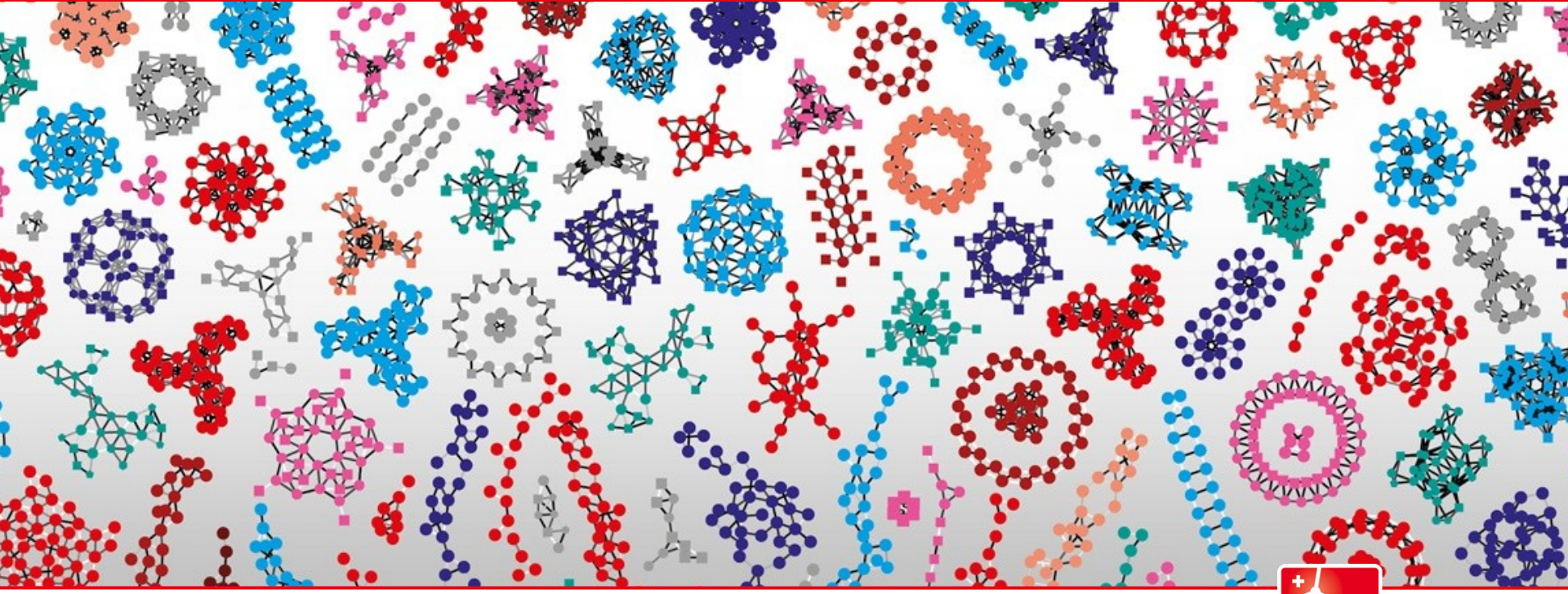
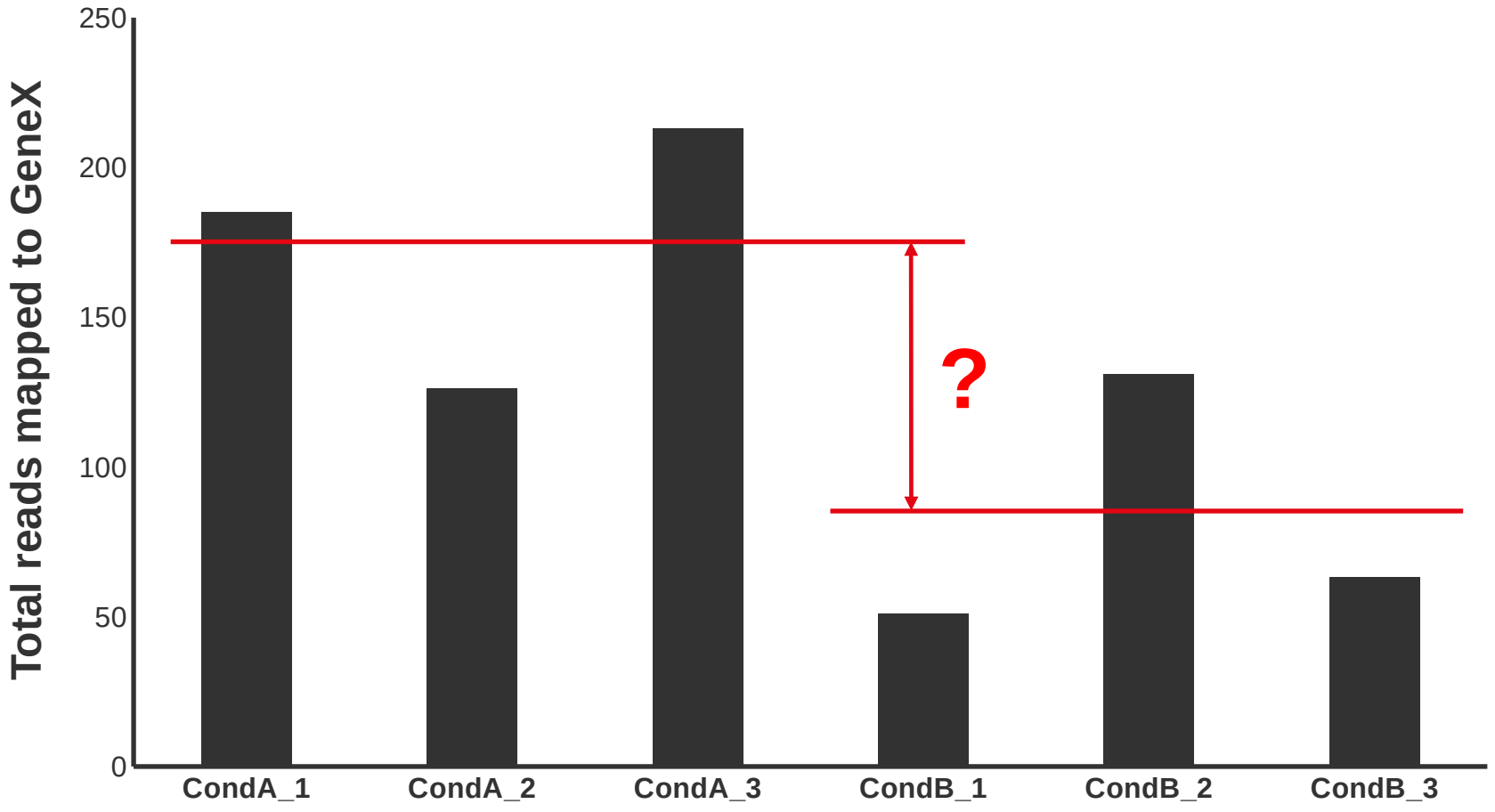


Swiss Institute of Bioinformatics

Introduction to RNA-Seq – Differential Expression

Wandrille Duchemin

How to define statistical significance?



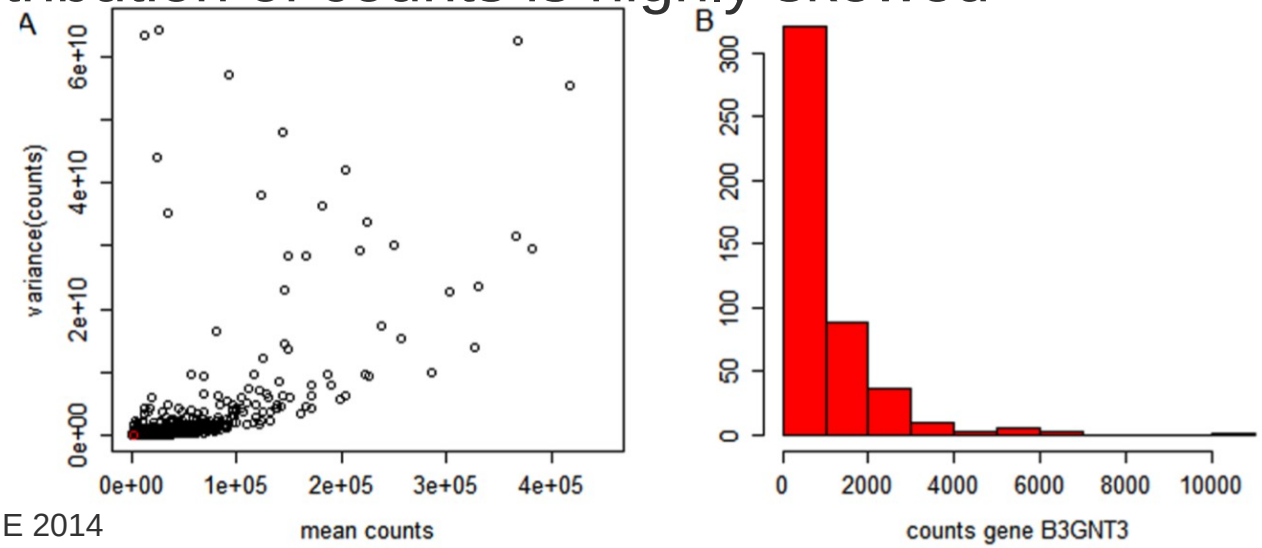
Statistical modeling of RNA-Seq data

- **Essentially, two approaches**
  - Non-parametric: *eg* voom (based on limma)
  - Parametric: *eg* edgeR and DESeq2

- **We will be working with the parametric approaches packaged in edgeR and DESeq2**
  - Important assumptions:
    - most genes are not differentially expressed
    - probability of a read mapping to geneX is the same for all samples in a class

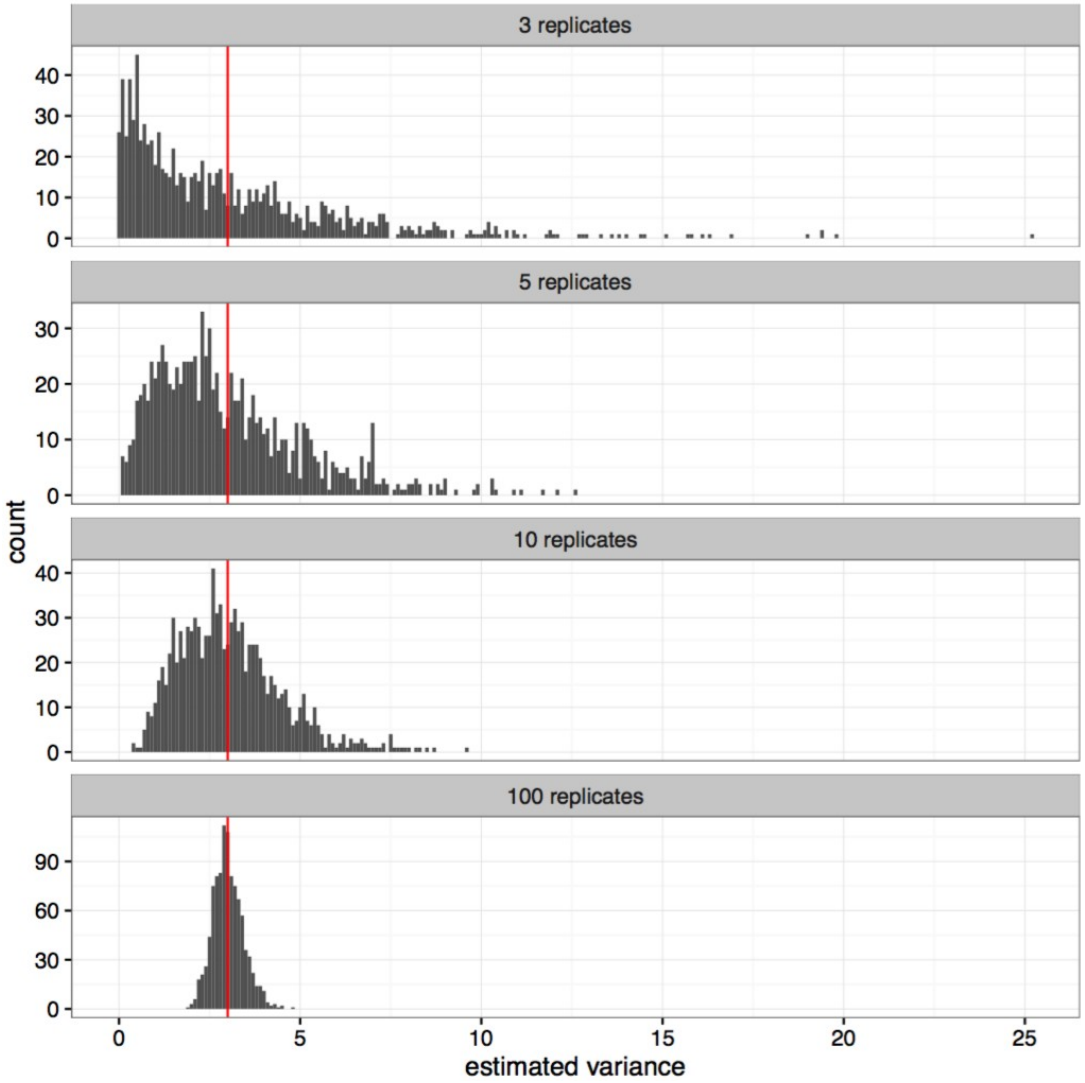
Challenges for RNA-Seq

- **Which statistical distribution is most appropriate?**
- **How to normalize read counts between samples?**
- **Estimating variance is difficult**
  - Typically, very few replicates
  - The variance depends on the mean count
  - Distribution of counts is highly skewed



Zweiner *et al*, PLoS ONE 2014

Estimating variance of a normally distributed variable



True variance = 3

What does this tell you about the ideal number of replicates for your RNA-Seq experiment?

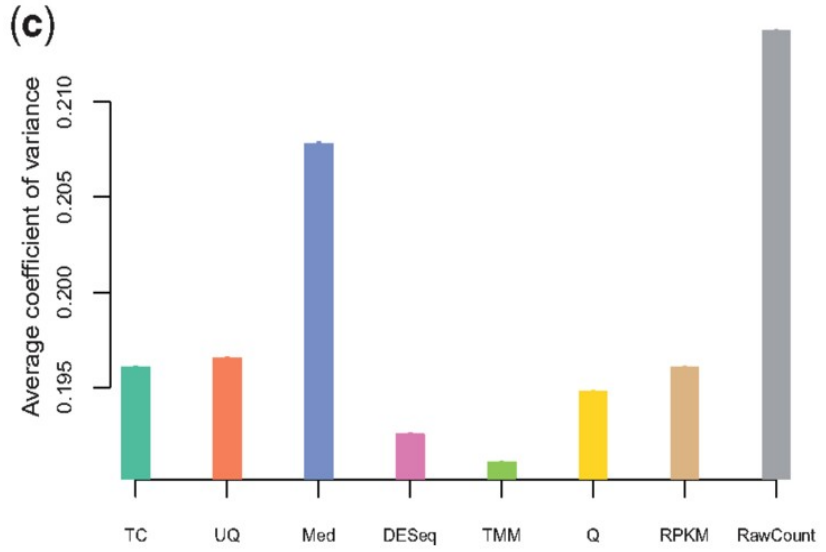
The negative binomial distribution

- Used by both edgeR and DESeq2, essentially a generalized Poisson distribution
- Variance is modeled as:
  - $var(X) = \mu + \theta\mu^2$
- θ = dispersion
- $\sqrt{\theta}$ = "biological coefficient of variation"
- Allows mRNA proportions to vary across samples, accurately capturing variability across biological replicates
- Implemented using a "generalized linear model"

Normalization of RNA-Seq data

- **Raw read counts are not directly comparable across samples. They depend on:**
  - Abundance in source material
  - Gene length
  - Sequencing depth
  - Sequencing biases
  - …
- **edgeR uses the “Trimmed Mean of M-Values” (TMM) method**
- **DESeq2 uses the “Relative Log Expression” (RLE) method**

TMM and RLE normalization yield comparable results



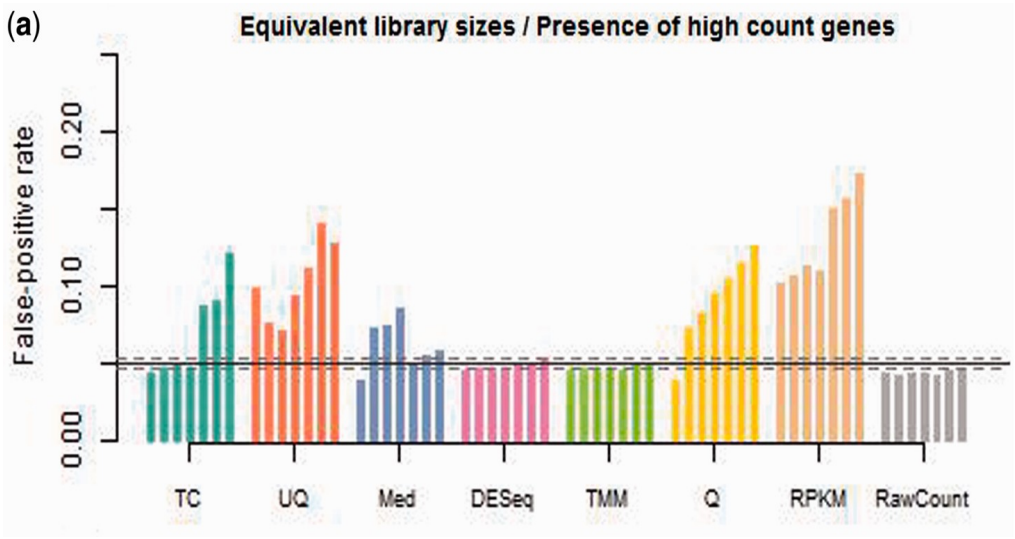


Table 3: Summary of comparison results for the seven normalization methods under consideration

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

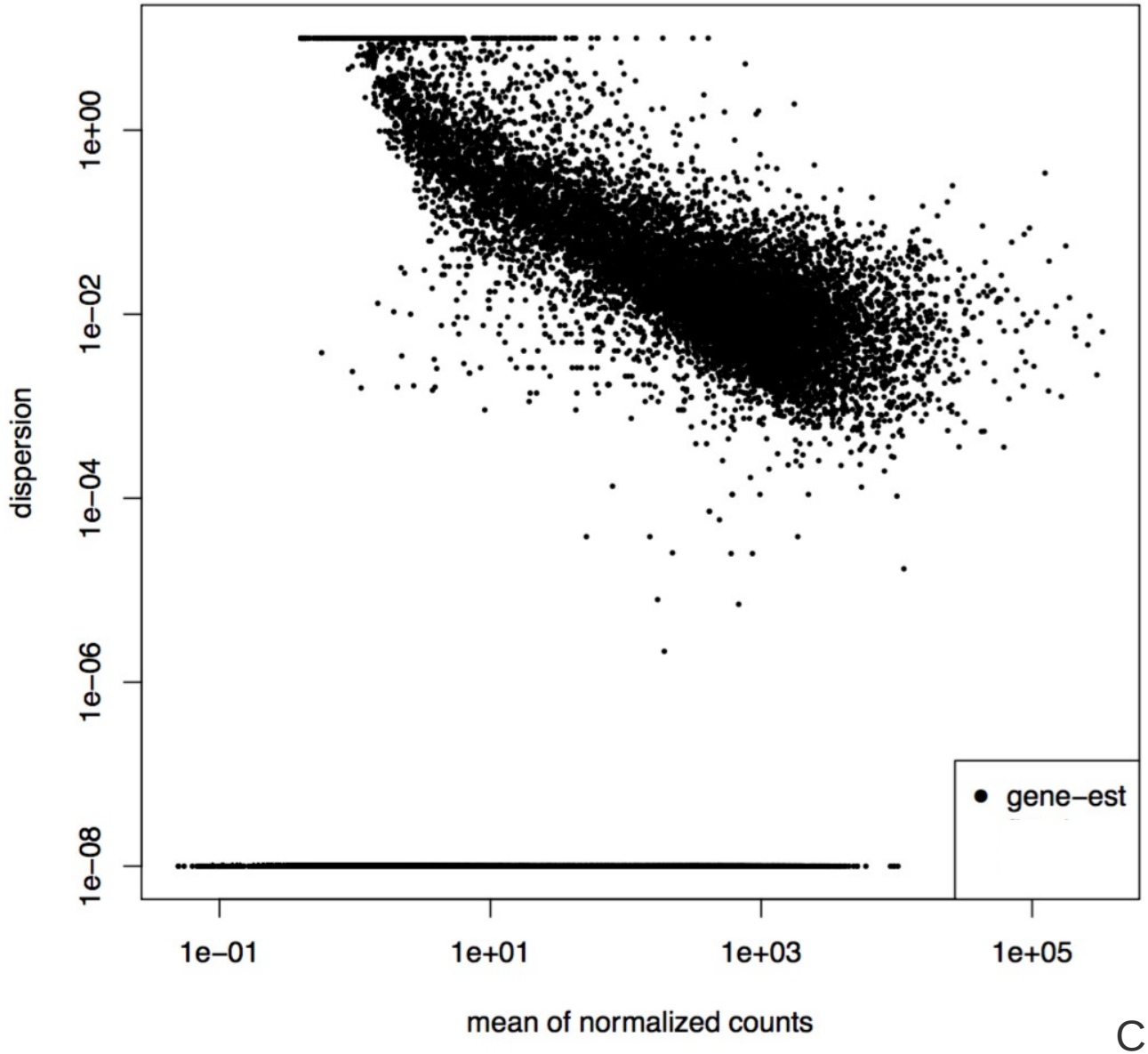
TMM and RLE normalization yield comparable results

- **Keep in mind that “normalization factors” from edgeR and “size factors” from DESeq2 are not equivalent theoretical parameters**
  - For a more detailed discussion, see Maza 2016

Shrinkage of dispersion estimates

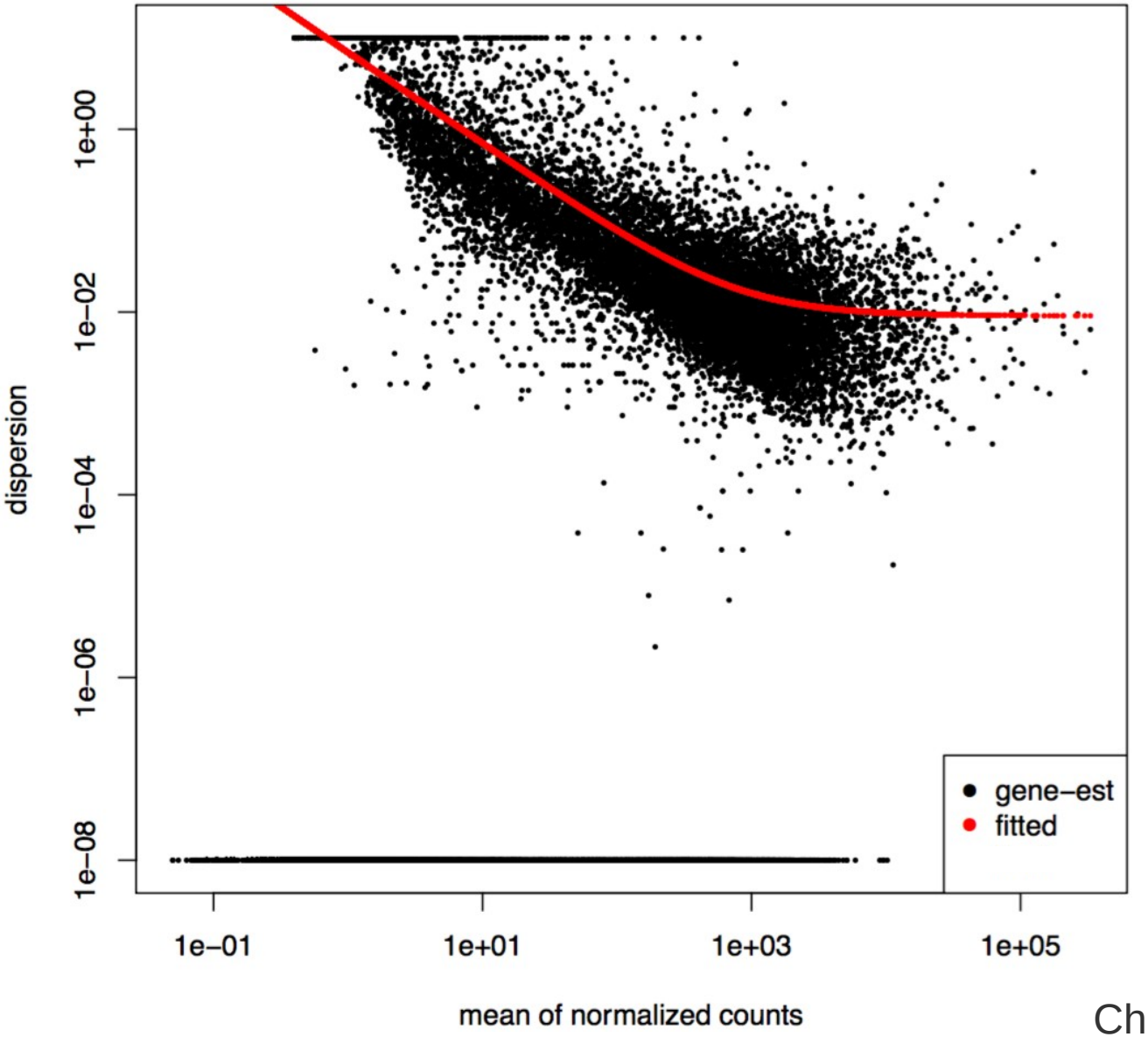
- **Problem: we often have few replicates**
- **Solution: take advantage of the large number of genes, shrink the gene-wise estimates towards a center value defined by the observed distribution of dispersions across:**
  - All genes (“common” dispersion estimate)
  - Genes with similar expression (“trended” dispersion estimate)

Shrinkage dispersion estimation



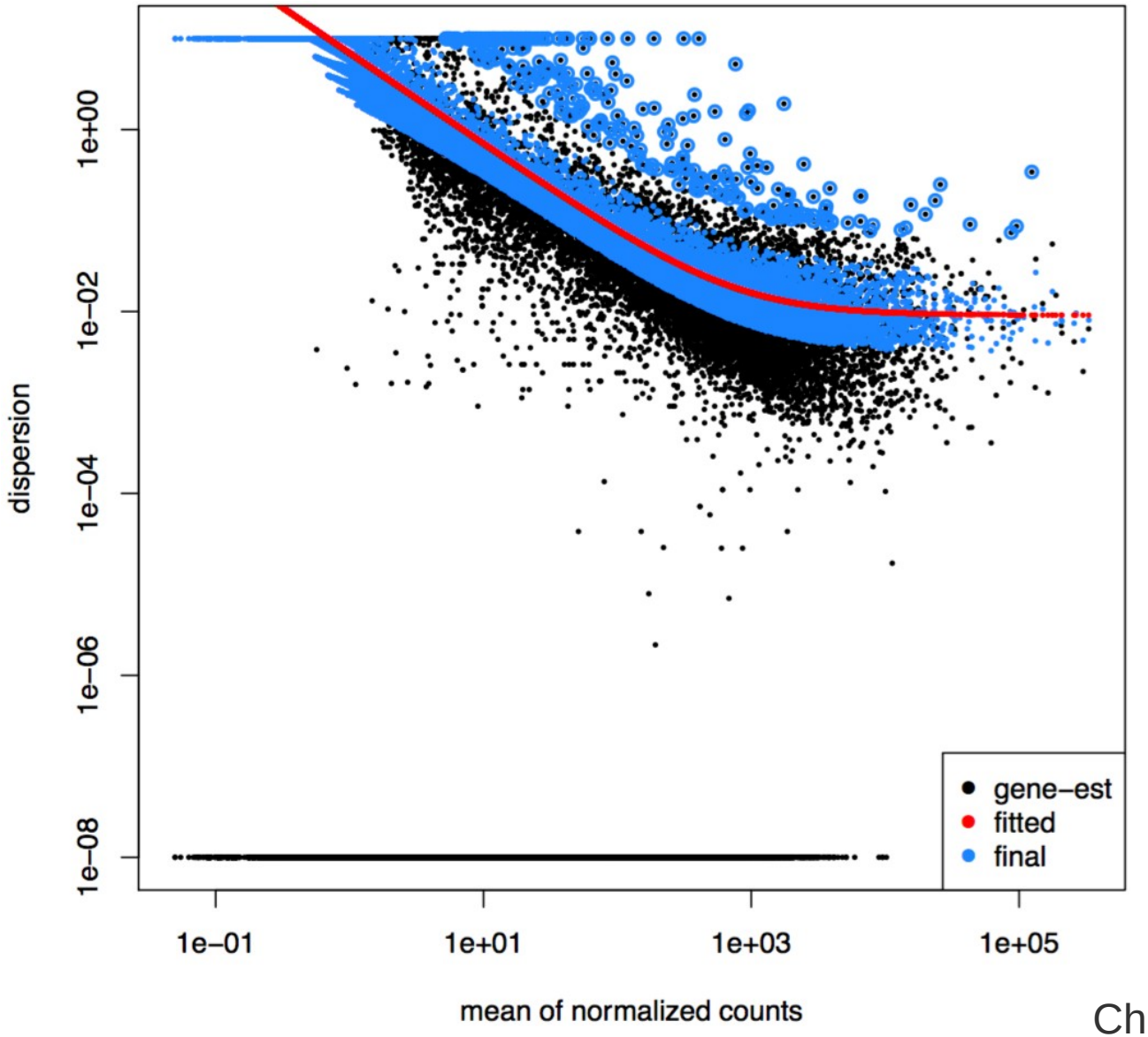
Charlotte Soneson

Shrinkage dispersion estimation



Charlotte Soneson

Shrinkage dispersion estimation

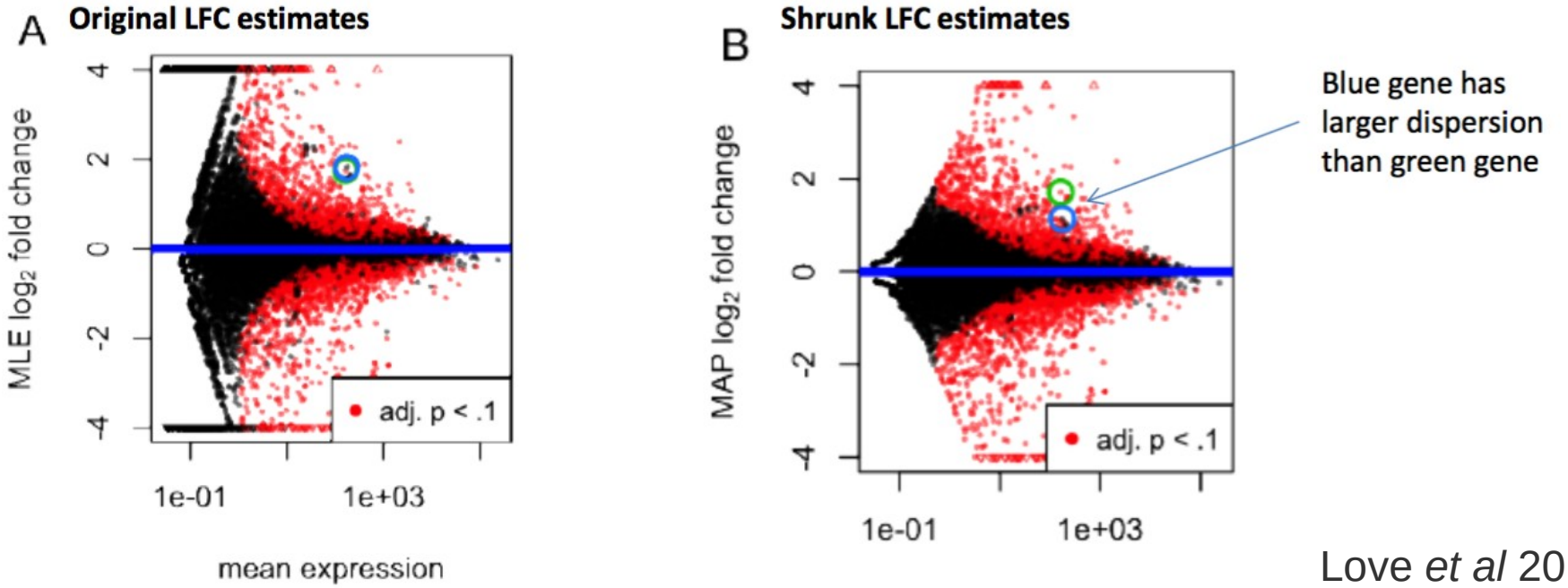


Charlotte Soneson

Shrinkage of log-fold change

- **Problem: weakly expressed genes tend to show much stronger differences between conditions, because count data are very noisy when counts are low**

- **Solution: shrink LFC estimates toward 0 such that shrinkage is stronger when less info is available (*eg* low counts, high dispersion, low replicates)**



Love *et al* 2014

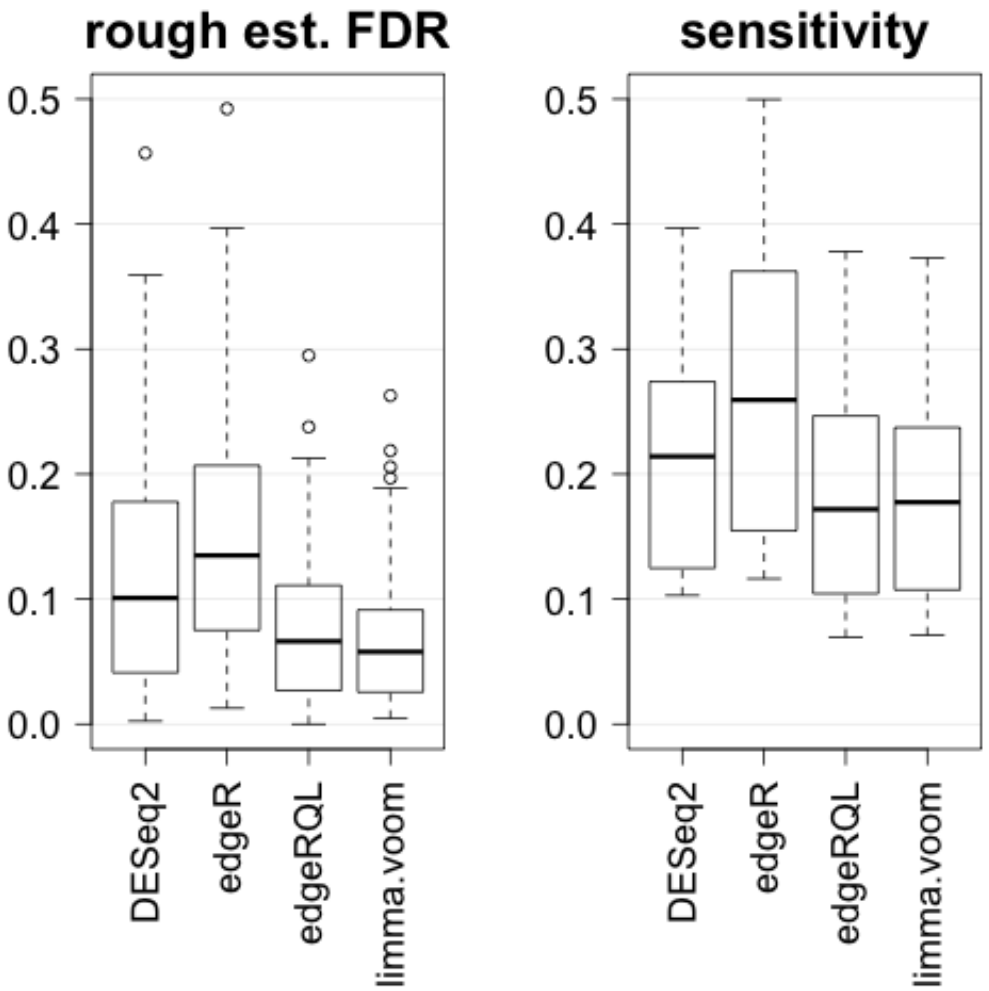
Tests for differential expression – DESeq2

- **A Z-score is calculated for each gene by dividing the shrunken LFC estimate by the standard error of the estimate**
- **The Z-score is compared to a standard normal distribution to obtain a P-value (Wald test)**
- **To correct for multiple testing, adjusted p-values are calculated using the Benjamin-Hochberg procedure**
  - Consider a genome with 20,000 genes
  - At a threshold of p = 0.05, we would expect 1000 significant tests even if there is NO differential expression

Tests for differential expression – edgeR

- **Different methods for different cases :**
- **“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 :** is preferred because it gives “stricter error rate control by accounting for the uncertainty in dispersion estimation… (which can be considerable when you have few replicates and/or the amount of shrinkage is low), whereas the other methods do not”
- **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”

edgeR vs DESeq2



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

edgeR “Robust” Mode

Dispersion estimates can be sensitive to outliers. If this is a significant aspect of your dataset, edgeR has a more robust implementation to estimate dispersions:

https://www.rdocumentation.org/packages/edgeR/versions/3.14.0/topics/estimateGLMRobustDisp

Practical

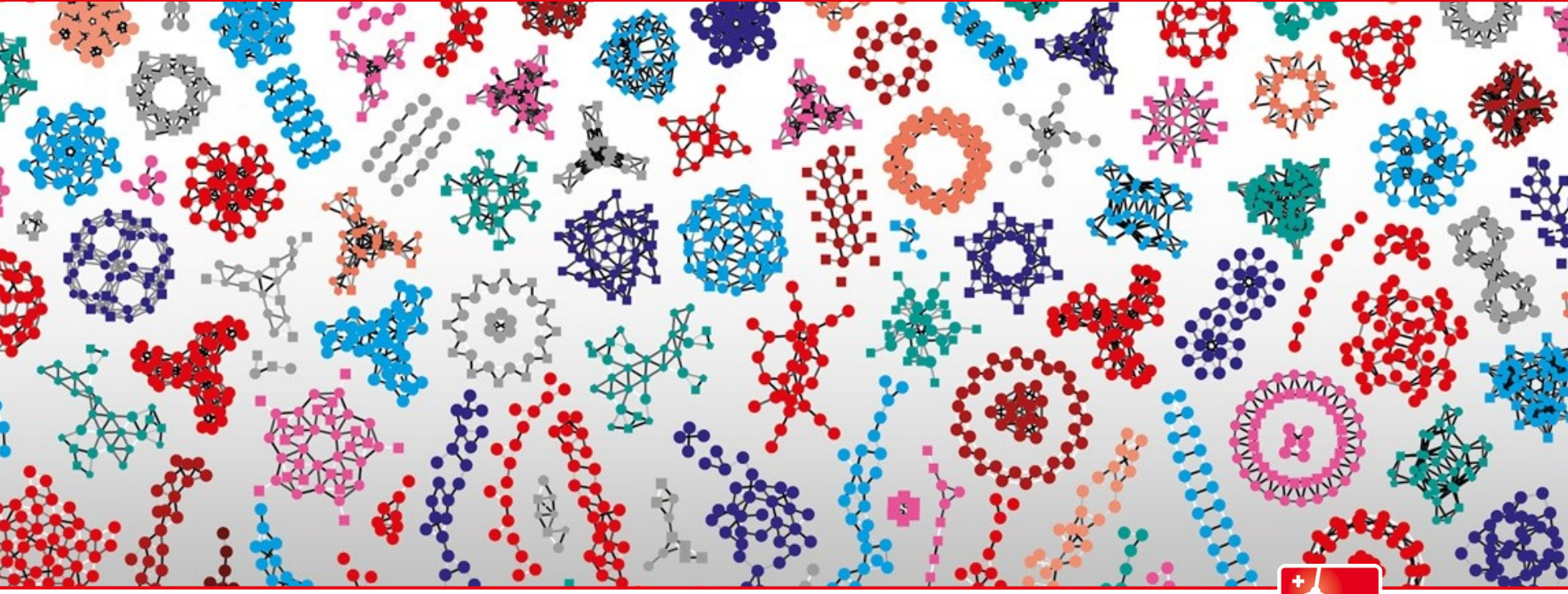
- Go to the website and follow the Differential Expression Inference practical

REFERENCES

Zwiener I, Frisch B, Binder H (2014) Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. PLOS ONE 9(1): e85150. https://doi.org/10.1371/journal.pone.0085150

Dillies et al. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. Nov;14(6):671-83. doi: 10.1093/bib/bbs046

Maza (2016) In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. Front. Genet. https://doi.org/10.3389/fgene.2016.00164



Swiss Institute of Bioinformatics

Contributors:

Geoffrey Fucile
Walid Gharib
Irene Keller
Pablo Escobar Lopez
Charlotte Soneson