Studying isoform regulation using RNA-seq data

Alejandro Reyes

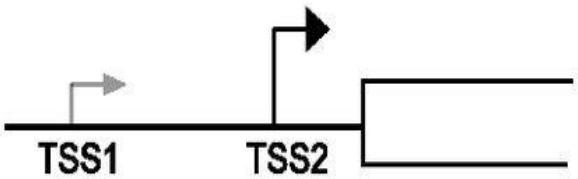
CSAMA 2014, Brixen

26.06.2014

OUTLINE

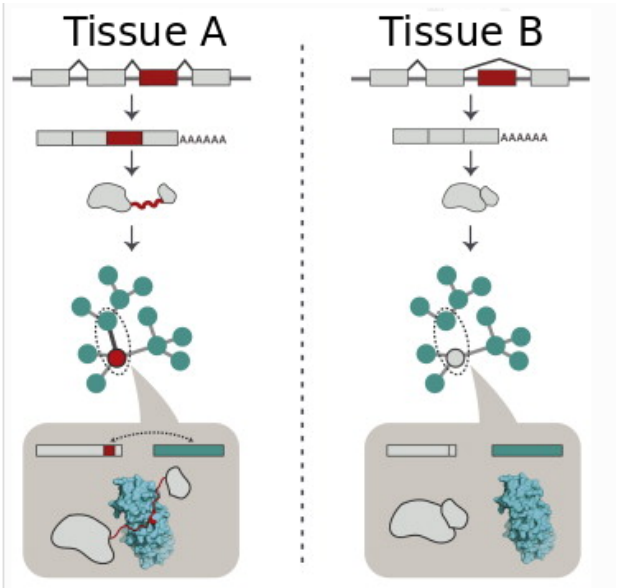
Differential usage of exons generates complexity

Alternative initiation



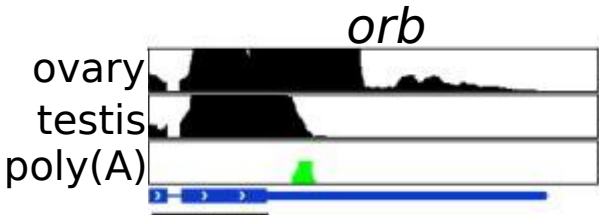
Rojas-Duran et al, 2012

Alternative splicing



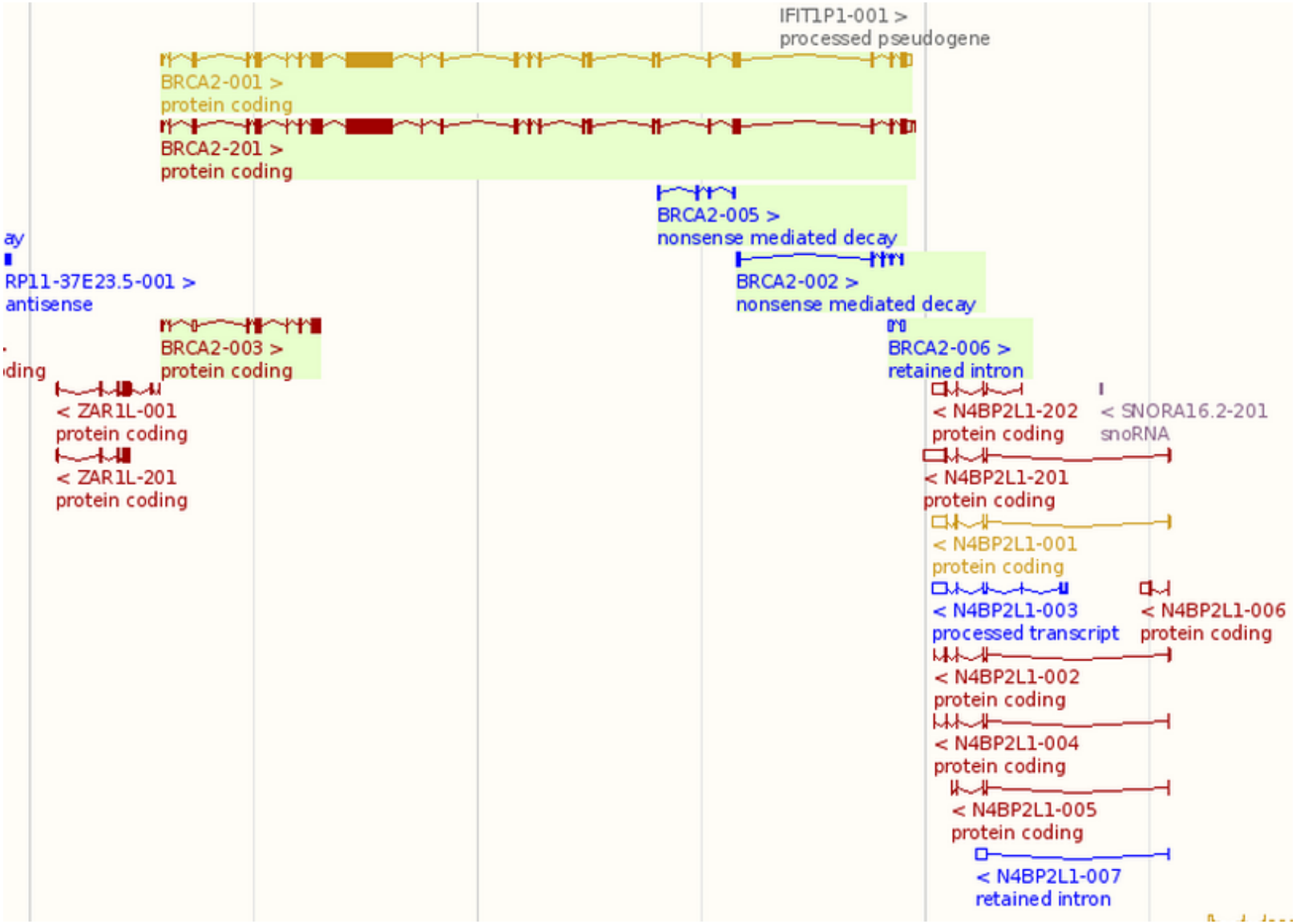
Buljan et al, 2012

Alternative polyA

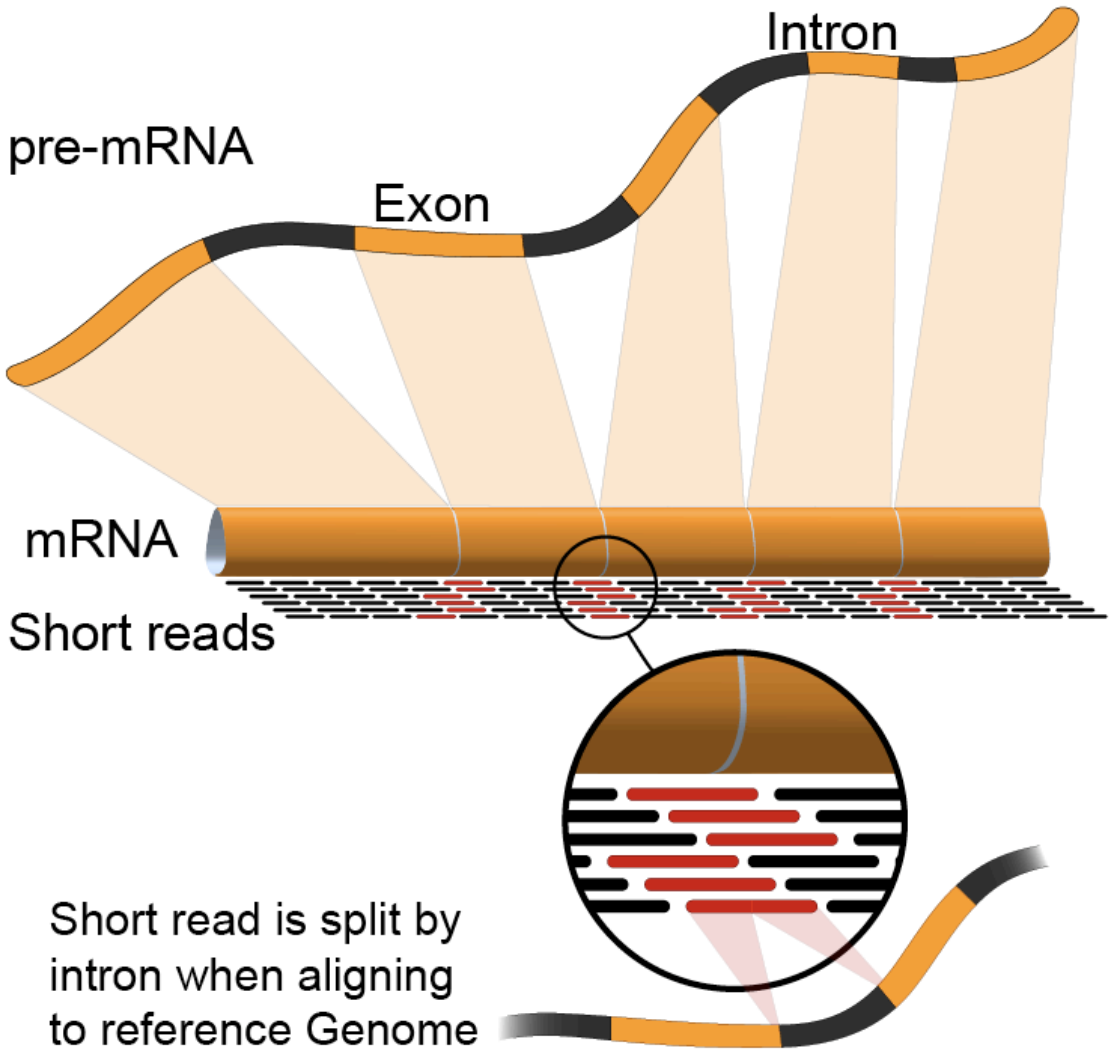


Smibert et al, 2012

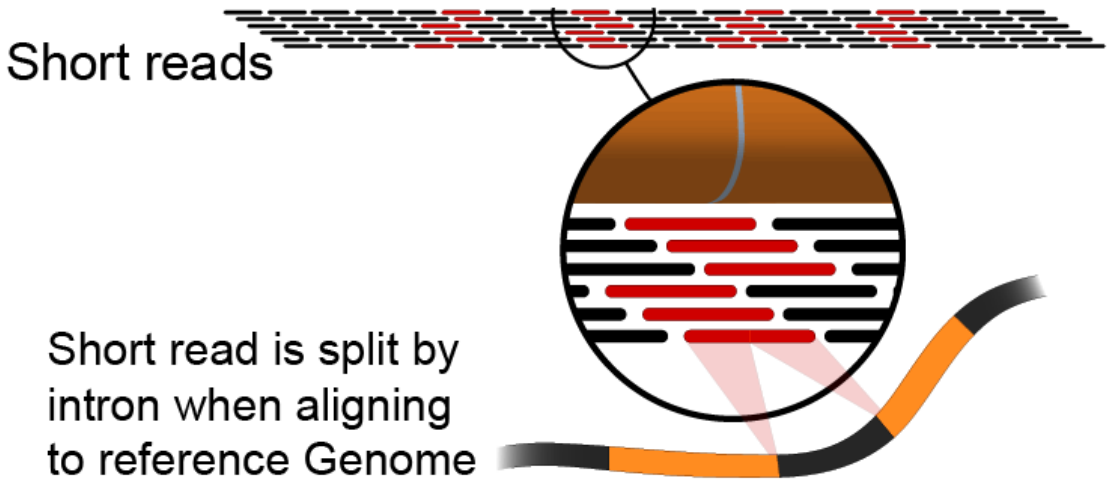
Genes have an average of ~ 6 annotated transcripts



Reads give partial information about transcripts



Reads give partial information about transcripts



Reads give partial information about transcripts

- Transcript assembly (Steijger et al, Nature Methods, 2013)

Reads give partial information about transcripts

- Transcript assembly (Steijger et al, Nature Methods, 2013)
- Isoform centric approaches

Reads give partial information about transcripts

- Transcript assembly (Steijger et al, Nature Methods, 2013)

- Isoform centric approaches
  - Hu et al. [2014], Bernard et al. [2014], Bohnert and Ratsch [2010], Huang et al. [2013], Li and Dewey [2011], Mezlini et al. (iReckon) [2013] , Li and Jiang (RSEM) [2012], Suo et al. [2014], **Trapnell et al (cufflinks). [2011]**

Reads give partial information about transcripts

- Transcript assembly (Steijger et al, Nature Methods, 2013)
- Isoform centric approaches
  - Hu et al. [2014], Bernard et al. [2014], Bohnert and Ratsch [2010], Huang et al. [2013], Li and Dewey [2011], Mezlini et al. (iReckon) [2013] , Li and Jiang (RSEM) [2012], Suo et al. [2014], **Trapnell et al (cufflinks). [2011]**
- Exon centric approaches

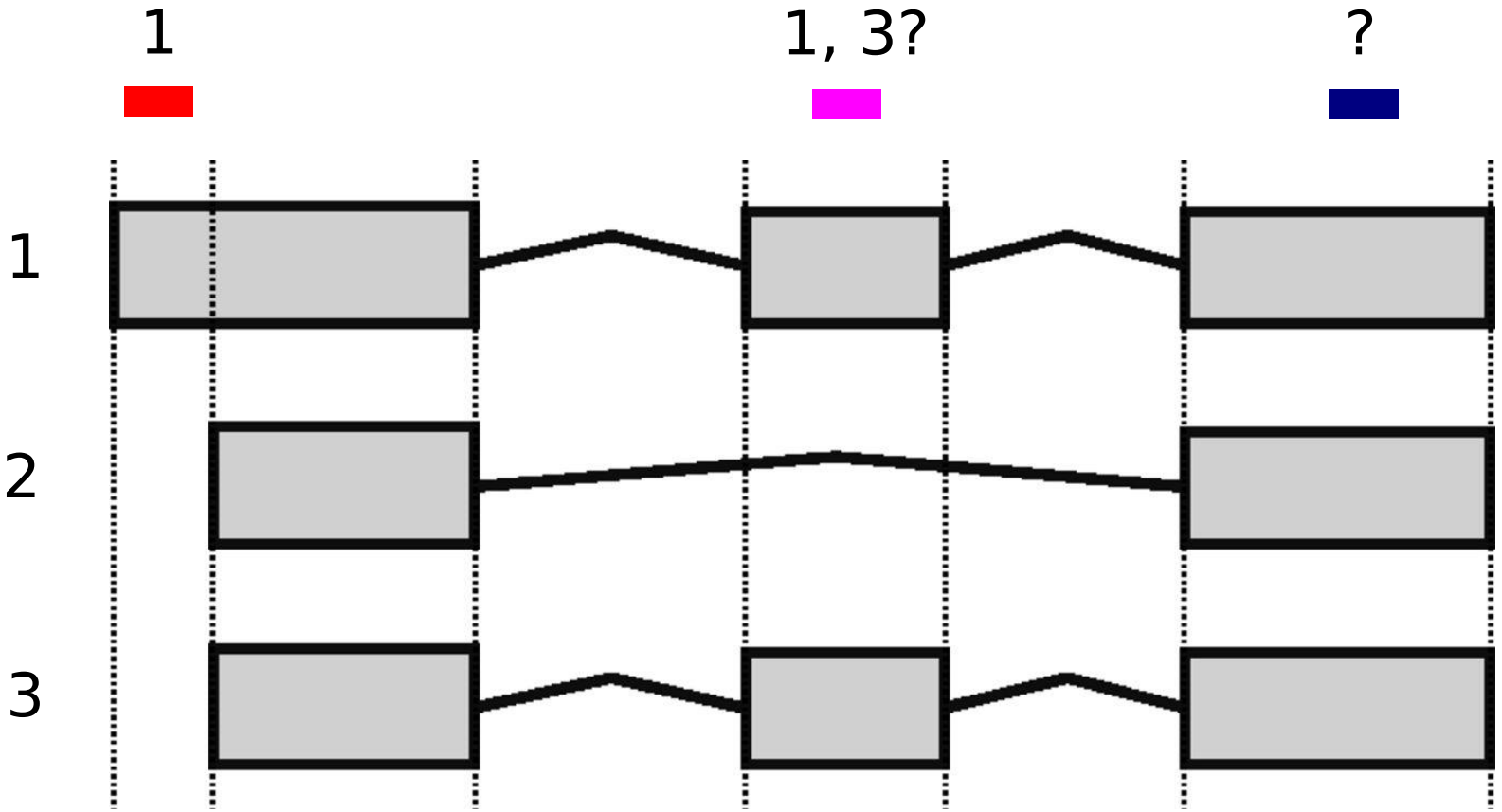
Reads give partial information about transcripts

- Transcript assembly (Steijger et al, Nature Methods, 2013)
- Isoform centric approaches
  - Hu et al. [2014], Bernard et al. [2014], Bohnert and Ratsch [2010], Huang et al. [2013], Li and Dewey [2011], Mezlini et al. (iReckon) [2013] , Li and Jiang (RSEM) [2012], Suo et al. [2014], **Trapnell et al (cufflinks). [2011]**
- Exon centric approaches
  - **Aschoff et al. (SplicingCompass) [2013], Katz et al. (MISO) [2010], Brooks et al.(BaseJunc) [2011], Anders et al. (DEXSeq) [2012]**

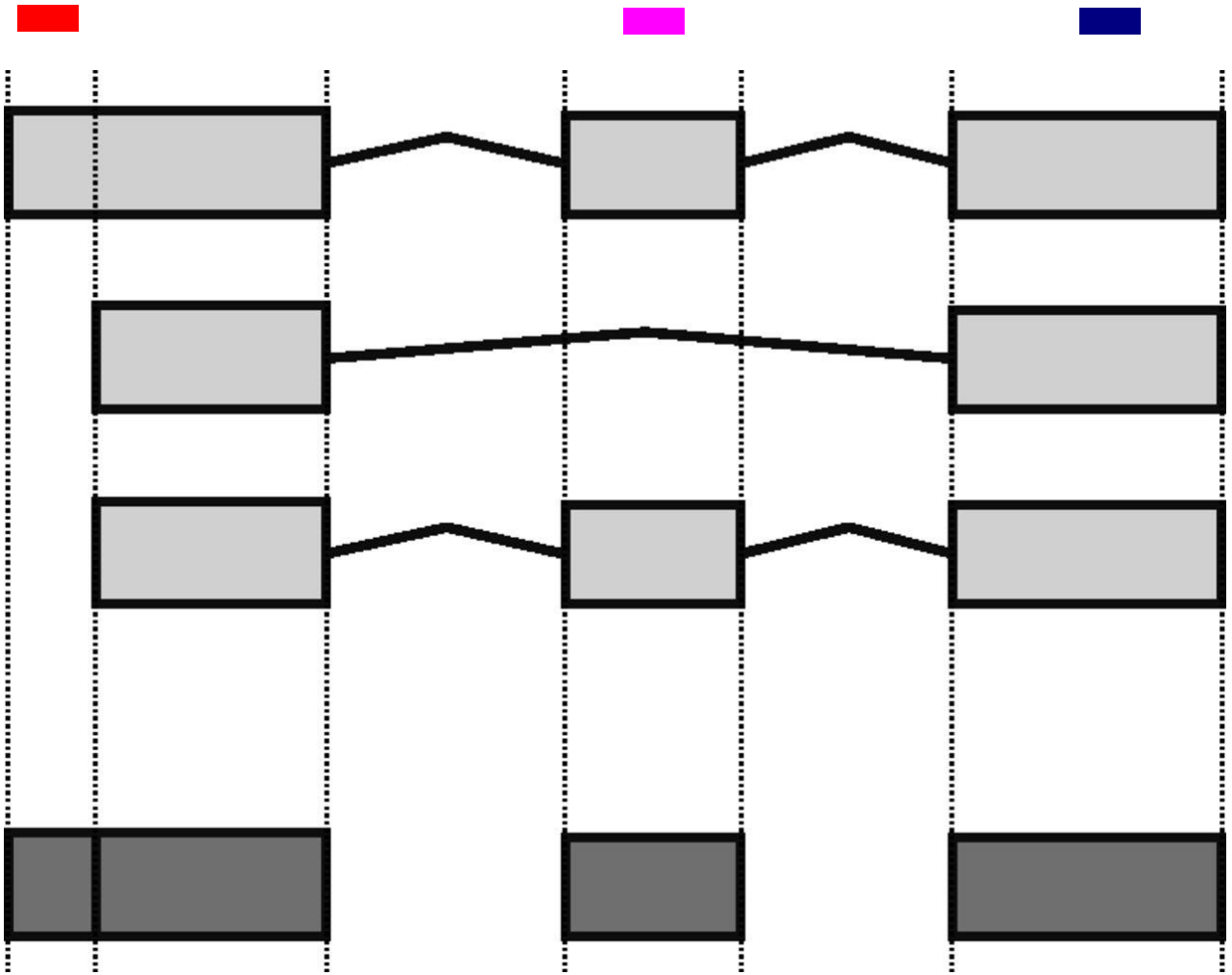
Reads give partial information about transcripts

- Transcript assembly (Steijger et al, Nature Methods, 2013)
- Isoform centric approaches
  - Hu et al. [2014], Bernard et al. [2014], Bohnert and Ratsch [2010], Huang et al. [2013], Li and Dewey [2011], Mezlini et al. (iReckon) [2013] , Li and Jiang (RSEM) [2012], Suo et al. [2014], **Trapnell et al (cufflinks). [2011]**
- Exon centric approaches
  - **Aschoff et al. (SplicingCompass) [2013], Katz et al. (MISO) [2010], Brooks et al.(BaseJunc) [2011], Anders et al. (DEXSeq) [2012]**

Isoform centric approach



Exon centric approach



OUTLINE

DEXSeq: exon centric approach

$$\text{exon usage} = \frac{\text{number of transcripts from the gene that contain an exon}}{\text{number of transcripts from the gene}}$$

DEXSeq inputs count tables

	treated1fb	treated2fb	treated3fb	untreated1fb	untreated2fb	untreated3fb	untreated4fb
E001	1997	494	562	1150	2514	570	547
E002	122	112	180	69	203	156	142
E003	276	293	305	190	398	312	259
E004	420	200	182	230	446	183	185
E005	416	217	279	146	170	237	231
E006	486	357	471	190	337	418	364
E007	574	465	536	469	805	480	496
E008	536	417	447	541	832	475	472
E009	191	237	216	217	427	286	222
E010	188	130	96	617	1177	520	508
E011	165	212	210	118	275	294	269
E012	536	437	414	441	792	619	504
E013	72	41	49	40	76	34	38
E014	3	0	33	5	0	2	42

Generalised linear model

$$K_{ijl} \sim \text{NB}(\text{mean} = s_j \mu_{ijl};\ \text{dispersion} = \alpha_{il})$$

Generalised linear model

$$K_{ijl} \sim \text{NB}(\text{mean} = s_j \mu_{ijl};\ \text{dispersion} = \alpha_{il})$$

$$\log_2 \mu_{ijl} = \beta_{ij}^{\text{S}} + l\beta_i^{\text{E}} + \beta_{i\rho_j}^{\text{EC}}$$

- for sample j, exon i, $l = 0$ for an exon and $l = 1$ for the sum of the rest of the exons of the same gene

Generalised linear model

$$K_{ijl} \sim \text{NB}(\text{mean} = s_j \mu_{ijl};\ \text{dispersion} = \alpha_{il})$$

$$\log_2 \mu_{ijl} = \beta^{\text{S}}_{ij} + l\beta^{\text{E}}_{i} + \beta^{\text{EC}}_{i\rho_j}$$

- for sample j, exon i, $l = 0$ for an exon and $l = 1$ for the sum of the rest of the exons of the same gene
- β^{S}_{ij}, sample specific contribution (expression strength)

Generalised linear model

$$K_{ijl} \sim \text{NB}(\text{mean} = s_j \mu_{ijl};\ \text{dispersion} = \alpha_{il})$$

$$\log_2 \mu_{ijl} = \beta^{\text{S}}_{ij} + l\beta^{\text{E}}_{i} + \beta^{\text{EC}}_{i\rho_j}$$

- for sample j, exon i, $l = 0$ for an exon and $l = 1$ for the sum of the rest of the exons of the same gene
- β^{S}_{ij}, sample specific contribution (expression strength)
- β^{E}_{i}, average exon usage in all the samples

Generalised linear model

$$K_{ijl} \sim \text{NB}(\text{mean} = s_j \mu_{ijl};\ \text{dispersion} = \alpha_{il})$$

$$\log_2 \mu_{ijl} = \beta^{\text{S}}_{ij} + l\beta^{\text{E}}_{i} + \beta^{\text{EC}}_{i\rho_j}$$

- for sample j, exon i, $l = 0$ for an exon and $l = 1$ for the sum of the rest of the exons of the same gene
- β^{S}_{ij}, sample specific contribution (expression strength)
- β^{E}_{i}, average exon usage in all the samples
- $\beta^{\text{EC}}_{i\rho_j}$, interaction coefficient between the exon and the treatment

DEXSeq uses a likelihood ratio test

- Full model

$$\log_2 \mu_{ijl} = \beta^{S}_{ij} + l\beta^{E}_{i} + \beta^{EC}_{i\rho_j}$$

- Reduced model

$$\log_2 \mu_{ijl} = \beta^{S}_{ij} + l\beta^{E}_{i}$$

DEXSeq allows working with complex models

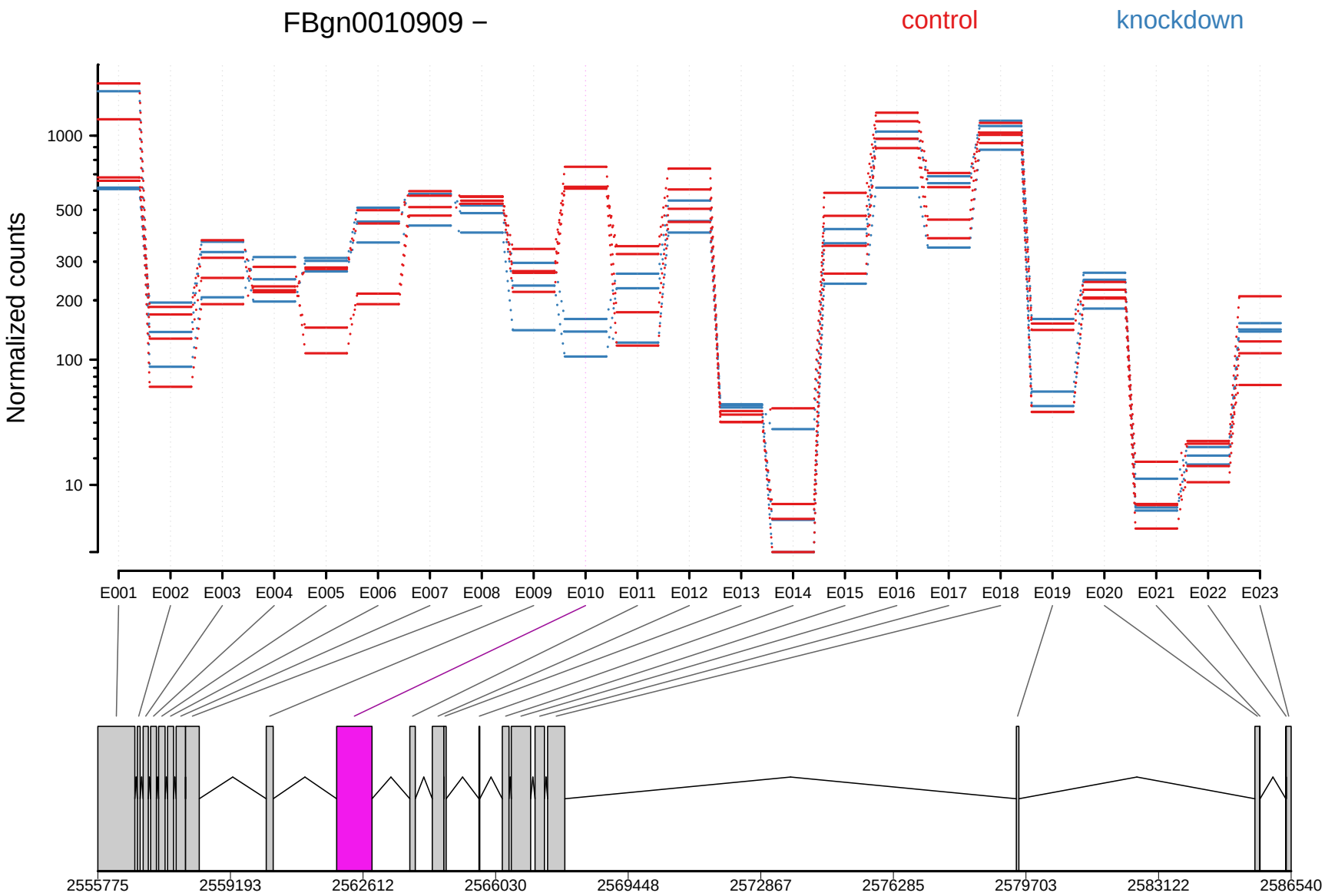
- Full model

$$\log_2 \mu_{ijl} = \beta_{ij}^{S} + +I\beta_{i}^{E} + \beta_{i\Gamma_j}^{ET} + \beta_{i\rho_j}^{EC}$$

- Reduced model

$$\log_2 \mu_{ijl} = \beta_{ij}^{S} + I\beta_{i}^{E} + \beta_{i\Gamma_j}^{ET}$$

DEXSeq output



DEXSeq implementation

- preparating annotation file (creating reduced models for each gene)
  - dexseq_prepare_annotation.py (python) or disjointExons (R)

DEXSeq implementation

- preparating annotation file (creating reduced models for each gene)
  - dexseq_prepare_annotation.py (python) or disjointExons (R)

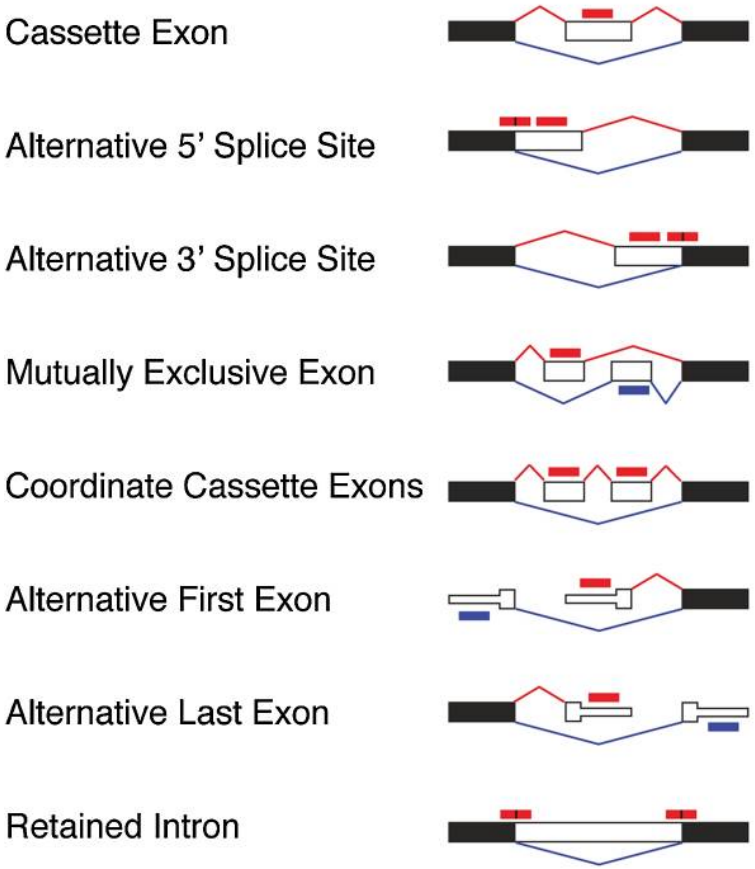
- tabulate the number of reads mapping to each exonic region
  - dexseq_count.py (python) or summarizeOverlaps (R)

DEXSeq implementation

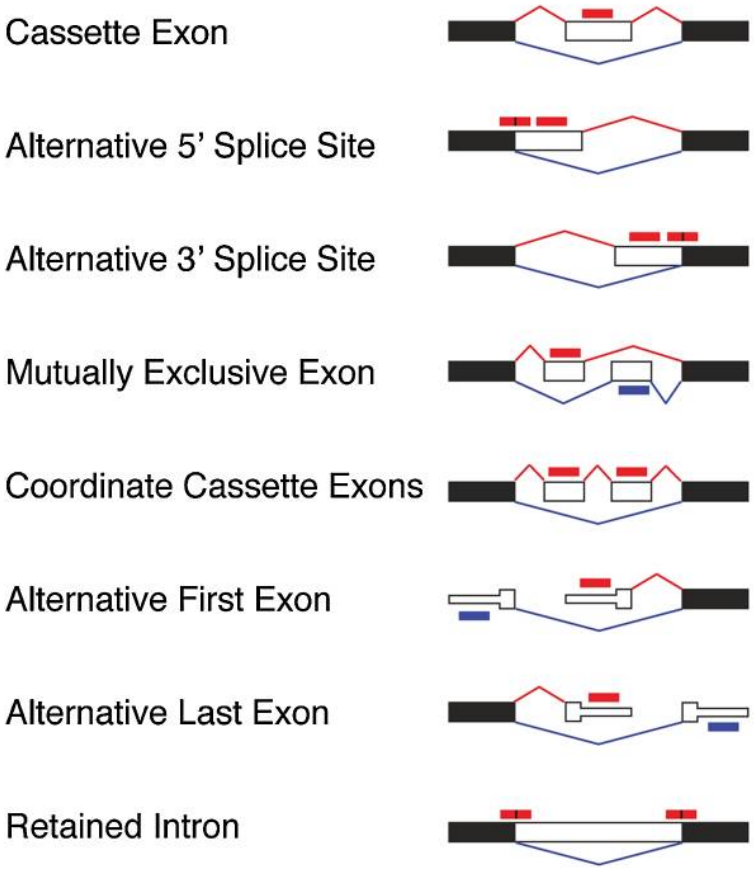
- preparating annotation file (creating reduced models for each gene)
  - dexseq_prepare_annotation.py (python) or disjointExons (R)
- tabulate the number of reads mapping to each exonic region
  - dexseq_count.py (python) or summarizeOverlaps (R)
- Testing, visualization, etc (R)

OUTLINE

Exon-exon junction approaches (MISO, juncBASE)

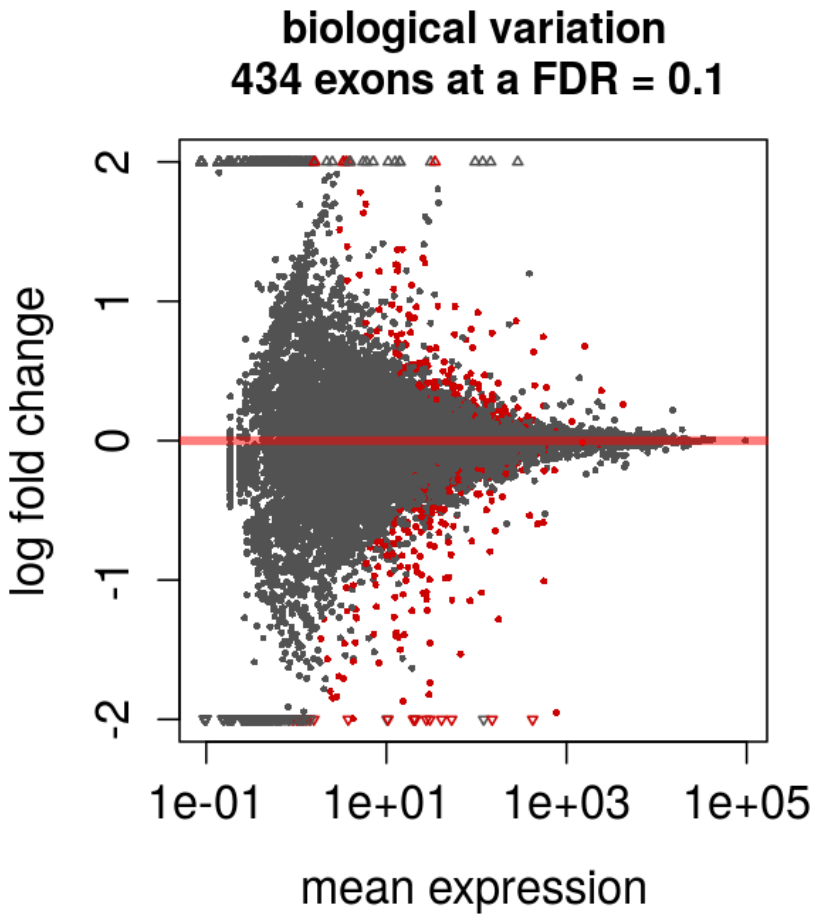


Exon-exon junction approaches (MISO, juncBASE)

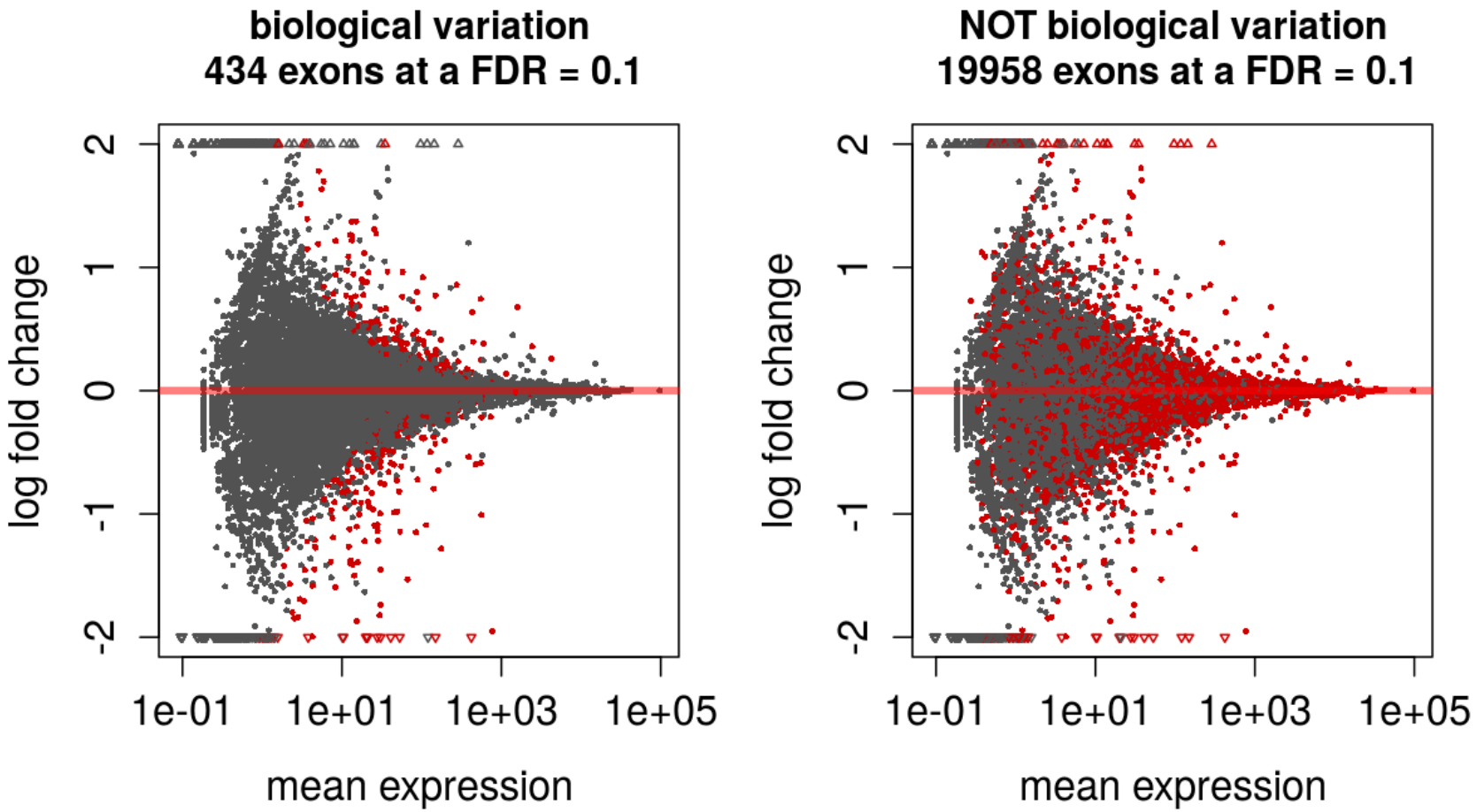


- intuitive counting scheme, but with reduced statistical power
- their testing does not consider biological variation!

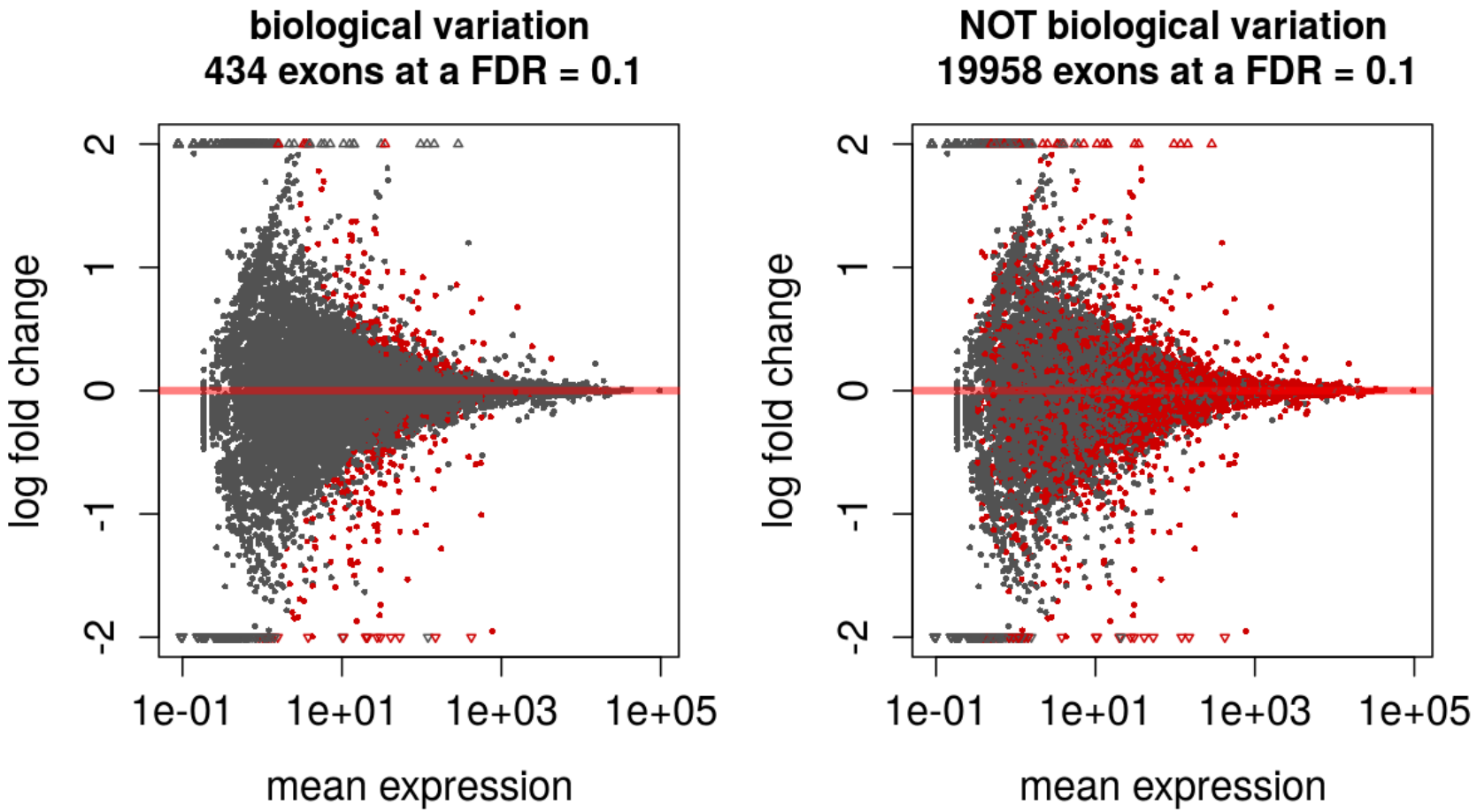
Not considering biological variation leads to high type I error



Not considering biological variation leads to high type I error

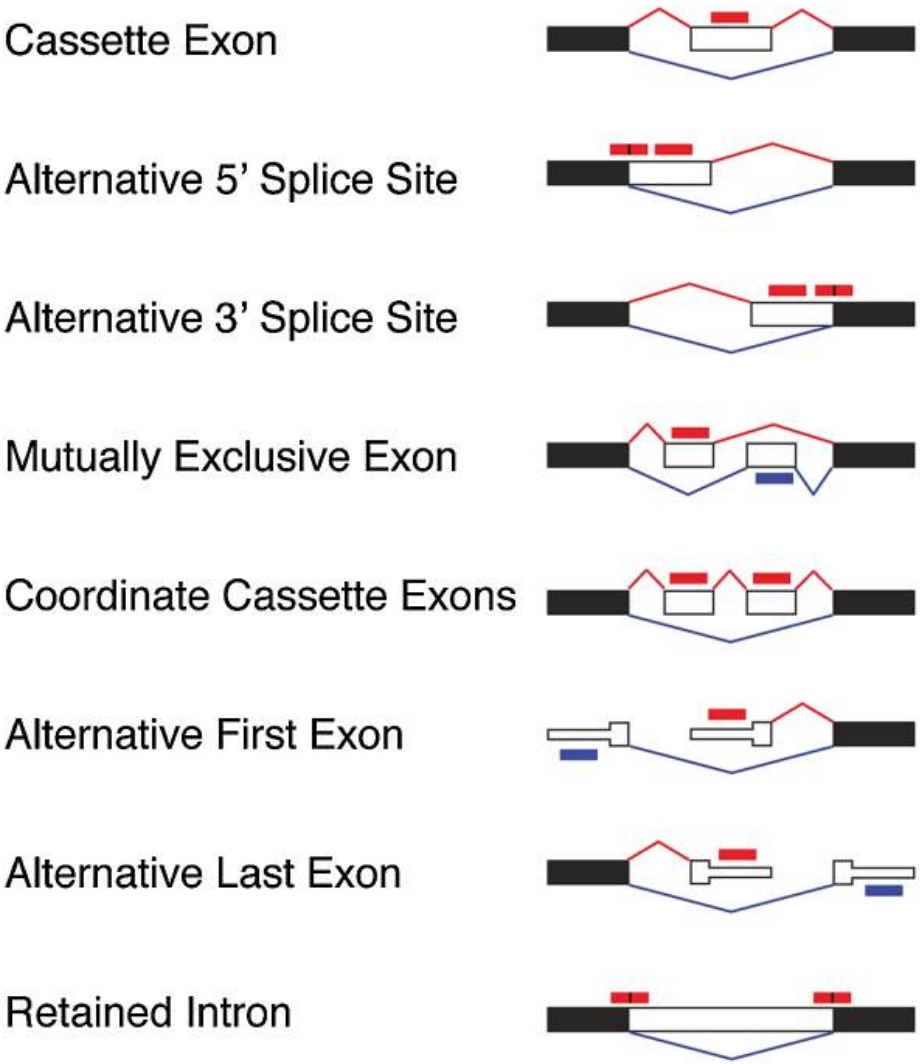


Not considering biological variation leads to high type I error



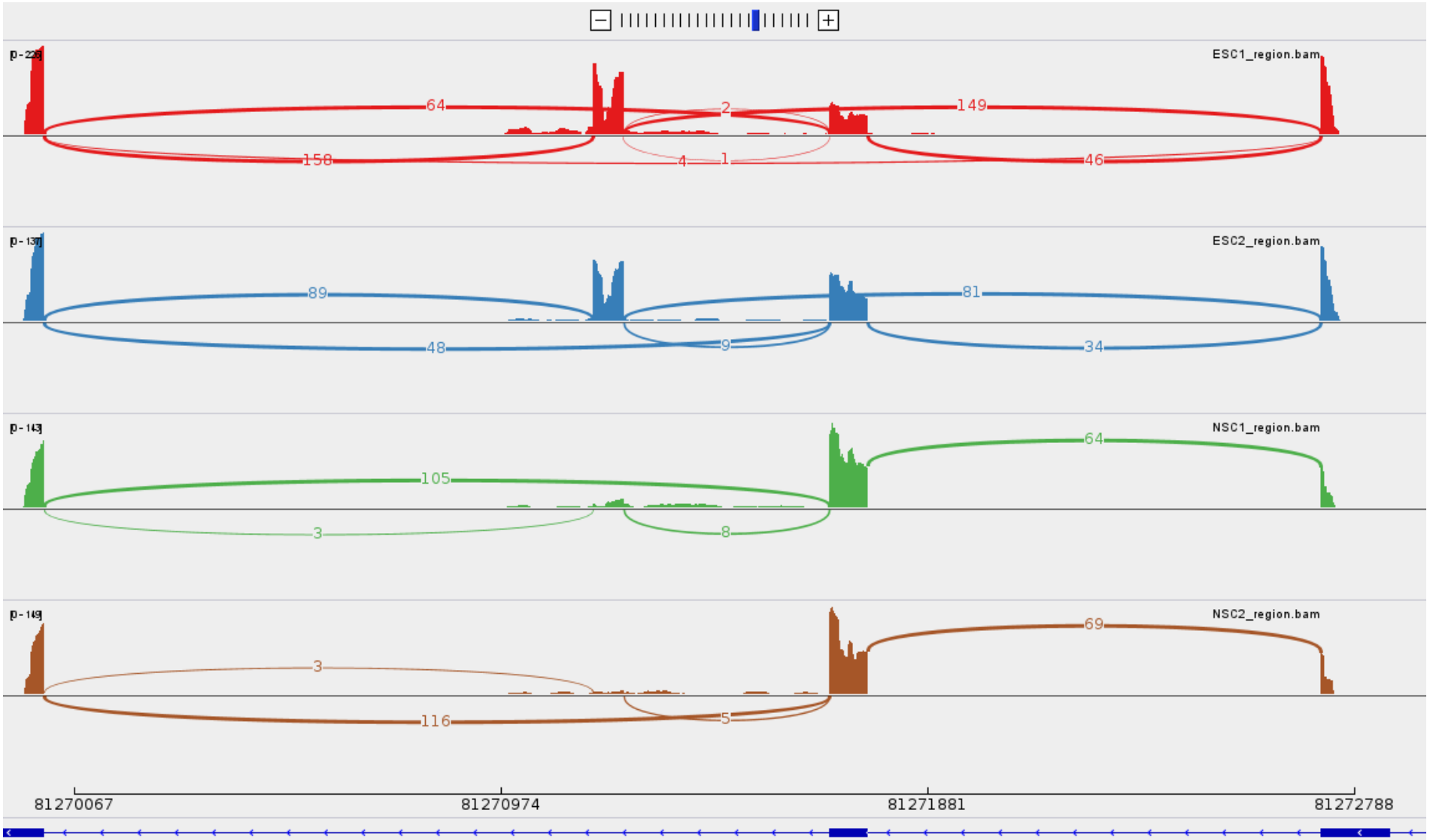
~ 46 times more hits!

Exon-exon junction approaches

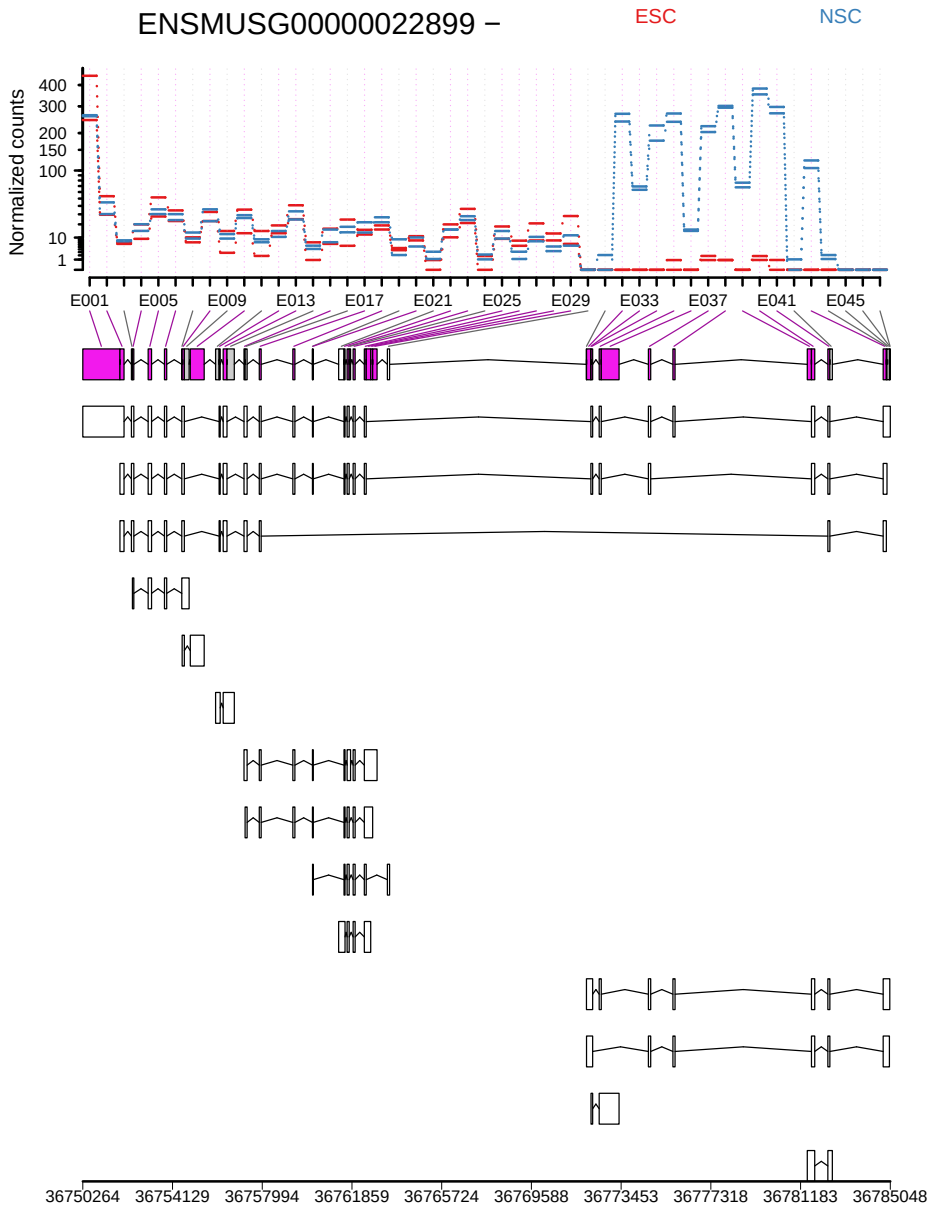


$$\text{exon usage} = \frac{\text{red reads}}{\text{blue reads}}$$

ESC to NSC splicing switches: example of DEXSeq in exon-exon junction reads



Large changes in isoform regulation are missed when we only focus in exon-exon junction reads



OUTLINE

DEXSeq vs cuffdiff

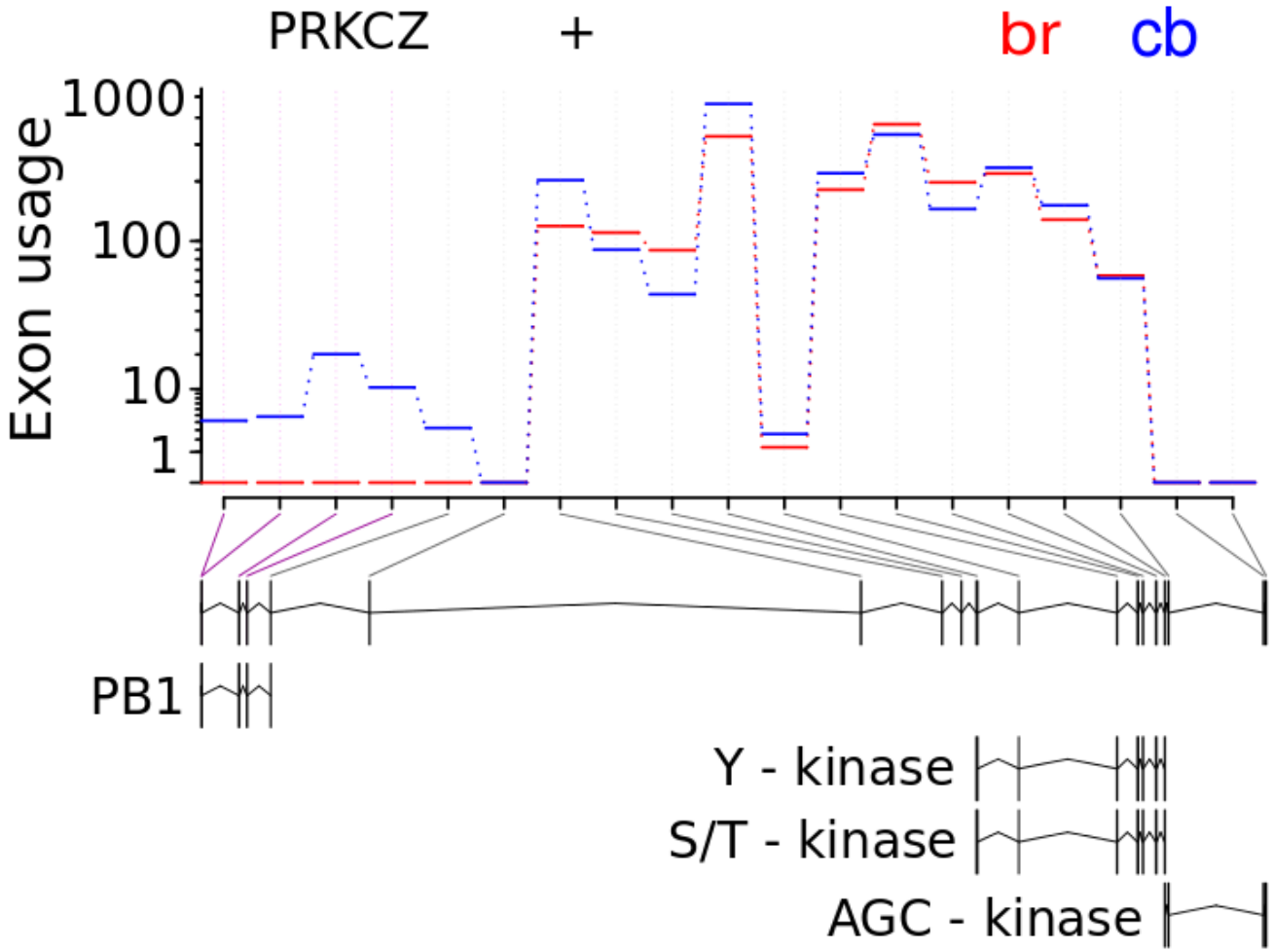
Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.1.0	cuffdiff 1.2.0	cuffdiff 1.3.0
proper comparisons, treatment (knock-down) vs control:					
T1 – T3	C1 – C4	159	145	69	50
T1, T2	C2, C3	52	323	120	578
mock comparisons, control vs control:					
C1, C3	C2, C4	8	314	650	639
C1, C4	C2, C3	7	392	724	728

DEXSeq vs cuffdiff

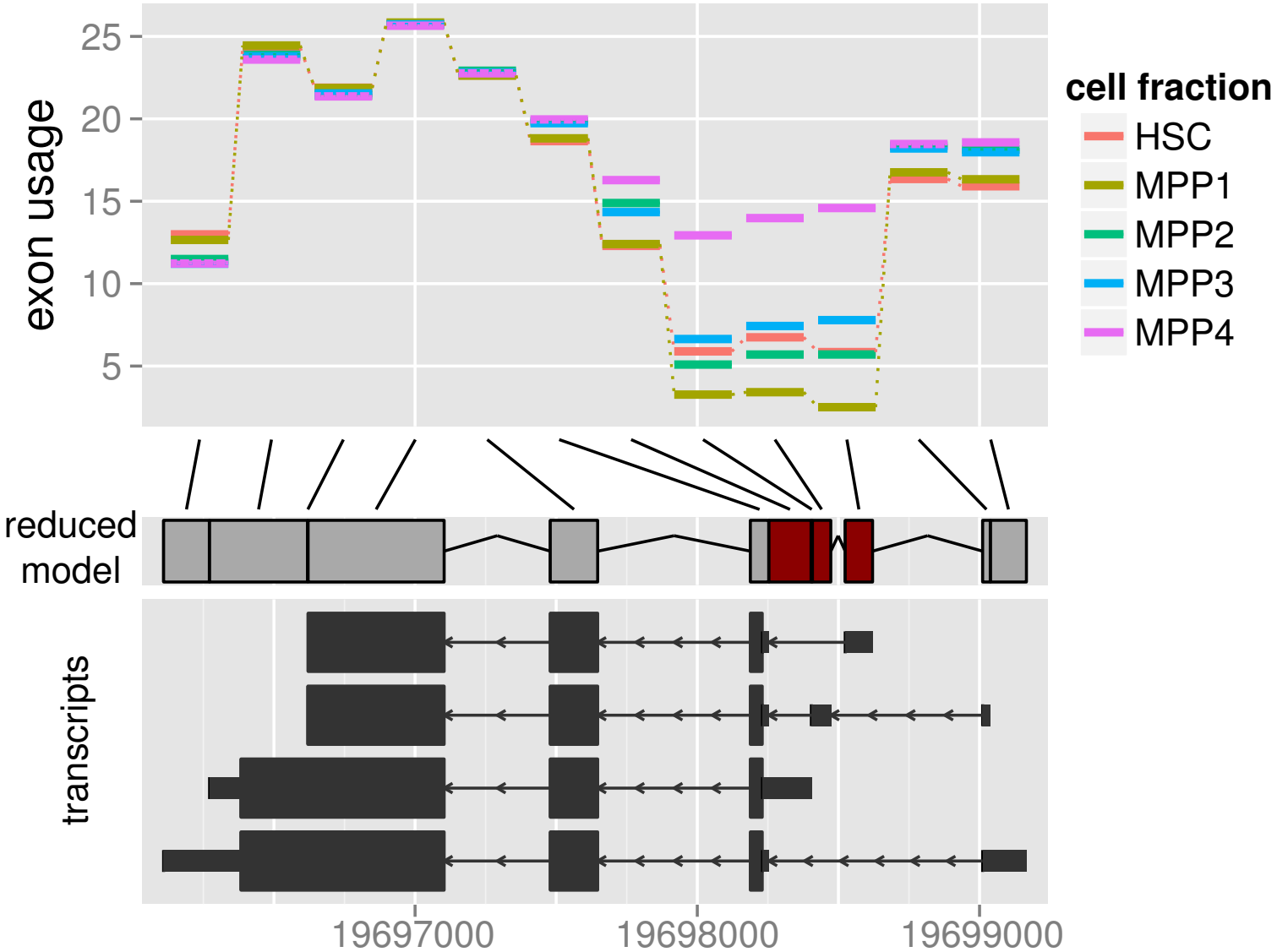
Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.1.0	cuffdiff 1.2.0	cuffdiff 1.3.0
proper comparisons, treatment (knock-down) vs control:					
T1 – T3	C1 – C4	159	145	69	50
T1, T2	C2, C3	52	323	120	578
mock comparisons, control vs control:					
C1, C3	C2, C4	8	314	650	639
C1, C4	C2, C3	7	392	724	728

Recent versions of cuffdiff are very conservative!

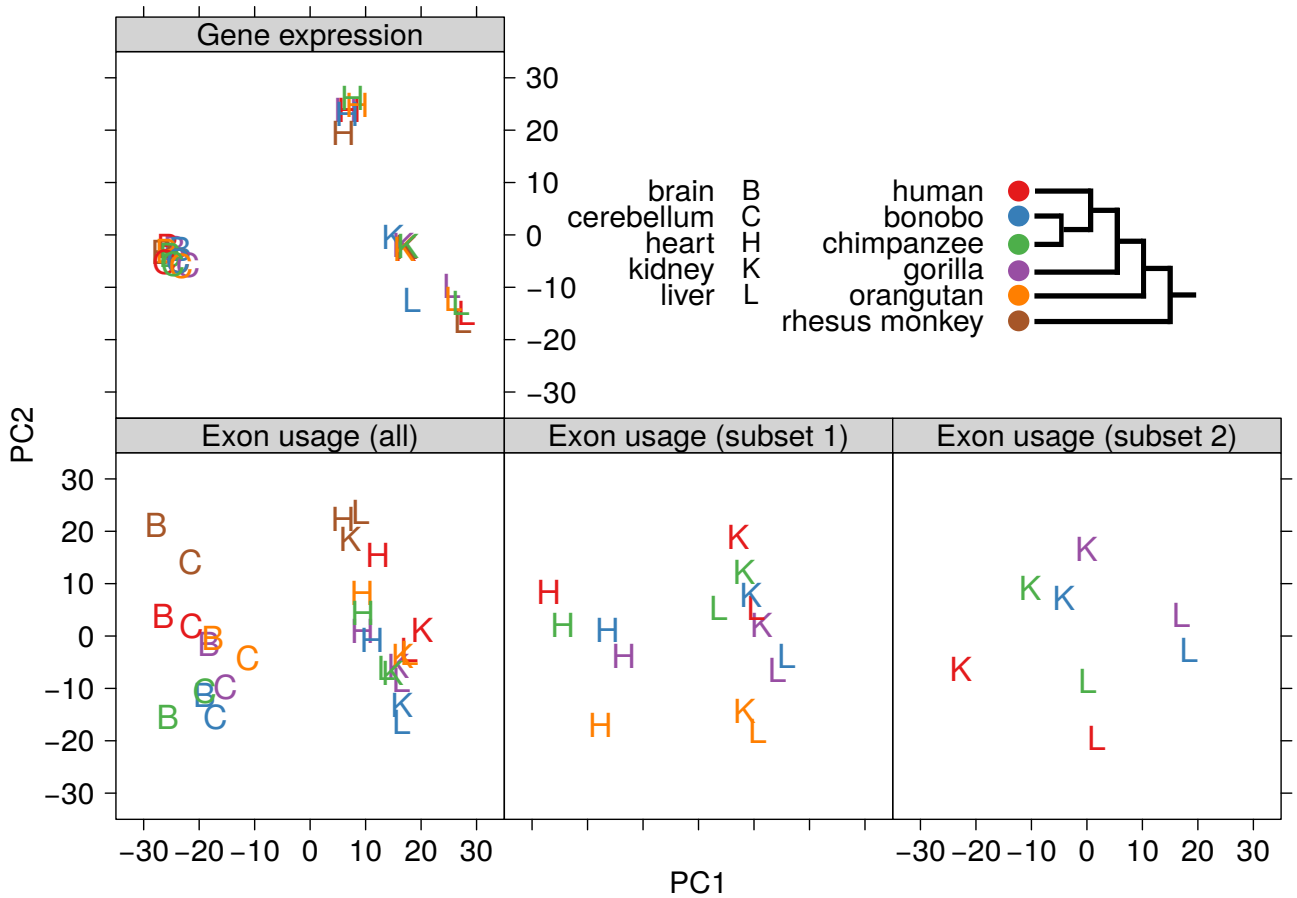
PRKCZ in brain



Apoe in Mouse HSC and MPP



Exon usage has more inter-species variation than gene expression



Reyes et al, PNAS, 2013

Co-authors

Simon Anders

Wolfgang Huber

Thanks to

Huber group

Mike Love