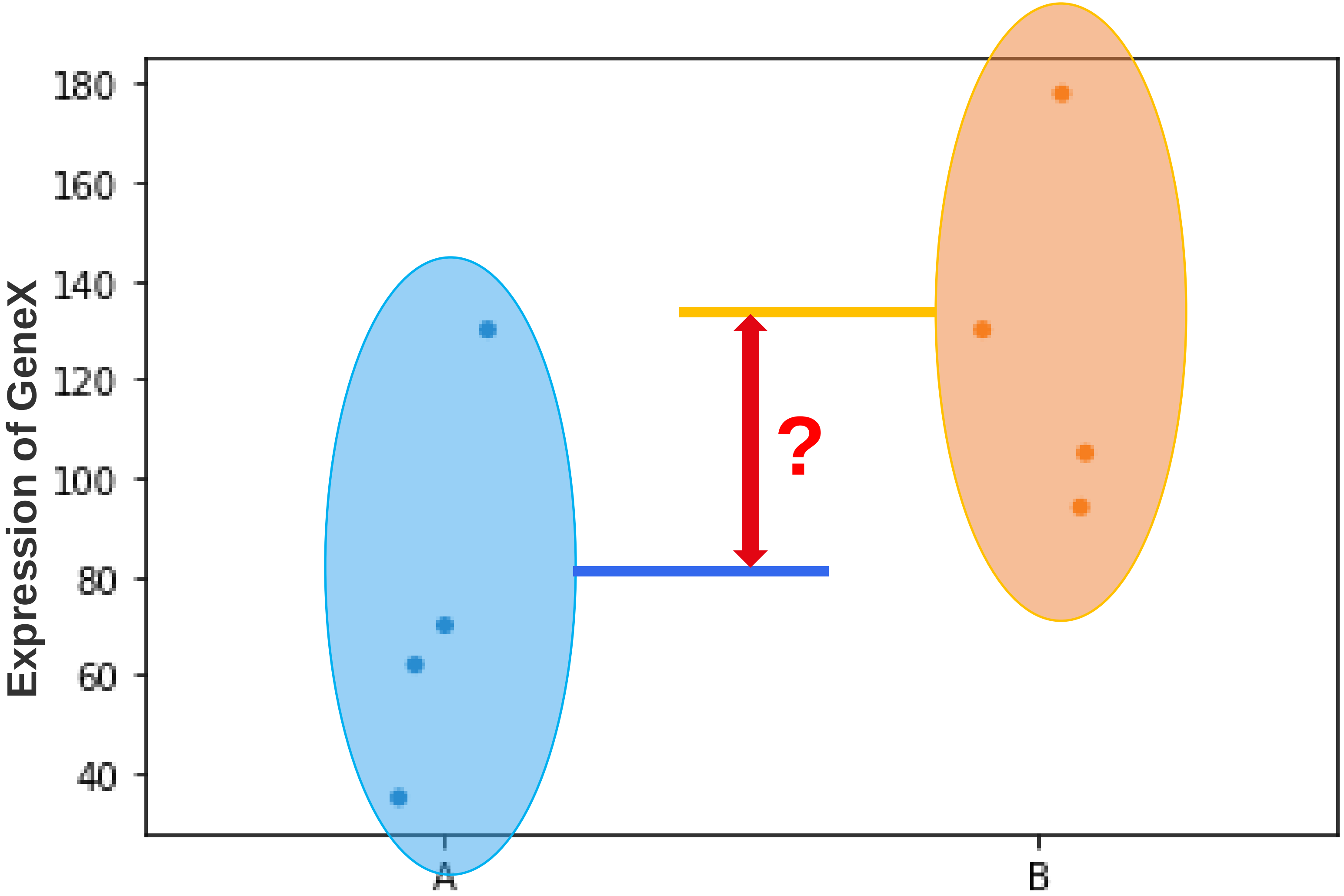
Swiss Institute of
Bioinformatics

Introduction to RNA-Seq: Differential Expression

Wandrille Duchemin

elixir SWITZERLAND

Differential Expression : the goal

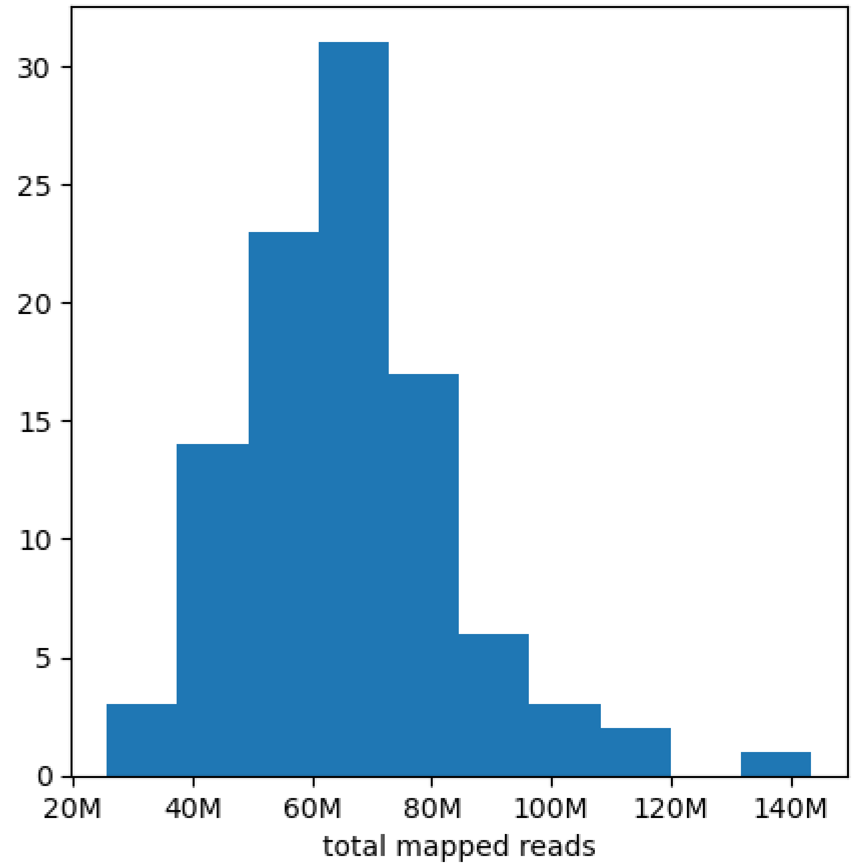


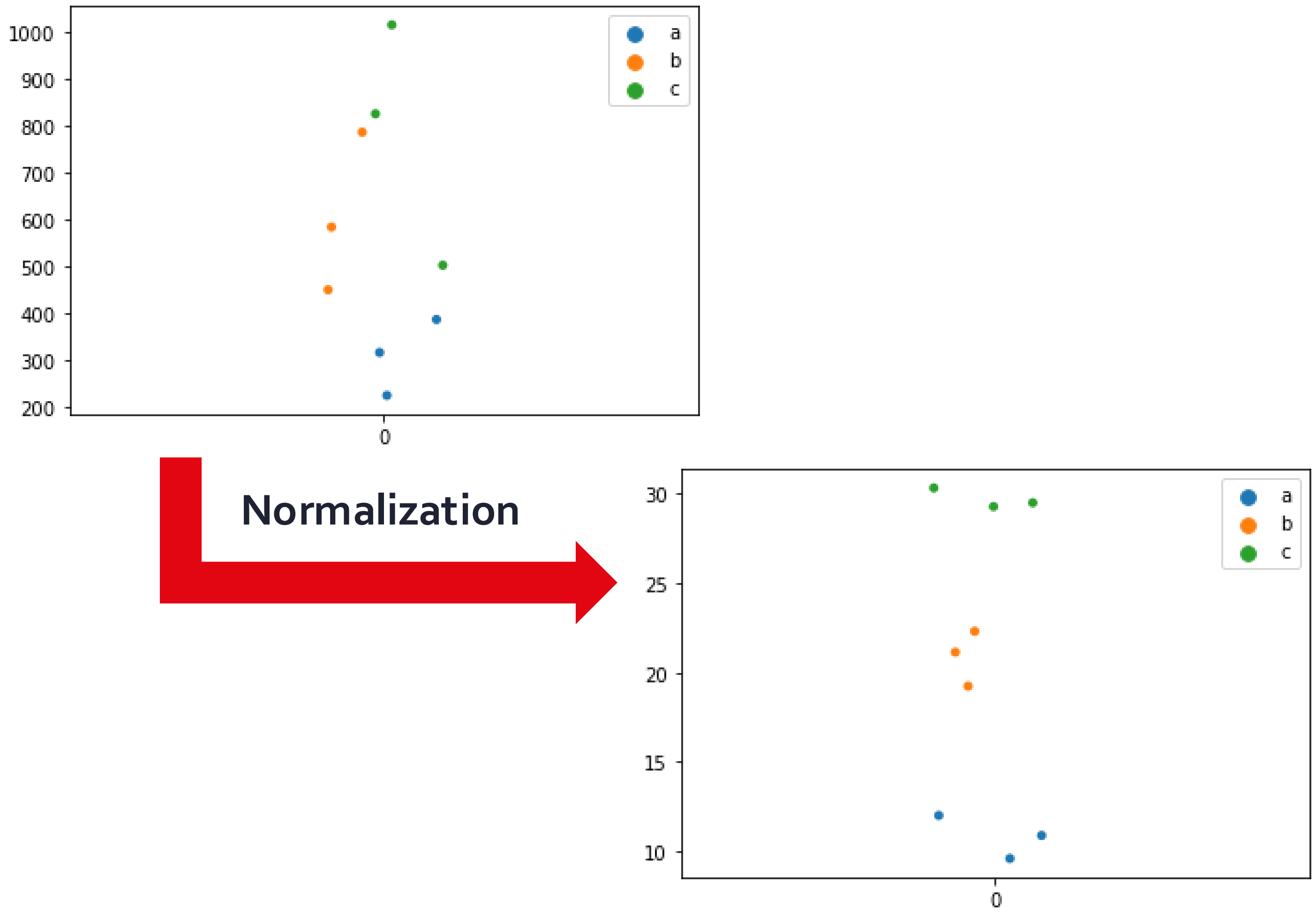
Differential Expression : challenges for RNAseq

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

Differential Expression : challenges for RNAseq

- Sequencing depth varies across libraries

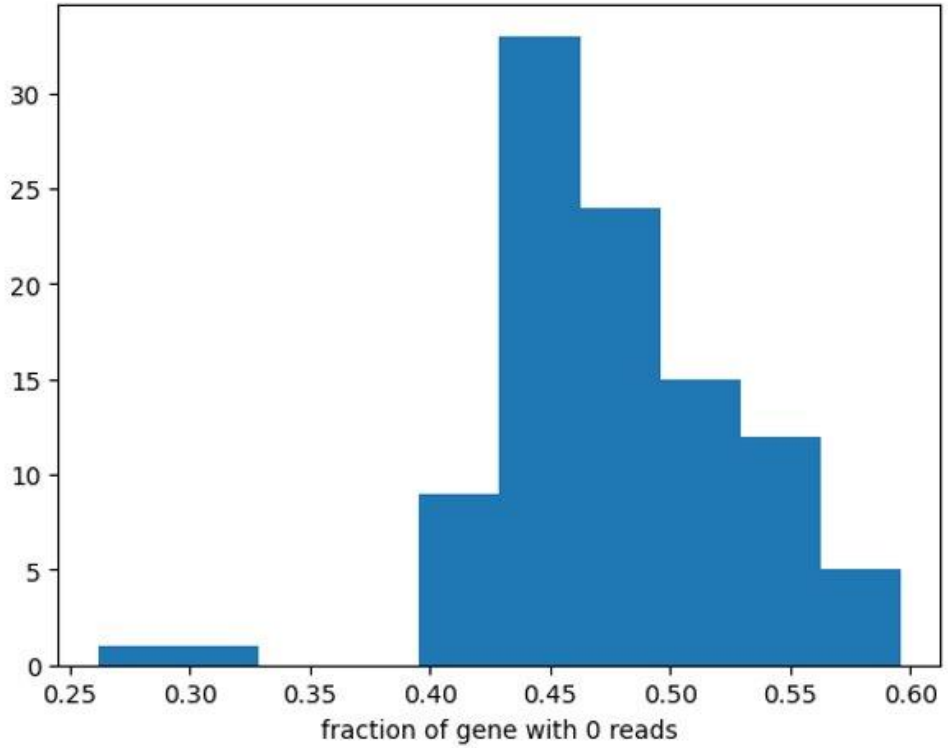


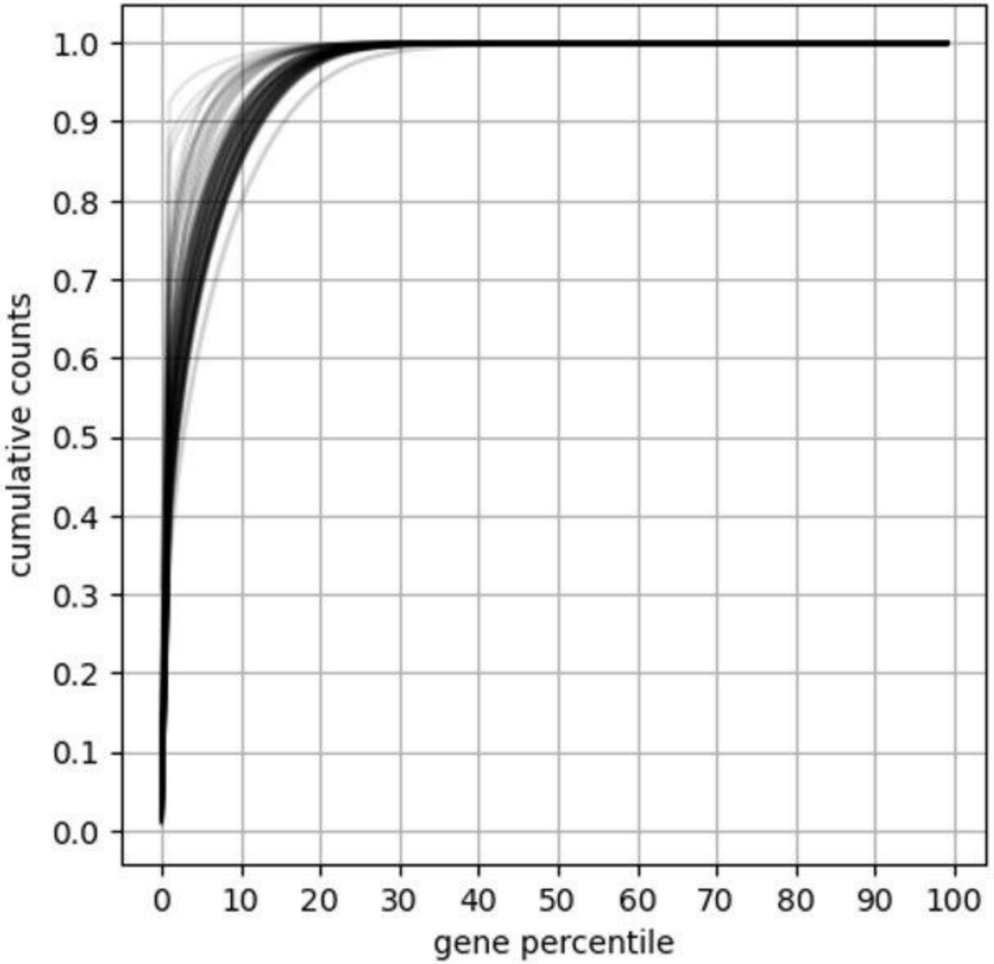


Normalization

Differential Expression : challenges for RNAseq

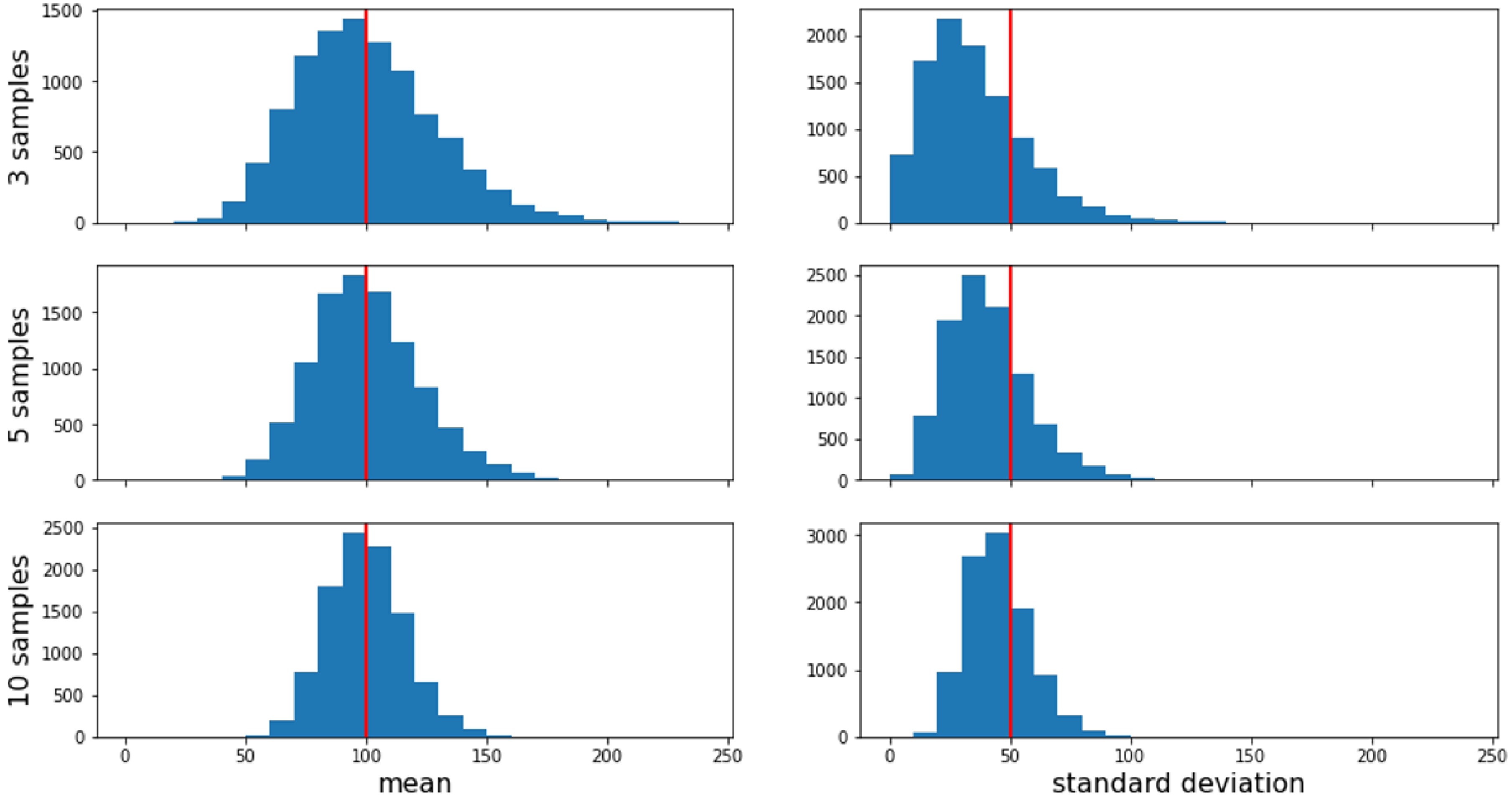
- High dynamic range of expression





Differential Expression : challenges for RNAseq

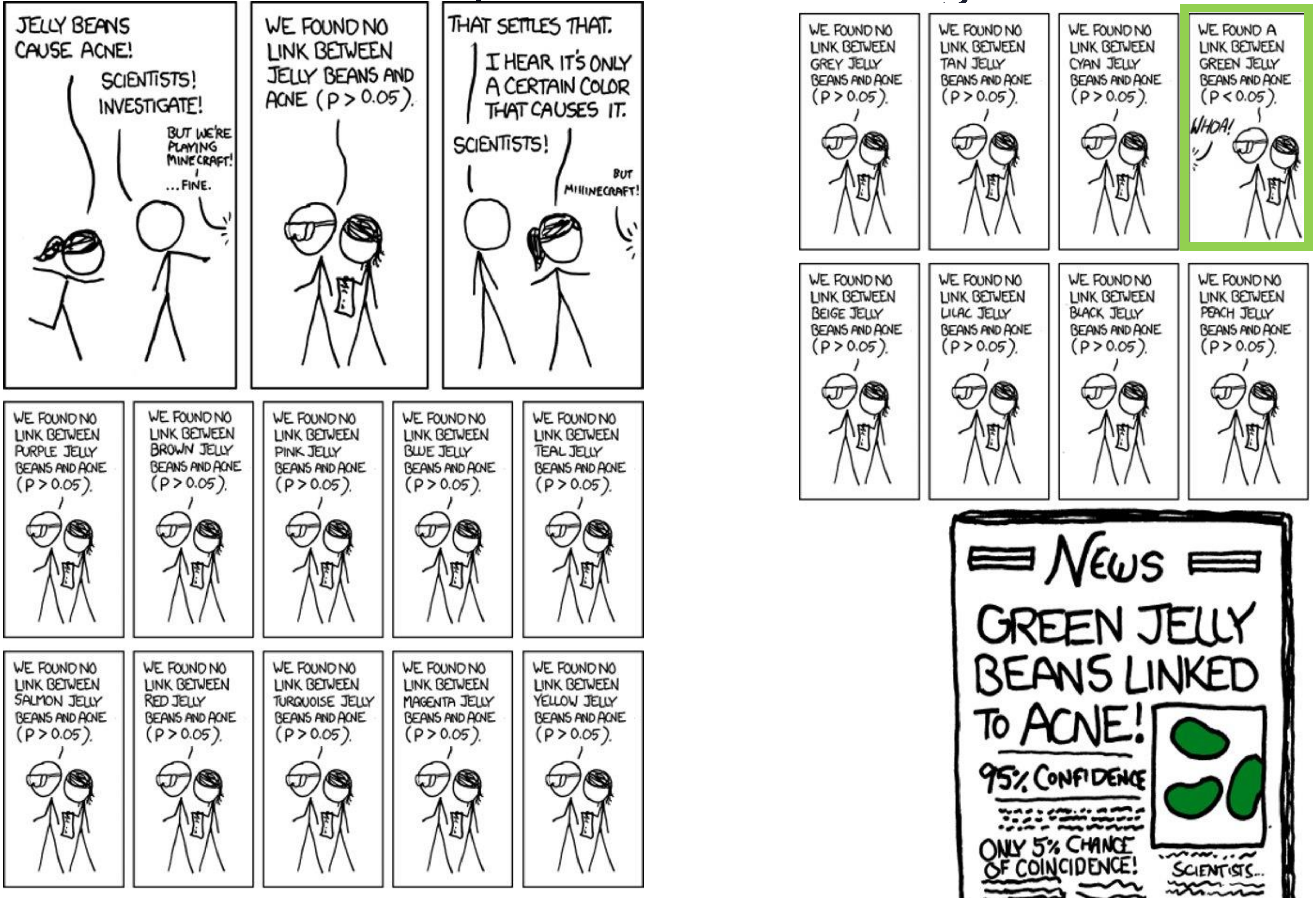
- Limited number of samples



Differential Expression : challenges for RNAseq

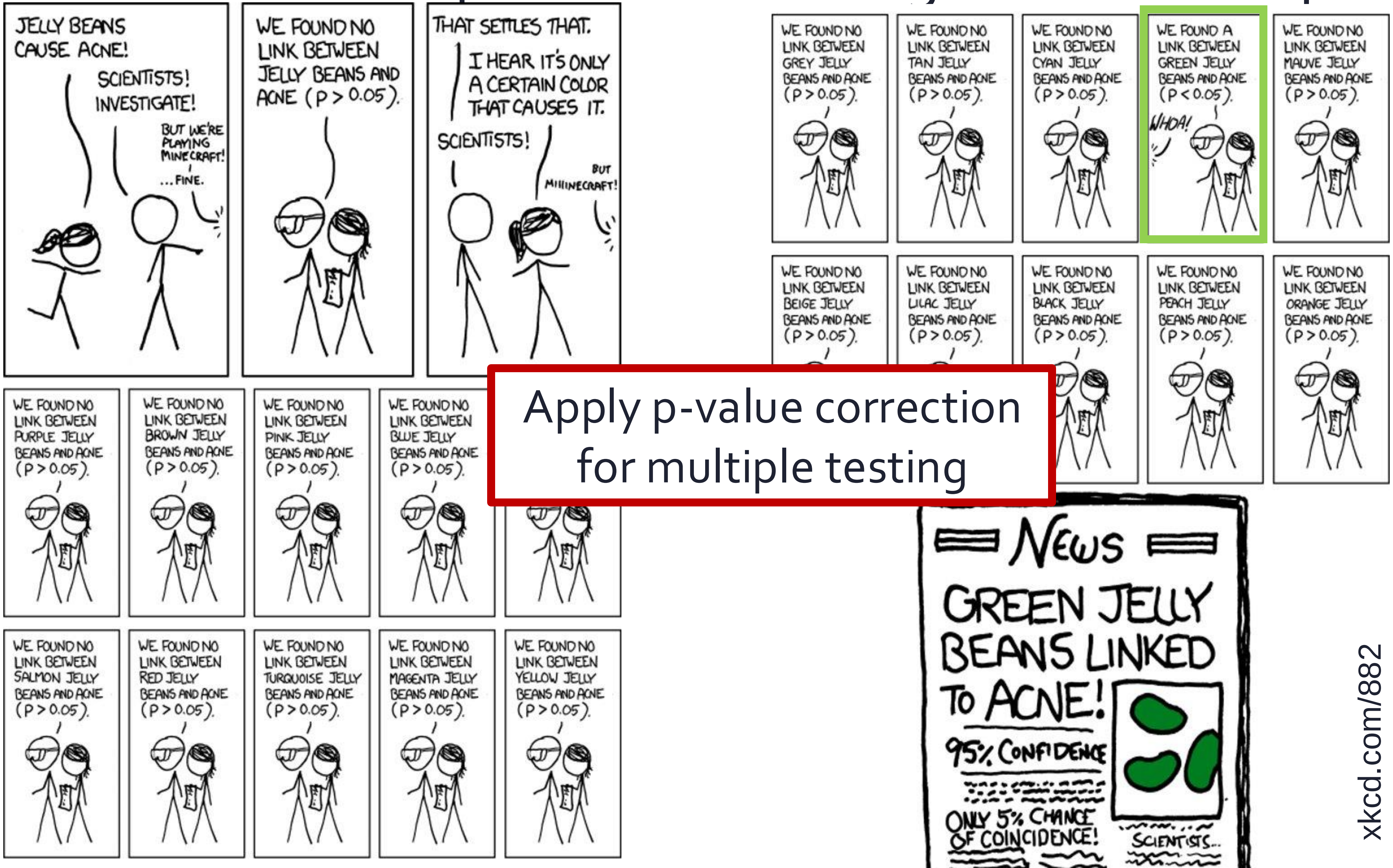
- Large number of genes

Differential Expression : challenges for RNAseq


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

xkcd.com/882

Differential Expression : challenges for RNAseq


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Apply p-value correction for multiple testing

xkcd.com/882

Input for Differential Expression

Counts from mapping

- Affected by library size

TPM from pseudo-aligners

- The R library tximport aggregates counts at the gene-level

Input for Differential Expression

Counts from mapping

- Affected by library size

TPM from pseudo-aligners

- The R library tximport aggregates counts at the gene-level

EdgeR and DESeq2 expect raw counts

Digression : "naïve" normalization

CPM (Count Per Million): count / library size * 10^6

RPKM (Read Per Kilobase per Million): CPM / gene length (kb)

TPM (Transcript Per Million):

- RPK : count / gene length (kb)
- TPM : RPK / sum(RPK) * 10^6

Digression : "naïve" normalization

CPM (Count Per Million): count / library size * 10^6

RPKM (Read Per Kilobase per Million): CPM / gene length (kb)

The sum of RPKM is different between samples

TPM (Transcript Per Million):

- RPK : count / gene length (kb)
- TPM : RPK / sum(RPK) * 10^6

The sum of TPM is constant between samples

Digression : "naïve" normalization

CPM (Count Per Million): count / library size * 10^6

RPKM (Read Per Kilobase per Million): CPM / **gene length** (kb)

TPM (Transcript Per Million):

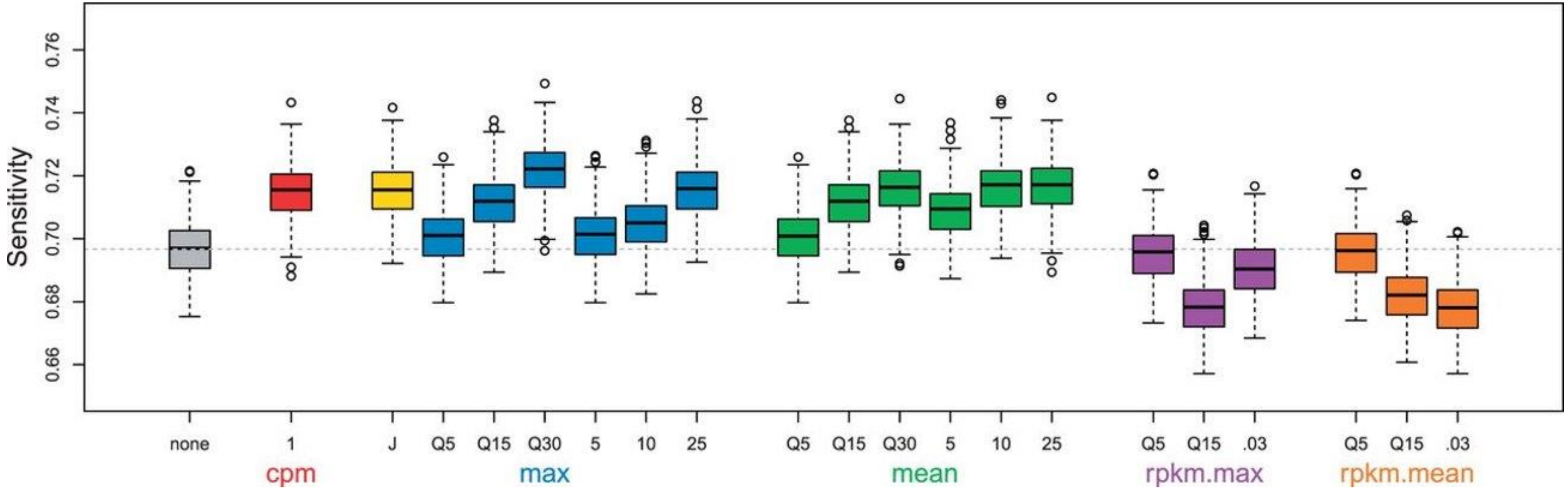
- RPK : count / **gene length** (kb)
- TPM : RPK / sum(RPK) * 10^6

How do you compute "gene length" ?

Differential Expression : filtering low count genes

Very low counts genes:

- Very little information. No chance of DE
- Filtering them out = less test = less p-value correction

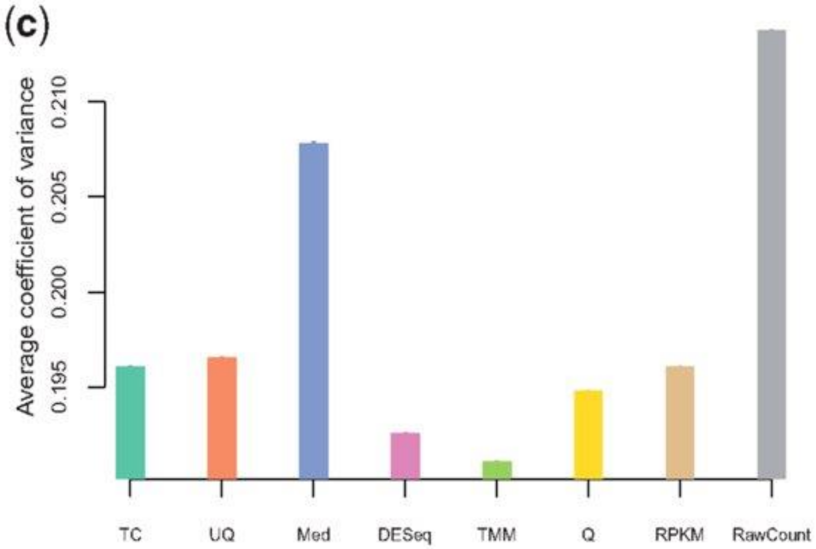


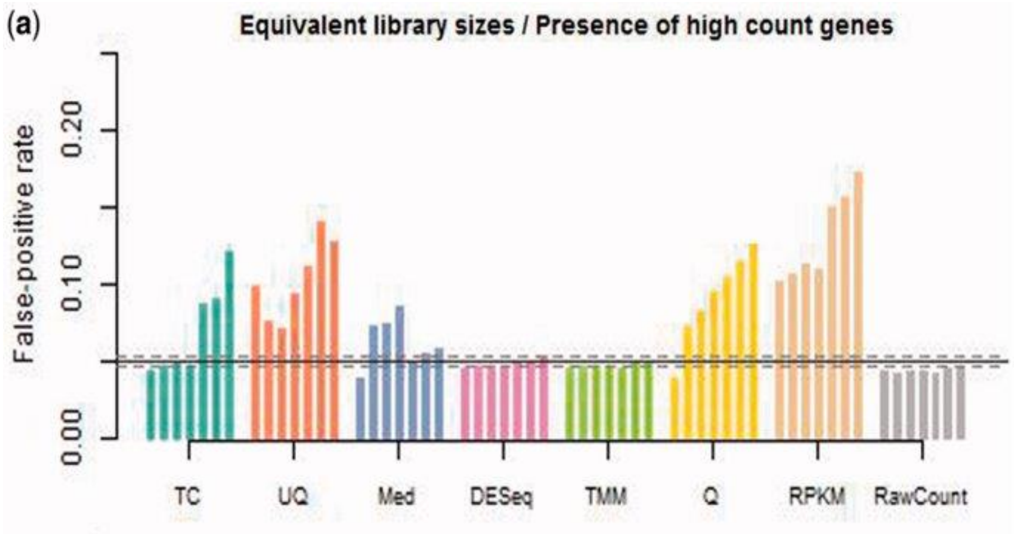
Andrea Rau et al., 2009 https://doi.org/10.1093/bioinformatics/btt350

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

DESeq2: mean normalized count optimizing # of DEG

Differential Expression : normalization





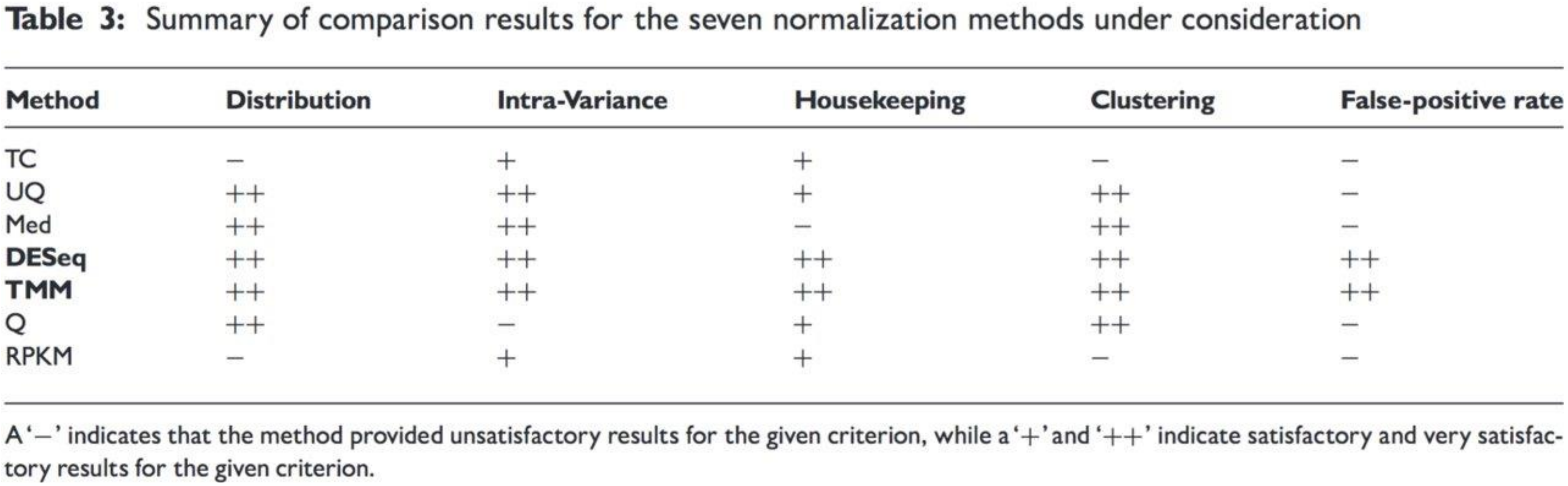


Table 3: Summary of comparison results for the seven normalization methods under consideration

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

Differential Expression : normalization

EdgeR: "Trimmed Mean of M-Values" (TMM)

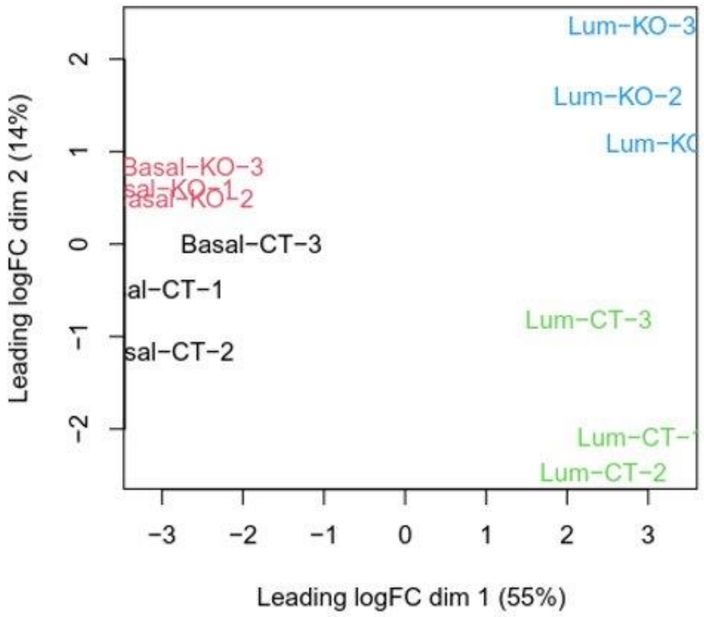
- Set one sample as reference
- For each sample, the TMM is computed as the weighted mean of log ratios between this test and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios.
- Compute the correction factor to get all TMMs to 1

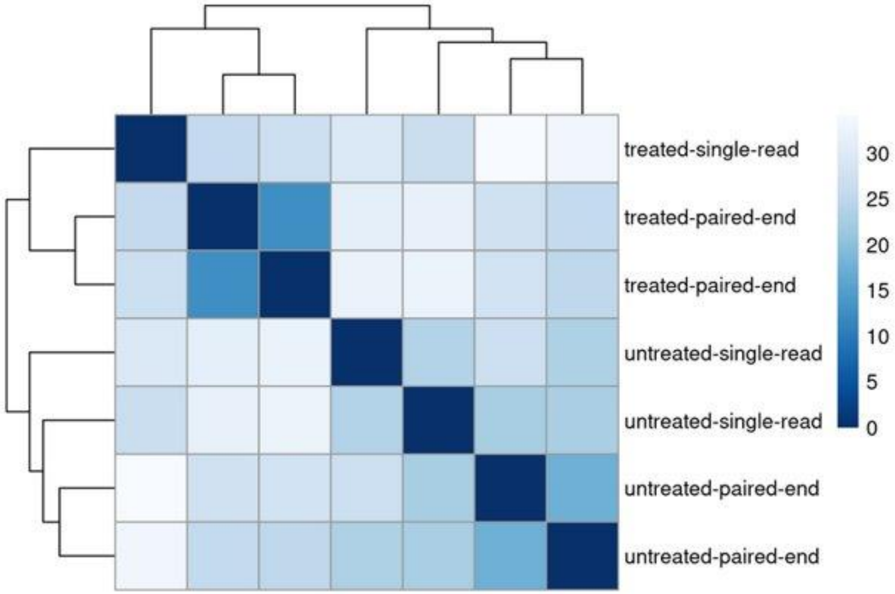
DESeq2: "Relative Log Expression" (RLE)

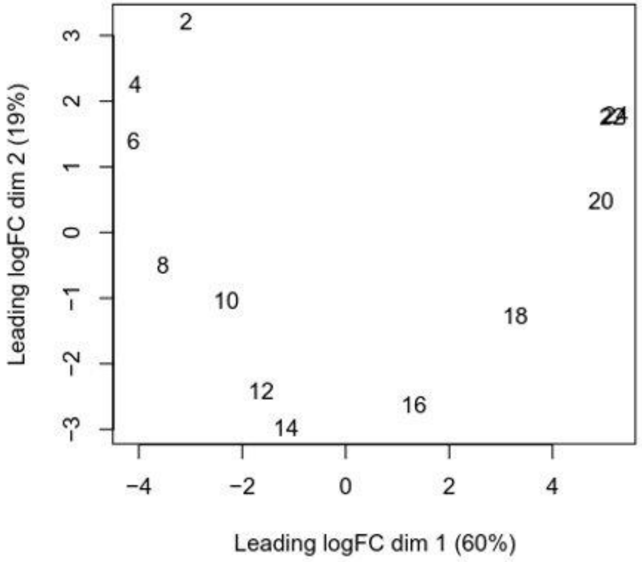
- For each sample: compute the median of the ratio of each gene read count over its geometric mean across all lanes.
- This provides the correction factor that should be applied to all read counts

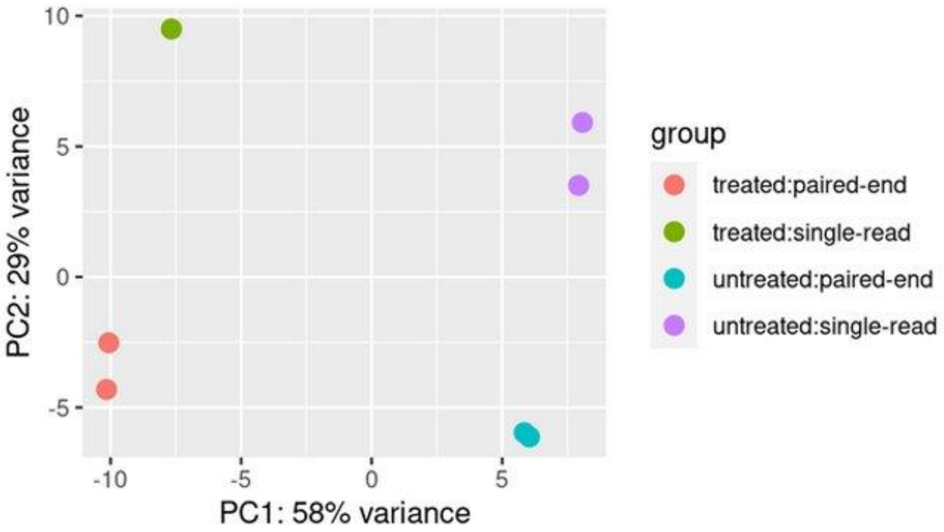
Both presume that most gene are not DE

Quality Control: PDS or PCA of the samples

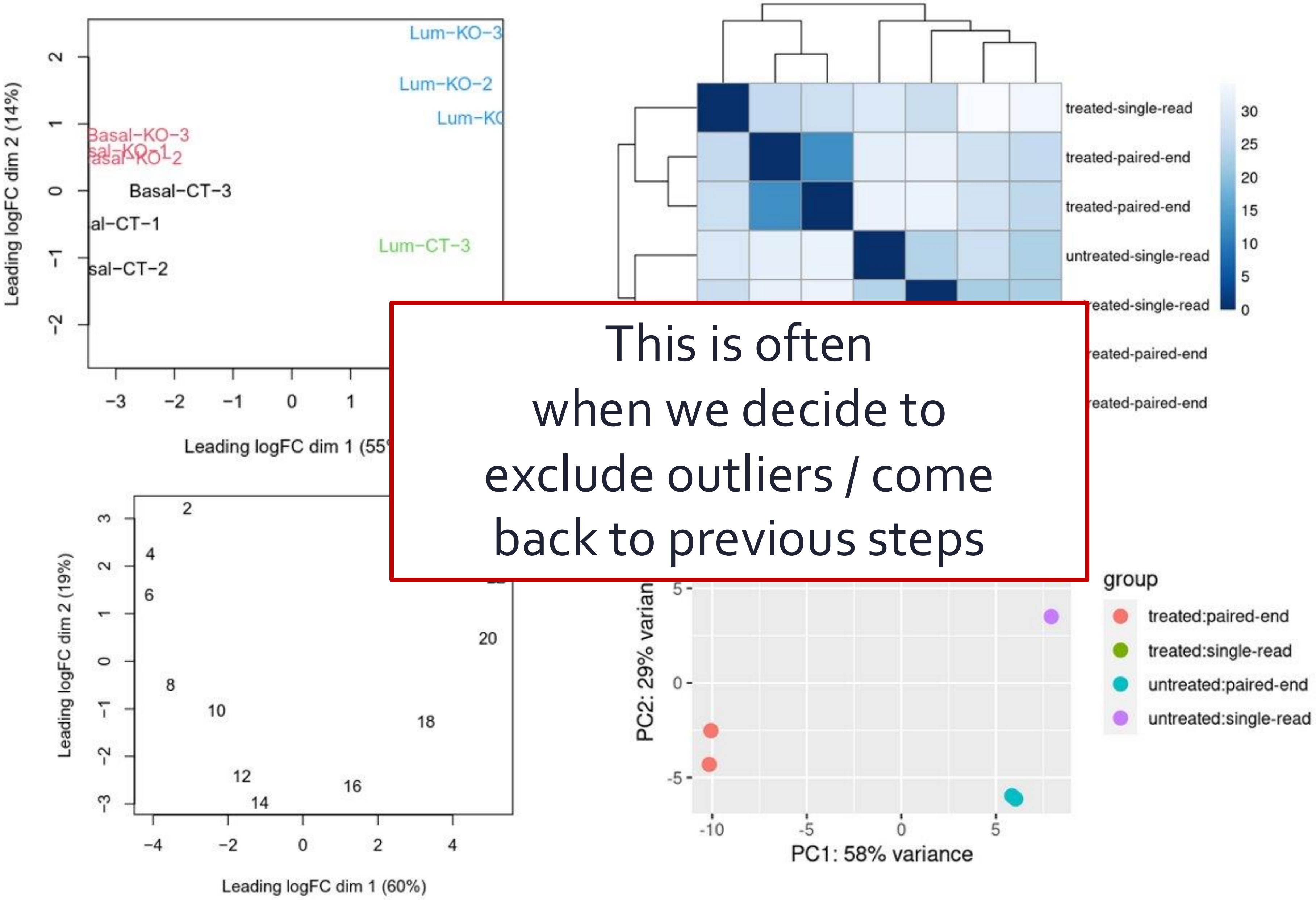








Quality Control: PDS or PCA of the samples



This is often
when we decide to
exclude outliers / come
back to previous steps

Differential Expression : statistical model

To conduct statistical testing, we need an adapted statistical model.

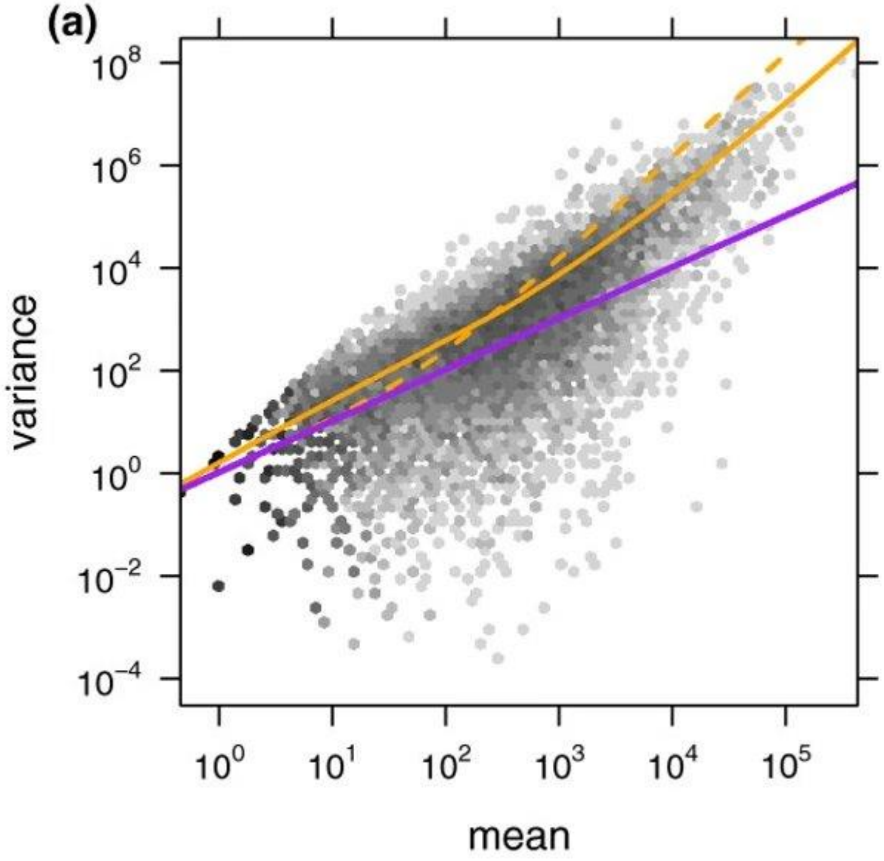
- Idea: expression corresponds to a number of transcripts, captured and sequenced independently from a given "space" (the sample) --> Poisson model

Differential Expression : statistical model

To conduct statistical testing, we need an adapted statistical model.

- Idea: expression corresponds to a number of transcripts, captured and sequenced independently from a given "space" (the sample) --> Poisson model

- There is an over-dispersion!



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

Differential Expression : statistical model

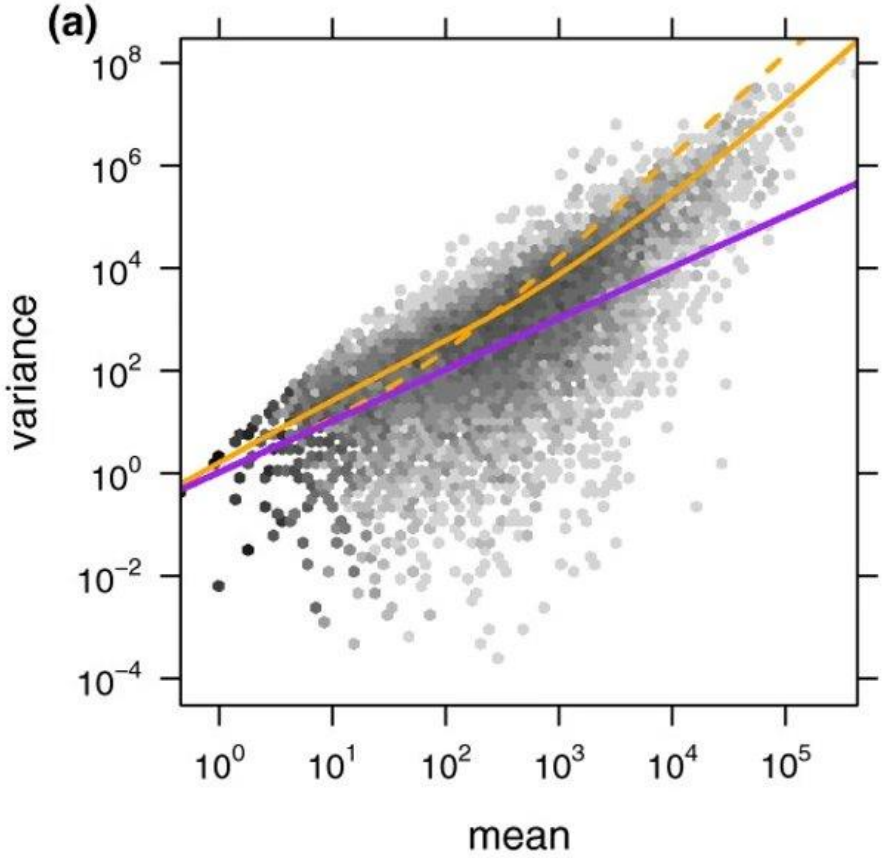
To conduct statistical testing, we need an adapted statistical model.

- Idea: expression corresponds to a number of transcripts, captured and sequenced independently from a given "space" (the sample) --> Poisson model

- There is an over-dispersion
  --> Negative Binomial model

  Variance = $\mu + \theta\mu^2$

  θ: dispersion parameter
  μ: (expected) expression



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

Differential Expression : statistical model

To conduct statistical testing, we need an adapted statistical model.

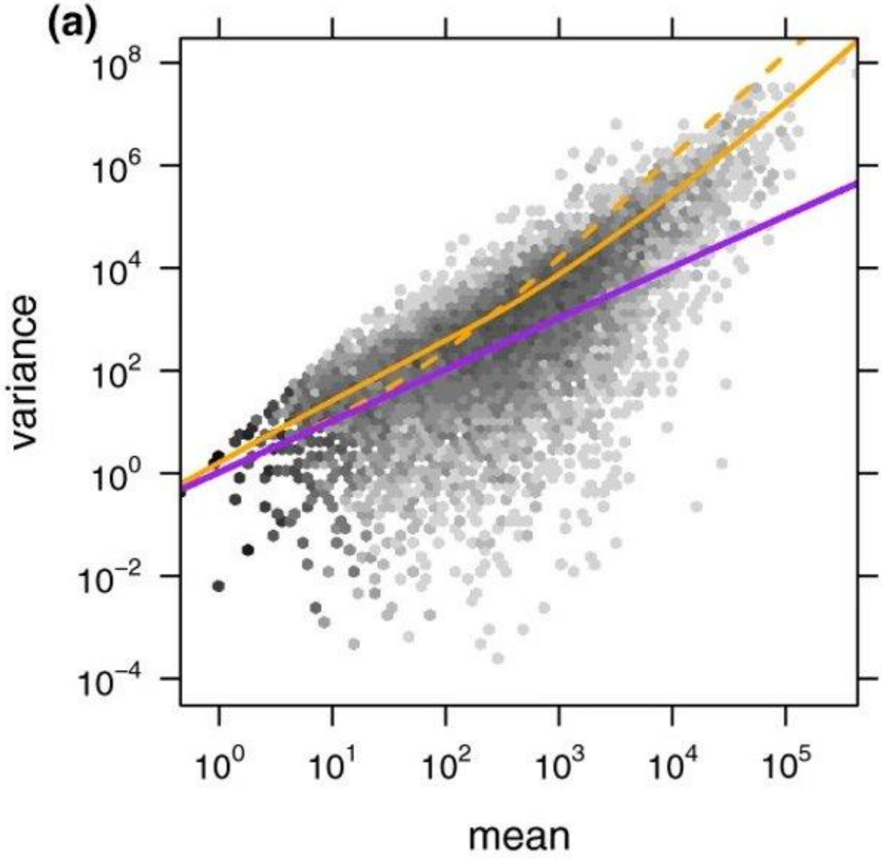
- Idea: expression corresponds to a number of transcripts, captured and sequenced independently from a given "space" (the sample) --> Poisson model

- There is an over-dispersion
  --> Negative Binomial model

  Variance = $\mu + \theta\mu^2$

  θ: dispersion parameter
  μ: (expected) expression

  Often modelled with a linear model:
  μ = base level + genotype effect + batch effect + treatment effect ....

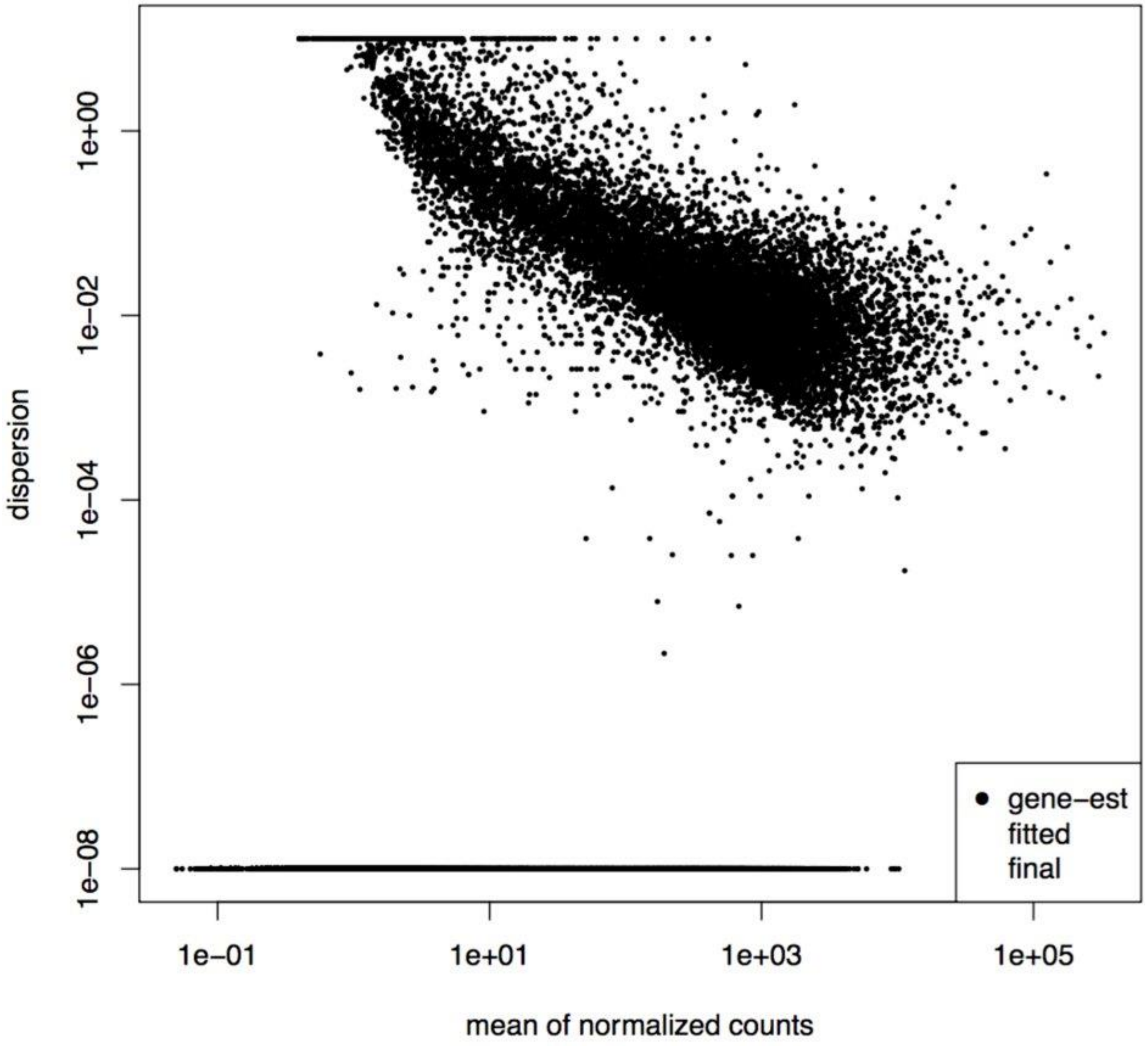


Anders, S., Huber, W. 2010 https://doi.org/10.1186/gb-2010-11-10-r106

Estimating dispersion

Problem:

we often have very few replicates

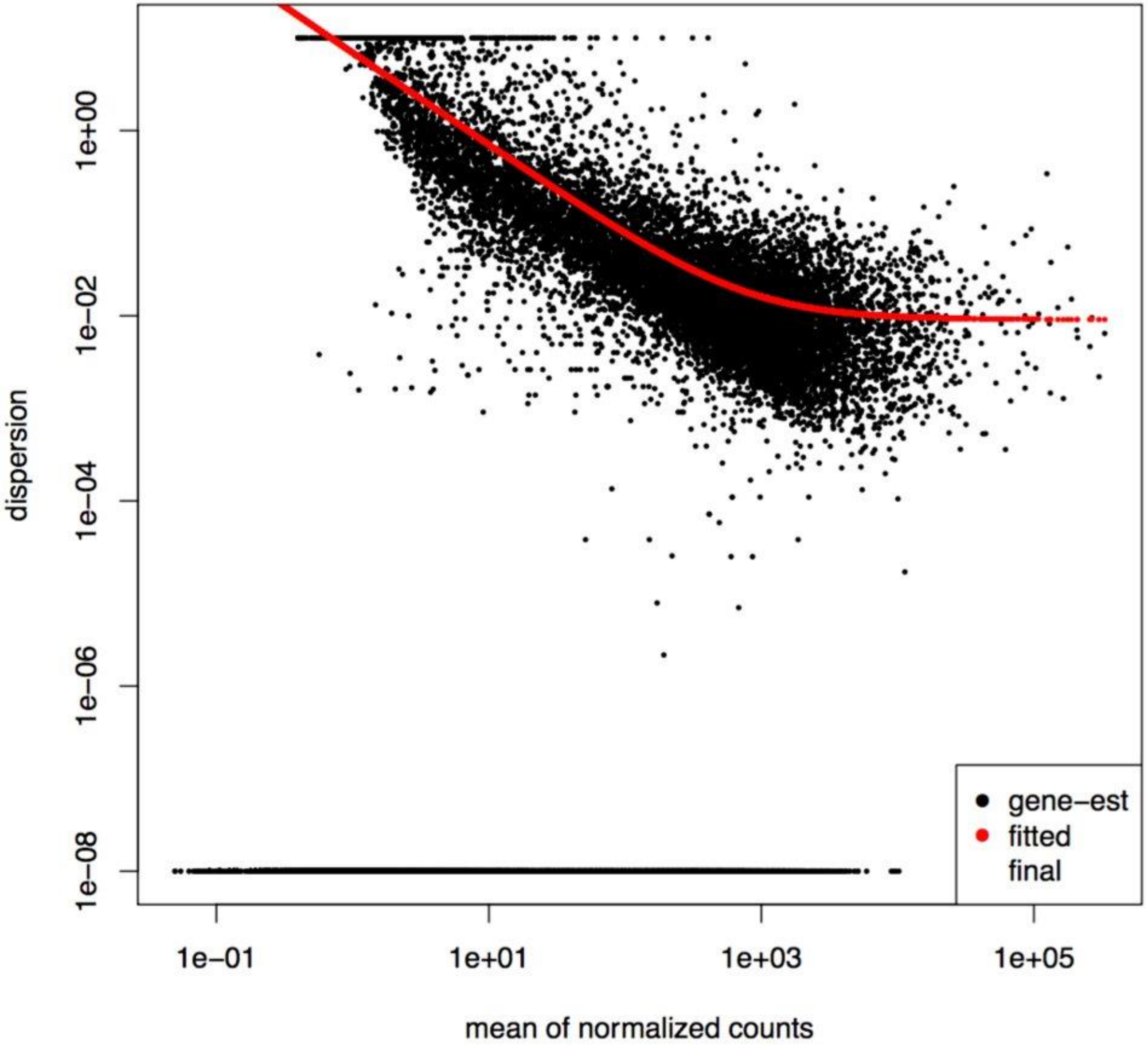


Estimating dispersion

Problem:
we often have very few replicates

Solution:
take advantage of the large number of genes

Shrink gene-wise estimates toward the center value observed of dispersion across **genes with similar expression**.

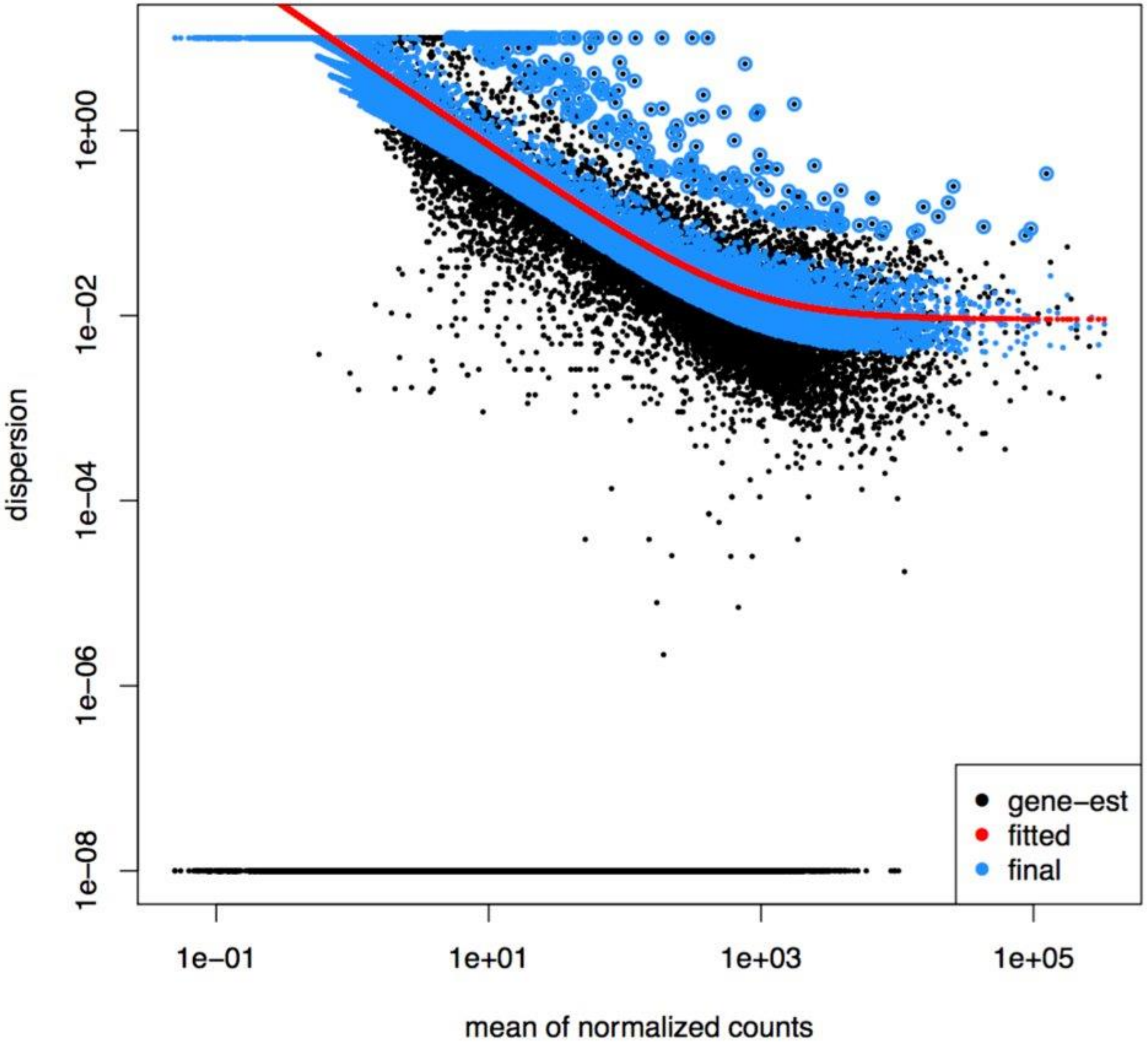


Estimating dispersion

Problem:
we often have very few replicates

Solution:
take advantage of the large number of genes

Shrink gene-wise estimates toward the center value observed of dispersion across **genes with similar expression**.



Testing for differential expression: DESeq2

For each gene:

Z-score = shrunken LFC / estimated standard error

Wald test:

Compare Z-score to a standard normal distribution to compute a **p-value**

Benjamini-Hochberg procedure to **adjust p-values**

Testing for differential expression: edgeR

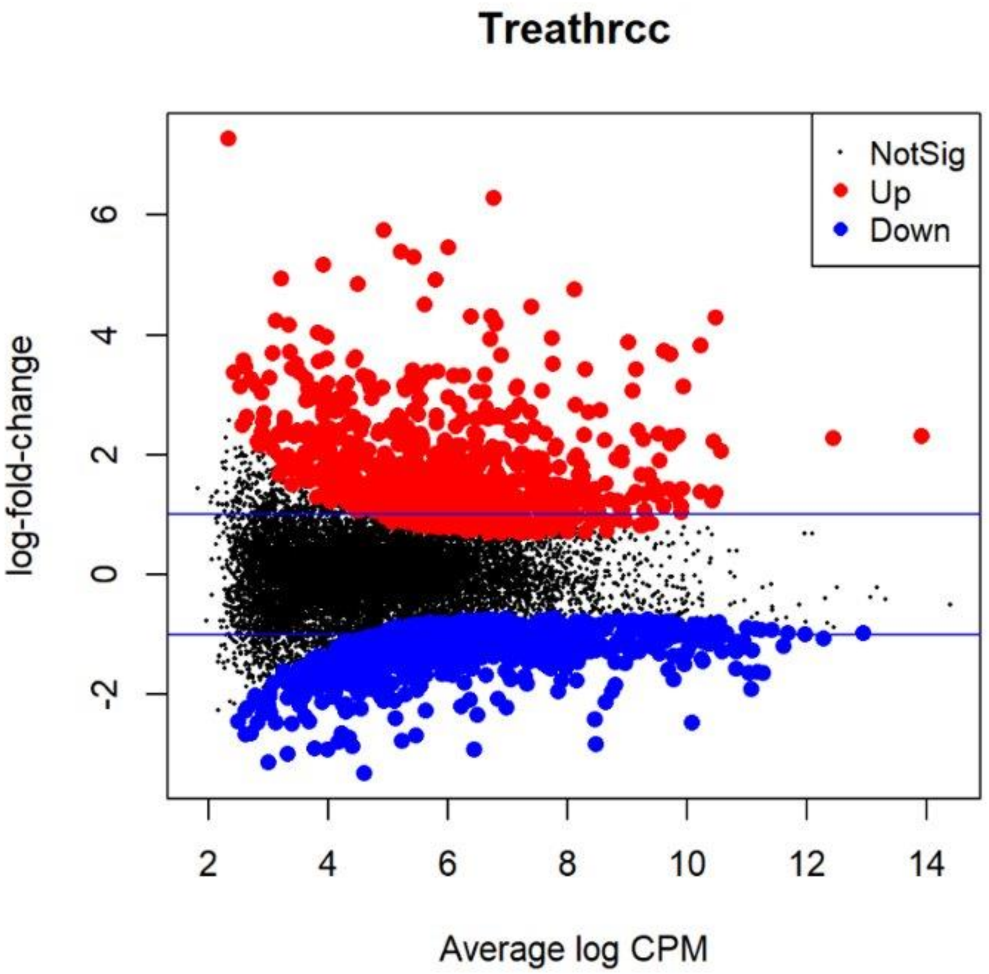
"simple": 1 factor : exactTest()
using the computed conditional distribution for the sum of counts in a group

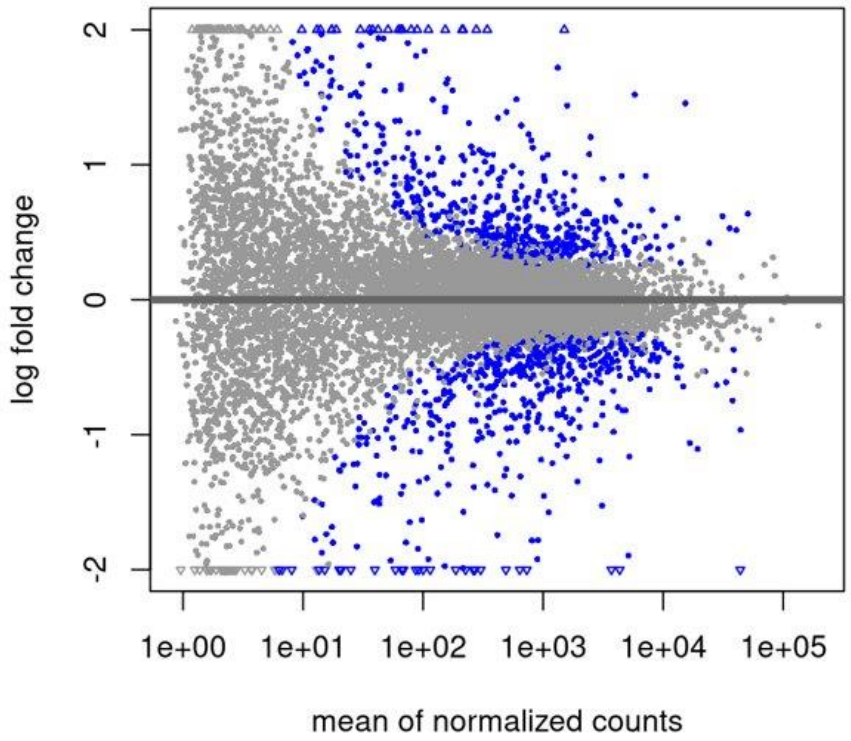
Otherwise: GLM framework

- **Quasi-likelihood F-test** : generally preferred

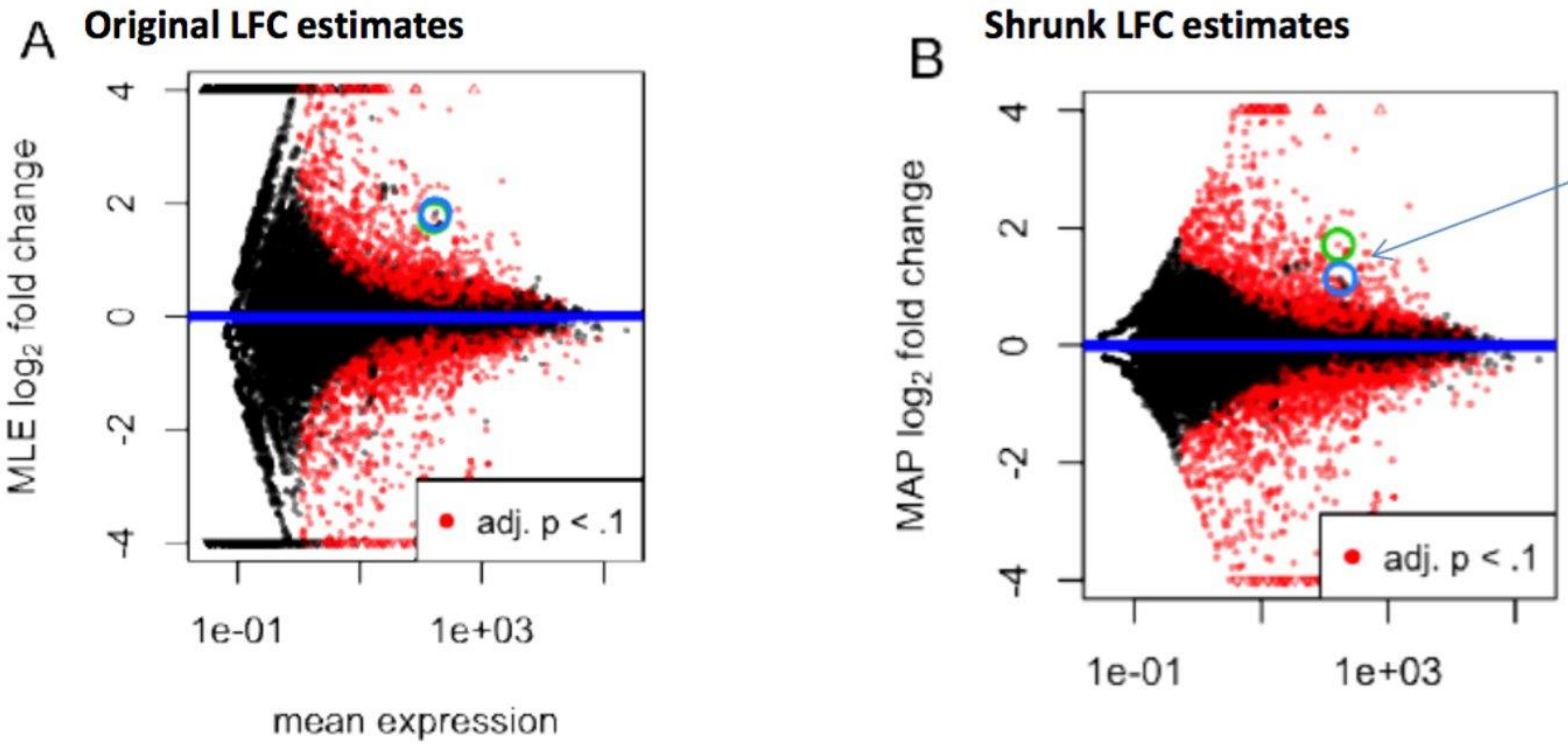
- **Likelihood Ratio Test** : when "the dispersions are very large and the counts are very small, whereby some of the approximations in the QL framework seem to fail"
  https://support.bioconductor.org/p/84291/

DE results: MA plot





DESeq2 : shrinkage of log-fold change



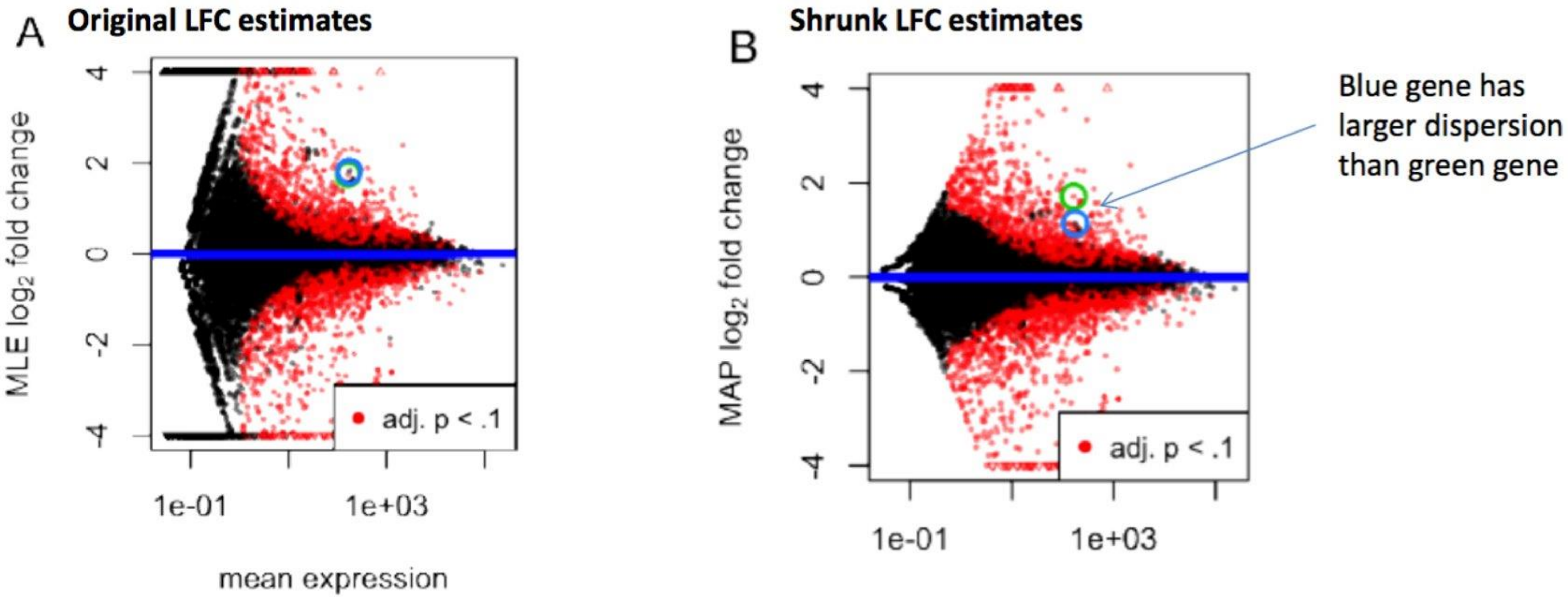
Original LFC estimates

Shrunk LFC estimates

Blue gene has larger dispersion than green gene

Love *et al.* 2014

DESeq2 : shrinkage of log-fold change



A Original LFC estimates

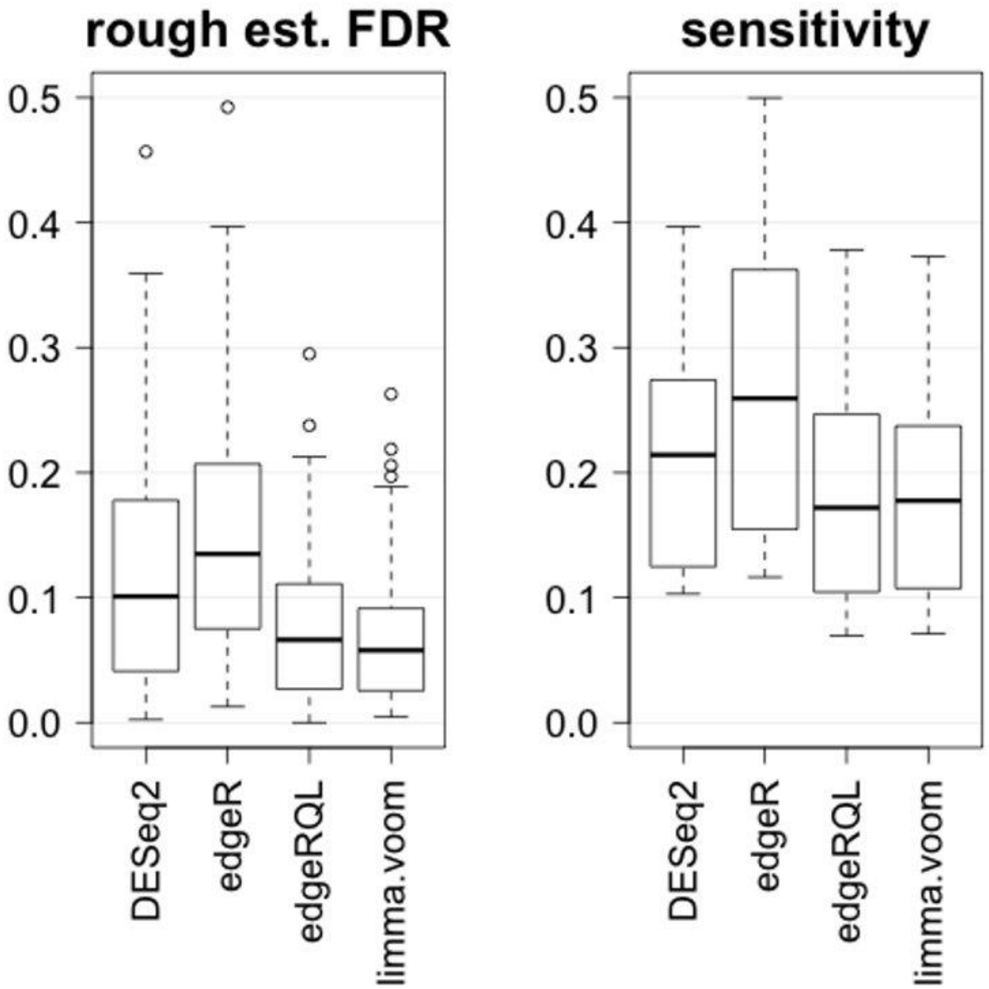
B Shrunk LFC estimates

Blue gene has larger dispersion than green gene

Makes log-fold change values more useful for down-stream analysis

Love *et al.* 2014

edgeR vs DESeq2

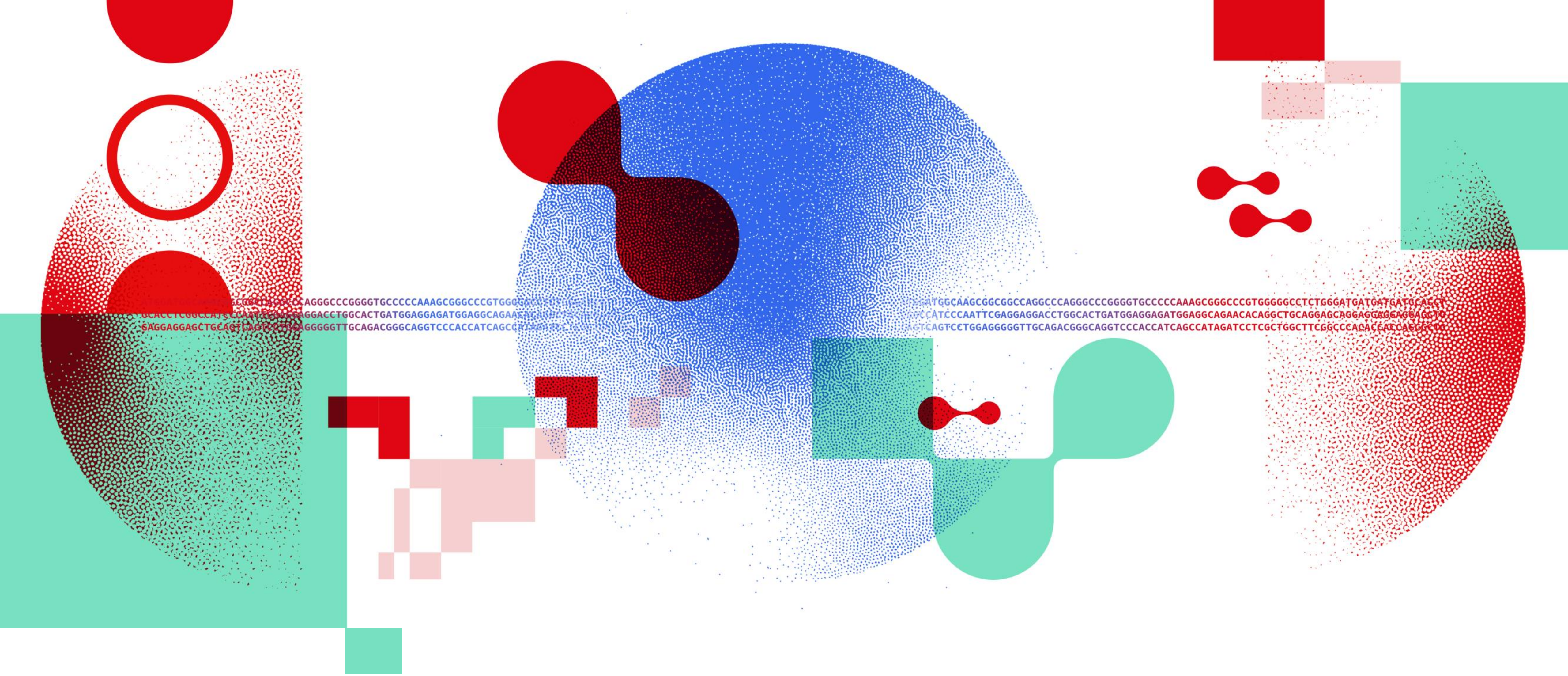


- edgeR exact test : more sensitive
- edgeR QL : more conservative
- DESeq2 : thight FDR control

Love *et al.* 2014

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

Practical



Thank you

elixir SWITZERLAND

SIB