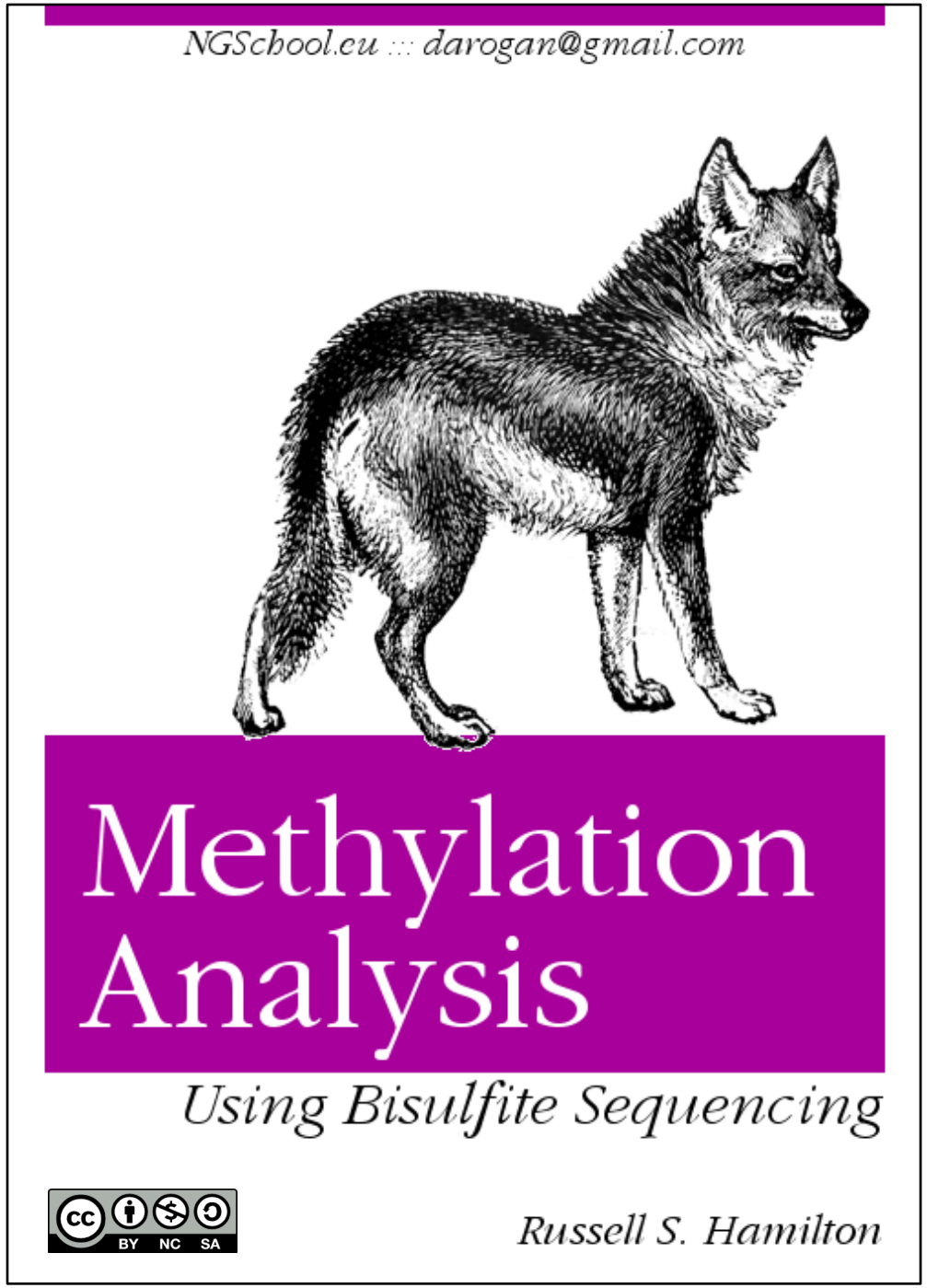


Dr Russell S. Hamilton

Email: rsh46@cam.ac.uk
Twitter: @drrshamilton

Version 0.1: 20160818

Who am I?

Bioinformatics Facility Manager
Centre for Trophoblast Research
University of Cambridge

Support ~30 Groups

Bioinformatician (R&D)
Cambridge Epigenetix

Bisulfite sequencing kits

Conflict of Interest
Shareholder in CEGX

Course Materials

```
rsync -aurvz USERNAME@ngschool.local:/ngschool/Methylation /ngschool --exclude='.git'
```

Presentation

- `Presentation/Hamilton_Methylation_Presentation.pdf`

FastQ files & simple bash scripts:

- `SimulatedData/`

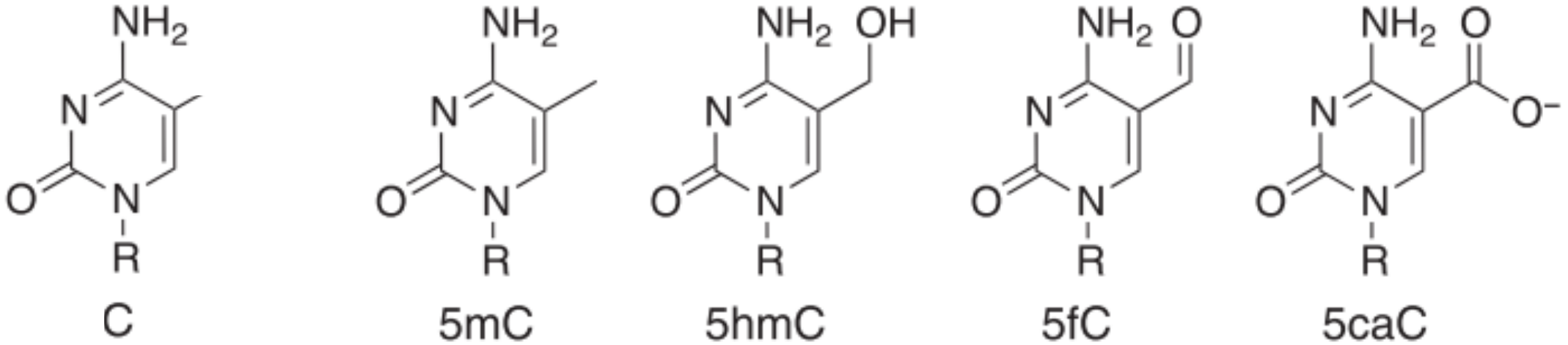
Reference genome:

- `NGSchool_GRCh38_Chr1_region/`

Pre Processed Data

- `SimulatedData_PreAnalysed/`

What is Methylation?

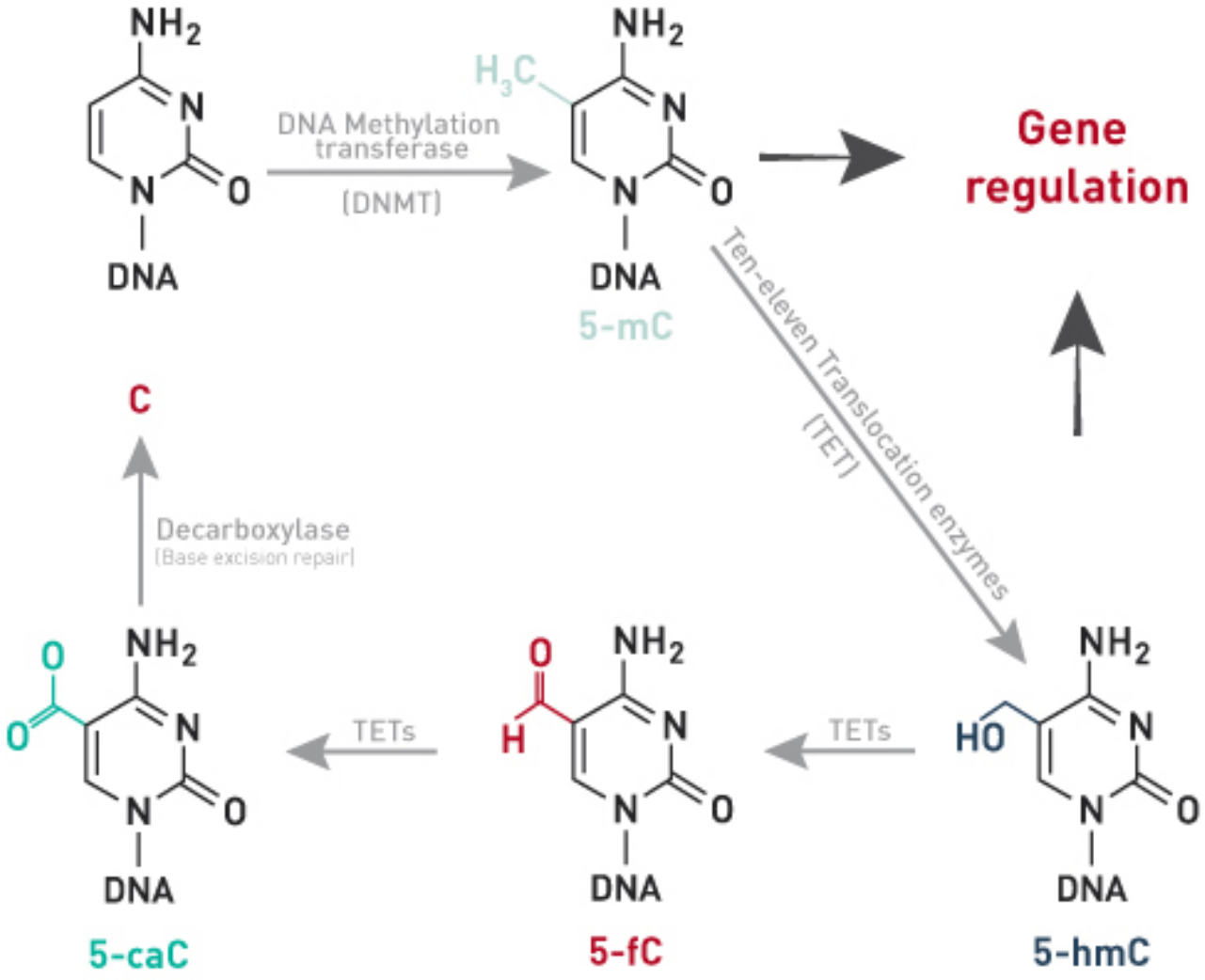


cytosine *5'-methyl-C* *5'-hydroxymethyl-C* *5'-formyl-C* *5'carboxy-C*

Modifications to cytosine are the most widely studied, however Adenosine also known to be methylated m6A

[Figure adapted from Booth et al., 2013]

What is Methylation?



[Figure from www.diagenode.com]

Methylation Function in Development

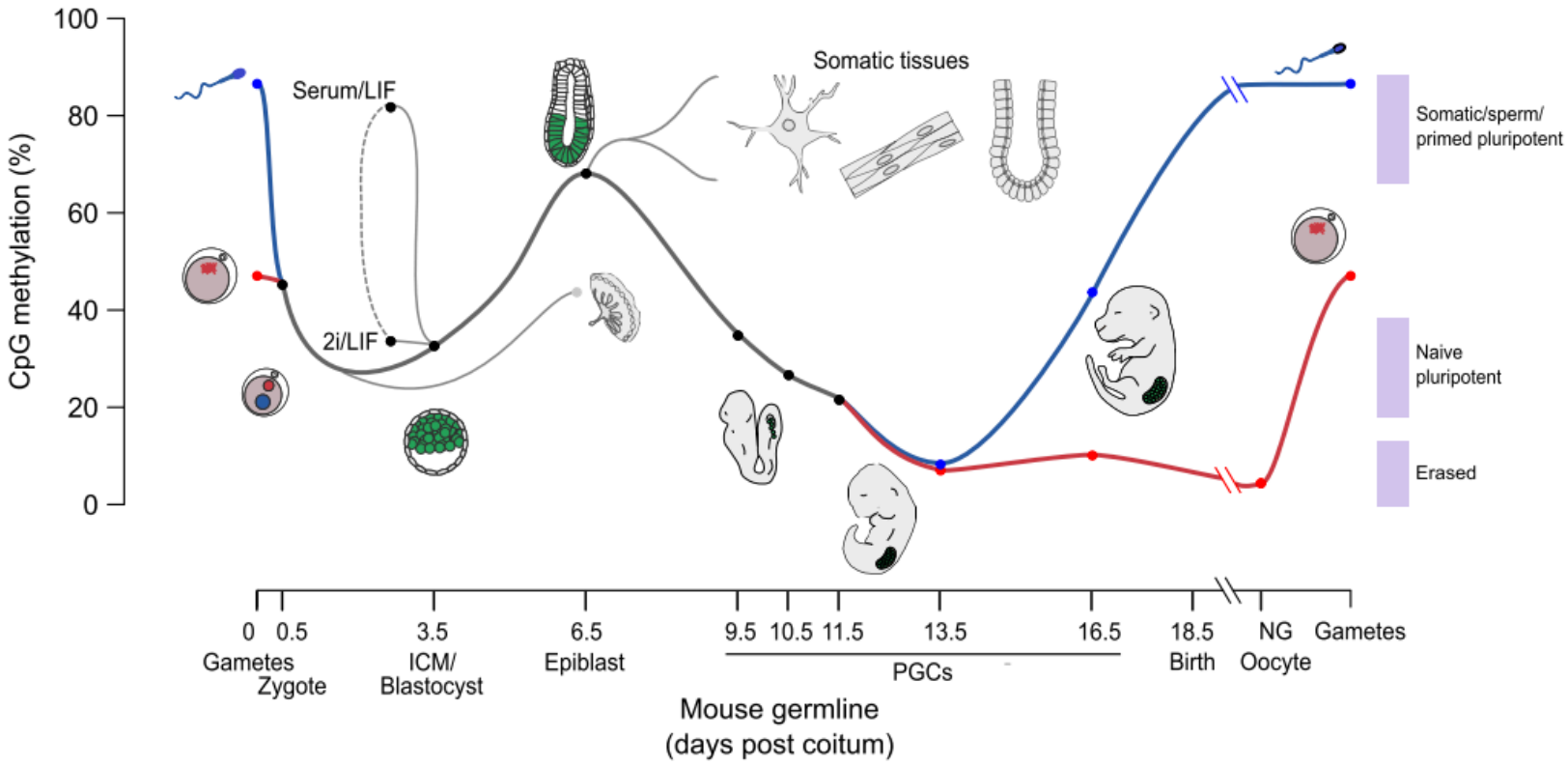
Clark et al. Genome Biology (2016) 17:72
DOI 10.1186/s13059-016-0944-x

Genome Biology

OPINION Open Access

Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity

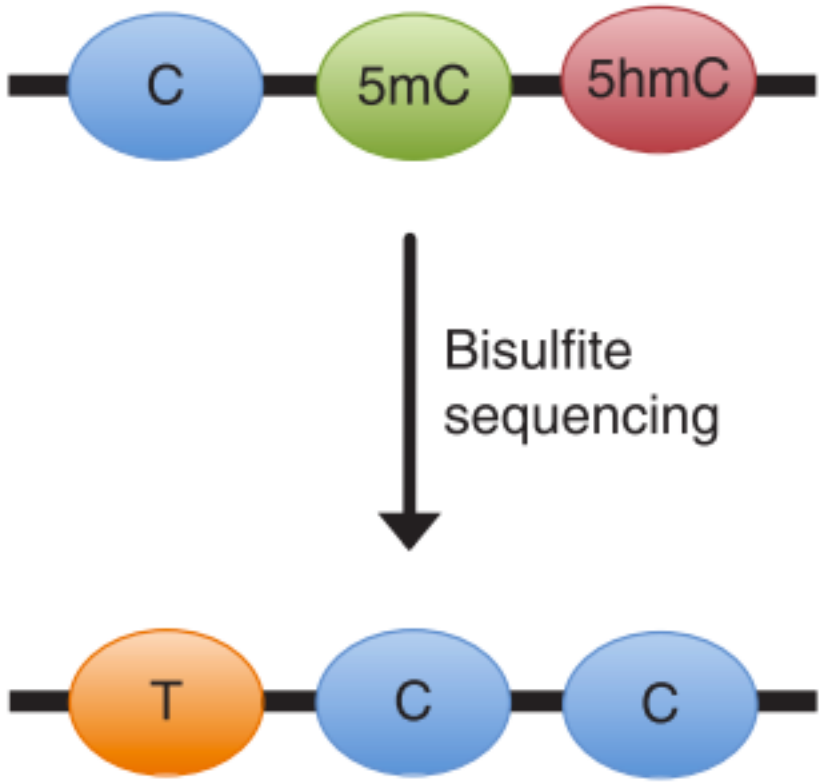
Stephen J. Clark[1], Heather J. Lee[1,2*], Sébastien A. Smallwood[1,3], Gavin Kelsey[1,4] and Wolf Reik[1,2,4,5]



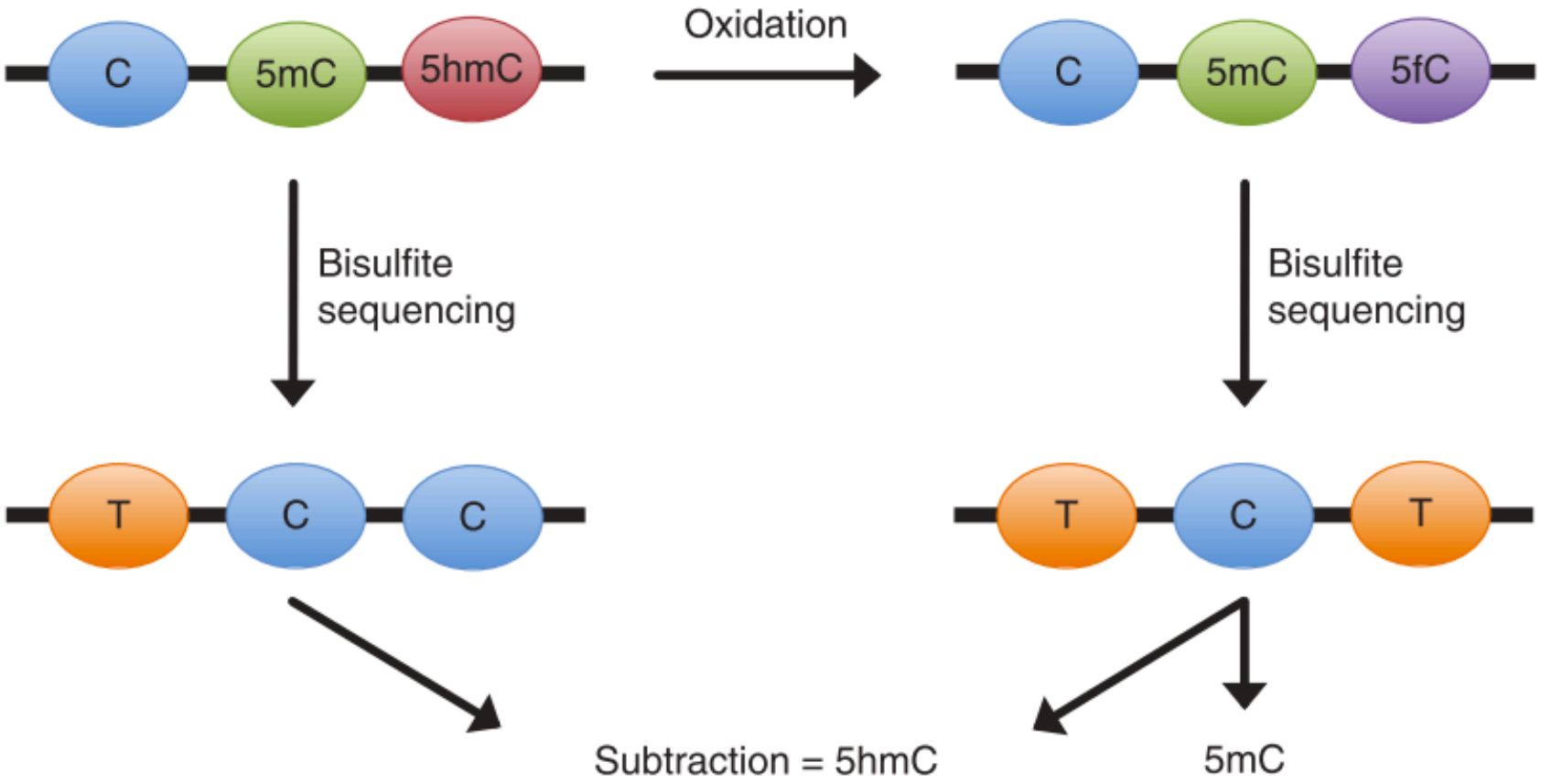
Bisulfite Sequencing (bs or mkbs)

Traditional bisulfite sequencing

Confounded signal 5mC + 5hmC

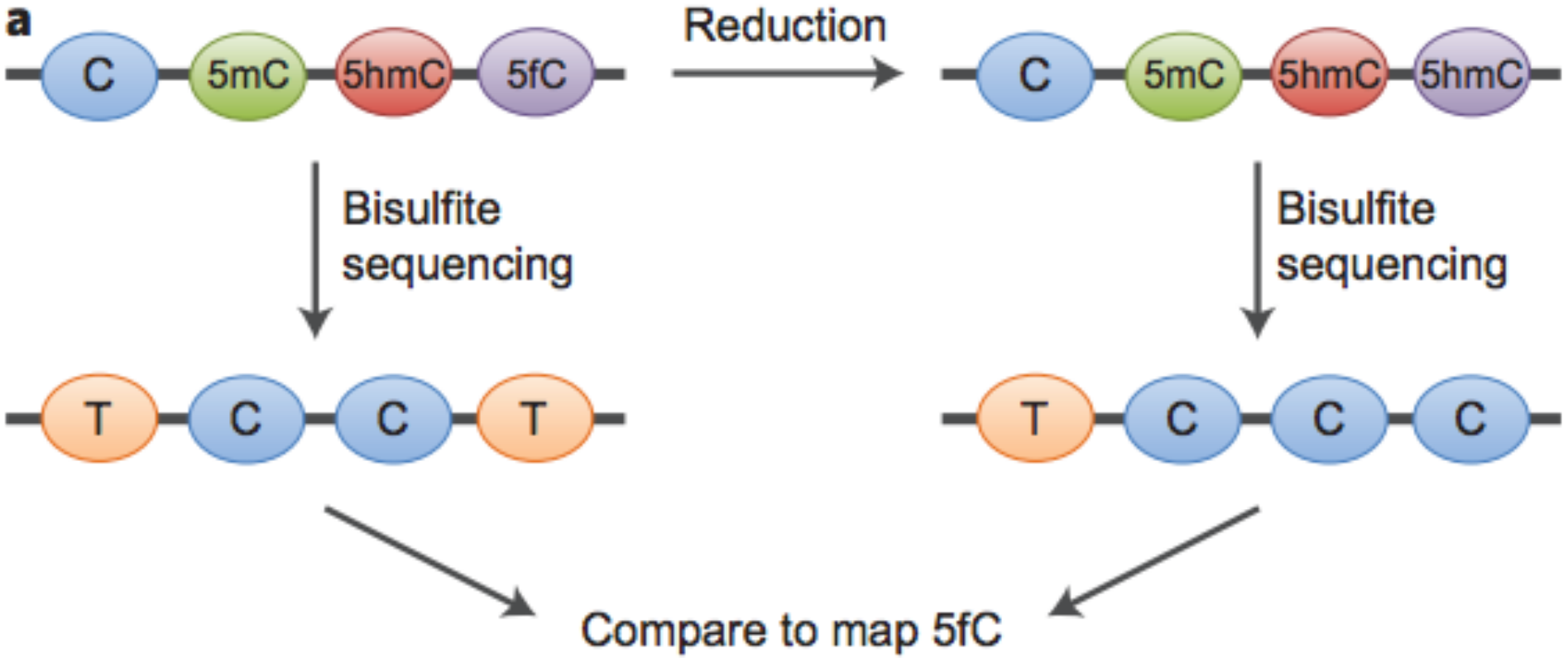


Oxidative bisulfite sequencing (oxbs)



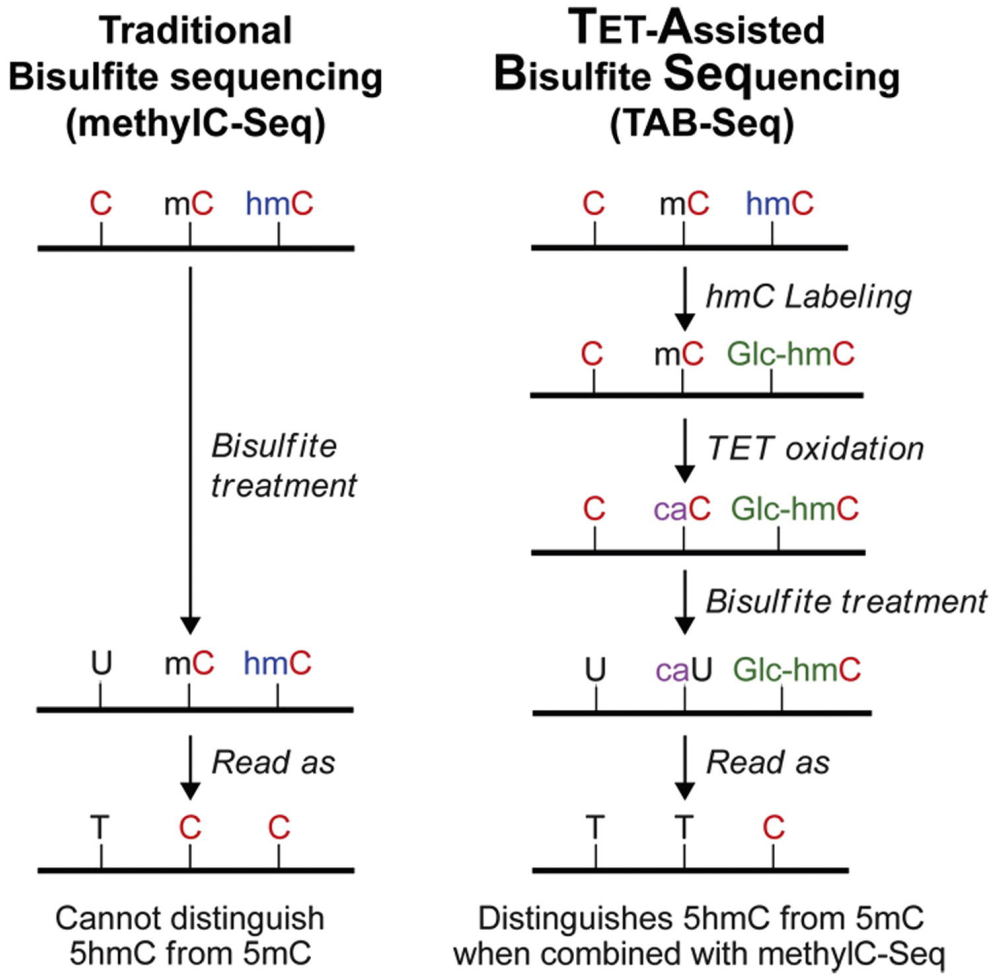
Commercialised by Cambridge Epigenetix (CEGX)

Reductive bisulfite sequencing



Commercialised by Cambridge Epigenetix (CEGX)

TAB-Seq



Commercialised by WiseGene

Assays using bisulfite sequencing

Illumina Array Based

- Illumina 450K array
- Illumina EPIC (850K) array

Illumina Sequencing Based

- Whole Genome Bisulfite Sequencing (WGBS) inc oxbs/redbs ⇐ *Current State of the Art*
- Antibody Pulldown (e.g. (h)MeDIP)
- Reduced Representation Bisulfite Sequencing (RRBS)
- Targeted Bisulfite Sequencing (e.g. CpGiant)
- Enzymatic conversion (TAB-Seq)

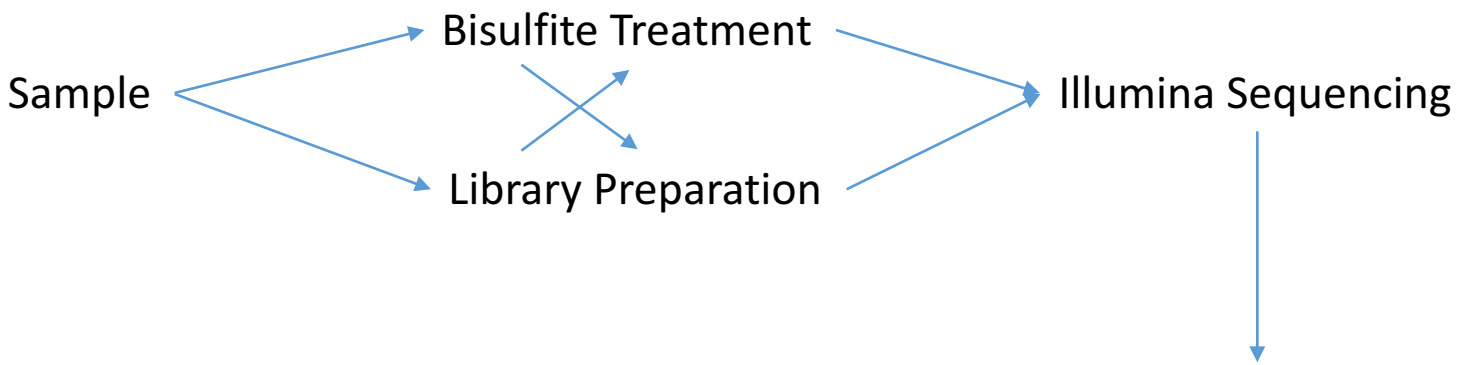
PacBio

- Direct reading of cytosine methylation

Nanopore

- Oxford Nanopore direct reading of cytosine methylation ⇐ *Future State of the Art*

Whole Genome Bisulfite Sequencing



Bisulfite aware genome aligner

• Bismark	*conservative*	*80+%*	*best supported*
• ERNE2	*balanced*	*90+%*	*fast*
• BWA-Meth	*leanient*	*95+%*	*soft-clipping*

Bismark Aligner

BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 11 2011, pages 1571–1572
doi:10.1093/bioinformatics/btr167

Sequence analysis

Advance Access publication April 14, 2011

Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications

Felix Krueger* and Simon R. Andrews

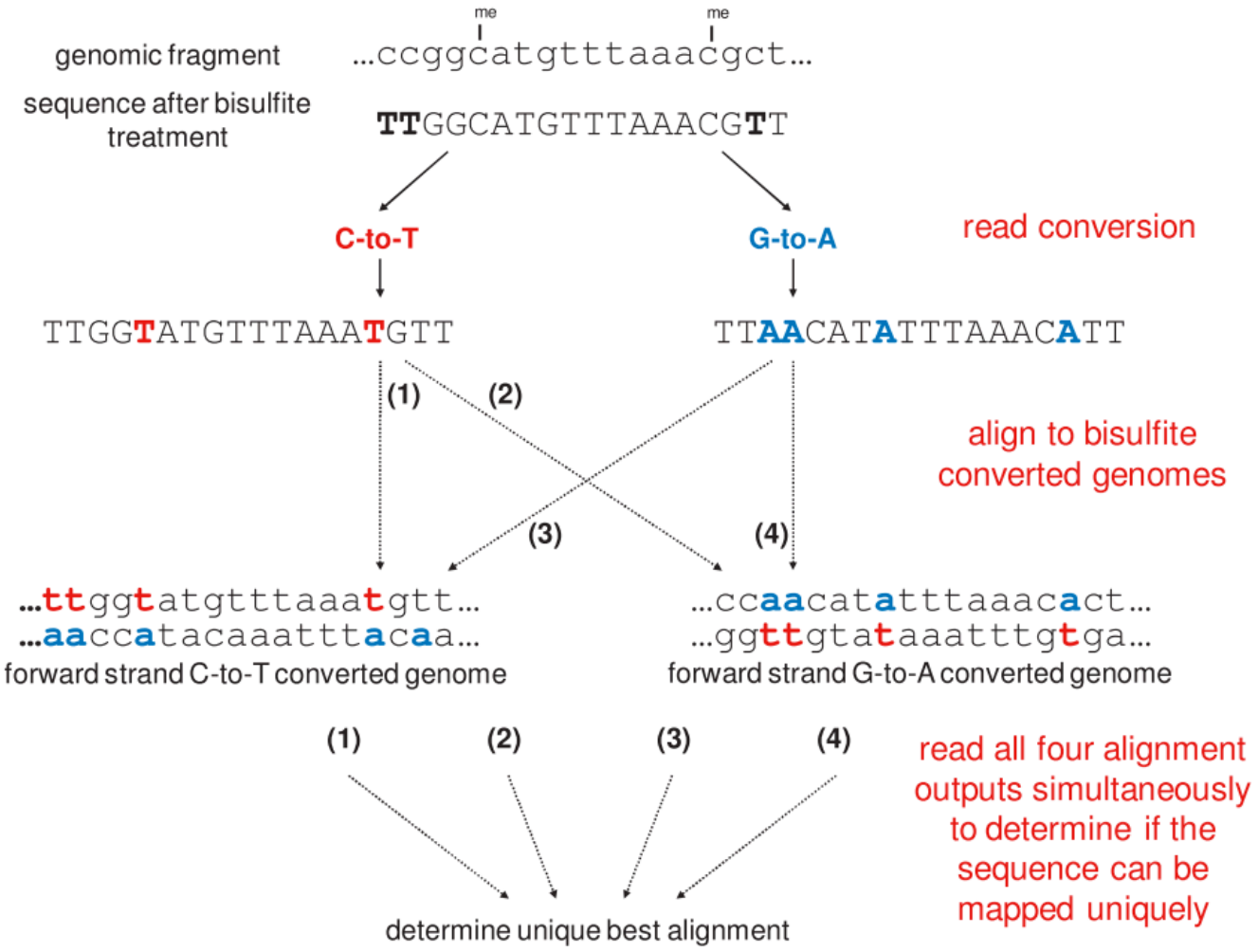
Bioinformatics Group, The Babraham Institute, CB22 3AT, Cambridge, UK

Associate Editor: Alfonso Valencia

- Arguable the most widely used bisulfite aware aligner…
- Very well supported (seqanswers.com, GitHub)
- Integrates with downstream tools

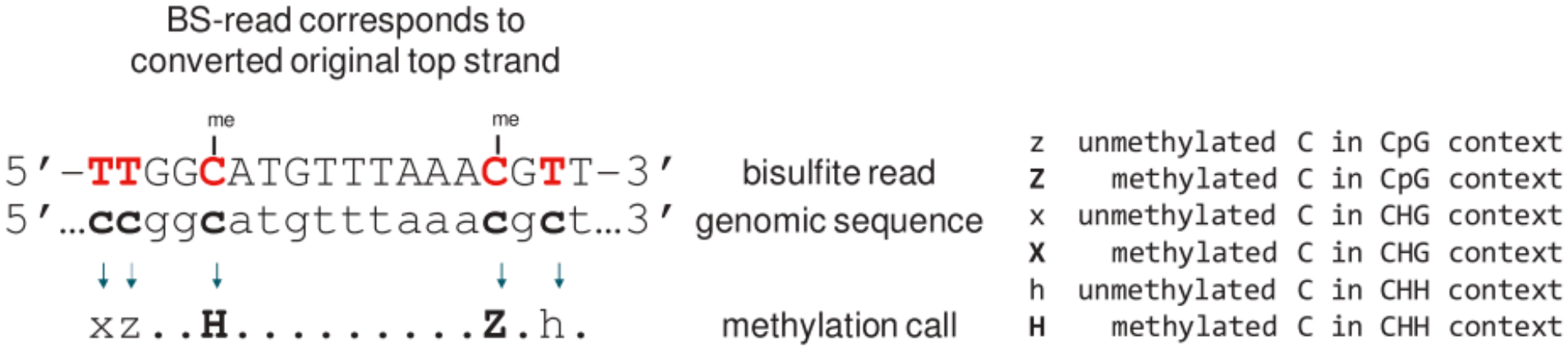
[Figure from Krueger F & Andrews SR (2011) Bioinformatics. 27(11):1571-2]

Bismark Aligner



[Figure from Krueger F & Andrews SR (2011) Bioinformatics. 27(11):1571-2]

Bismark Methylation Calling



[Figure from Krueger F & Andrews SR (2011) Bioinformatics. 27(11):1571-2]

Practical Pre-made data

1. Reference Genome Prepared for Bisulfite Alignment

Human Genome ~3Gb disk space
With preparation for bs-seq ~11Gb

Very CPU / Memory intensive…

Selected a random 5Mb region from Chr1
(no N content)

4.9Mb Homo_sapiens.GRCh38.dna.chromosome.1.region30000000-5000000.fa
39Mb Bisulfite_Genome

2. Bisulfite Treated and Sequenced Reads

Typically 30x coverage recommended to achieve accurate single base resolution methylation calls.

Very CPU / Memory intensive…

Simulated a data set using reference region and Sherman

	Read1	Read2
mkbs	1M/36Mb	1M/37Mb
oxbs	1M/38Mb	1M/37Mb

Bisulfite Genome Preparation

Reference genome must be prepared for bisulfite alignment

```
$ bismark_genome_prepare --bowtie2 NGSchool_GRCh38_Chr1_region
```

```
├── Homo_sapiens.GRCh38.dna.chromosome.1.region30000000-5000000.fa
├── Bisulfite_Genome
    ├── CT_conversion
    |   ├── BS_CT.1.bt2
    |   ├── BS_CT.2.bt2
    |   ├── BS_CT.3.bt2
    |   ├── BS_CT.4.bt2
    |   ├── BS_CT.rev.1.bt2
    |   ├── BS_CT.rev.2.bt2
    |   └── genome_mfa.CT_conversion.fa
    └── GA_conversion
        ├── BS_GA.1.bt2
        ├── BS_GA.2.bt2
        ├── BS_GA.3.bt2
        ├── BS_GA.4.bt2
        ├── BS_GA.rev.1.bt2
        ├── BS_GA.rev.2.bt2
        └── genome_mfa.GA_conversion.fa
```

Simulating Bisulfite Treated Reads

Simulate a set of 1M bisulfite reads to achieve ~30X coverage Chr1:30000000-5000000

```
$ Sherman \
    --length 100 \
    --number_of_seqs 1000000 \
    --genome_folder NGSchool_GRCh38_Chr1_region/ \
    --paired_end \
    --minfrag 70 \
    --maxfrag 400 \
    --conversion_rate 99 \
    --error_rate 0.25 \
    --variable_length_adapter 100

mv simulated_1.fastq mkbs_sim_1.fastq; gzip mkbs_sim_1.fastq
mv simulated_2.fastq mkbs_sim_2.fastq; gzip mkbs_sim_2.fastq
```

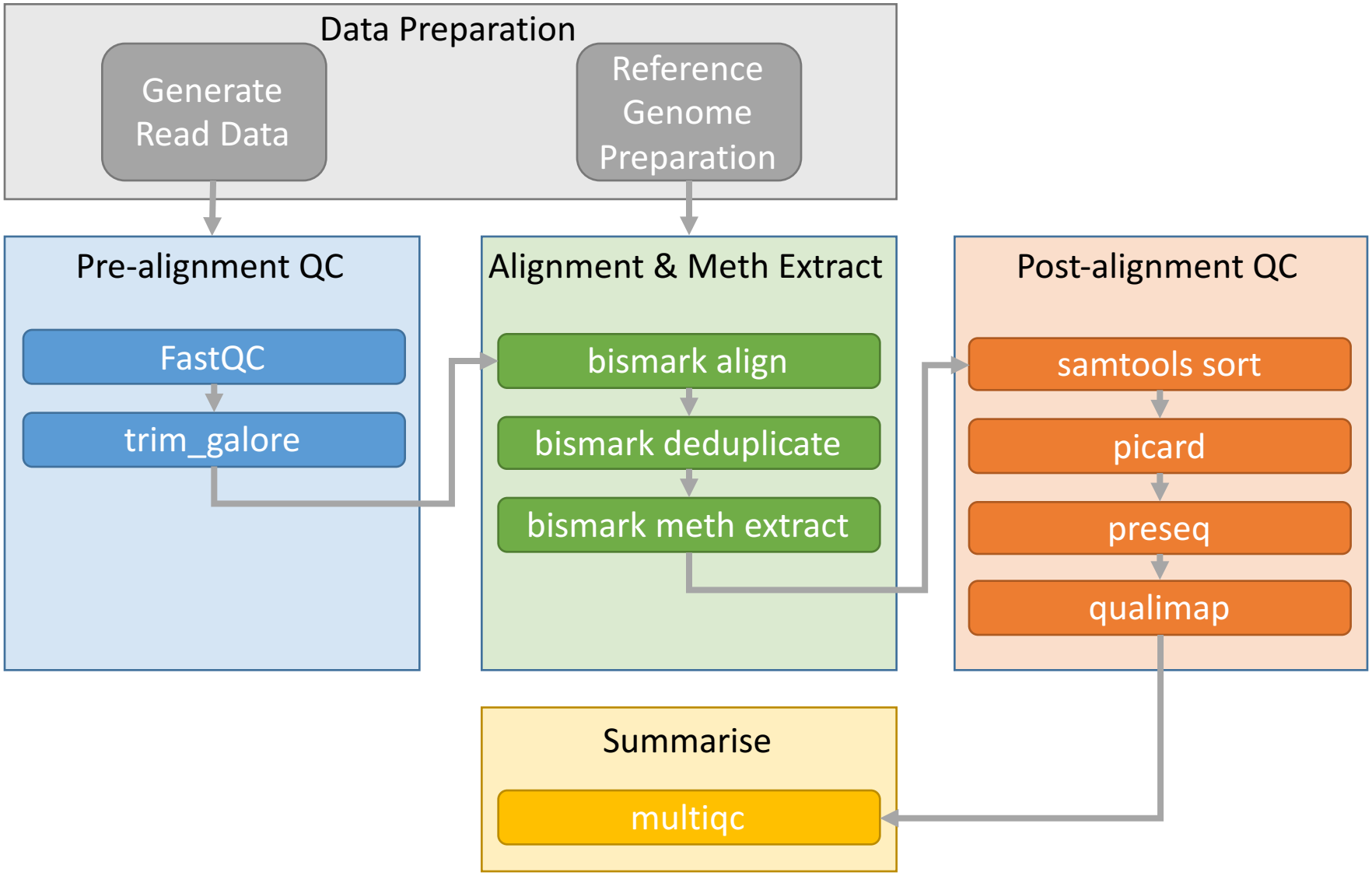
Simulate a set of 1M oxidative bisulfite reads

As above, but with

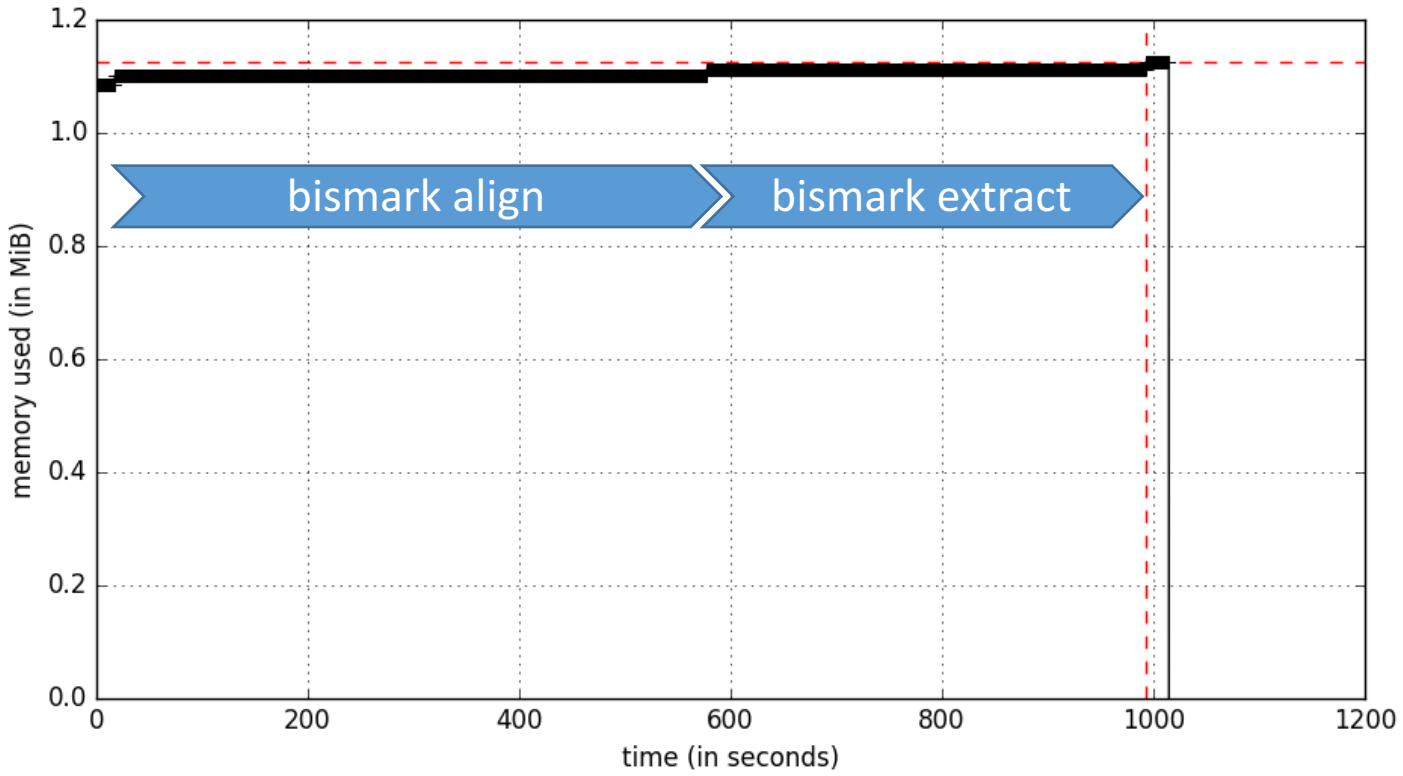
```
    --conversion_rate 90

mv simulated_1.fastq oxbs_sim_1.fastq; gzip oxbs_sim_1.fastq
mv simulated_2.fastq oxbs_sim_2.fastq; gzip oxbs_sim_2.fastq
```

http://www.bioinformatics.babraham.ac.uk/projects/sherman

Pipeline


Data Preparation
Generate Read Data
Reference Genome Preparation
Pre-alignment QC
FastQC
trim_galore
Alignment & Meth Extract
bismark align
bismark deduplicate
bismark meth extract
Post-alignment QC
samtools sort
picard
preseq
qualimap
Summarise
multiqc

Pipeline Run Time



Serial pipeline single core/thread
MacBook Pro

Bs-seq: 1M reads from 5Mb Region of Chr1

Total Memory:	~1.1Gb
Run Time:	1000 seconds
	17 minutes

Full Genome

Total Memory:	~11Gb RAM
Run Time:	~12 Hours
	depending on #reads

NGS Pipelines

- Processing a large number of samples in a consistent, documented and reproducible manner
- Optimise CPU usage with queuing system GRIDengine, SLURM etc
- Pipelines can be custom bash scripts, Docker containers or specific pipeline tools.
- Exact versions and command line options should be recorded in log files.

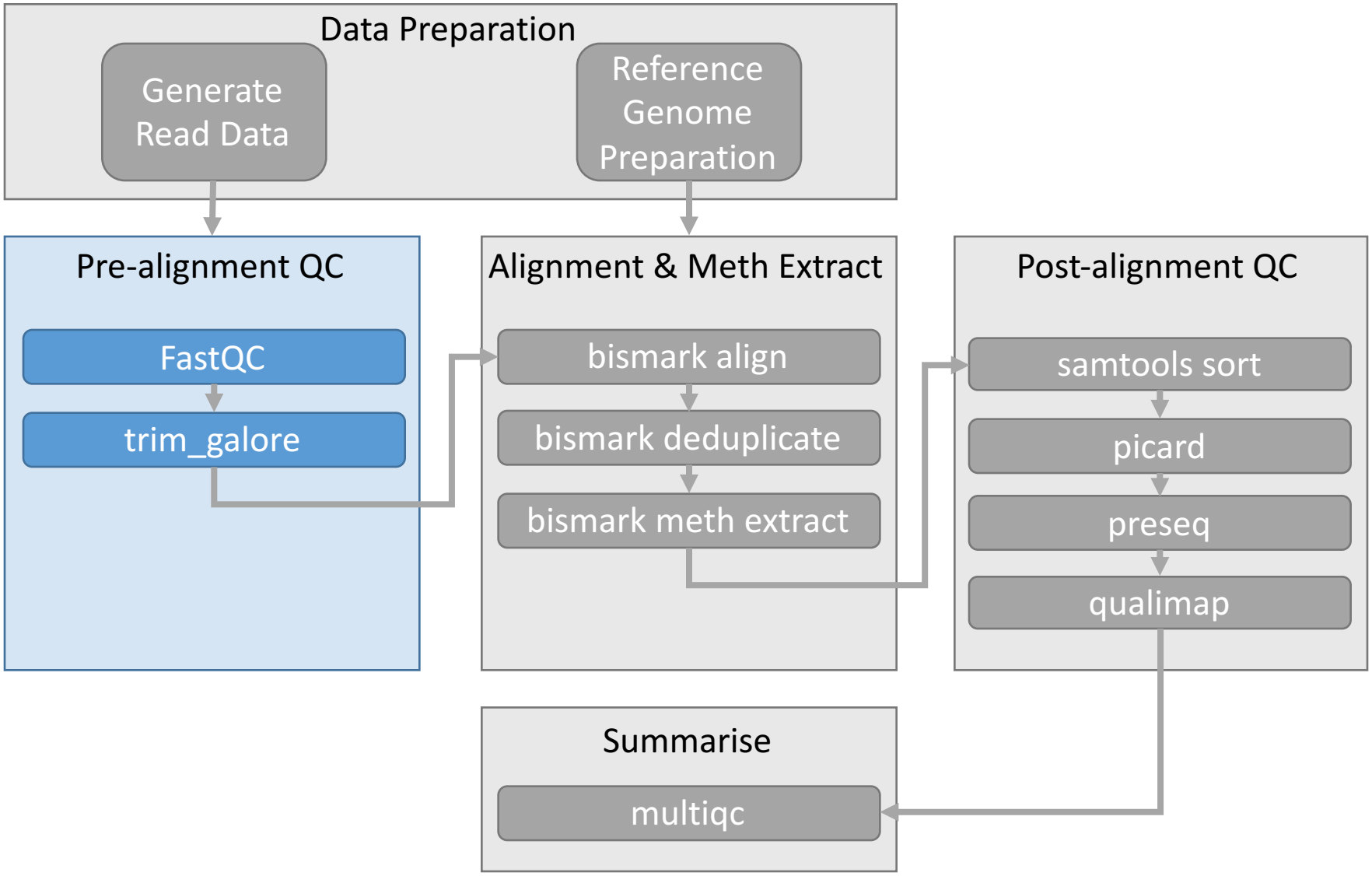
http://clusterflow.io

Single command: `$ cf --genome GRCh38 bismark_pipeline *.fa.gz`

Pipeline outline:

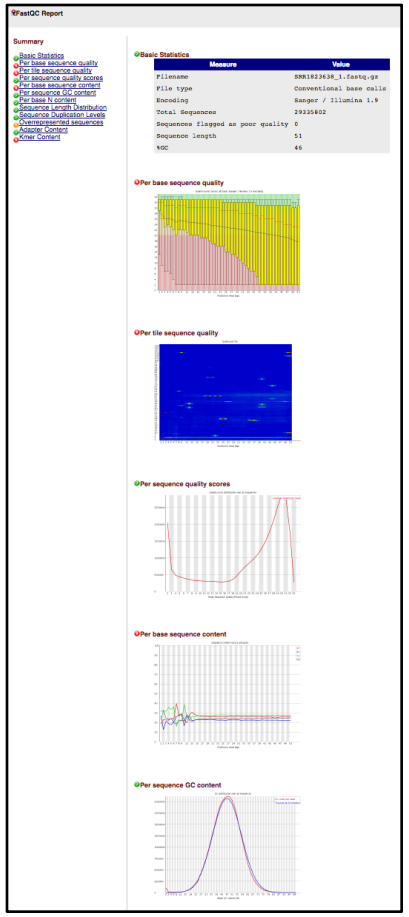
```
#fastqc
#trim_galore
          #bismark_align
                    #bismark_deduplicate
                              #samtools_sort coord
                                        #preseq_lc_extrap
                                        #preseq_bound_pop
                                        #qualimap_bamqc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

Pipeline


Data Preparation
Generate Read Data
Reference Genome Preparation
Pre-alignment QC
FastQC
trim_galore
Alignment & Meth Extract
bismark align
bismark deduplicate
bismark meth extract
Post-alignment QC
samtools sort
picard
preseq
qualimap
Summarise
multiqc

FastQC

FastQC	A quality control tool for high throughput sequence data
Version	0.11.5
Download	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Terminal:

```
$ fastqc –q mkbs_sim_1000000_0.25_1.fastq.gz
$ firefox mkbs_sim_1000000_0.25_1_fastqc.html
```

Read 1

```
$ fastqc –q mkbs_sim_1000000_0.25_2.fastq.gz
$ firefox mkbs_sim_1000000_0.25_1_fastqc.html
```

Read 2

Output:

HTML Reports

```
mkbs_sim_1000000_0.25_1_fastqc.html
mkbs_sim_1000000_0.25_2_fastqc.html
```

Archive of data/images

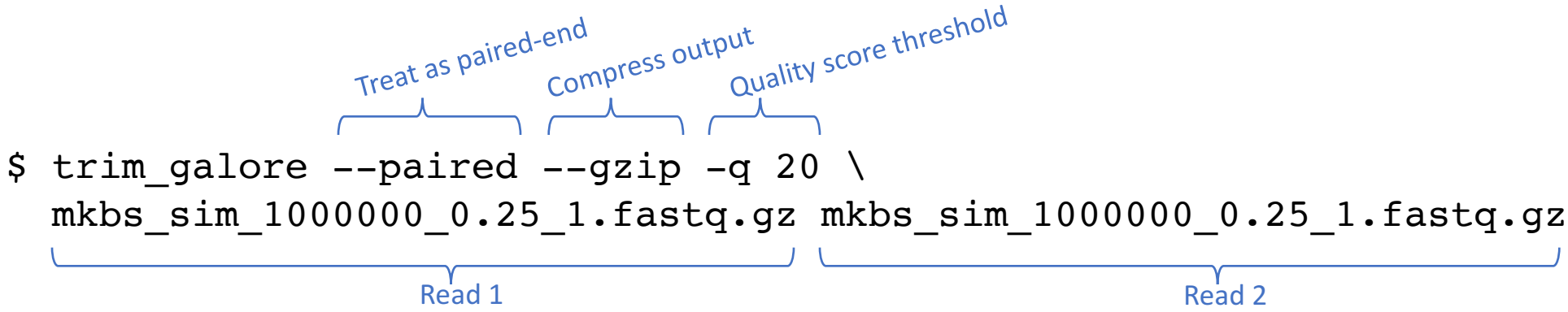
```
mkbs_sim_1000000_0.25_1_1_fastqc.zip
mkbs_sim_1000000_0.25_2_fastqc.zip
```

Bioinformatics Top Tip:
Simon Andrews' https://sequencing.qcfail.com/

trim_galore

trim_galore	A wrapper tool around Cutadapt to consistently apply quality and adapter trimming to FastQ files
Version	0.4.1
Download	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Terminal:



Output:

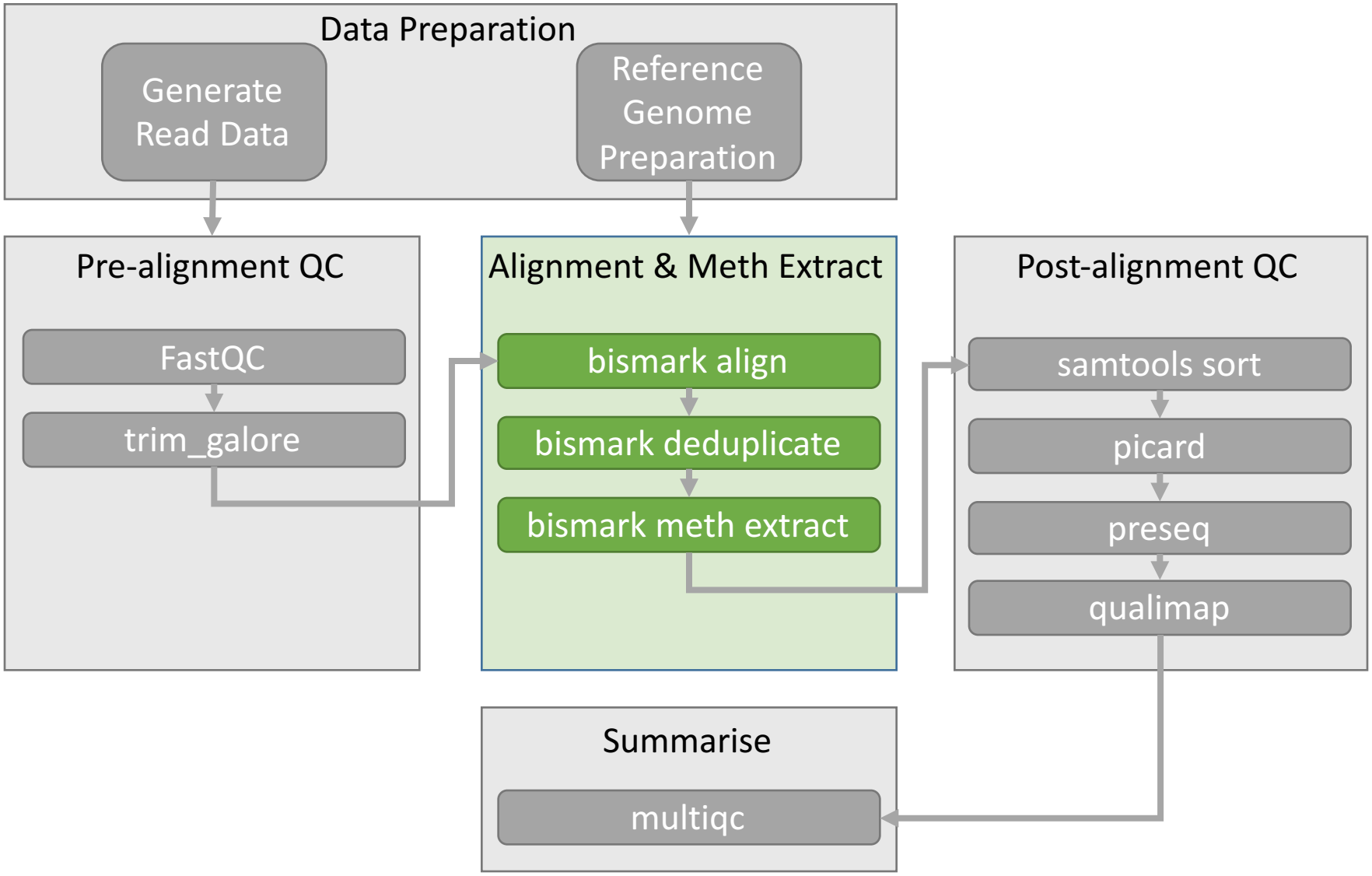
Trimmed Fastq files

```
mkbs_sim_1000000_0.25_1_val_1.fq.gz
mkbs_sim_1000000_0.25_2_val_2.fq.gz
```

Trimming report files

```
mkbs_sim_1000000_0.25_1.fastq.gz_trimming_report.txt
mkbs_sim_1000000_0.25_2.fastq.gz_trimming_report.txt
```

Pipeline


Data Preparation
Generate Read Data
Reference Genome Preparation
Pre-alignment QC
FastQC
trim_galore
Alignment & Meth Extract
bismark align
bismark deduplicate
bismark meth extract
Post-alignment QC
samtools sort
picard
preseq
qualimap
Summarise
multiqc

Bismark alignment

bismark	A tool to map bisulfite converted sequence reads and determine cytosine methylation states
Version	0.16.3
Download	http://www.bioinformatics.babraham.ac.uk/projects/bismark/

Terminal:

Use bowtie2 (>50nt)

Reference Genome Location

```
$ bismark --bowtie2 NGSchool_GRCh38_Chr1_region \
 -1 mkbs_sim_1000000_0.25_1_val_1.fq.gz -2 mkbs_sim_1000000_0.25_2_val_2.fq.gz
```

Read 1

Read 2

Output:

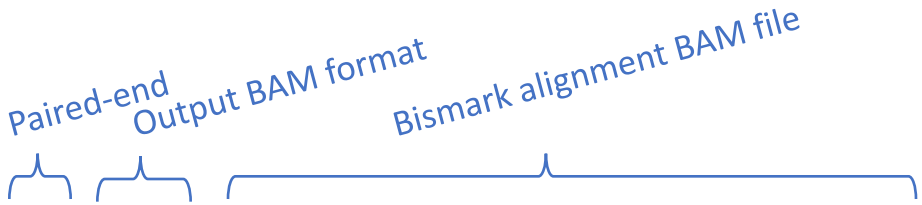
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_PE_report.txt
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.bam

OSX Trouble Shooting:
Some versions of bismark use zcat
Fix by using sed to replace zcat with gunzip -c

Bismark Deduplicate

bismark	A tool to map bisulfite converted sequence reads and determine cytosine methylation states
Version	0.16.3
Download	http://www.bioinformatics.babraham.ac.uk/projects/bismark/

Terminal:




```
$ deduplicate_bismark -p –bam mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.bam
```

Output:

```
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.bam
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplication_report.txt
```

Bismark methylation extractor

bismark	A tool to map bisulfite converted sequence reads and determine cytosine methylation states
Version	0.16.3
Download	http://www.bioinformatics.babraham.ac.uk/projects/bismark/

Terminal:

```
$ bismark_methylation_extractor --ignore_r2 1 --ignore_3prime_r2 2 \
  --bedGraph --gzip -p --no_overlap --report \
  mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.bam
```

Ignore first 5' bp Read 2

Ignore the last 2 3' bp Read 2

Cytosine Methylation

Compress

PE

Ignore PE overlap

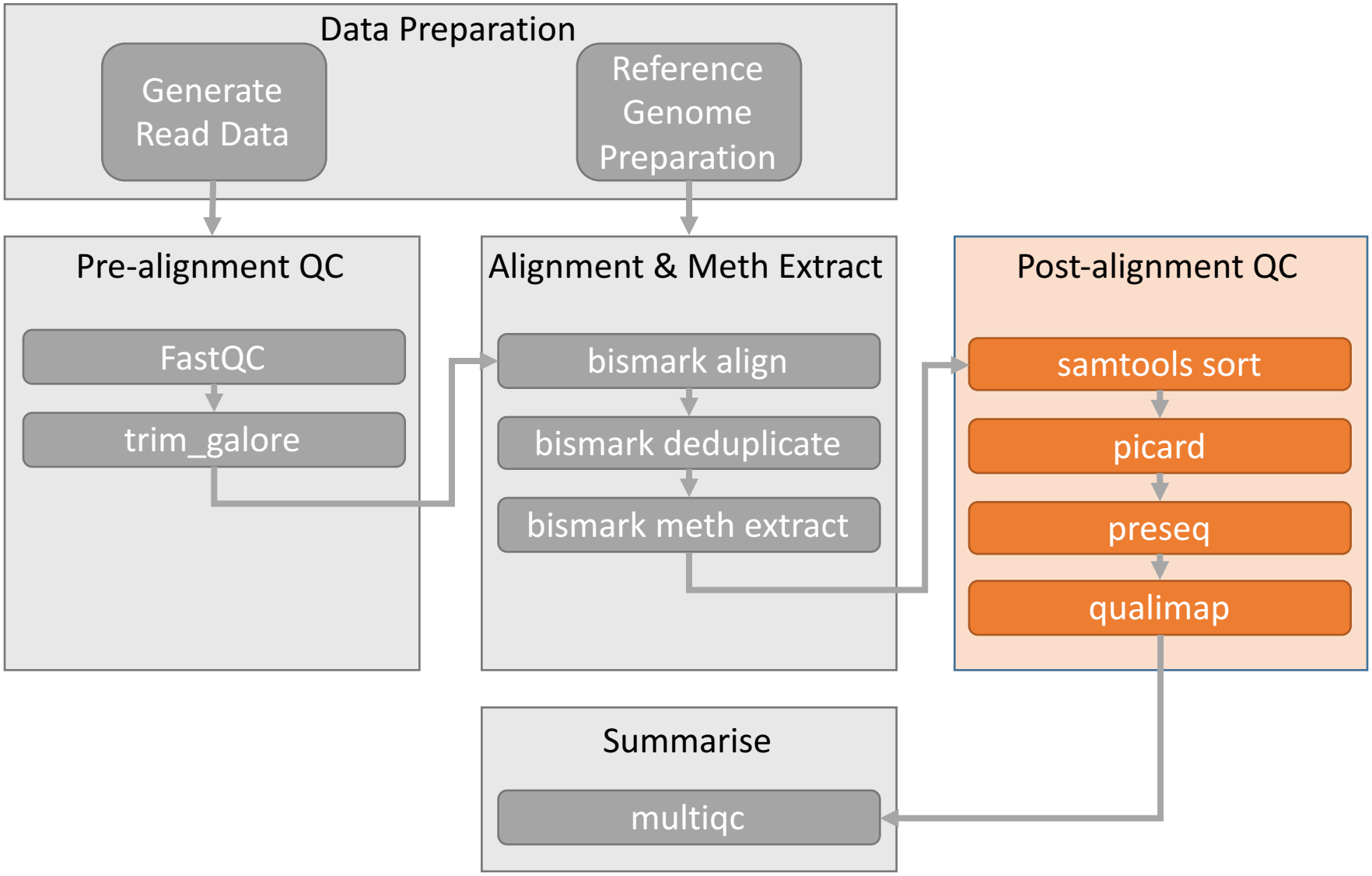
Final summary

Bismark deduplicated BAM file

Output:

mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.M-bias.txt
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.M-bias_R1.png
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.M-bias_R2.png
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.bedGraph.gz
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated.bismark.cov.gz
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.deduplicated_splitting_report.txt

Pipeline


Data Preparation
Generate Read Data
Reference Genome Preparation
Pre-alignment QC
FastQC
trim_galore
Alignment & Meth Extract
bismark align
bismark deduplicate
bismark meth extract
Post-alignment QC
samtools sort
picard
preseq
qualimap
Summarise
multiqc

Samtools sort

samtools	Utilities for the Sequence Alignment/Map (SAM) format
Version	1.3.1
Download	http://www.htslib.org/download

Terminal:

Output coordinate sorted BAM file

```
$ samtools sort -o mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam \
  mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.bam
```

Input name sorted BAM file

Output:

mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam

Samtools index

samtools	Utilities for the Sequence Alignment/Map (SAM) format
Version	1.3.1
Download	http://www.htslib.org/download

Terminal:

Input name sorted BAM file

```
$ samtools index mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam
```

Output:

```
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam.bai
```

Picard insert size metrics

Picard	Tools for manipulating high-throughput sequencing (HTS) data
Version	2.2.4+
Download	http://broadinstitute.github.io/picard/

Terminal:

Location of Picard Jar File

Picard Tool to run

```
$ java -jar /usr/local/picard-tools-2.2.4/picard.jar CollectInsertSizeMetrics \
   INPUT=mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam
   OUTPUT=mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd_picard_insert_size_metrics.txt
   HISTOGRAM_FILE=mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd_picard_insert_size_plot.pdf
   METRIC_ACCUMULATION_LEVEL=ALL_READS
```

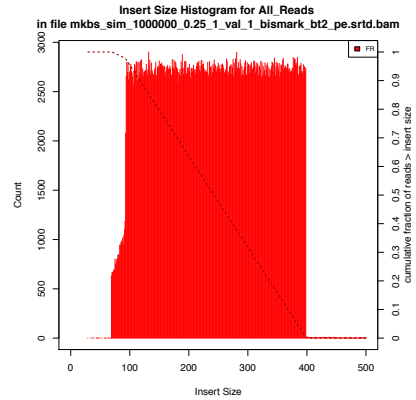
Input coord sorted BAM file

Outputs

Look at all reads

Output:

mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd_picard_insert_size_metrics.txt
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd_picard_insert_size_plot.pdf



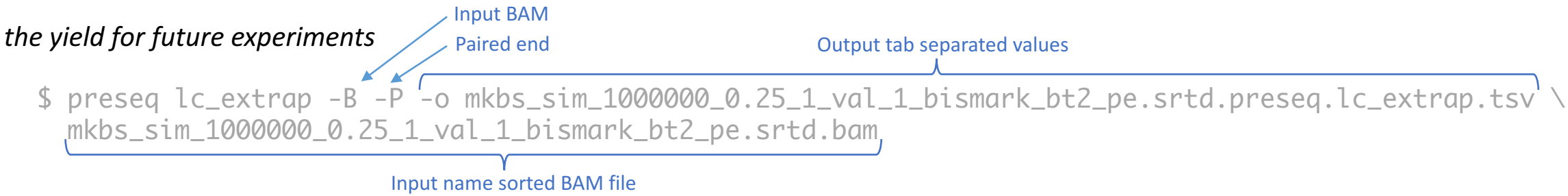
Usually a distribution
Due to simulated data

Preseq

preseq	Predicting and estimating the complexity of a genomic sequencing library
Version	2.0
Download	http://smithlabresearch.org/software/preseq/

Terminal:

predict the yield for future experiments



Output:

mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.preseq.lc_extrap.tsv

Plots unique molecules Vs molecules sequences

Qualimap bamqc

QualiMap	Evaluating next generation sequencing alignment data
Version	2.2
Download	http://qualimap.bioinfo.cipf.es/

Terminal:

Run as command line (useful for HPC!)

```
$ JAVA_OPTS="-Djava.awt.headless=true"
```

Paint chromosome limits inside charts

Input name sorted BAM file

```
$ qualimap bamqc -sd -c -bam mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam
```

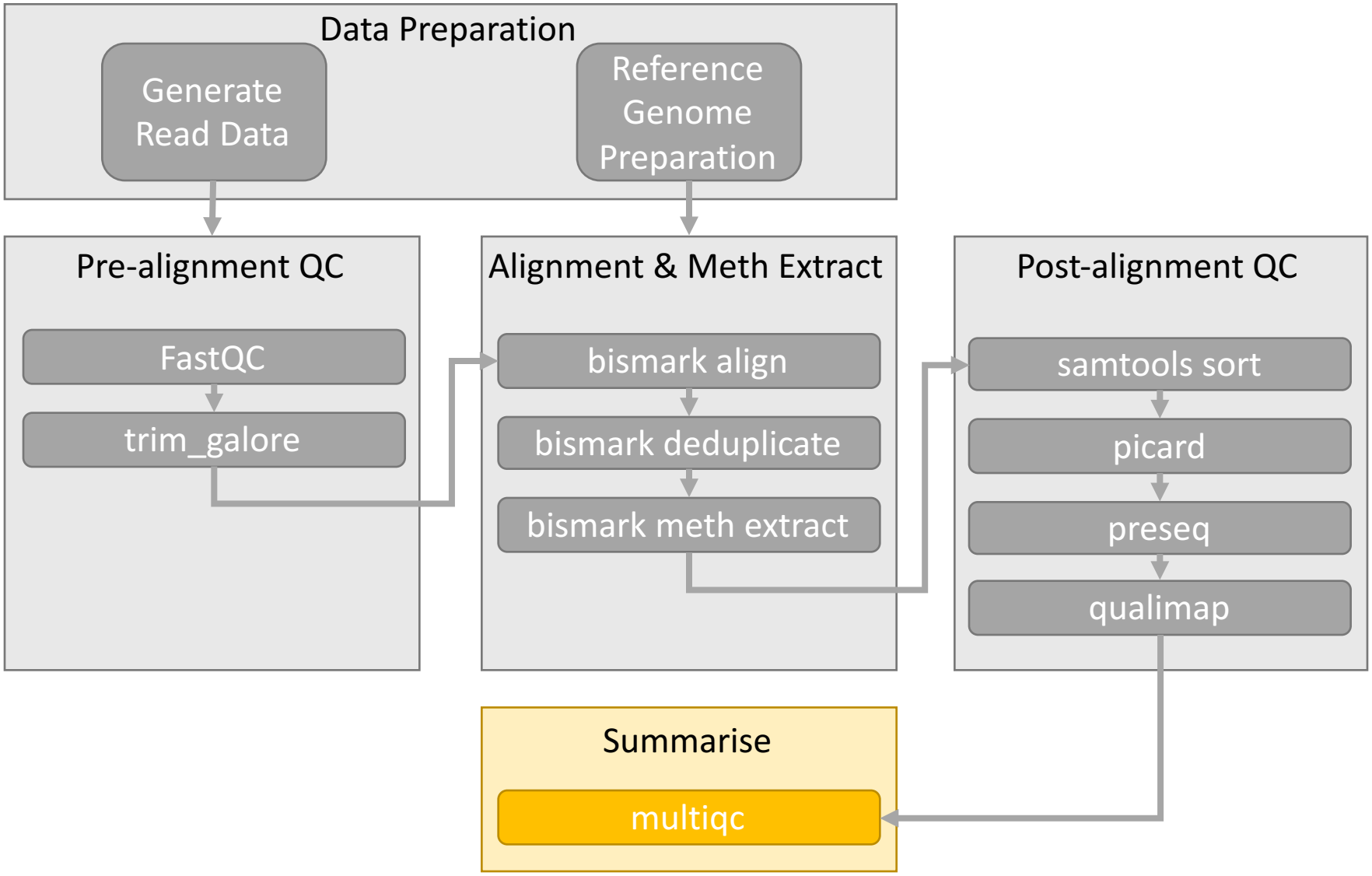
skip duplicated alignments

Input BAM format

Output:

```
mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd_stats/
└── qualimapReport.html
```

Pipeline


Data Preparation
Generate Read Data
Reference Genome Preparation
Pre-alignment QC
FastQC
trim_galore
Alignment & Meth Extract
bismark align
bismark deduplicate
bismark meth extract
Post-alignment QC
samtools sort
picard
preseq
qualimap
Summarise
multiqc

MultiQC

MultiQC	Aggregate results from bioinformatics analyses across many samples into a single report
Version	0.8dev
Download	http://multiqc.info/

Terminal:

A title for your report

Output filename

```
$ multiqc -f -i ”NGSchool.eu" --filename "NGSchool.eu.multiqc_report.html" .
```

Overwrite existing report

“.” Is a special Linux symbol which means the current directory

```
$ NGSchool.eu.multiqc_report.html
```

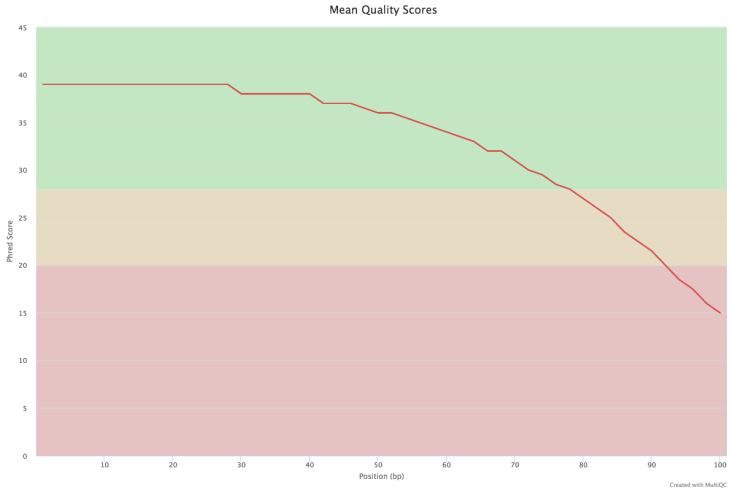
Output:

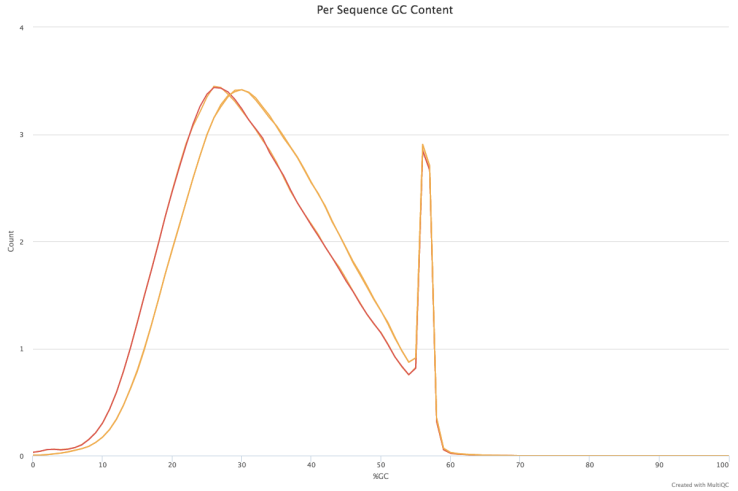
HTML Report

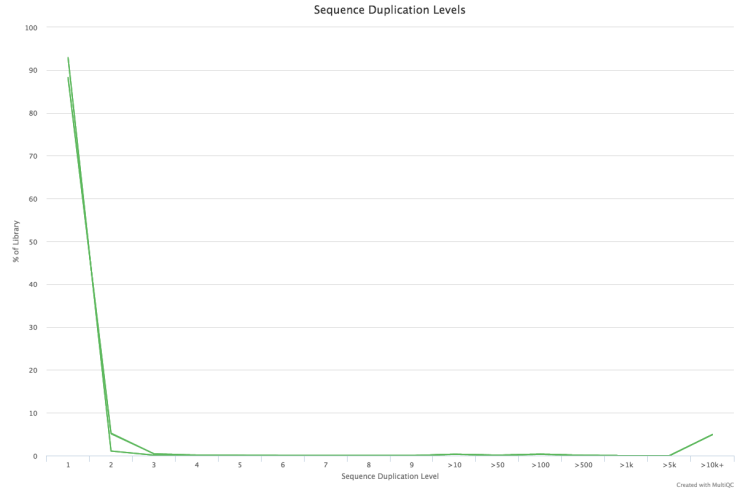
```
NGSchool.eu.multiqc_report.html
NGSchool.eu.multiqc_report_data
```

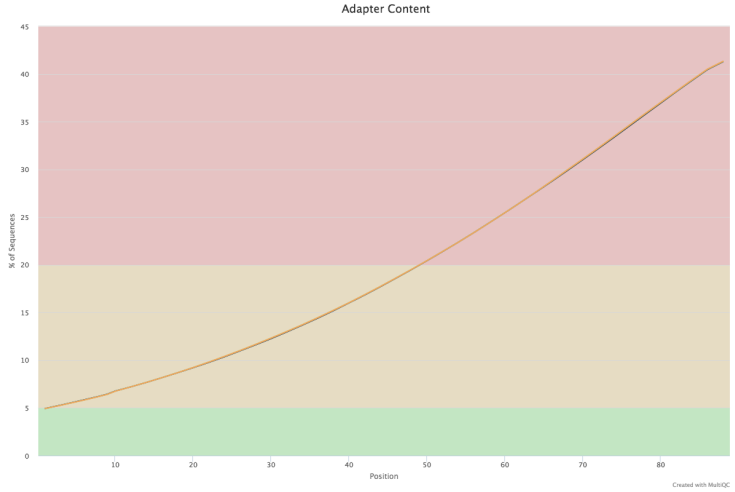
MultiQC ::: FastQC

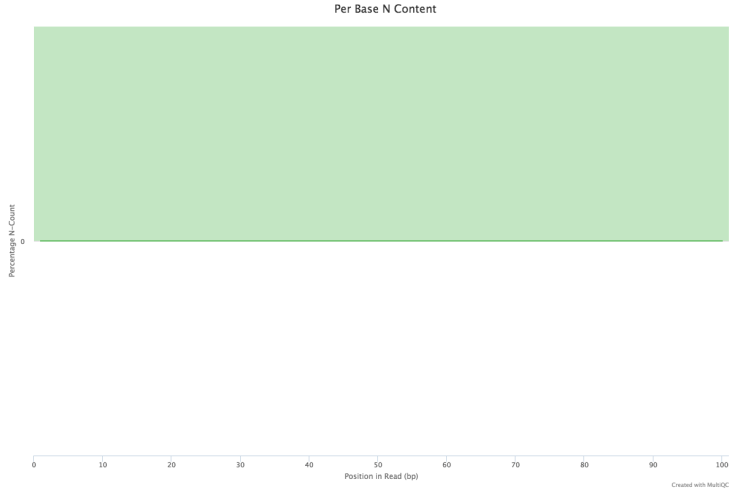
Includes mkbs & oxbs

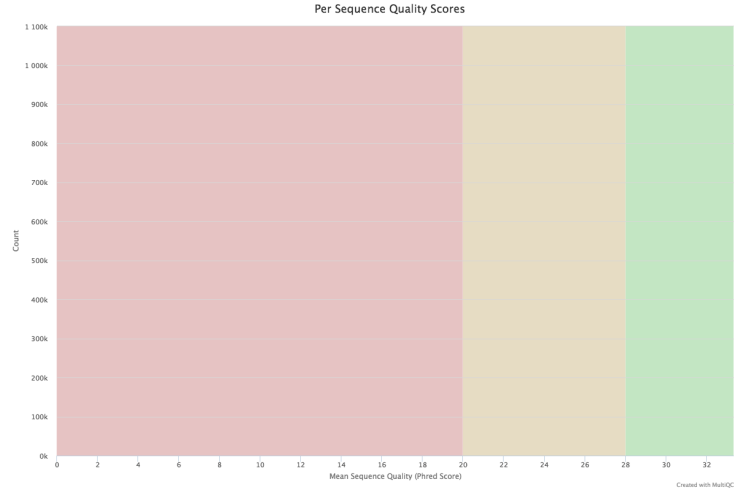






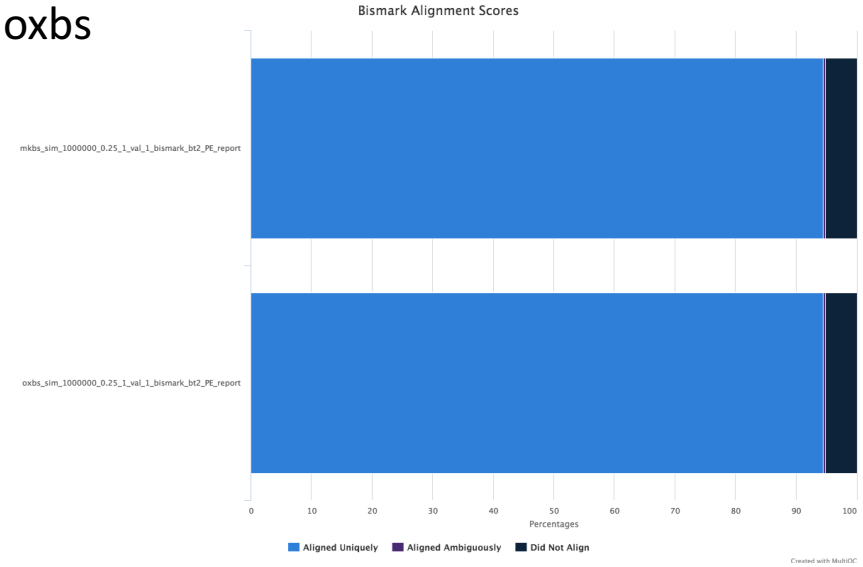


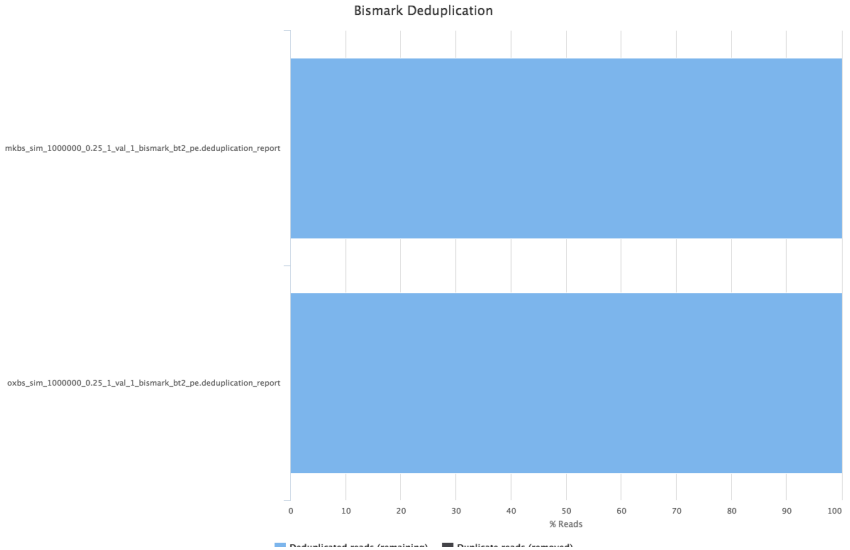


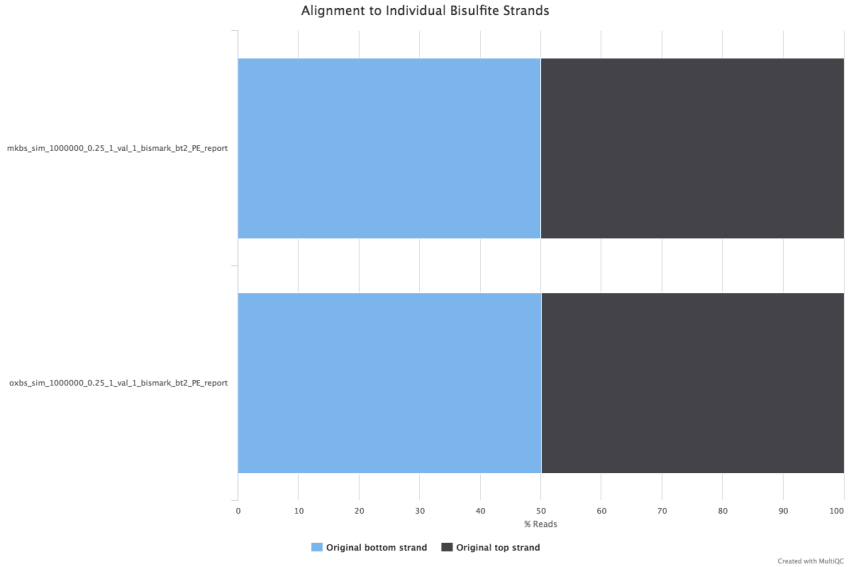


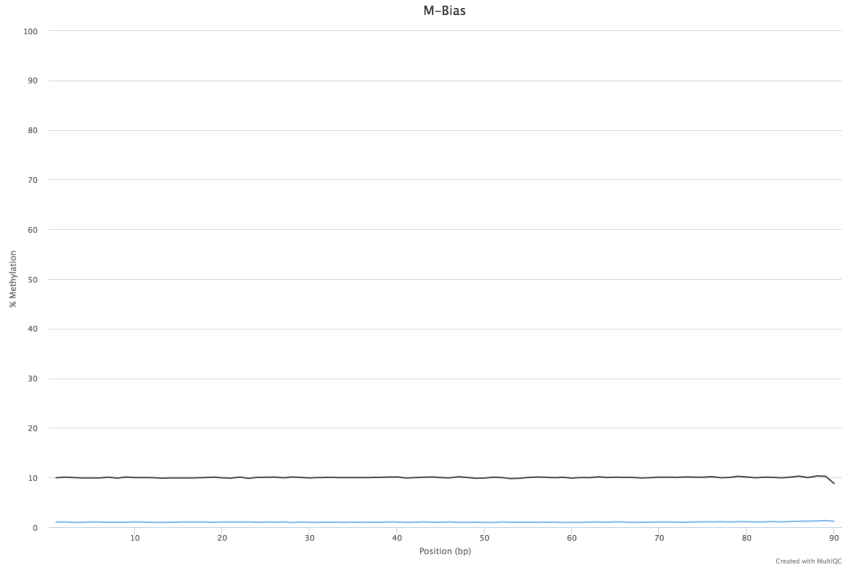
MultiQC ::: Bismark

Includes mkbs & oxbs









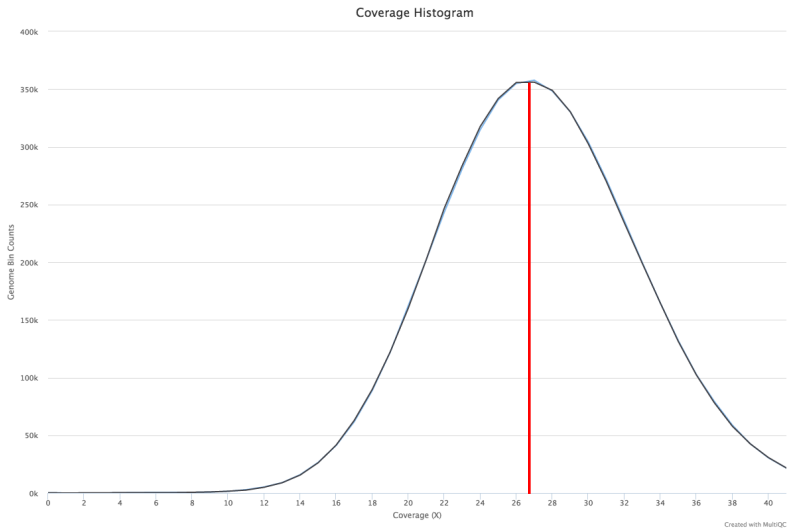
MultiQC ::: Qualimap

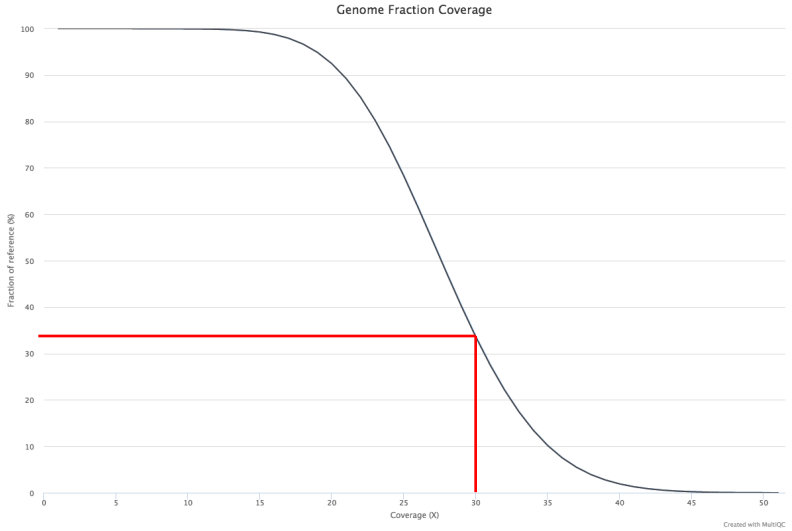
Includes mkbs & oxbs

Reads simulated for ~30X coverage

After trimming and alignment

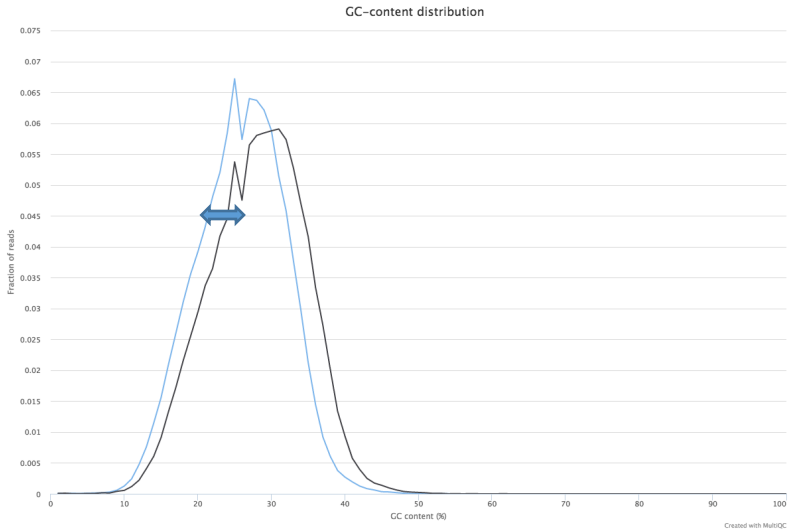
- *Average 27X*
- *~35% genome covered at 30X*

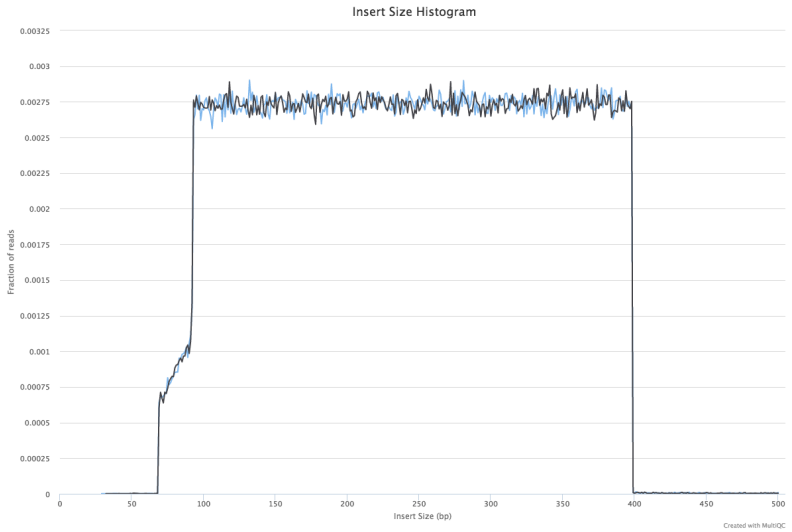




mkbs / oxbs difference = hmC

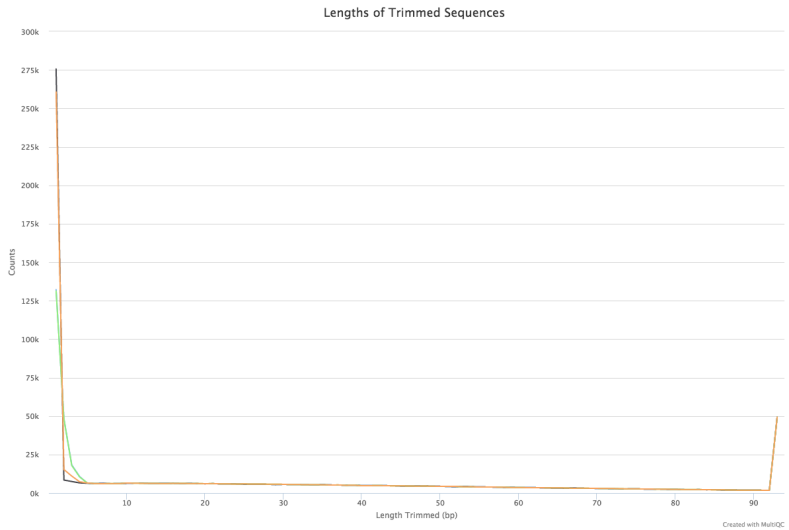
(In oxbs hmC oxidised to fC sequences as C, therefore more C expected in oxbs)



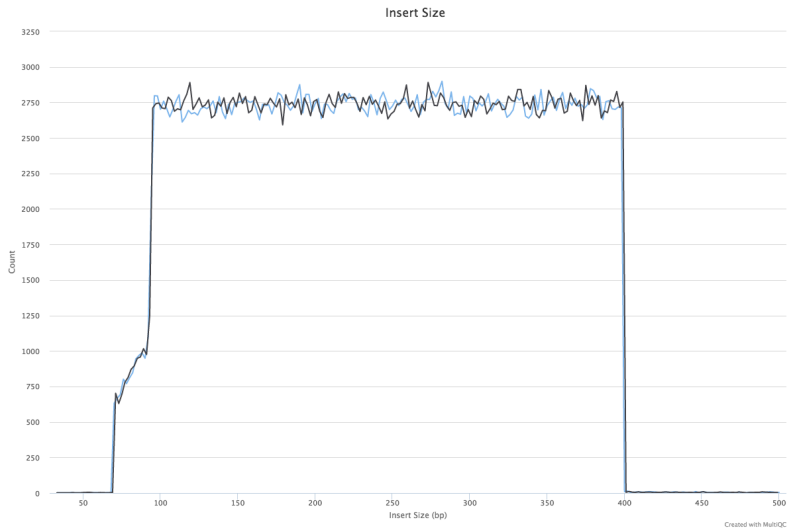


MultiQC ::: Cutadapt & Picard

Picard Insert Size Metrics



Cutadapt



Methyl-Kit

Methyl-Kit	R package for DNA methylation analysis
Version	v0.99.2
Download	https://github.com/al2na/methylKit

Terminal:

Custom R-script

```
$ Rscript ngschool.methylkit.R
```

Output:

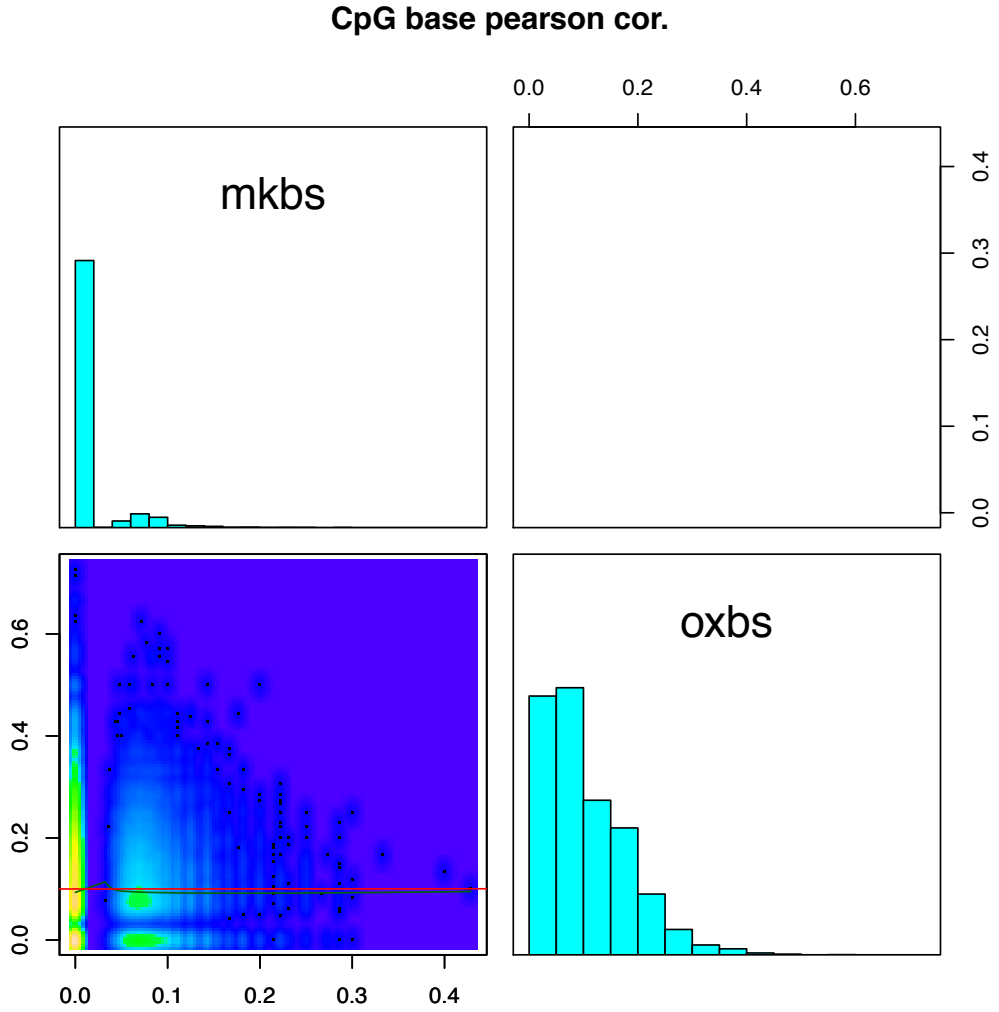
Plots

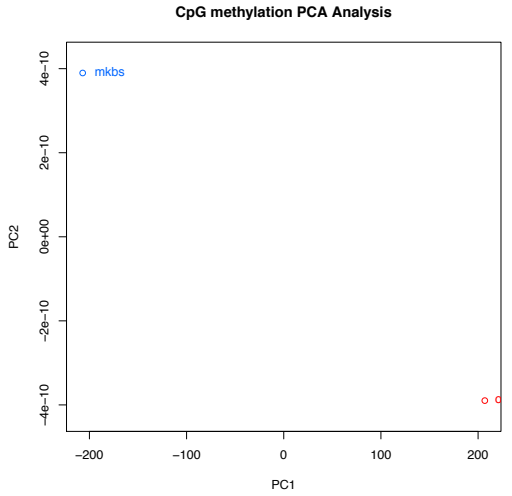
```
NGSchool.eu.methylkit.PCASamples.ward_corr_plot.pdf
NGSchool.eu.methylkit.CorrelationPlot.pdf
NGSchool.eu.methylkit.PCASamples.screeplot.pdf
NGSchool.eu.methylkit.PCASamples.pdf
NGSchool.eu.methylkit.diffMethPerChr.pdf
```

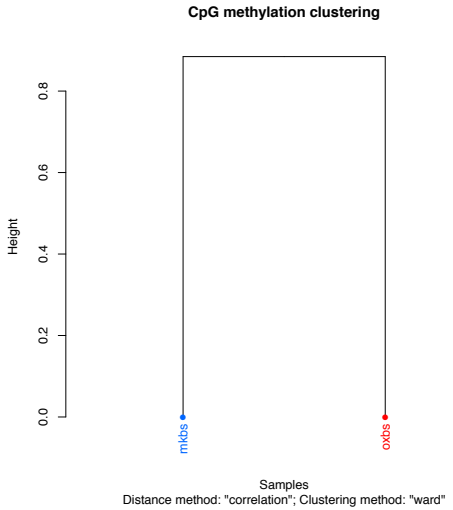
Tables

```
NGSchool.eu.methylkit.hyper_methylated.tsv
NGSchool.eu.methylkit.DiffMeth.tsv
NGSchool.eu.methylkit.hypo_methylated.tsv
NGSchool.eu.methylkit.differentialy_methylated.tsv
```

Methyl-Kit







Note: This is simulated data so biologically meaningless!

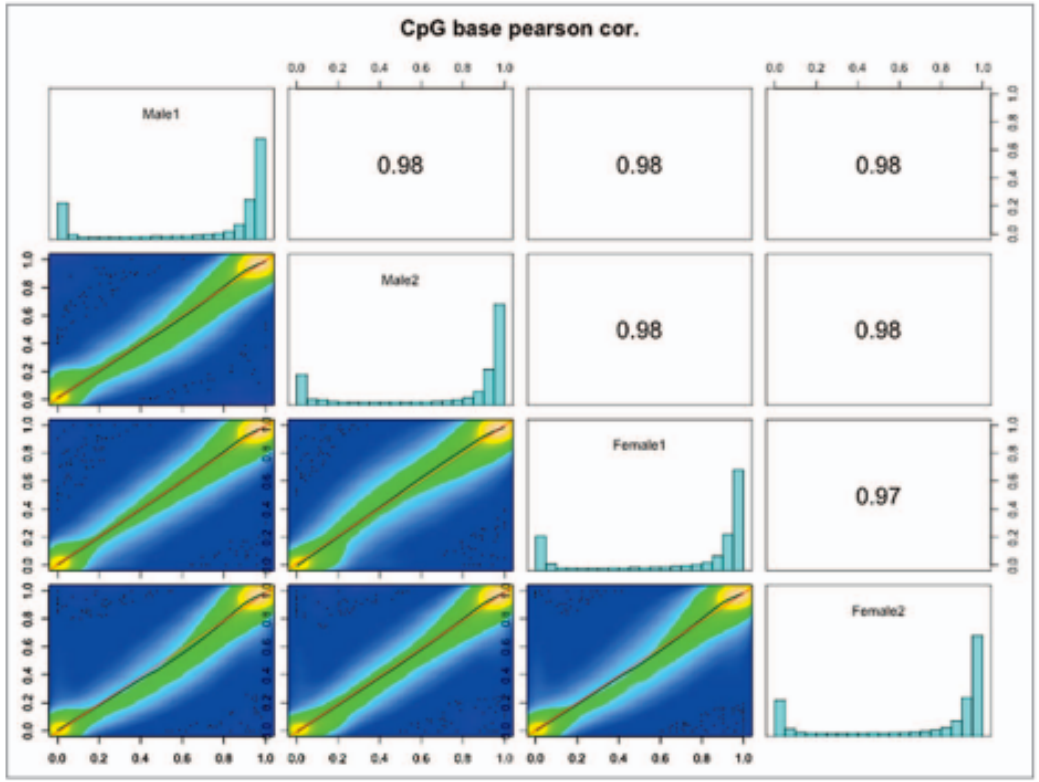
MethylKit ::: Example Data

RESEARCH PAPER

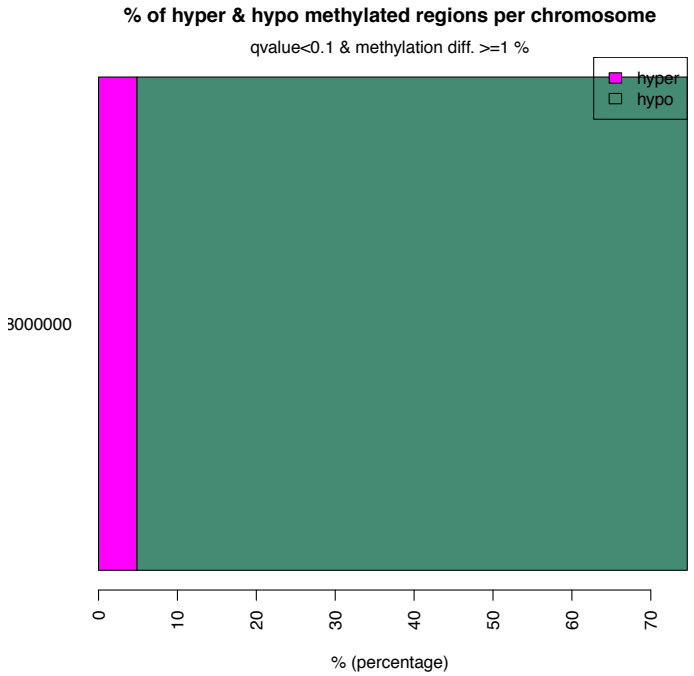
Epigenetics 8:9, 979–989; September 2013; © 2013 Landes Bioscience

Mapping the zebrafish brain methylome using reduced representation bisulfite sequencing

Aniruddha Chatterjee[1,2,*], Yuichi Ozaki[3], Peter A Stockwell[4,5], Julia A Horsfield[1,2], Ian M Morison[1,2], and Shinichi Nakagawa[2,3]



Comparing Samples / DMR Calling



Top 5 by location out of 122K [NGSchool.eu.methylkit.hypo_methylated.txt]

chr	start	end	strand	pvalue	qvalue	meth.diff
1:3000000-8000000	505	505	*	0.1586	0.0376	13.3333
1:3000000-8000000	794	794	*	0.3040	0.0400	8.3333
1:3000000-8000000	1058	1058	*	0.1974	0.0376	11.1111
1:3000000-8000000	1476	1476	*	0.3003	0.0397	6.2500
1:3000000-8000000	1811	1811	*	0.8925	0.0832	1.6667

Top 5 by location out of 8K [NGSchool.eu.methylkit.hyper_methylated.txt]

chr	start	end	strand	pvalue	qvalue	meth.diff
1:3000000-8000000	38	387	*	0.0821	0.0346	-22.2222
1:3000000-8000000	416	416	*	0.1526	0.0376	-15.3846
1:3000000-8000000	417	417	*	0.1667	0.0376	-12.5000
1:3000000-8000000	448	448	*	0.0255	0.0346	-33.3333
1:3000000-8000000	490	490	*	0.1736	0.0376	-11.1111

Note: This is simulated data so biologically meaningless!

IGV

Load the reference genome (only the 5Mb region)

```
IGV
└── Genomes
    └── Create .genome File...
        └── Select
            NGSchool_GRCh38_Chr1_region/Homo_sapiens.GRCh38.dna.chromosome.1.region30000000-5000000.fa
```

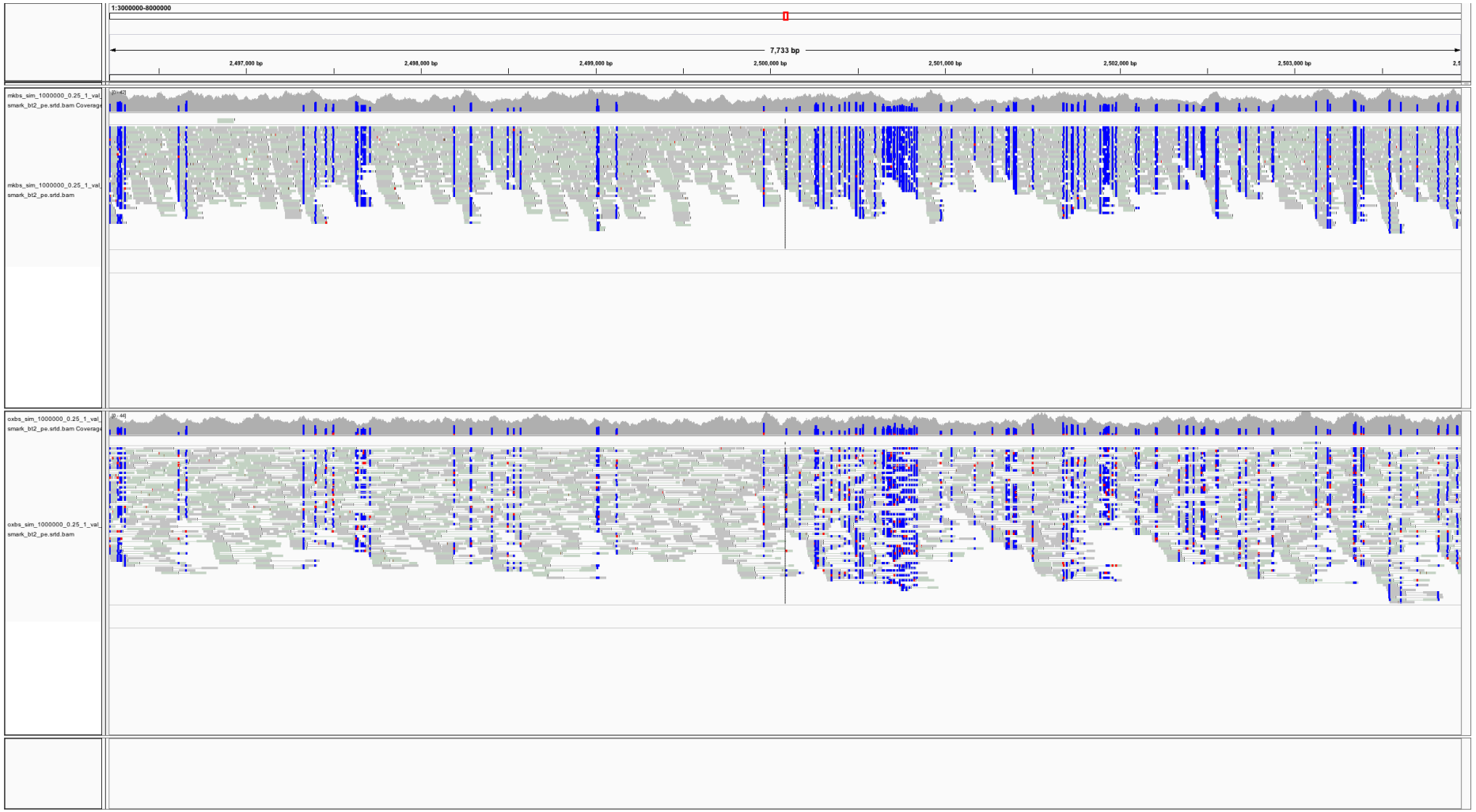
Load the aligned reads BAM files:

```
IGV
└── File
    └── Load From File...
        SimulatedData /mkbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam
        SimulatedData /oxbs_sim_1000000_0.25_1_val_1_bismark_bt2_pe.srtd.bam
```

IGV

mkbs

oxbs



Future of Single Base Resolution Methylation

Oxford Nanopore Technologies

Angewandte Communications

VIP Epigenetic Markers

DOI: 10.1002/anie.201300413

Single-Molecule Detection of 5-Hydroxymethylcytosine in DNA through Chemical Modification and Nanopore Analysis**

*Wen-Wu Li, Lingzhi Gong, and Hagan Bayley**

Cytosine Variant Calling with High-throughput Nanopore Sequencing

Arthur C. Rand*, Miten Jain*, Jordan Eizenga*, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson and Benedict Paten

Department of Biomolecular Engineering, University of California, Santa Cruz.

Genomics Institute, University of California, Santa Cruz.

*These authors contributed equally to this work.

Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer

Jared T. Simpson[1,2,*], Rachael Workman[3], P.C. Zuzarte[1], Matei David[1], L. J. Dursi[1], Winston Timp[3,*]

Affiliations:

1 Ontario Institute for Cancer Research, Toronto, Canada
2 Department of Computer Science, University of Toronto, Toronto, Canada
3 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland

Correspondence to:

jared.simpson@oicr.on.ca
wtimp@jhu.edu

Dr Russell S. Hamilton

Email: rsh46@cam.ac.uk

Web: http://www.trophoblast.cam.ac.uk/directory/Russell-Hamilton

UNIVERSITY OF CAMBRIDGE

Department of Physiology, Development and Neuroscience