DNA methylation data analysis

Yujia Qin

Bioinformatics Core, JABSOM

Department of Quantitative Health Science

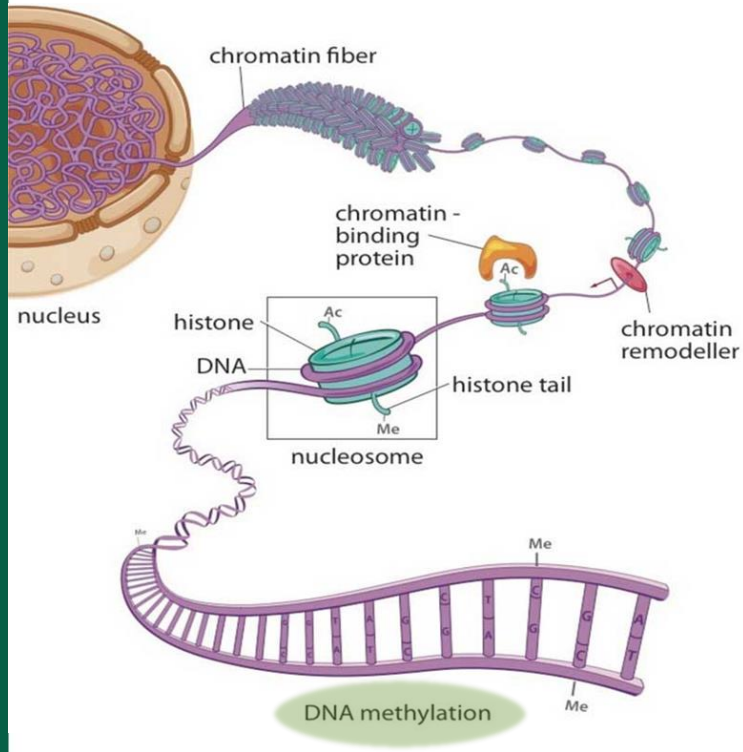
University of Hawaii

2022

1.
DNA methylation

Introduction

What is **DNA methylation**?

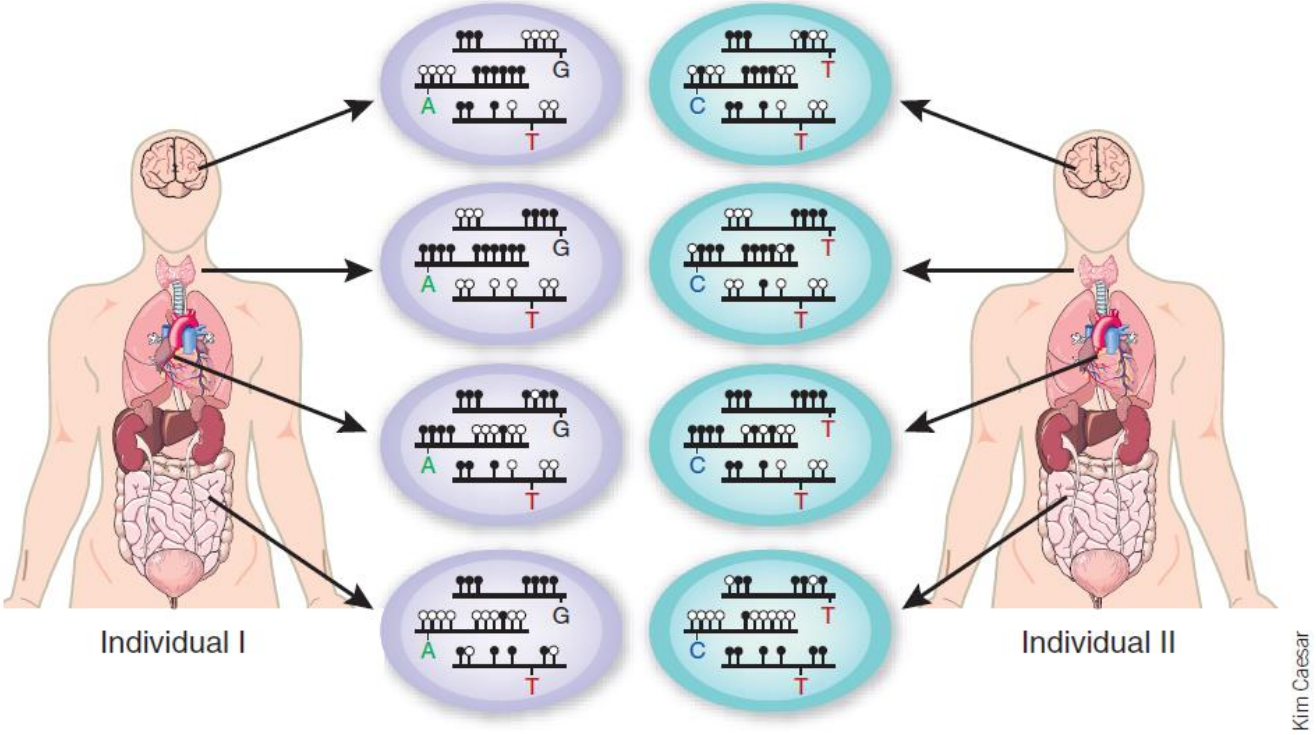


- **DNA methylation** is a major <u>epigenetic</u> mechanism involving direct chemical modification to the DNA

- **Epigenetics** is the study of the mechanisms that cause changes in gene expression but are not due to the change in the DNA sequences, including DNA methylation, histone modification, non-coding RNA expressions

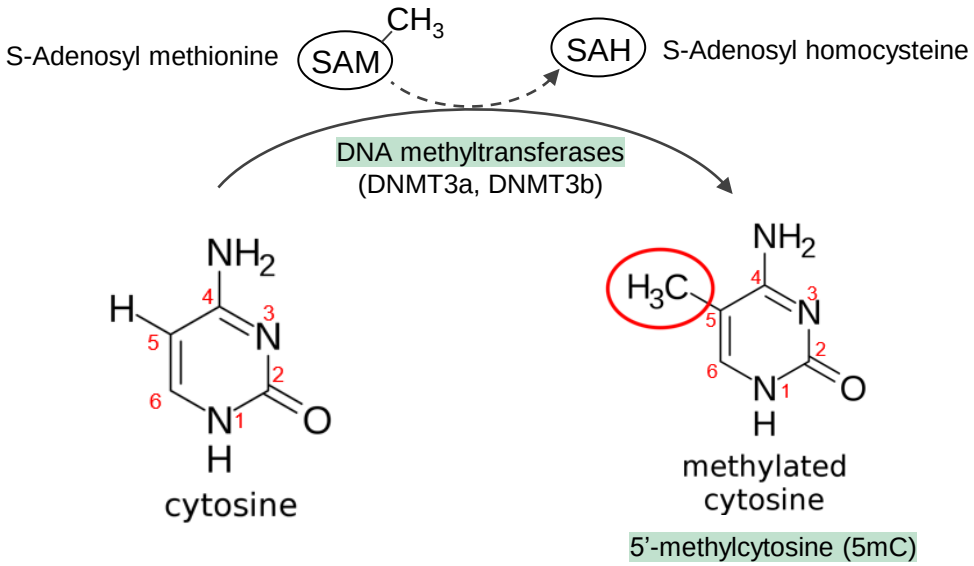
From The Cell Biology of Stem Cells (2010)

Tissue-specific **DNA methylation** and epigenetic heterogeneity among individuals



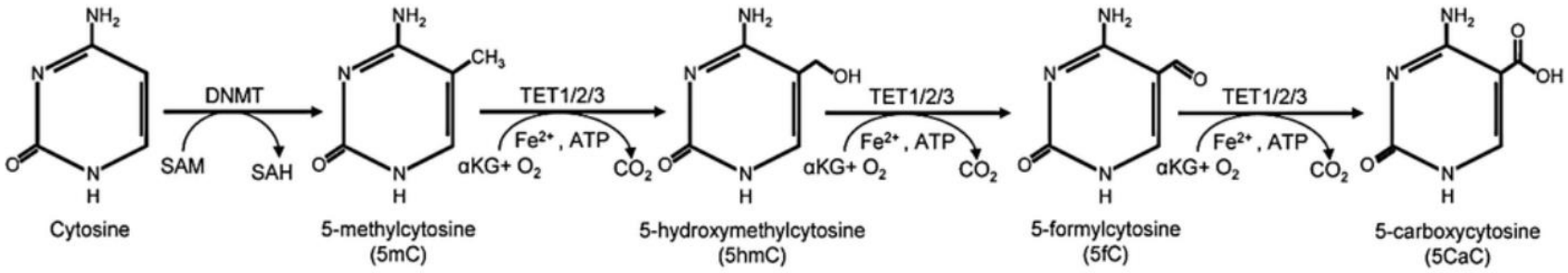
Brena RM, Huang TH, Plass C. Toward a human epigenome. Nature genetics. 2006 Dec;38(12):1359-60.

What is **DNA methylation**?

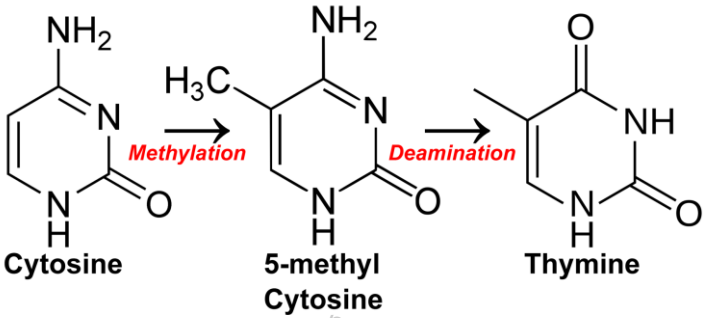


- Most common form of DNA methylation is 5-methylcytosine (5mC)
- Three types of DNA methyltransferases (DNMT's): DNMT1, DNMT3a, DNMT3b
- 5mC is inherently mutagenic: spontaneously undergo deamination, leading to C->T transitions.

Oxidized Derivatives of 5-MethylCytosine

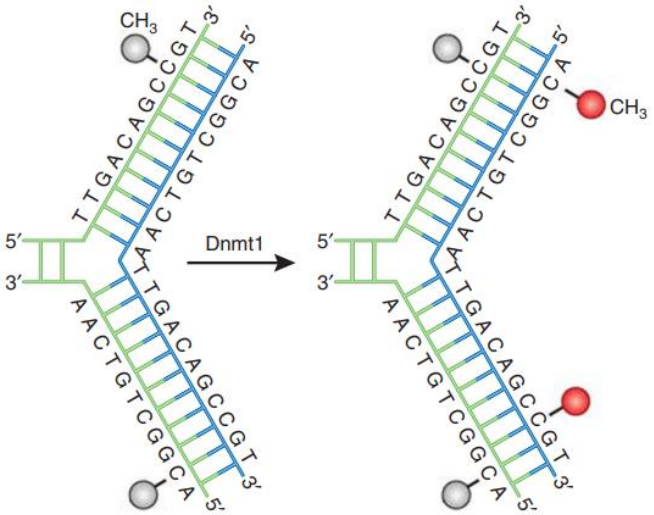


Cytosine Deamination

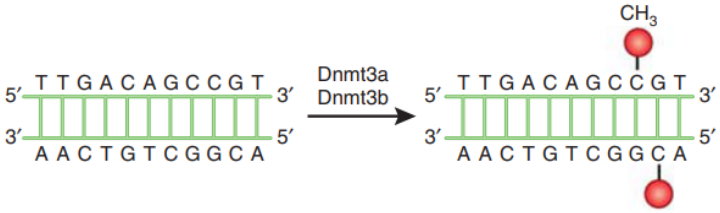


DNA methylation **enzymes**

DNMT1: **maintenance** DNA methyltransferase:



DNMT3a, DNMT3b: **de novo** DNA methyltransferase:

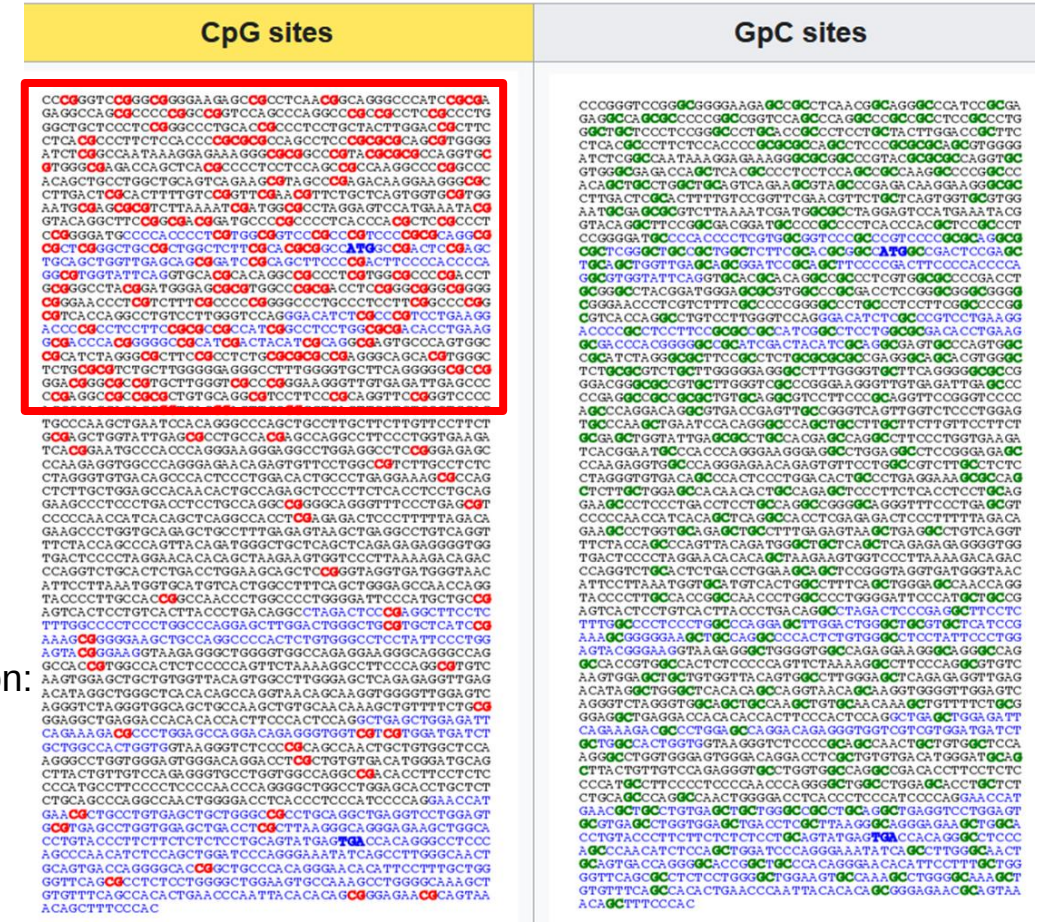


Moore, Lisa D., Thuc Le, and Guoping Fan. Neuropsychopharmacology 38.1 (2013): 23-38.

DNA methylation **locations**

- DNA methylation primarily happens at CpG sites.
- **CpG** dinucleotides/**sites**
  - DNA regions where a cytosine is followed by a guanine in the 5'-3' direction (5'-Cytosine-phosphate-Guanine-3')
  - CG dinucleotides present on single strand of DNA
  - <u>Under-representation</u>
- **CpG island** are regions with a high frequency of CpG sites
  - Length > 200bp
  - GC percentage > 50%
  - Observed-to-expected CpG ratio> 60%
- CpG island '**shores**' are regions of comparatively low CpG density, located approximately 2 kb from CpG islands

CpG island



In mammalian genome, 70-80% CpGs are **methylated**.

60% of mammalian **promoters** have CpG islands, but they are mostly **unmethylated**

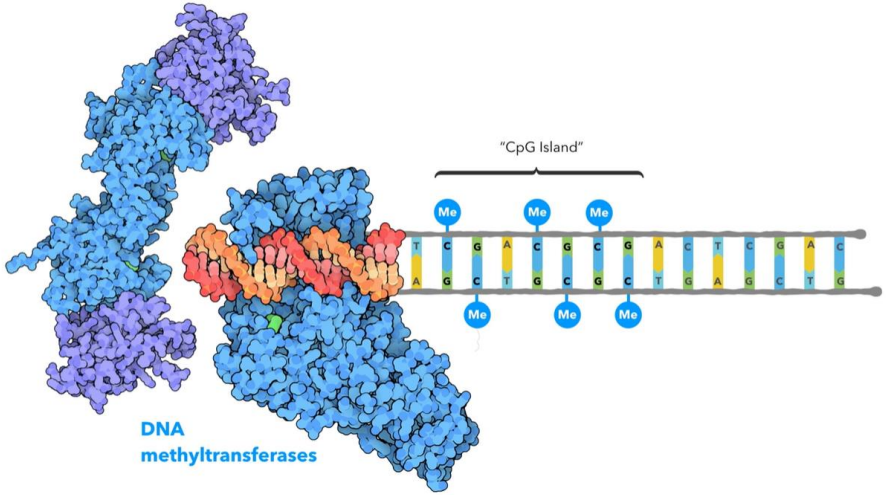
Under-representation: CG suppression

DNA methylation **functions** vs. **locations**

- DNA methylation effects on transcriptional regulation differ depending on the location of the CpG site (intragenic vs promoter region vs enhancer).
  - **Methylated** CpG island **promoters** are associated with **gene repression**.
  - CpG islands, occur and commonly span **promoters** of house-keeping genes. These promoter CpG islands typically remain unmethylated, resulting in active gene expression
  - **Gene bodies** tend to have intermediate CpG densities. Unlike CpG island promoters, extensive exonic or genic methylation is typically associated with **active gene expression**.
- Beyond these regions, the genome has a lower-than-expected frequency of CpG sites which are typically methylated

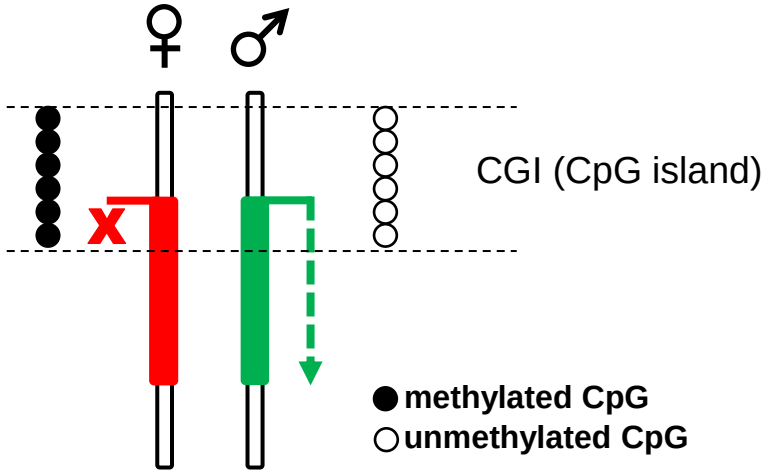
DNA methylation **functions**

- DNA methylation has been shown to be important for:
  - Genomic imprinting
  - Transposable element silencing
  - Stem cell differentiation
  - Embryonic development
  - Inflammation
  - Cancer



Imprinted Genes: mono-allelic expression

Differential allelic DNA methylation

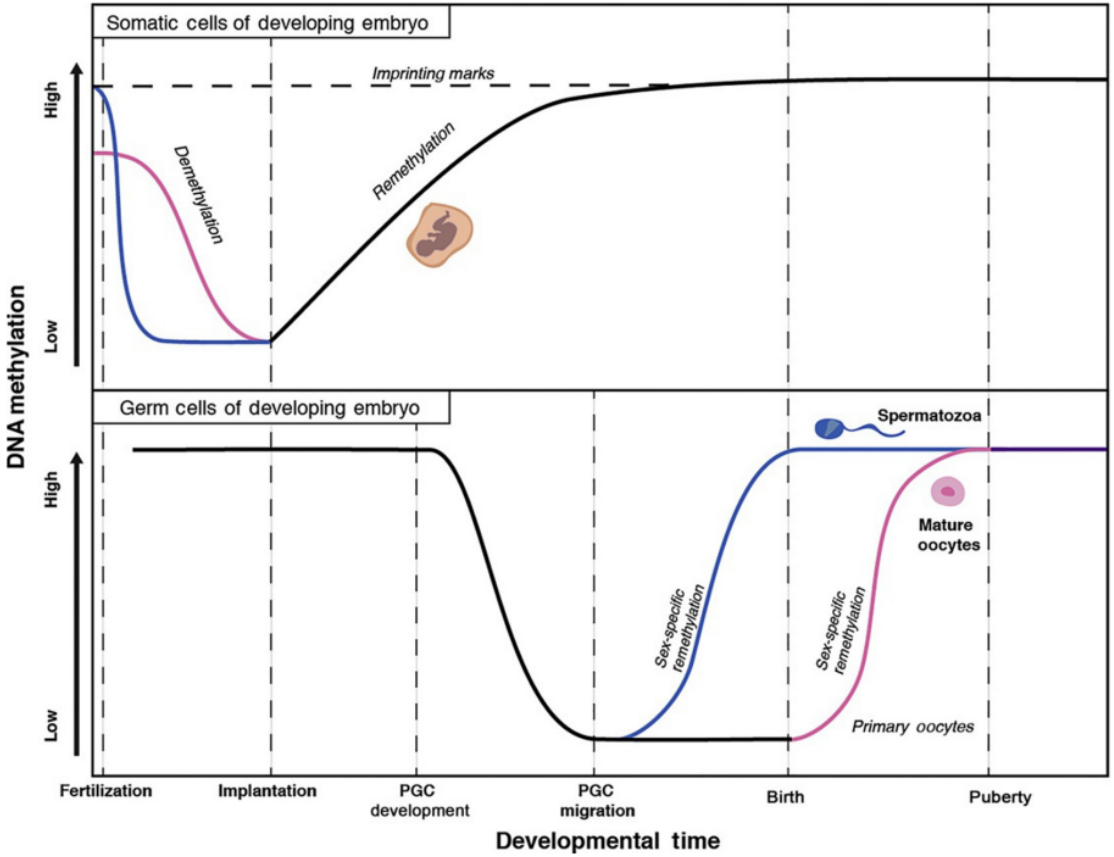


Imprinted Genes: Mono-allelic expression with parent-of-origin specificity. Have key roles in energy metabolism, placenta functions.

DNA methylation is reset during reprogramming

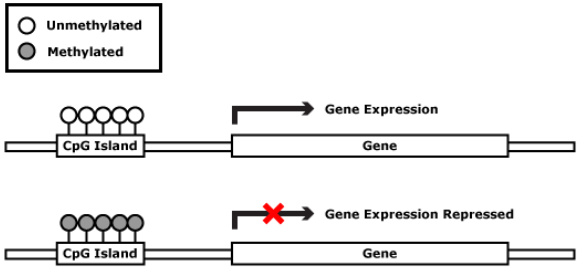
The genome undergoes two waves of global demethylation and re-methylation for the purpose of producing the next generation:

1. After fertilization
2. Germ cell development



Zeng et al,. Genes (Basel). 2019 Apr; 10(4): 257.

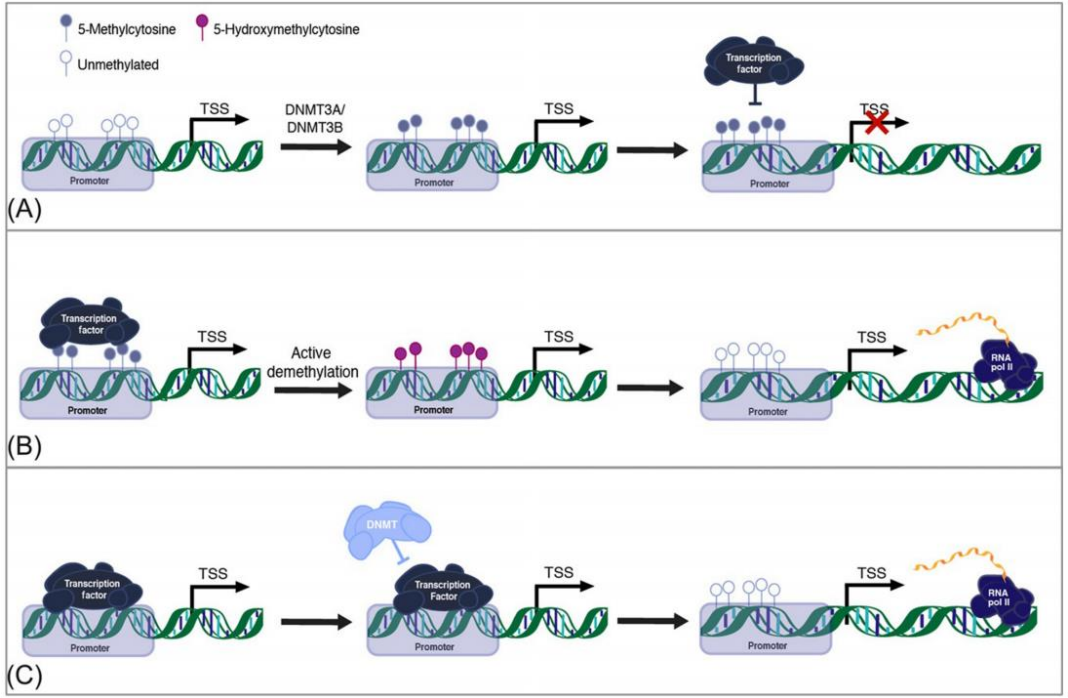
DNA methylation in gene **promoter**



Silencing of gene expression

Two possible ways:

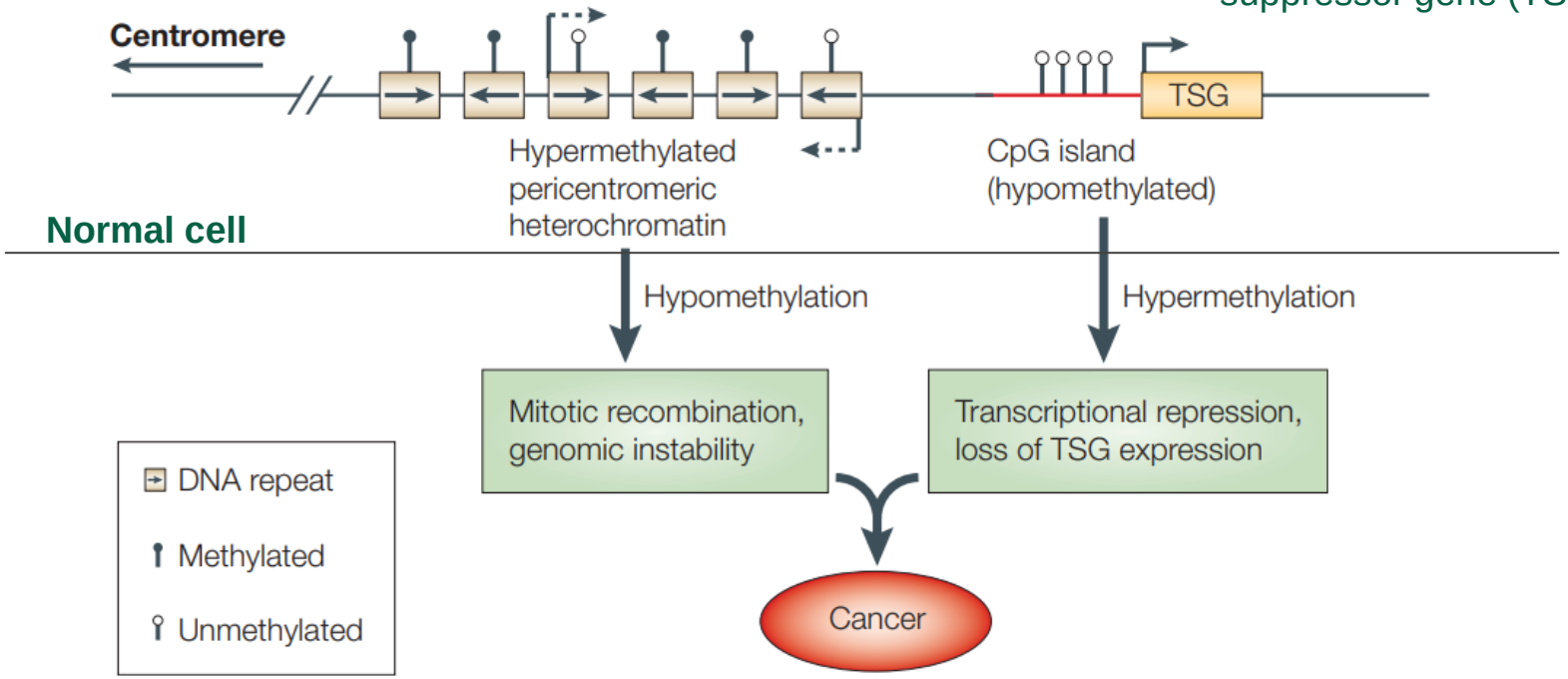
- Prevent TF from binding
- Wrap DNA up make it inaccessible.



Bommarito et al,. Toxicoepigenetics. Academic Press, 2019. 127-151.

DNA methylation and cancer

Transcribed tumor suppressor gene (TSG)

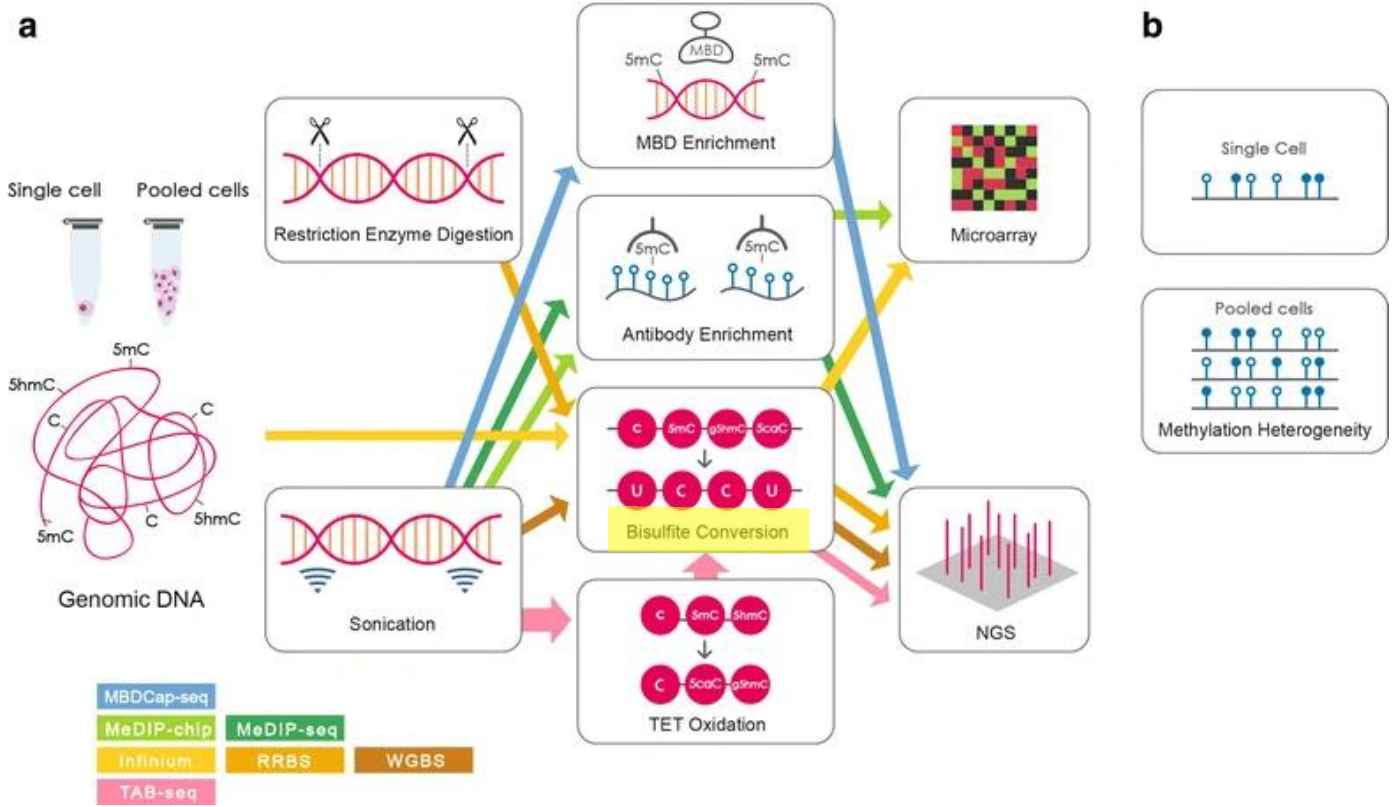


Robertson, Keith D. "DNA methylation and human disease." Nature Reviews Genetics 6.8 (2005): 597-610.

2.
How to measure DNA methylation?

Bisulfite coversion

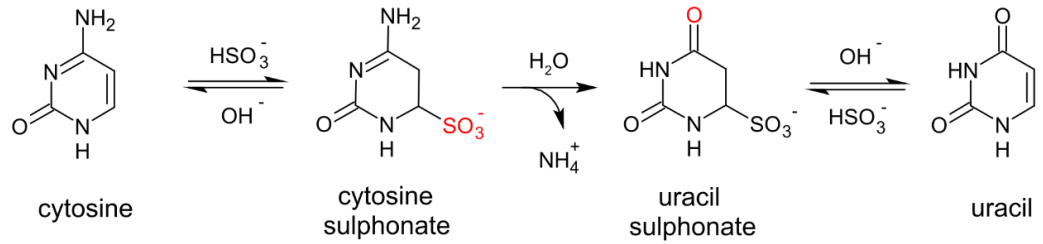
Assays for measuring DNA methylation

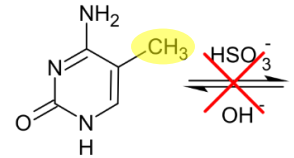


Yong, et al,. "Profiling genome-wide DNA methylation." Epigenetics & chromatin 9.1 (2016): 1-16.

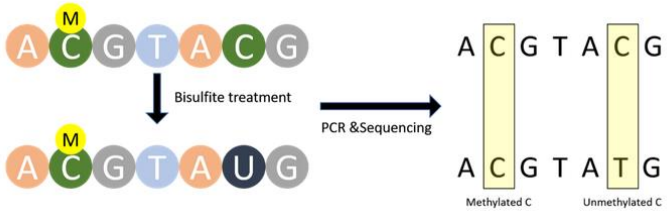
Bisulfite treatment

- Treatment of DNA with **bisulfite** converts cytosine residues to uracil but leaves <u>5-methylcytosine</u> residues <u>unaffected</u>.

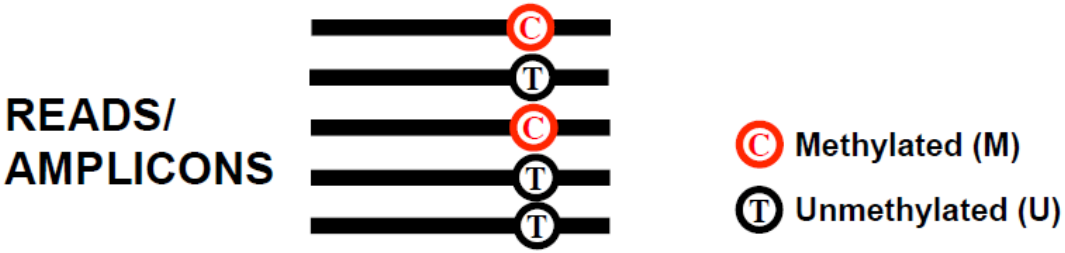




5-methylcytosine



Measuring DNA methylation



$$\%\text{Methylation} = \frac{M}{M + U} \times 100$$

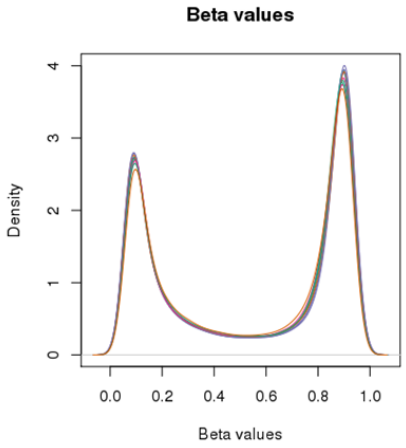
Measuring DNA methylation

- For **microarrays**, there are other measurements:

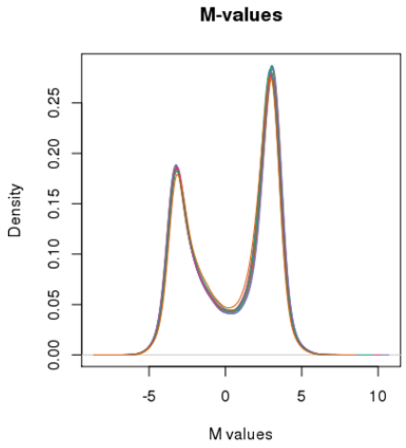
$$\beta = \frac{Meth}{Meth + Unmeth}$$

$$M = \log_2 \frac{Meth}{Unmeth}$$

$$M = \log_2 \frac{\beta}{1-\beta}$$

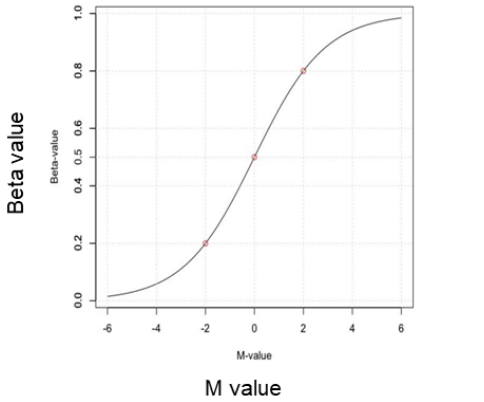


Intuitive, easy to interpret, great for visualisation



Better statistical properties, recommended for statistical testing

Du et al. 2011, BMC Bioinformatics



Can convert between them via a logit transformation

Illumina Infinium **Human** Methylation BeadChips

27k array (2009)

1 chip = 12 samples

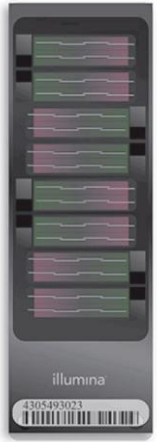
>27,000 unique **CpG sites** measured in each sample

450k array (2011)

1 chip = 12 samples

>450,000 unique **CpG sites** measured in each sample

EPIC array (Today)

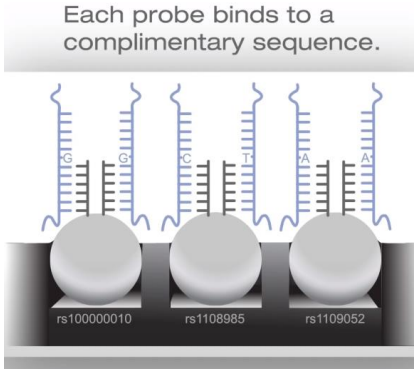


1 chip = 8 samples

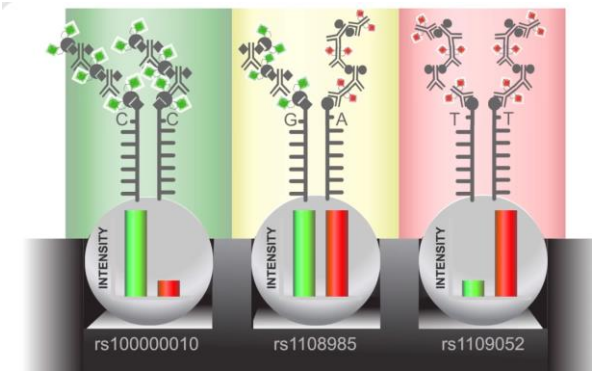
>850,000 unique **CpG sites** measured in each sample

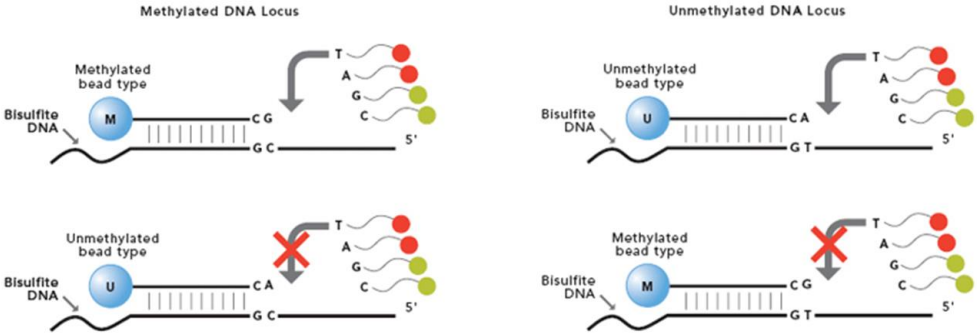
Modified slide from Belinda Phipson

Assays for measuring DNA methylation – **bisulfite microarrays**



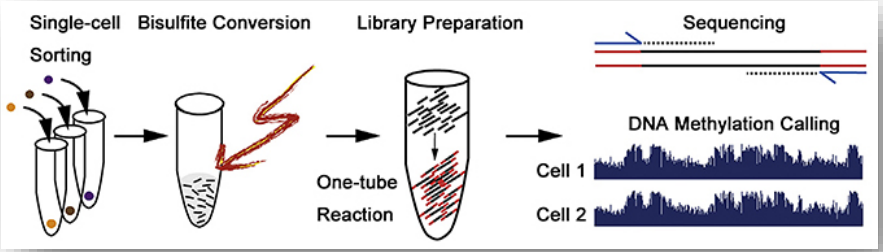
Illumina **Infinium** Methylation Assay





Whole-genome bisulfite sequencing (WGBS)

- WGBS experiment steps:
  - Sonication
  - End-repair
  - A-tailing
  - Adapter ligation
  - Bisulfite conversion
  - Amplification
  - Sequencing



- Advantages:
  - measure single-cytosine methylation levels genome-wide
  - directly estimate the ratio of molecules methylated rather than enrichment levels

3.
WGBS data analysis workflow

The real deal

WGBS data analysis steps

1. Quality control → 2. Bisulfite alignment → 3. Methylation Quantification → 4. Visualization → 5. DMRs detection

1. Quality control

   Quality trimming, adapter trimming

2. Bisulfite sequence alignment

   Two strategies: wild-card alignment, three letter alignment

3. Quantification of DNA methylation (methylation calling)

   Sequence deduplication, .bed file format

4. Visualization

5. Differentially methylated regions (DMRs) detection

Why do we need quality control?

NGS generates highly accurate data, but it can have certain types of errors:

- contamination with adapters
- technical duplication in the library
- failure at specific parts of the flowcell
- PCR duplicates

Reads without proper trimming reads may result in

- low mapping efficiency
- mis-alignments
- errors in methylation calls since adapters are methylated
- base-call errors tend toward 50%

Tools to use: FastQC, Trim Galore!, Trimmomatics ...

Sequencing data files - FASTQ

Pair-end reads

sample_1_**R1**_001.fastq

sample_1_**R2**_001.fastq

sample_2_**R1**_001.fastq

sample_2_**R2**_001.fastq

FASTQ Format

Identifier — `@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`

Sequence — `TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA`

+ sign & identifier — `+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`

Quality scores — `efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]]`[]dddd`

Base T

phred Quality] = 29

Quality control – **FASTQC**

Q-score: "Phred quality scores"

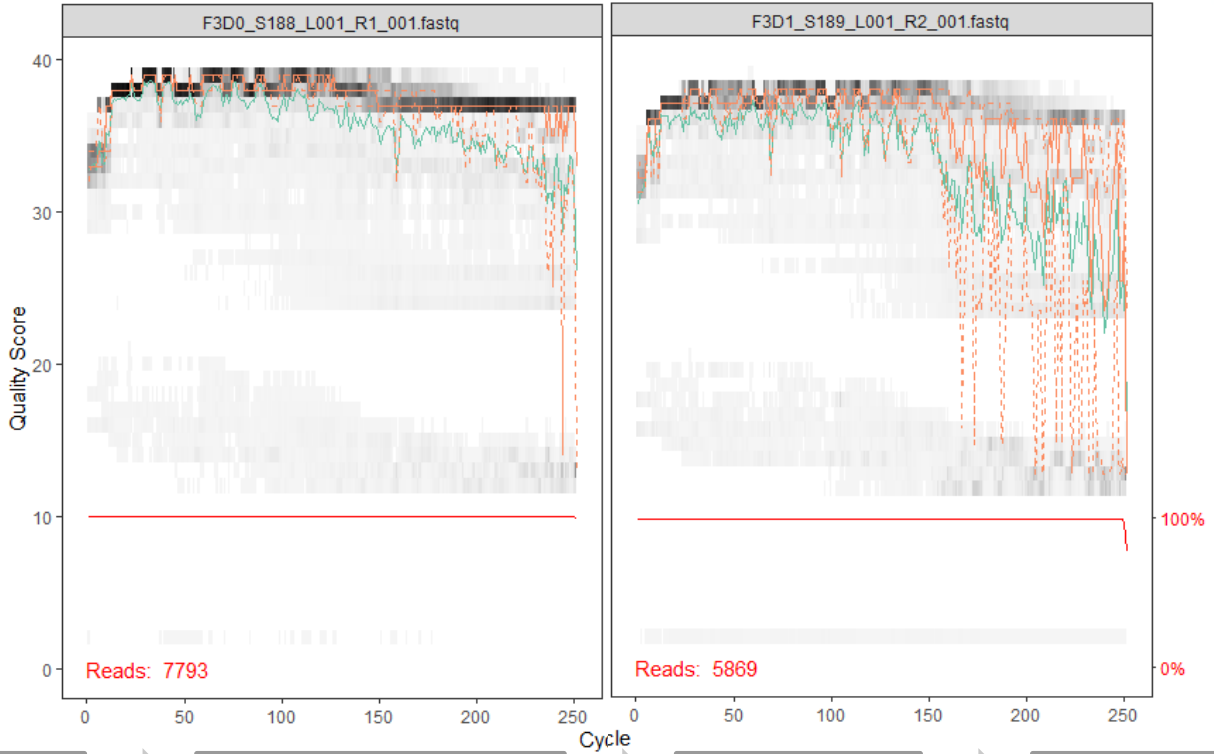
These scores represent the likelihood of the base being called wrong.

$$Q_{\text{phred}} = -10 \log_{10} e,$$

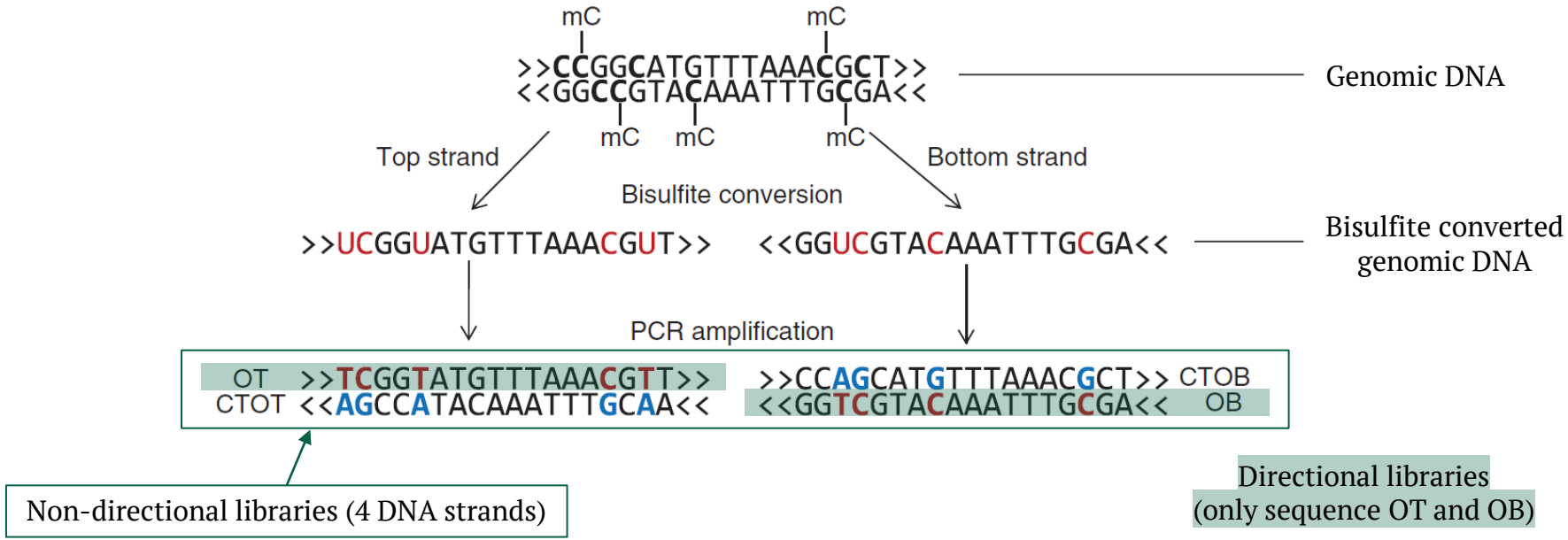
e is the probability that the base is called wrong.

When Q= 30, *e* = 0.1%

For Illumina, Q>30, the base call quality value is good enough



Effect of bisulfite treatment of DNA



Krueger et al,. *Nat Methods 9, 145-51* (2012)

Bisulfite sequence alignment – two strategies

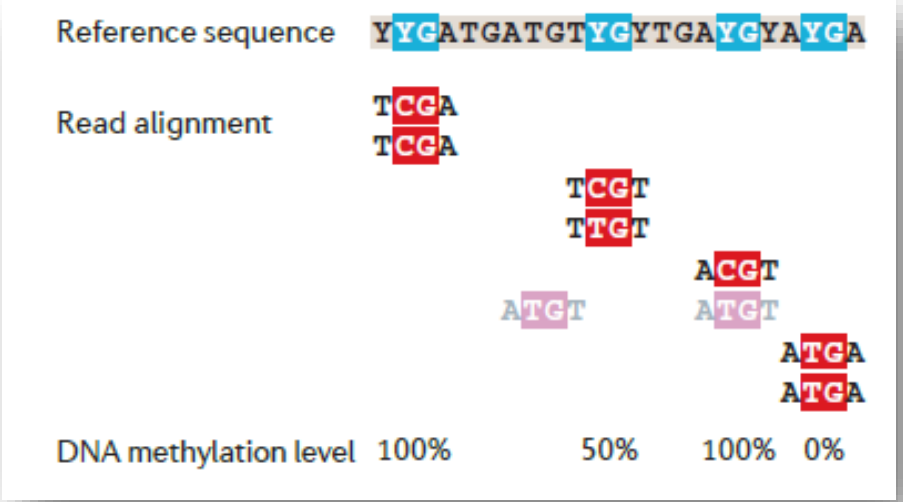
Genomic DNA sequence CCGATGATGTCGCTGACGCACGA
DNA methylation level 100% 50% 50% 0%

Bisulphite-sequencing reads ACGT,ATGA,ATGA,ATGT,
TCGA,TCGA,TCGT,TTGT

1.Wild-card aligner

Replace Cs in the genomic DNA sequence (reference) by the wild-card letter Y, which matches both Cs and Ts in the read sequence

The **DNA methylation level**: the percentage of aligning Cs among all uniquely mapped reads



Bock, *Nat Rev genet*, 2012

Bisulfite sequence alignment – two strategies

Genomic DNA sequence	CCGATGATGTCGCTGACGCACGA			
DNA methylation level	100%	50%	50%	0%

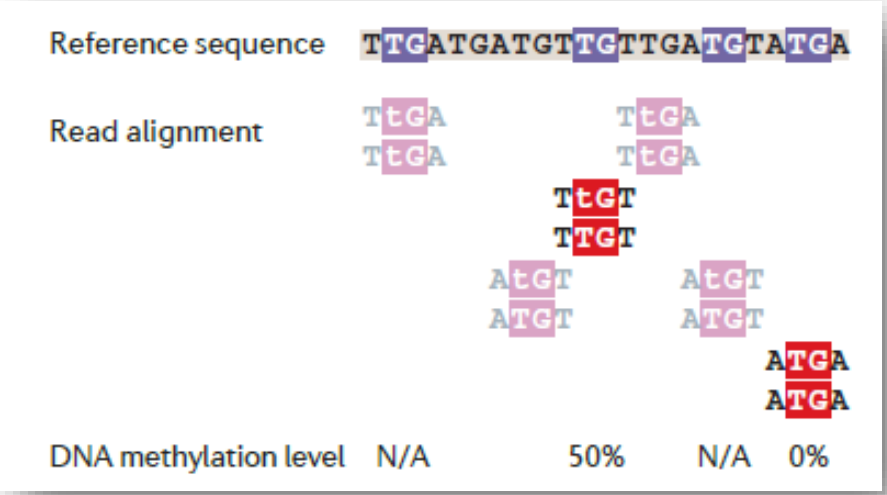
Bisulphite-sequencing reads	ACGT, ATGA, ATGA, ATGT, TCGA, TCGA, TCGT, TTGT

2. Three-letter aligner

Converting all Cs into Ts in the reads (in lower case t) and genomic DNA sequences

Carry out the alignment on a three-letter alphabet (A, G and T) using standard aligner, such as Bowtie.

The **DNA methylation level**: the percentage of aligning Cs among all uniquely mapped reads



Bock, *Nat Rev genet*, 2012

Bismark primary alignment output (BAM file)

chromosome position

Read 1

```
HISEQ2000-06:366:C3G4NACXX:3:1101:1316:2067_1:N:0:      99      16      71322125        255     100M    =
71322232        207
NTTATTTAGTTTTTTAGGGTTTGTGTGTAGGAGTGTGGGAATTATGTTTTTTATGGTTGATATTTATTTAAAAGTGAGTATAAATTATATATATTTTTTT
#1=DDDDDAAFFHIIIA:<FGHCCEFGHD?CFFBBBGEHHGHIII<FEHIIIII==DE??EHHFHEEEEEEEC>;>66;@CDEEEDCEEEEEEEDDDCBB
NM:i:14 XX:Z:G8C2C7C21C13C6CC1C17CC3C4CC4
XM:Z:.........h..h.......x.....................h.............x......hh.h.................hh...h....hh....
XR:Z:CT XG:Z:CT XA:Z:1
```

sequence

quality

```
HISEQ2000-06:366:C3G4NACXX:3:1101:1316:2067_1:N:0:      147     16      71322232        255     100M    =
71322125        -207
GGTTATTTTATTTAGGGTTATTGTTTTAGAGTTTTATTGTTGTGAACAGATATATGATTAAGGTAATTTTTATAAGGATAATATTTAATTGGAGTTGGTT
CCCEEECADCFFFFHHGHGHIIGIHFIJJIJIHFGHGGGEHIJIIJGIGFJJJJJJJJJJIGJJJJGJJJJIIIJJIJIJJJJJJIJHHHHHFFFFFCCC
NM:i:21 XX:Z:2G2CC1C1C1C11C11C2C10C1C4CC4C2C1C3C5C2C12C3C1
XM:Z:.....hh.h.h.x..........h..........x..x......X...h.h....hh....h..h.h...h.....h..h...........x...h.
XR:Z:GA XG:Z:CT XB:Z:1
```

Read 2

methylation call

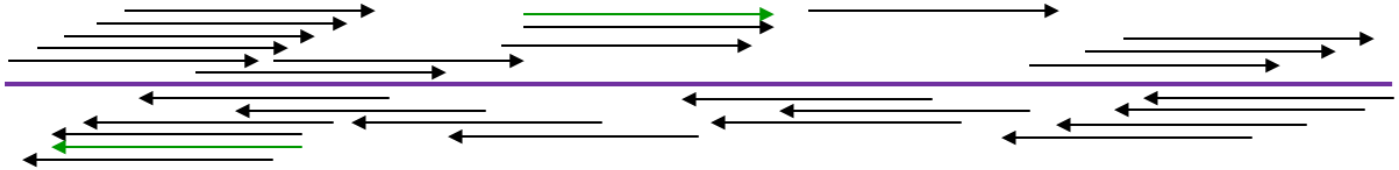
```
          me          me
5'-TTGGCATGTTTAAACGTT-3'   bisulfite read
5'...ccggcatgtttaaacgct...3'   genomic sequence
   ↓↓   ↓            ↓  ↓
   xz..H.........Z.h.      methylation call
```

z	unmethylated C in CpG context
Z	methylated C in CpG context
x	unmethylated C in CHG context
X	methylated C in CHG context
h	unmethylated C in CHH context
H	methylated C in CHH context

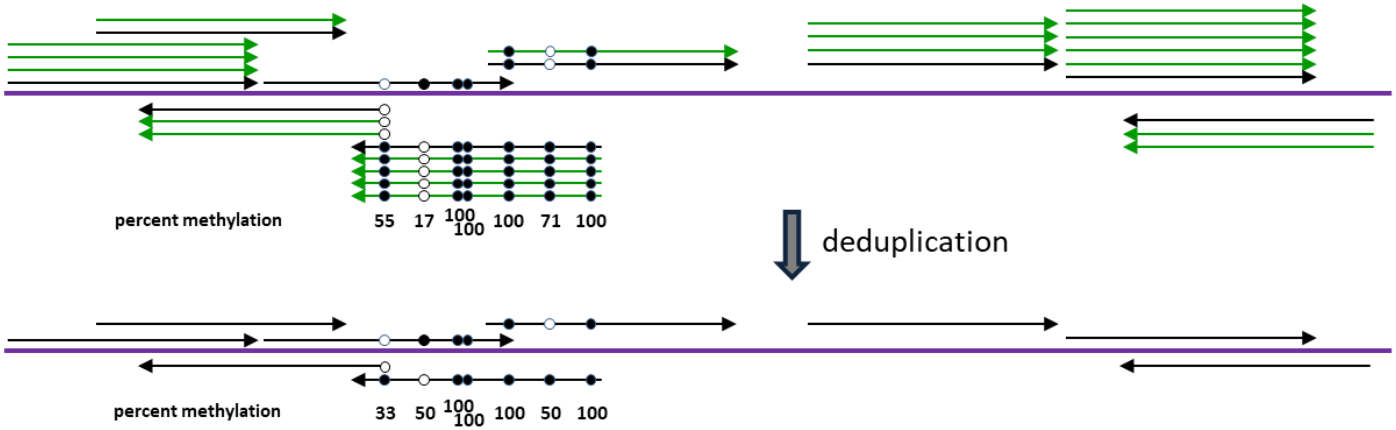
1. Quality control → 2. Bisulfite alignment → 3. Methylation Quantification → 4. Visualization → 5. DMRs detection

Sequence duplication

Complex/diverse library:



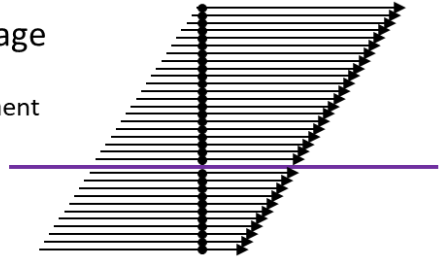
Duplicated library:



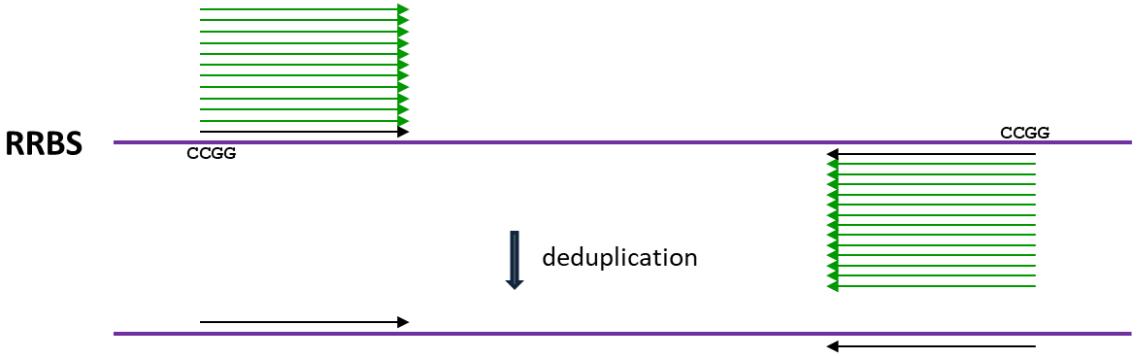
Sequence Deduplication - considerations

Advisable for large genomes and moderate coverage

- unlikely to sequence several genuine copies of the same fragment amongst >5bn possible fragments with different start sites
- maximum coverage with duplication may still be (read length)-fold (even more with paired-end reads)



NOT advisable for RRBS or other target enrichment methods
where higher coverage is either desired or expected



Methylation extraction – output .bed/bedGraph file

.bed/bedgraph file format (UCSC)

Following the track definition line are the track data in four column BED format:

```
chromA  chromStartA  chromEndA  dataValueA
chromB  chromStartB  chromEndB  dataValueB
```

Example of methylation extraction output (bedGraph/coverage output):

1	5705370	5705370	100	1	0		
1	5706335	5706335	60	3	2		
1	5706336	5706336	100	3	0		
1	5706453	5706453	75	3	1		
1	5706454	5706454	0	0	2		
1	5706845	5706845	71.4285714285714			5	2
1	5706846	5706846	66.666666666667			2	1
1	5707925	5707925	0	0	1		
1	5707926	5707926	66.666666666667			2	1
1	5709177	5709177	100	2	0		
1	5709178	5709178	0	0	1		
1	5710030	5710030	66.666666666667			4	2

Chromosome start end Methylation percentage meth unmeth

Decide early on which data to use

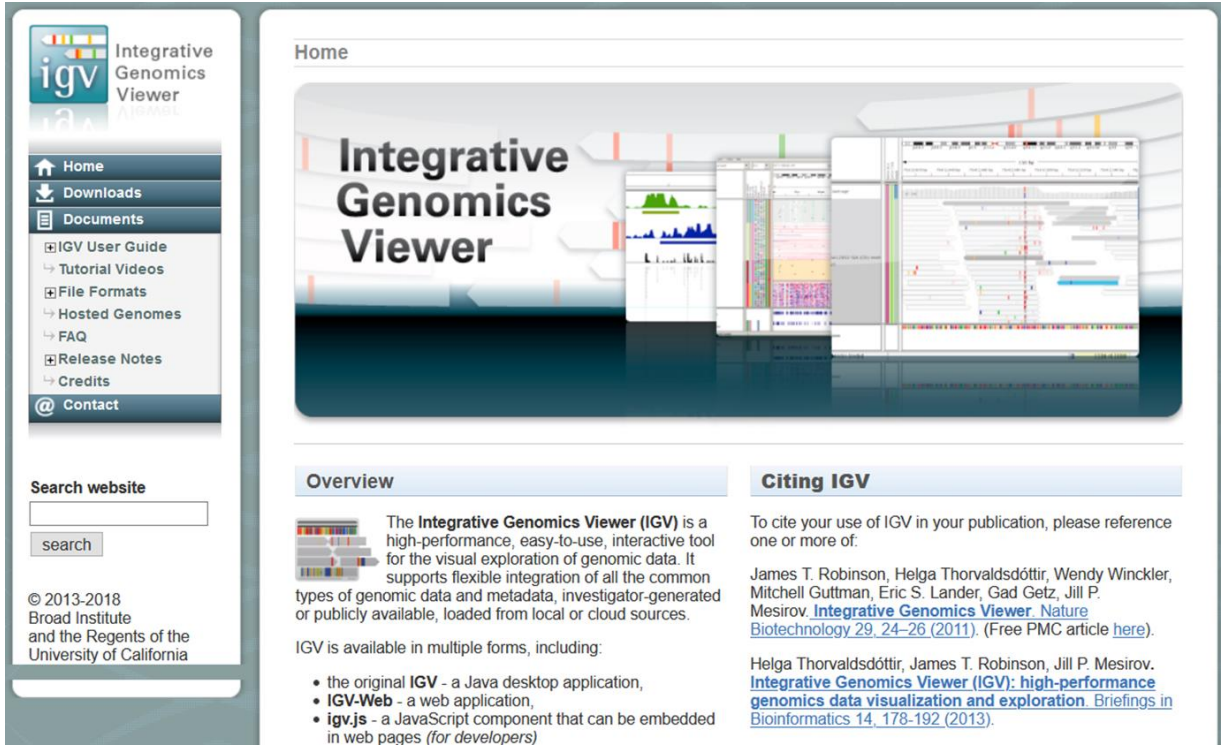
Methylation contexts

- CpG: Only generally relevant context for mammals
- CHG: Only known to be relevant in plants
- CHH: Generally unmethylated

Methylation strands

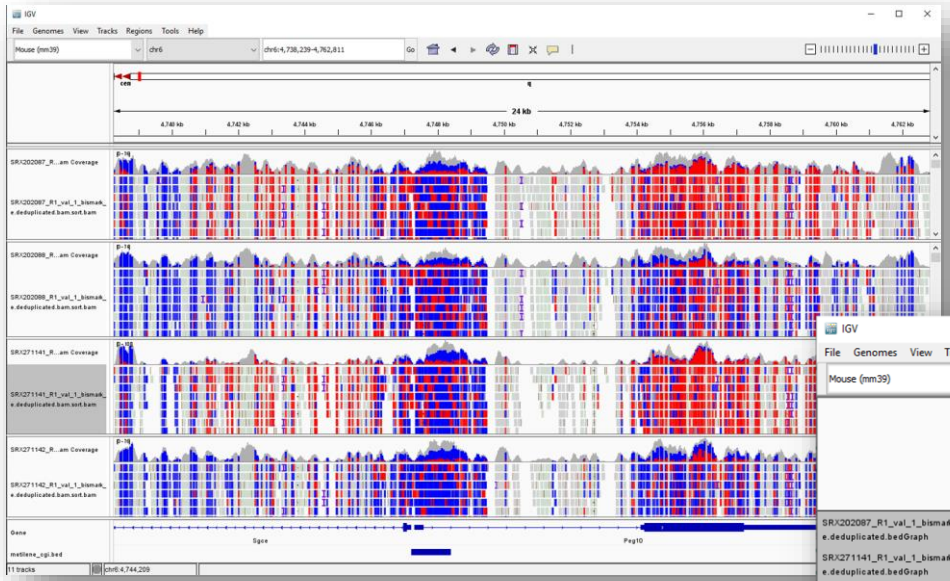
- CpG methylation is generally symmetric
- Normally makes sense to merge OT / OB strands

Always start by looking at your data …



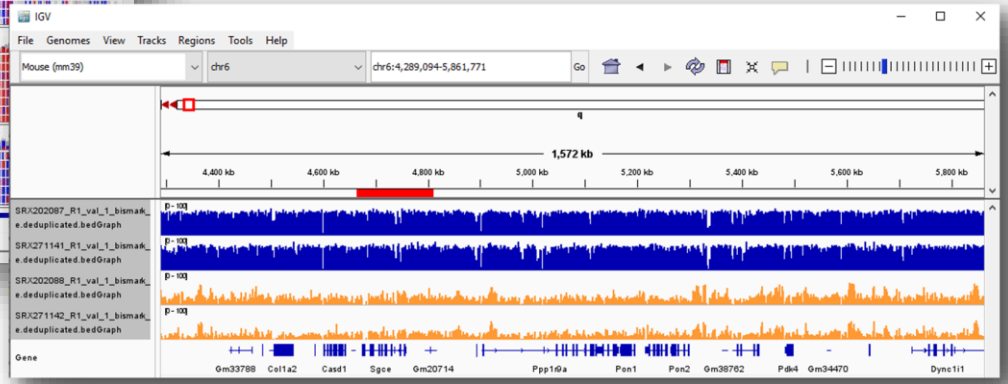
1. Quality control → 2. Bisulfite alignment → 3. Methylation Quantification → **4. Visualization** → 5. DMRs detection

Visualization of DNA Methylation: IGV



View the alignment (bisulfite colored mode)

View the methylation profiles



1. Quality control → 2. Bisulfite alignment → 3. Methylation Quantification → **4. Visualization** → 5. DMRs detection

Differentially Methylated Regions (DMRs)

- **DMRs**: regions that exhibit consistently different DNA methylation levels between sample groups (e.g., cases vs. control).
- DMRs can be a single C (differentially methylated cytosine, DMC) or as large as an entire gene locus.
- Size: a few hundred to a few thousand bases

Differentially Methylated Regions (DMRs)

DMR detection:

- Basic: t-test, Wilcoxon rank-sum test
- Advanced: mixture models, Shannon entropy, feature selection, logistic M values ....

Test at large number of genomic loci

- Correction for multiple hypothesis testing: false discovery rate (FDR: q-value)
- Only strongest single-CpG difference tend to remain significant

To improve the statistical power

- Larger genomic regions
- Pre-selected set of candidate genomic regions

Resolution levels

a

Genomic DNA sequence	CG	...	CG	...	...	...	CG	...	CG	...	...	...	CG	...	CG	...	CG	...	CG	...	CG	...	...	CG
Cases: Sample 1	3%		6%				80%		57%				1%		0%		1%		1%		42%			78%
Cases: Sample 2	2%		0%				50%		74%				0%		1%		0%		0%		38%			85%
Cases: Sample 3	0%		1%				95%		86%				2%		0%		0%		0%		41%			67%
Controls: Sample 4	0%		2%				8%		1%				12%		3%		15%		8%		36%			72%
Controls: Sample 5	1%		4%				5%		2%				15%		5%		33%		11%		39%			94%
Controls: Sample 6	0%		2%				13%		1%				19%		2%		24%		22%		33%			92%

Single CpG

b

Single-CpG analysis	CG1	CG2	CG3	CG4	CG5	CG6	CG7	CG8	CG9	CG10
Higher in cases (*q* value)	0.333	0.993	0.085	0.068	0.993	0.993	0.993	0.993	0.196	0.993
Higher in controls (*q* value)	0.993	0.732	0.993	0.993	0.070	0.104	0.104	0.110	0.993	0.351

Tilling region

Larger

c

Genome-wide tiling analysis	Tiling region 1	Tiling region 2	Tiling region 3	Tiling region 4	Tiling region 5	Tiling region 6	Tiling region 7	Tiling region 8
Higher in cases (*q* value)	0.549		0.048*		0.988	0.988	0.549	0.988
Higher in controls (*q* value)	0.768		0.993		0.067	0.067	0.299	0.299

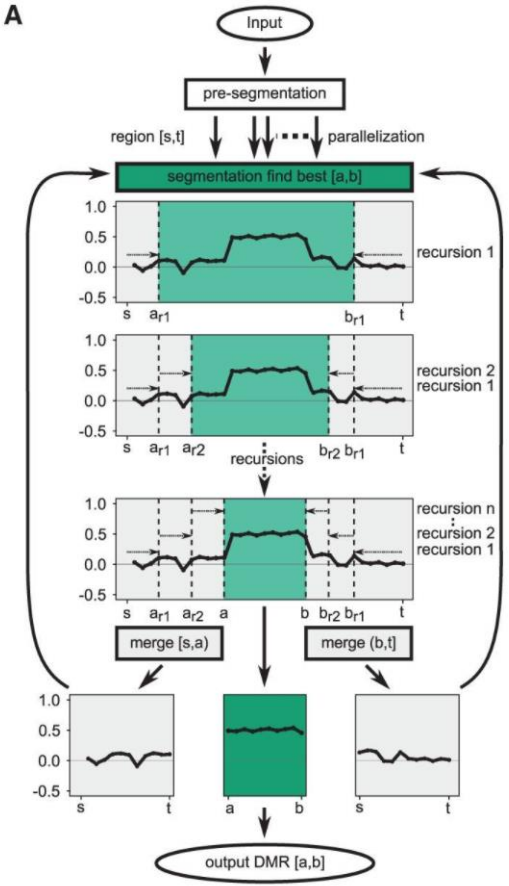
Annotated region

Largest

d

Annotated genome analysis	Enhancer	Promoter region	First exon
Higher in cases (*q* value)	0.024*	0.986	0.353
Higher in controls (*q* value)	0.993	0.045*	0.299

Bock, *Nat Rev genet*, 2012



DMR detection - metilene

- The objective problem of finding DMRs has two dimensions:
  - find a genomic region
  - the individuals of two groups are significantly distinct in their methylation levels.
- De-novo DMR detection

Differentially variability

VMCs/VMRs

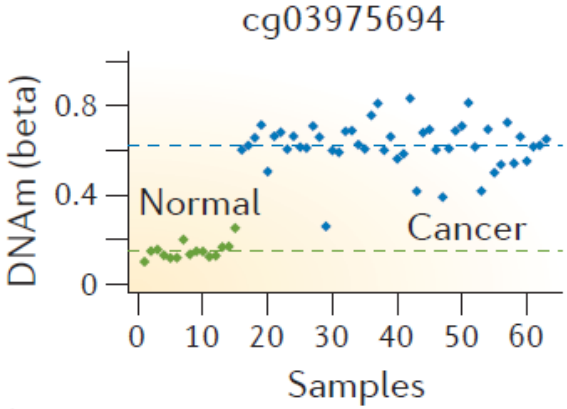
Variably methylated cytosines/regions

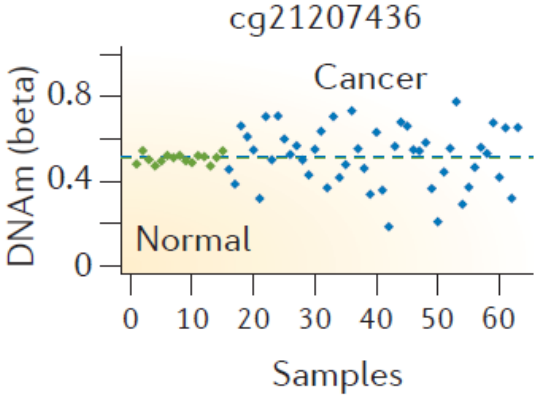
Differentially methylated cytosines/regions (Most common supervised features)

DMCs/**DMRs**

DVCs/**DVRs**

Differentially variable cytosines/regions





1. Quality control → 2. Bisulfite alignment → 3. Methylation Quantification → 4. Visualization → 5. DMRs detection

Teschendorff, A. E., & Relton, C. L., *Nat Rev genet*, 2018

The end.