Epigenetics

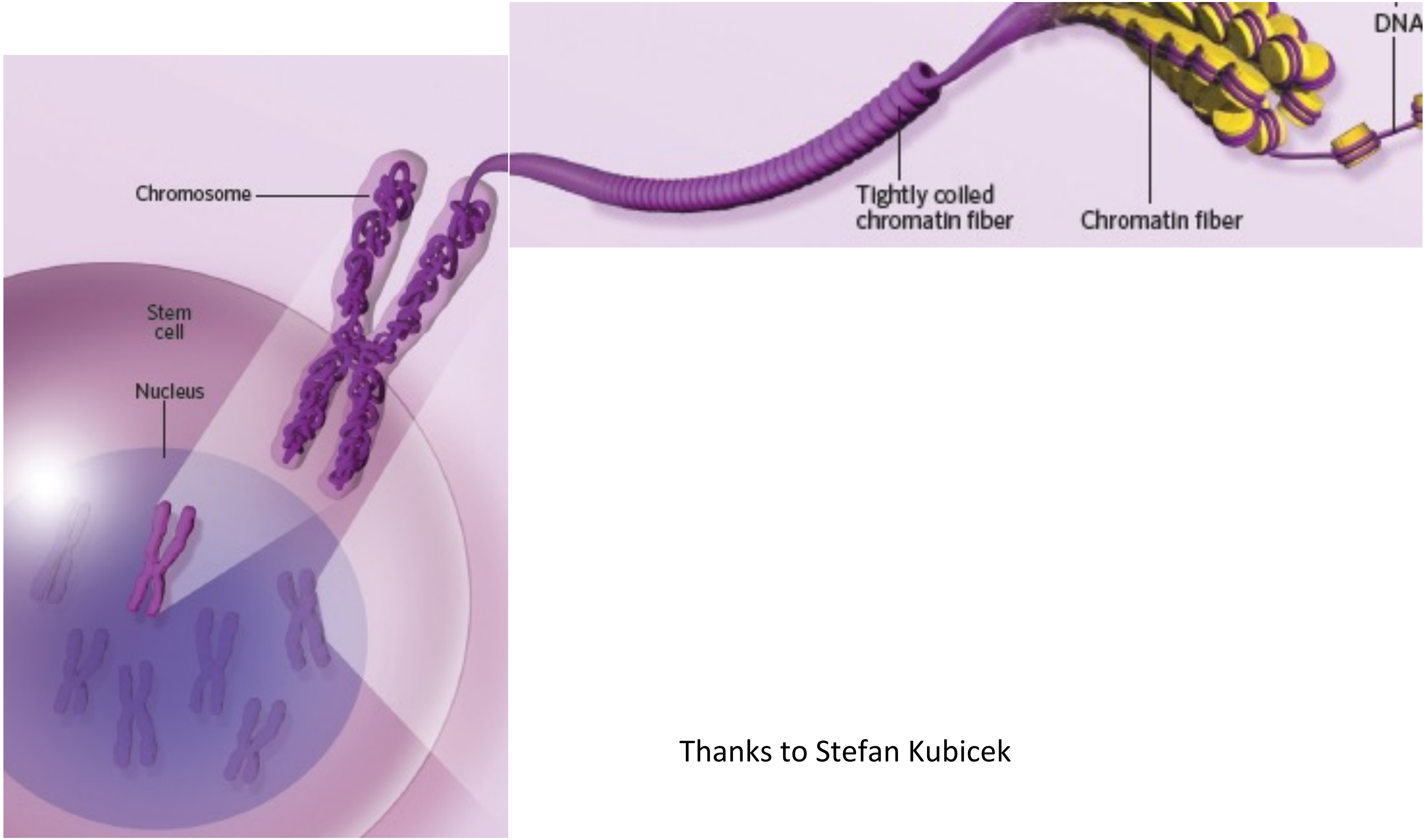
Resources

- Papers from Irizarry (Harvard), Fienberg (JHU), Esteller (Spain)

Overview

- Epigenetic events regulate activities of genes without changing DNA sequence.
- Different genes are expressed depending on the methyl-marks attached to DNA itself and by changes in the structure and/or composition of chromatin.
- Histones are main components of chromatin which come in bundle of 8 units.
- DNA winds around histones (146 bp) to form a structure called nucleosome

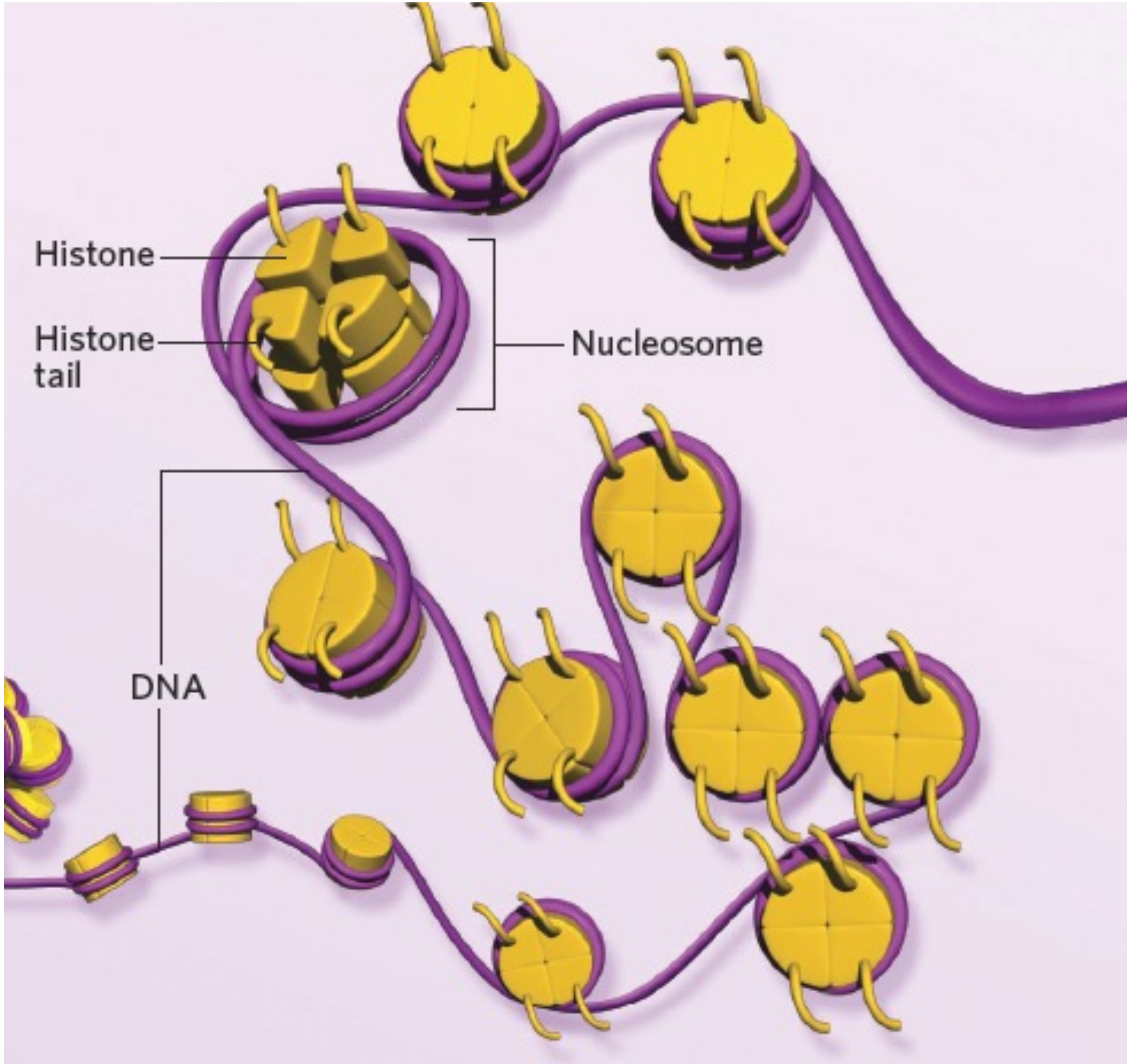
Epigenetics Primer



Thanks to Stefan Kubicek

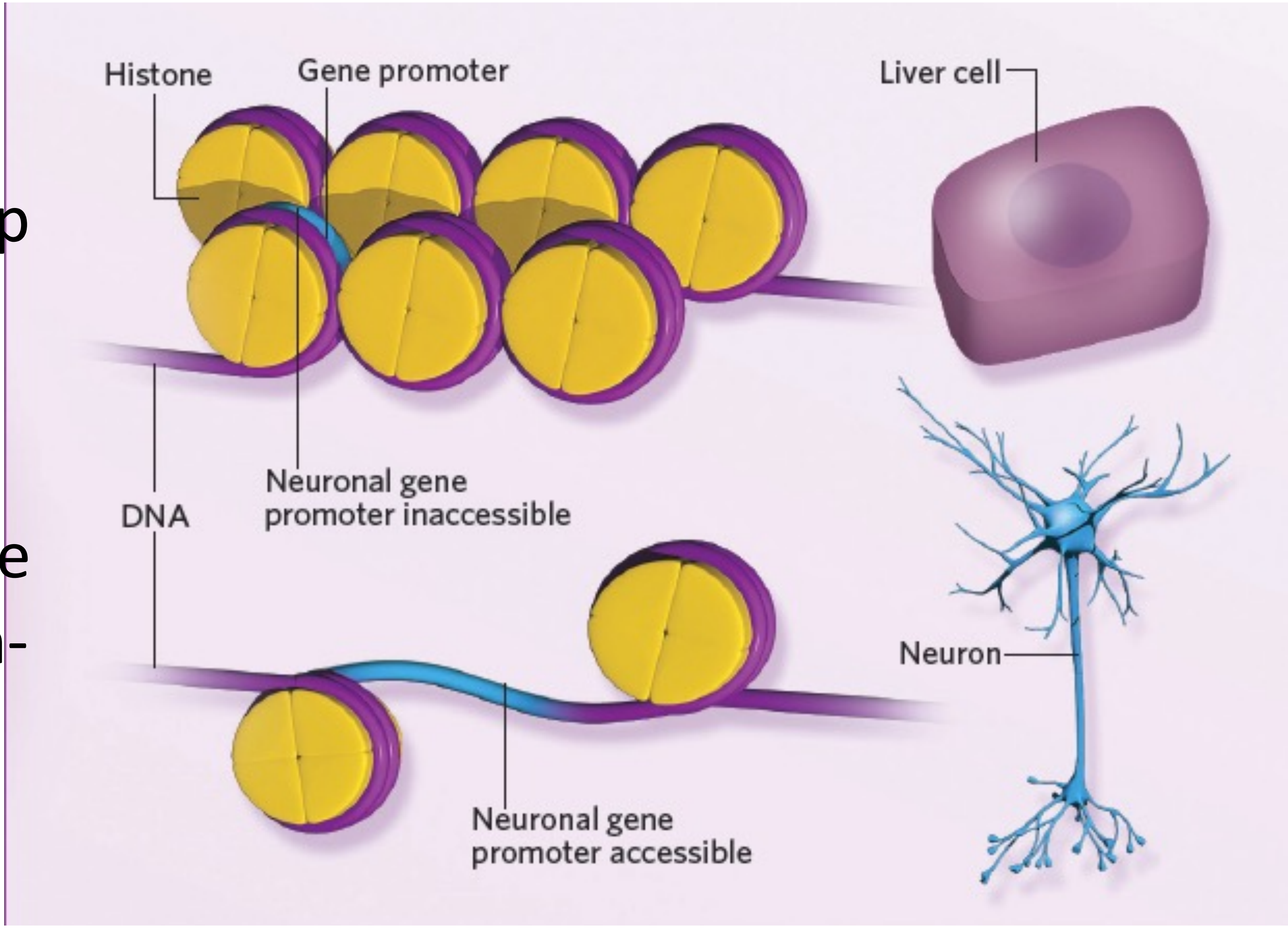
Epigenetic effects on nucleosome:

- Chemical modification via molecular additions to histone tails or DNA
- Position/shape changes via chromatin remodeling proteins
- Variation in histone subtypes



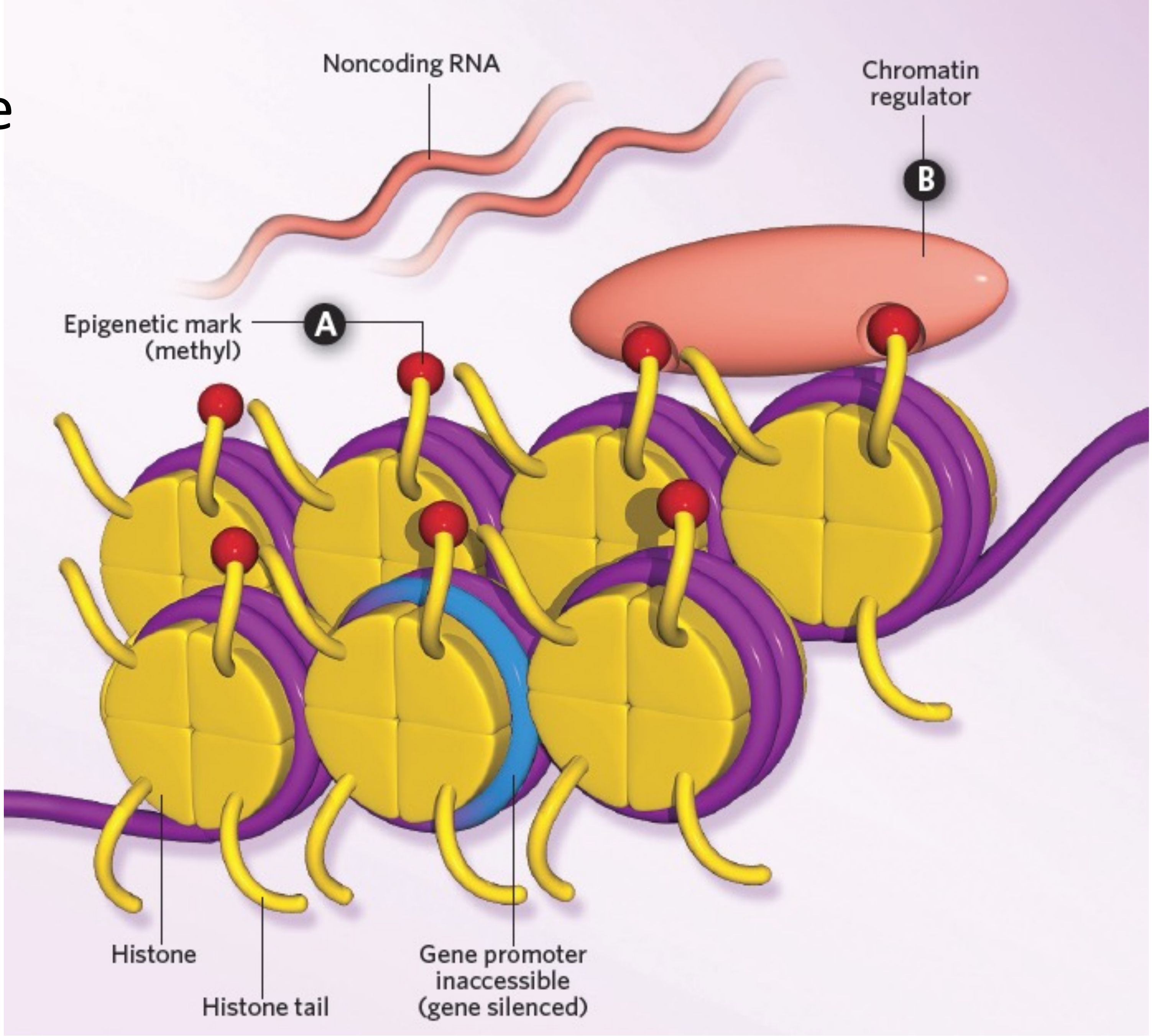
Epigenetic effects critical for cell differentiation:

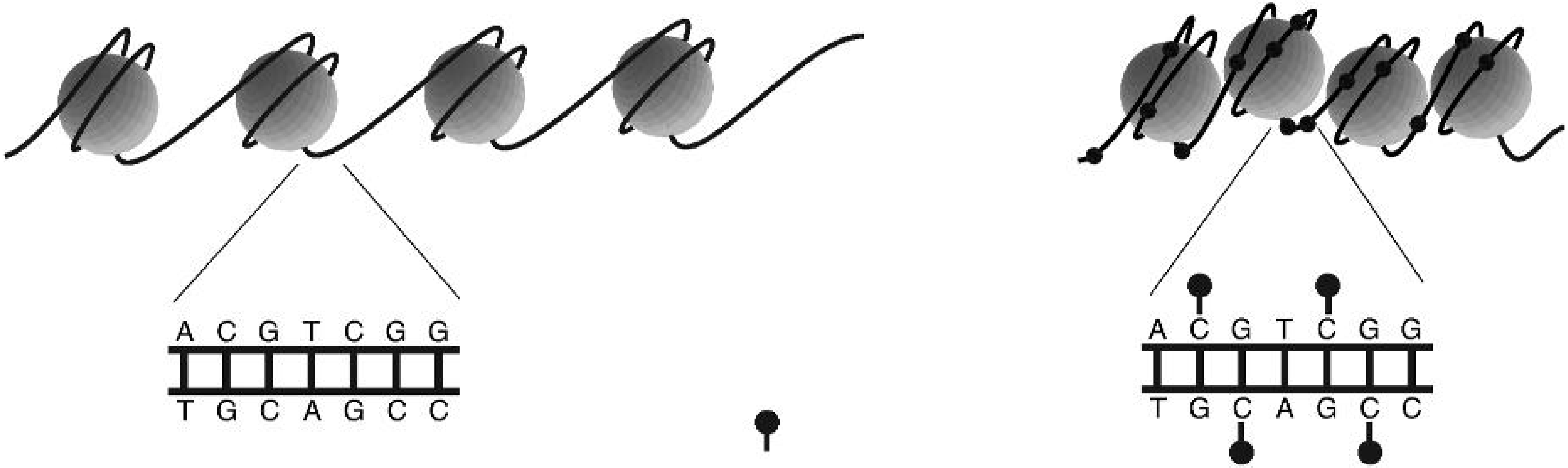
- Neuronal cell expresses genes that help it develop dendrites and axons.
- Liver cell will have same genes marked with epigenetic tags that cause tighter binding of neuron-specific DNA, making it inaccessible to transcription machinery.


Histone
Gene promoter
Liver cell
Neuronal gene promoter inaccessible
DNA
Neuron
Neuronal gene promoter accessible

Epigenetic marks to inactivate genes:

- Methylation at certain positions on histone tails made by histone-modifying enzymes and recognized by other chromatin regulators.
- Noncoding RNAs may be regulating these enzymes.
- Histone modifications that inactivate genes are reversible!!





Unmethylated DNA
Loosely packed
Active gene expression

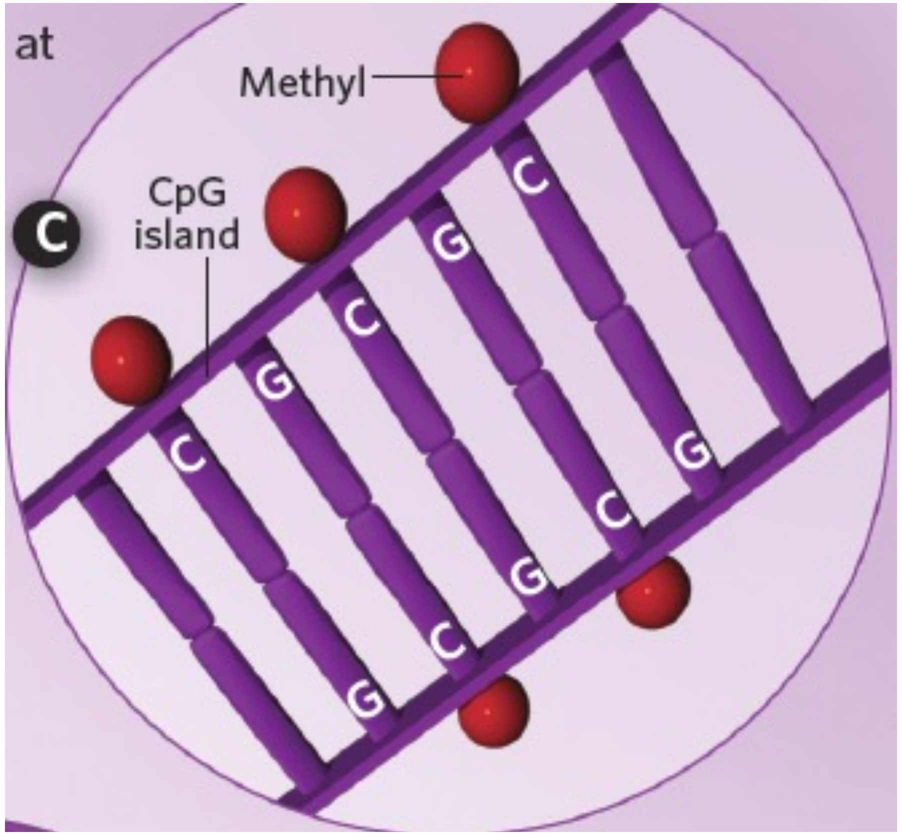
+

Methyl group molecules

=

Methylated DNA
Tightly packed
Silenced gene expression

- DNA methylation occurs at clusters of cytosine (CpG islands) that commonly occur within gene promoters.
- Direct methylation of DNA causes a permanent and heritable change in gene expression.
- Heritability occurs in early stages of development allowing cells to keep irrelevant genes silenced in successive generations.



- How to reactivate silenced genes?
- Acetylation, phosphorylation, methylation on certain positions on a histone tail can cause DNA to unwind, releasing genes that are otherwise inaccessible.
- Modifications occur at accessible positions of histone tail with the help of activating proteins.
- Histone-remodeling complexes can also make genes accessible to transcription.

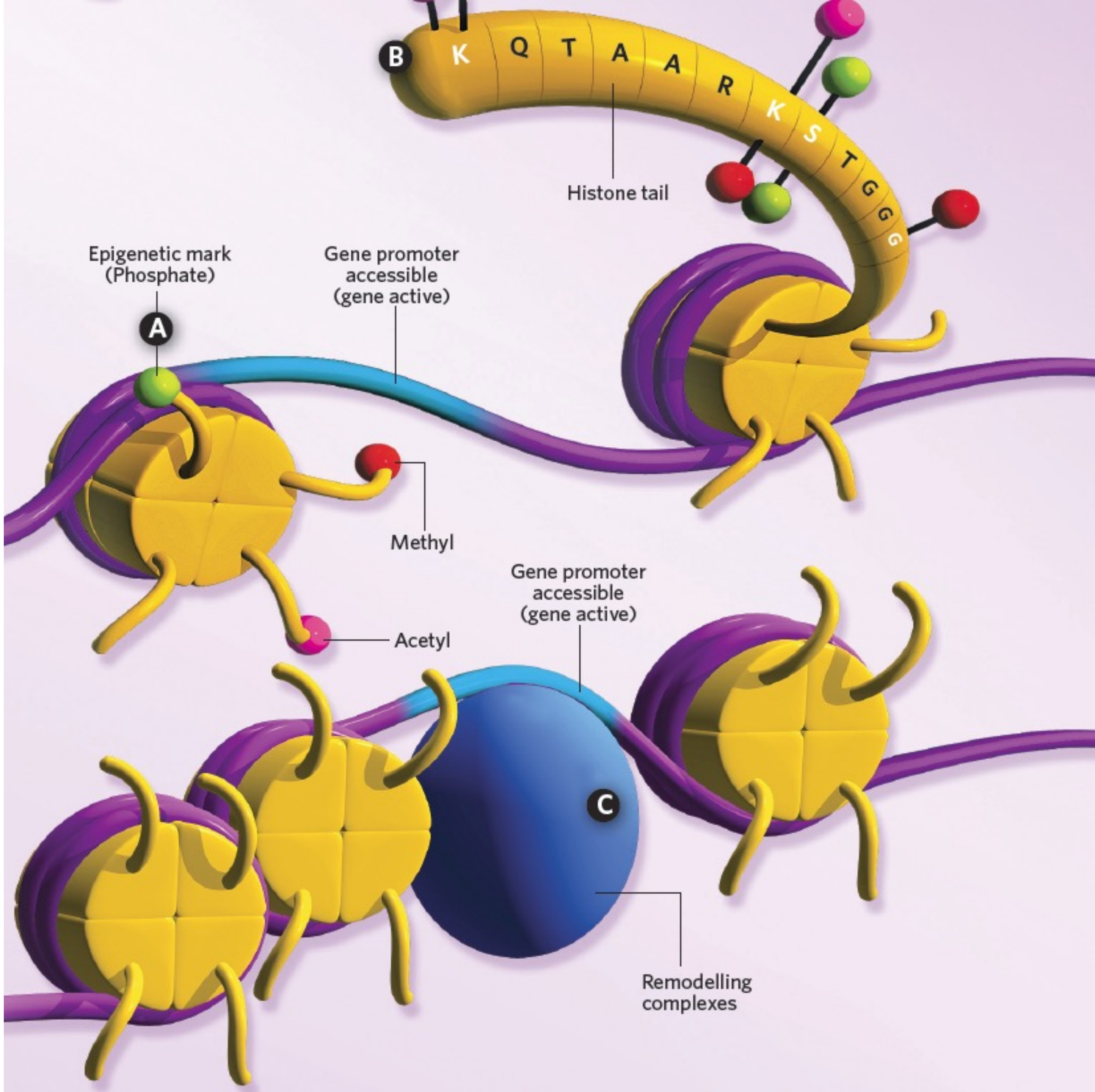
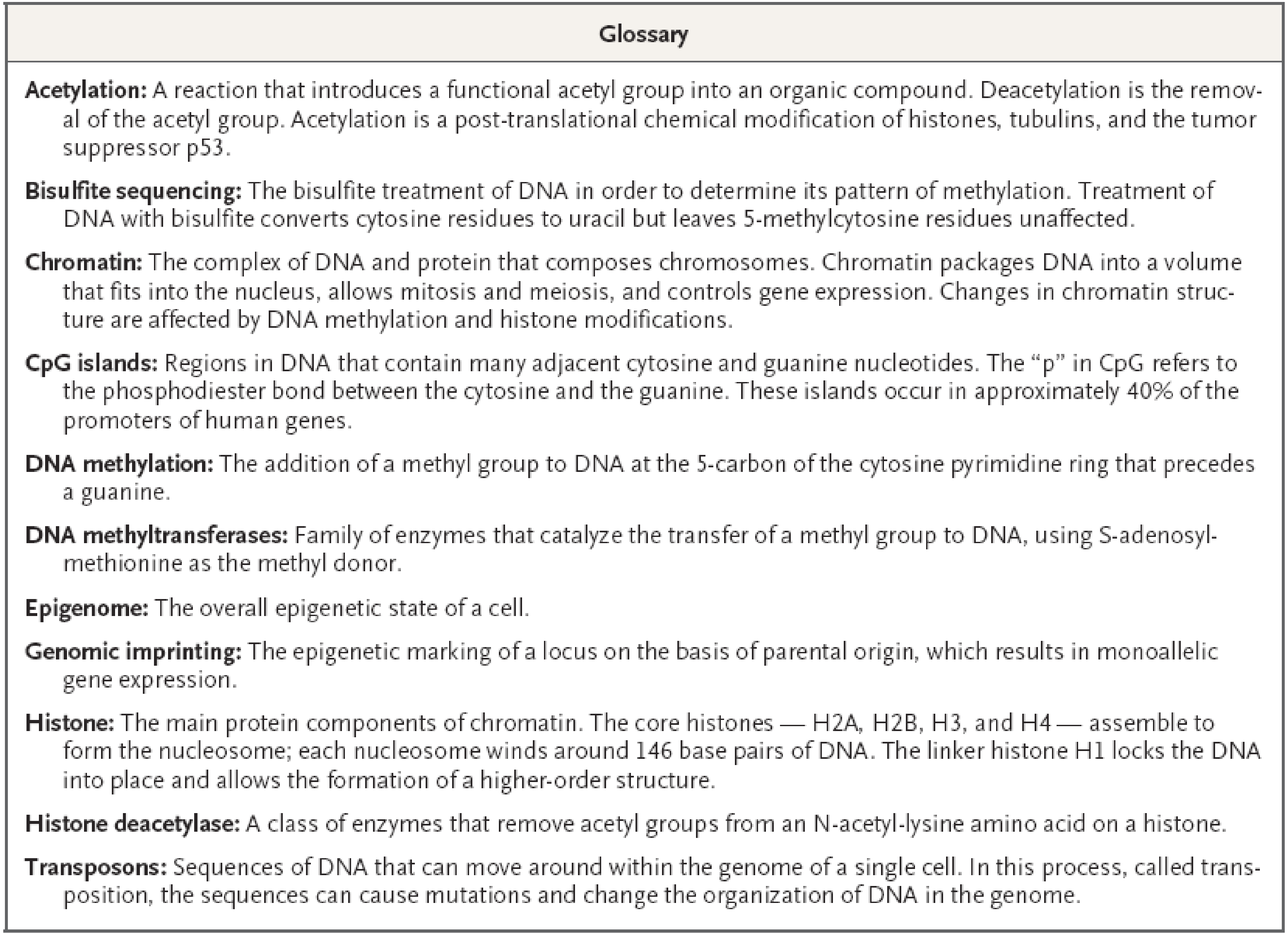

B
K
Q
T
A
A
R
K
S
T
G
G
Histone tail
Epigenetic mark (Phosphate)
A
Gene promoter accessible (gene active)
Methyl
Acetyl
Gene promoter accessible (gene active)
C
Remodelling complexes

Table 1. Glossary of Terms

Term	Definition
5-methylcytosine	Result of DNA methylation at a cytosine (usually at a CpG site)
Antisense oligonucleotide	Group of DNA nucleotides that complementarily binds to mRNA and prevents translation to protein
Chromatin	Groups of nucleosomes whose configuration (euchromatin vs heterochromatin) controls access of gene transcription machinery to genomic DNA
CpG site/island	Site within the genome where a cytosine is immediately followed by a guanine, linked by a phosphate group. This is often a potential site for DNA methylation. Areas with high percentages of CpG sites are termed *CpG islands*.
Epigenetics	Heritable changes in gene expression that do not involve variation in DNA sequence and are modifiable
Histone	Group of eight proteins (usually including H2A, H2B, H3 and H4) that helps to modify accessibility of DNA for gene transcription
Histone deacetylase (HDAC)	Enzyme that removes acetyl groups from histones, changing the epigenetic configuration
Messenger RNA (mRNA)	RNA formed by cellular transcription machinery that, after editing, is translated to protein
Methyltransferase	Enzyme that adds methyl groups to, for example, DNA or histones
microRNA (miRNA)	Group of ~22 RNA nucleotides that (within a protein complex) complementarily binds to target mRNA and prevents translation to protein
Nucleosome	Arrangement of genomic DNA wrapped around a histone octamer (eight proteins)
Promoter	Genomic region upstream of a gene that initiates gene transcription



Glossary

Acetylation: A reaction that introduces a functional acetyl group into an organic compound. Deacetylation is the removal of the acetyl group. Acetylation is a post-translational chemical modification of histones, tubulins, and the tumor suppressor p53.

Bisulfite sequencing: The bisulfite treatment of DNA in order to determine its pattern of methylation. Treatment of DNA with bisulfite converts cytosine residues to uracil but leaves 5-methylcytosine residues unaffected.

Chromatin: The complex of DNA and protein that composes chromosomes. Chromatin packages DNA into a volume that fits into the nucleus, allows mitosis and meiosis, and controls gene expression. Changes in chromatin structure are affected by DNA methylation and histone modifications.

CpG islands: Regions in DNA that contain many adjacent cytosine and guanine nucleotides. The “p” in CpG refers to the phosphodiester bond between the cytosine and the guanine. These islands occur in approximately 40% of the promoters of human genes.

DNA methylation: The addition of a methyl group to DNA at the 5-carbon of the cytosine pyrimidine ring that precedes a guanine.

DNA methyltransferases: Family of enzymes that catalyze the transfer of a methyl group to DNA, using S-adenosylmethionine as the methyl donor.

Epigenome: The overall epigenetic state of a cell.

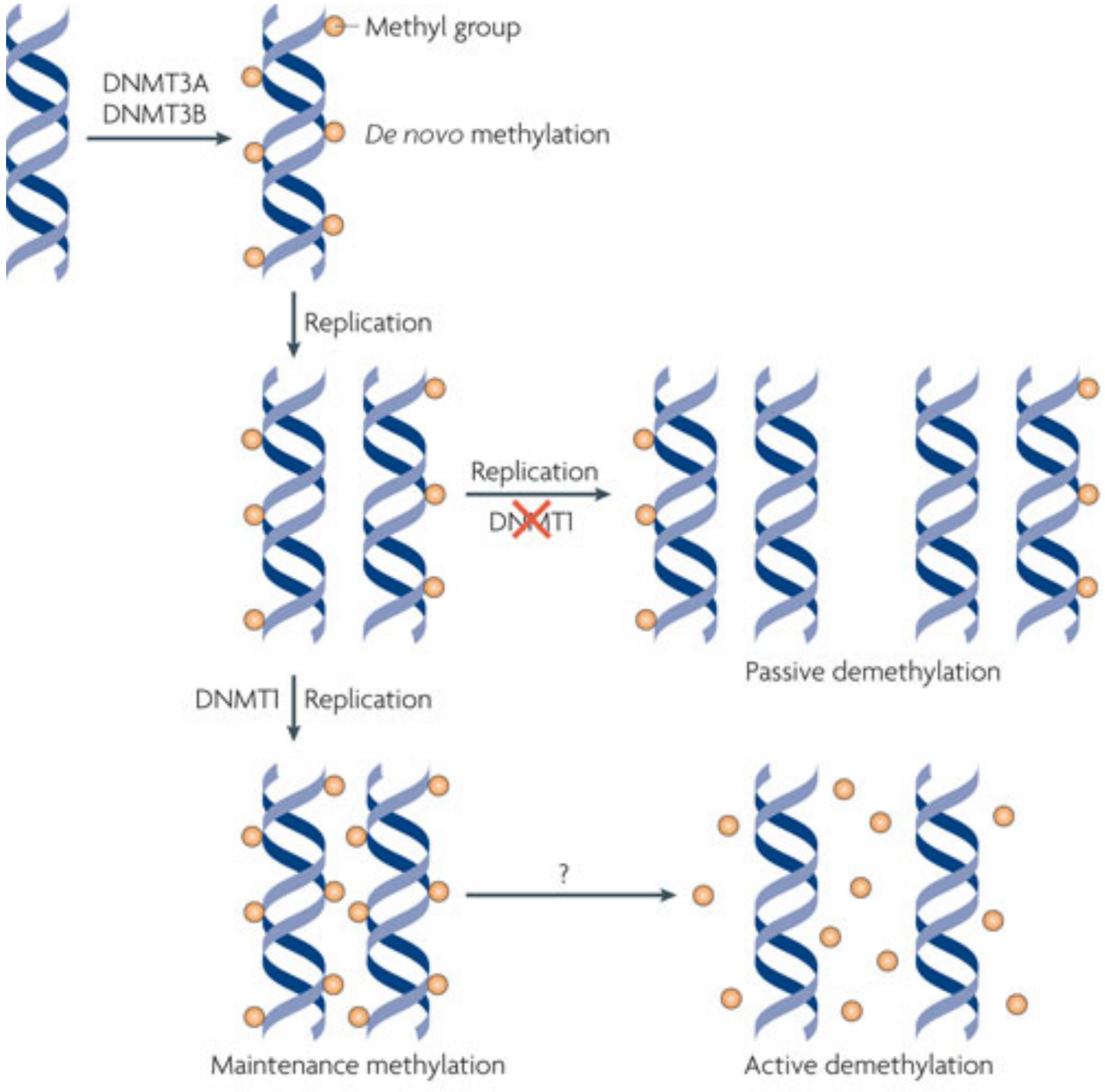
Genomic imprinting: The epigenetic marking of a locus on the basis of parental origin, which results in monoallelic gene expression.

Histone: The main protein components of chromatin. The core histones — H2A, H2B, H3, and H4 — assemble to form the nucleosome; each nucleosome winds around 146 base pairs of DNA. The linker histone H1 locks the DNA into place and allows the formation of a higher-order structure.

Histone deacetylase: A class of enzymes that remove acetyl groups from an N-acetyl-lysine amino acid on a histone.

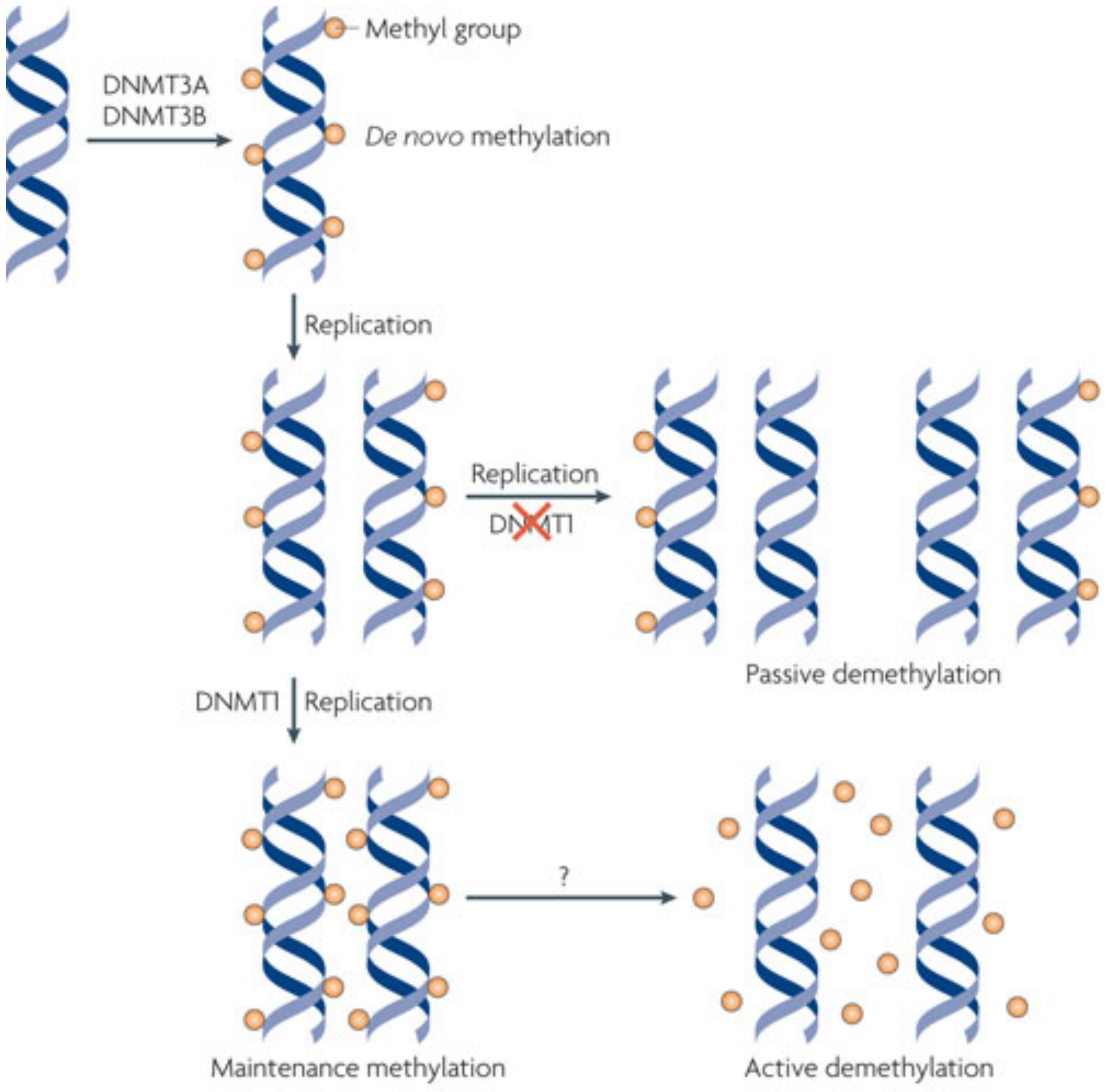
Transposons: Sequences of DNA that can move around within the genome of a single cell. In this process, called transposition, the sequences can cause mutations and change the organization of DNA in the genome.

- During early development, methylation patterns are initially established by the *de novo* DNA methyltransferases: DNMT3A and DNMT3B



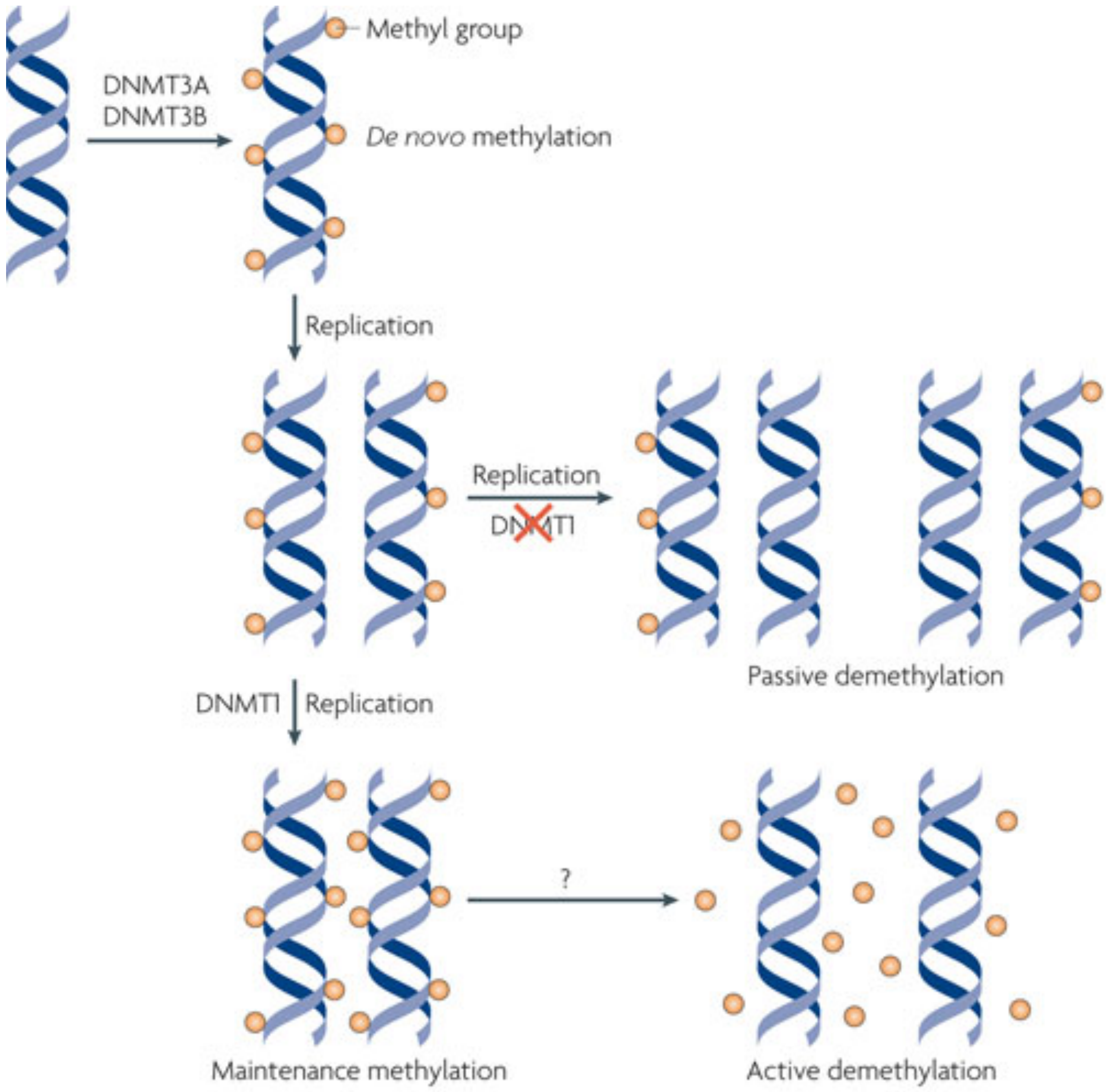
Nature Reviews | Molecular Cell Biology

- When DNA replication and cell division occurs, methyl marks are maintained in daughter cells by the maintenance methyltransferase, DNMT1, which has a preference for hemi-methylated DNA.



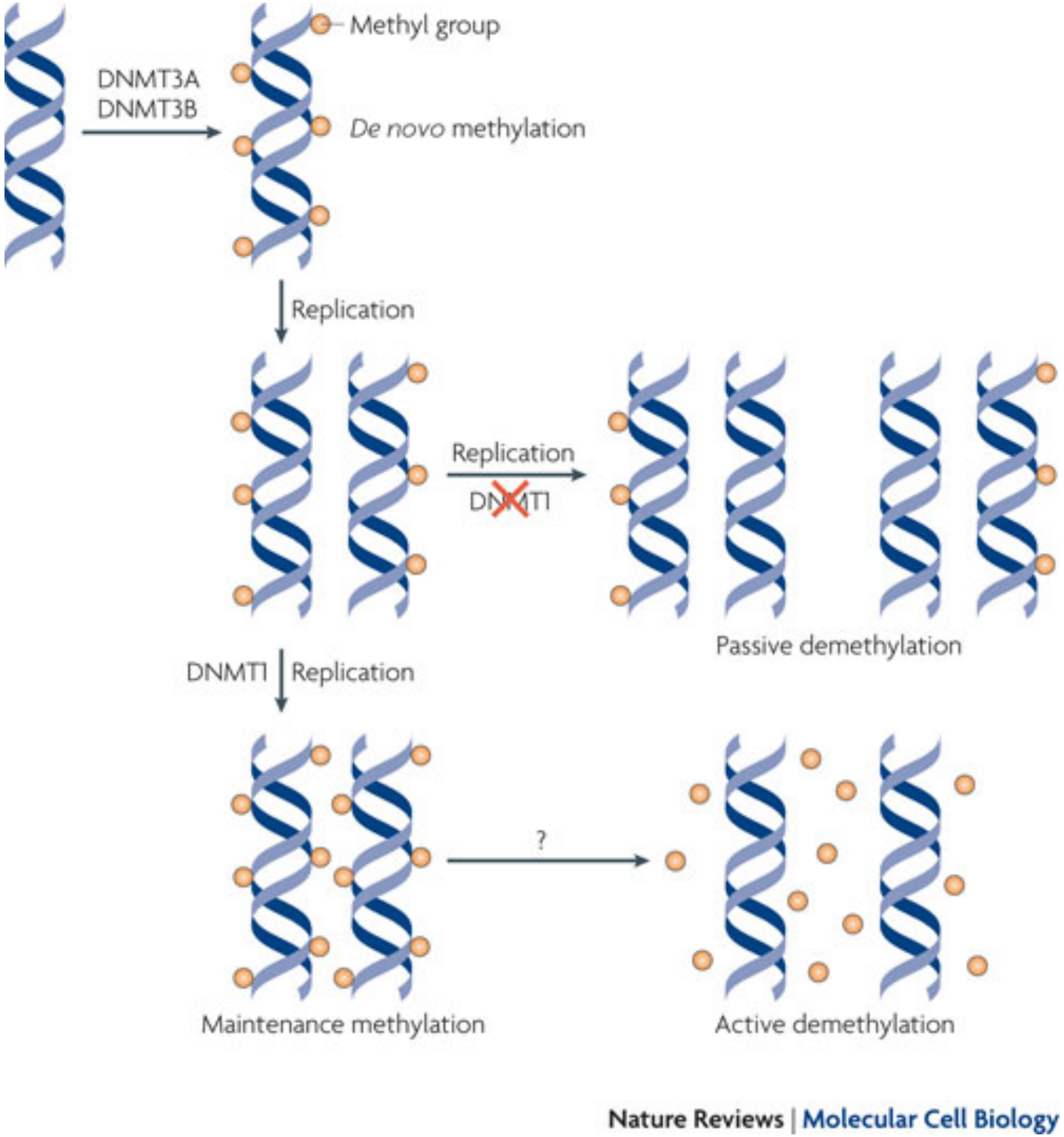
Nature Reviews | Molecular Cell Biology

- If DNMT1 is inhibited or absent when the cell divides, the newly synthesized strand of DNA will not be methylated and successive rounds of cell division will result in passive demethylation.



Nature Reviews | Molecular Cell Biology

- Active demethylation can occur through enzyme replacement of 5-methylcytosine (5meC) with C.



Nature Reviews | Molecular Cell Biology

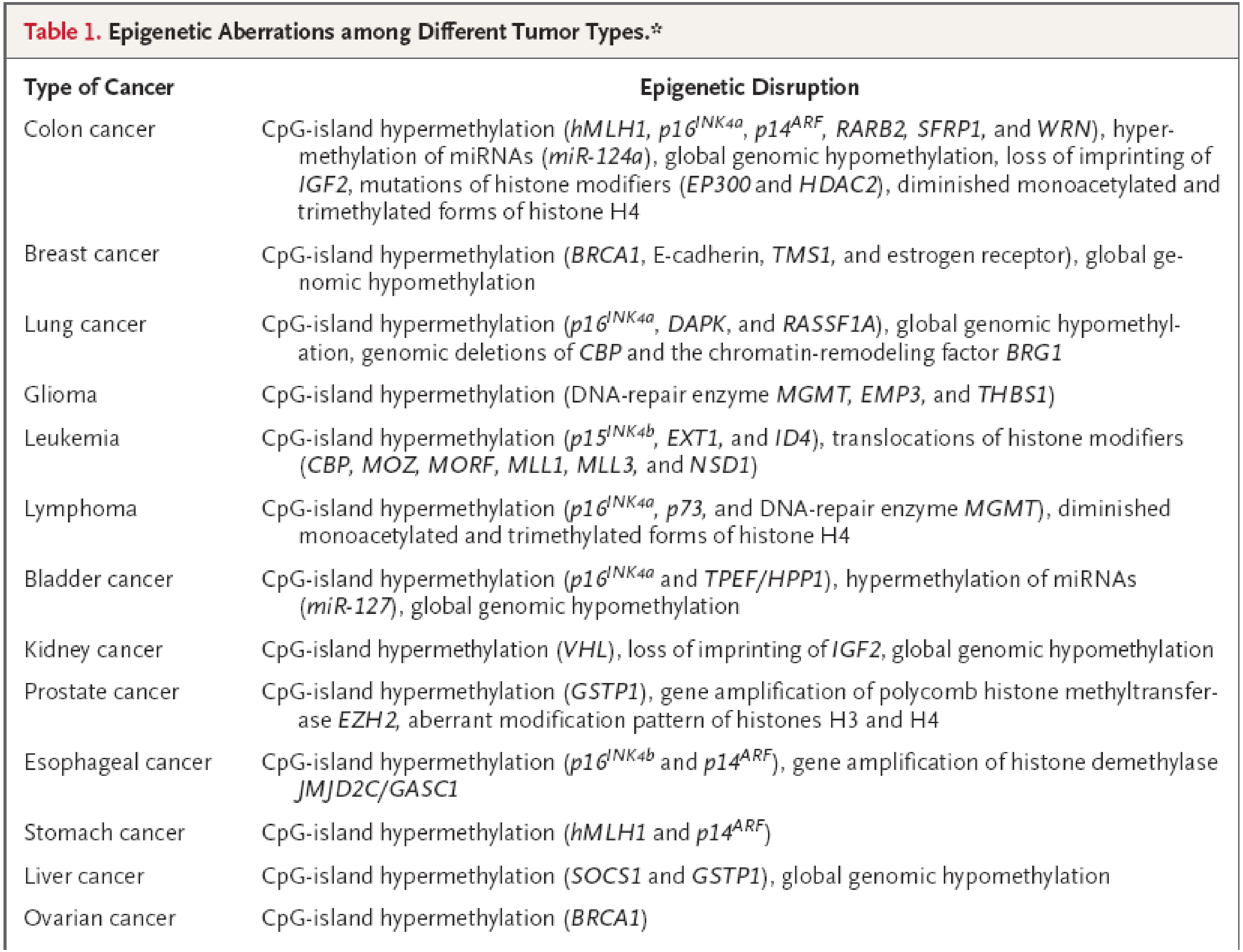
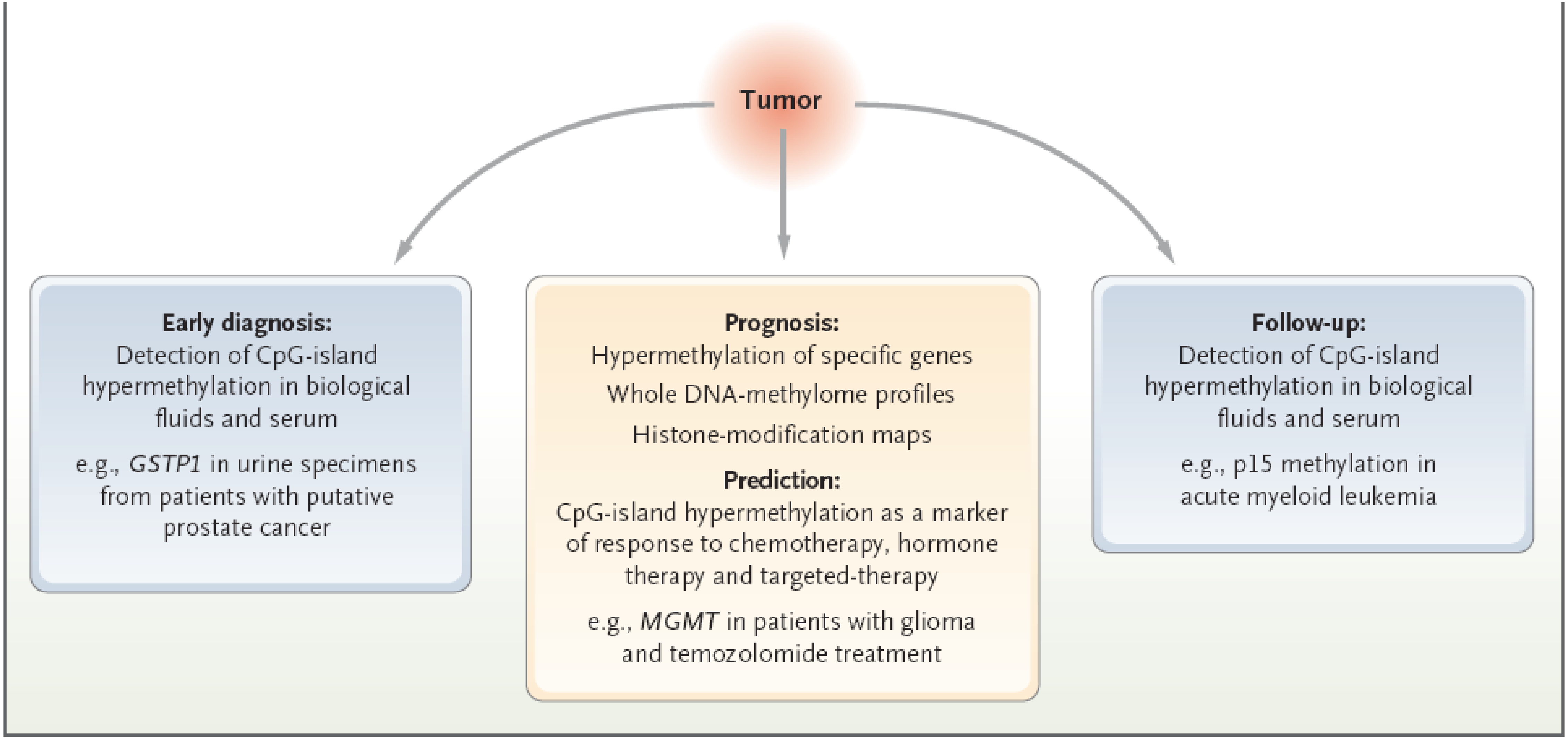


Table 1. Epigenetic Aberrations among Different Tumor Types.*

Type of Cancer	Epigenetic Disruption
Colon cancer	CpG-island hypermethylation (*hMLH1*, $p16^{INK4a}$, $p14^{ARF}$, *RARB2*, *SFRP1*, and *WRN*), hypermethylation of miRNAs (*miR-124a*), global genomic hypomethylation, loss of imprinting of *IGF2*, mutations of histone modifiers (*EP300* and *HDAC2*), diminished monoacetylated and trimethylated forms of histone H4
Breast cancer	CpG-island hypermethylation (*BRCA1*, E-cadherin, *TMS1*, and estrogen receptor), global genomic hypomethylation
Lung cancer	CpG-island hypermethylation ($p16^{INK4a}$, *DAPK*, and *RASSF1A*), global genomic hypomethylation, genomic deletions of *CBP* and the chromatin-remodeling factor *BRG1*
Glioma	CpG-island hypermethylation (DNA-repair enzyme *MGMT*, *EMP3*, and *THBS1*)
Leukemia	CpG-island hypermethylation ($p15^{INK4b}$, *EXT1*, and *ID4*), translocations of histone modifiers (*CBP*, *MOZ*, *MORF*, *MLL1*, *MLL3*, and *NSD1*)
Lymphoma	CpG-island hypermethylation ($p16^{INK4a}$, *p73*, and DNA-repair enzyme *MGMT*), diminished monoacetylated and trimethylated forms of histone H4
Bladder cancer	CpG-island hypermethylation ($p16^{INK4a}$ and *TPEF/HPP1*), hypermethylation of miRNAs (*miR-127*), global genomic hypomethylation
Kidney cancer	CpG-island hypermethylation (*VHL*), loss of imprinting of *IGF2*, global genomic hypomethylation
Prostate cancer	CpG-island hypermethylation (*GSTP1*), gene amplification of polycomb histone methyltransferase *EZH2*, aberrant modification pattern of histones H3 and H4
Esophageal cancer	CpG-island hypermethylation ($p16^{INK4b}$ and $p14^{ARF}$), gene amplification of histone demethylase *JMJD2C/GASC1*
Stomach cancer	CpG-island hypermethylation (*hMLH1* and $p14^{ARF}$)
Liver cancer	CpG-island hypermethylation (*SOCS1* and *GSTP1*), global genomic hypomethylation
Ovarian cancer	CpG-island hypermethylation (*BRCA1*)



Tumor

Early diagnosis:
Detection of CpG-island hypermethylation in biological fluids and serum

e.g., *GSTP1* in urine specimens from patients with putative prostate cancer

Prognosis:
Hypermethylation of specific genes

Whole DNA-methylome profiles

Histone-modification maps

Prediction:
CpG-island hypermethylation as a marker of response to chemotherapy, hormone therapy and targeted-therapy

e.g., *MGMT* in patients with glioma and temozolomide treatment

Follow-up:
Detection of CpG-island hypermethylation in biological fluids and serum

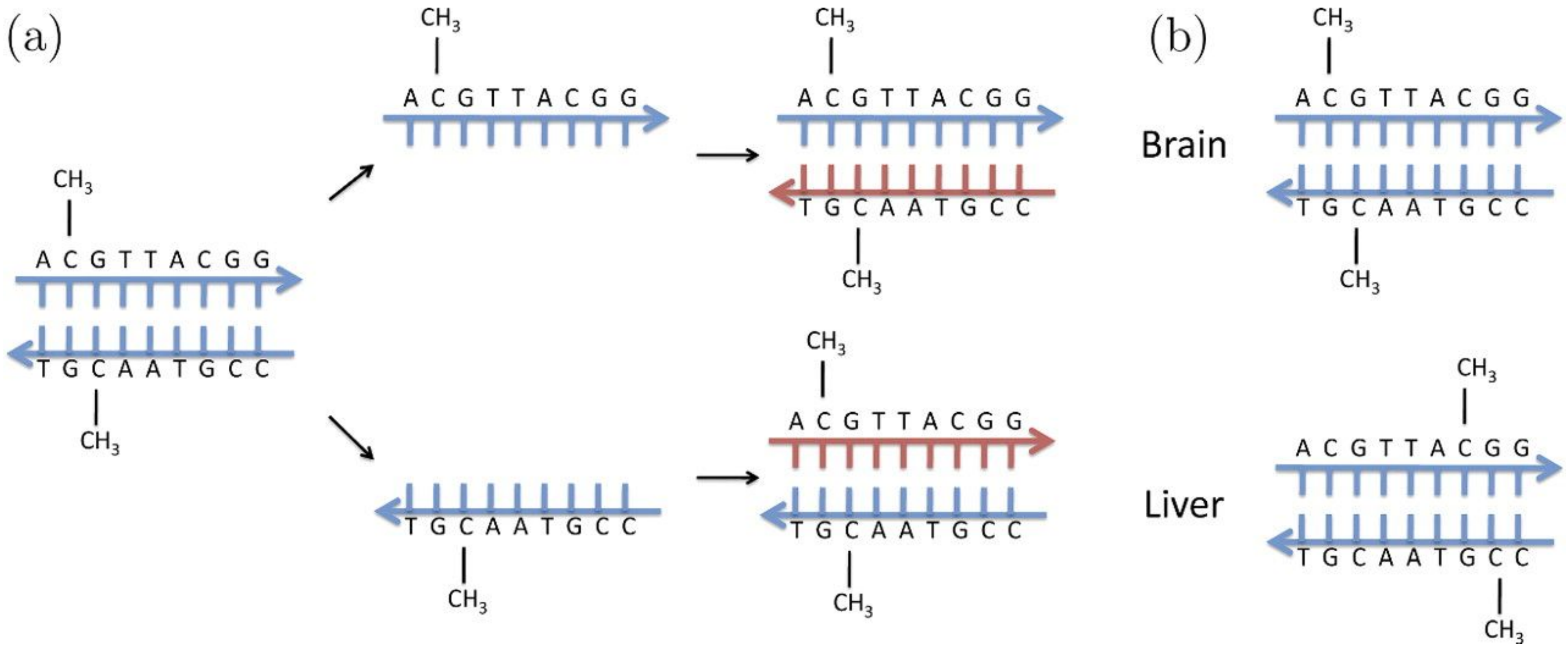
e.g., p15 methylation in acute myeloid leukemia

- Sensitive and quantitative detection of hypermethylated tumor-suppressor genes in all types of biologic fluids and biopsy specimens.
- Establishment of DNA methylation and histone-modification profiles of the primary tumor specimen might be a valuable tool in determining the prognosis and predicting the patient's response to therapies.

Computational Methods

- CpG island detection
- Bisulfite sequencing + alignment
    - bSmooth
- Deconvolution for heterogeneity
- CpG shores and cancer epi (possibly next week)

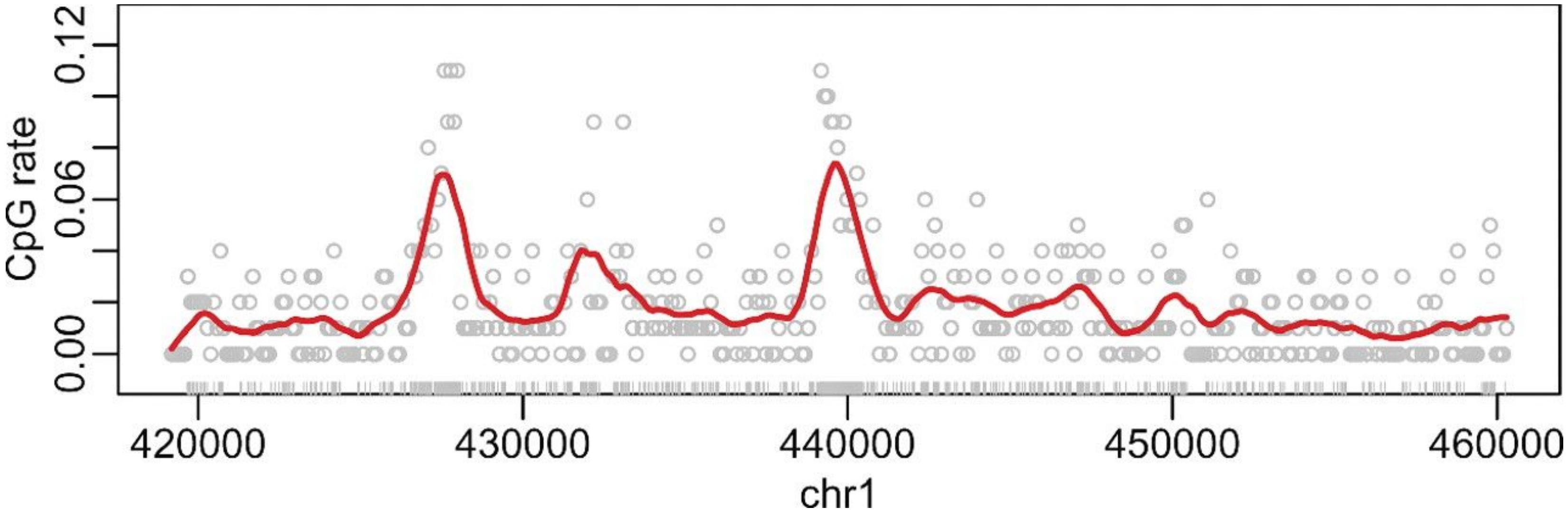
Cartoon illustrating how DNA methylation is inherited in cell division on how it could be involved in tissue differentiation.



Wu H et al. Biostat 2010;11:499-514

Biostatistics

A genomic region of 40 000 bases from chromosome 1 is shown.



A genomic region of 40 000 bases from chromosome 1 is shown. The ticks on the x-axis represent CpG locations. The points represent CpG rates in segments of length 256 bases The curve is the result of a kernel smoother of the points. Approximately 20% of the genome are Cs and 20% are Gs. Thus, we expect about 4% of dinucleotides to be CpG. However, most points are well below rates 4% with 2 clusters well above 4%. The latter are CGI.

Wu H et al. Biostat 2010;11:499-514

Biostatistics

CGI definitions

Gardiner-Garden and Frommer (1987)

- At least 200 bp with the proportion of Gs or Cs, ‘GC’ content, more than 50% and observed to expected CpG ratio (O/E) > 0.6

$$\mathrm{O/E} = \frac{\#\mathrm{CpG}/N}{\#C/N \times \#G/N}$$

- *N* number of bp in the segment under consideration
- UCSC Genome browser maintains an extensive list

CGI definitions

Takai and Jones

- Min length of 500 bp, min GC content >= 55% and O/E >= 0.65
- Alu-repetitive elements get excluded nicely (Alu sequence appears more than 1M times)

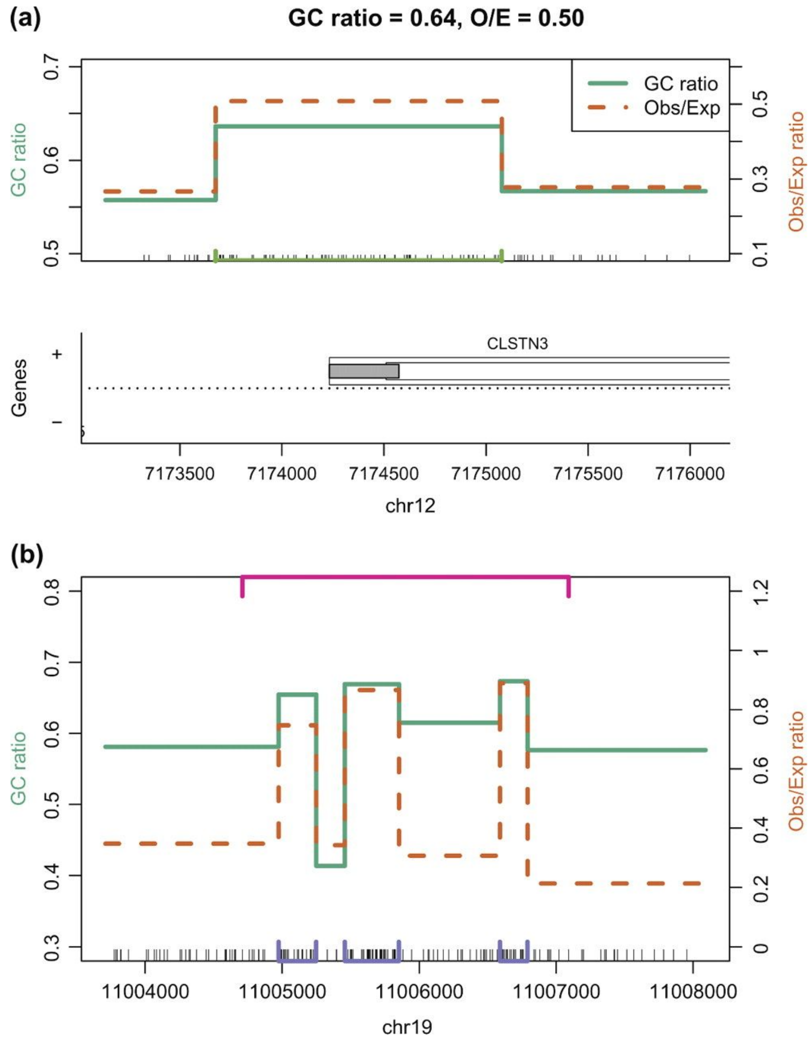
Glass *and others* (2007)

- For every CpG, they recorded the length of a segment needed to cover the nearest 27 CpGs. They then observed that, for certain species, a histogram of these lengths shows a bimodal distribution. The histogram was used to select a cutoff and regions associated with the first mode are defined as CpG "clusters" (their terminology for CGI).

The observed to expected ratio (green) and percentage of G + C (orange) are shown for 2 regions of the human genome.

Limitations of existing methods

The observed to expected ratio (green) and percentage of G + C (orange) are shown for 2 regions of the human genome. CpG clusters are denoted with bars along the bottom or top of the plot. (a) For a region covering the 5′ end of CLSTN3 a CpG dense region that is not in current CGI list is denoted by the lime green bar at the bottom. (b) The top (pink) bar denotes one of Glass and others (2007) CGI that engulfs 3 Genome Browser CGIs (denoted with purple bars at the bottom). The regions between the Genome Browser CGI have low observed to expected ratio.



Wu H et al. Biostat 2010;11:499-514

Biostatistics

HMMs that we studied before require modifications

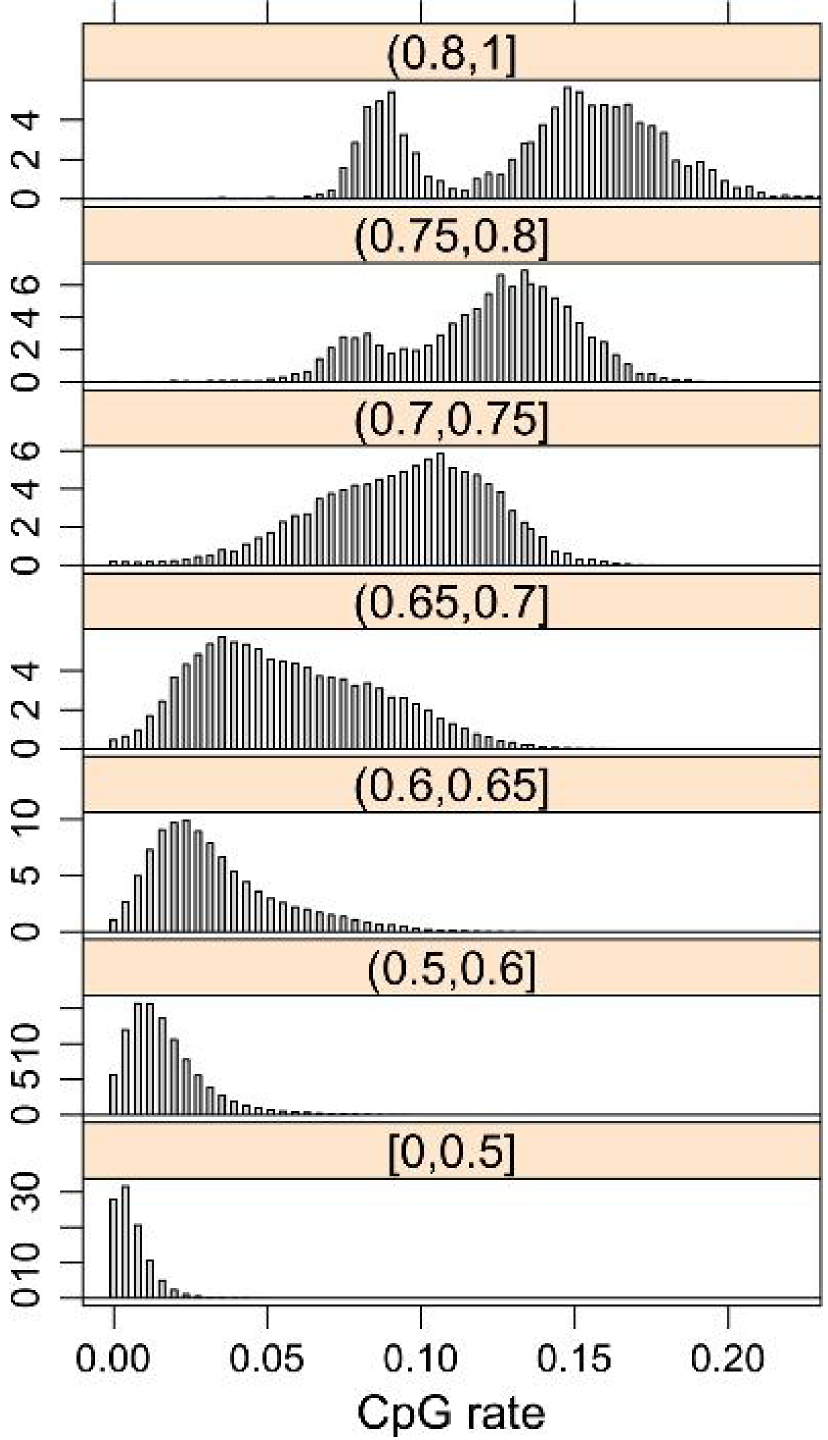
- If CGIs are simply a cluster of CpGs, then a procedure that scans through the genome searching for regions with larger than expected CpG rates would suffice.
- However, the evolutionary theory for CGI motivates a more sophisticated approach.
- Briefly, the human genome is depleted of CpG because the mutation rate for this specific dinucleotide is higher than others (Lander *and others*, 2001).
- CGIs are believed to be the result of certain segments of the genome being somewhat protected from the mechanism that leads to this mutation. This is a possible explanation for the association of CGI and locations relevant to development.

Rethinking CGIs

- This evolutionary argument, based on differing mutation rates, suggest that the fundamental property that defines a CGI is not the CpG density *per se* but the CpG density conditioned on GC content.
- This is because regions that originally had high GC content had more CpG dinucleotides which, even unprotected, resulted in relatively high CpG counts.
- Gardiner-Garden and Frommer's definition, based on the observed to expected ratio as opposed to just the number of CpG, agrees with the above described theory.

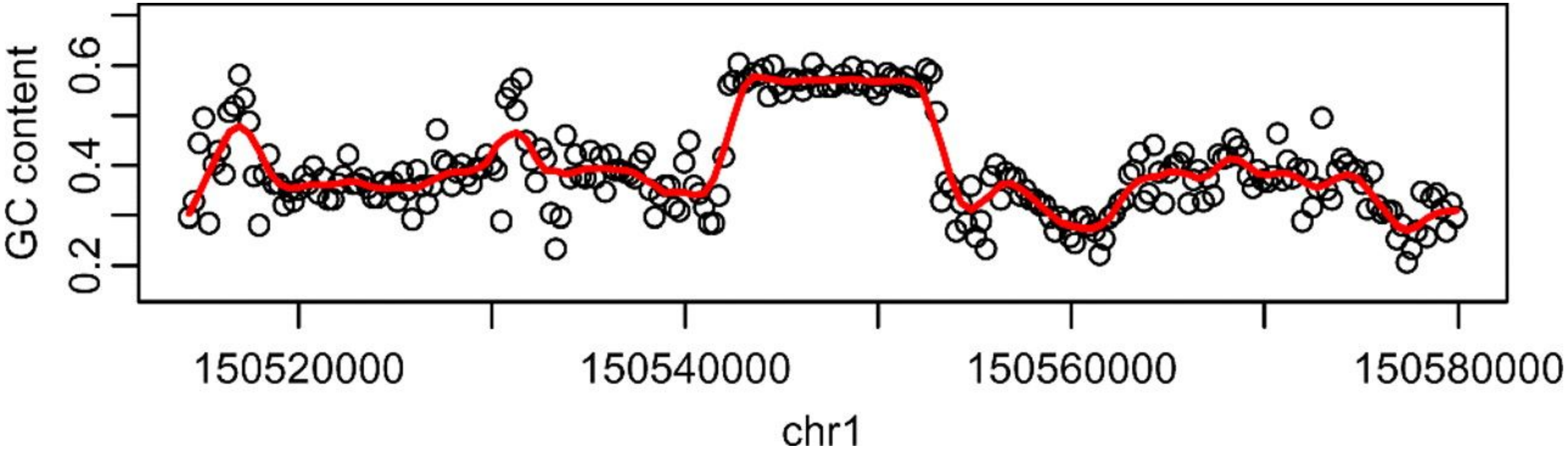
Insight from stratification

Histogram of CpG rates in nonoverlapping genomic segments of length 256 bases, stratified by the GC contents
in each segment. GC content strata is shown on top of each histogram. *x*-axis is the CpG rate. *y*-axis shows the percentage of segments belonging to each CpG rate category.



2-state HMM needed

GC content plots.



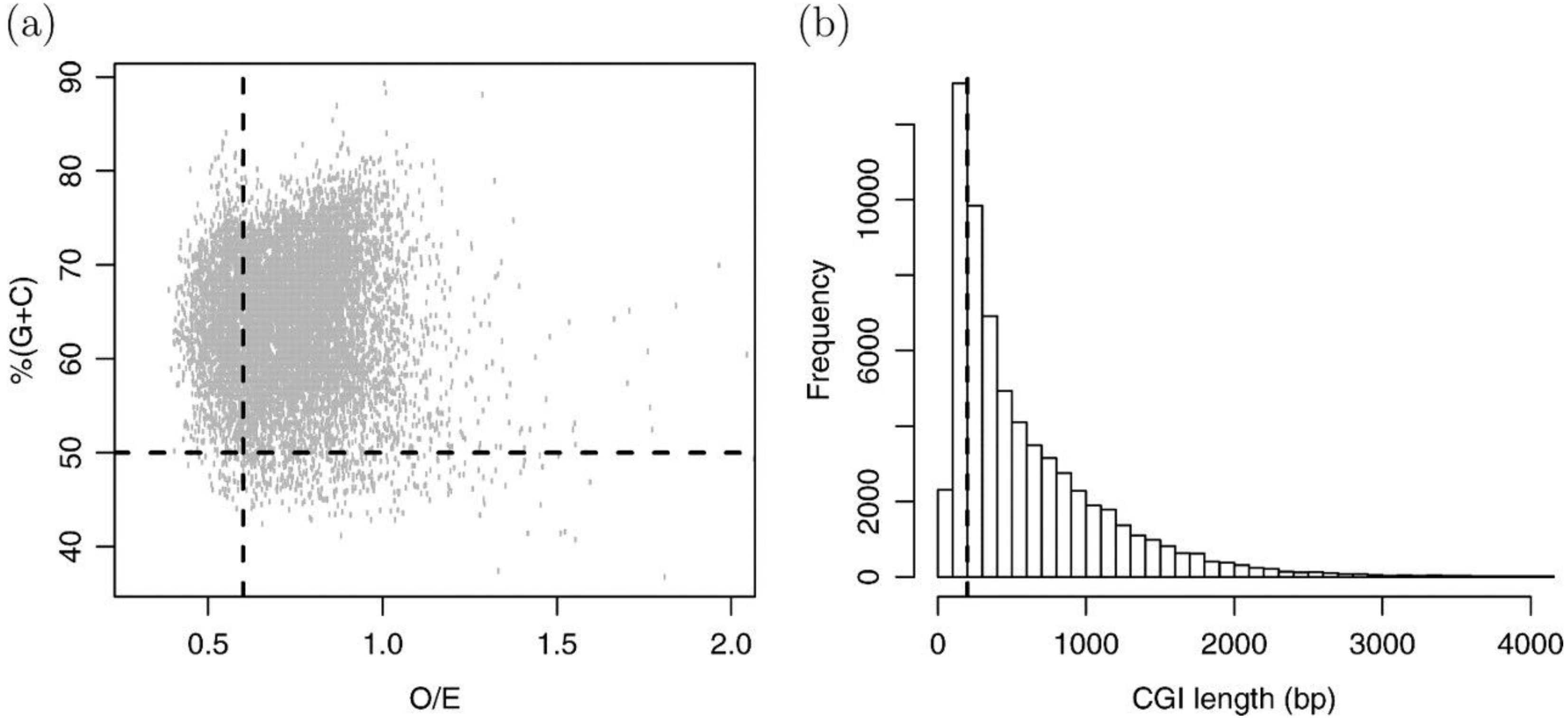
GC content plots. A region with no Alu repeats was divided into nonoverlapping segments of length 256 bases. The points are the GC content of each segment. The curve is the result of a kernel smoother of the points.

Wu H et al. Biostat 2010;11:499-514

Biostatistics

Statistical characteristics of model-based CGI list for human (hg18).

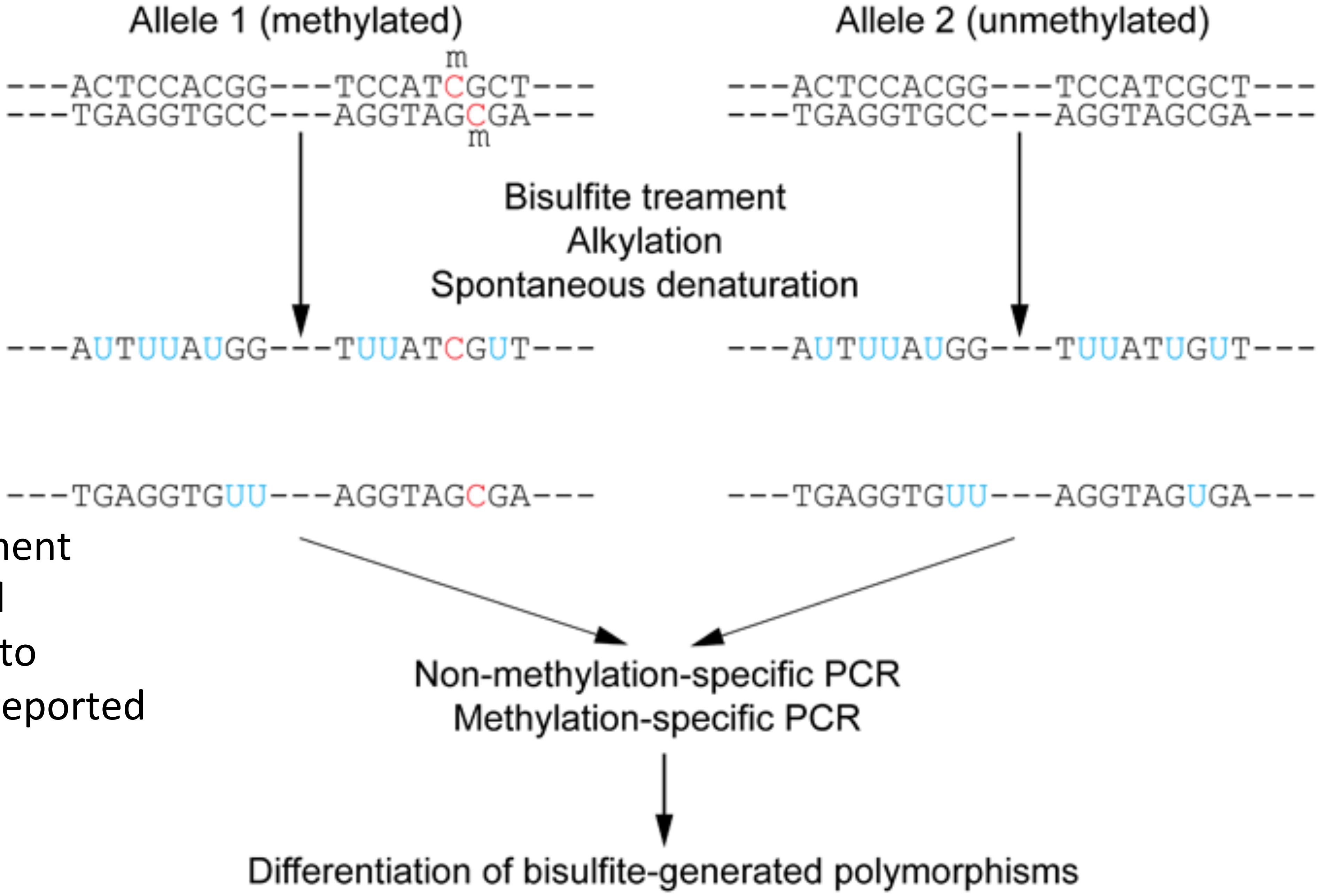
(a) GC content versus O/E. The red vertical and horizontal lines represent the cutoffs used by the Gardiner-Garden and Frommer definition: O/E > 0.6, GC content > 0.5. (b) Histogram of CGI lengths. The vertical line is at the minimum length requirement of Gardiner-Garden and Frommer CGI definition (200 bp).



Wu H et al. Biostat 2010;11:499-514

Biostatistics

Bisulfite sequencing



Allele 1 (methylated)

```
            m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
                      m
```

Allele 2 (unmethylated)

```
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
```

Bisulfite treament
Alkylation
Spontaneous denaturation

```
---AUTUUAUGG---TUUATCGUT---        ---AUTUUAUGG---TUUATUGUT---

---TGAGGTGUU---AGGTAGCGA---        ---TGAGGTGUU---AGGTAGUGA---
```

Non-methylation-specific PCR
Methylation-specific PCR

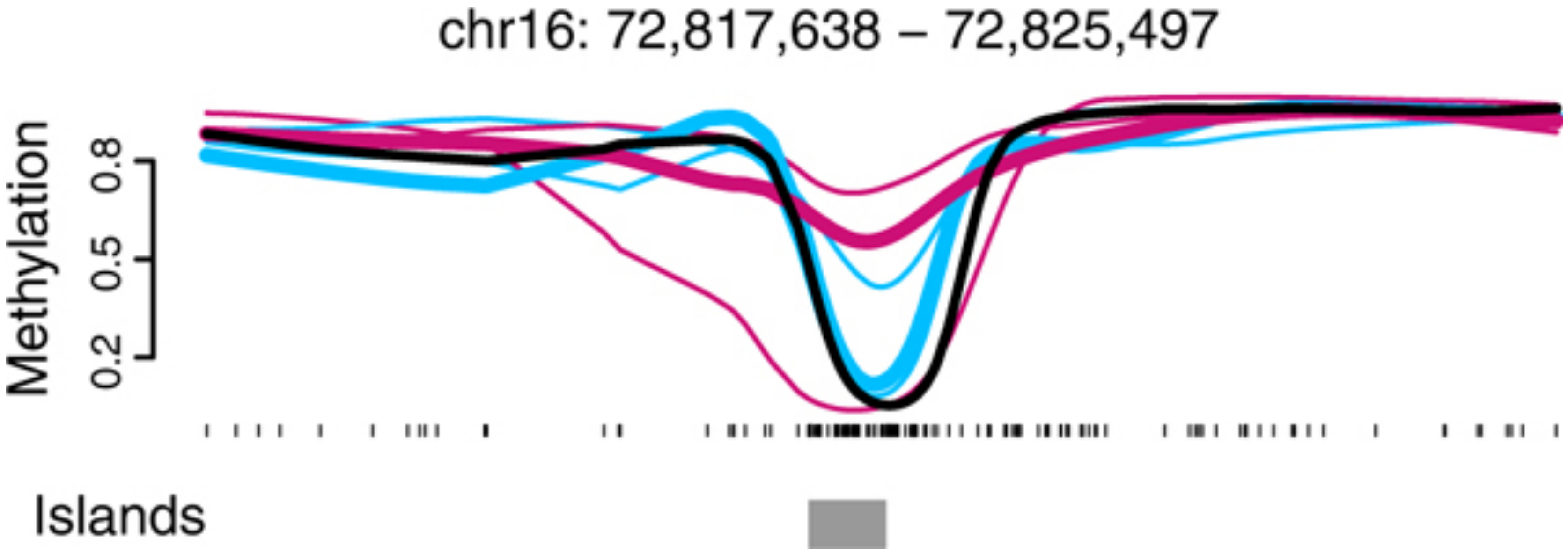
Differentiation of bisulfite-generated polymorphisms

Sodium bisulfite treatment converts unmethylated cytosine C nucleotides to uracils (U), which are reported as thymines (T) by the sequencers and leaves methylated cytosines unmodified.

Whole genome bisulfite sequencing (WGBS)

- WGBS comprehensive but costly
  - Lister et al compared DNA methylation profiles of an embryonic stem cell line and a fibroblast cell line.
  - Both sequenced to about 30x coverage
  - Use computational methods to reduce this to 4x
- Different genomic regions exhibit different levels of DNA methylation variation among individuals (see figure next), but WGBS can be expensive to run on all replicates

Cancer-specific differentially methylated regions: need for biological replicates



- methylation profiles for three normal samples (blue) and matched cancers (red) from the Hansen data [1].
- smoothed methylation profile for an IMR90 cell-line (black) from the Lister data [3].
- If only analyzed normal-cancer pair 3 (thick lines) were to be analyzed, there would appear to be a methylation difference between cancer and normal in this genomic region.
- When all three cancer-normal pairs are considered, however, this region does not appear to be a cancer-specific differentially methylated region.

WGBS Analysis

- Align bisulfite converted reads (lots of work done)
- Identify differentially methylated regions (DMRs) between two or more conditions
- Post-alignment analysis limited:
    - Identify DMRs and group them into regions (ad hoc rules)
    - Do not take biological variability into account
- What we need:
    - Take unbiased and bisulfite-aware read alignment step
    - Compile quality assessment metrics based on stratifying methylation estimates by read position
    - Apply local averaging to improve precision of methylation measurements
    - Detect DMRs accounting for biological variability when replicates are available

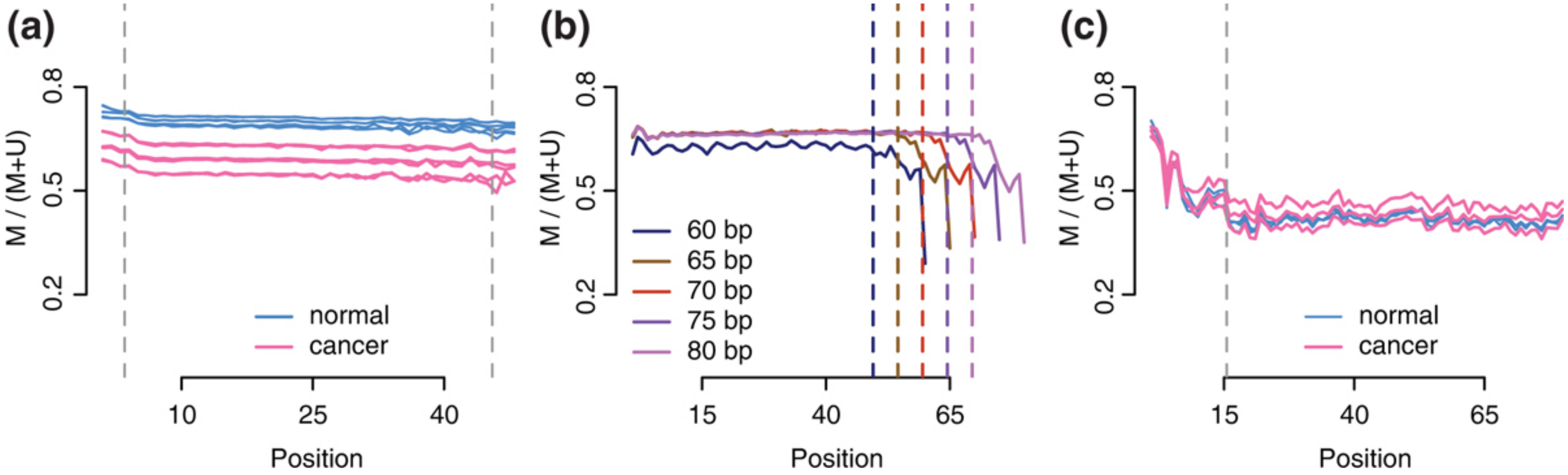
Alignment of WGBS data

- Sodium bisulfite treatment converts unmethylated cytosine C nucleotides to uracils (U), which are reported as thymines (T) by the sequencers and leaves methylated cytosines unmodified
- Align sequencing reads derived from treated DNA to a reference genome and infer methylation status of reference genome
    - C in bisulfite-treated read overlaps a C in the reference, this indicates that reference C is methylated in at least one molecule in the sample
- Reference C's methylation status affects the scores of alignment covering it => bias either toward or against alignments covering methylated cytosines

Alignment Bias

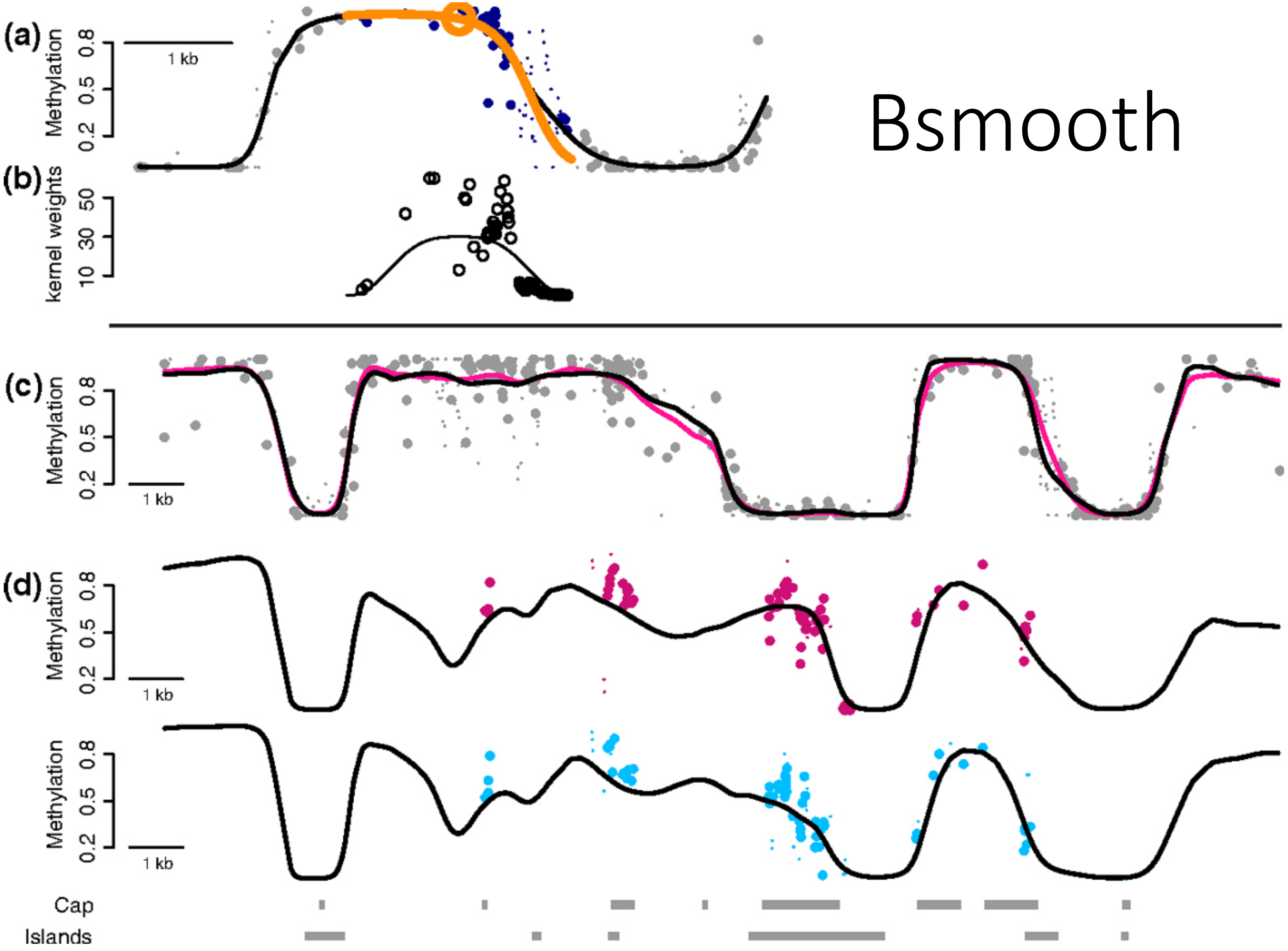
- Avoid bias by removing the penalty associated with aligning a C or T in the read to a C in the reference genome. One such approach is ‘in silico bisulfite conversion’, whereby C nucleotides both in the reads and in the reference genome are converted to T nucleotides prior to alignment
- Or convert only the reference genome in this way but this results in bias against reads overlapping both methylated and unmethylated cytosines

Quality control to offset systematic sequencing and base-calling errors that are common



(a) M-bias plot for the Hansen data, a WGBS experiment on cancer samples. Each sample was sequenced on two flowcells. We show the methylation proportion across each possible read position. This plot shows limited evidence of methylation bias across the read positions. Vertical lines indicate cutoffs used for M-bias filtering. **(b)** M-bias plots for the Lister data, a WGBS experiment in a fibroblast cell line. These data were aligned using iterative trimming and each read length is depicted separately (different colors). The plot shows methylation bias toward the end of reads for all read lengths. **(c)** M-bias plot for the Hansen-capture data, a capture bisulfite sequencing experiment on cancer samples. The plot shows methylation bias at the start of the reads.

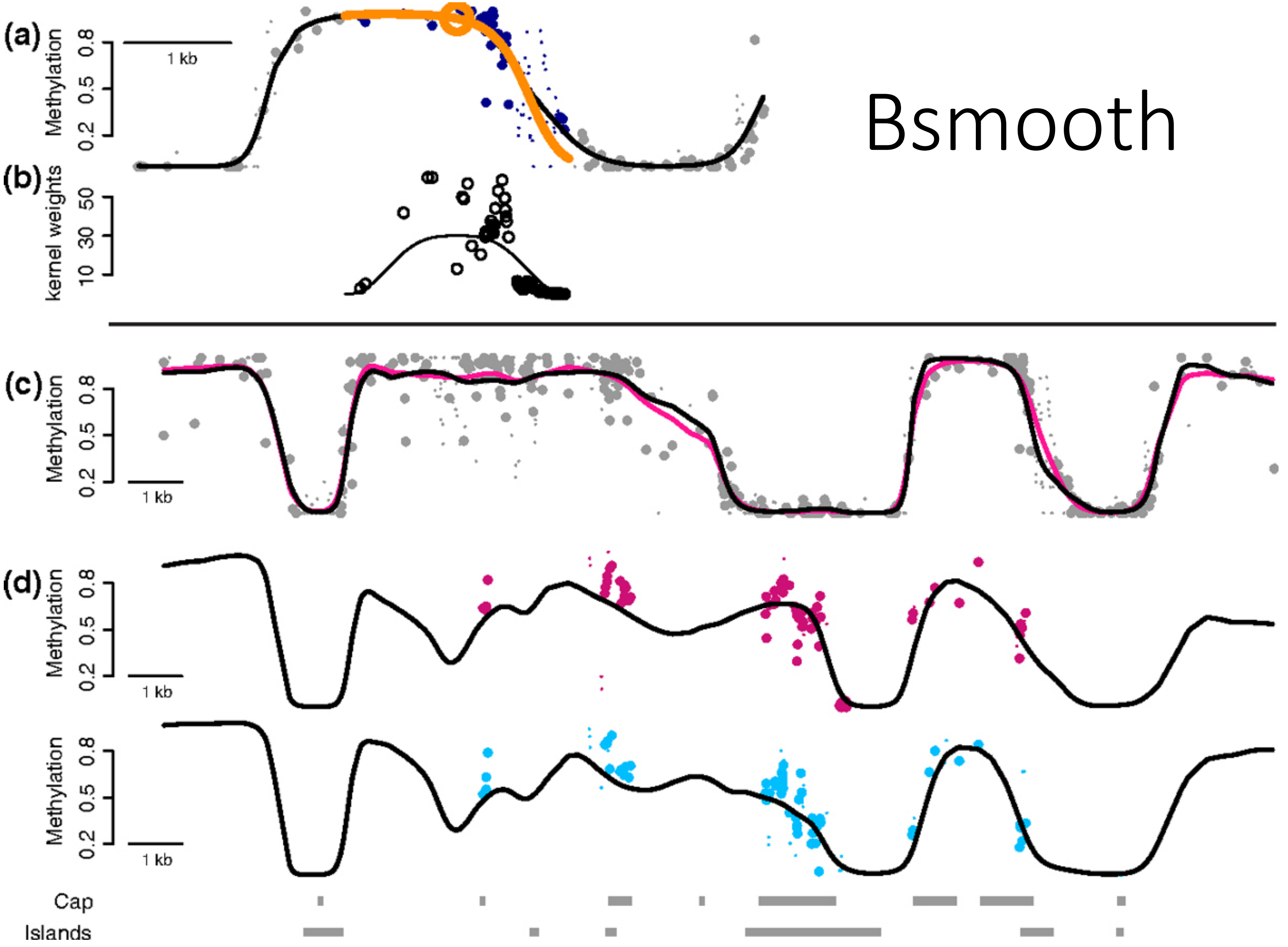
a. Points represent single-CpG methylation estimates plotted against their genomic location. Large points are based on greater than 20× coverage. The orange circle denotes the location for which we are estimating the methylation profile. The blue points are those receiving positive weight in the local likelihood estimation. The orange line is obtained from the fitted parabola. The black line is the methylation profile resulting from repeating the procedure for each location.



Bsmooth

Hansen *et al. Genome Biology* 2012 **13**:R83 doi:10.1186/gb-2012-13-10-r83

(b) The curve represents the kernel used in the weighted regression and the points are the actual weights, which are also influenced by coverage. **(c)** Points are as in (a) for the 25× coverage Lister data. The pink line is obtained by applying BSmooth to a the full data. The black line is the estimate from BSmooth based on a 5× subset of the Lister data.



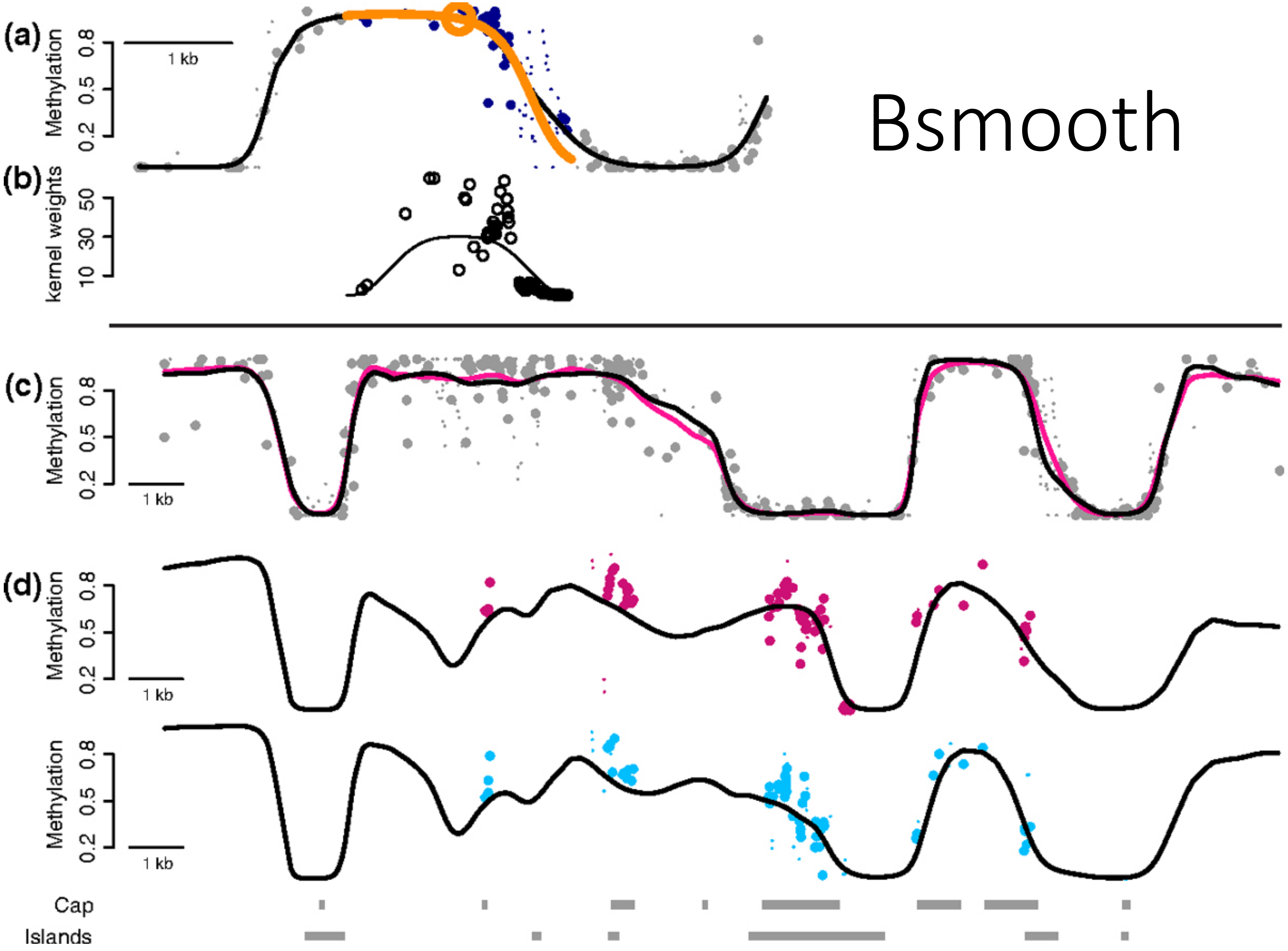
Bsmooth

Hansen *et al. Genome Biology* 2012 **13**:R83 doi:10.1186/gb-2012-13-10-r83

Bsmooth

(d) The points are as in (a) but for the Hansen-capture data with average 35× coverage, and average across three replicates. The black line is the BSmooth estimate obtained from the 4× Hansen data, averaged across three replicates.

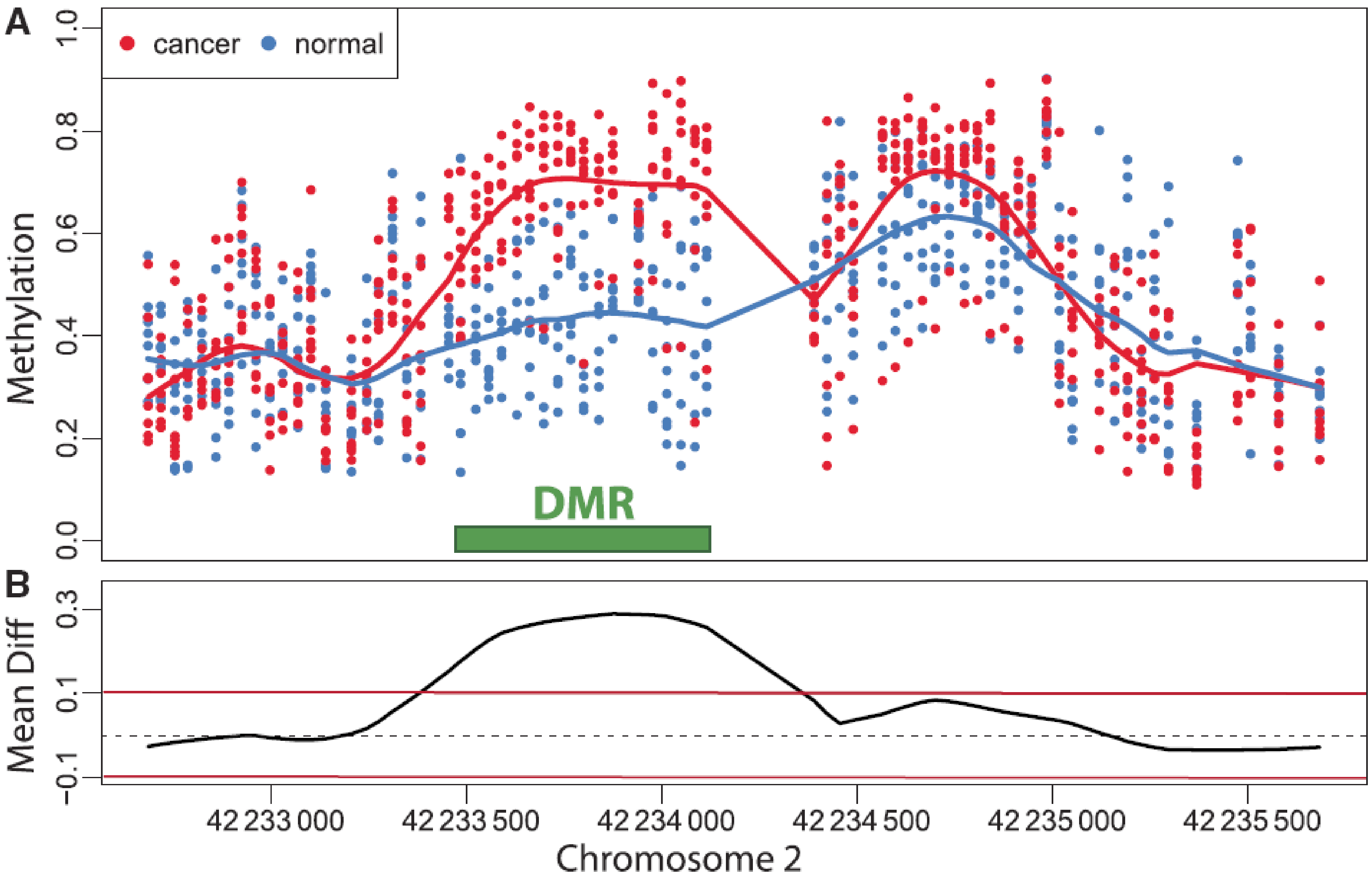
Hansen *et al. Genome Biology* 2012 **13**:R83 doi:10.1186/gb-2012-13-10-r83



Hansen *et al. Genome Biology* 2012 **13**:R83 doi:10.1186/gb-2012-13-10-r83

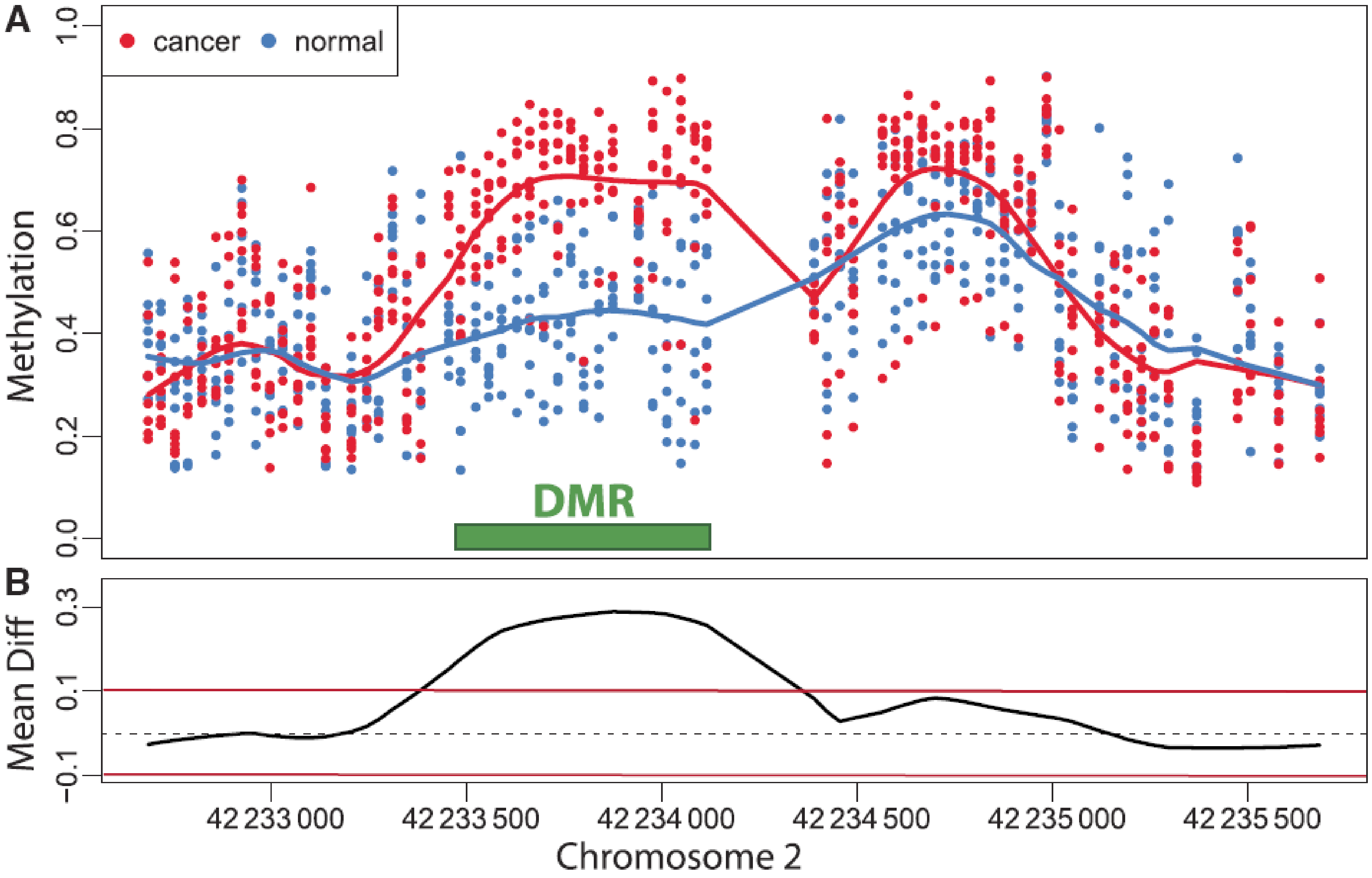
(A) The points show methylation measurements from the colon cancer dataset plotted against genomic location from illustrative region on chromosome 2. Eight normal and eight cancer samples are shown in this plot and represented by eight blue points and eight red points at each genomic location for which measurements were available. The curves represent the smooth estimate of the population-level methylation profiles for cancer (red) and normal (blue) samples. The green bar represents a region known to be a cancer DMR.20

Example of a differentially methylation region (DMR)



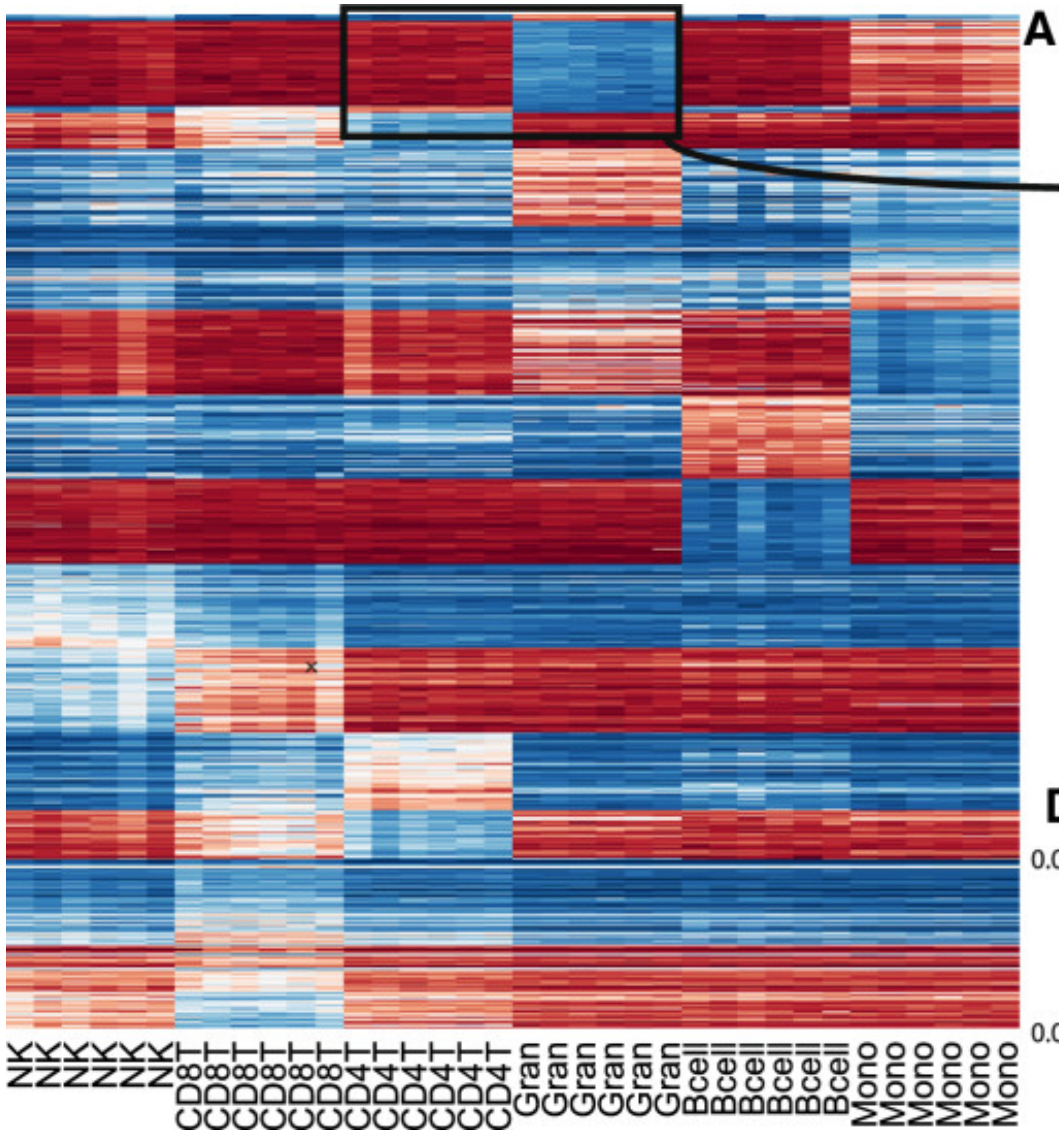
(B)The black curve is an estimate of the population-level difference between normal and cancer. We expect the curve to vary due to measurement error and biological variation but to rarely exceed a certain threshold, for example those represented by the red horizontal lines. Candidate DMRs are defined as the regions for which this black curve is outside these boundaries.
Note that the DMR manifests as a bump in the black curve

Example of a differentially methylation region (DMR)



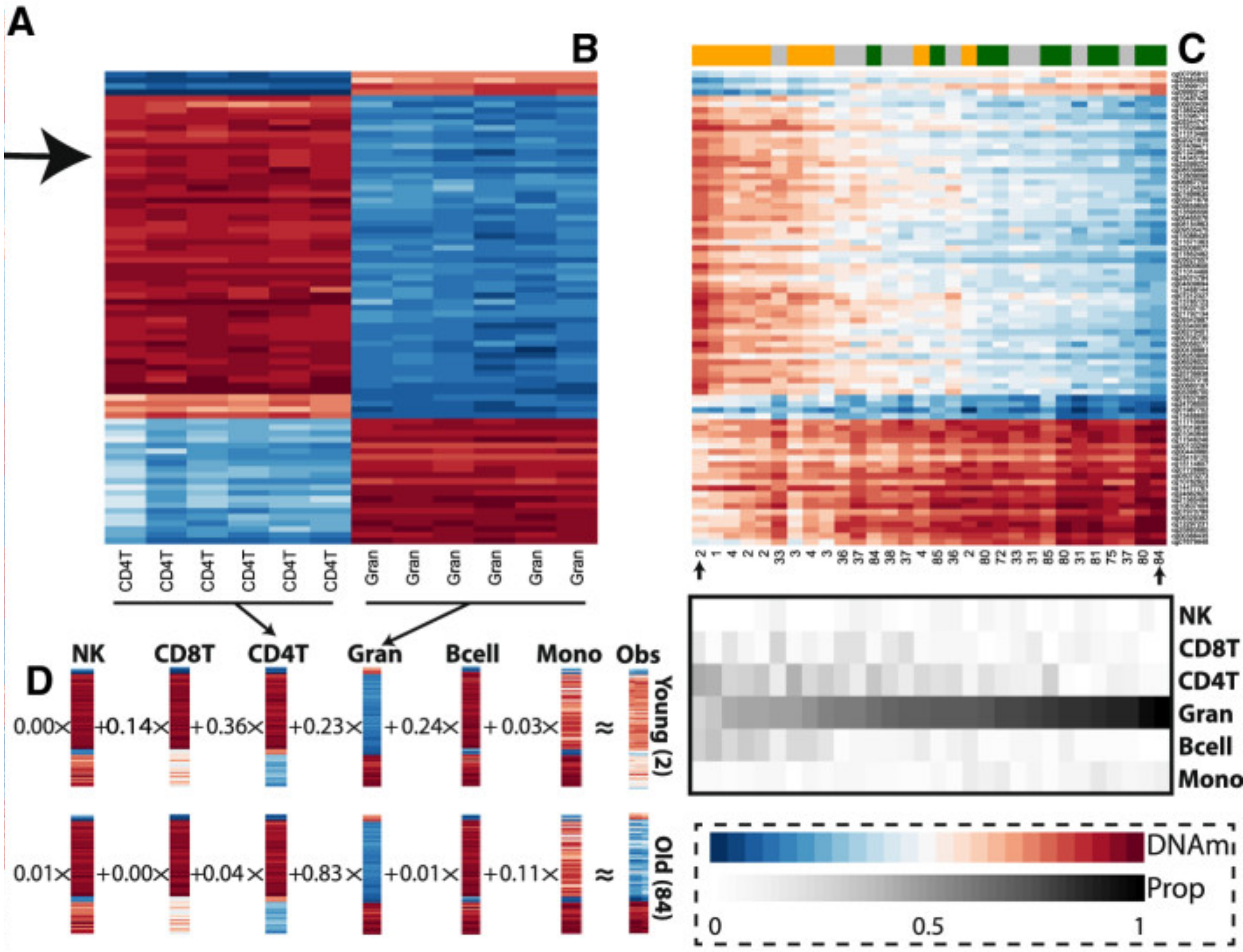
How blood composition drives observed age differences

(A) Heatmap of the cell sorted data shows very clear and consistent DNAm profiles for each cell type. We show 600 probes selected for estimating composition proportions used to demonstrate differences here.



How blood composition drives observed age differences

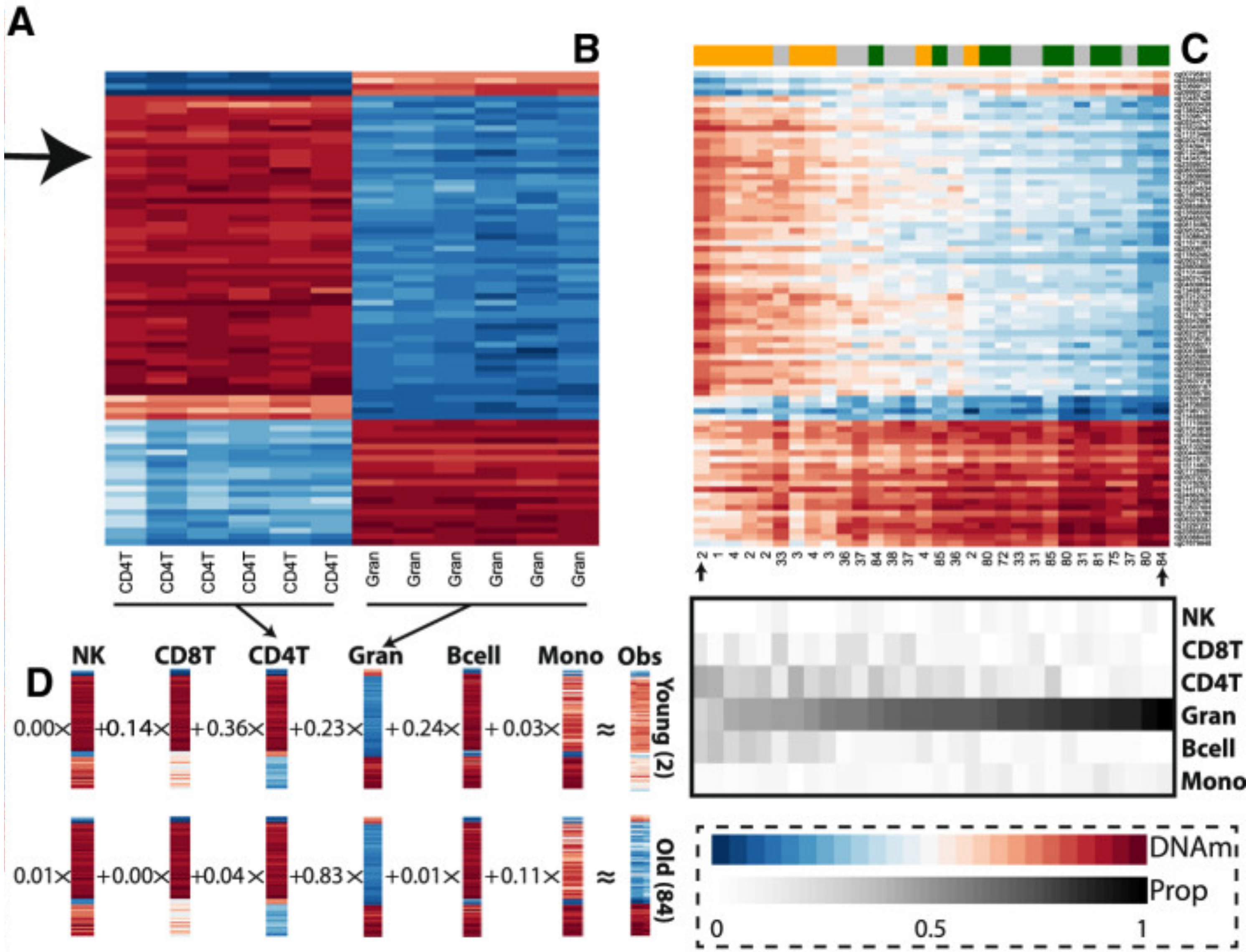
(B) To simplify the illustration we selected a section of **(A)** displaying only the two most abundant cell types: CD4+ T cells and granulocytes. **(C)** Heatmap of a randomly selected sample of 30 whole blood samples (from the data in Additional file 1) across three age groups (10 per group): between 1 and 5 years of age, between 30 and 40, greater than 60 years. The same probes as in **(B)** are used. When the samples are ordered by their estimated granulocyte proportion, the samples roughly cluster by age and a similar pattern to **(B)** is observed. The estimated cell count proportions for each of the samples are shown below. Note the strong confounding between age and cell composition.



How blood composition drives observed age differences

(D) For the two samples highlighted with an arrow in **(C)**, we show how a weighted average of the cell type profiles can reconstruct the observed DNAm profiles. The numbers shown are the estimated proportions. Note how different weights (cell counts) for old and young result in very different observed DNAm patterns. Note that the differences in CD4+ T cells and granulocytes drive much of the differences in DNAm. NK, CD56+ natural killer cells; CD8T, CD8+ T cells; CD4T, CD4+ T cells, Gran, granulocytes; Bcell, CD19+ B cells; Mono, CD14+ monocytes; DNAm, proportion of DNA methylation at individual CpGs (Illumina 'beta' values, bound between 0 and 1); Prop, cell count proportion, between 0 and 1 for each component, such that they sum to 1.

Jaffe and Irizarry *Genome Biology* 2014 **15**:R31 doi:10.1186/gb-2014-15-2-r31


A
B
C
CD4T
CD4T
CD4T
CD4T
CD4T
CD4T
Gran
Gran
Gran
Gran
Gran
Gran
D
NK
CD8T
CD4T
Gran
Bcell
Mono
Obs
0.00×
+0.14×
+0.36×
+0.23×
+0.24×
+0.03×
≈
Young (2)
0.01×
+0.00×
+0.04×
+0.83×
+0.01×
+0.11×
≈
Old (84)
NK
CD8T
CD4T
Gran
Bcell
Mono
DNAm
Prop