Differential expression analysis for sequencing count data

Simon Anders

EMBL

Two applications of RNA-Seq

- Discovery
  - find new transcripts
  - find transcript boundaries
  - find splice junctions

- Comparison

  Given samples from different experimental conditions, find effects of the treatment on
  - gene expression strengths
  - isoform abundance ratios, splice patterns, transcript boundaries

EMBL

Count data in HTS

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

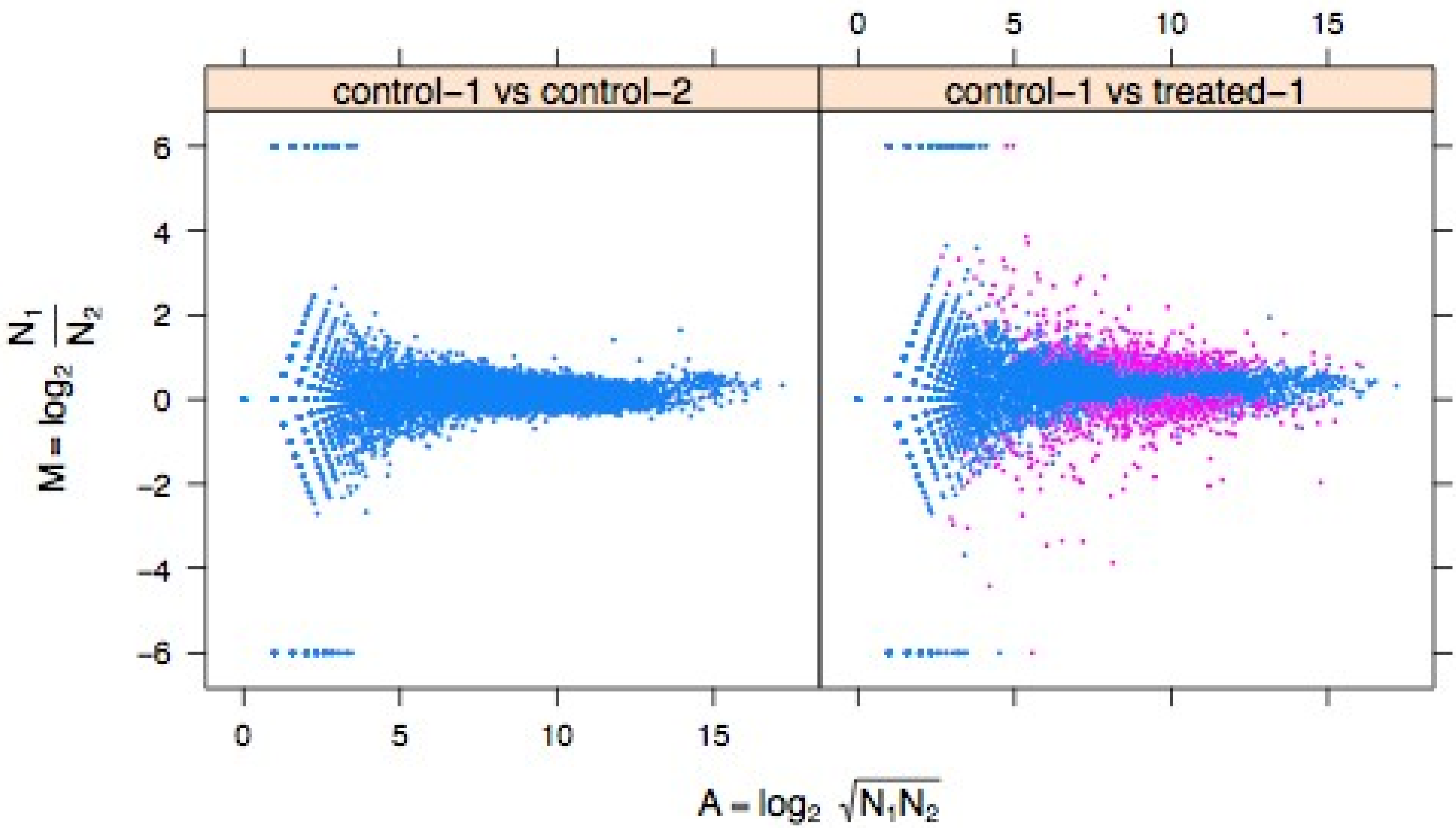
- RNA-Seq
- Tag-Seq
- ChIP-Seq
- HiC
- Bar-Seq
- ...

EMBL

Sample-to-sample variation

comparison of two replicates

comparison of treatment vs control



Sample-to-sample variability

- In RNA-Seq, the minimum variance given by the Poisson distribution.
- Taking only Poisson noise into account is insufficient, though.
- Many publications ignore this.

Differential expression: Two questions

Assume you use RNA-Seq to determine the concentration of transcripts from some gene in different samples. What is your question?

1. "Is the concentration in one sample different from the expression in another sample?"

or

2. "Can the difference in concentration between treated samples and control samples be attributed to the treatment?"

EMBL

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

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

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

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

Replicates

Two replicates permit to

- globally estimate variation

Sufficiently many replicates permit to

- estimate variation for each gene
- randomize out unknown covariates
- spot outliers
- improve precision of expression and fold-change estimates

Replication at what level?

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

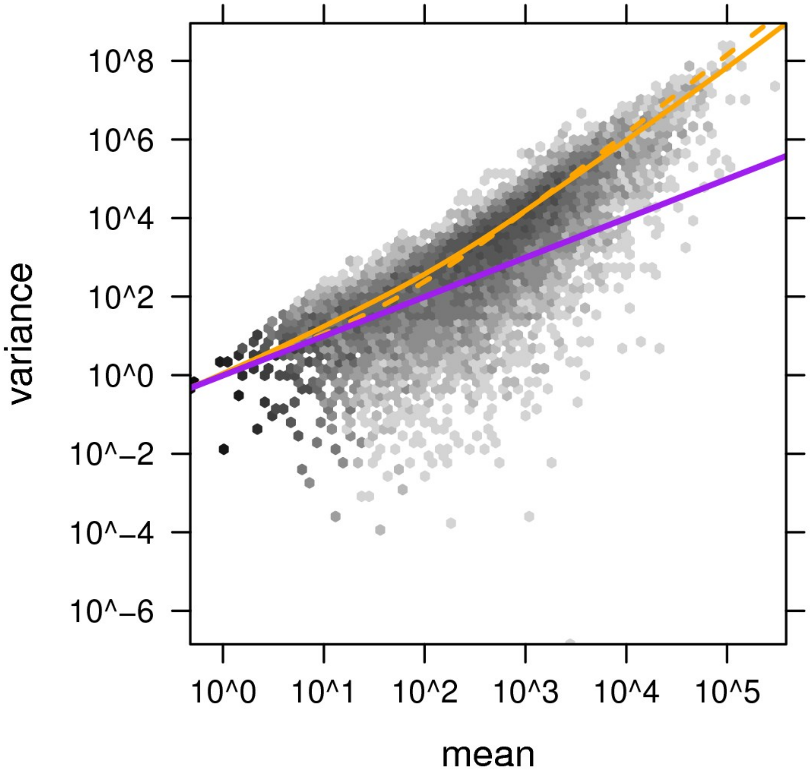
Estimating noise from the data

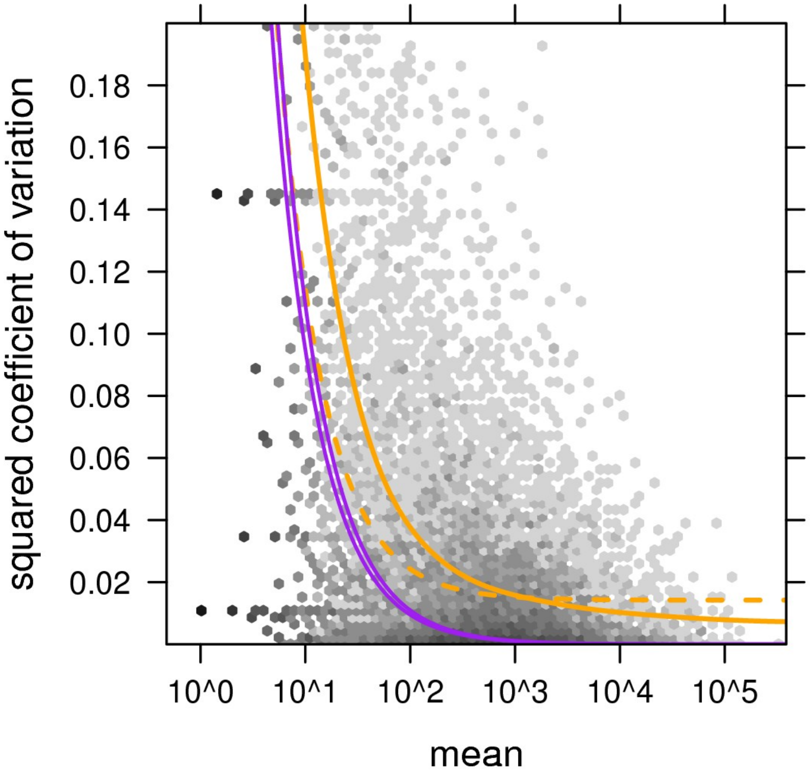
- If we have many replicates, we can estimate the variance for each gene.

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

EMBL

Variance depends strongly on the mean



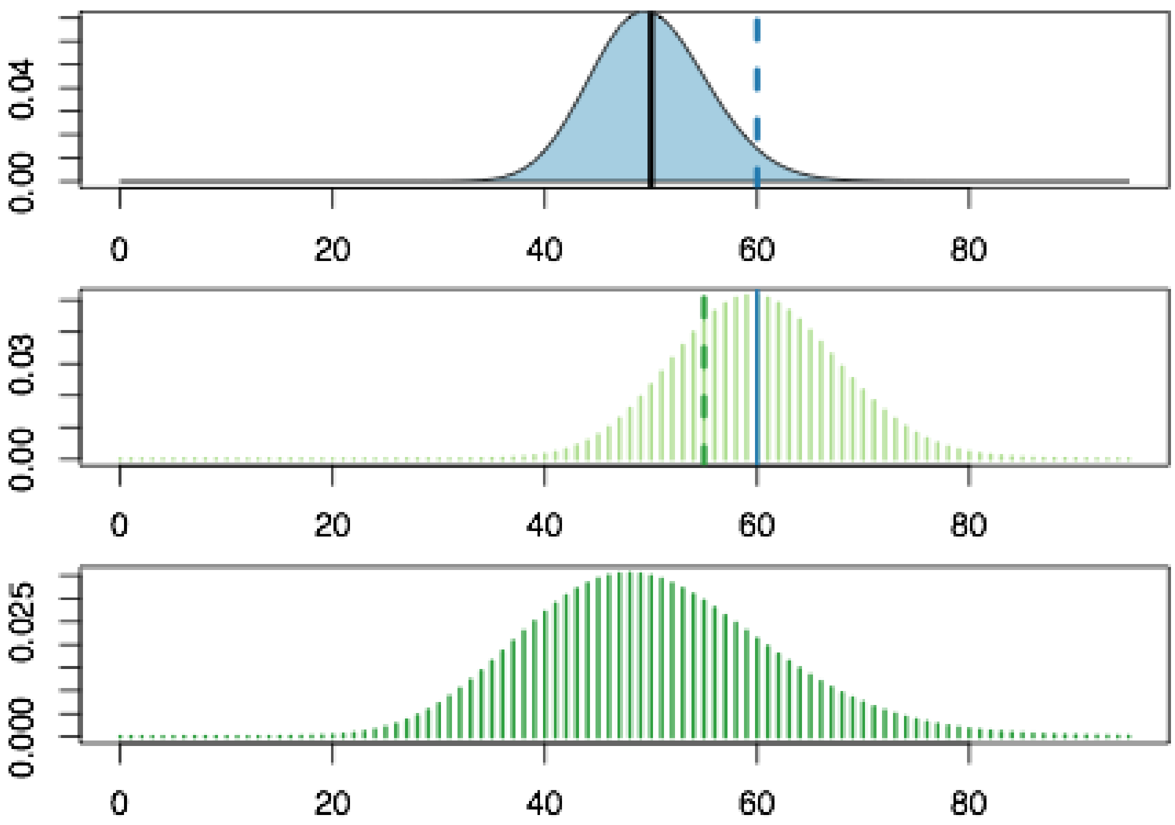


Variance calculated from comparing two replicates

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

EMBL

The NB distribution from a hierarchical model



Biological sample with mean μ and variance v

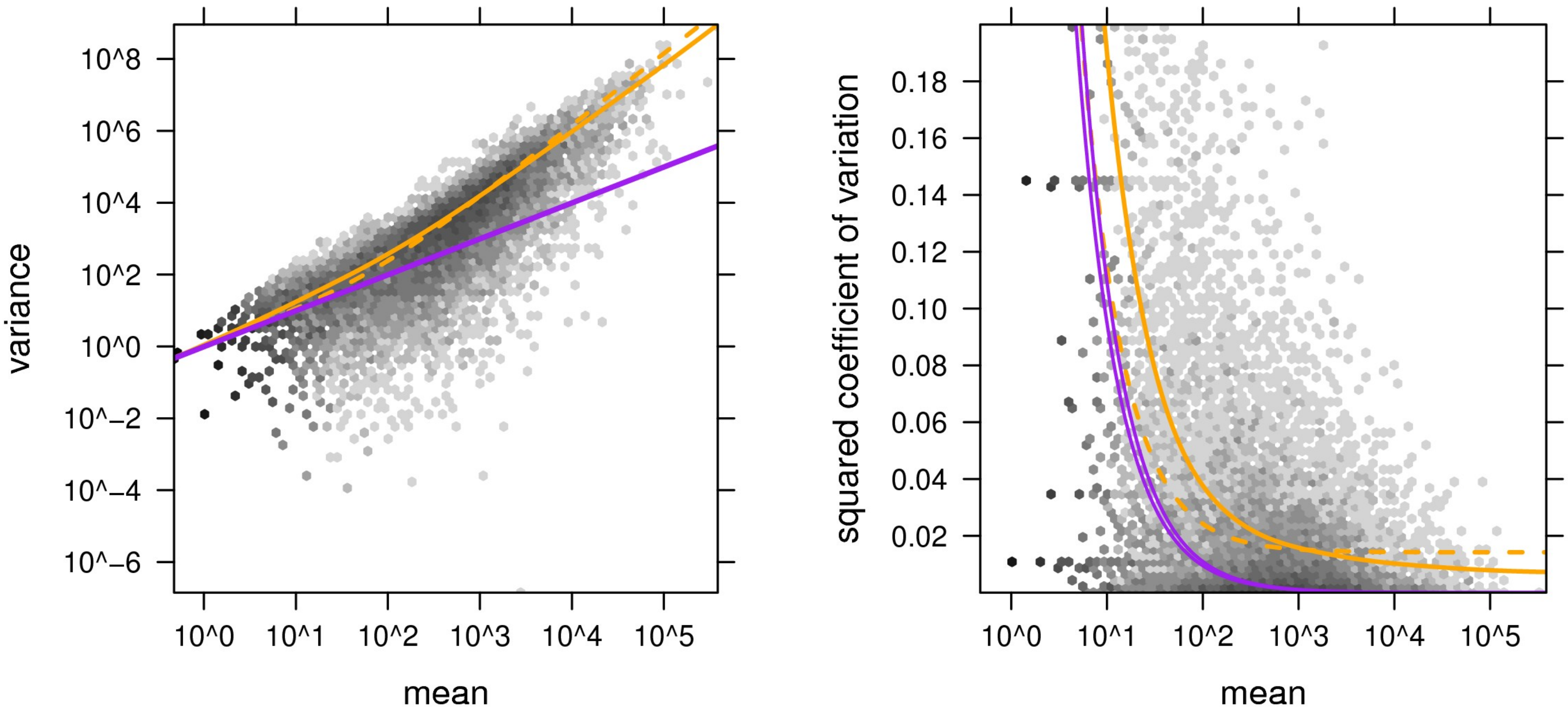
Poisson distribution with mean q and variance q.

Negative binomial with mean μ and variance $q+v$.

EMBL

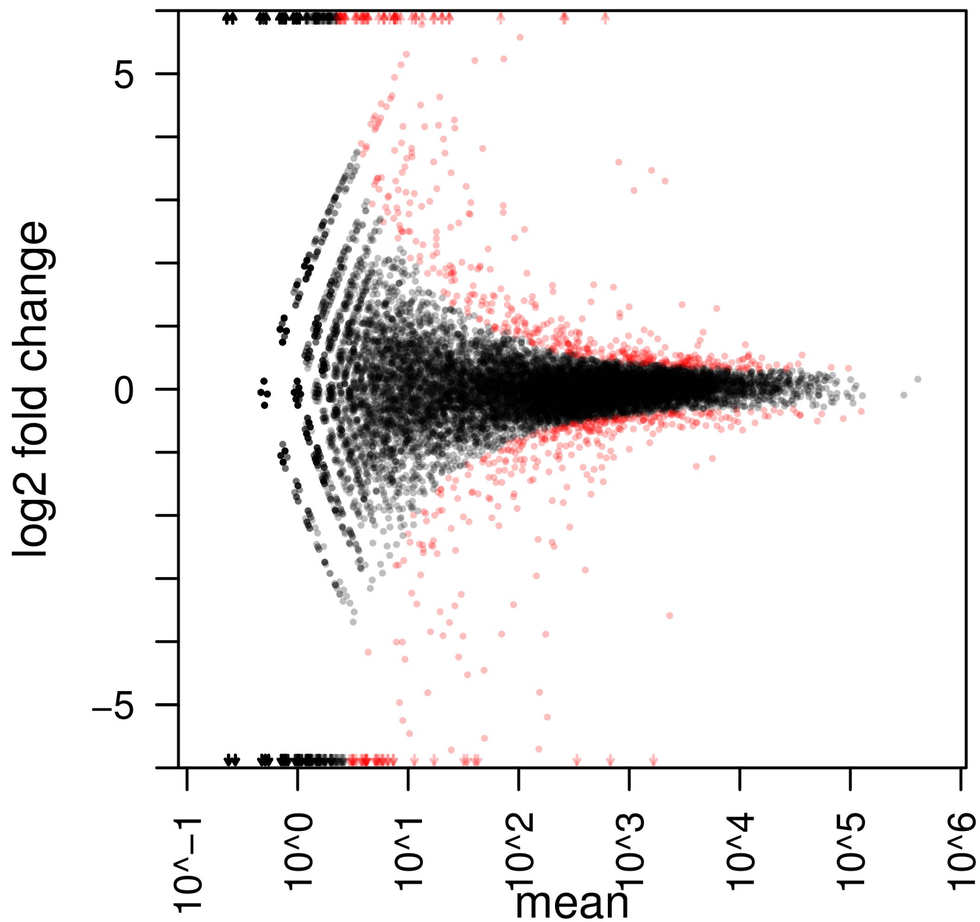
Model fitting

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



Dispersion fit

Differential expression

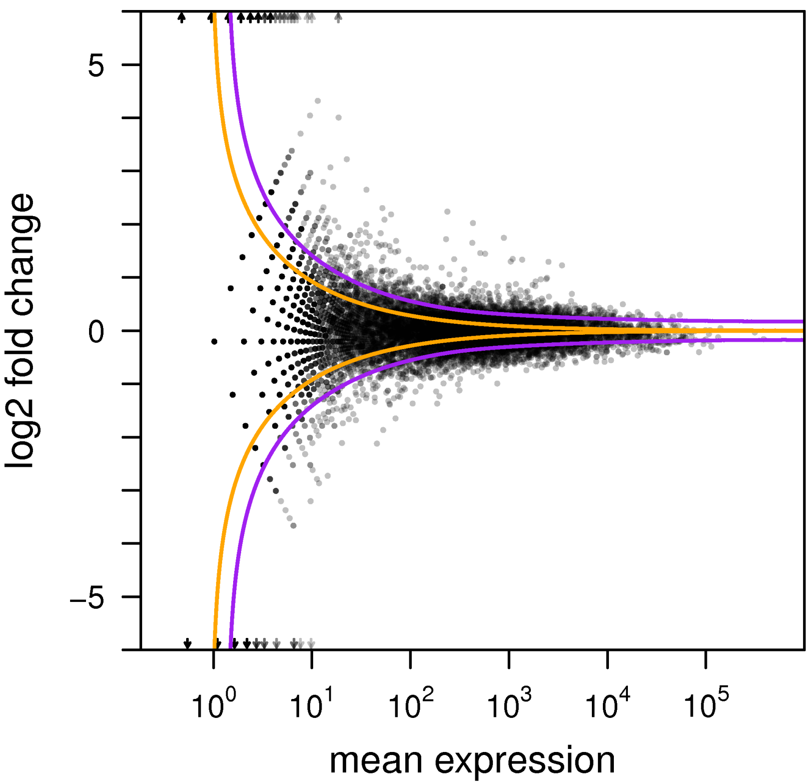


RNA-Seq data: overexpression of two different genes in flies [data: Furlong group]

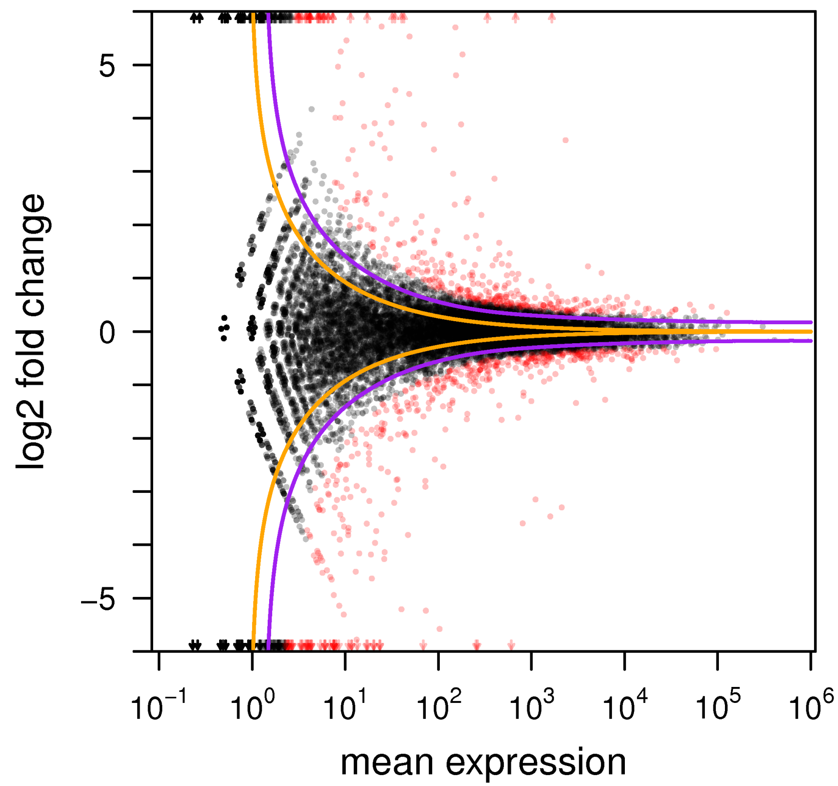
EMBL

Type-I error control

comparison of two replicates

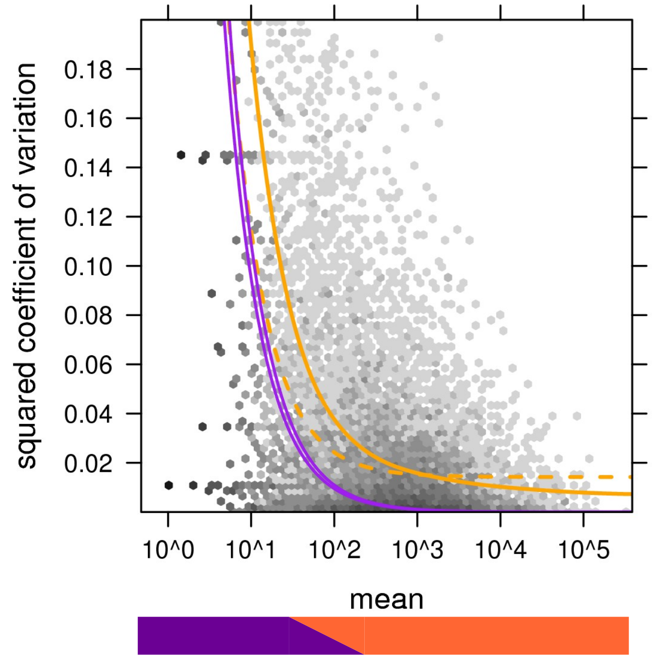


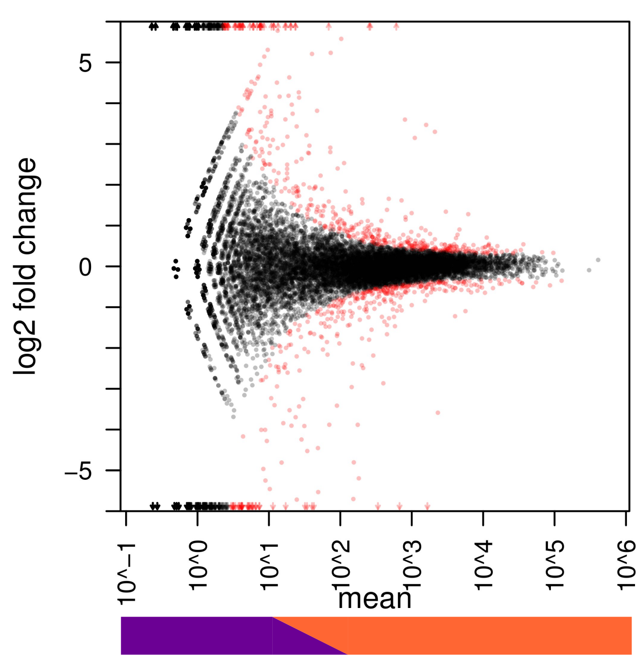
comparison of treatment vs control



EMBL

Two noise ranges





dominating noise

shot noise (Poisson)

biological noise

How to improve power?

deeper sampling

more biological replicates

EMBL

Further use cases

Similar count data appears in

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

and can be analysed with *DESeq* as well.

EMBL

Comparative ChIP-Seq with DESeq

Step 1: Get a list of counting bins by either

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

Step 2: Make a count table:

columns – samples; rows – counting bins

and use DESeq

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

EMBL

Generalized linear models

Simple design:

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

Common complex designs:

- Designs with blocking factors
- Factorial designs

EMBL

GLMs: Blocking factor

Sample	treated	sex
S1	no	male
S2	no	male
S3	no	male
S4	no	female
S5	no	female
S6	yes	male
S7	yes	male
S8	yes	female
S9	yes	female
S10	yes	female

EMBL

GLMs: Blocking factor

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

full model for gene *i*:

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

reduced model for gene *i*:

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

EMBL

GLMs: Blocking factor

```
cds <- newCountDataset( countTable, designTable )

cds <- estimateSizeFactors( cds )
cds <- estimateDispersions( cds, method="pooled-CR" )

fit0 <- fitNbinomGLMs( cds, count ~ sex )
fit1 <- fitNbinomGLMs( cds, count ~ sex + treatment )

pvals <- nbinomGLMTest( fit1, fit0 )
```

Dispersion estimation: Cox, Reid: J Roy Stat Soc B, 1987
McCarthy, Chen, Smyth: Nucl Acid Res, 2012

EMBL

GLMs: Interaction

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

full model for gene *i*:

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

reduced model for gene *i*:

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

EMBL

GLMs: paired designs

- Often, samples are paired (e.g., a tumour and a healthy-tissue sample from the same patient)
- Then, using pair identity as blocking factor improves power.

full model:

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

reduced model:

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

i	gene
j	subject
l	tissue state

EMBL

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.
- Proper inference has to take thin into account, and sample-to-sample variability

EMBL

Data set used for to demonstrate DEXSeq:

Genome Research

Research 21:193–202 © 2011

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

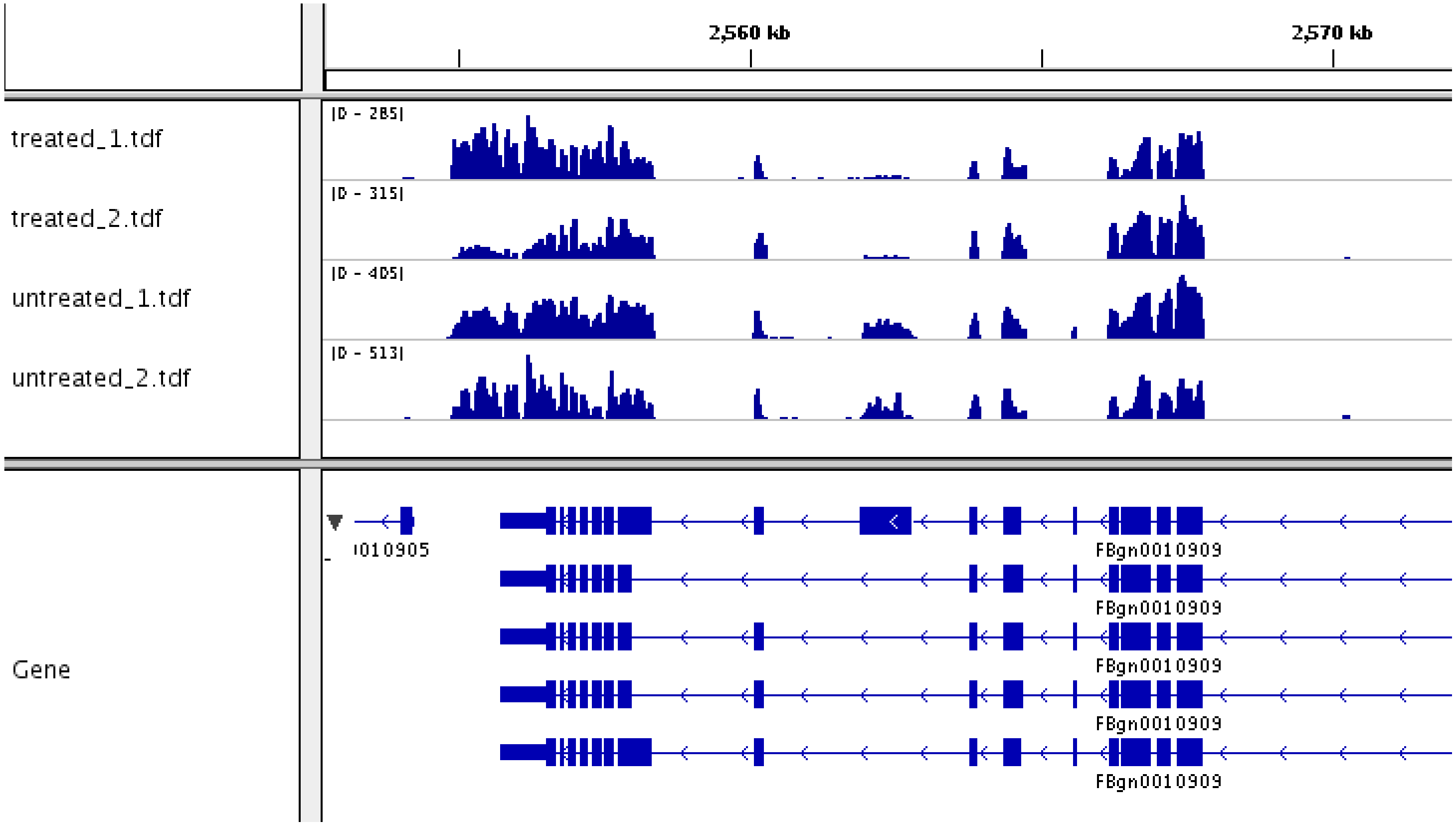
Angela N. Brooks,[1,7] Li Yang,[2,7] Michael O. Duff,[2,3] Kasper D. Hansen,[4] Jung W. Park,[2,3] Sandrine Dudoit,[4,5] Steven E. Brenner,[1,6,8] and Brenton R. Graveley[2,3,8]

Drosophila melanogaster S2 cell cultures:

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

EMBL

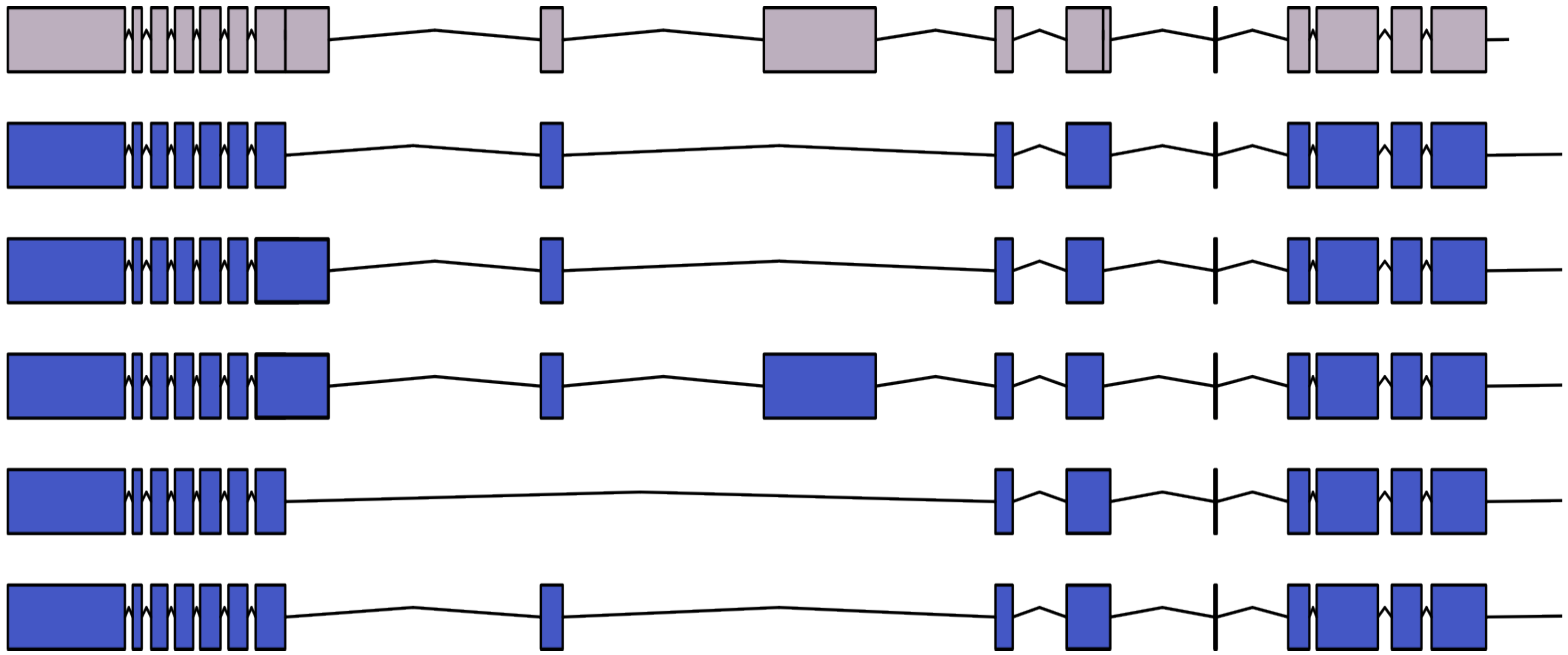
Alternative isoform regulation



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

EMBL

Exon counting bins



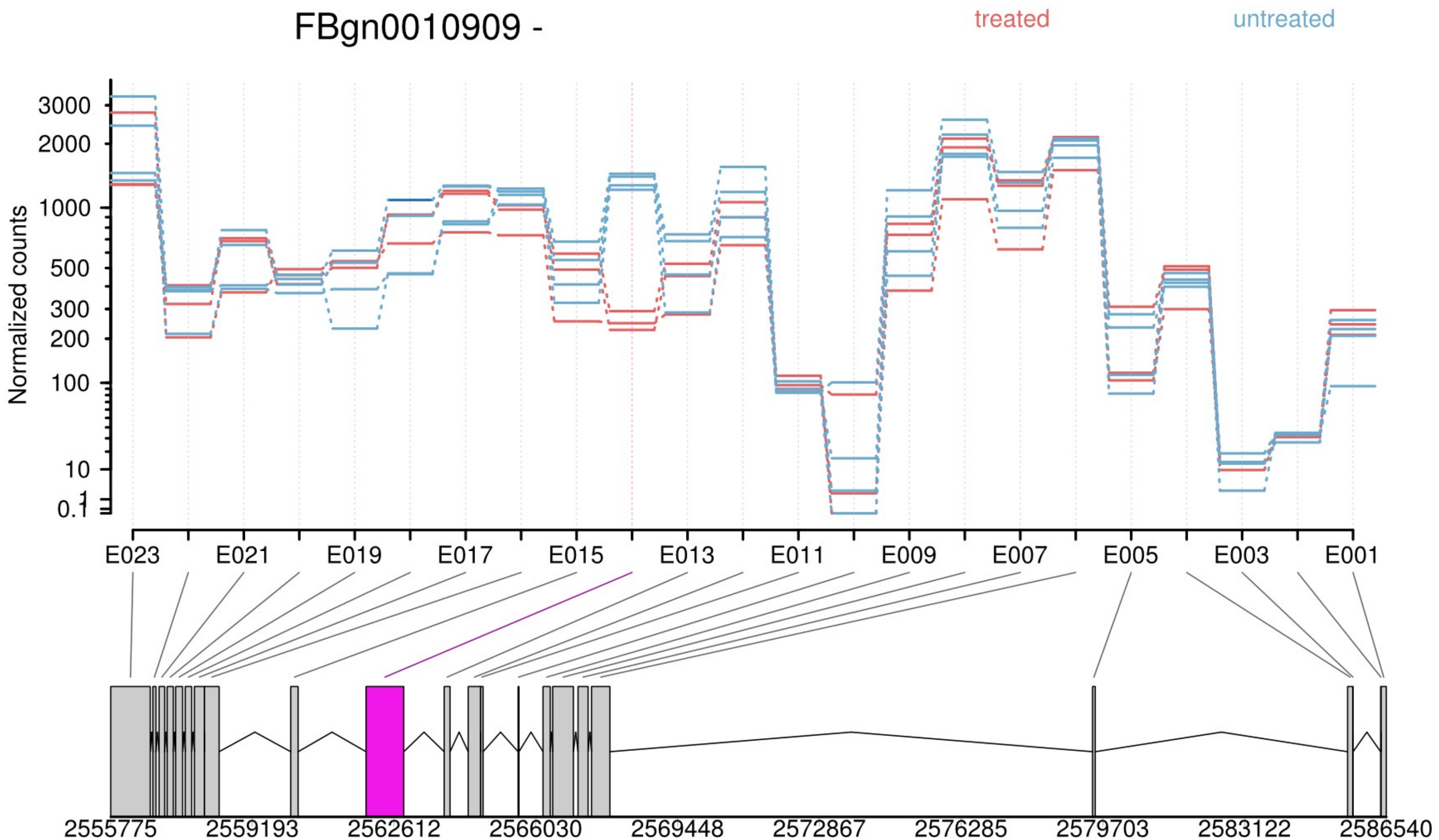
EMBL

Count table for a gene

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

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

[...]


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

Model

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

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

size factor

dispersion

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

expression strength in control

fraction of reads falling onto exon *l* in control

change in expression due to treatment

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

EMBL

Model, refined

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

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

size factor

dispersion

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

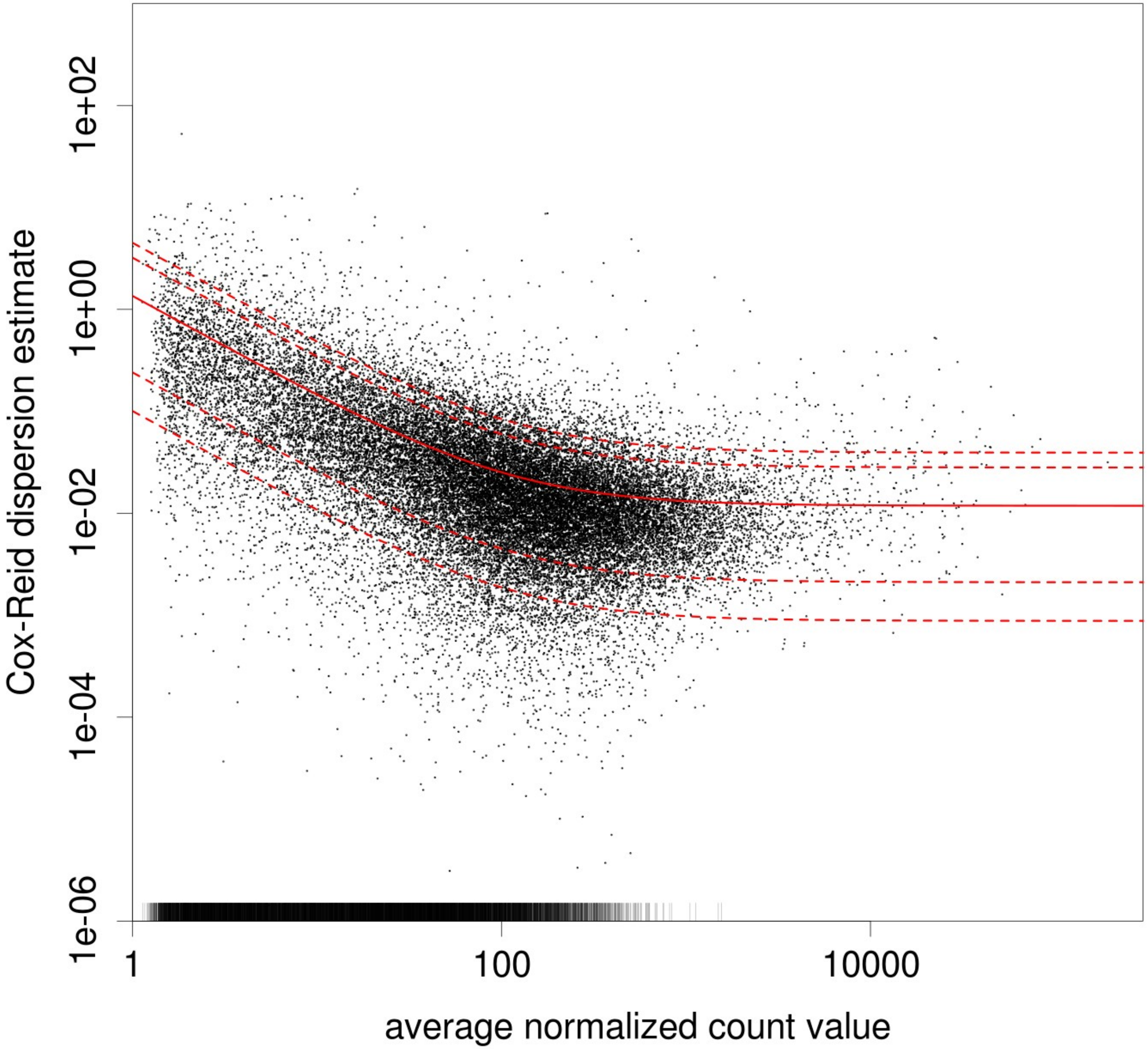
expression strength in sample *j*

fraction of reads falling onto exon *l* in control

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

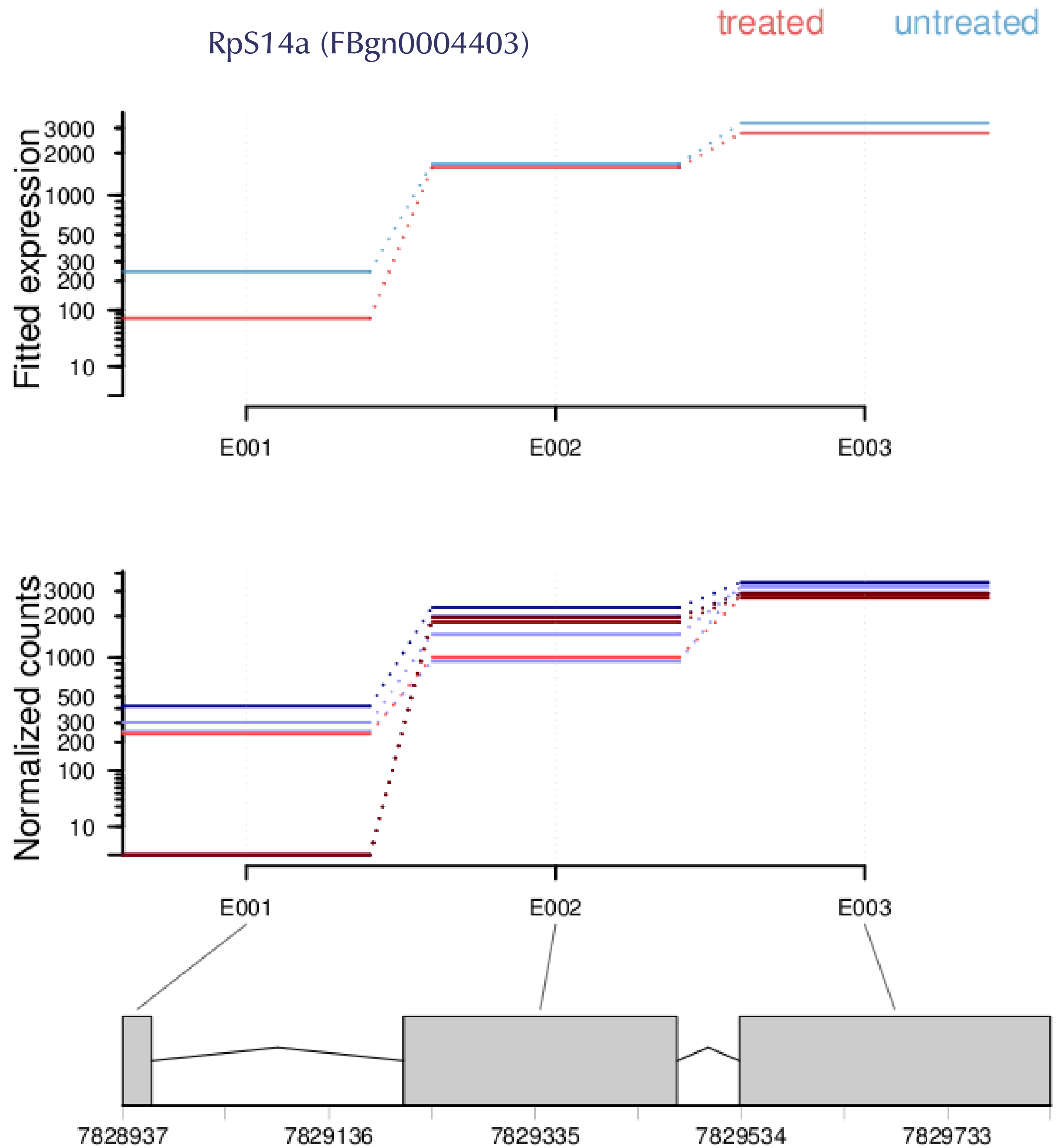
further refinement: fit an extra factor for library type (paired-end vs single)

Dispersion vs mean



EMBL

RpS14a (FBgn0004403)



EMBL

DEXSeq and other tools

- *MISO* and *ALEXA-Seq* do not account for biological variability.
- Neither does *cuffdiff,* as described in the authors' publications.
- New versions of *cuffdiff* claim to account for biological variability, however ...

- See also Glaus et al.'s EBSeq, though.

EMBL

DEXSeq and other tools

- *MISO* and *ALEXA-Seq* do not account for biological variability.
- Neither does *cuffdiff*, as described in the authors' publications.
- New versions of *cuffdiff* claim to account for biological variability, however …

- See also Glaus et al.'s BitSeq, though.

EMBL

Test *cuffdiff* vs *DEXSeq*

Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.1.0	cuffdiff 1.2.0	cuffdiff 1.3.0
proper comparisons, treatment (knock-down) vs control:					
T1 – T3	C1 – C4	159	145	69	50
T1, T2	C2, C3	52	323	120	578
mock comparisons, control vs control:					
C1, C3	C2, C4	8	314	650	639
C1, C4	C2, C3	7	392	724	728

Table S1: Results of the comparison for the Brooks et al. data.

EMBL

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

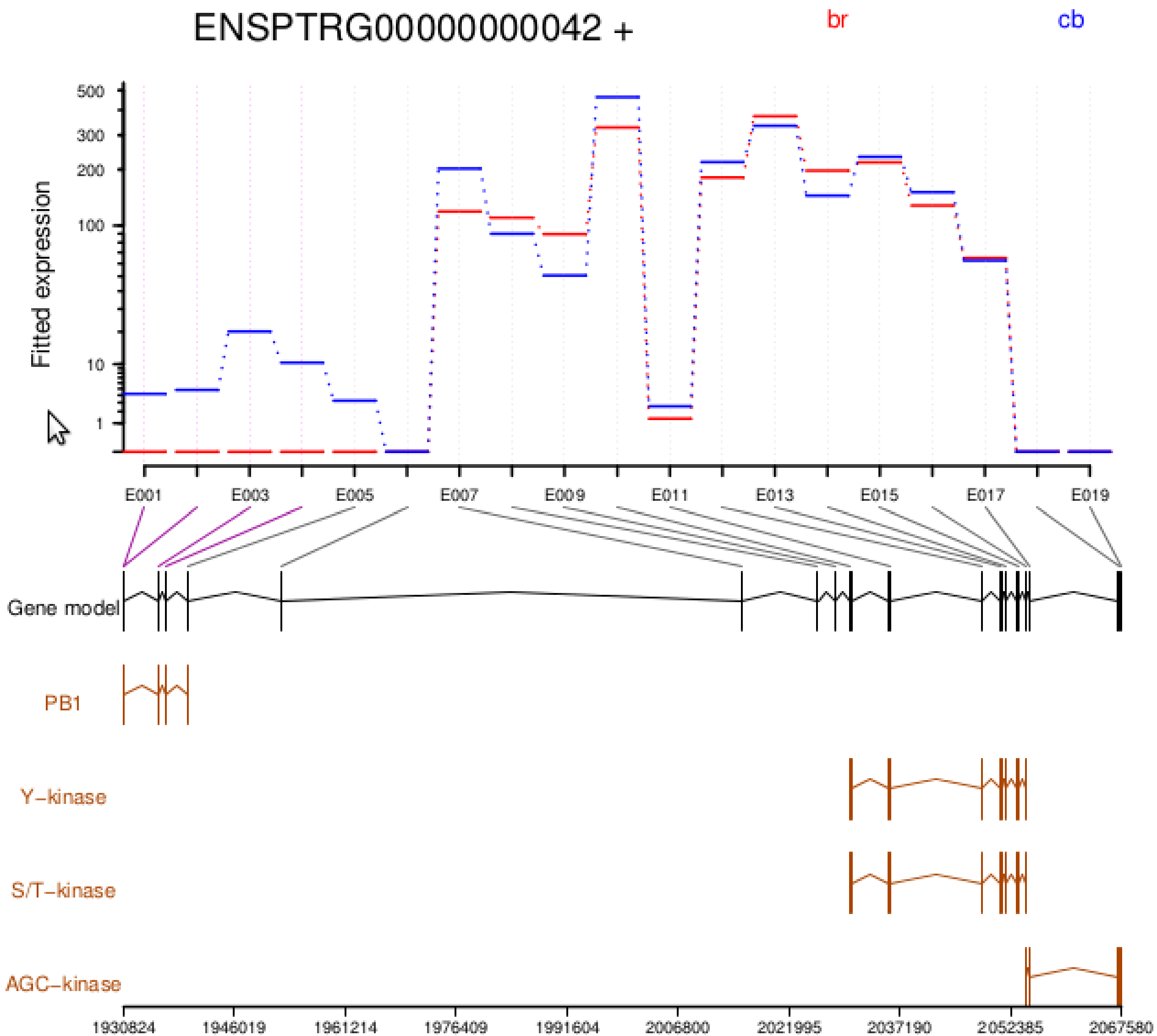
Table S2: Results of the comparison for the Brawand et al. data.

EMBL

Exons vs isoforms

- DEXSeq deliberately tests at the level of exons, not isoforms.
- This might be an advantage: We have more annotation on exons than on isoforms, anyway.

EMBL



ENSPTRG00000000042 +
br
cb
Fitted expression
500
300
200
100
10
1
E001
E003
E005
E007
E009
E011
E013
E015
E017
E019
Gene model
PB1
Y−kinase
S/T−kinase
AGC−kinase
1930824
1946019
1961214
1976409
1991604
2006800
2021995
2037190
2052385
2067580

EMBL

DEXSeq

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

Conclusion

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

- DEXSeq is availabe from Bioconductor, paper is published in Genome Research.

EMBL

Outlook: Current developments

Use of shrinkage estimators (empirical Bayes) for

- dispersion
- fold changes / GLM coefficients

Improvements to DEXSeq

- “splice graphs”
- junction reads

EMBL

Acknowledgements

Coauthors:

- Alejandro Reyes
- Wolfgang Huber
- Michael Love

Funding:

- EMBL

EMBL