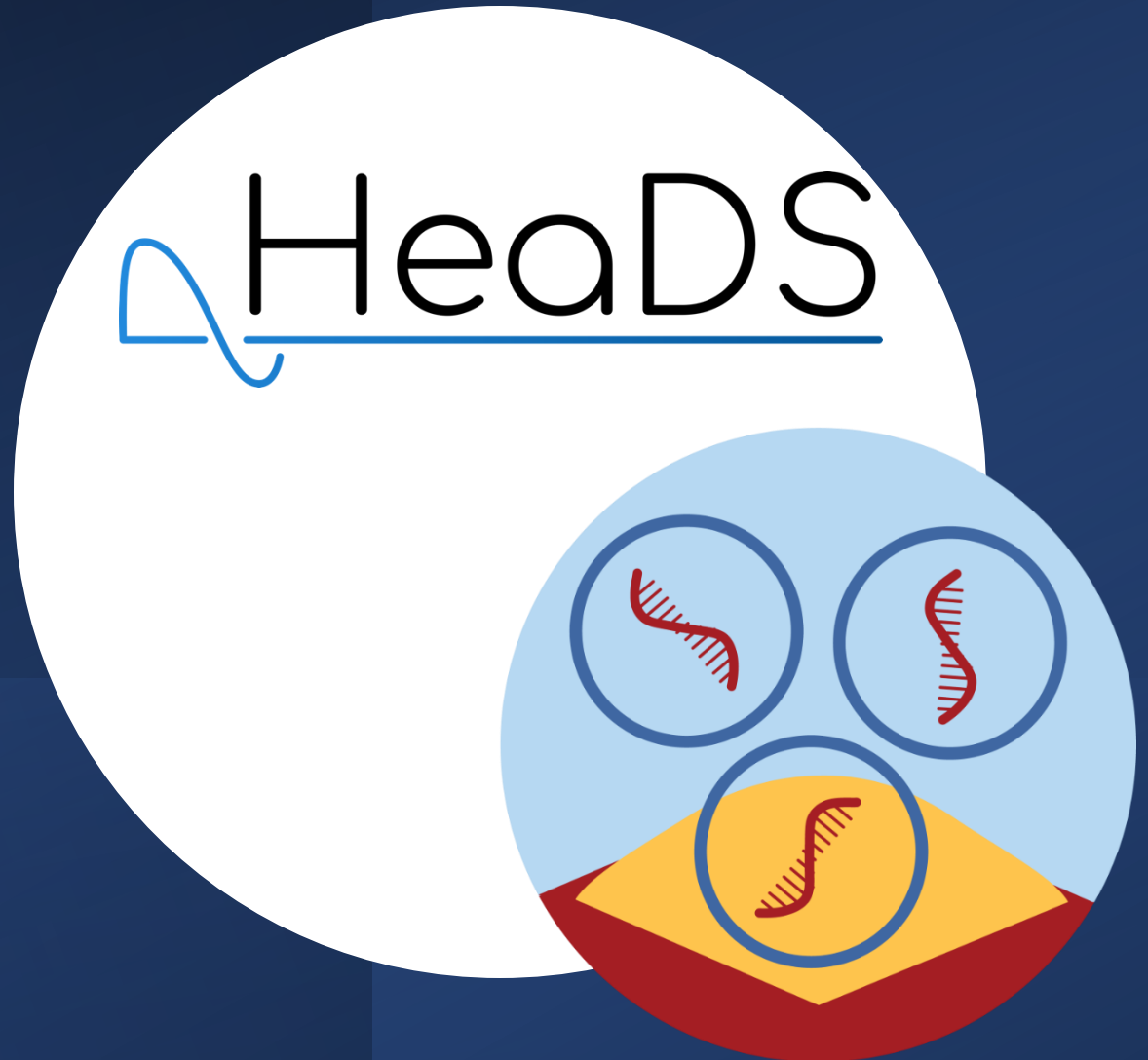


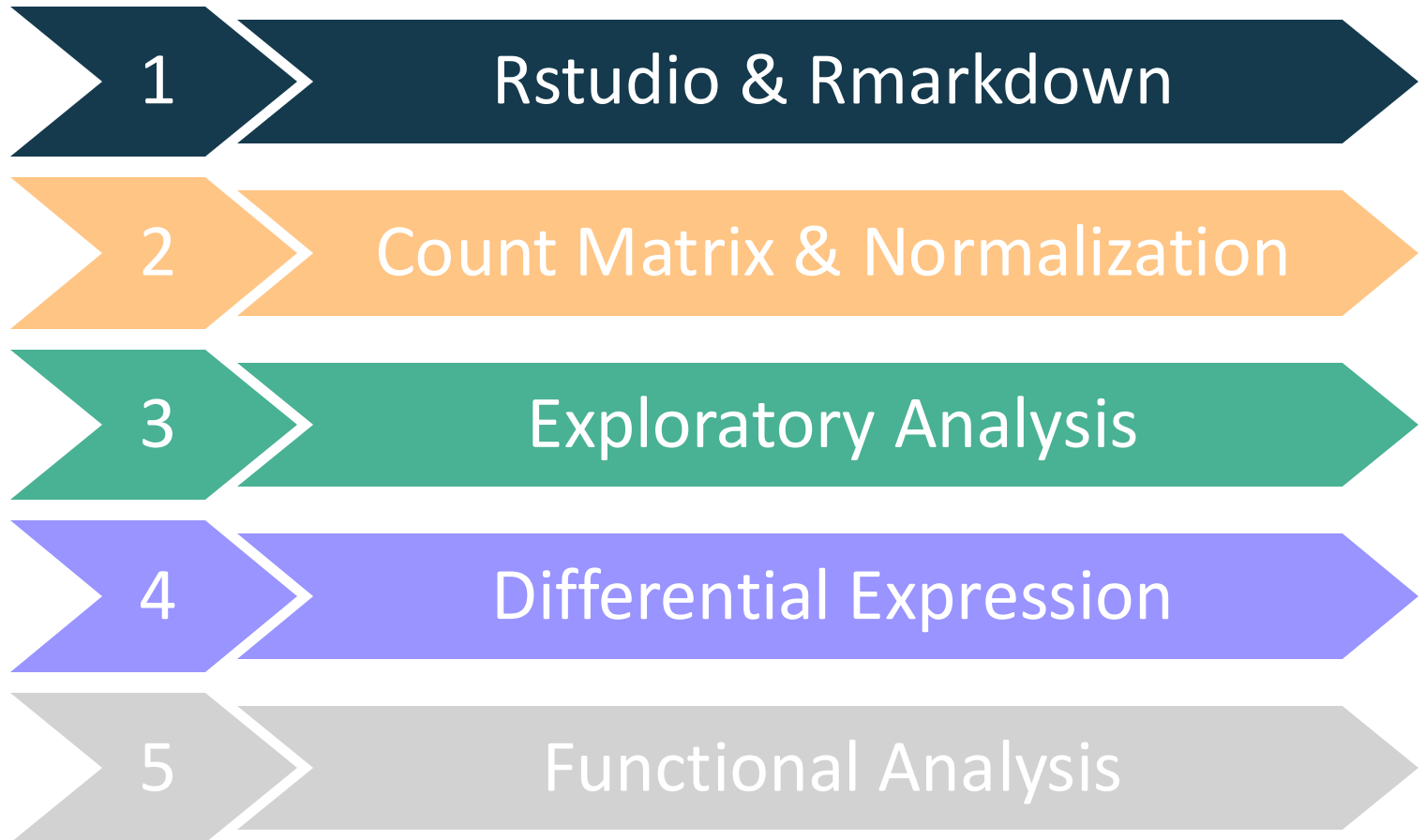
# Differential Expression Analysis

Center for Health Data Science



Health Data Science Sandbox

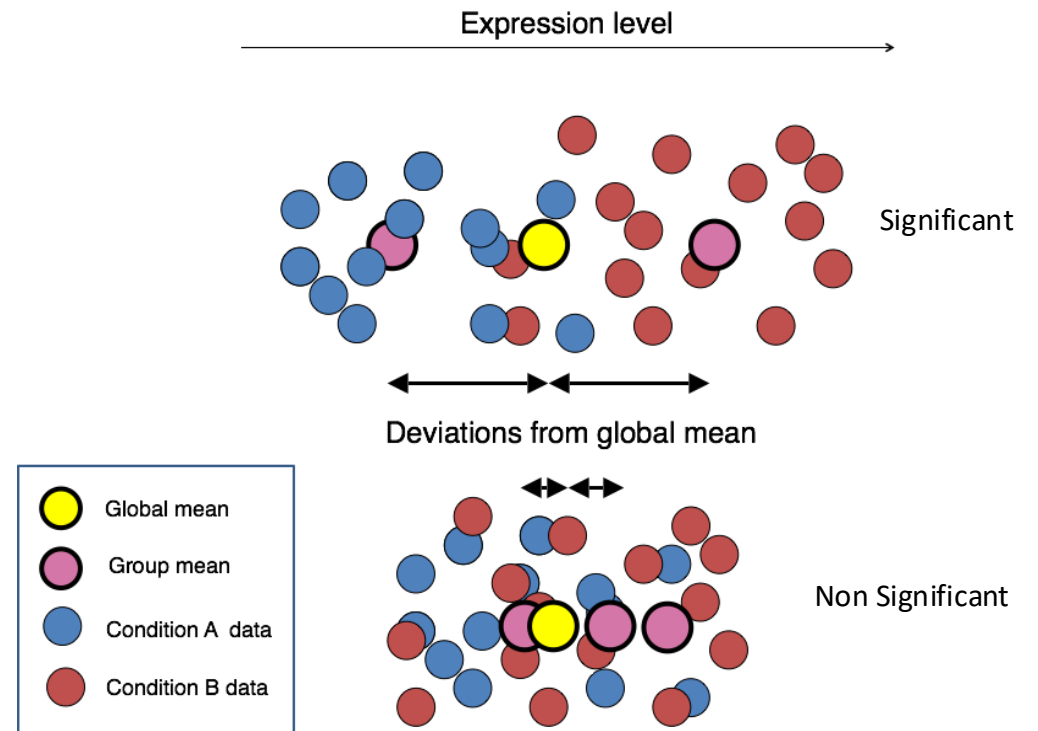
# Overview



# Differential Expression Analysis

Goal of differential expression analysis (DEA):

- Compare gene expression between conditions
  - Pathway -and ontology enrichment
  - Understand underlying biological mechanism
  - Explore effect of exposure or drug treatment
  - Gene target for translation medicine



# Differential Expression Analysis

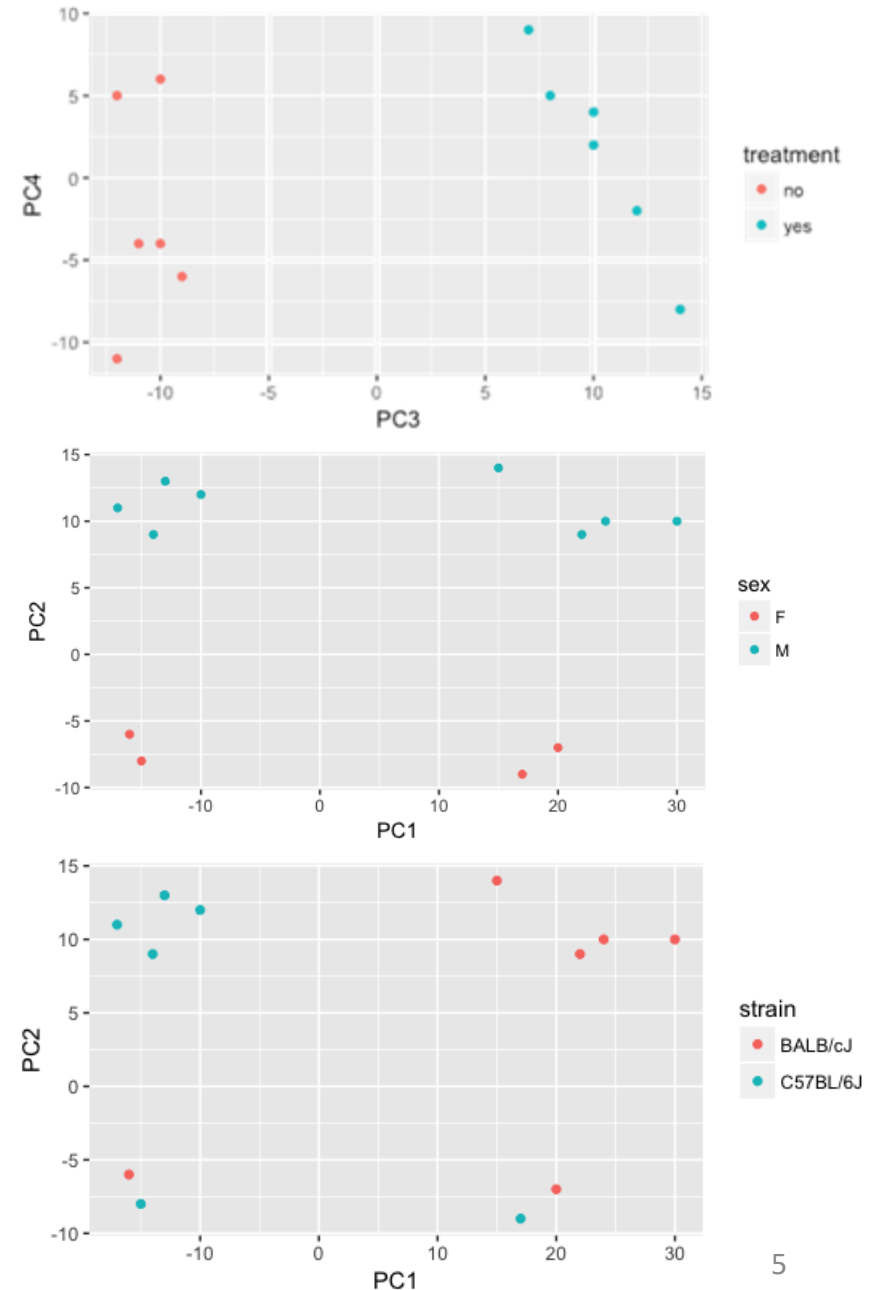
- What conditions you want to compare → **Contrast**
- Simple experiments are easy to compare (e.g. Treatment vs Control)
- Complicated experiments:
  - Multiple outcomes
  - Multiple explanatory variables to account for
  - Unpaired or paired samples
  - Batch and other technical artefacts

# DEA

- Define sources of variation for DE testing:
  - Meta data variables that defined PCs in your PCA
  - Your variable of interest (outcome variable)
  - Model counts and perform statistical test
- Make a **design matrix** including:
  - outcome variable (treatment)
  - explanatory variables (sex, strain)
  - confounding variables (batch)

```
design = ~ treatment
```

```
design = ~ sex + strain + treatment
```



# DEA contrast designs

- One factor with two levels

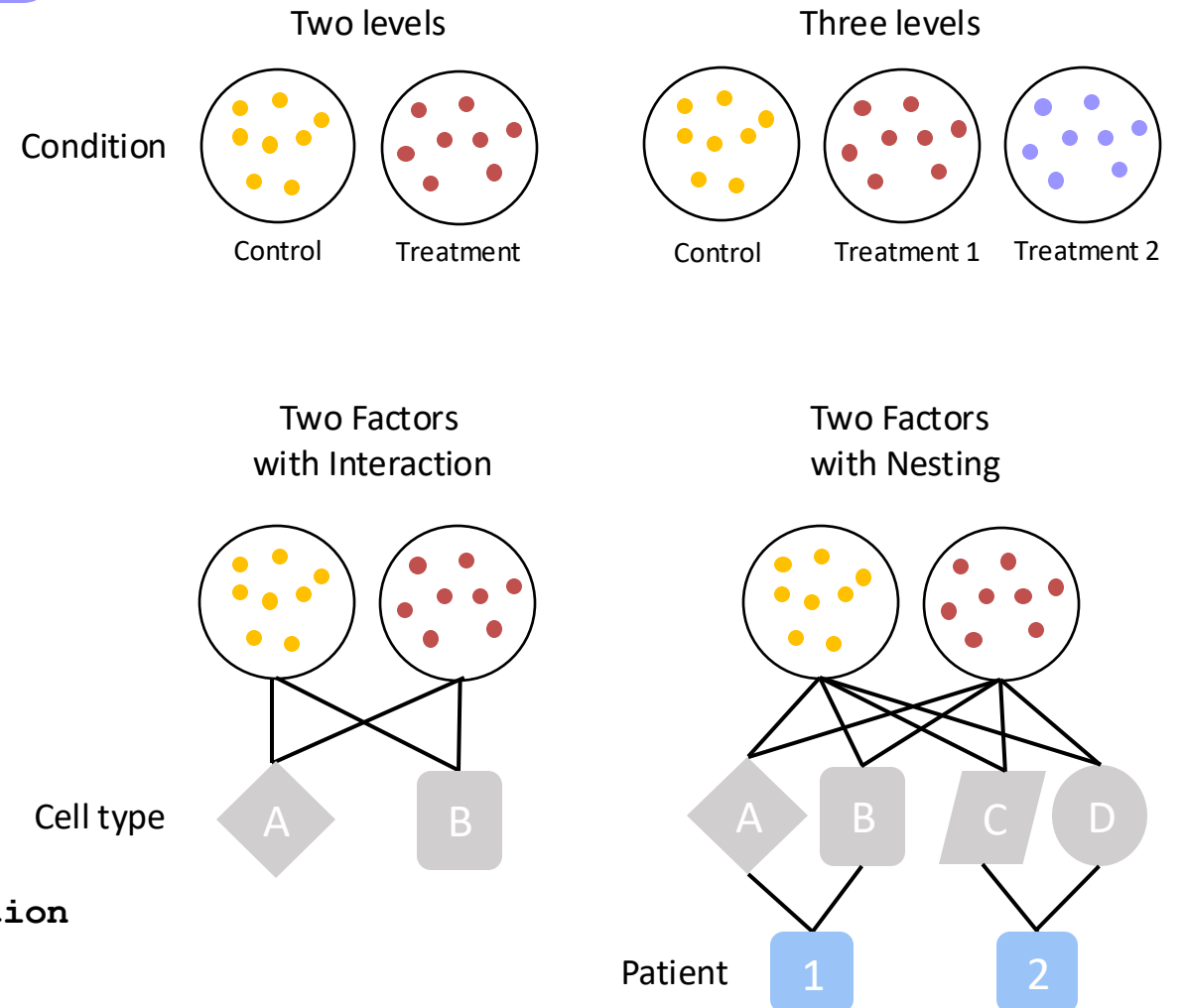
$\sim 0 + \text{Condition}$

- One factor with n levels
- Two factors with n levels

$\sim 0 + \text{Celltype} + \text{Condition}$

- Two factors with interaction
- Two factors with nesting

$\sim 0 + \text{Celltype} + \text{Condition} + \text{Celltype}:\text{Condition}$



# DESeq2 Workflow

- 1. Define sources of variation for testing (variables)
- **2. DESeq2 Differential Gene Expression Workflow**

**DESeq()**

- Four steps:
  1. Estimate size factors (same as normalization!)
  2. Estimate gene-wise dispersion
  3. Fit Negative Binomial **Generalized Linear Model**
  4. Post hoc test for **LFC** and **significance (p-val)**

1. Estimate Size Factors



2. Estimate Gene-Wise Dispersion



3. GLM Fit for Each Gene



4. Post Hoc Test for Significance

# DESeq2 Workflow

## 1. Estimate Size Factors

- Normalization for sample library size and RNA composition
- Size factors are used as parameters in the generalized linear models (GLMs)
- *We went through the steps of this in section 5 (RNAseq counts)*

```
estimateSizeFactors ()
```



# DESeq2 Workflow

## 2. Estimate Gene-Wise Dispersion

- To estimate DEG → to account for replicate variation
- **Dispersion** compares variance to the mean

**estimateDispersions ()**

$$dispersion = \frac{var - \mu}{\mu^2}$$

Which difference is more *striking* between genes?

Gene	Sample X	Sample Y
A	200	100
B	20	10

**Gene A:**

$$200 - 100 = 100$$

$$200 / 100 = 2$$

**Gene B:**

$$20 - 10 = 10$$

$$20 / 10 = 2$$

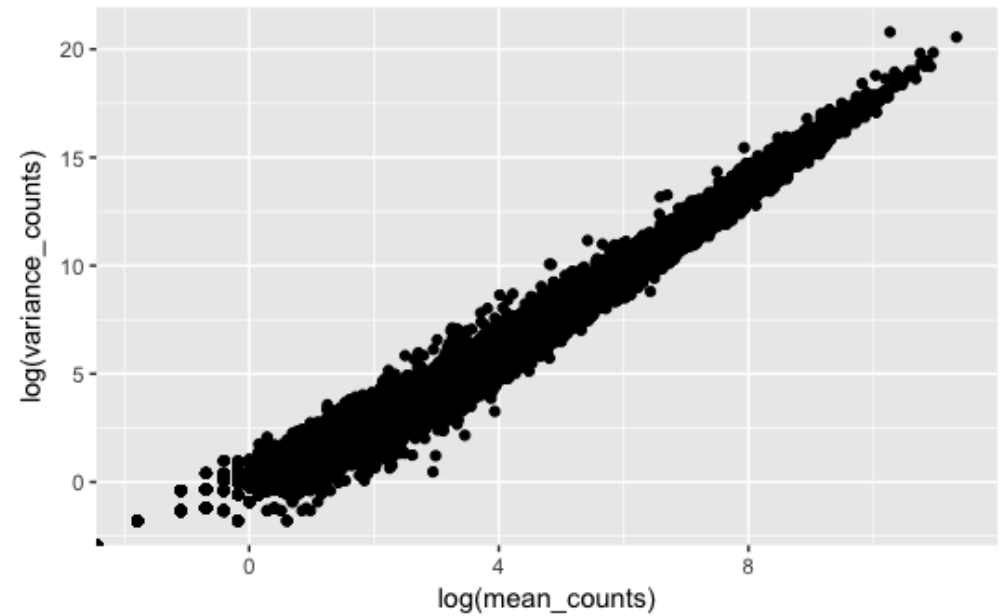
# DESeq2 Workflow

## 2. Estimate Gene-Wise Dispersion

- To estimate DEG → to account for replicate variation
- Replicate variation → account for dispersion

`estimateDispersions()`

$$dispersion = \frac{var - \mu}{\mu^2}$$



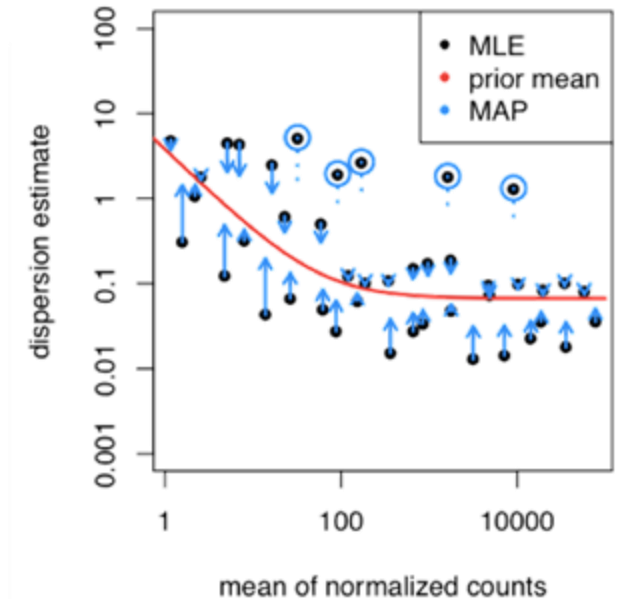
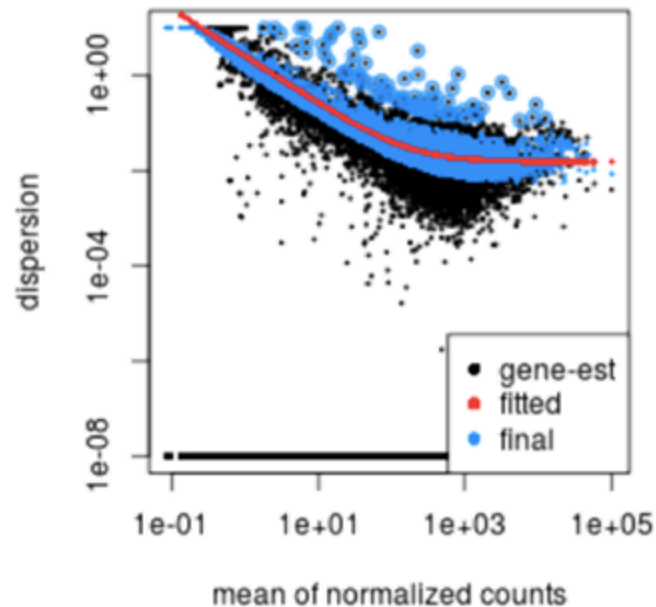
# DESeq2 Workflow

## 2. Estimate Gene-Wise Dispersion

- Dispersion is adjusted for lowly expressed genes

`estimateDispersions()`

$$dispersion = \frac{var - \mu}{\mu^2}$$



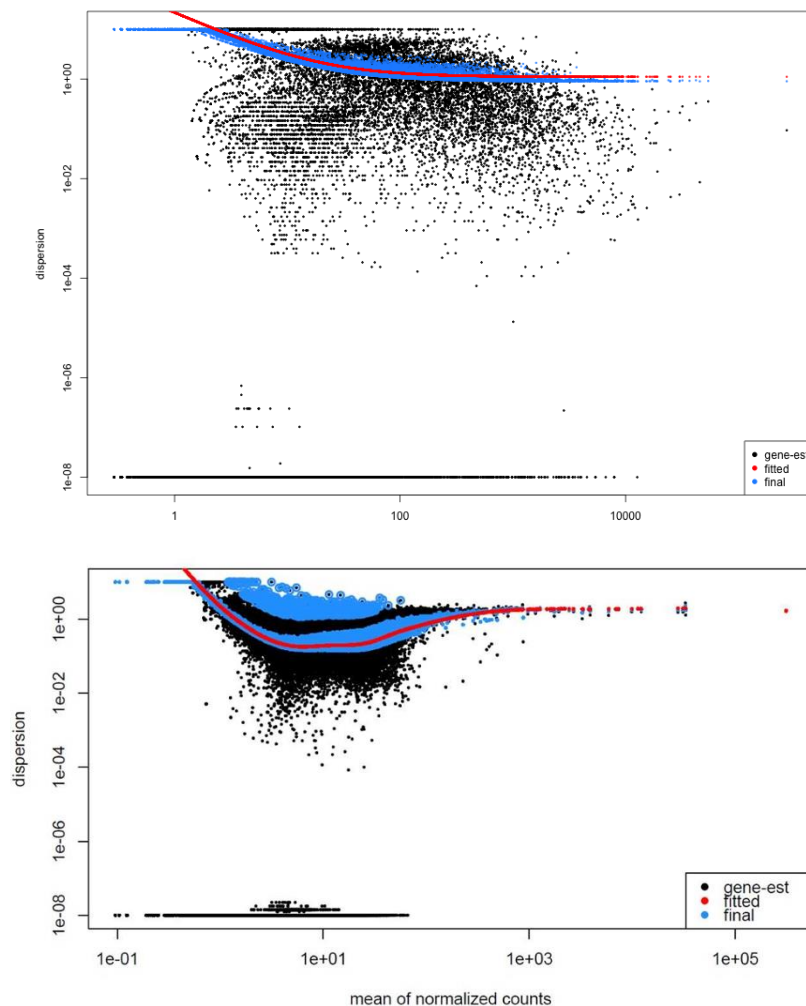
# DESeq2 Workflow

## 2. Estimate Gene-Wise Dispersion

- Worrisome dispersions

`estimateDispersions()`

$$dispersion = \frac{var - \mu}{\mu^2}$$



Bad fit!

Outlier or  
contamination?

# DESeq2 Workflow

## 3. GLM Fit for Each Gene

- Fit a Generalized Linear Model (Negative Binomial)
- Model takes into account dispersion and size factors
- Post hoc test: *Wald test* or *likelihood ratio test* to obtain fold changes

`nbinomWaldTest()`

raw count for gene  $i$ , sample  $j$

The mean is taken as “normalized counts” scaled by a normalization factor

one dispersion per gene

$$K_{ij} \sim \text{NB}(s_{ij}q_{ij}, \alpha_i)$$
$$\mu_{ij} = s_j q_{ij}$$

# Exercise

Let's Do Differential Expression Analysis:

Notebook:

- *07a\_DEA.Rmd*



# DESeq2 Workflow

## 4. Post Hoc Test (Wald Test)

Calculation of Log<sub>2</sub> Fold Changes (**LFC**) and their statistical significance:

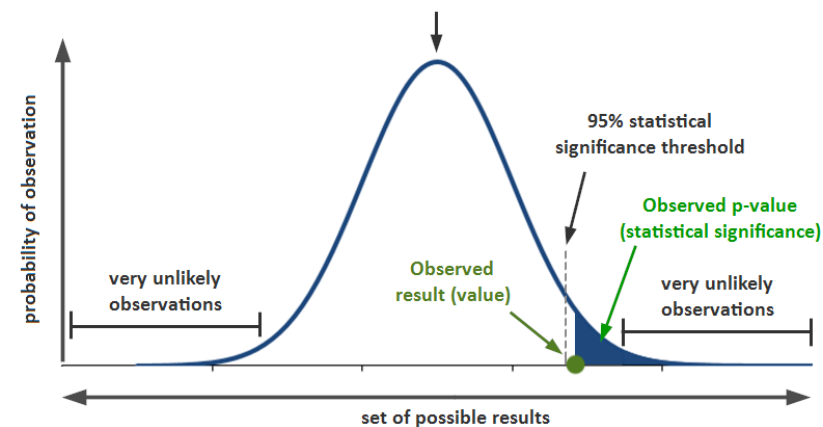
- **Fold change:** ratio of mean expression levels between conditions
  - How big the change is
- **Wald test P-value:** probability of observing the fold change if it was due by chance
  - How significant the change is

$$LFC = \log_2(\text{mean expr. gene}_i \text{ in cond}_A) - \log_2(\text{mean expr. gene}_i \text{ in cond}_B)$$

$$LFC = \log_2(q_{ij}) = \sum_r x_{ir} \beta_{ir}$$

GLM coefficient for each explanatory variable

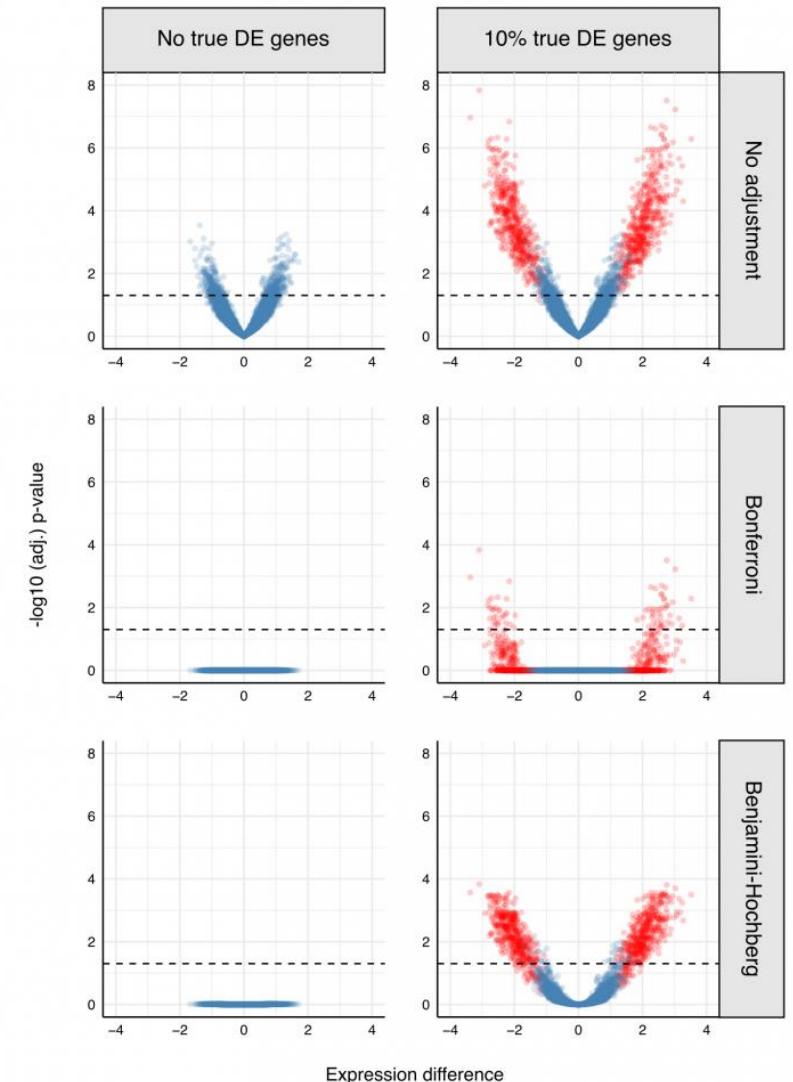
Design matrix



# DEA

## The multiple testing problem

- DE testing of thousands of genes will increase the number of false positives
- Bonferroni (or B-H) correction for adjusted p-values
- Filter significant DE between conditions
  - A very small change can be significant but is it biologically interesting?
  - Absolute LFC > 1 (or < -1) & adjusted p-value < 0.05
- Visualization with volcano plots, heatmaps, MA plots





# Exercise

Let's check out DEA results!

Notebook:

- *07b\_hypothesis\_testing.Rmd*

