Introduction to methylation analysis

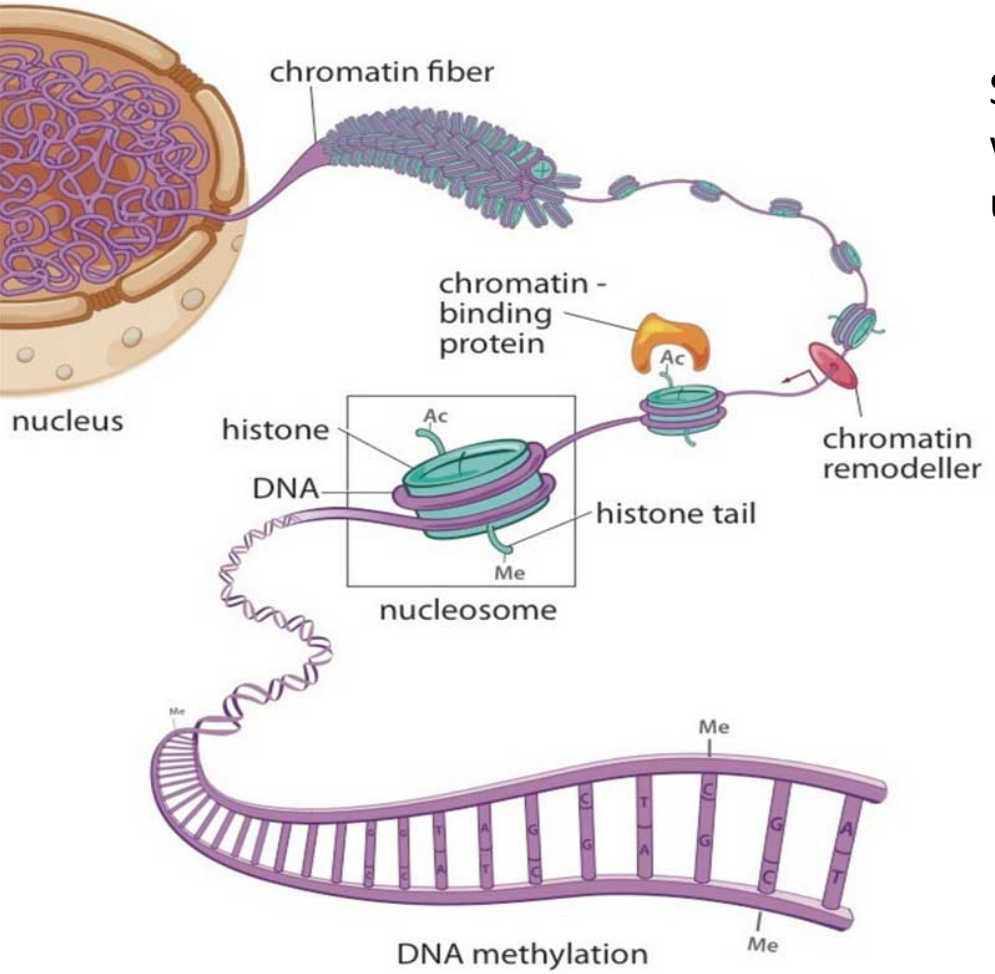
Data processing, QC and alignment

Sergio Martínez Cuesta
sermarcue@gmail.com

Employment:

Materials from:

Epigenetics



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

From The Cell Biology of Stem Cells (2010)

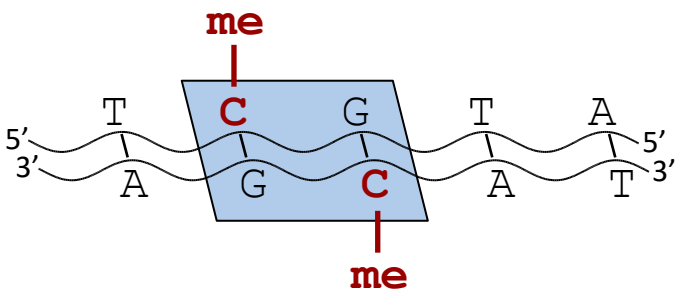
Babraham Bioinformatics

Types of DNA methylation

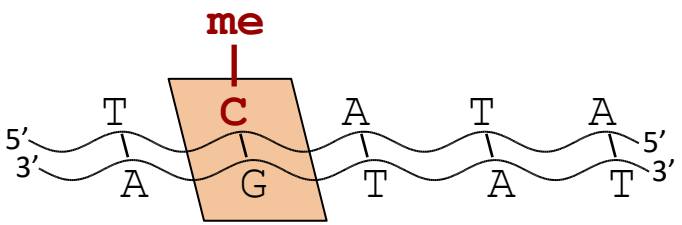
canonical

	Plants	Mammals
CG	symmetric	symmetric

CG context

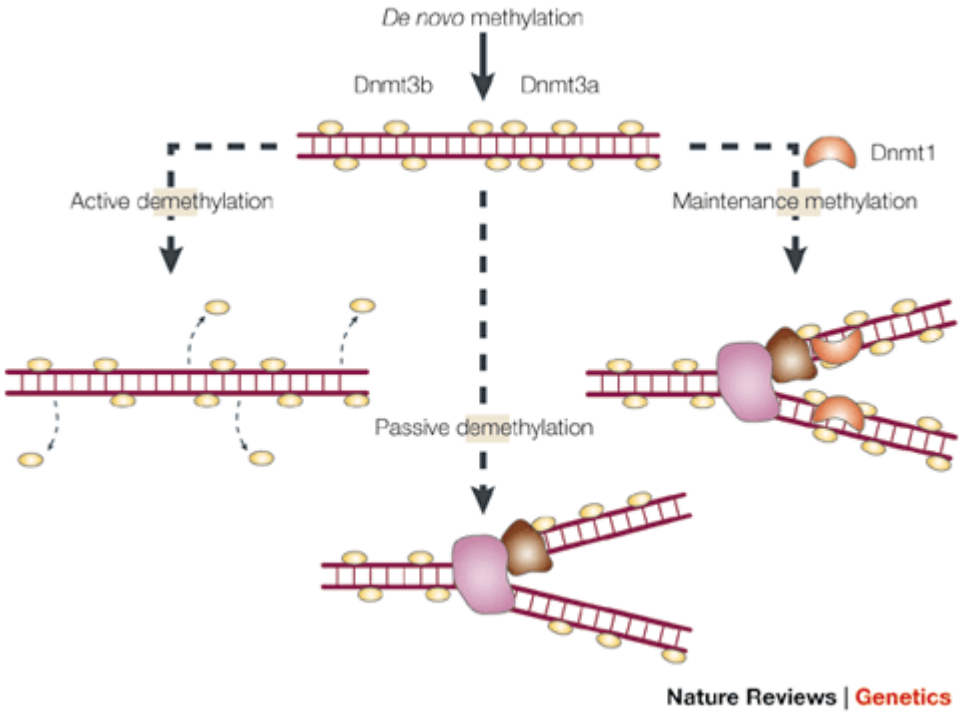


non-CG context



Babraham Bioinformatics

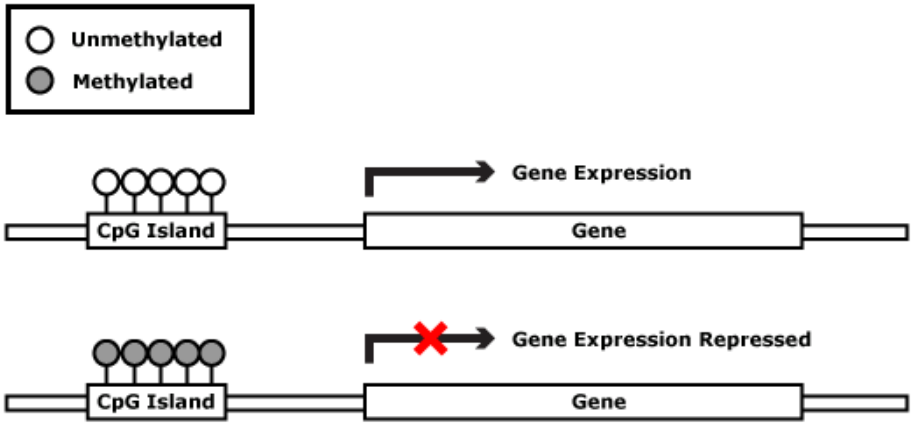
DNA methylation is maintained



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

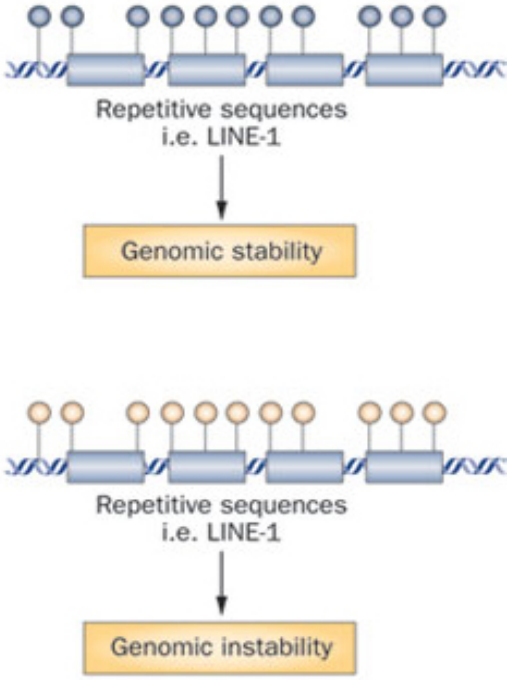
Babraham Bioinformatics

Regulation by DNA methylation



Silencing of gene expression

Tissue differentiation and embryonic development



Repeat activity

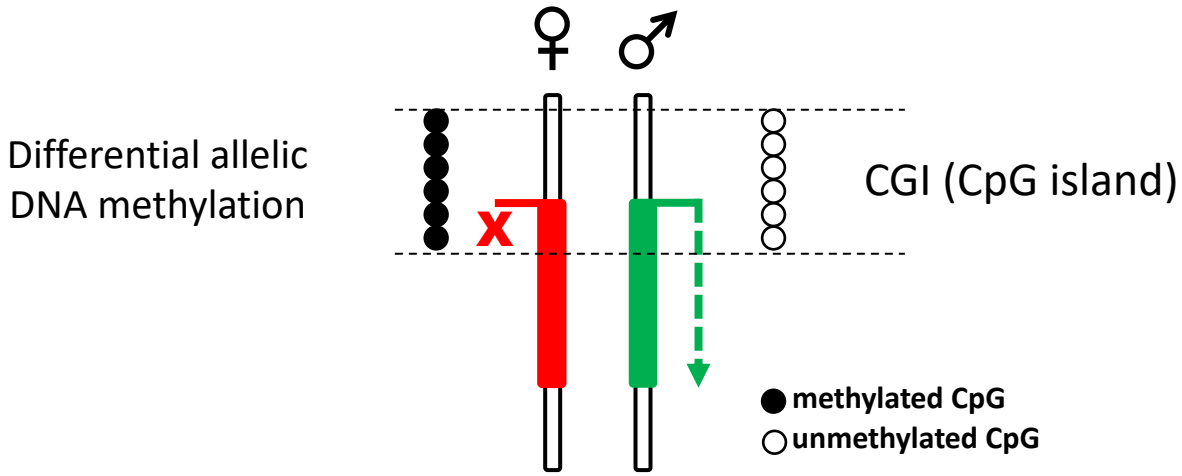
Genomic stability

Faults in correct DNA methylation may result in

- early development failure
- epigenetic syndromes
- cancer

Babraham Bioinformatics

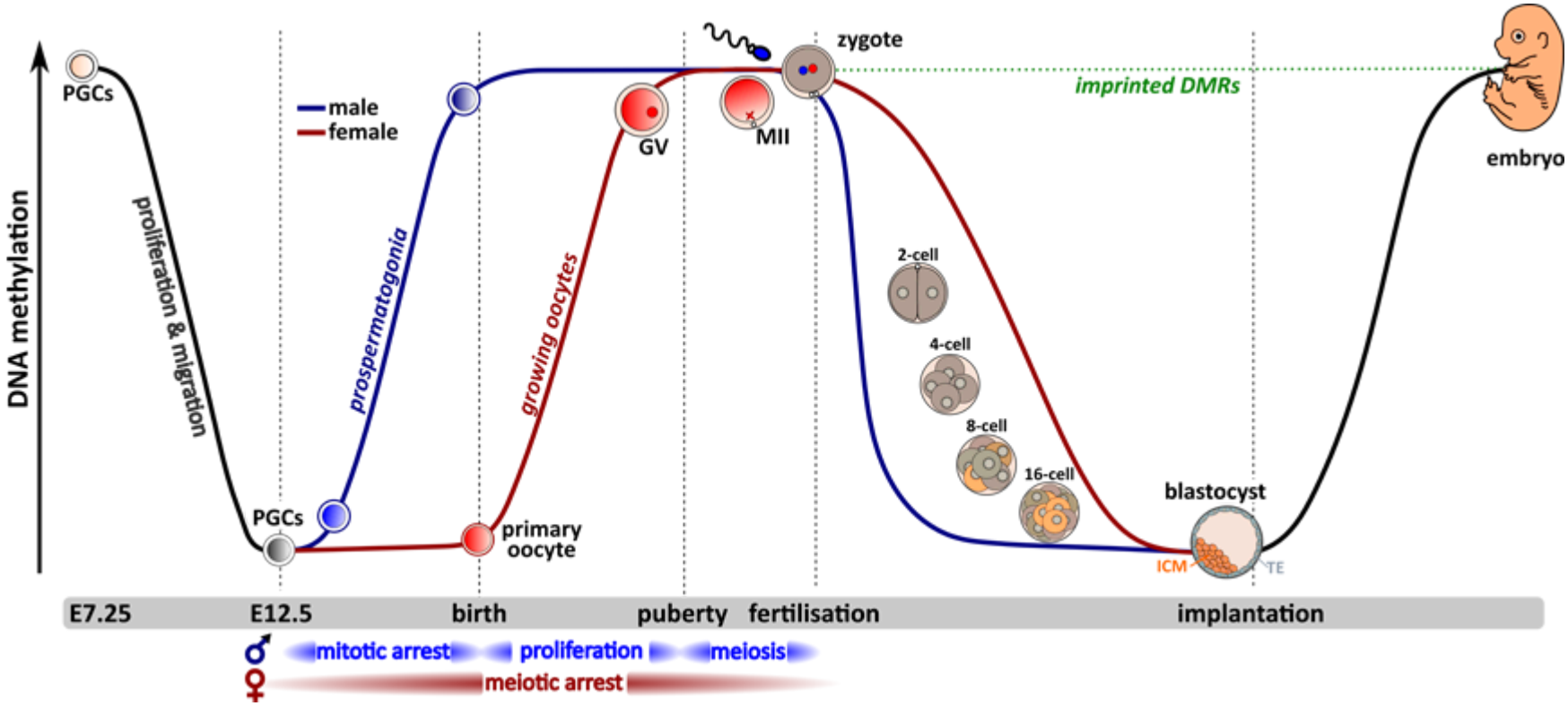
Imprinted Genes: mono-allelic expression



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

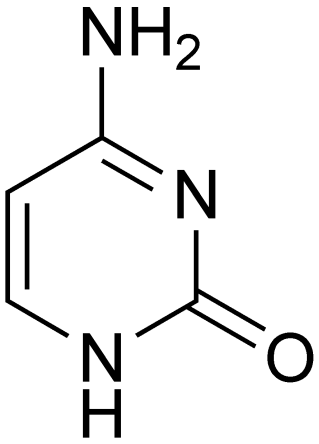
Babraham
Bioinformatics

DNA methylation is reset during reprogramming



Babraham Bioinformatics

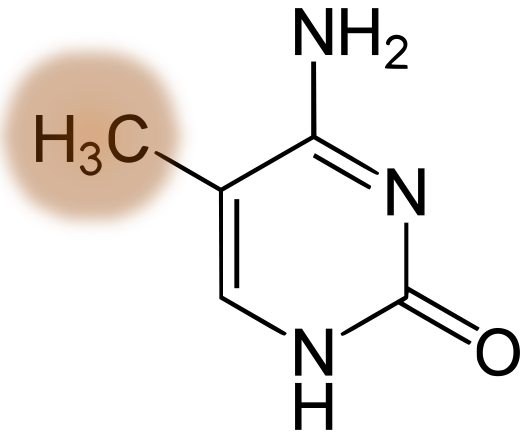
DNA Methylation



Cytosine

DNA methyl-transferases

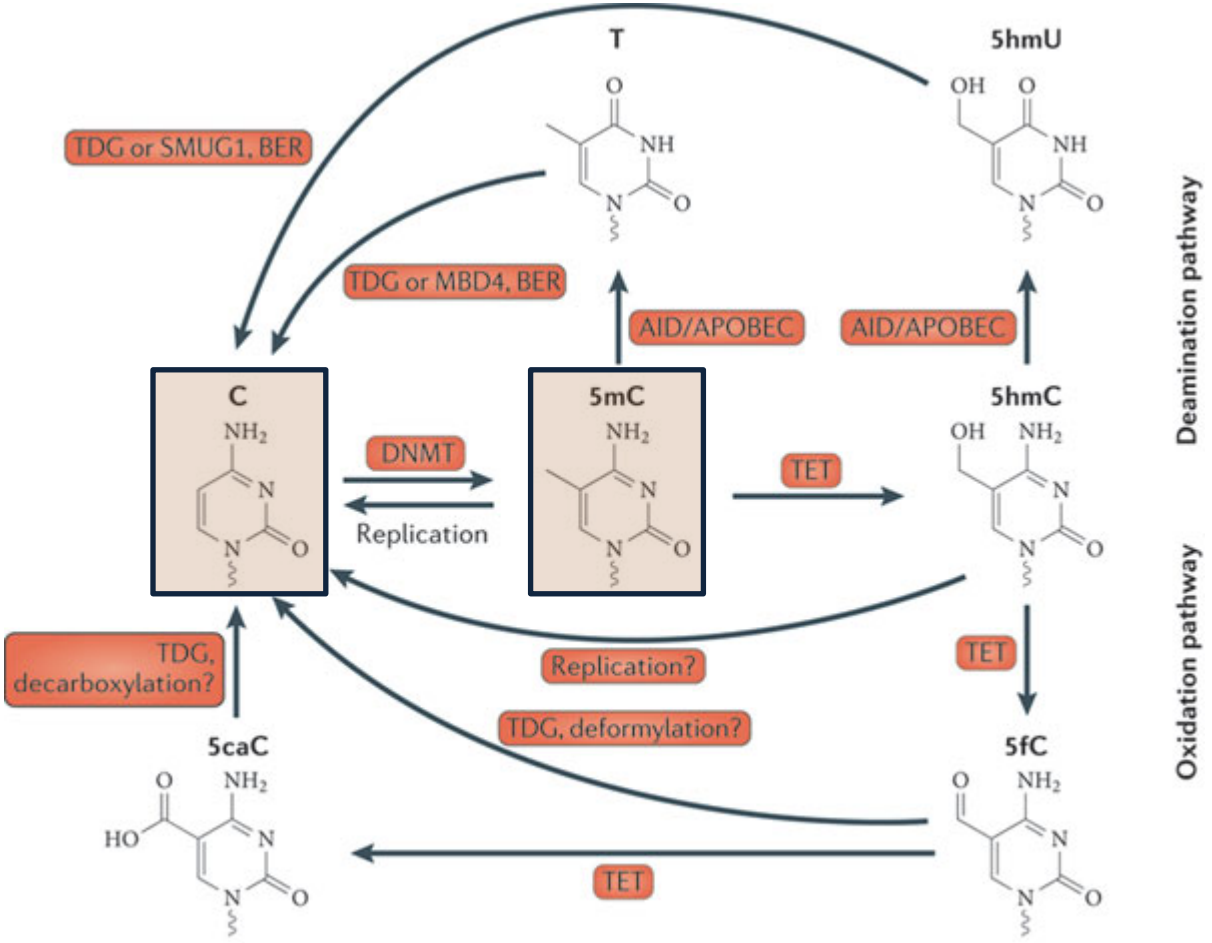
DNA-demethylase(s)?
TETs?
Passive demethylation?



5-methyl Cytosine

Babraham Bioinformatics

Other cytosine modifications



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

Nature Reviews | Genetics

Babraham Bioinformatics

Measuring DNA methylation by Bisulfite-sequencing

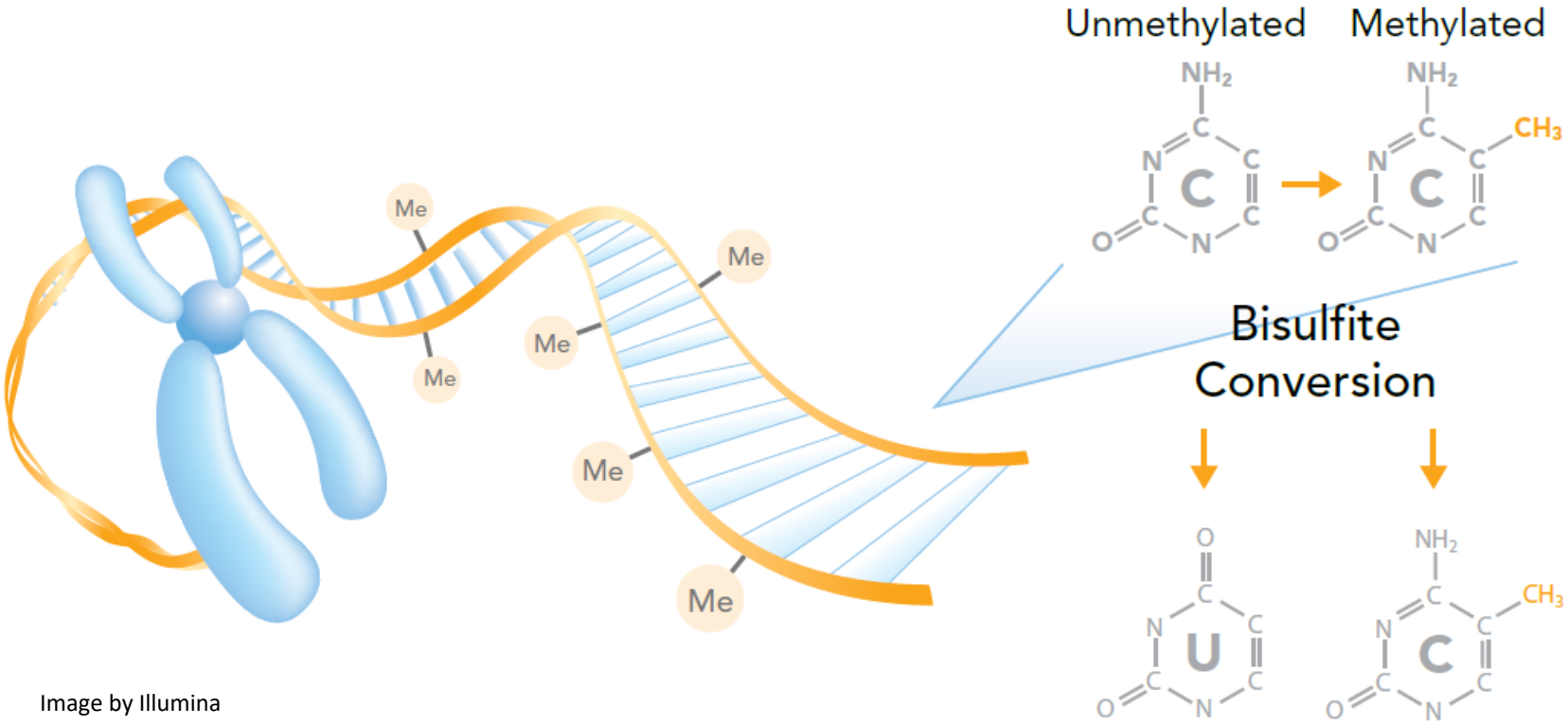
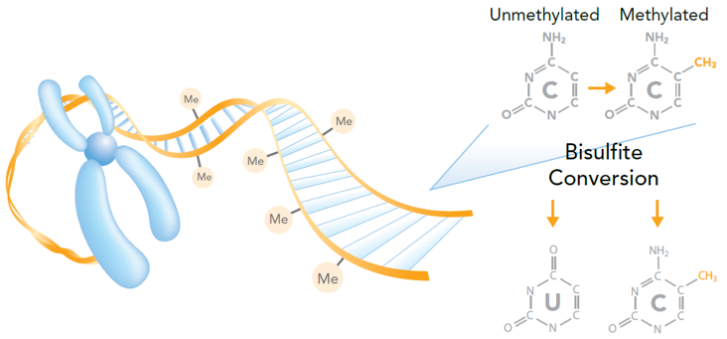


Image by Illumina

Babraham
Bioinformatics

Bisulfite Informatics



me me

CCAGTCGCTATAGCGCGATATCGTA

Convert

TTAGTTGCTATAGTGCGATATTGTA

Map

TTAGTTGCTATAGTGCGATATTGTA

. . . CCAGTCGCTATAGCGCGATATCGTA . . .

Babraham Bioinformatics

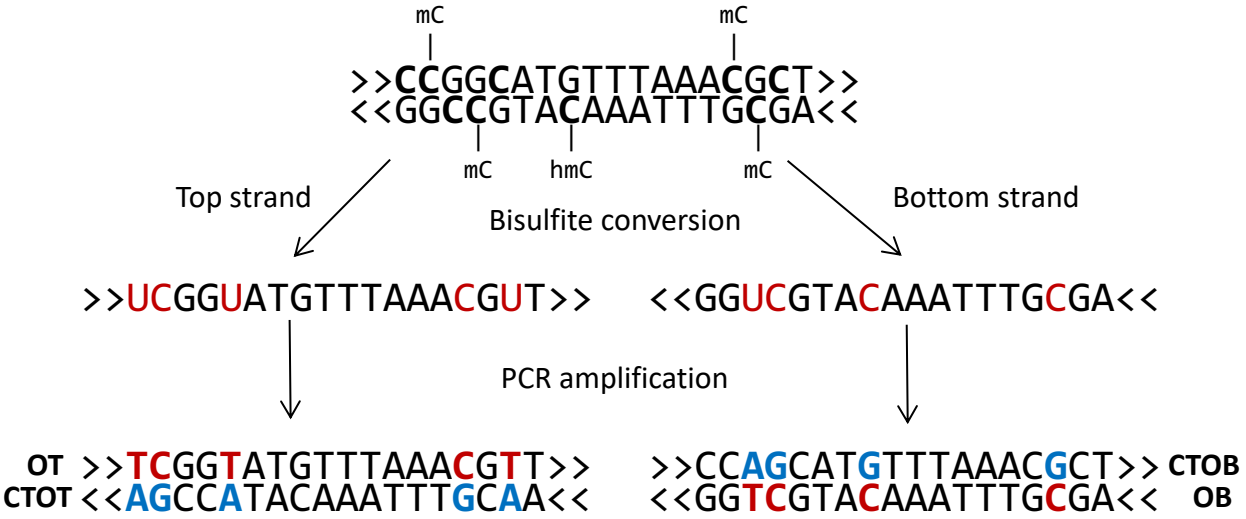
BS-Seq Analysis Workflow

Explore and understand your data

Sequencing → Processing pipeline → Methylation Analysis

Babraham Bioinformatics

Bisulfite conversion of a genomic locus



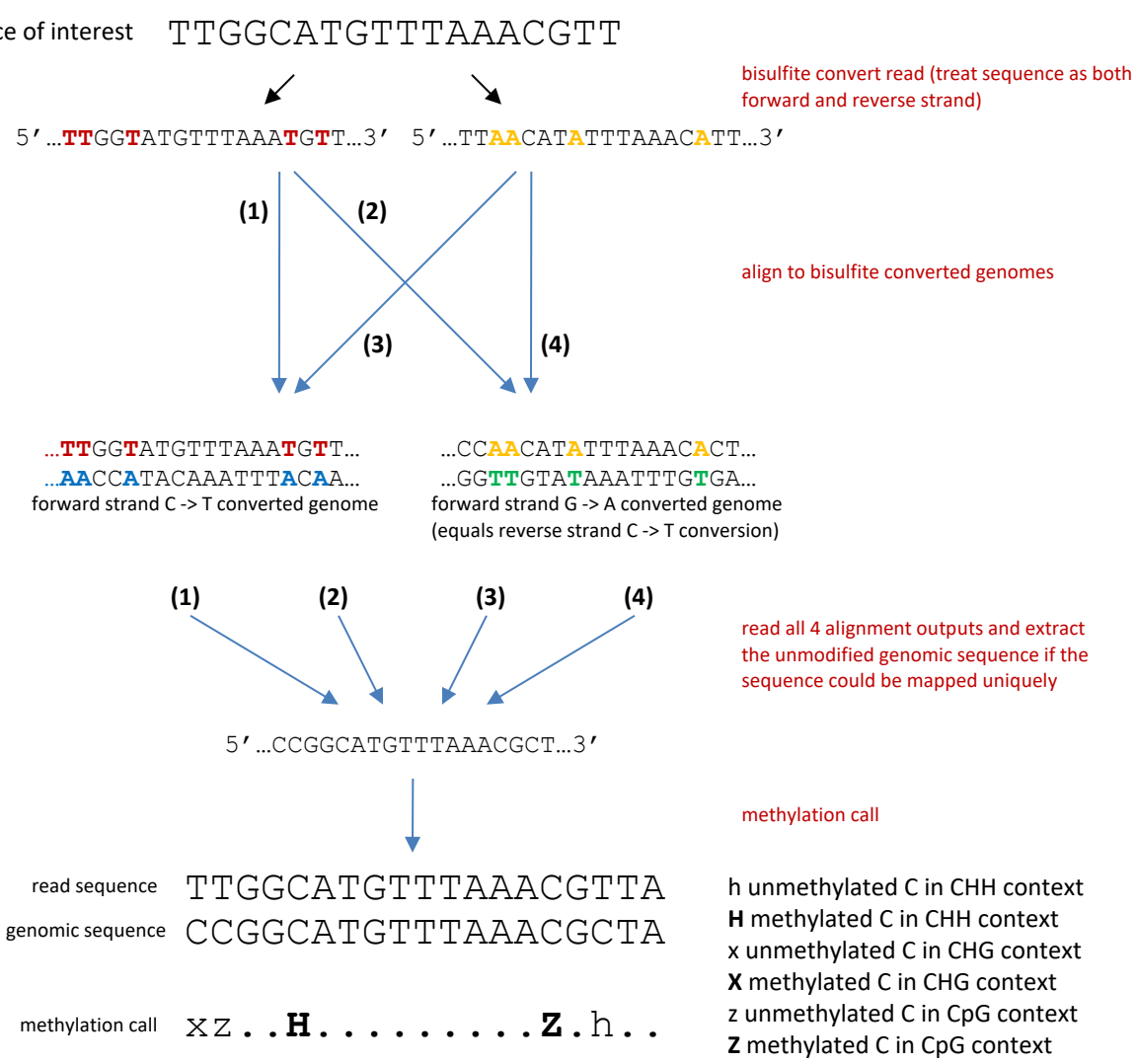
- 2 different PCR products and 4 possible different sequence strands from one genomic locus

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

Babraham Bioinformatics

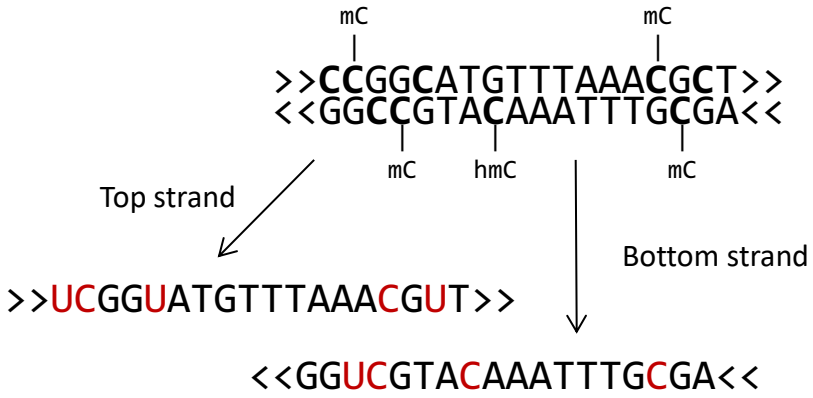
3-letter alignment of Bisulfite-Seq reads

Bismark



Babraham Bioinformatics

Common sequencing protocols



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

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

2) PBAT libraries

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

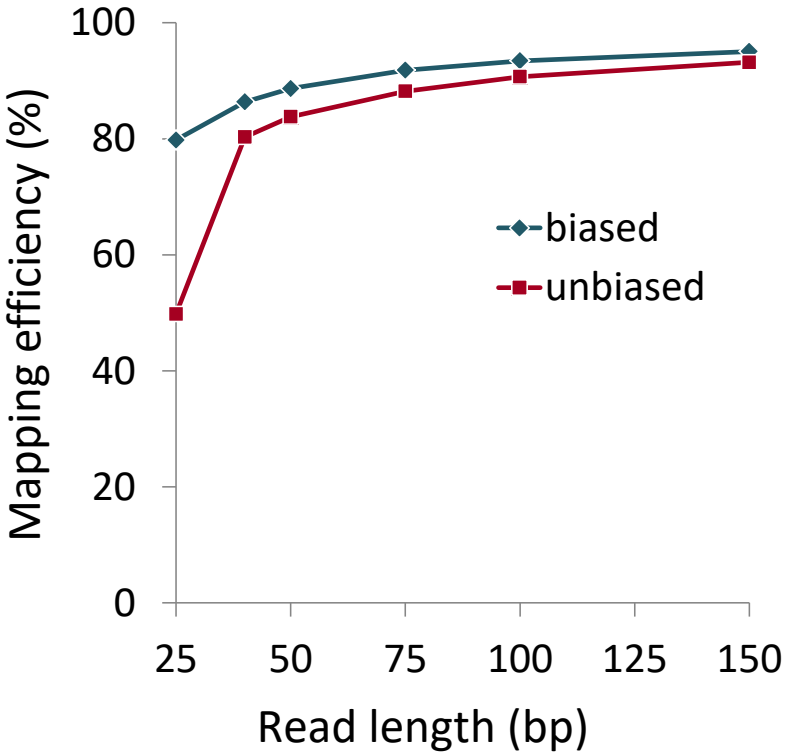
3) Non-directional libraries
(e.g. single-cell BS-Seq, Zymo Pico Methyl-Seq)

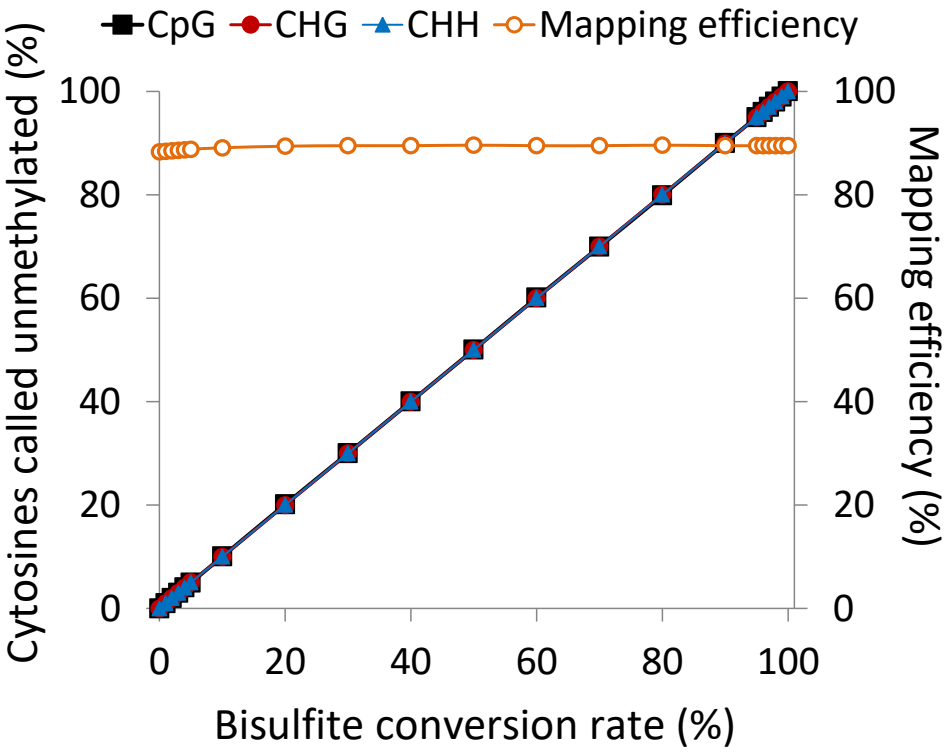
OT >>TCGGTATGTTTAAACGTT>>
CTOT <<AGCCATACAAATTTGCAA<<

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

Babraham Bioinformatics

Validation





Babraham Bioinformatics

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;>CDACCCCCCCCCCCCAACCCBCCC
@HS31_12166:1:1101:5276:2474#2/1
AGGTTTGTTGAGGTAATTTTTTTTTTTATATTTTATTATATTTTTATAGAATATATATTTTTTTTTTTTTTTTTTTTTATTAATATTTTAAATATTTTAT
+
BBCDDFDBHFDCDCGIIJJJJJJJJJJGIEIHJJIJJGH@GHIJJ=DAEEEHEFDFFFFFFEDDDDDDDDD-9BDDDDD(:@:>:(+(4>:C@((4:(((4
@HS31_12166:1:1101:5376:2480#2/1
GCCCTTCAAAAAAAAATAATATTAATTTTACTTACTAAAAACTTTAATATCCCATTAAAAAAATATTAAACTACATATACCGTTATCAATATTATTTAAA
+
CCCFFFFFHHHHHJJJJJJJIJJJJIJJJJJJJJJJJJJJGHJJHJJJJHIJIJJJJJHHFDEEEFEEDDDDDDDFEEEDDDDDDDEDDEDEEEEEED
@HS31_12166:1:1101:5674:2287#2/1
TAAACCTAGATATATAACCATCTACTAACTAAATTTCATTTACTAAAAACTATCATTTTTAAAAAATCTAAATTTTTCTCTCTCAACAACTAACAATTAT
+
BBCFDEFFHFHHHHIJJFHIIJJIHHHJGHHGHIJJJIJJJHJJJJJJJJJJJJJJJJJJIJJJJJIJJJJJJJJJJHHHHHHFFFFFDDEEEDDDDDDE
@HS31_12166:1:1101:5575:2309#2/1
ATCACACCCTACCTCTAAATACATCTATAAATCTCAATATACCACCGTCAAACATAATTCAAAACCAATCCTCCCACCACACTAAAAAAACAATAAACTA
+
BBCFDDDFHHFHDGHIHIIJHIHIFHHIIJHIHIHIIJIJJHGGHGIHIIIJJJJJJJJJJJJJJJJJJJIJHHHHHFFFDDDEEDDDDDDDDDDDDDDD
@HS31_12166:1:1101:5709:2315#2/1
GCCTCAAAACTTCAAAAAAAATAACCTAAATATCTATCTAATACCAAACATTAAATAAACTCCCATATCACAACAACAATATATTATCCTTAAAAACTCC
+
B@BFDDFFHFHHHFHIIJJJIJJIDHHGHIJJJHHGIGGHIJJGFGHJJJJJJJJJJJJJJJJHHHHHFFFFFFDDDDDDDDDEEFEFDDEDDDDDDDDD
@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

Babraham Bioinformatics

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

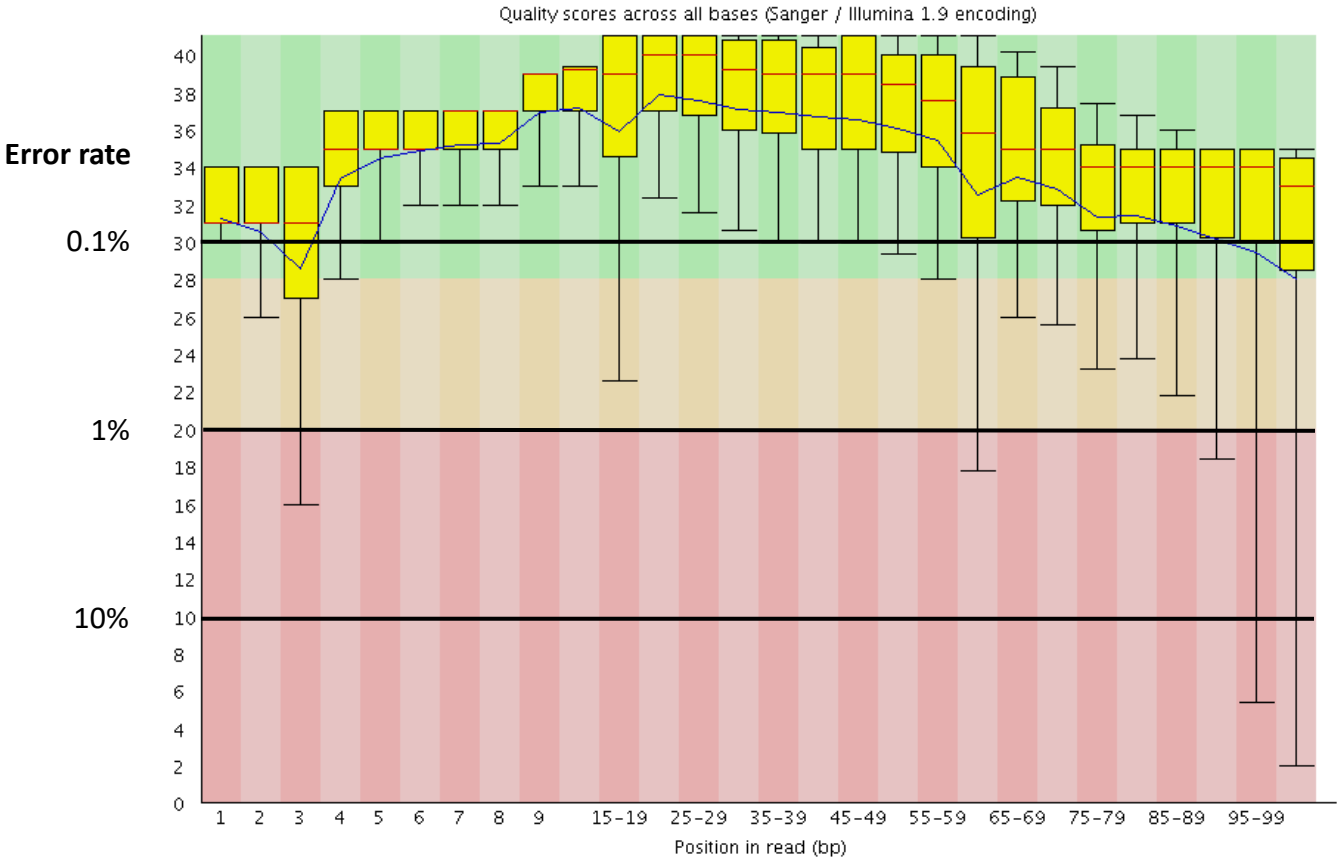
- # of sequences
- Basecall qualities
- Base composition
- Potential contaminants
- Expected duplication rate

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

Babraham Bioinformatics

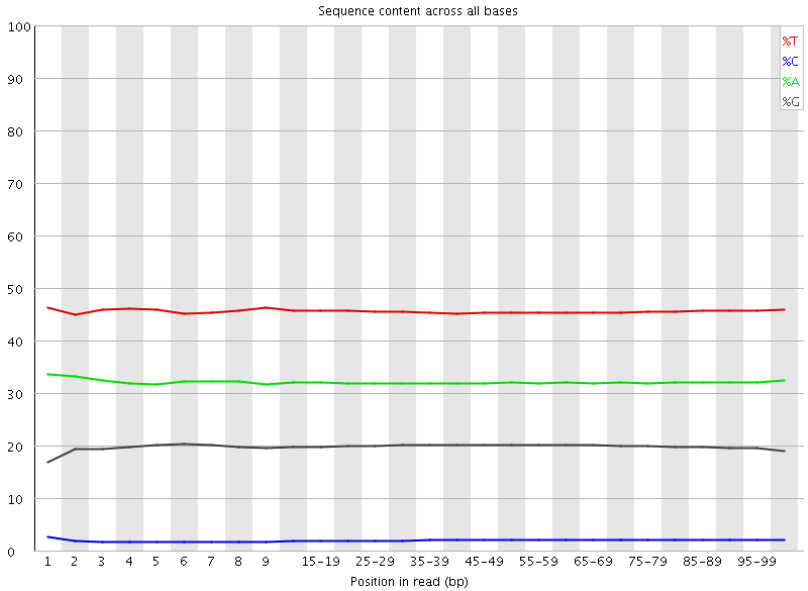
QC Raw data: Sequence Quality



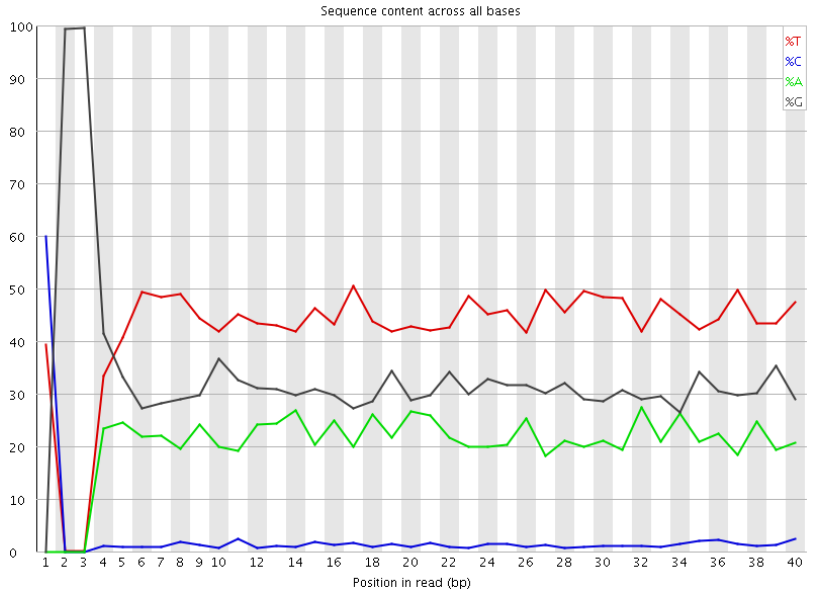
Babraham Bioinformatics

QC: Base Composition

WGSBS

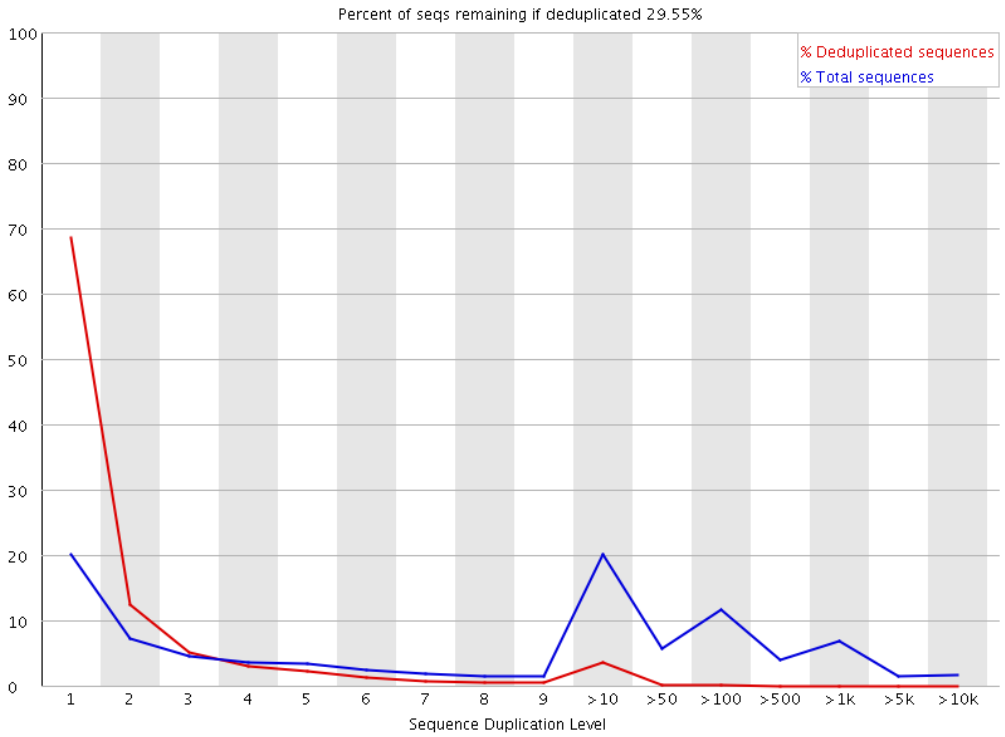


RRBS



Babraham Bioinformatics

QC: Duplication rate



Babraham Bioinformatics

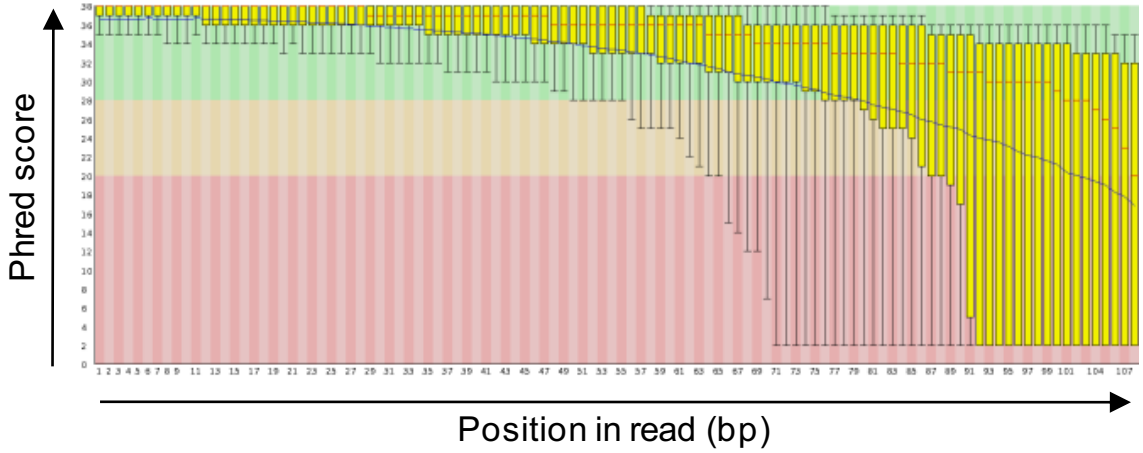
QC: Overrepresented sequences

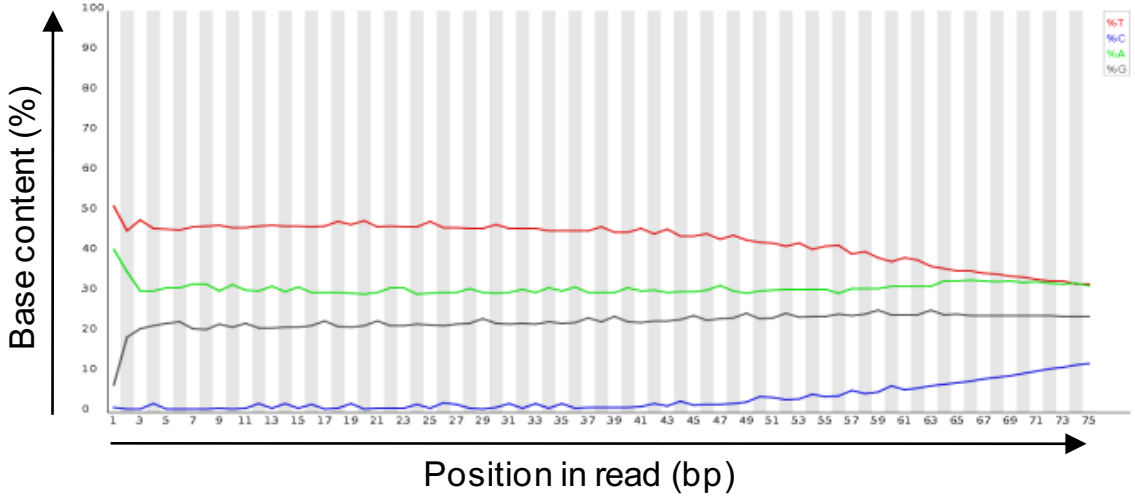
Overrepresented sequences

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

Babraham Bioinformatics

Common problems in BS-Seq





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

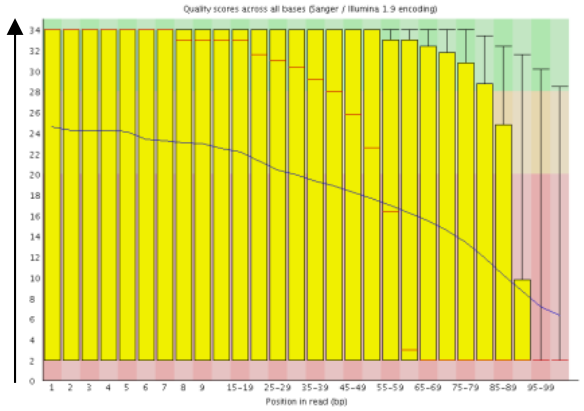
Babraham
Bioinformatics

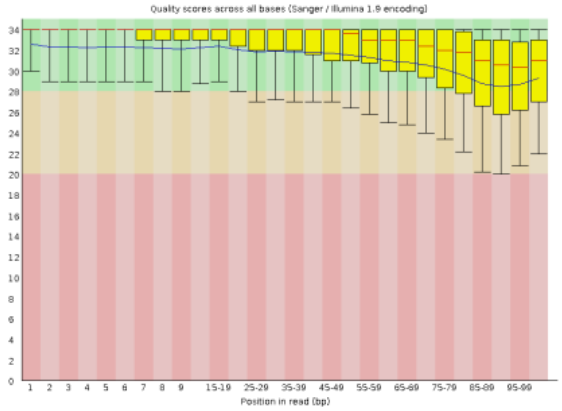
Removing poor quality basecalls

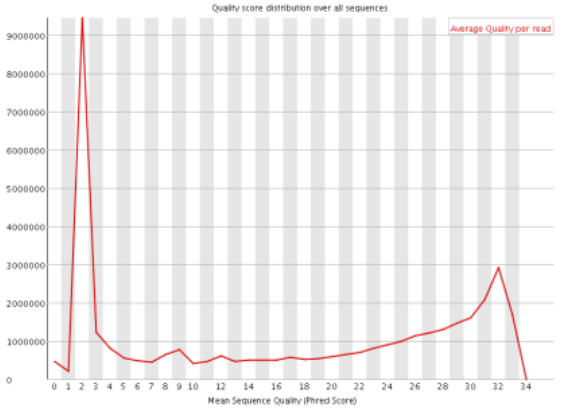
before trimming

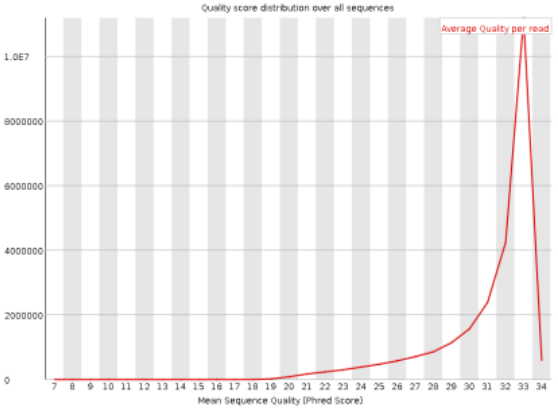
after trimming

Phred score





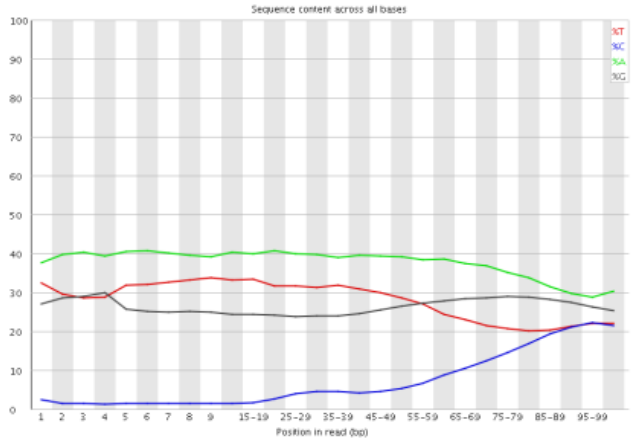




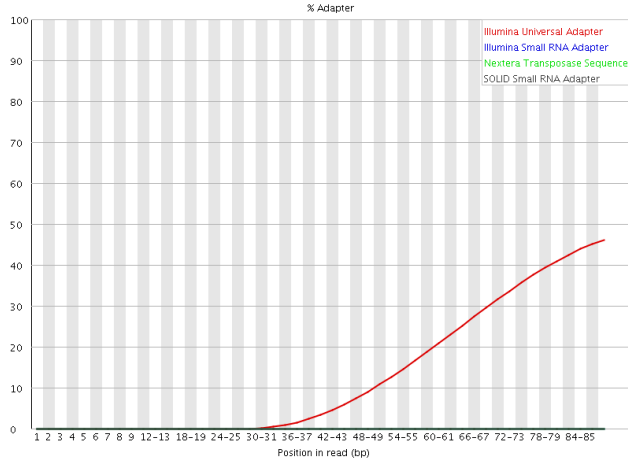
Babraham Bioinformatics

Removing adapter contamination

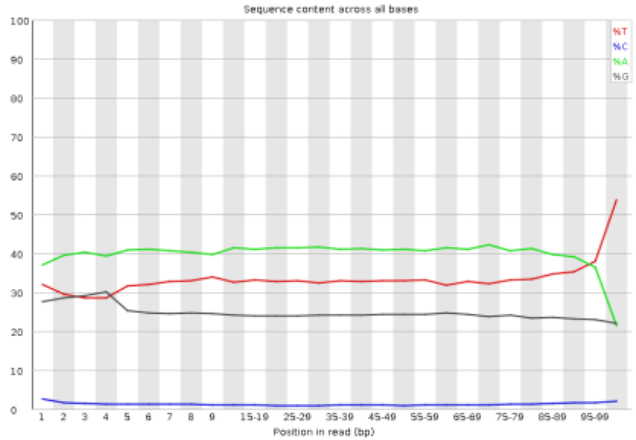
before trimming



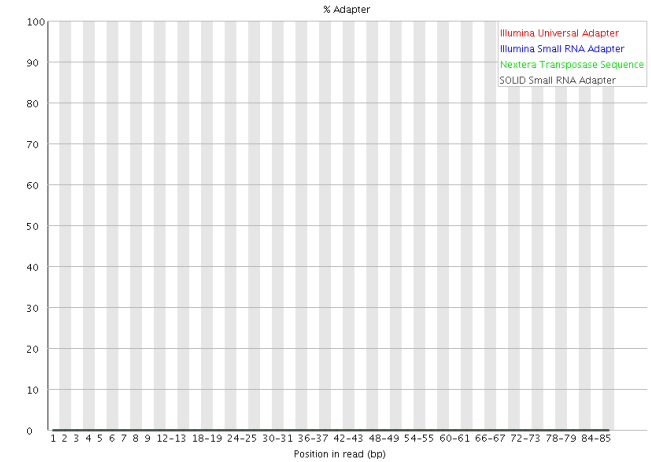
Adapter Content



after trimming



Adapter Content



Babraham Bioinformatics

Adapter trimming

(Illumina adapter: AGATCGGAAGAGC)

```
B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGAT
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGAT
```

partial match

full match

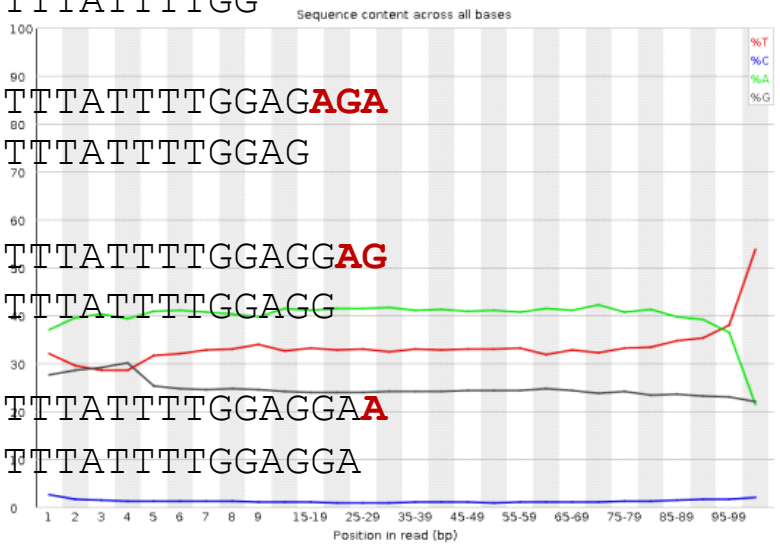
```
B: AGATCTTTTATTCGGTAGGATAGATCGGAAGAGCXXXXXXXXXXXXXXX
A: AGATCTTTTATTCGGTAGGAT

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGATC
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGG

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGAGA
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAG

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGAG
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGG

B: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGAA
A: AGATCTTTTATTCGGTAGGATTAGCGGTAGTTATTTTATTTTGGAGGA
```



Babraham Bioinformatics

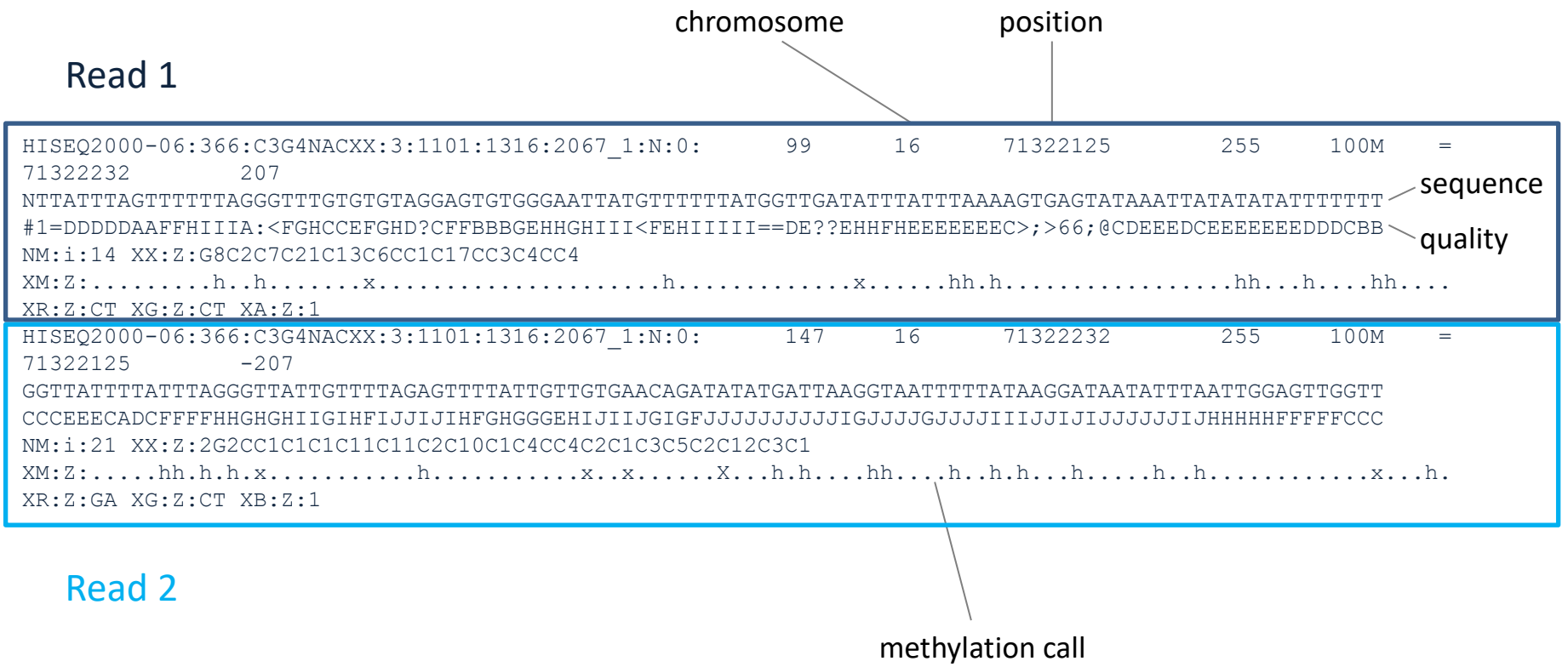
Summary Adapter/Quality Trimming

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)

Babraham Bioinformatics

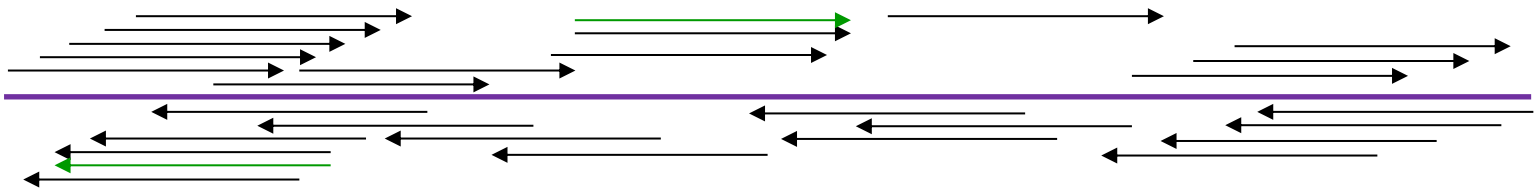
Part II: Sequence alignment – Bismark primary alignment output (BAM file)



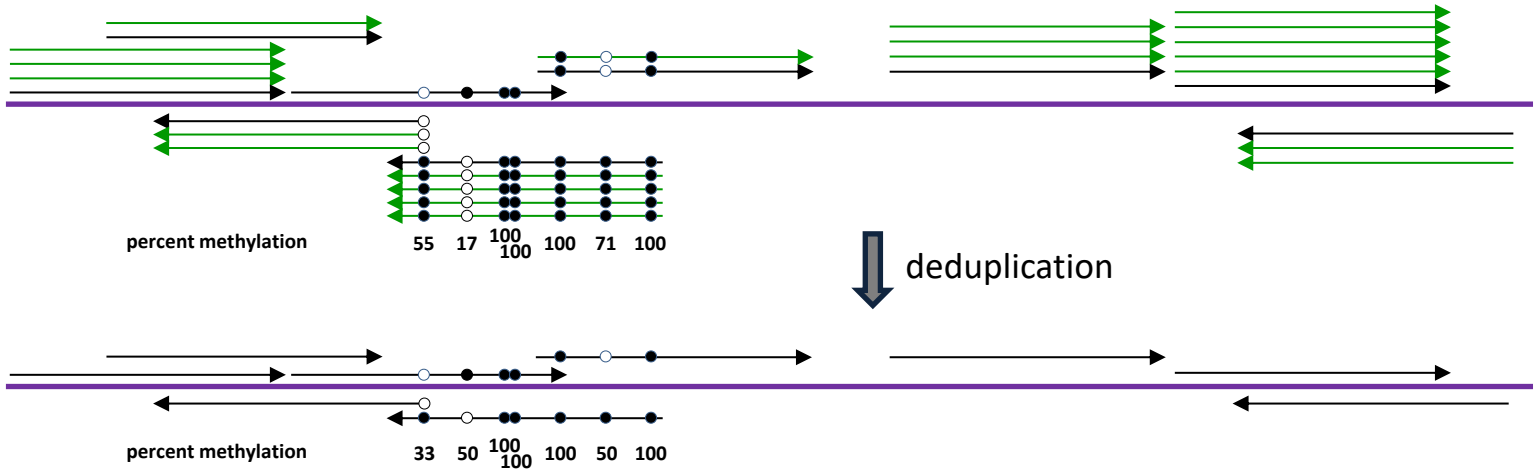
Babraham Bioinformatics

Sequence duplication

Complex/diverse library:



Duplicated library:

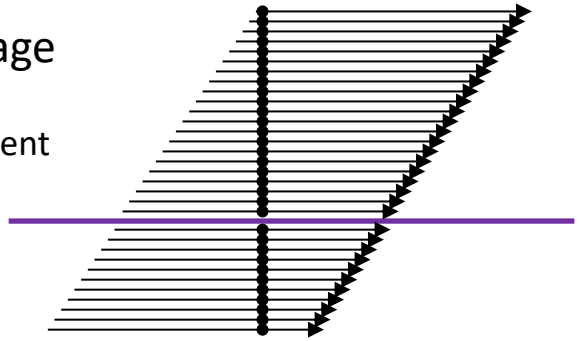


Babraham Bioinformatics

Deduplication - considerations

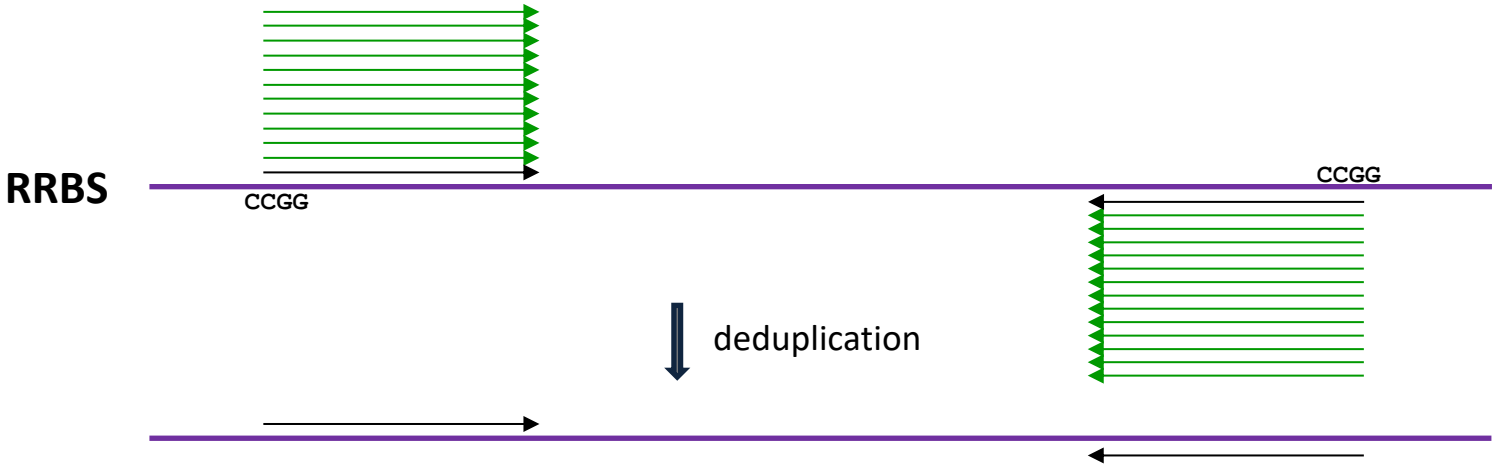
Advisable for large genomes and moderate coverage

- 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)



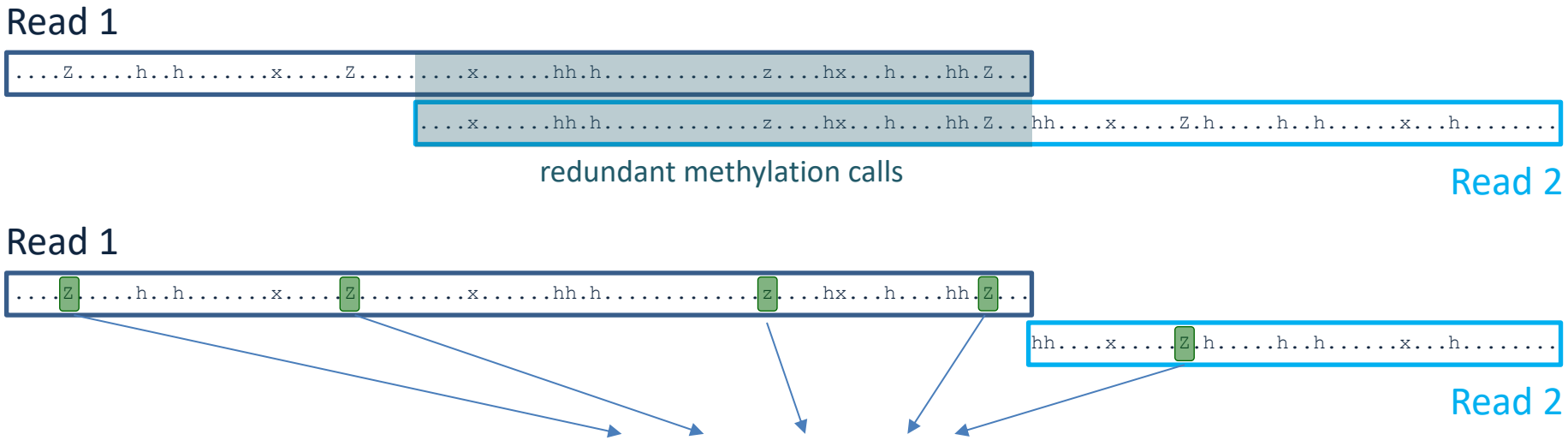
NOT advisable for RRBS or other target enrichment methods

where higher coverage is either desired or expected



Babraham Bioinformatics

Methylation extraction



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

Babraham Bioinformatics

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

Babraham Bioinformatics

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

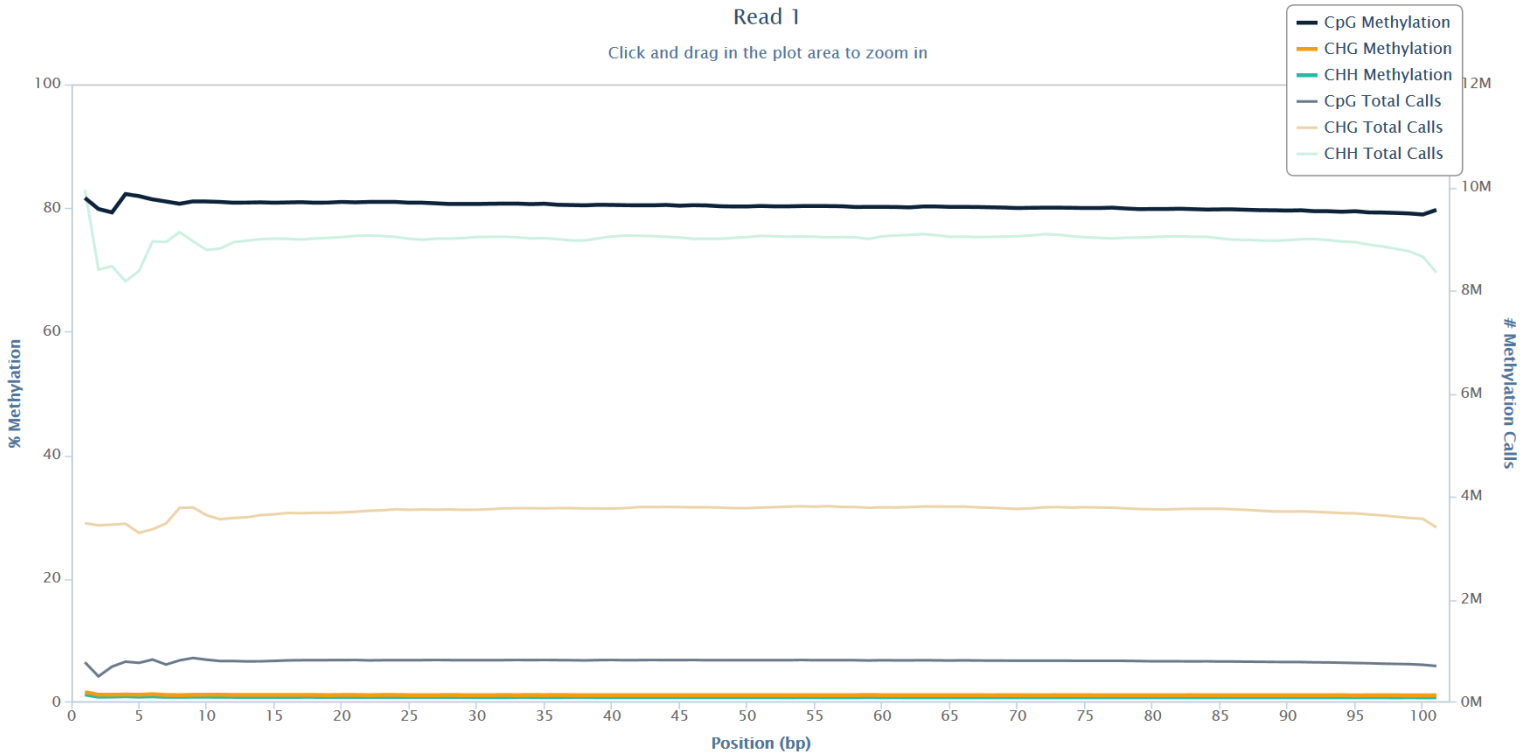
optional: merge into
CpG dinucleotide entities

Genome wide CpG report

Babraham Bioinformatics

Part III: Mapped QC - Methylation bias

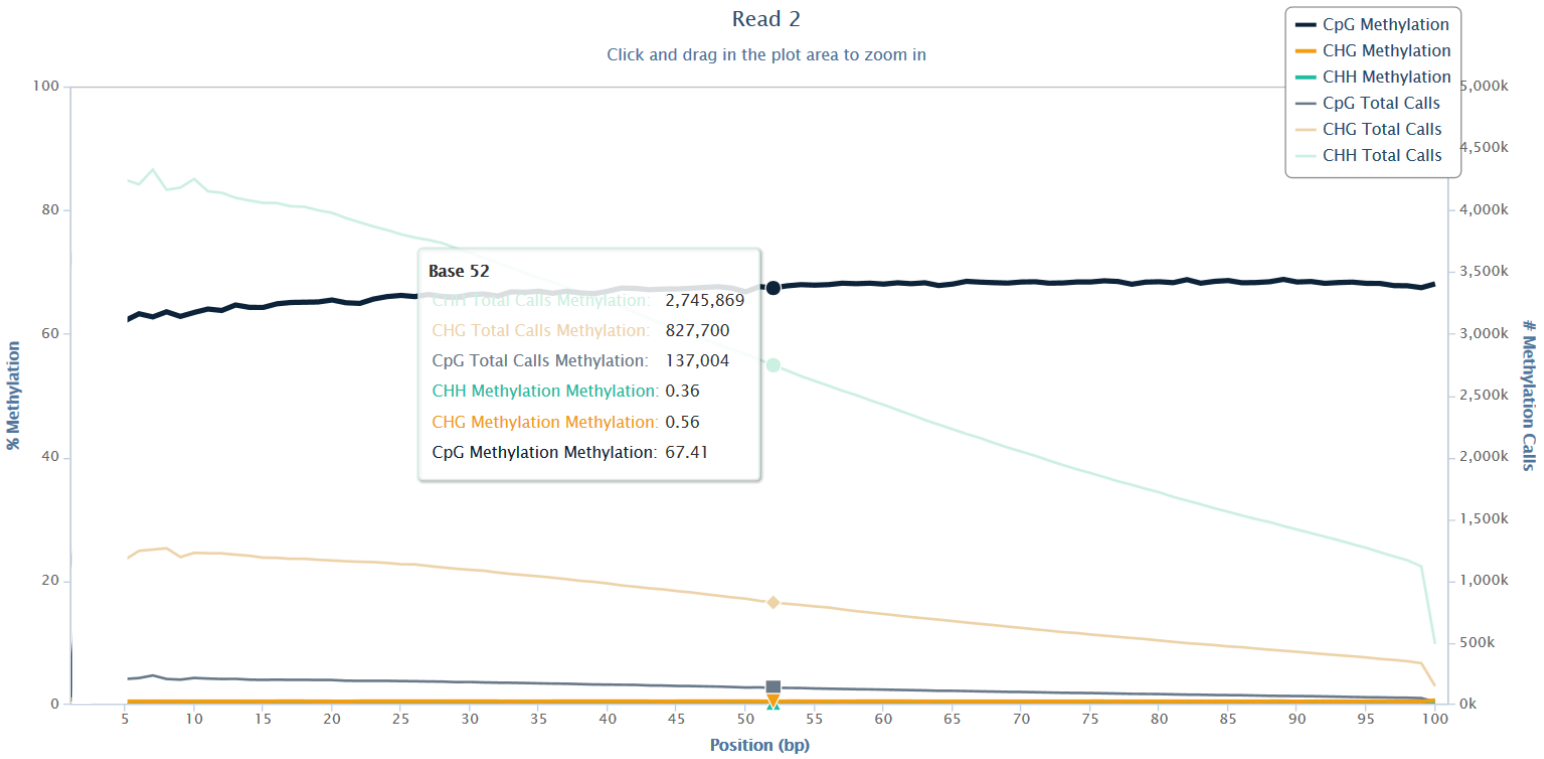
M-Bias Plot



good opportunity to look at conversion efficiency

Babraham Bioinformatics

Artificial methylation calls in paired-end libraries



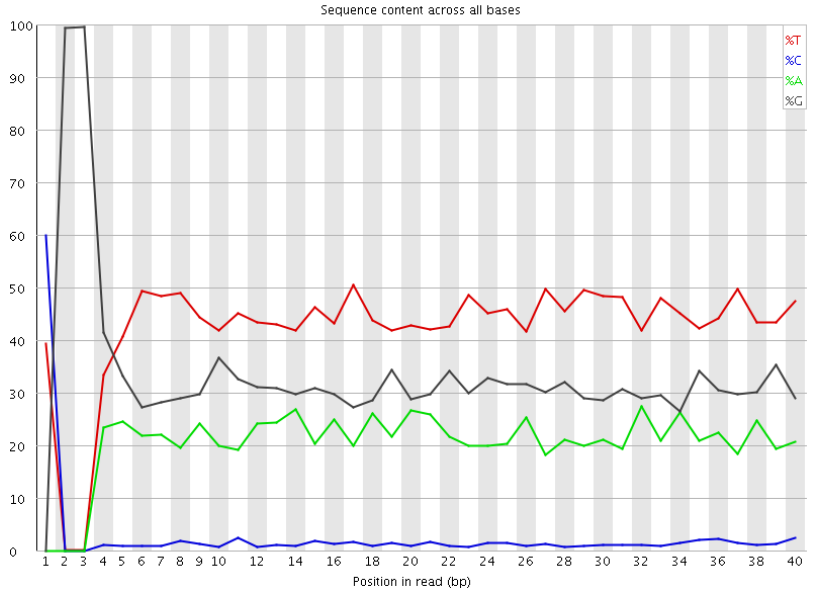
end repair + A-tailing

```
5′-        GGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCA        -3′
3′-        ACCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGG        -5′
```

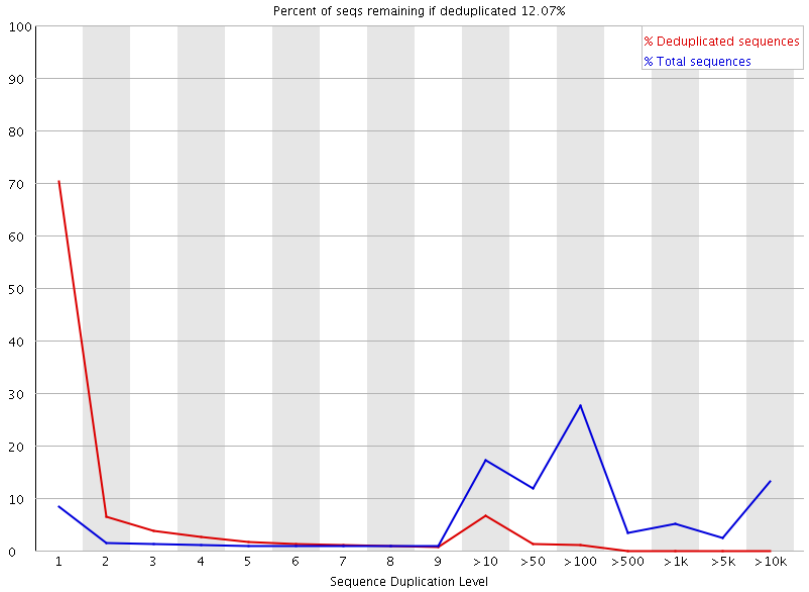
Babraham Bioinformatics

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

Sequence composition bias

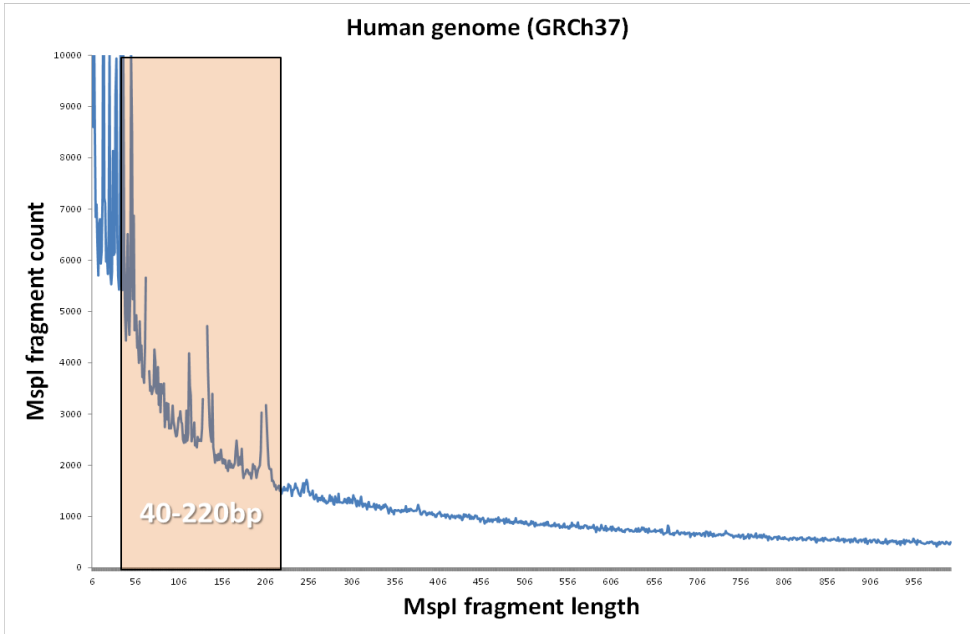


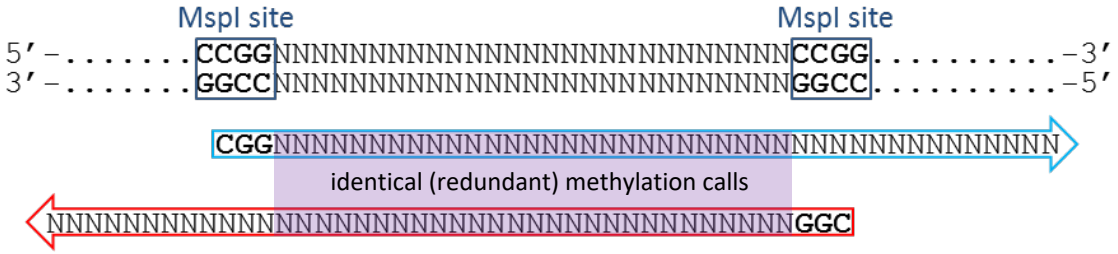
High duplication rate



Babraham Bioinformatics

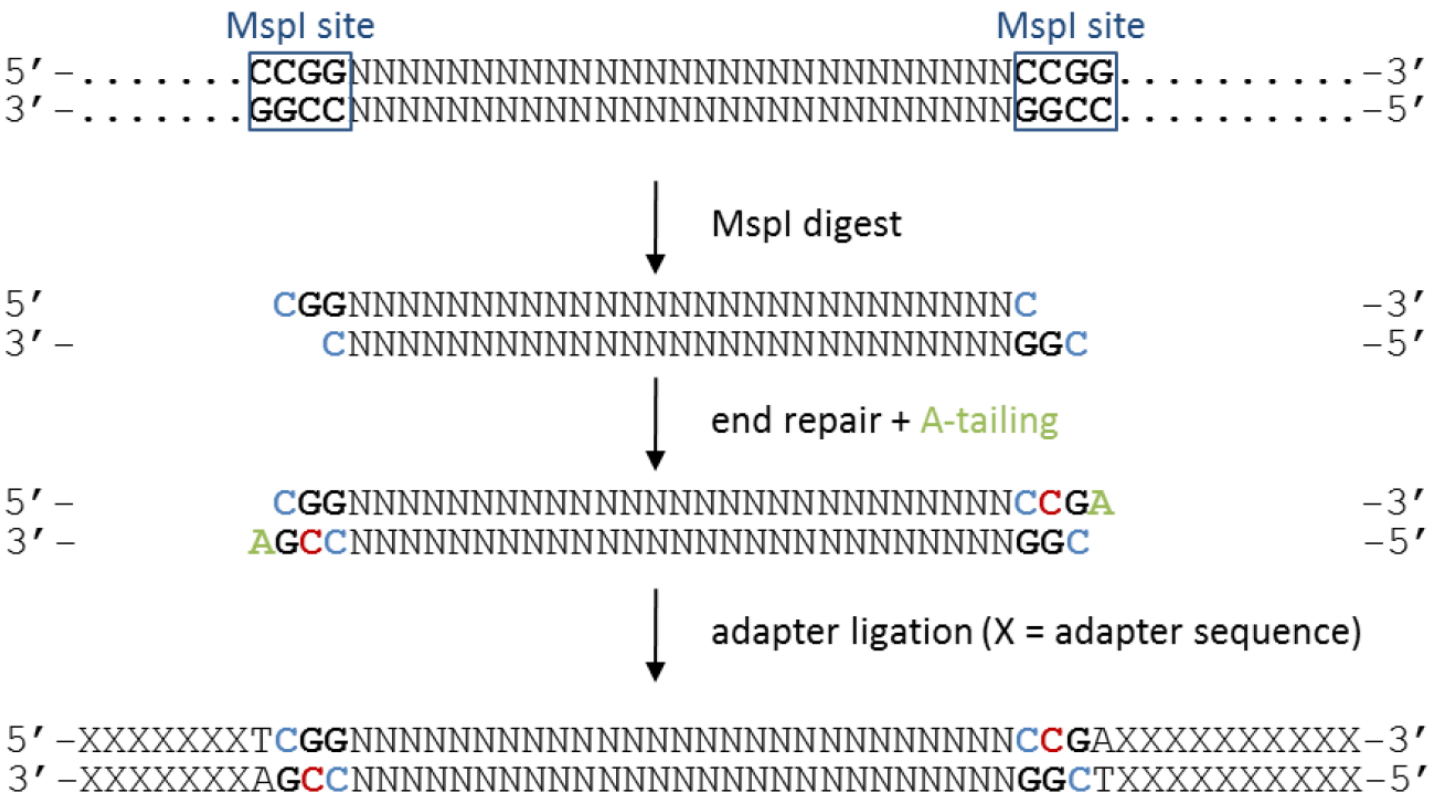
Fragment size distribution in RRBS





Babraham Bioinformatics

Artificial methylation calls in RRBS libraries



C genomic cytosine

C unmethylated cytosine

Babraham Bioinformatics

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*

Babraham Bioinformatics

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

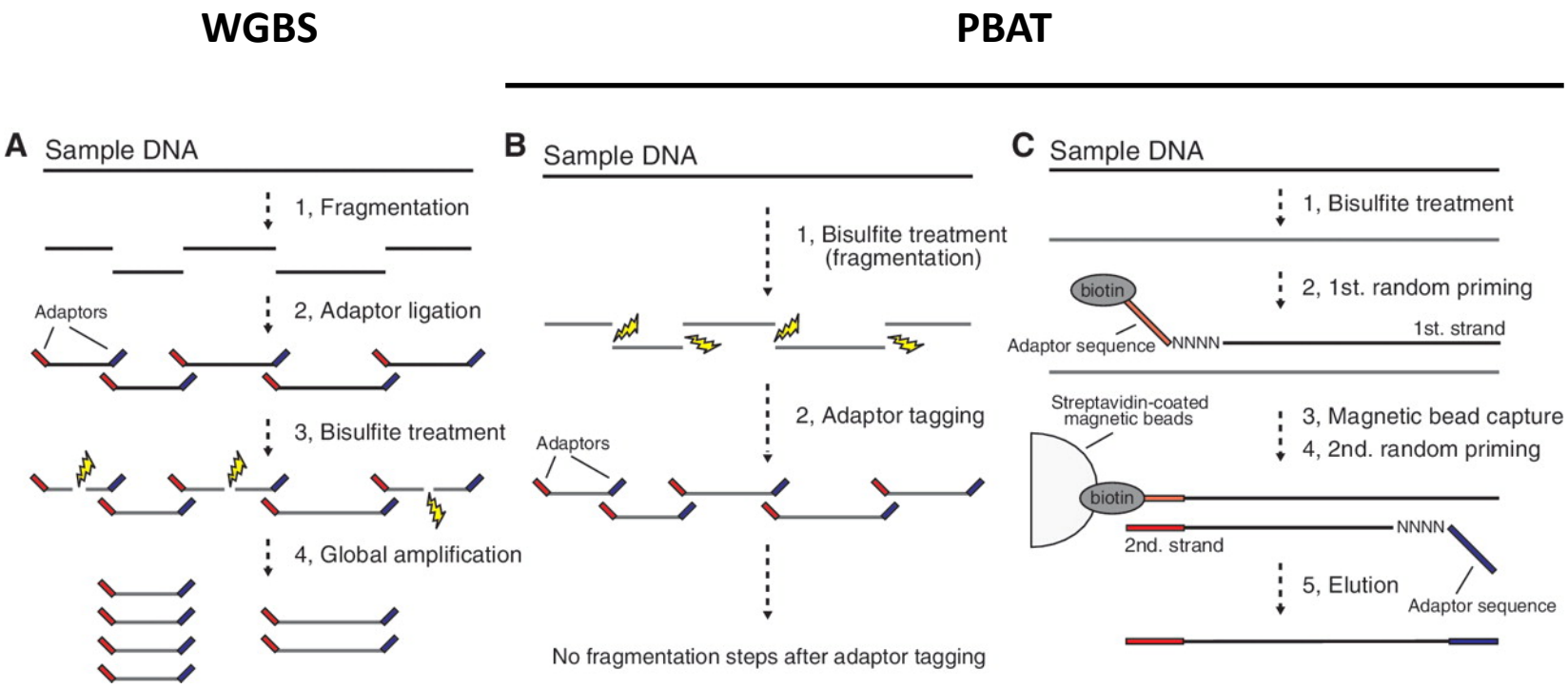
QCFAIL.com https://sequencing.qcfail.com/

Babraham Bioinformatics

Thank you for your attention

Questions?

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

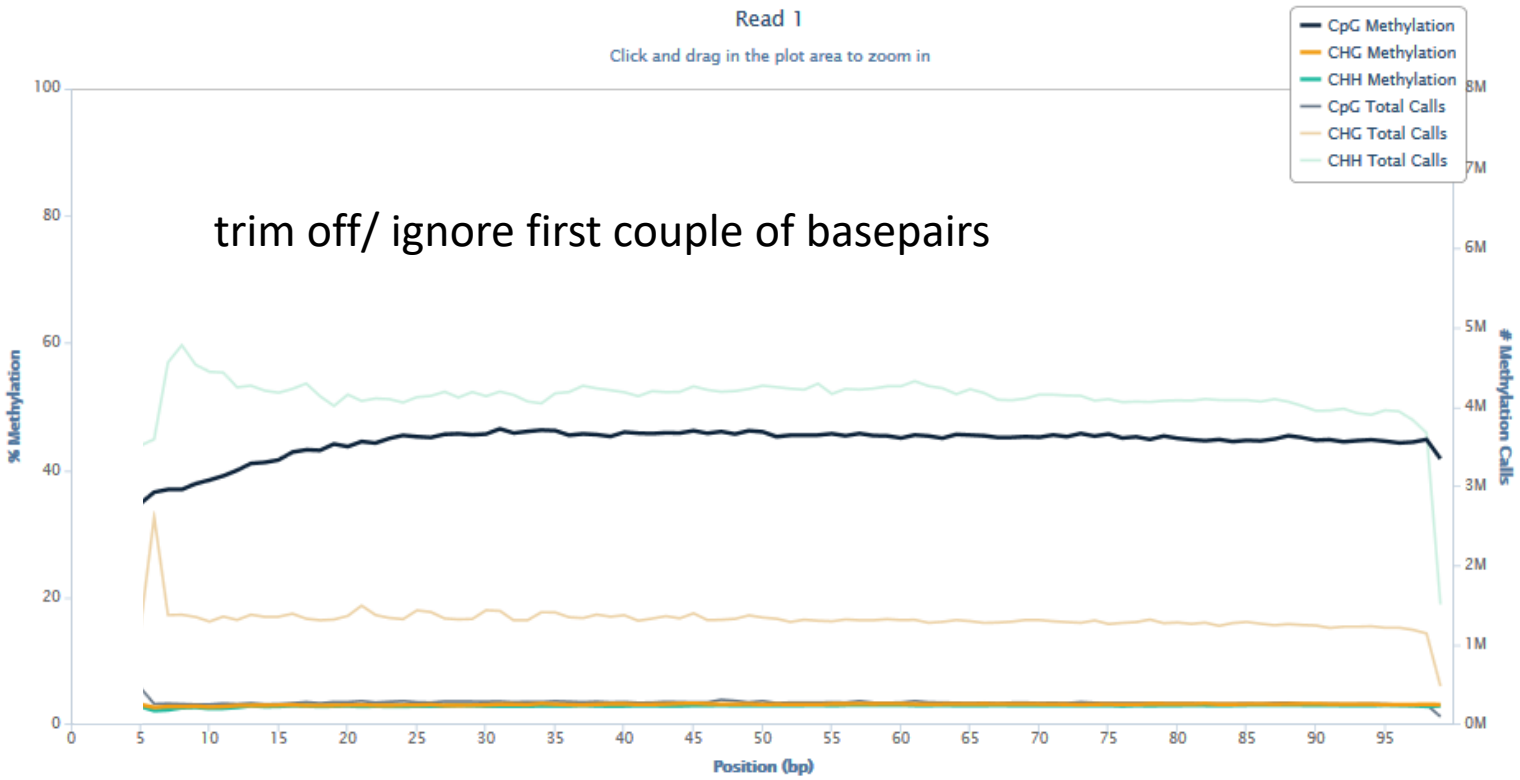


→ suitable for low input material

Babraham Bioinformatics

PBAT-Seq

M-Bias Plot



Babraham Bioinformatics