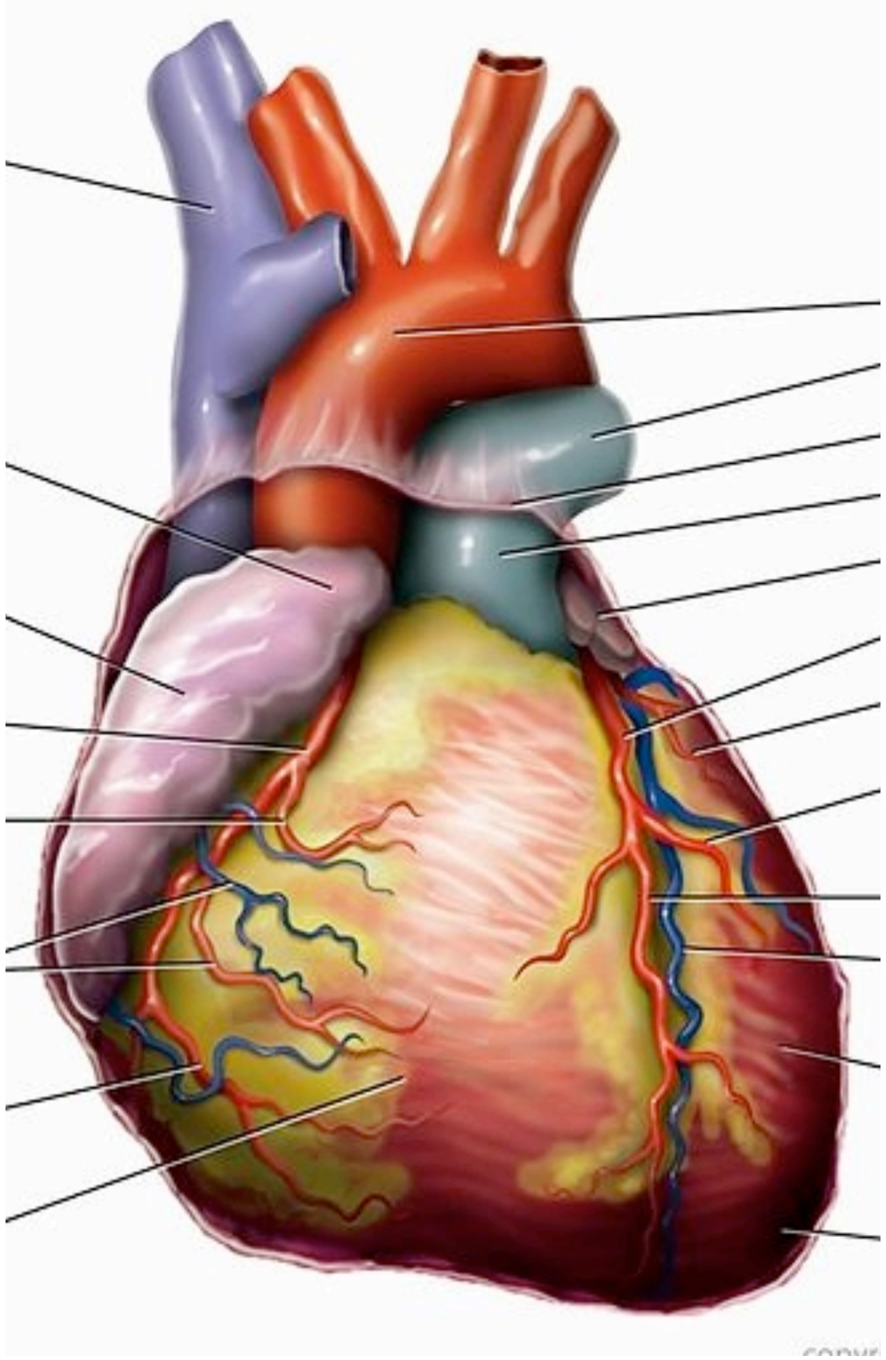
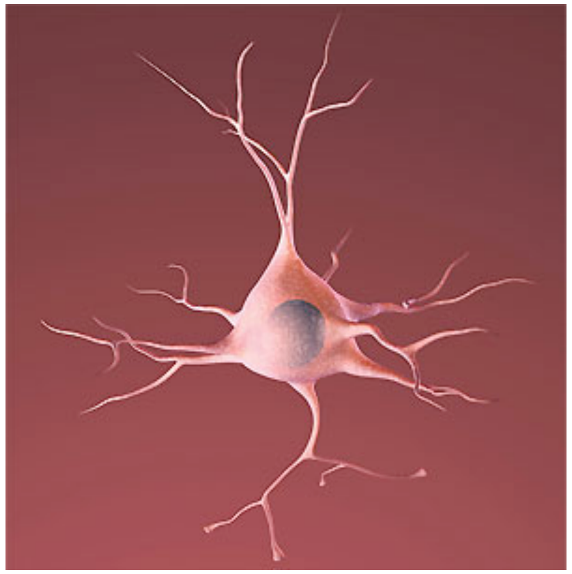
Analysis of shotgun bisulfite sequencing of cancer samples

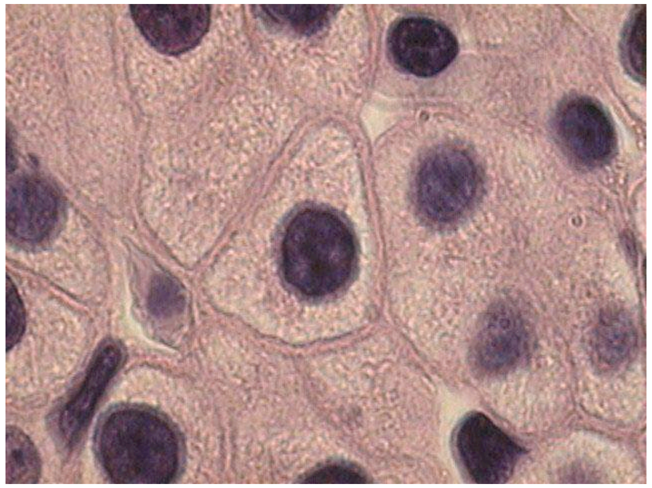
Kasper Daniel Hansen
<khansen@jhsph.edu>
Postdoc with Rafael Irizarry
Johns Hopkins Bloomberg School of Public Health
Brixen, July 1st, 2011

The basis of phenotypic variation: species

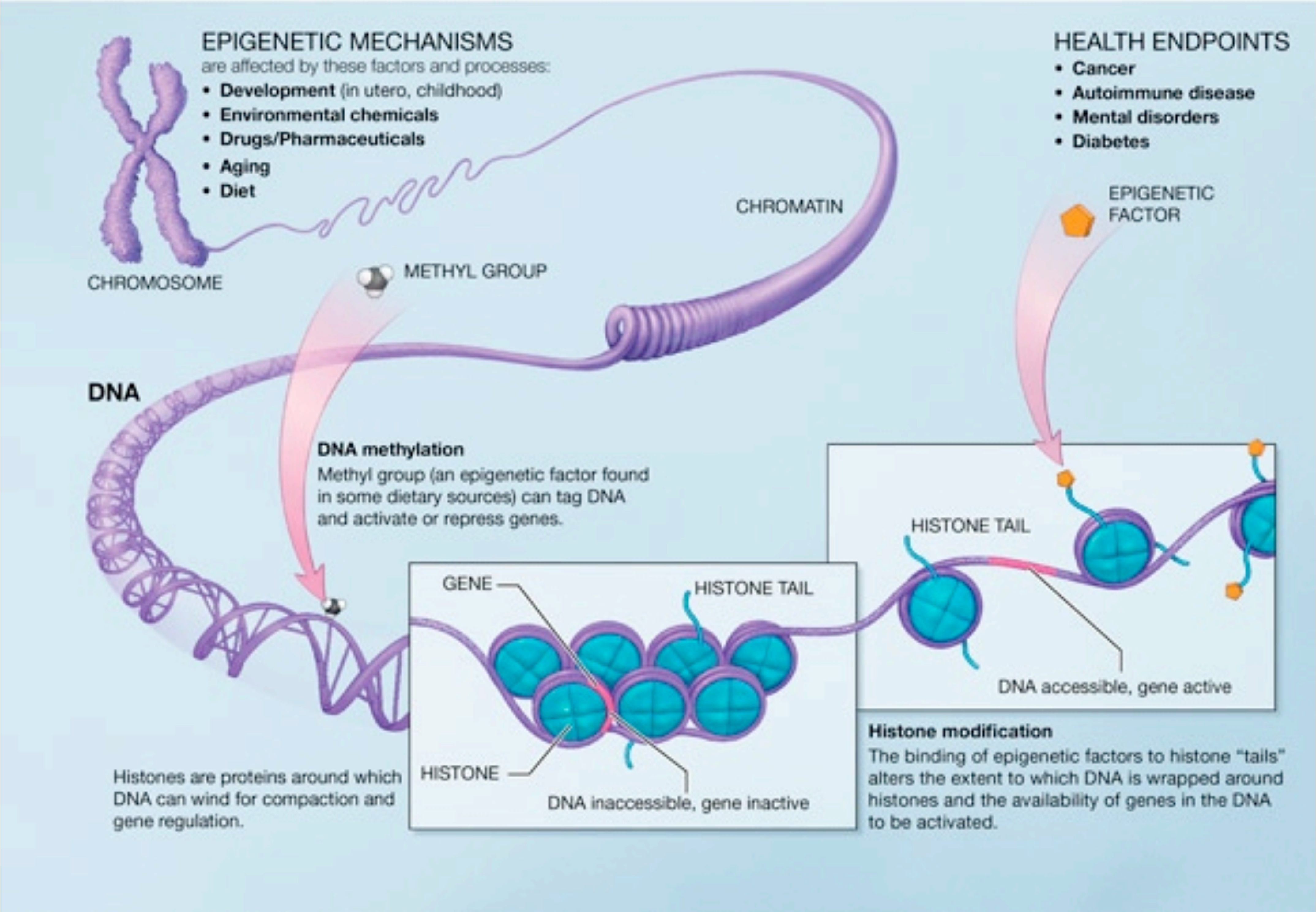
The basis of phenotypic variation: tissues







Epigenetics



EPIGENETIC MECHANISMS
are affected by these factors and processes:

- **Development** (in utero, childhood)
- **Environmental chemicals**
- **Drugs/Pharmaceuticals**
- **Aging**
- **Diet**

HEALTH ENDPOINTS

- **Cancer**
- **Autoimmune disease**
- **Mental disorders**
- **Diabetes**

DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

Histones are proteins around which DNA can wind for compaction and gene regulation.

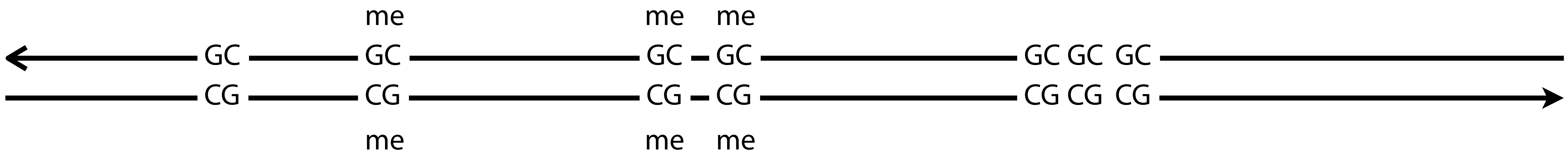
Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

Heritable changes in phenotype that are not caused by changes in DNA.

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

DNA Methylation

In humans: methylation occurs at CpG dinucleotides (28.2M)



CpGs are depleted genomewide.

CpGs tend to cluster together (clusters are termed CpG Islands), these clusters are enriched in or near promoters.

Methylation is associated with “openness” of the DNA.
Hypermethylation (high) is associated with gene silencing
Hypomethylation (low) is associated with active genes

Methylation is inherited (at least in cell division).

Measuring DNA methylation

PCR does not preserve methylation information.
Hybridization is not affected by methylation.

Pretreatment	**Analytical step**			
	Locus-specific analysis	**Gel-based analysis**	**Array-based analysis**	**NGS-based analysis**
Enzyme digestion	• *Hpa*II-PCR	• Southern blot • RLGS • MS-AP-PCR • AIMS	• DMH • MCAM • HELP • MethylScope • CHARM • MMASS	• Methyl–seq • MCA–seq • HELP–seq • MSCC
Affinity enrichment	• MeDIP-PCR		• MeDIP • mDIP • mCIP • MIRA	• MeDIP–seq • MIRA–seq
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC–seq • BSPP • WGSBS

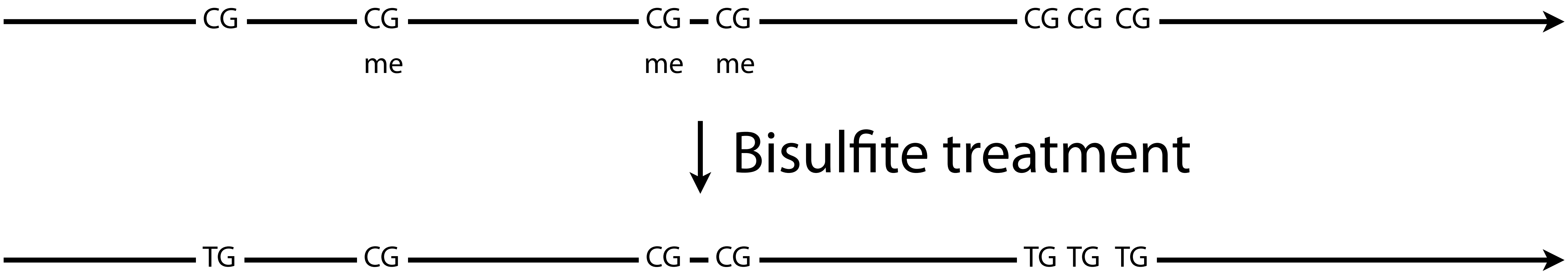
Illumina methylation arrays:
GoldenGate (early 2007, 1.5k CpGs), “27k” (late 2007), “450k” (2011)

Laird, *Nature Reviews Genetics* (2010)

Bisulfite treatment

The gold standard for measuring DNA methylation at single CpGs is bisulfite treatment followed by Sanger or Pyro sequencing

Bisulfite treatment converts unmethylated Cs to Us (= T)



CG — CG — CG – CG — CG CG CG →

me me me

↓ Bisulfite treatment

TG — CG — CG – CG — TG TG TG →

Can be used genome-wide, but requires the same sequencing effort as whole genome DNA sequencing (= expensive).

Cancer and DNA methylation

DNA methylation in cancer was the first epigenetic modification discovered in cancer (~25 years ago).

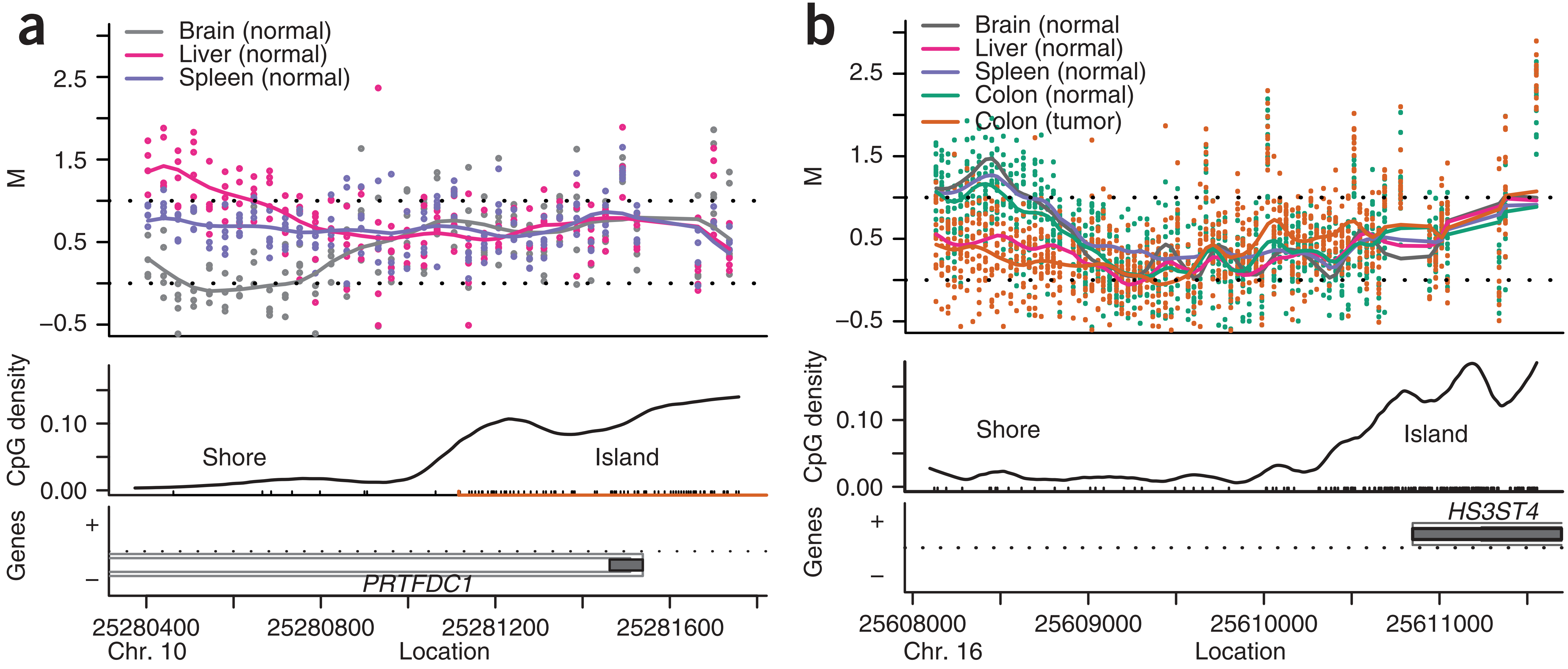
Focus (at least lately) in the literature have been on
  hyper methylation of CpG islands in promoters (tumor supprs)
  hypo methylation of select repeat elements

although
  global hypomethylation
  hypo methylation of selected genes (typically oncogenes)

have also been described.

Methylation terminology
  Hyper: goes up,   Hypo: goes down
  DMR: differentially methylated region

CpG Islands shores



Many changes are not in CpG islands, but in regions bordering CpG islands; termed CpG Island shores.

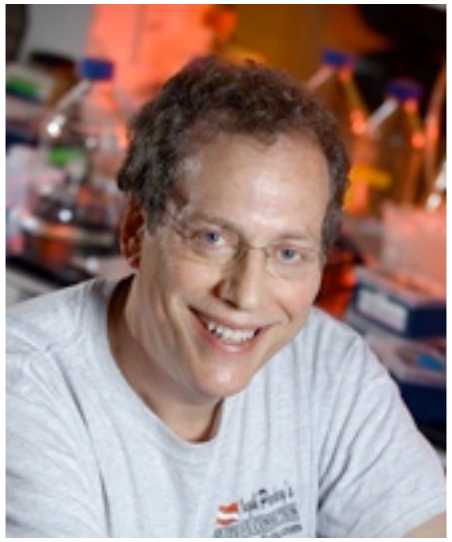
Irizarry et al. *Nature Genetics* (2009)

Acknowledgements

Increased methylation variation in epigenetic domains across cancer types

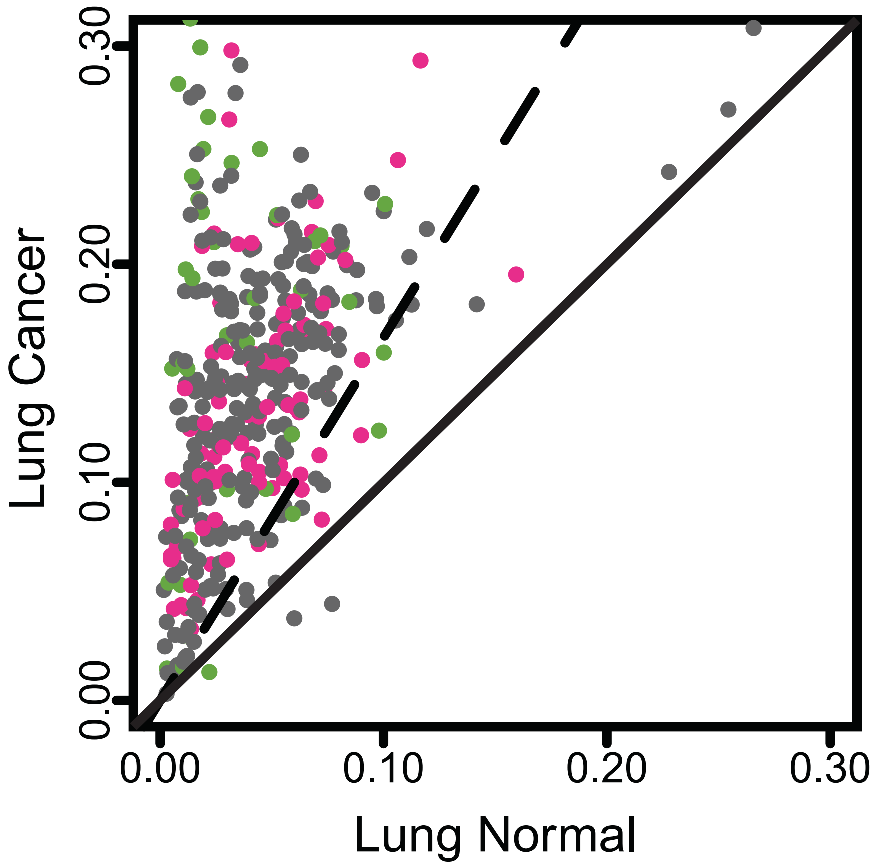
Kasper Daniel Hansen[1,2,10], Winston Timp[2–4,10], Héctor Corrada Bravo[2,5,10], Sarven Sabunciyan[2,6,10], Benjamin Langmead[1,2,10], Oliver G McDonald[2,7], Bo Wen[2,3], Hao Wu[8], Yun Liu[2,3], Dinh Diep[9], Eirikur Briem[2,3], Kun Zhang[9], Rafael A Irizarry[1,2] & Andrew P Feinberg[2,3]

Nature Genetics, Advance Online



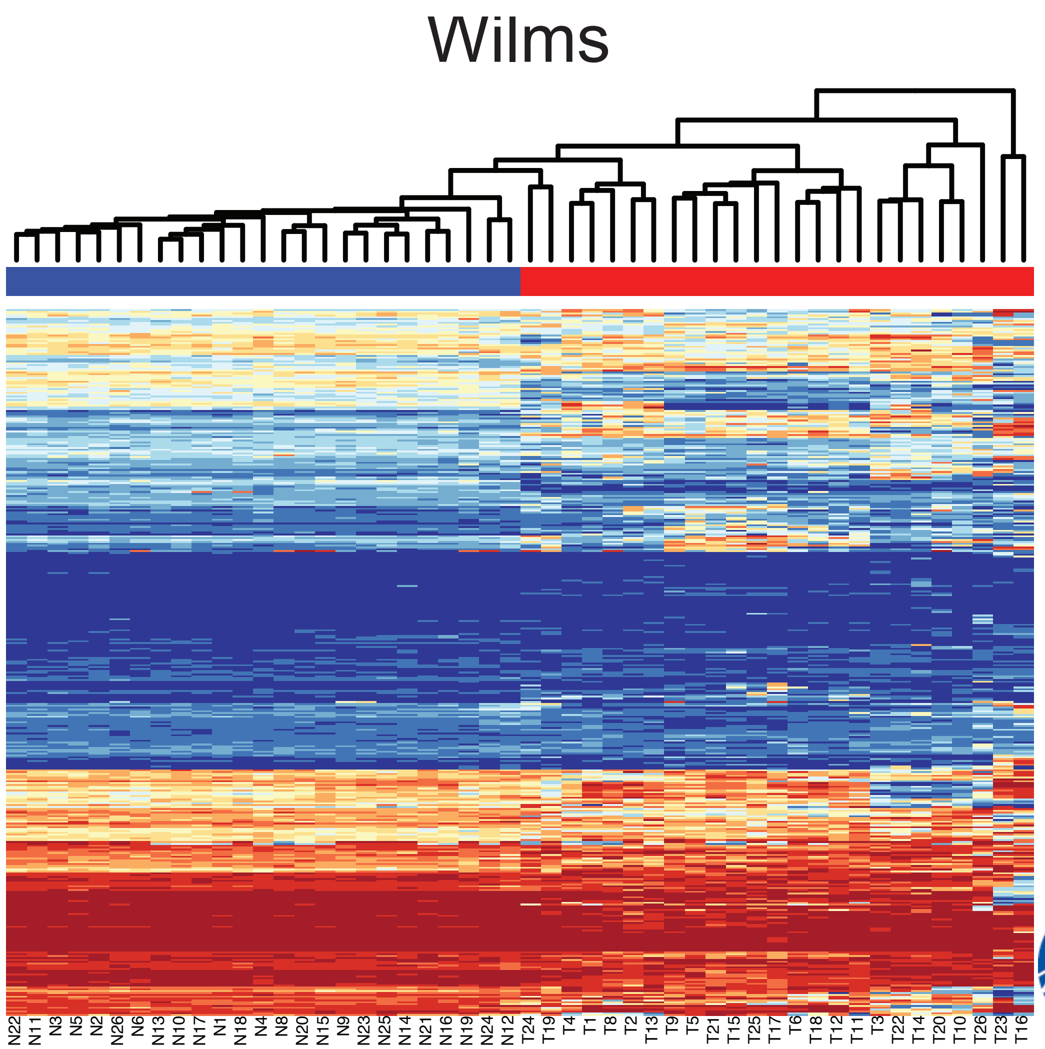
JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Increased methylation variation across all cancers



Increased variation between normals and cancers, for the *same* regions across *all* 5 cancer types (lung, colon, breast, thyroid and Wilms).

151 regions in 290 samples.



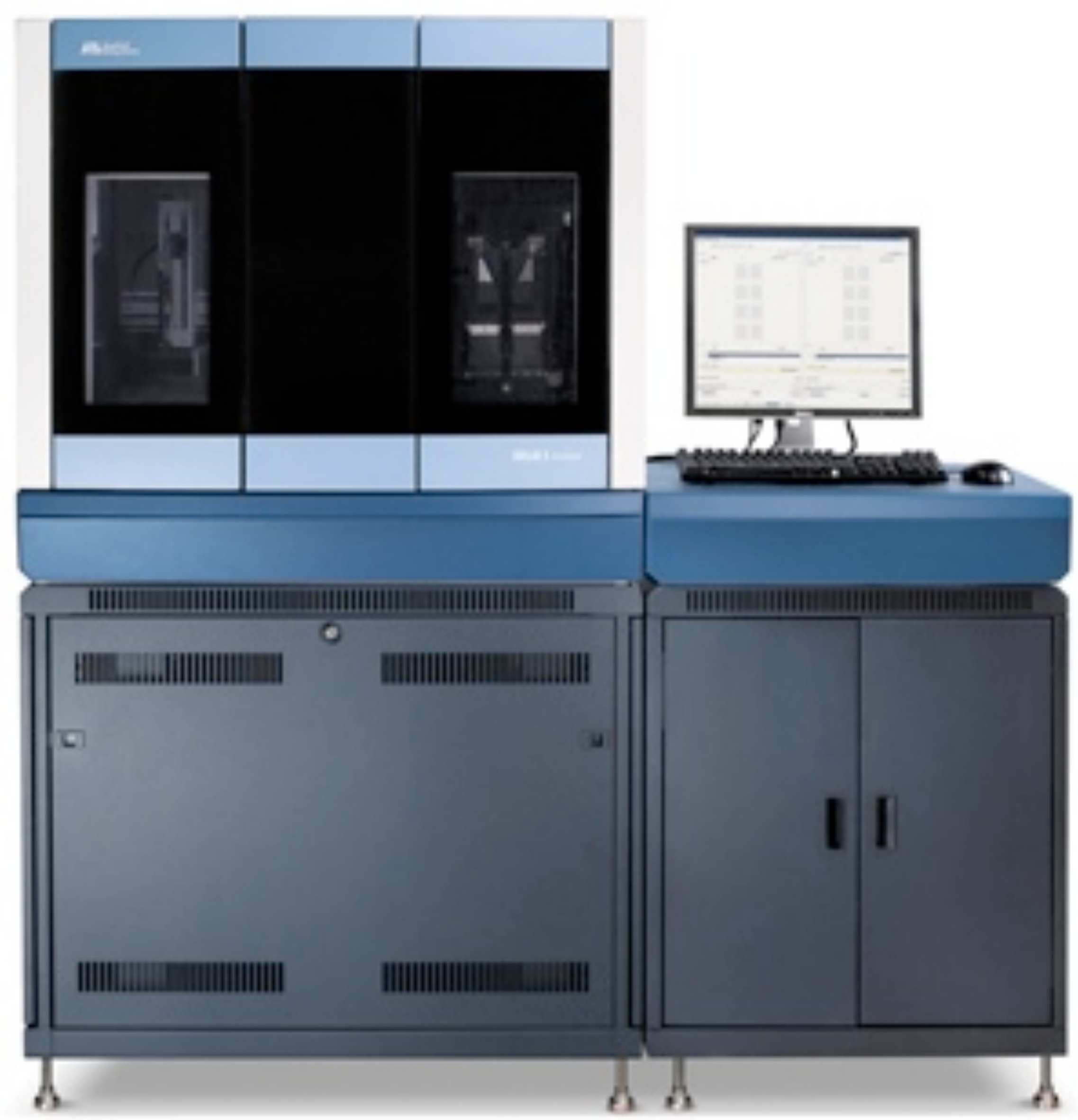
The *same* regions that distingush cancers from normals, distinguishes normal tissue types.

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Design

3 colon cancers and their matched normal mucosa

2 adenomas



ABI SOLiD, 50bp reads

~5x coverage on CpGs

We traded coverage for biological replicates.

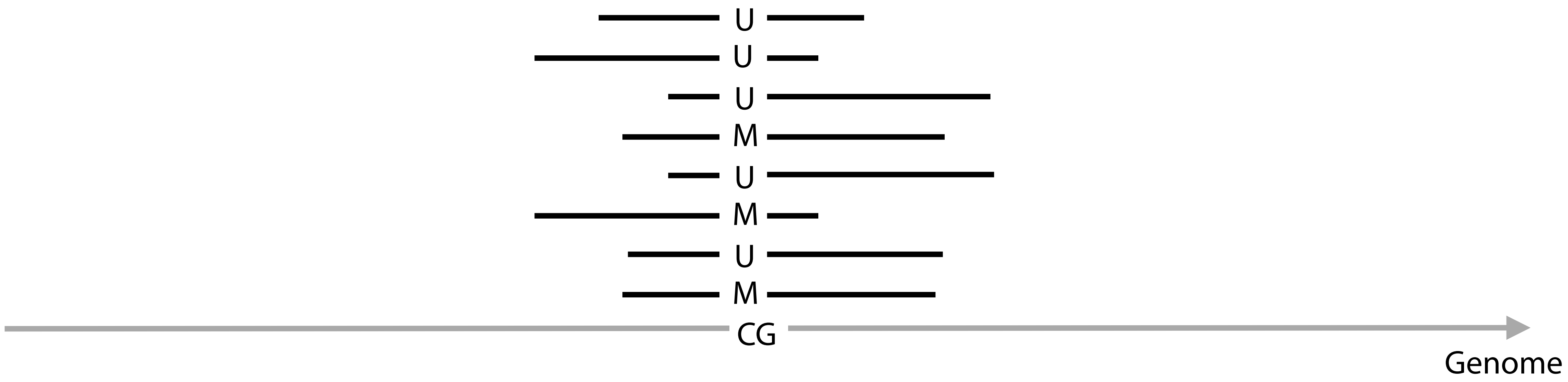
JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Mapping

Bisulfite conversion makes the genome into an (appr) 3 letter alphabet, making mapping hard.

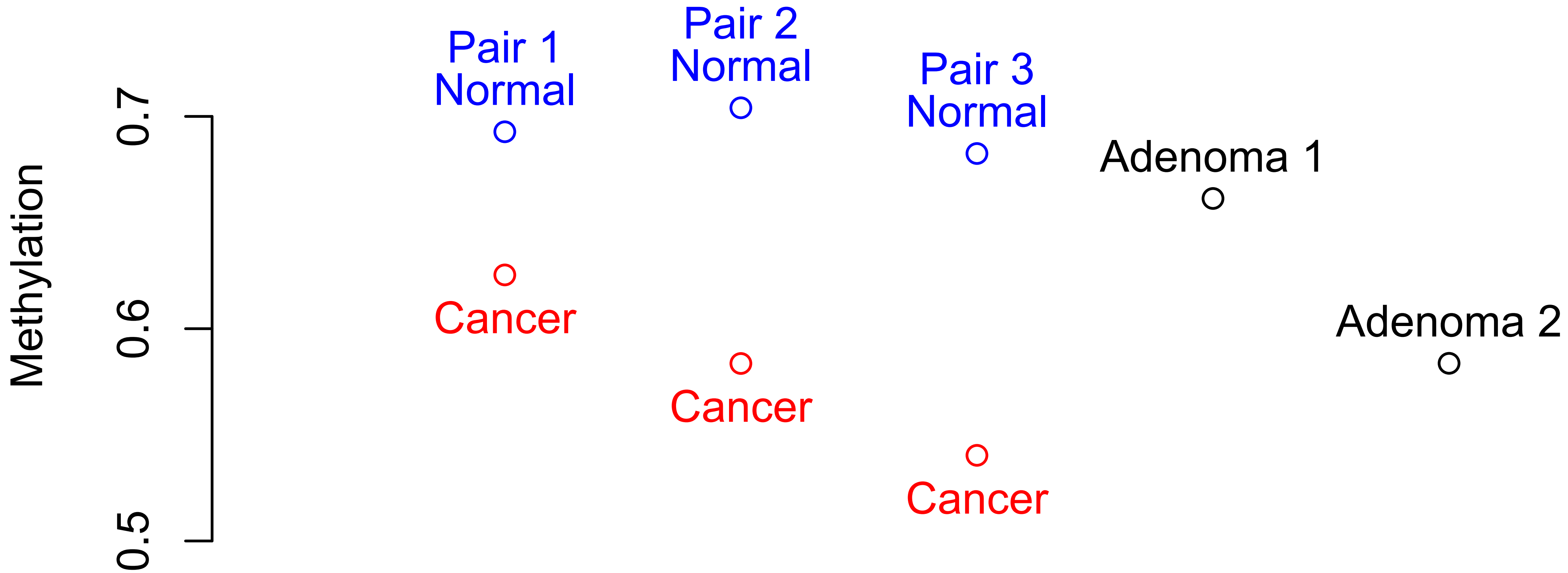
We could not use existing tricks for unbiased alignment of bisulfite sequencing data: we wrote a custom aligner, *Merman*.

We can map ~20M CpGs uniquely.



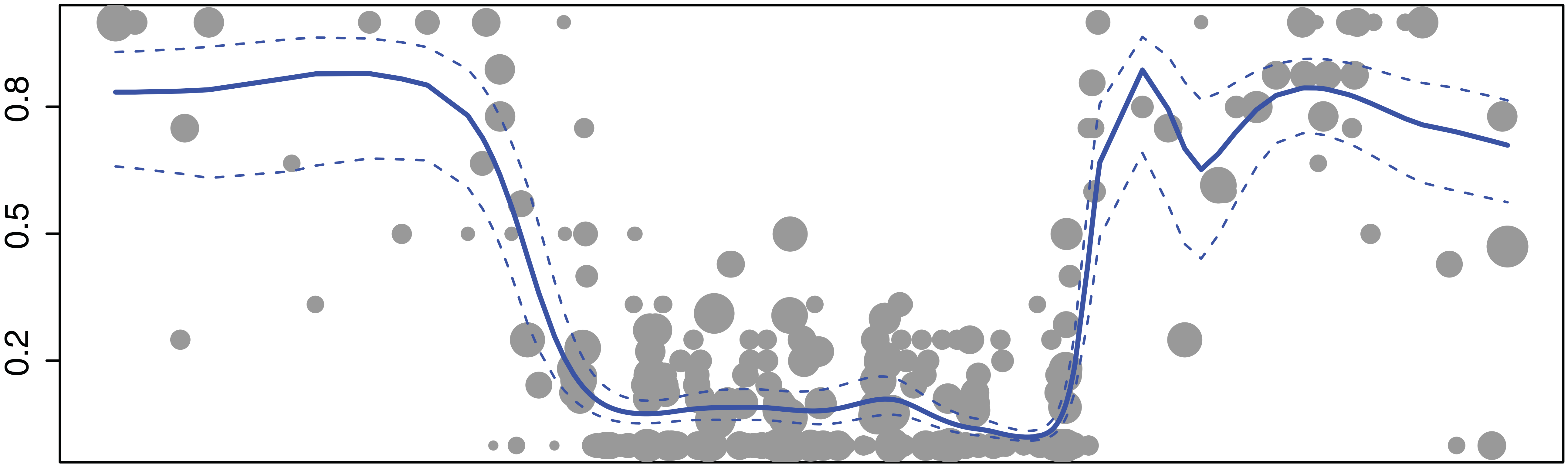
Coverage (for this CpG): 8
3 M's and 5 U's (Unmethylated)

Global levels of methylation



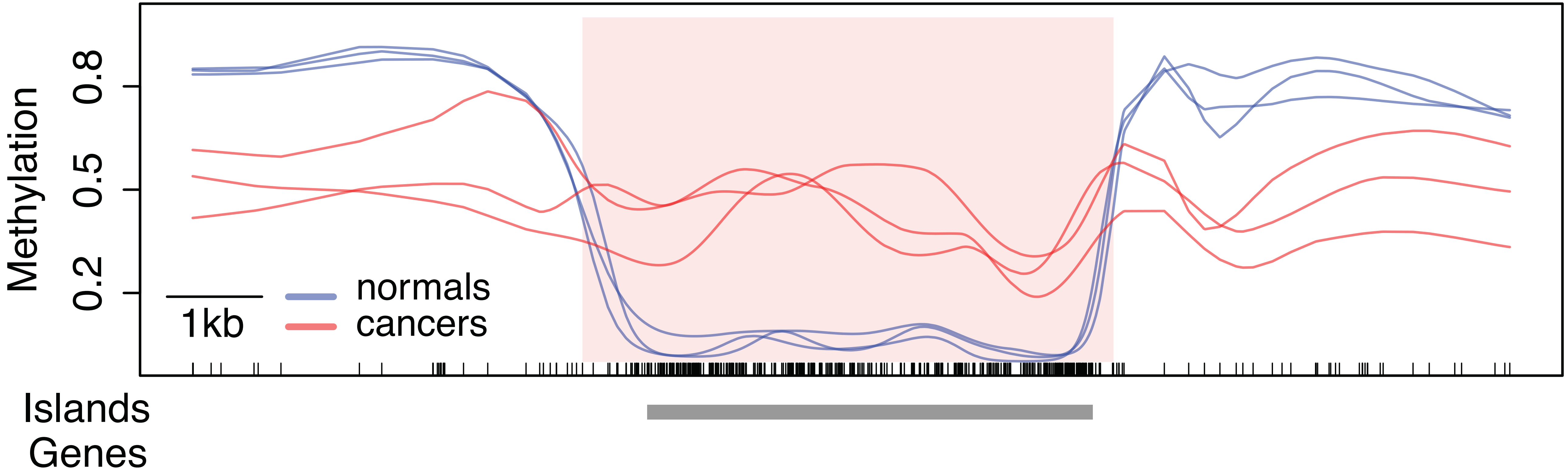
Bisulfite conversion rates estimated using λ phage spike-in to be 99.7-99.8%

One sample, small region ~ 14kb



Smoothing using a binomial model (local likelihood)
Adaptive bandwidth (⟵ important)

Small region



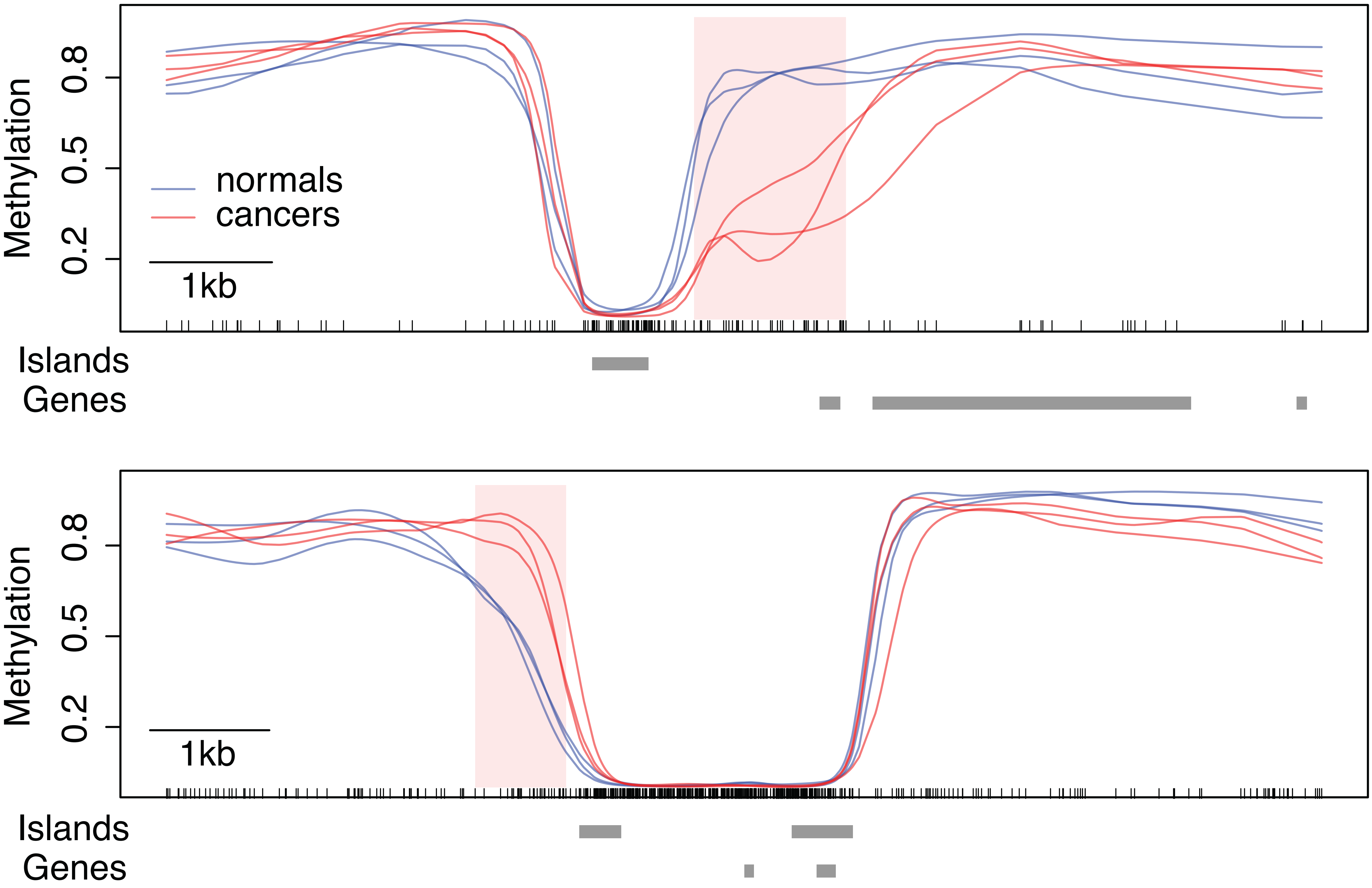
Shore
(hypo)

Island
(hyper)

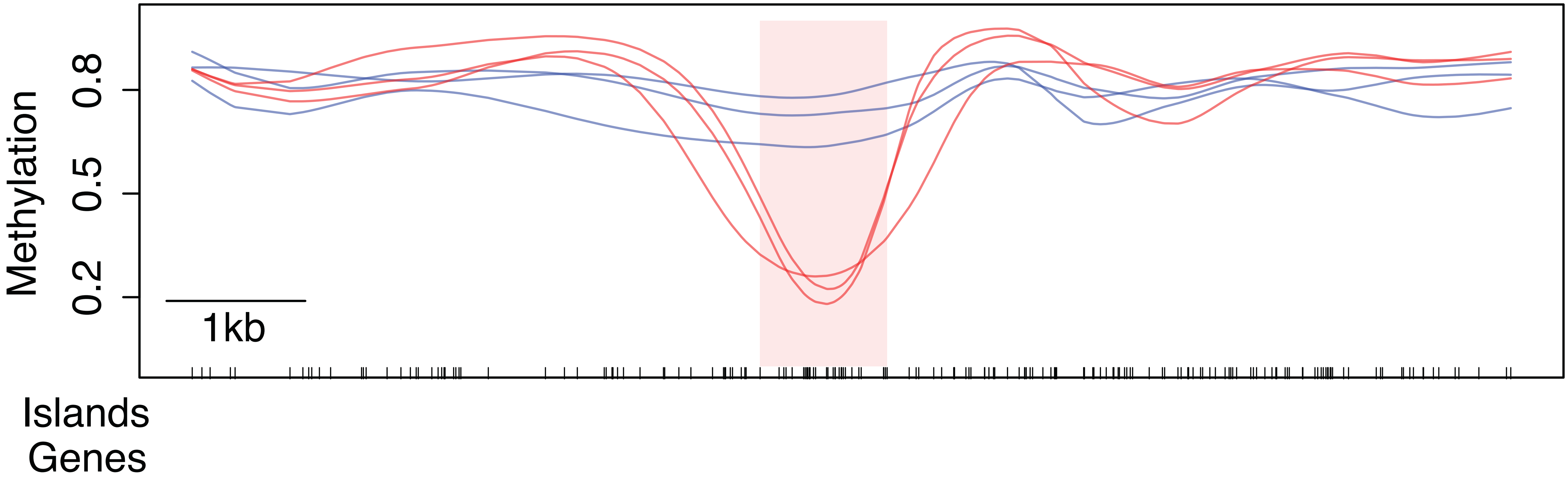
Shore
(hypo)

Loss of methylation boundaries in cancer

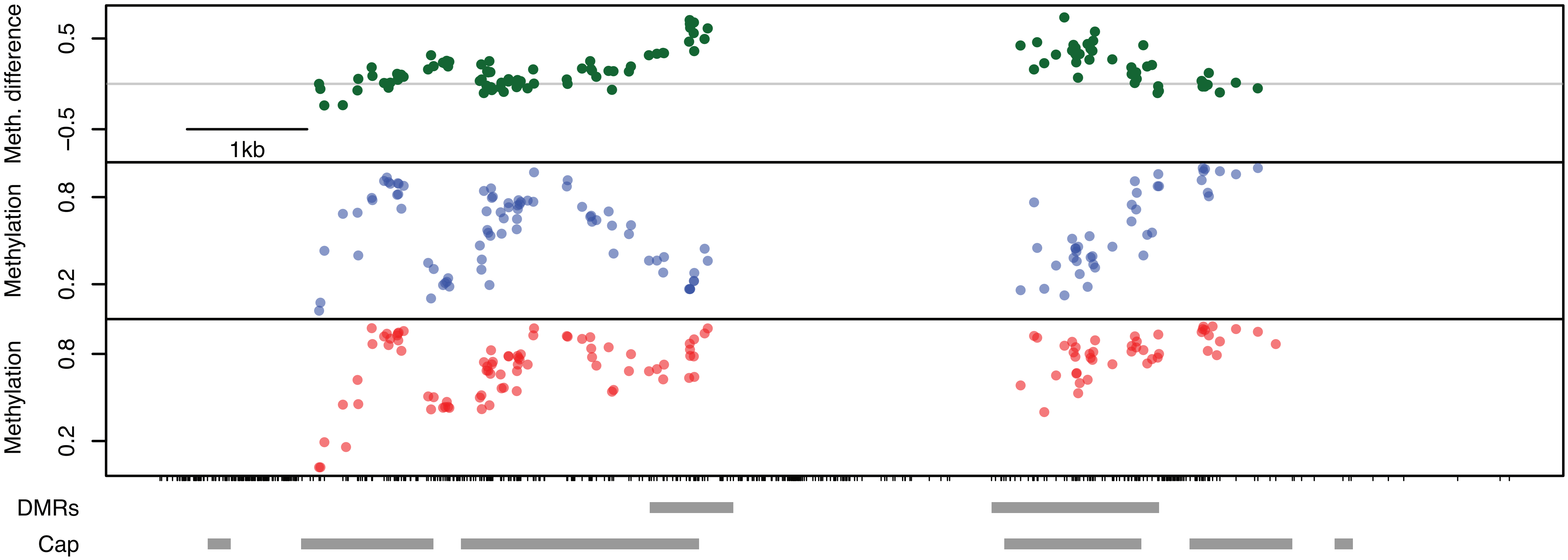
Boundary Shifts (inwards and outwards)



Novel hypomethylation



Capture bisulfite



~40,000 capture regions, ~400,000 CpGs

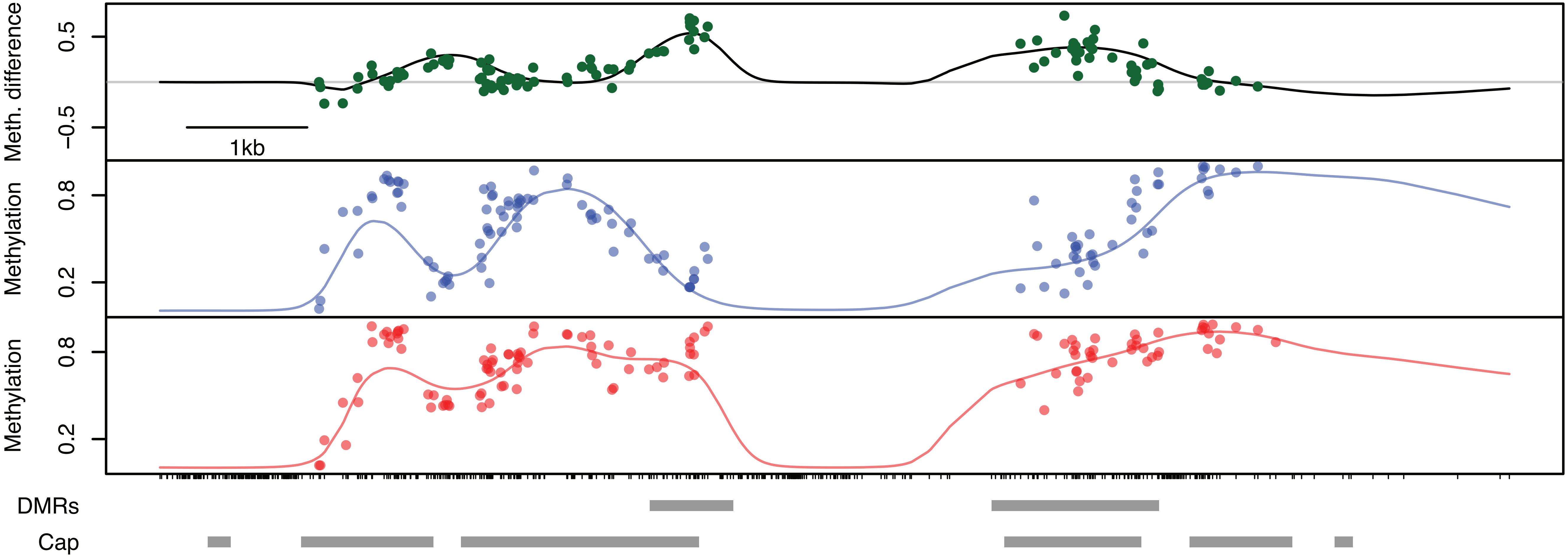
Red: Average of cancers

Blue: Average of normals

Green: Difference between cancers and normals

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Capture bisulfite



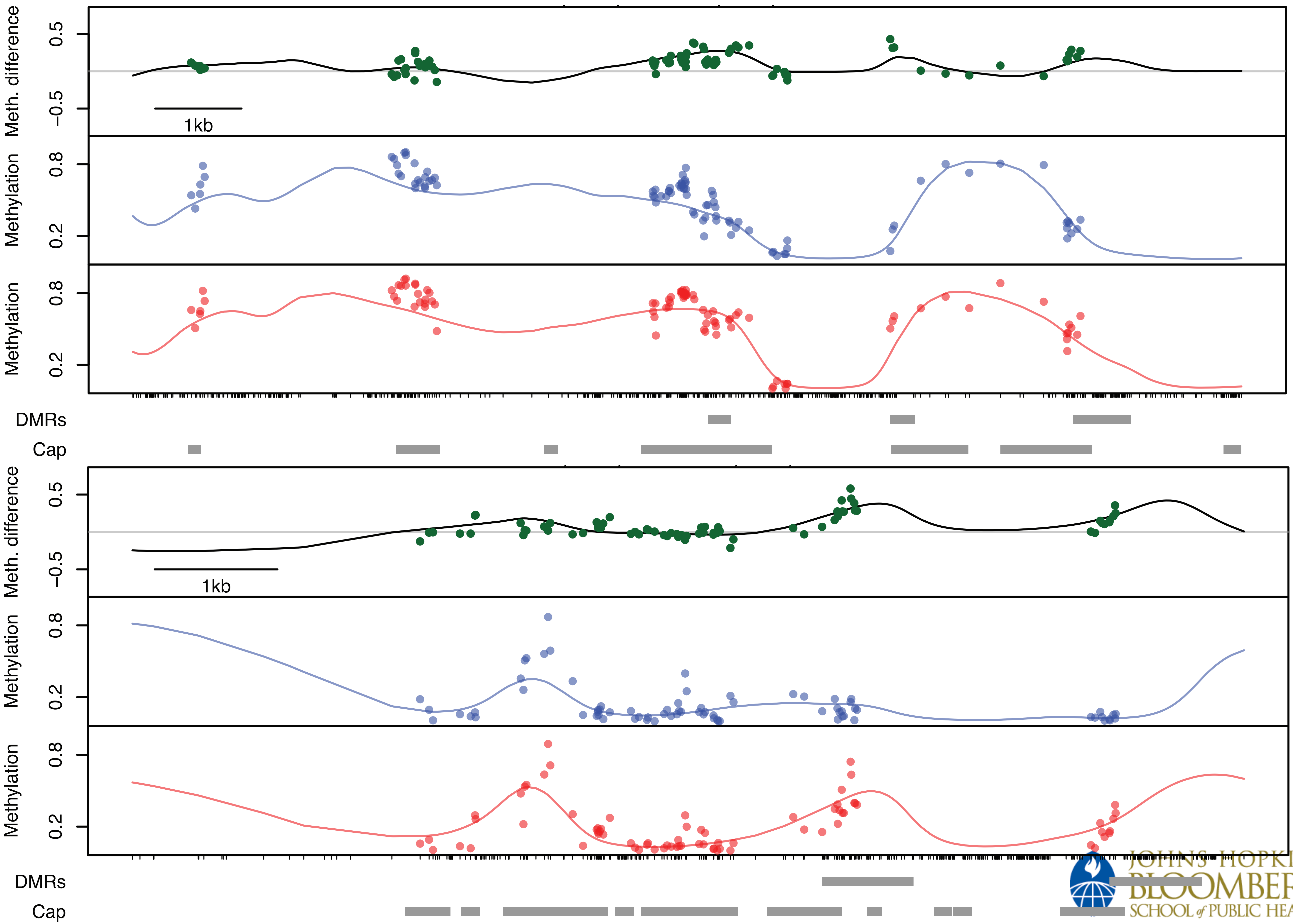
~40,000 capture regions, ~400,000 CpGs

Red: Average of cancers

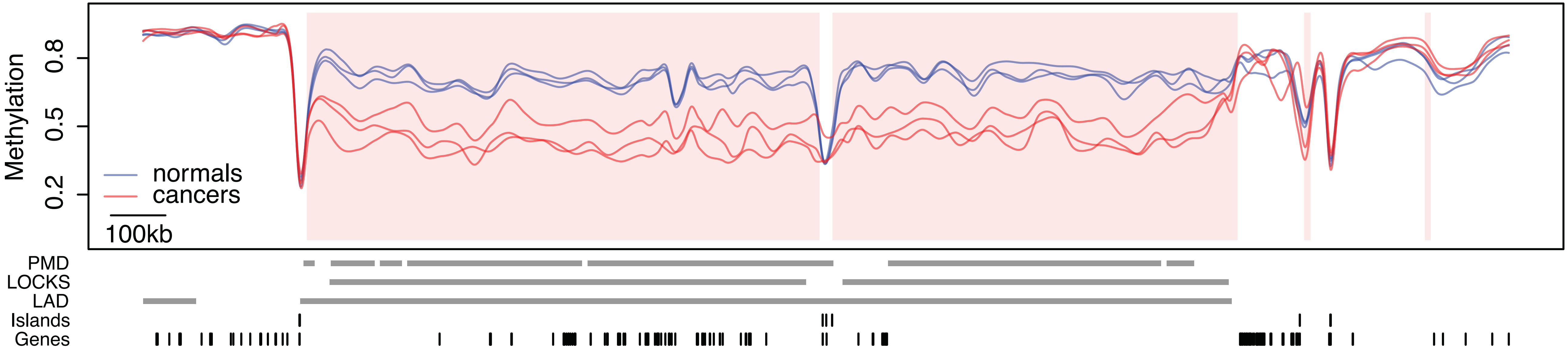
Blue: Average of normals

Green: Difference between cancers and normals

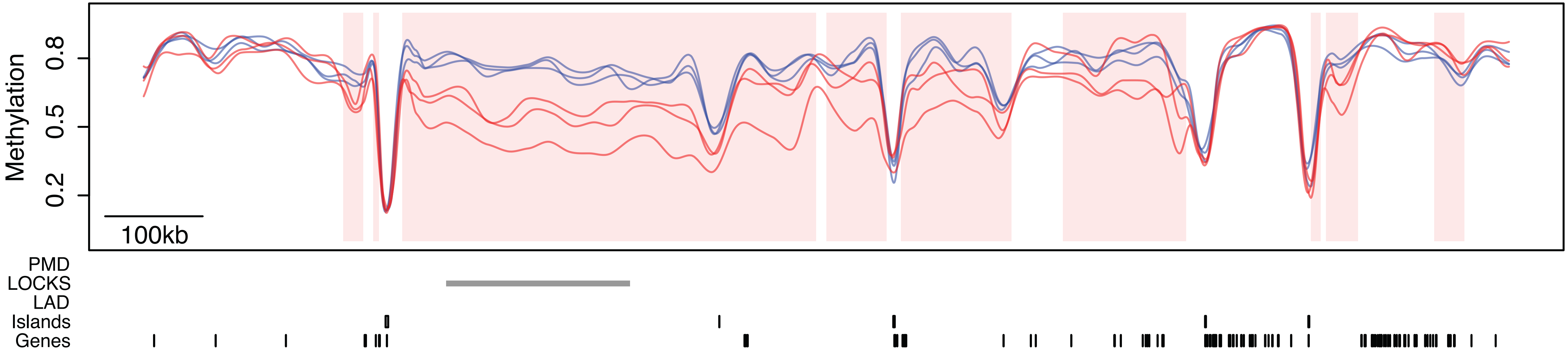
More capture



Large blocks of hypomethylation



Consistent boundaries

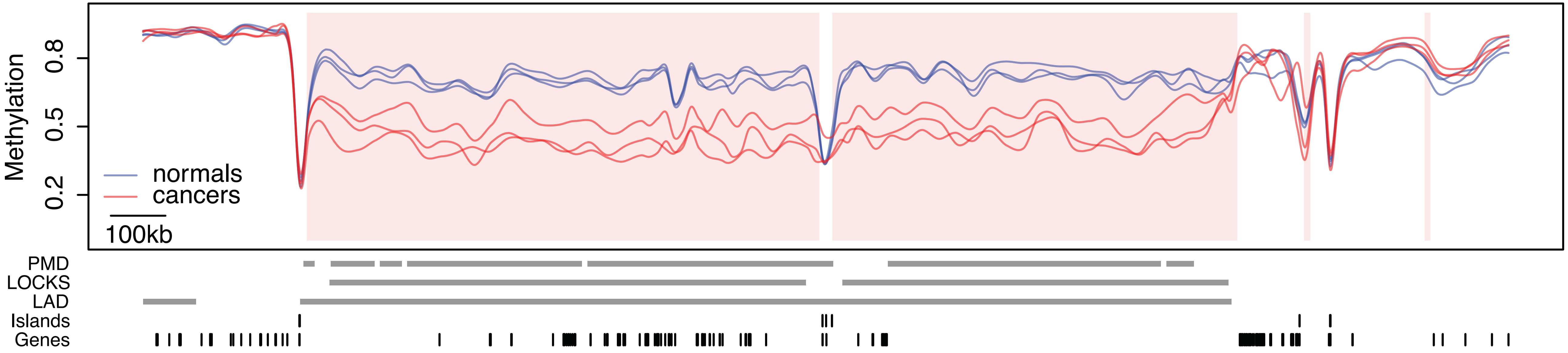


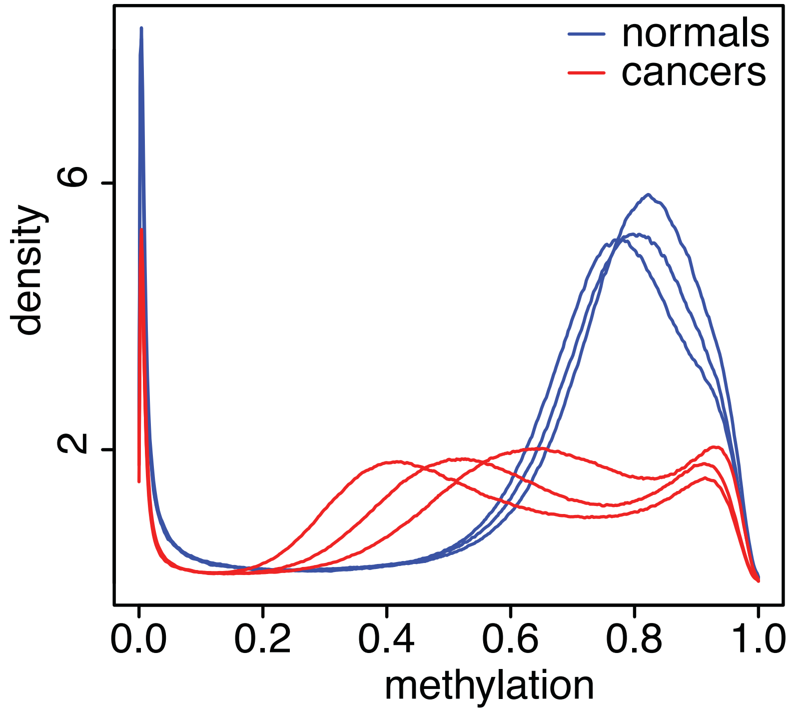
(Some) coincides with LADs, LOCKs and PMDs.

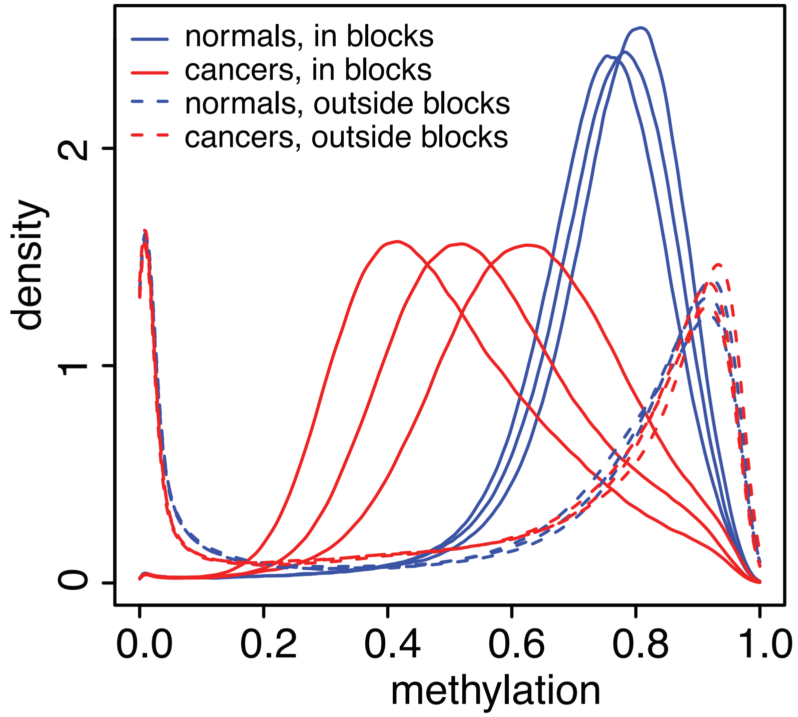
Related to structural conformation of the DNA in the nucleus

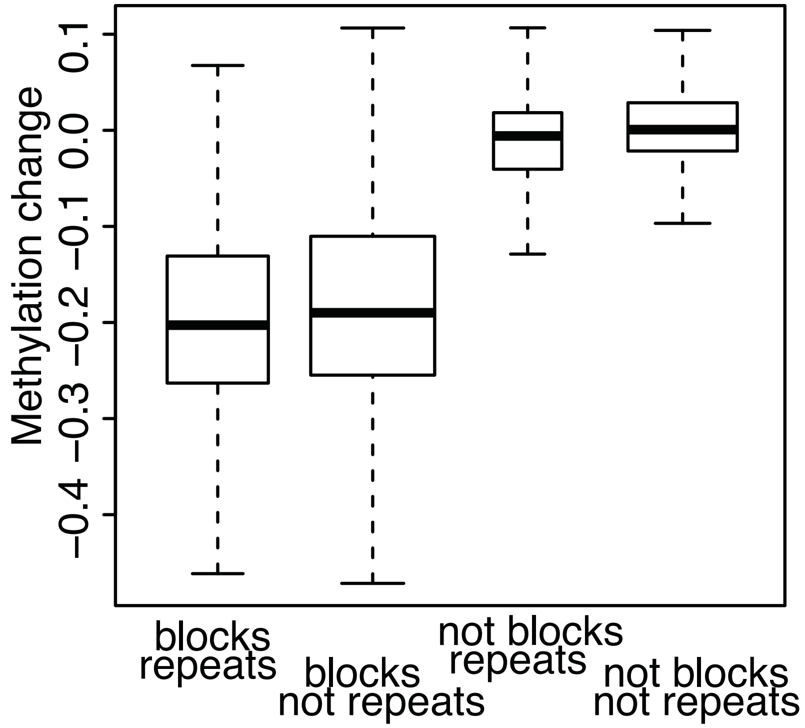
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

What predicts hypomethylation in blocks?

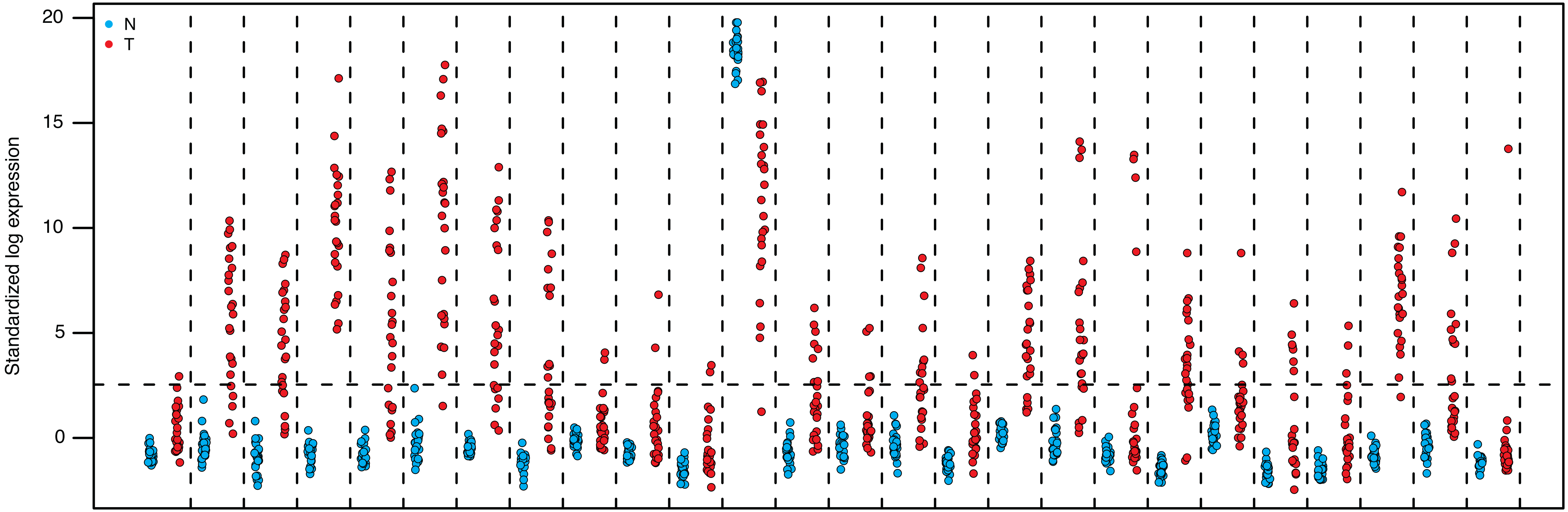






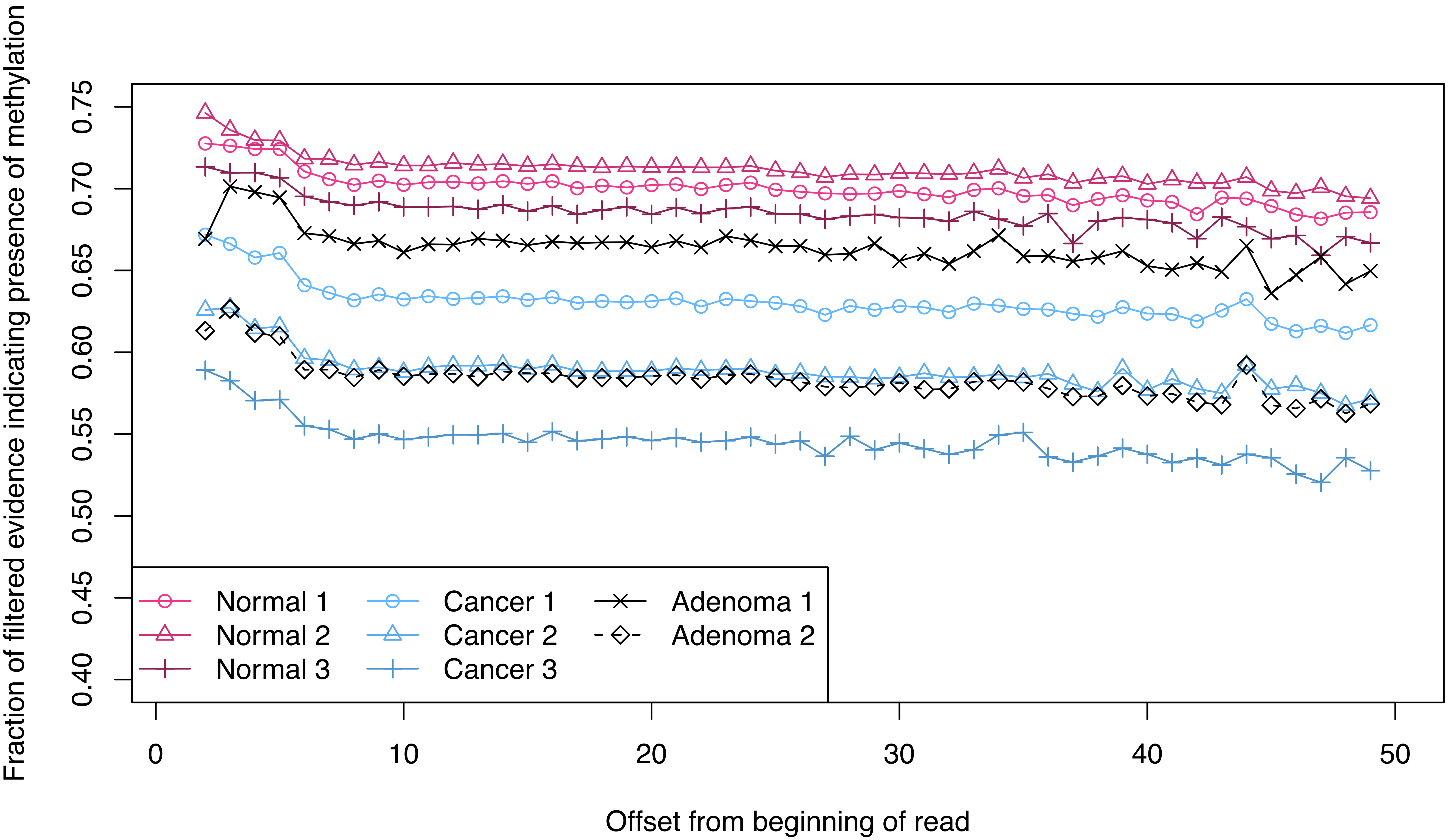


Blocks are enriched for hyper-variables genes



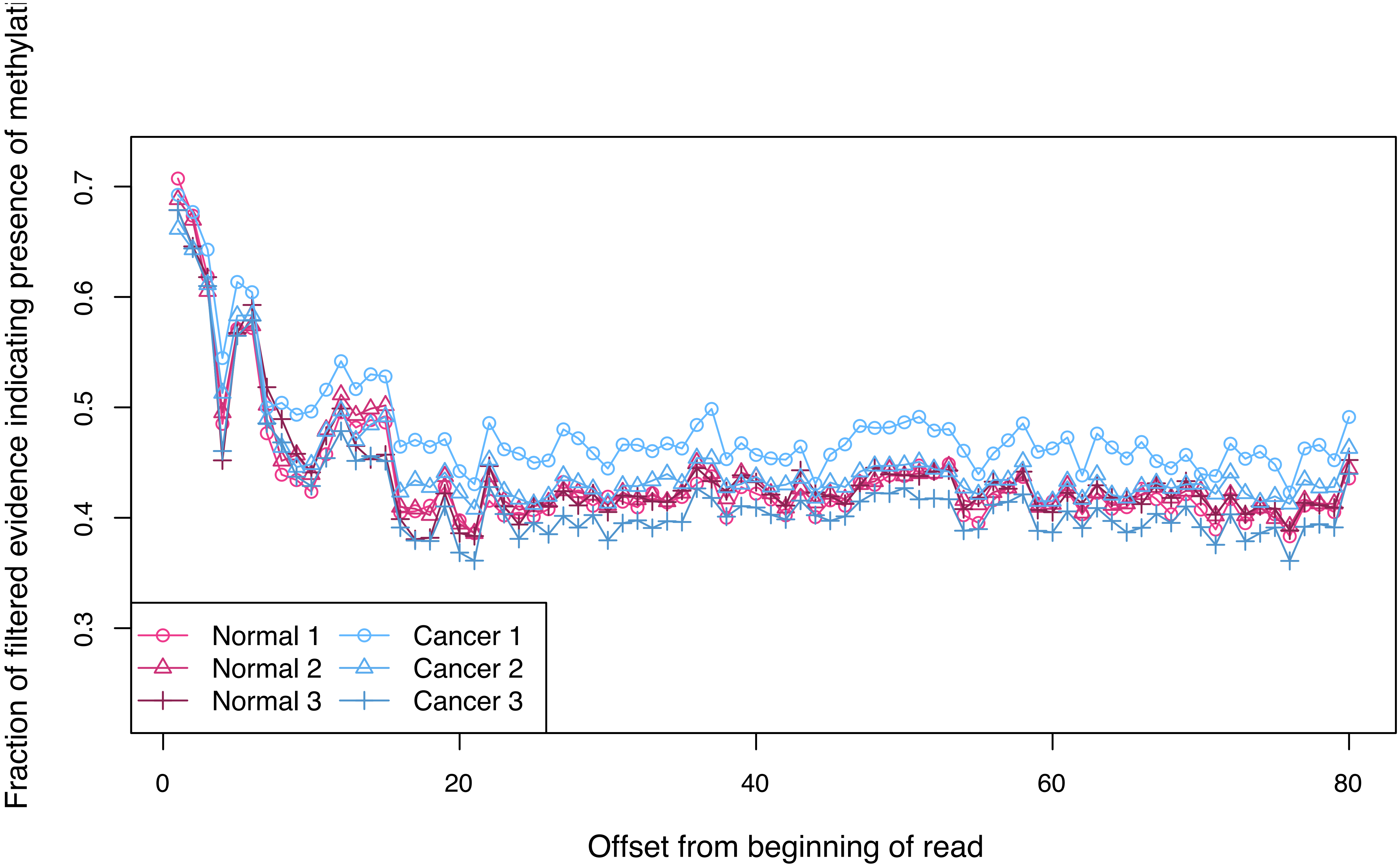
Some of these genes are associated with tumor progression

Quality control: M-bias WGBS



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Quality control: M-bias Capture



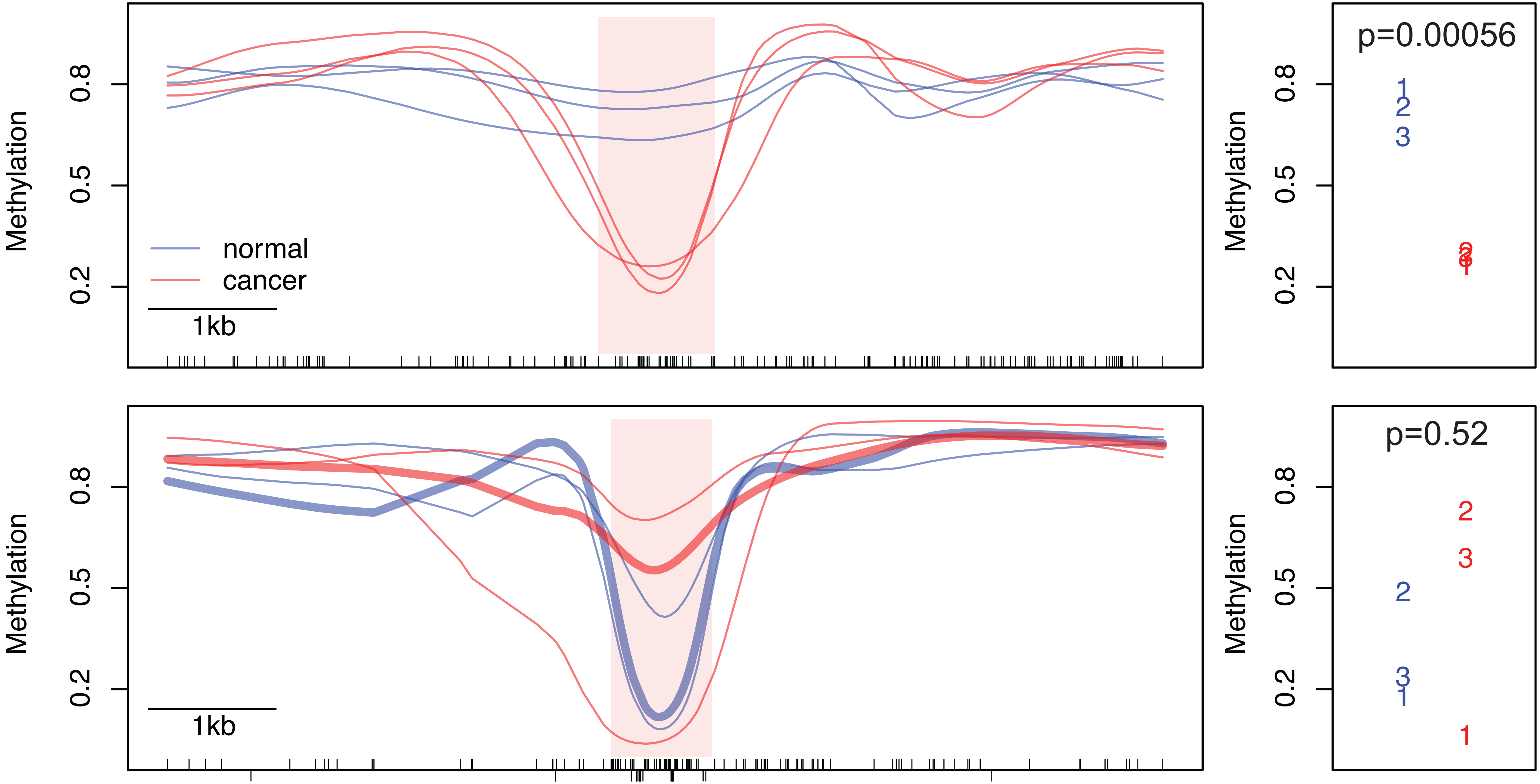
Based on this, we trim 15bp.

This improves the concordance

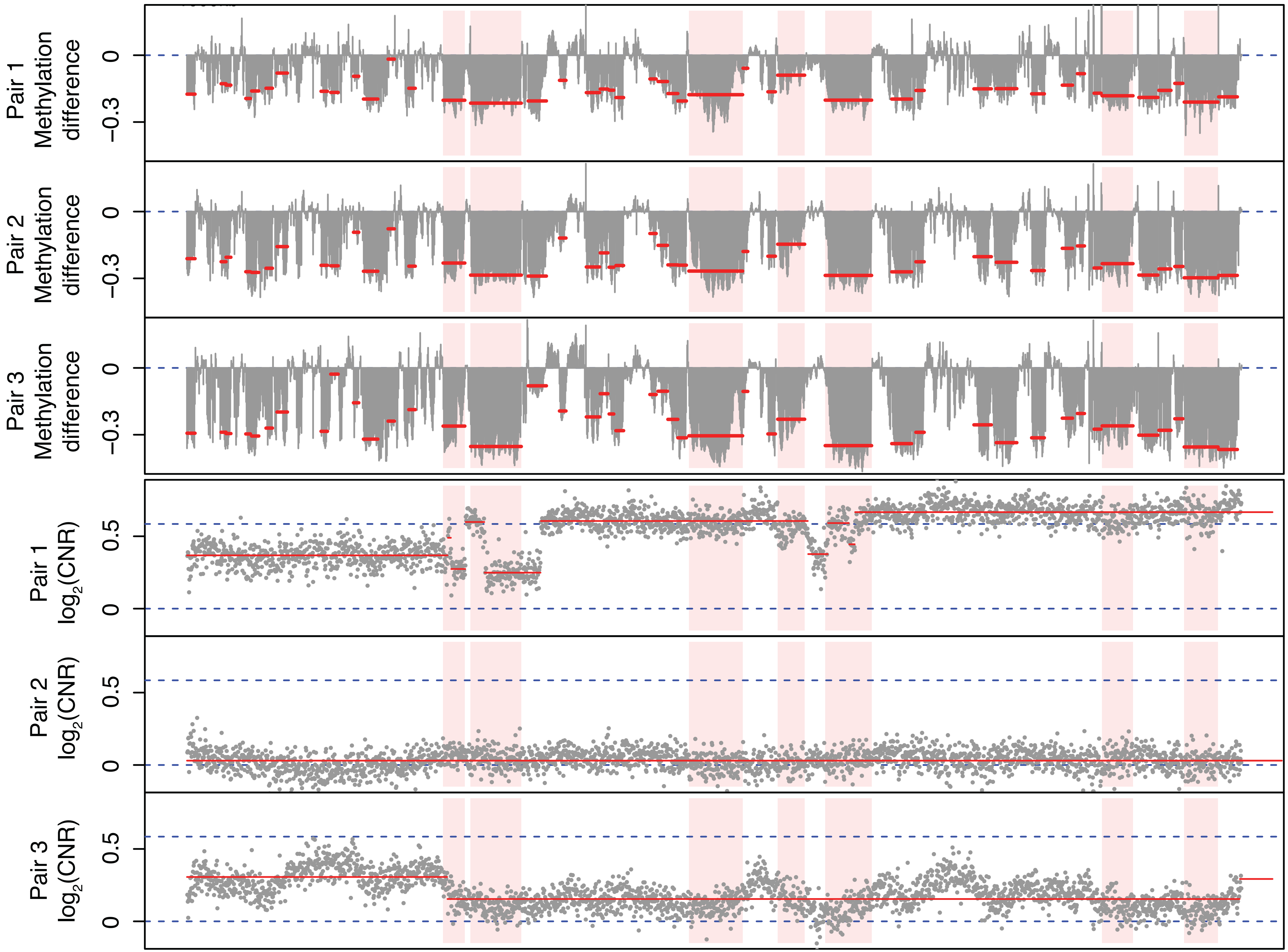
Conclusions

- Large blocks of hypomethylation in cancer
  Global hypomethylation, expression variability
  LOCKs/LADs
- Structural changes (boundaries) in small regions
  Unified framework for shore/islands hypo/hyper methylation
- With our smoothing technique, 4-5x is good enough
  Verified by high coverage padlock bs capture
- biological replicates are very useful
- Quality assessment (M-bias plots)

Advantages of biological replicates



The effect of copy number variation (CNV)



Chr 20

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH