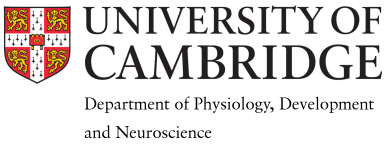
Methylation Analysis using Bisulfite Sequencing

Russell S. Hamilton[1]

[1] *Centre for Trophoblast Research, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Site, Cambridge, CB2 3DY*
[1] *Email: rsh46@cam.ac.uk ::: darogan@gmail.com*

2016-08-02
version 0.1



Lecture notes to accompany presentation

NGSchool.eu 2016

Contents

1 Introduction **4**
- 1.1 What is methylation? . . . 4
- 1.2 Identifying methylation . . . 4
- 1.3 What is hydroxymethyl-cytosine? . . . 4
- 1.4 What is formyl-cytosine and carboxy-cytosine? . . . 4

2 Analysis **5**
- 2.1 Pipeline . . . 5
- 2.2 Pre Alignment Quality Control . . . 6
- 2.3 Alignment . . . 6
- 2.4 Post Alignment Quality Control . . . 9
  - 2.4.1 QualiMap . . . 9
  - 2.4.2 Preseq . . . 10
  - 2.4.3 Picard Insert Size Metrics . . . 10
- 2.5 Deduplication . . . 10
- 2.6 Spike In Controls . . . 11
- 2.7 Methylation Calling . . . 12
  - 2.7.1 Bismark methylation extractor . . . 12
- 2.8 DMR Calling . . . 12

3 Prerequisites **12**
- 3.1 Software Required . . . 12
- 3.2 Reference Genome . . . 12
- 3.3 Data Sets Required . . . 13
  - 3.3.1 Aging Data . . . 13
  - 3.3.2 Amplicon Data . . . 13

4 Discussion / Concluding Remarks **13**

List of Figures

1 Cytosine Modifications 4
2 Bisulfite Sequencing 5
3 Oxidative Bisulfite Sequencing 6
4 Reduced Bisulfite Sequencing 7
5 FastQC 8
6 Bisulfite Sequencing Alignment 9
7 Bisulfite Alignment Metrics 10
8 Qualimap BAMQC Metrics 11
9 Preseq and Picard Metrics 11

List of Tables

1 Prerequisite Software 12
2 Data from (Hadad *et al.*, 2016) 13
3 Data from (Äijö *et al.*, 2016) 14

Listings

1 Bismark Clusterflow Pipeline 5
2 Calling the Bismark Clusterflow Pipeline 5
3 Bismark alignment: pbat samples 6
4 Example bismark Alignment Report 6
5 Bismark alignment: lux controls 7
6 Samtools coordinate sort 9
7 Qualimap bamqc 9
8 Preseq library complexity estimation 10
9 Picard insert size metrics 10
10 Bismark deduplication 10
11 Bismark methylation extraction 12
12 Bismark genome preparation 12

1 Introduction

1.1 What is methylation?

The most common form of methylation occurs at the 5' position of cytosines through the addition of a methyl group. In mammals this is most common at CpG sites, but also occurs at other cytosine positions. Some prokaryotes, such as *E. coli*, have pentameric methylation sites.

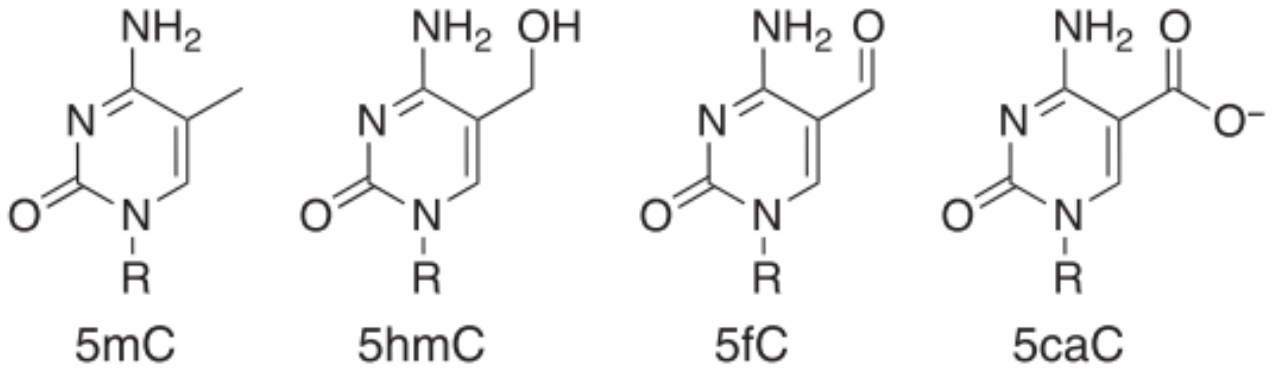


Figure 1: Cytosine Modifications. Structures of 5'-methyl-cytosine (5-mC), 5'-hydroxymethyl-cytosine (5-hmC), 5'-formyl-cytosine (fC) and 5'carboxy-cytosine(caC). [Figure from (Booth *et al.*, 2013)]

1.2 Identifying methylation

There are several techniques for assaying single base methylation levels. Cost and tissue availability are often limiting factors therefore reduced genome methods are very popular.

- Introduce bisulfite bismark (Krueger and Andrews, 2011)
- Introduce RRBS
- Introduce 450K / EPIC
- Introduce hmC calling BS/ox-bs subtraction
- Introduce TAB-Seq

1.3 What is hydroxymethyl-cytosine?

Introduce hmC and its biological significance, also highlight coverage requirement very costly 30x as subtraction required

1.4 What is formyl-cytosine and carboxy-cytosine?

Introduce fC, caC and their biological significance

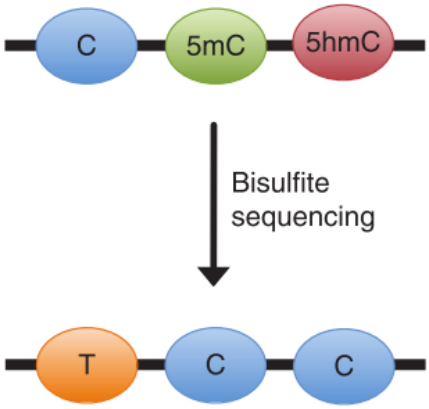


Figure 2: Bisulfite Sequencing. Cytosines are converted to thymine, 5-methyl-cytosine and 5-hydroxymethyl-cytosine are protected from conversion and are read as cytosine in sequencing. [Figure from (Booth *et al.*, 2013)]

2 Analysis

The recommendation for bisulfite sequencing and oxidative bisulfite sequencing to be able to call single base resolution 5-hmC is 30x. Therefore this is going to be an expensive experiment, requiring careful consideration of samples, replicates and tissue types.

2.1 Pipeline

To processing a large number of samples in a consistent, documented and reproducable manner it is advisable to use a pipeline system. Pipelines can be custom bash scripts, docker containers or specific pipeline tools such as clusterflow.io. Exact versions and command line options should be recorded in log files.

```
#fastqc
#trim_galore
        #bismark_align
                #bismark_deduplicate
                      #preseq_lc_extrap
                      #preseq_bound_pop
                      #qualimap_bamqc
                      #picard_insert_size_metrics
                      #featureCounts
                      #bismark_methXtract
                                #bismark_report
                                        >bismark_summary_report
>multiqc
```

Listing 1: Bismark Clusterflow Pipeline

```
$ cf --genome <PATH/TO/GENOME/INDEX> pipeline_name *.fq.gz
```

Listing 2: Calling the Bismark Clusterflow Pipeline

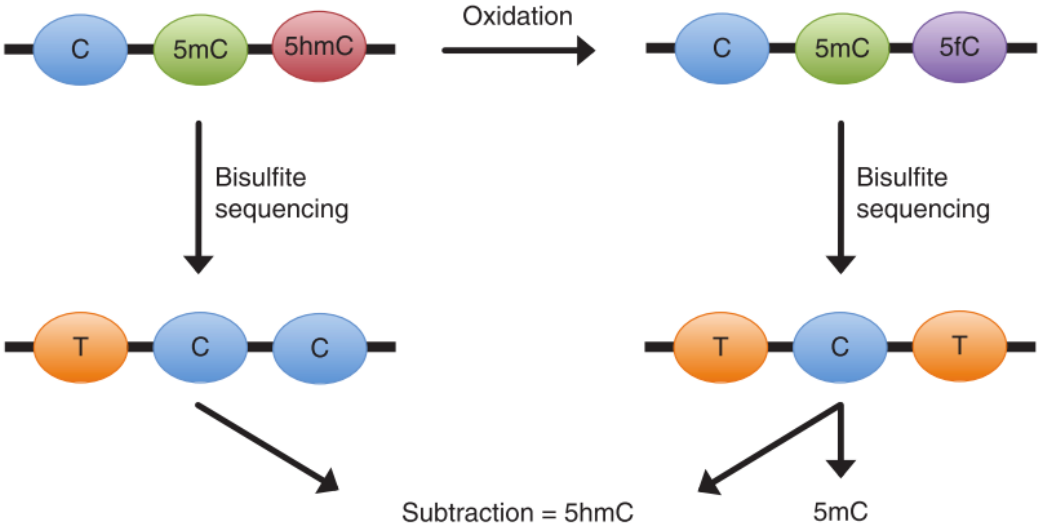


Figure 3: Oxidative bisulfite sequencing. Cytosines are converted to thymine, 5-methyl-cytosine is protected from conversion and is read as cytosine in sequencing. 5-hydroxymethyl-cytosine is converted to 5-formyl-cytosine during oxidation and then to thymine in bisulfite. A subtraction is required to read the 5-hydroxymethyl-cytosine component.[Figure from (Booth *et al.*, 2013)]

2.2 Pre Alignment Quality Control

Fastqc for the fastq files provides a comprehensive assessment of the sequencing quality and the adapter contamination.

2.3 Alignment

The alignment of bisulfite treated samples is broadly the same for whole genome, targetted and reduced-representation bisulfite sequencing (RRBS). There are specific options in bismark worth noting (RRBS, pbat, directional).

Bismark alignment of PBAT samples:

```
$ bismark /path/to/reference/GRCm38_Lambda/ --pbat \
    -1 read_1.fq.gz -2 read_2.fq.gz
```

Listing 3: Bismark alignment: pbat samples

The `--pbat` option is the only non-default option used as this is required as the samples were prepared with post-bisulfite adapter tagging. This is an attempt to reduce the number of fragments lost due to fragmentation, but adding the adapters after bisulfite sequencing (Miura *et al.*, 2012).

Example bismark Alignment Report:

```
Final Alignment report
======================
Sequence pairs analysed in total:   23456857
Number of paired-end alignments with a unique best hit:     16059532
Mapping efficiency: 68.5%

Sequence pairs with no alignments under any condition:      6471470
```

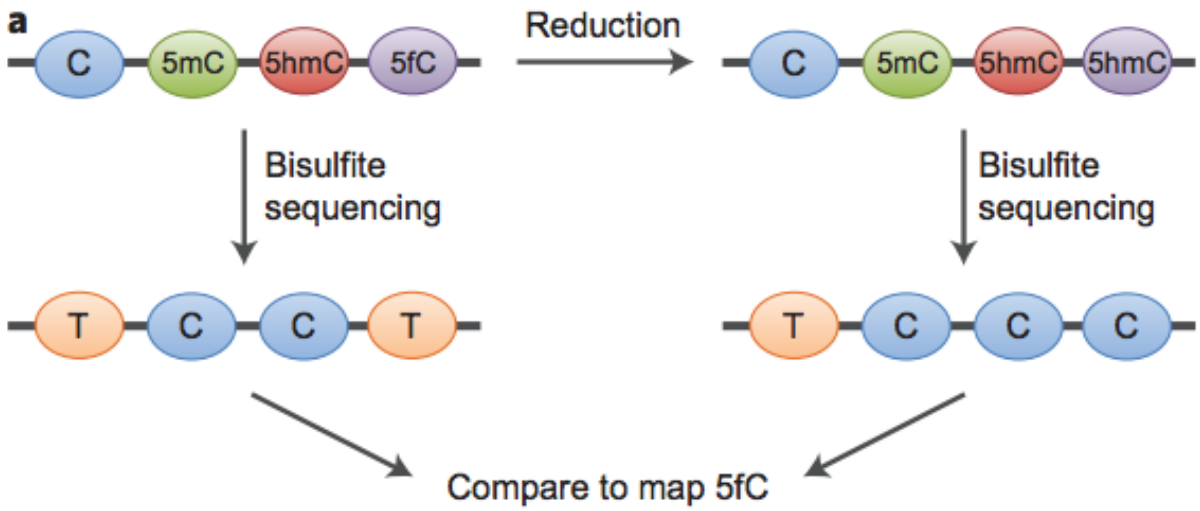


Figure 4: Reduced bisulfite sequencing. Cytosines are converted to thymine, 5-methyl-cytosine and 5-hydroxymethyl-cytosine are protected from conversion ands read as cytosine in sequencing. 5-formyl-cytosine is converted to thymine in bisulfite treatment, and converted to 5-hydroxymethyl-cytosine during reduction, read as cytosine. A subtraction is required to read the 5-formylmethyl-cytosine component. [Figure from (Booth *et al.*, 2014)]

```
Sequence pairs did not map uniquely:        925855
Sequence pairs which were discarded because genomic sequence could not
be extracted:        5

Number of sequence pairs with unique best (first) alignment came from the
bowtie output:
CT/GA/CT:   0       ((converted) top strand)
GA/CT/CT:   7983330 (complementary to (converted) top strand)
GA/CT/GA:   8076197 (complementary to (converted) bottom strand)
CT/GA/GA:   0       ((converted) bottom strand)

Final Cytosine Methylation Report
=================================
Total number of C's analysed:       896797899

Total methylated C's in CpG context:        25511027
Total methylated C's in CHG context:        1667723
Total methylated C's in CHH context:        6550281
Total methylated C's in Unknown context:    0

Total unmethylated C's in CpG context:      10166967
Total unmethylated C's in CHG context:      211368803
Total unmethylated C's in CHH context:      641533098
Total unmethylated C's in Unknown context:  22

C methylated in CpG context:        71.5%
C methylated in CHG context:        0.8%
C methylated in CHH context:        1.0%
C methylated in unknown context (CN or CHN):        0.0%
```

Listing 4: Example bismark Alignment Report

Bismark alignment of Lux control samples. As they are short, and with known methylation state, the alignments must be intollerant to mismatches.

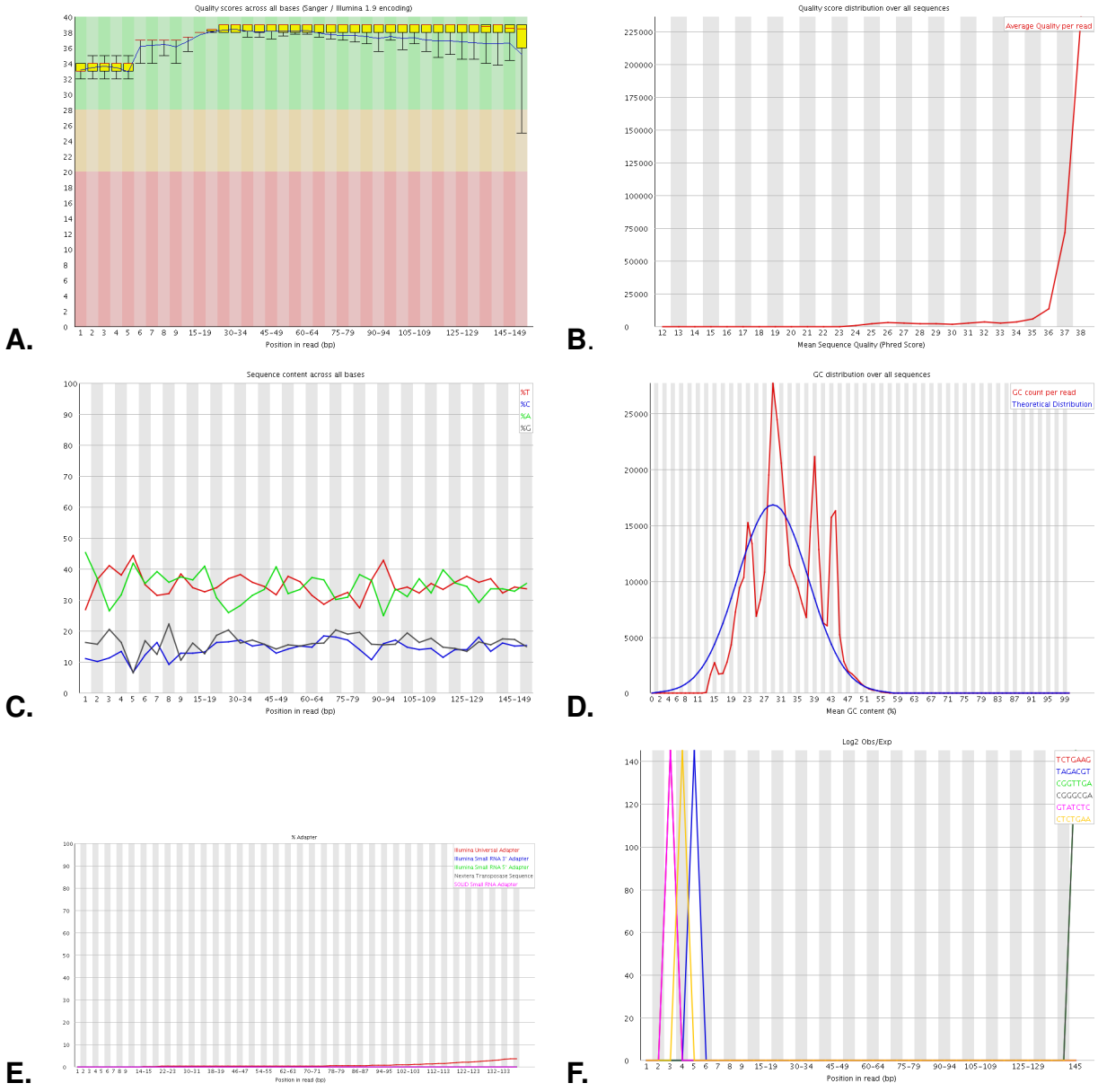


Figure 5: FastQC Metrics. A. Per base quality scores. B. Quality scores. C. Per base sequence content. D. GC content. E. Adapter content. F. Kmer content

```
$ bismark /path/to/reference/GRCm38_Lambda/ -I 0 -X 2000 -N 0 \
   --non_directional -1 read_1.fq.gz -2 read_2.fq.gz
```

Listing 5: Bismark alignment: lux controls

- `-I 0` : minimum insert size of zero - no overlapping R1 / R2
- `-X 2000` : maximum insert size of 2000nt (default is 500nt)
- `-N 0` : number of allowed mismatches
- `--non_directional` : selected for non directional library preps (not current illumina protocols)

Typically a bismark alignment of approx 70% is an to be expected. Below this there could be issues such as adapter contamination. The duplication rate and fragmentation are factors influencing the alignment rate.

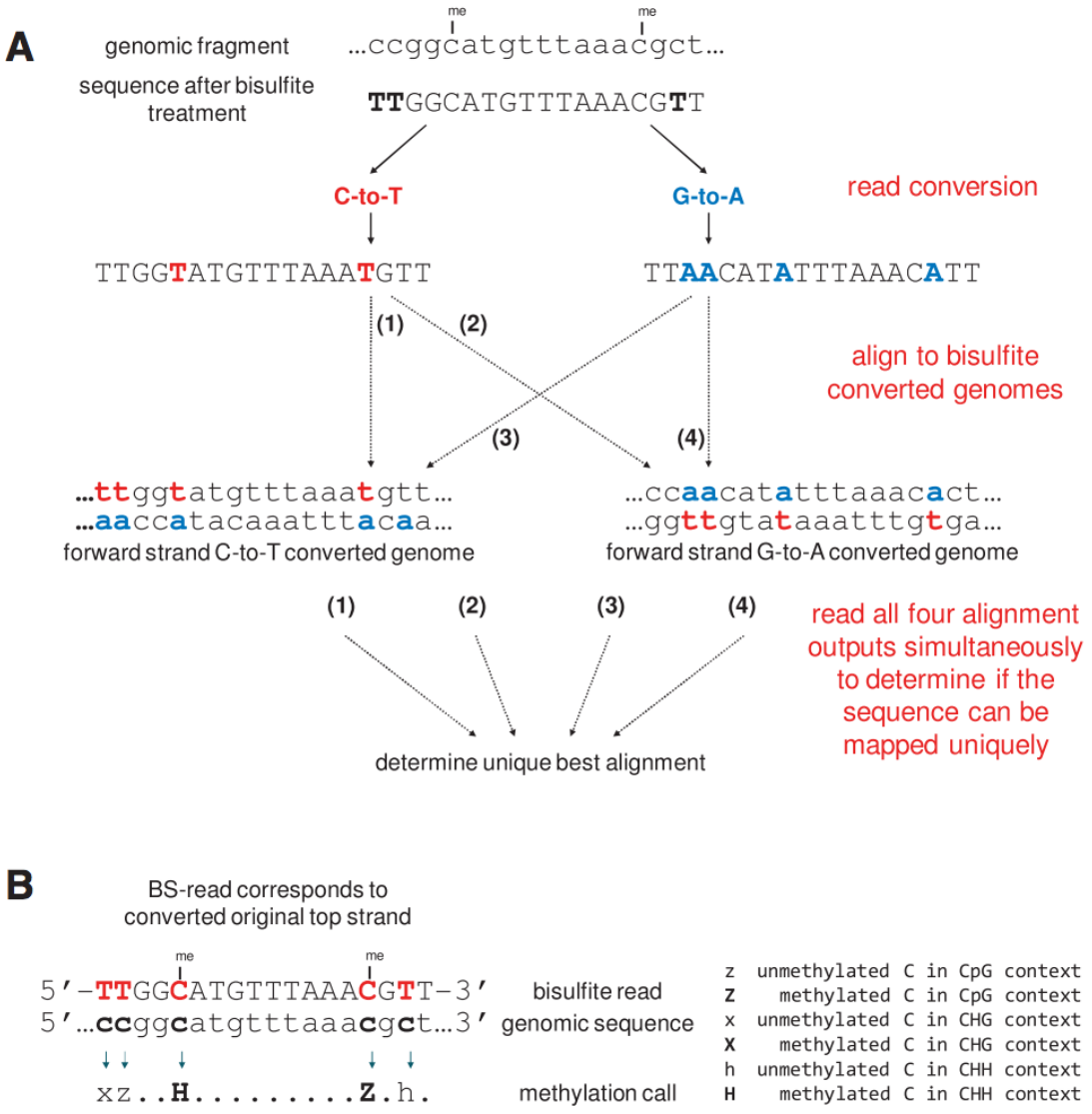


Figure 6: Aligning Bisulfite Sequencing Reads with Bismark. [Figure from (Krueger and Andrews, 2011)]

2.4 Post Alignment Quality Control

Bismark produces name sorted bam files to be compatible with the methylation extractor. To perform the post alignment QC, most analysis tools require coordinate sorted so an extra re-sort step is required.

```
$ samtools sort -o sample.coordsrt.bam sample.bam
```

Listing 6: Samtools coordinate sort

2.4.1 QualiMap

In particular, check the GC content, we are losing a base so important to check conversion accuracy.

```
$  qualimap bamqc -bam sample.bam -outfile result.pdf
```

Listing 7: Qualimap bamqc

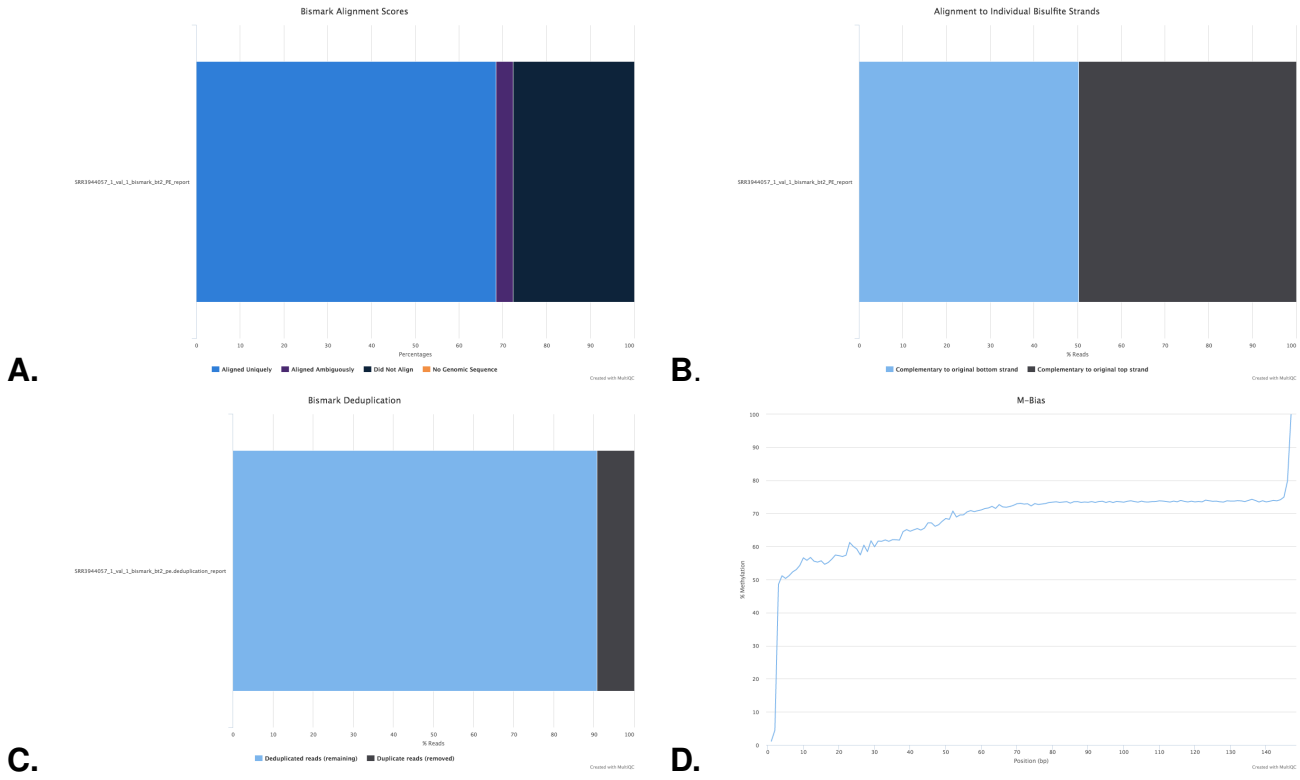


Figure 7: Bismark Alignment Metrics. A. Percentage alignment. B. Strand bias. C. Bismark Deduplication. D. Bismark M-Bias

2.4.2 Preseq

Saturation curves give an indication of how much sequencing is required in light of the bisulfite induced fragmentation.

```
$ preseq lc_extrap -e 1000000000 -P -l 999999999999 -v -Q -B sample.bam
```

Listing 8: Preseq library complexity estimation

2.4.3 Picard Insert Size Metrics

Due to bisulfite treatment causing fragmentation it is crucial to check the PE insert sizes. Picard, like Qualimap, calculates insert sizes of paired end data.

```
$ java -jar picard.jar CollectInsertSizeMetrics \
     I=sample.bam O=insert_size_metrics.txt \
     H=insert_size_histogram.pdf
```

Listing 9: Picard insert size metrics

2.5 Deduplication

Bismark includes tools for deduplication, based on identical genomic mapping. The advantage of using this tool rather than e.g. `samtools dedup` is that is is fully compatible with the bismark name sorted alignments.

```
$ deduplicate_bismark -p --bam sample_1_val_1_bismark_bt2_pe.bam
```

Listing 10: Bismark deduplication

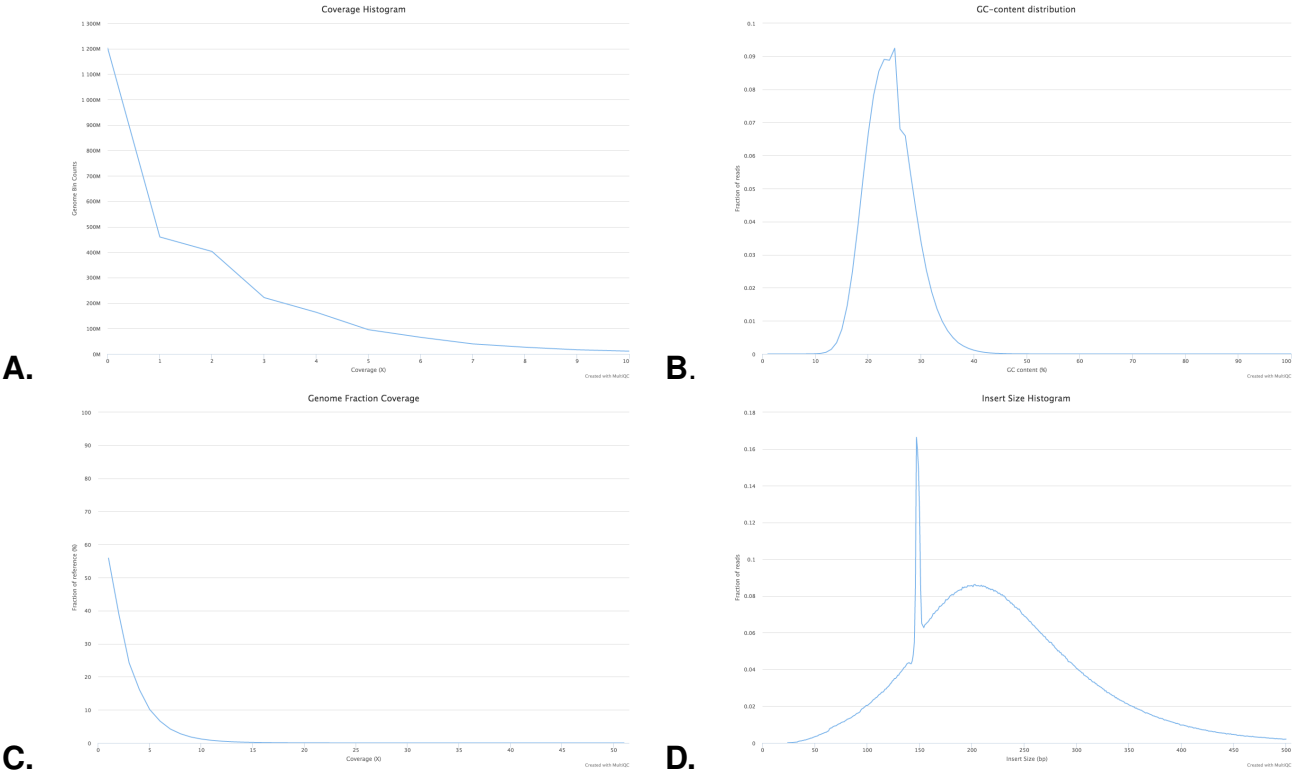


Figure 8: Qualimap bamqc metrics. A. Coverage Histogram. B. GC-Content. C. Geneome Fraction Coverage. D. Insert Size

2.6 Spike In Controls

To assess the conversion efficiency of the bisulfite, oxidation and reduction treatments it is advisable to include synthetic spike in controls with known methylation states in to the sequencing. These can either be designed by the individual research group, or companies like Cambridge Epigenetix include controls in the kits they sell. Analysing the controls, often due to their short length, can be non trivial so CEGX provide analysis software for assessing the methylation conversion rates of their controls.

In the Lux paper, (Äijö *et al.*, 2016), they designed their own controls (See table below)

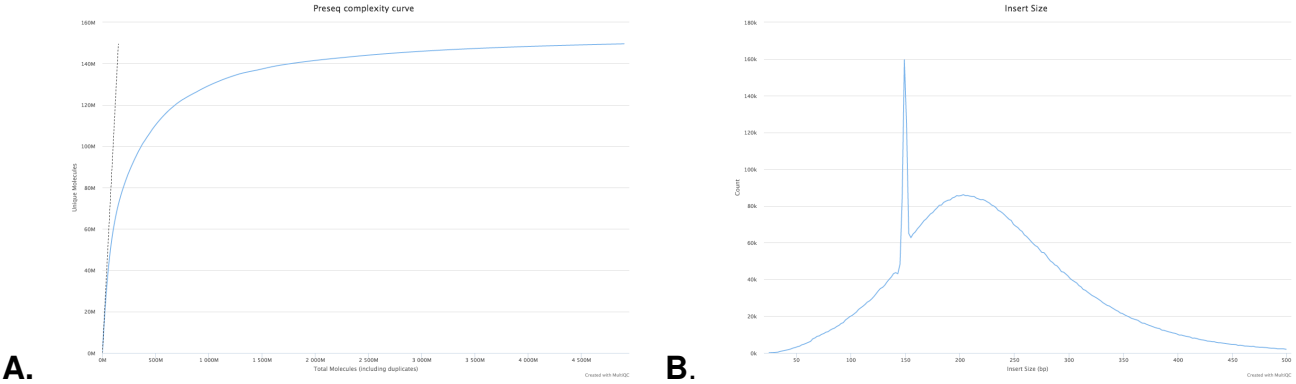


Figure 9: Preseq estimate library complexity and Picard insert size metrics. A. Percentage alignment. B. Strand bias.

Software	*Version*	*Link*
FastQC	v0.11.5	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
trim_galore	v0.4.1	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
samtools	1.3.1	http://www.htslib.org/download/
bowtie2	recent	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
bismark	0.16.1	https://github.com/FelixKrueger/Bismark/releases
Qualimap	2.2	http://qualimap.bioinfo.cipf.es/
preseq	2.0	http://smithlabresearch.org/software/preseq/
multiqc	v0.8dev	pip install git+https://github.com/ewels/MultiQC.git
R-Studio	recent	https://www.rstudio.com/products/RStudio/
R	recent	https://www.r-project.org/
Lux	recent	https://github.com/tare/Lux/
methyl-kit	v0.99.2	https://github.com/al2na/methylKit
picard	recent	https://broadinstitute.github.io/picard/

Table 1: Prerequisite Software

2.7 Methylation Calling

2.7.1 Bismark methylation extractor

Bismark comes packaged with its own methylation extractor, with the ability to call methylation in all cytosine environments, not just CpG. Also a variety of reports, and levels of verbosity can be specified.

```
$ bismark_methylation_extractor --multi 4 --ignore_r2 1 \
   --ignore_3prime_r2 2 --bedGraph --counts --gzip -p \
   --no_overlap --report sample.deduplicated.bam
```

Listing 11: Bismark methylation extraction

2.8 DMR Calling

- Introduce principals od DMR calling
- Focus on Methyl-Kit (Akalin *et al.*, 2012)

3 Prerequisites

3.1 Software Required

3.2 Reference Genome

The reference genome needs to be prepared for bisulfite sequencing. In the case of bismark the top and bottom strands need to be converted C to T and G to A and then indexed with bowtie2. In the example data sets, both are from mouse and in the Lux data Lamba derived spike in controls are used. Therefore for this practical a custom genome of GRCm38 with Lambda is used.

```
$ bismark_genome_prepare --bowtie2 /path/to/GRCm38_Lambda
```

Listing 12: Bismark genome preparation

	Replicate	Treatment	Sample Name
Female Aged 3 months	1	mkBS	SRR3944074:M4_1813
	2	mkBS	SRR3944075:M5_1815
	3	mkBS	SRR3944076:M6_1817
	1	oxBS	SRR3944064:M4_1813
	2	oxBS	SRR3944065:M5_1815
	3	oxBS	SRR3944066:M6_1817
Female Young 24 months	1	mkBS	SRR3944057:M1_1801
	2	mkBS	SRR3944058:M2_1805
	3	mkBS	SRR3944069:M3_1808
	1	oxBS	SRR3944061:M1_1801
	2	oxBS	SRR3944062:M2_1805
	3	oxBS	SRR3944063:M3_1808
Male Aged 24 months	1	mkBS	SRR3944080:M3_1813
	2	mkBS	SRR3944059:M4_1815
	3	mkBS	SRR3944060:M6_1817
	1	oxBS	SRR3944071:M3_1813
	2	oxBS	SRR3944072:M4_1815
	3	oxBS	SRR3944073:M6_1817
Male Young 3 months	1	mkBS	SRR3944077:M1_1801
	2	mkBS	SRR3944078:M2_1805
	3	mkBS	SRR3944079:M5_1808
	1	oxBS	SRR3944067:M1_1801
	2	oxBS	SRR3944068:M2_1805
	3	oxBS	SRR3944070:M5_1808

Table 2: Data from (Hadad *et al.*, 2016)

3.3 Data Sets Required

3.3.1 Aging Data

Paper (Hadad *et al.*, 2016) DOI:http://dx.doi.org/10.1186/s13072-016-0080-6

3.3.2 Amplicon Data

Amplicons from the Lux paper (Äijö *et al.*, 2016) DOI:http://dx.doi.org/10.1186/s13059-016-0911-6
Other sequences are often spike into the prep to increase library diversity due to the loss of cytosines - in this case lambda, but PhiX also used. Spike in controls with known methylation status are also used.

Note: Align to GRCm38 and Lambda simultaneously

4 Discussion / Concluding Remarks

Due to fragmentation and the statistical power required for bs/oxBS subtraction samples need to be sequenced at very high depth (30x). Also due to the dual treatment, double the

Sample	Replicate	Treatment	Sample Name
Tet2 KO	1	mkBS	SRR2009038:Tet2_KO_mkbs
	2	mkBS	SRR2009039:Tet2_KO_mkbs
	3	mkBS	SRR2009040:Tet2_KO_mkbs
	1	oxBS	SRR2009041:Tet2_KO_oxbs
	2	oxBS	SRR2009042:Tet2_KO_oxbs
	3	oxBS	SRR2009043:Tet2_KO_oxbs
v6.5 KD	1	mkBS	SRR2009044:v6.5_KD_mkbs
	2	mkBS	SRR2009045:v6.5_KD_mkbs
	3	mkBS	SRR2009046:v6.5_KD_mkbs
	1	oxBS	SRR2009047:v6.5_KD_oxbs
	2	oxBS	SRR2009048:v6.5_KD_oxbs
	3	oxBS	SRR2009049:v6.5_KD_oxbs

Table 3: Data from (Äijö *et al.*, 2016)

amount of sample is required, this may be a limiting factor in cases with low tissue availability (FFPE). Bisulfite and oxidative bisulfite treatments to get single base resulution methylcytosine and/or hydrodymethylcytosine are therefore expensive experiment. Some of the reduced genome approaches may therefore be more appropriate e.g. RRBS, 450K/EPIC. However these too heve their limitations (number of CpGs covered).

References

Äijö, T., Huang, Y., Mannerström, H., Chavez, L., Tsagaratou, A., Rao, A., and Lähdesmäki, H. (2016). A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome Biol*, **17**(1).

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylKit: a comprehensive r package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, **13**(10), R87.

Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W., and Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature protocols*, **8**(10), 1841–51.

Booth, M. J., Marsico, G., Bachman, M., Beraldi, D., and Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature Chemistry*, **6**(5), 435–440.

Hadad, N., Masser, D. R., Logan, S., Wronowski, B., Mangold, C. A., Clark, N., Otalora, L., Unnikrishnan, A., Ford, M. M., Giles, C. B., Wren, J. D., Richardson, A., Sonntag, W. E., Stanford, D. R., and Freeman, W. (2016). Absence of genomic hypomethylation or regulation of cytosine-modifying enzymes with aging in male and female mice. *Epigenetics & Chromatin*, **9**(1).

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**(11), 1571–1572.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, **40**(17), e136.