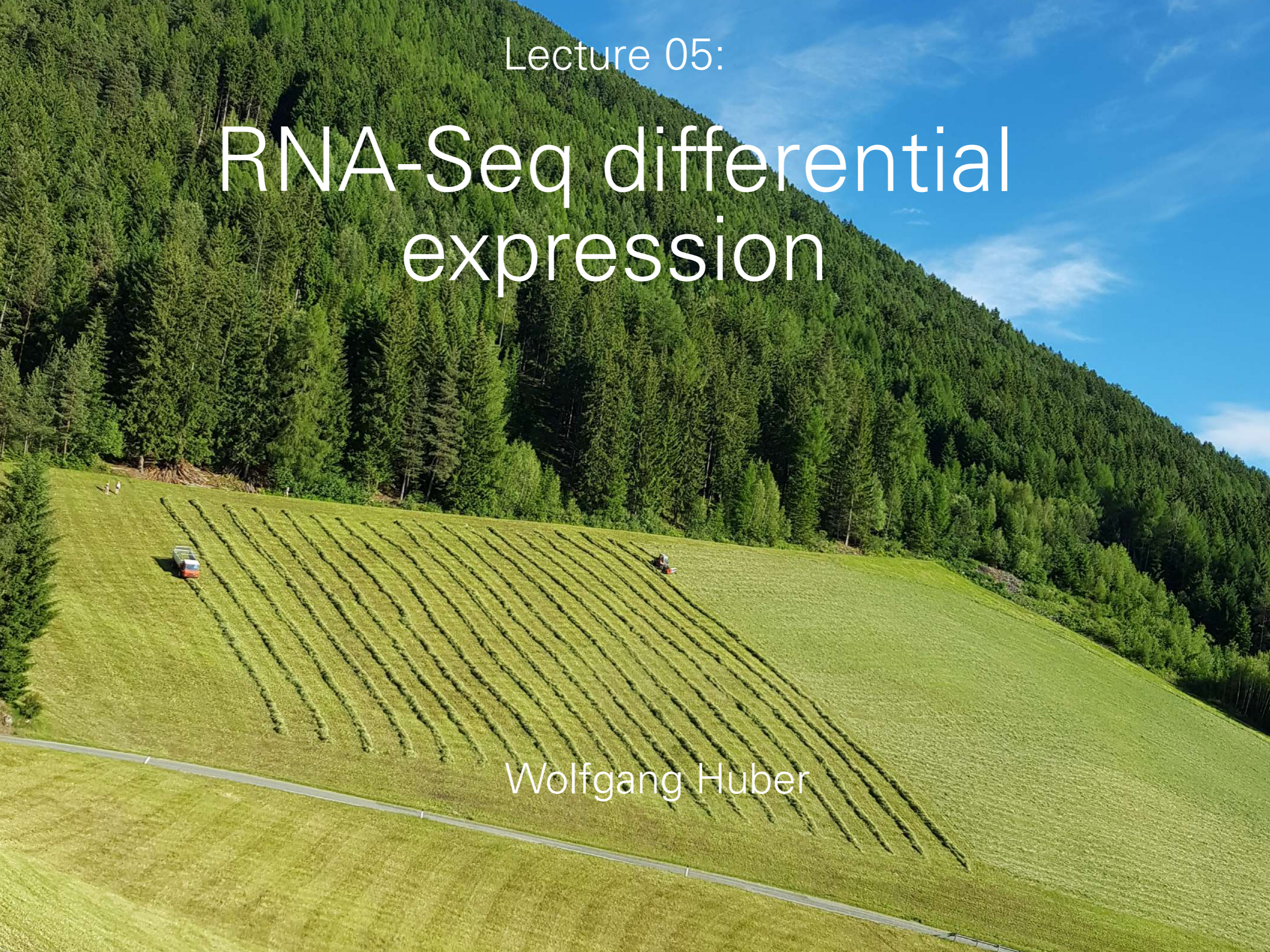


Lecture 05:

RNA-Seq differential expression

Wolfgang Huber



Aims for this lecture

Understand the peculiarities of count-data from high-throughput sequencing and the Gamma-Poisson model

Some basic concepts for generalized linear models (cf. Levi's lecture on Wed for more)

Understand shrinkage estimation and its application to modelling experiments using HT assays

Some further bells and whistles of DESeq2: transformation, outlier robustness, banded testing

Testing differential exon abundance (\rightarrow isoform usage)

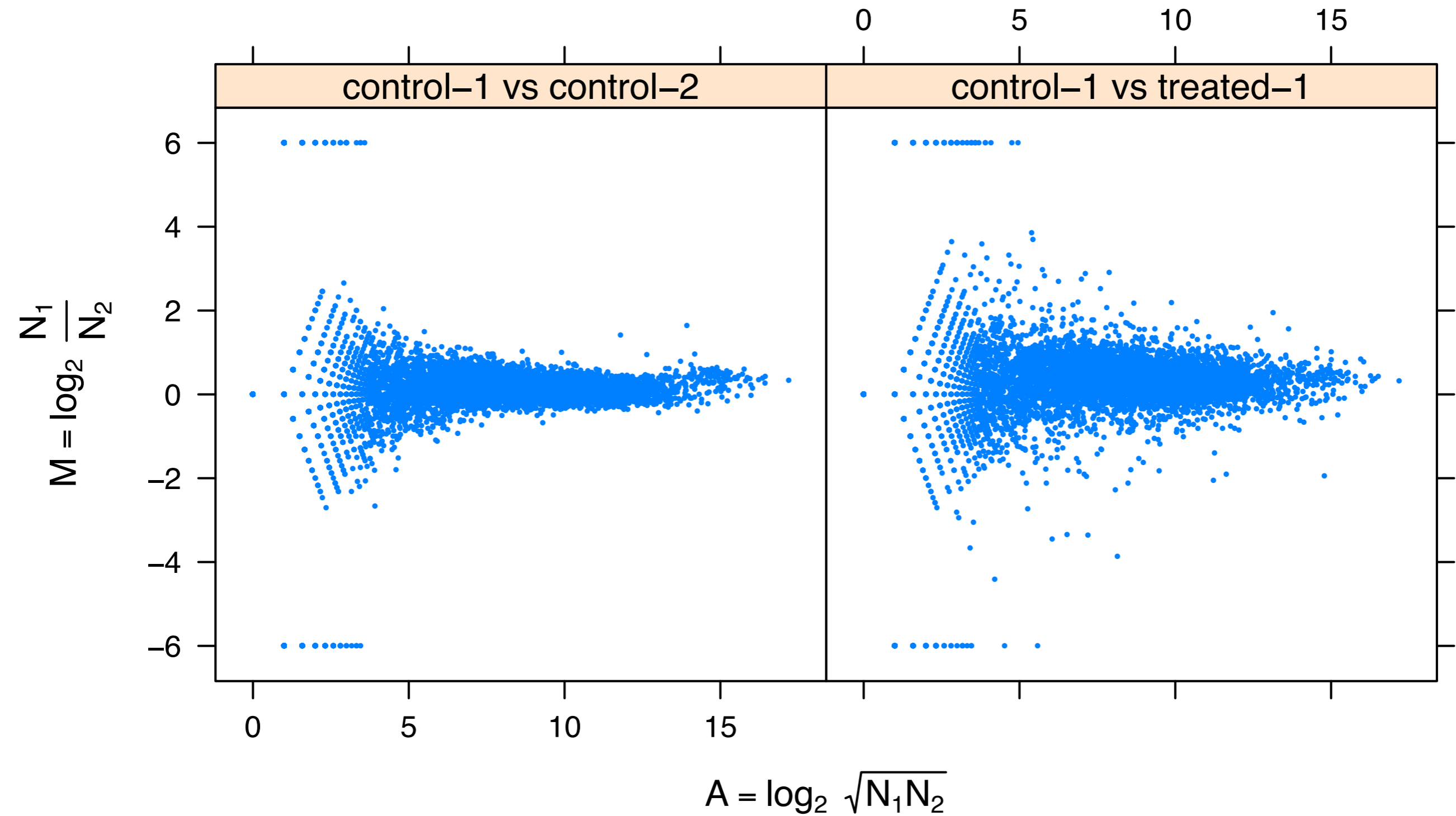
Count data in HTS

Gene	G1iNS1	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
- ...

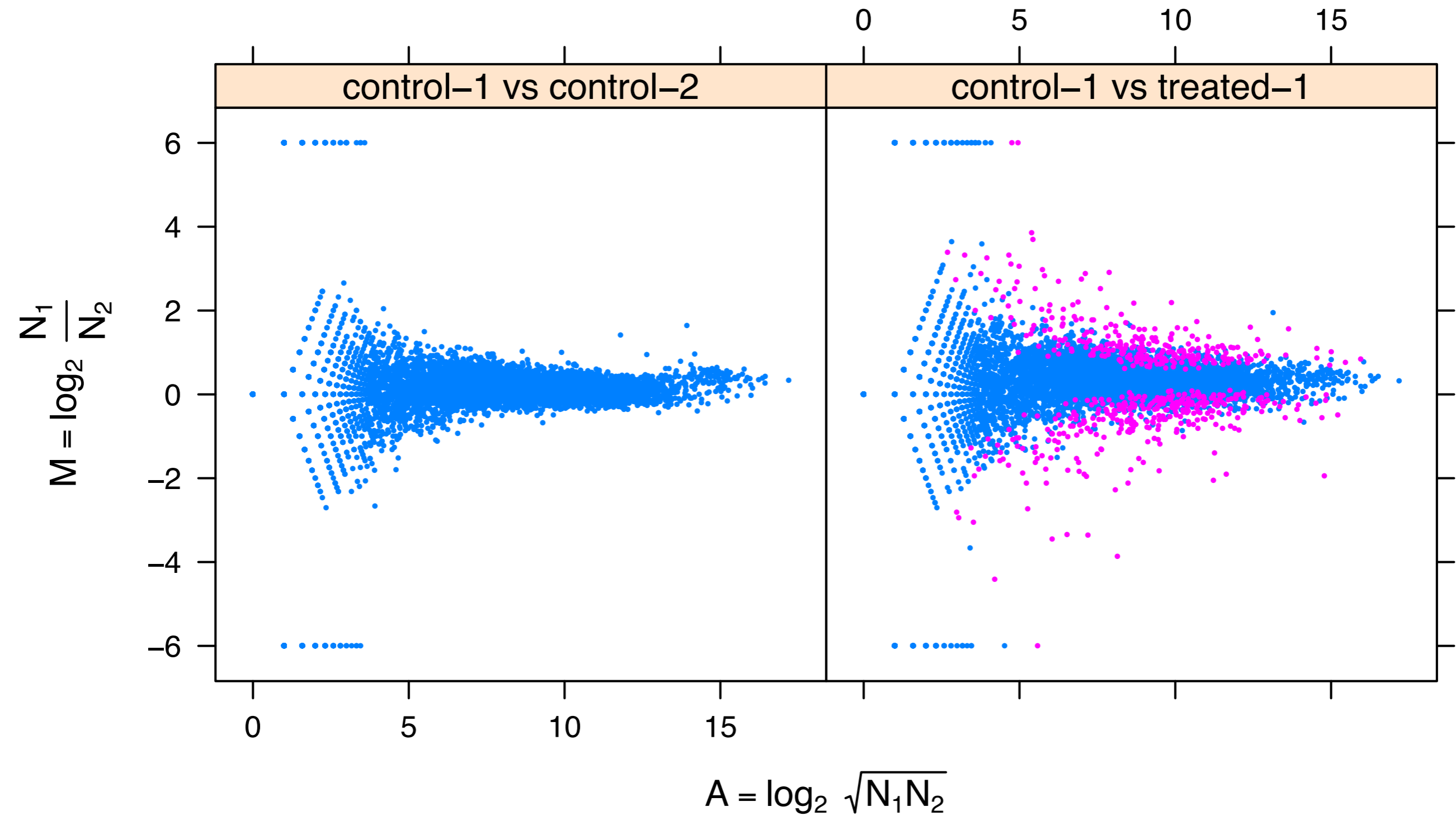
Simon Anders





two biological
replicates

treatment vs control



two biological
replicates

treatment vs control

Challenges

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

➔ heteroskedasticity matters

Data are discrete, positive, skewed

Small numbers of replicates (...)

➔ no (log-)normal model

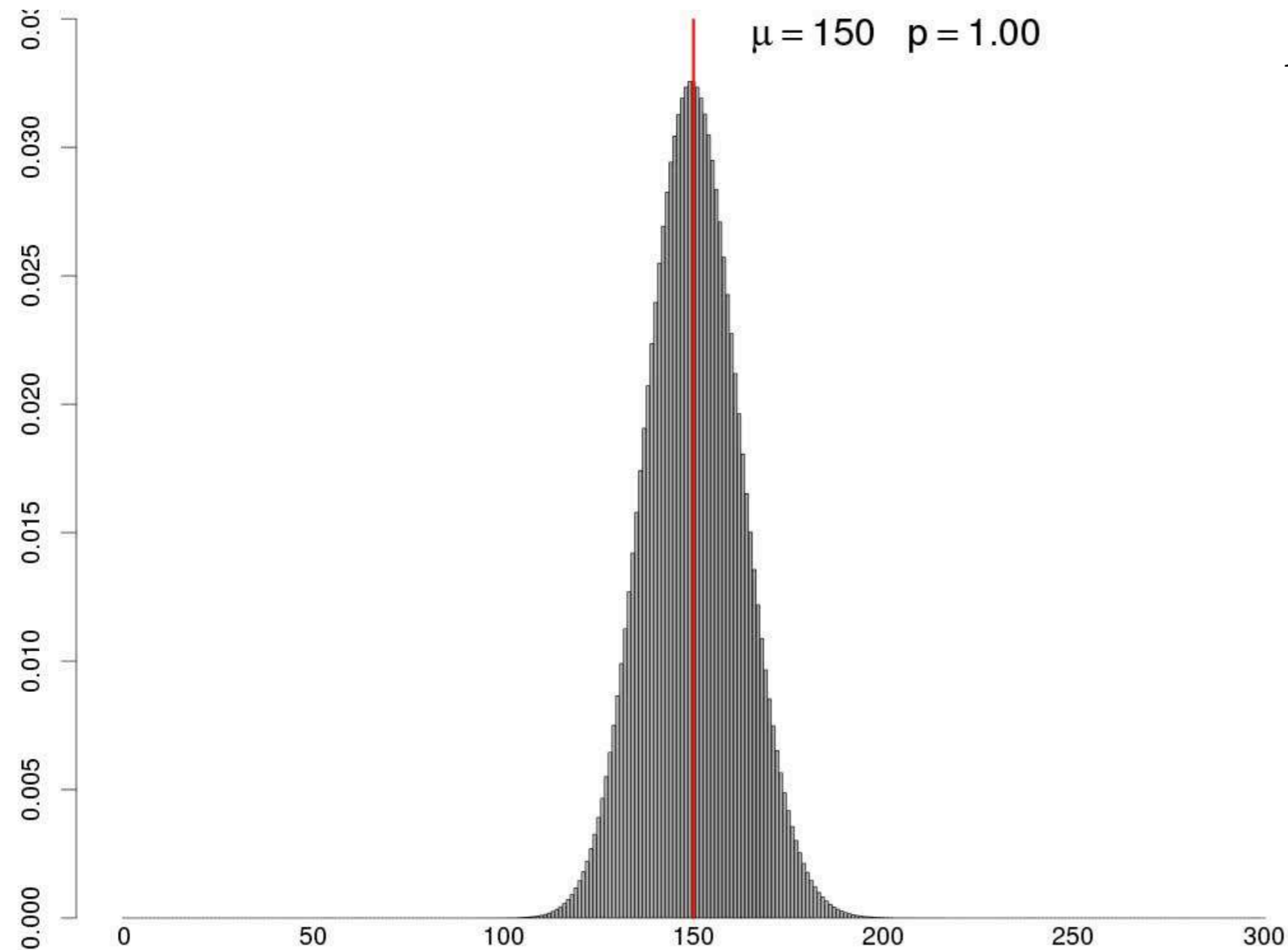
➔ no rank based or permutation methods

➔ use parametric stochastic model to infer tail behaviour (approximately) from low-order moments

➔ power sharing between genes ('large-p small n')

The Gamma-Poisson (a.k.a. Negative Binomial) distribution

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



Alternative parameterisation

$$\alpha = \frac{1}{r}$$
$$\mu = \frac{pr}{1 - p}$$

Moments

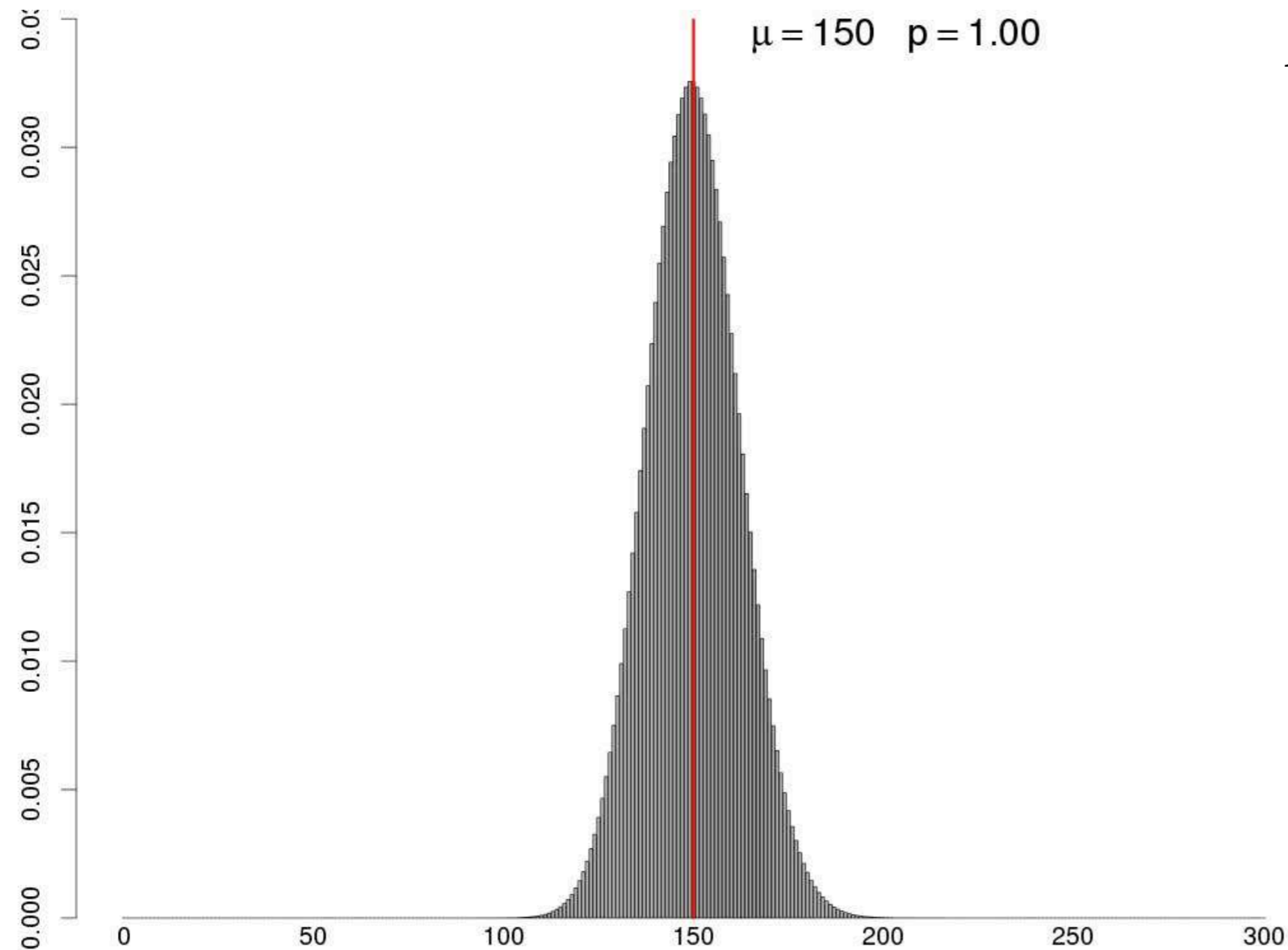
$$\text{mean} = \mu$$

$$\text{variance} = \mu + \alpha\mu^2$$

Bioconductor package
DESeq, since 2010

The Gamma-Poisson (a.k.a. Negative Binomial) distribution

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



Alternative parameterisation

$$\alpha = \frac{1}{r}$$
$$\mu = \frac{pr}{1 - p}$$

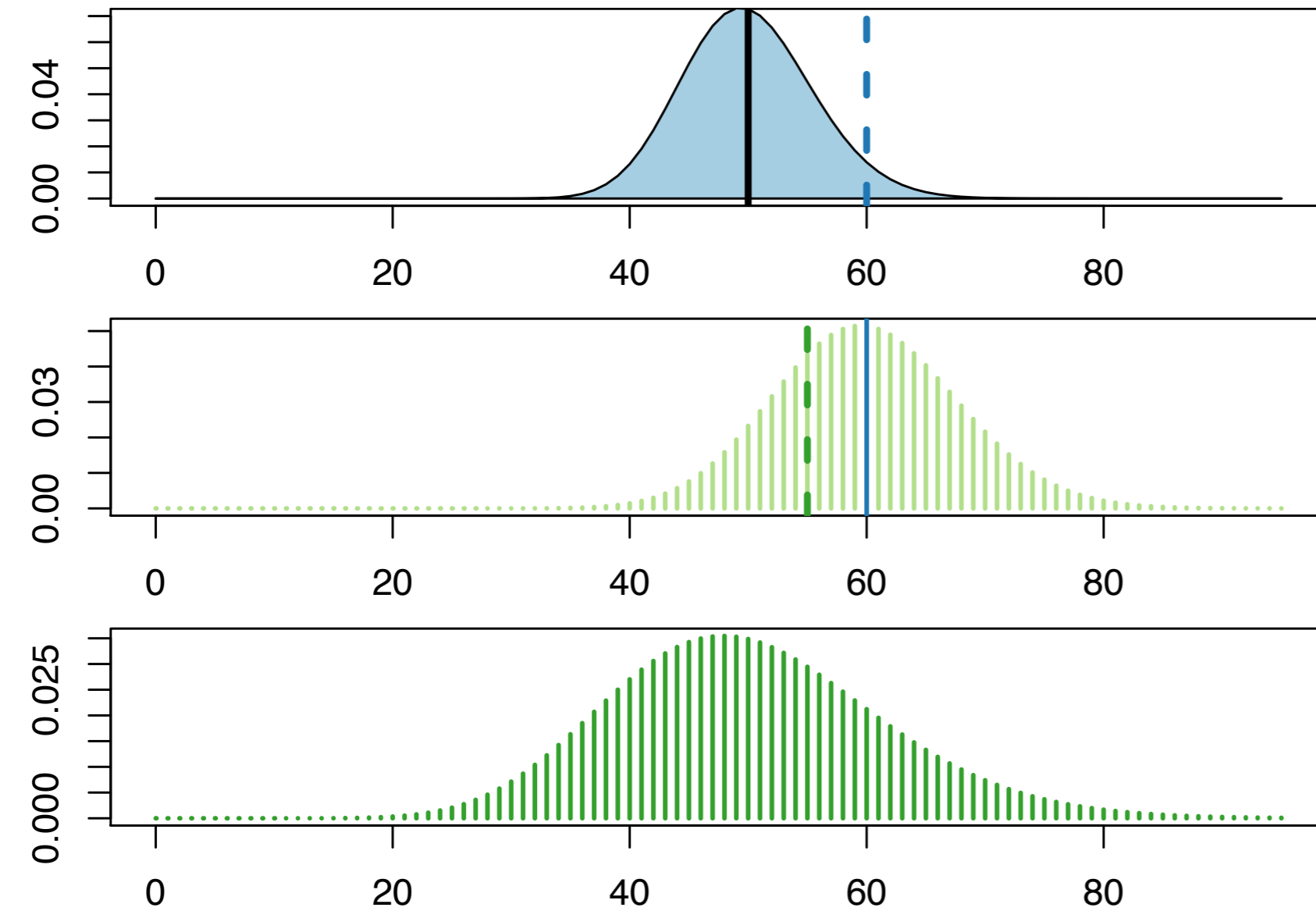
Moments

$$\text{mean} = \mu$$

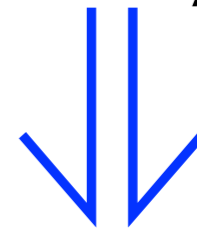
$$\text{variance} = \mu + \alpha\mu^2$$

Bioconductor package
DESeq, since 2010

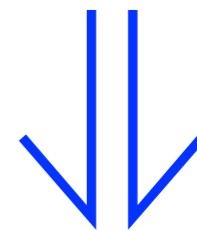
The Gamma-Poisson distribution models a Poisson process whose mean is itself randomly varying



Biological sample to sample
variability Γ



Poisson counting statistics Λ



Overall distribution GP

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

Two component noise model

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

shot noise (Poisson) biological noise

Small counts

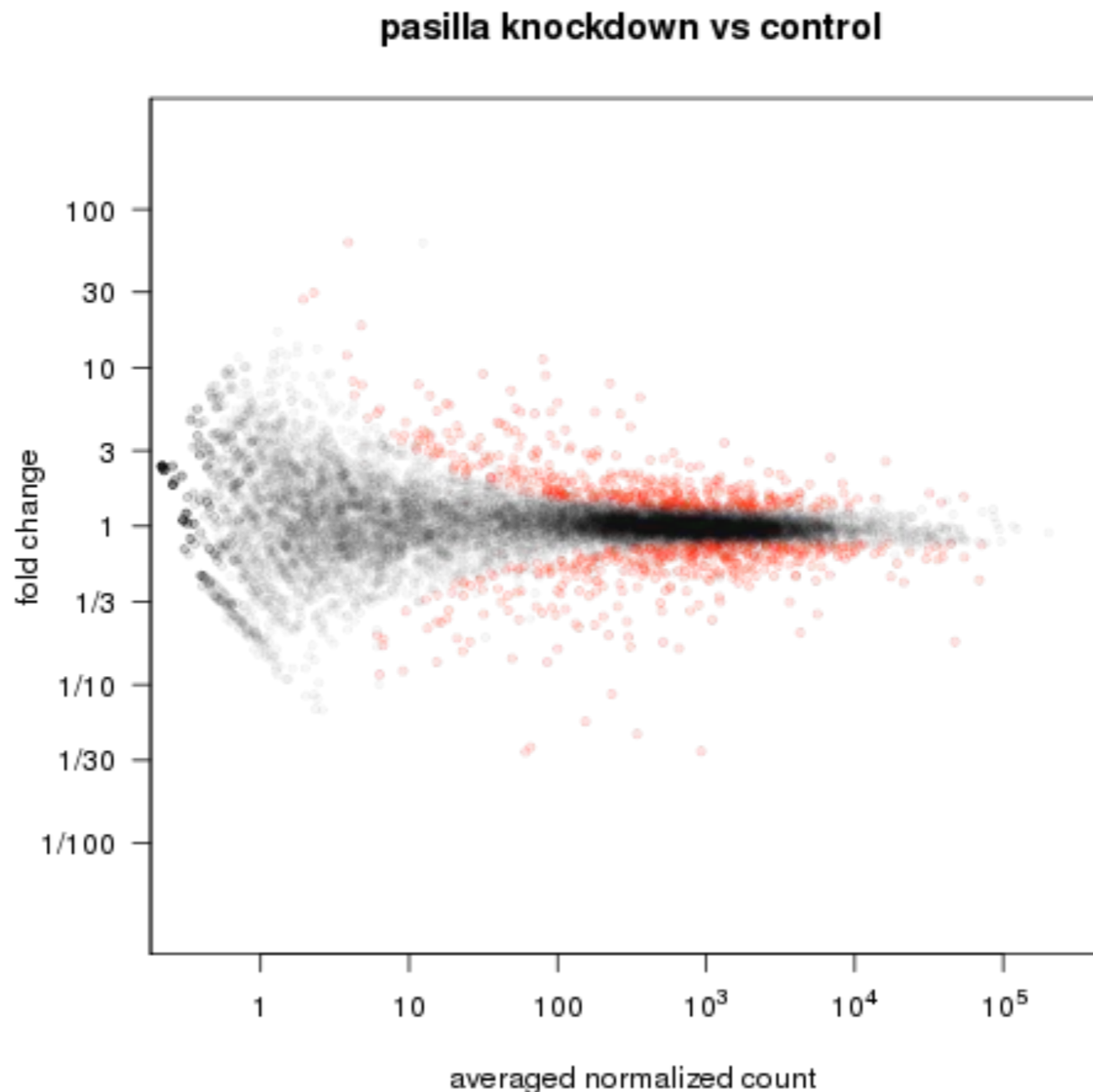
Sampling noise
dominant

Improve power:
deeper coverage

Large counts

Biological noise
dominant

Improve power:
more biol.
replicates



Generalised linear model of the Gamma-Poisson (or NB) family

$$N_{ij} \sim \text{NB}(\mu_{ij}, \alpha_{ij}) \quad \text{Noise part}$$

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

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

Generalised linear model of the Gamma-Poisson (or NB) family

$$N_{ij} \sim \text{NB}(\mu_{ij}, \alpha_{ij}) \quad \text{Noise part}$$

$$\log \mu_{ij} = \text{Important special case: two groups} \quad \text{mantic}$$

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?

$$Y \sim D(\mu, \sigma)$$

A GLM consists of three elements:

1. A linear predictor $\eta = X \beta$
2. A non-linear transformation (link function) g such that $g(\mu) = \eta$
3. A probability distribution D (from the exponential family),
with mean μ and scale σ

Ordinary linear model: $g = \text{identity}$, $D = \text{normal}$

DESeq(2), edgeR, ...: $g = \log$, $D = \text{Gamma-Poisson}$

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 :

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

reduced model for gene i :

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

GLM: Interactions

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

full model for gene i :

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

reduced model for gene i :

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

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:

$$\log \mu_{ijl} = \beta_i^0 + \begin{cases} 0 & \text{for } l = 1(\text{healthy}) \\ \beta_i^T & \text{for } l = 2(\text{tumour}) \end{cases}$$

reduced model:

$$\log \mu_{ij} = \beta_i^0$$

i gene
 j subject
 l tissue state

Recap: designs for 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

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:

$\text{count} \sim \text{assayType} + \text{treatment} + \text{assayType} : \text{treatment}$

reduced model:

$\text{count} \sim \text{assayType} + \text{treatment}$

See <https://support.bioconductor.org/p/61509/> ("DESeq2 testing ratio of ratios")

For an application example: e.g., Zarnack et al., Cell 2013

Benefitting from the many variables ('big data')

To assess signal and noise in the data from one gene, we have

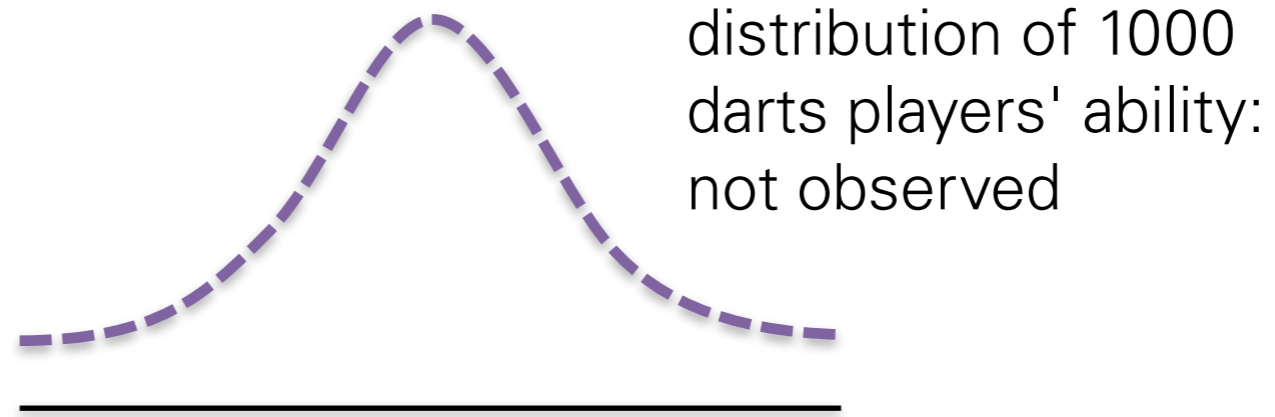
- the data for that gene
- that of all the other genes
- user-defined parameters (e.g. cutoffs)

⇒ regularisation, (empirical) Bayes

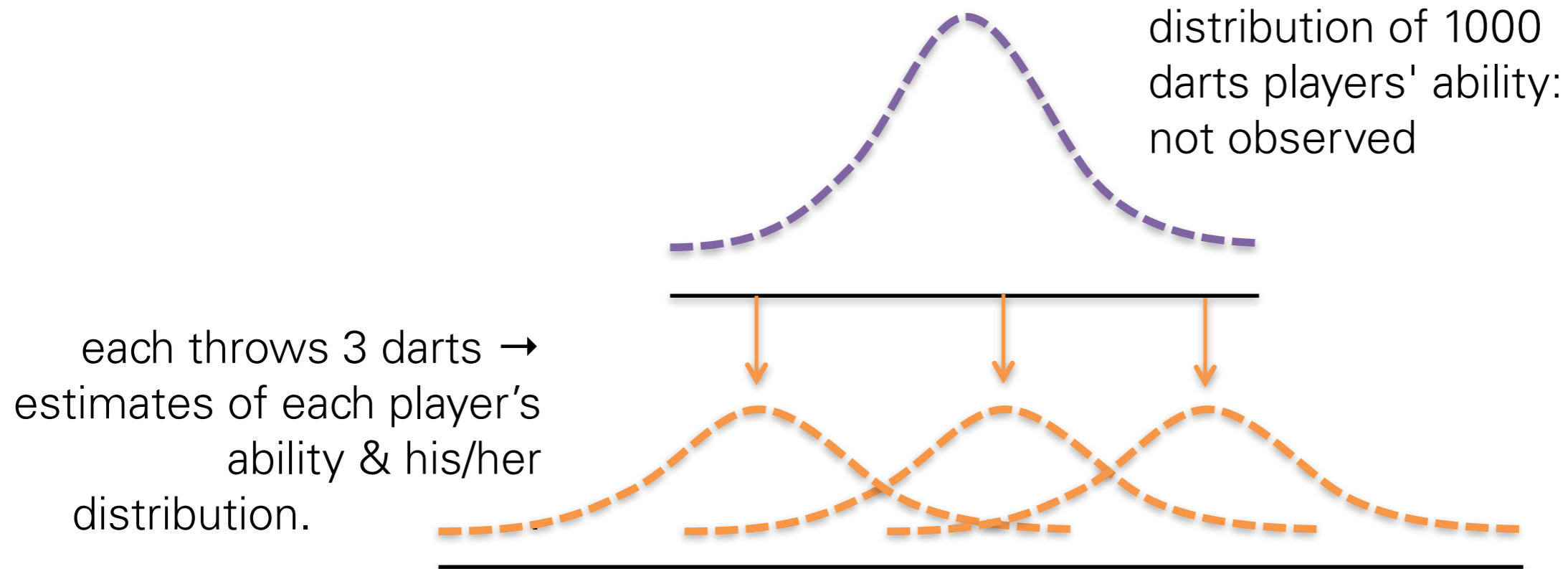


Michael Love

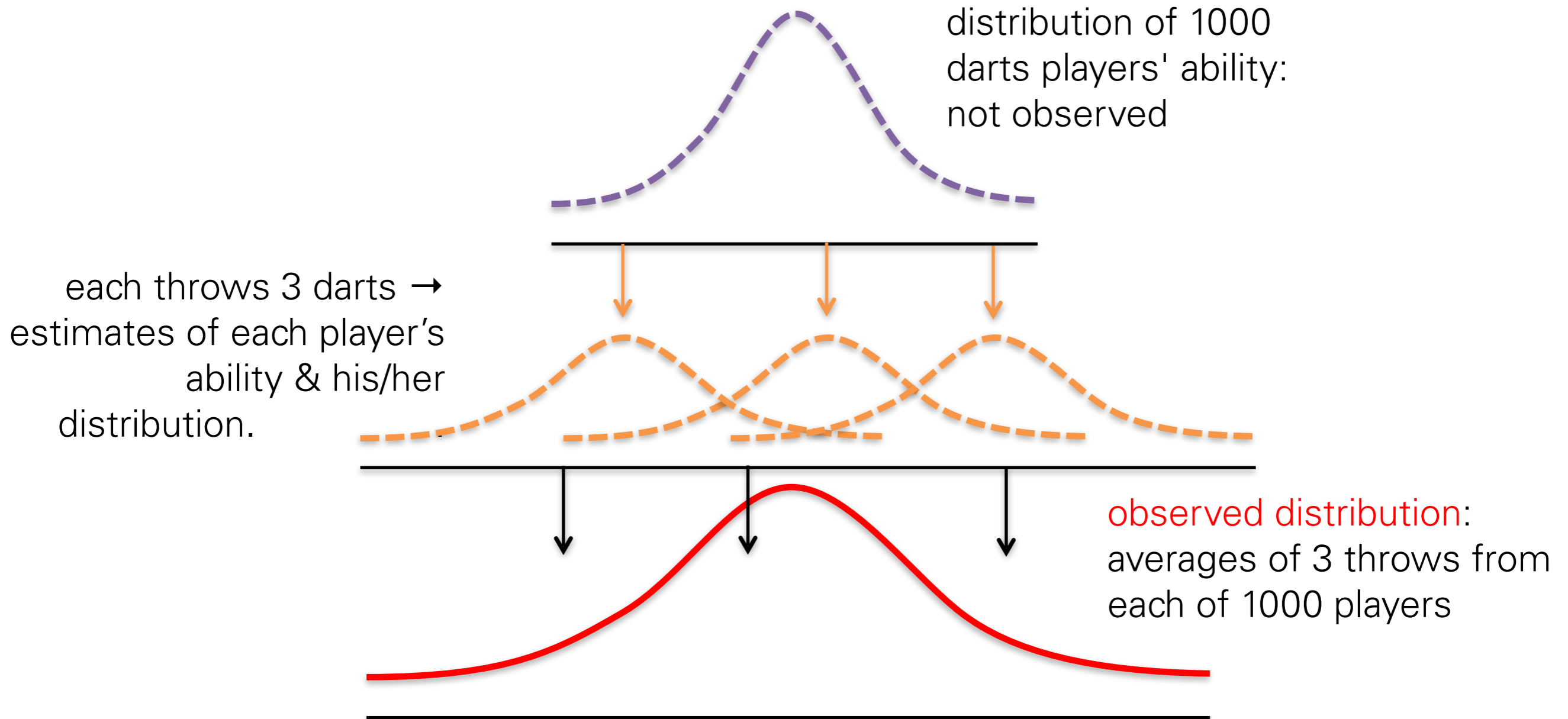
Theoretical Interlude: Shrinkage estimation



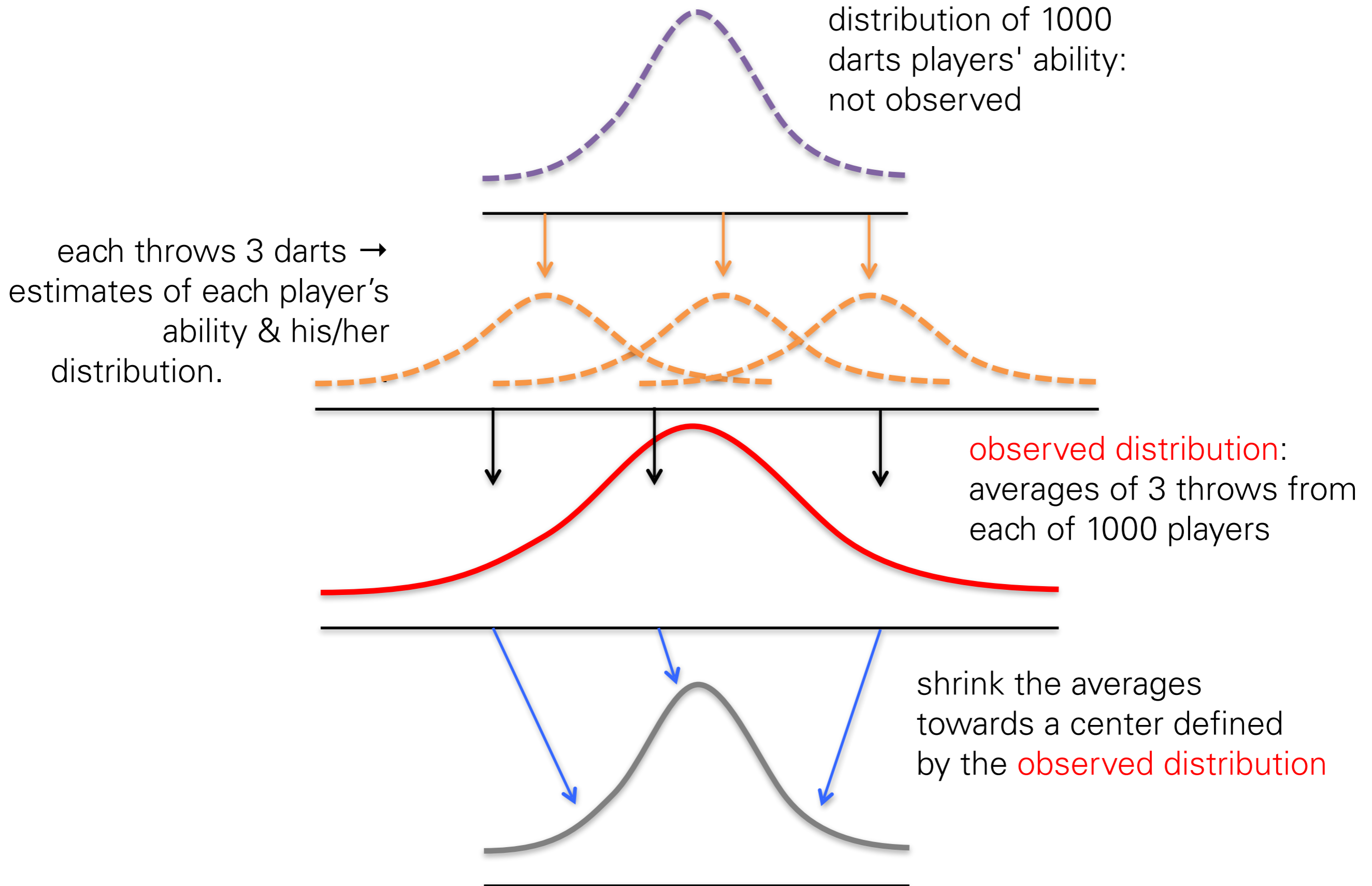
Theoretical Interlude: Shrinkage estimation



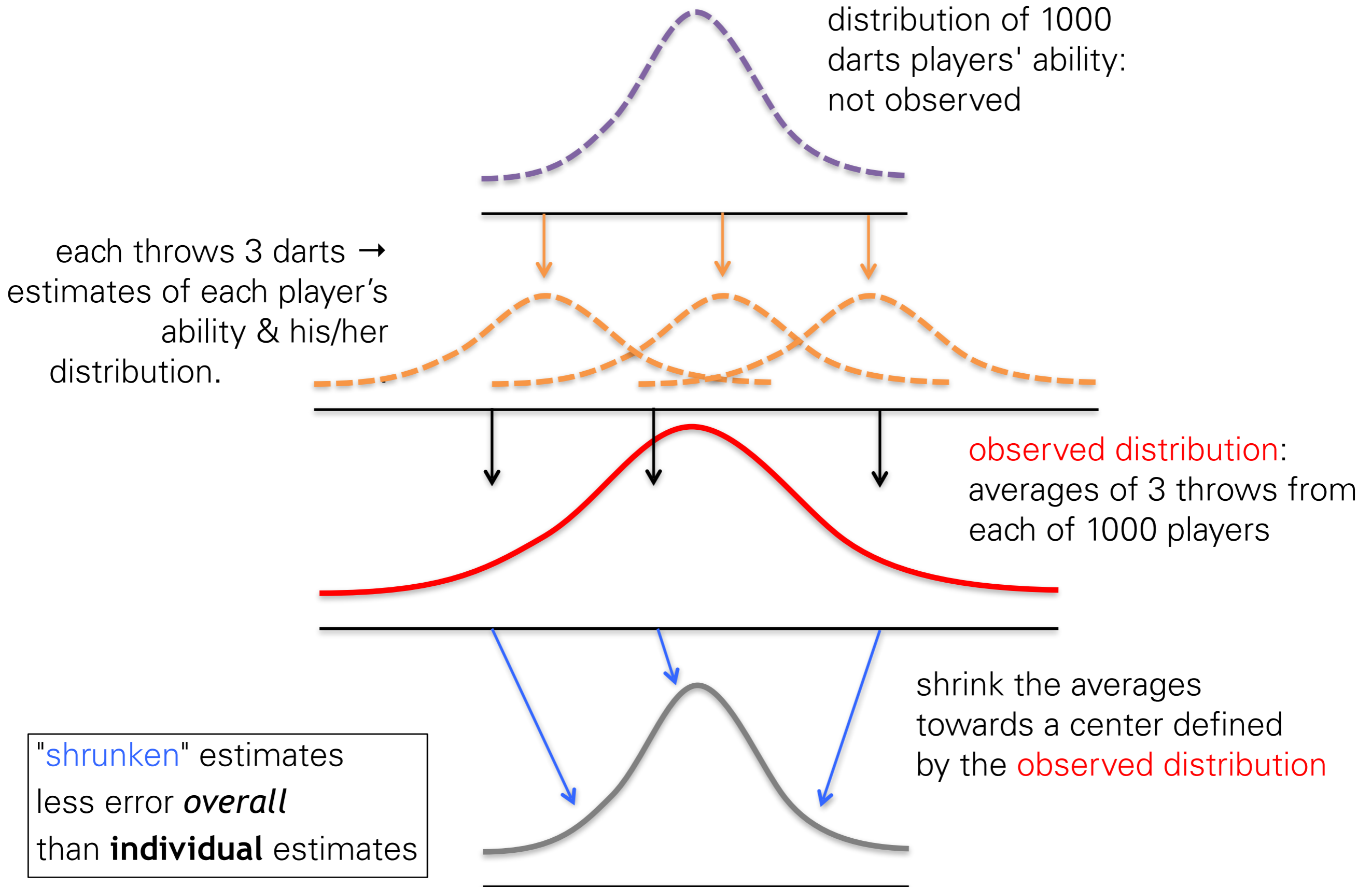
Theoretical Interlude: Shrinkage estimation



Theoretical Interlude: Shrinkage estimation



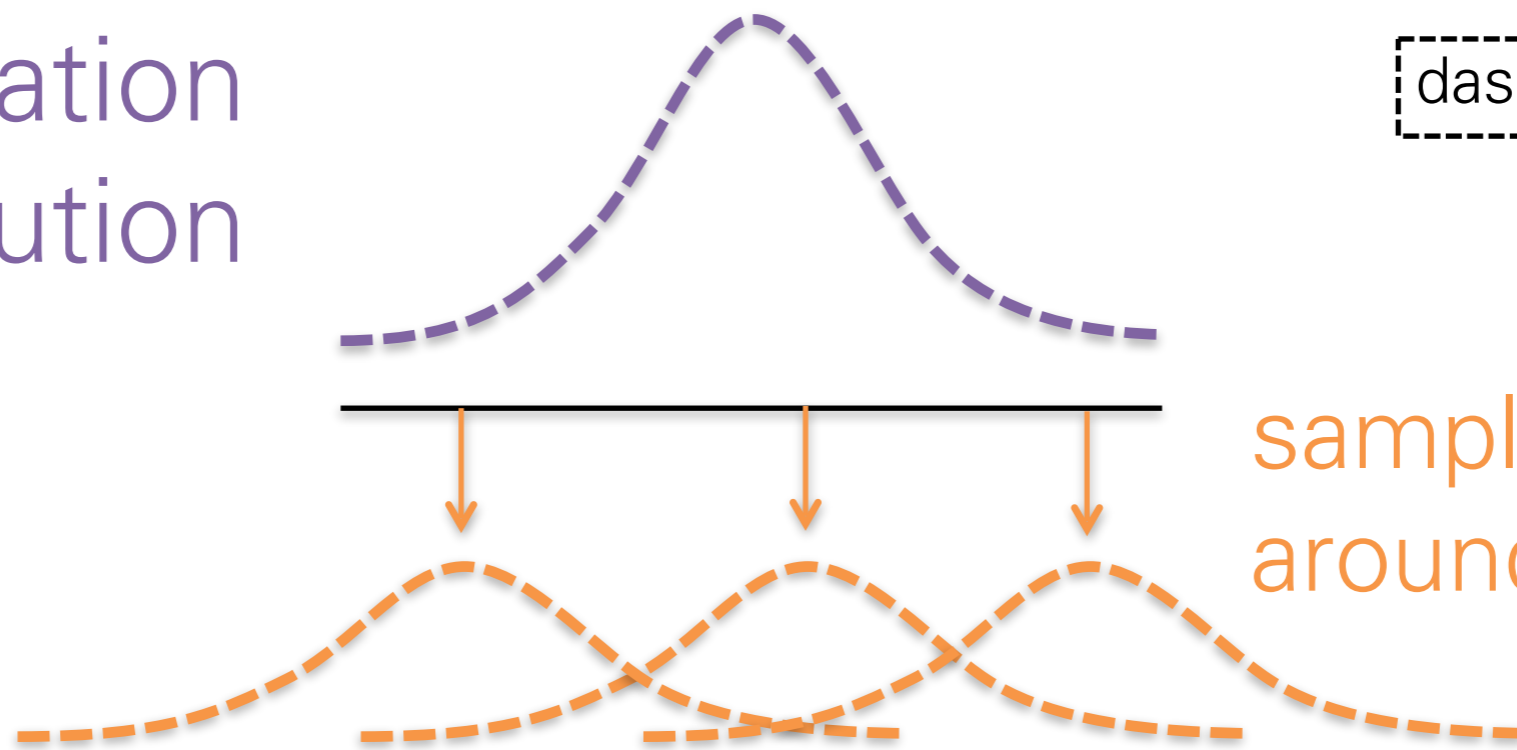
Theoretical Interlude: Shrinkage estimation



Shrinkage estimation

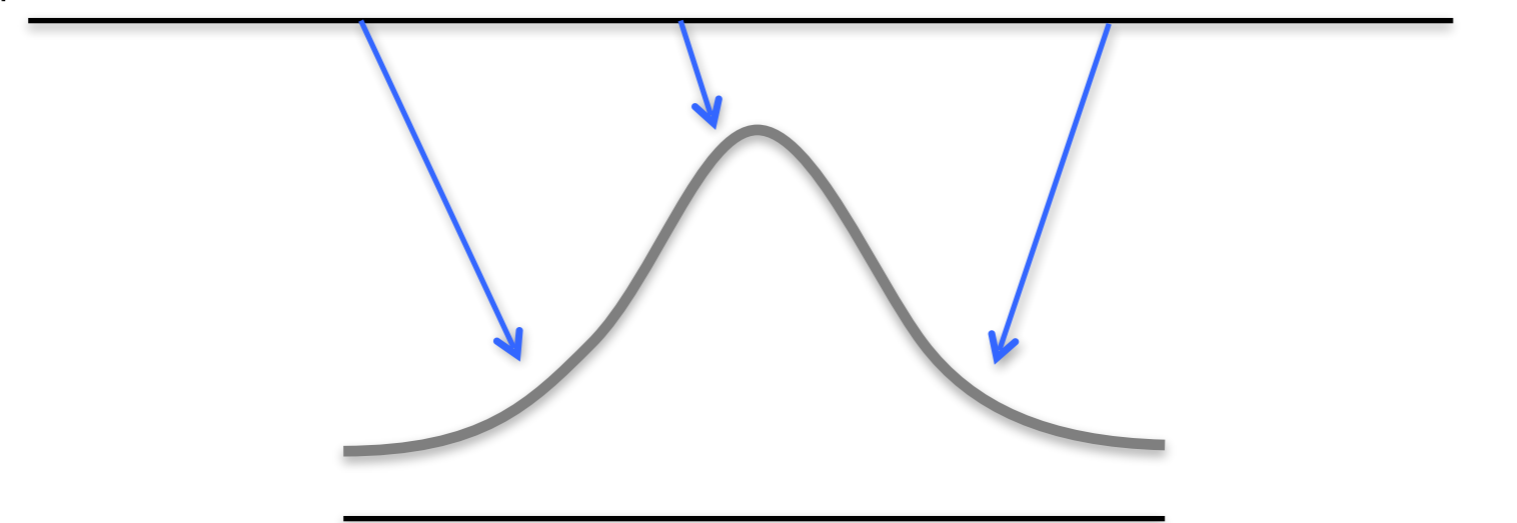
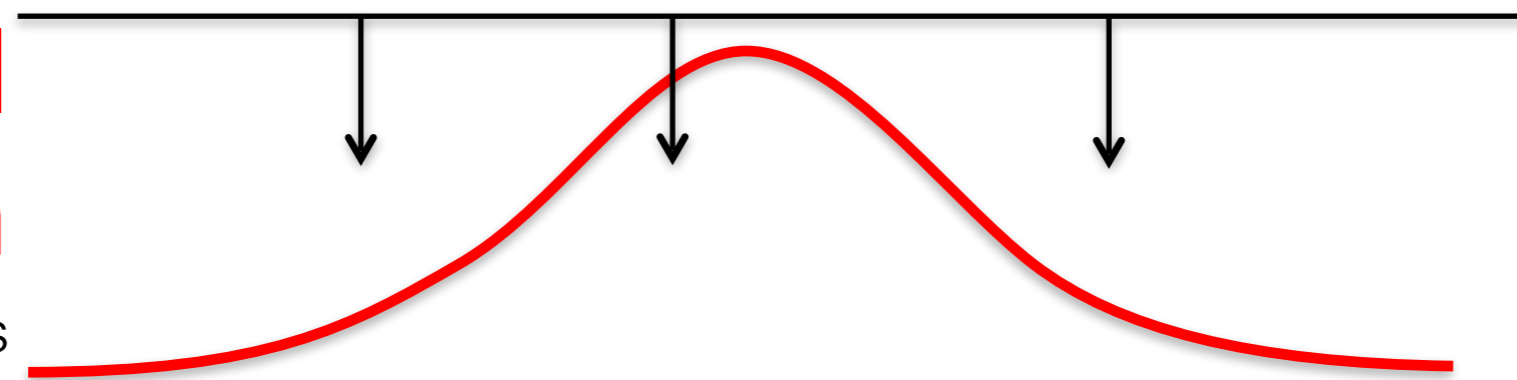
population distribution

dashed = unobserved



empirical distribution

the center defines the prior mean



Shrinkage estimation

population distribution

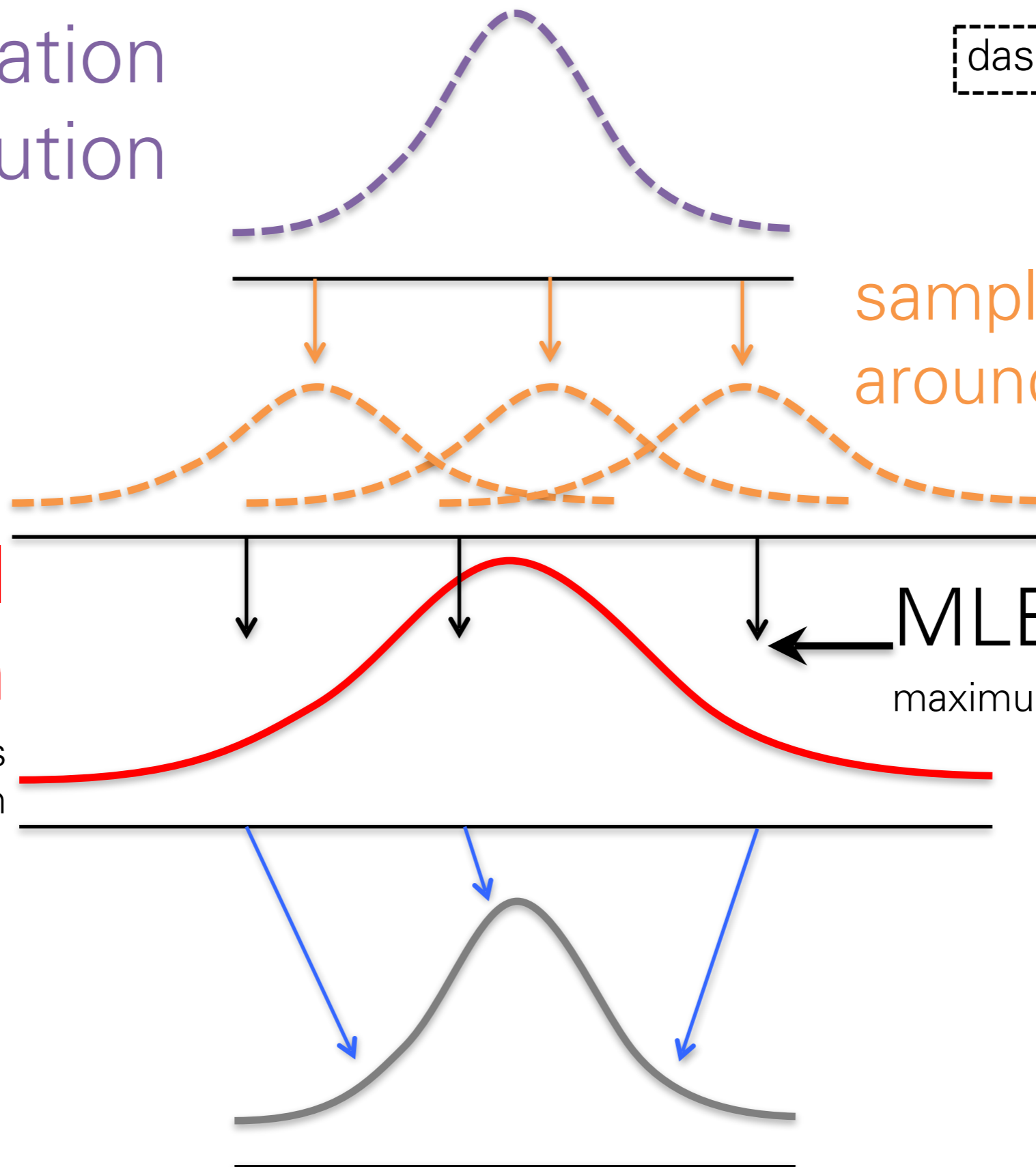
dashed = unobserved

sampling variance around true ability

empirical distribution

the center defines the prior mean

MLE
maximum likelihood estimates



Shrinkage estimation

population distribution

dashed = unobserved

sampling variance around true ability

empirical distribution

the center defines the prior mean

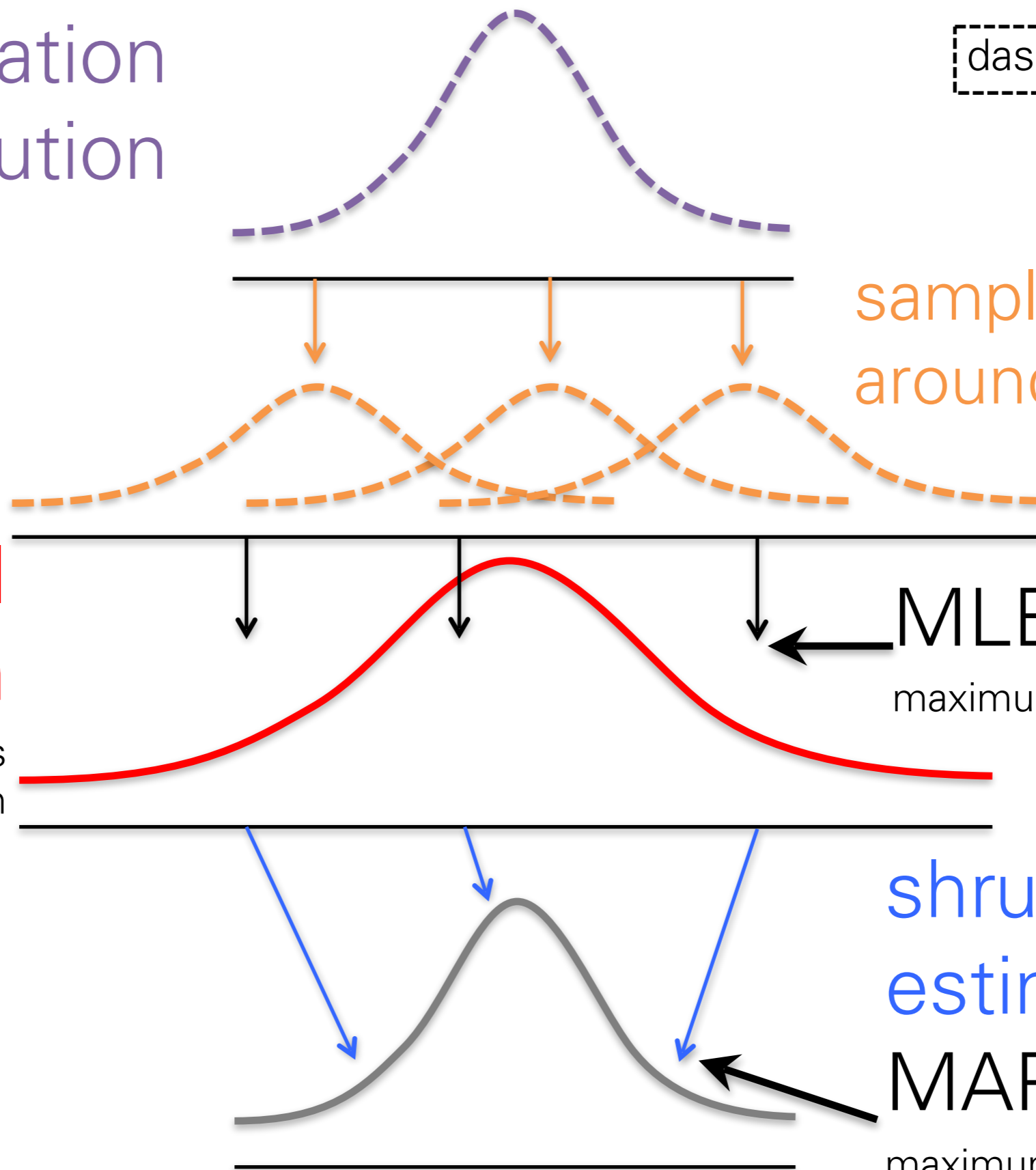
MLE

maximum likelihood estimates

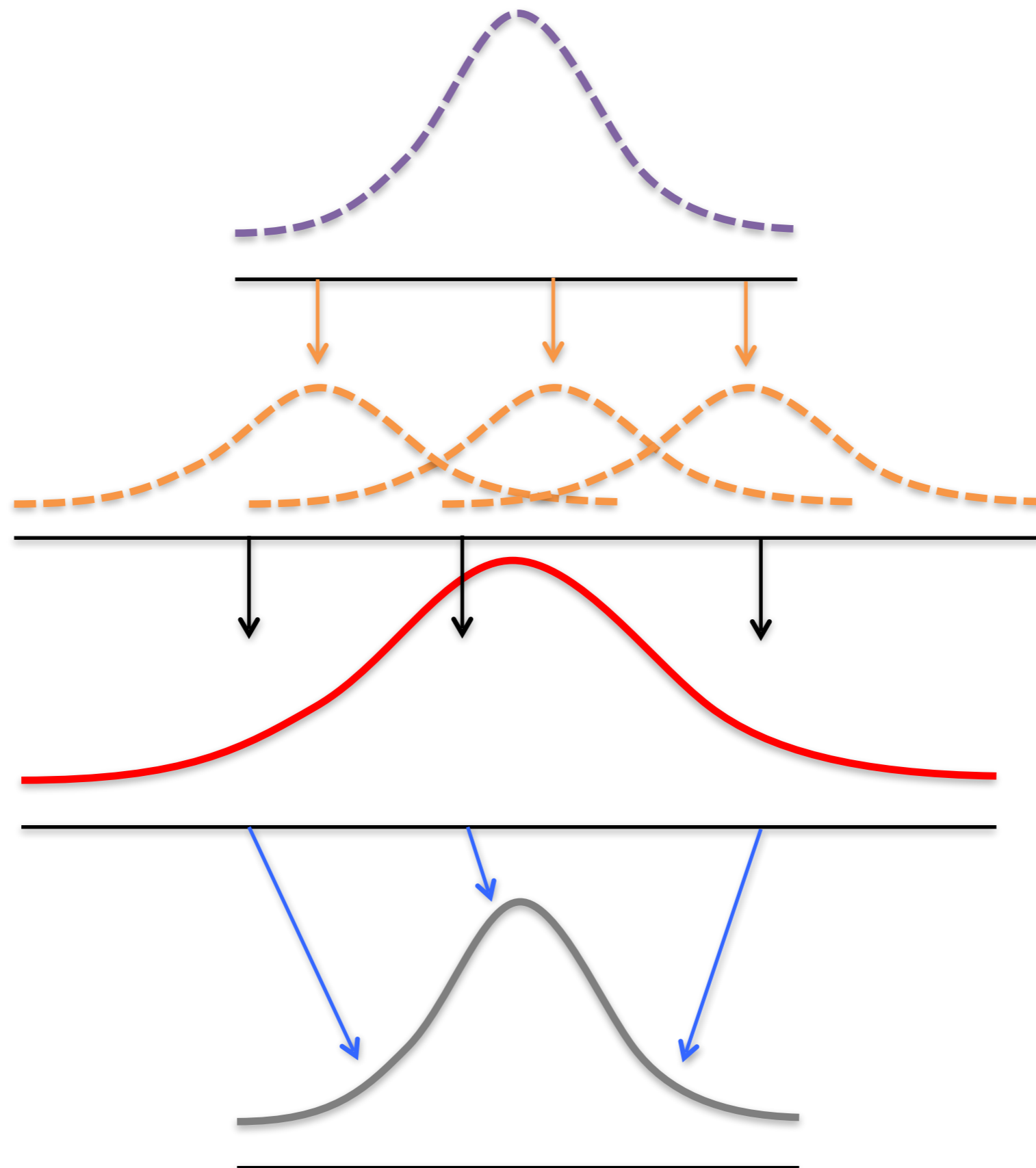
shrunk estimates or

MAP

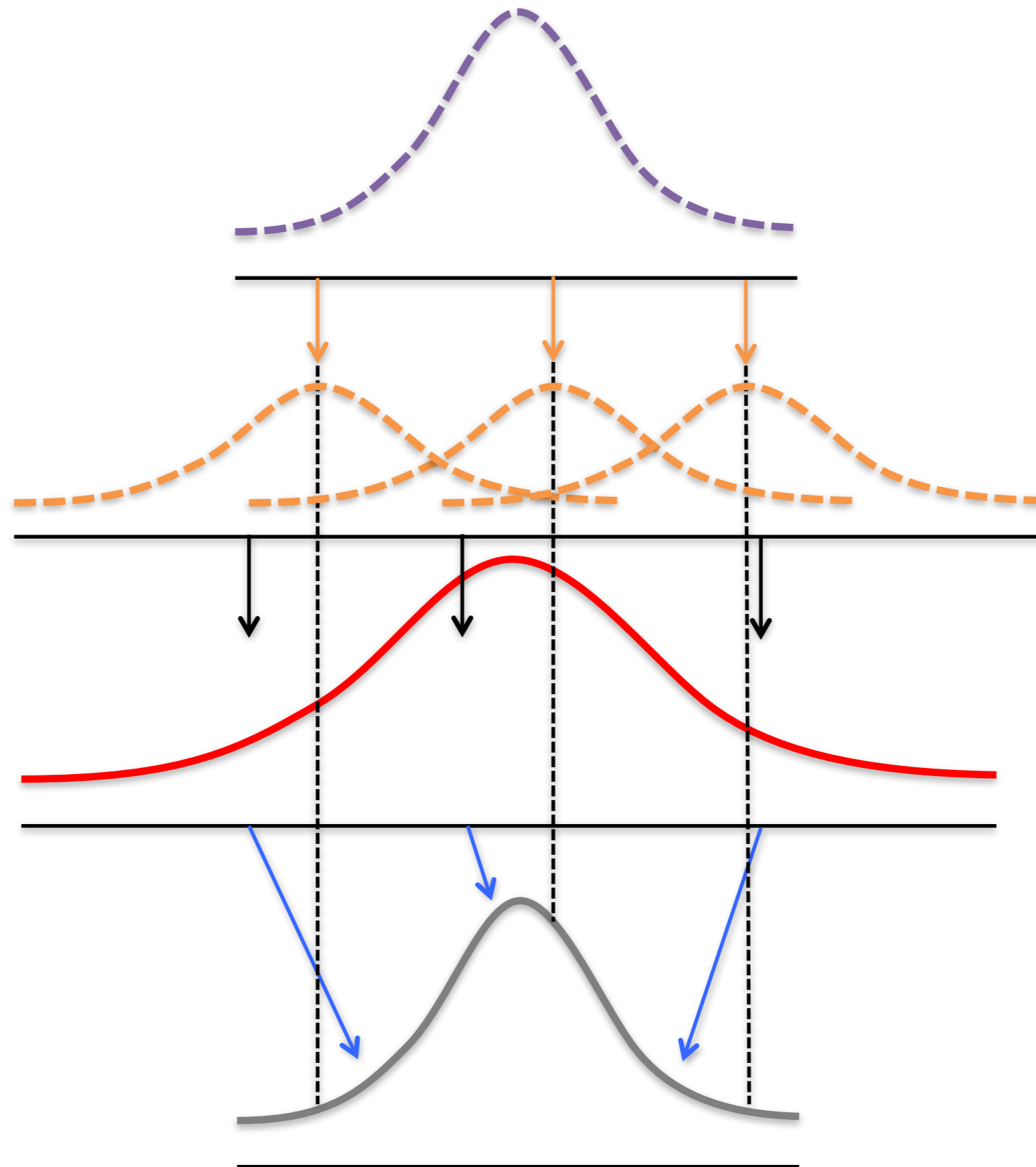
maximum a posteriori



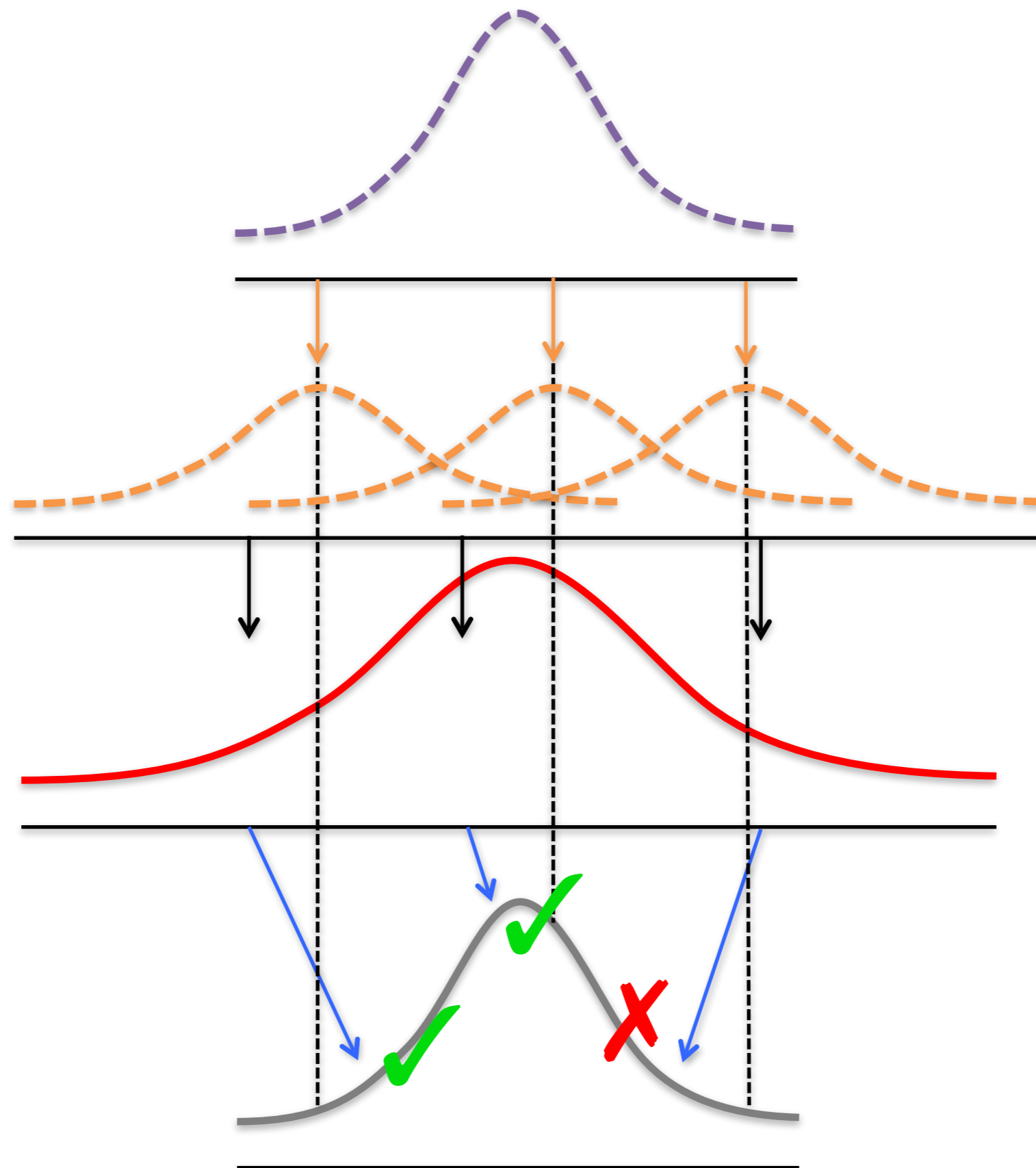
Shrinkage estimation



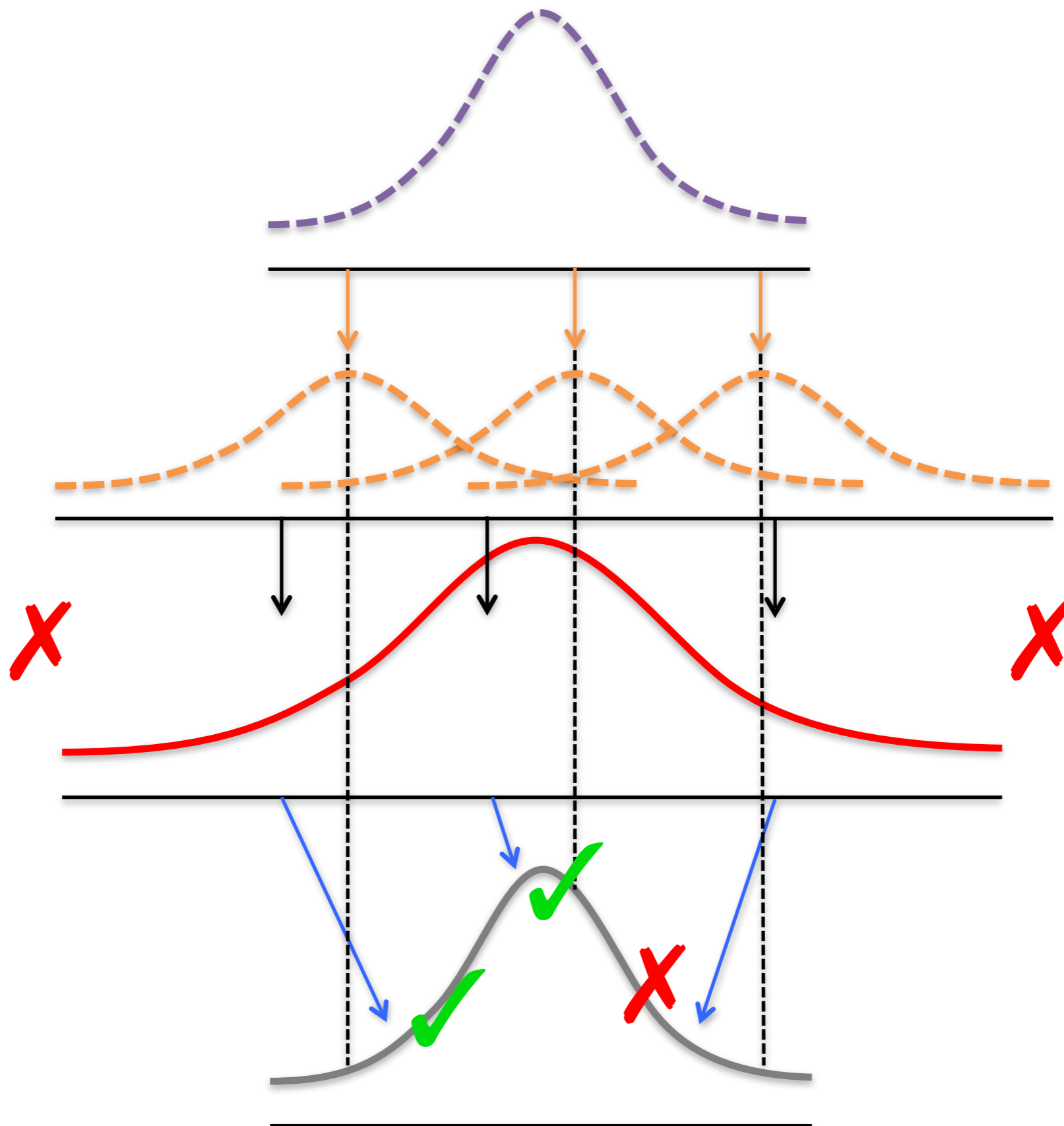
Shrinkage estimation



Shrinkage estimation



Shrinkage estimation

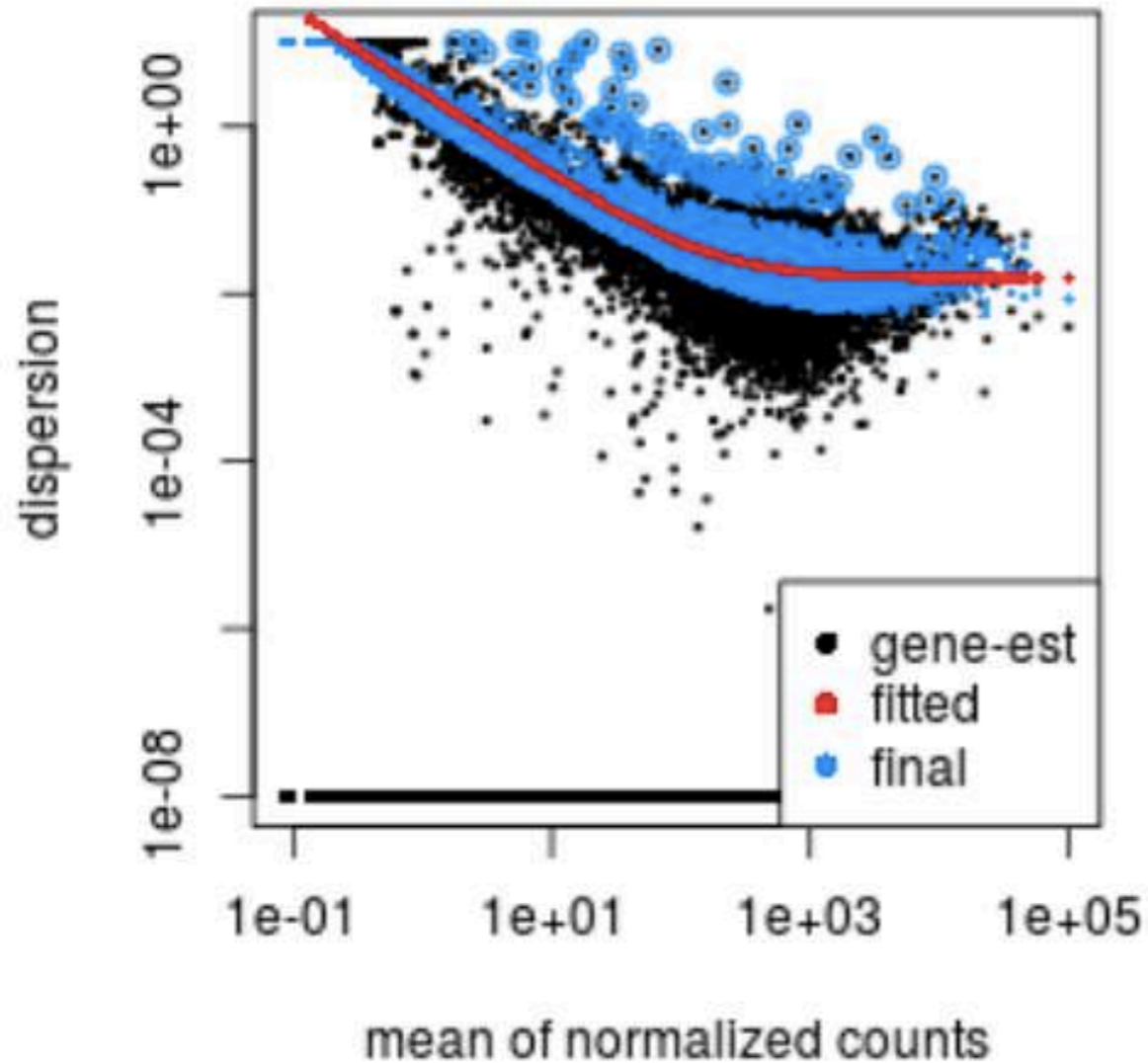


Shrinkage estimators in genomics

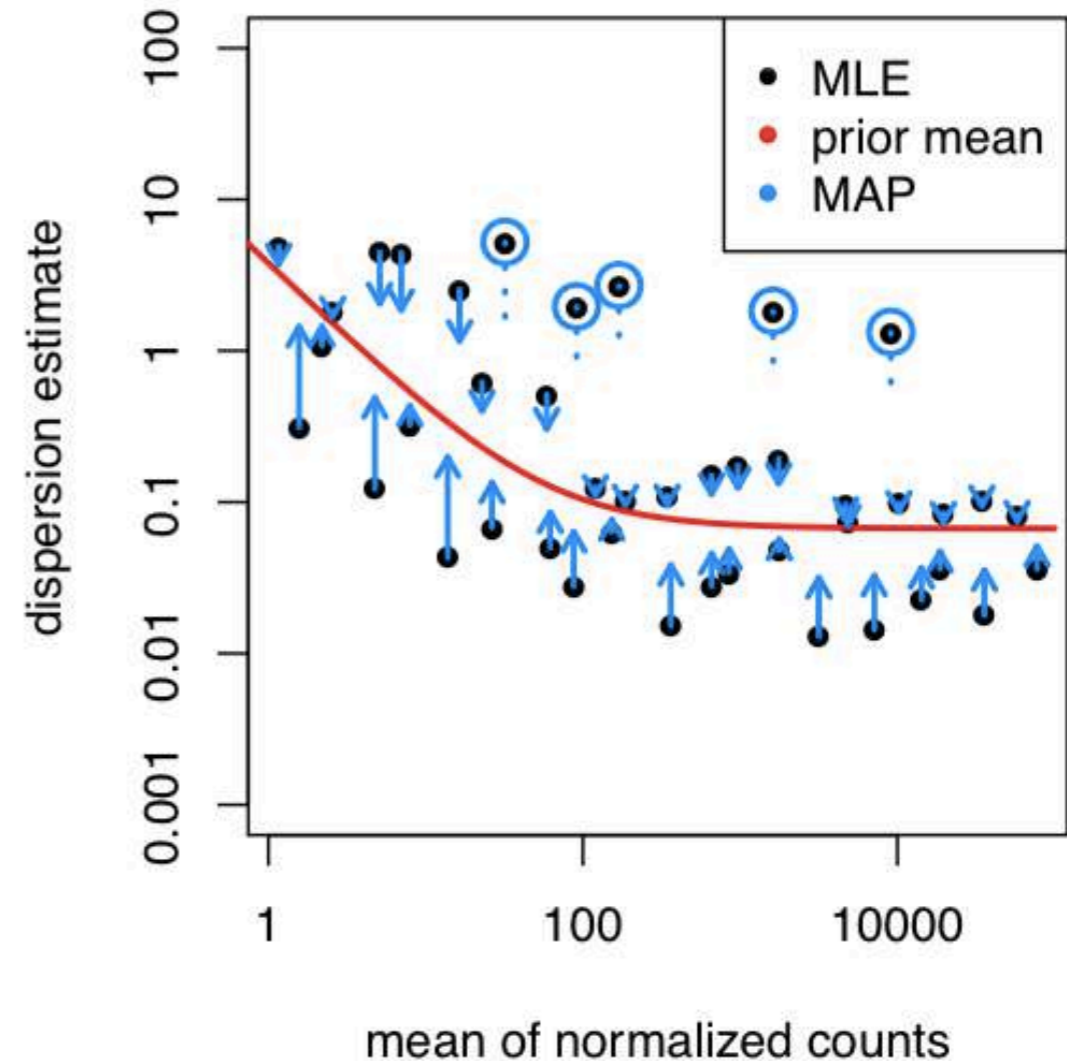
- Lönnstedt and Speed 2002: microarrays
- Smyth 2004: limma for microarrays
- Robinson and Smyth 2007: edgeR for SAGE and then applied to RNA-seq
- Many adaptations: DSS and DESeq2 use a similar approach, data-driven strength of shrinkage

Shrinkage of dispersion for RNA-seq

all genes (Pasilla)

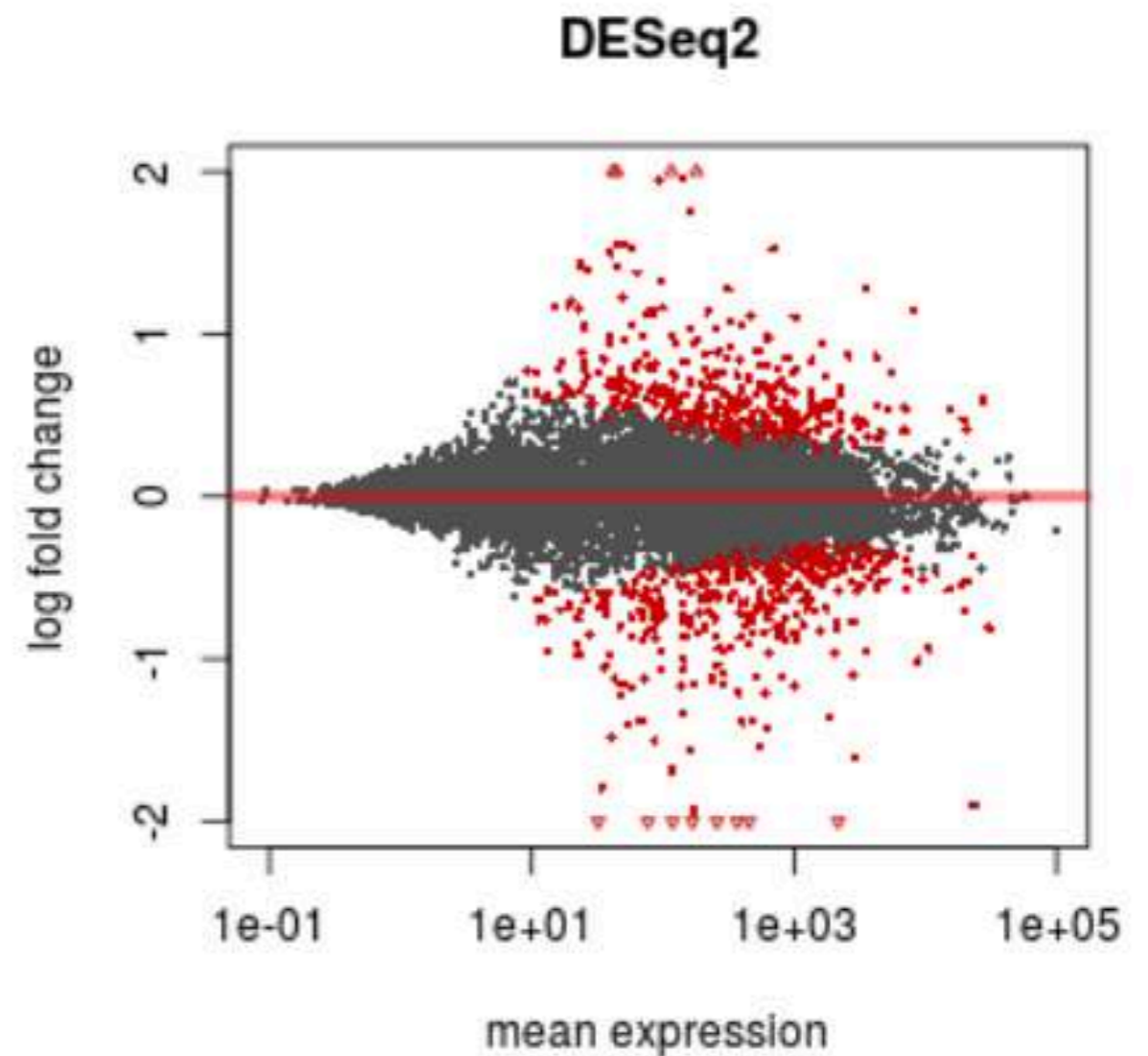
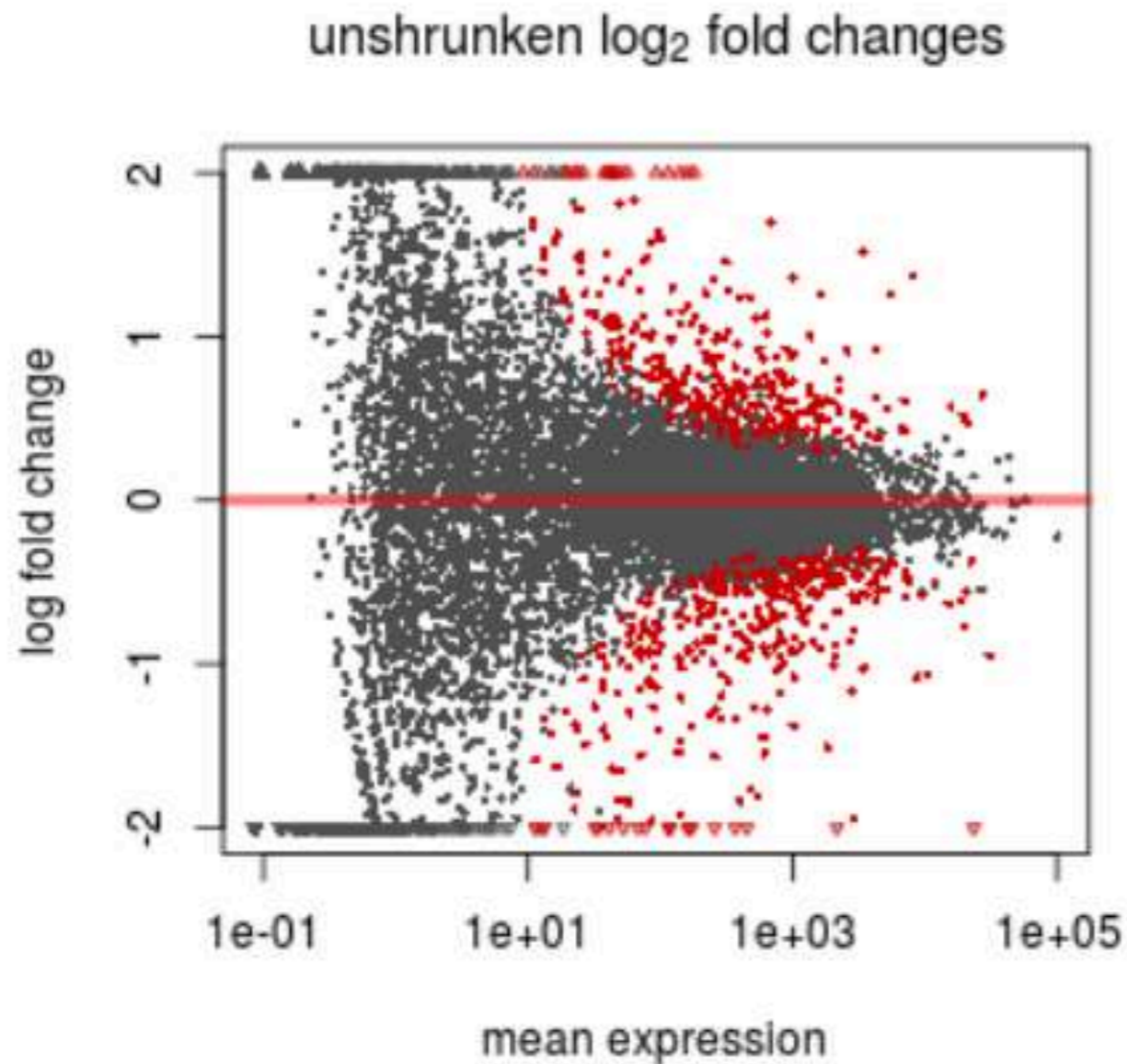


a subset of genes (Pickrell)

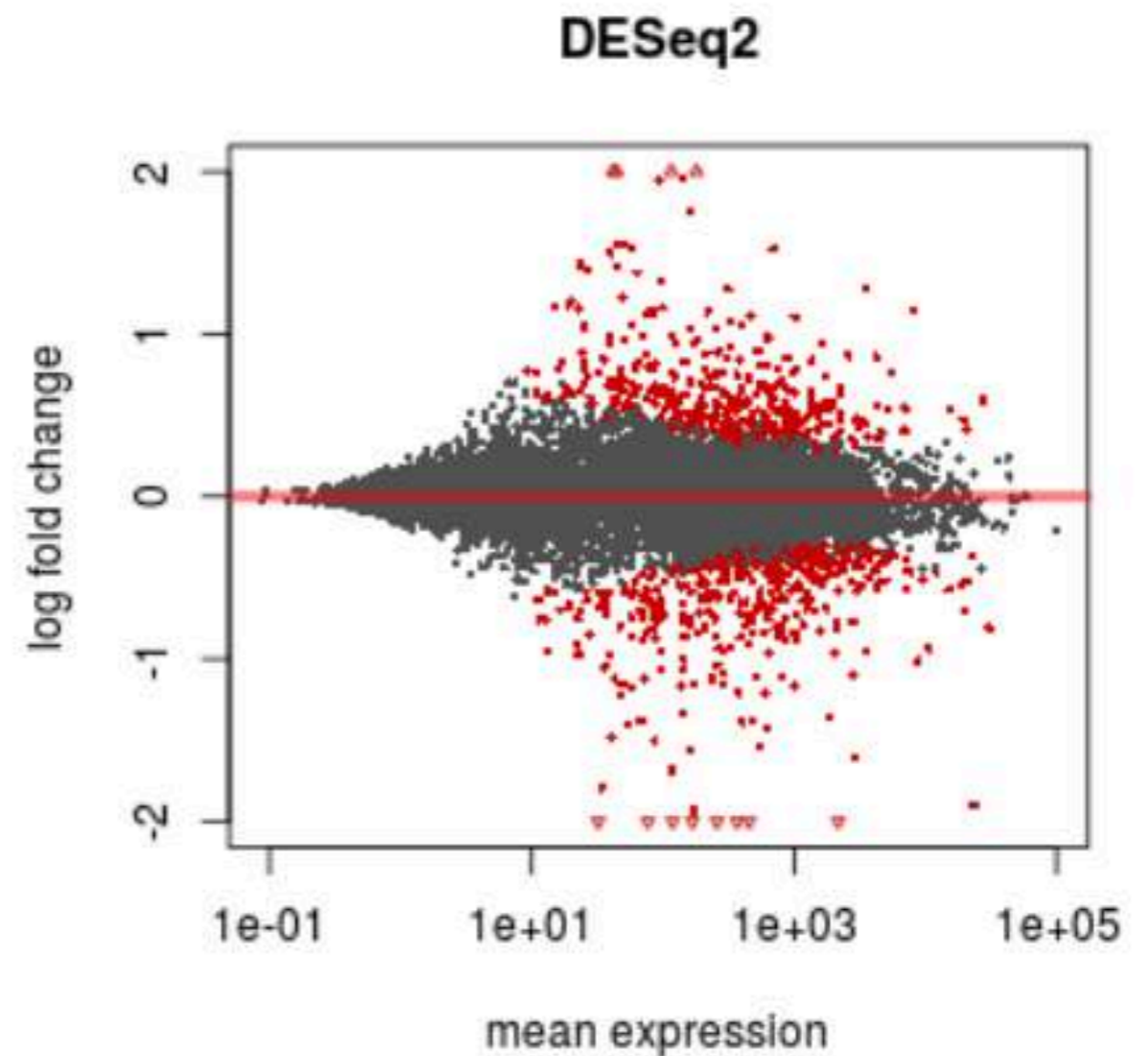
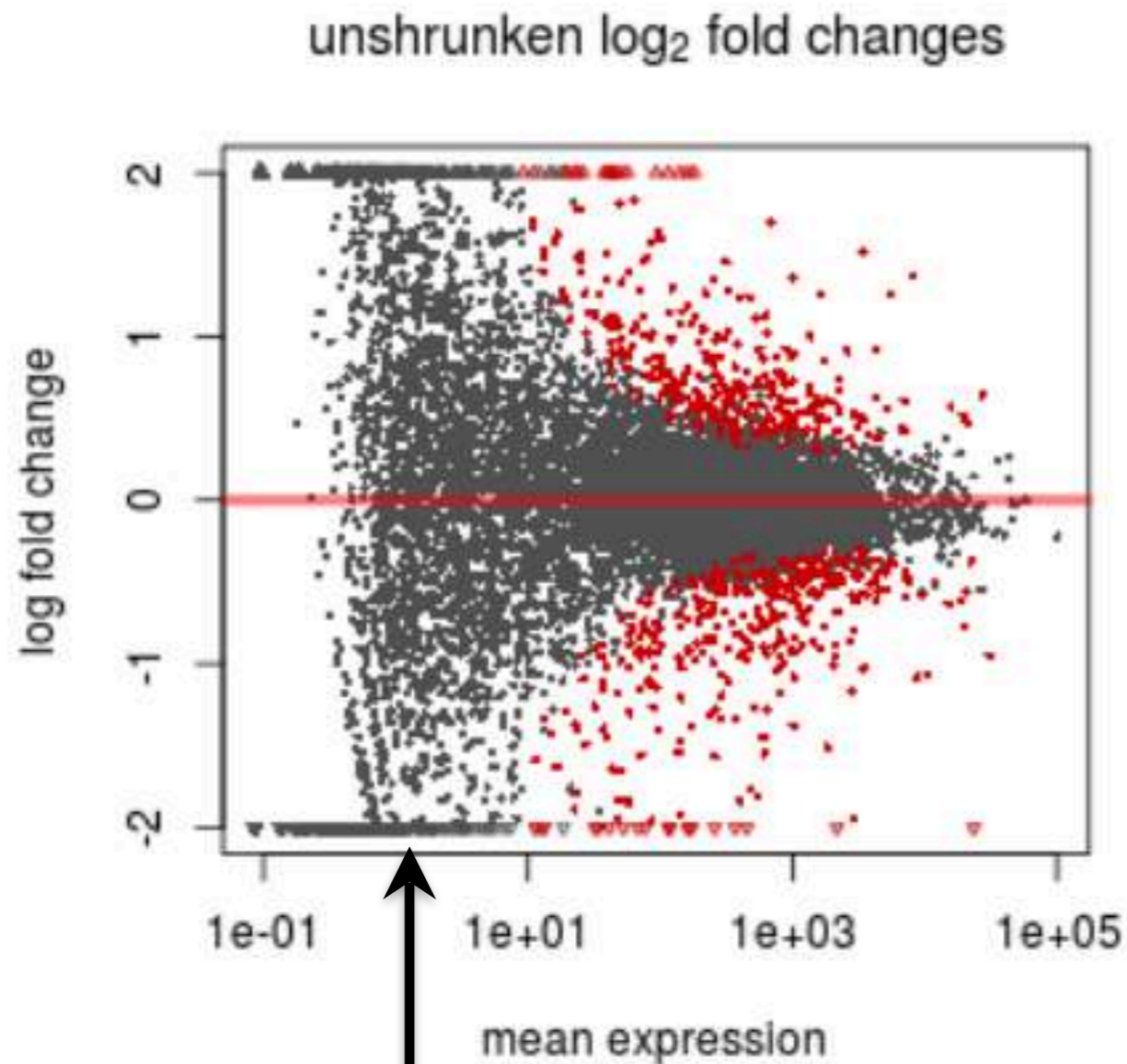


1. Gene estimate = maximum likelihood estimate (MLE)
2. Fitted dispersion trend = the mean of the prior
3. Final estimate = maximum a posteriori (MAP)

Shrinkage of fold changes for RNA-seq

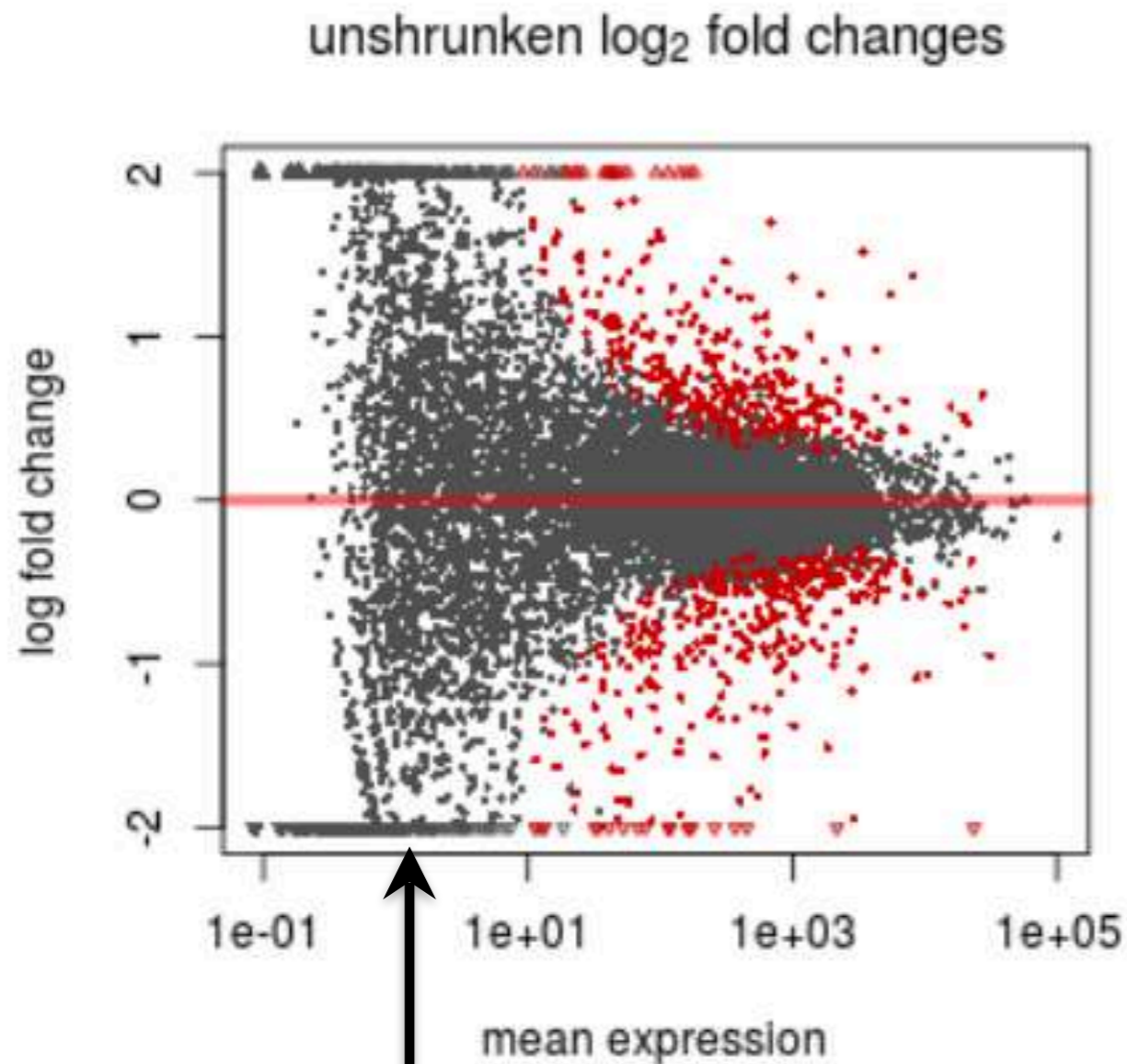


Shrinkage of fold changes for RNA-seq

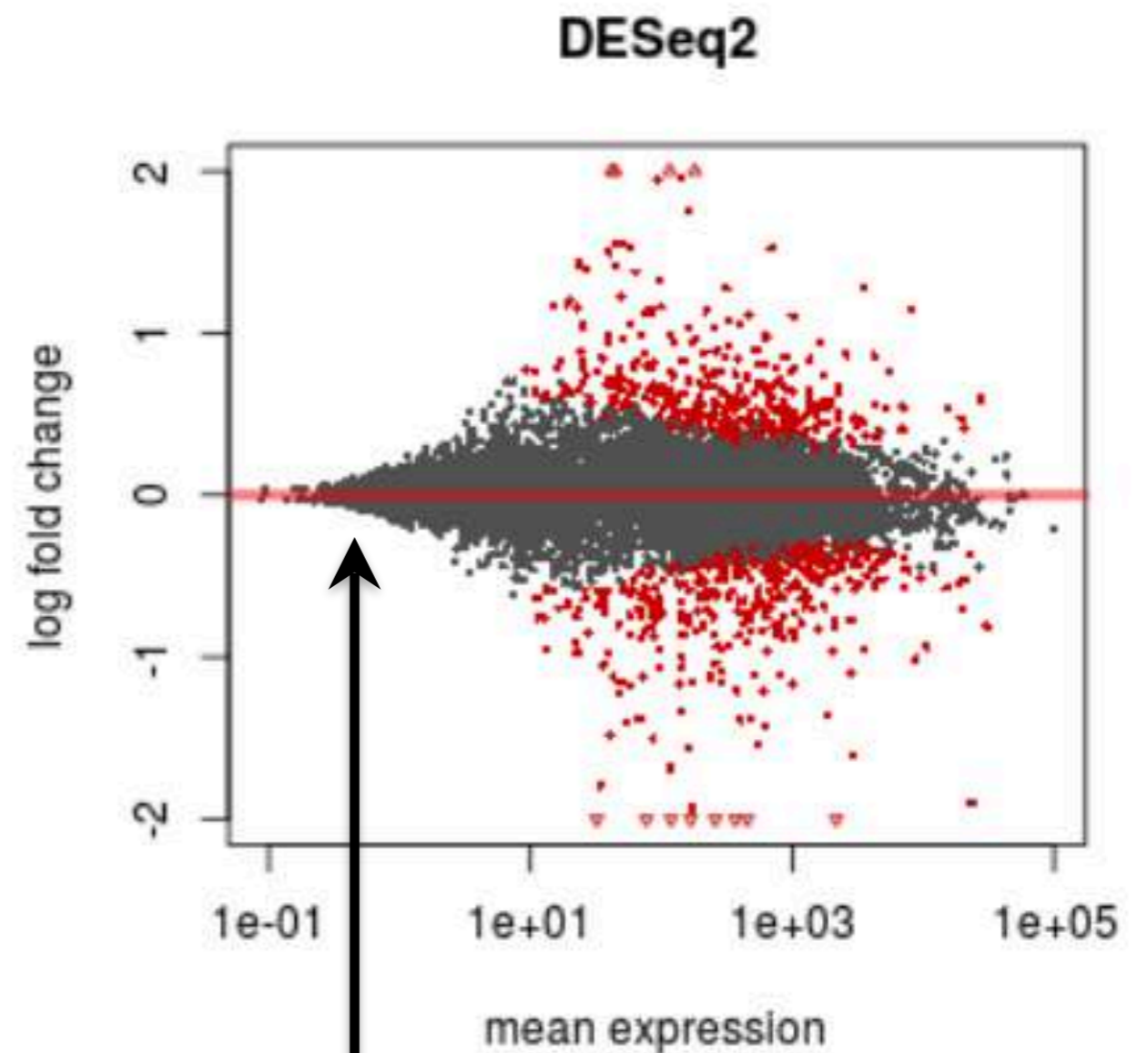


Noisy estimates due to low counts
statistical model will give large p-values,
but also the FC estimates themselves are
not trustworthy

Shrinkage of fold changes for RNA-seq

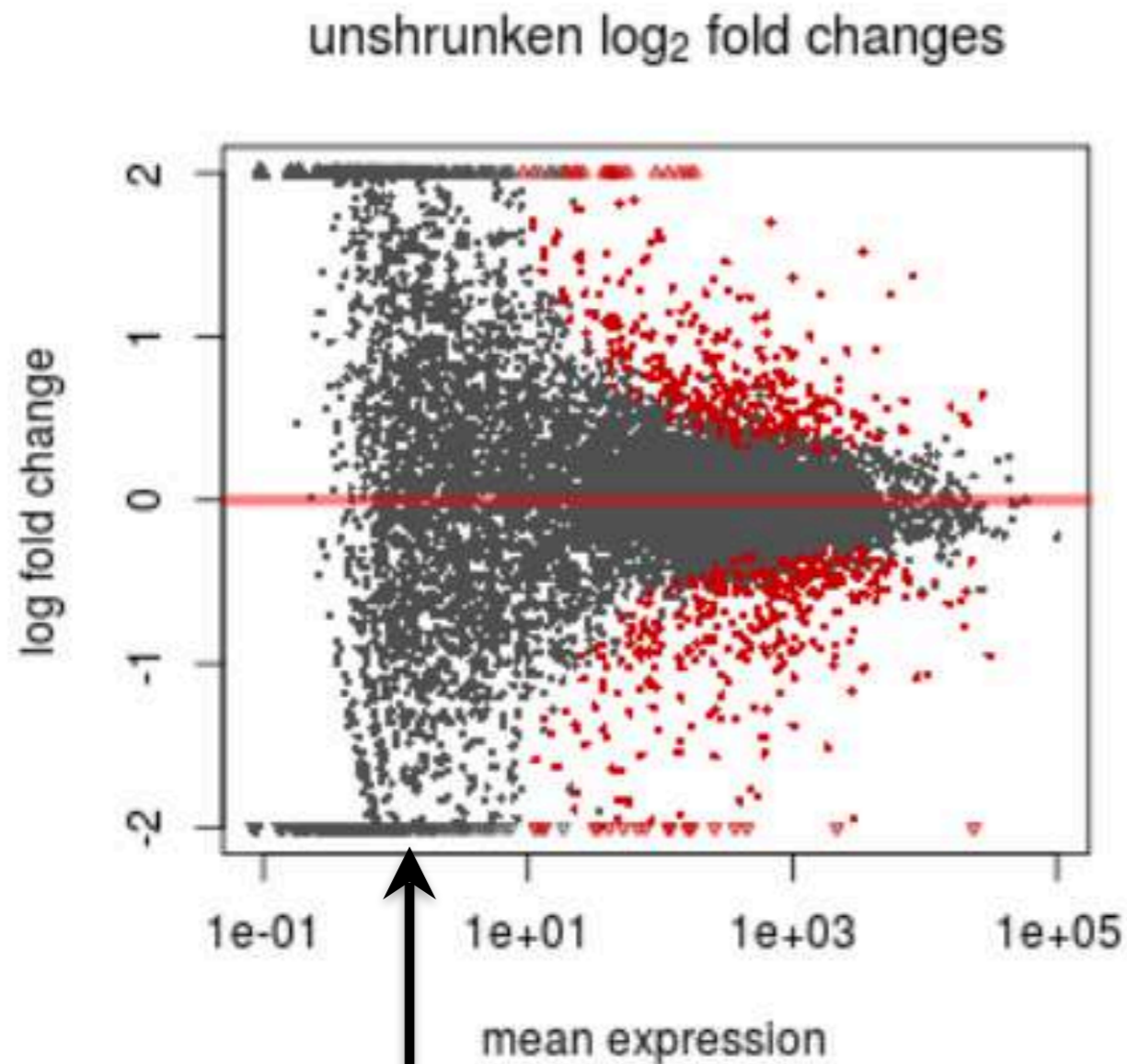


Noisy estimates due to low counts
statistical model will give large p-values, but also the FC estimates themselves are not trustworthy

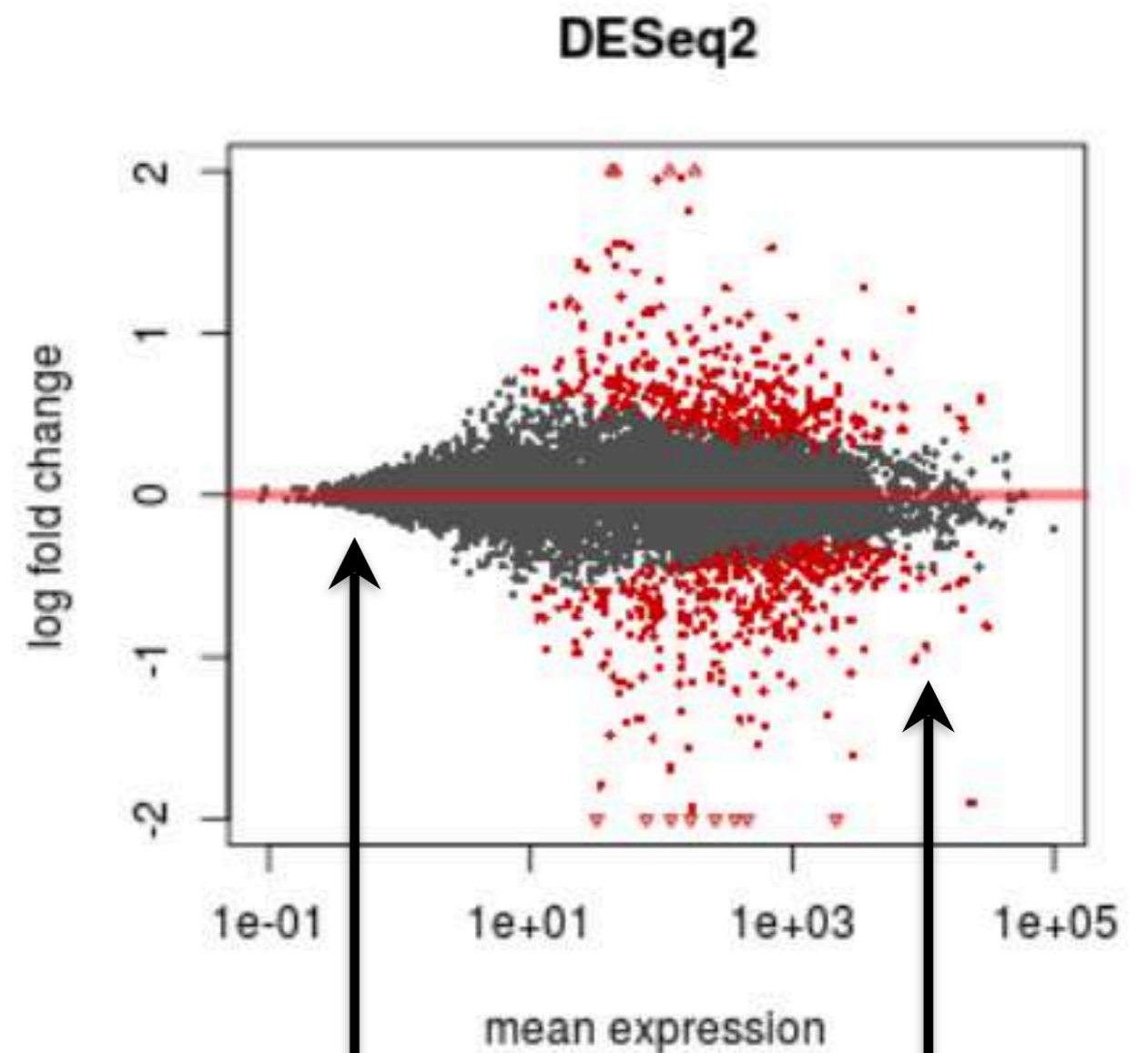


shrinkage is not equal.
strong moderation for low information genes: low counts

Shrinkage of fold changes for RNA-seq



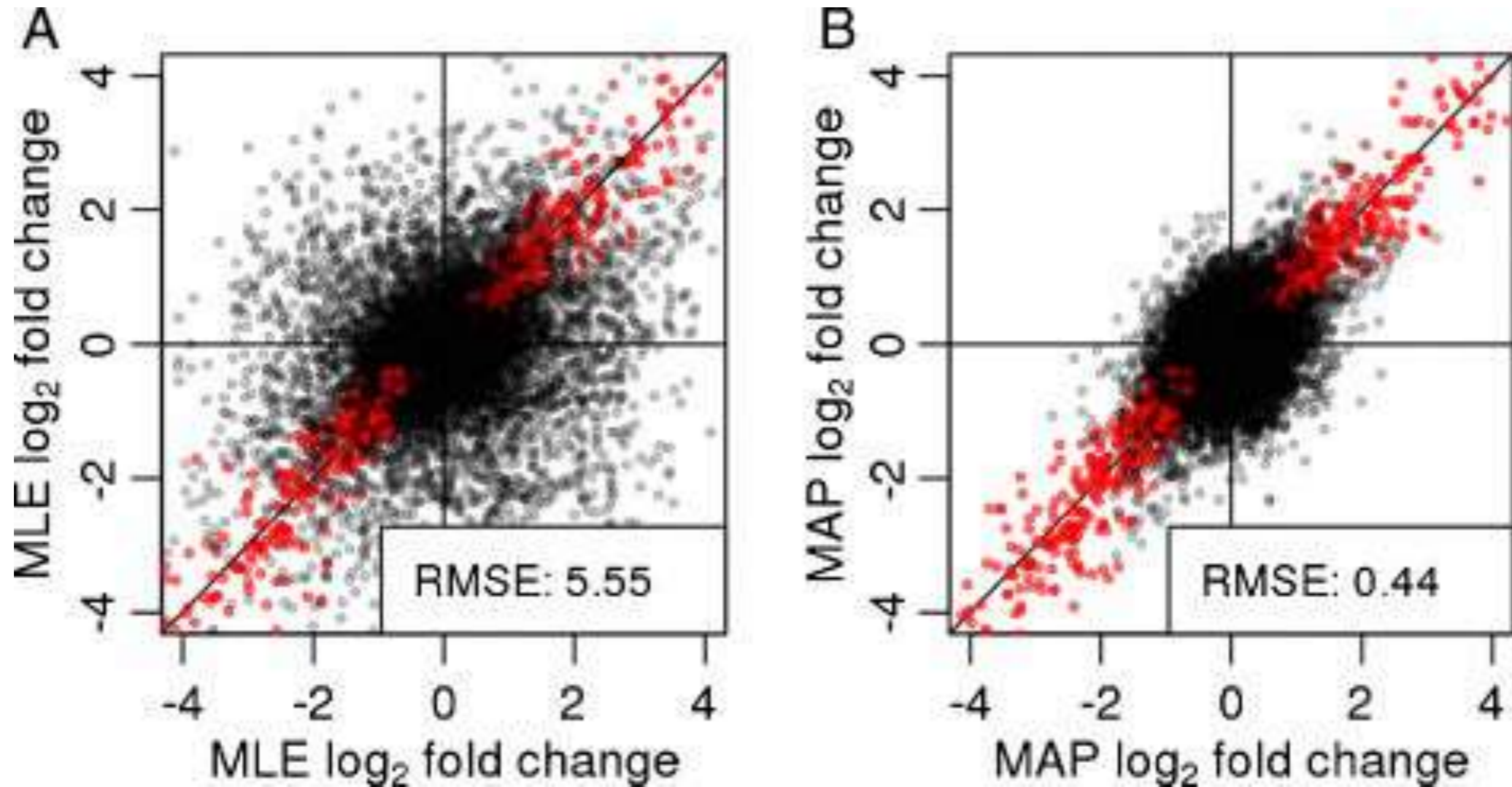
Noisy estimates due to low counts
statistical model will give large p-values, but also the FC estimates themselves are not trustworthy



shrinkage is not equal.
strong moderation for low information genes: low counts

almost no shrinkage

Why shrink fold changes?



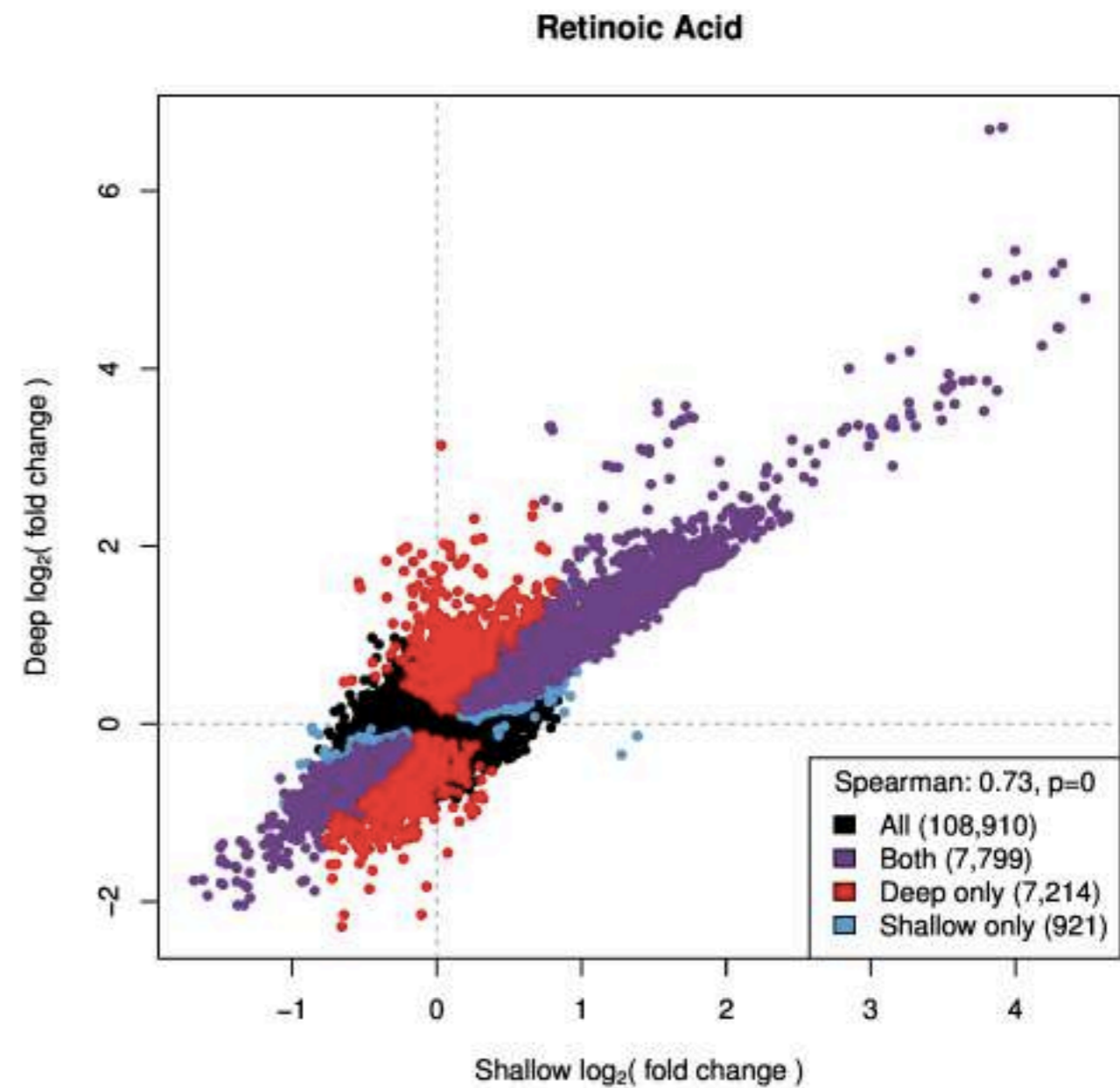
Split a dataset into two equal parts, compare LFC

Why shrink fold changes?

Comparison of log fold changes across two experiments.

"A new two-step high-throughput approach:

1. gene expression screening of a large number of conditions
2. deep sequencing of the most relevant conditions"



G. A. Moyerbrailean et al. "A high-throughput RNA-seq approach to profile transcriptional responses" <http://dx.doi.org/10.1101/018416>

The maths: empirical Bayes shrinkage of gene-wise dispersion estimates

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

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

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

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

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

The maths: empirical Bayes shrinkage of gene-wise dispersion estimates

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu})) \quad \longleftarrow \quad \text{"naive" GLM likelihood}$$

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

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

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

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

The maths: empirical Bayes shrinkage of gene-wise dispersion estimates

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu})) \quad \longleftarrow \quad \text{"naive" GLM likelihood}$$

$$\text{CR}(\alpha) = -\frac{1}{2} \log(\det(X^t W X)) \quad \longleftarrow \quad \text{Cox-Reid bias term}$$

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

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

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

The maths: empirical Bayes shrinkage of gene-wise dispersion estimates

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu})) \quad \longleftarrow \quad \text{"naive" GLM likelihood}$$

$$\text{CR}(\alpha) = -\frac{1}{2} \log(\det(X^t W X)) \quad \longleftarrow \quad \text{Cox-Reid bias term}$$

$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha)) \quad \longleftarrow \quad \text{bias-corrected likelihood}$$

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

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

The maths: empirical Bayes shrinkage of gene-wise dispersion estimates

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu})) \quad \leftarrow \text{“naive” GLM likelihood}$$

$$\text{CR}(\alpha) = -\frac{1}{2} \log(\det(X^t W X)) \quad \leftarrow \text{Cox-Reid bias term}$$

$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha)) \quad \leftarrow \text{bias-corrected likelihood}$$

$$\text{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma_{\alpha\text{-prior}}^2) \quad \leftarrow$$

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

prior on α by ‘information sharing’
across genes

The maths: empirical Bayes shrinkage of gene-wise dispersion estimates

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu})) \quad \leftarrow \text{“naive” GLM likelihood}$$

$$\text{CR}(\alpha) = -\frac{1}{2} \log(\det(X^t W X)) \quad \leftarrow \text{Cox-Reid bias term}$$

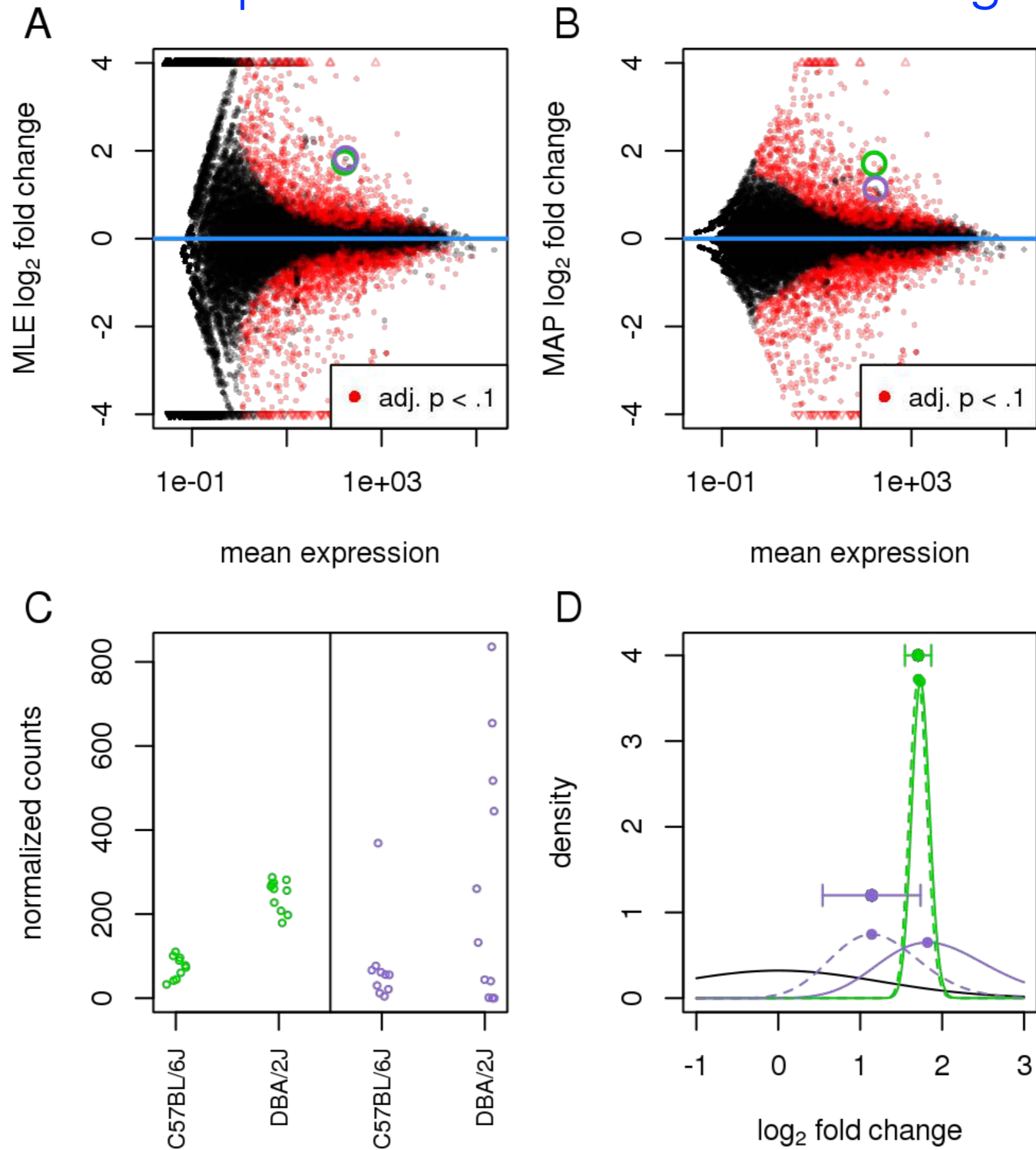
$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha)) \quad \leftarrow \text{bias-corrected likelihood}$$

$$\text{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma_{\alpha\text{-prior}}^2) \quad \leftarrow$$

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

prior on α by ‘information sharing’ across genes

Example: difference between maximum likelihood and maximum a posteriori estimate for two genes



Banded hypothesis testing: integrate testing with fold-change cutoff

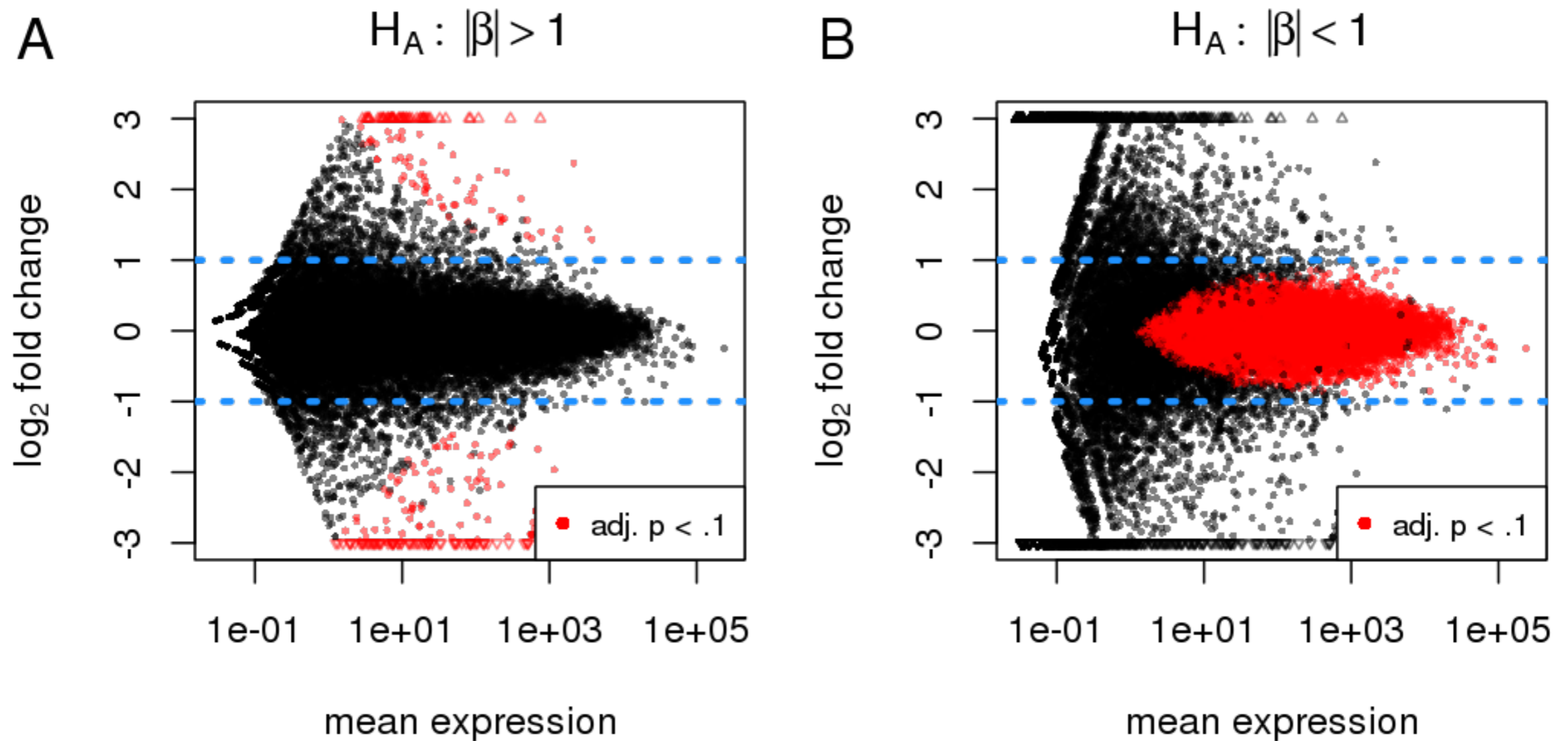
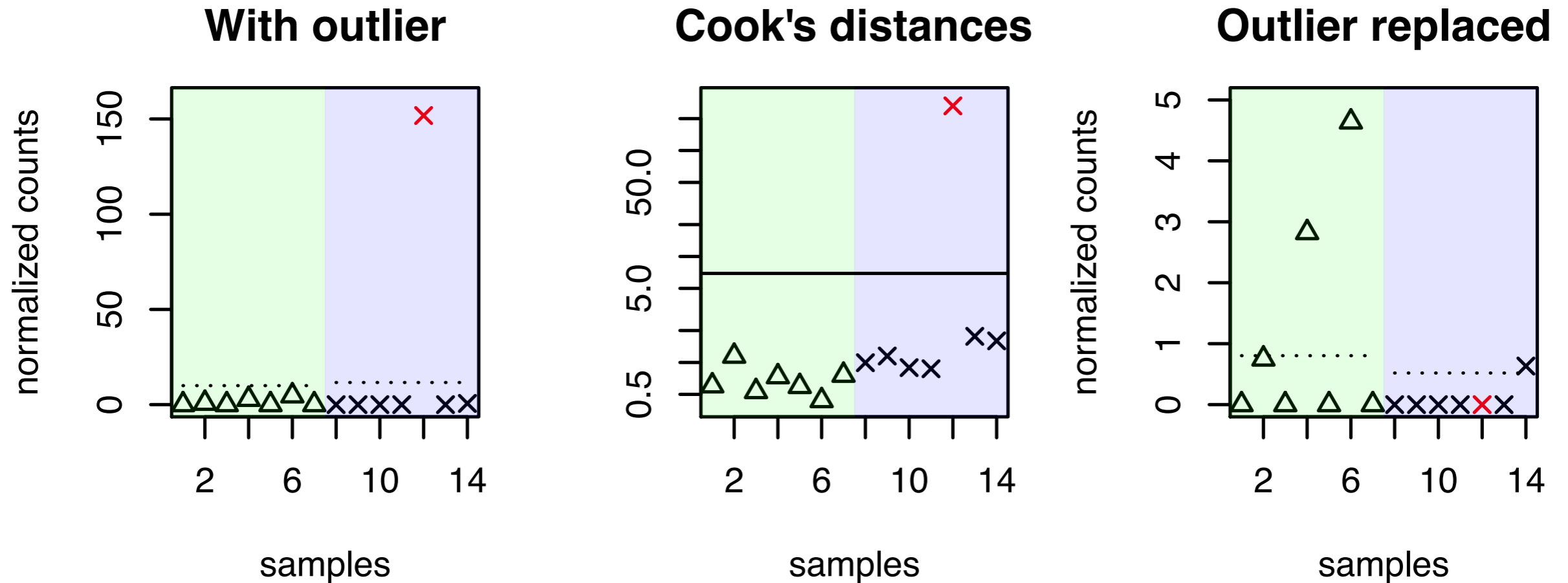


Figure 4 Hypothesis testing involving non-zero thresholds. Shown are MA-plots for a 10 vs 11 comparison using the Bottomly *et al.* [15] dataset, with highlighted points indicating low adjusted p -values. The alternate hypotheses are that logarithmic (base 2) fold changes are (A) greater than 1 in absolute value or (B) less than 1 in absolute value.

Outlier robustness

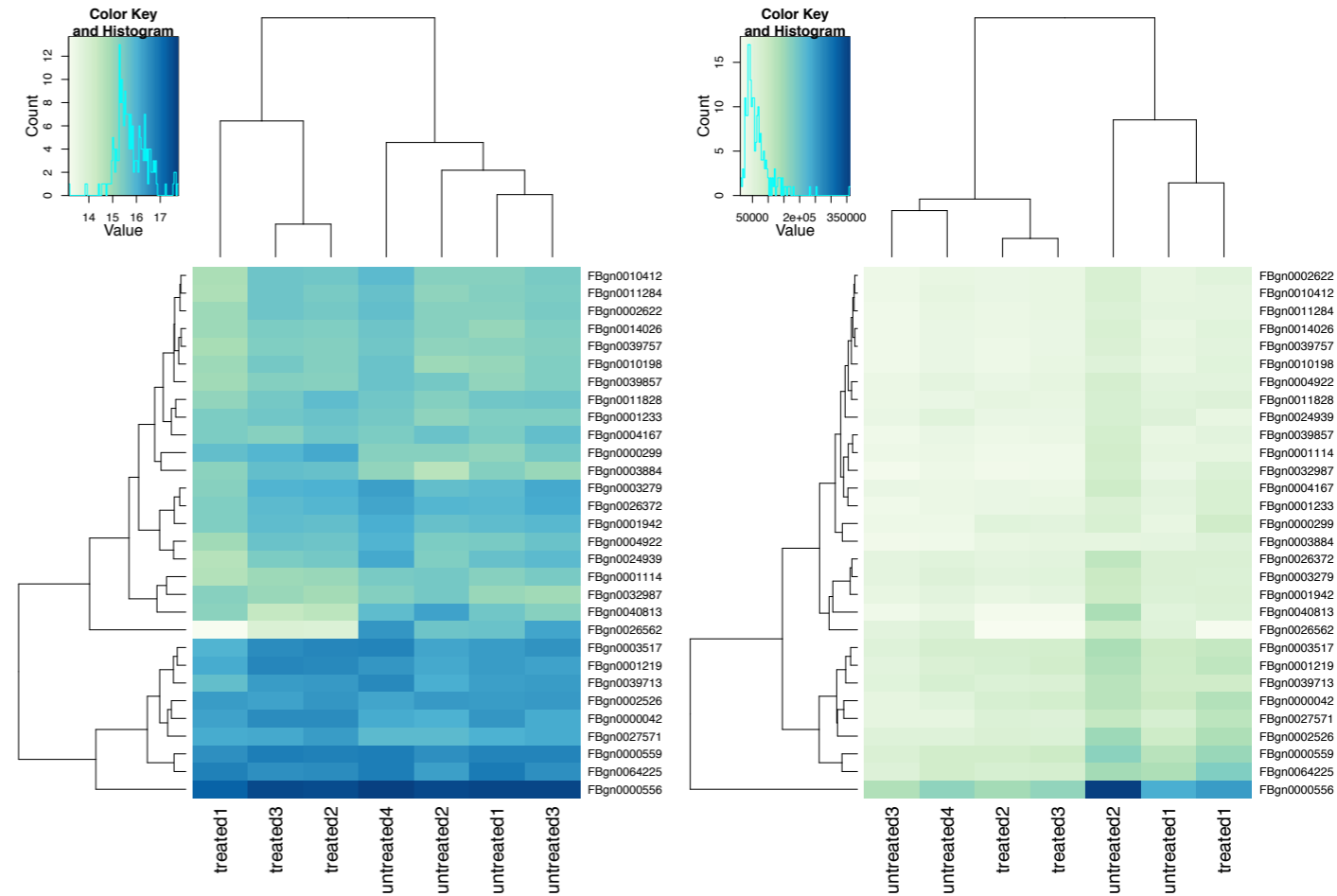
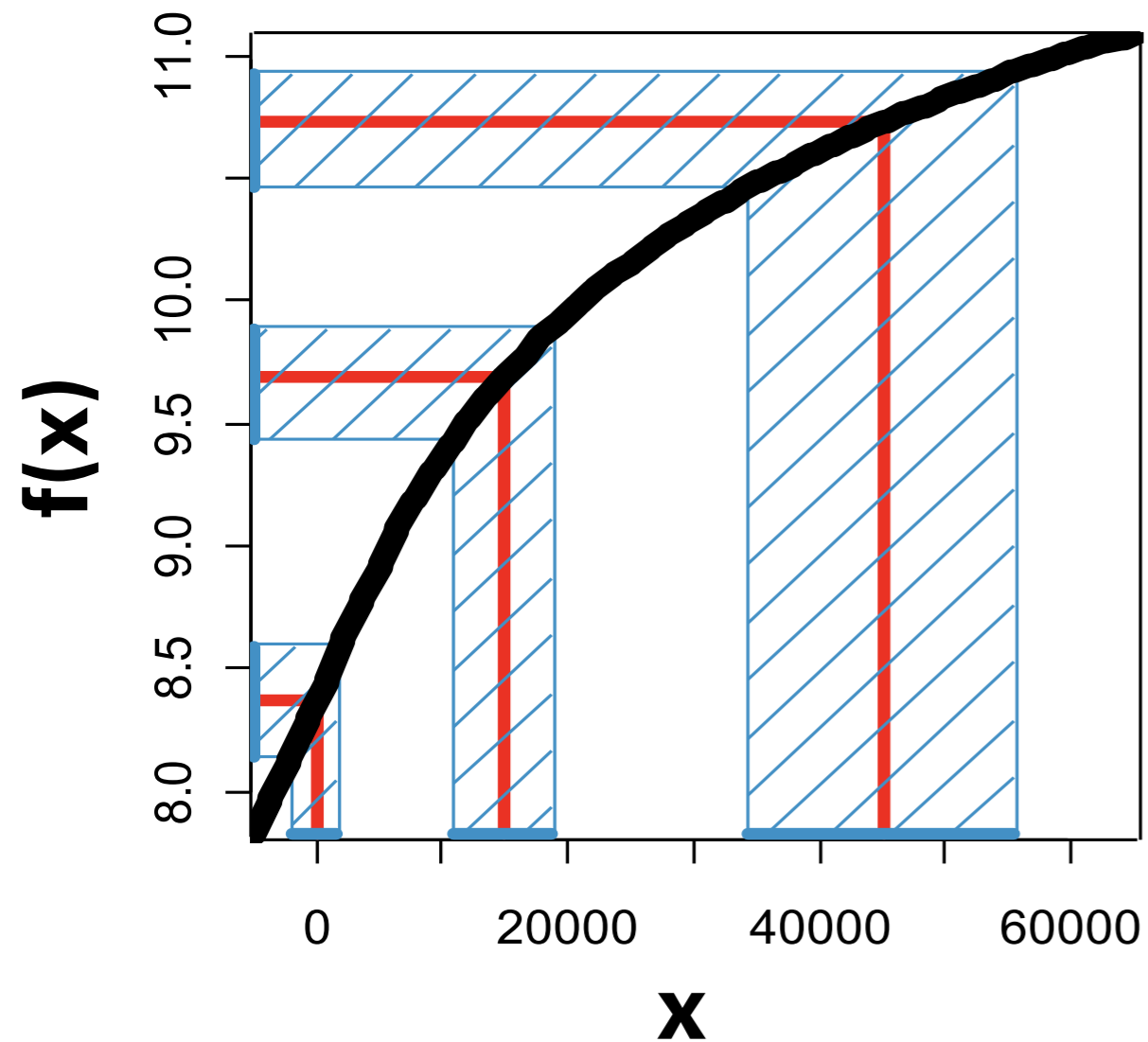


Cook's distance:

Change in fitted coefficients if the sample were removed

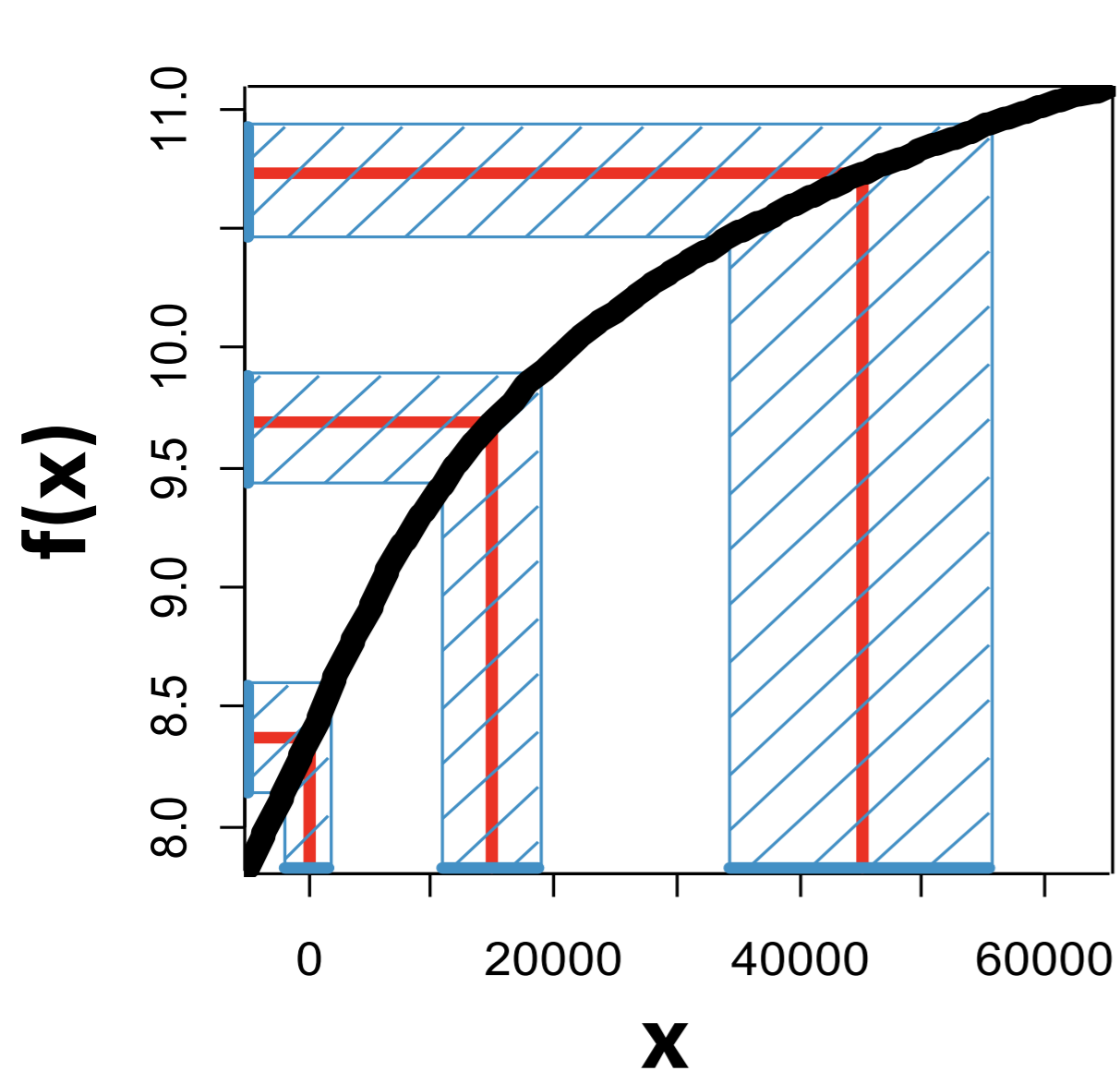
Variance-stabilizing transformation

$$f(x) = \int^x \frac{du}{\sqrt{v(u)}}$$



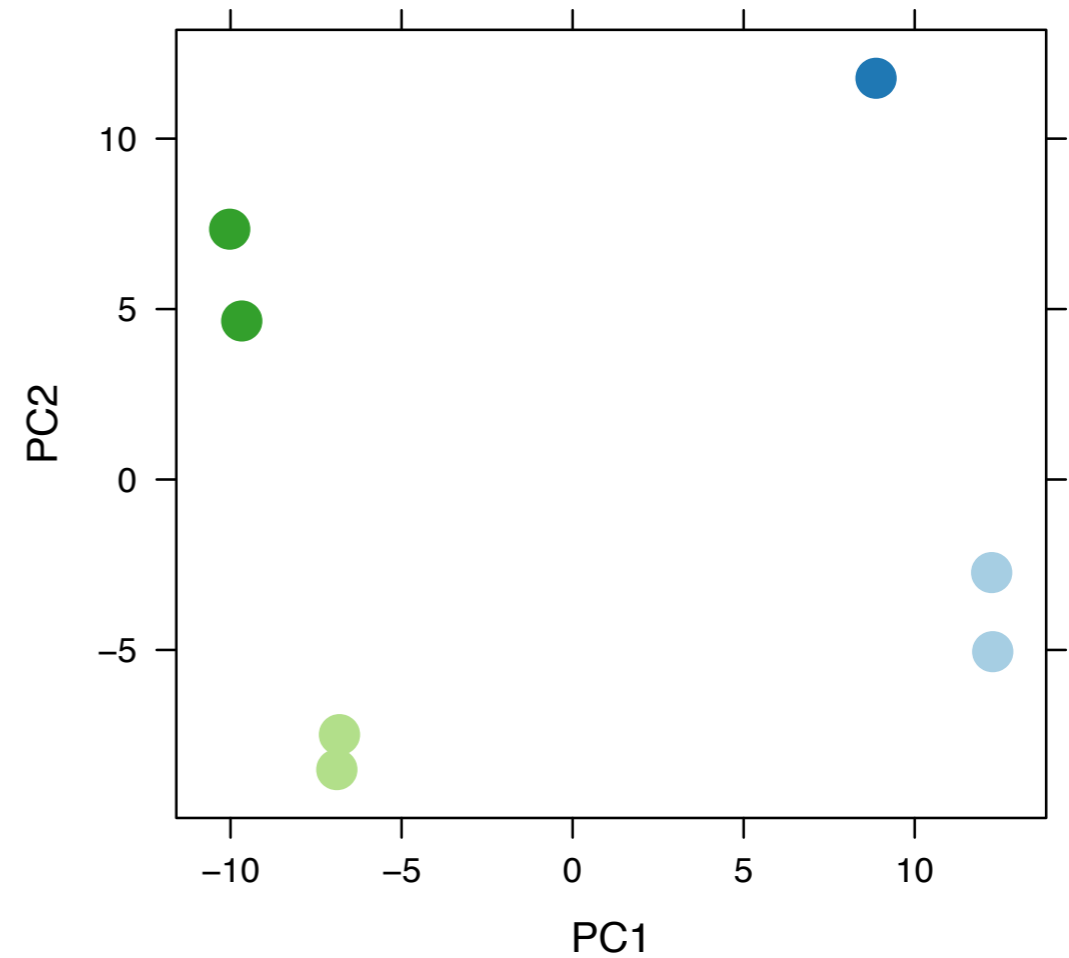
Variance-stabilizing transformation

$$f(x) = \int^x \frac{du}{\sqrt{v(u)}}$$

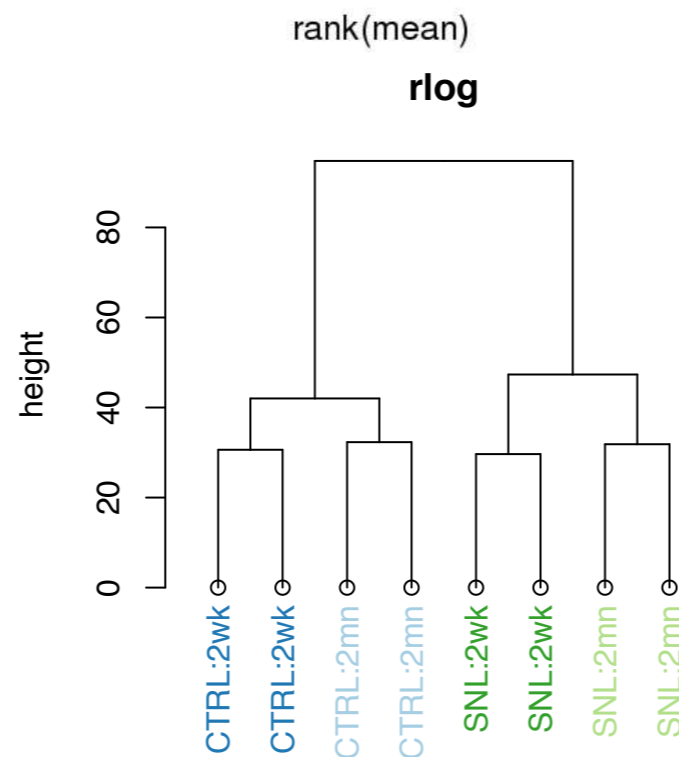
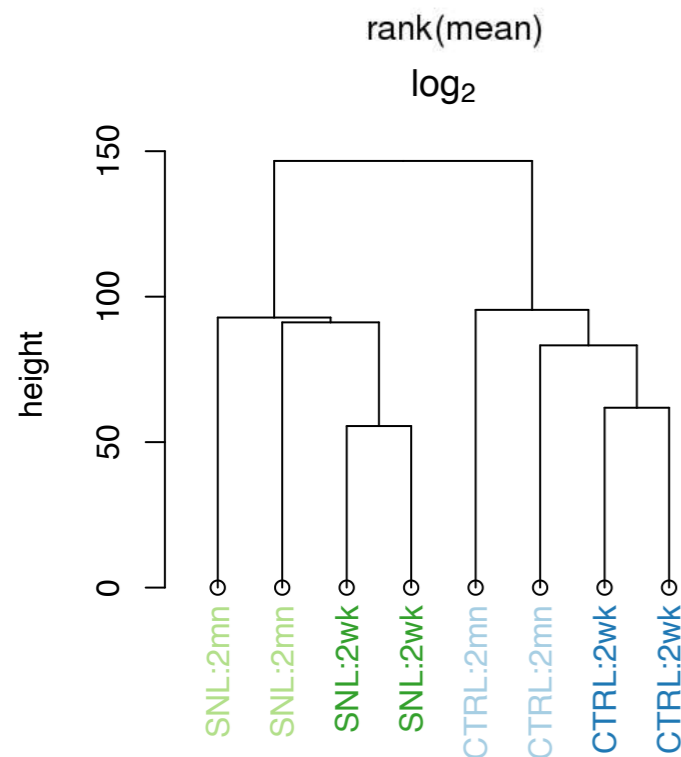
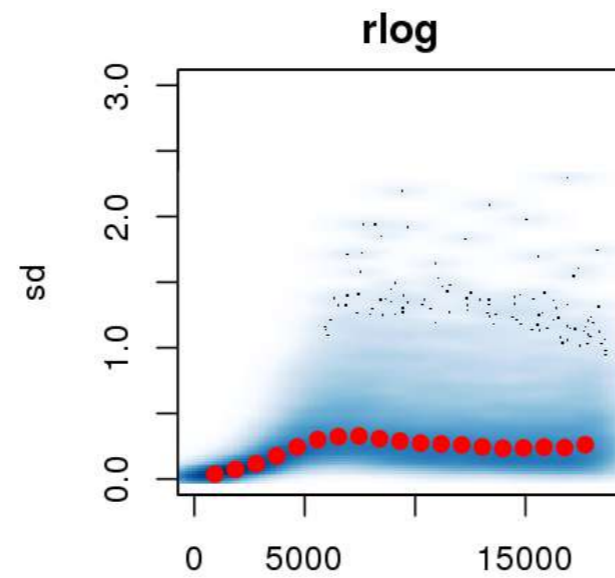
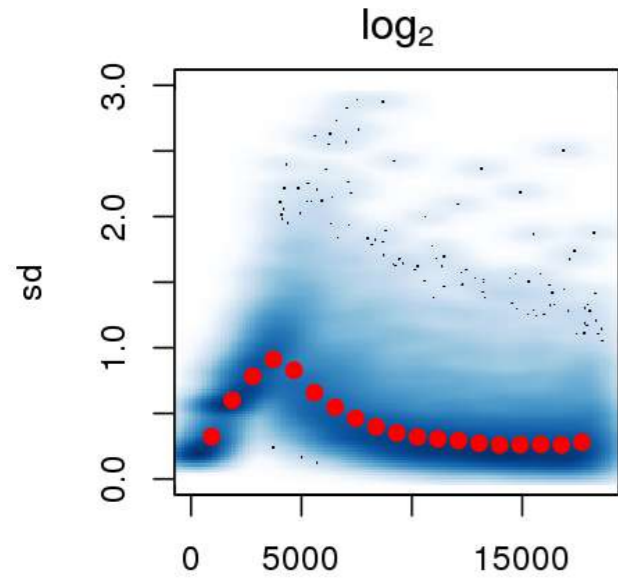


Legend for the histogram:

- treated : paired-end
- treated : single-end
- untreated : paired-end
- untreated : single-end



Regularized log-transformation: Visualization, Clustering, PCA

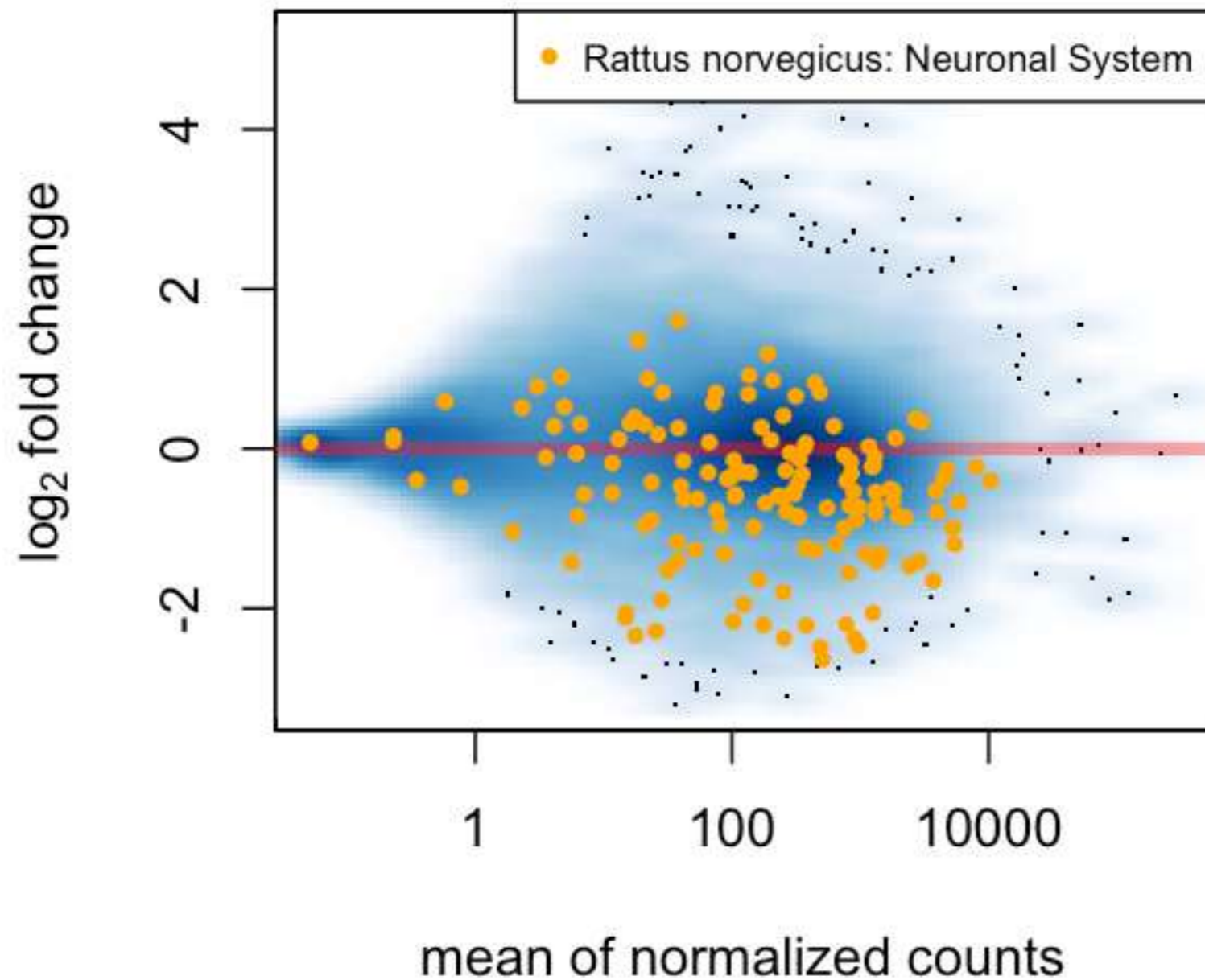


"rlog":

Shrunken log fold changes for every sample:
reduces effect of shot noise on inter-sample distances

RNA from the dorsal root ganglion of rats that underwent spinal nerve ligation and controls, 2 weeks & 2 months after the ligation. Hammer, ..., Beutler AS, Genome Research 2010.

GSEA with shrunken log fold changes



Reactome gene set
one-sample t-statistic

Neuronal System

144 genes

avg LFC: -0.55

adjusted p-value: 10^{-8}

RNA from the dorsal root ganglion of rats that underwent spinal nerve ligation and controls, 2 weeks & 2 months after the ligation. Hammer, ..., Beutler AS, Genome Research 2010.

Summary so far

- 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
- Outliers
- Data transformation: visualisation, clustering, classification
- Next up: exon-level analysis

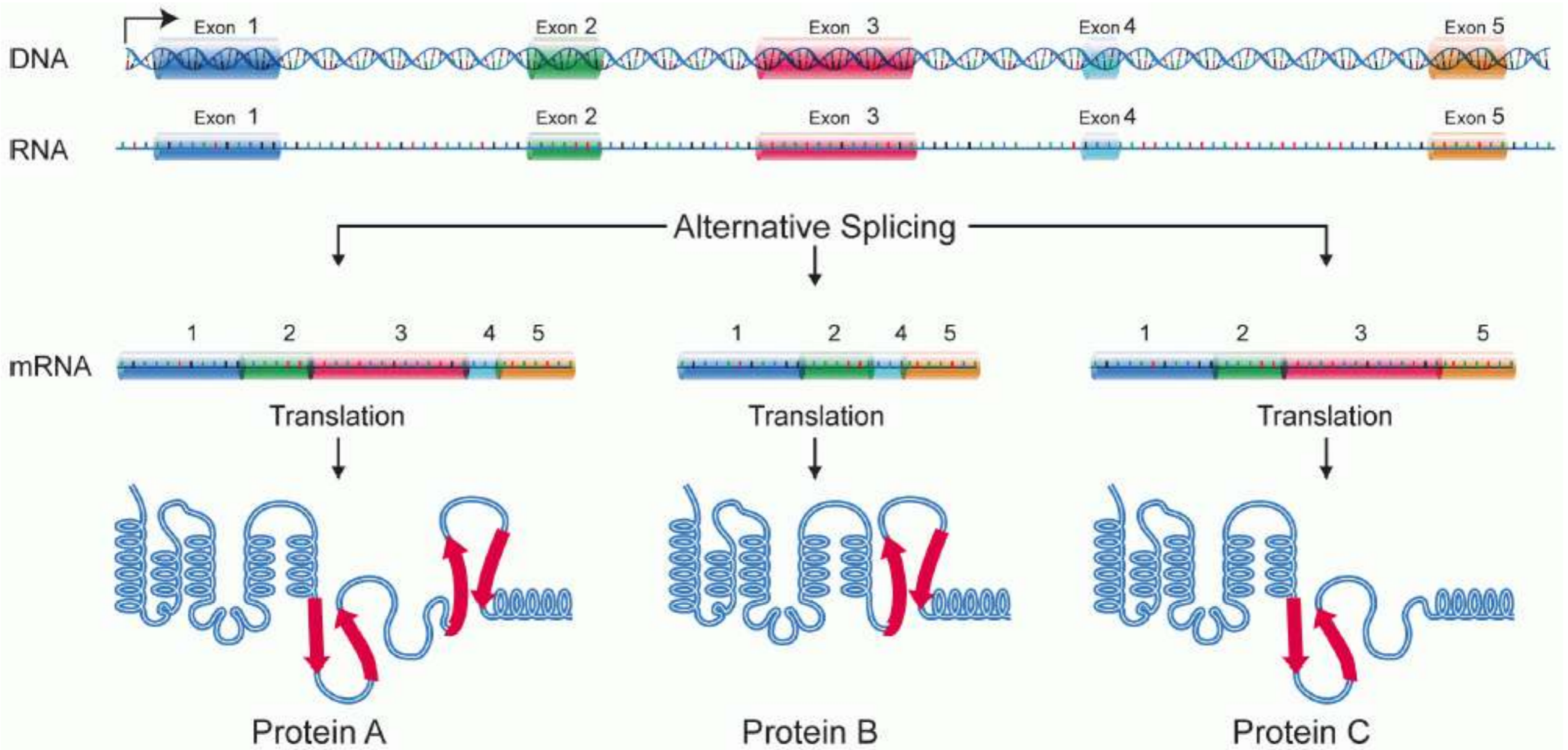
Genes and transcripts

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

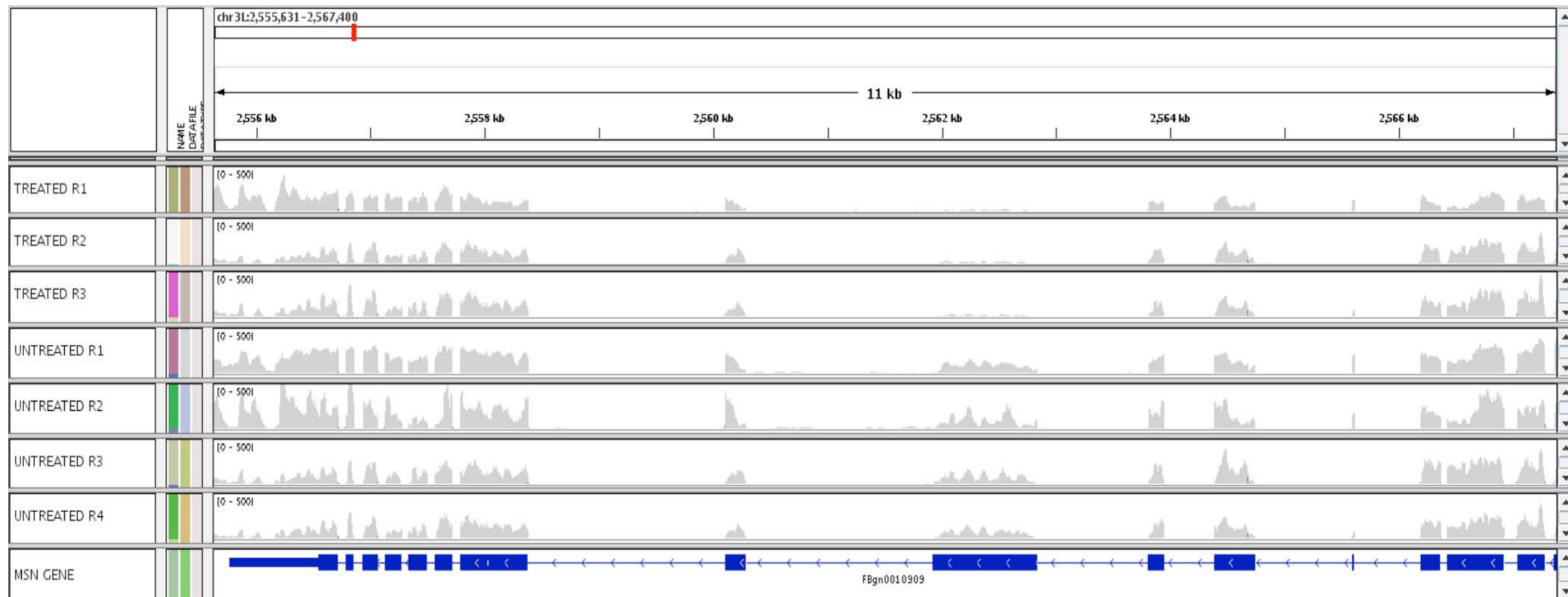
How to look at gene sub-structure?



Alternative isoform regulation



Alejandro Reyes



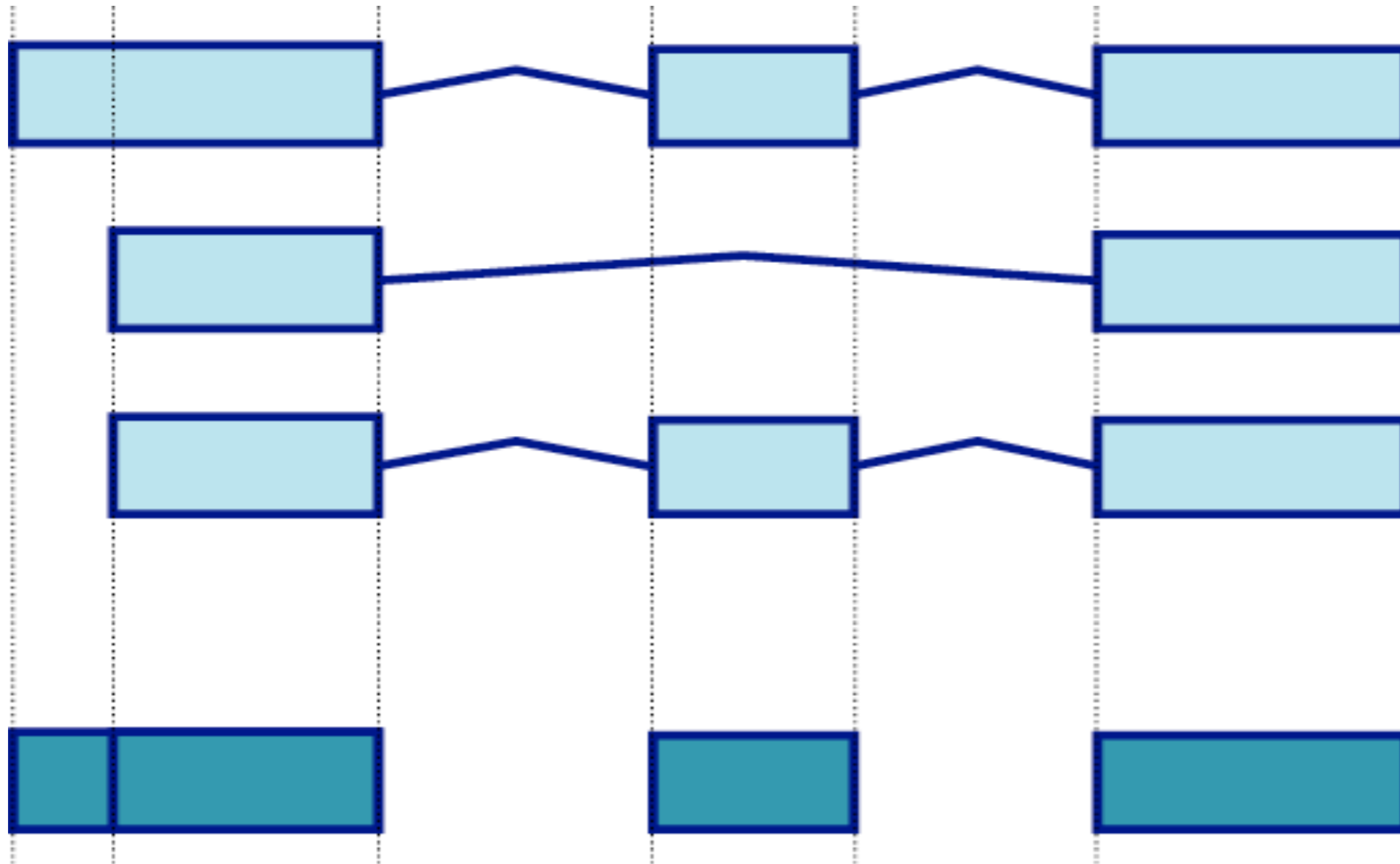
Data: Brooks, ..., Graveley, Genome Res., 2010

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		
[...]						

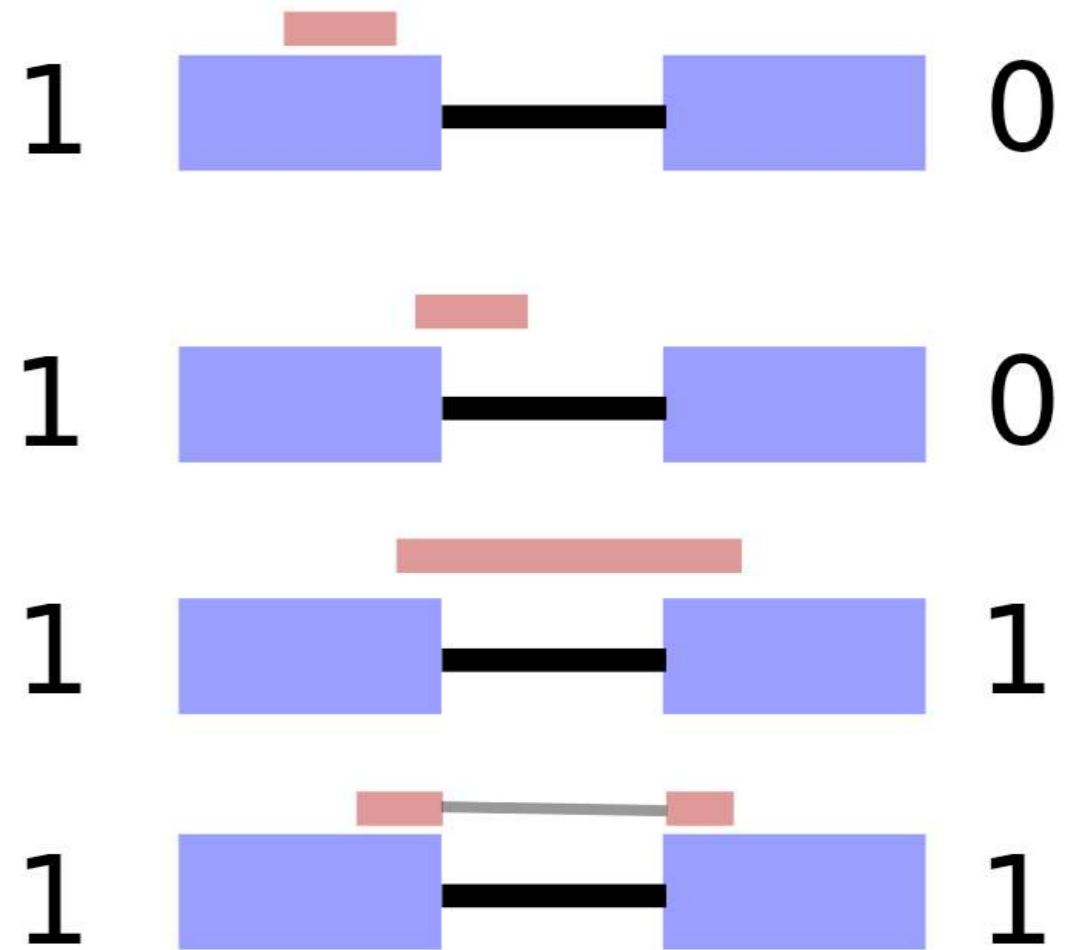
Exon counting bins



Anders, Reyes and Huber.
Genome Research 2012

Counting rules

- Alignment vs genome
- Uniquely aligned reads

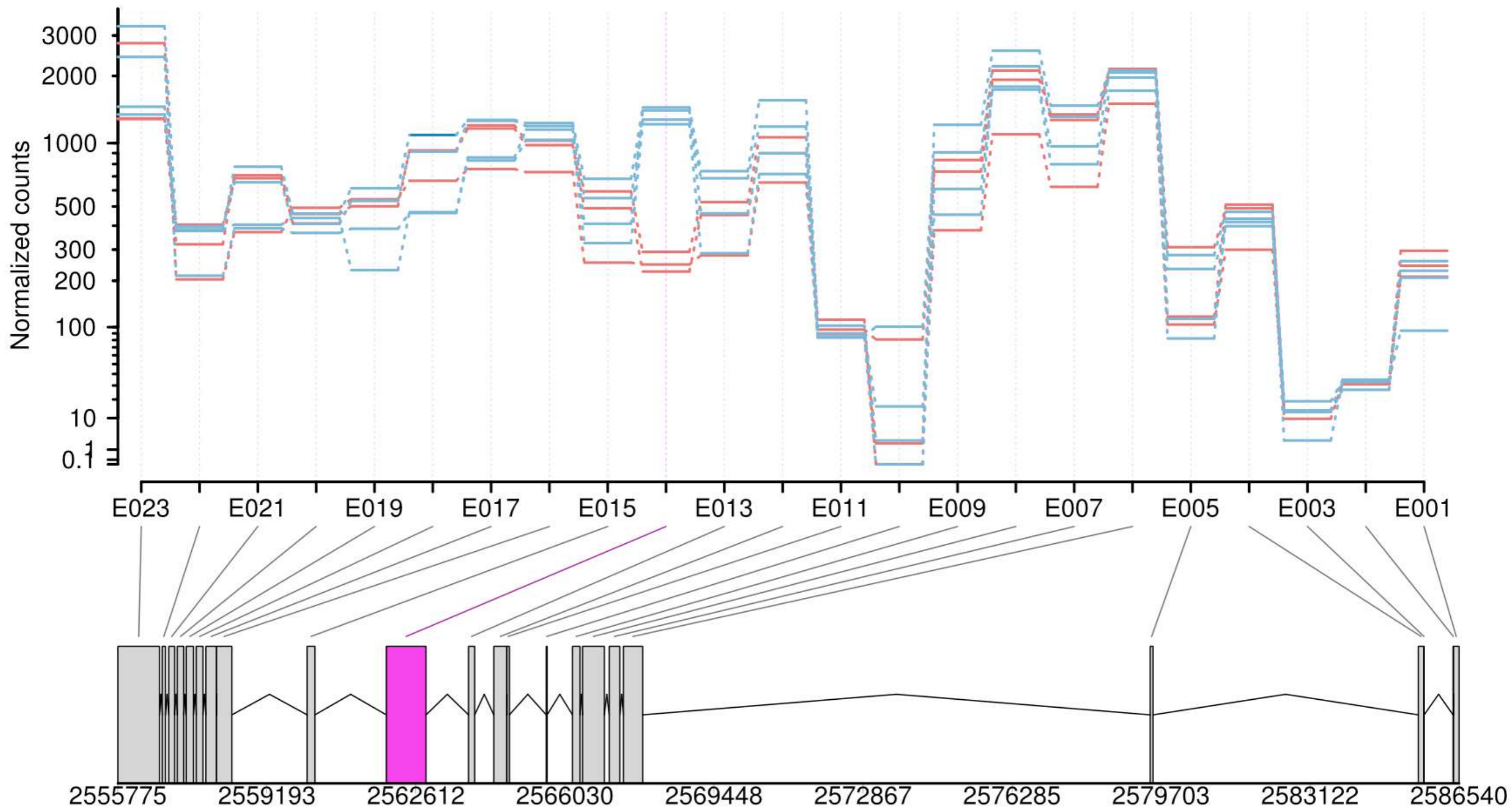


Differential exon usage

msn - mishappen

treated

untreated



DEXSeq

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

number of reads mapping to the exon

number of reads mapping to the other exons of
the same gene

DEXSeq

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

counts in gene i , sample j , exon l

size factor

dispersion

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

expression
strength in control

change in expression
of gene i due to
treatment

fraction of reads
falling onto exon
 l in control

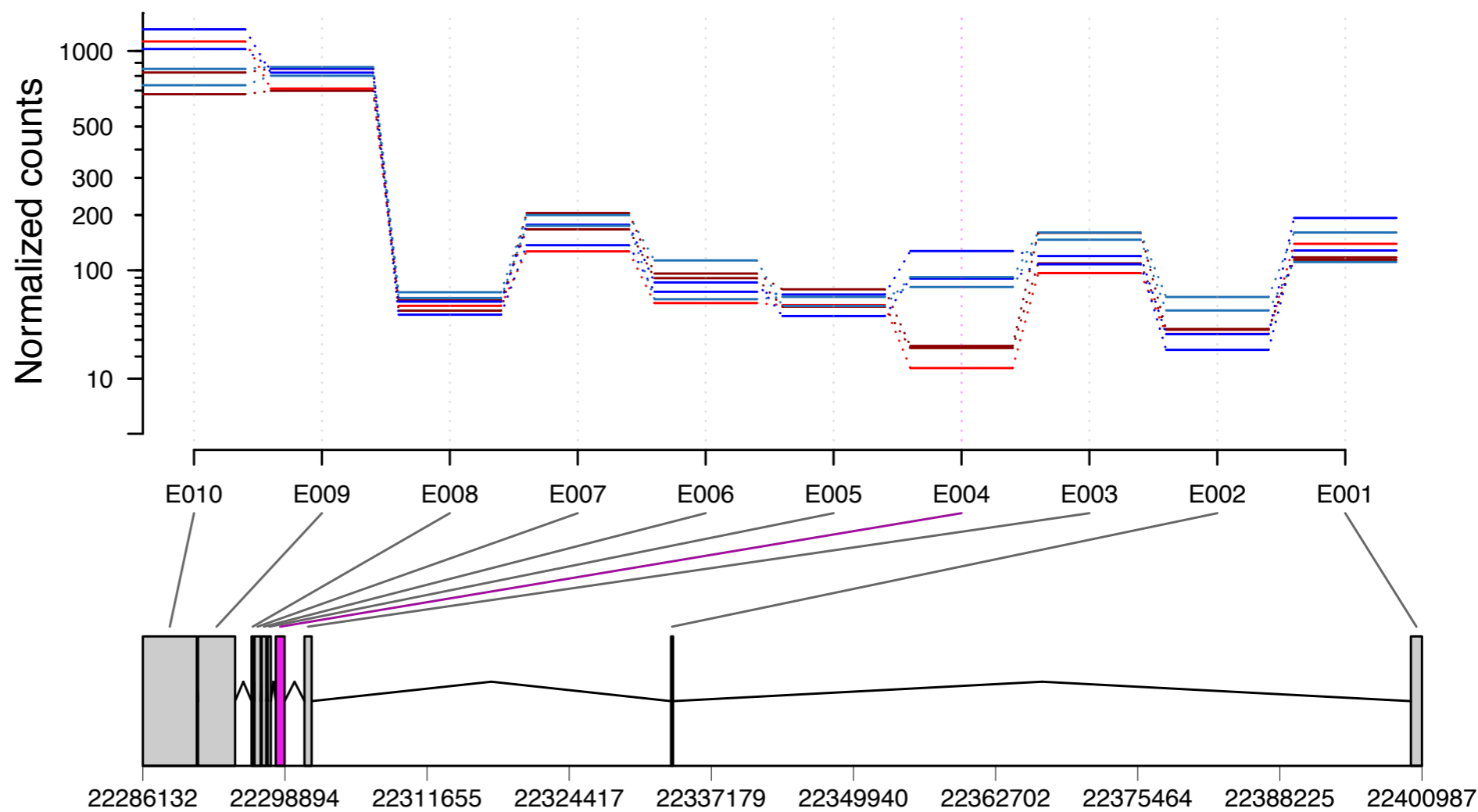
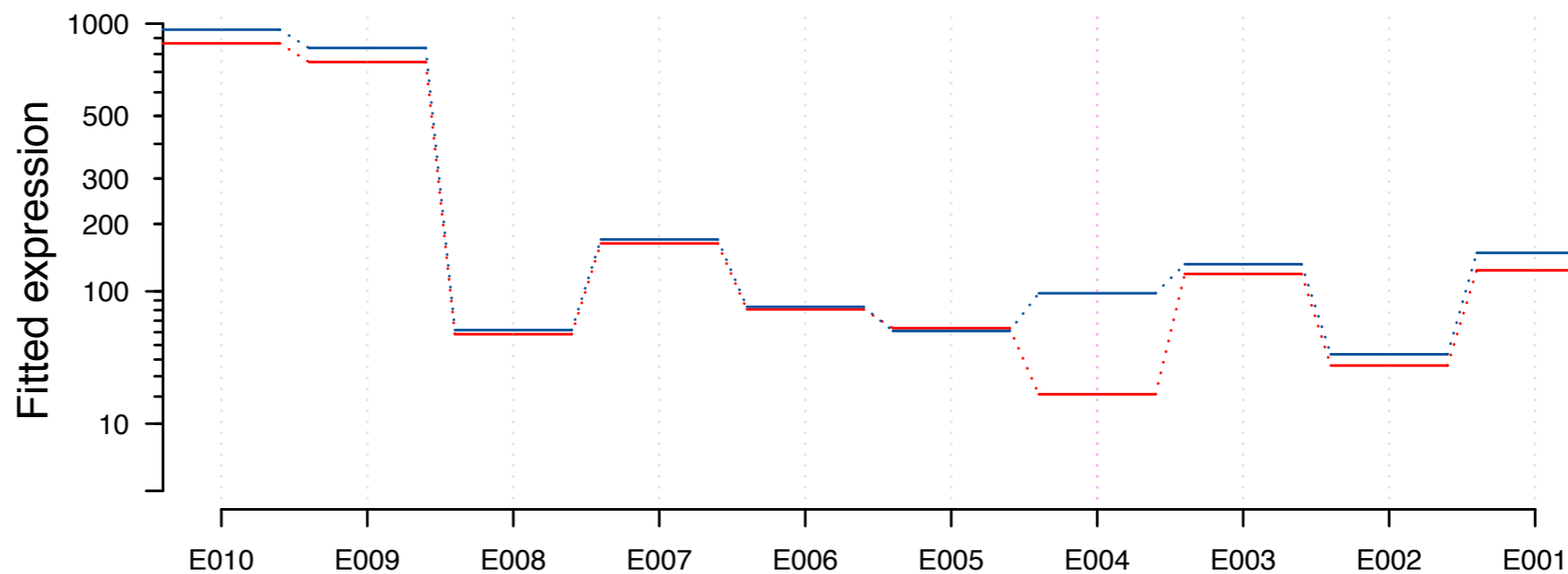
change to fraction
of reads for exon l
due to treatment

Ten-m

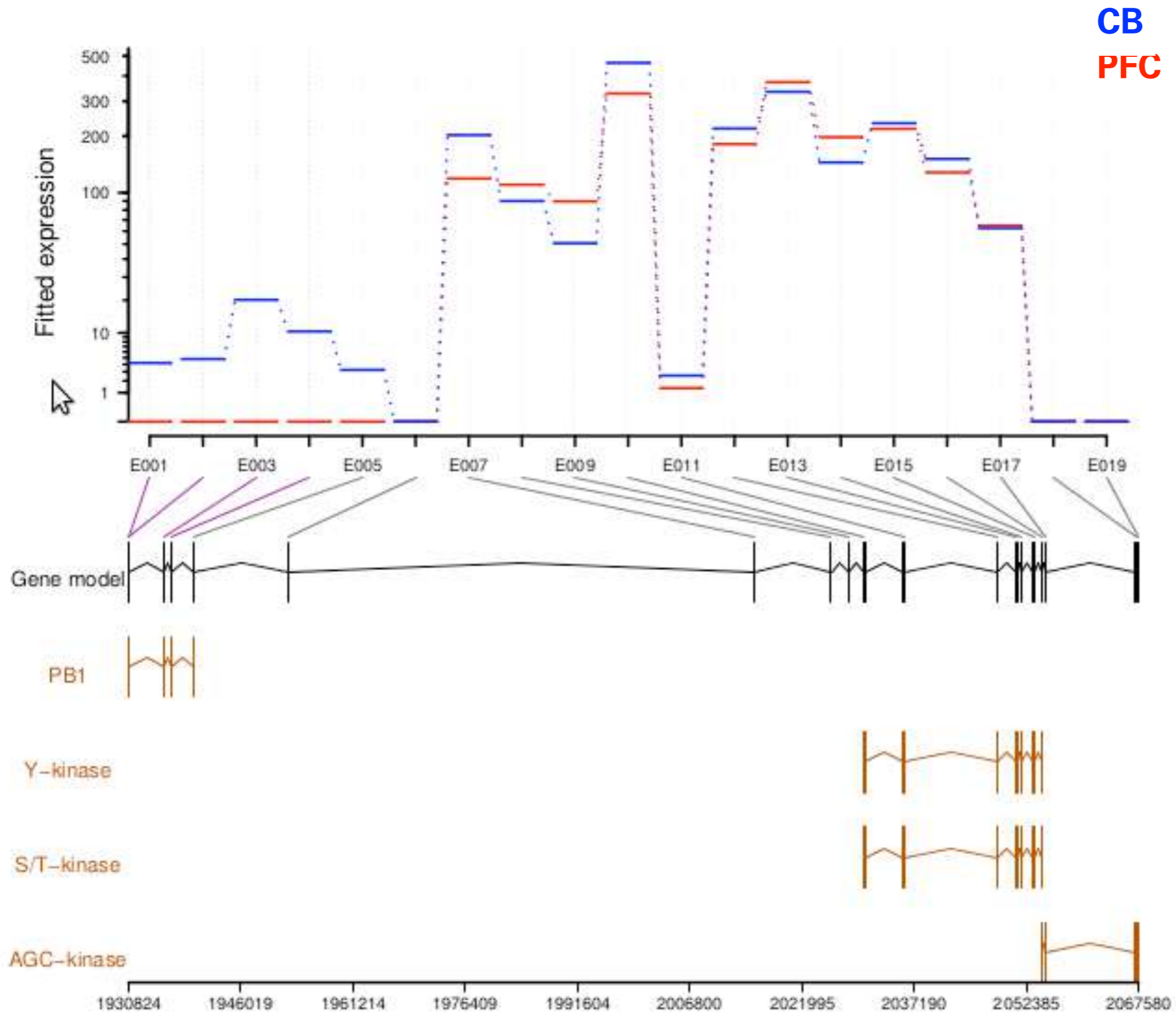
FBgn0004449 –

treated

untreated



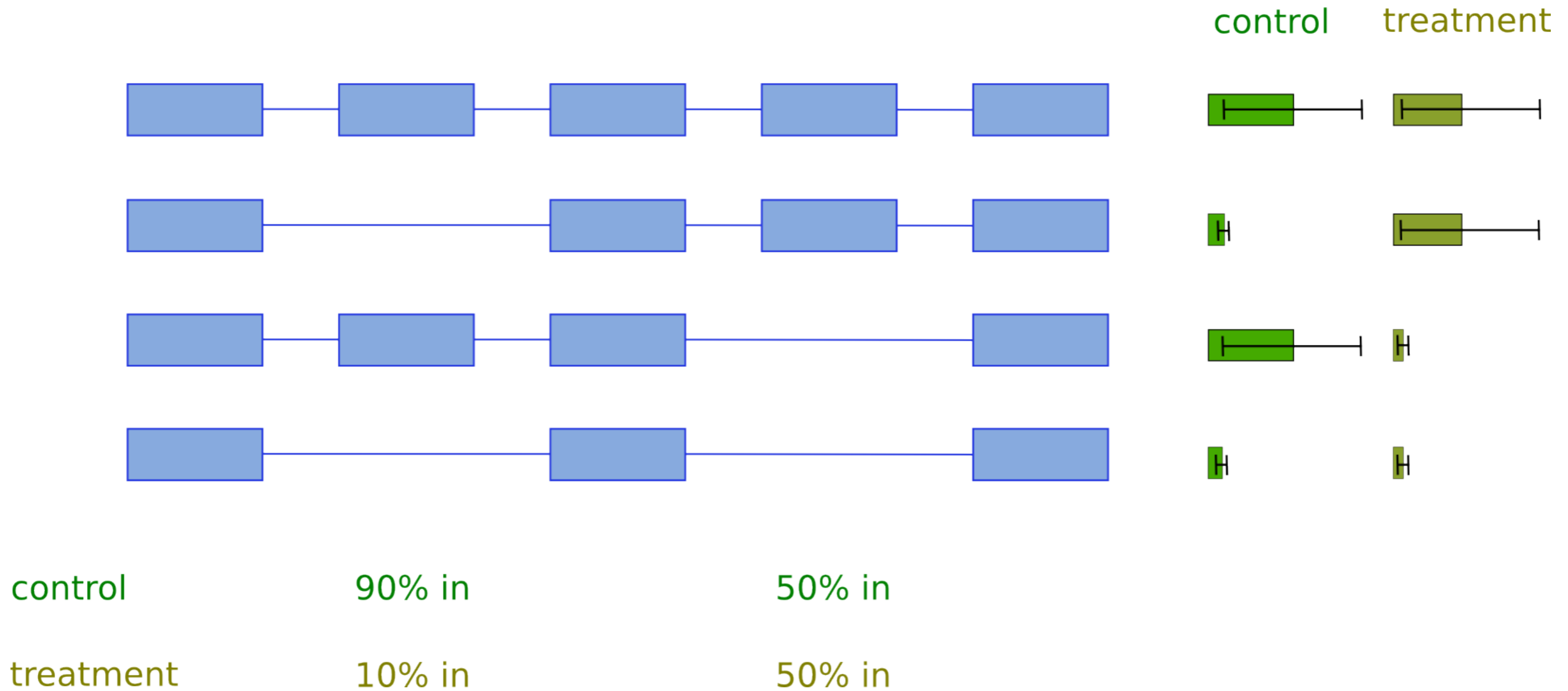
PKC ζ - PKM ζ



long form:
PKC-zeta

N-term.
truncated:
PKM-zeta

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

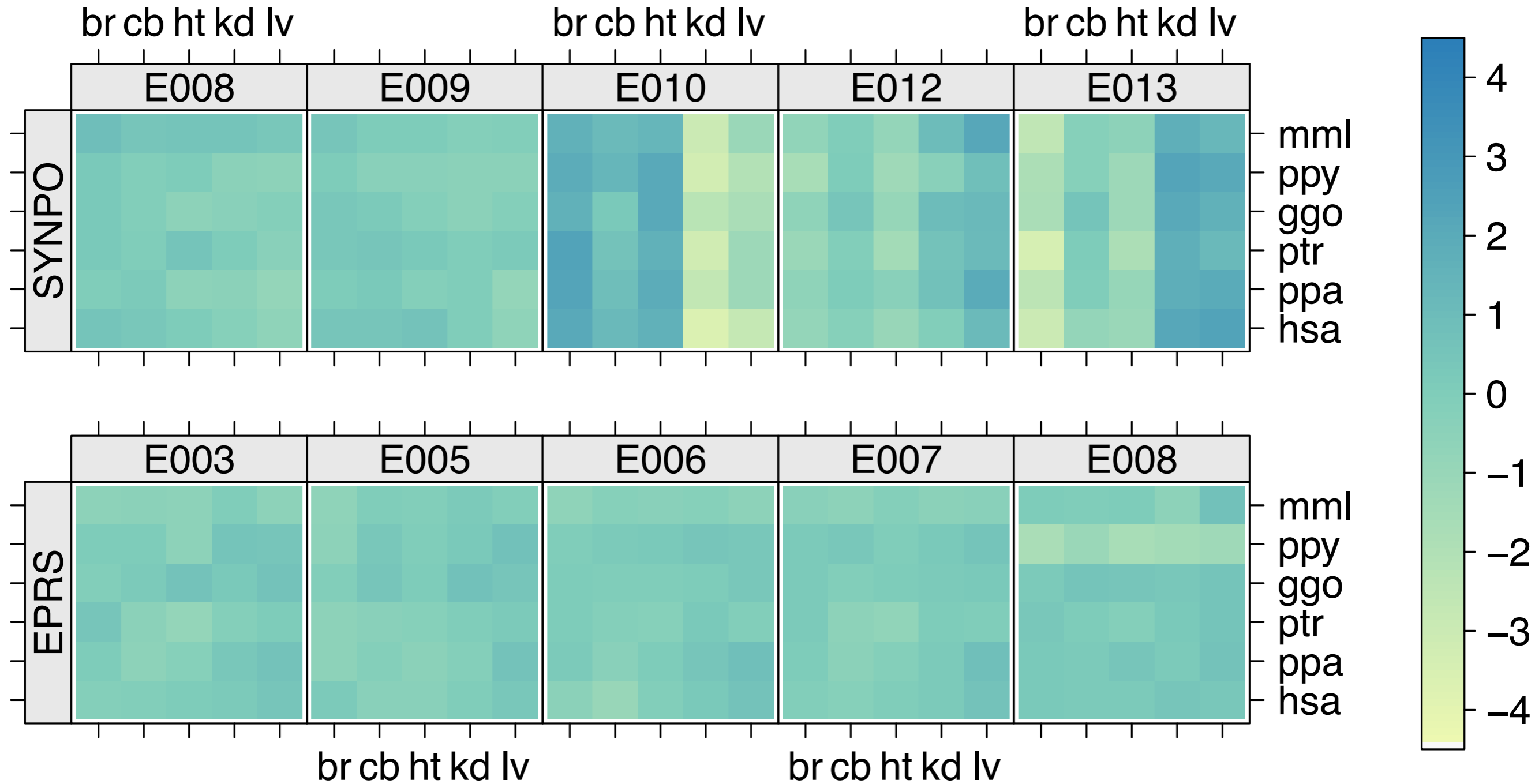


The evolution of gene expression levels in mammalian organs

David Brawand^{1,2*}, Magali Soumillon^{1,2*}, Anamaria Necsulea^{1,2*}, Philippe Julien^{1,2}, Gábor Csárdi^{2,3}, Patrick Harrigan⁴, Manuela Weier¹, Angélica Liechti¹, Ayinuer Aximu-Petri⁵, Martin Kircher⁵, Frank W. Albert^{5†}, Ulrich Zeller⁶, Philipp Khaitovich⁷, Frank Grützner⁸, Sven Bergmann^{2,3}, Rasmus Nielsen^{4,9}, Svante Pääbo⁵ & Henrik Kaessmann^{1,2}

- “Rate of gene expression evolution varies among organs, lineages and chromosomes, owing to differences in selective pressures”
- 9 species, 6 tissues, 2 individuals each: ~139 samples, ~414.145.196 high quality reads
- General goal: explore the functional and evolutionary aspect of the regulation of exon usage

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