Introduction to DNA Methylation

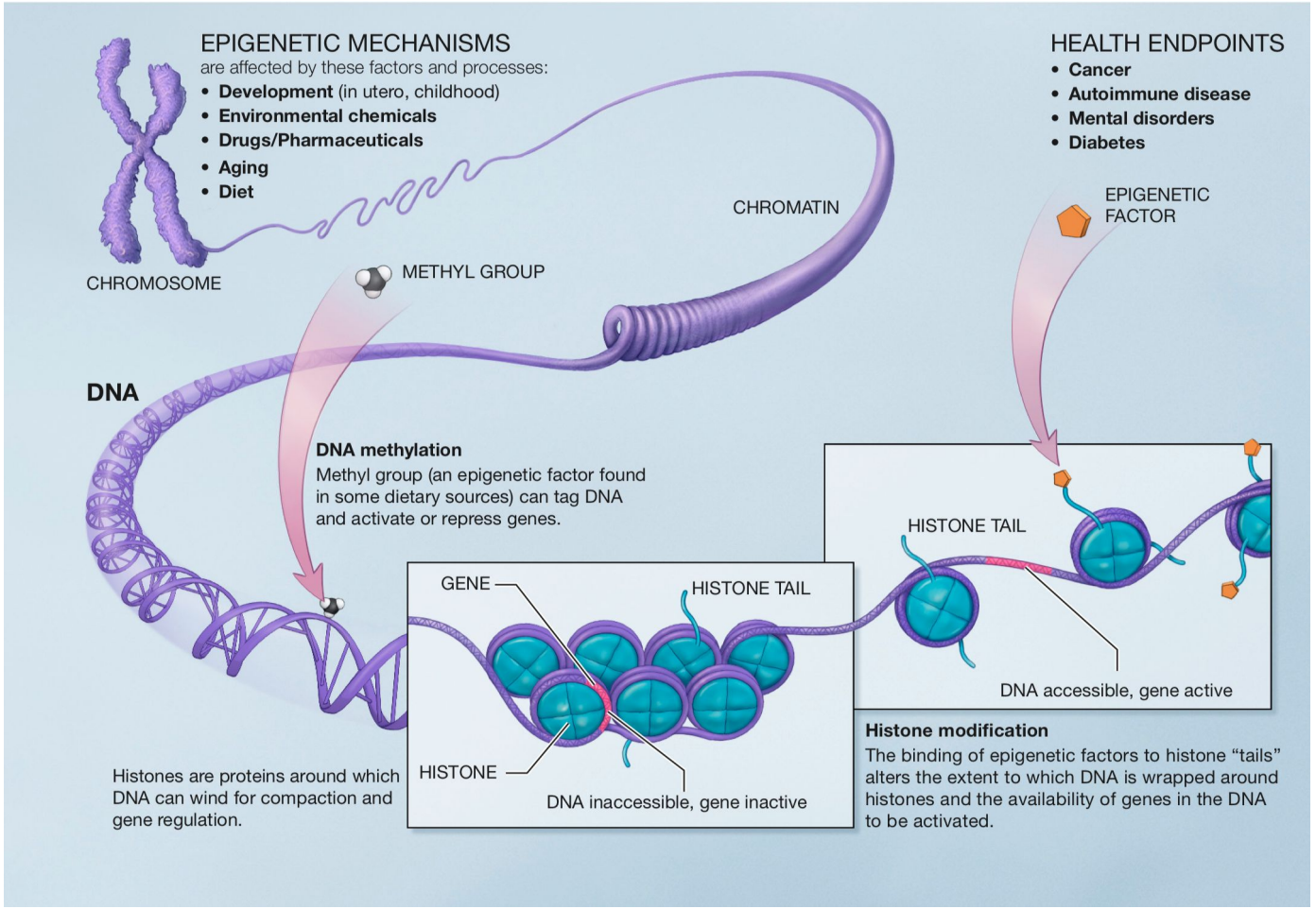
Roman Chernyatchik,
April 24th, Saint Petersburg

Epigenetics

Studies heritable phenotype changes that do not involve alterations in the DNA sequence

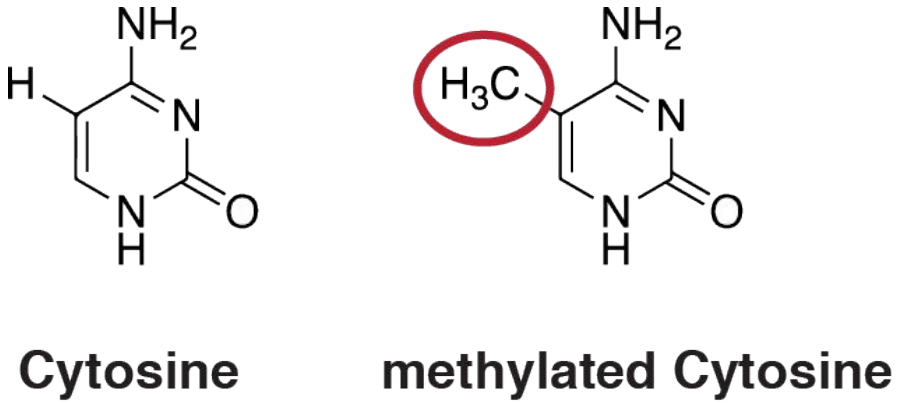
Epigenetics Regulation via:

- Histone Modifications
- Chromatin accessibility and 3D structure
- Non-Coding RNAs
- DNA cytosine methylation



https://commonfund.nih.gov/sites/default/files/epigeneticmechanisms.pdf

What is DNA Methylation?



Cytosine **methylated Cytosine**

Occurs at:

- Cytosine (much better studied)
- Adenine

Key Dates

- 1925: 5mC was first reported by **Johnson** and **Coghill** as one of the hydrolysis products of tuberculinic acid
- 1948: **Hotchkiss** conclusively identified 5mC in calf thymus DNA
- 1975: Two papers independently suggested that methylation of cytosine residues in the context of CpG dinucleotides could serve as an epigenetic mark in vertebrate (**Holliday** & **Pugh**; **Riggs**)
- 2006: iPS Cells technology (Shinya **Yamanaka**)
- 2003 **Horvath** methylation clock

Cytosine Contexts

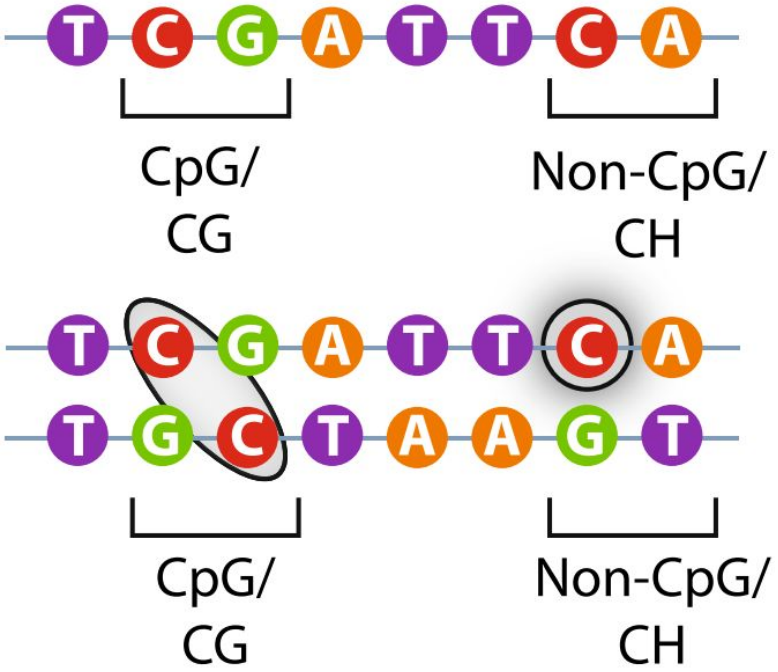
- CpG
- Non-CpG
  - Lister 2009: CHH, CHG
  - Now: CH (H=A,T,G)

Humans:

- Mostly CG methylation
- Non-CpG slightly methylated in ES cells, neurons, almost non methylated in differentiated cells

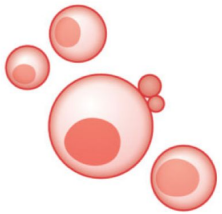
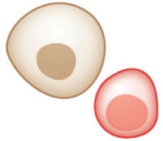
Plants:

- Both contexts could be methylated



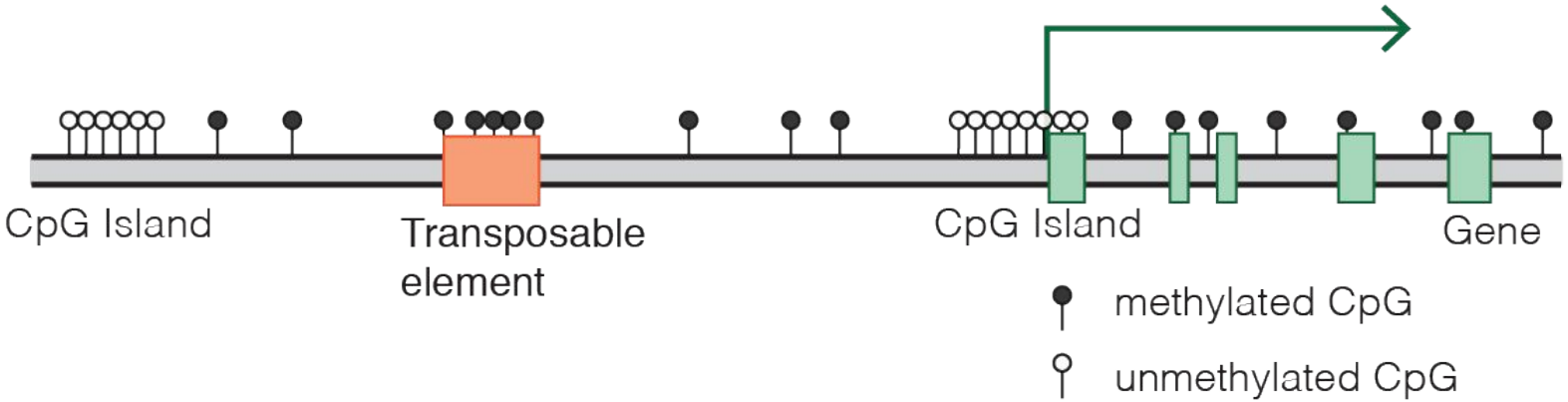
(Masser et al. 2018)

Non-CpG Methylation Levels in Tissues

mCH level		Me Blue: CAC	Me Red: CAG	Brown: unknown
Extremely abundant (>2%)	Neuron			
Abundant (1–2%)	Frontal cortex (>5 years old)	H1 ESC SCNT-SC iPSC Oocyte First polar body Second polar body		
Detectable (<1%)	Frontal cortex (1 month–5 years old) Glia	Heart Gastric Adipose Psoas Adrenal gland Aorta	H9 ESC ICM MSC Mesendoderm TBL	
Undetectable (~0%)	Frontal cortex (fetal) NPC	Sigmoid colon Small bowel	Sperm Fibroblast	

(He et al. 2015)

Typical Mammalian DNA methylation landscape



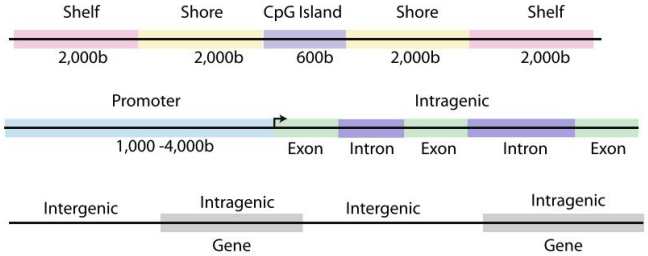
CpG Island (CGI) - DNA regions with frequency of the CG sequence is higher than other regions.Many genes

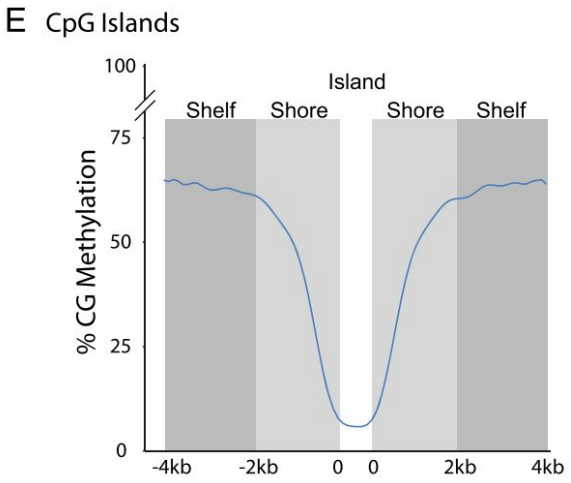
- 60% human genes has CGI in promoter

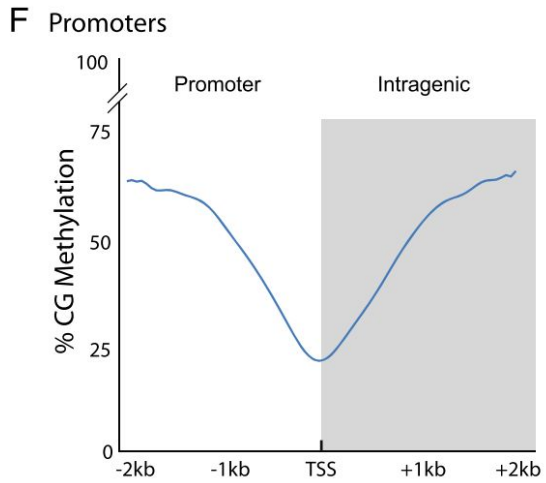
https://en.wikipedia.org/wiki/DNA_methylation

Methylation at TSS and CpG Islands

CG methylation across promoters often varies greatly across a relatively narrow region with lowest methylation typically observed around the transcription start site (TSS), even in un-expressed genes



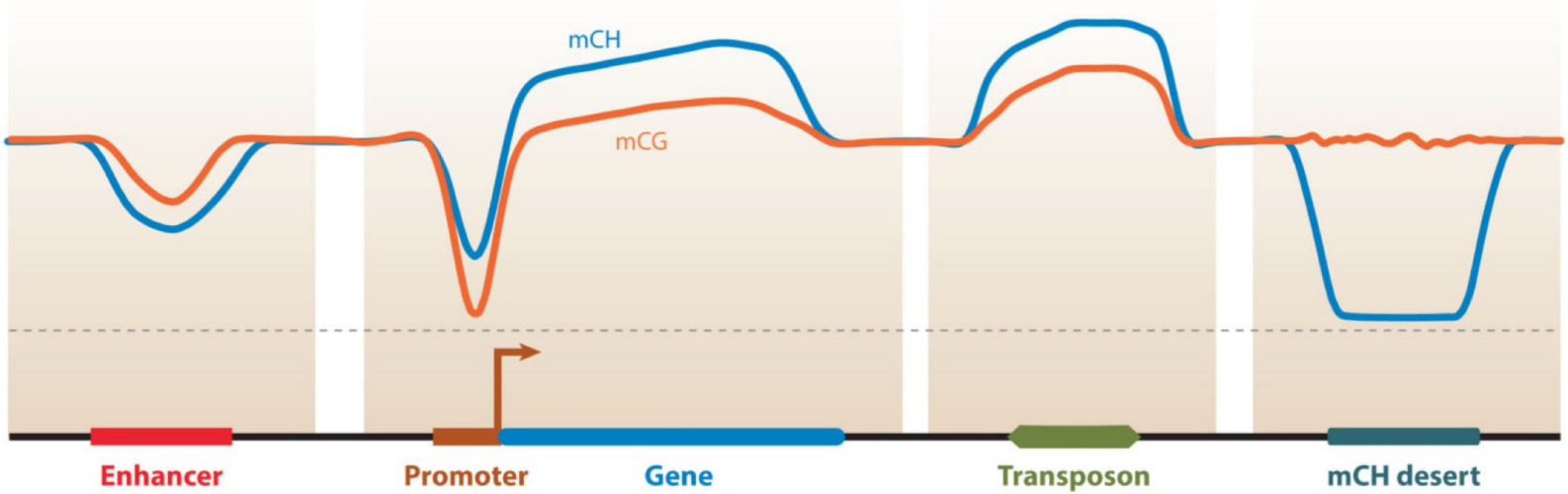




(Masser et al. 2018)

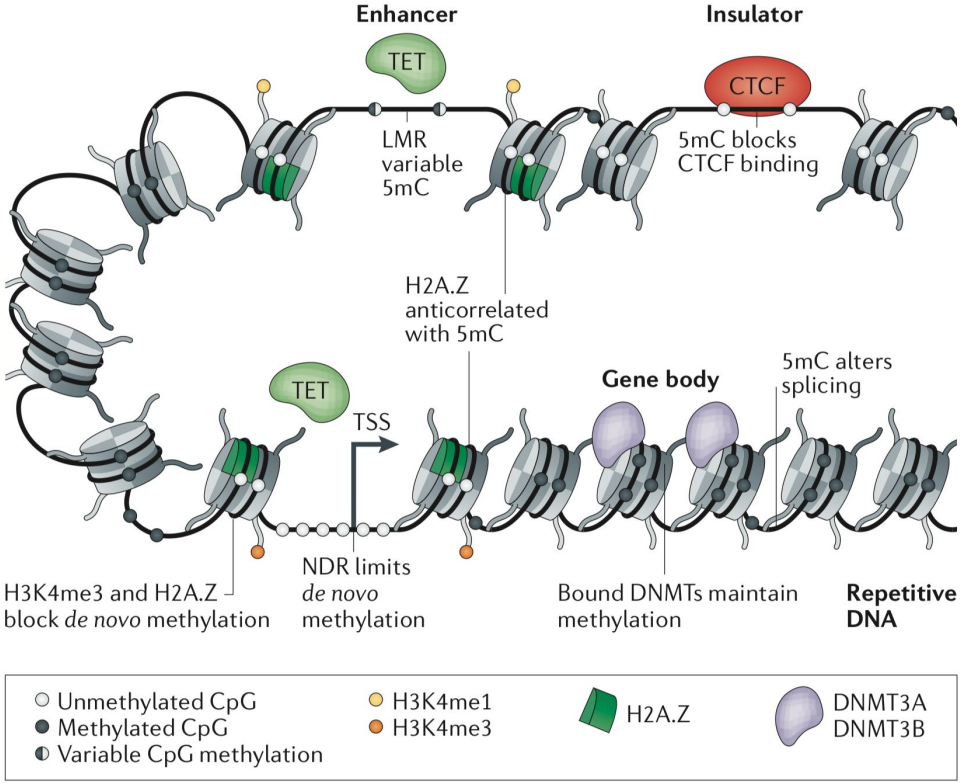
Average mCH and mCG at different genomic loci

The scale is schematic, and intergenic mCH and mCG levels have been scaled to be the same level.



(He et al. 2015)

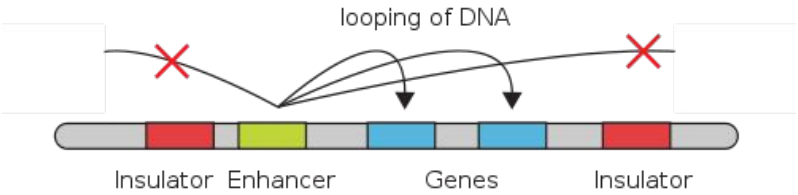
DNA Methylation Epigenetics Mechanisms



60% genes have CpG Islands

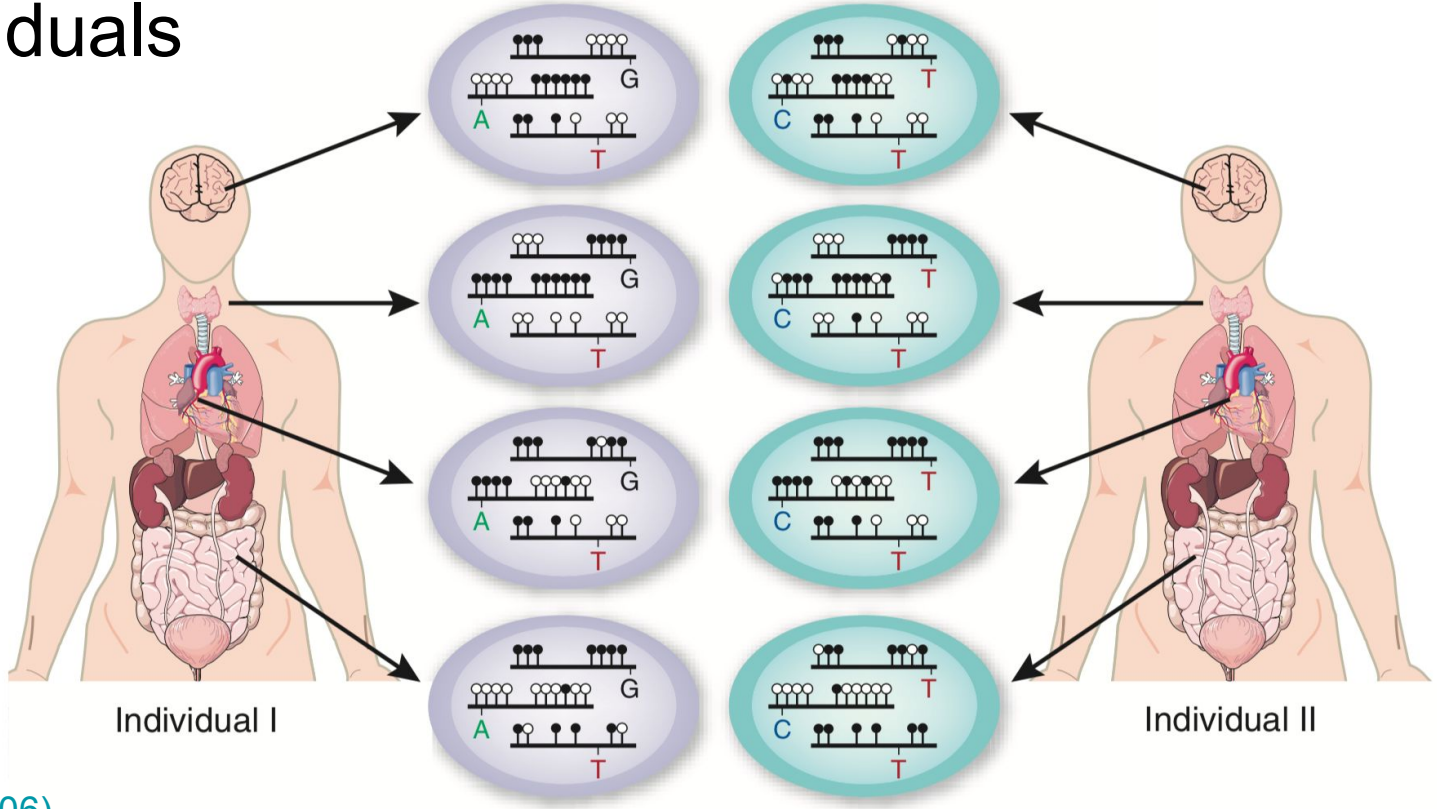
NDR - nucleosome depleted region

LMR - low methylated region



Jones (2012)

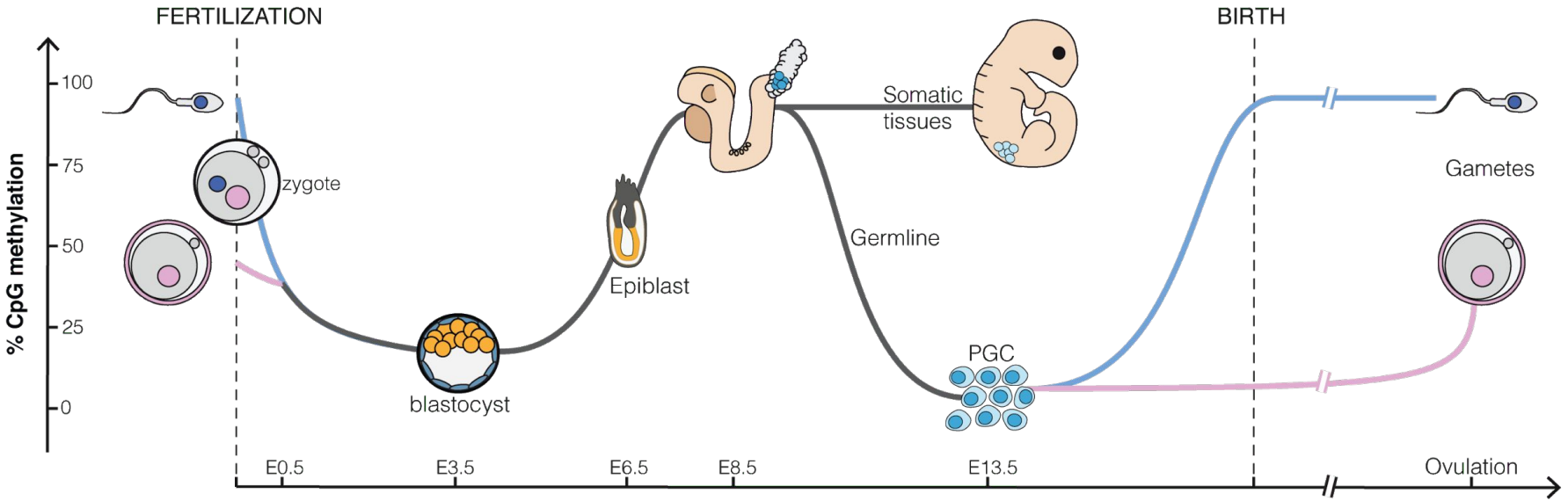
DNA Methylation is tissue specific & vary across individuals



Brena (2006)

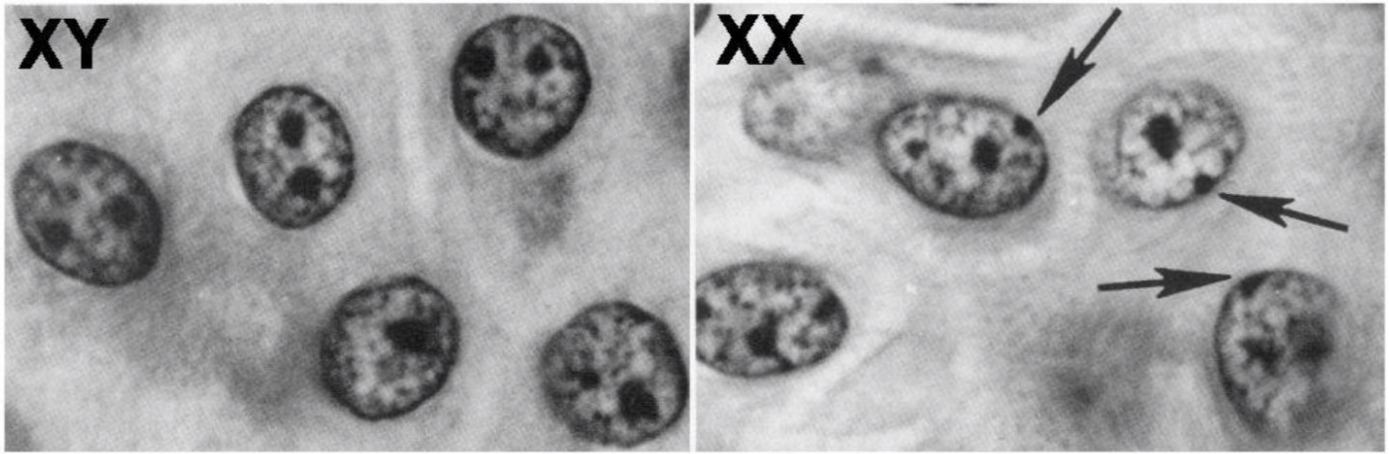
Kim Caesar

Embryonic Reprogramming



https://en.wikipedia.org/wiki/DNA_methylation

Barr Body - Inactivated X chromosome (Xi)

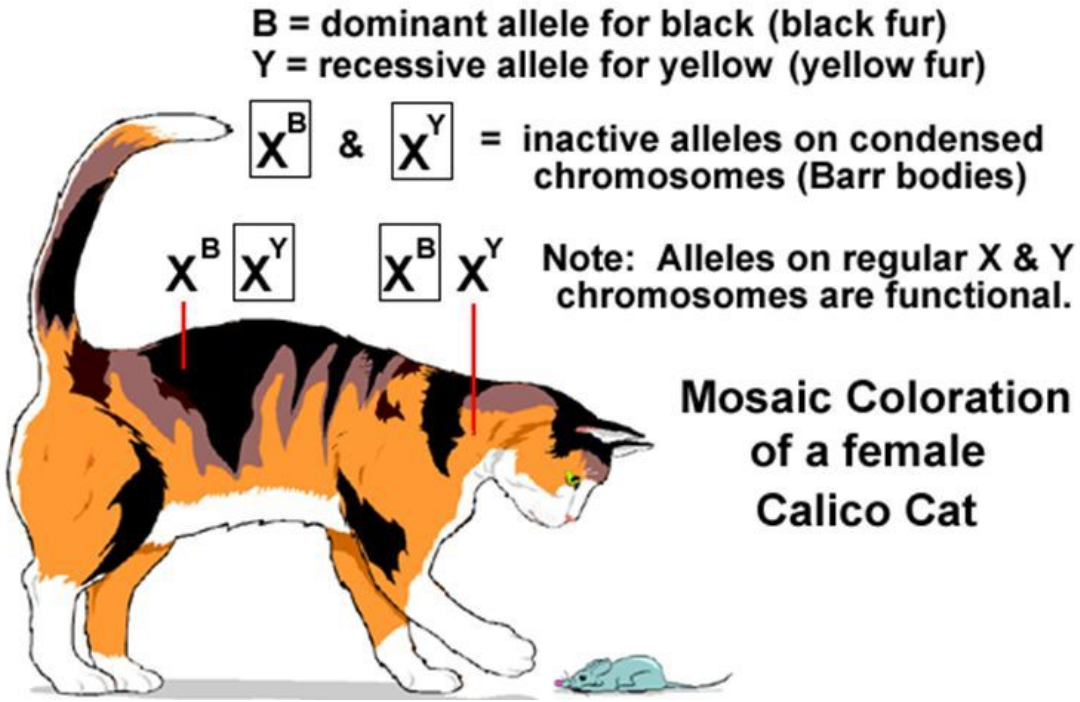


Xi:

- Do not express the majority of genes
- High DNA methylation level
- Low level: H3K4me3, H3K27ac
- High levels: H3K27me3, H3K9me3

Calico cats

Red, Black fur colors are X-linked

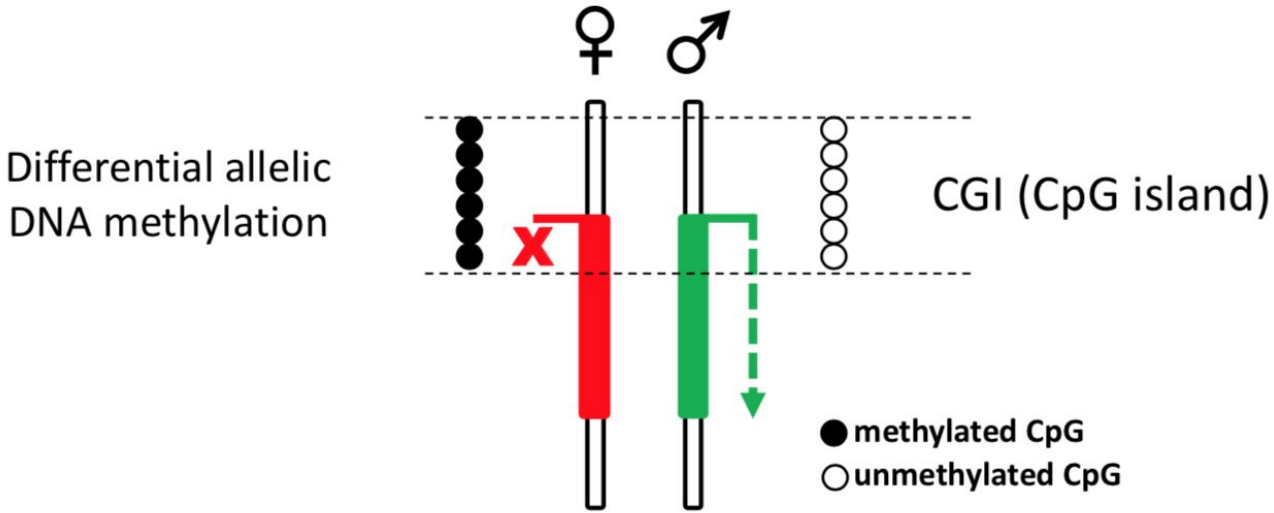


Transgenerational transmission of information

Epigenetics - heritable changes:

1. High temperature-induced change in expression in C. elegans endure at least 14 generations (Klosin et al, 2017)
2. Persistent epigenetic differences associated with prenatal exposure to famine in humans (Heijmans et al, 2008) – “Here we show that individuals who were prenatally exposed to famine during the Dutch Hunger Winter in 1944–45 had, 6 decades later, less DNA methylation of the imprinted IGF2 gene compared with their unexposed, same-sex siblings.”

Imprinted Genes: mono-allelic expression



Imprinted genes - key roles in energy metabolism, placenta functions.

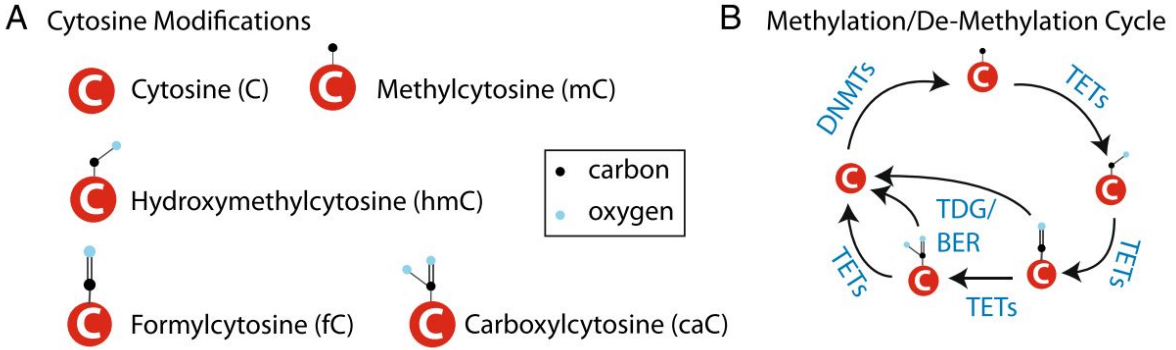
Imprinting failure leads to disease – e.g. Angelman syndrome, etc.

DNA Methylation / De-Methylation

C → mC:
DNA methyltransferases (DNMTs) add methyl groups

mC → C:
Tet methylcytosine dioxygenases (TETs) sequentially oxidize modifications

- back to an unmodified cytosine
- or include base excision repair (BER) through thymine DNA glycosylase (TDG).



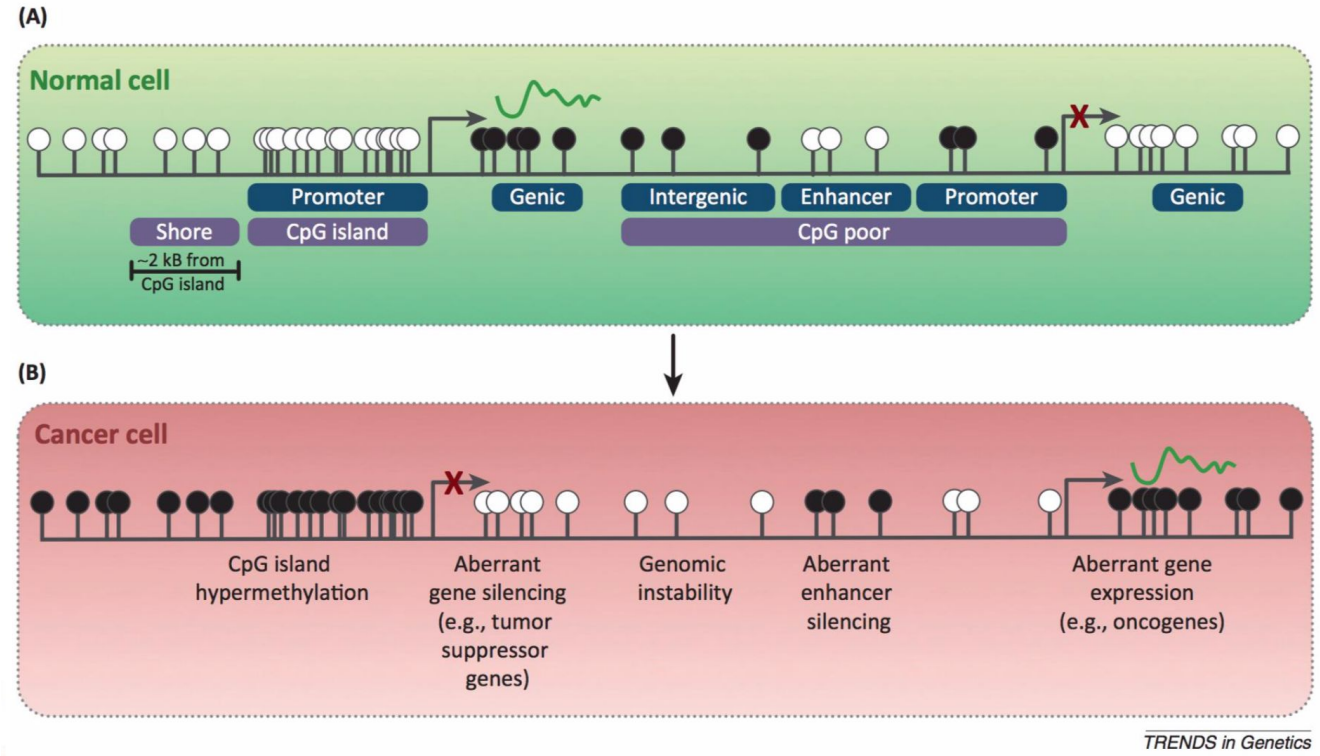
(Masser et al. 2018)

DNMT Types

- **DNMT1**
  Maintenance methyltransferase in mammals. It predominantly methylates hemimethylated CpG di-nucleotides in the mammalian genome
  - 3 nucleotides recognition motif: a C on one strand and CpG on the other
- **DNMT3A, DNMT3B, DNMT3L**
  De-novo methylation. Could methylate hemi methylated and unmethylated CpG at the same rate
  - DNMT3a prefers CpG methylation
  - DNMT3L is required for establishing maternal genomic imprints in gametogenesis.

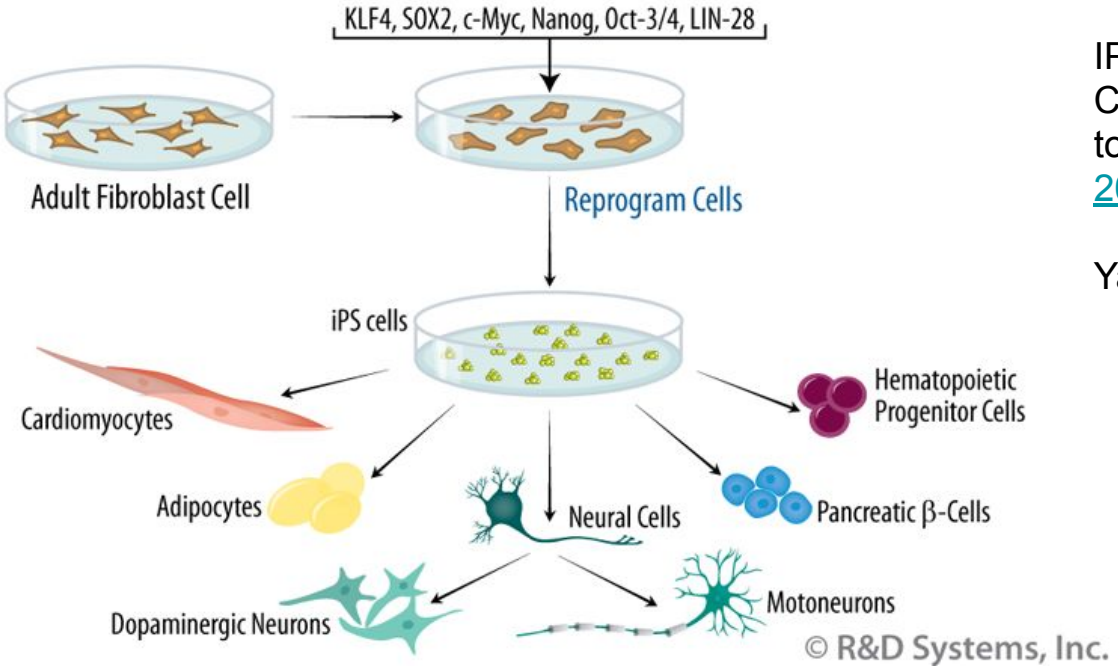
https://en.wikipedia.org/wiki/DNA_methyltransferase

Cancer Methylomes



(Stirzaker et al, 2014)

Cells Reprogramming



(Image source)

IPSc (Induced Pluripotent Stem Cells) - somatic cells reprogrammed to a pluripotent state (Yamanaka, 2006).

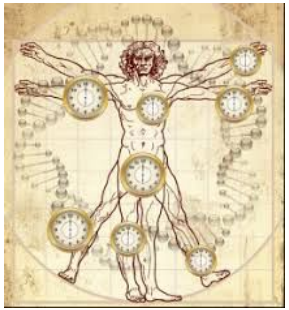
Yamanaka factors:

- Oct3/4
- Sox2
- c-Myc
- Klf4

Methylation Clock

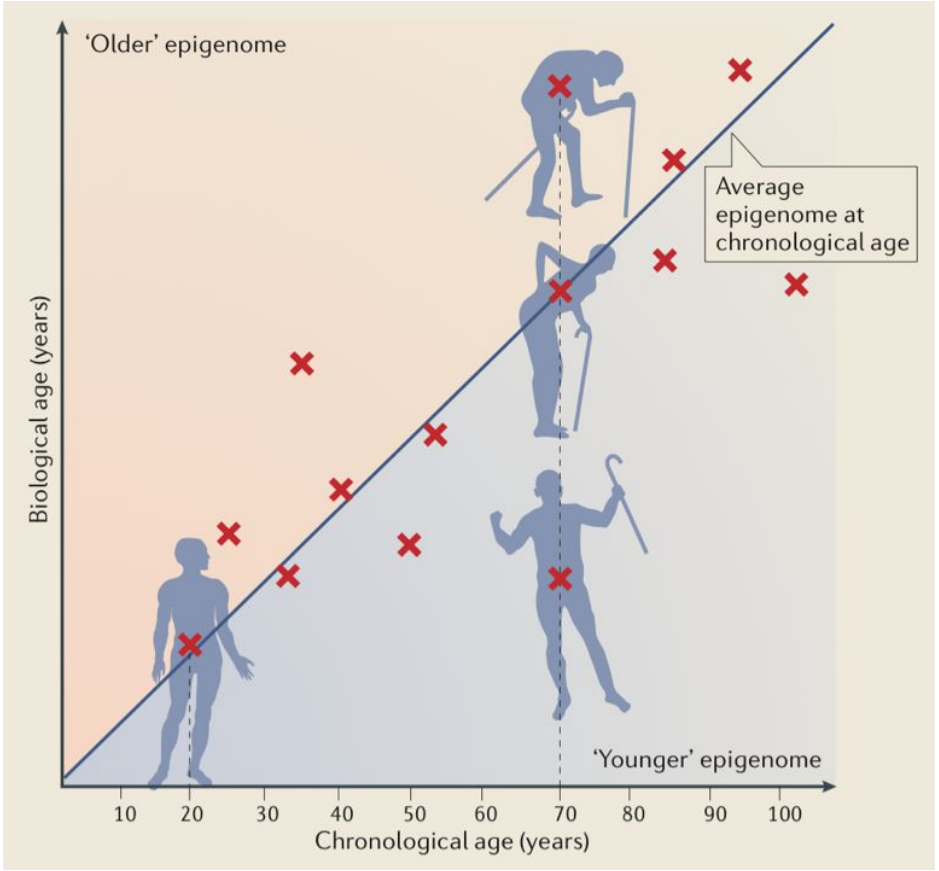
Clock models

- 2013: Horvath 353 CpG, multi tissue
- 2013: Hannum 71 CpG, blood
- 2012: Garagani, 1 CpG in ELOVL2 gene, whole blood
- 2018: PhenoAge (Levine), 513 CpG, whole blood



Consistent with:

- Long-livers
- Cancer
- All-cause mortality

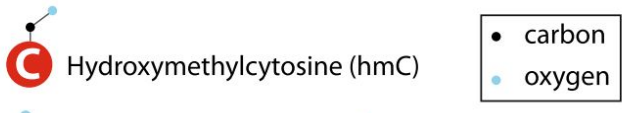


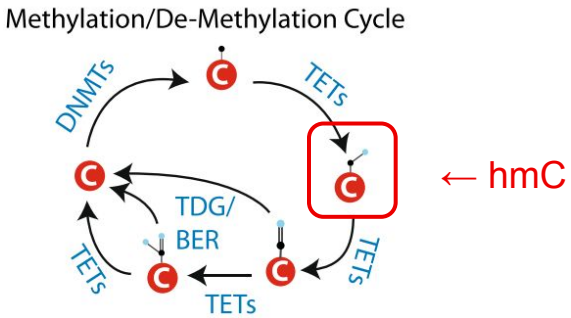
(image source)

5-hydroxymethylcytosine (hmC)

Demethylation intermediate product

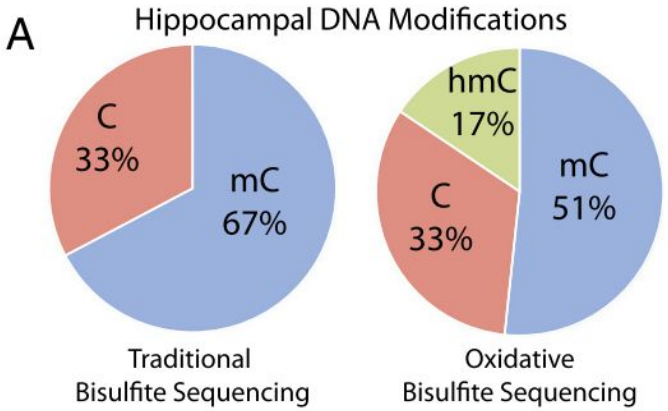
- Not enough studied

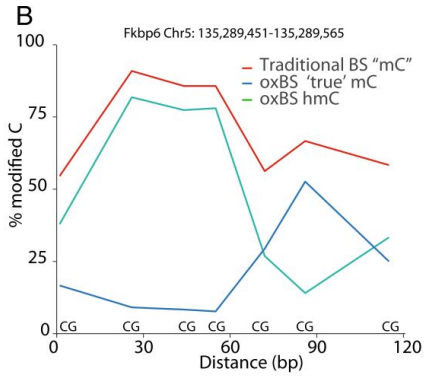


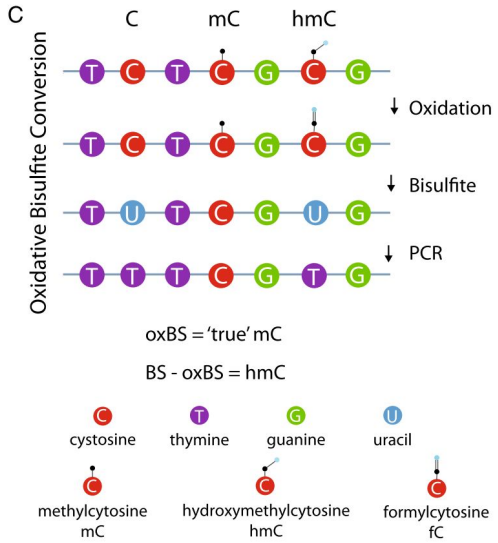


(Masser et al. 2018)

Hydroxy methylation (hmC)



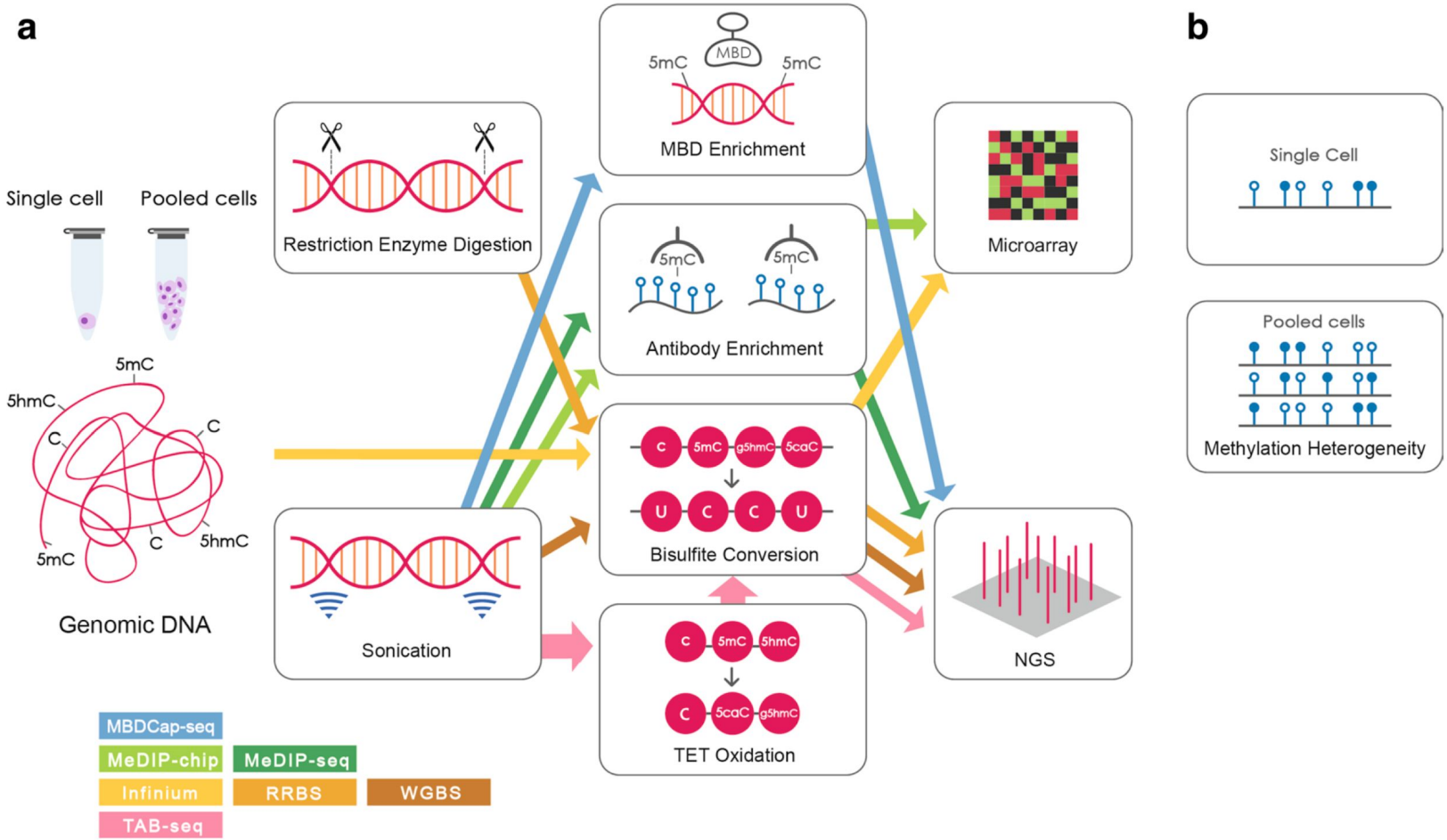




- hmC considered as mC by WGBS
- 5-hmc plays role in epigenetics
- Gene Body
  - mCG positively corr with gene expression
  - hmCH pos corr
  - mCH - both pos and negative
- Promoter
  - mCH - pos cor with expression
  - mCG - represses
  - hmCG - pos assoc with gene expression

(Masser et al. 2018)

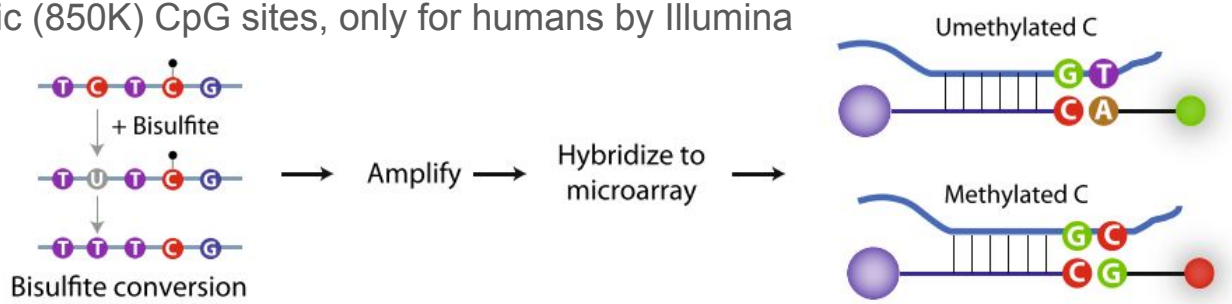
DNA Methylation Measuring Protocols



(Yong et al. 2016)

Methylation Detection

- Before NGS:
  - Liquid chromatography (HPLC, Reddy and Reddy 1990)
  - Enzyme-linked immunosorbent assays (ELISA) with antibodies for mc/hmC
  - Methylation microarrays (Bibikova et al. 2009) ~ SNP arrays, Bisulfite Conversion: 27K, 450K, Epic (850K) CpG sites, only for humans by Illumina



- NGS Affinity enrichment ~ immunoprecipitation
  - meDIP-Seq, MBD-Seq, ..
- NGS with single base resolution
  - WGBS, RRBS/ERRBS, BOCS, BSAS, SC-WGBS

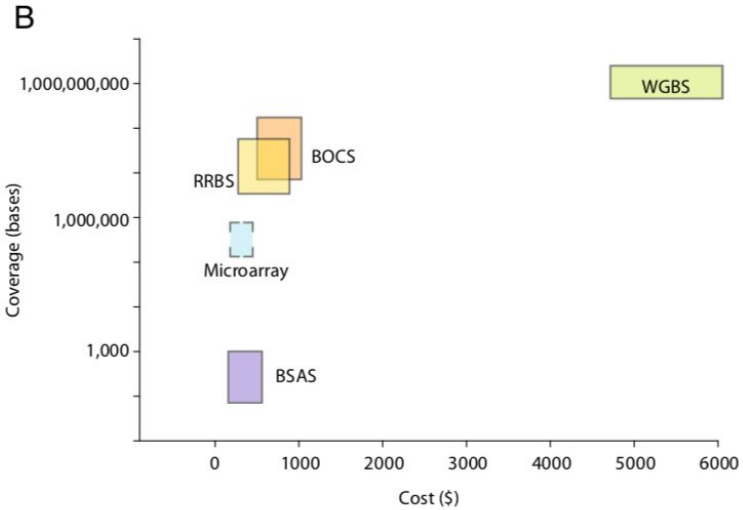
(Masser et al. 2018)

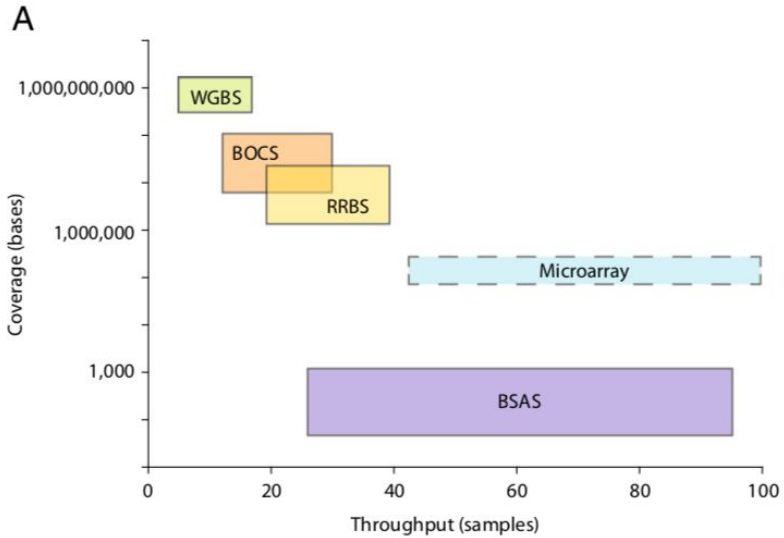
Methylome Resolution

Human ~ 28 millions CpG sites

- Microarrays:
  - 27.000, 450.000, 800.000 fixed CpG positions
- NGS Affinity enrichment
  - Low resolution (average signal at 1 kbp regions), legacy methods
- NGS with single base resolution
  - WGBS ~ 98% CpG sites + non-CpG
  - RRBS ~ 10% CpG sites + non-CpG (mostly CpG islands)

Protocols Comparison



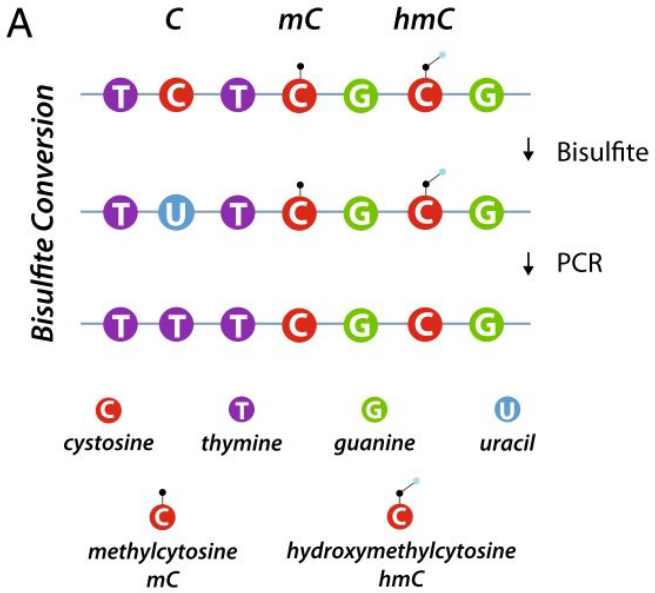


(Masser et al. 2018)

Bisulfite Conversion

Bisulfite Treatment

- C → U
- mC → mC
- hmC → hmC

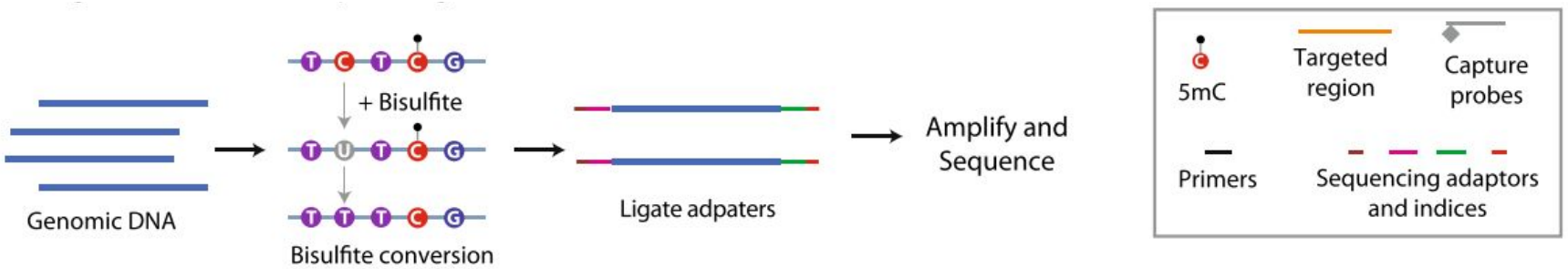


NB:

- Doesn’t differ: mC & hmC
- Bisulfite conversion could be over/under done => wrong results
  - Non-CpG as not methylated: n/a for stem cells, oocytes, brain; plant or insect DNA
  - chrMT: Actually seems to be methylated (Vitolacobazzi et al., 2013)
  - Under conversion:
    - lambda genome: accurate
    - PhiX (unmethylated spike-in)
  - Over conversion: In vitro methylated T7 dsDNA (by CpG mc transferase sssI)
  - ERRBS: using 'C' provided by 3'-end repairing (rrbs guide)

WGBS

- Whole-genome bisulfite sequencing (**WGBS**, Cokus et al. 2008; Lister et al. 2008)

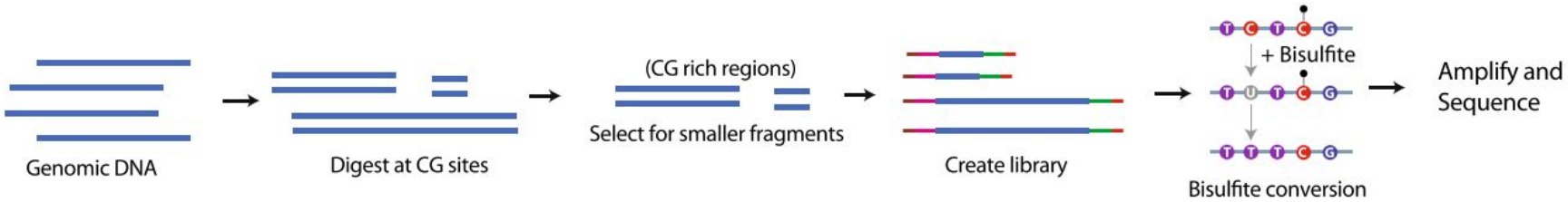


(Masser et al. 2018)

RRBS / ERRBS

- Reduced representation bisulfite sequencing **RRBS** / eRRBS (Meissner et al. 2005)
  - CCGG motif cleavage, 40-200 bp size selection ~ CGIs regions
  - Limited overlapping sites between different RRBS datasets (Stubbs et al. 2017)
  - Deduplication not applied to RRBS (cleavage by same sites)

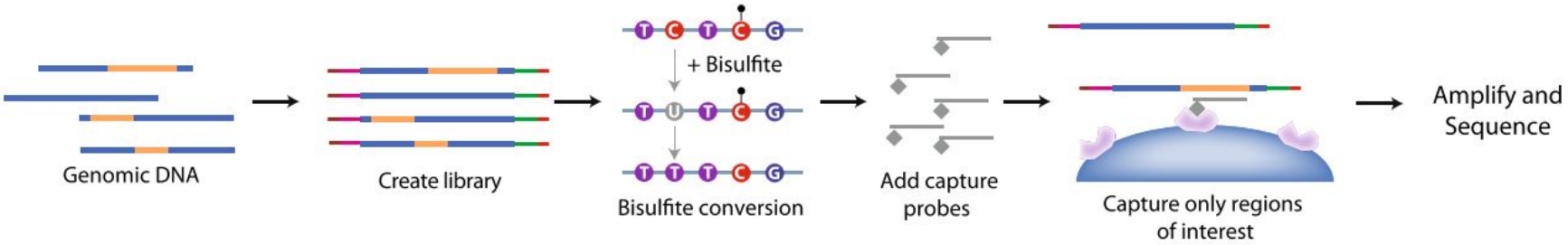
UMI-based modification against PCR-bias (Stubbs et al. 2017)



(Masser et al. 2018)

BOCS

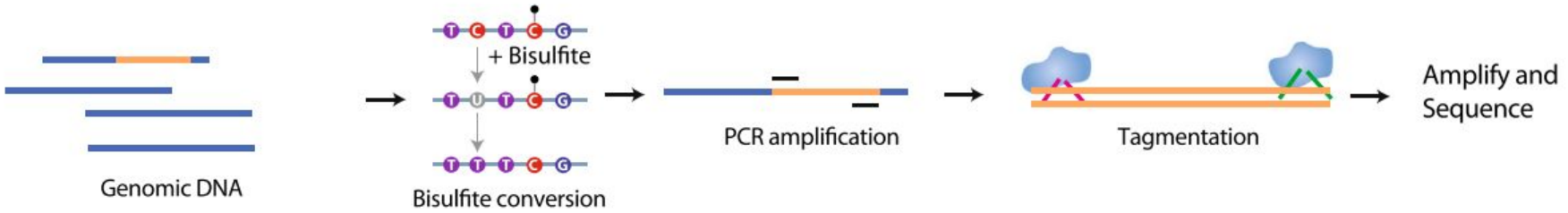
- Bisulfite oligonucleotide capture sequencing (**BOCS**, Wang et al. 2011)
  - ~ 80–200 Mb genomic regulatory elements



(Masser et al. 2018)

BSAS

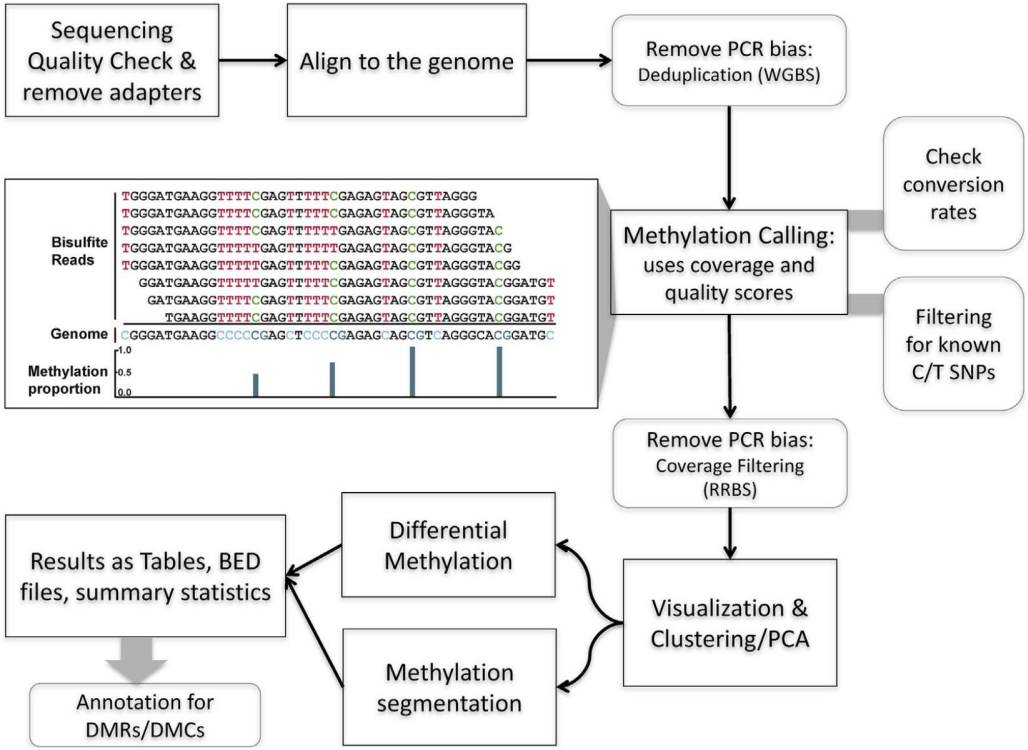
- Bisulfite amplicon sequencing (**BSAS**, Masser 2013)
  - ~ 1 Mb, using PCR primers



(Masser et al. 2018)

BS-Seq Data Processing Pipeline

1. QC & Adapters Trimming
2. BS Reads Alignment
3. WGBS Deduplication
4. Methylation calling
5. Methylation QC Metrics
6. Batch Effect correction
7. DMC, DMRs, Methylated segments
8. Annotate DMR / DMC / methylation segments

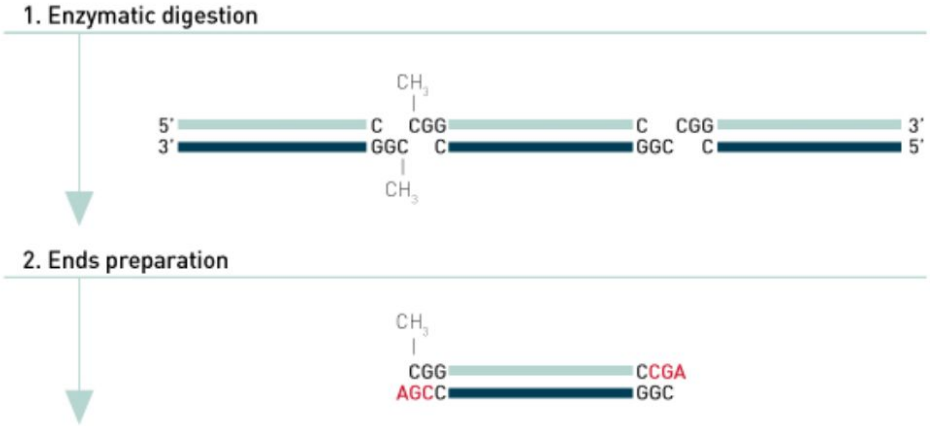


(K. Wreczycka et al. 2017)

RRBS Data Processing

~ BS-Seq pipeline

- Skip de-duplication step
  - Large % of duplicates due to CCGG cleavage
- Ends repairing bias:
  - “Unmethylated” bases used for end repairing
  - Need to trim R1 3’, R2 5’



Step 1: QC & Adapters Trimming

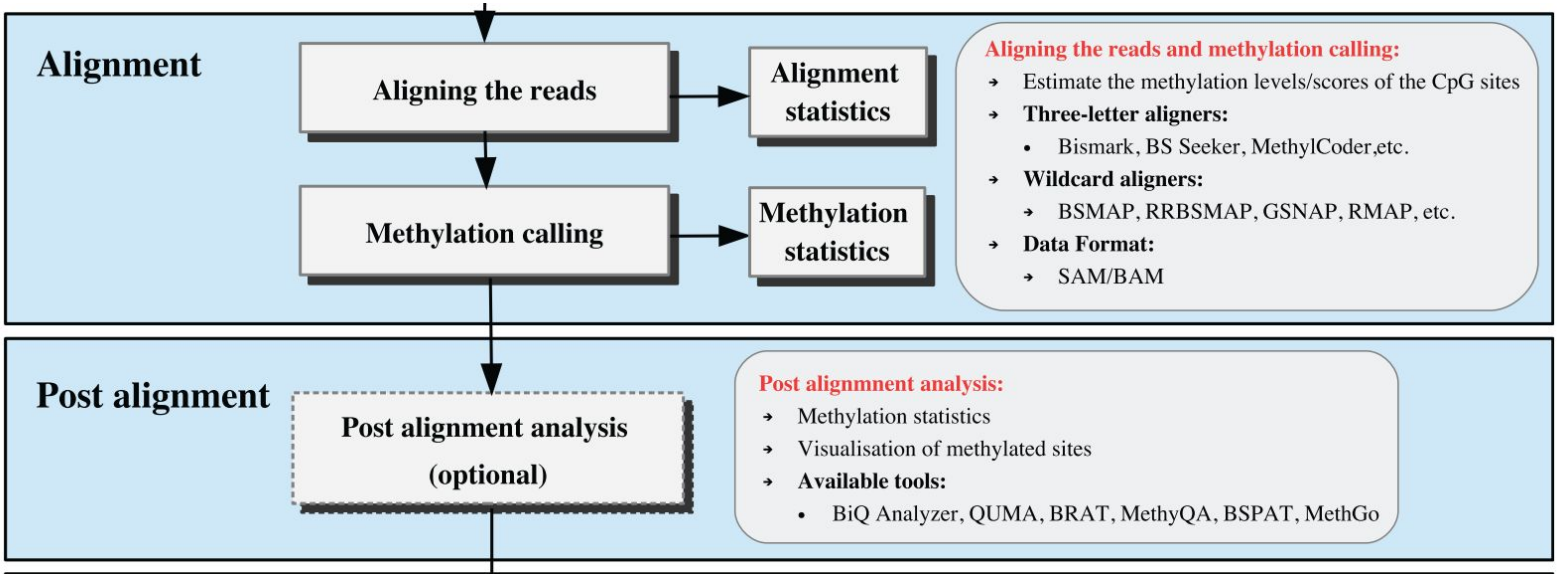
Reads QC: FastQC; AlterQC / FASTP

- Coverage is required 5–15x (Ziller et al. 2015) ~ 25 Gb per sample

Trimming, barcodes split:

- **Trim Galore (wrapper for Cutadapt)**
  - ENCODE WGBS PE Pipeline
- Mott Trimmer
  - ENCODE WGBS SE Pipeline
- AlterQC / FASTP
  - Trimming
  - PE reads qc (fastqc supports only SE reads)
- FASTX
  - Recommended by Book: Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing
- ...

Step 2: BS Reads Alignment



- 3 letter aligners
- Wildcard aligners

(Shafi et al. 2018)

3 letter aligners

Bowtie based aligners

- **Bismark**
  - By Babraham Bioinformatics Institute (FastQC, Trim Galore)
  - ENCODE Pipeline
  - De-facto Standard using Bismark tools
- BS-Seeker (3-letter alphabet + bowtie)

BWA aligner

- E.g. BWA-MEM aligner
- Faster than Bowtie2 for long reads

Others: Pash, Novoalign,LAST,... see (Tsuji et al. 2015)

Table 1.

Feature comparison of Bismark and BS Seeker

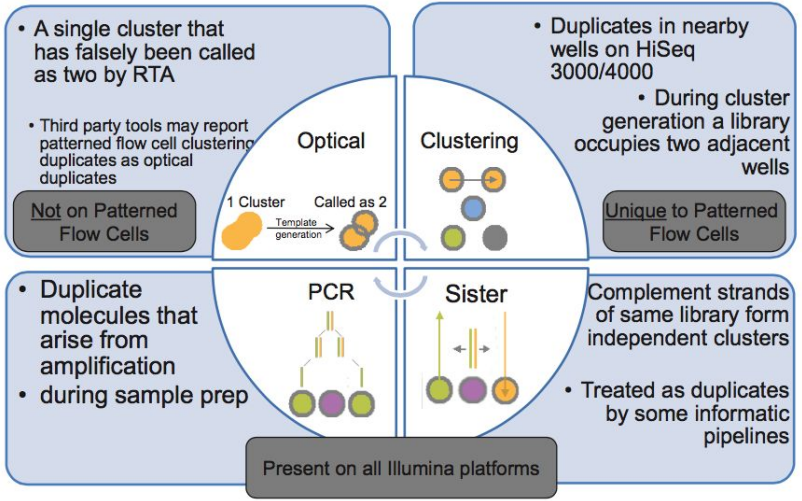
Feature	Bismark	BS Seeker
Bowtie instances (directional/non-directional)	4	2/4
Single-end (SE)/paired-end (PE) support	Yes/yes	yes/no
Variable read length (SE/PE)	Yes/yes	no/NA
Adjustable insert size (PE)	Yes	NA
Uses basecall qualities for FastQ mapping	Yes	No
Adjustable mapping parameters	5	2
Directional/non-directional library support	Yes/yes	Yes/yes[a]

(Krueger et al. 2011)

Step 3: WGBS Deduplication

Sources of duplicates:

- PCR
- Optical duplicates
- Clustering duplicates
- Sister duplicates



In WGBS deduplication step is applied (e.g in RRBS is skipped)

Tools:

- Bismark Deduplicate
  - ENCODE pipeline

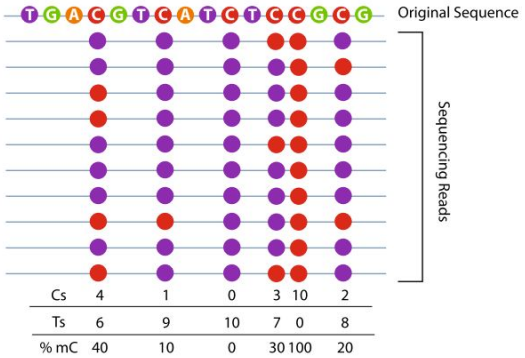
http://core-genomics.blogspot.com/2016/05/increased-read-duplication-on-patterned.html?m=1

Step 4: Methylation calling

Calculate methylated/unmethylated counts for each cytosine

Tools

- **Bismark**
  - ENCODE pipelines
- BS-Seeker, …



SNP issue: “C -> T” SNPs leads in DMC calling inaccuracies

- Remove SNP using genotyping info
- Bis-SNP bs-seq data genotyping & mc calling

Step 5: Methylation QC Metrics

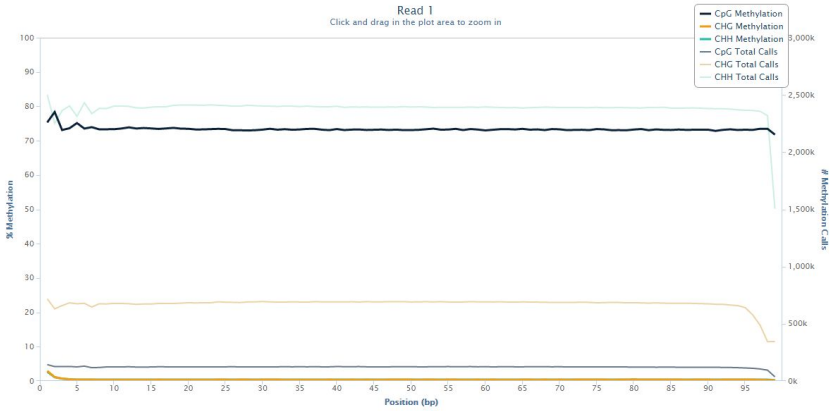
Methylation calling QC

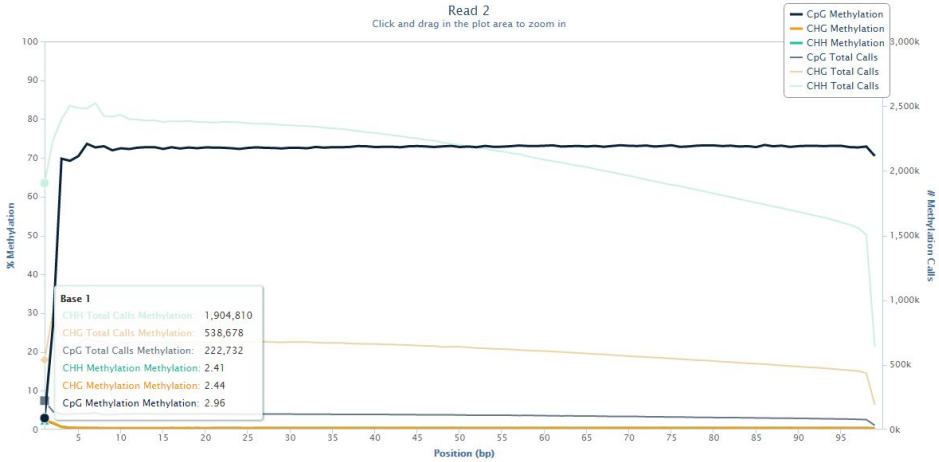
- C to T conversion rate should be >= 98% (ENCODE Pipeline)
- Check M-Bias Plots (generated by Bismark)

Measure BS-Seq conversion rate using:

- Consider Non-CpG sites as unmethylated
  - N/A for stem cells, oocytes, brain; plant or insect DNA
- Consider chrMT as unmethylated
  - actually seems to be methylated (Vitolacobazzi et al., 2013)
- Under conversion:
  - lambda genome: accurate
  - PhiX (unmethylated spike-in)
- Over conversion:
  - In vitro methylated T7 dsDNA (by CpG mc transferase sssI)
- ERRBS: using 'C' provided by 3'-end repairing (rrbs guide)

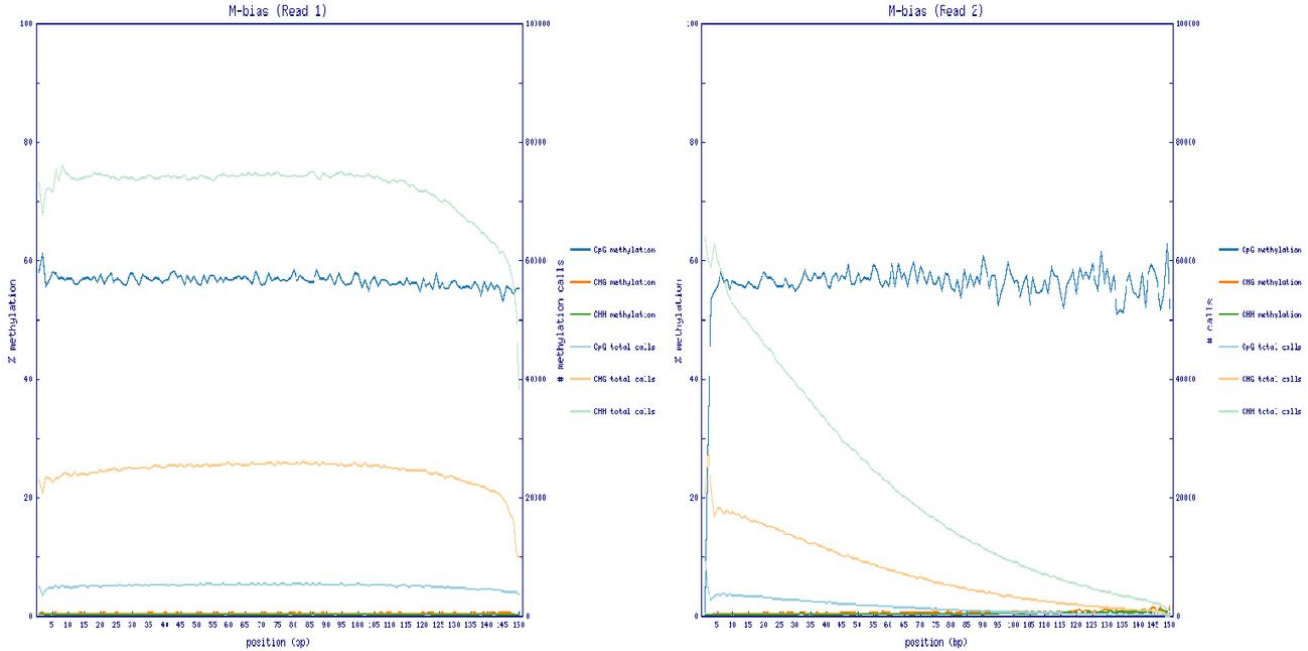
M-Bias Plots





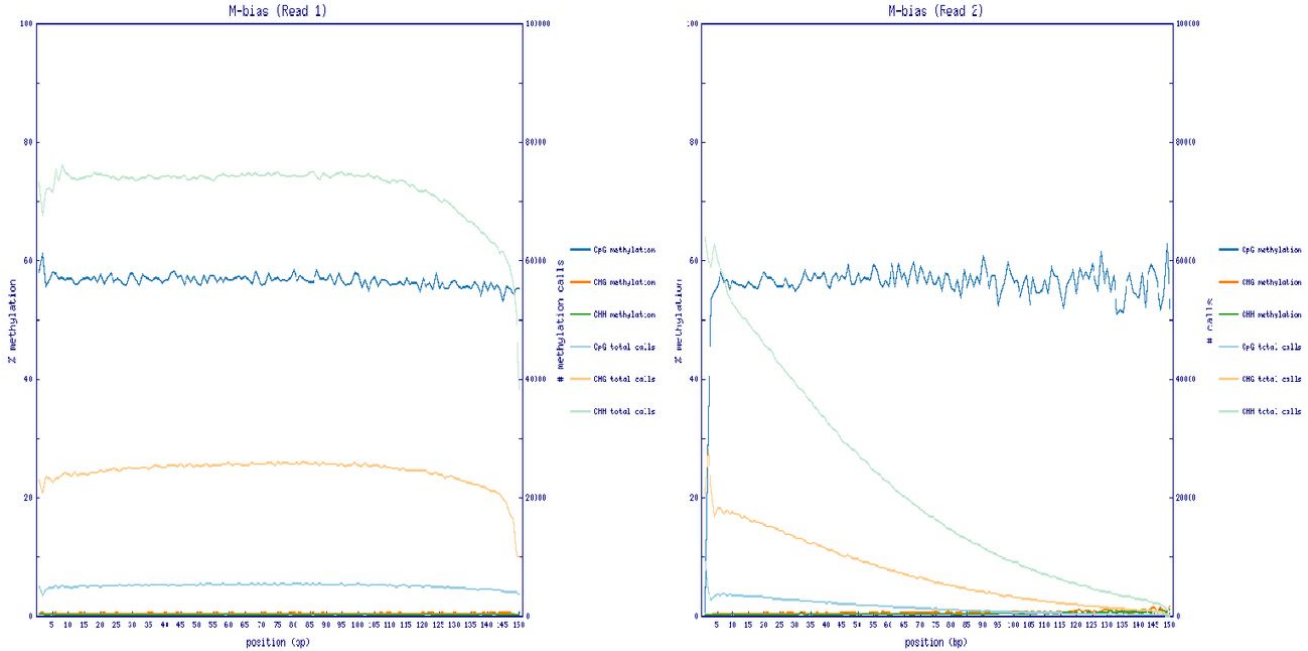
Ends Repairing bias, see https://sequencing.qcfail.com/articles/library-end-repair-reaction-introduces-methylation-biases-in-paired-end-pe-bisulfite-seq-applications/

M-Bias Plots



1) Bias at R1 3’ End?
2) Bias at R2 3’ End?

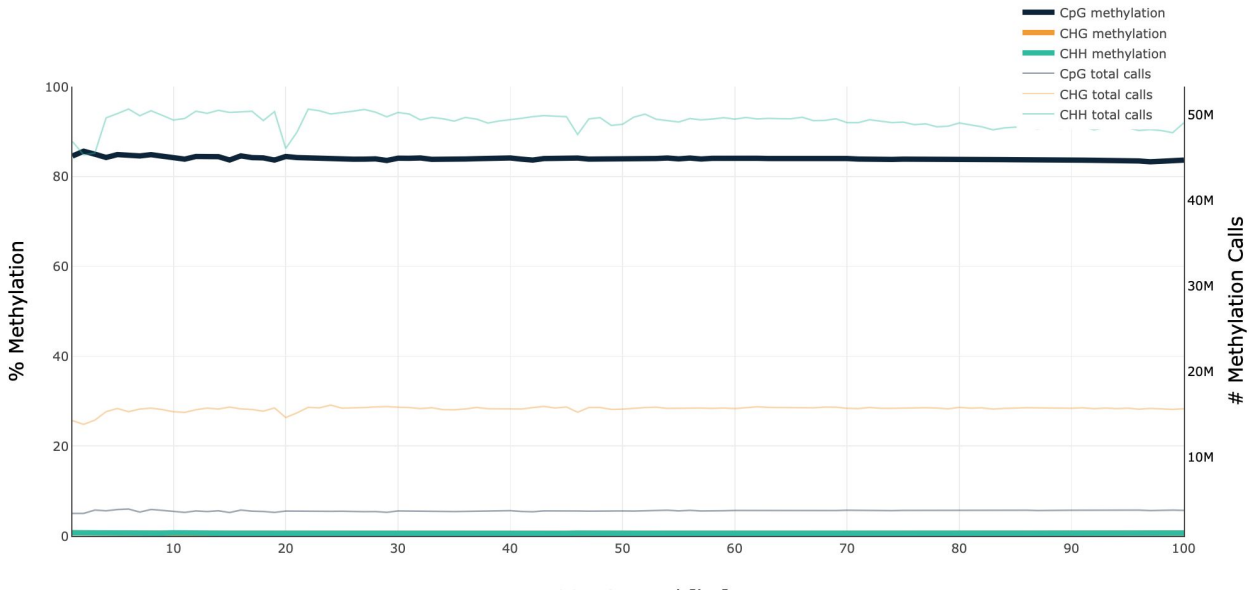
M-Bias Plots



1) Bias at R1 3’ End?
   R1 3’ end = R2 5’ end for short fragments
2) Signal goes down at R2?
   R2 & R1 intersection counted once

M-Bias Plots

In case of methylation bias at 3’ / 5’ ends - rerun mc calling with trimming options



Step 6: Batch Effect correction

Tools used in papers:

- ComBat
- In-house scripts

Sources of batch effect:

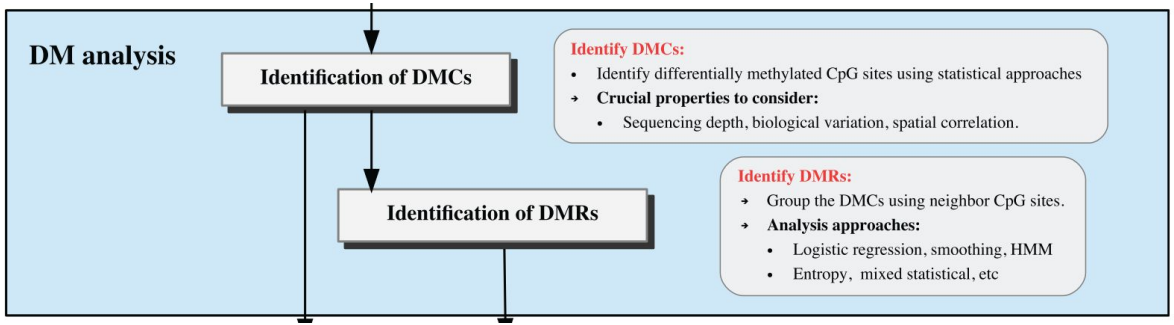
- Flowcell;
- Lane;
- Researcher

Step 7: DMC/DMR calling

DMC - differentially methylated cytosine
DMR - differentially methylated region

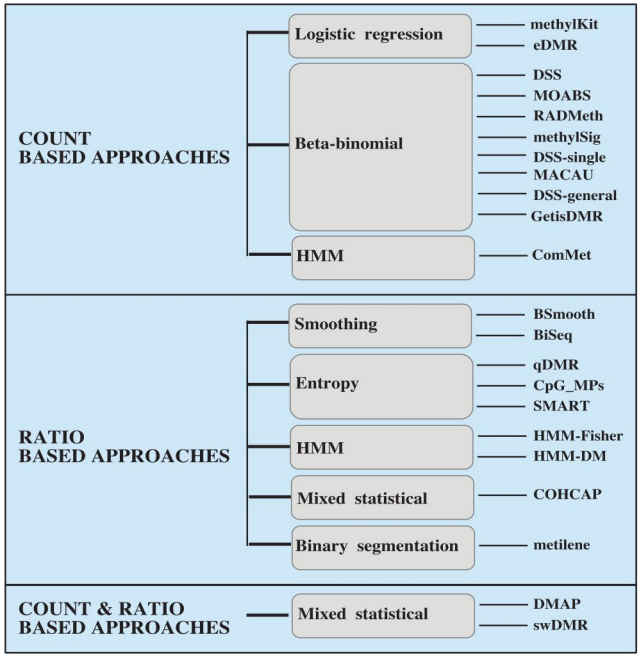
Popular tools:

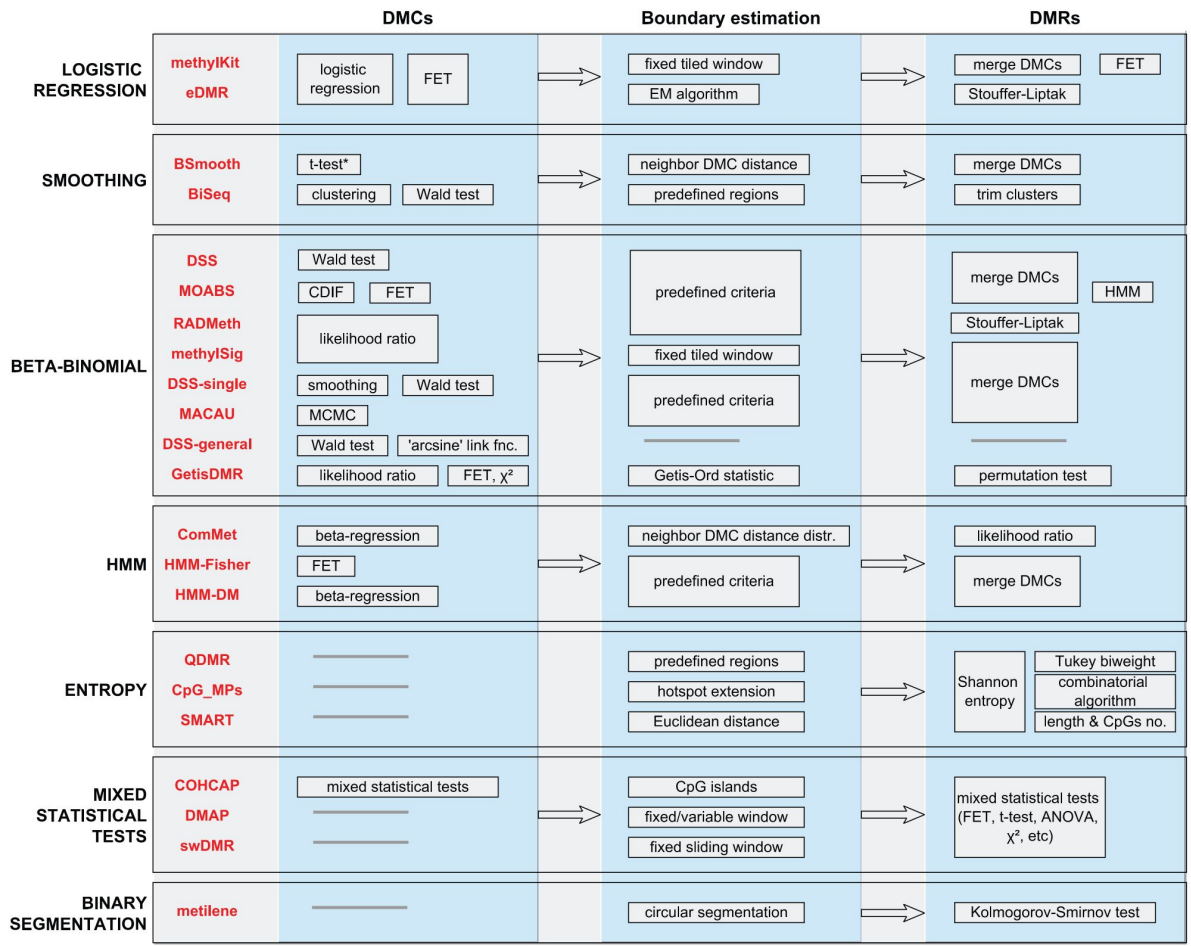
- MethPipe
- Metilene
- BS-SMooth (bs-seq)
- In-house scripts



(Shafi et al. 2018)

DMC Tools

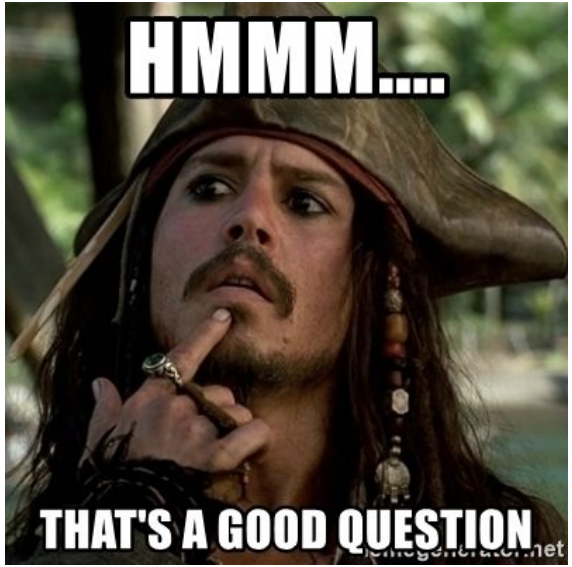




(Shafi et al. 2018)

Pros & Cons

What is the best?



LOGISTIC REGRESSION BASED APPROACHES	PROS: (i) Consider additional covariates; (ii) Provide error correction for multiple tests; (iii) Annotate the DMCs or DMRs. CONS: (i) Do not consider biological variation.
SMOOTHING BASED APPROACHES	PROS: (i) Consider biological variation; (ii) Consider spatial correlation; (iii) Can reduce required sequencing depth; (iv) Can estimate methylation level of missing CpGs. CONS: (i) Can not detect sharp methylation changes.
BETA-BINOMIAL BASED APPROACHES	PROS: (i) Consider biological variation; (ii) Some consider additional covariates; (iii) Some provide error correction for multiple tests; (iv) Mostly can detect *de novo* regions; (v) Can detect sharp methylation changes; (vi) Consider sequencing coverage. CONS: (i) Usually do not consider spatial correlation.
HMM BASED APPROACHES	PROS: (i) Can detect variable size DMRs; (ii) Consider spatial correlation; (iii) Usually consider biological variation. CONS: (i) Do not consider additional covariates; (ii) Do not provide error correction for multiple tests; (iii) Usually do not consider sequencing coverage.
ENTROPY BASED APPROACHES	PROS: (i) Directly identify DMRs; (ii) Can identify sample specific methylation patterns/DMRs; (iii) Consider biological variation. CONS: (i) Do not identify DMCs; (ii) Do not consider sequencing coverage.
MIXED STATISTICAL BASED APPROACHES	PROS: (i) Mostly can detect DMR boundaries; (ii) Provide error correction for multiple tests; (iii) Provide flexibility to chose different statistical approaches. CONS: (i) Usually do not consider biological variation; (ii) Do not consider spatial correlation; (iii) Do not consider additional covariates.
BINARY SEGMENTATION BASED APPROACHES	PROS: (i) Considers biological variation; (ii) Provides error correction for multiple tests; (iii) Can detect variable size DMRs. CONS: (i) Does not identify DMCs; (ii) Does not consider spatial correlation; (iii) Does not consider additional covariates.

(Shafi et al. 2018)

Step 8: Annotate Results

- GREAT
- Enrichr
- Methylation Clocks
  - 2013: Horvath 353 CpG, multi tissue
  - 2013: Hannum 71 CpG, blood
  - 2012: Garagani, 1 CpG in ELOVL2 gene, whole blood
  - 2018: PhenoAge (Levine), 513 CpG, whole blood

Takeaways

- Tools: Trim Galore, Bismark + Bowtie1/2, MethPipe, Metilene
- Be careful with bias
  - Duplicates
  - End-Repairing
  - Bisulfite treatment (over bs-converted / under bs-converted)
  - Bisulfite conversion rate is tissue dependent

Thank you