SOMATIC MUTATION BURDEN ANALYSIS BY CORRECTING MULTIPLE COVARIATES

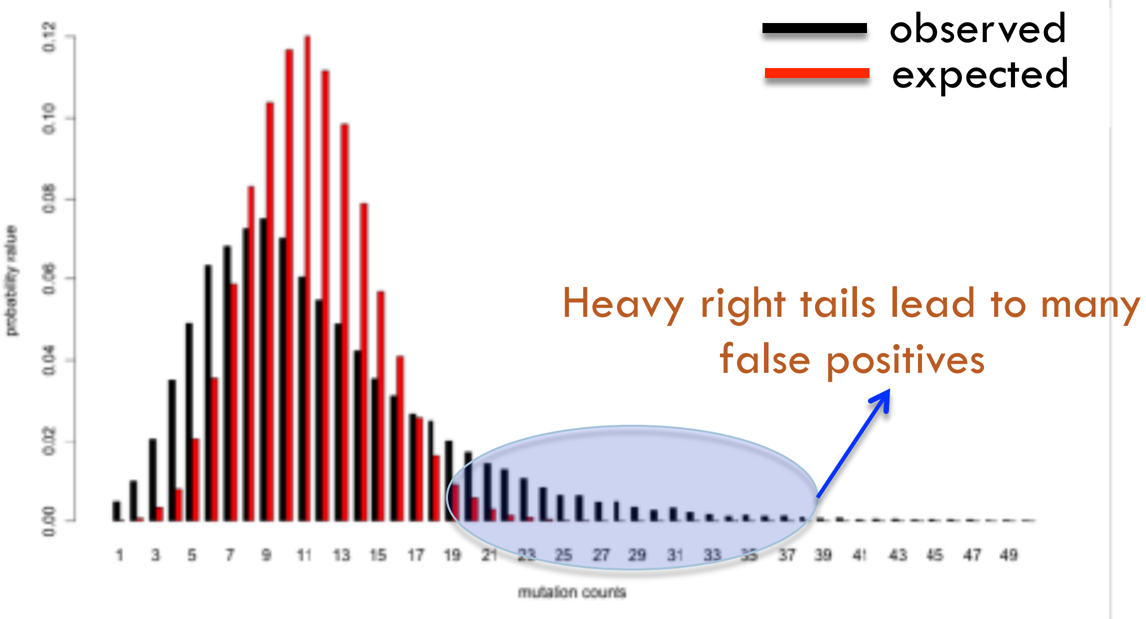
6/29/15

Jing Zhang
Gerstein Lab

Challenges in identifying noncoding drivers

- □ Noncoding variants may serve as drivers in many cancer types
  - TERT, PLEKHS1, WDR74 and SDHD promoters
  - miRNA-binding sites on BRCA1 and BRCA2
- □ Goal: identify highly mutated noncoding regions as driver candidates
- □ Challenge: mutation count data is usually over-dispersed

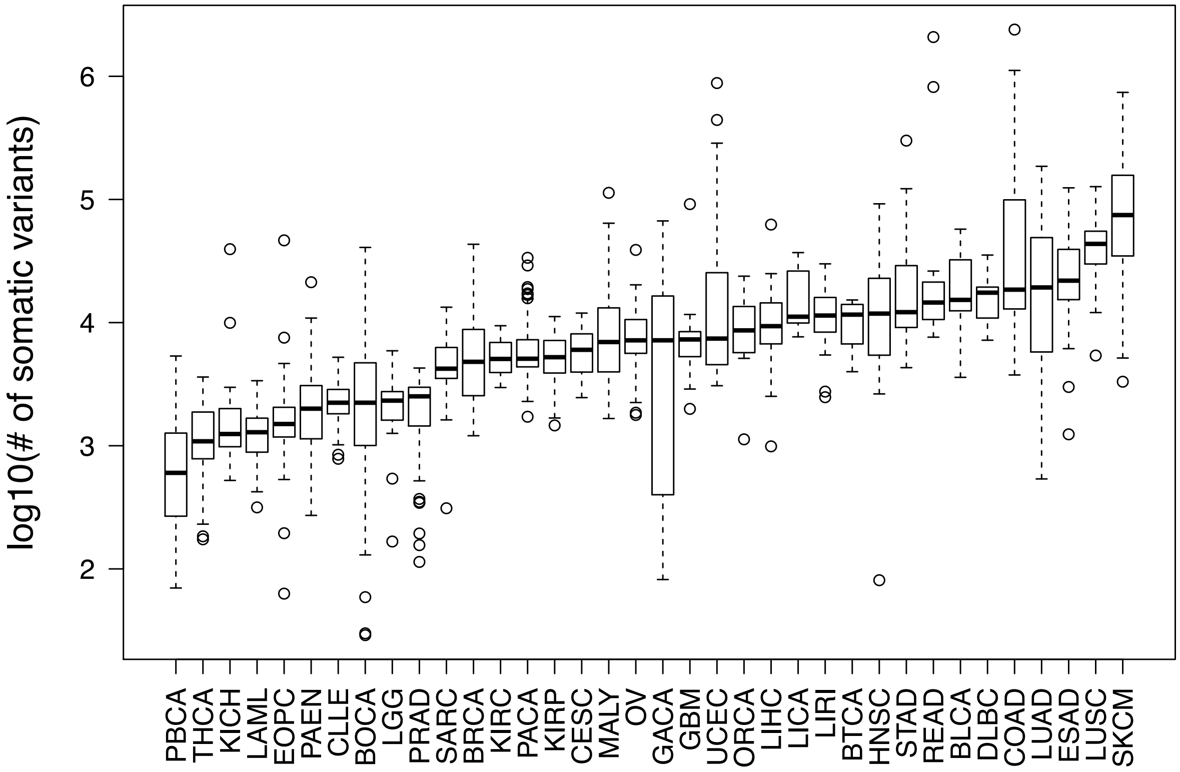
- ○ Reasons for overdispersion:
  - ▪ Mutation rate heterogeneity
  - ▪ Correlations among neighboring positions



Heavy right tails lead to many false positives

Sources of overdispersion

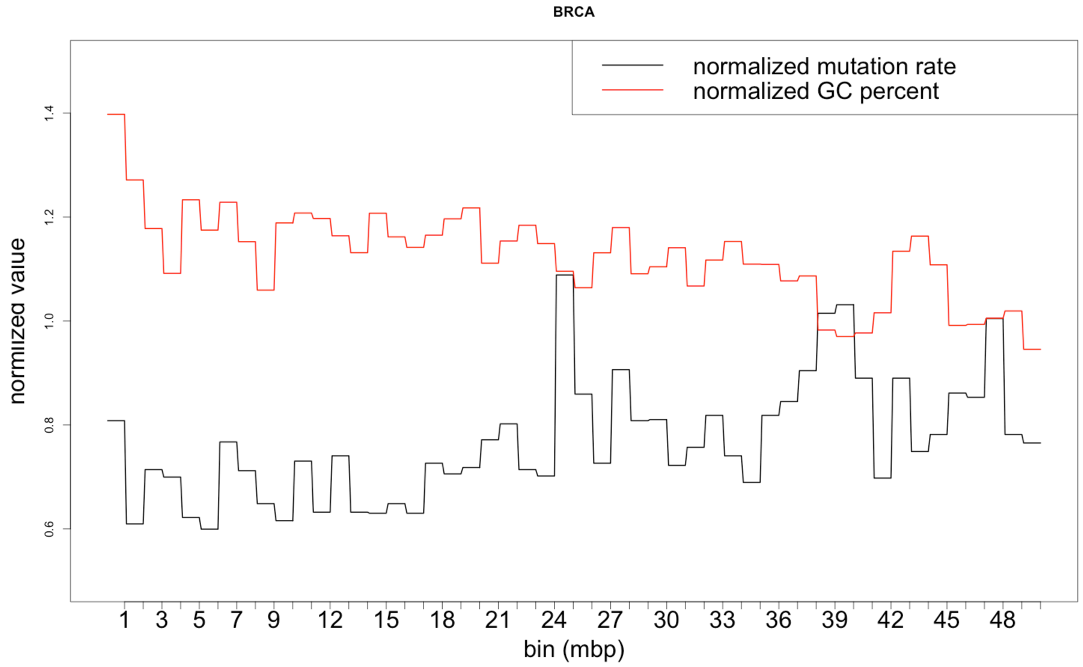
- Sources of mutation rate heterogeneity of :
  1. Mutation rate heterogeneity among different cancer types
  2. Mutation rate heterogeneity among different sample of the same cancer type



- ❖ SKCM: median number of mutations 74680
- ❖ PBCA: median number of mutations 602
- ❖ Max and min number of mutations of EOPC: 46540 and 63

Sources of overdispersion

- Sources of mutation rate heterogeneity of :
  3. Regional differences within the same sample



- Sources of mutation rate correlations:
  1. Correlations of SNVs due to existence of SV

Binomial and Beta-Binomial

- □ Binomial distribution: $\binom{n}{k} p^k (1-p)^{n-k}$

- ❑ Beta-binomial distribution:

$$x_i \big| p : Binomial\left(n_i, p\right)$$

$$p : Beta\left(\mu, \gamma\right)$$

  - ○ Assuming p is sampling from a beta distribution
  - ○ May be interpreted as sampling from different samples, regions, or cancer types(if there is)

$$\Pr\left\{Y = y \big| n, p, \gamma\right\} = \begin{pmatrix} n \\ y \end{pmatrix} \frac{\prod_{i=0}^{y-1}\left(p+\gamma i\right)\prod_{i=0}^{n-y-1}\left(1-p+\gamma i\right)}{\prod_{i=0}^{n-1}\left(1+\gamma i\right)}$$

Mean of the point mutation probability

Indicates the overdispersion of mutation counts

$$\log it\left(p_k\right) = \sum_{j=1}^{J} x_{kj} b_j, \gamma \sim \text{constant}$$

Poisson family

- □ Poisson distribution: $P\left(Y=y|p\right)=e^{-p}p^{y}/y!$
- □ Negative Binomial Distribution (type I):

$Y|\gamma \sim PO(\mu\gamma)$ and $\gamma \sim GA(1,\sigma^{\frac{1}{2}})$, $E(Y)=\mu$ and $Var(Y)=\mu+\sigma\mu^2$.

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

- □ Poisson inverse Gaussian Distribution:

$Y|\gamma \sim PO(\mu\gamma)$ and $\gamma \sim IG(1,\sigma^{\frac{1}{2}})$,

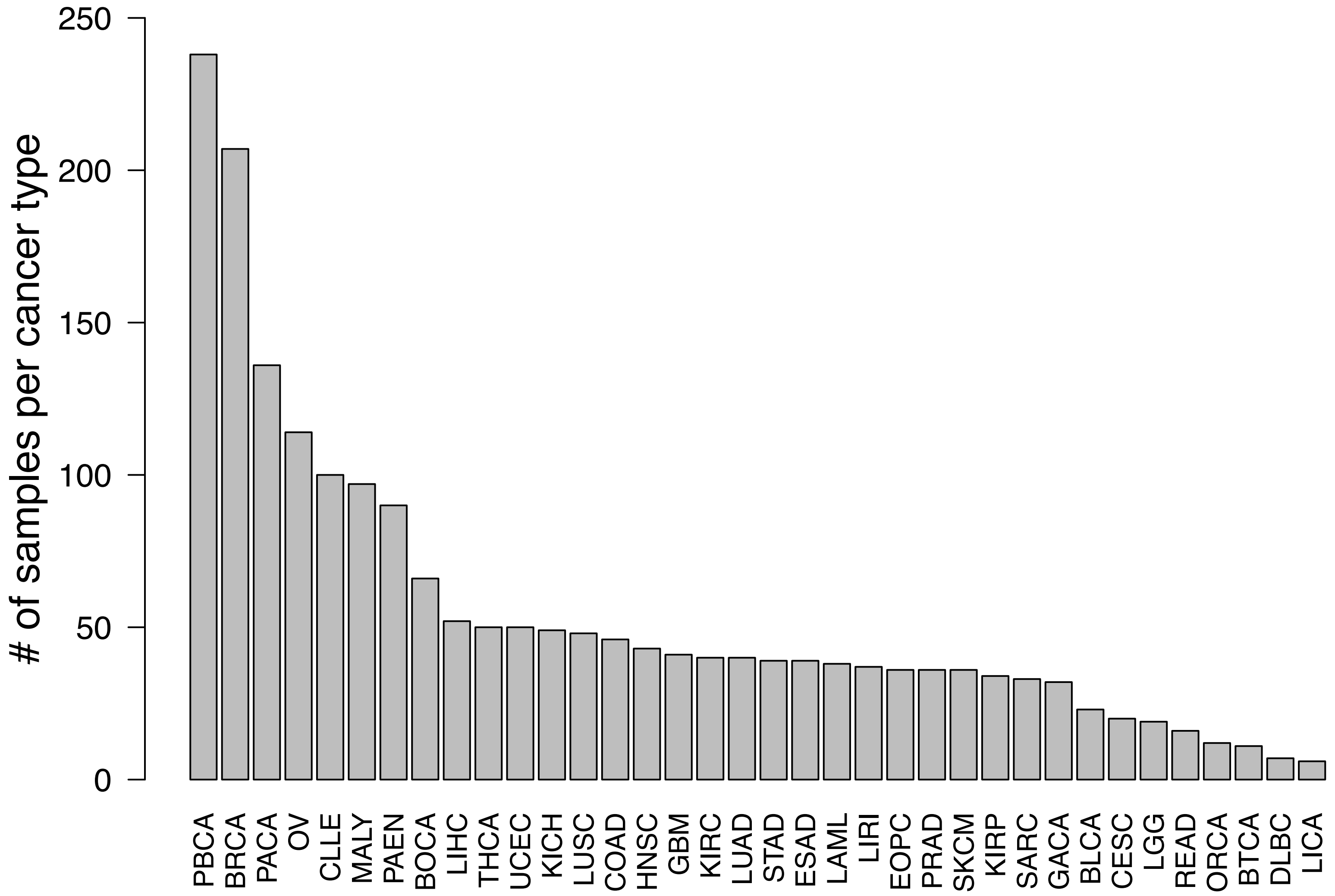
$$p_Y(y|\mu,\sigma)=\left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}}\frac{\mu^y e^{1/\sigma}K_{y-\frac{1}{2}}(\alpha)}{(\alpha\sigma)^y y!}$$

Computational Goal

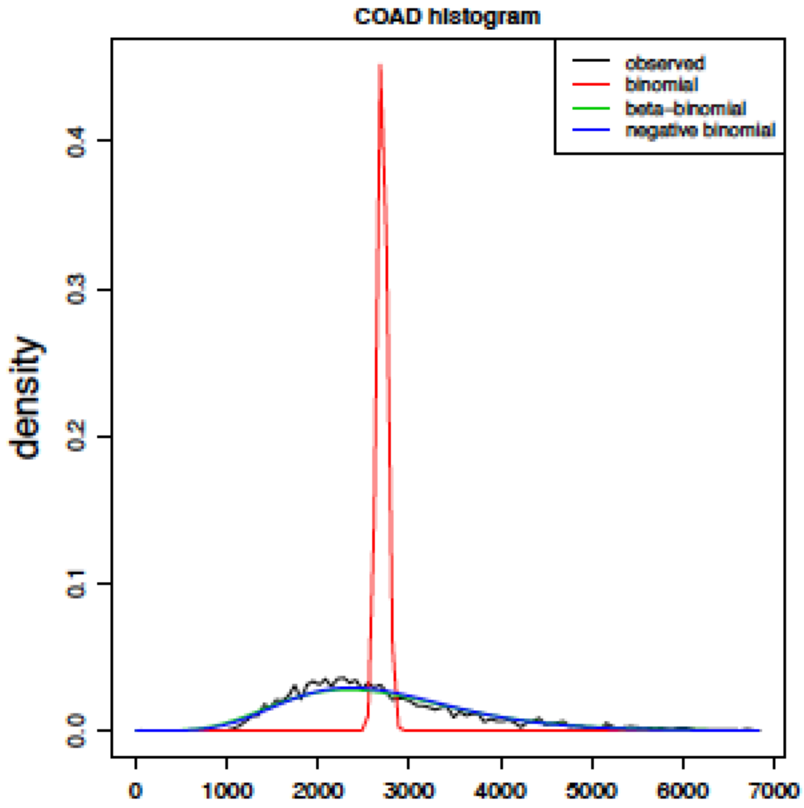
- □ Reasonable local ***noncoding*** mutation rate prediction
  - ○ Previous model: replication timing + GC content
  - ○ Current model: list of correlated genomic features
    - ✧ GC content, CpG content, Replication timing
    - ✧ Chromatin Accessibility, Histone modification marks
    - ✧ Expression level

$y_1, \cdots, y_k, \cdots y_K$	Mutation counts
$\mathbf{X}$	Covariant matrix
$n_1, \cdots, n_k, \cdots n_K$	Length of FNC elements

Summary of data used

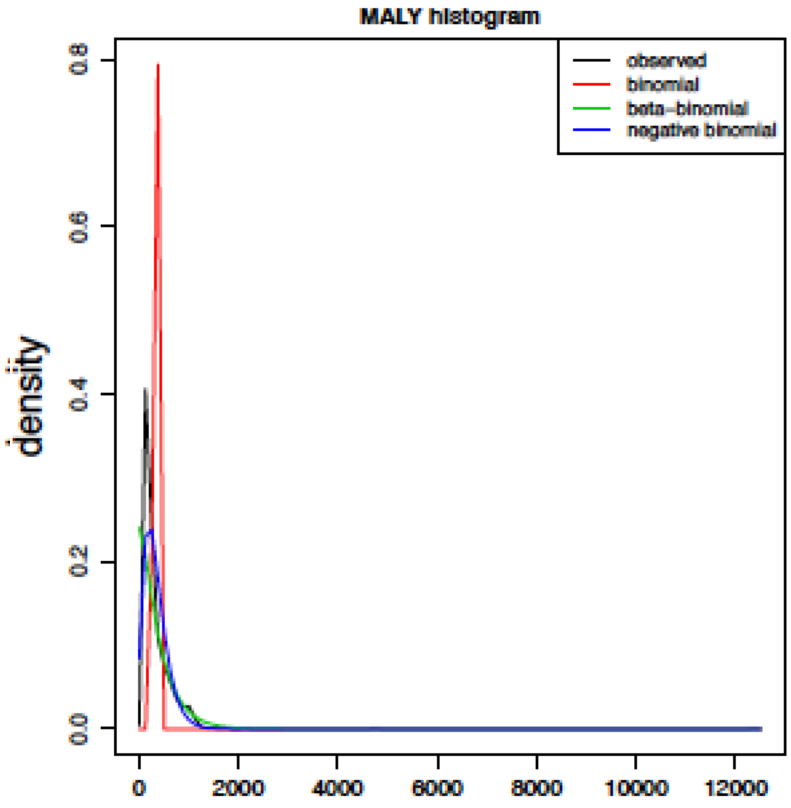


Distribution fitting comparison



Number of mutations in 1mb bins

Colon Adenocarcinoma



Number of mutations in 1mb bins

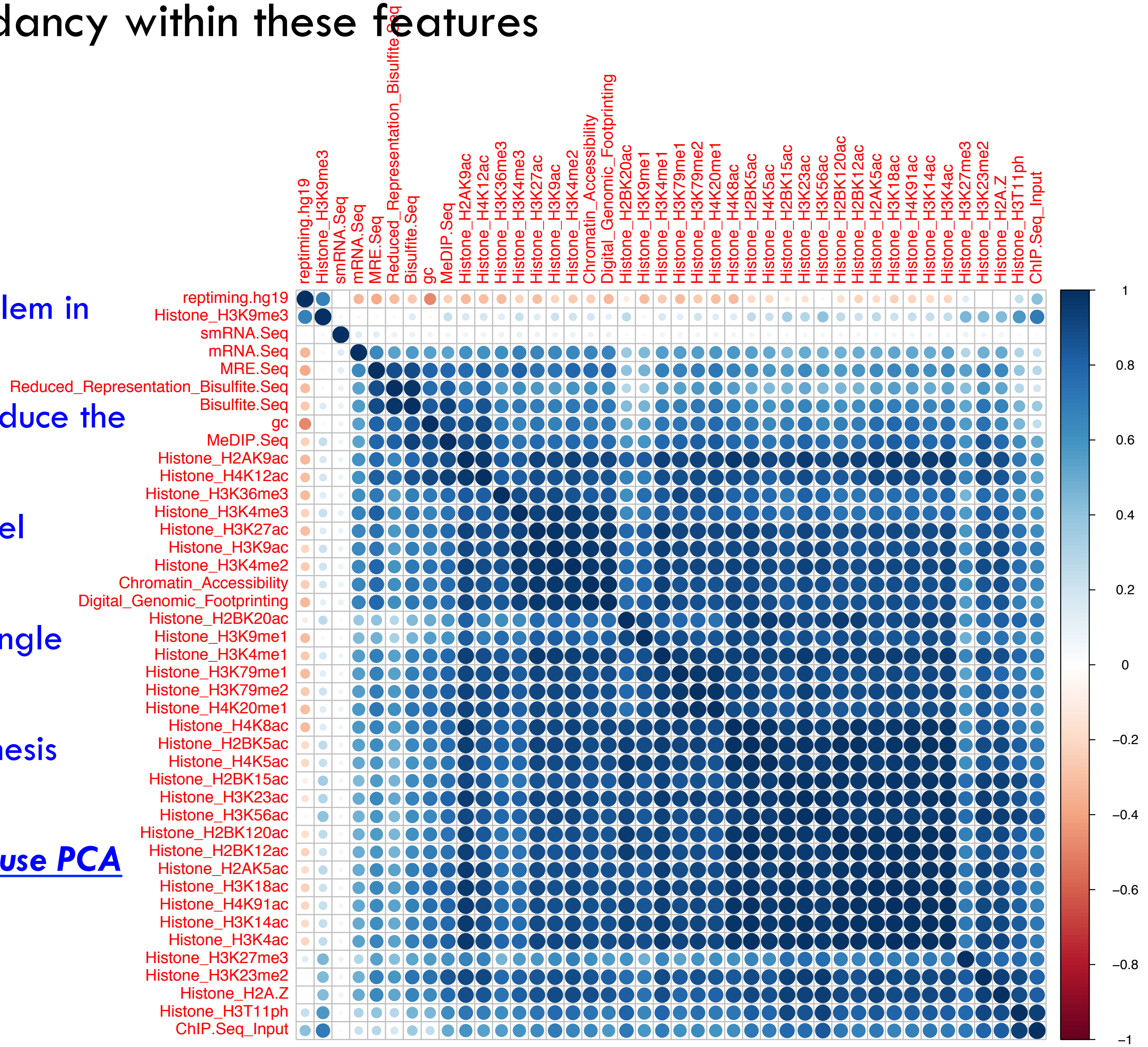
Malignant Lymphoma

Choice of feature list

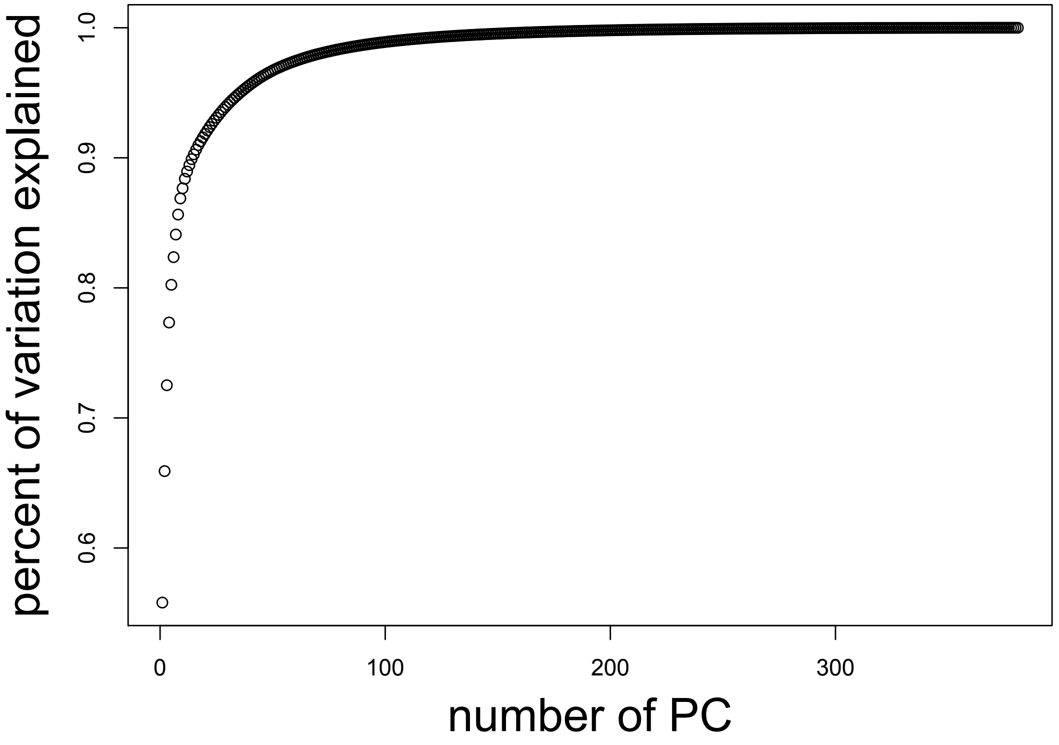
- **Full list:** 381 features from the Epigenetics Roadmap and ENCODE projects
  - 7 modification marks: Histone_H3K27ac, Histone_H3K27me3, Histone_H3K36me3, Histone_H3K4me1, Histone_H3K4me3, Histone_H3K9ac, Histone_H3K9me3
  - Expression data from mRNA-seq, Chromatin accessibility
  - GC content, CpG percentage, replication timing
- **Average list:** 42 features
  - 40 features averaged across Epigenetics Roadmap project,
  - GC content + replication timing

Problem: redundancy within these features

- Multicollinearity problem in coefficient estimation
- Will not affect the reduce the predictive power or reliability of the model
- Will only impact the understanding of a single predictor and its corresponding hypothesis testing
- ***<u>Solution: go with it or use PCA based regression</u>***



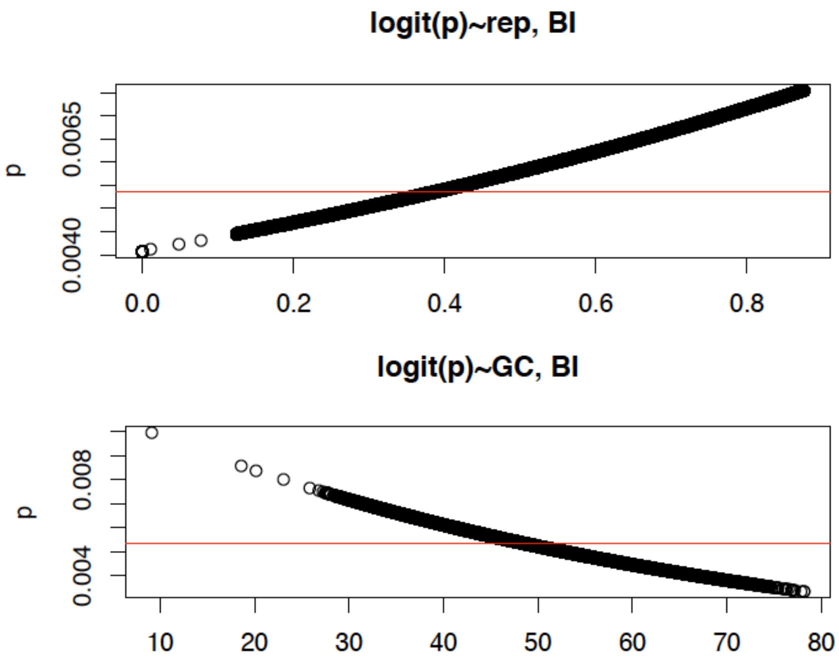
PCA analysis of covariates

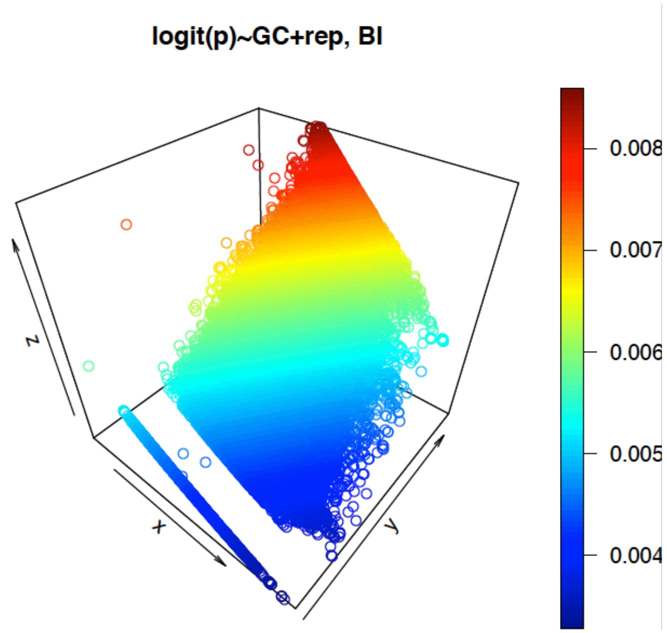


- Use PCA to project the features into orthogonal space and run regression on these independent components
- PCA based regression might be very sensitive for number of PCs selected
- To keep approximately the same performance, need at least 105 PCs that explains > 0.99 percent of variation

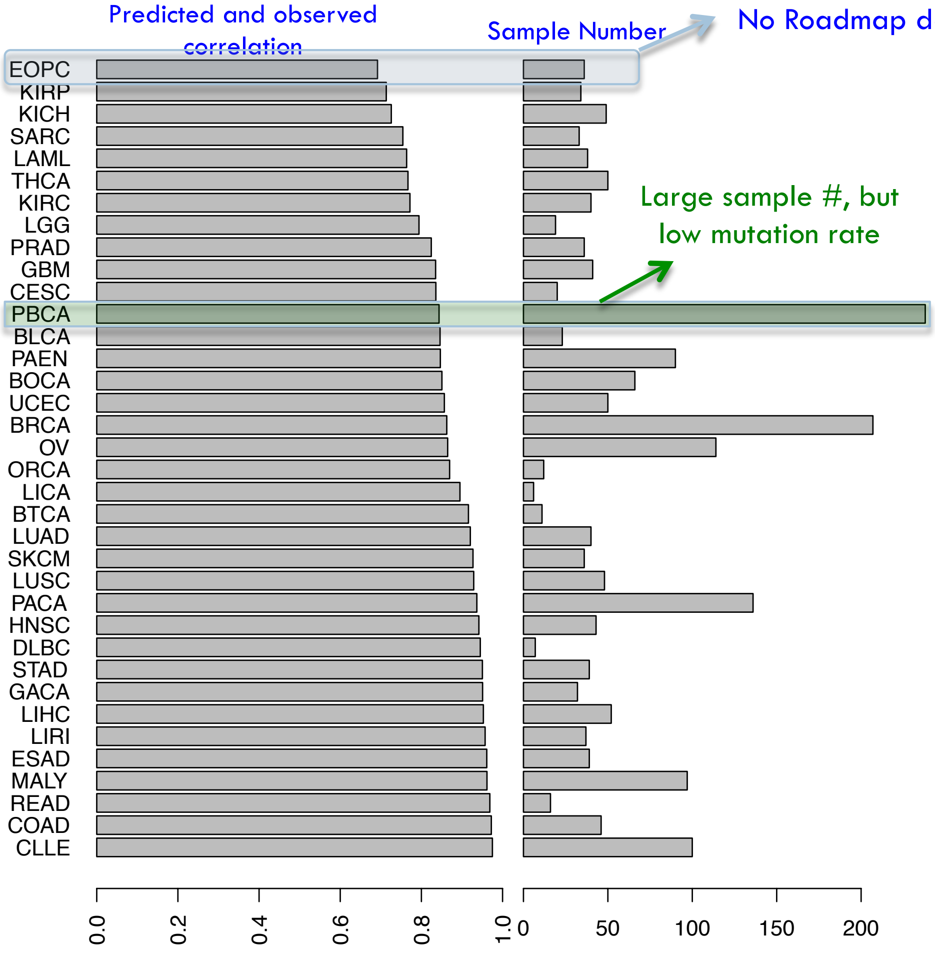
Virtualization Example

- Correlation of mutation rate and GC
  - -0.246 (Pearson) and -0.259 (spearman)
- Correlation of mutation rate and replication timing
  - 0.314 (Pearson) and 0.276 (spearman)





Binomial Family performance

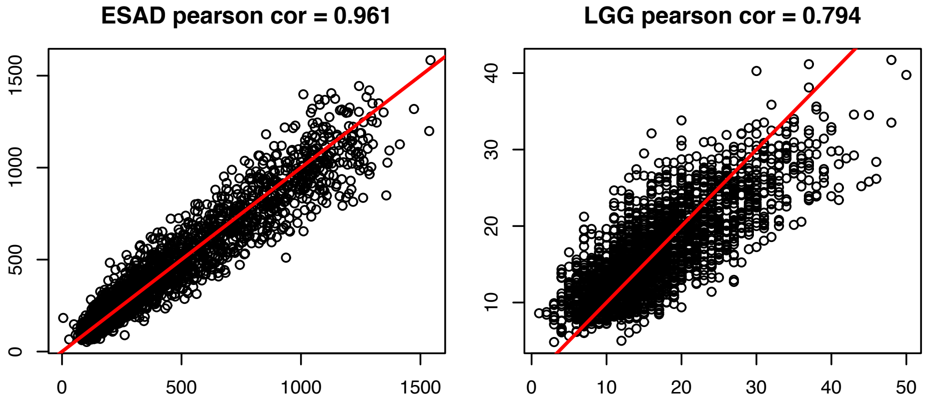


Predicted and observed correlation

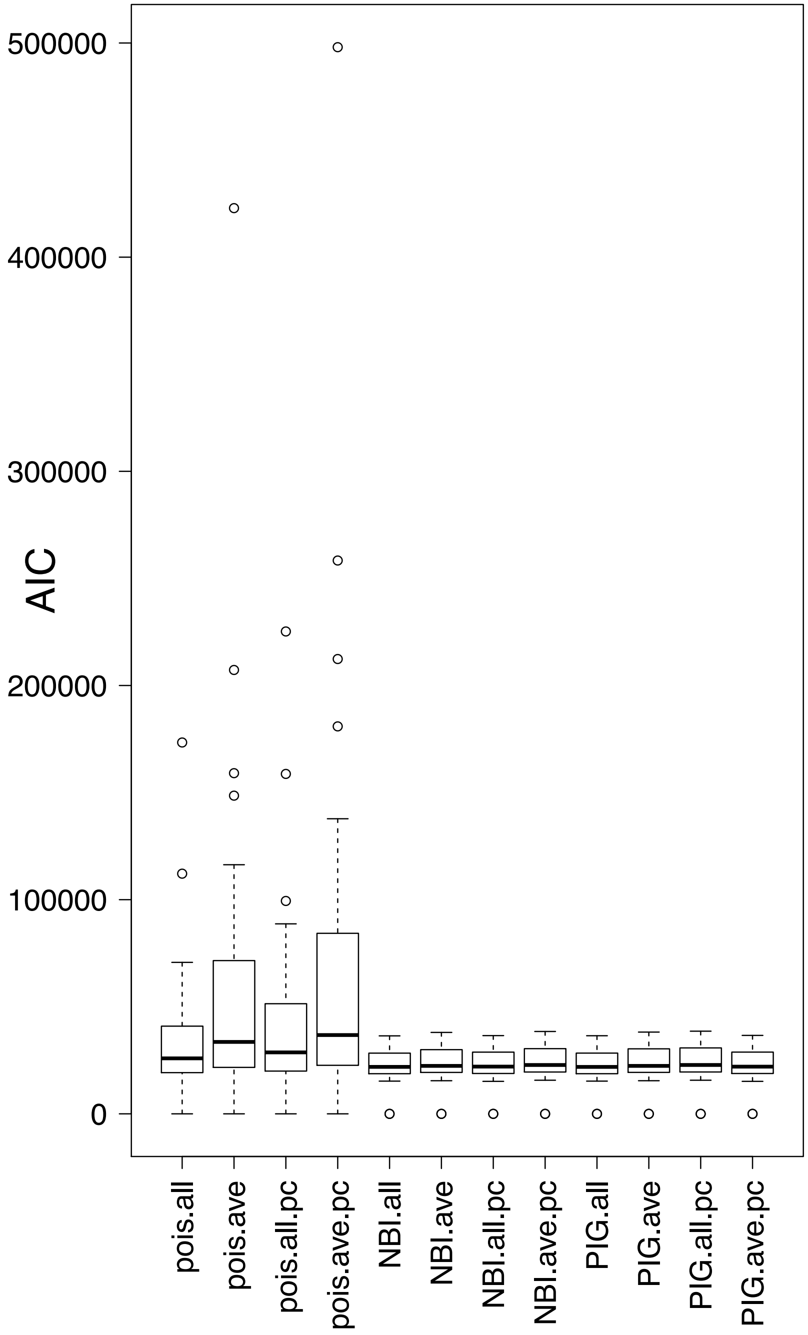
Sample Number

No Roadmap data

Large sample #, but low mutation rate



- These 381 features accurately predict the mutation rate in various cancer types
- Pearson correlation of the observed and predicted variant counts varies from 0.692 to 0.975
- Performance is not dominated by sample size effect



Performance comparison of the Poisson family

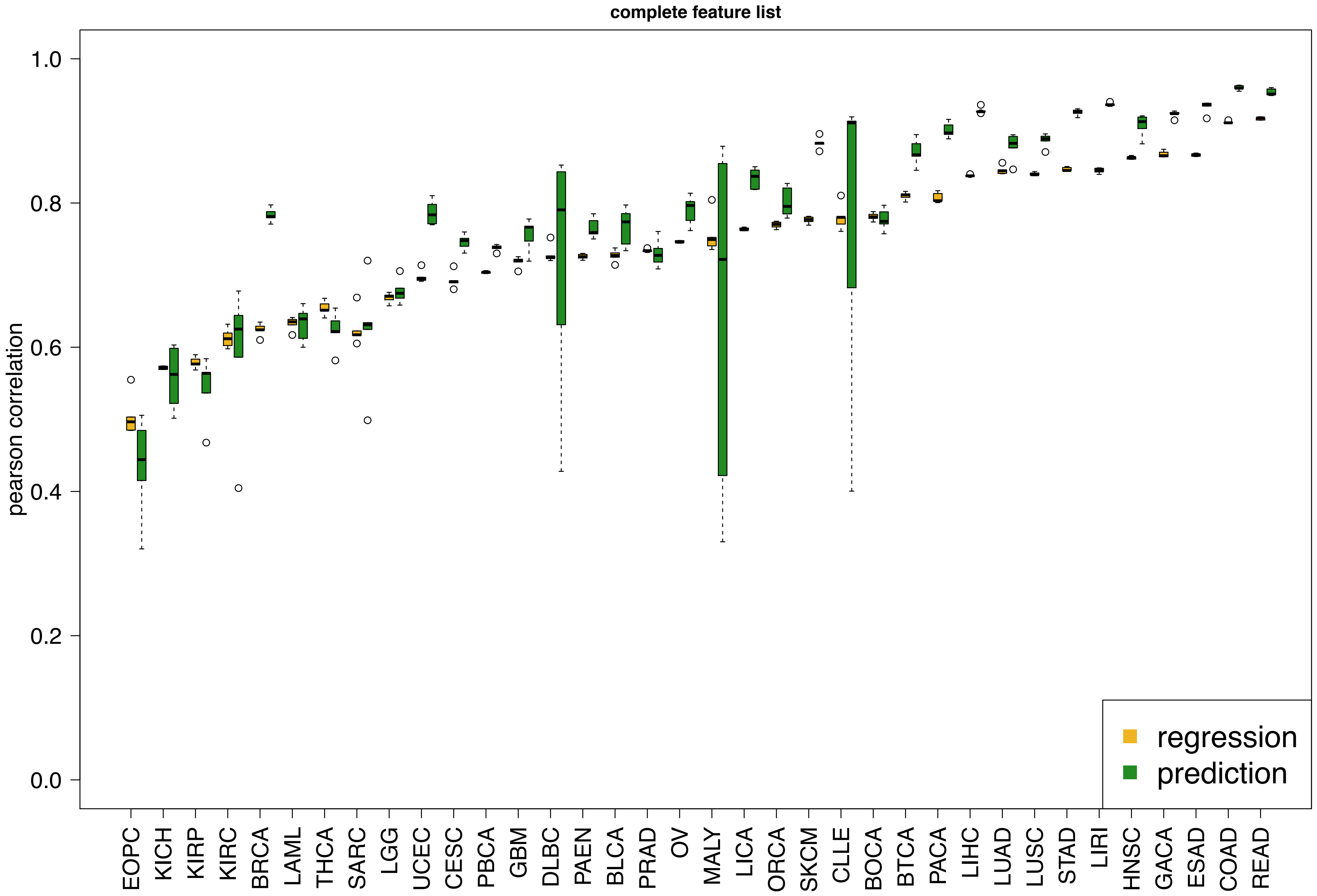
$$AIC = 2K - 2\ln(L)$$

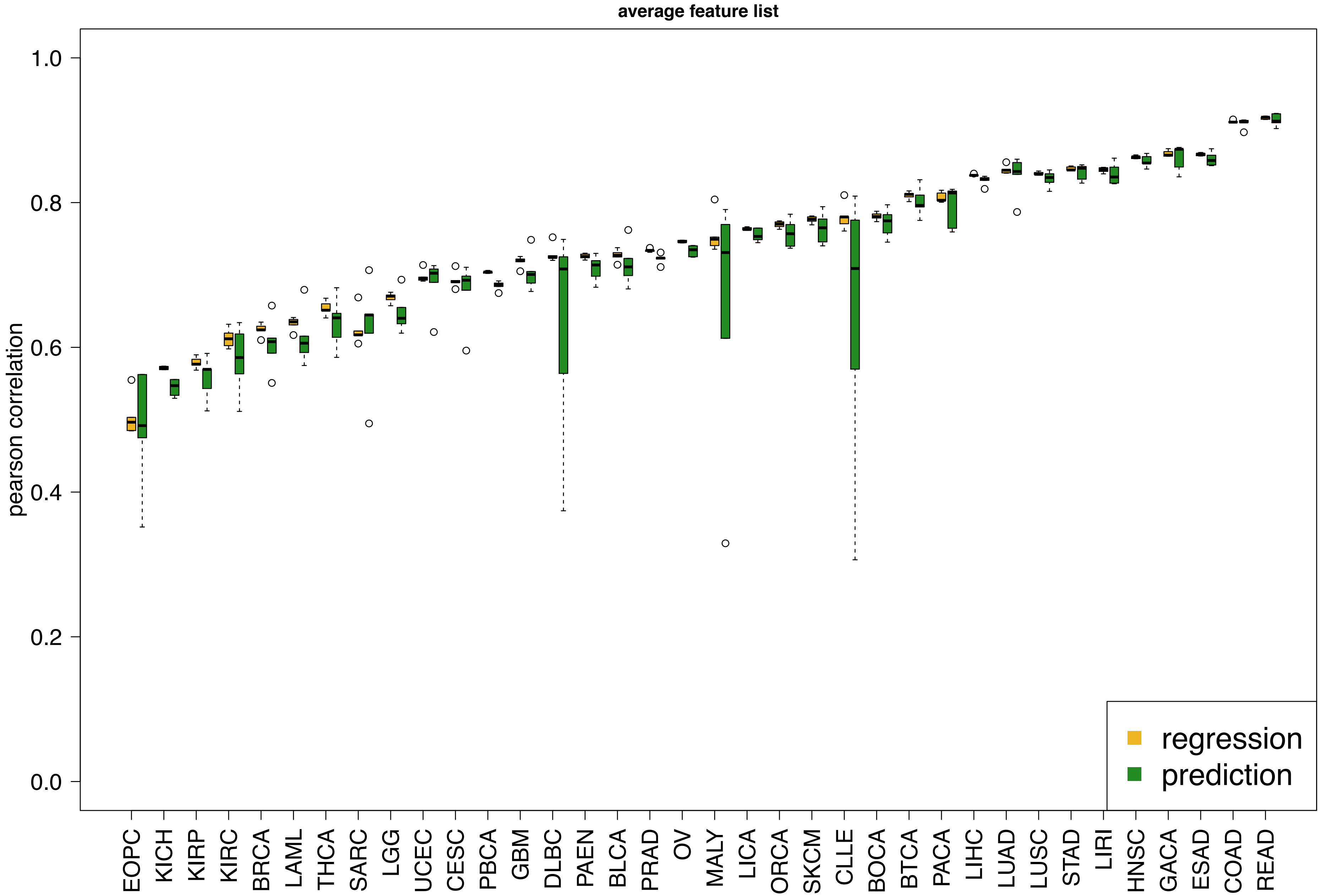
↓

Number of parameters in the model

- ✧ Poisson > NBI ≈ PIG
- ✧ Even with PCs explaining 99% of variation, performance is still can not be comparable to all feature list
- ✧ PC1 is not the most significance predictor of mutation counts, meaning the factor that explaining most covariates variation is NOT the one explaining the counts

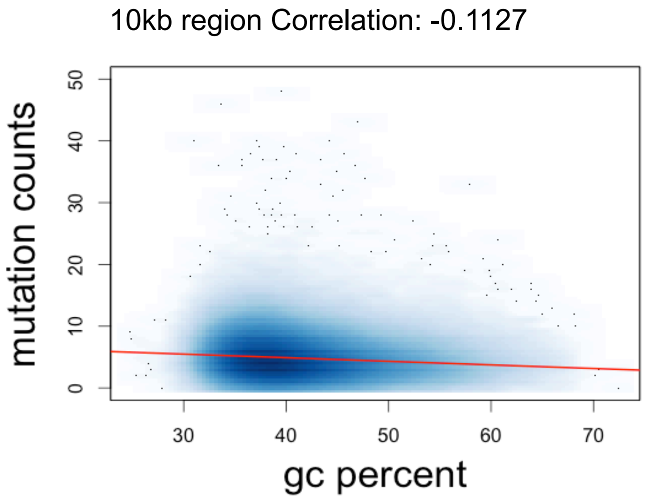
6/29/15

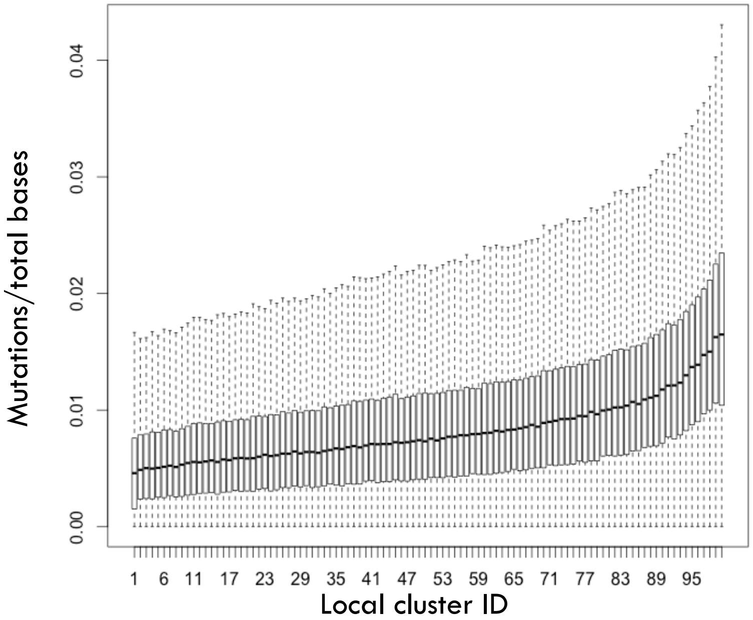

complete feature list
pearson correlation
1.0
0.8
0.6
0.4
0.2
0.0
regression
prediction
EOPC
KICH
KIRP
KIRC
BRCA
LAML
THCA
SARC
LGG
UCEC
CESC
PBCA
GBM
DLBC
PAEN
BLCA
PRAD
OV
MALY
LICA
ORCA
SKCM
CLLE
BOCA
BTCA
PACA
LIHC
LUAD
LUSC
STAD
LIRI
HNSC
GACA
ESAD
COAD
READ


average feature list
pearson correlation
0.0
0.2
0.4
0.6
0.8
1.0
regression
prediction
EOPC
KICH
KIRP
KIRC
BRCA
LAML
THCA
SARC
LGG
UCEC
CESC
PBCA
GBM
DLBC
PAEN
BLCA
PRAD
OV
MALY
LICA
ORCA
SKCM
CLLE
BOCA
BTCA
PACA
LIHC
LUAD
LUSC
STAD
LIRI
HNSC
GACA
ESAD
COAD
READ

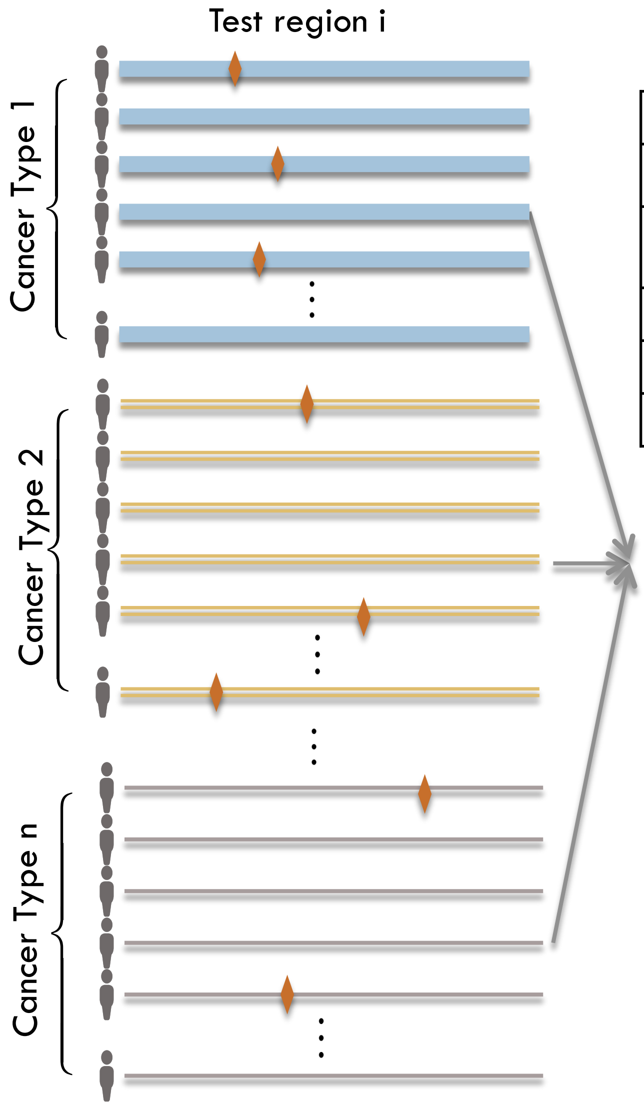
Lack of statistical power in smaller regions

- □ Challenge: target regions are usually not large enough for accurate background mutation rate estimation
- □ Nearest neighbor: in high dimensional space, difficult to find a neighbor
- □ Solution: test region clustering based on predictions





Flowchart of our mutational analysis



Calculate covariant matrix

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

PCA based regression for neighbor definitions

$$\hat{\mu}_{1i}: \text{ estimate of } \left\{k\,\middle|\,\left|g^{-1}\left(\mathbf{X}_i\hat{\mathbf{b}}_1\right)-g^{-1}\left(\mathbf{X}_k\hat{\mathbf{b}}_1\right)\right|<\xi\right\}\Rightarrow p_{i,1}$$

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

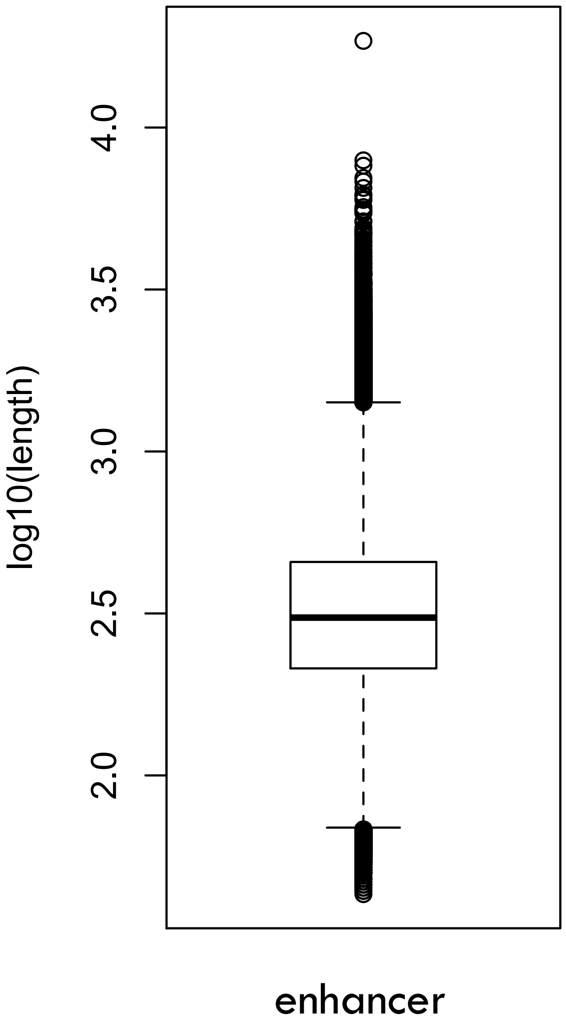
$$\vdots$$

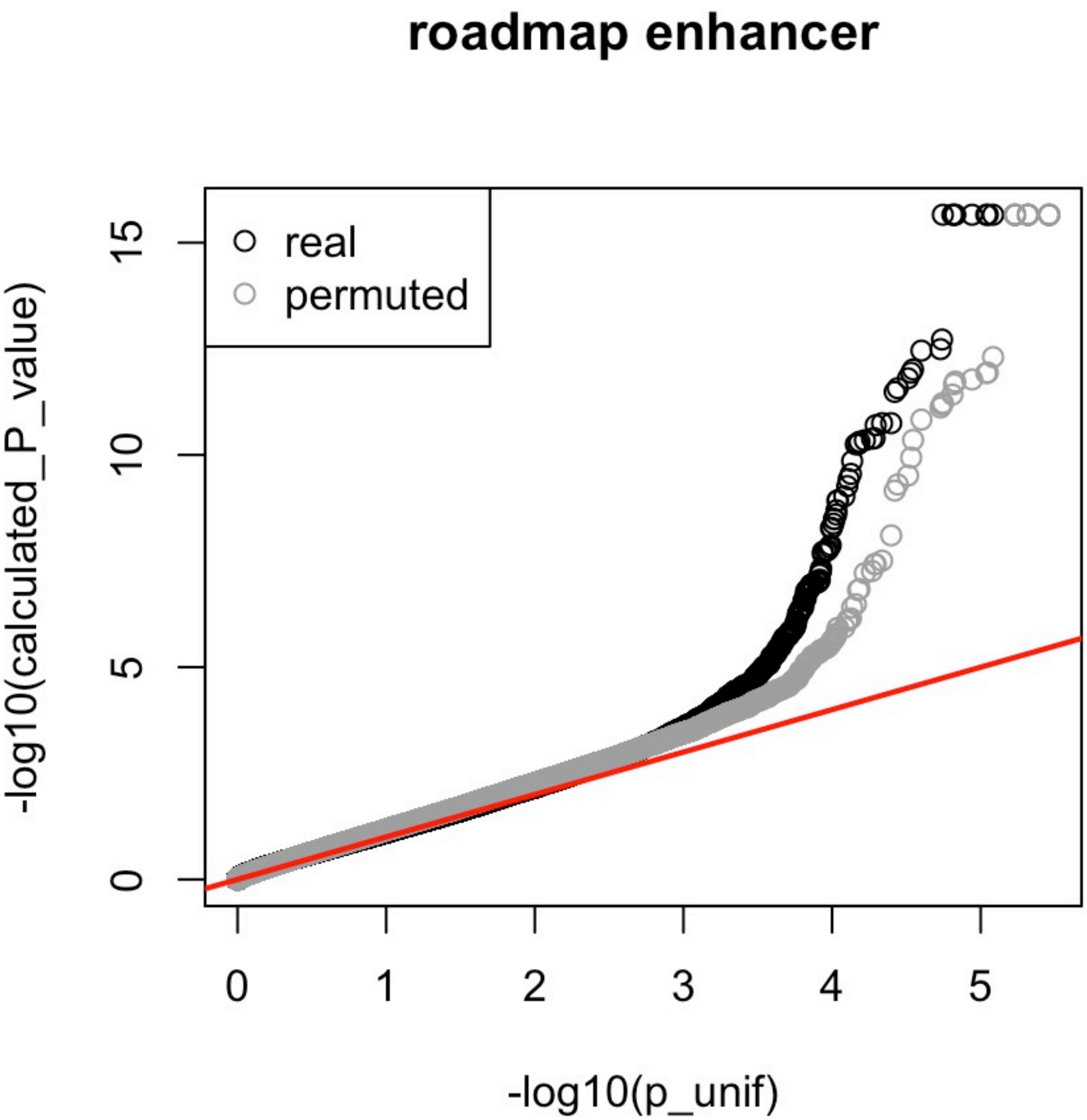
$$\hat{\mu}_{ki}: \text{ estimate of } \left\{k\,\middle|\,\left|g^{-1}\left(\mathbf{X}_i\hat{\mathbf{b}}_n\right)-g^{-1}\left(\mathbf{X}_k\hat{\mathbf{b}}_n\right)\right|<\xi\right\}\Rightarrow p_{i,n}$$

$$\longrightarrow \hat{p}_i$$

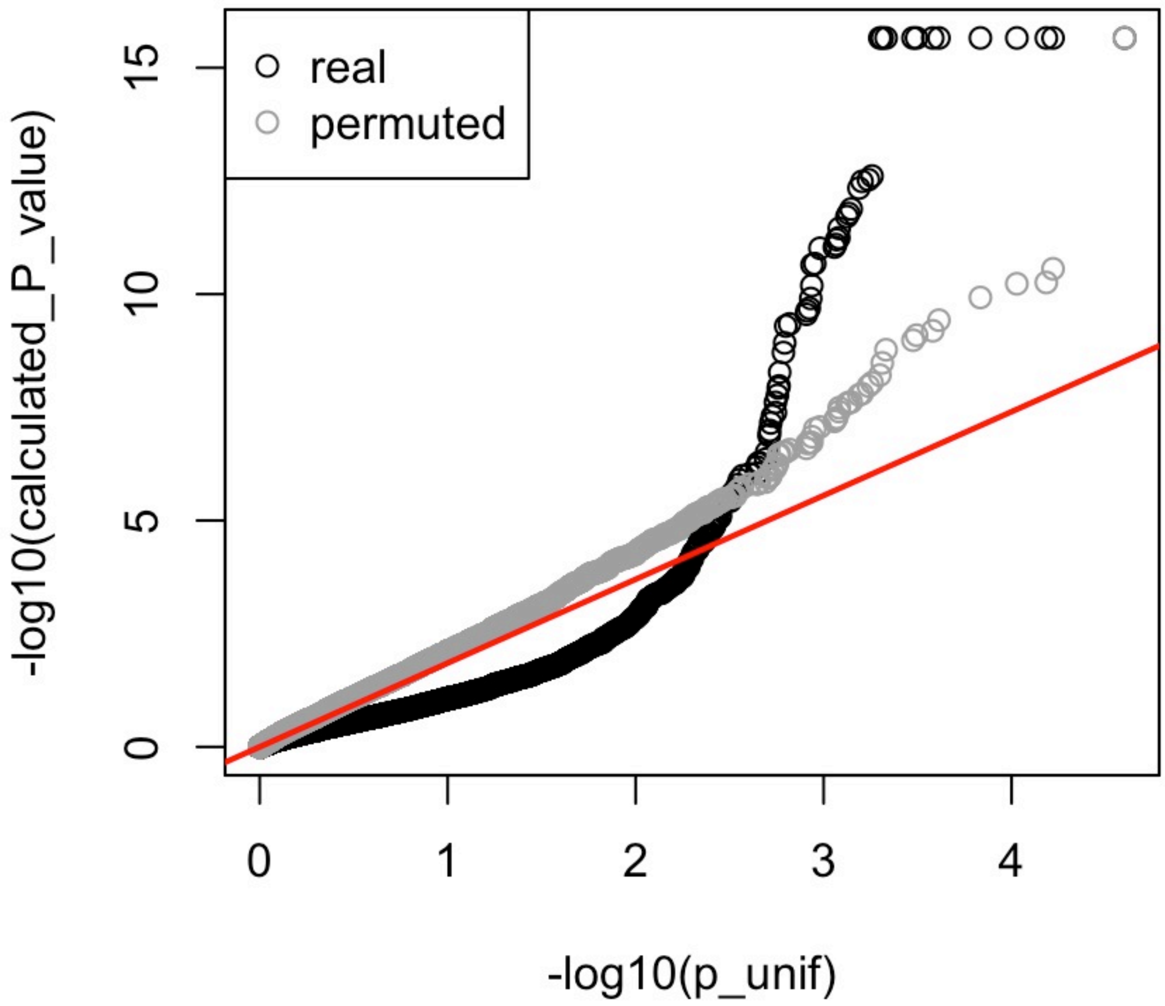
Final significance

Pan Cancer Analysis



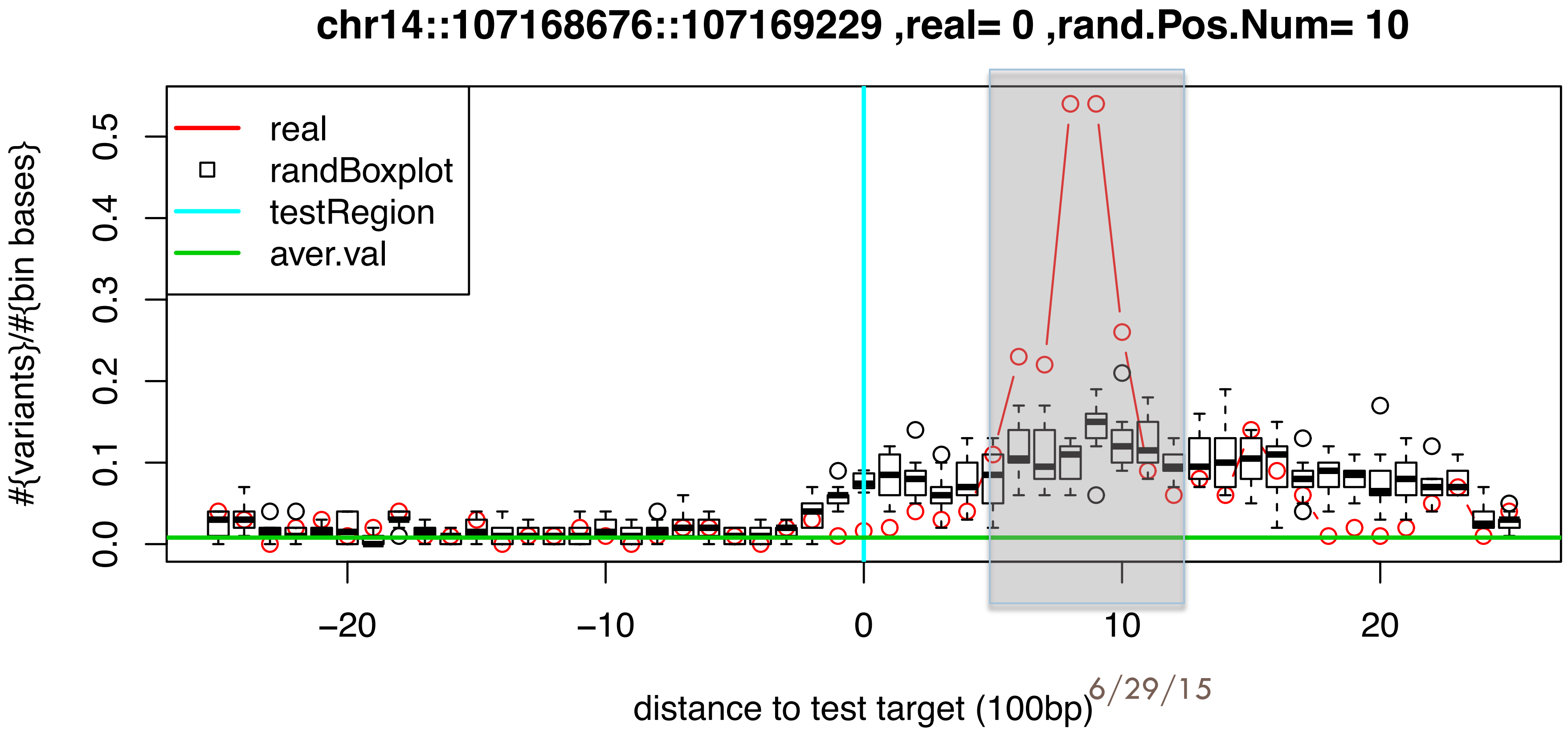


Protein coding genes



On going work

- Compare the P values from Poisson and Binomial family
- Efficient implementation of current method
- Annotation free analysis



Acknowledgement

- Mark Gerstein
  - Lucas Lochovsky
  - Jason Liu
- Ekta Khurana
  - Priyanka Dhingra

Questions: J.zhang@yale.edu