Epigenomics Data Analysis Workshop 2023

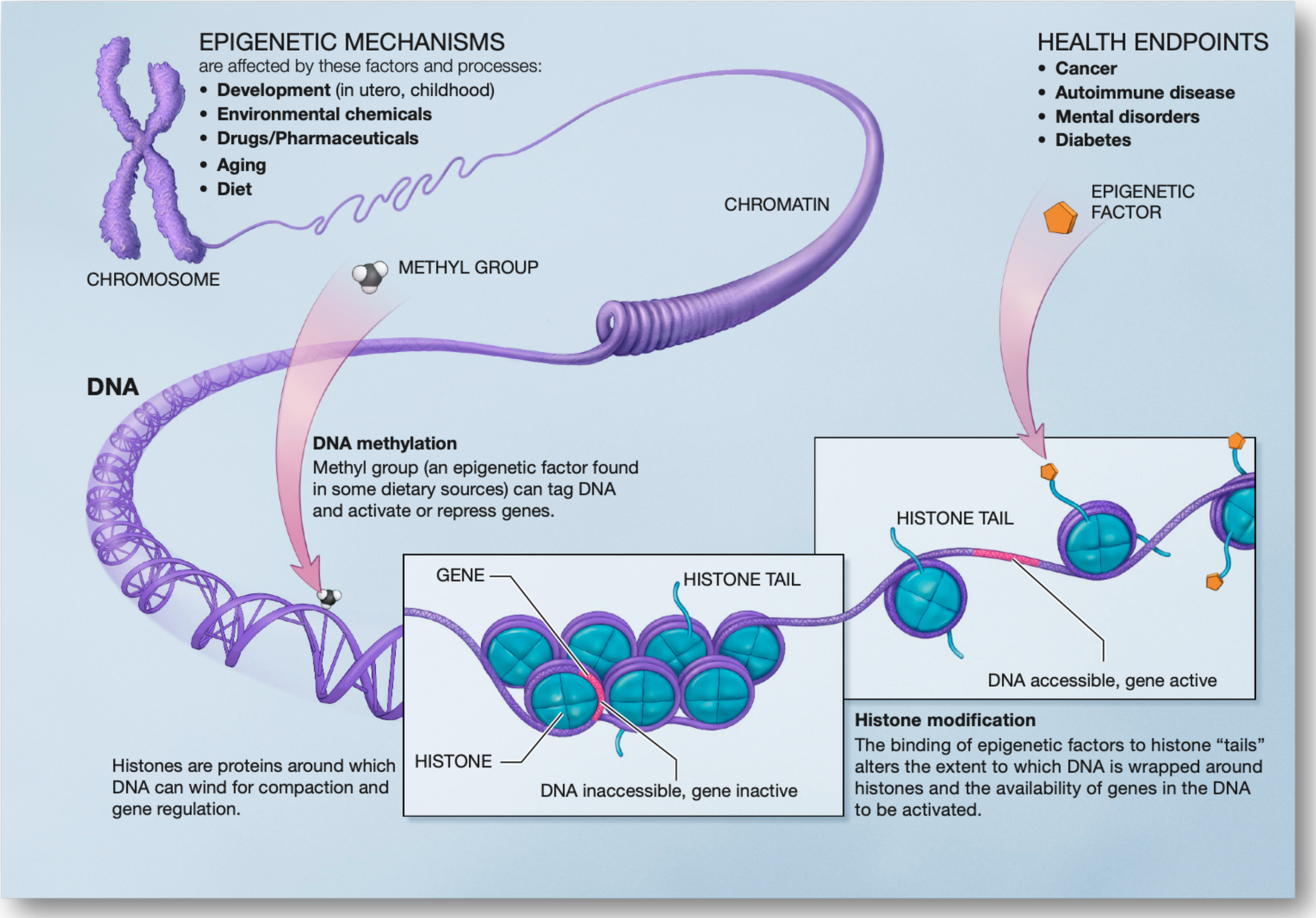
DNA Methylation

Schedule

- 09:30 - 10:15 Short introduction to DNA methylation + Overview Array exercises
- 10:15 - 10:30 UPPMAX set-up + break
- 10:30 - 12:00 Array exercises
- 10:00 - 13:00 Lunch
- 13:00 - 14:00 DNA Methylation: Methods & Technologies
- 14:00 - 14:15 Break
- 14:15 - 14:30 Overview Exercises Bisulfite Sequencing
- 14:30 - 16:30 Bisulfite Sequencing Exercise
- 16:30 - 17:00 Test Yourself

Introduction to DNA methylation

Epigenetics



EPIGENETIC MECHANISMS are affected by these factors and processes:

- **Development** (in utero, childhood)
- **Environmental chemicals**
- **Drugs/Pharmaceuticals**
- **Aging**
- **Diet**

HEALTH ENDPOINTS

- **Cancer**
- **Autoimmune disease**
- **Mental disorders**
- **Diabetes**

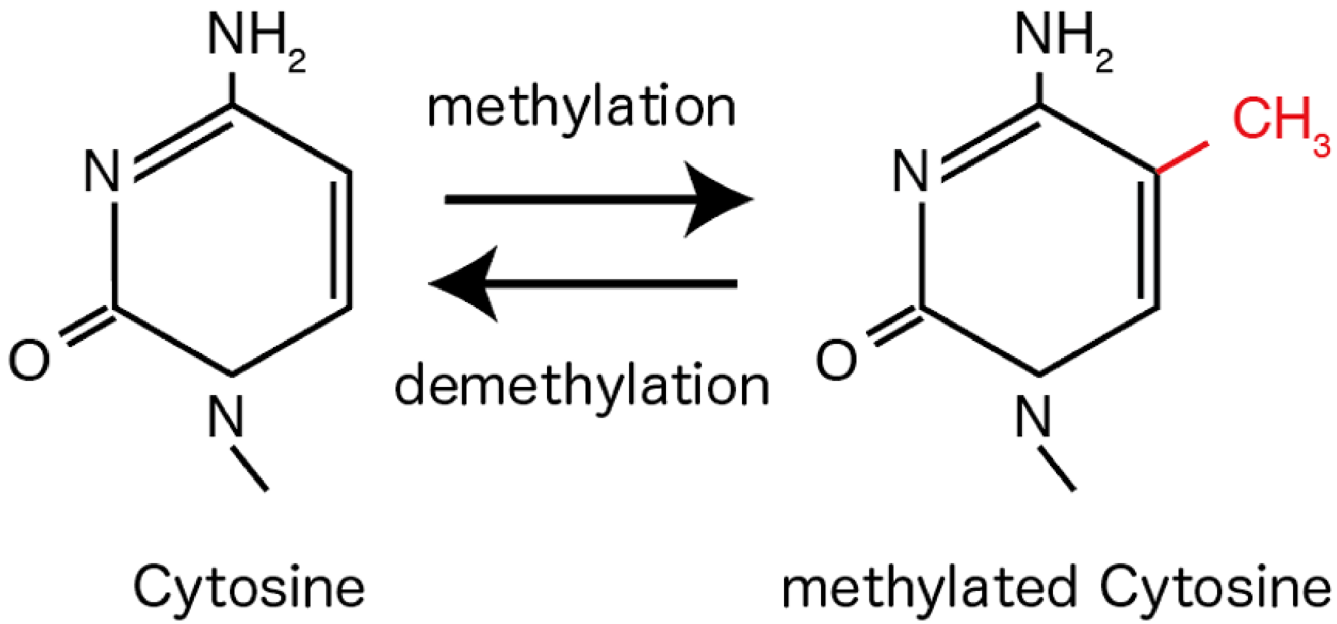
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

Histones are proteins around which DNA can wind for compaction and gene regulation.

Histone modification
The binding of epigenetic factors to histone “tails” alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

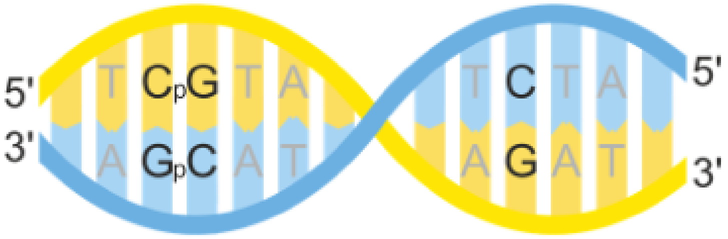
source: NIH

What is DNA methylation?



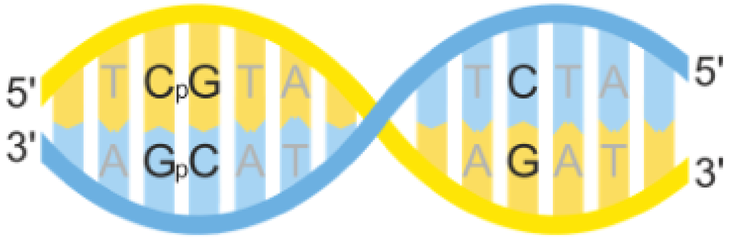
What is DNA methylation?

- Mostly found at cytosines followed by guanines
  - 90% in CpG sites
  - Default state is methylated
  - Prone to mutation -> depleted

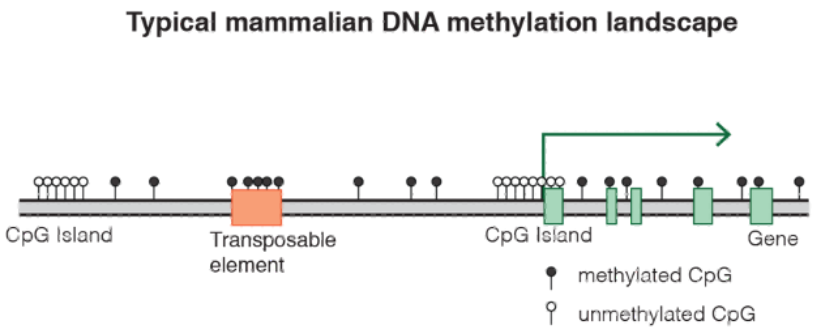


What is DNA methylation?

- Mostly found at cytosines followed by guanines
  - 90% in CpG sites
  - Default state is methylated
  - Prone to mutation -> depleted

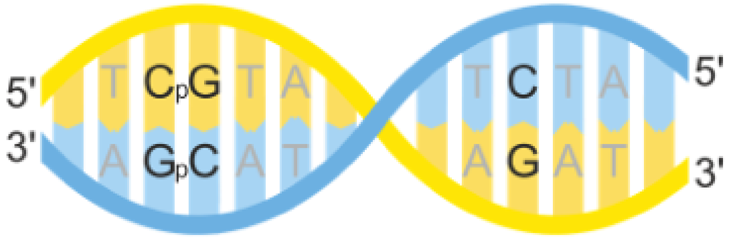


- CpG sites often occurs as clusters: CpG Islands
  - Region with high frequency of CpG
  - Often associated with promoters
  - Unmethylated if gene is expressed

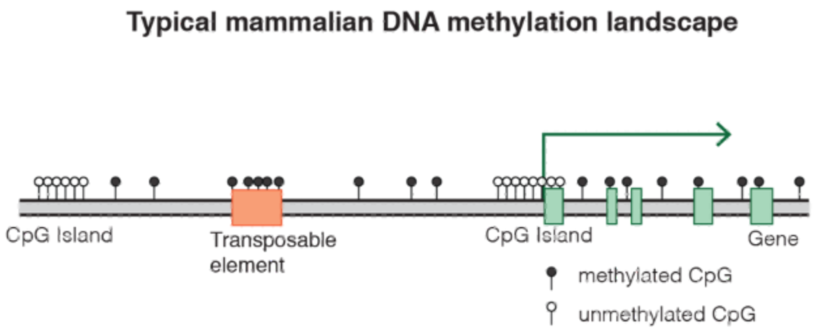


What is DNA methylation?

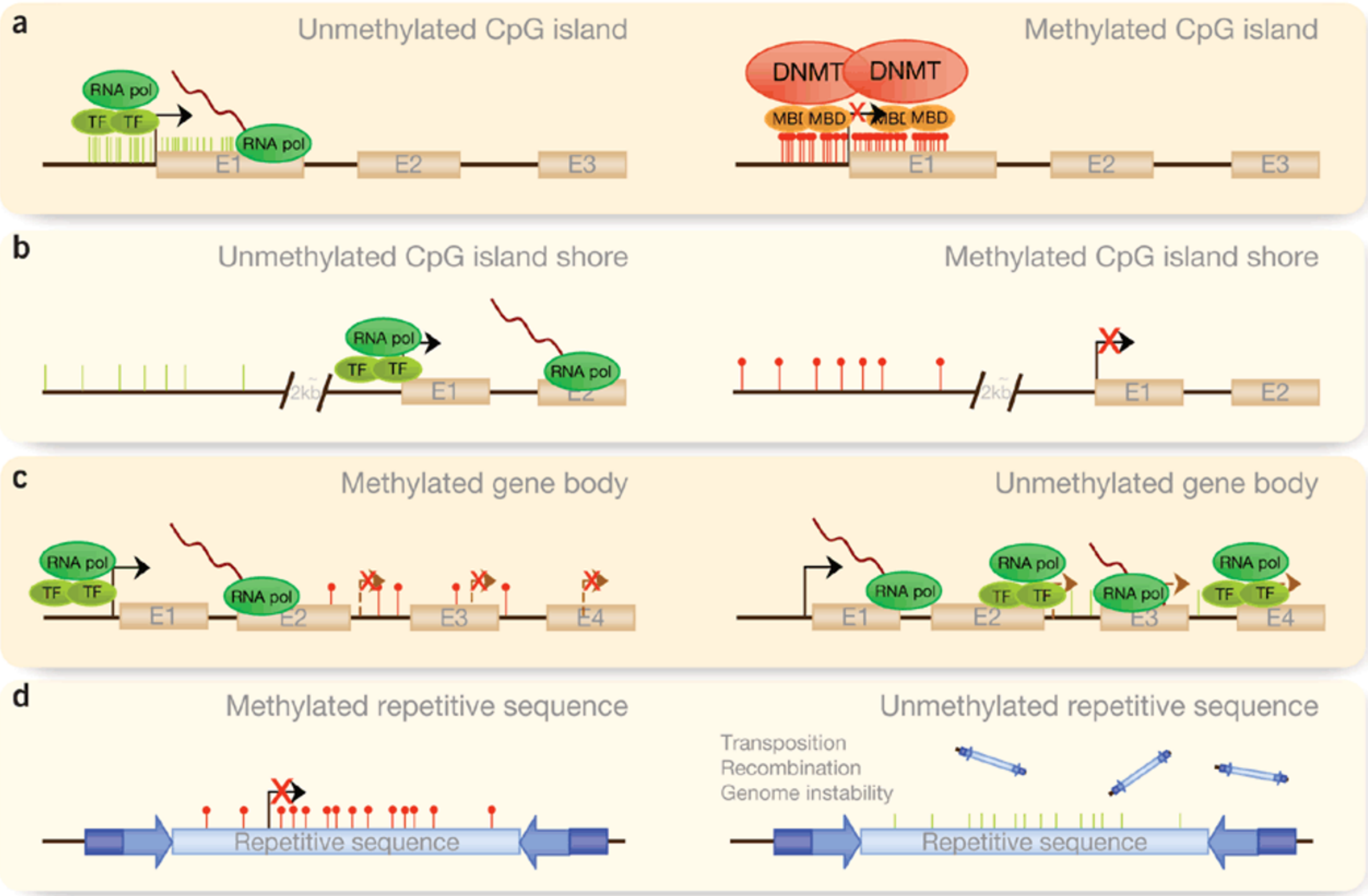
- Mostly found at cytosines followed by guanines
  - 90% in CpG sites
  - Default state is methylated
  - Prone to mutation -> depleted



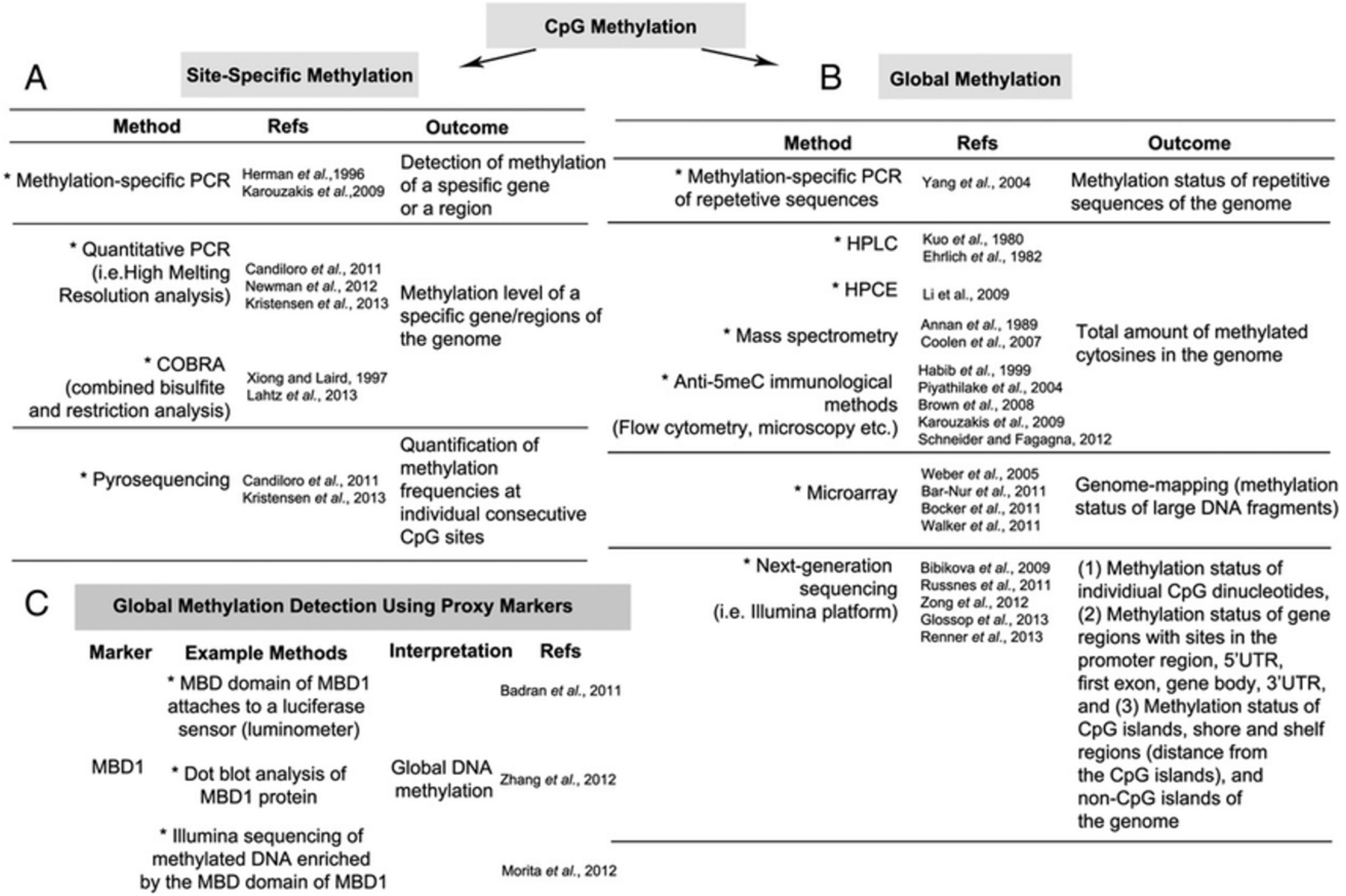
- CpG sites often occurs as clusters: CpG Islands
  - Region with high frequency of CpG
  - Often associated with promoters
  - Unmethylated if gene is expressed
- Role in development, aging, cancer, exercise, …



Effects of Methylation



Detection of DNA methylation



CpG Methylation

A Site-Specific Methylation

Method	Refs	Outcome
* Methylation-specific PCR	Herman *et al.*,1996 Karouzakis *et al.*,2009	Detection of methylation of a spesific gene or a region
* Quantitative PCR (i.e.High Melting Resolution analysis)	Candiloro *et al.*, 2011 Newman *et al.*, 2012 Kristensen *et al.*, 2013	Methylation level of a specific gene/regions of the genome
* COBRA (combined bisulfite and restriction analysis)	Xiong and Laird, 1997 Lahtz *et al.*, 2013	
* Pyrosequencing	Candiloro *et al.*, 2011 Kristensen *et al.*, 2013	Quantification of methylation frequencies at individual consecutive CpG sites

B Global Methylation

Method	Refs	Outcome
* Methylation-specific PCR of repetetive sequences	Yang *et al.*, 2004	Methylation status of repetitive sequences of the genome
* HPLC	Kuo *et al.*, 1980 Ehrlich *et al.*, 1982	Total amount of methylated cytosines in the genome
* HPCE	Li et al., 2009	
* Mass spectrometry	Annan *et al.*, 1989 Coolen *et al.*, 2007	
* Anti-5meC immunological methods (Flow cytometry, microscopy etc.)	Habib *et al.*, 1999 Piyathilake *et al.*, 2004 Brown *et al.*, 2008 Karouzakis *et al.*, 2009 Schneider and Fagagna, 2012	
* Microarray	Weber *et al.*, 2005 Bar-Nur *et al.*, 2011 Bocker *et al.*, 2011 Walker *et al.*, 2011	Genome-mapping (methylation status of large DNA fragments)
* Next-generation sequencing (i.e. Illumina platform)	Bibikova *et al.*, 2009 Russnes *et al.*, 2011 Zong *et al.*, 2012 Glossop *et al.*, 2013 Renner *et al.*, 2013	(1) Methylation status of individiual CpG dinucleotides, (2) Methylation status of gene regions with sites in the promoter region, 5'UTR, first exon, gene body, 3'UTR, and (3) Methylation status of CpG islands, shore and shelf regions (distance from the CpG islands), and non-CpG islands of the genome

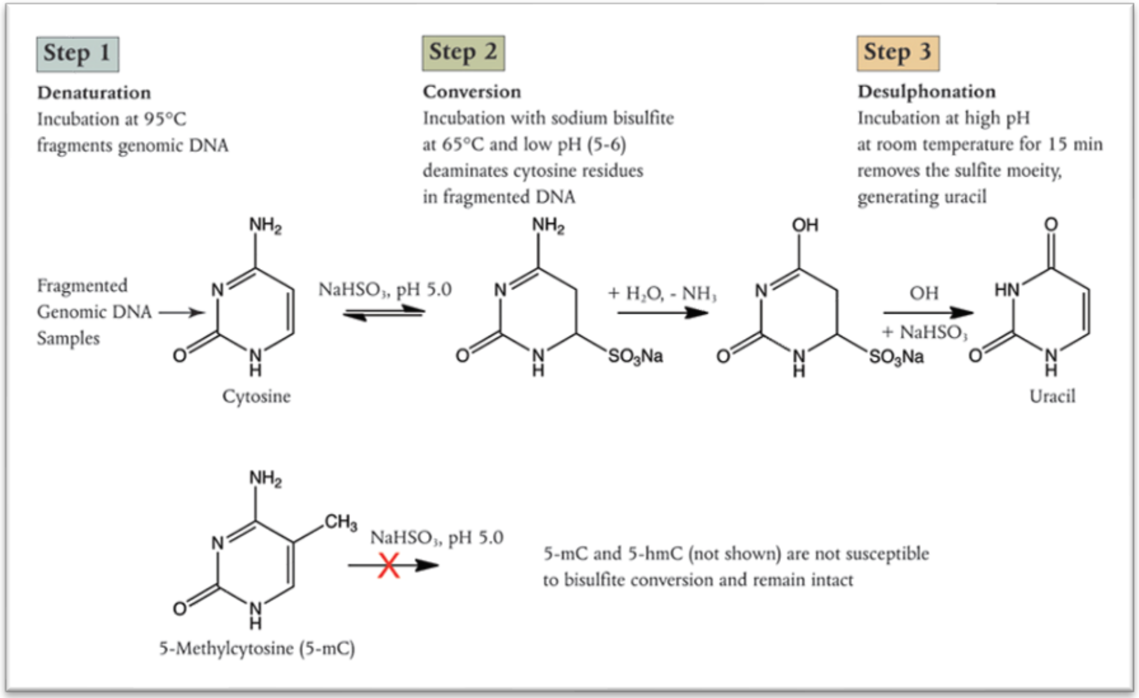
C Global Methylation Detection Using Proxy Markers

Marker	Example Methods	Interpretation	Refs
MBD1	* MBD domain of MBD1 attaches to a luciferase sensor (luminometer)	Global DNA methylation	Badran *et al.*, 2011
	* Dot blot analysis of MBD1 protein		Zhang *et al.*, 2012
	* Illumina sequencing of methylated DNA enriched by the MBD domain of MBD1		Morita *et al.*, 2012

Celik et al. (2014), Journal of Immunological Methods

Bisulfite Conversion

- Bisulfite conversion crucial for both arrays and sequencing
- C -> U (->T)
- mC -> mC (-> C)
- methylation-specific PCR, high resolution melting curve analysis, micro-array based approaches and next generation sequencing

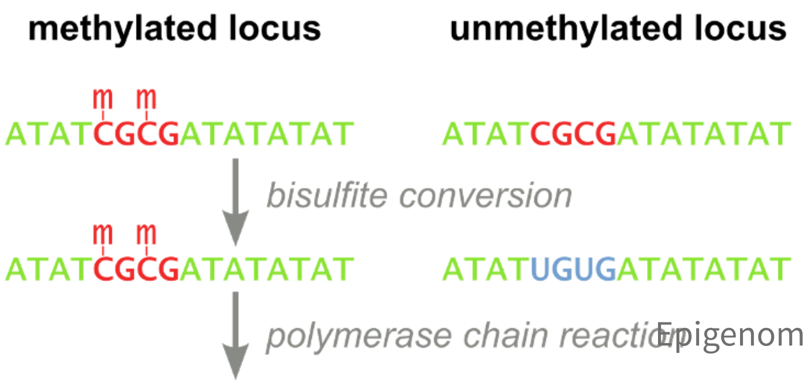


Step 1
Denaturation
Incubation at 95°C fragments genomic DNA

Step 2
Conversion
Incubation with sodium bisulfite at 65°C and low pH (5-6) deaminates cytosine residues in fragmented DNA

Step 3
Desulphonation
Incubation at high pH at room temperature for 15 min removes the sulfite moeity, generating uracil

5-mC and 5-hmC (not shown) are not susceptible to bisulfite conversion and remain intact



methylated locus **unmethylated locus**

ATATCGCGATATATAT ATATCGCGATATATAT

bisulfite conversion

ATATCGCGATATATAT ATATUGUGATATATAT

polymerase chain reacti

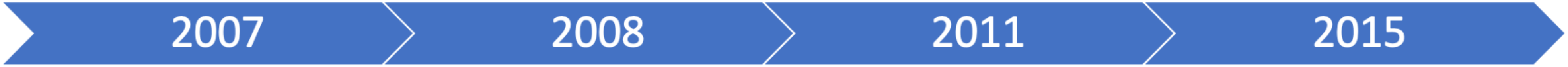
Illumina Methylation Arrays

GoldenGate

1500 CpGs,
cancer focused

Infinium
HumanMethylation450

480K CpGs, 99% RefSeq genes

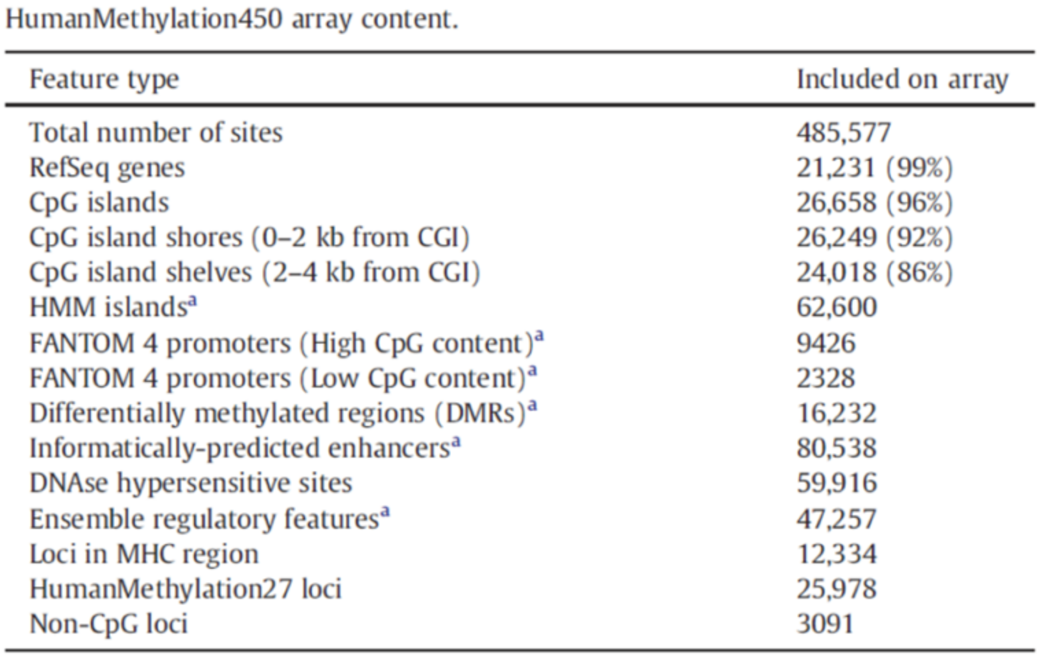


2007

2008

2011

2015



HumanMethylation450 array content.

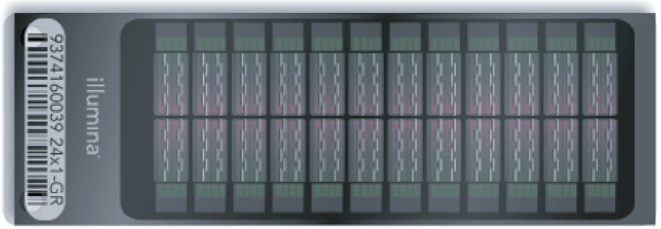
Feature type	Included on array
Total number of sites	485,577
RefSeq genes	21,231 (99%)
CpG islands	26,658 (96%)
CpG island shores (0–2 kb from CGI)	26,249 (92%)
CpG island shelves (2–4 kb from CGI)	24,018 (86%)
HMM islands[a]	62,600
FANTOM 4 promoters (High CpG content)[a]	9426
FANTOM 4 promoters (Low CpG content)[a]	2328
Differentially methylated regions (DMRs)[a]	16,232
Informatically-predicted enhancers[a]	80,538
DNAse hypersensitive sites	59,916
Ensemble regulatory features[a]	47,257
Loci in MHC region	12,334
HumanMethylation27 loci	25,978
Non-CpG loci	3091

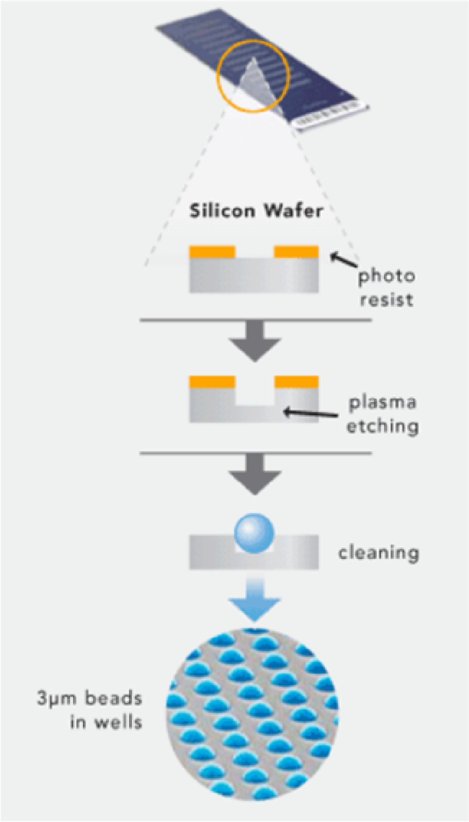
MethylationEPIC

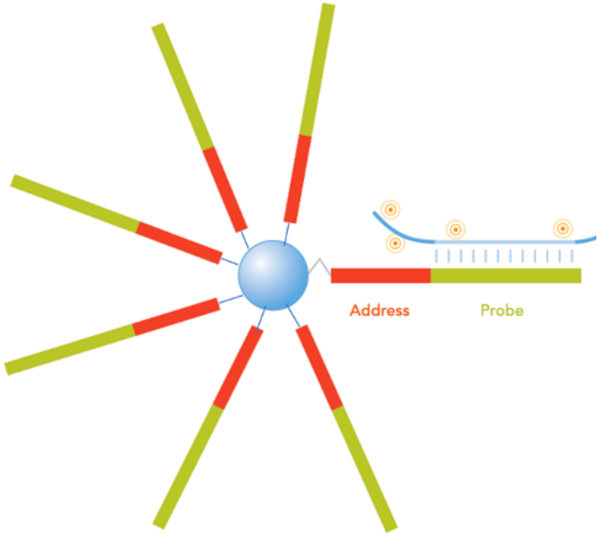
850K CpGs, >90% 450 + additional regulatory regions

450 Array

- 50bp single stranded DNA oligos (“probes”) attached to silica beads
- 2 detection channels: red and green
- Hybrid of 2 different probe designs

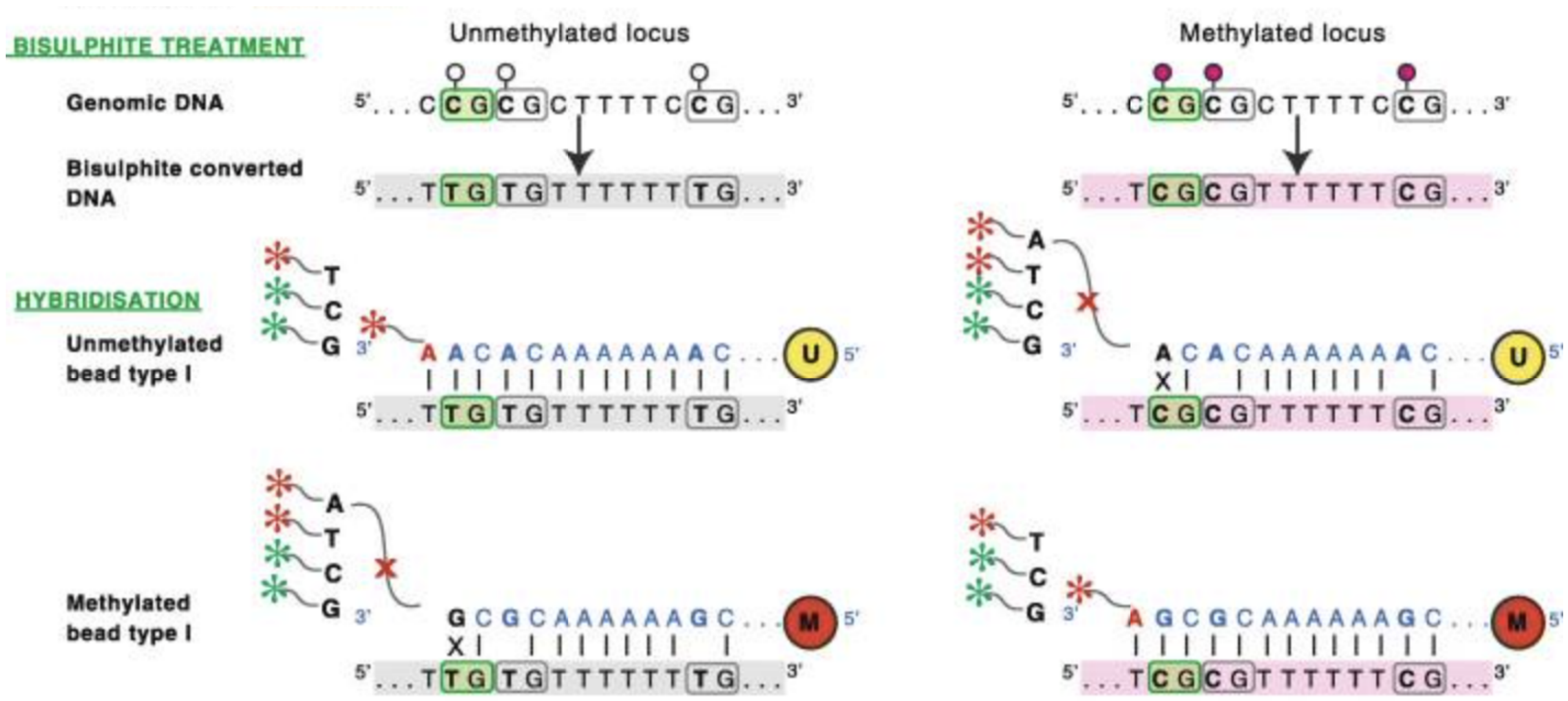






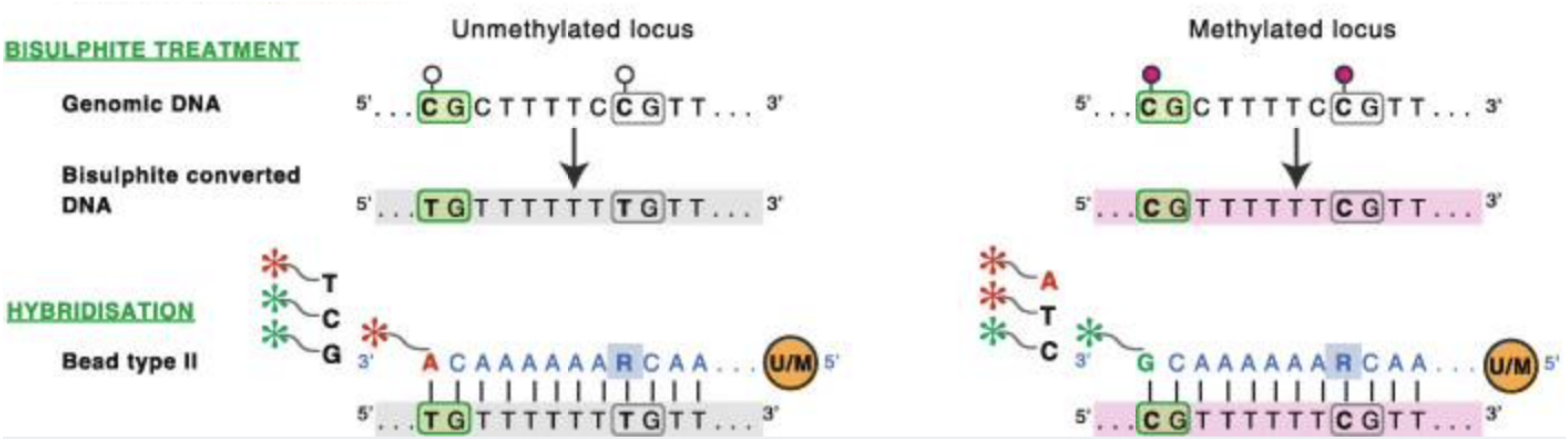
Infinium: Type I vs II design

- Type I: single color detection, two beads



Infinium: Type I vs II design

- Type II: two color detection, single bead



Infinium: Type I vs II design

Type I	Type II
Same chemistry as 27K	New from 450K on
2 beads/CpG	1 bead/CpG (fits more)
Better for CpG dense regions	better for less dense regions
More stable/reproducible	lower dynamic range

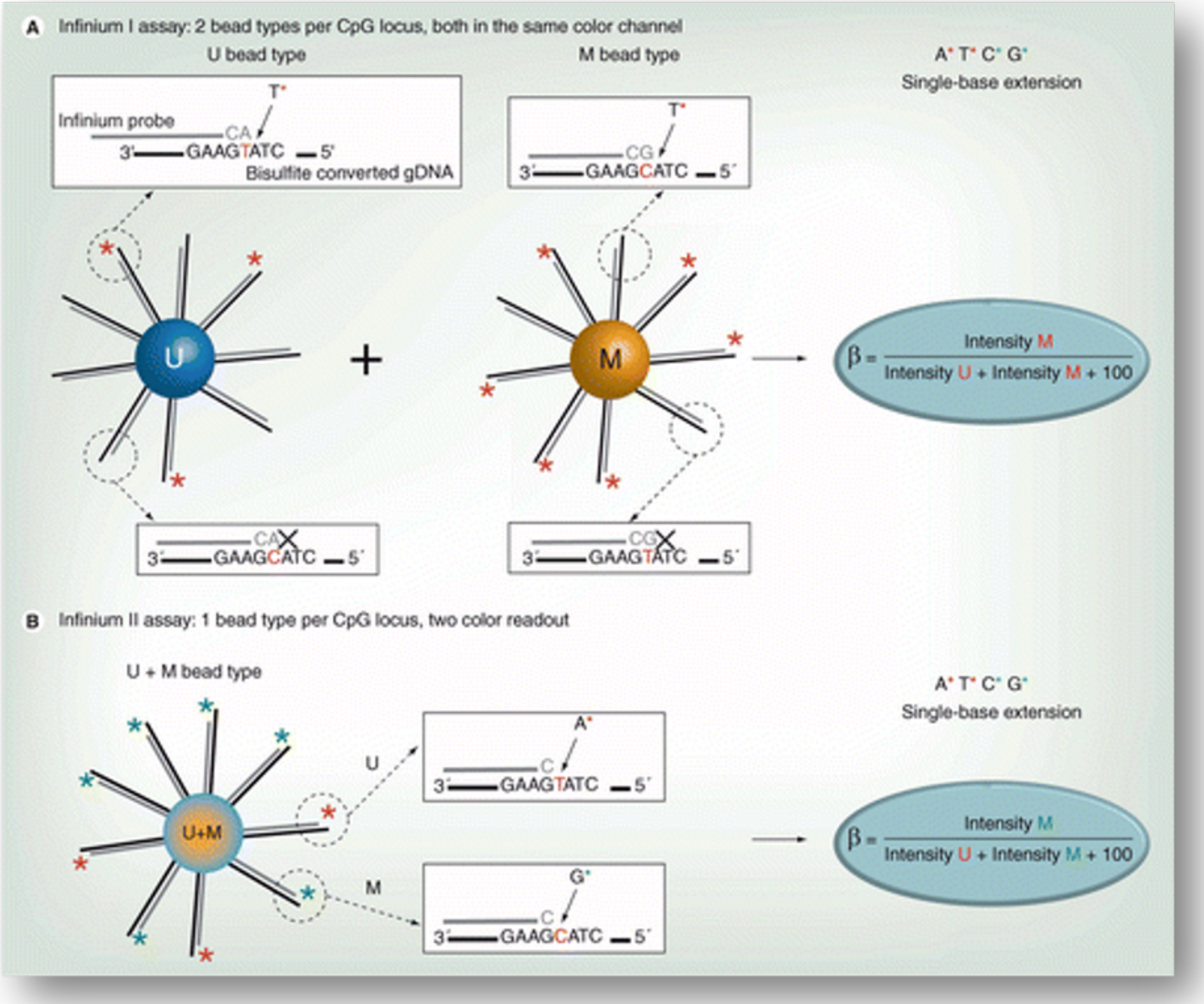
From red/green to methylation level

- Intensities are used to estimate *Beta* values; for both probe designs

$$\text{beta} = M/(M + U + 100)$$

- *Beta* value between 0 and 1
- Easily interpretable, but related M-value has better statistical properties

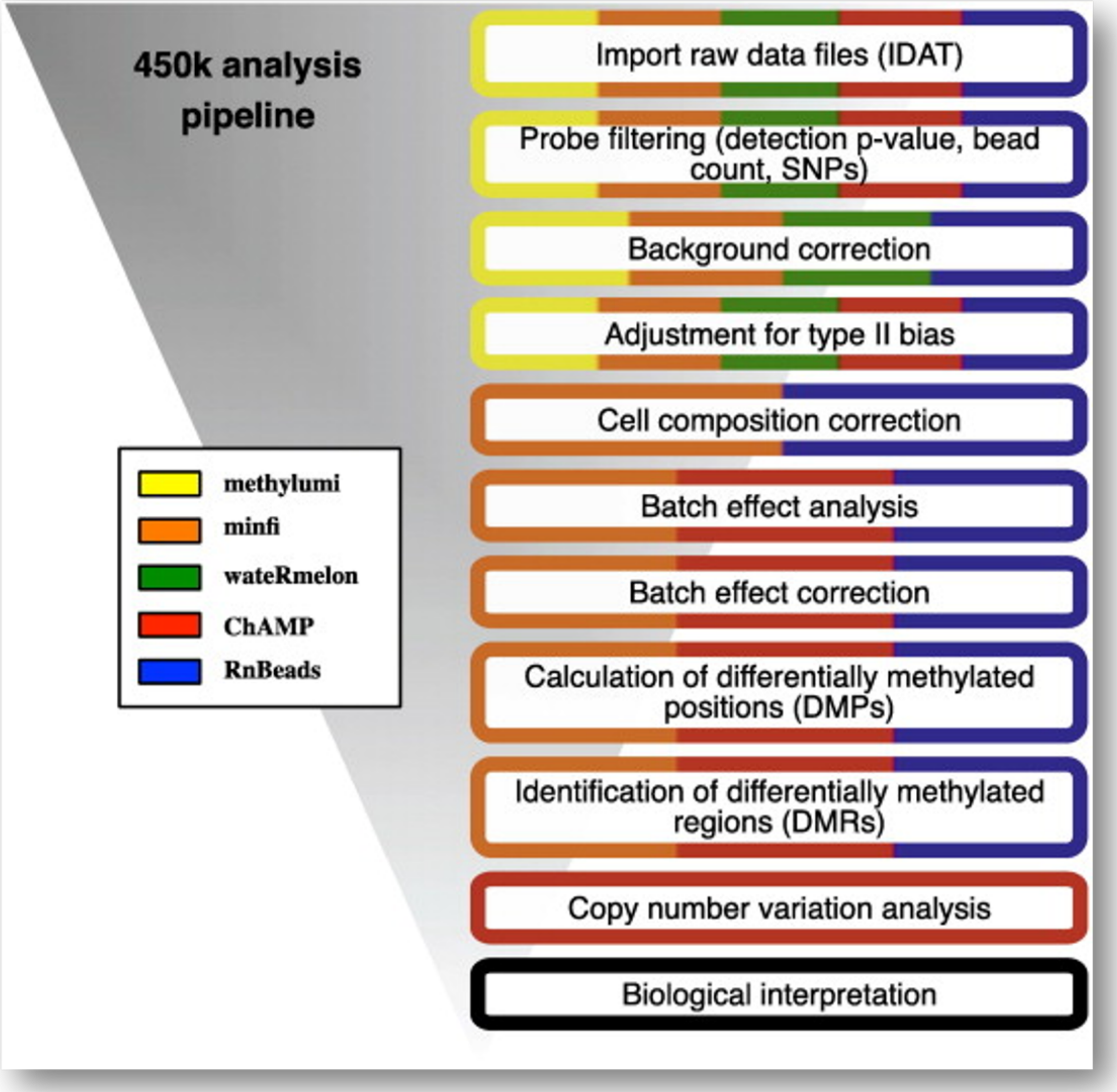
$$\text{Mvalue} = \log 2(M/U)$$



Analysis Workflow

- Typical analysis consists of different steps…
- Many tools for analyzing Illumina arrays
- R package minfi

```r
library(minfi)
```


450k analysis pipeline
Import raw data files (IDAT)
Probe filtering (detection p-value, bead count, SNPs)
Background correction
Adjustment for type II bias
Cell composition correction
Batch effect analysis
Batch effect correction
Calculation of differentially methylated positions (DMPs)
Identification of differentially methylated regions (DMRs)
Copy number variation analysis
Biological interpretation
🟨 methylumi
🟧 minfi
🟩 wateRmelon
🟥 ChAMP
🟦 RnBeads

Import data

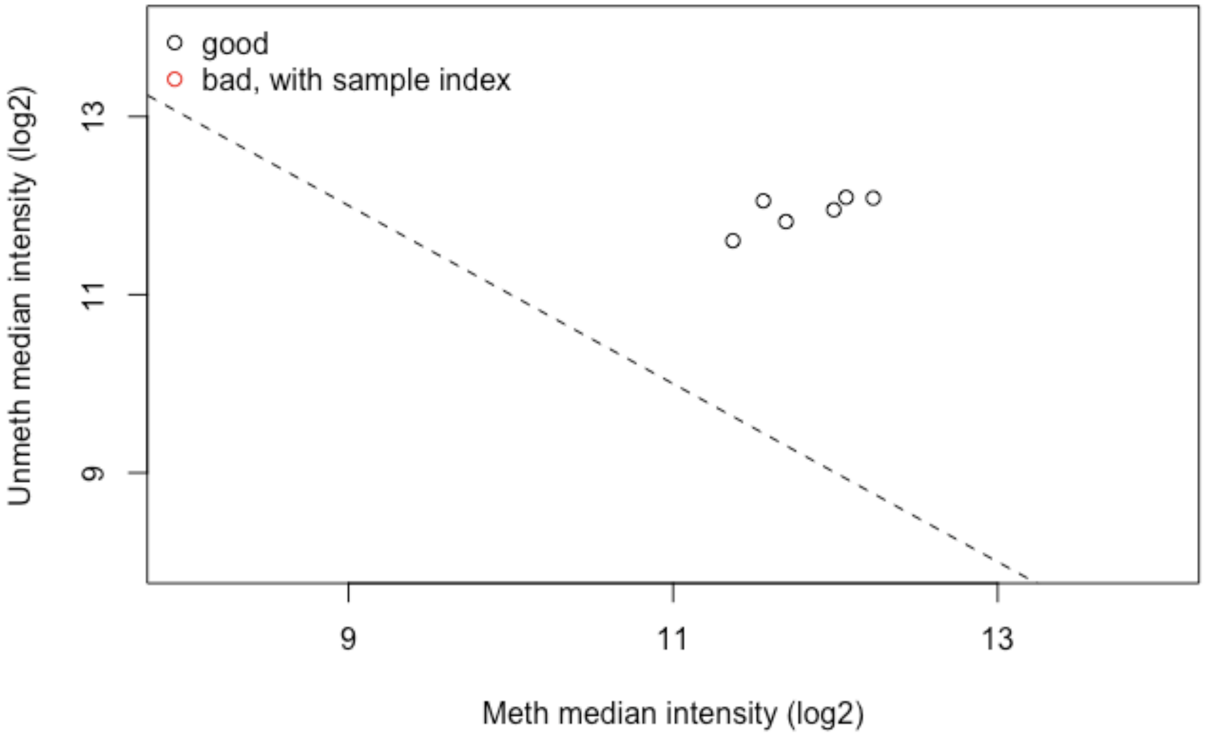
- IDAT files; slide scanner output
  - 5859594006_R01C01_Grn.idat

```r
dataDirectory <- "/sw/courses/epigenomics/DNAmethylation/array_data/"
# read in the sample sheet for the experiment
targets <- read.metharray.sheet(dataDirectory, pattern="SampleSheet.csv")
# read in the raw data from the IDAT files
rgSet <- read.metharray.exp(targets=targets)
# Go from intensity data to methylation levels
MSet <- preprocessRaw(rgSet)
```

Initial Quality Control

- Plot median intensity in M vs U

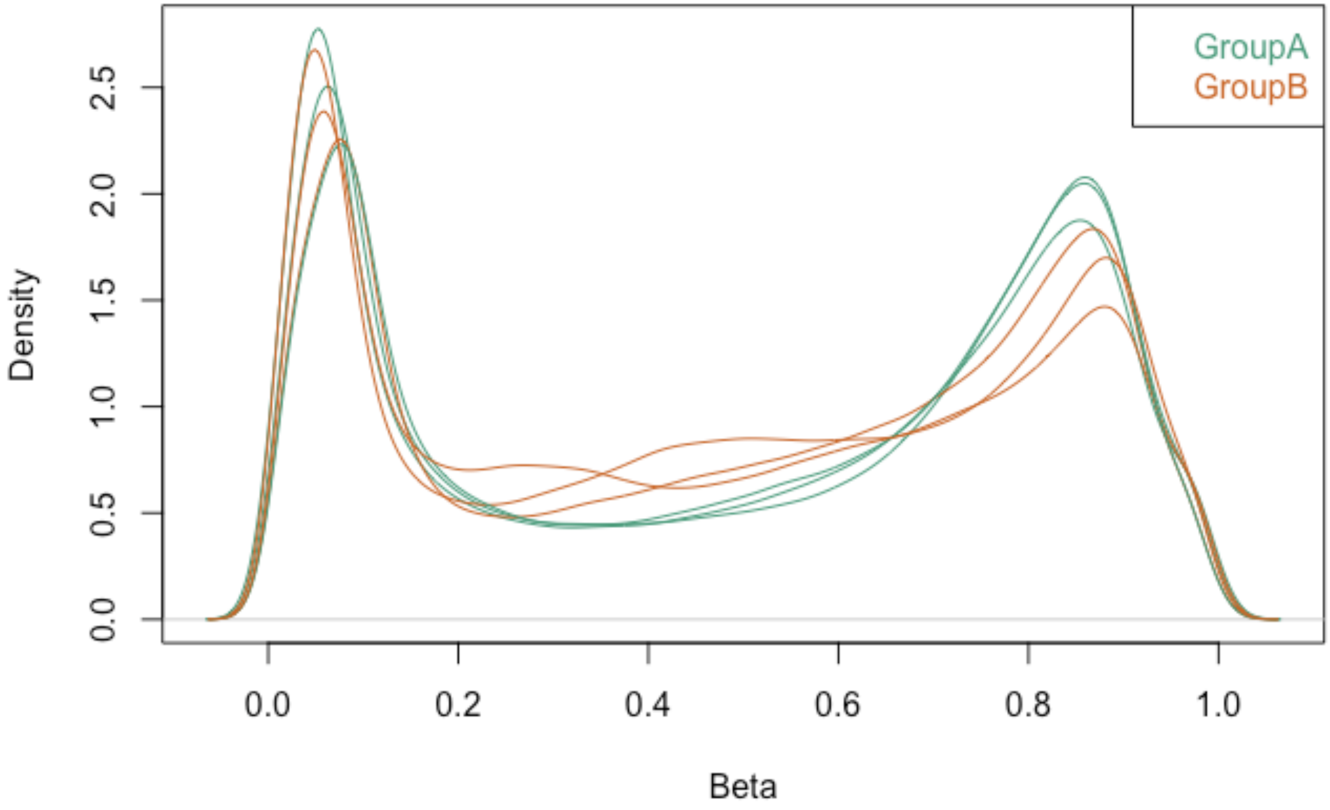
```r
qc <- getQC(MSet)
plotQC(qc)
```



Initial Quality Control

- *Beta* value density distribution

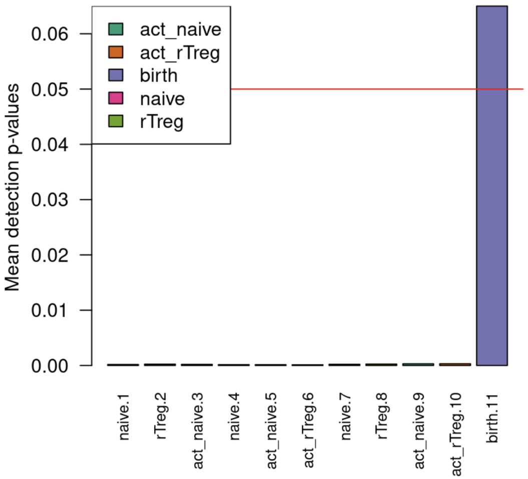
```r
densityPlot(MSet, sampGroups = phenoData$Sample_Group)
```



Initial Quality Control

- Detection p-value: Are the intensities significantly above background?

```
# Calculate the detection p-values
detP <- detectionP(rgSet)
# examine mean detection p-values across all samples to identify any failed
barplot(colMeans(detP), las=2, cex.names=0.8, ylab="Mean detection p-values
abline(h=0.05,col="red")
```

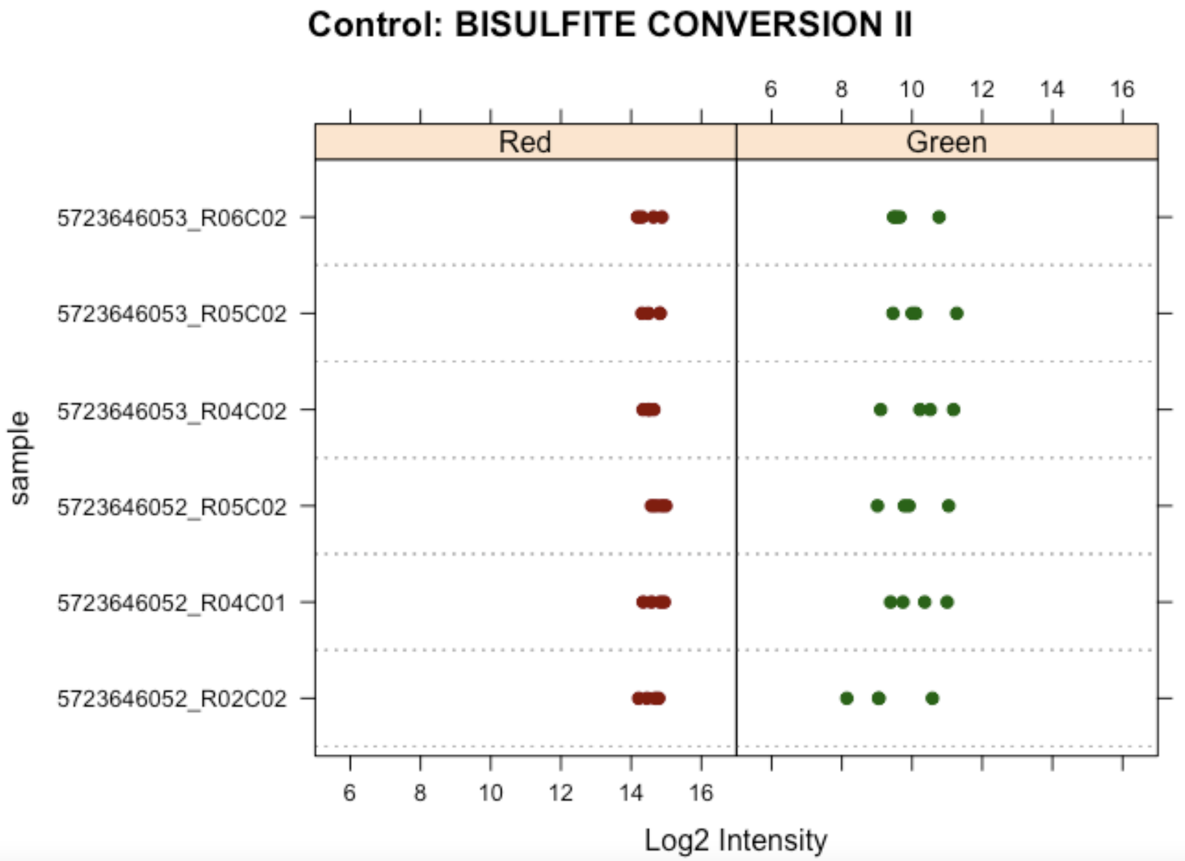


Potentially remove bad samples and/or probes.

Initial Quality Control

- Several internal control probes for different sample preparation steps (bisulfite conversion, hybridization, …)

```r
controlStripPlot(RGSet, controls="BISULFITE CONVERSION II")
```



- Staining control
- Bisulfite conversion
- extension controls
- specificity controls
- hybridization controls
- target removal controls
- negative controls

Description in Illumina manual

Other considerations…

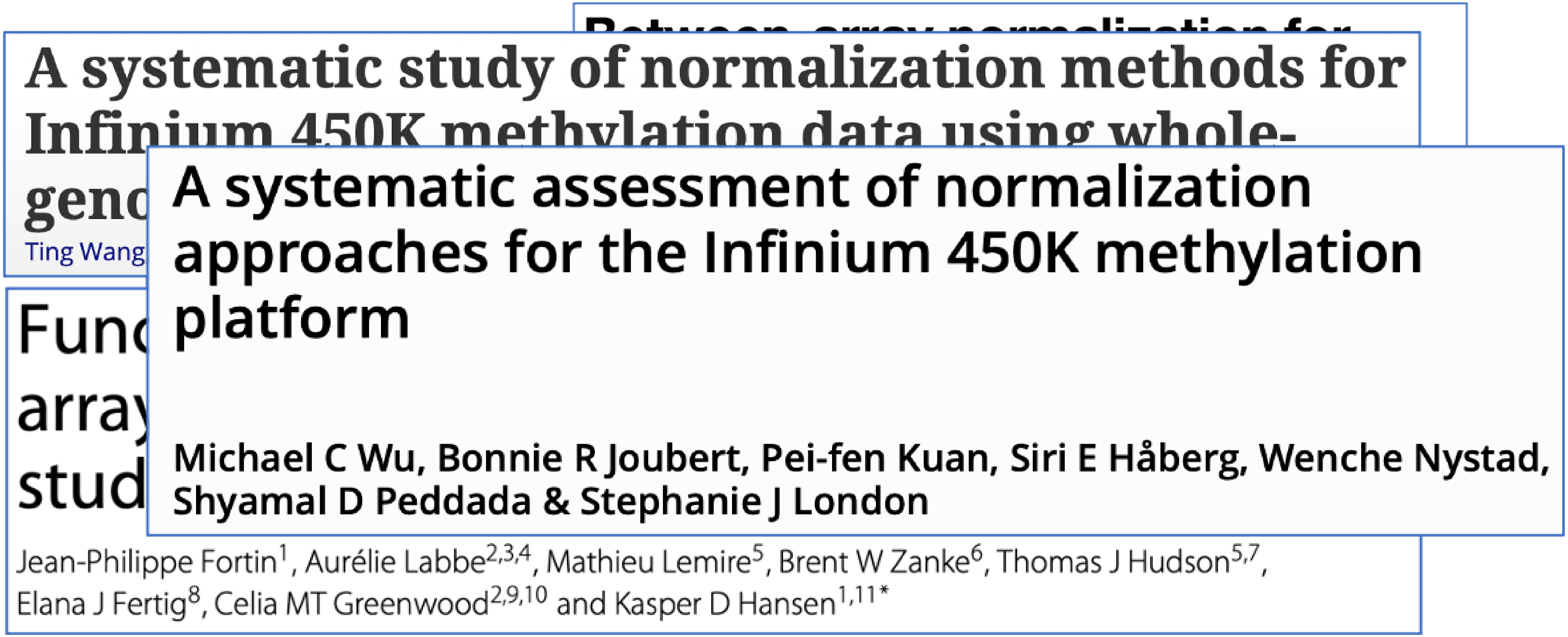
- Remove X/Y Chromosome CpGs?
- Remove CpG overlapping with known SNP and/or cross reactive probes
- Check sample structure with PCA

Many of the previous plots can be looked at interactively with shinyMethyl.

paper: A comprehensive overview of Infinium HumanMethylation450 data processing.

Normalization

- Within and across array normalization



Between-array normalization for

A systematic study of normalization methods for Infinium 450K methylation data using whole-geno

Ting Wang

Func array stud

Jean-Philippe Fortin[1], Aurélie Labbe[2,3,4], Mathieu Lemire[5], Brent W Zanke[6], Thomas J Hudson[5,7], Elana J Fertig[8], Celia MT Greenwood[2,9,10] and Kasper D Hansen[1,11*]

A systematic assessment of normalization approaches for the Infinium 450K methylation platform

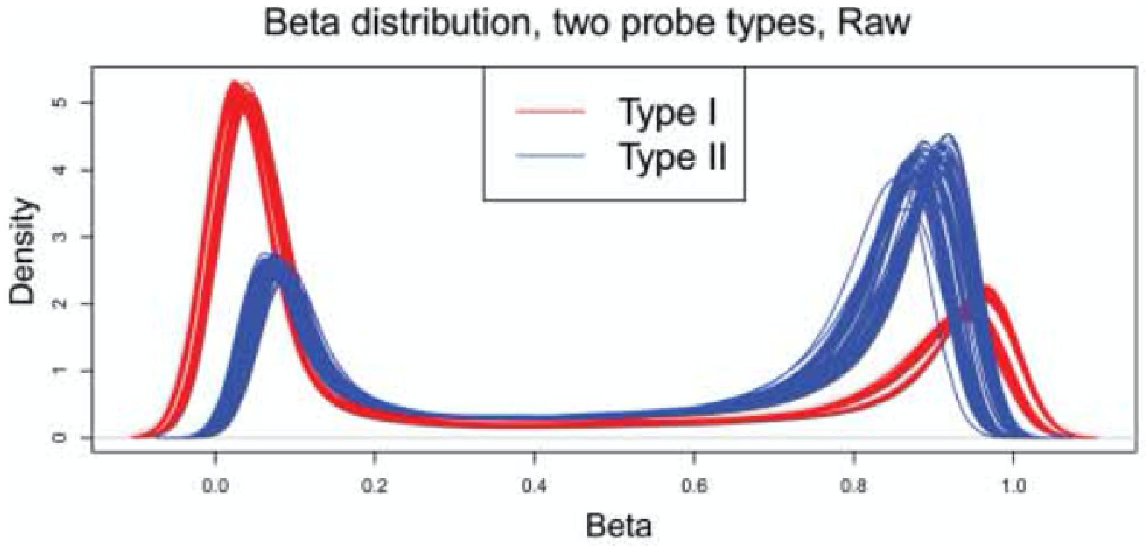
Michael C Wu, Bonnie R Joubert, Pei-fen Kuan, Siri E Håberg, Wenche Nystad, Shyamal D Peddada & Stephanie J London

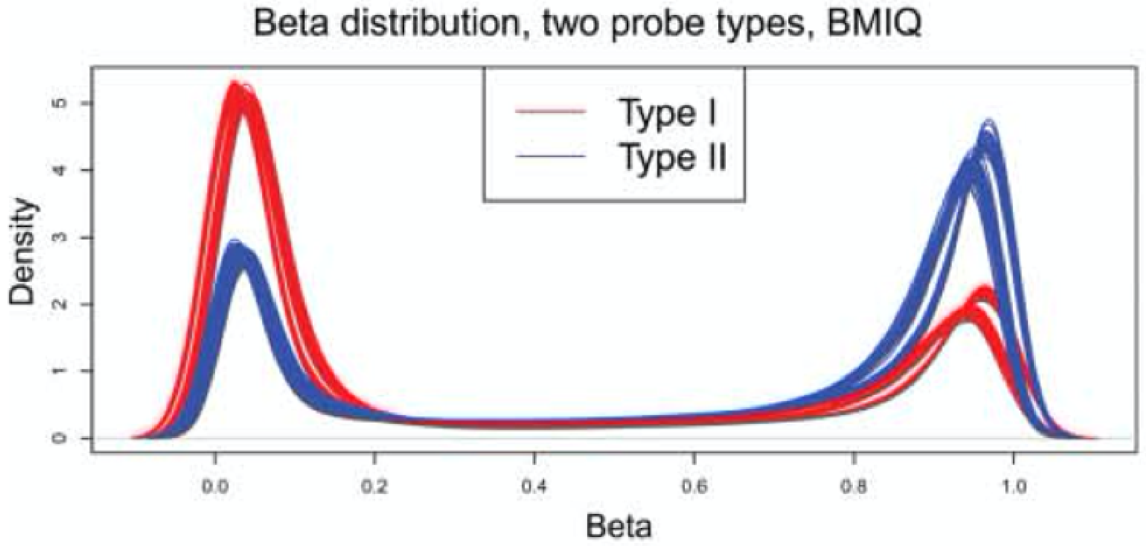
Normalization

- Within and across array normalization
- Within array:
    - background correction
    - dye bias adjustment
    - Type I/II bias correction
- Between array:
    - starting material
    - labeling efficiency
- Good overview + described in lab
- An evaluation of processing methods for HumanMethylation450 BeadChip data

Assess normalization case by case

- Within and across array normalization not always necessary
- Depends on biological signal


Beta distribution, two probe types, Raw


Beta distribution, two probe types, BMIQ

Differential Methylation

- Identification of systematic differences in methylation between groups of samples (case vs control, smokers vs non-smokers, …)

- Usually starts on a per CpG basis

- Many ways to approach this
  - Questions being asked of data, available information on potential confounders, nature/structure of the data (repeated measures, …)

- Some possible approaches
  - T-test and ANOVA models
  - Wilcoxon rank-sum and Kruskall Wallis test
  - Linear, logistic and Cox regression or mixed effect models

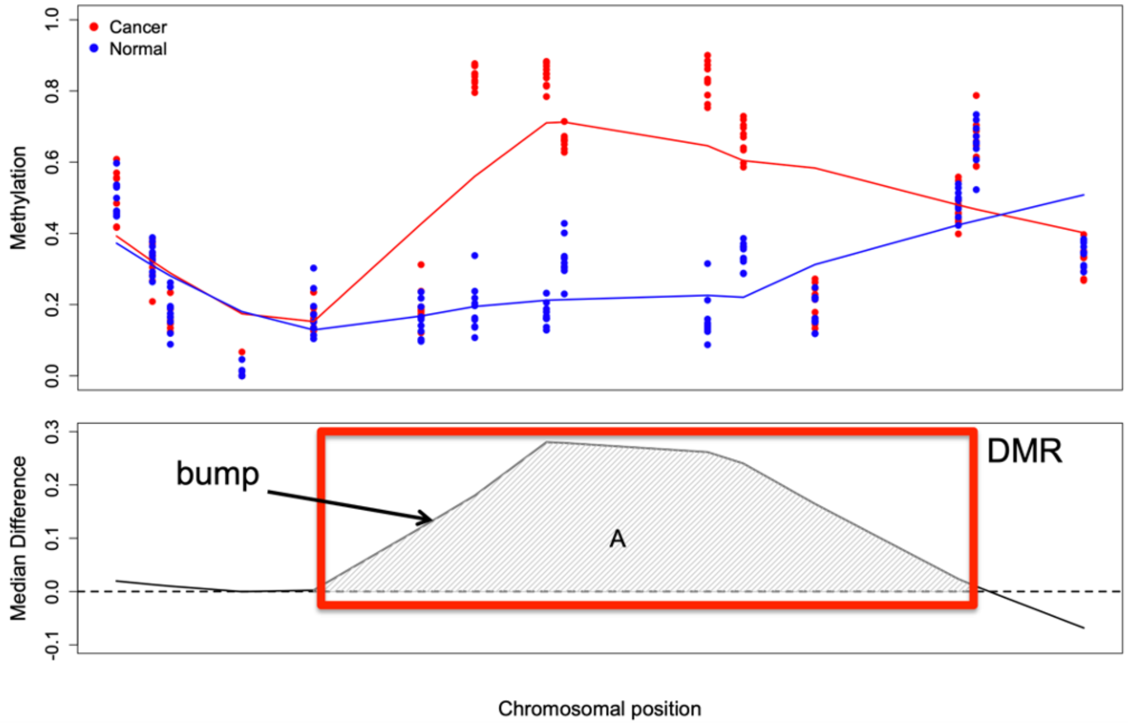
- Use M-values: M = log2(M/U) and *Beta* minimal difference cutoff

Differential Methylation

- Single CpG often less informative than region (DMR)

- How to define region?
  - Sliding window
  - Heuristic cutoff
  - Functional units

- We will try last two in the lab

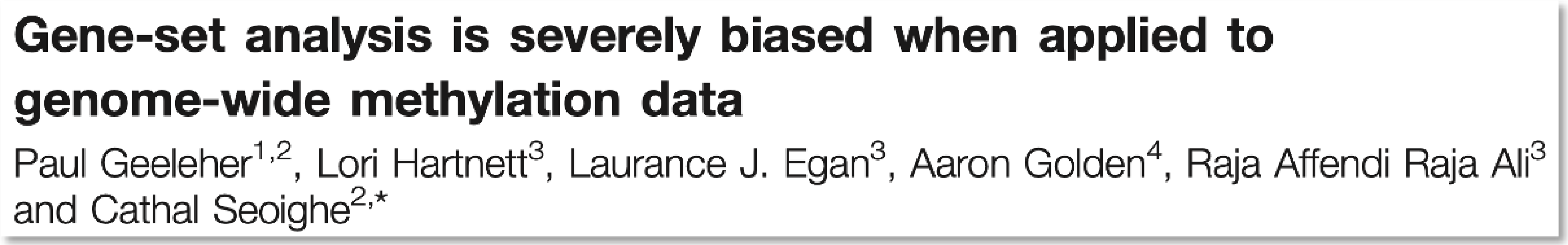


Gene Set Enrichment

- Long list of DMP or DMR…. What does it mean?

- Gene expression -> GO analysis

- Not so straightforward for methylation data!
  - CpG link to genes unclear
  - Directionality
  - Bias! Number of CpG per gene differs



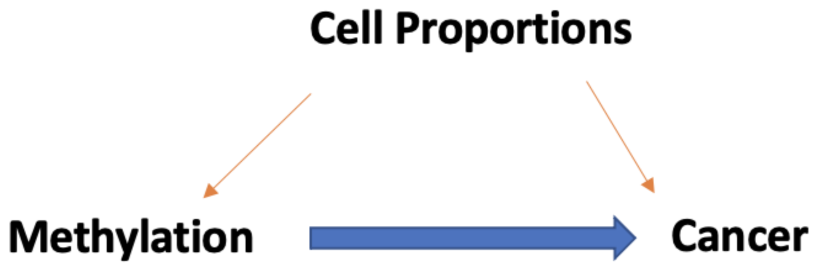
Gene-set analysis is severely biased when applied to genome-wide methylation data

Paul Geeleher[1,2], Lori Hartnett[3], Laurance J. Egan[3], Aaron Golden[4], Raja Affendi Raja Ali[3] and Cathal Seoighe[2,*]

- missMethyl, methylGSA, BioMethyl

Cell Type Deconvolution

- Estimates the relative proportion of pure cell types within a sample
- Most cohort studies use data from blood samples



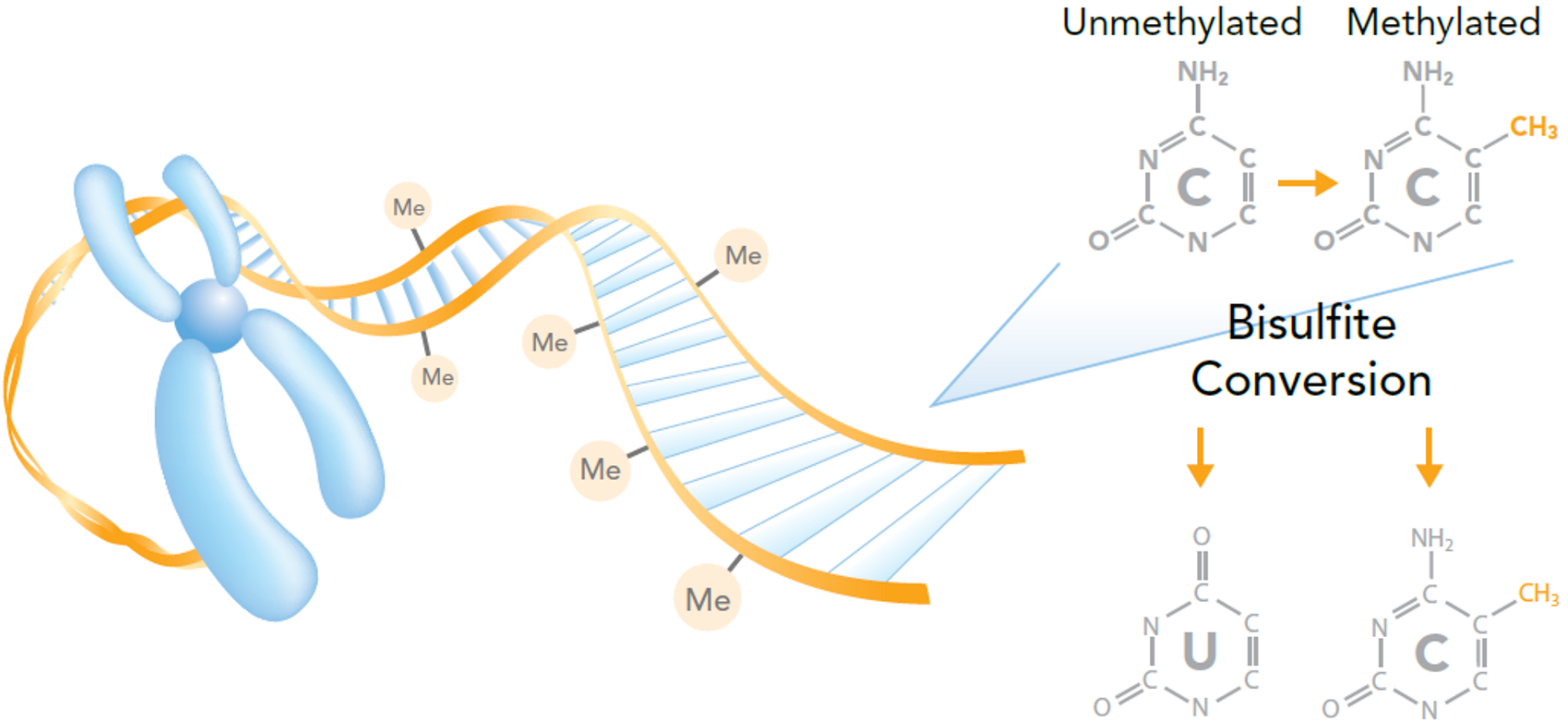
- Minfi: RGChannelSet returns relative proportions of CD4+ and CD8+ T-cells, NK cells, monocytes, granulocytes and B-cells in each sample

Datasets

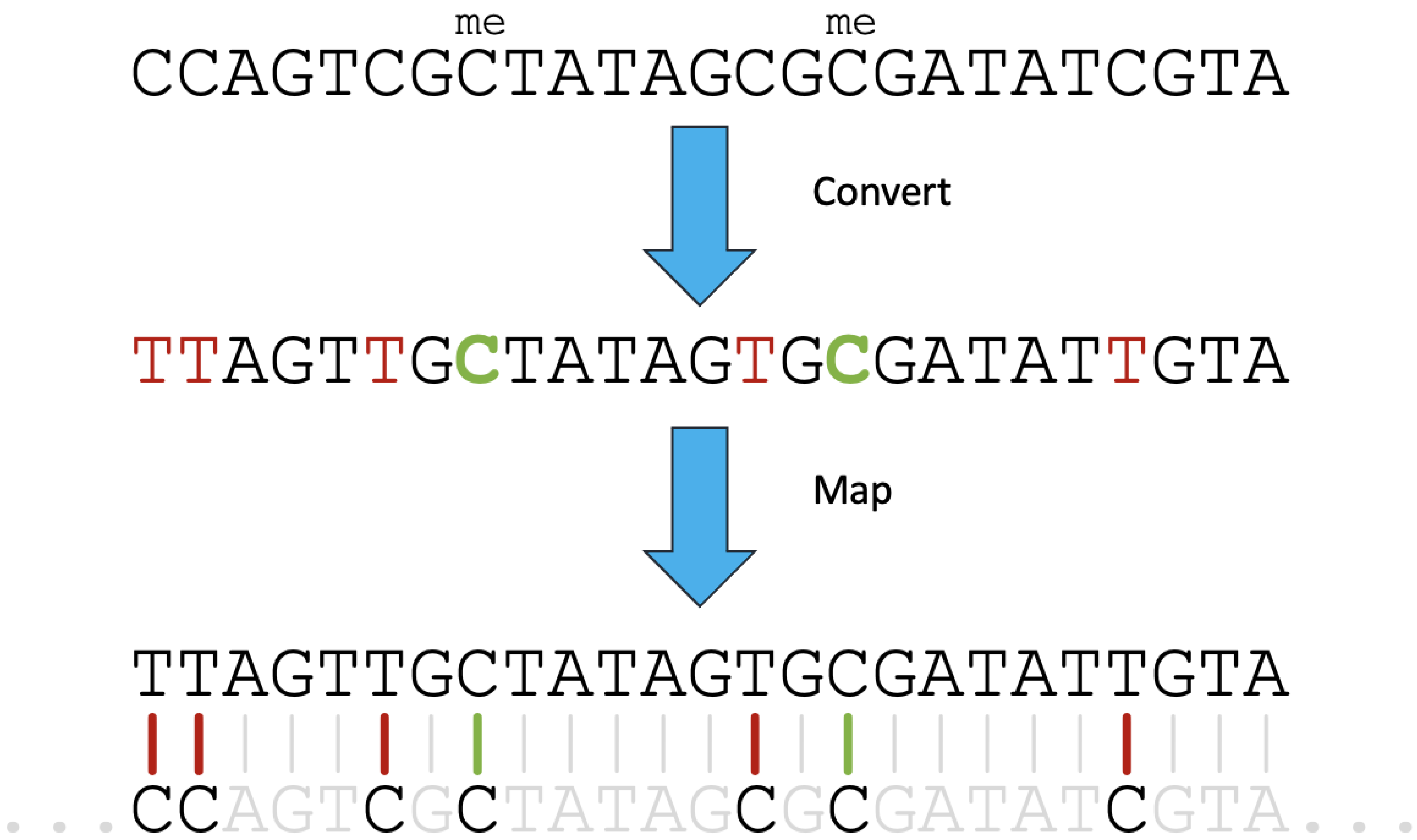
- Small toy data
- IDAT files
- 10 samples: 4 different T-cell types from 3 individuals
  - Naive
  - Treg
  - act_naive
  - act_Treg
- An additional sample has been added from another study GSE51180, to illustrate approaches for identifying poor quality samples.

Bisulfite Sequencing

Bisulfite Sequencing

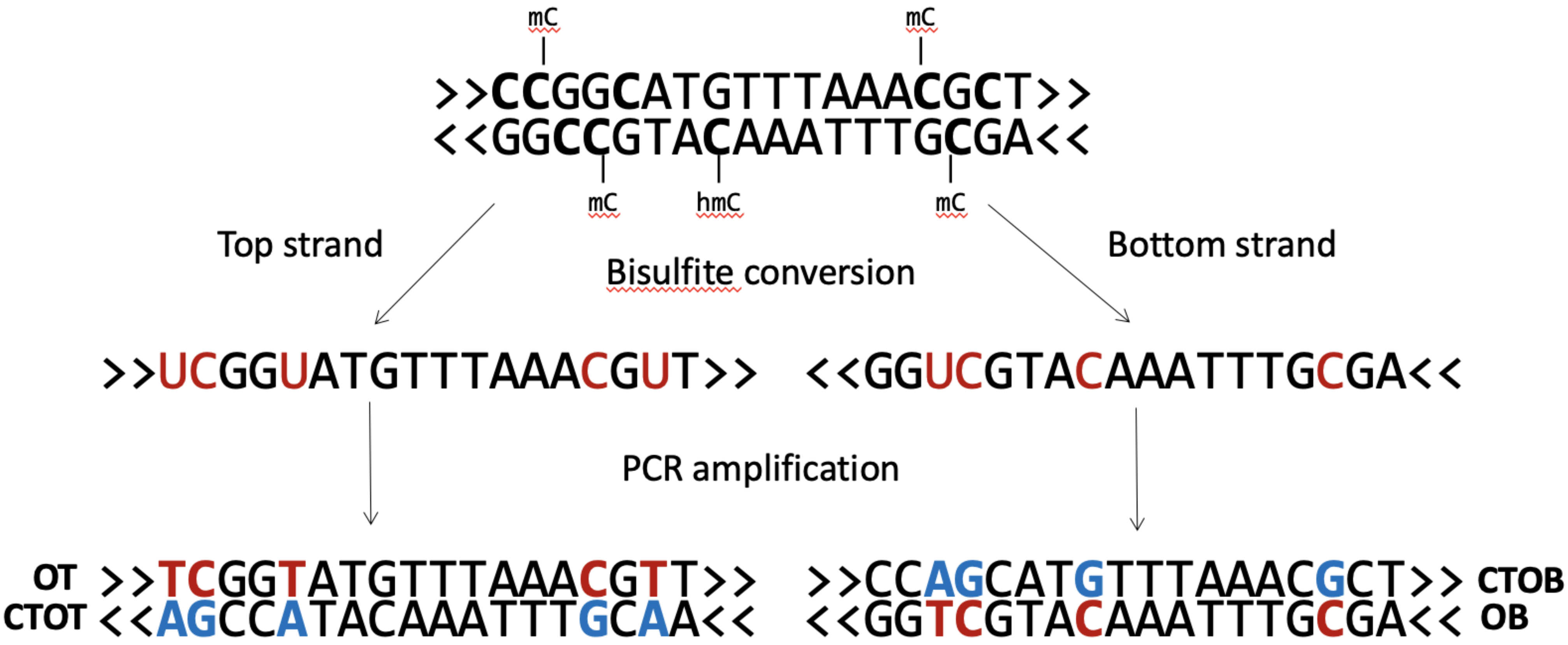


Easy readout… in theory



me
me
CCAGTCGCTATAGCGCGATATCGTA
Convert
TTAGTTGCTATAGTGCGATATTGTA
Map
TTAGTTGCTATAGTGCGATATTGTA
...CCAGTCGCTATAGCGCGATATCGTA...

… but not in reality



mC mC

>>CCGGCATGTTTAAACGCT>>

<<GGCCGTACAAATTTGCGA<<

mC hmC mC

Top strand

Bisulfite conversion

Bottom strand

>>UCGGUATGTTTAAACGUT>>

<<GGUCGTACAAATTTGCGA<<

PCR amplification

OT >>TCGGTATGTTTAAACGTT>>

CTOT <<AGCCATACAAATTTGCAA<<

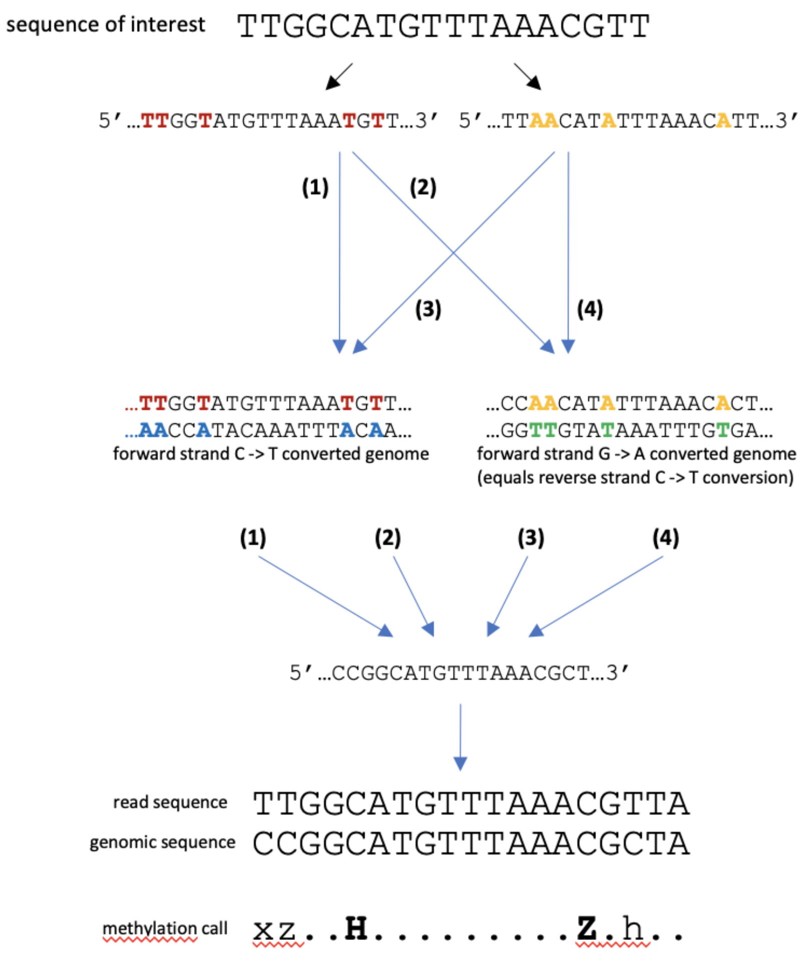
>>CCAGCATGTTTAAACGCT>> CTOB

<<GGTCGTACAAATTTGCGA<< OB

- 2 different PCR product and 4 possible different sequence strands from one genomic locus
- Each of these 4 can exist in any possible conversion state

3-letter alignment

Bismark



sequence of interest TTGGCATGTTTAAACGTT

5′ …TTGGTATGTTTAAATGTT…3′ 5′ …TTAACATATTTAAACATT…3′

…TTGGTATGTTTAAATGTT…
…AACCATACAAATTTACAA…
forward strand C -> T converted genome

…CCAACATATTTAAACACT…
…GGTTGTATAAATTTGTGA…
forward strand G -> A converted genome
(equals reverse strand C -> T conversion)

5′ …CCGGCATGTTTAAACGCT…3′

read sequence TTGGCATGTTTAAACGTTA
genomic sequence CCGGCATGTTTAAACGCTA

methylation call xz..H.........Z.h..

bisulfite convert read (treat sequence as both forward and reverse strand)

align to bisulfite converted genomes

read all 4 alignment outputs and extract the unmodified genomic sequence if the sequence could be mapped uniquely

methylation call

h unmethylated C in CHH context

H methylated C in CHH context

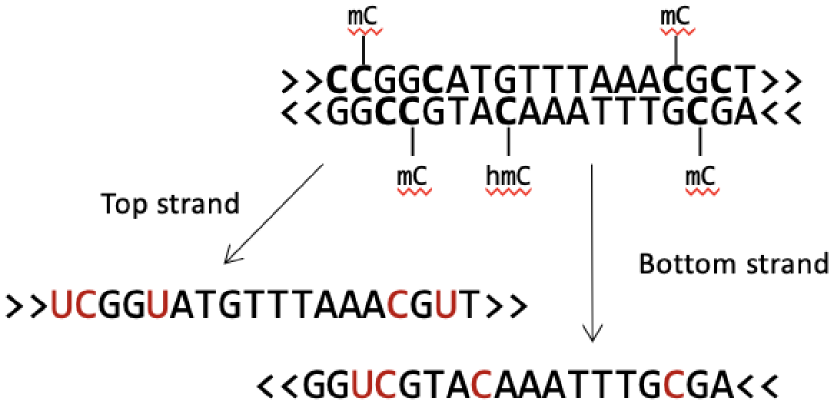
x unmethylated C **in CHG context**

X methylated C **in CHG context**

z unmethylated C in CpG context

Z methylated C in CpG context

Common library preparations



mC mC

>>CCGGCATGTTTAAACGCT>>
<<GGCCGTACAAATTTGCGA<<

mC hmC mC

Top strand

Bottom strand

>>UCGGUATGTTTAAACGUT>>

<<GGUCGTACAAATTTGCGA<<

1) Directional libraries
(vast majority of kits, also EpiGnome/Truseq)

```
OT  >>TCGGTATGTTTAAACGTT>>
    <<GGTCGTACAAATTTGCGA<<  OB
```

2) PBAT libraries

```
CTOT <<AGCCATACAAATTTGCAA<<
     >>CCAGCATGTTTAAACGCT>>  CTOB
```

3) Non-directional libraries
(e.g. single-cell BS-Seq, Zymo Pico Methyl-Seq)

```
  OT  >>TCGGTATGTTTAAACGTT>>
CTOT  <<AGCCATACAAATTTGCAA<<
      >>CCAGCATGTTTAAACGCT>>  CTOB
      <<GGTCGTACAAATTTGCGA<<  OB
```

Quality Control is essential

- Accurate C >T detection
- Pre-alignment
    - Base quality/composition
    - Duplication levels
    - Adapter removal

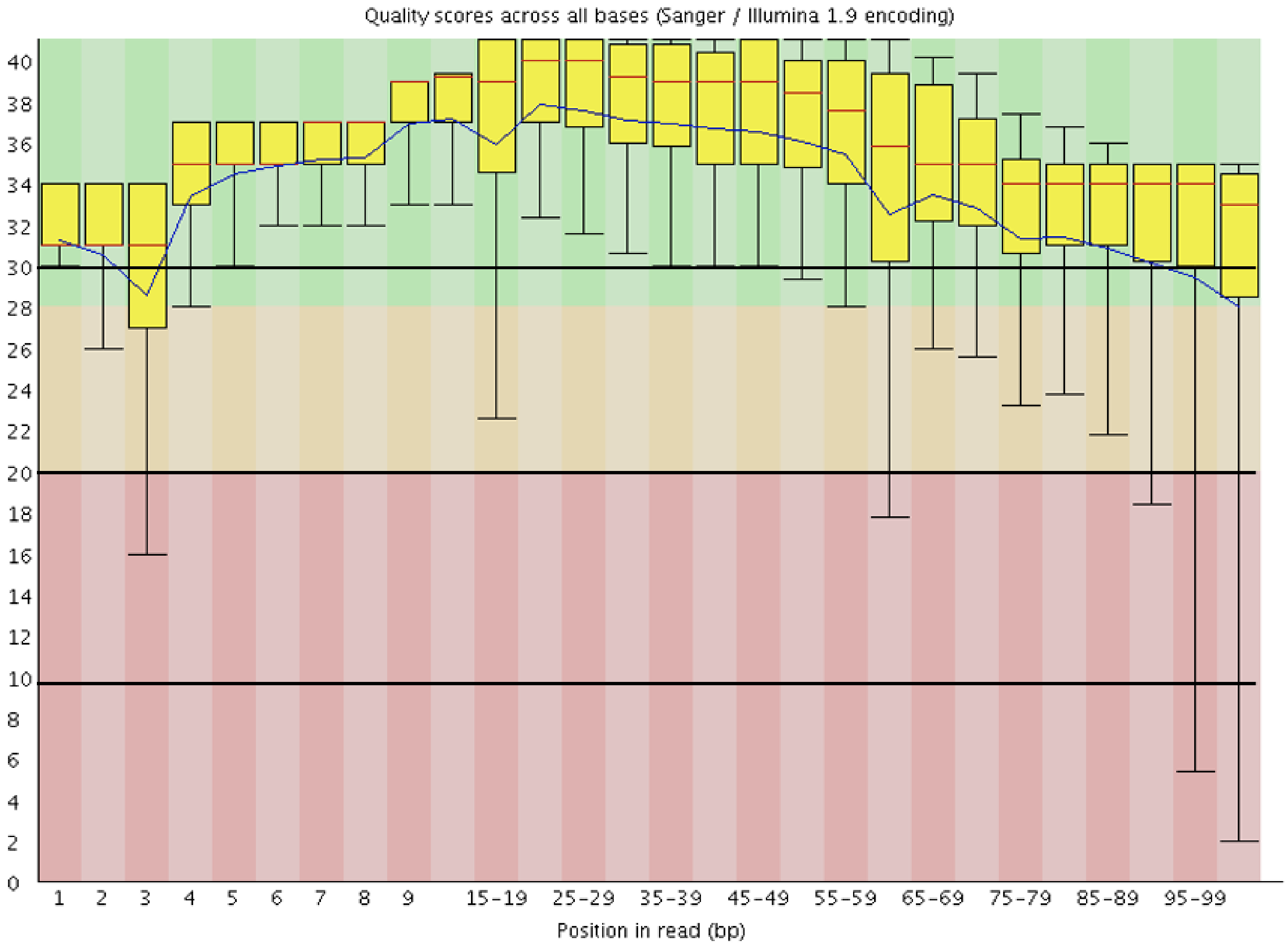
Average Base Quality

Error rate

0.1%

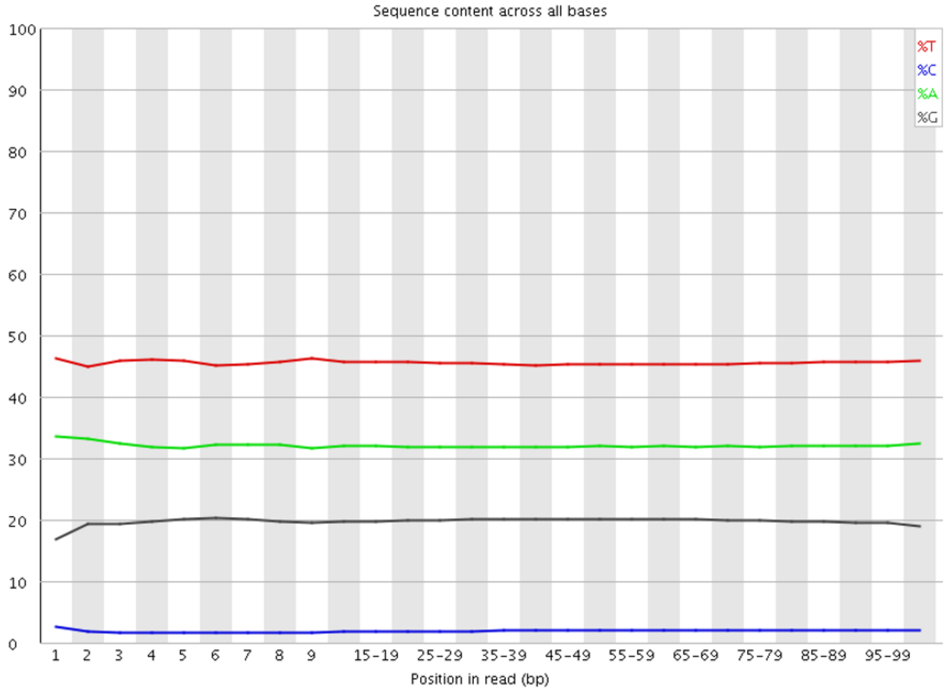
1%

10%

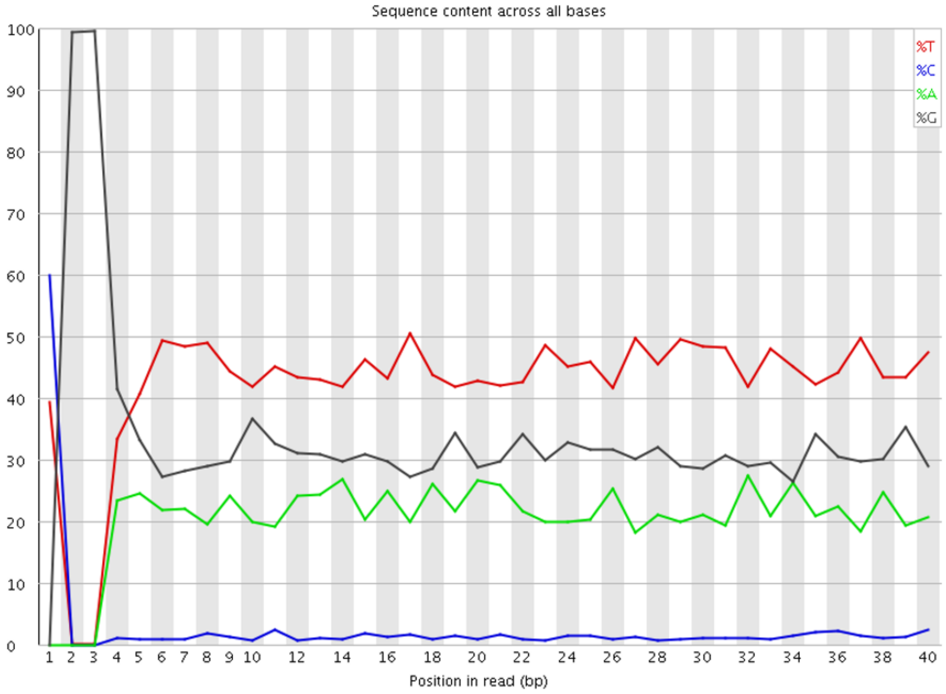


Base Composition

WGBS



RRBS



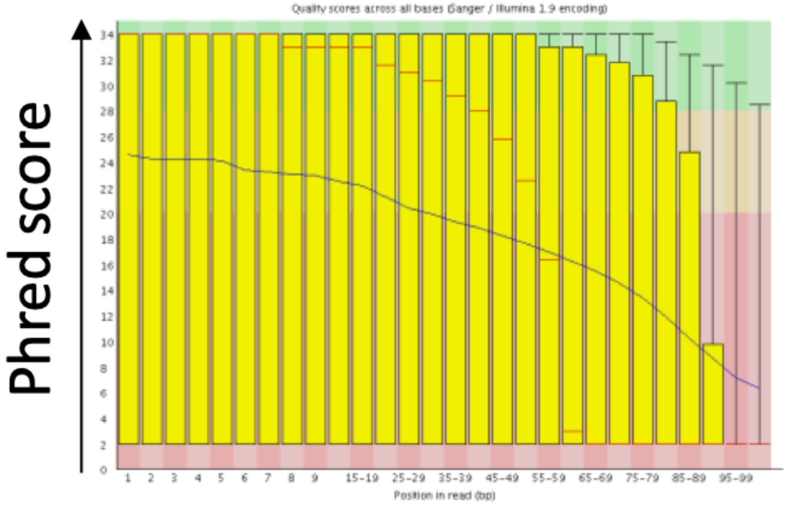
Common bisulfite sequencing QC issues

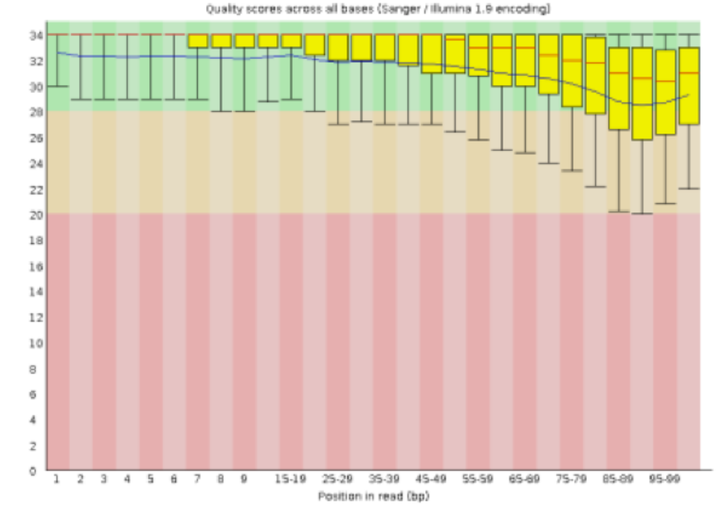
Not observed in ChIP or RNA-Seq

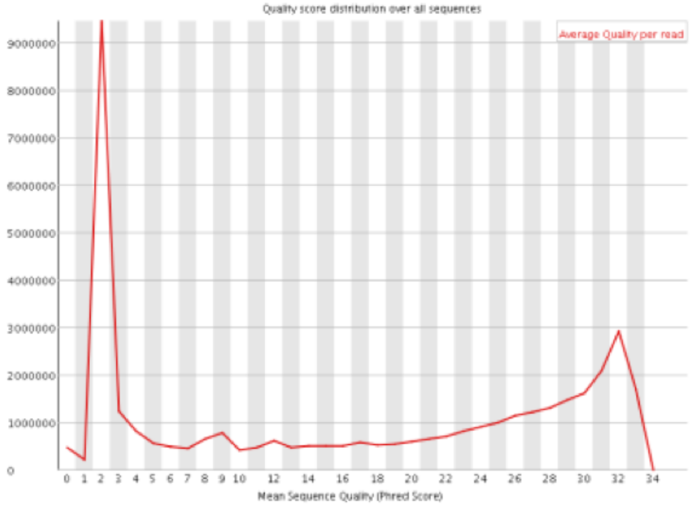
Remove poor quality basecalls

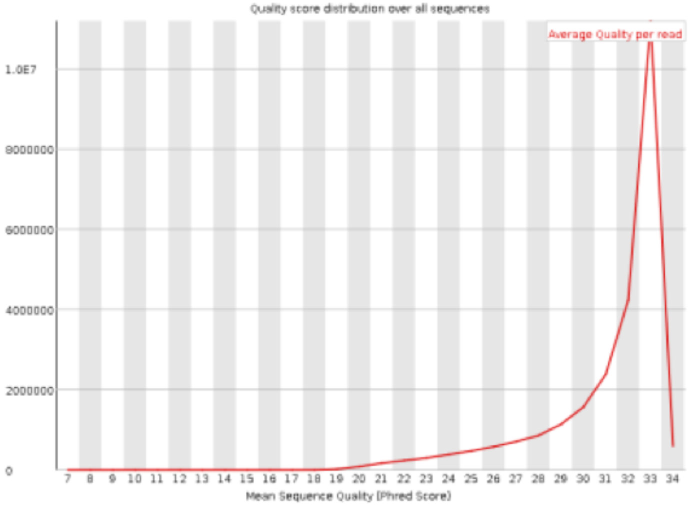
before trimming

after trimming



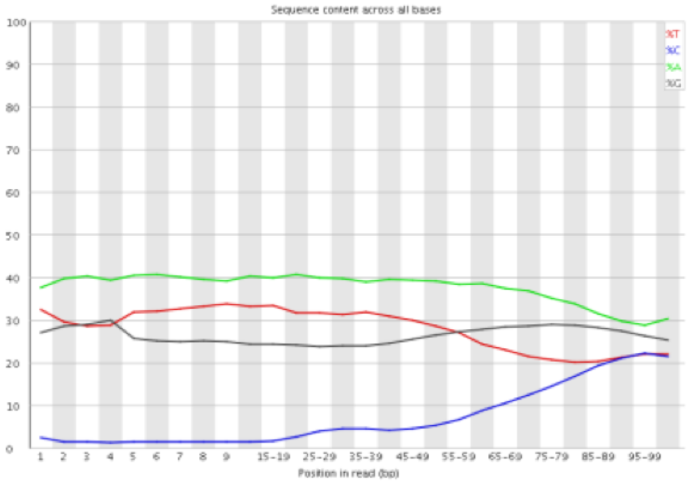




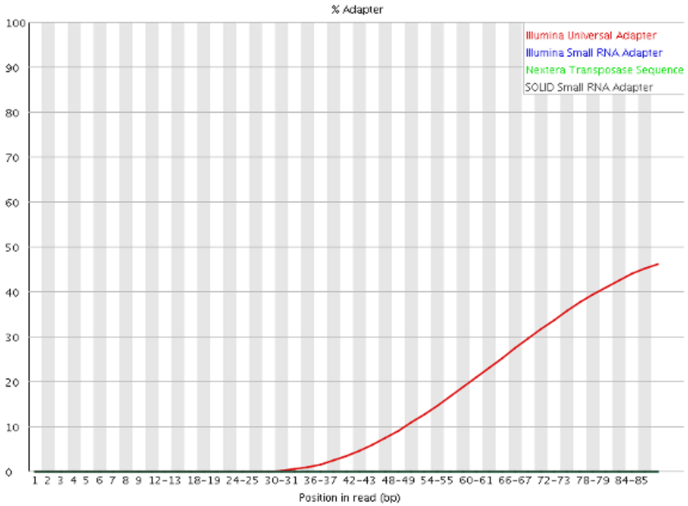


Remove adapter contamination

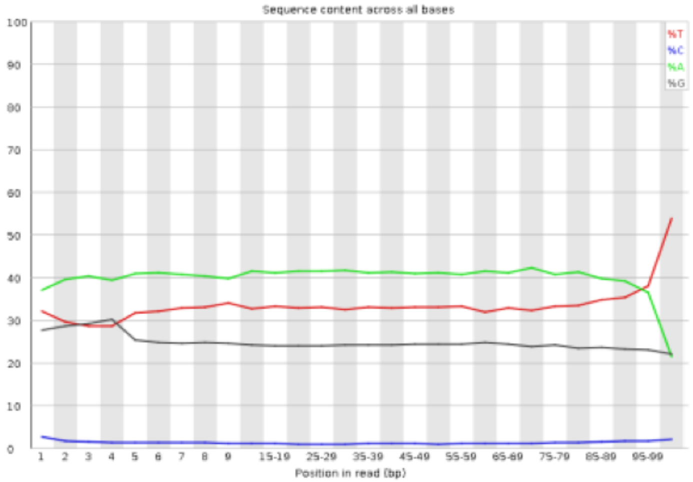
before trimming



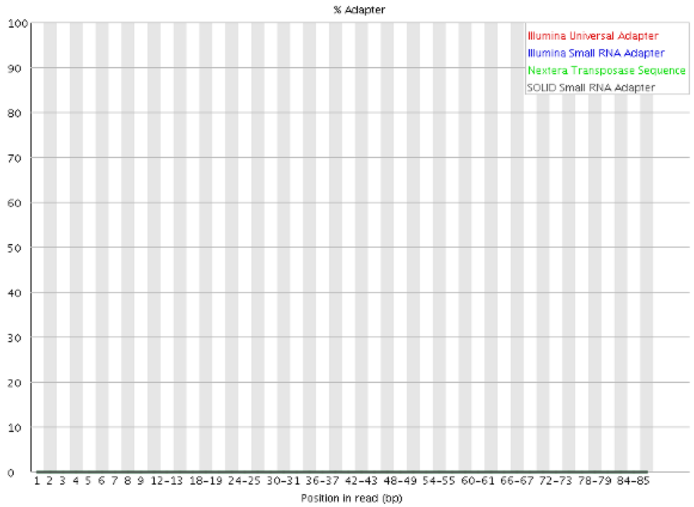
✗ Adapter Content



after trimming



✓ Adapter Content



Summary Adapter/Quality trimming

Important to trim, if not:

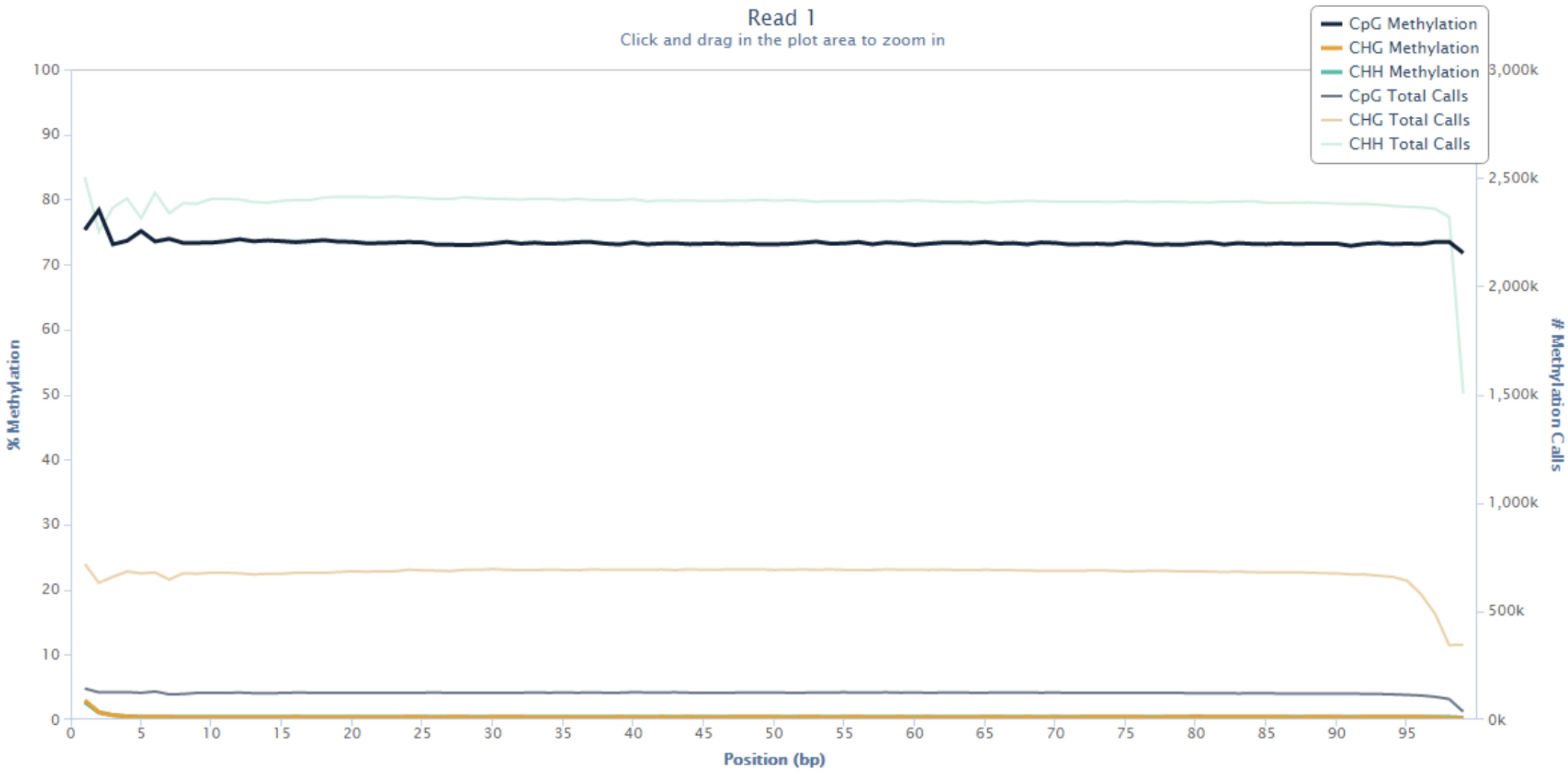
- Low mapping efficiency
- misalignments
- errors in methylation calls (adapters are methylated)
- basecall errors

Quality Control is essential

- Post-alignment
    - Incomplete conversion? non-CpG should be near 100%
    - Degradation? Check alignment rates and insert length
    - Average methylation levels
    - PCR bias? Deduplicate?

M-bias

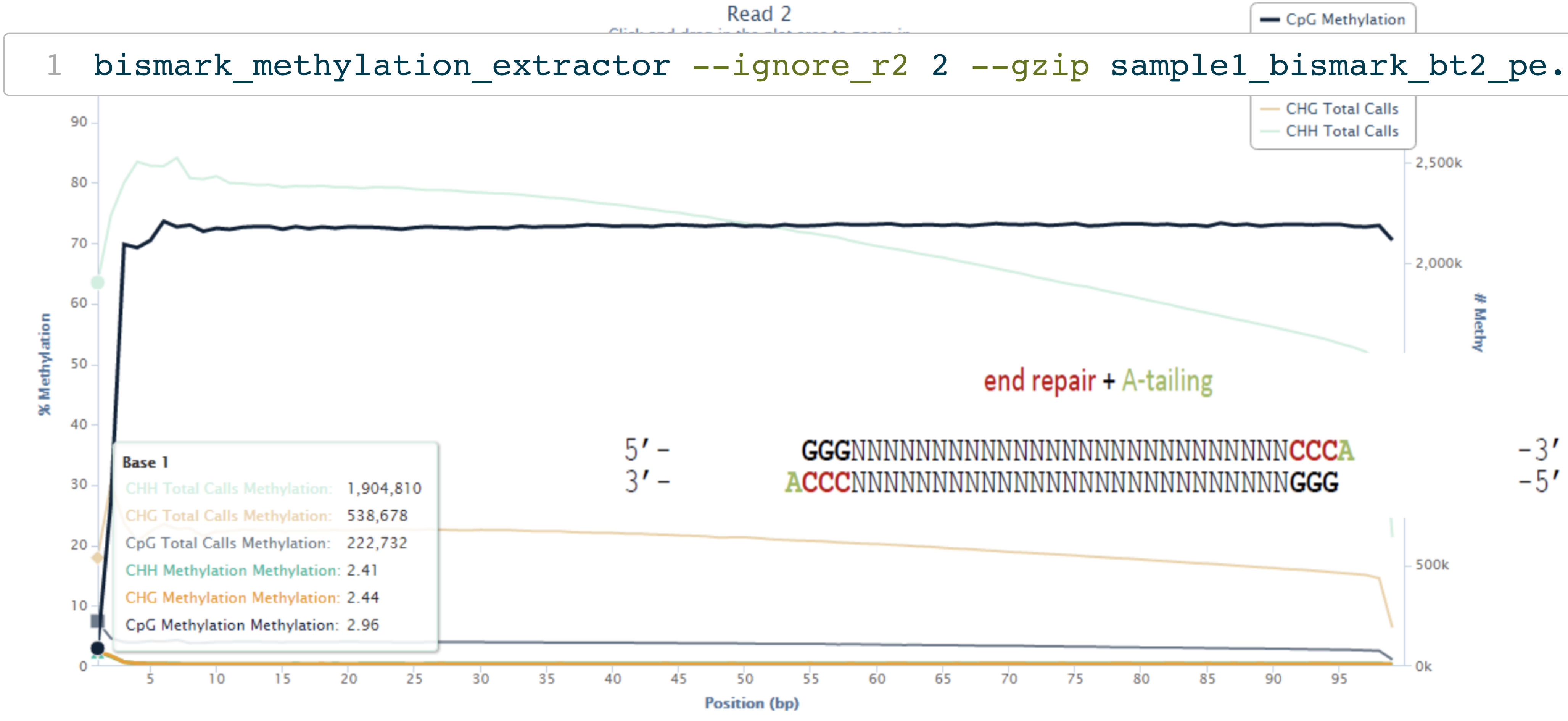
Average methylation levels across the entire length of the read



M-bias

Average methylation levels across the entire length of the read

```
bismark_methylation_extractor --ignore_r2 2 --gzip sample1_bismark_bt2_pe.bam
```



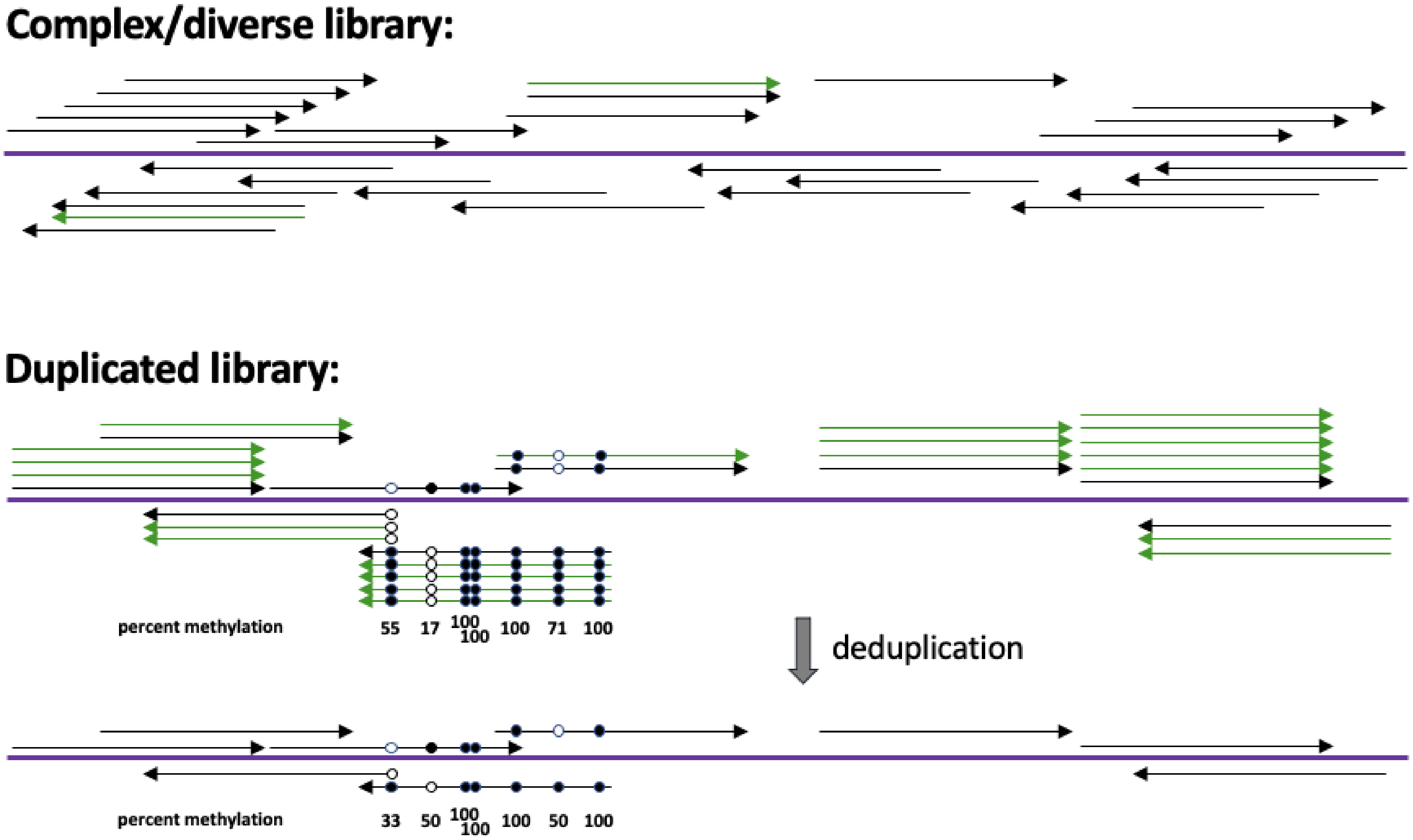
Base 1

CHH Total Calls Methylation: 1,904,810
CHG Total Calls Methylation: 538,678
CpG Total Calls Methylation: 222,732
CHH Methylation Methylation: 2.41
CHG Methylation Methylation: 2.44
CpG Methylation Methylation: 2.96

end repair + A-tailing

5′ – GGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCA –3′
3′ – ACCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGG –5′

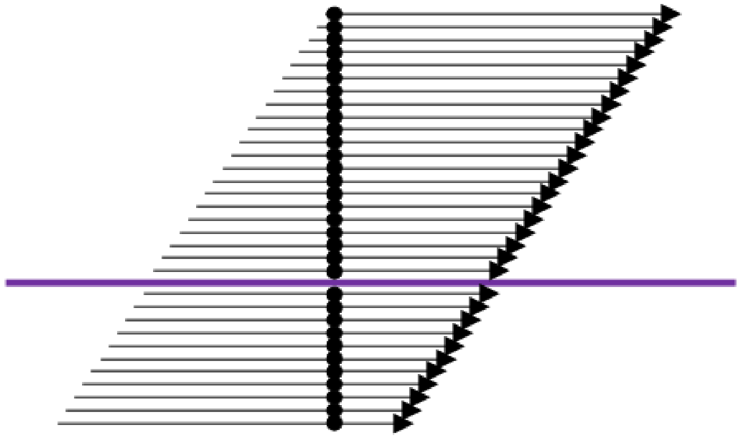
Sequence duplication



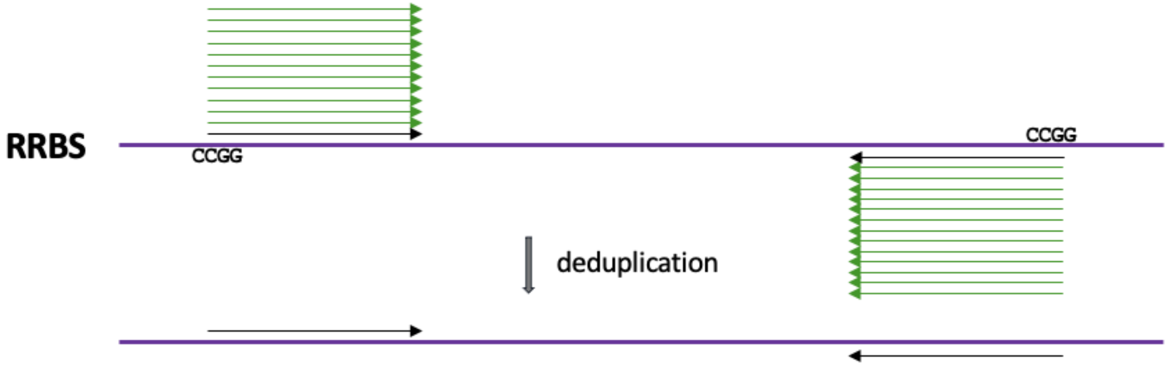
Deduplication?

Advisable for large genomes and moderate coverage

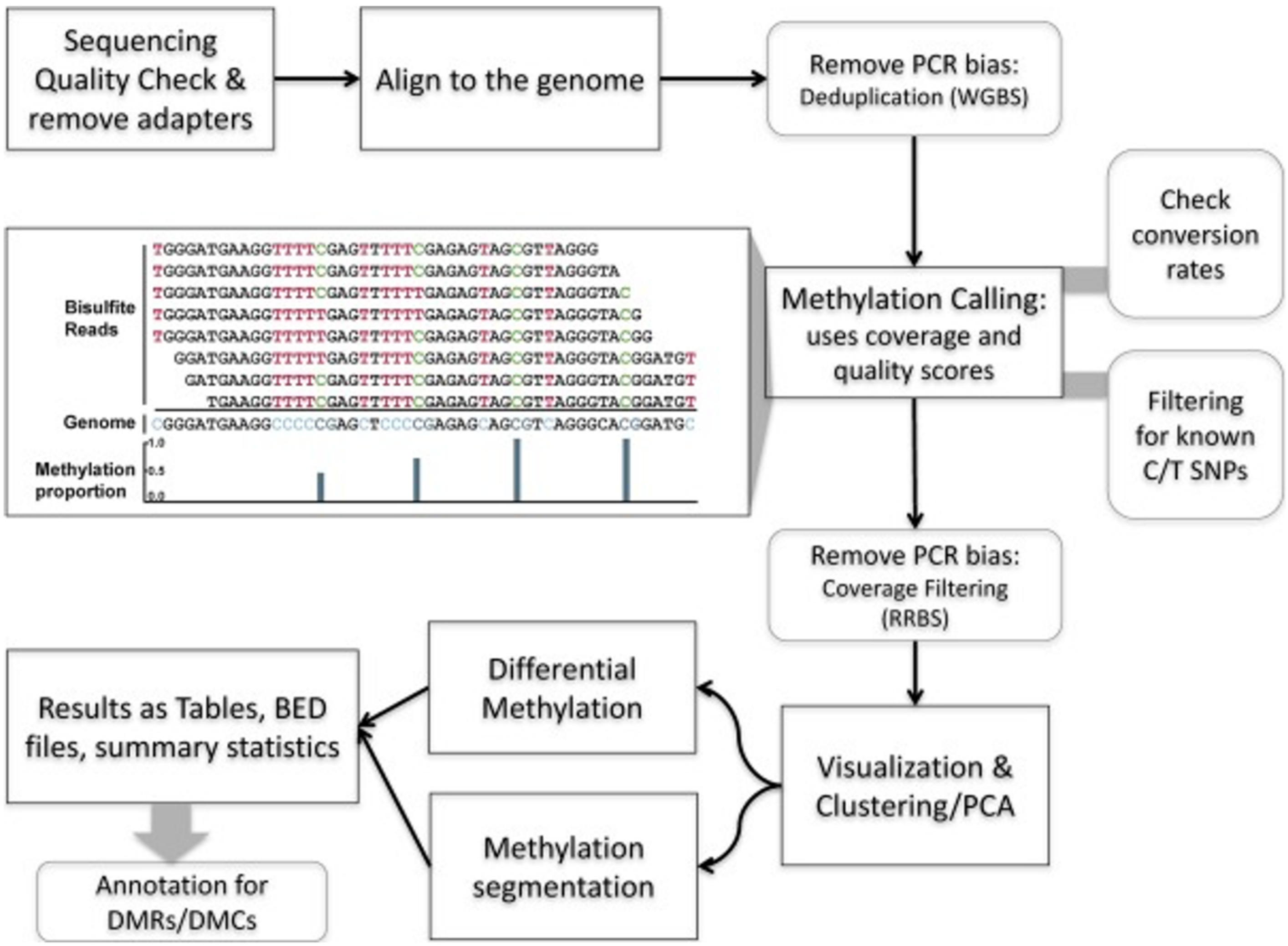
- Unlikely to sequence several genuine copies
- Should have sufficient coverage, even after dedup



NOT advisable for RRBS or other target enrichment methods - high coverage expected



Workflow



Sequencing Quality Check & remove adapters

Align to the genome

Remove PCR bias: Deduplication (WGBS)

Methylation Calling: uses coverage and quality scores

Check conversion rates

Filtering for known C/T SNPs

Remove PCR bias: Coverage Filtering (RRBS)

Visualization & Clustering/PCA

Differential Methylation

Methylation segmentation

Results as Tables, BED files, summary statistics

Annotation for DMRs/DMCs

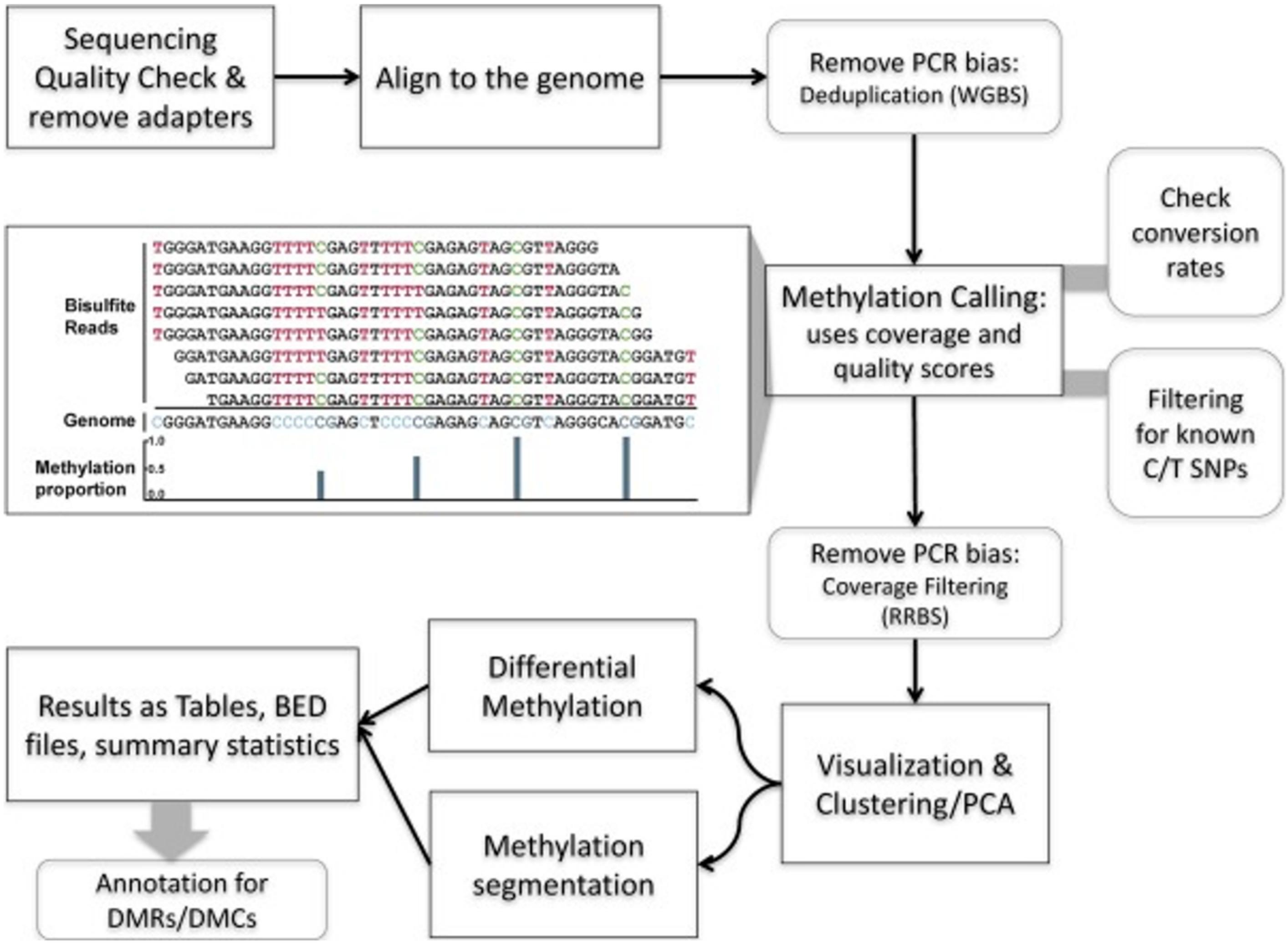
Bisulfite Reads
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGG
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTA
TGGGATGAAGGTTTTCGAGTTTTTTGAGAGTAGCGTTAGGGTAC
TGGGATGAAGGTTTTTGAGTTTTTTGAGAGTAGCGTTAGGGTACG
TGGGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGG
GGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
GATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
TGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
Genome
CGGGATGAAGGCCCCCGAGCTCCCCGAGAGCAGCGTCAGGGCACGGATGC
Methylation proportion
1.0
0.5
0.0

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

- nf-core pipeline: methylseq (see Thursday)
- Preprocessing + QC
  - 2 aligners: Bismark or bwa/meth/MethylDackel
  - QC: qualimap, preseq an multiqc
- Output ready for downstrea analysis

MethylKit: R package



Sequencing Quality Check & remove adapters

Align to the genome

Remove PCR bias: Deduplication (WGBS)

Methylation Calling: uses coverage and quality scores

Check conversion rates

Filtering for known C/T SNPs

Remove PCR bias: Coverage Filtering (RRBS)

Visualization & Clustering/PCA

Differential Methylation

Methylation segmentation

Results as Tables, BED files, summary statistics

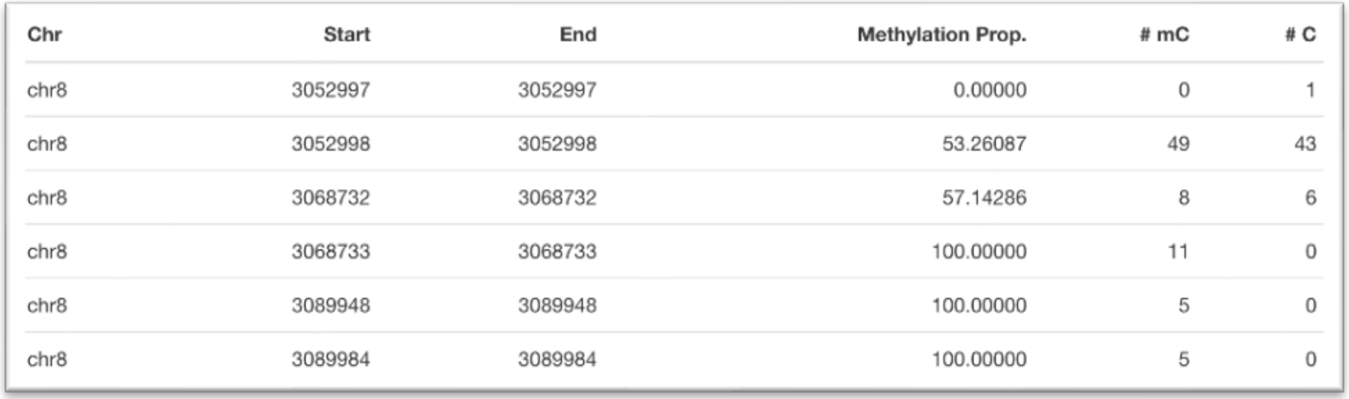
Annotation for DMRs/DMCs

Bisulfite Reads
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGG
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTA
TGGGATGAAGGTTTTCGAGTTTTTTGAGAGTAGCGTTAGGGTAC
TGGGATGAAGGTTTTTGAGTTTTTTGAGAGTAGCGTTAGGGTACG
TGGGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGG
GGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
GATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
TGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
Genome
CGGGATGAAGGCCCCCGAGCTCCCCGAGAGCAGCGTCAGGGCACGGATGC
Methylation proportion
1.0
0.5
0.0

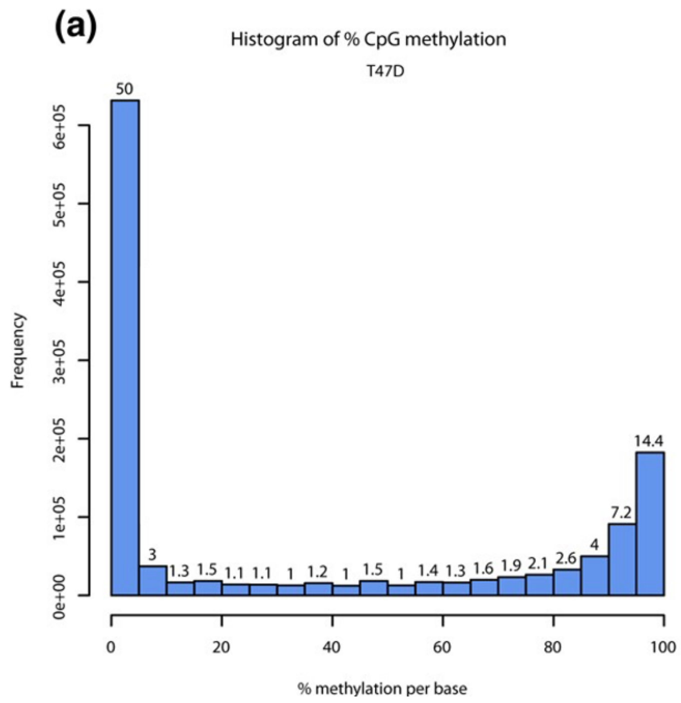
Can read Bismark coverage files as input

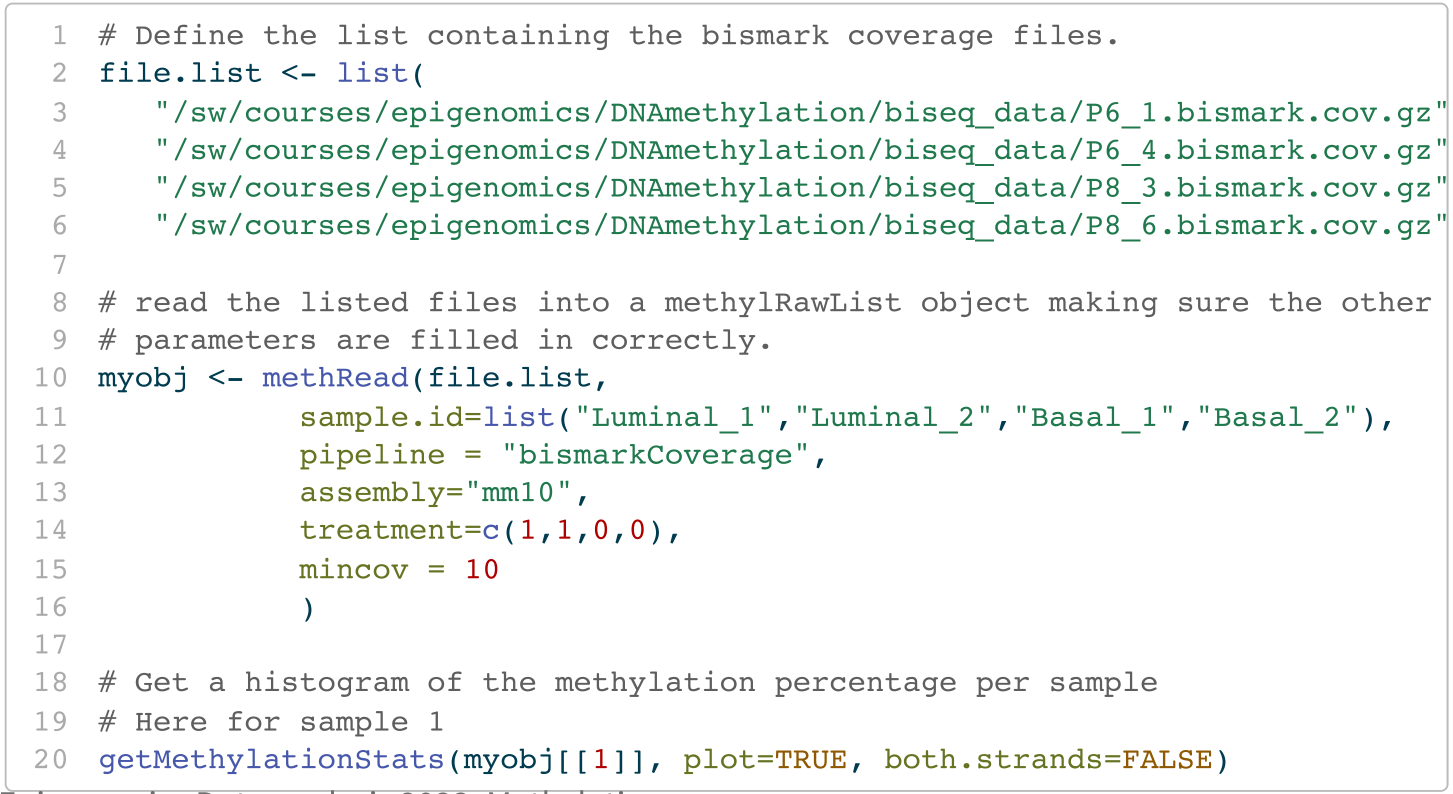
Descriptive statistics

Coverage file Bismark



Chr	Start	End	Methylation Prop.	# mC	# C
chr8	3052997	3052997	0.00000	0	1
chr8	3052998	3052998	53.26087	49	43
chr8	3068732	3068732	57.14286	8	6
chr8	3068733	3068733	100.00000	11	0
chr8	3089948	3089948	100.00000	5	0
chr8	3089984	3089984	100.00000	5	0



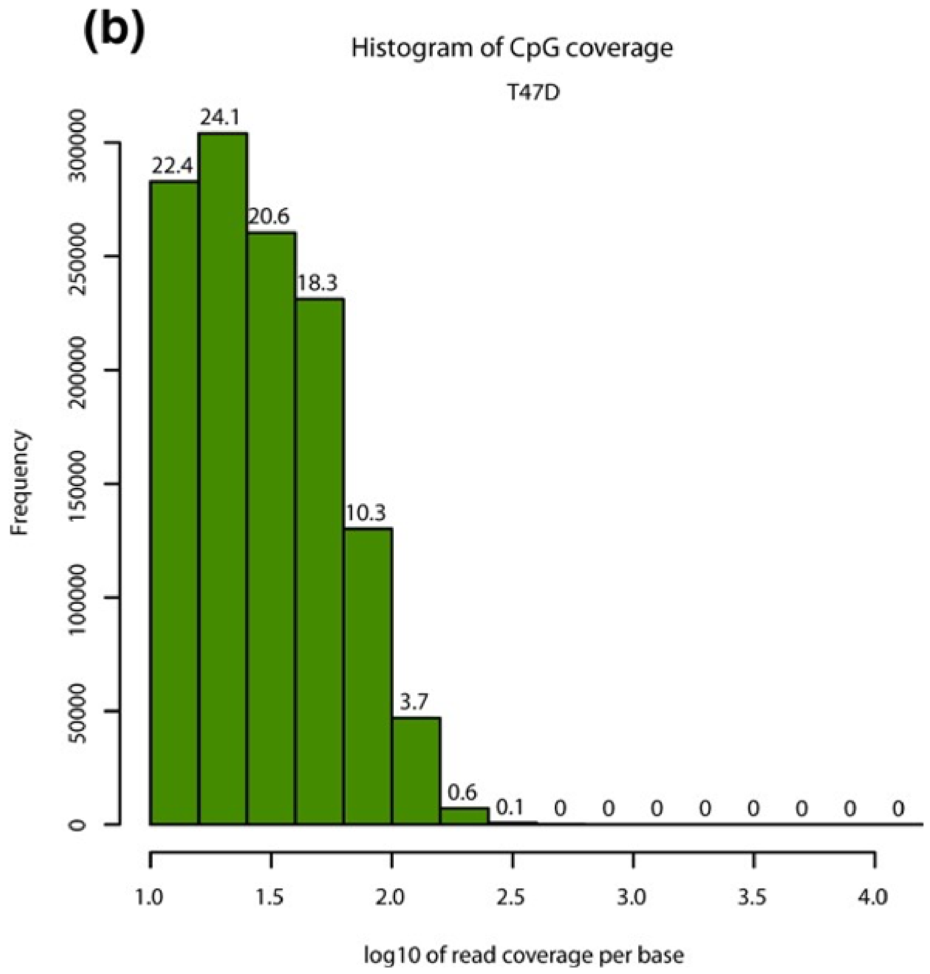


```
# Define the list containing the bismark coverage files.
file.list <- list(
   "/sw/courses/epigenomics/DNAmethylation/biseq_data/P6_1.bismark.cov.gz"
   "/sw/courses/epigenomics/DNAmethylation/biseq_data/P6_4.bismark.cov.gz"
   "/sw/courses/epigenomics/DNAmethylation/biseq_data/P8_3.bismark.cov.gz"
   "/sw/courses/epigenomics/DNAmethylation/biseq_data/P8_6.bismark.cov.gz"

# read the listed files into a methylRawList object making sure the other
# parameters are filled in correctly.
myobj <- methRead(file.list,
           sample.id=list("Luminal_1","Luminal_2","Basal_1","Basal_2"),
           pipeline = "bismarkCoverage",
           assembly="mm10",
           treatment=c(1,1,0,0),
           mincov = 10
           )

# Get a histogram of the methylation percentage per sample
# Here for sample 1
getMethylationStats(myobj[[1]], plot=TRUE, both.strands=FALSE)
```

Descriptive statistics

Coverage Distribution

```
# Get a histogram of the read coverage per sample
getCoverageStats(myobj[[1]], plot=TRUE, both.strands=FALSE)
```



- Secondary peak towards the right -> PCR duplication?
- Filter cutoff?

```
myobj.filt <- filterByCoverage(myobj,
                      lo.count=10,
                      lo.perc=NULL,
                      hi.count=NULL,
                      hi.perc=99.9)
```

Filtering

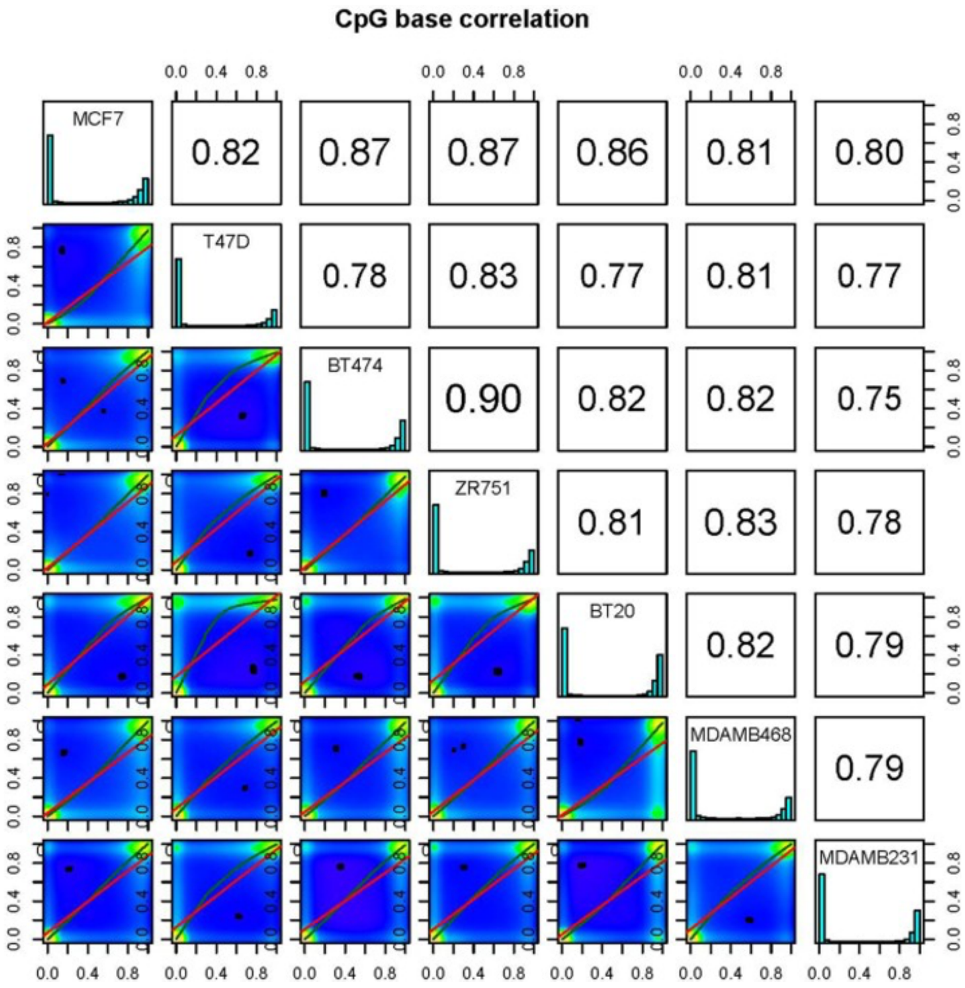
Remove CpG that have no variation

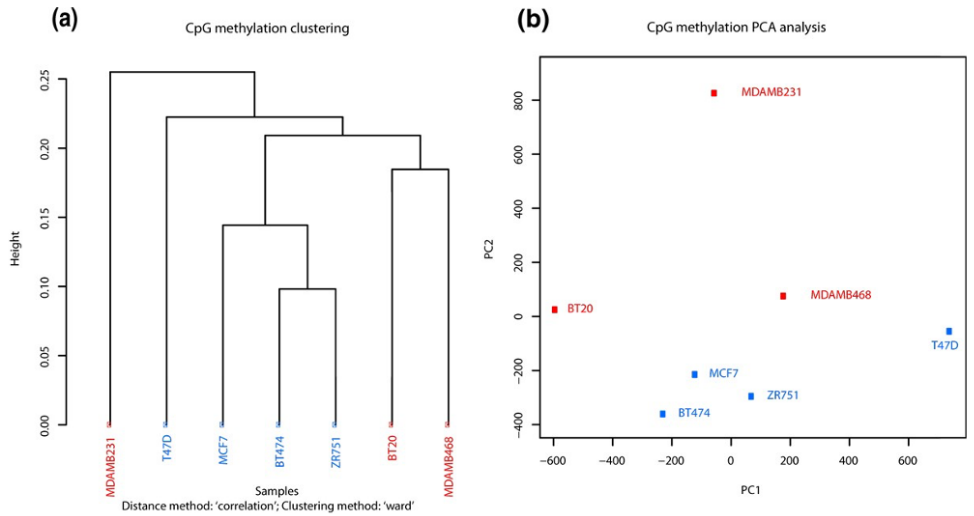
```r
# get percent methylation matrix
pm=percMethylation(meth)

# calculate standard deviation of CpGs
sds=matrixStats::rowSds(pm)

# Visualize the distribution of the per-CpG standard deviation
# to determine a suitable cutoff
hist(sds, breaks = 100)

# keep only CpG with standard deviations larger than 2%
meth <- meth[sds > 2]
```

Remove SNP overlap

```r
# give the locations of 2 example SNPs
mut <- GRanges(seqnames=c("chr21","chr21"),
         ranges=IRanges(start=c(9853296, 9853326),
                        end=c( 9853296,9853326)))

# select CpGs that do not overlap with mutations
meth <- meth[!as(meth,"GRanges") %over% mut, ]
```

Sample Structure


CpG base correlation


(a) CpG methylation clustering

Distance method: 'correlation'; Clustering method: 'ward'

(b) CpG methylation PCA analysis

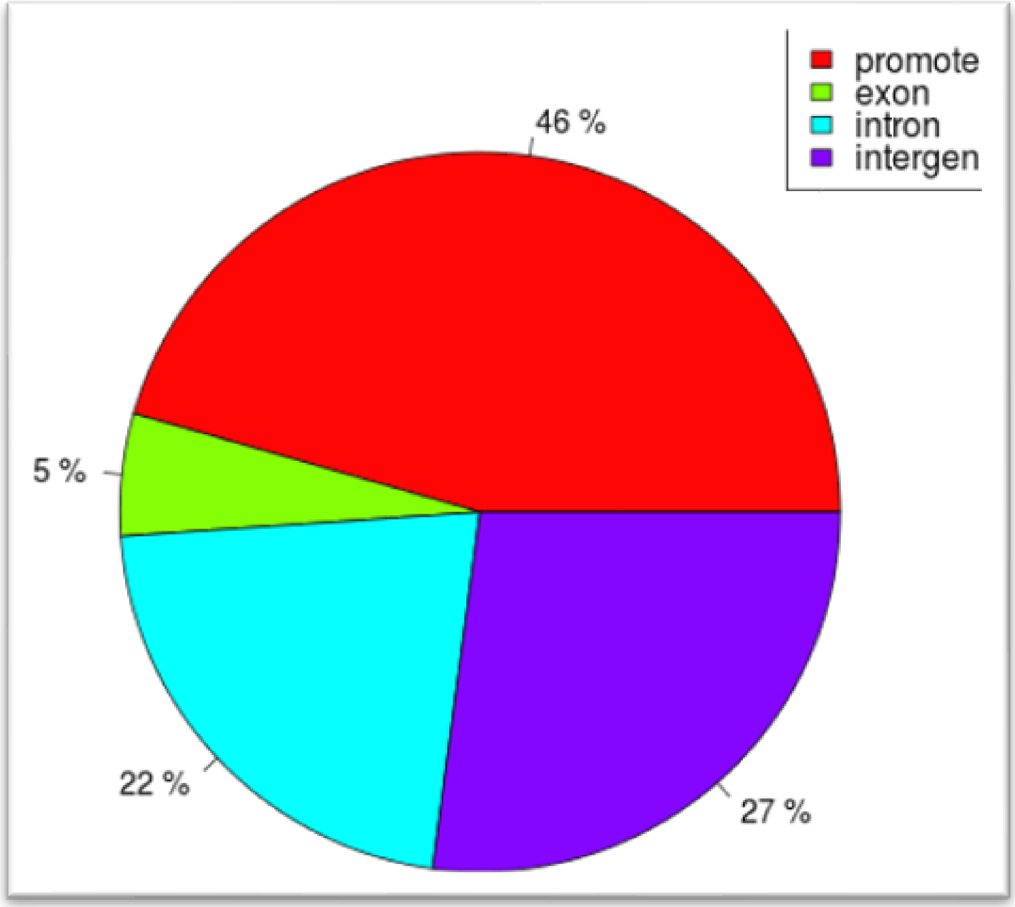
```r
getCorrelation(meth,plot=TRUE)
clusterSamples(meth, dist="correlation",
                method="ward", plot=TRUE)
PCASamples(meth)
```

Differential Methylation

- Many choices; often calculated by comparing proportion in methylated Cs in a test relative to control
- No replicates: Fisher's exact test
- Replicates:
  - linear regression
  - logistic regression (works with [0-1] data)
  - Beta-binomial (count data)
- Regression models can add covariantes/confounders
- Aggregate in regions (see lab)

Annotate results

- How to interpret the DMR/DMPs?
- Integrate with genome annotation datasets
    - Where in relation to gene/regulatory region?
- Genomation R packge: toolkit for annotation
- Lab: basic annotation transcripts and CpG islands
- Requires some knowledge of R (GenomicRanges package)

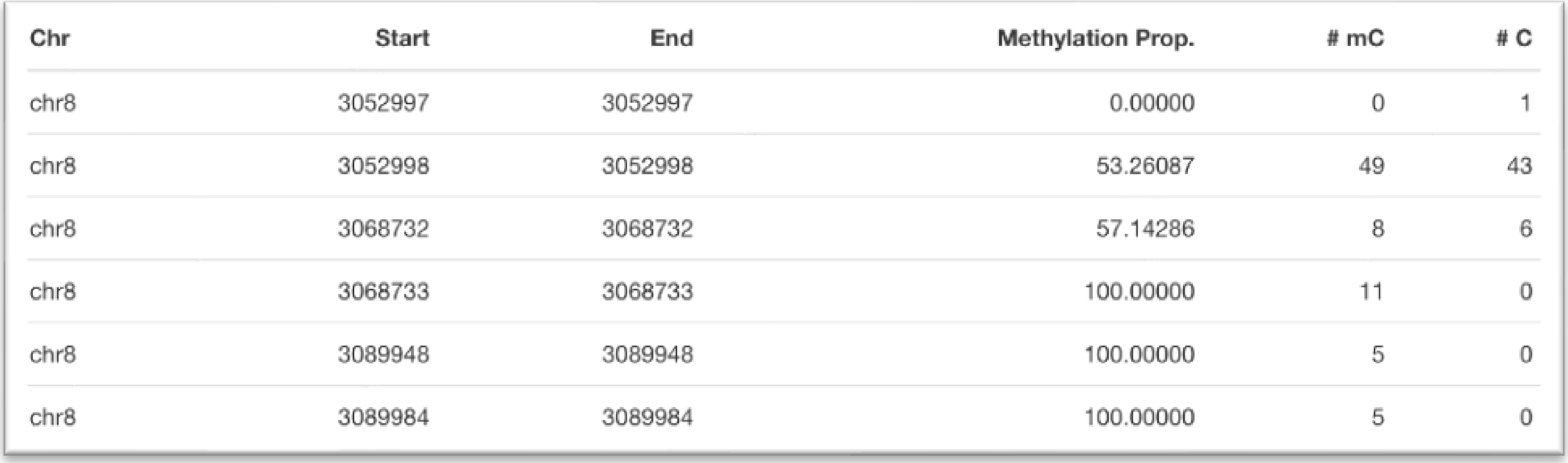


Remarks

- Normalization somewhat less important for bisulfite sequencing (Fisher's exact is sensitive to sequencing depth)
- Gene enrichments is as difficult as for arrays, not many implemented methods (rGREAT, Goseq)

Lab

- Small dataset of mammary gland cells in mouse
- 4 samples: 2 luminal, 2 basal
- Bismark coverage files



Chr	Start	End	Methylation Prop.	# mC	# C
chr8	3052997	3052997	0.00000	0	1
chr8	3052998	3052998	53.26087	49	43
chr8	3068732	3068732	57.14286	8	6
chr8	3068733	3068733	100.00000	11	0
chr8	3089948	3089948	100.00000	5	0
chr8	3089984	3089984	100.00000	5	0