Comparative analysis of RNA-Seq data with DESeq and DEXSeq

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

Alignment

Should one align to the genome or the transcriptome?

to transcriptome

- easier, because no gapped alignment necessary

  (but: splice-aware aligners are mature by now)

but:

- risk to miss possible alignments!

  (transcription is more pervasive than annotation claims)

→ Alignment to genome preferred.

Count data in HTS

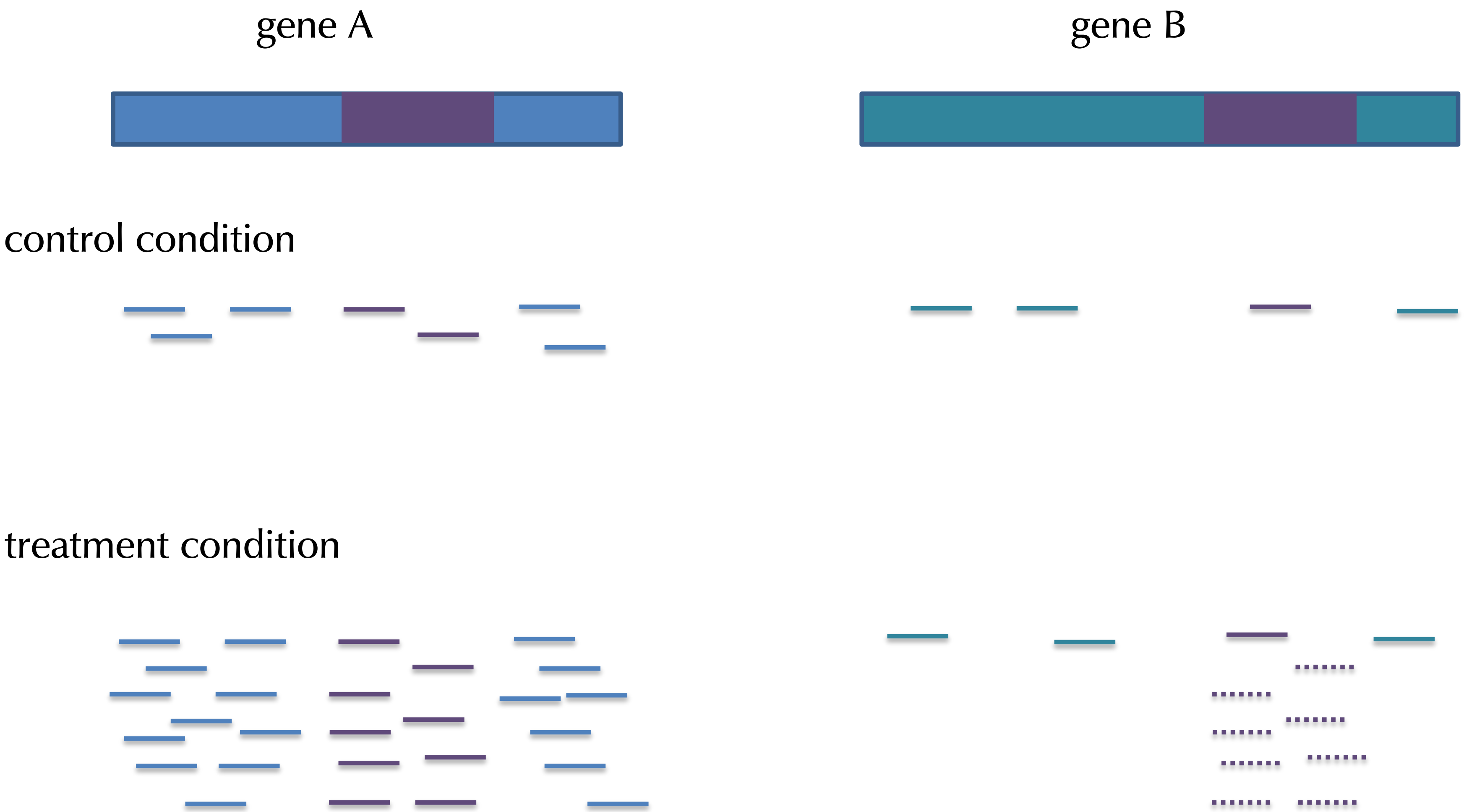
```
            control-1  control-2  control-3   treated-1   treated-2
FBgn0000008        78         46         43          47          89
FBgn0000014         2          0          0           0           0
FBgn0000015         1          0          1           0           1
FBgn0000017      3187       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

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.

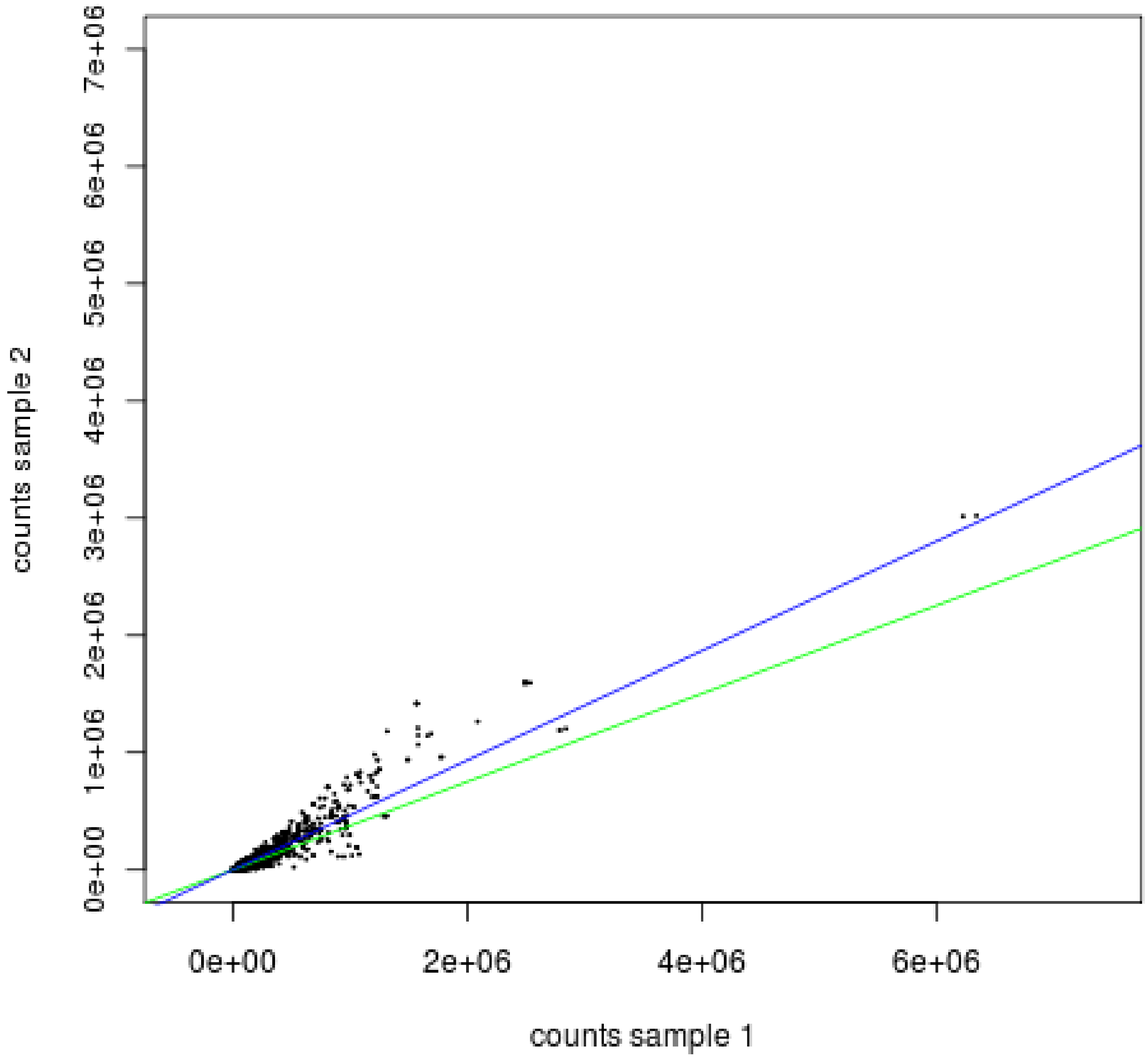
Normalization for library size

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

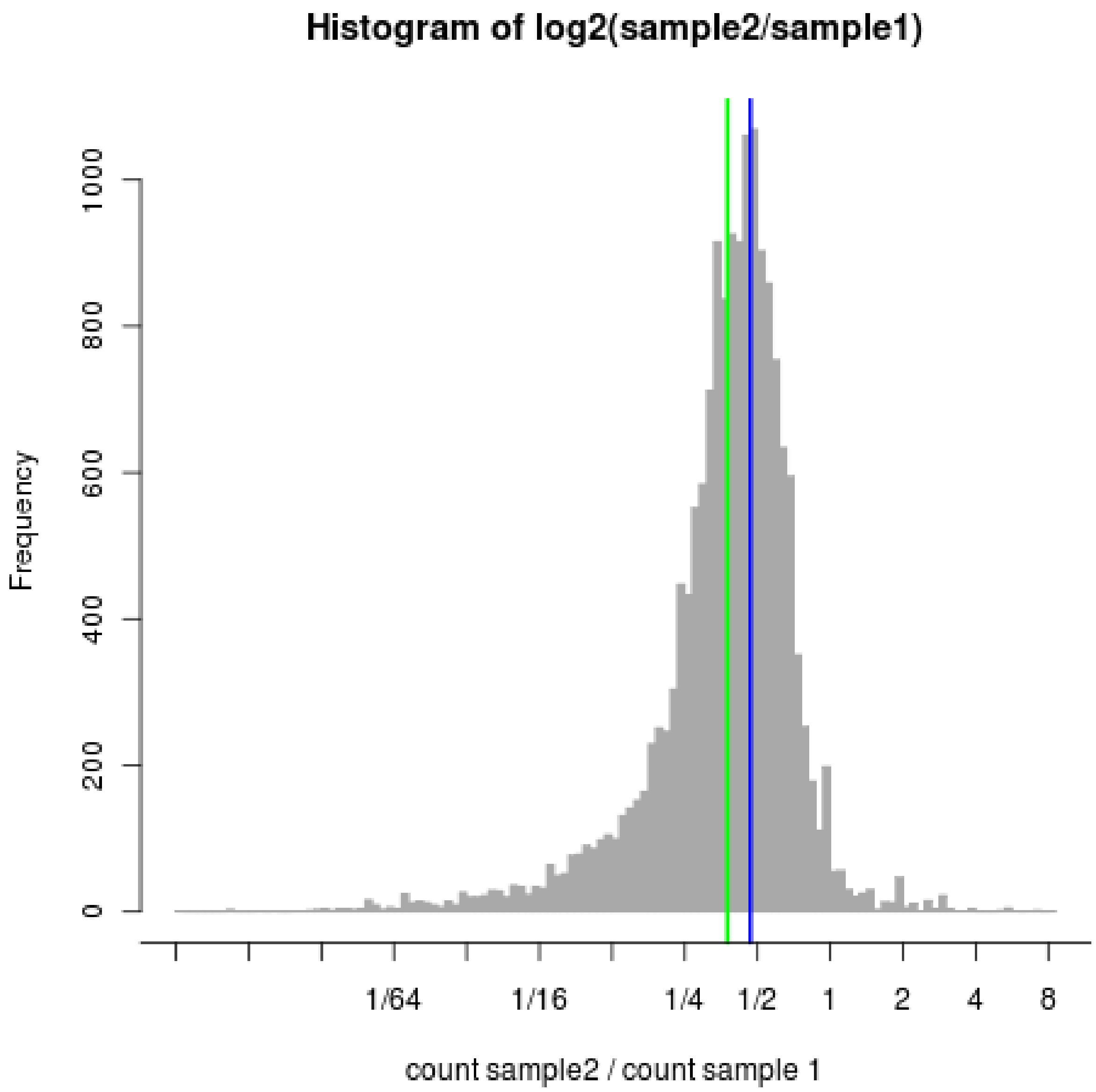
- Naive approach: Divide by the total number of reads per sample
- Problem: Genes that are strongly and differentially expressed may distort the ratio of total reads.

- By dividing, for each gene, the count from sample A by the count for sample B, we get one estimate per gene for the size ratio or sample A to sample B.
- We use the median of all these ratios.

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.

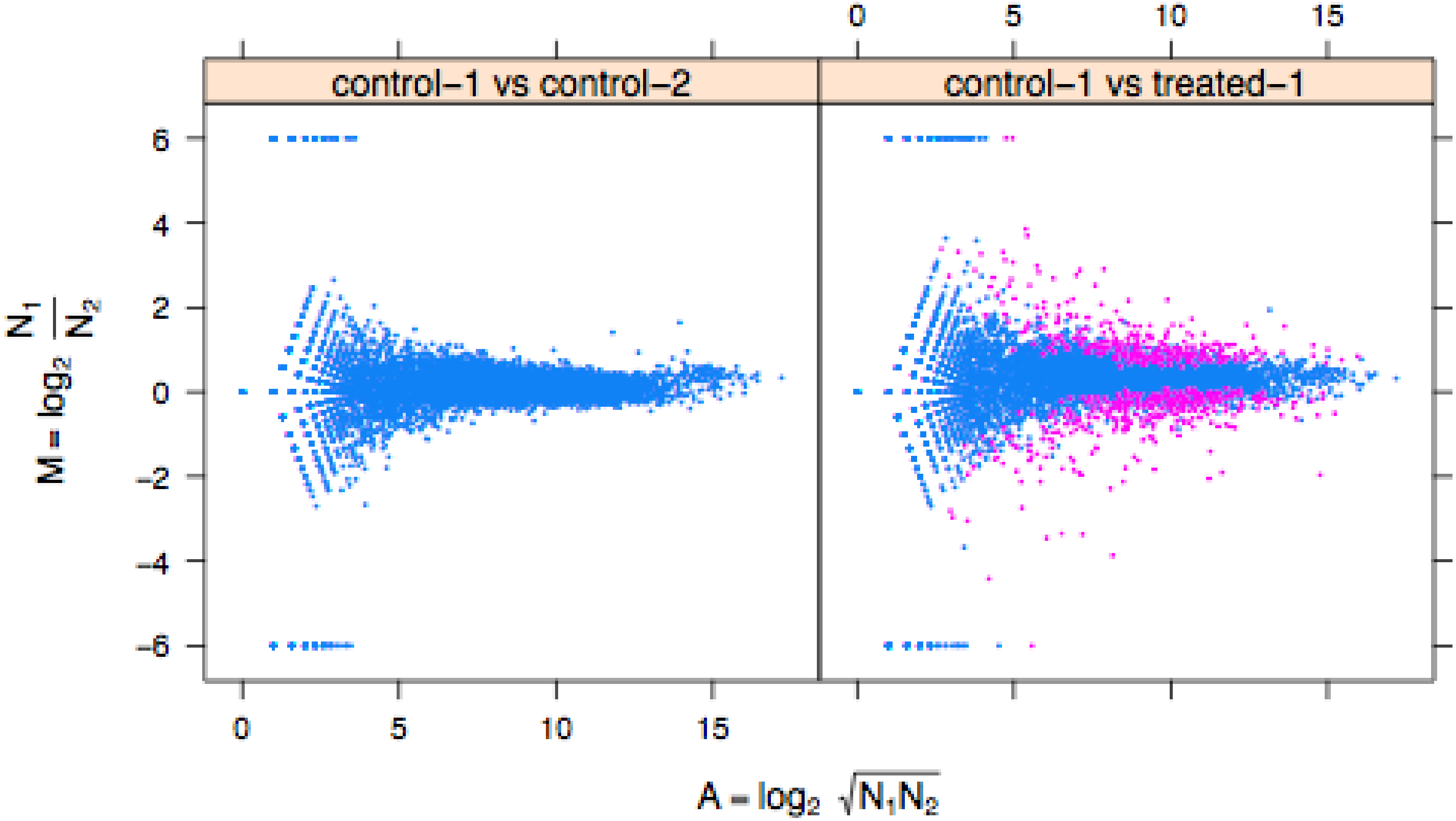
Anders and Huber, 2010

similar approach: Robinson and Oshlack, 2010

Sample-to-sample variation

comparison of
two replicates

comparison of
treatment vs control



Effect size and significance

Fundamental rule:

- We may attribute a change in expression to a treatment *only if* this change is large compared to the expected noise.

To estimate what noise to expect, we need to compare replicates to get a variance v.

If we have m replicates, the standard error of the mean is $\sqrt(v/m)$.

What do we mean by differential expression?

A treatment affects some gene, which in turn affect other genes.

In the end, all genes change, albeit maybe only slightly.

What do we mean by differential expression?

A treatment affects some gene, which in turn affect other genes.

In the end, all genes change, albeit maybe only slightly.

Potential stances:

- *Biological significance:* We are only interested in changes of a certain magnitude. (effect size > some threshold)
- *Statistical significance:* We want to be sure about the direction of the change. (effect size $\gg$ noise)

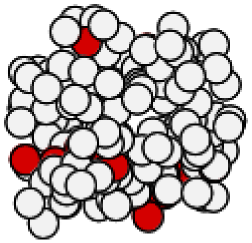
Counting noise

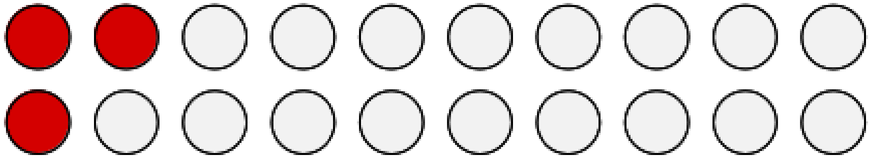
In RNA-Seq, noise (and hence power) depends on count level.

Why?

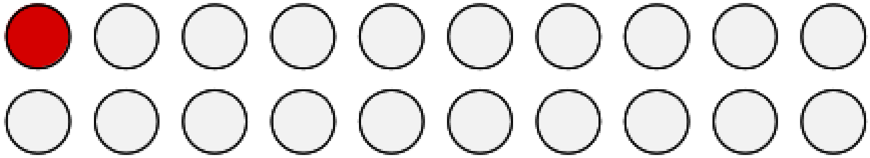
The Poisson distribution

- 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.

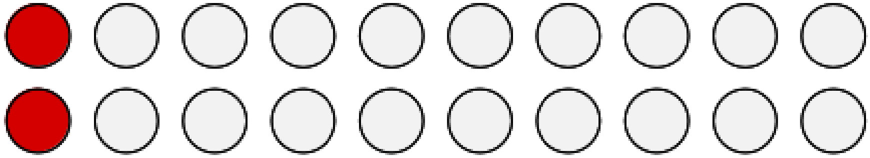




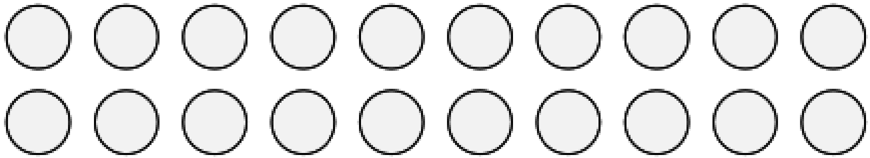
3 / 20 = 15%



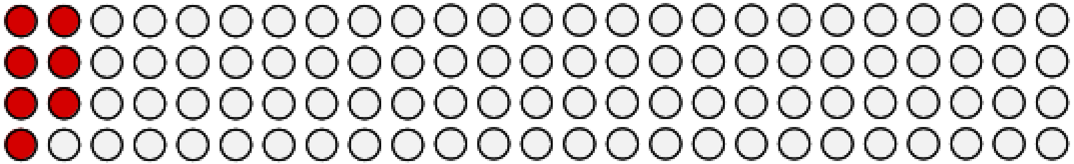
1 / 20 = 5%



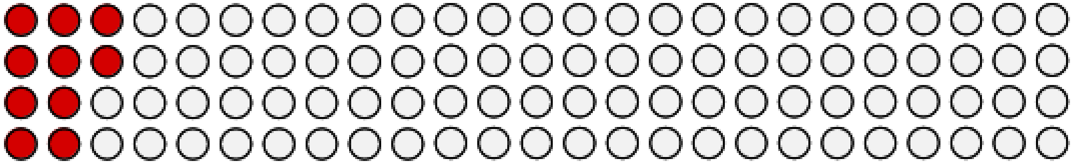
2 / 20 = 10%



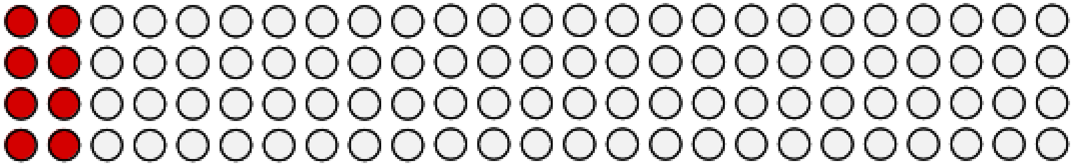
0 / 20 = 0%



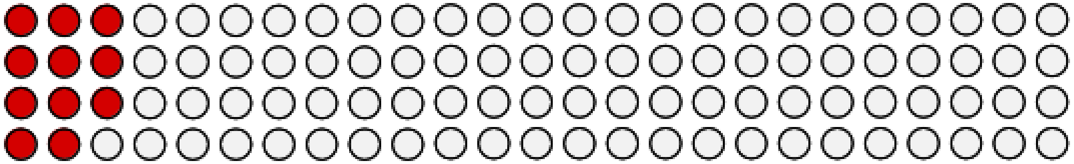
7 / 100 = 7%



10 / 100 = 10%



8 / 100 = 8%



11 / 100 = 11%

Poisson distribution

- If p is the proportion of red balls in the bag, and we draw n balls, we expect $\mu=pn$ balls to be red.
- The actual number k of red balls follows a *Poisson* distribution, and hence k varies around its expectation value μ with standard deviation $\sqrt{\mu}$.

Poisson distribution

- If p is the proportion of red balls in the bag, and we draw n balls, we expect $\mu = pn$ balls to be red.
- The actual number k of red balls follows a *Poisson* distribution, and hence k varies around its expectation value μ with standard deviation $\sqrt{\mu}$.
- Our estimate of the proportion $p = k/n$ hence has the expected value $\mu/n = p$ and the standard error
- $\Delta p = \sqrt{\mu}/n = p/\sqrt{\mu}$. The relative error is $\Delta p/p = 1/\sqrt{\mu}$.

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	$\sqrt{10{,}000} = 100$	$1 / \sqrt{10000} = 1.0\%$

- For Poisson-distributed data, the variance is equal to the mean.

- Hence, no need to estimate the variance, according to many papers

Really?

Counting noise

- Consider this situation:
  - Several flow cell lanes are filled with aliquots of the *same* prepared library.
  - The concentration of a certain transcript species is *exactly* the same in each lane.
  - We get the same total number of reads from each lane.
- For each lane, count how often you see a read from the transcript. Will the count all be the same?

Shot noise

- Consider this situation:
  - Several flow cell lanes are filled with aliquots of the *same* prepared library.
  - The concentration of a certain transcript species is *exactly* the same in each lane.
  - We get the same total number of reads from each lane.
- For each lane, count how often you see a read from the transcript. Will the count all be the same?
- Of course not. Even for equal concentration, the counts will vary. This *theoretically unavoidable* noise is called *shot noise*.

Shot noise

- Shot noise: The variance in counts that persists even if everything is exactly equal. (Same as the evenly falling rain on the paving stones.)

- Stochastics tells us that shot noise follows a *Poisson distribution*.
- The standard deviation of shot noise can be *calculated*: it is equal to the square root of the average count.

Sample-to-sample noise

Now consider

- Several lanes contain samples from biological replicates.
- The concentration of a given transcript varies around a mean value with a certain standard deviation.
- This standard deviation cannot be calculated, it has to be *estimated* from the data.

Differential expression: Two questions

Assume you use RNA-Seq to determine the concentration of transcripts from some gene in different samples. What is your question?

- 1. “Is the concentration in one sample different from the expression in another sample?”

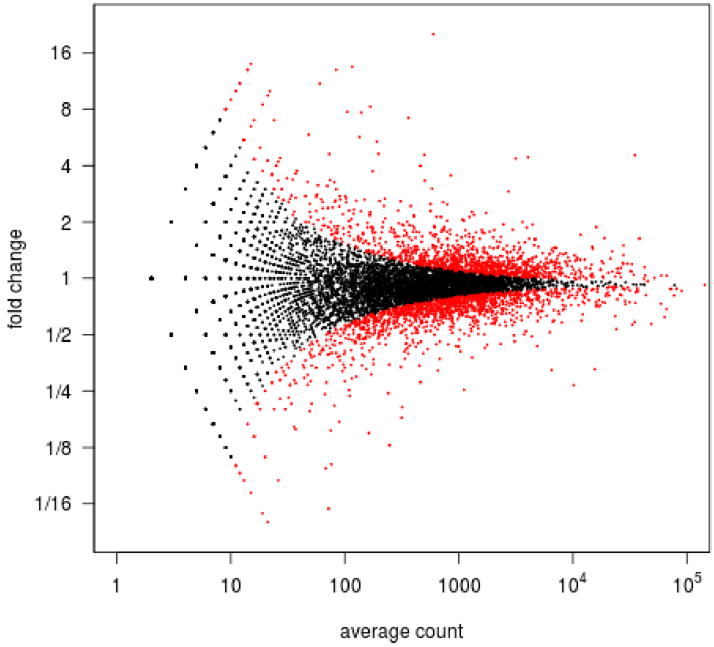
or

- 2. “Can the difference in concentration between treated samples and control samples be attributed to the treatment?”

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



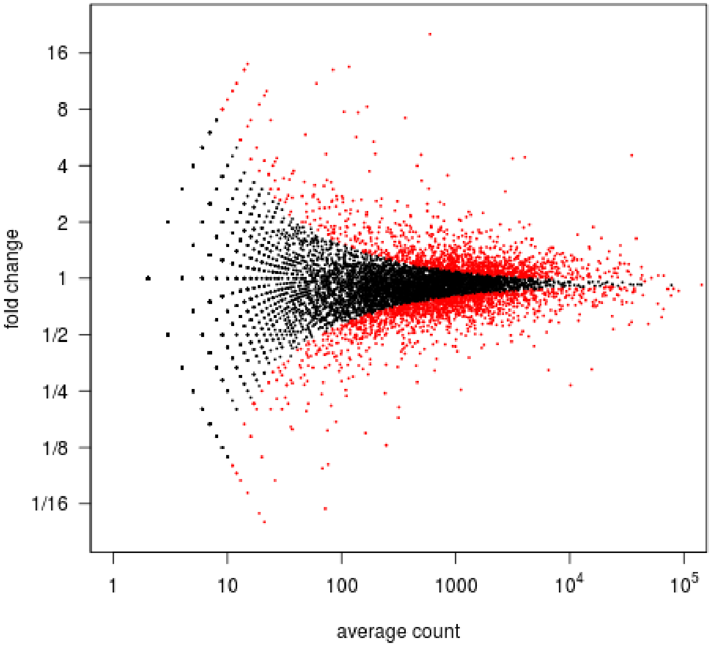
red: significant genes according to Fisher test (at 10% FDR)

Fisher's exact test between two samples

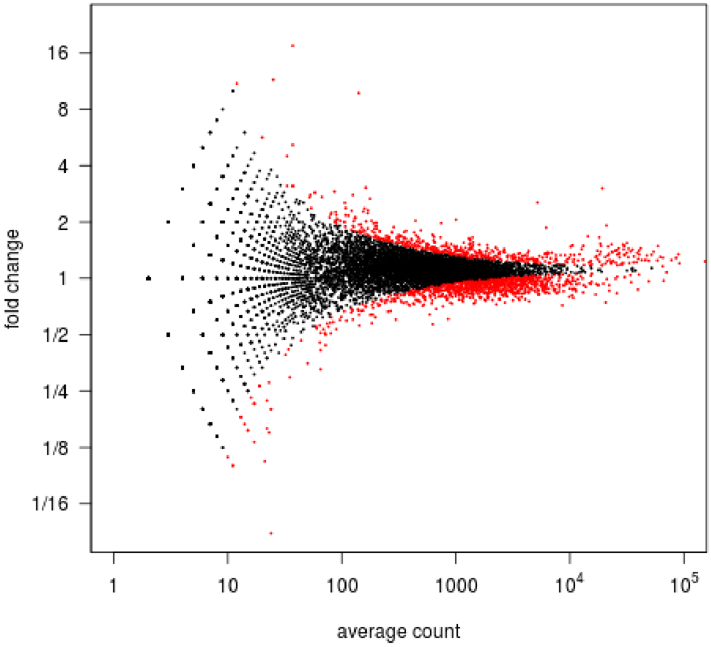
Example data: fly cell culture, knock-down of pasilla

(Brooks et al., Genome Res., 2011)

knock-down sample T2
versus
control sample U3



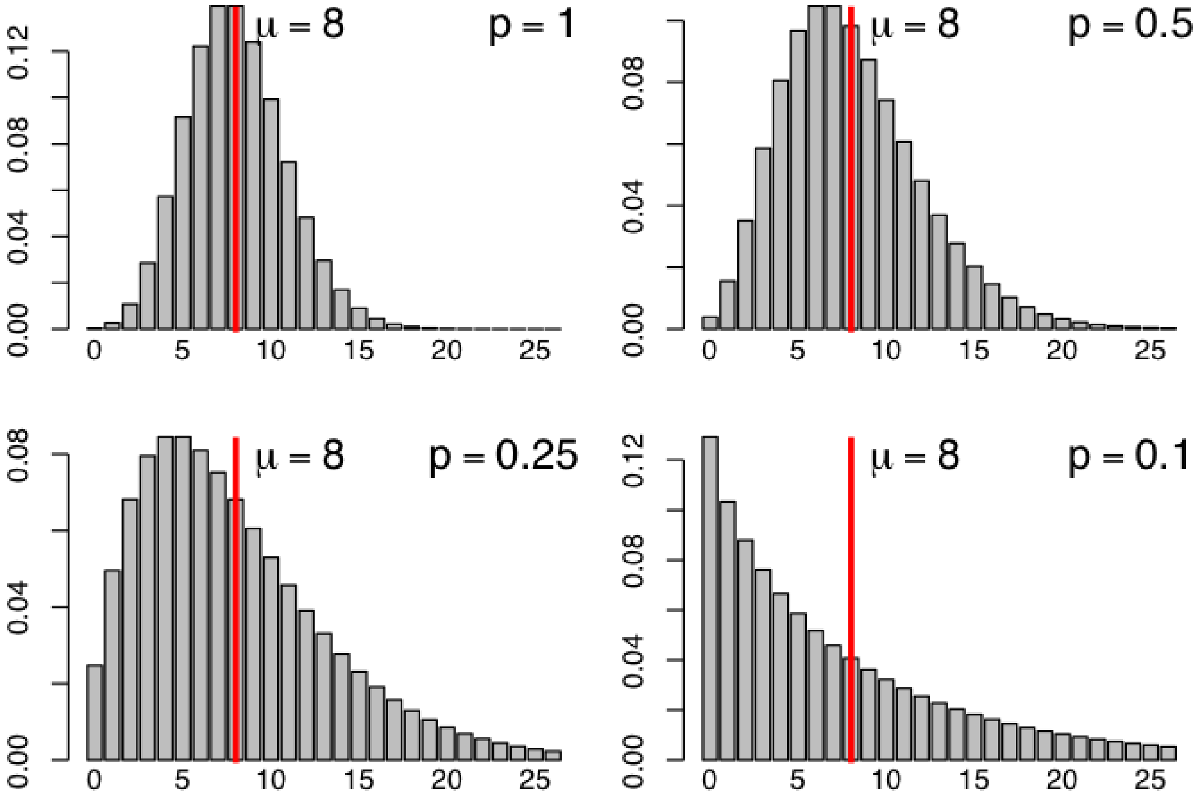
control sample U2
versus
control sample U3



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

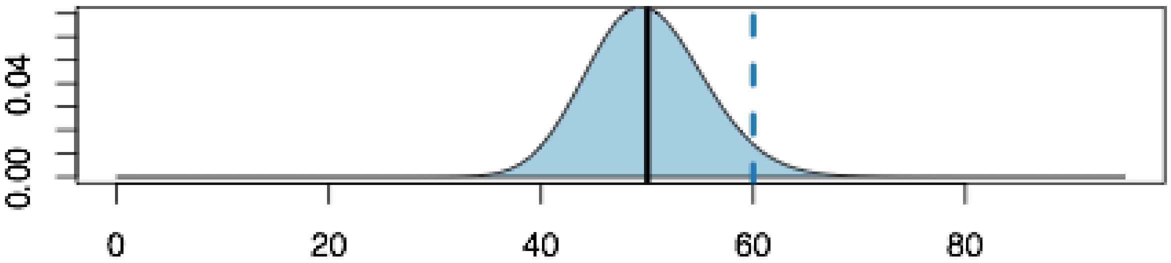
The negative binomial distribution

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

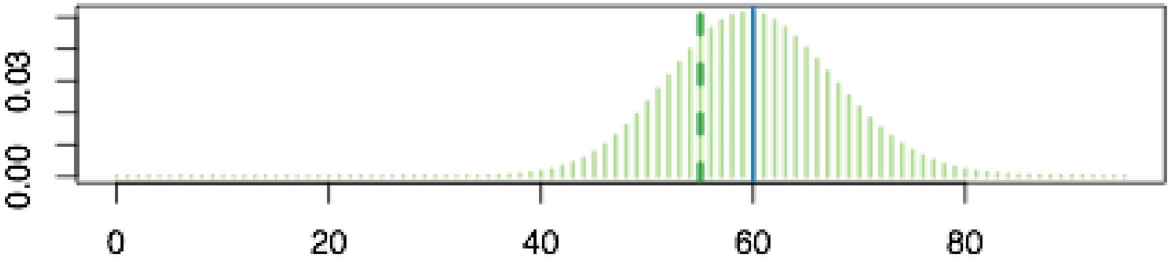


$$\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

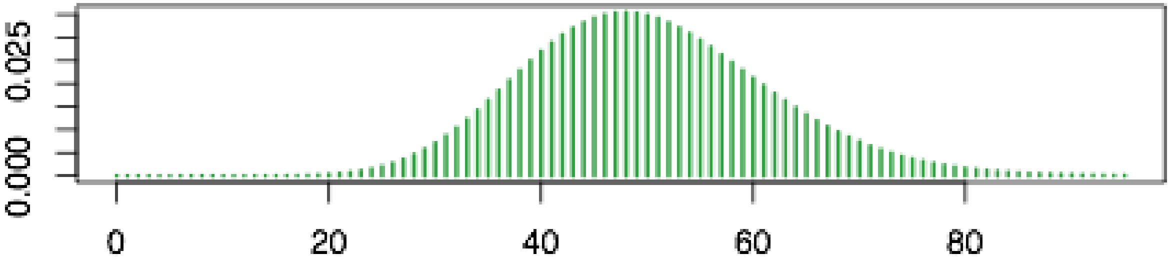
The NB from a hierarchical model



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.

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.

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

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

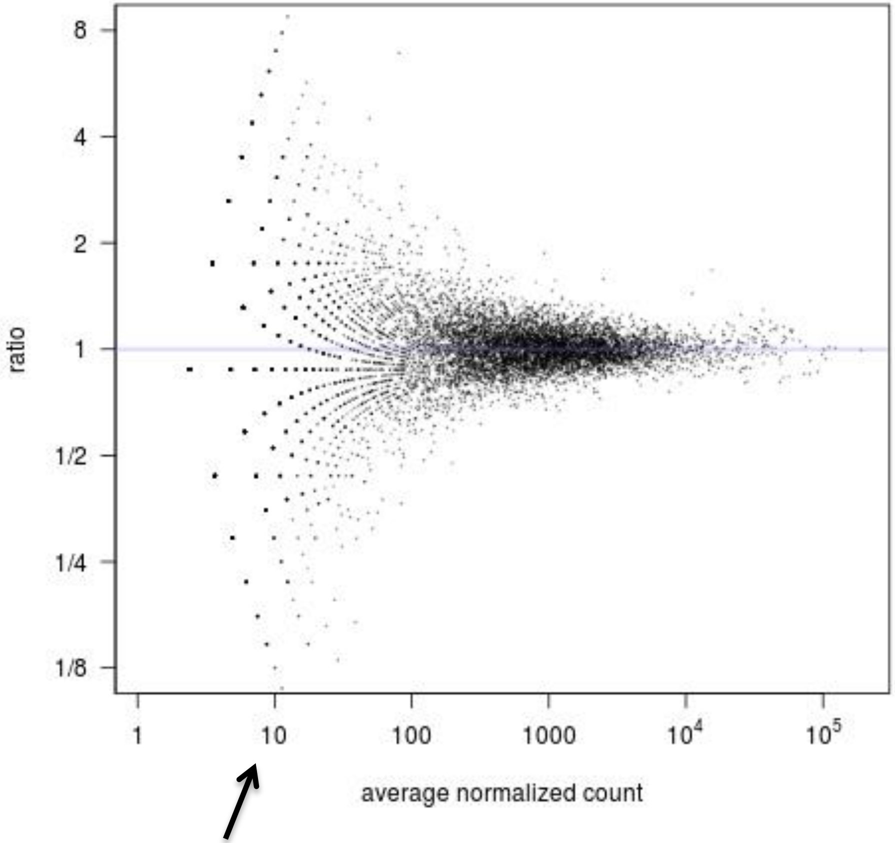
Example: A gene differs by 20% between samples within a group (CV=0.2)

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 variability



Comparison of normalized counts between two replicate samples

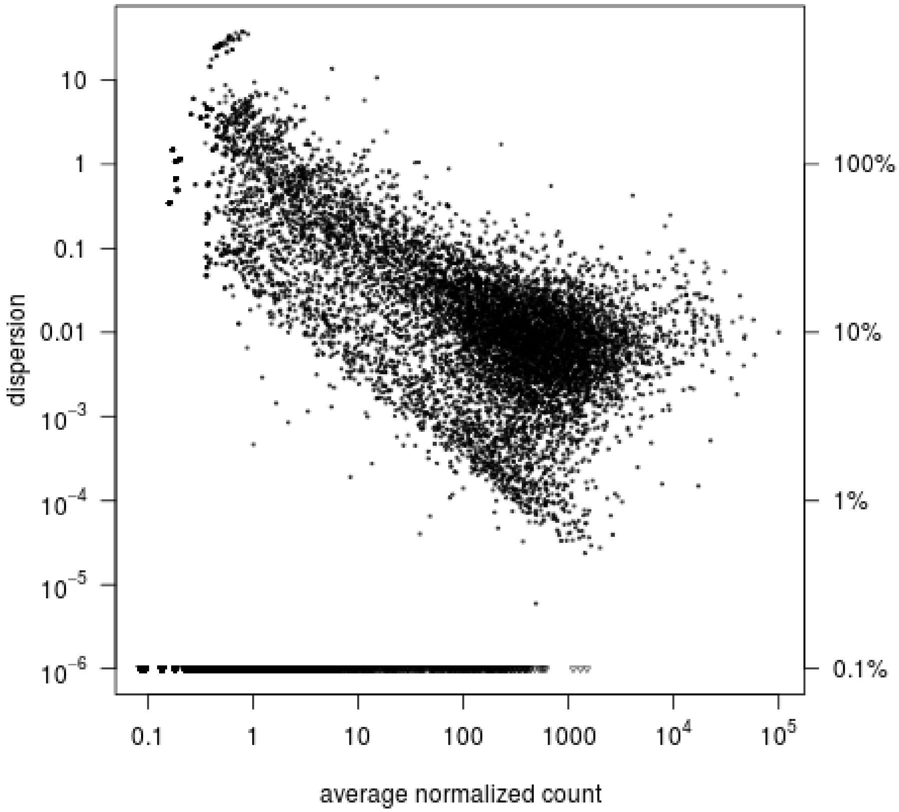
(Drosophila cell culture, treated with siRNA, data by Brooks et al., 2011)

Core assumption:
Genes of similar expression strength have similar sample-to-sample variance.

Under this assumption, we can estimate variance with more precision.

Baldi & Long (2001); Lönnsted & Speed (2002); Smyth (2004); Robinson, McCarthy & Smyth (2010); Wu et al (2013);...

Shrinkage estimation of variability



Dispersion

- Minimum variance of count data:

  $v = \mu$ (Poisson)

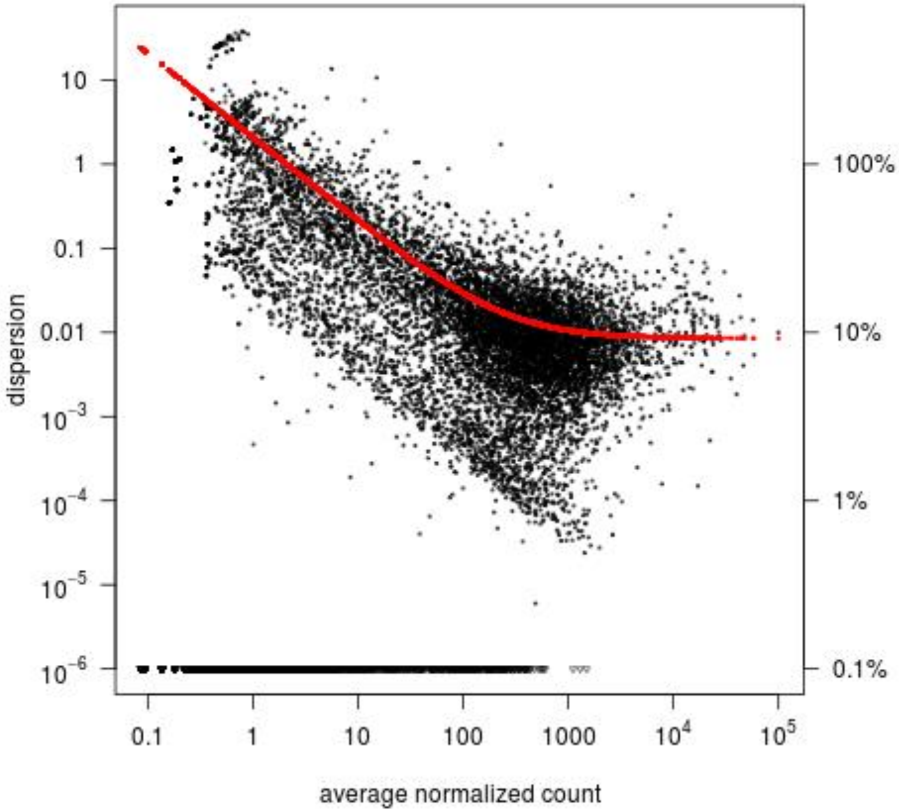
- Actual variance:

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

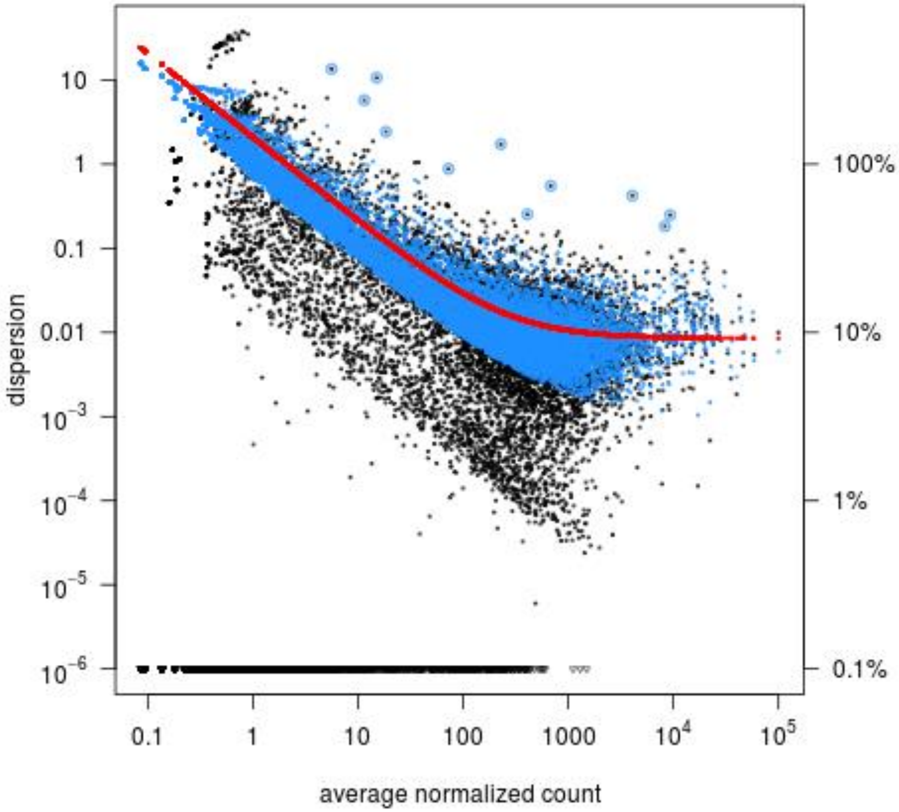
- α : “dispersion” $\alpha = (\mu - v) / \mu^2$

  (squared coefficient of variation of extra-Poisson variability)

Shrinkage estimation of variability



Shrinkage estimation of variability

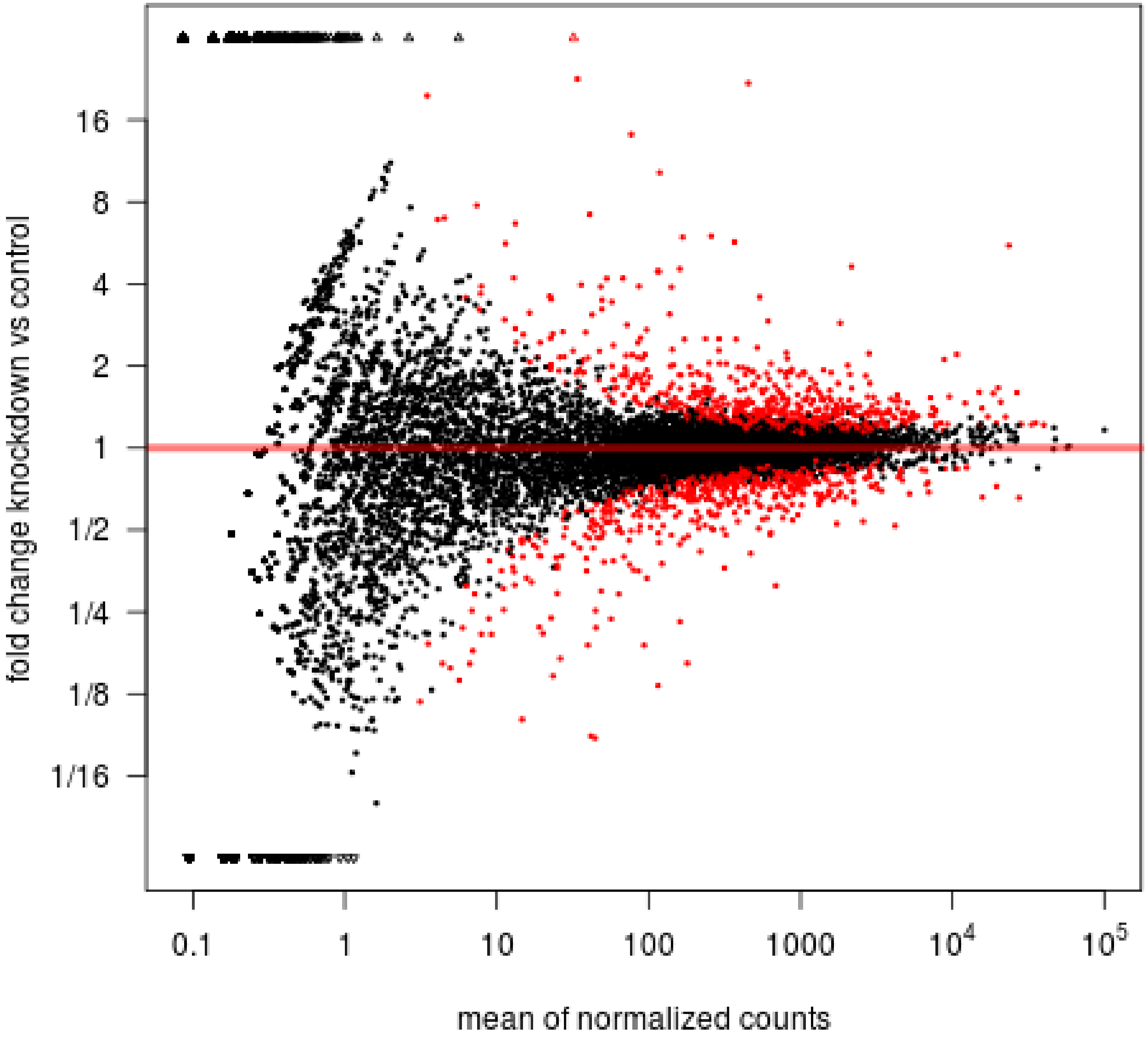


Dispersion shrinkage in DESeq2

- Estimate dispersion for each gene (using only that gene's count data)
- Fit dependence on mean.
- Fit log-normal empirical prior for true dispersion scatter around fitted values.
- Narrow prior to account for sampling width.
- Calculate maximum a-posteriori values as final dispersion estimates.
- Use raw values for high-dispersion outliers.

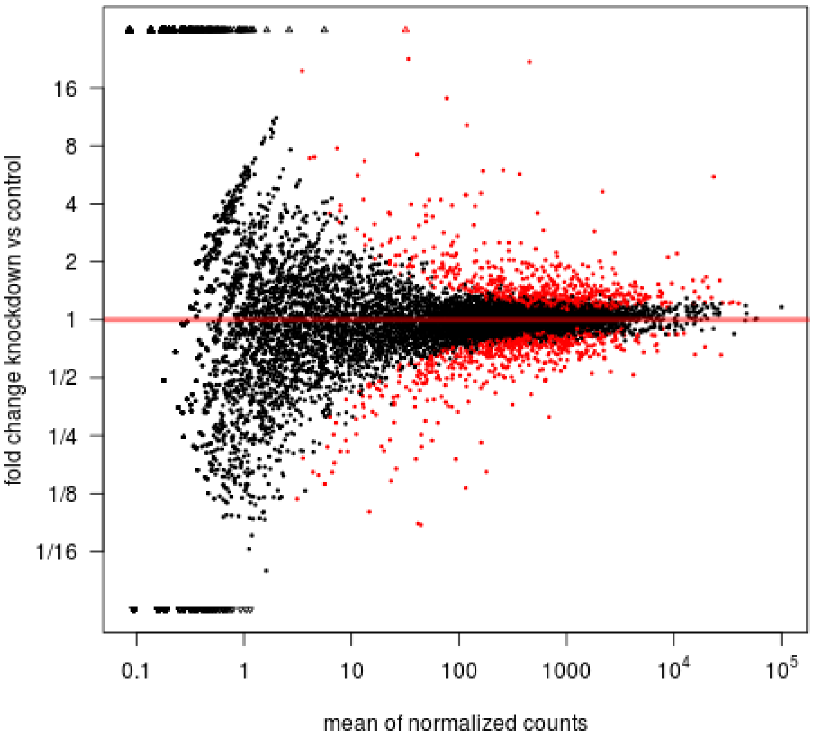
(Similar approach: DSS by Wu, Wang & Wu, 2013)

Weak genes have exaggerated effect sizes

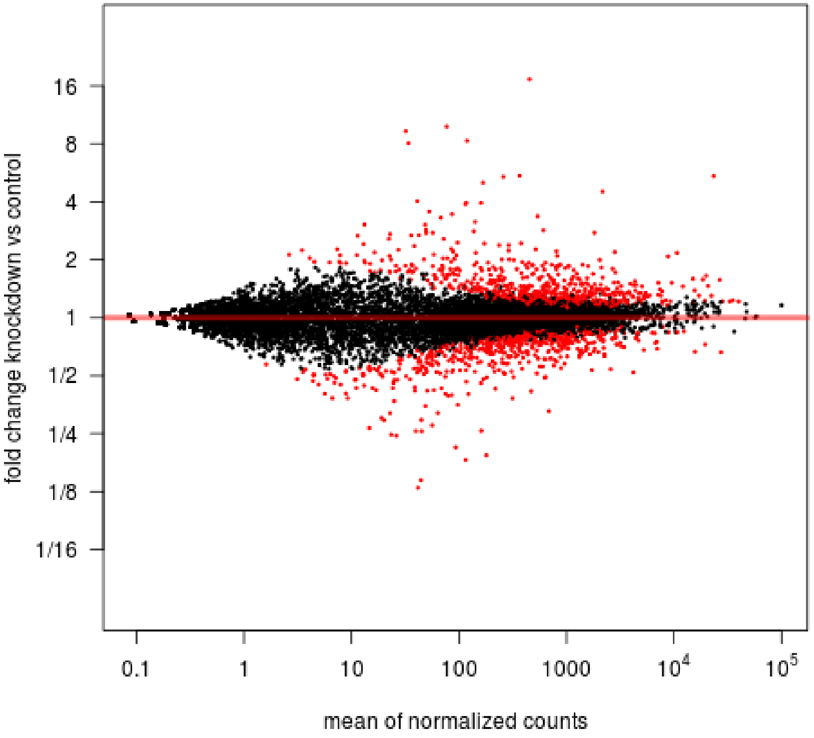


Shrinkage estimation of effect sizes

without shrinkage



with shrinkage



Shrinkage estimation of effect sizes

Procedure:

- Fit GLMs for all genes without shrinkage.
- Estimate normal empirical-Bayes prior from non-intercept coefficients.
- Adding log prior to the GLMs′ log likelihoods results in a ridge penalty term.
- Fit GLMs again, now with the penalized likelihood to get shrunken coefficients.

From testing to estimating

- **Testing:** Is the gene's change noticeably different from zero?
  Can we say whether it is *up or down*?

- **Estimation:** *How strong* is the change?

From testing to estimating

- **Testing:** Is the gene's change noticeably different from zero?
  Can we say whether it is *up or down*?

- **Estimation:** *How strong* is the change?
  How precise is this estimate?

→ Fold change estimates need information on their standard error.

From testing to estimating

→ Fold change estimates need information on their standard error.

It is convenient to have the same precision for all fold-change estimates.

Hence: Shrinkage. (variance-bias trade-off)

Gene ranking

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

- by p value?
- by log 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!

Gene-set enrichment analysis

Given the list of genes with strong effects in an experiment ("hits"): What do they mean?

Common approach: Take a collection of gene sets (e.g., GO, KEGG, Reactome, etc.), look for sets that are enriched in hits.

Gene-set enrichment analysis

Given the list of genes with strong effects in an experiment (“hits”): What do they mean?

Common approach: Take a collection of gene sets (e.g., GO, KEGG, Reactome, etc.), look for sets that are enriched in hits.

Gene-set enrichment analysis

Two approaches:

Categorical test: Is the gene set enriched for *significantly* differentially-expressed genes?

Continuous test: Are the fold changes of the genes in the set particularly strong?

Gene-set enrichment analysis: Worries

Power in RNA-Seq depends on counts.

Hit lists are enriched for genes with high count values: *strong* genes, and genes with *long* transcripts.

This causes bias in categorical tests.

(e.g., Oshlack & Wakefield, 2009)

Gene-set enrichment analysis: Worries

Fold-change estimates in RNA-Seq depends on counts.

Genes with low counts have exaggerated fold changes.

This causes bias in continuous tests.

(e.g., Oshlack & Wakefield, 2009)

Gene-set enrichment analysis: Shrinkage to the rescue

After shrinkage, log-fold-changes a re homoskedastic.
This makes a continuous test easy:

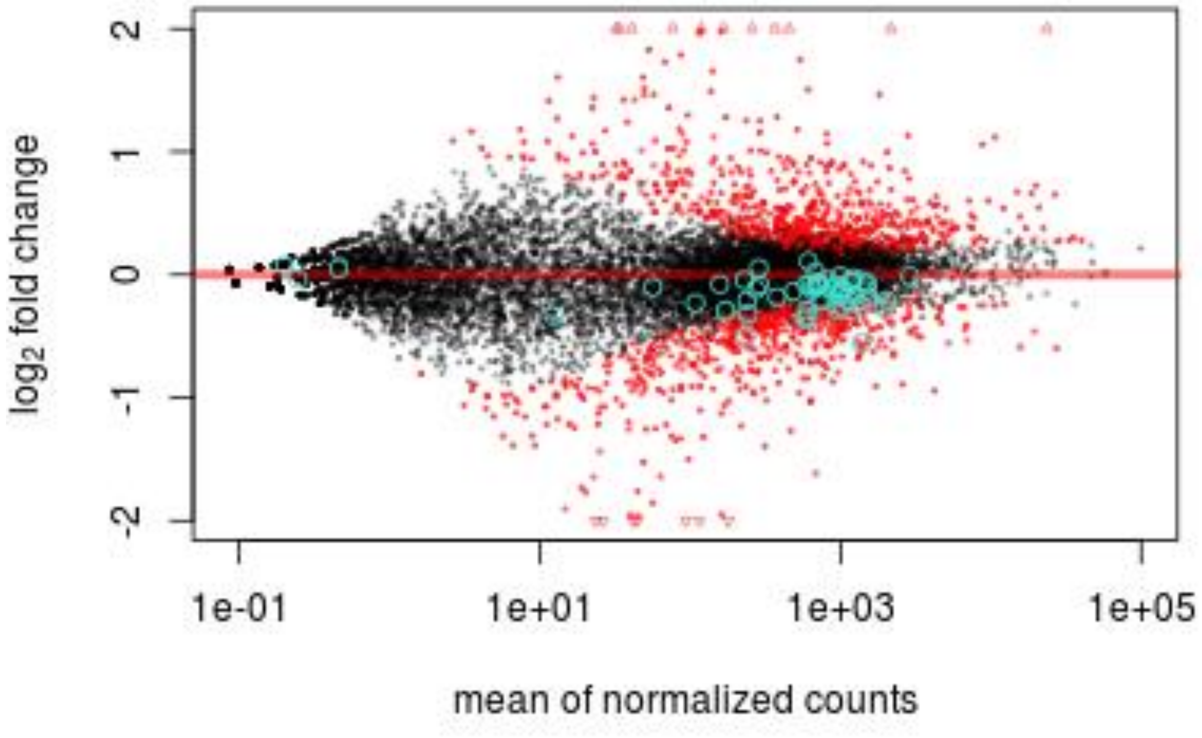
Gene-set enrichment analysis: Shrinkage to the rescue

After shrinkage, log-fold-changes (LFCs) are homoskedastic. This makes a continuous test easy:

Perform an ordinary t test:

- Is the mean of the LFCs of all the genes in the set non-zero?

GSEA with shrunken log fold changes



fly cell culture, knock-down of *pasilla* versus control (Brooks et al., 2011)

turquoise circles: genes in Reactome Path 3717570
“APC/C-mediated degradation of cell cycle proteins”
56 genes, avg LFC: -0.15, p value: $4 \cdot 10^{-11}$ (t test)

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.

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.

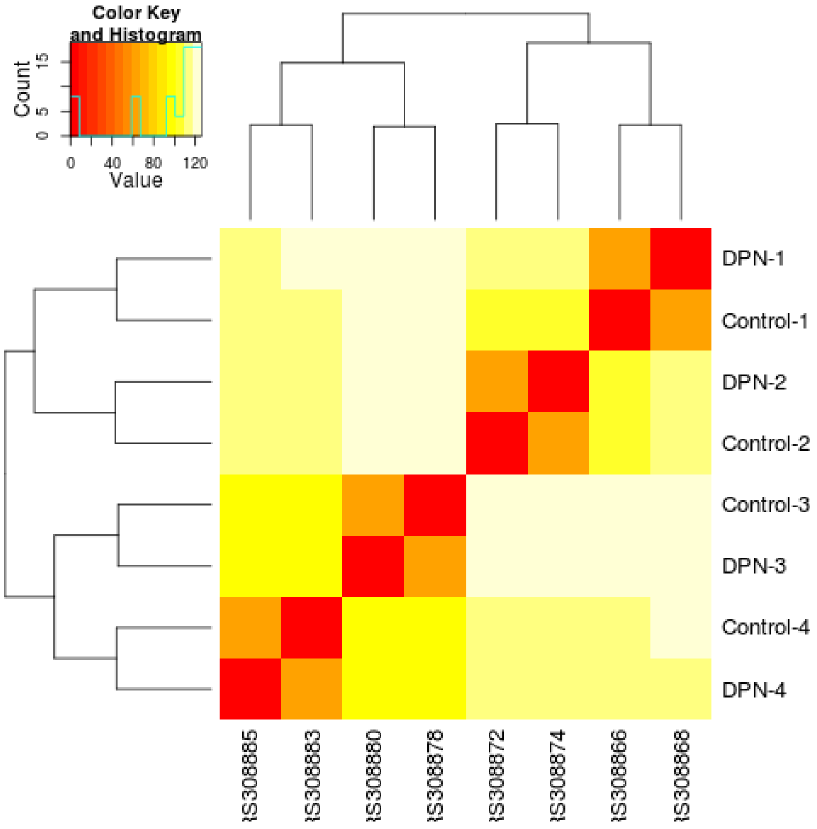
More things to do with shrinkage:
The rlog transformation

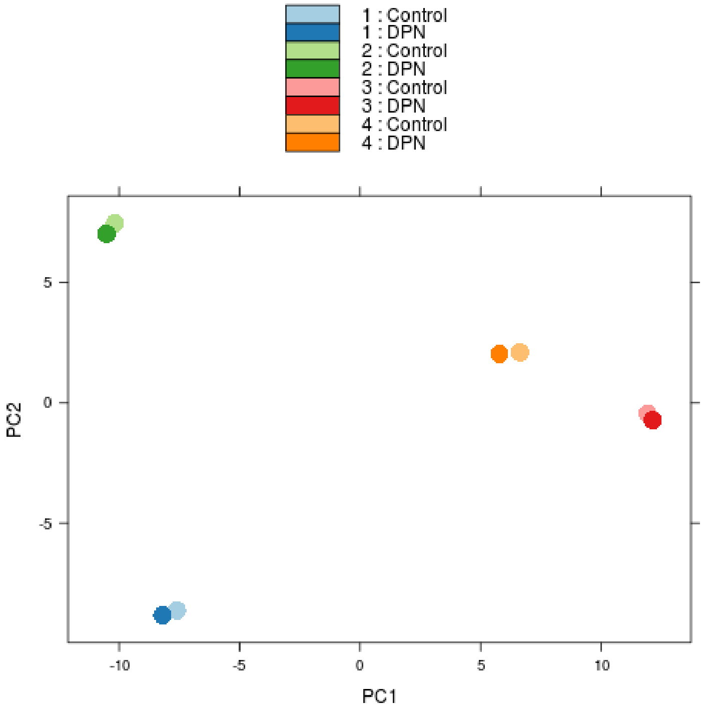
RNA-Seq data is not homoscedastic.

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

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

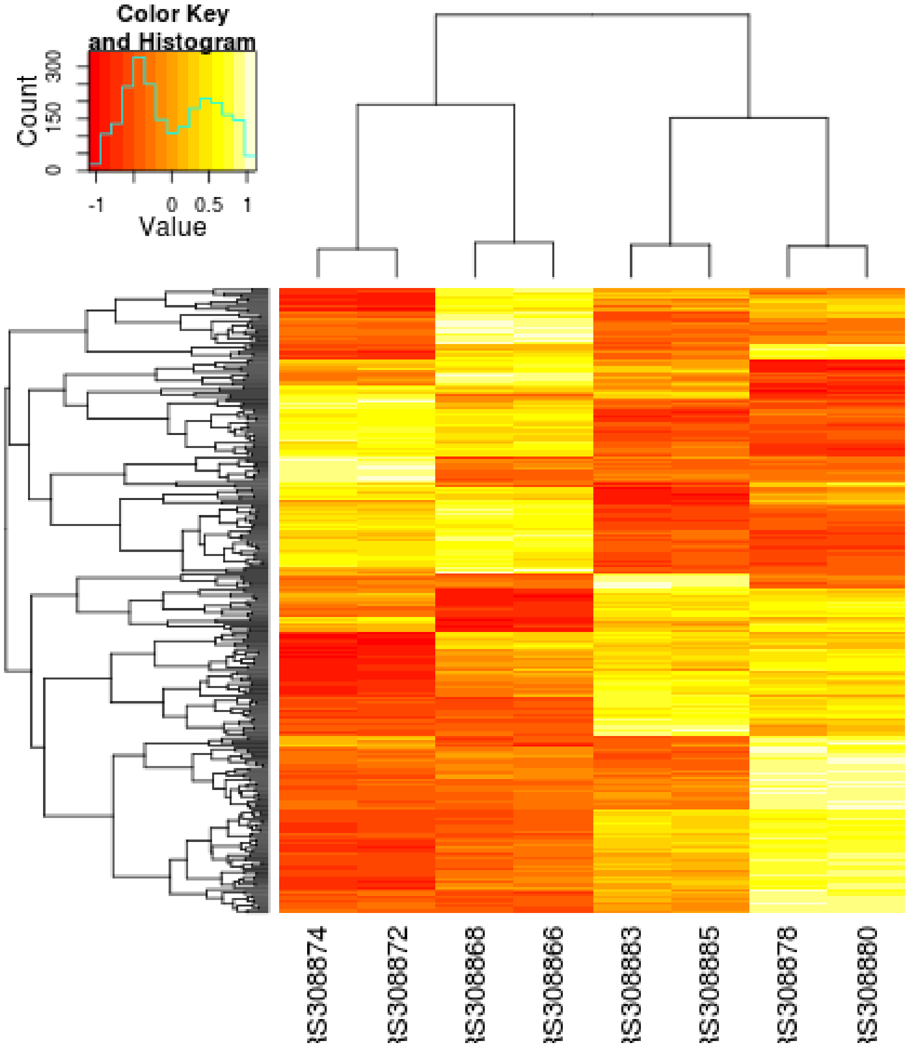
Visualization of rlog-transformed data: Sample clustering and PCA





Data: Parathyroid samples from Haglung et al., 2012

Visualizationof rlog-transformed data: Gene clustering



More things to do with shrinkage:
The rlog transformation

Conceptual idea of the rlog transform:

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

Then “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 wth 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

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

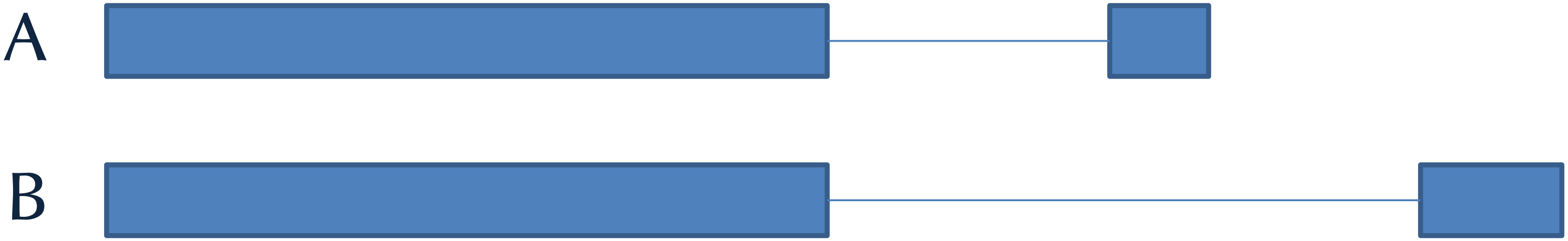
- So far, we looked at read counts *per gene*.

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

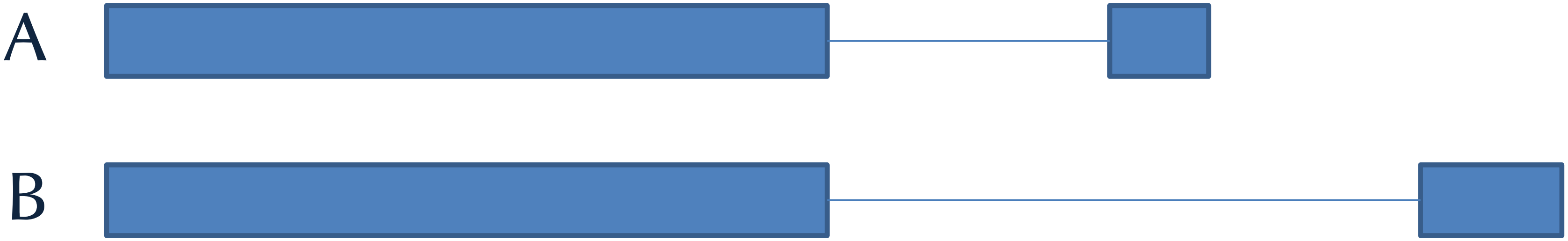
10 reads

30 reads

from A

from B

Assigning reads to transcripts



200 reads

(50 from A,
150 from B?)

5 reads

from A

15 reads

from B

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

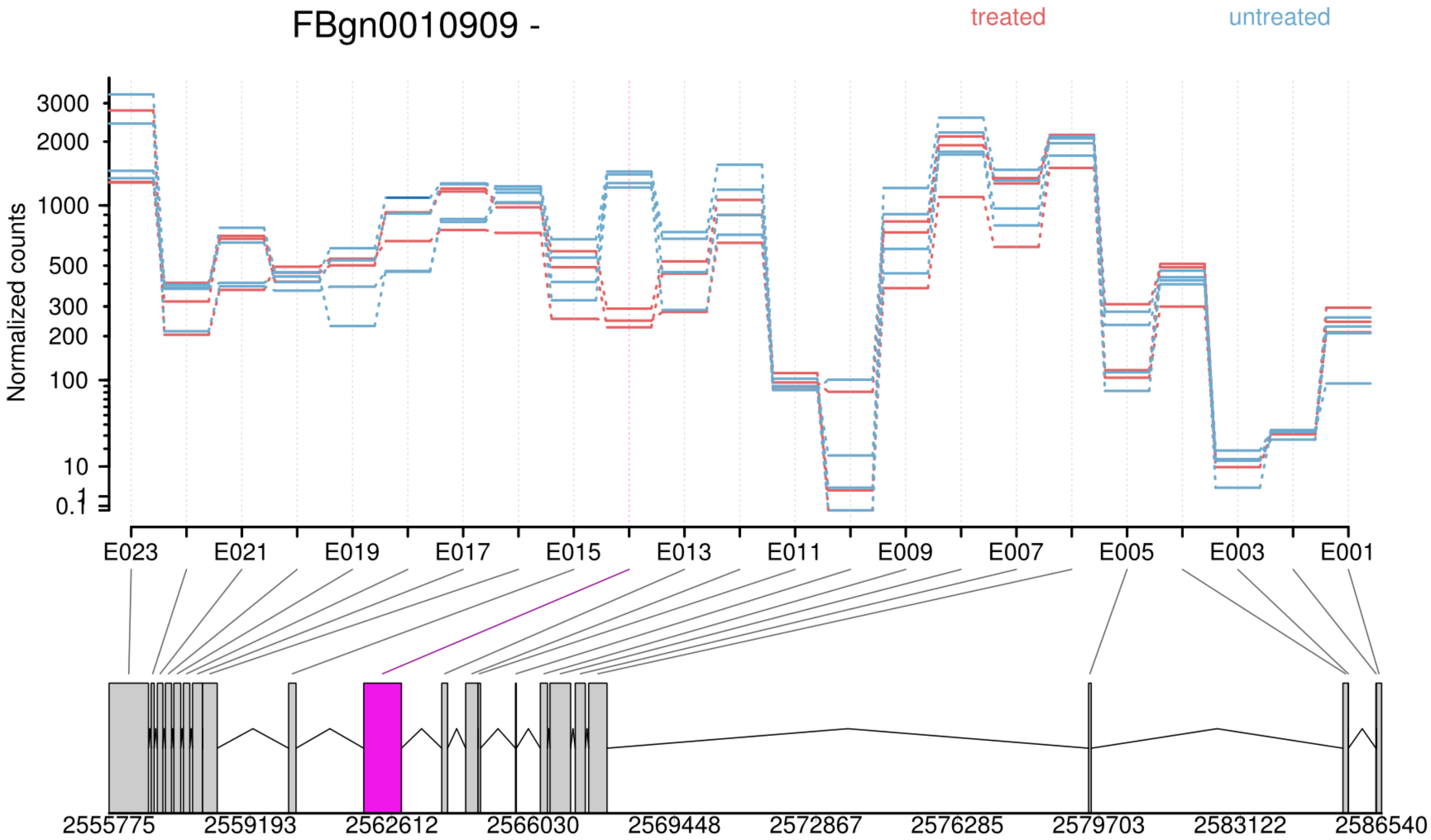
One step back:
Differential exon usage

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

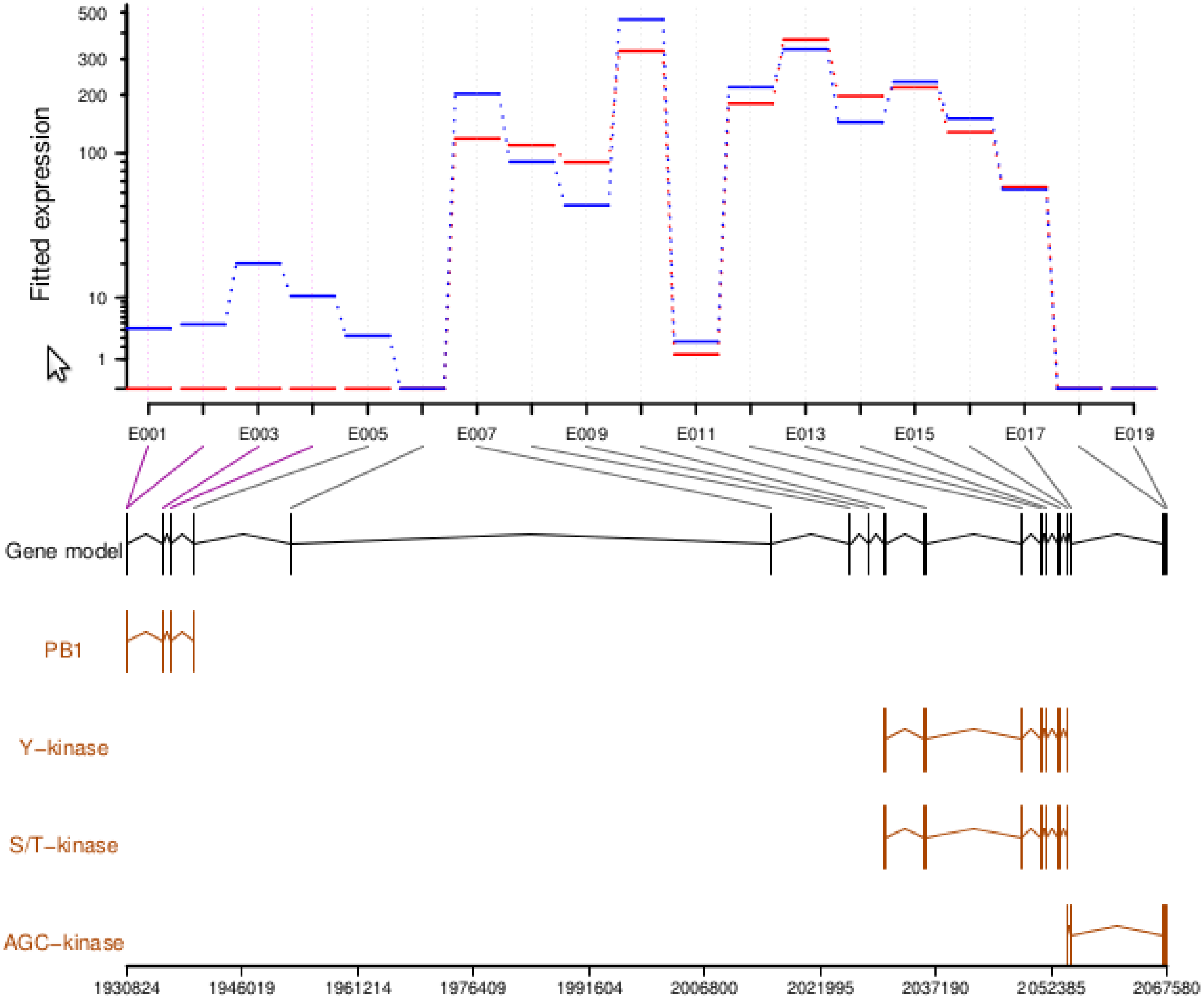
Usage on an exon =

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

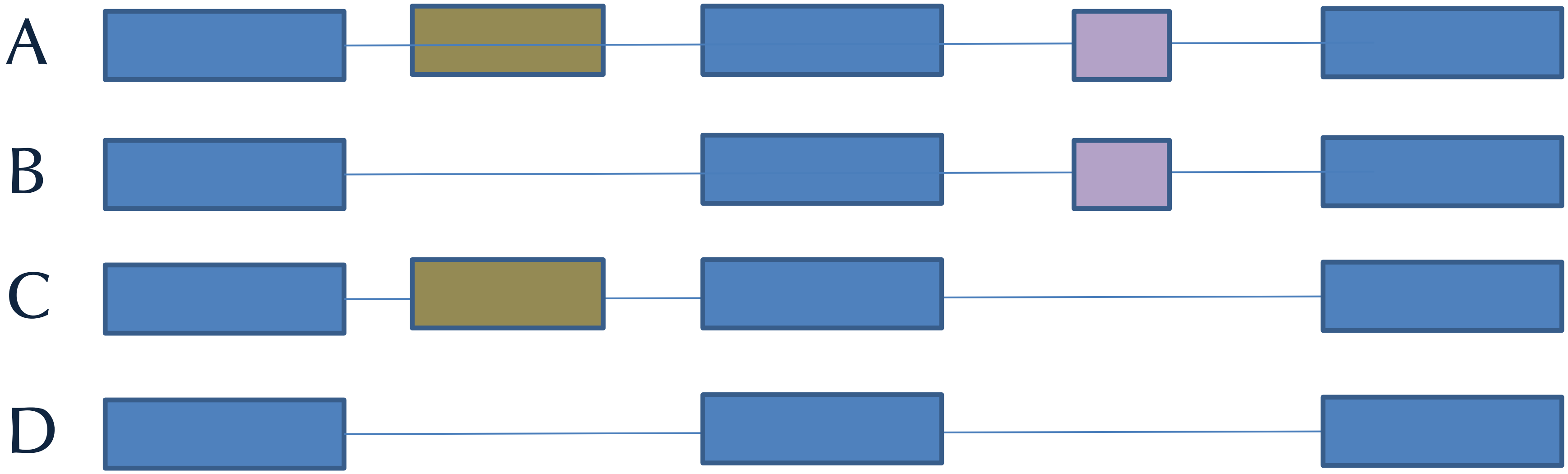
Differential exon usage -- Example



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, GSEA, clustering, ordination, etc.
- Testing for exon usage sheds light on alternative isoform regulation (DEXSeq)

Acknowledgements

Co-authors:

- Wolfgang Huber
- Alejnadro 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 *Radiant*