Analysis of DNA Methylation

Shamith Samarajiwa

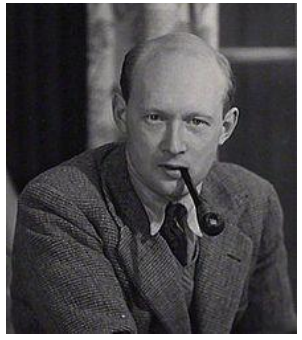
Group Leader (Computational Biology & Data Science)
MRC Cancer Unit,
University of Cambridge.

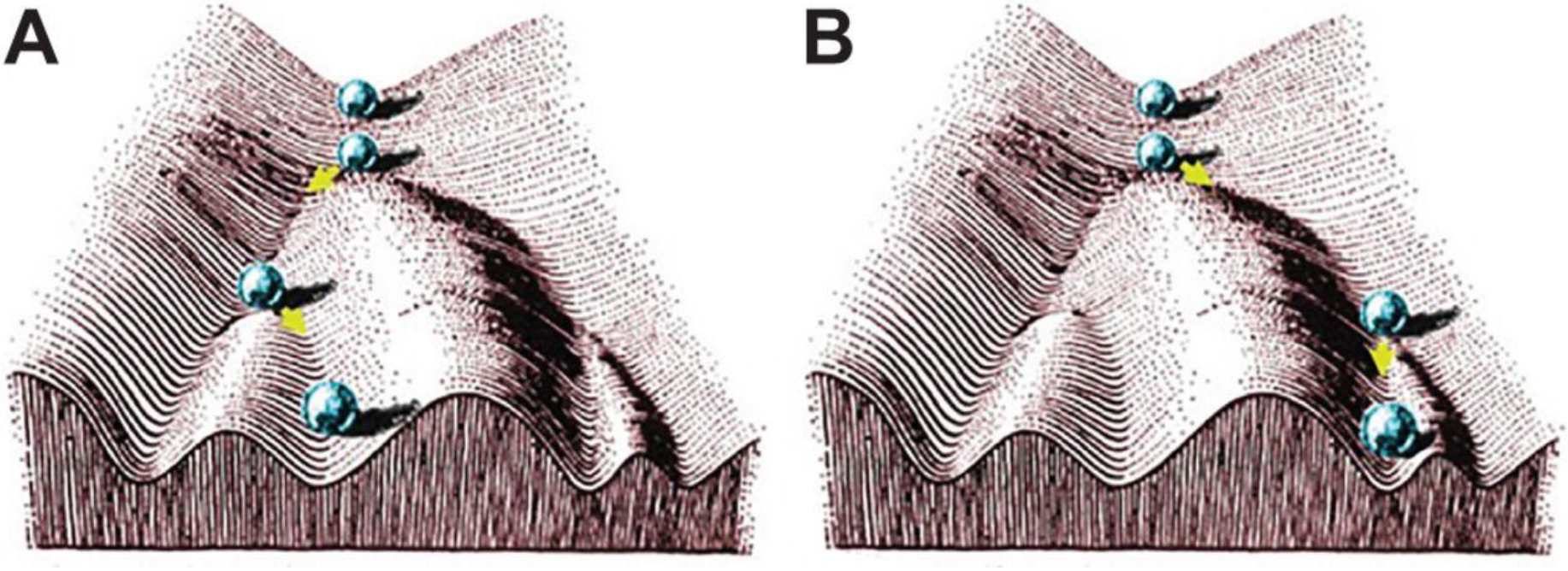
Computational Biology MPhil: Genomics II Lectures
March 10th 2021
Most slides are modified from Dr Rajbir Batra's Lectures in 2018/19

Overview

- Definitions of epigenetics
- Role of DNA methylation during development, somatic cells differentiation and in disease
- Cancer and DNA methylation
- Strength and weaknesses of different DNA methylation profiling technologies
- Bisulfite sequencing analysis workflow
- Applications of DNA methylation sequencing

Waddington's Diagram





Waddington's Canalised developmental landscape (modified): Conceptualization of development as a series of successive branching decisions, whereby genes act at critical points to govern tissue development along one path or another

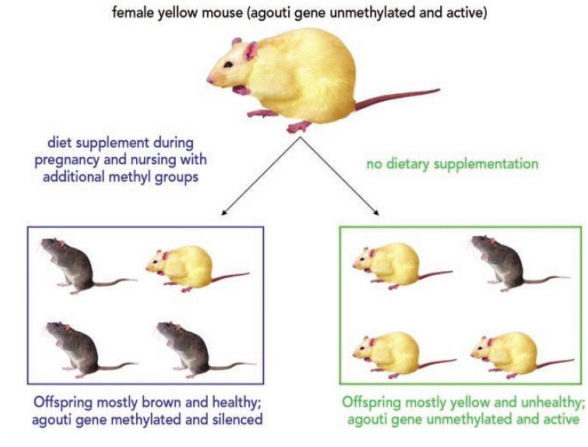
Definitions

Epigenetics: "*The inheritance of variation (-genetics) above and beyond (epi-) changes in the DNA sequence*"

- Environmental or behavioural changes that impart heritable changes (mitotic or meiotic) in gene activity and expression (phenotype) without changing the underlying DNA sequence.
  - Epigenetics changes are reversible.
  - Single genome, multiple epigenomes.
  - DNA methylation states can be transmitted from parent to daughter cells (memory beyond cellular perturbations)

Agouti viable yellow mice

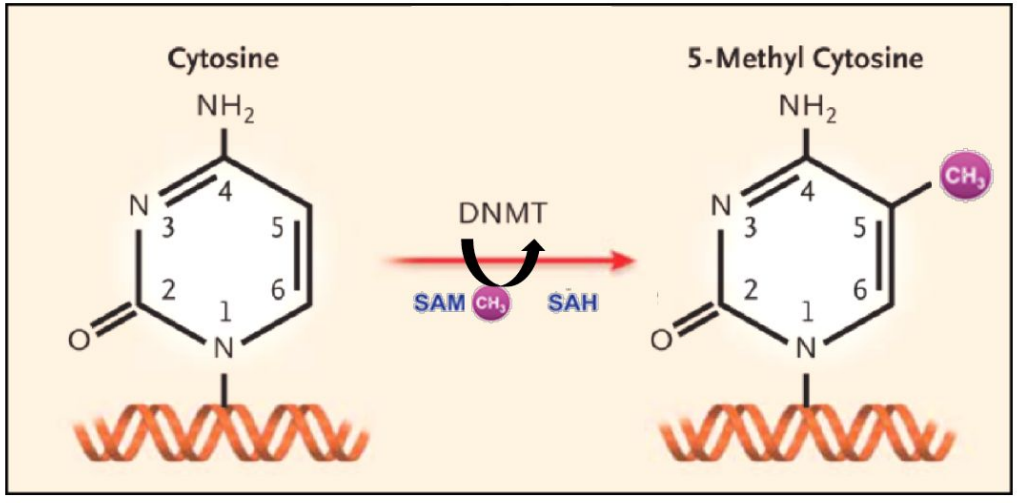
Waterland and Jirtle 2003



Epigenetic Agents

- DNA methylation and hydroxymethylation
- Histone modifications: post-translational modifications of histone tails
- Nucleosome occupancy and chromatin accessibility
- Chromatin architectural changes:
  - Chromatin interactions – e.g. distal enhancer-promoter interactions, TAD rearrangements
  - Chromatin domains - e.g. long range epigenetic silencing (LRES), A/B compartment dynamics
  - Chromatin remodelling agents – e.g. polycomb group of proteins (PcG)
- Noncoding RNAs

DNA Methylation



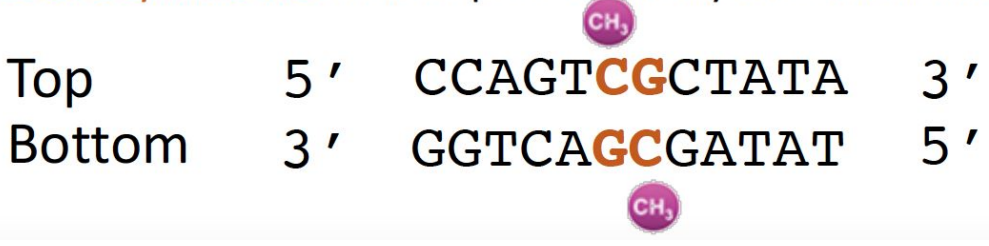
Herman & Baylin, 2003.

Can DNA methylation occur on all Cs?

sequence	context	mammals	plants
CCAGTCGCTATA	CpG	YES	YES
CCAGTCGCTATA	CHG	no*	YES
CCAGTCGCTATA	CHH	no*	YES

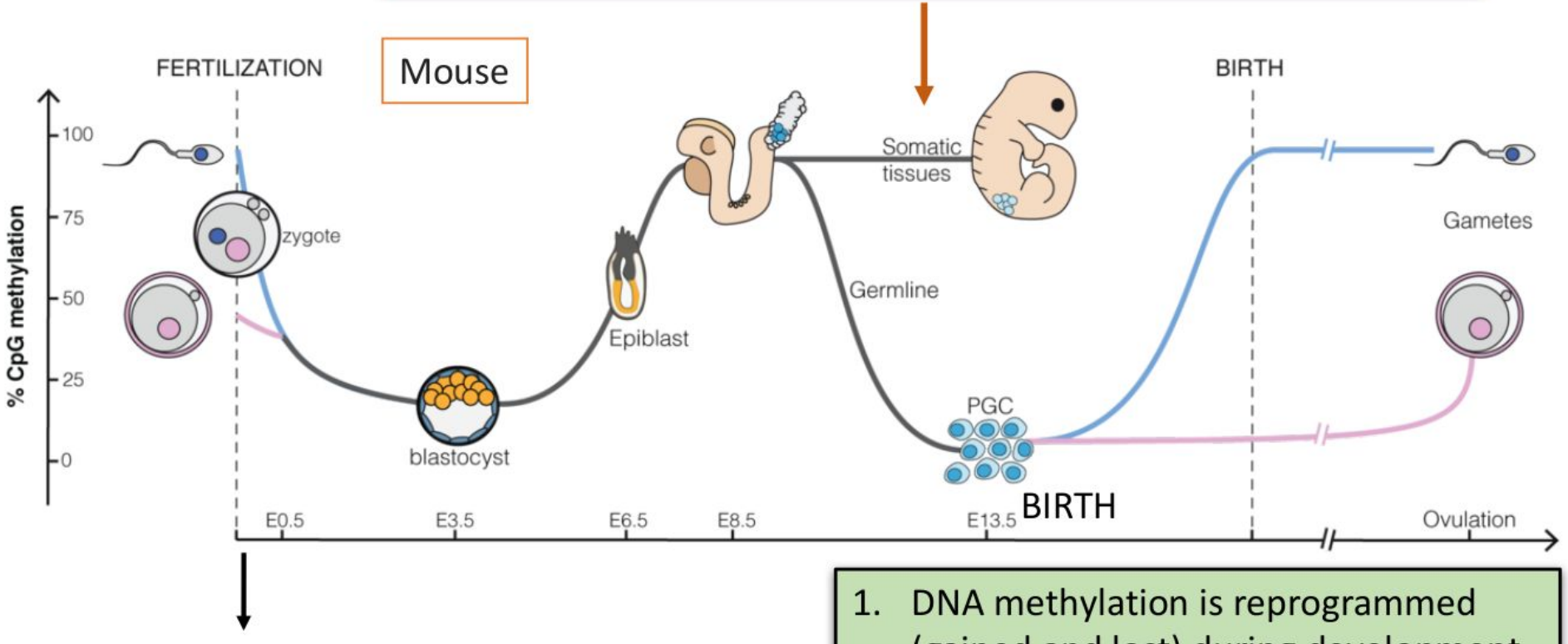
H being anything but G
* Occurs rarely in certain cell types

In humans, DNA methylation occurs almost exclusively at CpG dinucleotides
70%-80% of CpG sites are methylated in the human genome
CpG dinucleotides are symmetric with respect to methylation in humans



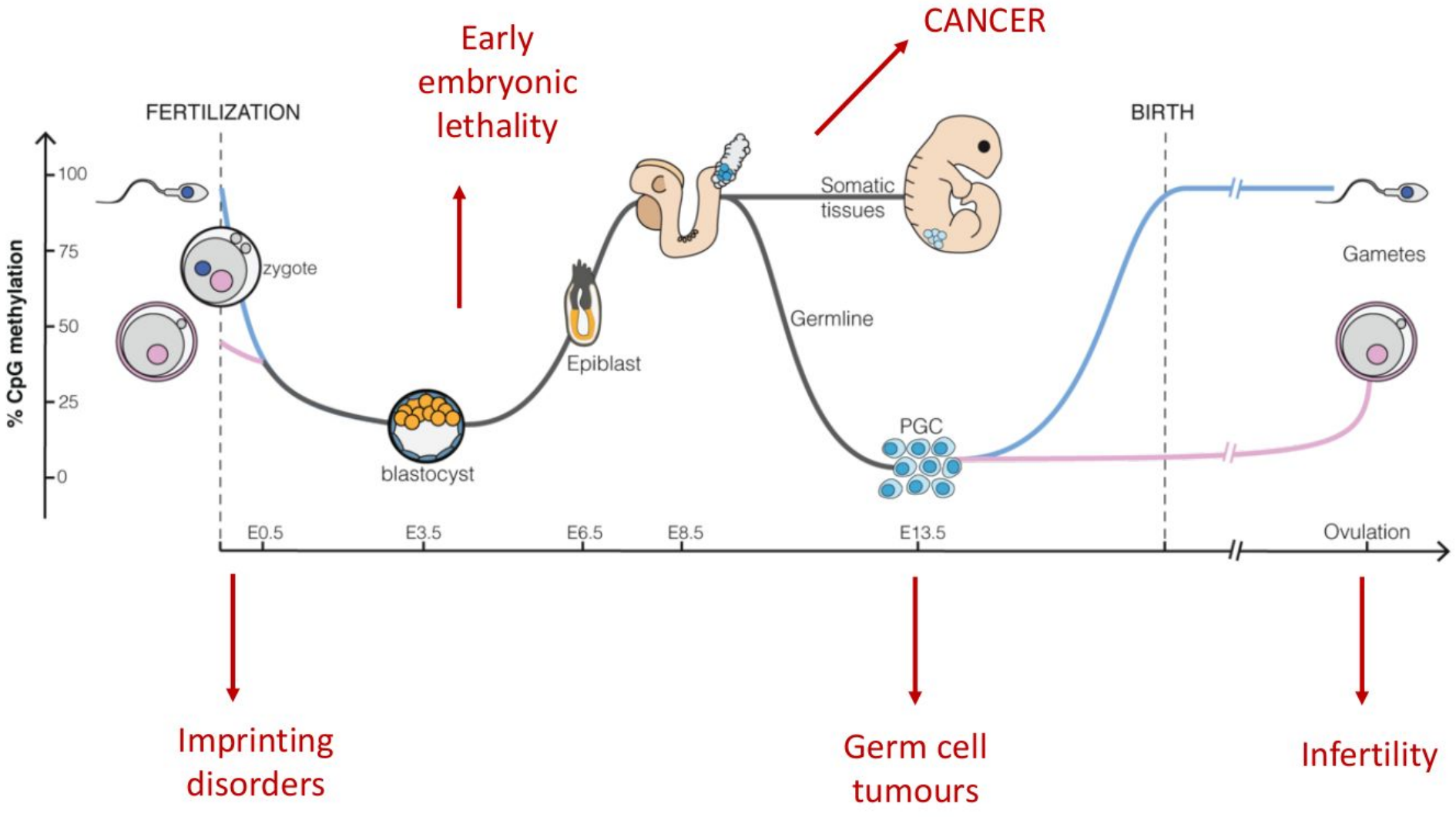
DNA methylation dynamic during cell differentiation and embryonic development

Cell-type specific DNA methylation is maintained in somatic cells

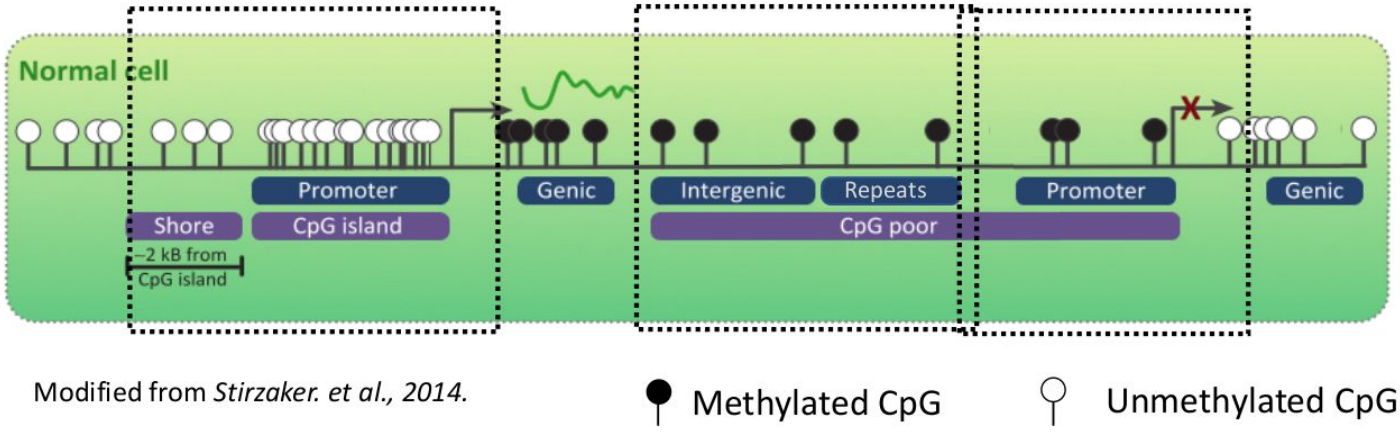


Imprinting
Some genes exhibit distinct epigenetic marks in maternal and paternal alleles resulting in allele-specific gene expression. Critical for embryo viability.

1. DNA methylation is reprogrammed (gained and lost) during development.
2. DNA methylation is mitotically heritable in somatic cells.


CANCER
Early embryonic lethality
FERTILIZATION
BIRTH
% CpG methylation
100
75
50
25
0
zygote
blastocyst
Epiblast
Somatic tissues
Germline
PGC
Gametes
E0.5
E3.5
E6.5
E8.5
E13.5
Ovulation
Imprinting disorders
Germ cell tumours
Infertility

DNA methylation in normal tissue



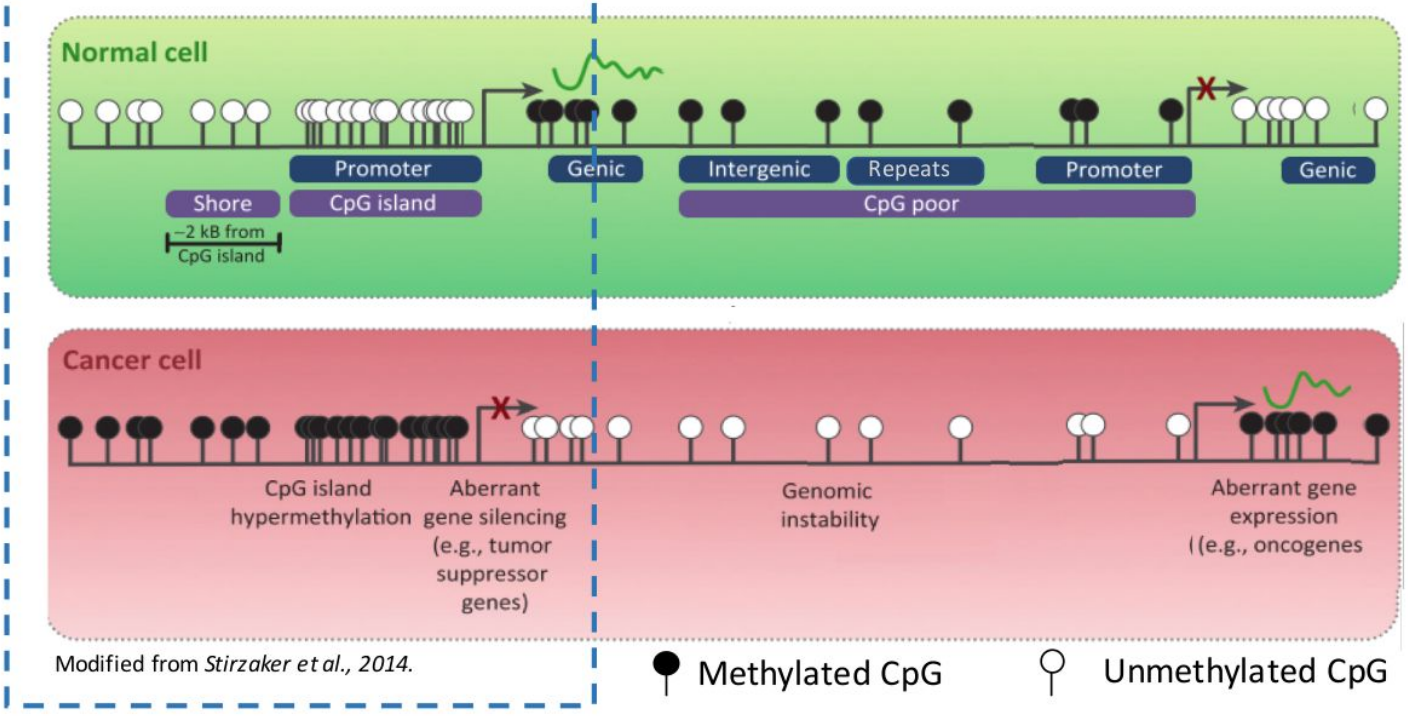
- ~28 million CpG sites in human genome. Majority not evenly distributed.
- CpG islands (CGIs) are clustered CpG sites at gene promoters. These are usually not methylated and correlate with active gene expression during differentiation.
- Methylation of CpG island promoters is associated with gene repression.
- Extensive exonic or genic methylation associated with active gene expression.
- CGI shores (2kb from CGIs) exhibit tissue and cancer specific differential methylation associated with gene expression.
- Remainder of CpG sites (CpG poor promoters and enhancers etc) in the genome are usually methylated.

DNA methylation in normal tissue

- Normal development and differentiation
- Tissue specific gene expression
- **X chromosome inactivation:** In female mammalian cells one of their two X chromosomes is transcriptionally silenced in a complex and highly coordinated manner (Lyon, 1961). The inactivated X chromosome then condenses into a compact structure called a Barr body, and it is stably maintained in a silent state (Boumil & Lee, 2001).
- **Genomic imprinting:** Epigenetic marks that are established in the parental germline and maintained in the somatic cells of the progeny. The genes in imprinted areas of an organism's genome are expressed depending on the parent of origin. As a result, the inheritance of both the maternal and paternal genes is required for normal development to proceed (McGrath & Solter, 1984; Surani *et al.*, 1984).
- Transgenerational Inheritance

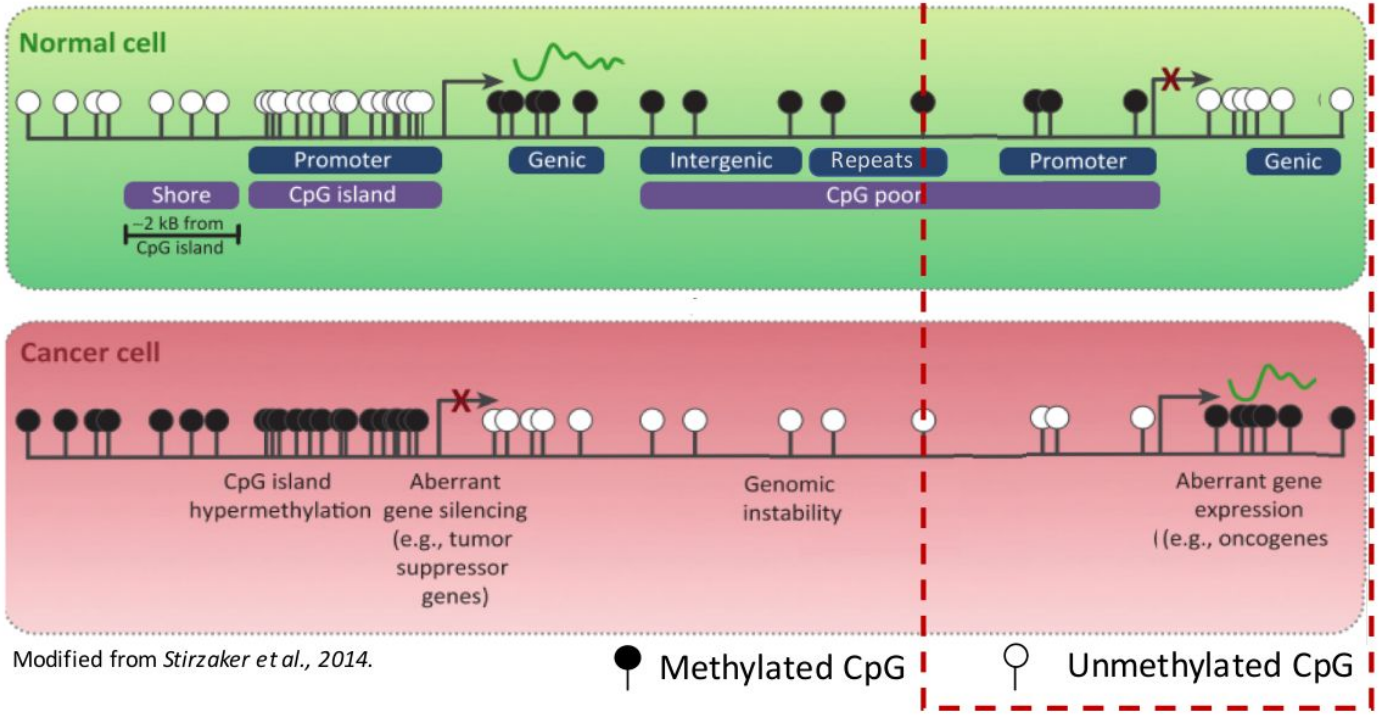
Calico cats

Cancer: Tumour Suppressor Silencing



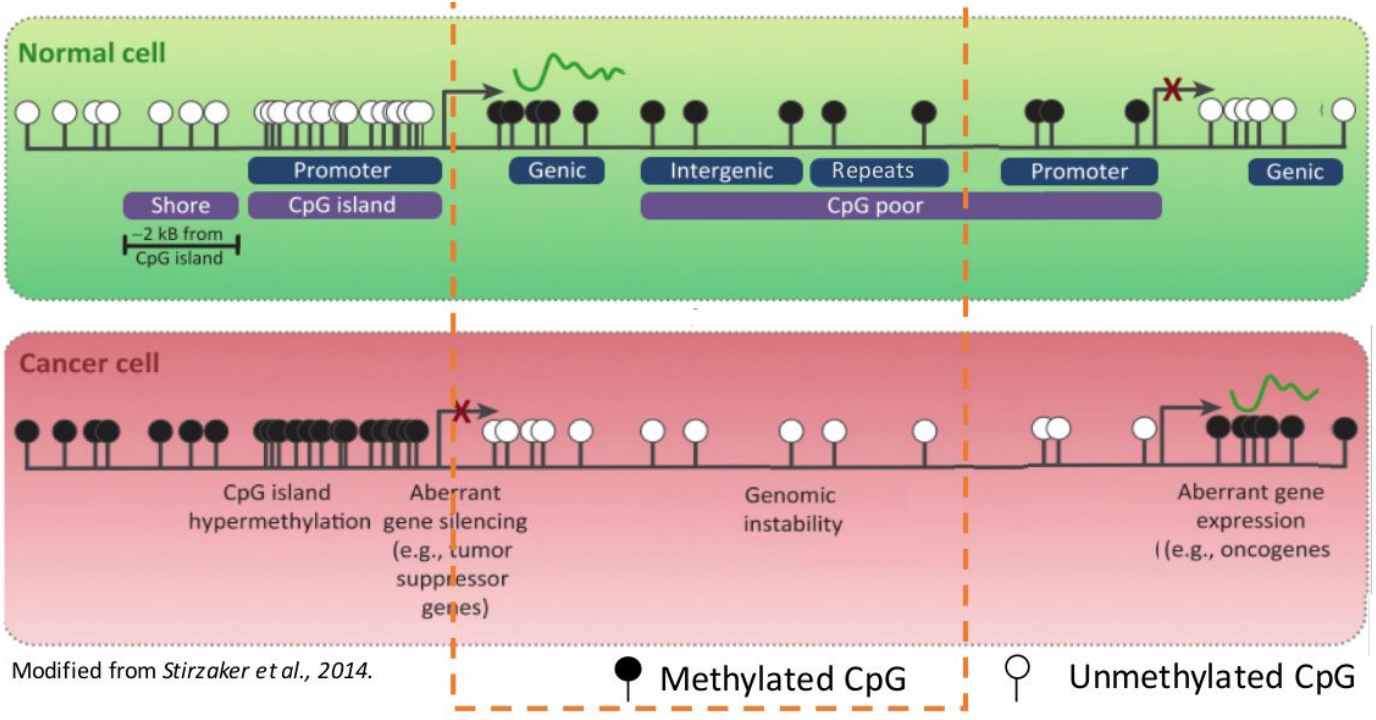
- Cancer genomes are Hypomethylated (less 5-meC than normal counterparts). However, promoters of tumour suppressor genes in these cancer cells can show increased DNA methylation at CGI or shores (Feinberg et al., 1983; Gama-Sosa et al., 1983) and loss of gene expression (recruitment of transcriptional repressors and blocking TF binding)
- More frequent than genetic mutations

Cancer: Oncogene Activation



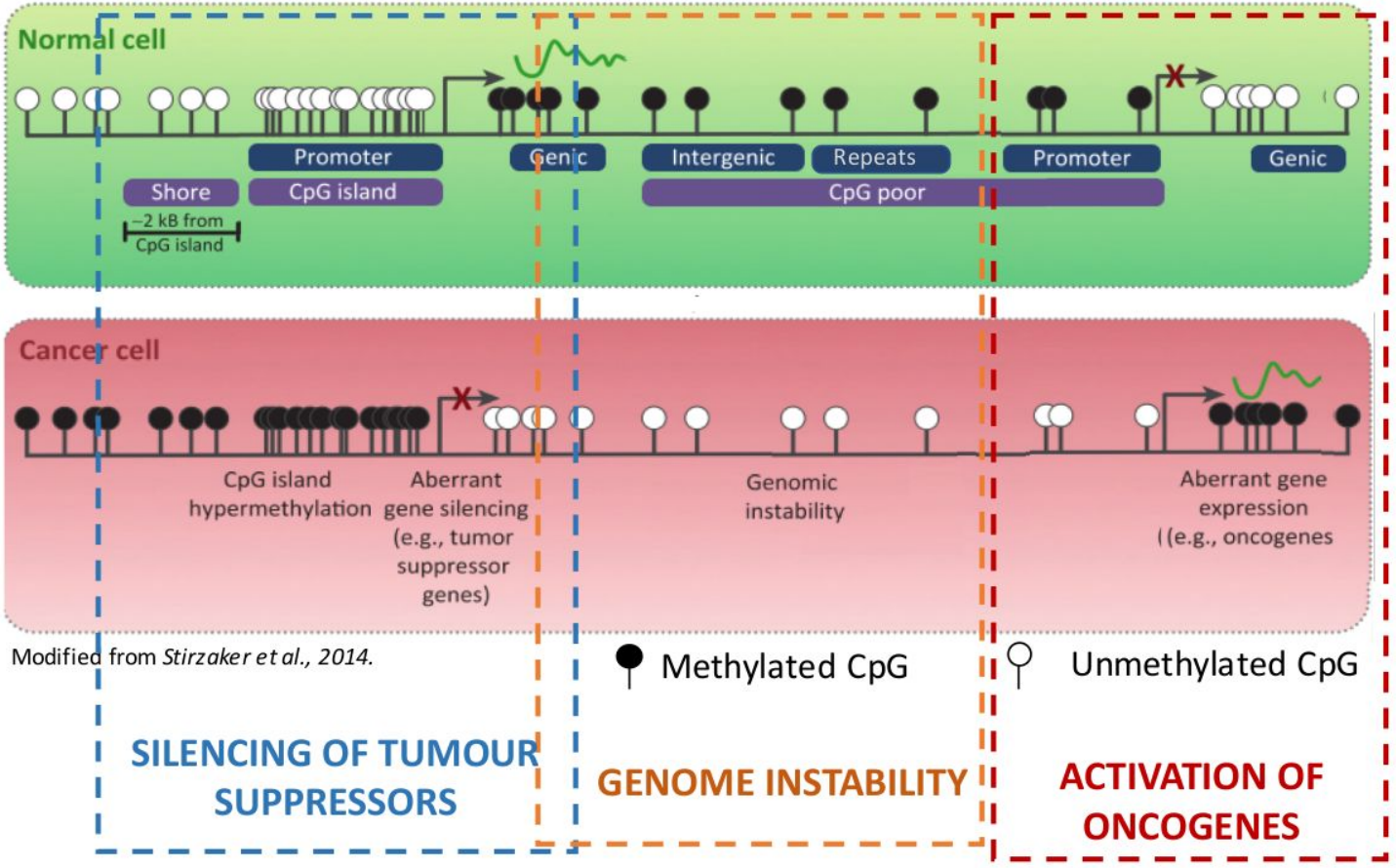
- Normal tissues: CpG poor promoters are methylated leading to gene silencing.
- Cancer: Hypomethylation of these regions is associated with the expression of oncogenes

Cancer: Genomic Instability



- Normal tissues: Repetitive and intergenic elements are methylated.
- Cancer: Hypomethylation of repetitive elements leads to their activation and transposition over the genome.
- Cancer: Hypomethylation of intergenic regions may lead to activation of unwanted exons.

Methylation and Hallmarks of Cancer



DNA Methylation profiling platforms

Table 1 | **Main principles of DNA methylation analysis**

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 • MBDCap-Seq (aka MethylCap-Seq)
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC–seq • BSPP • WGBS • PBAT • Single cell (scBS) • ···

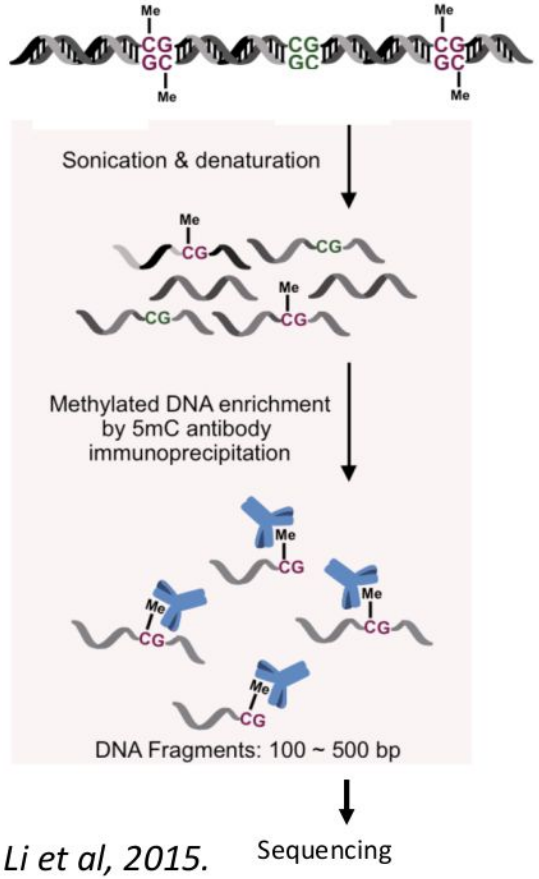
Popular Genome-wide methods

Modified from *Bock et al, 2010.*

GENOME WIDE DNA METHYLATION PROFILING	Microarray-based Most commonly used	Capture/ Affinity based	Bisulphite sequencing Gold standard
Discrimination - meth & unmeth	Bisulphite treatment	Antibody capture	Bisulphite treatment
NGS	Yes	No	Yes
Examples	Illumina HM27 Illumina HM450 Illumina EPIC	MeDip-seq MBDCap-Seq	WGBS RRBS PBAT (No PCR)
Compatible with low input DNA	Yes	No	Yes
Cost	£	£££	WGBS (£££££)/ RRBS (££)
Coverage	Illumina HM27 (0.1%) Illumina HM450 (~1.5%) Illumina EPIC (3%)	15-25%	WGBS (100% = 28 M CpG) RRBS (5-10%)
Bioinformatics analysis	Similar to gene expr Illumina microarrays (later in course)	Similar to Chip-seq (earlier in course)	Explained in today's lecture
Ease of bioinformatics analysis	Easy	Medium	Hard
Single nucleotide resolution	Yes	No	Yes
Quantification of DNA methylation	Absolute	Relative	Absolute
Cancer epiclonal analysis possible	No	No	Yes
Free from Copy number bias	Yes	No	Yes
Free from incomplete bisulphite sequencing bias	No	Yes	No (can be QC)

Capture based techniques (MeDIP-seq)

Enrichment with antibody + Sequencing



Li et al, 2015.

Data normalisation + analysis

Processing enrichment-based data	
BATMAN	Command-line tool for methylated DNA immunoprecipitation (MeDIP) data normalization
Bowtie	General-purpose aligner based on the Burrows–Wheeler transform
BWA	General-purpose aligner based on the Burrows–Wheeler transform
MEDIPS	User-friendly R package for MeDIP data normalization
MEDME	R package for MeDIP data normalization
MeDUSA	Command-line software pipeline for MeDIP read alignment, data normalization, quality control and DMR identification
MetMap	Command-line tool for normalization of DNA methylation data obtained using restriction enzymes that specifically cut unmethylated DNA
MeQA	Command-line software pipeline for MeDIP read alignment, data normalization and quality control
Repitools	R/Bioconductor package for quality control and visualization of enrichment-based DNA methylation data

Bock et al, 2010.

Analysis similar (but not equivalent) to Chip-seq analysis (earlier in the course)

Bisulfite Sequencing Platforms (WGBS & RRBS)

WGBS

Whole Genome Bisulphite sequencing

Fragmentation of DNA
Size selection of fragments
Bisulphite conversion
Library amplification
Sequencing

RRBS

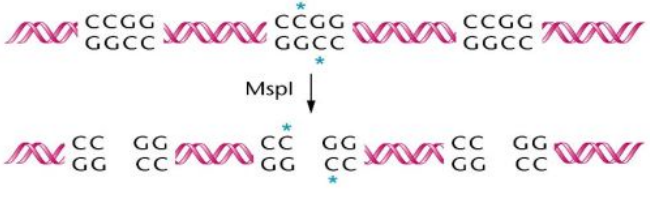
Reduced Representation Bisulphite sequencing

Restriction enzyme *(MspI)* digestion
Size selection of fragments
Bisulphite conversion
Library amplification
Sequencing

In bold – specific to bisulphite sequencing profiling

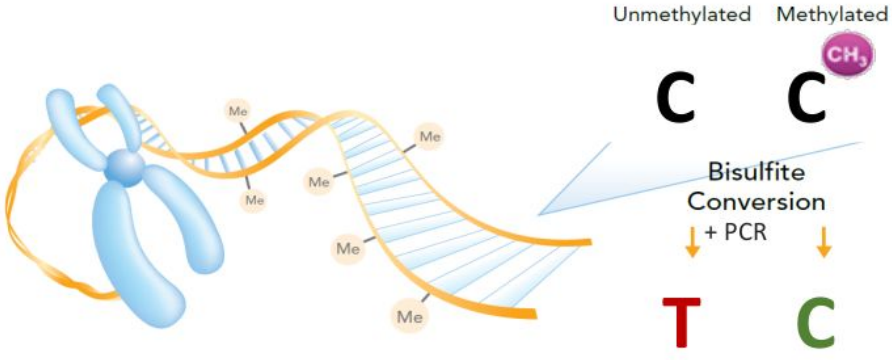
Restriction enzyme *(MspI)* digestion

Does not fragment DNA randomly
MspI targets **CCGG** motifs.
Enrich for regions of high CpG content



But what is Bisulphite conversion?

Bisulfite Conversion



Unmethylated Methylated

C CH_3C

Bisulfite Conversion + PCR

T C

Discriminates between methylated and unmethylated cytosines

T = unmethylated C

C = methylated C

CH_3 CH_3

CCAGTCGCTATAGCGCGATATCGTA

Selective bisulphite conversion of unmethylated C into T + Sequencing

TTAGTCGTTATAGTGCGATATTGTA

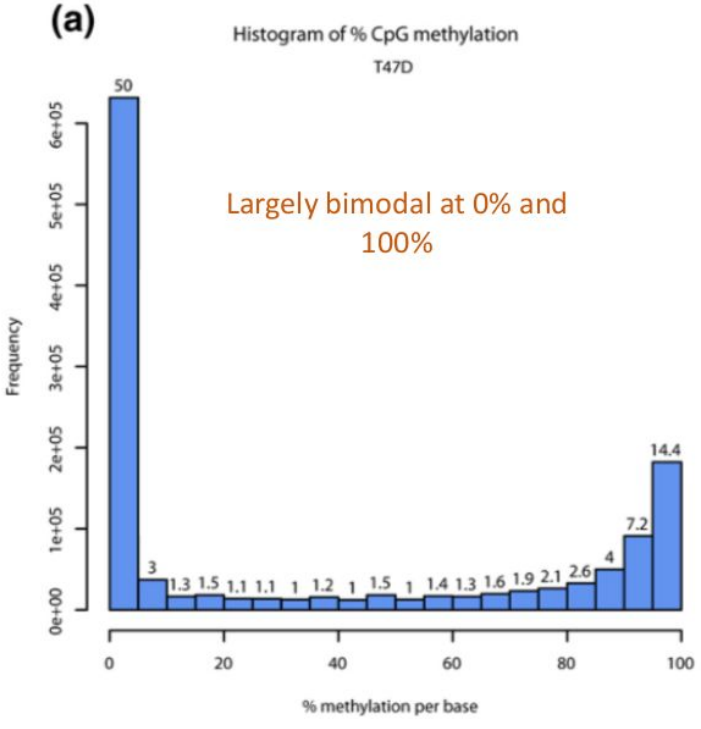
CCAGTCGCTATAGCGCGATATCGTA...

But aligning to the reference genome is a computational challenge

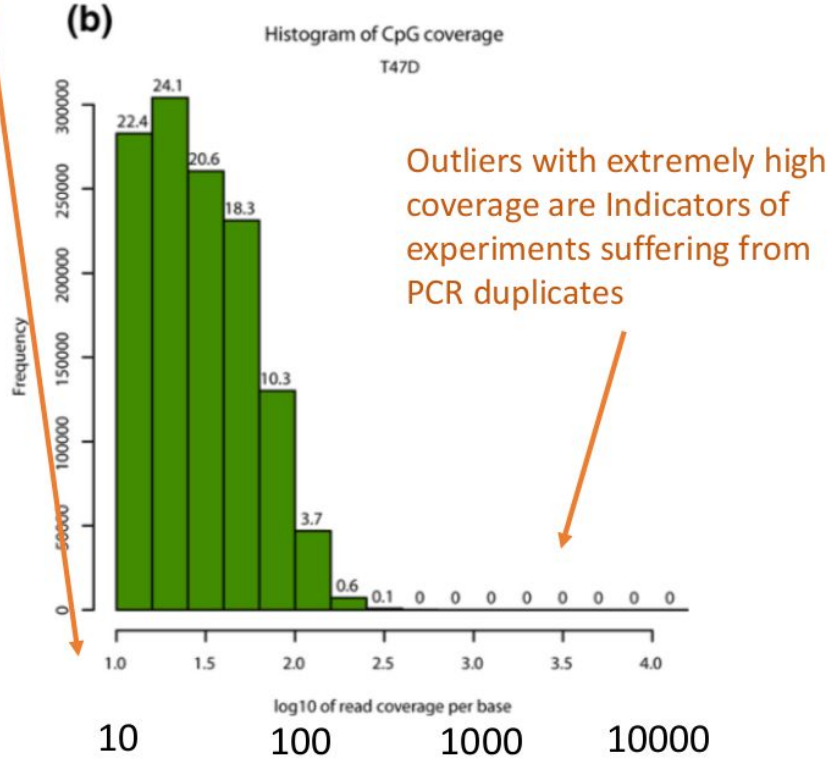
Comparison with reference genome

Sample level descriptive statistics

CpG sites with extremely low coverage (<5 or <10 reads) are insufficient



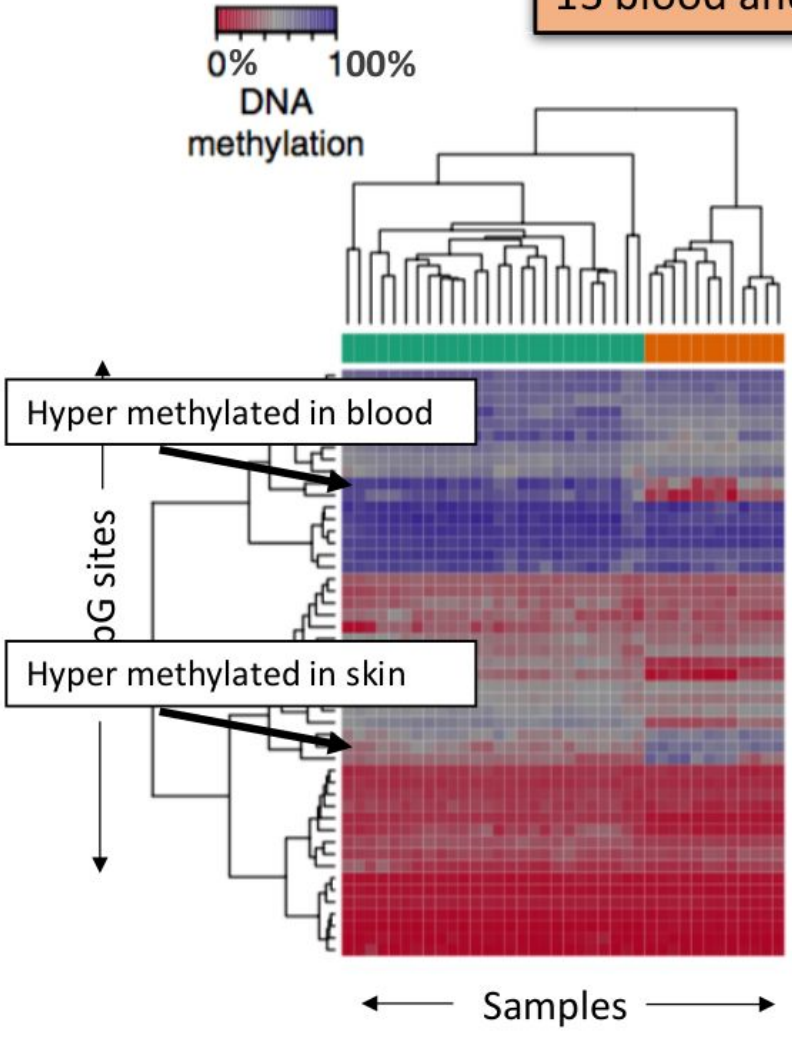
Distribution of methylation over all CpG sites



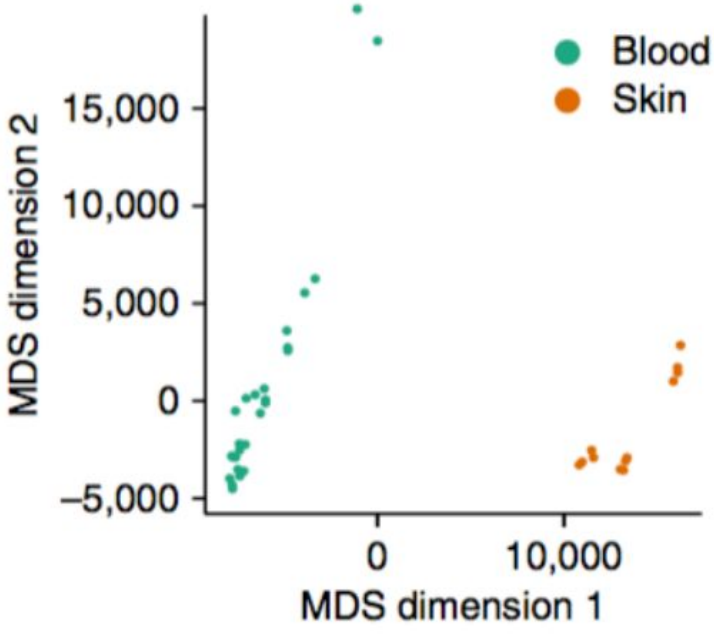
Distribution of coverage for all CpG sites

methylKit: *Akalin et al, 2012.*

DNA methylation profiling of
13 blood and 6 skin cell populations (with replicates)



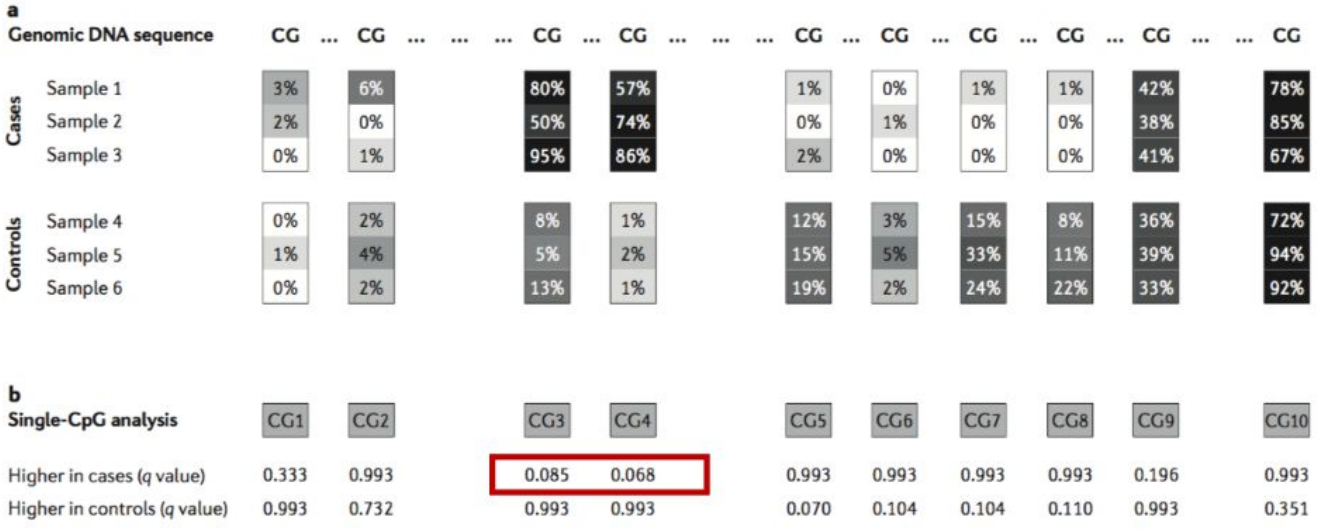
PCA or MDS. Dimension reduction methods



Tissue type specific methylation patterns

RnBeads: *Assenov et al, 2014.*

Differential Methylation Analysis



Modified from *Bock et al, 2012.*

Microarrays:

1. Linear regression (logit transform first)
2. Beta regression (0-100%)

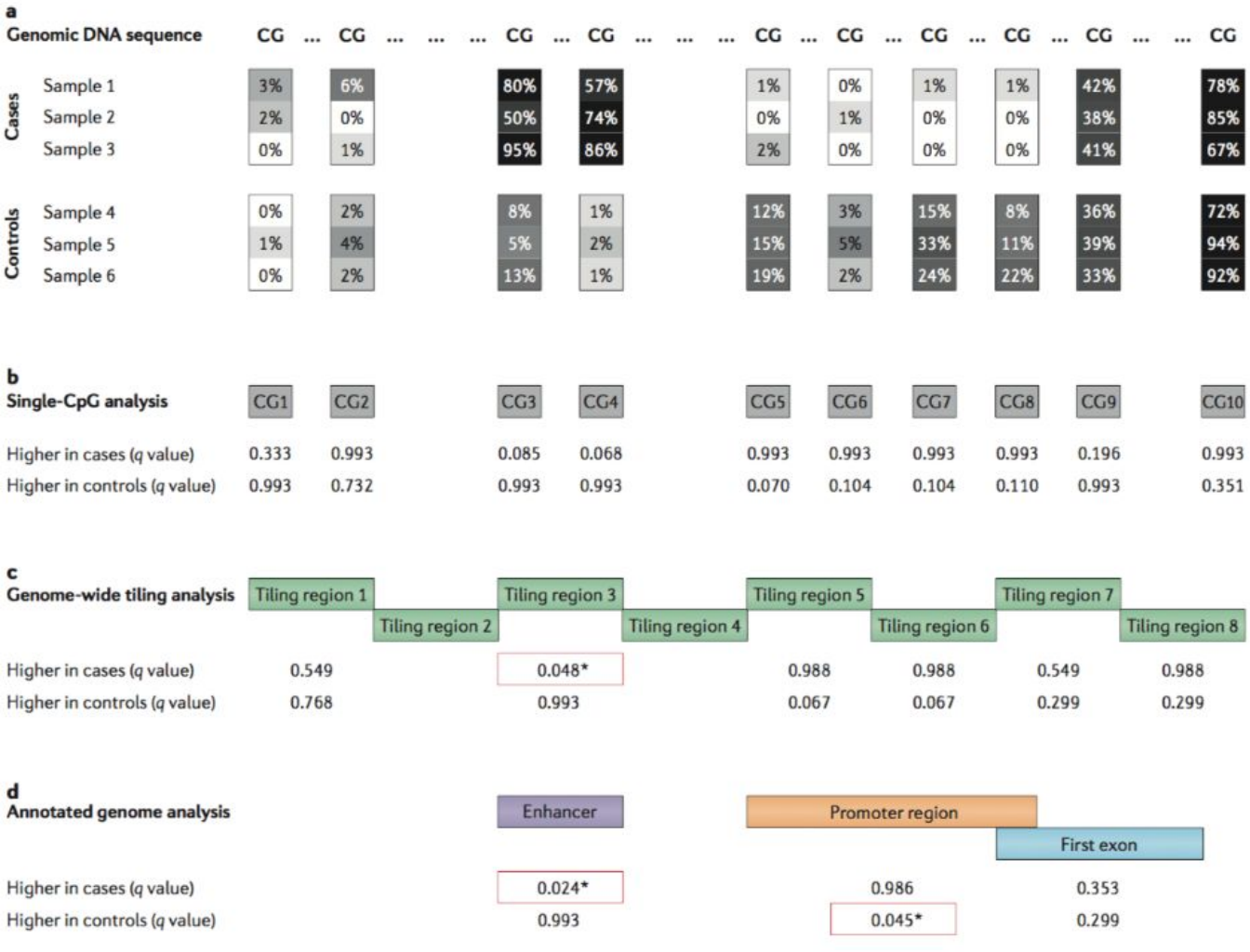
Bisulphite sequencing:
(need to take into account number of reads)

1. Logistic regression (Tool: methylKit)
2. Beta-binomial regression

SINGLE CPG ANALYSIS:
depending on platform you can profile 450K to 28M CpG sites
Accounting for multiple testing (Benjamini Hochberg).

Limitations

- Only strongest single CpG differences remain.
- For bisulphite sequencing, need high coverage (30+) to detect single CpG differences.



Use predefined regions for comparison

Statistically better

- Lower number of tests (e.g. 20000 gene promoters)
- Neighbouring CpGs with similar differences in methylation reinforce each other to improve power

Biologically relevant

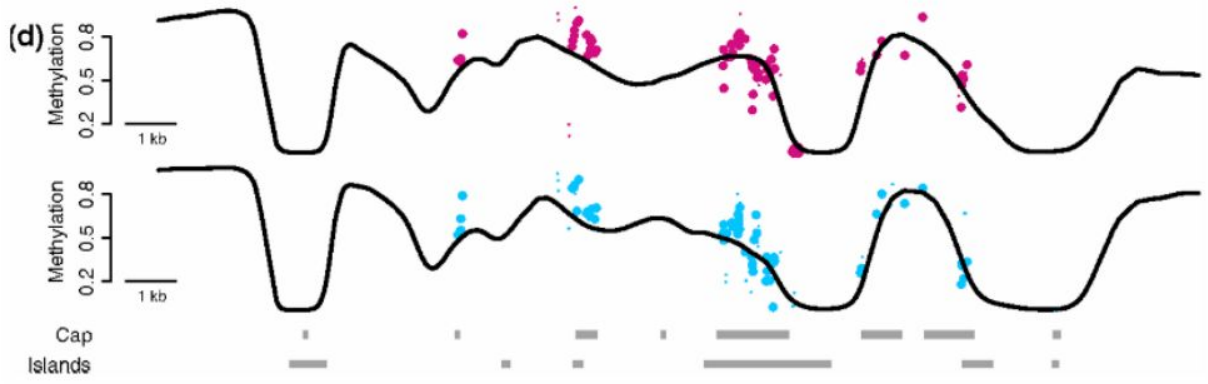
- functionally relevant findings are generally associated with genomic features

Modified from *Bock, 2012.*

Differentially Methylated Regions

BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions

Kasper D Hansen[1*†], Benjamin Langmead[1,2*†] and Rafael A Irizarry[1,2*]



GROUP 1

GROUP 2

For WGBS. Smoothing to determine differentially methylated regions.

This works because in WGBS

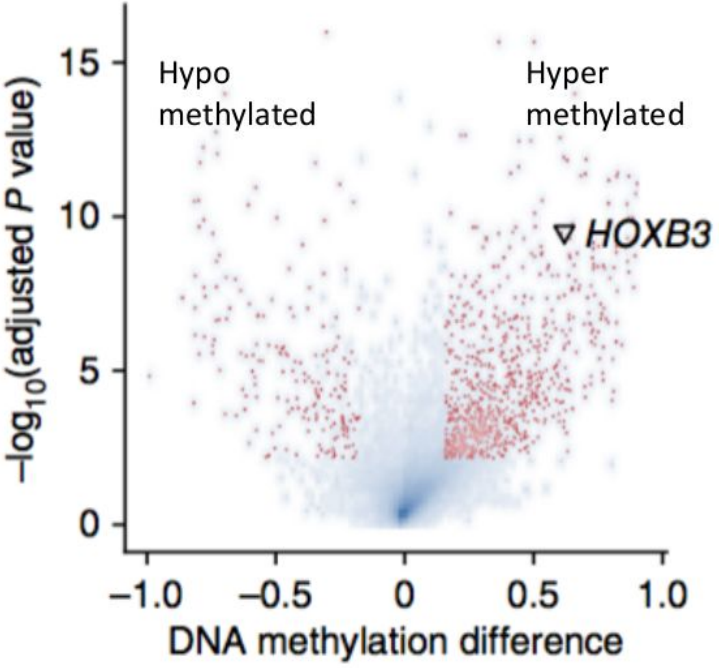
1. Methylation estimates are available for every CpG site in genome.
2. Neighbouring CpG sites are correlated with respect to methylation.

So, Neighbouring CpGs with similar differences in methylation reinforce each other to improve power.

Identifying and annotating DMRs

Two thresholds for identifying CPGs/ regions differential methylation analysis

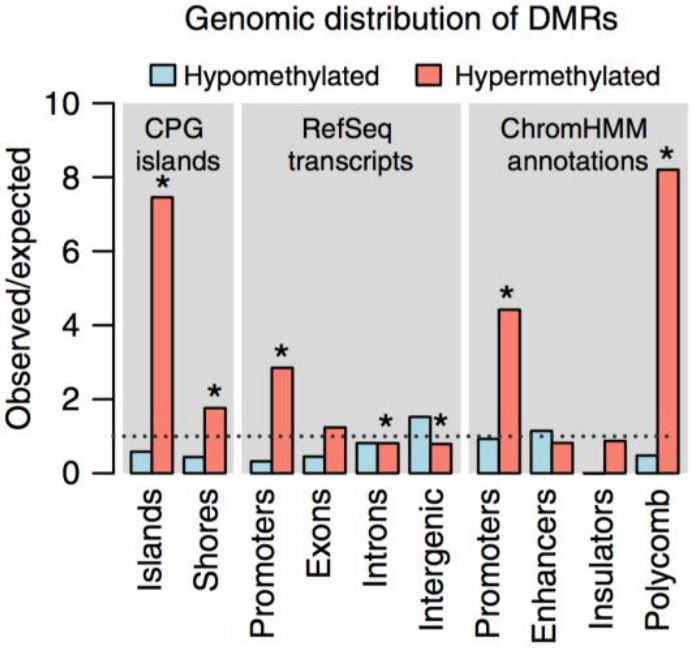
1. Minimum average methylation difference between two groups. Usually >=20%
2. Benjamini Hochberg (FDR) controlled *p-values*. Usually use alpha = 0.05.



RnBeads: *Assenov et al, 2014.*

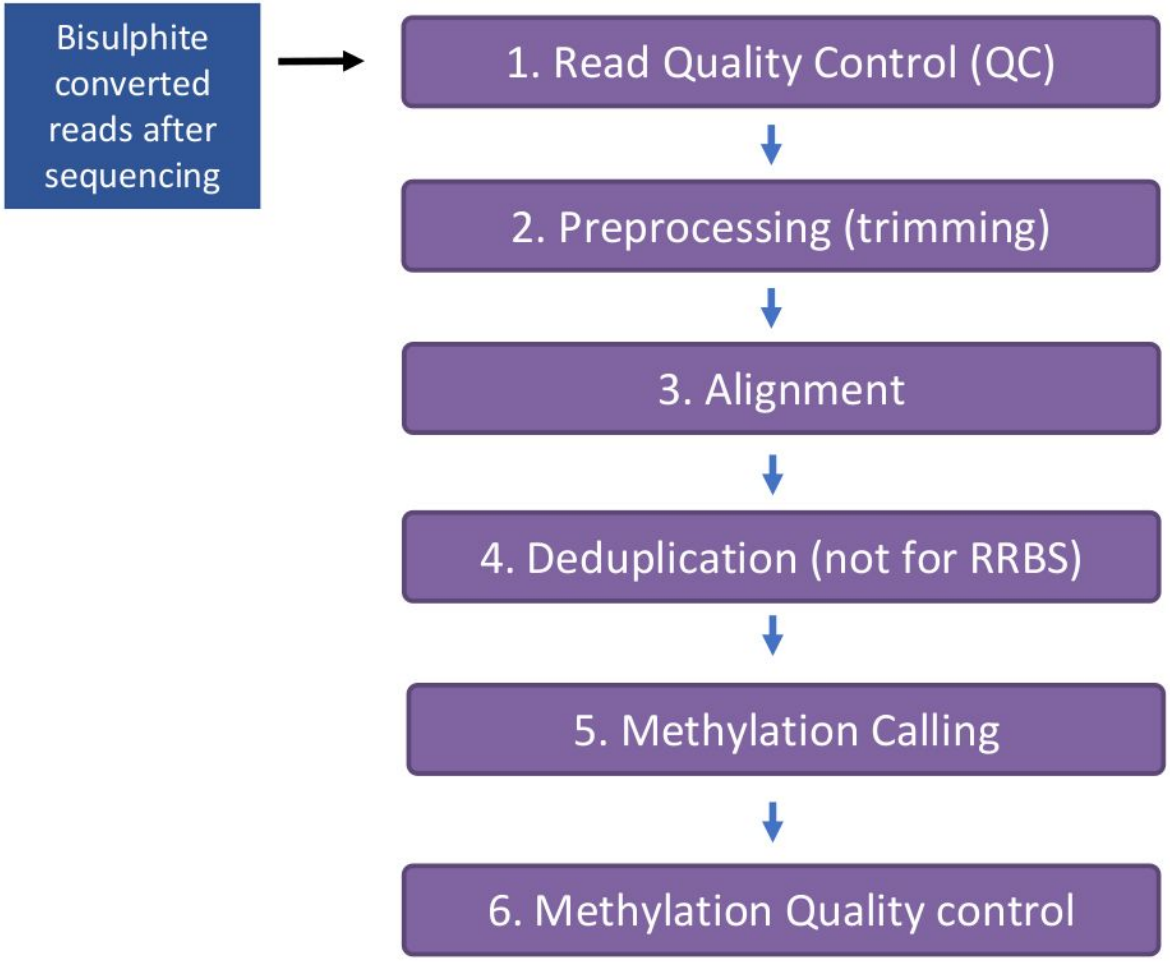
Annotate differentially methylated CpG sites or regions

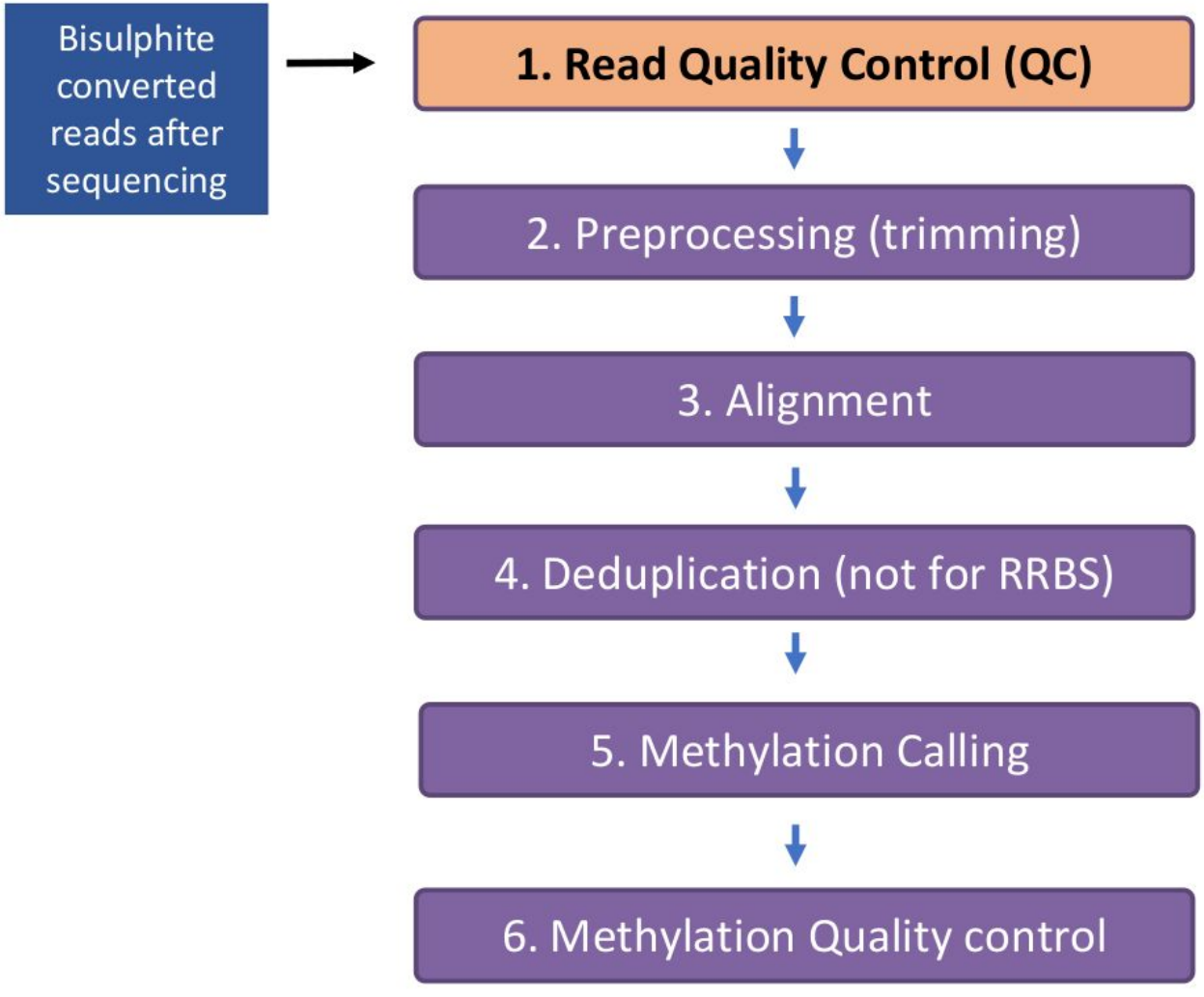
Cancer versus normal tissue



Stirzaker et al, 2015.

Workflow: Bisulfite Sequencing Data




Bisulphite converted reads after sequencing
1. Read Quality Control (QC)
2. Preprocessing (trimming)
3. Alignment
4. Deduplication (not for RRBS)
5. Methylation Calling
6. Methylation Quality control

QC: Base quality

TOOL

FastQC

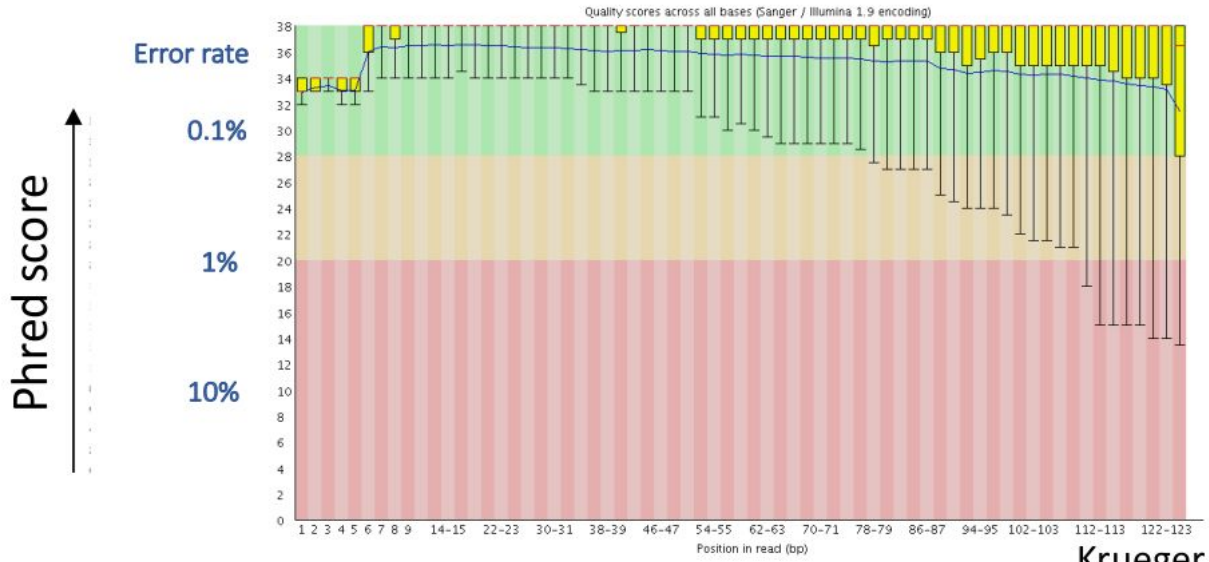
Babraham Institute

A quality control tool for **ALL** high throughput sequencing data

Basic Statistics

Measure	Value
Filename	SLX-10303.C6GM2ANXX.s_3.r_1
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	215996148
Sequences flagged as poor quality	0
Sequence length	125
%GC	32

Per base sequence quality



Krueger & Andrews, Babraham

QC: Base composition

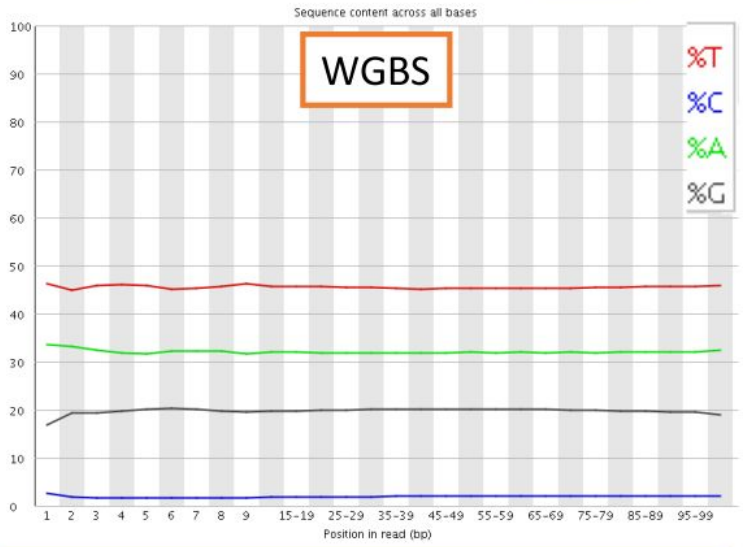
TOOL

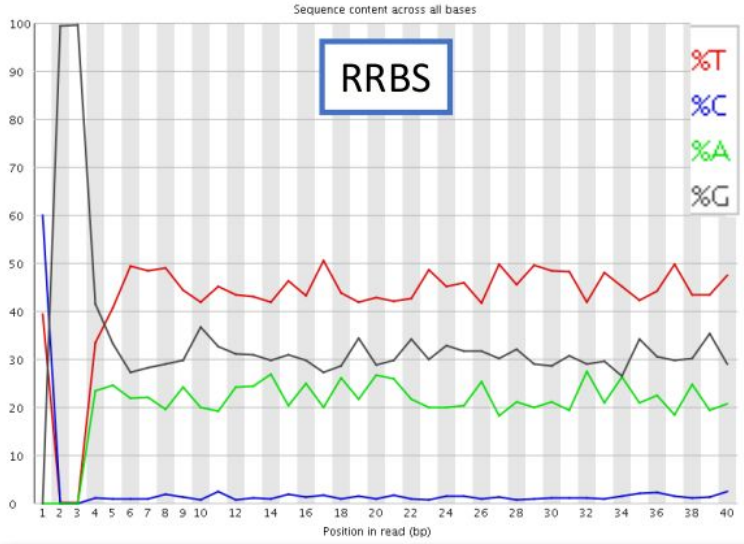
FastQC

Babraham Institute

Potential contaminants

Krueger & Andrews, Babraham





Why are there so few Cs in the reads?

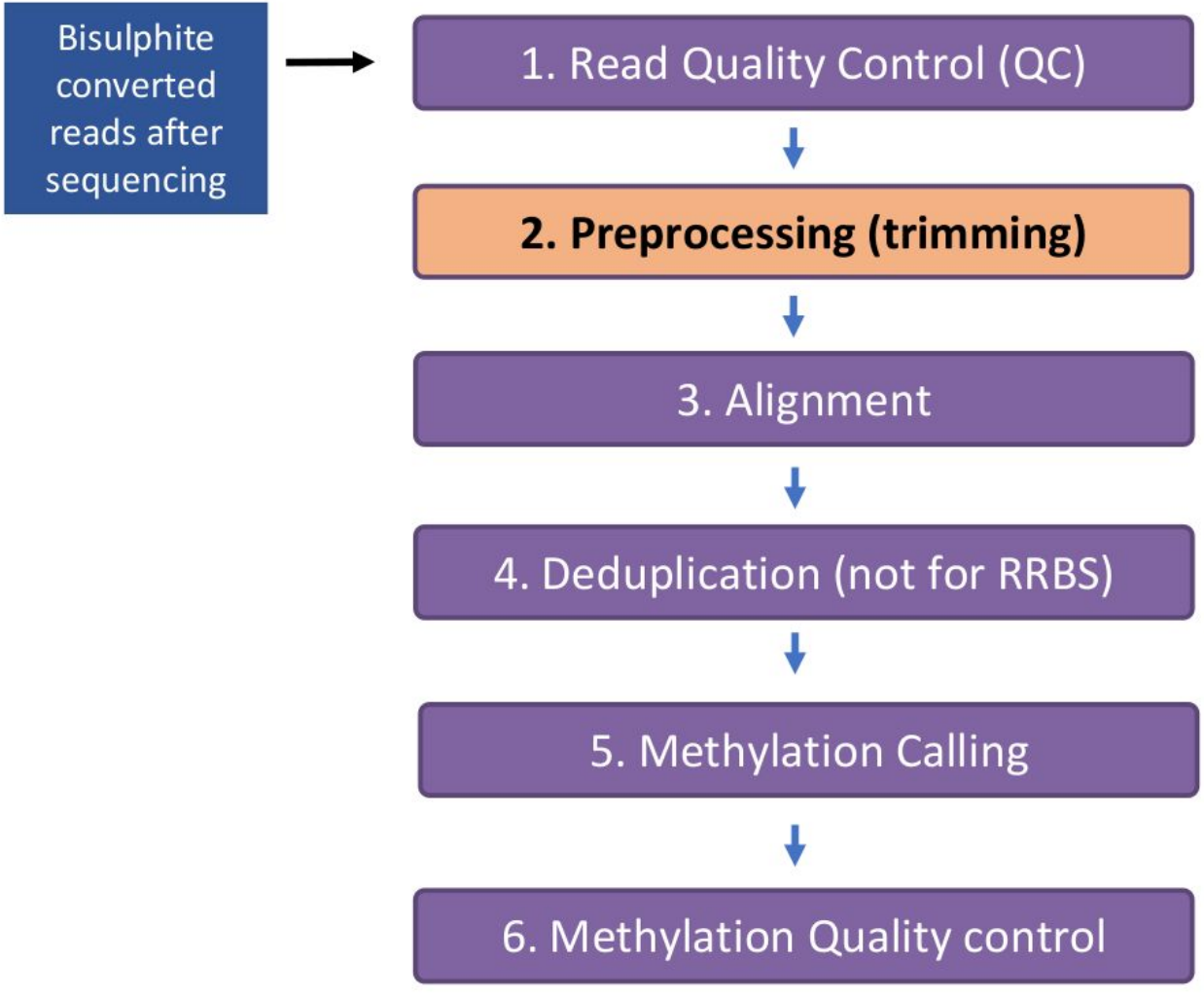
1. 20% of bases in genome are C.
2. Only 10% of C = 2% overall, are in CpG sites. Rest are always unmethylated. So converted to T.
3. Only 70%-80% of CpG sites in humans = 1.5% overall are methylated. Rest are converted to T.

What is the irregularity at the beginning?

Restriction enzyme cuts at CCGG.
Reads always start with CGG
1st C can be methylated (C) or unmethylated (T)

Why more Gs?

RRBS is enriched for CpG sites.


Bisulphite converted reads after sequencing
1. Read Quality Control (QC)
2. Preprocessing (trimming)
3. Alignment
4. Deduplication (not for RRBS)
5. Methylation Calling
6. Methylation Quality control

Preprocessing (Trimming)

TOOL **cutadapt** for WGBS

Trim Galore! wrapper for RRBS

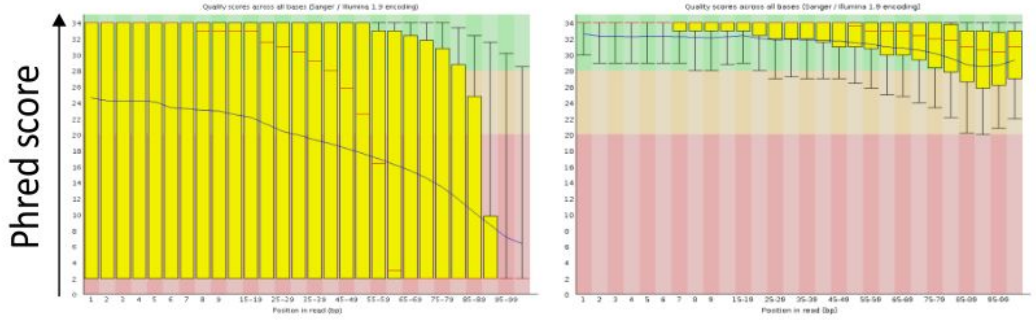
Why trimming

1. Remove poor quality basecalls at 3'
2. Remove adaptor contamination at 3'
3. **Only for RRBS:** removes extra artificially unmethylated cytosine at 3' due to end repair at restriction site

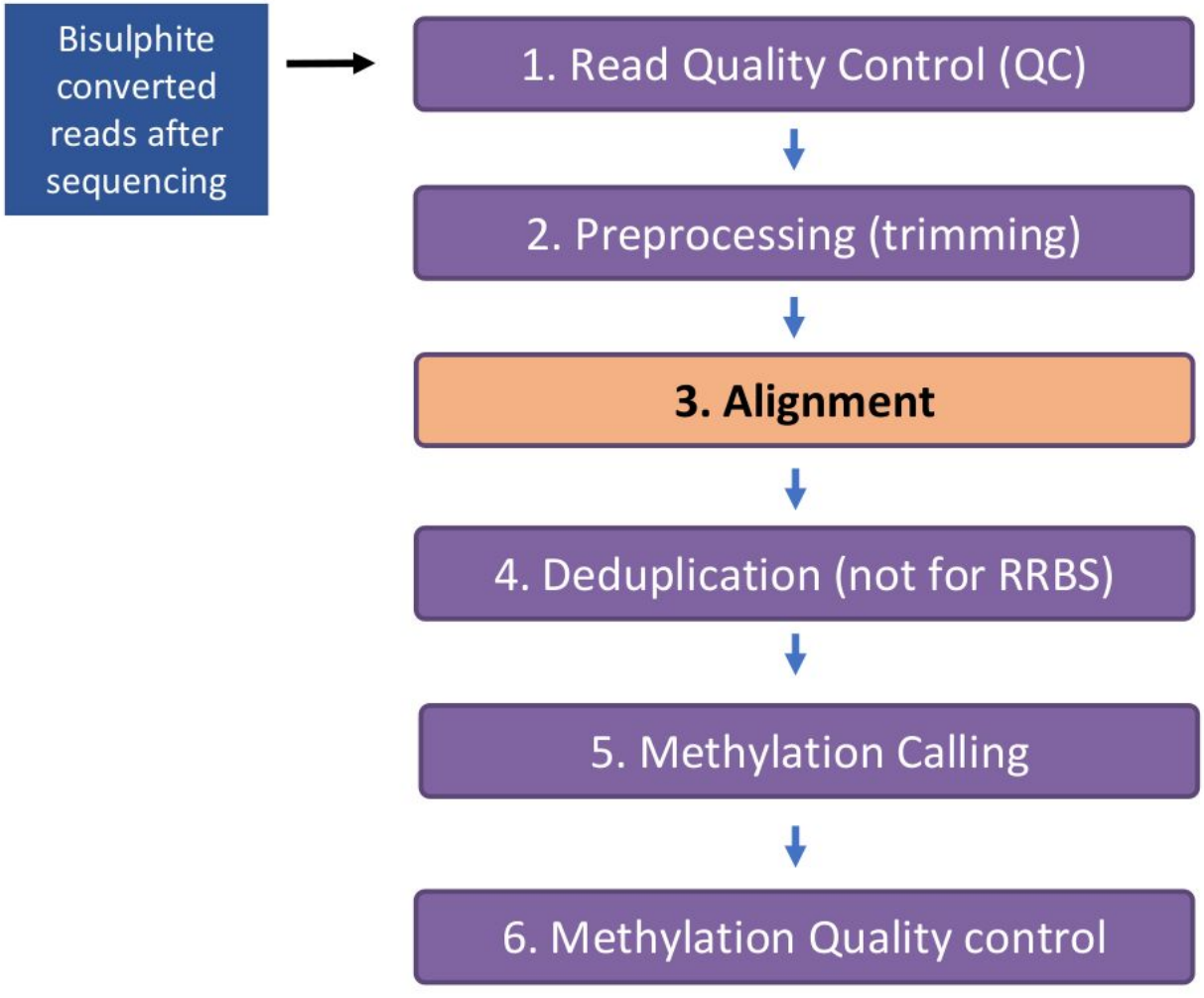
Failure to trim will result in

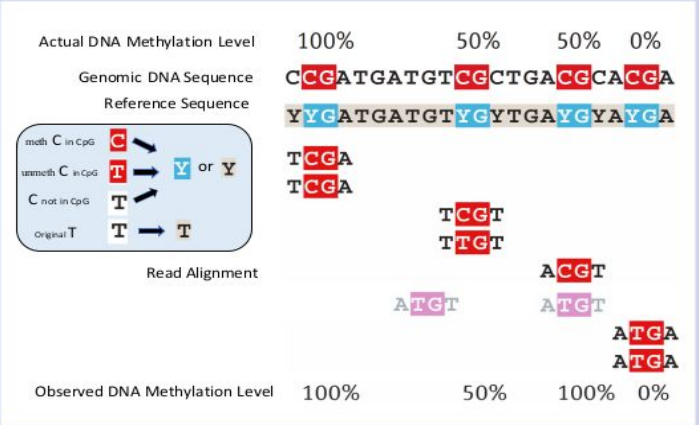
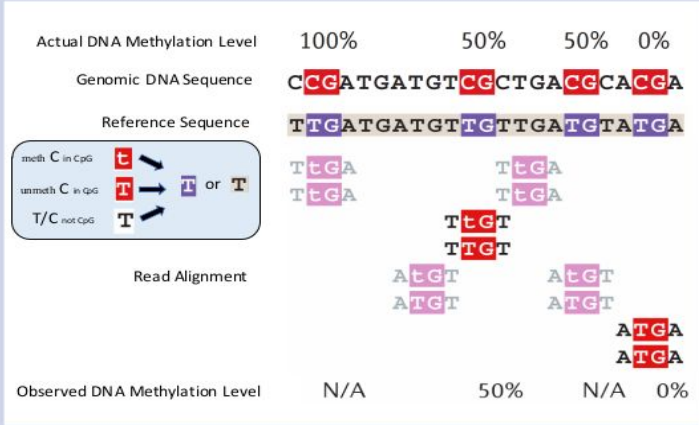
- Low mapping efficiency
- Misalignments
- Errors in methylation calls since adapters are always methylated and additional cytosine is always unmethylated (only for RRBS)

before trimming **after trimming**



Krueger & Andrews, Babraham


Bisulphite converted reads after sequencing
1. Read Quality Control (QC)
2. Preprocessing (trimming)
3. Alignment
4. Deduplication (not for RRBS)
5. Methylation Calling
6. Methylation Quality control

STRATEGIES FOR ALIGNING BISULPHITE SEQUENCING READS	Wild card aligner	Three-base aligner
Change in read sequence	None	Replace all Cs into Ts
Change in reference genome	Replace Cs into wild card letter Y	Replace all Cs into Ts
Strategy	Any Cs and Ts in read sequence match to Y in reference genome	Cs and Ts mean become one base. Effectively leaves three bases: A,G and T.
Strengths	Higher genomic coverage	No bias
Cost	Bias towards increased DNA methylation levels due to mismatch between unmethylated Cs and original Ts in read sequence	Lower genomic coverage due to mismatches due to lower sequence complexity. This can be alleviated with longer reads.
Examples	BSMAP, GSMAP, Lastm, Pash, RMAP, RRBSMAP, segmehl	Bismark, BS-Seeker, MethylCoder
Schematic examples *Bock, 2012.*		

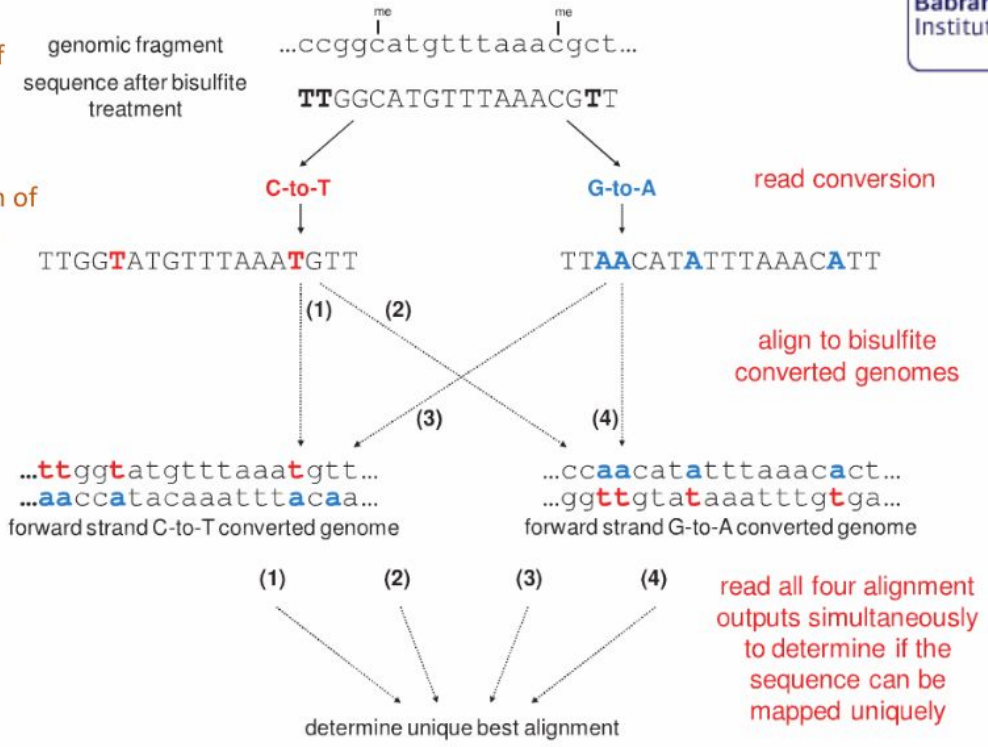
Note: Reference free approaches for non-model organisms/ unsequenced genomes. *Klughammer et al, 2015.*

Alignment

TOOL **Bismark**

Biological conversion of unmethylated C to T

In silico conversion of all C to T in reads and reference



Also have suite of tools for bisulphite sequencing analysis

Widely used 3-base aligner

After alignment reverse the *in silico* conversion to call methylation

Krueger & Andrews, 2011.

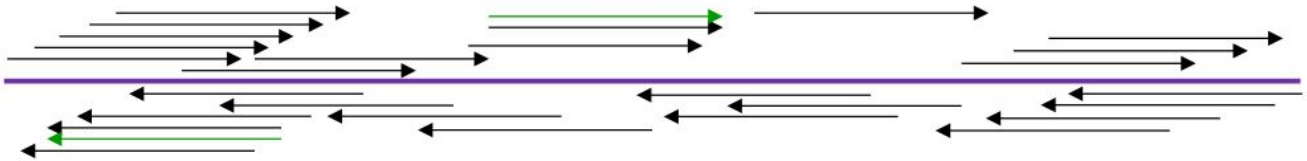
Deduplication (only for WGBS)

TOOL **Bismark**

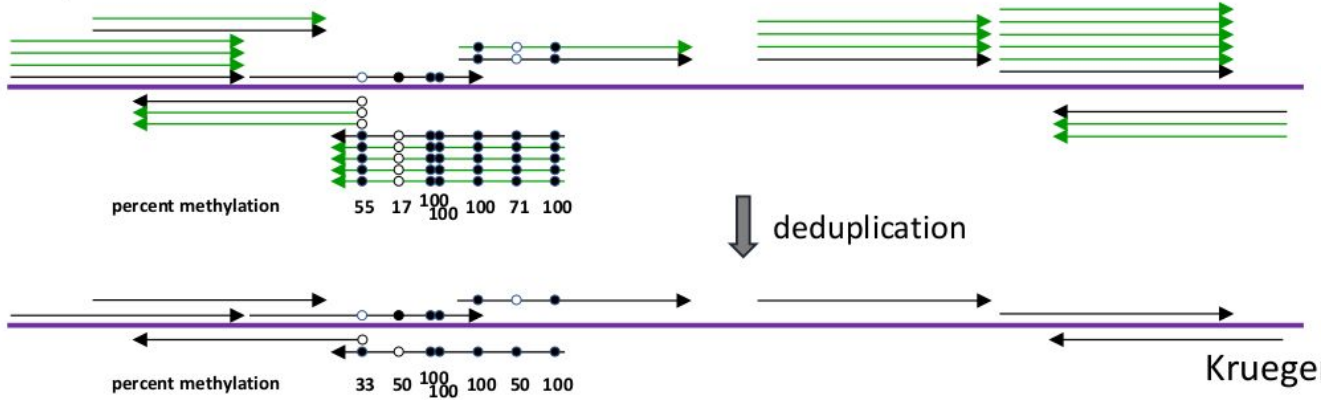
WGBS: Random DNA fragmentation

unlikely to sequence several cells with exact copies of the same fragment amongst >5bn possible fragments with different start sites for WGBS
➔ Duplicates are likely PCR amplifications (artificial copies of same fragment introduced during library preparation)

Complex/diverse library:



Duplicated library:



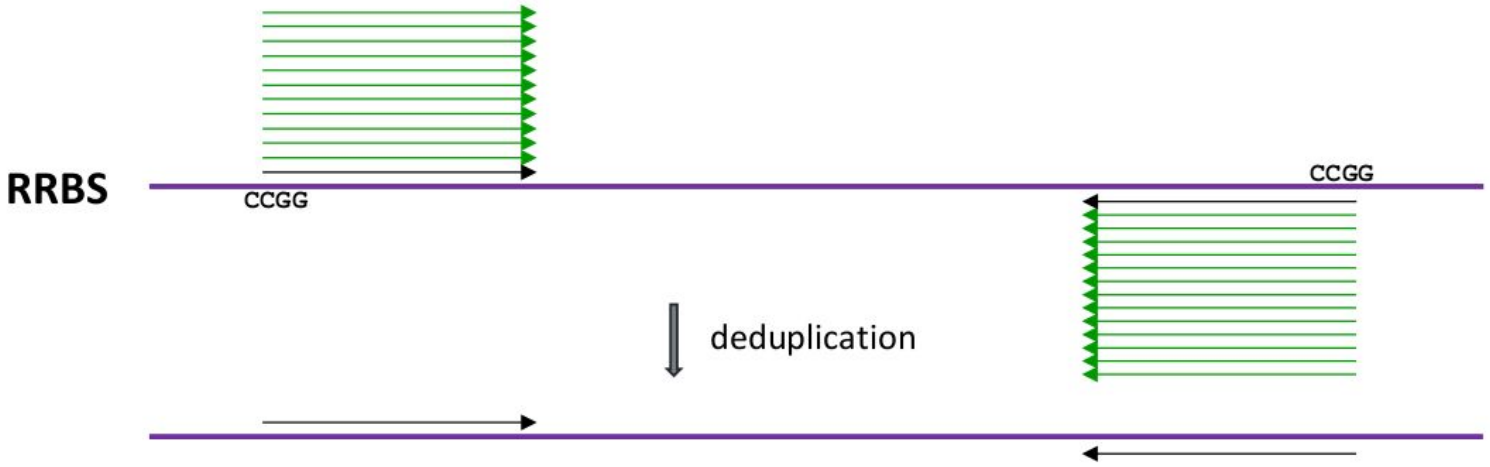
Duplicates provide redundant information which will bias your results

Krueger & Andrews, Babraham

Deduplication (not suitable for RRBS)

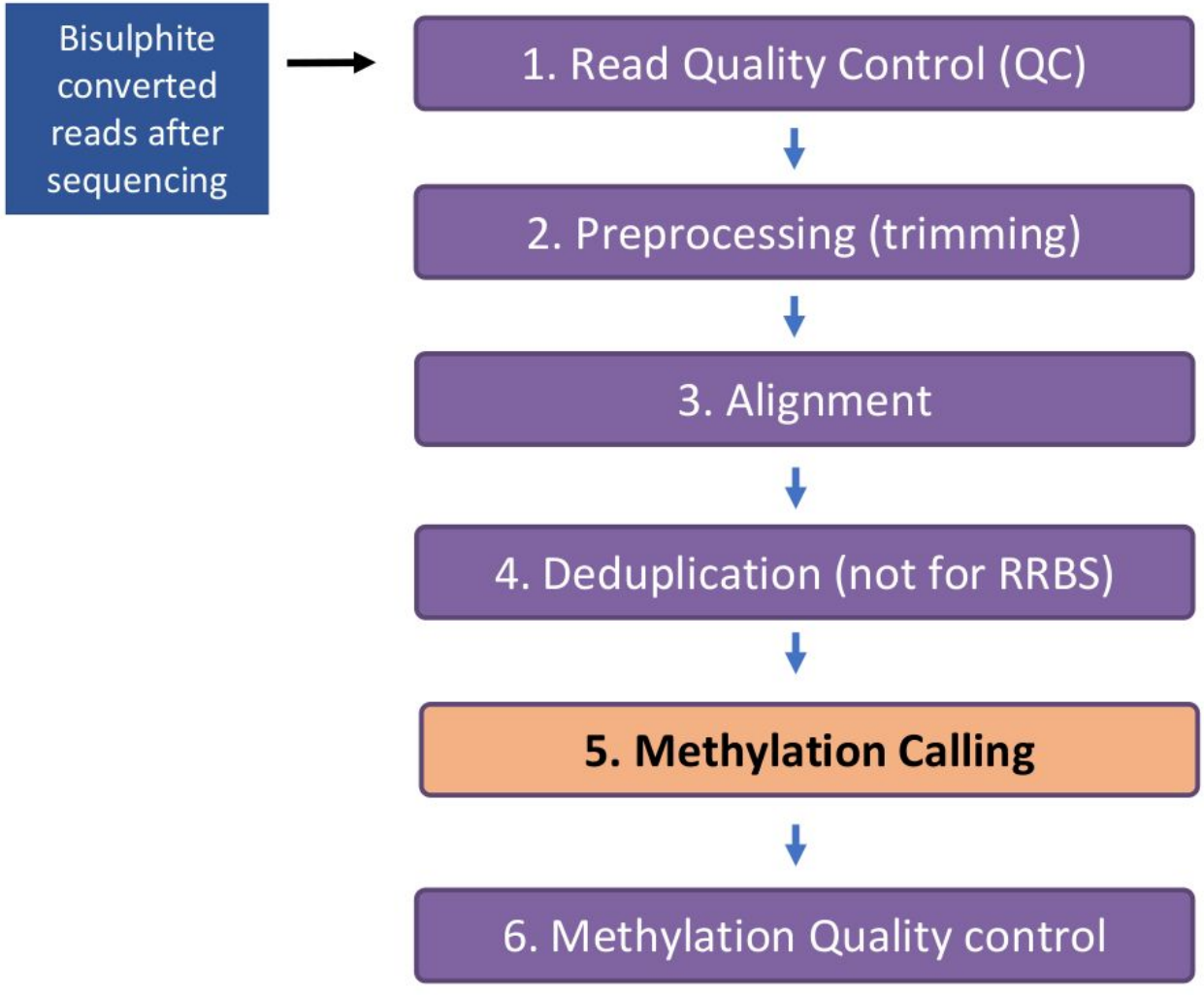
NOT advisable for RRBS or other target enrichment methods

Restriction enzyme digestion means that start and end sites for DNA fragments are identical by design
Restriction enzyme (*MspI)* always cuts at CCGG
➔ Duplicates can represent different cells



Deduplication results in loss of reads from different DNA fragments representing distinct cells. Not advisable.

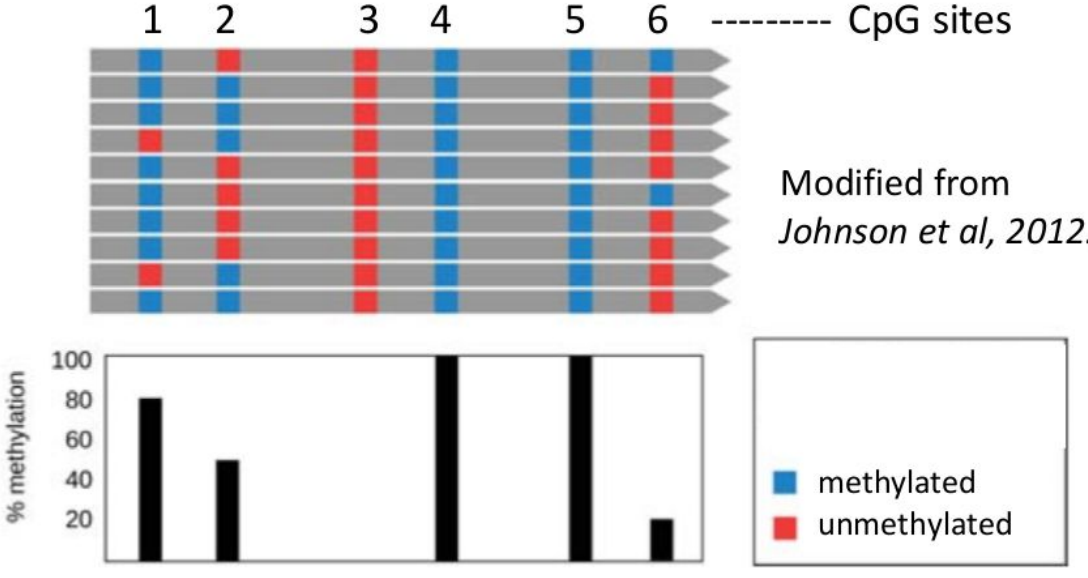
Krueger & Andrews, Babraham


Bisulphite converted reads after sequencing
1. Read Quality Control (QC)
2. Preprocessing (trimming)
3. Alignment
4. Deduplication (not for RRBS)
5. Methylation Calling
6. Methylation Quality control

Extraction of methylation calls

TOOL **Bismark**

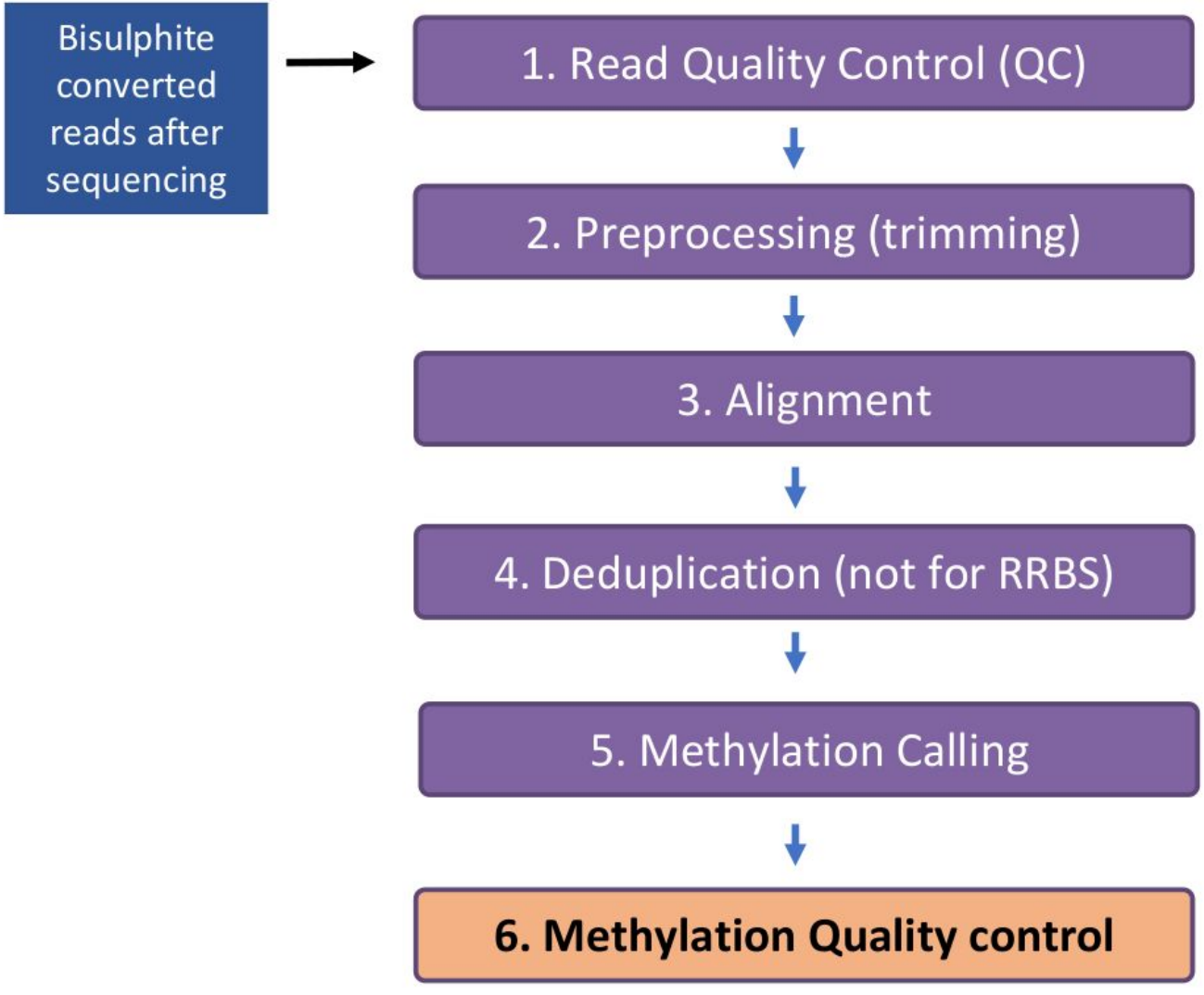
Bismark methylation extractor



Modified from *Johnson et al, 2012.*

Additional steps for mammals

- Only keep CpG sites
- Merge methylation information from top and bottom strand
  - Accurate methylation estimates
  - Twice the coverage

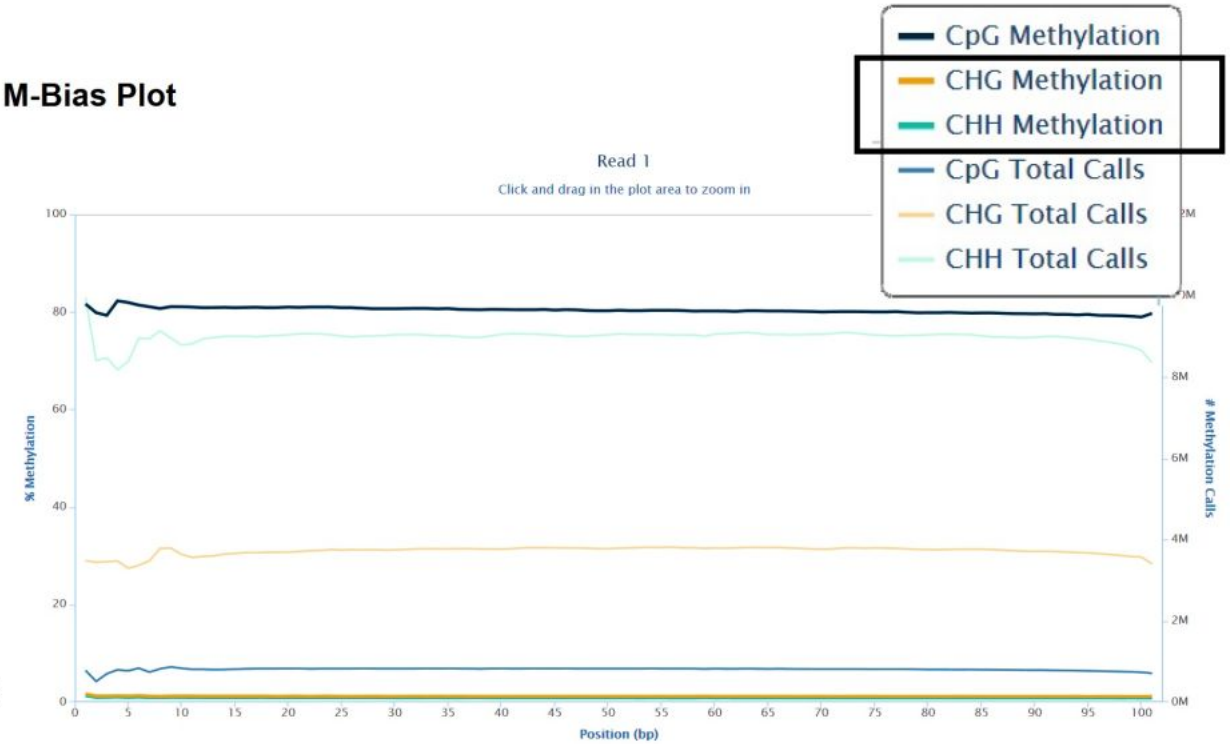

Bisulphite converted reads after sequencing
1. Read Quality Control (QC)
2. Preprocessing (trimming)
3. Alignment
4. Deduplication (not for RRBS)
5. Methylation Calling
6. Methylation Quality control

Methylation call QC (Bisulfite conversion)

TOOL **Bismark**

Bisulphite conversion converts all unmethylated Cs to Ts
A mark of complete bisulphite conversion is:
CHH and CHG methylation should be ~0% in mammals

M-Bias Plot



H= A,T or C

Krueger & Andrews, Babraham

Methylation call QC (C/T SNP filtering)

TOOL **Bis-SNP**

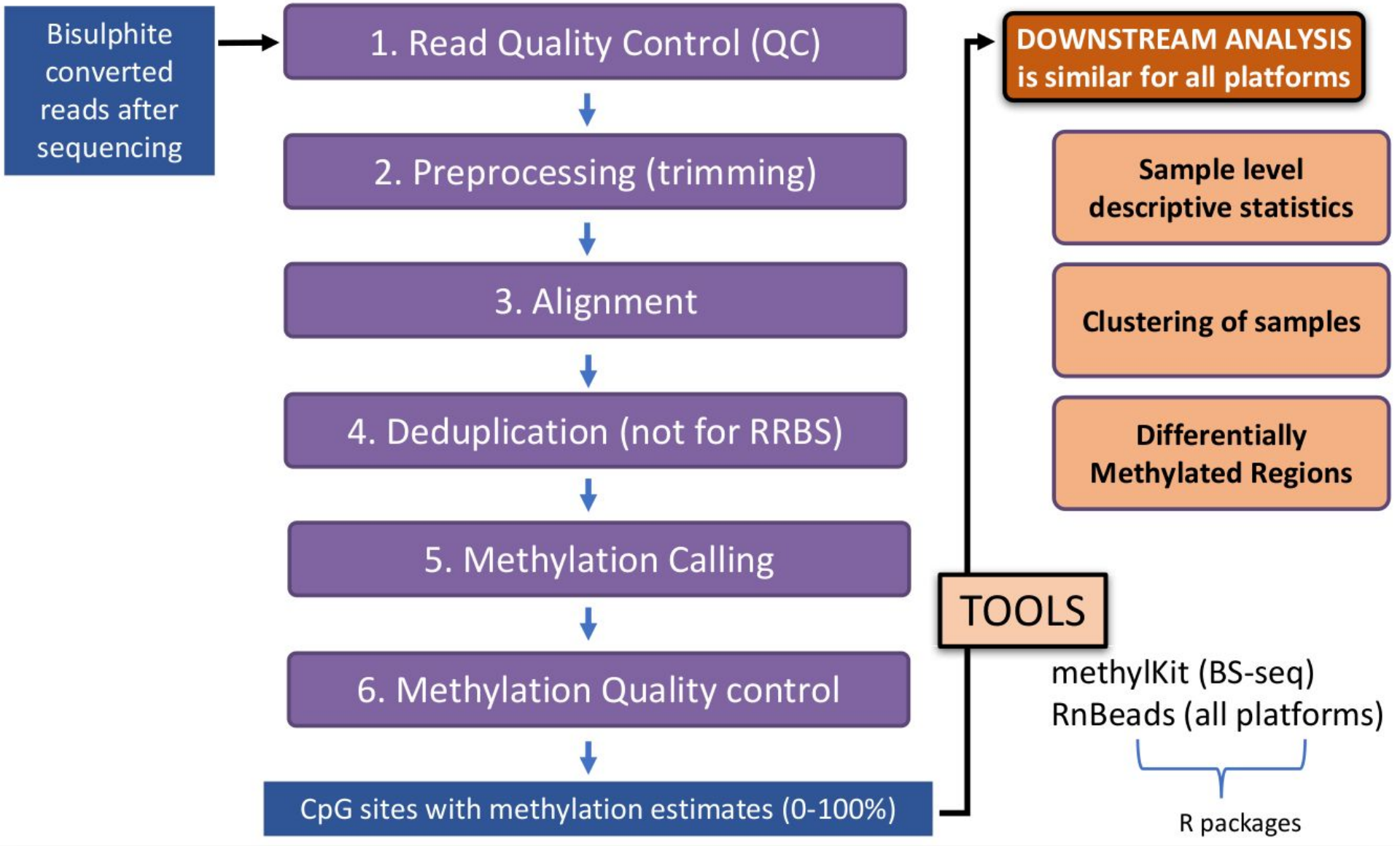
METHOD Open Access

Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data

Yaping Liu[1,2], Kimberly D Siegmund[3], Peter W Laird[1] and Benjamin P Berman[1,3*]

Bisulphite conversion converts all unmethylated Cs to Ts. **C → T**
C/T SNPs at CpG sites will look similar. **C → T**. Can confound methylation calls.

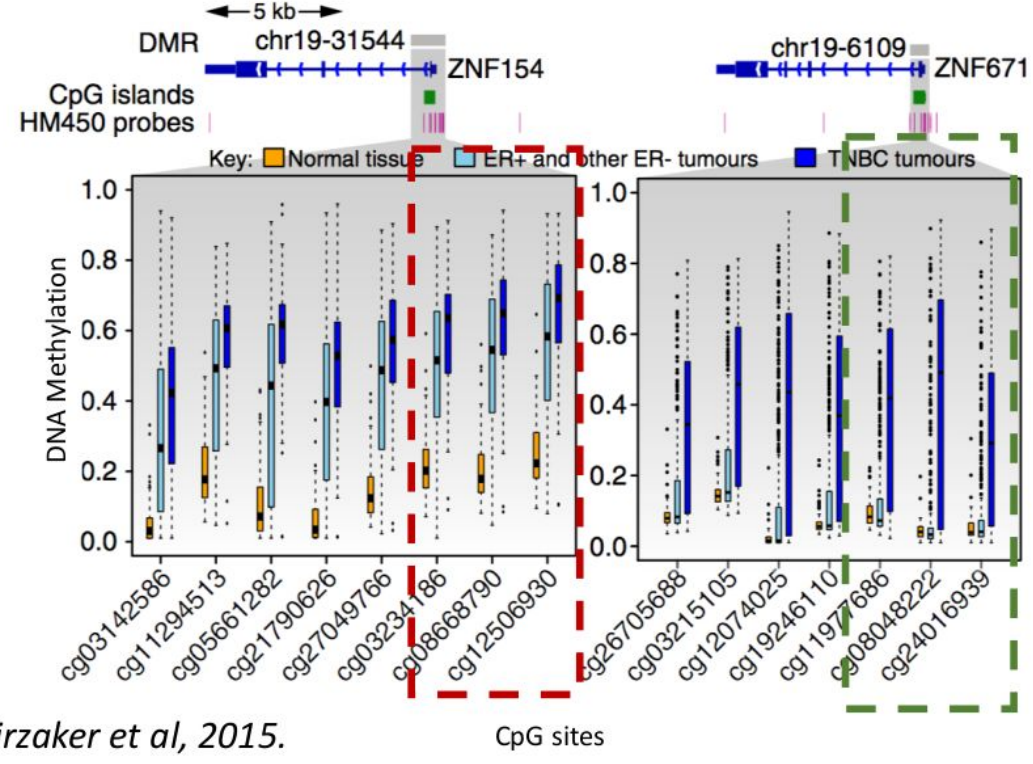
- *Bis-SNP* can discriminate between the two by investigating the complementary strand.
- *Remove SNPs from results.*


Bisulphite converted reads after sequencing
1. Read Quality Control (QC)
2. Preprocessing (trimming)
3. Alignment
4. Deduplication (not for RRBS)
5. Methylation Calling
6. Methylation Quality control
CpG sites with methylation estimates (0-100%)
DOWNSTREAM ANALYSIS is similar for all platforms
Sample level descriptive statistics
Clustering of samples
Differentially Methylated Regions
TOOLS
methylKit (BS-seq)
RnBeads (all platforms)
R packages

Cancer diagnosis

Identify differentially methylated regions. These can be used as biomarkers to aid in

1. **Diagnosis of tumour tissues**
2. **Identification of tumour subtype**



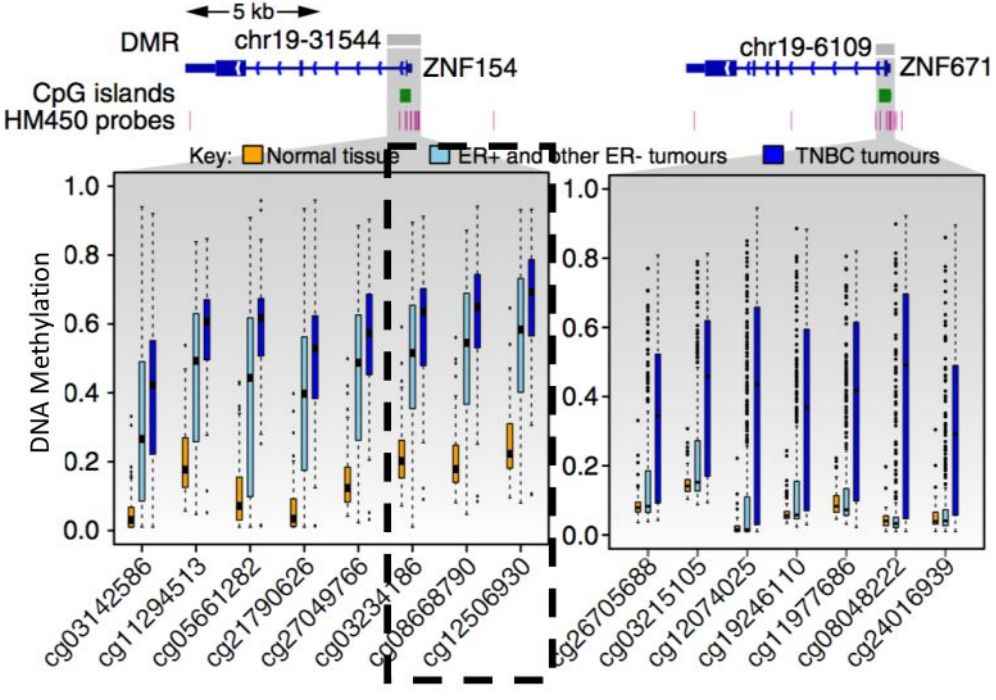
Stirzaker et al, 2015.

- DNA methylation is a robust biomarker.
- Represents a stable readout of the epigenetic state and transcriptional activity of sample.
- Easy to profile compared to other epigenetic marks.

Cancer monitoring

Identify differentially methylated regions. These can be used as biomarkers to aid in

1. **Diagnosis of tumour tissues**
2. **Identification of tumour subtype**
3. **Monitor tumour decline or recurrence in circulating tumour DNA (ctDNA) found in blood**

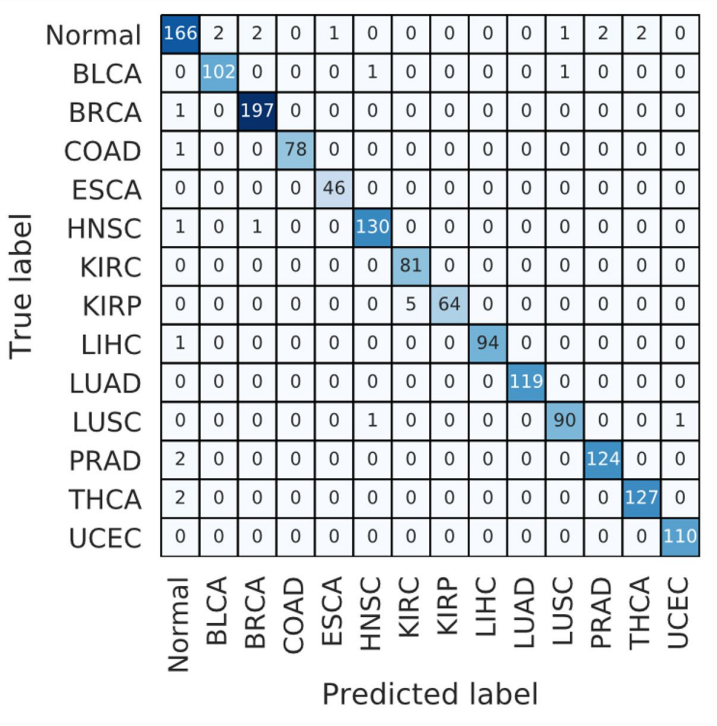


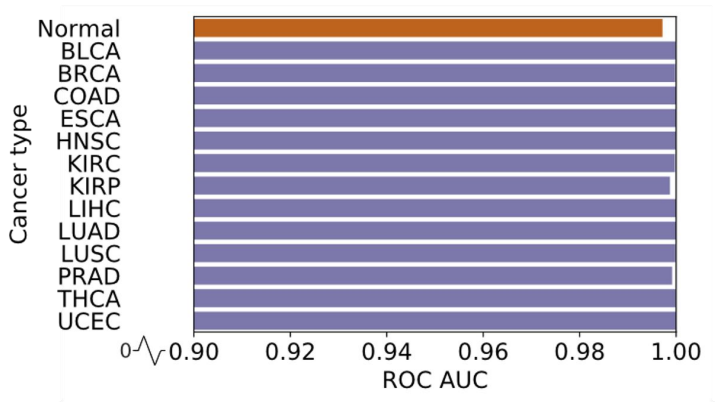
Stirzaker et al, 2015.

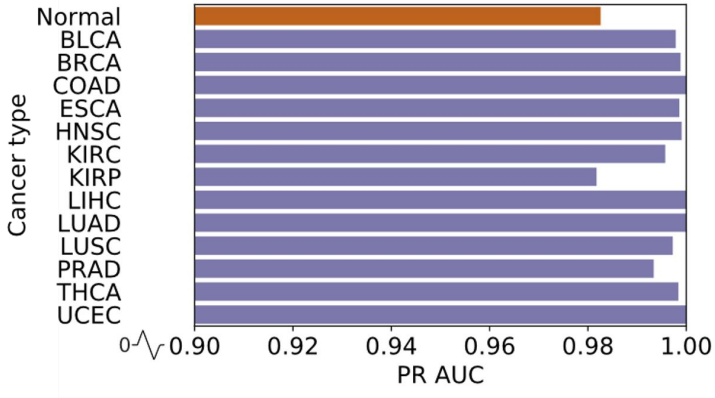
- DNA methylation is a robust biomarker.
- Represents a stable readout of the epigenetic state and transcriptional activity of sample.
- Easy to profile compared to other epigenetic marks.

Machine Learning for cancer detection

Multiclass Gradient Boosted decision trees





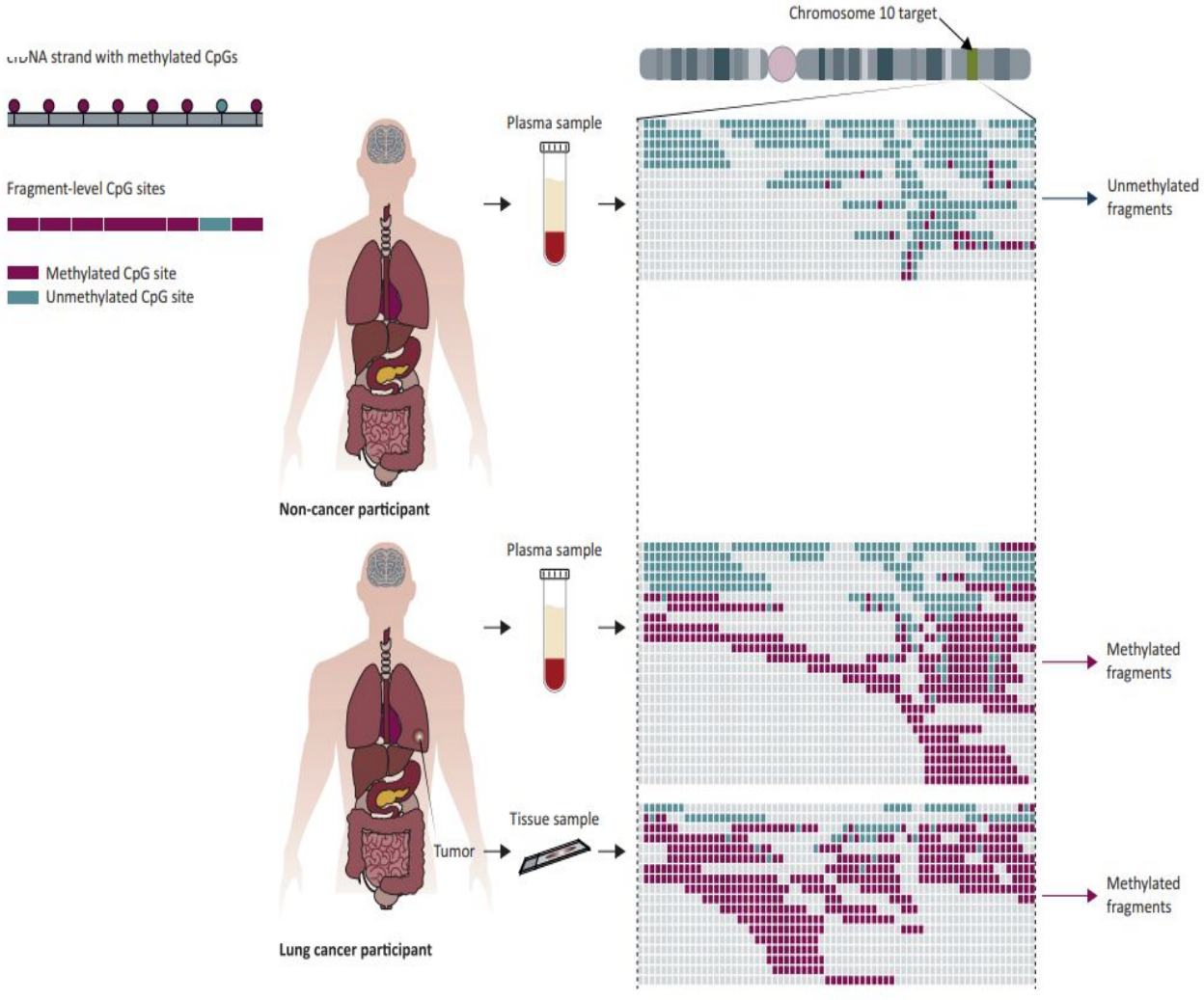


KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
PRAD	Prostate adenocarcinoma
THCA	Thyroid carcinoma
UCEC	Uterine Corpus Endometrial Carcinoma

Newsham et al bioRxiv 2020

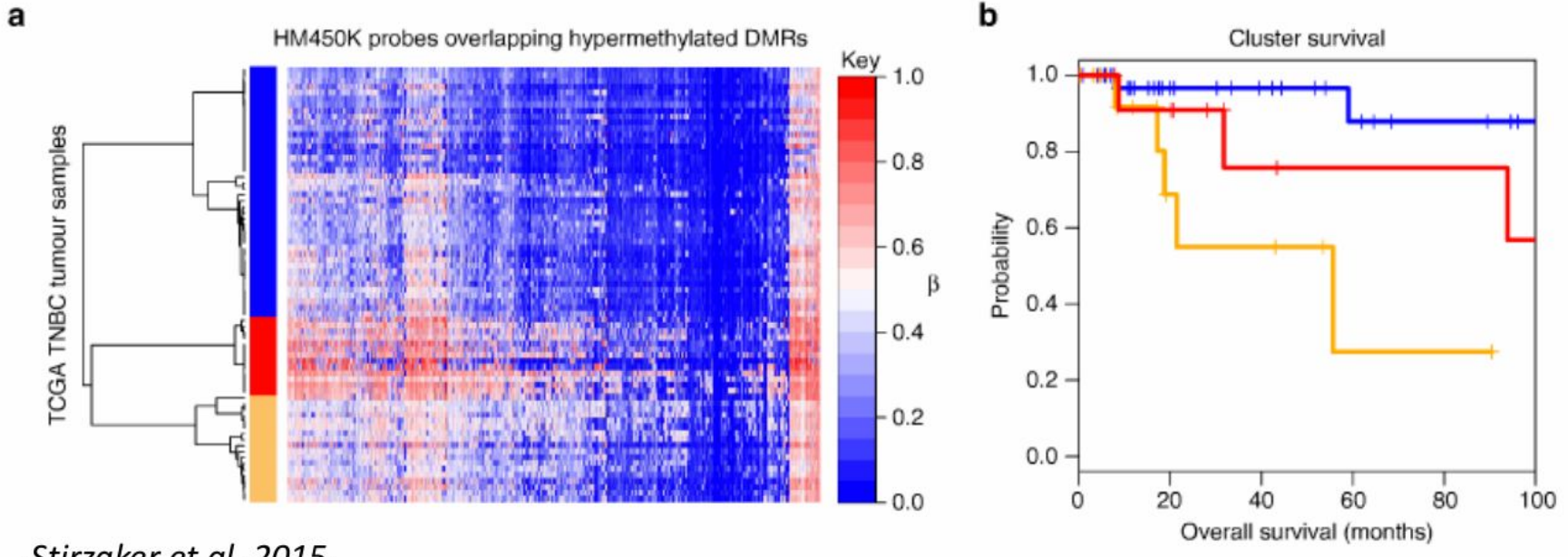
Cancer Early Detection

- Liquid biopsy based methylated ctDNA detection



Liu et al 2020

Novel Subtype Detection and Prognostics



Stirzaker et al, 2015.

- Clustering of DNA methylation profiles over all tumours to classify tumours into novel groups with distinct prognosis. Leads to construction of tumour classifier.
- Can apply this classifier on new tumours to identify subgroup which will enable treatment stratification and estimation of survival.