Differential expression analysis

Alternative exon usage

Wolfgang Huber
EMBL

31 October 2013 - Recife

European Molecular Biology Laboratory (EMBL)

European Intergovernmental Research Organisation

- **20 Member States**
- **Founded in 1974**
- **Sites in Heidelberg (D), Cambridge (GB), Roma (I), Grenoble (F), Hamburg (D)**
- **ca. 1400 staff (⊃1100 scientists) representing more than 60 nationalities**

EMBL's five missions

- **Basic research**
- **Development of new technologies and instruments**
- **Technology transfer**
- **Services to the member states**
- **Advanced training**

What can *you* do at EMBL?

Biology

Chemistry

Physics

Mathematics

Informatics

Engineering

www.embl.org/phdprogramme
www.embl.org/postdocs
www.embl.org/jobs

Progress in science is driven by technology

Sequencing - DNA-Seq, RNA-Seq, ChiP-Seq, HiC

Microscopy & remote sensing- molecular interactions and life-cycles in single, live cells

Large scale perturbation libraries - RNAi, drugs

We work on the methods in statistical computing, integrative bioinformatics and mathematical modelling to turn these data into biology.

Research areas

Gene expression

- Statistics - differential expression; alternative exon usage
- 3D structure of DNA (HiC & Co.)
- Single-cell transcriptomics and noise

Simon Anders, Aleksandra Pekoswka, Alejandro Reyes, Jan Swedlow; Tibor Pakozdi

collaborations with L. Steinmetz, P. Bertone, E. Furlong, T. Hiiragi

Cancer Genomics & Precision Oncology

- Somatic mutation detection (incl subclonal)
- Phylogeny inference

Julian Gehring, Paul Pyl

collaborations with C.v.Kalle/M.Schmid, H. Glimm (NCT); J. Korbel

Genetic Interactions, pharmacogenetics (reverse genetics)

- Large-scale combinatorial RNAi & automated microscopy phenotyping
- Cancer mutations & drugs

Joseph Barry, Bernd Fischer, Felix Klein, Malgorzata Oles

collaborations with M.Boutros (DKFZ), T.Zenz (NCT), M. Knop (Uni)

Basics of statistics

- Tools & infrastructure for software ‘publication’
- Teaching

Bernd Klaus, Andrzej Oles

collaborations M.Morgan (FHCRC), R.Gentleman (Genentech)

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**

Count data in HTS

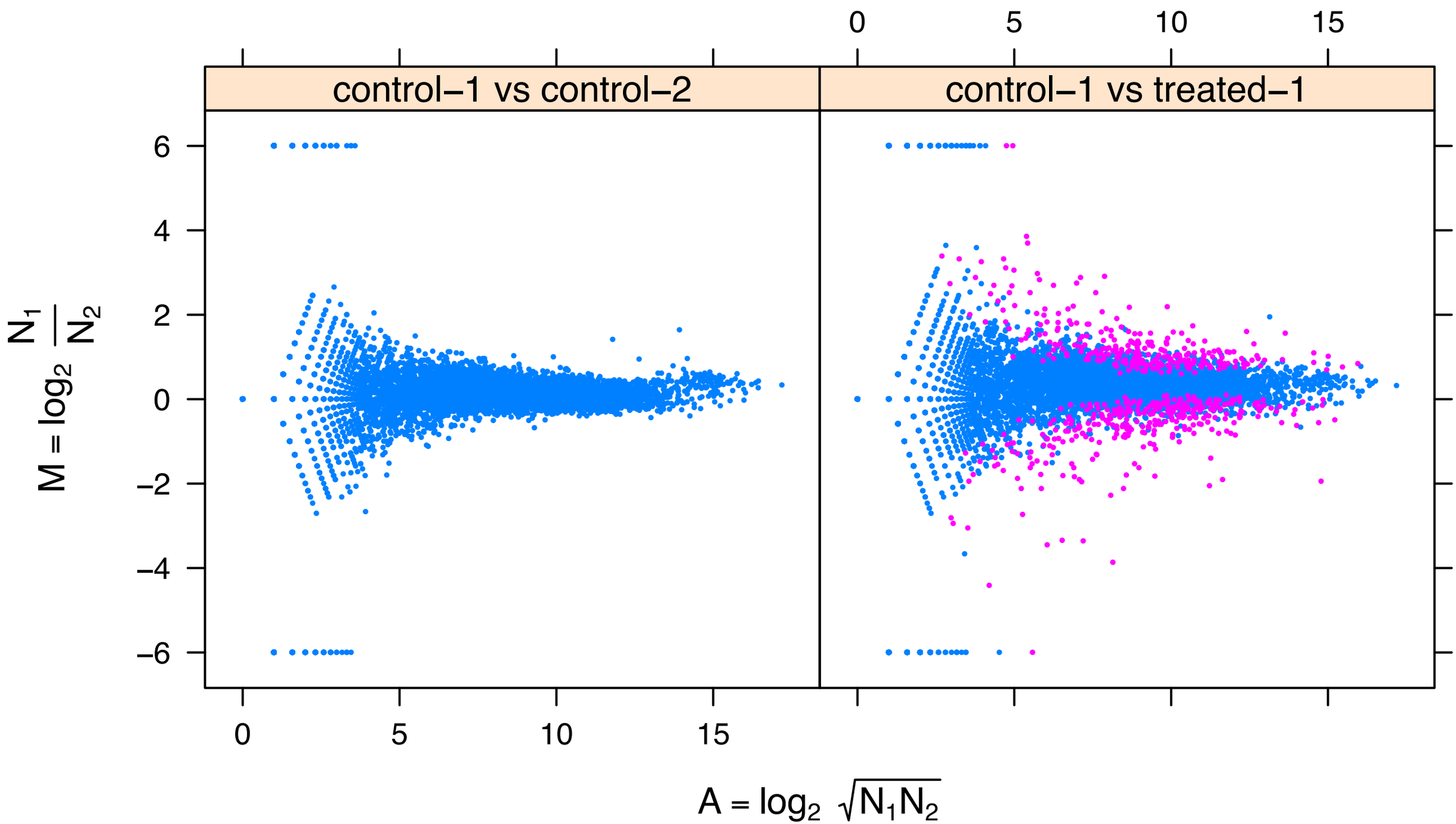
Gene	GliNS1	G144	G166	G179	CB541	CB660
13CDNA73	4	0	6	1	0	5
A2BP1	19	18	20	7	1	8
A2M	2724	2209	13	49	193	548
A4GALT	0	0	48	0	0	0
AAAS	57	29	224	49	202	92
AACS	1904	1294	5073	5365	3737	3511
AADACL1	3	13	239	683	158	40
[...]						

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

Simon Anders

Counting rules

- Count reads, not bases
- 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



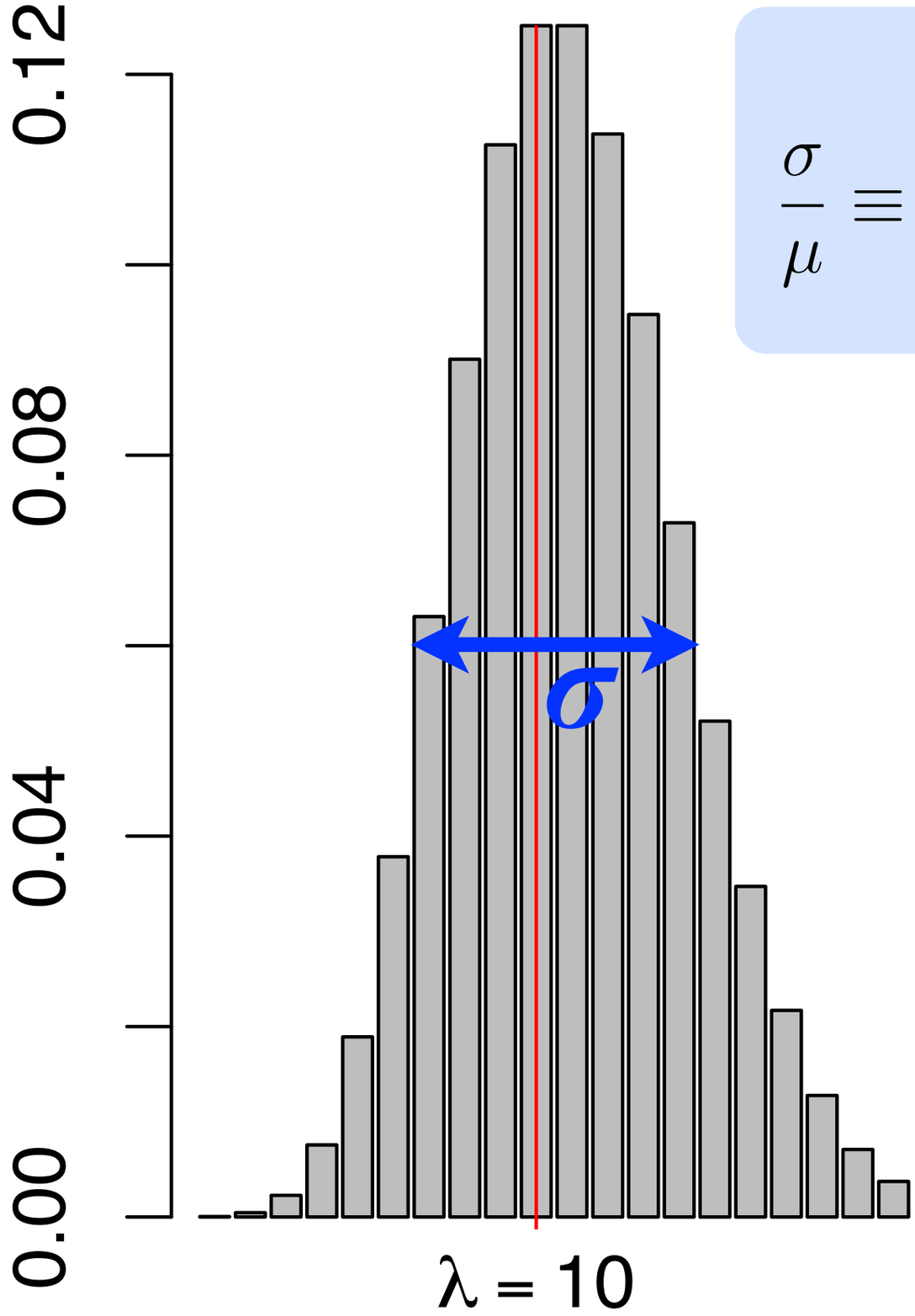
two biological replicates

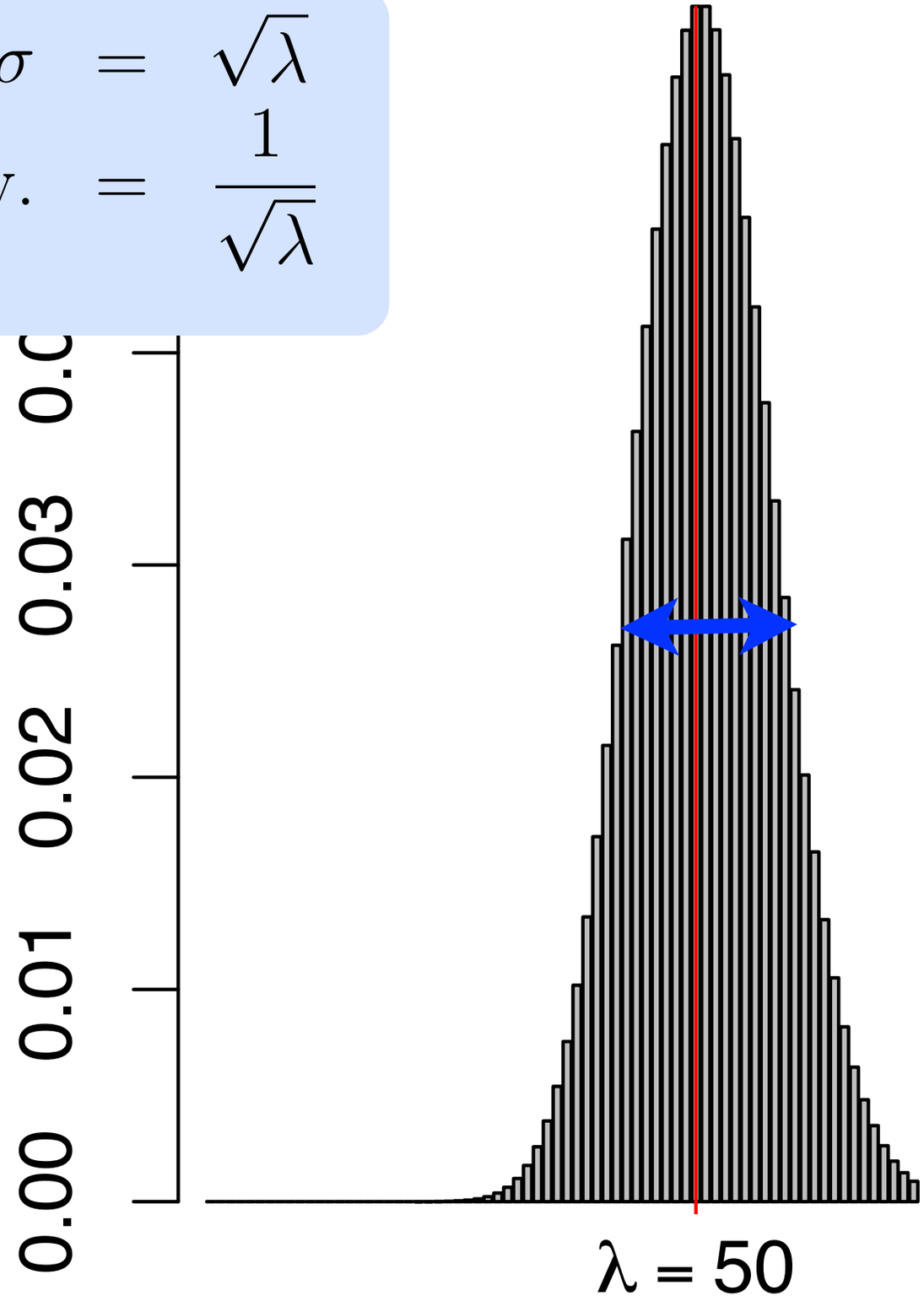
treatment vs control

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

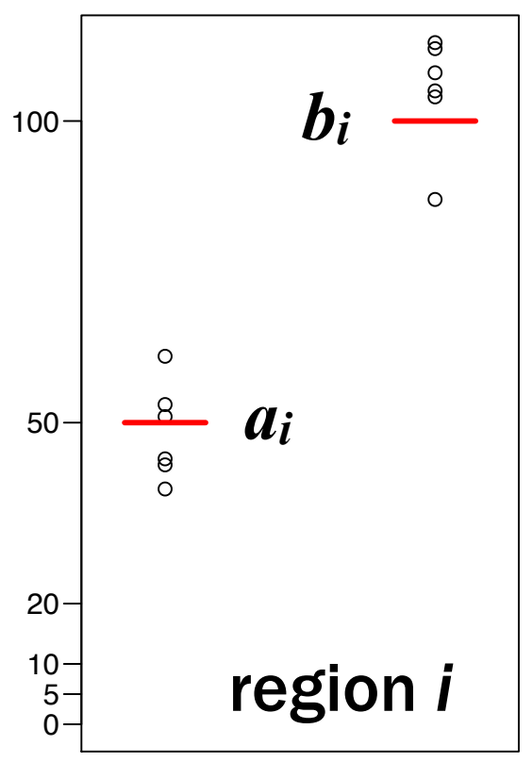
$$N_{ij} \sim \text{Poisson}(\mu_{ij})$$

Noise part

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

Systematic part

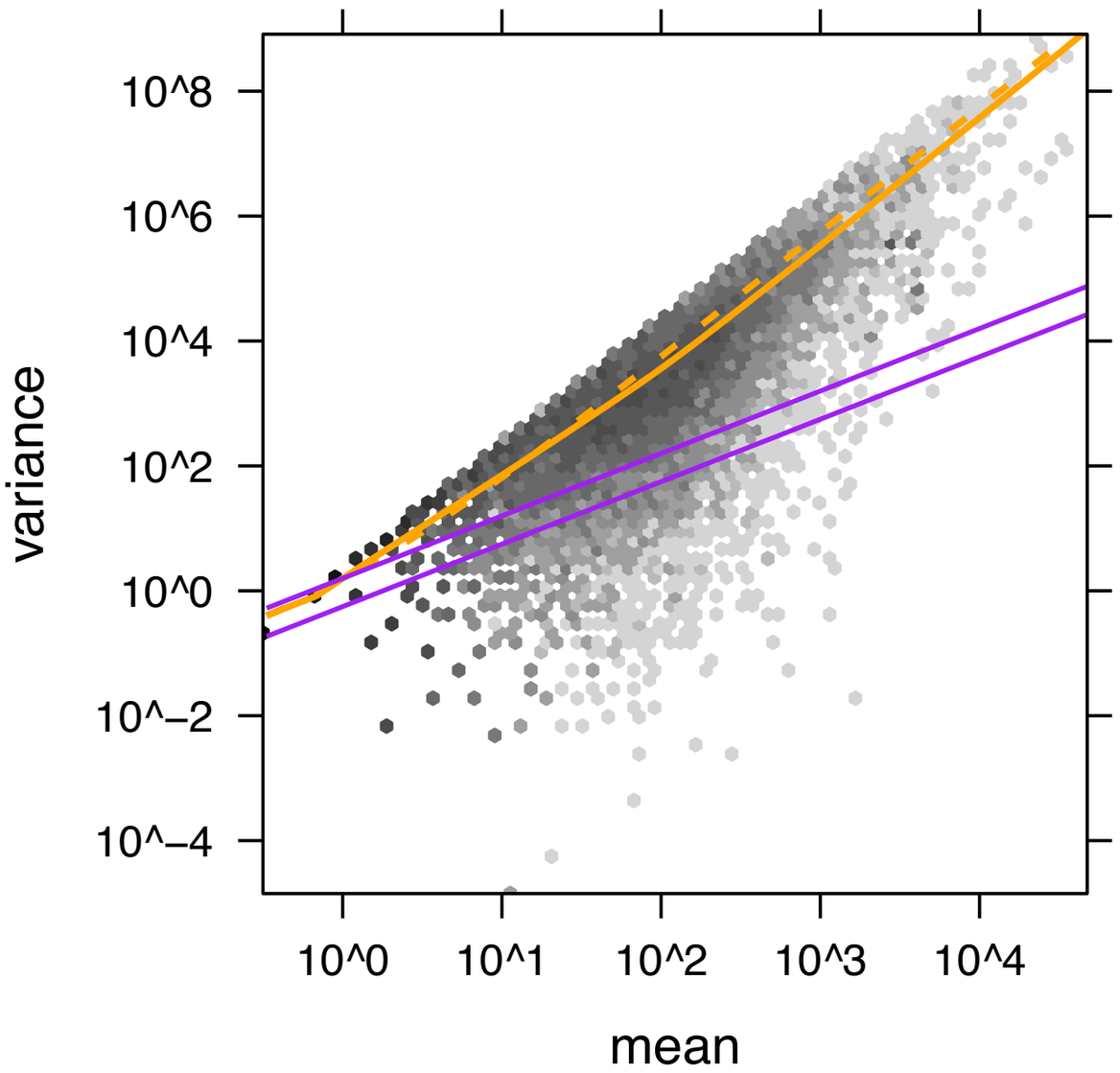
- μ_{ij} **expected count of region *i* in sample *j***
- s_j **library size factor**
- x_{kj} **design matrix**
- β_{ik} **(differential) effect for region *i***



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), ...



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

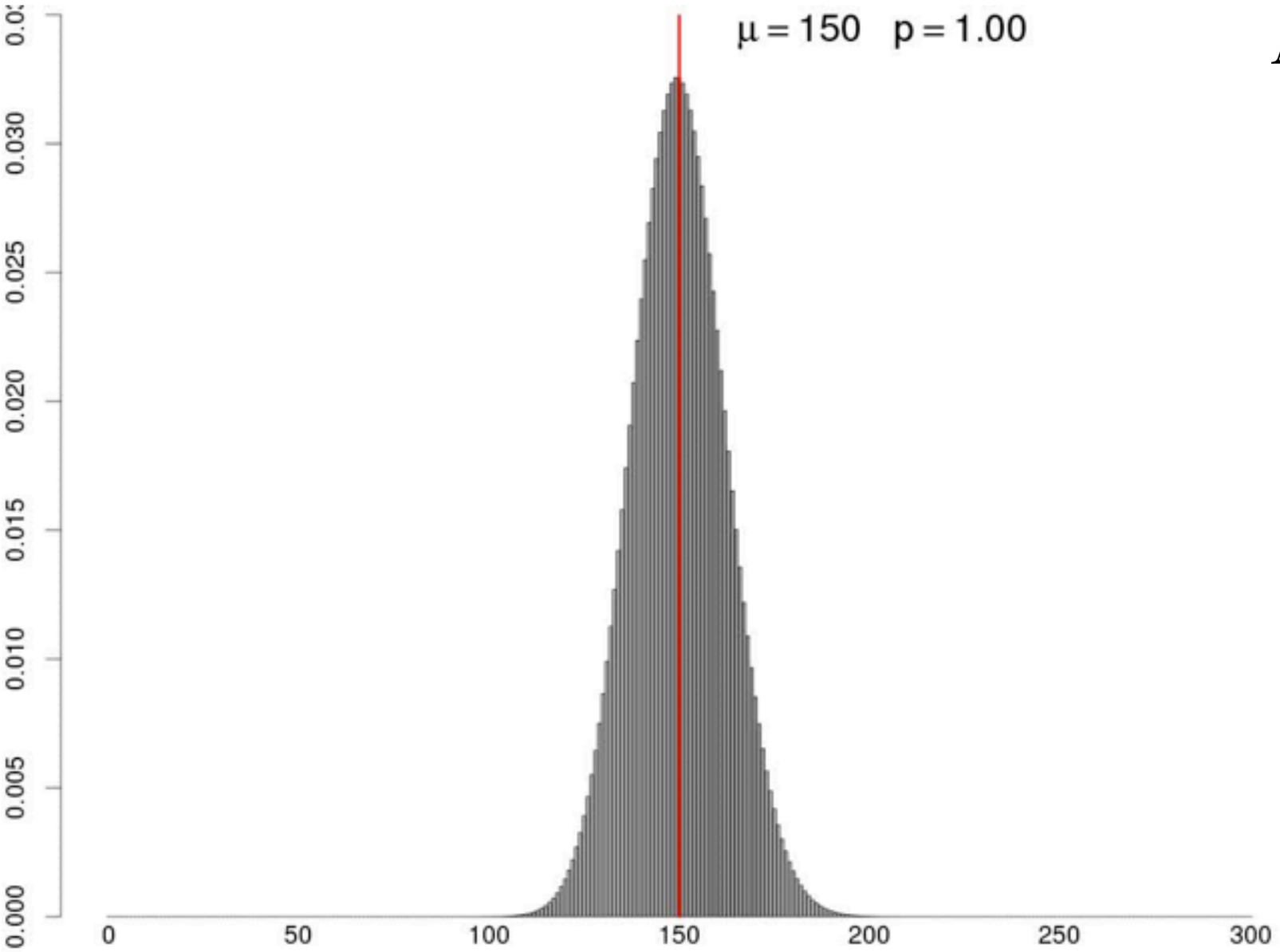
➡ want to use parametric stochastic model to infer tail behaviour (approximately) from low-order moments (mean, variance)

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

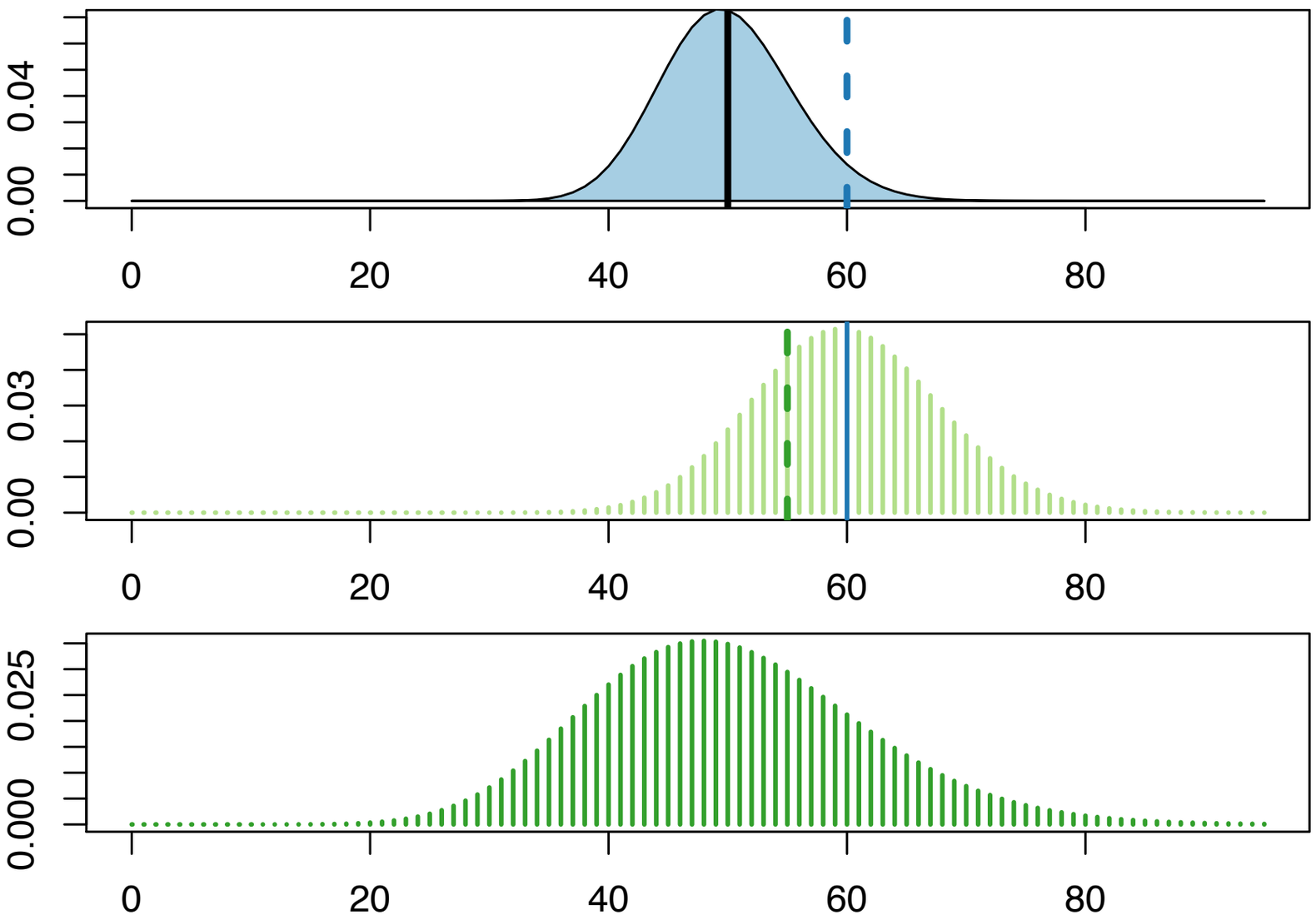
$$\begin{aligned} \alpha &= \frac{1}{r} \\ \mu &= \frac{pr}{1-p} \end{aligned}$$

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

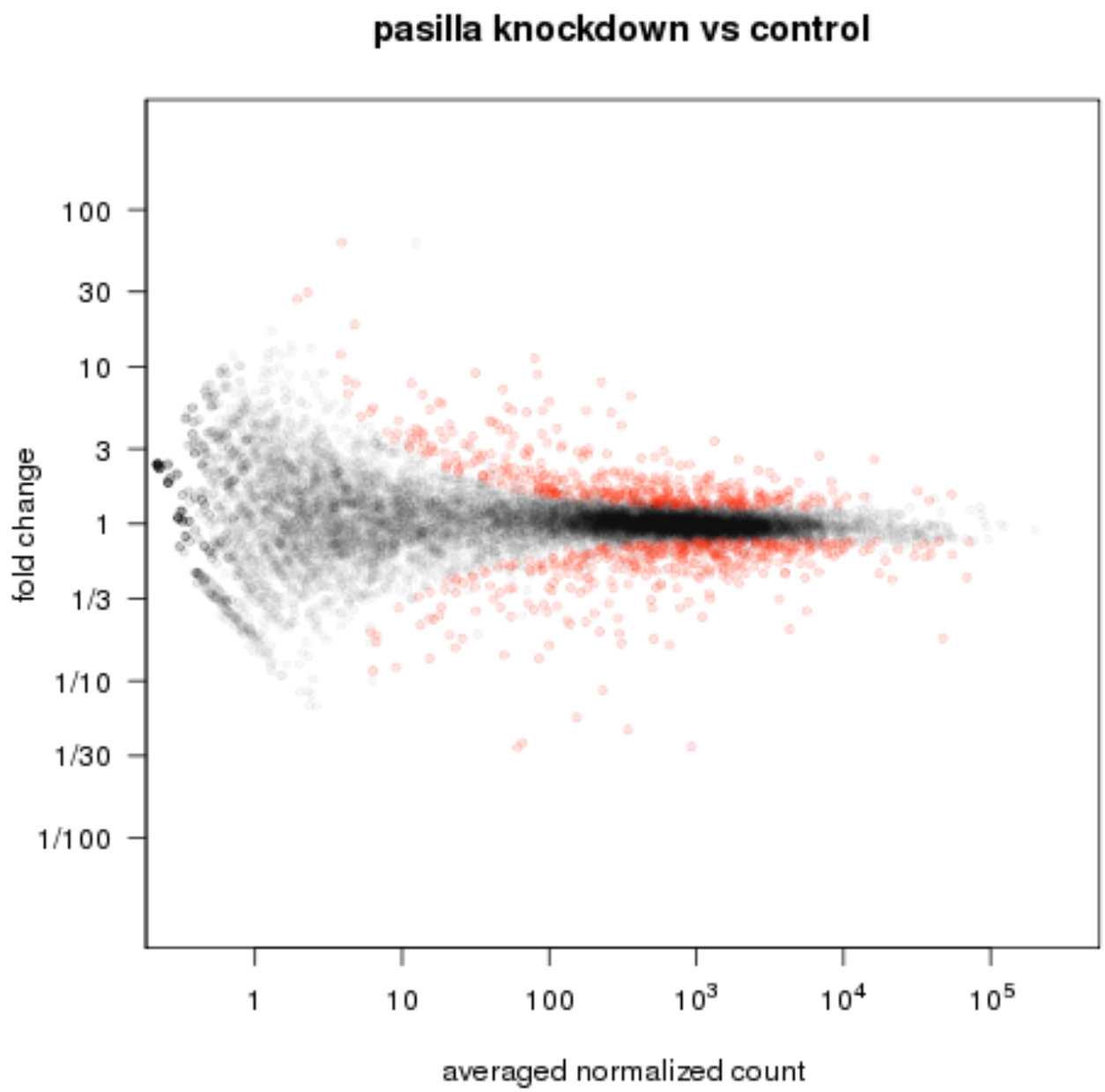
$$\text{var} = \mu + c\,\mu^2$$

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:

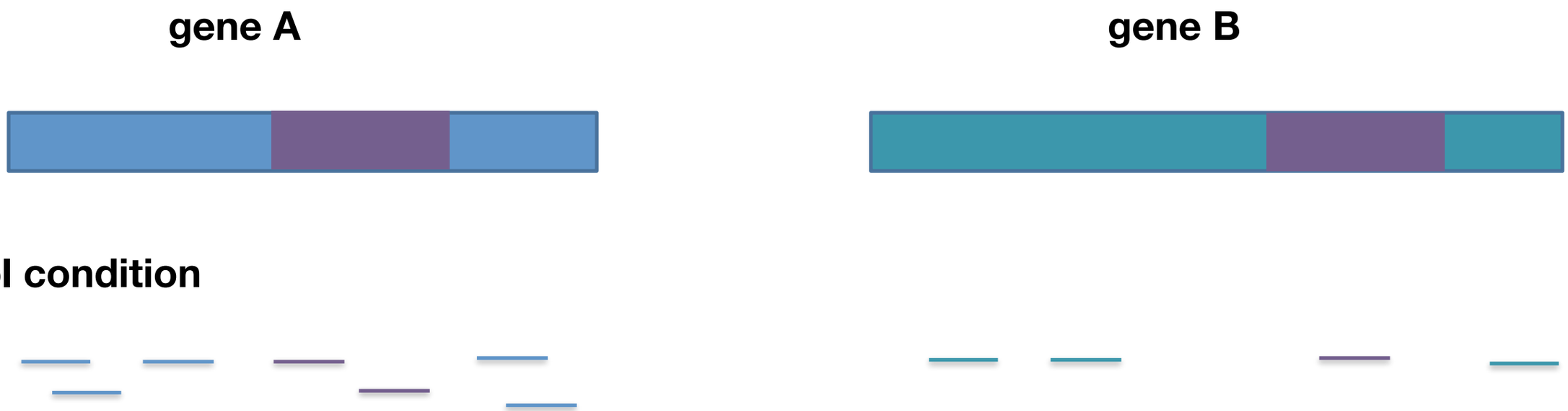
count ~ assayType + treatment + assayType : treatment

reduced model:

count ~ assayType + treatment

Zarnack et al., Cell 2013

Why we discard non-unique alignments



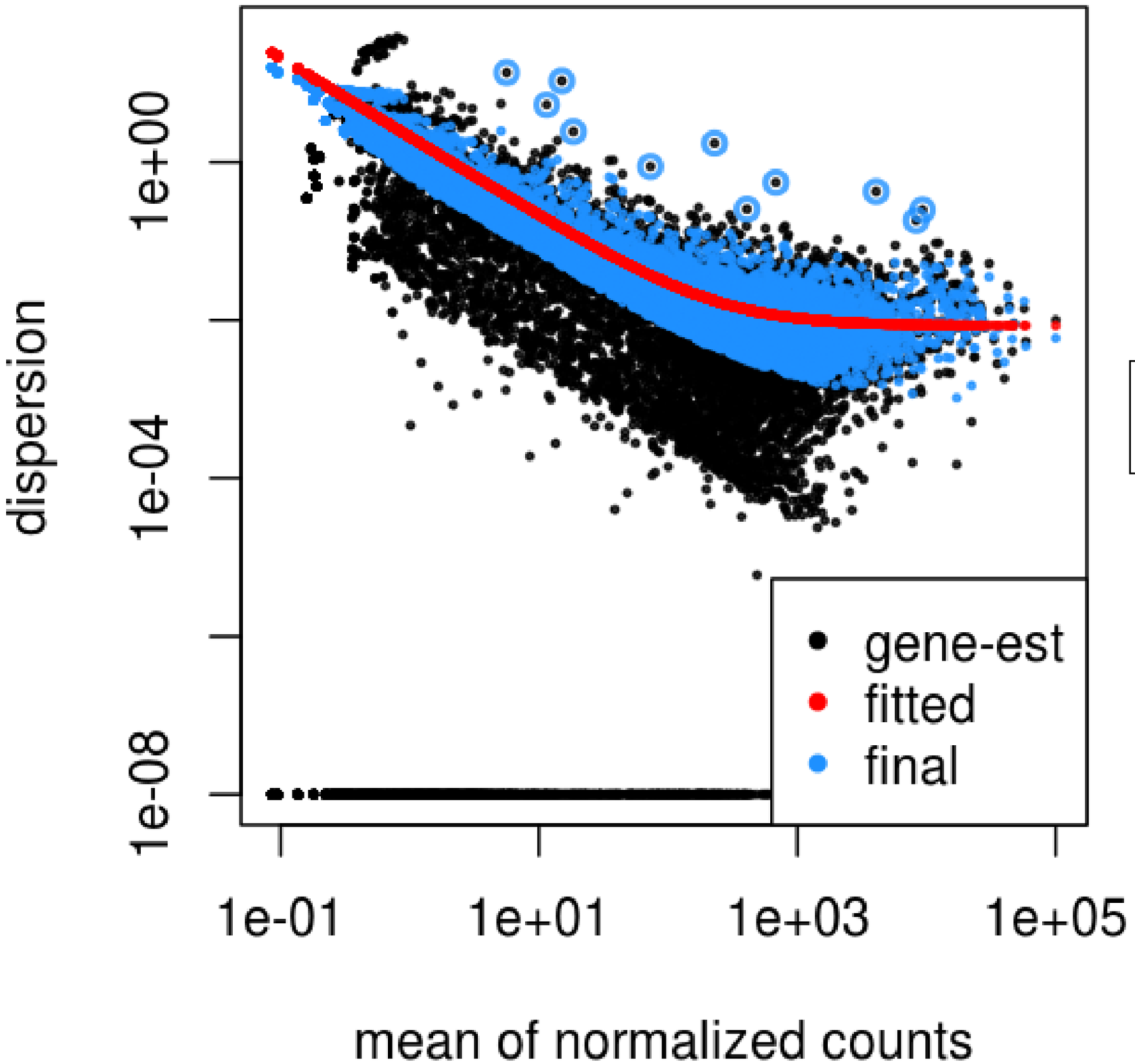
Modelling Variance

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

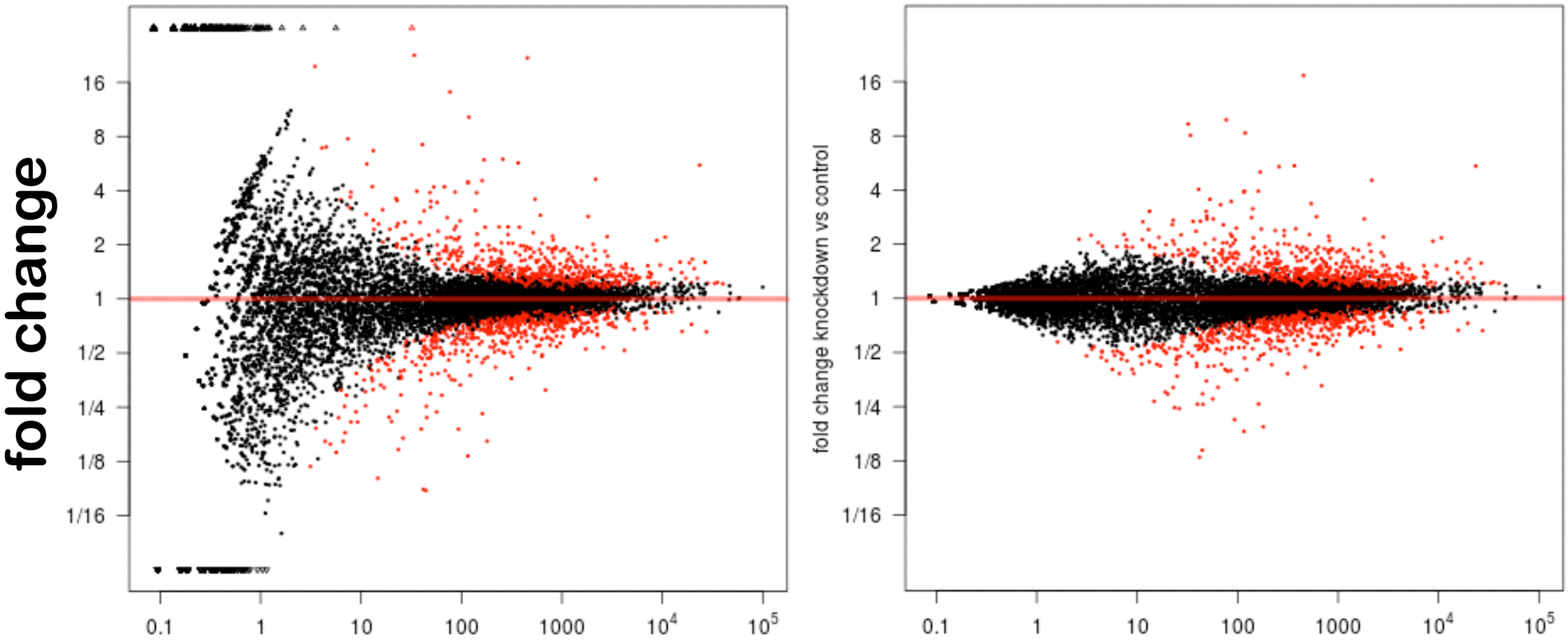
Dispersion estimation: shrinkage



dispersion outliers:

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

Beta (estimated effects): shrinkage



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

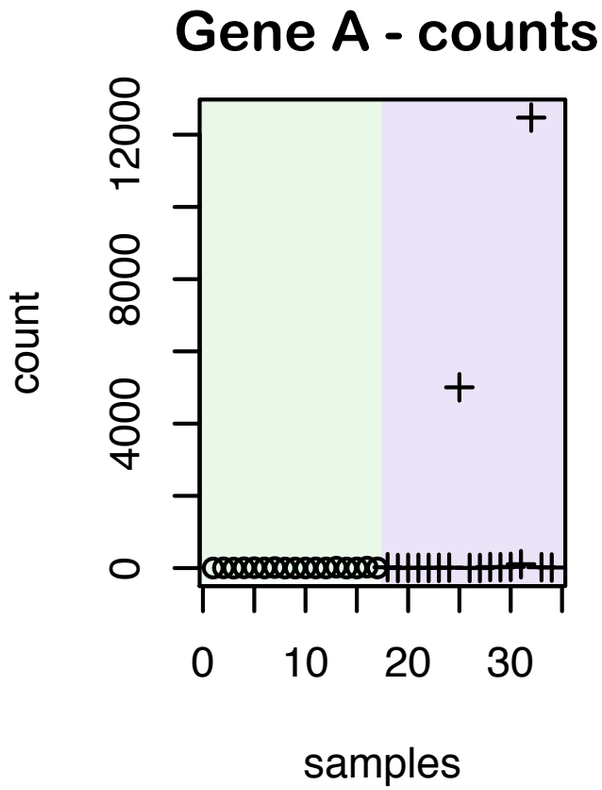
$$\text{prior}(\alpha) = \log(f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma^2_{\text{prior}})$$

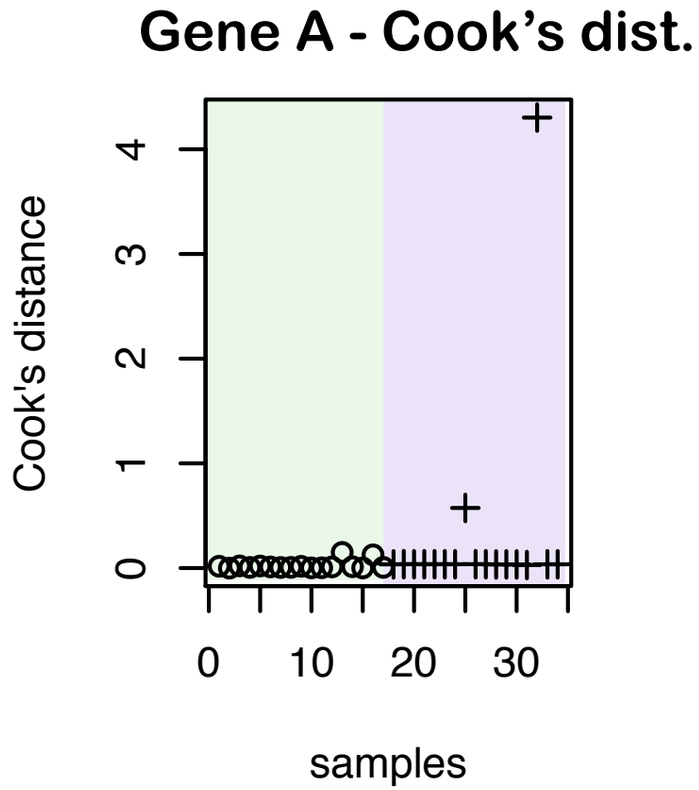
prior on α by ‘information sharing’ across genes

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

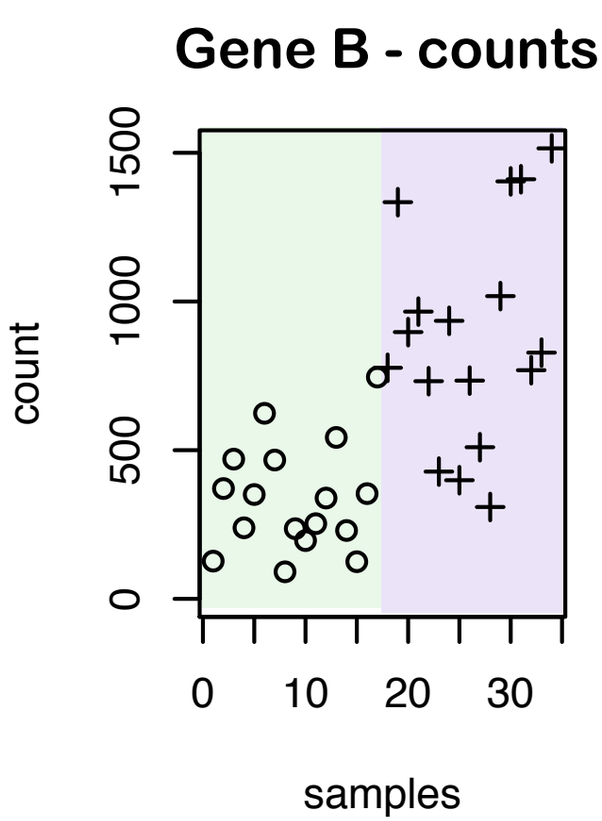
penalized likelihood

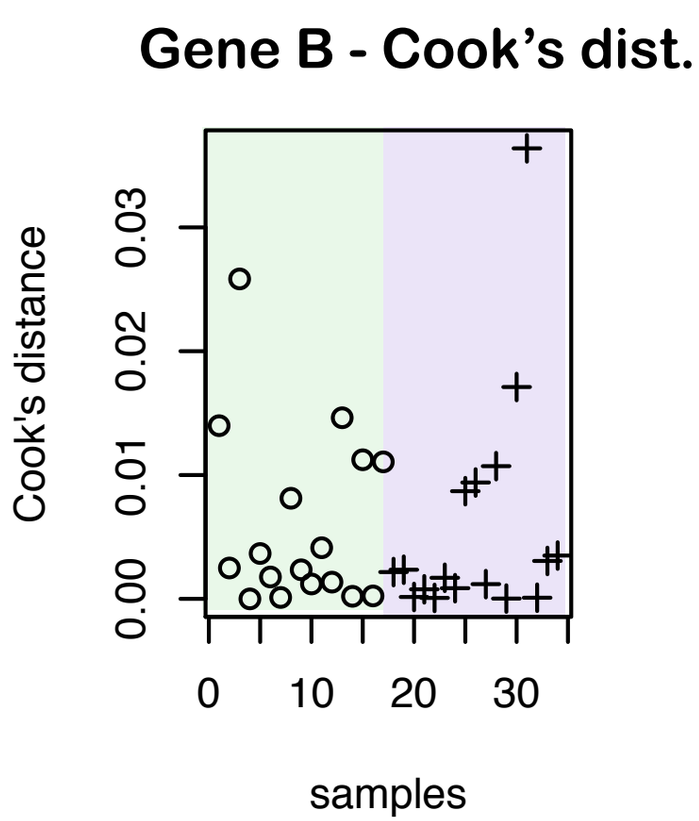
Outlier robustness



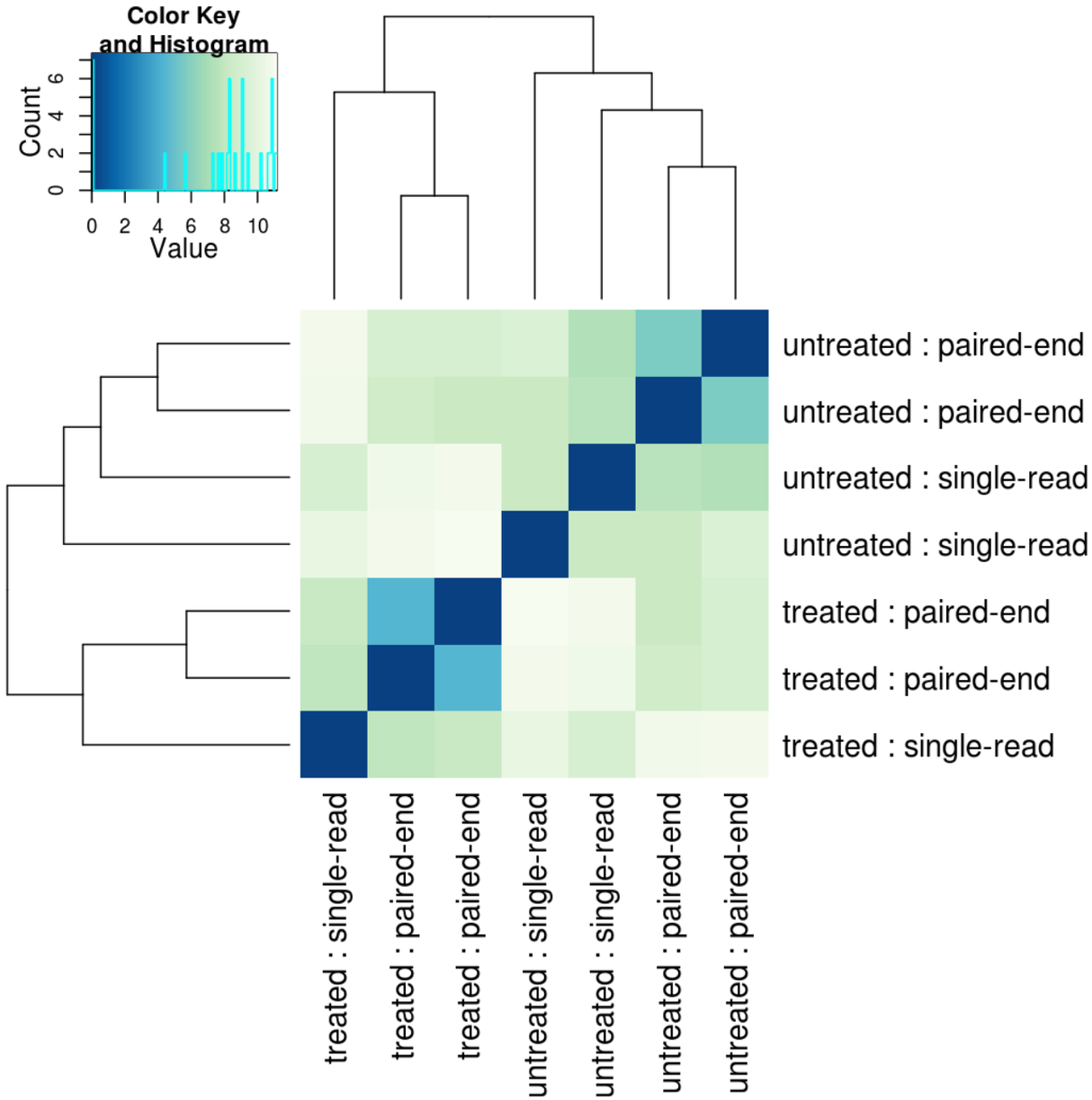


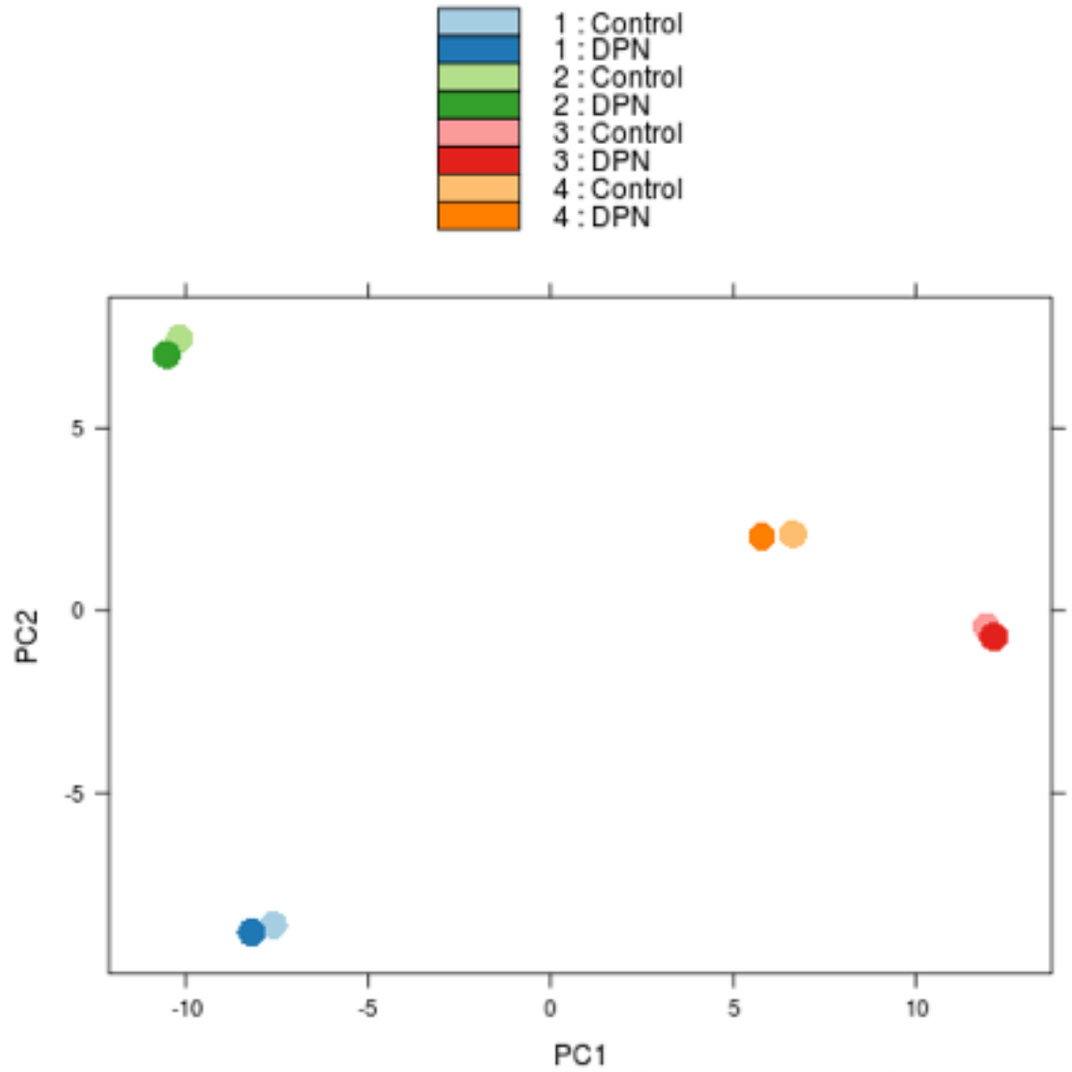
Cook’s distance:
Change in fitted coefficients if the sample were removed





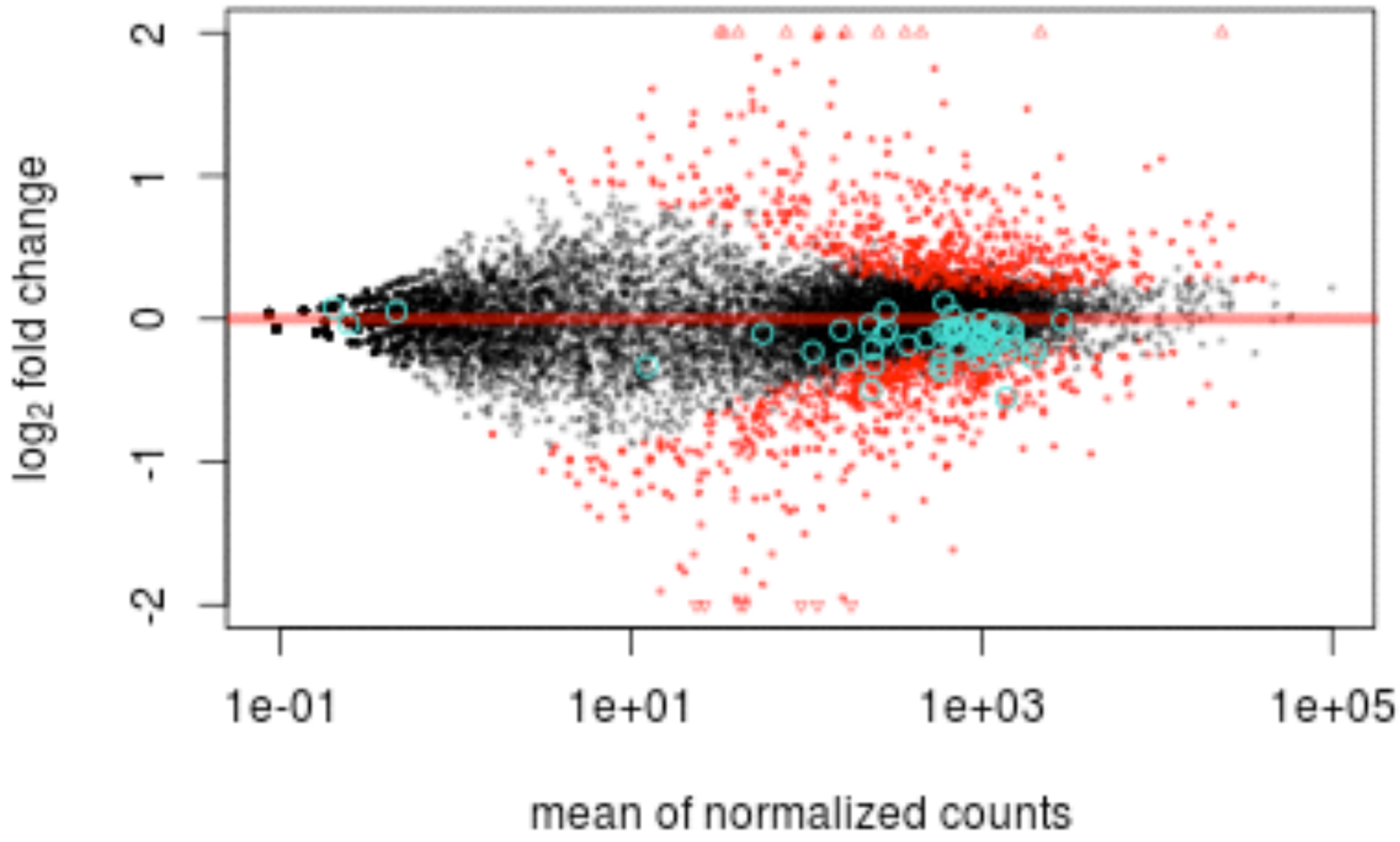
regularized log-transformation: visualization, clustering, PCA





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*.

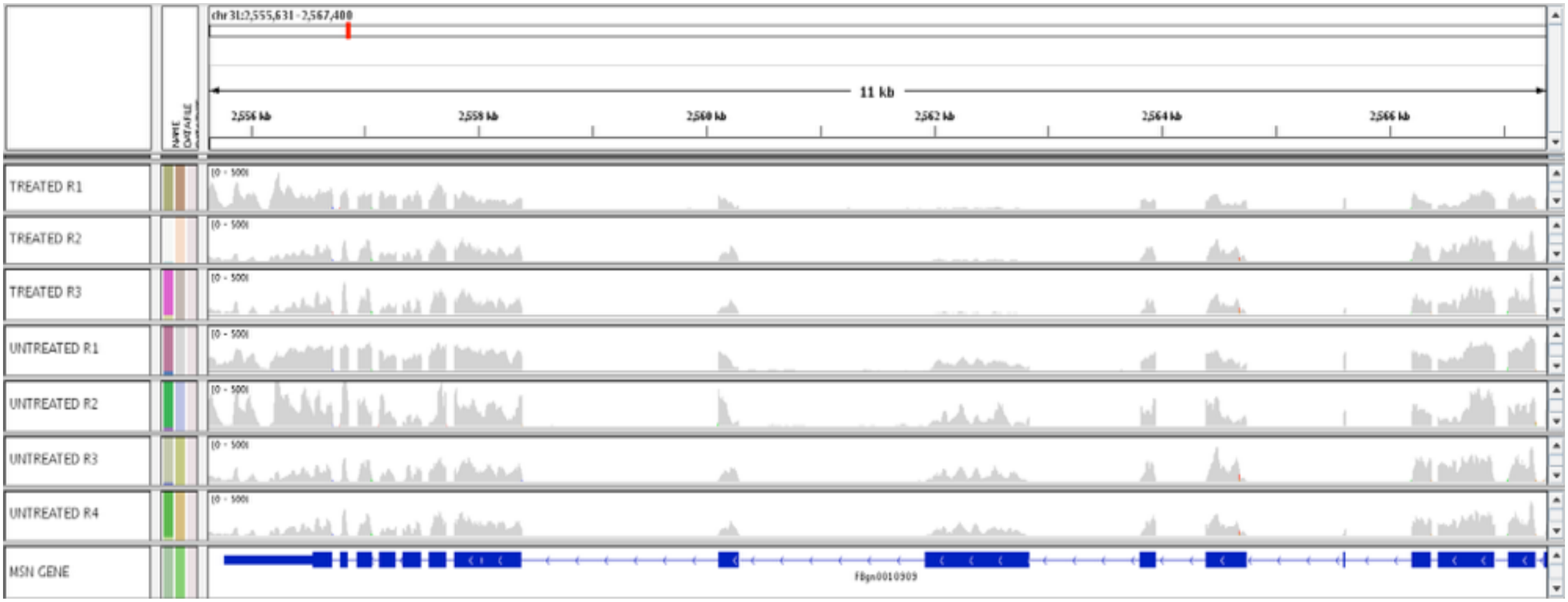
A gene's read count may increase

because the gene produces *more* transcripts

because the gene produces *longer* transcripts

How to look at gene sub-structure?

Alternative isoform regulation



Data: Brooks, ..., Graveley, Genome Res., 2010

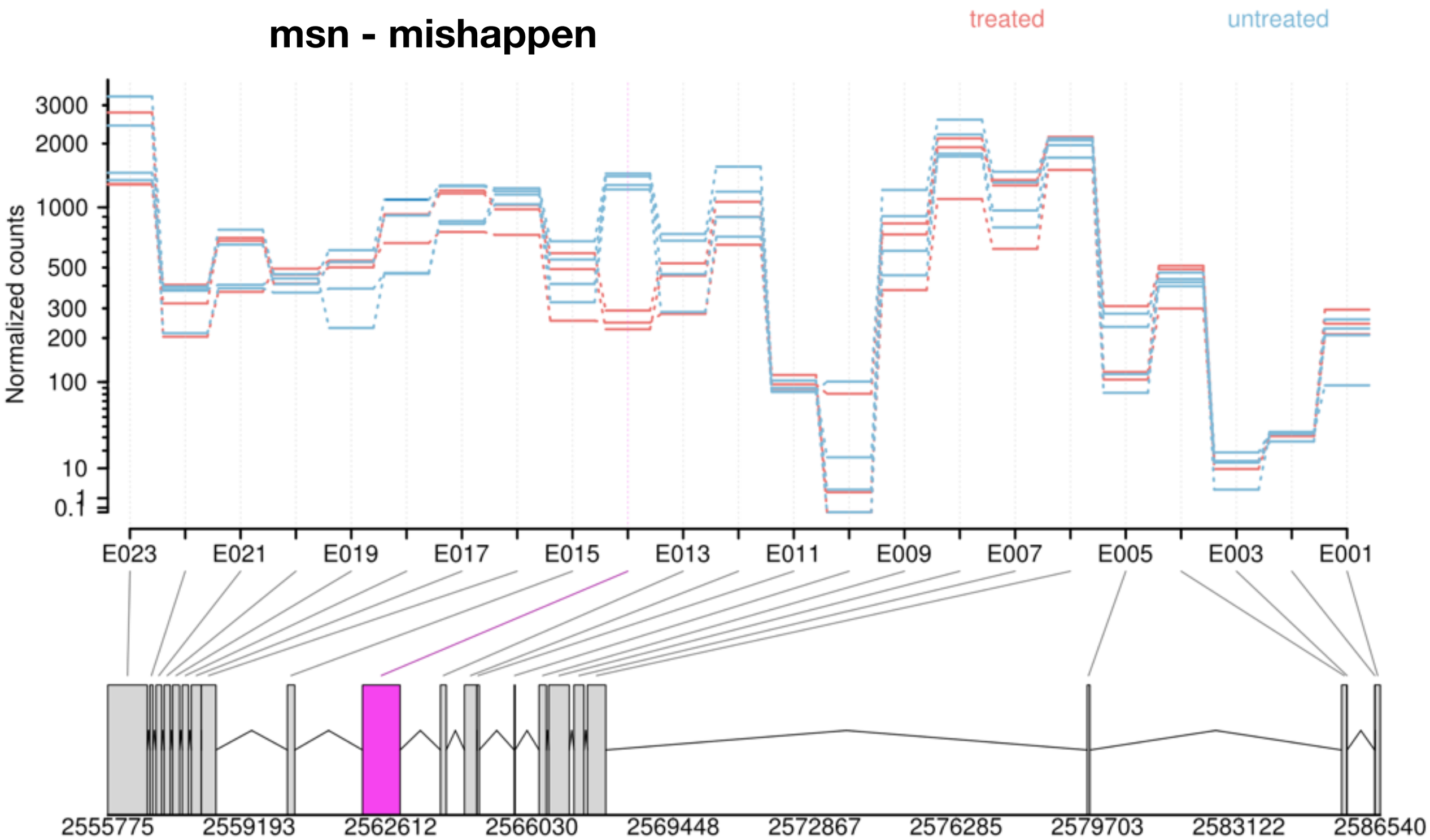
Count table for a gene

number of reads mapped to each exon in a gene

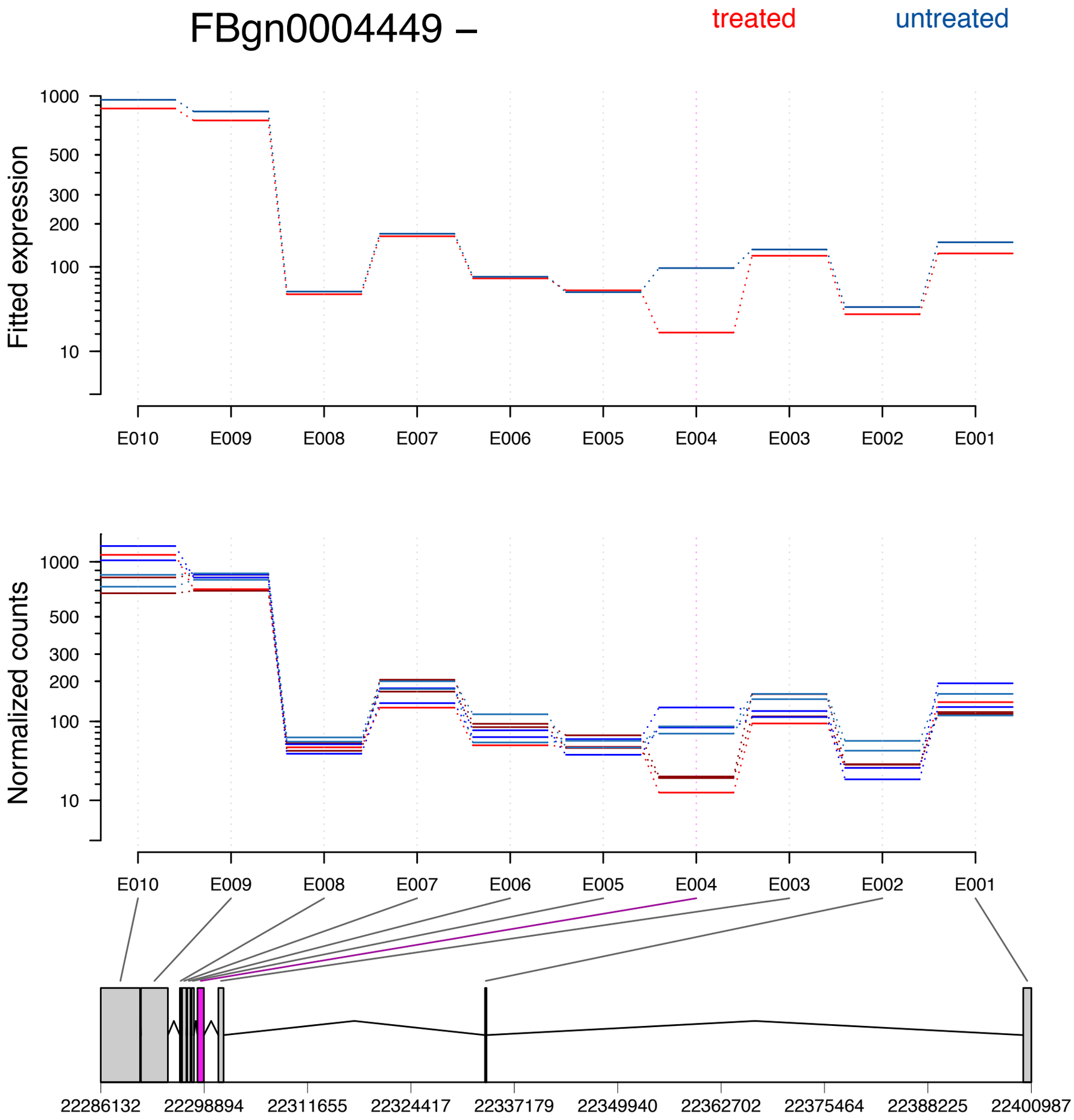
	treated_1	treated_2	control_1	control_2	
E01	398	556	561	456	
E02	112	180	153	137	
E03	238	306	298	226	
E04	162	171	183	146	
E05	192	272	234	199	
E06	314	464	419	331	
E07	373	525	481	404	
E08	323	427	475	373	
E09	194	213	273	176	
E10	90	90	530	398	<--- !
E11	172	207	283	227	
E12	290	397	606	368	<---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

[...]

Differential exon usage



Ten-m


FBgn0004449 –
treated
untreated
Fitted expression
1000
500
300
200
100
10
E010
E009
E008
E007
E006
E005
E004
E003
E002
E001
Normalized counts
1000
500
300
200
100
10
E010
E009
E008
E007
E006
E005
E004
E003
E002
E001
22286132
22298894
22311655
22324417
22337179
22349940
22362702
22375464
22388225
22400987

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

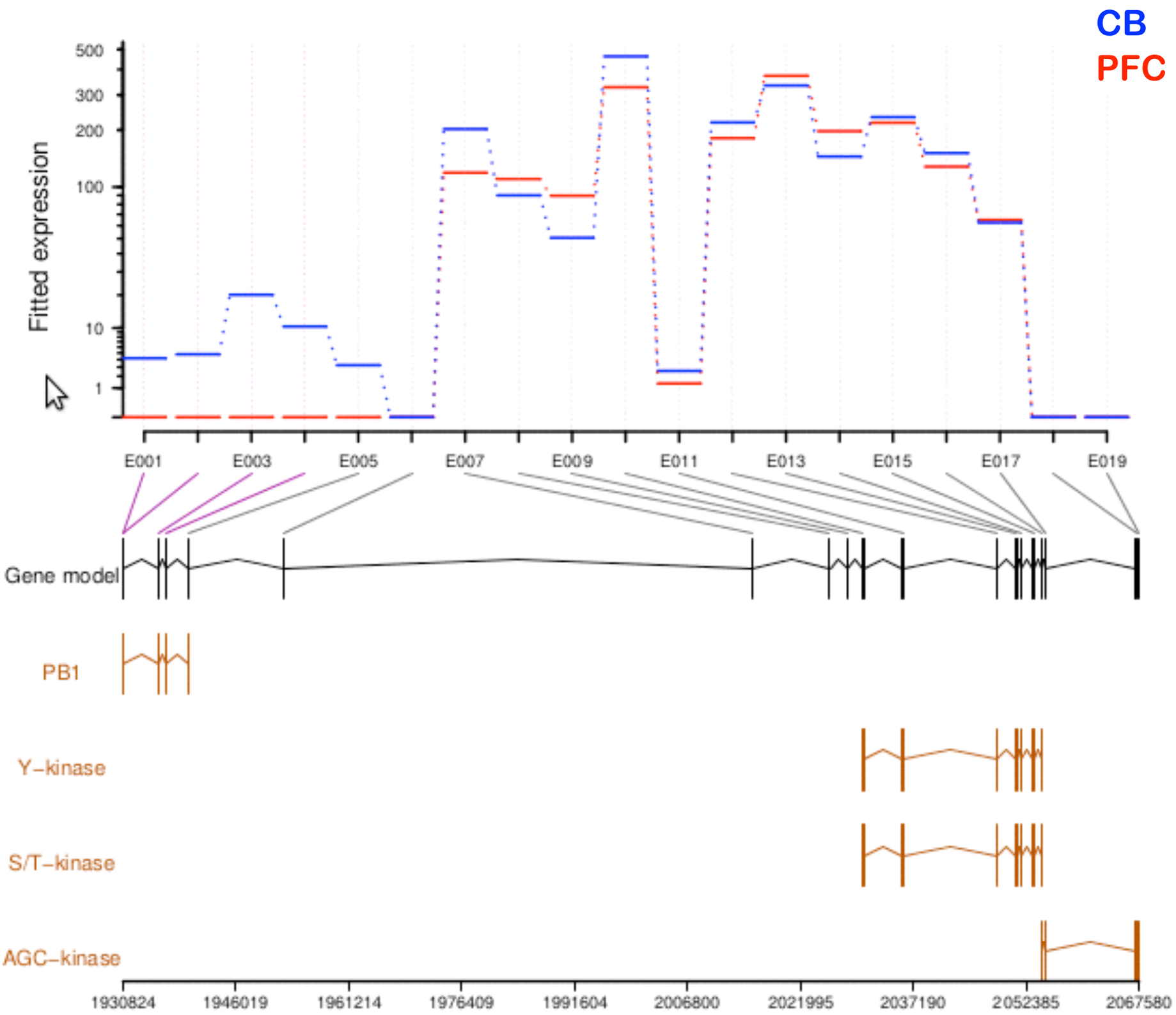
change to fraction of reads for exon l due to treatment

DEXSeq

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

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

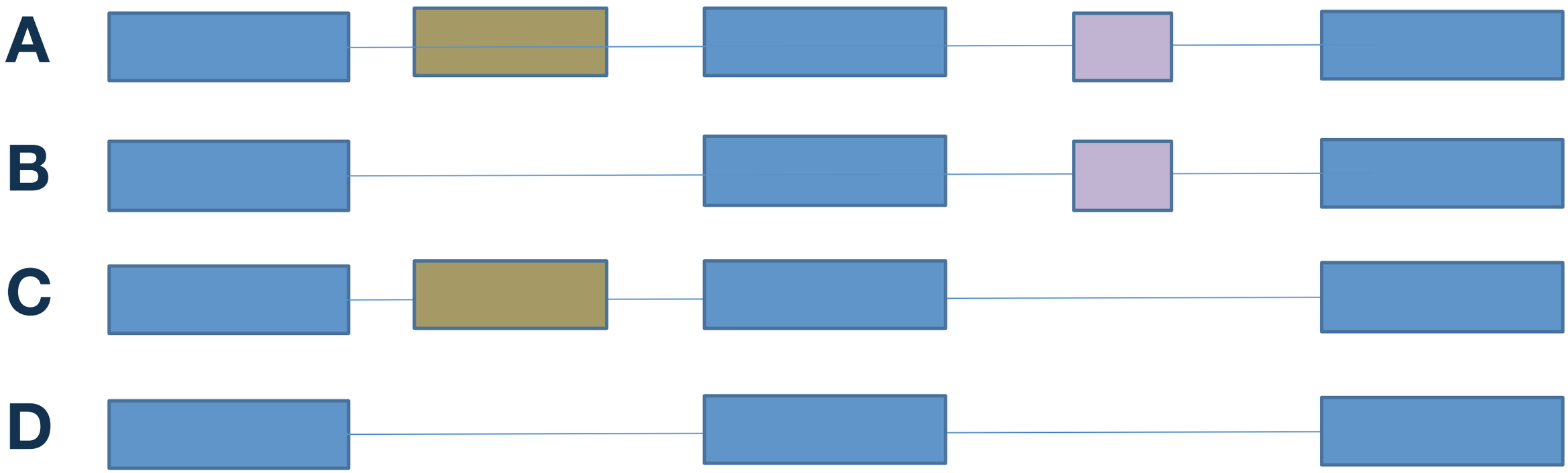
PKC ζ - PKM ζ



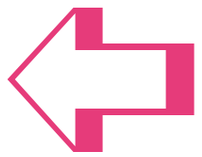
long form:
PKC-zeta

N-term.
truncated:
PKM-zeta

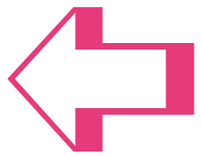
Differential usage of exons or of isoforms?



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

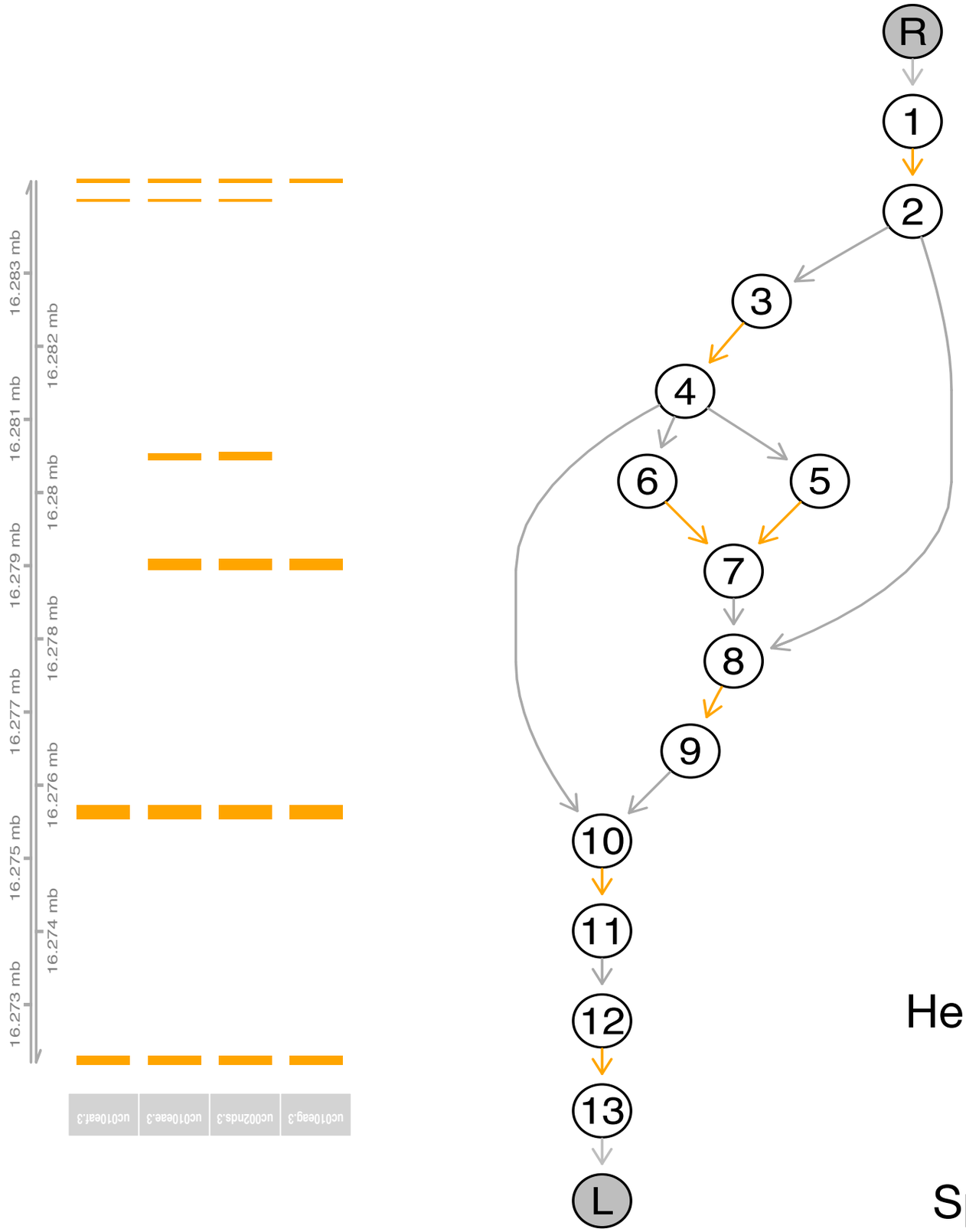


More genes with less replicates



More genes with same-same comparison

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



Splicing Graphs

Heber, Steffen ... Pevzner, Pavel A. Splicing graphs and EST assembly problem Bioinformatics, 18, S181-S188, 2002.

SplicingGraphs package on Bioconductor

Figure 1: Splicing graph representation of the four transcript variants of gene CIB3 (Entrez ID 117286). Left: transcript representation. Right: splicing graph representation. Orange arrows are edges corresponding to exons

Noisy Splicing Drives mRNA Isoform Diversity in Human Cells

Joseph K. Pickrell[1]*, Athma A. Pai[1]*, Yoav Gilad[1]*, Jonathan K. Pritchard[1,2]*

1 Department of Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, **2** Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois, United States of America

Abstract

While the majority of multiexonic human genes show some evidence of alternative splicing, it is unclear what fraction of observed splice forms is functionally relevant. In this study, we examine the extent of alternative splicing in human cells using deep RNA sequencing and *de novo* identification of splice junctions. We demonstrate the existence of a large class of low abundance isoforms, encompassing approximately 150,000 previously unannotated splice junctions in our data. Newly-identified splice sites show little evidence of evolutionary conservation, suggesting that the majority are due to erroneous splice site choice. We show that sequence motifs involved in the recognition of exons are enriched in the vicinity of unconserved splice sites. We estimate that the average intron has a splicing error rate of approximately 0.7% and show that introns in highly expressed genes are spliced more accurately, likely due to their shorter length. These results implicate noisy splicing as an important property of genome evolution.

PLoS Genetics 2010

"... we extrapolate that the majority of different mRNA isoforms present in a cell are not functionally relevant, though most copies of a pre-mRNA produce truly functional isoforms."

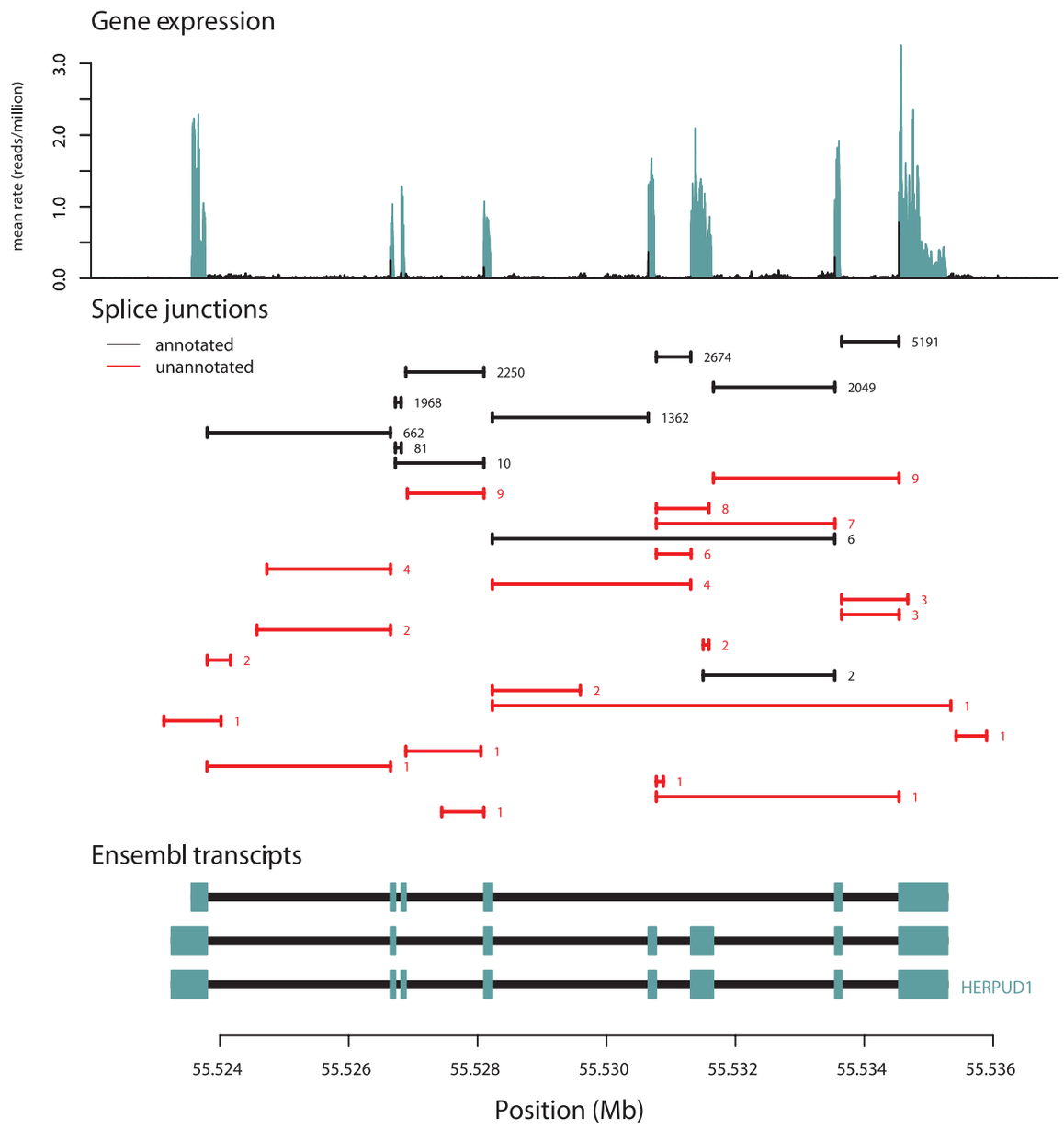
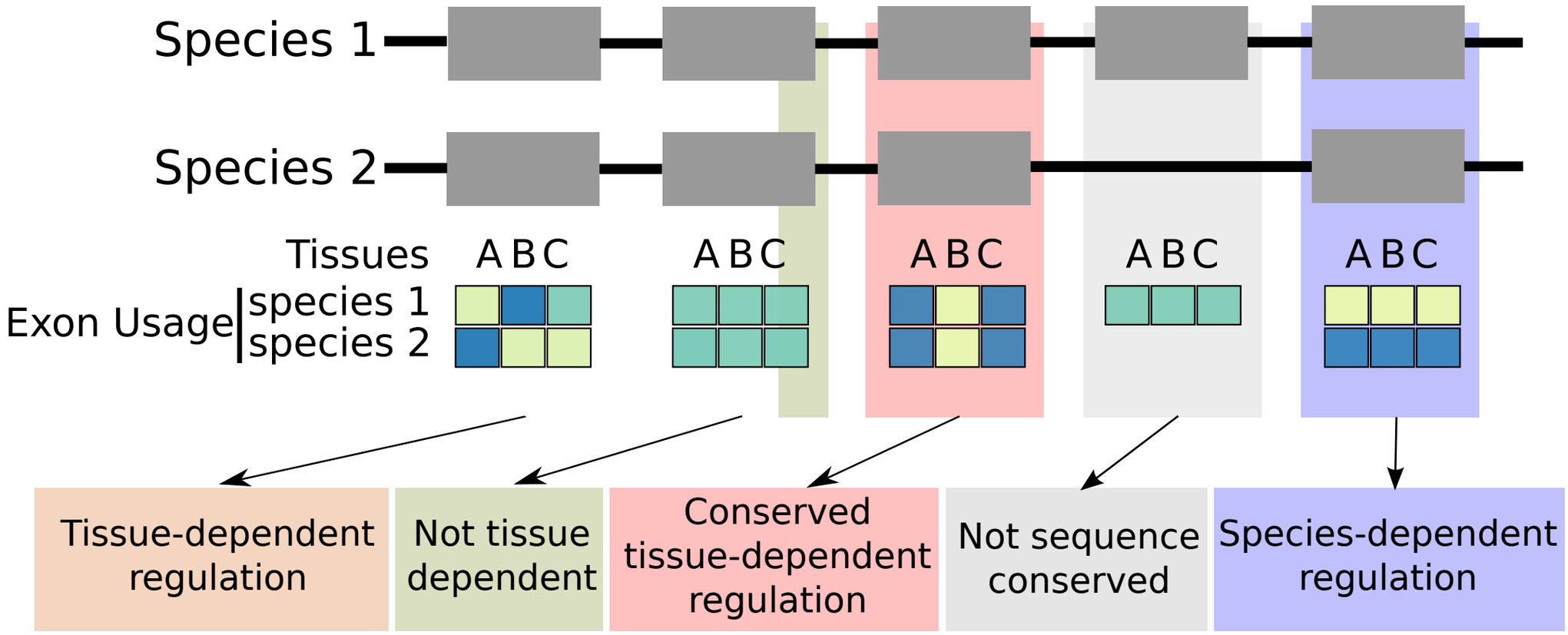


Figure 2. An example of splice junctions identified in a gene. In the top panel, we plot the average expression level at each base in a regior surrounding *HERPUD1*. In blue are bases annotated as exonic, and in black are those annotated as not exonic. In the middle panel, we plot the positions of all splice junctions in the region identified in our data. In black are splice junctions that are present in gene databases; in red are those that are not. The number of sequencing reads supporting each junction is written to the right of each junction, and junctions are ordered from top to bottom of the plot according to their coverage. In the bottom panel, we show the gene models in the region from Ensembl. The blue boxes show the positions of exons, and the black lines the positions of introns.
doi:10.1371/journal.pgen.1001236.g002

Regulation of (alternative) exon usage

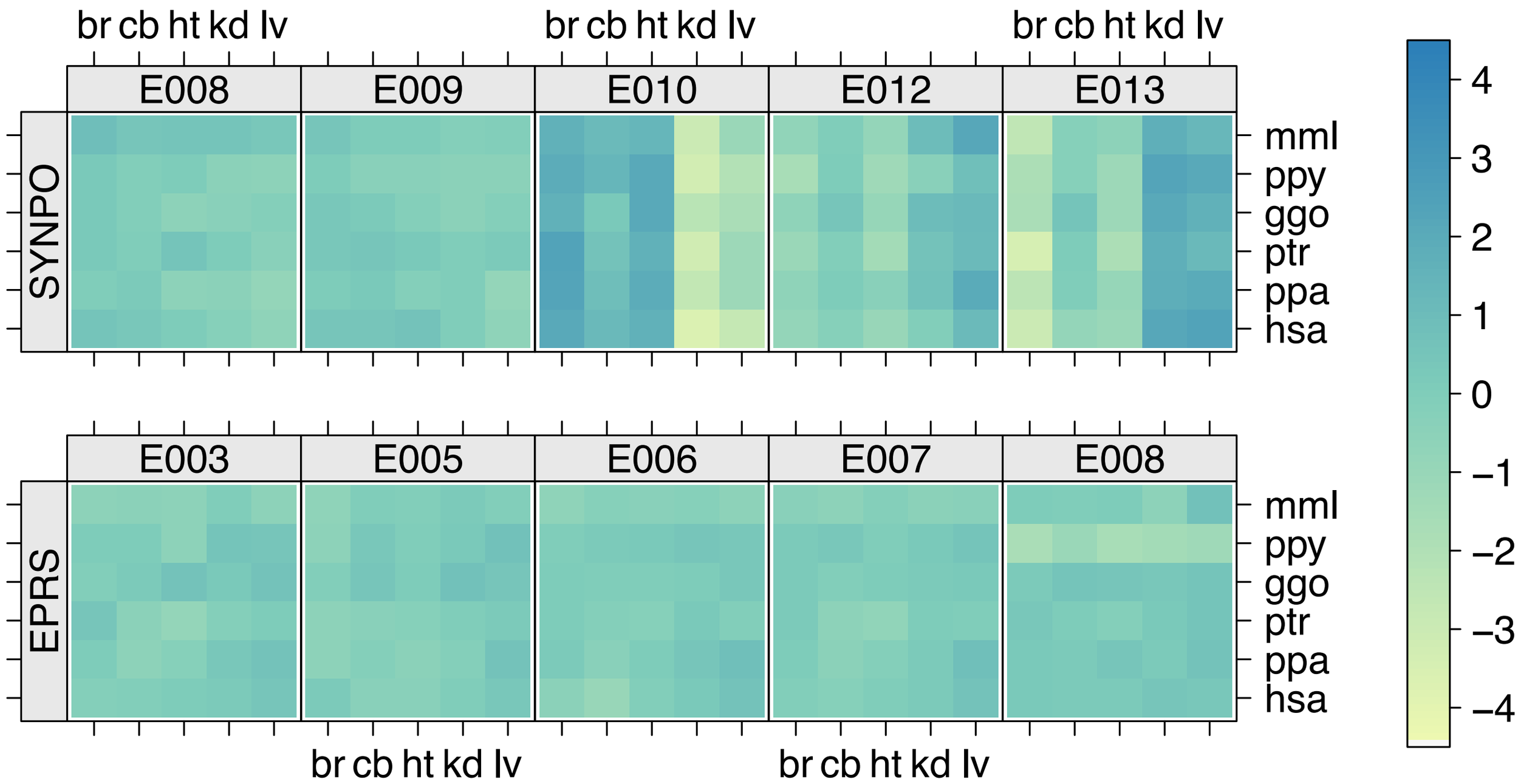


Data: multiple replicate samples each from:

- **6 primate species (hsa, ppa, ptr, ggo, ppy, mml) X**
- **5 tissues (heart, kidney, liver, brain, cerebellum)**

Brawand et al. Nature 2011 (Kaessmann Lab, Lausanne, CH)

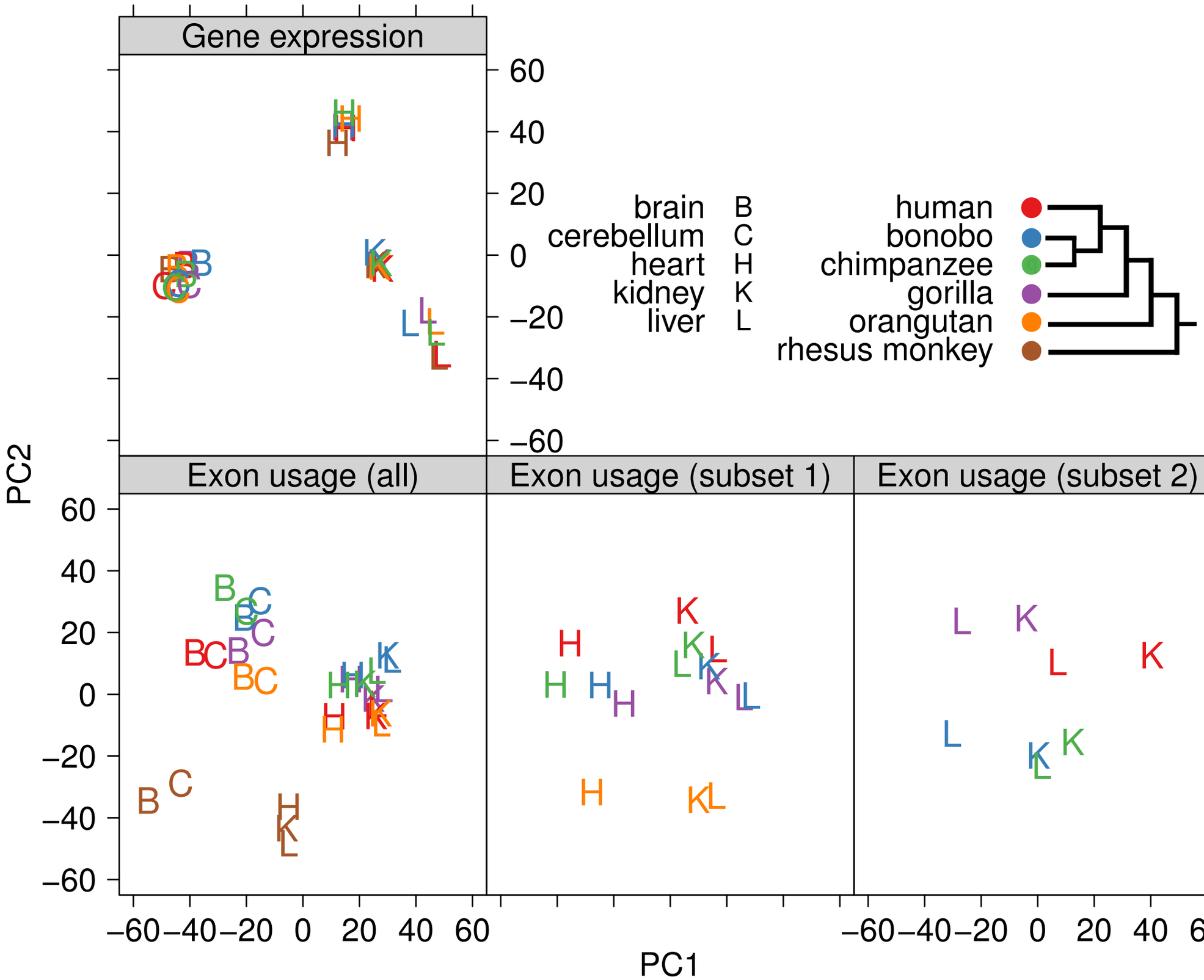
Tissue and species dependence of relative exon usage



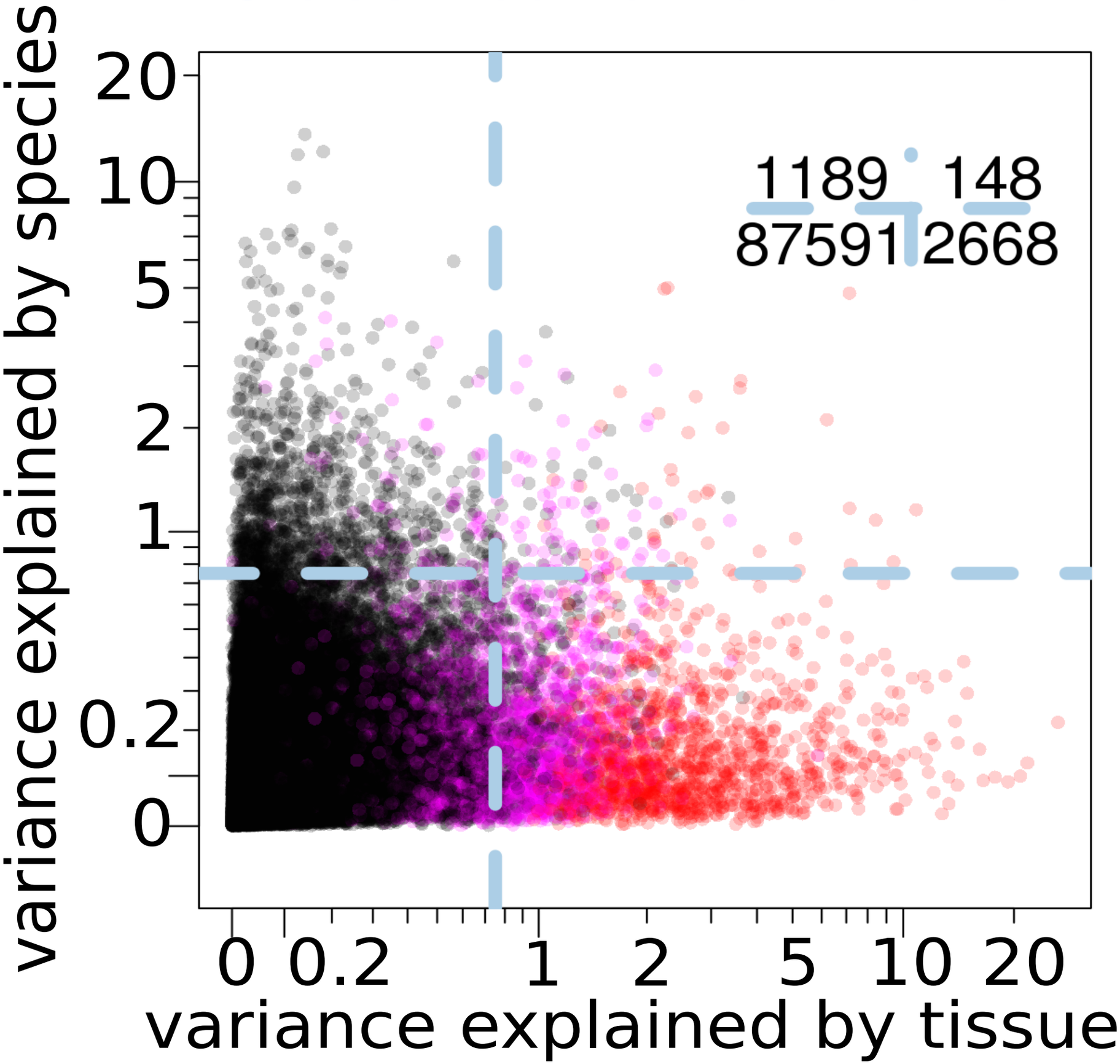
Drift and conservation of differential exon usage across tissues in primate species

Alejandro Reyes[a,1], Simon Anders[a,1], Robert J. Weatheritt[b,2], Toby J. Gibson[b], Lars M. Steinmetz[a,c], and Wolfgang Huber[a,3]

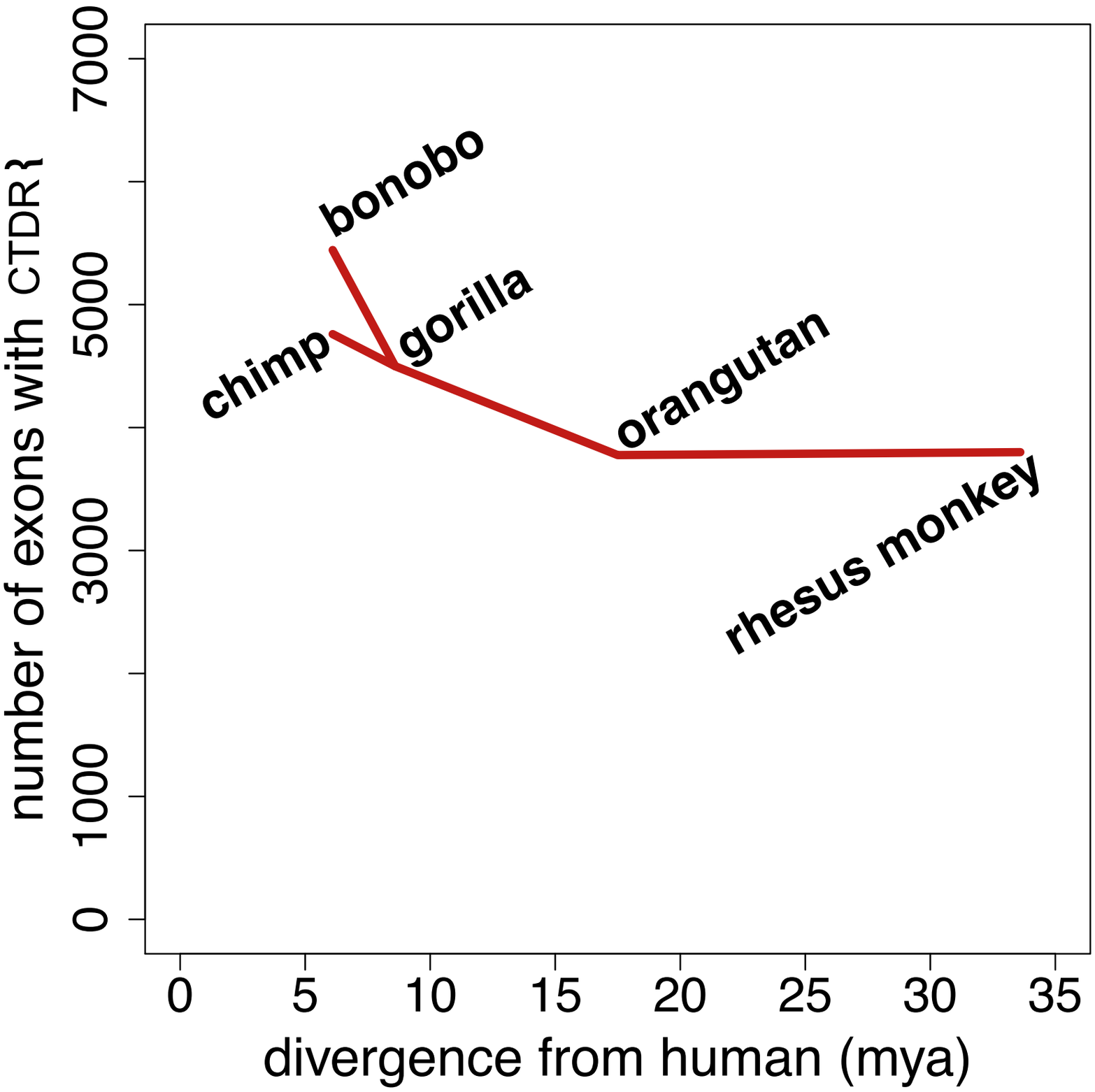
PNAS 2013


Gene expression
Exon usage (all)
Exon usage (subset 1)
Exon usage (subset 2)
PC1
PC2
brain B
cerebellum C
heart H
kidney K
liver L
human
bonobo
chimpanzee
gorilla
orangutan
rhesus monkey

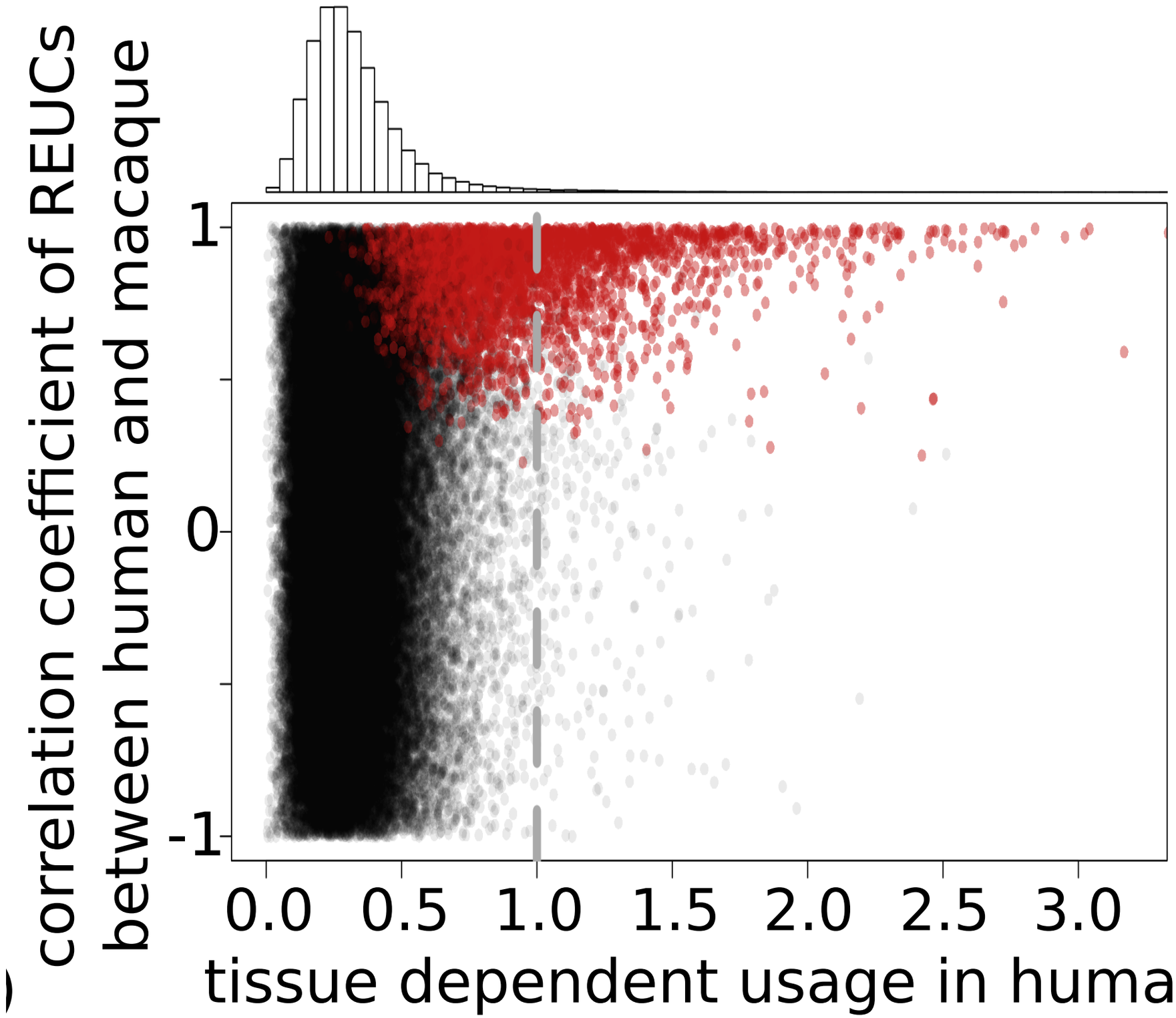
Classification of exons



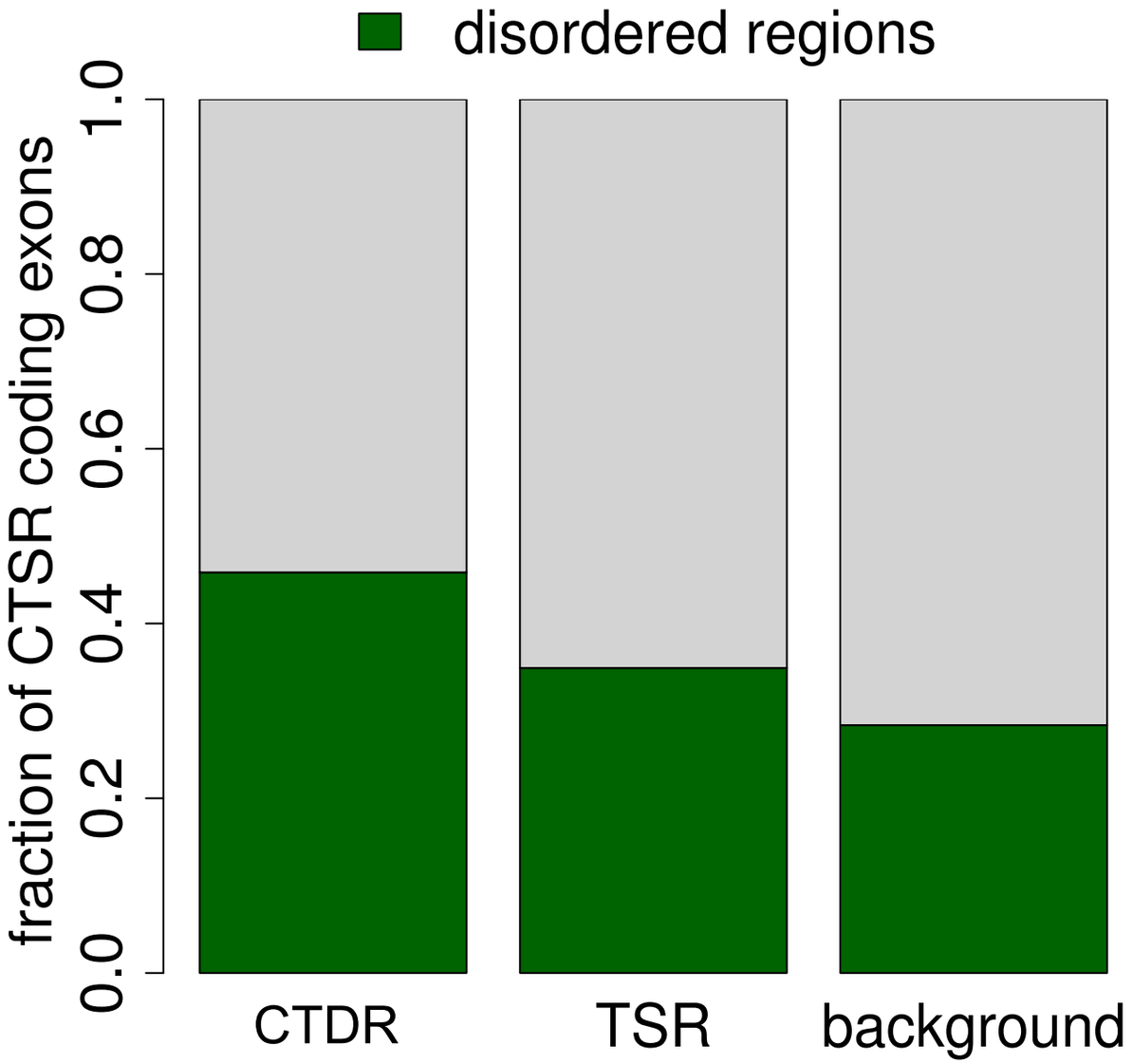
Conservation: a core set of tissue-dependent exons across primates

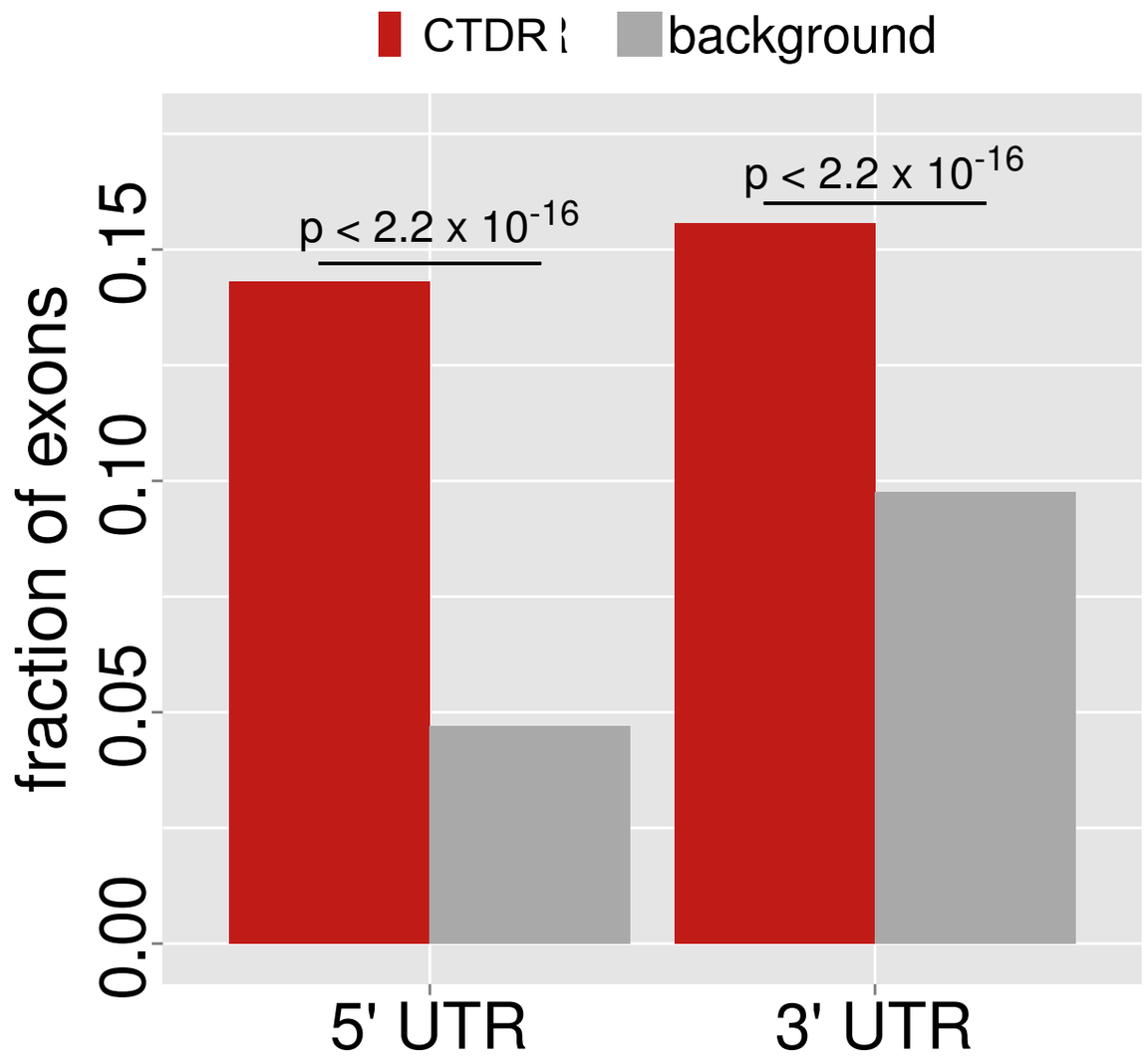


Strong patterns of tissue-dependent exon usage are frequently conserved

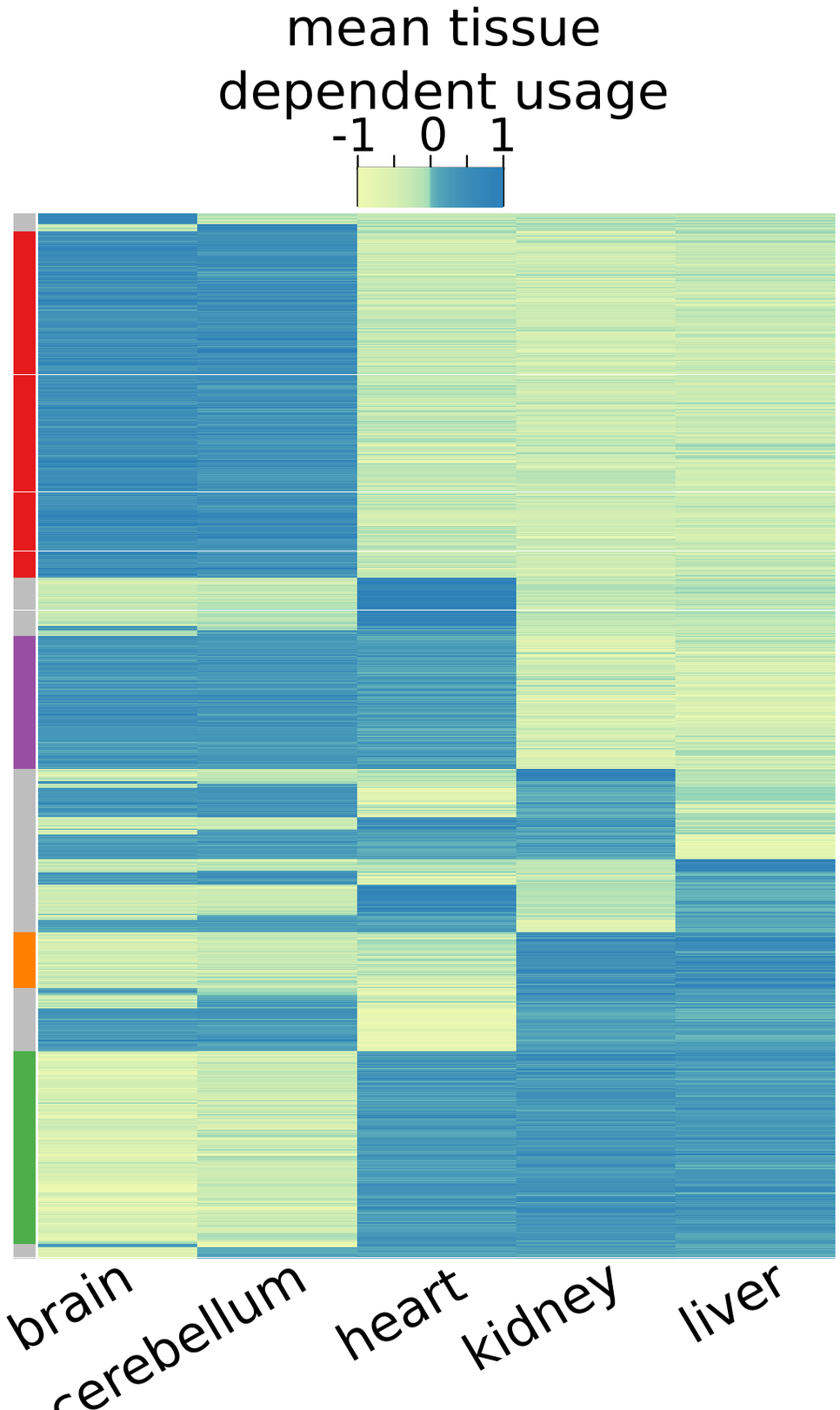


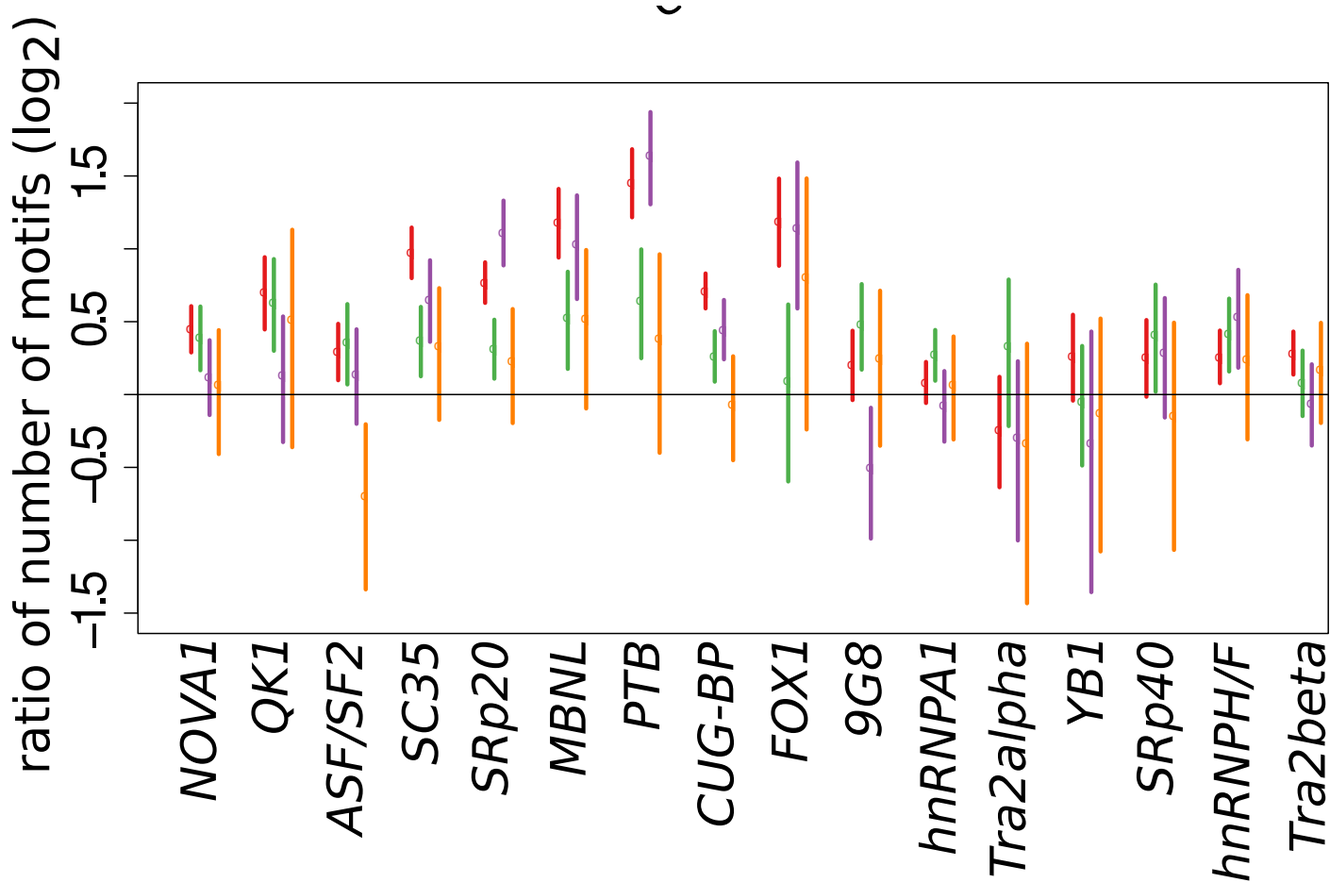
Functional associations of conserved tissue-dependent exons





Tissue-dependent usage patterns are associated with splicing factor binding motifs and suggest a cis-regulatory code





Summary tissue-dependent exon usage

Detection of tissue-dependent regulation and its conservation across species at unprecedented scale and precision.

Most of tissue-dependent alternative exon usage in primates is

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

However, a significant fraction is

- **high amplitude**
- **conserved**
- **associated with function in mRNA life-cycle & localisation, translation regulation, protein interaction & function**

Summary differential expression

- **Text-book statistical concepts are (almost) sufficient for differential expression: ANOVA, hypothesis testing, generalized linear models**
- **In addition: small-n large-p - information sharing across genes, empirical Bayes, shrinkage**
- **In practice, visualisation (“drill down”) and quality control (batch effects) are very important**
- **Exon-level analysis**

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