Bisulfite-Sequencing Theory and Quality Control

felix.krueger@babraham.ac.uk
v2021-04
Bisulfite-Seq theory and Quality Control

a.m.

Mapping and QC practical
Visualising and Exploring talk

Lunch

Visualising and Exploring practical

p.m.

coffee

Differential methylation talk & practical
Epigenetics

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

From The Cell Biology of Stem Cells (2010)
Types of DNA methylation

<table>
<thead>
<tr>
<th></th>
<th>Mammals</th>
<th>Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG context</td>
<td>present</td>
<td>present</td>
</tr>
<tr>
<td>non-CG context</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DNA methylation is stably maintained
Silencing of gene expression
Tissue differentiation and embryonic development

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

Regulation by DNA methylation

Silencing of gene expression
Genomic stability
Repeat activity
Genomic instability
Imprinted Genes: Mono-allelic expression with parent-of-origin specificity. Have key roles in energy metabolism, placenta functions.
DNA Methylation

Cytosine

DNA methyl-transferases

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

5-methyl Cytosine
Other cytosine modifications
Measuring DNA methylation by Bisulfite-sequencing
Bisulfite Informatics

\[
\begin{align*}
  \text{me} & : \text{CCAGTCGCTATAGCGCGATATCGTA} \\
  \text{Convert} & : \text{TTAGTTGCTATAGTGCGATATTGTA} \\
  \text{Map} & : \text{TTAGTTGCTATAGTGCGATATTGTA} \\
\end{align*}
\]
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

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

align to bisulfite converted genomes

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

**methylation call**

h unmethylated C in CHH context
H methylated C in CHH context
x unmethylated C in CHG context
X methylated C in CHG context
z unmethylated C in CpG context
Z methylated C in CpG context
Common sequencing protocols

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

2) PBAT libraries

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

OT
>>TCGGTATGTTTAAACGTT>>
<<GGTGCTACAAATTTGCAA<< OB

CTOT
<<AGCCATACAAATTTGCAA<<
>>CCAGCATGTTAAACGCT>> CTOB

OT
>>TCGGTATGTTTAAACGTT>>
<<AGCCATACAAATTTGCAA<<

OT
>>TCGGTATGTTTAAACGTT>>
<<GGTGCTACAAATTTGCAA<< OB
Validation

Cytosines called unmethylated (%)
Bisulfite conversion rate (%)
Mapping efficiency (%)

- CpG
- CHG
- CHH
- Mapping efficiency
BS-Seq Analysis Workflow

- QC
- Trimming
- Mapping
- Analysis
- Mapped QC
- Methylation extraction
Raw Sequence Data (FastQ file)

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>s_4_1_sequence.txt</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Illumina 1.5</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>35290120</td>
</tr>
<tr>
<td>Sequence length</td>
<td>40</td>
</tr>
<tr>
<td>%GC</td>
<td>46</td>
</tr>
</tbody>
</table>
QC Raw data: Sequence Quality

Error rate

0.1%

1%

10%
QC: Base Composition

WGSBS

Sequence content across all bases

RRBS

Sequence content across all bases
QC: Duplication rate

Percent of seqs remaining if deduplicated 29.55%

% Deduplicated seqences
% Total sequences
# QC: Overrepresented sequences

## Overrepresented sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAAGACGGTTCAAGAATGGACGGACGTCTGATA</td>
<td>6254891</td>
<td>23.5273909691508</td>
<td>Illumina Paired End PCR Primer 2 (100% over 40bp)</td>
</tr>
<tr>
<td>GATCGGAAAGCGGCTTACAGAATGGACGGACGTCTC</td>
<td>1956005</td>
<td>7.35739350331777</td>
<td>Illumina Paired End PCR Primer 2 (100% over 40bp)</td>
</tr>
<tr>
<td>GAAGACGGTTCAAGAATGGACGGACGTCTGATA</td>
<td>774763</td>
<td>2.9142237687587667</td>
<td>Illumina Paired End PCR Primer 2 (96% over 31bp)</td>
</tr>
<tr>
<td>GAAGACGGTTCAAGAATGGACGGACGTCTGATA</td>
<td>140148</td>
<td>0.5271581538405985</td>
<td>Illumina Paired End Adapter 2 (100% over 27bp)</td>
</tr>
<tr>
<td>AAGAGCCGTTCAAGAATGGACGGACGTCTGATA</td>
<td>105720</td>
<td>0.387653317352233</td>
<td>Illumina Paired End PCR Primer 2 (96% over 30bp)</td>
</tr>
<tr>
<td>NAAAGACGGTTCAAGAATGGACGGACGTCTGATA</td>
<td>98639</td>
<td>0.3710245821323372</td>
<td>Illumina Paired End PCR Primer 2 (97% over 40bp)</td>
</tr>
<tr>
<td>AAGAGCCGTTCAAGAATGGACGGACGTCTGATA</td>
<td>82413</td>
<td>0.3099147281777295</td>
<td>Illumina Paired End PCR Primer 2 (100% over 40bp)</td>
</tr>
<tr>
<td>GATCGGAAAGCGGCTTACAGAATGGACGGACGTCTG</td>
<td>53872</td>
<td>0.2026362421418372</td>
<td>Illumina Paired End Adapter 2 (97% over 36bp)</td>
</tr>
<tr>
<td>NAAAGACGGTTCAAGAATGGACGGACGTCTGATA</td>
<td>36541</td>
<td>0.13744674275471</td>
<td>Illumina Paired End PCR Primer 2 (100% over 36bp)</td>
</tr>
<tr>
<td>ATCGGAAAGCGGCTTACAGAATGGACGGACGTCTC</td>
<td>35781</td>
<td>0.13458004908076072</td>
<td>Illumina Paired End PCR Primer 2 (100% over 40bp)</td>
</tr>
<tr>
<td>CGGAGACGGTTCAAGAATGGACGGACGTCTGAT</td>
<td>33905</td>
<td>0.1275315895051329</td>
<td>Illumina Paired End PCR Primer 2 (100% over 40bp)</td>
</tr>
<tr>
<td>NACGGAAGCCGTTCAAGAATGGACGGACGTCTGAT</td>
<td>30564</td>
<td>0.1149646217854272</td>
<td>Illumina Paired End PCR Primer 2 (97% over 40bp)</td>
</tr>
<tr>
<td>GAAGACGGTTCAAGAATGGACGGACGTCTGATA</td>
<td>28274</td>
<td>0.1063509264123442</td>
<td>Illumina Paired End PCR Primer 2 (97% over 40bp)</td>
</tr>
<tr>
<td>CAAACAGCTTCTAAAAACAAAAAGACAAAAACACTAA</td>
<td>27952</td>
<td>0.10513974310123076</td>
<td>No Hit</td>
</tr>
</tbody>
</table>
Common problems in BS-Seq

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

before trimming

after trimming

Phred score
Removing adapter contamination

before trimming

after trimming

[Graphs showing sequence content across all bases before and after trimming]
Adapter trimming
(Illumina adapter: AGATCGGAAGAGC)

B: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT
A: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT

partial match
B: AGATC
A: AGATC

full match
B: AGATCGGAAGAGC
A: AGATCGGAAGAGC

B: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT
A: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT

B: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT
A: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT

B: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT
A: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAT

B: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAGA
A: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAGA

B: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAGA
A: AGATCTTTTATTCGGTAGGATAGCCTGAGTTATTTTATTTTGGAGGAGA
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)

<table>
<thead>
<tr>
<th>chromosome</th>
<th>position</th>
<th>sequence</th>
<th>quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>HISEQ2000-06:366:C3G4NACXX:3:1101:1316:2067_1:N:0:</td>
<td>99 16 71322125 255 100M</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>HISEQ2000-06:366:C3G4NACXX:3:1101:1316:2067_1:N:0:</td>
<td>147 16 7132232 255 100M</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Read 1**

**Read 2**

**methylination call**
Sequence duplication

Complex/diverse library:

Duplicated library:

Deduplication

percent methylation

55 17 100 100 71 100

33 50 100 100 50 100
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.........Z.......x..h...h..h........

Read 2

....x......hh..h................z.....hx...h...h..h.......

redundant methylation calls

Read 1

....x......hh..h................z.....hx...h...h..h.......

Read 2

hh.x...h...h...h...h........

CpG methylation output

<table>
<thead>
<tr>
<th>read ID</th>
<th>meth state</th>
<th>chr</th>
<th>pos</th>
<th>context</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS9 1118:8:2023:313:4022:38681#13/1</td>
<td>+</td>
<td>1</td>
<td>3029229</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:1205:13025:9541#19/1</td>
<td>+</td>
<td>1</td>
<td>3079409</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:1301:11752:81850#13/1</td>
<td>-</td>
<td>1</td>
<td>3104640</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2201:12469:84166#19/1</td>
<td>+</td>
<td>1</td>
<td>3104642</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2206:15661:25006#13/1</td>
<td>+</td>
<td>1</td>
<td>3104502</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2308:1530:10033#13/1</td>
<td>-</td>
<td>1</td>
<td>3124392</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2308:1530:10033#13/1</td>
<td>+</td>
<td>1</td>
<td>3124416</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2308:1530:10033#13/1</td>
<td>+</td>
<td>1</td>
<td>3124416</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2105:9522:91783#13/1</td>
<td>+</td>
<td>1</td>
<td>3124392</td>
<td>Z</td>
</tr>
<tr>
<td>HS9 1118:8:2105:9522:91783#13/1</td>
<td>+</td>
<td>1</td>
<td>3124416</td>
<td>Z</td>
</tr>
</tbody>
</table>
Methylation extraction I

CpG methylation output

bedGraph/coverage output
### Methylation extraction II

**coverage output**

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>strand</th>
<th>meth</th>
<th>unmeth</th>
<th>di-nuc</th>
<th>tri-nuc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10525</td>
<td>10525</td>
<td>6.66666666666667</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10542</td>
<td>10542</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10565</td>
<td>10565</td>
<td>6.66666666666667</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10571</td>
<td>10571</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10677</td>
<td>10677</td>
<td>6.66666666666667</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10679</td>
<td>10679</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10589</td>
<td>10589</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10620</td>
<td>10620</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10617</td>
<td>10617</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**coverage2cytosine**

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>strand</th>
<th>meth</th>
<th>unmeth</th>
<th>di-nuc</th>
<th>tri-nuc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10526</td>
<td>10526</td>
<td>0</td>
<td>0</td>
<td>CG</td>
<td>CG</td>
</tr>
<tr>
<td>1</td>
<td>10542</td>
<td>10542</td>
<td>3</td>
<td>0</td>
<td>CG</td>
<td>CGA</td>
</tr>
<tr>
<td>1</td>
<td>10543</td>
<td>10543</td>
<td>0</td>
<td>0</td>
<td>CG</td>
<td>CG</td>
</tr>
<tr>
<td>1</td>
<td>10563</td>
<td>10563</td>
<td>2</td>
<td>1</td>
<td>CG</td>
<td>CG</td>
</tr>
<tr>
<td>1</td>
<td>10564</td>
<td>10564</td>
<td>0</td>
<td>0</td>
<td>CG</td>
<td>CG</td>
</tr>
<tr>
<td>1</td>
<td>10571</td>
<td>10571</td>
<td>3</td>
<td>0</td>
<td>CG</td>
<td>CG</td>
</tr>
<tr>
<td>1</td>
<td>10572</td>
<td>10572</td>
<td>0</td>
<td>0</td>
<td>CG</td>
<td>CG</td>
</tr>
<tr>
<td>1</td>
<td>10577</td>
<td>10577</td>
<td>2</td>
<td>1</td>
<td>CG</td>
<td>CG</td>
</tr>
</tbody>
</table>

**optional: merge into CpG dinucleotide entities**

### Genome wide CpG report

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>strand</th>
<th>meth</th>
<th>unmeth</th>
<th>di-nuc</th>
<th>tri-nuc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10525</td>
<td>10525</td>
<td>6.66666666666667</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10542</td>
<td>10542</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10565</td>
<td>10565</td>
<td>6.66666666666667</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10571</td>
<td>10571</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10677</td>
<td>10677</td>
<td>6.66666666666667</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10679</td>
<td>10679</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10589</td>
<td>10589</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10620</td>
<td>10620</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10617</td>
<td>10617</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### columns
- chr: chromosome
- pos: position
- strand: strand orientation (+ for forward, - for reverse)
- meth: methylation coverage
- unmeth: unmethylation coverage
- di-nuc: di-nucleotide
- tri-nuc: tri-nucleotide
Part III: Mapped QC - Methylation bias

good opportunity to look at conversion efficiency
Artificial methylation calls in paired-end libraries

end repair + A-tailing

5' - GGGNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCA
3' - ACCCNNNNNNNNNNNNNNNNNNNNNNNNNNNGGG

- 3'
- 5'
Specialist applications

(e)RRBS

WGBS

PBAT

NOMe-seq
target
enrichment

amplicon

NMT-seq

non-directional

single-cell

+ different library kit protocols
Reduced representation BS-Seq (RRBS)

Sequence composition bias

High duplication rate
Artificial methylation calls in RRBS libraries

Use appropriate trimming (trim_galore --rrbs)
Use appropriate trimming \((\text{trim\_galore} \ --\text{clip\_r1} \ 10 \ --\text{clip\_r2} \ 15)\)
Post-bisulfite adapter tagging (PBAT)

WGBS

A. Sample DNA
   - 1. Fragmentation
   - 2. Adaptor ligation
   - 3. Bisulfite treatment
   - 4. Global amplification

PBAT

B. Sample DNA
   - 1. Bisulfite treatment (fragmentation)
   - 2. Adaptor tagging

C. Sample DNA
   - 1. Bisulfite treatment
   - 2. 1st. random priming
   - 3. Magnetic bead capture
   - 4. 2nd. random priming
   - 5. Elution

No fragmentation steps after adaptor tagging

→ suitable for low input material
PBAT-Seq

M-Bias Plot

trim off first few basepairs *before* alignment
Bismark User Guide


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 troubleshooting section at QGalls.com.

1) Quick Reference

Bismark needs a working version of Perl and it is run from the command line. Furthermore, Bowtie or Breathe 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 --multithread for more details below.

First you need to download a reference genome and place it in a genome folder. Genomas can be obtained e.g., from the Ensembl or UCSC websites. For the example below you would need to download the Homo sapiens genome. Bismark supports reference genome sequence files in Fastx format, allowing file extensions are either either .fa or .fai,.fa.gz or .fai.gz. Both single- and multiple-FASTX files are supported.

The following examples will use the file: /test_dataset.fas.gz, 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 B-Seq library). An example report for use with Bowtie 1 and Bowtie can be found in Appendix E.

(I) Running bismark_genome_preparation

USAGE:
bismark_genome_preparation [options] [species_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/H40H37/

(II) Running bismark

USAGE:
# (VIII) Notes about different library types and commercial kits

Here is a table summarising general recommendations for different library types and/or different commercially available kits. Some more specific notes can be found below.

<table>
<thead>
<tr>
<th>Technique</th>
<th>5' Trimming</th>
<th>3' Trimming</th>
<th>Mapping</th>
<th>Deduplication</th>
<th>Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS-Seq</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RRBS</td>
<td>--res (R2 only)</td>
<td>--res (R1 only)</td>
<td></td>
<td></td>
<td>(✓)</td>
</tr>
<tr>
<td>RRBS (NuGEN Ovation)</td>
<td>special processing</td>
<td>special processing</td>
<td></td>
<td></td>
<td>(✓)</td>
</tr>
<tr>
<td>PBAT</td>
<td>6N / 9N</td>
<td>(6N / 9N)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>single-cell (scBS-Seq)</td>
<td>6N</td>
<td>(6N)</td>
<td>--non_directional; single-end mode</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>TruSeq (Epigenome)</td>
<td>8 bp</td>
<td>(8 bp)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accel-NGS (Swift)</td>
<td>10 bp</td>
<td>(10 bp)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Zymo Pico-Methyl</td>
<td>10 bp</td>
<td>(10 bp)</td>
<td></td>
<td>--non_directional</td>
<td>✓</td>
</tr>
</tbody>
</table>

- - Default settings (nothing in particular is required, just use Trim Galore or Bismark default parameters)
- ✓ - Yes, please!
- ✗ - No, absolutely not!

5' Trimming can be accomplished with Trim Galore using:

```
--clip_r1 <NUMBER> (Read 1) or
--clip_r2 <NUMBER> (Read 2)
```

3' Trimming can be accomplished with Trim Galore using:

```
--three_primers_clip_r1 <NUMBER> (Read 1) or
--three_primers_clip_r2 <NUMBER> (Read 2)
```

**SPECIFIC LIBRARY/KIT NOTES**

RRBS

[GitHub Link](https://github.com/FelixKrueger/Bismark/tree/master/Docs#viii-notes-about-different-library-types-and-commercial-kits)
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 analyses 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, BaitQC, Bismark, FastQC, TrimGalore!

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

**protocol:** Quality Control, trimming and alignment of Bisulfite-Seq data
Bismark workflow using a workflow manager
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/
- Cluster Flow  www.bioinformatics.babraham.ac.uk/projects/clusterflow/

https://sequencing.qcfail.com/
Sierra: A web-based LIMS system for small sequencing facilities

SeqMonk: Genome browser, quantitation and data analysis

FastQ Screen: organism and contamination detection

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

Bismark: Bisulfite-sequencing alignments and methylation calls

HiCUP: Hi-C mapping

FastQC: quality control for high throughput sequencing

ASAP: Allele-specific alignments

https://www.bioinformatics.babraham.ac.uk
DNA methylation is reset during reprogramming

DNA methylation

PGCs

progermatogenic proliferation & migration

progermatogenic growing oocyte

GV

MII

zygote

imprinted DMRs

embryo

E7.25

E12.5

birth

puberty

fertilisation

implantation

mitotic arrest

proliferation

meiosis

mitotic arrest

male

female
Validation

- **Mapping efficiency (%)**
  - Read length (bp)
  - Bisulfite conversion rate (%)
- **Cytosines called unmethylated (%)**
- **Mapping efficiency (%)**
  - Cytosines called unmethylated (%)
  - Read length (bp)
  - Bisulfite conversion rate (%)

**Graphs:**
- Bias vs. Read length (bp)
- Bias vs. Bisulfite conversion rate (%)
- Mapping efficiency vs. Bisulfite conversion rate (%)

**Legend:**
- CpG
- CHG
- CHH
- Mapping efficiency
Fragment size distribution in RRBS

Human genome (GRCh37)

Mspl site

$5'$ - $\text{CGGG...GGCC}$

$3'$ - $\text{GGCC...GGCC}$

identical (redundant) methylation calls

Mspl site

$5'$ - $\text{GGCC...GGCC}$

$3'$ - $\text{CGGG...CGGG}$

$40-220bp$