Overcoming Bias and Batch Effects in RNA-Seq Data

Rafael A Irizarry
Professor of Biostatistics and Computational Biology,
Dana Farber Cancer Institute
Professor of Biostatistics, Harvard School of Public Health

@rafalab

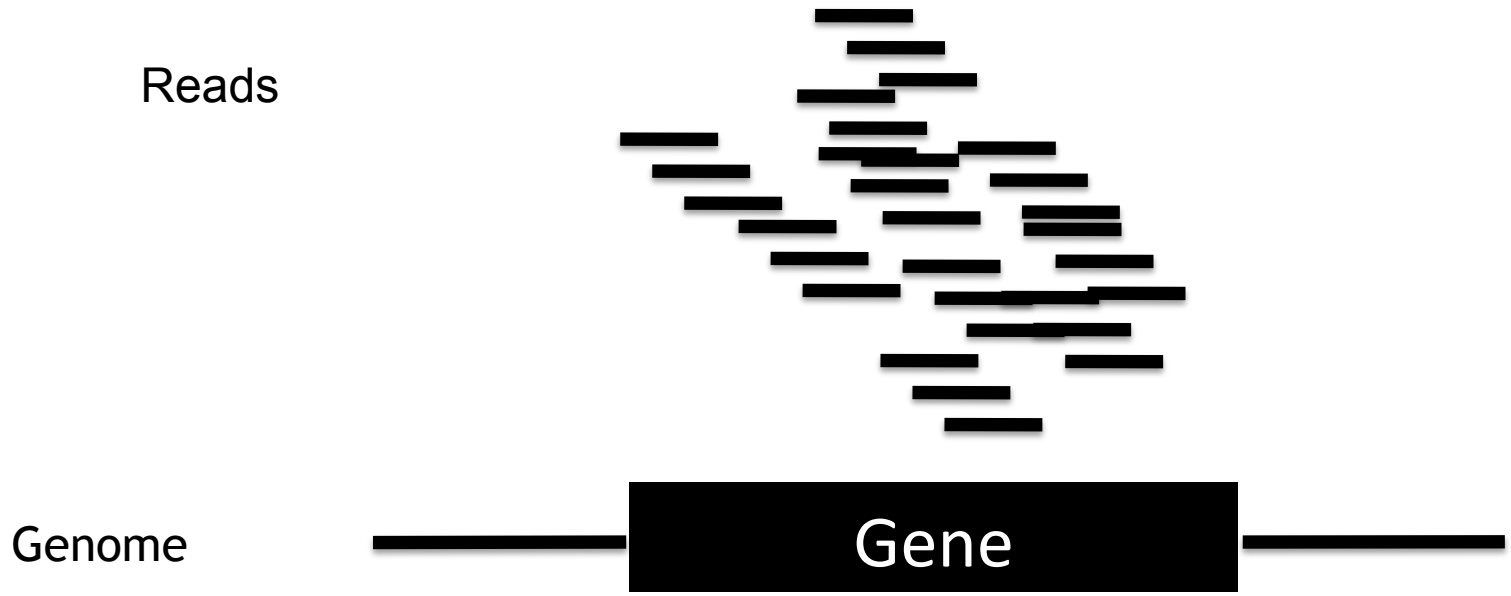
http://www.bioconductor.org

Co-authors

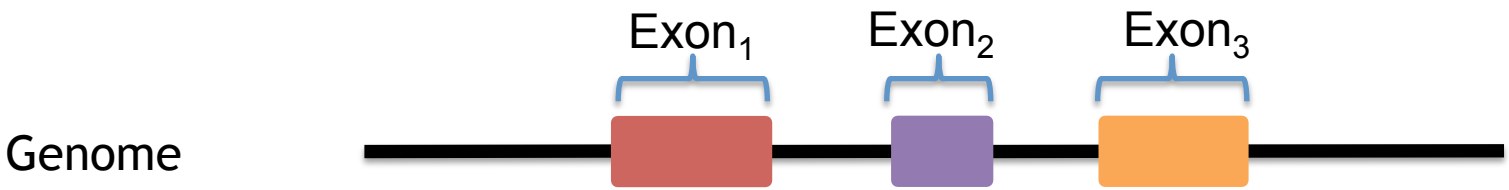
Mike Love

John Hogenesch

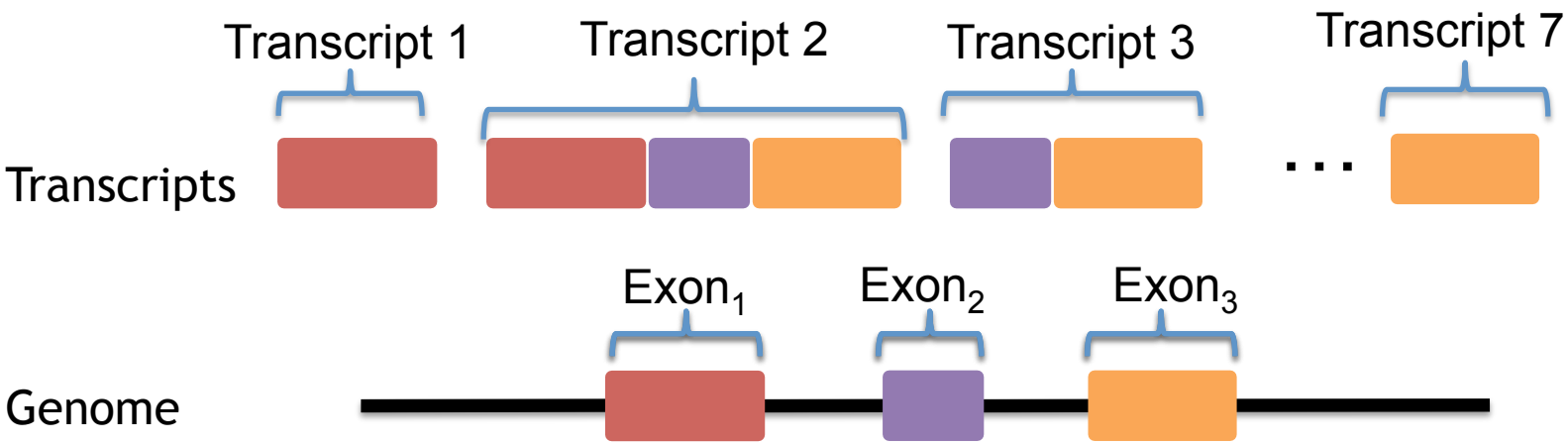
Gene expression



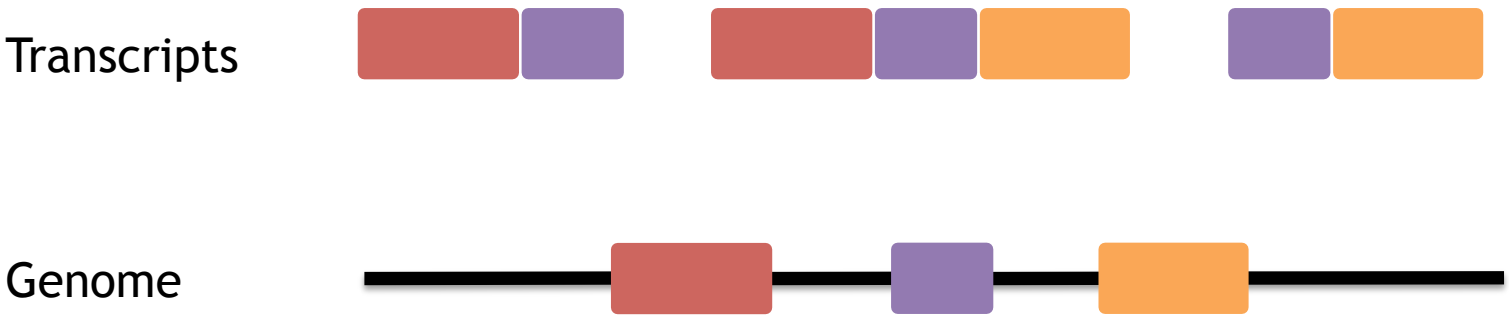
More complicated



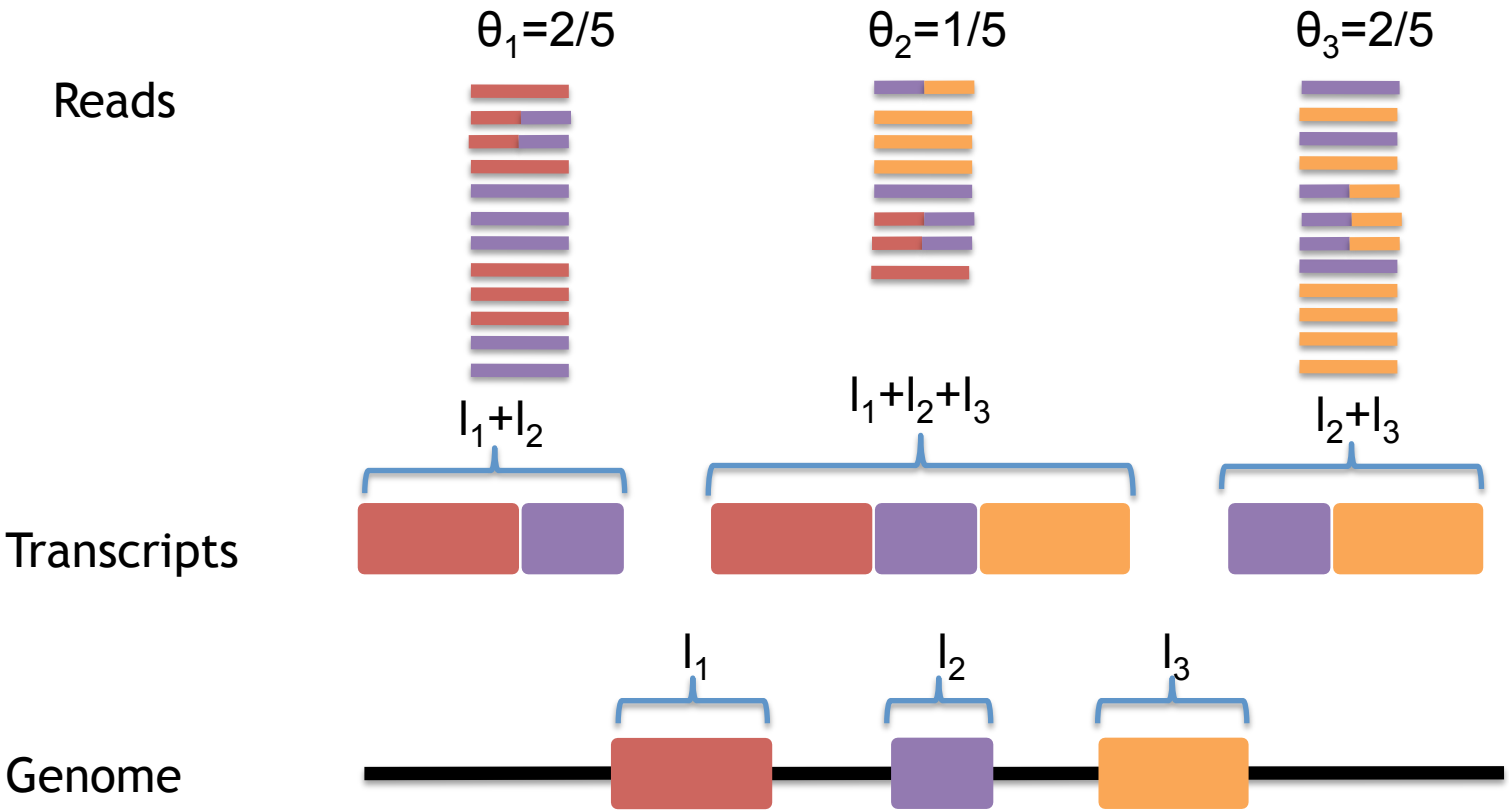
More complicated



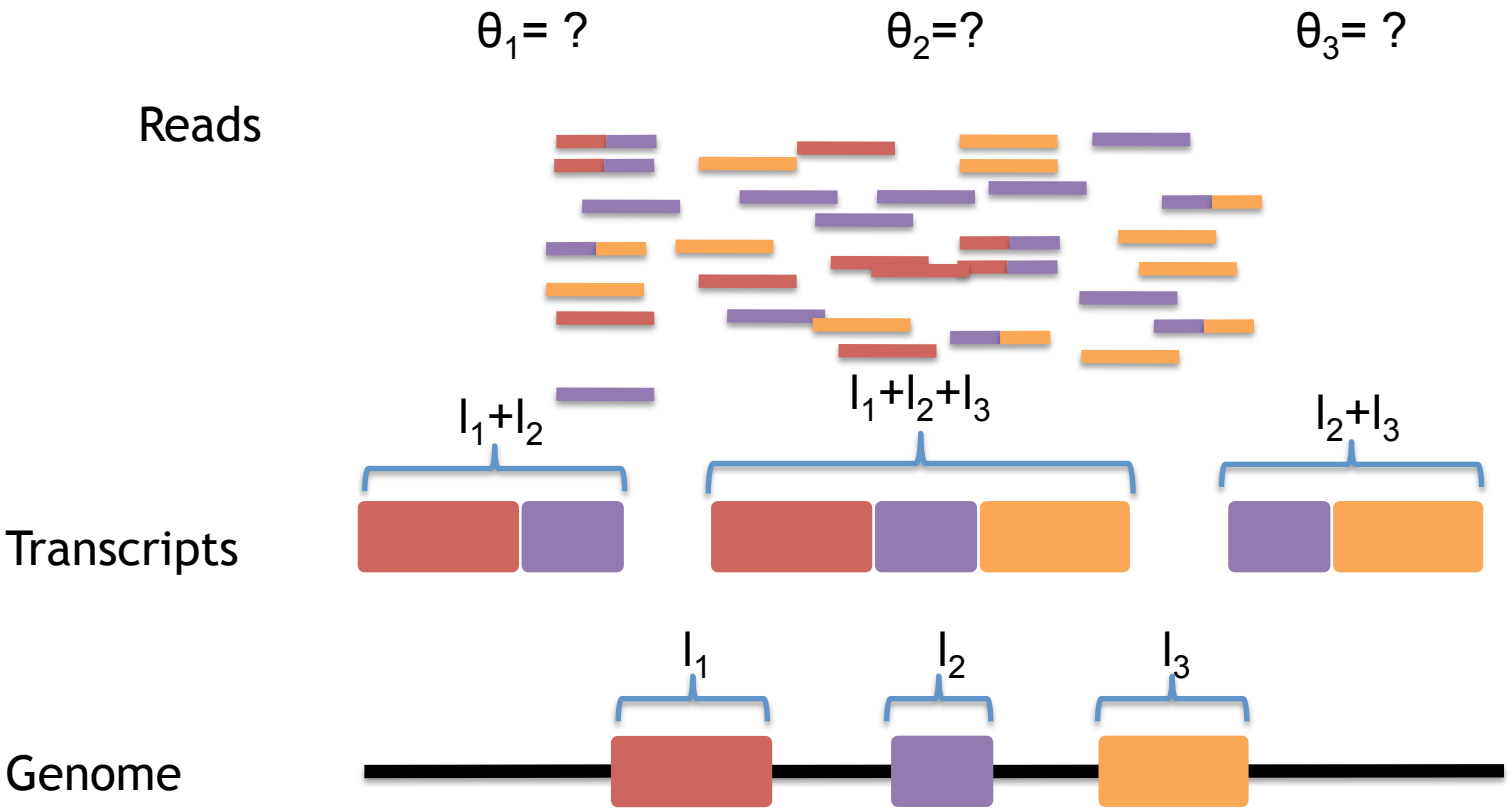
Data generation (illustration with 3 transcripts)



Data generation



We see



Statistical model

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_{1,2} \\ Y_{2,3} \end{pmatrix} = \begin{pmatrix} \text{Exon 1 count} \\ \text{Exon 2 count} \\ \text{Exon 3 count} \\ \text{Junction 1,2 count} \\ \text{Junction 2,3 count} \end{pmatrix}$$

For example $Y_1 \sim \text{Poisson}(l_1\theta_1 + l_1\theta_2)$

Jiang and Wong (2009) *Bioinformatics*

Statistical model

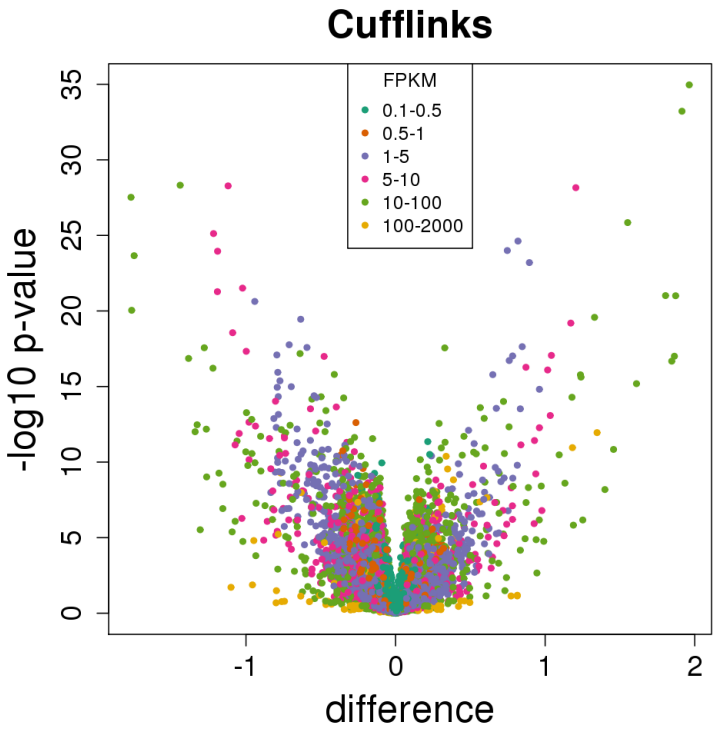
Y's are independent Poisson and transcript quantification is MLE of θs

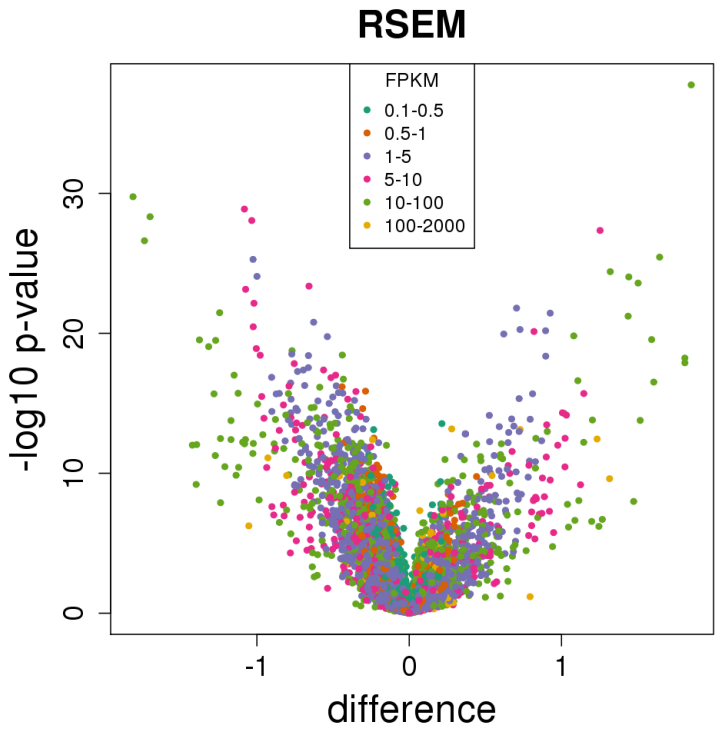
$$\mathrm{E}\begin{pmatrix} Y_1/l_1 \\ Y_2/l_2 \\ Y_3/l_3 \\ Y_{1,2}/l_{1,2} \\ Y_{2,3}/l_{2,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

Notes:
1) this not standard GLM as link is not log
2) Empirical Bayes approach are commonly used

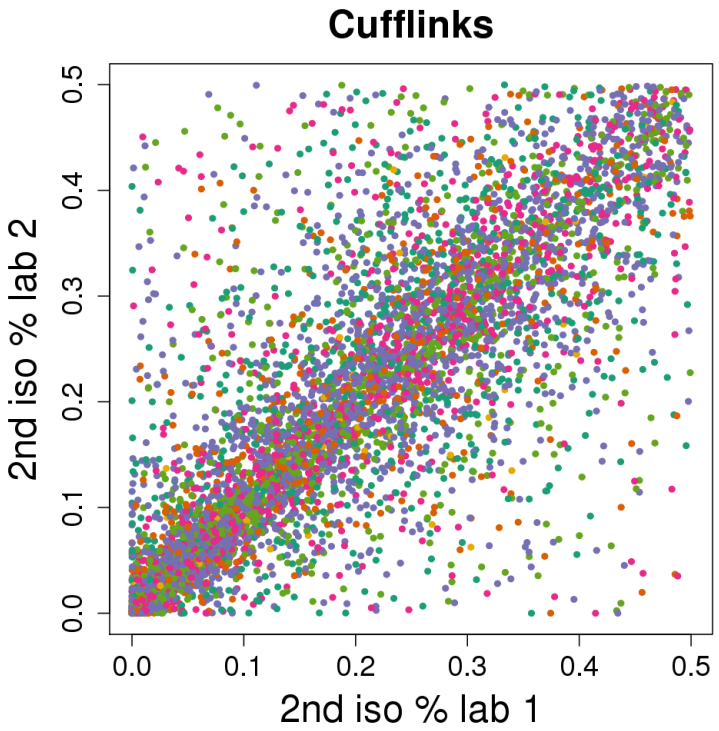
Jiang and Wong (2009) *Bioinformatics*

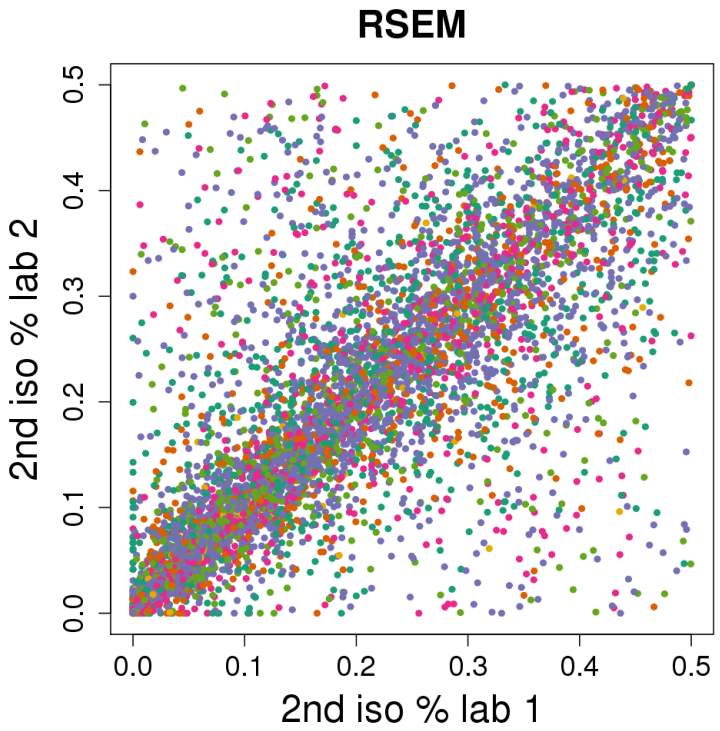
Comparison of two groups sampled from one population (from GEUVADIS)

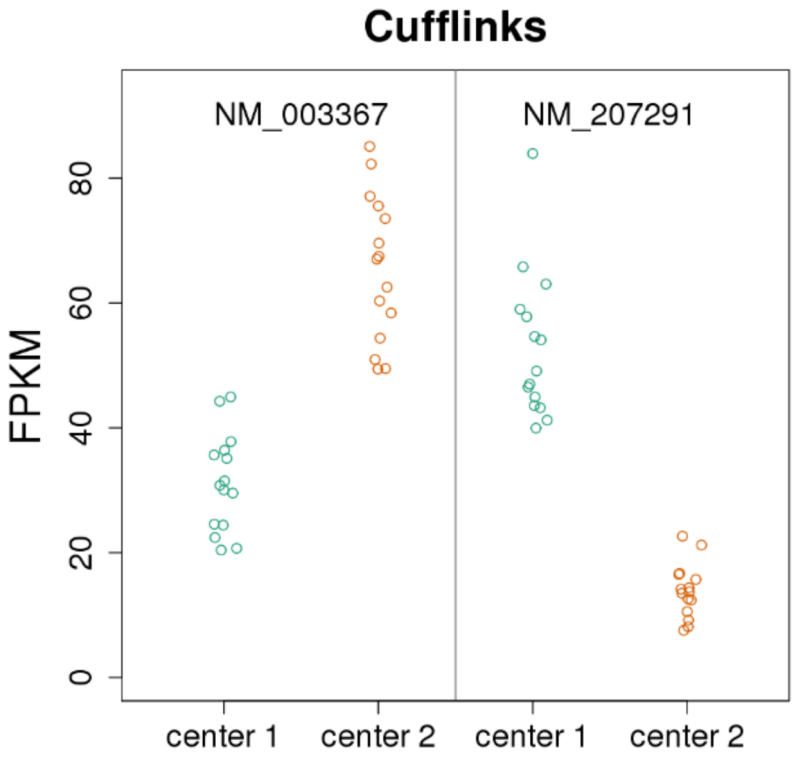


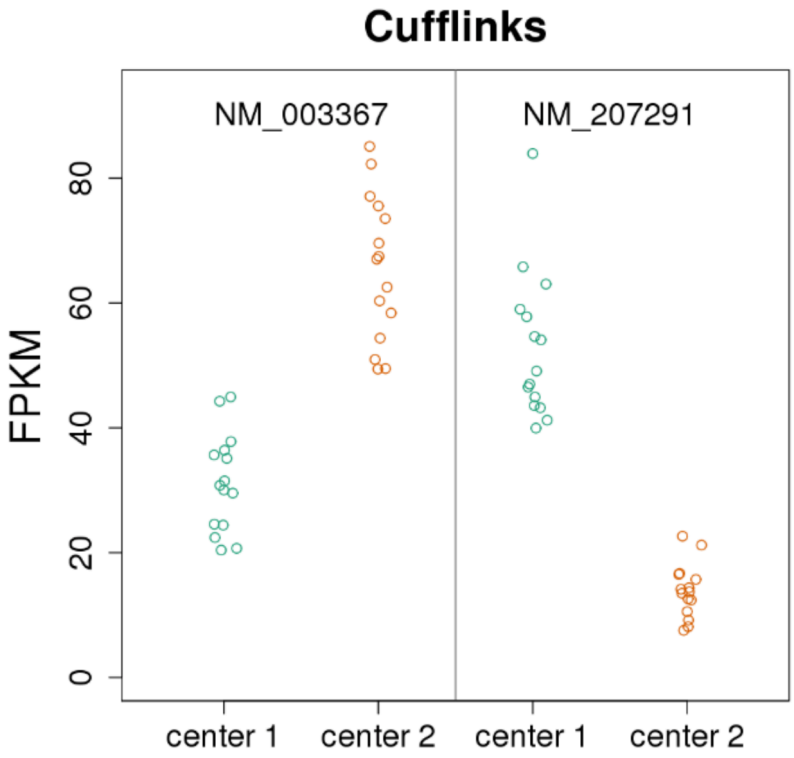


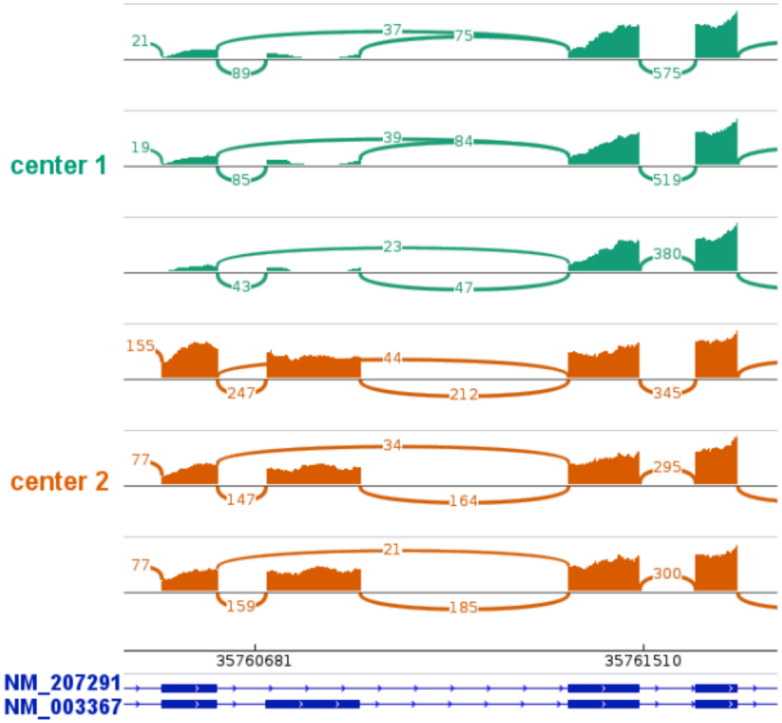
% Expression assigned to second highest isoform






Cufflinks
NM_003367
NM_207291
FPKM
0
20
40
60
80
center 1
center 2
center 1
center 2


Cufflinks
NM_003367
NM_207291
FPKM
0
20
40
60
80
center 1
center 2
center 1
center 2


center 1
center 2
21
37
75
89
575
19
39
84
85
519
23
380
43
47
155
44
247
212
345
77
34
295
147
164
77
21
300
159
185
35760681
35761510
NM_207291
NM_003367

This statistical model is wrong

Y's are independent Poisson and transcript quantification is MLE of θs

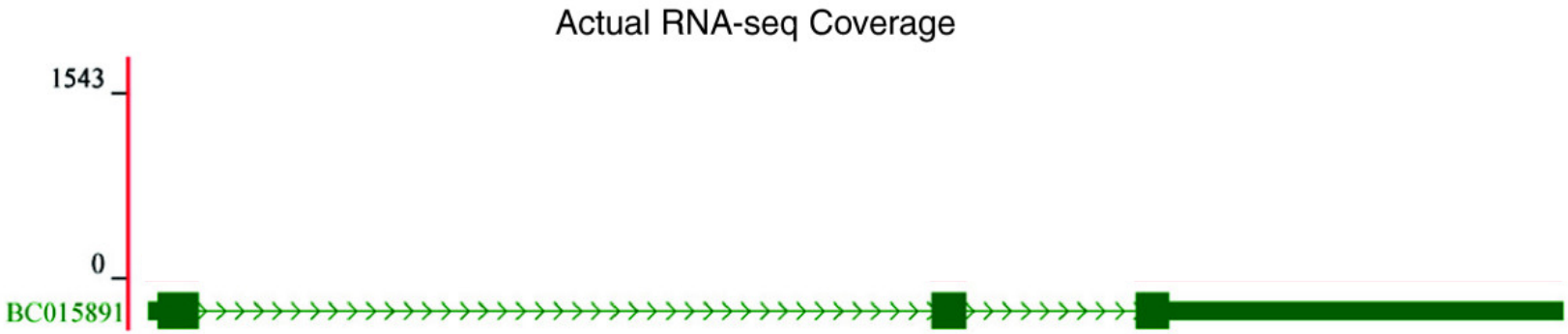
$$\mathrm{E}\begin{pmatrix} Y_1/l_1 \\ Y_2/l_2 \\ Y_3/l_3 \\ Y_{1,2}/l_{1,2} \\ Y_{2,3}/l_{2,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

Notes:
1) this not standard GLM as link is not log
2) Empirical Bayes approach are commonly used

Jiang and Wong (2009) *Bioinformatics*

Technical artifacts in “coverage”

- What should this look like?



In vitro transcription, followed by sequencing: Lahens et al (2014)

Statistical model with bias incorporated

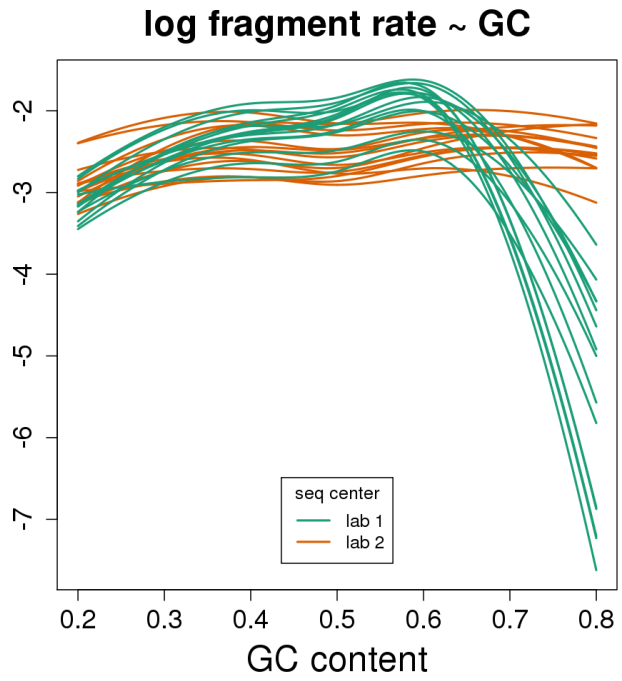
Y's are independent Poisson and transcript quantification is MLE of θs

$$\mathrm{E}\begin{pmatrix} Y_1/l_1 \\ Y_2/l_2 \\ Y_3/l_3 \\ Y_{1,2}/l_{1,2} \\ Y_{2,3}/l_{2,3} \end{pmatrix} = \begin{pmatrix} b_1 & b_1 & 0 \\ b_2 & b_2 & b_2 \\ 0 & b_3 & b_3 \\ b_4 & b_4 & 0 \\ 0 & b_5 & b_5 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

Notes:
1) this not standard GLM as link is not log
2) Empirical Bayes approach are commonly used

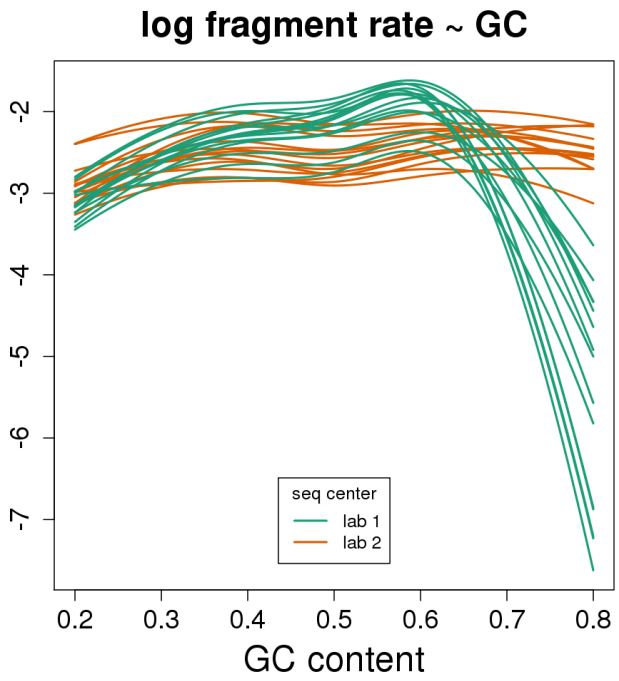
Parameters no longer identifiable

Bias varies from sample to sample



Plots from CQN: Hansen, Irizarry, Wu *Biostatistics 2012*

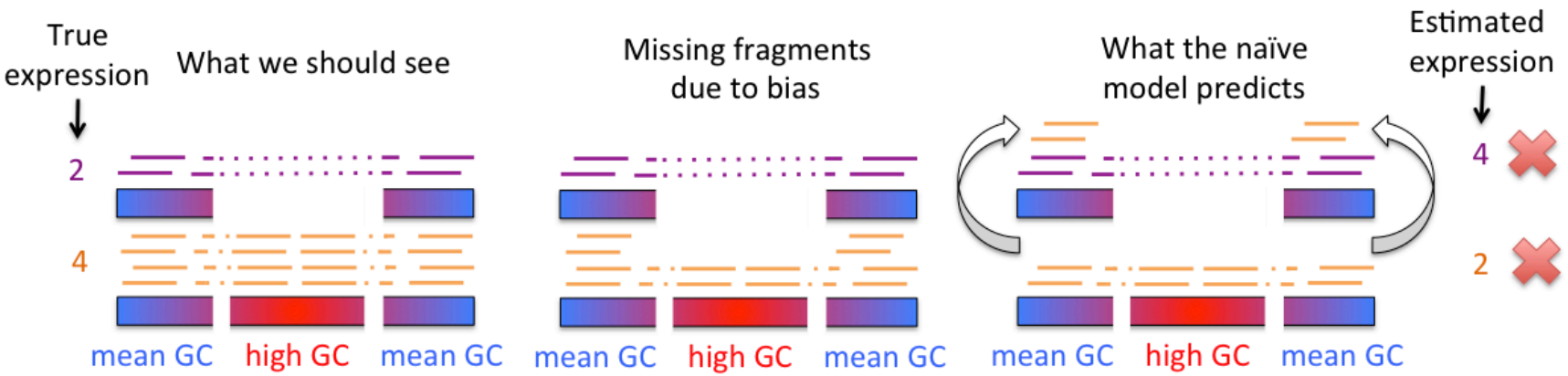
Bias varies from sample to sample

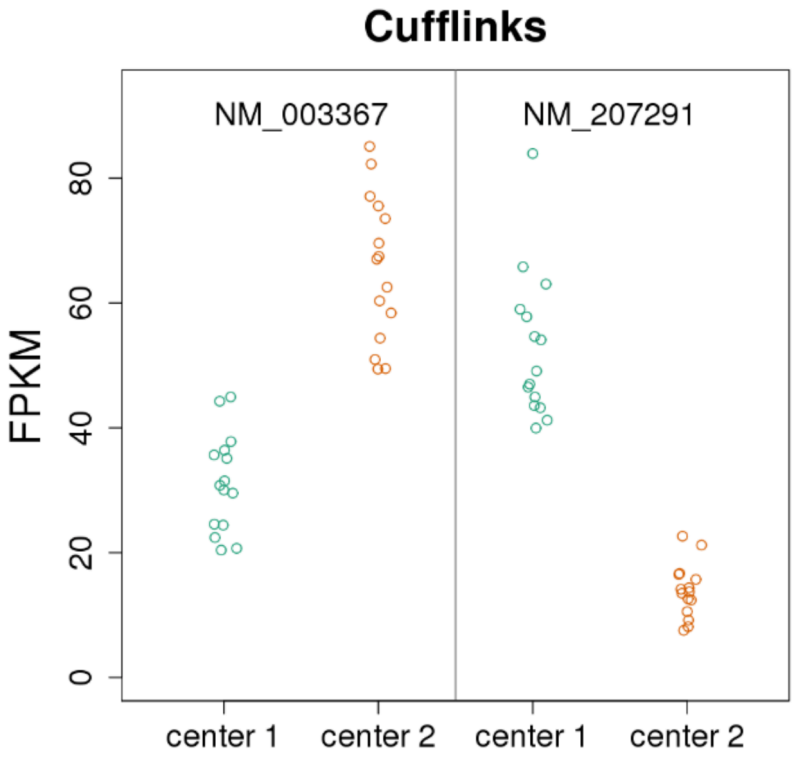


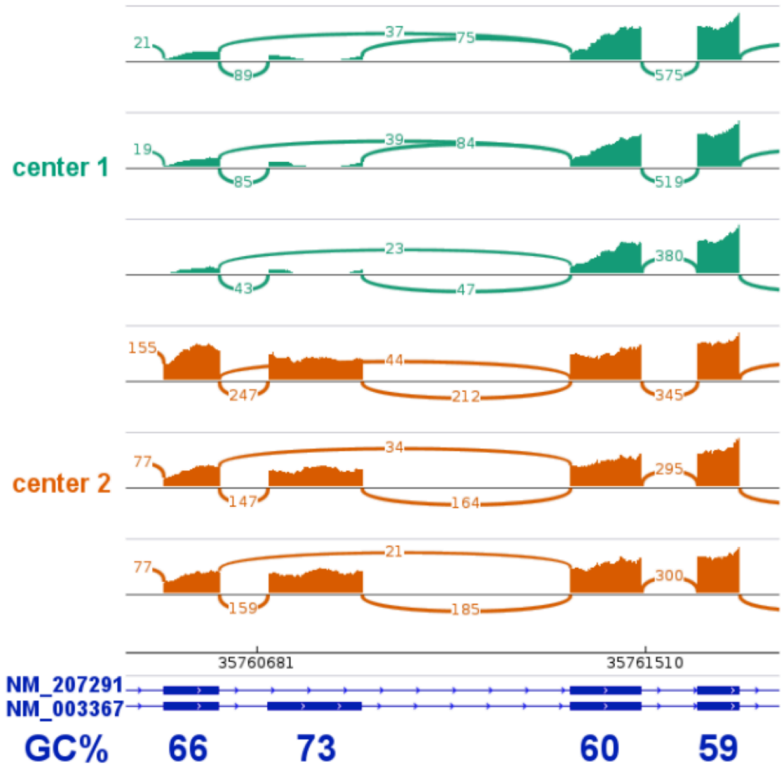
$$Y_j \sim \text{Poisson}(\lambda_j^b)$$

$$\log(\lambda_j^b) = \sum_k X_{jk}\beta_k + o_j + g_j$$

Plots from CQN: Hansen, Irizarry, Wu *Biostatistics 2012*


True expression
What we should see
2
4
mean GC
high GC
mean GC
Missing fragments due to bias
mean GC
high GC
mean GC
What the naïve model predicts
mean GC
high GC
mean GC
Estimated expression
4
2

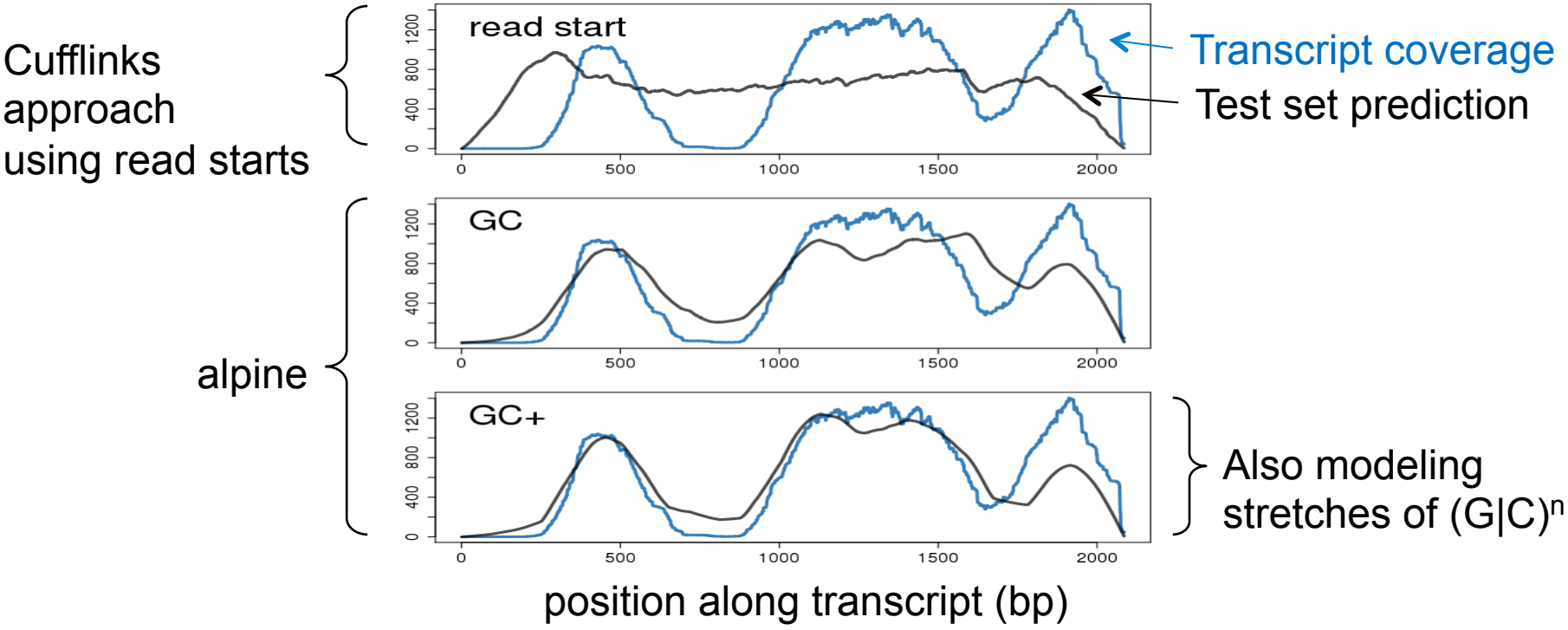

Cufflinks
NM_003367
NM_207291
FPKM
0
20
40
60
80
center 1
center 2
center 1
center 2


center 1
center 2
21
37
75
89
575
19
39
84
85
519
23
380
43
47
155
44
247
212
345
77
34
295
147
164
77
21
300
159
185
35760681
35761510
NM_207291
NM_003367
GC%
66
73
60
59

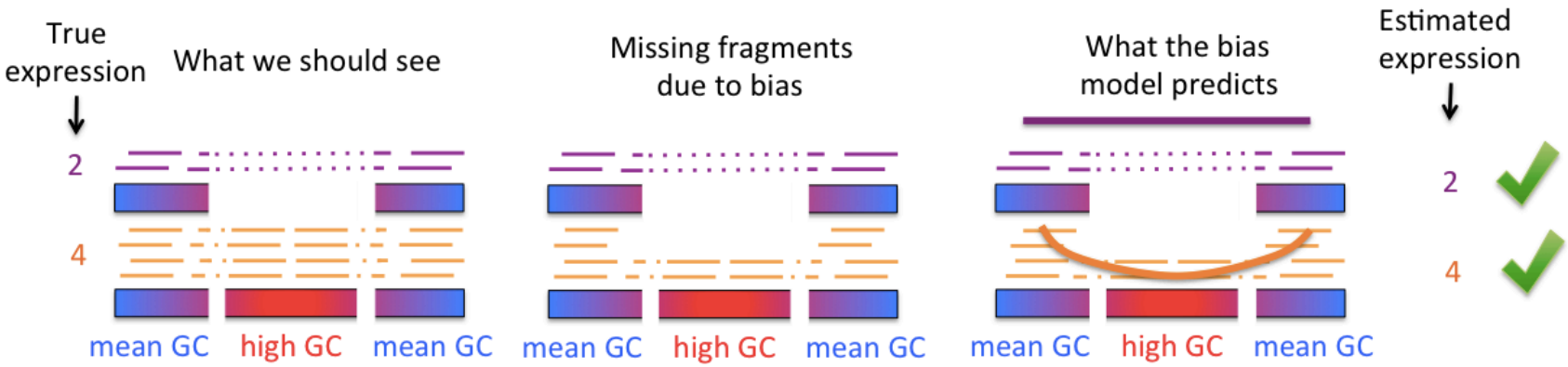
Alpine

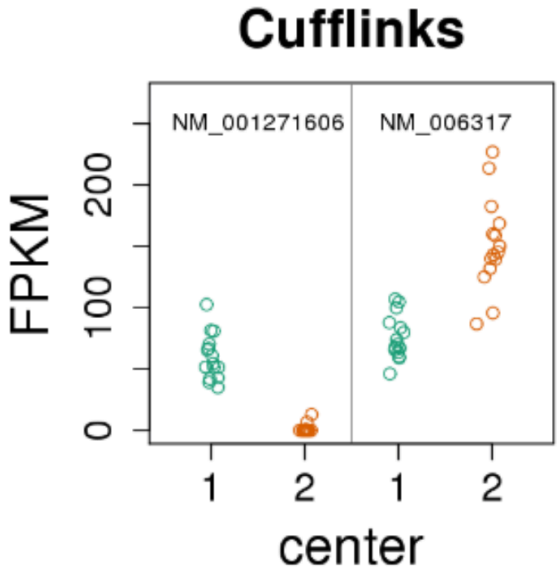
Alpine

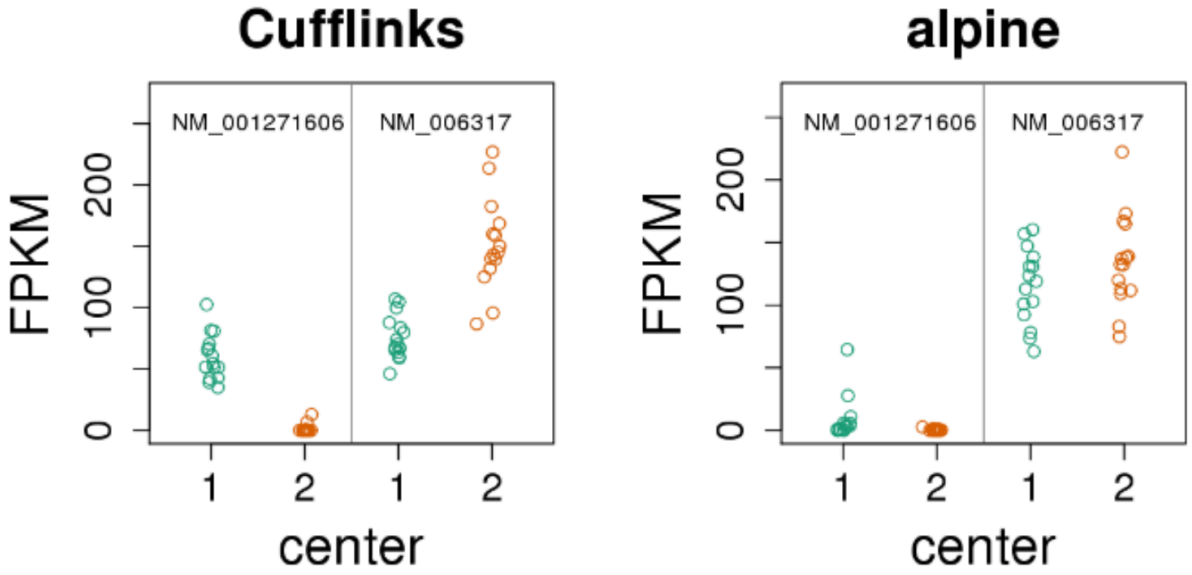
Better predict coverage artifacts



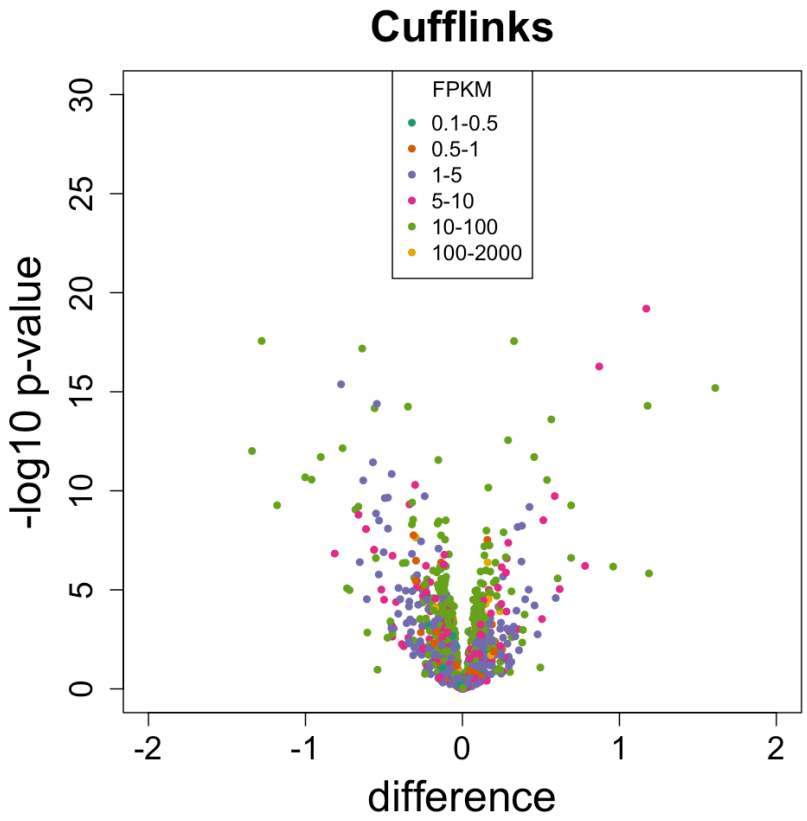
Parsimonious model

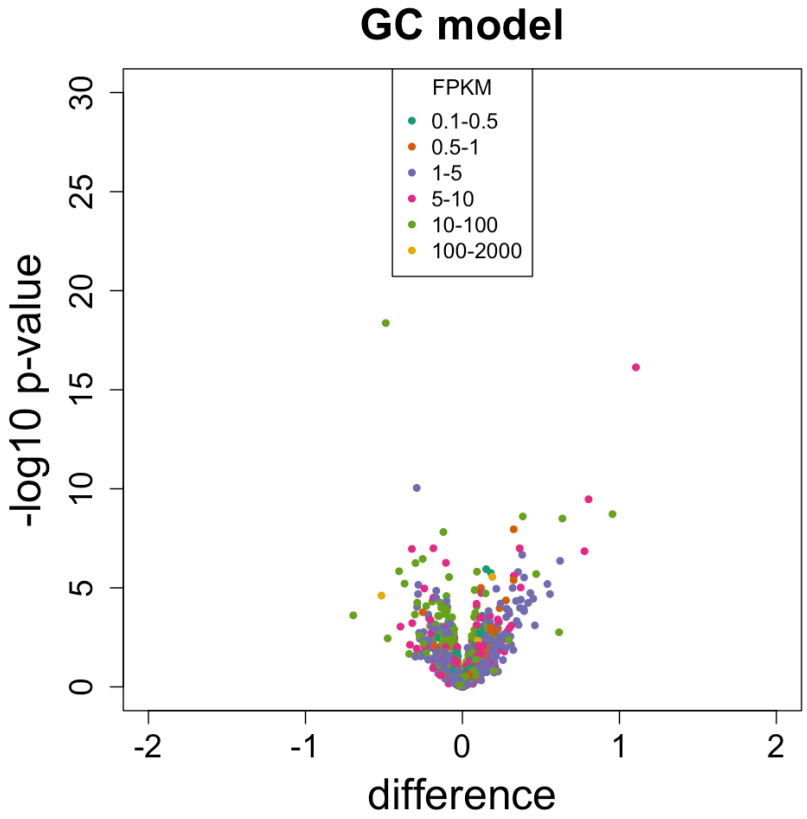



Cufflinks
NM_001271606
NM_006317
FPKM
0
100
200
1
2
1
2
center

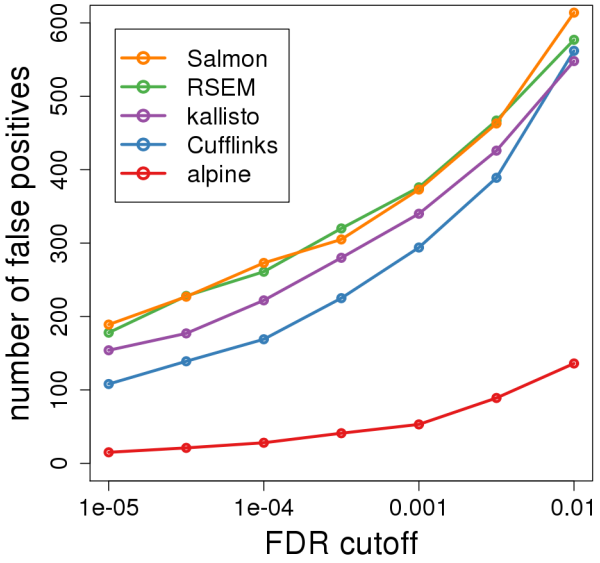

Cufflinks
alpine
NM_001271606
NM_006317
FPKM
0
100
200
1
2
center

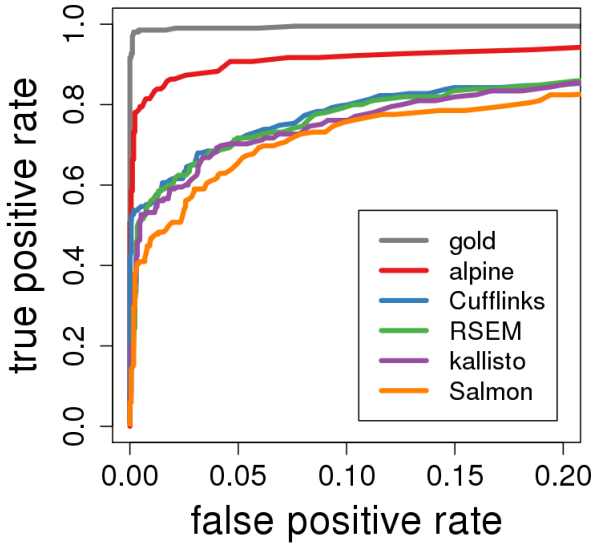
Current improvement




GC model
FPKM
0.1-0.5
0.5-1
1-5
5-10
10-100
100-2000
-log10 p-value
0
5
10
15
20
25
30
-2
-1
0
1
2
difference

More comparisons





MC simulations used for ROC

Acknowledgments

Michael Love

John Hogenesch

NIH R01 grants

- HG005220
- GM083084
- RR021967/GM103552

NIH P41 grant

- HG004059

http://rafalab.org @rafalab