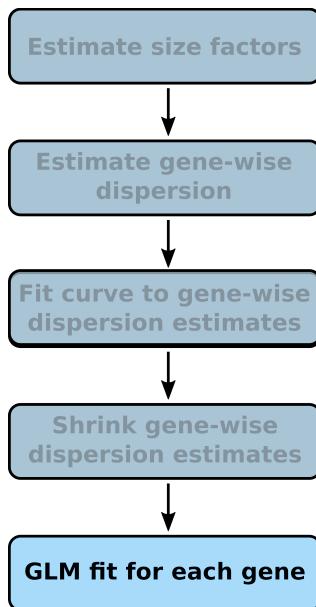
# Differential Expression - linear models

- Calculate coefficients describing change in gene expression
- Linear Model → General Linear Model

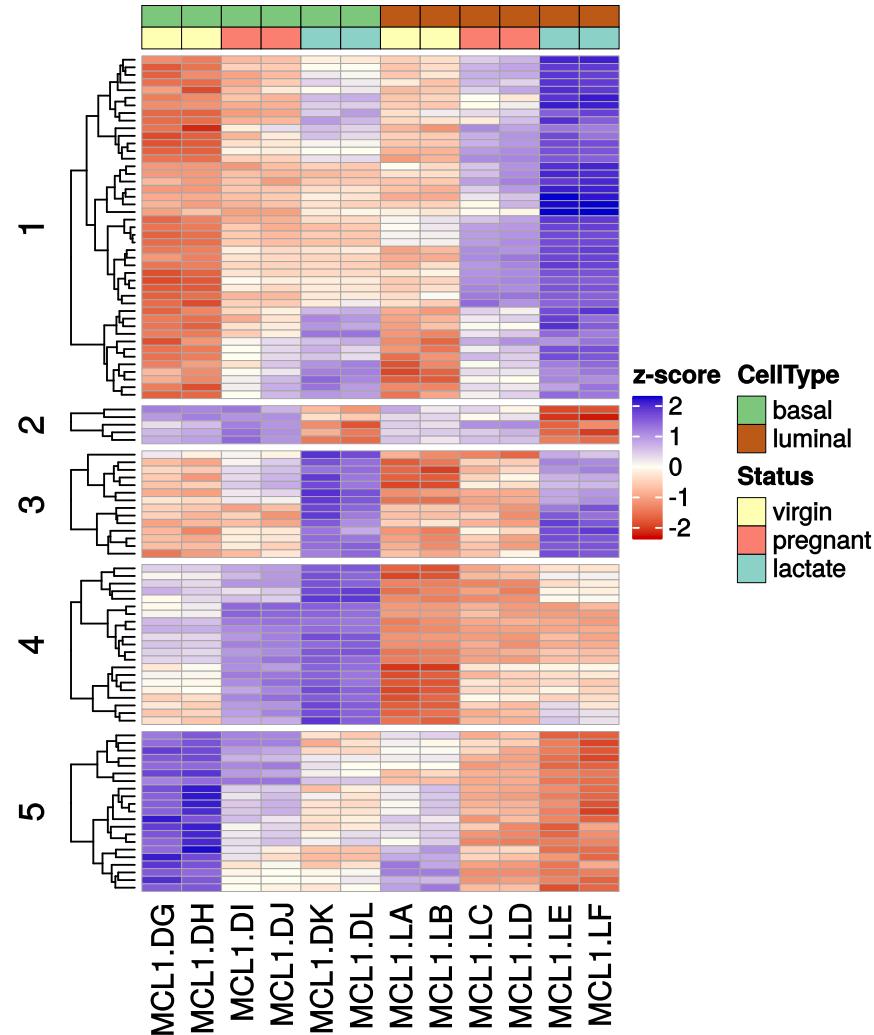


# Differential Expression - linear models

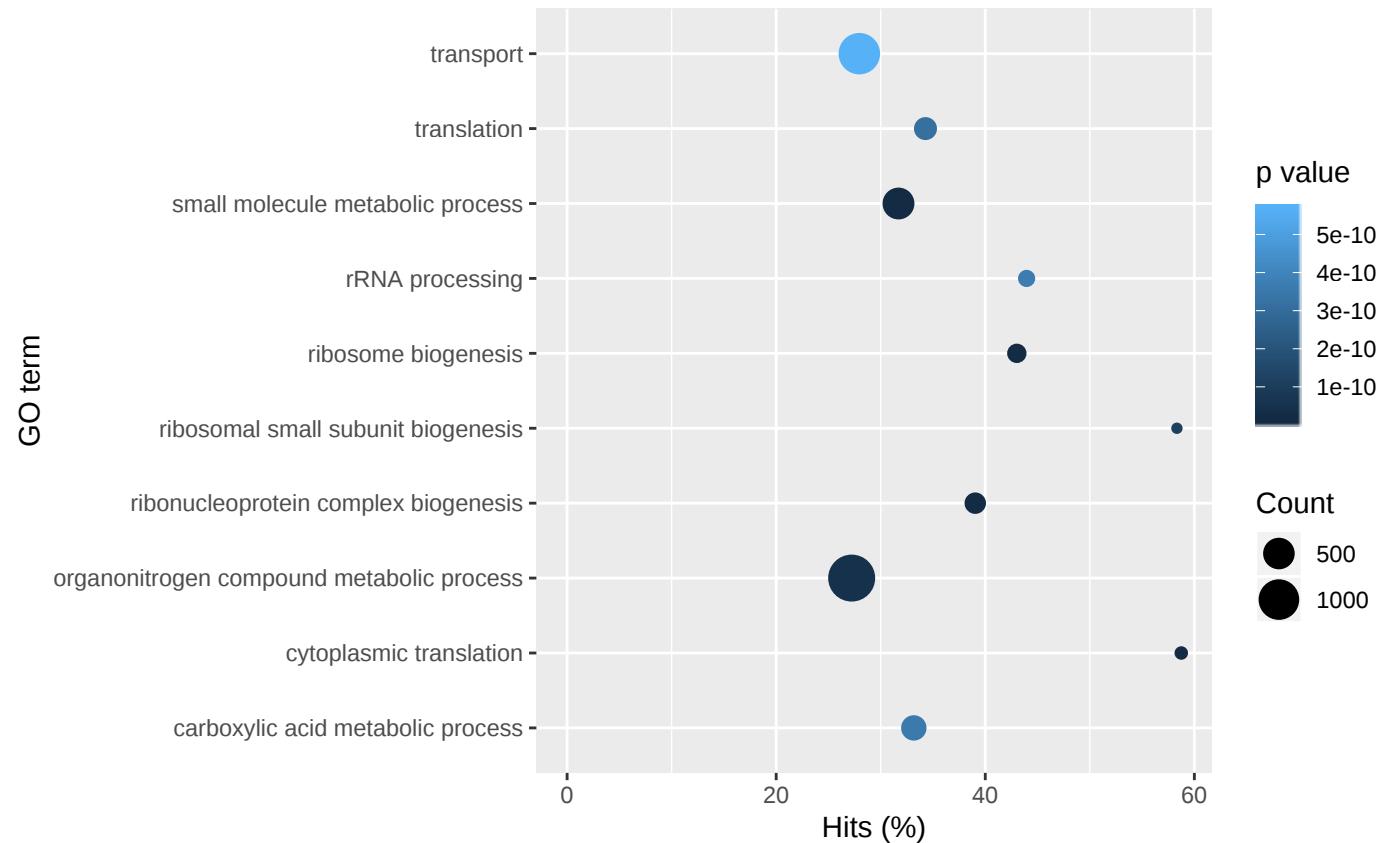
- Calculate coefficients describing change in gene expression
- Linear Model → General Linear Model



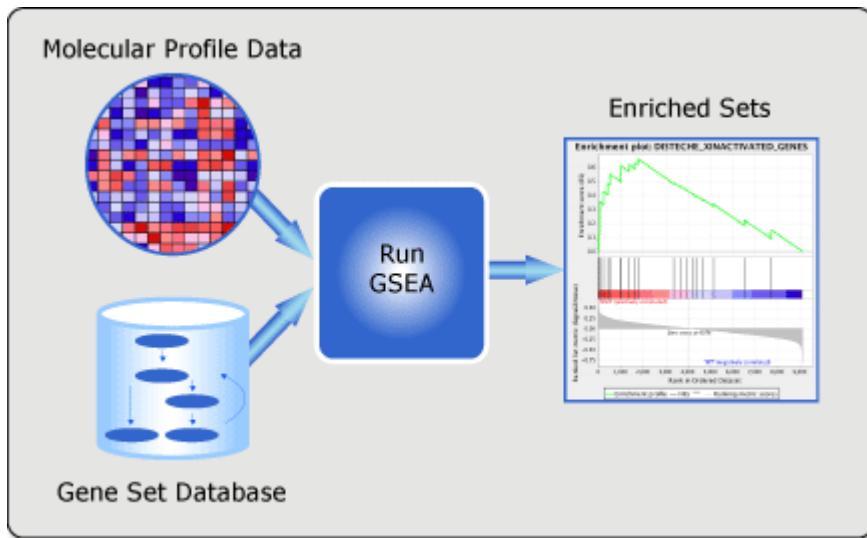
# Towards biological meaning - hierachical clustering



# Towards biological meaning - Gene Ontology testing



# Towards biological meaning - Gene Set Enrichment Analysis



<http://software.broadinstitute.org/gsea>

- ▶ **H** (hallmark gene sets, 50 gene sets)
- ▶ **C1** (positional gene sets, 326 gene sets)
  - ▶ by chromosome: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [X](#) [Y](#)
- ▶ **C2** (curated gene sets, 4762 gene sets)
  - ▶ **CGP** (chemical and genetic perturbations, 3433 gene sets)
  - ▶ **CP** (Canonical pathways, 1329 gene sets)
  - ▶ **CP:BIOCARTA** (BioCarta gene sets, 217 gene sets)
  - ▶ **CP:KEGG** (KEGG gene sets, 186 gene sets)
  - ▶ **CP:REACTOME** (Reactome gene sets, 674 gene sets)
- ▶ **C3** (motif gene sets, 836 gene sets)
  - ▶ **MIR** (microRNA targets, 221 gene sets)
  - ▶ **TFT** (transcription factor targets, 615 gene sets)
- ▶ **C4** (computational gene sets, 858 gene sets)
  - ▶ **CGN** (cancer gene neighborhoods, 427 gene sets)
  - ▶ **CM** (cancer modules, 431 gene sets)
- ▶ **C5** (GO gene sets, 5917 gene sets)
  - ▶ **BP** (GO biological process, 4436 gene sets)
  - ▶ **CC** (GO cellular component, 580 gene sets)
  - ▶ **MF** (GO molecular function, 901 gene sets)
- ▶ **C6** (oncogenic signatures, 189 gene sets)
- ▶ **C7** (immunologic signatures, 4872 gene sets)

# Towards biological meaning - Pathway Analysis

