Bisulfite-Sequencing Theory and Quality Control

Felix Krueger, Simon Andrews
felix.krueger@babraham.ac.uk
simon.andrews@babraham.ac.uk

v2022-04

Babraham Bioinformatics

a.m.

Bisulfite-Seq theory and Quality Control

coffee

Mapping and QC practical

Visualising and Exploring talk

Lunch

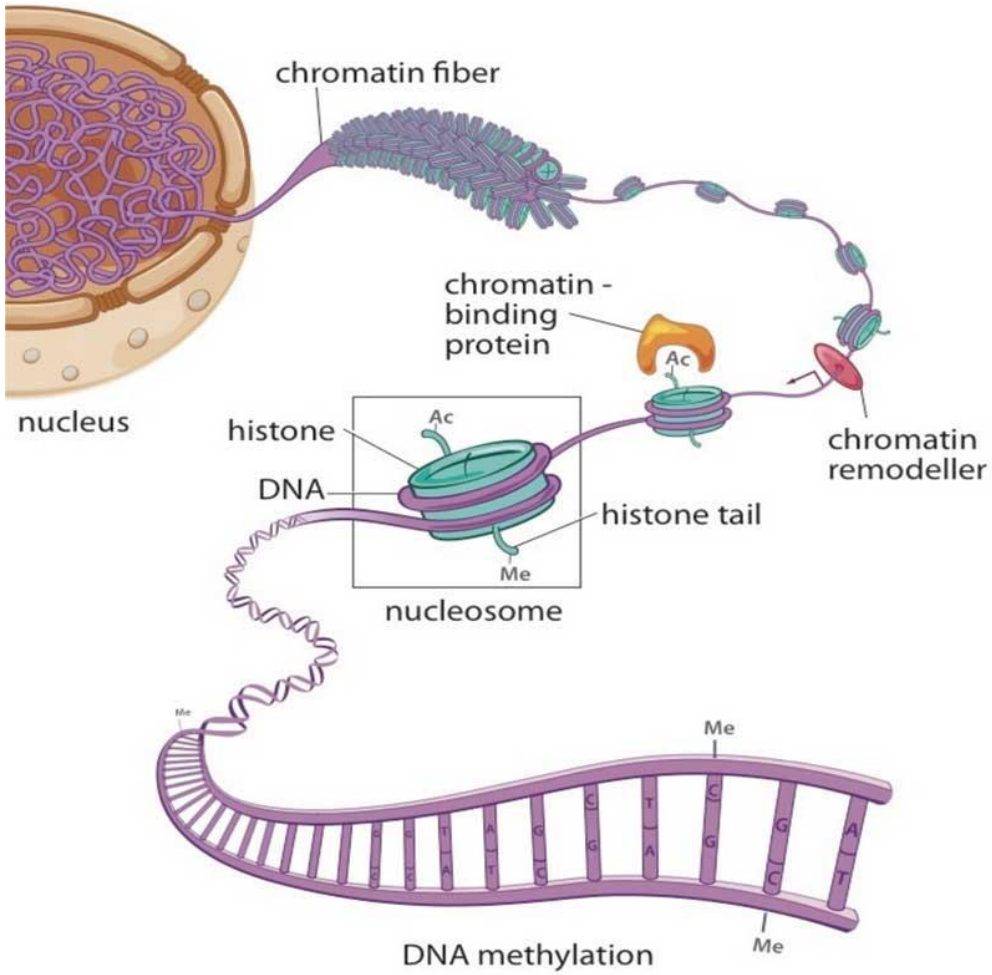
p.m.

Visualising and Exploring practical

coffee

Differential methylation talk & practical

Epigenetics



From The Cell Biology of Stem Cells (2010)

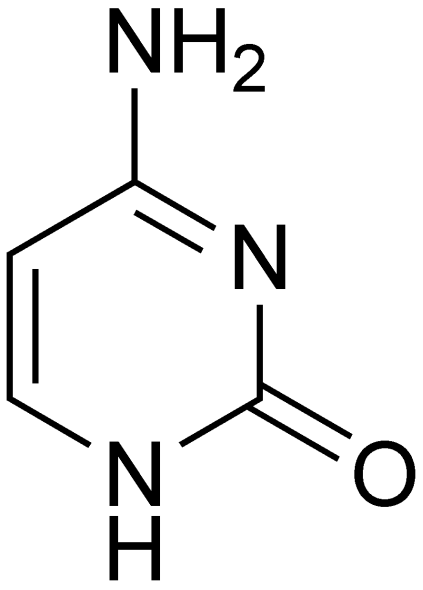
Studies changes in gene expression which are not encoded by the underlying DNA sequence

Chromatin

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

DNA cytosine methylation

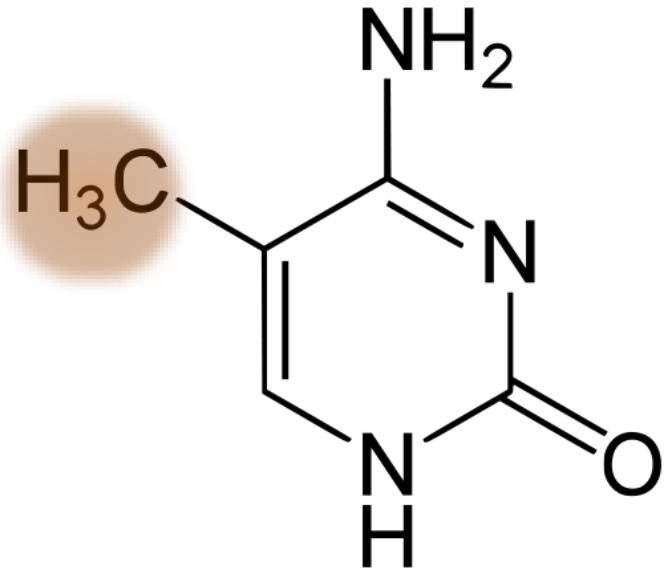
DNA Methylation



Cytosine

DNA methyl-transferases

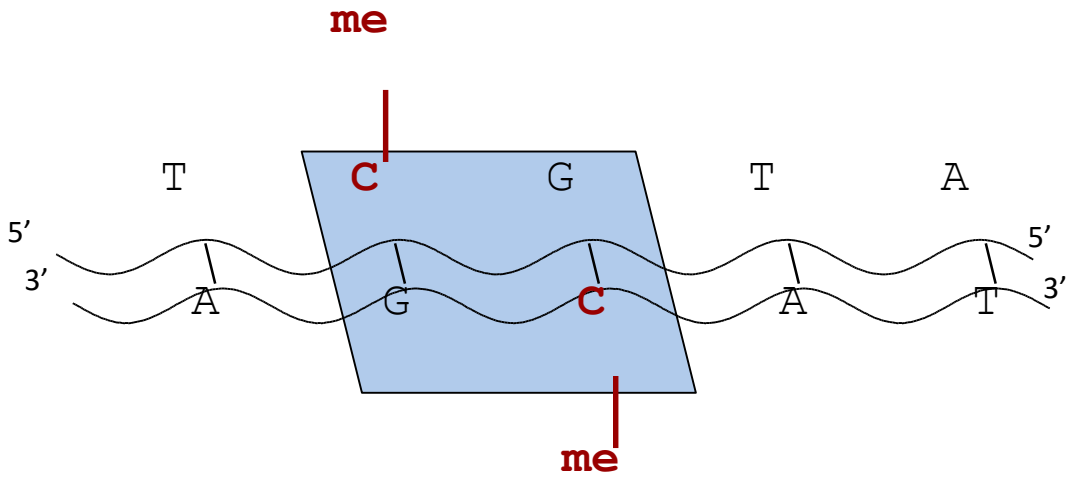
DNA-demethylase(s)
TET enymes
Passive demethylation



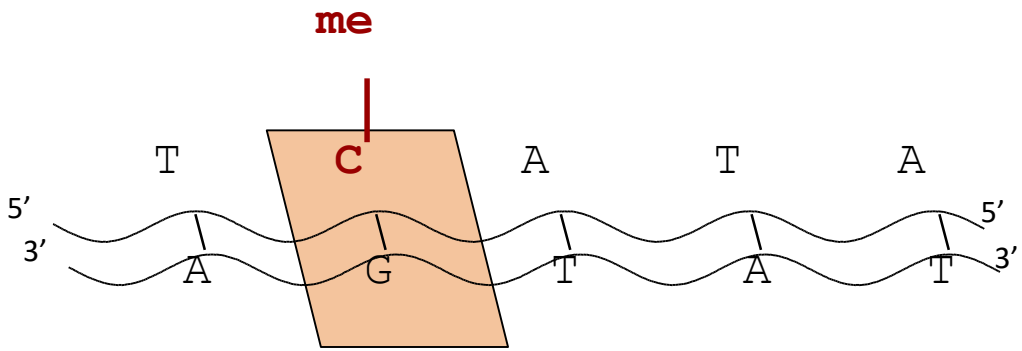
5-methyl Cytosine

Context of DNA methylation

CG context

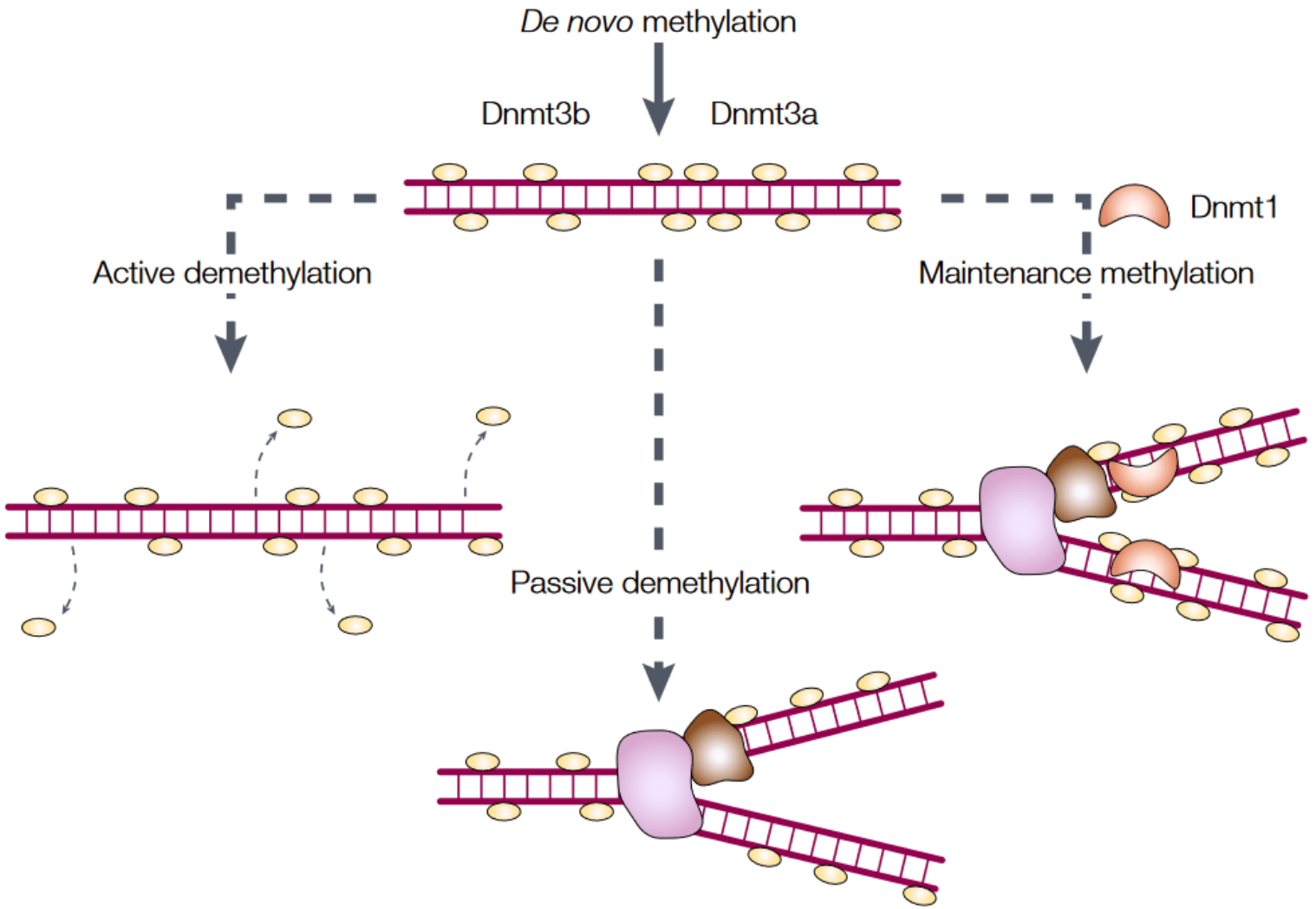


non-CG context



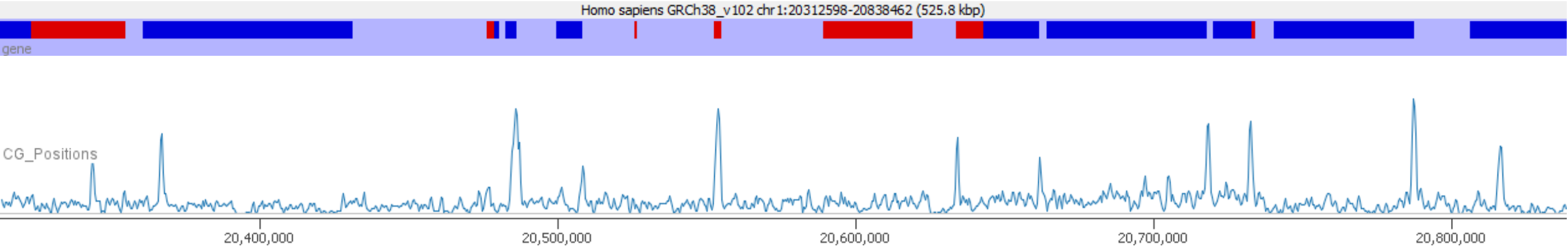
	Mammals	Plants
CG	present	present

DNA methylation is maintained through cell division



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

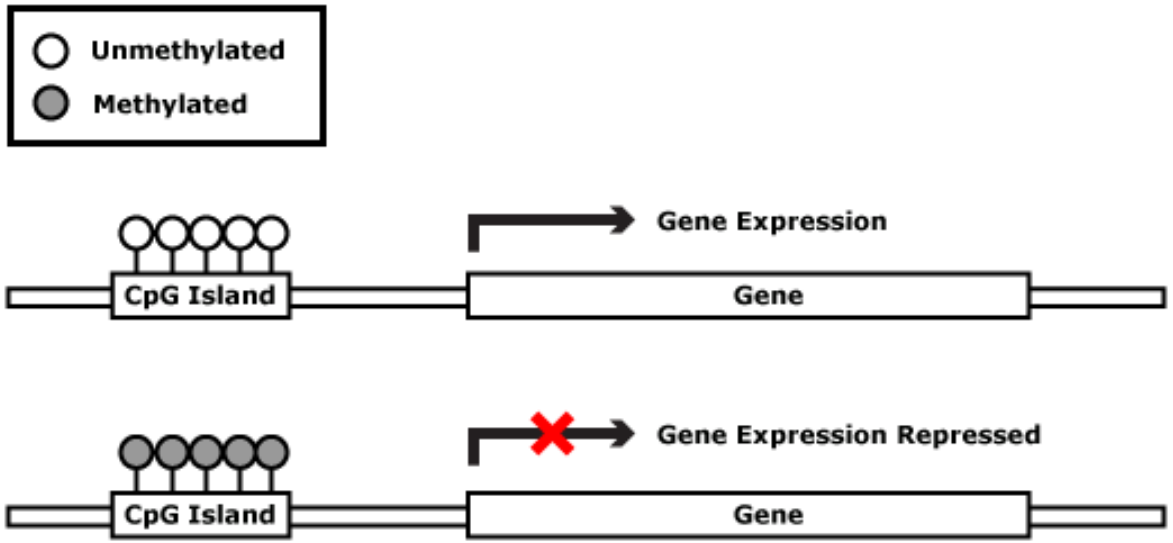
Distribution of CG



- CpG dinucleotides are not evenly distributed
- Most occur within high density regions called CpG Islands

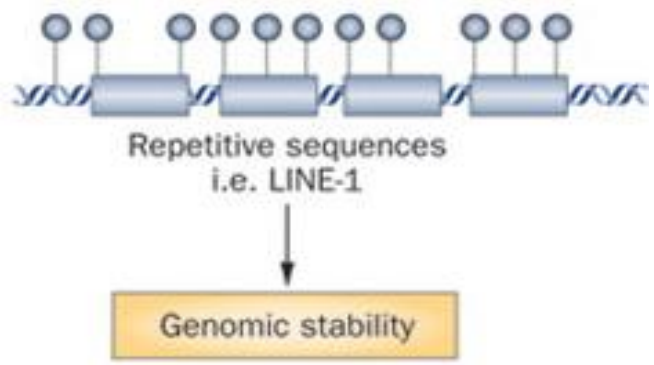
- Most of the genome contains methylated CpGs
- CpGs in CpG islands are largely unmethylated

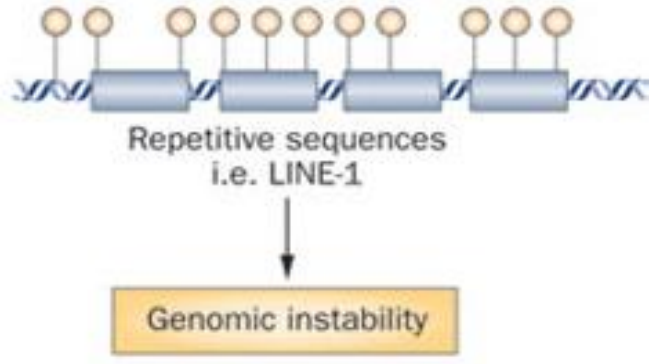
Regulation by DNA methylation



Silencing of gene expression
Tissue differentiation and embryonic development

Faults in correct DNA methylation may result in
- early development failure
- epigenetic syndromes
- cancer

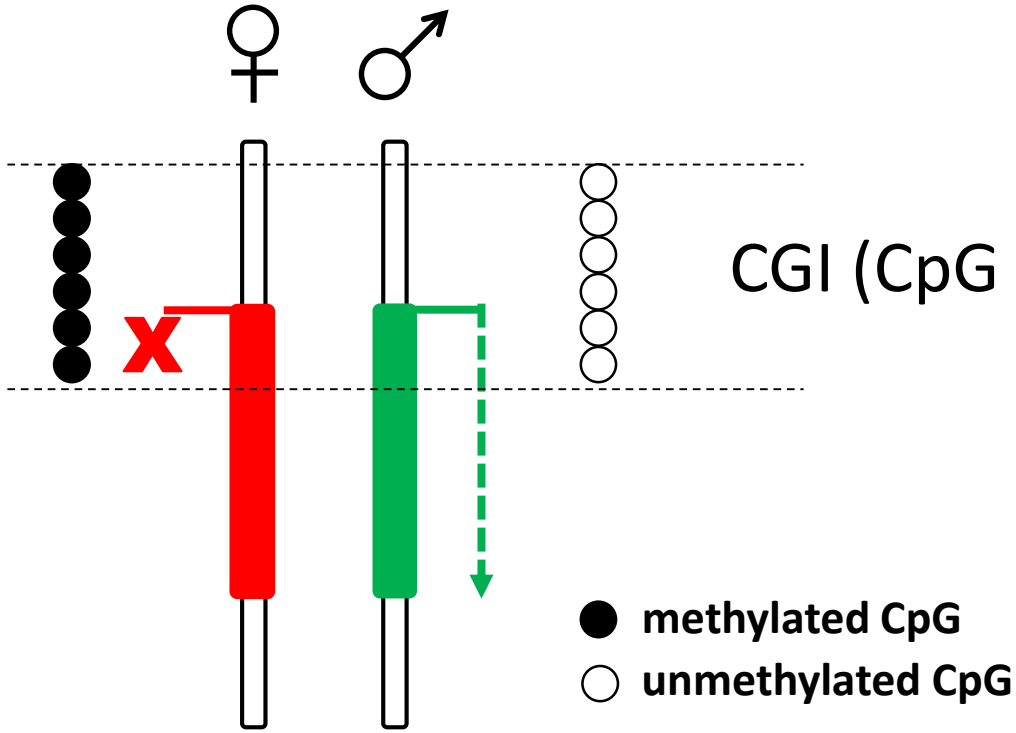




Repeat activity
Genomic stability

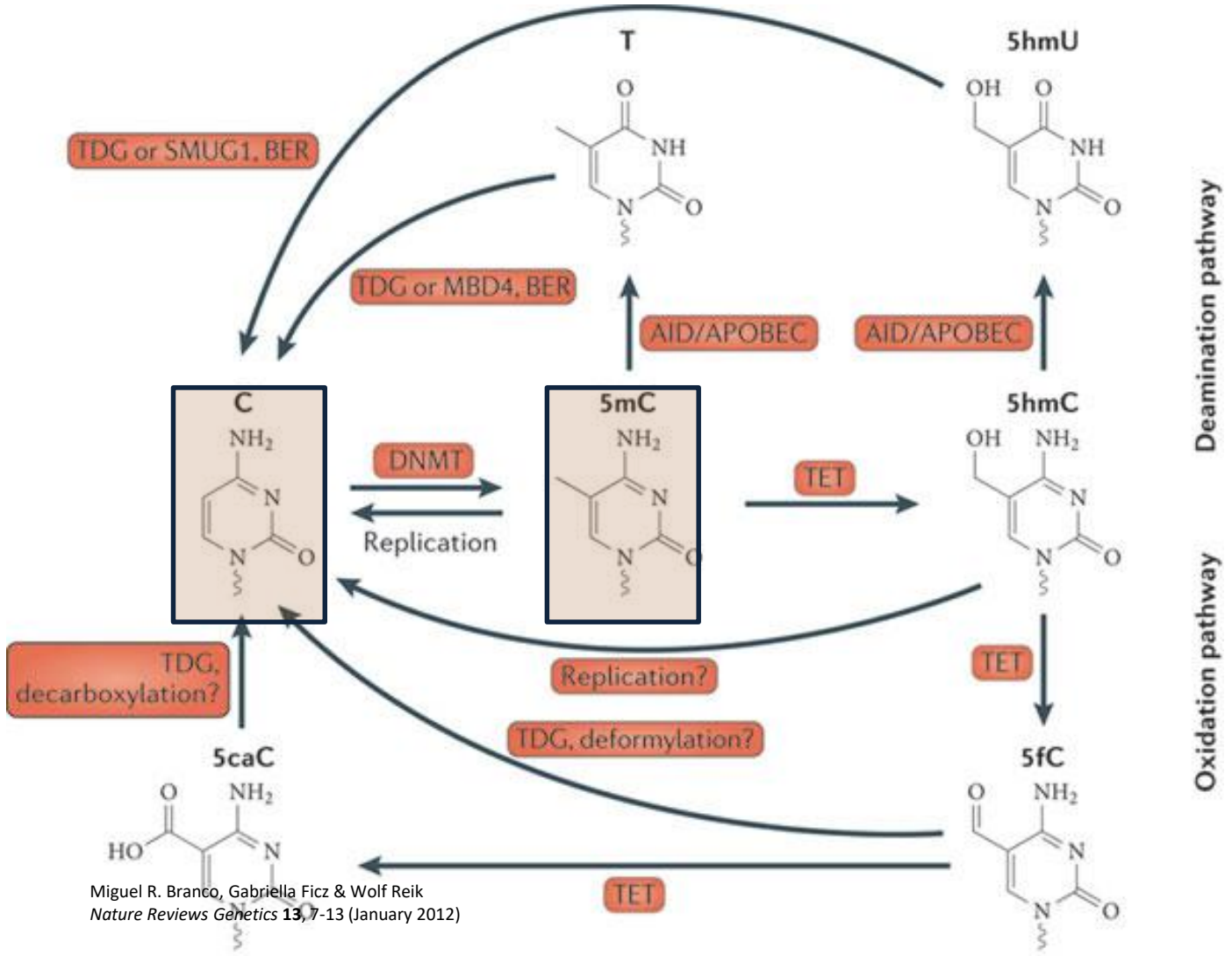
Genomic Imprinting: mono-allelic expression

Differential allelic
DNA methylation



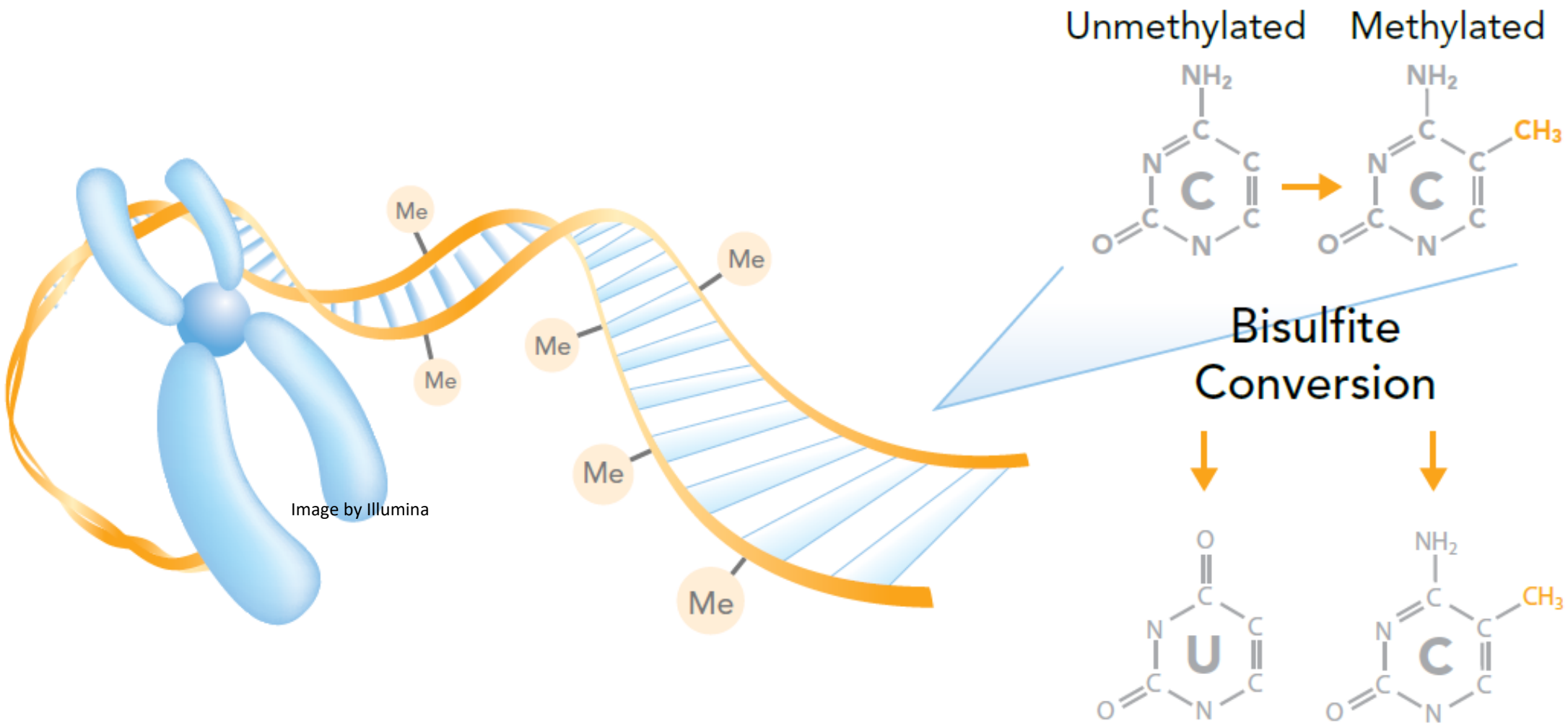
Imprinted Genes: Mono-allelic expression with parent-of-origin specificity.
Have key roles in energy metabolism, placenta functions.

Cytosine Modifications

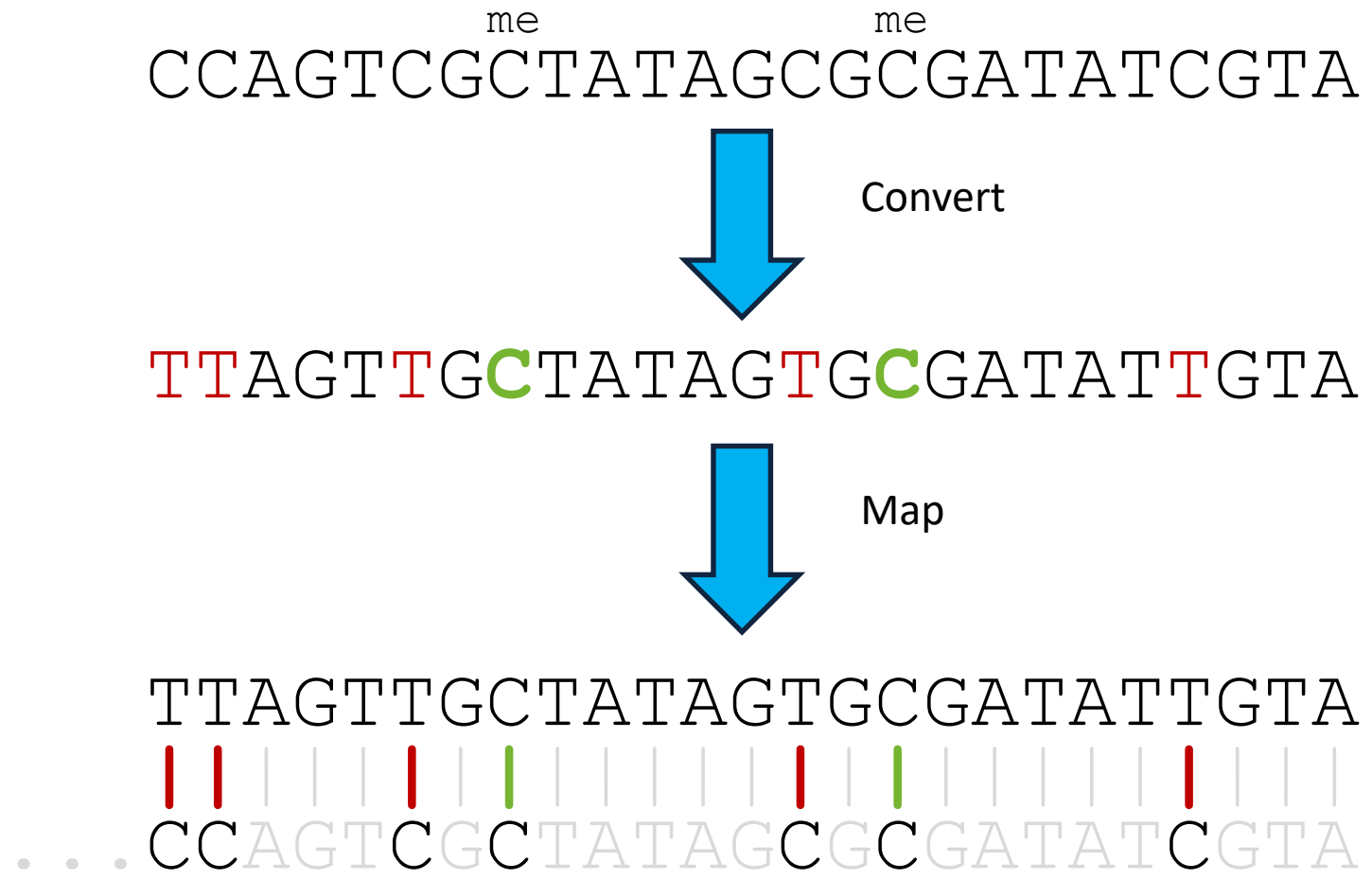


Miguel R. Branco, Gabriella Ficz & Wolf Reik
Nature Reviews Genetics **13**, 7-13 (January 2012)

Measuring DNA methylation by Bisulfite-sequencing

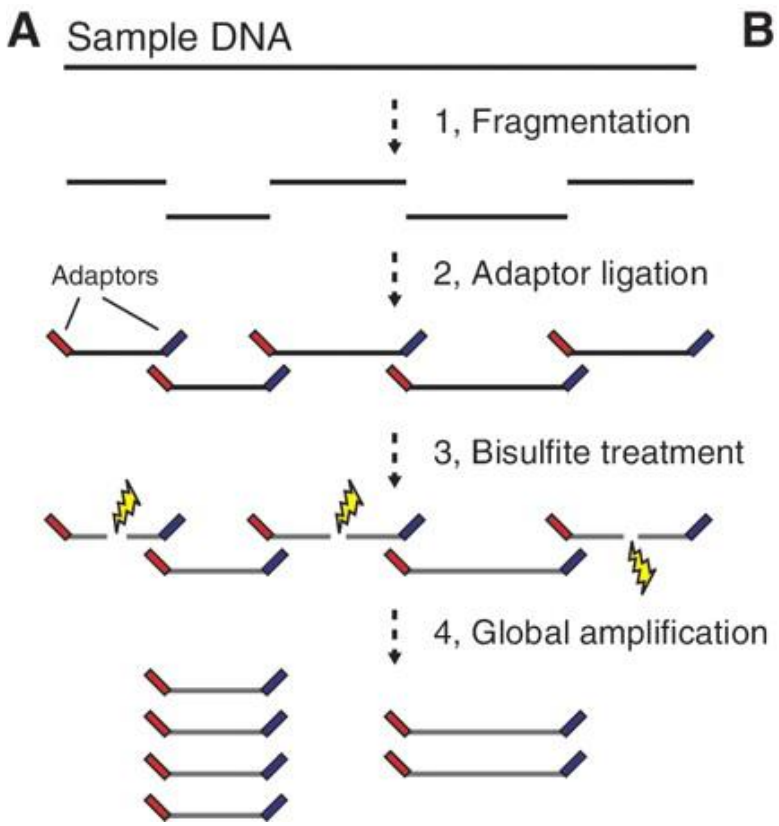


Bisulfite Sequencing

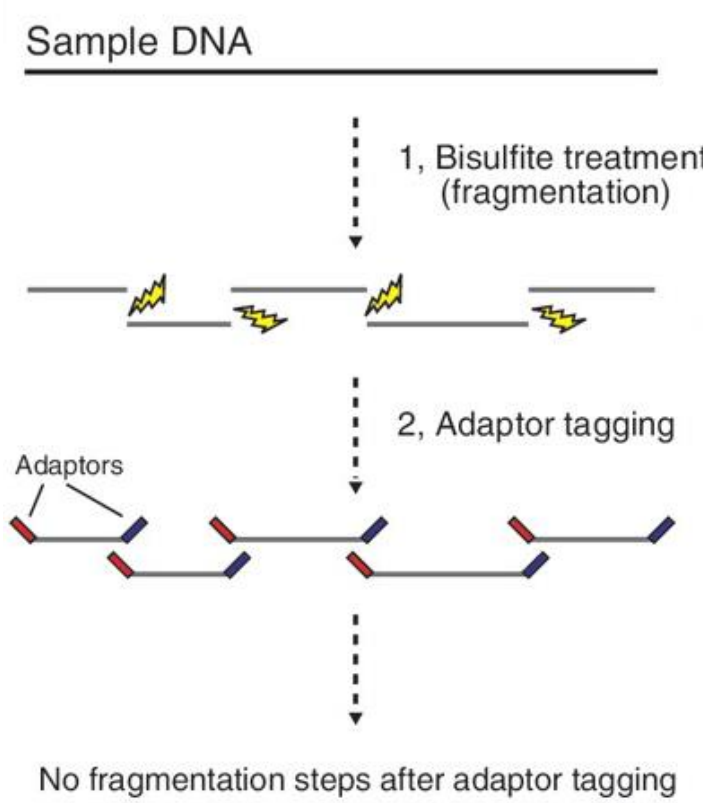


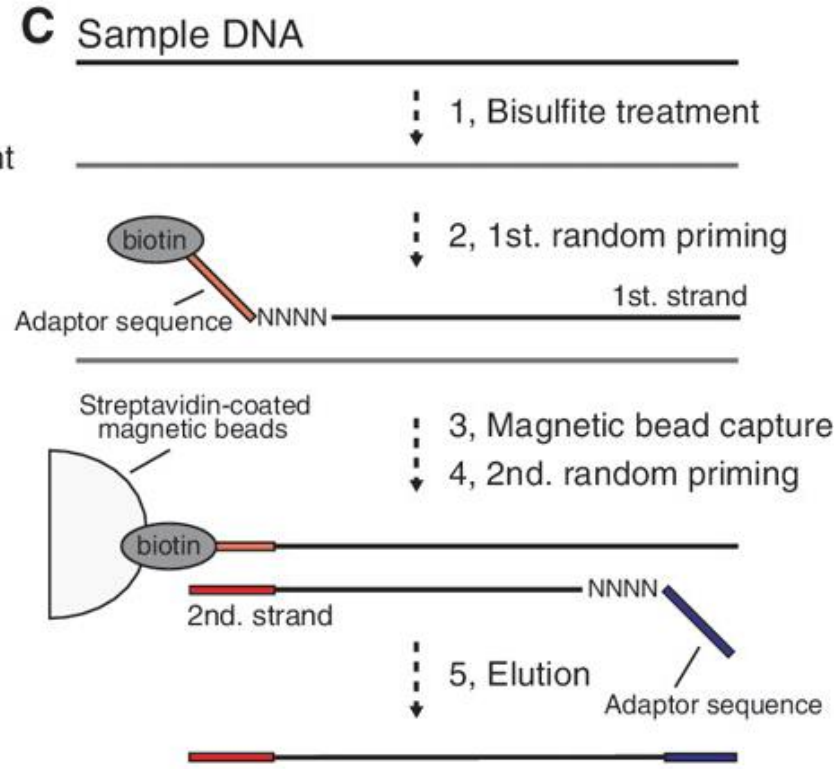
Bisulphite Library Preparation

Whole Genome Bisulphite Sequencing (WGBS)



Post Bisulphite Adapter Tagging (PBAT)





Enriched Bisulphite Sequencing

Reduced Representation (RRBS)

Cut with a restriction enzyme (MspI) whose recognition site contains a CpG

Size select for short fragments to enrich for CpG dense regions

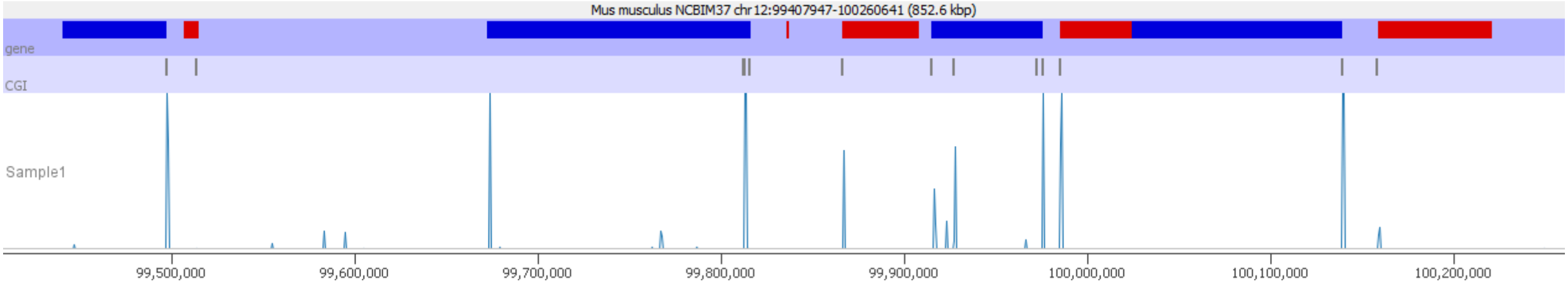
Enrichment Kits

Oligonucleotide hybridisation pull down systems

Can enrich for anything, special kits for methylation

CpG Islands and 'shores'

Relevant promoters and enhancers

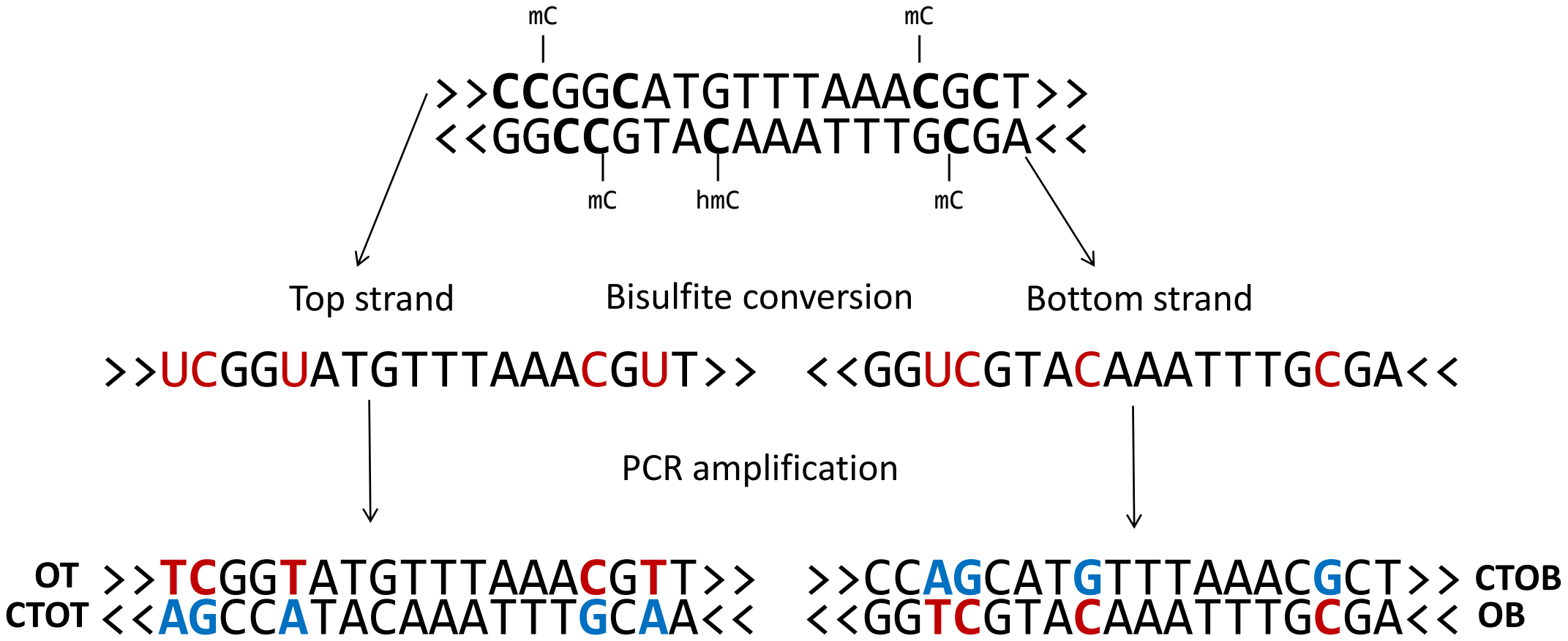


BS-Seq Analysis Workflow

Explore and understand your data

Sequencing → Processing pipeline → Methylation Analysis

Bisulfite conversion of a genomic locus

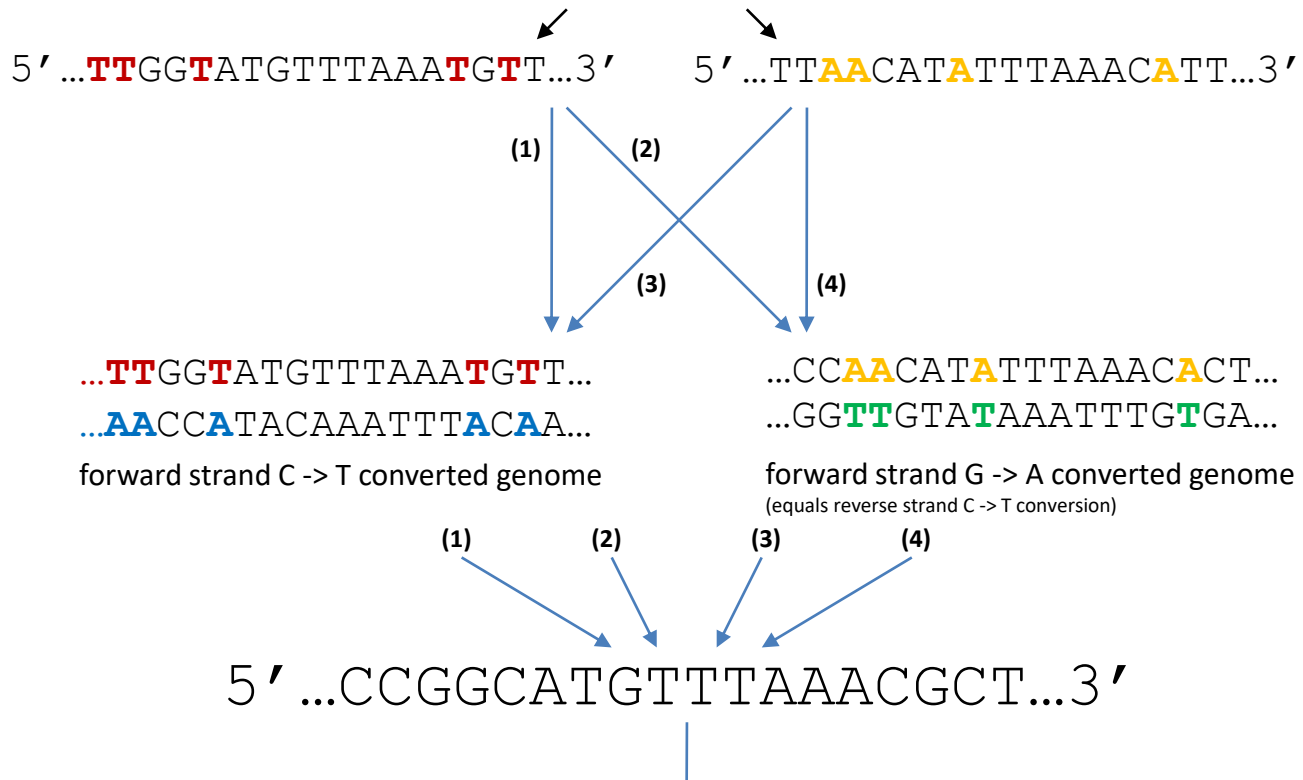


- 2 different PCR products and 4 possible different sequence strands from one genomic locus

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

3-letter alignment of Bisulfite-Seq reads

sequence of interest TTGGCATGTTTAAACGTT



Fully bisulfite convert read
(as both forward and reverse strand)

Align to bisulfite converted genomes

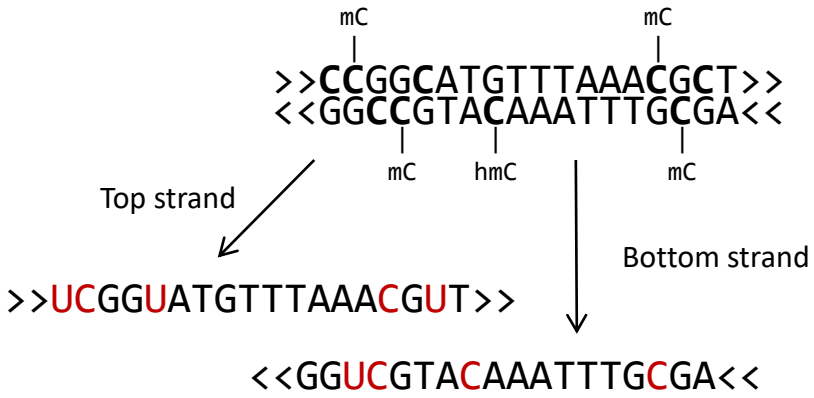
Read all 4 alignment outputs and extract the unmodified genomic sequence if the sequence could be mapped uniquely

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

Bismark

Common sequencing protocols



1) Directional libraries

(vast majority of kits, also EpiGnome/Truseq)

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

2) PBAT libraries

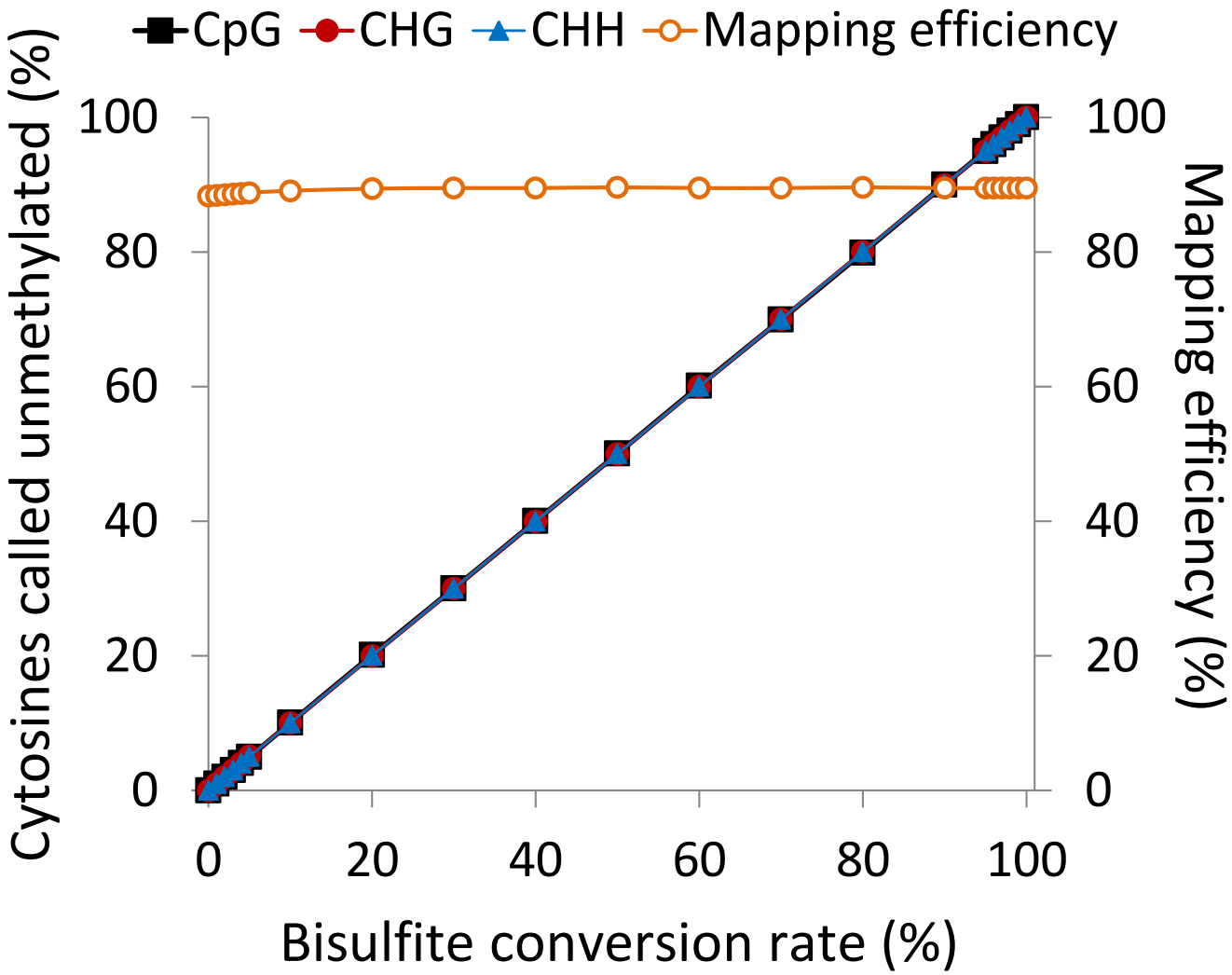
CTOT <<AGCCATACAAATTTGCAA<<
>>CCAGCATGTTTAAACGCT>> CTOB

3) Non-directional libraries

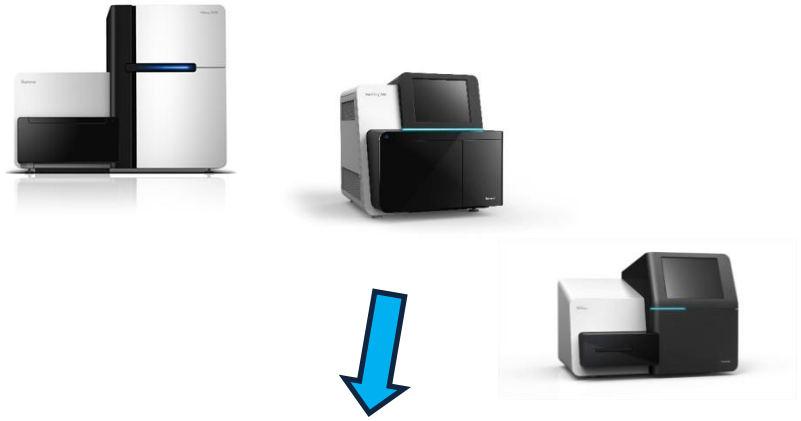
(e.g. single-cell BS-Seq, Zymo Pico Methyl-Seq)

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

Validation


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

BS-Seq Analysis Workflow



Raw Sequence Data (FastQ file)

```
@HS31_12166:1:1101:5279:2453#2/1
ATTTTCCCCTAAAAAAACCACTTCCGCCACTCCCAACTTTACTCAATTTCTTATAAATCTTTATATATAAAATTAAAATCTCCTAAACTTTCCCCTATTC
+
@C@DFD;DFFDHFHGEEECGHFHHICDFFGHDGHGDGHIGDFGG@8CHIHHIHIGIGHCHHHCHBE@D>BCEEEC;>CDACCCCCCCCCCCCAACCCBCCC
@HS31_12166:1:1101:5276:2474#2/1
AGGTTTGTTGAGGTAATTTTTTTTTTTATATTTTATTATATTTTTATAGAATATATATTTTTTTTTTTTTTTTTTTTTTATTAATATTTTAAATATTTTAT
+
BBCDDFDBHFDCDCGIIJJJJJJJJJJGIEIHJJIJJGH@GHIJJ=DAEEEHEFDFFFFFFEDDDDDDDD-9BDDDDD(:@:>:(+(4>:C@((4:(((4
@HS31_12166:1:1101:5376:2480#2/1
GCCCTTCAAAAAAAAATAATATTAATTTTACTTACTAAAAACTTTAATATCCCATTAAAAAAATATTAAACTACATATACCGTTATCAATATTATTTAAA
+
CCCFFFFFHHHHHJJJJJJJIJJJJIJJJJJJJJJJJJJJGHJJHJJJJHIJIJJJJJHHFDEEEFEEDDDDDDDFEEEDDDDDDDEDDEDEEEEEED
@HS31_12166:1:1101:5674:2287#2/1
TAAACCTAGATATATAACCATCTACTAACTAAATTTCATTTACTAAAAACTATCATTTTTAAAAAATCTAAATTTTTCTCTCTCAACAACTAACAATTAT
+
BBCFDEFFHFHHHHIJJFHIIJJIHHHJGHHGHIJJJIJJJHJJJJJJJJJJJJJJJJJJIJJJJJIJJJJJJJJJJHHHHHHHFFFFFDDEEEDDDDDDE
@HS31_12166:1:1101:5575:2309#2/1
ATCACACCCTACCTCTAAATACATCTATAAATCTCAATATACCACCGTCAAACATAATTCAAAACCAATCCTCCCACCACACTAAAAAAACAATAAACTA
+
BBCFDDDFHHFHDGHIHIIJHIHIFHHIIJHIHIHIIJIJJHGGHGIHIIIJJJJJJJJJJJJJJJJJJIJHHHHHFFFDDDEEDDDDDDDDDDDDDDD
@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

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

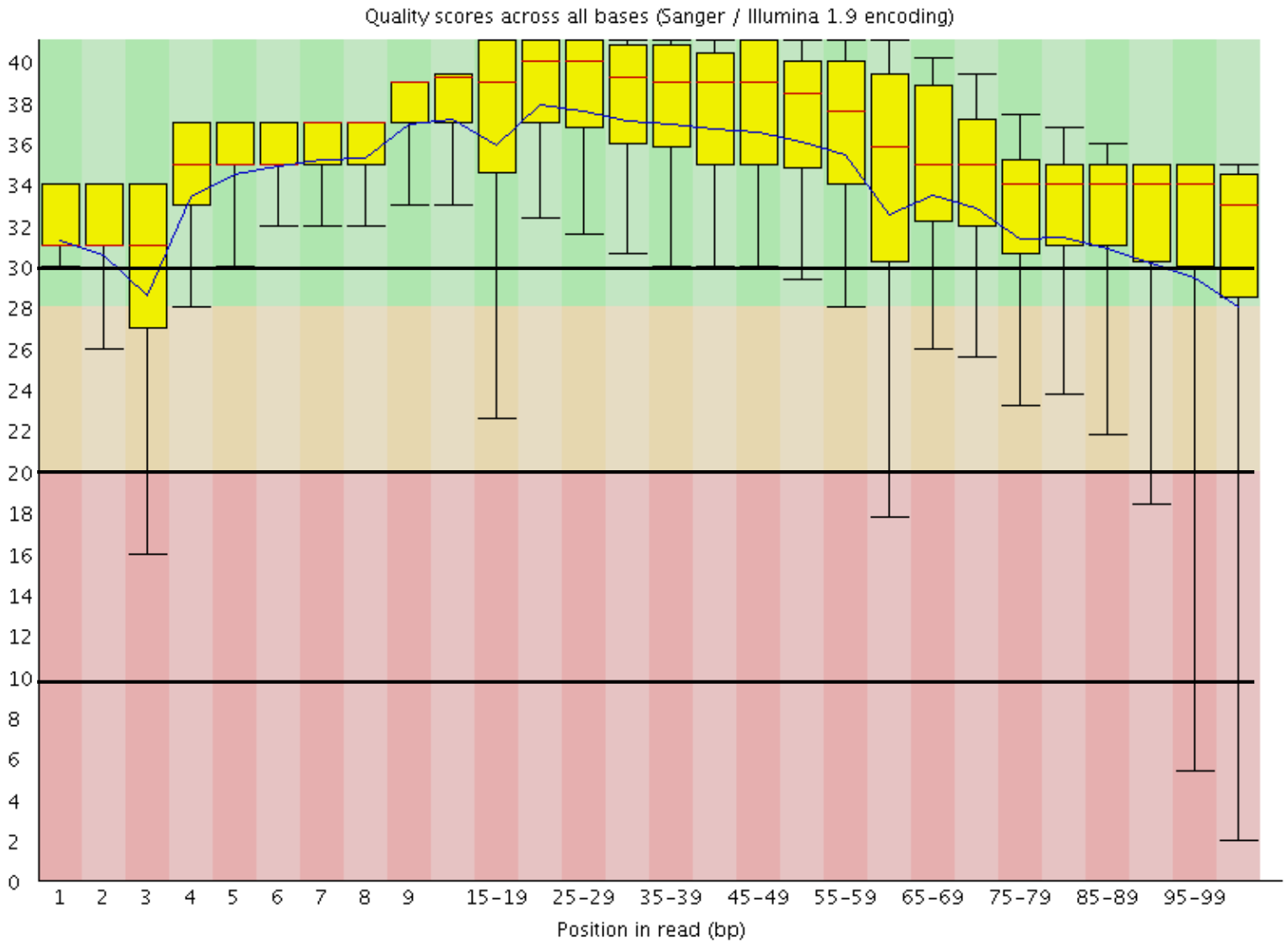
QC Raw data: Sequence Quality

Error rate

0.1%

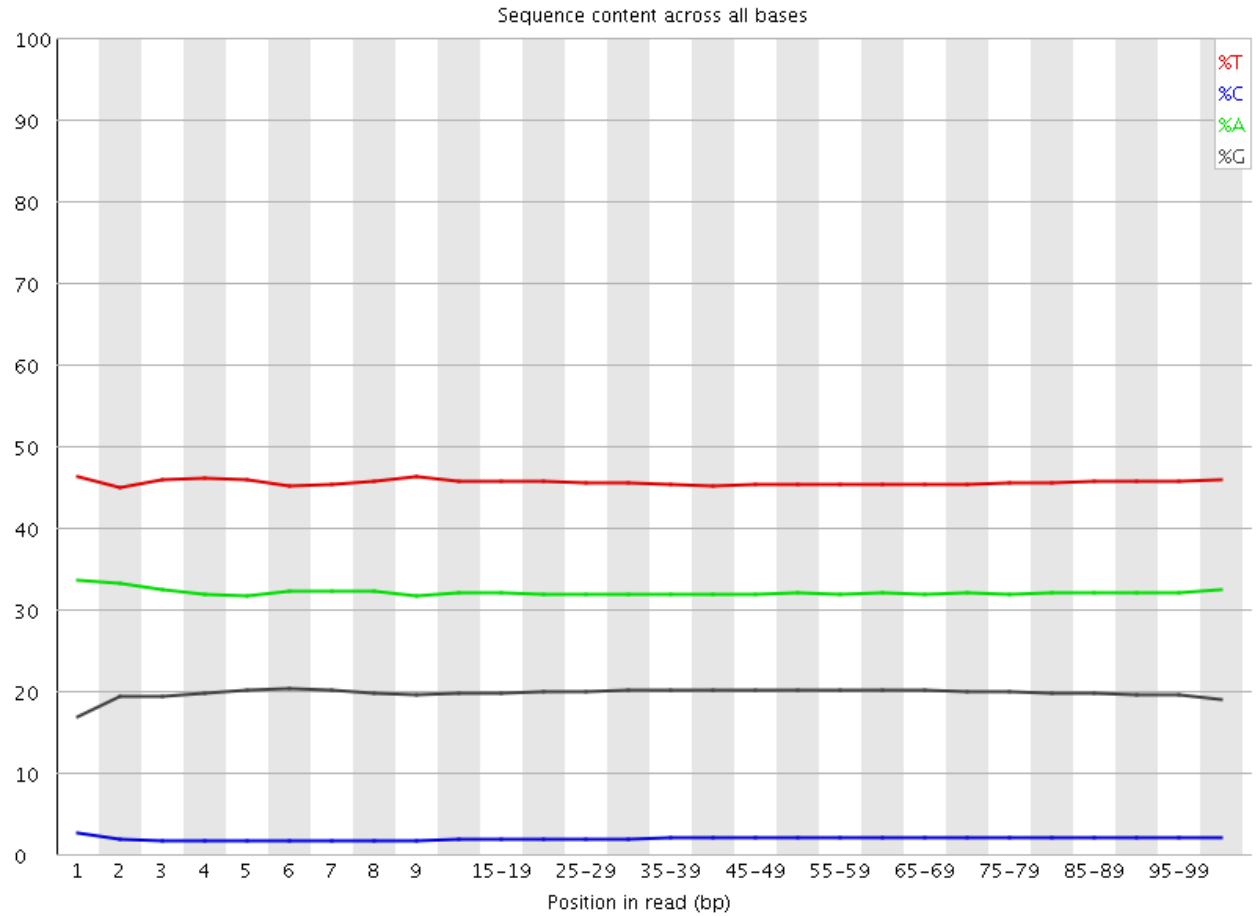
1%

10%

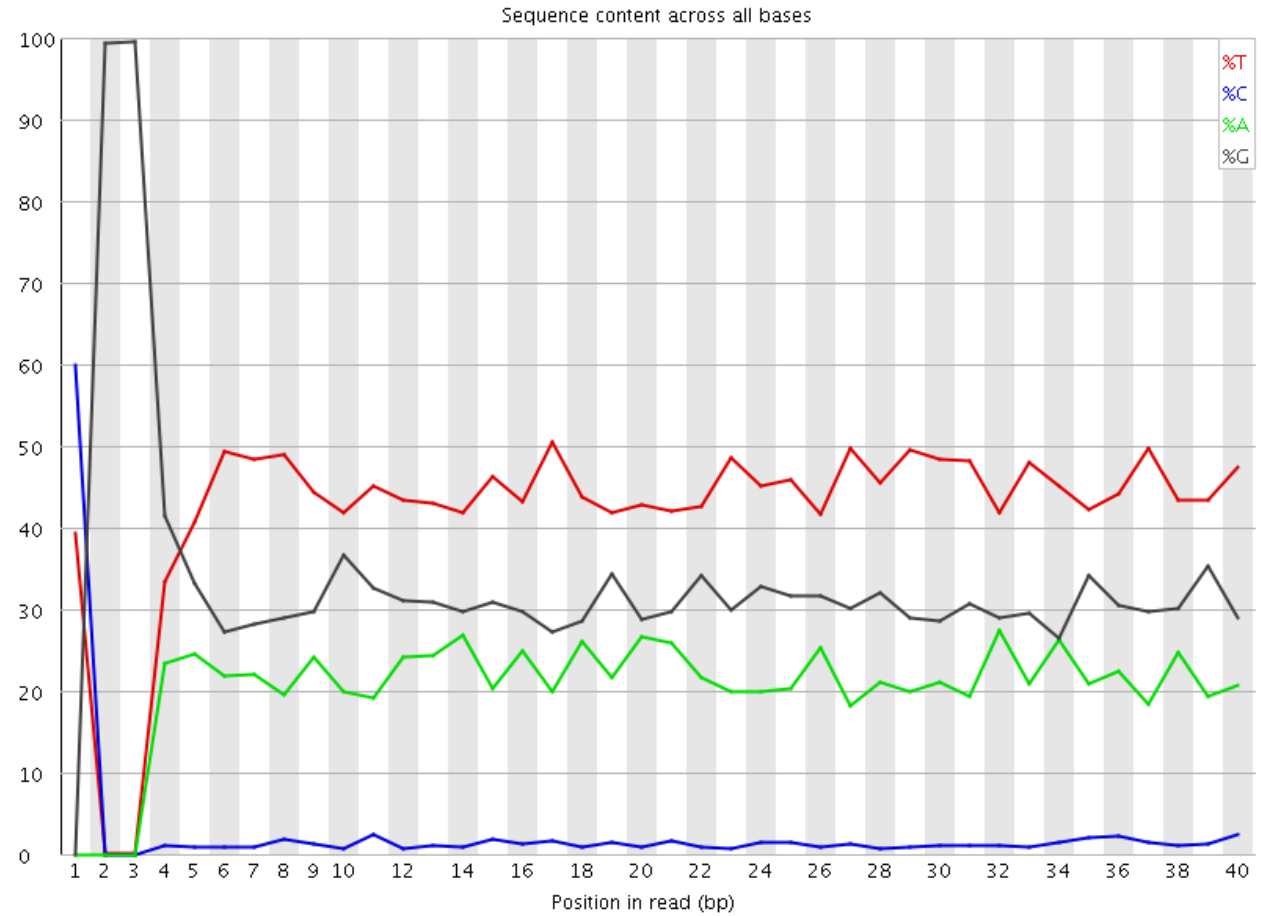


QC: Base Composition

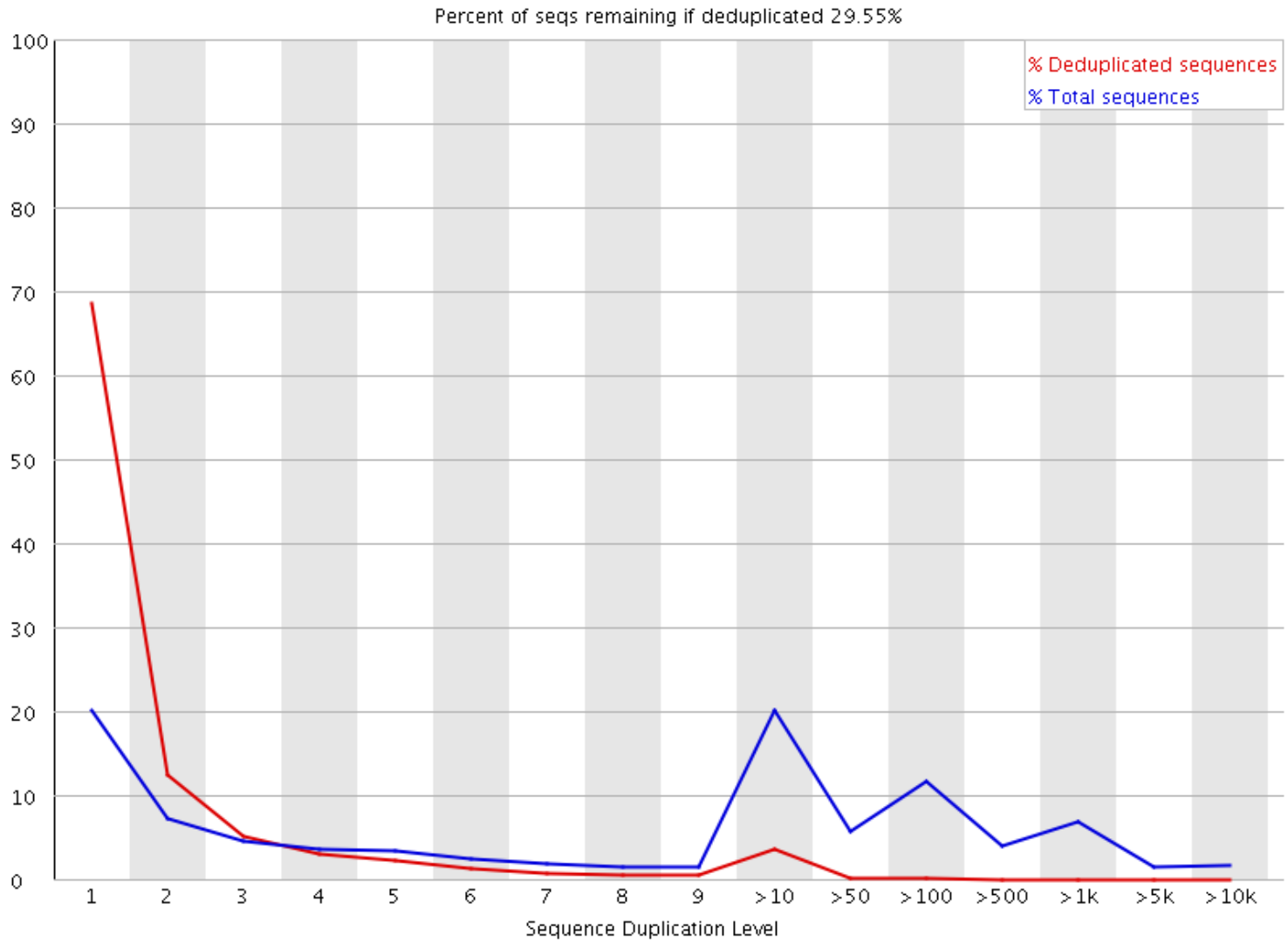
WGSBS



RRBS



QC: Duplication rate

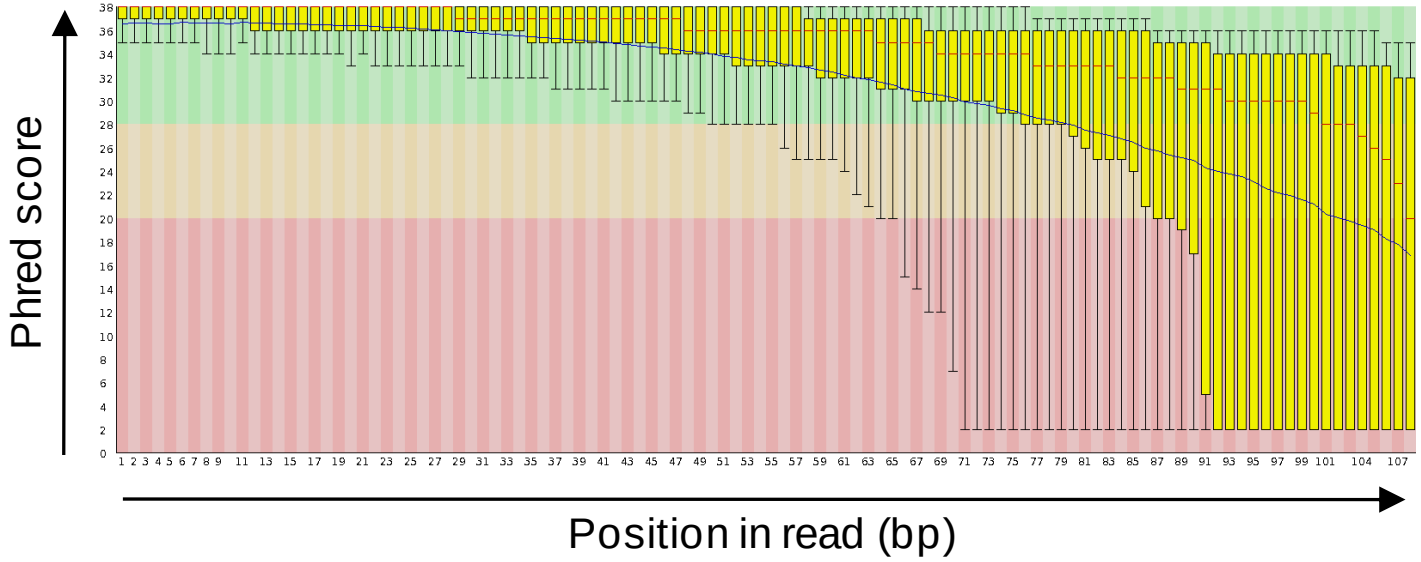


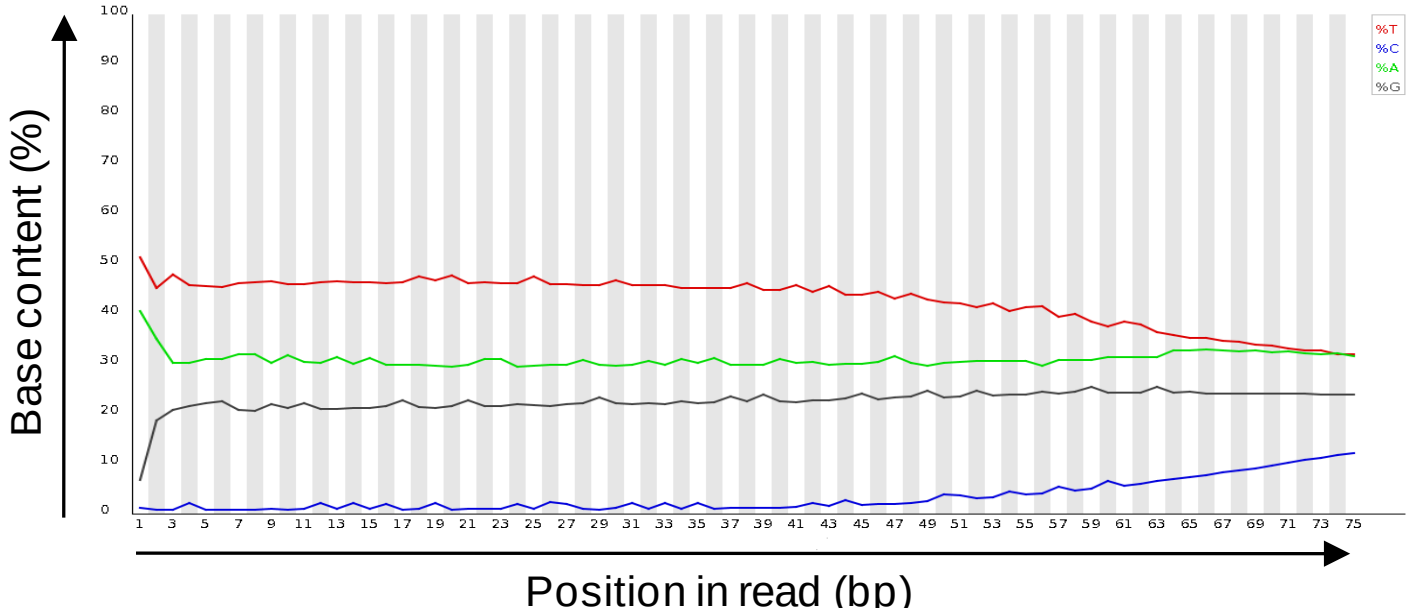
QC: Overrepresented sequences

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

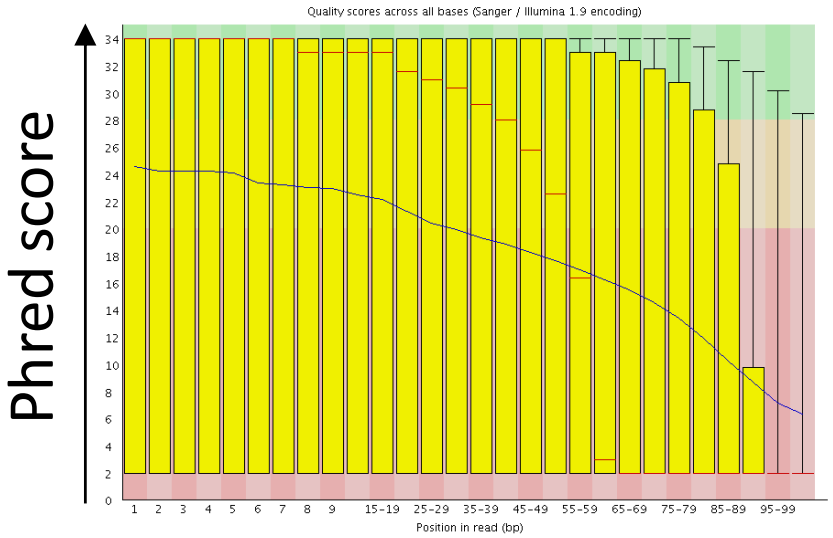


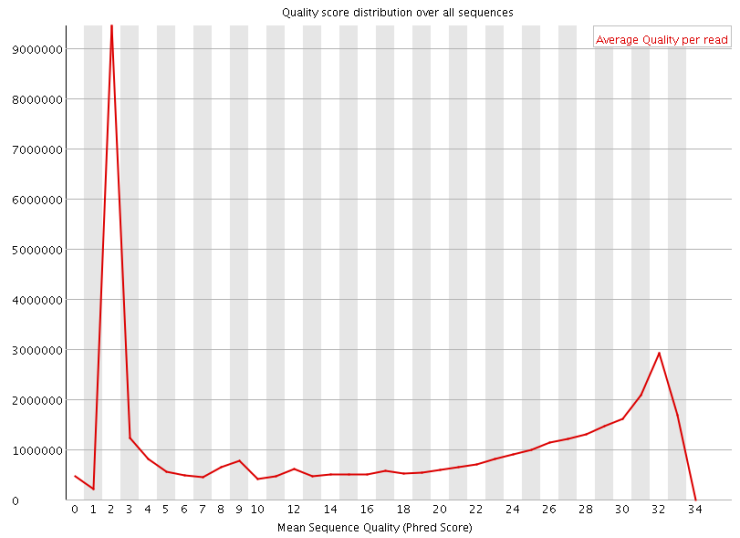


Not observed
in ‘normal’ libraries,
e.g. ChIP or RNA-Seq

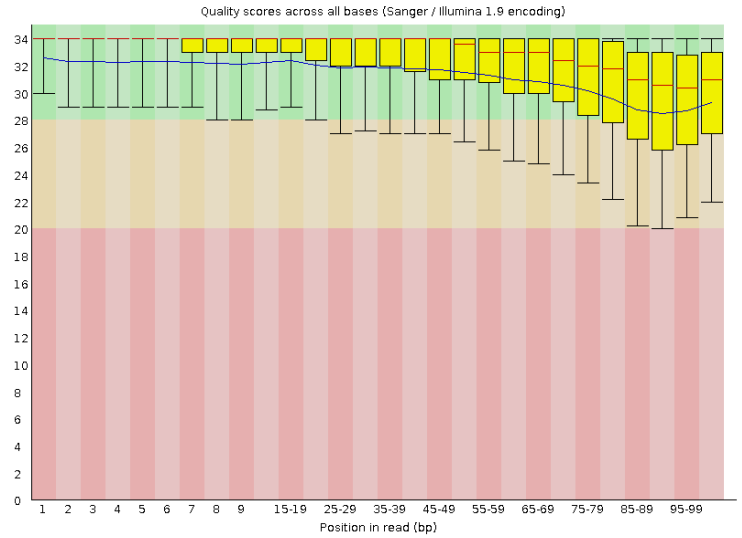
Removing poor quality basecalls

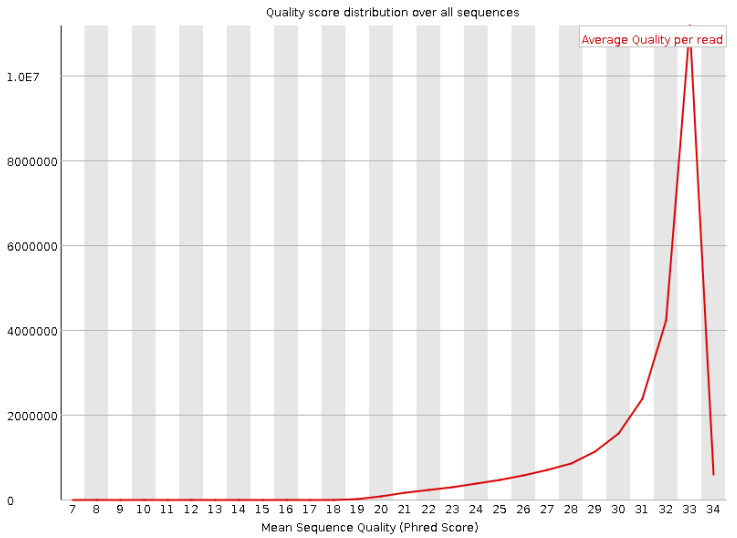
before trimming





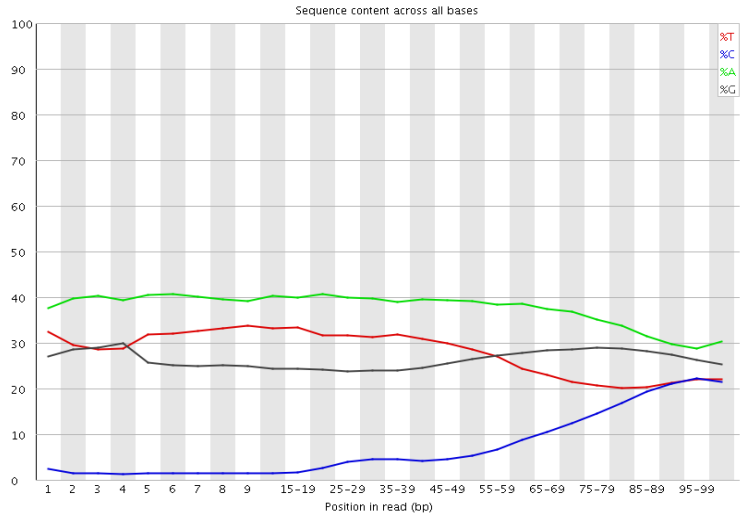
after trimming

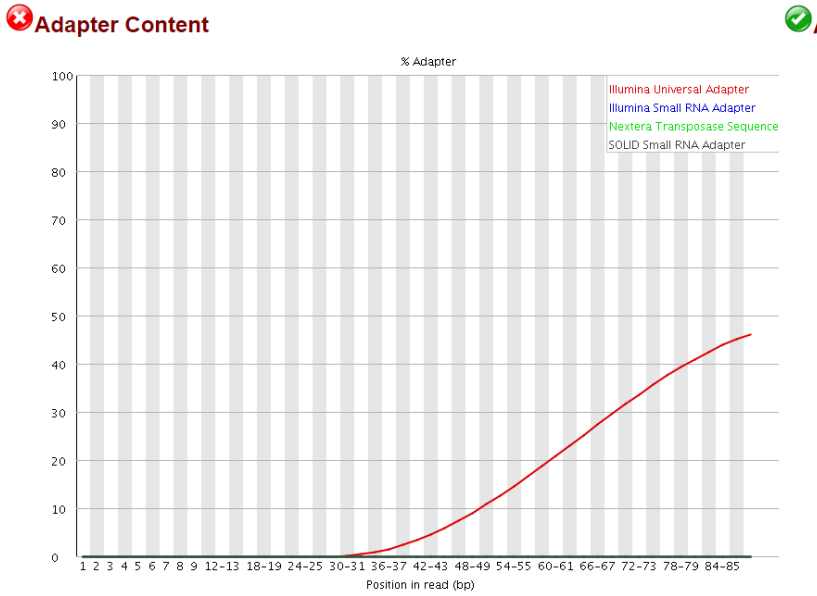




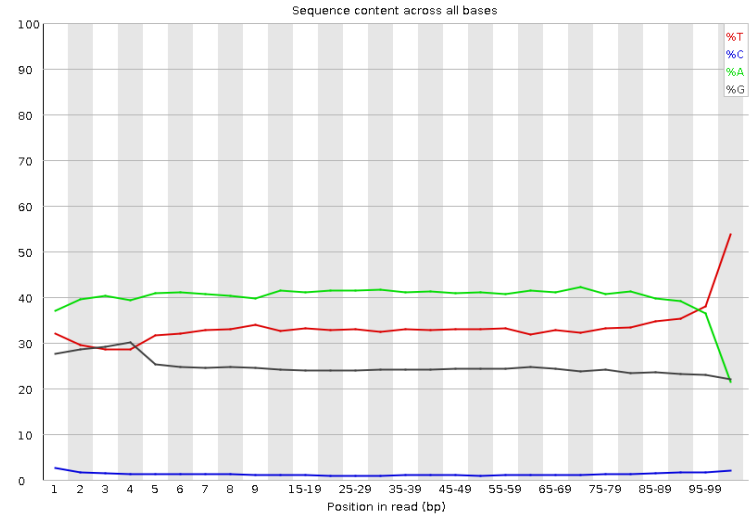
Removing adapter contamination

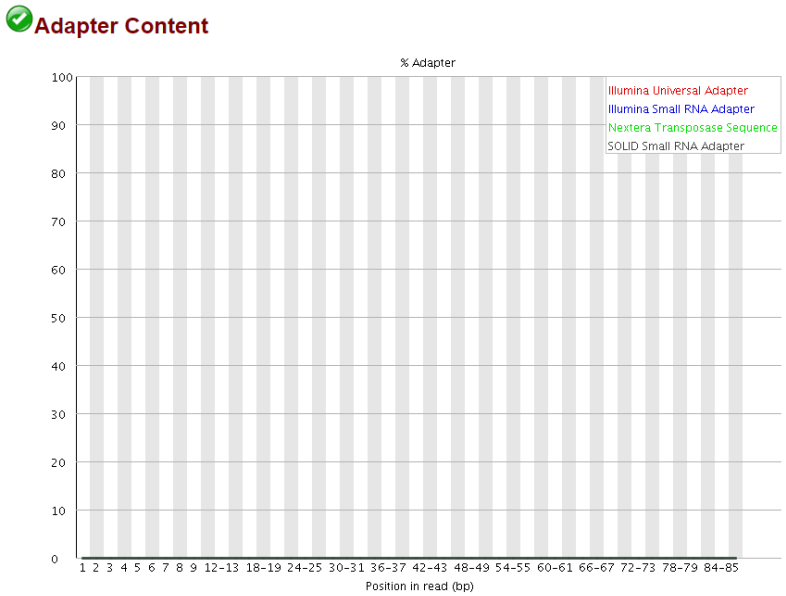
before trimming





after trimming



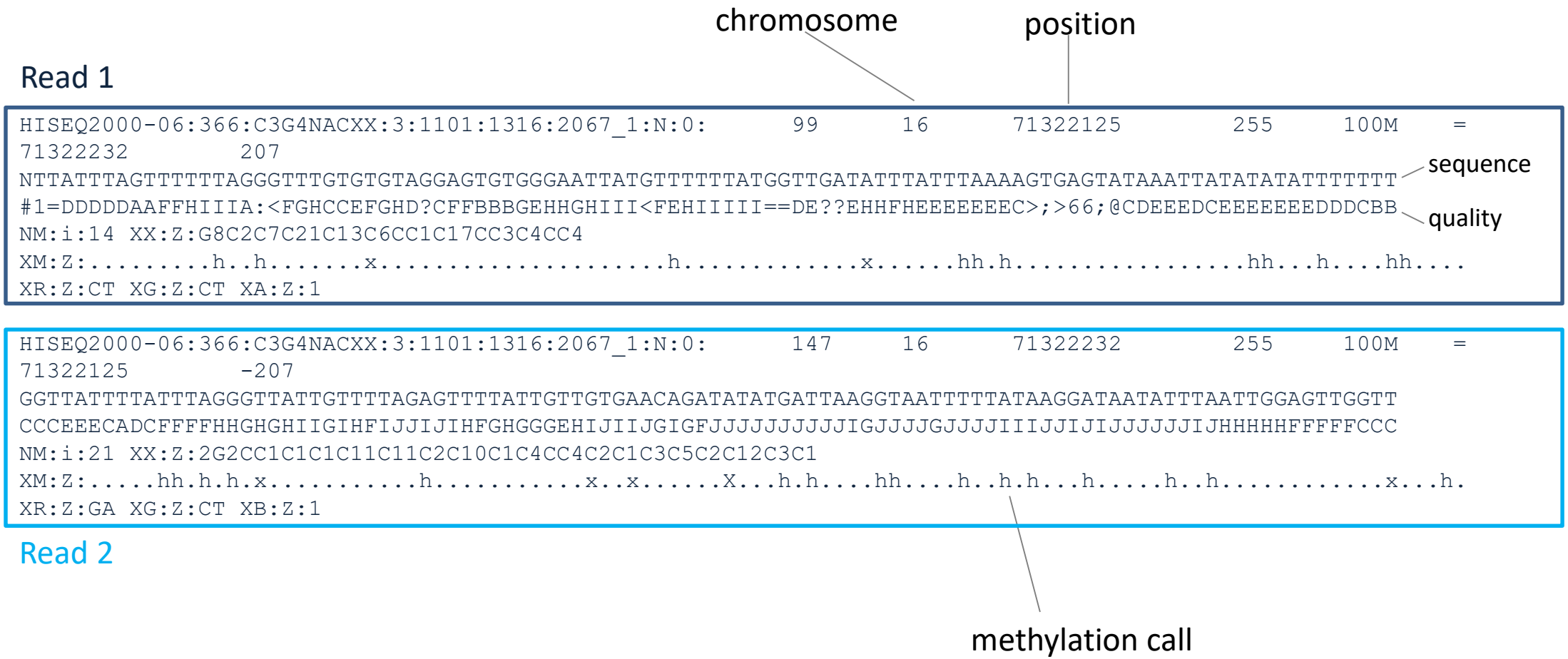


Summary Adapter/Quality Trimming

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

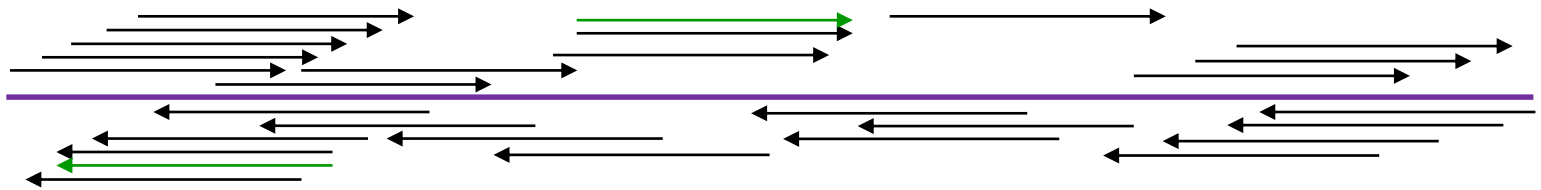
- Low mapping efficiency
- Mis-alignments
- Errors in methylation calls since adapters are methylated
- Basecall errors tend toward 50% (C:mC)

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

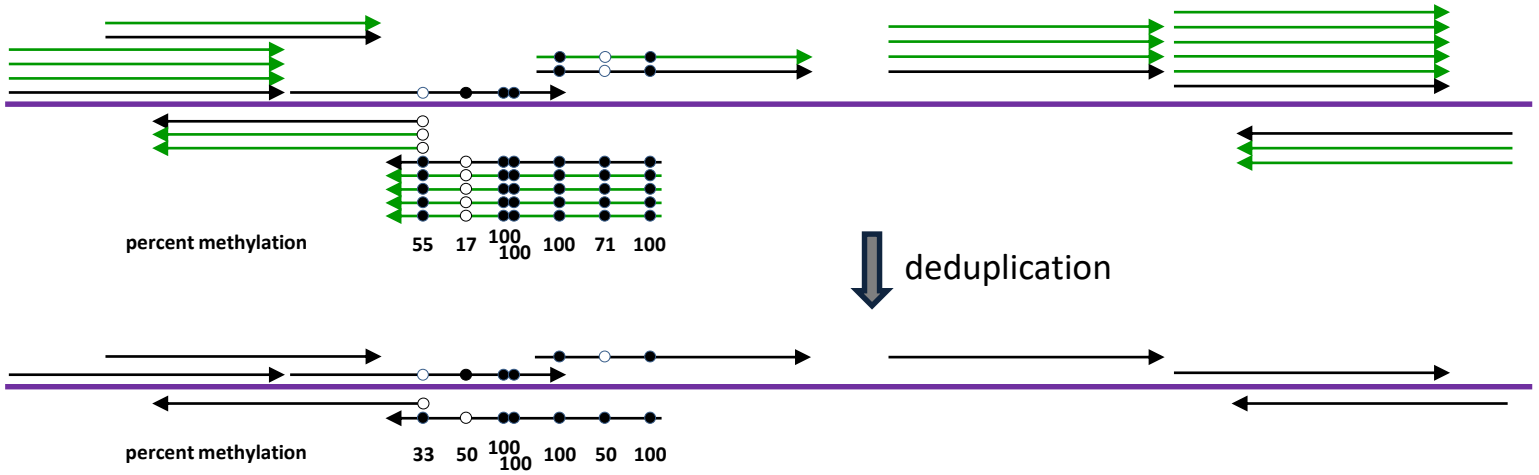


Sequence duplication

Complex/diverse library:



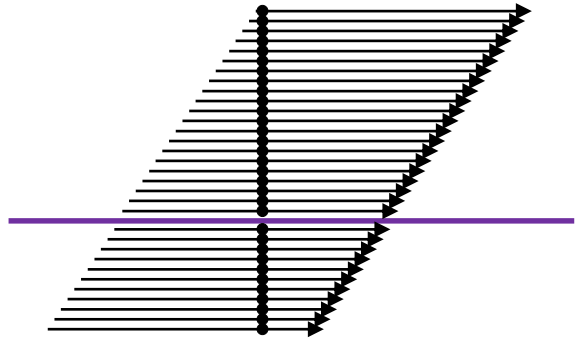
Duplicated library:

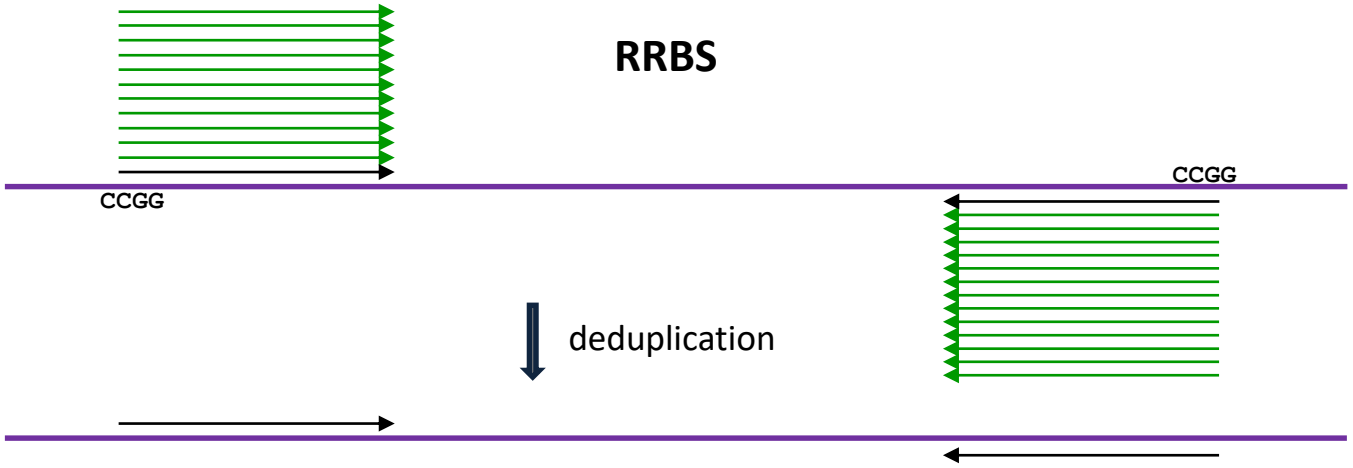


Deduplication - considerations

Advisable for large genomes and moderate coverage

- Unlikely to sequence several genuine copies of the same fragment by chance
- Could limit coverage in high coverage studies, but would need to be very deep

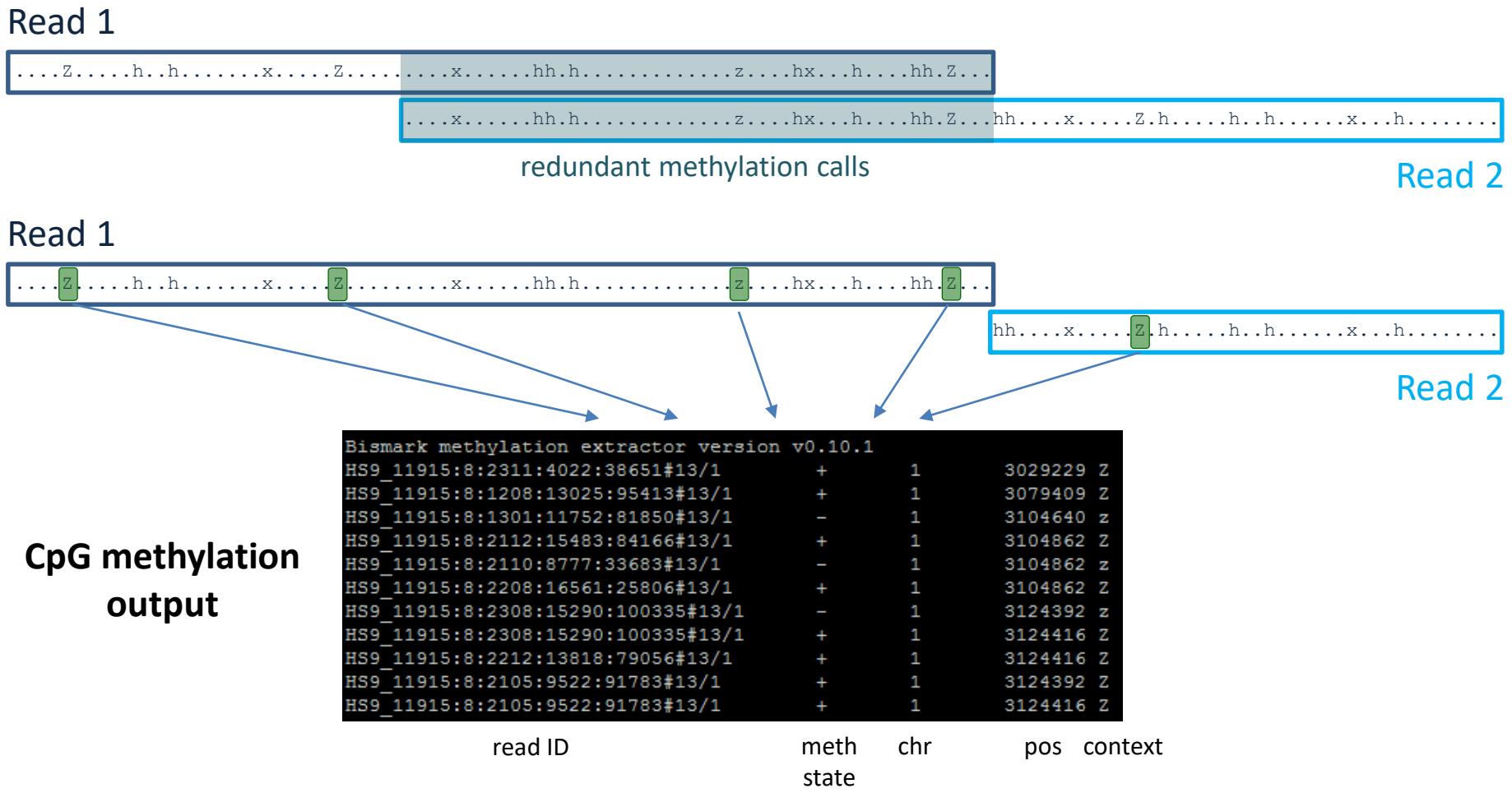




NOT advisable for RRBS or target enrichment methods

- Non-random start positions (restriction sites)
- Higher local density means random collisions are likely

Methylation extraction



Methylation extraction

```
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

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

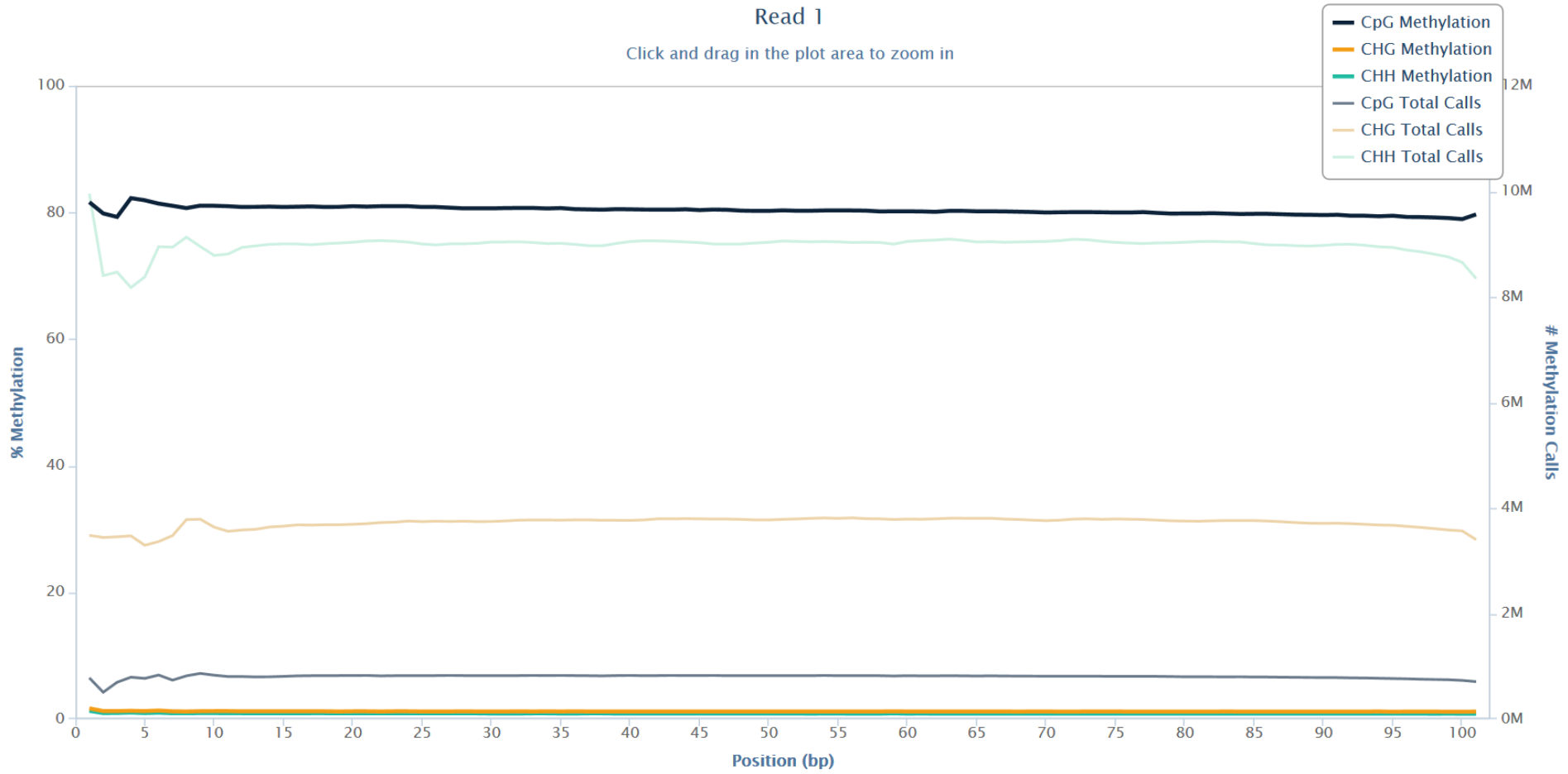
chr pos strand meth unmeth di-nuc tri-nuc

Optional: merge into CpG dinucleotide entities

Genome wide CpG report

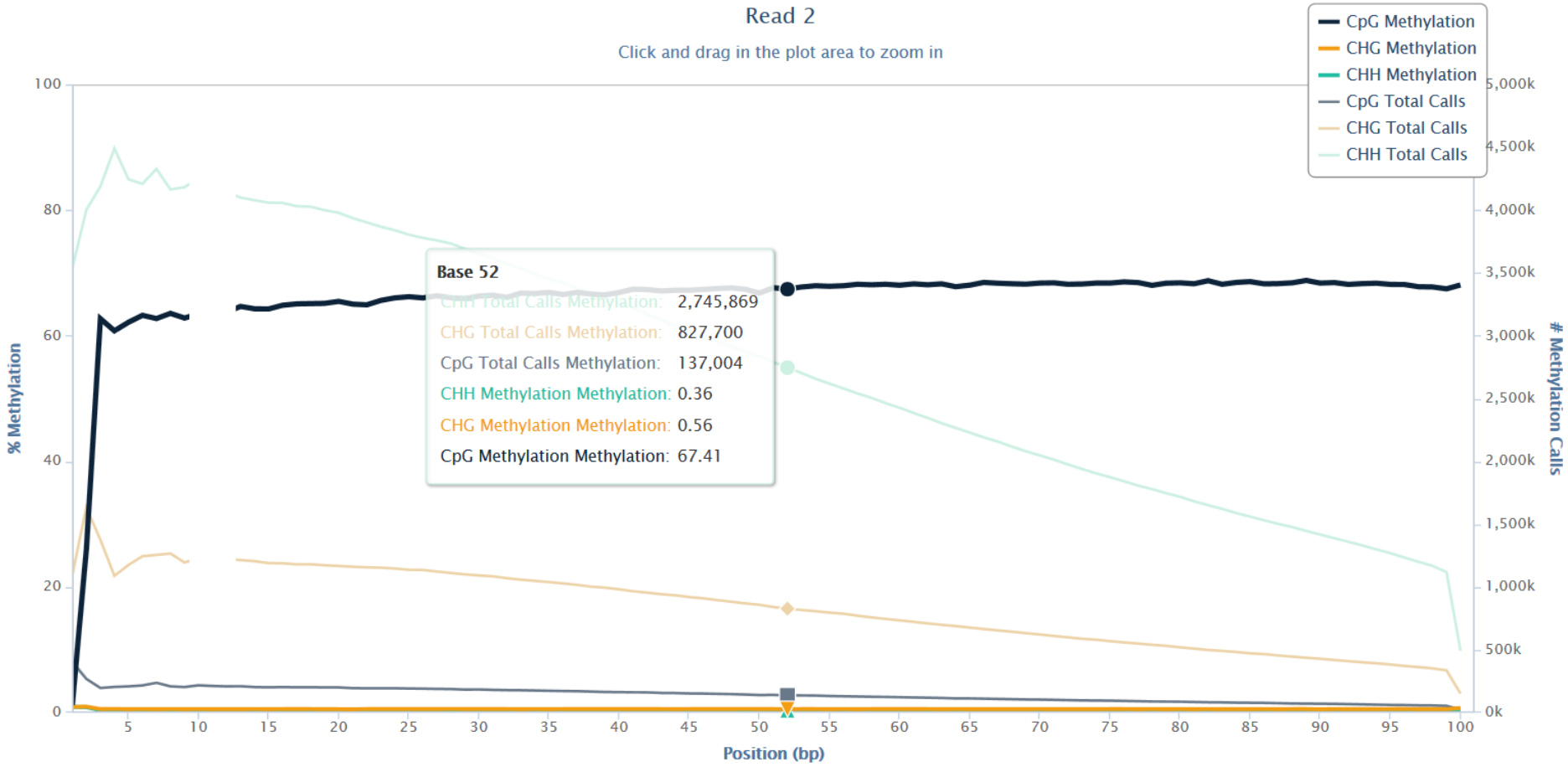
Part III: Mapped QC - Methylation bias

M-Bias Plot



good opportunity to look at conversion efficiency

Artificial methylation calls in paired-end libraries



end repair + A-tailing

5′– **GGG**NNNNNNNNNNNNNNNNNNNNNNNNNNNN**CCCA** –3′
3′– **ACCC**NNNNNNNNNNNNNNNNNNNNNNNNNNNN**GGG** –5′

Specialist applications

(e)RRBS

WGBS

PBAT

NOMe-seq

target enrichment

amplicon

single-cell

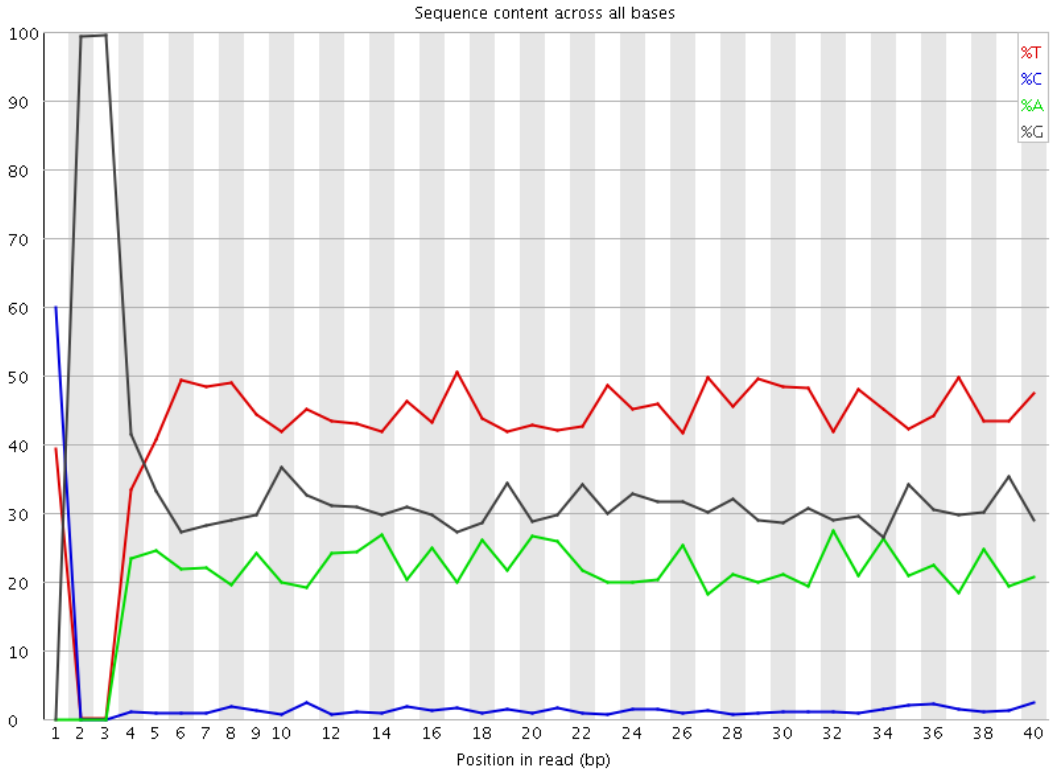
NMT-seq

non-directional

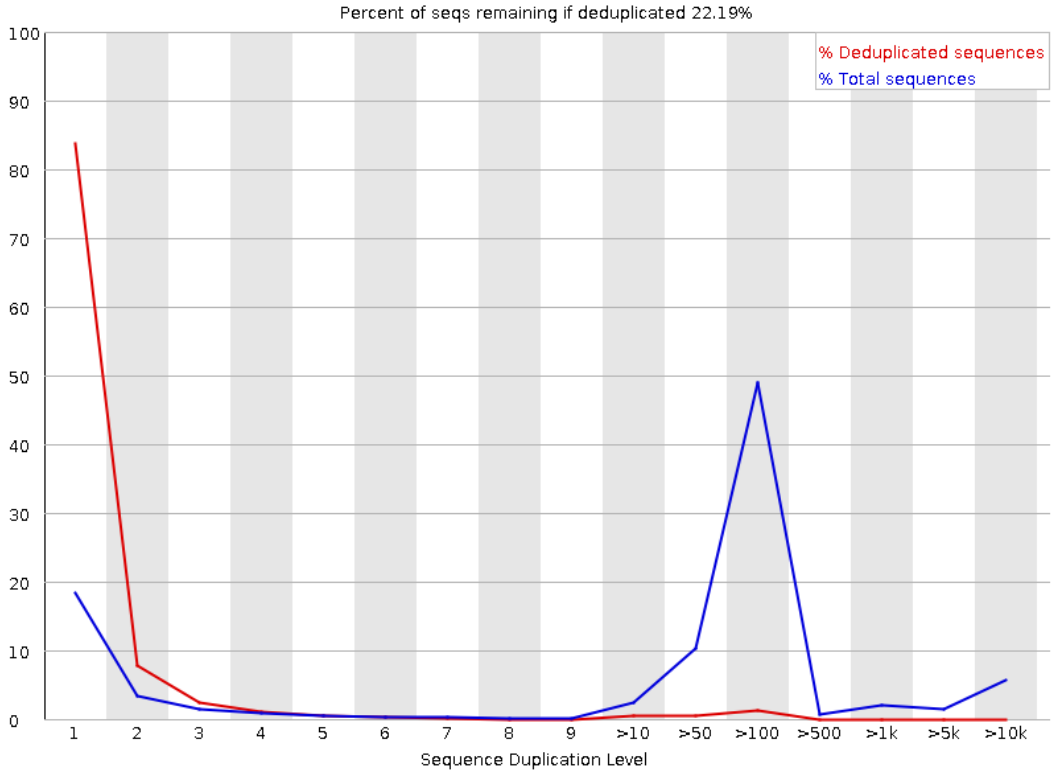
+ different library kit protocols

Reduced representation BS-Seq (RRBS)

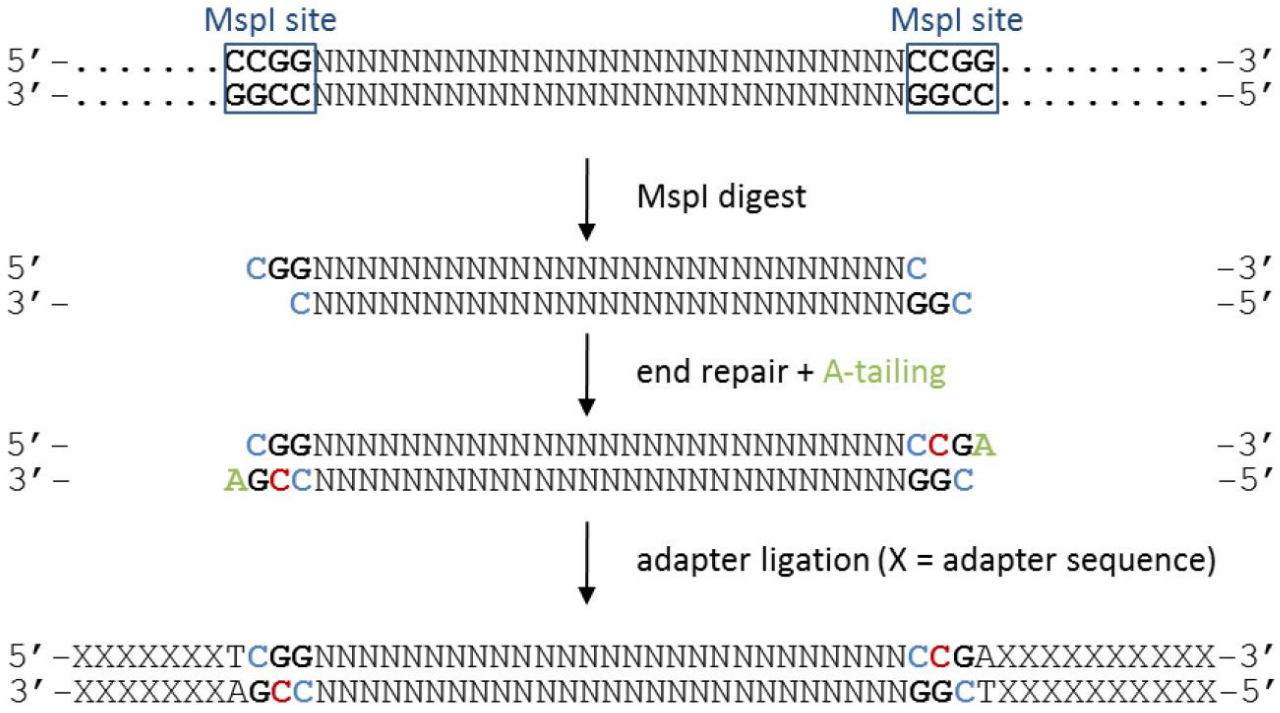
Sequence composition bias



High duplication rate



Artificial methylation calls in RRBS libraries

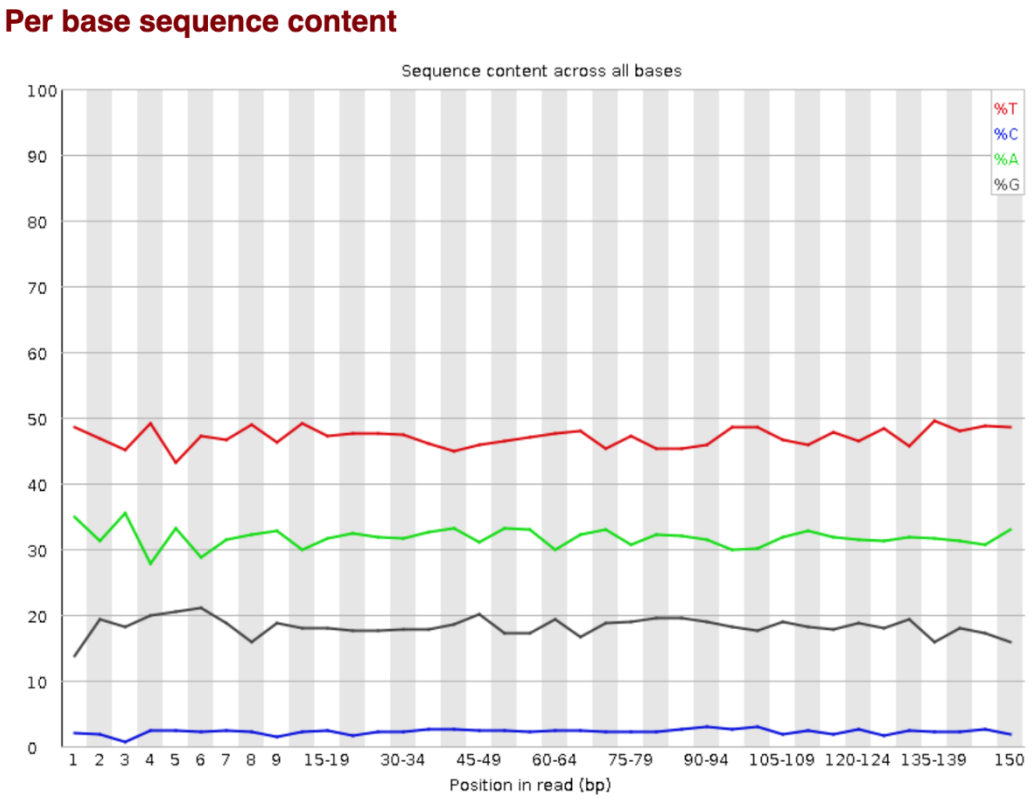


C genomic cytosine
C unmethylated cytosine

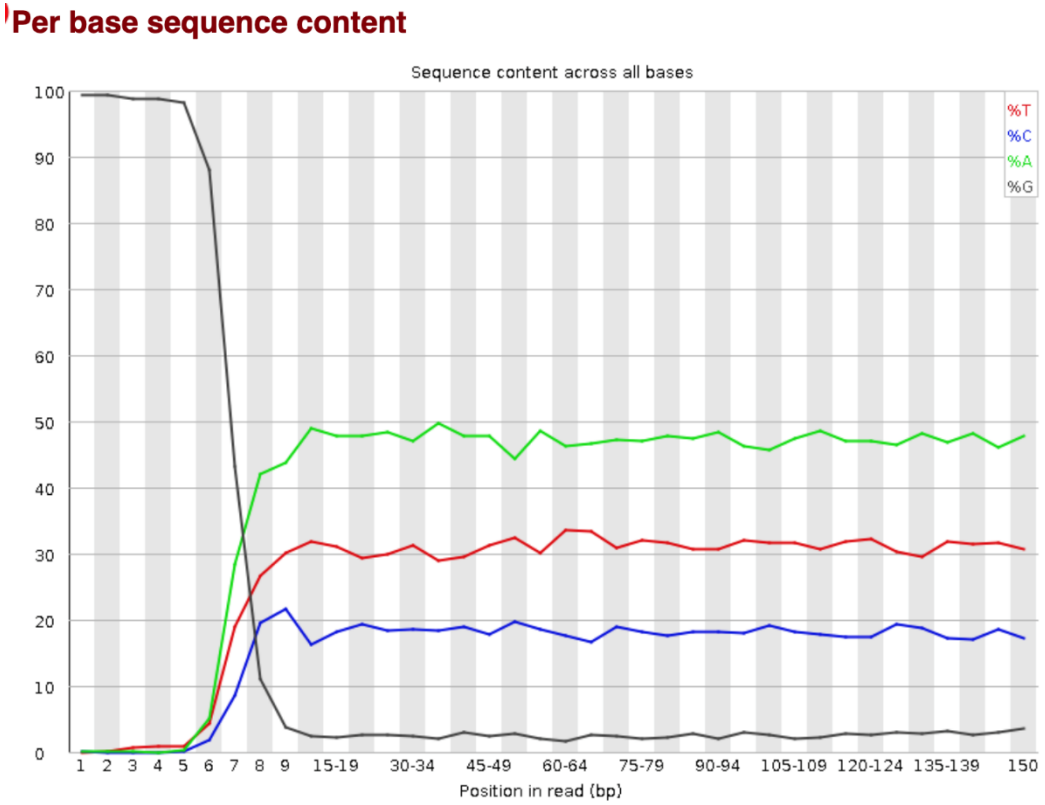
Use appropriate trimming (`trim_galore --rrbs`)

Accel Swift kit

Read 1



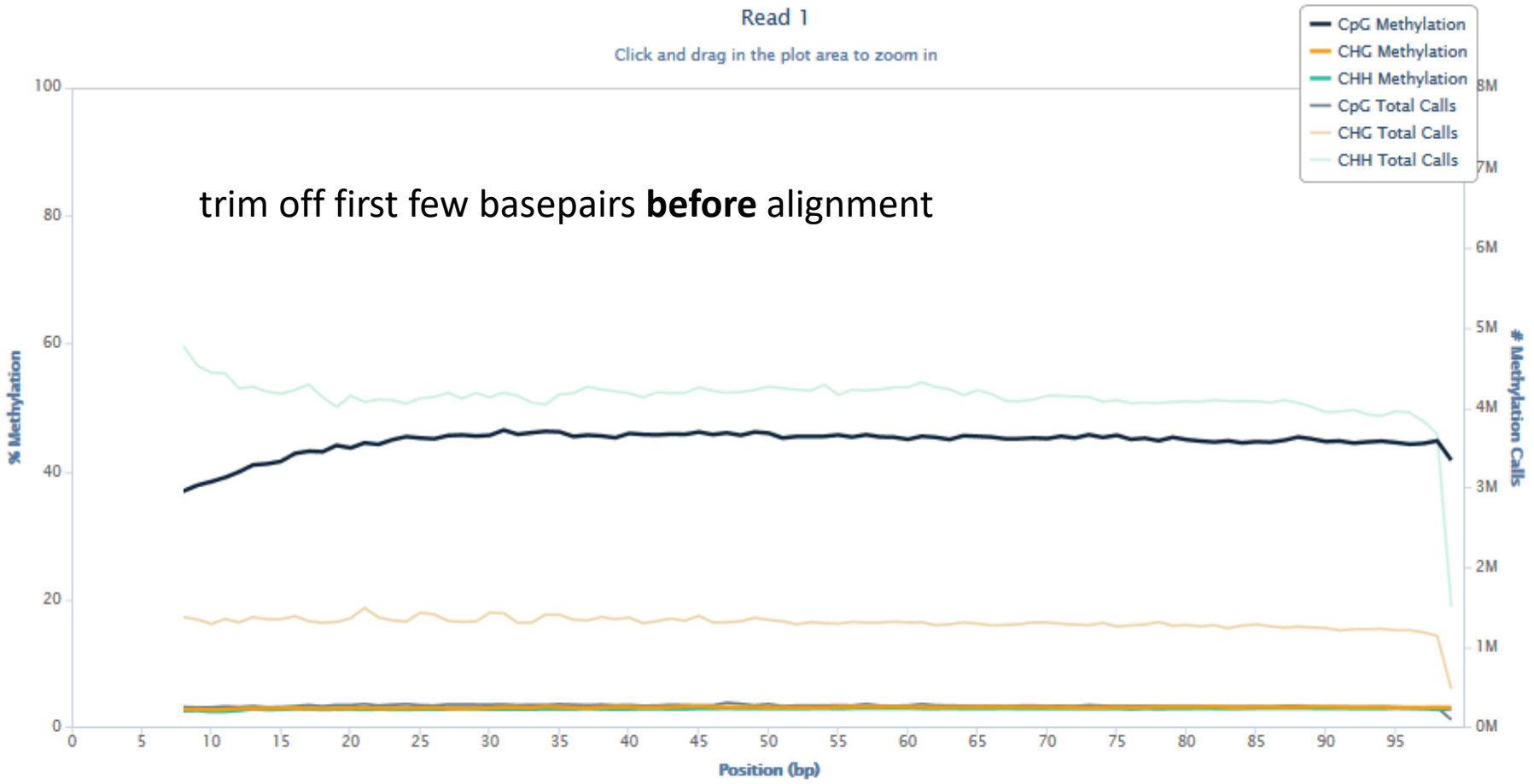
Read 2



Use appropriate trimming (`trim_galore --clip_r1 10 --clip_r2 15`)

PBAT-Seq

M-Bias Plot



trim off first few basepairs **before** alignment

Bismark Bisulfite Mapper

A tool to map bisulfite converted sequence reads and determine cytosine methylation states

View on GitHub

Bismark Bisulfite Mapper

User Guide - v0.23.0

30 September, 2020

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.

Technique	5' Trimming	3' Trimming	Mapping	Deduplication	Extraction
BS-Seq	☐	☐	☐	✅	`--ignore_r 2 2`
RRBS	`--rrbs` (R2 only)	`--rrbs` (R1 only)	☐	❌	☐
RRBS (NuGEN Ovation)	special processing	special processing	☐	❌	`--ignore_r 2 2`
PBAT	6N / 9N	(6N / 9N)	`--pbat`	✅	☐
single-cell (scBS-Seq)	6N	(6N)	`--non_directional`; single-end mode	✅	☐
TruSeq (EpiGnome)	8 bp	(8 bp)	☐	✅	☐
Accel-NGS (Swift)	R1: 10, R2:15bp	(10 bp)	☐	✅	☐
Zymo Pico-Methyl	10 bp	(10 bp)	`--non_directional`	✅	☐

http://felixkrueger.github.io/Bismark/Docs/

QCFAIL.com

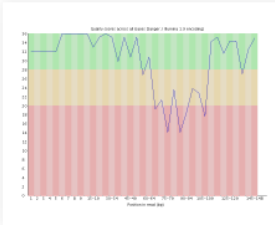
https://sequencing.qcfail.com/

Genomic sequence not in the genome assembly creates mapping artefacts

Probably the single biggest problem with the mapping of reads to a reference sequence is dealing with reads which come from parts of the genome which aren't in the assembly. These reads can cause significant amounts of noise in anlayses performed on genomic data.

March 21, 2016 | Simon Andrews | All Technologies, All Applications

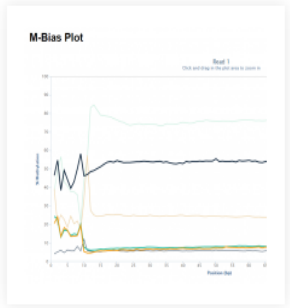
Illumina 2 colour chemistry can overcall high confidence G bases



With the introduction of the NextSeq system Illumina changed the way their image data was acquired so that instead of capturing 4 images per cycle they needed only 2. This speeds up image acquisition significantly but also introduces a problem where high quality calls for G bases can be made where there is actually no signal on the flowcell.

May 4, 2016 | Simon Andrews | NextSeq, All Applications, Cutadapt, FastQC

Mispriming in PBAT libraries causes methylation bias and poor mapping efficiencies



Random priming in PBAT libraries introduces drastic biases in the base composition and methylation levels especially at the 5′ end of all reads. As a result, affected bases should be removed from the libraries before the alignment step.

March 11, 2016 | Felix Krueger | Illumina, Methylation, PBAT, BamQC, Bismark, FastQC, Trim Galore!

Library end-repair reaction introduces methylation biases in paired-end (PE) Bisulfite-Seq applications

Library construction of standard directional BS-Seq samples often consist of several steps including sonication, end-repair, A-tailing and adapter ligation. Since the end-repair step typically uses unmethylated cytosines for the fill-in reaction the filled-in bases will generally appear unmethylated after bisulfite conversion irrespective of their true genomic methylation state.

February 12, 2016 | Felix Krueger | Illumina, BS-Seq, Methylation, Bismark, Data Processing

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

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

Epigenesys

`nf_bisulfite_WGBS`

FastQC

Trim Galore

FastQ Screen --bisulfite

FastQC

Bismark

Bismark dedup

Bismark methXtract

Bismark reports

Bismark summary

MultiQC

Bismark workflow using a workflow manager

nextflow

```
[ce/4a2468] process > FASTQC (lane7561_TTGGTATG_i5F_del_GFP_35_L001_R3)       [100%] 96 of 96, cached: 96 ✔
[2c/b83638] process > TRIM_GALORE (lane7561_TTGGTATG_i5F_del_GFP_35_L001_R1)  [ 94%] 90 of 96, cached: 90
[be/b14097] process > FASTQ_SCREEN (lane7561_TCCTCAAT_i5F_del_GFP_29_L001_R1) [  0%] 0 of 90
[94/8a2957] process > FASTQC2 (lane7561_TCCAGTCG_i5F_del_GFP_41_L001_R1)      [ 46%] 41 of 90, cached: 41
[01/56550d] process > BISMARK (lane7561_GCCAATGT_i5F_del_GFP_6_L001_R1)       [  1%] 1 of 90, cached: 1
[-        ] process > BISMARK_DEDUPLICATION                                   [  0%] 0 of 1
[-        ] process > BISMARK_METHYLATION_EXTRACTOR                           -
[-        ] process > BISMARK2REPORT                                          -
[-        ] process > BISMARK2SUMMARY                                         -
[-        ] process > MULTIQC                                                 -
```

https://github.com/s-andrews/nextflow_pipelines

Useful links

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

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

Sierra: A web-based LIMS system for small sequencing facilities

SRA Downloader

reStrainingOrder

SeqMonk: Genome browser, quantitation and data analysis

Trim Galore! Quality and adapter trimming for (RRBS) sequencing libraries

FastQ Screen: organism and contamination detection

Bismark: Bisulfite-sequencing alignments and methylation calls

Cluster Flow

HiCUP Hi-C mapping

ASAP: Allele-specific alignments

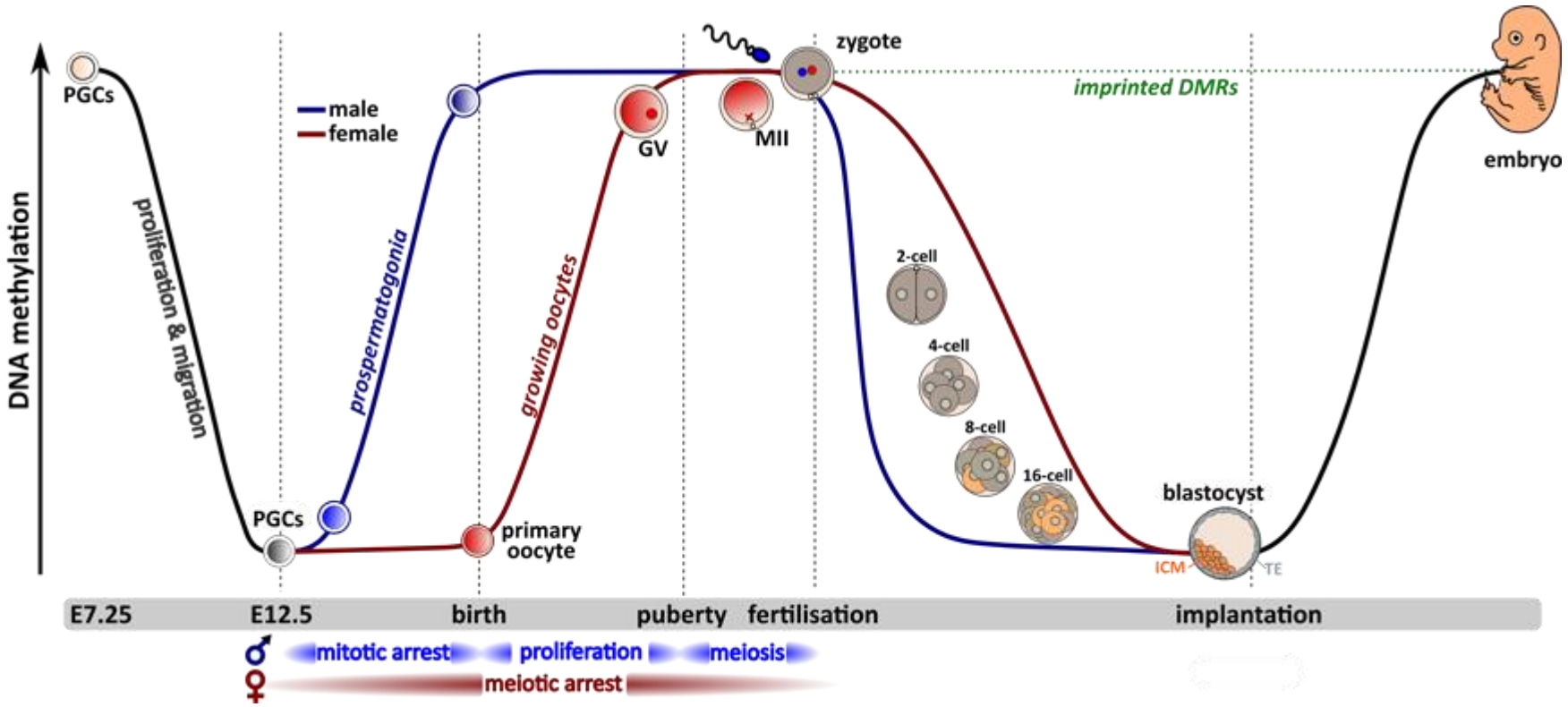
CELL LINE SLEUTH

FastQC: quality control for high throughput sequencing

SNPsplit

https://www.bioinformatics.babraham.ac.uk

DNA methylation is reset during reprogramming



Fragment size distribution in RRBS