PERFORMANCE COMPARISON USING GENOMIC FEATURES FROM DIFFERENT PROJECTS

Jing Zhang, Gerstein Lab, Yale

02/22/2016

Mixing samples of the same cancer type

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

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

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

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

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

Accurate local mutation rate estimation

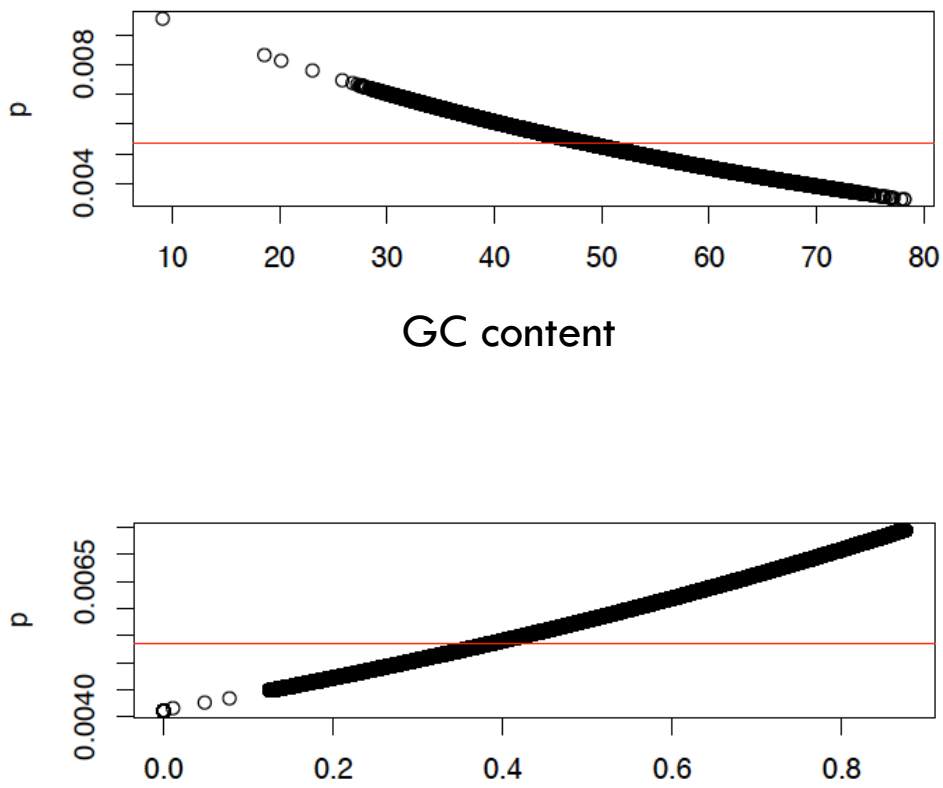
$$g_1(\mu_i) = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,k}\beta_k + \cdots x_{i,K}\beta_K$$

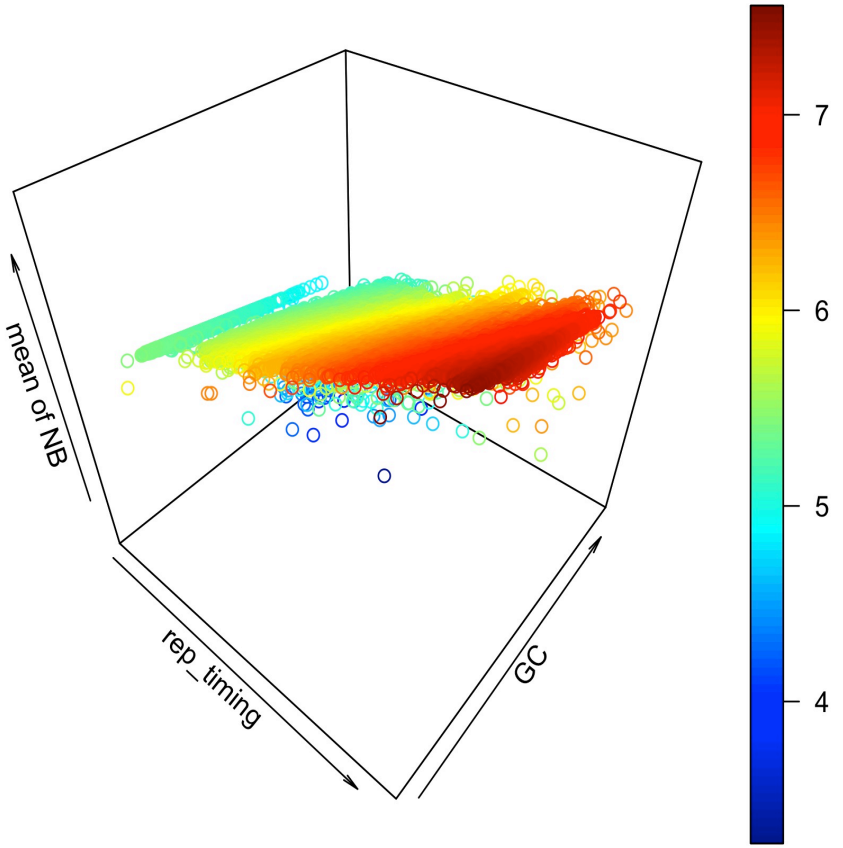
By pooling the variants together we are assuming the same covariate coefficients

Covariate list:

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

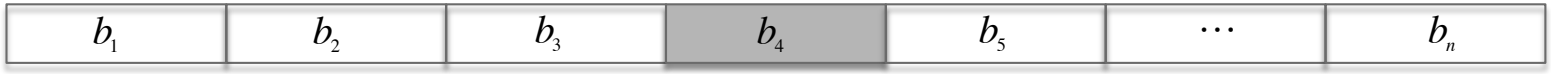
Toy Example of how the fitting works





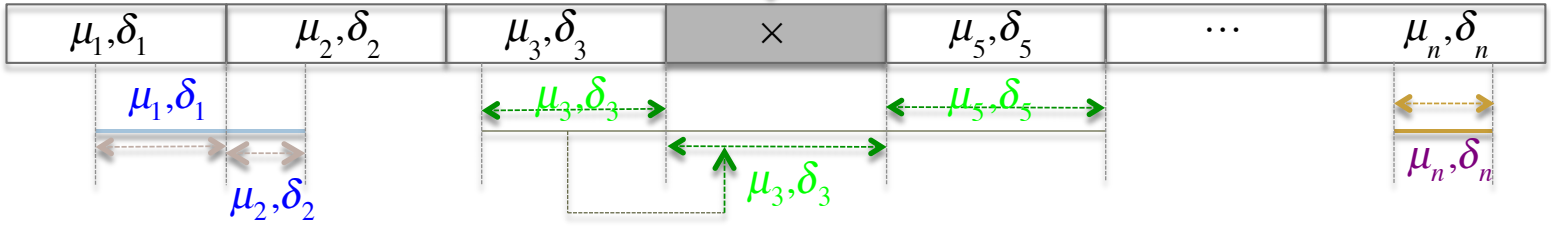
How to find the local mutation rate

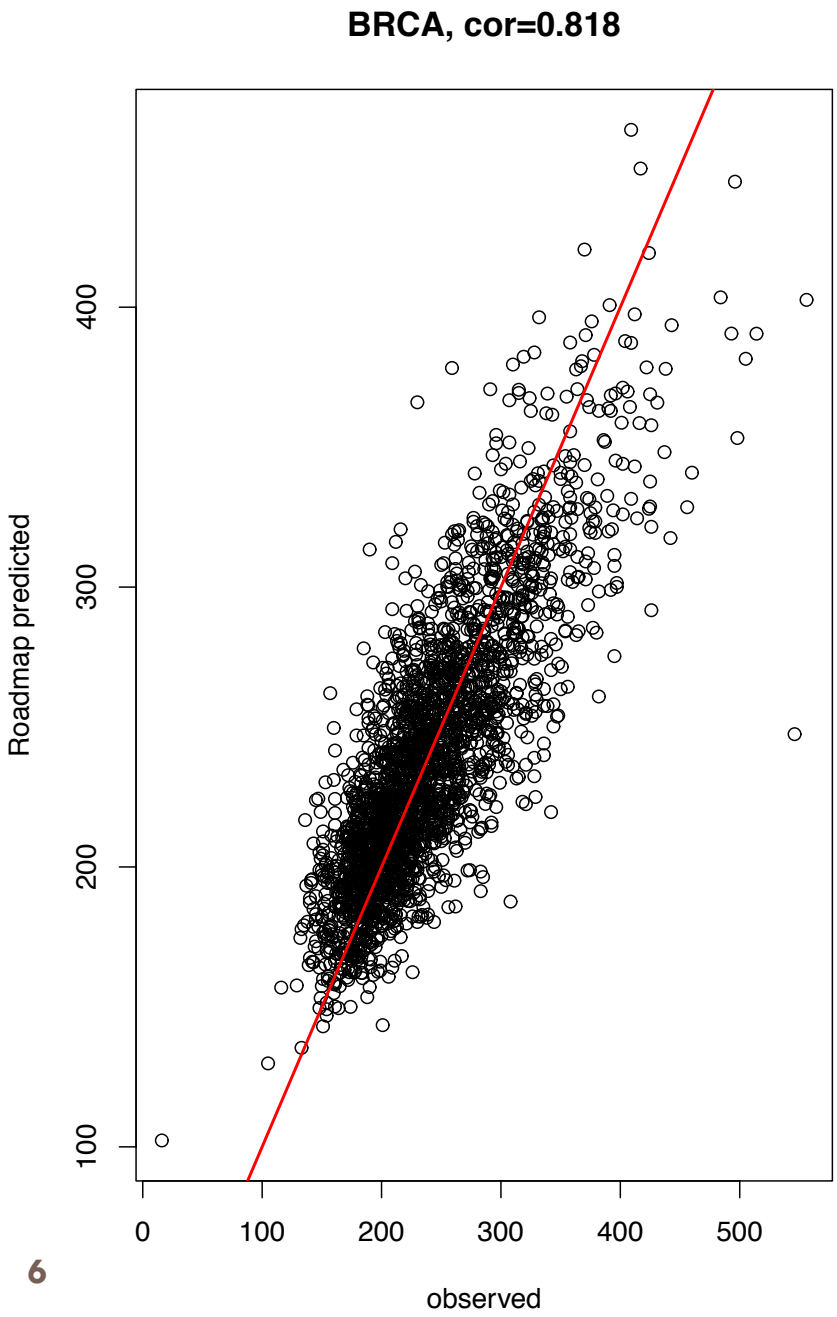
Step 1. Chopping the chromosome into bins, remove the gap regions

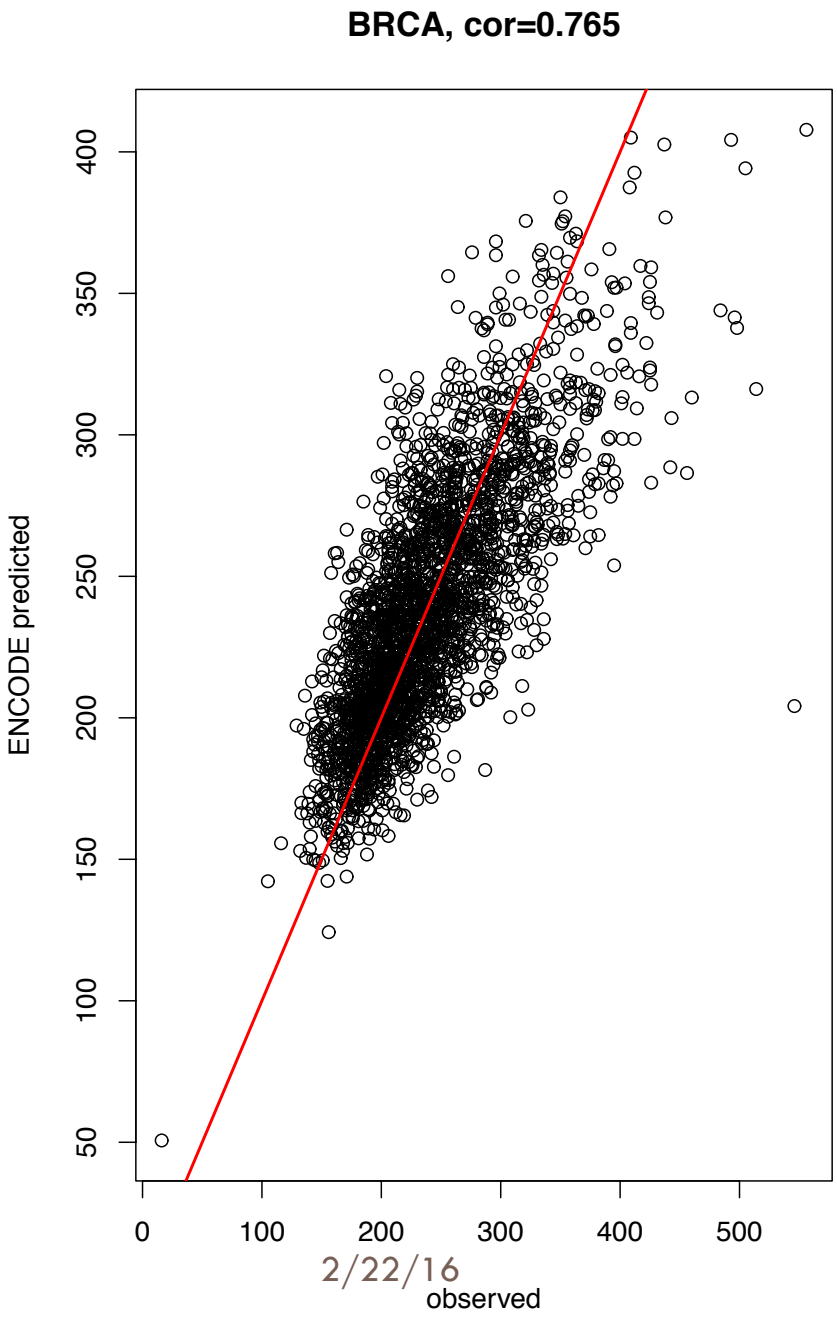


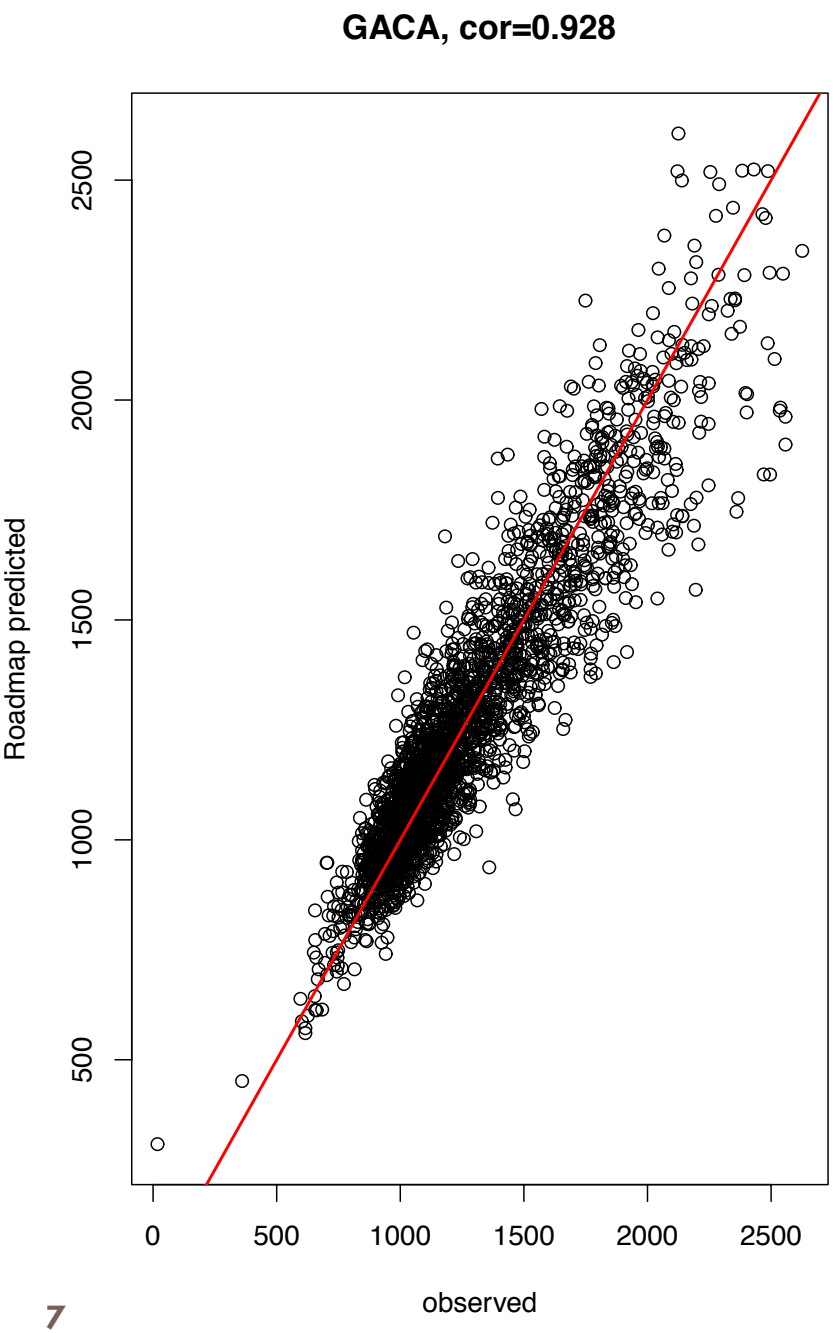
Step 2. Parameter estimation

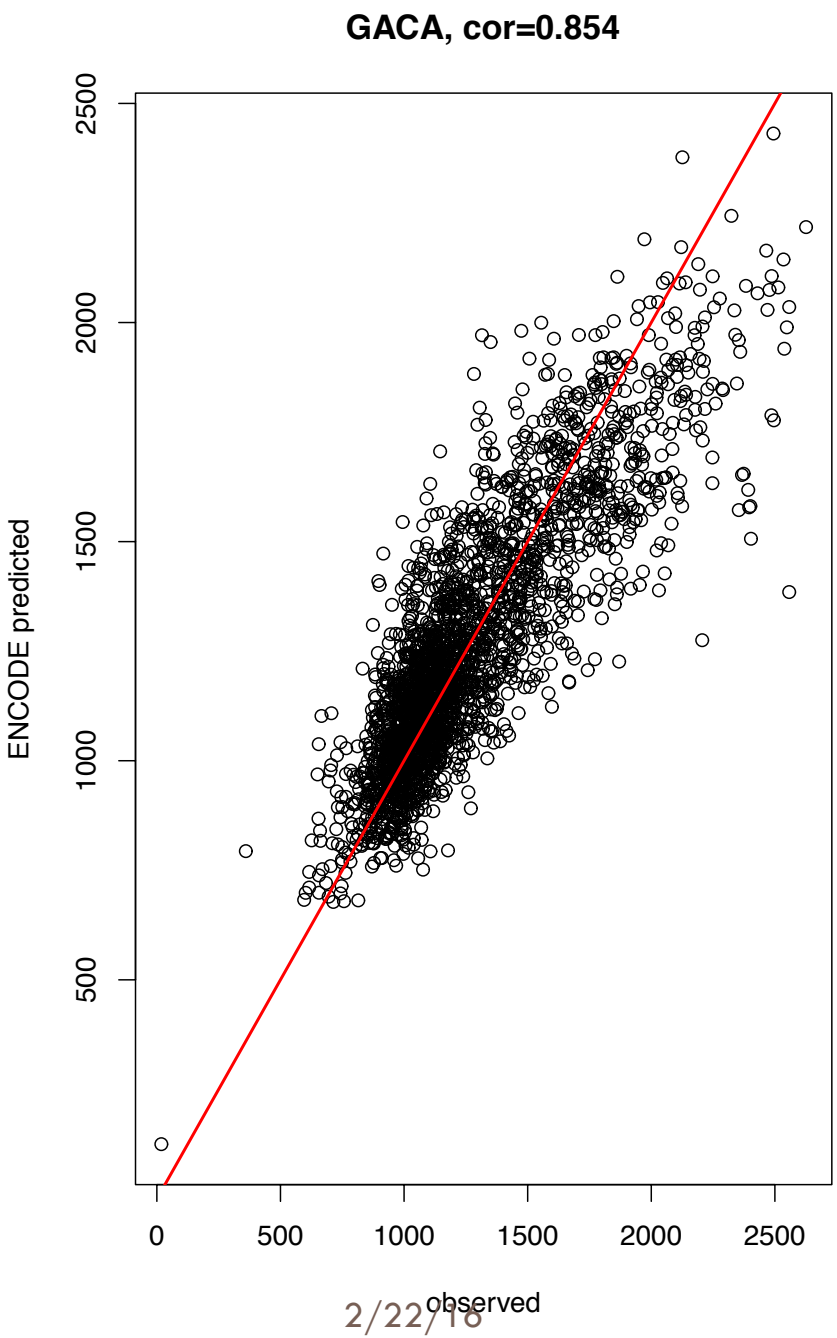
Step 3. Target gridding

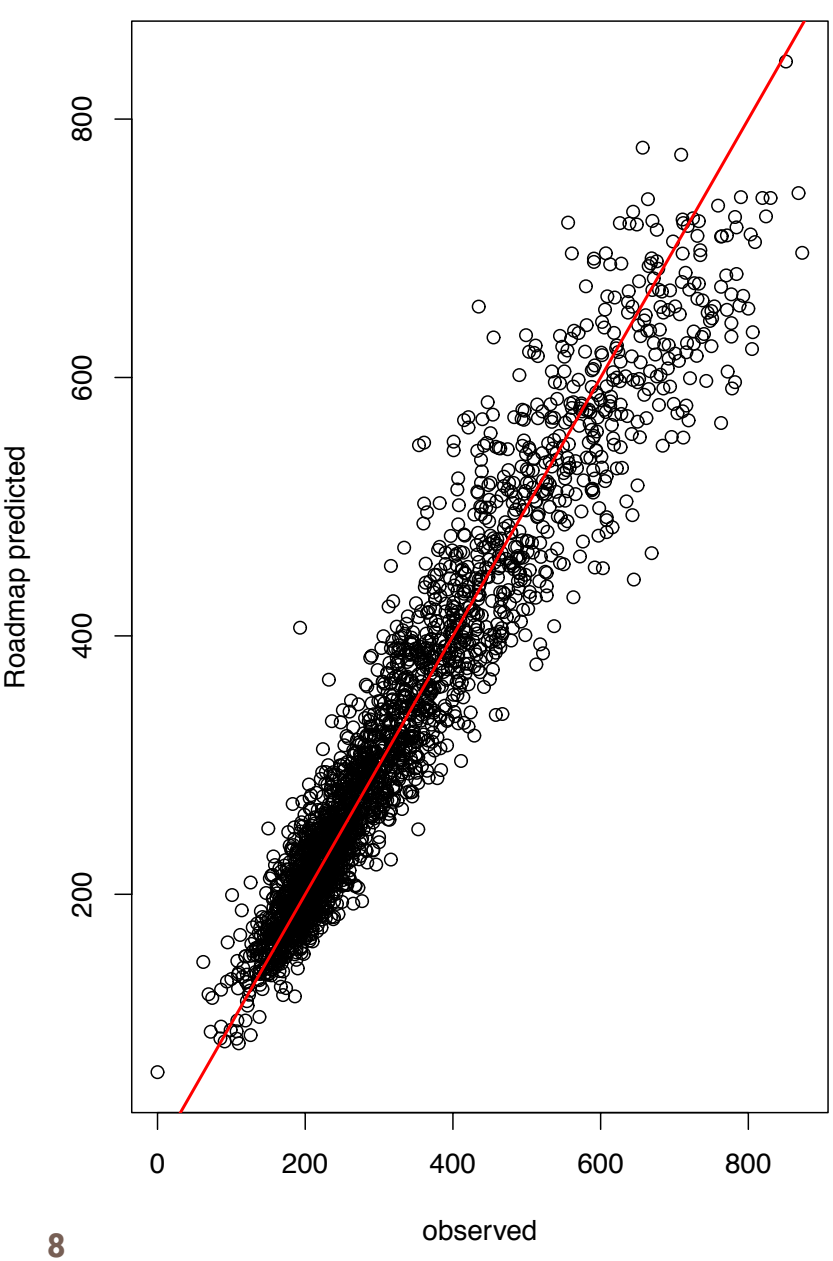


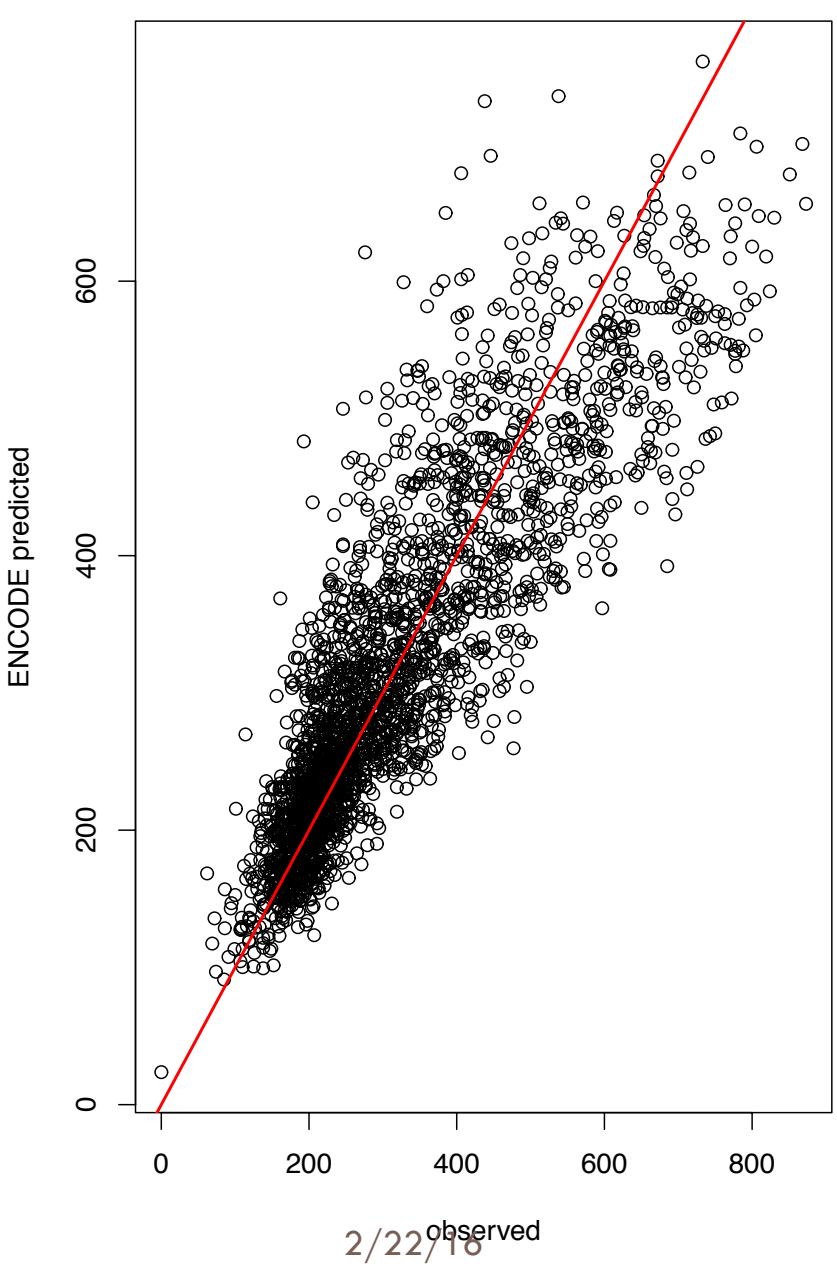

BRCA, cor=0.818
Roadmap predicted
observed
100
200
300
400
0
100
200
300
400
500

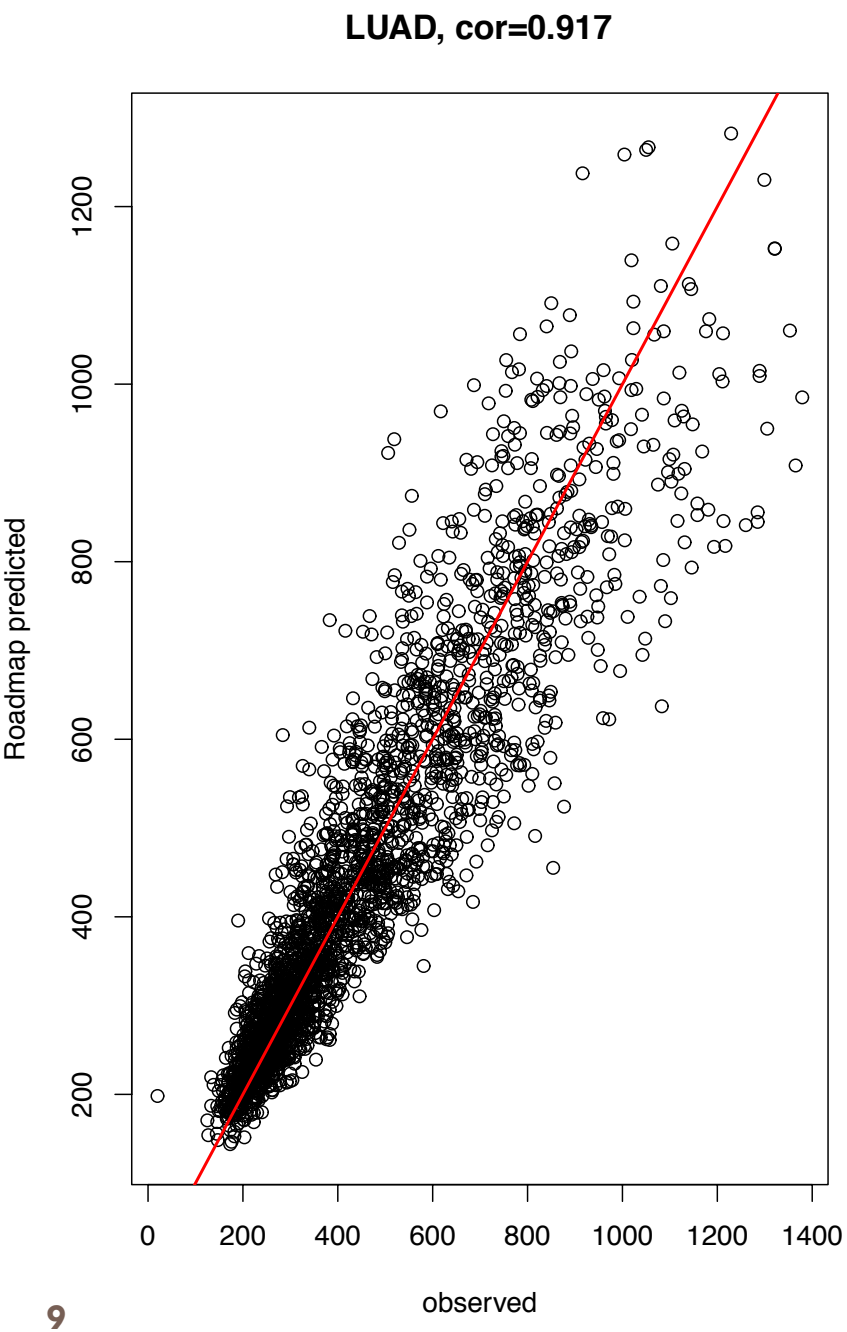

BRCA, cor=0.765
ENCODE predicted
observed
50
100
150
200
250
300
350
400
0
100
200
300
400
500

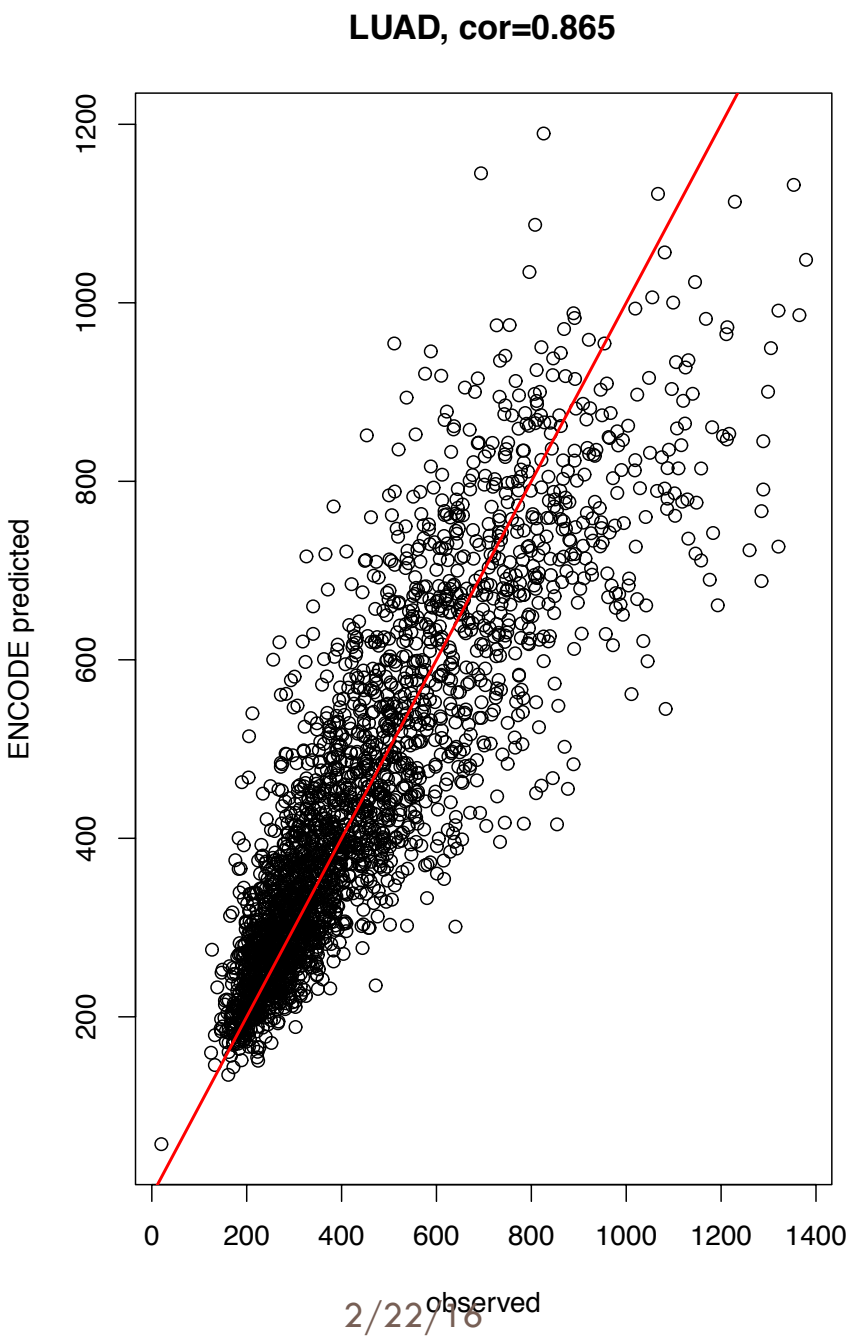

GACA, cor=0.928
Roadmap predicted
observed
0
500
1000
1500
2000
2500

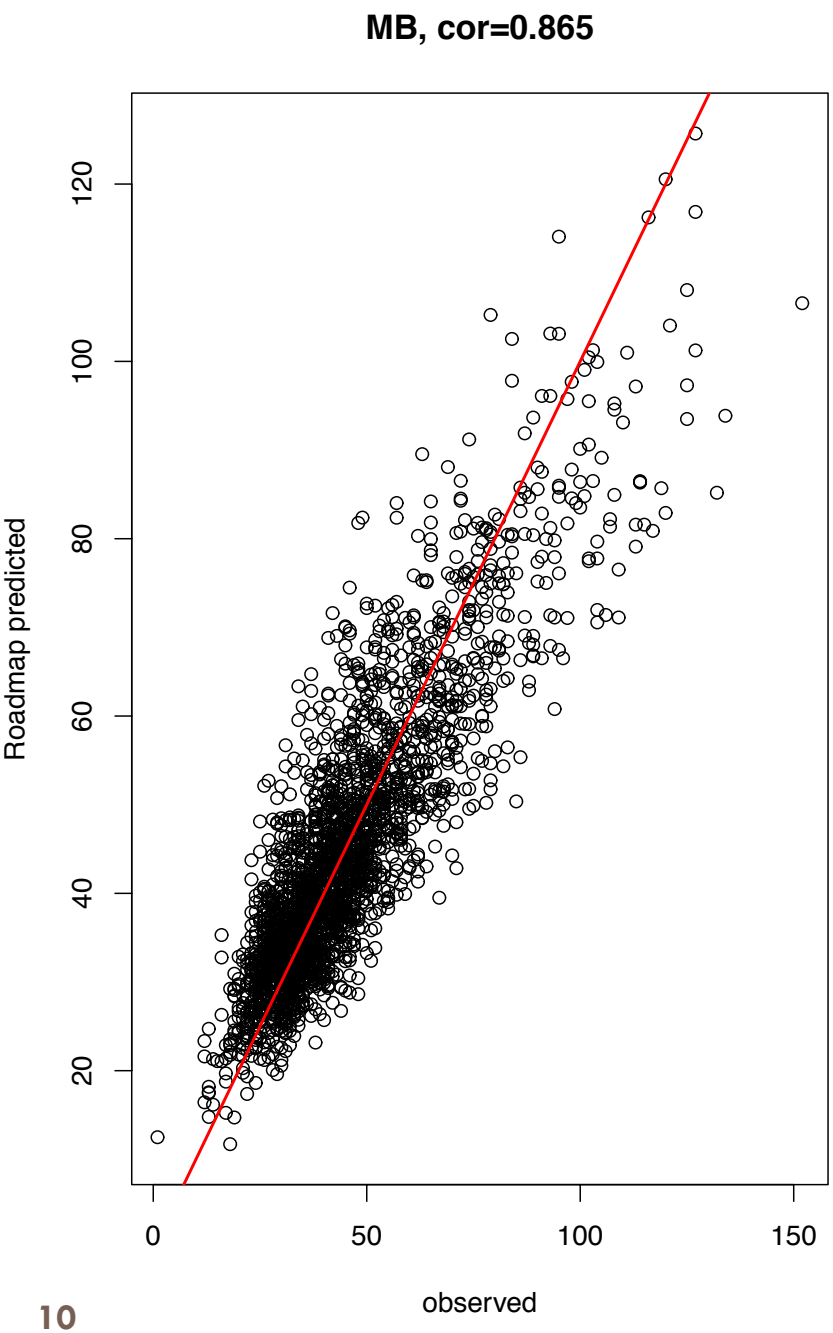

GACA, cor=0.854
ENCODE predicted
observed
0
500
1000
1500
2000
2500

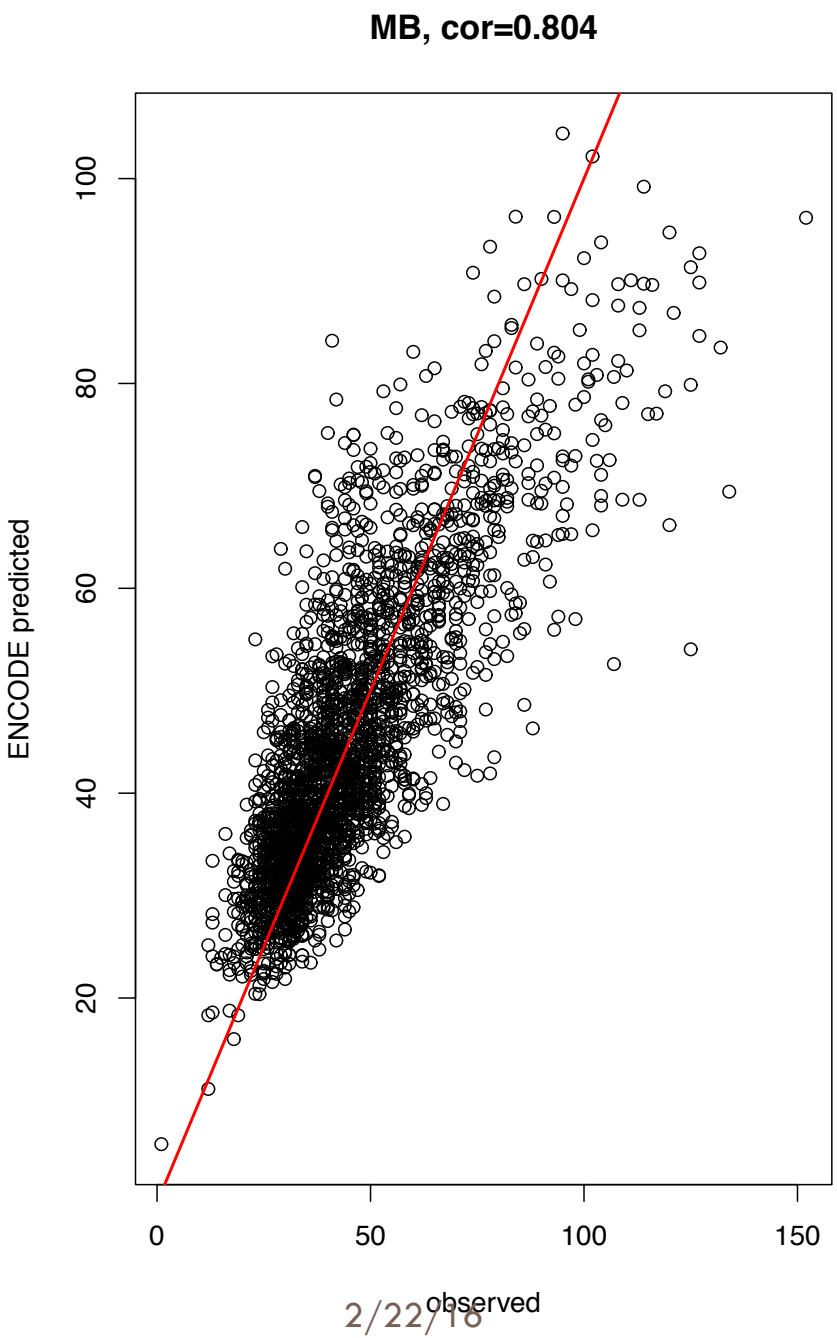

LICA, cor=0.958
Roadmap predicted
observed
0
200
400
600
800

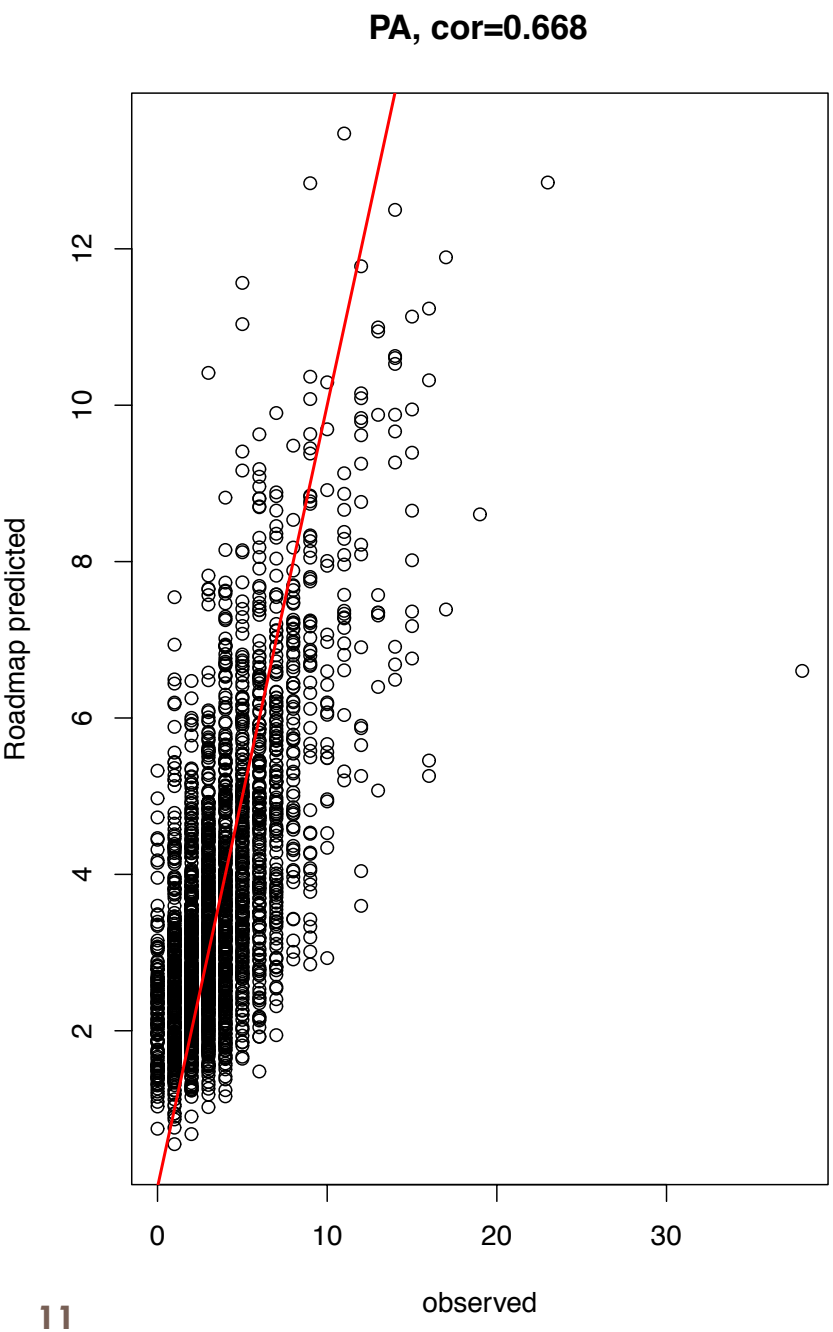

LICA, cor=0.869
ENCODE predicted
observed
0
200
400
600
800

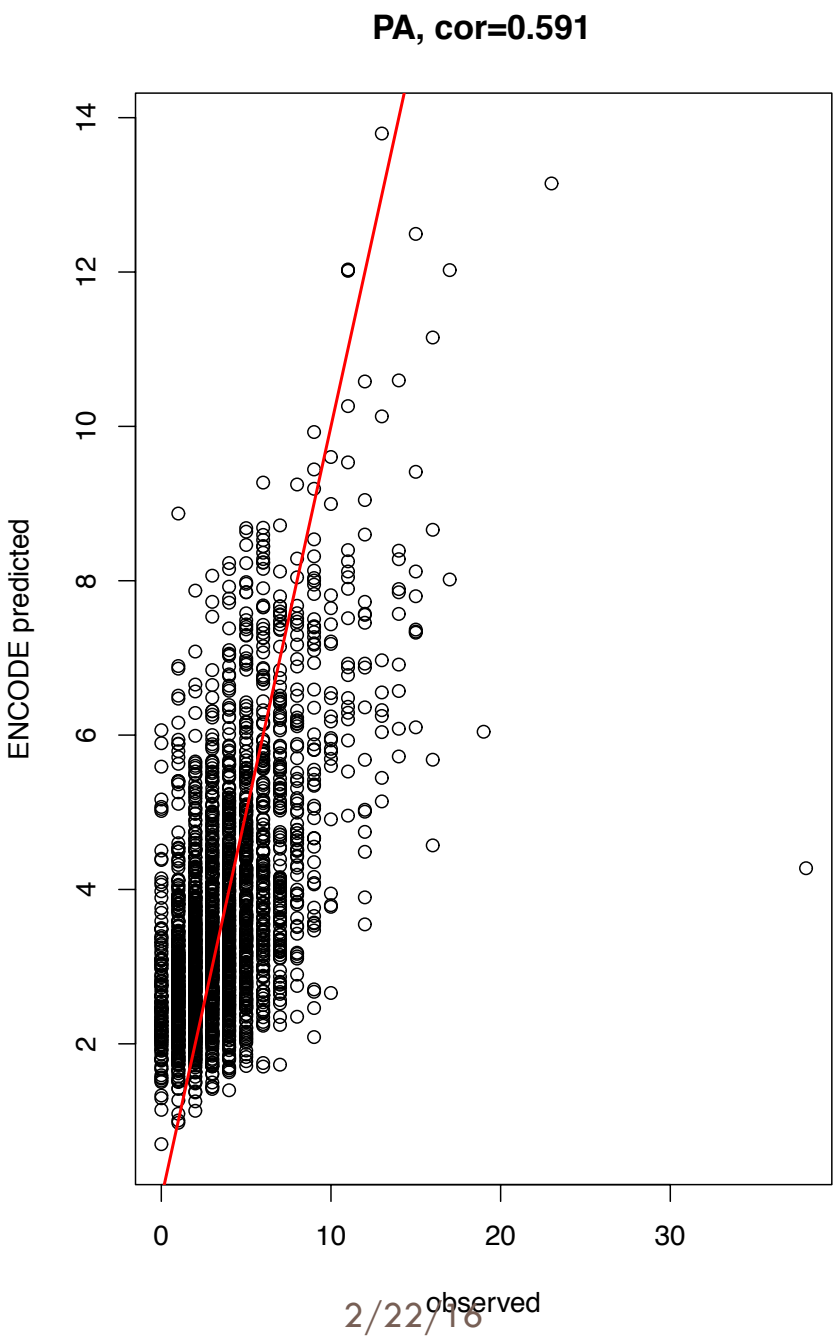

LUAD, cor=0.917
Roadmap predicted
observed
0
200
400
600
800
1000
1200
1400

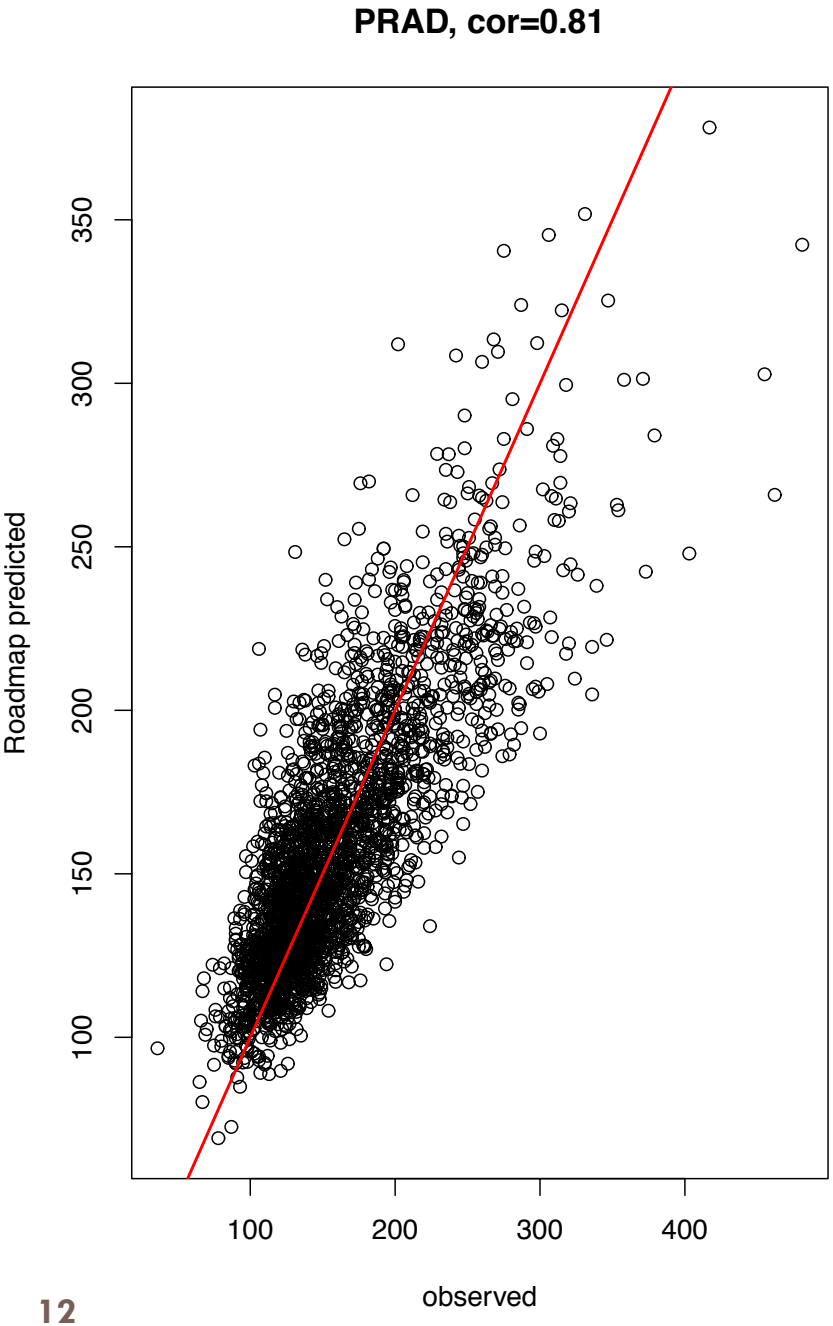

LUAD, cor=0.865
ENCODE predicted
observed
200
400
600
800
1000
1200
1400

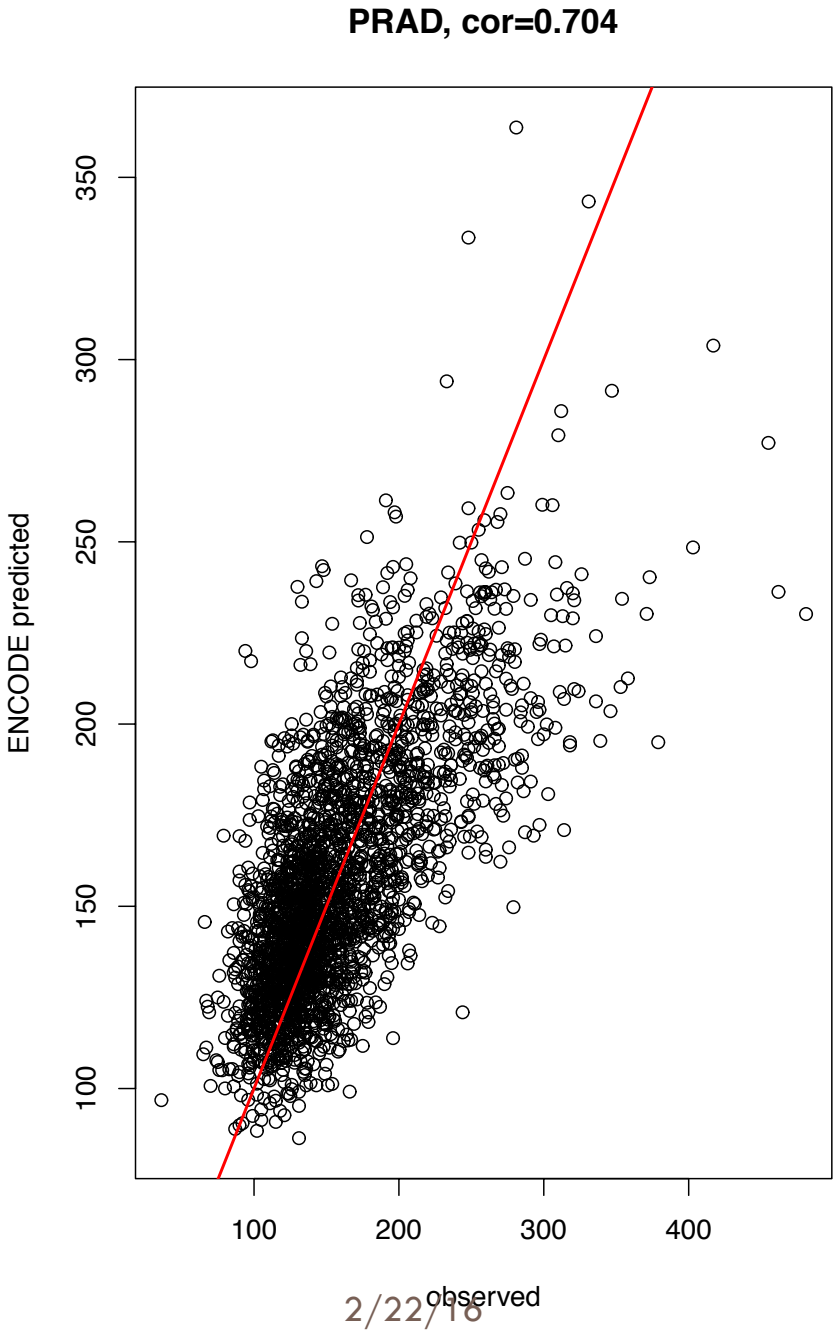

MB, cor=0.865
Roadmap predicted
observed
0
50
100
150
20
40
60
80
100
120


MB, cor=0.804
ENCODE predicted
observed
0
50
100
150
20
40
60
80
100


PA, cor=0.668
Roadmap predicted
observed
0
10
20
30
2
4
6
8
10
12


PA, cor=0.591
ENCODE predicted
observed
0
10
20
30
2
4
6
8
10
12
14


PRAD, cor=0.81
Roadmap predicted
observed
100
150
200
250
300
350
100
200
300
400


PRAD, cor=0.704
ENCODE predicted
observed
100
150
200
250
300
350
100
200
300
400