UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Comparative analysis of high-throughput sequencing data with *DESeq 2*

Simon Anders

University of 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

DESeq / DESeq2

- Method for count data regression
- R/Bioconductor package
- widely used,
  part of many standard workflows

Anders and Huber, Genome Biology, 2010

Love, Huber, Anders, Genome Biology, 2014

Count data in high-throughput sequencing

samples

genes

`	control_1	control_2	control_3	treated_1	treated_2	tre
ENSG00000000003	792	1064	444	953	519	
ENSG00000000005	4	1	2	3	3	
ENSG00000000419	294	282	164	263	179	
ENSG00000000457	156	184	93	145	75	
ENSG00000000460	396	207	210	212	221	
ENSG00000000938	3	8	2	5	0	
ENSG00000000971	12	23	10	12	4	
ENSG00000001036	2536	2349	1438	2307	1339	
ENSG00000001084	385	411	244	457	243	
ENSG00000001167	374	464	218	396	274	
ENSG00000001460	78	103	48	73	42	
ENSG00000001461	441	560	256	495	276	
ENSG00000001497	497	467	289	443	350	
ENSG00000001561	500	644	299	521	295	
ENSG00000001617	67	114	29	94	45	
ENSG00000001626	1	1	0	1	0	
ENSG00000001629	1151	1382	620	1229	791	
ENSG00000001630	450	501	284	547	255	
ENSG00000001631	463	515	251	525	309	
ENSG00000002016	129	157	65	137	78	

[switch to live demo]

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
[...]					

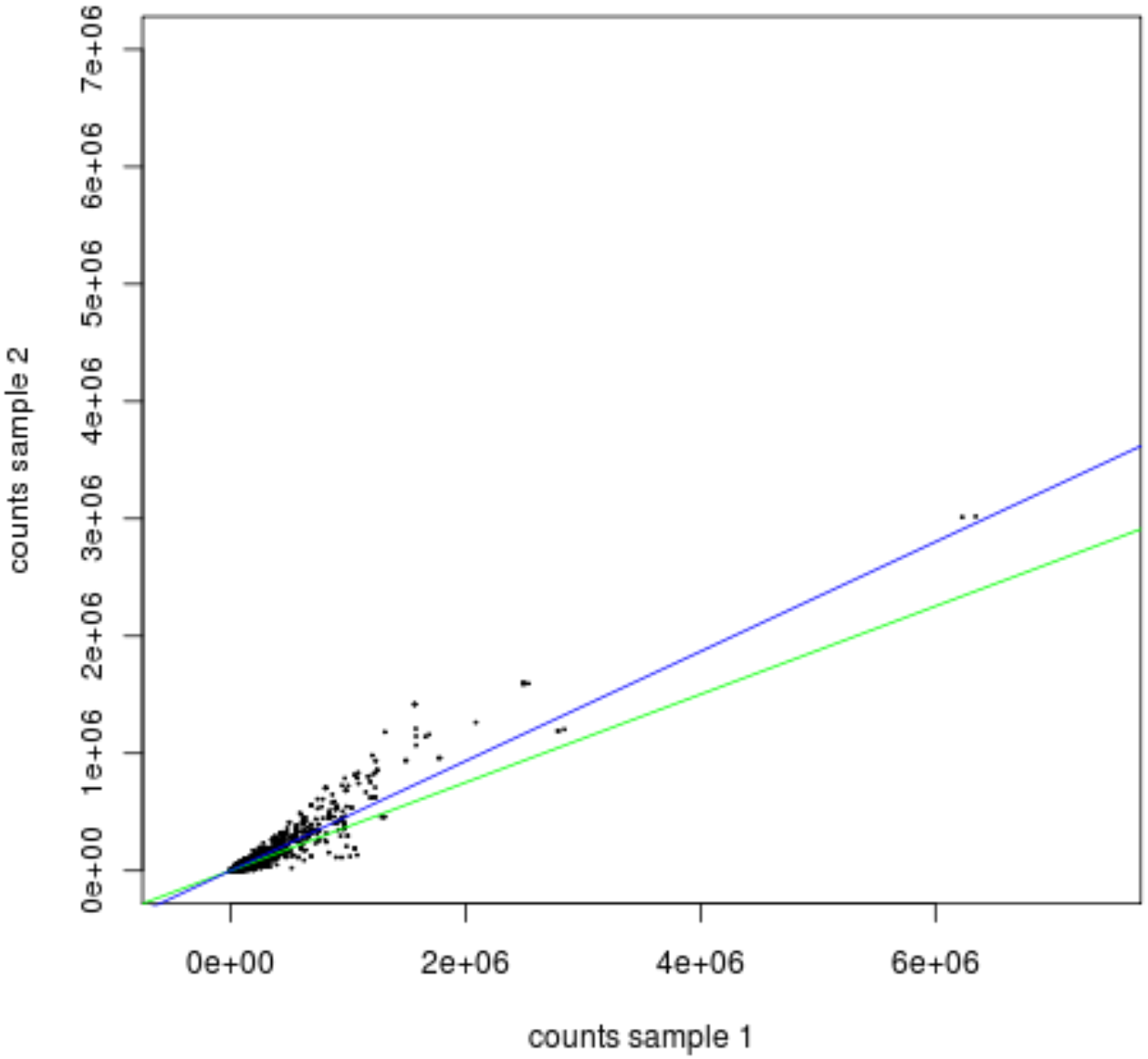
- RNA-Seq
- Tag-Seq
- ChIP-Seq
- HiC
- Bar-Seq
- ...

Normalization for library size

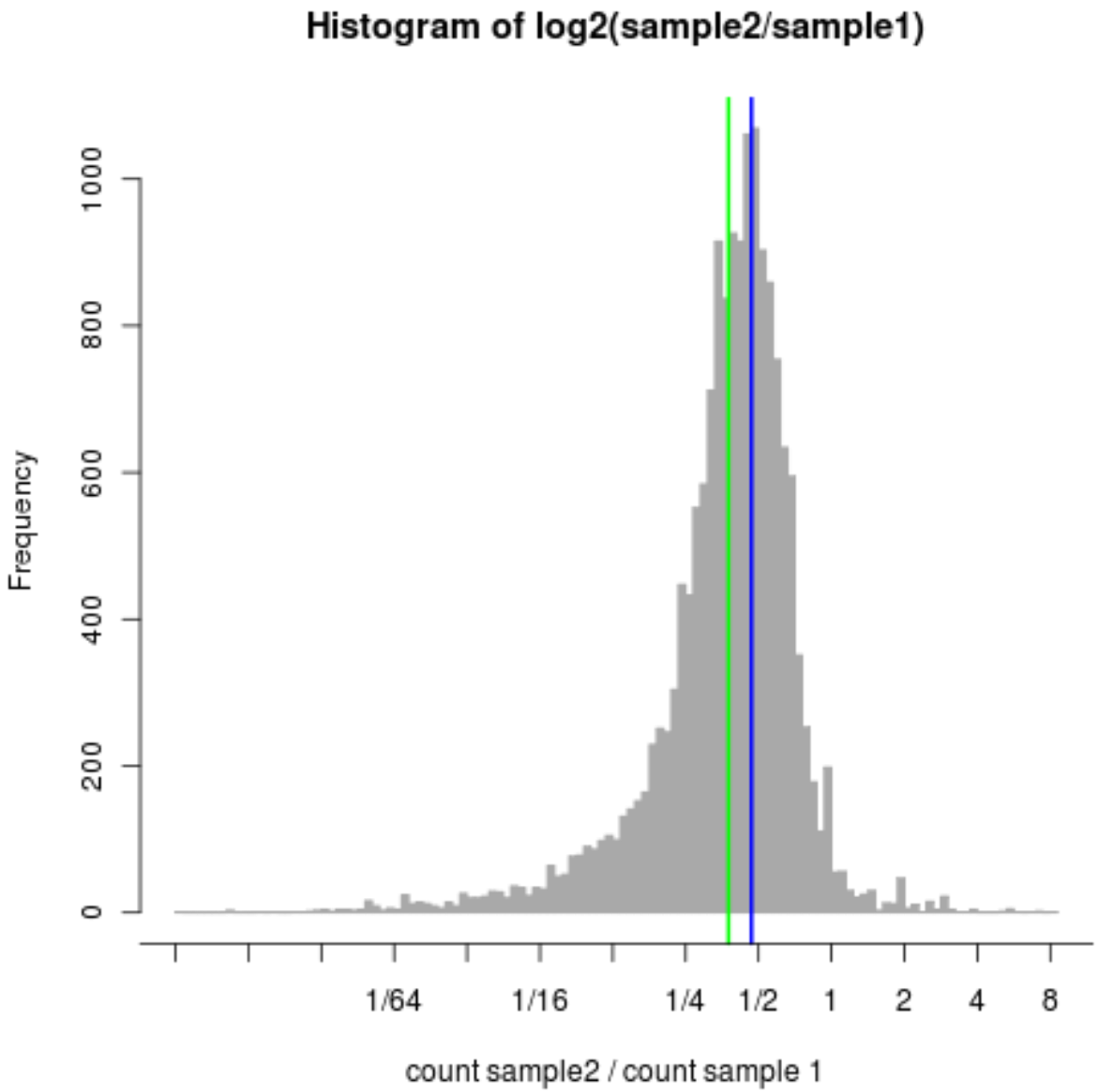
- If sample A has been sampled deeper than sample B, we expect counts to be higher.

- 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



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.

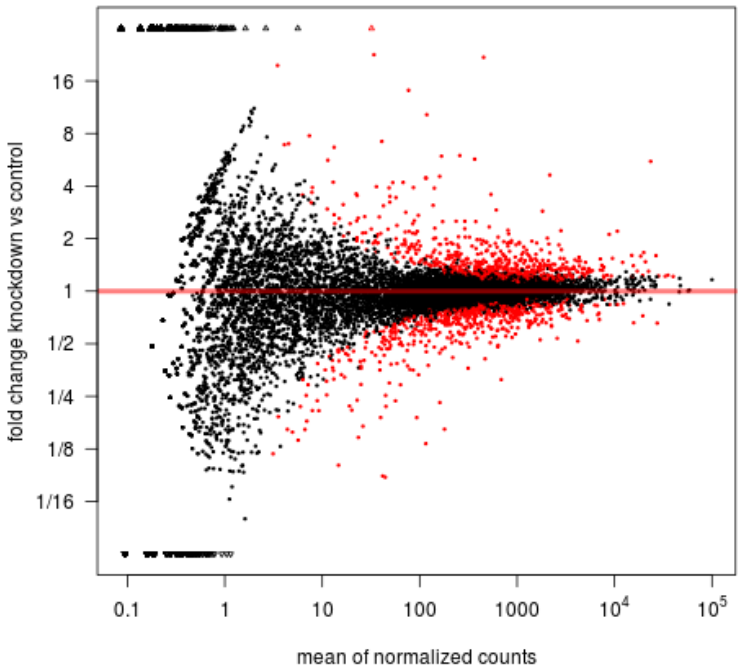
Anders and Huber, 2010

similar approach: Robinson and Oshlack, 2010

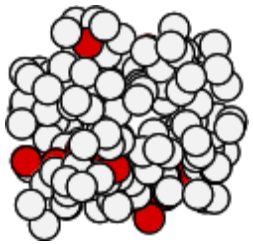
Counting noise

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

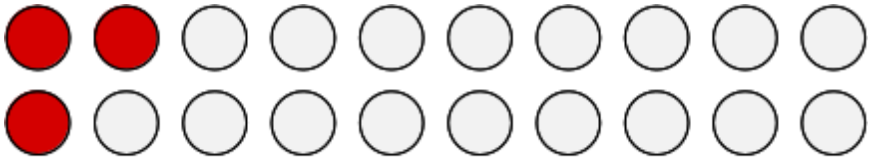
Why?



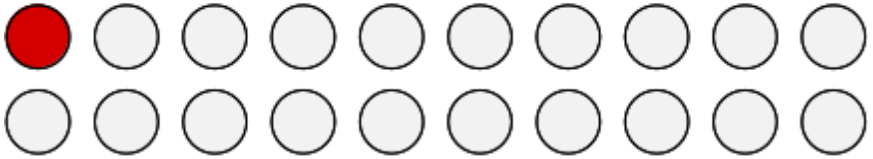
The Poisson distribution



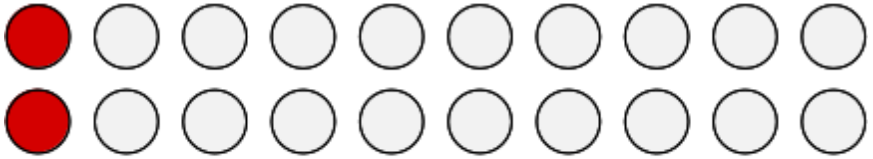
- This bag contains very many small balls, 10% of which are red.
- Several experimenters are tasked with determining the percentage of red balls.
- Each of them is permitted to draw 20 balls out of the bag, without looking.



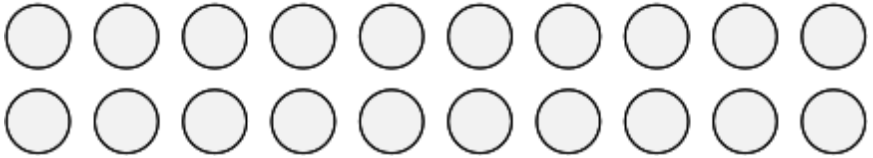
3 / 20 = 15%



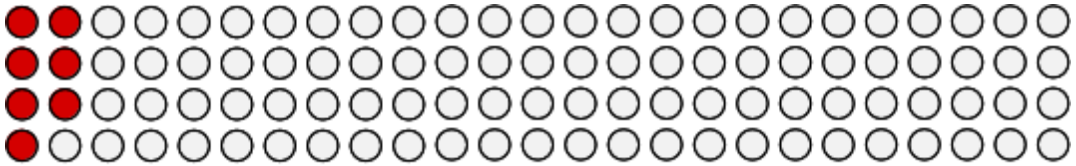
1 / 20 = 5%



2 / 20 = 10%



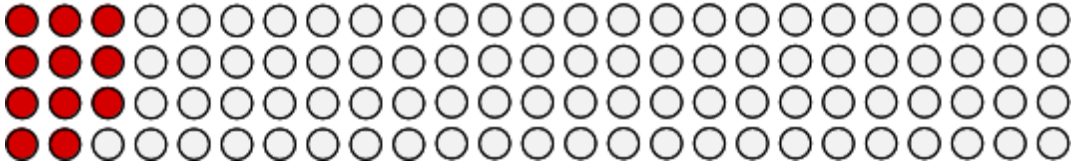
0 / 20 = 0%



7 / 100 = 7%

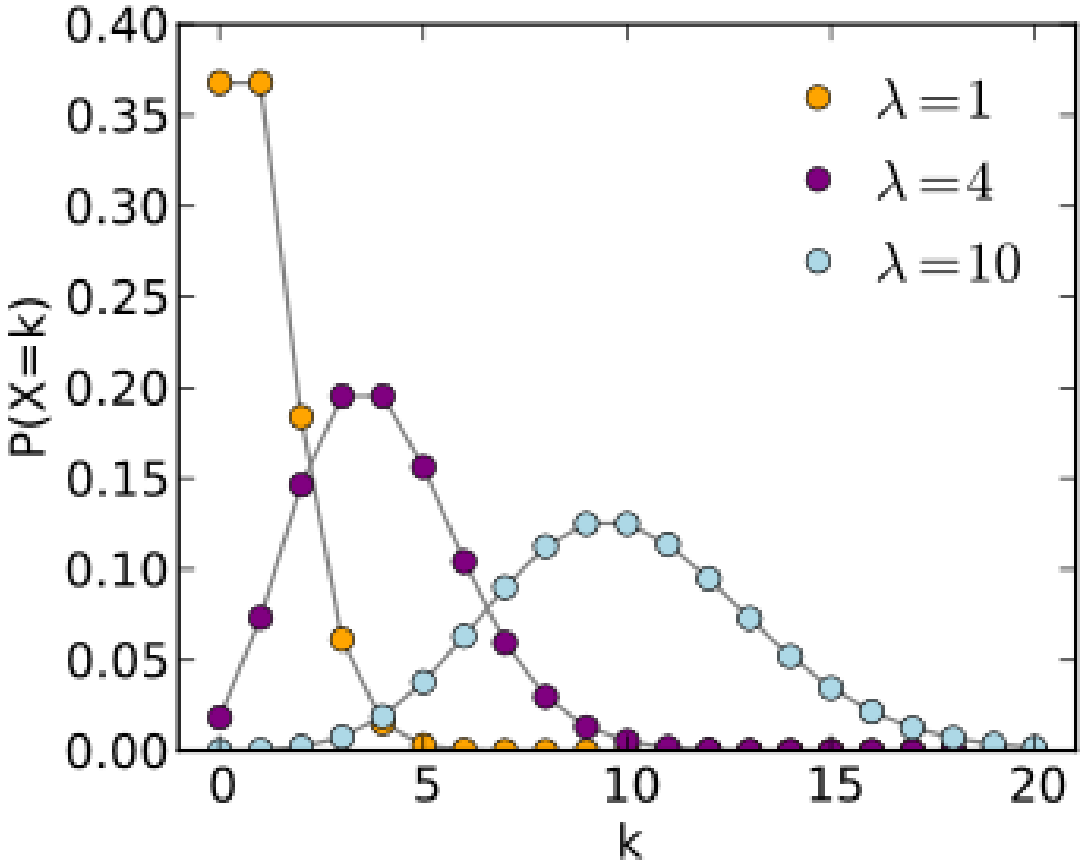
10 / 100 = 10%

8 / 100 = 8%



11 / 100 = 11%

The Poisson distribution



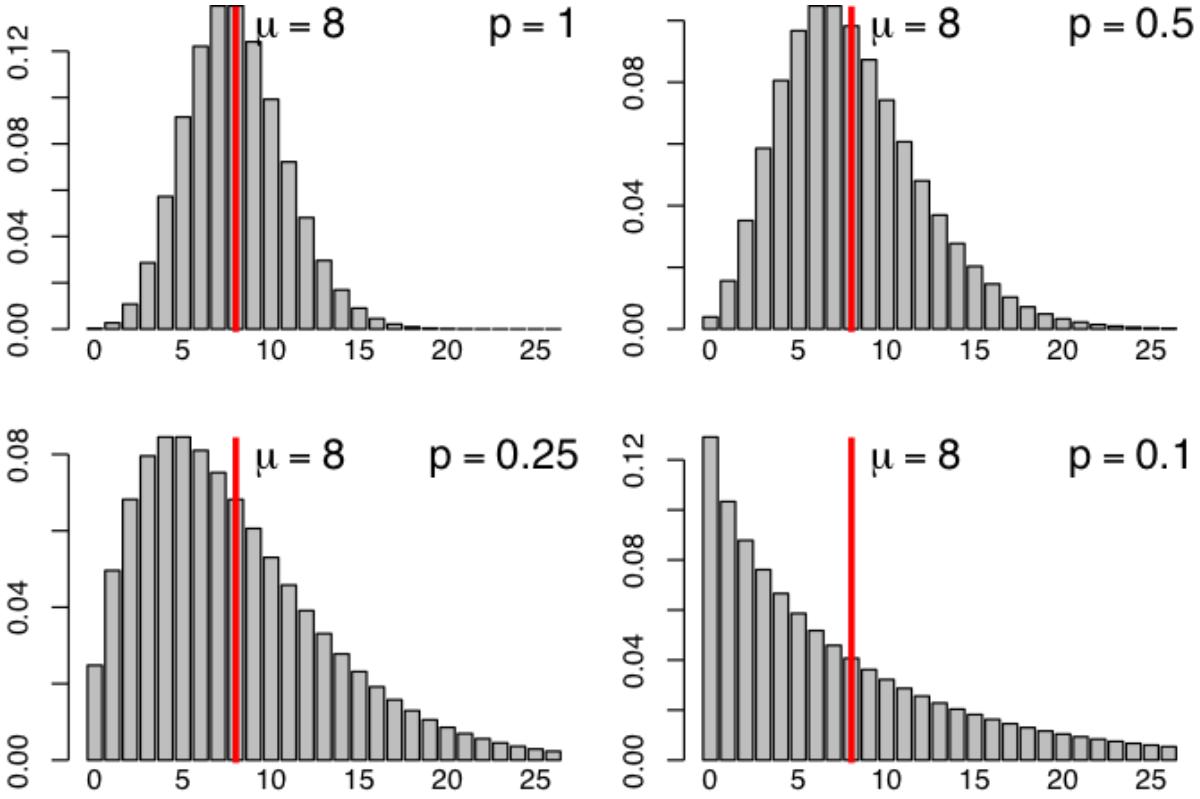
$$\Pr(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

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	$\sqrt{10} = 3$	$1 / \sqrt{10} = 31.6\%$
100	$\sqrt{100} = 10$	$1 / \sqrt{100} = 10.0\%$
1,000	$\sqrt{1{,}000} = 32$	$1 / \sqrt{1000} = 3.2\%$
10,000	$\sqrt{10{,}000} = 100$	$1 / \sqrt{10000} = 1.0\%$

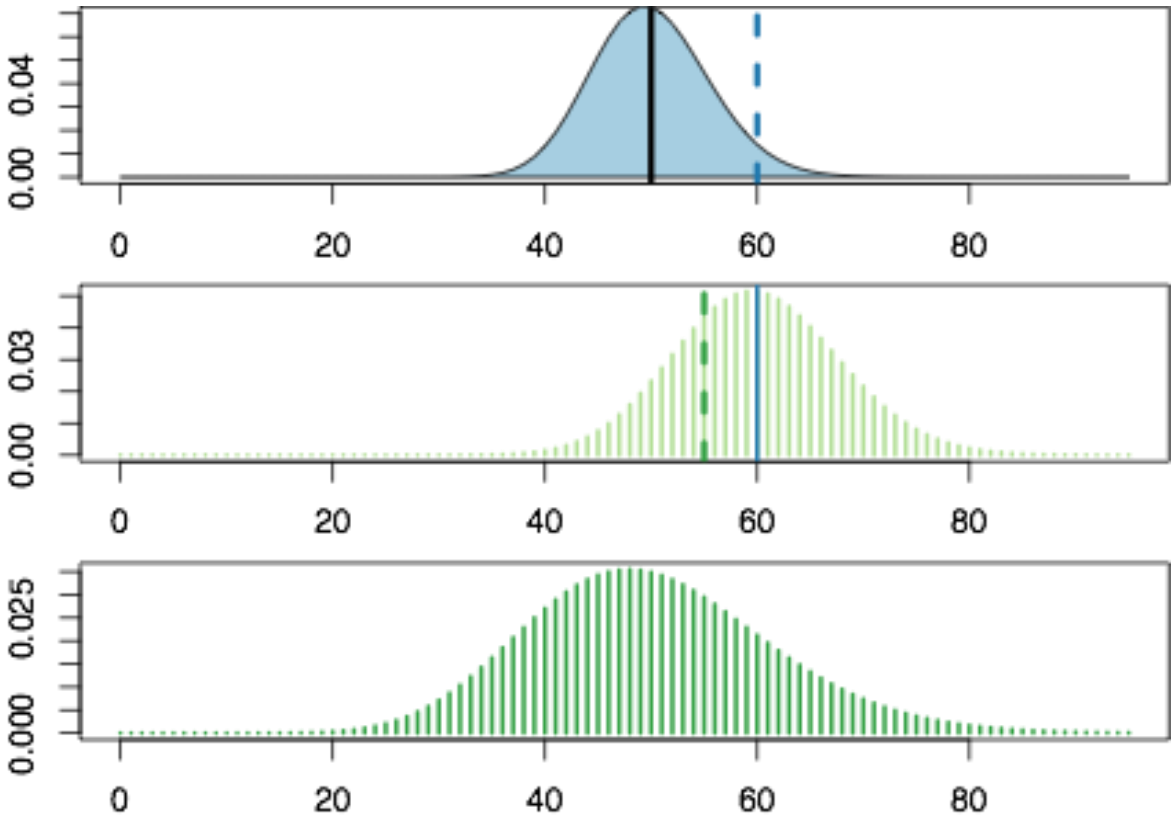
The negative binomial distribution

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.

Assumption:

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

Null hypothesis:

- All samples have the same μ_j.

Alternative hypothesis:

- Mean is the same only within groups:

$$\log \mu_j = \beta_0 + x_j \beta_T$$

$x_j = 0$ for if *j* is control sample

$x_j = 1$ for if *j* is treatment sample

Testing: Generalized linear models

$$\log \mu_j = \beta_0 + x_j \beta_T$$

$x_j = 0$ 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.

Is the value for β_T significantly different from null?

Can we reject the null hypothesis that it is merely cause by noise?

The Wald test gives us a *p* value.

p values

The p value from the Wald test indicates the probability that the observed difference between treatment and control (as indicated by β_T), or an even stronger one, is observed even though the there is no true treatment effect.

p values

Assuming that there is no true effect,

what is the probability

of seeing the observed effect

or an even stronger one?

Multiple testing

- Consider: A genome with 10,000 genes
- We compare treatment and control. Unbeknownst to us, the treatment had no effect at all.
- How many genes will have $p < 0.05$?

Multiple testing

- Consider: A genome with 10,000 genes
- We compare treatment and control. Unbeknownst to us, the treatment had no effect at all.
- How many genes will have $p < 0.05$?

- $0.05 \times 10{,}000 = 500$ genes.

Multiple testing

- Consider: A genome with 10,000 genes
- We compare treatment and control
- Now, the treatment is real.

- 1,500 genes have $p < 0.05$.
- How many of these are false positives?

Multiple testing

- Consider: A genome with 10,000 genes
- We compare treatment and control
- Now, the treatment is real.

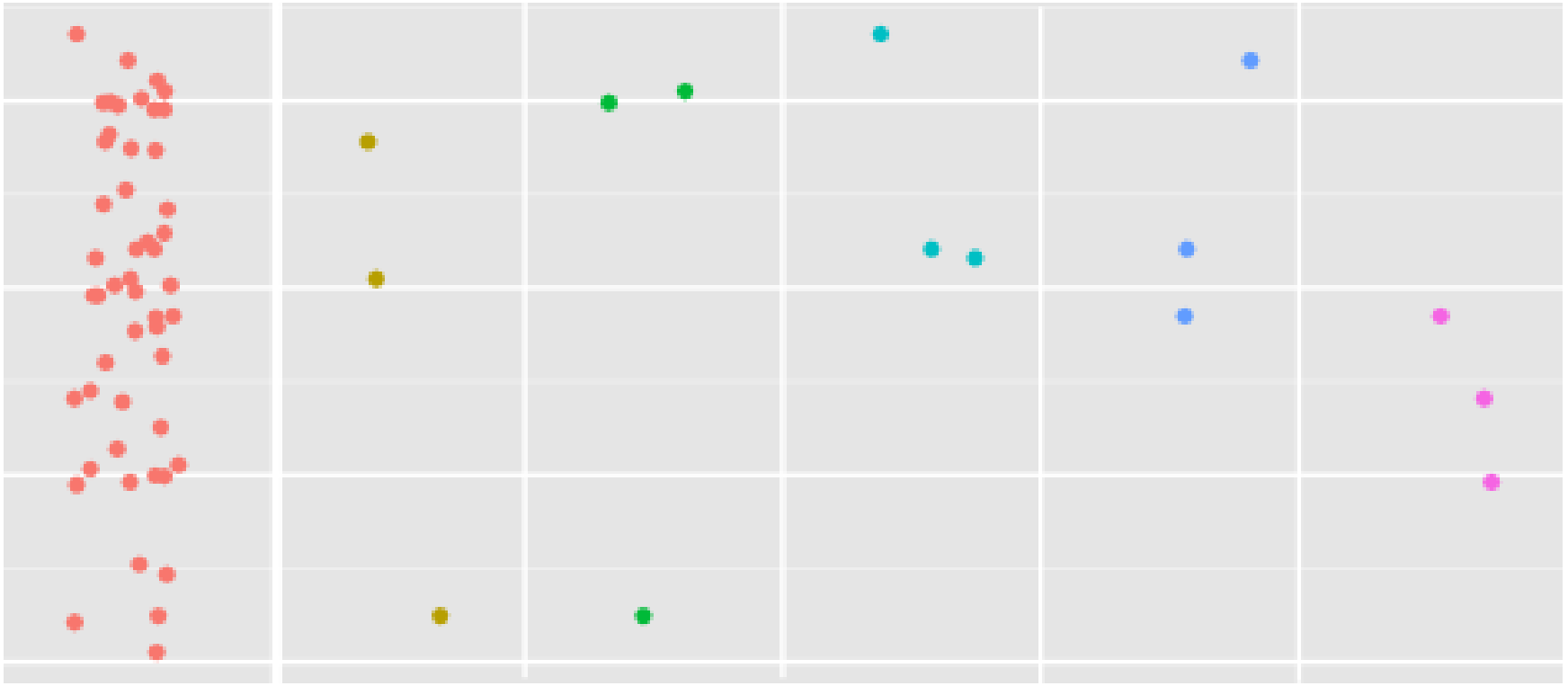
- 1,500 genes have $p < 0.05$.
- How many of these are false positives?

- 500 genes, i.e., 33%

Dispersion

- A crucial input to the GLM procedure and the Wald test is the estimated strength of within-group variability.

- Getting this right is the hard part.



Estimation of variability is the bottleneck

Example: A gene differs by 20% between samples within a group (CV=0.2)

What fold change gives rise to p=0.0001?

Number of samples	4	6	8	10	20	100
CV known	55%	45%	39%	35%	35%	11%
CV estimated						

(assuming normality and use of z or t test, resp.)

Estimation of variability is the bottleneck

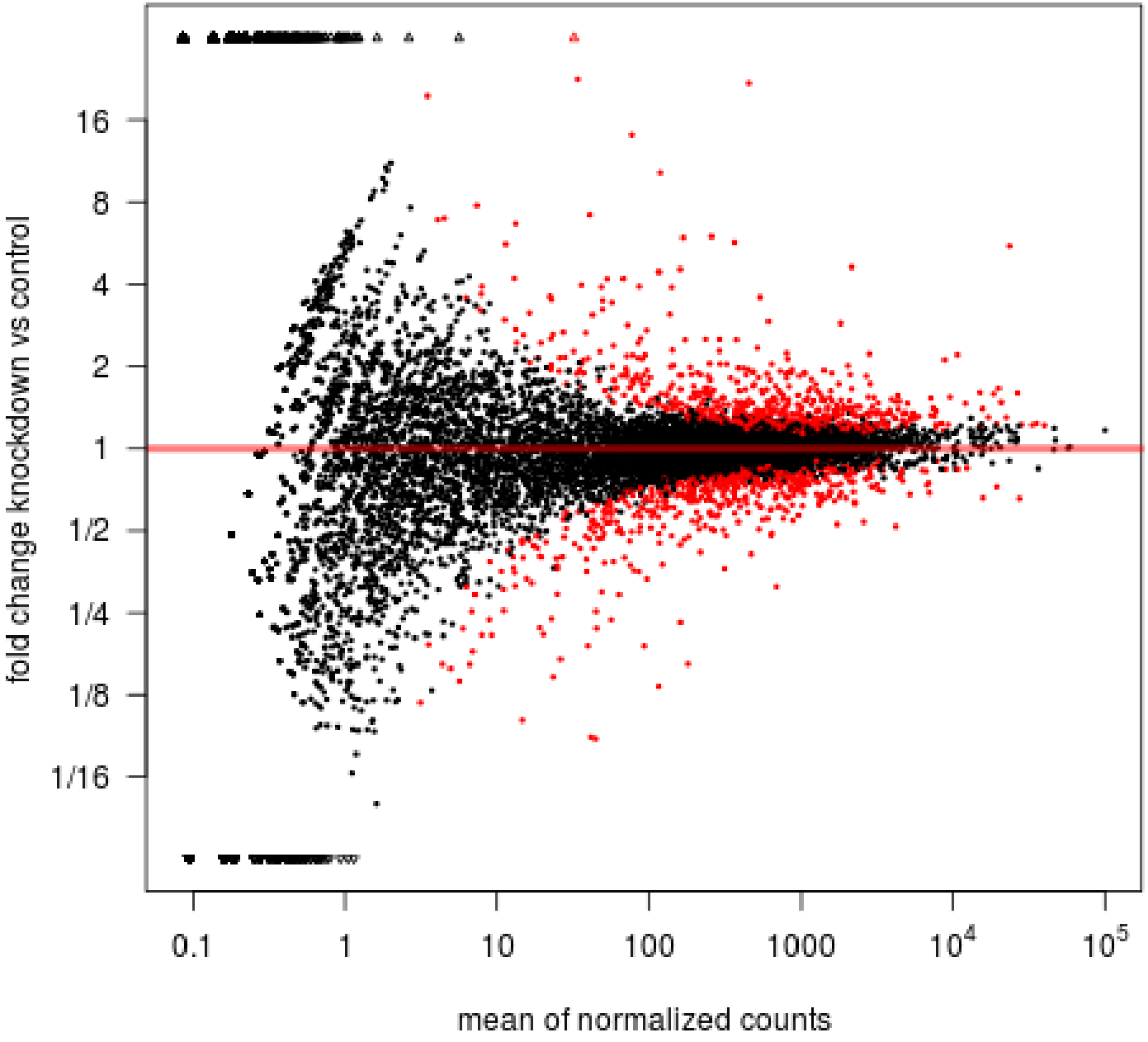
Example: A gene differs by 20% between samples within a group (CV=0.2)

What fold change gives rise to p=0.0001?

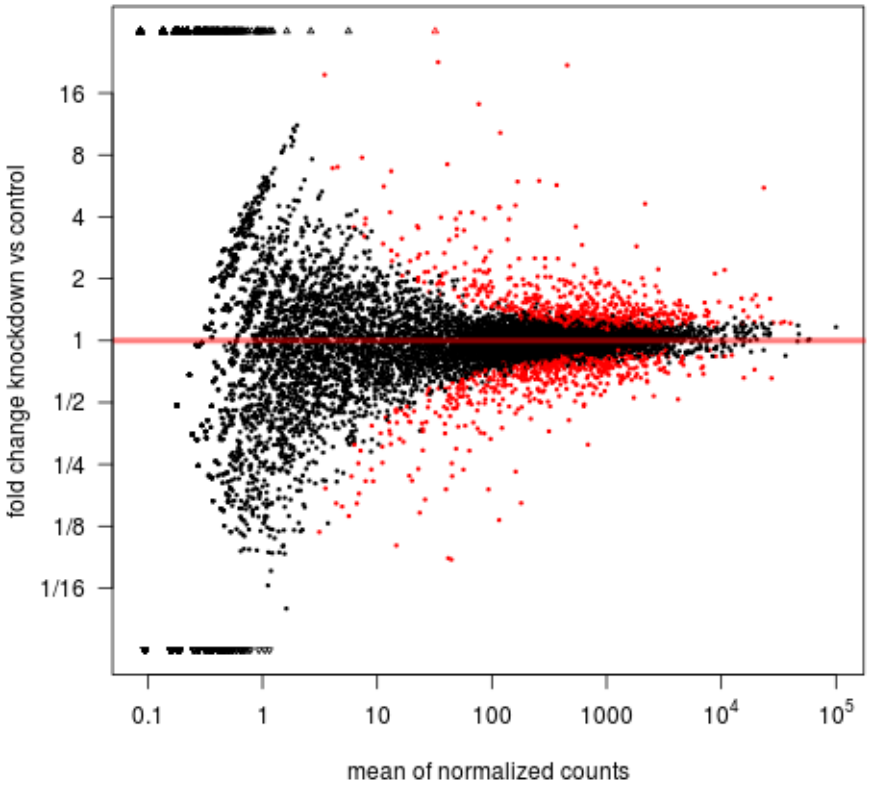
Number of samples	4	6	8	10	20	100
CV known	55%	45%	39%	35%	35%	11%
CV estimated	1400% (14x)	180% (1.8x)	91%	64%	31%	11%

(assuming normality and use of z or t test, resp.)

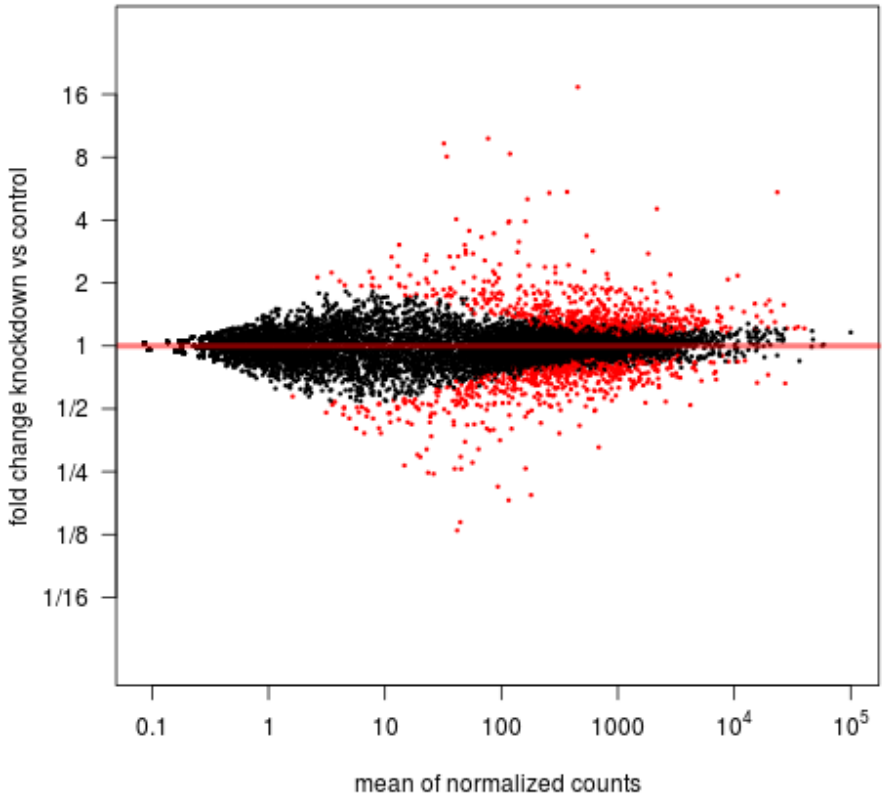
Shrinkage estimation of log fold changes

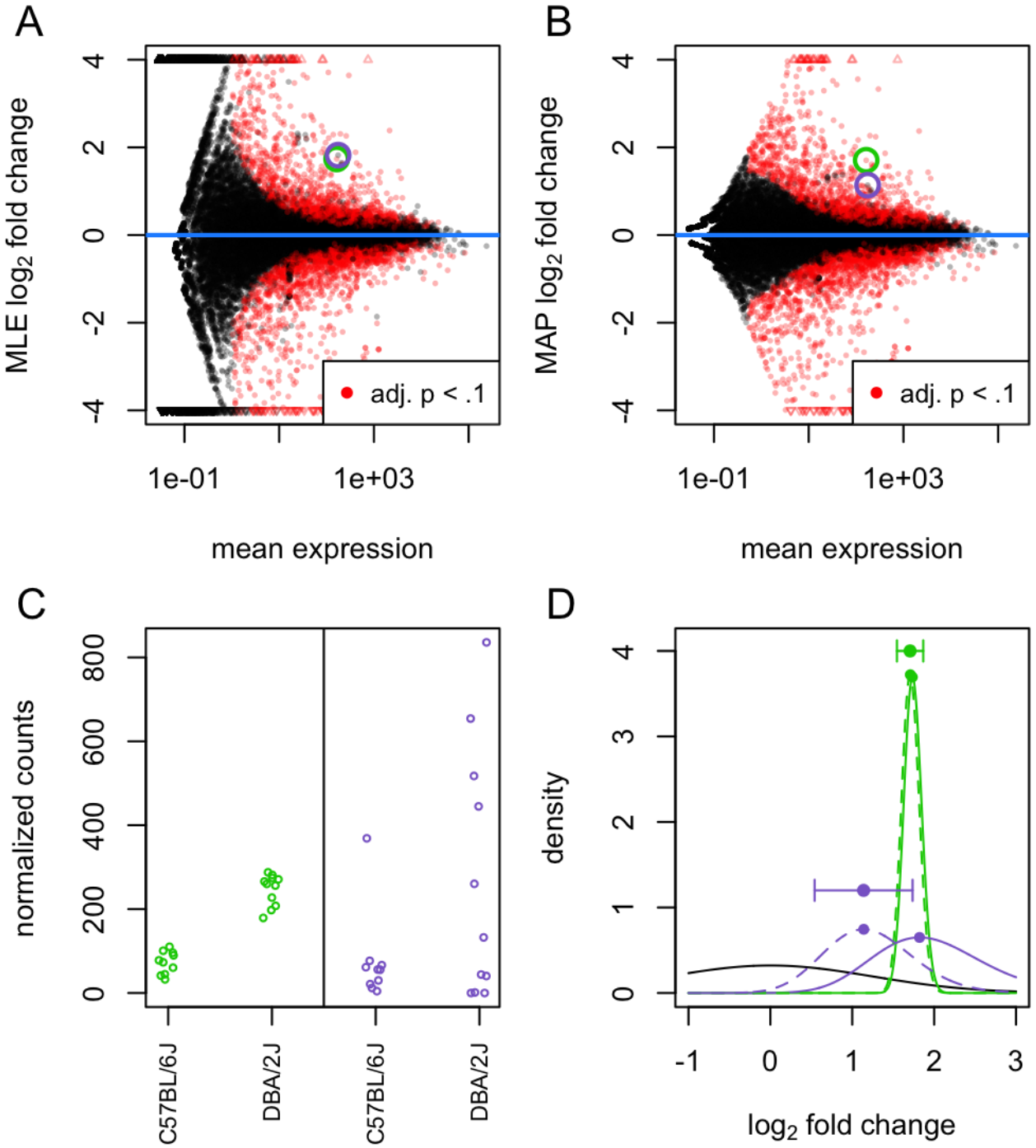

fold change knockdown vs control
16
8
4
2
1
1/2
1/4
1/8
1/16
0.1
1
10
100
1000
10^4
10^5
mean of normalized counts

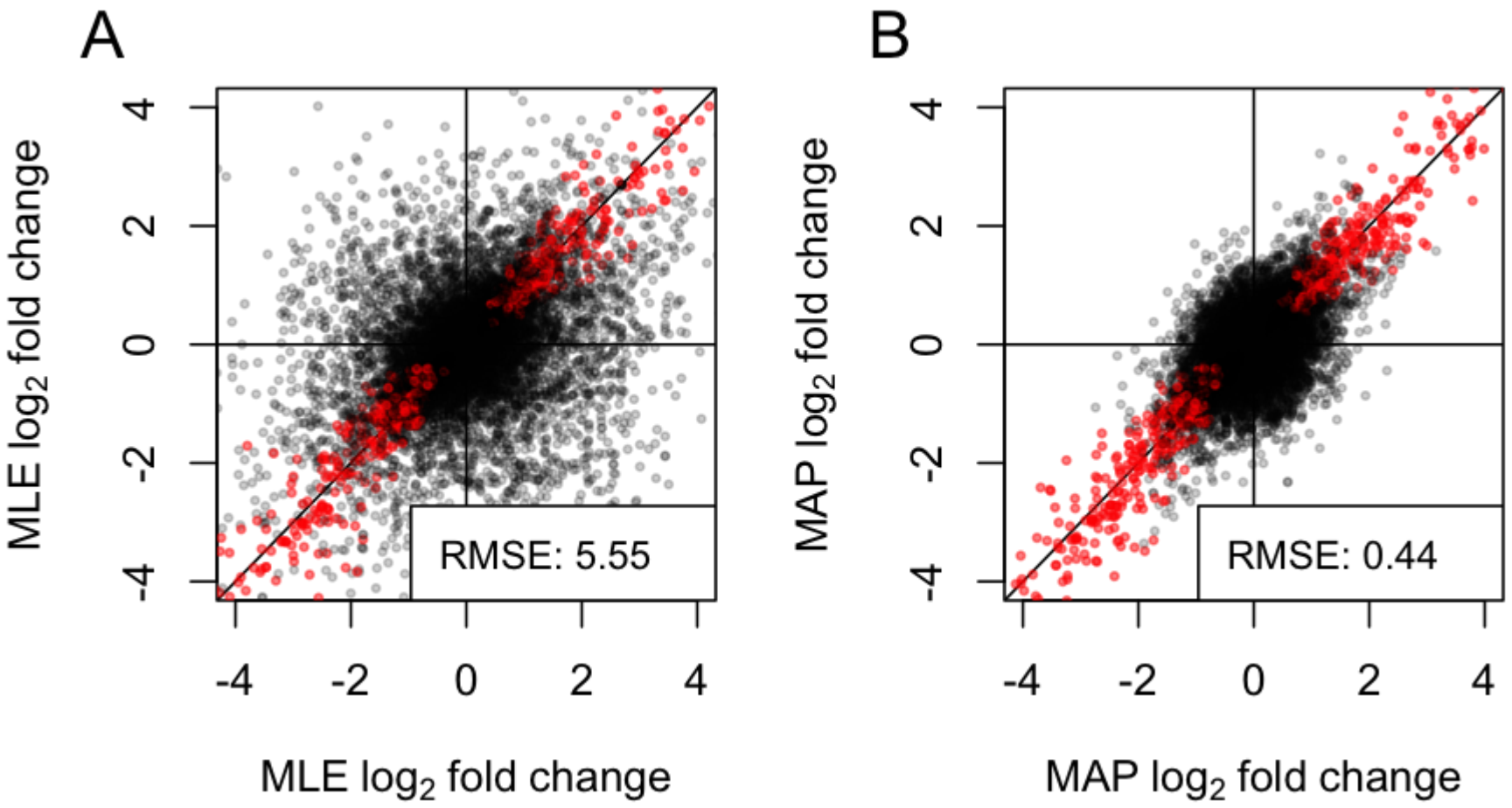
without shrinkage



with shrinkage




A
MLE log2 fold change
4
2
0
-2
-4
adj. p < .1
1e-01
1e+03
mean expression
B
MAP log2 fold change
4
2
0
-2
-4
adj. p < .1
1e-01
1e+03
mean expression
C
normalized counts
0
200
400
600
800
C57BL/6J
DBA/2J
C57BL/6J
DBA/2J
D
density
0
1
2
3
4
-1
0
1
2
3
log2 fold change


A
MLE log2 fold change
4
2
0
-2
-4
RMSE: 5.55
-4
-2
0
2
4
MLE log2 fold change
B
MAP log2 fold change
4
2
0
-2
-4
RMSE: 0.44
-4
-2
0
2
4
MAP log2 fold change

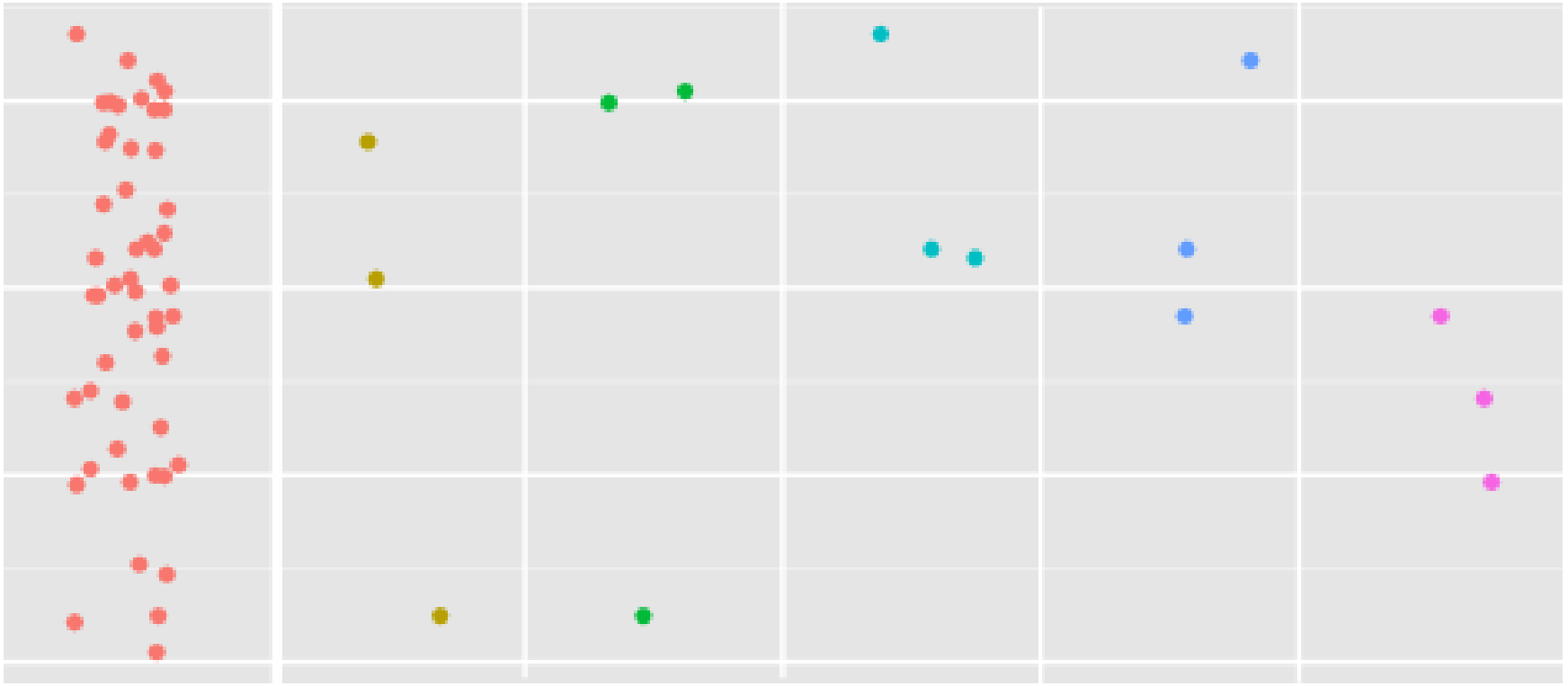
Gene ranking

How to rank a gene list to prioritize down-stream experiments?

- by p value?
- by log fold change?

- by *shrunken* log fold change!

Shrinkage estimation of dispersions

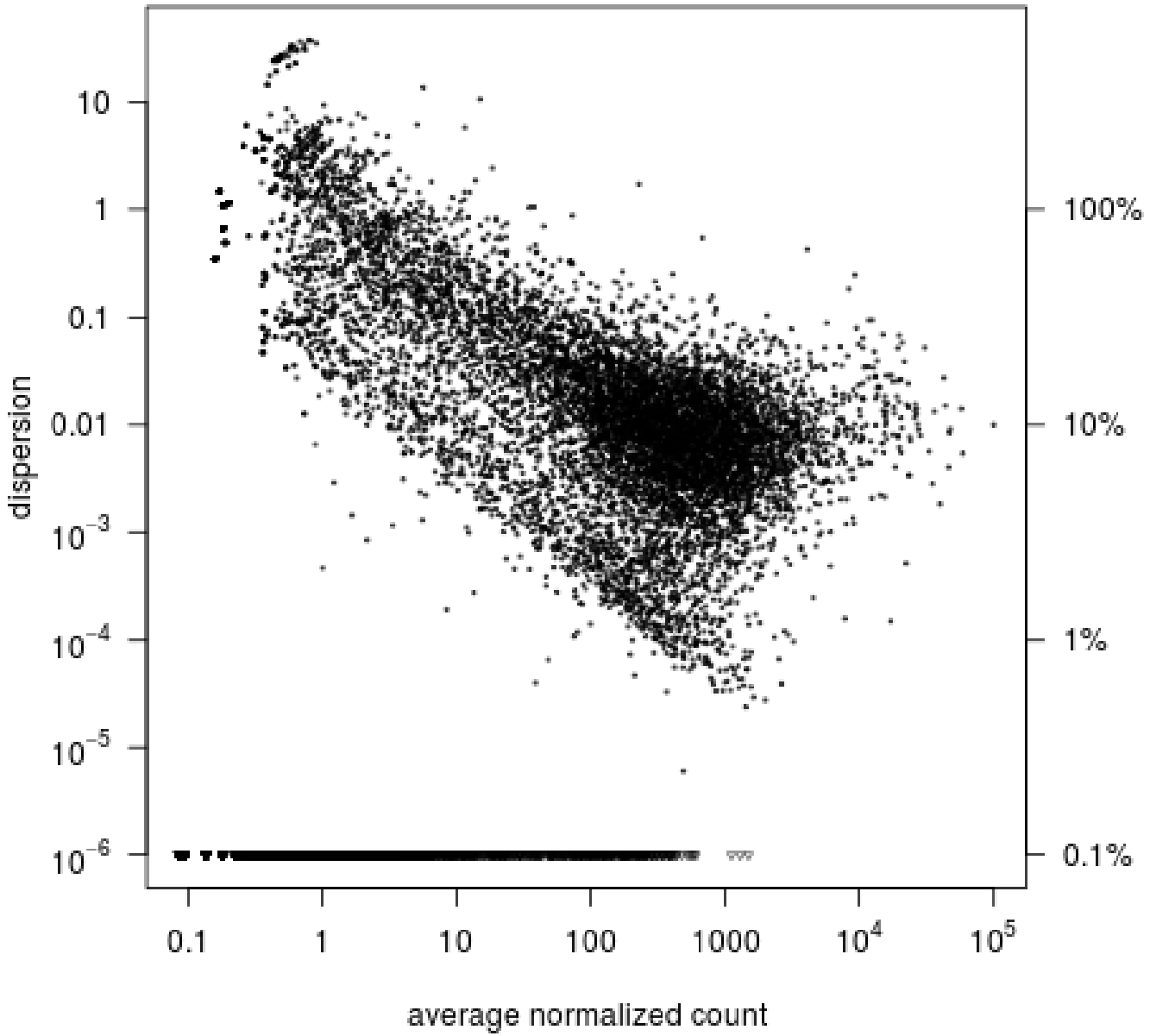


Dispersion

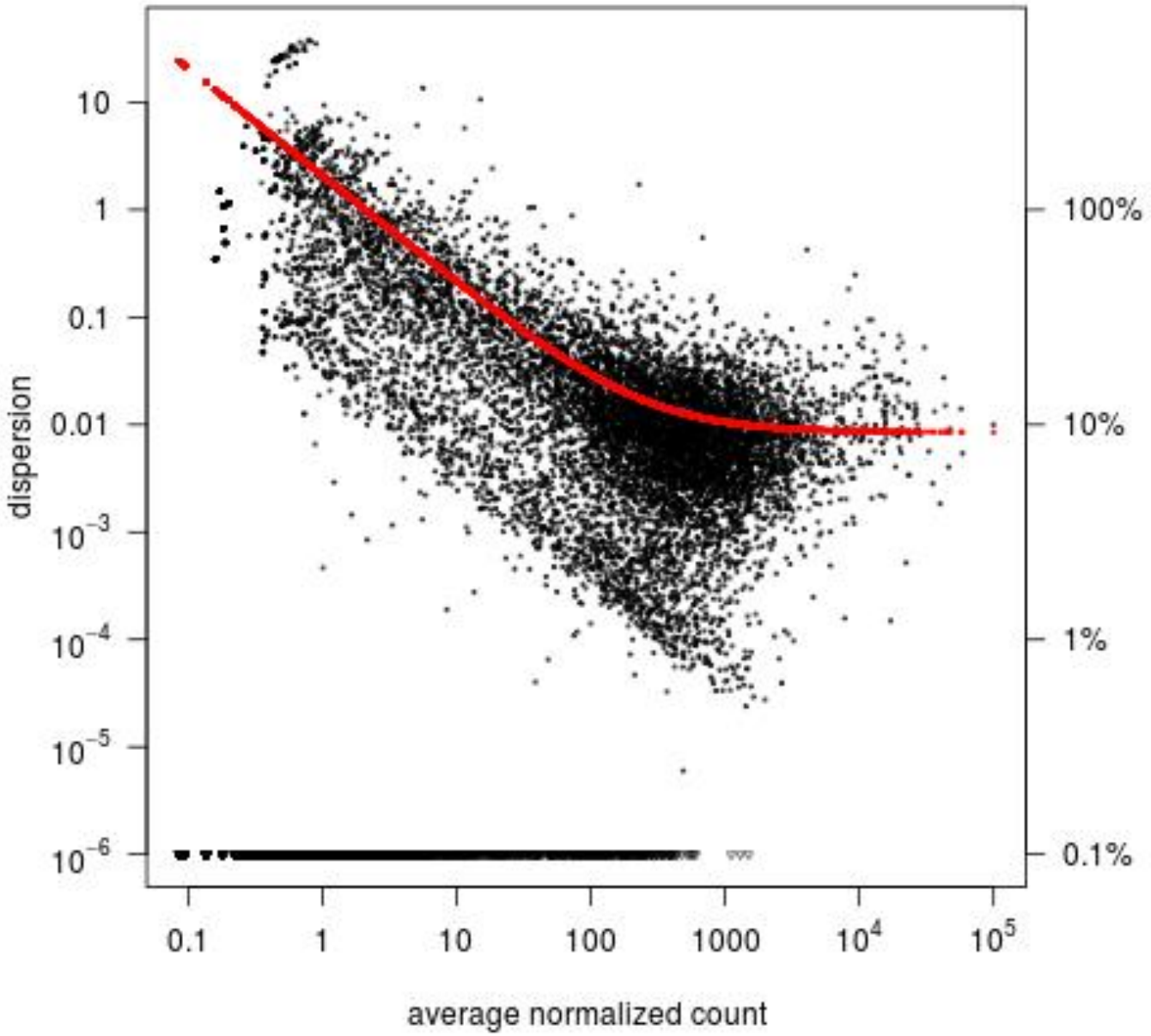
- quantifies within-group variability
- reliable estimation is crucial
- hard to estimate from few samples

➢ Use empirical-Bayes shrinkage estimation

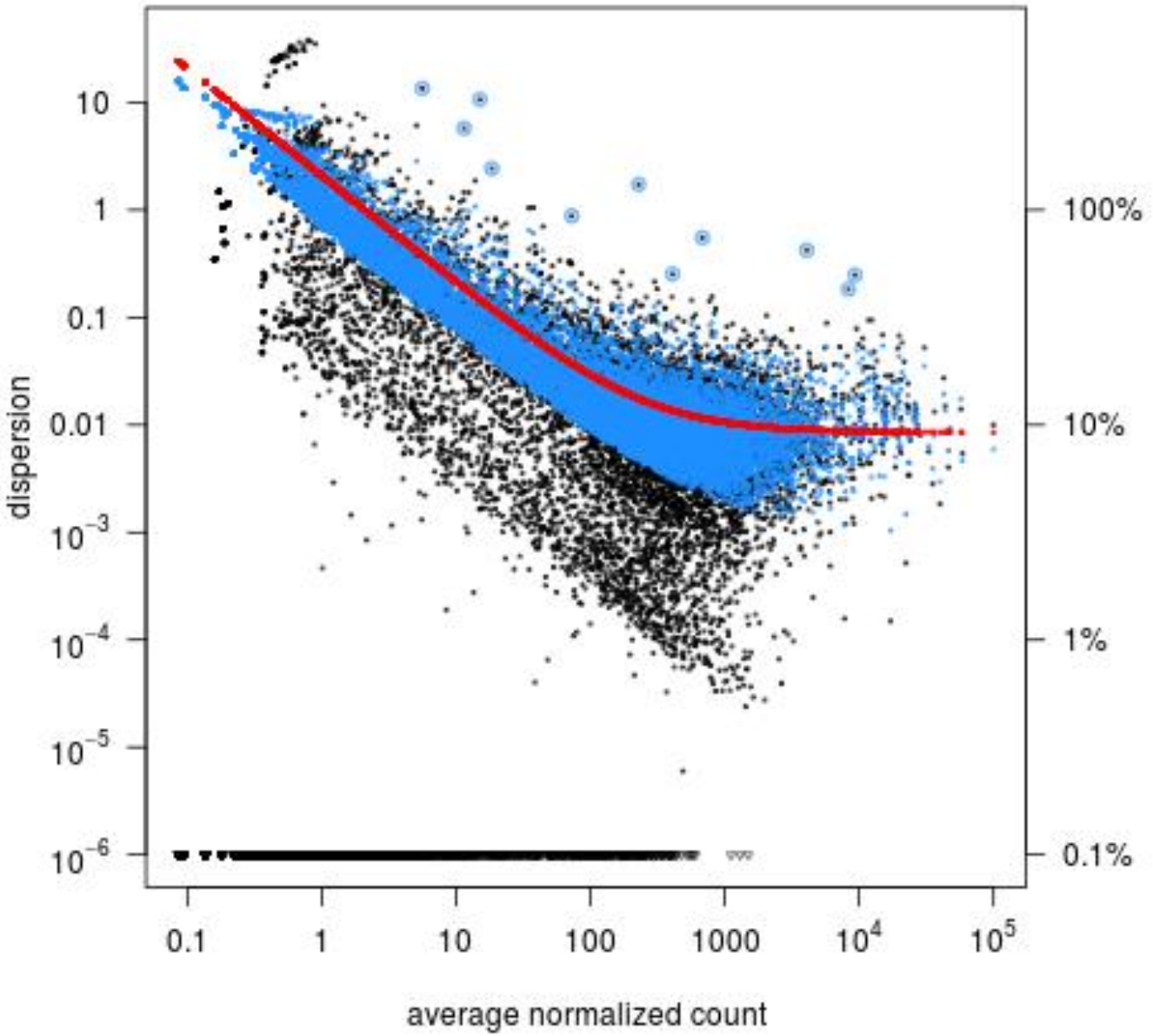
Shrinkage estimation of dispersion (within-group variability)



Shrinkage estimation of dispersion (within-group variability)



Shrinkage estimation of dispersion (within-group variability)



More things to do with shrinkage:

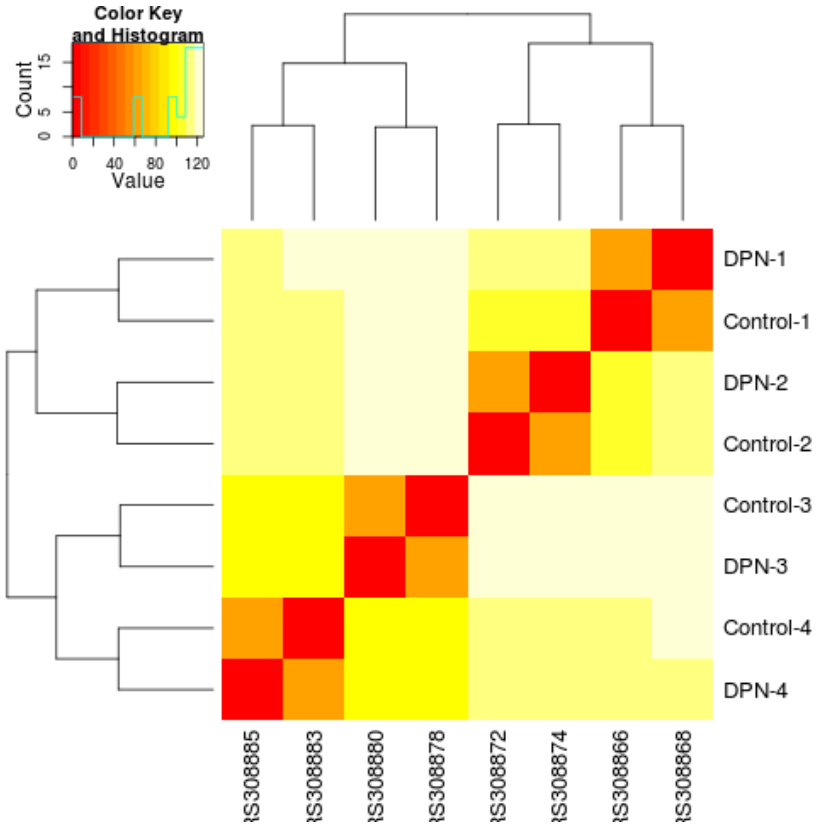
The rlog transformation

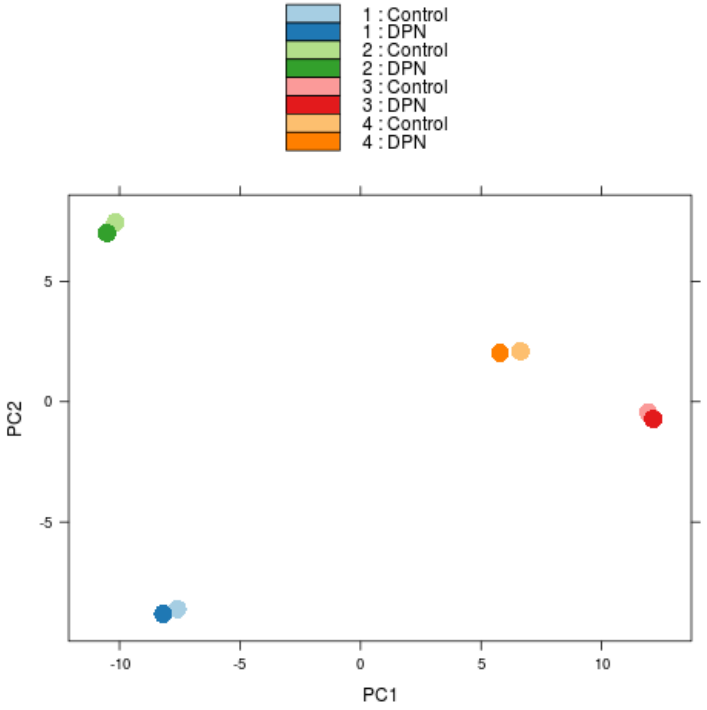
Many useful methods want homoscedastic data:

- Hierarchical clustering
- PCA and MDS

But: RNA-Seq data is not homoscedastic.

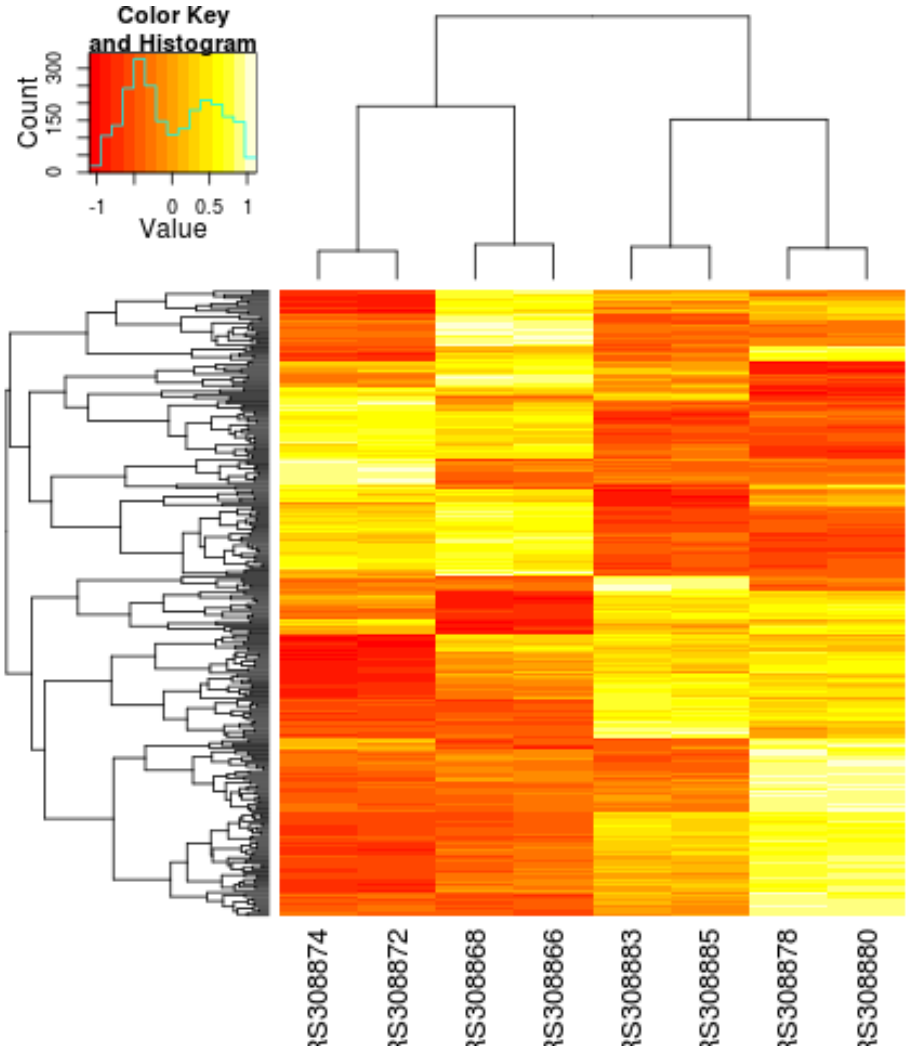
Visualization of rlog-transformed data: Sample clustering and PCA





Data: Parathyroid samples from Haglung et al., 2012

Visualizationof rlog-transformed data: Gene clustering



More things to do with shrinkage:

The rlog transformation

RNA-Seq data is not homoscedastic.

- On the count scale, large counts have large (absolute) variance.

- After taking the logarithm, small counts show excessive variance.

More things to do with shrinkage:

The rlog transformation

Conceptual idea of the rlog transform:

Log-transform the average across samples of each gene's normalized count.

The "pull in" the log normalized counts towards the log averages. Pull more for weaker genes.

More things to do with shrinkage:

The rlog transformation

Procedure:

- Fit log-link GLM with intercept for average and one coefficient per sample.
- Estimate empirical-Bayes prior from sample coefficients.
- Fit again, now with ridge penalty from EB prior.
- Return fitted linear predictors.

Summary: Effect-size shrinkage

A simple method that makes many things easier, including:

- visualizing and interpreting effect sizes
- ranking genes
- performing GSEA
- performing clustering and ordination analyses

Complex designs

Generalized linear models

- read count for gene i in sample j:

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

- expected expression from linear model

$$\log \mu_{ij} = \beta_{i0} + \beta_{i1} x_{j1} + \beta_{i2} x_{j2}$$

  with design-matrix elements $x_{j.}$ and to-be-determined coefficients $\beta_{i.}$.

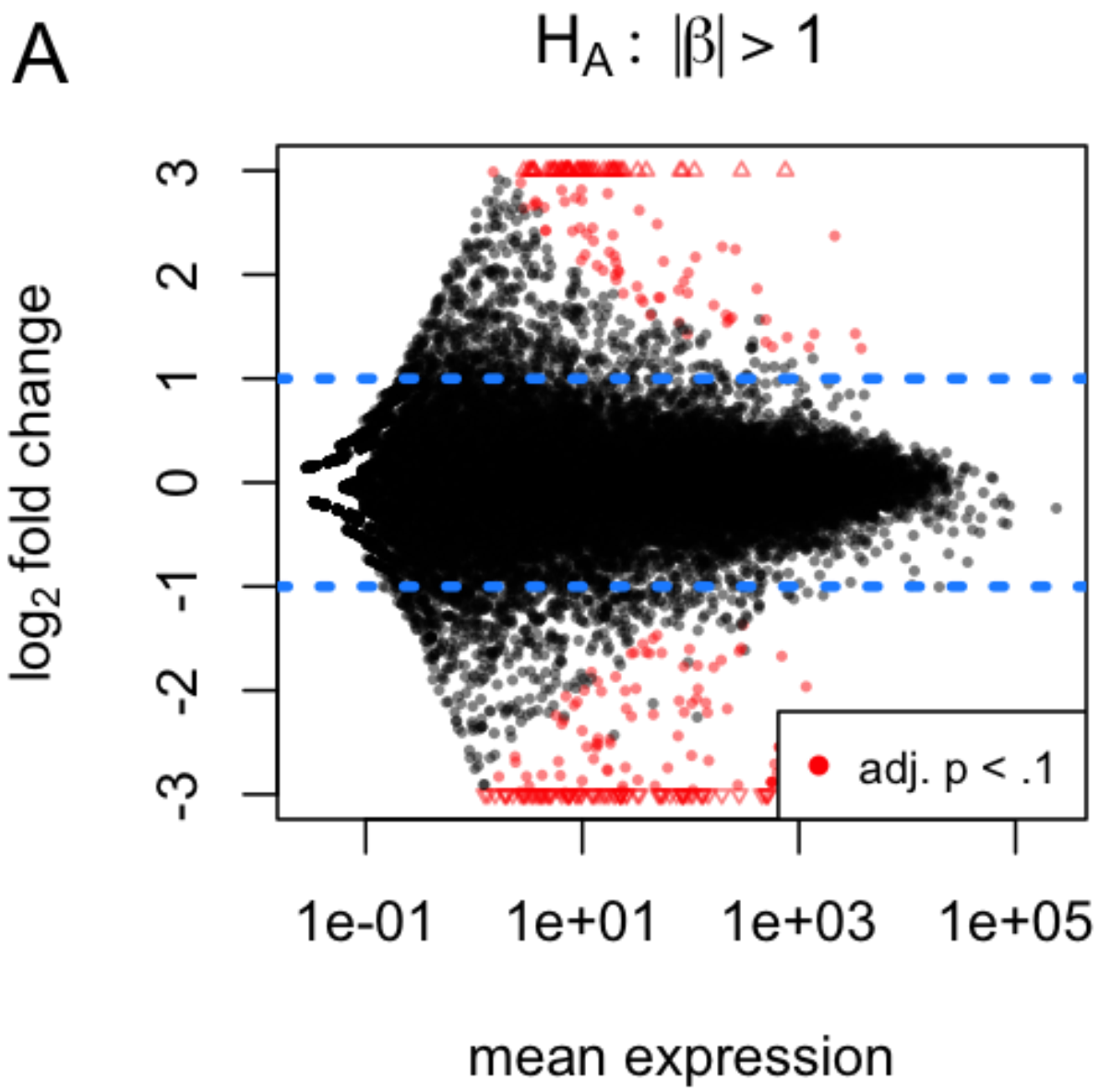
- dispersion α_i.

DESeq2 is not only for RNA-Seq

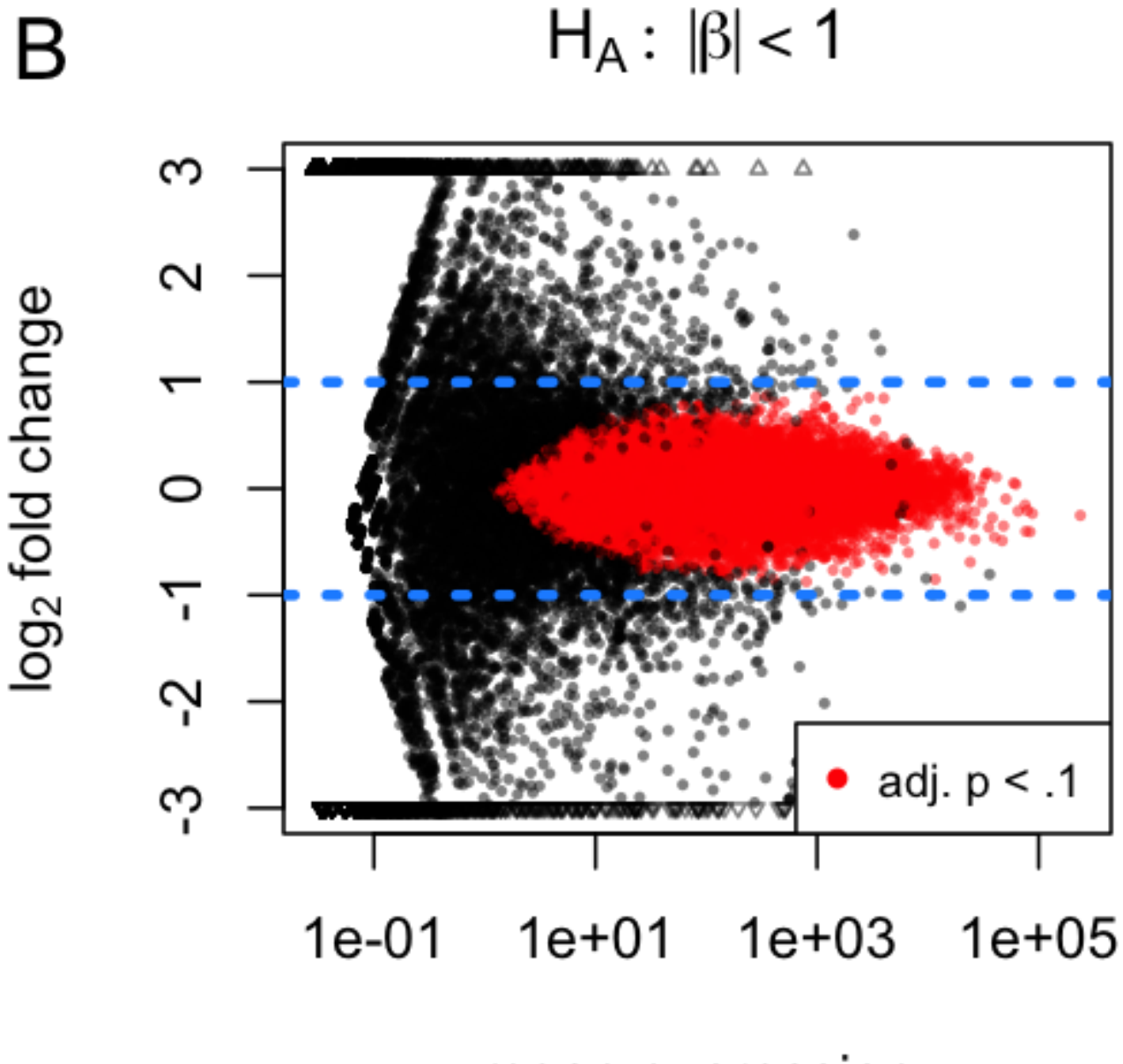
- RNA-Seq	1000+ papers
- ChIP-Seq	Ross-Ines et al., Nature, 2012 Avangani et al., Nature, 2014
- barcode-based assays	e.g., Robinson, G3, 2013
- metagenomics data	McMurdie et al., PLoS Comp Biol , 2014
- ribosome profiling	Vasquez et al., Nucl Acids Res, 2014
- shRNA and CRISPR/Cas9 screen	Zhou et al., Nature, 2014

What does
“differentially expressed”
actually mean?

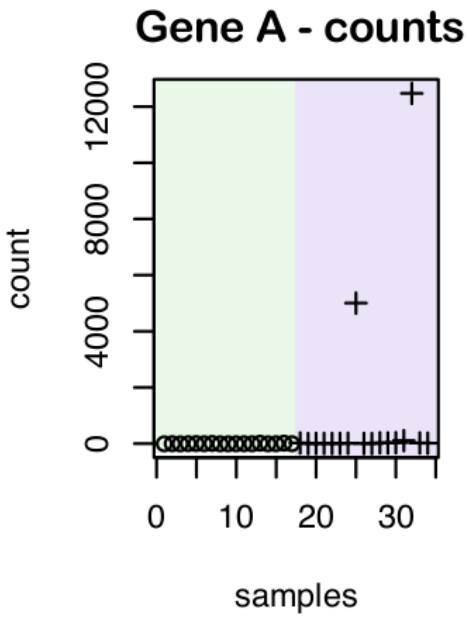
Genes changing *significantly* more than 2-fold:

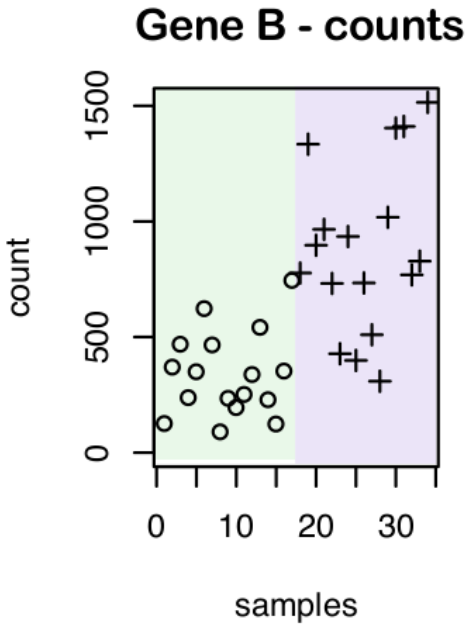


Genes changing *significantly* less than 2-fold:

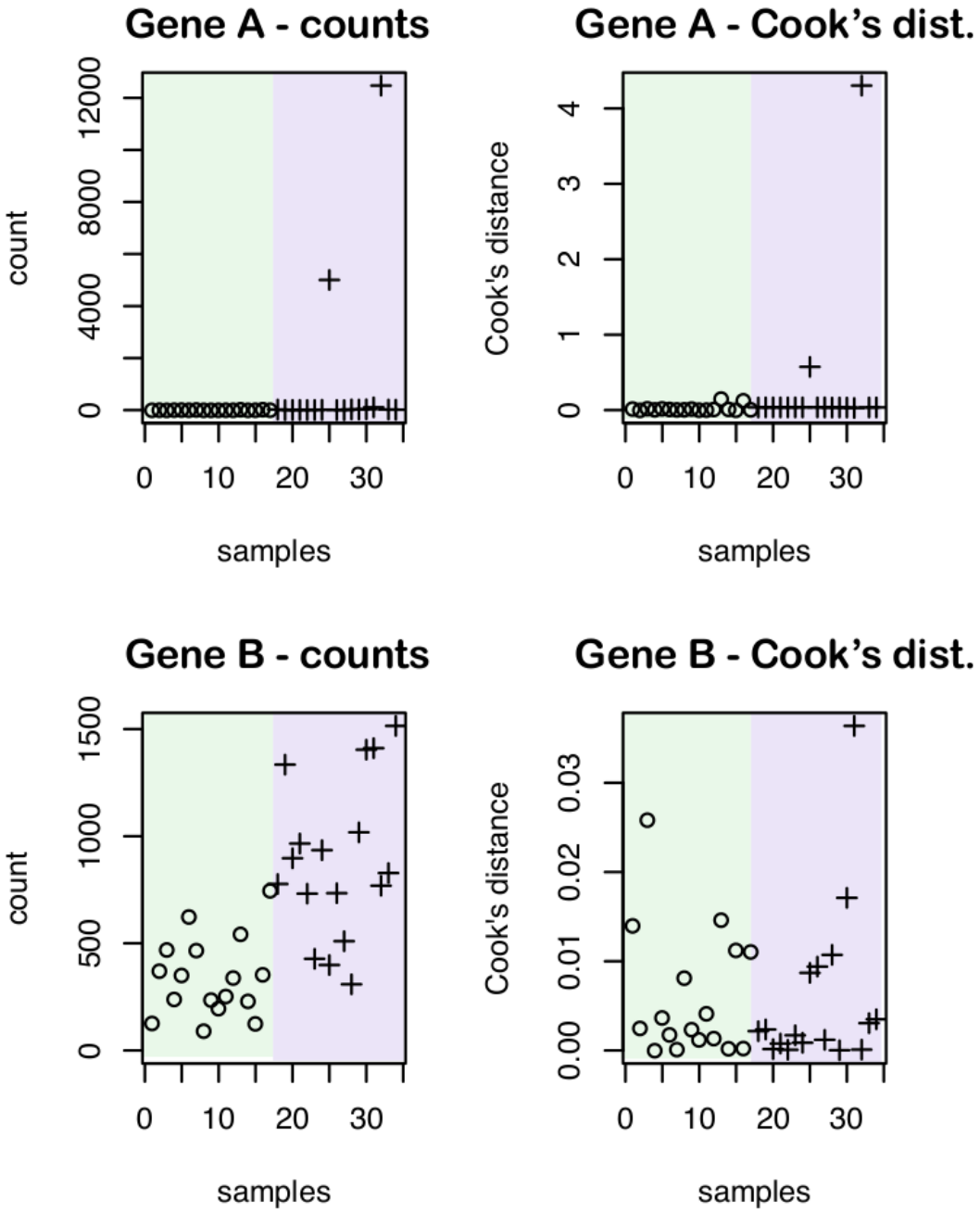


Outlier robustness



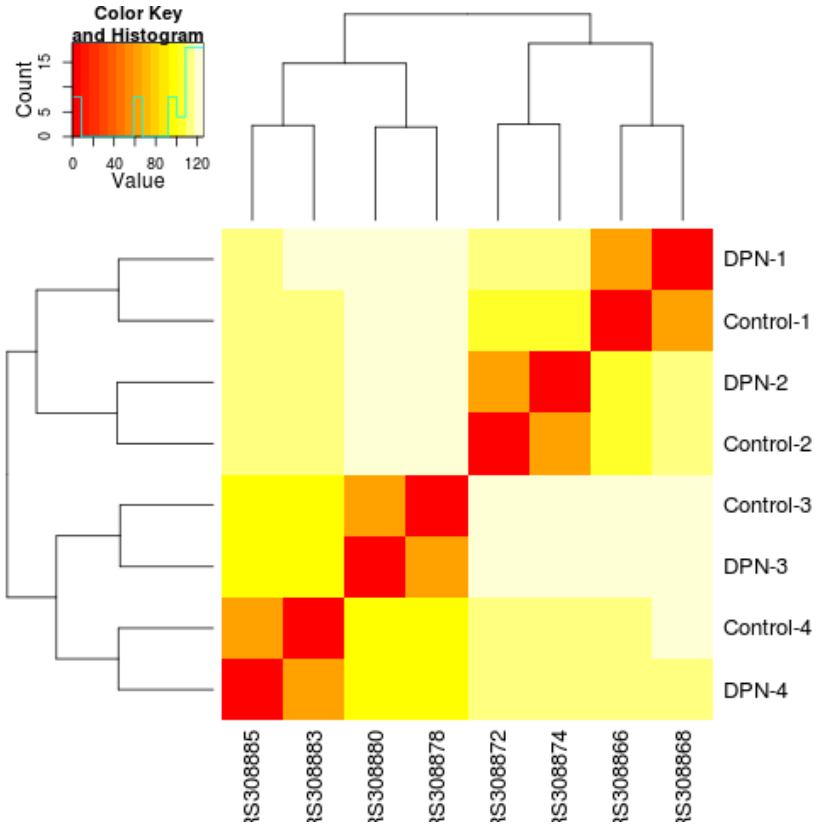


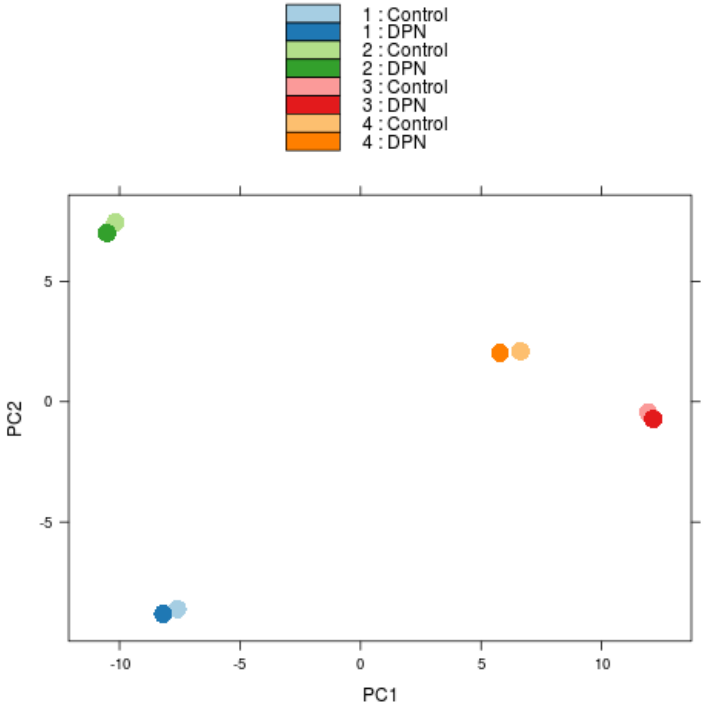
Outlier robustness



The “regularized log”(“rlog”) transformation

renders RNA-Seq data suitable for clustering and ordination methods





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

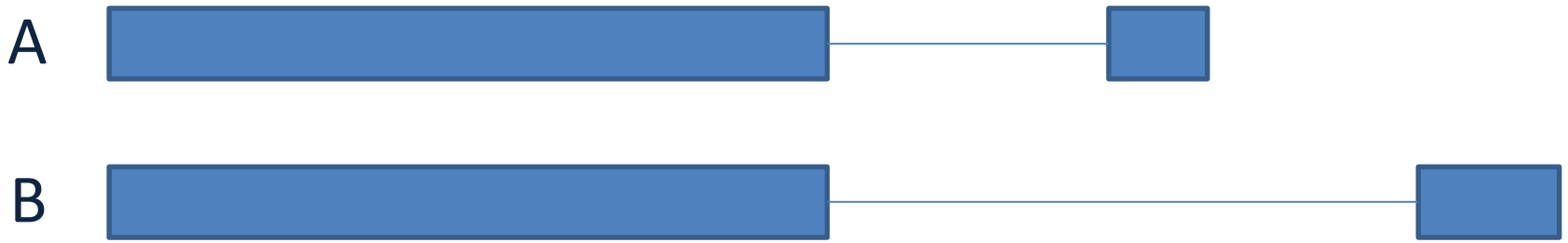
- So far, we looked at read counts *per gene*.

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

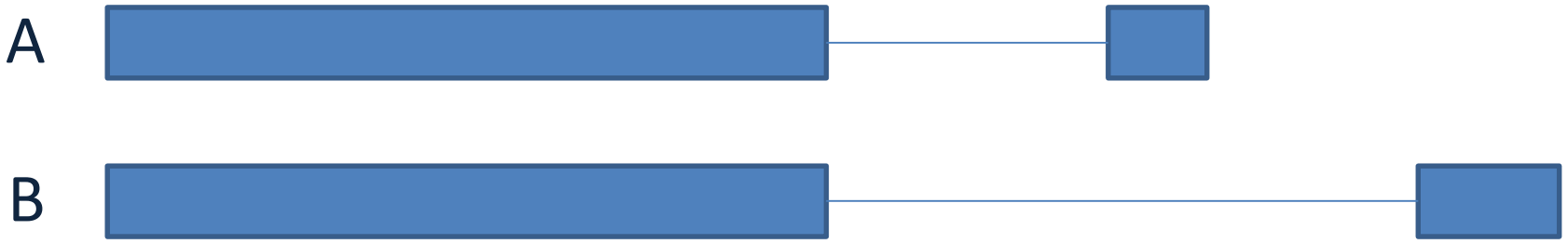
10 reads

from A

30 reads

from B

Assigning reads to transcripts



200 reads

(50 from A,
150 from B?)

5 reads

from A

15 reads

from B

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

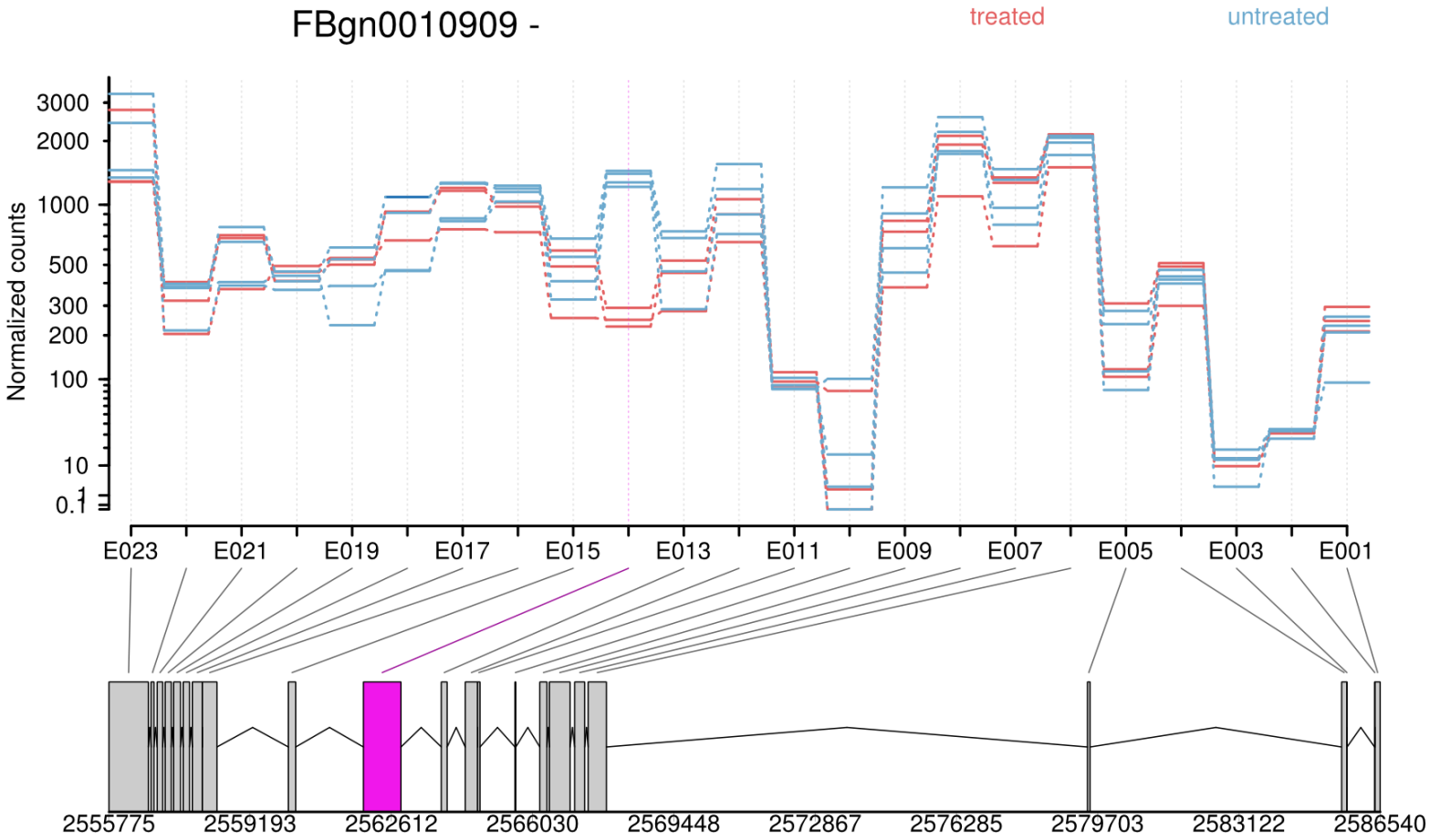
One step back:
Differential exon usage

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

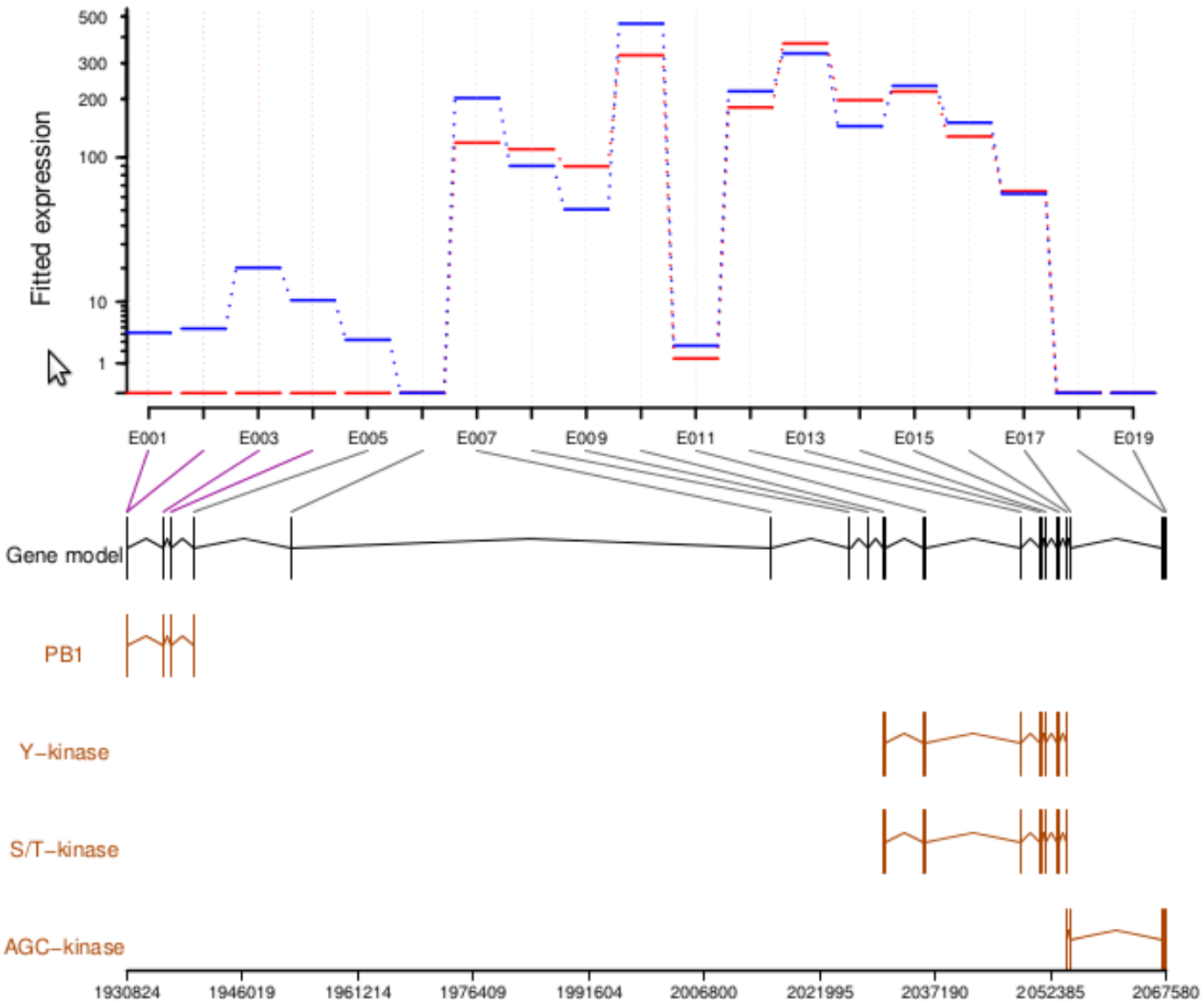
Usage on an exon =

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

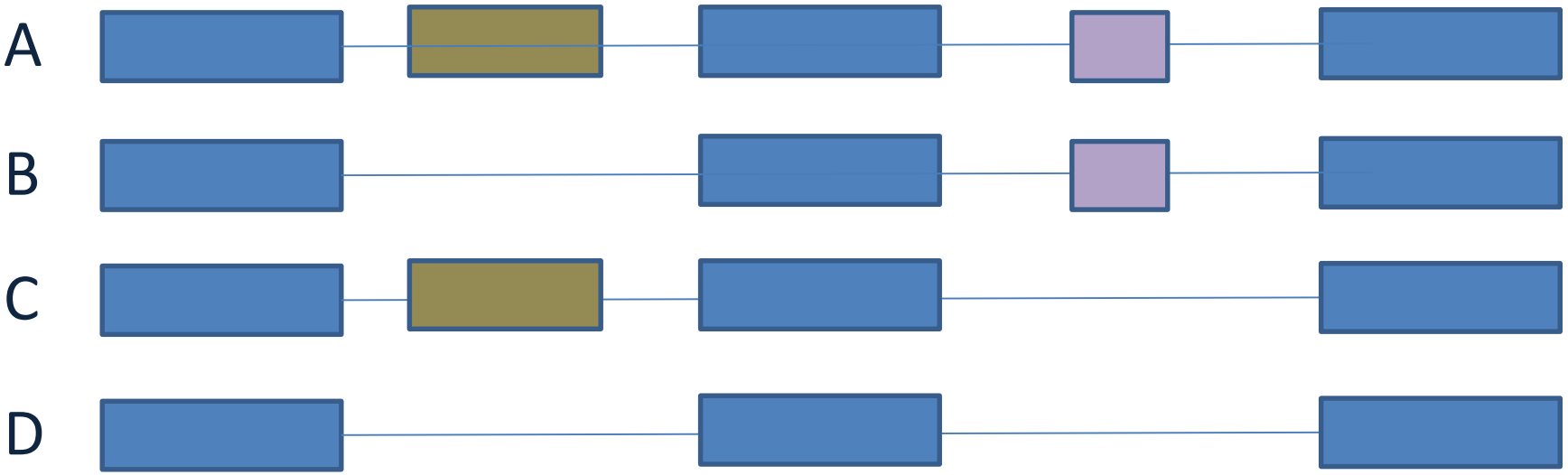
Differential exon usage -- Example



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)
- This helps with interpretation, visualization, clustering, ordination, etc.

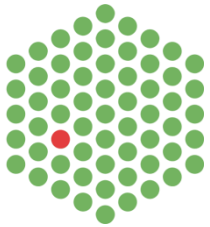
Acknowledgments

- Michael Love
- Alejandro Reyes
- Wolfgang Huber

Thanks also to

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

Funding:



EMBL

European Union:
FP7-health Project *Radiant*

Replication at what level?

- Prepare several libraries from the same sample (**technical replicates**).
  → controls for measurement accuracy
  → allows conclusions about just this sample

Replication at what level?

- Prepare several samples from the same cell-line (**biological replicates**).

  → controls for measurement accuracy *and* variations in environment an the cells' response to them.

  → allows for conclusions about the specific cell line

Replication at what level?

- Derive samples from different individuals (**independent samples**).

  → controls for measurement accuracy, variations in environment *and* variations in genotype.

  → allows for conclusions about the species

How much replication?

Two replicates permit to

- globally estimate variation

Sufficiently many replicates permit to

- estimate variation for each gene
- randomize out unknown covariates
- spot outliers
- improve precision of expression and fold-change estimates