WGBS data processing pipeline

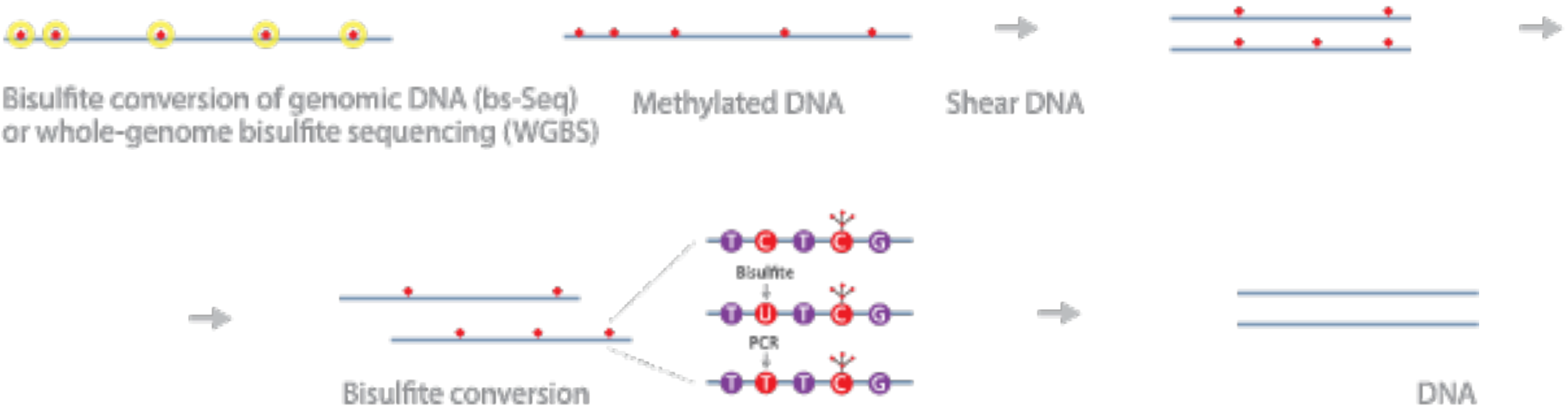
Hyung Joo Lee

Ting Wang Lab

May 26, 2021

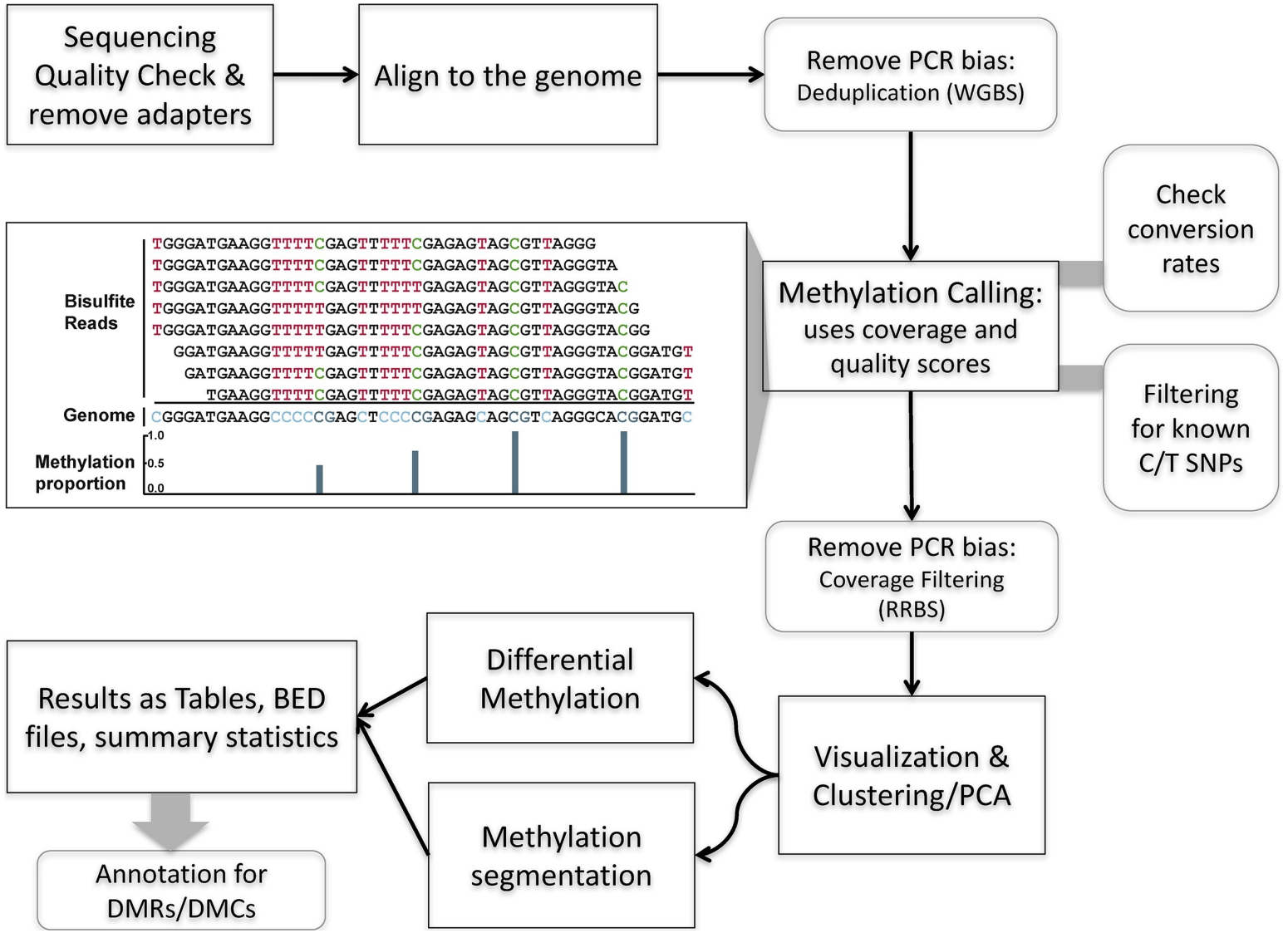
What is whole genome bisulfite genome sequencing?

BS-seq, Bisulfite-seq, WGBS,
MethylC-seq



From Illumina website

Overview of WGBS data processing and analysis



Wreczycka et al. 2017

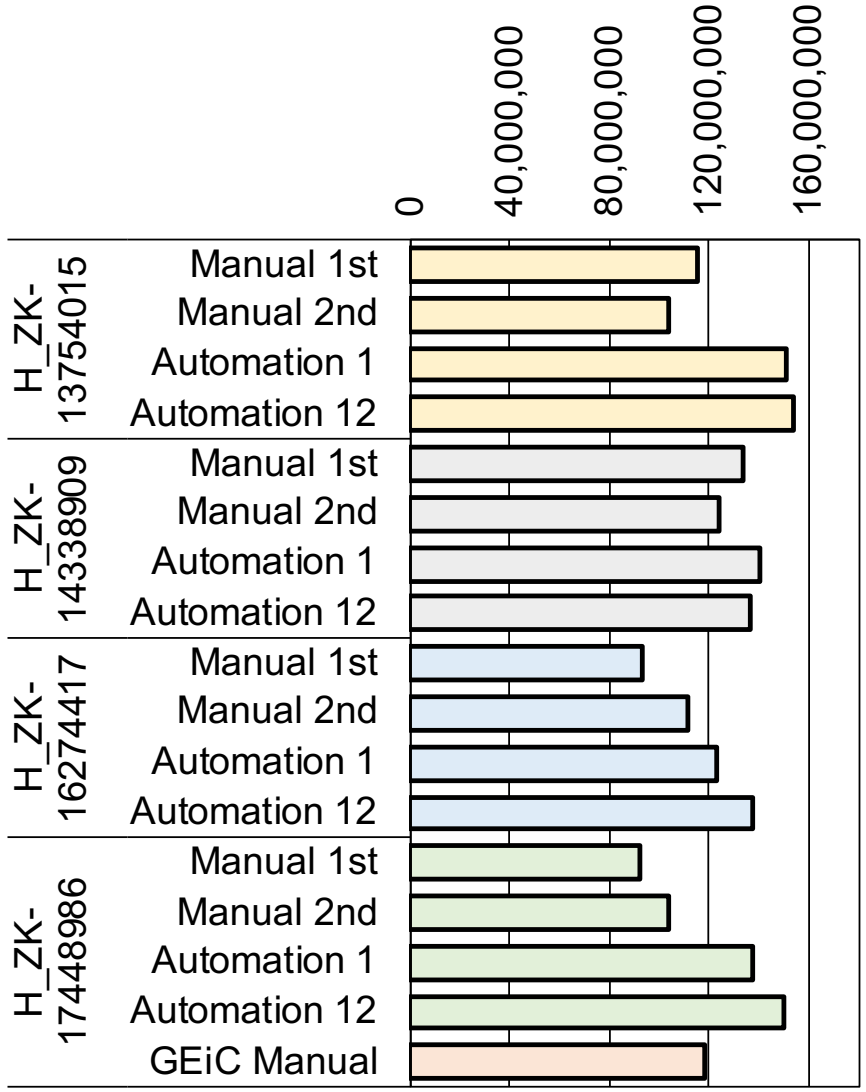
Toy dataset: Long Life Family Study (LLFS) WGBS data

- Human blood samples
- Swift Accel-NGS Bisulfite DNA Library Kit
- PE (2x151) sequencing (NovaSeq)
- 4 DNA samples
  - Manual library preparation vs automation
  - 2 replicates per preparation

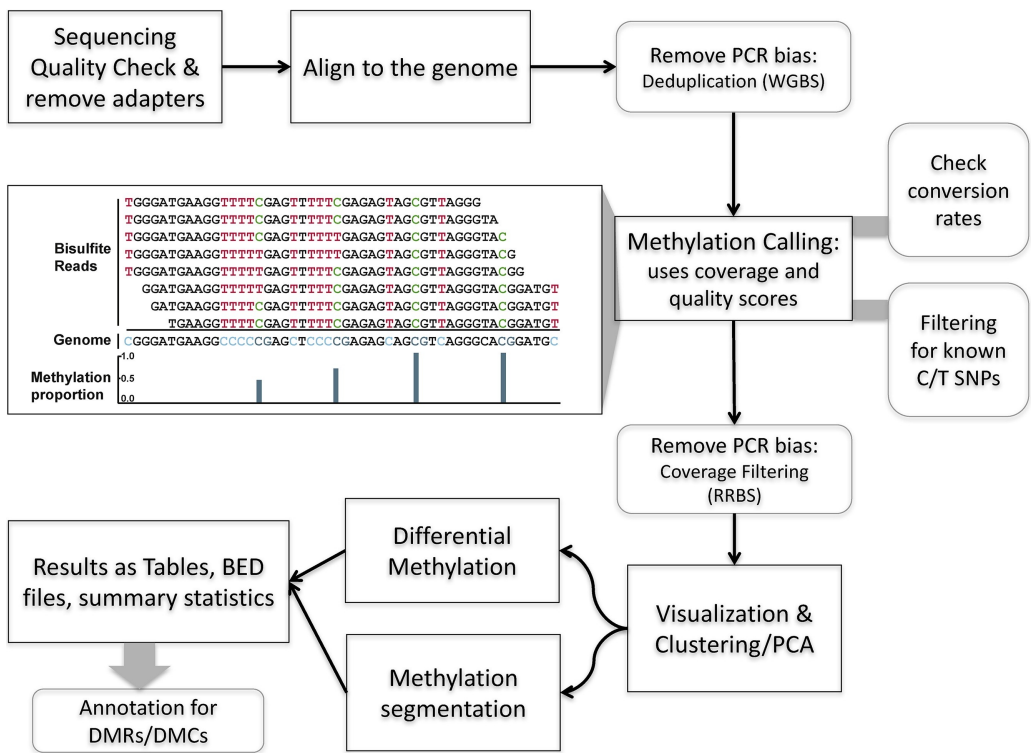
- Let's play with 100K pairs of one sample (0.1%)
  - `/scratch/wgbs/0_fastq/toy_R[12].fq.gz`

Scripts: https://wangftp.wustl.edu/~hlee/wgbs_scripts/
Data and Results: https://wangftp.wustl.edu/~hlee/wgbs/

Full data and results: https://htcf.wustl.edu/files/nXVYkoeO



Overview of WGBS data processing and analysis



Wreczycka et al. 2017

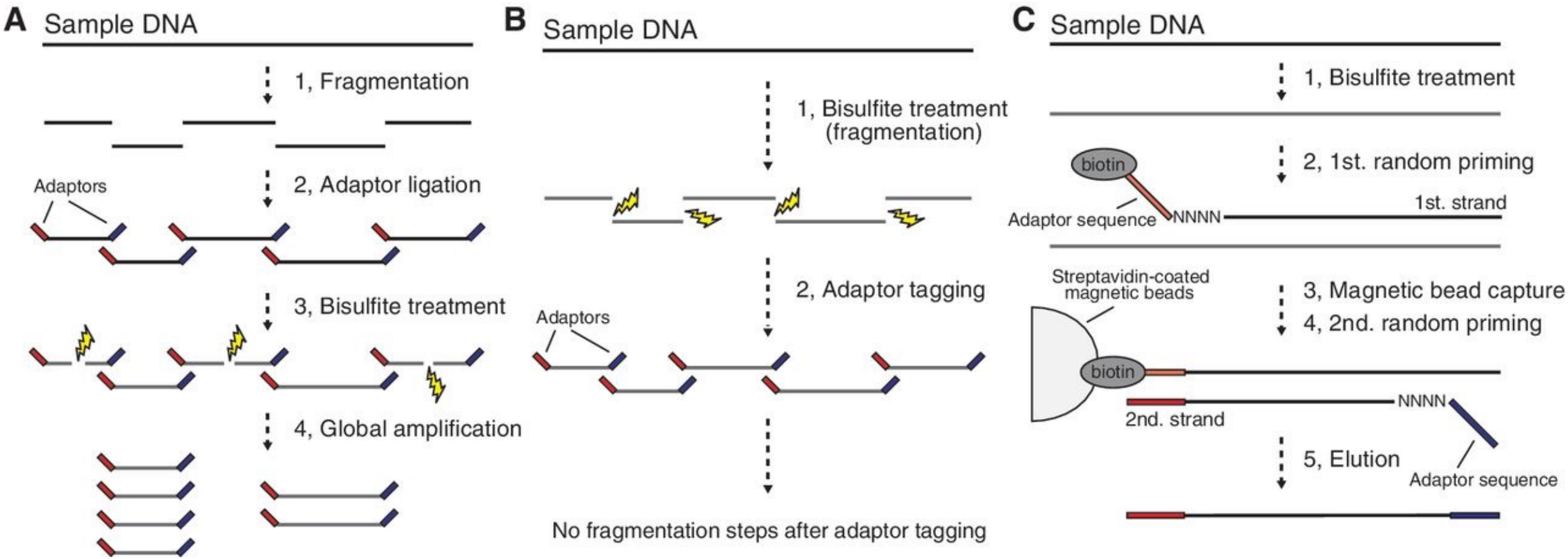
0. Sequencing quality check: FastQC v.0.11.5
1. Adapter trimming: Trim Galore! v.0.6.6
1. phiX contamination test: bwa
2. Read alignment: bismark v.0.18.1
   a. lambda genome (bisulfite conversion rate)
   b. hg38 genome
   c. PCR duplicate removal
   d. methylation extraction
   e. C coverage and methylation levels
3. Quality assurance
   a. Library complexity: preseq v.3.1.2
   b. Insertion size estimation
   c. GC-bias and GC content
   d. Methylation level check (CG/CH)

Which preparation method was used for your library?

Traditional WGBS

PBAT
(post-bisulfite adaptor tagging)

Random-priming
mediated PBAT



Miura et al. 2012

Bismark recommendations for different kits

Technique	5' Trimming	3' Trimming	Mapping	Deduplication	Extraction
BS-Seq	⬜	⬜	⬜	✅	`--ignore_r2 2`
RRBS	`--rrbs` (R2 only)	`--rrbs` (R1 only)	⬜	❌	⬜
RRBS (NuGEN Ovation)	special processing	special processing	⬜	❌	`--ignore_r2 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`	✅	⬜

- ⬜ - Default settings (nothing in particular is required
- ✅ - Yes, please!
- ❌ - No, absolutely not!

The random priming of post-bisulfite methods (such as PBAT, scBS-Seq, EpiGnome, Pico Methyl, Accel etc.) introduces errors, indels and methylation biases that may detrimentally affect your mapping efficiencies and methylation calls. The Accel-NGS Methyl-Seq protocol uses **Adaptase technology** for capturing single-stranded DNA in an unbiased (again, not that unbiased actually...) manner. Also here, the first ~10 bp show extreme biases in sequence composition and M-bias, so trimming off at *least* 10 bp is advisable (please check the M-bias plot if even more is needed).

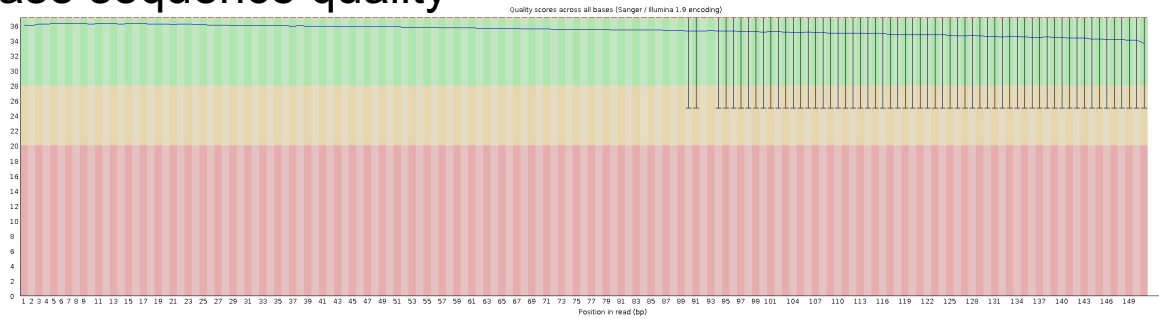
https://github.com/FelixKrueger/Bismark/tree/master/Docs

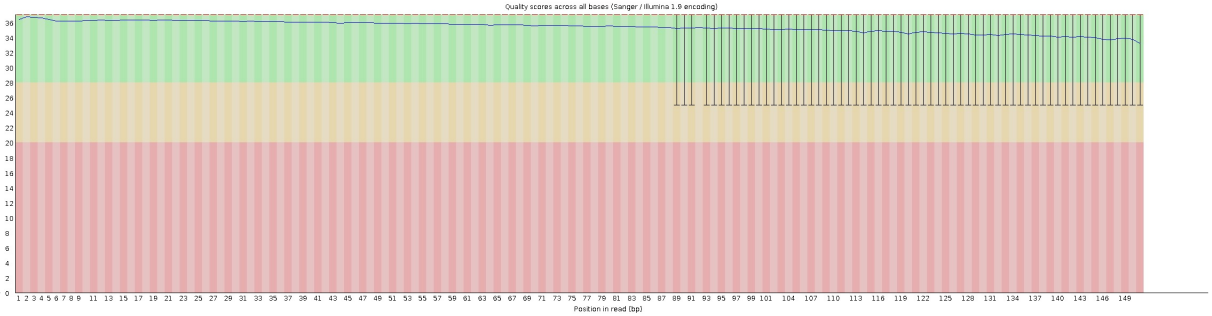
Quality check of fastq files

Read 1

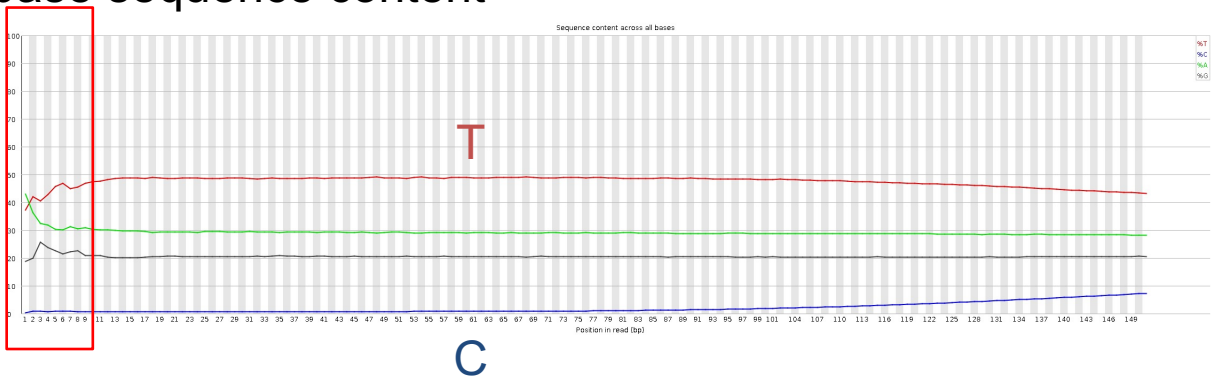
Read 2

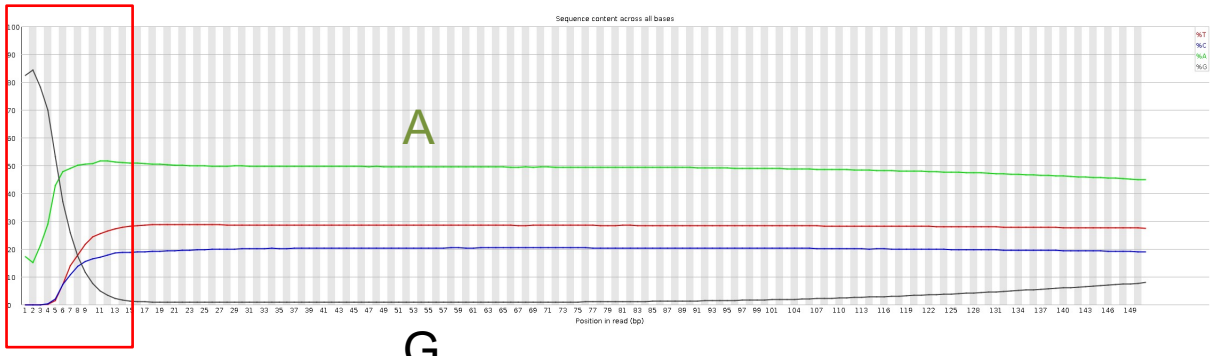
Per base sequence quality





Per base sequence content





Read 1: C>T converted
Read 2: G>A converted

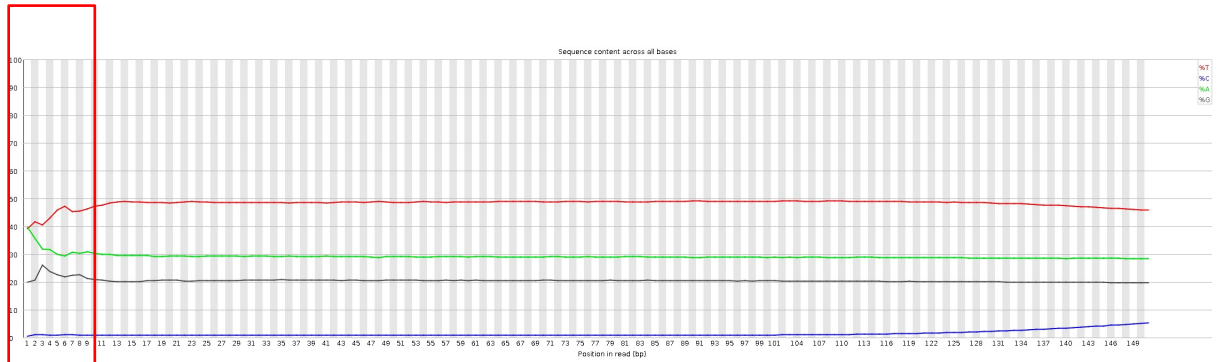
Quality check of fastq files

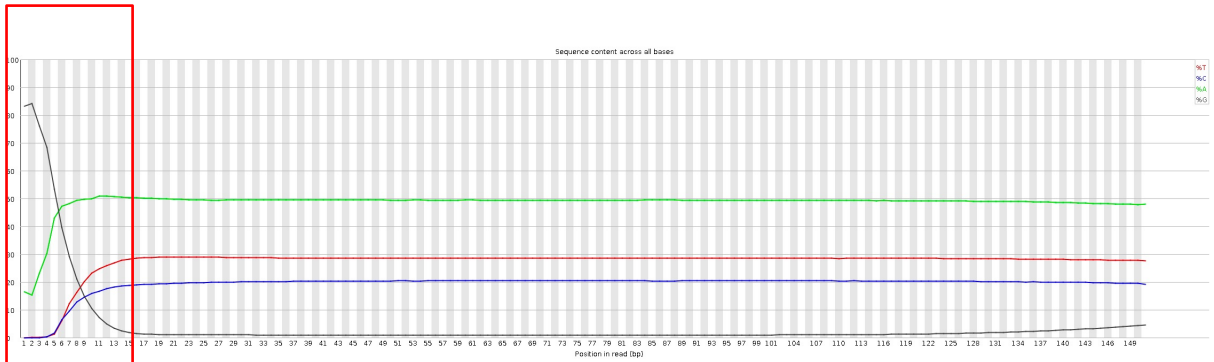
Per base sequence content

Read 1

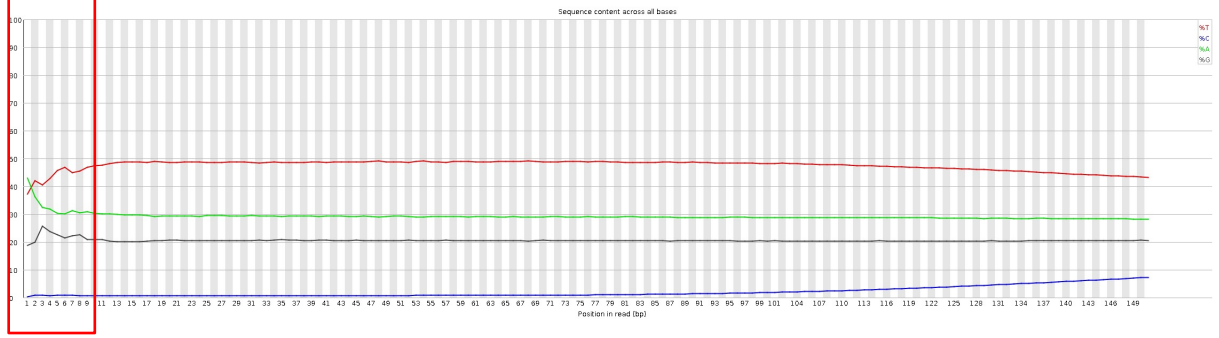
Read 2

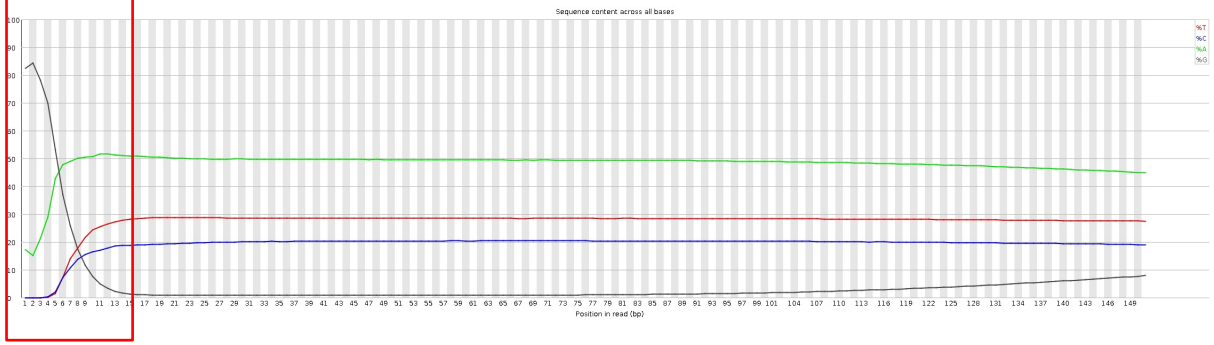
GEiC_WGBS_A1





H_ZK-14338909_auto_rep2





Typical feature of Swift Accel-NGS Bisulfite DNA Library Kit
5’ trimming is required due to the bias

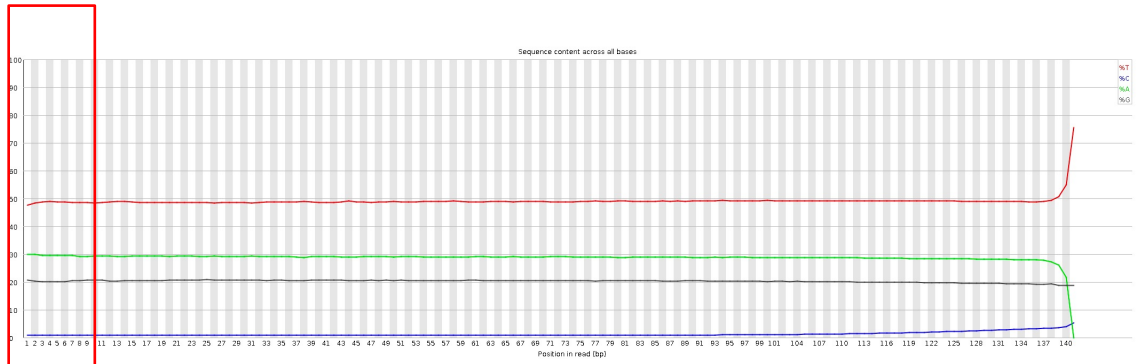
Quality check of fastq files after 5’ trimming

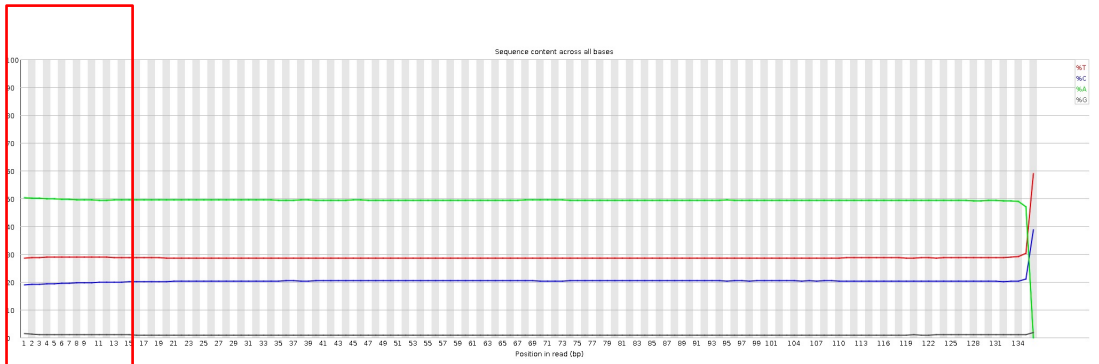
Per base sequence content

Read 1

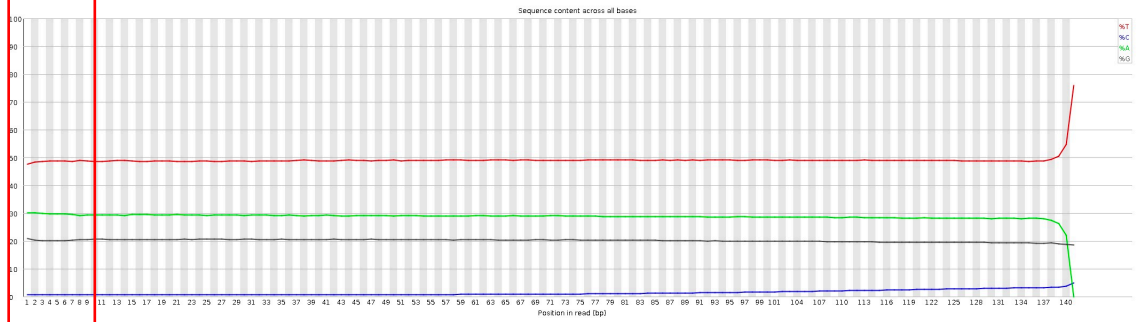
Read 2

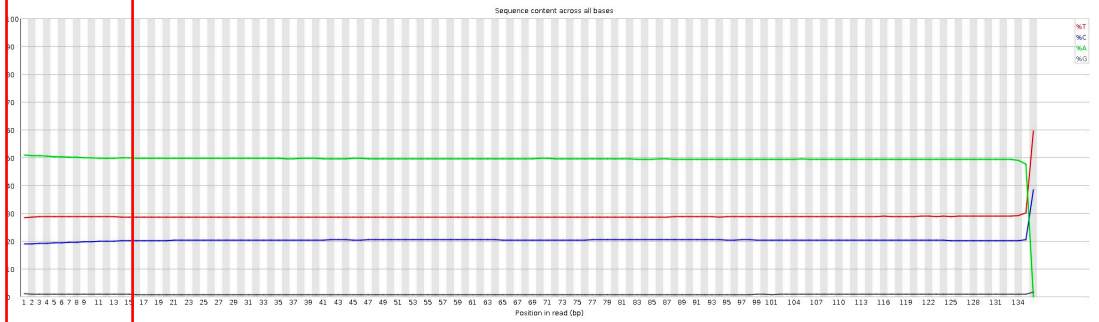
GEiC_WGBS_A1





H_ZK-14338909_auto_rep2





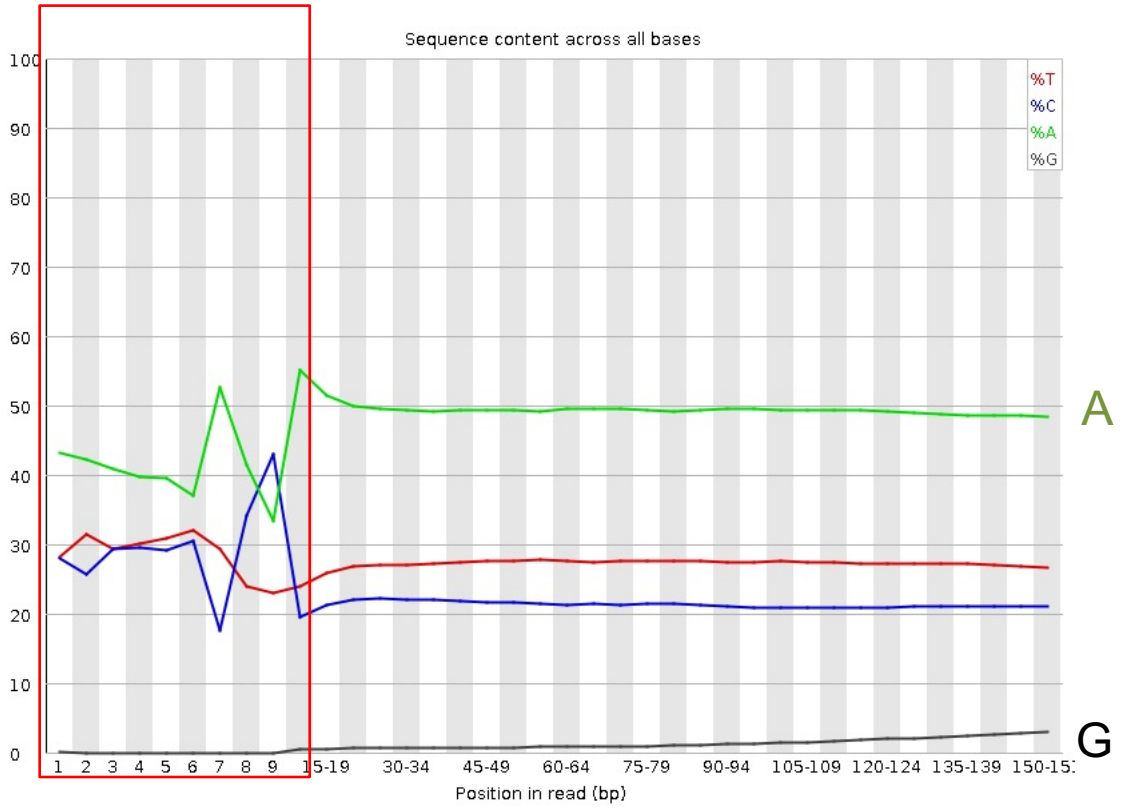
5’ bases + 3’ adapter trimming

Read 1: 10 bases, Read 2: 15 bases

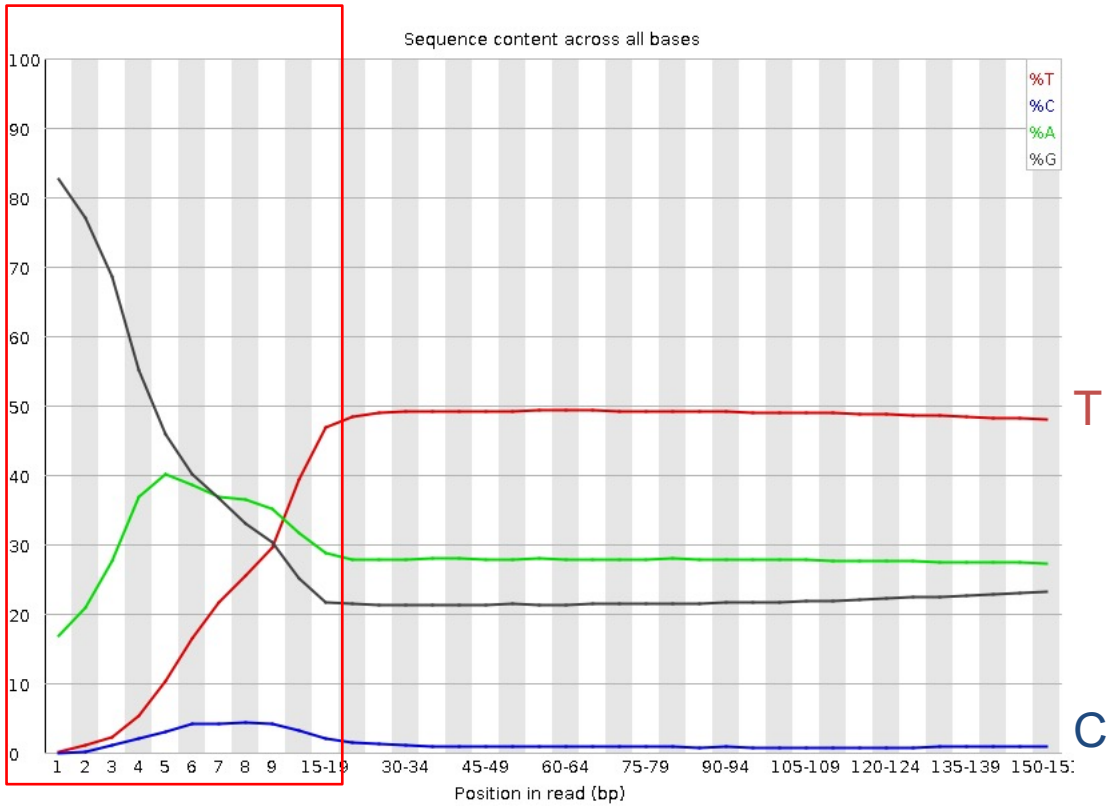
Quality check of PBAT libraries

Per base sequence content

Read 1



Read 2



Read 1: G>A converted
Read 2: C>T converted

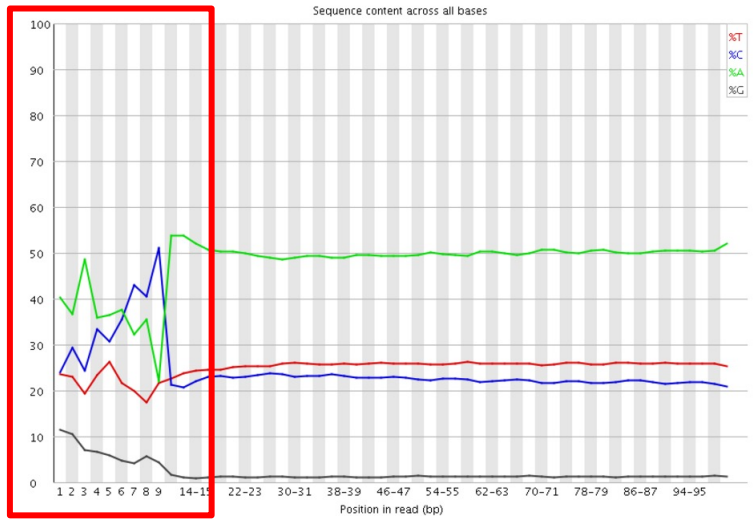
Susan Fisher lab (by courtesy of Jessica)

Mispriming in library preparation causes bias

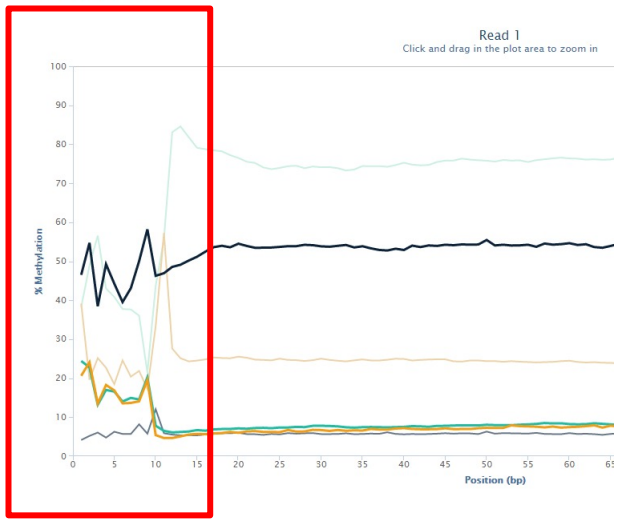
Mispriming in PBAT libraries causes methylation bias and poor mapping efficiencies

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.

Per base sequence content



M-Bias Plot



PBAT: Post-Bisulfite Adapter Tagging library

"We never really got to root cause of the problem but were convinced that this has to be a technical artefact rather than a biological effect. We suggested to simply remove the first bases either by 5′ trimming the raw sequences, or by ignoring the first N bases during the methylation extraction."

- Felix Krueger (Bismark author)

https://sequencing.qcfail.com/articles/mispriming-in-pbat-libraries-causes-methylation-bias-and-poor-mapping-efficiencies/

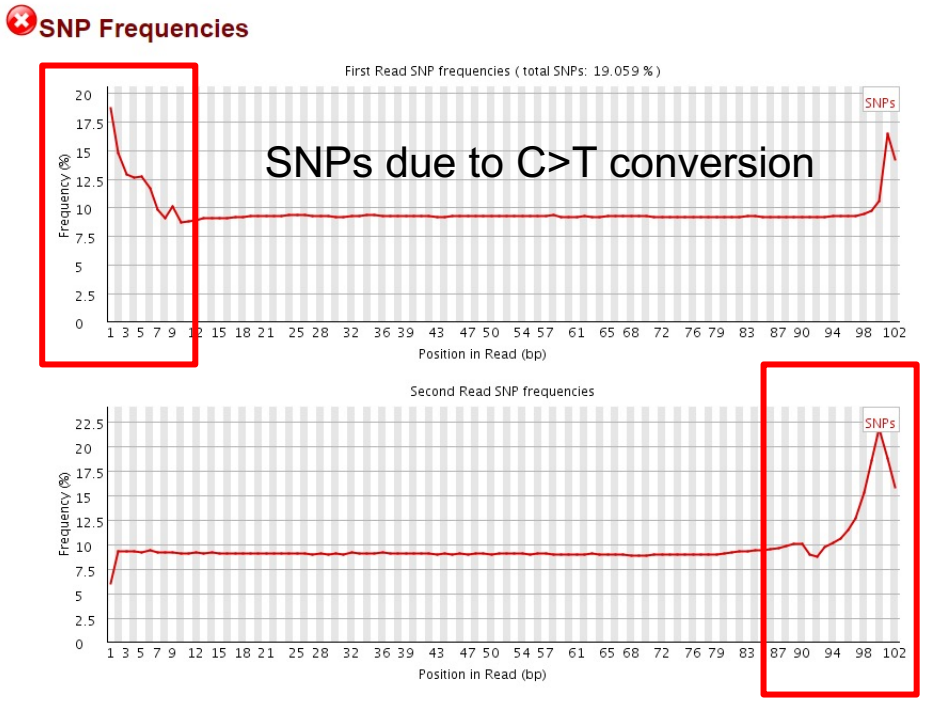
Mispriming in library preparation causes bias

Mispriming in PBAT libraries causes methylation bias and poor mapping efficiencies

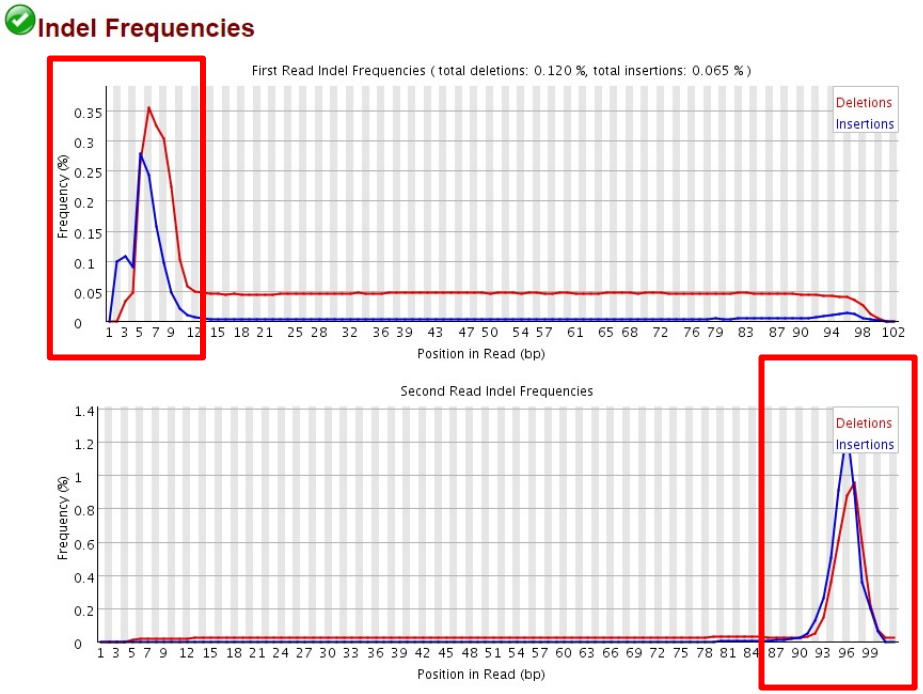
Random priming in PBAT libraries introduces drastic biases in the base composition and methylation levels especially at the 5' end of all reads. As a result, affected bases should be removed from the libraries before the alignment step.

"There are weird things going on during the "random priming" step of the PBAT protocol, resulting in a drastically increased rate of SNPs and InDels."
- Felix Krueger (Bismark author)

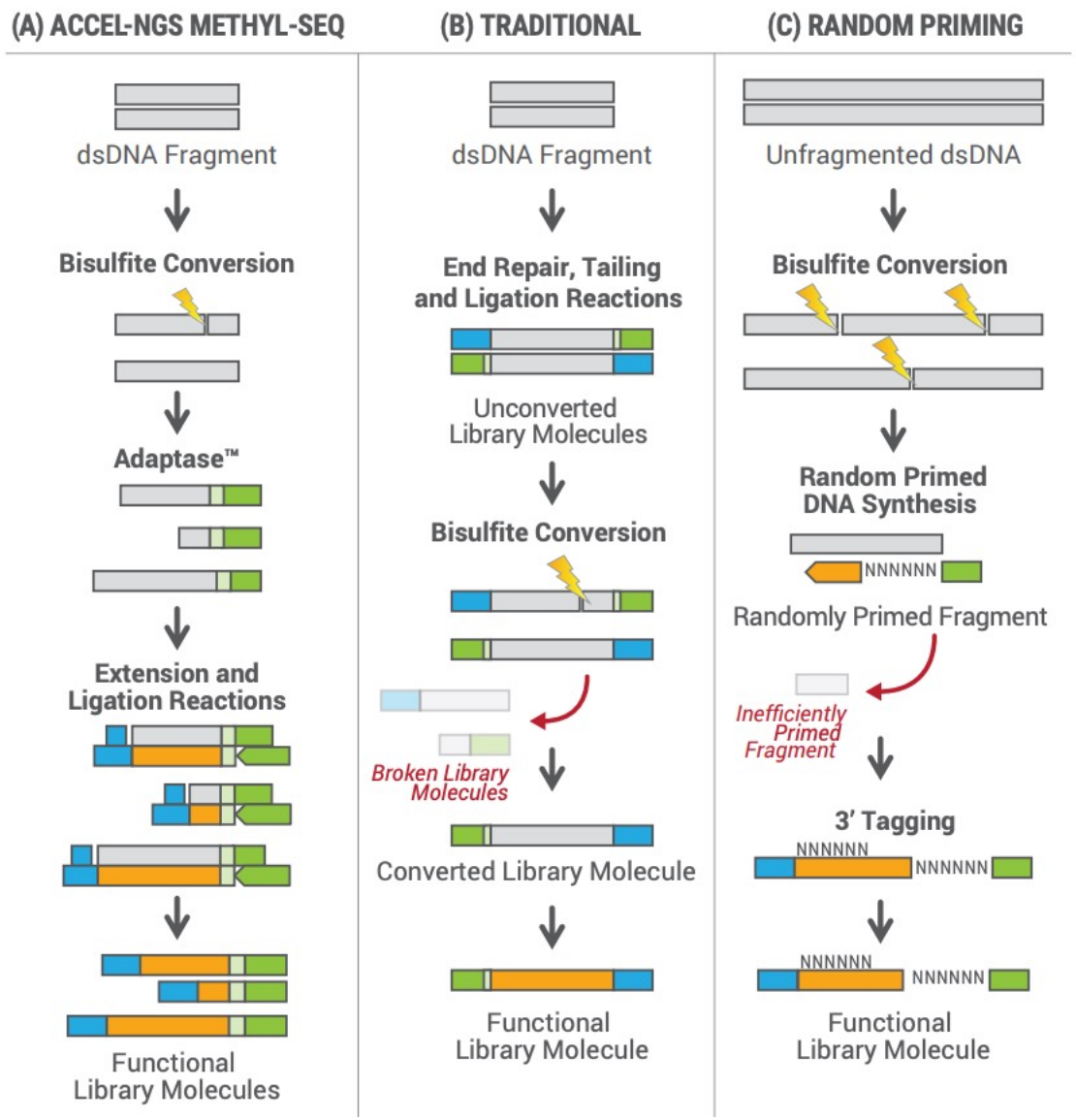
Increased SNP frequencies at the 5' end



Increased indel frequencies at the 5' end



Swift Accel-NGS Bisulfite DNA Library Kit



- Swift Biosciences' Adaptase® technology, used in the Accel-NGS® Methyl-Seq DNA Library Kit, adds a low complexity polynucleotide tail with an average length of 8 bases to the 3' end of each fragment during the addition of the first NGS adapter molecule. If these tails are not trimmed bioinformatically from the sequencing data, it is normal and expected to observe them at the beginning of Read 2 (R2). When read length is close to fragment size, the tail may also be observed toward the end of Read 1 (R1) data.
- The Accel-NGS Methyl-Seq Kit adds bases to 3' termini during the Adaptase tailing step, including unmethylated cytosines. This tail adds a synthetic sequence, adding methylation information to the dataset. Therefore, trimming is required for Accel-NGS Methyl-Seq libraries to obtain improved mapping efficiency (with tools like Bismark or BSMAP) and precise methylation information and bisulfite conversion efficiency.
- Many informatics pipelines already include trimming of up to 10 bases from the beginning of both R1 and R2 to eliminate any synthetic cytosine methylation introduced as a result of filling in overhangs during end repair steps of conventional dsDNA library preparation and low-quality bases due to bisulfite treatment.

- From the manufacturer's instruction

PhiX reads, lambda DNA, and bisulfite conversion rates

Library Name	Sample	Library prep	# Total read pairs	# Removed pairs trimming	# Trimmed read pairs	# PhiX reads	% PhiX reads	# Lambda read pairs	% Labmda read pairs	Bisulfite conversion rate
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	150,592,761	241,425	150,351,336	4	0.000001%	126,196	0.08%	98.9%
H_ZK-13754015-lib5		Automation 12	153,585,029	274,185	153,310,844	1	0.000000%	147,854	0.10%	98.8%
H_ZK-13754015-lib2		Manual 1st	115,448,443	146,571	115,301,872	0	0.000000%	176,520	0.15%	98.5%
H_ZK-13754015-lib3		Manual 2nd	103,826,384	425,707	103,400,677	0	0.000000%	172,693	0.17%	98.6%
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	140,562,154	354,915	140,207,239	7	0.000002%	183,966	0.13%	99.0%
H_ZK-14338909-lib5		Automation 12	136,853,914	234,610	136,619,304	3	0.000001%	144,407	0.11%	98.9%
H_ZK-14338909-lib2		Manual 1st	133,211,507	433,348	132,778,159	1	0.000000%	248,109	0.19%	98.5%
H_ZK-14338909-lib3		Manual 2nd	123,896,710	369,794	123,526,916	3	0.000001%	179,573	0.15%	98.7%
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	122,794,756	245,881	122,548,875	6	0.000002%	196,792	0.16%	98.9%
H_ZK-16274417-lib5		Automation 12	137,723,144	341,165	137,381,979	4	0.000001%	191,669	0.14%	98.9%
H_ZK-16274417-lib2		Manual 1st	93,256,705	193,786	93,062,919	3	0.000002%	195,522	0.21%	98.7%
H_ZK-16274417-lib3		Manual 2nd	111,316,902	277,457	111,039,445	0	0.000000%	271,866	0.24%	98.6%
H_ZK-17448986-lib4	H_ZK-17448986	Automation 1	137,698,365	351,875	137,346,490	4	0.000001%	316,975	0.23%	98.9%
H_ZK-17448986-lib5		Automation 12	150,437,804	310,606	150,127,198	7	0.000002%	277,401	0.18%	98.8%
H_ZK-17448986-lib2		Manual 1st	92,525,438	230,554	92,294,884	2	0.000001%	237,230	0.26%	98.7%
H_ZK-17448986-lib3		Manual 2nd	103,340,524	364,797	102,975,727	0	0.000000%	245,464	0.24%	98.6%
GEiC_WGBS_A1		GEiC Manual	118,539,862	56,839	118,483,023	0	0.000000%	49	0.00004%	99.0%

ENCODE WGBS standards (conversion rate)

Current Standards

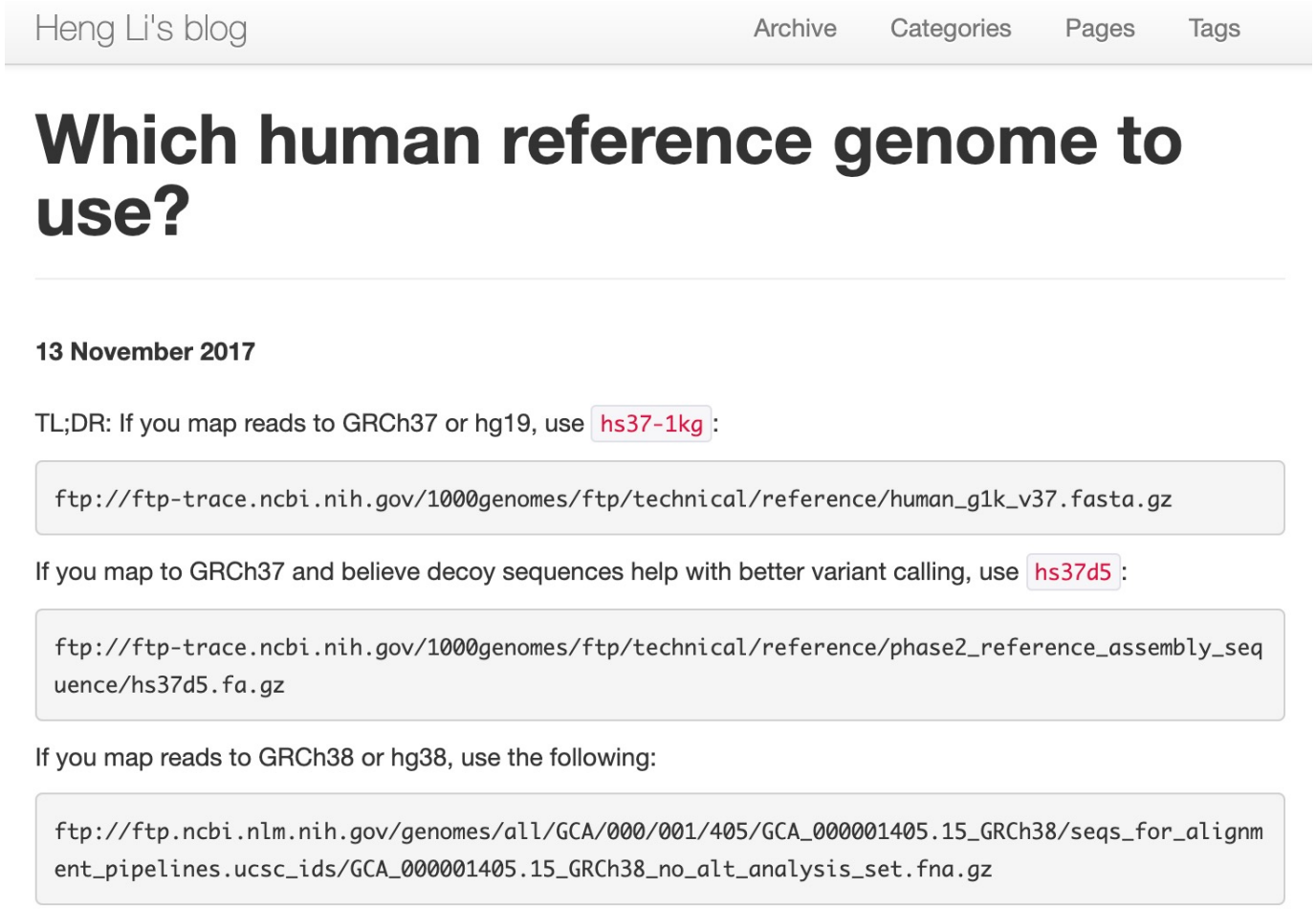
Experimental guidelines for WGBS experiments can be found here.

- Experiments should have two or more biological replicates; they may have two technical replicates per biological replicate. Assays performed using EN-TEx samples may be exempted due to limited availability of experimental material.
- The C to T conversion rate should be ≥98%
- The CpG quantification should have a Pearson correlation of ≥0.8 for sites with ≥10X coverage.
- Sequencing may be paired- or single-ended, as long as sequencing type is specified and paired sequences are indicated.
- The experiment must pass routine metadata audits in order to be released.

									mda airs	Bisulfite conversion rate
H_									0.08%	98.9%
H_									0.10%	98.8%
H_									0.15%	98.5%
H_									0.17%	98.6%
H_									0.13%	99.0%
H_									0.11%	98.9%
H_									0.19%	98.5%
H_									0.15%	98.7%
H_									0.16%	98.9%
H_									0.14%	98.9%
H_									0.21%	98.7%
H_									0.24%	98.6%
H_ZK-17448986-lib1	H_ZK-17448986	Automation 1	150,437,804	310,606	150,127,198	7	0.000002%	277,401	0.23%	98.9%
H_ZK-17448986-lib5		Automation 12	150,437,804	310,606	150,127,198	7	0.000002%	277,401	0.18%	98.8%
H_ZK-17448986-lib2		Manual 1st	92,525,438	230,554	92,294,884	2	0.000001%	237,230	0.26%	98.7%
H_ZK-17448986-lib3		Manual 2nd	103,340,524	364,797	102,975,727	0	0.000000%	245,464	0.24%	98.6%
GEiC_WGBS_A1		GEiC Manual	150,437,804	56,839	150,380,965	0	0.000000%	49	0.00003%	99.0%

https://www.encodeproject.org/data-standards/wgbs/

Which reference genome do you have to use?



Heng Li's blog Archive Categories Pages Tags

Which human reference genome to use?

13 November 2017

TL;DR: If you map reads to GRCh37 or hg19, use `hs37-1kg`:

```
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz
```

If you map to GRCh37 and believe decoy sequences help with better variant calling, use `hs37d5`:

```
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
```

If you map reads to GRCh38 or hg38, use the following:

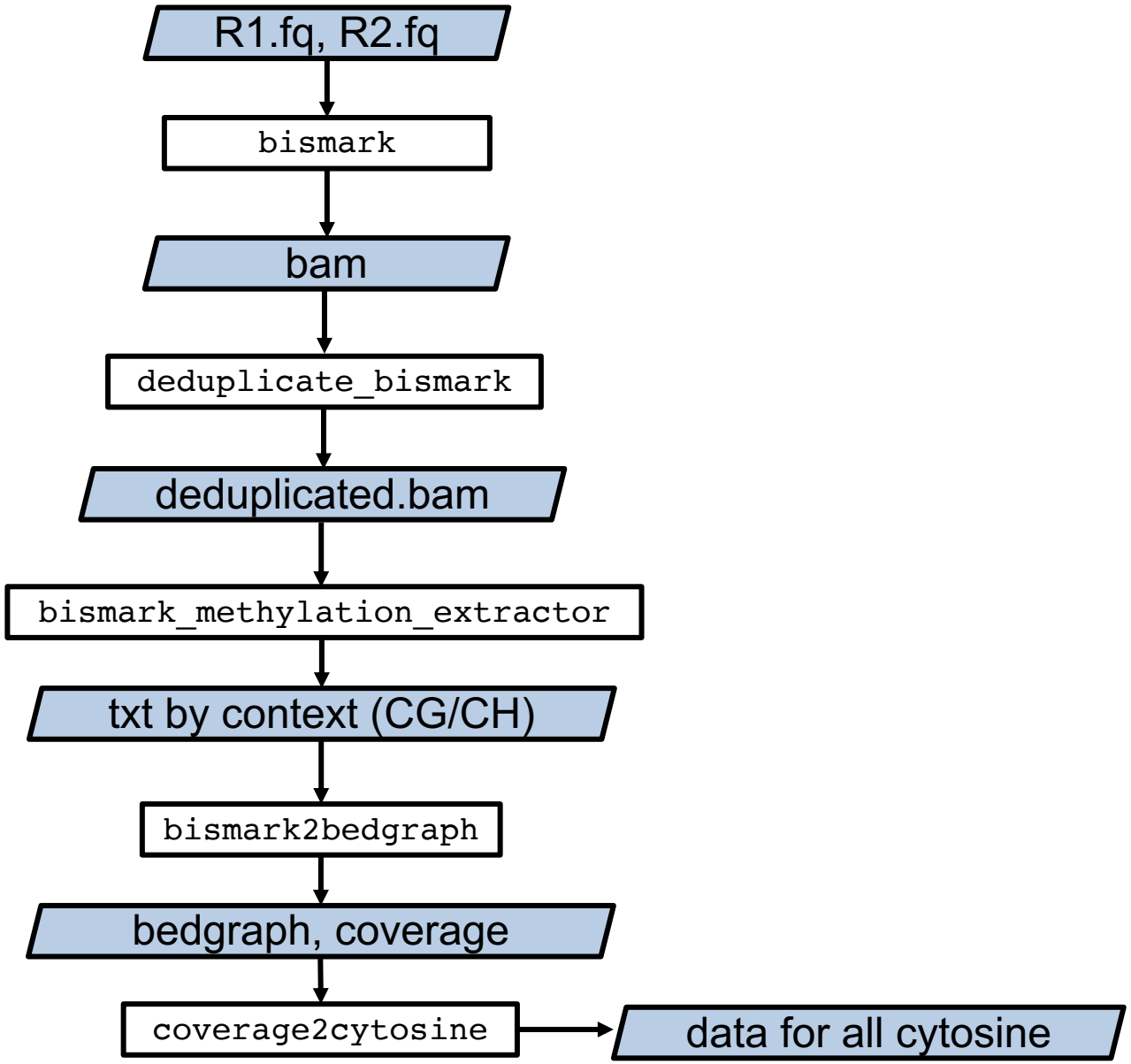
```
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz
```

https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use

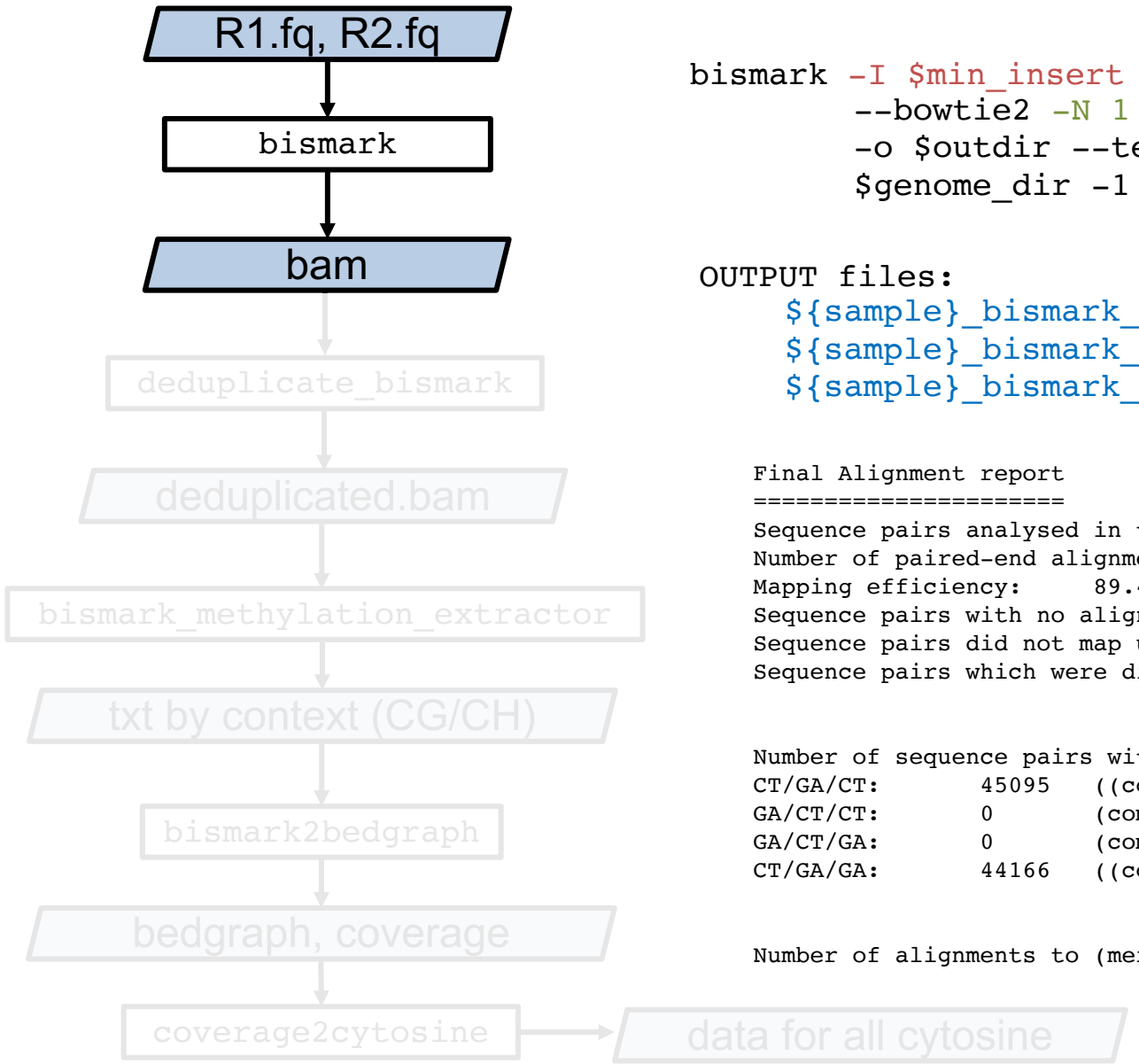
Potential issues to consider

1. Inclusion of ALT contigs
2. Padding ALT contigs with long “N”s
3. Inclusion of multi-placed sequences
4. Using accession numbers instead of chromosome names
5. Not including unplaced and unlocalized contigs

Bismark processes of WGBS data



Bismark processes of WGBS data



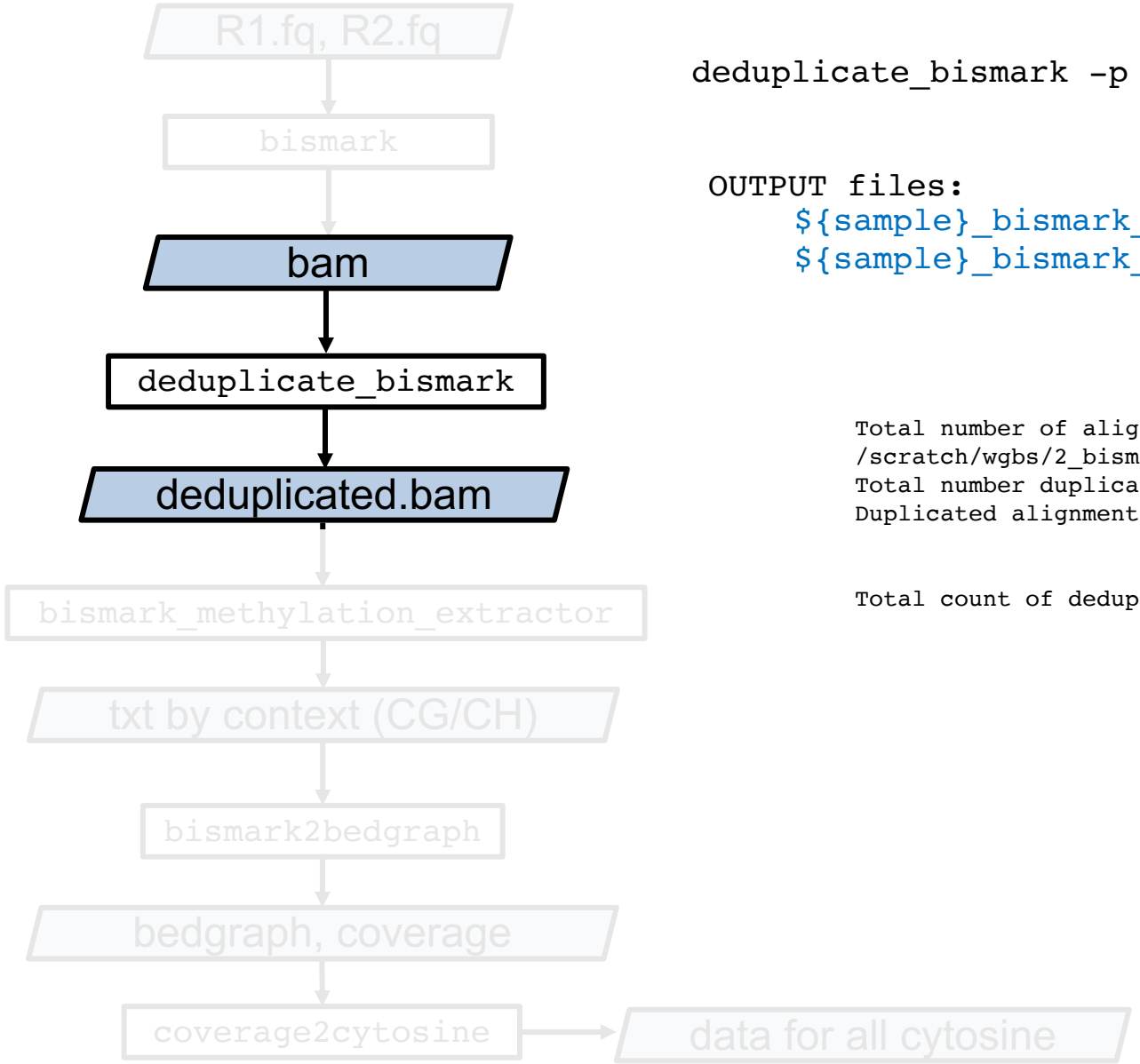

```
bismark -I $min_insert -X $max_insert --parallel 2 -p $cpus \
        --bowtie2 -N 1 -L 28 --score_min L,0,-0.6 \
        -o $outdir --temp_dir $tmp_dir --gzip (--nucleotide_coverage) \
        $genome_dir -1 $read1_fq -2 $read2_fq
```

OUTPUT files:

- ${sample}_bismark_bt2_pe.bam
- ${sample}_bismark_bt2_PE_report.txt
- ${sample}_bismark_bt2_pe.nucleotide_stats.txt (optional)

```
Final Alignment report
======================
Sequence pairs analysed in total:       99849
Number of paired-end alignments with a unique best hit: 89261
Mapping efficiency:     89.4%
Sequence pairs with no alignments under any condition:  4578
Sequence pairs did not map uniquely:    6010
Sequence pairs which were discarded because genomic sequence could not be extracted:    0


Number of sequence pairs with unique best (first) alignment came from the bowtie output:
CT/GA/CT:       45095   ((converted) top strand)
GA/CT/CT:       0       (complementary to (converted) top strand)
GA/CT/GA:       0       (complementary to (converted) bottom strand)
CT/GA/GA:       44166   ((converted) bottom strand)


Number of alignments to (merely theoretical) complementary strands being rejected in total:     0
```

Bismark processes of WGBS data

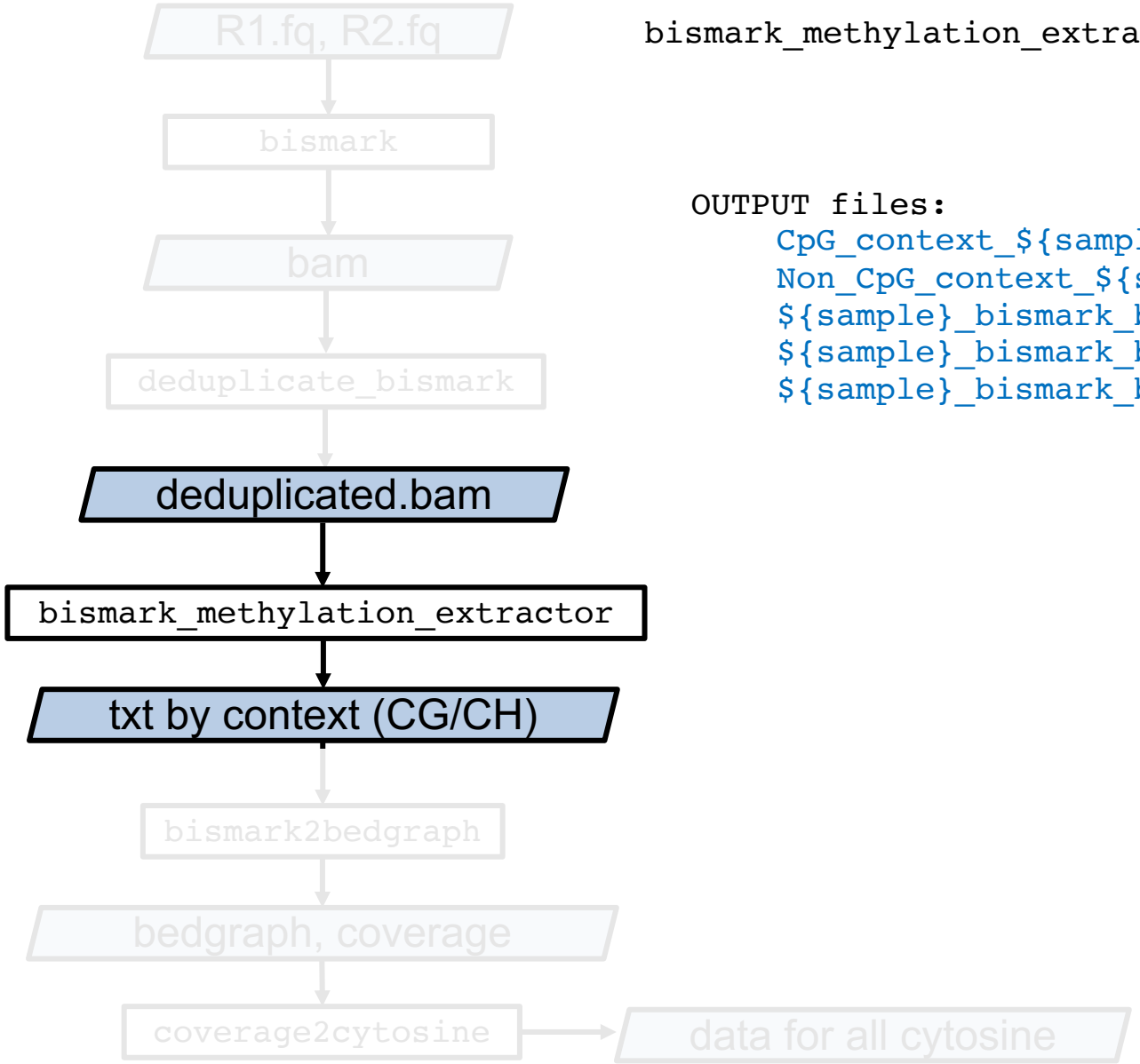



```
deduplicate_bismark -p --bam $bam_pe
```

```
OUTPUT files:
    ${sample}_bismark_bt2_pe.deduplicated.bam
    ${sample}_bismark_bt2_pe.deduplication_report.txt
```

```
Total number of alignments analysed in
/scratch/wgbs/2_bismark/toy_bismark_bt2_pe.bam:  89261
Total number duplicated alignments removed:     9169 (10.27%)
Duplicated alignments were found at:    8434 different position(s)


Total count of deduplicated leftover sequences: 80092 (89.73% of total)
```

Bismark processes of WGBS data



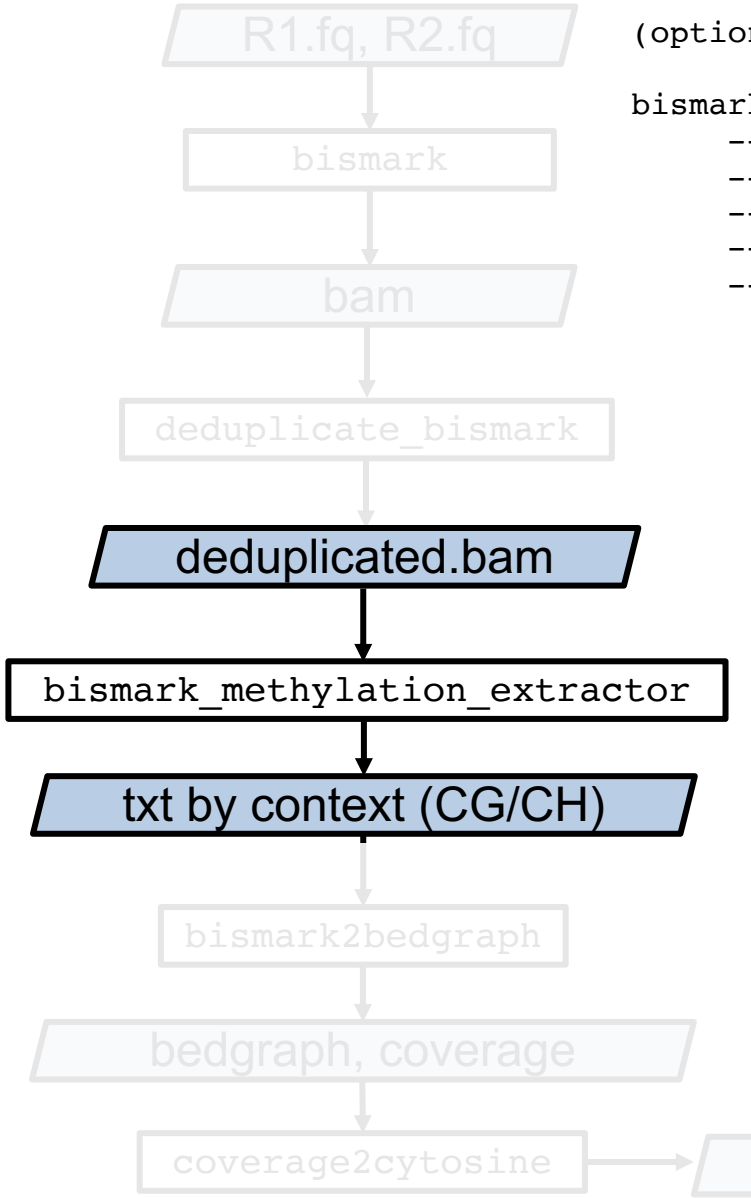

```
bismark_methylation_extractor --paired-end --no_overlap \
                              --comprehensive --merge_non_CpG --report \
                              -o $outdir --gzip --parallel $cpus \
                              $bam_dedup_pe
```

OUTPUT files:

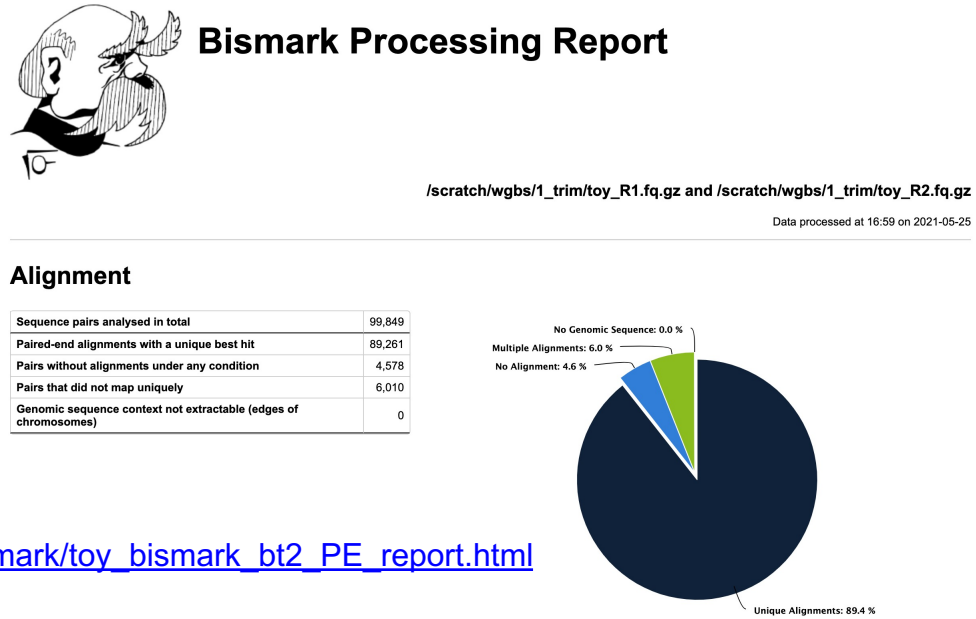
- CpG_context_${sample}_bismark_bt2_pe.deduplicated.txt.gz
- Non_CpG_context_${sample}_bismark_bt2_pe.deduplicated.txt.gz
- ${sample}_bismark_bt2_pe.deduplicated_splitting_report.txt
- ${sample}_bismark_bt2_pe.deduplicated.M-bias.txt
- ${sample}_bismark_bt2_pe.deduplicated.M-bias_R[12].png

```
Processed 80092 lines in total
Total number of methylation call strings processed: 160184


Final Cytosine Methylation Report
=================================
Total number of C's analysed:   3068976


Total methylated C's in CpG context:    114873
Total methylated C's in CHG context:    4906
Total methylated C's in CHH context:    18241


Total C to T conversions in CpG context:        28726
Total C to T conversions in CHG context:        656710
Total C to T conversions in CHH context:        2245520


C methylated in CpG context:    80.0%
C methylated in non-CpG context:        0.8%
```

Bismark processes of WGBS data




```
(optional)

bismark2report -o ${base}_bismark_bt2_PE_report.html --dir $outdir \
     --alignment_report ${outdir}/${base}_bismark_bt2_PE_report.txt \
     --dedup_report ${outdir}/${base}_bismark_bt2_pe.deduplication_report.txt \
     --splitting_report ${outdir}/${base}_bismark_bt2_pe.deduplicated_splitting_report.txt \
     --mbias_report ${outdir}/${base}_bismark_bt2_pe.deduplicated.M-bias.txt \
     --nucleotide_report ${outdir}/${base}_bismark_bt2_pe.nucleotide_stats.txt


  OUTPUT files:
      ${sample}_bismark_bt2_PE_report.html
```



Bismark Processing Report

/scratch/wgbs/1_trim/toy_R1.fq.gz and /scratch/wgbs/1_trim/toy_R2.fq.gz

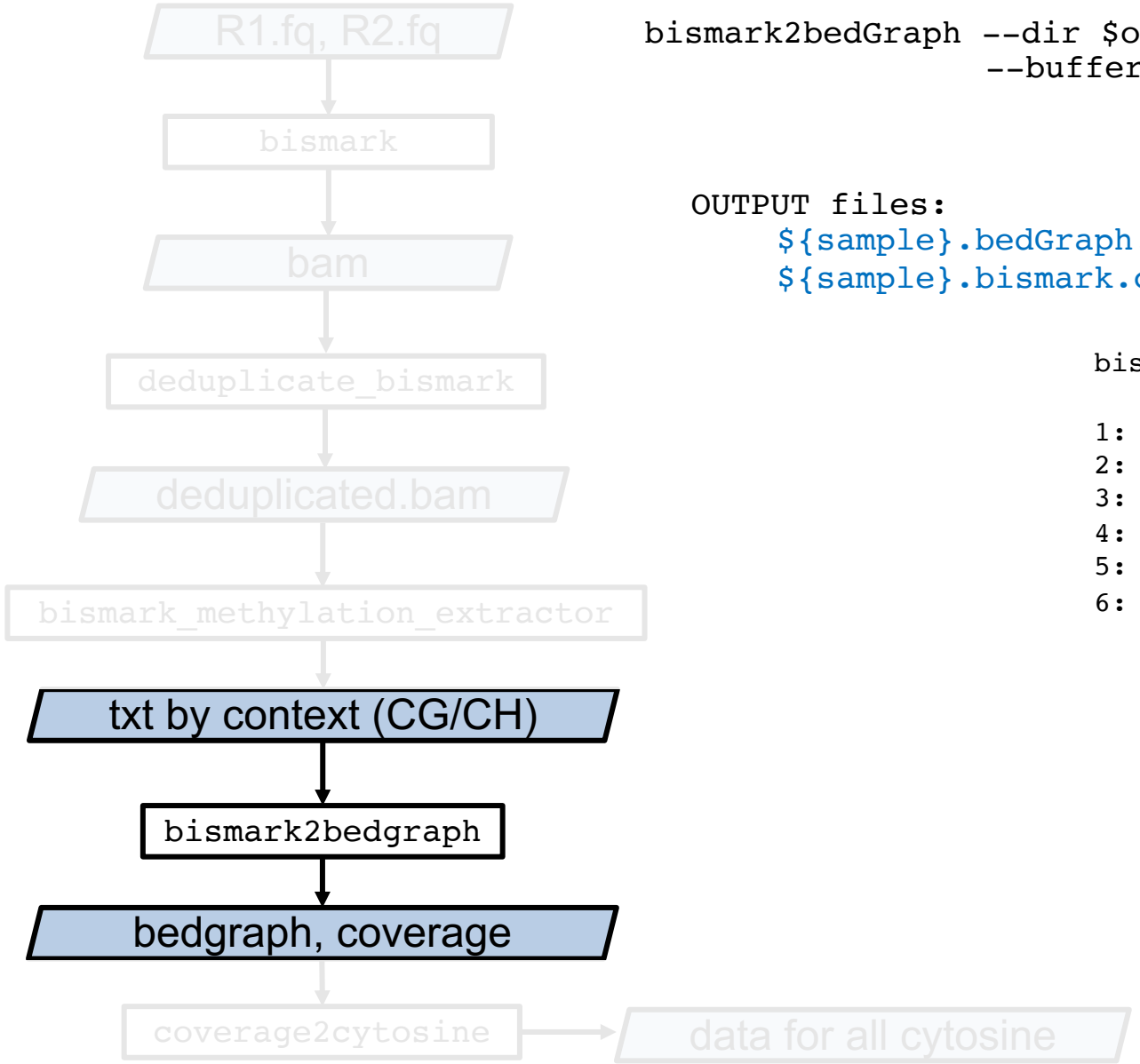
Data processed at 16:59 on 2021-05-25

Alignment

Sequence pairs analysed in total	99,849
Paired-end alignments with a unique best hit	89,261
Pairs without alignments under any condition	4,578
Pairs that did not map uniquely	6,010
Genomic sequence context not extractable (edges of chromosomes)	0

https://wangftp.wustl.edu/~hlee/wgbs/2_bismark/toy_bismark_bt2_PE_report.html

Bismark processes of WGBS data

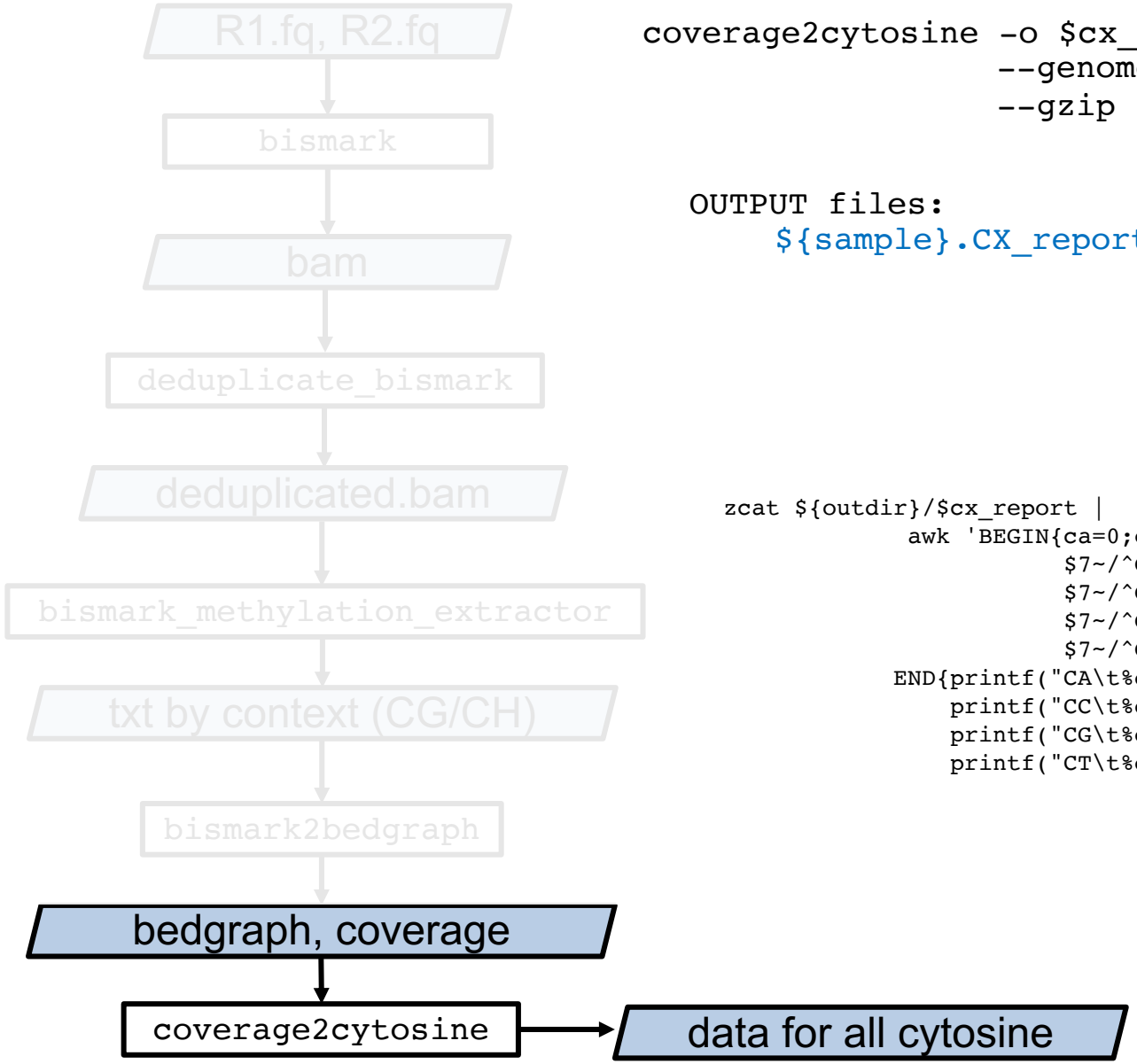



```
bismark2bedGraph --dir $outdir --cutoff 1 --CX_context \
                 --buffer_size=75G --scaffolds -o $bedGraph $merged
```

```
OUTPUT files:
    ${sample}.bedGraph.gz
    ${sample}.bismark.cov.gz
```

```
bismark.cov.gz (1-based)

1: chromosome
2: start position
3: end position
4: methylation percentage
5: count methylated
6: count non-methylated
```

Bismark processes of WGBS data




```
coverage2cytosine -o $cx_report --dir $outdir \
                  --genome_folder $genome_dir --CX_context \
                  --gzip $cov
```

OUTPUT files:

${sample}.CX_report.txt.gz

CX_report.txt.gz (1-based)

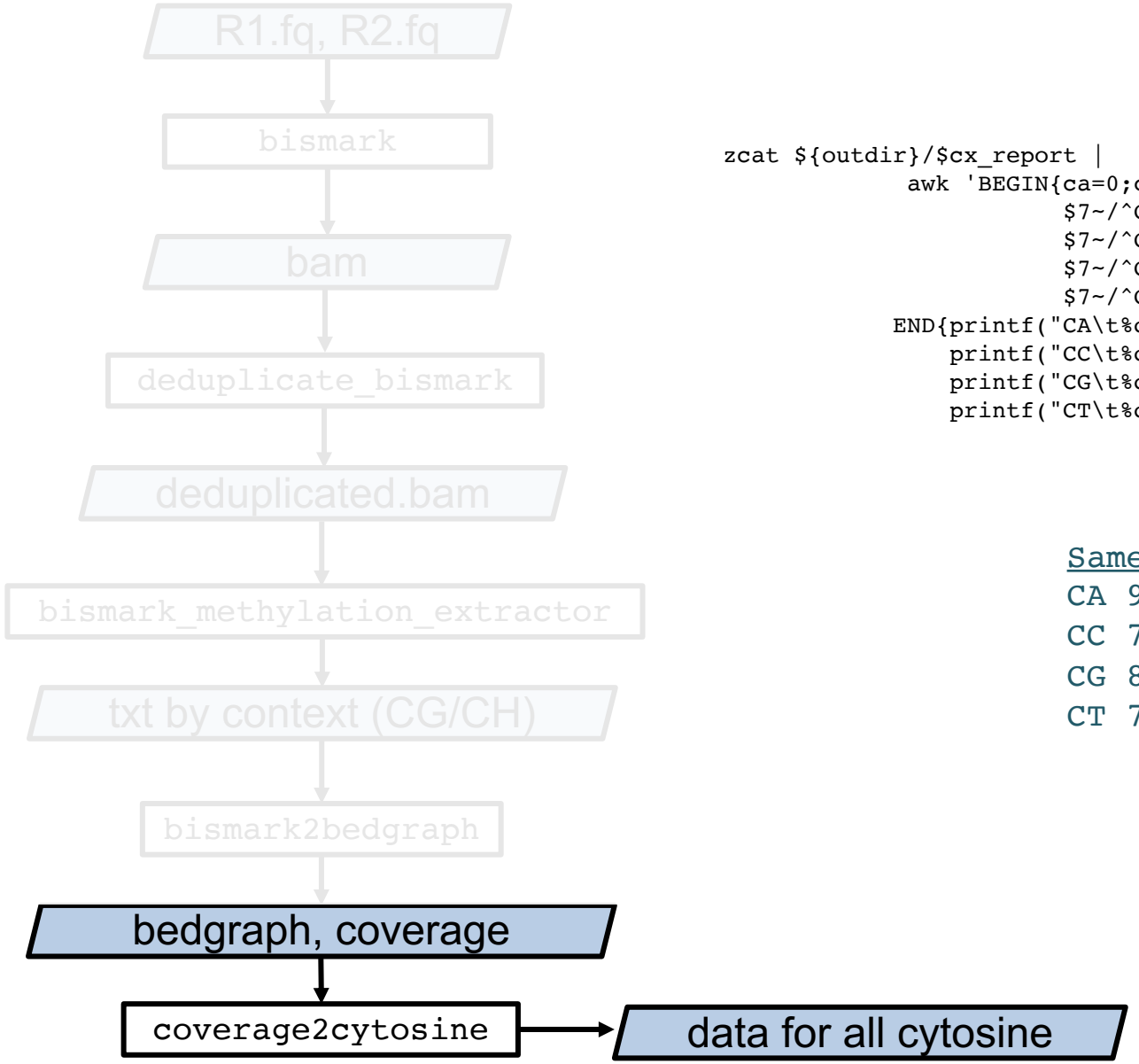
1: chromosome
2: position
3: strand
4: count methylated
5: count non-methylated
6: C-context
7: trinucleotide context

```
zcat ${outdir}/$cx_report |
        awk 'BEGIN{ca=0;cc=0;cg=0;ct=0;mca=0;mcc=0;mcg=0;mct=0}
                $7~/^CA/ {ca+=$5; mca+=$4}
                $7~/^CC/ {cc+=$5; mcc+=$4}
                $7~/^CG/ {cg+=$5; mcg+=$4}
                $7~/^CT/ {ct+=$5; mct+=$4}
            END{printf("CA\t%d\t%d\t%.3f\n", ca, mca, mca/(ca+mca));
                printf("CC\t%d\t%d\t%.3f\n", cc, mcc, mcc/(cc+mcc));
                printf("CG\t%d\t%d\t%.3f\n", cg, mcg, mcg/(cg+mcg));
                printf("CT\t%d\t%d\t%.3f\n", ct, mct, mct/(ct+mct));}' >$cx_me
```

CA	1094036	9714	0.009
CC	746623	5716	0.008
CG	28665	114932	0.800
CT	1061632	7658	0.007

CH methylation = 0.79%
conversion rate = 99.21%

Bismark processes of WGBS data




```
zcat ${outdir}/$cx_report |
            awk 'BEGIN{ca=0;cc=0;cg=0;ct=0;mca=0;mcc=0;mcg=0;mct=0}
                      $7~/^CA/ {ca+=$5; mca+=$4}
                      $7~/^CC/ {cc+=$5; mcc+=$4}
                      $7~/^CG/ {cg+=$5; mcg+=$4}
                      $7~/^CT/ {ct+=$5; mct+=$4}
          END{printf("CA\t%d\t%d\t%.3f\n", ca, mca, mca/(ca+mca));
              printf("CC\t%d\t%d\t%.3f\n", cc, mcc, mcc/(cc+mcc));
              printf("CG\t%d\t%d\t%.3f\n", cg, mcg, mcg/(cg+mcg));
              printf("CT\t%d\t%d\t%.3f\n", ct, mct, mct/(ct+mct));}' >$cx_me
```

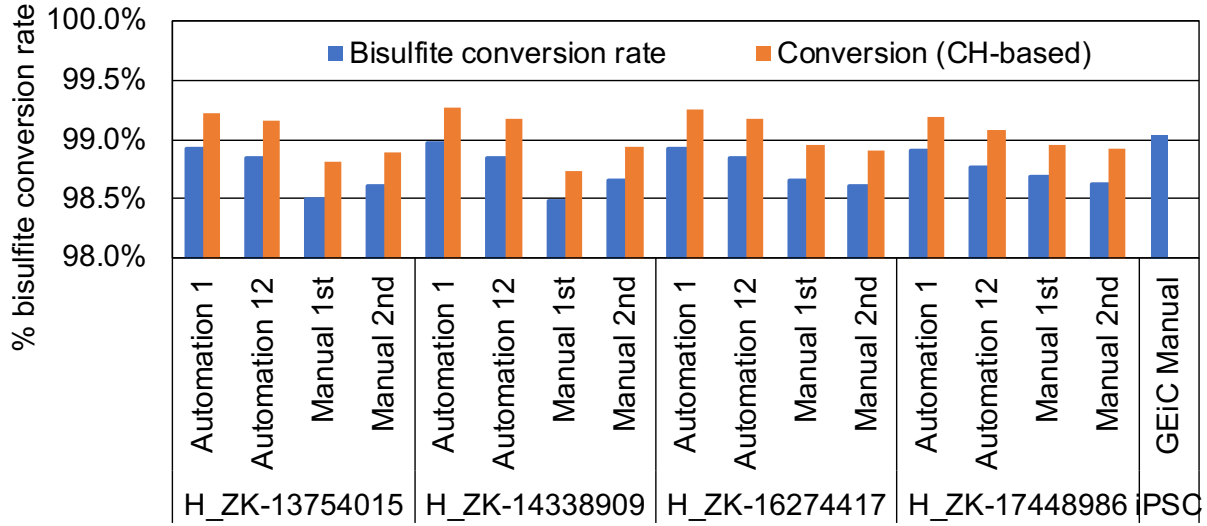
Same run on the lambda DNA

```
CA 907 21 0.023
CC 716 11 0.015
CG 829 11 0.013
CT 749 11 0.014
```

lambda DNA methylation = 1.66%
conversion rate = 98.34%

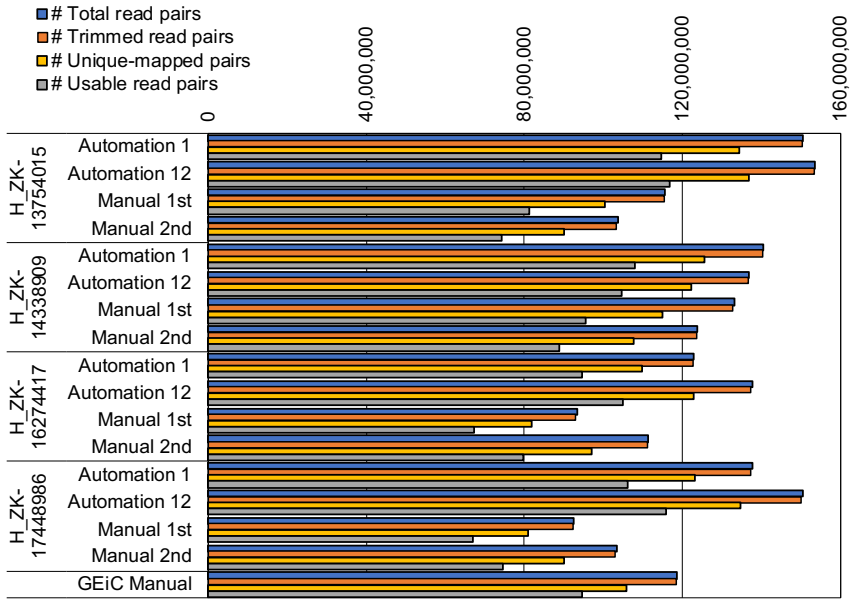
CH (non-CpG) methylation levels

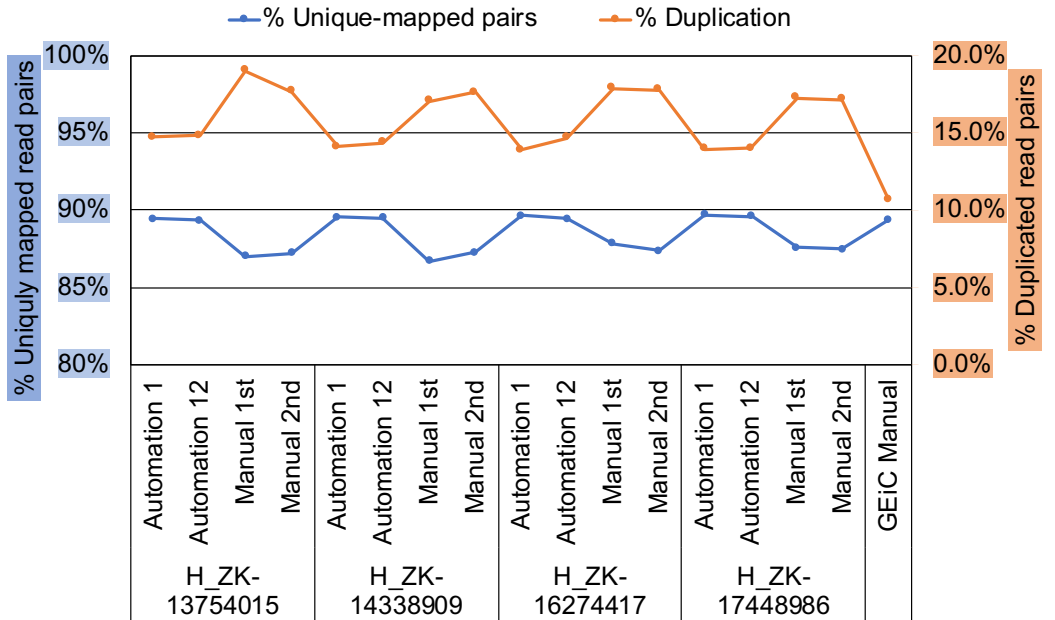
Library Name	Sample	Library prep	Bisulfite conversion rate	Mean mCH/CH	Mean mCA/CA	Mean mCC/CC	Mean mCT/CT	Conversion (CH-based)
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	0.989	0.008	0.009	0.007	0.007	0.992
H_ZK-13754015-lib5		Automation 12	0.988	0.008	0.010	0.008	0.008	0.992
H_ZK-13754015-lib2		Manual 1st	0.985	0.012	0.013	0.012	0.011	0.988
H_ZK-13754015-lib3		Manual 2nd	0.986	0.011	0.012	0.011	0.010	0.989
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	0.990	0.007	0.008	0.007	0.006	0.993
H_ZK-14338909-lib5		Automation 12	0.989	0.008	0.009	0.008	0.008	0.992
H_ZK-14338909-lib2		Manual 1st	0.985	0.013	0.013	0.013	0.012	0.987
H_ZK-14338909-lib3		Manual 2nd	0.987	0.011	0.012	0.010	0.010	0.989
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	0.989	0.007	0.009	0.007	0.007	0.993
H_ZK-16274417-lib5		Automation 12	0.989	0.008	0.009	0.008	0.007	0.992
H_ZK-16274417-lib2		Manual 1st	0.987	0.011	0.012	0.010	0.010	0.989
H_ZK-16274417-lib3		Manual 2nd	0.986	0.011	0.012	0.011	0.010	0.989
H_ZK-17448986-lib4	H_ZK-17448986	Automation 1	0.989	0.008	0.009	0.008	0.007	0.992
H_ZK-17448986-lib5		Automation 12	0.988	0.009	0.010	0.009	0.008	0.991
H_ZK-17448986-lib2		Manual 1st	0.987	0.010	0.012	0.010	0.010	0.990
H_ZK-17448986-lib3		Manual 2nd	0.986	0.011	0.012	0.010	0.010	0.989
GEiC_WGBS_A1	iPSC	GEiC Manual	0.990	0.013	0.022	0.006	0.009	0.987



Alignment statistics (Bismark)

Library Name	Sample	Library prep	# Total read pairs	# Removed pairs trimming	# Trimmed read pairs	# Mapped pairs	% Mapping	# Multi-mapped pairs	# Unique-mapped pairs	% Unique-mapped pairs	# Usable read pairs	% Duplication
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	150,592,761	241,425	150,351,336	143,614,953	95.5%	9,131,694	134,483,259	89.4%	114,704,936	14.7%
H_ZK-13754015-lib5		Automation 12	153,585,029	274,185	153,310,844	146,295,127	95.4%	9,357,591	136,937,536	89.3%	116,661,305	14.8%
H_ZK-13754015-lib2		Manual 1st	115,448,443	146,571	115,301,872	107,363,098	93.1%	7,049,891	100,313,207	87.0%	81,257,988	19.0%
H_ZK-13754015-lib3		Manual 2nd	103,826,384	425,707	103,400,677	96,508,077	93.3%	6,344,067	90,164,010	87.2%	74,221,802	17.7%
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	140,562,154	354,915	140,207,239	133,834,566	95.5%	8,290,103	125,544,463	89.5%	107,856,247	14.1%
H_ZK-14338909-lib5		Automation 12	136,853,914	234,610	136,619,304	130,551,612	95.6%	8,298,688	122,252,924	89.5%	104,731,323	14.3%
H_ZK-14338909-lib2		Manual 1st	133,211,507	433,348	132,778,159	123,357,128	92.9%	8,258,343	115,098,785	86.7%	95,470,203	17.1%
H_ZK-14338909-lib3		Manual 2nd	123,896,710	369,794	123,526,916	115,346,819	93.4%	7,531,162	107,815,657	87.3%	88,853,302	17.6%
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	122,794,756	245,881	122,548,875	116,975,961	95.5%	7,133,837	109,842,124	89.6%	94,602,181	13.9%
H_ZK-16274417-lib5		Automation 12	137,723,144	341,165	137,381,979	131,199,824	95.5%	8,331,175	122,868,649	89.4%	104,850,691	14.7%
H_ZK-16274417-lib2		Manual 1st	93,256,705	193,786	93,062,919	87,118,676	93.6%	5,387,837	81,730,839	87.8%	67,143,931	17.8%
H_ZK-16274417-lib3		Manual 2nd	111,316,902	277,457	111,039,445	103,614,374	93.3%	6,574,702	97,039,672	87.4%	79,797,561	17.8%
H_ZK-17448986-lib4	H_ZK-17448986	Automation 1	137,698,365	351,875	137,346,490	131,095,303	95.4%	7,873,525	123,221,778	89.7%	106,033,442	13.9%
H_ZK-17448986-lib5		Automation 12	150,437,804	310,606	150,127,198	143,339,207	95.5%	8,766,055	134,573,152	89.6%	115,748,895	14.0%
H_ZK-17448986-lib2		Manual 1st	92,525,438	230,554	92,294,884	86,157,923	93.4%	5,335,345	80,822,578	87.6%	66,890,935	17.2%
H_ZK-17448986-lib3		Manual 2nd	103,340,524	364,797	102,975,727	97,049,195	94.2%	6,981,274	90,067,921	87.5%	74,642,952	17.1%
GEiC_WGBS_A1		GEiC Manual	118,539,862	56,839	118,483,023	112,722,360	95.1%	6,880,983	105,841,377	89.3%	94,538,071	10.7%



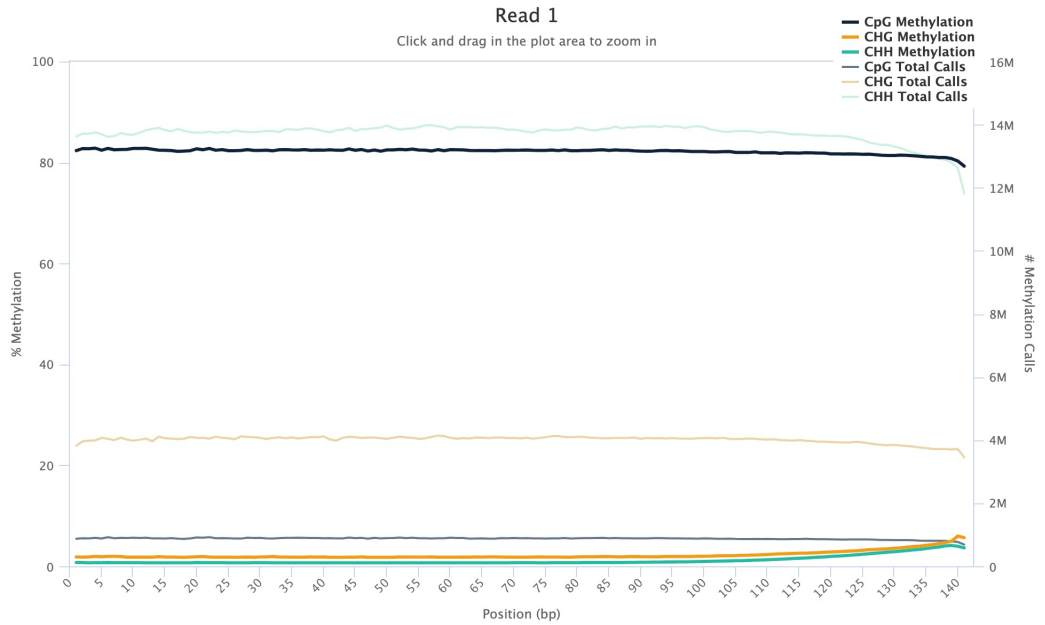


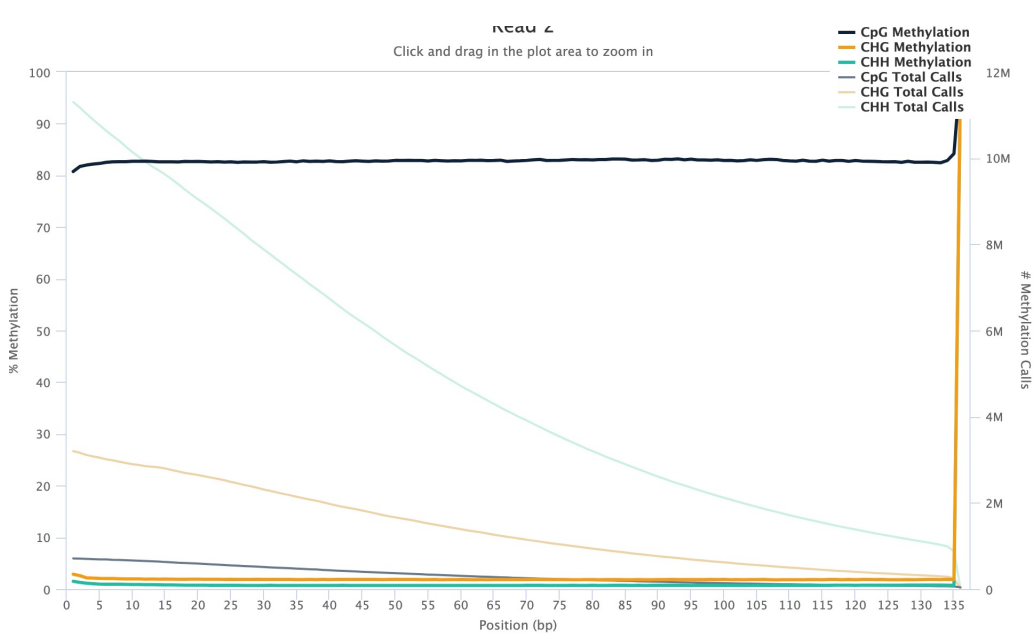
M-bias plot (bias at the 5’ end of the reads)

Read 1

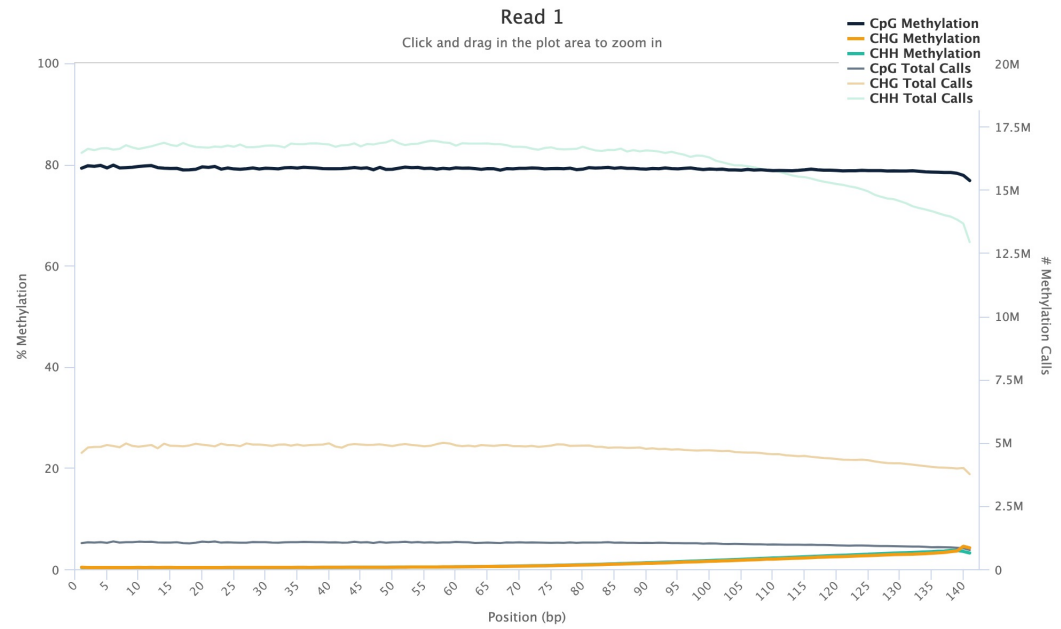
Read 2

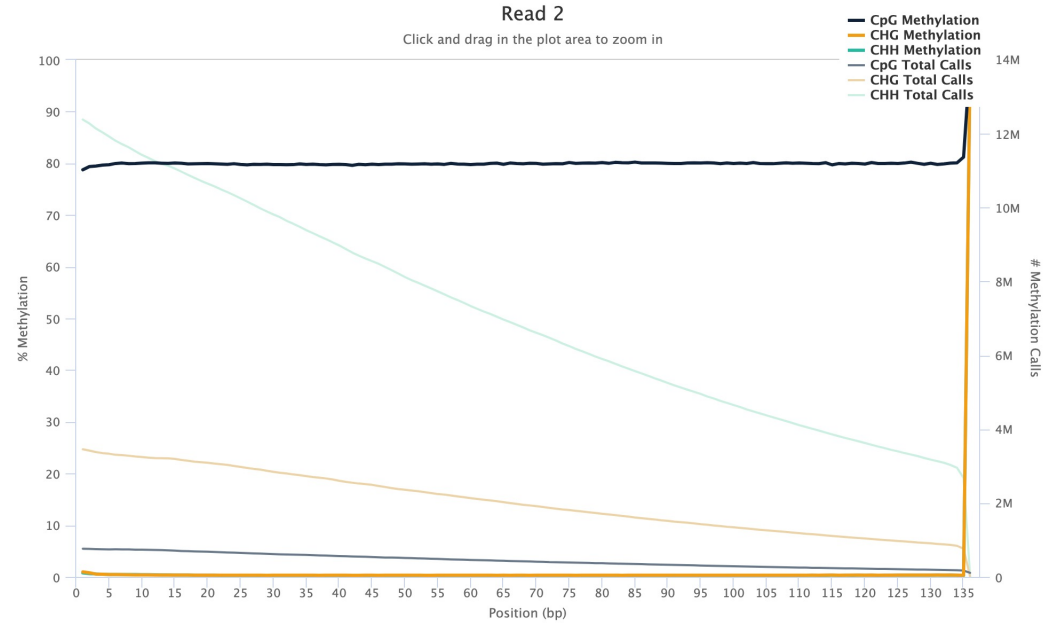
GEiC_WGBS_A1



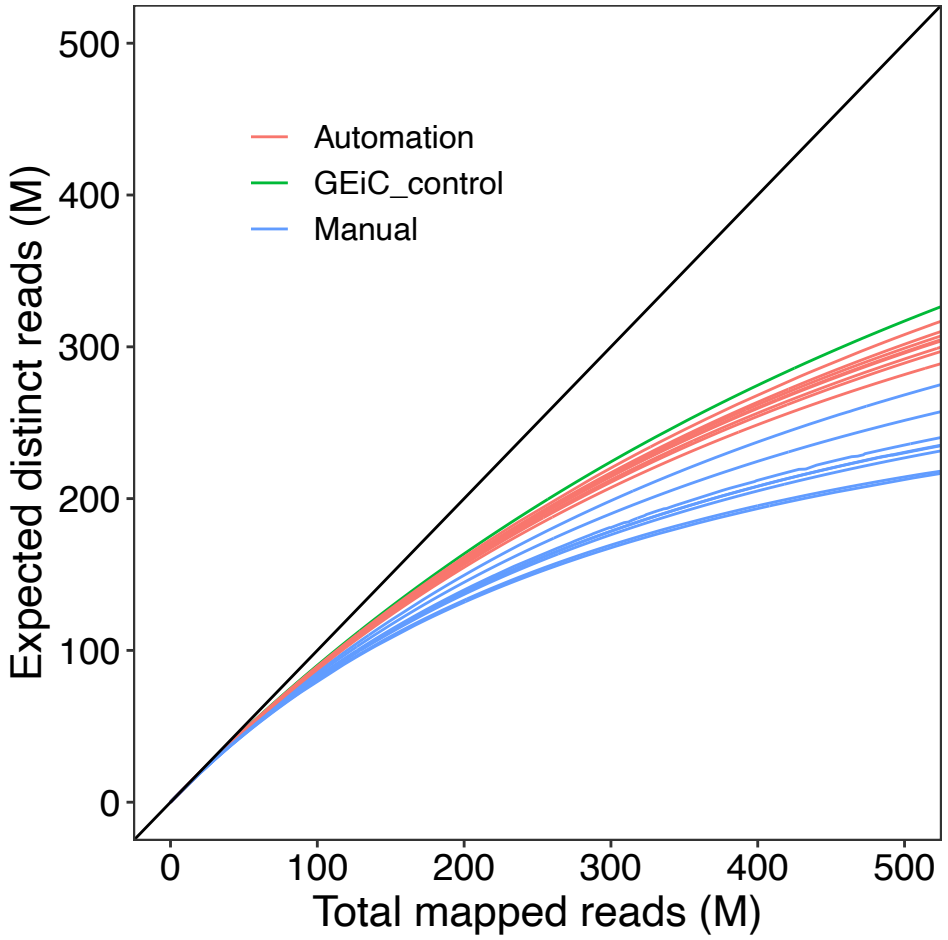


H_ZK-17448986_auto_rep2





Library complexity



Preseq v.2.0.2

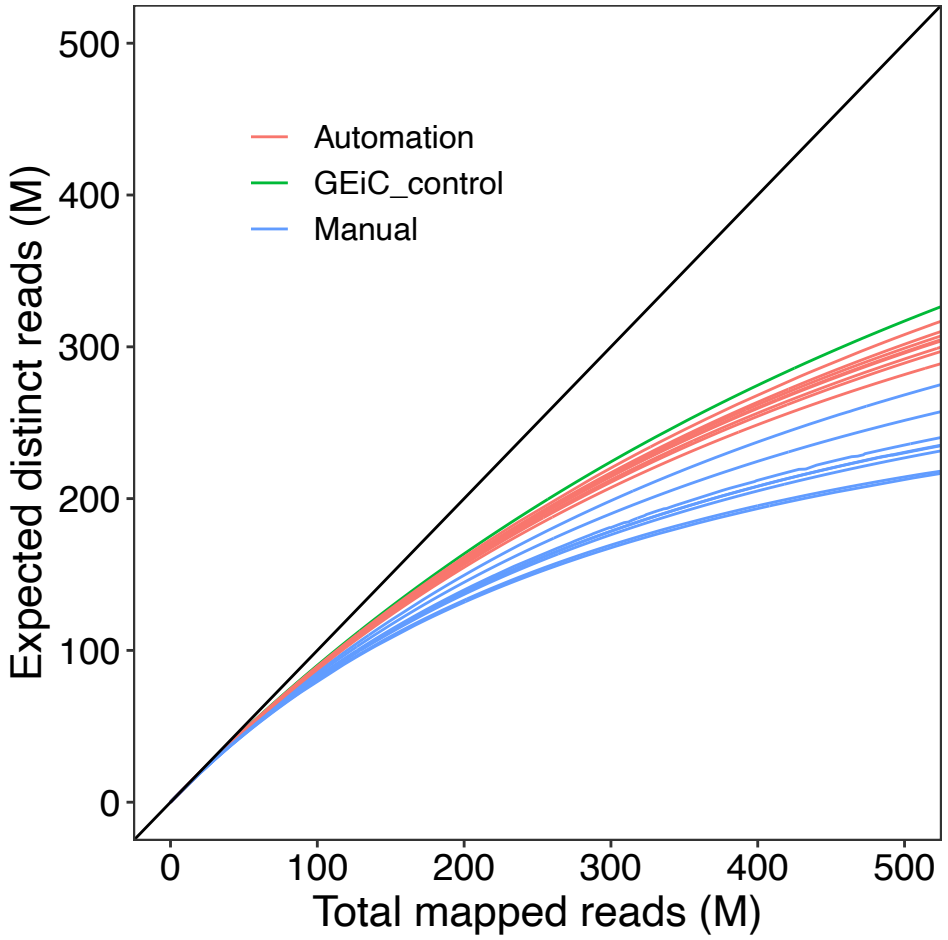
```
samtools sort -m 2G -o $bam_sorted -T /tmp/$sample -@ $CPU $bam

preseq lc_extrap -o $output -B -P -D $bam_sorted
```

INPUT file:
${sample}_bismark_bt2_pe.bam (before deduplication)

OUTPUT file:
${sample}.preseq_lc_extrap.txt

TOTAL_READS	EXPECTED_DISTINCT	LOWER_0.95CI	UPPER_0.95CI
0	0	0	0
1000000.0	392033.8	217277.9	423259.6
2000000.0	434387.5	-667779.7	1771596.6
3000000.0	443890.9	-1301258.6	1919133.3
4000000.0	444192.7	-10597131.7	2407845.7
5000000.0	478680.5	-1241946.1	4536247.4
6000000.0	472283.5	-2650920.1	3068362.6
7000000.0	466143.0	-6226620.2	2547373.3
8000000.0	476648.6	-3323457.5	2501649.3
9000000.0	488658.1	-3381684.0	6370650.1
10000000.0	487053.2	-4727512.6	4850590.4
.....			
9995000000.0	535163.7	-1767186.8	4478377.3
9996000000.0	535163.7	-1767187.2	4478376.9
9997000000.0	535163.7	-1767187.5	4478376.5
9998000000.0	535163.7	-1767187.8	4478376.1
9999000000.0	535163.7	-1767188.1	4478375.6

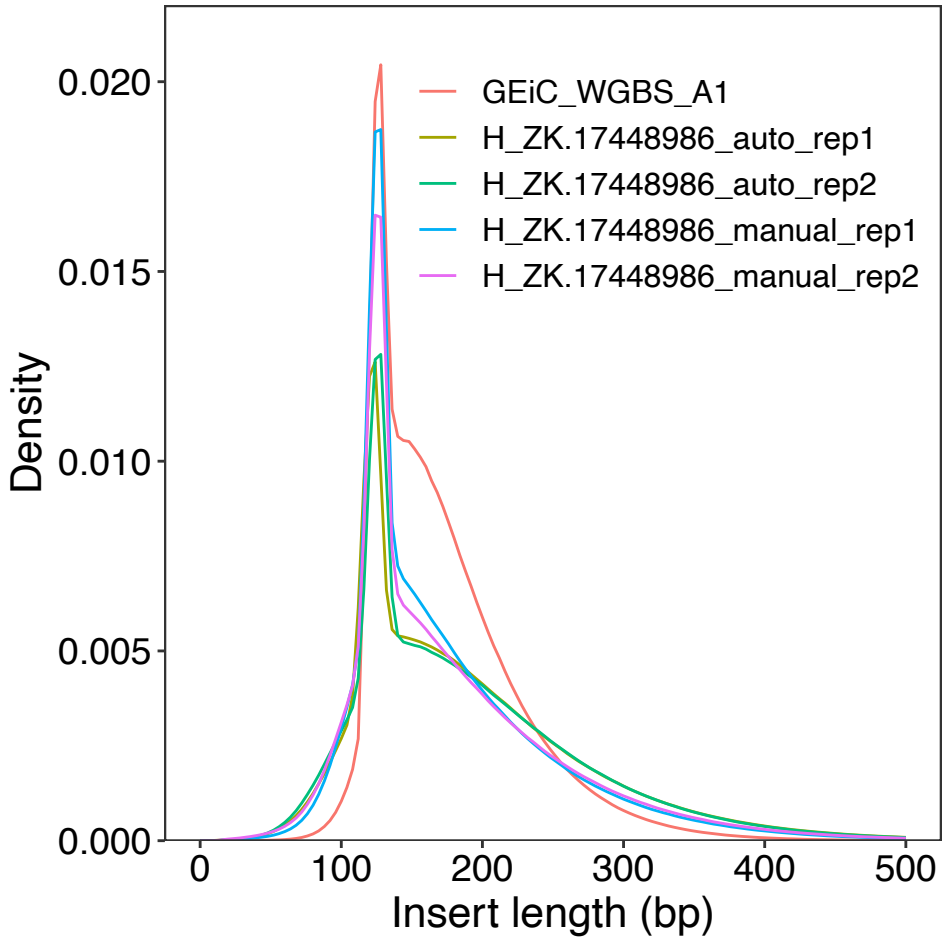
Library complexity



Library Name	Sample	Library prep	# Unique-mapped pairs	# Usable read pairs	% Duplication
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	134,483,259	114,704,936	14.7%
H_ZK-13754015-lib5		Automation 12	136,937,536	116,661,305	14.8%
H_ZK-13754015-lib2		Manual 1st	100,313,207	81,257,988	19.0%
H_ZK-13754015-lib3		Manual 2nd	90,164,010	74,221,802	17.7%
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	125,544,463	107,856,247	14.1%
H_ZK-14338909-lib5		Automation 12	122,252,924	104,731,323	14.3%
H_ZK-14338909-lib2		Manual 1st	115,098,785	95,470,203	17.1%
H_ZK-14338909-lib3		Manual 2nd	107,815,657	88,853,302	17.6%
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	109,842,124	94,602,181	13.9%
H_ZK-16274417-lib5		Automation 12	122,868,649	104,850,691	14.7%
H_ZK-16274417-lib2		Manual 1st	81,730,839	67,143,931	17.8%
H_ZK-16274417-lib3		Manual 2nd	97,039,672	79,797,561	17.8%
H_ZK-17448986-lib4	H_ZK-17448986	Automation 1	123,221,778	106,033,442	13.9%
H_ZK-17448986-lib5		Automation 12	134,573,152	115,748,895	14.0%
H_ZK-17448986-lib2		Manual 1st	80,822,578	66,890,935	17.2%
H_ZK-17448986-lib3		Manual 2nd	90,067,921	74,642,952	17.1%
GEiC_WGBS_A1		GEiC Manual	105,841,377	94,538,071	10.7%

Preseq v.2.0.2

Insert length distribution




```
# select the first 100K alignments
samtools view -h -@ $CPU $bam_dedup |
    head -100000197 | samtools view -b -o $bam_tmp -@ $CPU

# make temporary insert length txt file
bamToBed -bedpe -i $bam_tmp | awk -vOFS="\t" '{print $1,$2,$6,$6-$2}' | gzip -nc > $insert_tmp

Rscript $rscript_insert $insert_tmp $out_insert &> $rlog_insert
```

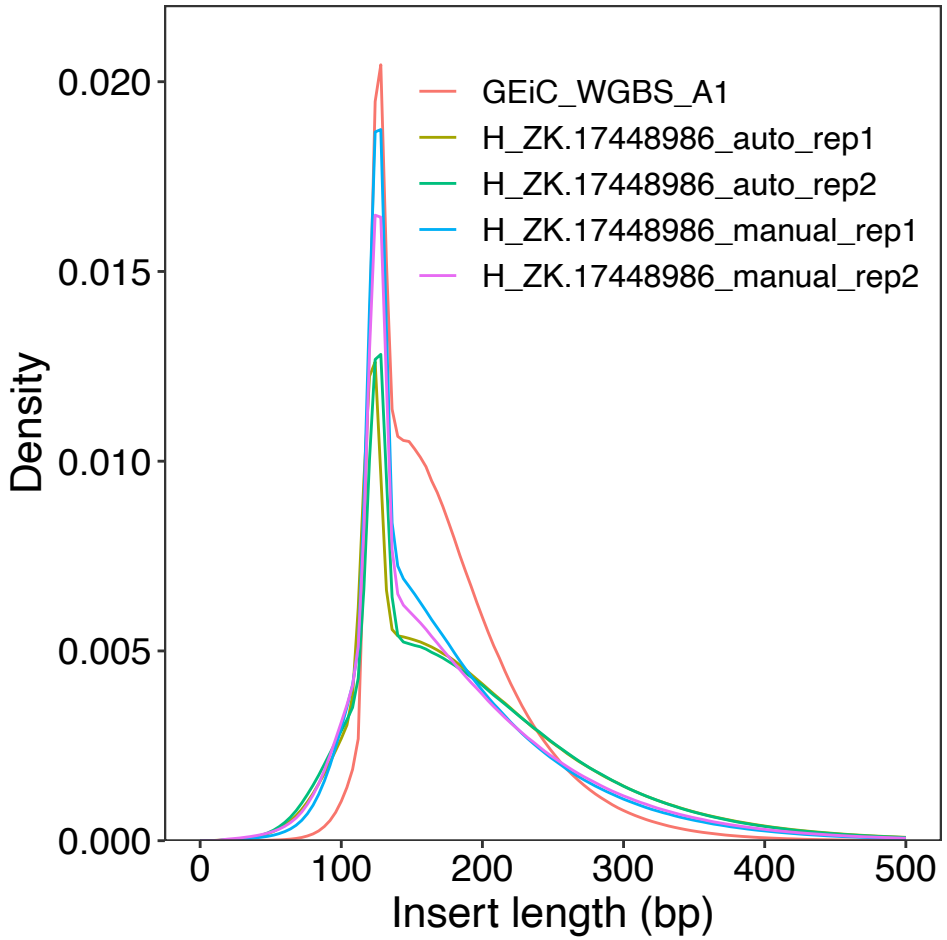
[$rscript_insert]

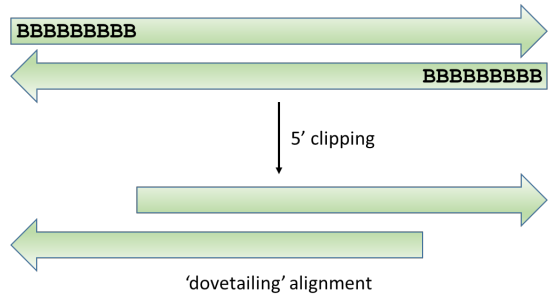
```
insert <- read.delim(file = file, as.is = TRUE, header=F)

### density graph
g <- ggplot(insert, aes(insert[,4])) + geom_density(n=126)

### plot coordiantes
g <- ggplot_build(g)
density <- c( base, g$data[[1]][,1] )
write(density, wfile, sep ="\t")
```

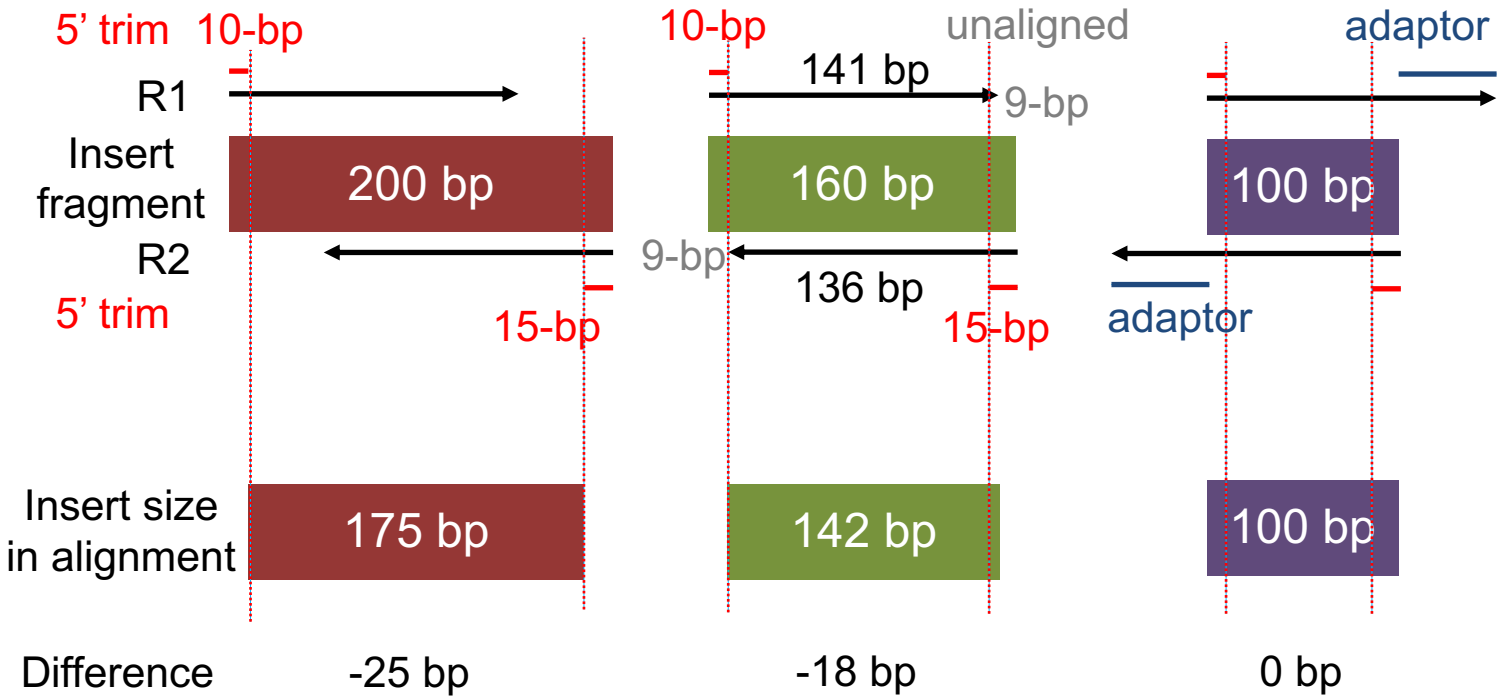
OUTPUT file:
${sample}.insert_length.txt

Insertion length of 100-300 bp, median ~150 bp (10 (R1) + 15 (R2) 5’ end trimming)
PE reads of 2x151 bp will have some overlaps

Insert length distribution



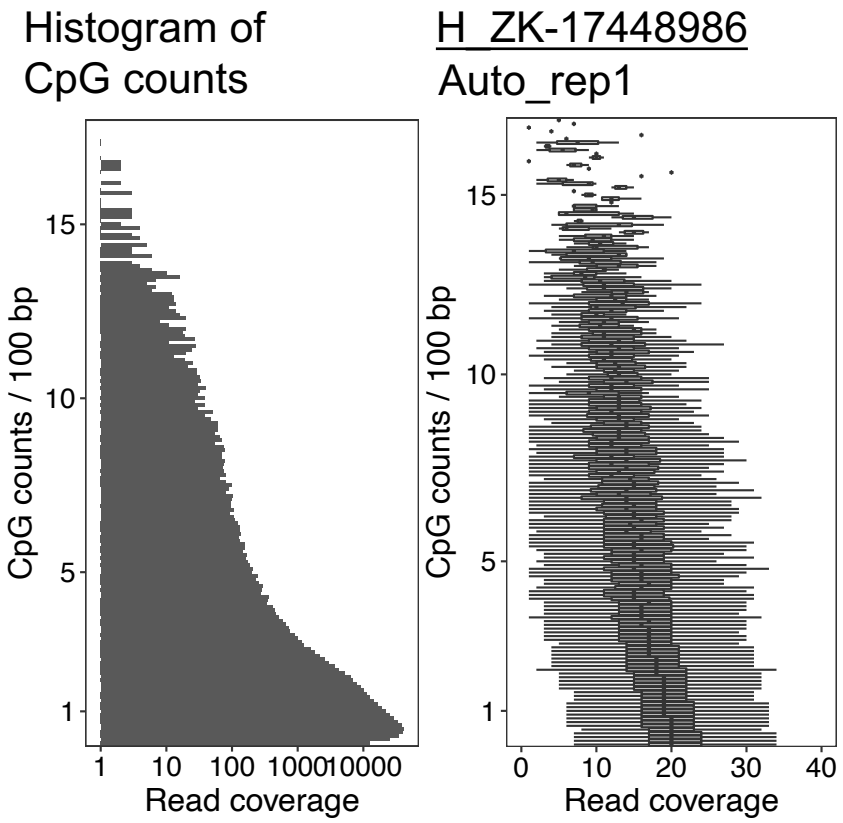




- Insert fragments with the lengths of 130-170 bp got trimmed with different numbers of bases (0-25).

Insertion length of 100-300 bp, median ~150 bp (10 (R1) + 15 (R2) 5' end trimming)
PE reads of 2x151 bp will have some overlaps

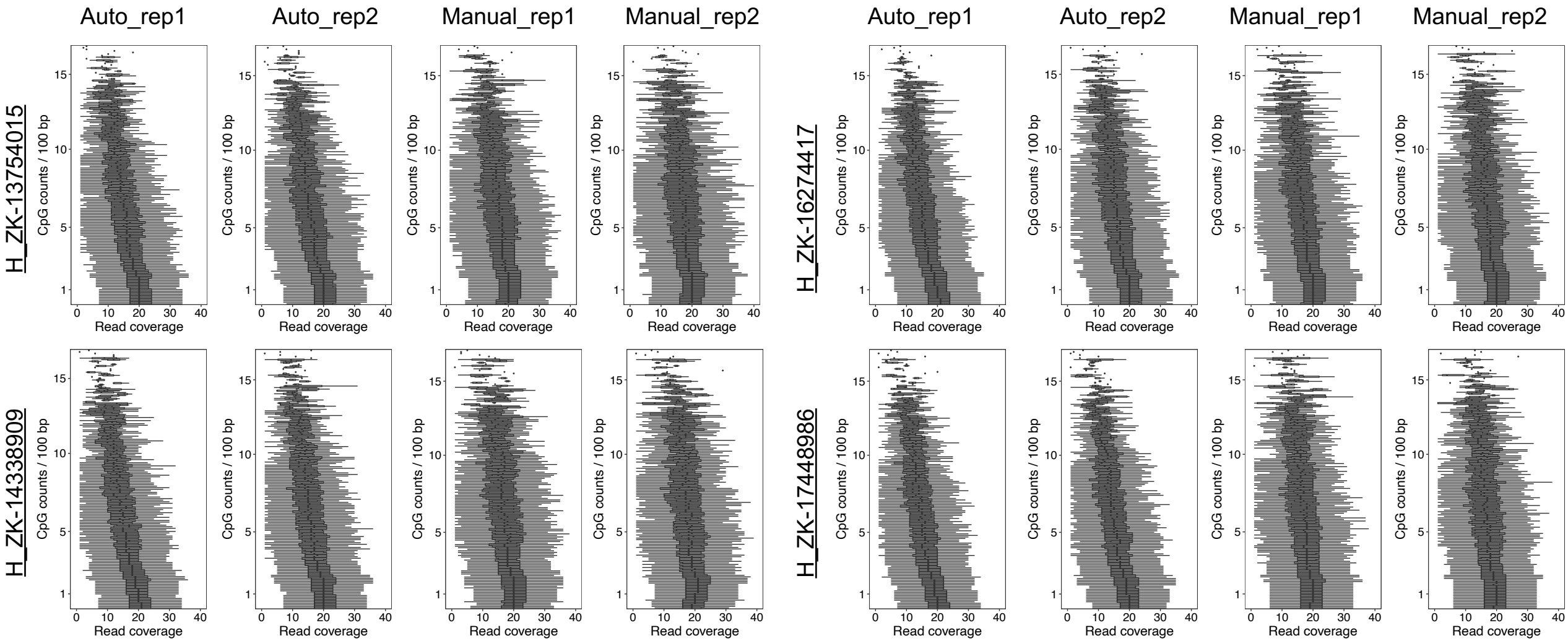
GC bias (CpG contents vs. coverages)




```
# select the first 100K alignments
samtools view -h -@ $CPU $bam_dedup |
    head -100000197 | samtools view -b -o $bam_tmp -@ $CPU

# make temporary insert length txt file
bamToBed -bedpe -i $bam_tmp | awk -vOFS="\t" '{print $1,$2,$6,$6-$2}' | gzip -nc >
$insert_tmp

Rscript $rscript_cpgbias $cov_chr1 $sample &> $rlog_cpgbias
```

1 kb bins with 500 bp sliding window in chr1

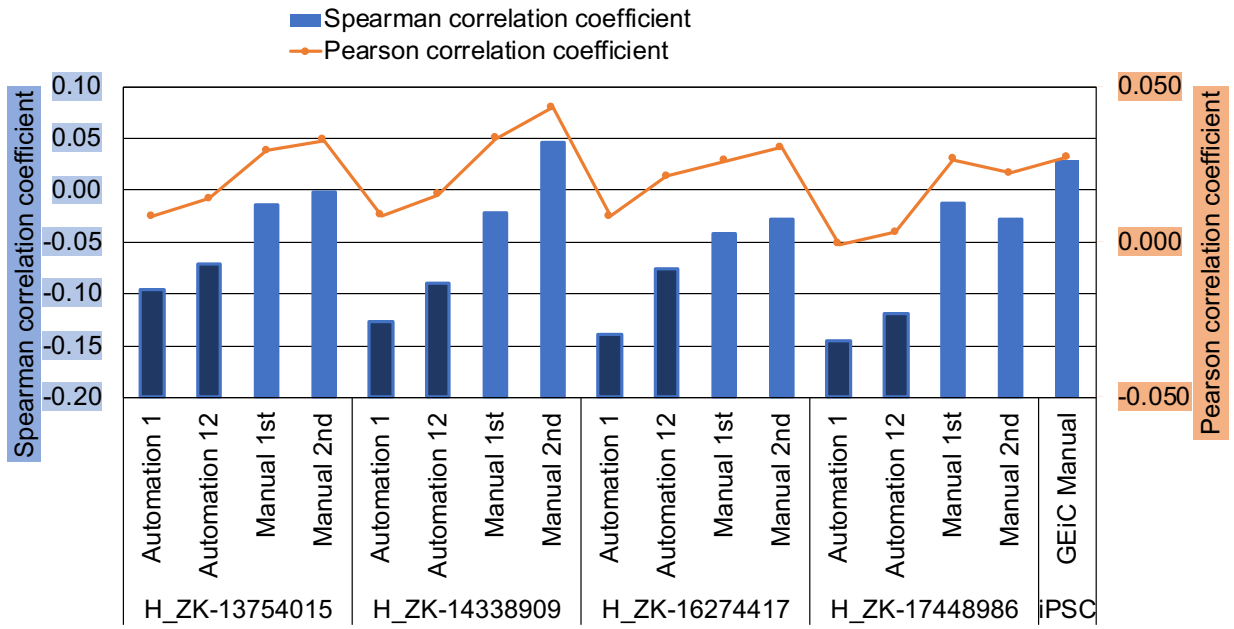
GC bias (CpG contents vs. coverages)



1 kb bins with 500 bp sliding window in chr1

GC bias (CpG contents vs. coverages)

Library Name	Sample	Library prep	Spearman correlation coefficient	Pearson correlation coefficient
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	-0.095	0.008
H_ZK-13754015-lib5		Automation 12	-0.072	0.014
H_ZK-13754015-lib2		Manual 1st	-0.014	0.029
H_ZK-13754015-lib3		Manual 2nd	-0.002	0.033
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	-0.127	0.008
H_ZK-14338909-lib5		Automation 12	-0.089	0.015
H_ZK-14338909-lib2		Manual 1st	-0.022	0.033
H_ZK-14338909-lib3		Manual 2nd	0.045	0.043
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	-0.139	0.008
H_ZK-16274417-lib5		Automation 12	-0.076	0.021
H_ZK-16274417-lib2		Manual 1st	-0.042	0.026
H_ZK-16274417-lib3		Manual 2nd	-0.029	0.030
H_ZK-17448986-lib2	H_ZK-17448986	Automation 1	-0.146	-0.001
H_ZK-17448986-lib3		Automation 12	-0.119	0.003
H_ZK-17448986-lib4		Manual 1st	-0.012	0.026
H_ZK-17448986-lib5		Manual 2nd	-0.028	0.022
GEiC_WGBS_A1	iPSC	GEiC Manual	0.030	0.027



Genome, C, and CpG coverage

Library Name	Sample	Library prep	# Usable pairs	Max coverage	C coverage in genome	C coverage in CpGs	C coverage in CGI	CpG coverage	CpG coverage in CGI
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	114,704,936	5.12	3.68	3.54	2.36	7.07	4.72
H_ZK-13754015-lib5		Automation 12	116,661,305	5.21	3.77	3.65	2.52	7.30	5.04
H_ZK-13754015-lib2		Manual 1st	81,257,988	3.63	2.37	2.34	1.77	4.67	3.54
H_ZK-13754015-lib3		Manual 2nd	74,221,802	3.32	2.24	2.21	1.74	4.43	3.48

```
CA      1094036  9714     0.009
CC      746623   5716     0.008
CG      28665    114932   0.800
CT      1061632  7658     0.007
```

```
cnt_c=$(( 598683433+600854940 ))        # Watson strand + Crick strand
cnt_c=$(( $cnt_c - 171823*2 ))          # Discard chrEBV
cnt_cg=$(( 29303965 * 2 ))

c_cov=$( cat $cx_me | awk -F"\t" -v c=$cnt_c 'BEGIN{s=0} {s+=$2+$3} END{print s/c}' )
cg_cov=$( cat $cx_me | awk -F"\t" -v c=$cnt_cg 'BEGIN{s=0} $1=="CG" {s+=$2+$3} END{print s/c}' )

echo -e "$sample\t$c_cov\t$cg_cov" > $cov_genome
```

Max coverage =

$$\frac{\text{(\# usable pairs) X (read length)}}{\text{(genome size) X 2 strands}}$$

Read length = (151 X 2) – 10 – 15 (trimming)

Genome size ~ 3.1 X 10^9

Genome, C, and CpG coverage

Library Name	Sample	Library prep	# Usable pairs	Max coverage	C coverage in genome	C coverage in CpGs	C coverage in CGI	CpG coverage	CpG coverage in CGI
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	114,704,936	5.12	3.68	3.54	2.36	7.07	4.72
H_ZK-13754015-lib5		Automation 12	116,661,305	5.21	3.77	3.65	2.52	7.30	5.04
H_ZK-13754015-lib2		Manual 1st	81,257,988	3.63	2.37	2.34	1.77	4.67	3.54
H_ZK-13754015-lib3		Manual 2nd	74,221,802	3.32	2.24	2.21	1.74	4.43	3.48
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	107,856,247	4.82	3.58	3.41	2.23	6.82	4.45
H_ZK-14338909-lib5		Automation 12	104,731,323	4.68	3.32	3.19	2.20	6.38	4.40
H_ZK-14338909-lib2		Manual 1st	95,470,203	4.27	2.75	2.71	2.07	5.43	4.13
H_ZK-14338909-lib3		Manual 2nd	88,853,302	3.97	2.69	2.70	2.12	5.41	4.24
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	94,602,181	4.23	3.16	2.98	1.99	5.96	3.97
H_ZK-16274417-lib5		Automation 12	104,850,691	4.68	3.41	3.30	2.29	6.60	4.58
H_ZK-16274417-lib2		Manual 1st	67,143,931	3.00	2.10	2.05	1.50	4.10	3.00
H_ZK-16274417-lib3		Manual 2nd	79,797,561	3.57	2.44	2.40	1.82	4.80	3.63
H_ZK-17448986-lib4	H_ZK-17448986	Automation 1	106,033,442	4.74	3.41	3.21	2.11	6.41	4.23
H_ZK-17448986-lib5		Automation 12	115,748,895	5.17	3.65	3.48	2.26	6.95	4.52
H_ZK-17448986-lib2		Manual 1st	66,890,935	2.99	2.04	2.00	1.54	4.01	3.08
H_ZK-17448986-lib3		Manual 2nd	74,642,952	3.33	2.29	2.23	1.71	4.47	3.41
GEiC_WGBS_A1	iPSC	GEiC Manual	94,538,071	4.22	2.93	2.83	1.96	5.66	3.93

Max coverage =

$$\frac{\text{(\# usable pairs) X (read length)}}{\text{(genome size) X 2 strands}}$$

Read length = (151 X 2) – 10 – 15 (trimming)

Genome size ~ 3.1×10^9

Making browser methylc tracks

INPUT (bismark CX report file)

```
chr1 10004 + 0 0 CHH CCC
chr1 10005 + 0 0 CHH CCT
chr1 10006 + 0 0 CHH CTA
chr1 10010 + 0 0 CHH CCC
chr1 10011 + 0 0 CHH CCT
chr1 10012 + 0 0 CHH CTA
chr1 10016 + 0 0 CHH CCC
chr1 10017 + 0 0 CHH CCT
chr1 10018 + 0 0 CHH CTA
chr1 10022 + 0 0 CHH CCC
```

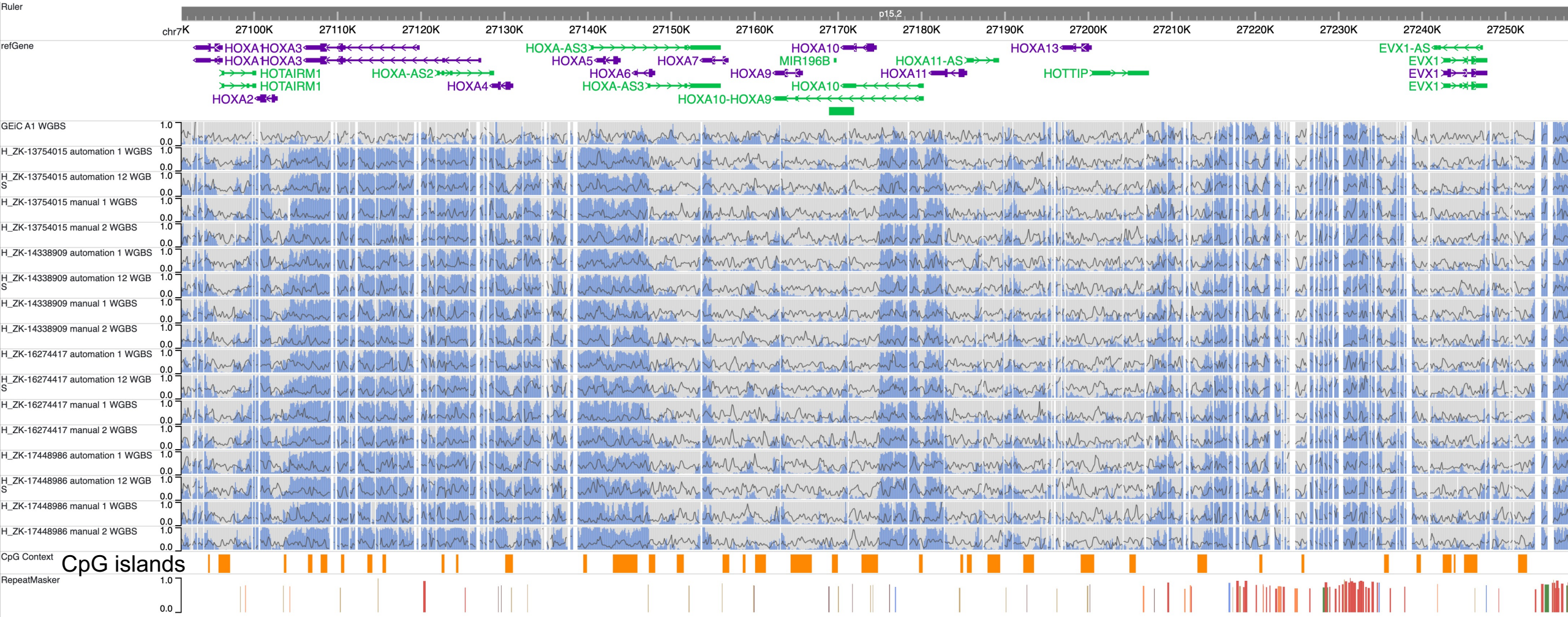

OUTPUT (browser methylc track)

```
chr1 10468 10470 CG 0.000 + 1
chr1 10470 10472 CG 1.000 + 1
chr1 10488 10490 CG 1.000 + 1
chr1 10492 10494 CG 0.000 + 1
chr1 10496 10498 CG 1.000 + 1
chr1 10524 10526 CG 1.000 + 1
chr1 10541 10543 CG 1.000 + 1
chr1 10562 10564 CG 1.000 + 1
chr1 10570 10572 CG 1.000 + 1
chr1 10576 10578 CG 0.500 + 2
```



```
zcat $cx_report |
    awk -F"\t" 'BEGIN{OFS=FS} $6=="CG" && $4+$5>0 { if ($3=="+") {print $1,$2-1,$2+1,$4,$5} if ($3=="-") {print $1,$2-2,$2,$4,$5} }' |
    sort -k1,1 -k2,2n |
    groupBy -g 1,2,3 -c 4,5 -o sum,sum |
    awk -F"\t" 'BEGIN{OFS=FS} {mcg=sprintf("%.3f", $4/($4+$5)); print $1,$2,$3,"CG",mcg,"+",$4+$5 }' |
    bgzip >$methylc
```

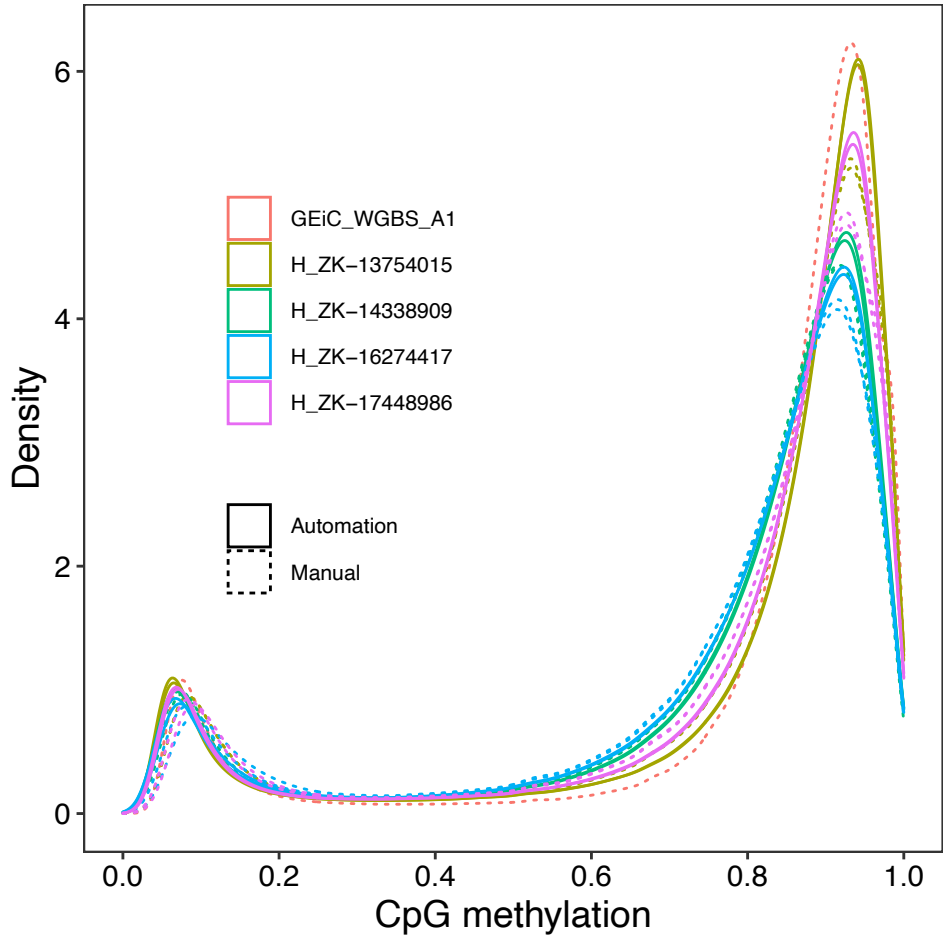
Browser view of WGBS data



http://epigenomegateway.wustl.edu/browser/?genome=hg38&hub=https://wangftp.wustl.edu/~hlee/mgi/mgi.json

CpG methylation: average level and bimodal distribution

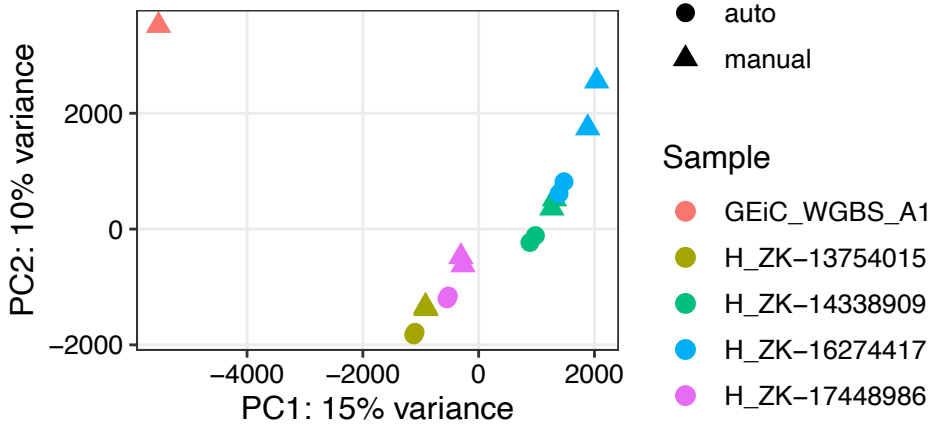
Library Name	Sample	Library prep	Mean mCG/CG
H_ZK-13754015-lib4	H_ZK-13754015	Automation 1	0.805
H_ZK-13754015-lib5		Automation 12	0.804
H_ZK-13754015-lib2		Manual 1st	0.792
H_ZK-13754015-lib3		Manual 2nd	0.789
H_ZK-14338909-lib4	H_ZK-14338909	Automation 1	0.777
H_ZK-14338909-lib5		Automation 12	0.773
H_ZK-14338909-lib2		Manual 1st	0.763
H_ZK-14338909-lib3		Manual 2nd	0.763
H_ZK-16274417-lib4	H_ZK-16274417	Automation 1	0.770
H_ZK-16274417-lib5		Automation 12	0.769
H_ZK-16274417-lib2		Manual 1st	0.762
H_ZK-16274417-lib3		Manual 2nd	0.759
H_ZK-17448986-lib4	H_ZK-17448986	Automation 1	0.793
H_ZK-17448986-lib5		Automation 12	0.794
H_ZK-17448986-lib2		Manual 1st	0.781
H_ZK-17448986-lib3		Manual 2nd	0.779
GEiC_WGBS_A1	iPSC	GEiC Manual	0.824

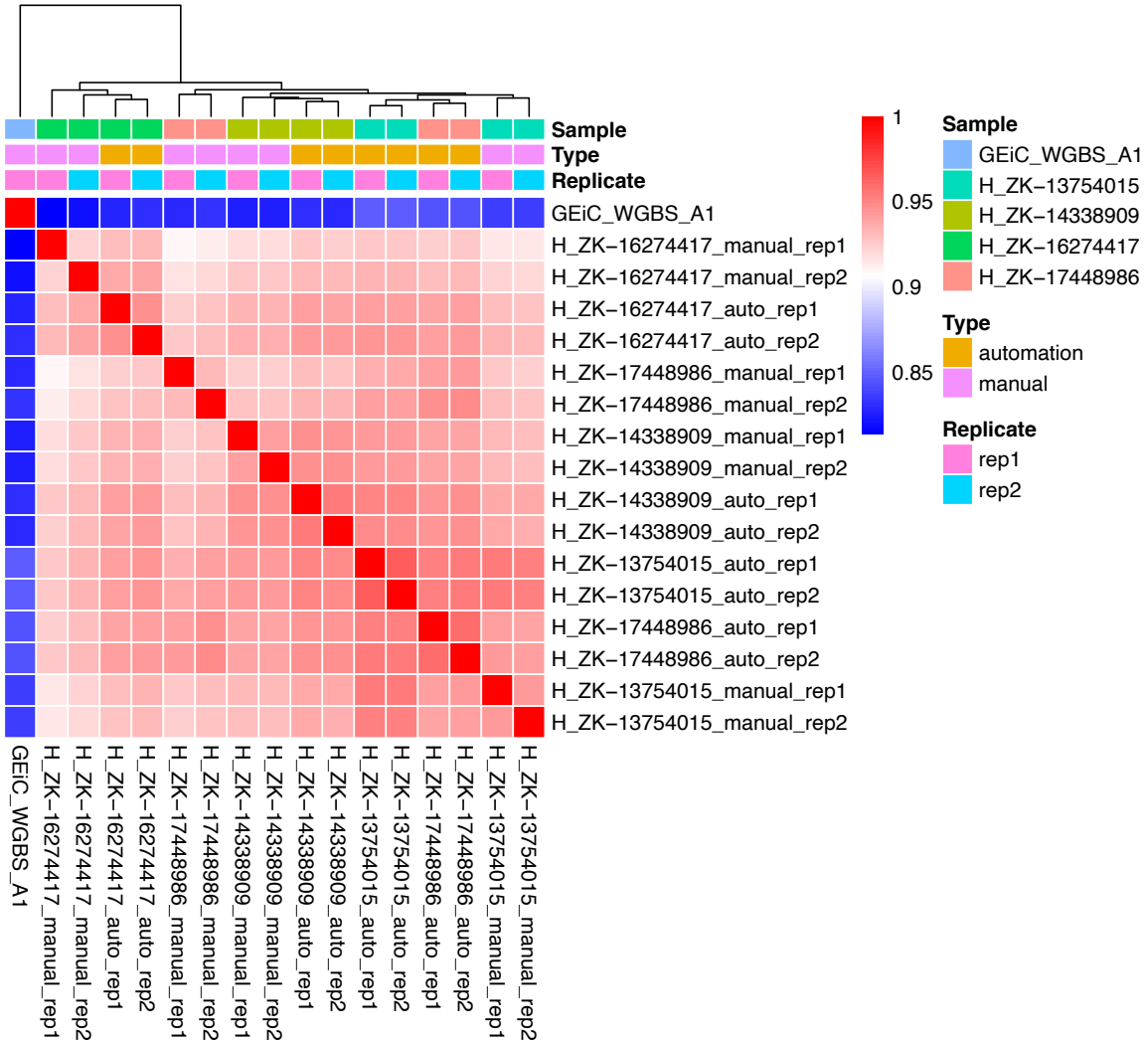


Calculated using all methylation calls

Smoothed CpG methylation values, total 22,308,910 CpGs

PCA and correlation among the WGBS samples

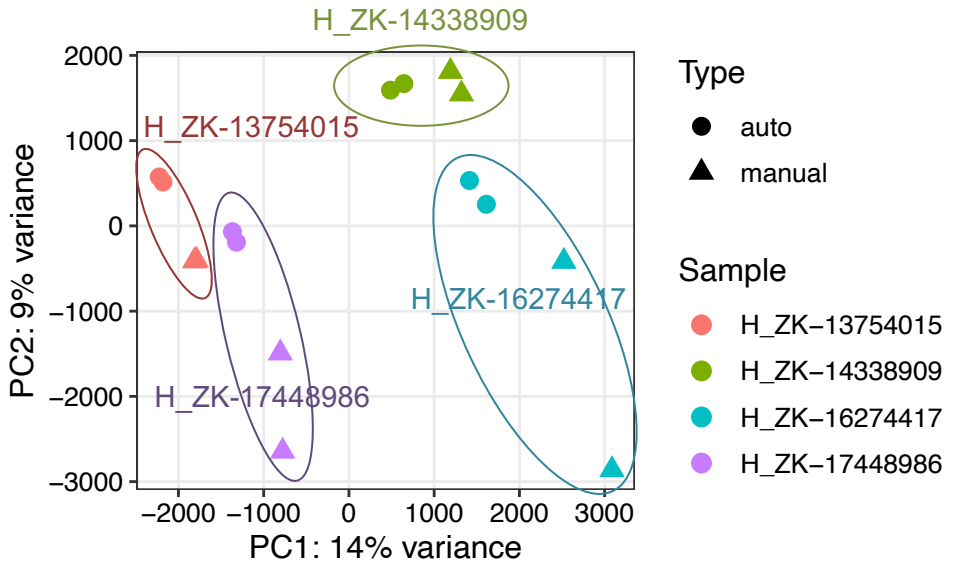


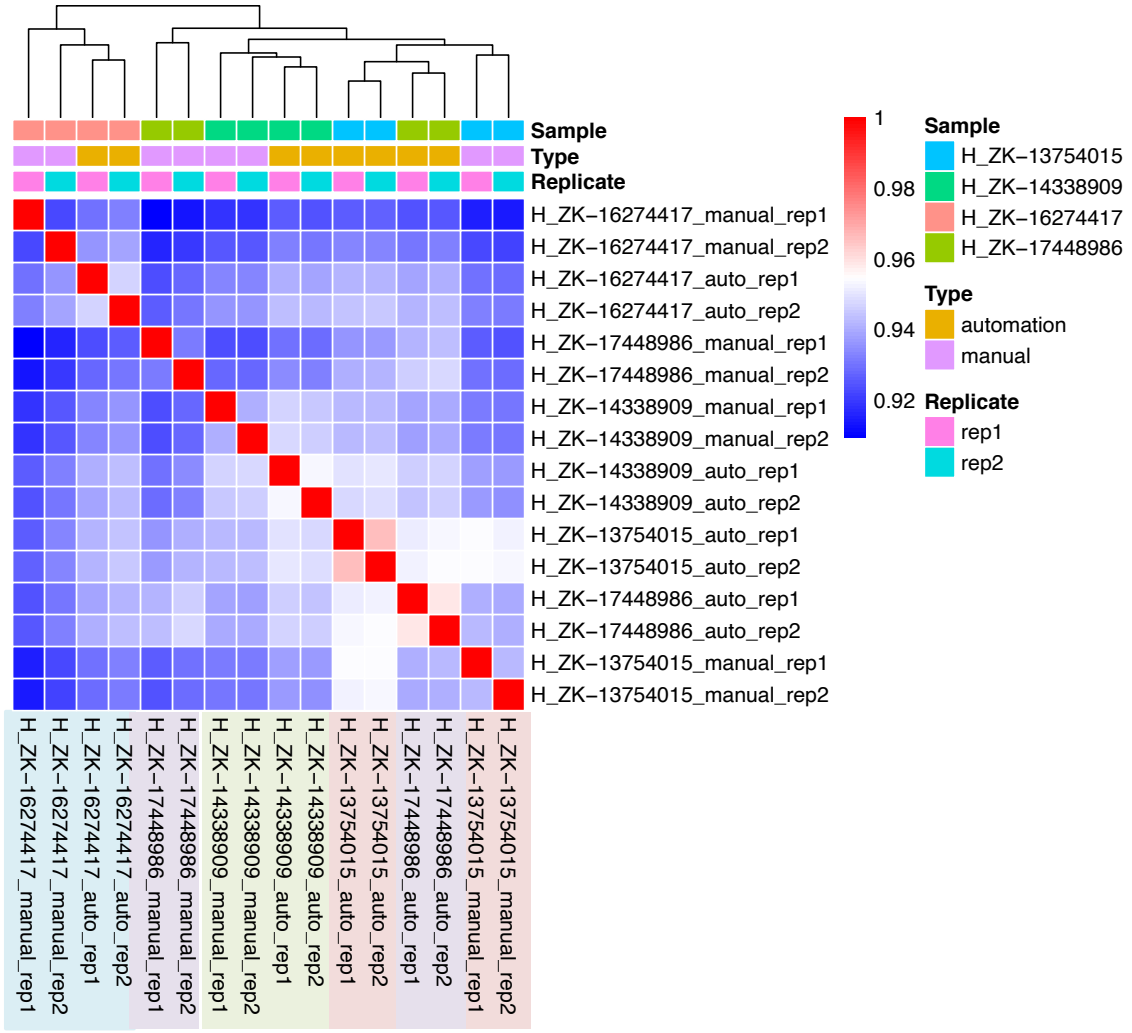


Smoothed methylation values,
total 22,308,910 CpGs

Pearson’s correlation coefficient

PCA and correlation among the WGBS samples

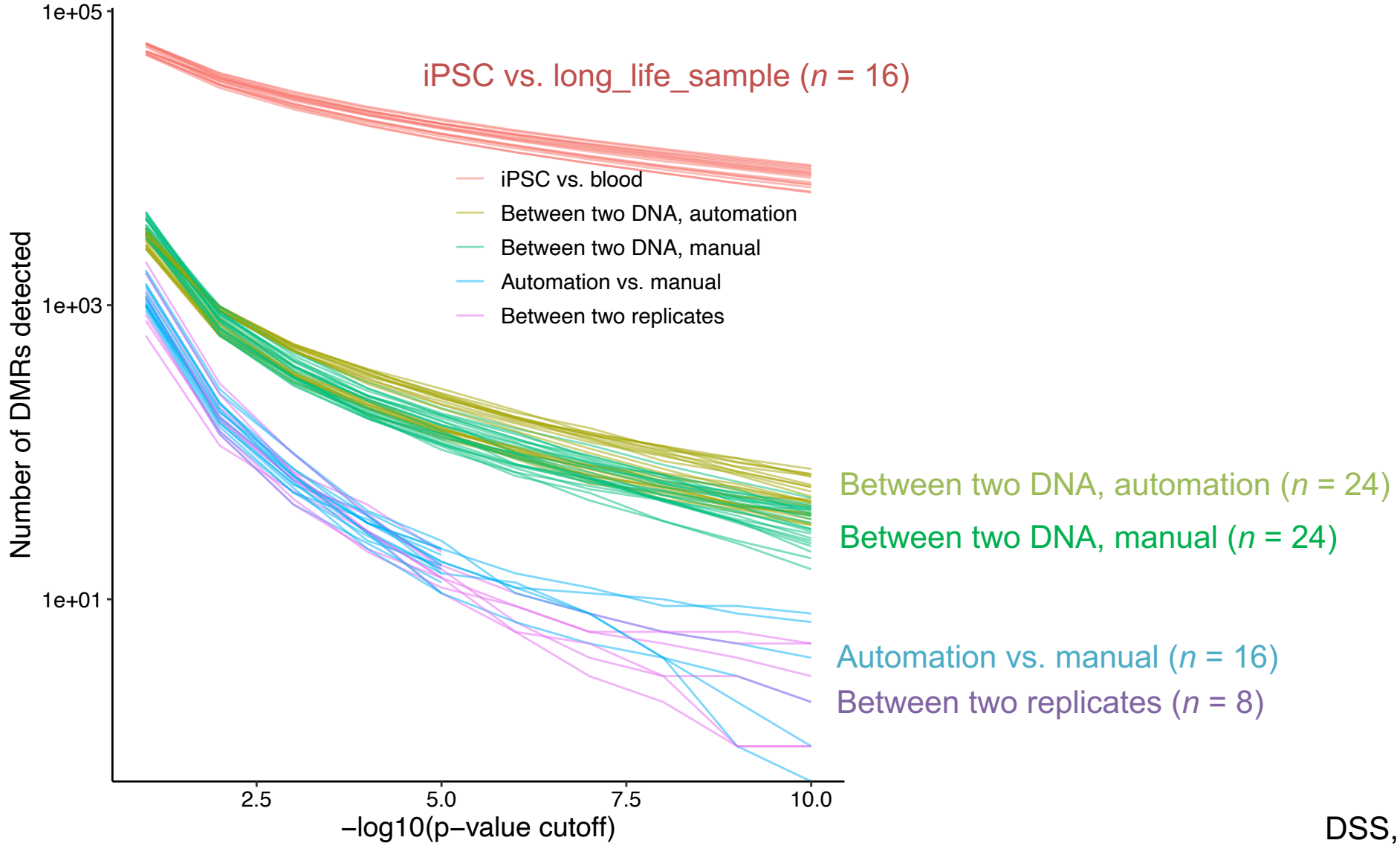




Smoothed methylation values,
total 22,308,910 CpGs

Pearson's correlation coefficient

Number of differentially methylated regions (DMRs)



DSS, delta = 0.2