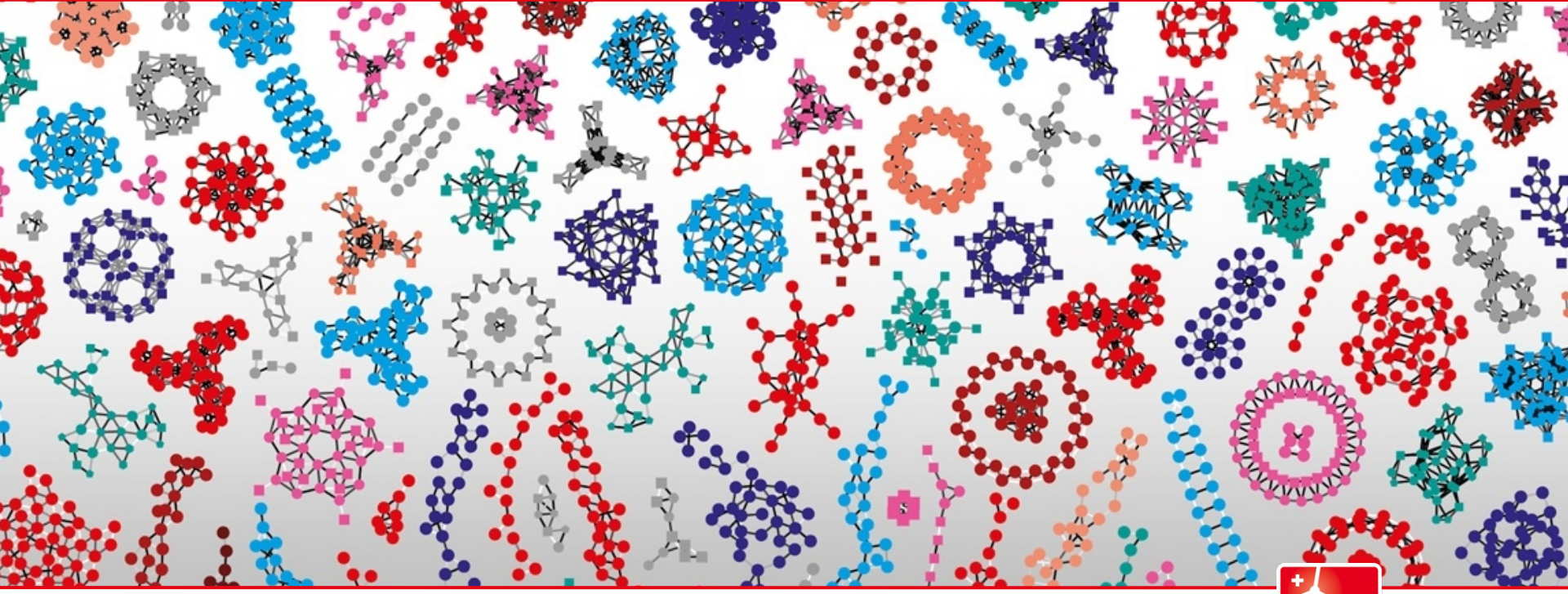
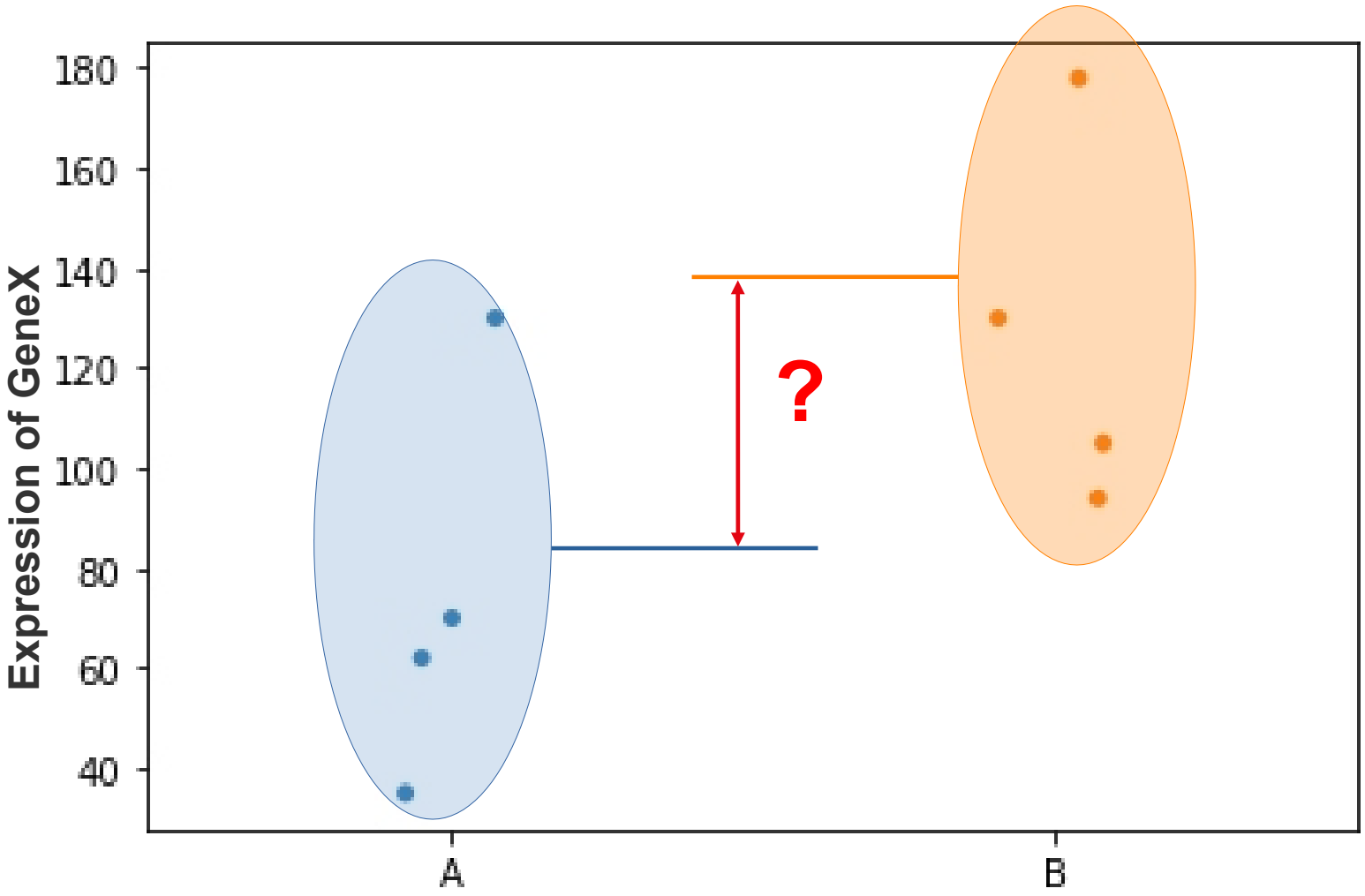


Swiss Institute of Bioinformatics

Introduction to RNA-Seq – Differential Expression

Wandrille Duchemin

DE – the goal

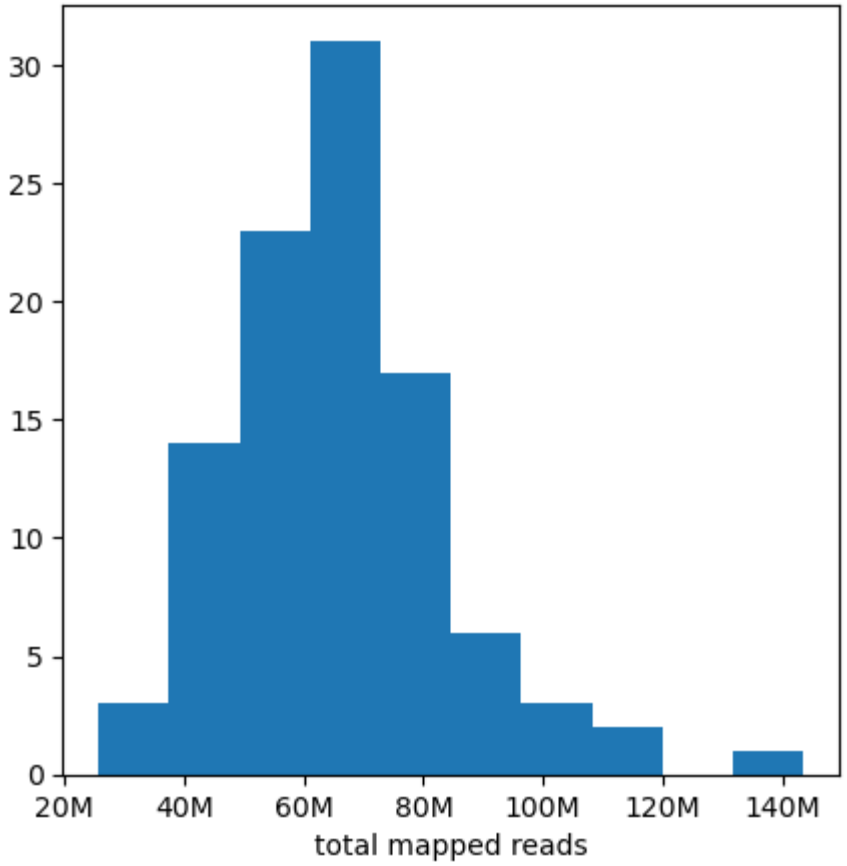


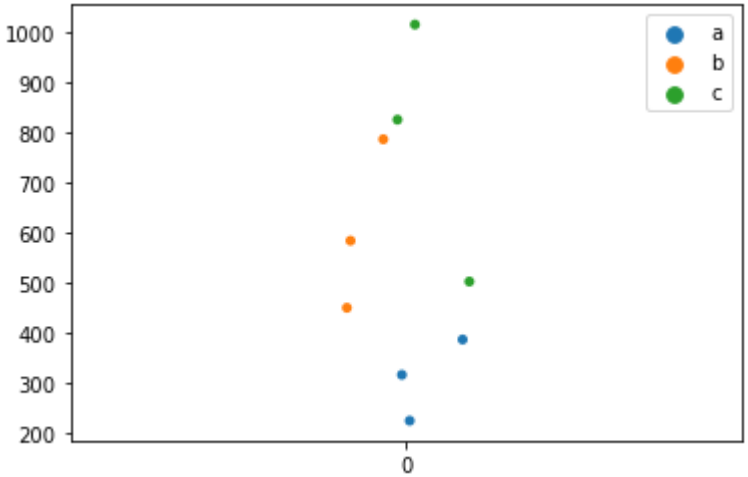
DE - challenges for RNA-Seq

- **Sequencing depth varies across libraries**
- **High dynamic range**
- **Limited number of samples**
- **Large number of genes**

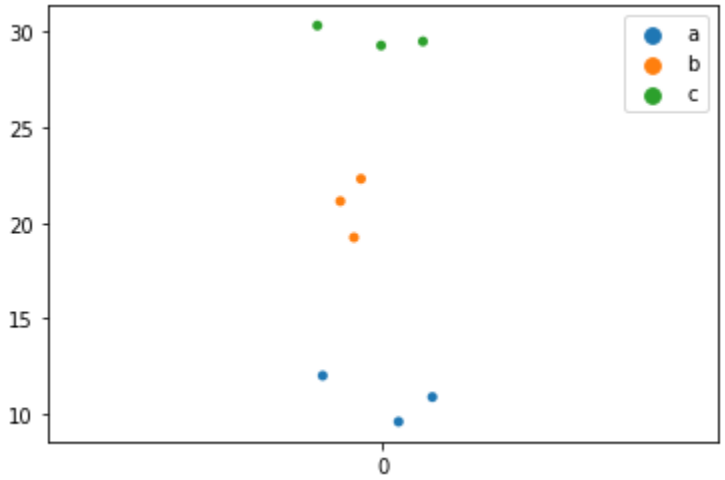
DE - challenges for RNA-Seq

- **Sequencing depth varies across libraries**



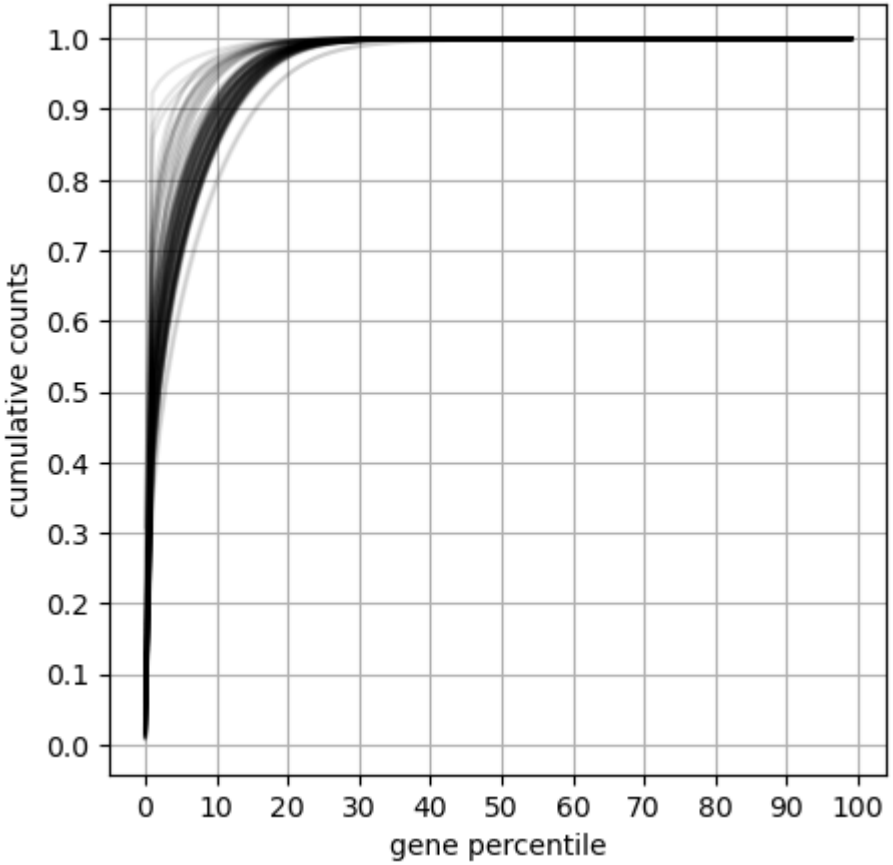


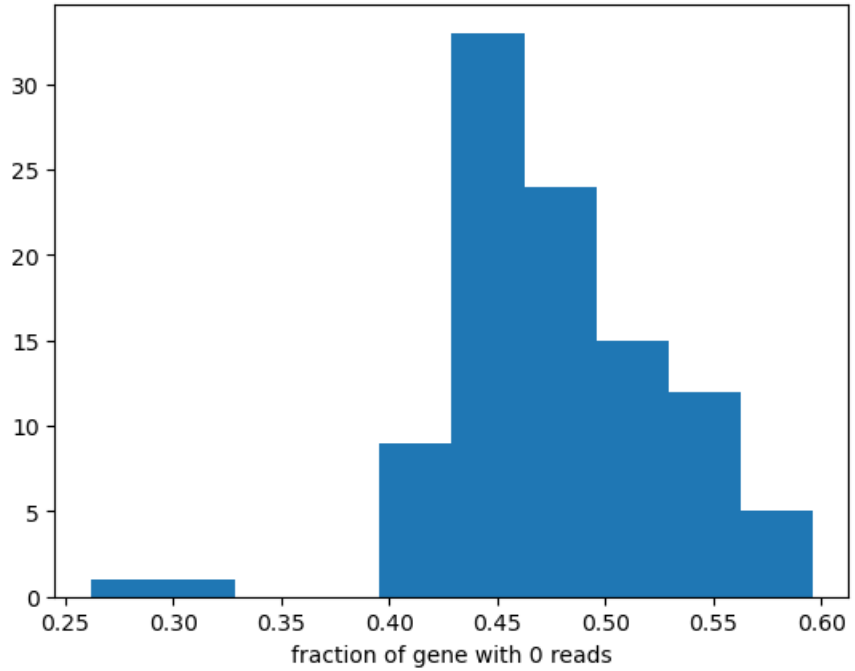
Normalization



DE - challenges for RNA-Seq

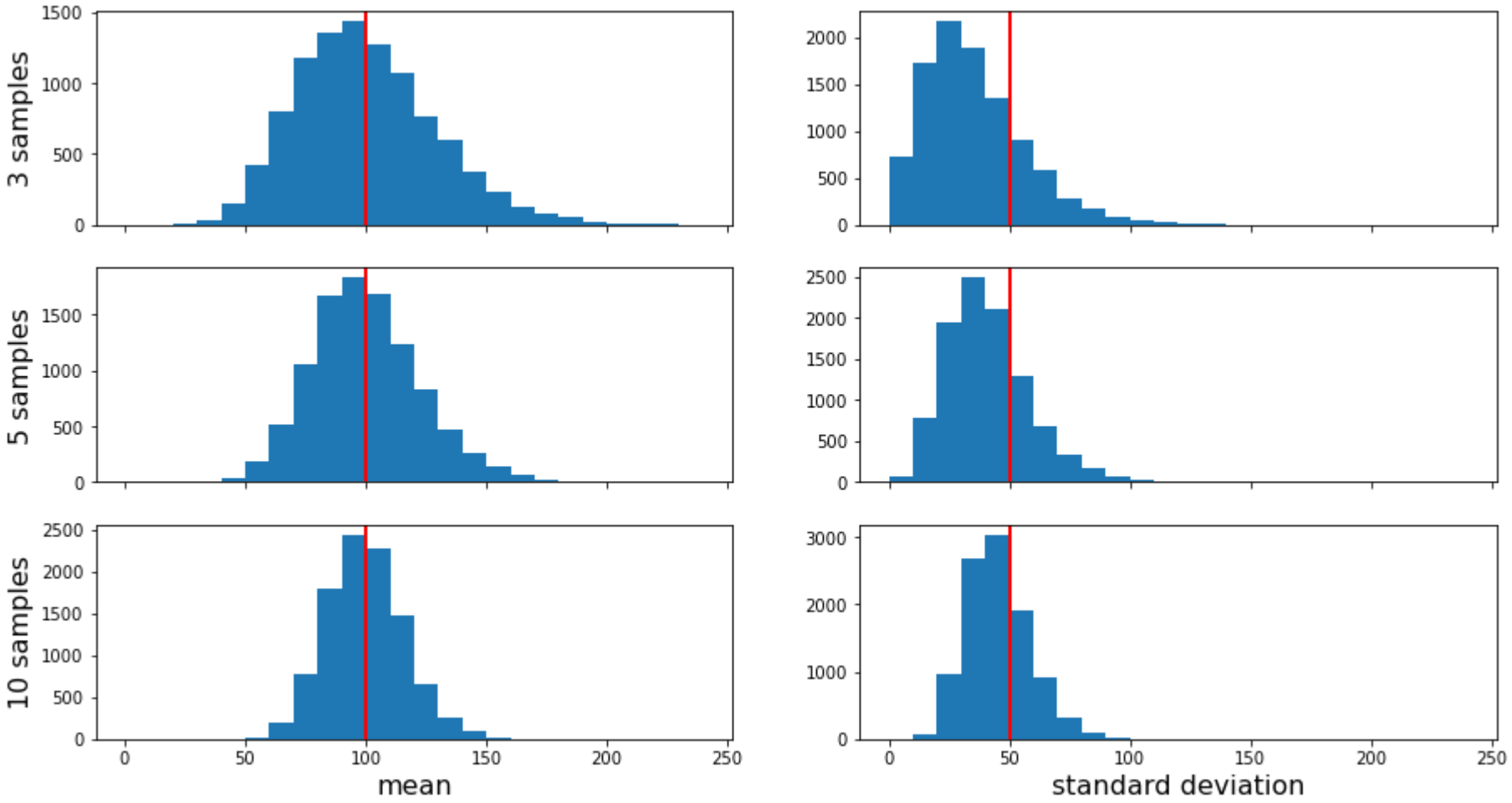
- **High dynamic range**





DE - challenges for RNA-Seq

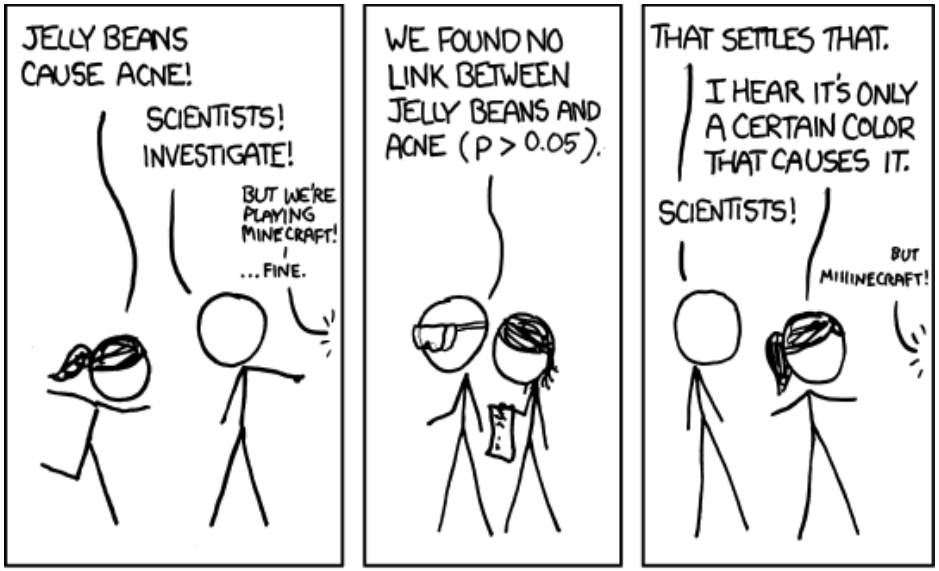
- **Limited number of samples**

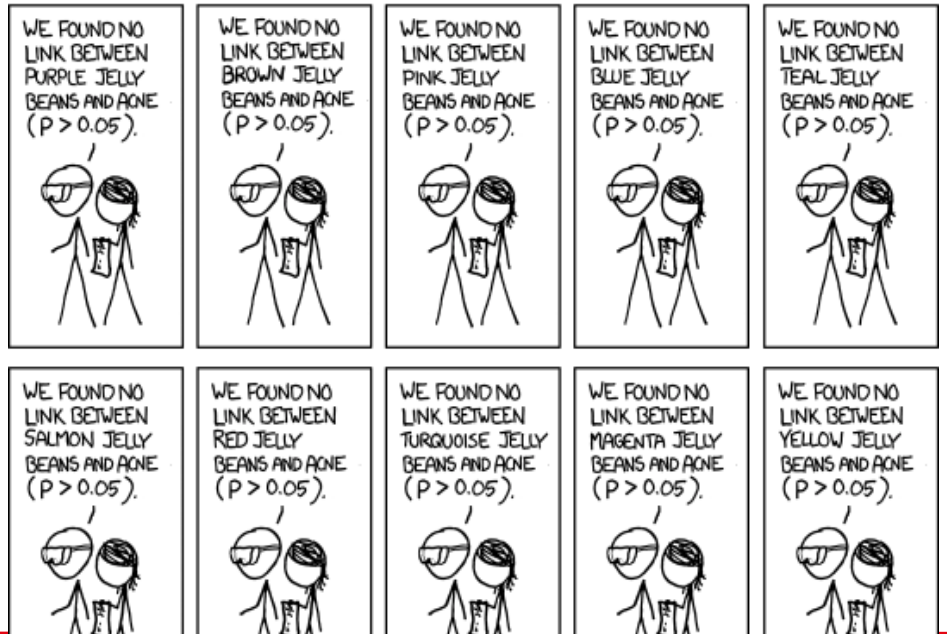


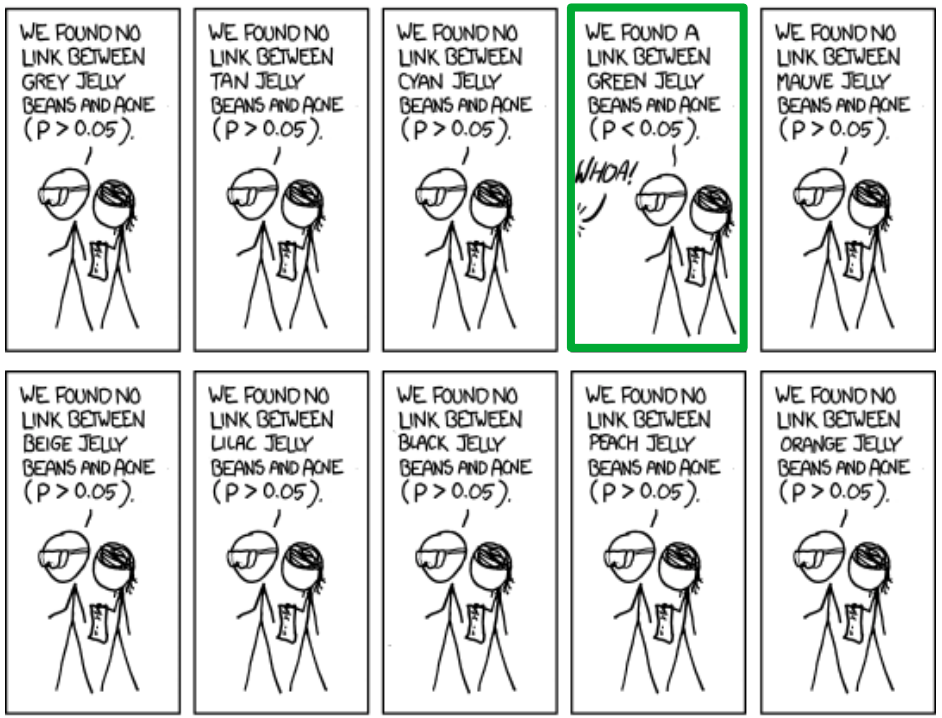
DE - challenges for RNA-Seq

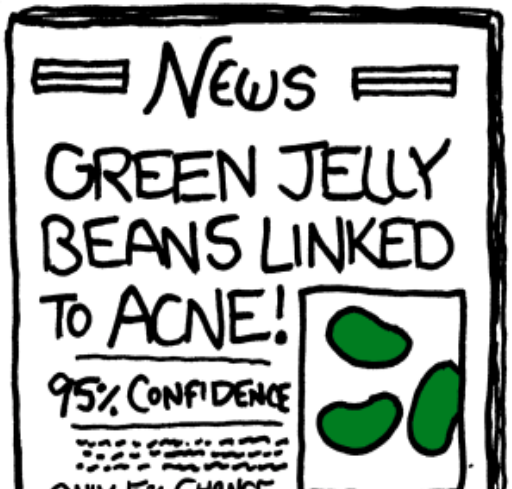
- **Large number of genes to test**

DE - challenges for RNA-Seq



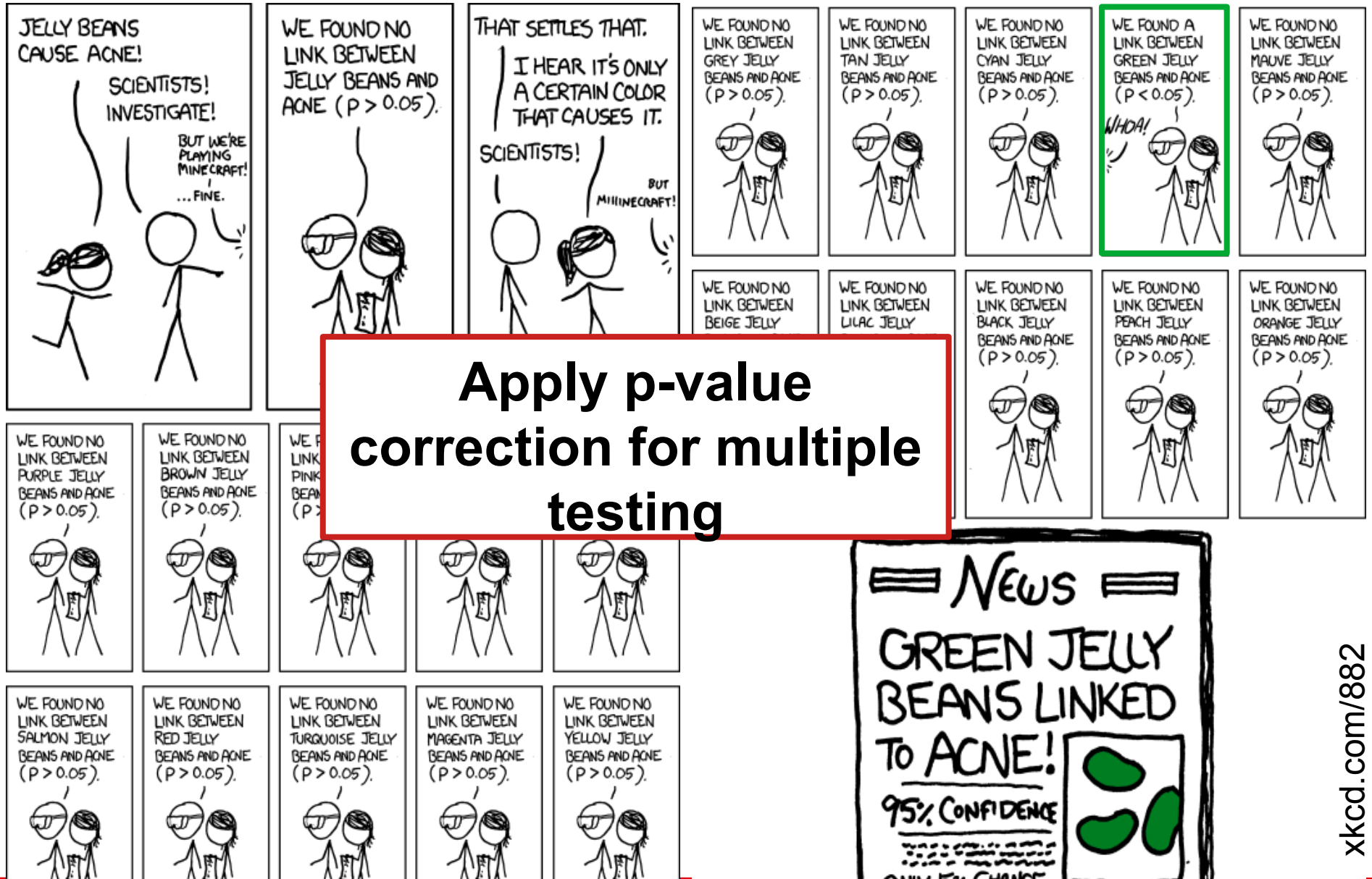






xkcd.com/882

DE - challenges for RNA-Seq



Apply p-value correction for multiple testing

xkcd.com/882

DE – Input for Gene Differential Expression

- **Counts from mapping**
  - **Handling of overlap? Stranding?**
  - **Multi-mapping reads?**
  - **Affected by library size**
- **TPM from pseudo-aligners**
  - **Tximport aggregates counts at the gene-level**

DE – Input for Gene Differential Expression

- **Counts from mapping**
  - **Handling of overlap? Stranding?**
  - **Multi-mapping reads?**
  - **Affected by library size**
- **TPM from pseudo-aligners**
  - **Tximport aggregates counts at the gene-level**

EdgeR and DESeq2 expect raw counts

DE – digression: “naive” normalization

- **CPM (Count Per Million) : count / library size * 10^6**
- **RPKM (Reads Per Kilobase per Million):**
  - **CPM / gene length (kb)**
- **TPM (Transcript Per Million):**
  - **RPK = Count / gene length (kb)**
  - **RPK / sum(RPK) * 10^6**

DE – “naive” normalization

- **CPM (Count Per Million) : count / library size * 10^6**
- **RPKM (Reads Per Kilobase per Million):**
  - **CPM / gene length (kb)**

Sum RPKM different between samples

- **TPM (Transcript Per Million):**
  - **RPK = Count / gene length (kb)**
  - **RPK / sum(RPK) * 10^6**

Sum TPM constant between samples

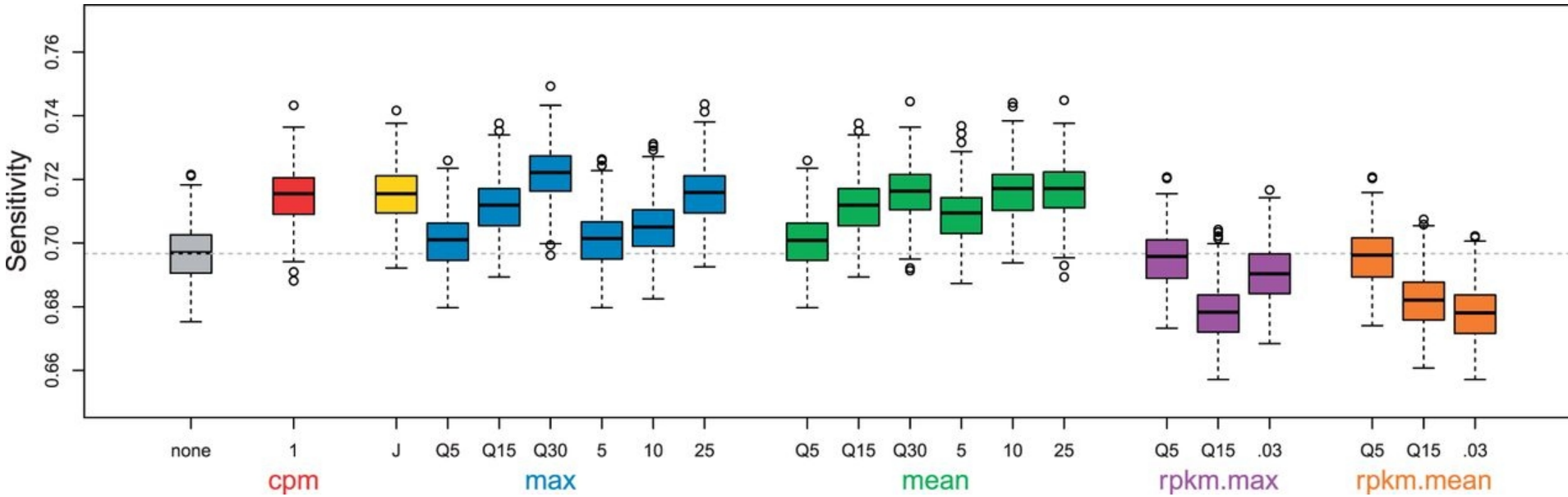
DE – “naive” normalization

- **CPM (Count Per Million) : count / library size * 10^6**
- **RPKM (Reads Per Kilobase per Million):**
  - **CPM / gene length (kb)**
- **TPM (Transcript Per Million):**
  - **RPK = Count / gene length (kb)**
  - **RPK / sum(RPK) * 10^6**

How do you compute “gene length” ?

DE – Filtering low count genes

- **Very low counts genes :**
  - **Very little information. No chance of DE.**
  - **Filter : less p-value correction**

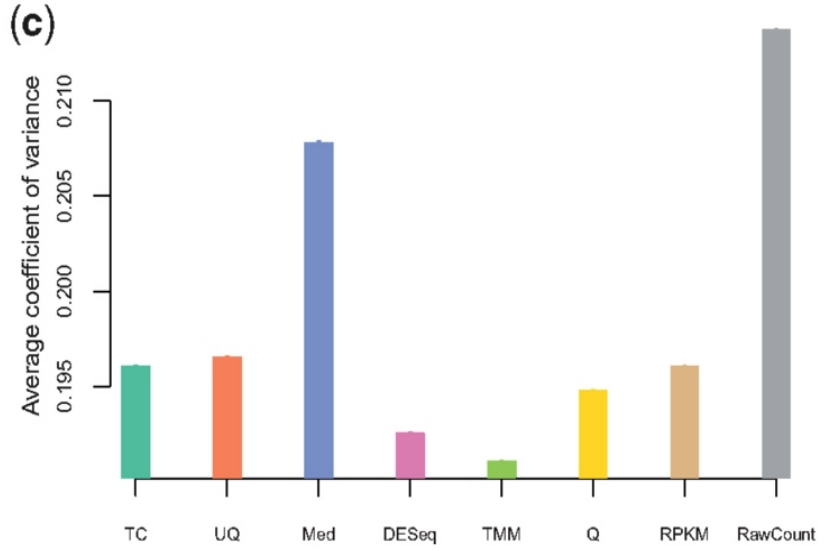


Andrea Rau et al., 2009 https://doi.org/10.1093/bioinformatics/btt350

EdgeR: CPM > 10/(min lib size) in at least N samples

DESeq2 : mean normalized count optimizing # of DEG

DE – Normalization



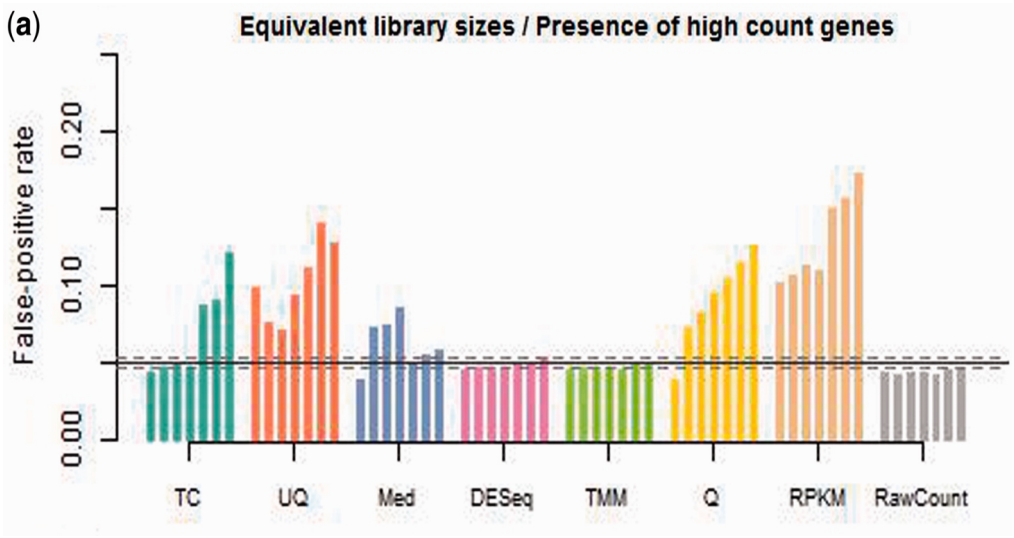


Table 3: Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	−	+	+	−	−
UQ	++	++	+	++	−
Med	++	++	−	++	−
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	−	+	++	−
RPKM	−	+	+	−	−

A '−' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

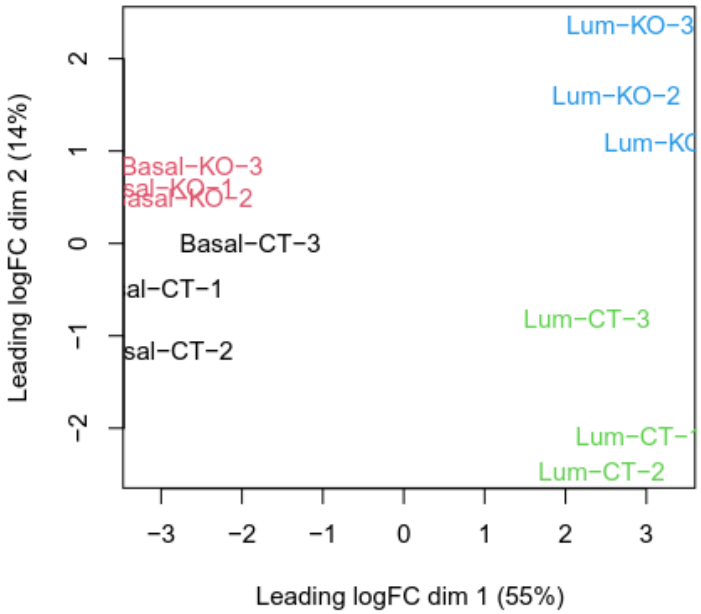
Dillies *et al* 2013 https://doi.org/10.1093/bib/bbs046

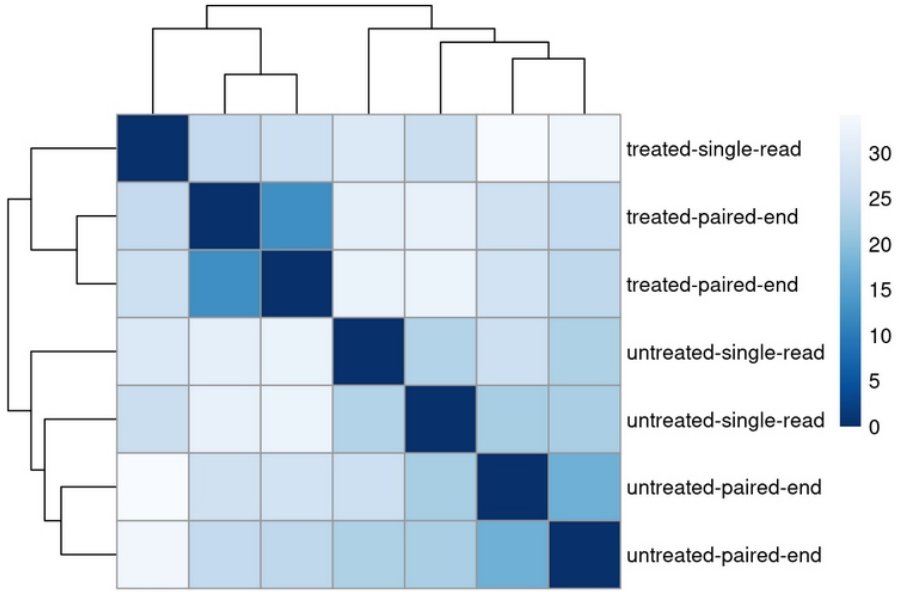
DE – Normalization

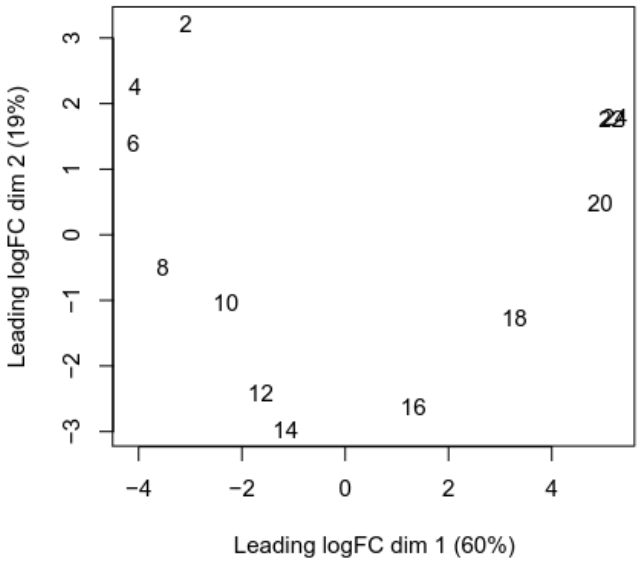
- **EdgeR : “Trimmed Mean of M-Values” (TMM)**
- **DESeq2 : “Relative Log Expression” (RLE)**

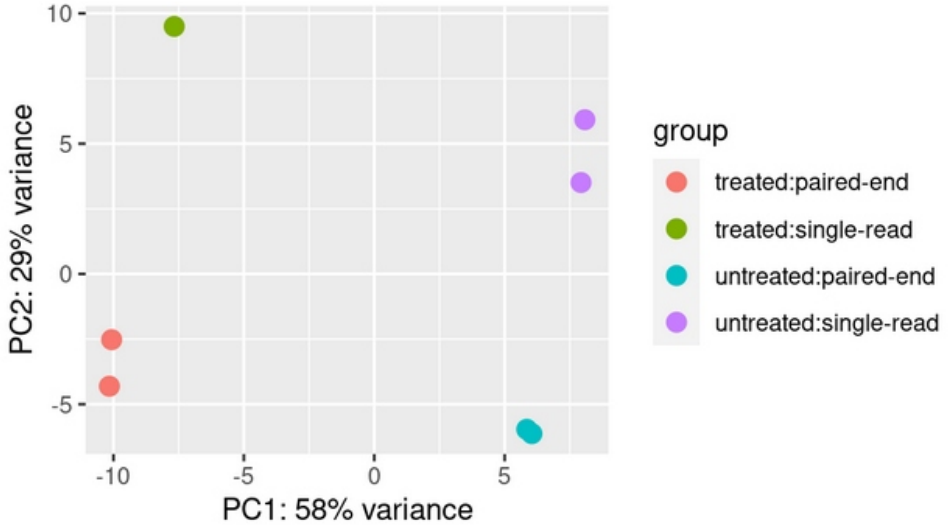
- **Both presume that most genes are not DE!**

QC – MDS or PCA of the samples



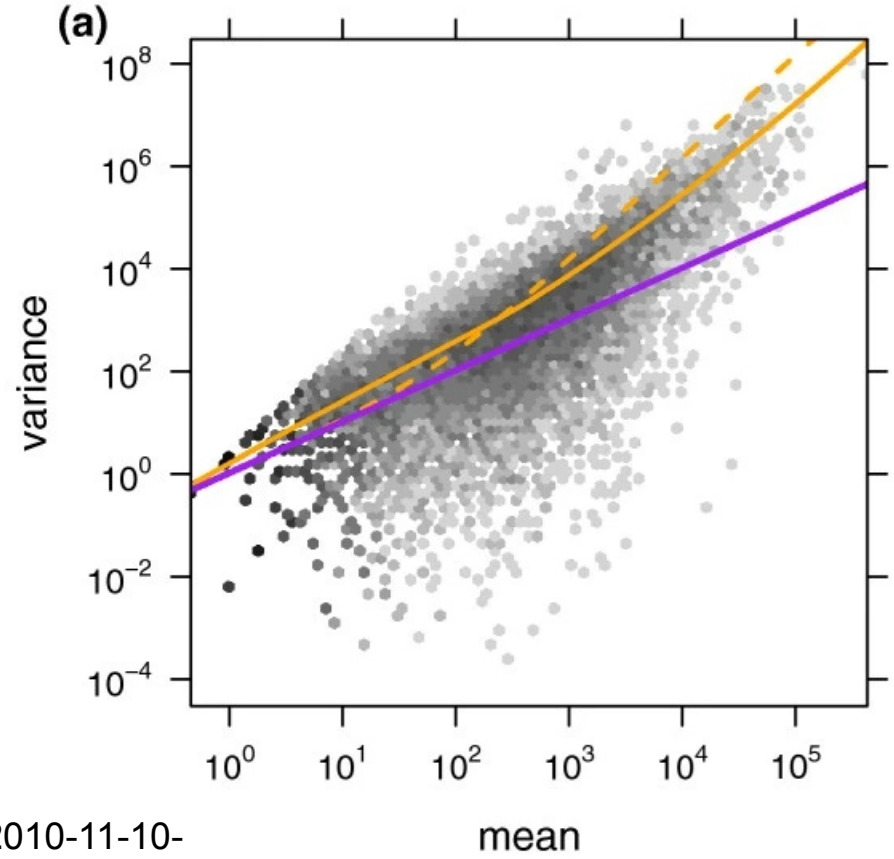






DE – negative binomial model

- **Generalized Poisson with over-dispersion**
- **count of a gene in a sample ~ NB(μ , θ)**
  - **Variance = μ + θμ²**
  - **θ: dispersion parameter**

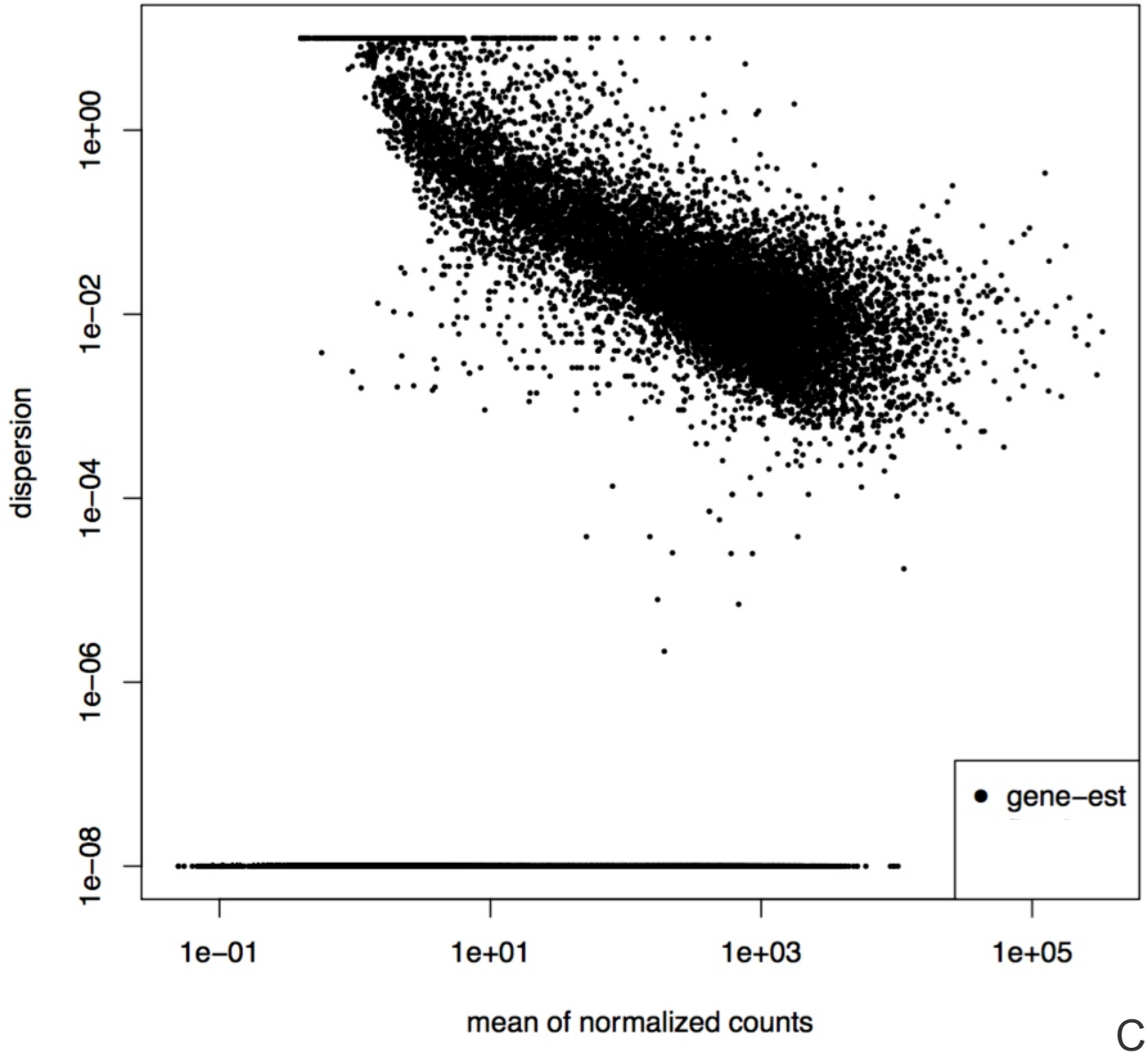


Anders, S., Huber, W. 2010 https://doi.org/10.1186/gb-2010-11-10-r106

Shrinkage of dispersion estimates

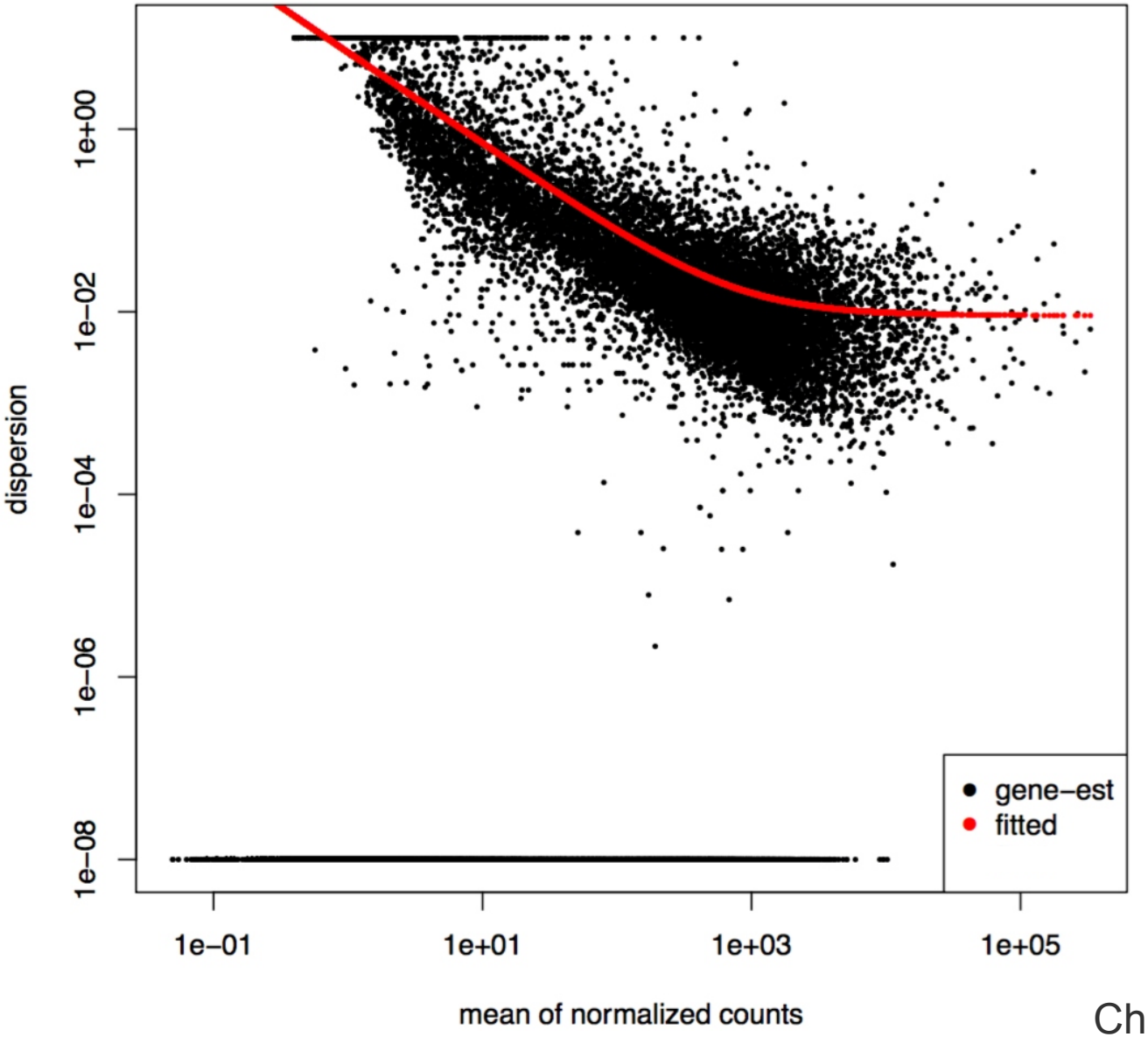
- **Problem: we often have few replicates**
- **Solution: take advantage of the large number of genes**
  - **shrink gene-wise estimates towards** the center value observed of dispersion across **genes with similar expression**

Shrinkage dispersion estimation



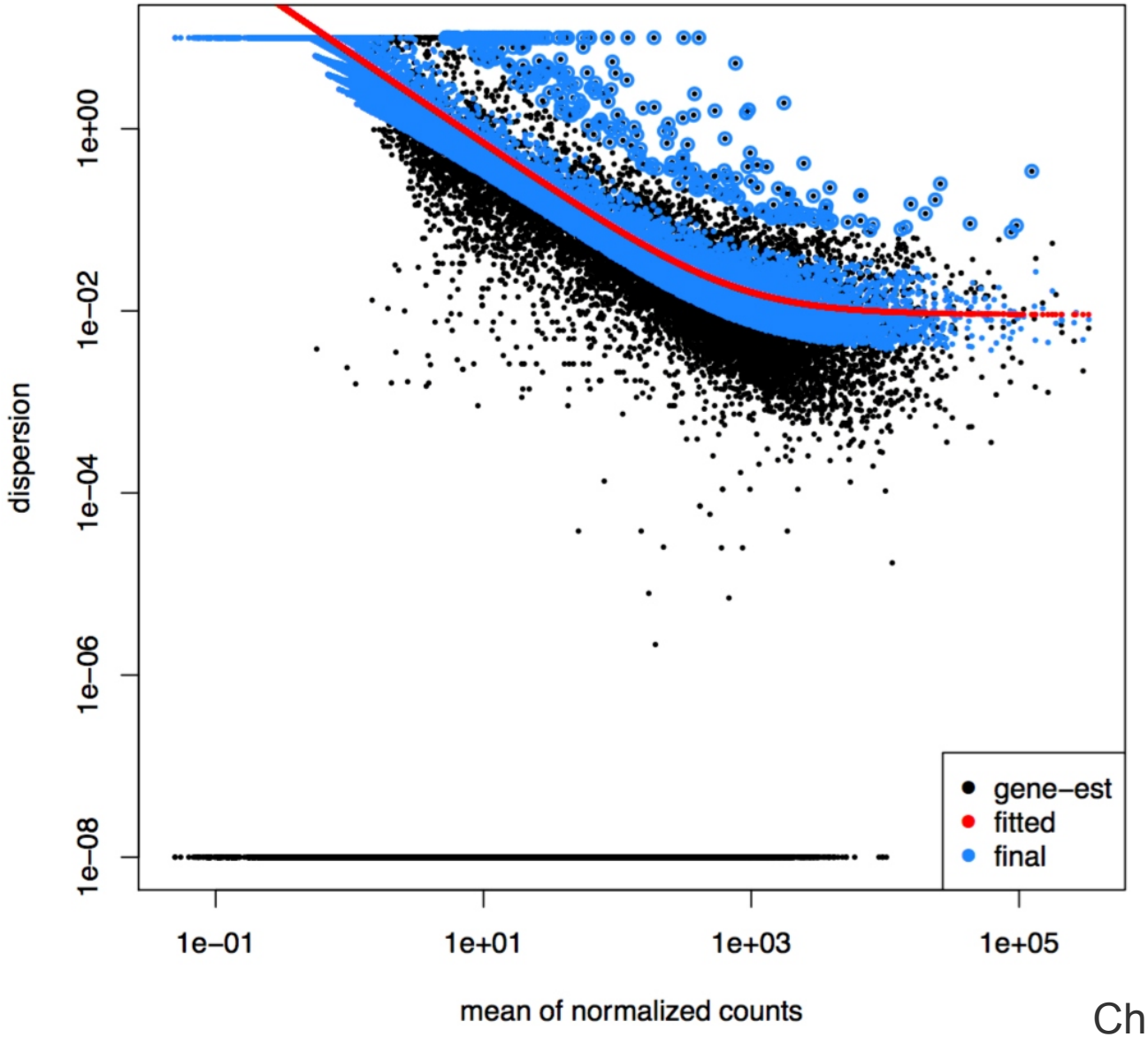
Charlotte Soneson

Shrinkage dispersion estimation



Charlotte Soneson

Shrinkage dispersion estimation



Charlotte Soneson

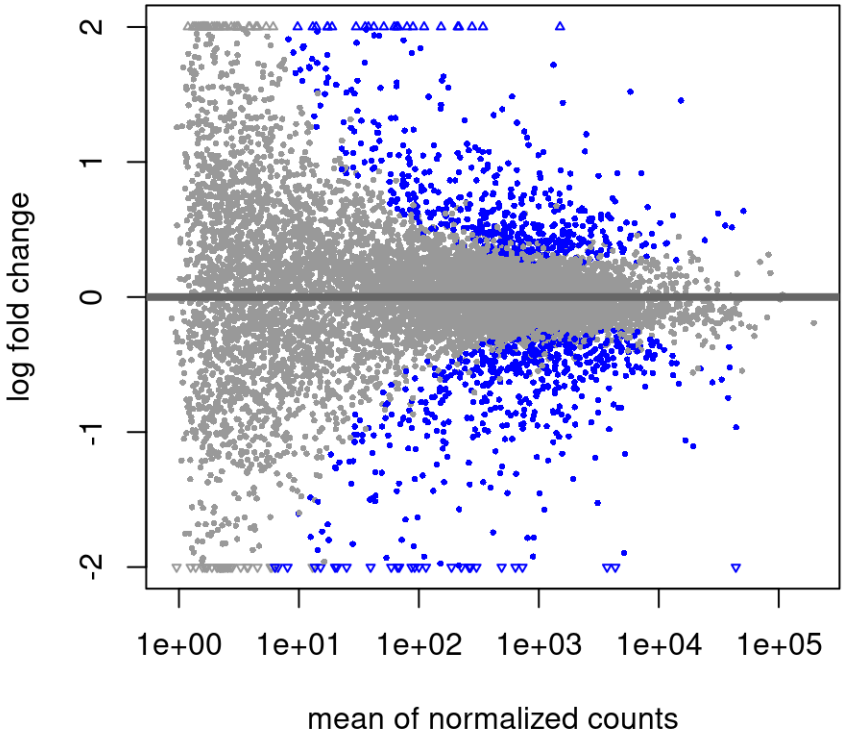
Tests for differential expression – DESeq2

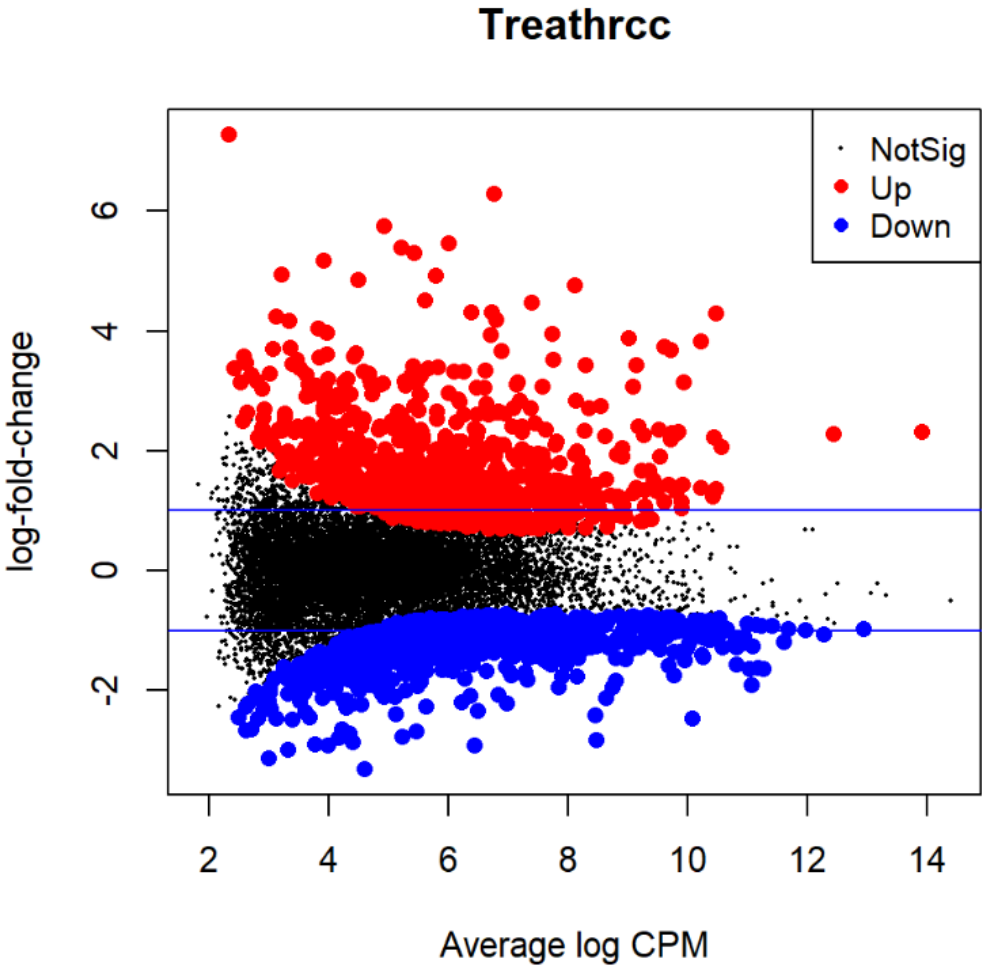
- **For each gene:**
  - **Z-score = shrunken LFC / estimate standard error**
- **Z-score → standard normal distribution → p-value (Wald test)**
- **Benjamini-Hochberg procedure to adjust p-values**

Tests for differential expression – edgeR

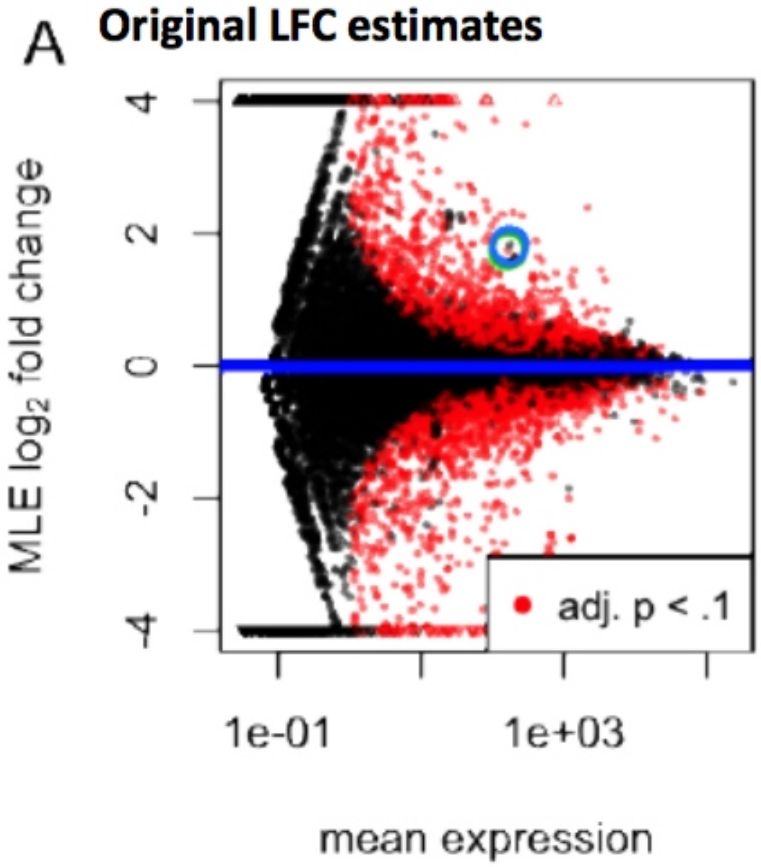
- **“simple” 1 factor : exactTest()** ,
  - using the computed conditional distribution for the sum of counts in a group
- **Otherwise a GLM framework in used :**
  - **QL F-test :** preferred → normally stricter error rate control
  - **LRT :** when “the dispersions are very large and the counts are very small, whereby some of the approximations in the QL framework seem to fail”

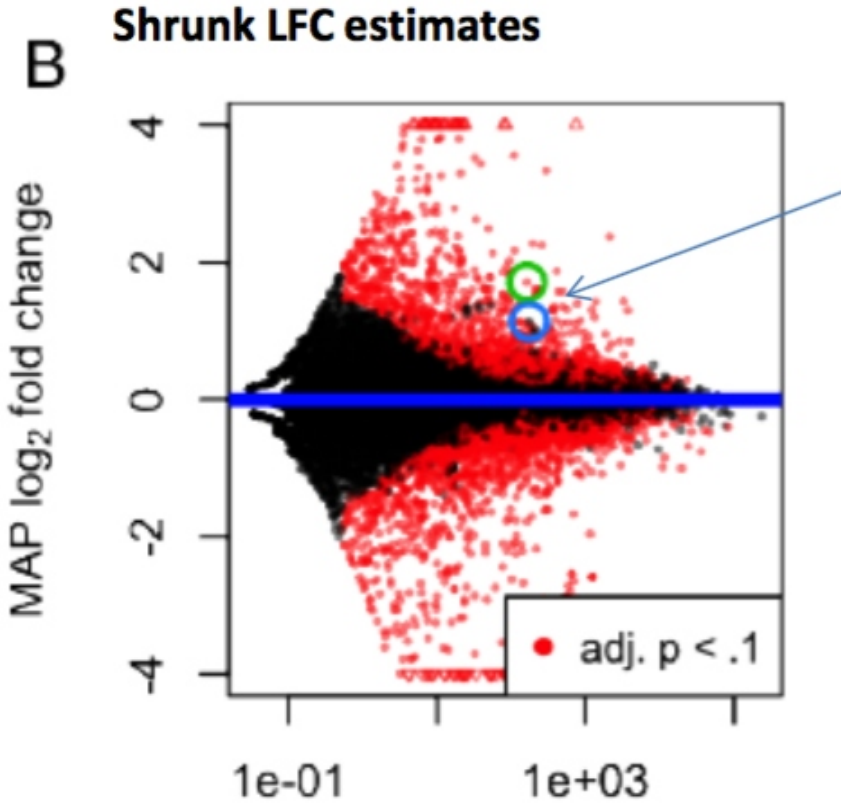
MA plot


log fold change
2
1
0
-1
-2
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
mean of normalized counts


Treathrcc
NotSig
Up
Down
log-fold-change
6
4
2
0
-2
2
4
6
8
10
12
14
Average log CPM

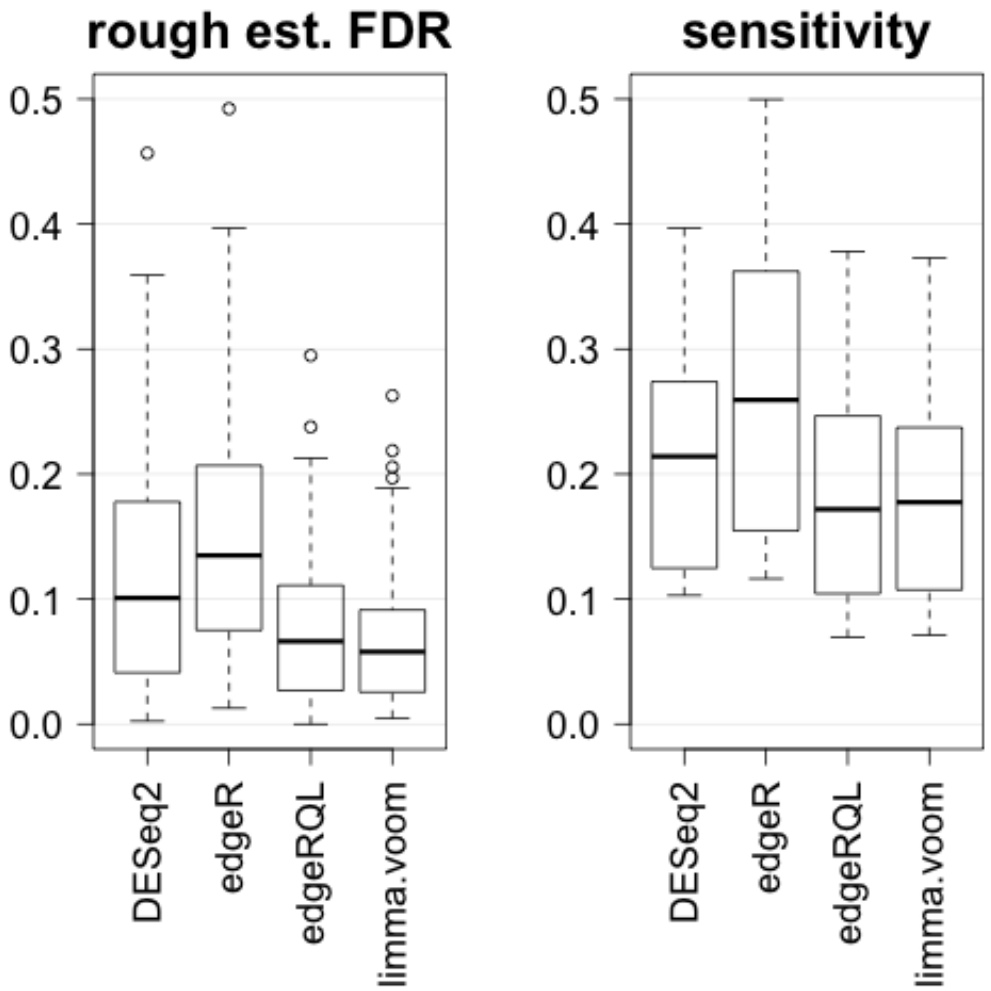
Shrinkage of log-fold change





Love *et al* 2014

edgeR vs DESeq2

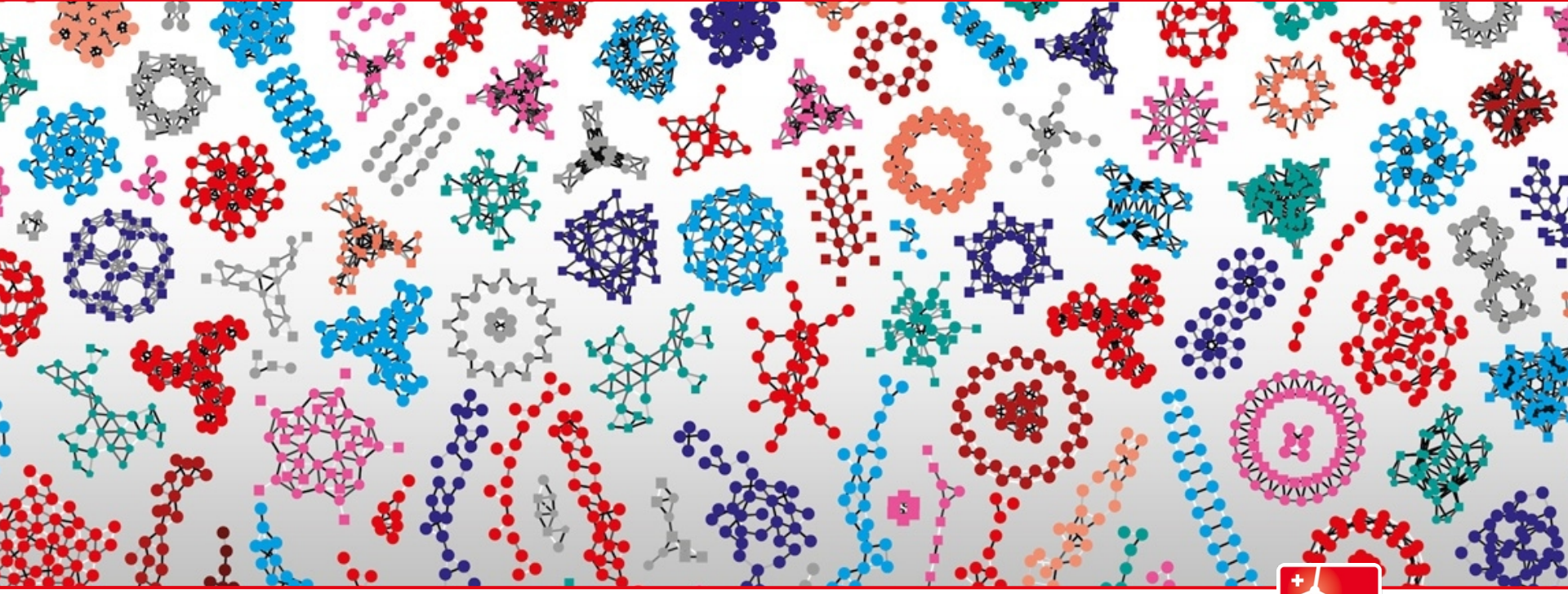


- **EdgeR-exactTest() :** more sensitive
- **EdgeR-QL :** more conservative
- **DESeq2 :** tight FDR control

https://mikelove.wordpress.com/2016/09/28/deseq2-or-edger/

Practical

- Go to the website and follow the Differential Expression practical



Swiss Institute of Bioinformatics

Contributors:

Geoffrey Fucile
Walid Gharib
Irene Keller
Pablo Escobar Lopez
Charlotte Soneson