Differential expression analysis for sequencing count data

Simon Anders

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

EMBL

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.

EMBL

Count data in HTS

Gene	GliNS1	G144	G166	G179	CB541	CB660
13CDNA73	4	0	6	1	0	5
A2BP1	19	18	20	7	1	8
A2M	2724	2209	13	49	193	548
A4GALT	0	0	48	0	0	0
AAAS	57	29	224	49	202	92
AACS	1904	1294	5073	5365	3737	3511
AADACL1	3	13	239	683	158	40
[...]						

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

EMBL

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

EMBL

Normalisation for library size

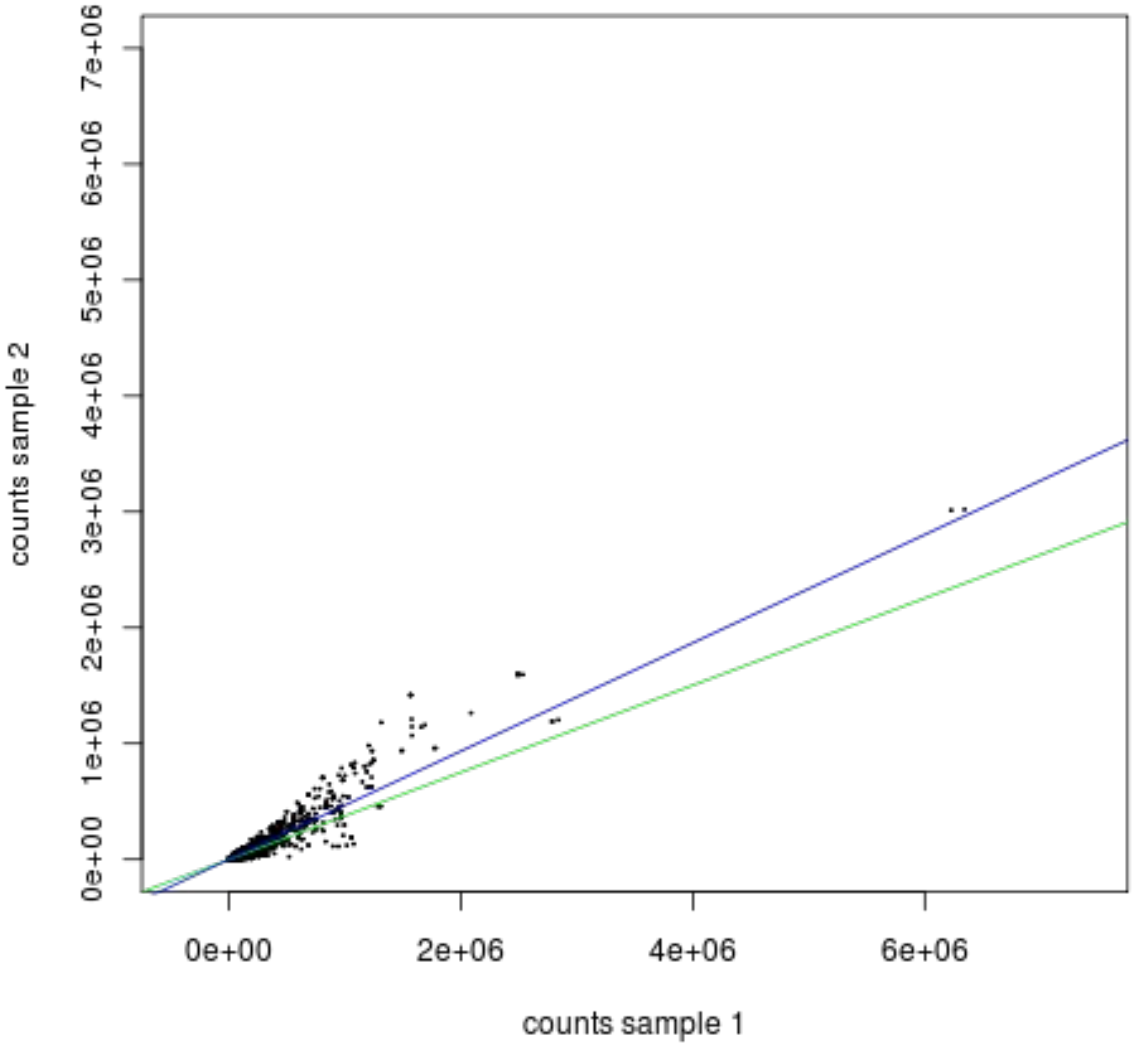
- If sample A has been sampled deeper than sample B, we expect counts to be higher.

- 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.

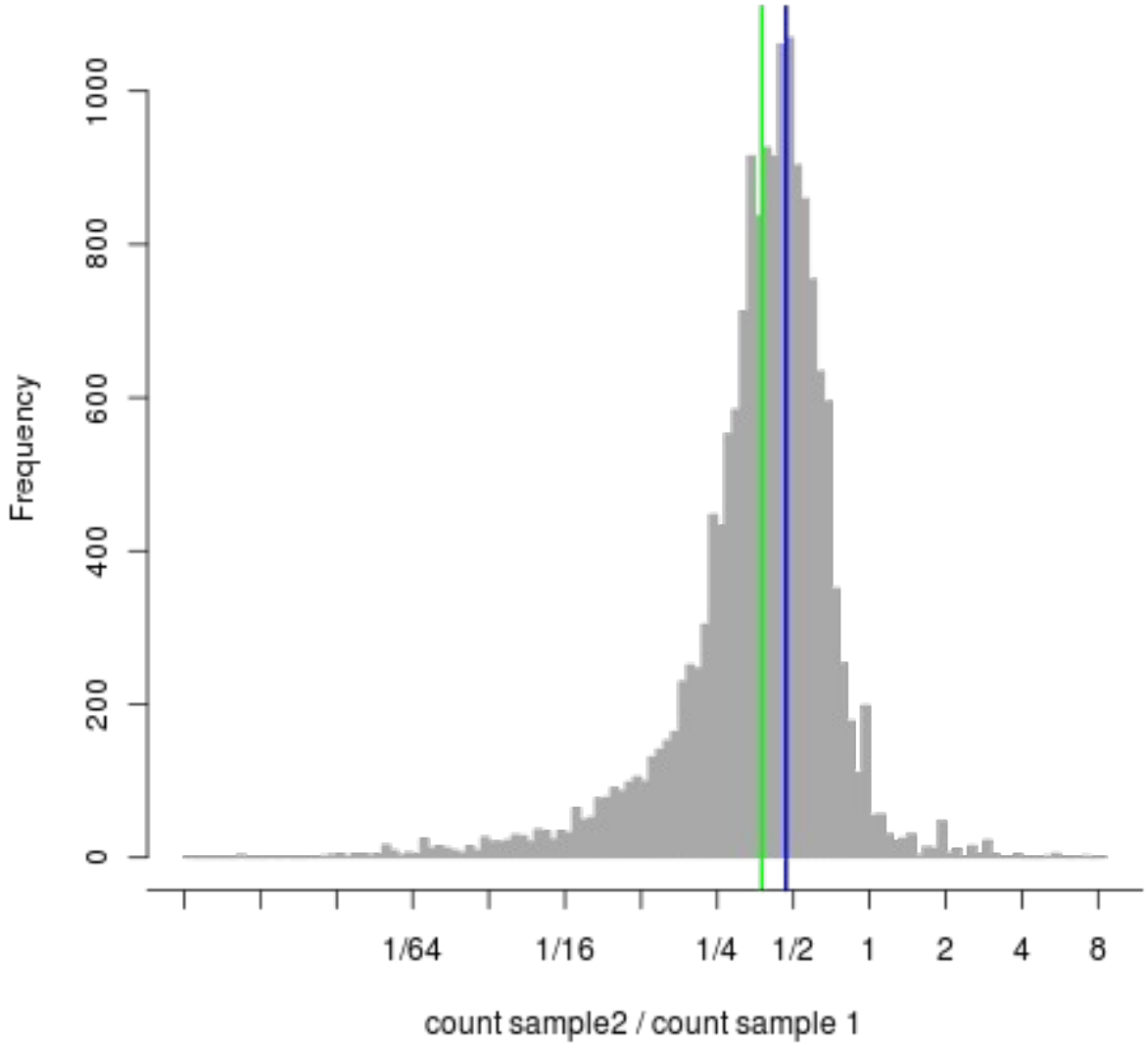
EMBL

Normalisation for library size



EMBL

Normalisation for library size



EMBL

Normalizing for more than two samples

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.

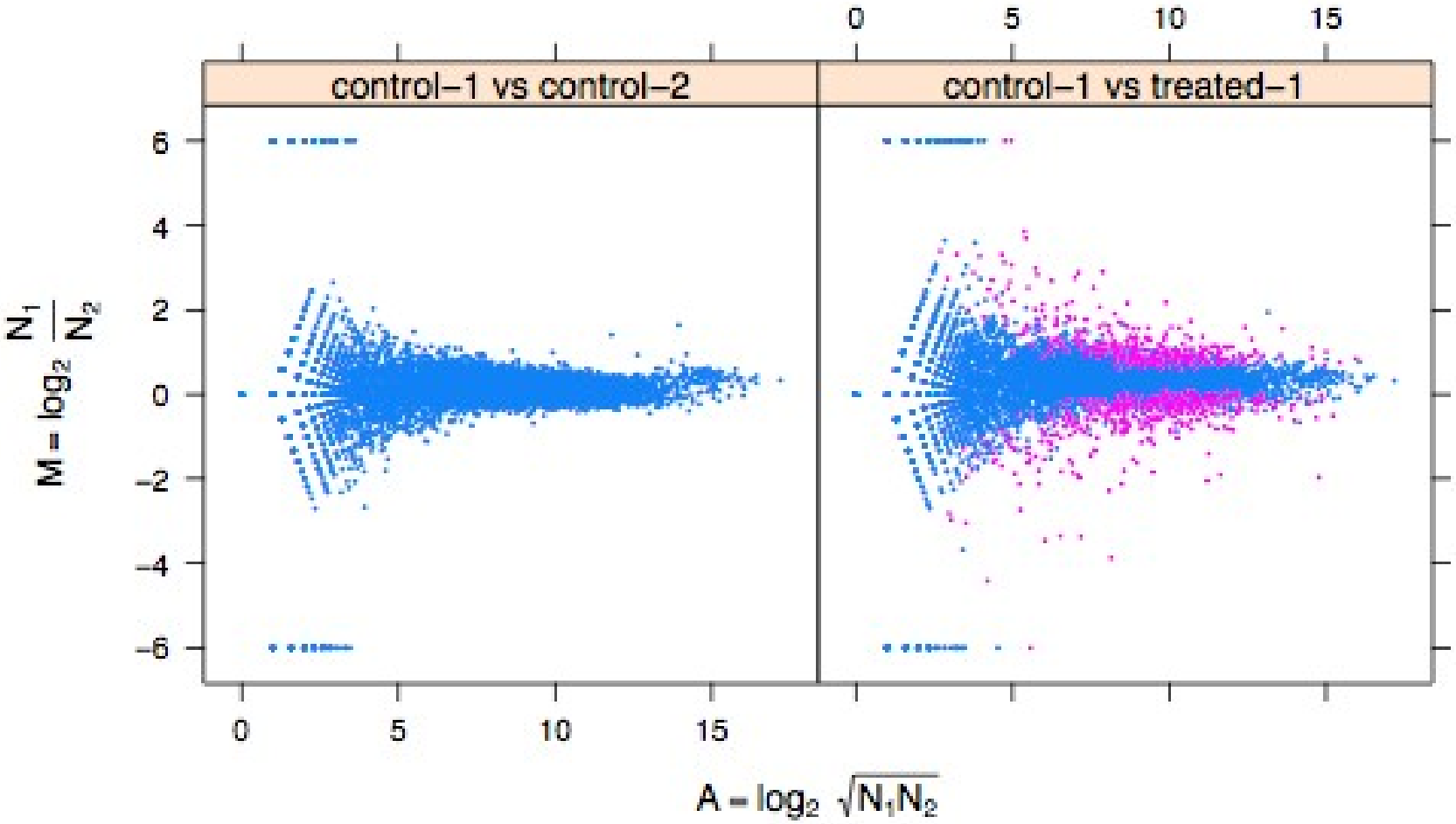
Anders and Huber, 2010
similar approach: Robinson and Oshlack, 2010

EMBL

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 $\sqrt{v}/\sqrt{m}$.

EMBL

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)

EMBL

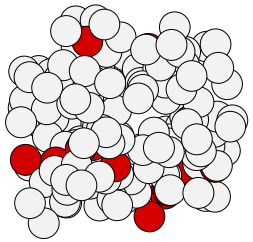
Counting noise

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

- Why?

EMBL

The Poisson distribution

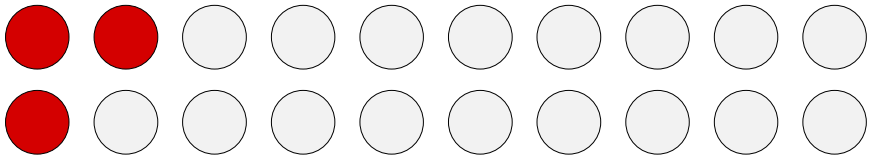


This bag contains very many small balls, 10% of which are red.

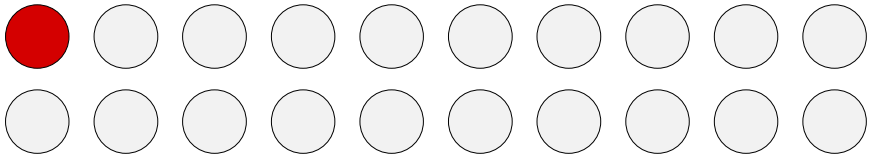
Several experimenters are tasked with determining the percentage of red balls.

Each of them is permitted to draw 20 balls out of the bag, without looking.

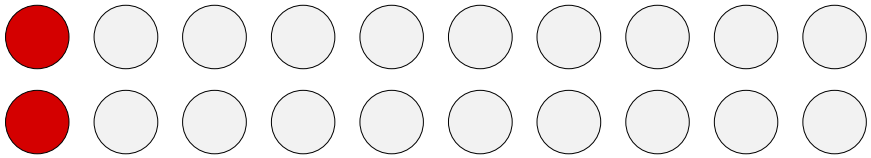
EMBL



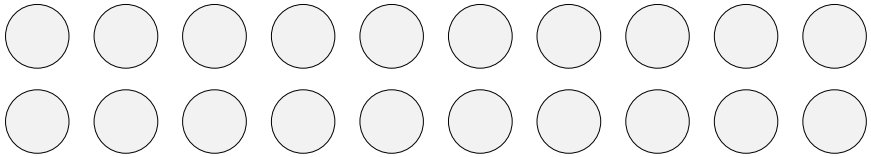
3 / 20 = 15%



1 / 20 = 5%

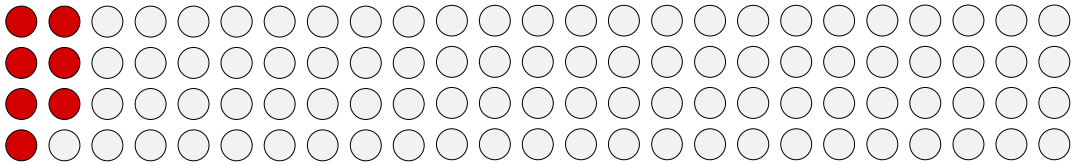


2 / 20 = 10%

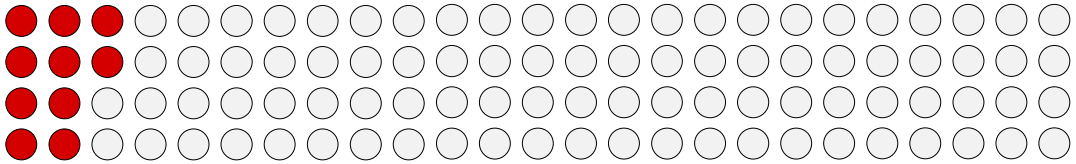


0 / 20 = 0%

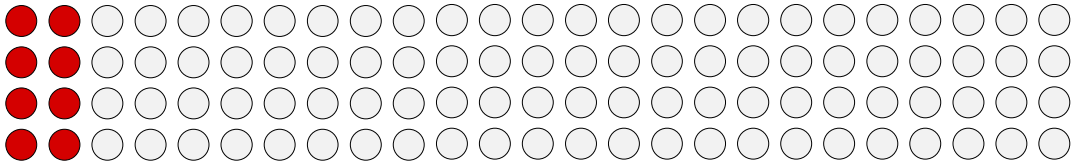
EMBL



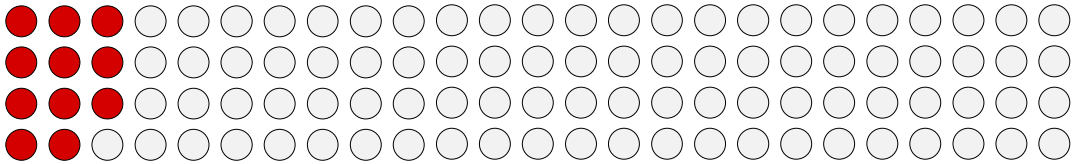
7 / 100 = 7%



10 / 100 = 10%



8 / 100 = 8%



11 / 100 = 11%

EMBL

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 $\hat{p} = k/n$ hence has the expected value $\mu/n = p$ and the standard error $\Delta p = \sqrt{\mu}/n = p/\sqrt{\mu}$. The relative error is $\Delta p/p = 1/\sqrt{\mu}$.

balls drawn	expected number of red balls	relative error of estimate
20	2	$1/\sqrt{2} = 71\%$
100	10	$1/\sqrt{10} = 32\%$

EMBL

Poisson distribution: Counting uncertainty

expected number of red balls	standard deviation of number of red balls	relative error in estimate for fraction of red balls
10	$\sqrt{10} = 3.2$	$1/\sqrt{10} = 31.6\%$
100	$\sqrt{100} = 10.0$	$1/\sqrt{100} = 10.0\%$
1,000	$\sqrt{1{,}000} = 31.6$	$1/\sqrt{1{,}000} = 3.2\%$
10,000	$\sqrt{10{,}000} = 100.0$	$1/\sqrt{10{,}000} = 1.0\%$

EMBL

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

Hence, no need to estimate the variance

according to several authors: Marioni et al. (2008), Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010)

Really?

Is HTS count data Poisson-distributed?

To sort this out, we have to distinguish *two* sources of noise.

EMBL

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?

EMBL

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*.

EMBL

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.

EMBL

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.

EMBL

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?"

EMBL

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

Look at the differences between replicates? They show how much variation occurs without difference in treatment.

Could it be that the treatment has no effect and the difference between treatment and control is just a fluctuation of the same kind as between replicates?

To answer this, we need to assess the strength of this sample noise.

Summary: Noise

We distinguish:

- Shot noise
  - unavoidable, appears even with perfect replication
  - dominant noise for weakly expressed genes
- Technical noise
  - from sample preparation and sequencing
  - negligible (if all goes well)
- Biological noise
  - unaccounted-for differenced between samples
  - Dominant noise for strongly expressed genes

can be computed

needs to be estimated from the data

EMBL

Replicates

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

EMBL

Replication at what level?

Replicates should differ in *all* aspects in which control and treatment samples differ, except for the actual treatment.

EMBL

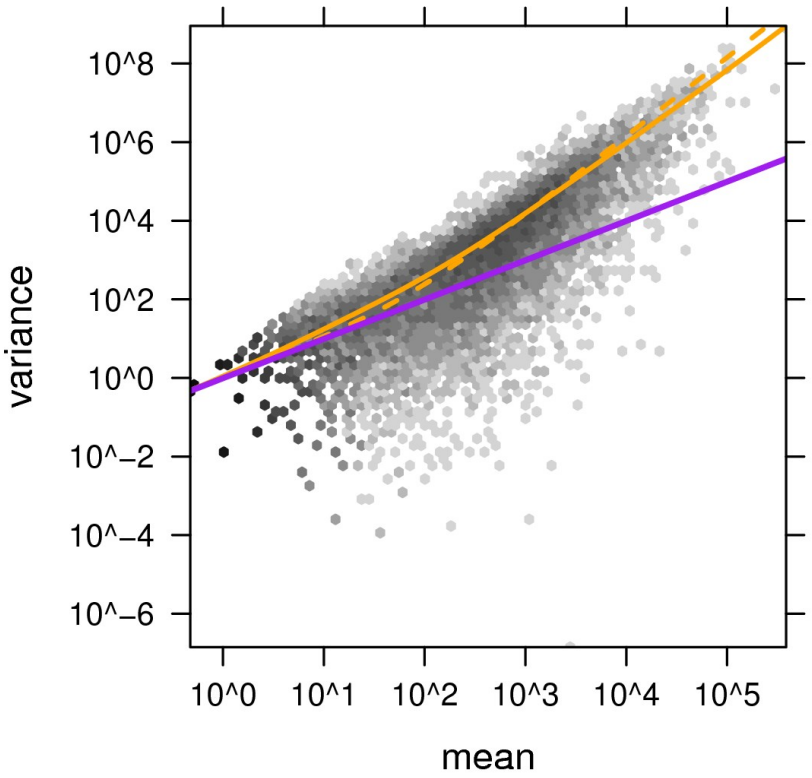
Estimating noise from the data

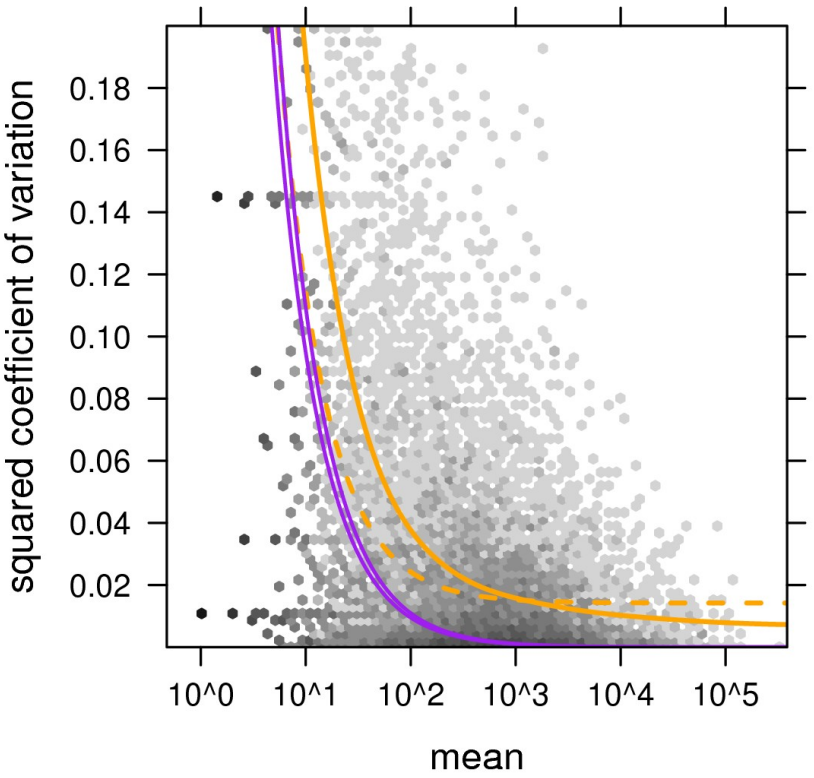
- If we have many replicates, we can estimate the variance for each gene.

- With only few replicates, we need an additional assumption. We use: “Genes with similar expression strength have similar variance.”

EMBL

Variance depends strongly on the mean





Variance calculated from comparing two replicates

Poisson	$v = \mu$	———
Poisson + constant CV	$v = \mu + \alpha \mu^2$	- - - -
Poisson + local regression	$v = \mu + f(\mu^2)$	———

EMBL

Technical and biological replicates

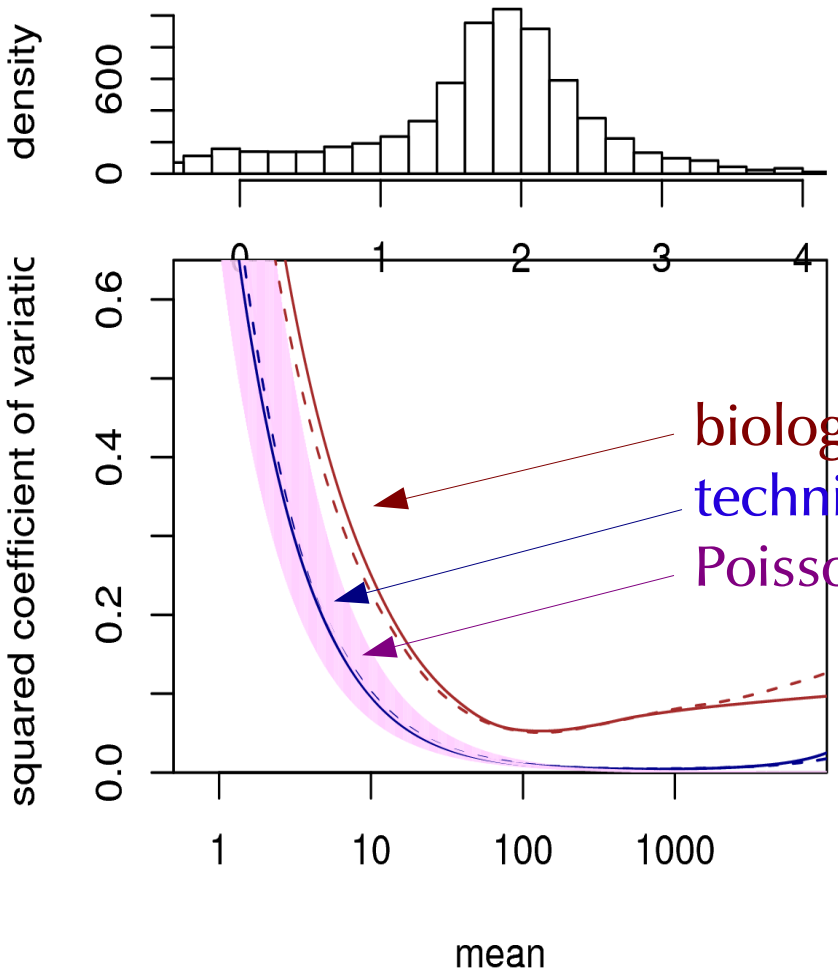
Nagalakshmi *et al.* (2008) have found that

- counts for the same gene from different *technical* replicates have a variance equal to the mean (Poisson).
- counts for the same gene from different *biological* replicates have a variance exceeding the mean (overdispersion).

Marioni *et al.* (2008) have looked confirmed the first fact (and caused some confusion about the second fact).

EMBL

Technical and biological replicates

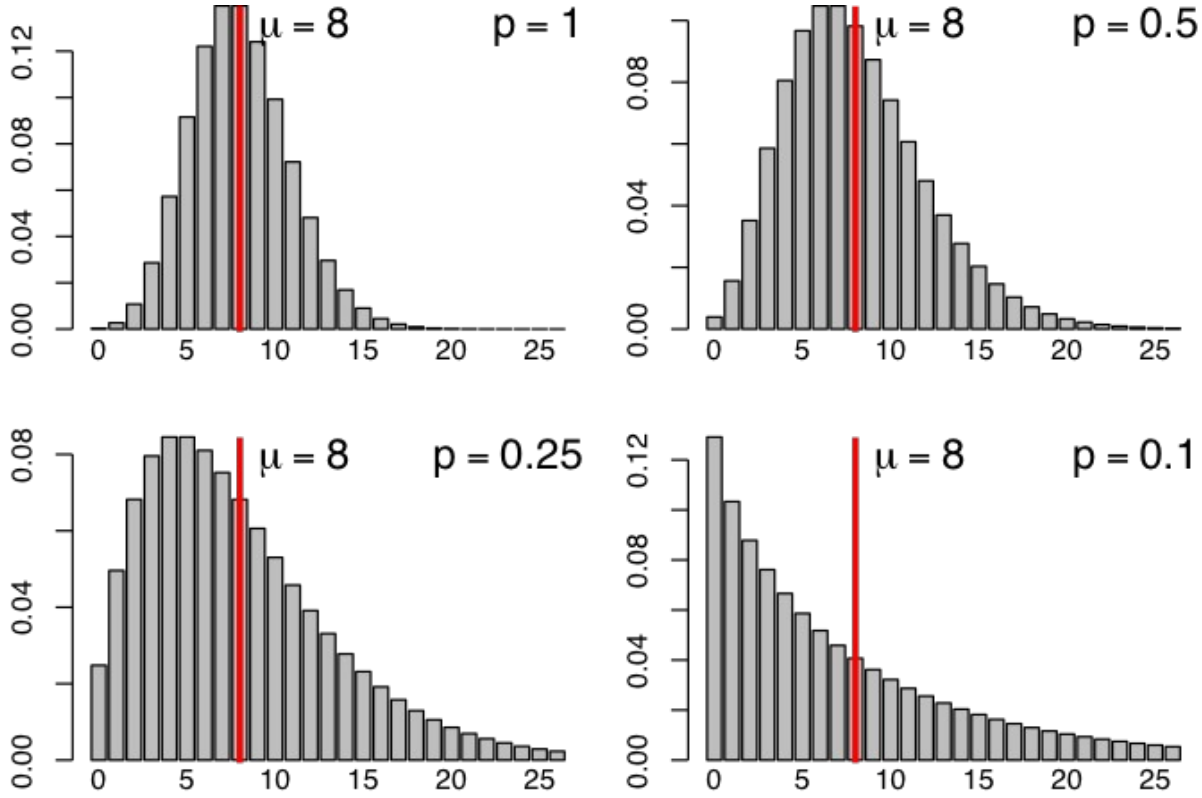


RNA-Seq of yeast [Nagalakshmi et al, 2008]

EMBL

The negative-binomial distribution

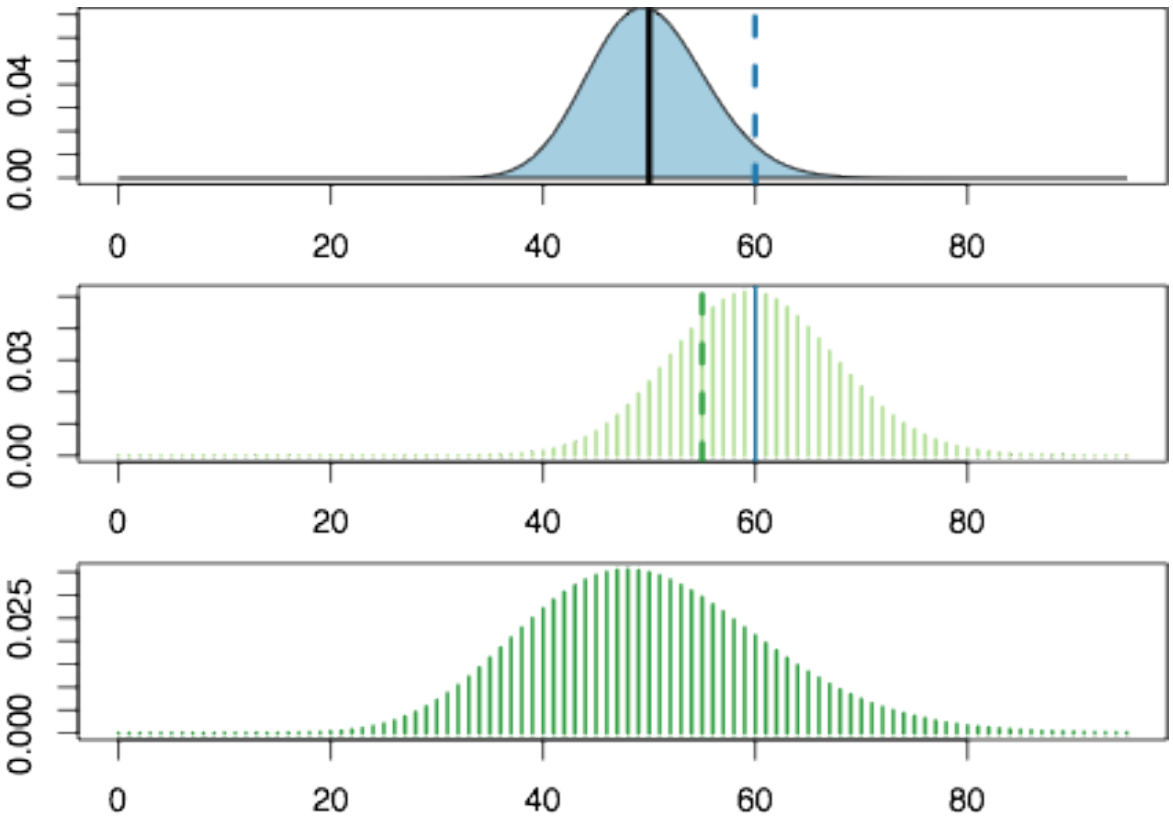
A commonly used generalization of the Poisson distribution with *two* parameters



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

EMBL

The NB distribution 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 $q+v$.

EMBL

Testing: Null hypothesis

Model:

The count for a given gene in sample *j* come from negative binomial distributions with the mean $s_j \mu_\rho$ and variance $s_j \mu_\rho + s_j^2 v(\mu_\rho)$.

s_j	relative size of library *j*
μ_ρ	mean value for condition ρ
$v(\mu_\rho)$	fitted variance for mean μ_ρ

Null hypothesis:

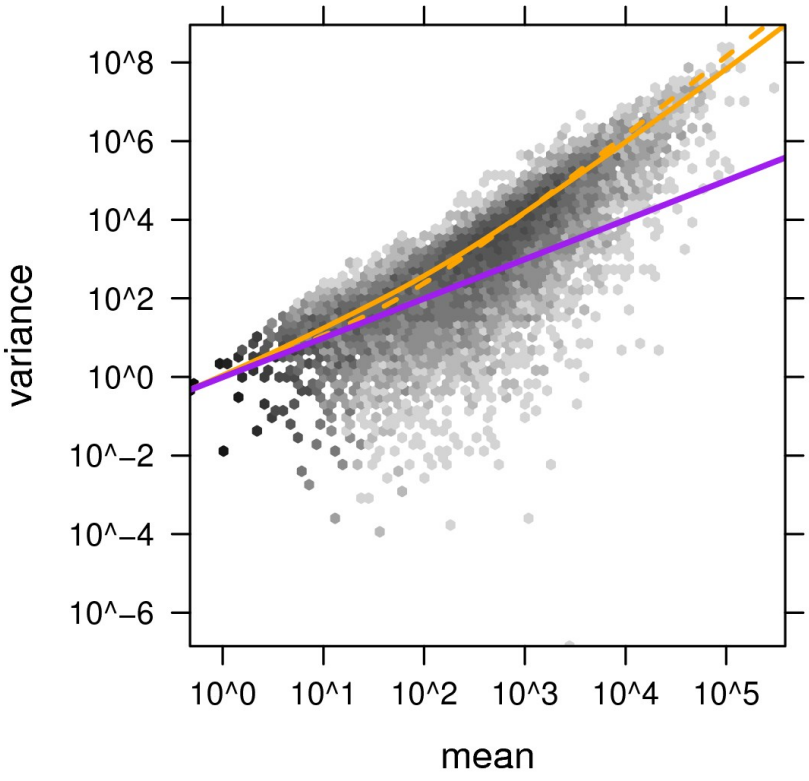
The experimental condition r has no influence on the expression of the gene under consideration:

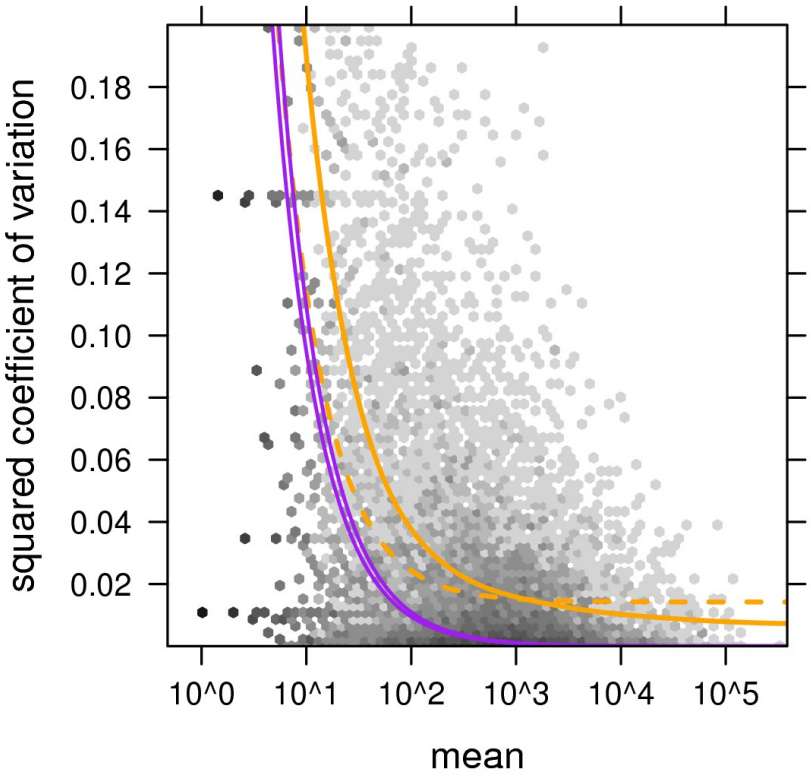
$$\mu_{\rho_1} = \mu_{\rho_2}$$

EMBL

Model fitting

- Estimate the variance from replicates
- Fit a line to get the variance-mean dependence $v(\mu)$
  (local regression for a gamma-family generalized linear model, extra math needed to handle differing library sizes)





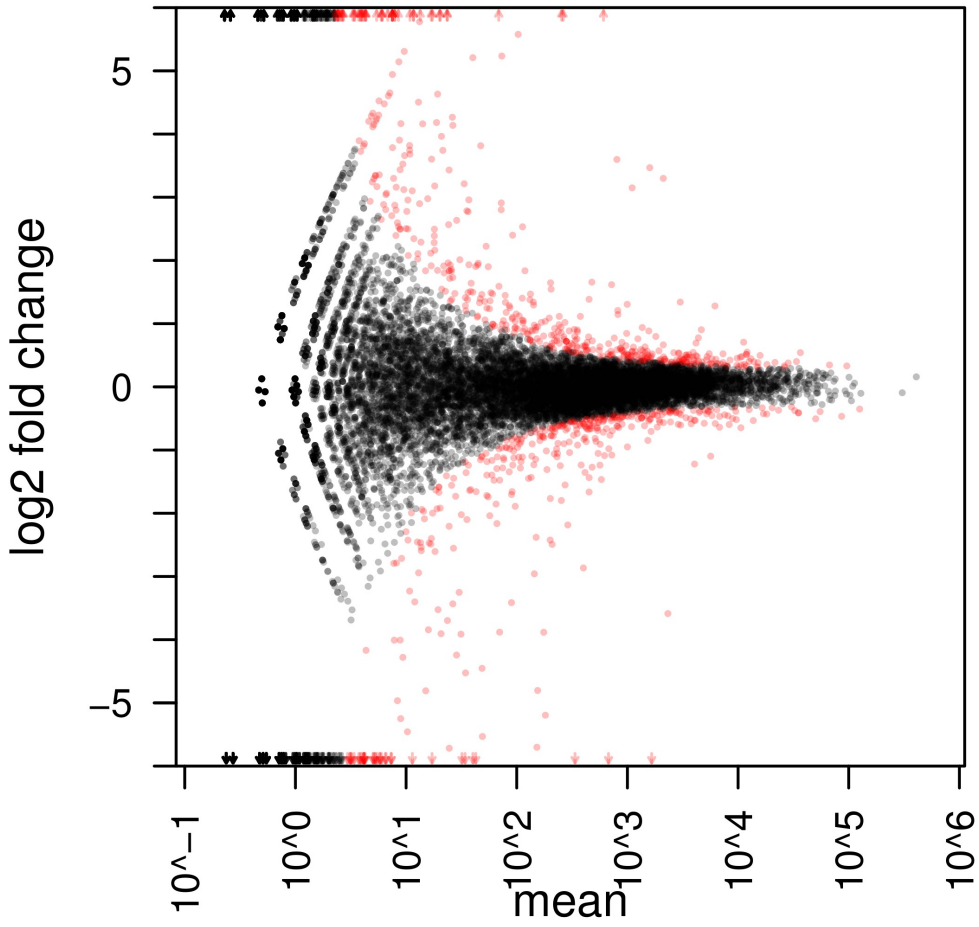
Testing for differential expression

- For each of two conditions, add the count from all replicates, and consider these sums K_{iA} and K_{iB} as NB-distributed with moments as estimated and fitted.
- Then, we calculate the probability of observing the actual sums or more extreme ones, conditioned on the sum being $k_{iA}+k_{iA}$, to get a p value.

(similar to the test used in Robinson and Smyth's *edgeR)*

EMBL

Differential expression

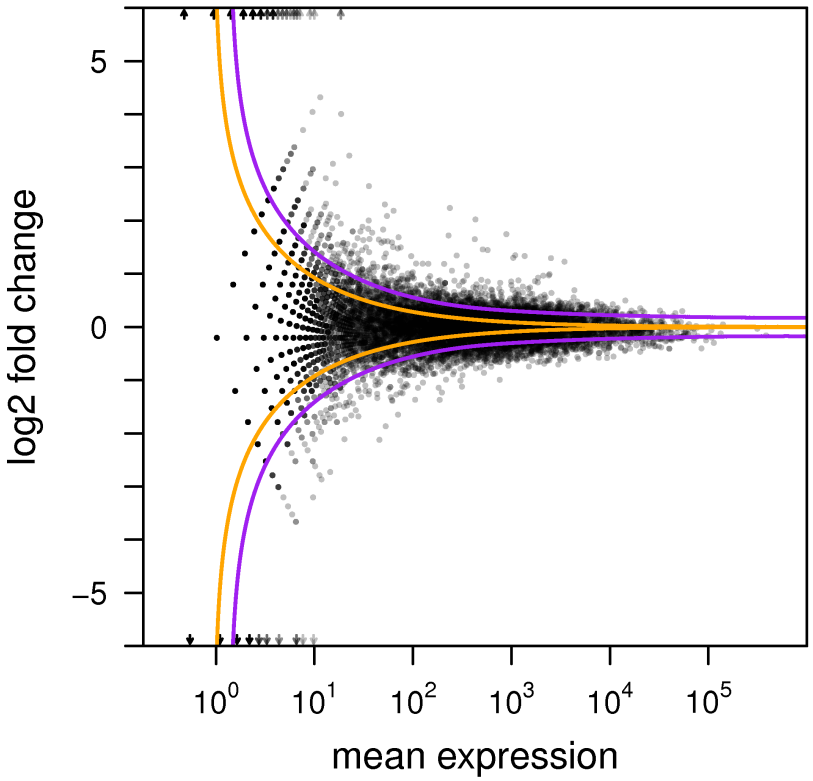


RNA-Seq data: overexpression of two different genes in flies [data: Furlong group]

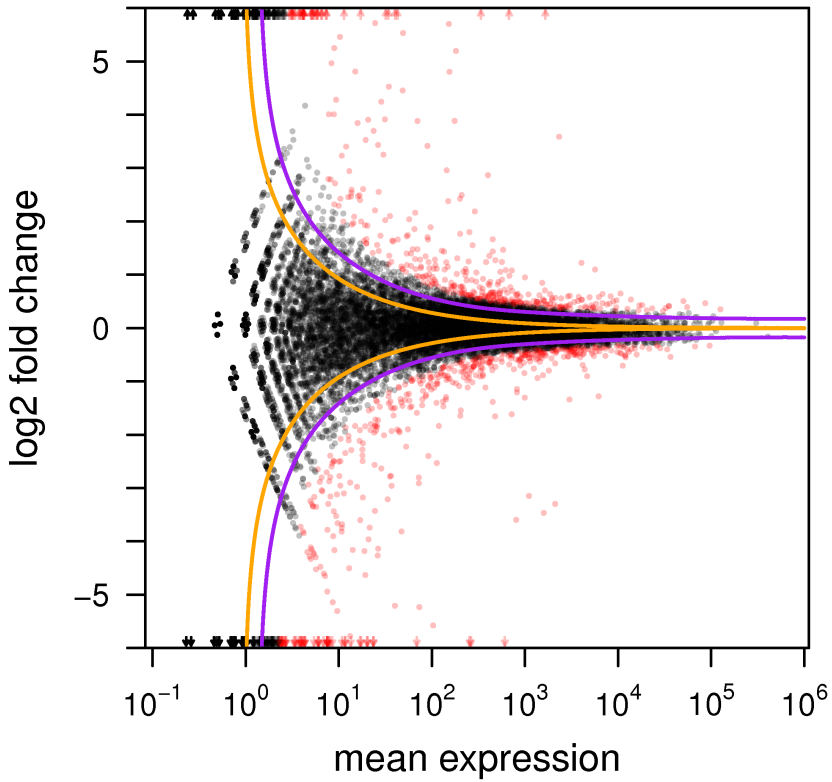
EMBL

Type-I error control

comparison of
two replicates

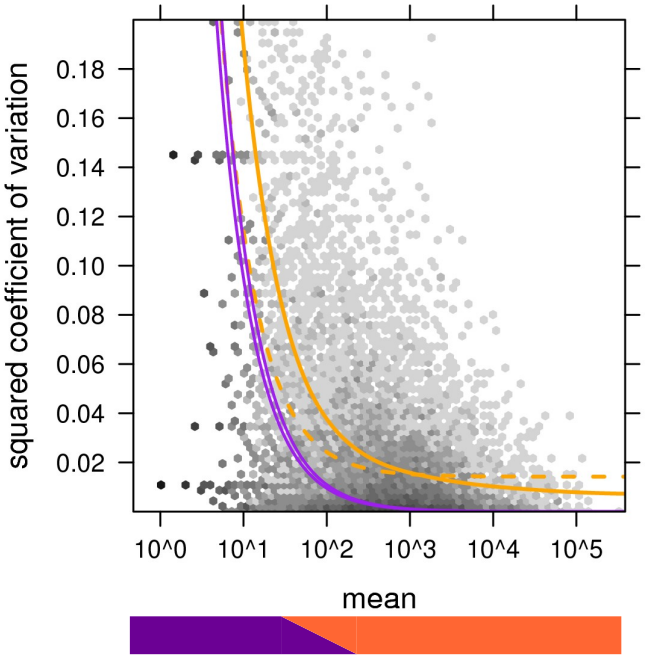


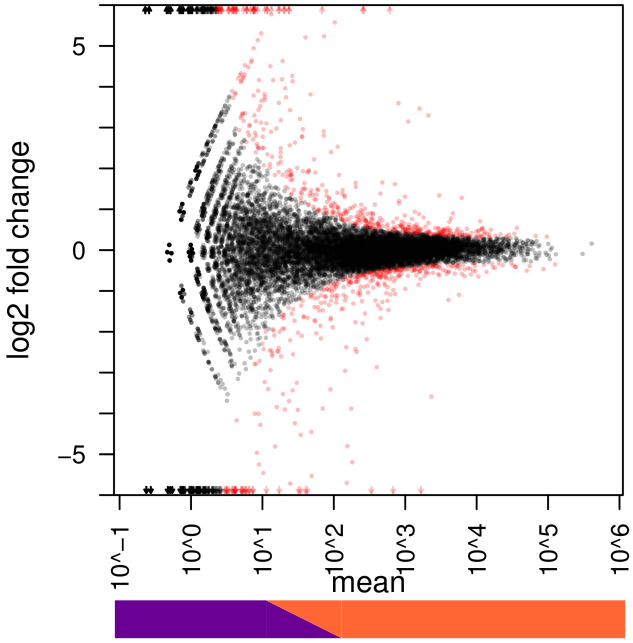
comparison of
treatment vs control



EMBL

Two noise ranges





dominating noise

shot noise (Poisson)

biological noise

How to improve power?

deeper sampling

more biological replicates

EMBL

Conclusions I

- Proper estimation of variance between *biological* replicates is vital. Using Poisson variance is incorrect.
- Estimating variance-mean dependence with local regression works well for this purpose.
- The negative-binomial model allows for a powerful test for differential expression

- S. Anders, W. Huber: "Differential expression analysis for sequence count data", Genome Biol **11** (2010) R106
- Software (*DESeq*) available from Bioconductor and EMBL web site.

EMBL

Further use cases

Similar count data appears in

- comparative ChiP-Seq
- barcode sequencing
- ...

and can be analysed with *DESeq* as well.

EMBL

Comparative ChIP-Seq

How does binding of a certain transcription factor differ between two conditions?

Too naive approach: Use a peak finder on each sample, look for binding sites with peaks in one but not the other condition

Comparaive ChIP-Seq with DESeq

Step 1: Get a list of counting bins by either

- running a peak finder on each samples and merging the peak lists, or
- merging the reads and running the finder on the pooled reads, or
- using windows around annotated features

Step 2: Make a count table:
columns – samples; rows – counting bins

and use DESeq

Note: The input samples are used in Step 1 only.

EMBL

Generalized linear models

Simple design:

- Two groups of samples (“control” and “treatment”), no sub-structure within each group.

Common complex designs:

- Designs with blocking factors
- Factorial designs

EMBL

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

EMBL

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*:

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

EMBL

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*:

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

EMBL

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

EMBL

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

EMBL

GLMs: CLIP-Seq/RNA-Seq assay

full model:

count ~ sample + assayType + assayType:treatment

reduced model:

count ~ sample + assayType

EMBL

Alternative splicing

- So far, we counted reads in *genes*.
- To study alternative splicing, reads have to be assigned to *transcripts*.
- This introduces ambiguity, which adds uncertainty.
- Current tools (e.g., *cufflinks*) allow to quantify this uncertainty.
- However: To assess the significance of differences to isoform ratios between conditions, the assignment uncertainty has to be combined with the noise estimates.
- This is not yet possible with existing tools.

EMBL

Regulation of isoform abundance ratios

- In higher eukaryotes, most genes have several isoforms.
- RNA-Seq is better suited than microarrays to see which isoforms are present in a sample.
- This opens the possibility to study regulation of isoform abundance ratios, e.g.: Is a given exon spliced out more often in one tissue type than in another one?

- We recently released *DEXSeq*, a tool to test for *differential isoform expression* in RNA-Seq data.

EMBL

Data set used for to demonstrate DEXSeq:

Genome Research

Research 21:193–202 © 2011

Conservation of an RNA regulatory map between *Drosophila* and mammals

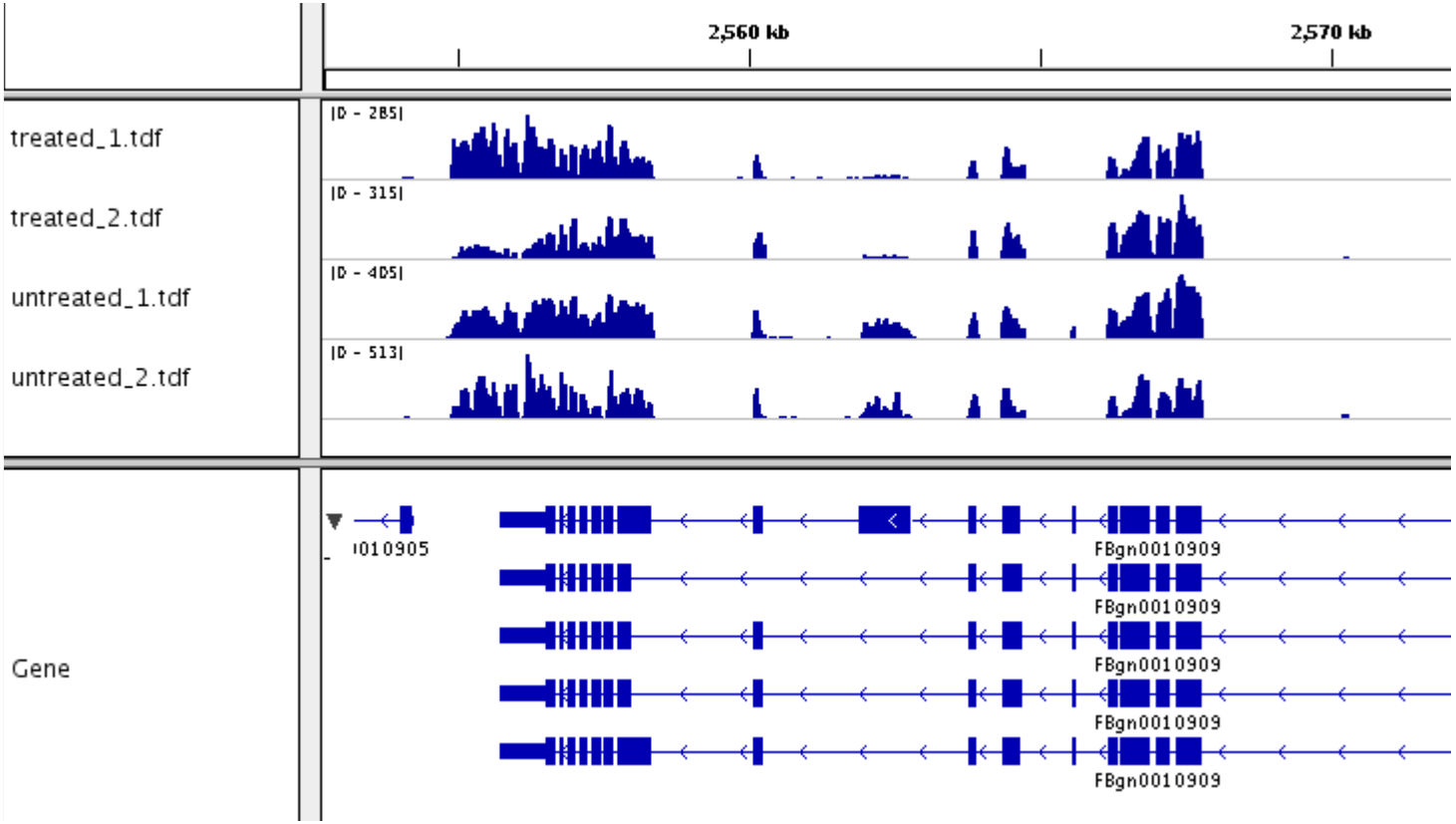
Angela N. Brooks,[1,7] Li Yang,[2,7] Michael O. Duff,[2,3] Kasper D. Hansen,[4] Jung W. Park,[2,3] Sandrine Dudoit,[4,5] Steven E. Brenner,[1,6,8] and Brenton R. Graveley[2,3,8]

Drosophila melanogaster S2 cell cultures:

- control (no treatment):
  4 biological replicates (2x single end, 2x paired end)
- treatment: knock-down of pasilla (a splicing factor)
  3 biological replicates (1x single end, 2x paired end)

EMBL

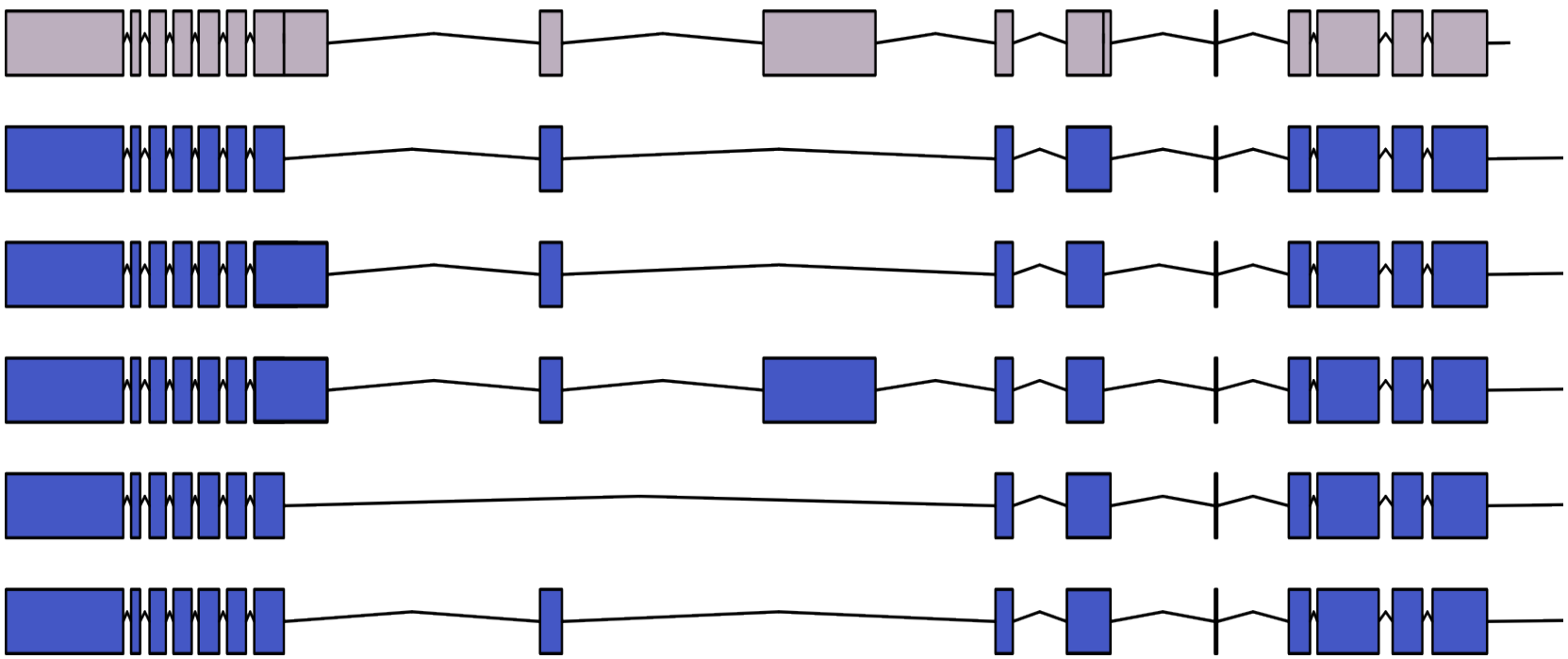
Alternative isoform regulation



Data: Brooks et al., Genome Res., 2010

EMBL

Exon counting bins



EMBL

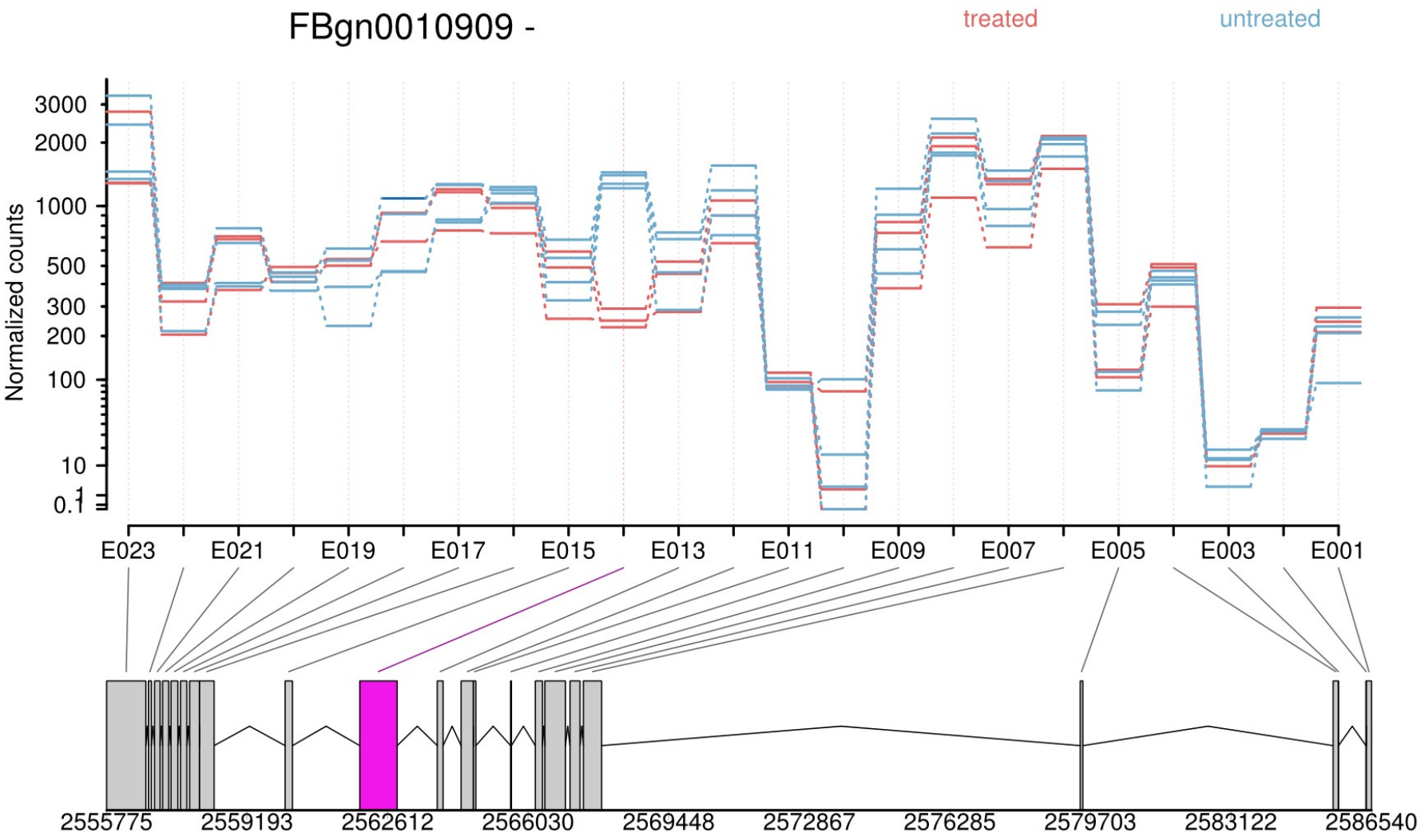
Count table for a gene

number of reads mapped to each exon (or part of exon) in gene *msn*:

	treated_1	treated_2	control_1	control_2	
E01	398	556	561	456	
E02	112	180	153	137	
E03	238	306	298	226	
E04	162	171	183	146	
E05	192	272	234	199	
E06	314	464	419	331	
E07	373	525	481	404	
E08	323	427	475	373	
E09	194	213	273	176	
E10	90	90	530	398	<--- !
E11	172	207	283	227	
E12	290	397	606	368	<--- ?
E13	33	48	33	33	
E14	0	33	2	37	
E15	248	314	468	287	
E16	554	841	1024	680	

[...]

EMBL


FBgn0010909 -
treated
untreated
Normalized counts
3000
2000
1000
500
300
200
100
10
1
0.1
E023
E021
E019
E017
E015
E013
E011
E009
E007
E005
E003
E001
2555775
2559193
2562612
2566030
2569448
2572867
2576285
2579703
2583122
2586540

EMBL

Model

The expected count rate for exon *l* of gene *I* in sample *j* can be modelled as product of

- the baseline (control) expression strength of gene *i*
- the fraction of the reads from gene *i* that overlap with exon *l* under control condition
- the effect of the treatment of sample *j* on the expression strength of gene *I*
- the effect of the treatment of sample *j* on the fraction of the reads from gene *i* that overlap with exon *l*
- the sequencing depth (normalization factor) of sample *j*

EMBL

Model

$$K_{ijl} = NB\left(s_j \mu_{ijl}, \alpha_{il}\right)$$

counts in gene *i*, sample *j*, exon *l*

size factor

dispersion

$$\log \mu_{ijl} = \beta_i^0 + \beta_{il}^{\mathrm{E}} x_l^{\mathrm{E}} + \beta_{ij}^{\mathrm{T}} x_j^{\mathrm{T}} + \beta_{ijl}^{\mathrm{ET}} x_l^{\mathrm{E}} x_j^{\mathrm{T}}$$

expression strength in control

fraction of reads falling onto exon *l* in control

change in expression due to treatment

change to fraction of reads for exon *l* due to treatment

EMBL

Model, refined

$$K_{ijl} = NB\left(s_j \mu_{ijl}, \alpha_{il}\right)$$

counts in gene *i*, sample *j*, exon *l*

size factor

dispersion

$$\log \mu_{ijl} = \beta^{S}_{ij} + \beta^{\mathrm{E}}_{il} x^{\mathrm{E}}_{l} + \beta^{\mathrm{ET}}_{ijl} x^{\mathrm{E}}_{l} x^{\mathrm{T}}_{j}$$

expression strength in sample *j*

fraction of reads falling onto exon *l* in control

change to fraction of reads for exon *l* due to treatment

further refinement: fit an extra factor for library type (paired-end vs single)

EMBL

Dispersion estimation

- Standard maximum-likelihood estimates for dispersion parameters have very strong bias in case of small sample size.
- A method-of-moments estimator (as used in *DESeq*) cannot be used due to crossed factors.

- We take over the solution from the new *edgeR* version: Cox-Reid conditional-maximum-likelihood estimation

[Cox, Reid, J Roy Stat Soc B, 1987]
[McCarthy, Chen, Smyth, Nucl Acid Res, 2012]

EMBL

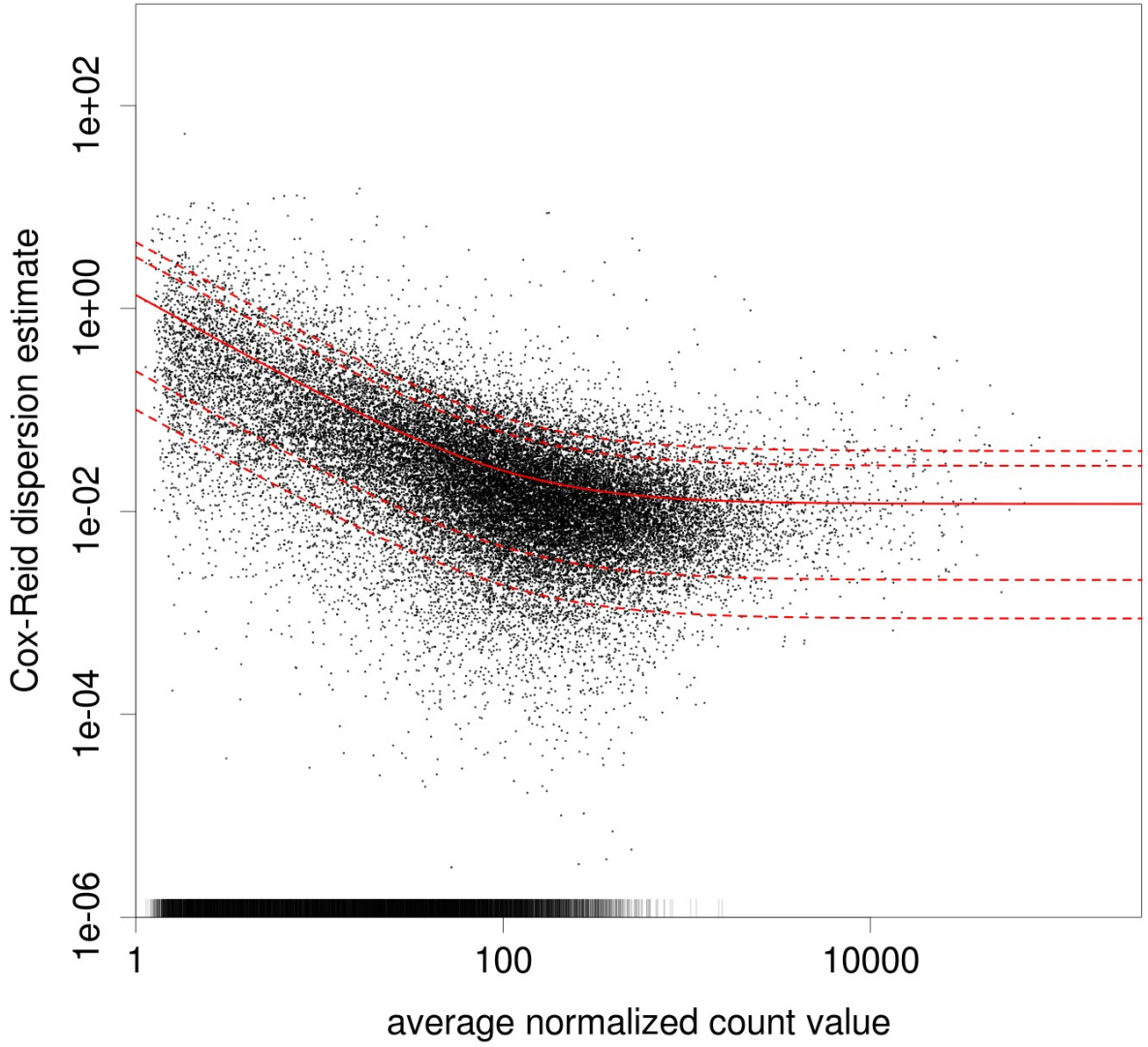
Dispersion estimation

Small sample size, so some data sharing is necessary to get power.

- one value fits all?
- one value for each gene?
- one value for each exon?

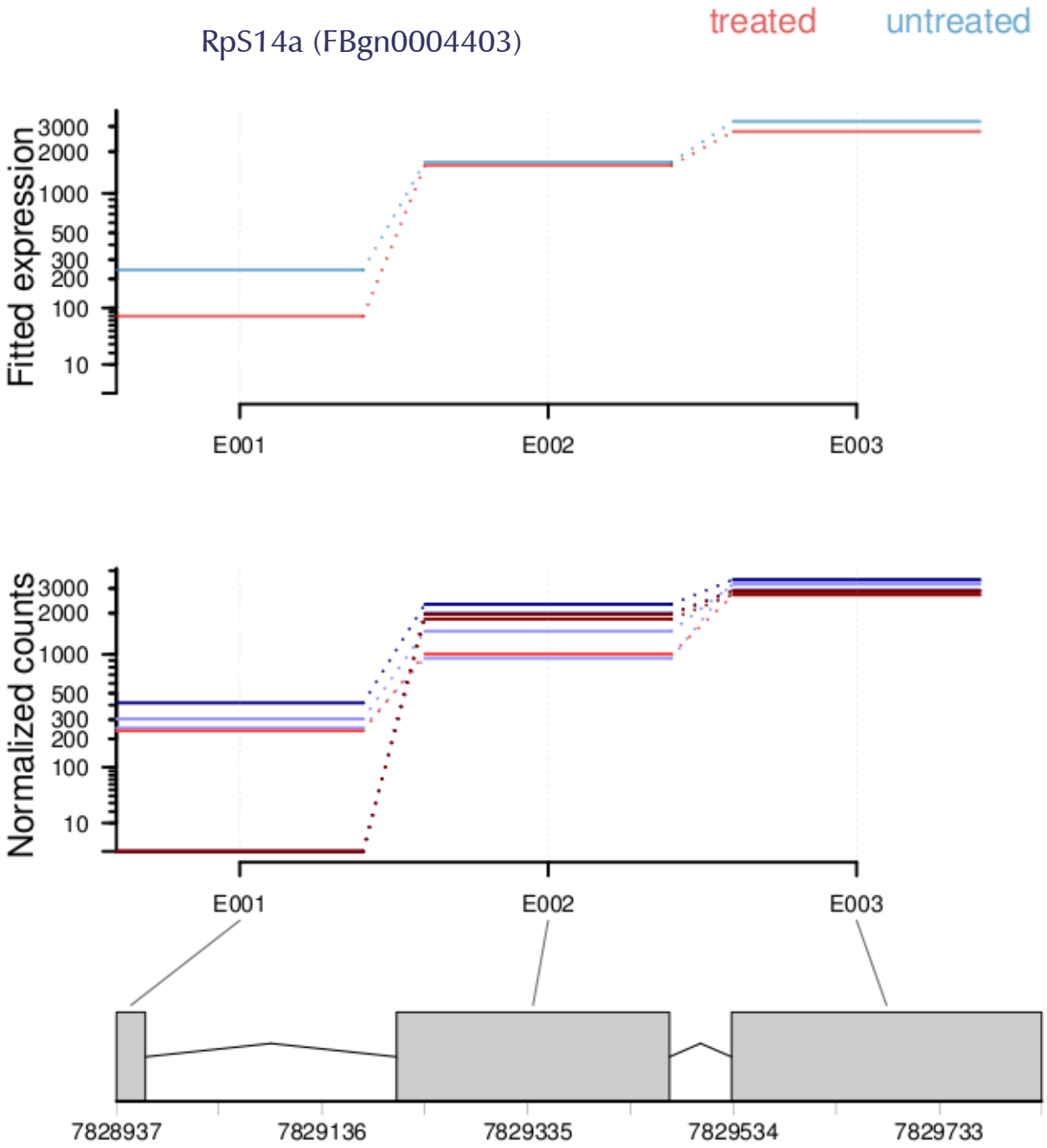
EMBL

Dispersion vs mean



EMBL

RpS14a (FBgn0004403)



EMBL

DEXSeq

- combination of Python scripts and an R package
- Python script to get counting bins from a GTF file
- Python script to get count table from SAM files
- R functions to set up model frames and perform GLM fits and ANODEV
- R functions to visualize results and compile an HTML report

- nearly ready for release

EMBL

DEXSeq

- combination of Python scripts and an R package
- Python script to get counting bins from a GTF file
- Python script to get count table from SAM files
- R functions to set up model frames and perform GLM fits and ANODEV
- R functions to visualize results and compile an HTML report

- nearly ready for release

EMBL

Conclusion II

- Counting within exons and NB-GLMs allows to study isoform regulation.
- Proper statistical testing allows to see whether changes in isoform abundances are just random variation or may be attributed to changes in tissue type or experimental condition.
- Testing on the level of individual exons gives power and might be helpful to study the mechanisms of alternative isoform regulation.

- DEXSeq is nearly ready for release.

EMBL

Acknowledgements

Coauthors:

- Alejandro Reyes
- Wolfgang Huber

Funding:

- EMBL

EMBL