Epigenomics: Some Statistical Applications

Rafael A. Irizarry

Department of Biostatistics

John Hopkins Bloomberg School of Public Health

Acknowledgements

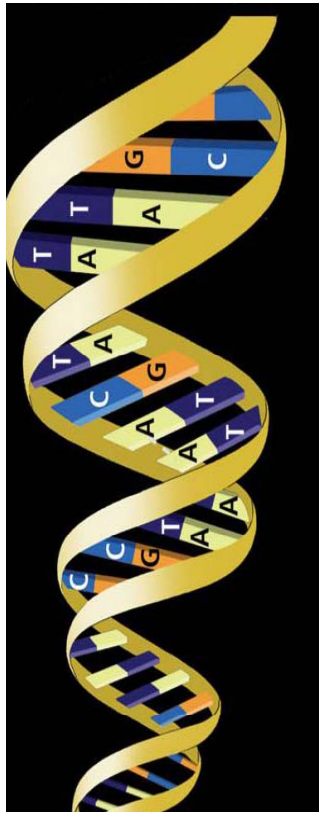
- Tom Albert, Nimblegen
- Benilton Carvalho, JHU Biostatistics
- Andy Feinberg Lab, JHU Medicine
- Todd Richmond, Nimblegen
- Hao Wu, JHU Biostatistics
- Jean Wu, Brown Biostat
- Vasan Yegnasubramanian, JHU Oncology

Outline

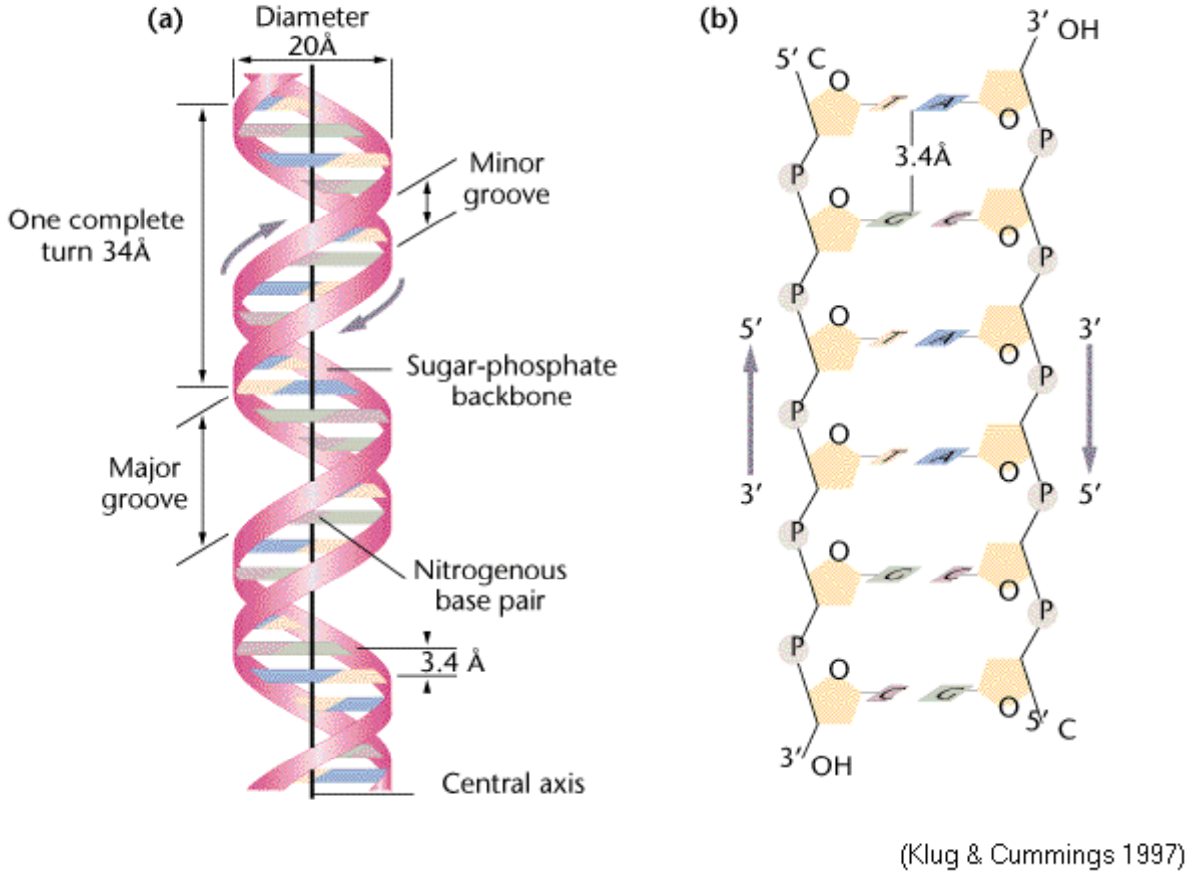
- Quick Introduction to Epigenetics
- Introduction to Methylation
- Overview of competing technologies
- Review: Expression arrays lessons
- Comparison
- Role of statisticians

Genetics: the alphabet of life

- **Letters of DNA sequence carry the information**



Epigenetics

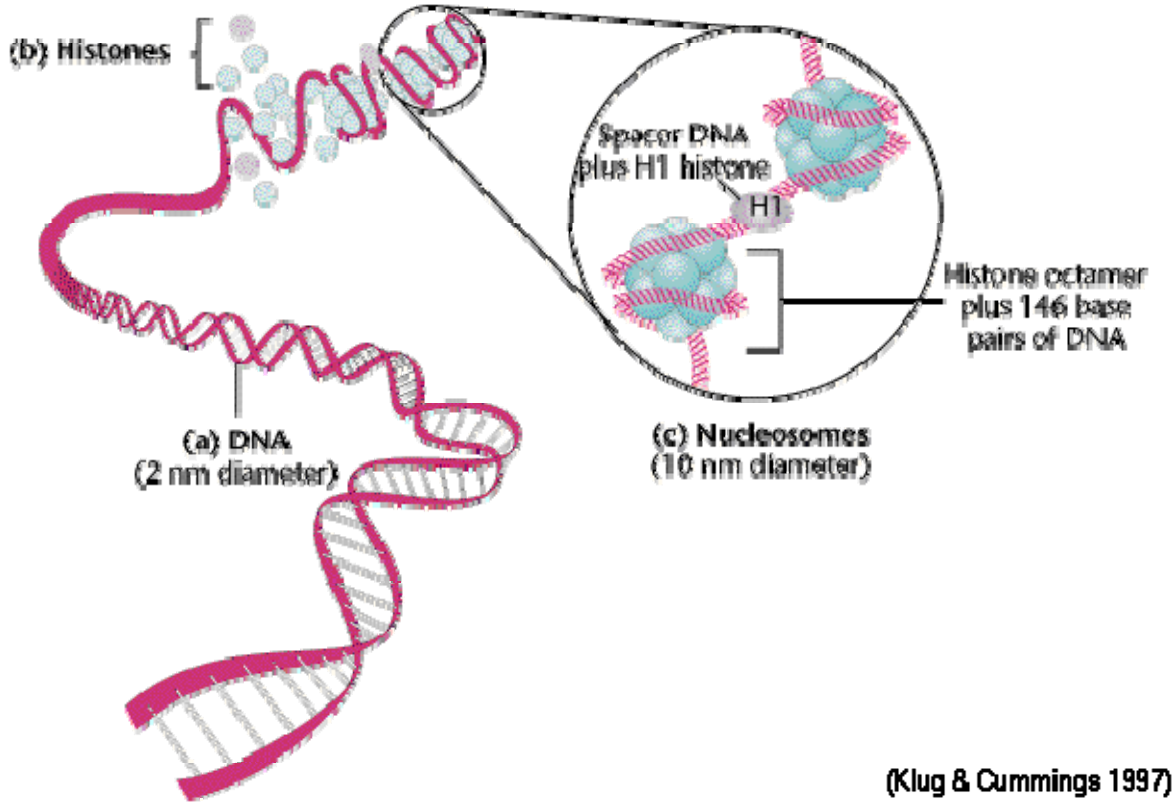


(Klug & Cummings 1997)

(3.4x10^{-10} meters/bp) x (6x10^{9} bp/genome) = ~2 meters/genome

Radius of the nucleus is ~ 10 µM !!!

Klug and Cummings, 1997

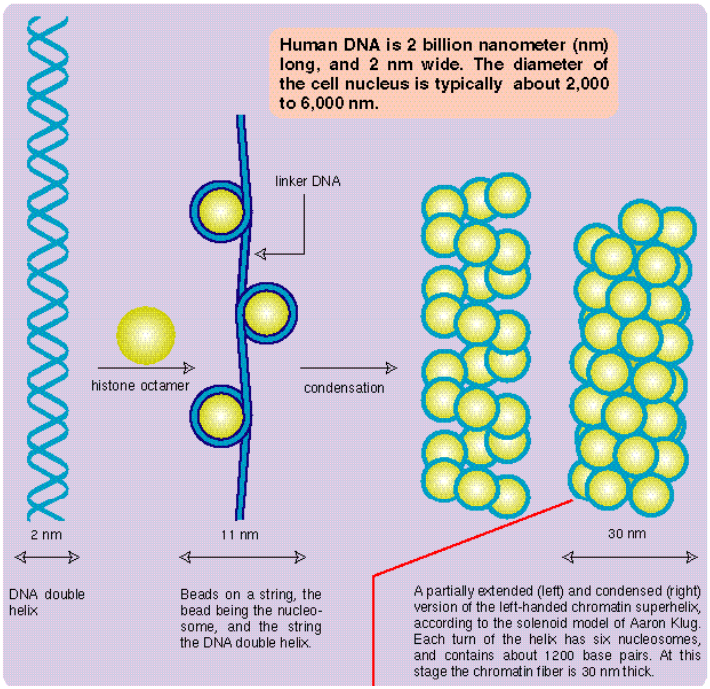


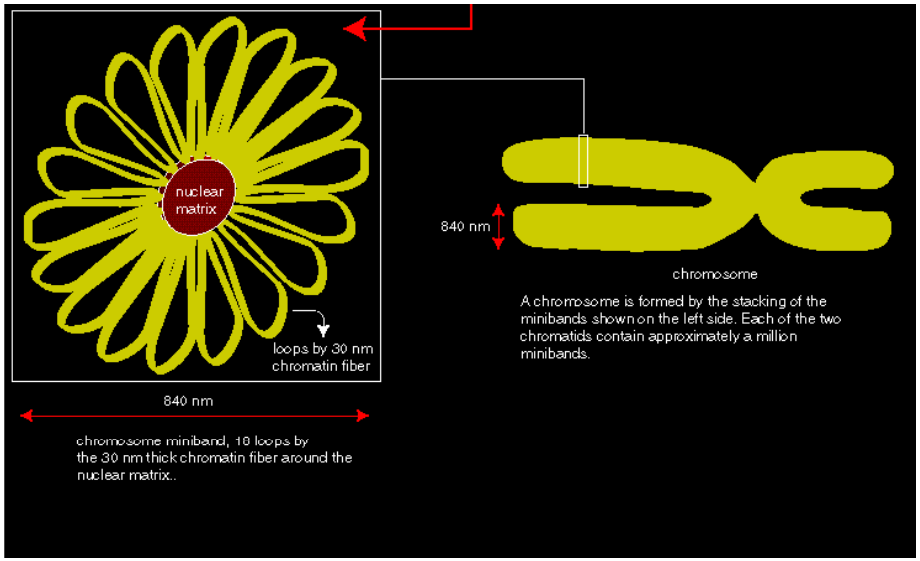
(Klug & Cummings 1997)

[(6 x 10^9 bp/genome) / (195 bp/nucleosome)] = ~ 30.8 x 10^6 nucleosomes/genome

~ 5 % of nuclear volume

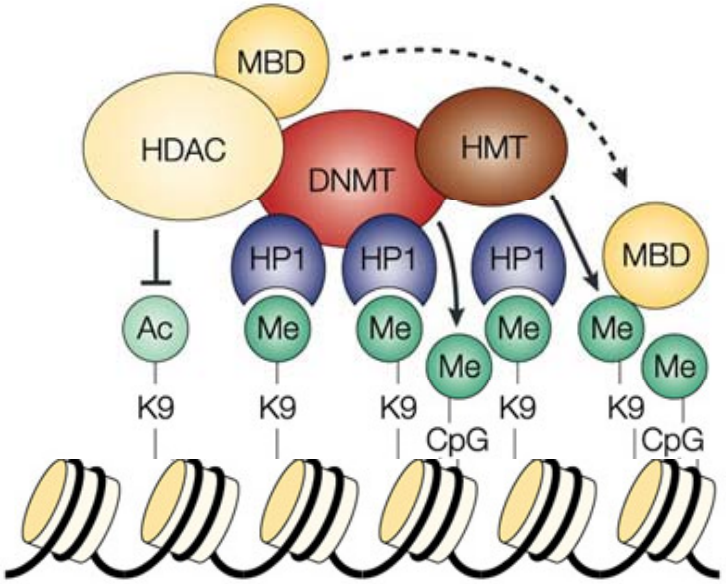
Nucleosome, Solenoid Model of Chromatin and Chromosome



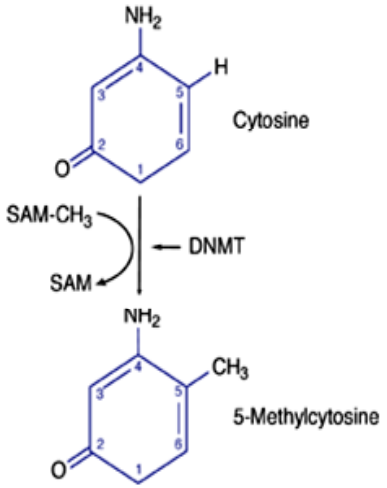


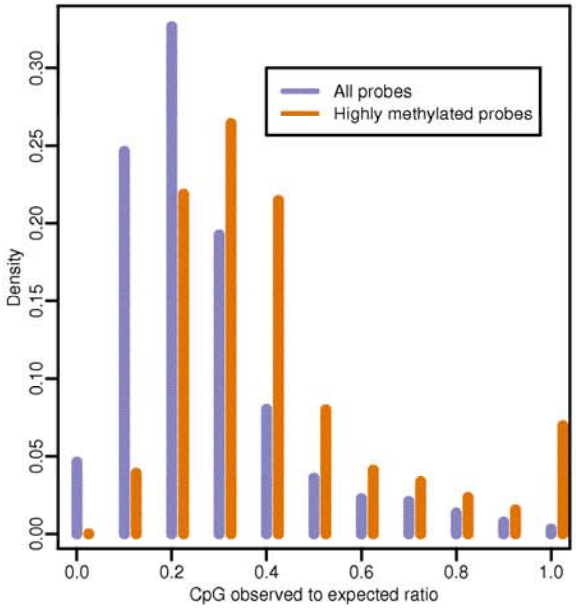
http://www.albany.edu/~achm110/solenoidchriomatin.html

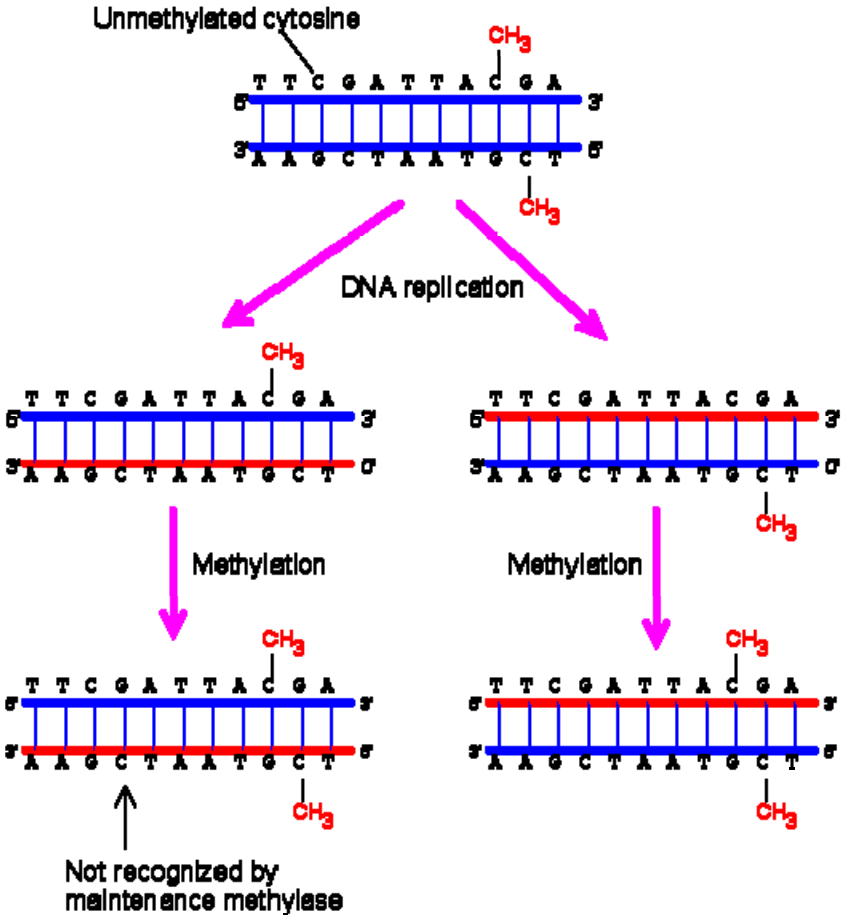
Epigenetics: the grammar of life



DNA methylation

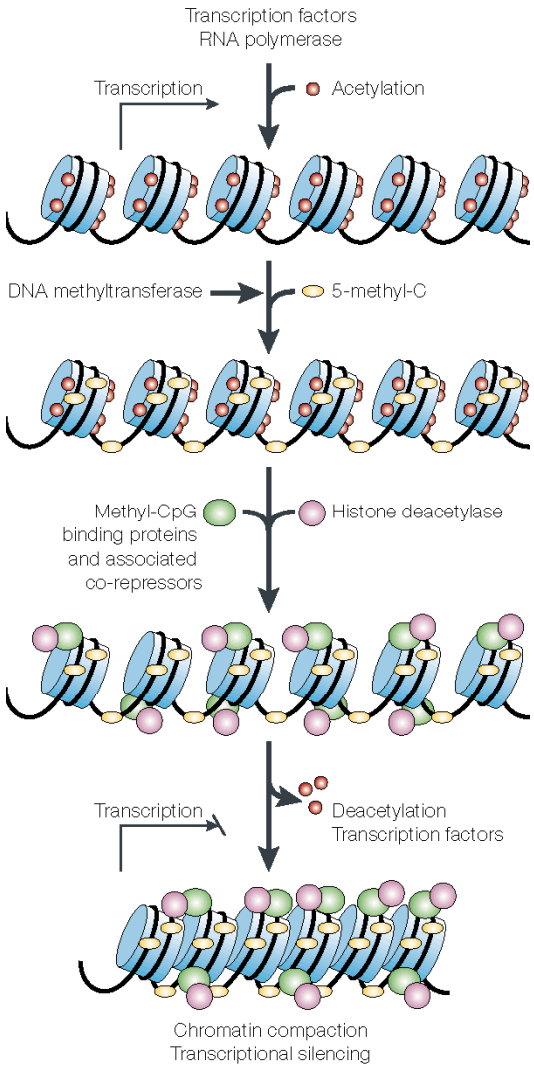


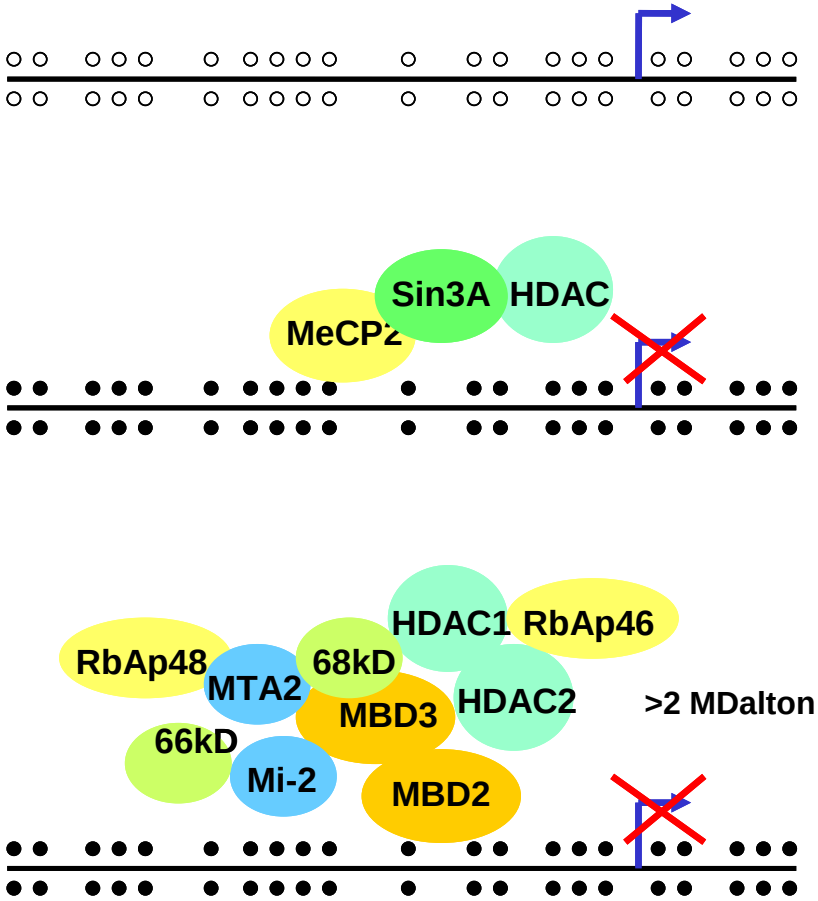




Obsered to expeced = Pr(CG) / { Pr(C) Pr(G) }

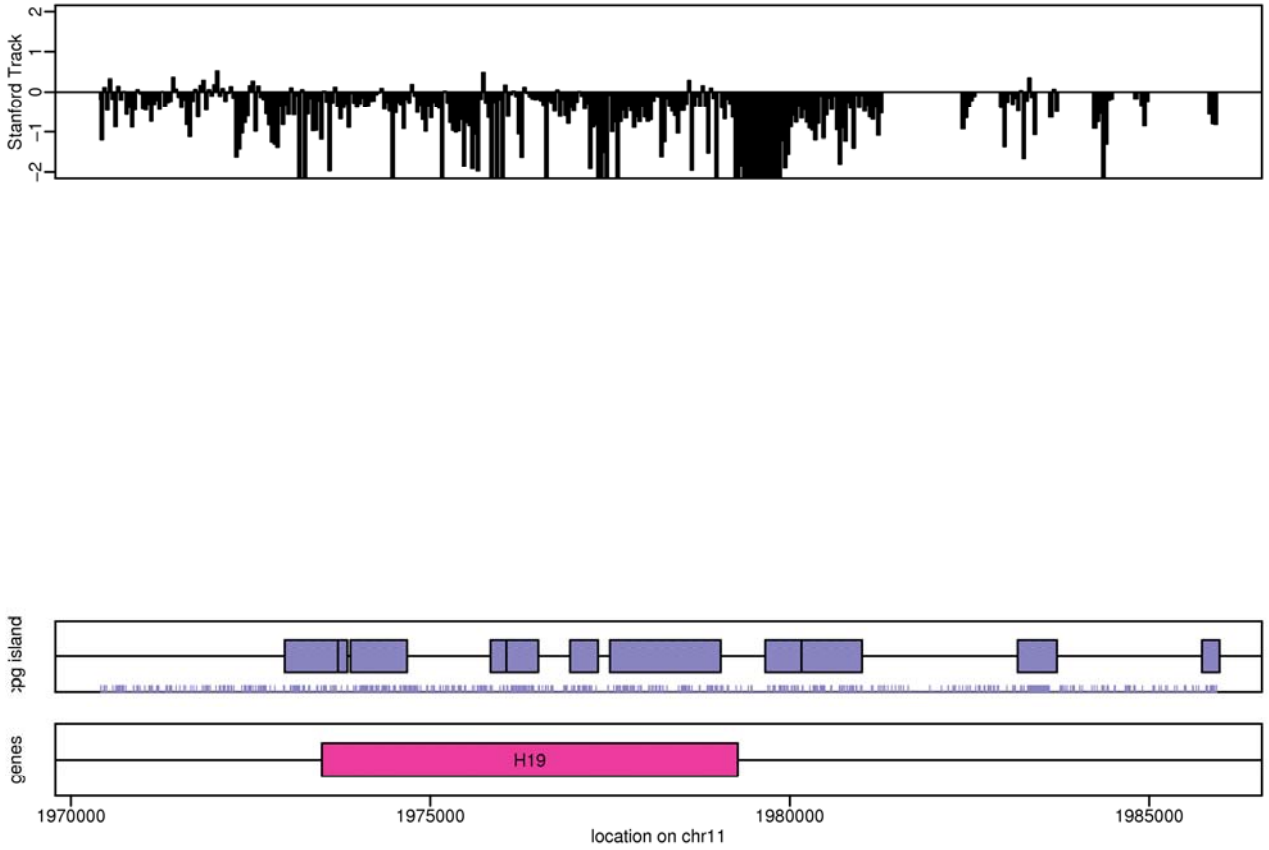
DNA methylation can lead to silencing of gene expression





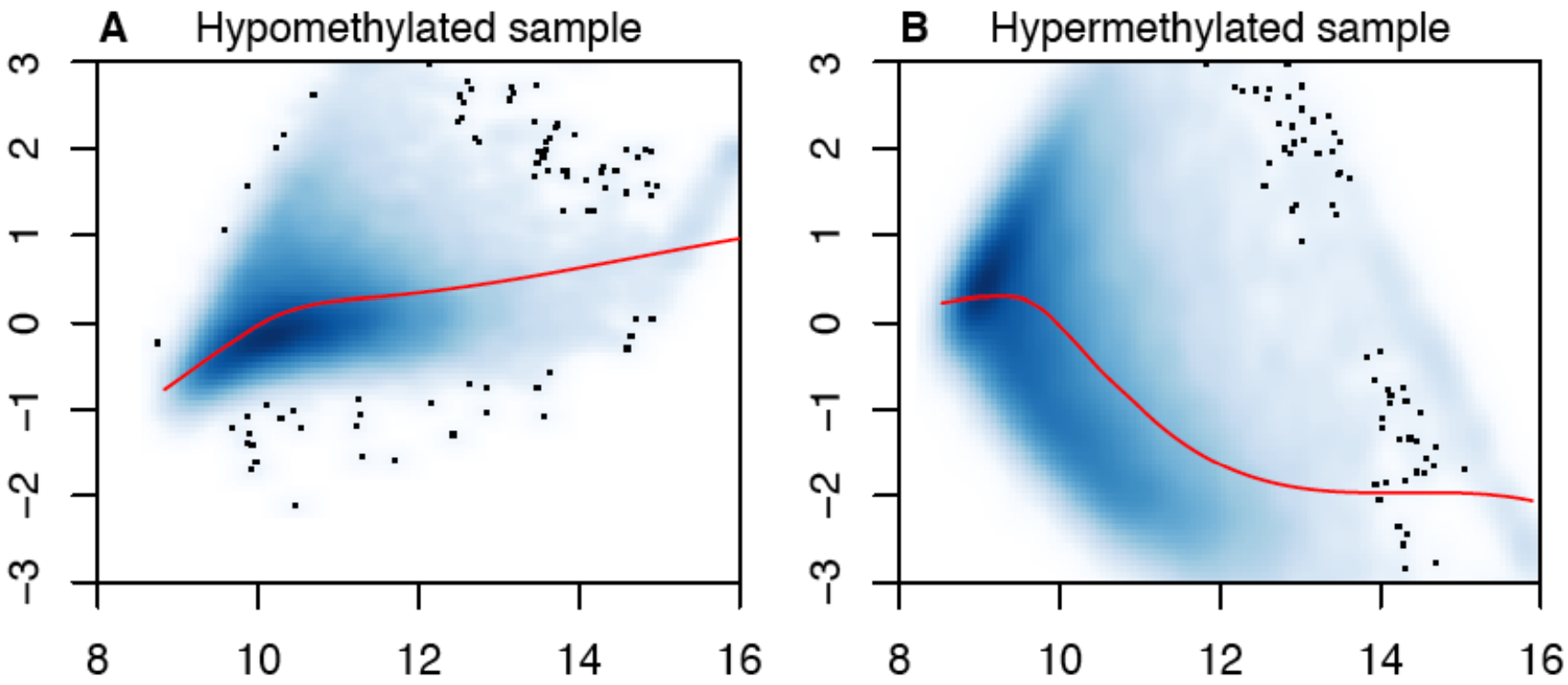
Robertson and Wolffe, Nat Rev Genet, 2000

ENCODE Track

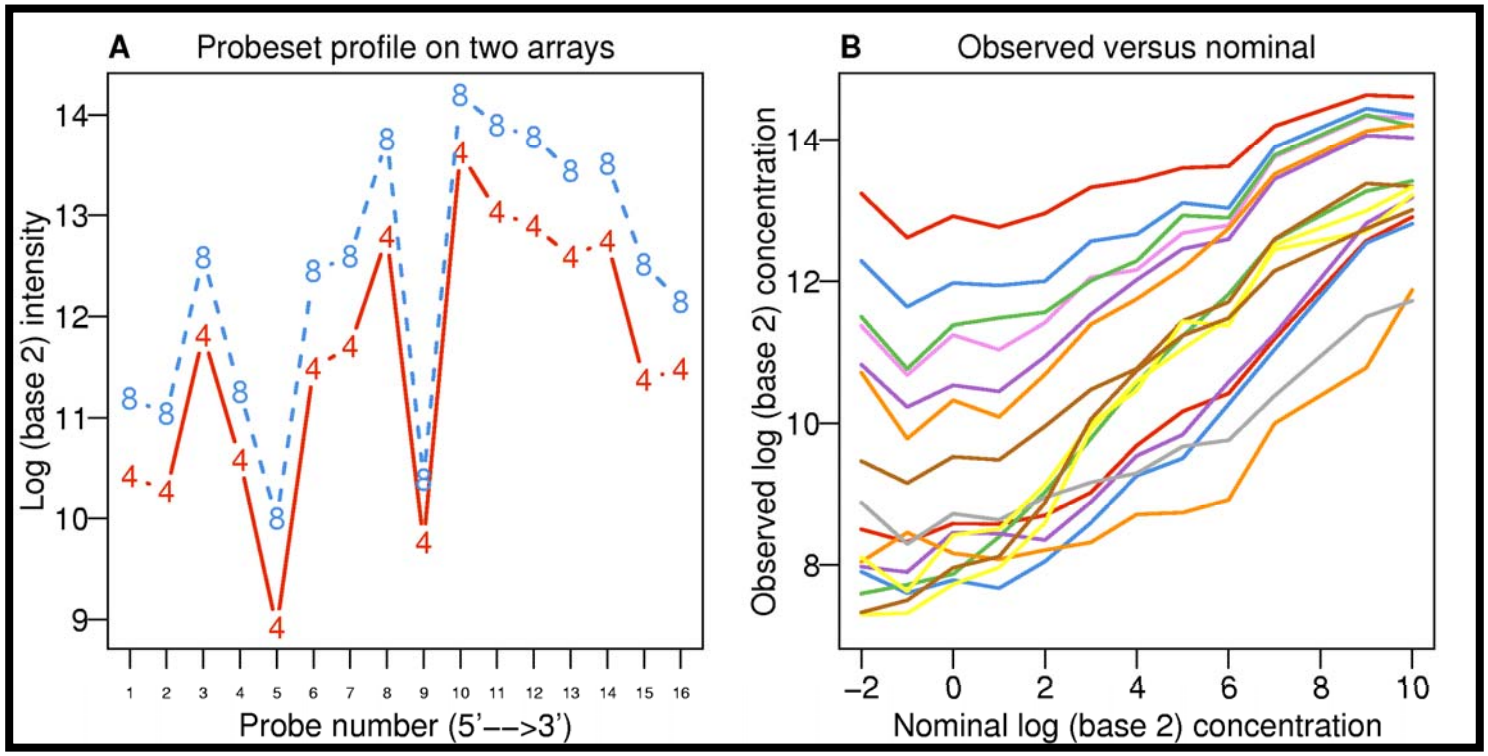


Expression Array Lessons

Normalization



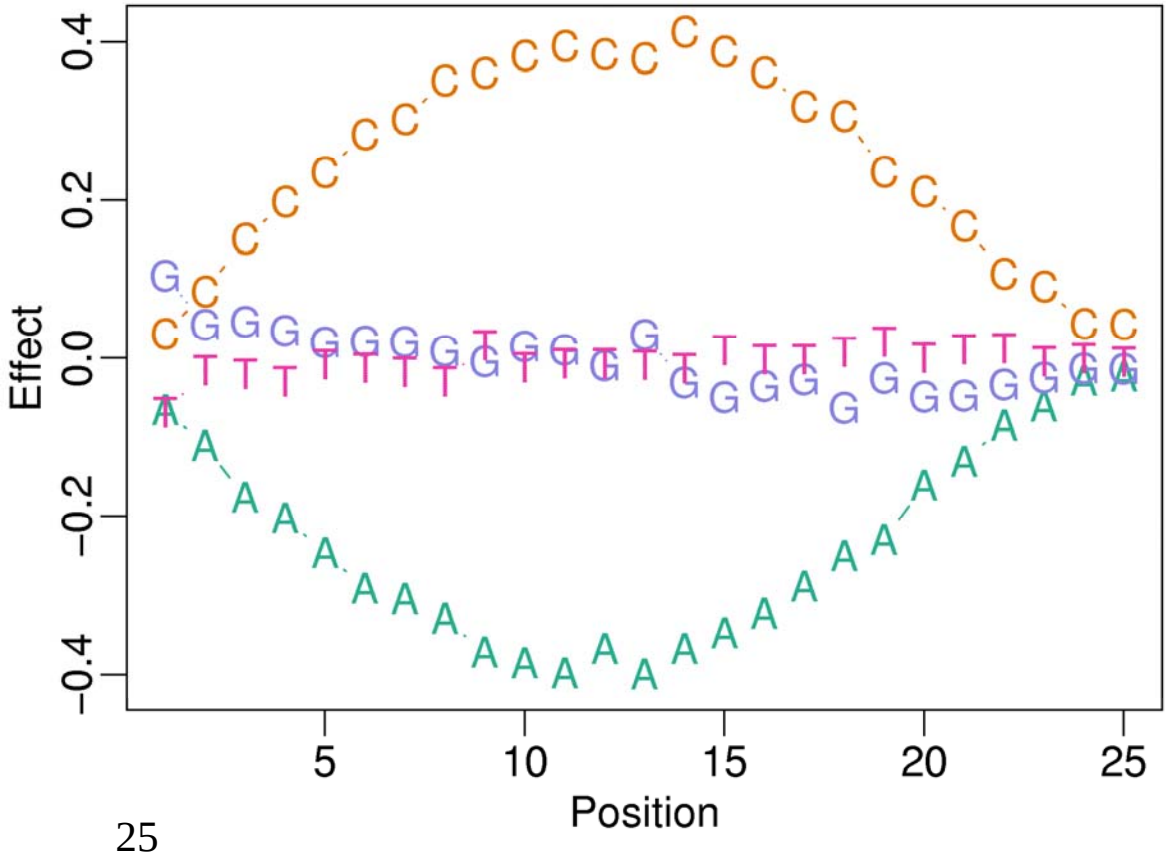
Probe effect



Intensity = Background + Probe Effect x Quantity x Erro

Sequence effect for BG

Wu et al. (2004) JASA 99(468) 909



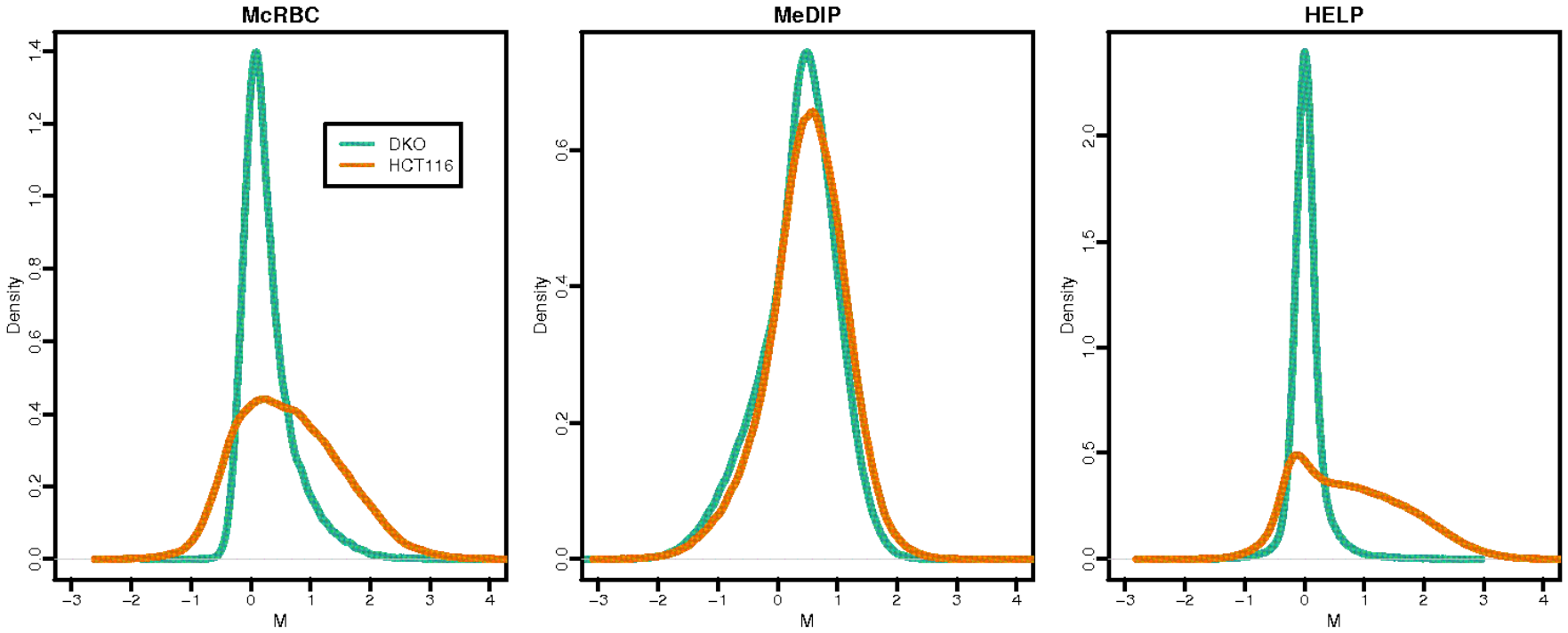
$$Affinity = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k = j}$$

$\mu_{j,k}$ ~ *smooth function of k*

Back to Methylation

High throughput of course….

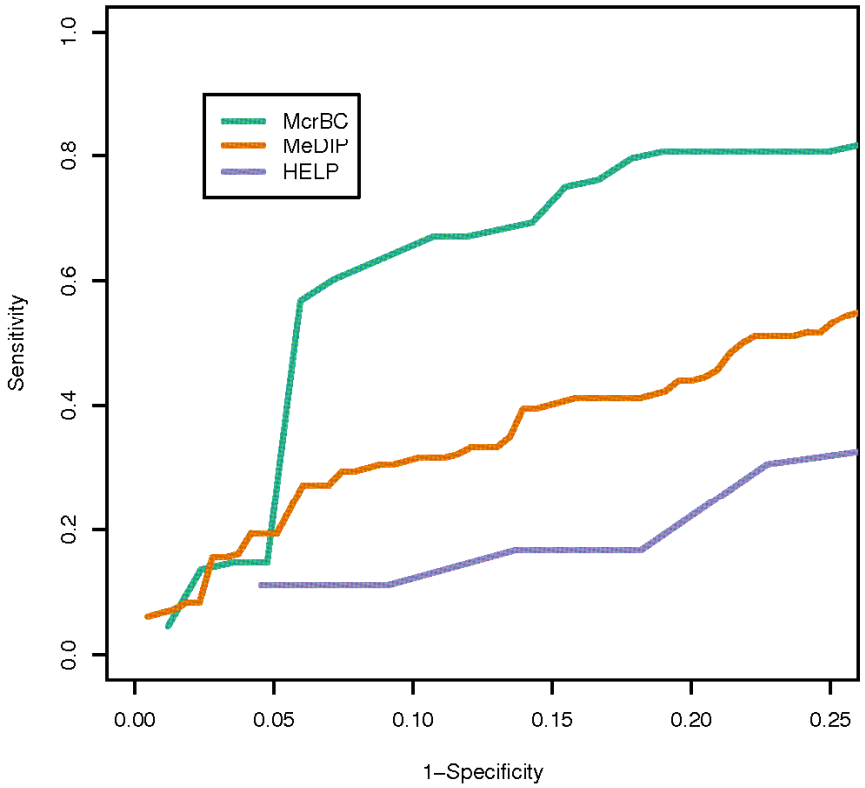
Densities for three methods



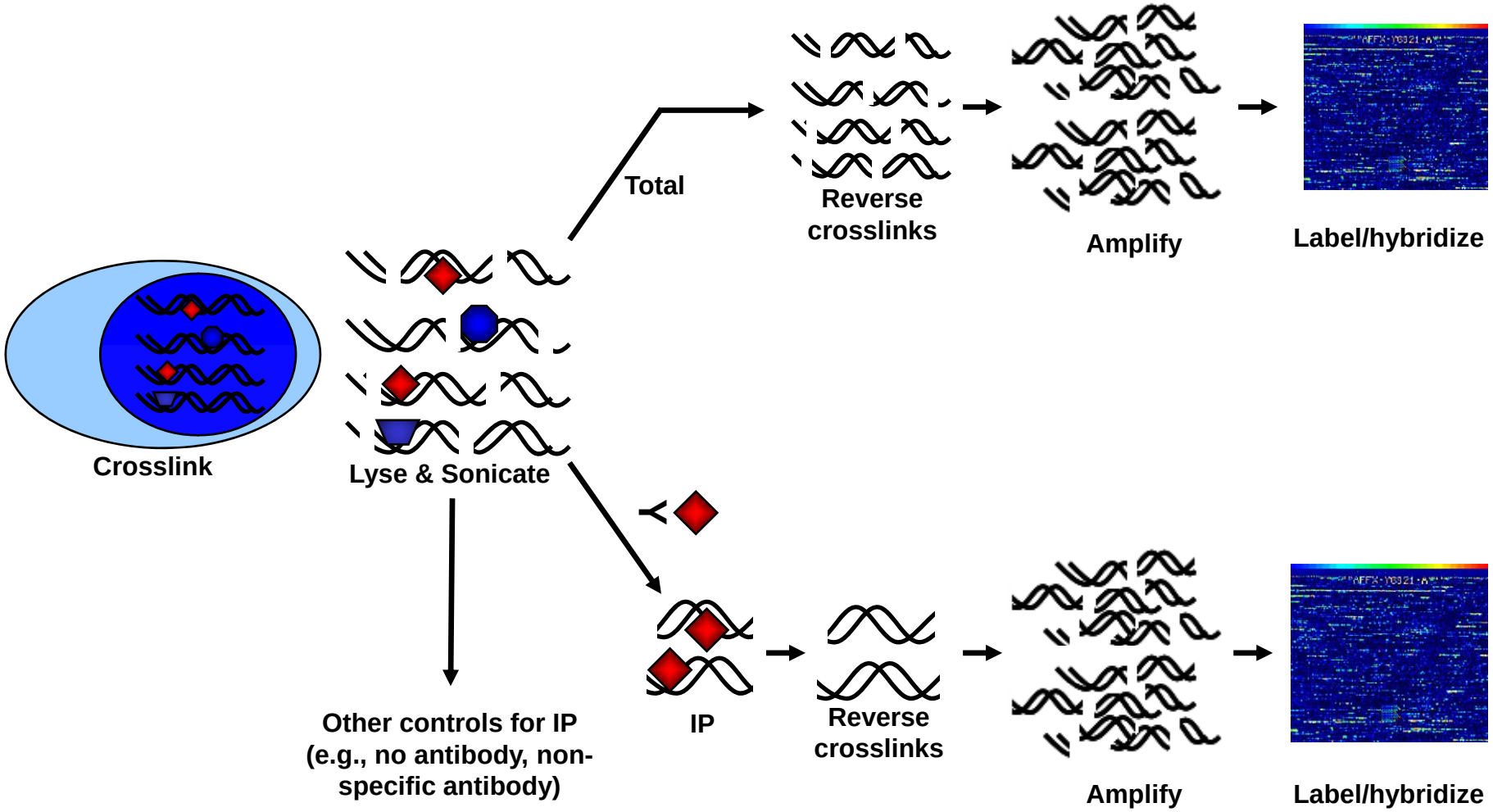
HCT116 lots of methylation

DKO very little methylation

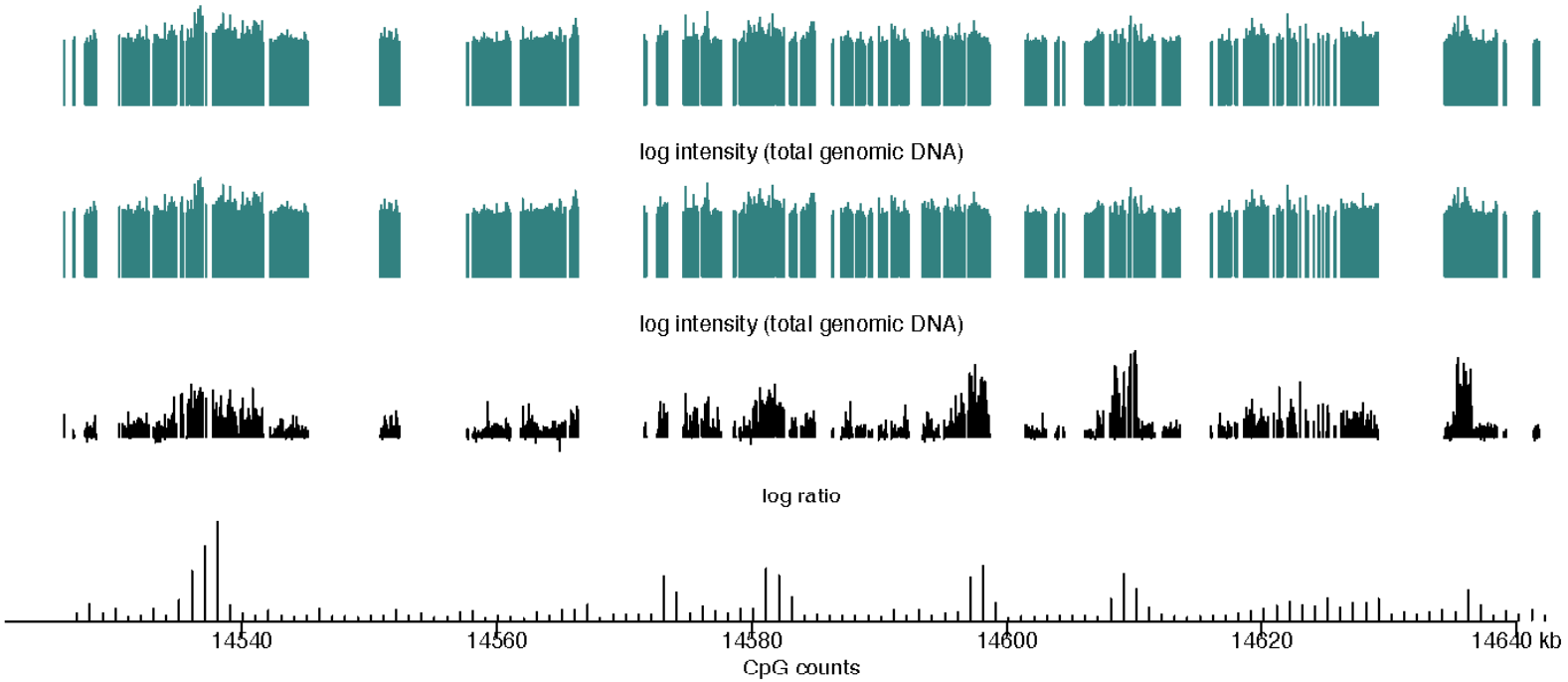
Hunh?


McrBC
MeDIP
HELP
Sensitivity
1–Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25

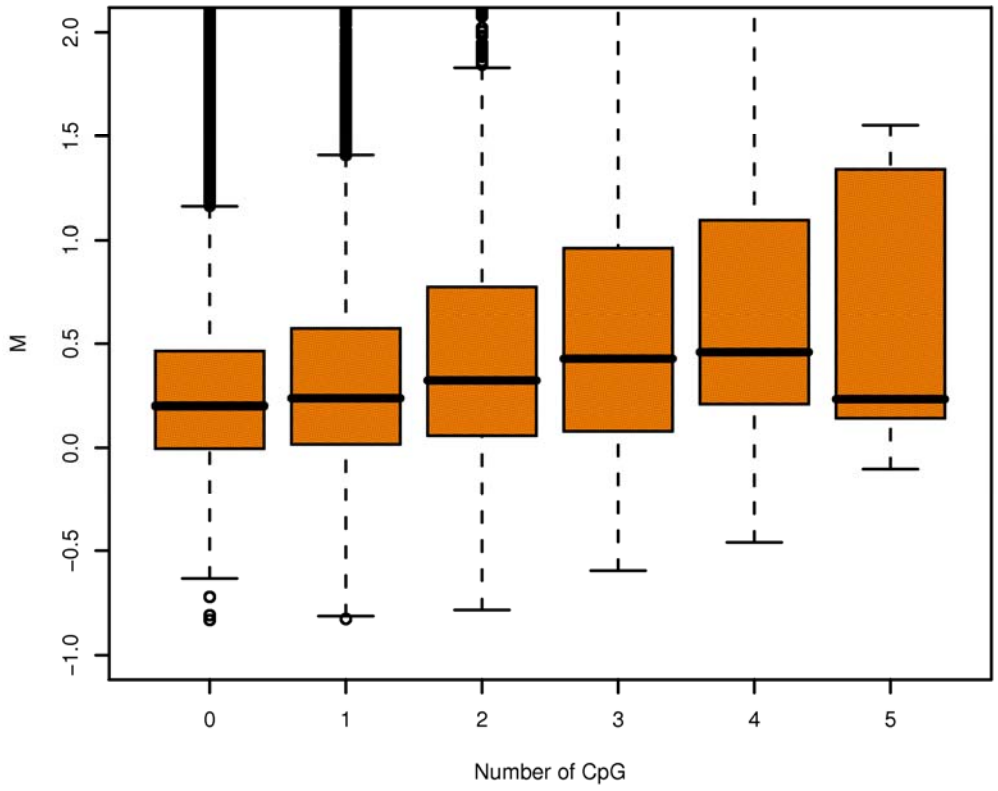
MeDIP (like ChIPchip)



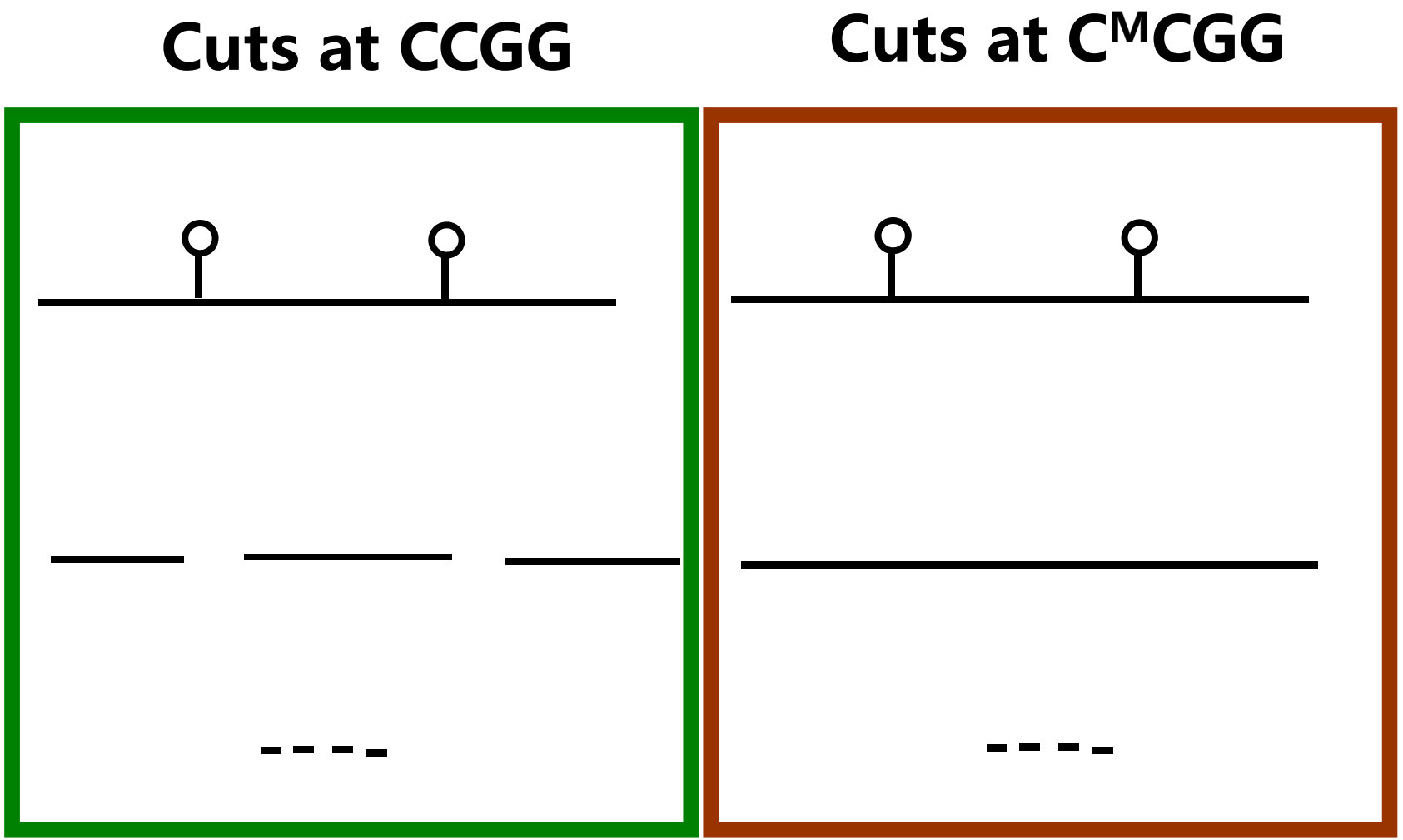
Some Data


log intensity (total genomic DNA)
log intensity (total genomic DNA)
log ratio
14540
14560
14580
14600
14620
14640 kb
CpG counts

Problem: Not specific

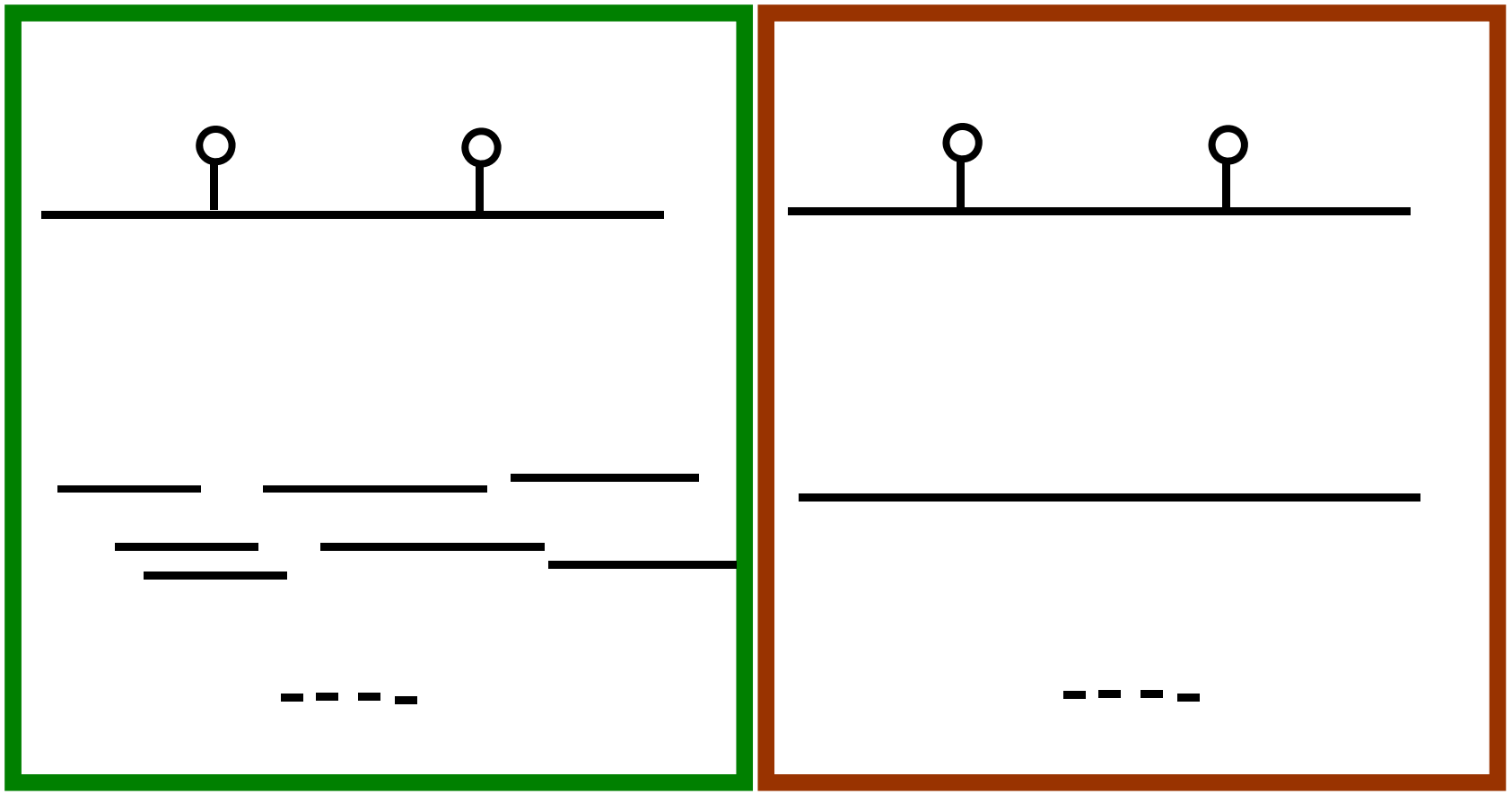


HELP: Two enzymes



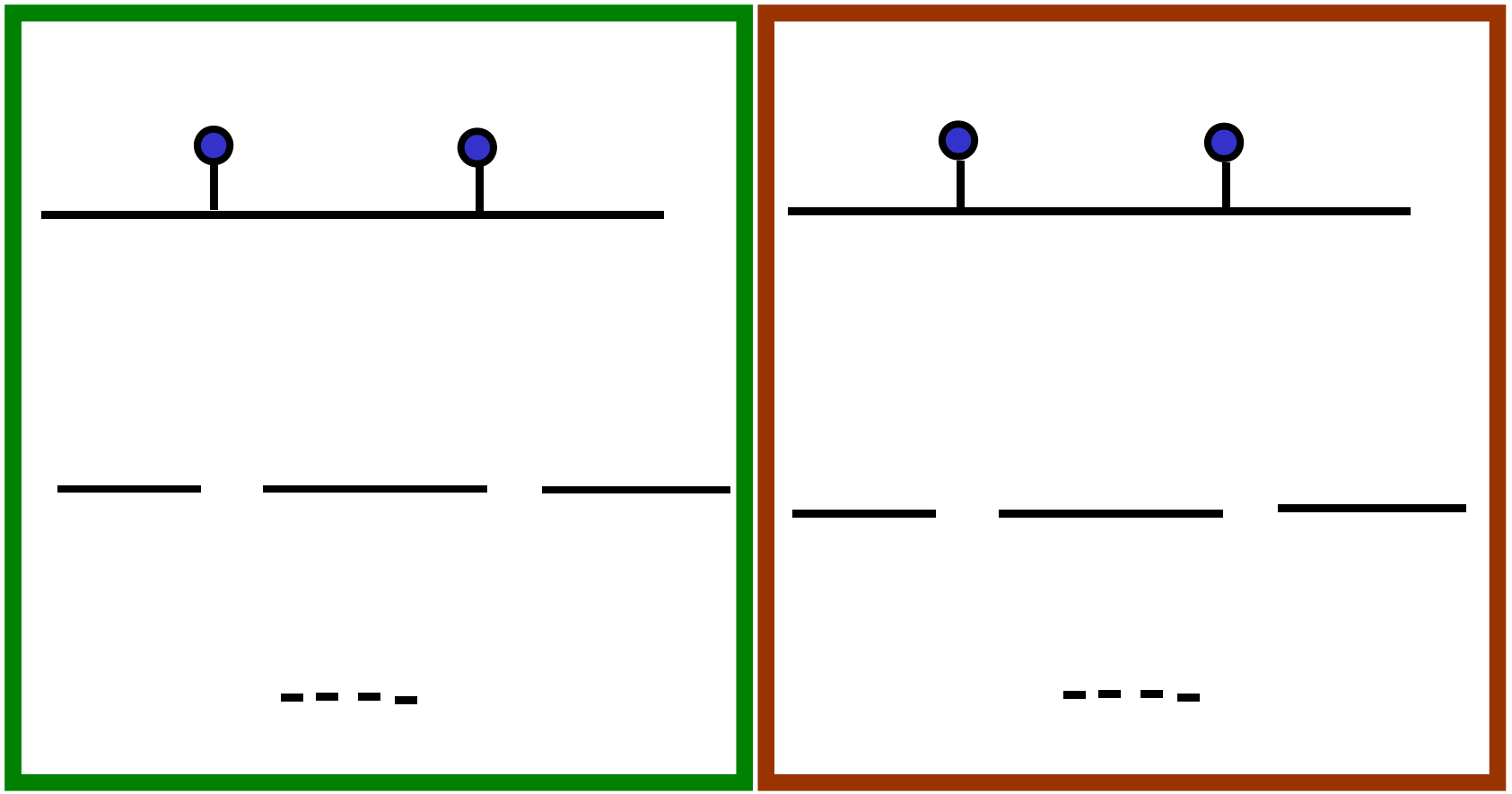
No Methylation

HELP after PCR



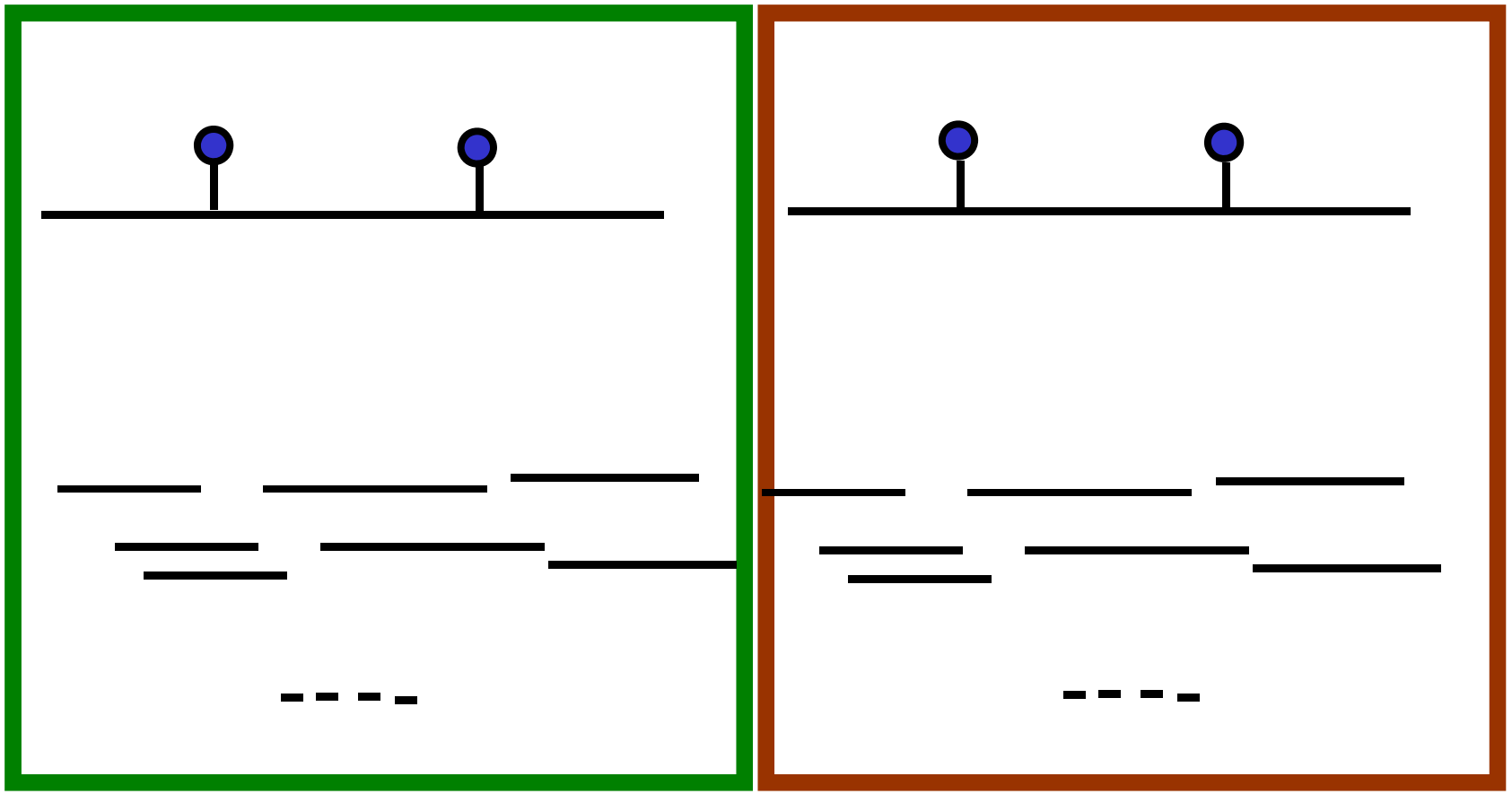
No Methylation

HELP



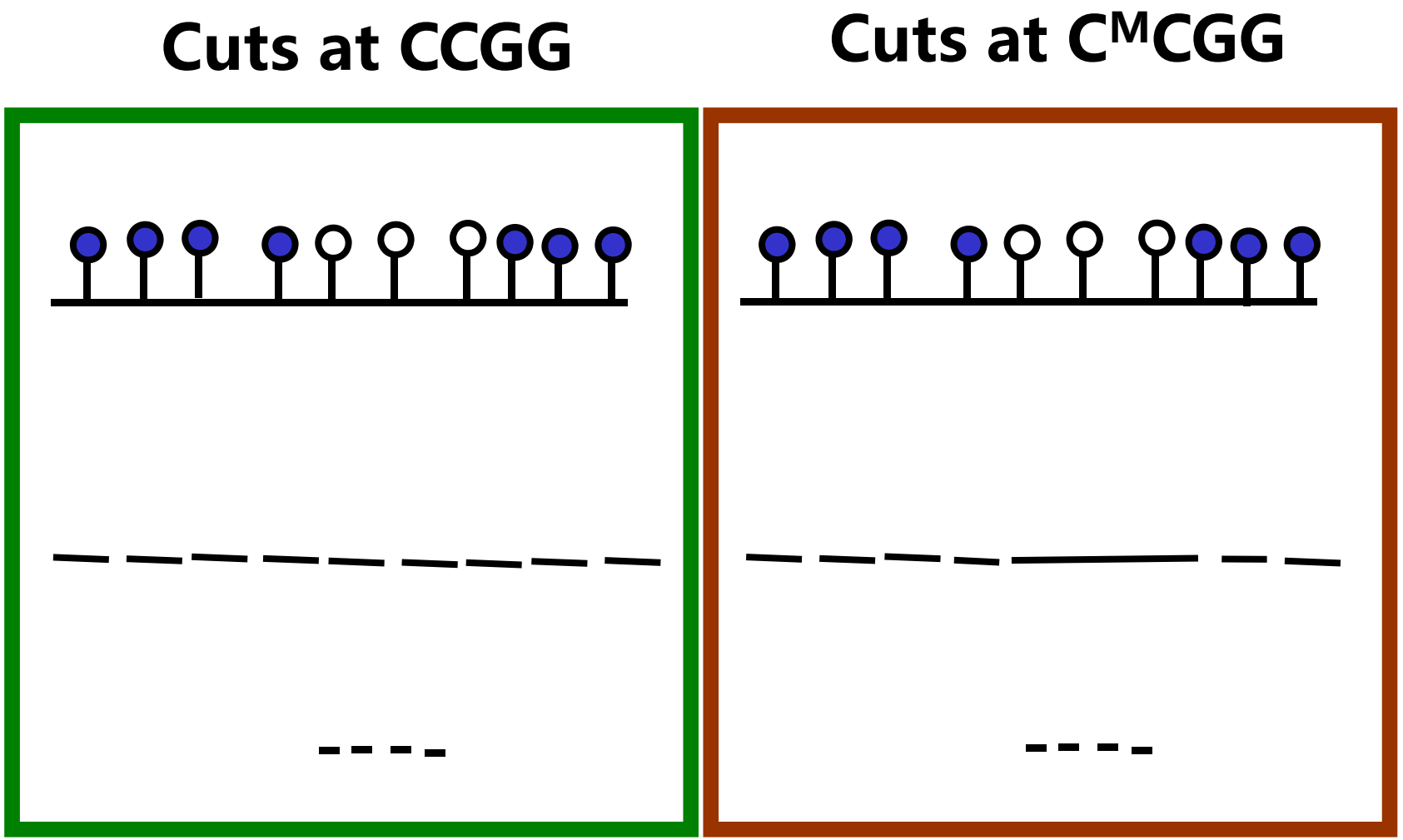
Methylation

HELP



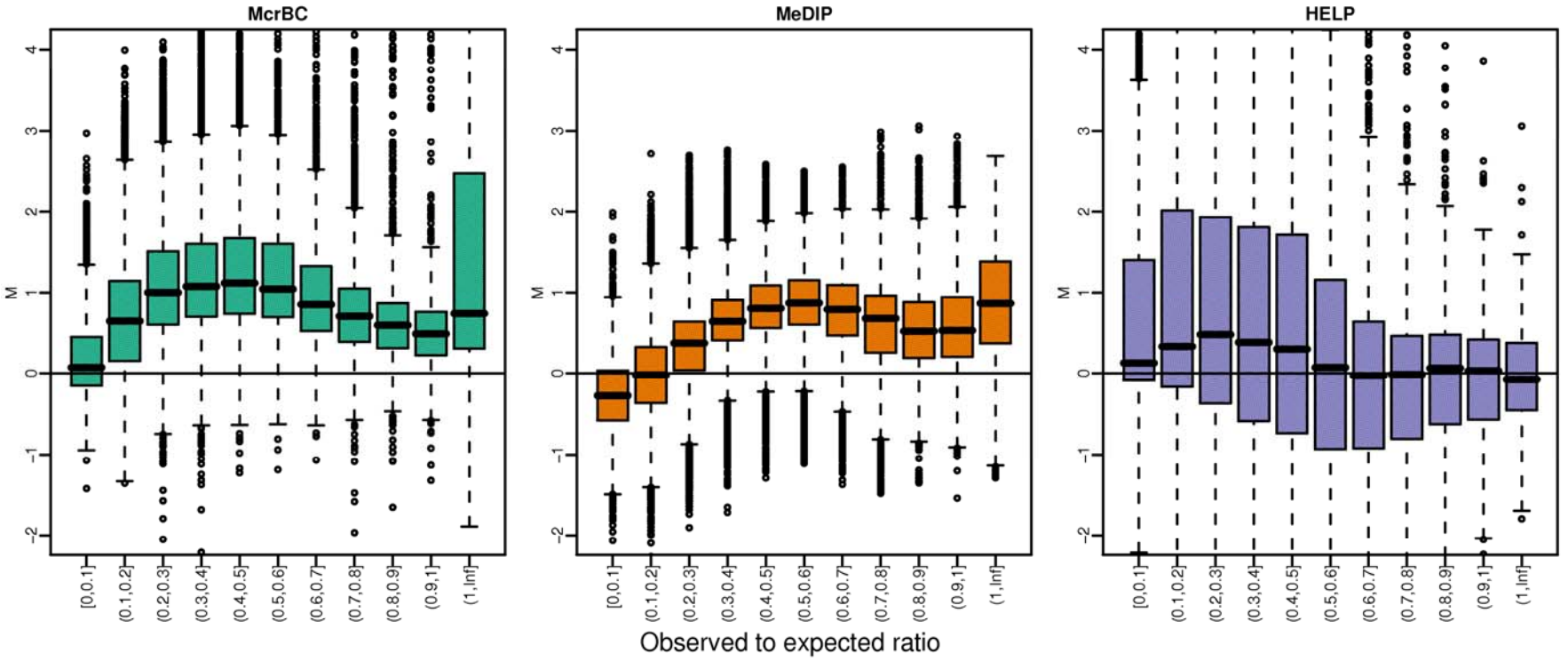
No Methylation

Problem with HELP



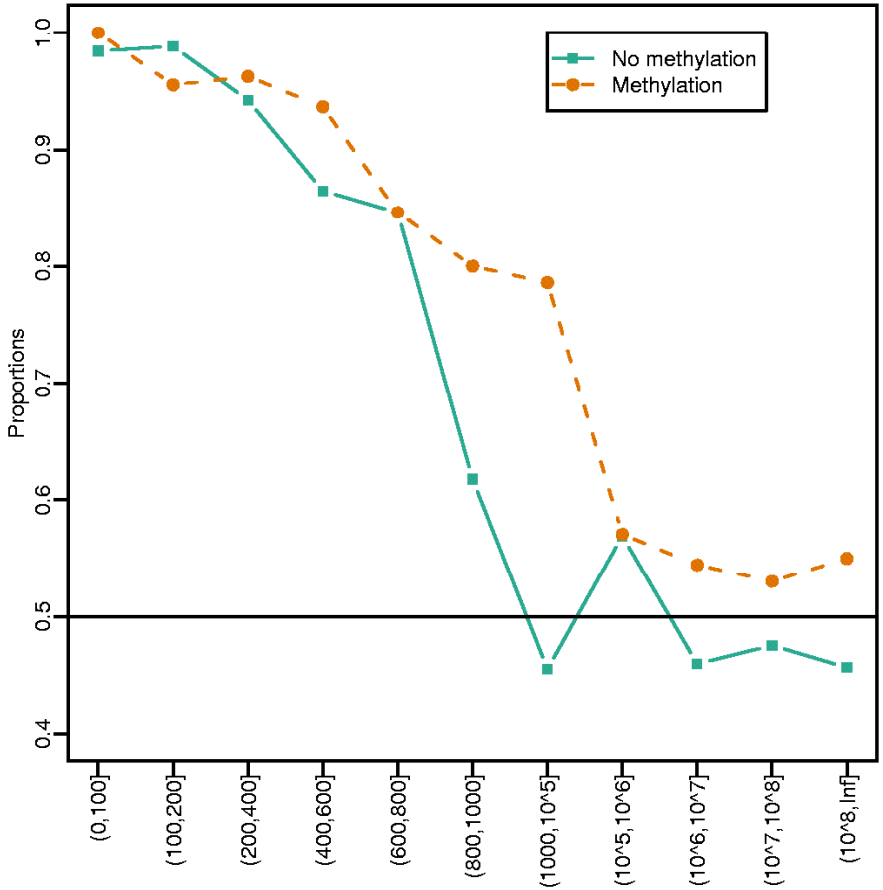
No Methylation

The Problem

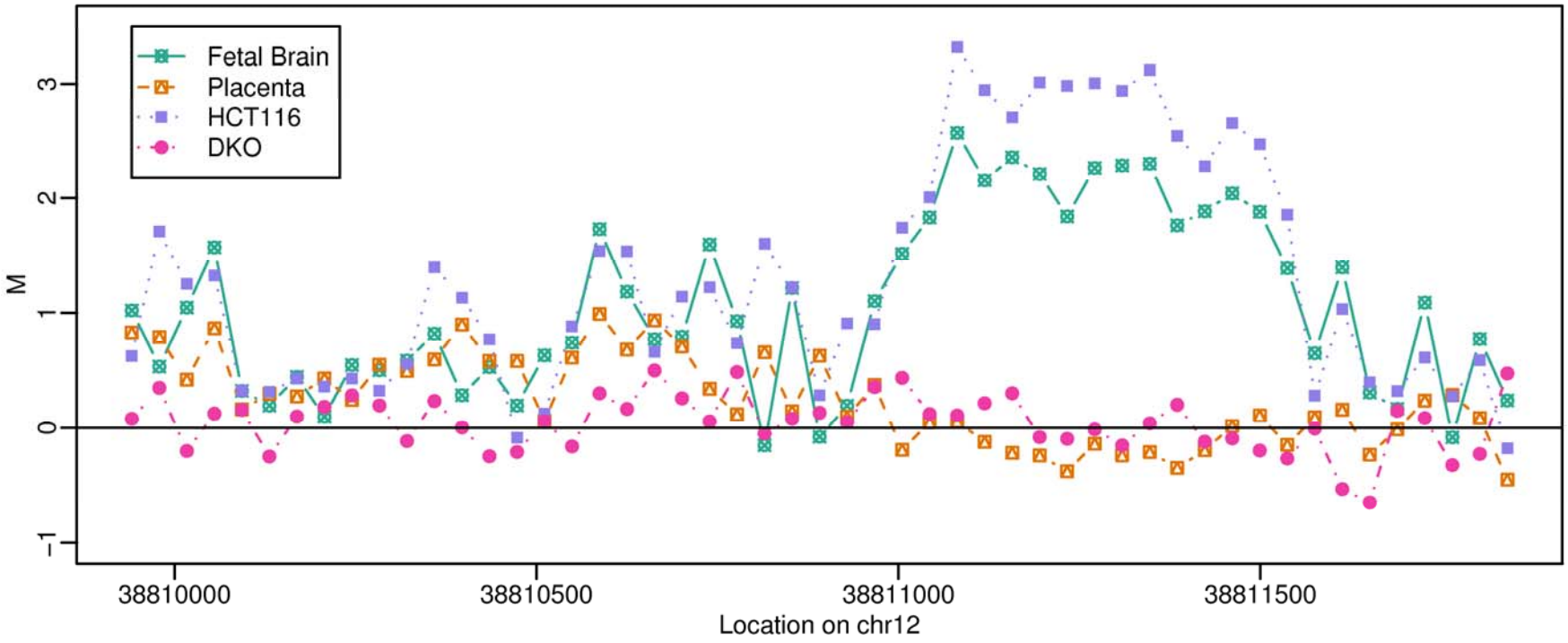


Obsered to expeced = Pr(CG) / { Pr(C) Pr(G) }

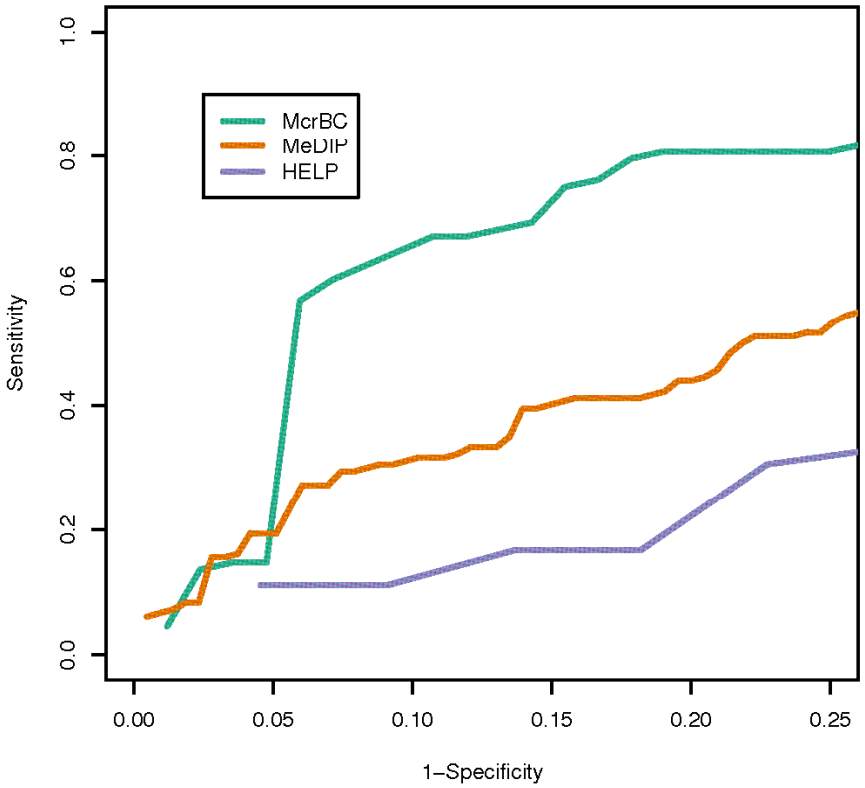
Proportion of neighboring CpG also methylated/not methylated

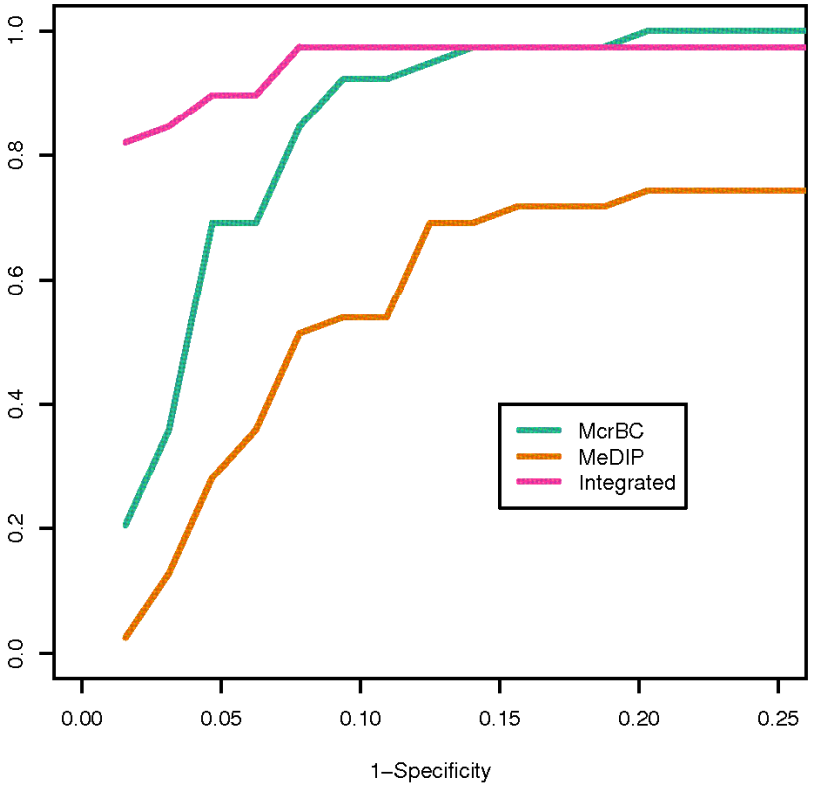


McRBC on Tiling array

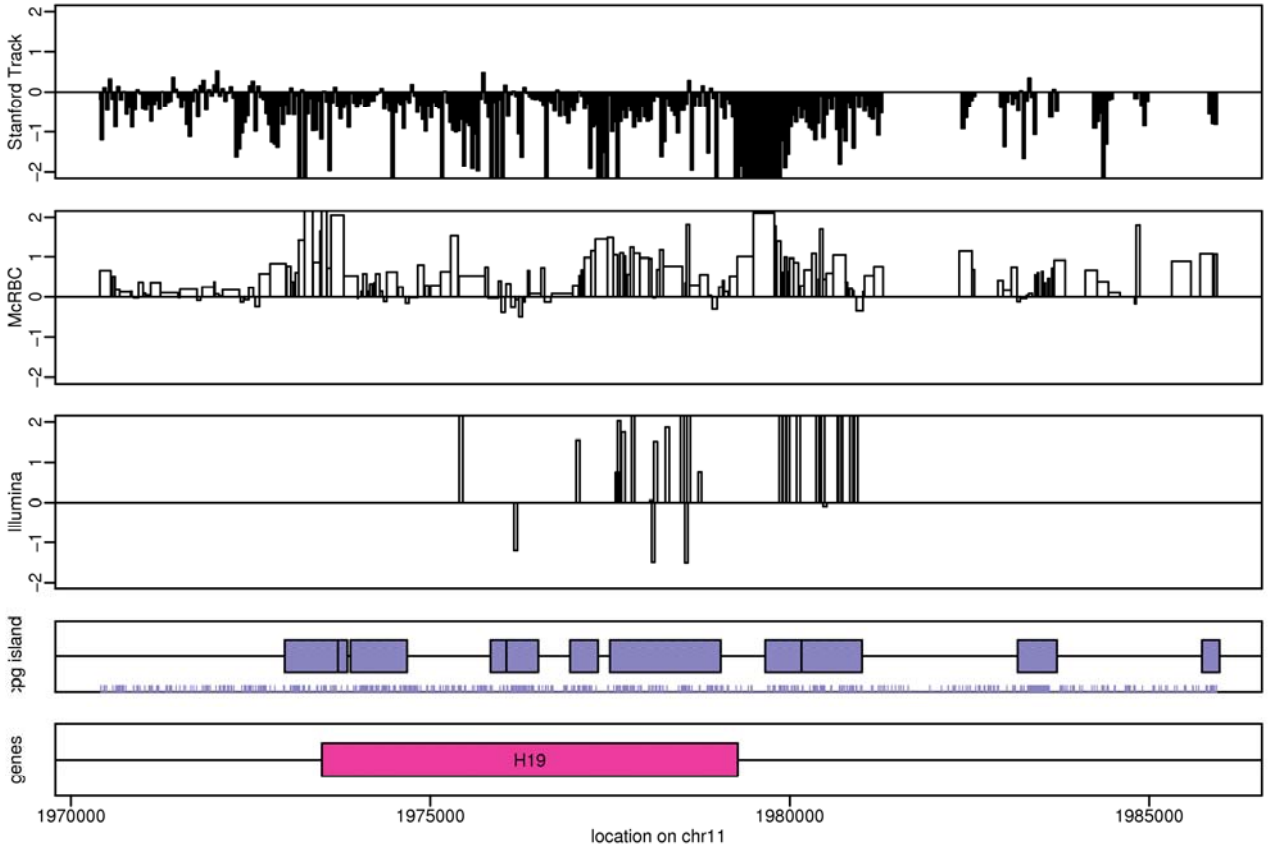


ROC now


McrBC
MeDIP
HELP
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25


McrBC
MeDIP
Integrated
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25

ENCODE Track

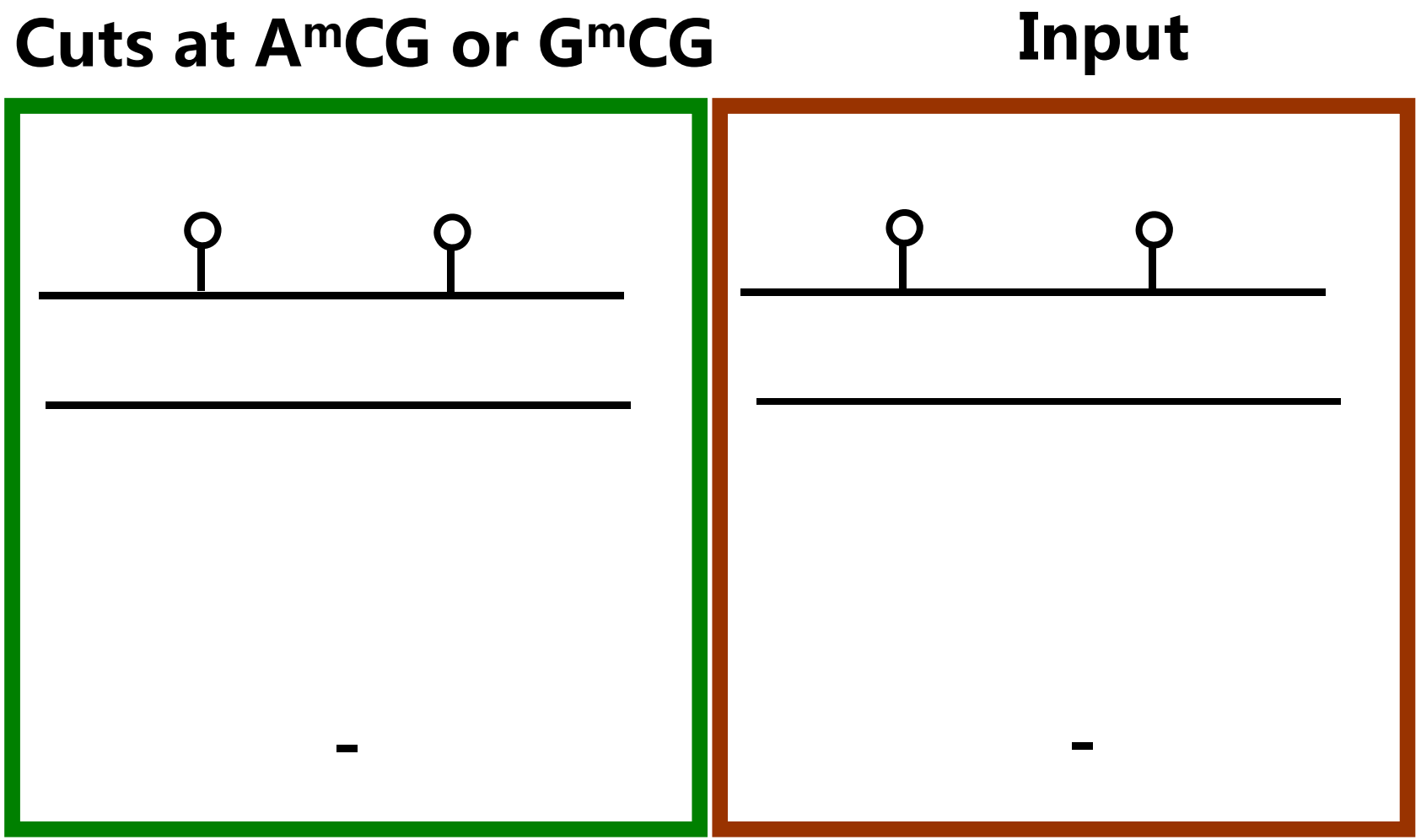


Problems for Statisticians

- **Background Correction + Normalization**
- **Probability Model for Segments**
- **Use these to from null and alternative models… we need power!**
- **Use these to create bump finding algorithms**
- **Adapt to high-throughput sequencing**

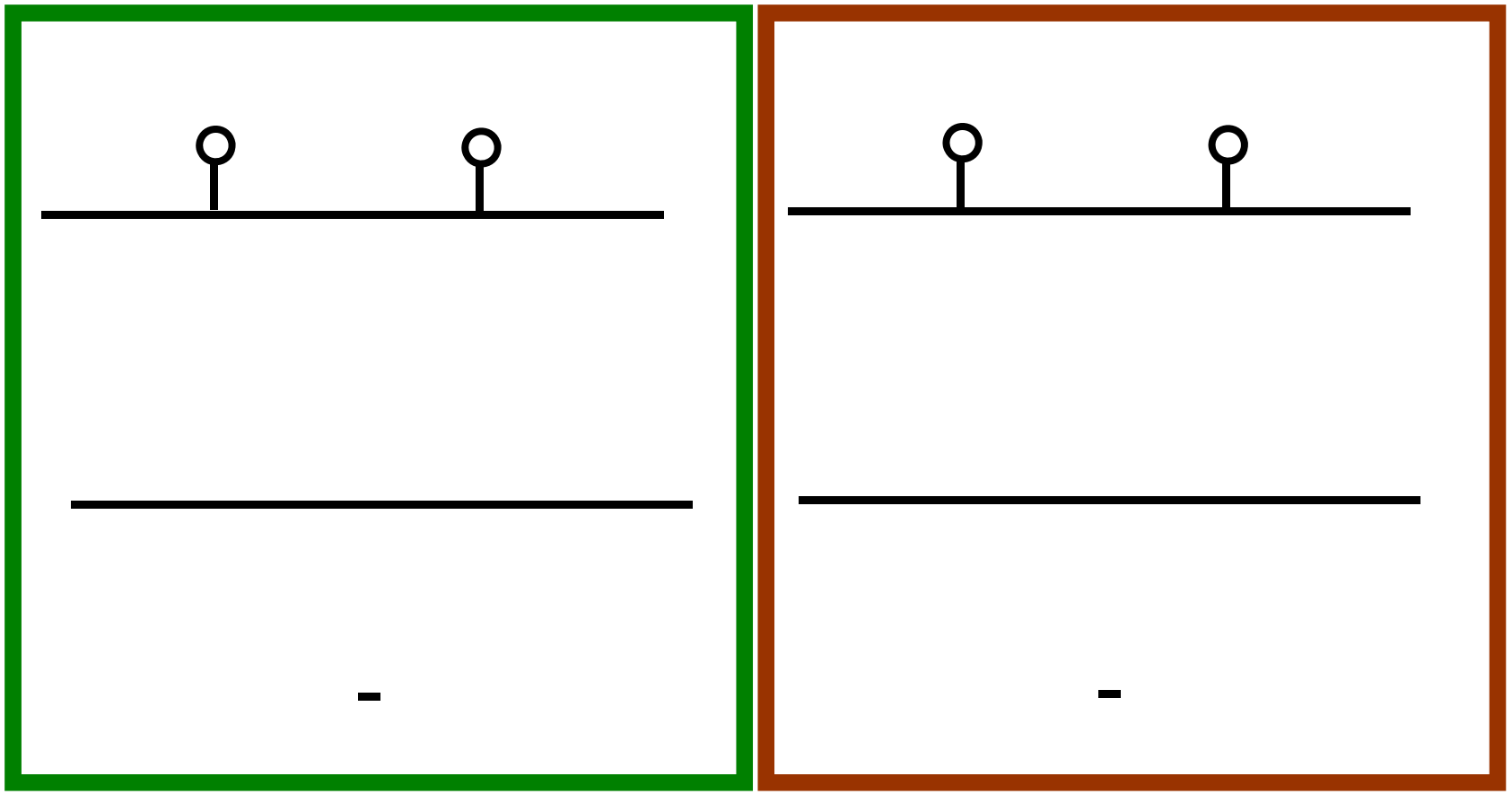
Supplemental Slides

McRBC: One enzyme



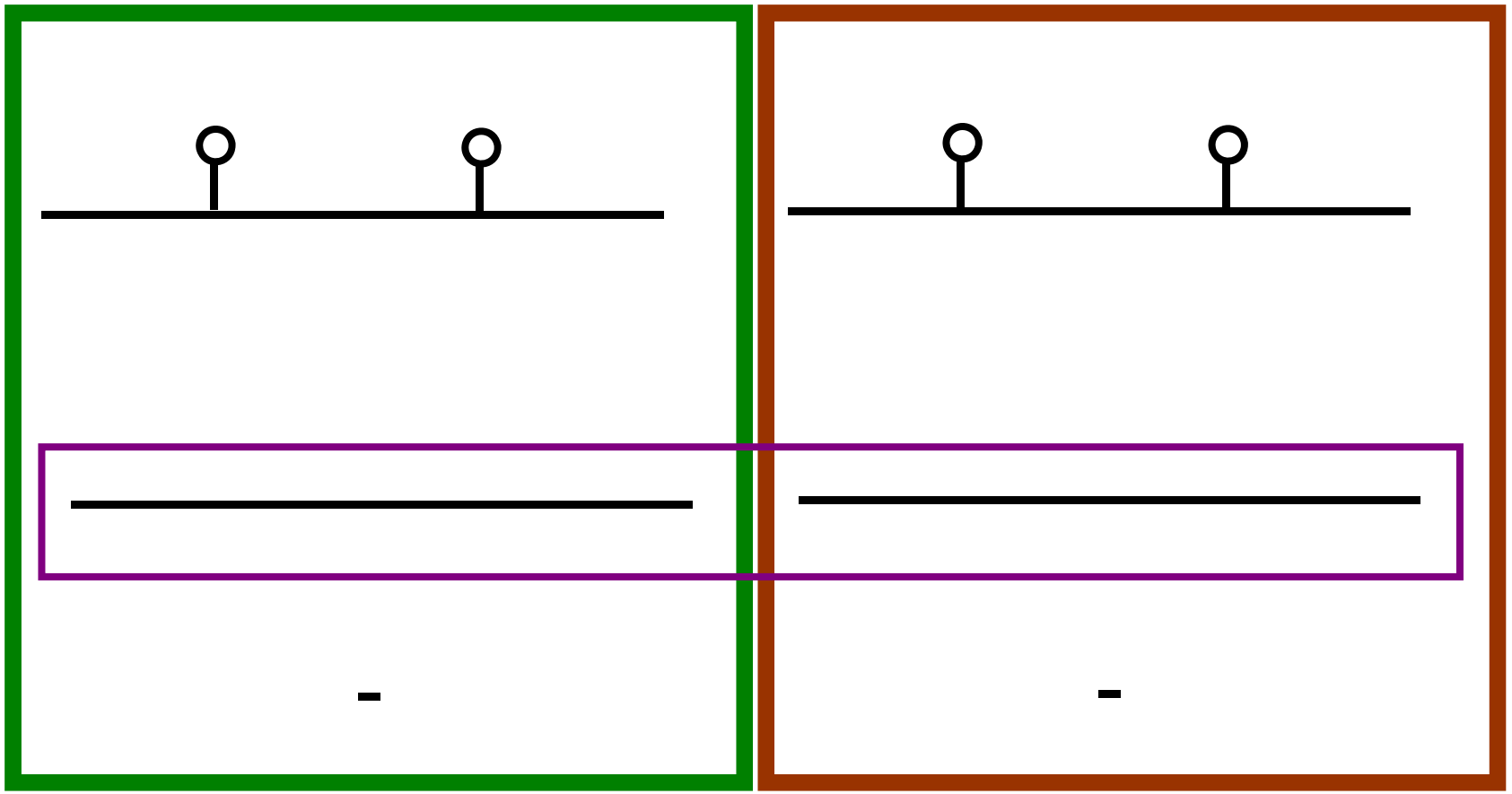
No Methylation

McRBC after Gel



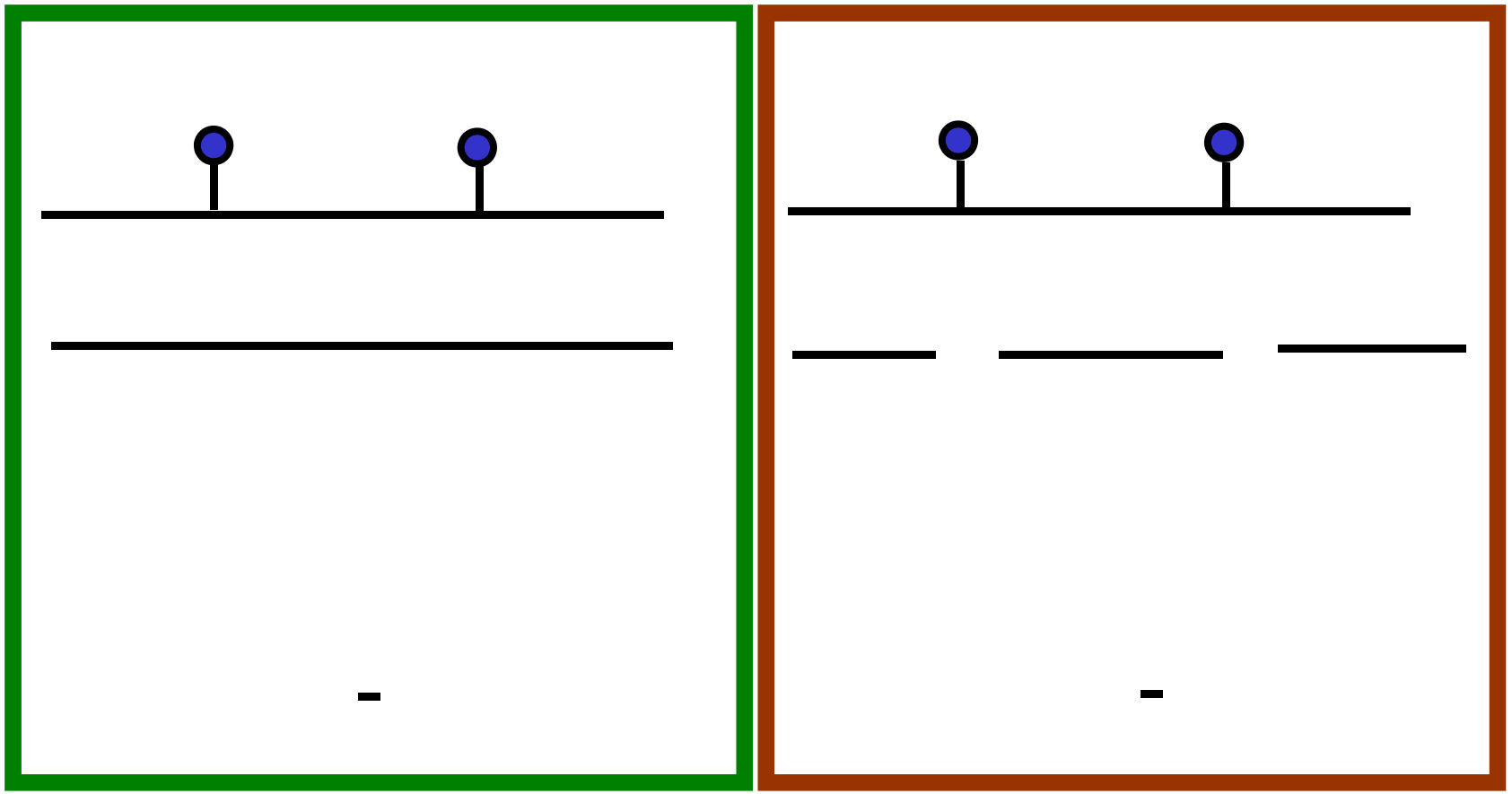
No Methylation

McRBC after Gel



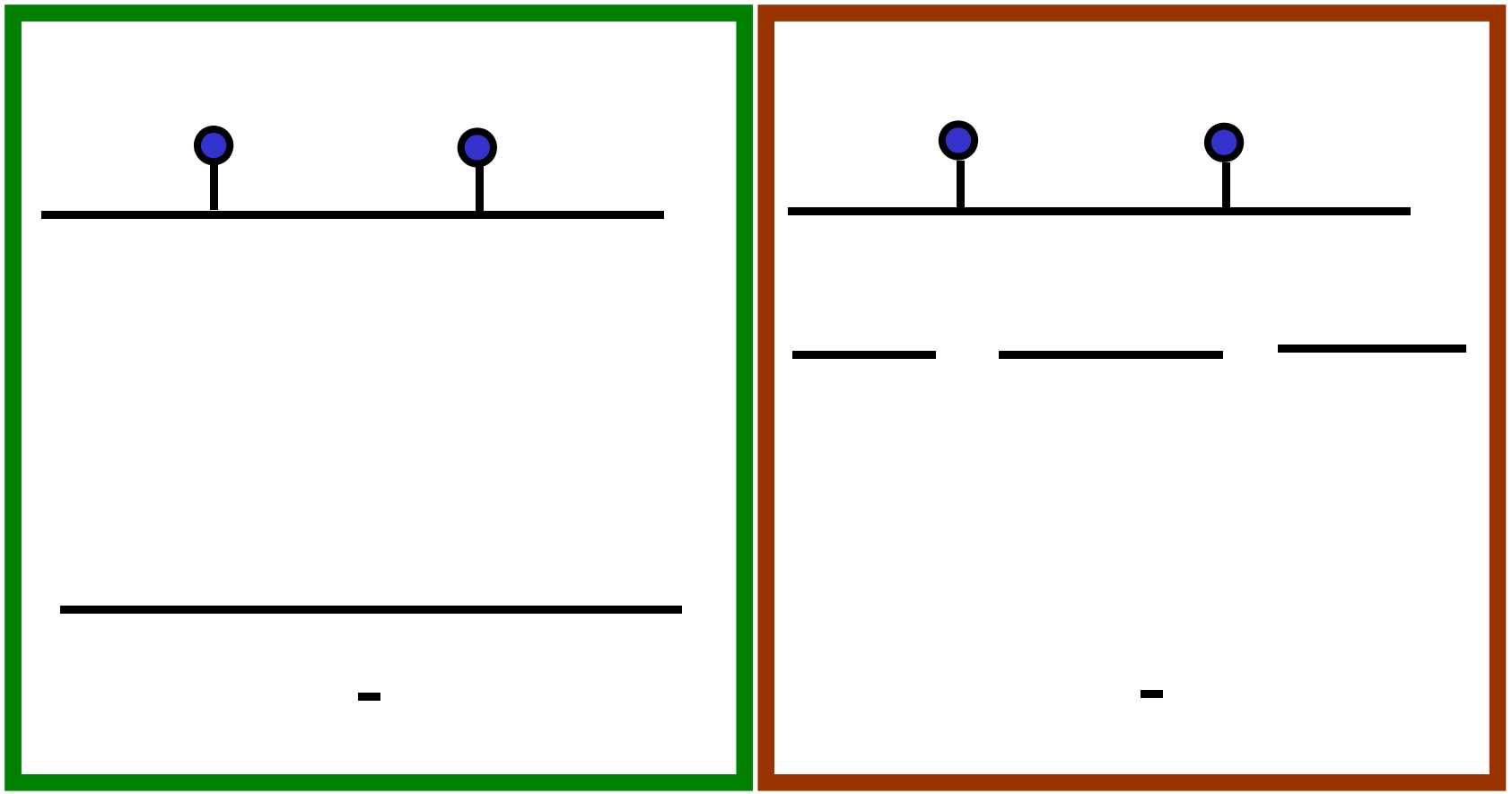
No Methylation

McRBC



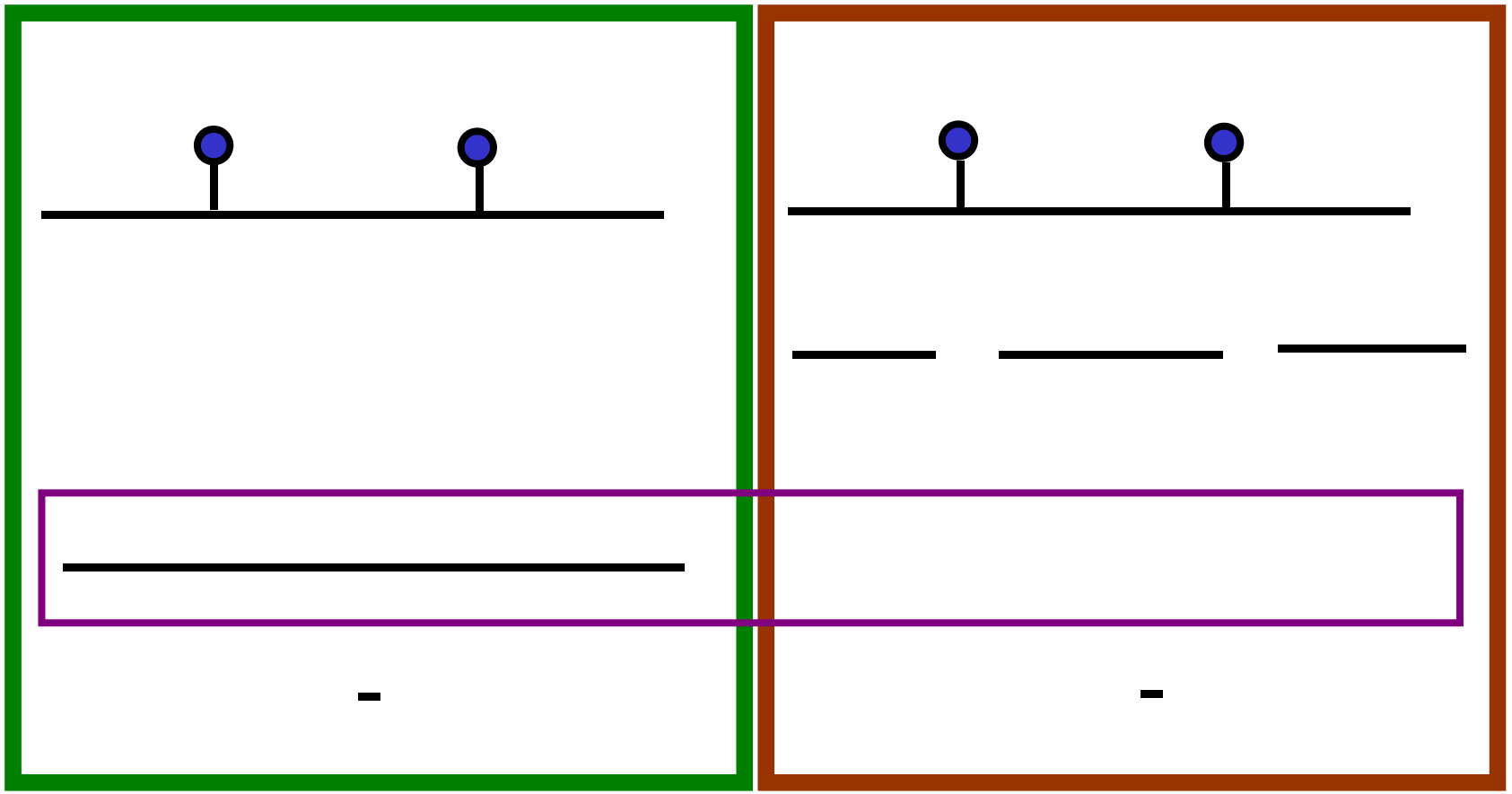
Methylation

McRBC after GEL

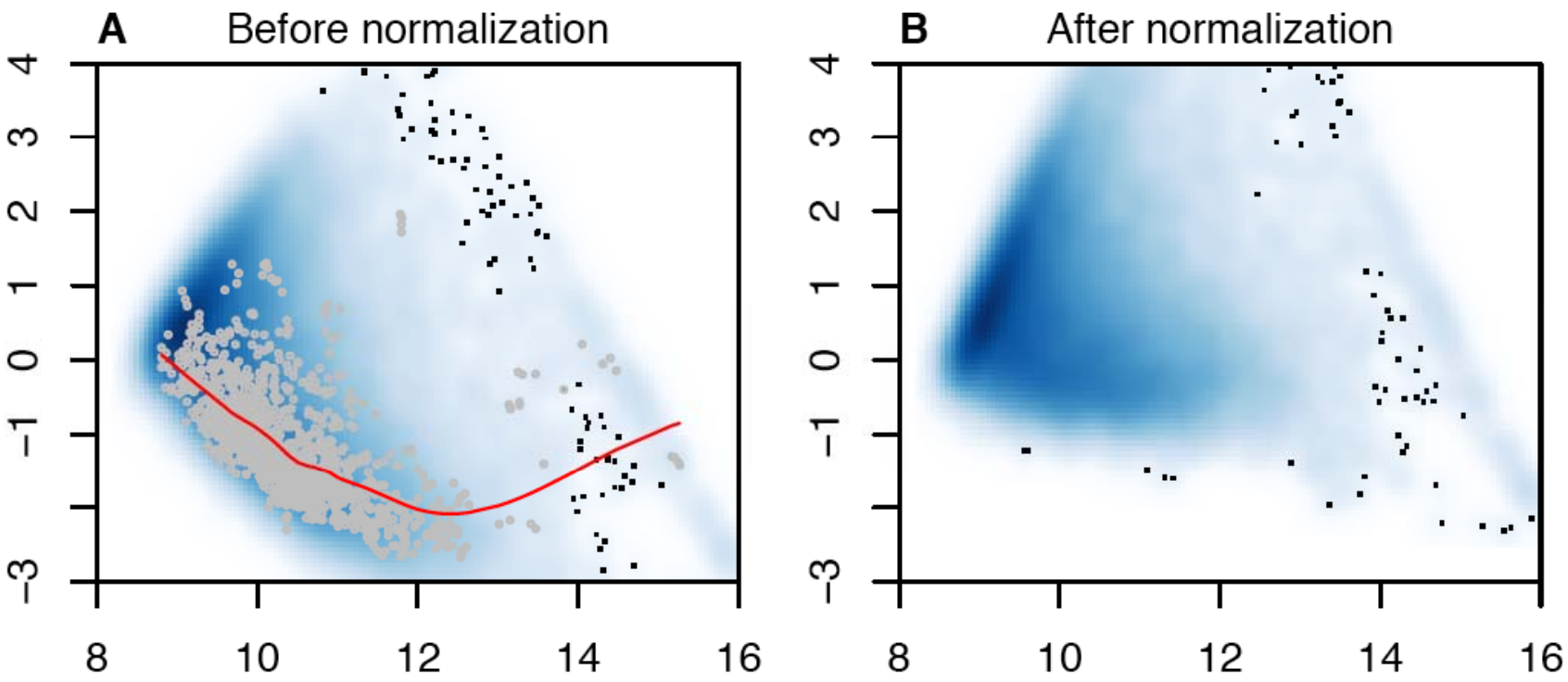


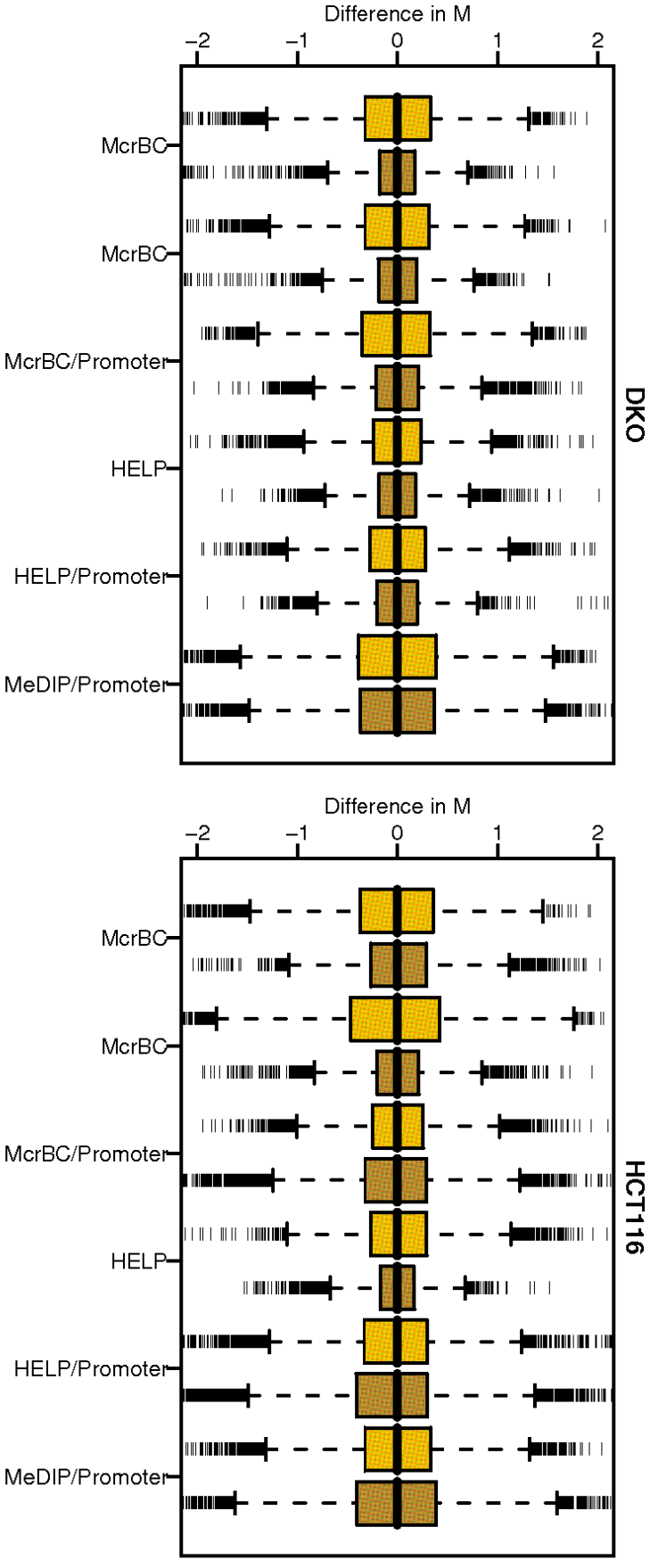
Methylation

McRBC after GEL



Methylation


A
Before normalization
B
After normalization
4
3
2
1
0
−1
−3
8
10
12
14
16


Difference in M
−2
−1
0
1
2
McrBC
McrBC
McrBC/Promoter
HELP
HELP/Promoter
MeDIP/Promoter
DKO
Difference in M
−2
−1
0
1
2
McrBC
McrBC
McrBC/Promoter
HELP
HELP/Promoter
MeDIP/Promoter
HCT116