Differential analysis of count data

Wolfgang Huber
EMBL

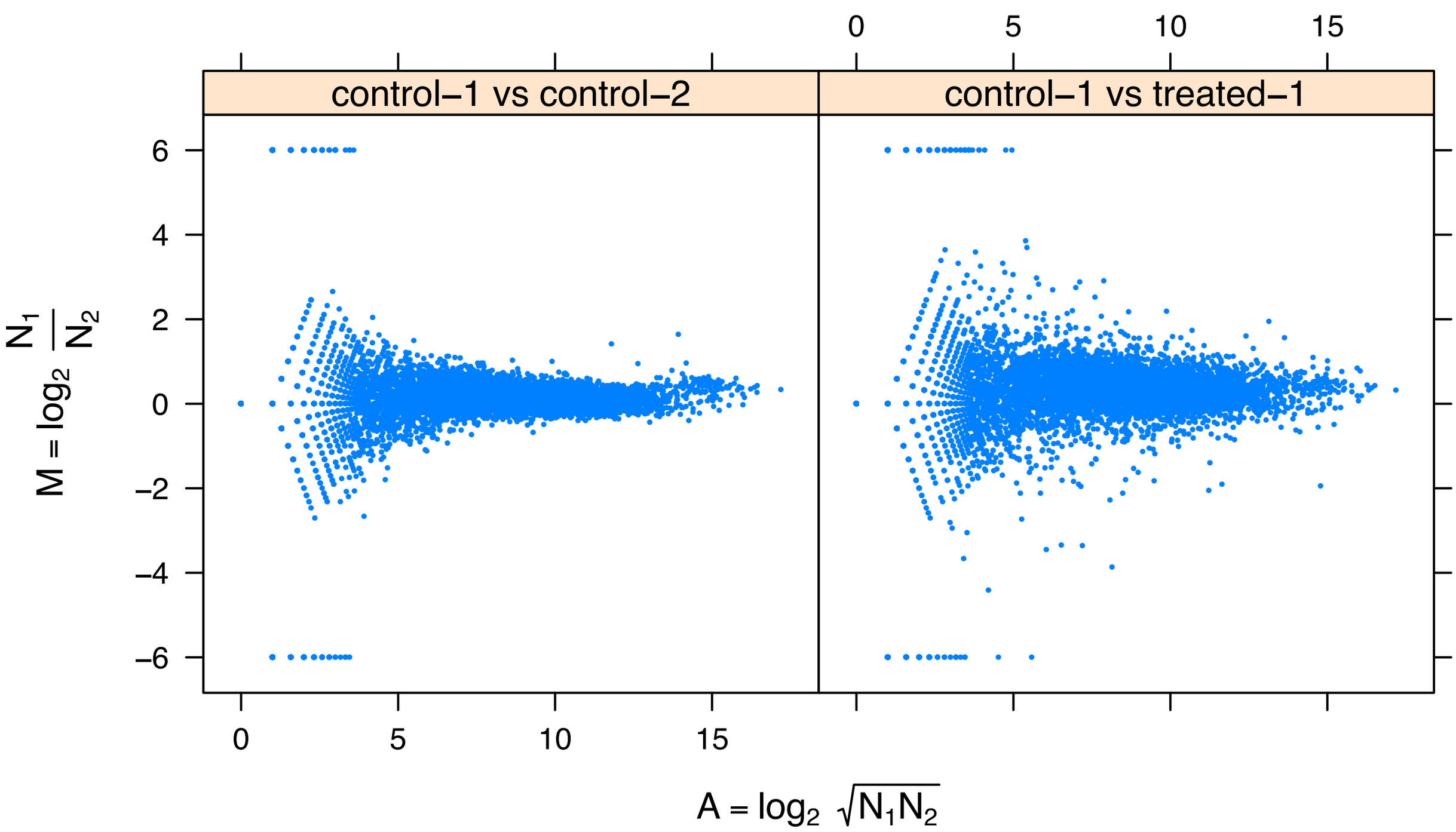
23 September 2013 - Tübingen

Count data

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

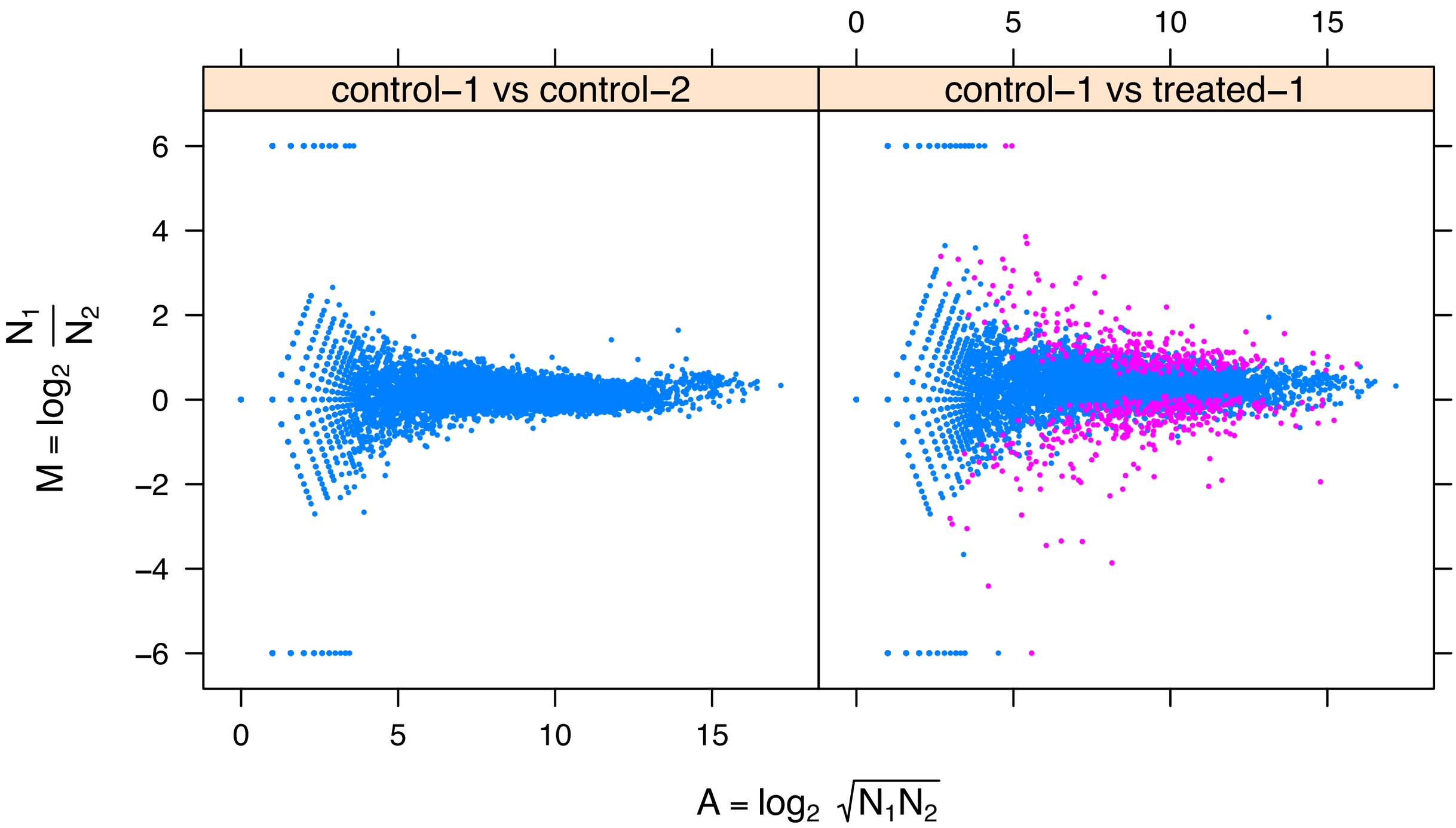
[...]

- RNA-Seq
- ChIP-Seq
- HiC
- Barcode-Seq
- Peptides in mass spec
- ...



two biological replicates

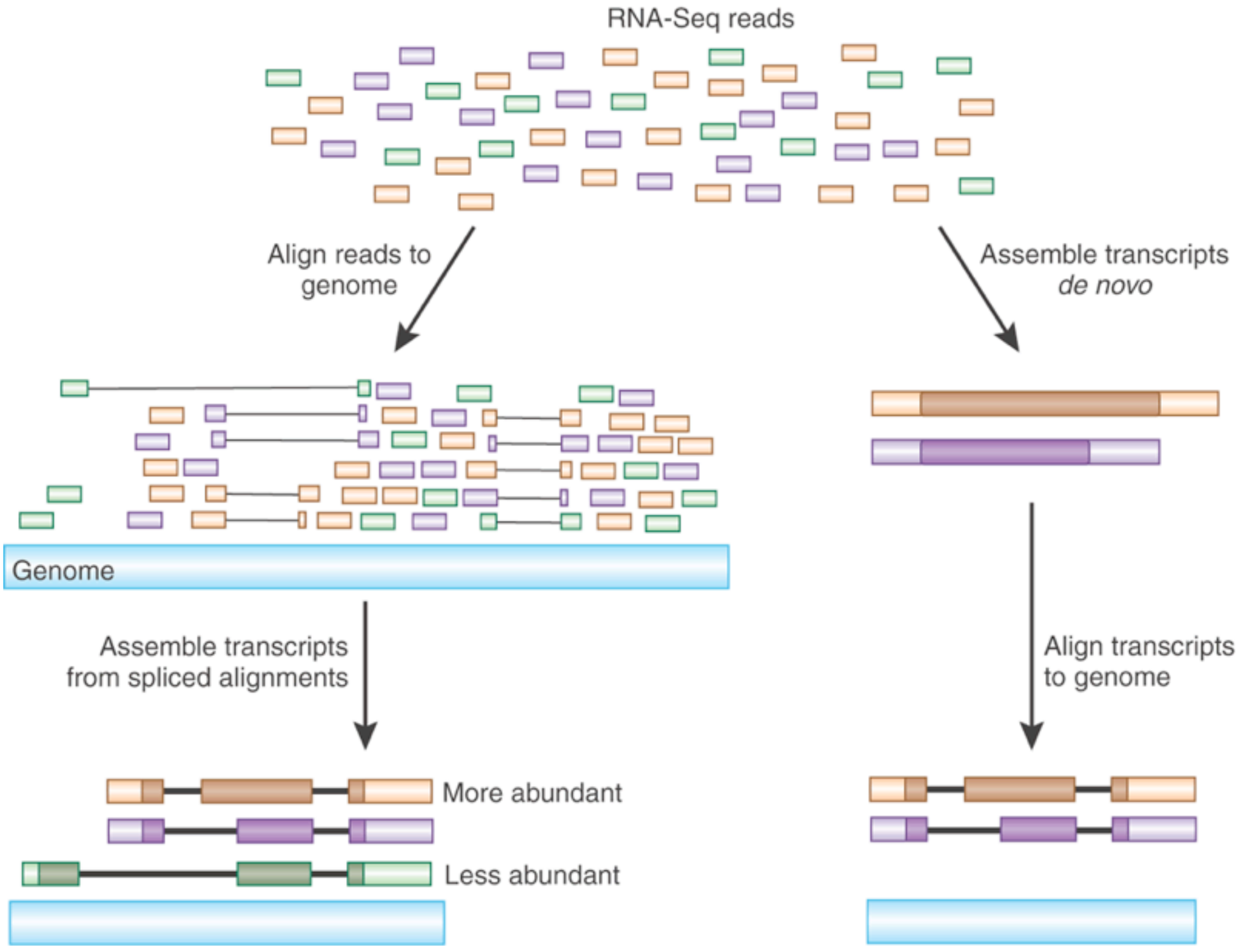
treatment vs control



two biological replicates

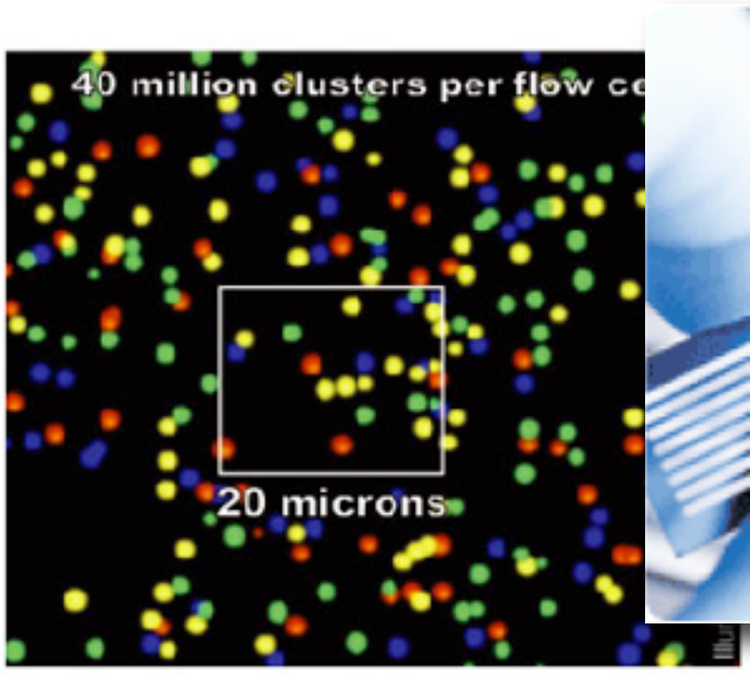
treatment vs control

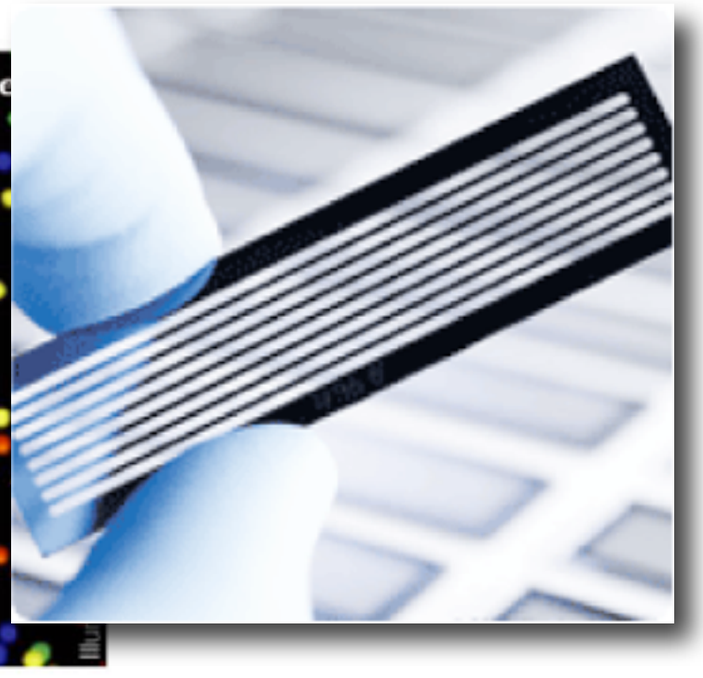
“2nd generation” sequencing



Haas & Zody, Nat. Biotech 2010

“2nd generation” sequencing



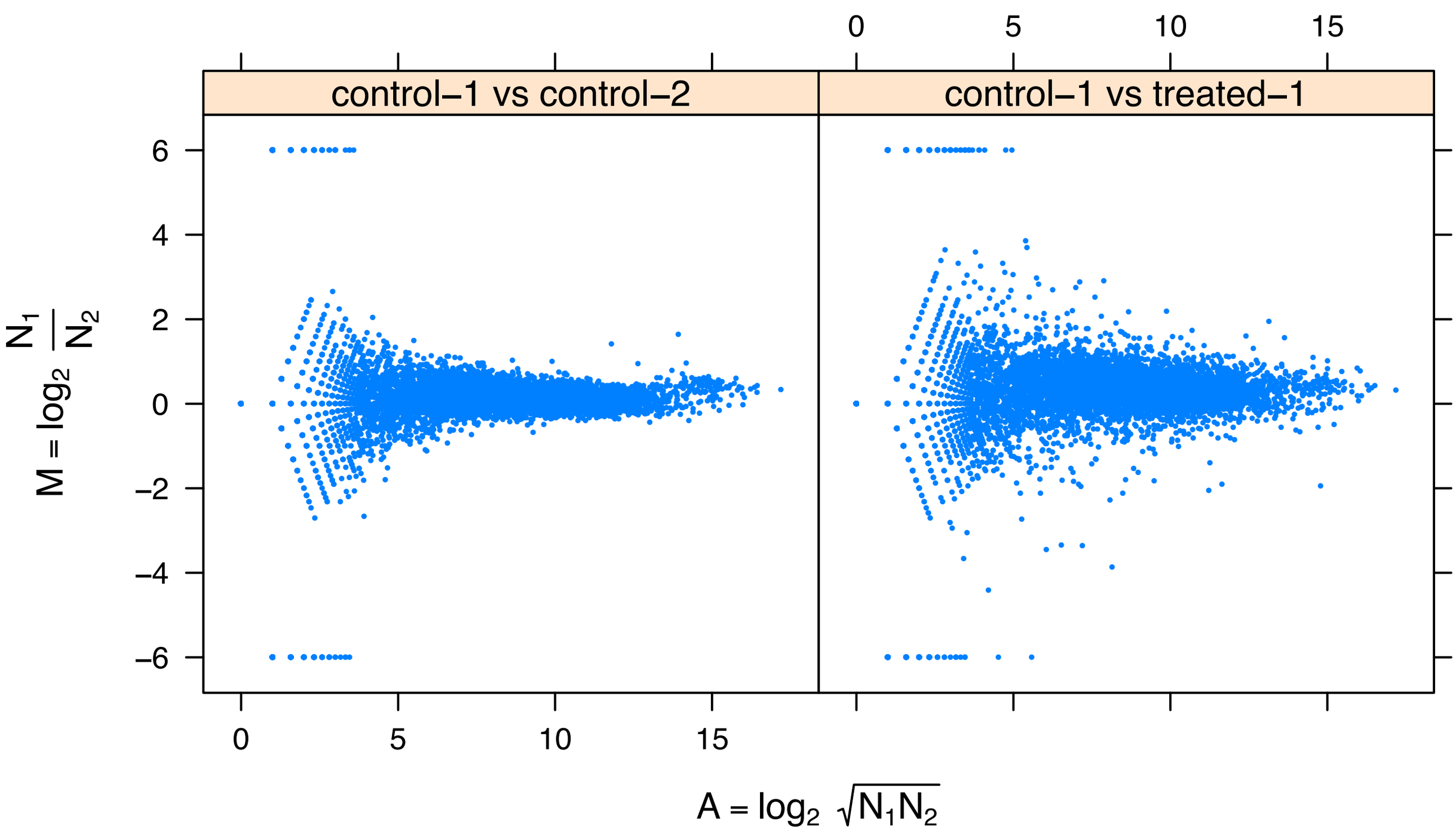


Solexa HiSeq 2500
1 run (11 days):
ca. 3x10^9 reads @ 2x100nt
(≙ 6 human genomes 30x)
ca. 5 k€ (marginal cost)

Applications

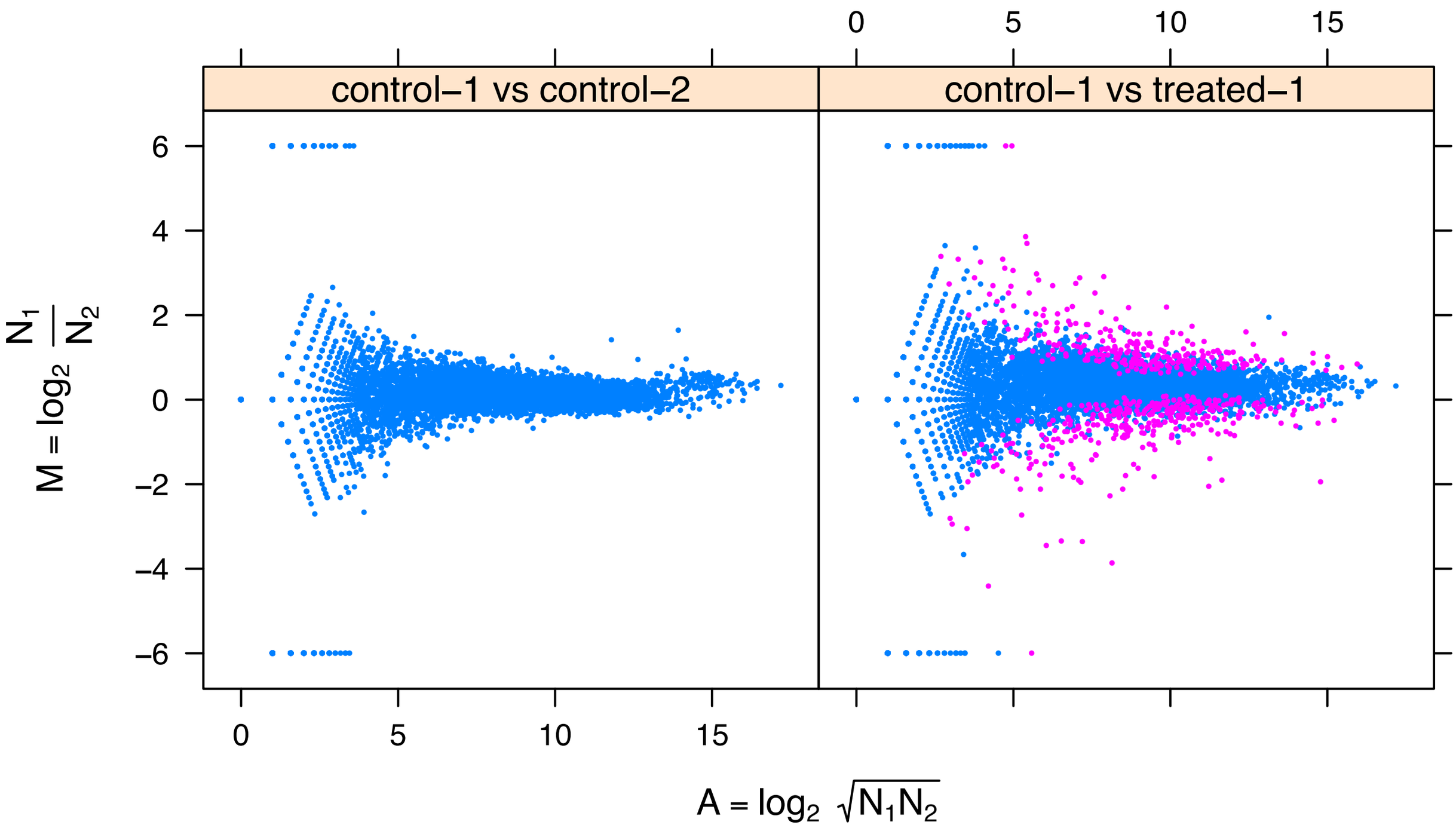
Genomes of new species
Individual genomes
Metagenomes
Cancer genomes
Transcriptome sequencing (RNA-Seq)
Protein-DNA binding (ChIP-Seq)
Protein-RNA binding (CLIP-Seq, RIP-Seq)
3D-structure of the nuclear DNA (Hi-C & Co.)

On the horizon: much longer reads
ability to assign reads to individual chromosomes, cells



two biological replicates

treatment vs control



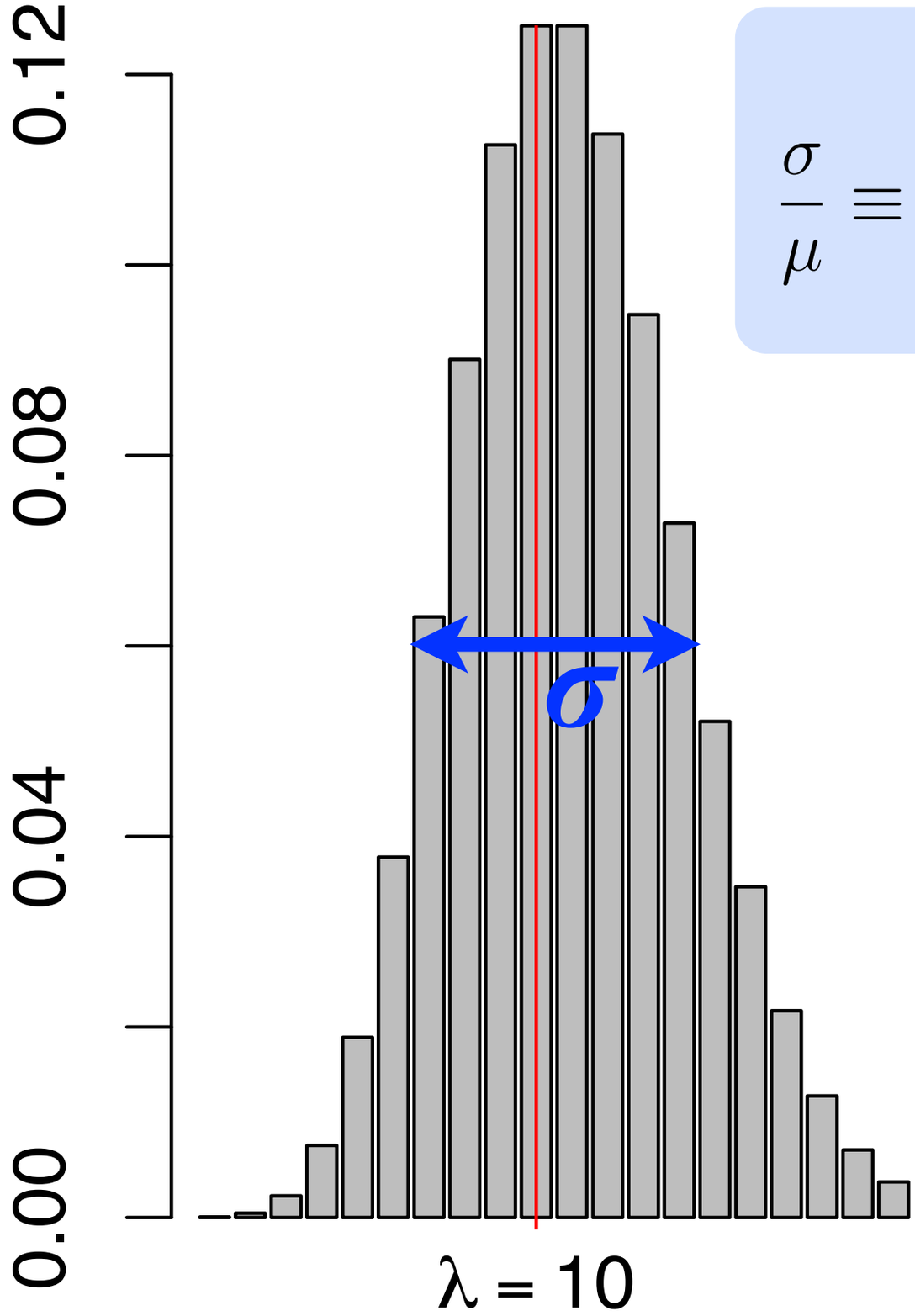
two biological replicates

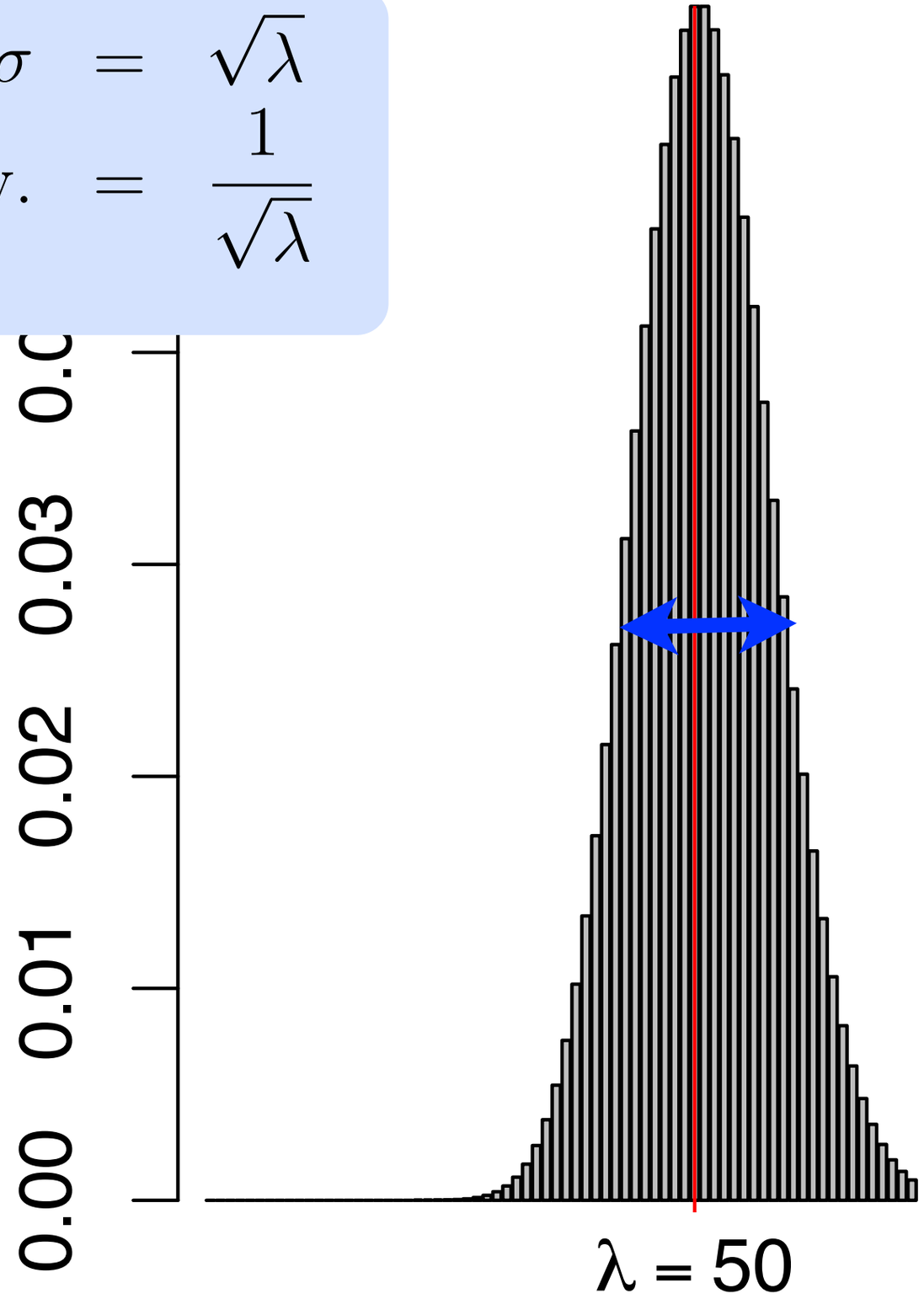
treatment vs control

The Poisson distribution is used for counting processes

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

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

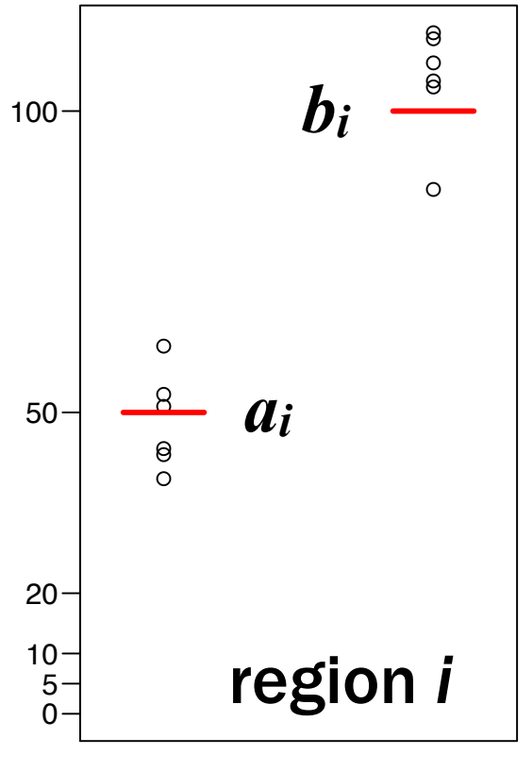




Analysis method: ANOVA

$$N_{ij} \sim \text{Poisson}(\mu_{ij}) \qquad \textbf{Noise part}$$

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



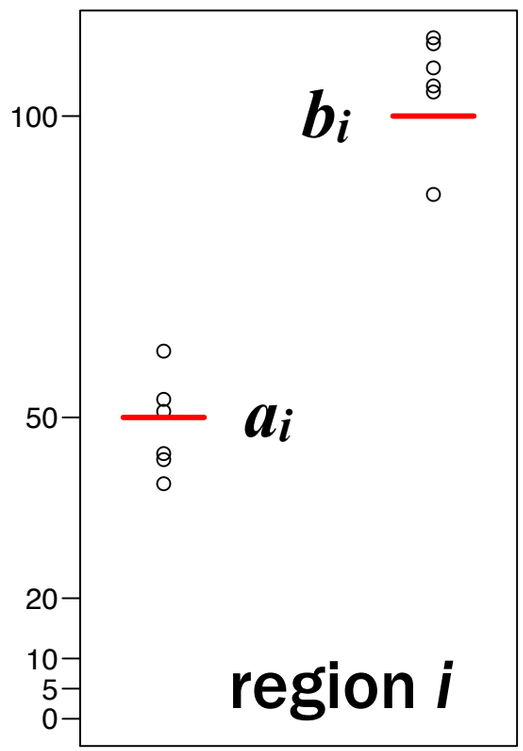
μ_{ij} expected count of region *i* in sample *j*
s_j library size factor
x_{kj} design matrix
β_{ik} (differential) effect for region *i*

Analysis method: ANOVA

$$N_{ij} \sim \text{Poisson}(\mu_{ij})$$ **Noise part**

$$\log \mu_{ij} = s_j + \sum_k \beta_{ik} x_{kj}$$ **Systematic part**

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



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

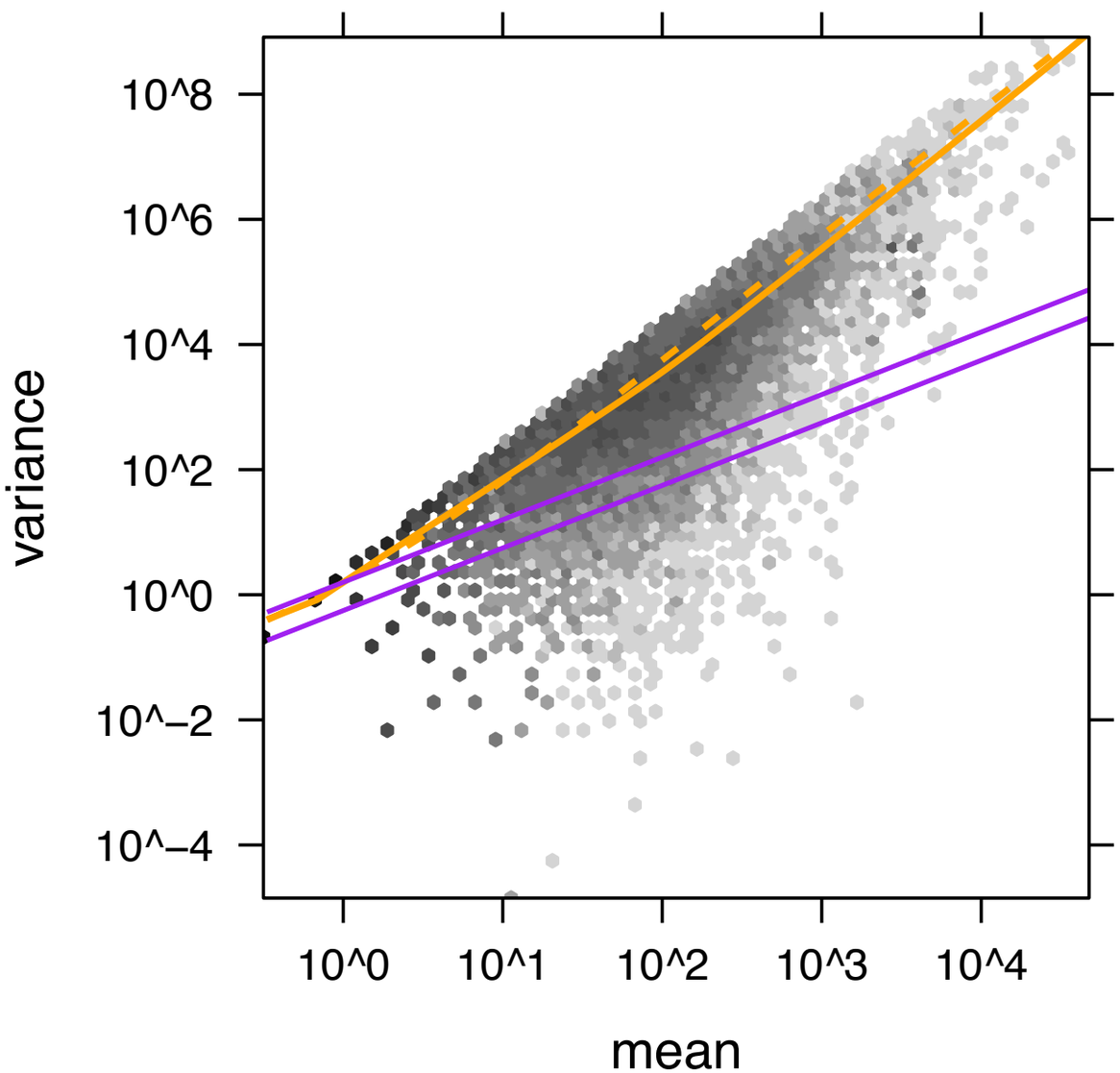
No need to estimate the variance. This is convenient.

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



NB: $v \sim \mu^2$

Poisson: $v \sim \mu^1$

Data: Nagalakshmi et al. Science 2008

So we need a better way

data are discrete, positive, skewed

➡ no (log-)normal model

small numbers of replicates

➡ no rank based or permutation methods

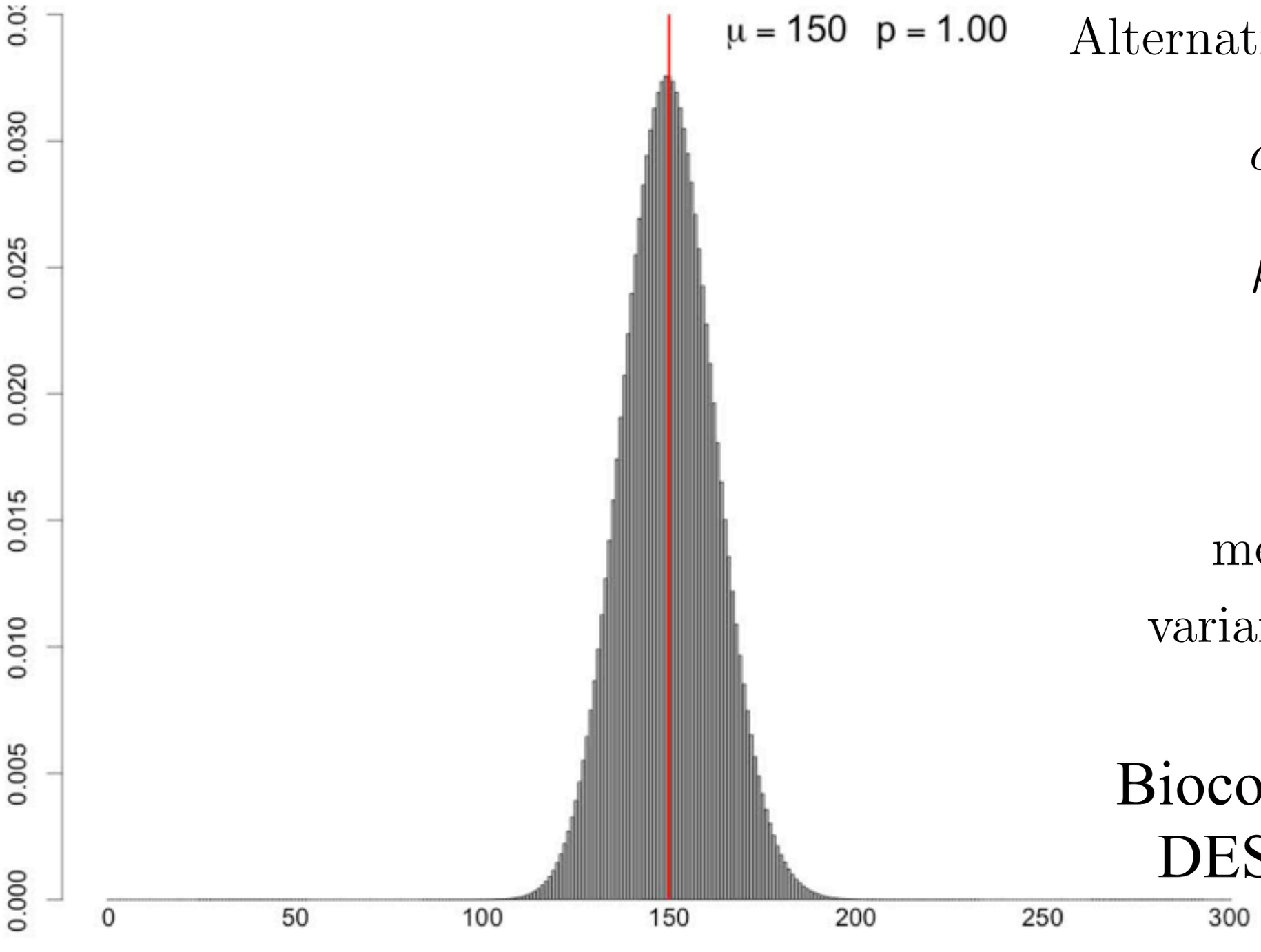
➡ want to use parametric stochastic model to infer tail behaviour (approximately) from low-order moments (mean, variance)

large dynamic range (0 … 10^5)

➡ heteroskedasticity matters

The negative-binomial distribution

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



Alternative parameterisation

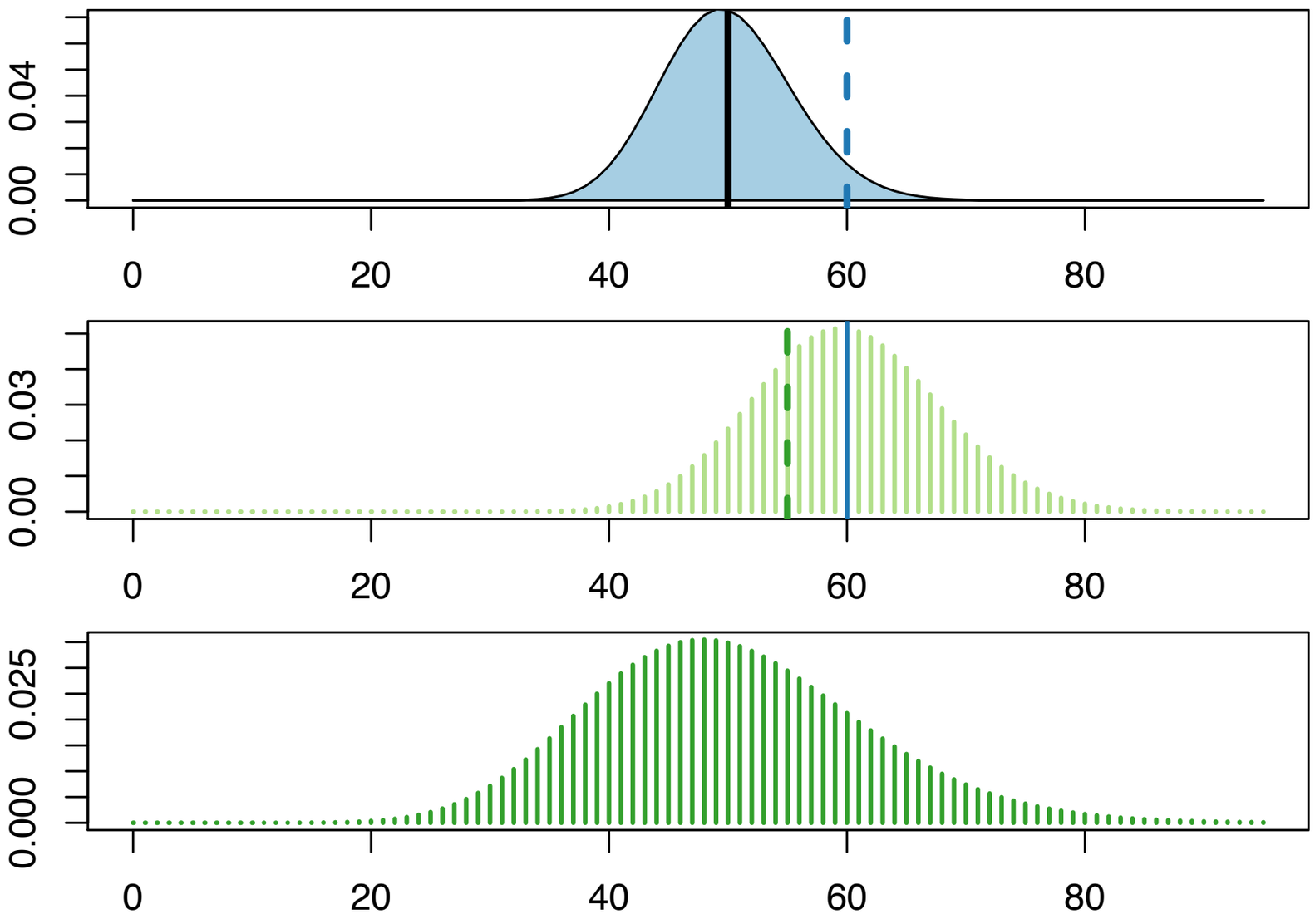
$$\begin{aligned} \alpha &= \frac{1}{r} \\ \mu &= \frac{pr}{1-p} \end{aligned}$$

Moments

$$\begin{aligned} \text{mean} &= \mu \\ \text{variance} &= \mu + \alpha\mu^2 \end{aligned}$$

Bioconductor package
DESeq, since 2010

The NB distribution models a Poisson process whose rate is itself randomly varying



Biological sample to sample variability Γ

Poisson counting statistics Λ

Overall distribution NB

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

Two component noise model

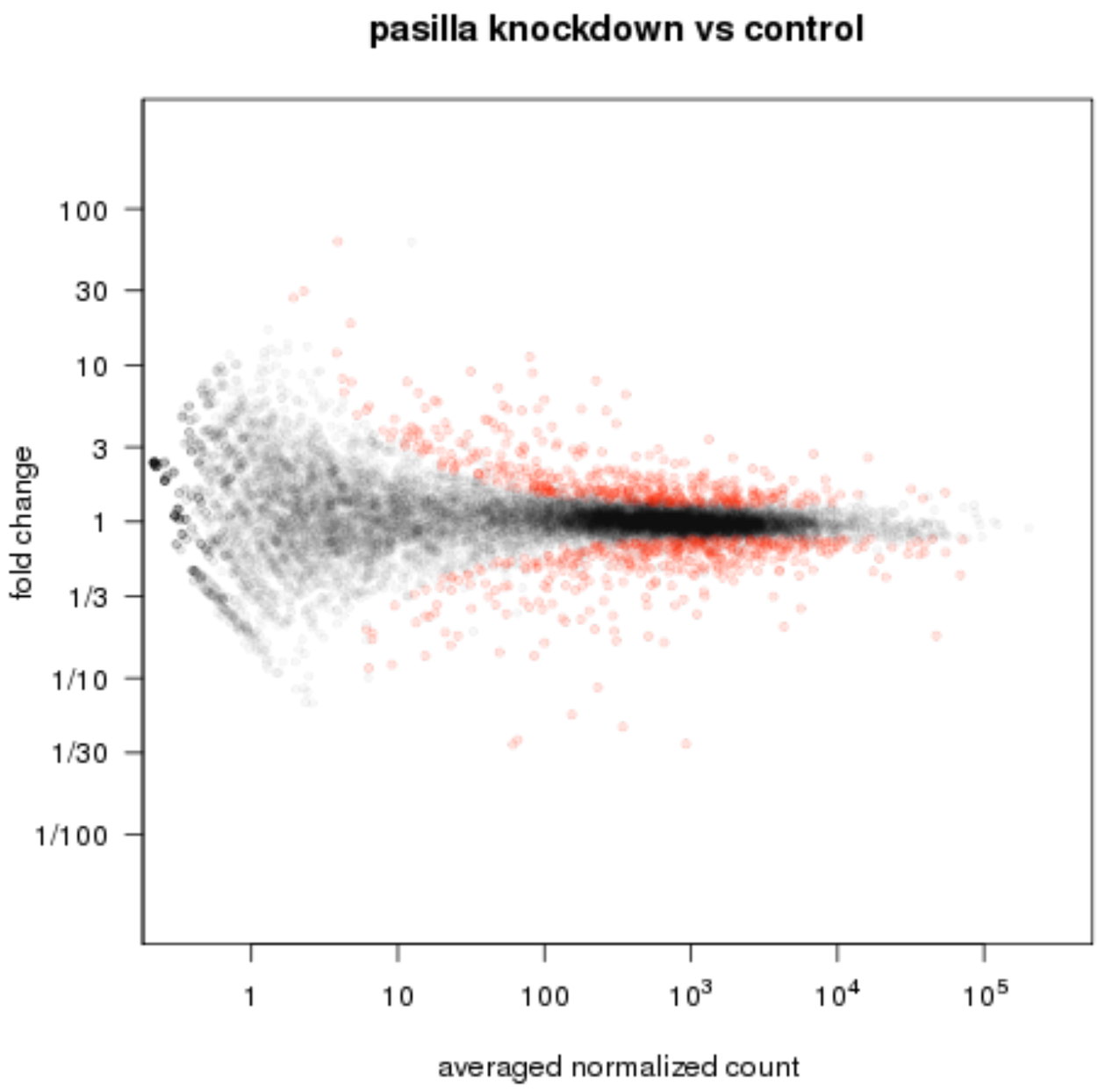
$$\text{var} = \mu + c\,\mu^2$$

shot noise (Poisson) — biological noise

Small counts

Sampling noise dominant

Improve power: deeper coverage



Large counts

Biological noise dominant

Improve power: more biol. replicates

Generalised linear model of the negative binomial family

$$N_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_{ij})$$ **Noise part**

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

What is a generalized linear model?

$$Y \sim D(m, s)$$

A GLM consists of three elements:

1. A probability distribution *D* (from the exponential family), with mean E[*Y*] = *m* and dispersion s

2. A linear predictor η = *X* β

3. A link function *g* such that *g*(*m*) = η.

Ordinary linear model: *g* = identity, *D* = Normal
DESeq(2), edgeR, ...: *g* = log, *D* = Negative Binomial

design with a 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

GLM with blocking factor

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

***i*: genes**
***j*: samples**

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

Generalized linear models

Simple design:

Two groups, e.g. *control* and *treatment*

Common complex designs:

- **Designs with blocking factors**
- **Factorial designs**
- **Designs with interactions**
- **Paired designs**

GLMs: Dual-assay designs (e.g.: CLIP-Seq + RNA-Seq)

How does affinity of an RNA-binding protein to mRNA change under a (drug, RNAi) treatment?

For each sample, we are interested in the ratio of CLIP-Seq to RNA-Seq reads. How is it affected by treatment?

full model:

count ~ assayType + treatment + assayType : treatment

reduced model:

count ~ assayType + treatment

Zarnack et al., Cell 2013

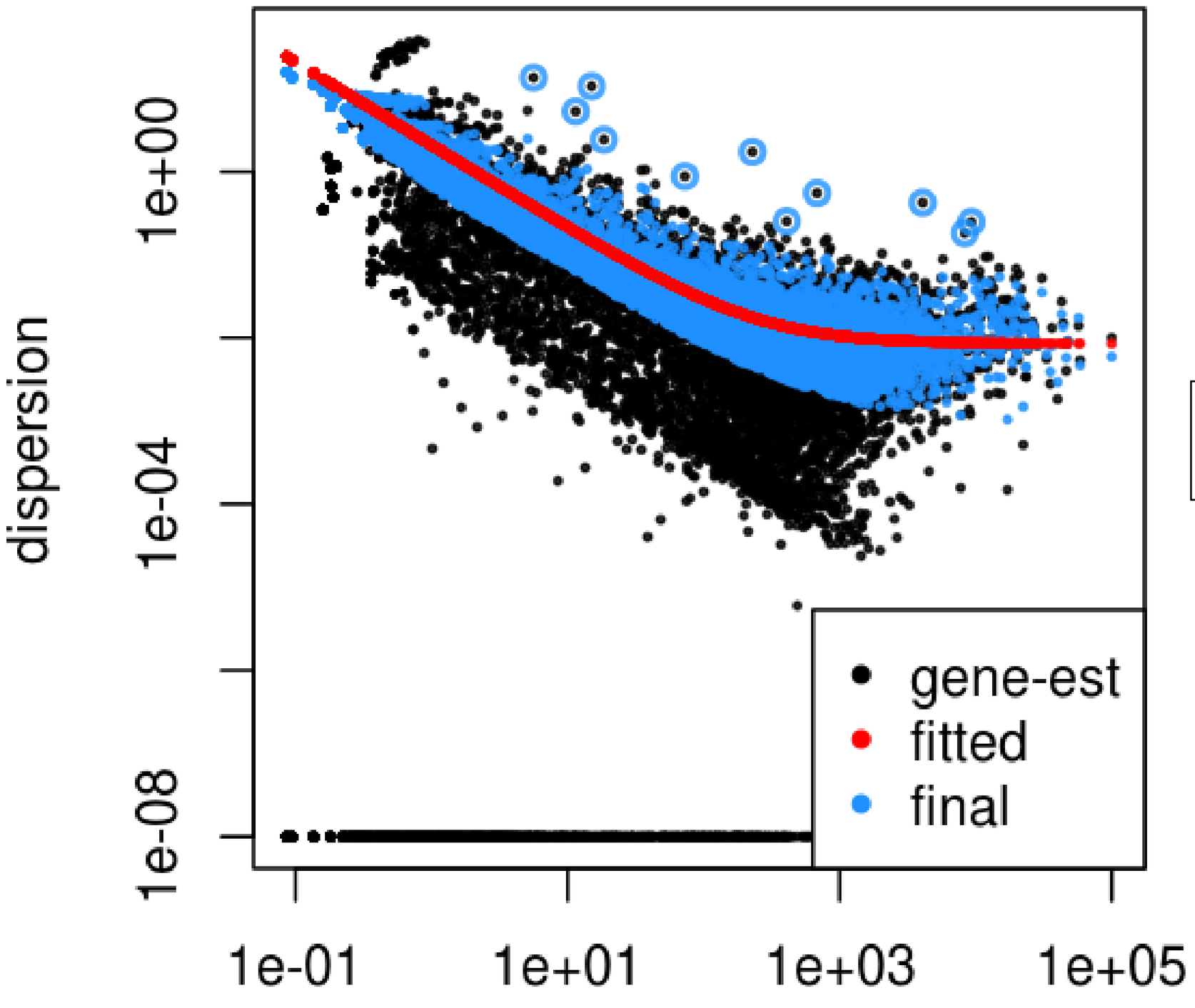
Modelling Variance

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

⇒ridge (Tikhonov) regularisation, empirical Bayes

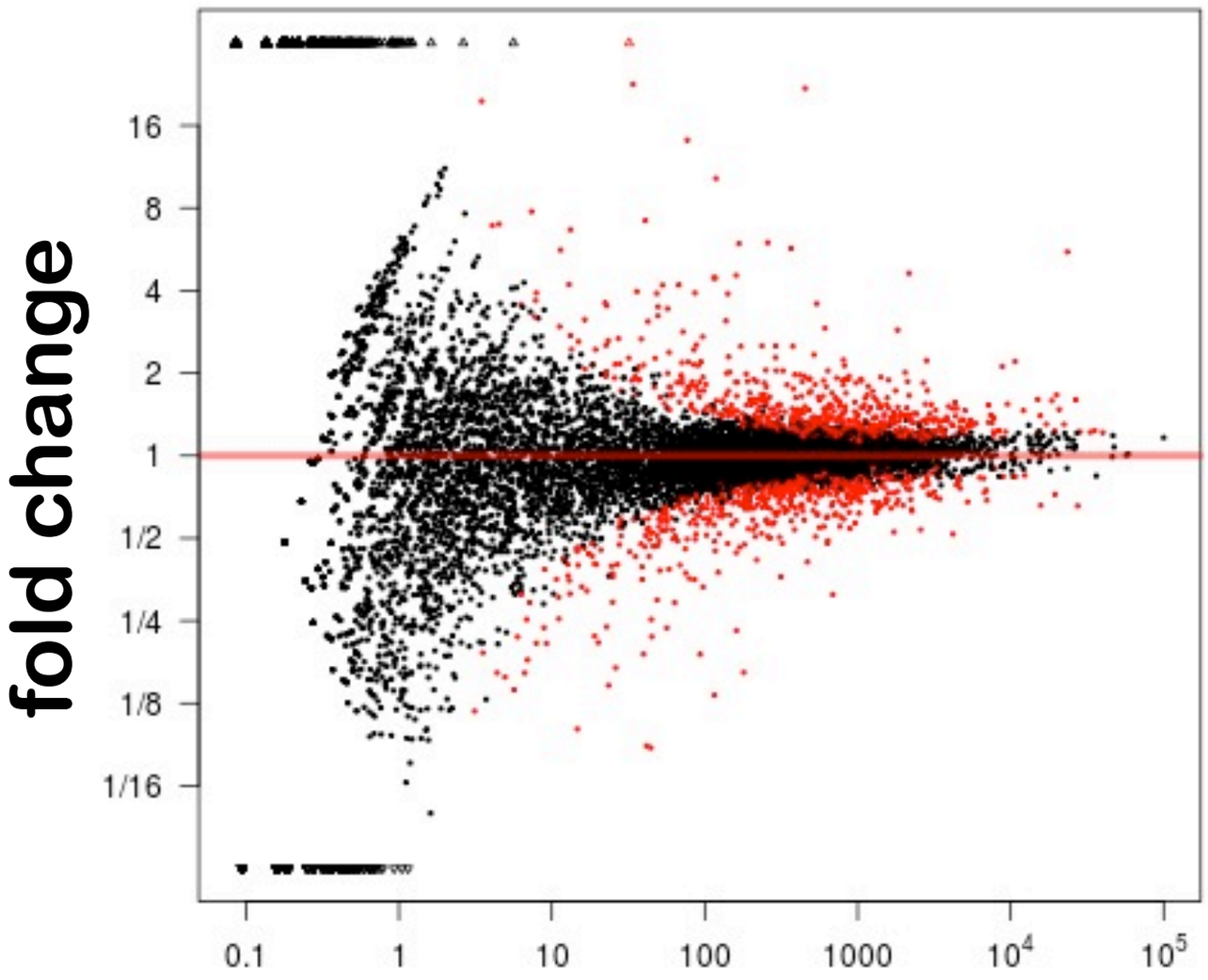
Dispersion estimation: shrinkage

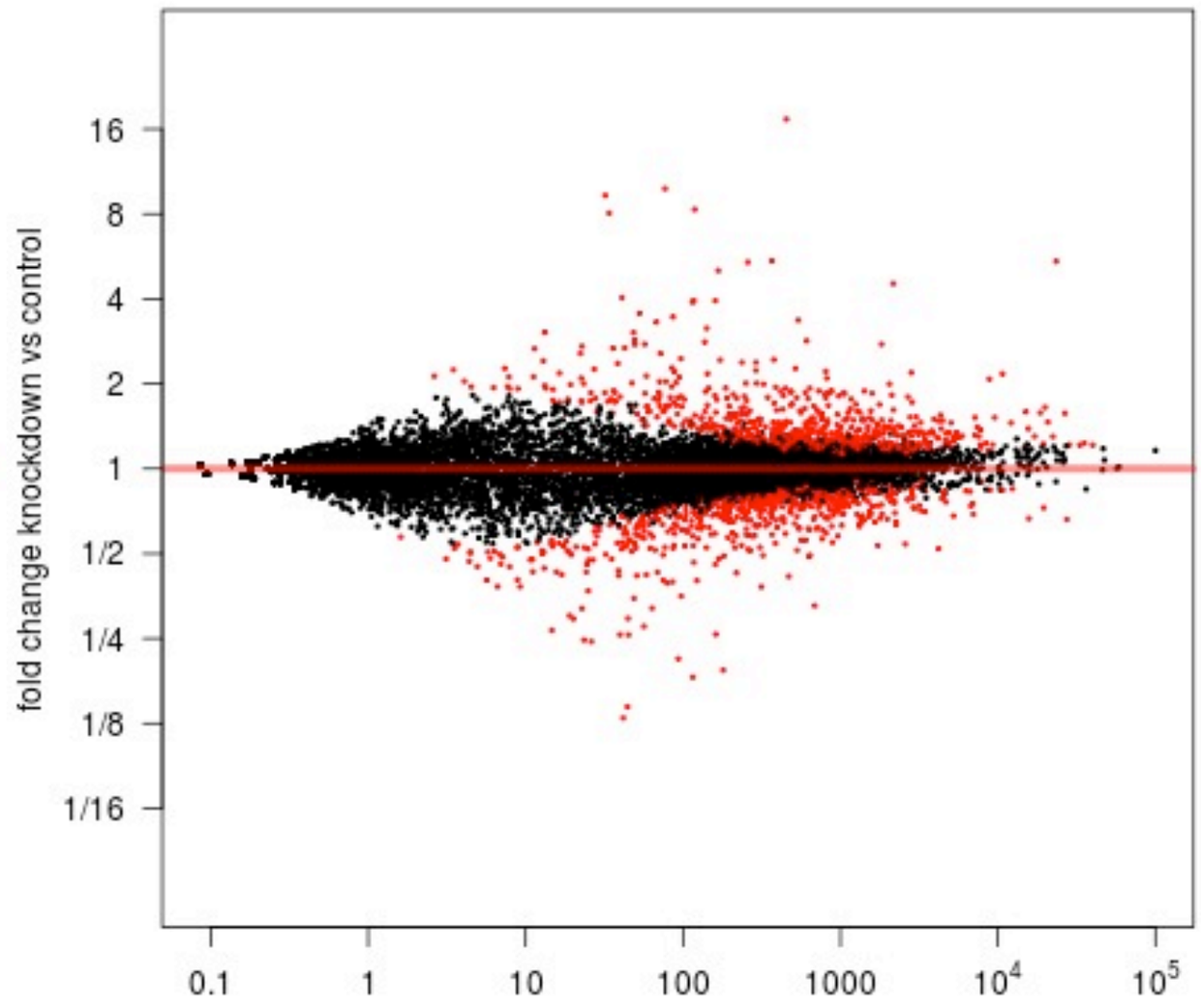


dispersion outliers:

$\log(\alpha_{\text{gene-est}}) - \log(\alpha_{\text{fit}}) > 2\ \sigma_{\text{rob}}$

Beta (estimated effects): shrinkage





mean of normalized counts

The mechanics: empirical Bayes shrinkage of gene-wise dispersion estimates and of (non-intercept) βs

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\text{argmax}}\, \ell(\alpha | y, \hat{\mu})$$

“naive” GLM likelihood

$$\text{CR}(\alpha) = -\frac{1}{2} \log(\det(X^t W X))$$

Cox-Reid bias term

$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\text{argmax}}\, (\ell(\alpha | y, \hat{\mu}) + \text{CR}(\alpha))$$

bias-corrected likelihood

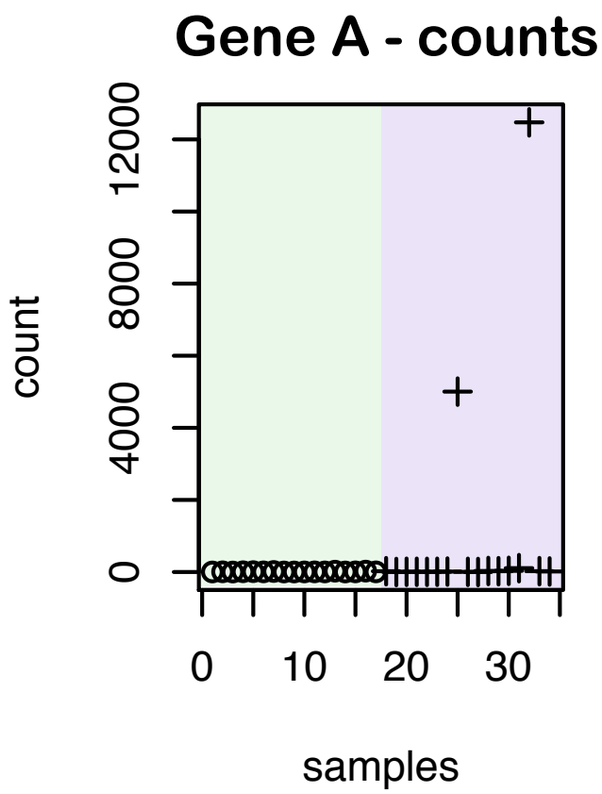
$$\text{prior}(\alpha) = \log(f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma^2_{\text{prior}})$$

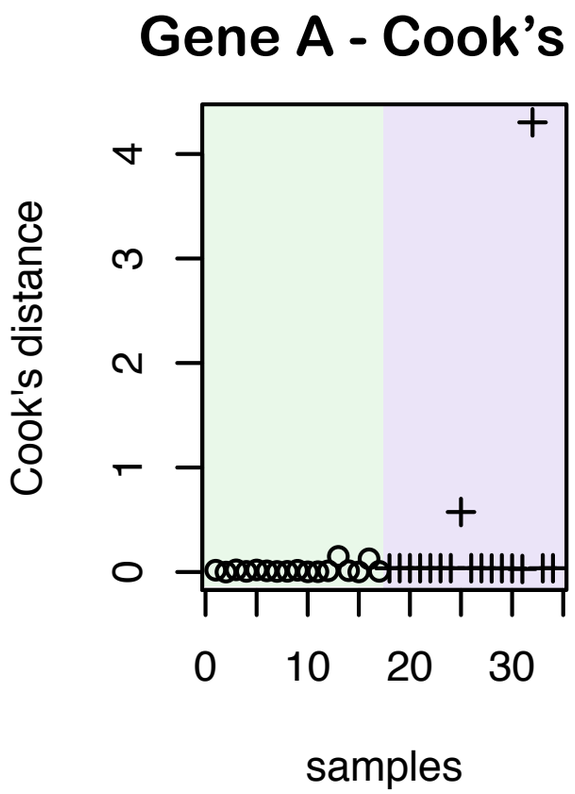
prior on α by ‘information sharing’ across genes

$$\hat{\alpha}_{\text{CR-MAP}} = \underset{\alpha}{\text{argmax}}\, (\ell(\alpha | y, \hat{\mu}) + \text{CR}(\alpha) + \text{prior}(\alpha))$$

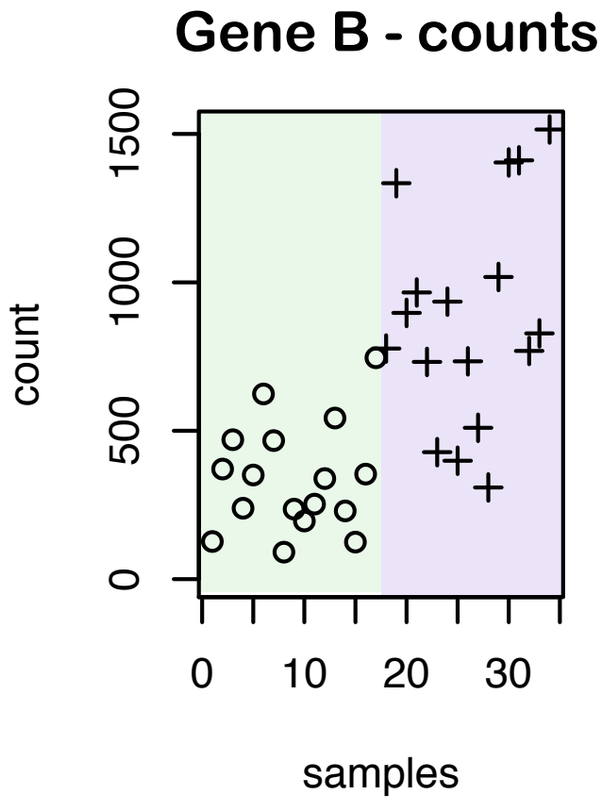
penalized likelihood

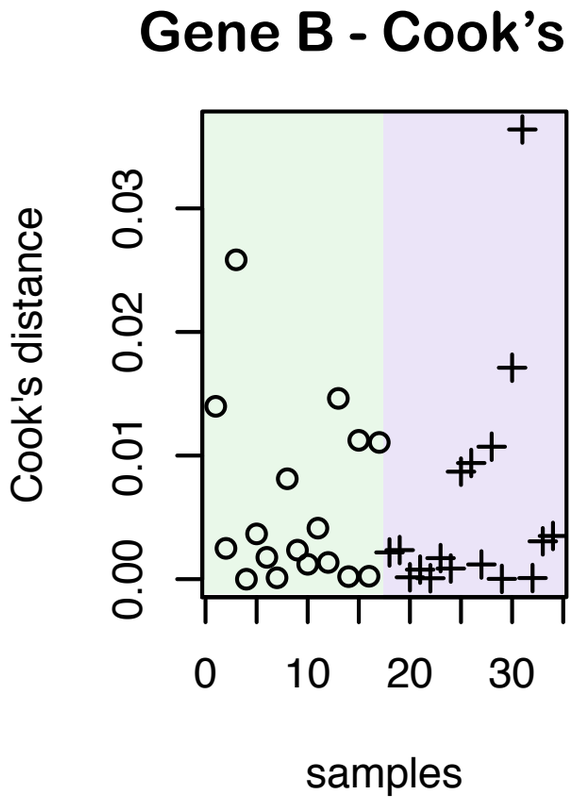
Outlier robustness



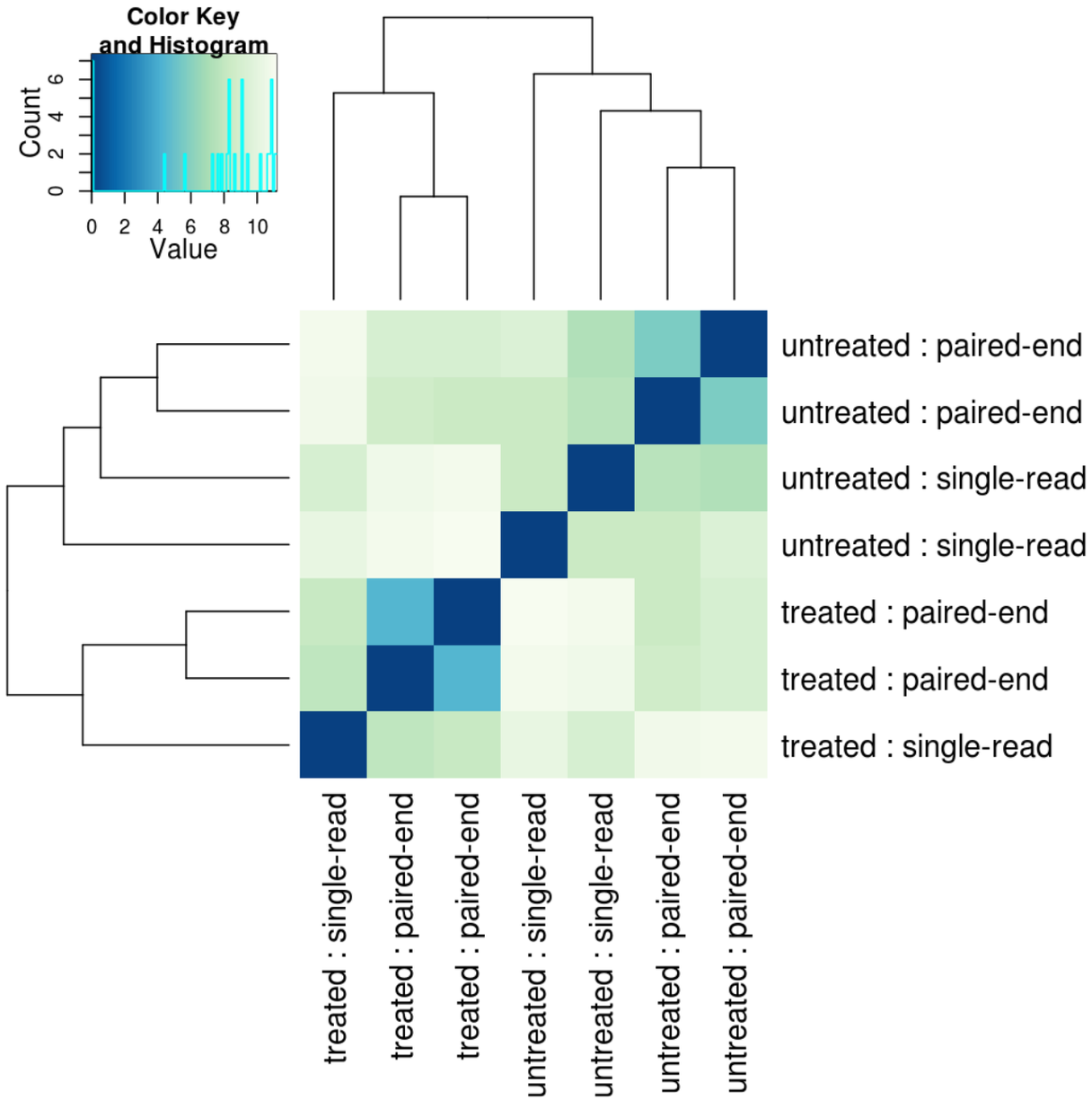


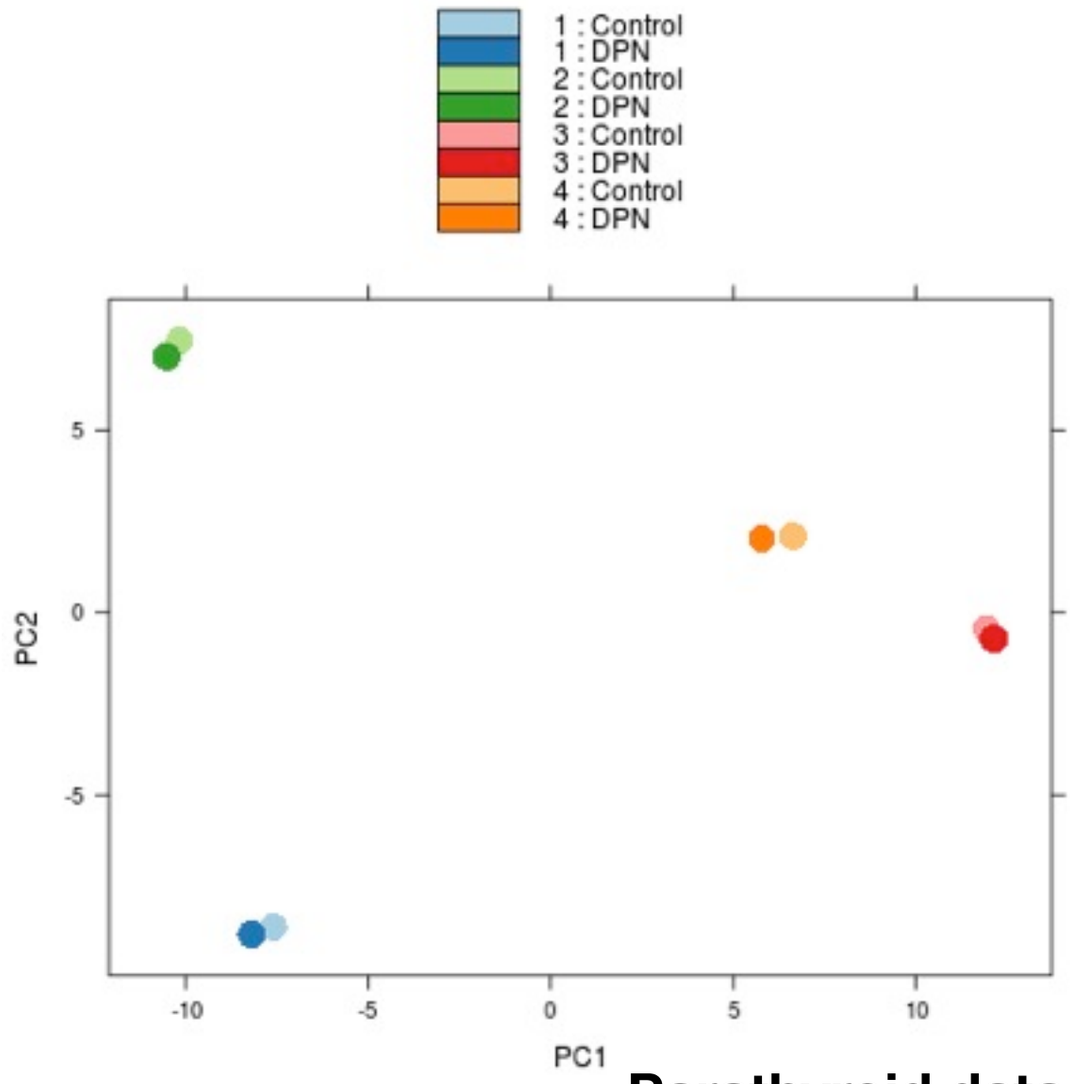
Cook’s distance:
Change in fitted coefficients if the sample were removed





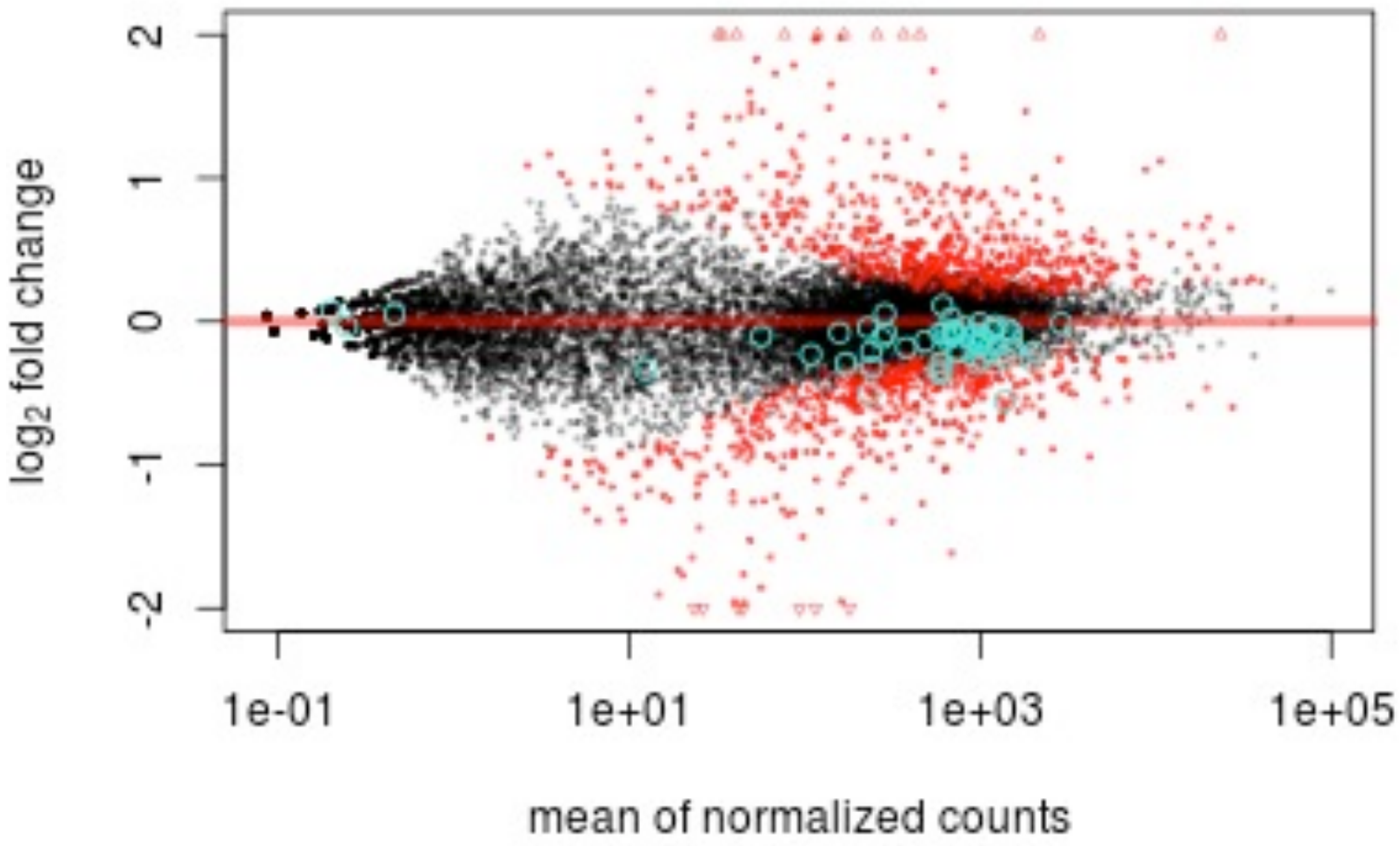
regularized log-transformation: visualization, clustering, PCA





Parathyroid data, Haglung et al. 2012

GSEA with shrunken log fold changes



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

turquoise circles:

Reactome Path "APC/C-mediated degradation of cell cycle proteins"

56 genes, avg LFC: -0.15, p value: $4 \cdot 10^{-11}$ (t test)

Genes and transcripts

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

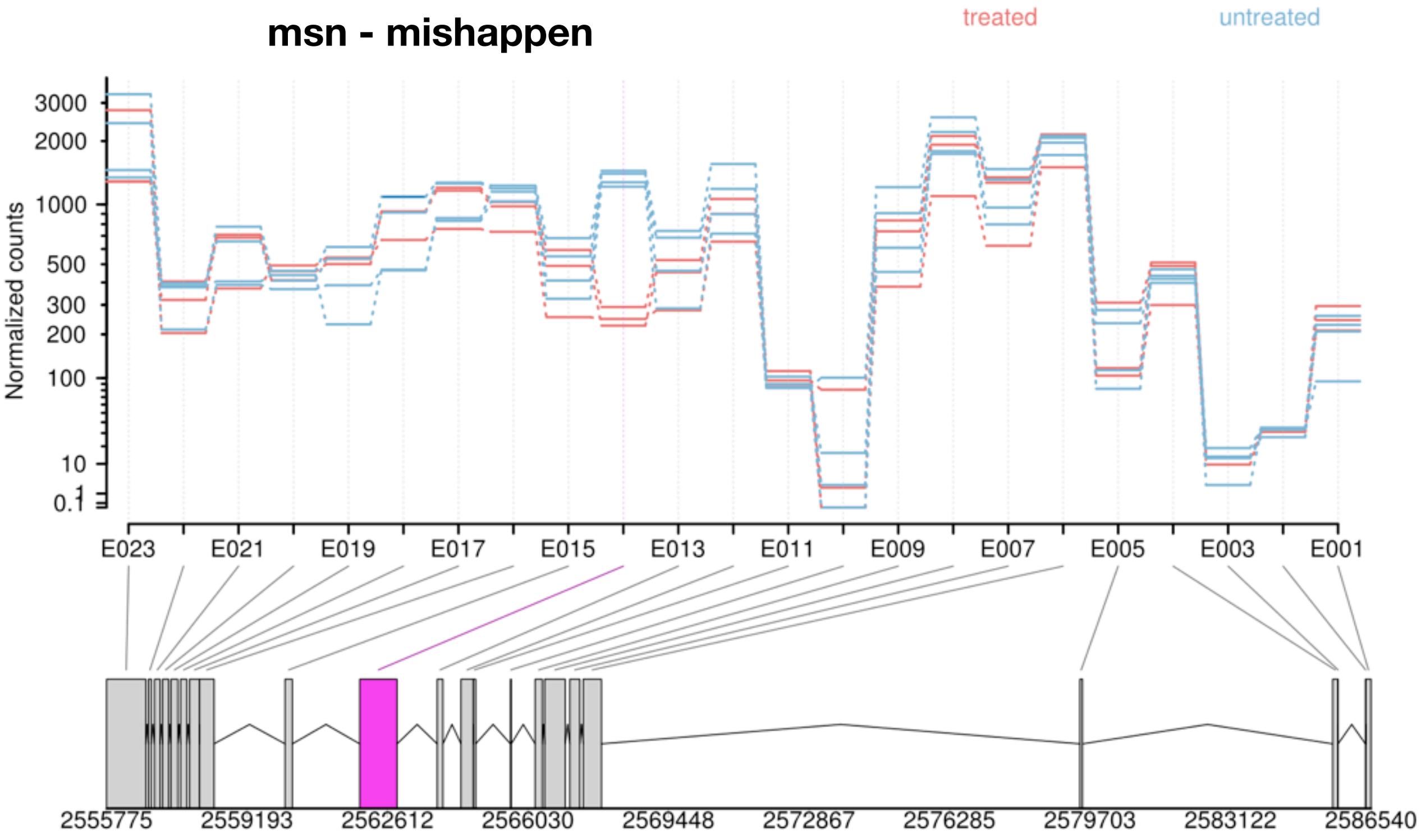
A gene's read count may increase

because the gene produces *more* transcripts

because the gene produces *longer* transcripts

How to look at gene sub-structure?

Differential exon usage



DEXSeq

$$K_{ijl} \sim \mathrm{NB}(s_j \mu_{ijl}, \alpha_{il})$$

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

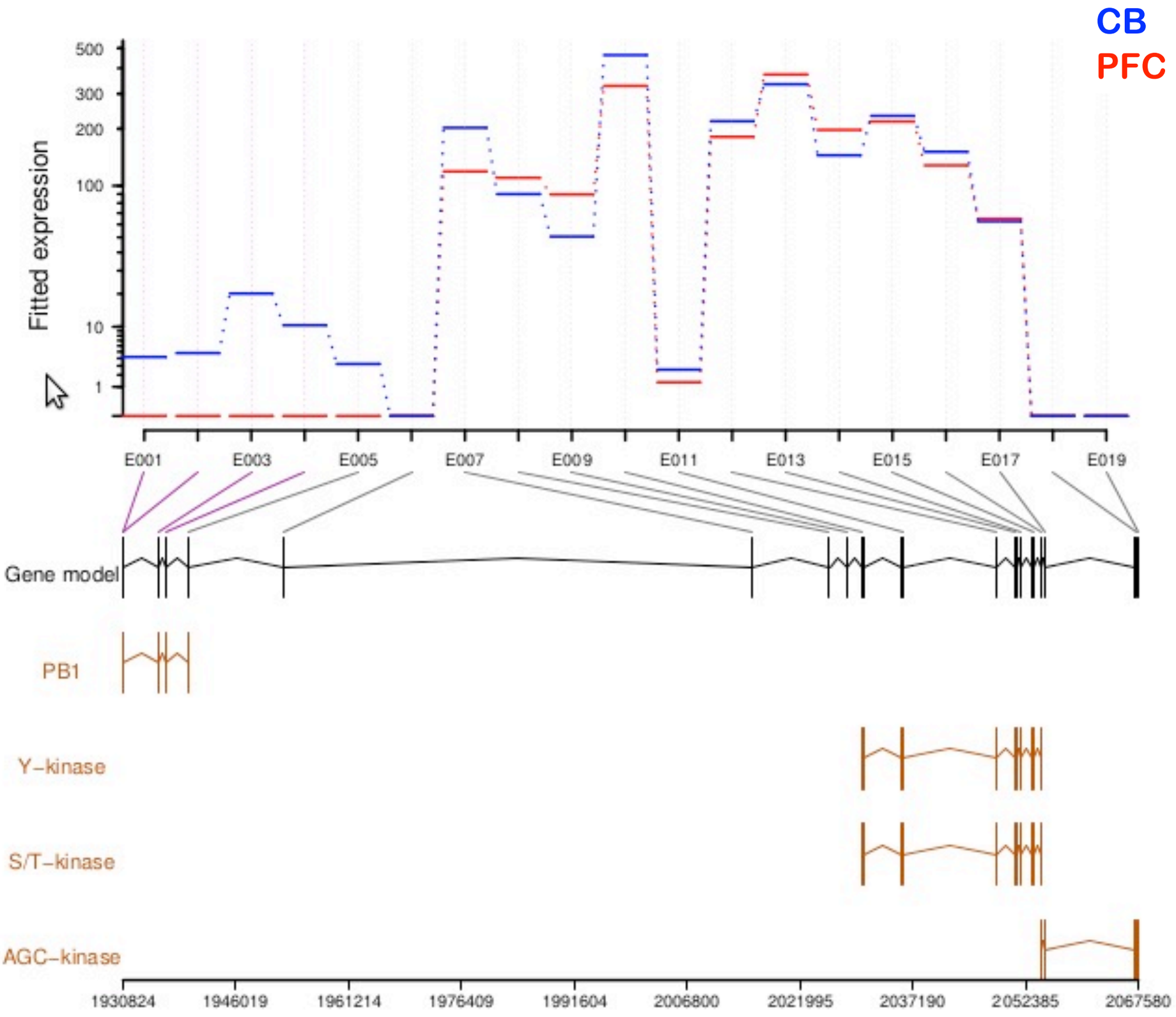
change to fraction of reads for exon l due to treatment

DEXSeq

test for changes in the (relative) usage of exons:

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

Differential exon usage - example



long form:
PKC-zeta

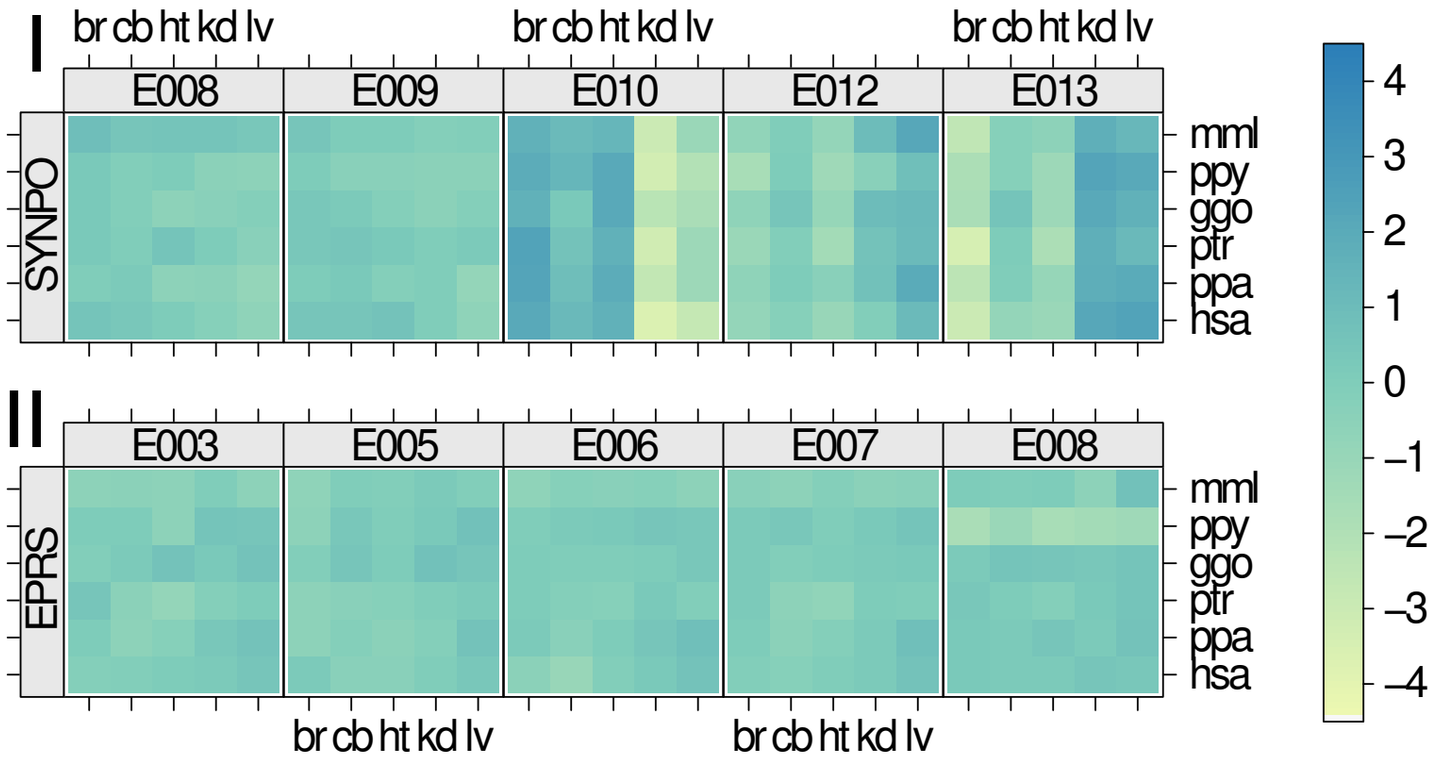
N-term.
truncated:
PKM-zeta

Drift and conservation of differential exon usage across tissues in primate species

Alejandro Reyes* [†], **Simon Anders*** [†], **Robert J. Weatheritt** [‡], **Toby Gibson** [‡], **Lars M. Steinmetz** [† §] **and Wolfgang Huber** [†]

[†]Genome Biology Unit,[‡]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, and [§]Stanford Genome Technology Center, Stanford University, 855 California Ave, Palo Alto, CA 94304, USA

PNAS (accepted)



Most tissue-dependent alternative exon usage in primates is

- **low amplitude,**
- **noise**
- **little evidence for conservation**

A significant fraction is

- **high amplitude**
- **conserved**
- **often associated with UTRs (mRNA life-cycle & localisation, translation regulation) and unstructured protein regions (PPI)**

Simon Anders
Joseph Barry
Bernd Fischer
Julian Gehring
Bernd Klaus
Felix Klein
Michael Love
Malgorzata Oles
Aleksandra Pekowska
Paul-Theodor Pyl
Alejandro Reyes
Jan Swedlow

Collaborators
Lars Steinmetz
Robert Gentleman (Genentech)
Michael Boutros (DKFZ)
Martin Morgan (FHCRC)
Jan Korbel
Magnus Rattray (Manchester)

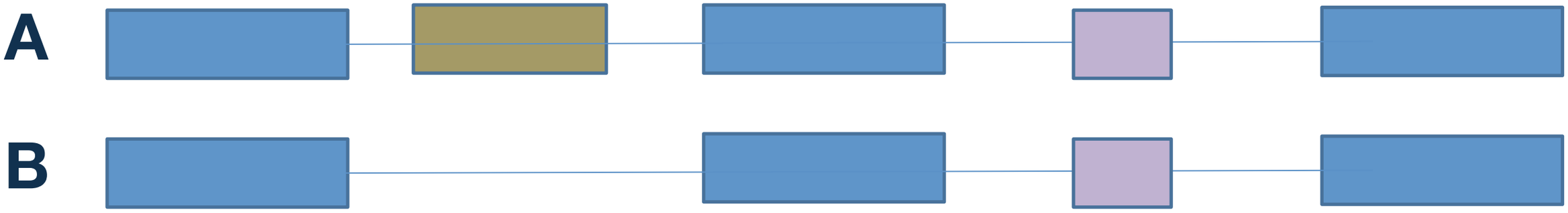
Special thanks
to all users who provided feed-back

Counting rules (RNA-Seq)

- Count unique fragments, not bases
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality is bad
  - (for paired-end reads) the mates do not map to the same gene

Why we discard non-unique alignments

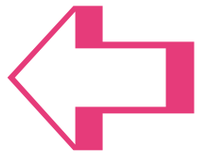
Differential usage of exons or of isoforms?



Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.3.0
proper comparison, PFC vs CB:			
PFC 1 – PFC 6	CB 1, CB 2	650	114
PFC 1, PFC 2	CB 1, CB 2	56	230
PFC 1, PFC 3	CB 1, CB 2	18	361
PFC 1, PFC 4	CB 1, CB 2	26	370
PFC 1, PFC 5	CB 1, CB 2	32	215
PFC 1, PFC 6	CB 1, CB 2	27	380
mock comparisons, PFC vs PFC :			
PFC 1, PFC 3	PFC 2, PFC 4	3	405
PFC 1, PFC 2	PFC 3, PFC 4	0	399
PFC 1, PFC 4	PFC 2, PFC 3	244	590
PFC 1, PFC 3	PFC 2, PFC 5	2	628
PFC 1, PFC 2	PFC 3, PFC 5	1	499
PFC 1, PFC 5	PFC 2, PFC 3	2	555
PFC 1, PFC 4	PFC 2, PFC 5	2	460
PFC 1, PFC 2	PFC 4, PFC 5	2	504
PFC 1, PFC 5	PFC 2, PFC 4	2	308
PFC 1, PFC 4	PFC 3, PFC 5	10	497
PFC 1, PFC 3	PFC 4, PFC 5	5	554
PFC 1, PFC 5	PFC 3, PFC 4	0	353
PFC 2, PFC 4	PFC 3, PFC 5	1	476
PFC 2, PFC 3	PFC 4, PFC 5	10	823
PFC 2, PFC 5	PFC 3, PFC 4	0	526

Table S2: Results of the comparison for the Brawand et al. data.

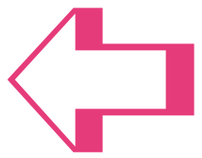
Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.3.0
proper comparison, PFC vs CB:			
PFC 1 – PFC 6	CB 1, CB 2	650	114
PFC 1, PFC 2	CB 1, CB 2	56	230
PFC 1, PFC 3	CB 1, CB 2	18	361
PFC 1, PFC 4	CB 1, CB 2	26	370
PFC 1, PFC 5	CB 1, CB 2	32	215
PFC 1, PFC 6	CB 1, CB 2	27	380
mock comparisons, PFC vs PFC :			
PFC 1, PFC 3	PFC 2, PFC 4	3	405
PFC 1, PFC 2	PFC 3, PFC 4	0	399
PFC 1, PFC 4	PFC 2, PFC 3	244	590
PFC 1, PFC 3	PFC 2, PFC 5	2	628
PFC 1, PFC 2	PFC 3, PFC 5	1	499
PFC 1, PFC 5	PFC 2, PFC 3	2	555
PFC 1, PFC 4	PFC 2, PFC 5	2	460
PFC 1, PFC 2	PFC 4, PFC 5	2	504
PFC 1, PFC 5	PFC 2, PFC 4	2	308
PFC 1, PFC 4	PFC 3, PFC 5	10	497
PFC 1, PFC 3	PFC 4, PFC 5	5	554
PFC 1, PFC 5	PFC 3, PFC 4	0	353
PFC 2, PFC 4	PFC 3, PFC 5	1	476
PFC 2, PFC 3	PFC 4, PFC 5	10	823
PFC 2, PFC 5	PFC 3, PFC 4	0	526



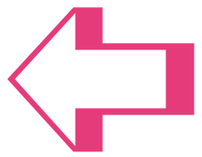
More genes with less replicates

Table S2: Results of the comparison for the Brawand et al. data.

Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.3.0
proper comparison, PFC vs CB:			
PFC 1 – PFC 6	CB 1, CB 2	650	114
PFC 1, PFC 2	CB 1, CB 2	56	230
PFC 1, PFC 3	CB 1, CB 2	18	361
PFC 1, PFC 4	CB 1, CB 2	26	370
PFC 1, PFC 5	CB 1, CB 2	32	215
PFC 1, PFC 6	CB 1, CB 2	27	380
mock comparisons, PFC vs PFC :			
PFC 1, PFC 3	PFC 2, PFC 4	3	405
PFC 1, PFC 2	PFC 3, PFC 4	0	399
PFC 1, PFC 4	PFC 2, PFC 3	244	590
PFC 1, PFC 3	PFC 2, PFC 5	2	628
PFC 1, PFC 2	PFC 3, PFC 5	1	499
PFC 1, PFC 5	PFC 2, PFC 3	2	555
PFC 1, PFC 4	PFC 2, PFC 5	2	460
PFC 1, PFC 2	PFC 4, PFC 5	2	504
PFC 1, PFC 5	PFC 2, PFC 4	2	308
PFC 1, PFC 4	PFC 3, PFC 5	10	497
PFC 1, PFC 3	PFC 4, PFC 5	5	554
PFC 1, PFC 5	PFC 3, PFC 4	0	353
PFC 2, PFC 4	PFC 3, PFC 5	1	476
PFC 2, PFC 3	PFC 4, PFC 5	10	823
PFC 2, PFC 5	PFC 3, PFC 4	0	526



More genes with less replicates



More genes with same-same comparison

Table S2: Results of the comparison for the Brawand et al. data.

likelihood ratio vs Wald test

- **LRT reported to be more powerful in many applications**

- but difficult to reconcile with the β-shrinkage. With null data:

Wald statistics ≈ Normal → uniform p-values

Differences in deviance: not X^2 (pile up at zero) → non-uniform p-values (piling up at 1)

(Without β-shrinkage: similar)

- **Wald tests allow (easily) banded hypotheses tests**

($|\beta| < c$, $|\beta| > c$)