Introduction to methylation analysis

Data processing, QC and alignment

Sergio Martínez Cuesta
sermarcue@gmail.com

Employment disclosures:

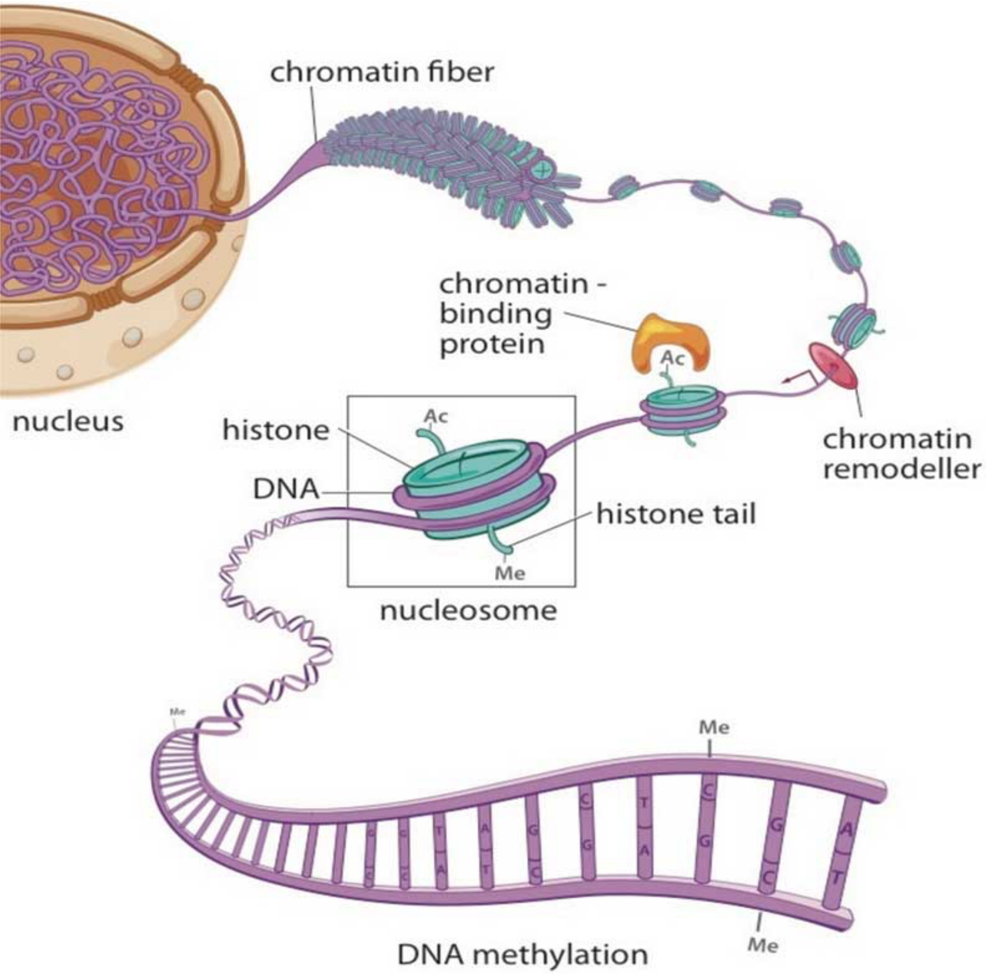
UNIVERSITY OF CAMBRIDGE

AstraZeneca

Materials obtained from:

Babraham Bioinformatics

Epigenetics



From The Cell Biology of Stem Cells (2010)

Studies changes in gene expression which are not encoded by the underlying DNA sequence

- histone modification
- non-coding RNAs
- higher order structure (accessibility/compaction)

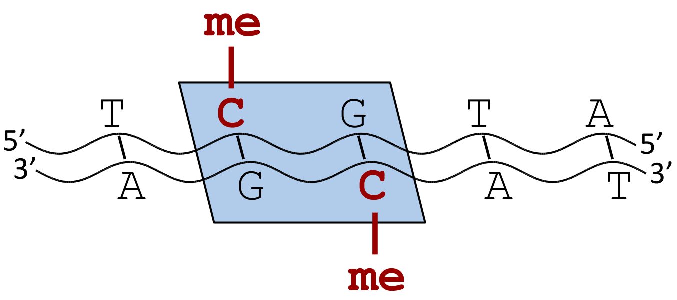
- DNA cytosine methylation

Types of DNA methylation

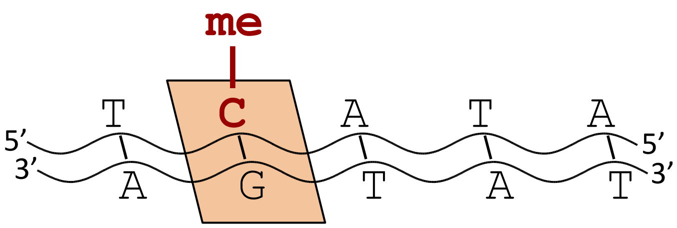
canonical

	Plants	Mammals
CG	symmetric	symmetric

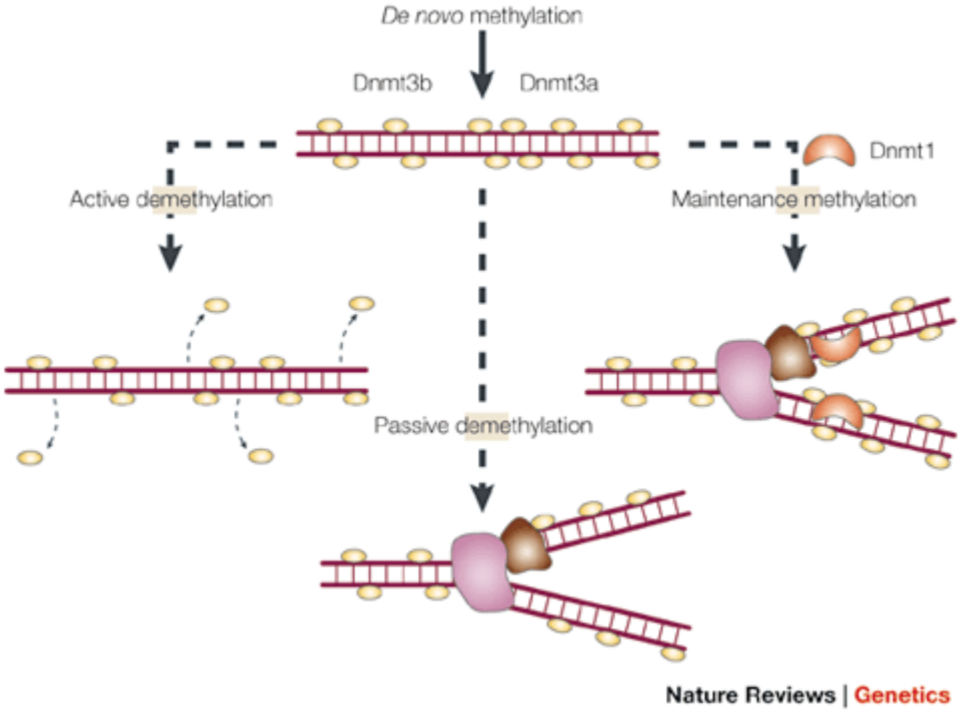
CG context



non-CG context

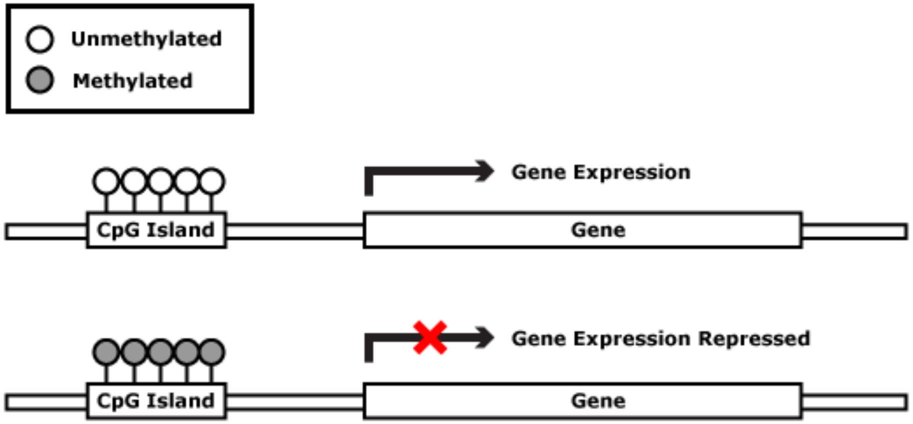


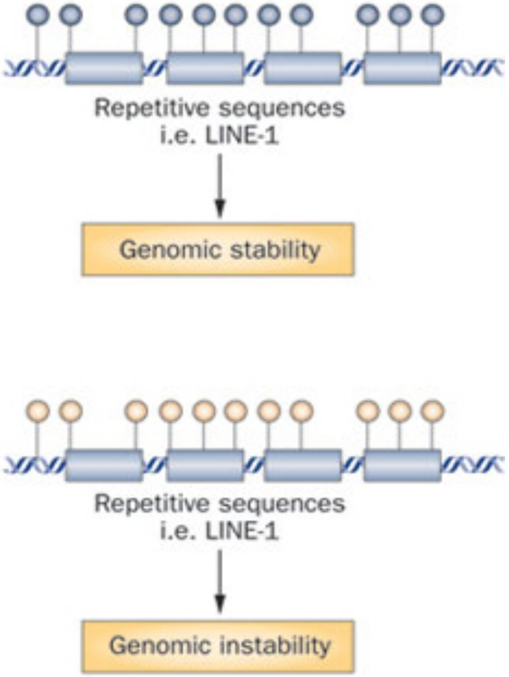
DNA methylation is maintained



from W. Reik & J. Walter, Nat. Rev. Genet. 2001

Regulation by DNA methylation





Silencing of gene expression
Tissue differentiation and embryonic development

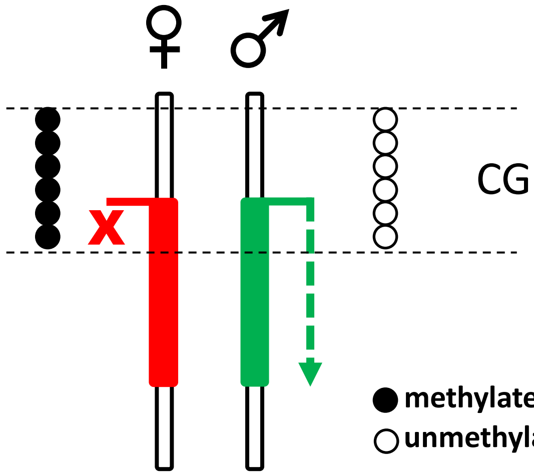
Repeat activity
Genomic stability

Faults in correct DNA methylation may result in

- early development failure
- epigenetic syndromes
- cancer

Babraham Bioinformatics

Imprinted Genes: mono-allelic expression



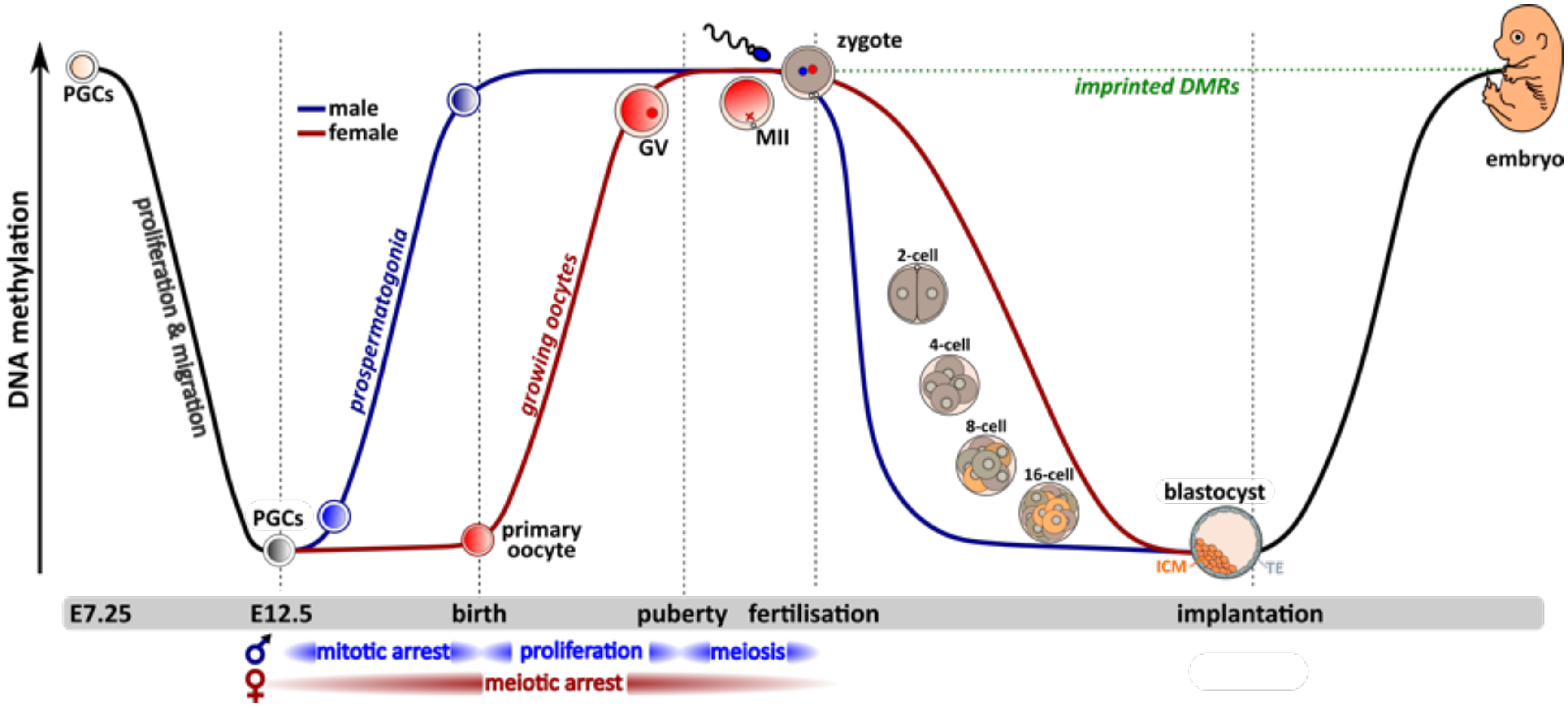
Differential allelic DNA methylation

CGI (CpG island)

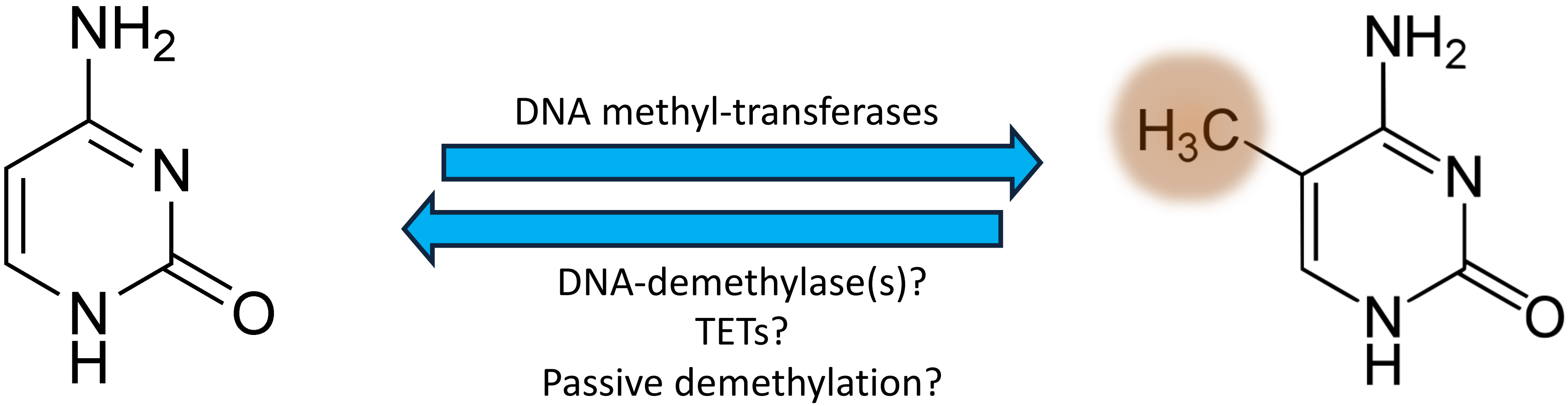
● methylated CpG
○ unmethylated CpG

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

DNA methylation is reset during reprogramming



DNA Methylation



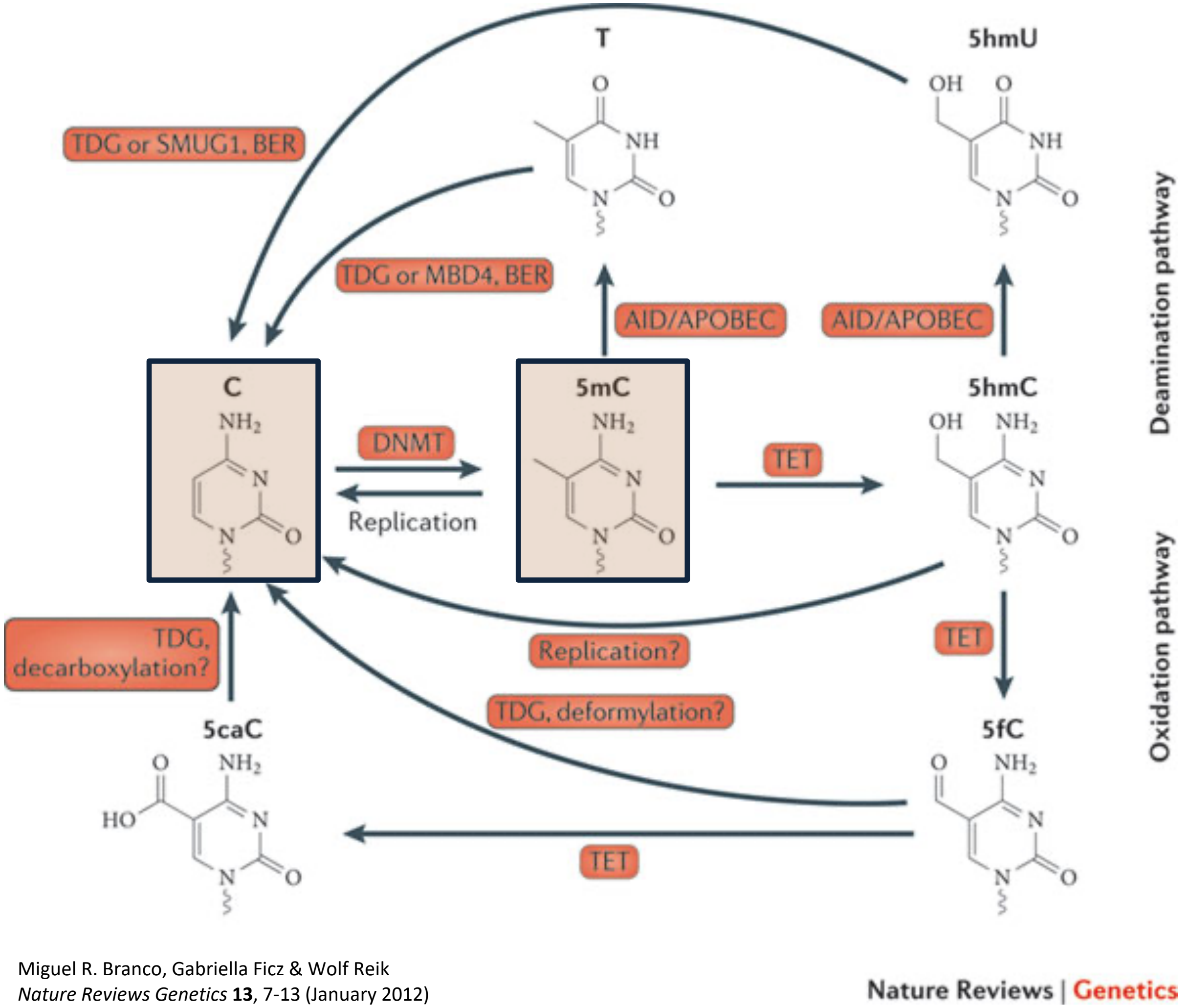
DNA methyl-transferases

DNA-demethylase(s)?
TETs?
Passive demethylation?

Cytosine

5-methyl Cytosine

Other cytosine modifications



Miguel R. Branco, Gabriella Ficz & Wolf Reik
Nature Reviews Genetics **13**, 7-13 (January 2012)

Measuring DNA methylation by Bisulfite-sequencing

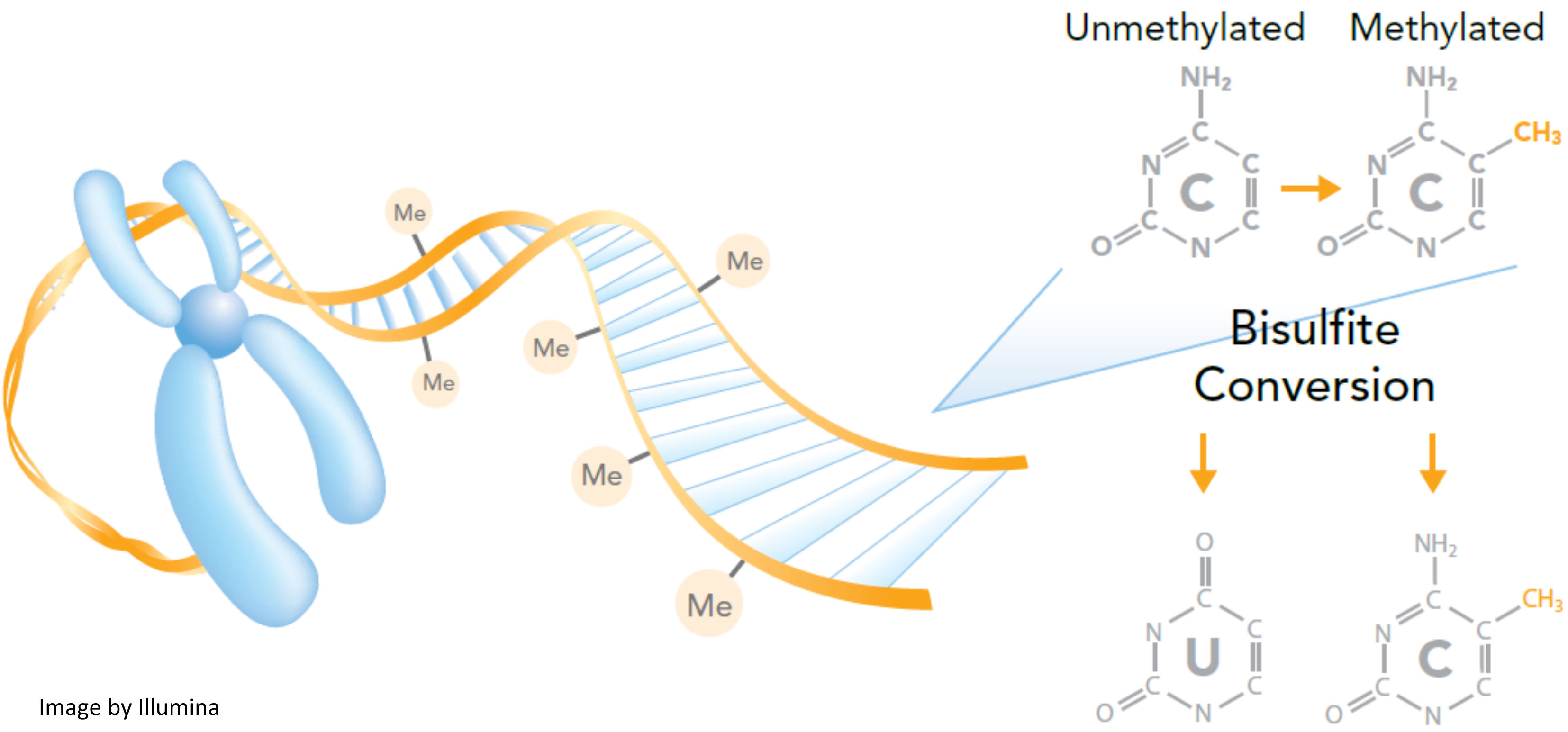
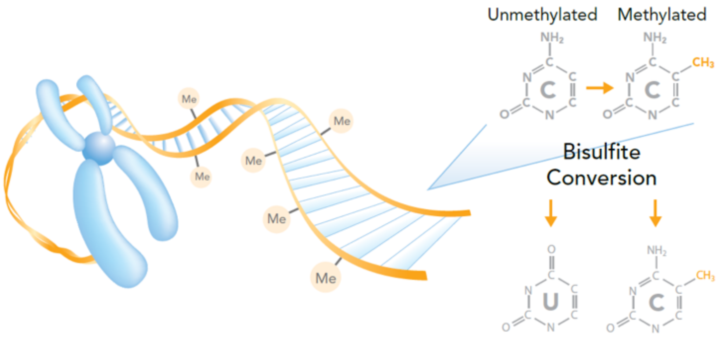


Image by Illumina

Bisulfite Informatics



```
          me                  me
CCAGTCGCTATAGCGCGATATCGTA
```

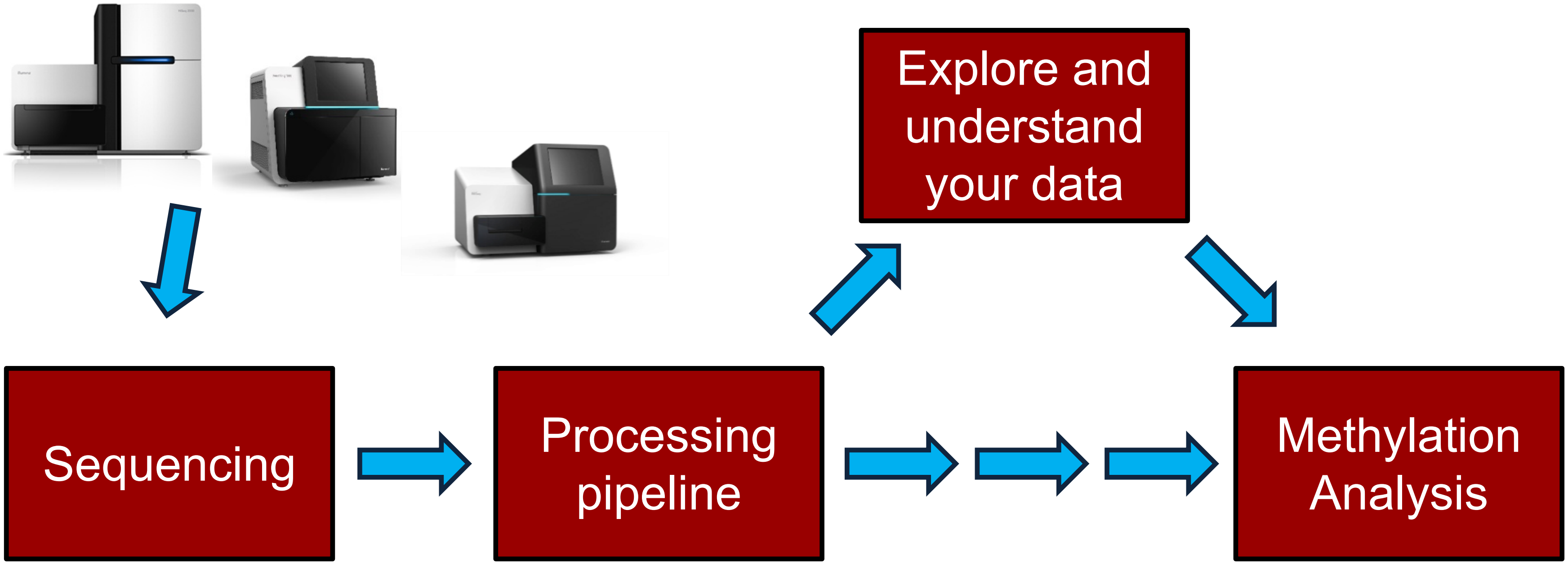
Convert

```
TTAGTTGCTATAGTGCGATATTGTA
```

Map

```
   TTAGTTGCTATAGTGCGATATTGTA
   | | | | | | | | | | | | | | | | | | | | | | | | |
...CCAGTCGCTATAGCGCGATATCGTA...
```

BS-Seq Analysis Workflow



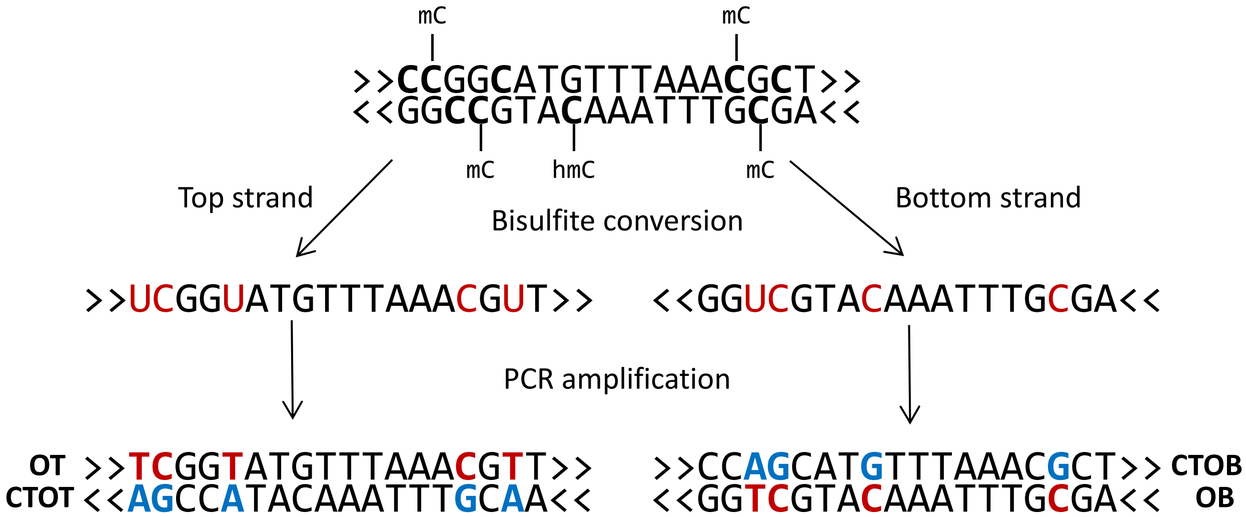
Sequencing

Processing pipeline

Explore and understand your data

Methylation Analysis

Bisulfite conversion of a genomic locus



mC mC

```
>>CCGGCATGTTTAAACGCT>>
<<GGCCGTACAAATTTGCGA<<
```

mC hmC mC

Top strand

Bisulfite conversion

Bottom strand

```
>>UCGGUATGTTTAAACGUT>>     <<GGUCGTACAAATTTGCGA<<
```

PCR amplification

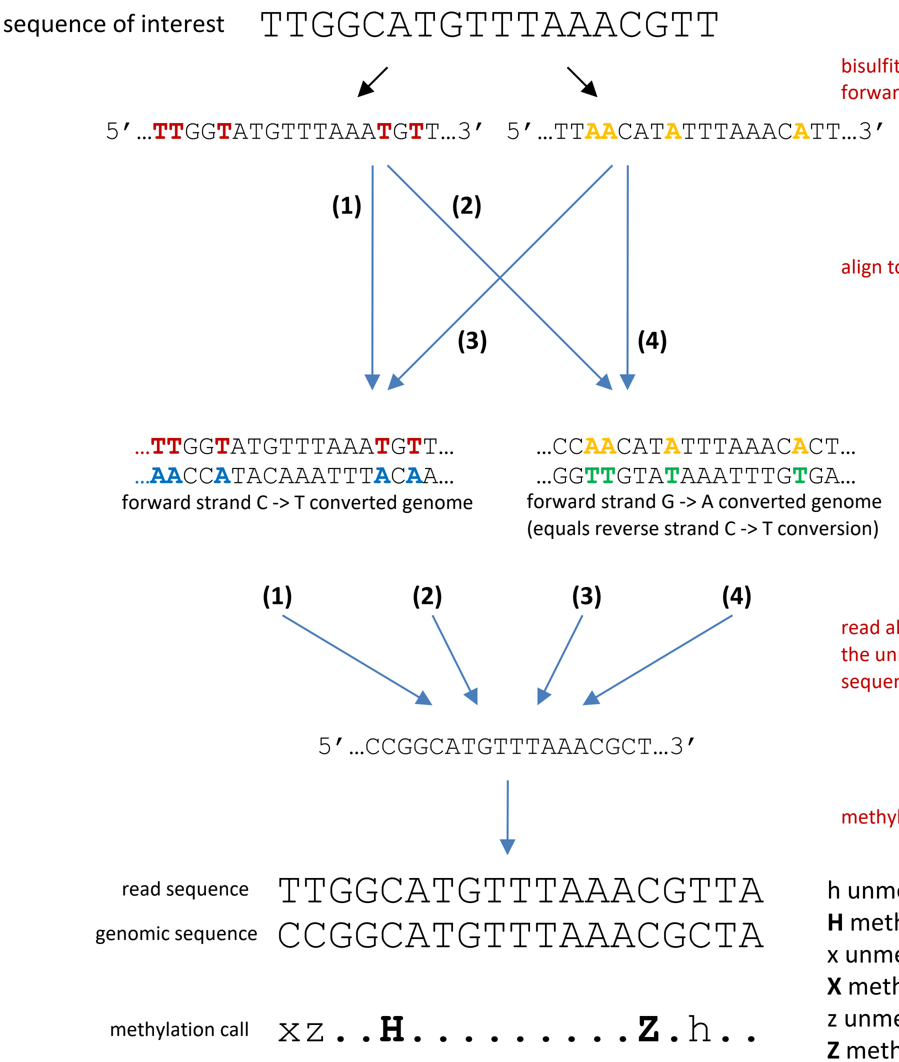
```
OT   >>TCGGTATGTTTAAACGTT>>     >>CCAGCATGTTTAAACGCT>> CTOB
CTOT <<AGCCATACAAATTTGCAA<<     <<GGTCGTACAAATTTGCGA<<  OB
```

- 2 different PCR products and 4 possible different sequence strands from one genomic locus

- each of these 4 sequence strands can theoretically exist in any possible conversion state

3-letter alignment of Bisulfite-Seq reads

Bismark



sequence of interest TTGGCATGTTTAAACGTT

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

5’…TTGGTATGTTTAAATGTT…3’ 5’…TTAACATATTTAAACATT…3’

align to bisulfite converted genomes

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

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

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

5’…CCGGCATGTTTAAACGCT…3’

methylation call

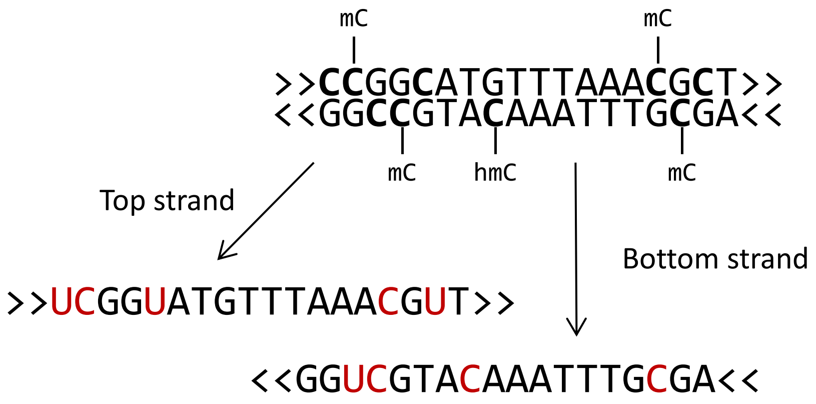
read sequence TTGGCATGTTTAAACGTTA
genomic sequence CCGGCATGTTTAAACGCTA

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

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

Babraham Bioinformatics

Common sequencing protocols



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
```

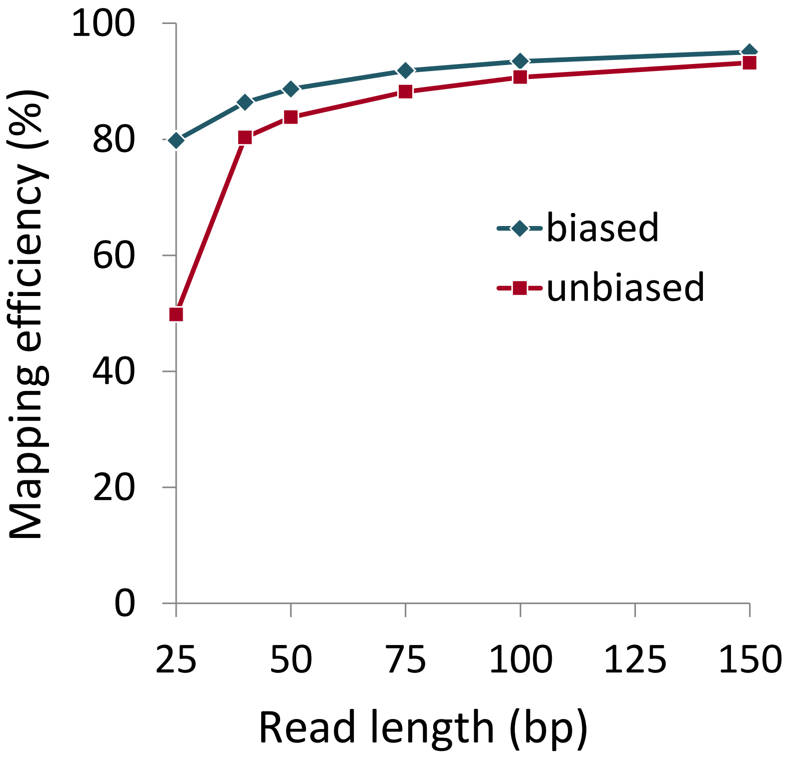
3) Non-directional libraries

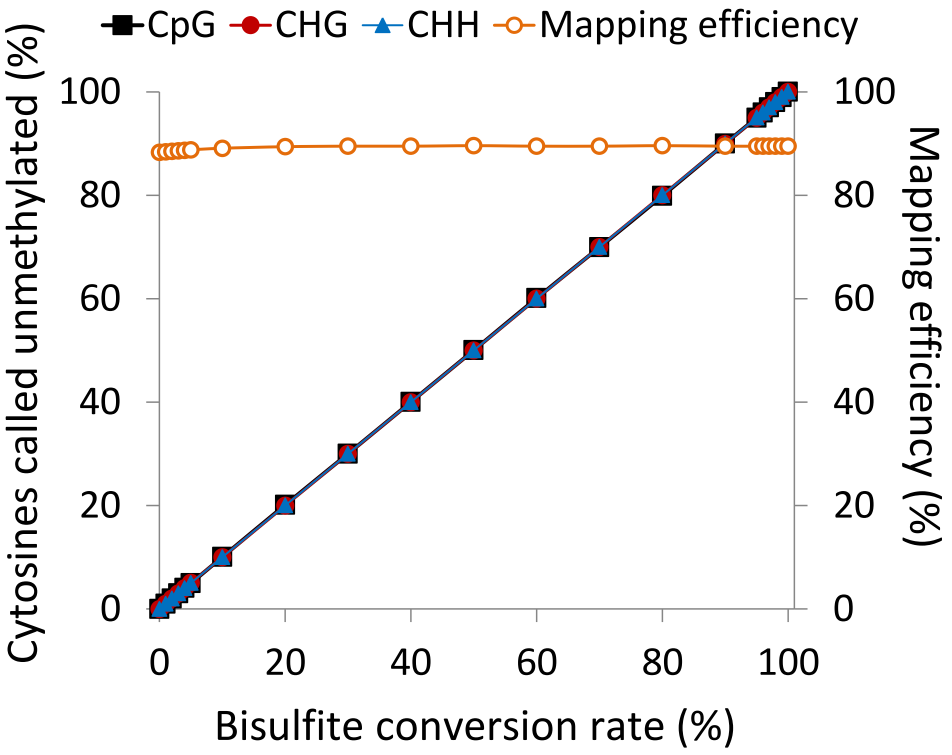
(e.g. single-cell BS-Seq, Zymo Pico Methyl-Seq)

```
  OT >>TCGGTATGTTTAAACGTT>>
CTOT <<AGCCATACAAATTTGCAA<<

     >>CCAGCATGTTTAAACGCT>> CTOB
     <<GGTCGTACAAATTTGCGA<<  OB
```

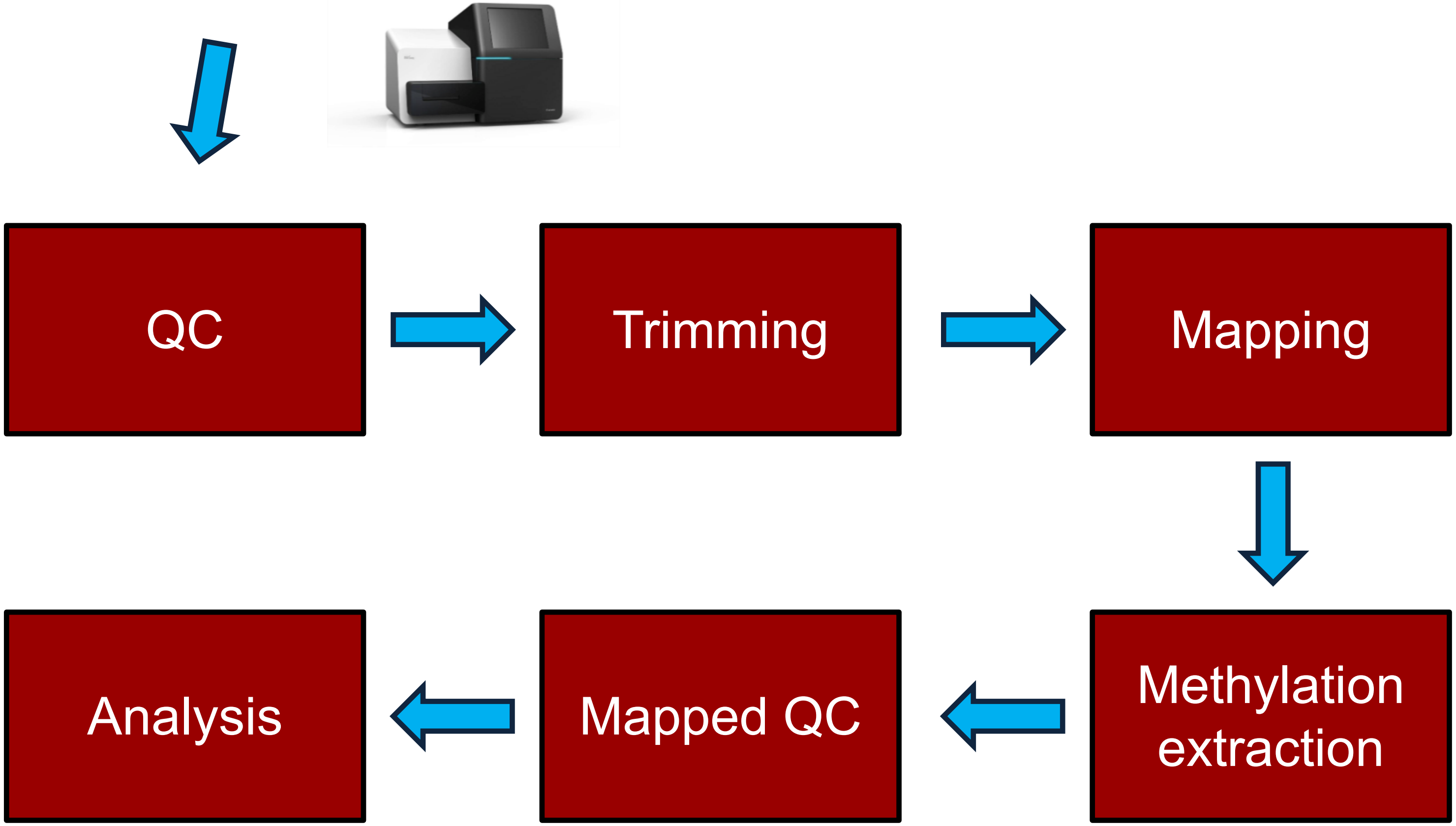
Babraham Bioinformatics

Validation


100
80
60
40
20
0
25
50
75
100
125
150
Mapping efficiency (%)
Read length (bp)
biased
unbiased


CpG
CHG
CHH
Mapping efficiency
100
80
60
40
20
0
0
20
40
60
80
100
100
80
60
40
20
0
Cytosines called unmethylated (%)
Mapping efficiency (%)
Bisulfite conversion rate (%)

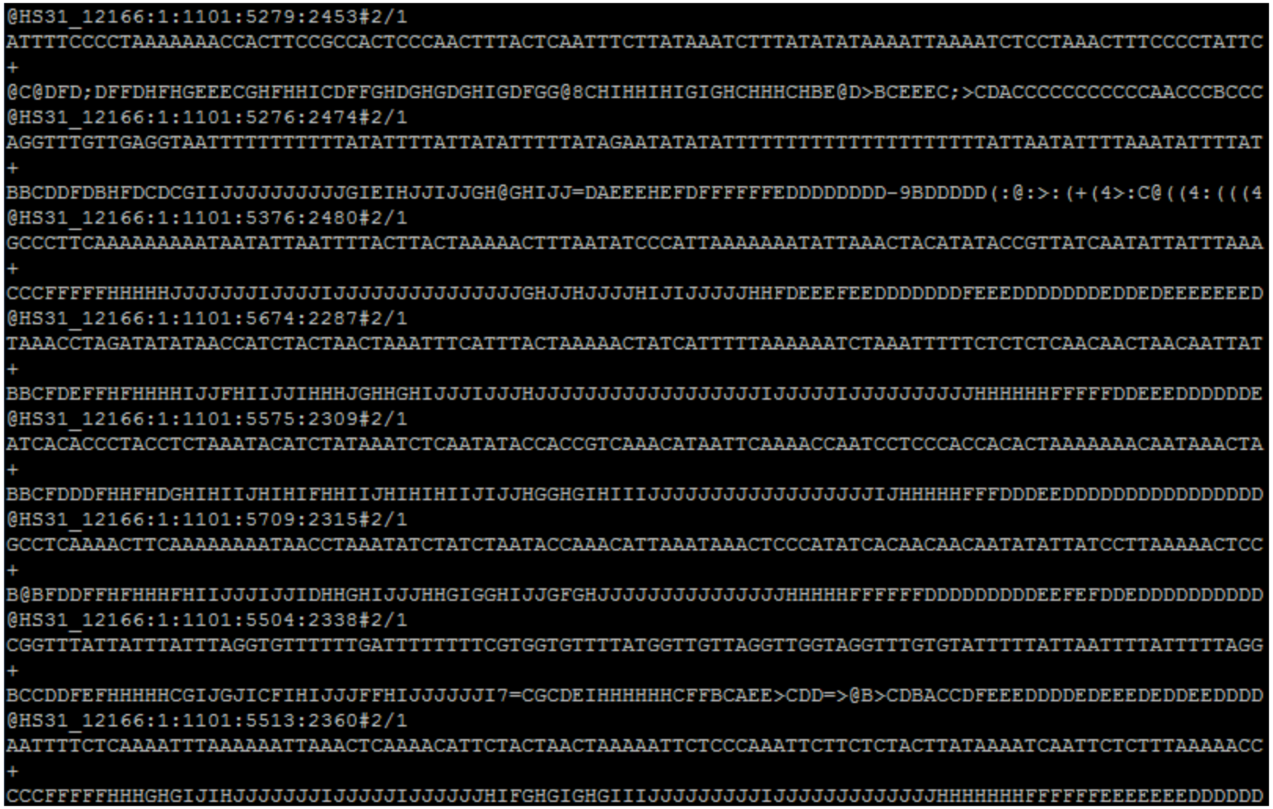
BS-Seq Analysis Workflow



QC → Trimming → Mapping → Methylation extraction → Mapped QC → Analysis

Babraham Bioinformatics

Raw Sequence Data



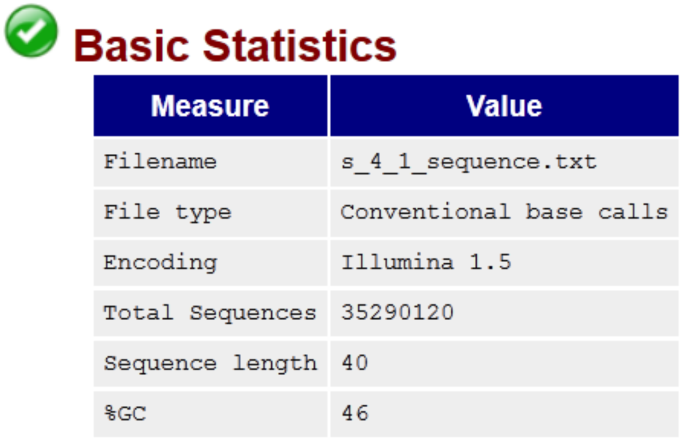
```
@HS31_12166:1:1101:5279:2453#2/1
ATTTTCCCCTAAAAAAACCACTTCCGCCACTCCCAACTTTACTCAATTTCTTATAAATCTTTATATATAAAATTAAAATCTCCTAAACTTTCCCCTATTC
+
@C@DFD;DFFDHFHGEEECGHFHHICDFFGHDGHGDGHIGDFGG@8CHIHHIHIGIGHCHHHCHBE@D>BCEEEC;>CDACCCCCCCCCCCAACCCBCCC
@HS31_12166:1:1101:5276:2474#2/1
AGGTTTGTTGAGGTAATTTTTTTTTTTATATTTTATTATATTTTTATAGAATATATATTTTTTTTTTTTTTTTTTTTTATTAATATTTTAAATATTTTAT
+
BBCDDFDBHFDCDCGIIJJJJJJJJJJGIEIHJJIJJGH@GHIJJ=DAEEEHEFDFFFFFFEDDDDDDDD-9BDDDDD(:@:>:(+(4>:C@((4:(((4
@HS31_12166:1:1101:5376:2480#2/1
GCCCTTCAAAAAAAAATAATATTAATTTTACTTACTAAAAACTTTAATATCCCATTAAAAAAATATTAAACTACATATACCGTTATCAATATTATTTAAA
+
CCCFFFFFHHHHHJJJJJJIJJJJIJJJJJJJJJJJJJJGHJJHJJJJHIJIJJJJJHHFDEEEFEEDDDDDDDFEEEDDDDDDDEDDEDEEEEEEED
@HS31_12166:1:1101:5674:2287#2/1
TAAACCTAGATATATAACCATCTACTAACTAAATTTCATTTACTAAAAACTATCATTTTTAAAAAATCTAAATTTTTCTCTCTCAACAACTAACAATTAT
+
BBCFDEFFHFHHHHIJJFHIIJJIHHHJGHHGHIJJJIJJJHJJJJJJJJJJJJJJJJJJIJJJJJIJJJJJJJJJHHHHHHFFFFFDDEEEDDDDDDE
@HS31_12166:1:1101:5575:2309#2/1
ATCACACCCTACCTCTAAATACATCTATAAATCTCAATATACCACCGTCAAACATAATTCAAAACCAATCCTCCCACCACACTAAAAAAACAATAAACTA
+
BBCFDDDFHHFHDGHIHIIJHIHIFHHIIJHIHIHIIJIJJHGGHGIHIIIJJJJJJJJJJJJJJJJJIJHHHHHFFFDDDEEDDDDDDDDDDDDDDDD
@HS31_12166:1:1101:5709:2315#2/1
GCCTCAAAACTTCAAAAAAAAATAACCTAAATATCTATCTAATACCAAACATTAAATAAACTCCCATATCACAACAACAATATATTATCCTTAAAAACTCC
+
B@BFDDFFHFHHHFHIIJJJIJJIDHHGHIJJJHHGIGGHIJJGFGHJJJJJJJJJJJJJJHHHHHFFFFFFDDDDDDDDDEEFEFDDEDDDDDDDDDD
@HS31_12166:1:1101:5504:2338#2/1
CGGTTTATTATTTATTTAGGTGTTTTTTGATTTTTTTTCGTGGTGTTTTATGGTTGTTAGGTTGGTAGGTTTGTGTATTTTTATTAATTTTATTTTTAGG
+
BCCDDFEFHHHHHCGIJGJICFIHIJJJFFHIJJJJJJI7=CGCDEIHHHHHHCFFBCAEE>CDD=>@B>CDBACCDFEEEDDDDEDEEEDEDDEEDDDD
@HS31_12166:1:1101:5513:2360#2/1
AATTTTCTCAAAATTTAAAAAATTAAACTCAAAACATTCTACTAACTAAAAATTCTCCCAAATTCTTCTCTACTTATAAAATCAATTCTCTTTAAAAACC
+
CCCFFFFFHHHGHGIJIHJJJJJJJIJJJJJIJJJJJJHIFGHGIGHGIIIJJJJJJJJJIJJJJJJJJJJJJHHHHHHHFFFFFFEEEEEEEDDDDDD
```

...

up to 1,000,000,000 lines per lane

Part I: Initial QC -
What does QC tell you about your library?

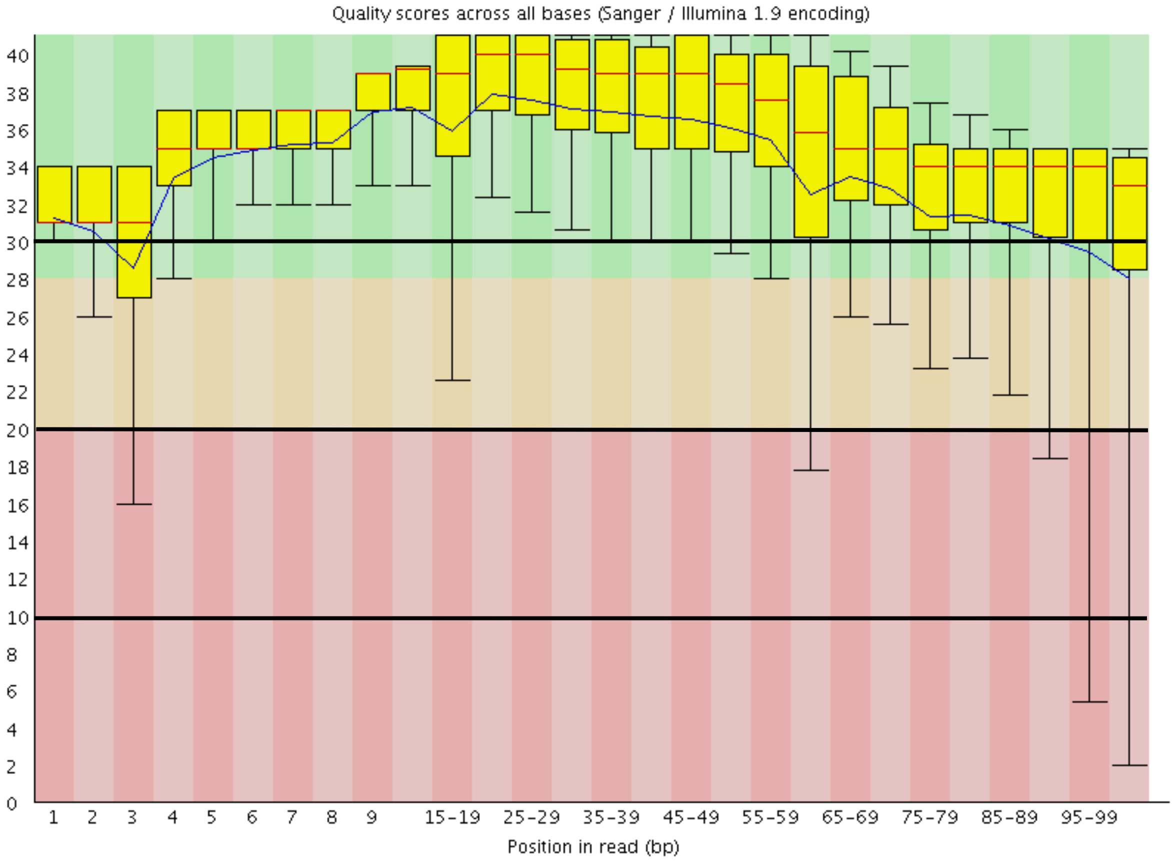
- # of sequences
- Basecall qualities
- Base composition
- Potential contaminants
- Expected duplication rate



✓ **Basic Statistics**

Measure	Value
Filename	s_4_1_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	35290120
Sequence length	40
%GC	46

QC Raw data: Sequence Quality



Error rate

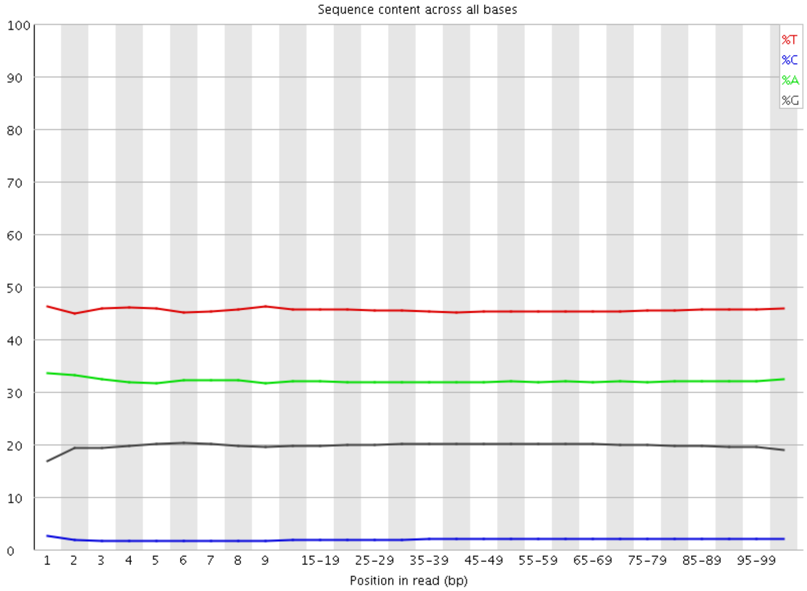
0.1%

1%

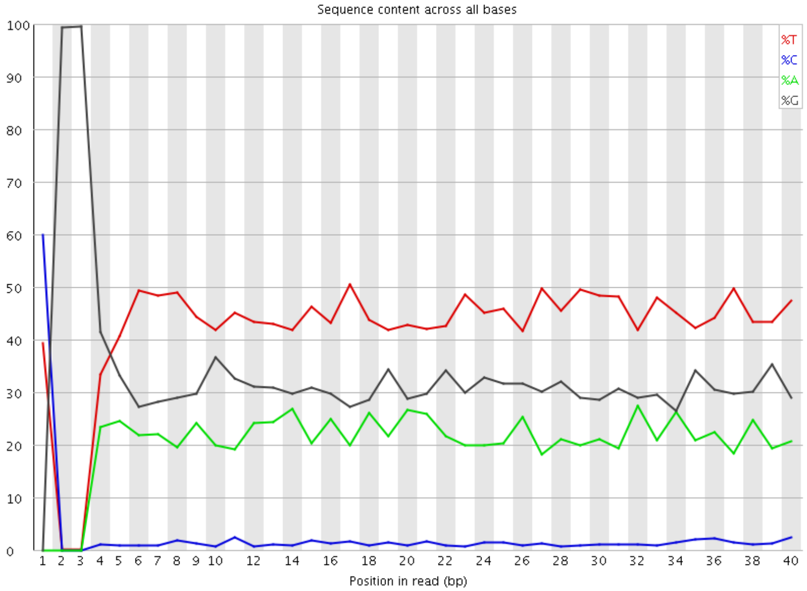
10%

QC: Base Composition

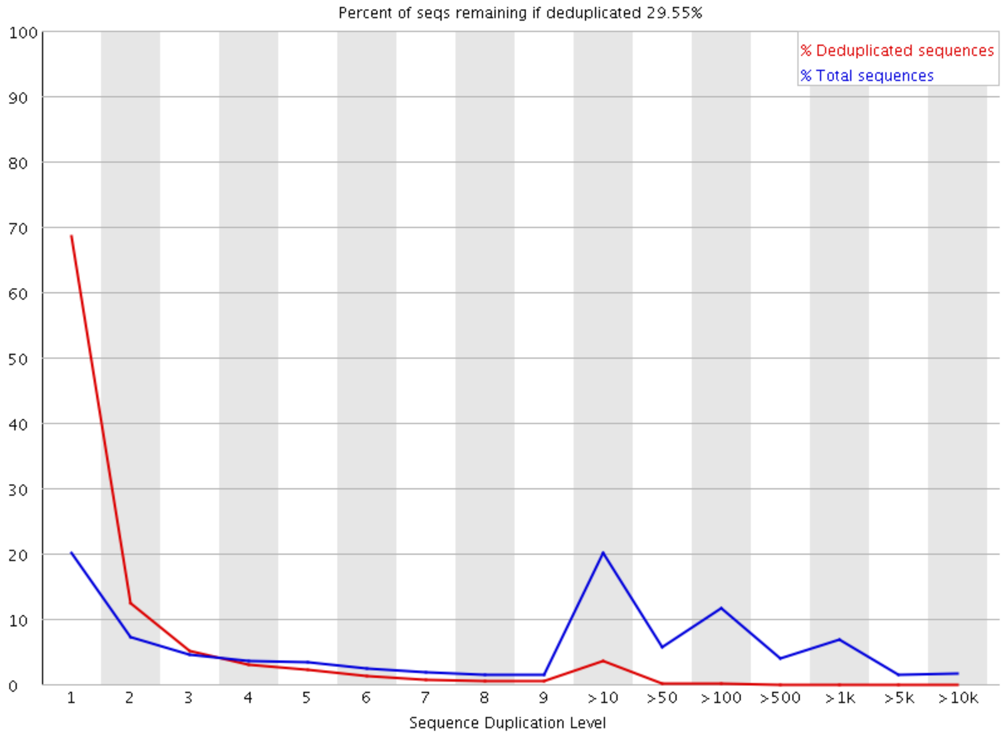
WGSBS



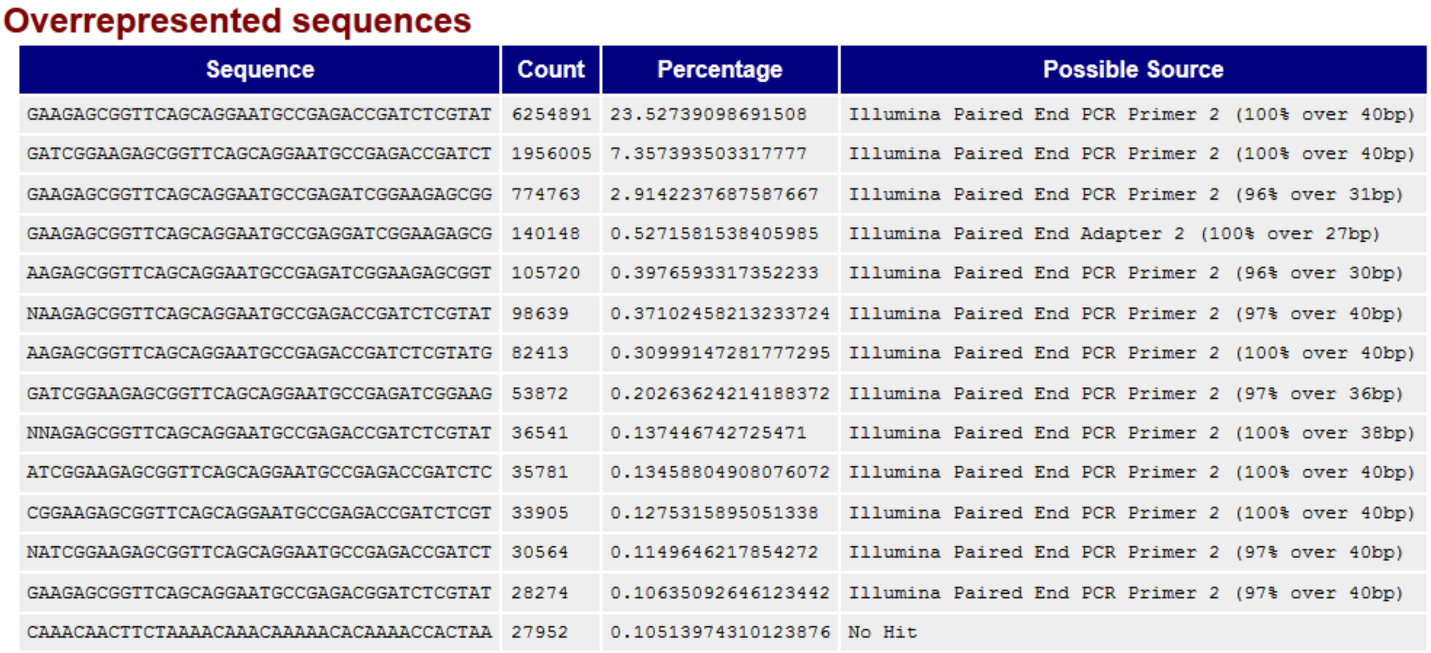
RRBS



QC: Duplication rate



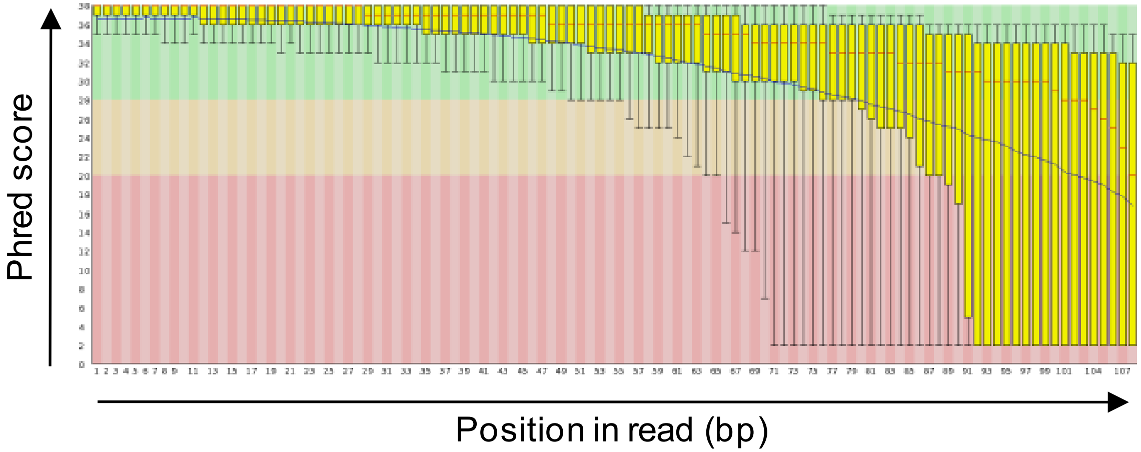
QC: Overrepresented sequences

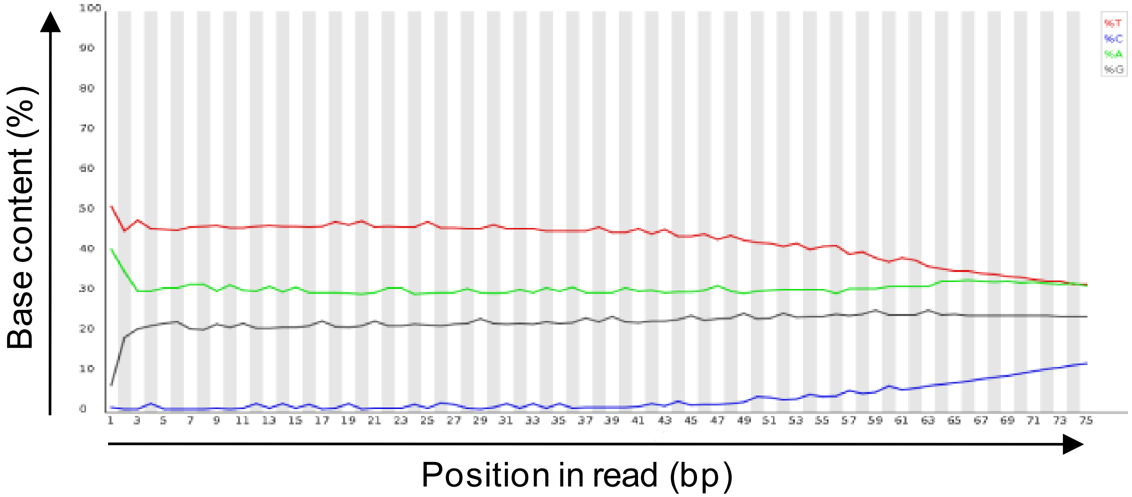


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	6254891	23.52739098691508	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	1956005	7.357393503317777	Illumina Paired End PCR Primer 2 (100% over 40bp)
GAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGG	774763	2.9142237687587667	Illumina Paired End PCR Primer 2 (96% over 31bp)
GAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAAGAGCG	140148	0.5271581538405985	Illumina Paired End Adapter 2 (100% over 27bp)
AAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGT	105720	0.3976593317352233	Illumina Paired End PCR Primer 2 (96% over 30bp)
NAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	98639	0.37102458213233724	Illumina Paired End PCR Primer 2 (97% over 40bp)
AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATG	82413	0.30999147281777295	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	53872	0.20263624214188372	Illumina Paired End PCR Primer 2 (97% over 36bp)
NNAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	36541	0.137446742725471	Illumina Paired End PCR Primer 2 (100% over 38bp)
ATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC	35781	0.13458804908076072	Illumina Paired End PCR Primer 2 (100% over 40bp)
CGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	33905	0.1275315895051338	Illumina Paired End PCR Primer 2 (100% over 40bp)
NATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	30564	0.1149646217854272	Illumina Paired End PCR Primer 2 (97% over 40bp)
GAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTAT	28274	0.10635092646123442	Illumina Paired End PCR Primer 2 (97% over 40bp)
CAAACAACTTCTAAAACAAACAAAAACACAAAACCACTAA	27952	0.10513974310123876	No Hit

Common problems in BS-Seq





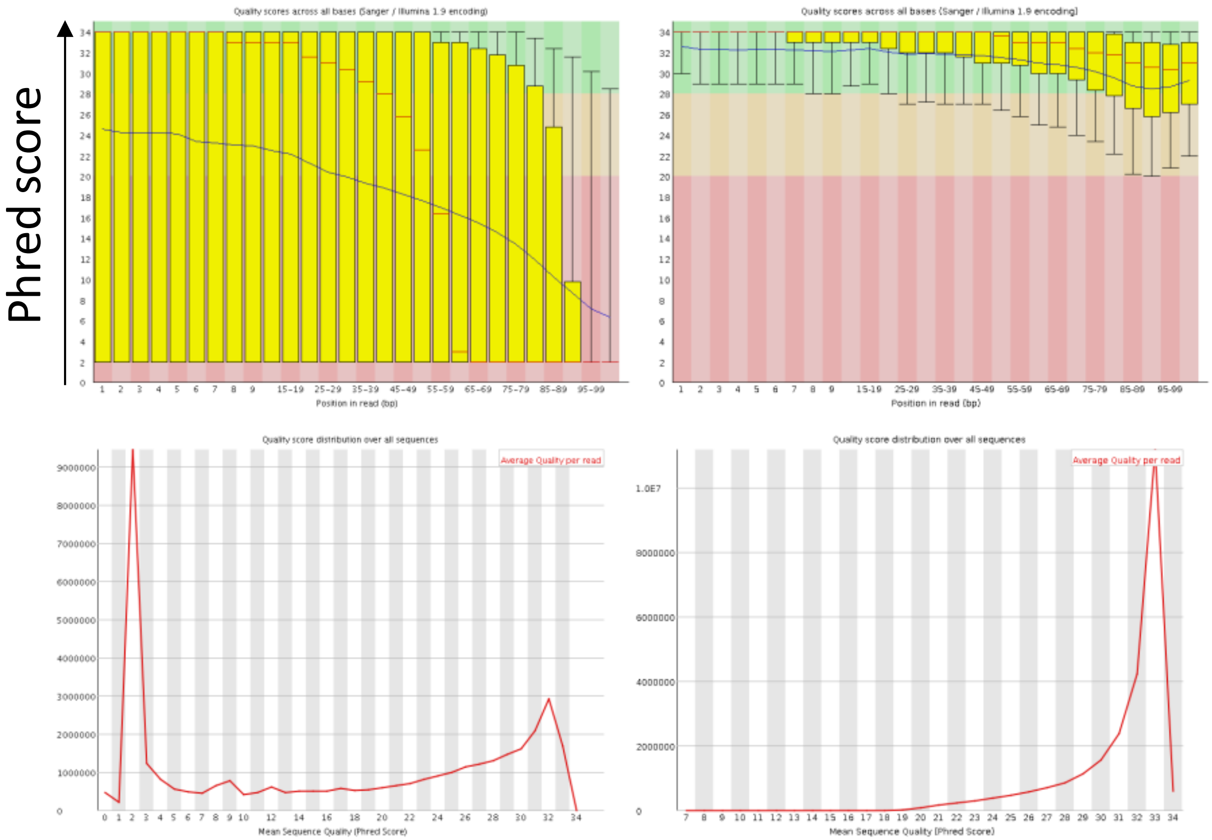
Not observed
in ‘normal’ libraries,
e.g. ChIP or RNA-Seq

Babraham
Bioinformatics

Removing poor quality basecalls

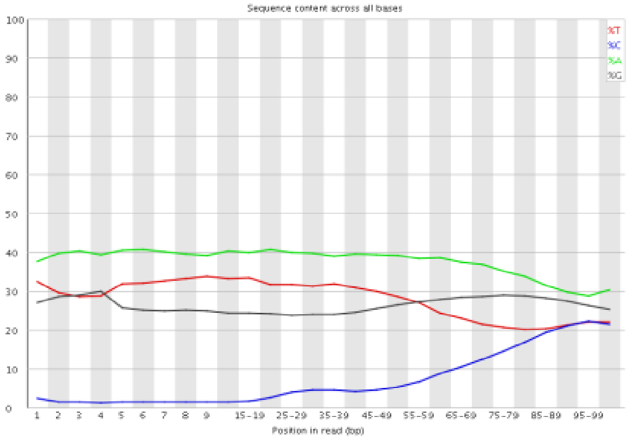
before trimming

after trimming

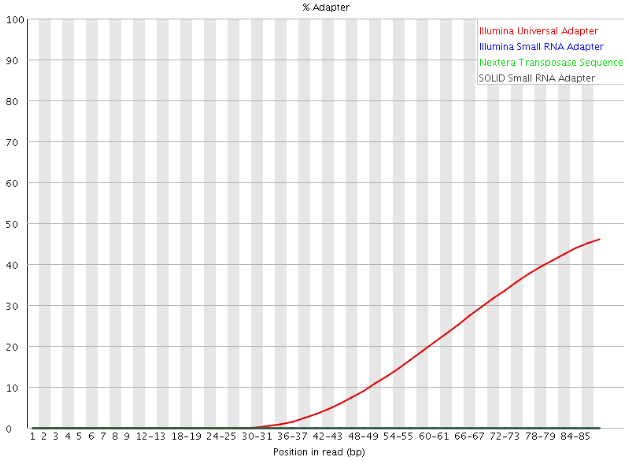


Removing adapter contamination

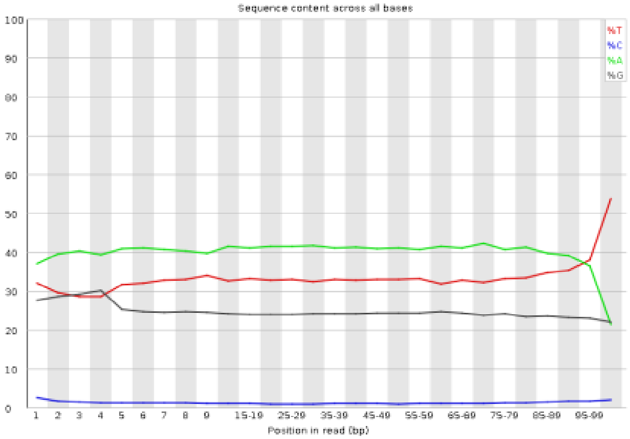
before trimming



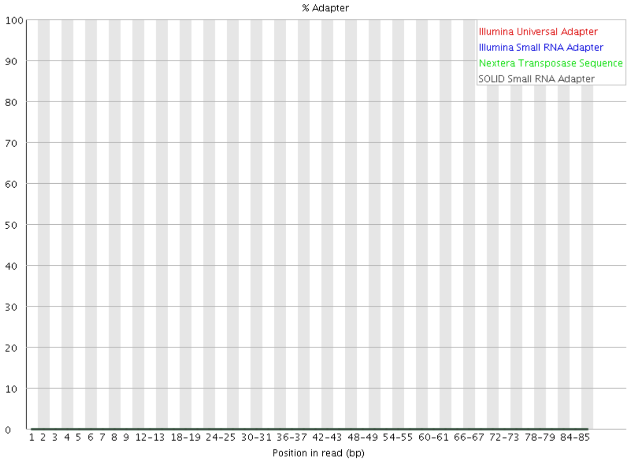
✗ Adapter Content



after trimming



✓ Adapter Content



Babraham Bioinformatics

Adapter trimming

(Illumina adapter: AGATCGGAAGAGC)

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGAT
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGAT

partial match full match

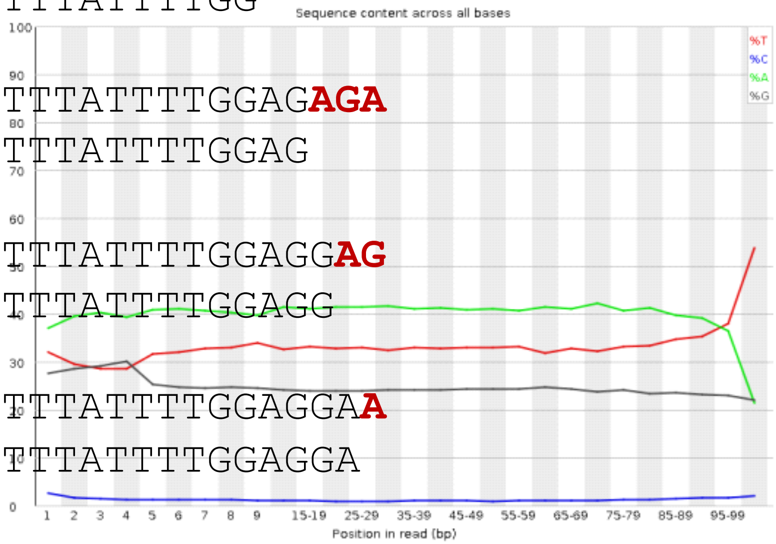
B: **AGATC**TTTTATTCGGTAGGAT**AGATCGGAAGAGC**XXXXXXXXXXXXXXX
A: AGATCTTTTATTCGGTAGGAT

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGG**AGATC**
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGG

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAG**AGA**
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAG

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGG**AG**
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGG

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGA**A**
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGA



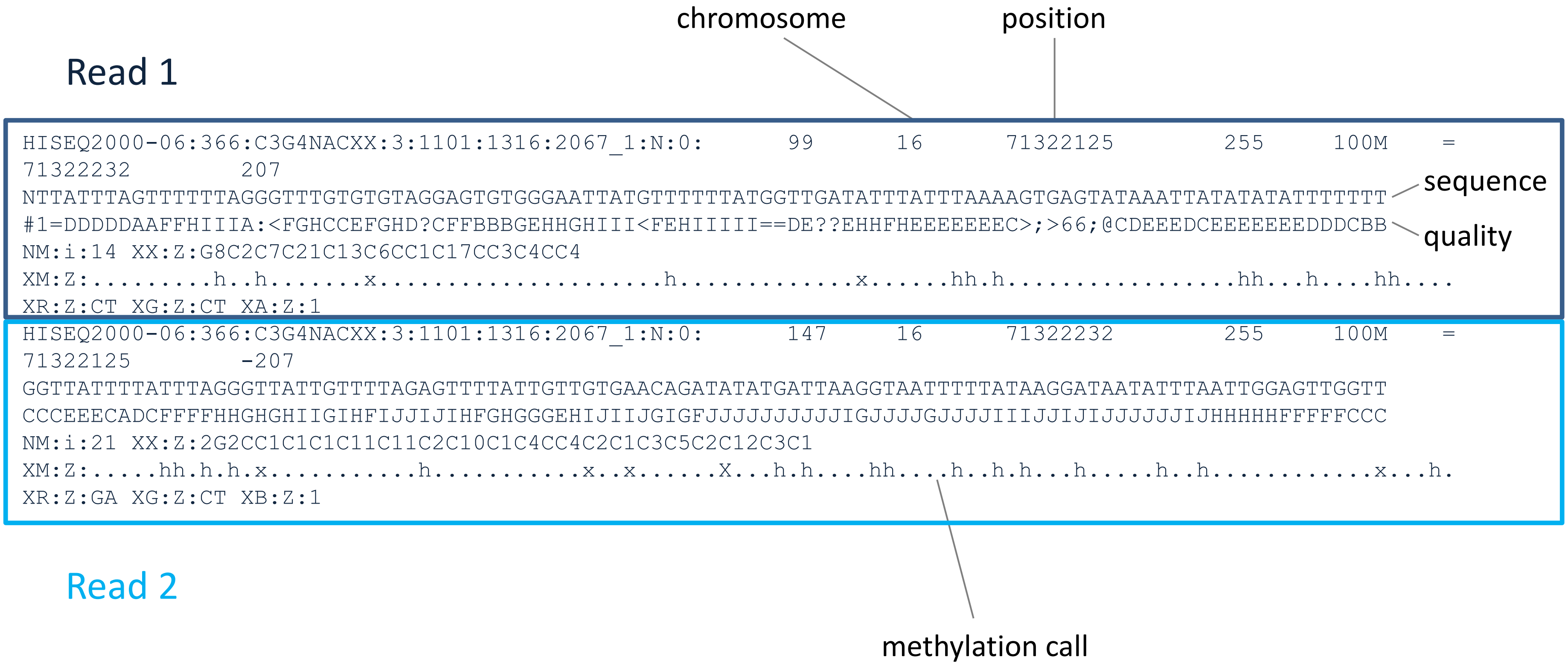
Babraham Bioinformatics

Summary Adapter/Quality Trimming

Important to trim because failure to do so might result in:

- Low mapping efficiency
- Mis-alignments
- Errors in methylation calls since adapters are methylated
- Basecall errors tend toward 50% (C:mC)

Part II: Sequence alignment –
Bismark primary alignment output (BAM file)



Read 1

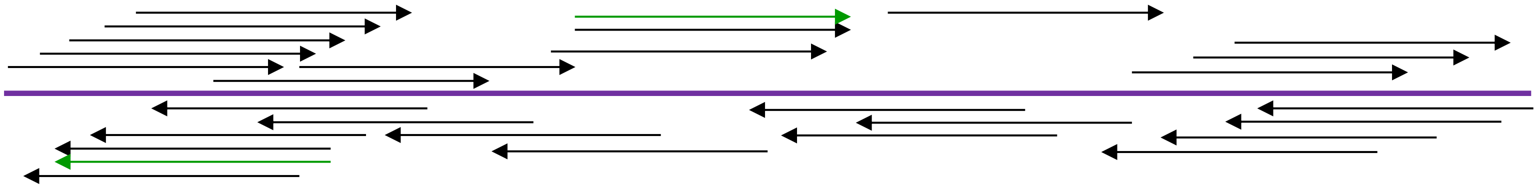
```
HISEQ2000-06:366:C3G4NACXX:3:1101:1316:2067_1:N:0:      99      16      71322125        255     100M    =
71322232        207
NTTATTTAGTTTTTTAGGGTTTGTGTGTAGGAGTGTGGGAATTATGTTTTTTATGGTTGATATTTATTTAAAAGTGAGTATAAATTATATATATTTTTTT
#1=DDDDDAAFFHIIIA:<FGHCCEFGHD?CFFBBBGEHHGHIII<FEHIIIII==DE??EHHFHEEEEEEEC>;>66;@CDEEEDCEEEEEEEEDDDCBB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

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

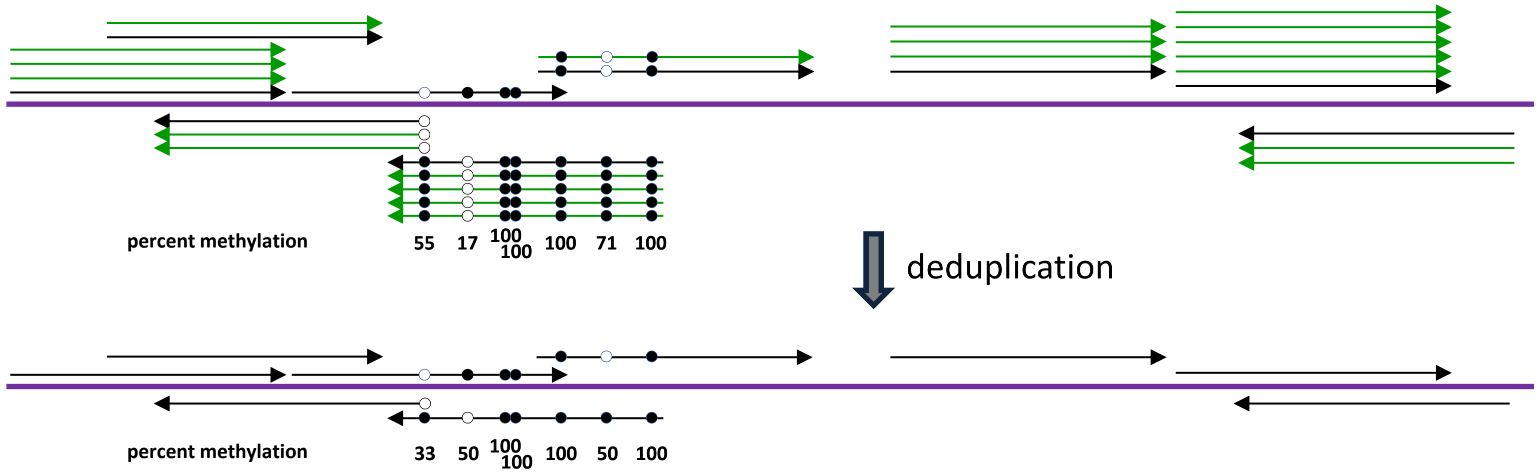
Read 2

Sequence duplication

Complex/diverse library:



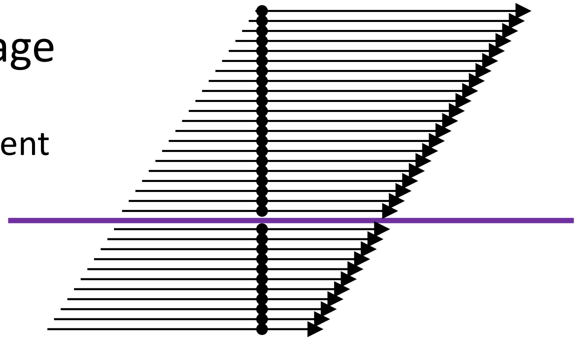
Duplicated library:



Deduplication - considerations

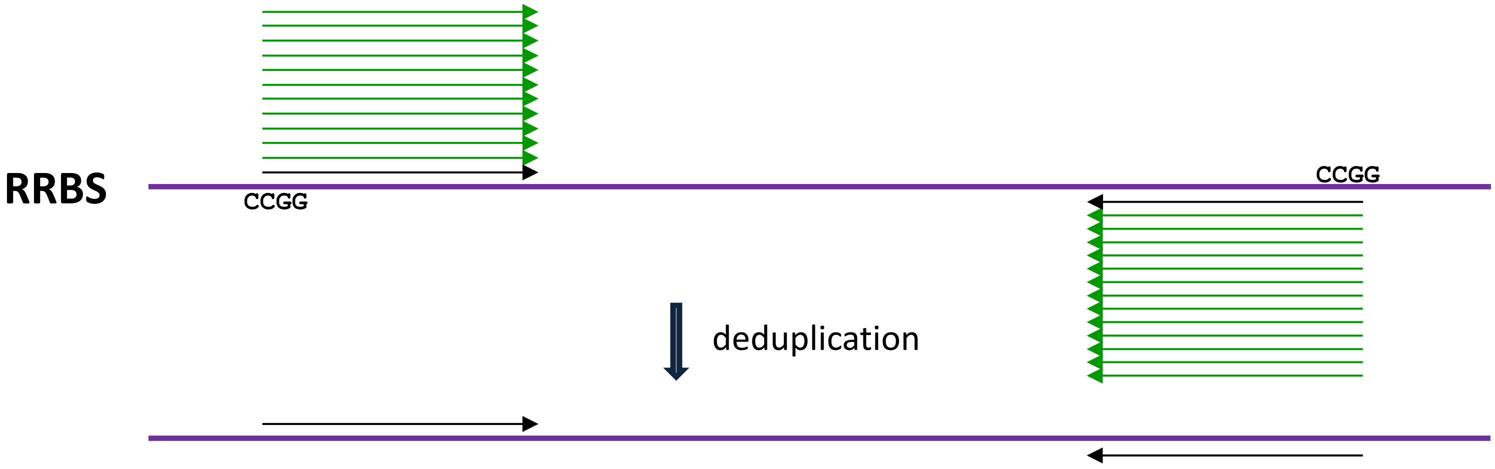
Advisable for large genomes and moderate coverage

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

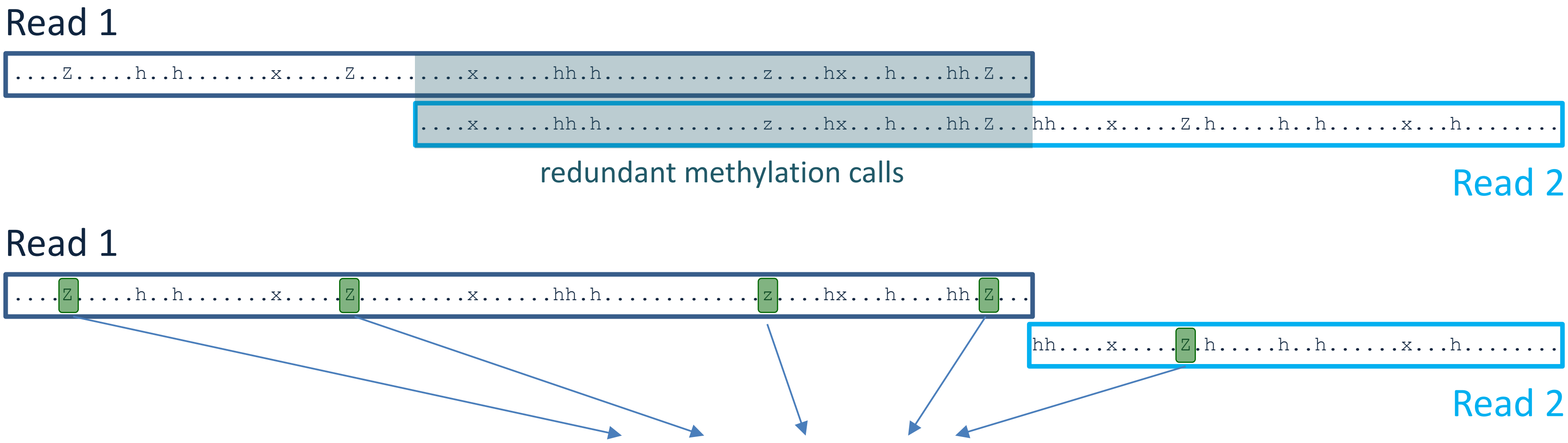


NOT advisable for RRBS or other target enrichment methods

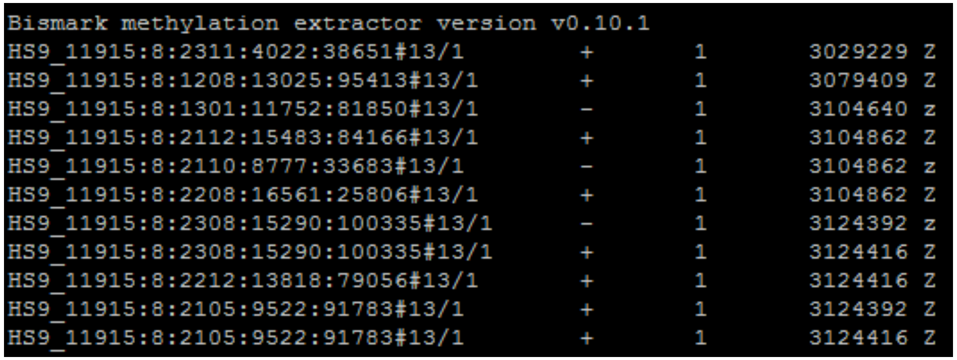
where higher coverage is either desired or expected



Methylation extraction


Read 1
....Z.....h..h.......x.....Z.........x......hh.h.............z....hx...h....hh.Z...
....x......hh.h.............z....hx...h....hh.Z...hh....x.....Z.h.....h..h......x...h........
redundant methylation calls
Read 2
Read 1
....Z.....h..h.......x.....Z.........x......hh.h.............z....hx...h....hh.Z...
hh....x.....Z.h.....h..h......x...h........
Read 2

CpG methylation output



```
Bismark methylation extractor version v0.10.1
HS9_11915:8:2311:4022:38651#13/1        +       1       3029229 Z
HS9_11915:8:1208:13025:95413#13/1       +       1       3079409 Z
HS9_11915:8:1301:11752:81850#13/1       -       1       3104640 z
HS9_11915:8:2112:15483:84166#13/1       +       1       3104862 Z
HS9_11915:8:2110:8777:33683#13/1        -       1       3104862 z
HS9_11915:8:2208:16561:25806#13/1       +       1       3104862 Z
HS9_11915:8:2308:15290:100335#13/1      -       1       3124392 z
HS9_11915:8:2308:15290:100335#13/1      +       1       3124416 Z
HS9_11915:8:2212:13818:79056#13/1       +       1       3124416 Z
HS9_11915:8:2105:9522:91783#13/1        +       1       3124392 Z
HS9_11915:8:2105:9522:91783#13/1        +       1       3124416 Z
```

read ID

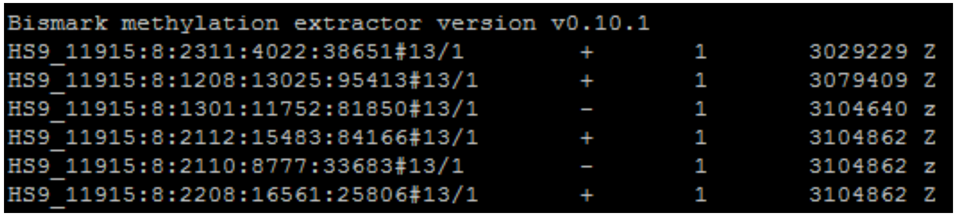
meth state

chr

pos

context

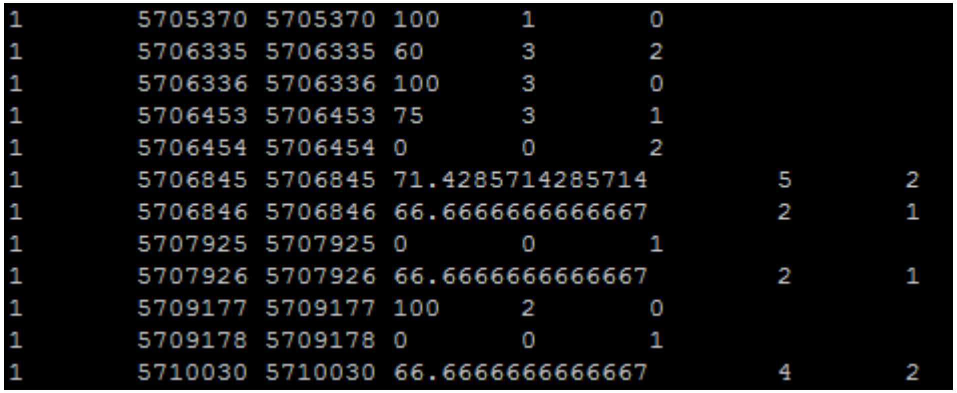
Methylation extraction I



```
Bismark methylation extractor version v0.10.1
HS9_11915:8:2311:4022:38651#13/1        +       1       3029229 Z
HS9_11915:8:1208:13025:95413#13/1       +       1       3079409 Z
HS9_11915:8:1301:11752:81850#13/1       -       1       3104640 z
HS9_11915:8:2112:15483:84166#13/1       +       1       3104862 Z
HS9_11915:8:2110:8777:33683#13/1        -       1       3104862 z
HS9_11915:8:2208:16561:25806#13/1       +       1       3104862 Z
```

CpG methylation output

bismark2bedGraph

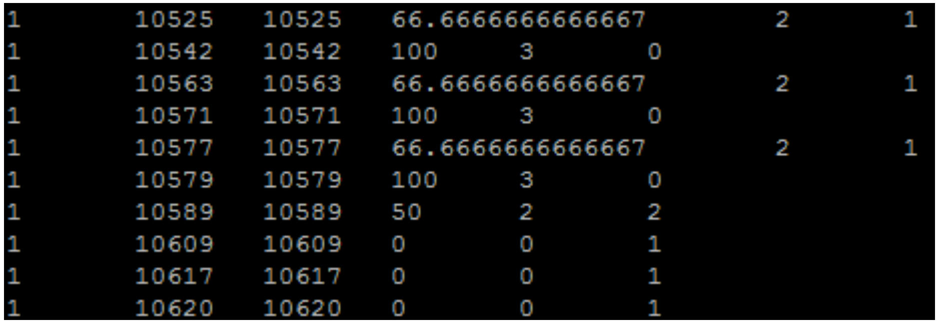


```
1       5705370 5705370 100     1       0
1       5706335 5706335 60      3       2
1       5706336 5706336 100     3       0
1       5706453 5706453 75      3       1
1       5706454 5706454 0       0       2
1       5706845 5706845 71.4285714285714        5       2
1       5706846 5706846 66.6666666666667        2       1
1       5707925 5707925 0       0       1
1       5707926 5707926 66.6666666666667        2       1
1       5709177 5709177 100     2       0
1       5709178 5709178 0       0       1
1       5710030 5710030 66.6666666666667        4       2
```

chr pos methylation percentage meth unmeth

bedGraph/coverage output

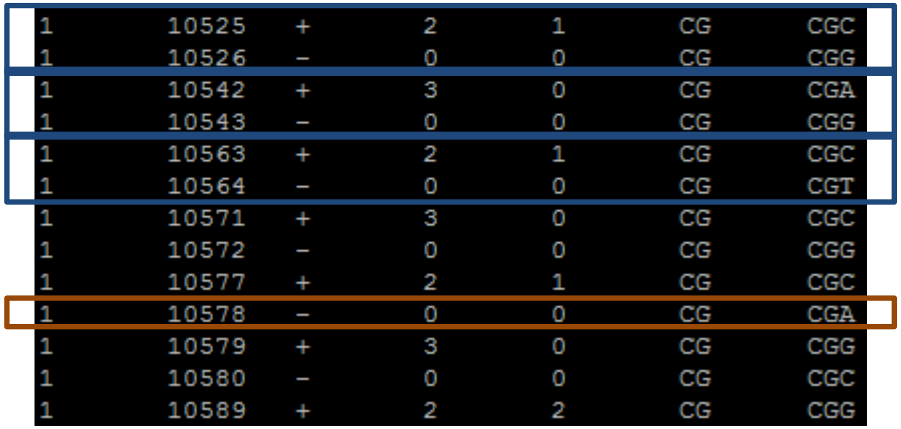
Methylation extraction II



```
1       10525   10525   66.6666666666667        2       1
1       10542   10542   100     3       0
1       10563   10563   66.6666666666667        2       1
1       10571   10571   100     3       0
1       10577   10577   66.6666666666667        2       1
1       10579   10579   100     3       0
1       10589   10589   50      2       2
1       10609   10609   0       0       1
1       10617   10617   0       0       1
1       10620   10620   0       0       1
```

coverage output

coverage2cytosine



```
1       10525   +       2       1       CG      CGC
1       10526   -       0       0       CG      CGG
1       10542   +       3       0       CG      CGA
1       10543   -       0       0       CG      CGG
1       10563   +       2       1       CG      CGC
1       10564   -       0       0       CG      CGT
1       10571   +       3       0       CG      CGC
1       10572   -       0       0       CG      CGG
1       10577   +       2       1       CG      CGC
1       10578   -       0       0       CG      CGA
1       10579   +       3       0       CG      CGG
1       10580   -       0       0       CG      CGC
1       10589   +       2       2       CG      CGG
```

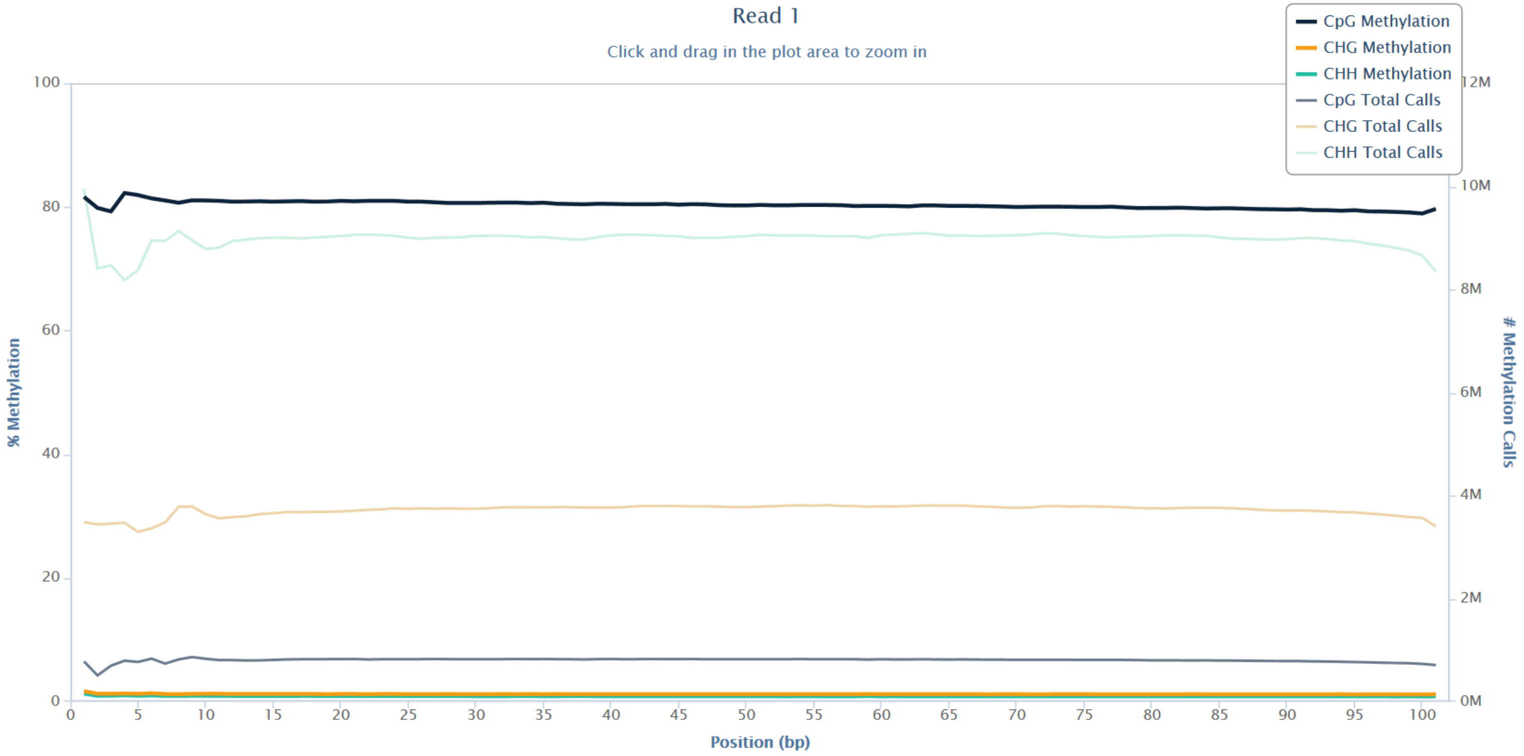
chr pos strand meth unmeth di-nuc tri-nuc

optional: merge into
CpG dinucleotide entities

Genome wide CpG report

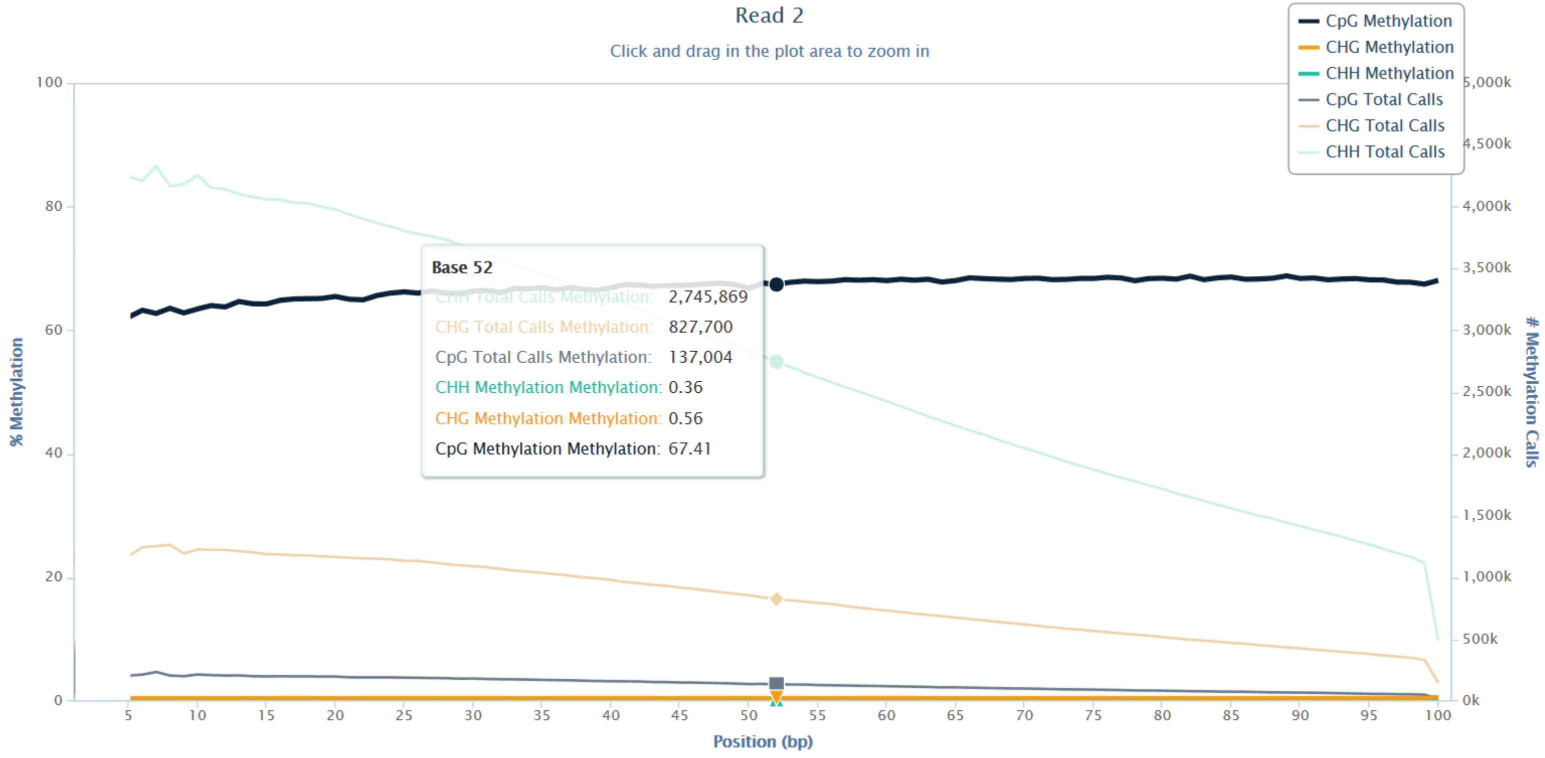
Part III: Mapped QC - Methylation bias

M-Bias Plot



good opportunity to look at conversion efficiency

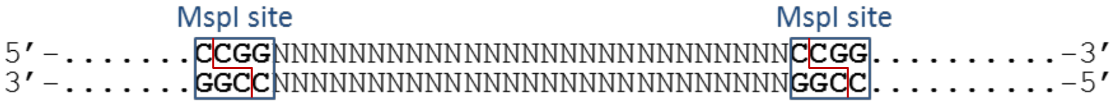
Artificial methylation calls in paired-end libraries



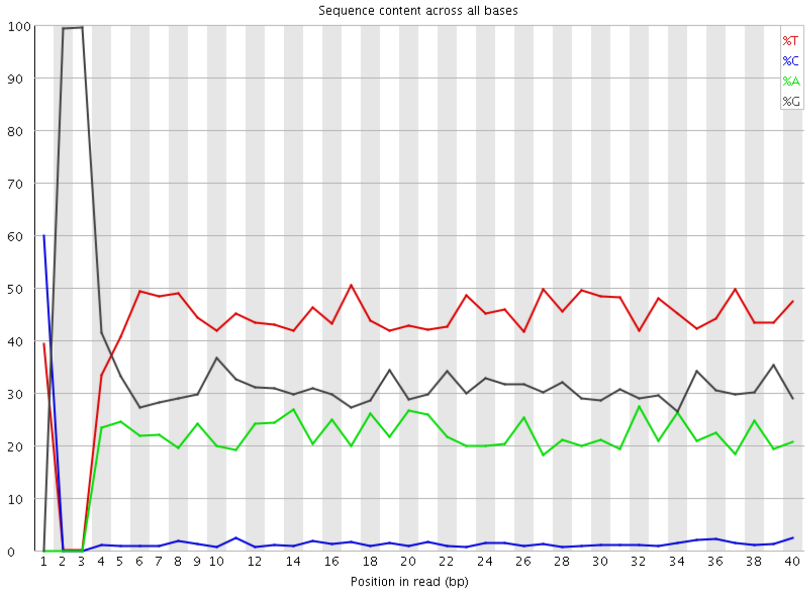
end repair + A-tailing

```
5′-   GGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCA   -3′
3′-  ACCCNNNNNNNNNNNNNNNNNNNNNNNNNNNGGG    -5′
```

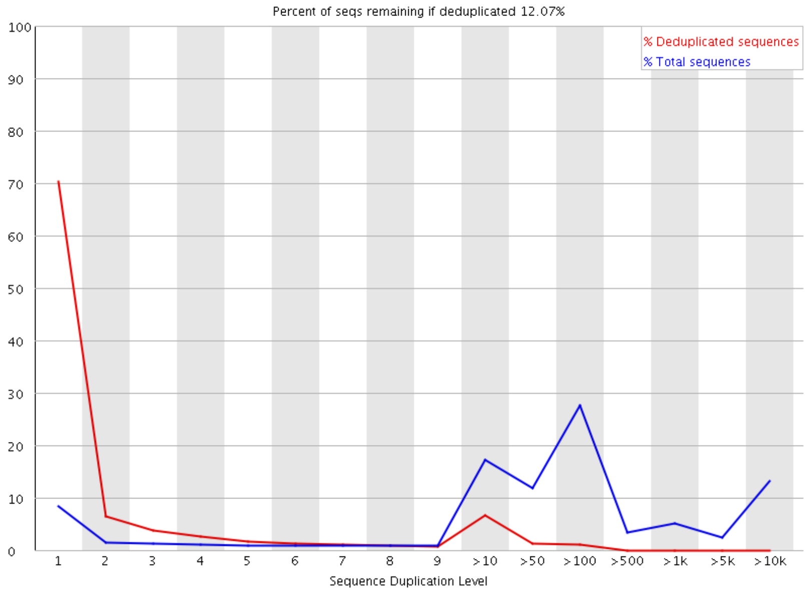
Specialist applications (I):
Reduced representation BS-Seq (RRBS)


MspI site
MspI site
5′–........CCGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGG...........–3′
3′–........GGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGCC...........–5′

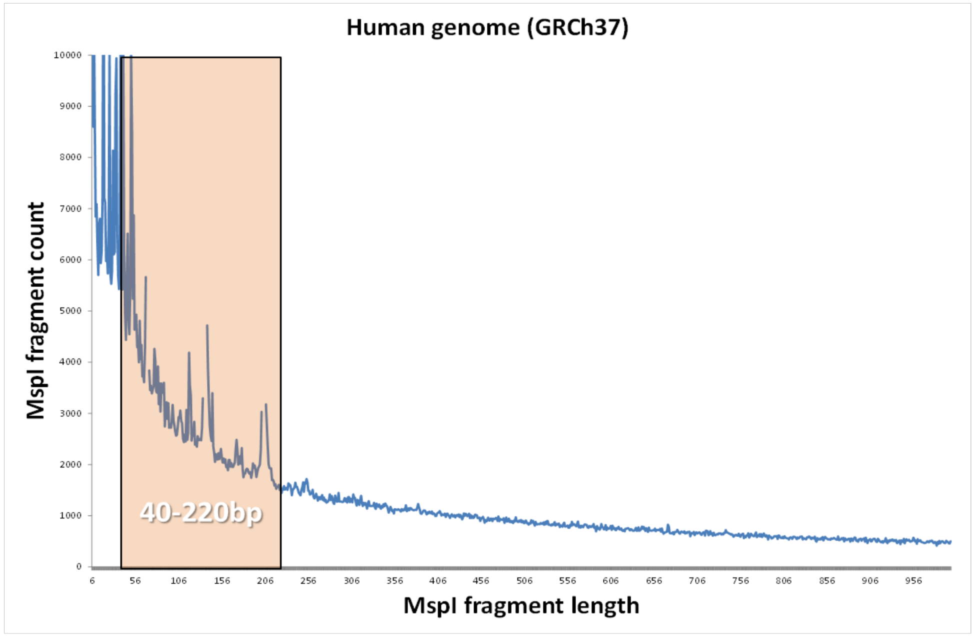
Sequence composition bias

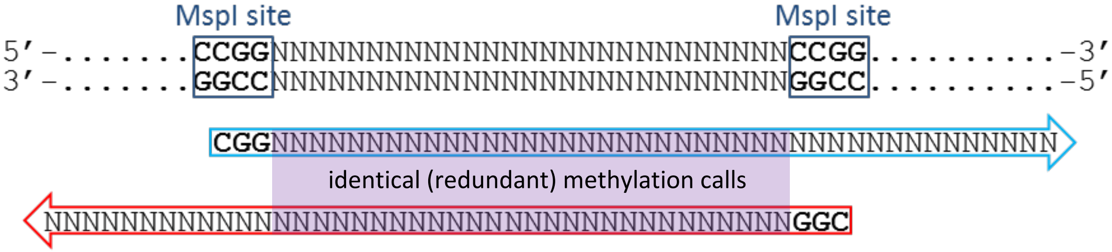


High duplication rate



Fragment size distribution in RRBS




MspI site

MspI site

5′ –........CCGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGG...........–3′
3′ –........GGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNGGCC...........–5′

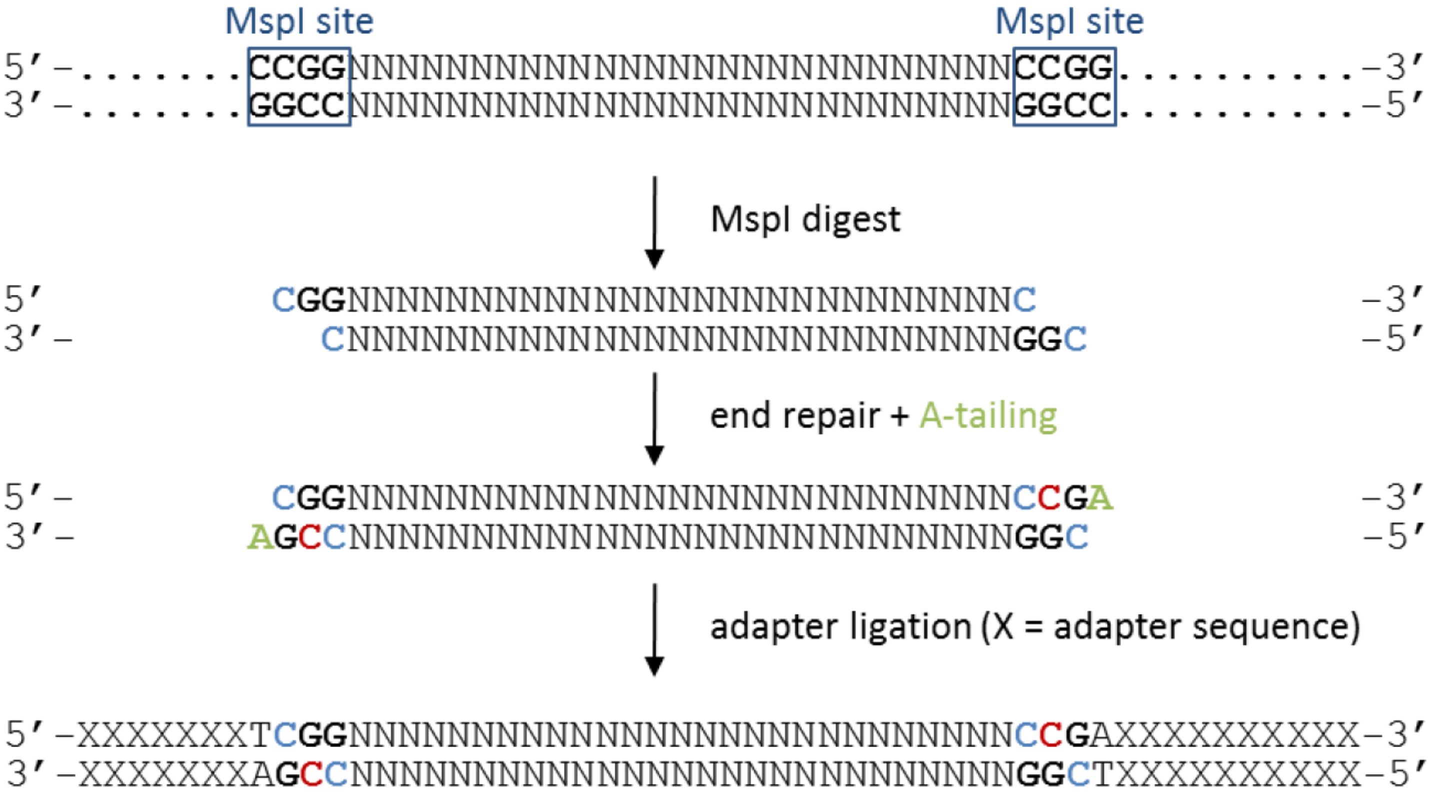
CGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

identical (redundant) methylation calls

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGC

Babraham Bioinformatics

Artificial methylation calls in RRBS libraries



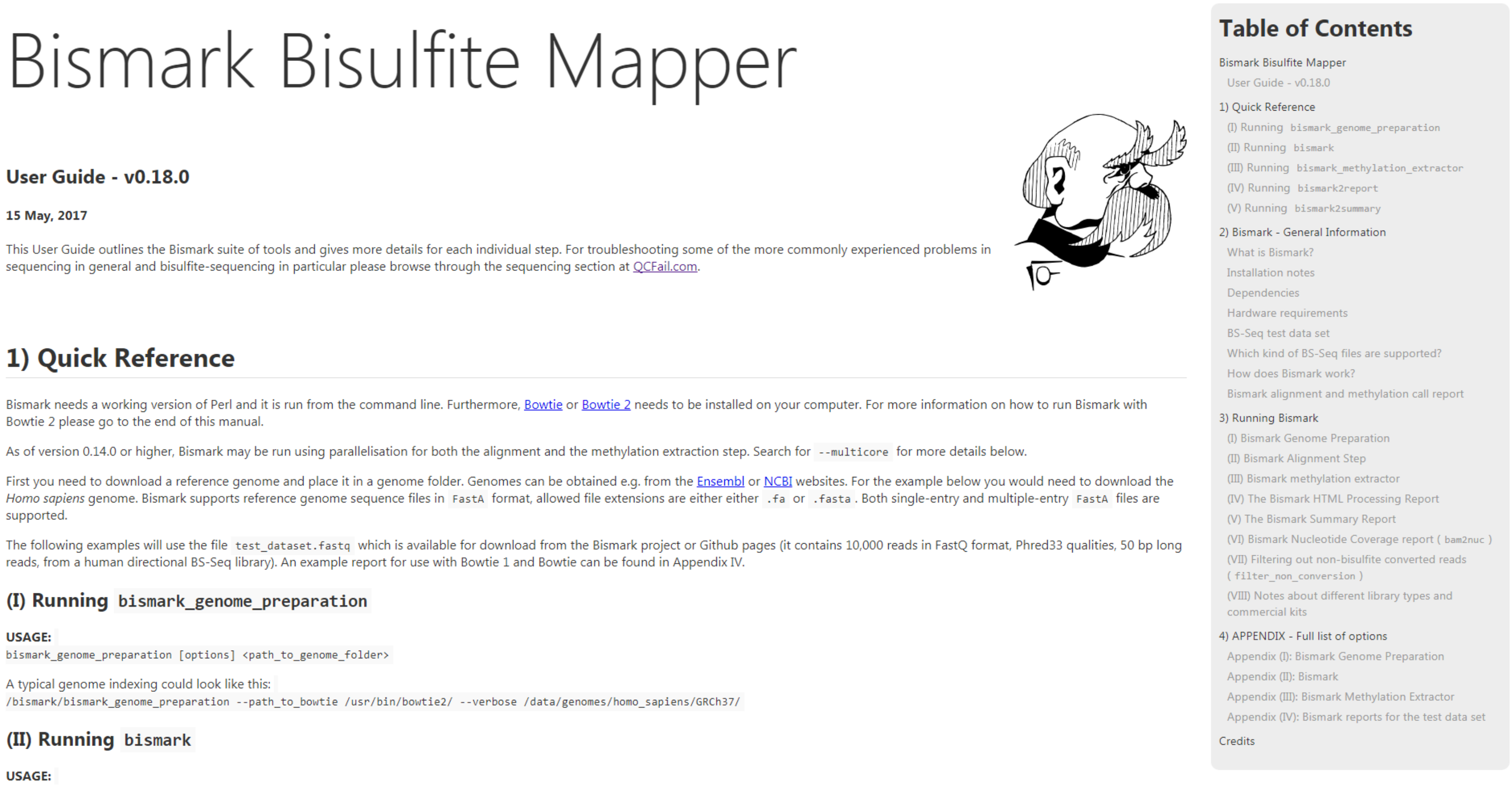
```
                 MspI site                       MspI site
5'–.......CCGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGG..........–3'
3'–.......GGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNGGCC..........–5'

                      MspI digest

5'         CGGNNNNNNNNNNNNNNNNNNNNNNNNNNNC              –3'
3'–          CNNNNNNNNNNNNNNNNNNNNNNNNNNNGGC            –5'

                      end repair + A-tailing

5'–        CGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGA           –3'
3'–       AGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNGGC            –5'

                      adapter ligation (X = adapter sequence)

5'–XXXXXXXTCGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGAXXXXXXXXXX–3'
3'–XXXXXXXAGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNGGCTXXXXXXXXXX–5'
```

C genomic cytosine

C unmethylated cytosine

Bismark User Guide

https://rawgit.com/FelixKrueger/Bismark/master/Docs/Bismark_User_Guide.html



Bismark Bisulfite Mapper

User Guide - v0.18.0

15 May, 2017

This User Guide outlines the Bismark suite of tools and gives more details for each individual step. For troubleshooting some of the more commonly experienced problems in sequencing in general and bisulfite-sequencing in particular please browse through the sequencing section at QCFail.com.

1) Quick Reference

Bismark needs a working version of Perl and it is run from the command line. Furthermore, Bowtie or Bowtie 2 needs to be installed on your computer. For more information on how to run Bismark with Bowtie 2 please go to the end of this manual.

As of version 0.14.0 or higher, Bismark may be run using parallelisation for both the alignment and the methylation extraction step. Search for `--multicore` for more details below.

First you need to download a reference genome and place it in a genome folder. Genomes can be obtained e.g. from the Ensembl or NCBI websites. For the example below you would need to download the *Homo sapiens* genome. Bismark supports reference genome sequence files in `FastA` format, allowed file extensions are either either `.fa` or `.fasta`. Both single-entry and multiple-entry `FastA` files are supported.

The following examples will use the file `test_dataset.fastq` which is available for download from the Bismark project or Github pages (it contains 10,000 reads in FastQ format, Phred33 qualities, 50 bp long reads, from a human directional BS-Seq library). An example report for use with Bowtie 1 and Bowtie can be found in Appendix IV.

(I) Running `bismark_genome_preparation`

USAGE:

```
bismark_genome_preparation [options] <path_to_genome_folder>
```

A typical genome indexing could look like this:

```
/bismark/bismark_genome_preparation --path_to_bowtie /usr/bin/bowtie2/ --verbose /data/genomes/homo_sapiens/GRCh37/
```

(II) Running `bismark`

USAGE:

Table of Contents

Babraham Bioinformatics

Bismark workflow

Pre Alignment

FastQC	Initial quality control
Trim Galore	Adapter/quality trimming using Cutadapt; handles RRBS and paired-end reads; Trim Galore and RRBS User guide

Alignment

Bismark	Output BAM

Post Alignment

Deduplication	optional
Methylation extractor	Output individual cytosine methylation calls; optionally bedGraph or genome-wide cytosine report
	M-bias analysis
bismark2report	Graphical HTML report generation

Example: http://www.bioinformatics.babraham.ac.uk/projects/bismark/PE_report.html

Epigenesys protocol: *Quality Control, trimming and alignment of Bisulfite-Seq data*

Useful links

- **FastQC** www.bioinformatics.babraham.ac.uk/projects/fastqc/
- **Trim Galore** www.bioinformatics.babraham.ac.uk/projects/trim_galore/
- **Cutadapt** https://code.google.com/p/cutadapt/
- **Bismark** www.bioinformatics.babraham.ac.uk/projects/bismark/
- **Bowtie** http://bowtie-bio.sourceforge.net/
- **Bowtie 2** http://bowtie-bio.sourceforge.net/bowtie2/
- **SeqMonk** www.bioinformatics.babraham.ac.uk/projects/seqmonk/
- **Cluster Flow** www.bioinformatics.babraham.ac.uk/projects/clusterflow/

Epigenesys **protocol:** *Quality control, trimming and alignment of Bisulfite-Seq data*

http://www.epigenesys.eu/en/protocols/bio-informatics/483-quality-control-trimming-and-alignment-of-bisulfite-seq-data-prot-57

QCFAIL.com https://sequencing.qcfail.com/

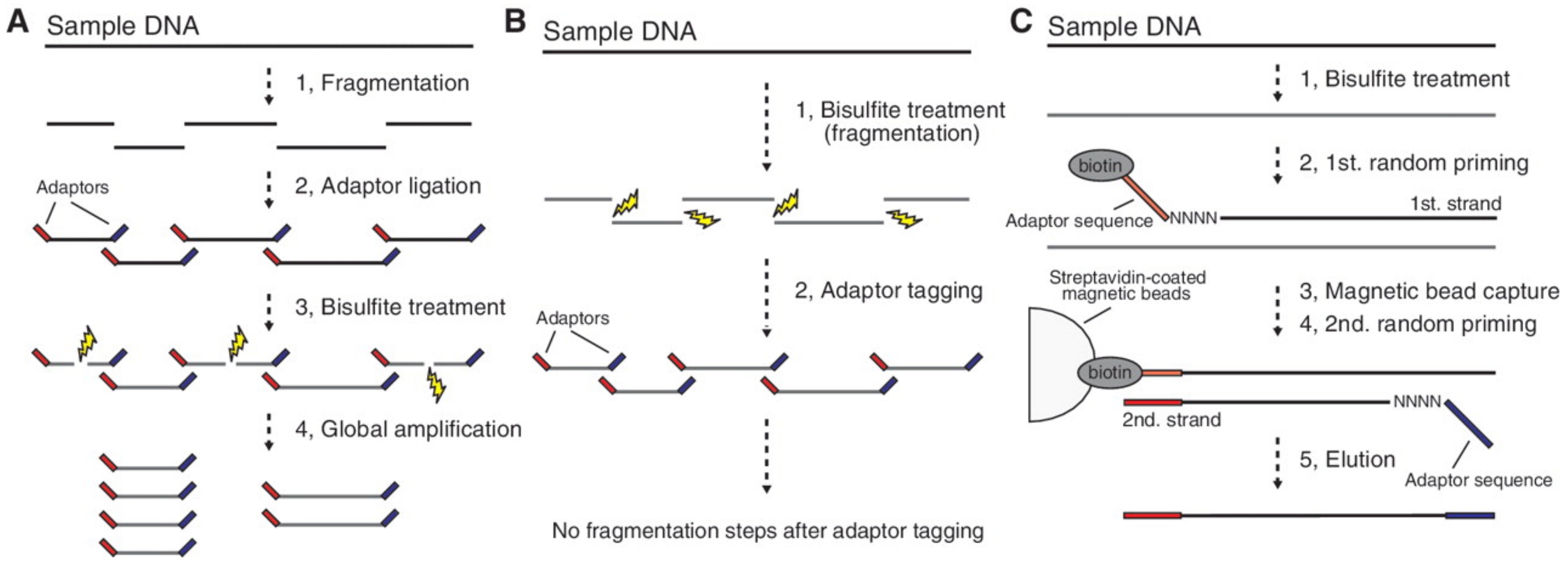
Thank you for your attention

Questions?

UNIVERSITY OF CAMBRIDGE AstraZeneca Babraham Bioinformatics

Specialist application (II):
Post-bisulfite adapter tagging (PBAT)

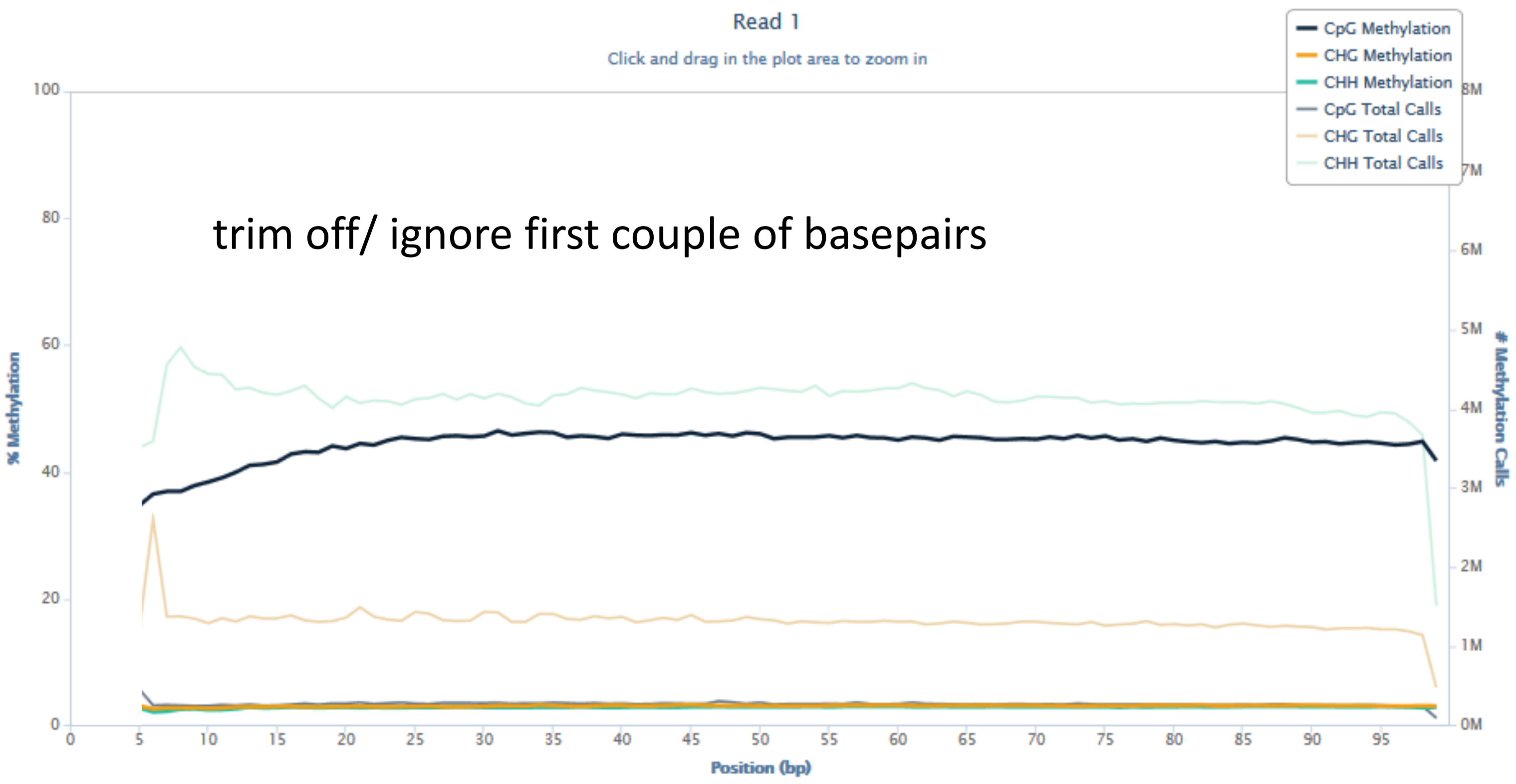
WGBS **PBAT**



→ suitable for low input material

PBAT-Seq

M-Bias Plot



trim off/ ignore first couple of basepairs