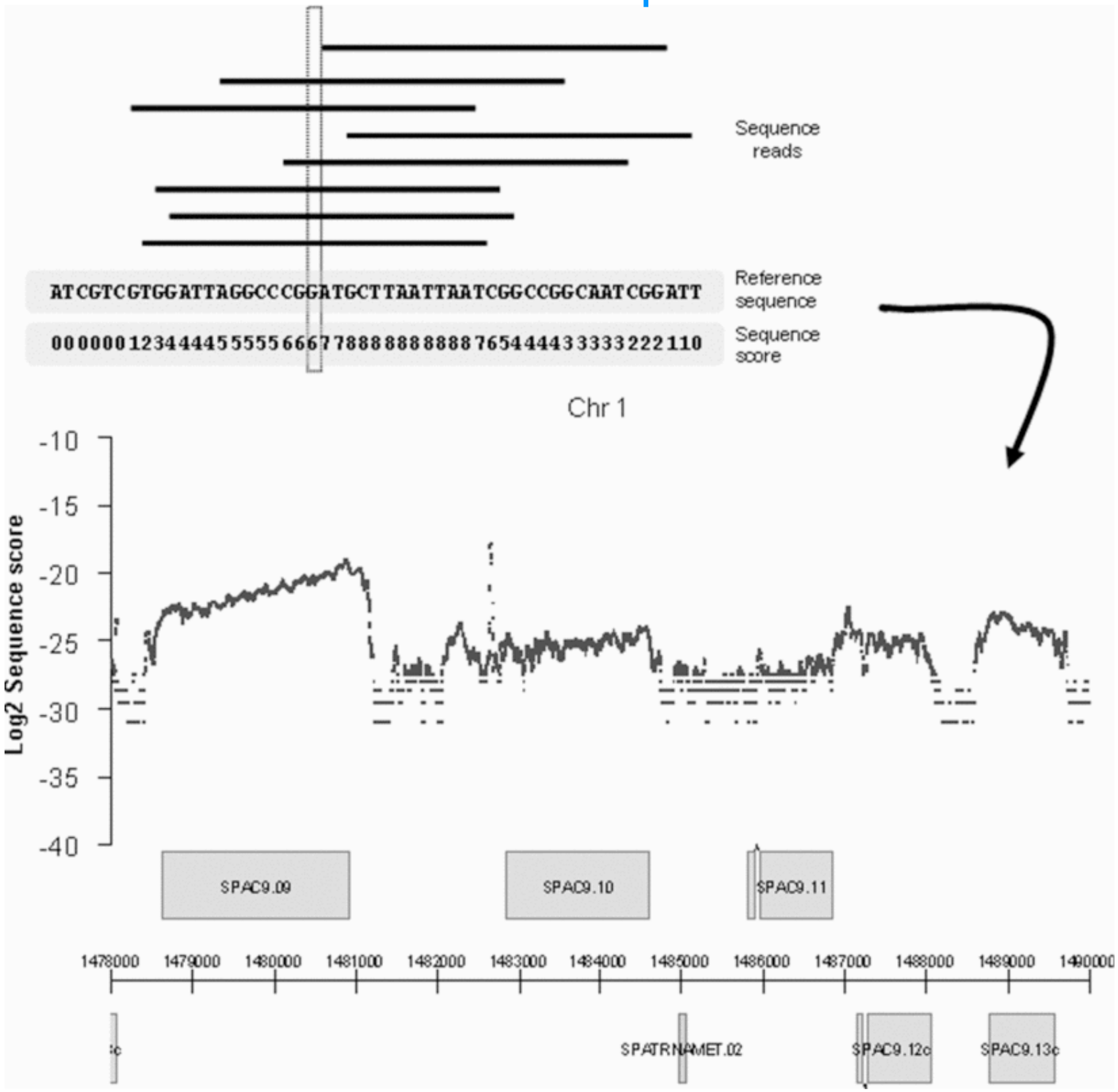
RNA-Seq & Differential Expression analysis

Susan Holmes, Nikos Ignatiadis and Wolfgang Huber

RNA-Seq


Sequence reads
ATCGTCGTGGATTAGGCCCGGATGCTTAATTAATCGGCCGGCAATCGGATT
Reference sequence
00000012344445555566677888888888876544433333222110
Sequence score
Chr 1
Log2 Sequence score
-10
-15
-20
-25
-30
-35
-40
SPAC9.09
SPAC9.10
SPAC9.11
1478000
1479000
1480000
1481000
1482000
1483000
1484000
1485000
1486000
1487000
1488000
1489000
1490000
SPATRNAMET.02
SPAC9.12c
SPAC9.13c

The count matrix

$$
\begin{array}{l}
 \\
\text{gene 1} \\
\text{gene 2} \\
\text{gene 3} \\
\vdots \\
\text{gene m}
\end{array}
\begin{array}{c}
\begin{array}{cccc}
\text{sample 1} & \text{sample 2} & \dots & \text{sample n}
\end{array} \\
\begin{pmatrix}
y_{11} & y_{12} & \dots & y_{1n} \\
y_{21} & y_{22} & \dots & y_{2n} \\
y_{31} & y_{32} & \dots & y_{3n} \\
\vdots & \vdots & \ddots & \vdots \\
y_{m1} & y_{m2} & \dots & y_{mn}
\end{pmatrix}
\end{array}
$$

The pasilla data

```
dim(counts)
```

```
## [1] 14599     7
```

```
counts[ 2000+(0:3), ]
```

```
##              untreated1 untreated2 untreated3 untreated4
## FBgn0020369       3387       4295       1315       1853
## FBgn0020370       3186       4305       1824       2094
## FBgn0020371          1          0          1          1
## FBgn0020372         38         84         29         28
##              treated1 treated2 treated3
## FBgn0020369     4884     2133     2165
## FBgn0020370     3525     1973     2120
## FBgn0020371        1        0        0
## FBgn0020372       63       28       27
```

```
annotationFile = system.file("extdata",
  "pasilla_sample_annotation.csv",
  package = "pasilla", mustWork = TRUE)
pasillaSampleAnno = readr::read_csv(annotationFile)
pasillaSampleAnno
```

```
## # A tibble: 7 x 6
##           file condition       type `number of lanes`
##          <chr>     <chr>      <chr>             <int>
## 1   treated1fb   treated single-read                 5
## 2   treated2fb   treated  paired-end                 2
## 3   treated3fb   treated  paired-end                 2
## 4 untreated1fb untreated single-read                 2
## 5 untreated2fb untreated single-read                 6
## 6 untreated3fb untreated  paired-end                 2
## 7 untreated4fb untreated  paired-end                 2
## # ... with 2 more variables: `total number of reads` <chr>,
## #   `exon counts` <int>
```

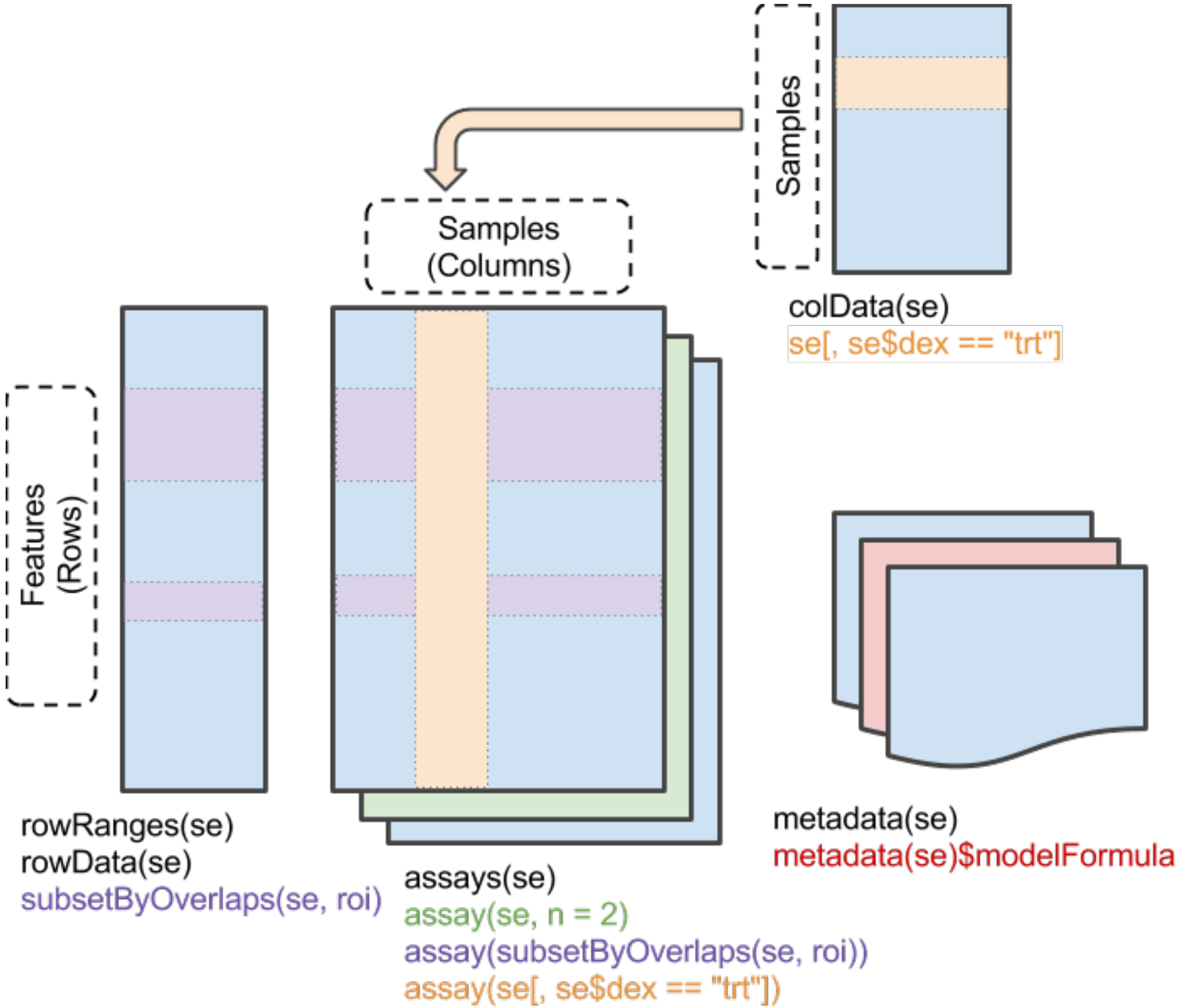
Analogous data structures in

- RNA-Seq
- ChIP-Seq
- Peptides in mass spec
- Microbiomes
- ...

```
with(pasillaSampleAnno,
       table(condition, type))
```

```
##             type
## condition    single-read paired-end
##   untreated           2          2
##   treated             1          2
```

The SummarizedExperiment class



Overview of this lecture

- Understand how DESeq2 works
- Do this in a step-by-step way starting from the basics
- Model things simplistically, see what statistical questions we care about and how to solve them.
- Add layers of complexity and repeat.
- See how far we can get!

Challenge 1: Count data

- Model:

$$Y_1^{\text{treat}}, \ldots, Y_{n_1}^{\text{treat}} \sim \text{Poisson}(\mu_{\text{treat}})$$
$$Y_1^{\text{control}}, \ldots, Y_{n_2}^{\text{control}} \sim \text{Poisson}(\mu_{\text{control}})$$

- Estimate:

$$\mu_{\text{treat}}, \mu_{\text{control}}$$

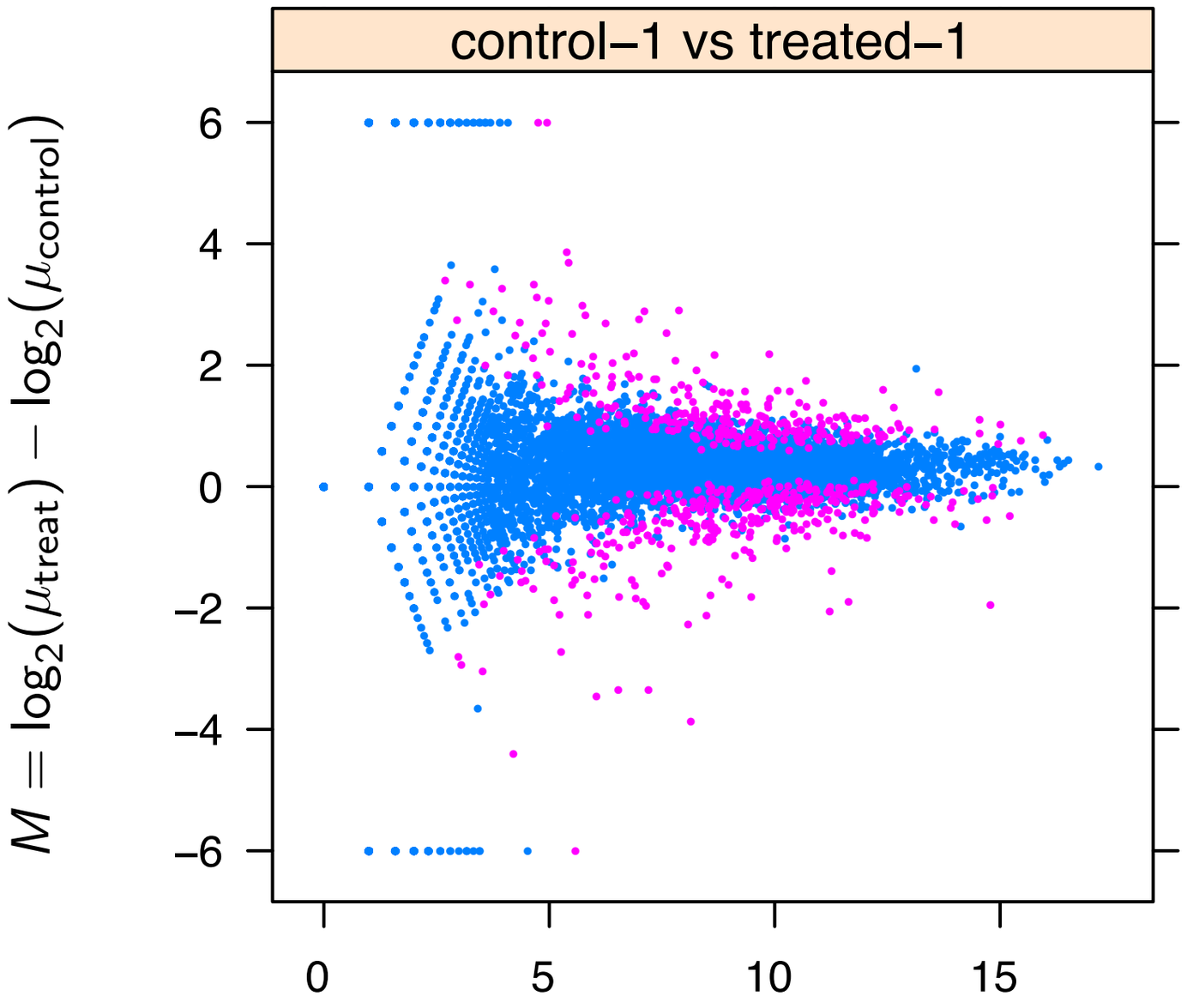
- Estimate the lfc (log-fold change):

$$\text{lfc} = \log_2\left(\frac{\mu_{\text{treat}}}{\mu_{\text{control}}}\right) = \log_2(\mu_{\text{treat}}) - \log_2(\mu_{\text{control}})$$

- Test:

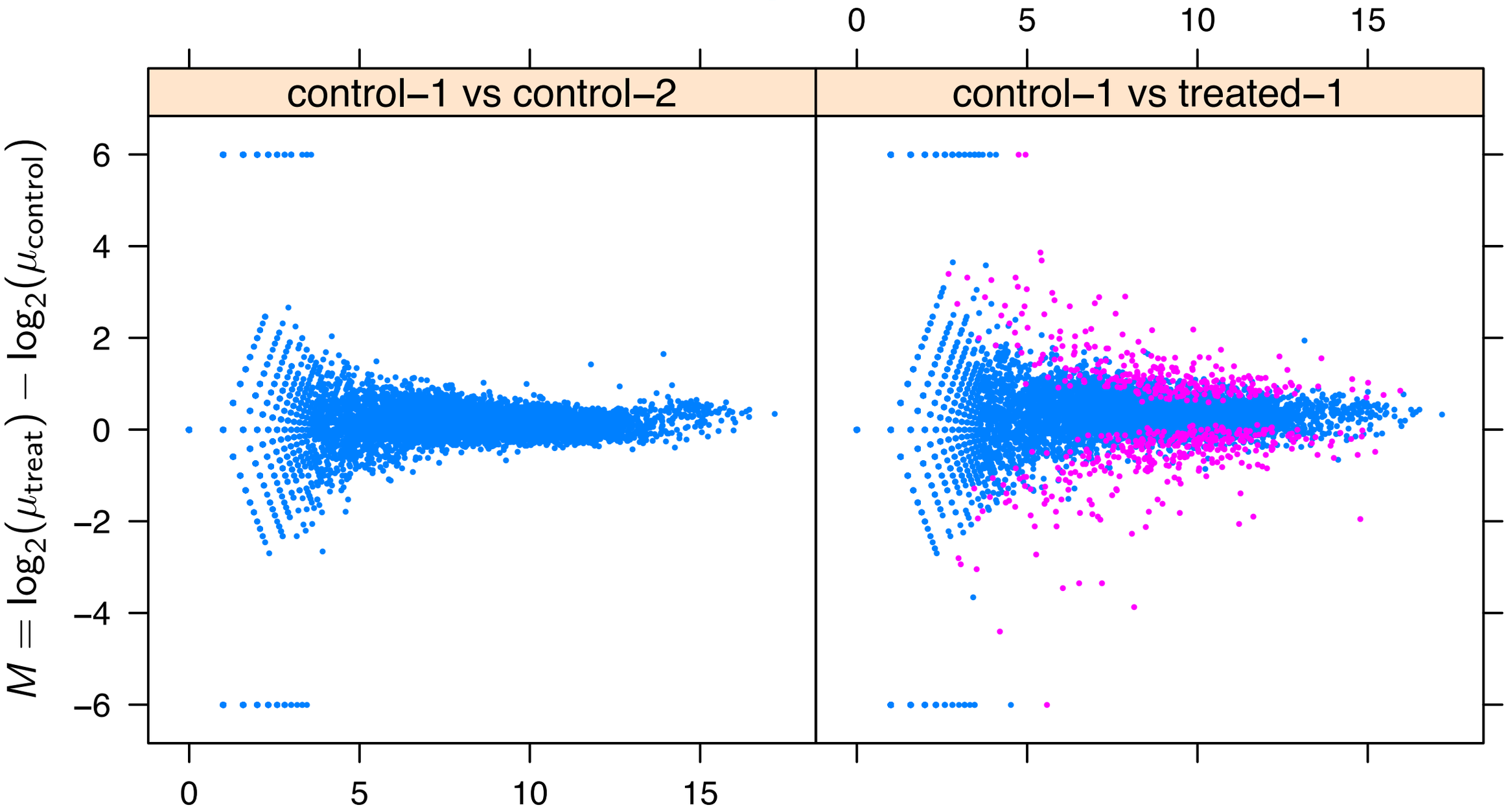
$$H_0 : \mu_{\text{treat}} = \mu_{\text{control}} \quad (H_0 : \text{lfc} = 0)$$

MA-plots



$$A = \frac{1}{2}\left[\log_2(\mu_{\text{treat}}) + \log_2(\mu_{\text{control}})\right]$$

MA-plots



$$A = \frac{1}{2}\left[\log_2(\mu_{\text{treat}}) + \log_2(\mu_{\text{control}})\right]$$

two biological replicates

treatment vs control

Challenge 1: Poisson model, estimation

- Recall that we can estimate $\mu_{\text{treat}}, \mu_{\text{control}}$ using Maximum Likelihood.
- This yields the estimators which are just the sample averages:

$$\hat{\mu}_{\text{treat}} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i^{\text{treat}}$$

$$\hat{\mu}_{\text{control}} = \frac{1}{n_2} \sum_{i=1}^{n_1} Y_i^{\text{control}}$$

- $$\widehat{\text{lfc}} = \log_2 \left(\frac{\hat{\mu}_{\text{treat}}}{\hat{\mu}_{\text{control}}} \right)$$

Challenge 1: Poisson model, testing

- For our test statistic we would like to use $\widehat{\text{lfc}}$, large values (in magnitude) would provide evidence against the null.
- Issue: How to get the distribution under the null hypothesis?
- By simulation, for example we could let

$$\hat{\mu}_0 = \frac{1}{2}\left(\hat{\mu}_{\text{treat}} + \hat{\mu}_{\text{control}}\right)$$

- Then for $b = 1 \ldots, B$ draw:

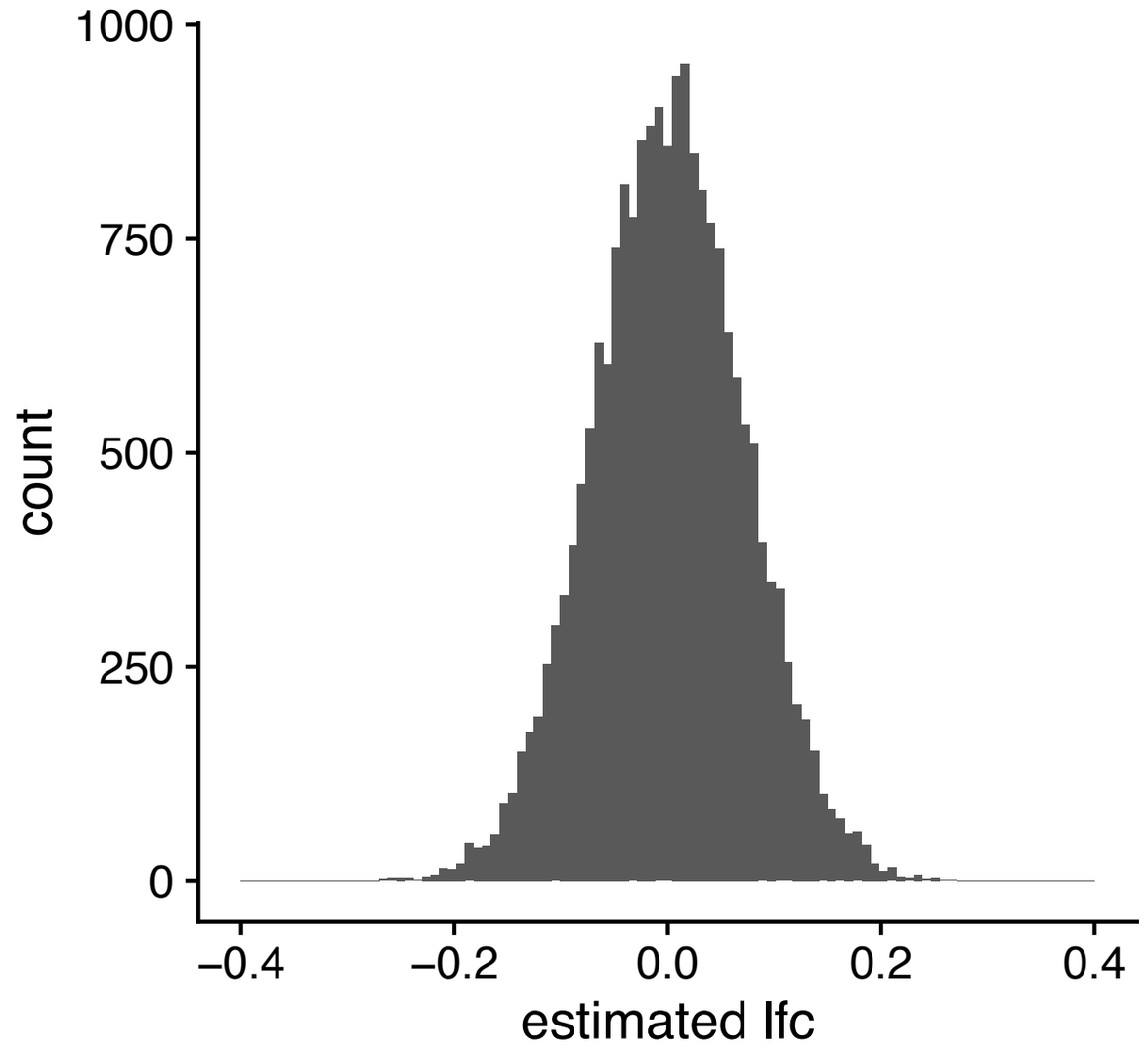
$$Y_{1,b}^{\text{treat}}, \ldots, Y_{n_1,b}^{\text{treat}} \sim \text{Poisson}(\hat{\mu}_0)$$
$$Y_{1,b}^{\text{control}}, \ldots, Y_{n_2,b}^{\text{control}} \sim \text{Poisson}(\hat{\mu}_0)$$

- Calculate $\widehat{\text{lfc}}_b$ based on these.

The (simulated) null distribution

2 treated samples and
2 control samples
average count 200

20000 replications



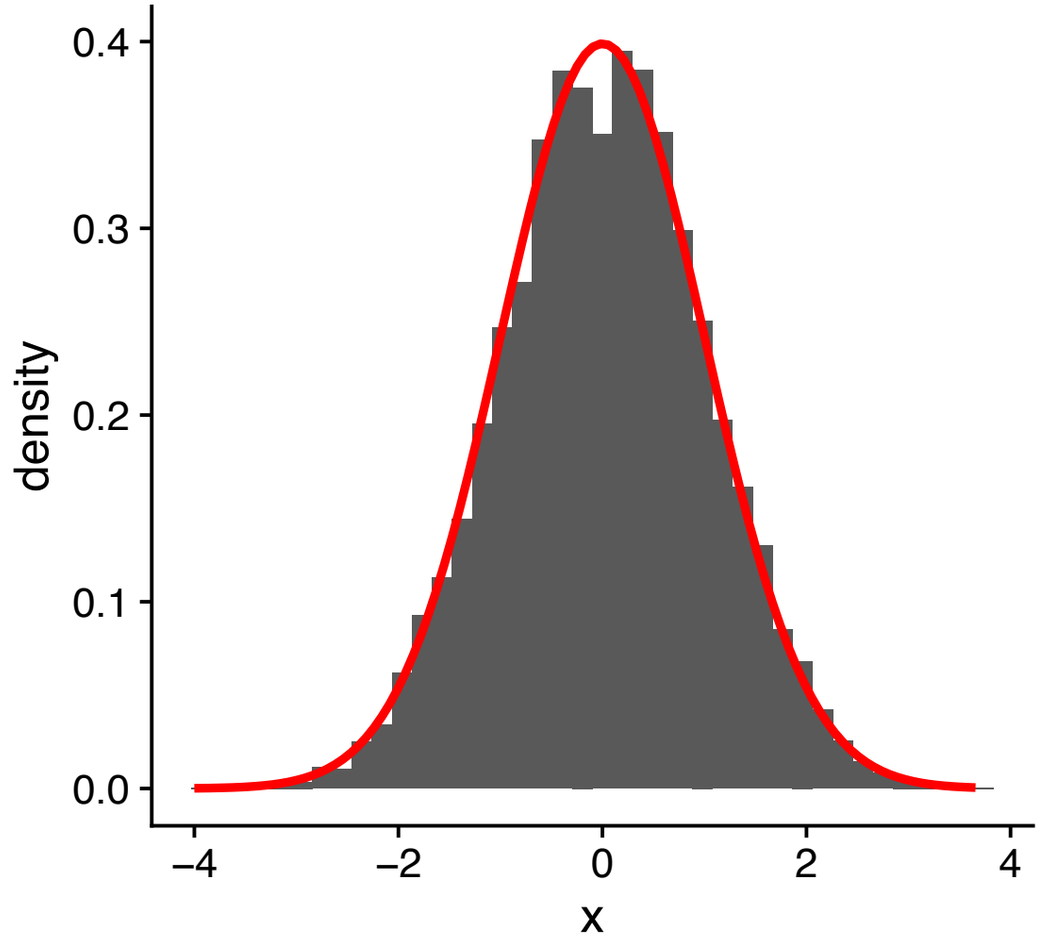
The Wald test

- Use simulation (or other way) to estimate the standard error $\widehat{se}$, i.e. the standard deviation of $\widehat{\text{lfc}}$, then under the null:

$$\frac{\widehat{lfc}}{\widehat{se}} \approx \mathcal{N}(0,1)$$

2 treated samples and
2 control samples
average count 200

20000 replications for histogram



Challenge 2: What about low counts?

For many genes we will have small counts.

Recall our estimator:

$$\widehat{\text{lfc}} = \log_2 \left(\frac{\hat{\mu}_{\text{treat}}}{\hat{\mu}_{\text{control}}} \right)$$

What if we got 3 counts for treatment, 0 for control?

Is the log-fold change infinity?

Occam's razor tells us otherwise.

Occam's razor

"Entities should not be multiplied without necessity."

"The simplest solution is most likely the right one."

https://en.wikipedia.org/wiki/Occam%27s_razor

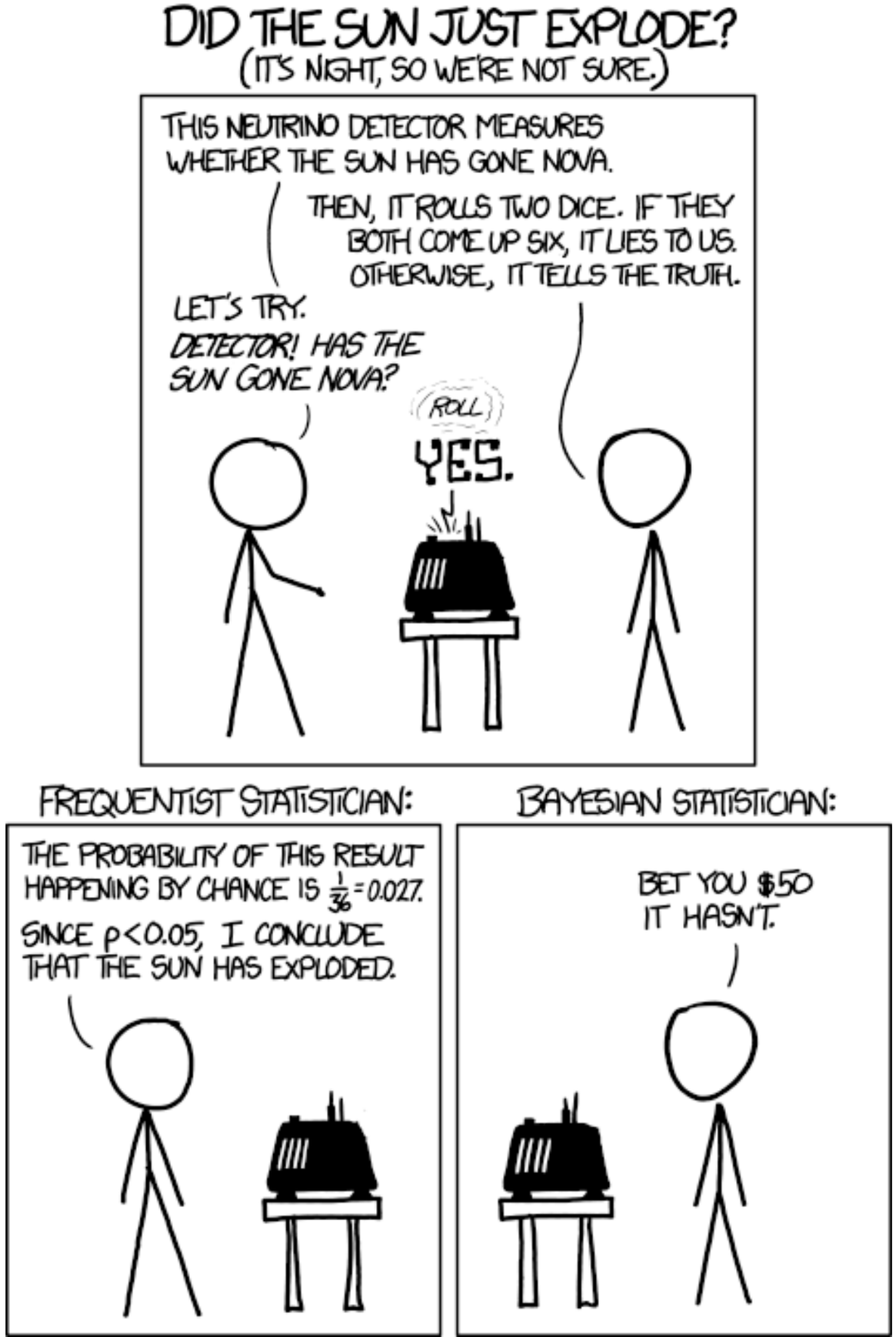
Pseudocounts

- Pseudocounts:

$$\widehat{\text{lfc}} = \log_2 \left(\frac{1 + \hat{\mu}_{\text{treat}}}{1 + \hat{\mu}_{\text{control}}} \right)$$

- LIMMA-Voom (Law, Chen, Shi, Smyth 2014)
- PoiClaClu package (Witten, 2011)

Bayesian statistics



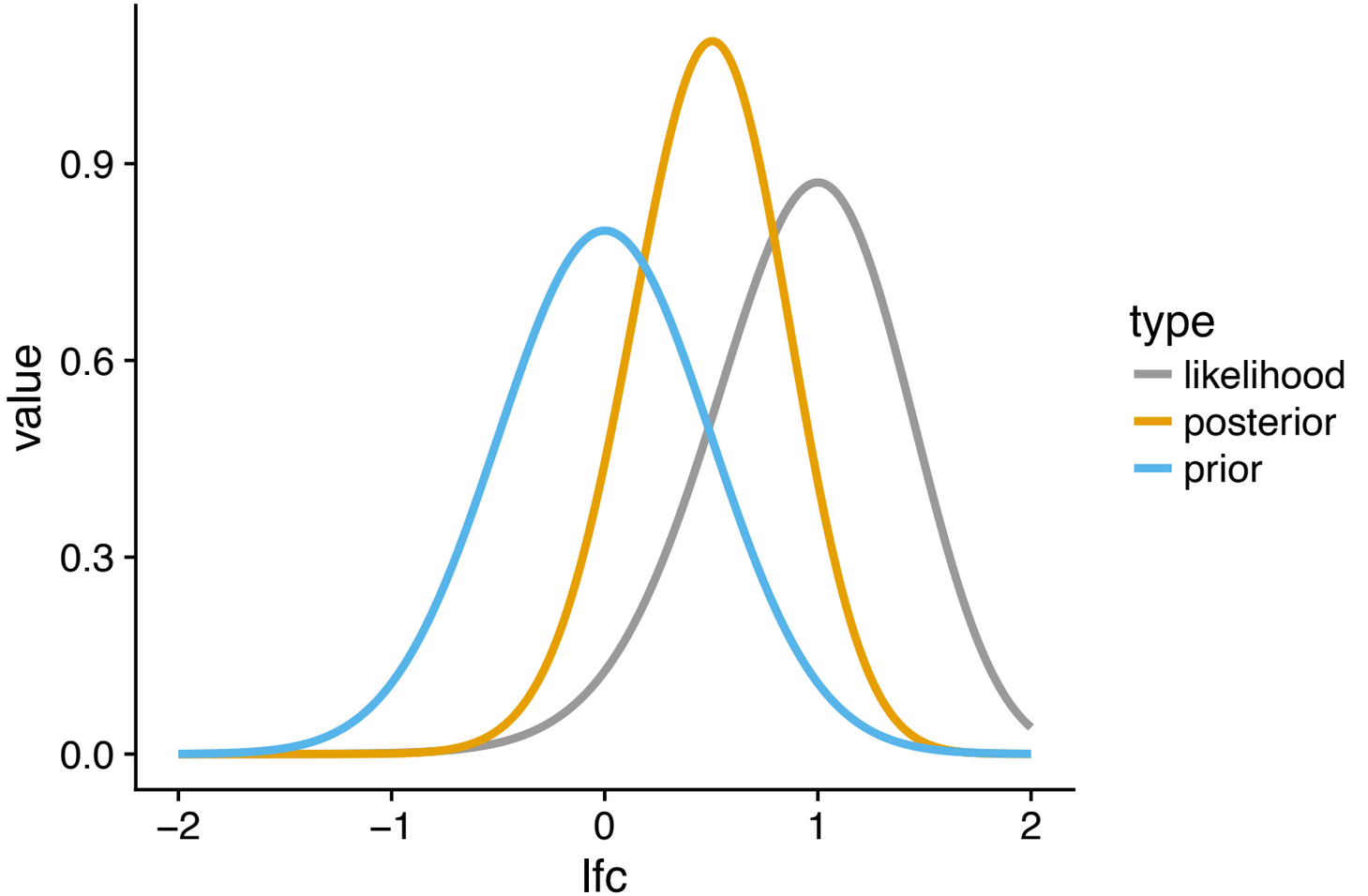
Bayesian approach

- Define prior on lfc

$$\text{lfc} \sim \mathcal{N}\left(0, \sigma^2\right)$$

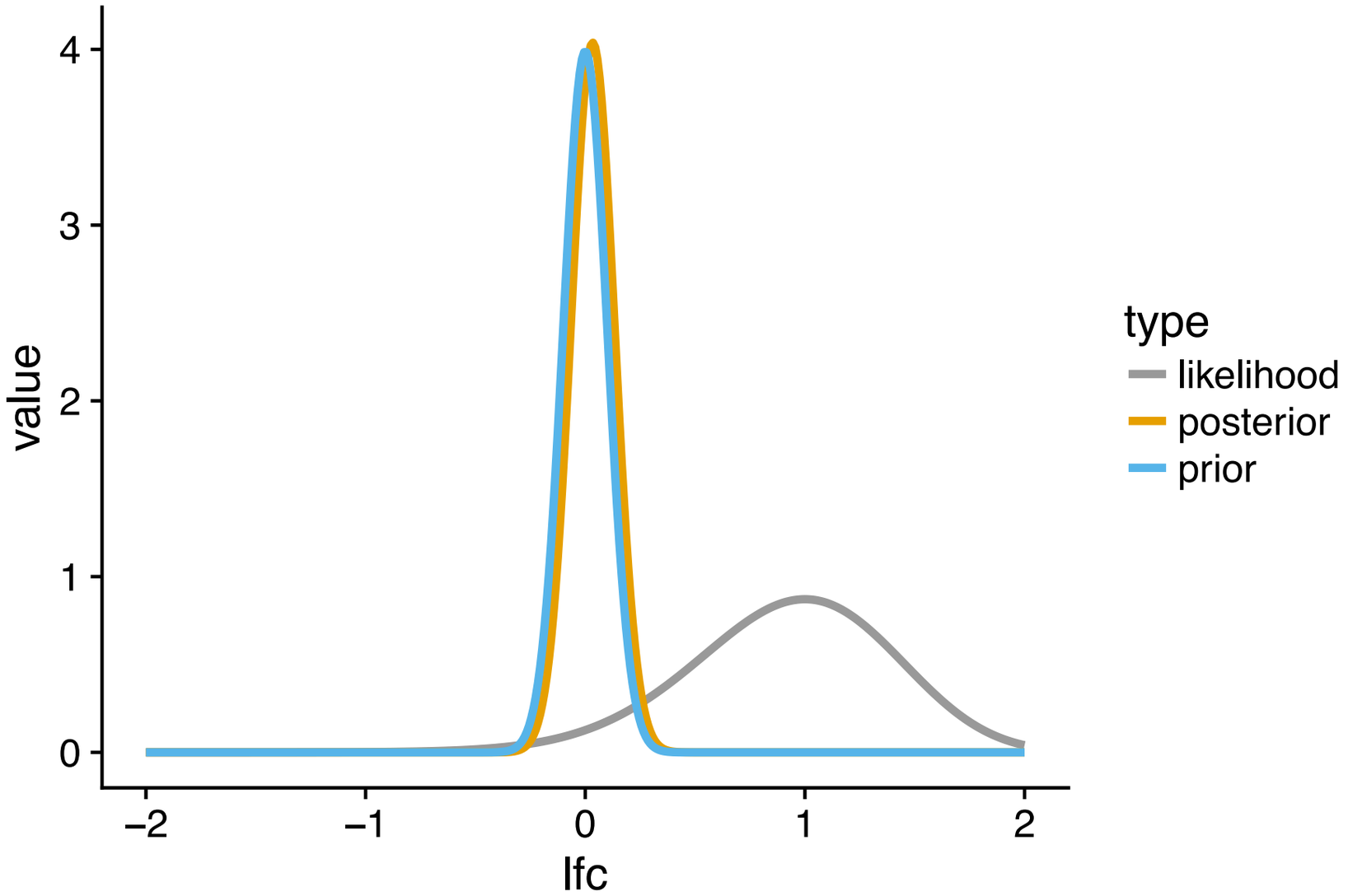
- Recall posterior = prior*likelihood
- Do not look at the maximum of the likelihood
- Look at the maximum of the posterior instead

- Example: σ=0.5,
- 5 counts for control
- 10 counts for treatment



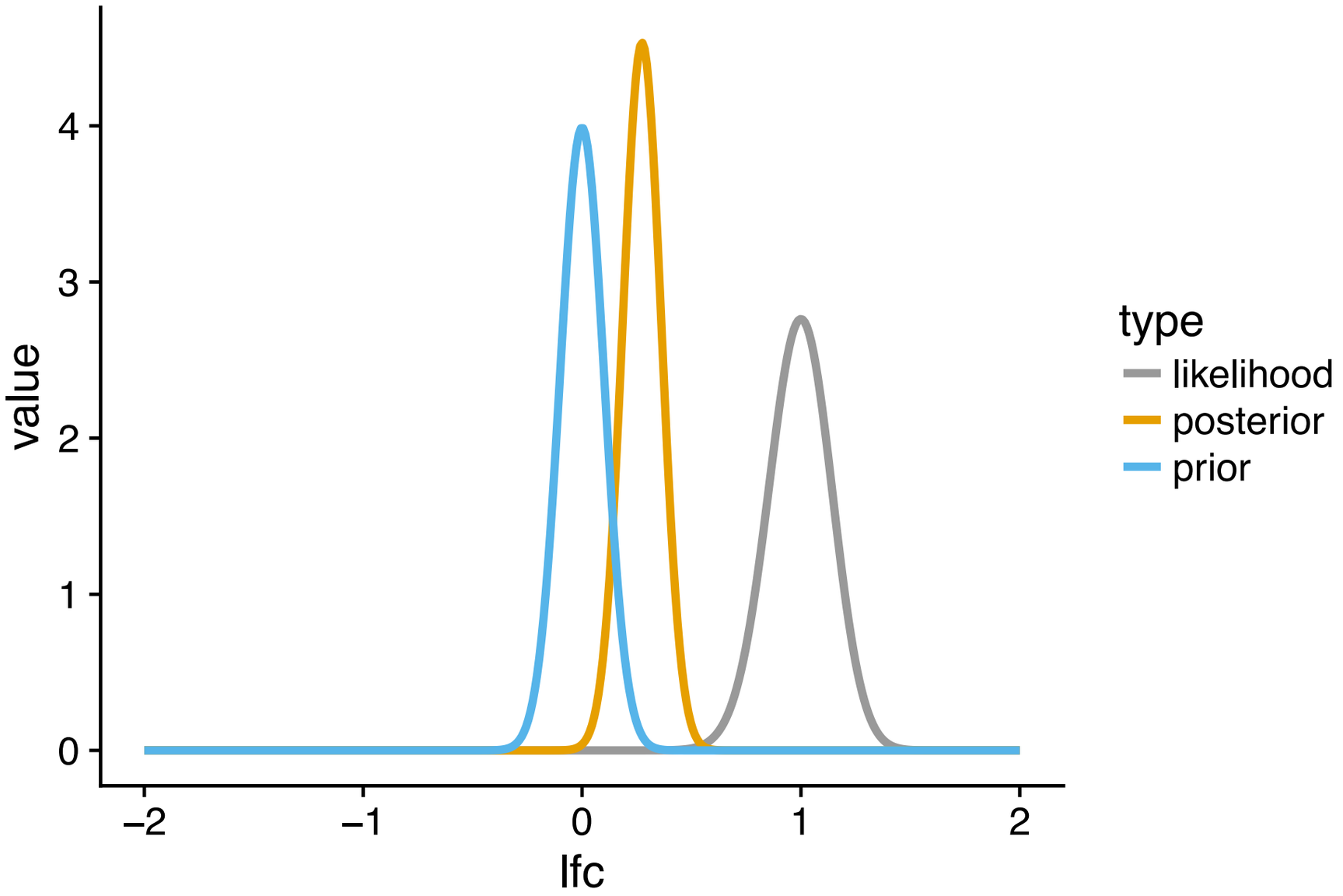
Stronger prior

- Example: σ=0.1,
- 5 counts for control
- 10 counts for treatment



More informative data

- Example: σ=0.1,
- 50 counts for control
- 100 counts for treatment



Remarks on Bayesian approach

- Once we chose prior, adaptive to signal in the data.
- But how to choose the prior?
- (Or how to choose the pseudocounts?)

Challenge 3: Biological variability

Poisson:

$$\text{var} = \text{mean} = \boldsymbol{\lambda}$$

This does not fit real data from biological replicates.

Variance-mean relationship in the pasilla data

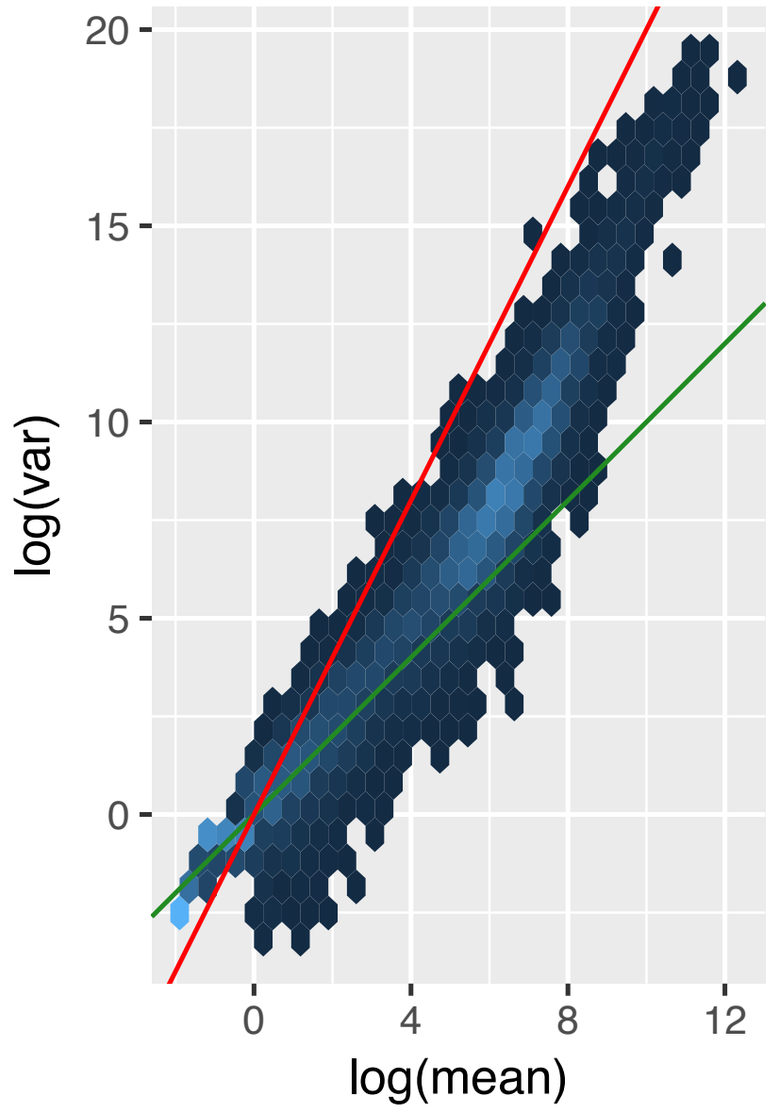


Figure 8.4: Variance versus mean for the (size factor adjusted) `counts` data. The axes are logarithmic. Also shown are lines through the origin with slopes 1 (green) and 2 (red).

$$v = c \cdot m^k$$

$$\Updownarrow$$

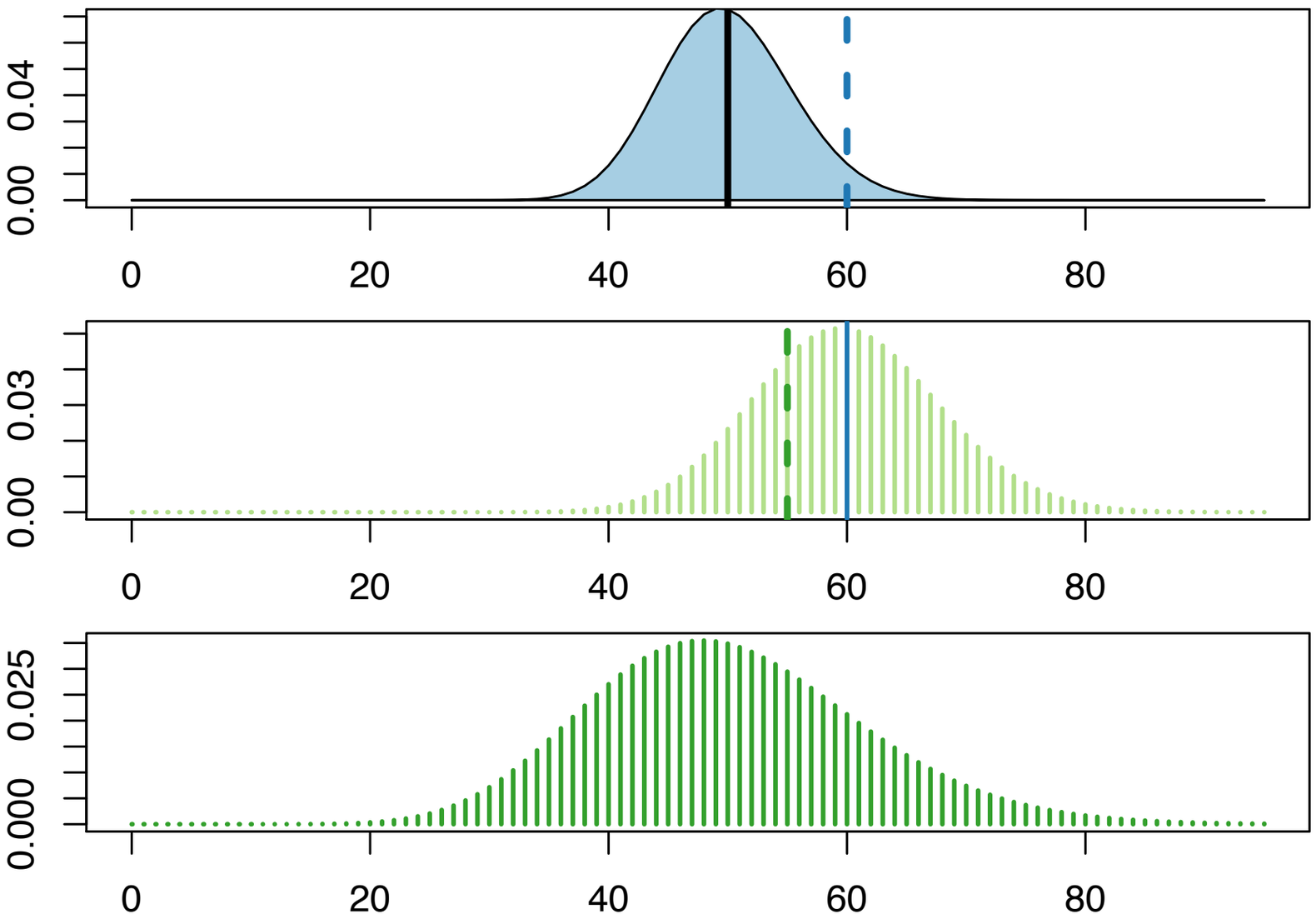
$$\log(v) = k \cdot \log(m) + \log(c)$$

We see evidence for k=1 and k=2, i.e.

$$v = c_1 \cdot m + c_2 \cdot m^2$$

Variance and mean are computed for each row (gene), across the columns (samples)

Challenge 3: Biological variability



Biological sample to sample variability Γ

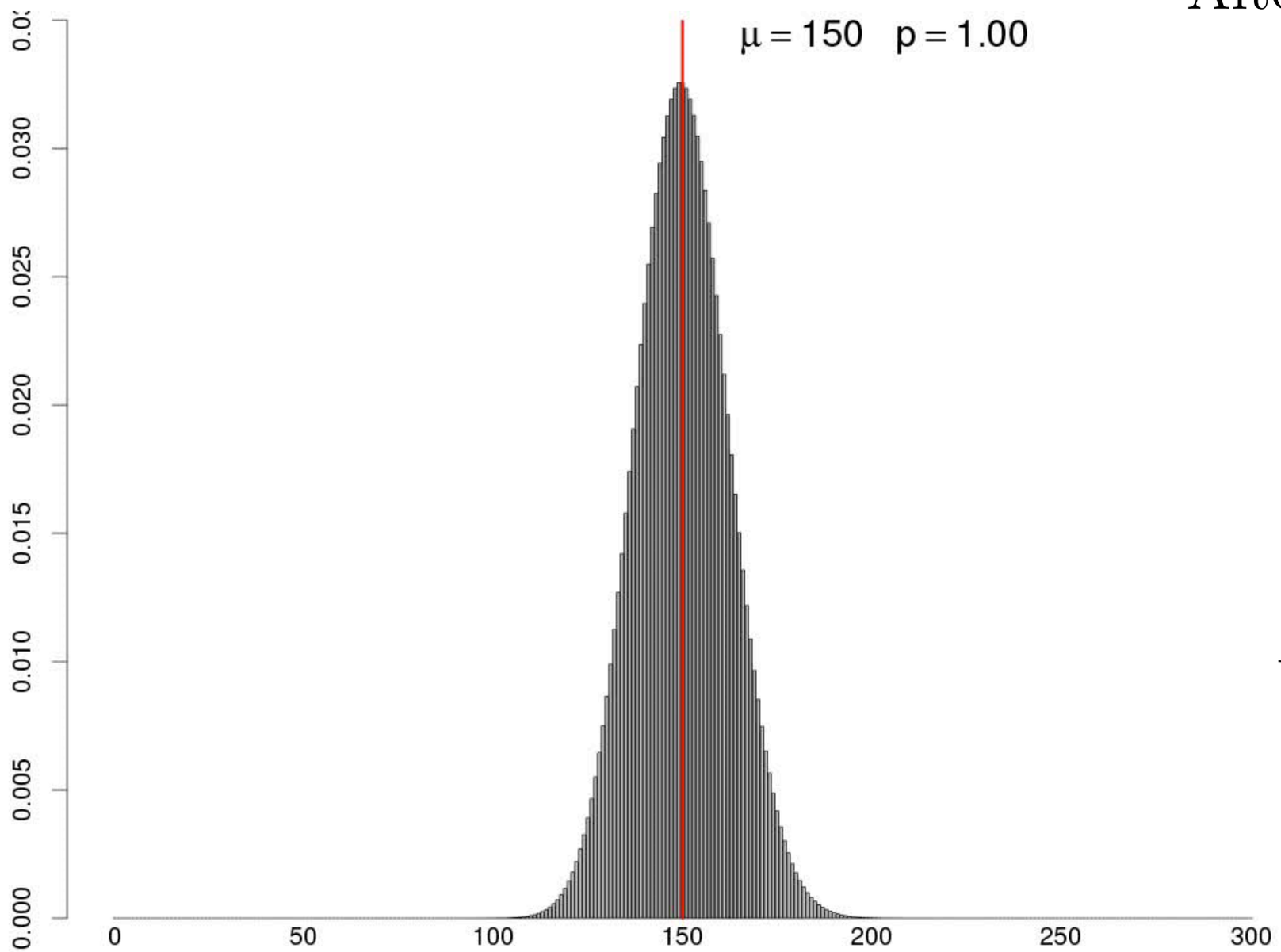
Poisson counting statistics Λ

Overall distribution GP

$$NB(\mu, \sigma^2 + \mu) = \Lambda(\Gamma(\mu, \sigma^2))$$

The Gamma-Poisson (a.k.a. Negative Binomial) distribution

$$\mathrm{P}(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \qquad r \in \mathbb{R}^+,\ p \in [0, 1]$$



Alternative parameterisation

$$\begin{aligned} \alpha &= \frac{1}{r} \\ \mu &= \frac{pr}{1 - p} \end{aligned}$$

Moments

$$\begin{aligned} \text{mean} &= \mu \\ \text{variance} &= \mu + \alpha\mu^2 \end{aligned}$$

Two component noise model

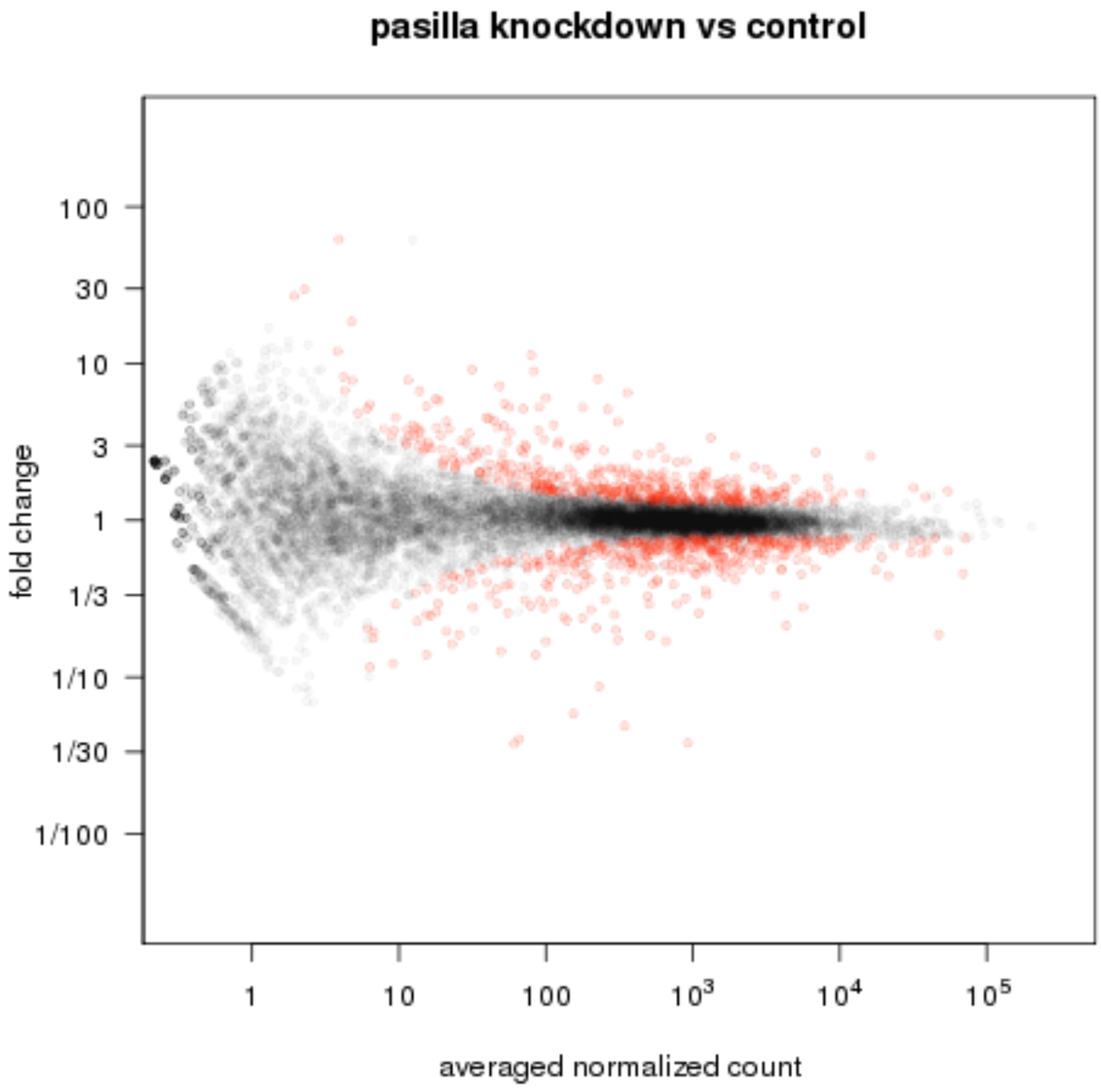
$$\text{var} = \mu + \alpha\,\mu^2$$

shot noise (Poisson) — biological noise

Small counts

Sampling noise dominant

Improve power: deeper coverage



Large counts

Biological noise dominant

Improve power: more biol. replicates

Gamma-Poisson

- Model:

$$Y_1^{\text{treat}}, \ldots, Y_{n_1}^{\text{treat}} \sim \text{GP}(\mu_{\text{treat}}, \alpha)$$
$$Y_1^{\text{control}}, \ldots, Y_{n_2}^{\text{control}} \sim \text{GP}(\mu_{\text{control}}, \alpha)$$

- In principle can test again with $\frac{\widehat{lfc}}{\widehat{se}}$
- If we want to estimate the standard error, we need to estimate α too.
- Possible by maximum likelihood, if we have many replicates (and individual counts not too small)...
- For small number of replicates?

Challenge 4: Systematic biases

- Each gene was measured on a different library prep & sequencing run.
- Maybe there is a systematic bias on the counts we observed (depth of the libraries). For example, there could be numbers s_1, s_2, s_3, s_4 such that:

$$Y_1^{\text{treat}} \sim \text{Poisson}(s_1 \mu_{\text{treat}})$$

$$Y_2^{\text{treat}} \sim \text{Poisson}(s_2 \mu_{\text{treat}})$$

$$Y_1^{\text{control}} \sim \text{Poisson}(s_3 \mu_{\text{control}})$$

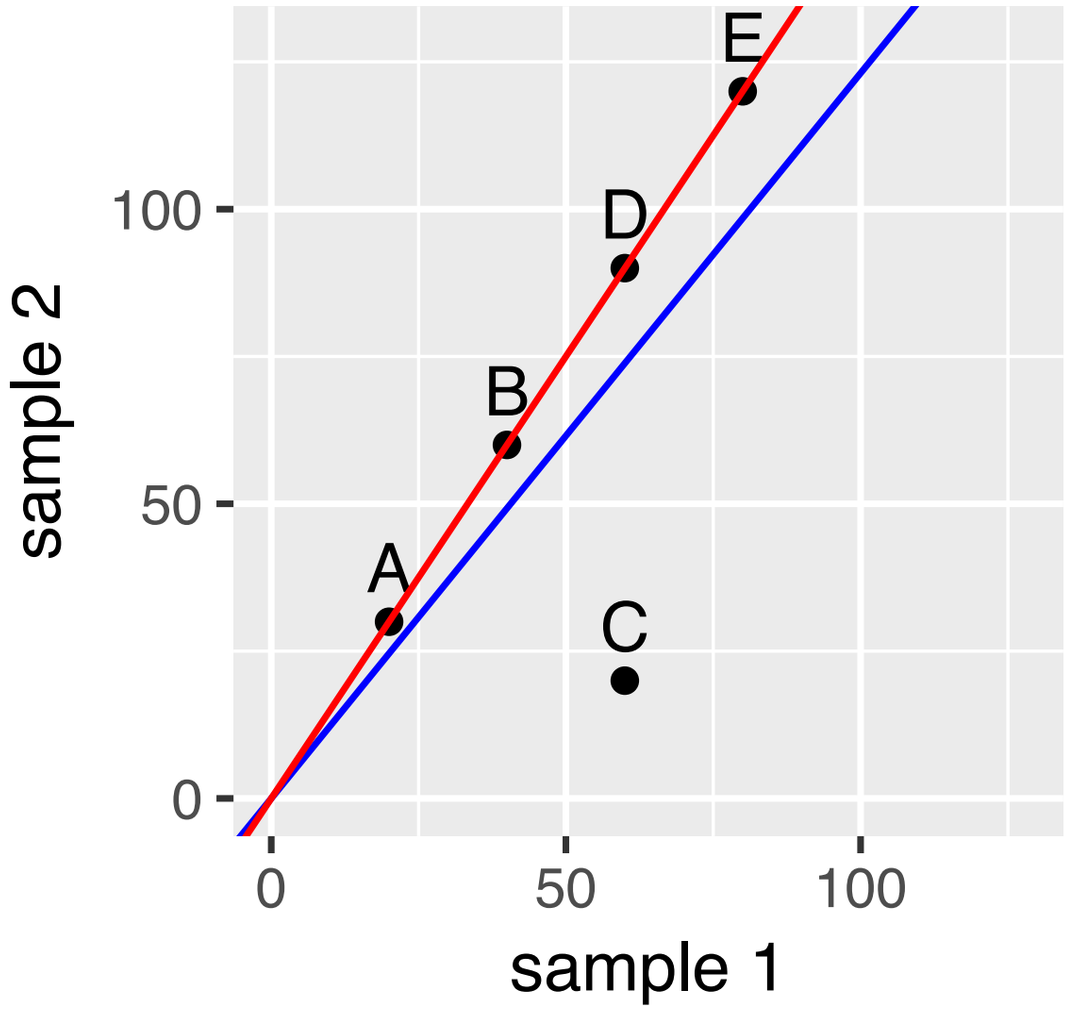
$$Y_2^{\text{control}} \sim \text{Poisson}(s_4 \mu_{\text{control}})$$

- What can we do?

Challenge 4: "Normalization"

- Goal: "To render counts from different samples, which may have been sequenced to different depths, comparable." (DESeq, 2012)
- Assumption: The same scaling factor s_i applies to all genes from sample *i*.
- For example, one could let s_i be equal to the total number of counts on sample *i* across all genes.
- In practice, we want something more robust.

“Normalization”



$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}}.$$

Anders and Huber,
Genome Biology (2010)

Figure 8.1: Size factor estimation. The points correspond to hypothetical genes whose counts in two samples are indicated by their x- and y-coordinates. The lines indicate two different ways of size factor estimation explained in the text.

"Normalization"

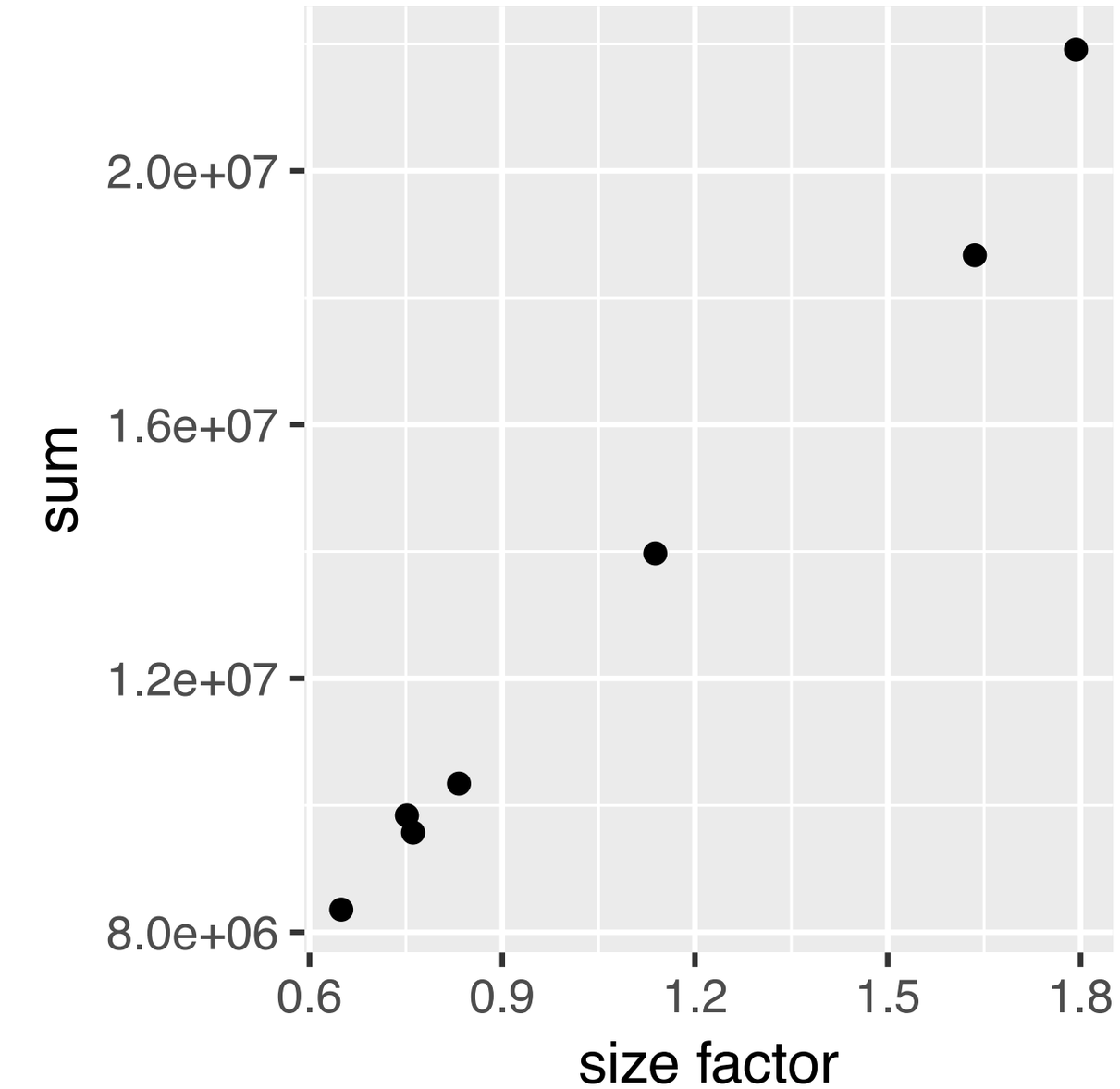


Figure 8.2: Size factors versus sums for the `counts` data.

Recap: Challenges and solutions so far

1. We have count data: model with Poisson distribution.
2. Estimation difficult at low counts: use a Bayesian approach, but how to choose prior?
3. Biological variability: model with Gamma-Poisson mixture distribution, but now we need many biological replicates.
4. Systematic biases: "normalization" (calibration)

Interlude: Opportunity of multiplicity

$$
\begin{array}{c} \\ \text{gene 1} \\ \text{gene 2} \\ \text{gene 3} \\ \vdots \\ \text{gene m} \end{array}
\begin{array}{c}
\begin{array}{cccc} \text{sample 1} & \text{sample 2} & \dots & \text{sample n} \end{array} \\
\begin{pmatrix}
y_{11} & y_{12} & \cdots & y_{1n} \\
y_{21} & y_{22} & \cdots & y_{2n} \\
y_{31} & y_{32} & \cdots & y_{3n} \\
\vdots & \vdots & \ddots & \vdots \\
y_{m1} & y_{m2} & \cdots & y_{mn}
\end{pmatrix}
\end{array}
\longrightarrow
\begin{pmatrix}
\delta_1 \\ \delta_2 \\ \delta_3 \\ \vdots \\ \delta_m
\end{pmatrix}
$$

Benefitting from the many variables ('big data')

To assess signal and noise in the data from one gene, we have

- the data for that gene
- that of all the other genes

⇒ regularisation, (empirical) Bayes

Challenge 2: Finding the prior

- Recall that in our Bayesian lfc estimation we posited:

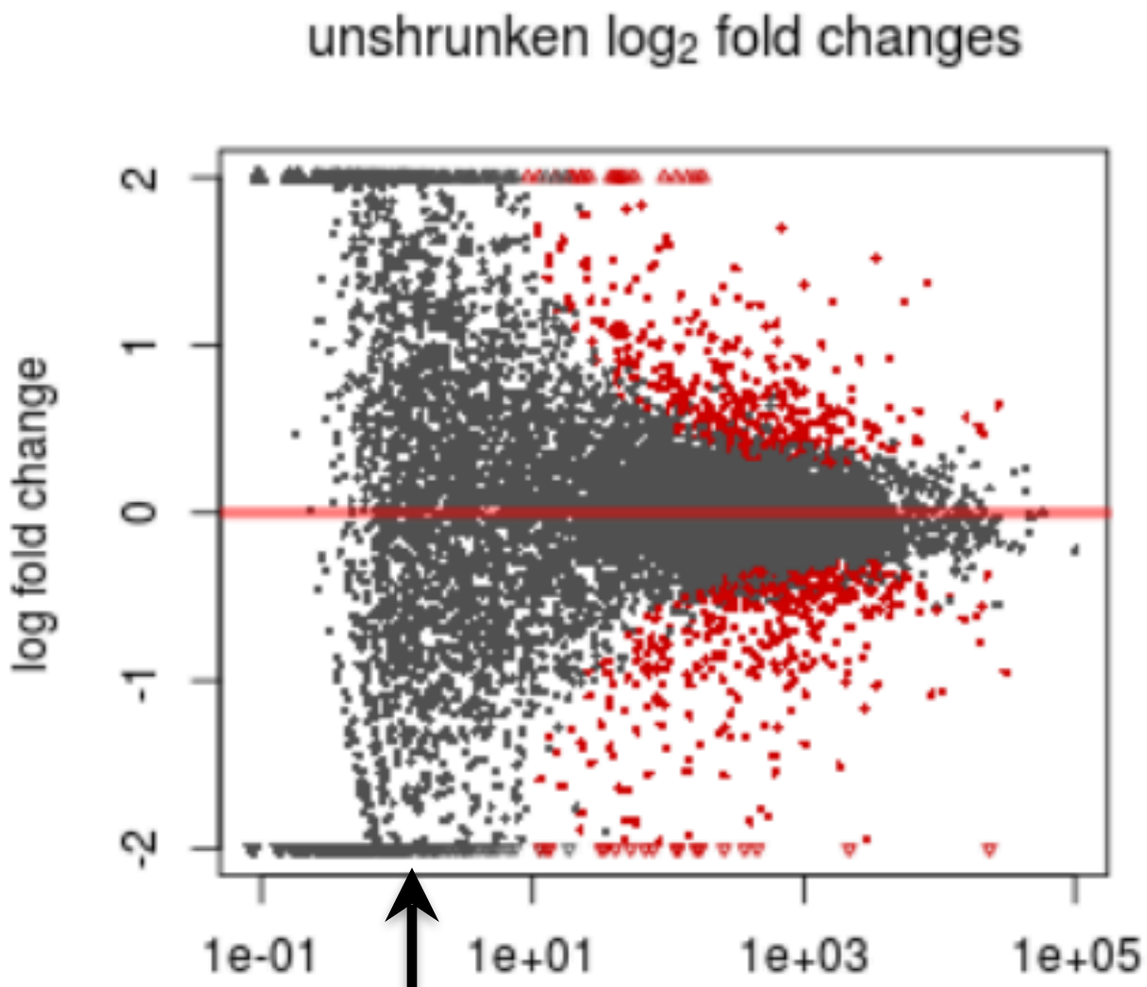
$$\text{lfc} \sim \mathcal{N}\left(0, \sigma^2\right)$$

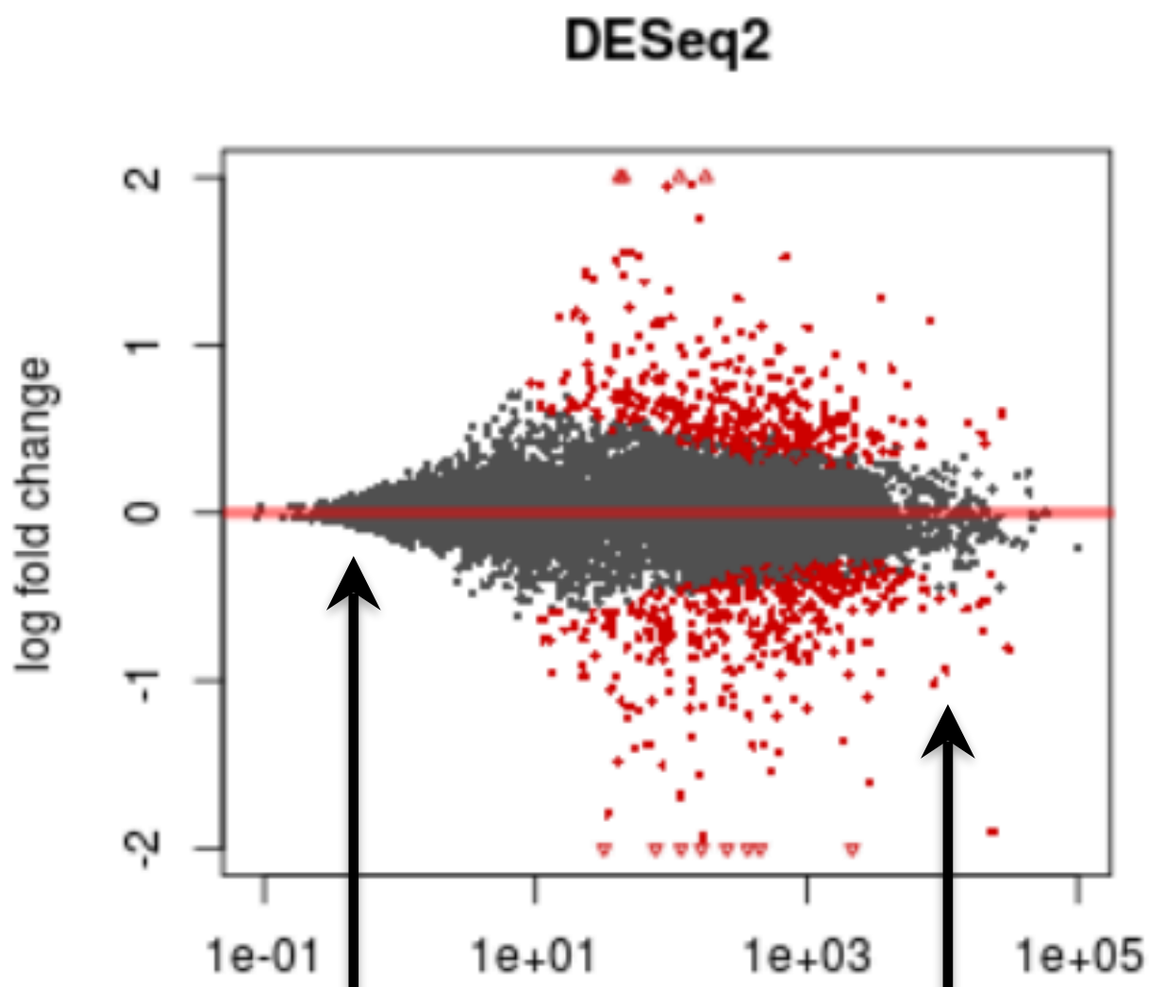
- Now we can posit this for all genes:

$$\text{lfc}_g \sim \mathcal{N}\left(0, \sigma^2\right)$$

- But since we have 20000 genes, we can now estimate σ from the data!
- Intuition: We can plot the distribution of $\widehat{lfc}_g$, if it is wide then σ is larger, otherwise small.

Effect of the empirical Bayes approach: "shrinkage" of the log fold change estimates



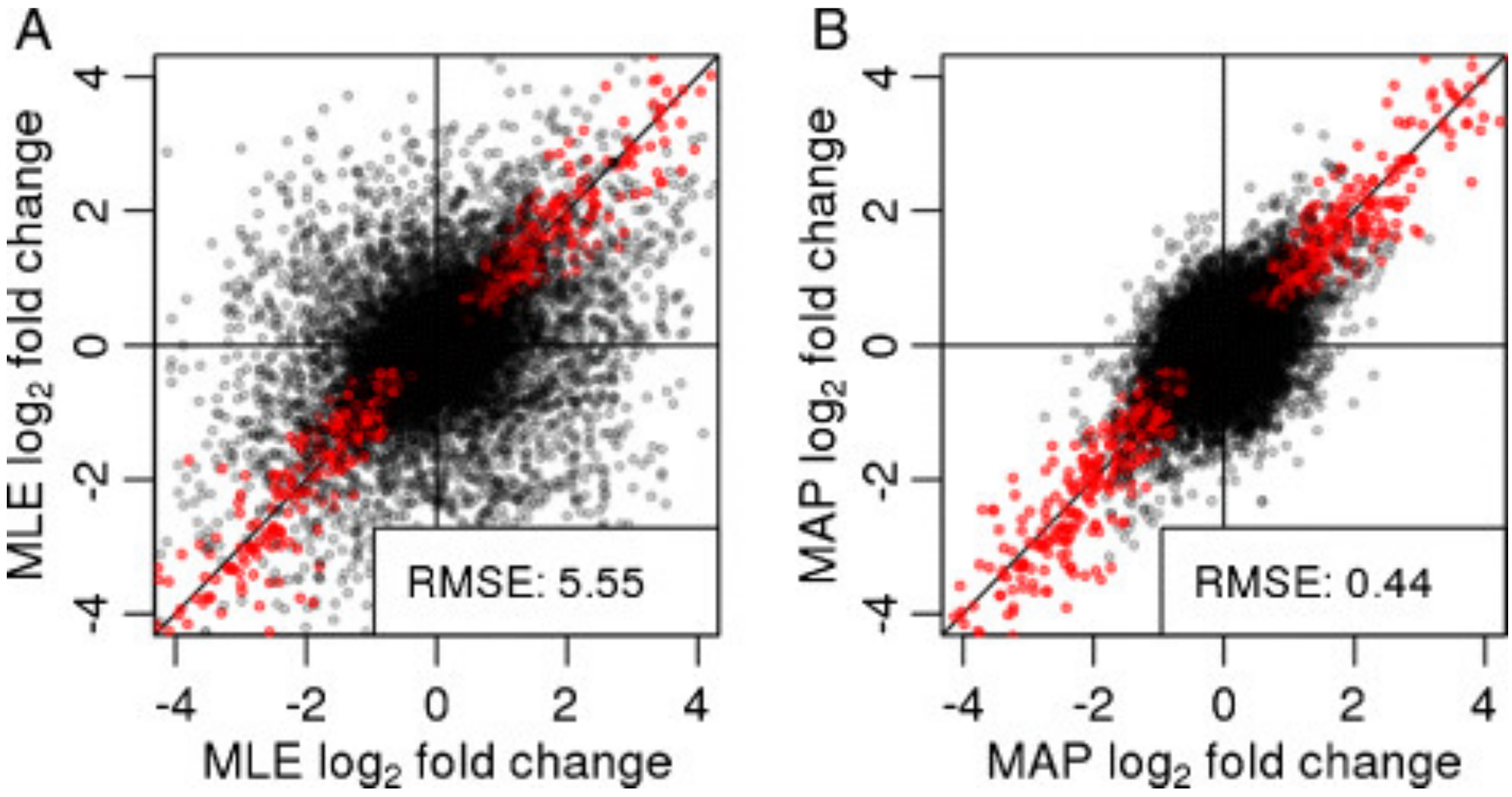


Noisy estimates due to low counts
statistical model will give large p-values, but also the FC estimates themselves are not trustworthy

shrinkage is not equal.
strong moderation for low information genes: low counts

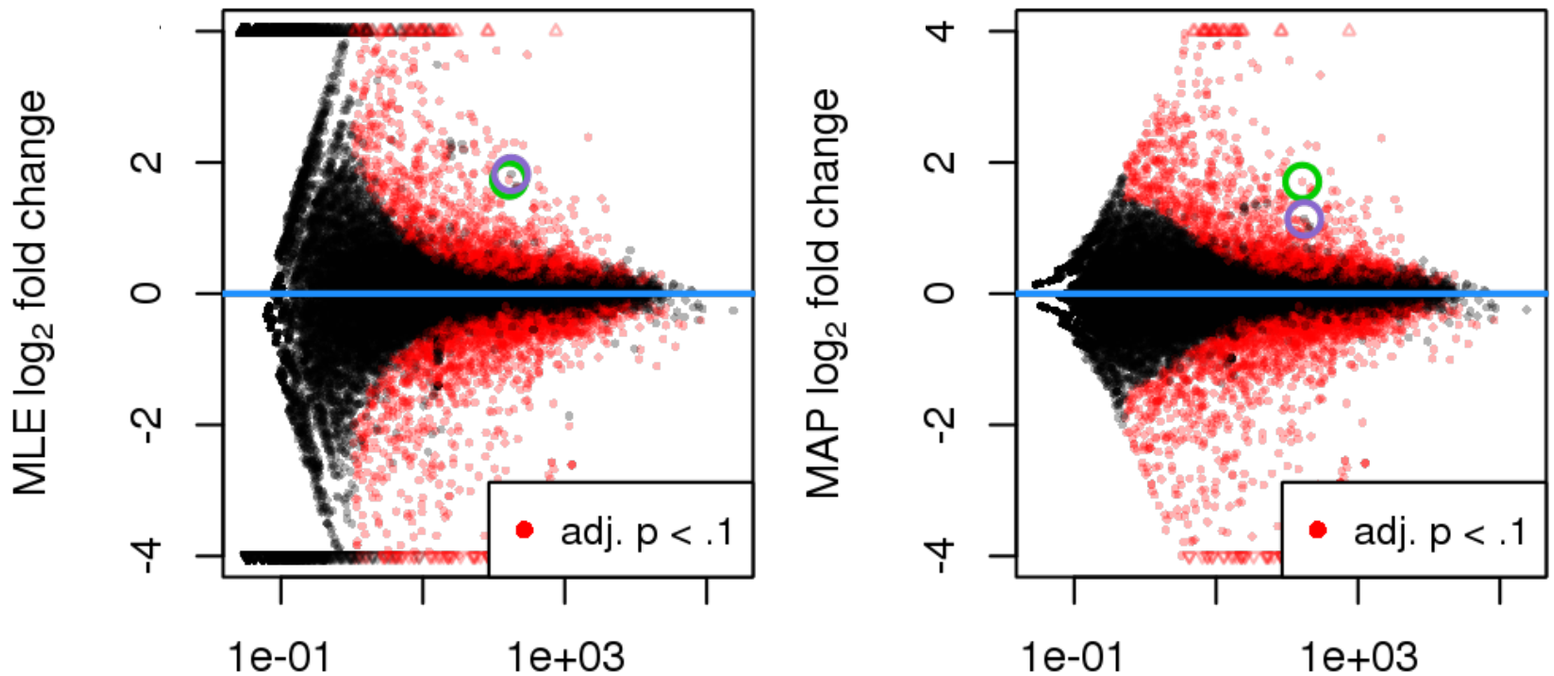
almost no shrinkage

Why shrink fold changes?



Split a dataset into two equal parts, compare LFC

Example: difference between maximum likelihood and maximum a posteriori estimate for two genes



Empirical Bayes in genomics

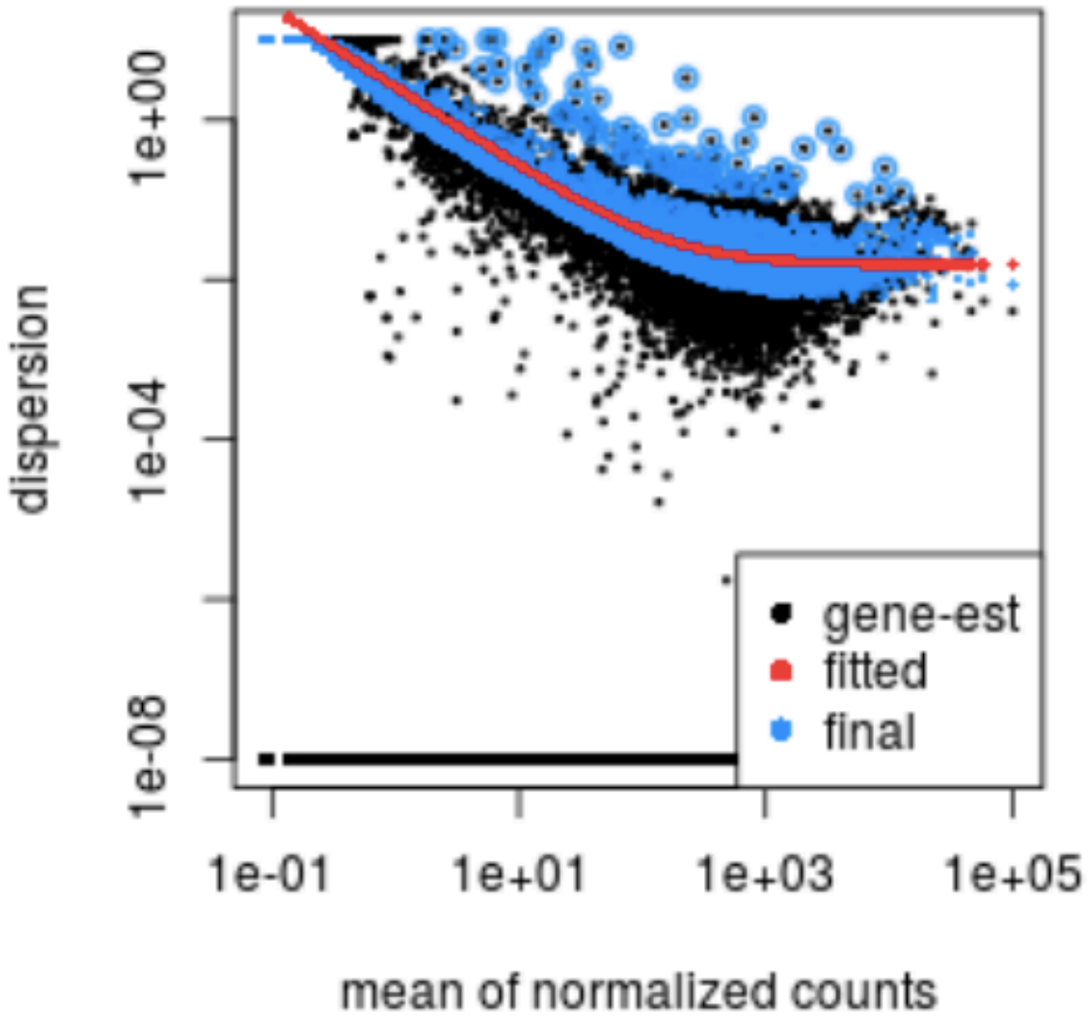
- Lönnstedt and Speed 2002: microarrays
- Smyth 2004: limma for microarrays
- Robinson and Smyth 2007:
  edgeR for SAGE and then applied to RNA-seq
- Many adaptations: DSS and DESeq2 use a similar approach, data-driven strength of shrinkage
- But also outside genomics... (basketball, baseball)

Challenge 3: Estimating the dispersion

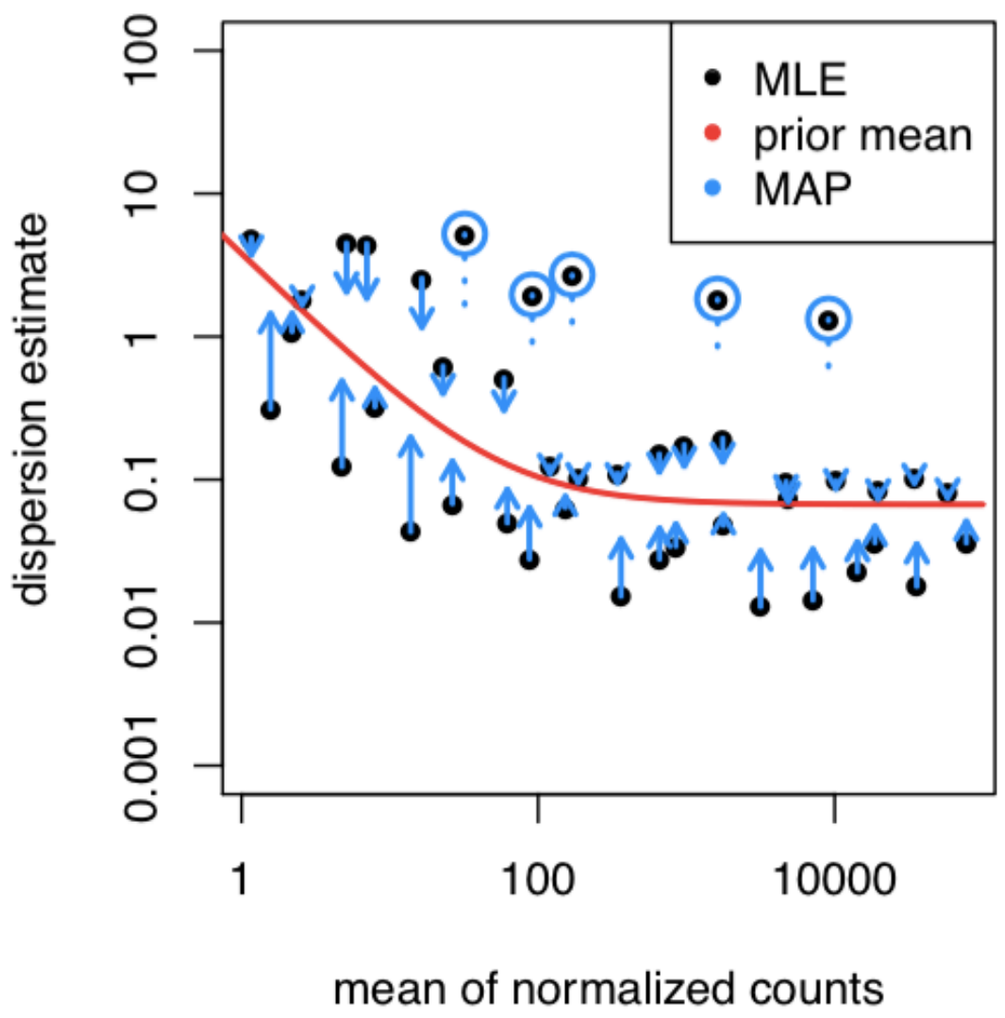
- Recall that the Gamma-Poisson distribution has variance $\mu + \alpha\mu^2$.
- Essentially impossible to estimate based on few replicates.
- Idea: α_g of different genes are related, hence we can use Empirical Bayes to shrink!

Shrinkage of dispersion for RNA-seq

all genes (Pasilla)



a subset of genes (Pickrell)



1. Gene estimate = maximum likelihood estimate (MLE)
2. Fitted dispersion trend = the mean of the prior
3. Final estimate = maximum a posteriori (MAP)

Running DESeq2 (or edgeR, etc.)

```
dds <- DESeqDataSetFromMatrix(countData = cts,
                              colData = coldata,
                              design= ~ batch + condition)
dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
res <- results(dds, name="condition_trt_vs_untrt")
```

- Integrated product that pulls in lots of different concepts and tools from statistics
- Black box vs understanding
- Smart phone analogy

Diagnostic plots after fitting a GP-GLM to RNA-Seq data (e.g., with DESeq2)

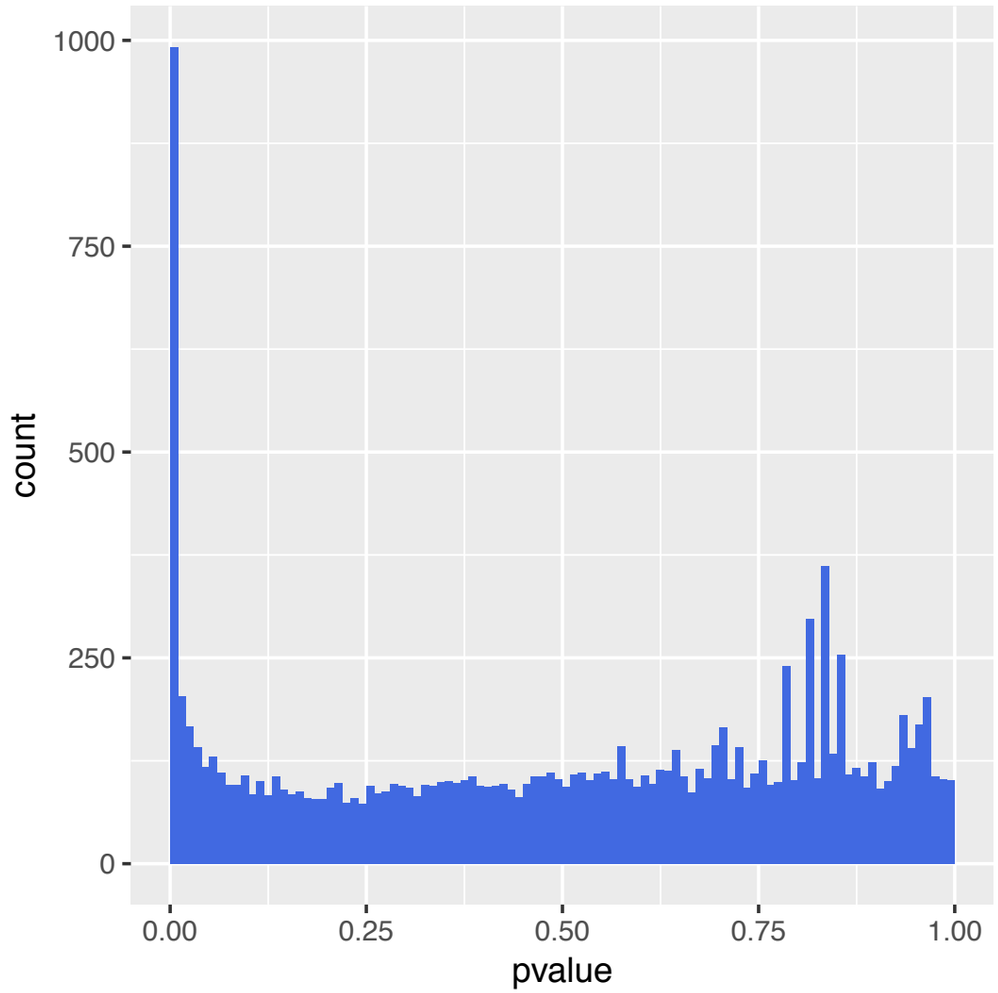


Figure 8.4: Histogram of p-values of a differential expression analysis.

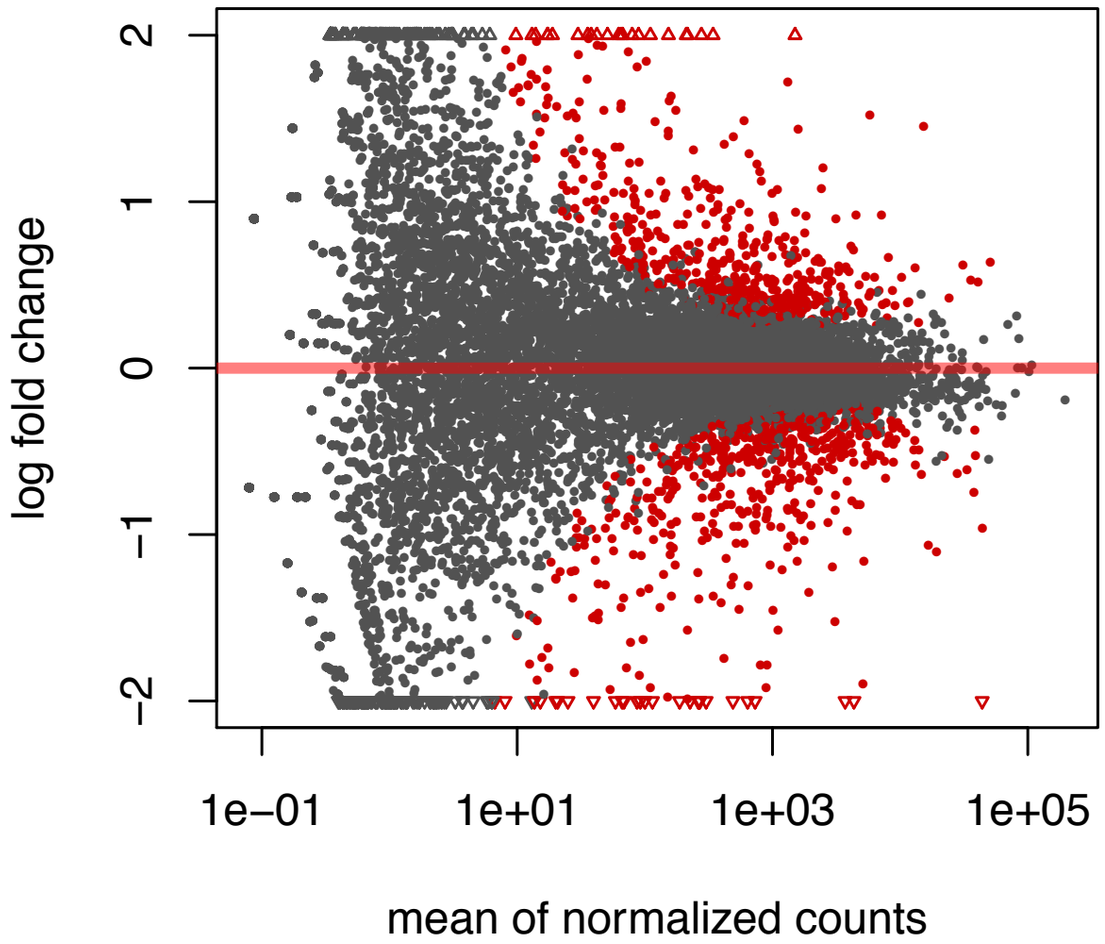


Figure 8.5: MA plot: fold change versus mean of size-factor normalized counts. Logarithmic scaling is used for both axes. By default, points are colored red if the adjusted p-value is less than 0.1. Points which fall out of the y-axis range are plotted as triangles.

Challenge 5: 'downstream' statistical analyses

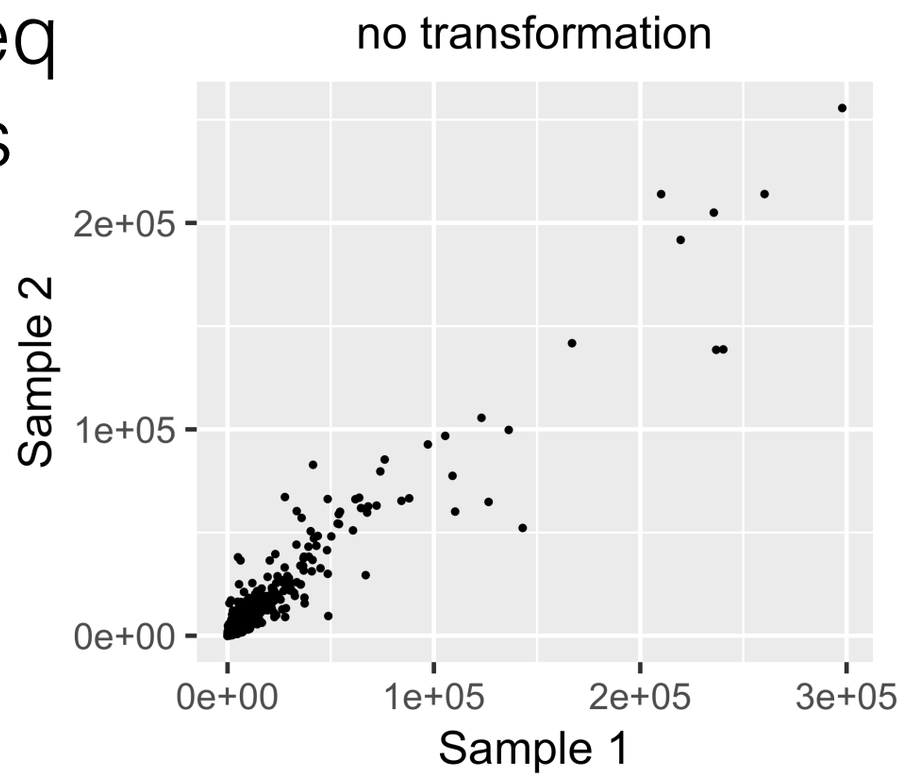
- Now we know how to test for differential expression between two conditions and how to estimate the log-fold changes
- What if we want to do:

    1. Classification
    2. Clustering
    3. PCA
    4. etc.

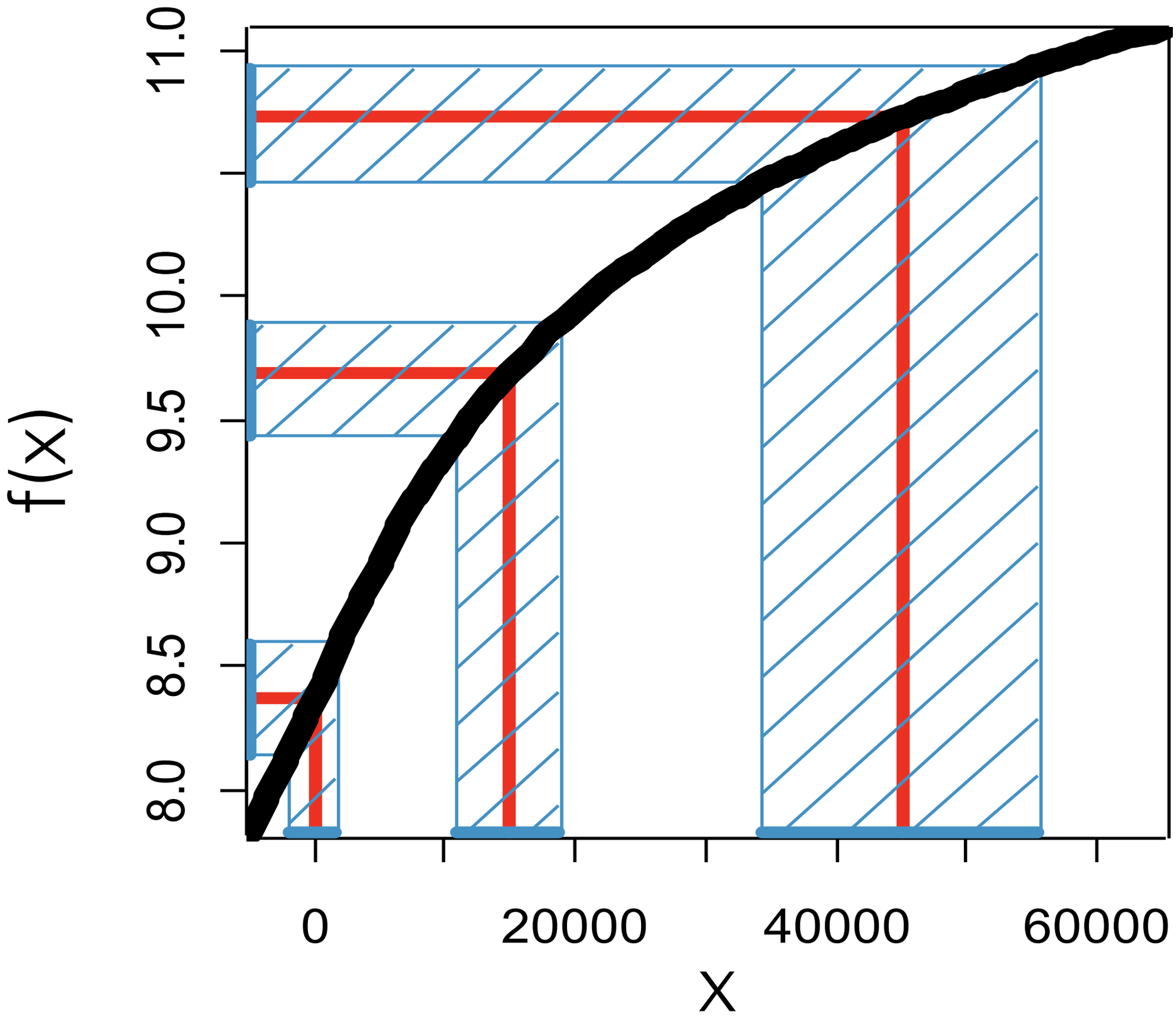
- Often can transform data (e.g.,with variance stabilizing transformation), then apply methods developed for regular (e.g., iid normal) data.

Transformations

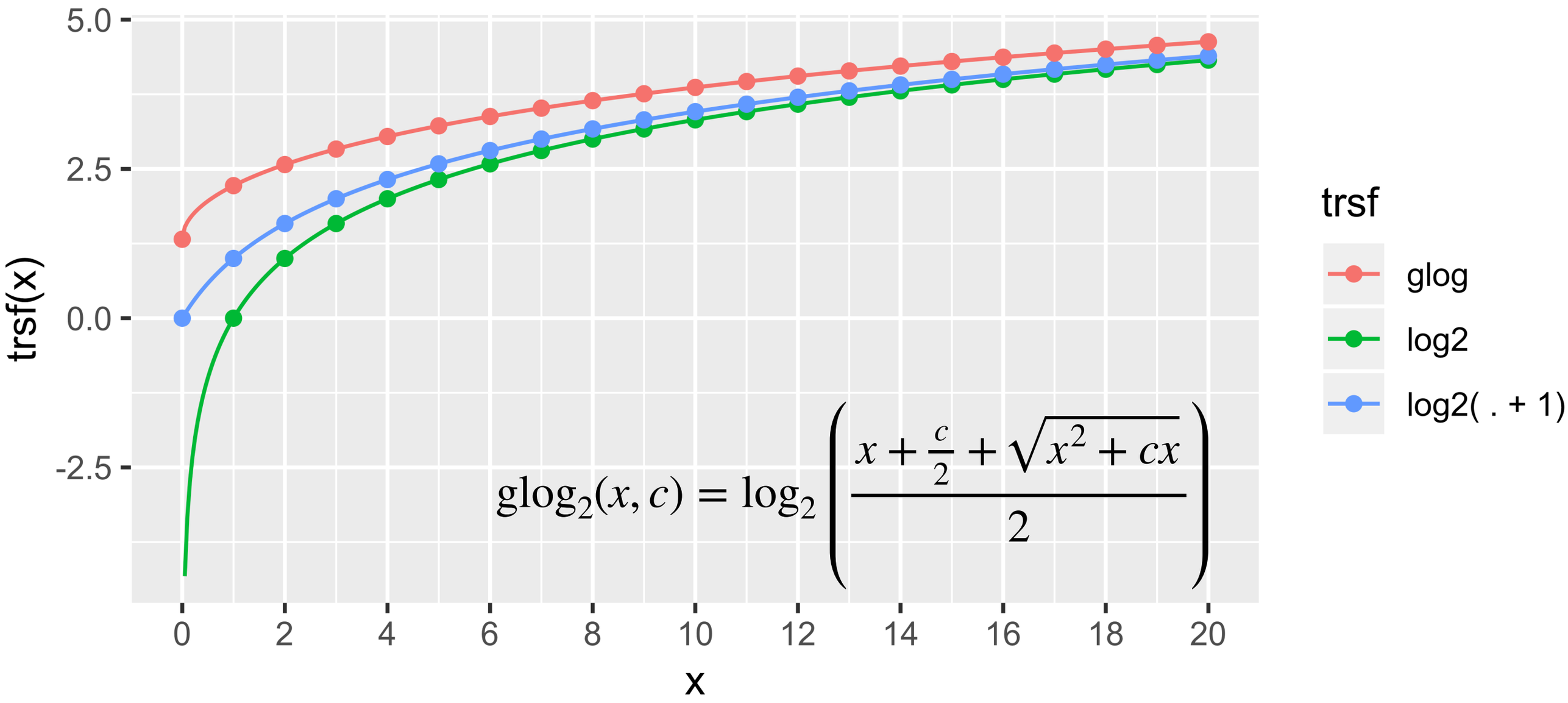
RNA-Seq counts



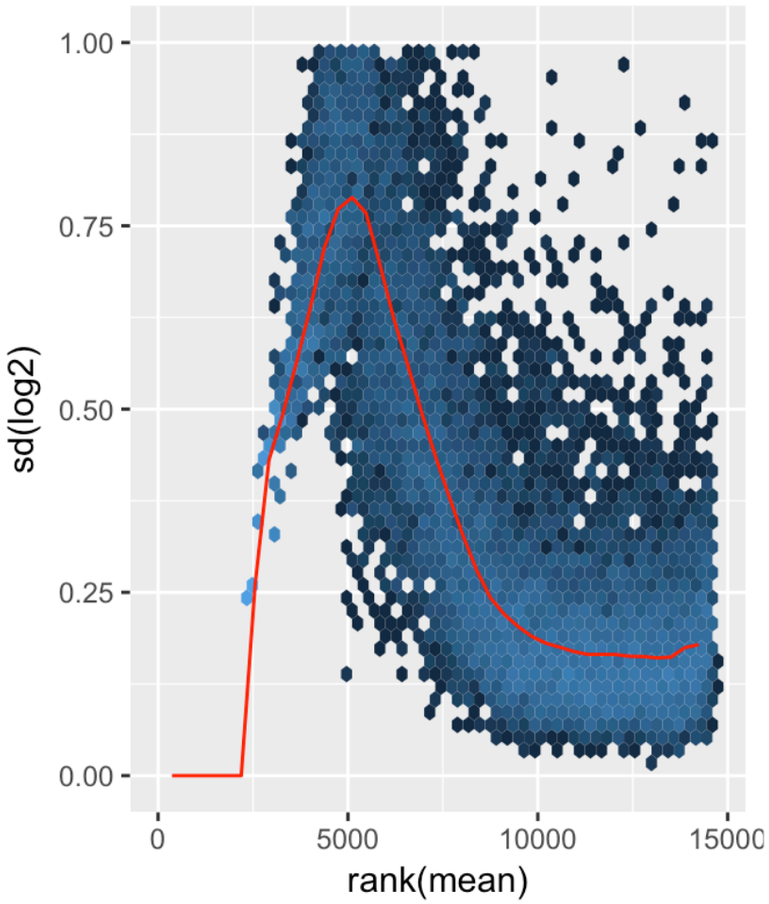
Variance stabilizing transformation

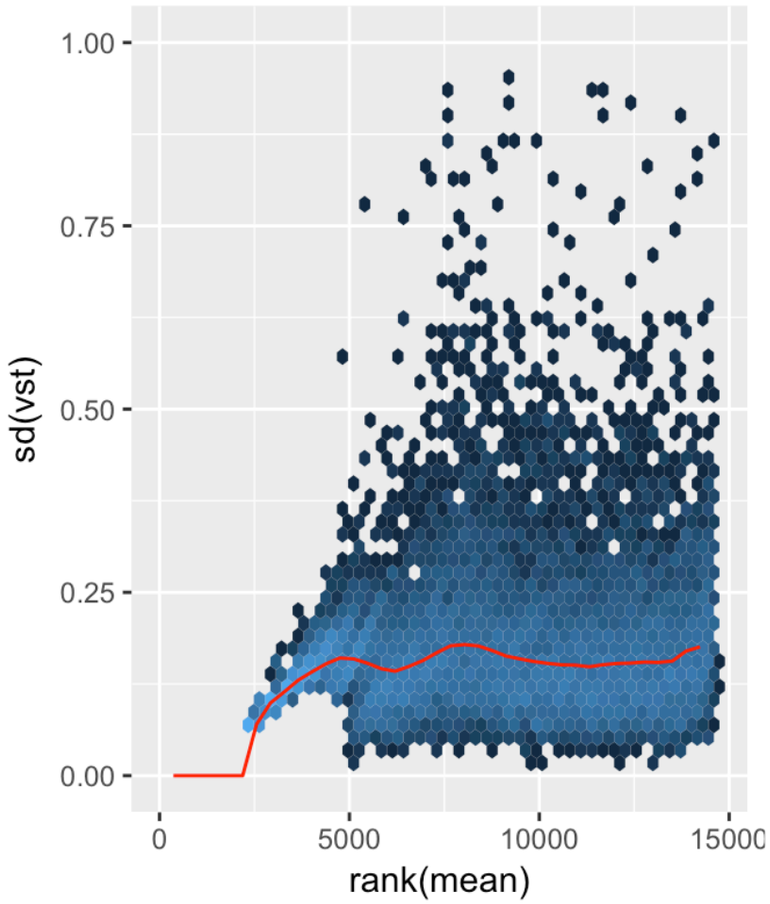


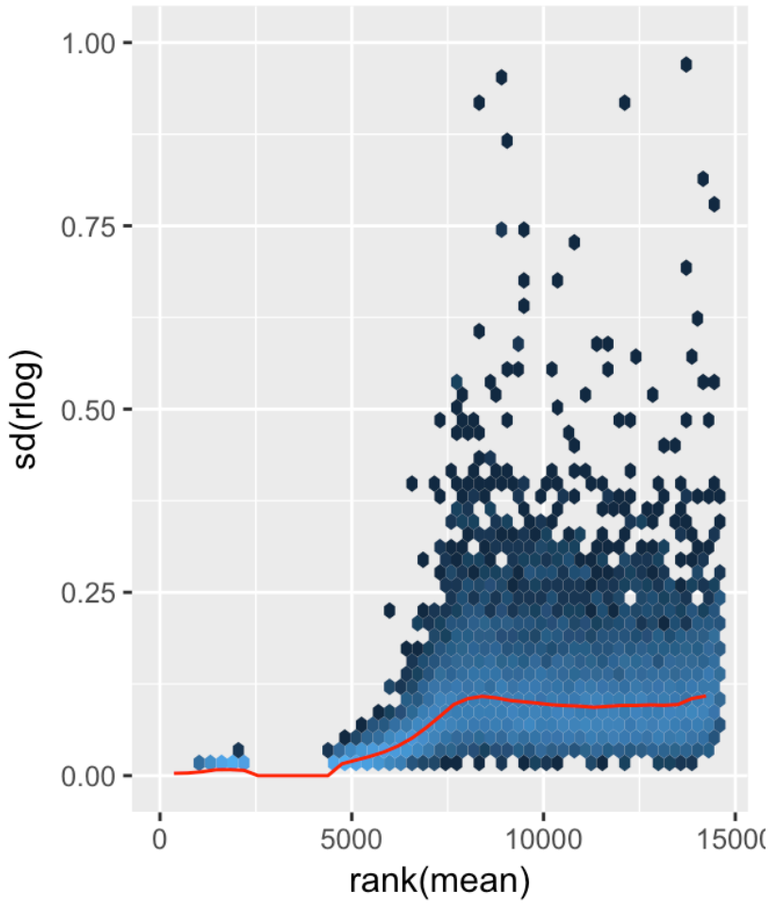
Variance-stabilizing transformation interpolates between $\sqrt{}$ and $\log_2$



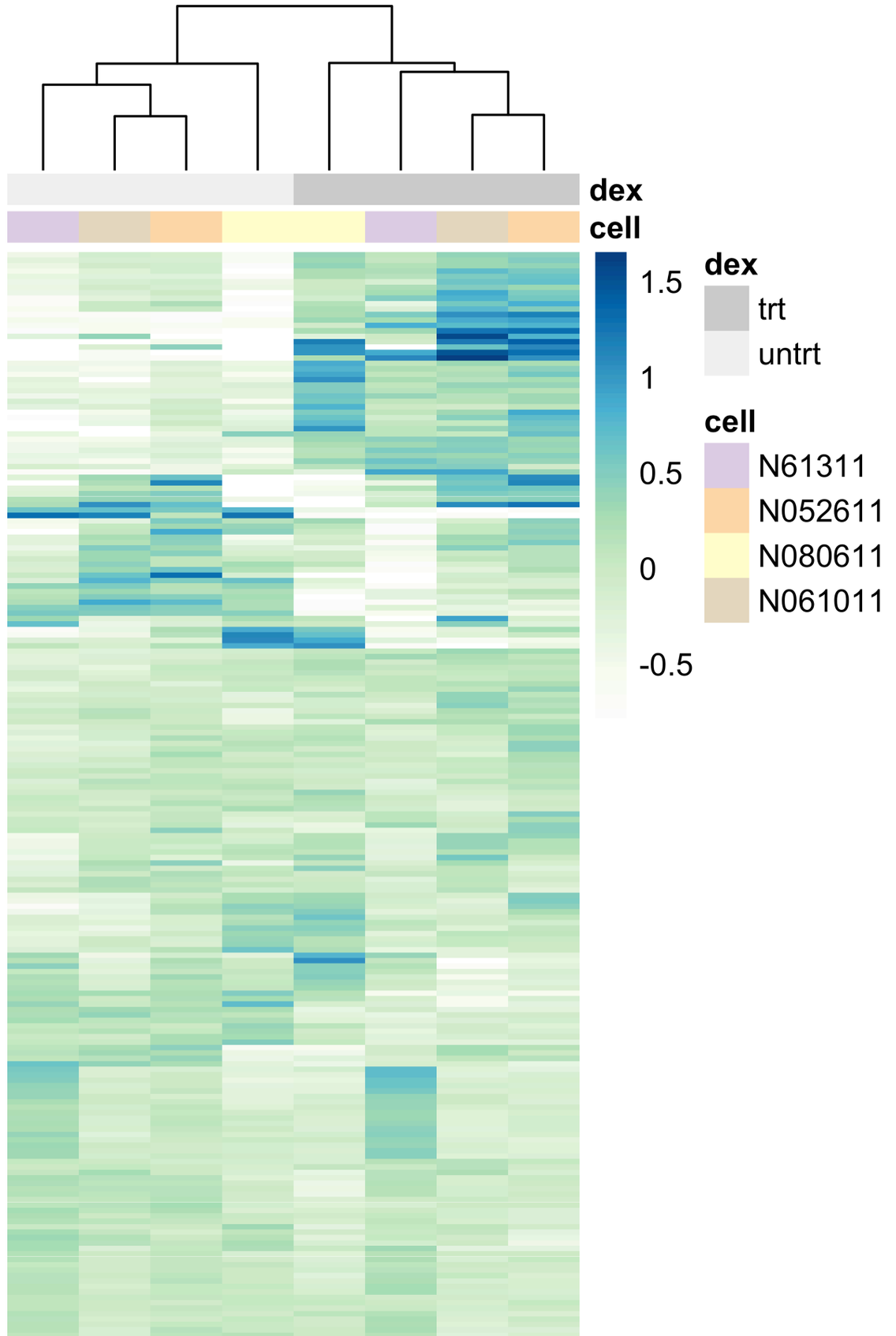
Variance stabilization







Variance-stabilizing transformation for color aesthetic



Recap: Transformations

Choosing the right transformation for your data is crucial.

The scale at which your data are recorded is not necessarily the one at which they should be visualised, analysed.

There is more than the logarithm.

Often, the variance-mean relationship is a good guide.

PS Such awareness exists in physics (radius vs volume, dezibels, Richter scale, critical fluctuations, Lyapunov exponents)

Challenge 6: More complicated designs

- We already know how to test for differential expression between two conditions and how to estimate the log-fold changes
- But reality is more complicated: factorial designs, batch effects

```
annotationFile = system.file("extdata",
  "pasilla_sample_annotation.csv",
  package = "pasilla", mustWork = TRUE)
pasillaSampleAnno = readr::read_csv(annotationFile)
pasillaSampleAnno
```

```
## # A tibble: 7 x 6
##           file condition       type `number of lanes`
##          <chr>     <chr>      <chr>             <int>
## 1   treated1fb   treated single-read                 5
## 2   treated2fb   treated  paired-end                 2
## 3   treated3fb   treated  paired-end                 2
## 4 untreated1fb untreated single-read                 2
## 5 untreated2fb untreated single-read                 6
## 6 untreated3fb untreated  paired-end                 2
## 7 untreated4fb untreated  paired-end                 2
## # ... with 2 more variables: `total number of reads` <chr>,
## #   `exon counts` <int>
```

Design: Example

Imagine we sequenced:

- 5 treated samples out of which 4 paired-end, 1 single-read
- 5 control samples out of which 1 paired-end, 4 single-read

What does it mean if a gene comes up as differentially expressed?

Imagine we have

- a cell line pair: "wild type" and BRD3-KO
- treat both with DMSO or iBET

Design

- Let us write:

$$\begin{aligned}\log_2(\mu_{\text{treat}}) &= \log_2(\mu_{\text{control}}) + \log_2(\mu_{\text{treat}}) - \log_2(\mu_{\text{control}}) \\ &= \log_2(\mu_{\text{control}}) + \log_2\left(\frac{\mu_{\text{treat}}}{\mu_{\text{control}}}\right) \\ &= \beta_0 + \beta_1\end{aligned}$$

- So we can say that for sample i:

$$\log_2(\mu_i) = \begin{cases}\beta_0, & \text{if control} \\ \beta_0 + \beta_1, & \text{if treated}\end{cases}$$

Design

- $$\log_2(\mu_i) = \begin{cases} \beta_0, \text{ if control} \\ \beta_0 + \beta_1, \text{ if treated} \end{cases}$$
- Now we want to include the technology (paired-end vs single-read) in the analysis as well. Let us define the log-fold change between paired-end and single-read:

$$\beta_2 = \log_2 \left(\frac{\mu_{\text{paired-end}}}{\mu_{\text{single-read}}} \right)$$

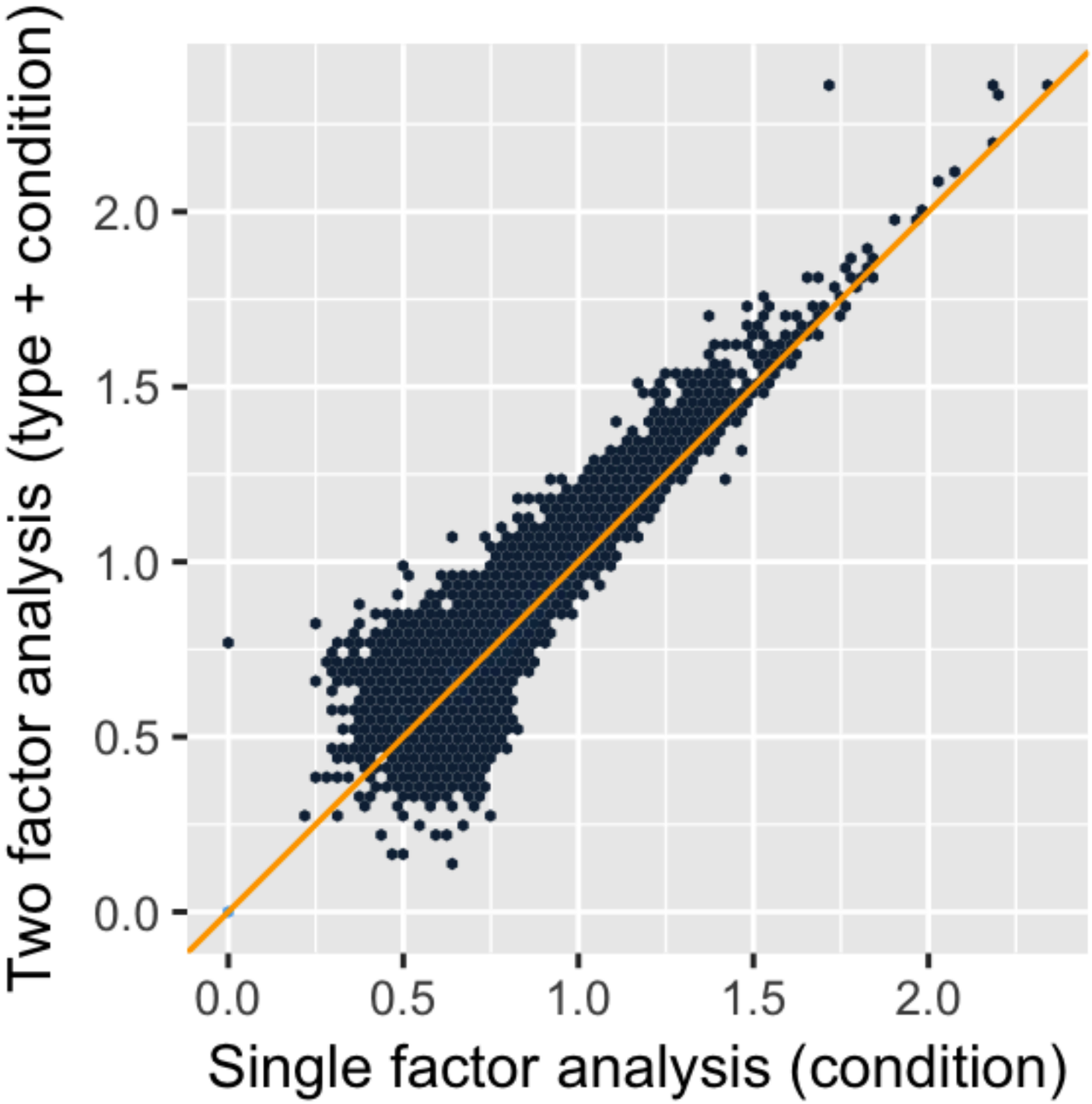
- Then

$$\log_2(\mu_i) = \begin{cases} \beta_0, \text{ if control and single-read} \\ \beta_0 + \beta_1, \text{ if treated and single-read} \\ \beta_0 + \beta_2, \text{ if control and paired-end} \\ \beta_0 + \beta_1 + \beta_2 \text{ if treated and paired-end} \end{cases}$$

Some notes on factorial designs

- We can inform DESeq2 of these designs by using the formula notation: `~ type + condition`
- If we then test for the log-fold change between treated and control, we say that we are *adjusting* or *blocking* or *controlling* for the sequencing technology.
- If every treated sample was sequenced on paired-end and every control sample was sequenced on single-read, then the model is **not identifiable**!

Comparison of the two analyses



On x and y-axis:
Transformation of p-values such that large values indicate small p-values!

How did power increase?

Sometimes specifying the design can improve power.
Common example: Paired designs

patient	treatment
1	before
1	after
2	before
2	after
3	before
3	after
4	before
4	after

Design: Advanced

$$\log_2(\mu_i) = \begin{cases} \beta_0, \text{ if control and single-read} \\ \beta_0 + \beta_1, \text{ if treated and single-read} \\ \beta_0 + \beta_2, \text{ if control and paired-end} \\ \beta_0 + \beta_1 + \beta_2 \text{ if treated and paired-end} \end{cases}$$

- Compact notation: Write $x_{i1} = 1$ if treated and 0 otherwise, and $x_{i2} = 1$ if paired-end and 0 otherwise, then

$$\log_2(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Design: Generalized Linear Models

- Can generalize this even further to:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

- Upshot: "Generalized Linear Models" are well studied, all methods described generalize to this setting.
- Usually expressed in terms of a design matrix

Design matrix for paired designs

patient	treatment
1	before
1	after
2	before
2	after
3	before
3	after
4	before
4	after

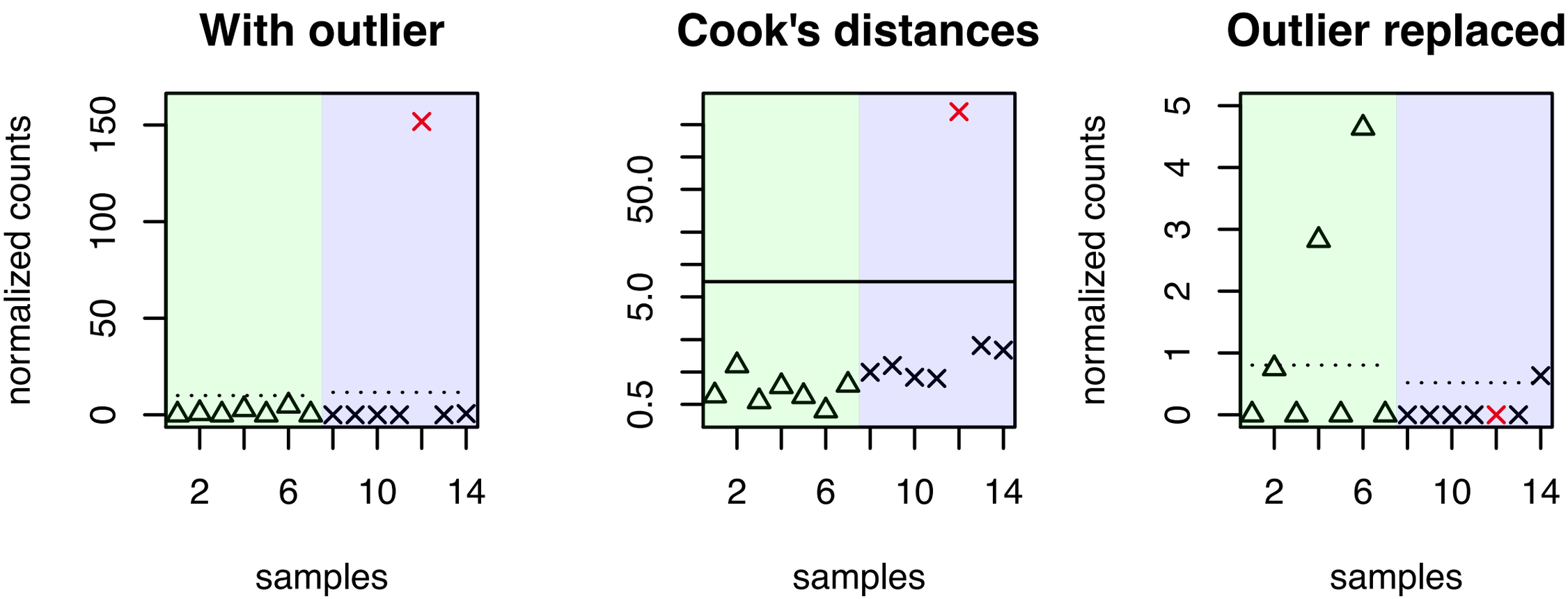
$\Longleftrightarrow$

x_0	x_1	x_2	x_3	x_4
1	0	0	0	0
1	0	0	0	1
0	1	0	0	0
0	1	0	0	1
0	0	1	0	0
0	0	1	0	1
0	0	0	1	0
0	0	0	1	1

Further extensions

- Because of the flexibility of underlying GLMs, we can deal with interactions, continuous covariates, time, etc.
- **`DESeq2`** workflow, **`limma`** vignette, Bioconductor support forum
- There are also methods that try to infer batch-effects/ confounders when we did not actually measure them:
  - RUV-Seq (Remove Unwanted Variation from RNA-Seq Data)
  - SVA (Surrogate Variable Analysis)

Challenge 7: Outlier robustness



Cook's distance:

Change in fitted coefficients if the sample were removed

Challenge 8: Banded hypothesis testing: integrate testing with fold-change cutoff

- So far our null hypothesis has been:

$$H_0 : \text{lfc} = 0$$

- But what if we do not care about $|\text{lfc}| \leq 1$? We can express this as a new null hypothesis:

$$H_0 : |\text{lfc}| \leq 1$$

Banded hypothesis testing: integrate testing with fold-change cutoff

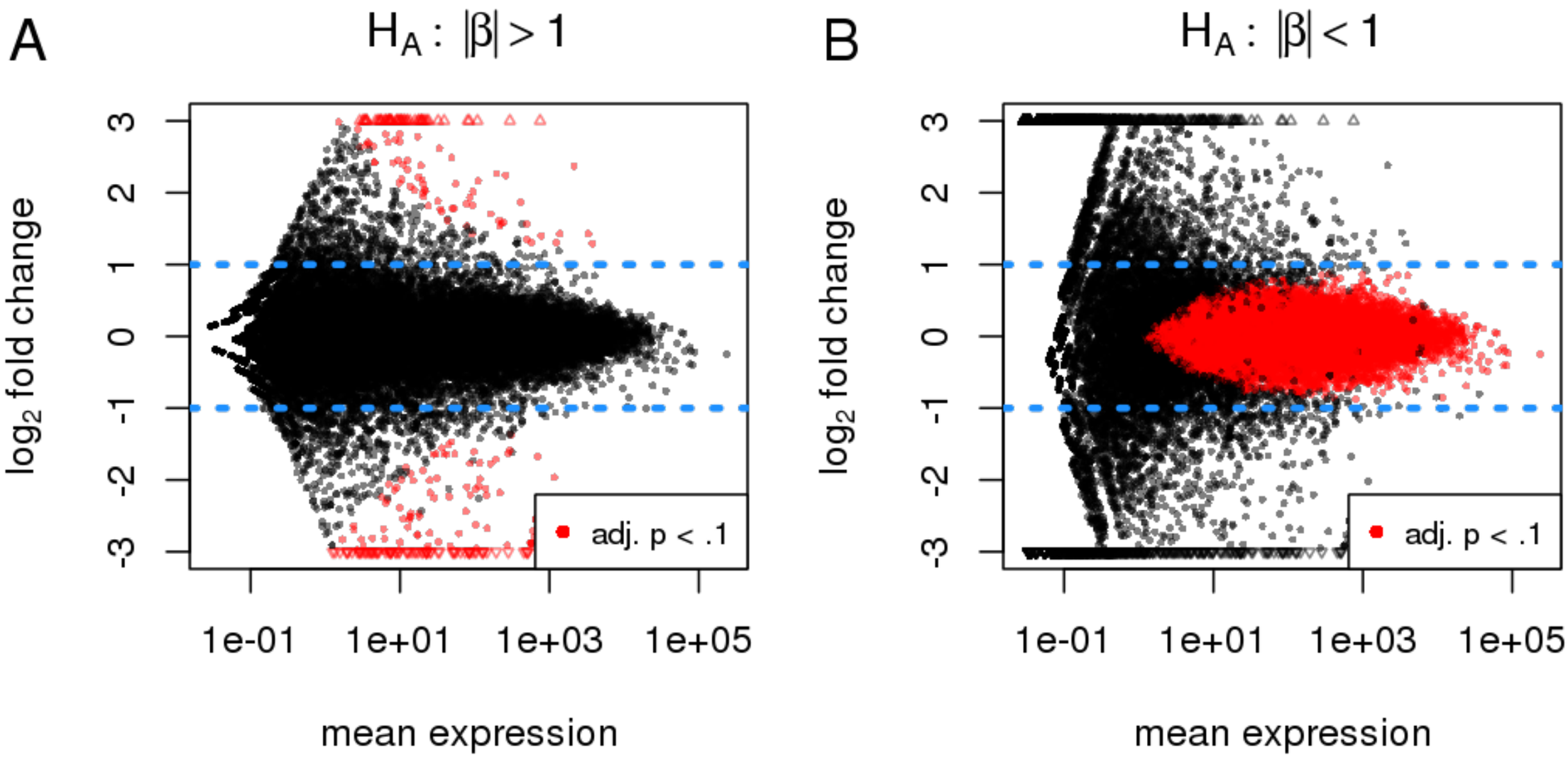


Figure 4 Hypothesis testing involving non-zero thresholds. Shown are MA-plots for a 10 vs 11 comparison using the Bottomly *et al.* [15] dataset, with highlighted points indicating low adjusted p-values. The alternate hypotheses are that logarithmic (base 2) fold changes are (A) greater than 1 in absolute value or (B) less than 1 in absolute value.

Recap: Challenges we addressed

1. Count data - Poisson distribution
2. Low counts - data-driven choice of prior for shrinkage
3. Biological variability - estimating dispersion by sharing of information across genes
4. Sampling bias - normalization
5. Variance stabilizing transformation
6. Dealing with more complicated experimental designs
7. Deal with outliers
8. Banded hypothesis tests

Acknowledgements

Simon Anders

Michael Love