NIMBUS: a *N*egative-binomial regression based *I*ntegrative *M*ethod for *BU*rden analysis of *S*omatic variants in cancer genomes.

JZ

Part 1

About the name

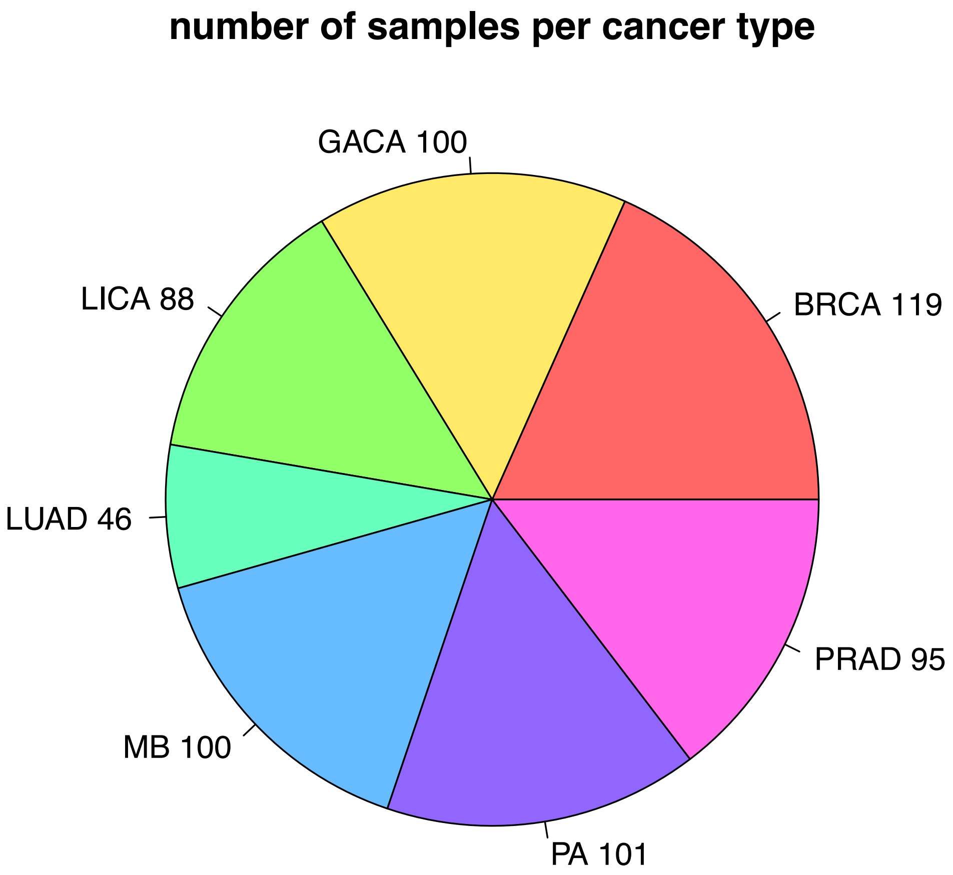
- NIMBUS (proposed by Jason Liu):
- 1. a luminous cloud or a halo surrounding a supernatural being or a saint
- 2. a large gray rain cloud

Other candidates

- *CRISIS*: *C*ancer Mutational Hotspot Detection using *R*egression based *I*ntegrative *S*omatIc Buden Analys*is*
- *BRISC*: a Negative *B*inomial *R*egression based *I*ntegrative *S*cheme for *C*ancer mutation burden analysis
- *REsIST*: A *RE*gression ba*s*ed *I*ntegrative *S*omatIc buden analysis Tool in Cancer genomes

Other candidates

- *aROMA*: *a* *R*egression based Framework s*O*matic *M*utation Burden *A*nalysis Tool
- *aRISE*: *a* *R*Egression based *I*ntegrative Framework for *S*omatic Mutation Burd*e*n analysis in Cancer genomes
- *REFOrM*: A *RE*gression based *F*ramework s*O*matic *M*utation Burden analysis



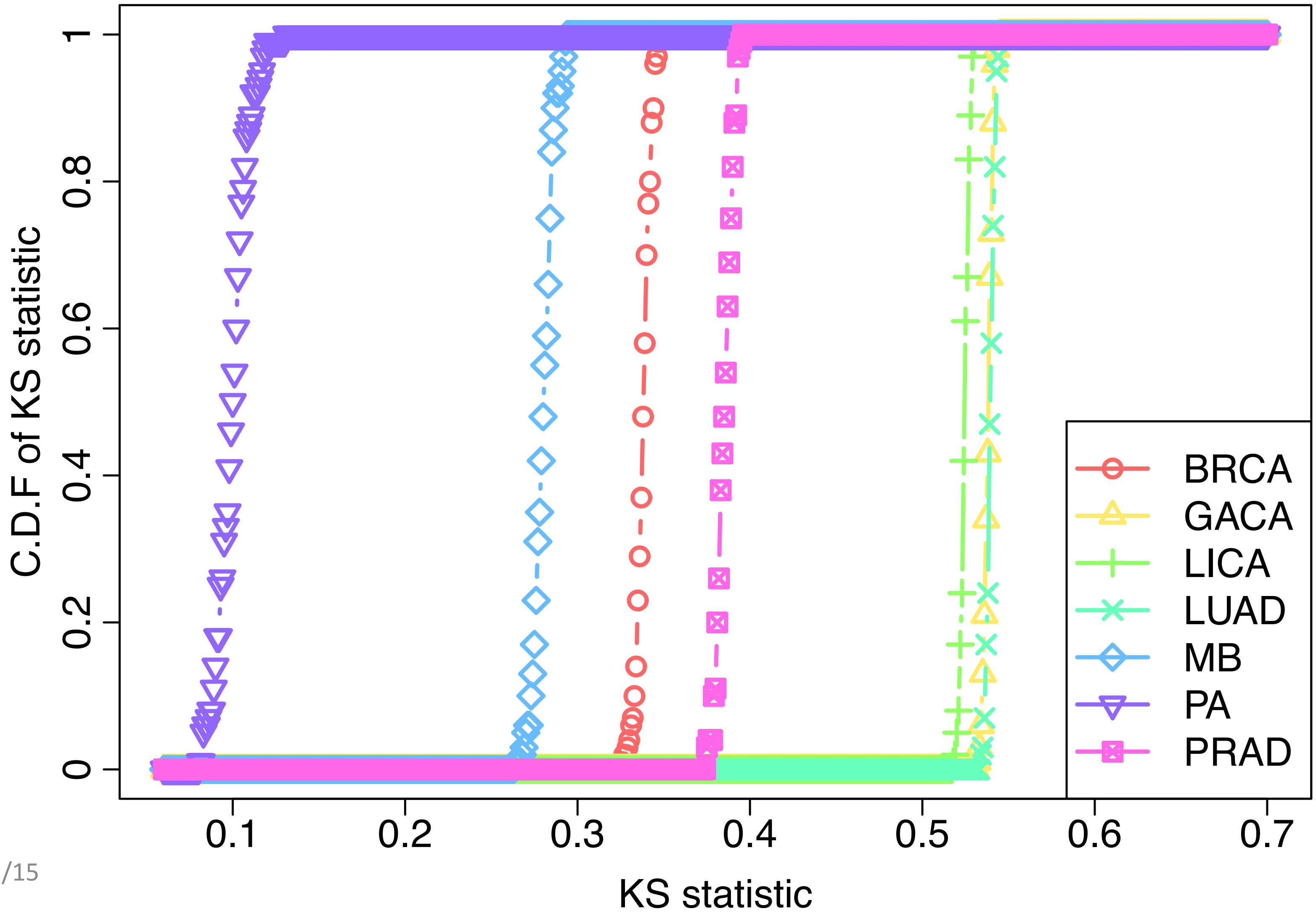
- Supp. 1 cancer type and sample number
- Selection Criterion:
  - Decent number of samples
  - Decent number of total variants
- Otherwise no need to do background correction!

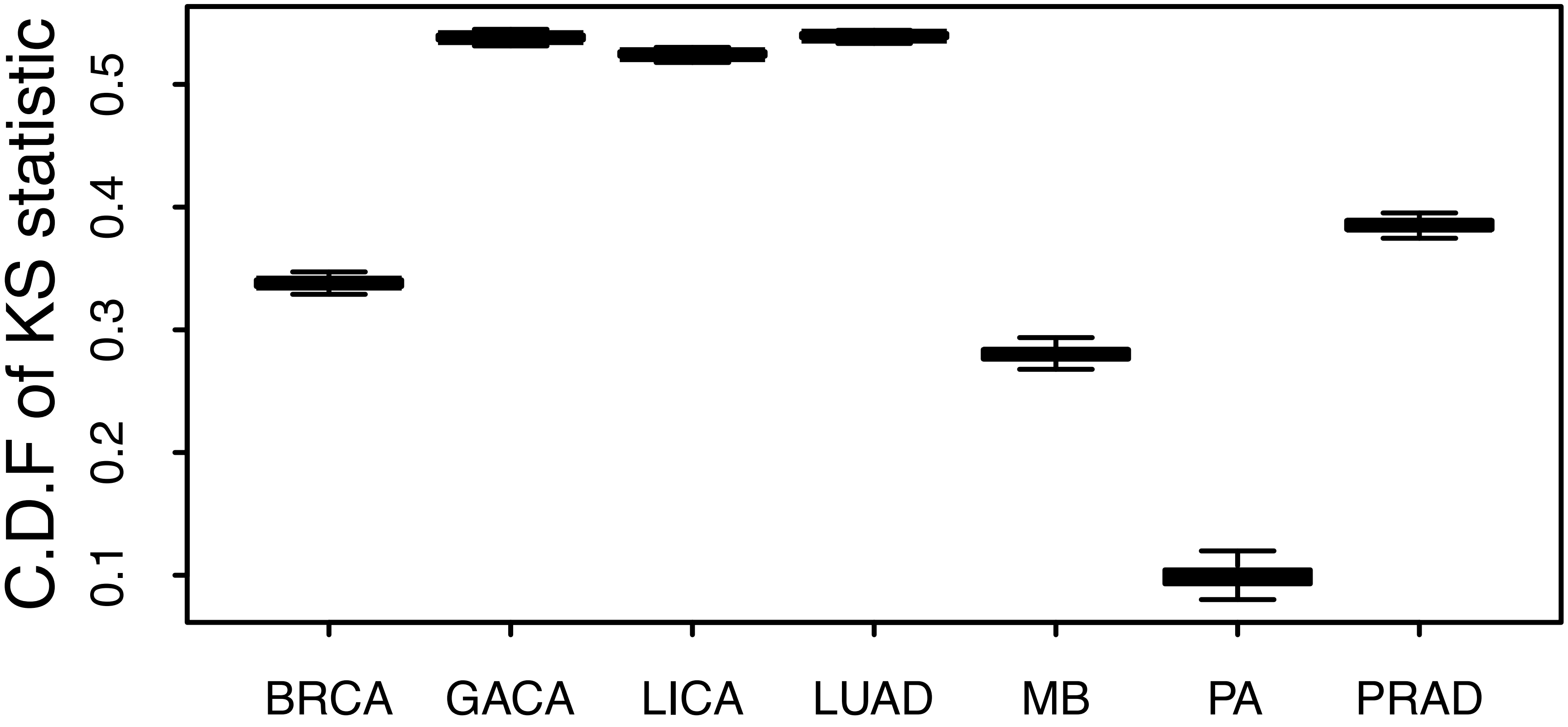
cancer	median	sd	max	min	variantNum
BRCA	3705	7300.526	67,506	1113	675897
GACA	14429.5	71372.080	407524	893	3522792
LICA	8706.5	5522.917	28446	1439	881141
LUAD	21287	35610.839	145548	1743	1509946
MB	965	1196.036	9272	44	125333
PA	70	114.414	769	2	10626
PRAD	4927	2764.873	18225	931	472176

Problem: lots of overdispersion in the observed variant count data

Method: per cancer type

1. Training data n bins, estimate Poisson lambda
2. Randomly generate n variant counts, calculate KS statistic of observed VS. generated
3. Repeat 2 for 100 times and plot C.D.F




C.D.F of KS statistic
0.1
0.2
0.3
0.4
0.5
BRCA
GACA
LICA
LUAD
MB
PA
PRAD

Key Problem in driver detection

- Control heterogeneity for strict false positive and negative control
- Source of heterogeneity
  - – Challenge 1: Cancer heterogeneity: separate cancer types
  - – Challenge 2: Sample heterogeneity: negative binomial
  - – Challenge 3: Baseline changes due to external covariates : regression

*NIMBUS: N*egative-binomial regression based *I*ntegrative *M*ethod for *BU*rden analysis

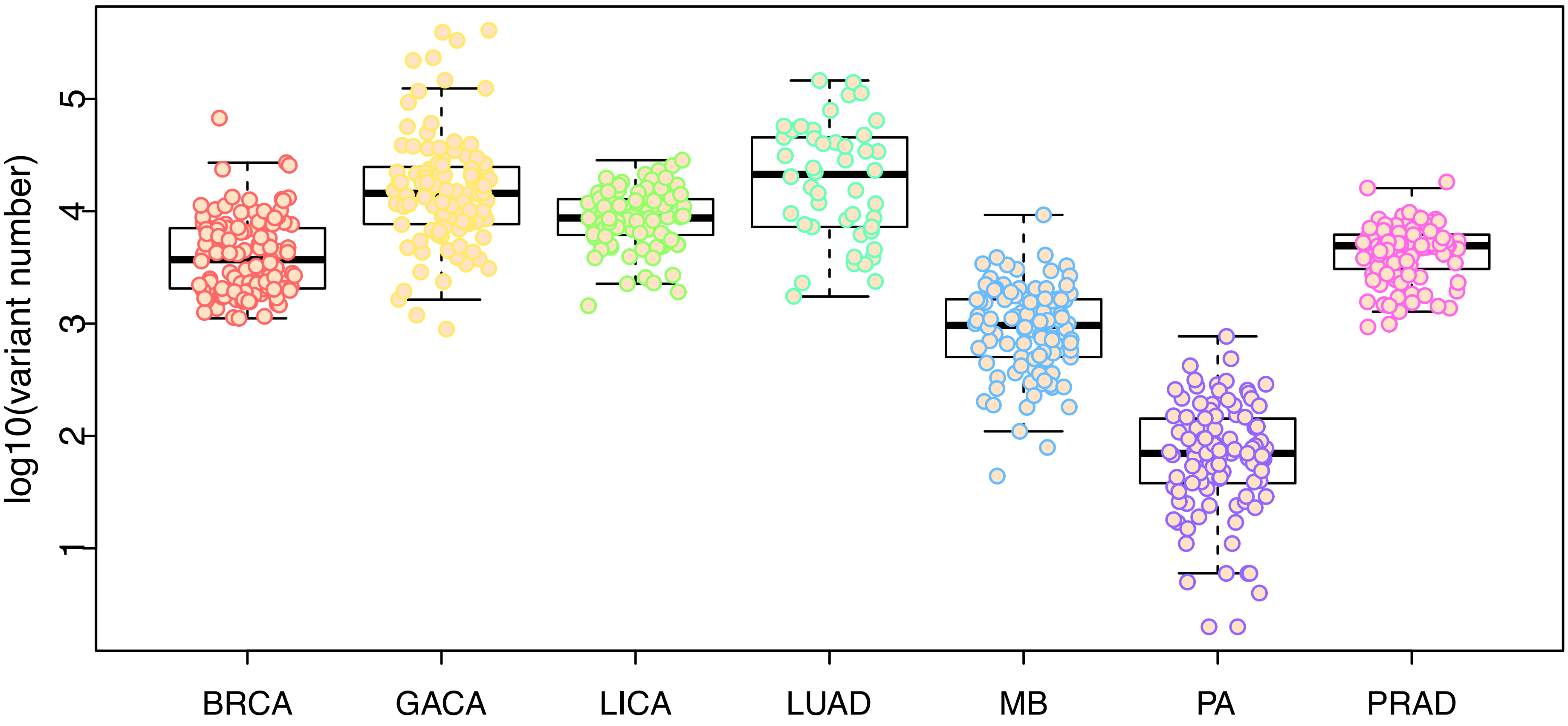
Key Problem in driver detection

- Control heterogeneity for strict false positive and negative control
- Source of heterogeneity
  - Challenge 1: Cancer heterogeneity: separate cancer types
  - Challenge 2: Sample heterogeneity: negative binomial
  - Challenge 3: Baseline changes due to external covariates : regression

Why not separate cancer samples?

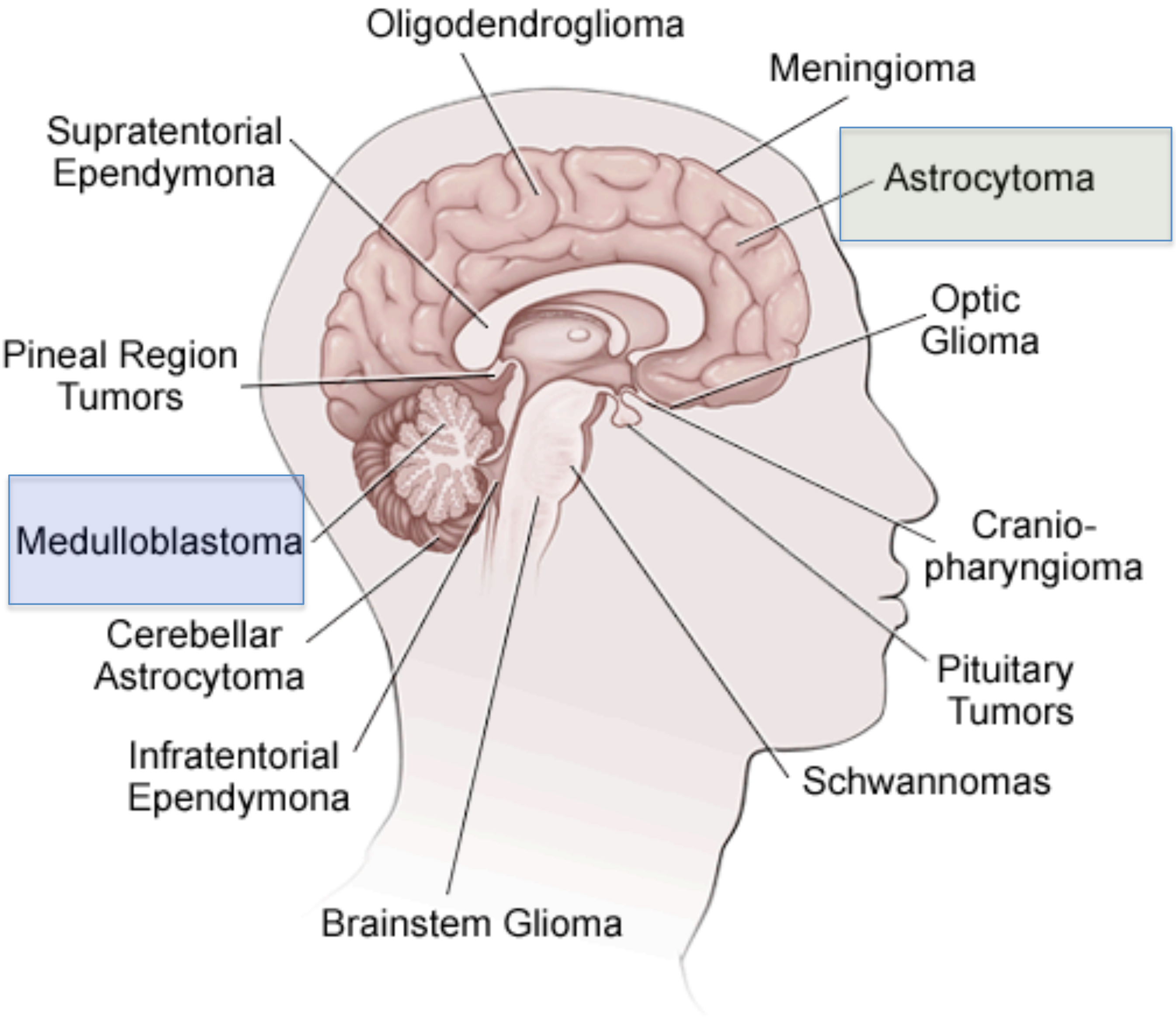
*NIMBUS: N*egative-binomial regression based *I*ntegrative *M*ethod for *BU*rden analysis

Challenge1 & 2: Lots of cancer type & sample heterogeneity

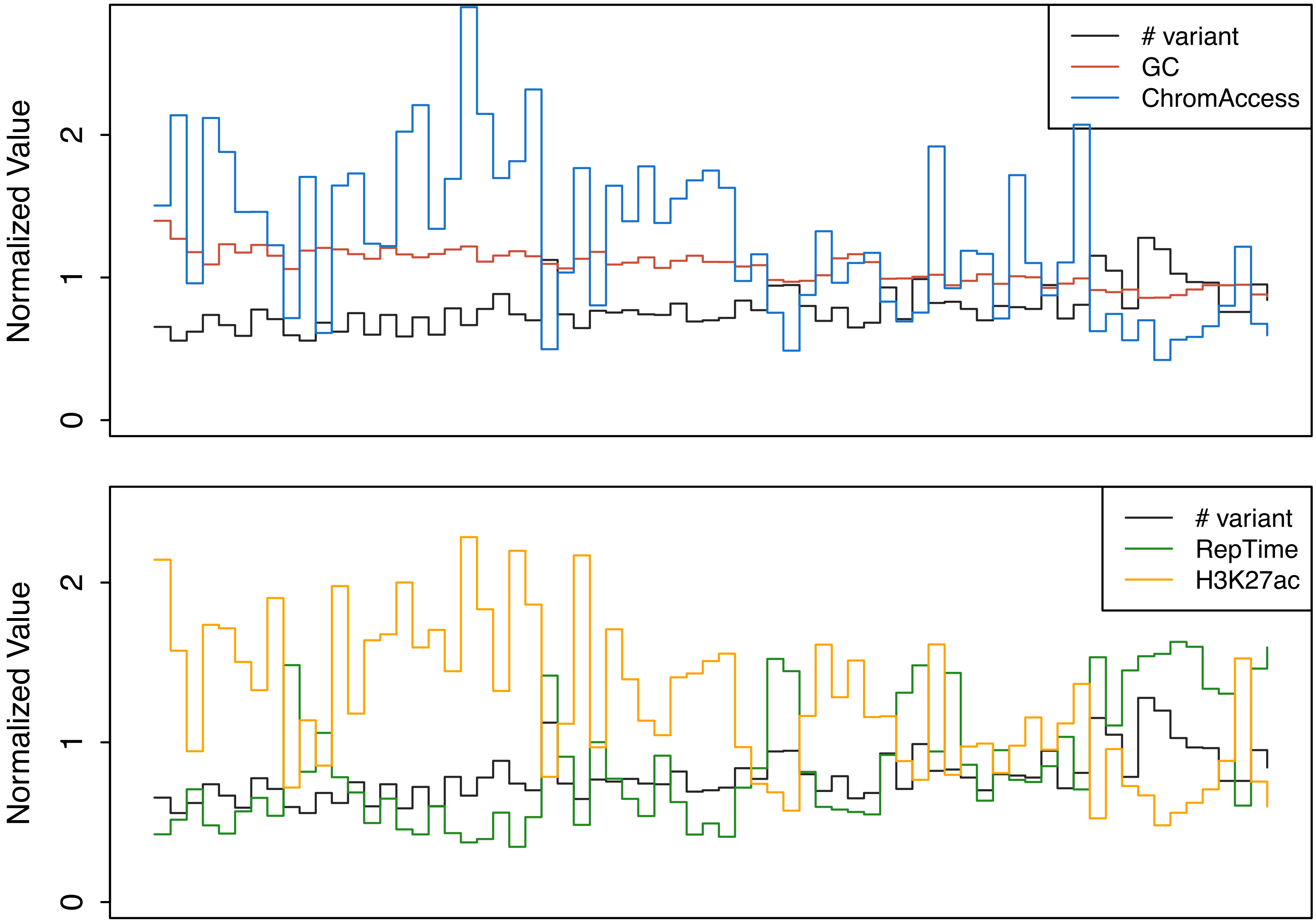


- MB: Medulloblastoma, #1 children brain cancer
- PA: Pilocytic_Astrocytoma, more in children and young, very quite genome

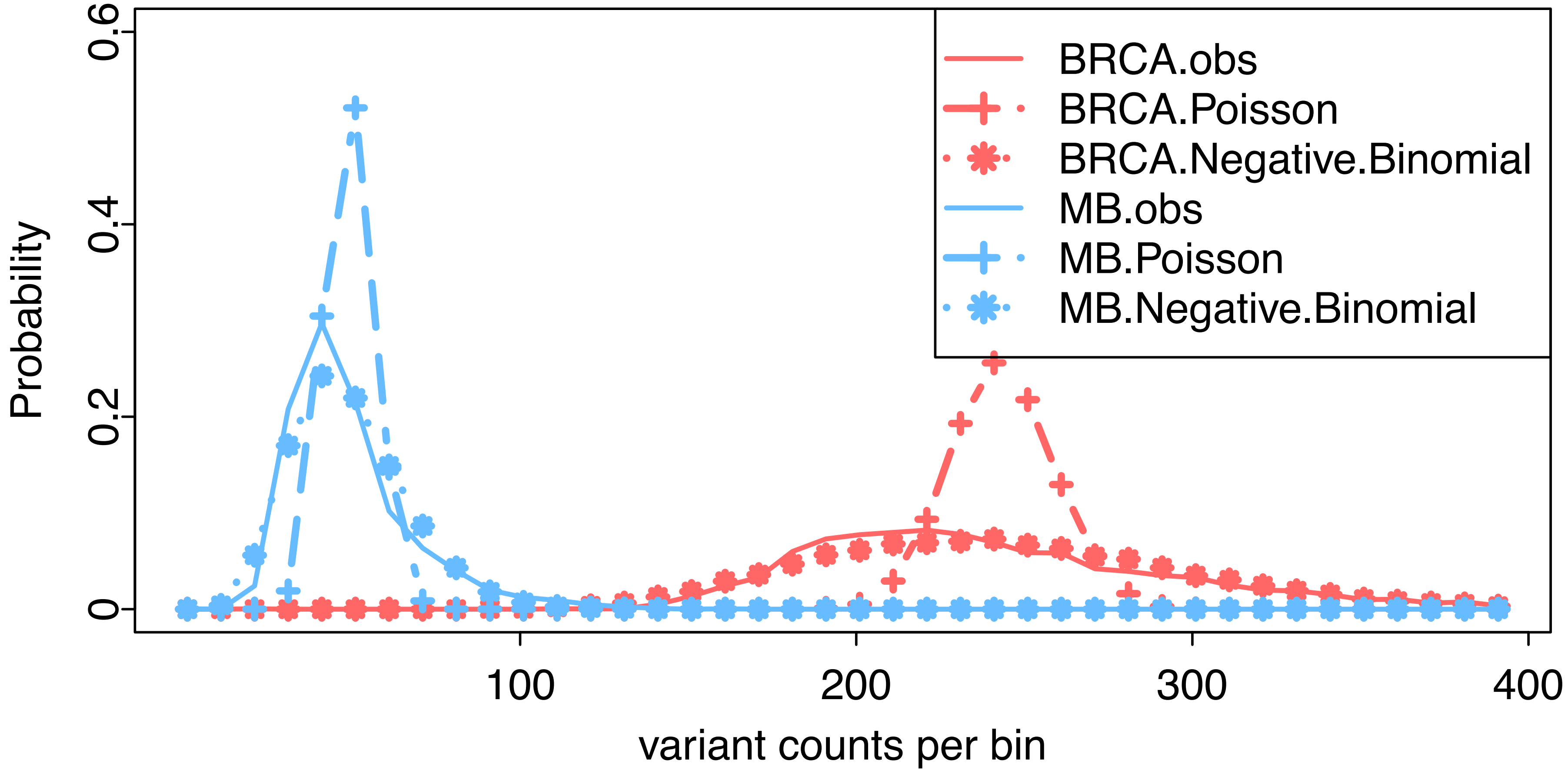
Location of Different Types of Brain Tumors

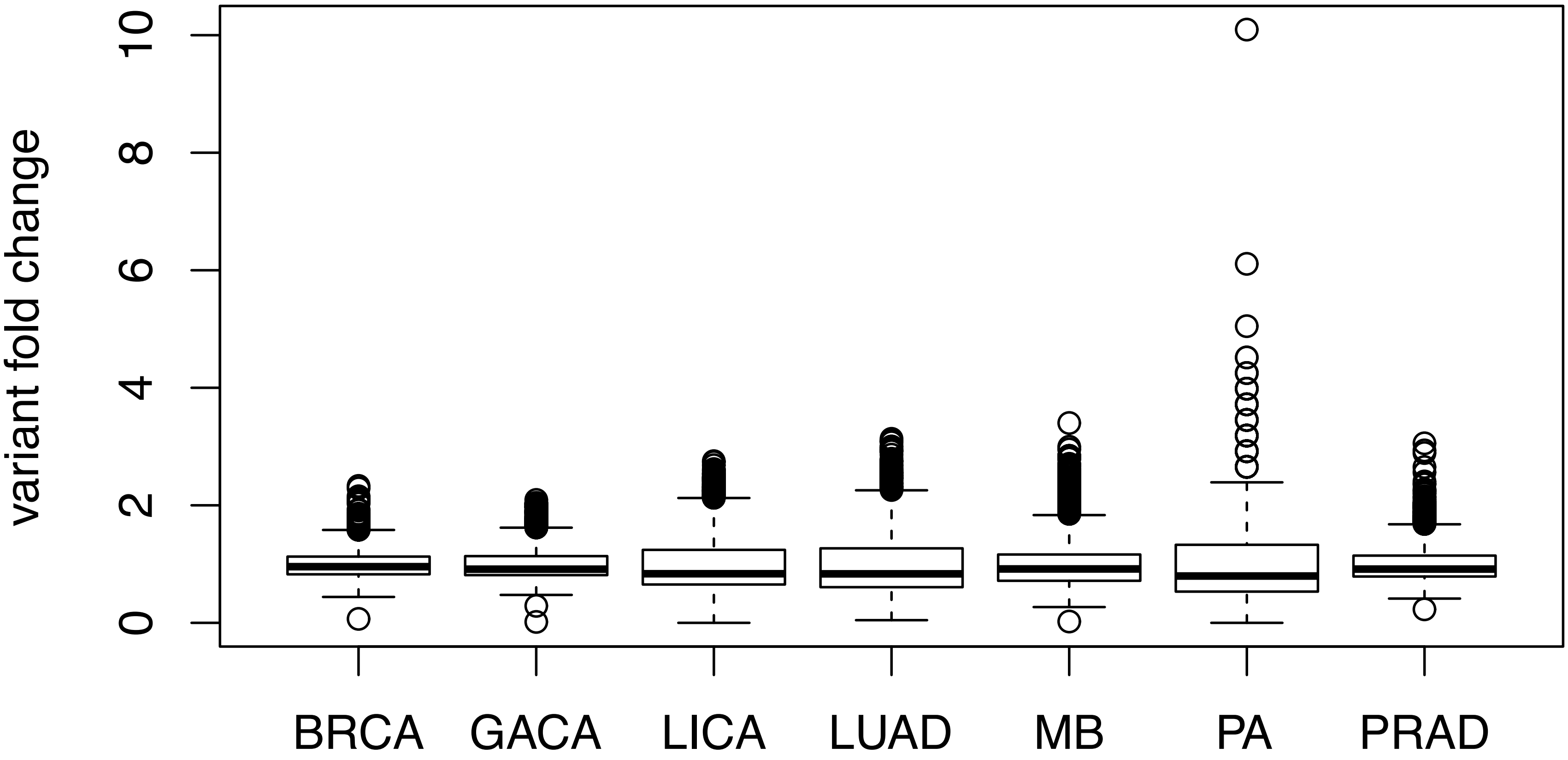


Challenge3: regional heterogeneity due to external variants



Result of challenge 1-3: overdispersion of mutation counts




variant fold change
0
2
4
6
8
10
BRCA
GACA
LICA
LUAD
MB
PA
PRAD

NIMBUS Model

$$NBI(y,\mu,\sigma)=\frac{\Gamma\left(\frac{1}{\sigma}+y\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(1+y)}\left(\frac{\sigma\mu}{1+\mu\sigma}\right)^{y}\left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}}$$

Negative Binomial distribution of Type I

Incoming rate is a Gamma random variable

$$Y|z\sim Pois(\lambda z)=\frac{e^{-\lambda z}(\lambda z)^{y}}{y!}$$

$$z\sim Gamma(\theta,\theta)=\frac{\theta^{\theta}z^{\theta-1}e^{-\theta z}}{\Gamma(\theta)}$$

Marginal distribution of Y

$$f_{Y}(y)=\int_{0}^{+\infty}\frac{e^{-\lambda z}(\lambda z)^{y}}{y!}\frac{\theta^{\theta}z^{\theta-1}e^{-\theta z}}{\Gamma(\theta)}dz$$

$$=\frac{\Gamma(\theta+y)}{\Gamma(y+1)\Gamma(\theta)}\left(\frac{\theta}{\lambda+\theta}\right)^{\theta}\left(\frac{\lambda}{\lambda+\theta}\right)^{y}$$

$$\boxed{\mu=\lambda,\sigma=\frac{1}{\theta},\sigma\uparrow\Rightarrow\theta\downarrow\Rightarrow\text{overdispersion}\uparrow}$$

- Mutation rate varies among cancer types
- Different cancer types have quite different mutational patterns

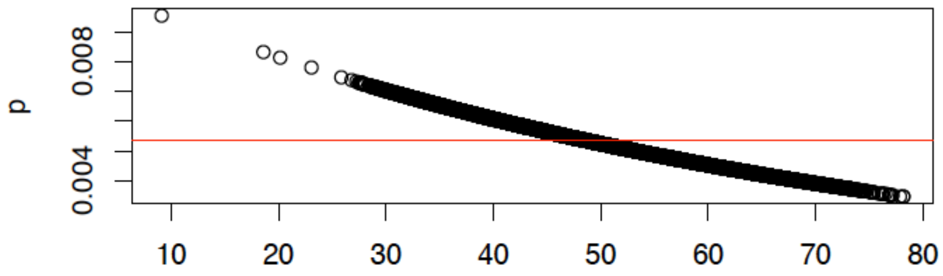
$$g_1\left(\underbrace{\mu_i}_{\downarrow}\right) = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,k}\beta_k + \cdots x_{i,K}\beta_K$$

By pooling the variants together we are assuming the same covariate coefficients

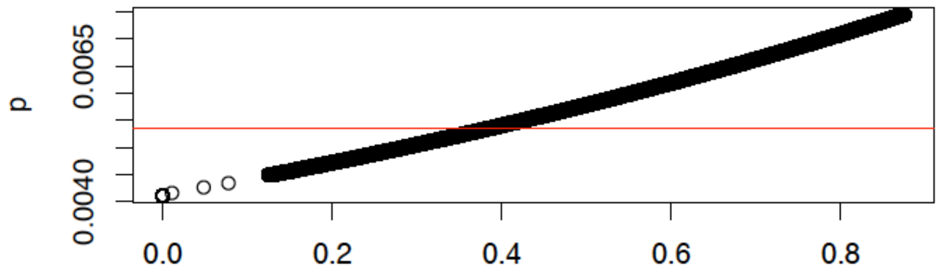
Covariate list:

- GC content
- replication timing,
- 7 histone modifications marks from Roadmap
- chromatin status from Roadmap
- mRNA-seq
- DNA-Methylation

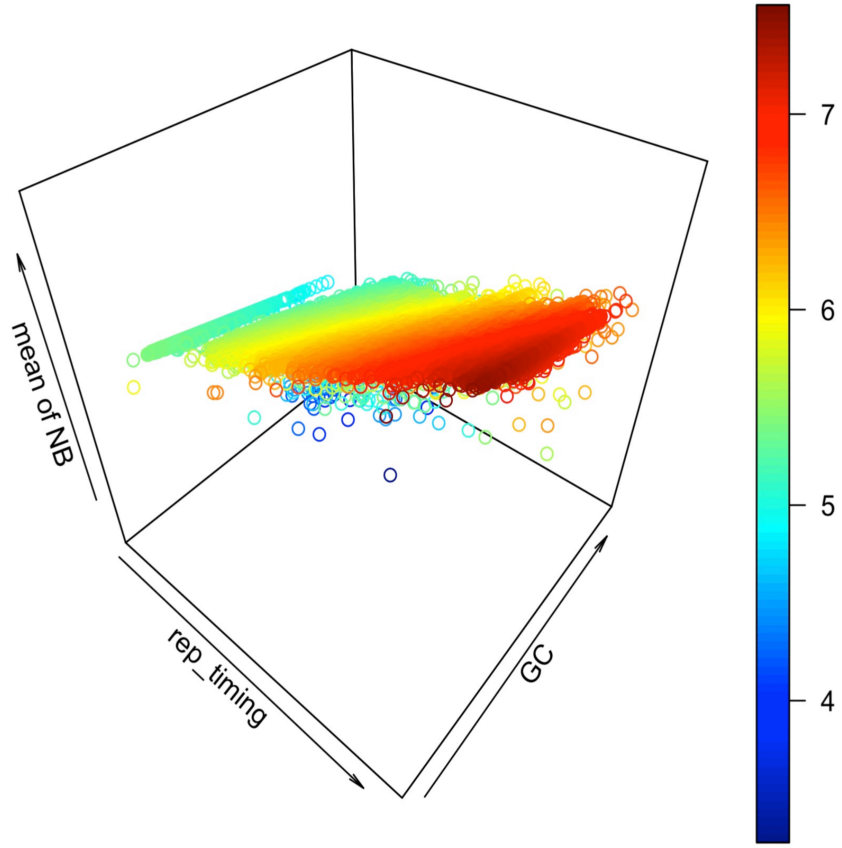
Toy Example of how the fitting works



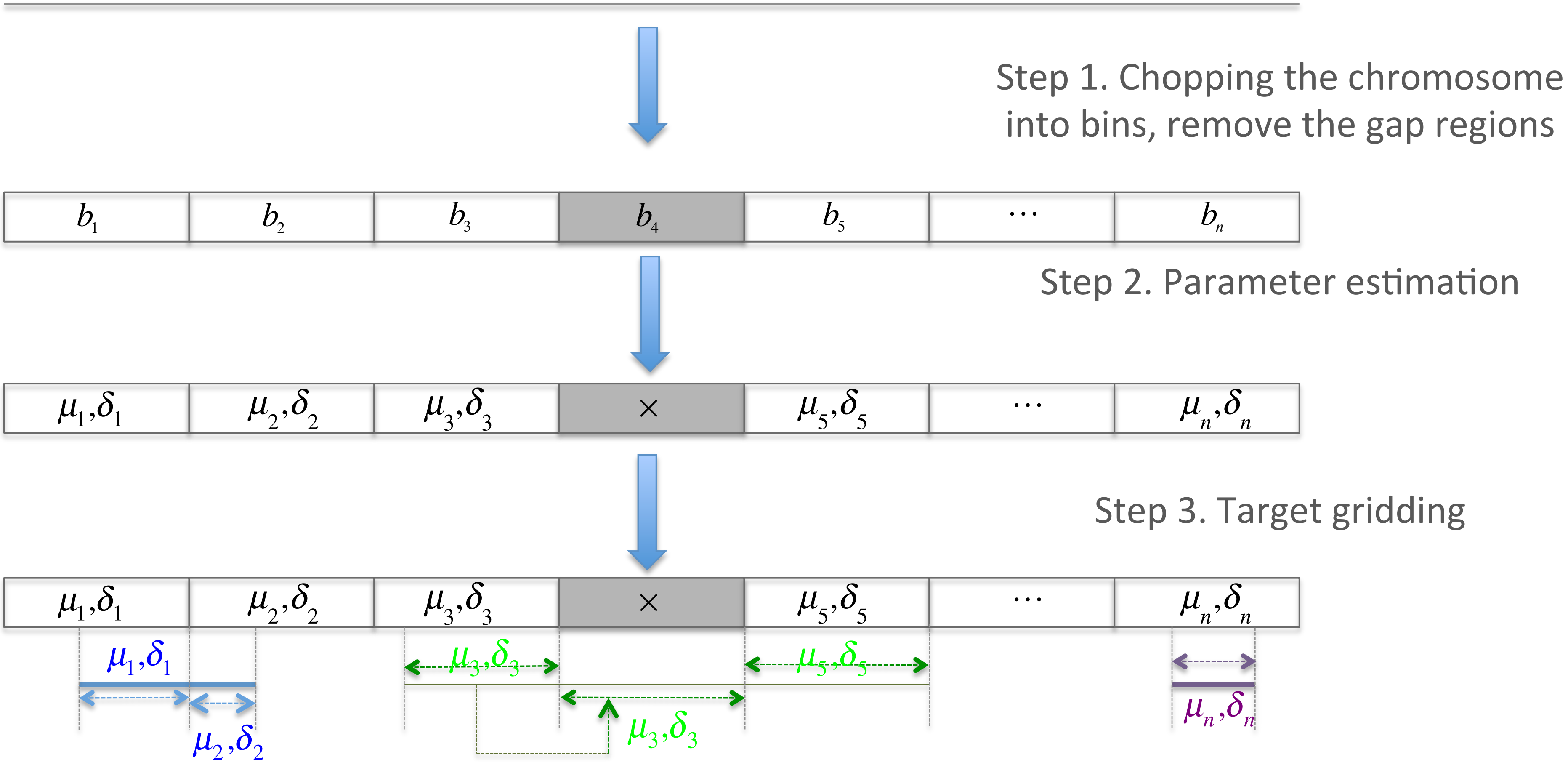
GC content



Replication timing



Scheme of the gridding method in NIMBUS



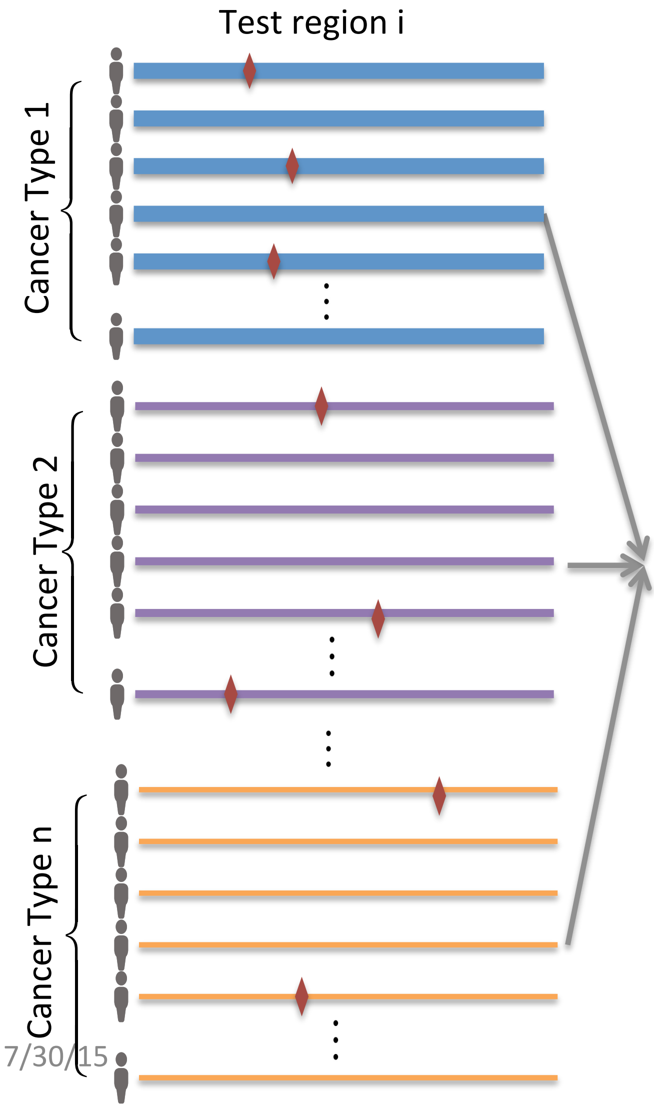
Step 1. Chopping the chromosome into bins, remove the gap regions

Step 2. Parameter estimation

Step 3. Target gridding

b_1 b_2 b_3 b_4 b_5 $\cdots$ b_n
μ_1,δ_1 μ_2,δ_2 μ_3,δ_3 $\times$ μ_5,δ_5 $\cdots$ μ_n,δ_n
μ_1,δ_1 μ_2,δ_2 μ_3,δ_3 $\times$ μ_5,δ_5 $\cdots$ μ_n,δ_n
μ_1,δ_1
μ_2,δ_2
μ_3,δ_3
μ_5,δ_5
μ_3,δ_3
μ_n,δ_n

Flowchart of NIMBUS



Calculate covariant matrix

	Tissue 1	Tissue2	...
Histone Modifications			
Chromatin Accessibility			
GC, CpG			
expression			
Replication timing			

PCA based regression for neighbor definitions

$$\hat{\mu}_{1i}:\ \text{estimate of}\ \left\{k \middle| \left| g^{-1}\left(\mathbf{X}_i\hat{\mathbf{b}}_{\mathbf{1}}\right) - g^{-1}\left(\mathbf{X}_k\hat{\mathbf{b}}_{\mathbf{1}}\right)\right| < \xi\right\} \Rightarrow p_{i,1}$$

$$\hat{\mu}_{2i}:\ \text{estimate of}\ \left\{k \middle| \left| g^{-1}\left(\mathbf{X}_i\hat{\mathbf{b}}_2\right) - g^{-1}\left(\mathbf{X}_k\hat{\mathbf{b}}_2\right)\right| < \xi\right\} \Rightarrow p_{i,2}$$

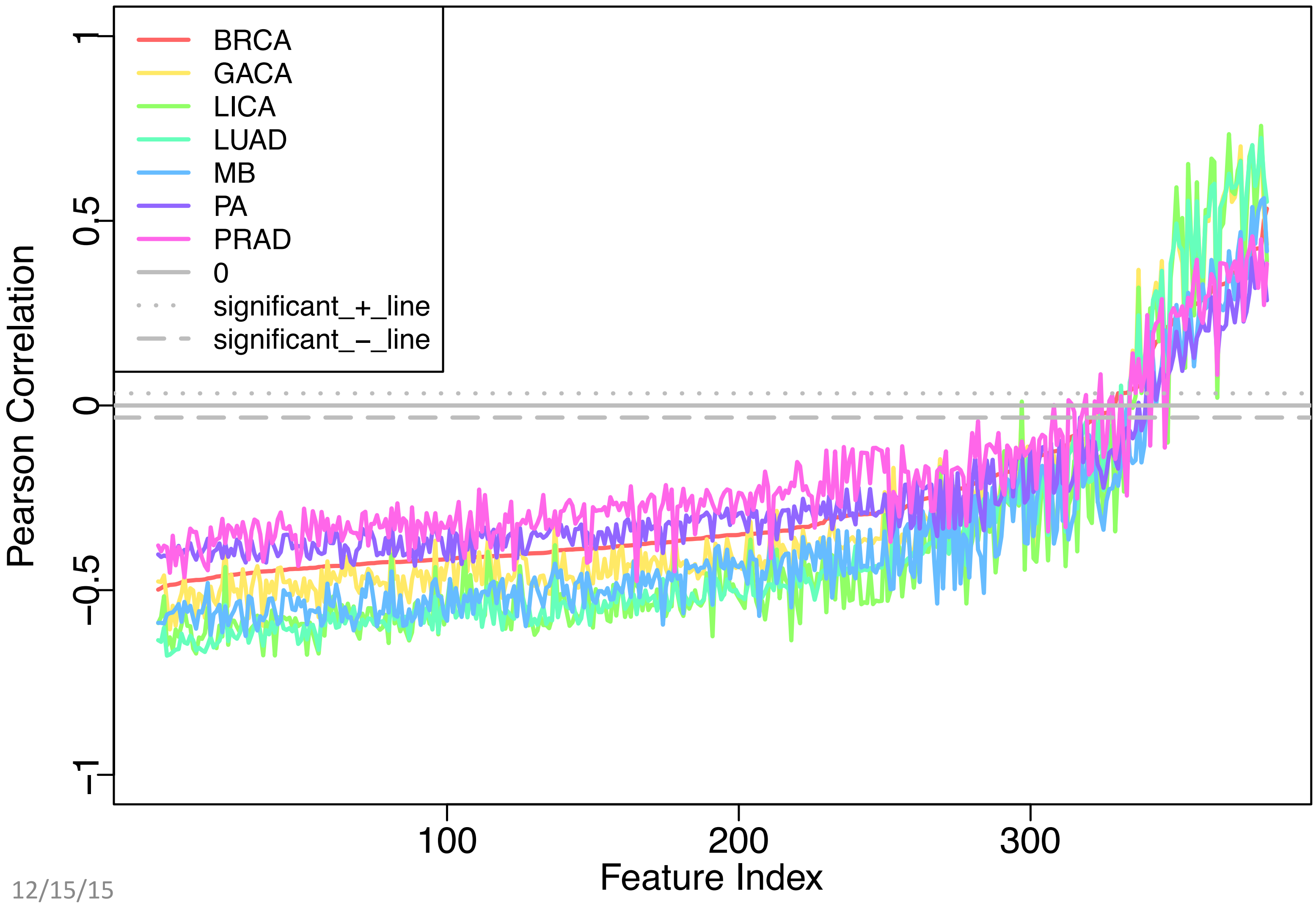
$$\vdots$$

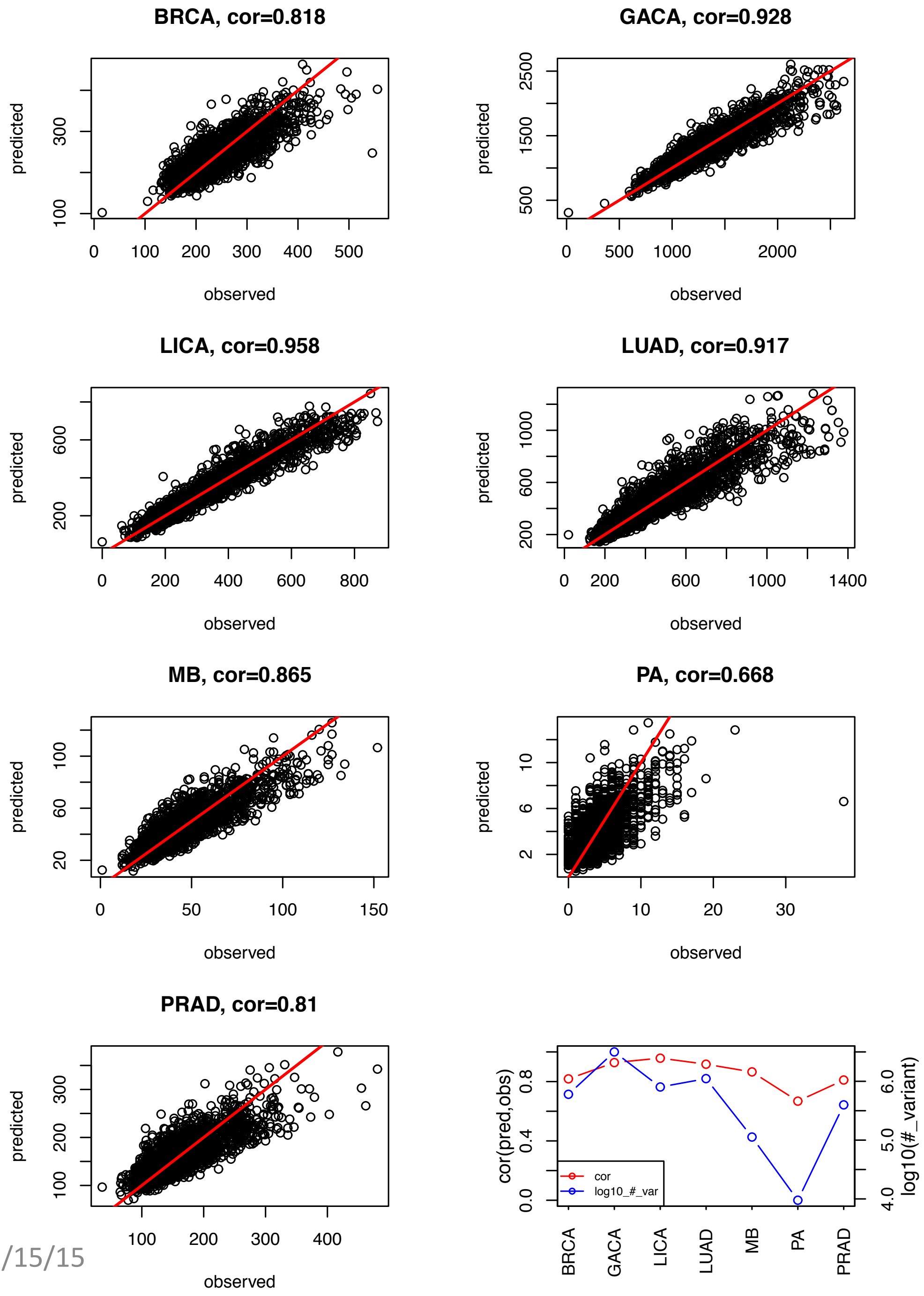
$$\hat{\mu}_{ki}:\ \text{estimate of}\ \left\{k \middle| \left| g^{-1}\left(\mathbf{X}_i\hat{\mathbf{b}}_n\right) - g^{-1}\left(\mathbf{X}_k\hat{\mathbf{b}}_n\right)\right| < \xi\right\} \Rightarrow p_{i,n}$$

$$\hat{p}_i$$

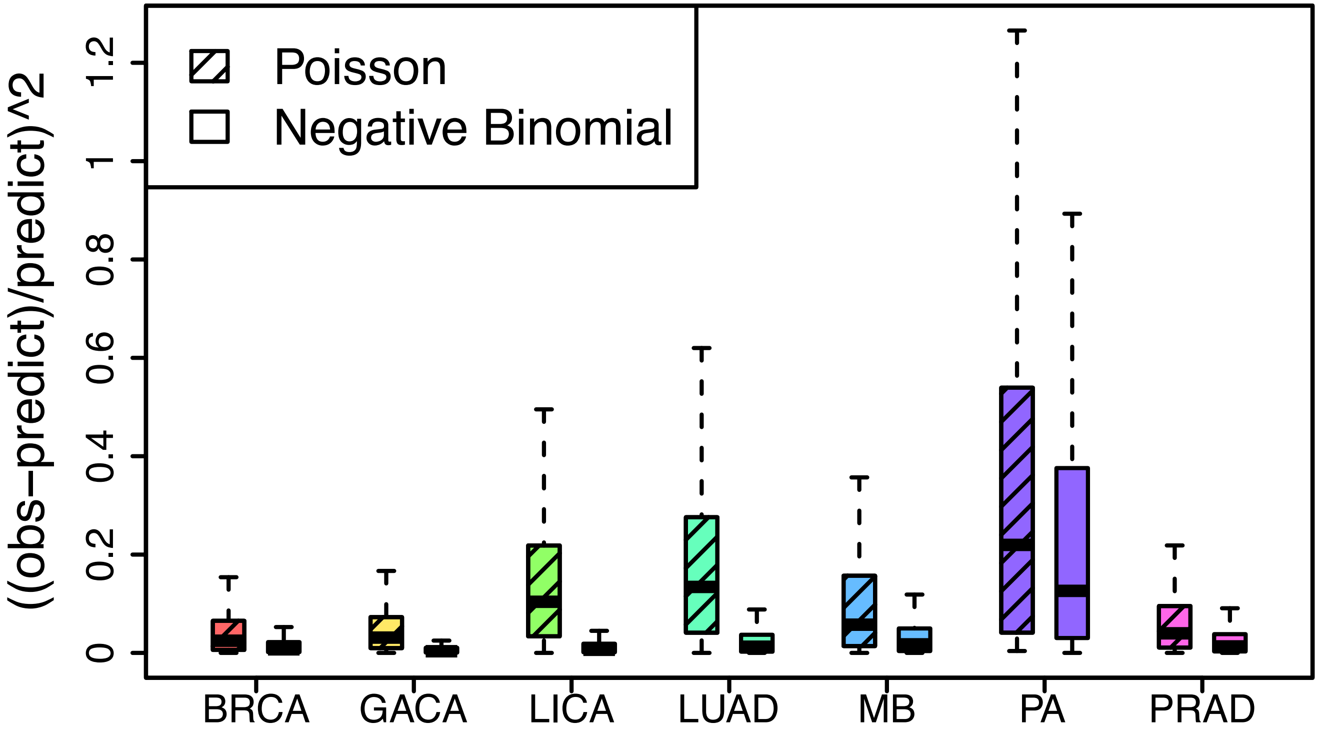
Final significance

- Features are sorted according to its correlation with BRCA mutation counts
- Correlation with different cancer types fluctuates, so regression need to be run separate

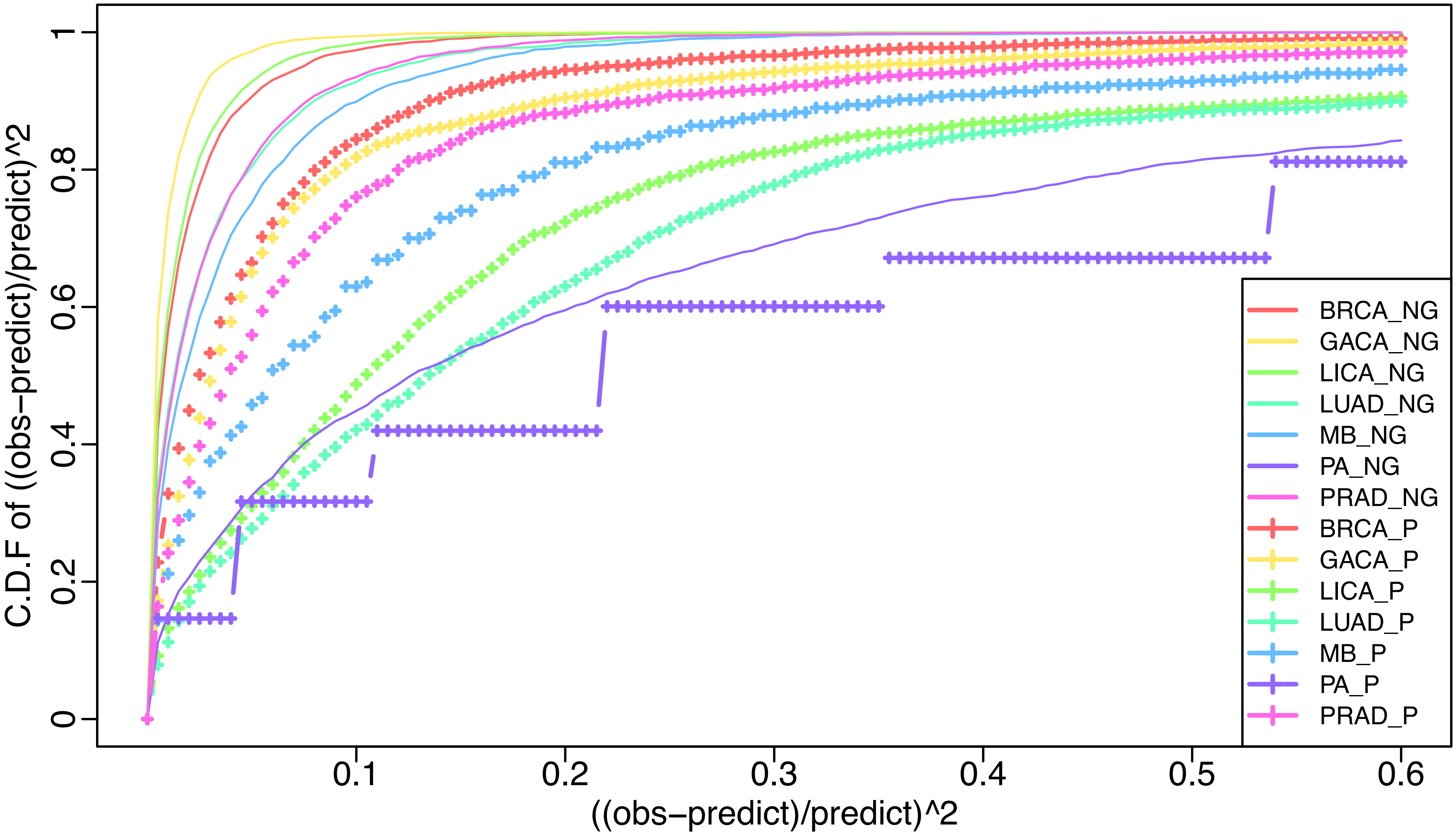




- Total variant counts affect the regression performance
- Matched data is good, but non match data still helps quite a lot due to the correlated nature of features

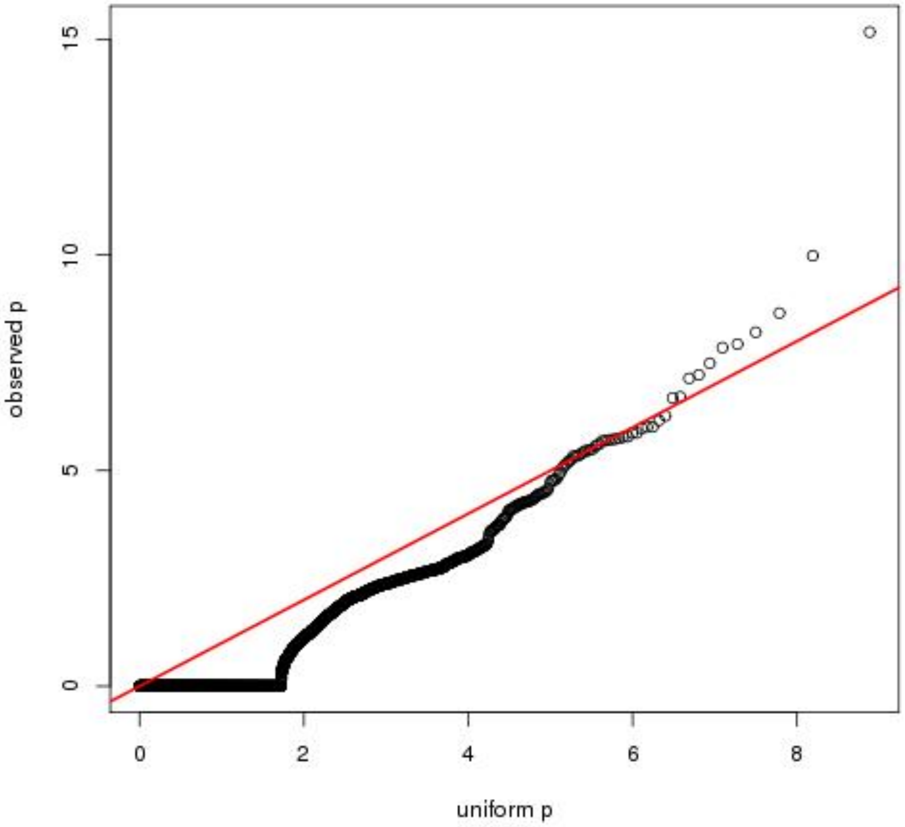


- Negative binomial regression can successfully reduce the variance in the mutation count data
- ((Obs-predict)/predict) ^2 value is much smaller after regression

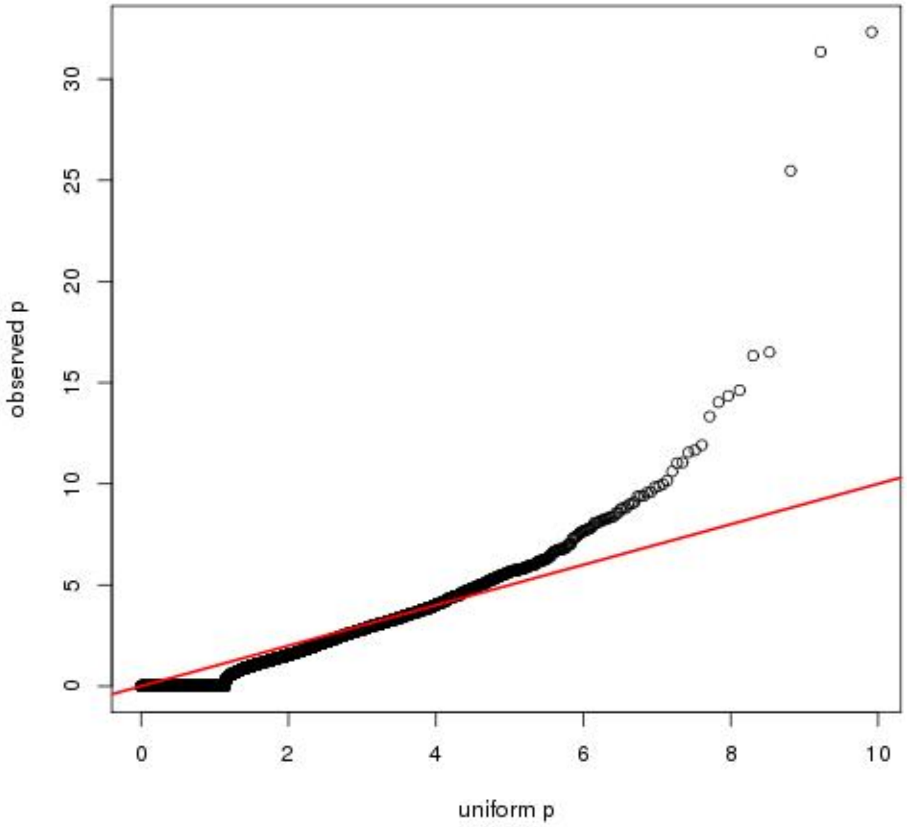

C.D.F of ((obs−predict)/predict)^2
0
0.2
0.4
0.6
0.8
1
0.1
0.2
0.3
0.4
0.5
0.6
((obs−predict)/predict)^2
BRCA_NG
GACA_NG
LICA_NG
LUAD_NG
MB_NG
PA_NG
PRAD_NG
BRCA_P
GACA_P
LICA_P
LUAD_P
MB_P
PA_P
PRAD_P

Cancer	Sigma	Variant	Correlation
BRCA	0.01417	601312	0.818
GACA	0.00847	3161741	0.928
LICA	0.01206	799681	0.958
LUAD	0.02866	1109556	0.917
MB	0.01583	112687	0.865
PA	0.01583	9487	0.668
PRAD	0.02781	396673	0.81

BRCA lincRNAs



BRCA promoters



Acknowledgement

- Jason Liu
- Lucas Lochovsky
- Jayanth Krishnan
- Donghoon Lee