Comparative analysis of RNA-Seq data with DESeq2

Simon Anders

EMBL Heidelberg

Two applications of RNA-Seq

Discovery

- find new transcripts
- find transcript boundaries
- find splice junctions

Comparison

Given samples from different experimental conditions, find effects of the treatment on

- gene expression strengths
- isoform abundance ratios, splice patterns, transcript boundaries

Sequencing count data

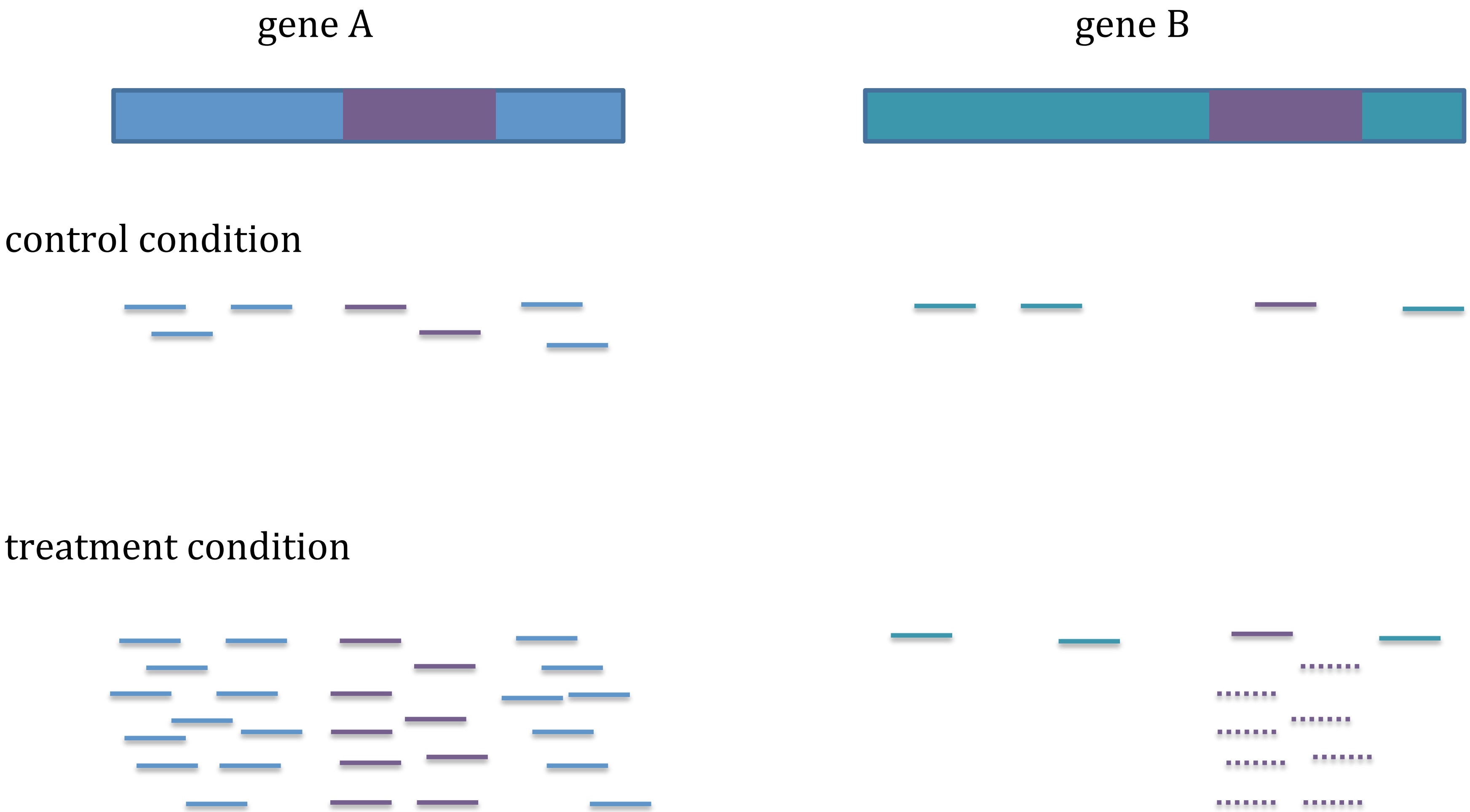
```
             control-1  control-2  control-3    treated-1    treated-2
FBgn0000008         78         46         43           47           89
FBgn0000014          2          0          0            0            0
FBgn0000015          1          0          1            0            1
FBgn0000017       3187       1672       1859         2445         4615
FBgn0000018        369        150        176          288          383
[...]
```

- RNA-Seq
- Tag-Seq
- ChIP-Seq
- HiC
- Bar-Seq
- …

Counting rules

- Count reads, not base-pairs
- Count each read at most once.
- Discard a read if
    - it cannot be uniquely mapped
    - its alignment overlaps with several genes
    - the alignment quality score is bad
    - (for paired-end reads) the mates do not map to the same gene

Why we discard non-unique alignments

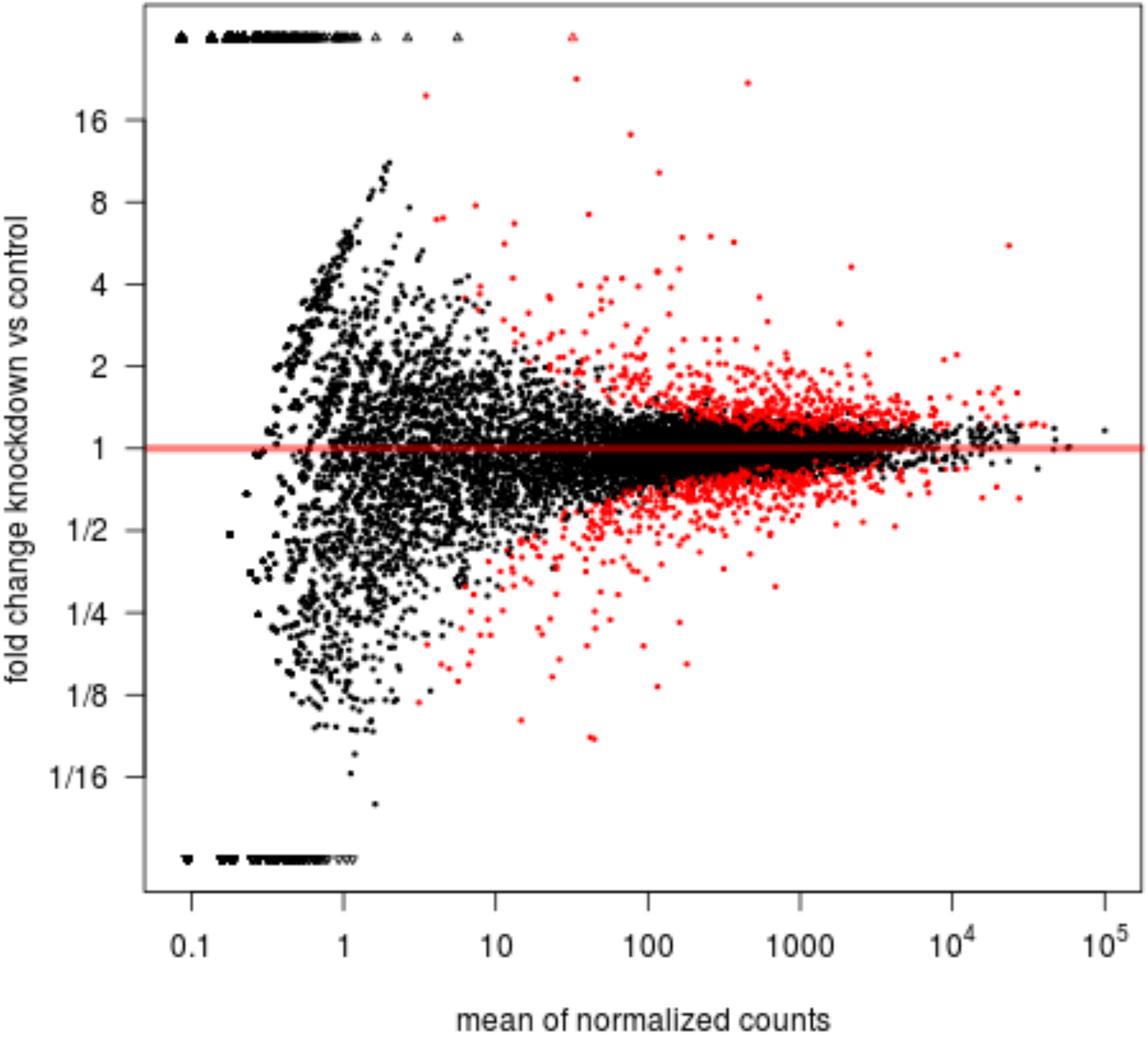


gene A

gene B

control condition

treatment condition


fold change knockdown vs control
16
8
4
2
1
1/2
1/4
1/8
1/16
0.1
1
10
100
1000
10^4
10^5
mean of normalized counts

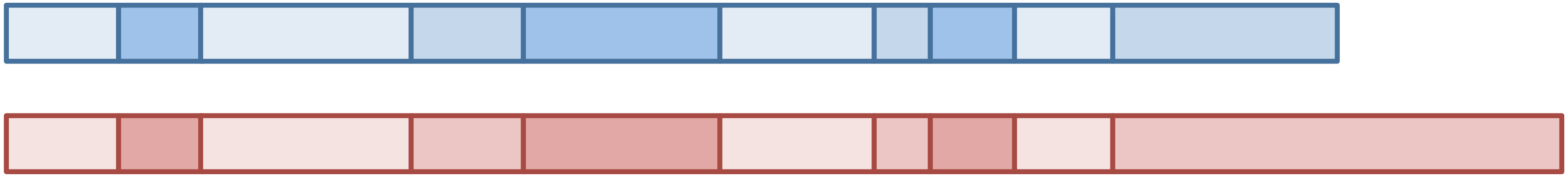
Normalization for library size

- If sample A has been sampled deeper than sample B, we expect counts to be higher.

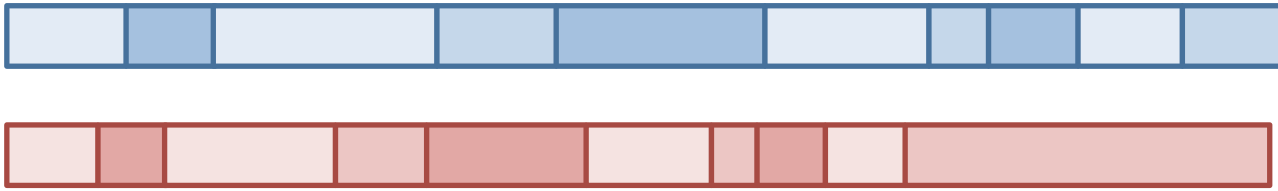
- Naive approach: Divide by the total number of reads per sample
- Problem: Genes that are strongly and differentially expressed may distort the ratio of total reads.

Normalization for library size

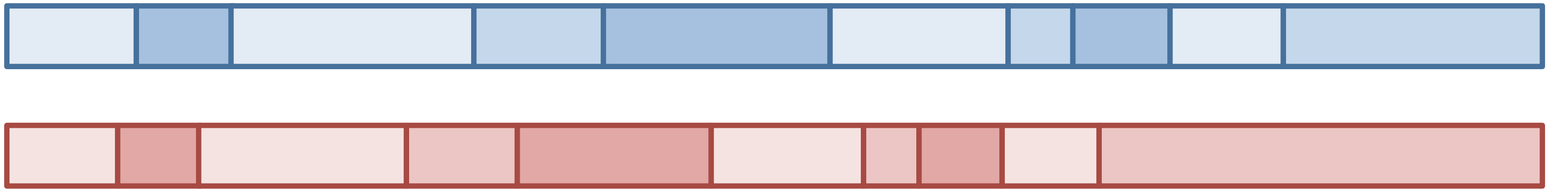
actual expression



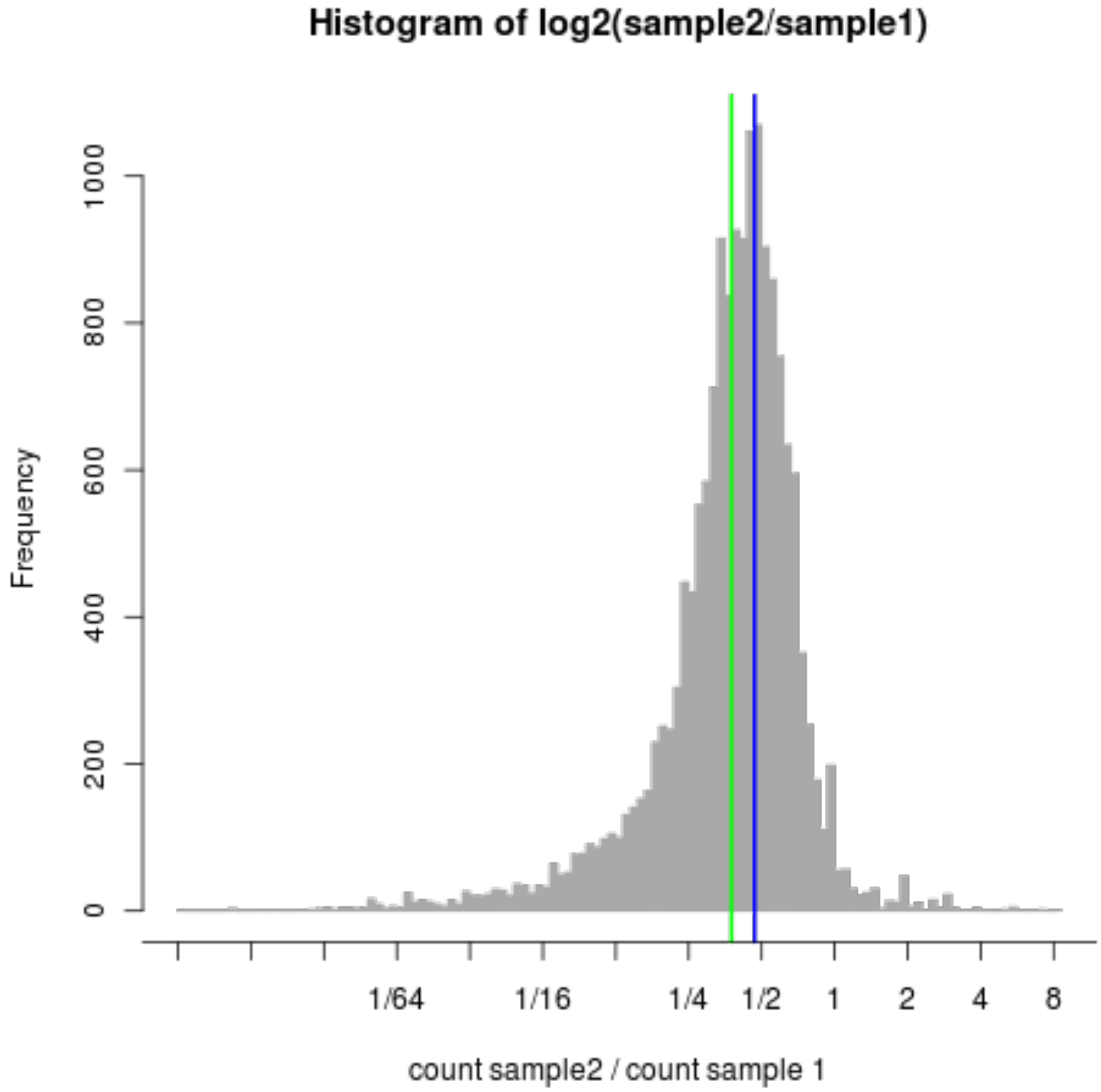
sequenced reads



naivly normalized



Normalization for library size



Normalization for library size

To compare more than two samples:

- Form a “virtual reference sample” by taking, for each gene, the geometric mean of counts over all samples
- Normalize each sample to this reference, to get one scaling factor (“size factor”) per sample.

Anders and Huber, 2010

similar approach: Robinson and Oshlack, 2010

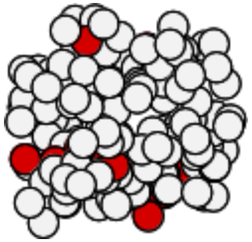
Counting noise

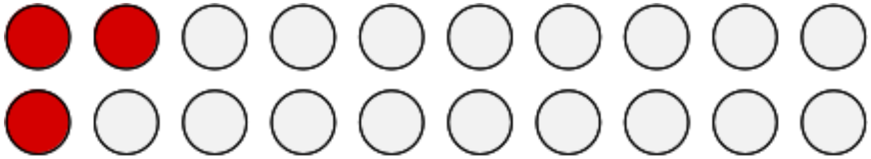
In RNA-Seq, noise (and hence power) depends on count level.

Why?

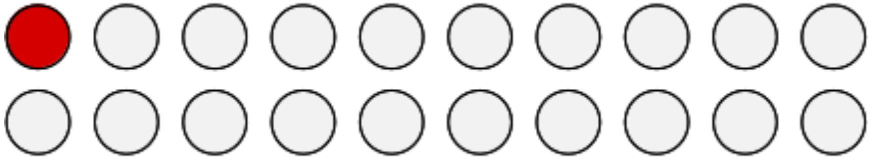
The Poisson distribution

- This bag contains very many small balls, 10% of which are red.
- Several experimenters are tasked with determining the percentage of red balls.
- Each of them is permitted to draw 20 balls out of the bag, without looking.

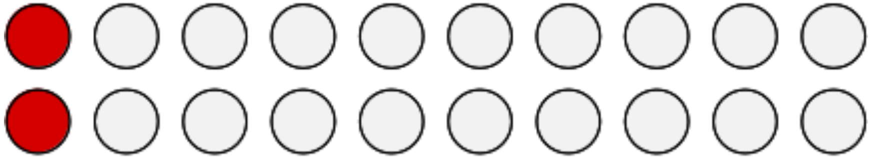




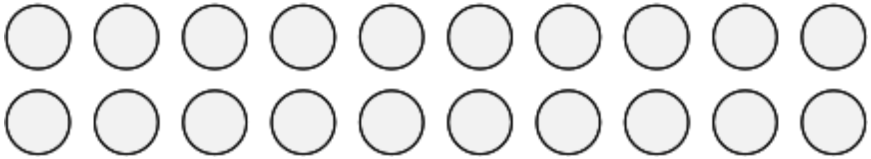
3 / 20 = 15%



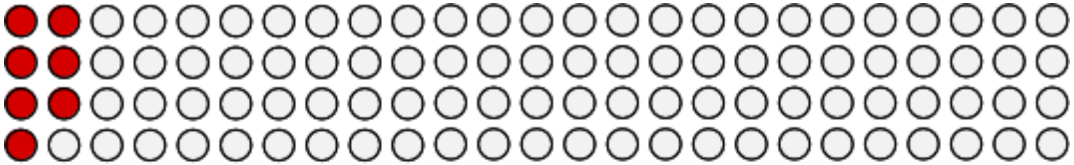
1 / 20 = 5%



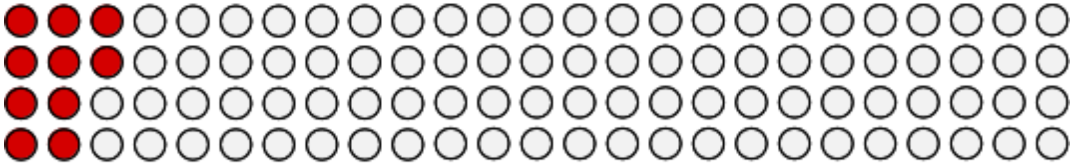
2 / 20 = 10%



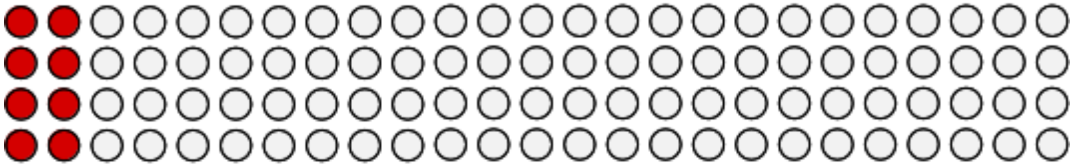
0 / 20 = 0%



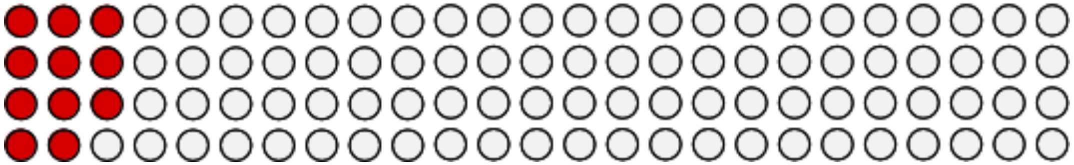
7 / 100 = 7%



10 / 100 = 10%



8 / 100 = 8%



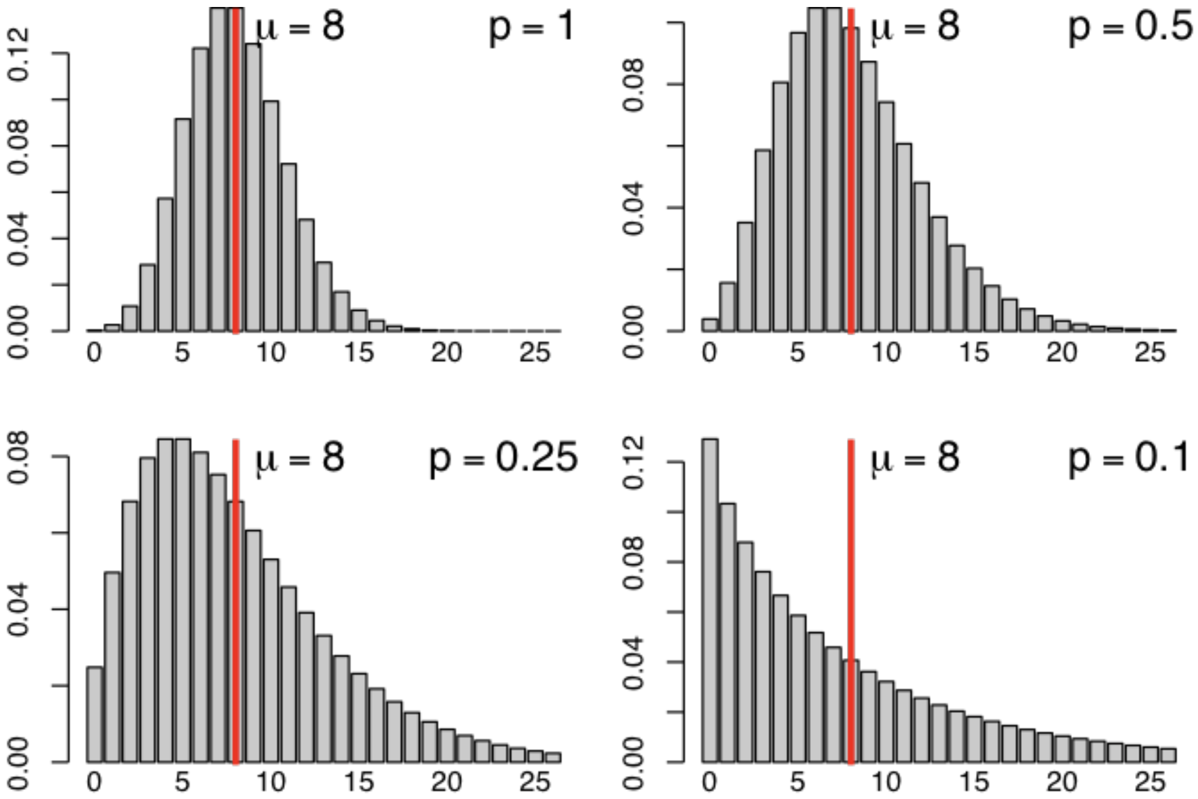
11 / 100 = 11%

Poisson distribution: Counting uncertainty

expected number of red balls	standard deviation of number of red balls	relative error in estimate for the fraction of red balls
10	√10 = 3	1 / √10 = 31.6%
100	√100 = 10	1 / √100 = 10.0%
1,000	√1,000 = 32	1 / √1000 = 3.2%
10,000	√10,000 = 100	1 / √10000 = 1.0%

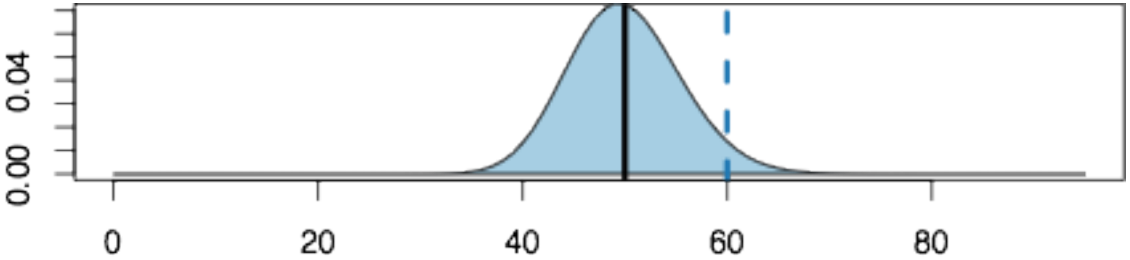
The negative binomial distribution

A commonly used generalization of the Poisson distribution with *two* parameters

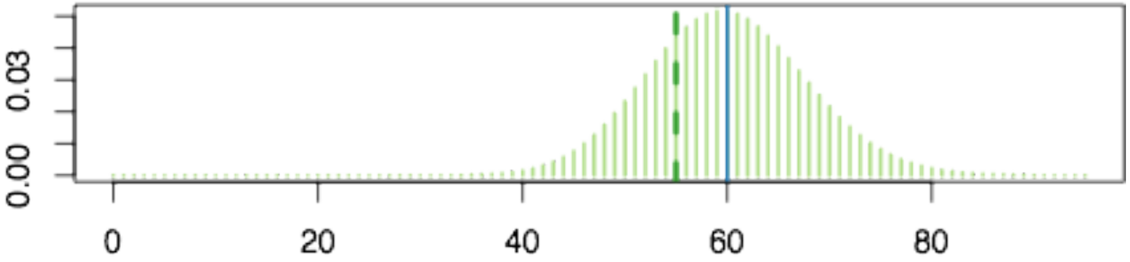


$$\Pr(Y = k) = \binom{k+r-1}{r-1} p^r (1-p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

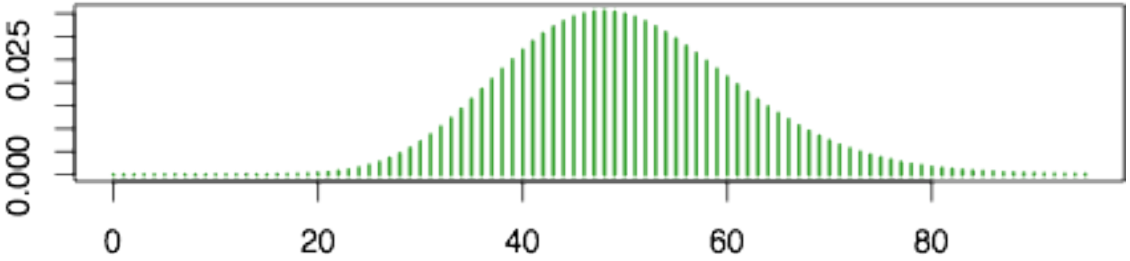
The NB from a hierarchical model



Biological sample with mean μ and variance v



Poisson distribution with mean q and variance q.



Negative binomial with mean μ and variance $q+v$.

Testing: Generalized linear models

Two sample groups: treatment and control.

Model:

- Count value K_{ij} for a gene in sample j is generated by NB distribution with mean $s_j \mu_j$ and dispersion α.

- The expected expression strength is:
  $\log \mu_j = \beta_{i0} + x_j \beta_{i\mathrm{T}}$

  $x_j = 0$ if j is control sample
  $x_j = 1$ if j is treatment sample

Null model:

$\beta_{i\mathrm{T}} = 0$, i.e., expectation is the same for all samples

Alternative model:

$\beta_{i\mathrm{T}} \neq 0$, i.e., expected expression changes from control to treatment, with log fold change (LFC) β_{T}

Testing: Generalized linear models

$$K_{ij} \sim \mathrm{NB}\,(s_j\, \mu_{ij,}\, \alpha_i)$$

$$\log \mu_{ij} = \beta_{i0} + x_j\, \beta_{i\mathrm{T}}$$

$x_j = 0$ for if j is control sample

$x_j = 1$ for if j is treatment sample

Calculate the coefficients β that fit best the observed data K.

Is the value for $\beta_{i\mathrm{T}}$ significantly different from null?

Can we reject the null hypothesis that it is merely cause by noise (as given by the dispersion α_i)?

We use a Wald test to get a p value.

Tasks in comparative RNA-Seq analysis

- Estimate fold-change between control and treatment

- Estimate variability within groups

- Determine significance

the hard part

Dispersion

- Minimum variance of count data:

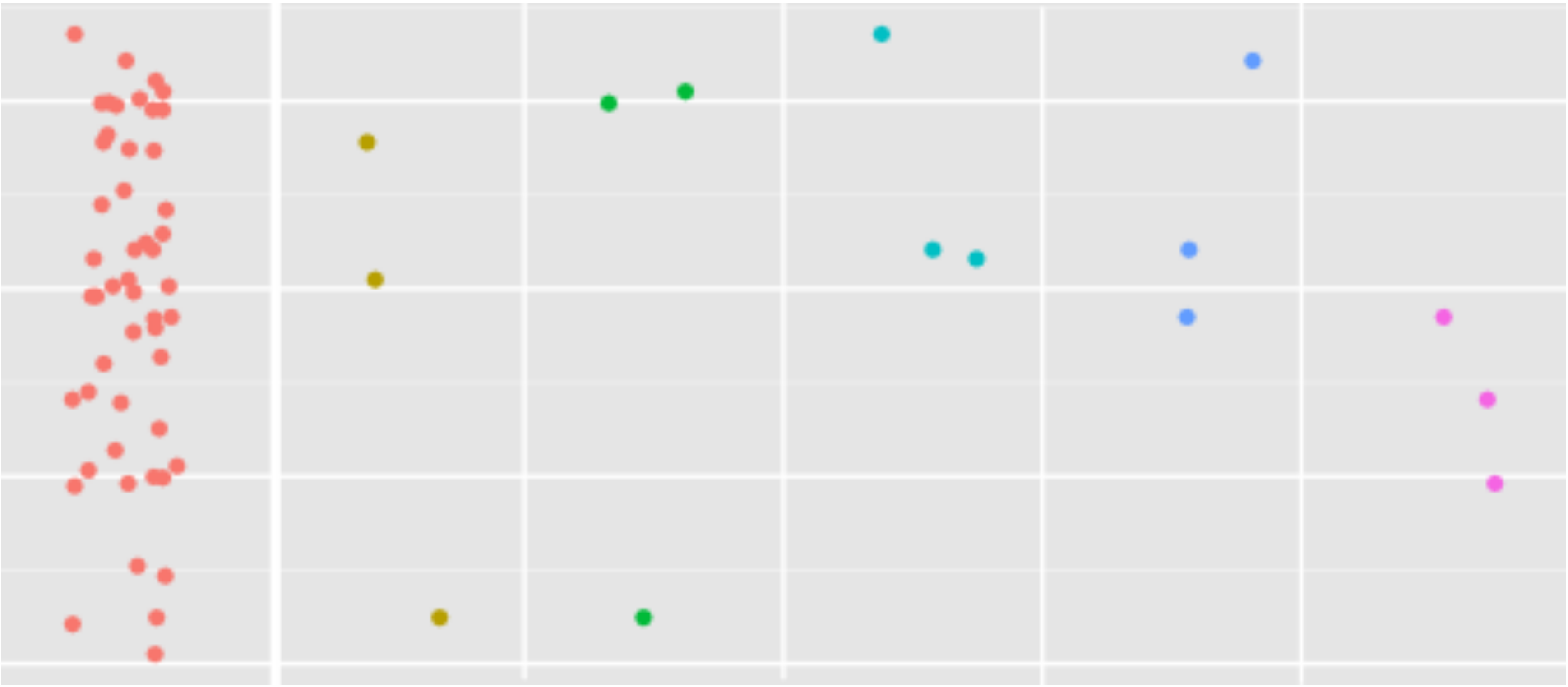
  $v = \mu$ (Poisson)

- Actual variance:

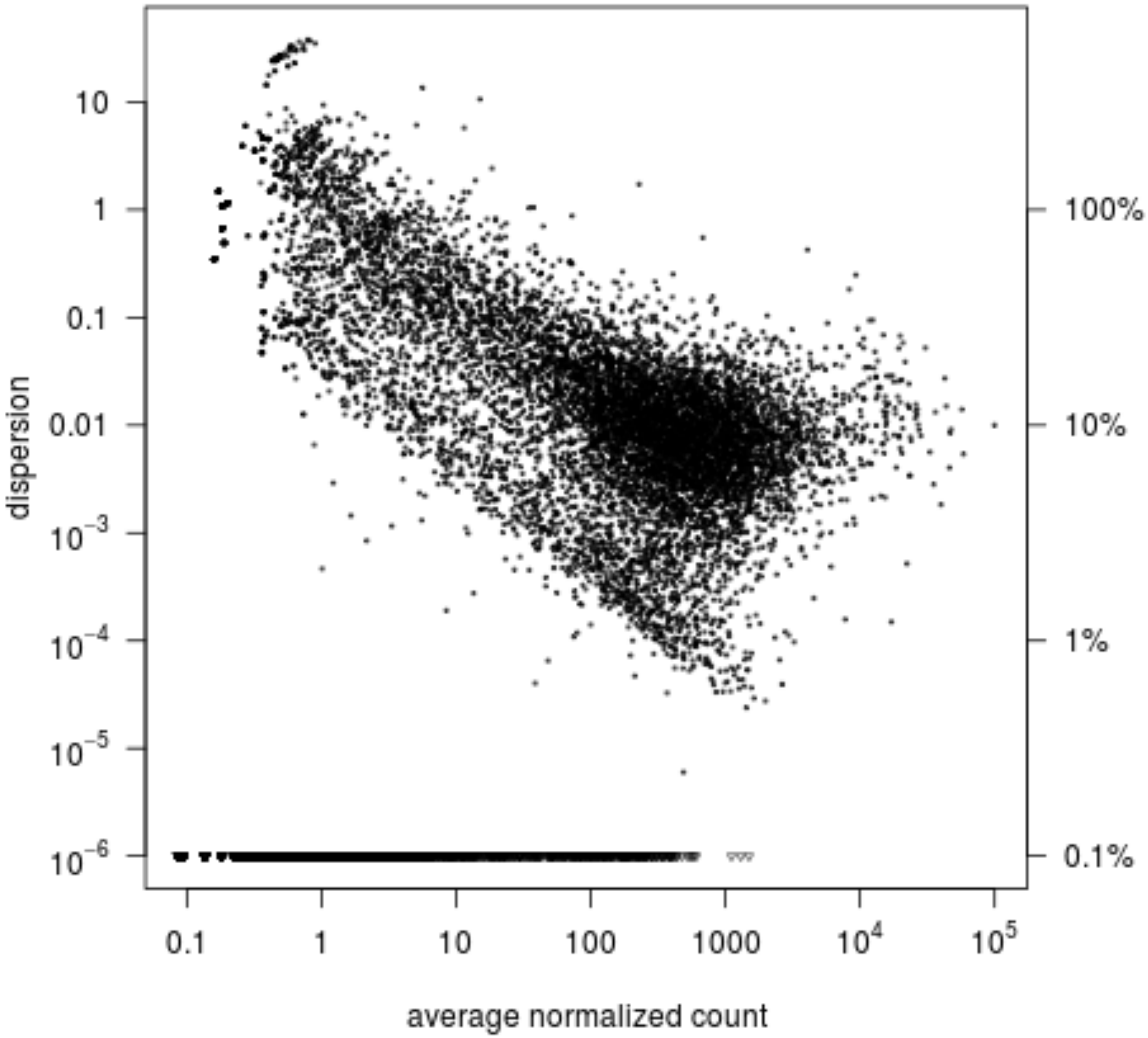
  $v = \mu + \alpha \mu^2$

- α : "dispersion" $\alpha = (\mu - v) / \mu^2$

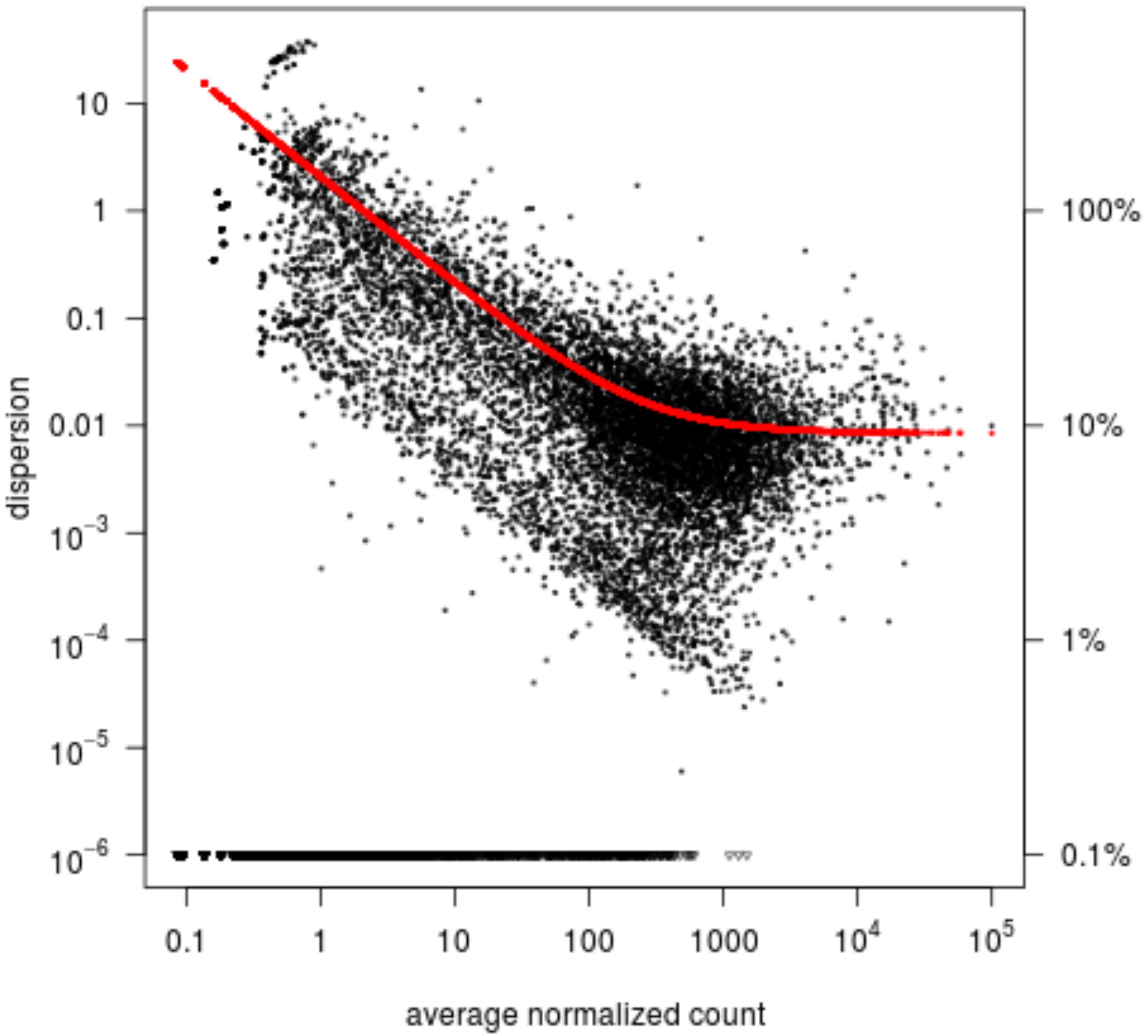
  (squared coefficient of variation of extra-Poisson variability)



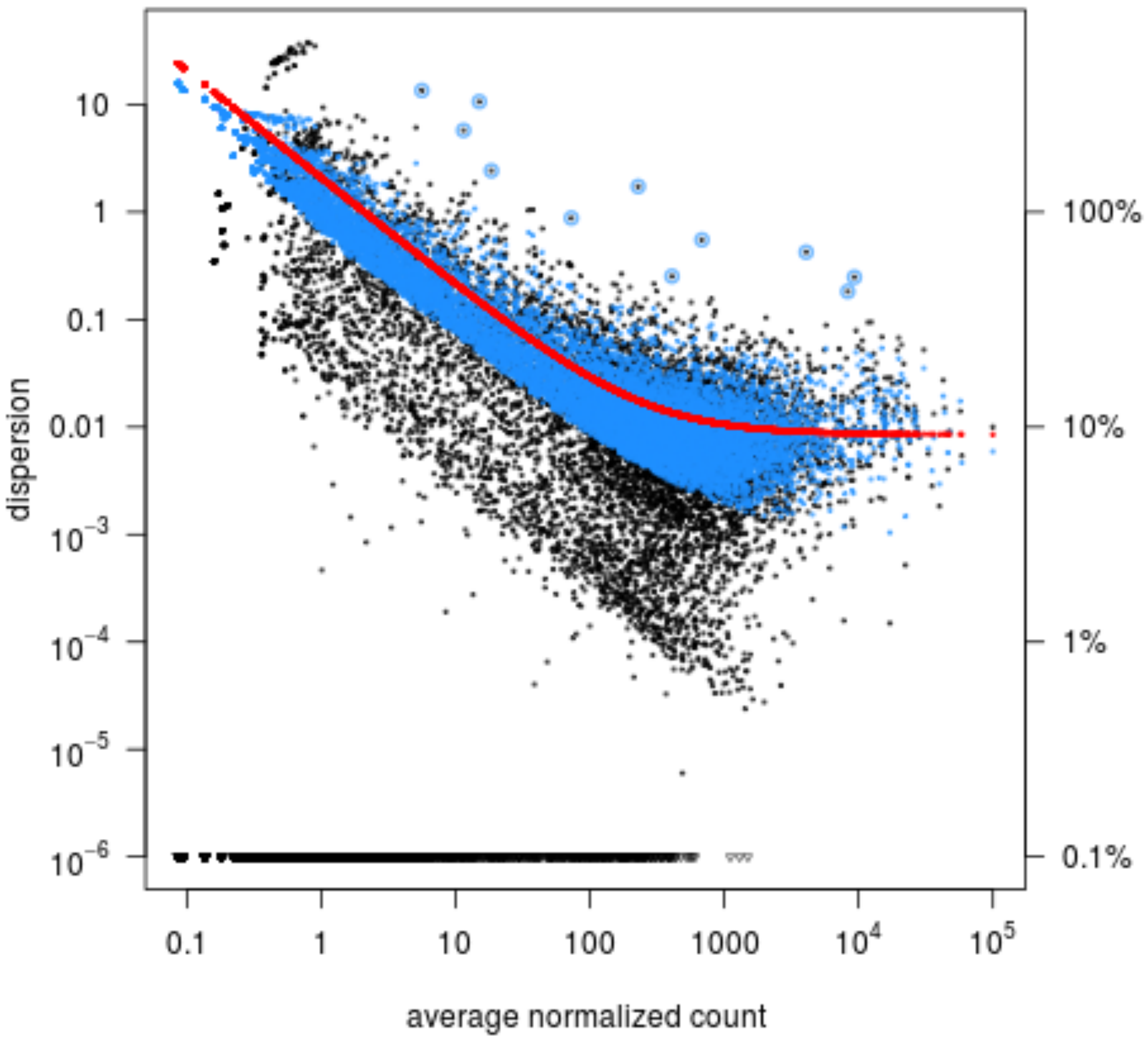
Shrinkage estimation of dispersion (within-group variability)



Shrinkage estimation of dispersion (within-group variability)

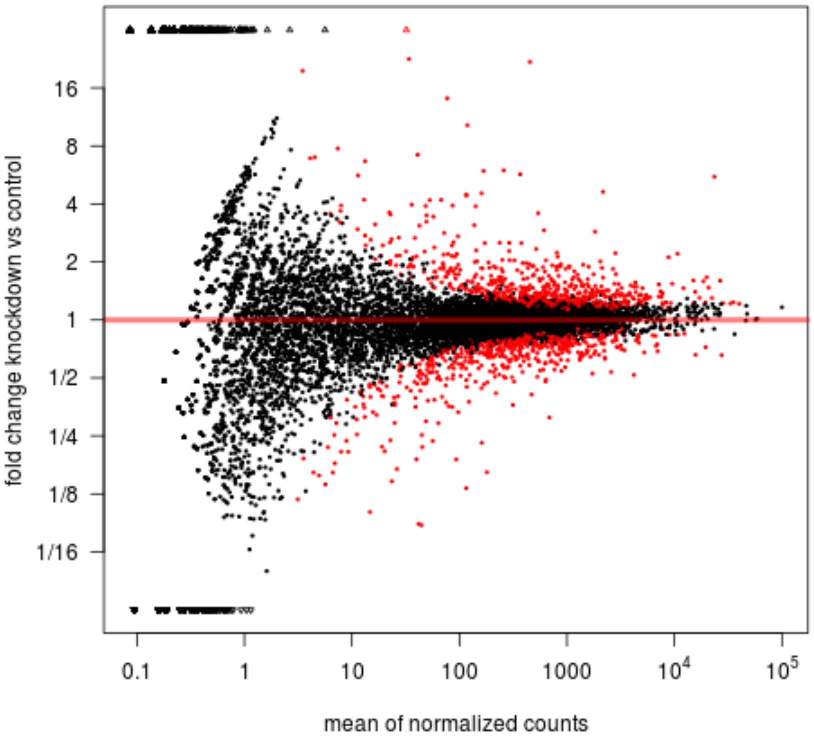


Shrinkage estimation of dispersion (within-group variability)

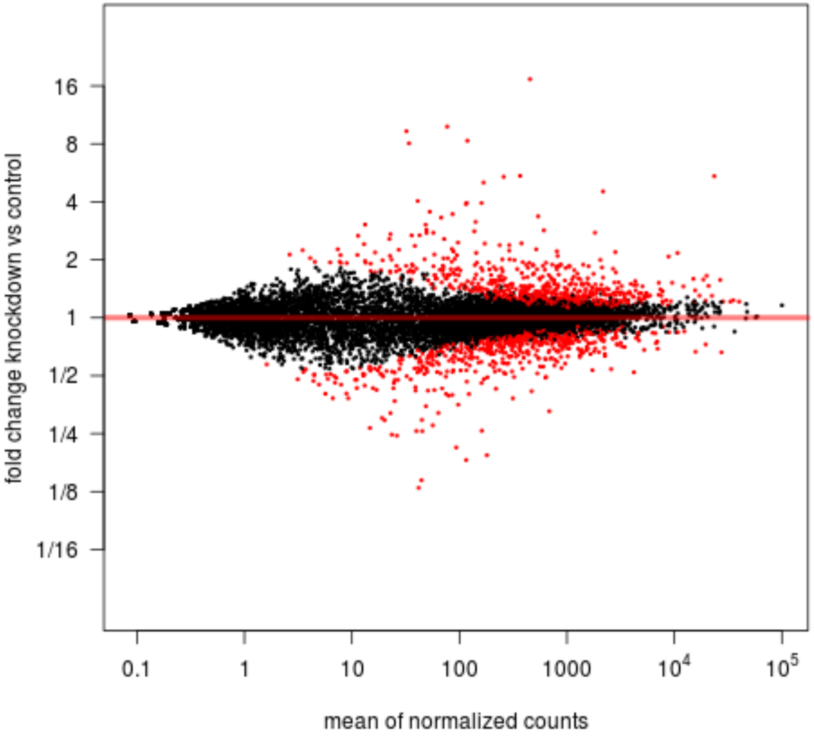


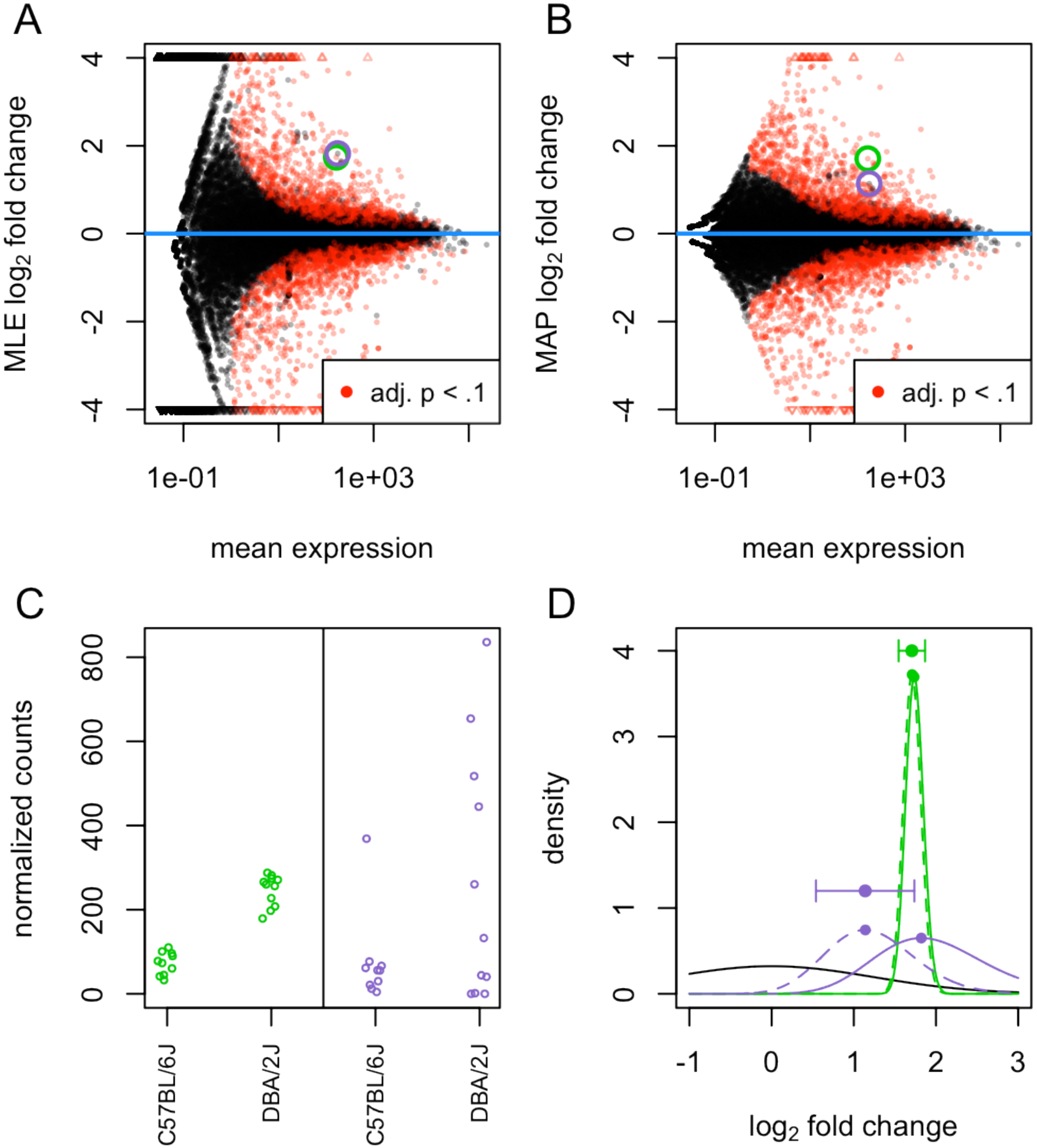
Shrinkage estimation of effect sizes

without shrinkage



with shrinkage




A
MLE $\log_2$ fold change
4
2
0
-2
-4
adj. p < .1
1e-01
1e+03
mean expression
B
MAP $\log_2$ fold change
4
2
0
-2
-4
adj. p < .1
1e-01
1e+03
mean expression
C
normalized counts
800
600
400
200
0
C57BL/6J
DBA/2J
C57BL/6J
DBA/2J
D
density
4
3
2
1
0
-1
0
1
2
3
$\log_2$ fold change

Complex designs

Simple: Comparison between two groups.

More complex:

- paired samples
- testing for interaction effects
- accounting for nuisance covariates
- …

GLMs: Blocking factor

Sample	treated	sex
S1	no	male
S2	no	male
S3	no	male
S4	no	female
S5	no	female
S6	yes	male
S7	yes	male
S8	yes	female
S9	yes	female
S10	yes	female

GLMs: Blocking factor

$$K_{ij} \sim NB(s_j \mu_{ij}, \alpha_{ij})$$

full model for gene *i*:

$$\log \mu_{ij} = \beta_i^0 + \beta_i^{\mathrm{S}} x_j^{\mathrm{S}} + \beta_i^{\mathrm{T}} x_j^{\mathrm{T}}$$

reduced model for gene *i*:

$$\log \mu_{ij} = \beta_i^0 + \beta_i^{\mathrm{S}} x_j^{\mathrm{S}}$$

GLMs: Interaction

$$K_{ij} \sim NB(s_j \mu_{ij}, \alpha_{ij})$$

full model for gene *i*:

$$\log \mu_{ij} = \beta_i^0 + \beta_i^{\mathrm{S}} x_j^{\mathrm{S}} + \beta_i^{\mathrm{T}} x_j^{\mathrm{T}} + \beta_i^{\mathrm{I}} x_j^{\mathrm{S}} x_j^{\mathrm{T}}$$

reduced model for gene *i*:

$$\log \mu_{ij} = \beta_i^0 + \beta_i^{\mathrm{S}} x_j^{\mathrm{S}} + \beta_i^{\mathrm{T}} x_j^{\mathrm{T}}$$

GLMs: paired designs

- Often, samples are paired (e.g., a tumour and a healthy-tissue sample from the same patient)

- Then, using pair identity as blocking factor improves power.

full model:

$$\log \mu_{ijl} = \beta_i^0 + \begin{cases} 0 & \text{for } \ l = 1(\text{healthy}) \\ \beta_i^{\mathrm{T}} & \text{for } \ l = 2(\text{tumour}) \end{cases}$$

reduced model:

$$\log \mu_{ij} = \beta_i^0$$

i	gene
j	subject
l	tissue state

GLMs: Dual-assay designs

How does the affinity of an RNA-binding protein to mRNA change under some drug treatment?

Prepare control and treated samples (in replicates) and perform on each sample RNA-Seq and CLIP-Seq.

For each sample, we are interested in the ratio of CLIP-Seq to RNA-Seq reads.

How is this ratio affected by treatment?

GLMs: CLIP-Seq/RNA-Seq assay

full model:

count ~ assayType + treatment + assayType:treatment

reduced model:

count ~ assayType + treatment

GLMs: CLIP-Seq/RNA-Seq assay

full model:

count ~ sample + assayType + assayType:treatment

reduced model:

count ~ sample + assayType

Genes and transcripts

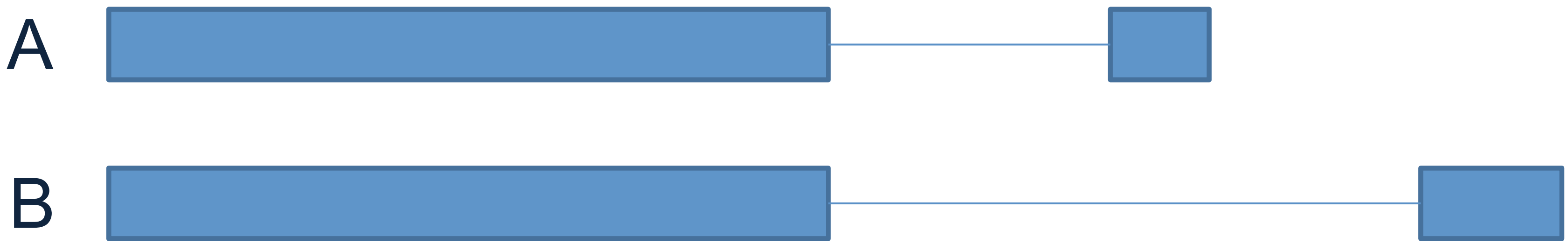
- So far, we looked at read counts *per gene*.

A gene's read count may increase

- because the gene produces *more* transcripts
- because the gene produces *longer* transcripts

How to look at gene sub-structure?

Assigning reads to transcripts



100 reads

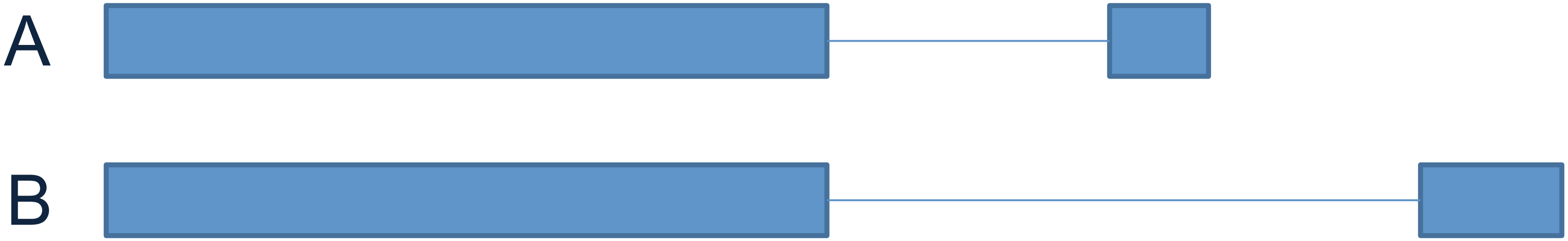
10 reads

30 reads

from A

from B

Assigning reads to transcripts



200 reads	5 reads	15 reads
(50 from A, 150 from B?)	from A	from B

total: A: 55 reads
B: 165 reads (accuracy?)

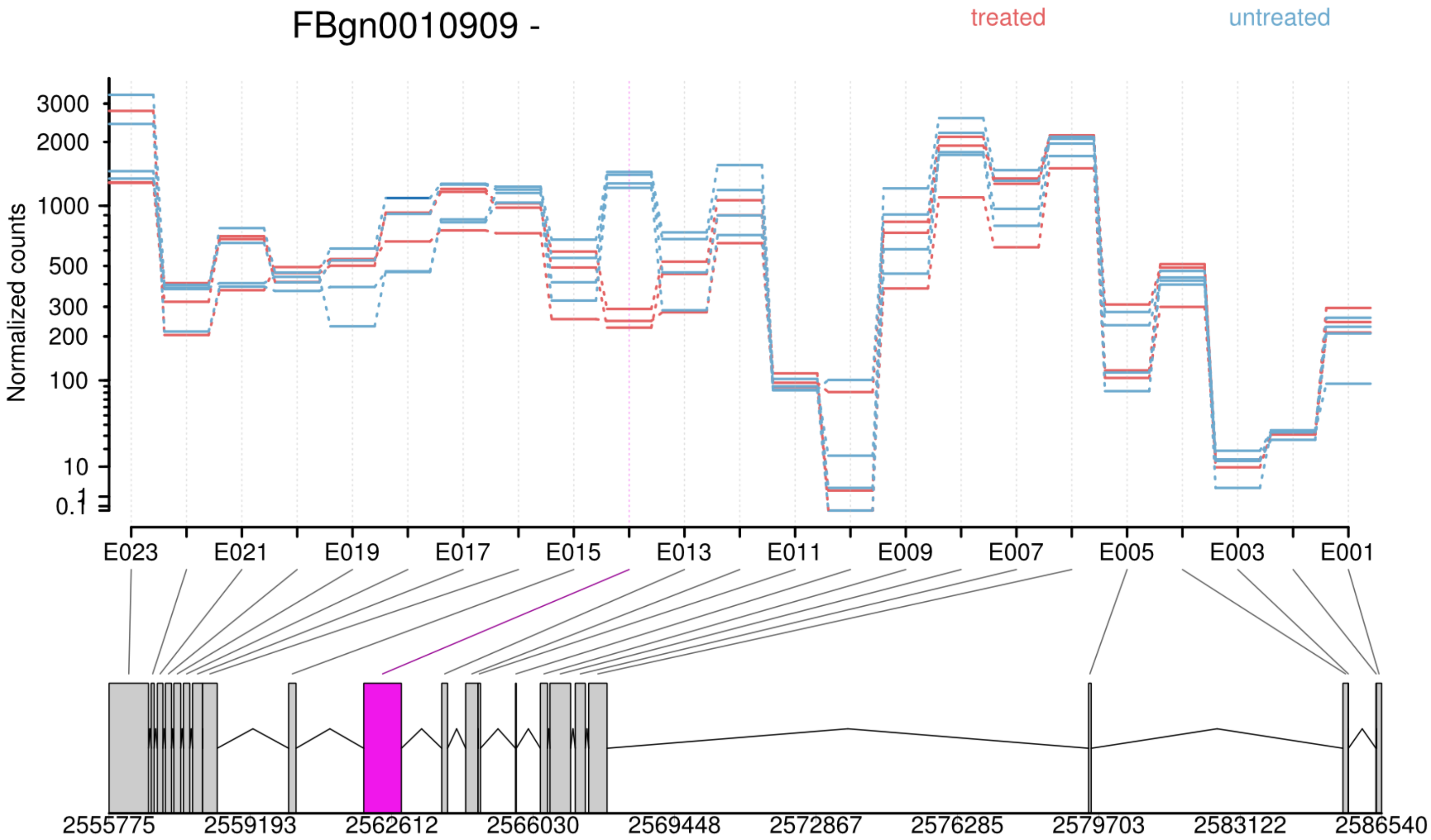
One step back:
Differential exon usage

Our tool, *DEXSeq*, tests for differential usage of exons.

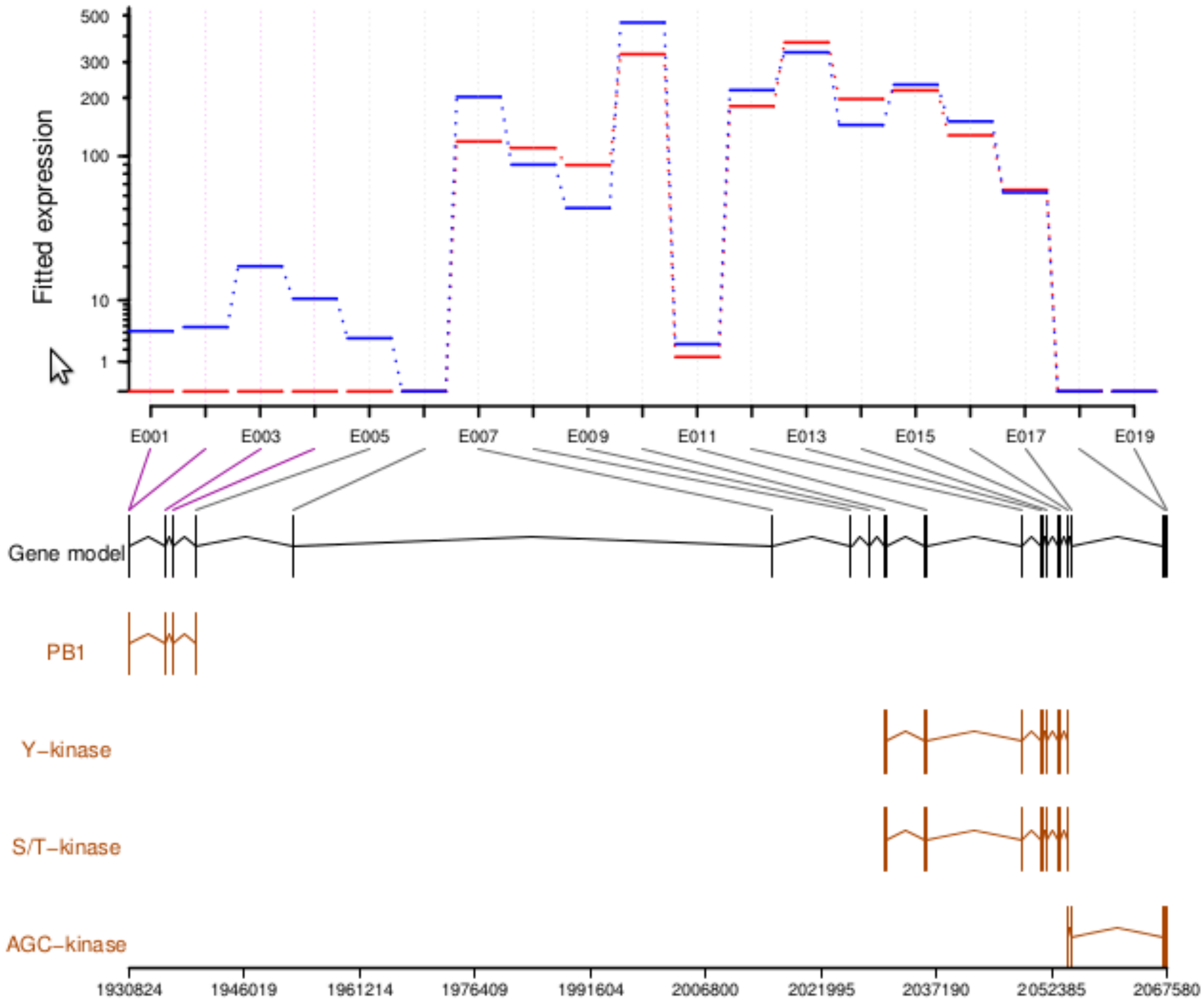
Usage on an exon =

$$\frac{\text{number of reads mapping to the exon}}{\text{number of reads mapping to any other exon of the same gene}}$$

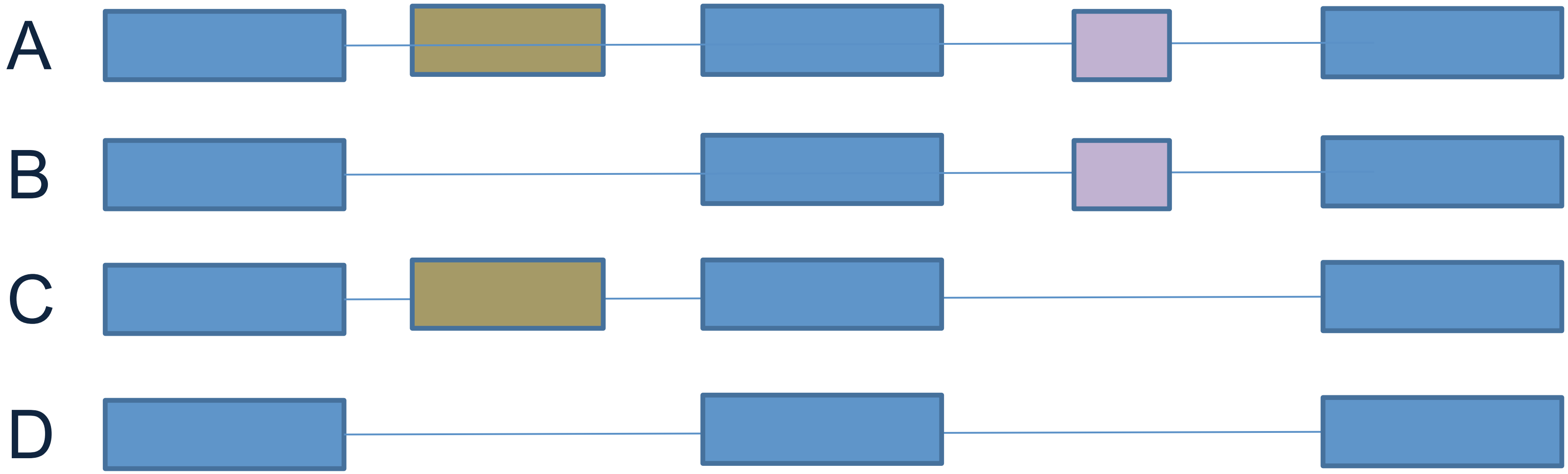
Differential exon usage -- Example



Differential exon usage -- Example



Differential usage of exons or of isoforms?



casette exon with well-understood function

casette exon with uncharacterized function

Summary

- Estimating fold-changes without estimating variability is pointless.
- Estimating variability from few samples requires information sharing across genes (shrinkage)
- Shrinkage can also regularize fold-change estimates. (New in DESeq2)

Acknowledgements

Co-authors:

- Wolfgang Huber
- Alejandro Reyes
- Mike Love (MPI-MG Berlin)

Thanks also to

- the rest of the Huber group
- all users who provided feed-back

Funding:

EMBL

SEVENTH FRAMEWORK PROGRAMME

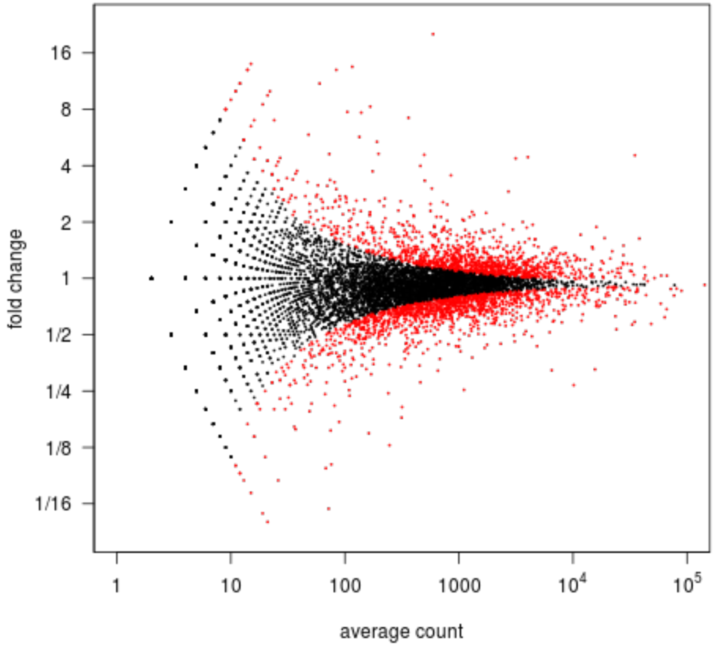
European Union:
FP7-health Project *Radia*

*

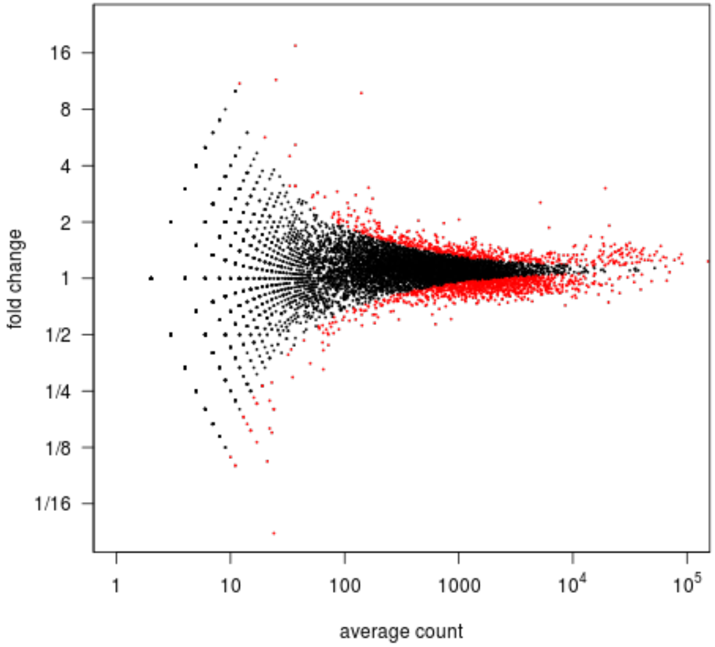
Fisher's exact test between two samples

Example data: fly cell culture, knock-down of pasilla
(Brooks et al., Genome Res., 2011)

knock-down sample T2
versus
control sample U3



control sample U2
versus
control sample U3



red: significant genes according to Fisher test (at 10% FDR)