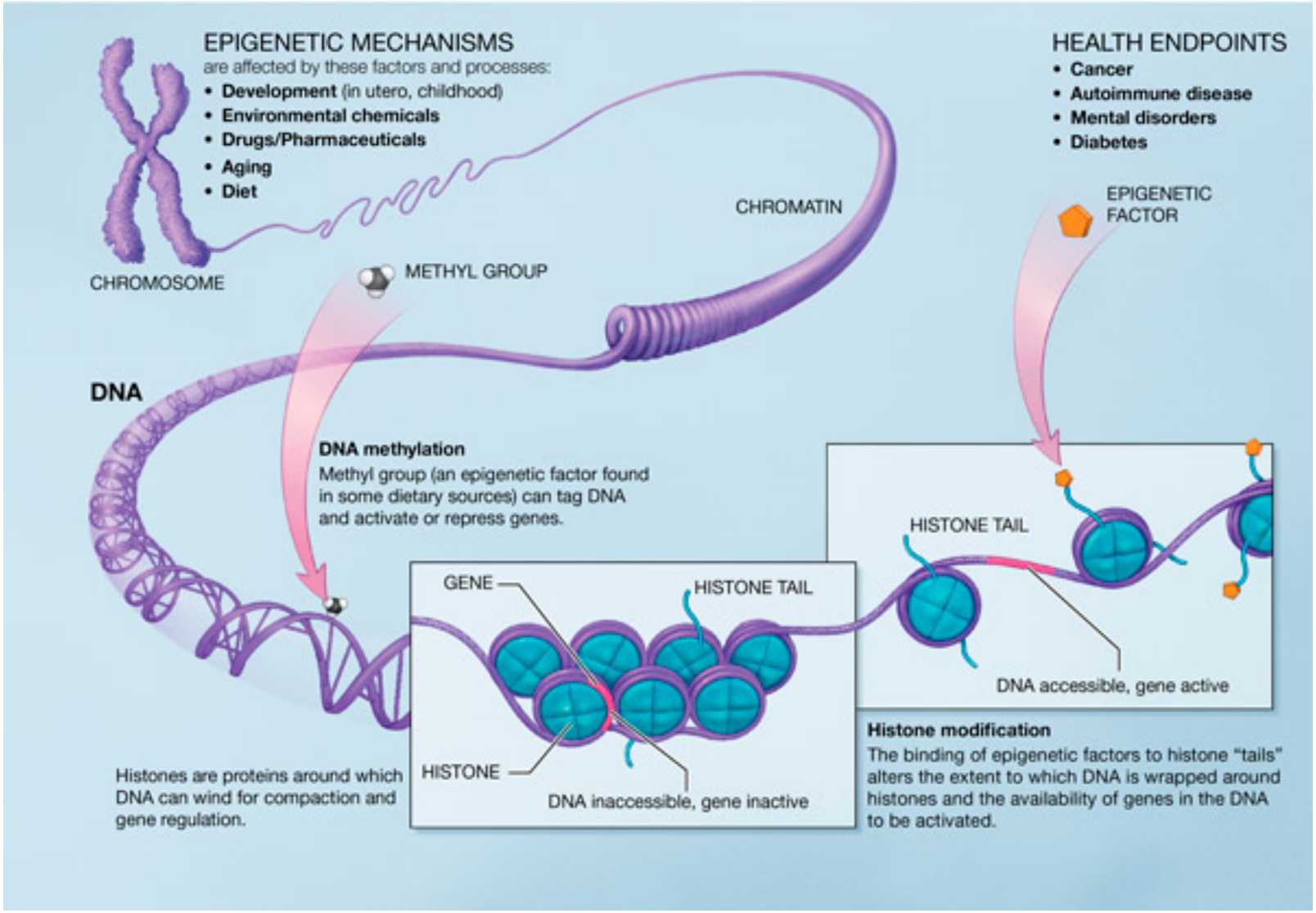
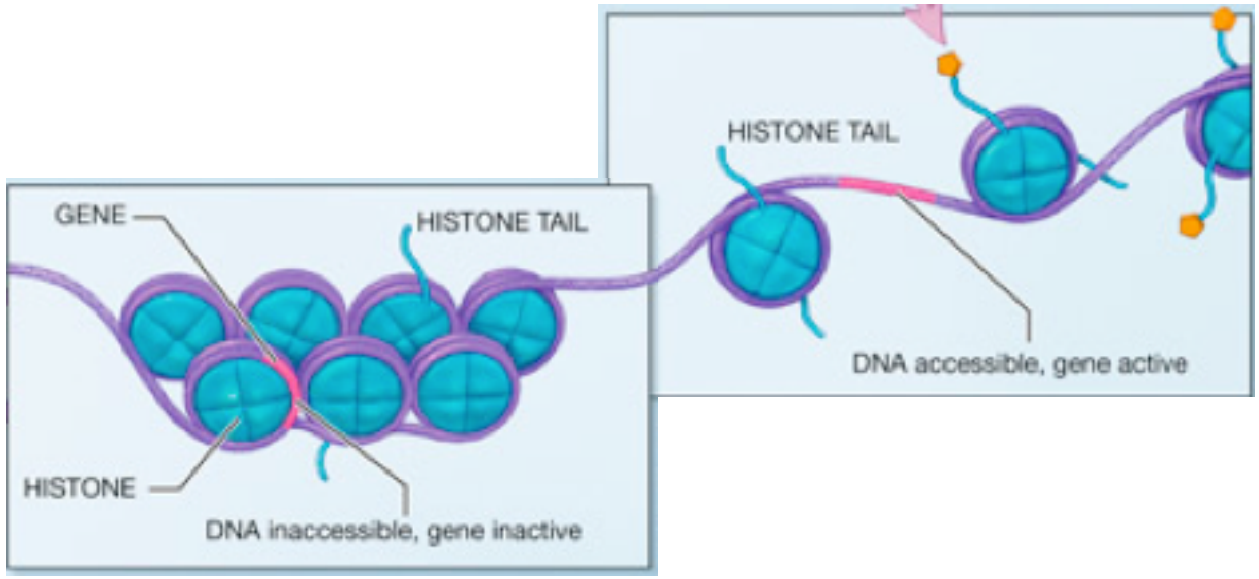
DNA methylation

Epigenetics



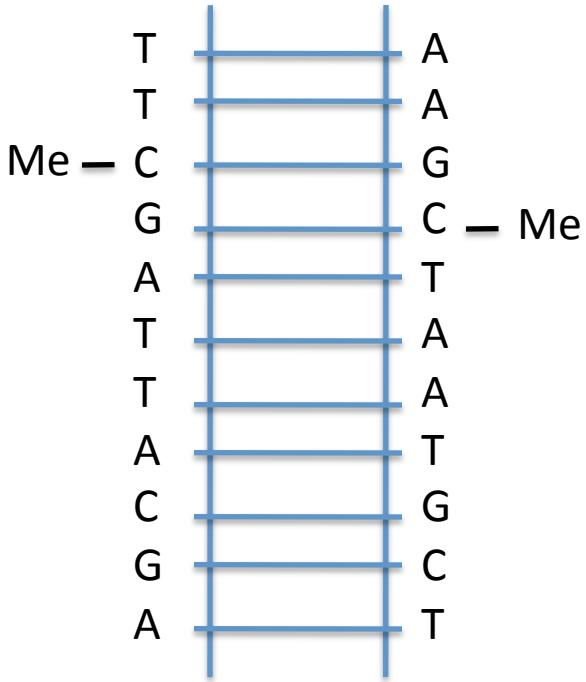
http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg

Epigenetics

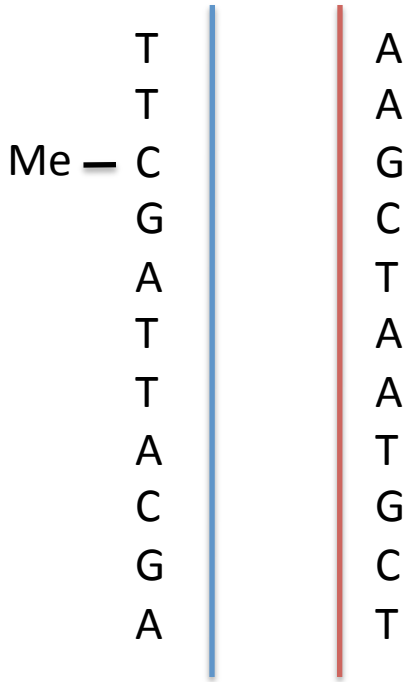


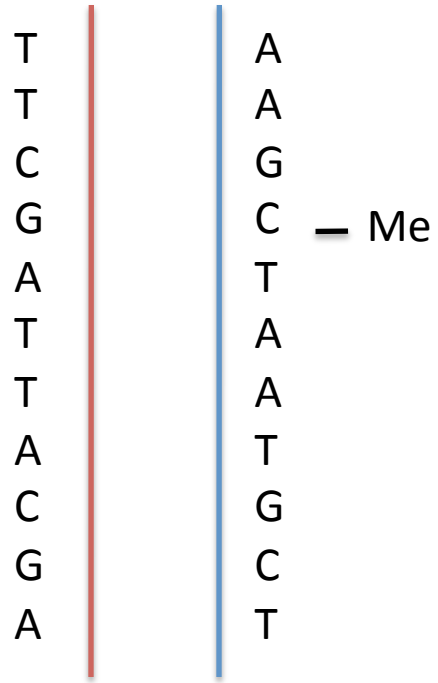
http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg

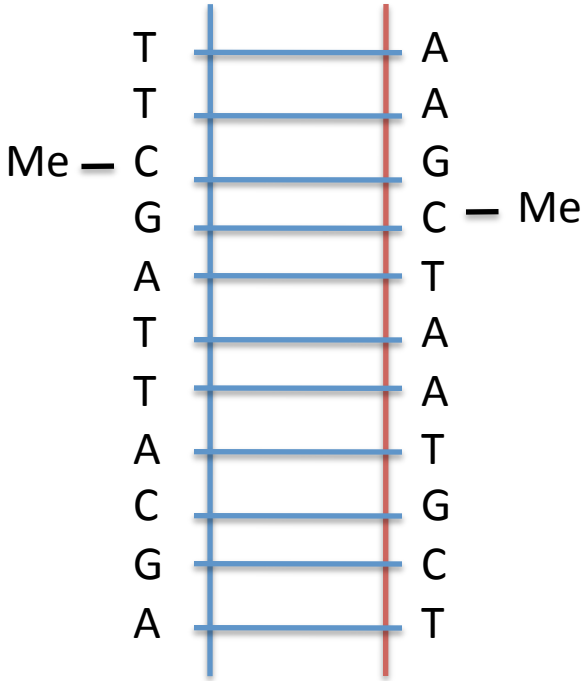
DNA Methylation

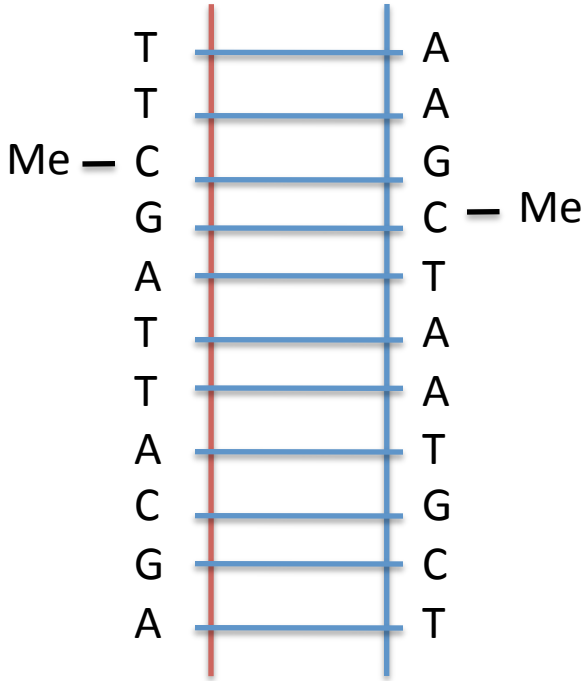



T
T
Me C
G
A
T
T
A
C
G
A
A
A
G
C Me
T
A
A
T
G
C
T

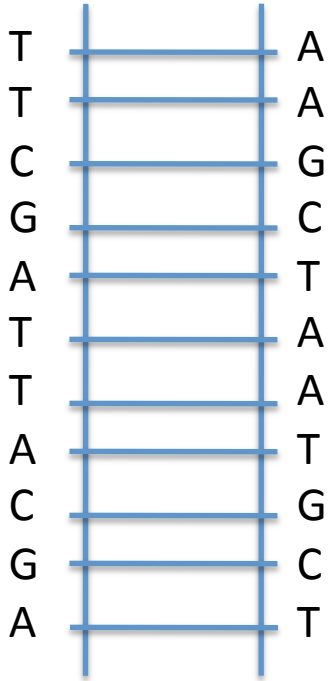

Me
T
T
C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T


T
T
C
G
A
T
T
A
C
G
A
A
A
G
C
Me
T
A
A
T
G
C
T

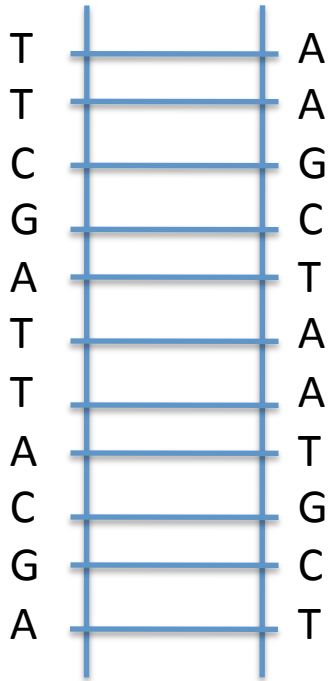

T
T
Me – C
G
A
T
T
A
C
G
A
A
A
G
C – Me
T
A
A
T
G
C
T


T
T
Me – C
G
A
T
T
A
C
G
A
A
A
G
C – Me
T
A
A
T
G
C
T

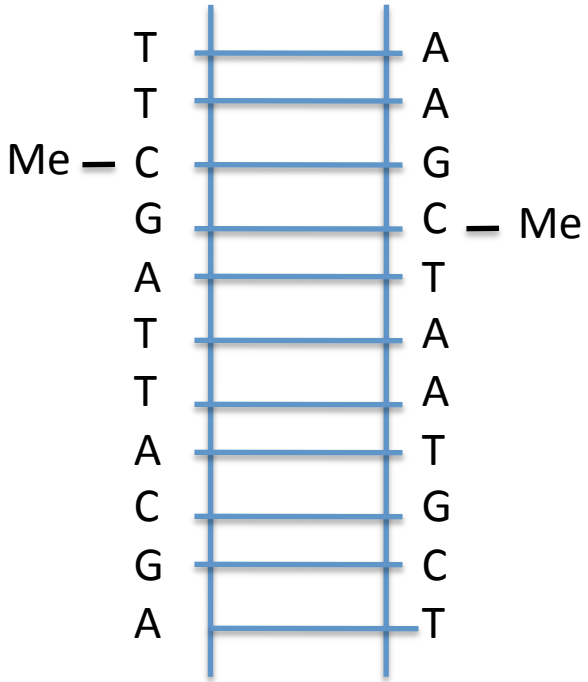
Liver



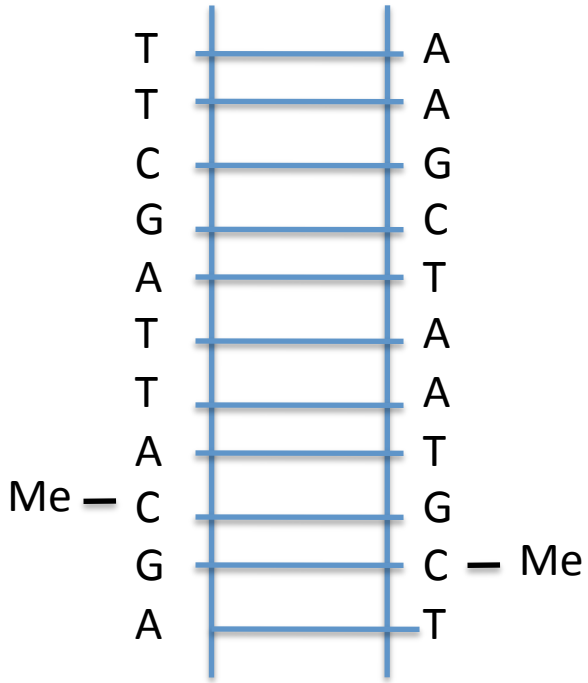
Brain



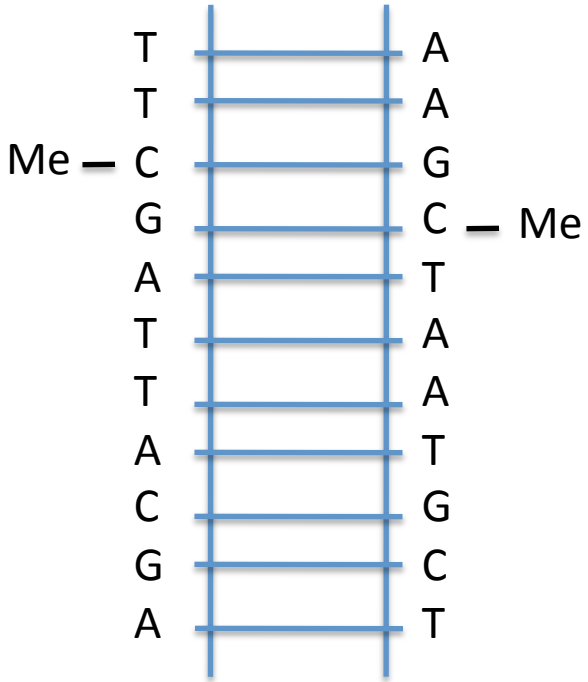
Liver

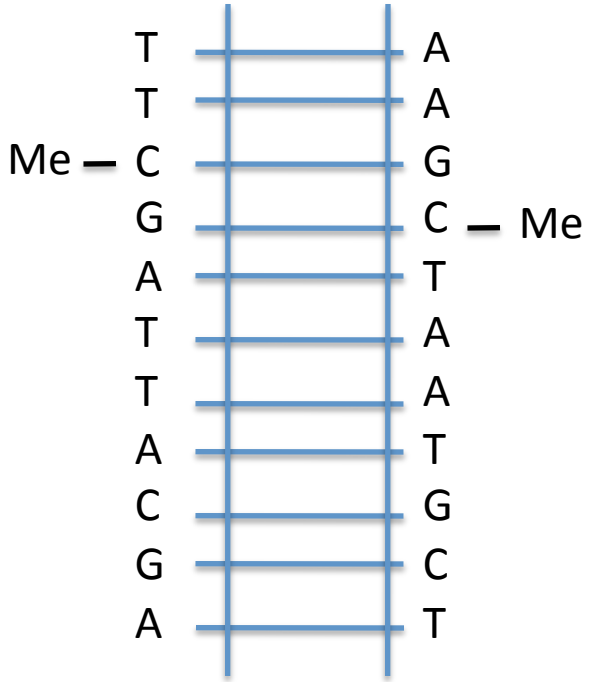

T A
T A
Me C G
G C Me
A T
T A
T A
A T
C G
G C
A T

Brain

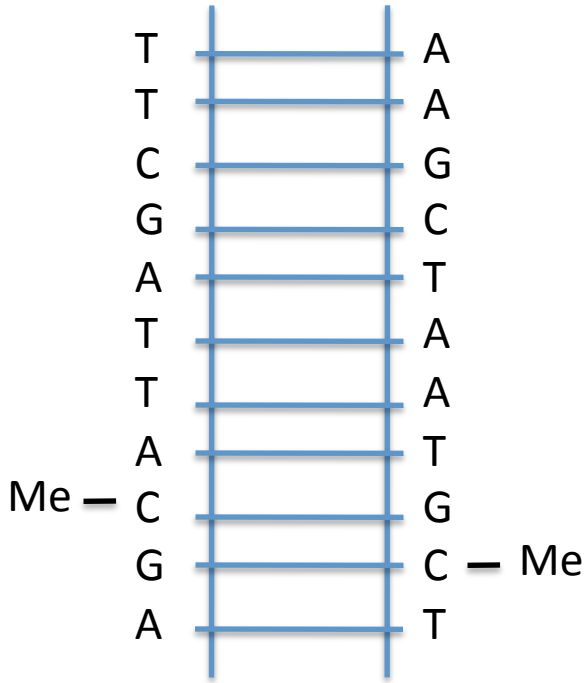

T A
T A
C G
G C
A T
T A
T A
A T
Me C G
G C Me
A T

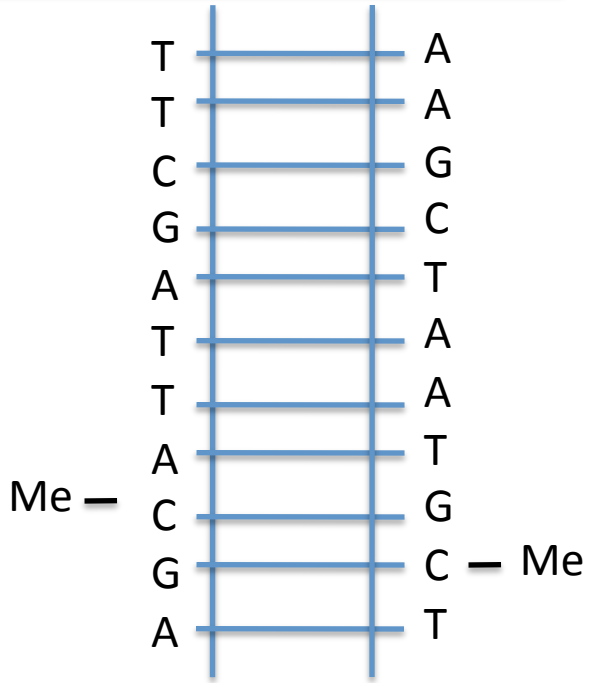
Liver


T
T
Me – C
G
A
T
T
A
C
G
A
A
A
G
C – Me
T
A
T
A
T
G
C
T

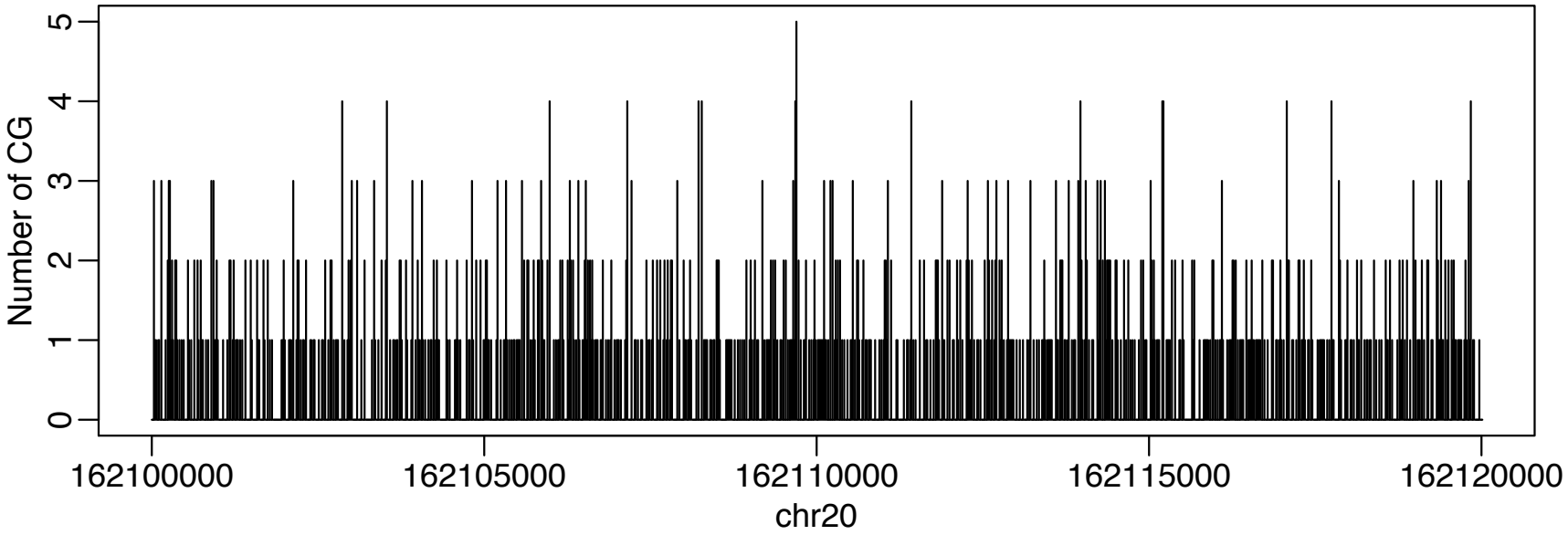

T
T
Me – C
G
A
T
T
A
C
G
A
A
A
G
C – Me
T
A
T
A
T
G
C
T

Brain


T
T
C
G
A
T
T
A
Me – C
G
A
A
A
G
C
T
A
T
G
C – Me
T

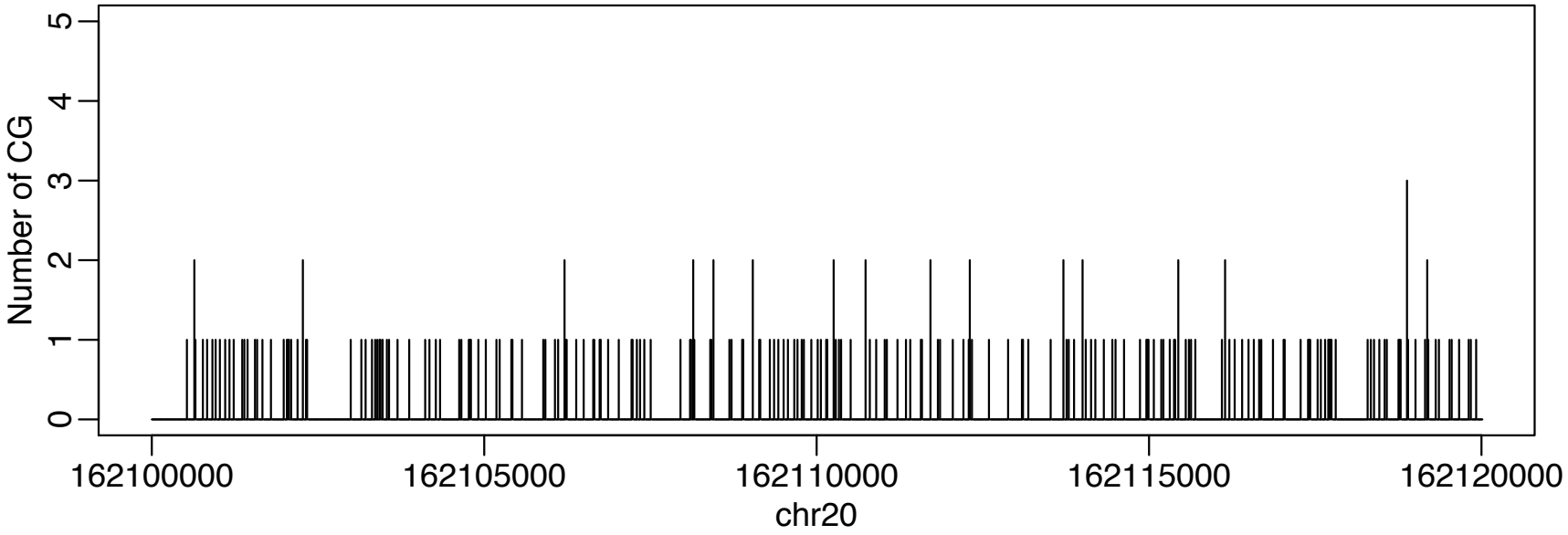

T
T
C
G
A
T
T
A
Me –
C
G
A
A
A
G
C
T
A
T
G
C – Me
T

GC counts on the genome



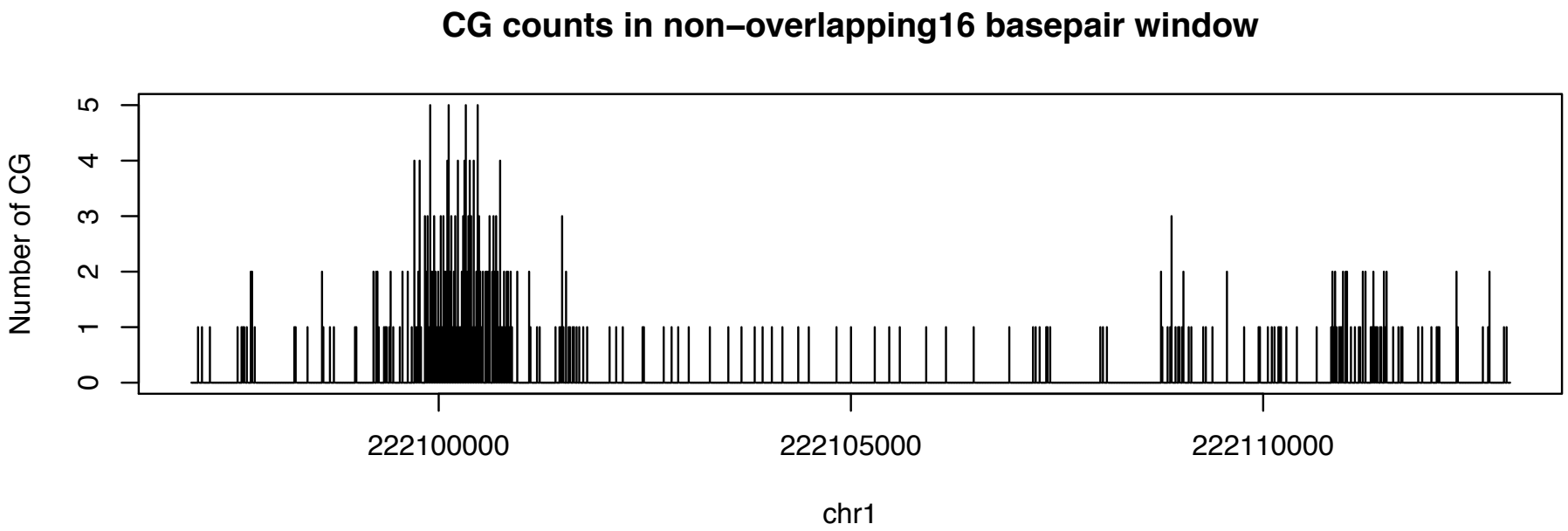
These are counts in 16 basepair bins

CpG are depleted



- These are counts in 16 basepair bins
- We see rate of about 1 in 100

CpG Islands



- But CpGs cluster into *islands* enriched near promoter

Irizarry et al. (2009) Mammalian Genome
Wu et al (2010) Biostatistics,
New illumina CpG array will use our CGI

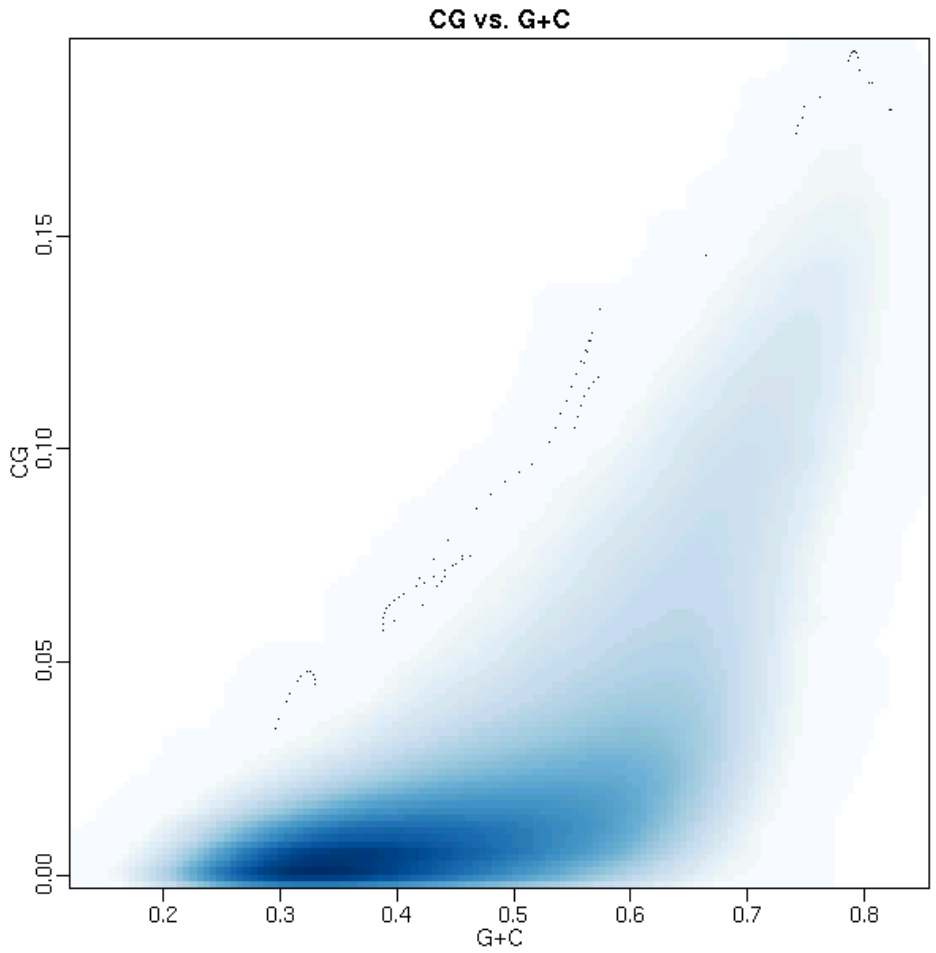
Gardiner-Garden and Frommer CpG Island definition

- N > 200
- GC-content > 50%
- obs/exp > 0.6
- Lists contain 20,000 CGI

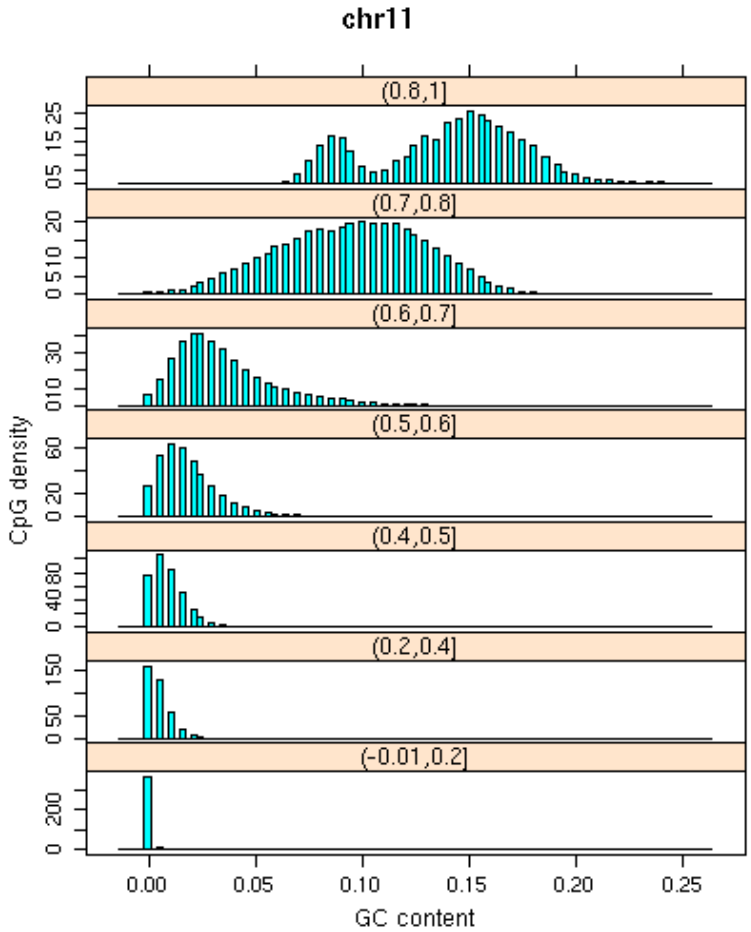
HMM based definition

- Problems:
  - leaves out many clusters
  - Not applicable to other species

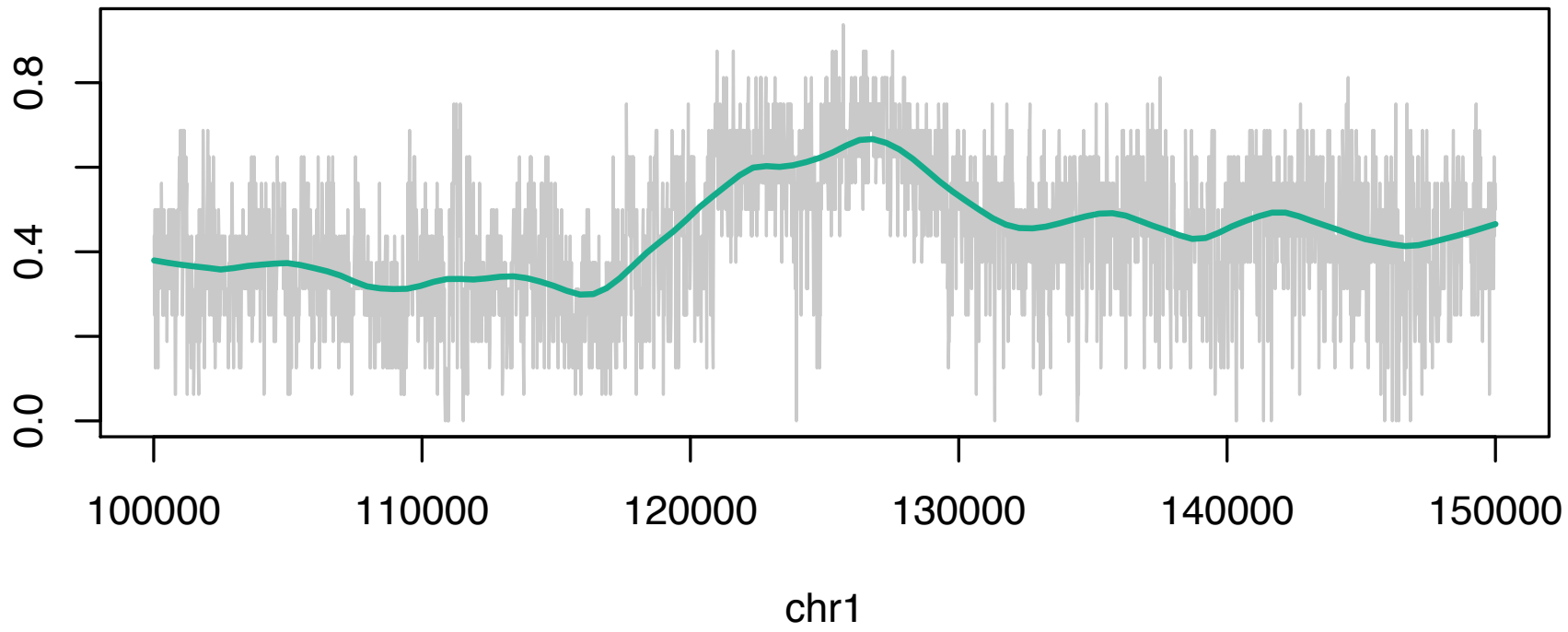
Whole genome view…



Why observed/expected and not counts?



GC content varies



Hidden Markov Model Approach

- Assume that GC content is smooth.
- Estimate and assume known: $p_C(t)$ and $p_G(t)$
- Assume probability of CpG is $\alpha_i\, p_C(t)p_G(t)$ for two states i = 0, 1.
- To avoid correlation problem, assume counts in bins of size L is Poisson with rate is $\alpha_i\, p_C(t)p_G(t)\, L$
- We use L=16
- Use EM to estimate α_0 and α_1 from data and fit HMM

Irizarry et al. (2009) Mammalian Genome, Wu et al (2010) Biostatistics,
New illumina CpG array will use our CGI

Conventional wisdom in 2004

- **Hyper**methylated CpG islands silence tumor suppressor genes

- Cancer cells are globally **hypo**methyated

High throughput measurement permitted us to observe the entire genome:

Irizarry et al. (2008) Genome Research
Aryee el al. (2010) Biostatistics

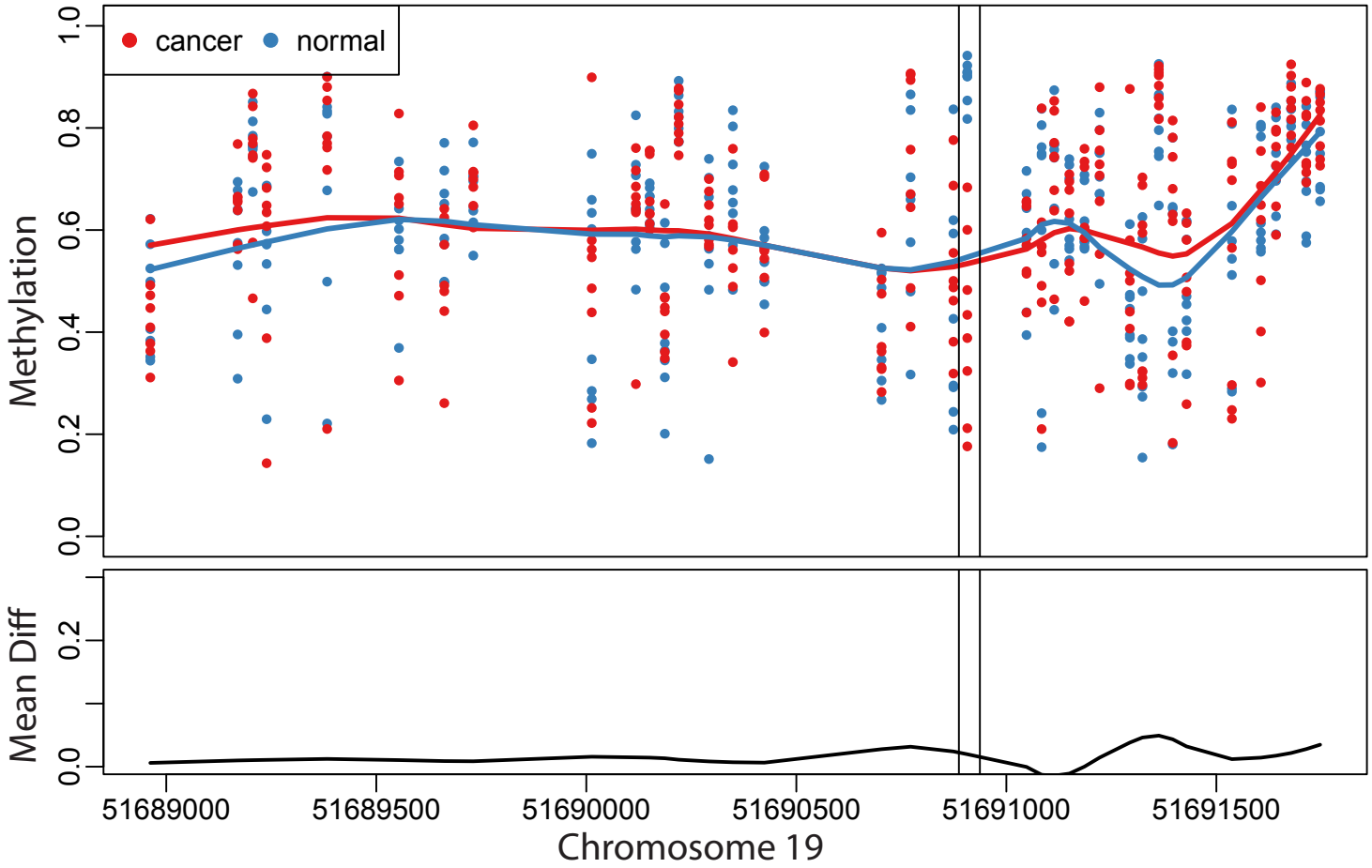
Finding differentially methyalted regions (DMRs)

Irizarry et al. (2008) Genome Research

Aryee el al. (2010) Biostatistics

Jaffe et al (2012) IJE

Genomic traceplot



Microarray data after much preprocessing

General Model

Baseline methylation level

Effect at j-th position

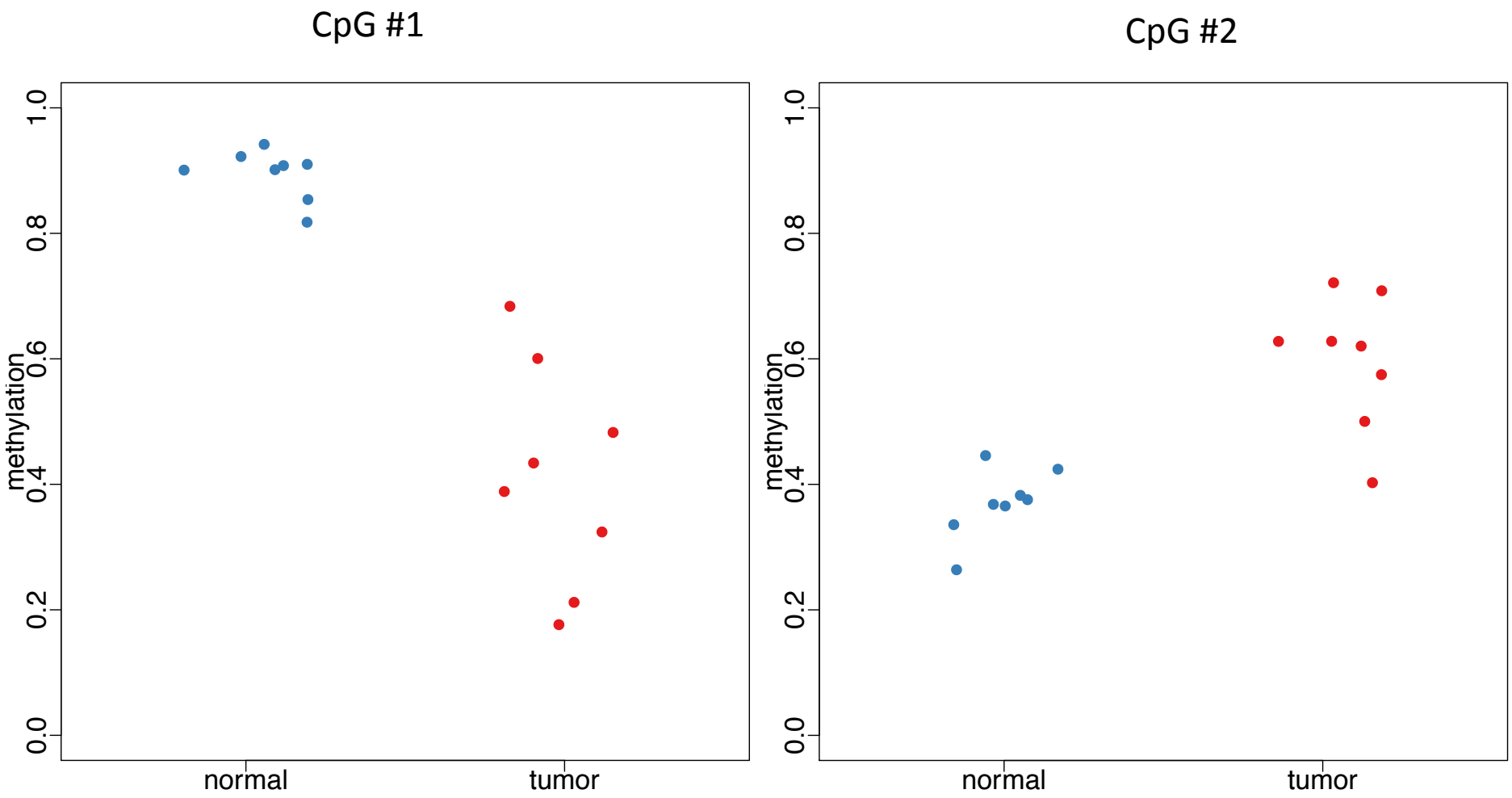
Measurement error

$$Y_{ij} = \beta_0(l_j) + X_i\beta_1(l_j) + \varepsilon_{ij}$$

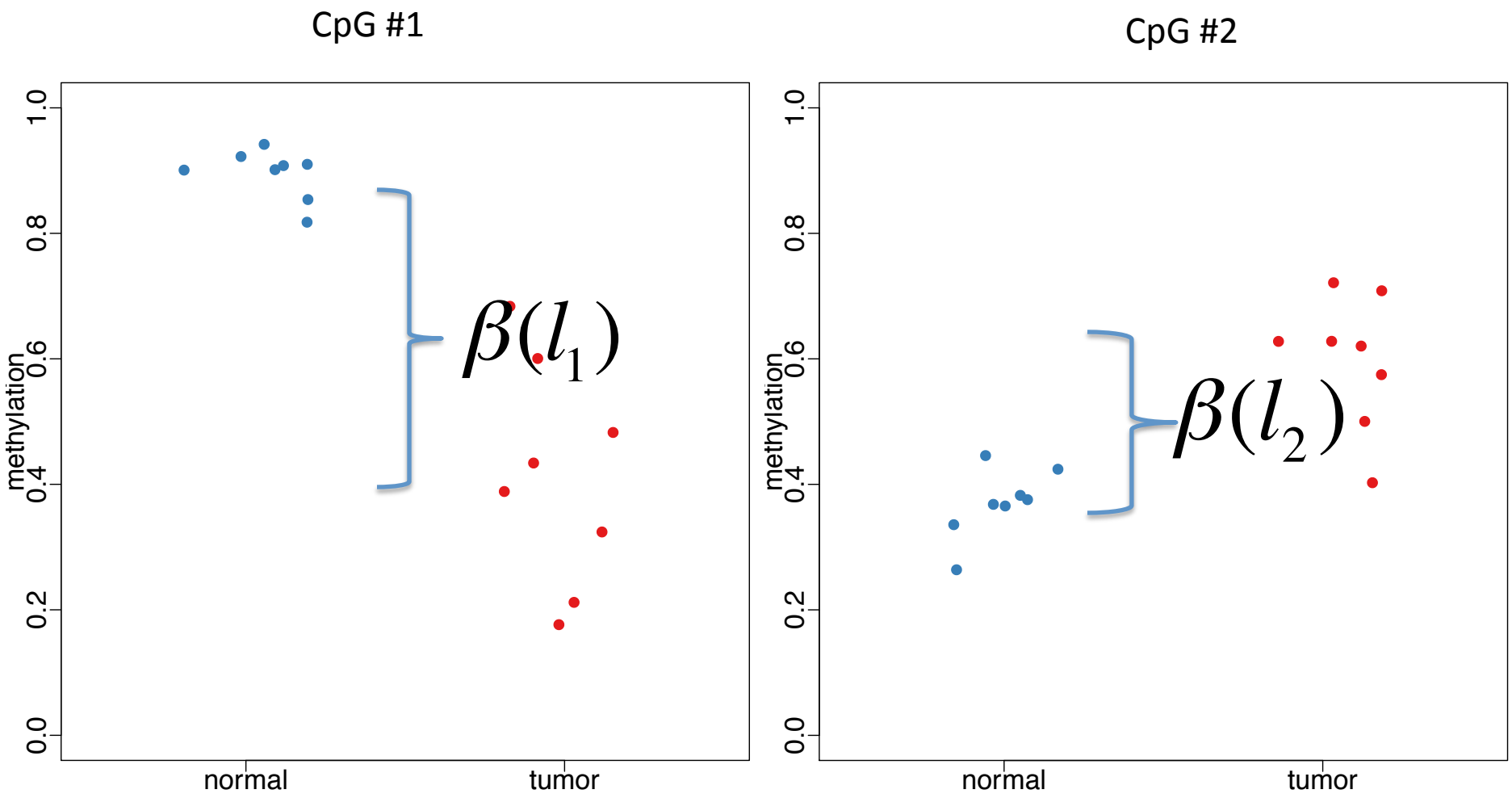
Observed Data

Outcome of interest

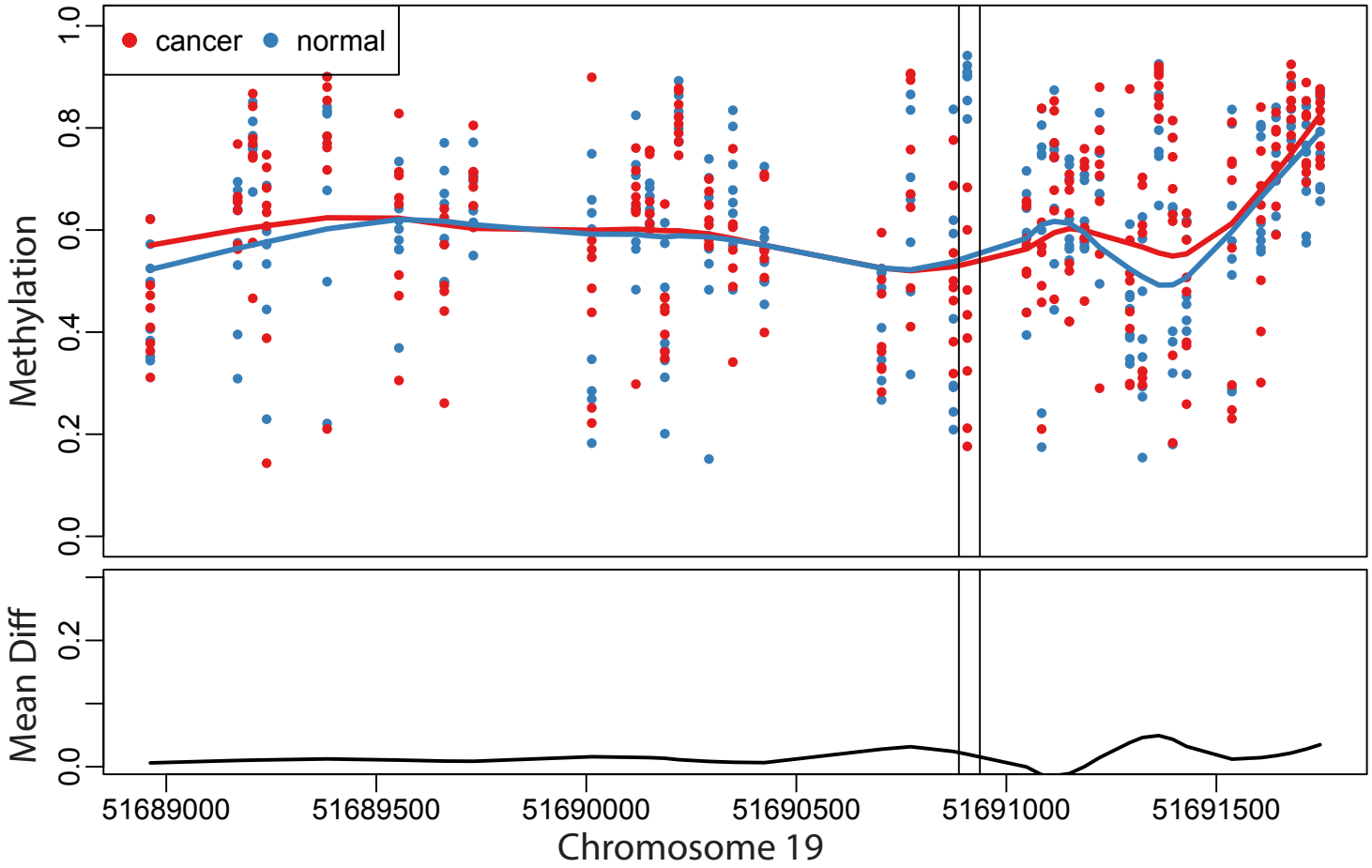
Do we trust single measurements?



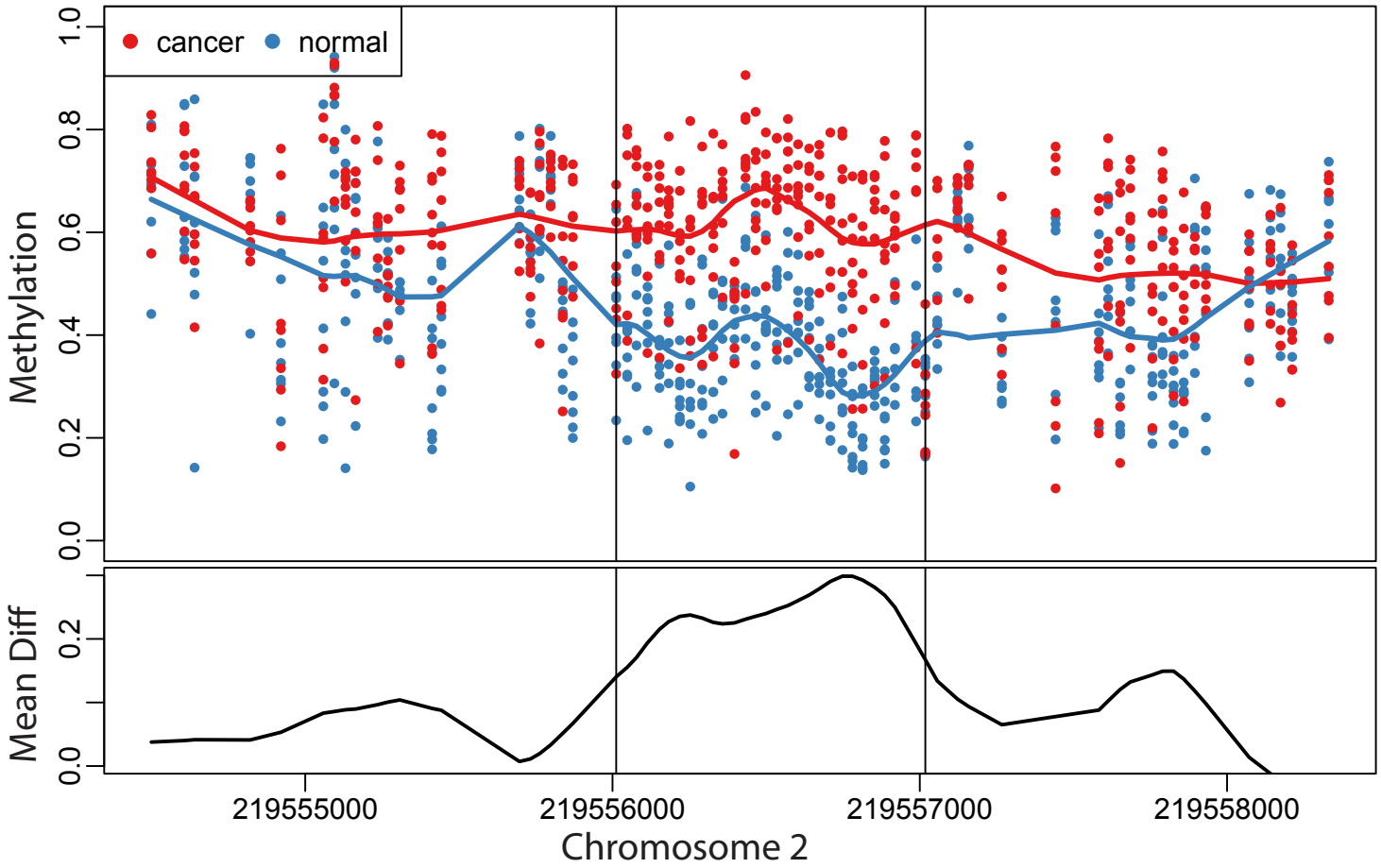
Do we trust single measurements?
Note X is 1 (cancer) or 0 (normal)



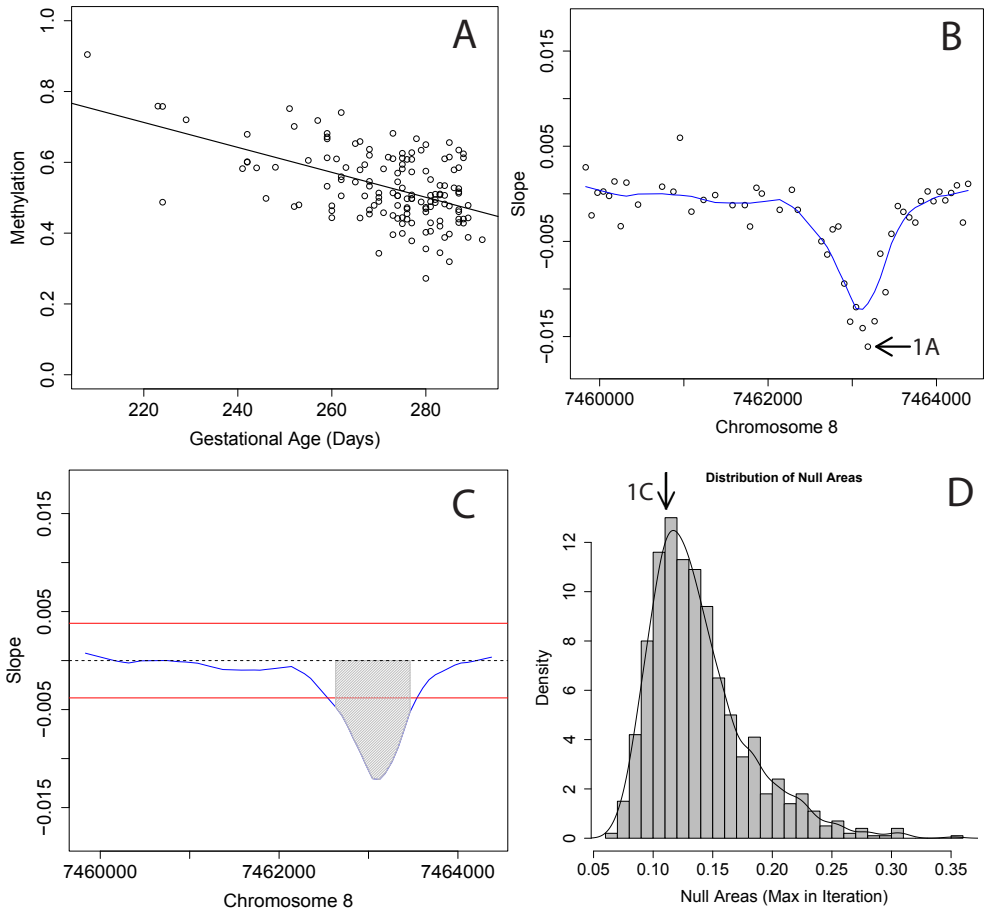
CpG #1


cancer
normal
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
0.0
0.2
51689000
51689500
51690000
51690500
51691000
51691500
Chromosome 19

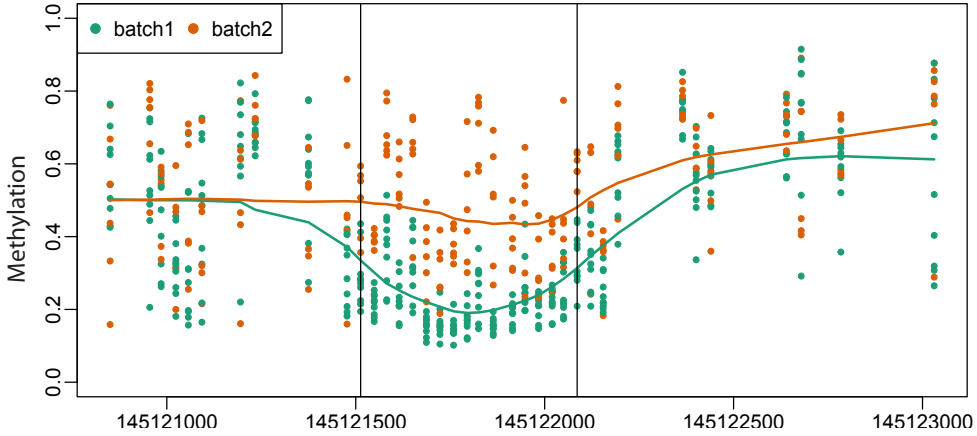
CpG #2


cancer
normal
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
0.0
0.2
219555000
219556000
219557000
219558000
Chromosome 2

Current general approach



Beware of batch effects



OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

There is hope

NATURE REVIEWS | GENETICS

OPINION

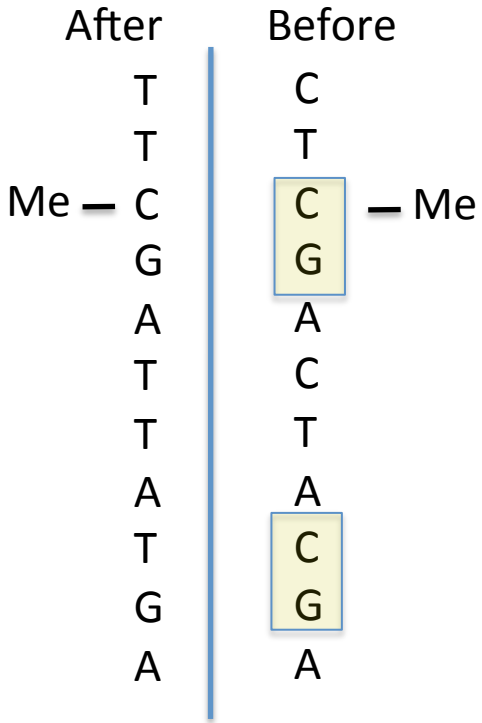
Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Next generation sequencing

Hansen et al. (2011) Nature Genetics

Bisulfite Treatment



Whole Genome Bisulfite Sequencing

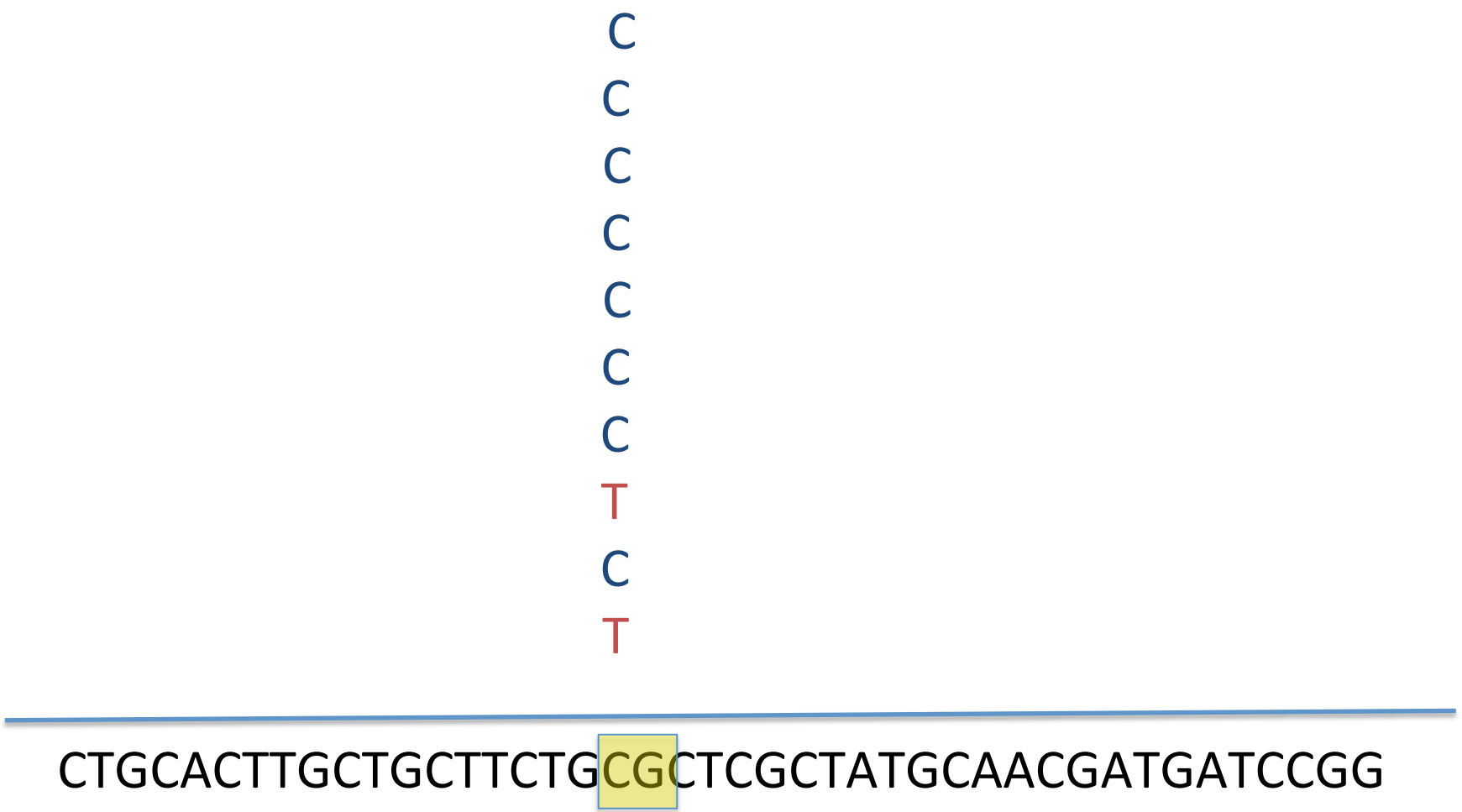
Whole Genome Bisulfilte Sequencing

Count Cs and Ts at CpG location

CTTGCTGCTTCTGCGCTCGCTATGCAACGATGAT
CTGCTTCTGCGCTCGCTATGCAACGATGATCCGGCT
TTGCTGCTTCTGCGCTCGCTATGCAACGATGATCCGGCTGC
ACTTGCTGCTTCTGCGCTCGCTATGCAACGATGA
TTGCTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCCG
TGCTGCTTCTGCGCTCGCTATGCAACGATGATC
CTGCTTCTGCGCTTGCTATGCAACGATGATCCG
TGCTGCTTCTGCGCTCGCTATGCAACGATGATC
TTGCTGCTTCTGCGCTTGCTATGCAACGATGATCCG

CTGCACTTGCTGCTTCTGCGCTCGCTATGCAACGATGATCCGG

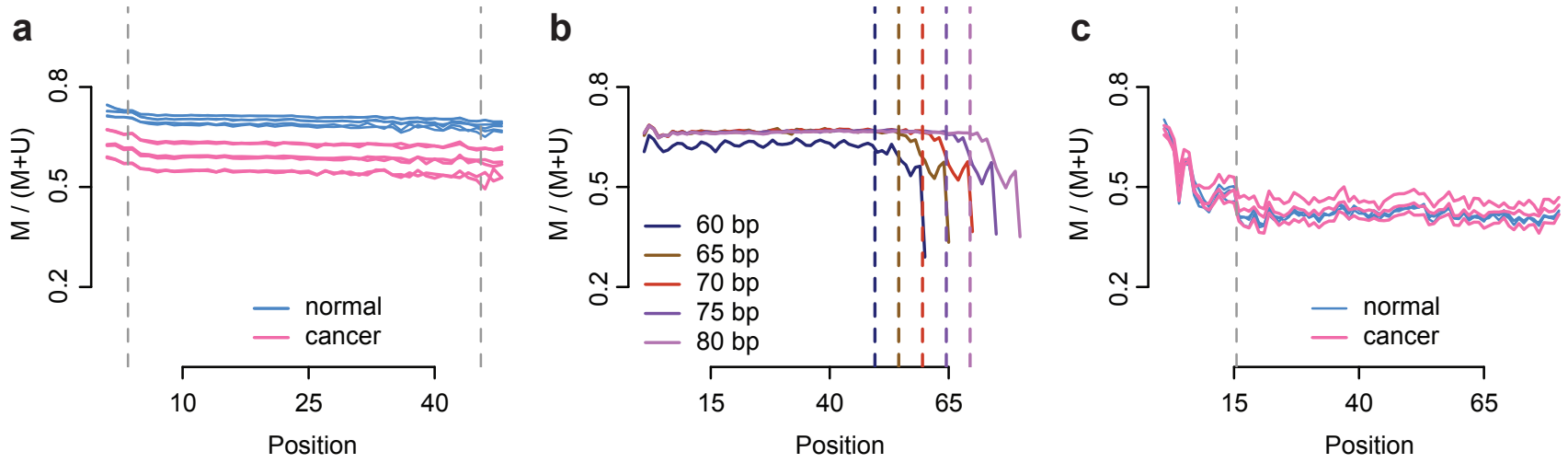
Quantitative Measurement: 80%



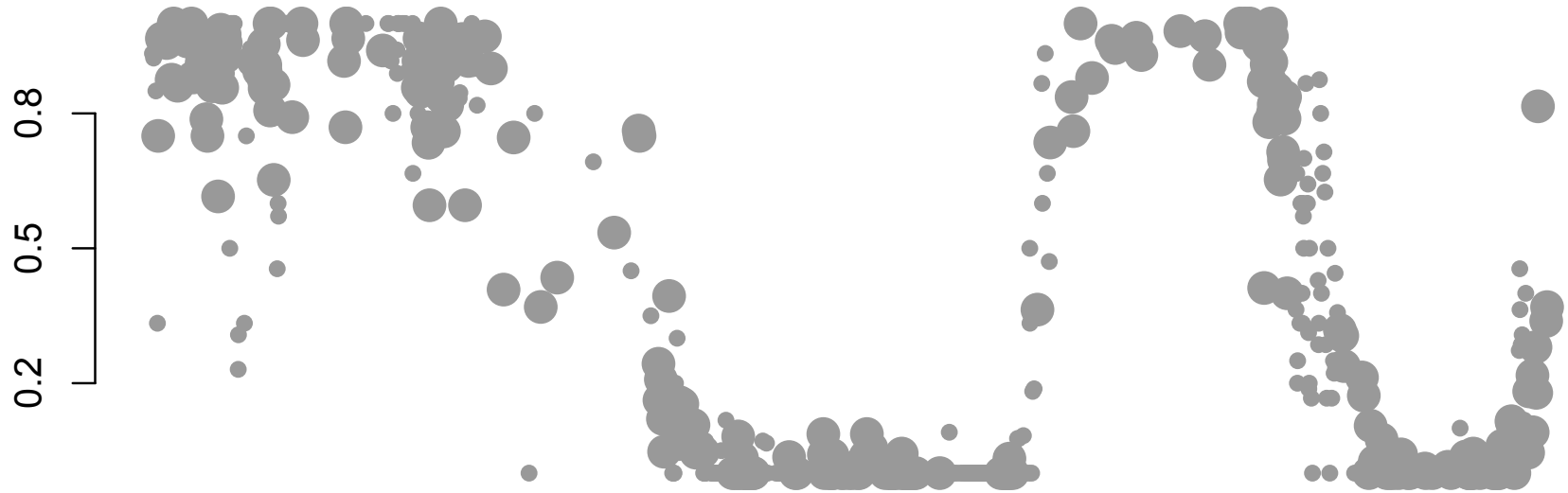
The cost of 30x

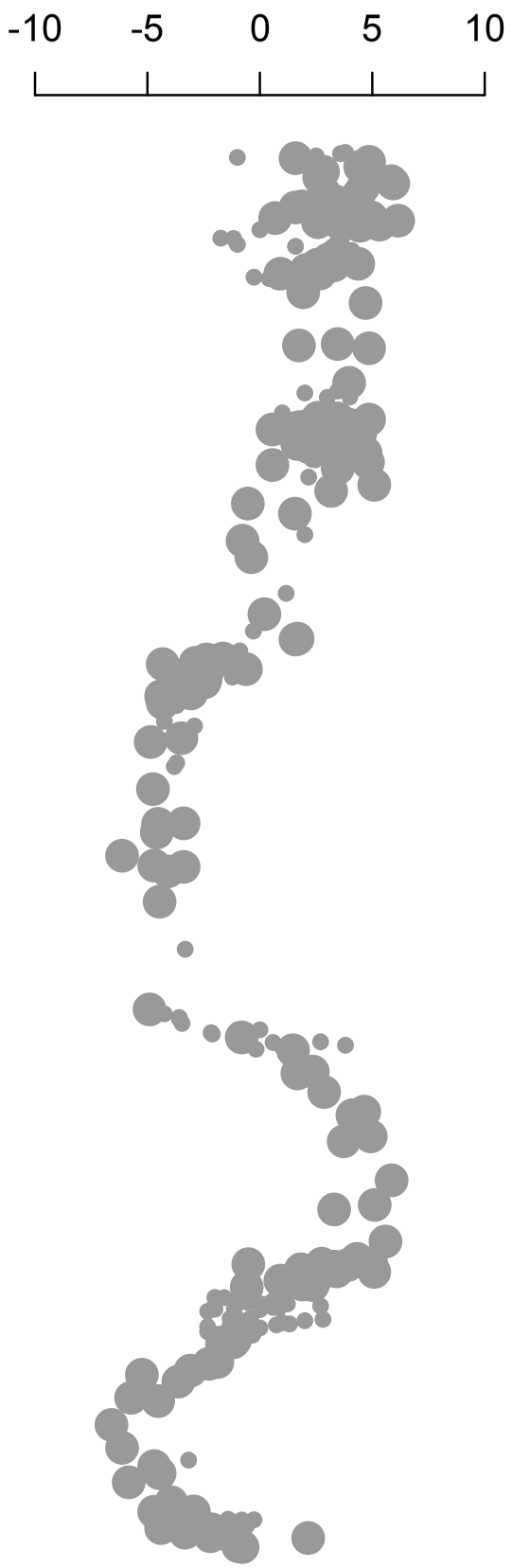
- We need biological replicates
- 3×10^9 x bases x ($ per base) x # samples = more $ than collaborator has
- Can we smooth to save $?

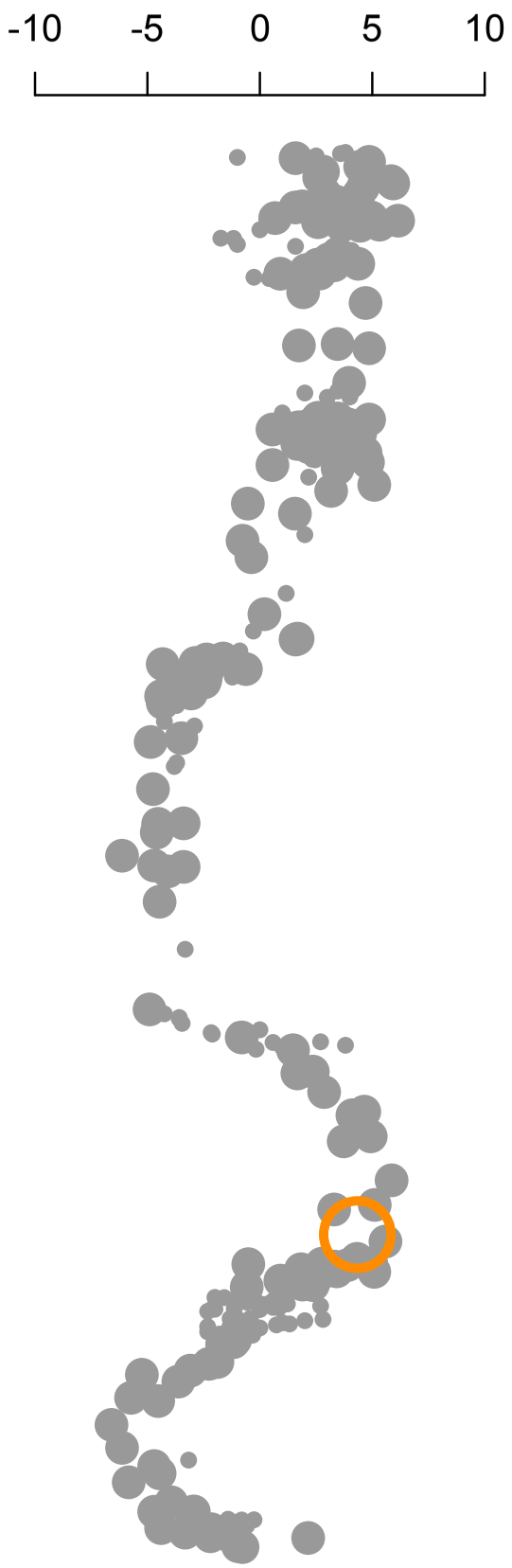
M-bias plots for sequencing

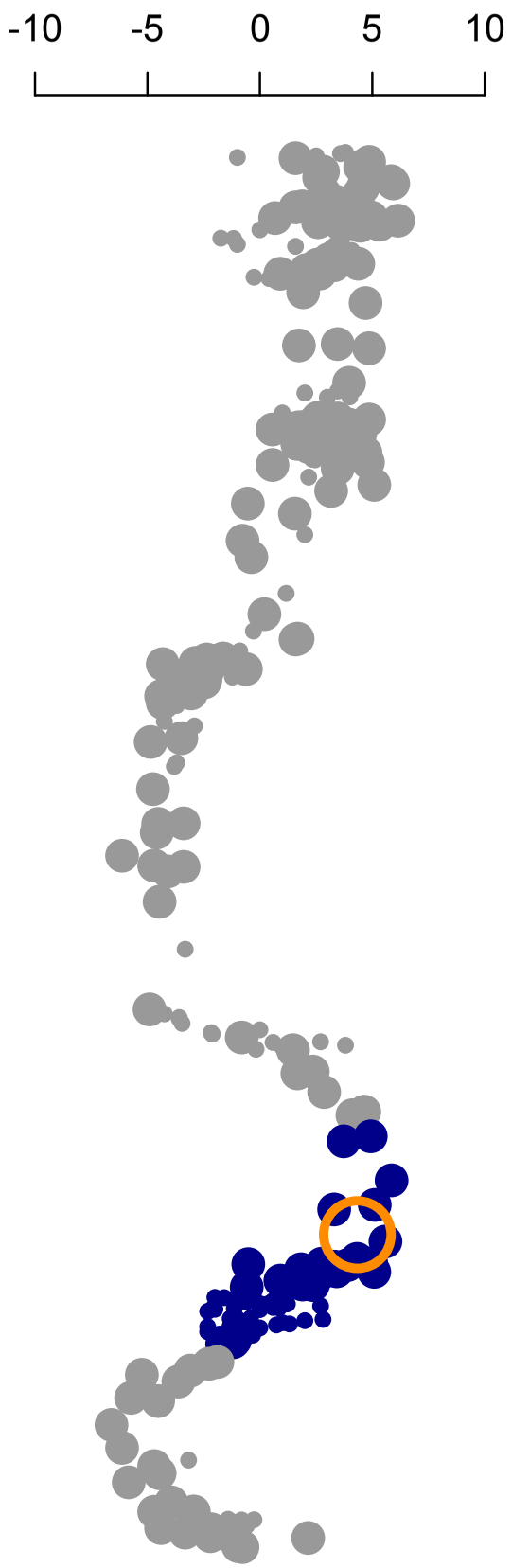


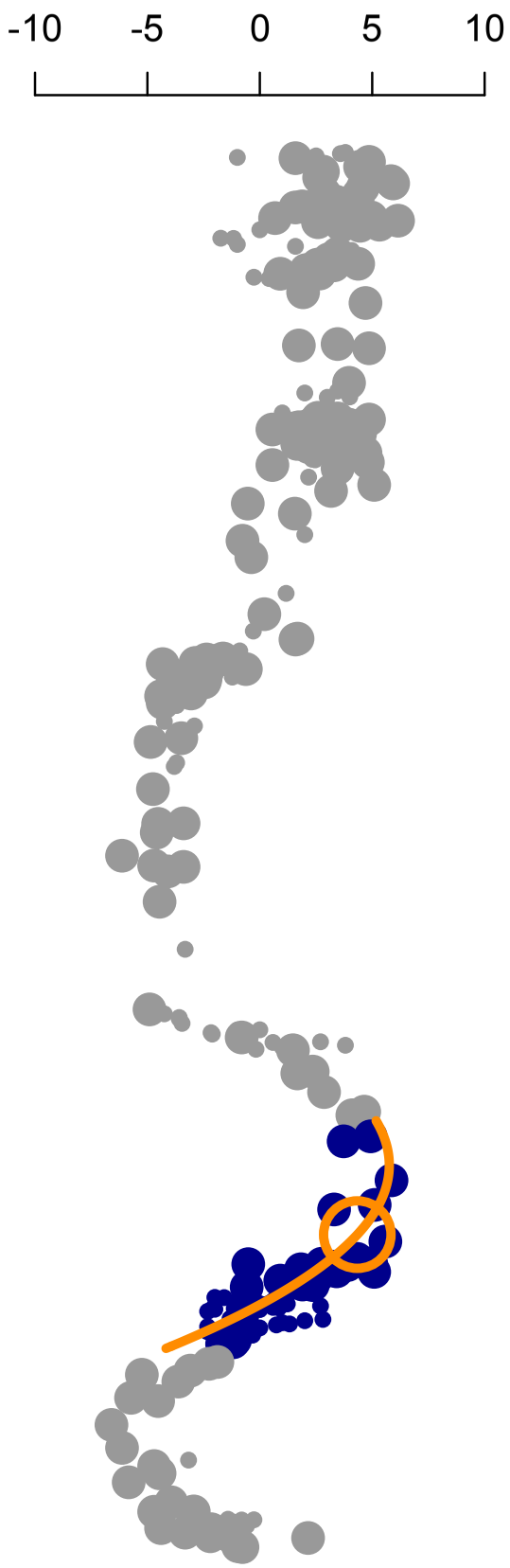
The Data

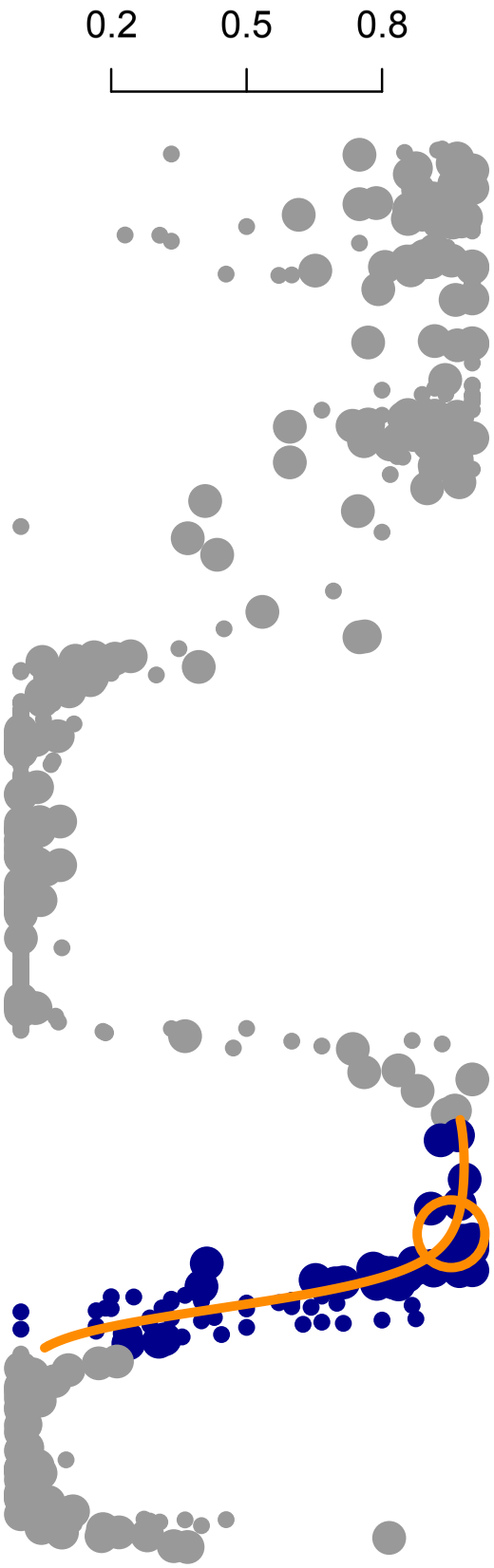


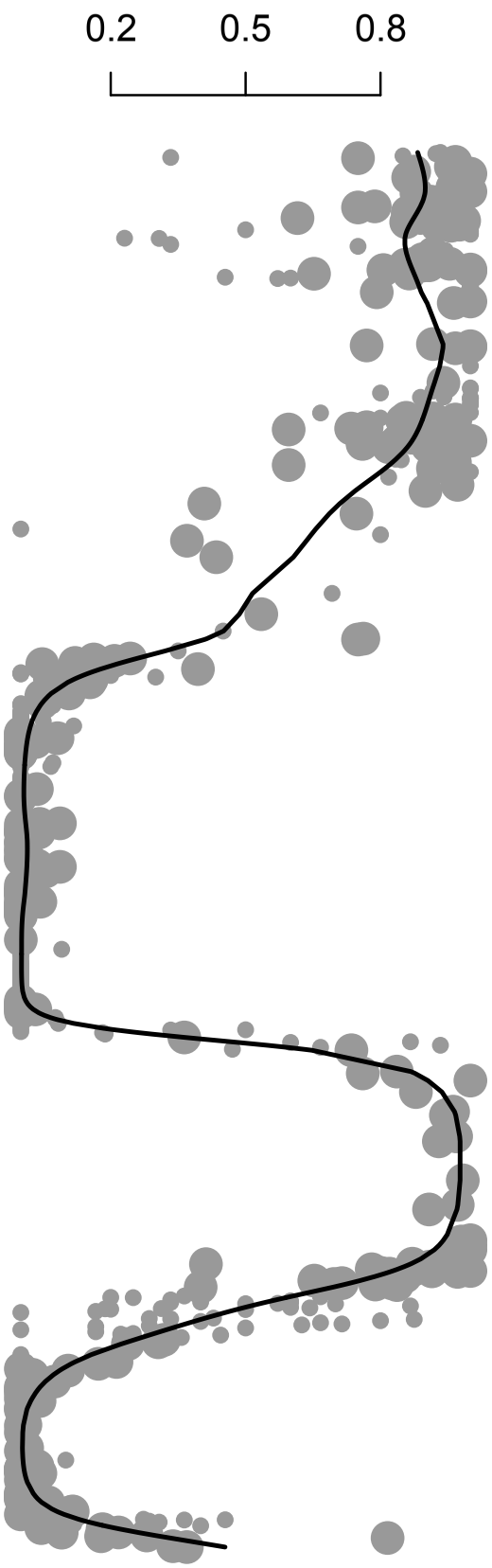

-10
-5
0
5
10


-10
-5
0
5
10

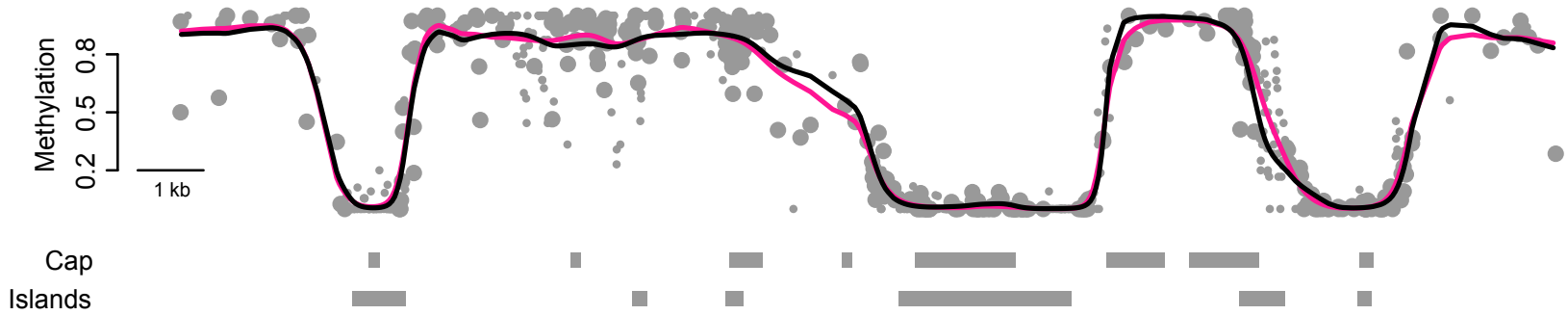

-10
-5
0
5
10


-10
-5
0
5
10

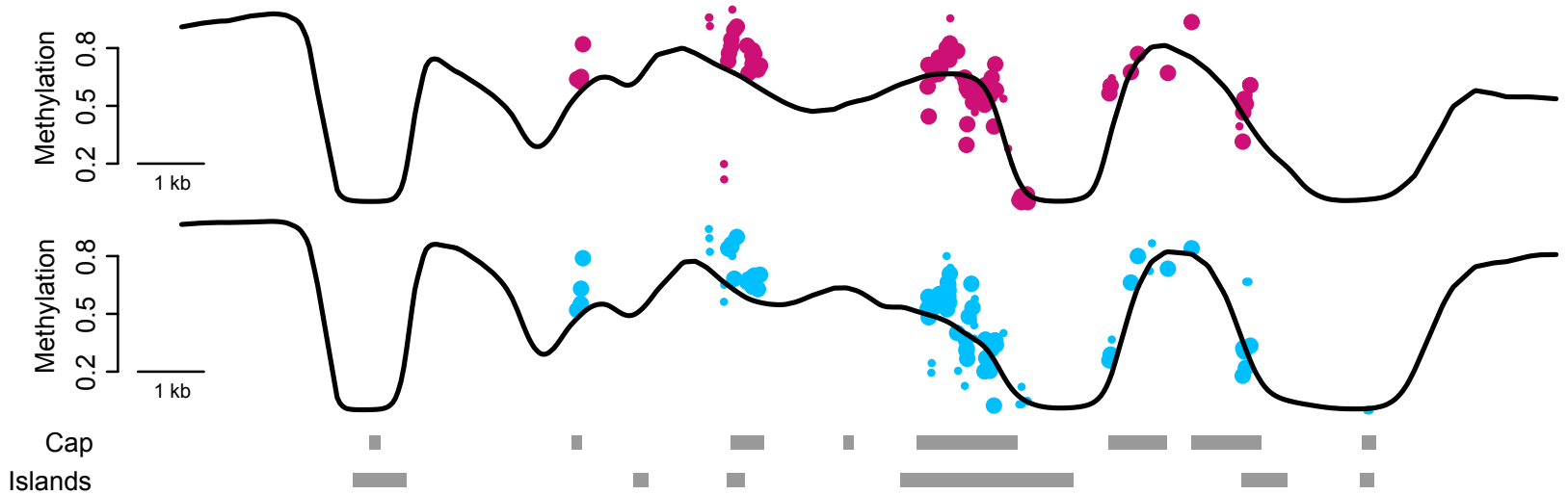

0.2
0.5
0.8


0.2
0.5
0.8

Smoothing on 4x vs 30x

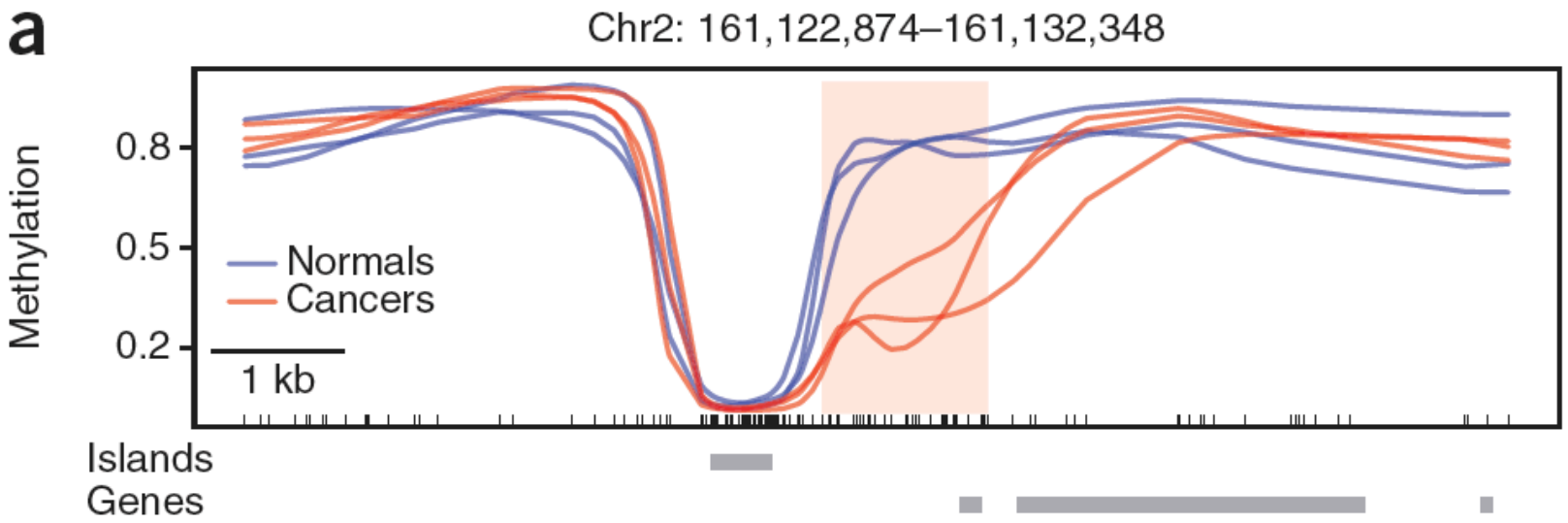


Smoothing on 4x vs capture data

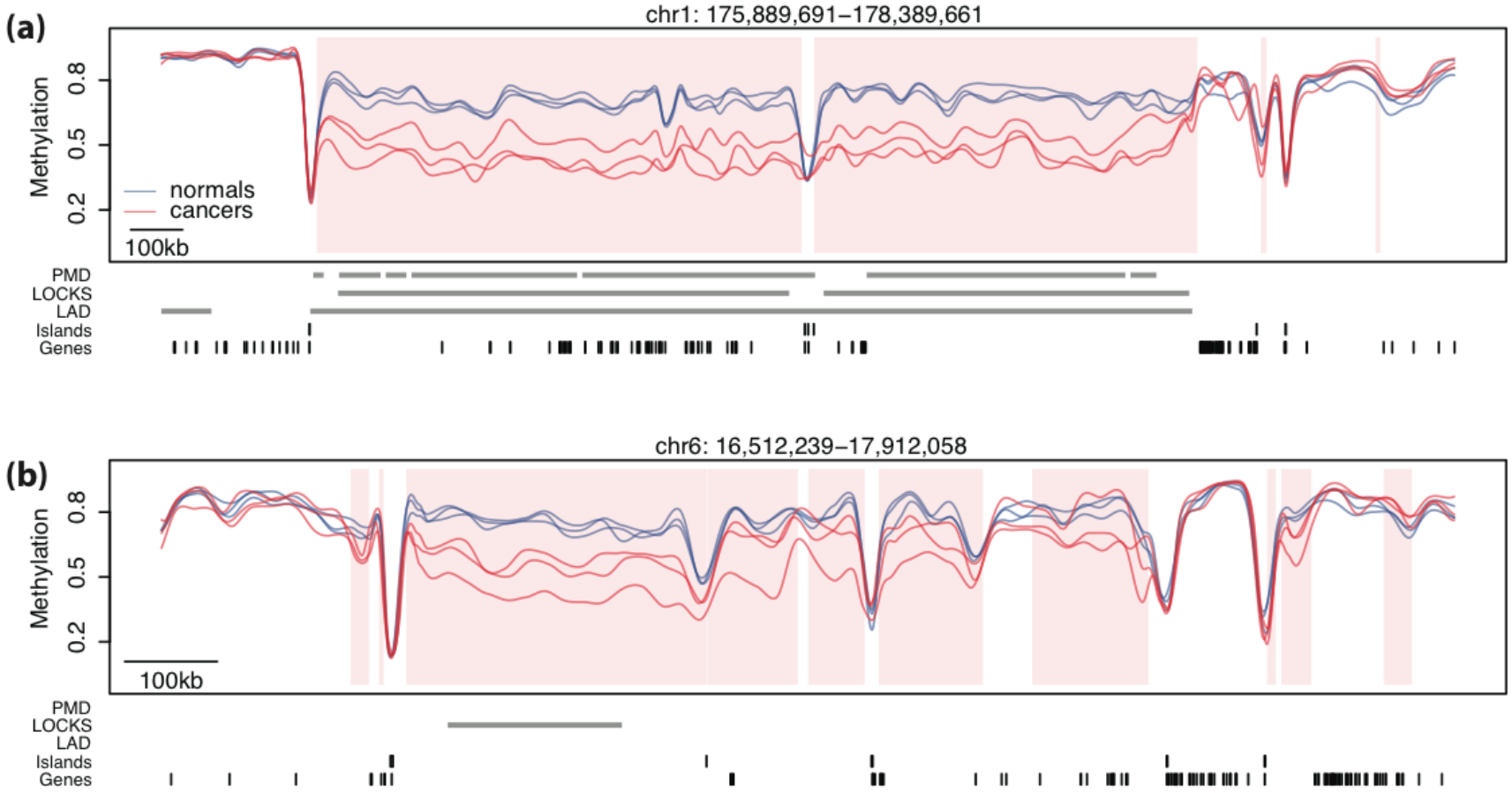


Two levels

Differentially methylated region



Hypomethylated blocks



End