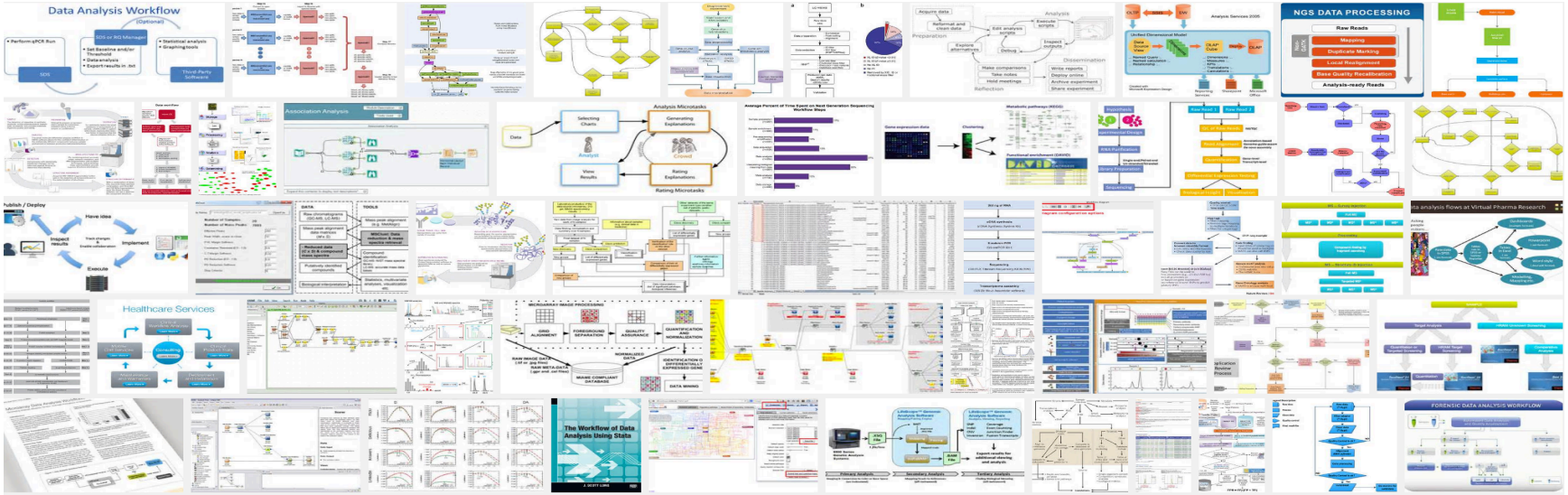
Bias, Systematic Error and Unwanted Variability in High-Throughput Technologies

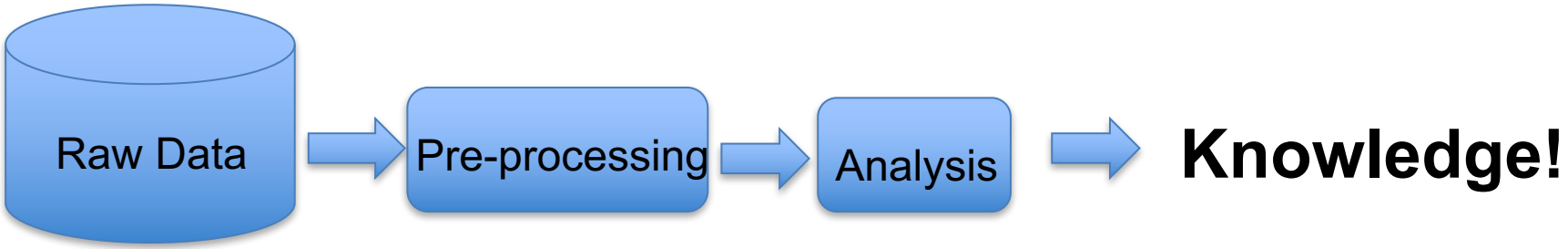
Rafael A Irizarry
Professor of Biostatistics and Computational Biology,
Dana Farber Cancer Institute
Professor of Biostatistics, Harvard School of Public Health

@rafalab

Workflows Everywhere



Typical Workflow

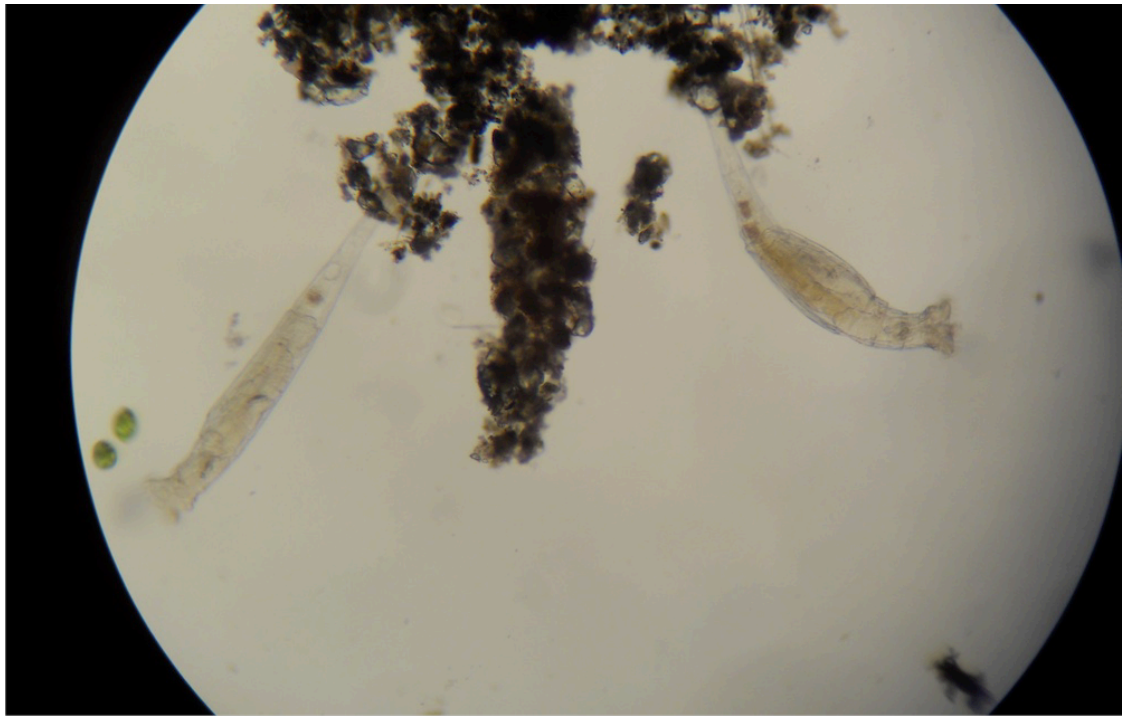


“Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the flexibility of the informed human mind...”

– John W. Tukey & Martin B. Wilk, Data Analysis & Statistics, 1966

Out of focus microscope

In Focus

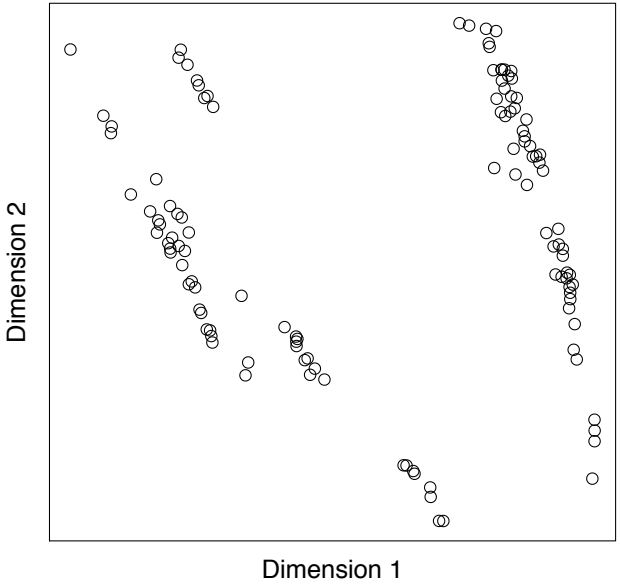


Another Tukey Quote

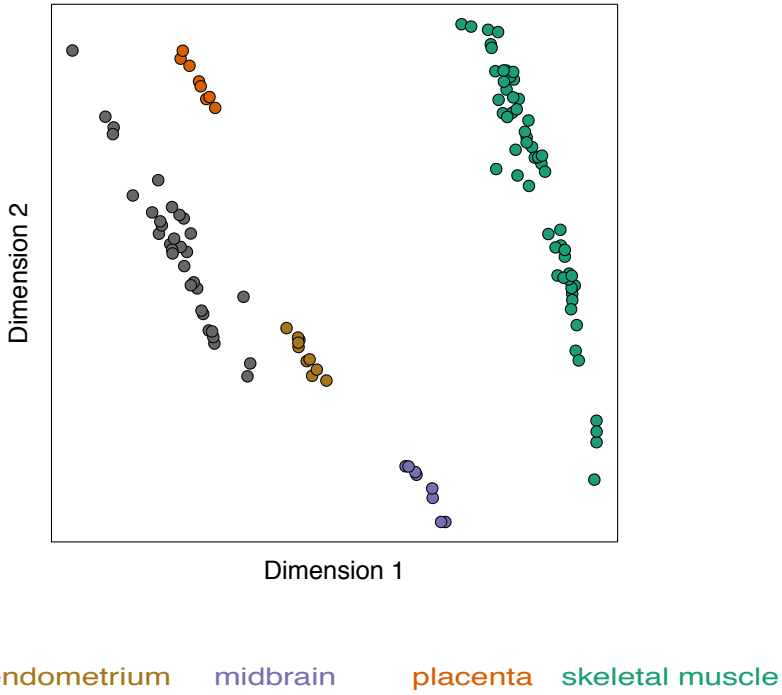
"The greatest value of a picture is when it forces us to notice what we never expected to see."

John W Tukey, EDA book

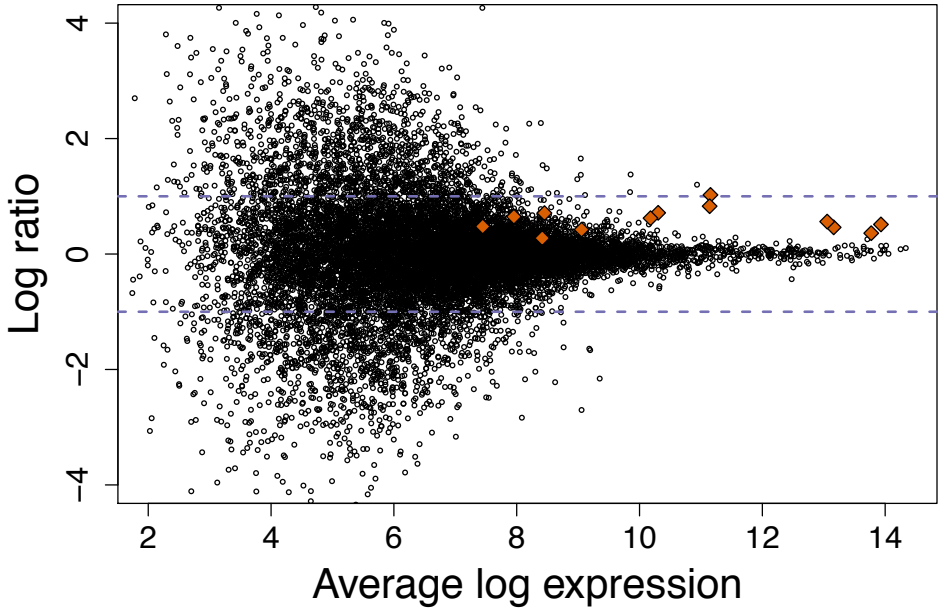
Data representing five tissues



Out of focus gene expression data



Why so much noise?



Deterministic approach

We want to estimate β, the hope is that:

$$Y_1 = \alpha + \beta$$

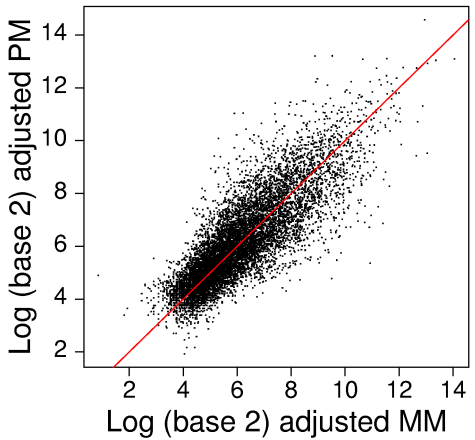
$$Y_2 = \alpha$$

$$Y_1 - Y_2 = \beta$$

But this is not correct!

Notice

- We care about ratios
- We usually take log of Y



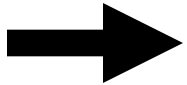
Stochastic approach

Better to assume:

$Y_1 = \alpha_1 + \beta$

$Y_2 = \alpha_2$

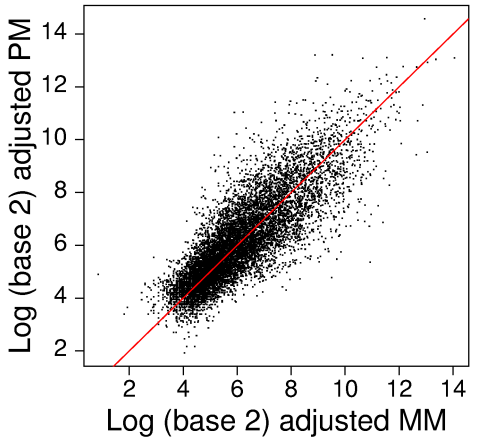
$Corr[\ log(\alpha_1), log(\alpha_2)] = 0.7$



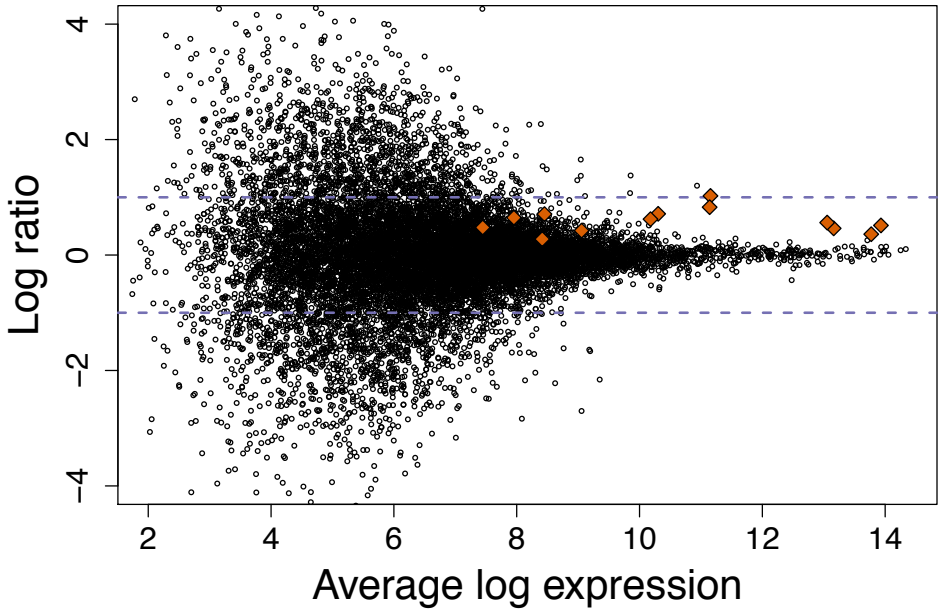
$Var[log(Y_1 - Y_2^*)] \sim 1/\beta^2$

Alternative solution:

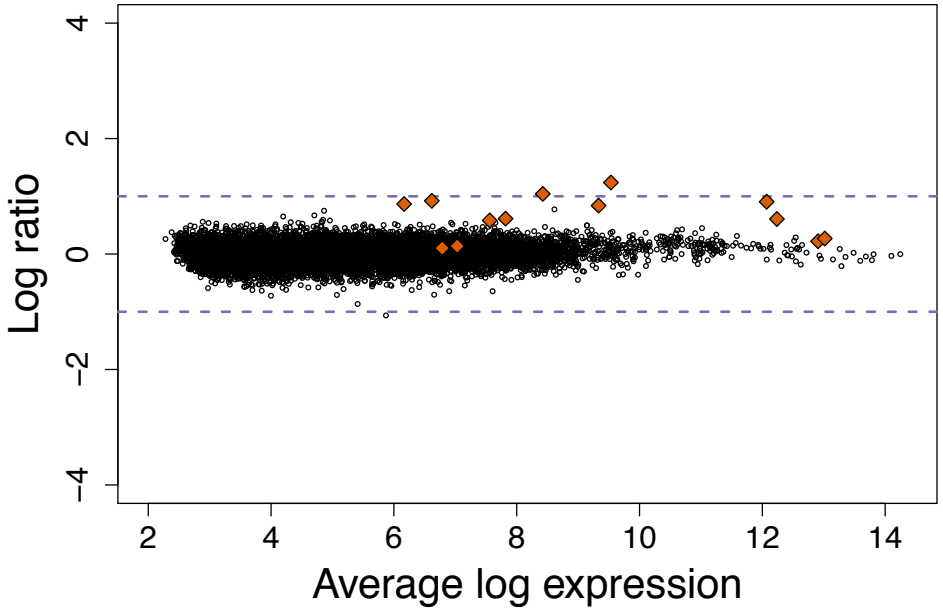
$E[\ \beta \mid Y_1, Y_2\ ]$



Why so much noise?

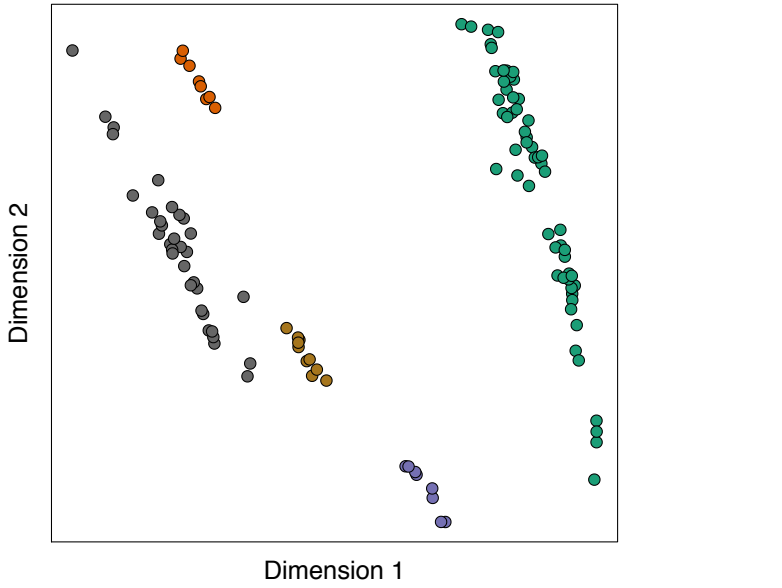


Our approach (RMA)

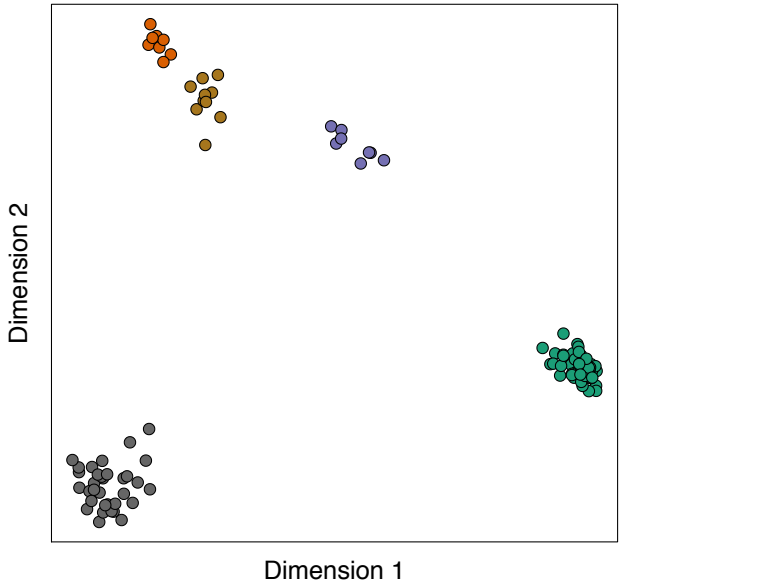


Irizarry et al. (2003) *Biostatistcs*

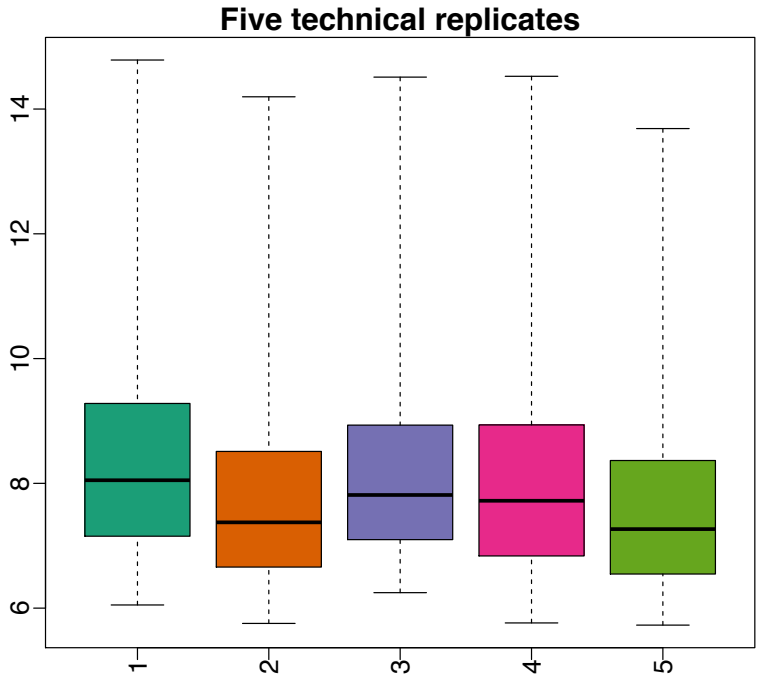
Out of focus


Dimension 2
Dimension 1
colon
endometrium
midbrain
placenta
skeletal muscle

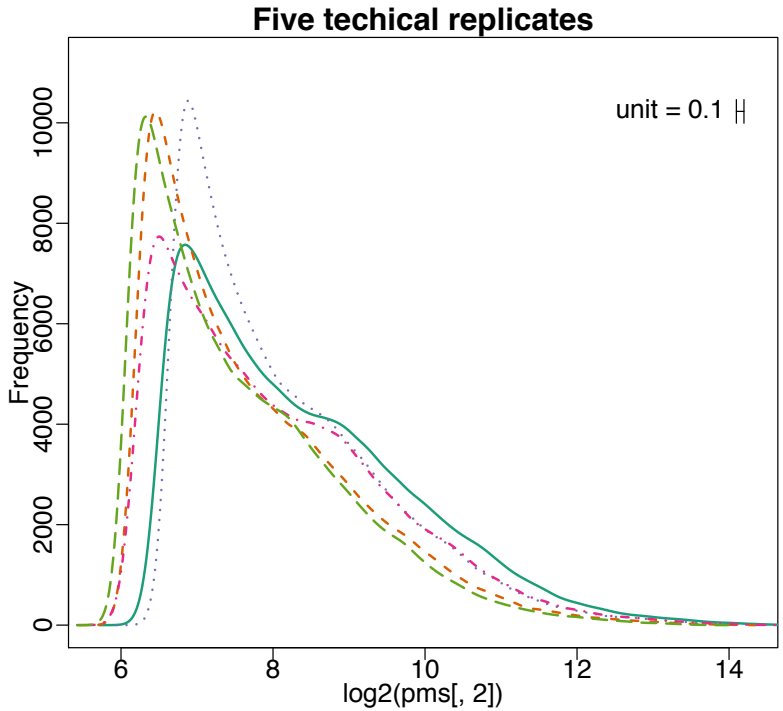
In focus


Dimension 2
Dimension 1
colon
endometrium
midbrain
placenta
skeletal muscle

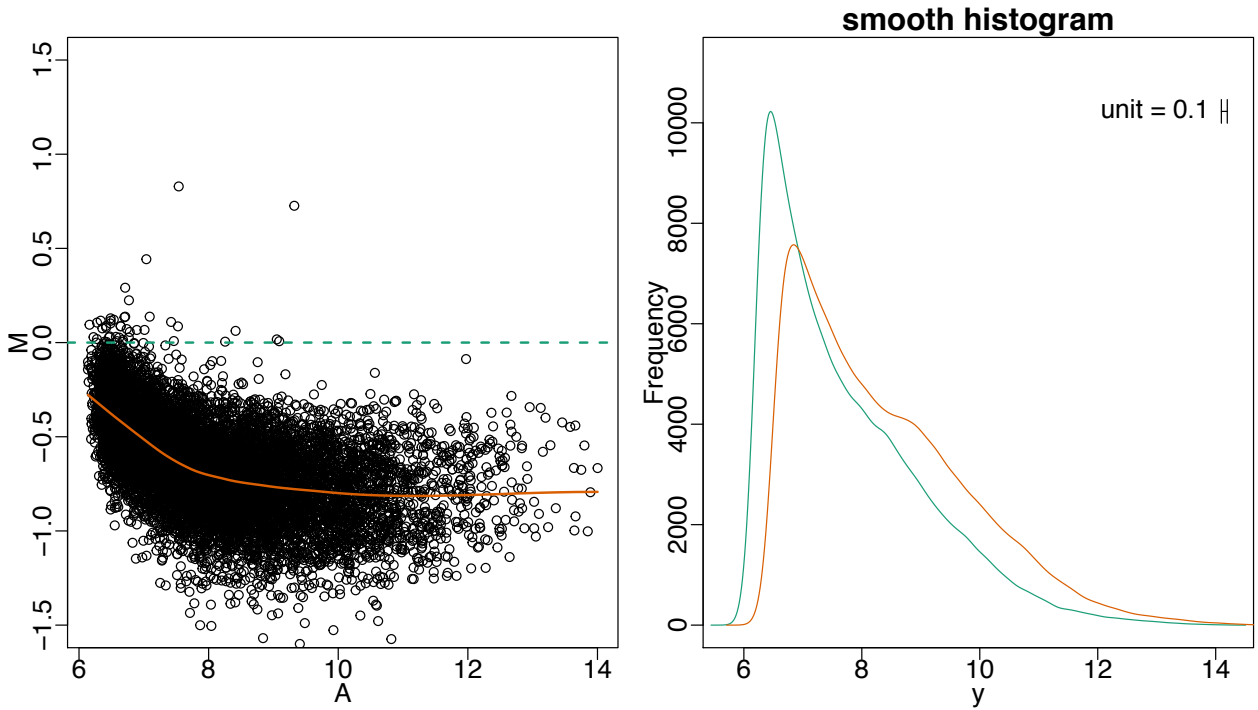
Values that are expected to be the same are not



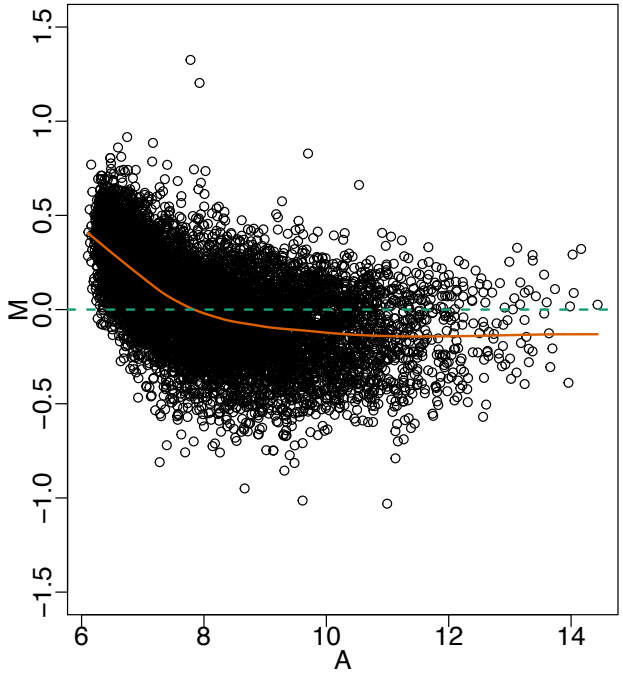
Distributions are different

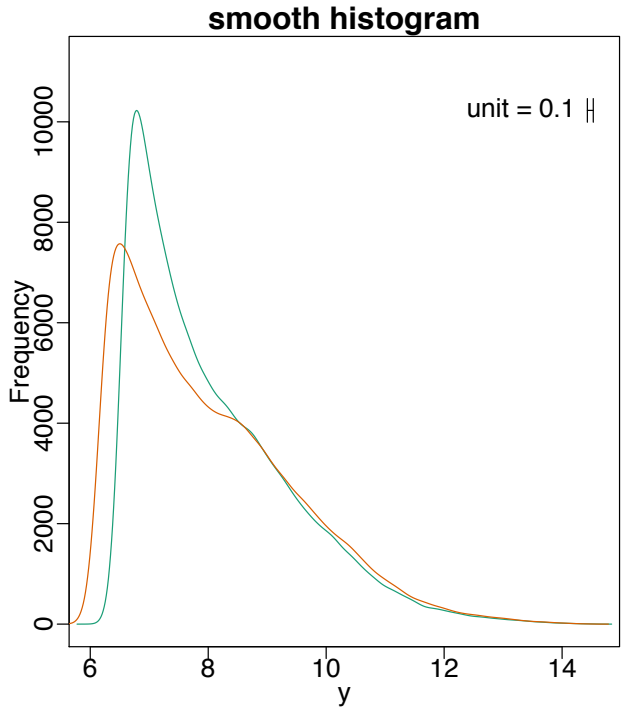


More than location and scale changes

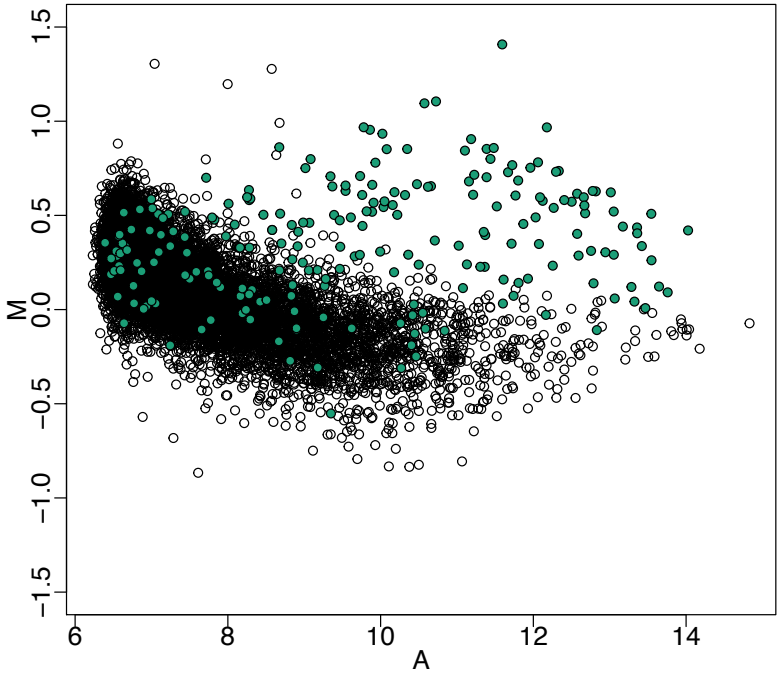


Median shifts do not solve the problem

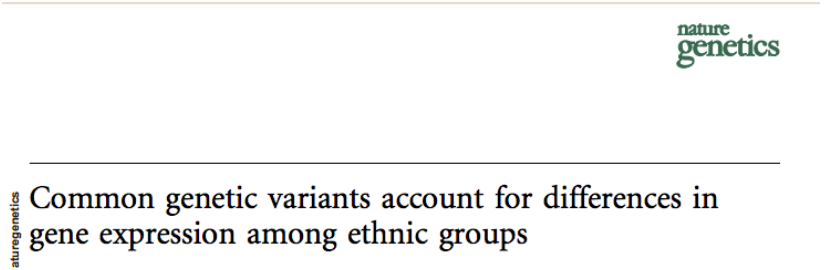




Non-linear effects

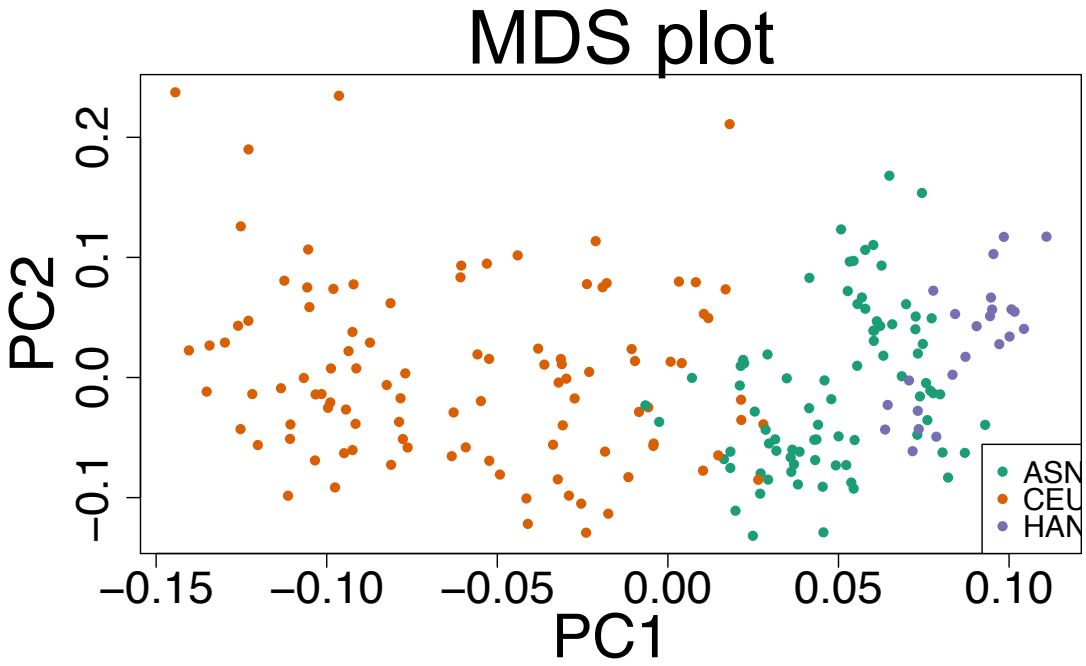


More data, more problems…

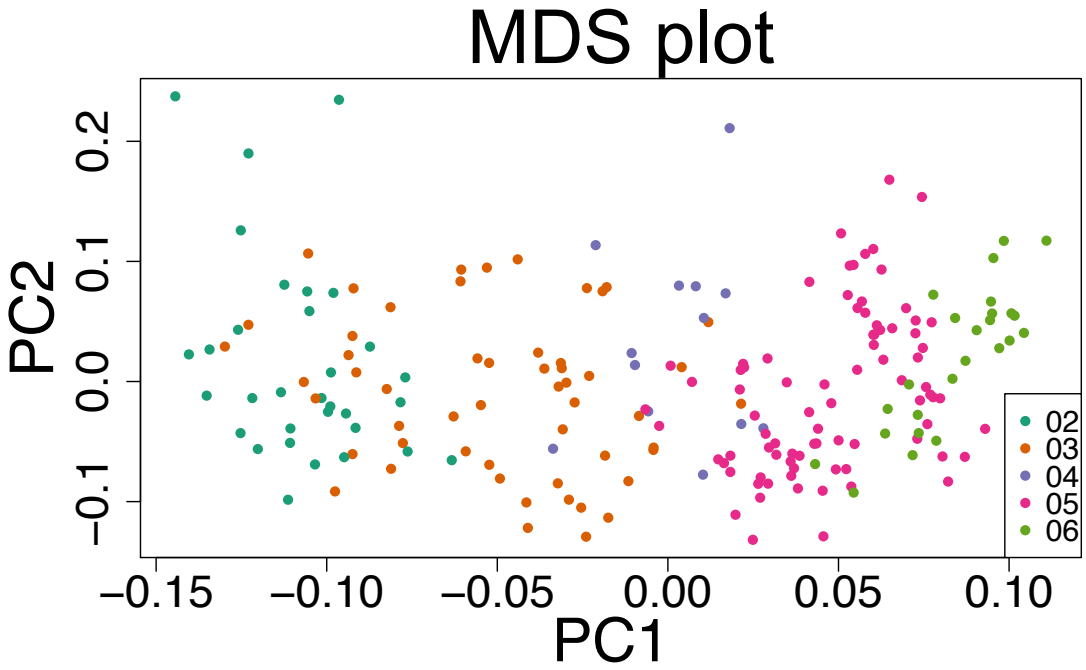


"This quantitative phenotype differs significantly between European derived and Asian-derived populations for 1,097 of 4,197 genes tested."

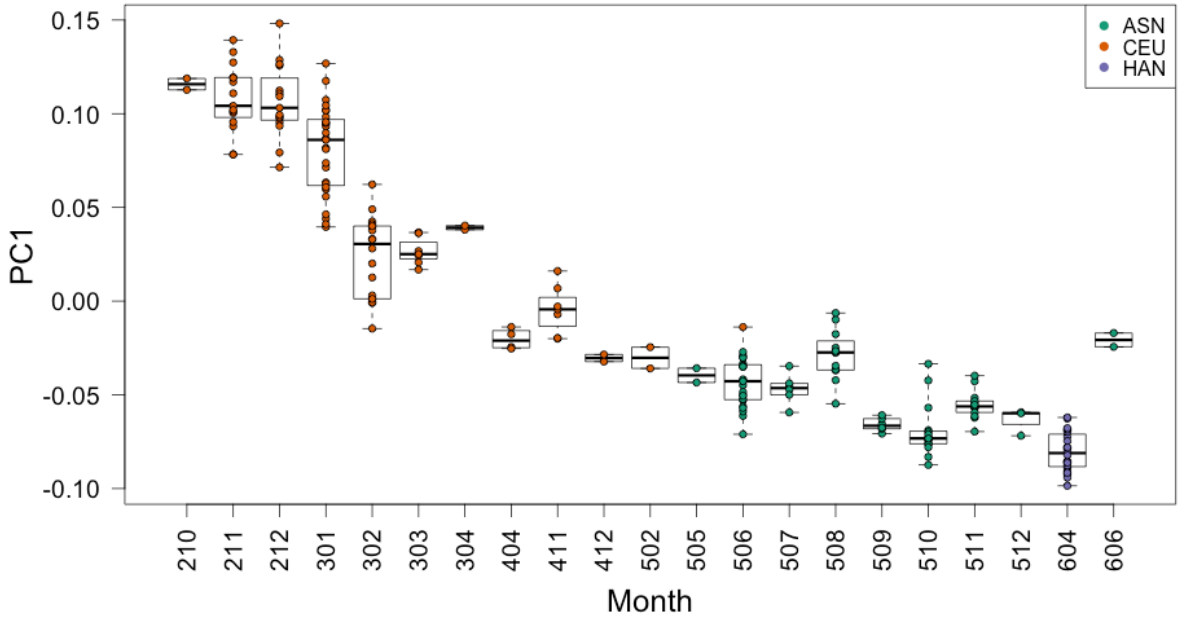
Blood RNA from different ethnicities



Now color is date



Boxplot of first PC by month



General problem

$$Y_{1,i} = \alpha + \varepsilon_{1,i}$$
$$Y_{2,i} = \alpha + \beta + \varepsilon_{2,i}$$

General problem

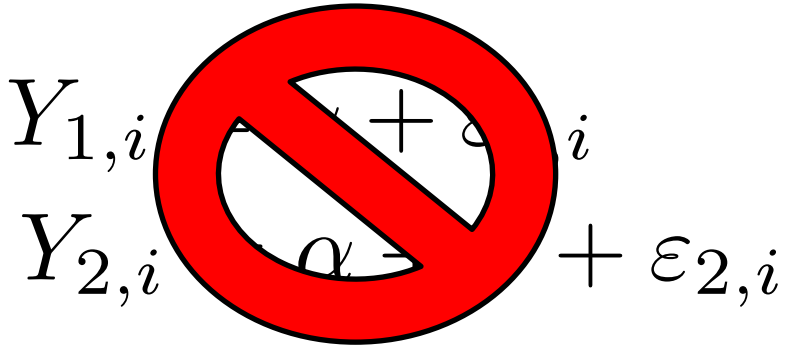
$$Y_{1,i} = \alpha + \varepsilon_{1,i}$$

$$Y_{2,i} = \alpha + \beta + \varepsilon_{2,i}$$

$$\bar{Y}_2 - \bar{Y}_1 \approx \beta$$

$$\bar{Y}_2 - \bar{Y}_1 \approx \beta \pm \sigma/\sqrt{N}$$

General problem



$$Y_{1,i} = \beta + W_1 + \varepsilon_{1,i}$$

$$Y_{2,i} = \alpha + \beta + W_2 + \varepsilon_{2,i}$$

$$Y_{1,i} = \beta + W_1 + \varepsilon_{1,i}$$

$$Y_{2,i} = \alpha + \beta + W_2 + \varepsilon_{2,i}$$

$$\bar{Y}_2 - \bar{Y}_1 \not\approx \beta \pm \sigma/\sqrt{N}$$

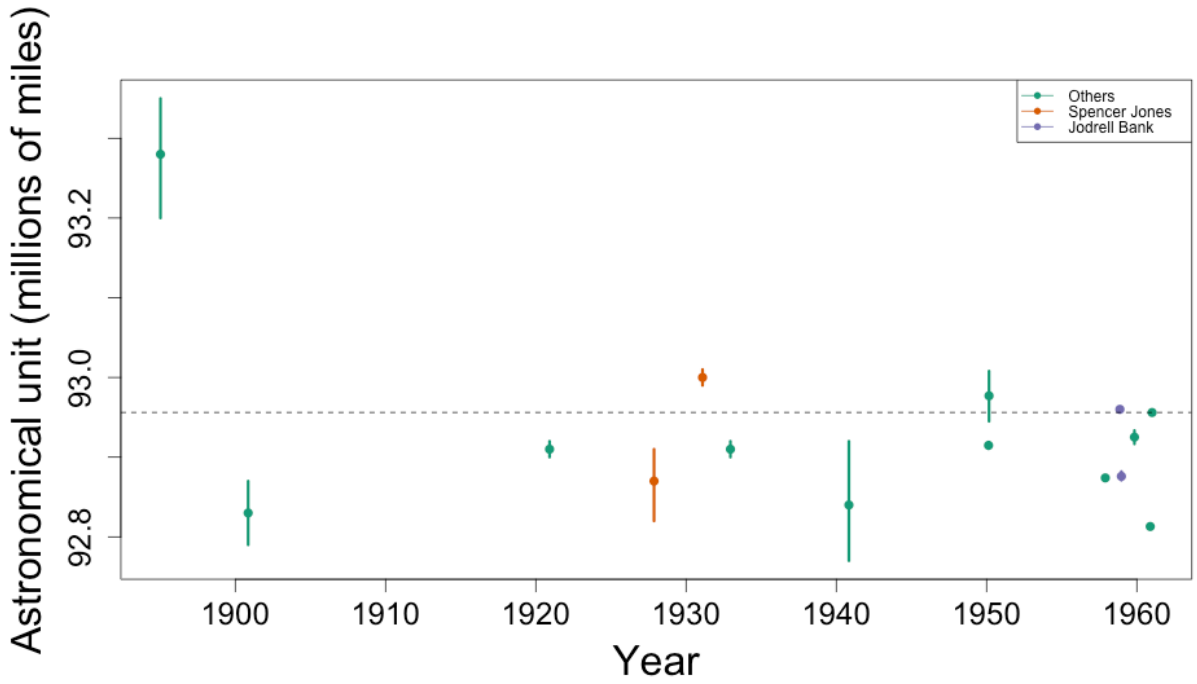
$$Y_{1,i} = \beta + W_1 + \varepsilon_{1,i}$$

$$Y_{2,i} = \alpha + \beta + W_2 + \varepsilon_{2,i}$$

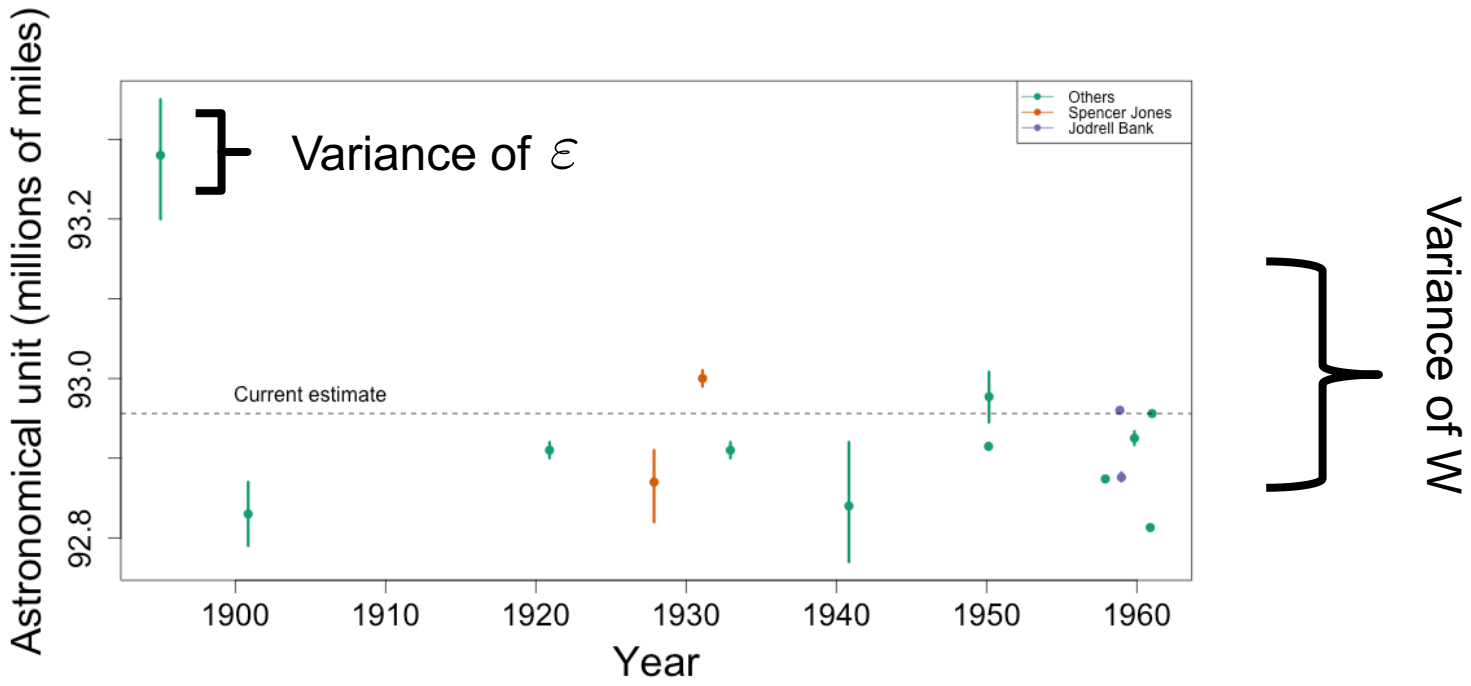
$$\bar{Y}_2 - \bar{Y}_1 \approx \beta + (W_2 - W_1) \pm \sigma/\sqrt{N}$$

$$\bar{Y}_2 - \bar{Y}_1 \approx \beta \pm (\sigma/\sqrt{N} + \text{variance due to } W)$$

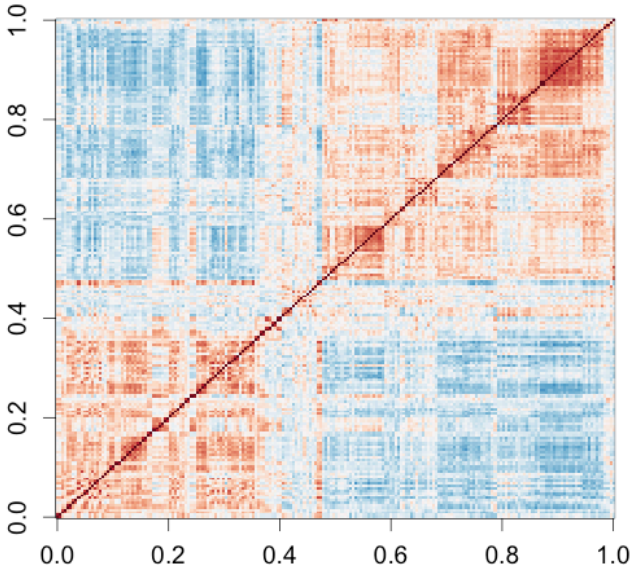
Batch Effects in Physics



Batch Effects in Physics



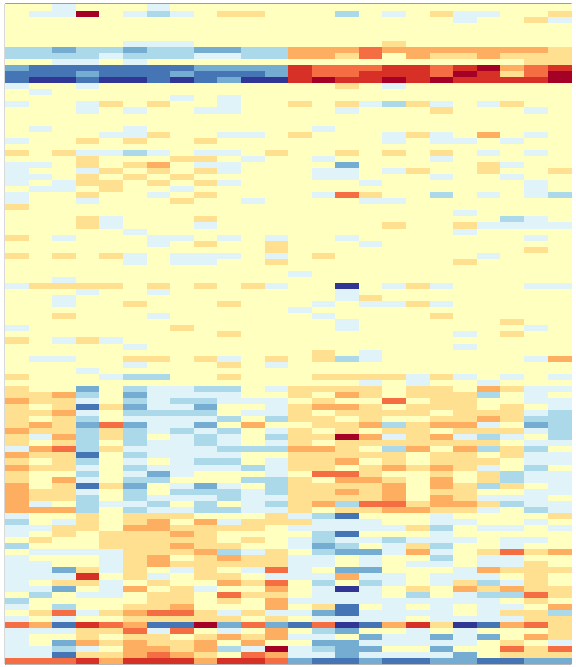
Motivates Factor Analysis Approach



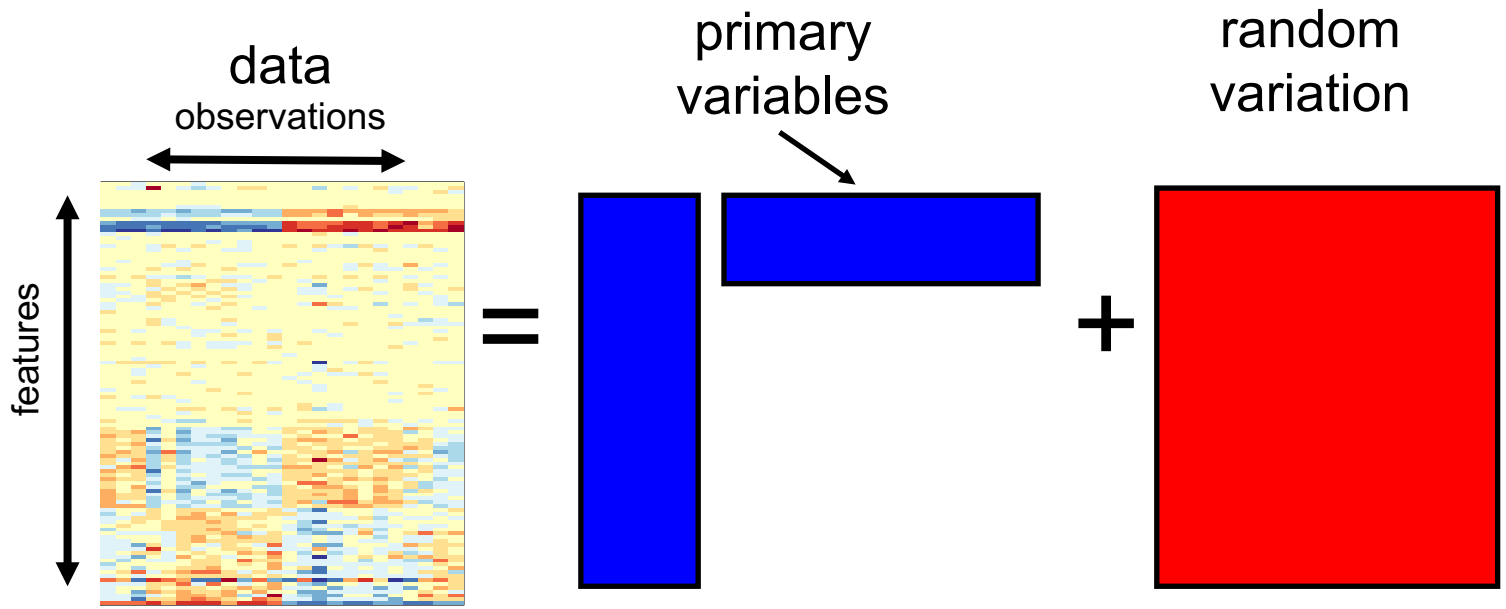
Leek and Storey 2007,2008
Listgarten et al. 2010
Gagnon-Bartsch and Speed- Biostatistics, 2012

12 males, 12 females, two months, 109 genes

	Female	Male
June 2005	3	9
October 2005	9	3

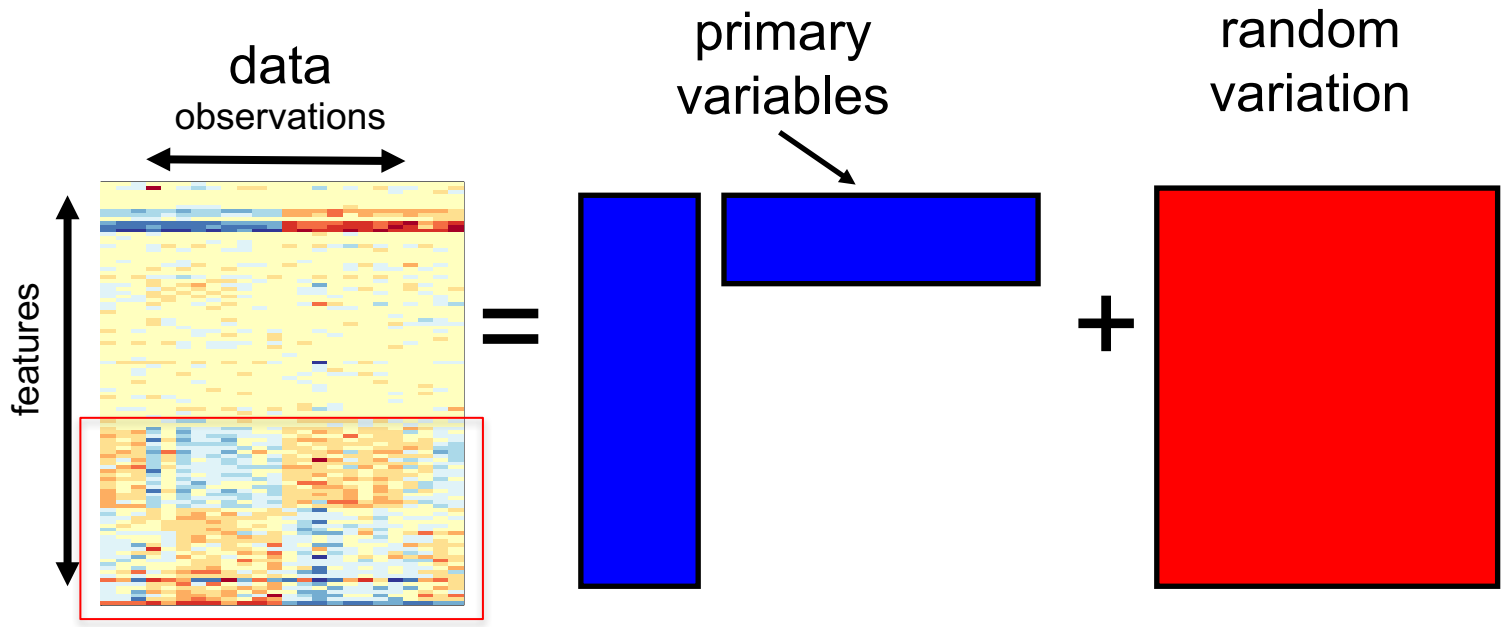


Decomposing variability



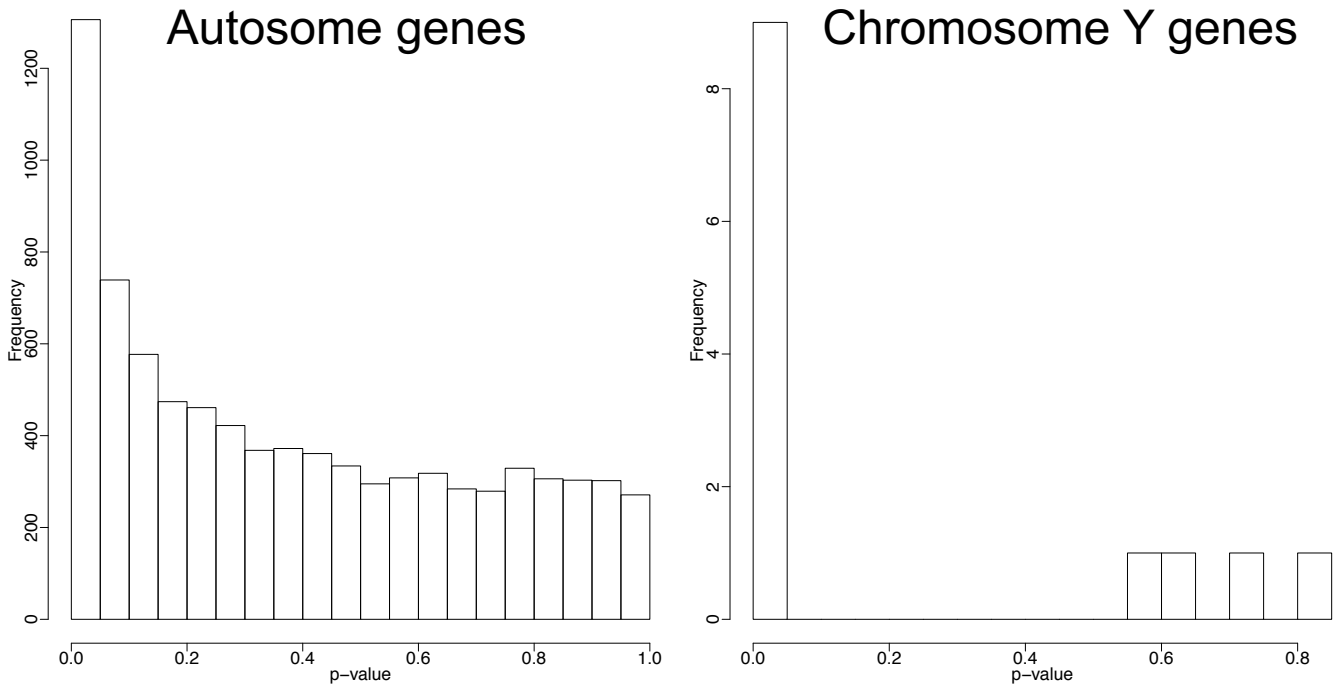
$$Y_{m\times n} = \beta_{m\times p} X_{p\times n} + \varepsilon_{m\times n}$$

This model does not account for batch

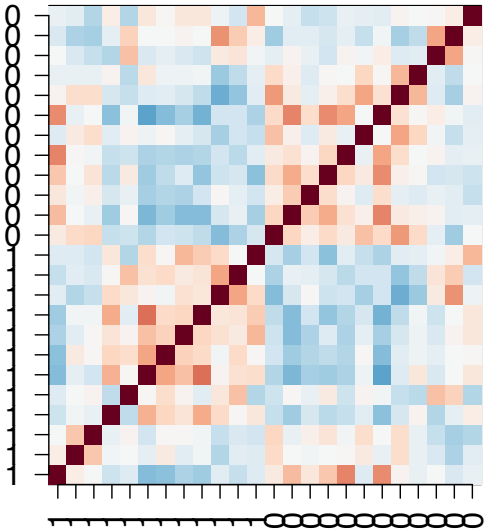


$$Y_{m,n} = \beta_{m,p} X_{p,n} + \varepsilon_{m,n}$$

p-value histograms

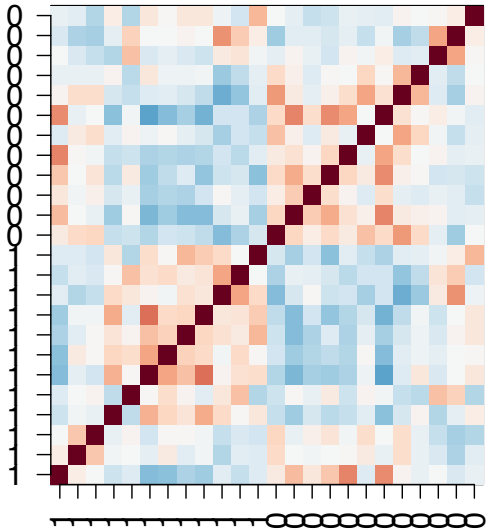


Sample correlations

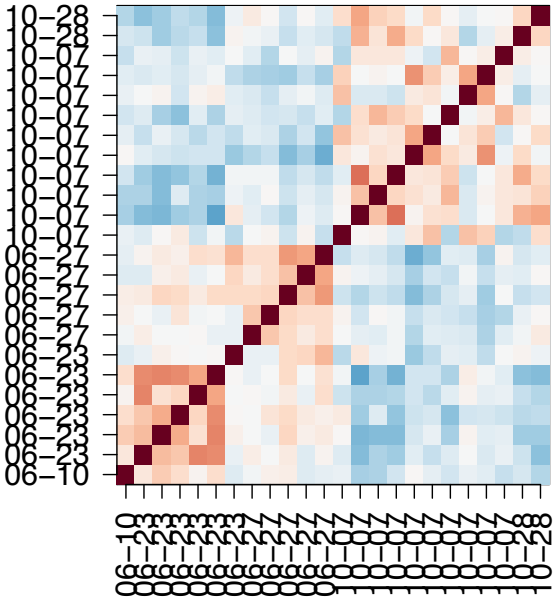


Sample correlations

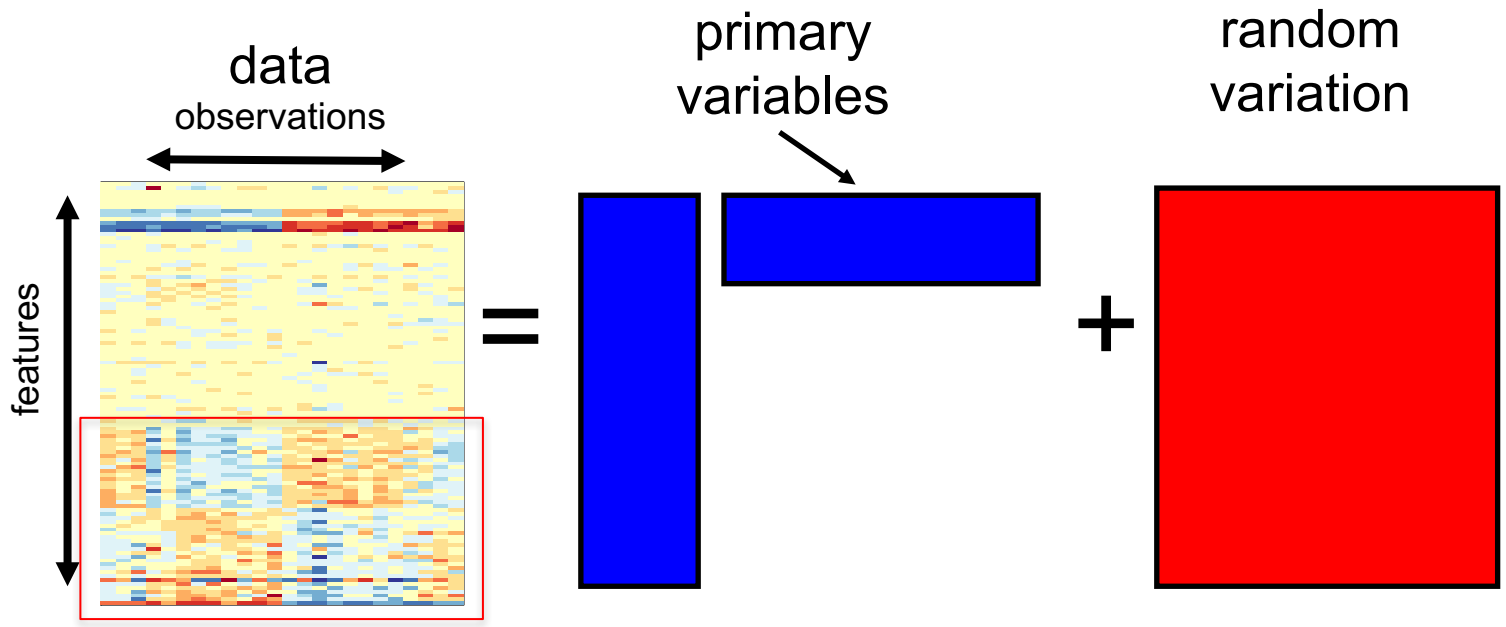
Original



Order by date

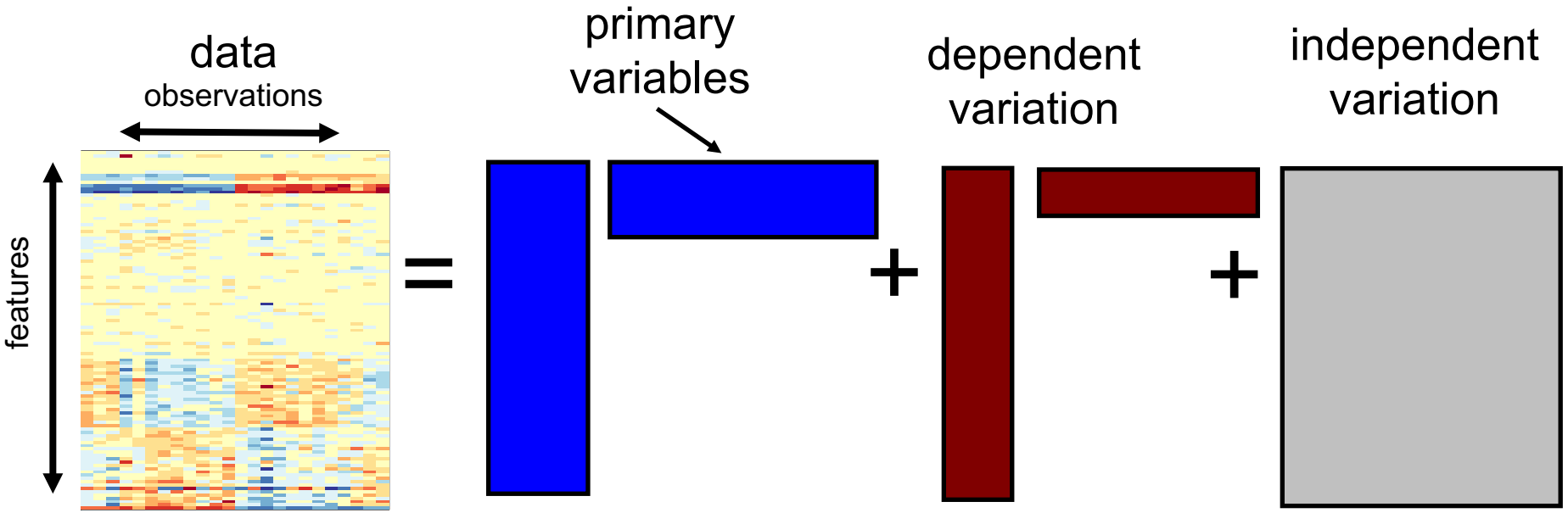


Note structure



$$Y_{m\times n} = \beta_{m\times p} X_{p\times n} + \varepsilon_{m\times n}$$

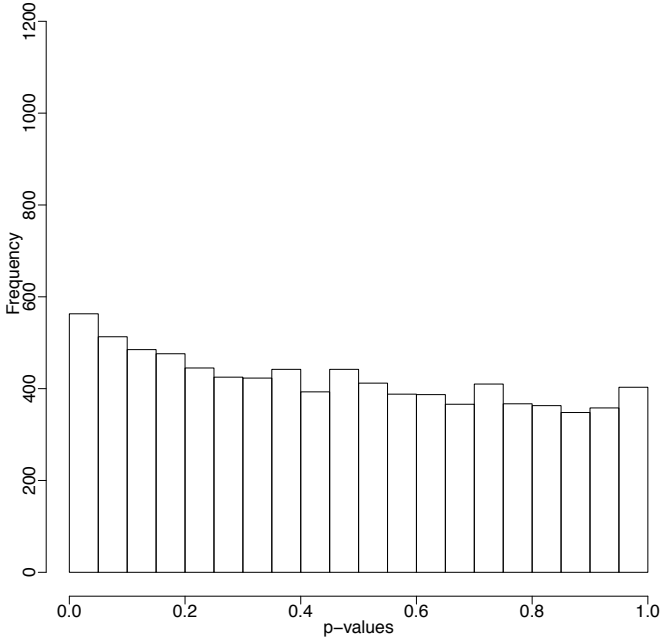
Factor analysis

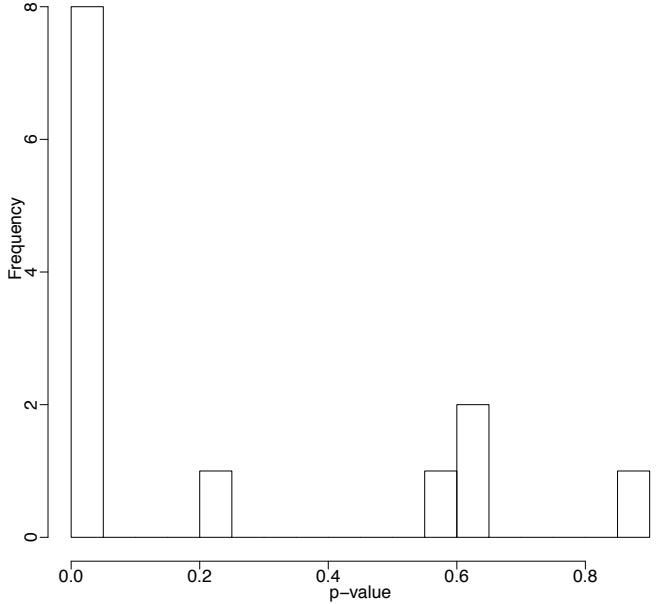


$$Y_{m\times n} = \beta_{m\times p} X_{p\times n} + \alpha_{m\times k} W_{k\times n} + \varepsilon_{m\times n}$$

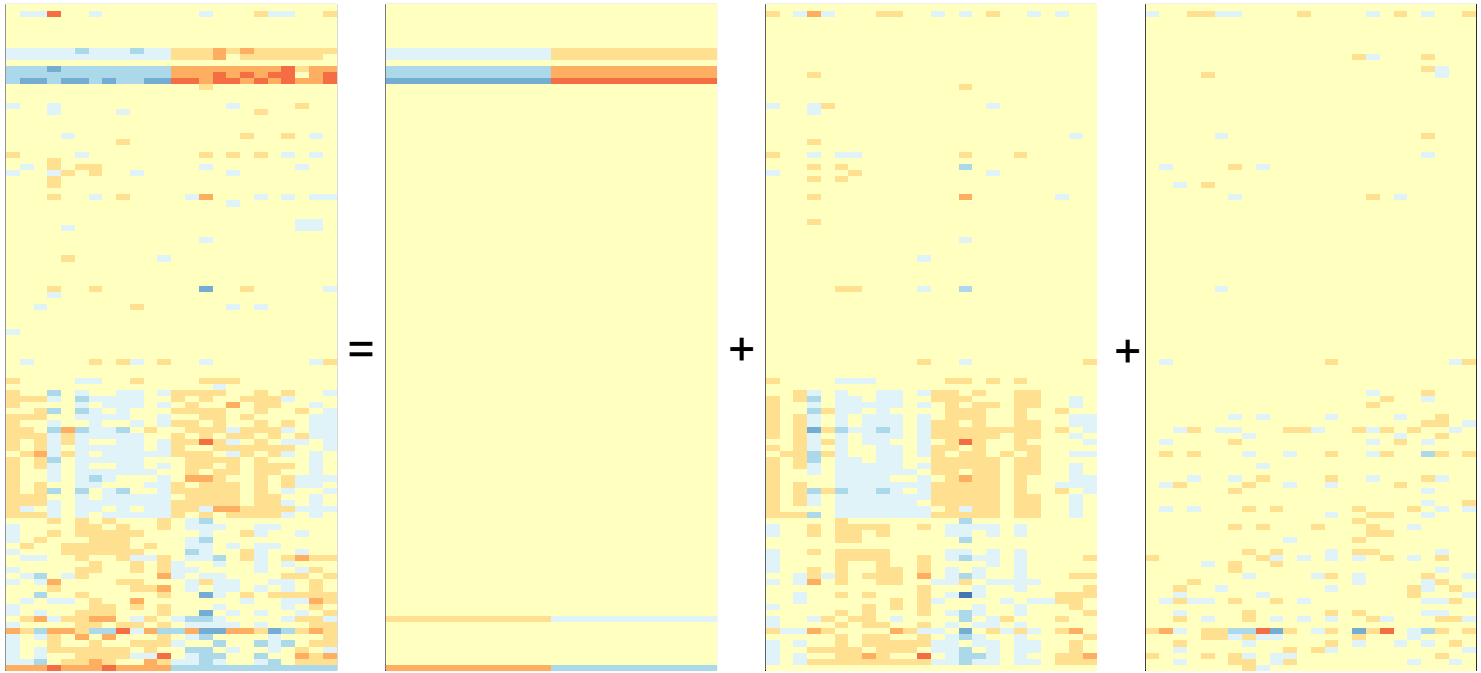
Leek and Storey 2007,2008

After Factor Analysis





Decomposed data



Sites that change with age ?

Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, et al: **Human aging-associated DNA hypermethylatio occurs preferentially at bivalent chromatin domains.** *Genome Res* 2010, **20:**434-439.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, et al: **Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.** *Genome Res* 2010, **20:**440-446.

Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST: **Age-associated DNA methylation in pediatric populations.** *Genome Res* 2012, **22:**623-632.

Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, et al: **Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.** *PLoS Genet* 2012, **8:**e1002629.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, et al: **Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates.** *Mol Cell* 2012.

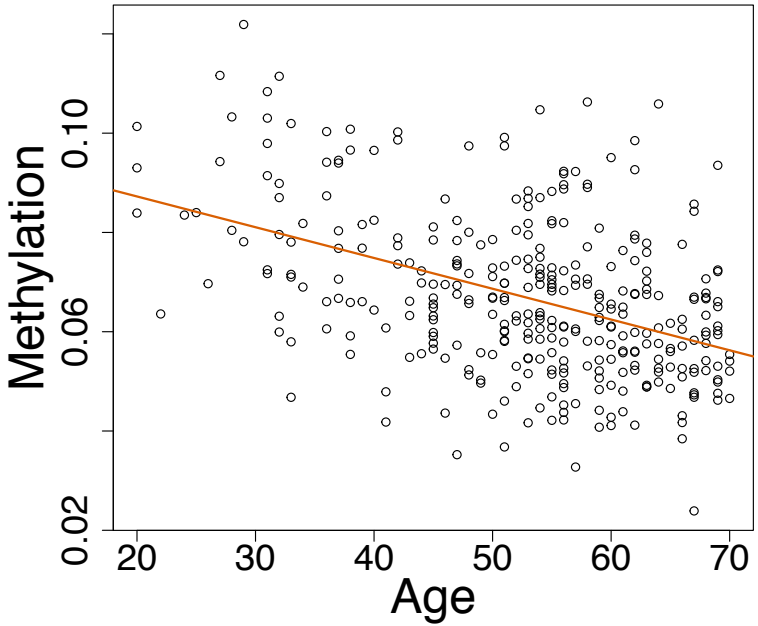
Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, et al: **Distinct DNA methylomes of newborns and centenarians.** *Proc Natl Acad Sci U S A* 2012, **109:**10522-10527.

Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, van den Berg LH, Ophoff RA: **Aging effects on DNA methylation modules in human brain and blood tissue.** *Genome Biol* 2012, **13:**R97.

Lee H, Jaffe AE, Feinberg JI, Tryggvadottir R, Brown S, Montano C, Aryee MJ, Irizarry RA, Herbstman J, Witter FR, et al: **DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth.** *Int J Epidemiol* 2012, **41:**188-199.

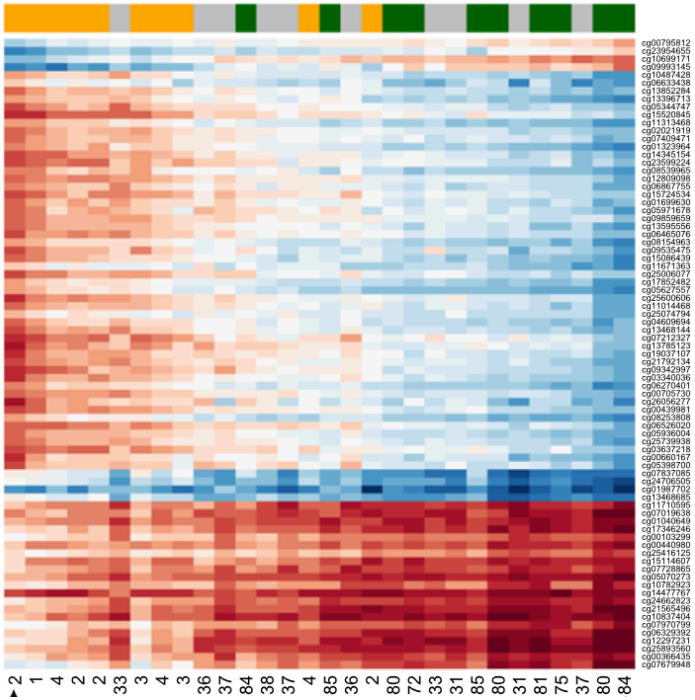
Johansson A, Enroth S, Gyllensten U: **Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan.** *PLoS One* 2013, **8:**e67378.

DNA methylation correlates with Age for this CpG

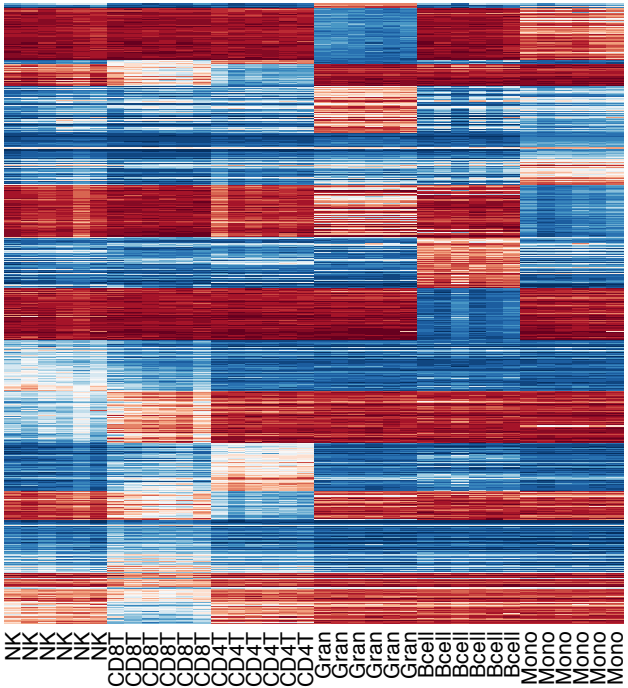


Data from GSE32148

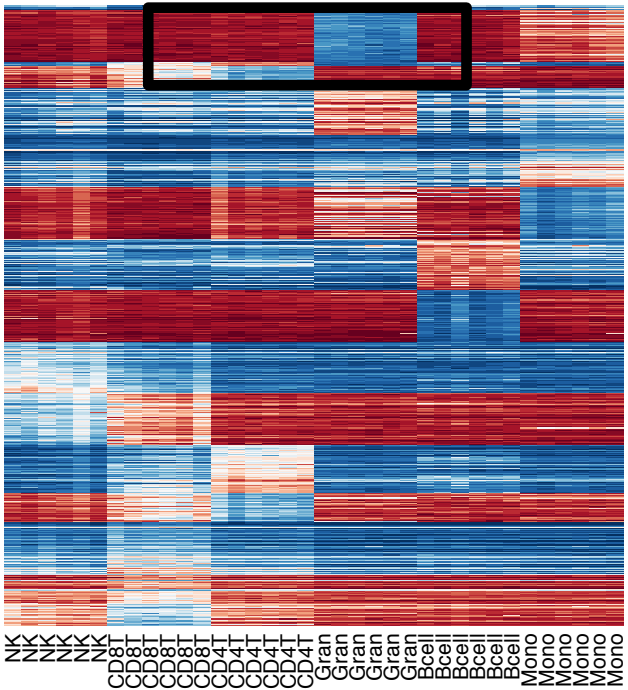
Whole Blood


cg00795812
cg23954655
cg10699171
cg09993145
cg10487428
cg06633438
cg13852284
cg13396713
cg05344747
cg15520845
cg11313468
cg02021919
cg07409471
cg01323964
cg14345154
cg23599224
cg08539965
cg12809098
cg06867755
cg15724534
cg01699630
cg05971678
cg09859659
cg13595556
cg06465076
cg08154963
cg09535475
cg15086439
cg11671363
cg25006077
cg17852482
cg05627557
cg25600606
cg11014468
cg25074794
cg04609694
cg13468144
cg07212327
cg13785123
cg19037107
cg21792134
cg09342997
cg03340036
cg06270401
cg00705730
cg26056277
cg00439981
cg08253808
cg06526020
cg05936004
cg25739938
cg03637218
cg00660167
cg05398700
cg07837085
cg24706505
cg01987702
cg13468685
cg11710595
cg07019638
cg01040649
cg17346246
cg00103299
cg00440980
cg25416125
cg15114607
cg07728865
cg05070273
cg10782923
cg14477767
cg24662623
cg21565496
cg10837404
cg07970799
cg06329392
cg12297231
cg25893560
cg00366435
cg07679948
2
1
4
2
2
33
3
4
3
36
37
84
38
37
4
85
36
2
80
72
33
31
85
80
31
81
75
37
80
84

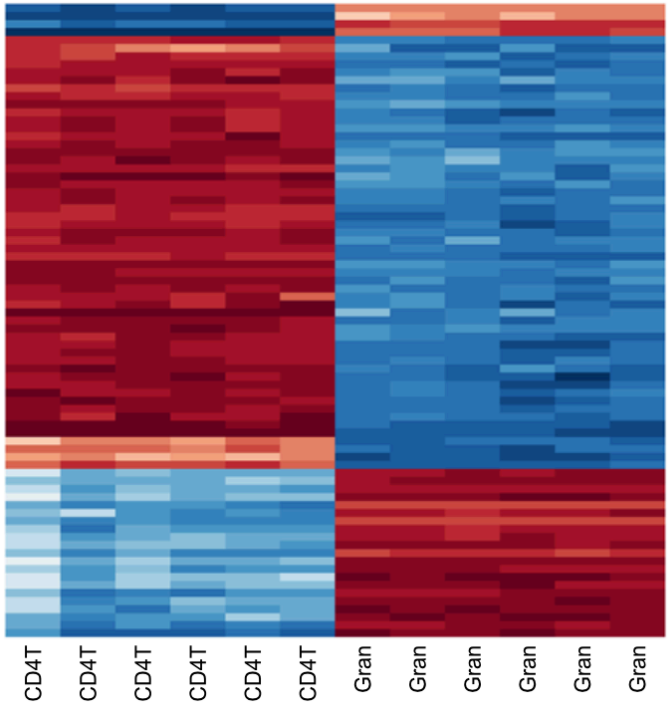
Blood is a mixture of many cell types



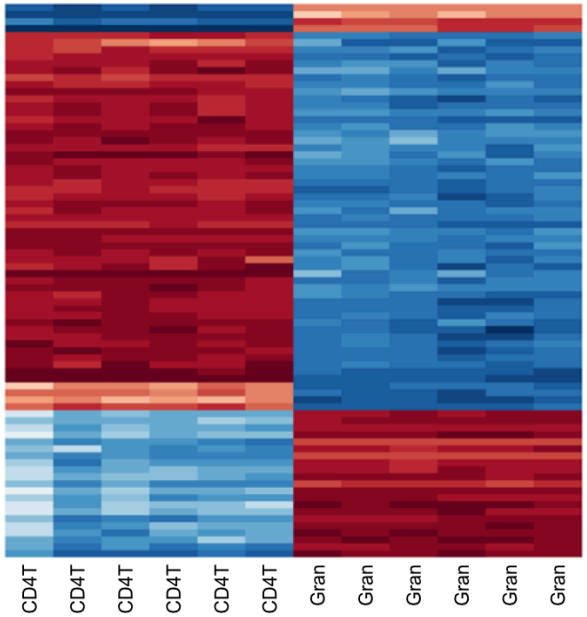
Blood is a mixture of many cell types



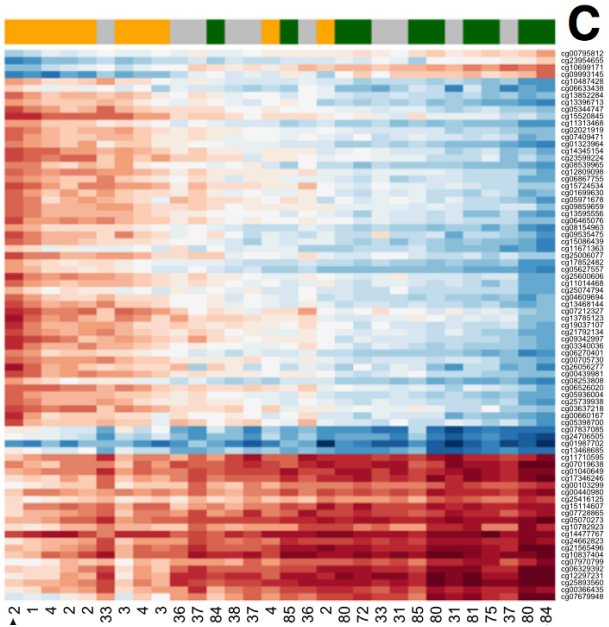
Close up



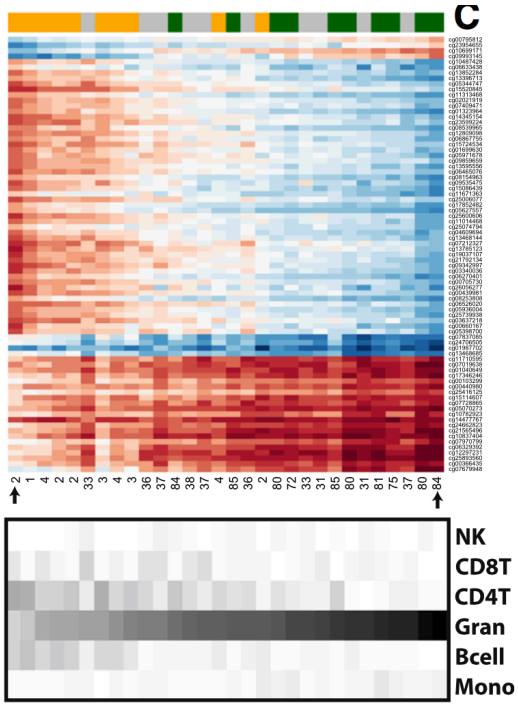
Sorted cell types



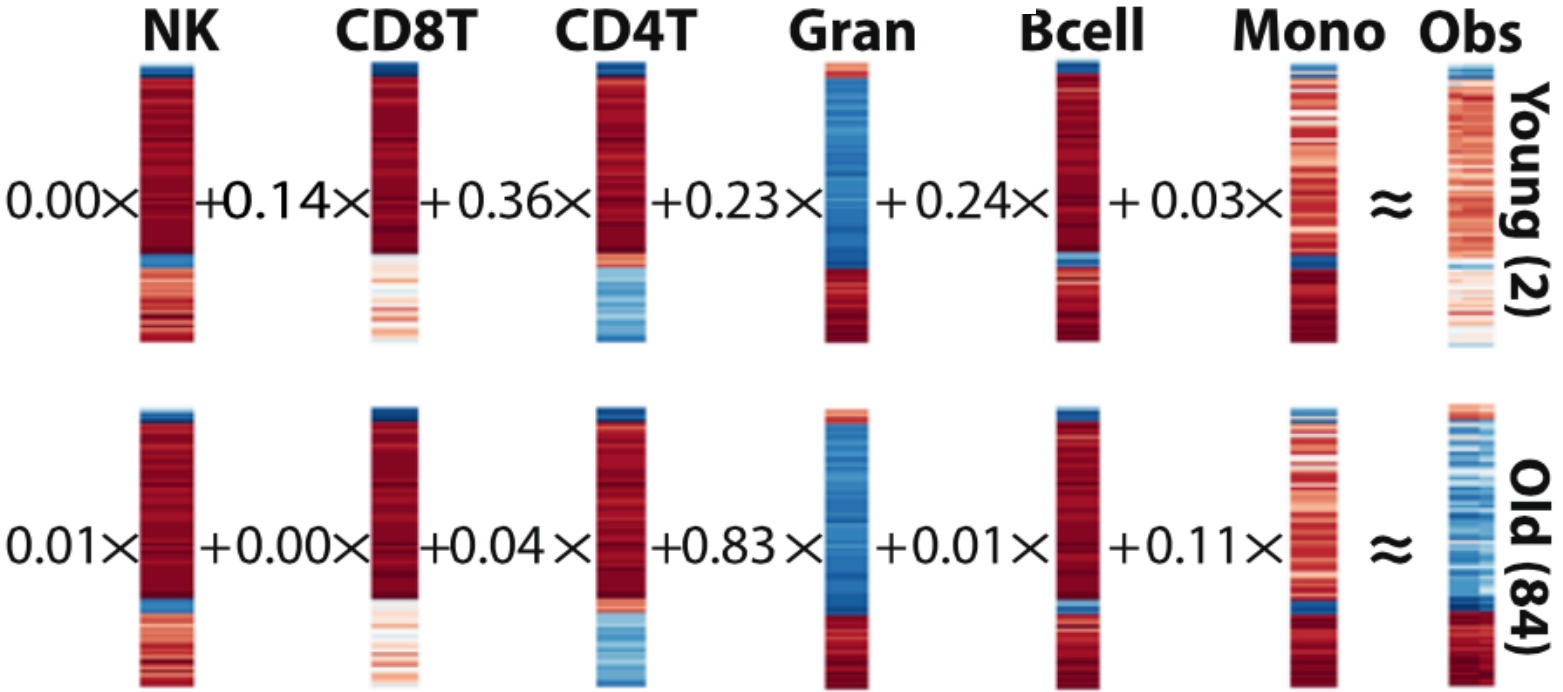
Whole blood



Cell composition changes with age

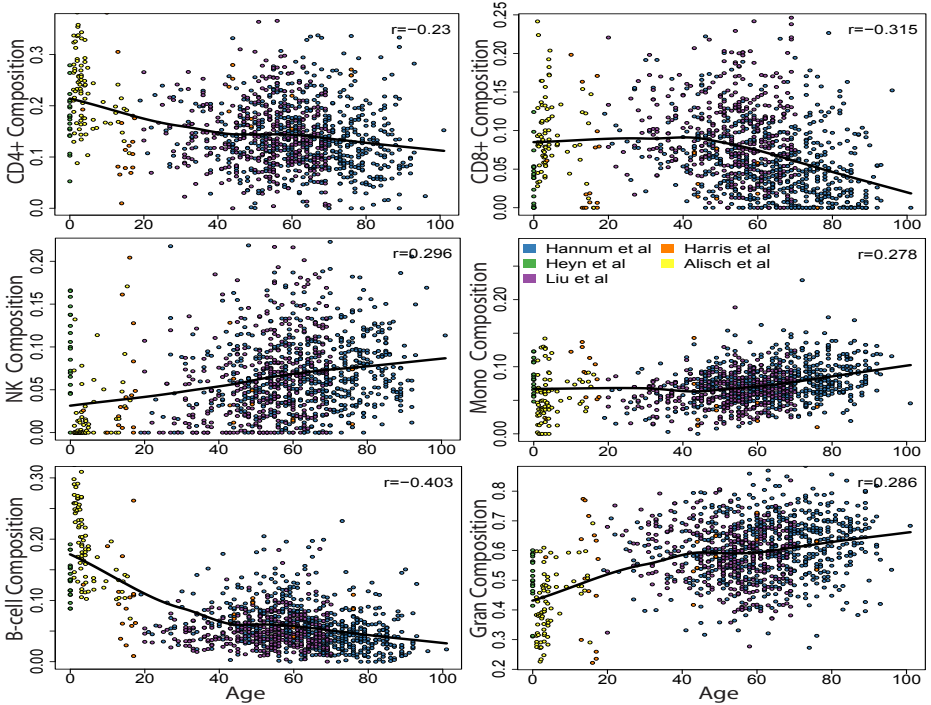


Different composition implies different profile



Jaffe and Irizarry (2014) Genome Biology

Cell composition versus age



Confounding differences in brain methylation

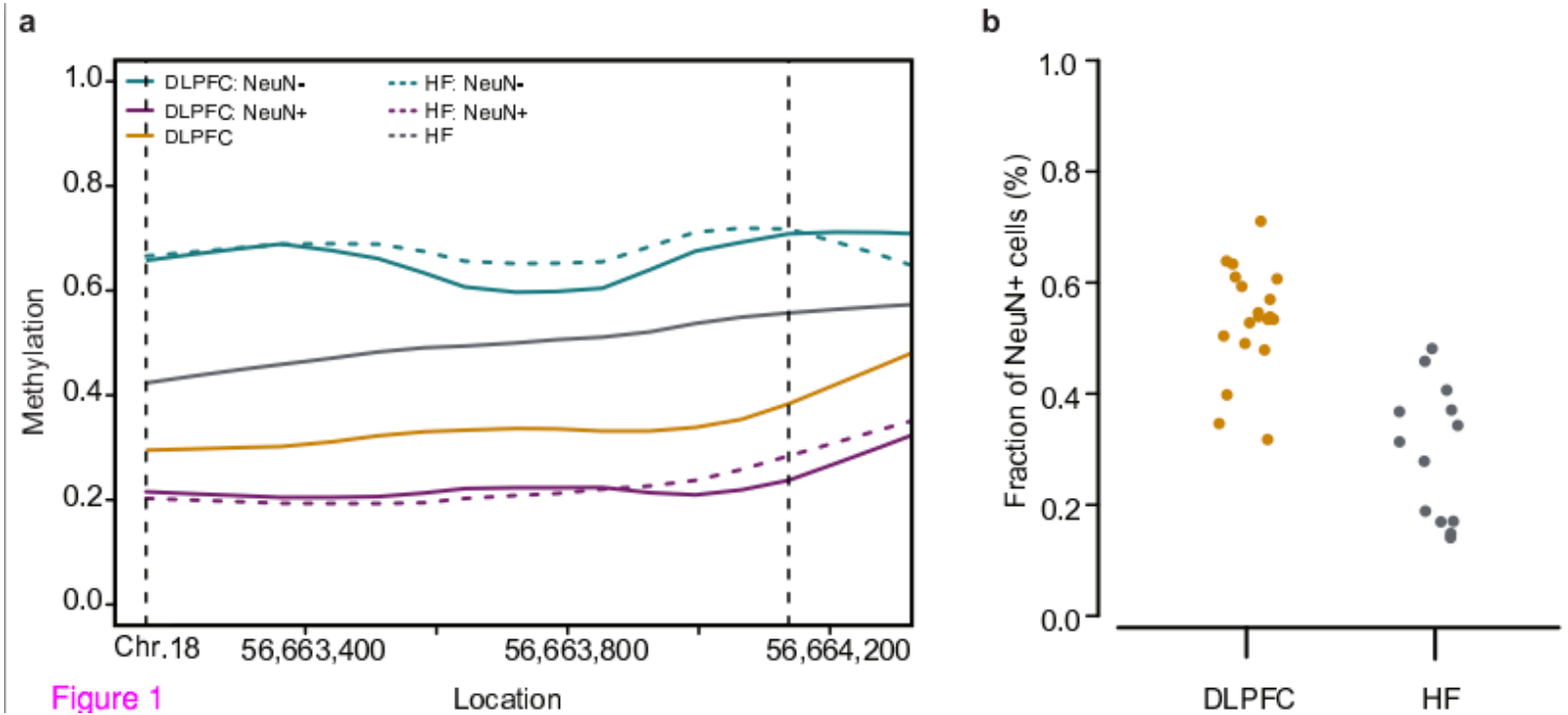


Figure 1

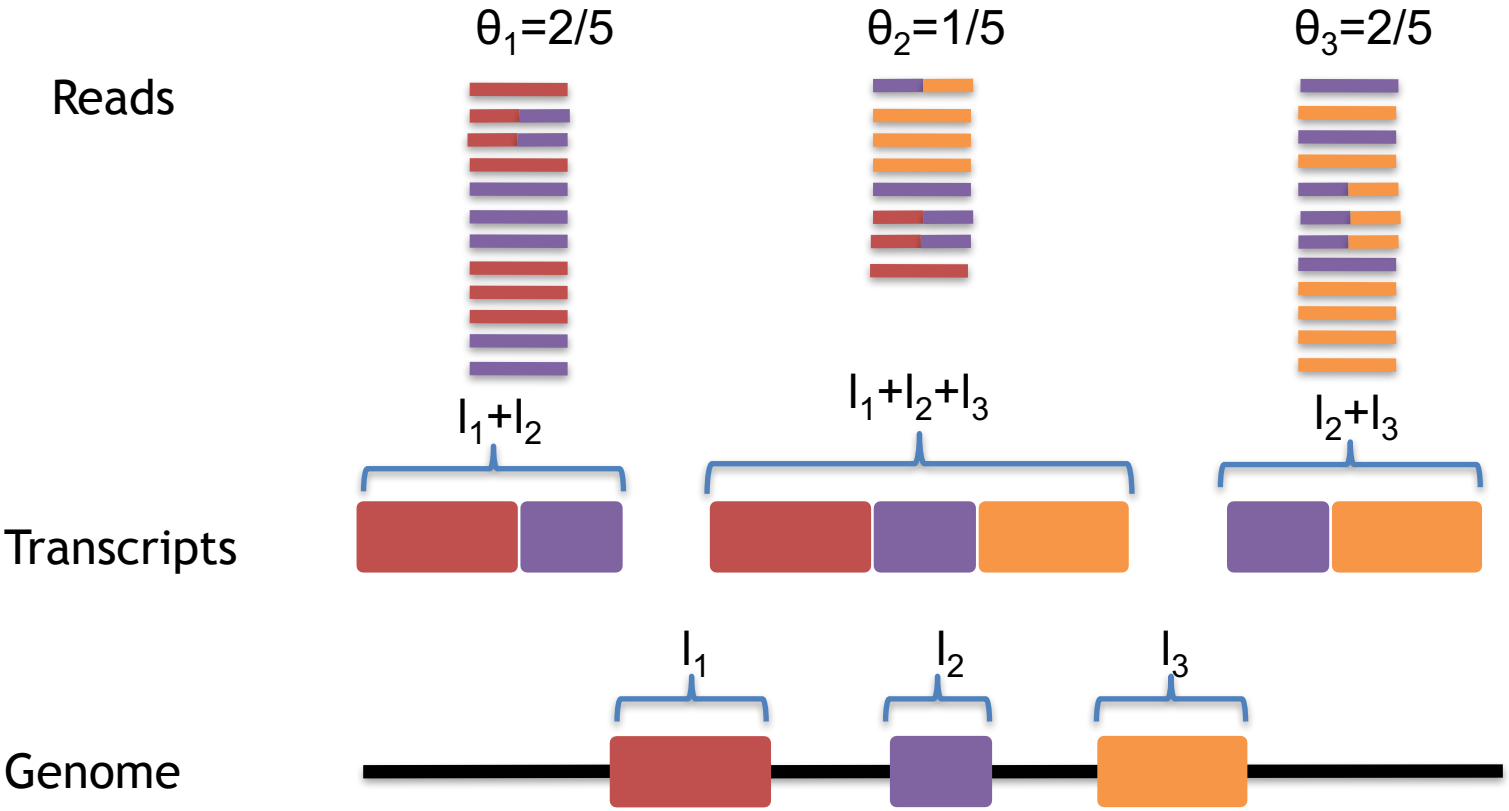
Montaño et al (2013) *Genome Biology*

Co-authors

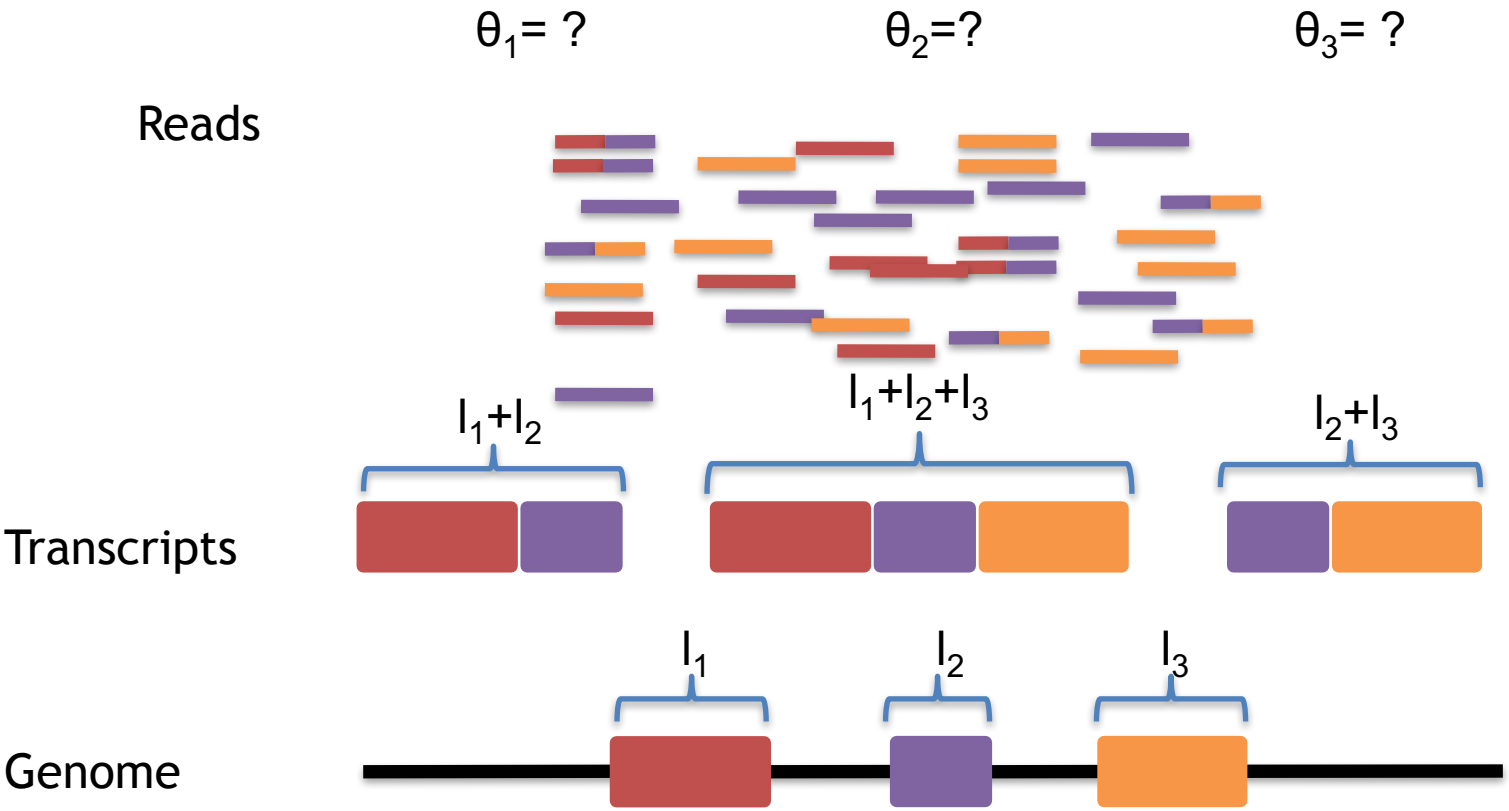
Mike Love

John Hogenesch

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

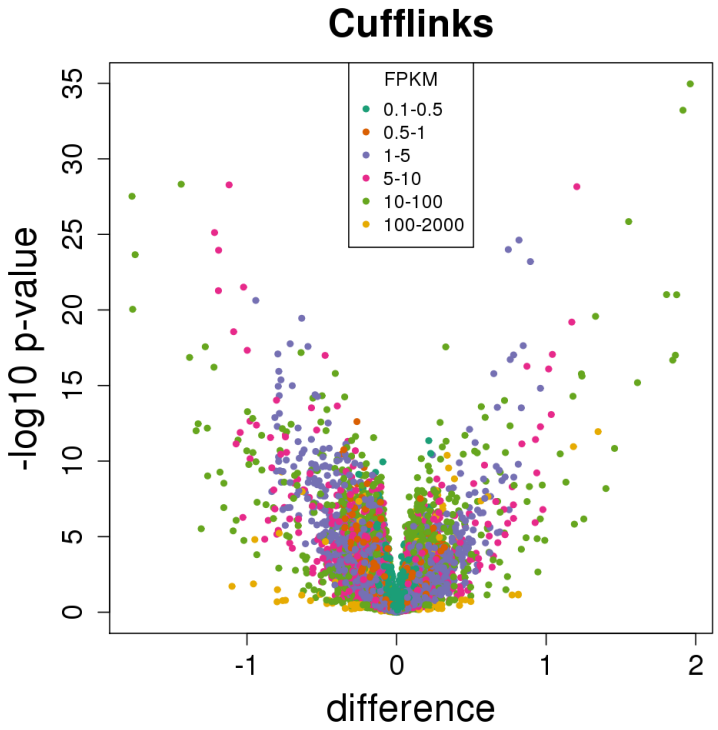
Y's are independent Poisson and transcript quantification is MLE of θs

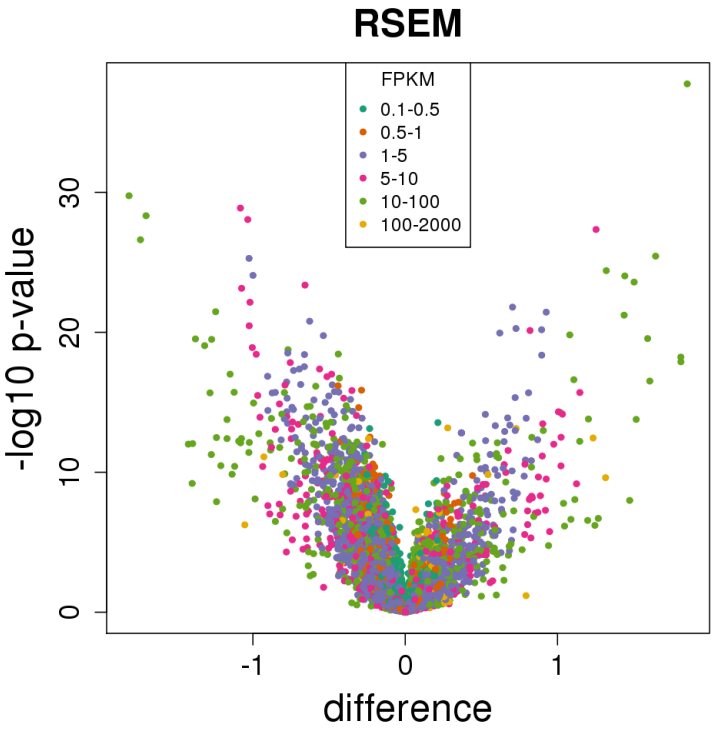
$$\mathrm{E}\begin{pmatrix} Y_1/l_1 \\ Y_2/l_2 \\ Y_3/l_3 \\ Y_{1,2}/l_{1,2} \\ Y_{2,3}/l_{2,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

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

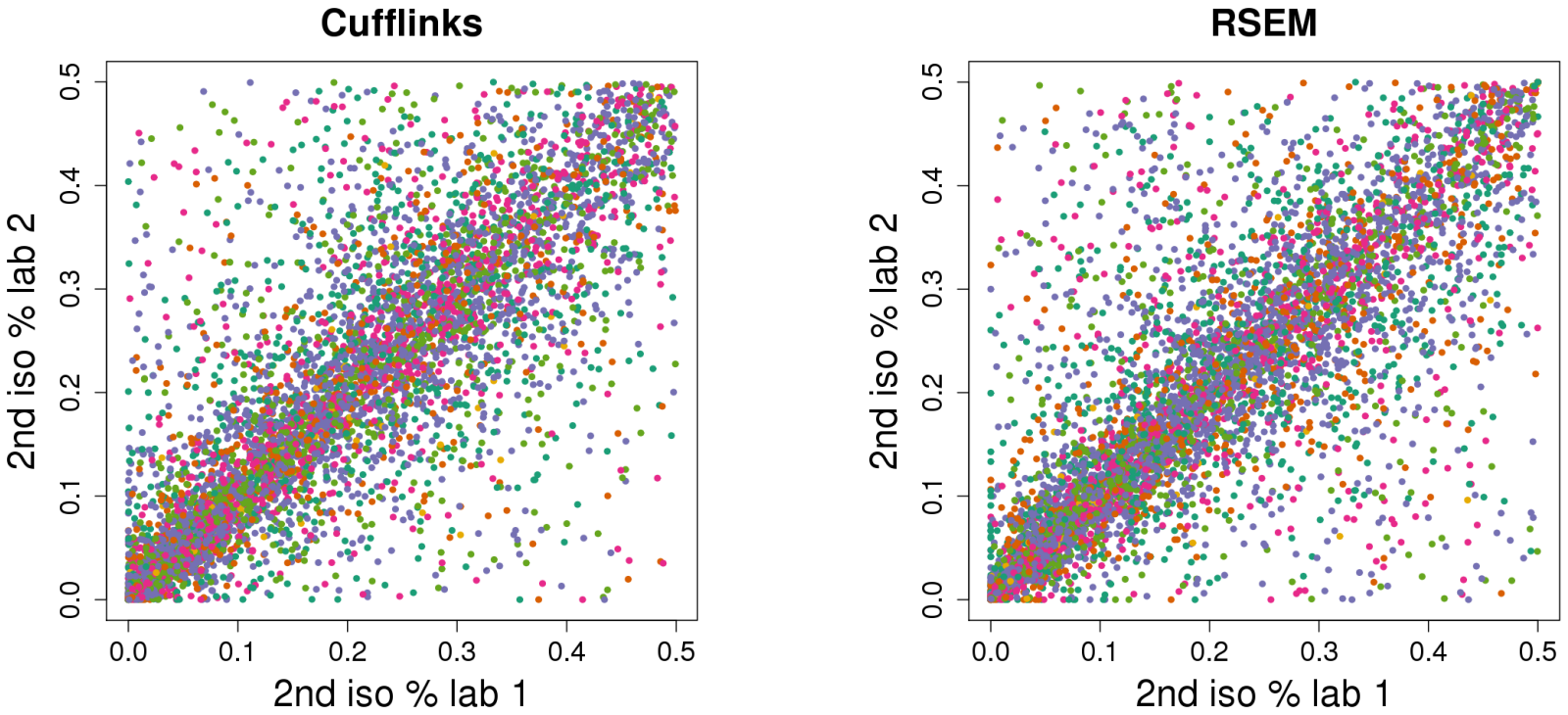
Jiang and Wong (2009) *Bioinformatics*

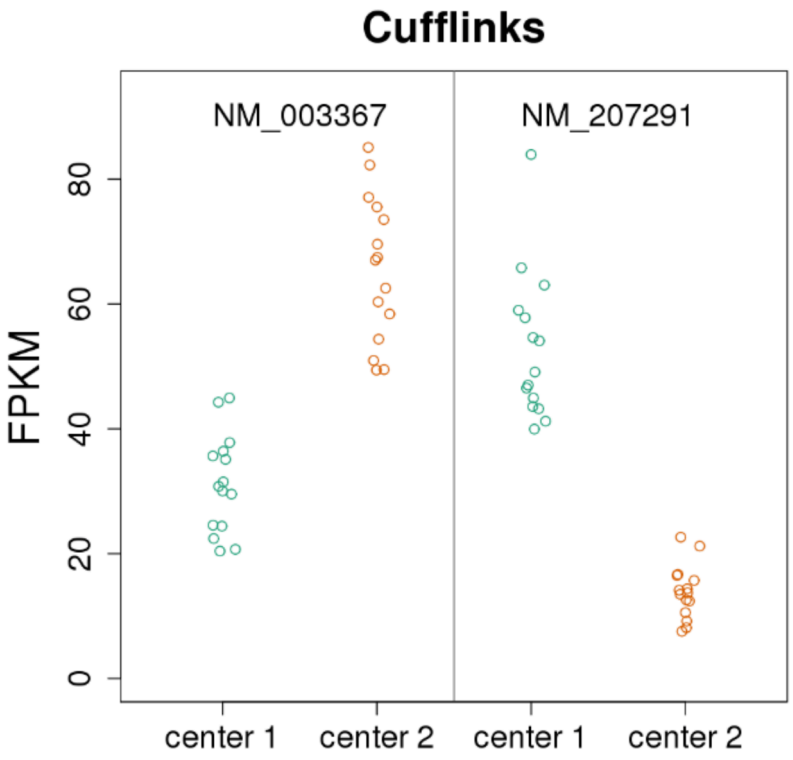
Comparison of two groups sampled from one population (from GEUVADIS)

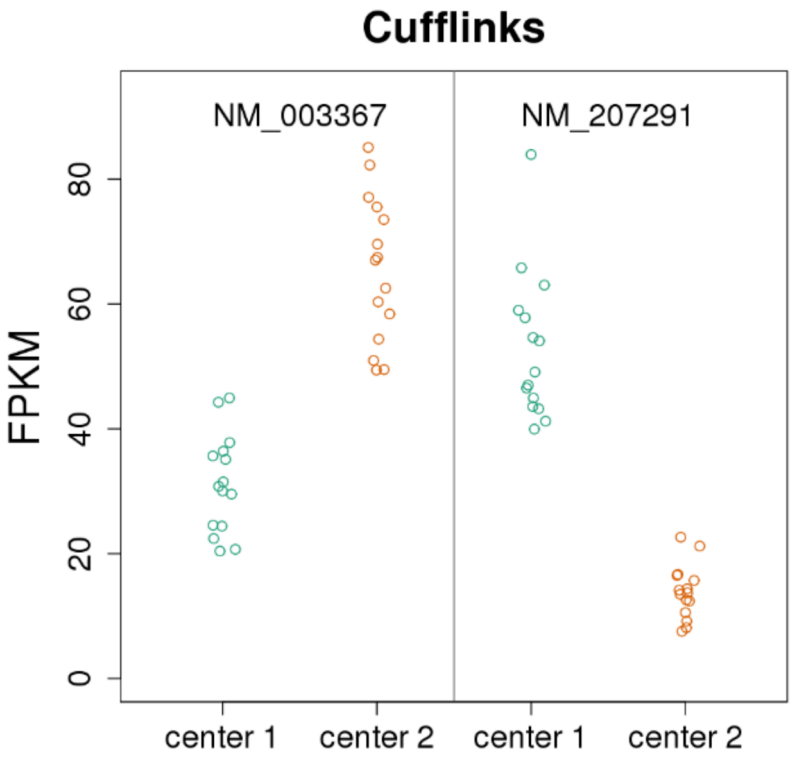


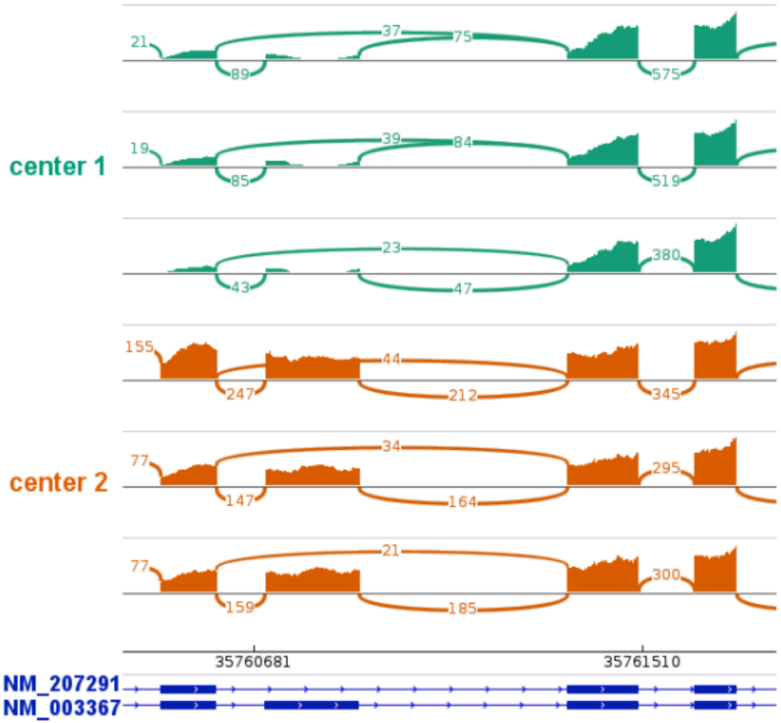


% Expression assigned to second highest isoform




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

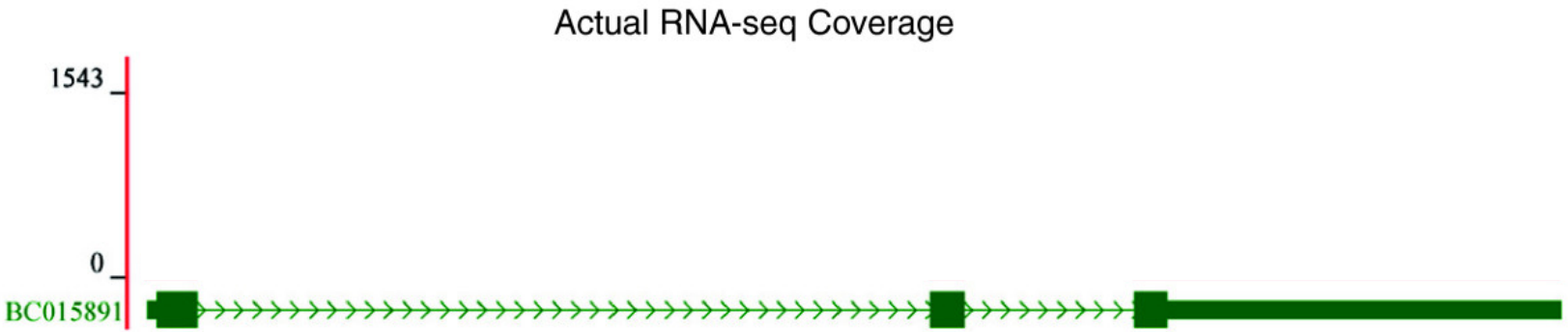
$$\mathrm{E}\begin{pmatrix} Y_1/l_1 \\ Y_2/l_2 \\ Y_3/l_3 \\ Y_{1,2}/l_{1,2} \\ Y_{2,3}/l_{2,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

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

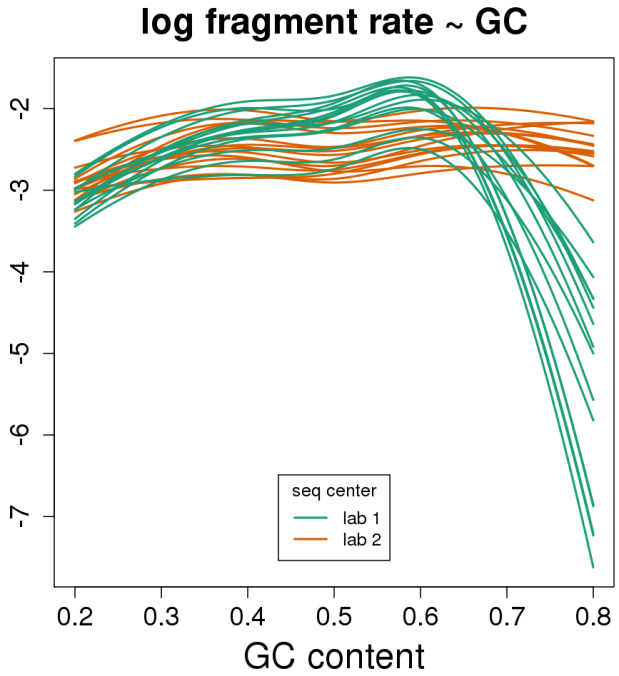
Y's are independent Poisson and transcript quantification is MLE of θs

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

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

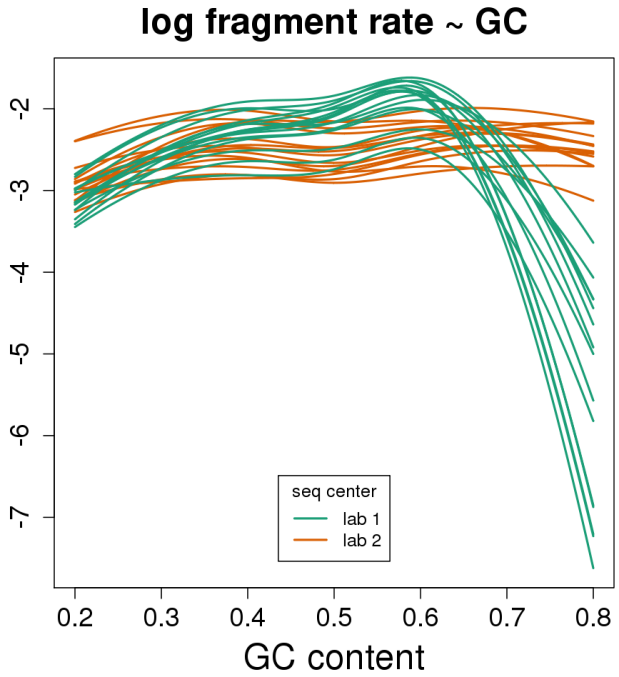
Parameters no longer identifiable

Bias varies from sample to sample



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

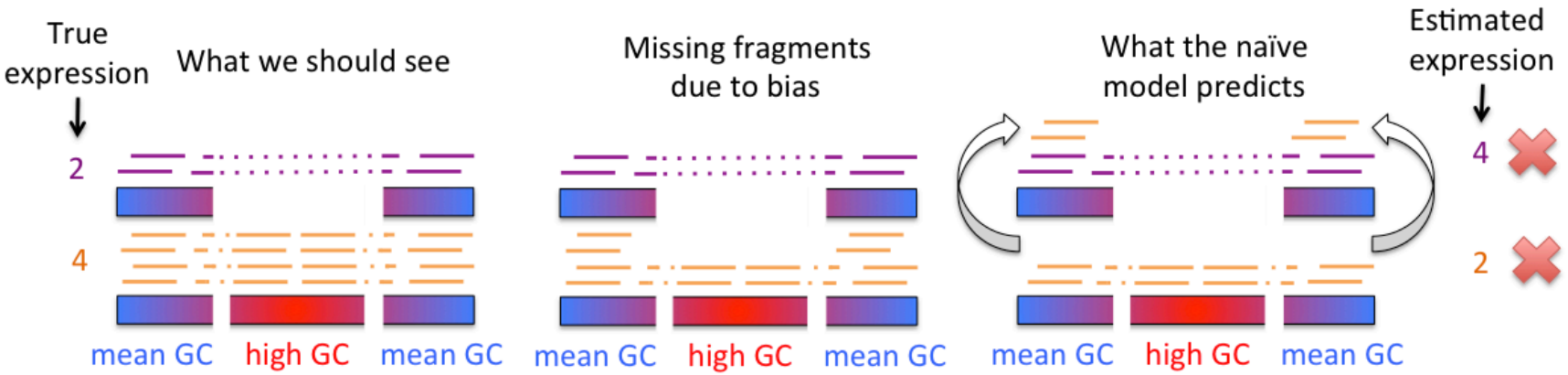
Bias varies from sample to sample

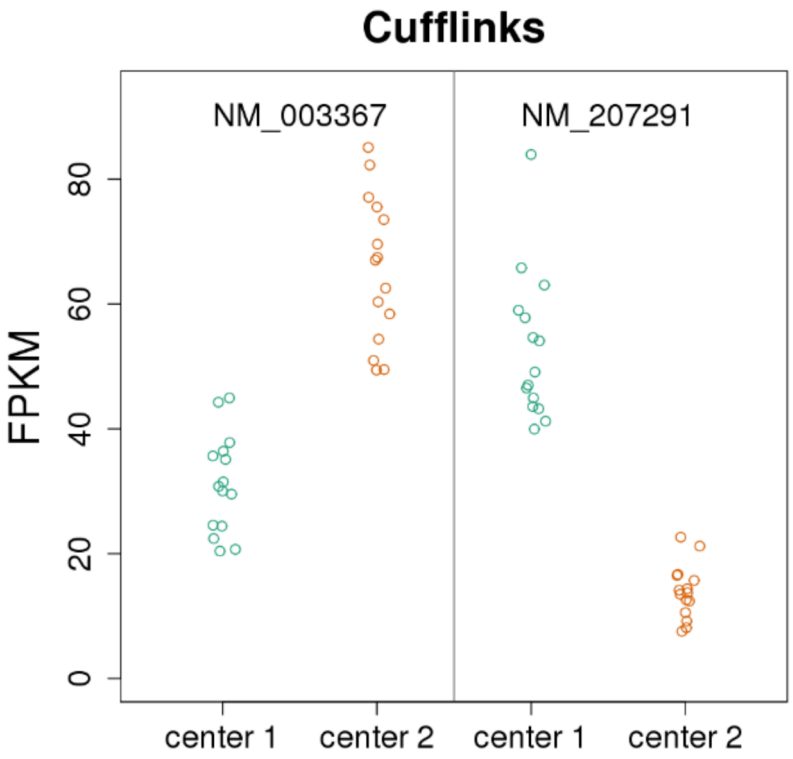


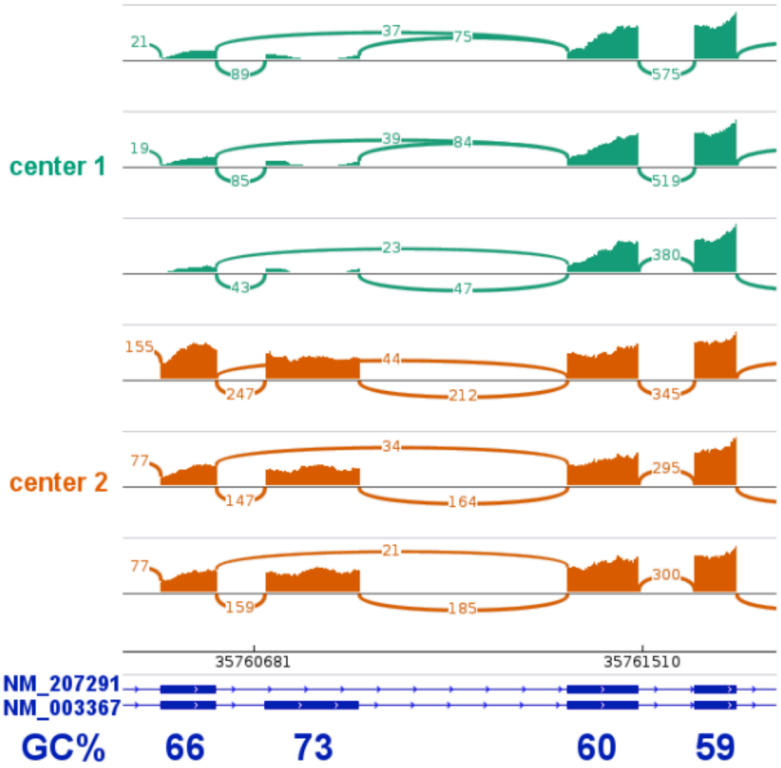
$$Y_j \sim \text{Poisson}(\lambda_j^b)$$

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

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

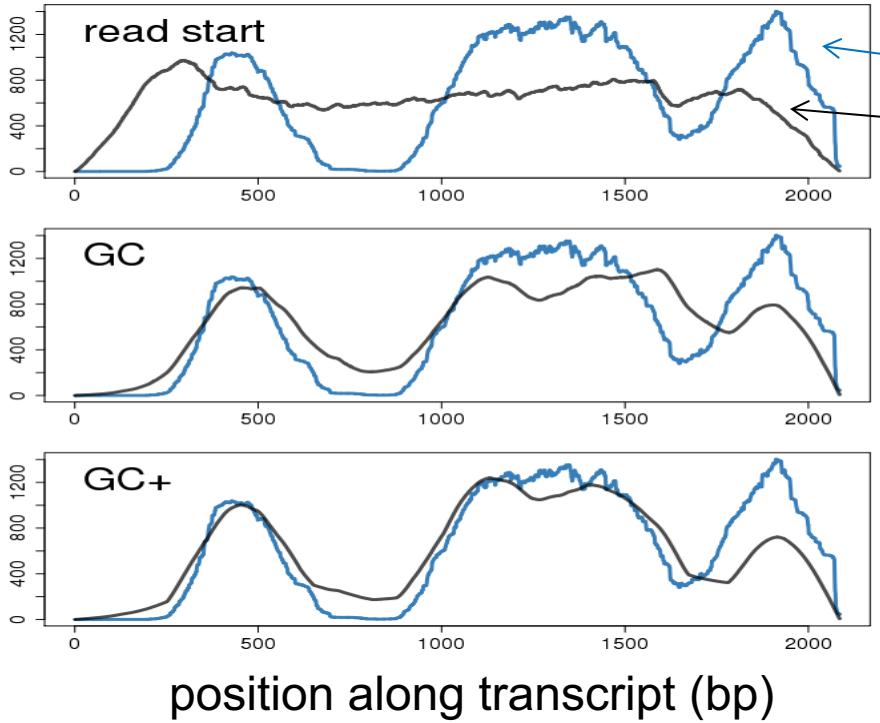
Alpine

Alpine

Better predict coverage artifacts

Cufflinks approach using read starts

alpine

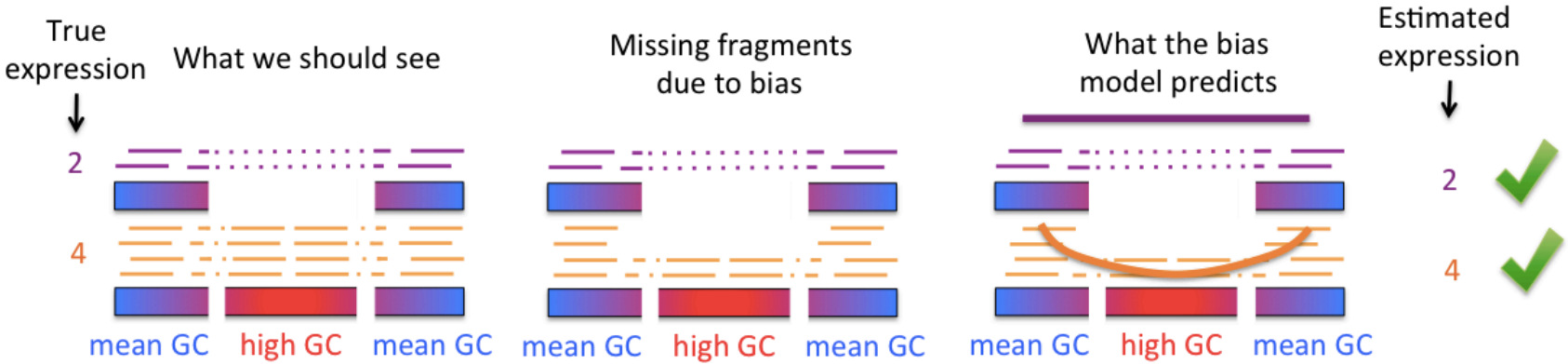


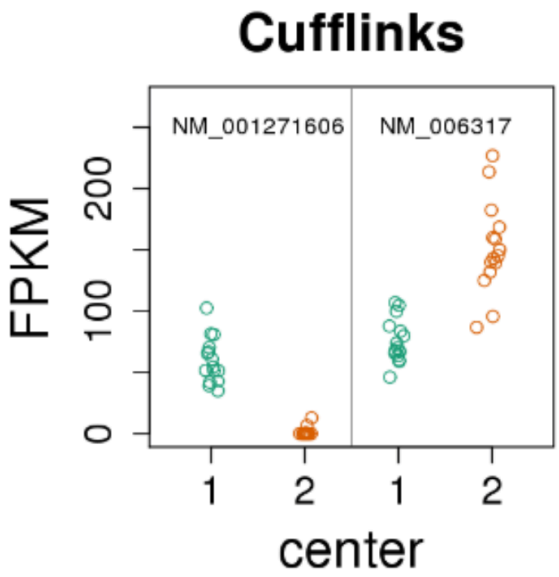
Transcript coverage

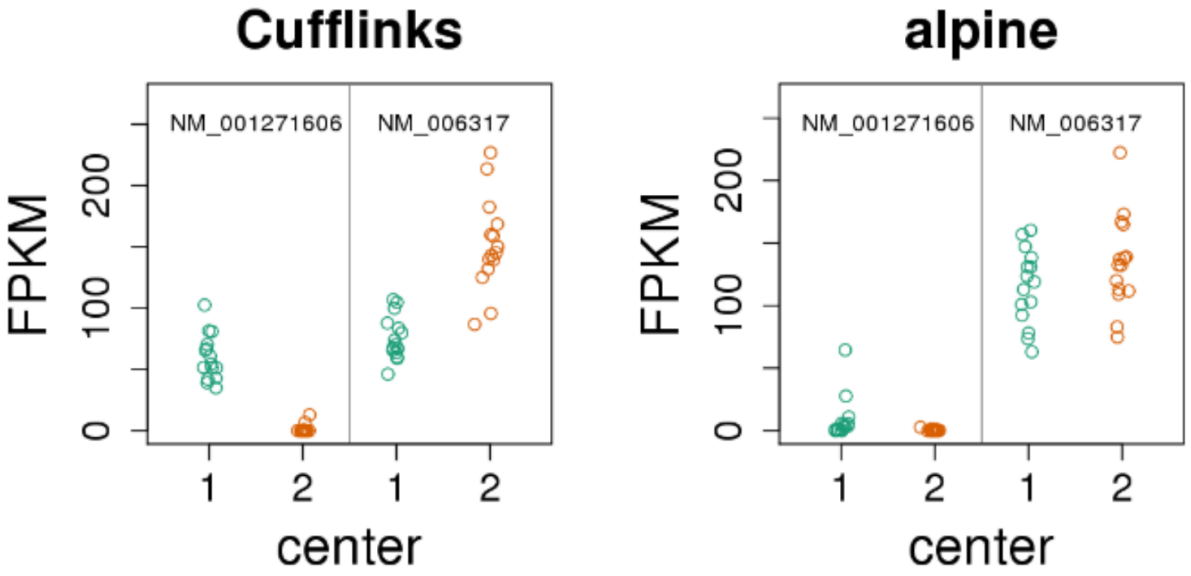
Test set prediction

Also modeling stretches of $(G|C)^n$

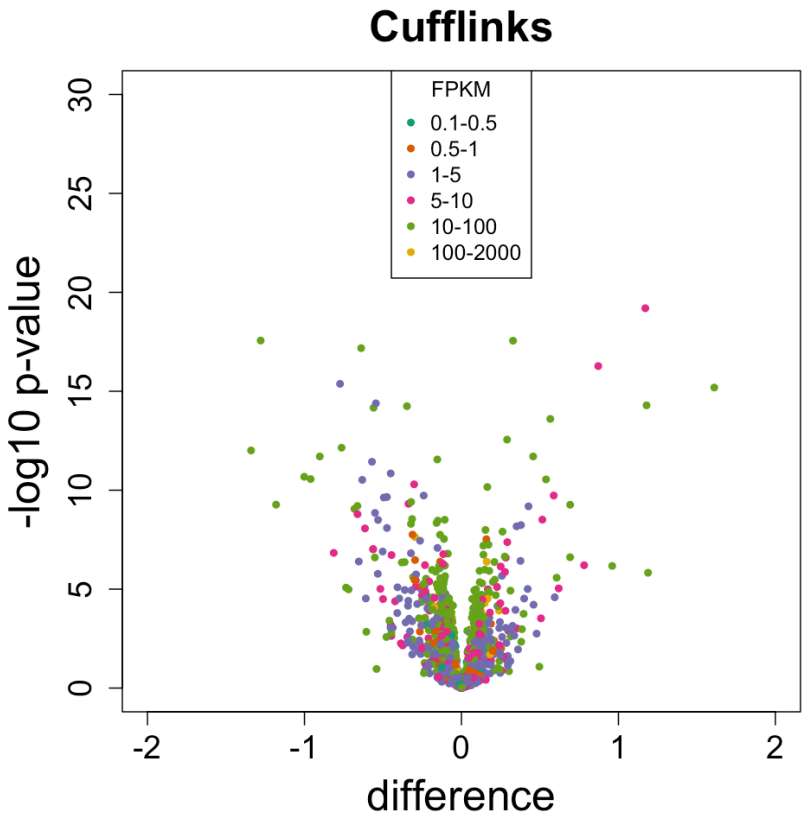
Parsimonious model

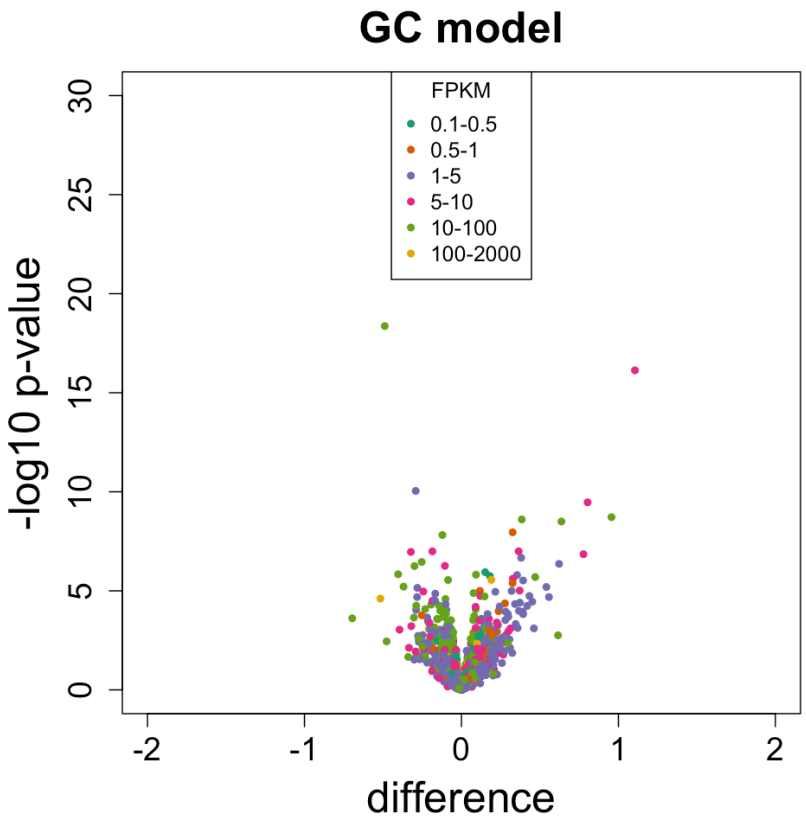



Cufflinks
NM_001271606
NM_006317
FPKM
0
100
200
1
2
1
2
center


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

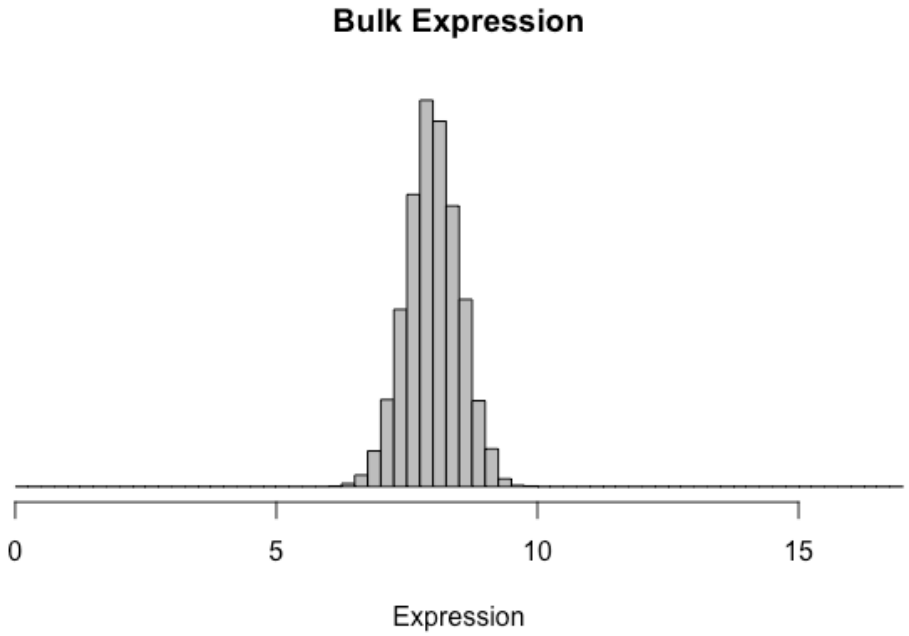
Single Cell RNA-Seq

Stephanie Hicks
@stephaniehicks

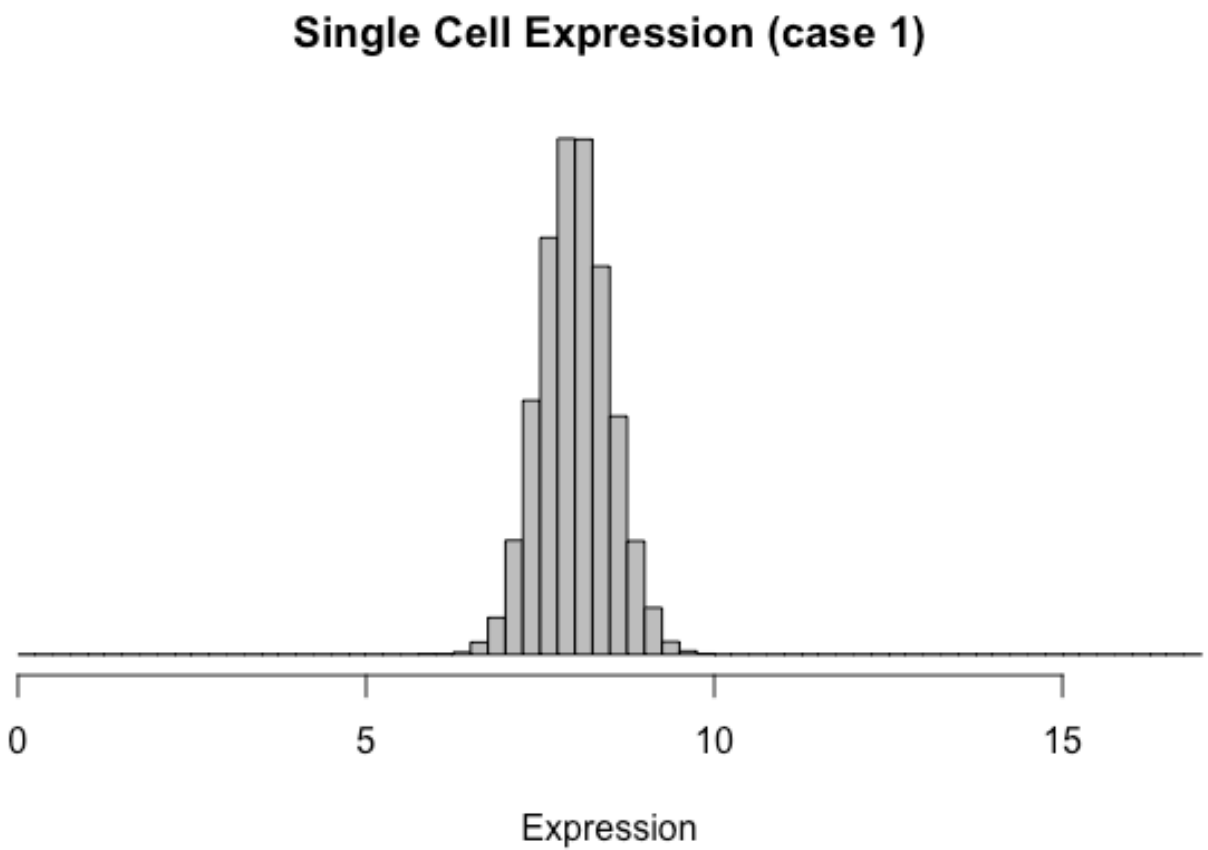
Will Townes
@sandakano

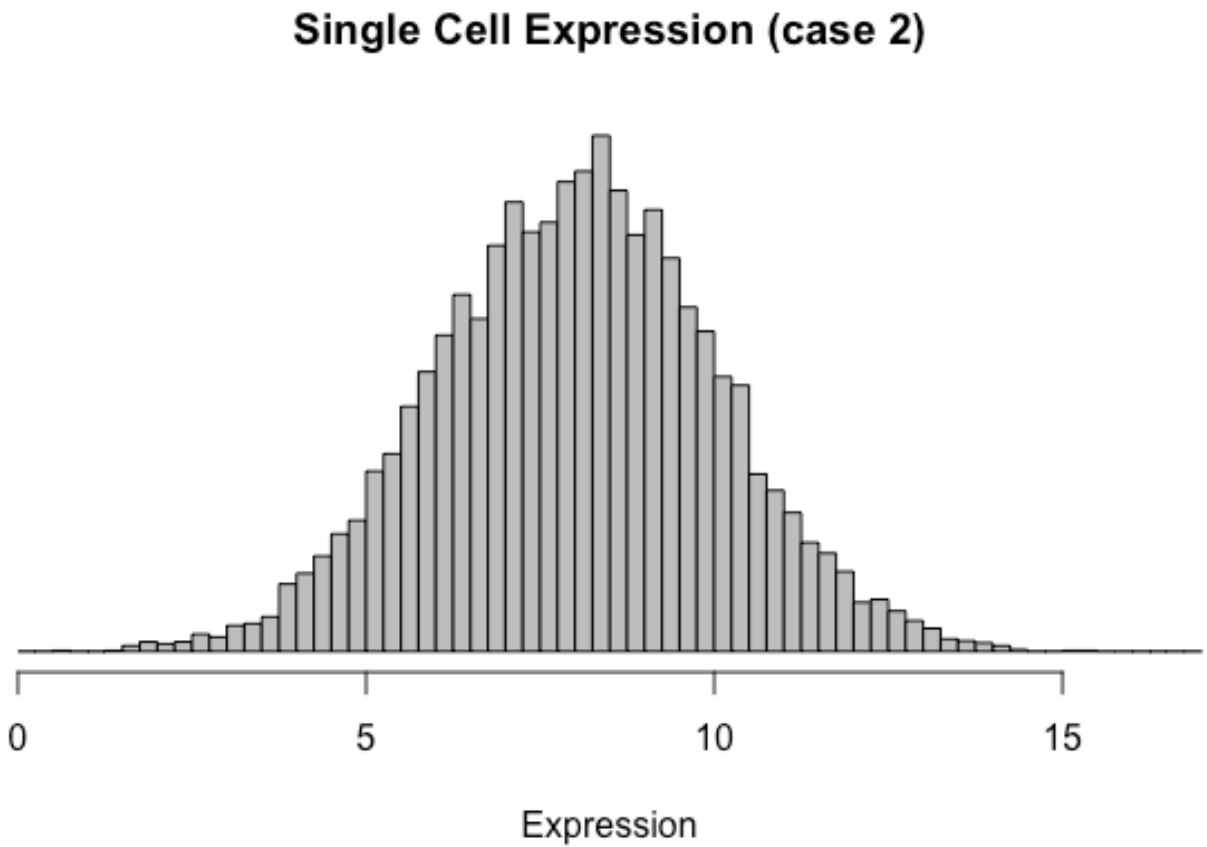
Single cell versus Bulk RNA-Seq measurements

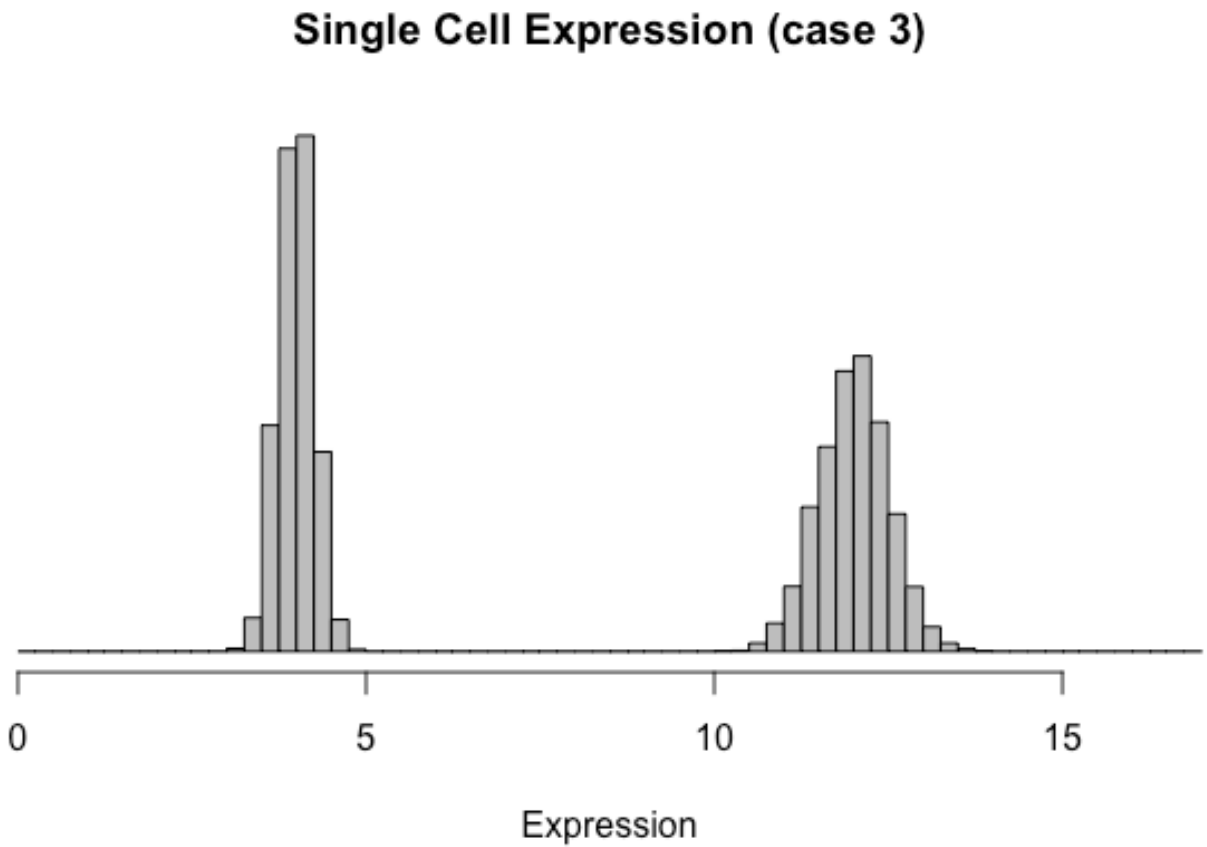
Bulk expression measured across several samples

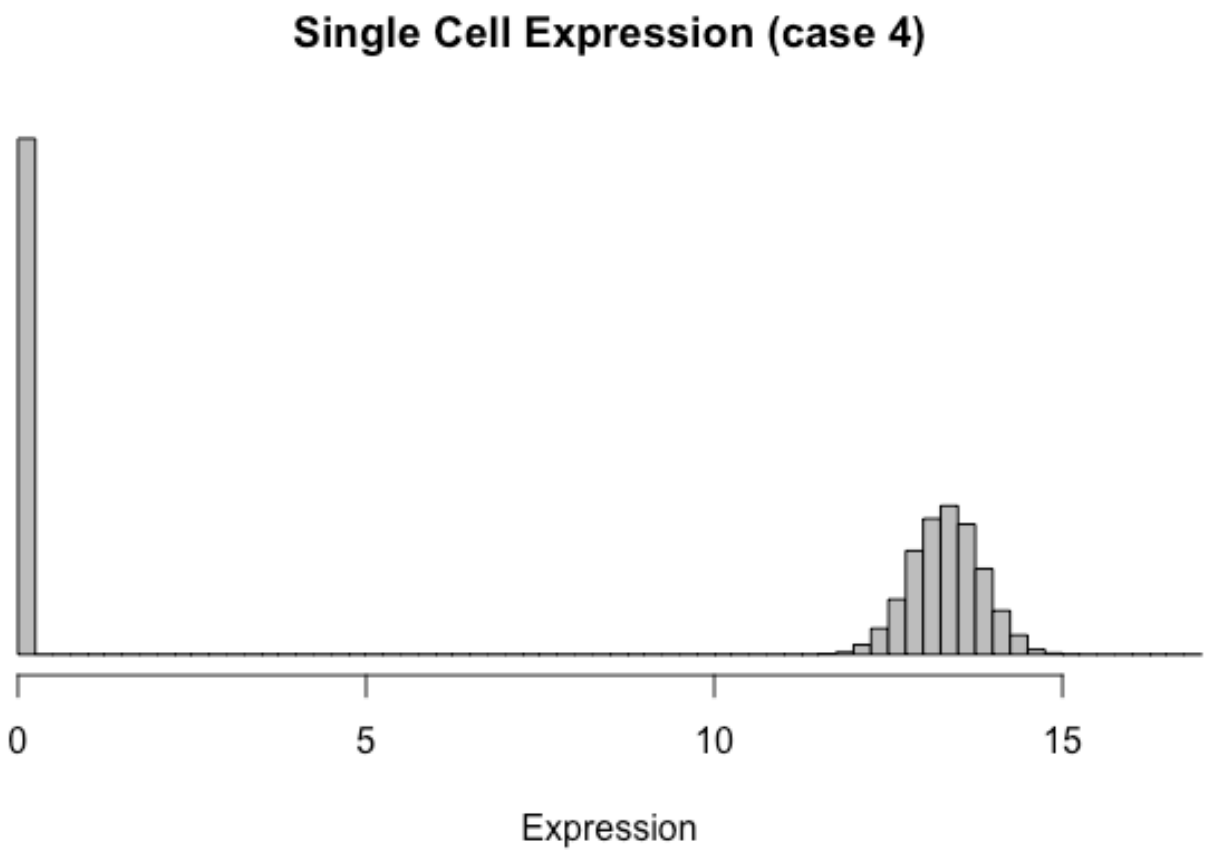


What will single cell measurements look like?

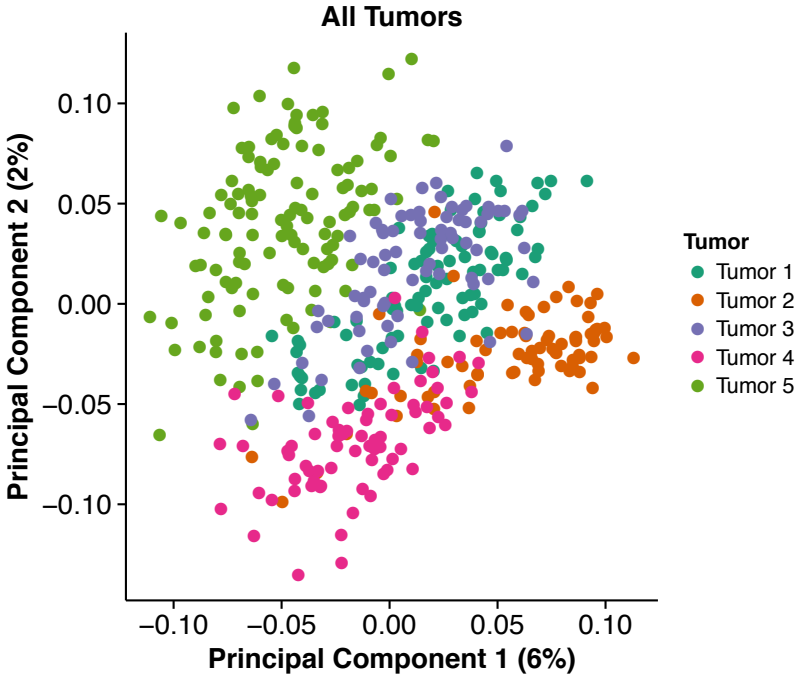

Single Cell Expression (case 1)
0
5
10
15
Expression


Single Cell Expression (case 2)
0
5
10
15
Expression

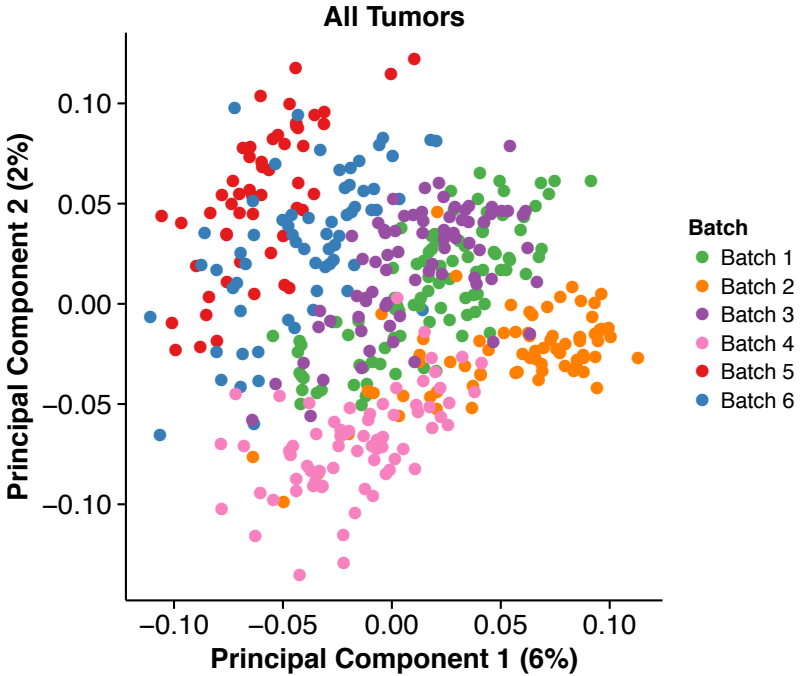

Single Cell Expression (case 3)
0
5
10
15
Expression


Single Cell Expression (case 4)
0
5
10
15
Expression

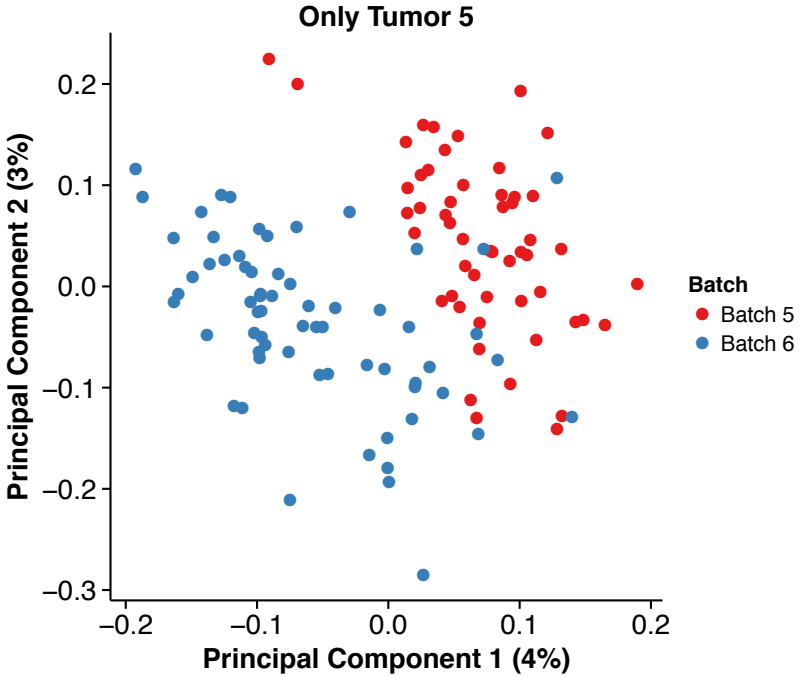
Discovering new cell types



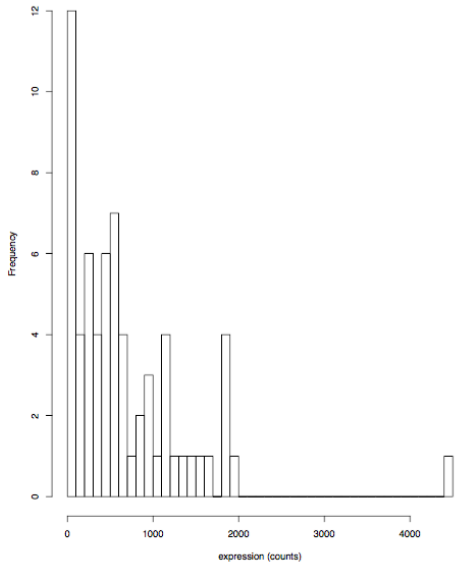
Or is it a batch effect?



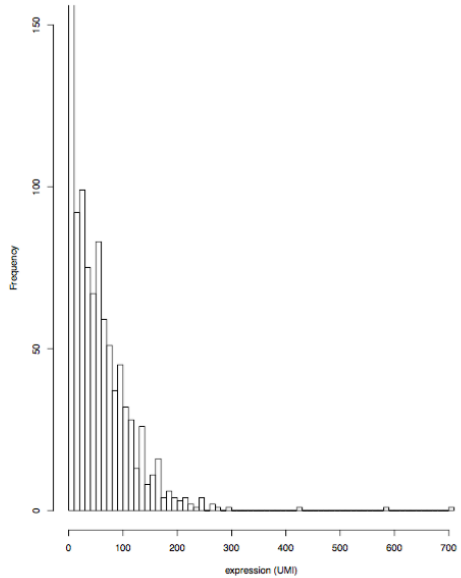
Same group measured on two sequencers



Zeros are very common

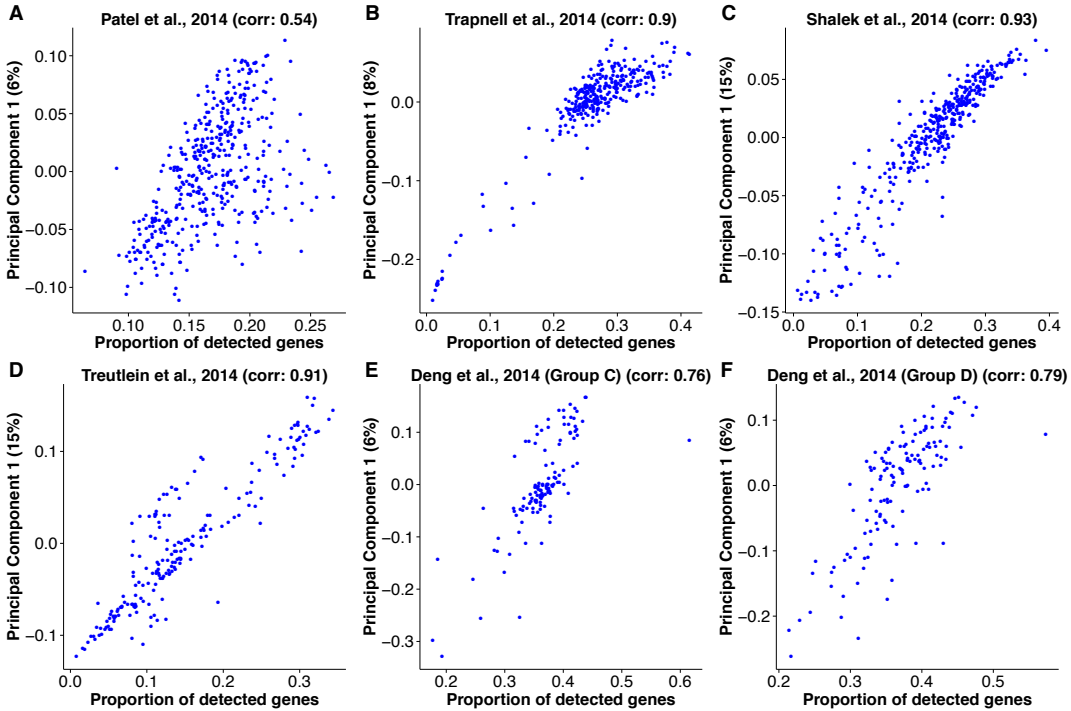


Patel et al 2014 Science

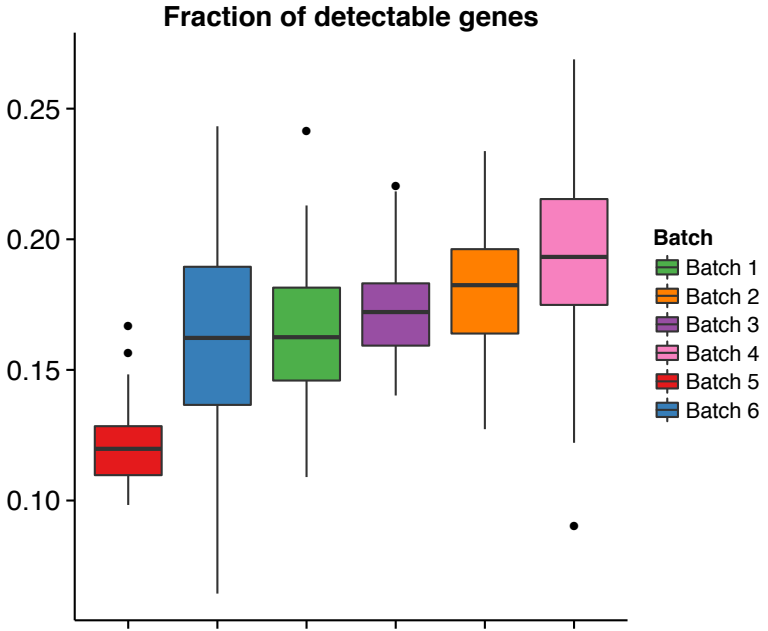


Zeisel et al 2015 Science

The proportion of zeros varies across sample

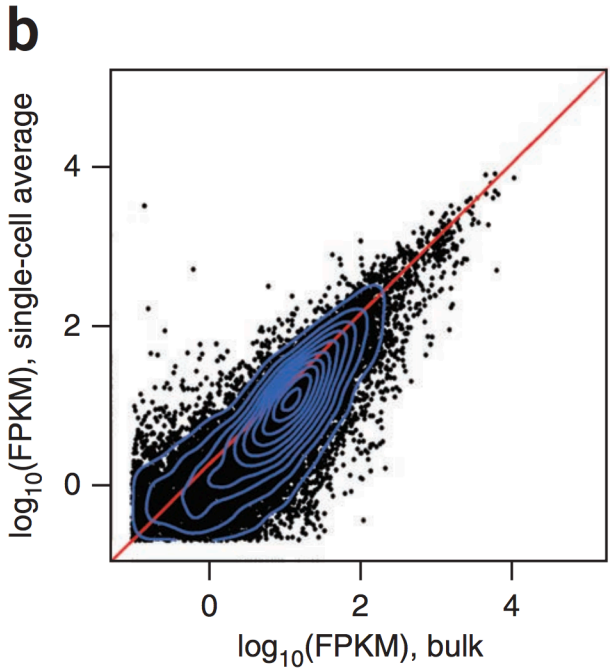


The proportion of zeros changes



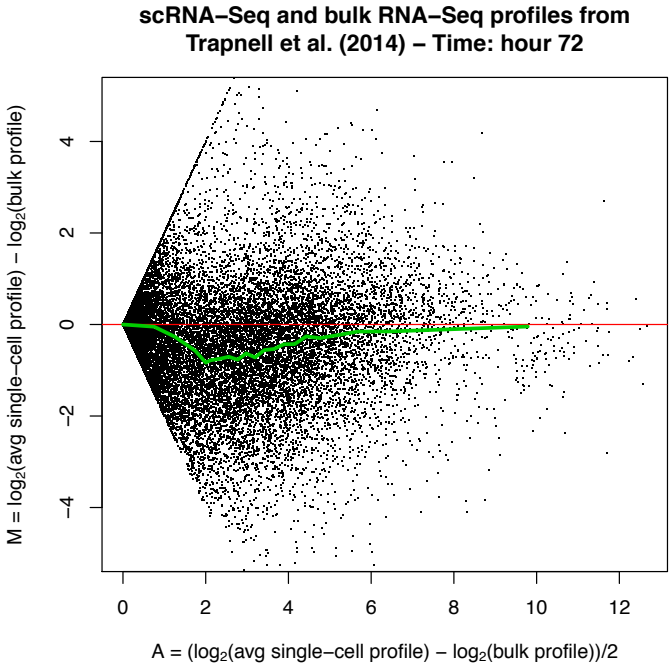
Are differences in proportion of zero real biology or an technical artifact?

Agreement between bulk and single cell?



Trapnell et al. (2014) showing concordance between single-cell and bulk

Here is an MA-plot version



If X is a bulk measurement and

Y_1, … Y_N are single cells measuremetns from bulk Y

If X and Y are samples from same population

Then regardless of distibution of Y,

$$E[Y] = E[X]$$

If we stratify the X into small bins then in bin X=x the average

$$A = (Y_1 + \ldots + Y_N)/N$$

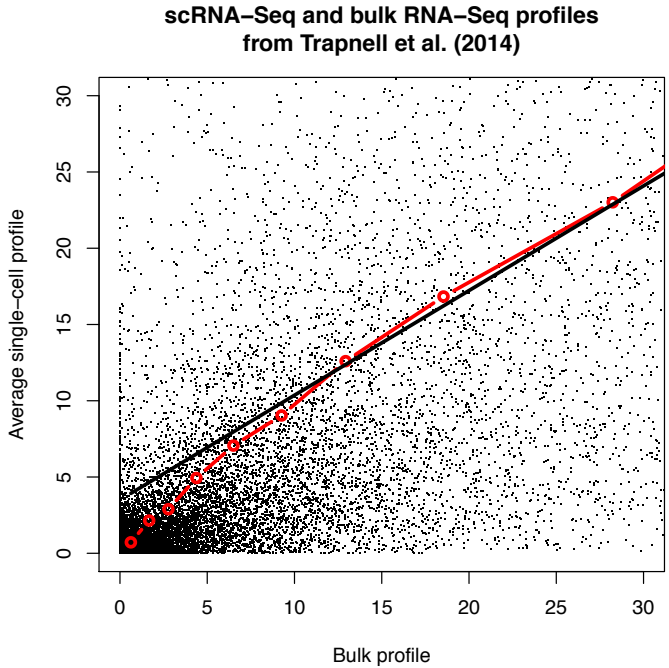
has expected value

$$E[A] = a + rX$$

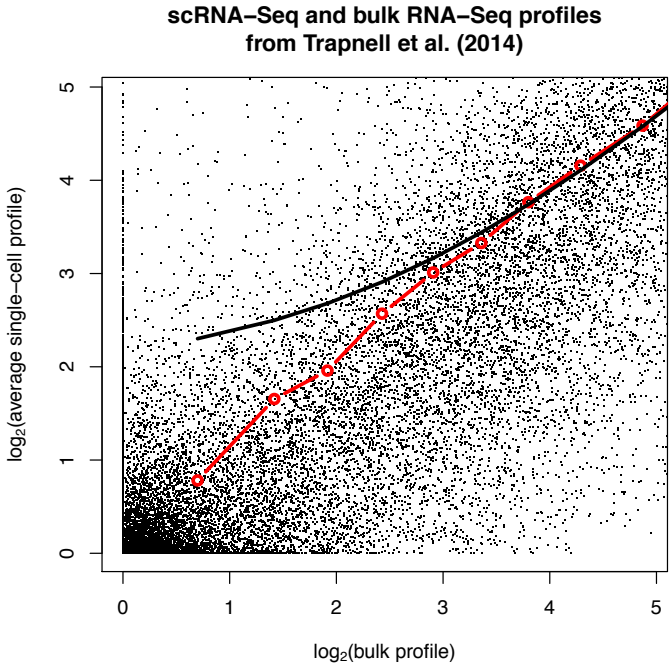
with r the correlation and a term that depends on averges, SDs and r

This is the regression line.

The proportion of 0s is higher than expected for low expressed gens

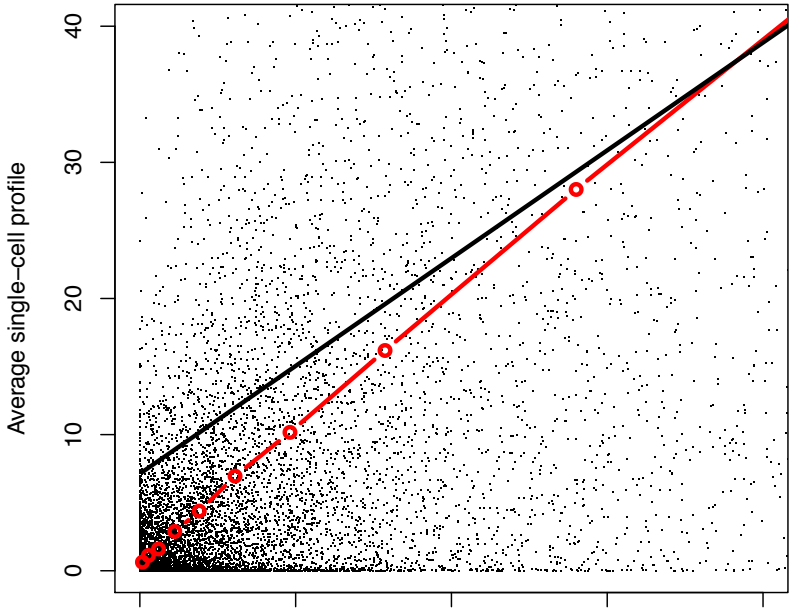


The proportion of 0s is higher than expected for low expressed gens

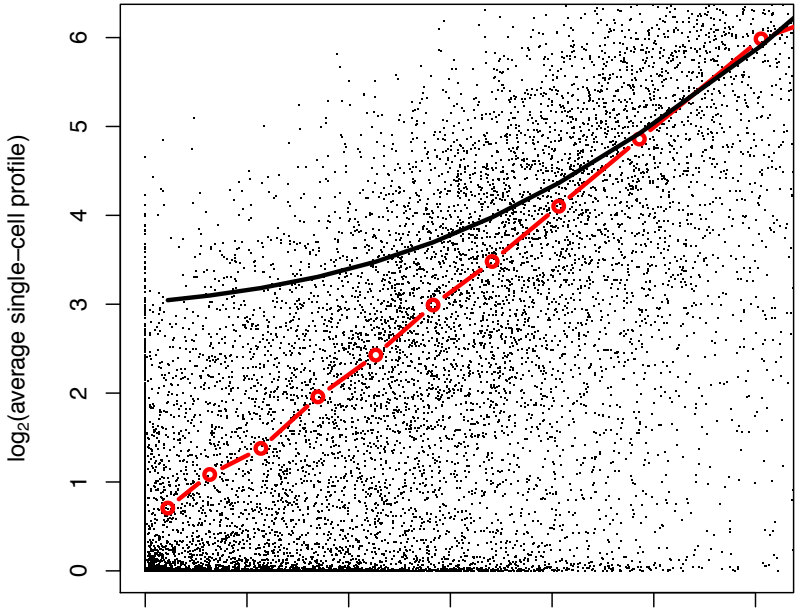


Other examples

Original scale

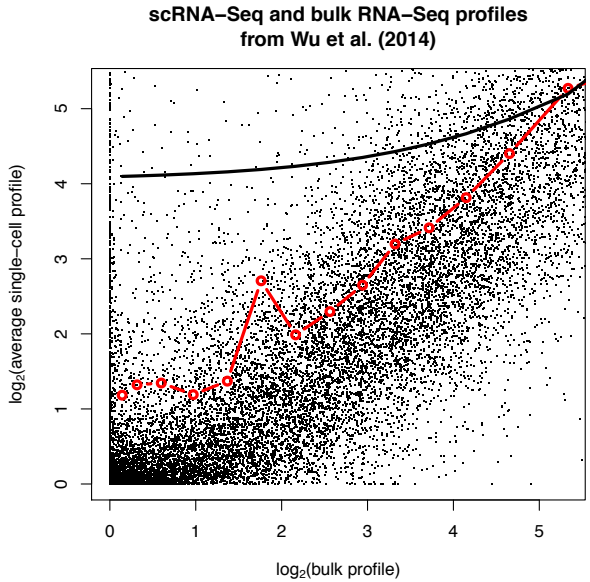


Log (base 2) scale



Another example

Original Scale



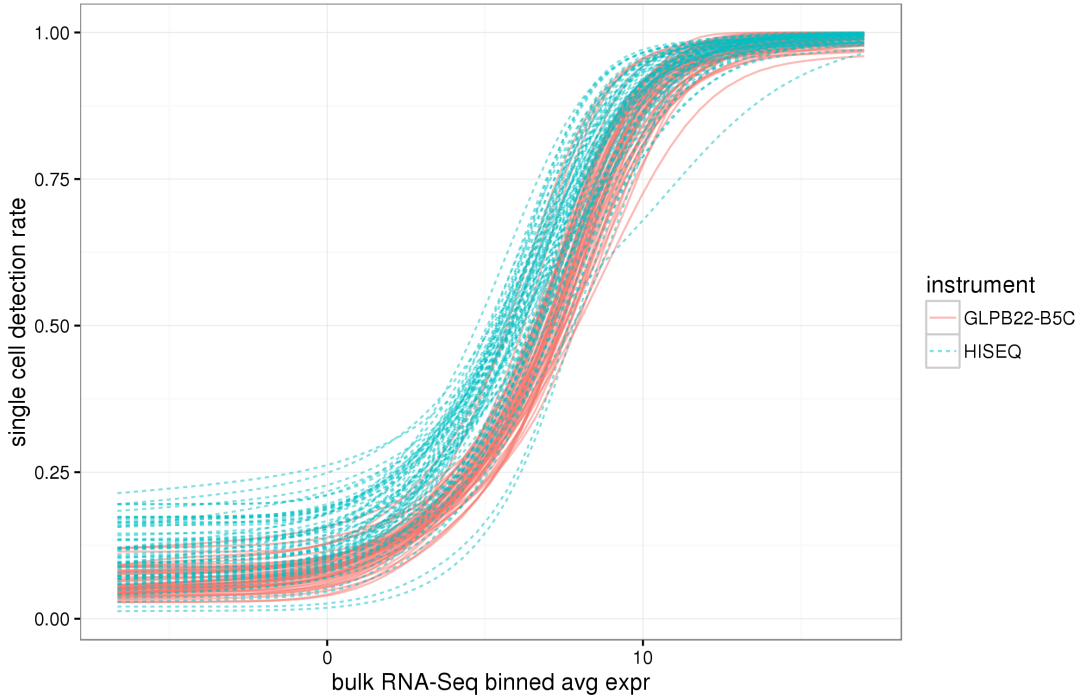
Zero Inflated models have been proposed

For example, Pearson and Yau (2016) Genome Biology

But there is a problem, these models assume the missing probability is the same across samples

This does not appear to be the case

Curves fit to 1 (detected) 0 (not-detected) data across several samples from two batches



ZIFA model won't account for this

Missing not at random model

Probabilistic PCA + Censoring

$$Y_{ng} \sim \mathcal{N}(\eta_{ng}, \sigma_y^2)$$

$$\eta_{ng} = y_0 + a_n + w_g + u_n^T v_g$$

$$\Pr(Z_{ng} = 1) = f(\eta_{ng})$$

$$f(\cdot) = ???$$

Missing not at random model

Competing method Zero Inflated Factor Analysis[1] (ZIFA)

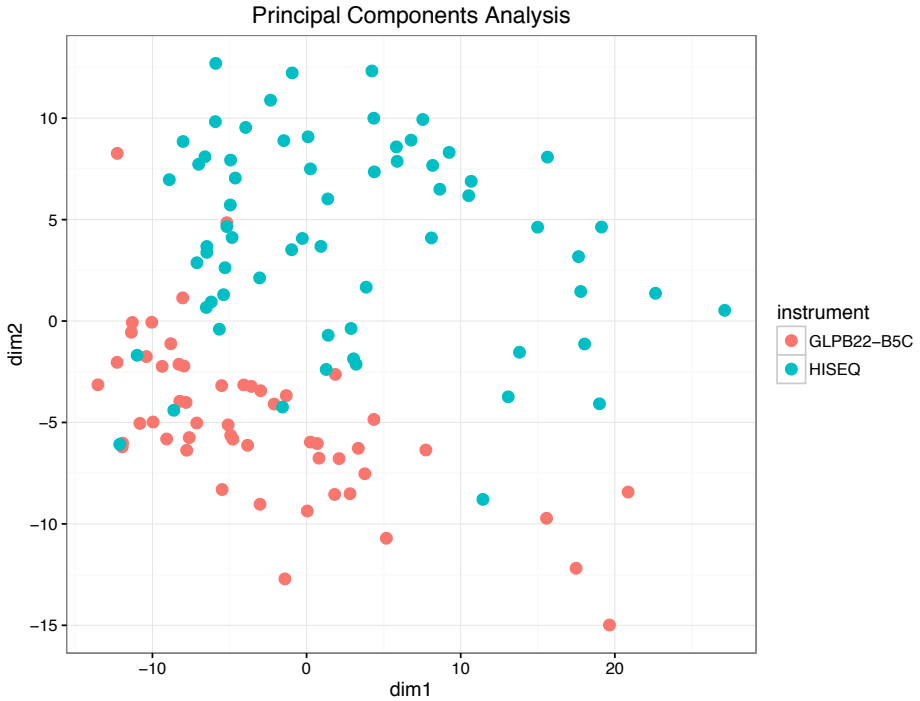
$$\Pr(Z_{ng} = 1) = 1 - \exp\{-\lambda \eta_{ng}^2\}$$

Our model

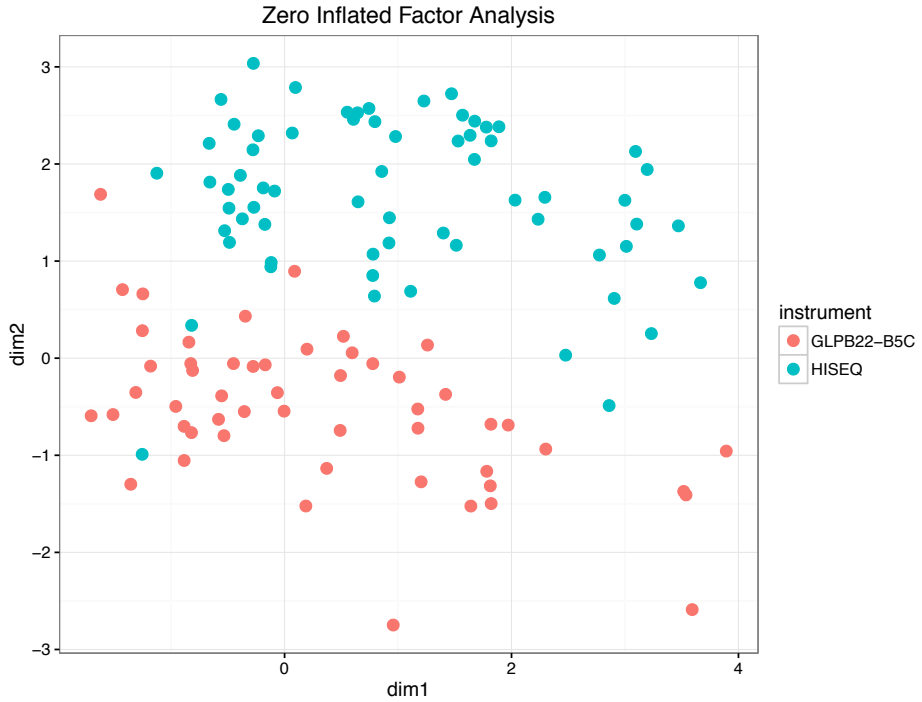
$$\Pr(Z_{ng} = 1) = \frac{\kappa_n}{1 + \exp\{-(\eta_{ng} - d_n)\}}$$

has between-cell variable censoring

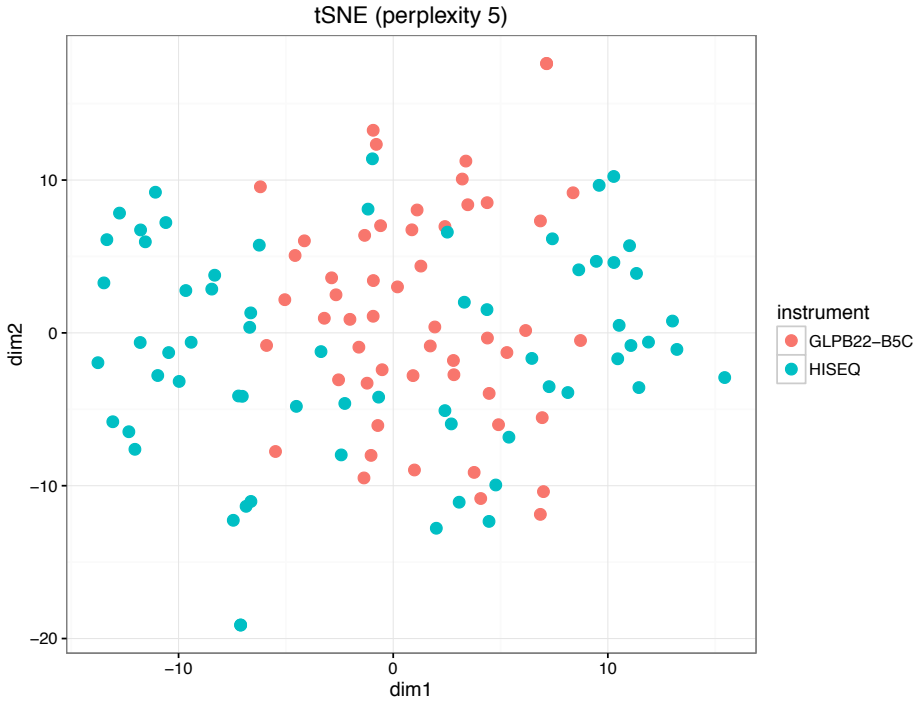
Back to the tumor cells run in two scanners



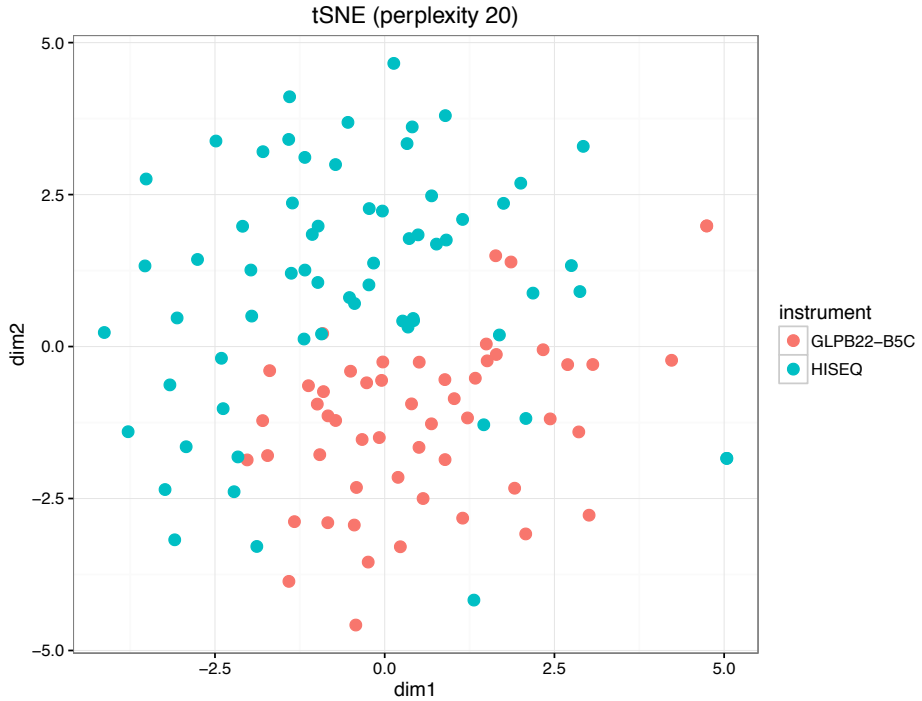
ZIFA does not remove batch effect



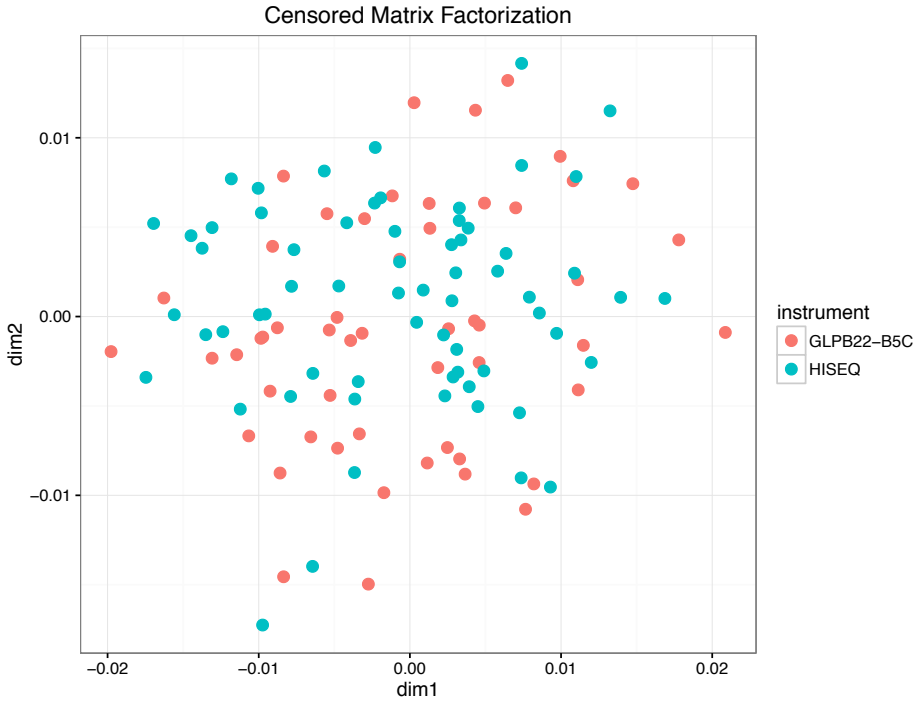
tSNE does not either



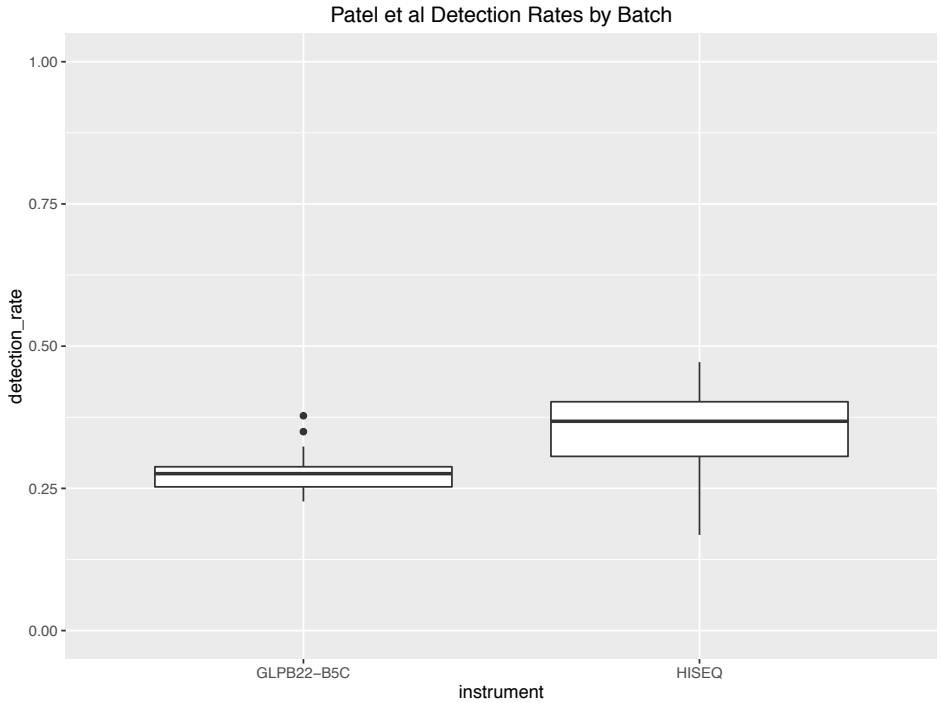
Regardless of choice of parameter



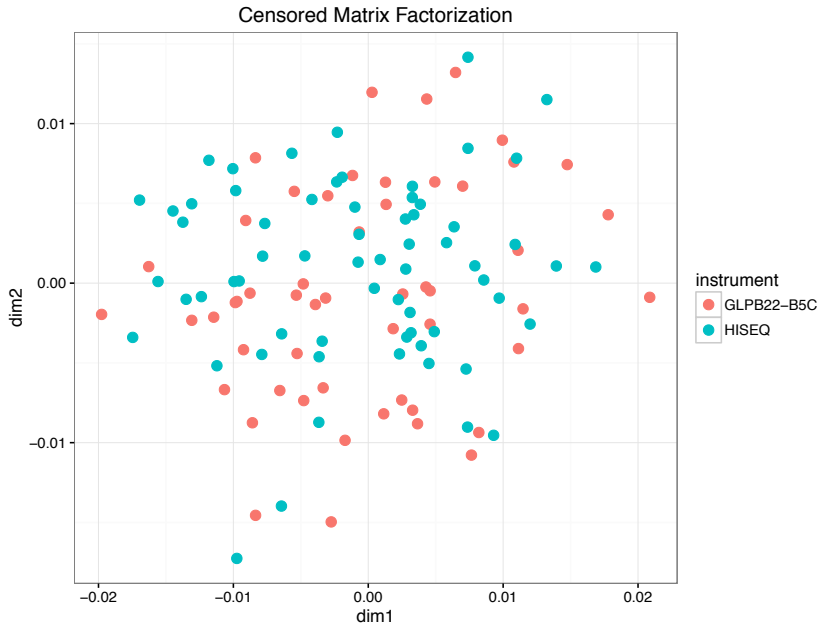
Our model does

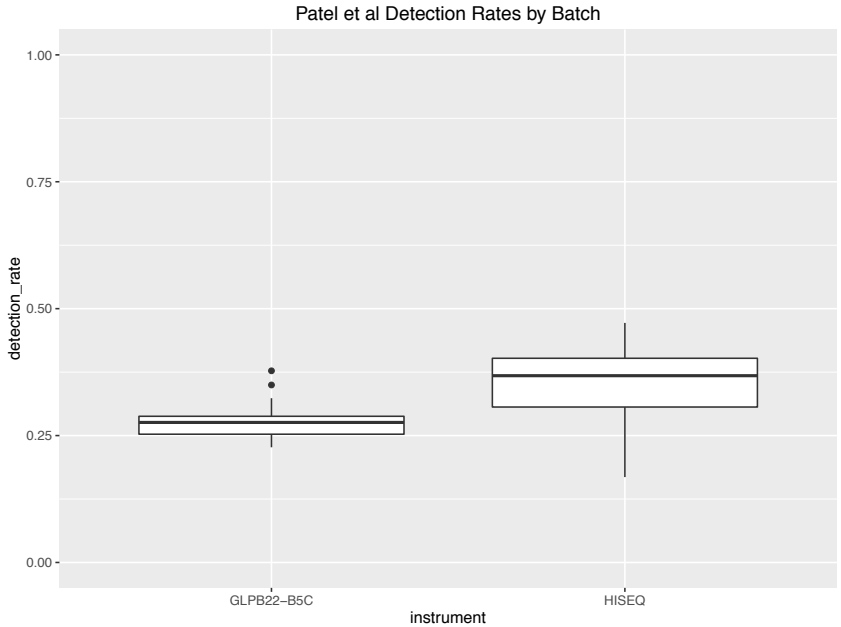


It correctly attributed variability to different detection rates:

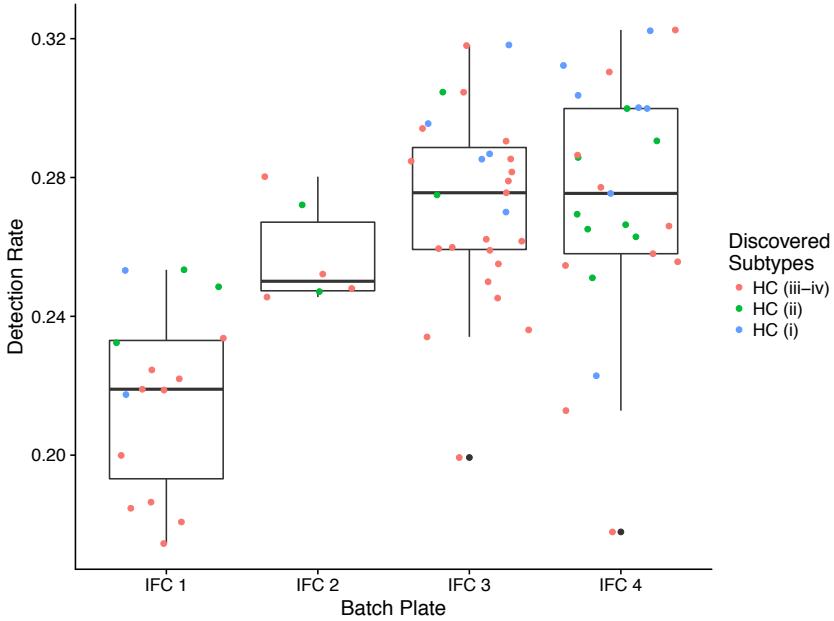


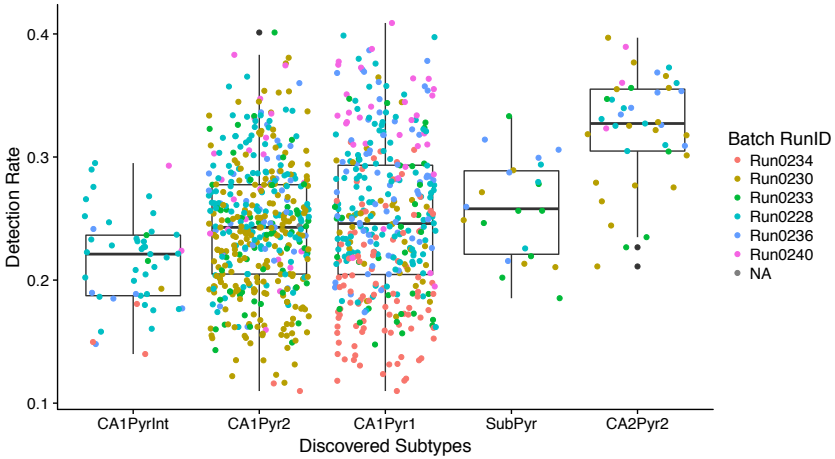
We propose reporting these separately





Group discovery can be driven by detection rates





Acknowledgments

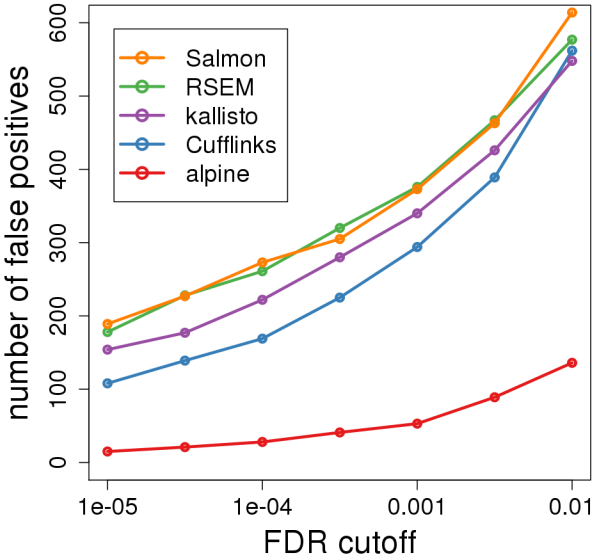
Michael Love
Stephanie Hicks
John Hogenesch
Andrew Jaffe
Jeff Leek
Mingxiang Teng

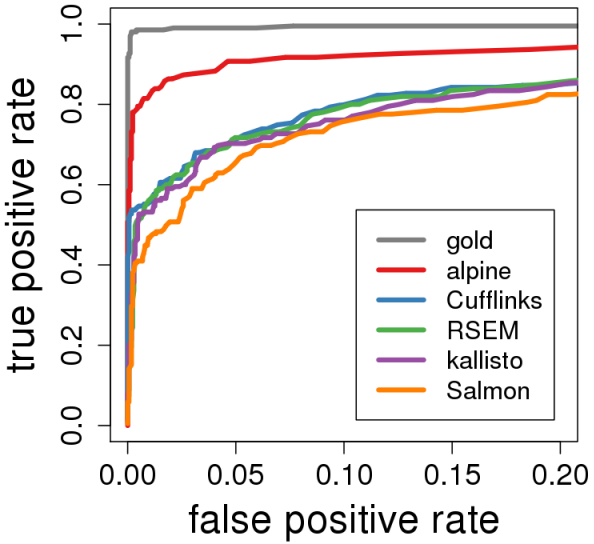
NIH R01 grants
- HG005220
- GM083084
- RR021967/GM103552

NIH P41 grant
- HG004059

http://rafalab.org @rafalab

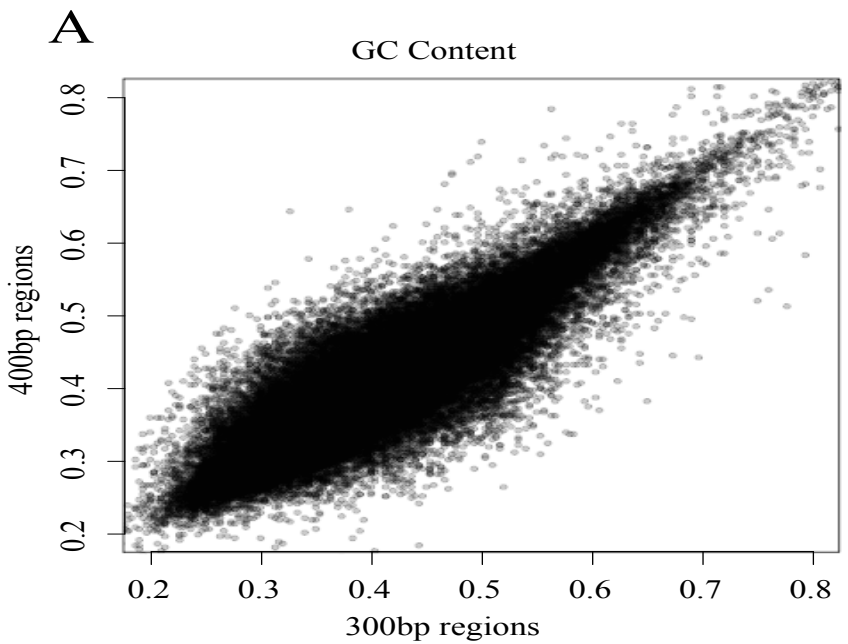
More comparisons

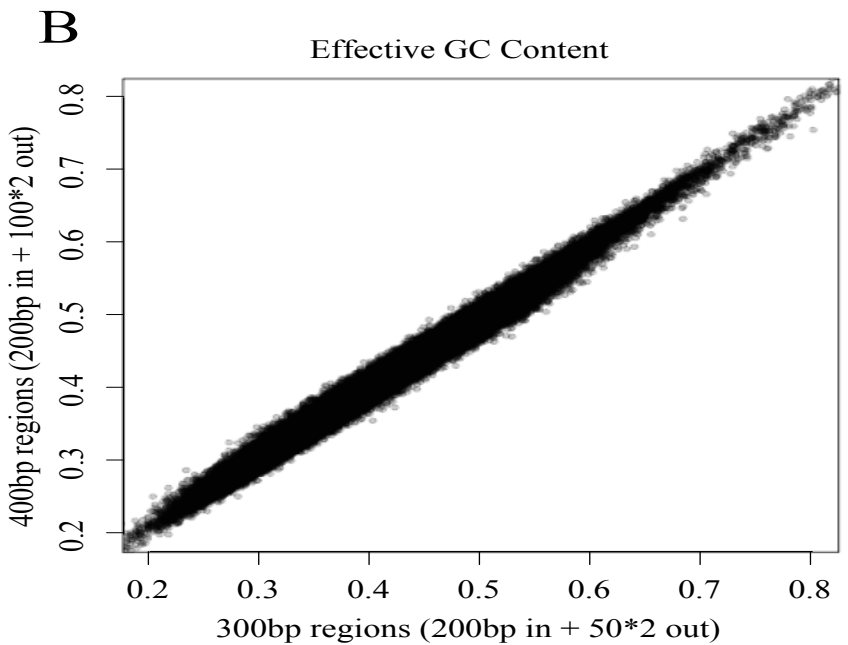




MC simulations used for ROC

Effective GC-content provides robust estimates towards GC-bias modeling

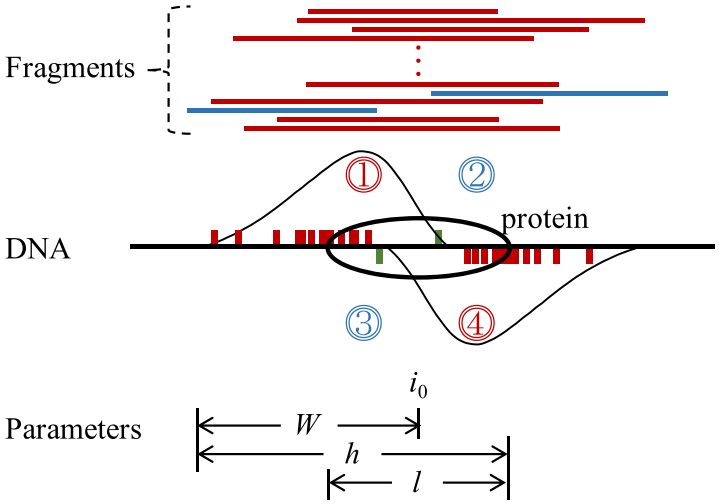


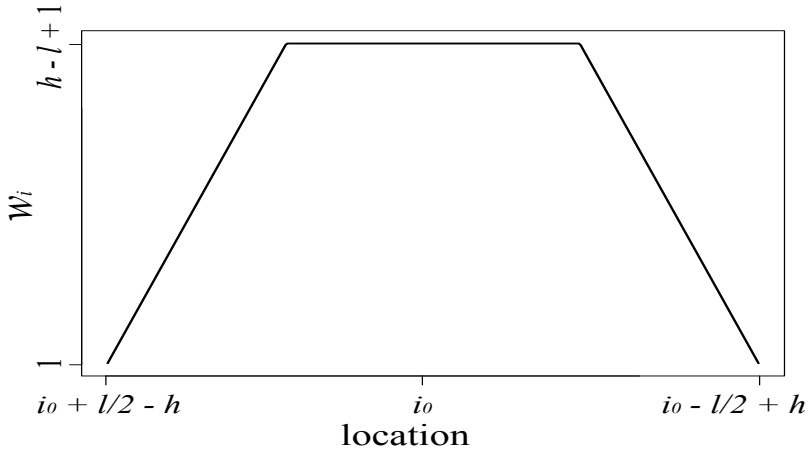


Normal GC-content

$$EGCC = \frac{1}{h(h-l+1)} \sum_{i=i_0+\frac{l}{2}-h}^{i_0-\frac{l}{2}+h} w_i x_i$$

Effective GC-content





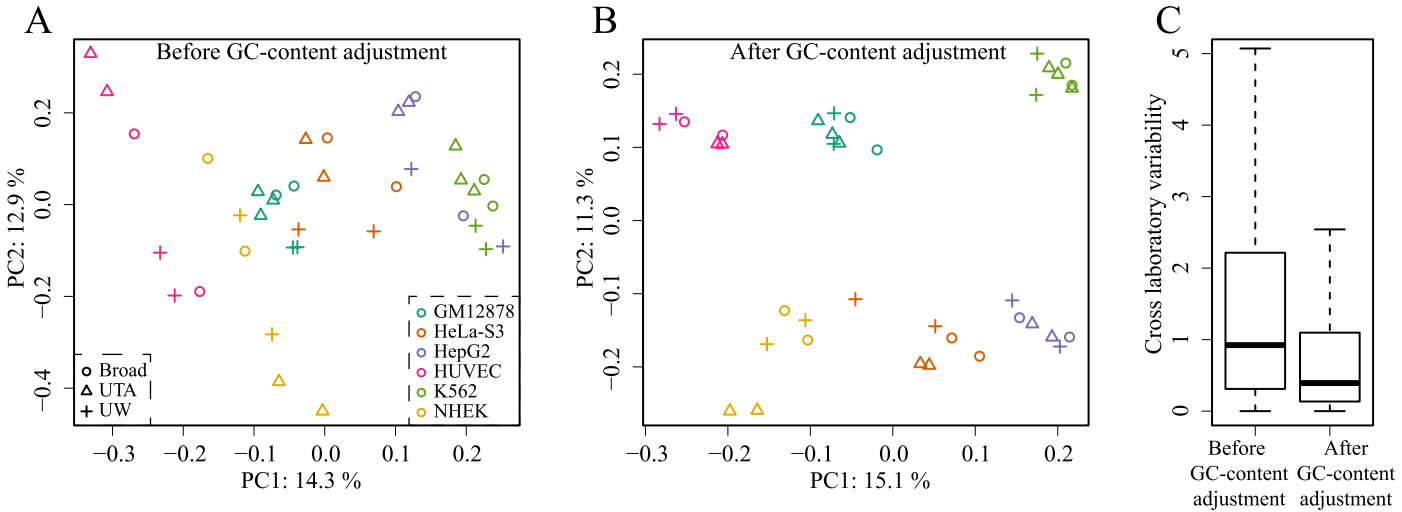
$$w_i = \begin{cases} i - i_0 - \frac{l}{2} + h & \text{if } i \in [i_0 + \frac{l}{2} - h, i_0 - \frac{l}{2}] \\ h - l + 1 & \text{if } i \in (i_0 - \frac{l}{2}, i_0 + \frac{l}{2}) \\ i_0 - \frac{l}{2} + h - i & \text{if } i \in [i_0 + \frac{l}{2}, i_0 - \frac{l}{2} + h] \end{cases}$$

GC effects adjustment in 6 cell lines based on ENCODE reported binding sites

https://www.encodeproject.org/data/annotations/v2/

PCA analysis

Mean square



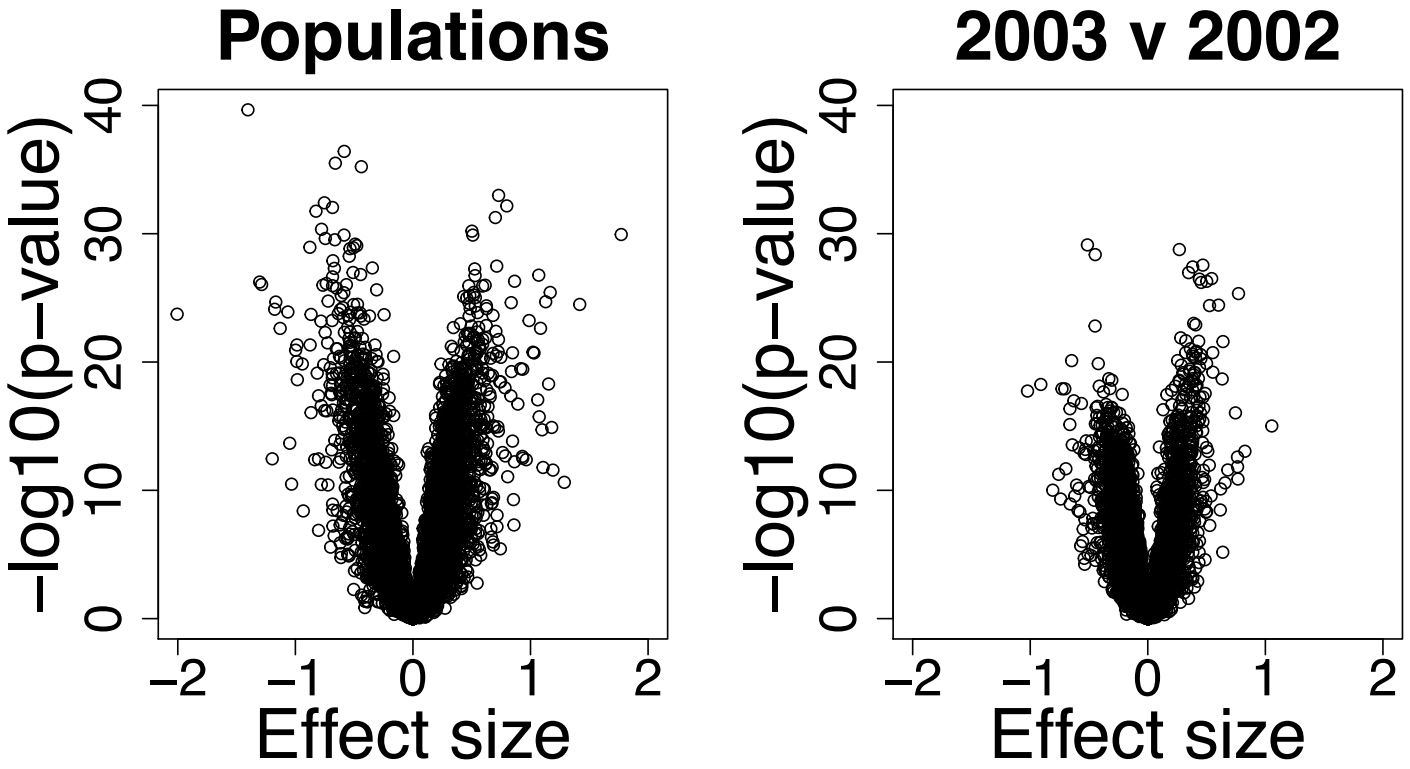
GC adjustment by: $(1 - \hat{Z}_i)\hat{f}_0(x_i) + \hat{Z}_i f_1(x_i)$

Links

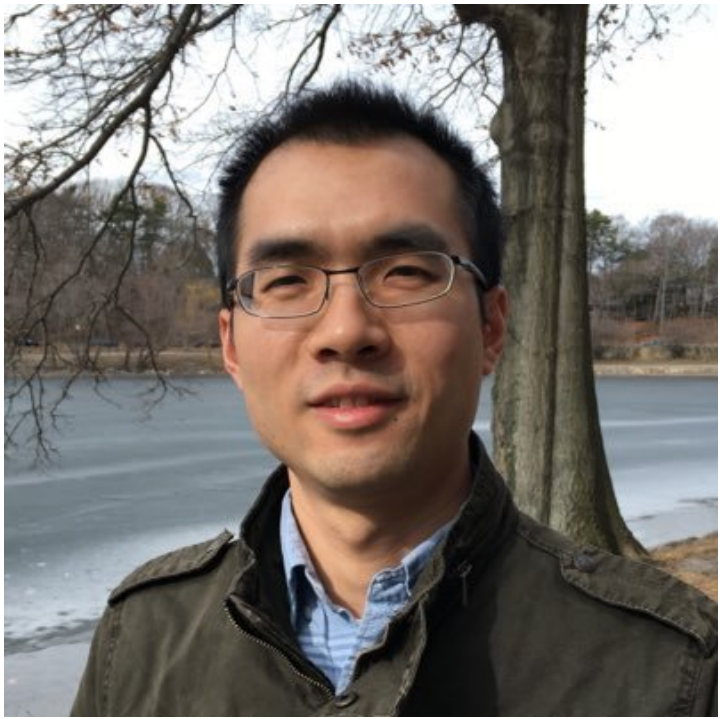
http://biorxiv.org/content/early/2017/01/15/090704

http://bioconductor.org/packages/devel/bioc/html/gcapc.html

Blood RNA from different ethnicities



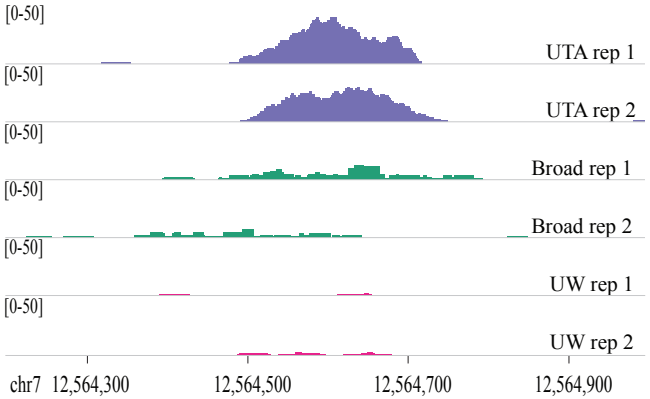
ChIP-Seq



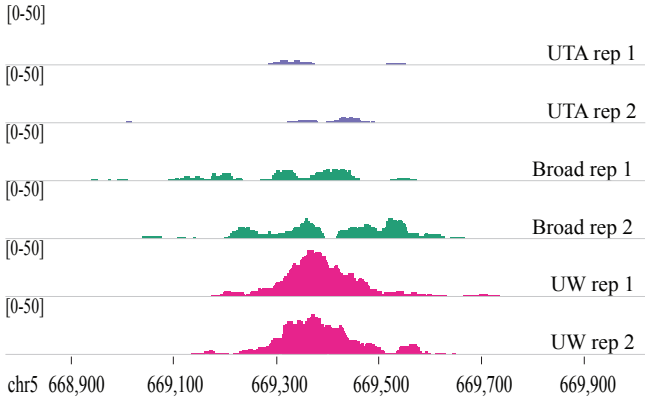
http://biorxiv.org/content/early/2016/12/01/090704

Coverage variability across replicates

Example 1



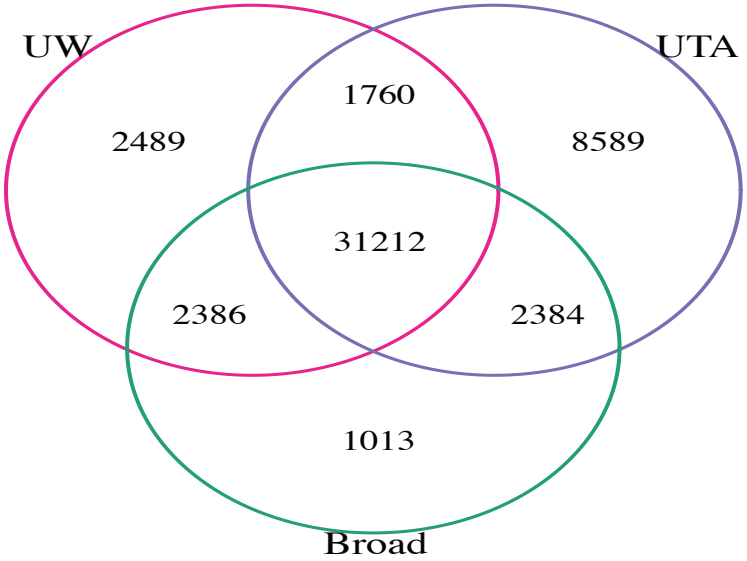
Example 2



These are CTCF binding sites for HUVEC cell-line

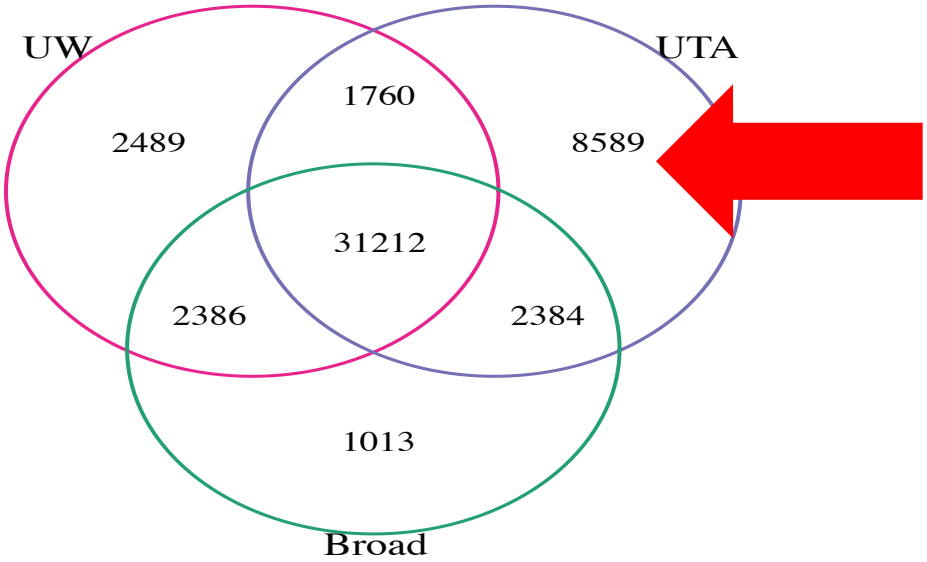
Reported peaks by ENCODE processing pipeline

24% of regions reported by only one lab



Reported peaks by ENCODE processing pipeline

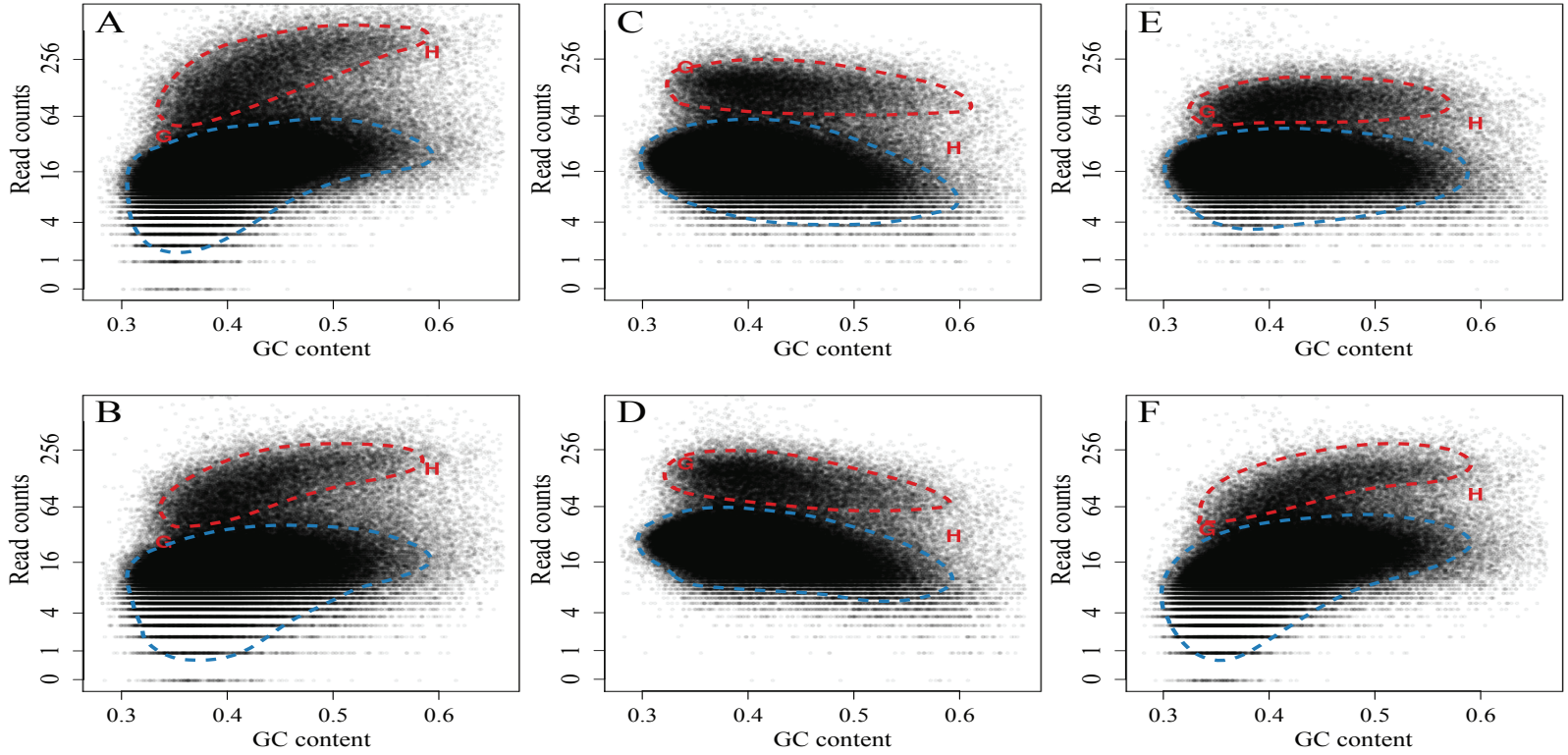
24% of regions reported by only one lab.



10kb Bin counts versus GC bias in CTCF

Rep 1

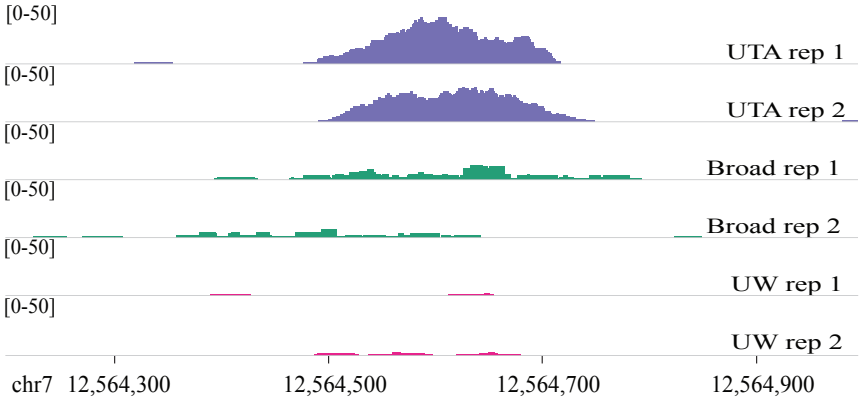
Rep 2



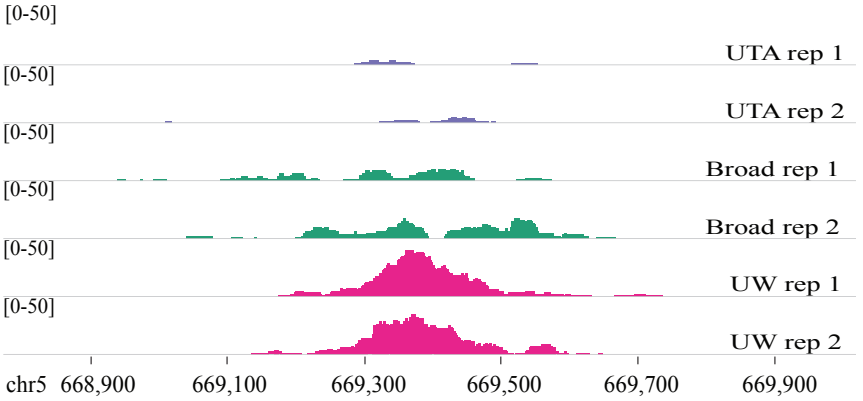
UW **UTA** **Broad**

Peak coverage variability across replicates

Region with low GC-content



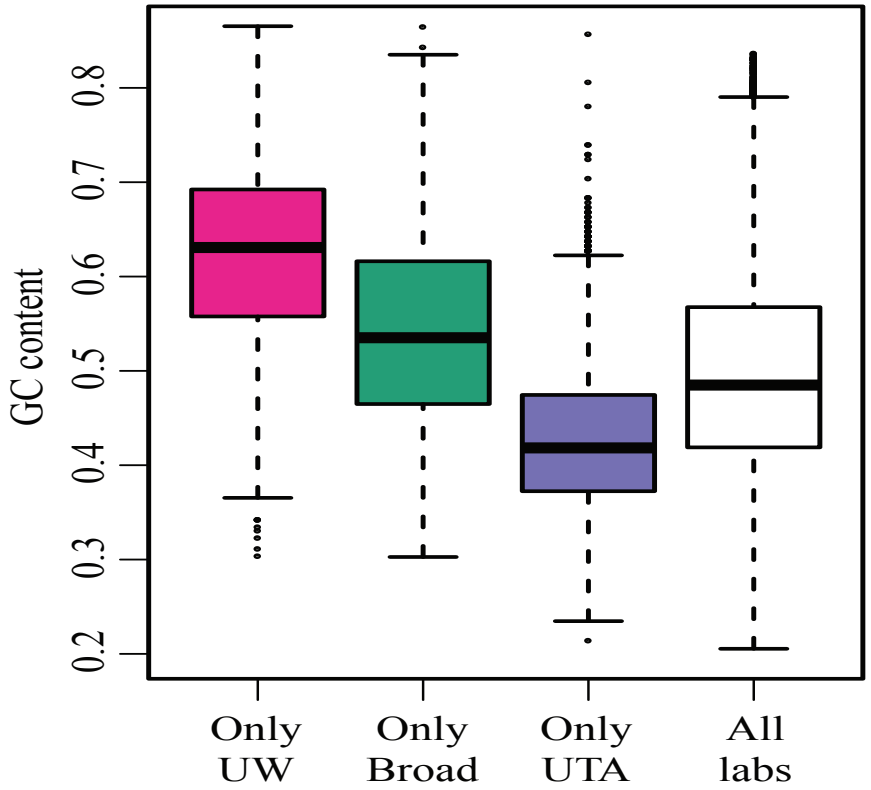
Region with high GC-content



Low GC-content results in higher coverage fo UTA samples

High GC-content resuts in higher coverage for UW samples

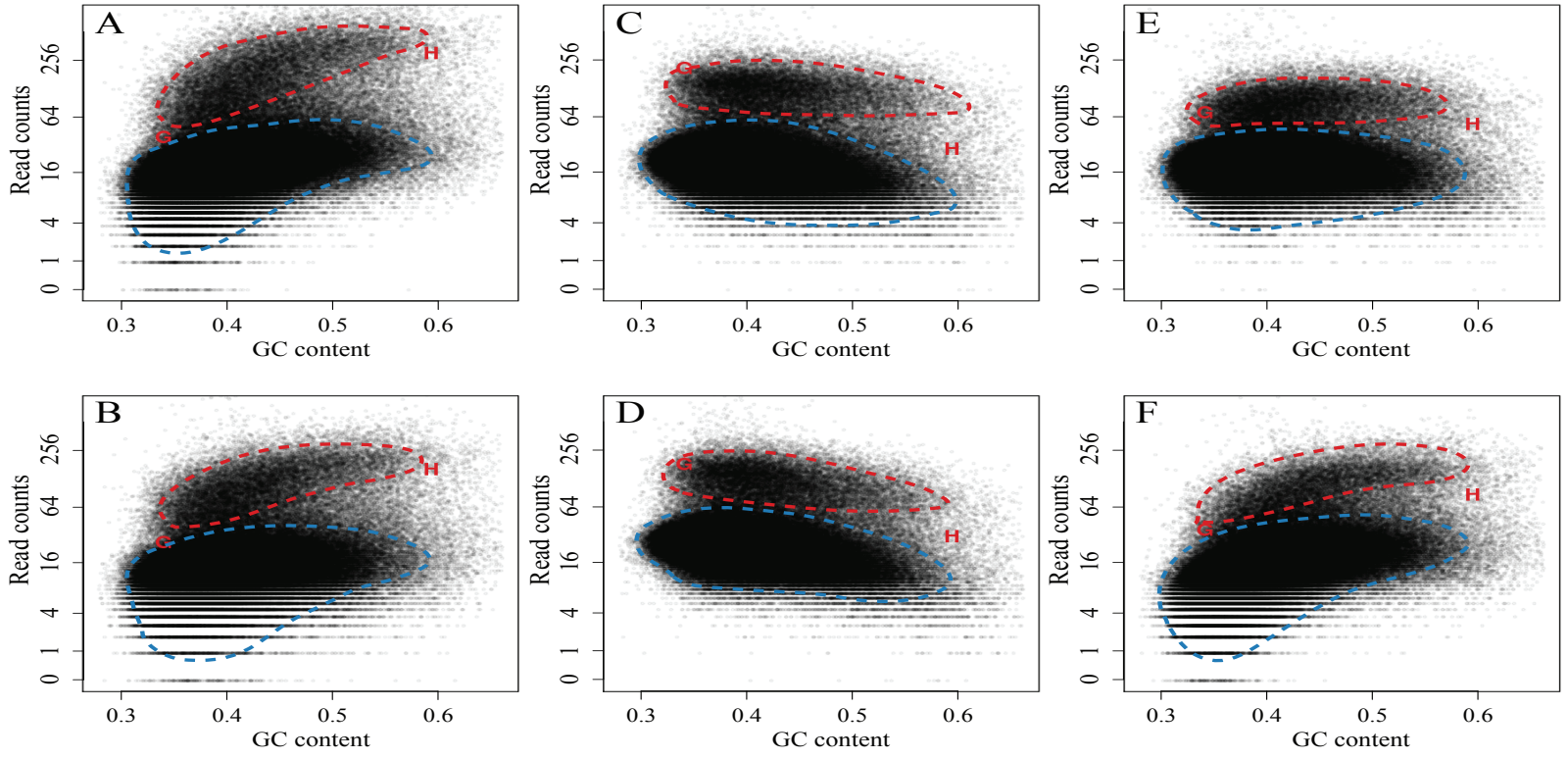
GC-content of peaks reported by only one lab



How do we adjust for GC-content bias?

Challenge: real binding regions tend to have higher GC

Both background and signal have bias



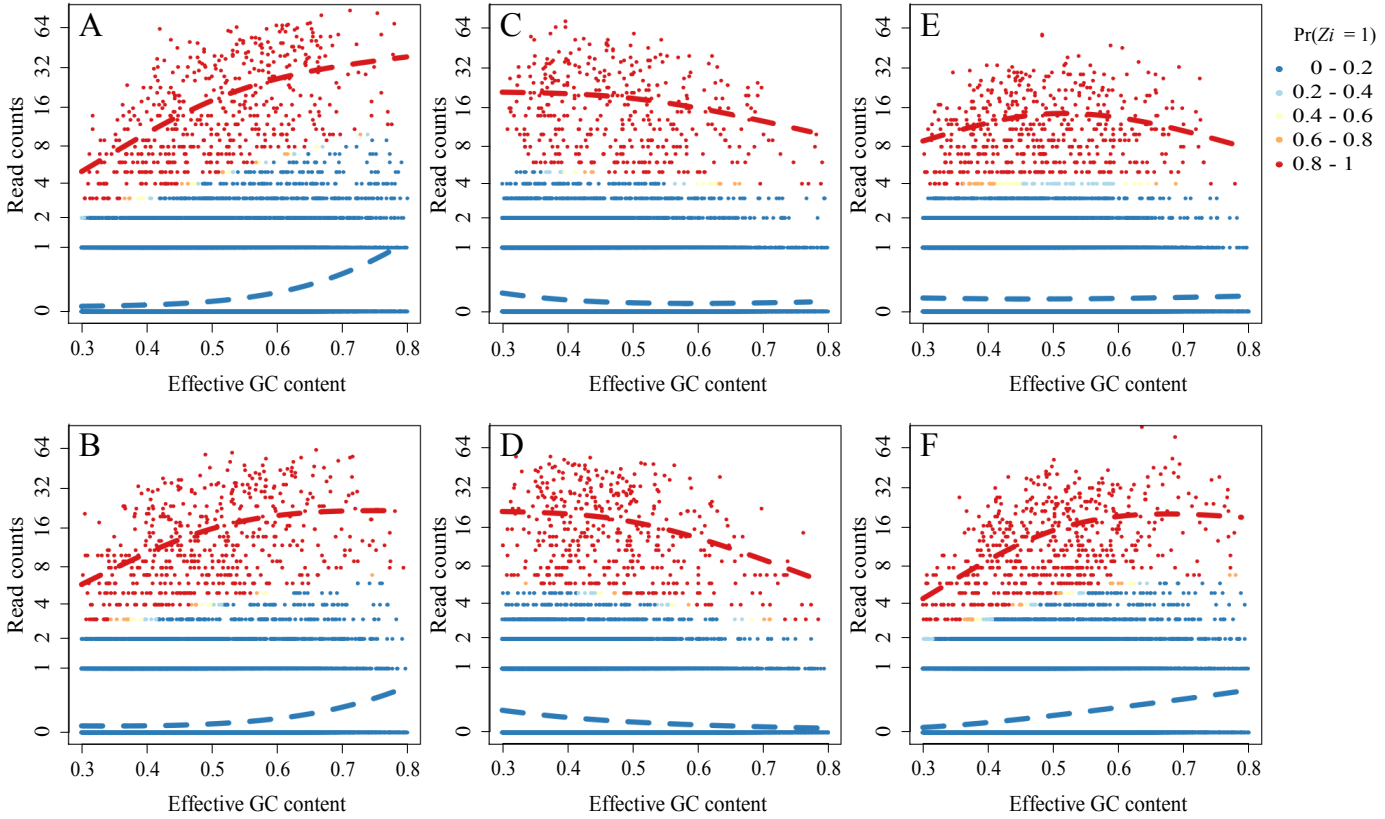
Correct bacEstimate GC-effects using mixture models

$$\log(\,\mathrm{E}[Y_i|Z_i = 0, X_i = x_i]\,) = \mu_0 + f_0(x_i)$$
$$\log(\,\mathrm{E}[Y_i|Z_i = 1, X_i = x_i]\,) = \mu_1 + f_1(x_i)$$

Assumptions:

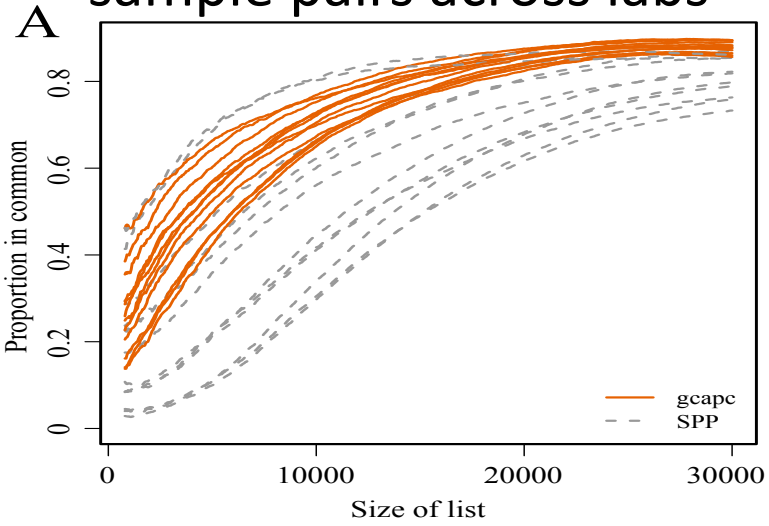
- Condition on *X* and *Z*, read counts Y_i follow Poisson distribution
- If region is in background $Z_i = 0$
- If region is in real peak $Z_i = 1$
- x_i is effective GC-content
- *f* is a smooth function

GC-effects estimated in CTCF samples

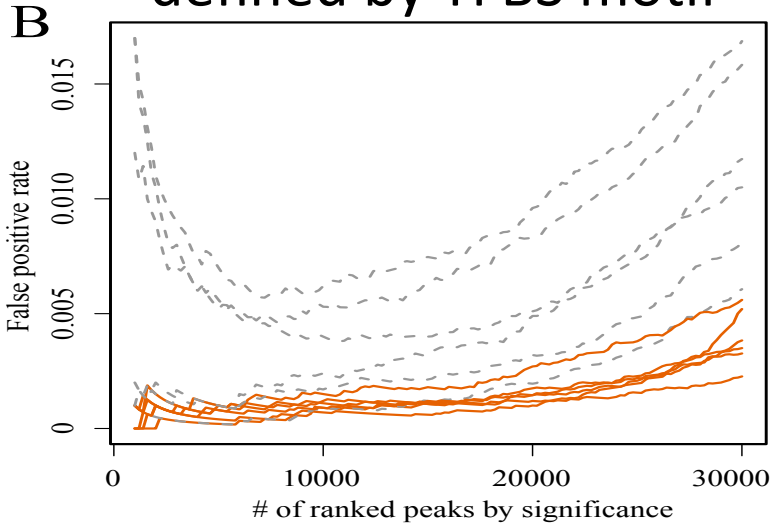


Improvements on consistency and false positive rate

Peak consistency between sample pairs across labs

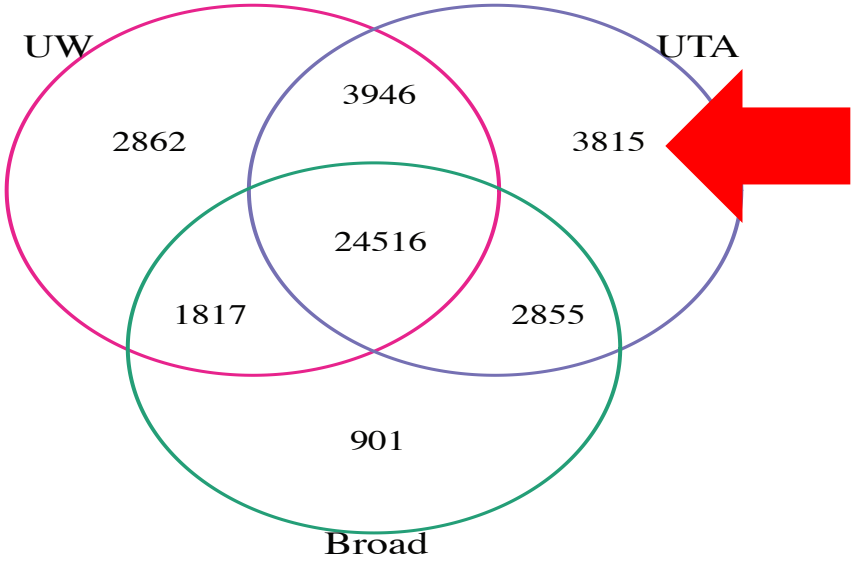


False positive peaks defined by TFBS motif



Other Improvements

Overlap between labs



GC-content of peaks reported by only one lab