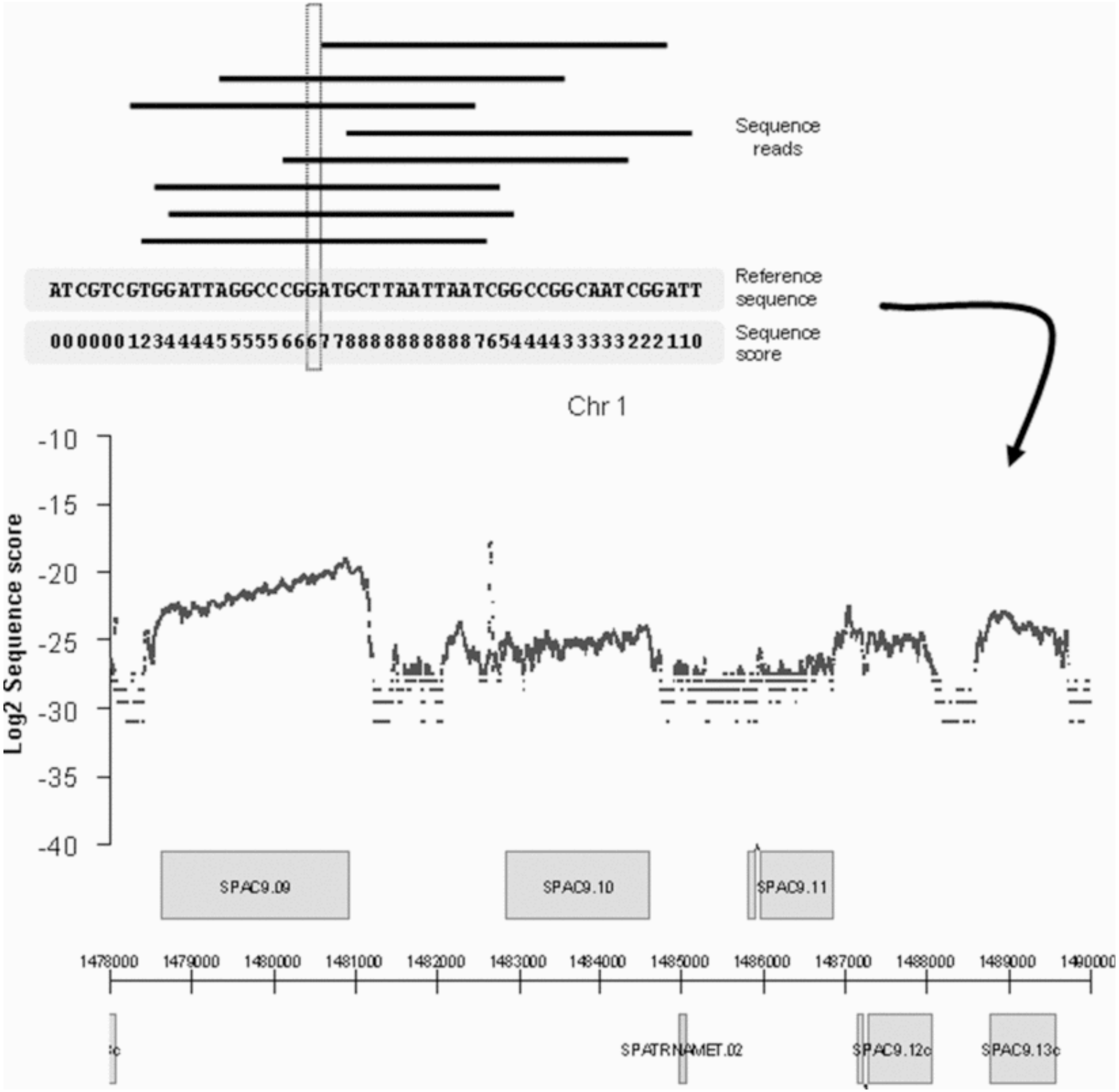
Differential expression for RNA-Seq

Wolfgang Huber

Genome Biology Unit, EMBL Heidelberg
and
European Bioinformatics Institute Cambridge, UK

RNA-Seq


Sequence reads
ATCGTCGTGGATTAGGCCCGGATGCTTAATTAATCGGCCGGCAATCGGATT
Reference sequence
000000123444455555666778888888888876544443333322211O
Sequence score
Chr 1
Log2 Sequence score
-10
-15
-20
-25
-30
-35
-40
SPAC9.09
SPAC9.10
SPAC9.11
1478000
1479000
1480000
1481000
1482000
1483000
1484000
1485000
1486000
1487000
1488000
1489000
1490000
SPATRNAMET.02
SPAC9.12c
SPAC9.13c

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

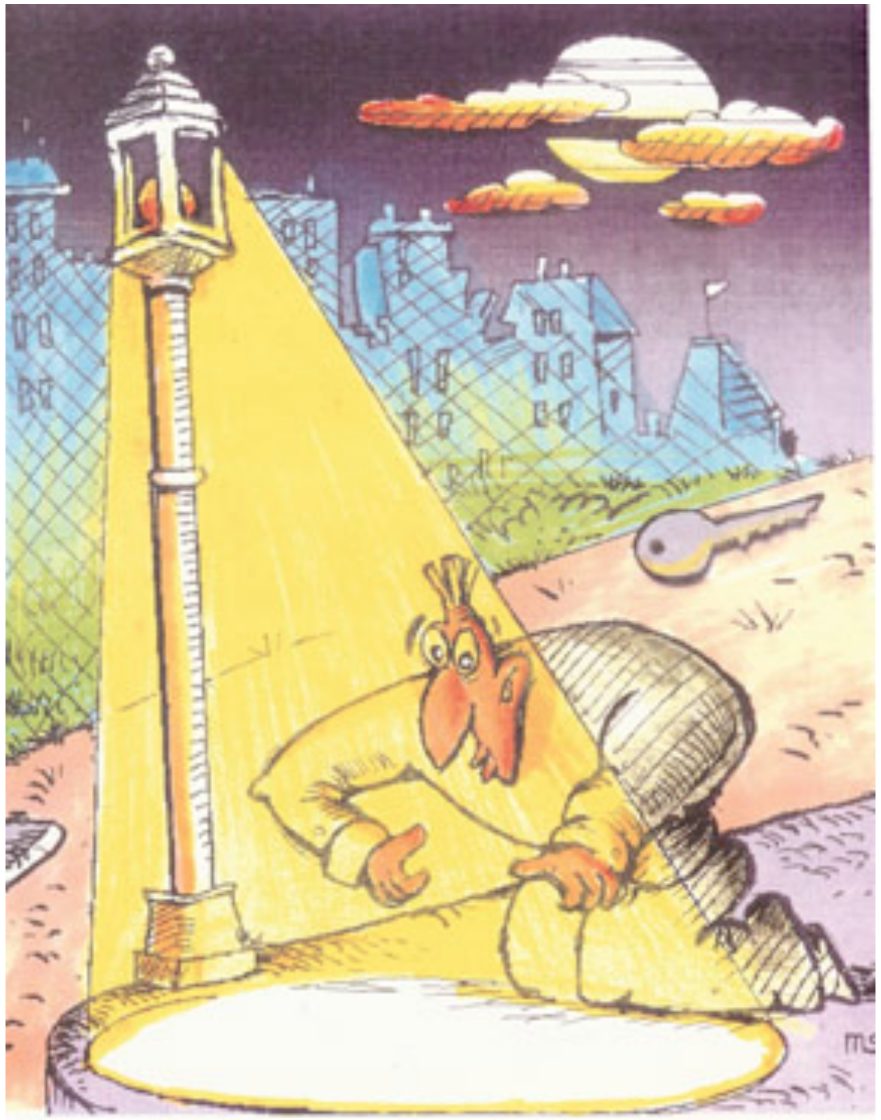
Should one align against the genome or the transcriptome?

against transcriptome

- **easier, because no gapped alignment necesssary**

but:

- **risk to miss possible alignments!**



Count data in HTS

- RNA-Seq
- Tag-Seq

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

[...]

- ChIP-Seq
- Bar-Seq
- ...

Counting rules

- Count reads, not nucleotides
- 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

Counting rules

- Count reads, not nucleotic
- Count each read at most c
- Discard a read if
  - it cannot be uniquely ma
  - its alignment overlaps w
  - the alignment quality sc
  - (for paired-end reads) th
    the same gene

Challenges with count data from high-throughput sequencing

discrete, positive, skewed

➡ no (log-)normal model

small numbers of replicates

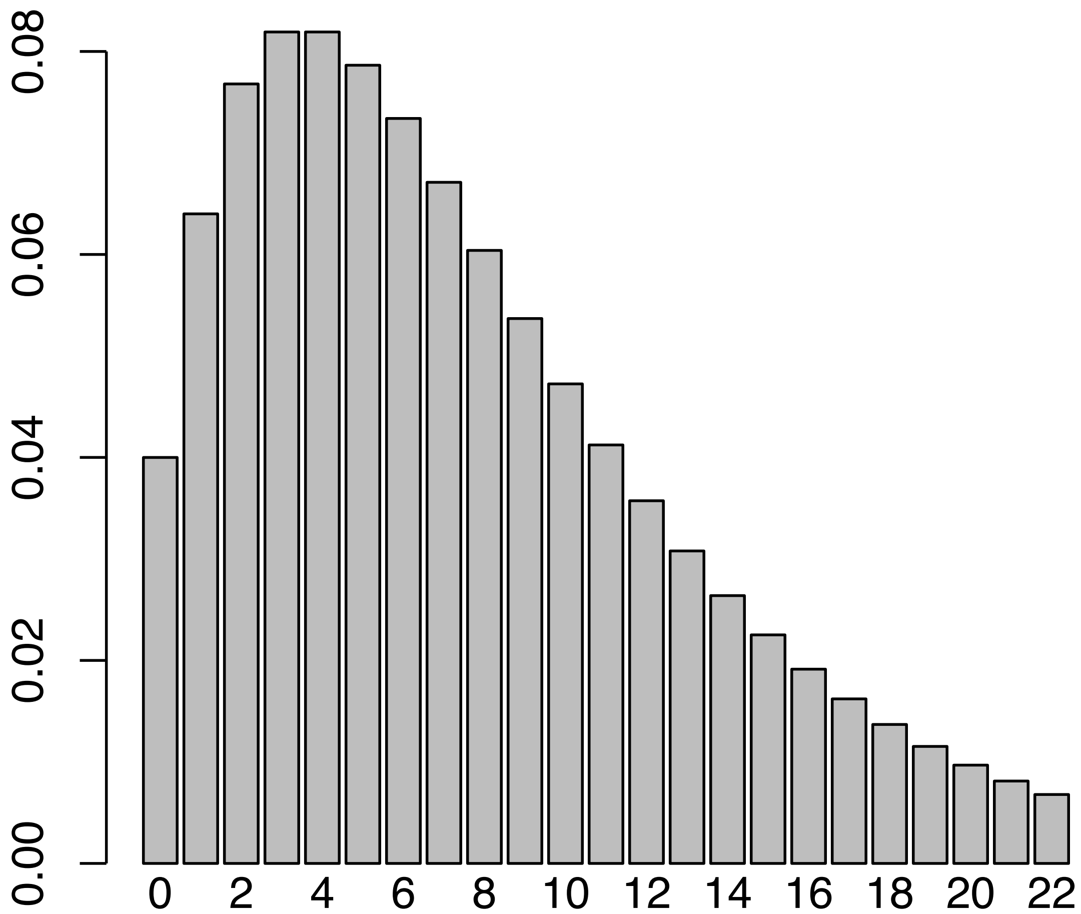
➡ no rank based or permutation methods

large dynamic range (0 ... 10^5)

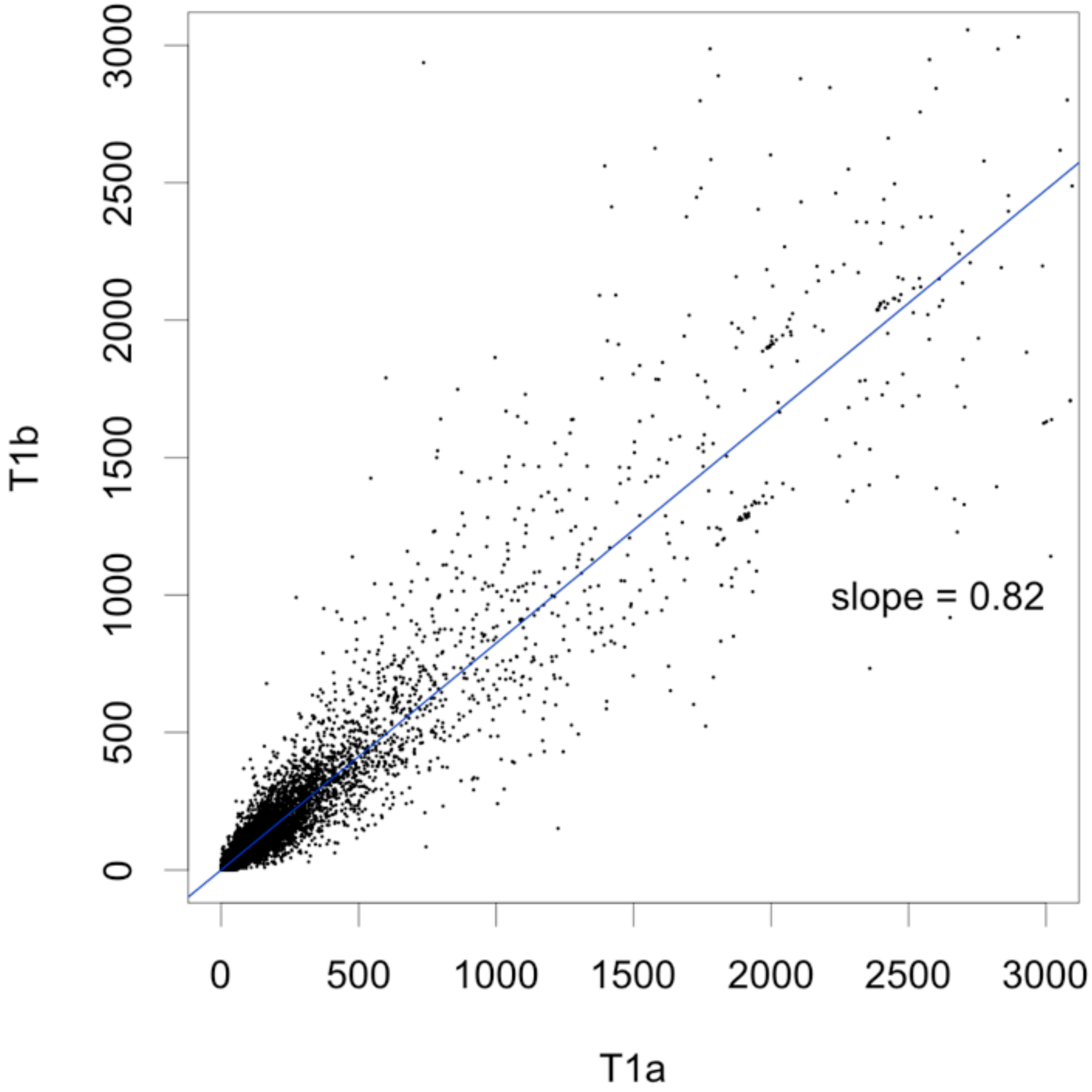
➡ heteroskedasticity matters

sequencing depth (coverage) varies between samples

➡ ”normalisation”



sequencing depth (library size) effect



Normalisation for library size

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

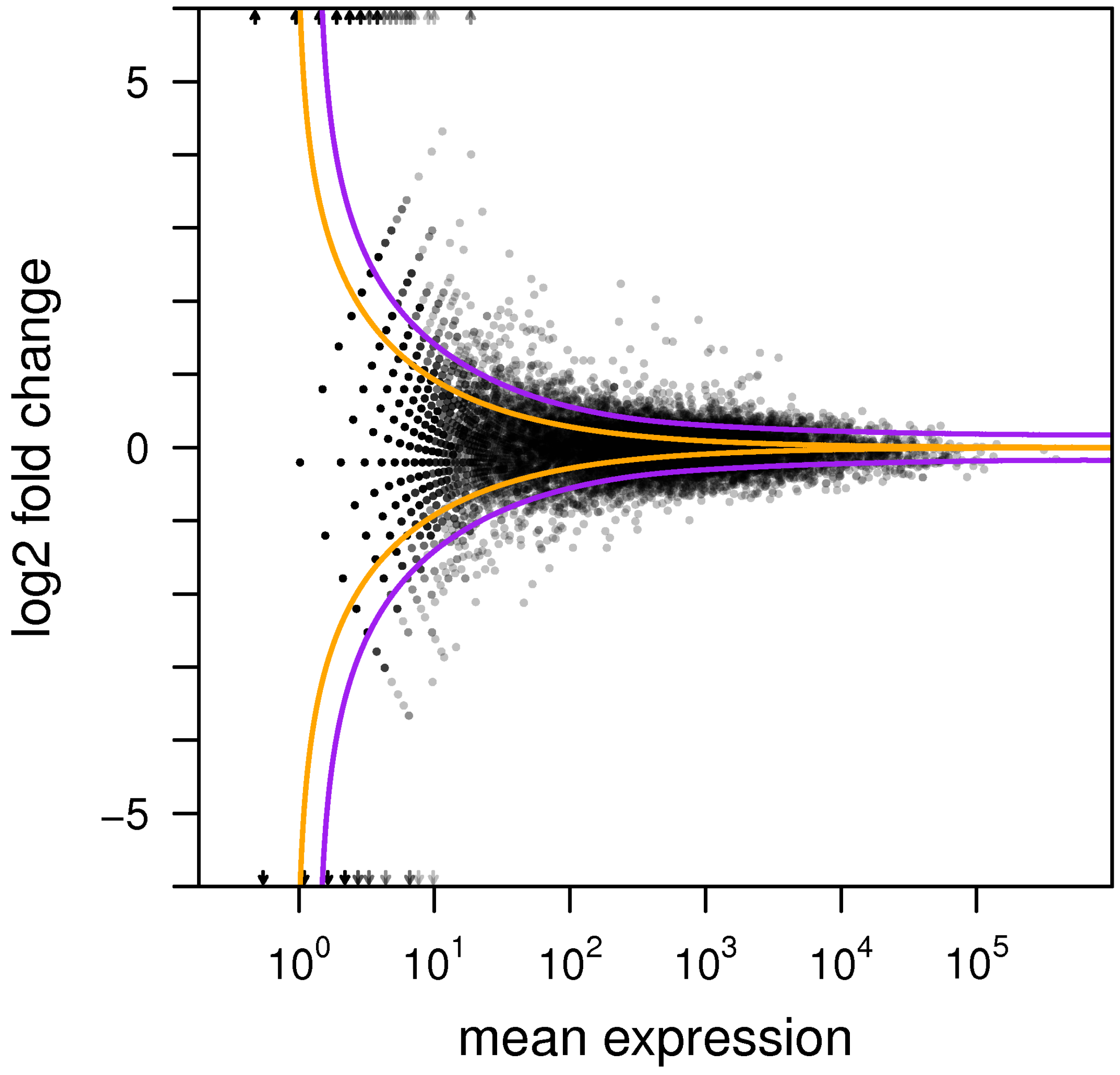
- Simply using the total number of reads per sample is not a good idea; genes that are strongly and differentially expressed may distort the ratio of total reads.

- 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.

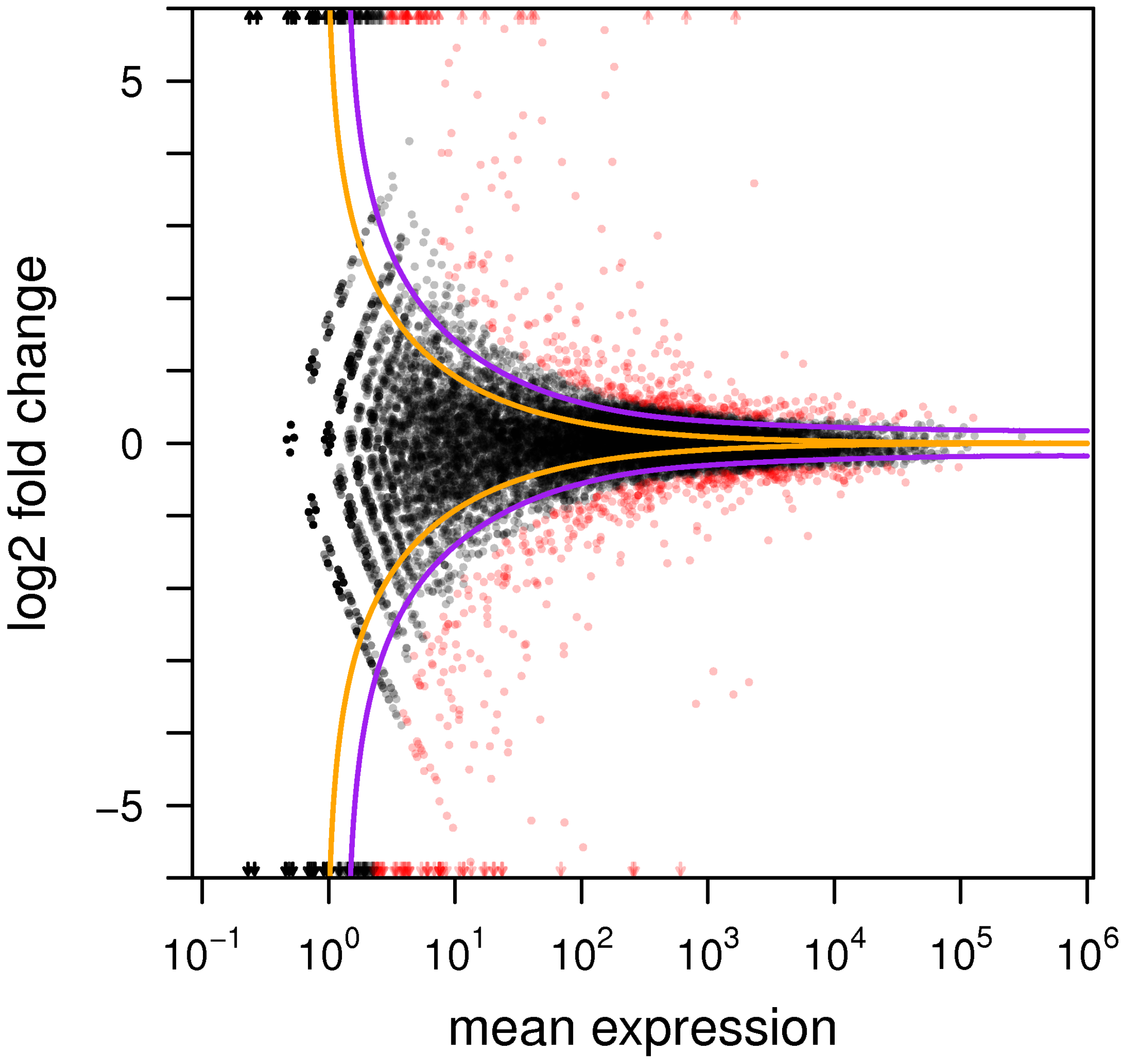
Anders & Huber, Genome Biology 2010 (DESeq package)

Sample-to-sample variation

comparison of two replicates



comparison of treatment vs control

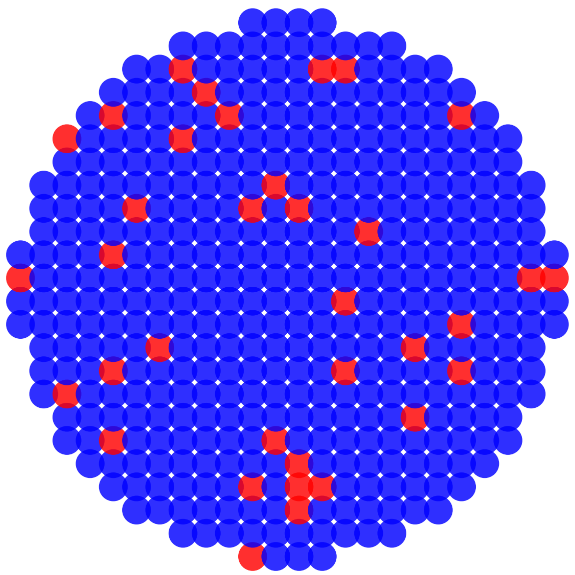


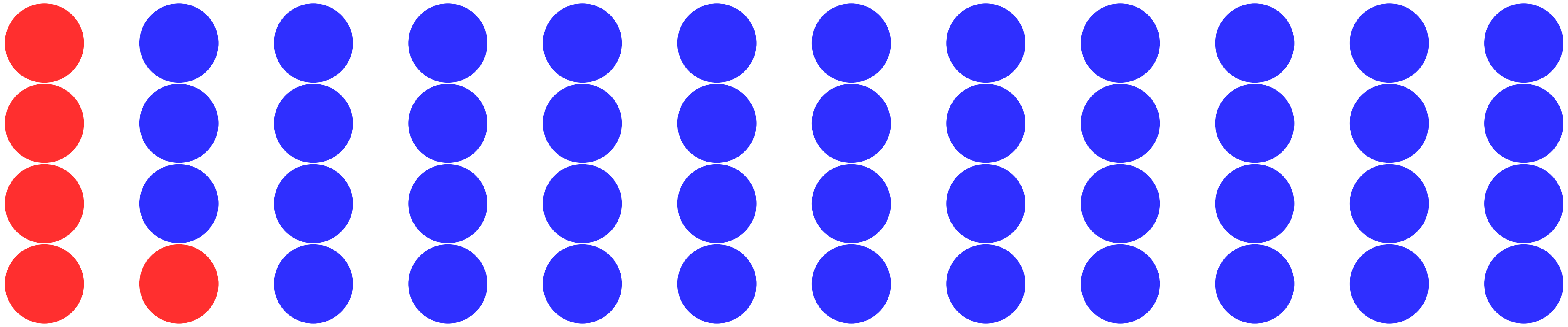
The Poisson distribution

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

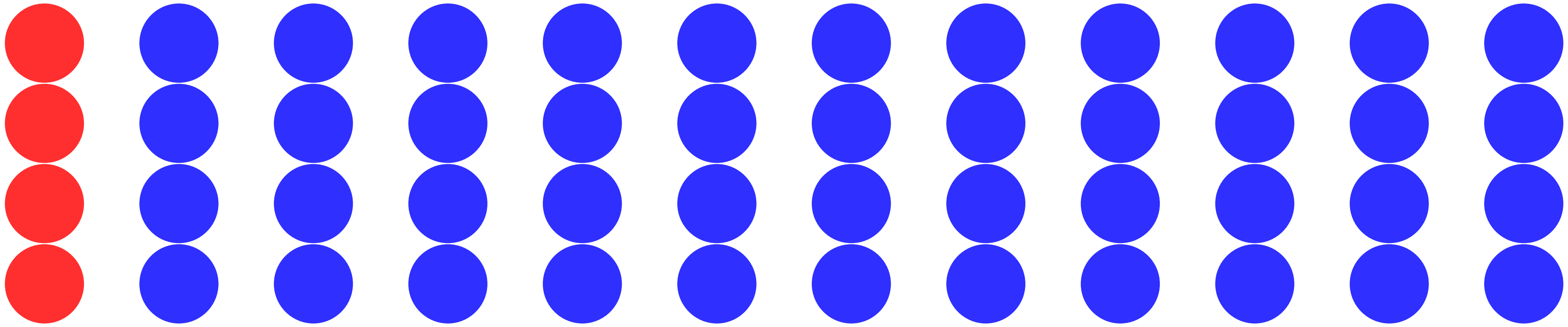
Several experimenters are tasked with determining the percentage of red balls.

Each of them is permitted to draw 50 balls out of the bag, without looking.

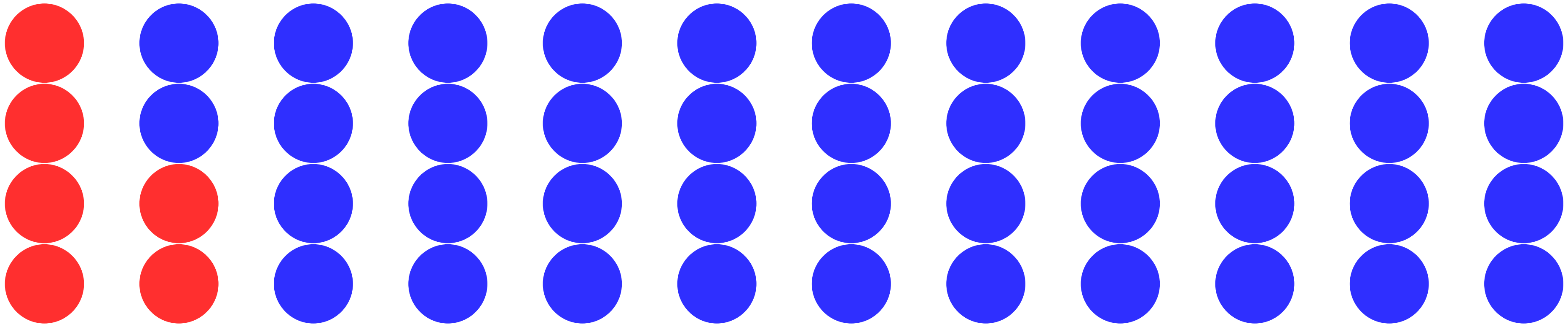




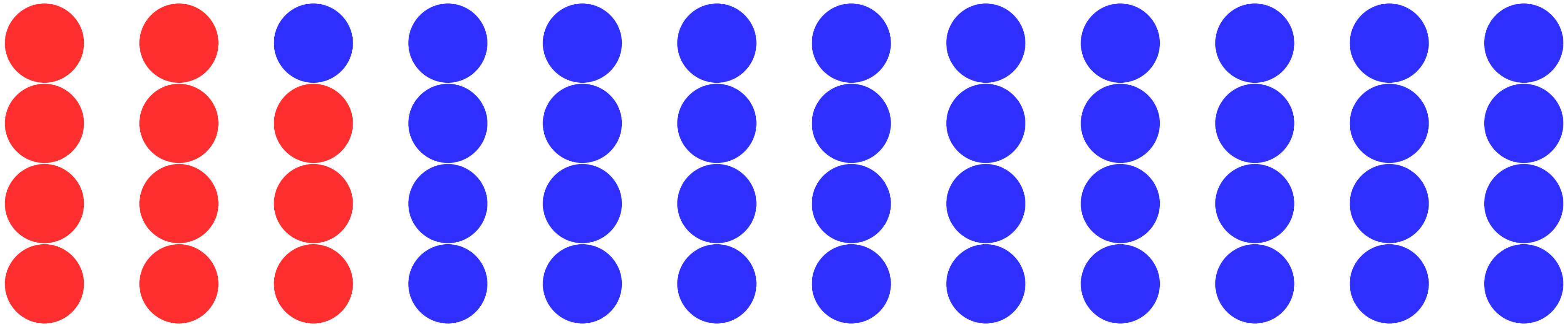
5 / 50 = 10%



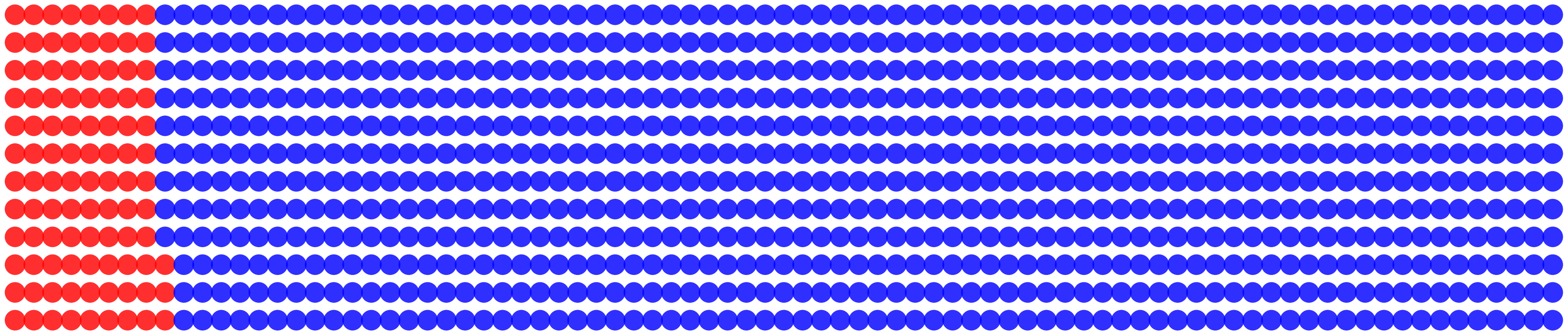
4 / 50 = 8%



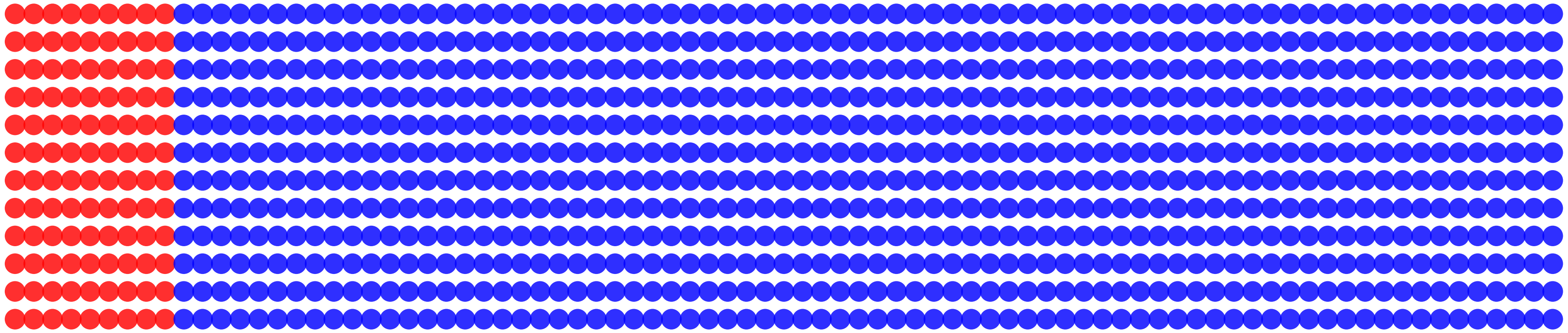
6 / 50 = 12%



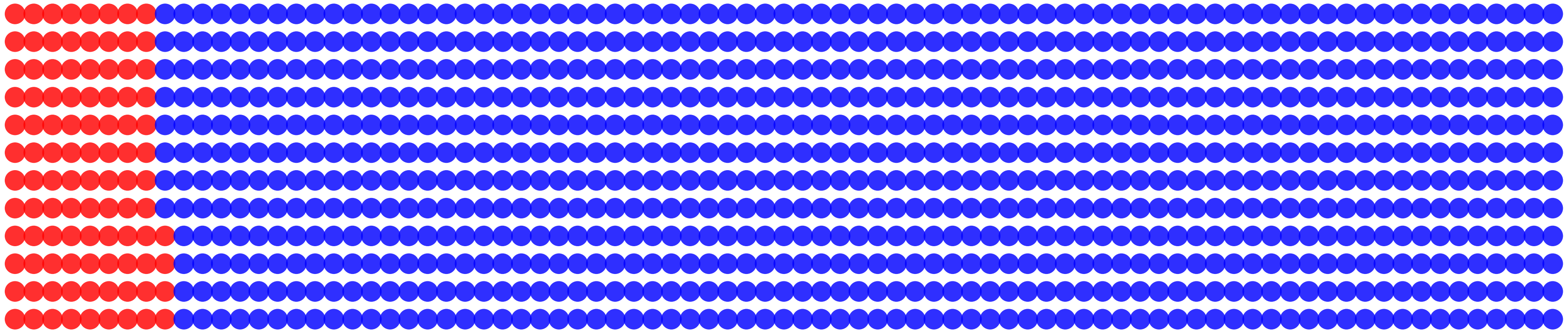
11 / 50 = 22%



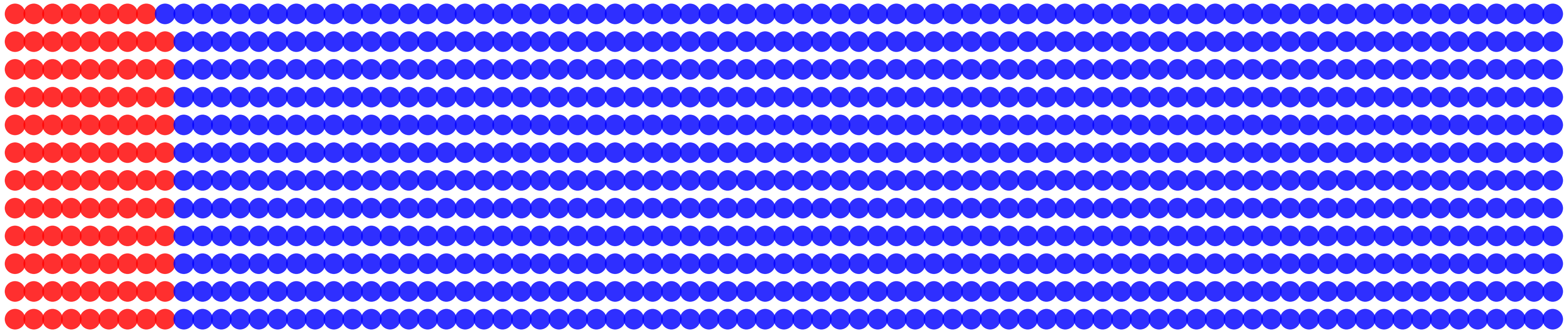
99/1000 = 9.9%



108/1000 =
10.8%



100/1000 =
10.0%

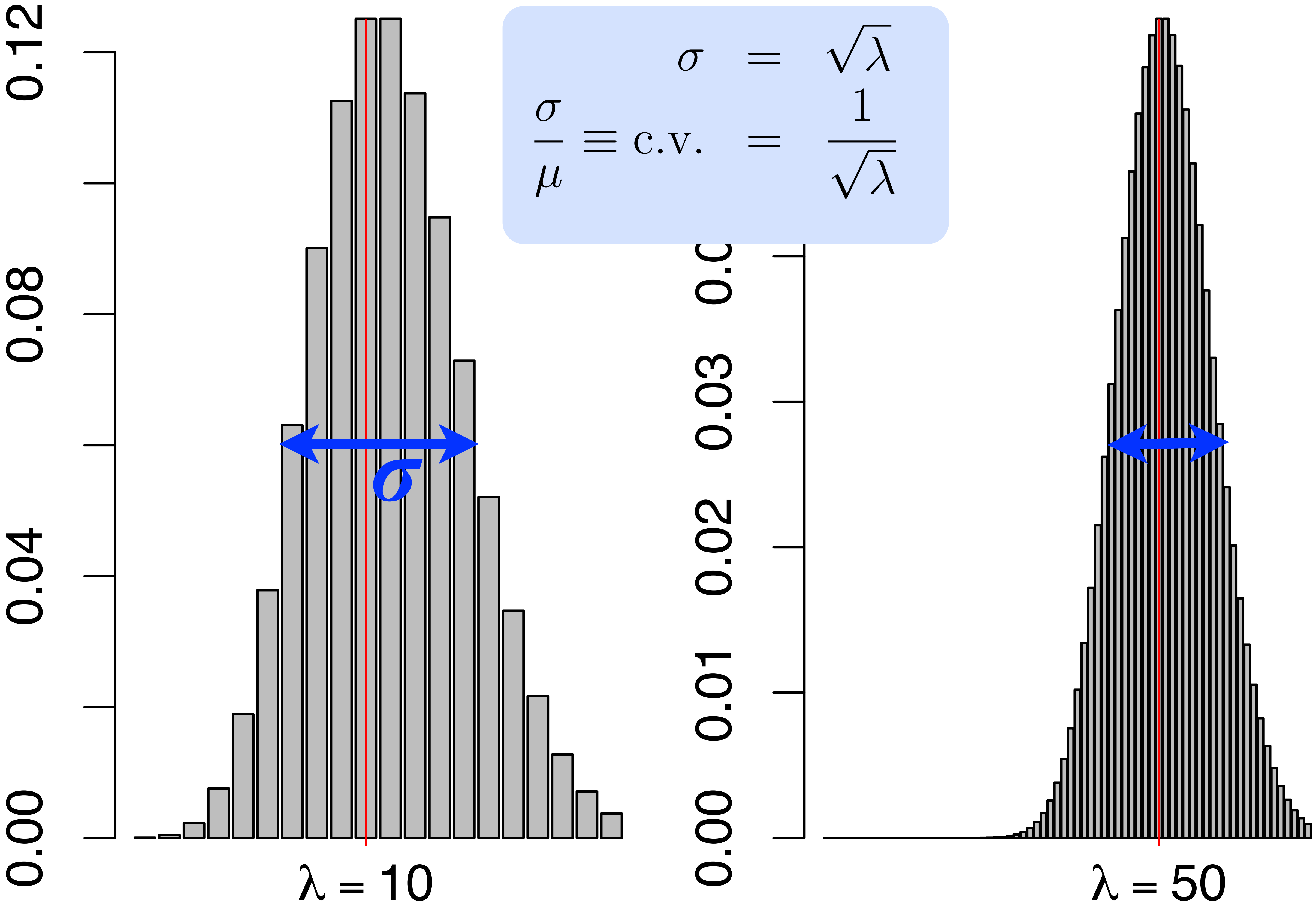


107 / 1000 =
10.7%

Poisson distribution:
the uncertainty of random sampling

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

The Poisson distribution is used for counting processes



$$\sigma = \sqrt{\lambda}$$

$$\frac{\sigma}{\mu} \equiv \text{c.v.} = \frac{1}{\sqrt{\lambda}}$$

Analysis method: ANOVA

$$N_{ij} \quad \sim \quad \text{Poisson}(\mu_{ij})$$

Noise part

$$\log \mu_{ij} \quad = \quad s_j + \sum_k \beta_{ik} x_{kj}$$

Systematic part

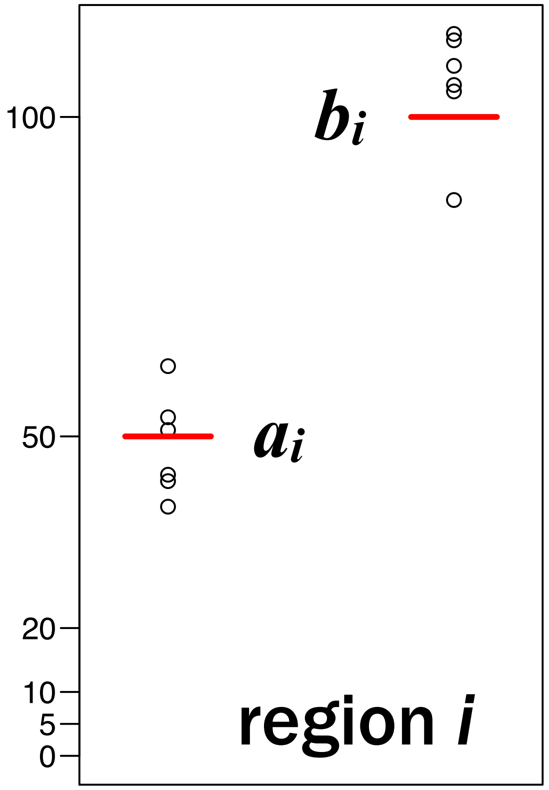
μ_{ij} **expected count of region *i* in sample *j***
s_j library size effect
x_{kj} **design matrix**
β_{ik} **(differential) effect for region *i***

Analysis method: ANOVA

$$N_{ij} \quad \sim \quad \text{Poisson}(\mu_{ij})$$

Noise part

$$\mu_{ij} = s_j \times \begin{cases} a_i & \text{if } j \in \text{group A} \\ b_i & \text{if } j \in \text{group B} \end{cases}$$



μ_{ij} expected count of region *i* in sample *j*

s_j library size effect

x_{kj} design matrix

β_{ik} (differential) effect for region *i*

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

No need to estimate the variance. This is convenient.

E.g. Marioni et al. (2008), Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010), …

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

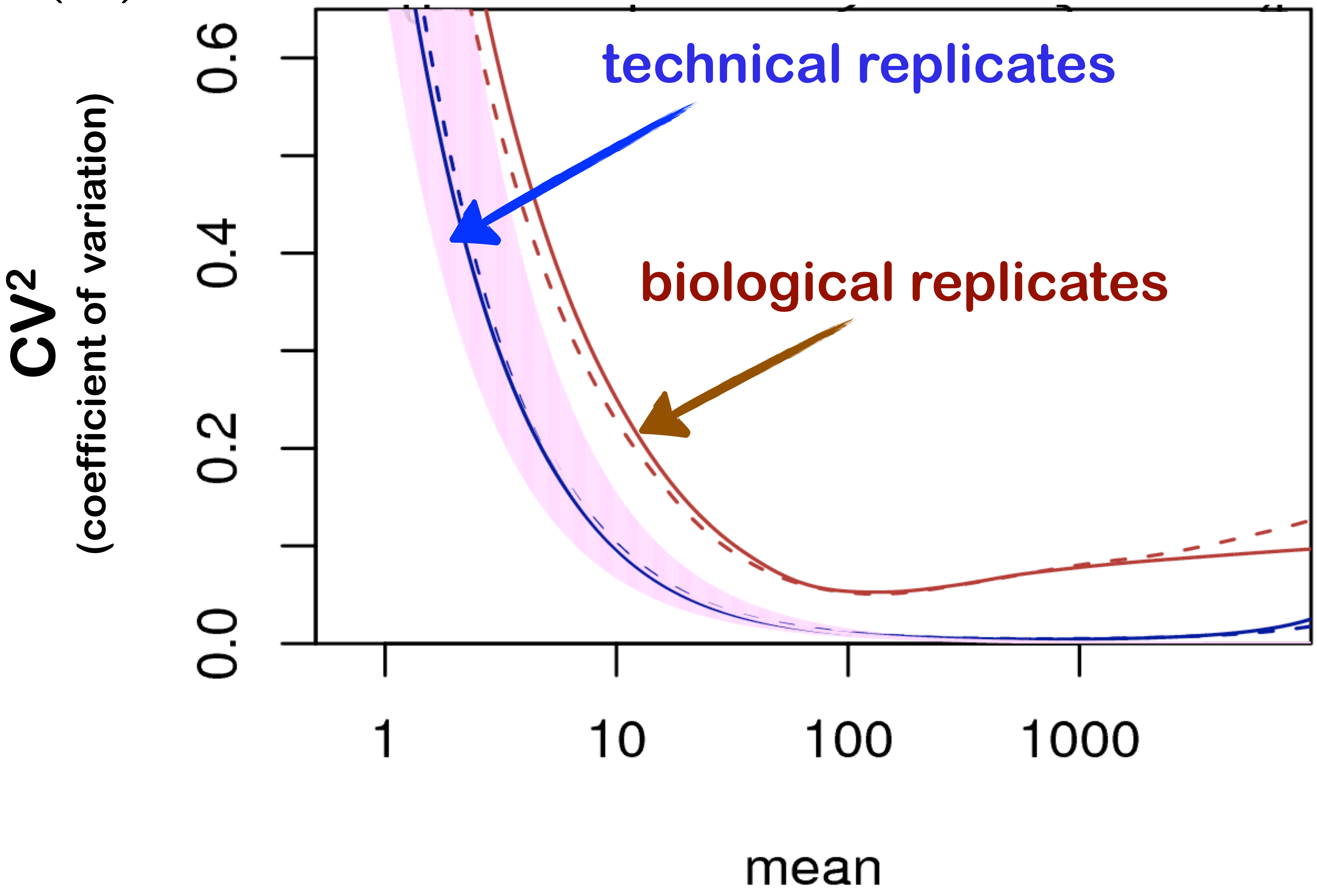
No need to estimate the variance. This is convenient.

E.g. Marioni et al. (2008), Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010), ...

Really?
Are HTS count data Poisson distributed?

To figure this out, we have to take a closer look at replicates and the nature of the noise in the data.

$$\left(\frac{\sigma}{\mu}\right)^2$$

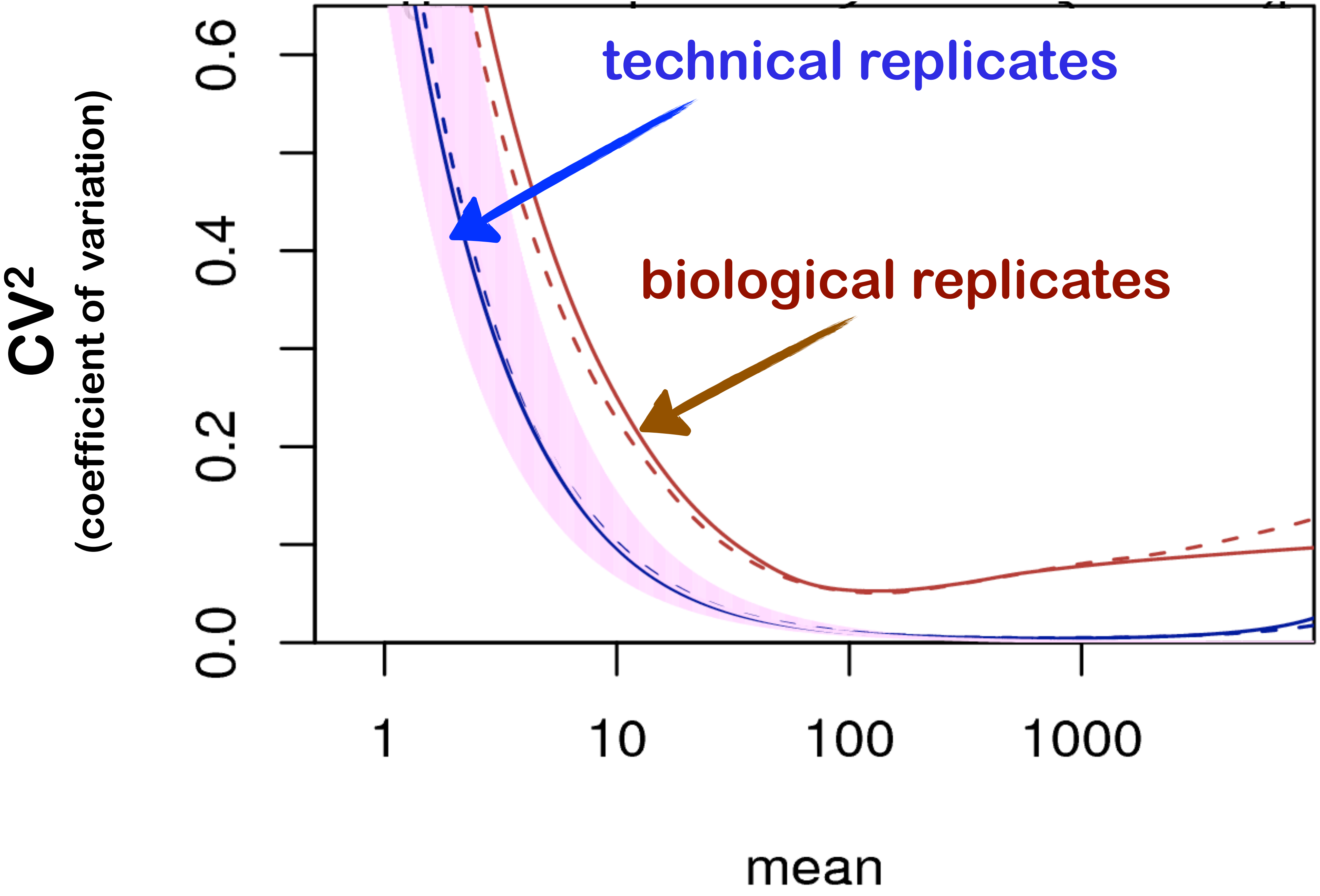


technical replicates

biological replicates

Based on the data of Nagalakshmi et al., Science 2008

$$\left(\frac{\sigma}{\mu}\right)^2$$



technical replicates

biological replicates

Much larger than Poisson

Consistent with Poisson

Based on the data of Nagalakshmi et al., Science 2008

So we need a better model

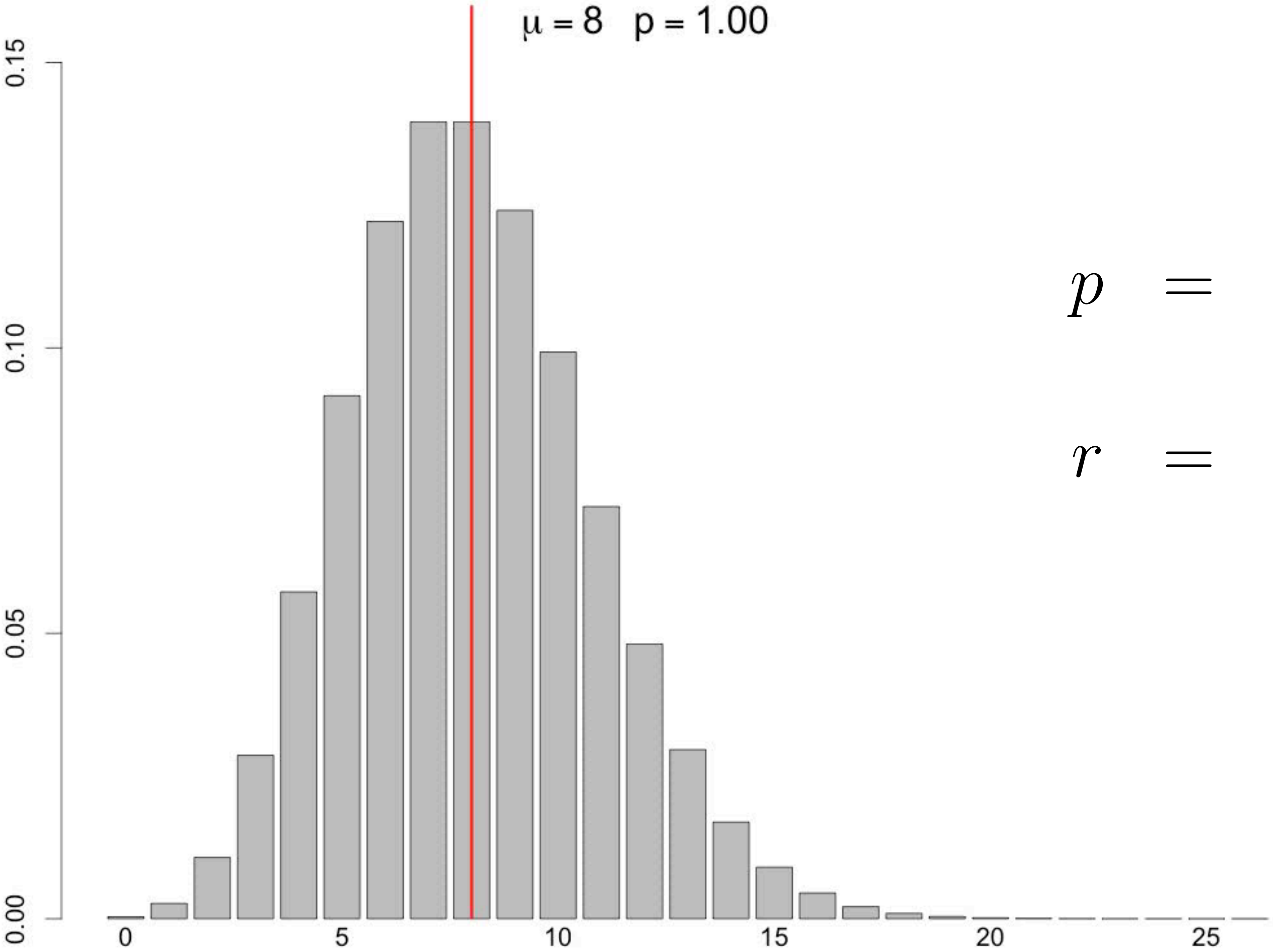
data are discrete, positive, skewed
➡ no (log-)normal model

small numbers of replicates
➡ no rank based or permutation methods
➡ want to use parametric stochastic model to infer tail behaviour (approximately) from low-order moments (mean, variance)

large dynamic range (0 ... 10^5)
➡ heteroskedasticity matters

Model building block I: the negative-binomial distribution

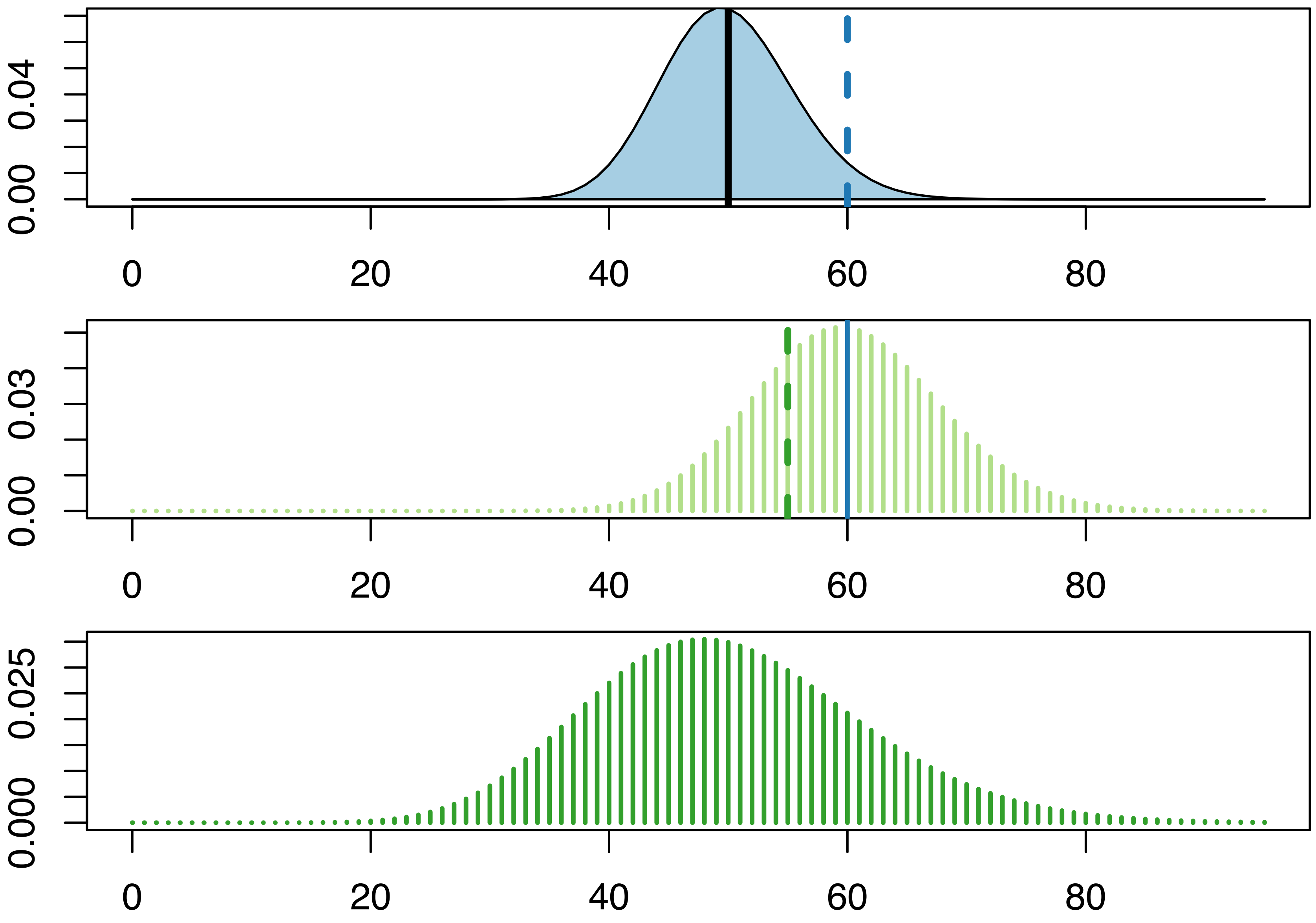
$$\mathrm{P}(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \qquad r \in \mathbb{R}^+, \ p \in [0, 1]$$



$$p = \frac{\mu}{\sigma^2}$$ **overdispersion**

$$r = \frac{\mu^2}{\sigma^2 - \mu}$$ **location**

The NB distribution is used when the rate of a Poisson process is itself randomly varying



Biological sample to sample variability Γ

⇓

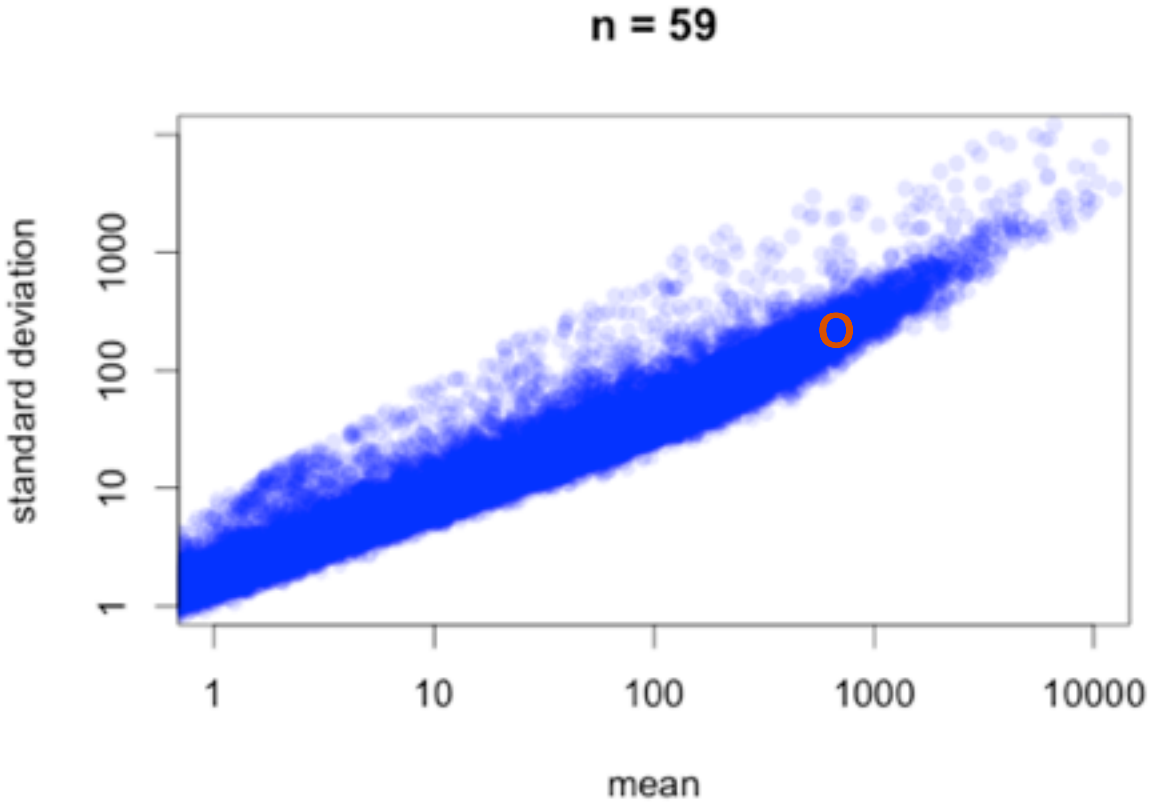
Poisson counting statistics Λ

⇓

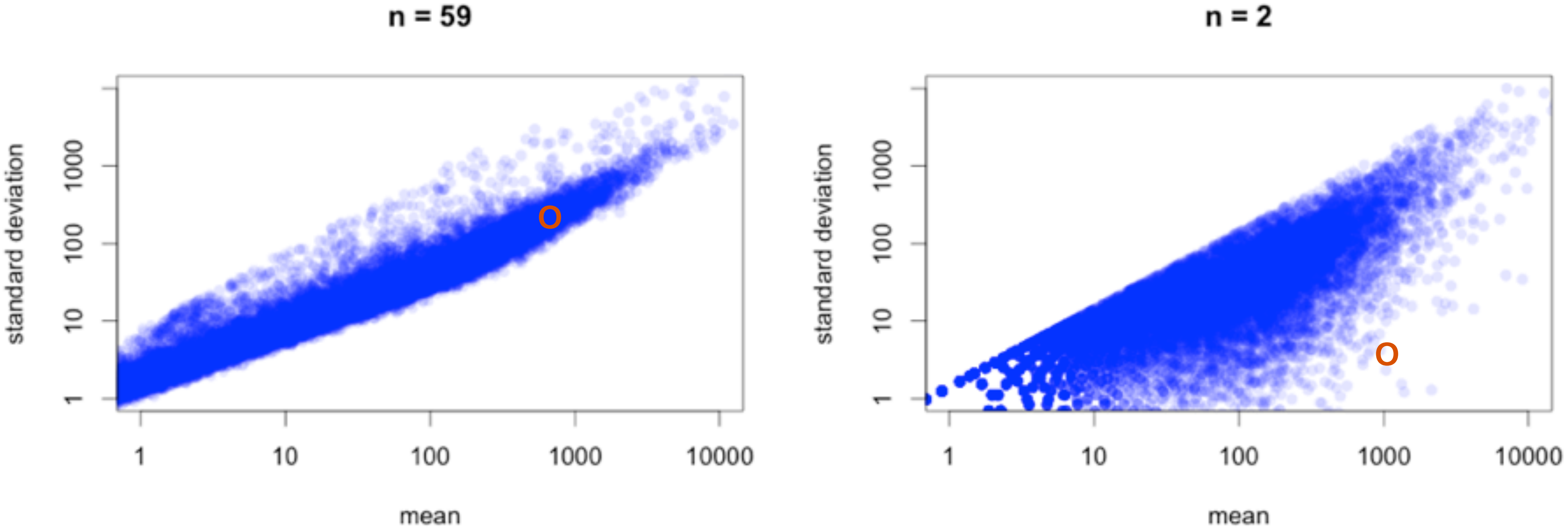
Overall distribution NB

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

Model building block II: variance regularisation and local regression on the mean

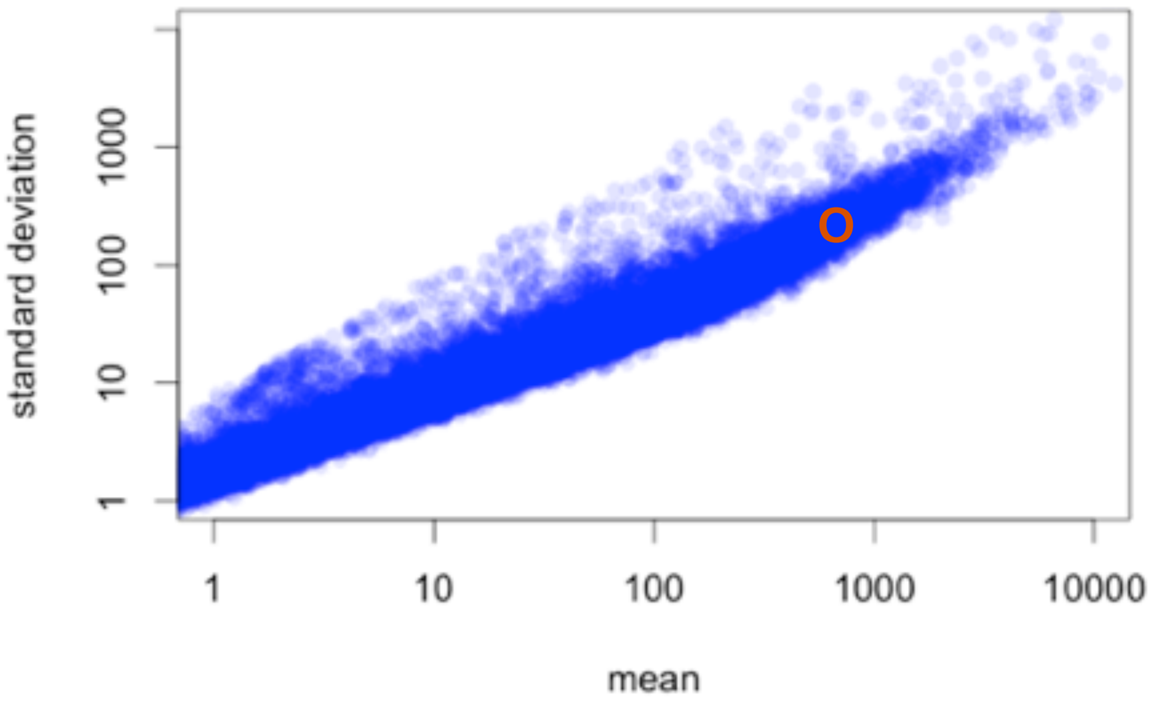


Model building block II: variance regularisation and local regression on the mean

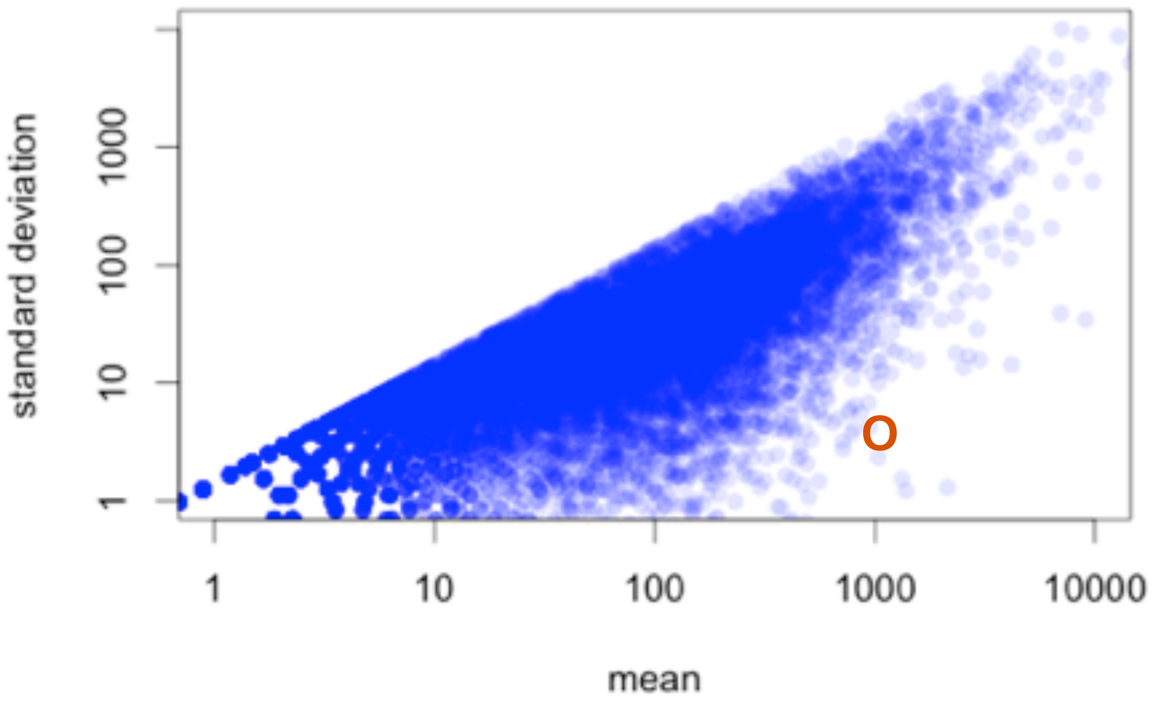


Model building block II: variance regularisation and local regression on the mean

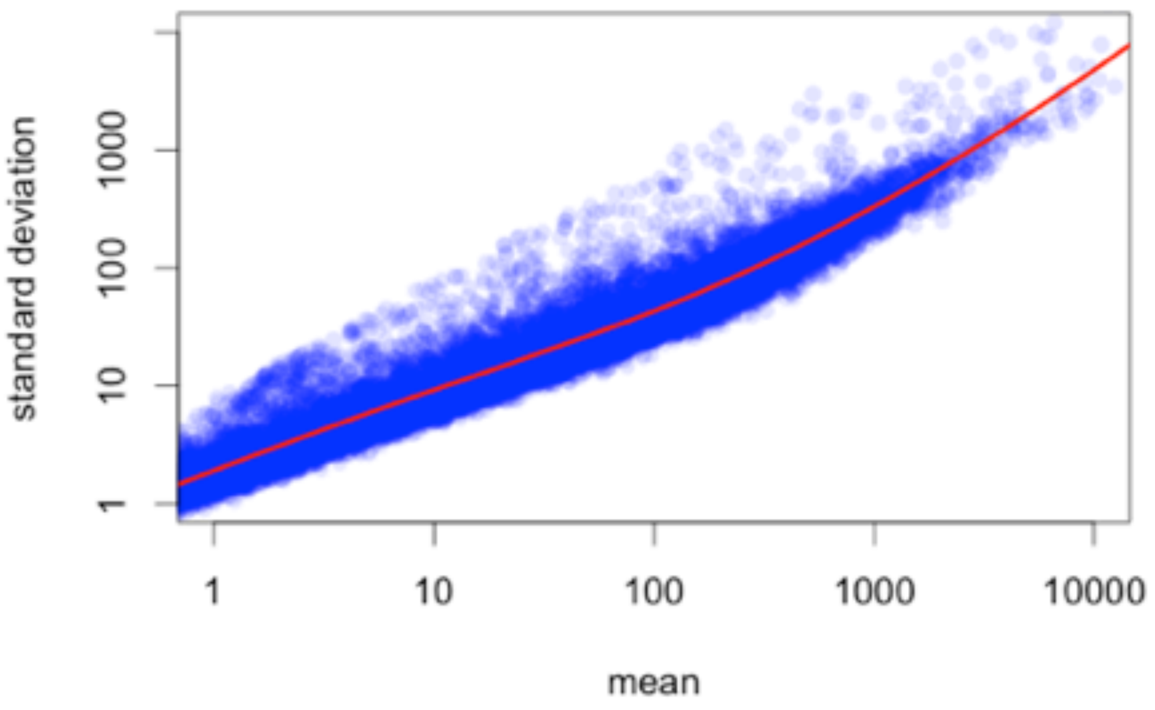
n = 59



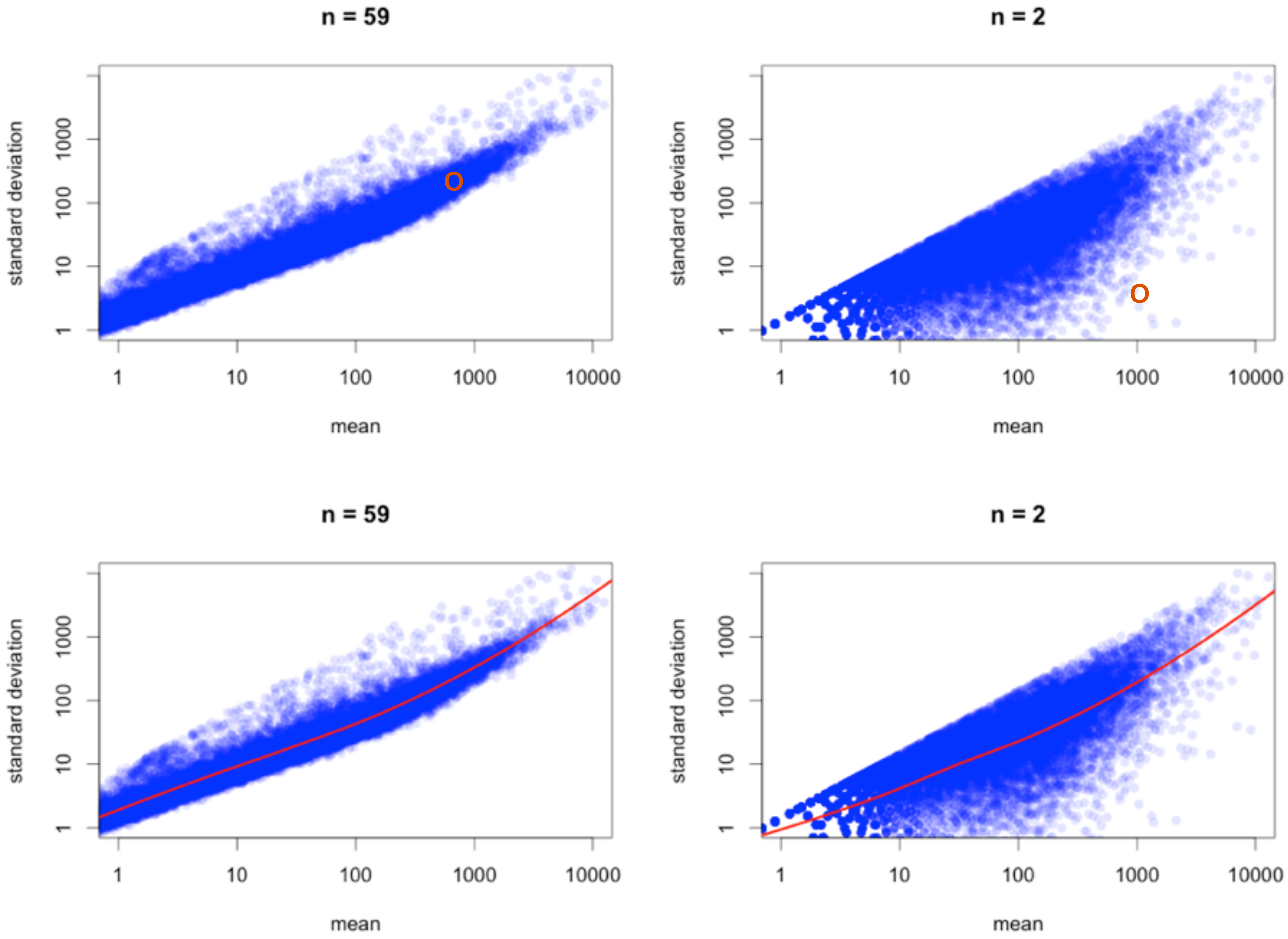
n = 2



n = 59



Model building block II: variance regularisation and local regression on the mean



n = 59

n = 2

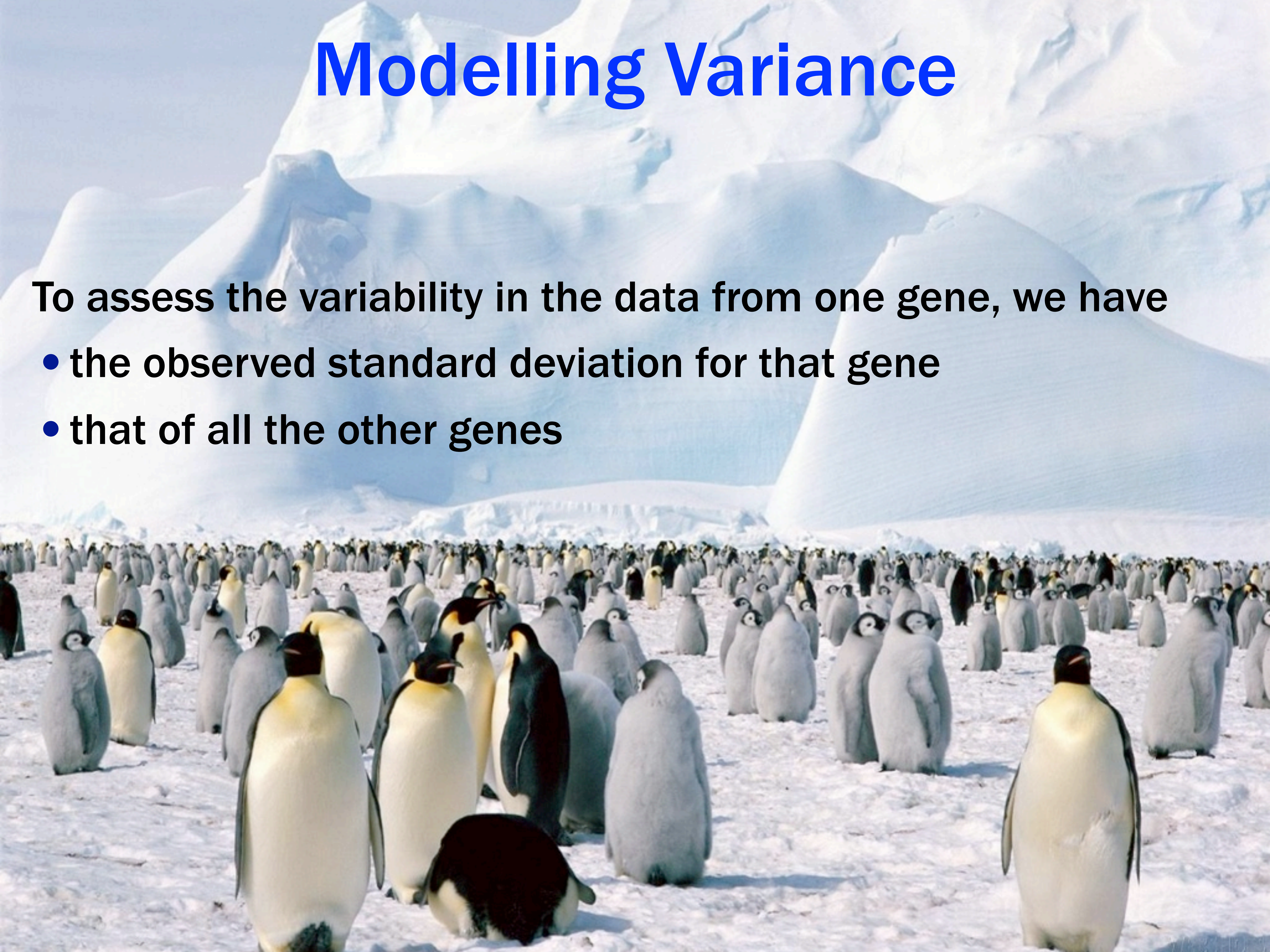
n = 59

n = 2

Modelling Variance

To assess the variability in the data from one gene, we have

- the observed standard deviation for that gene
- that of all the other genes



Putting it all together

$$N_{ij} \quad \sim \quad \mathrm{Poisson}(\mu_{ij})$$

Noise part

$$\log \mu_{ij} \quad = \quad s_j + \sum_k \beta_{ik} x_{kj}$$

Systematic part

μ_{ij} **expected count of gene *i* in sample *j***

s_j **library size effect**

x_{kj} **design matrix**

β_{ik} **(differential) expression effects for gene *i***

Putting it all together

$$N_{ij} \quad \sim \quad \mathrm{NB}(\mu_{ij}, \alpha(\mu_{ij}))$$

Noise part

$$\log \mu_{ij} \quad = \quad s_j + \sum_k \beta_{ik} x_{kj}$$

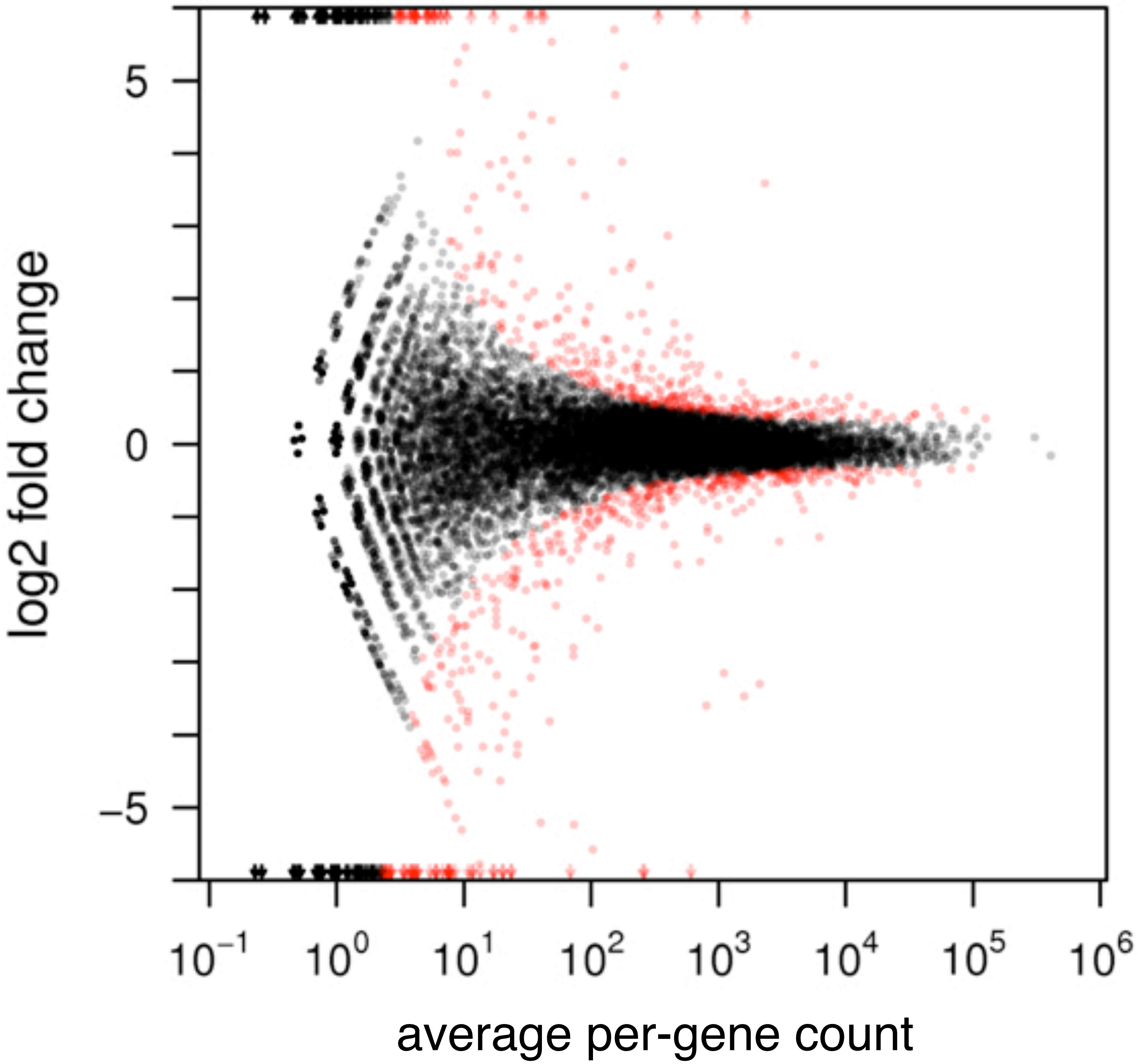
Systematic part

μ_{ij}	expected count o
s_j	library size effec
x_{kj}	design matrix
β_{ik}	(differential) exp

Generalised linear model of the negative binomial family with smooth dispersion-mean relation α

The DESeq package

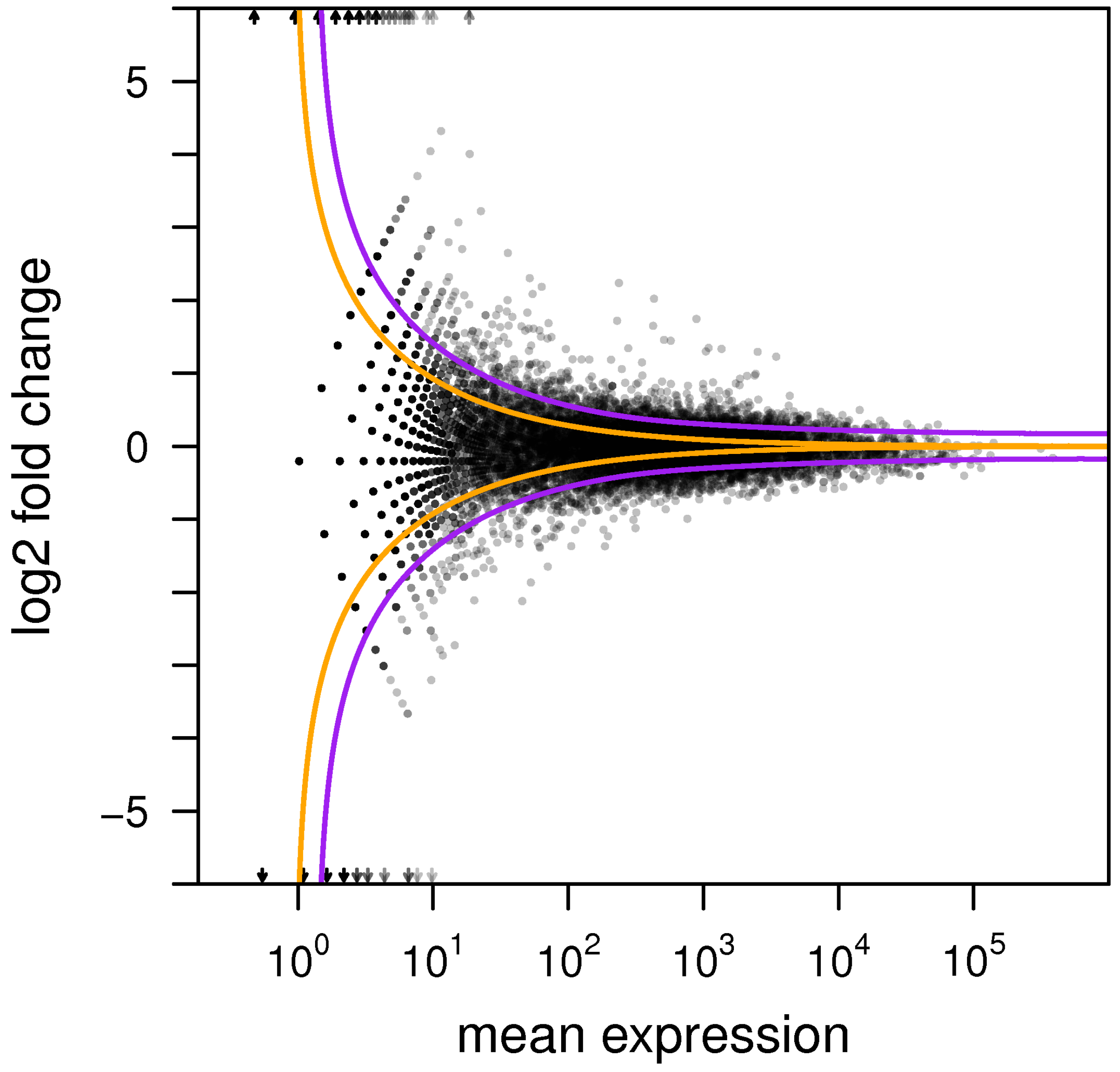
Negative binomial error modeling with intensity dependent dispersion



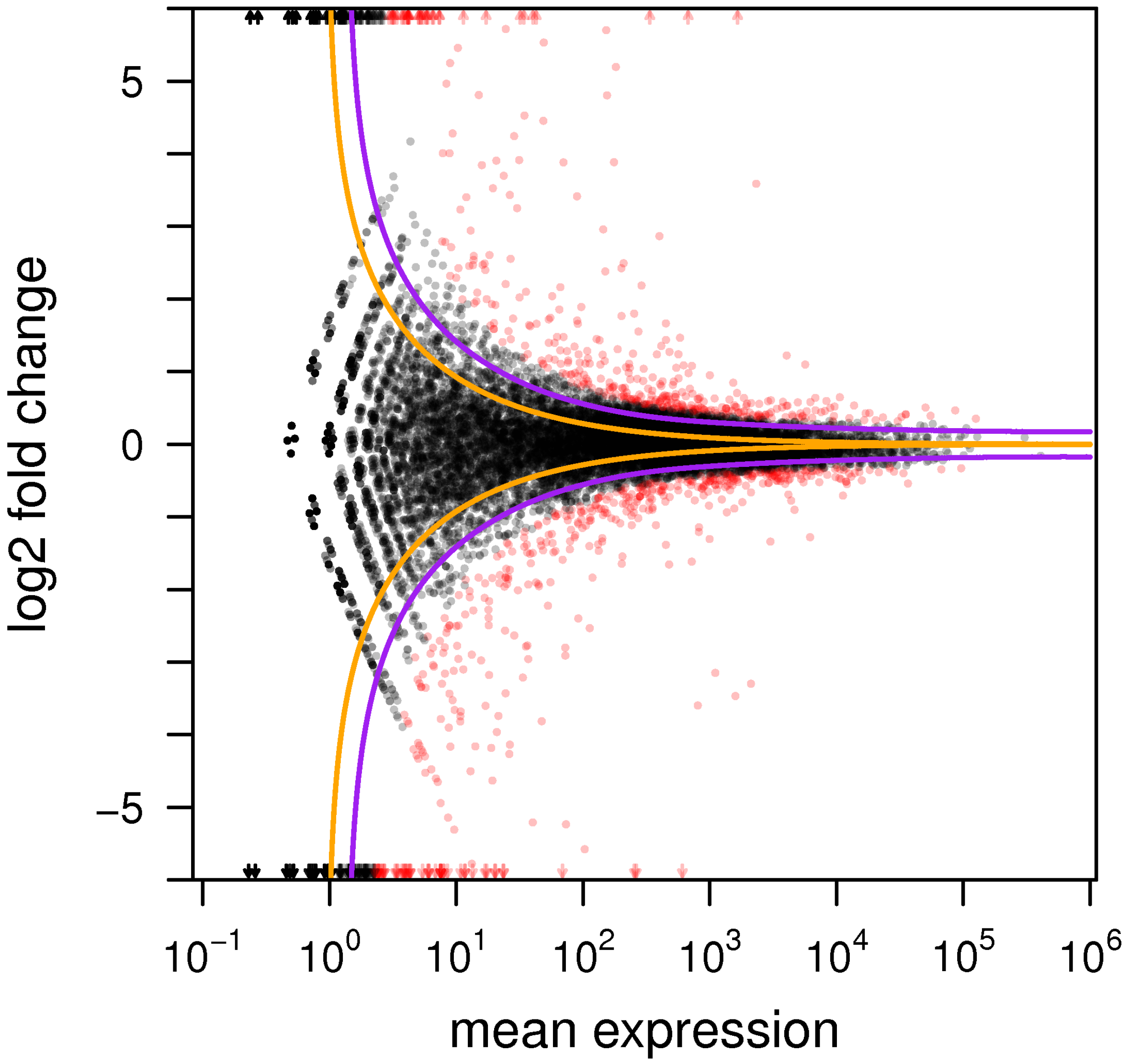
Anders and Huber, Genome Biol. 2010

Type-I error control

comparison of two replicates



comparison of treatment vs control



Two component noise model aids experimental design

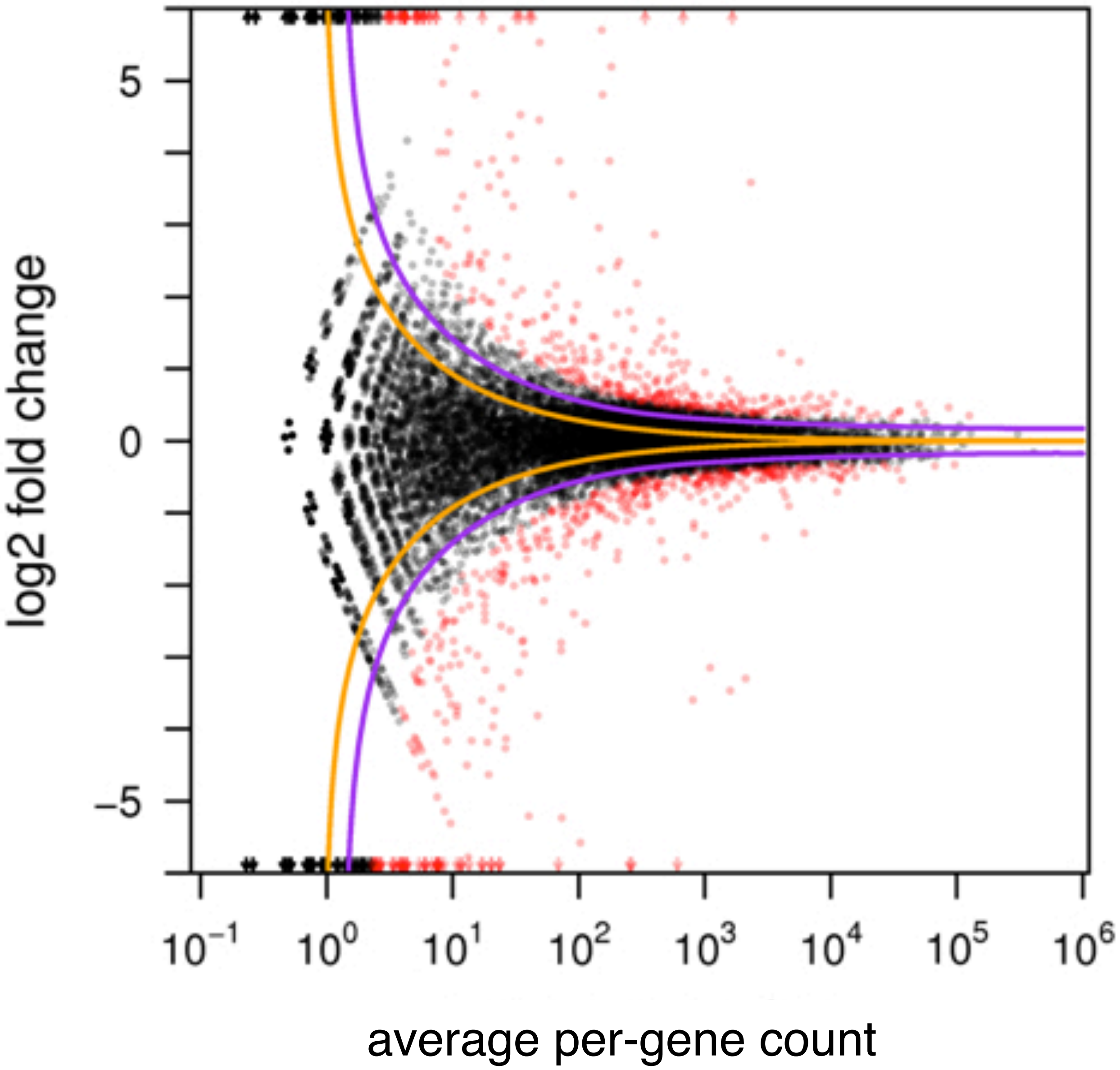
$$\mathrm{var} = \mu + c\,\mu^2$$

shot noise (Poisson)   biological noise

Small counts

Sampling noise dominant

Improve power: deeper coverage



Large counts

Biological noise dominant

Improve power: more biol. replicates

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*) in Bioconductor.

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.

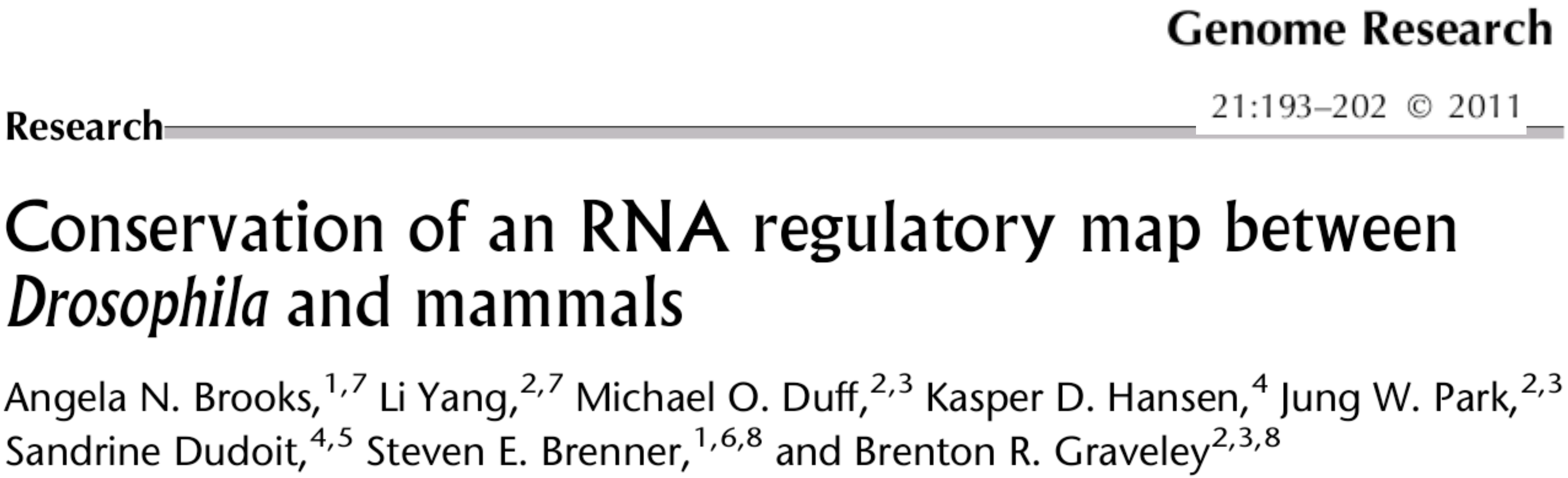
This is not yet possible with existing tools.

Regulation of isoform abundance

- 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?

- *DEXSeq*, a tool to test for *differential exon usage* in RNA-Seq data - see labs.

Data set used 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:

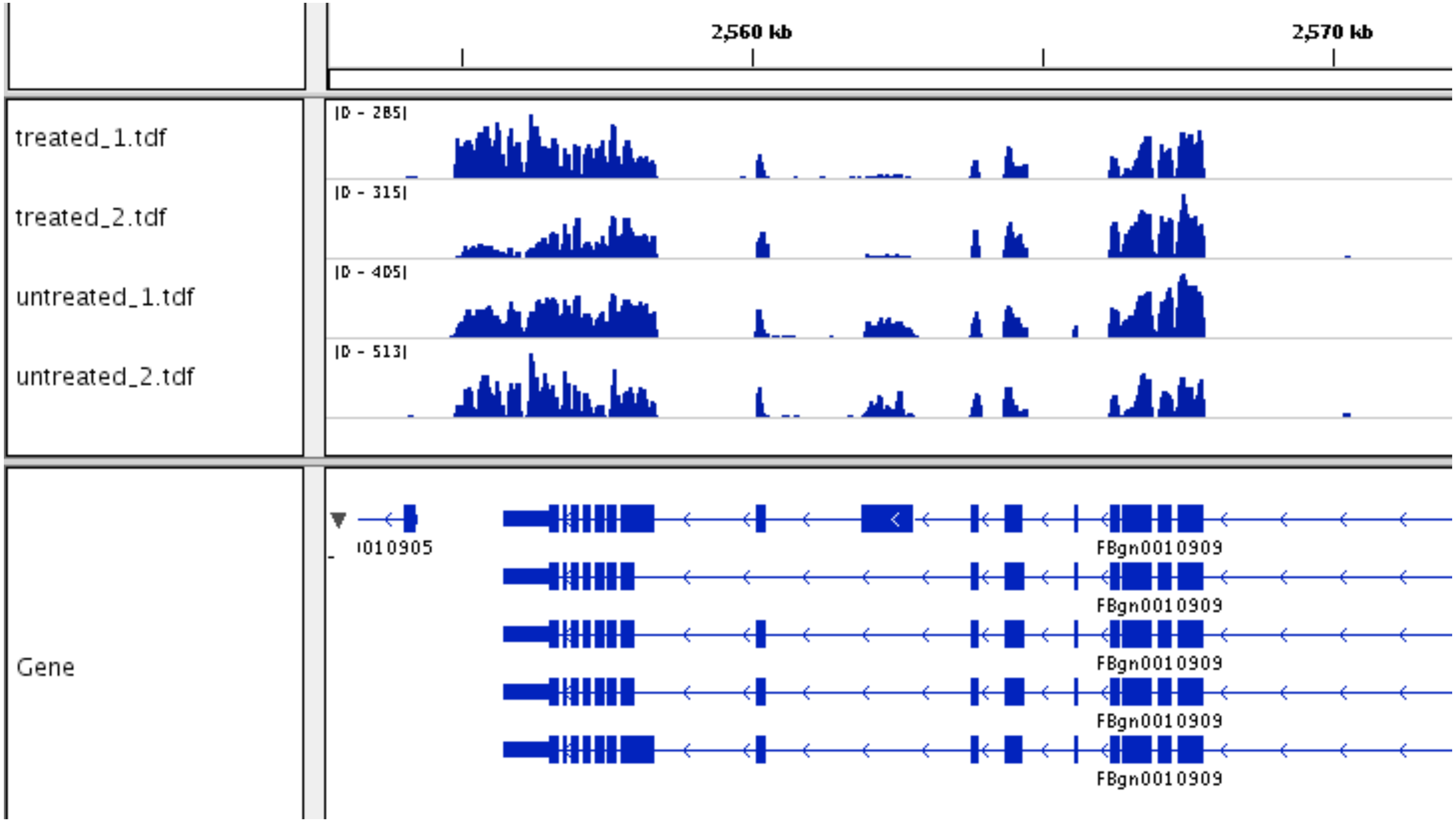
- **control (no treatment):**

  4 biological replicates (2x single end, 2x paired end)

- **treatment: knock-down of pasilla (a splicing factor)**

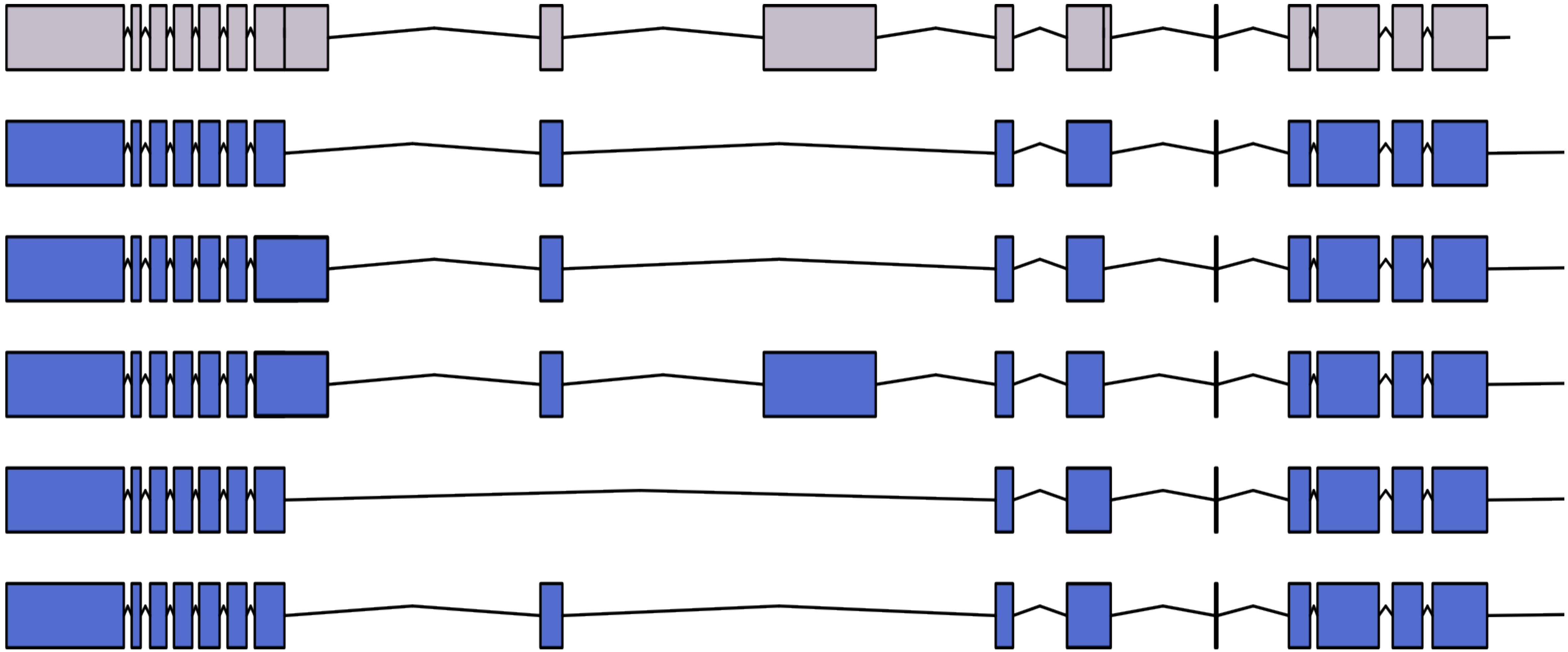
  3 biological replicates (1x single end, 2x paired end)

Alternative isoform regulation

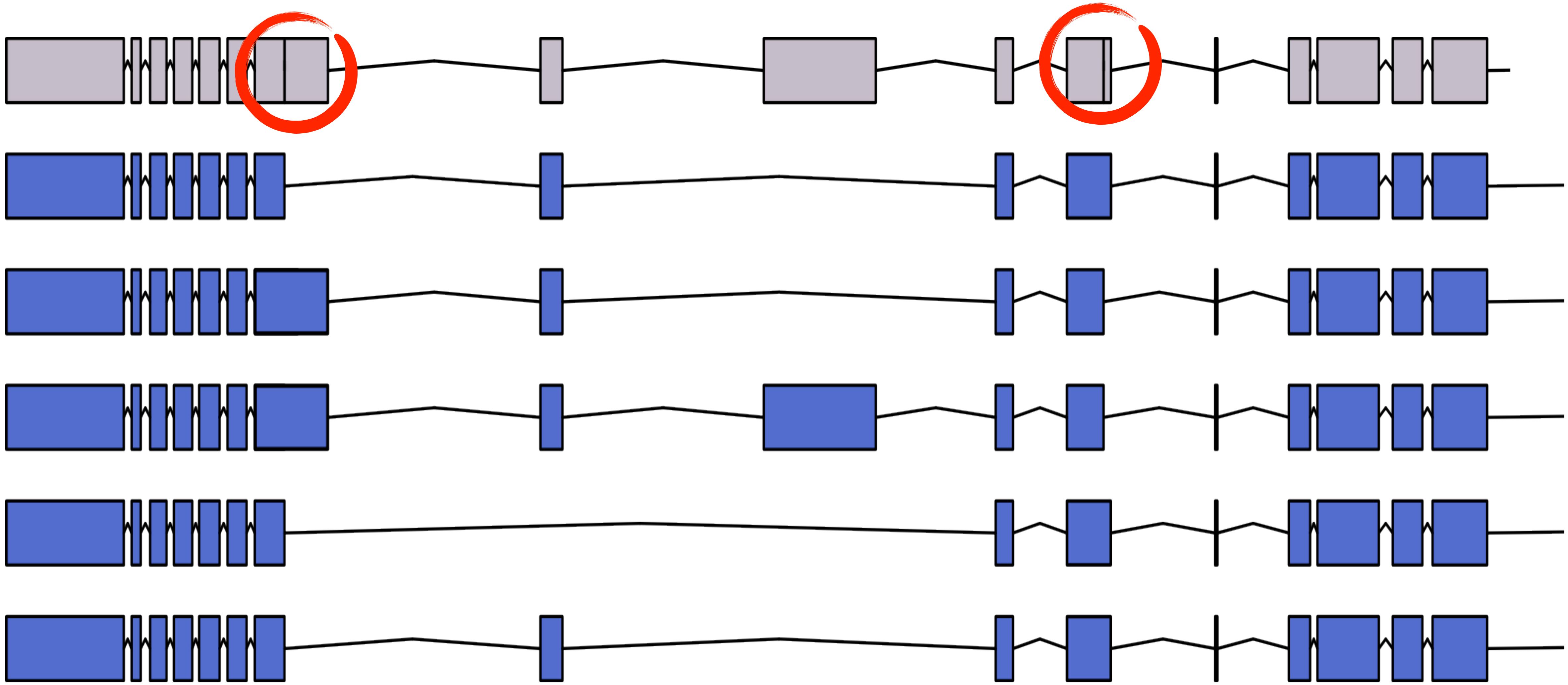


Data: Brooks et al., Genome Res., 2010

Exon counting bins



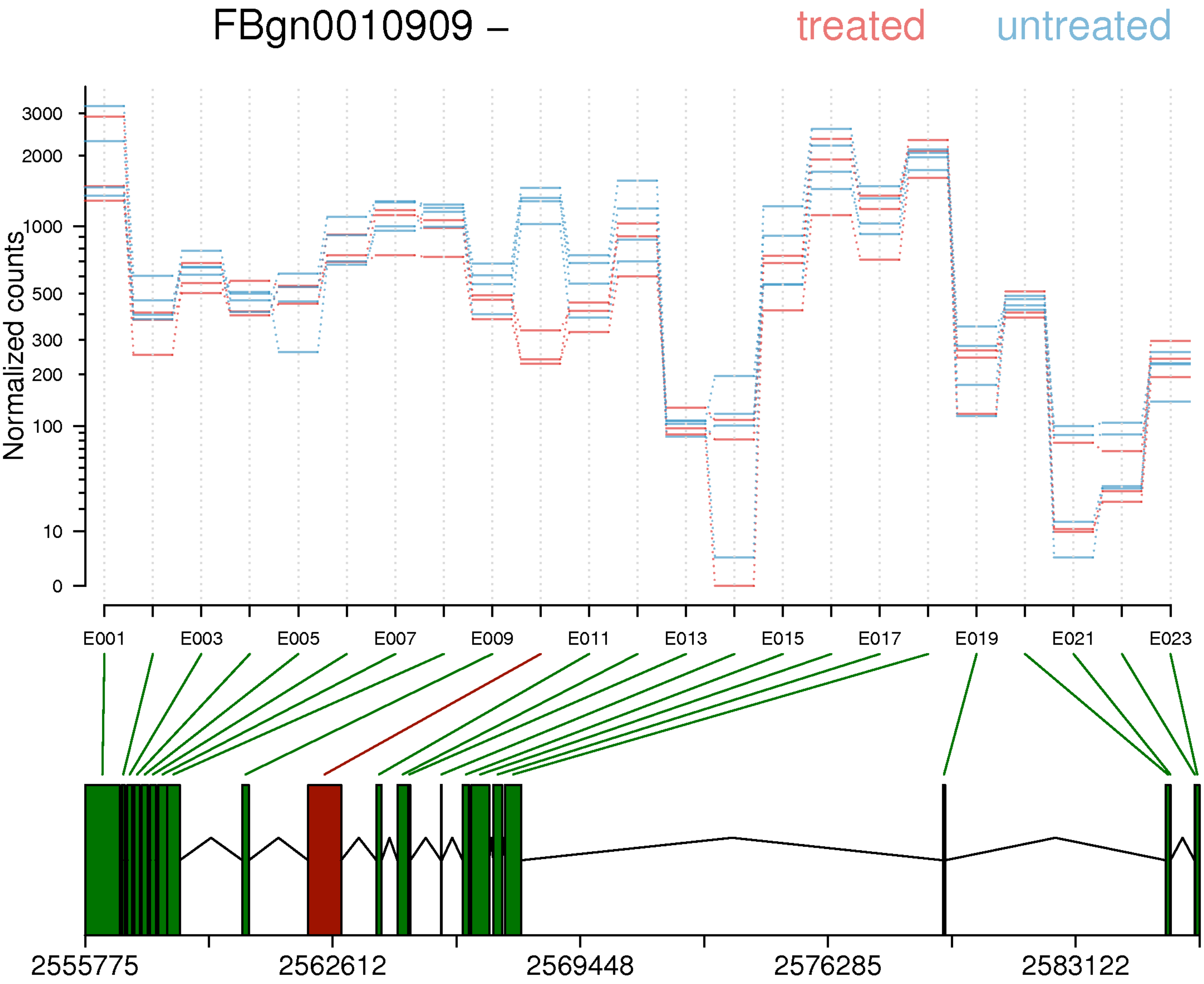
Exon counting bins



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
[...]
```


FBgn0010909 –
treated
untreated
Normalized counts
3000
2000
1000
500
300
200
100
10
0
E001
E003
E005
E007
E009
E011
E013
E015
E017
E019
E021
E023
2555775
2562612
2569448
2576285
2583122

Model

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

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

size factor

dispersion

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

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

Model, refined

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

$$\log \mu_{ijl} = \sum_j \beta^S_{ij} + \sum_l \beta^{\mathrm{E}}_{il} x^{\mathrm{E}}_l + \sum_{jl} \beta^{\mathrm{ET}}_{ijl} x^{\mathrm{E}}_l x^{\mathrm{T}}_j$$

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

Model, refined

$$K_{ijl} = NB\left(s_j \mu_{ijl}, \alpha\right.$$

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

$$\log \mu_{ijl} = \sum_j \beta^{S}_{ij} + \sum_l \beta^{\mathrm{E}}_{il} x^{\mathrm{E}}_l + \sum_{jl} \beta^{\mathrm{ET}}_{ijl} x^{\mathrm{E}}_l x^{\mathrm{T}}_j$$

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

Dispersion estimation

- Standard maximum-likelihood estimate for dispersion parameter has (unacceptably) strong bias in the case of small sample size.

- A method-of-moments estimator (as used in DESeq) cannot be used due to crossed factors.

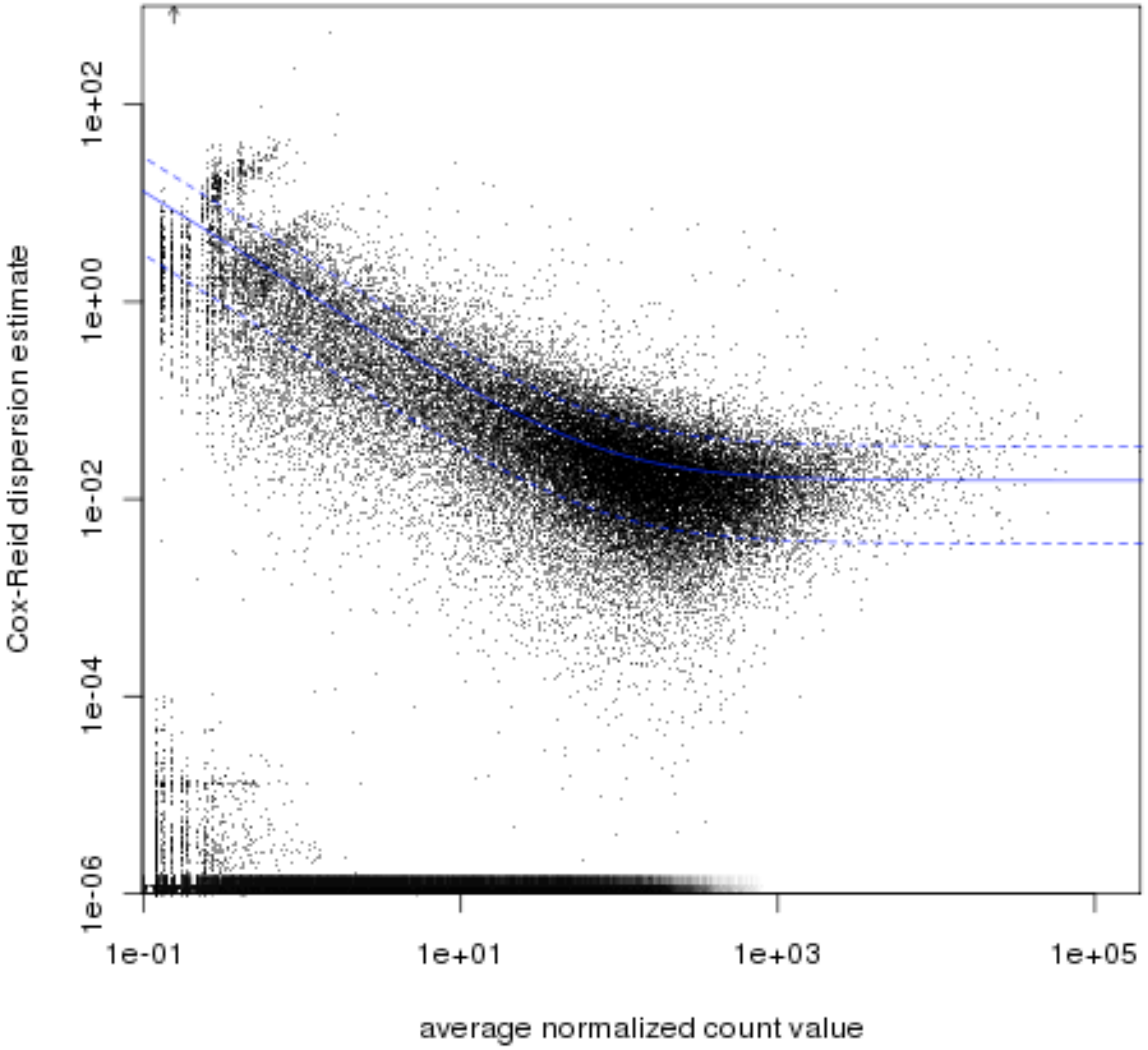
- We adapt the solution from the recent edgeR: Cox-Reid conditional-maximum-likelihood estimation (edgeR: Robinson, McCarthy, Smyth (2010))

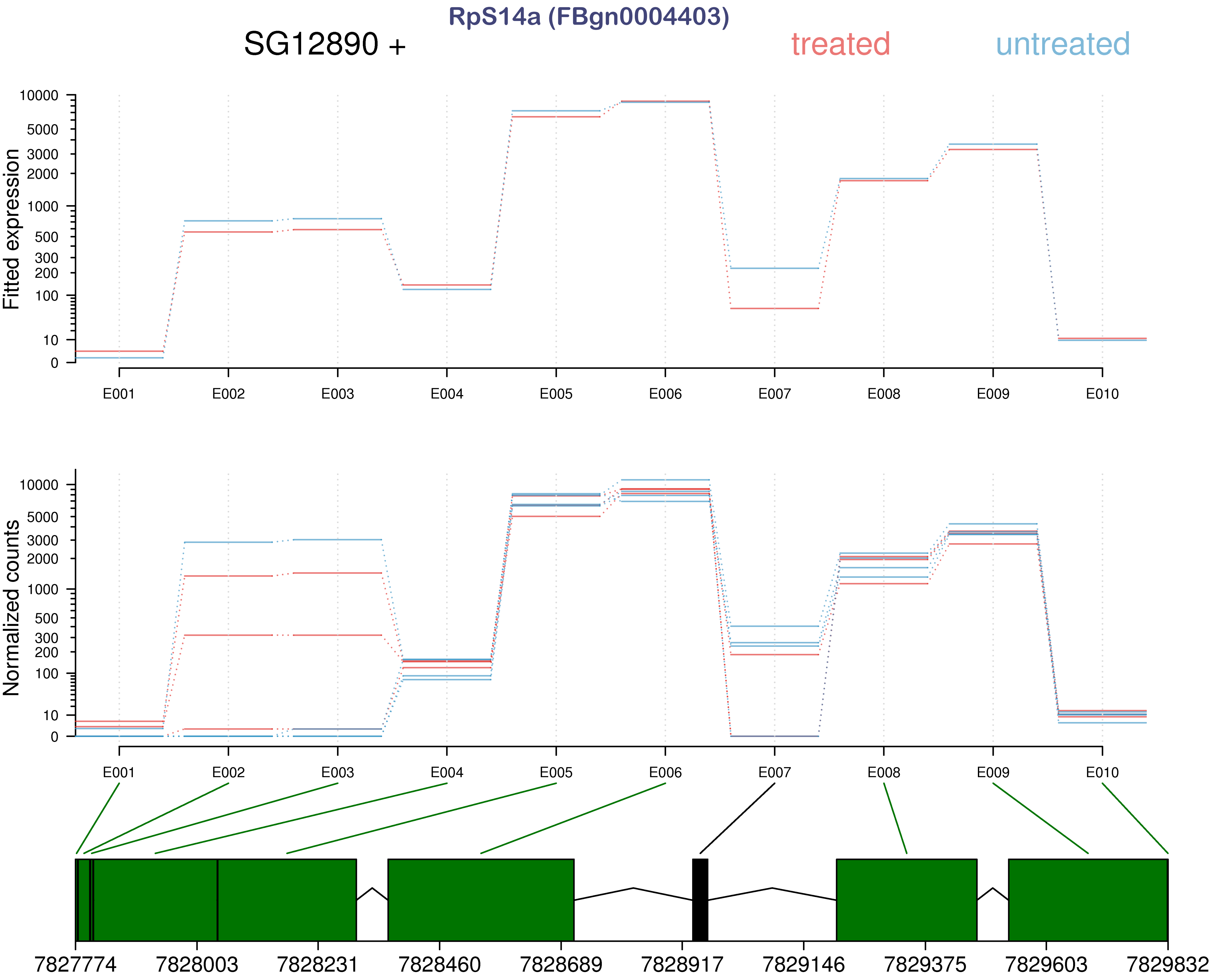
Dispersion estimation

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?

Dispersion vs mean




RpS14a (FBgn0004403)
SG12890 +
treated
untreated
Fitted expression
10000
5000
3000
2000
1000
500
300
200
100
10
0
E001
E002
E003
E004
E005
E006
E007
E008
E009
E010
Normalized counts
10000
5000
3000
2000
1000
500
300
200
100
10
0
E001
E002
E003
E004
E005
E006
E007
E008
E009
E010
7827774
7828003
7828231
7828460
7828689
7828917
7829146
7829375
7829603
7829832

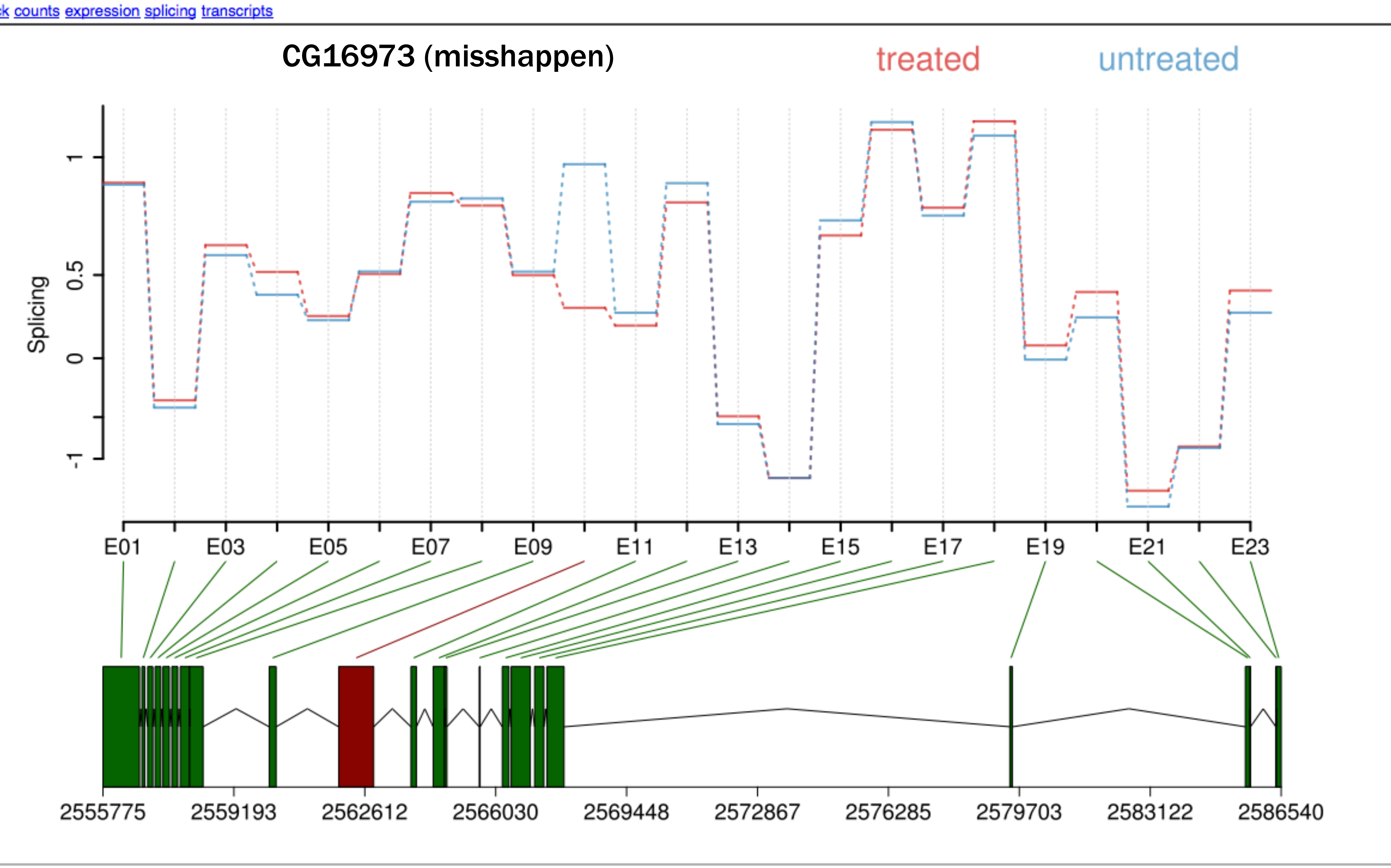
Conclusion II

- Counting within exons and NB-GLMs allows studying 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 a helpful component for the study of alternative isoform regulation.

Alternative exon expression detected by ANOVA - GLM

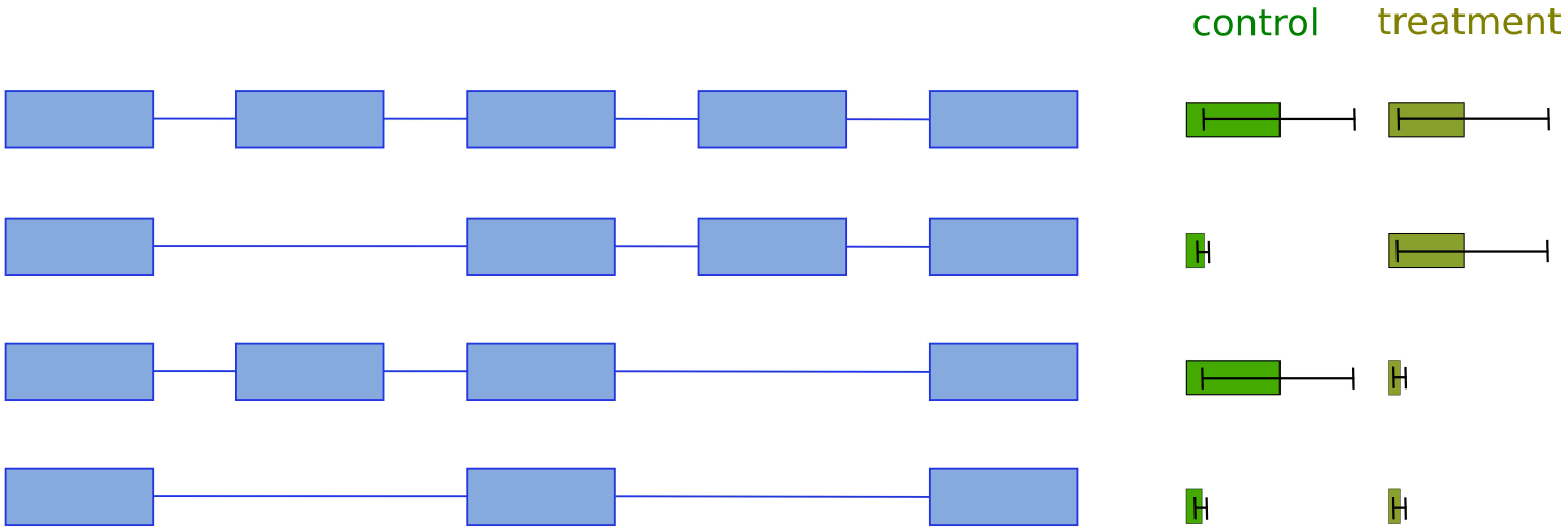


Simon Anders
Alejandro Reyes

Joseph Barry
Bernd Fischer
Ishaan Gupta
Felix Klein
Gregoire Pau
Aleksandra Pekowska
Paul-Theodor Pyl

Lars Steinmetz
Eileen Furlong
Paul Bertone
Robert Gentleman
Jan Korbel

Why testing for differential exon usage rather than for isoform abundance changes?



control	90% in	50% in
treatment	10% in	50% in