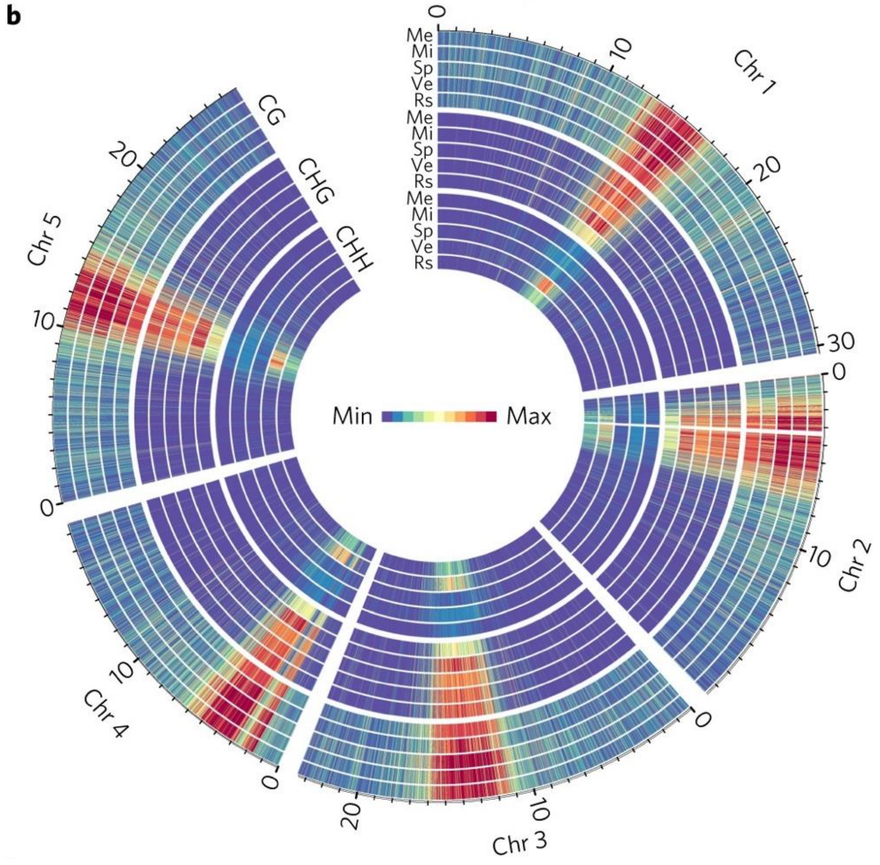
Introduction to epigenetics (DNA methylation)



Judit Tálas

Data Science Meeting

May 2025

What is epigenetics?

“stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence”*

Eukaryotes

Primary function - gene expression regulation and genome integrity

Chromatin modifications such as:

- Histone modifications
- DNA methylation
  - DNA methylation associated pathways:
    - animals: piRNA pathway, other small RNA pathways (RNAi), TE silencing
    - plants: RNA directed DNA methylation, other small RNA pathways (RNAi), TE silencing

Prokaryotes

RM systems

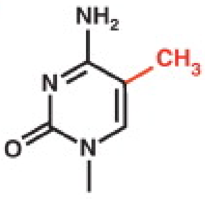
Transcriptional regulation (e.g. phase variation)

Cell cycle (*Caulobacter, E. coli*)

Caveat – I am not an expert on prokaryotic DNA methylation!

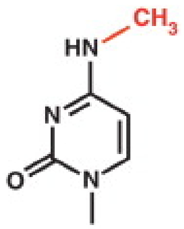
*Berger et al. (2009)

What is DNA methylation?



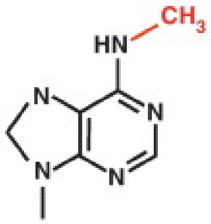
5mC

5-methylcytosine (prokaryotic and eukaryotic) (5mC)
The main methylation mark that you will find in animals and plants



4mC

N4-methylcytosine (prokaryotic and rarely eukaryotic) (4mC)

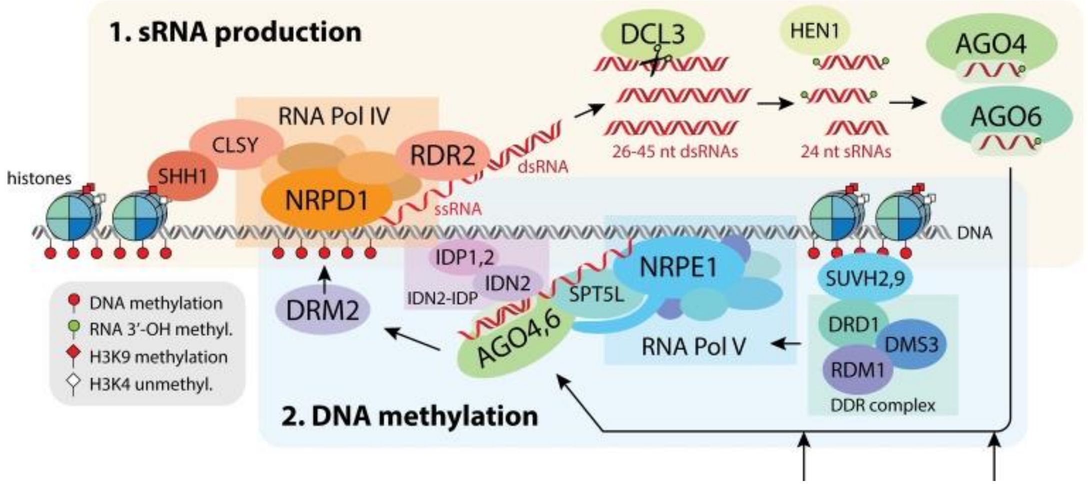


6mA

N6-methyladenine (prokaryotic) (6mA)

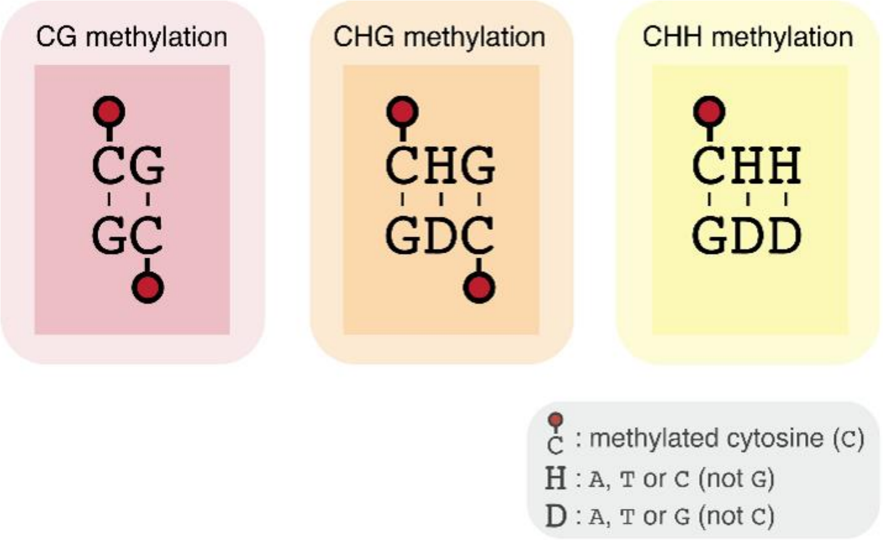
Deposited and maintained by DNA methyltransferases that recognise a motif (in prokaryotes) or in eukaryotes they are recruited by:

- chromatin remodelers
- histone modifications
- Non-coding RNAs
- DNA methylation itself



Erdmann et al. (2020)

What is DNA methylation?



DNA methylation is established and maintained by different pathways in different sequence contexts

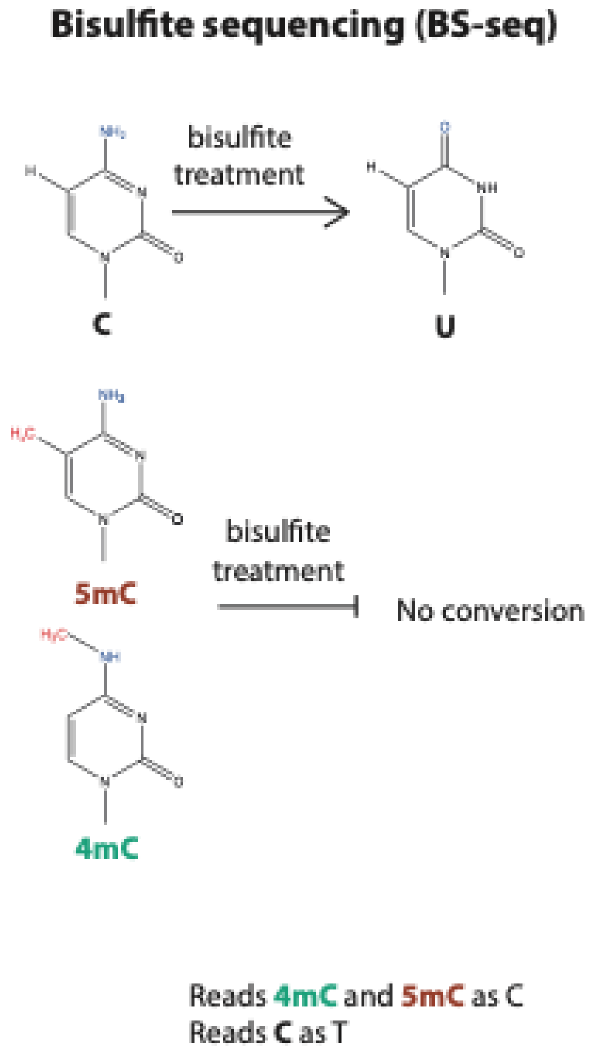
The main form of methylation in animals is CpG methylated islands

CG and CHG are symmetric, and CHH is asymmetric

They are established and maintained in different but related pathways

Erdmann et al. (2020)

How do we detect DNA methylation?



Reads 4mC and 5mC as C
Reads C as T

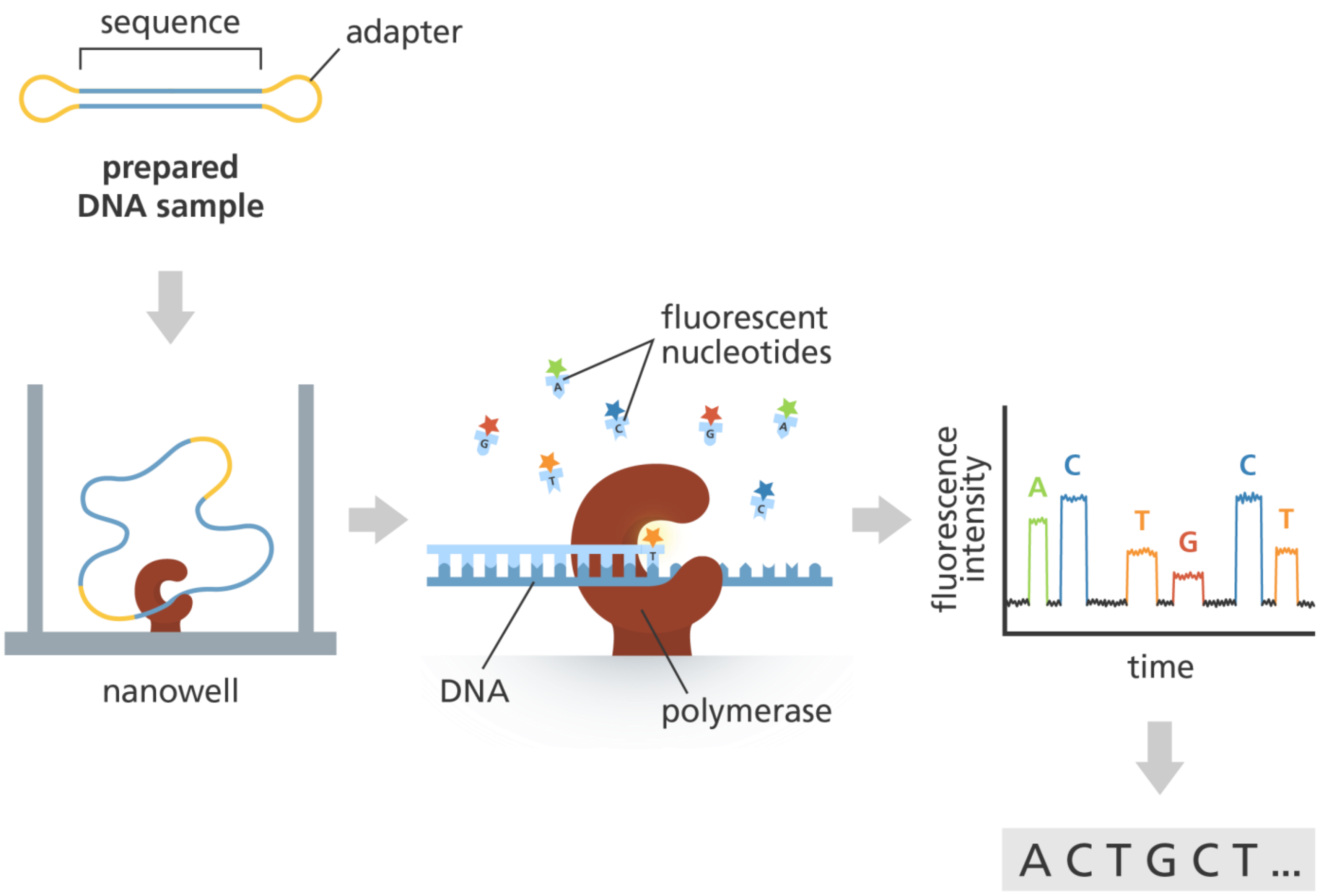
In the sequencing readout, every C is a methylated cytosine

Unmethylated cytosines are distinguished by aligning to a reference genome

Methylated bases are called in every aligned read

Other related sequencing technologies are also used which rely on enzymatic conversion of unmethylated cytosines
(AMD-seq, EM-seq, TAB-seq)

How do we detect DNA methylation?



Long read sequencing for modified bases

Single Molecule Real-Time sequencing (SMRT-seq) by PacBio

Nanopore sequencing

SMRT-seq (PacBio)

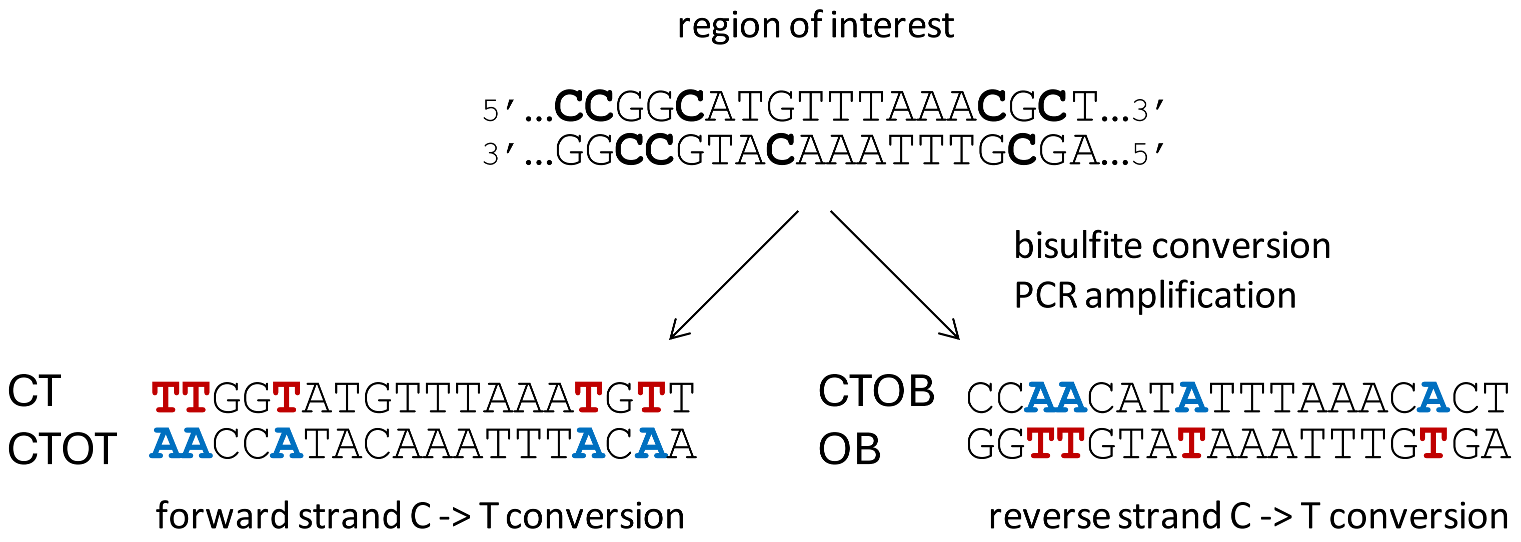
So you have the data – now what?

Get ready for some great tool names

1. **QC (quality filtering, adapter removal (e.g. fastqc, cutadapt, deduplication))**
    1. Check or correct for bisulfite conversion
    2. Use controls for unmethylated DNA such as chloroplast, or spike-in (e.g. lambda phage), and optionally in-vitro methylated sequences
2. **Mapping and methylation calling**

    Most popular tool is bismark (which we will explore next) or less commonly abismal
3. **Extract sequence context specific methylation data** from mapped BAM file conversion correction (bismark, MethylDackel – formerly PileOMeth)
4. **Downstream analysis such as**:
    1. visualise methylated regions using a genome browser (IGV or karyoplotR, pyGenomeTracks)
    2. differentially methylated regions (DMRs) (methpipe (CLI), MethylKit (R), DMRcaller (R), or custom scripts)
    3. Conduct ends analysis (look at binned average methylation levels around TSS of genes or TEs)

Mapping and methylation calling with bismark



region of interest

5′...**CC**GG**C**ATGTTTAAA**C**G**C**T...3′
3′...GG**CC**GTA**C**AAATTTG**C**GA...5′

bisulfite conversion
PCR amplification

CT **TT**GG**T**ATGTTTAAA**T**G**T**T
CTOT **AA**CC**A**TACAAATTT**A**C**A**A

forward strand C -> T conversion

CTOB CC**AA**CAT**A**TTTAAAC**A**CT
OB GG**TT**GTA**T**AAATTTG**T**GA

reverse strand C -> T conversion

Mapping is performed by bowtie2 or hisat2

Directional or non-directional libraries

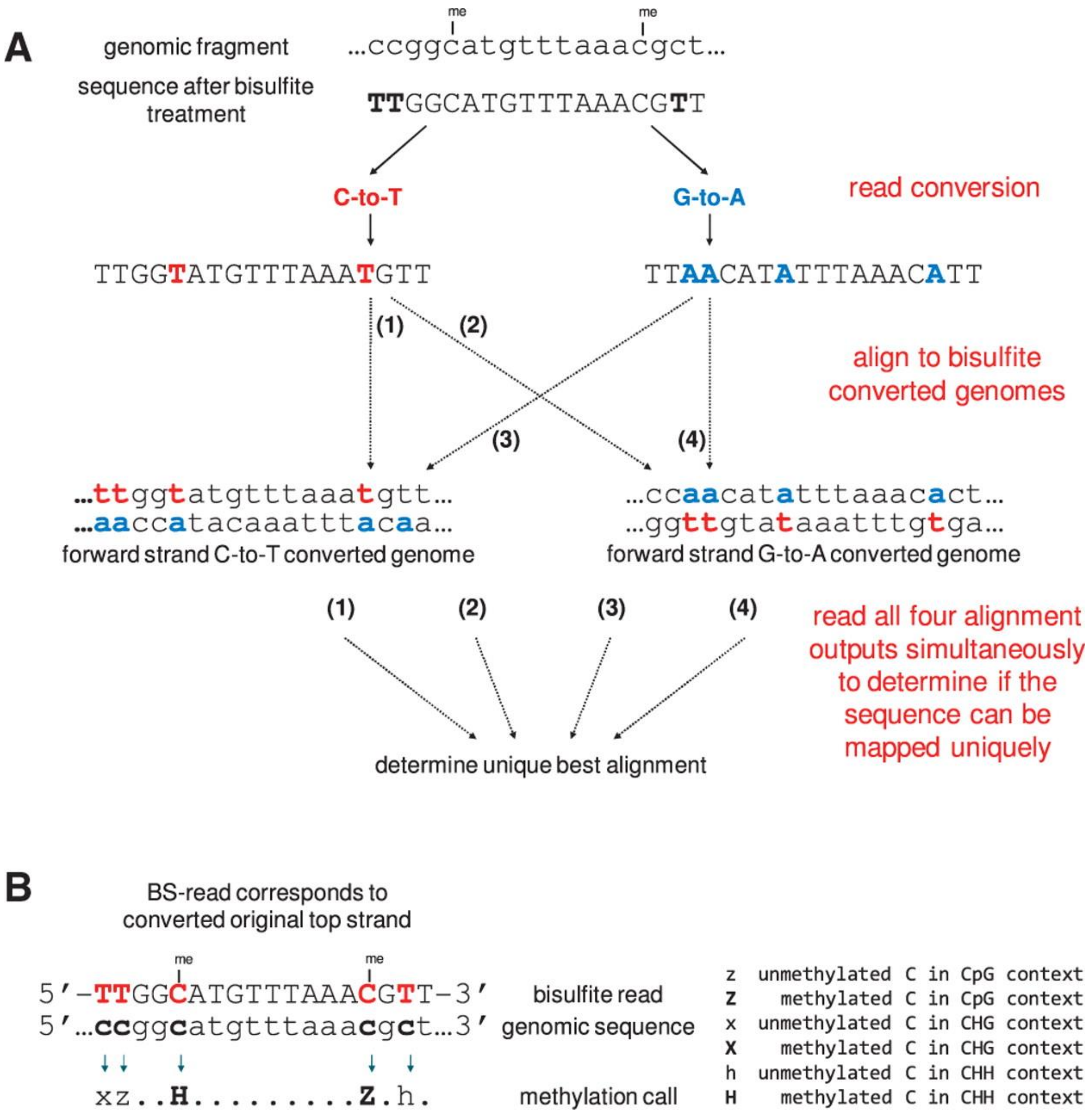
If non-directional, we could sequence 4 different reads for a given region of interest

The original top or bottom strands (CT or OB)

Or their complements (CTOT or CTOB)

Krueger et al. (2011)

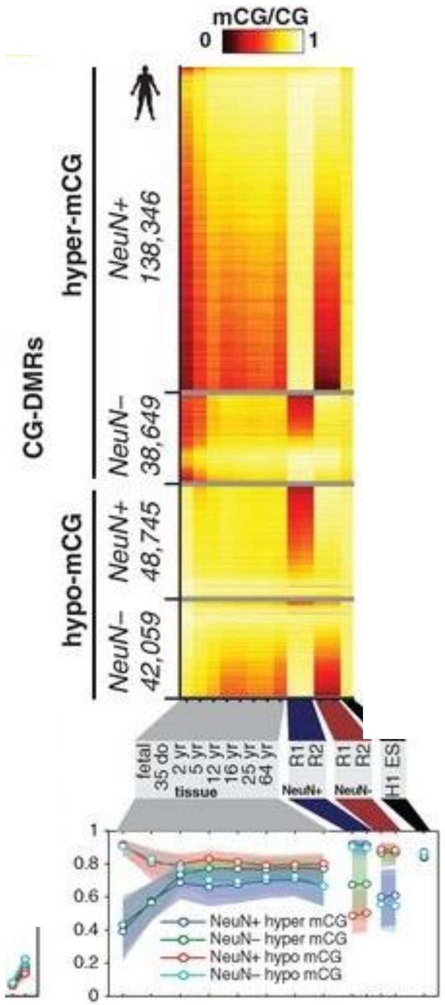
Mapping and methylation calling with bismark


A
genomic fragment
me
me
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment
B
BS-read corresponds to converted original top strand
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

Krueger et al. (2011)

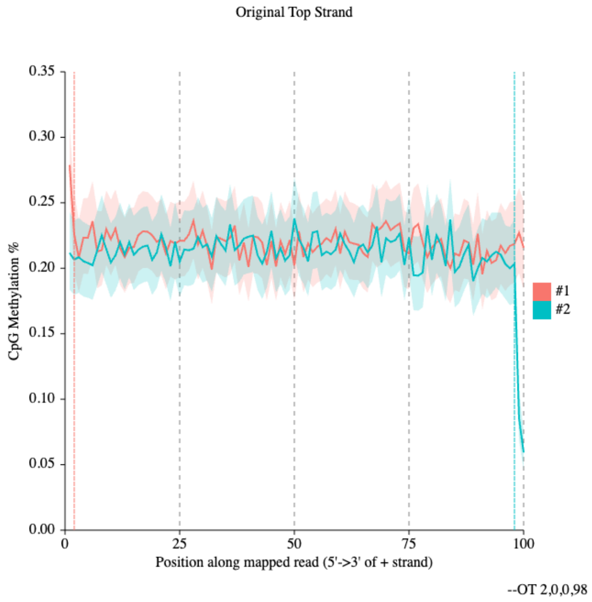
Downstream analysis

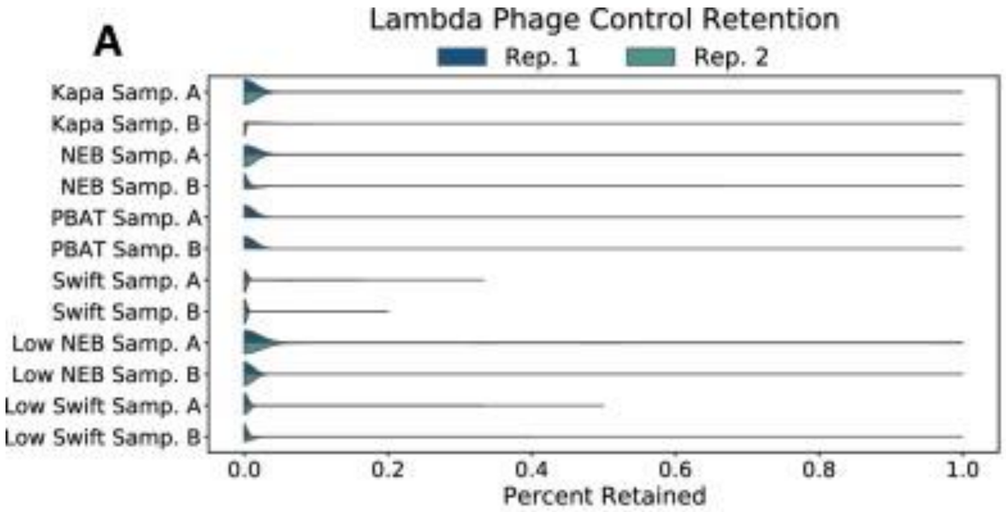
1. **bismark** methylation extractor, MethylDackel, or **custom parsers** to create bedgraph files
   1. Split files into CpG CHG and CHH contexts
   2. Bin methylation information to a specified window (for example every 50bp or 100bp to aid visualisation)
   3. Or convert to bigwig files for easy visualisation (deepTools)
2. **Visualise methylated regions** using a genome browser (IGV or karyoplotR)
3. **Determine differentially methylated regions** (DMRs) (methpipe/DNMTools (CLI), MethylKit (R), DMRcaller (R), BSmooth algorithm, or custom scripts with secondary statistical validation)
4. **Conduct ends analysis** (look at binned average methylation levels around TSS of genes, TEs or regions of interest)



Lister et. al. (2013)

Data interpretation - controls





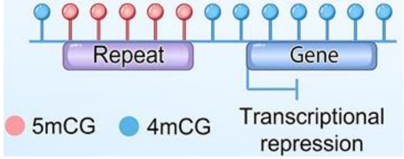
Bias/noise at the beginning of reads
Due to incomplete conversion or amplification bias
MethylDackel/bismark can visualise this with M bias plots

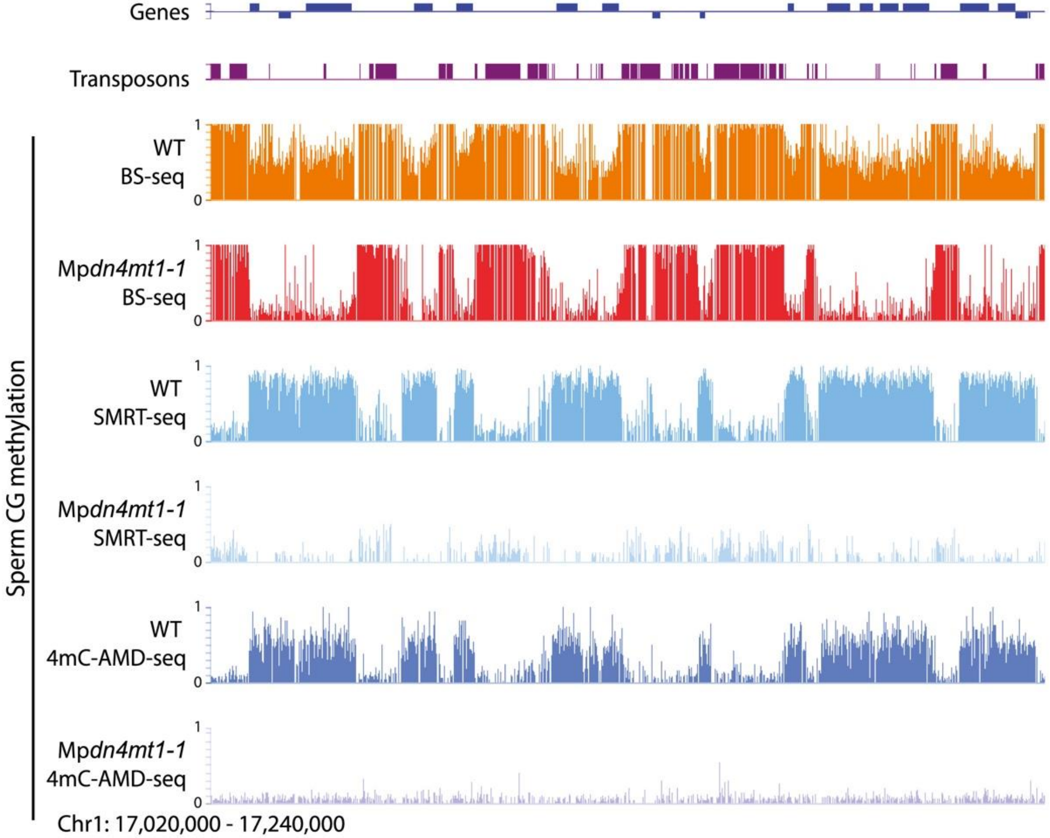
Bisulfite conversion rates

- Spike-in controls (e.g. lambda phage) and in vitro methylated controls
- Internal controls (less powerful)
  - Plants: chloroplast (unmethylated)
  - Mammals: non-CG methylation (in theory very low)
- Adjust error rates based on conversion rates

MethylDackel
Morrison et al. (2021)

Data interpretation





Even though we know there is no 4mC deposited in transposons, the detected levels aren't zero.

BS-seq: 5mC that wasn't fully deaminated will appear as C in sequencing

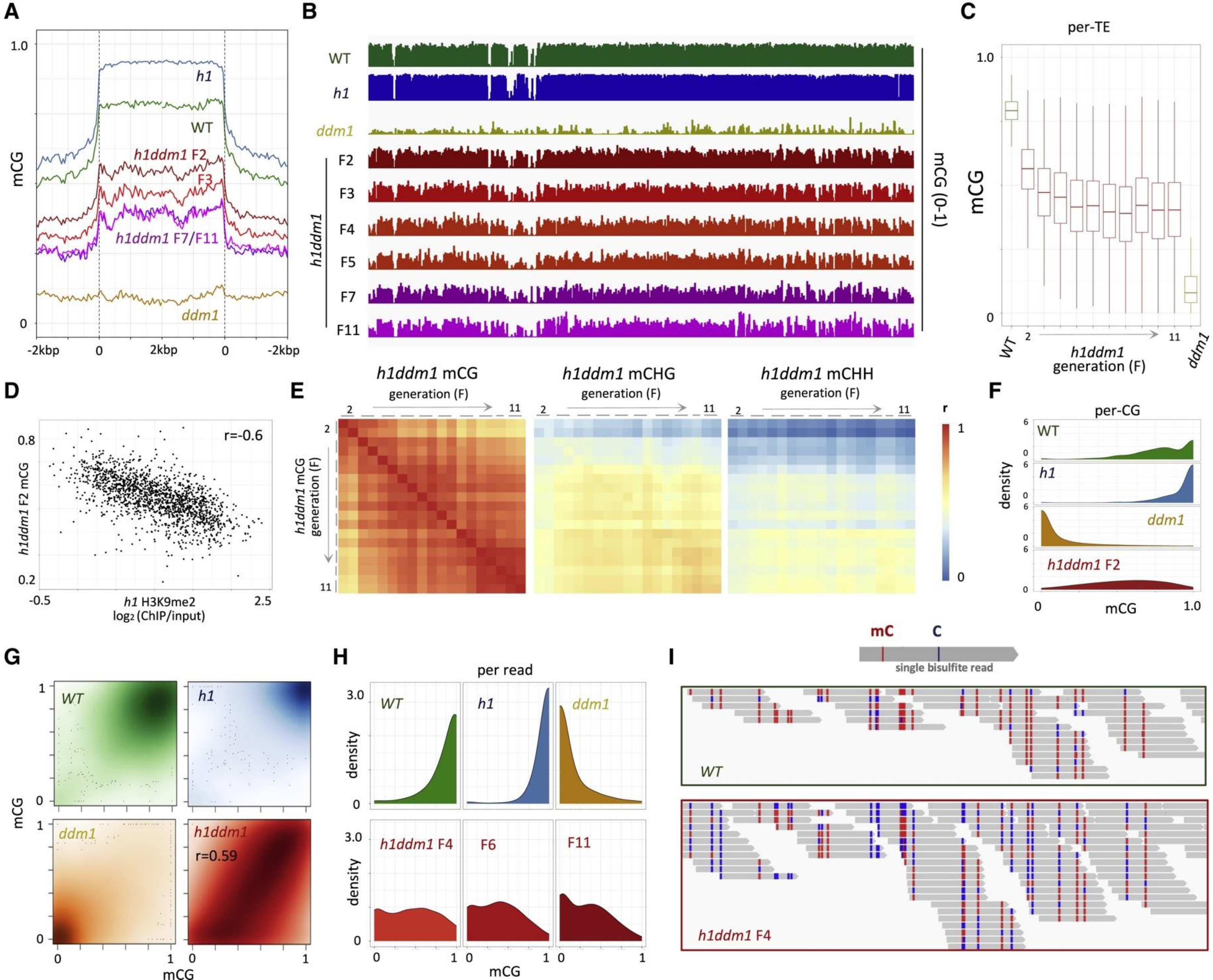
SMRT-seq: spurious methylation calls within TEs:

The PacBio algorithm is trained on bacterial 4mC methylation – this can result is spurious base calling as detection is determined by effect of modified base on polymerase kinetics, so high coverage is needed for confidence

Use controls, and validate using other methods (in this case, LC-MS, other sequencing technologies (TAB-seq, AMD-seq) immunoblot

Walker et al. (2025)

What will the data look like?



Ends analysis

Methylation traces

Box plots

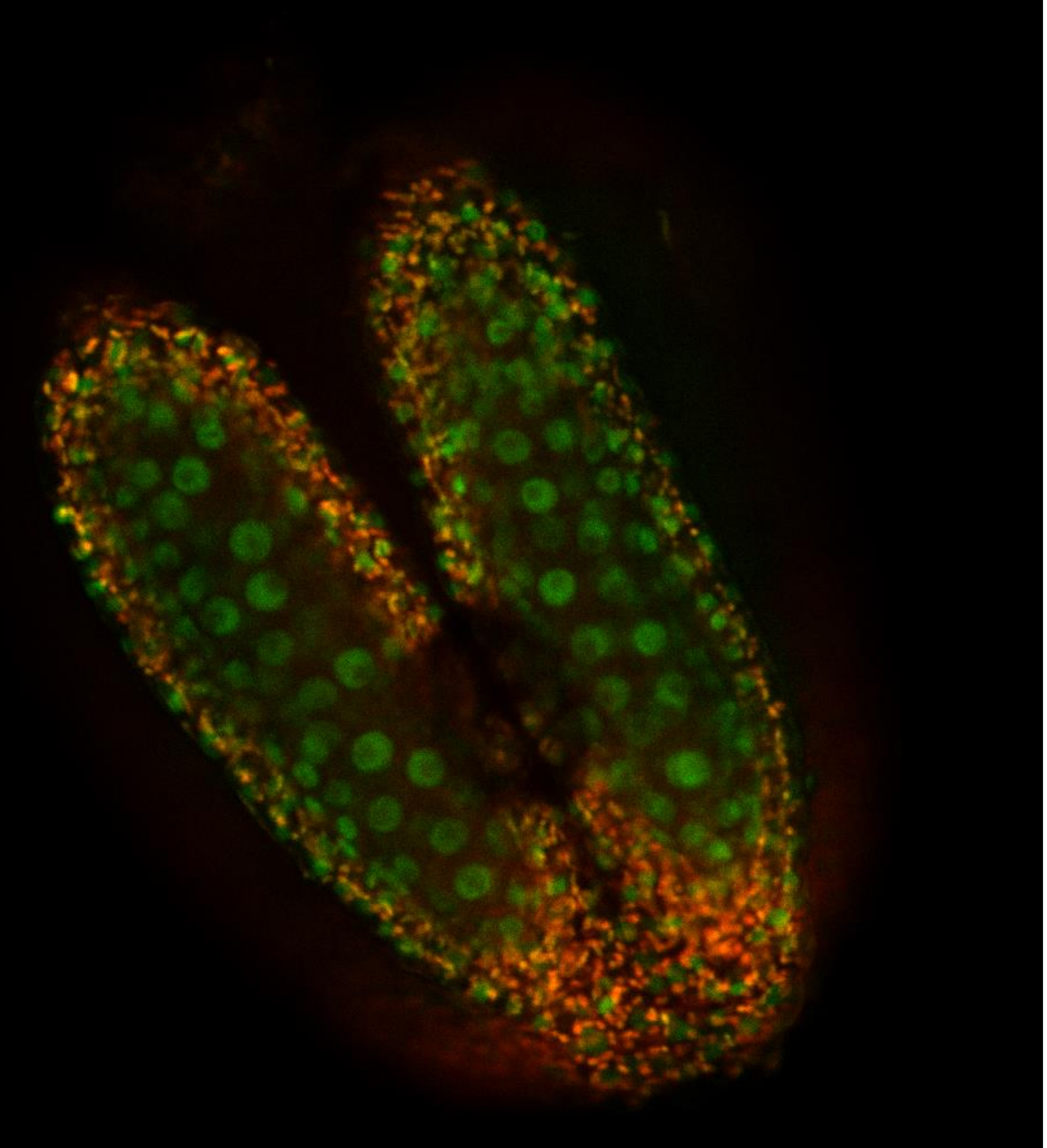
Density plots

Heatmaps

Genome browser snapshots

Often integrates with transcriptomics or ChIP-seq data

Lyons et al. (2023)



Thank you!