DNA METHYLATION IN EARLY EMBRYONIC DEVELOPMENT OF MOUSE

by Jude Aneke

This Whole Genome Bisulfite Sequence (WGBS) bioinformatics experiment was conducted on paired layout sequencing data, retrieved from the National Center for Biotechnology Information (NCBI) under the filenames SRR5836475, SRR5836476, SRR3824222, and SRR5836479, originates from research outlined in the article titled "Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer." Employing the Illumina HiSeq 2000 platform with a Bisulfite-Seq strategy, the experiment delves into the genomic intricacies of DNA methylation, shedding light on the epigenetic regulation underlying the transition from embryonic development to disease states like cancer.

***Note:** It is important to start that the mapping for this experiment was limited to **mouse chr18.**

Brief Sample Description

SRR5836475
ICM_rep1_WGBS
ICM_rep1_WGBS; Mus musculus; Bisulfite-Seq
Attributes:
Source name: Inner Cell Mass
Strain: B6D2 F1
Development stage: E3.5
Tissue: Inner Cell Mass

SRR5836476
ICM_rep2_WGBS
ICM_rep2_WGBS; Mus musculus; Bisulfite-Seq
Attributes:
Source name: Inner Cell Mass
Strain: B6D2 F1
Development stage: E3.5
Tissue: Inner Cell Mass

SRR3824222
Epiblast_rep1_WGBS
Epiblast_rep1_WGBS; Mus musculus; Bisulfite-Seq
Attributes:
Source name: proximal Epiblast
Strain: B6D2 F1
Development stage: E6.5
Tissue: proximal Epiblast

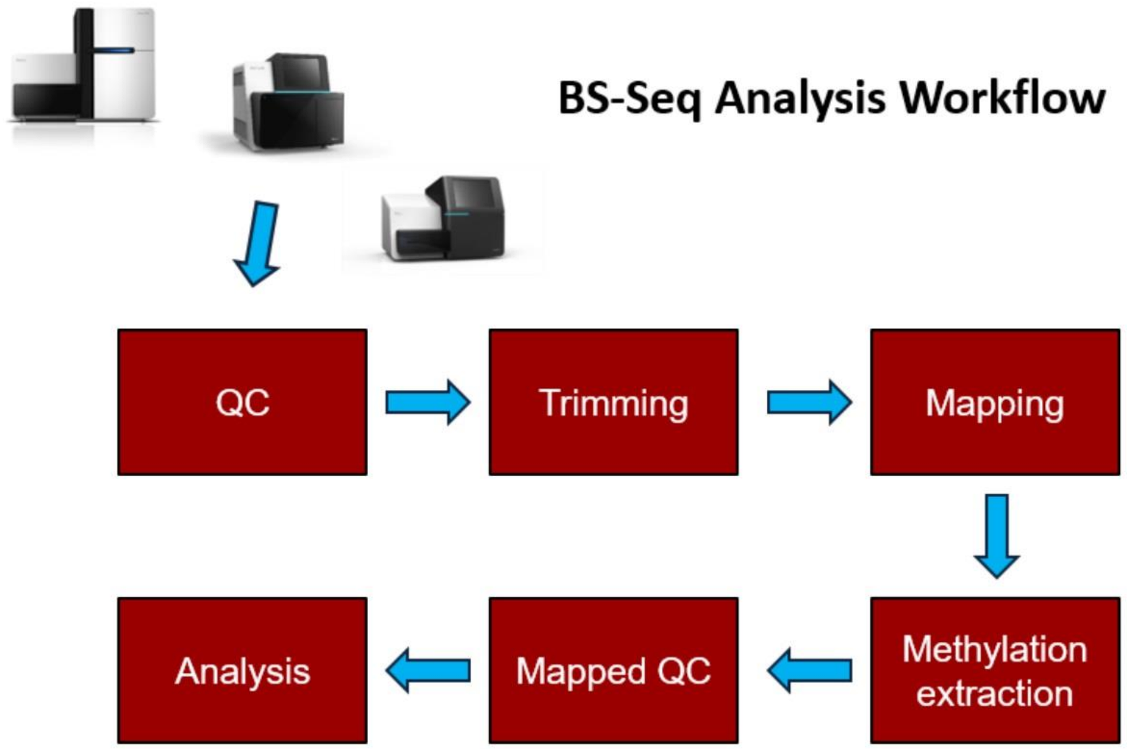
SRR5836479

Epiblast_rep2_WGBS
Epiblast_rep2_WGBS; Mus musculus; Bisulfite-Seq
Source name: proximal Epiblast
Strain: B6D2 F1
Development stage: E6.5
Tissue: proximal Epiblast

For the given assignment I used the above samples to experiment methylation of WGBS data (whole genome bisulfite sequencing) at various stages of mouse embryonic development.

Analysis Workflow

Fig 1a: BS-Seq Analysis Workflow



Source: https://www.bioinformatics.babraham.ac.uk/training/Methylation_Course/BS-Seq

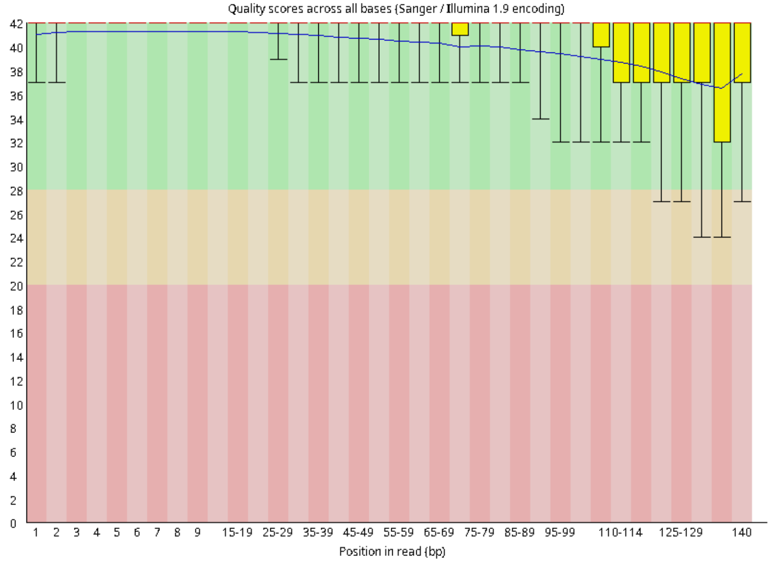
Quality Control Analysis

The fastqc of both sample shows our data is okay. The bimodal distribution in the CG mean values observed in FastQC results of ICM sample (fig 1b) typically indicates the presence of two distinct methylation states within the sample. The bimodal distribution suggests the presence of two distinct populations of CpG sites with different levels of methylation. This could indicate the presence of different cell types or subpopulations within the ICM sample, each exhibiting different methylation patterns.

On the other hand, if the Epiblast (fig 1c) sample does not exhibit a bimodal distribution in the CG mean values, it suggests that the methylation levels across CpG sites are more uniform or homogeneous within this sample. This could indicate a more homogenous cell population or a more consistent methylation pattern across CpG sites within the Epiblast sample compared to the ICM sample. There was no obvious difference in the plots of fig 1b and c after trimming. This pattern was consistent for all replicates.

In bisulfite sequencing, unmethylated cytosines are converted to thymines (T), leading to an increased T content compared to cytosine (C) in the sequenced DNA fragments. This conversion bias results in an elevated T content across the sequence. Additionally, the adenine (A) and guanine (G) content remains unaffected by bisulfite treatment. However, observations (fig 1d) indicate a sharp decline in adenine (A) content towards the end of the sequence. This decline may be attributed to the presence of adapter sequences, such as AGATCGGAAGAGC, which are commonly used in sequencing library preparation. The adapter sequence may introduce biases during sequencing, resulting in variations in base content along the sequenced fragments. The observed patterns reflect the effects of bisulfite treatment and potential adapter biases on the base composition of the sequenced DNA fragments and was observed across the four samples.

Fig 1b: Quality Control of ICM_rep1



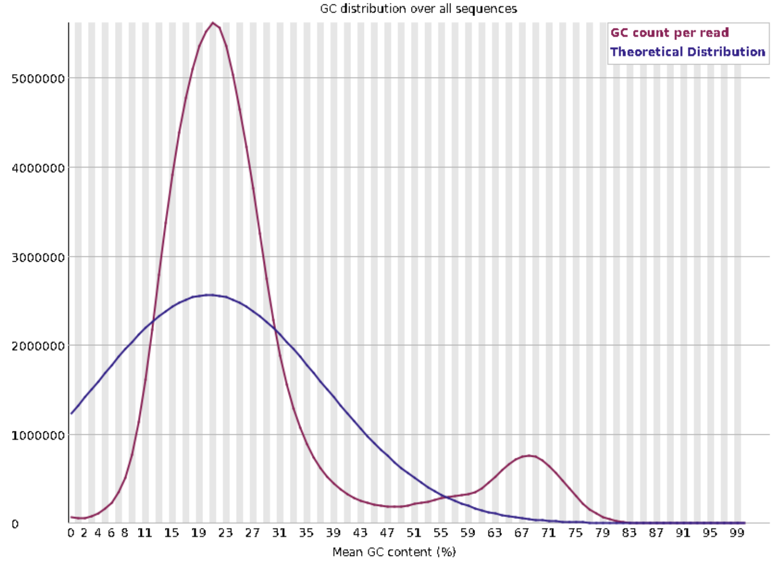
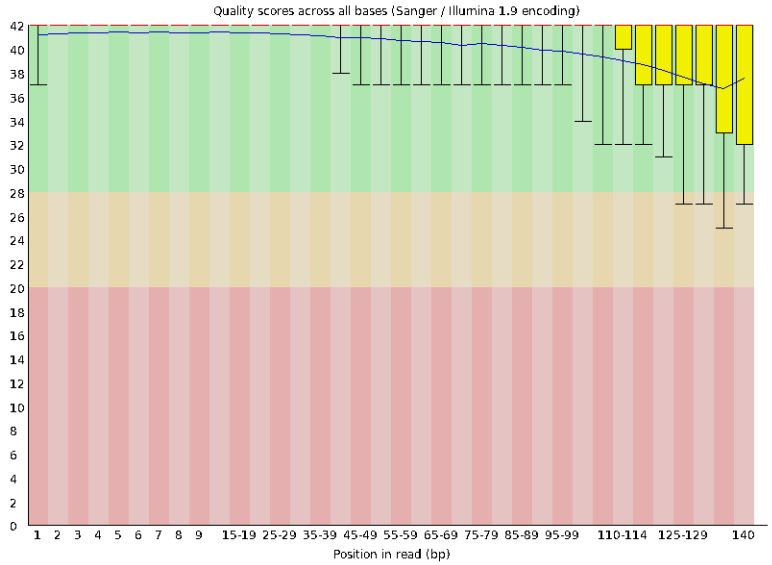


Fig 1c: Quality Control of Epiblast_rep2



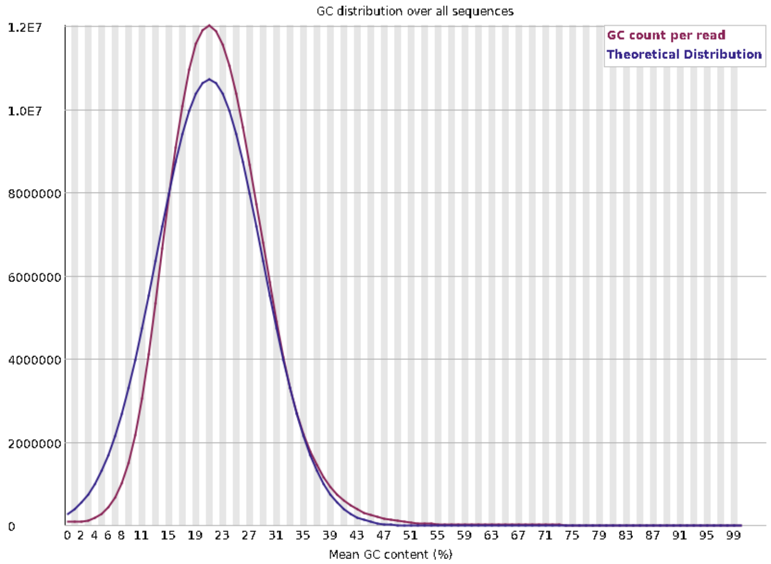


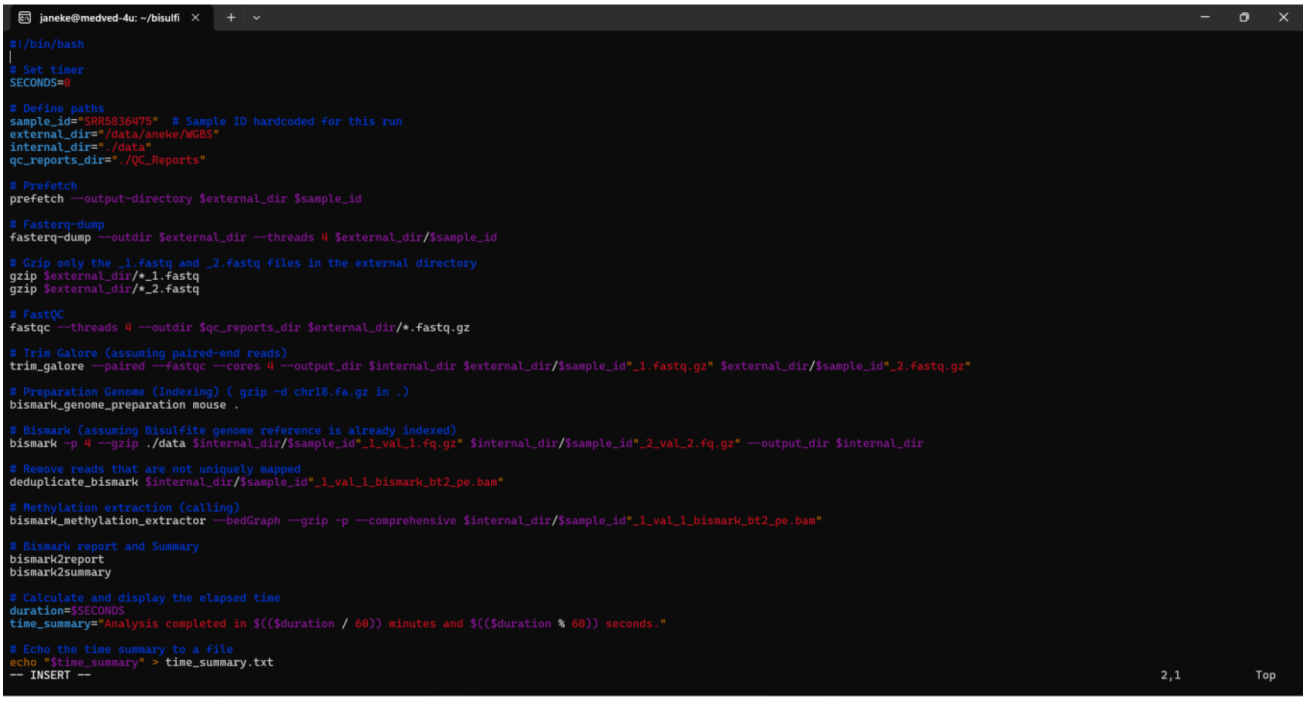
Fig 1d: Per base sequence content of Epiblast_rep1

- ATCG Ratio -> Unmethylated C gets converted to T. Thus, T > C in Bisulfite Sequencing
- ATCG Ratio -> A & G ratio is not affected by Bisulfite Sequencing
- Adapter Sequence -> AGATCGGAAGAGC

Mapping Analysis

The WGBS pipeline in fig 2a used for this experiment is designed to efficiently process bisulfite sequencing data, starting from raw reads retrieval to methylation calling and analysis. Initially, the script prefetches the sequencing data and then utilizes Fasterq-dump to convert it into fastq format, followed by FastQC for quality assessment. Subsequently, Trim Galore is employed for adapter and quality trimming, ensuring high-quality data for downstream analysis. The pipeline then prepares the genome index using Bismark and subsequently aligns the trimmed reads to the bisulfite-converted reference genome. After deduplication of mapped reads, methylation extraction is performed using Bismark, generating bedGraph files for methylation calling. Finally, the pipeline generates comprehensive reports and summaries for methylation profiling. The efficiency of our pipeline is underscored by its robustness and ability to handle large-scale WGBS datasets while providing detailed insights into DNA methylation patterns.

Fig 2a: WGBS Mapping and Methylation Extraction Analysis Pipeline



```
#!/bin/bash

# Set timer
SECONDS=0

# Define paths
sample_id="SRR5836475"  # Sample ID hardcoded for this run
external_dir="/data/aneke/WGBS"
internal_dir="./data"
qc_reports_dir="./QC_Reports"

# Prefetch
prefetch --output-directory $external_dir $sample_id

# Fasterq-dump
fasterq-dump --outdir $external_dir --threads 4 $external_dir/$sample_id

# Gzip only the _1.fastq and _2.fastq files in the external directory
gzip $external_dir/*_1.fastq
gzip $external_dir/*_2.fastq

# FastQC
fastqc --threads 4 --outdir $qc_reports_dir $external_dir/*.fastq.gz

# Trim Galore (assuming paired-end reads)
trim_galore --paired --fastqc --cores 4 --output_dir $internal_dir $external_dir/$sample_id"_1.fastq.gz" $external_dir/$sample_id"_2.fastq.gz"

# Preparation Genome (Indexing) ( gzip -d chr18.fa.gz in .)
bismark_genome_preparation mouse .

# Bismark (assuming Bisulfite genome reference is already indexed)
bismark -p 4 --gzip ./data $internal_dir/$sample_id"_1_val_1.fq.gz" $internal_dir/$sample_id"_2_val_2.fq.gz" --output_dir $internal_dir

# Remove reads that are not uniquely mapped
deduplicate_bismark $internal_dir/$sample_id"_1_val_1_bismark_bt2_pe.bam"

# Methylation extraction (calling)
bismark_methylation_extractor --bedGraph --gzip -p --comprehensive $internal_dir/$sample_id"_1_val_1_bismark_bt2_pe.bam"

# Bismark report and Summary
bismark2report
bismark2summary

# Calculate and display the elapsed time
duration=$SECONDS
time_summary="Analysis completed in $(($duration / 60)) minutes and $(($duration % 60)) seconds."

# Echo the time summary to a file
echo "$time_summary" > time_summary.txt
-- INSERT --
```

Fig 2b: Bismark Report for ICM_rep1

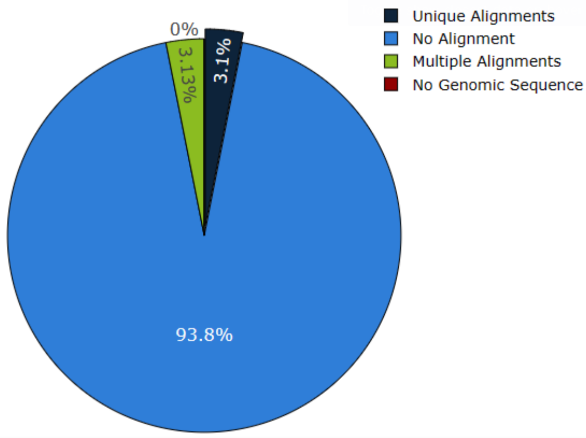
Bismark Processing Report

SRR5836475_1_val_1.fq.gz and SRR5836475_2_val_2.fq.gz

Data processed at 14:05:38 on 2024-04-16

Alignment Stats

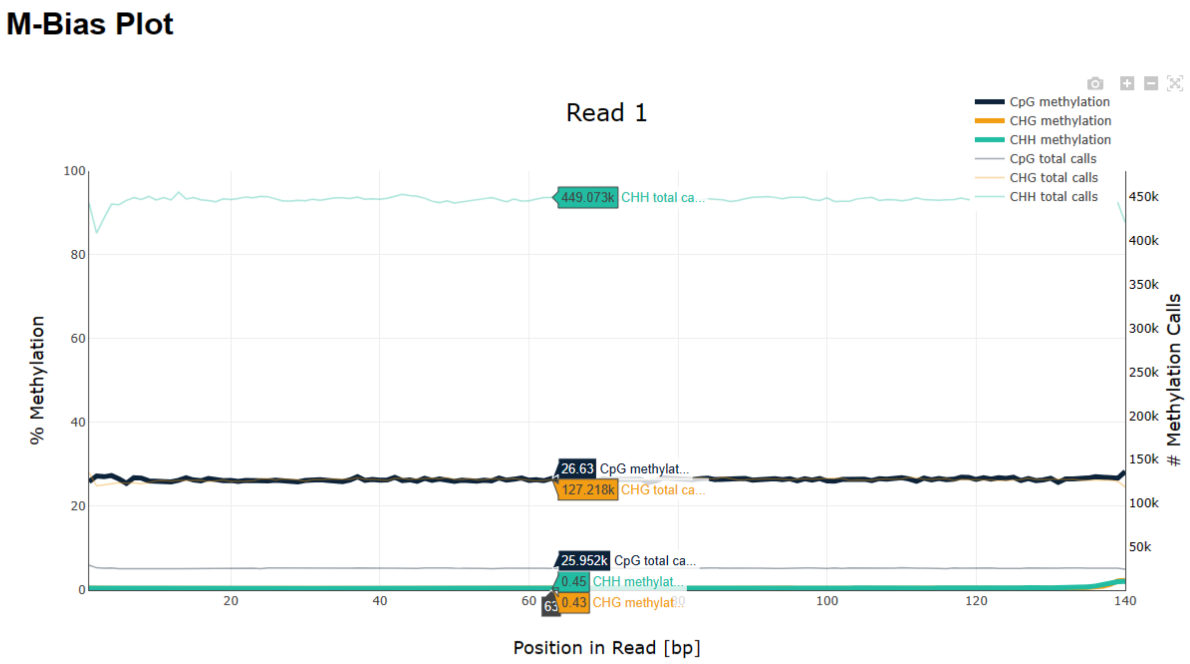
Sequence pairs analysed in total	108230174
Paired-end alignments with a unique best hit	3353259
Pairs without alignments under any condition	101493043
Pairs that did not map uniquely	3383872
Genomic sequence context not extractable (edges of chromosomes)	0



In Fig 2b, representing the DNA methylation profile of ICM_rep1, a Bismark report was generated to assess the bisulfite sequencing results. A total of 108,230,174 sequence pairs were analyzed, with a mapping efficiency of 3.1%. Among these, 3,353,259 pairs had a unique best-hit alignment. Notably, 101,493,043 pairs did not align under any condition, and 3,383,872 pairs did not map uniquely. Alignment details revealed that the majority of unique best-hit alignments originated from the converted top and bottom strands, with no alignments rejected from theoretical complementary strands.

Regarding cytosine methylation, a total of 189,298,124 C's were analyzed. Methylation levels varied across different sequence contexts, with 28.7% methylated C's in the CpG context, 0.7% in the CHG context, and 0.8% in the CHH context. Additionally, 9.5% of C's in the unknown context (CN or CHN) were methylated. These findings provide insights into the methylation patterns of ICM_rep1, highlighting substantial methylation in CpG sites compared to other contexts.

Fig 2c: Methylation Bias Plot for ICM_rep1 (Read 1)



The bias plot you sent is an M-Bias plot, which is a quality control metric used in Whole Genome Bisulfite Sequencing (WGBS) analysis. It helps assess the accuracy of methylation calling at different positions within a sequencing read 1 (fig 2c and f).

In your plot, the x-axis represents the position in read 1 (bp), and the y-axis represents the number of methylation calls (reads). The different lines represent different methylation types:

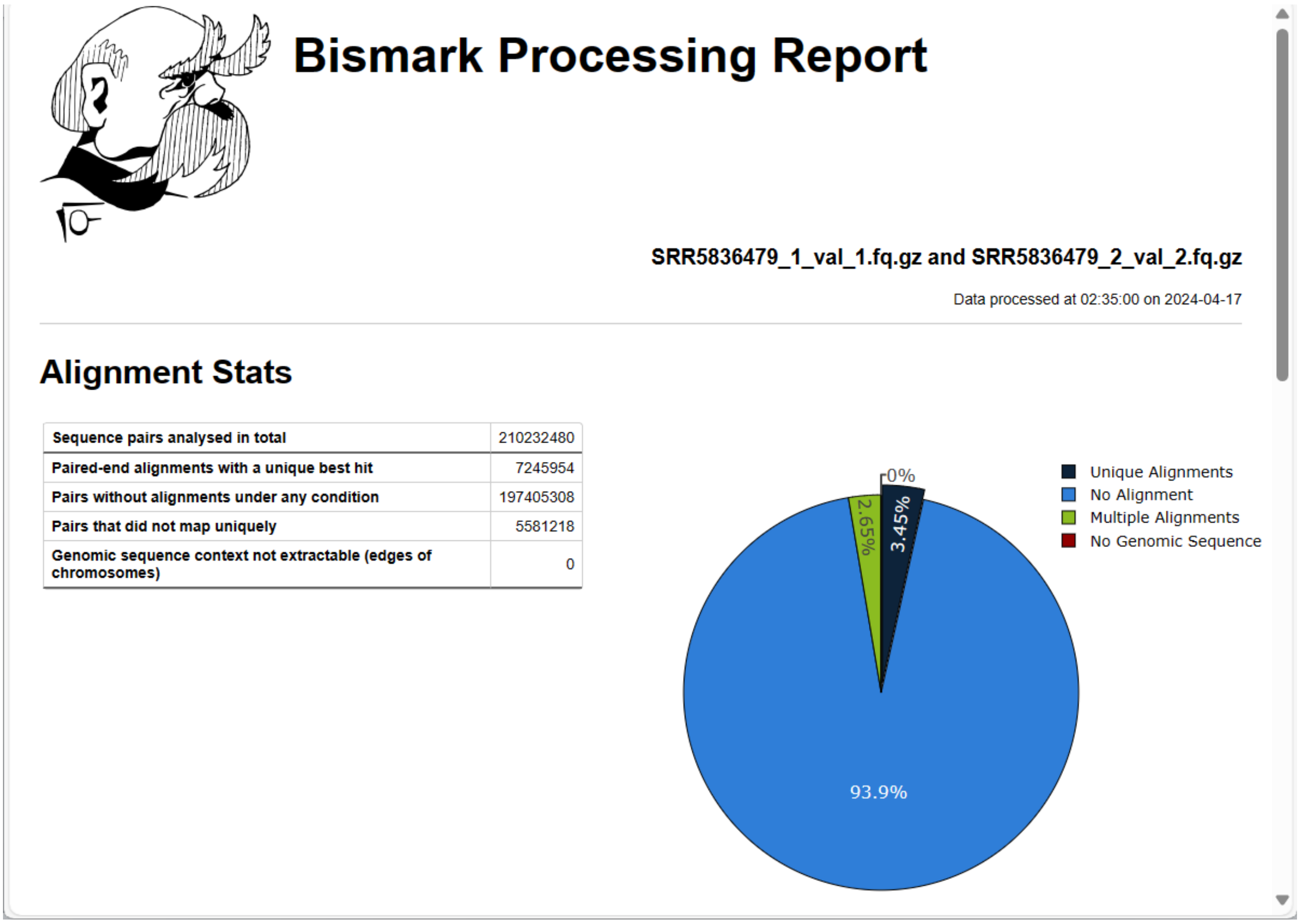
- CpG methylation (blue)
- CHG methylation (green)
- CHH methylation (red)

The plot shows a clear bias towards the ends of the reads (towards 0 bp and 140 bp). This is because the bisulfite conversion process, which is a crucial step in WGBS, can be inefficient at the ends of reads 2 (fig 2d and g). This can lead to an underestimation of methylation levels at the ends.

While all three variations (CpG, CHG, and CHH) represent cytosine followed by another nucleotide, CpG is typically more frequent in mammals due to its role in gene regulation. Conversely, plants tend to have more CHG and CHH methylation, possibly as a defense mechanism.

Fig 2d: Methylation Bias Plot for ICM_rep1 (Read 2)

Fig 2e: Bismark Report for Epiblast_rep2



Bismark Processing Report

SRR5836479_1_val_1.fq.gz and SRR5836479_2_val_2.fq.gz

Data processed at 02:35:00 on 2024-04-17

Alignment Stats

Sequence pairs analysed in total	210232480
Paired-end alignments with a unique best hit	7245954
Pairs without alignments under any condition	197405308
Pairs that did not map uniquely	5581218
Genomic sequence context not extractable (edges of chromosomes)	0

For Epiblast_rep2 (fig, the Bismark analysis was conducted on a total of 210,232,480 sequence pairs. Out of these, 7,245,954 pairs had a unique best alignment, resulting in a mapping efficiency of 3.4%. The majority of the aligned pairs originated from the top and bottom strands, with 3,629,116 pairs aligning to the top strand and 3,616,838 pairs aligning to the bottom strand.

Regarding cytosine methylation, a total of 406,778,142 C's were analyzed. In the CpG context, 14,096,470 C's were methylated, accounting for 79.3% of the total CpG sites analyzed. In the CHG context, 1,846,083 C's were methylated (2.1%), and in the CHH context, 4,562,983 C's were methylated (1.5%). Additionally, 32,116 C's were methylated in an unknown context. The overall methylation patterns indicate a high level of methylation in CpG sites compared to CHG and CHH contexts, consistent with expected methylation patterns in bisulfite sequencing data.

Fig 2f: Methylation Bias Plot for Epiblast_rep2 (Read 1)

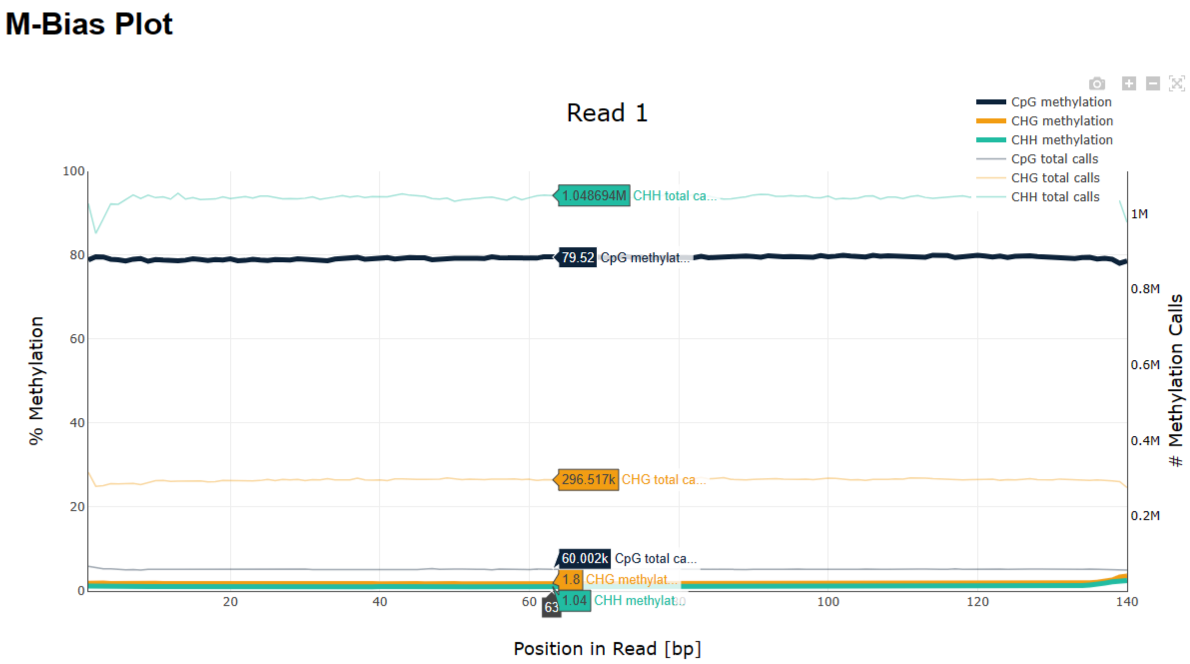


Fig 2f: Methylation Bias Plot for Epiblast_rep2 (Read 2)

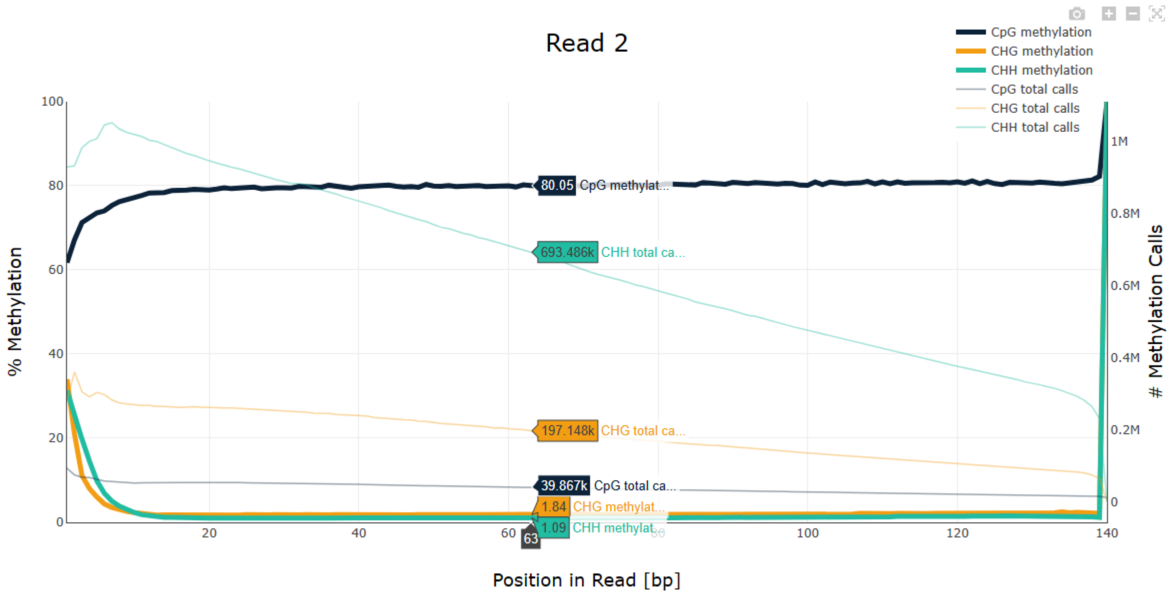
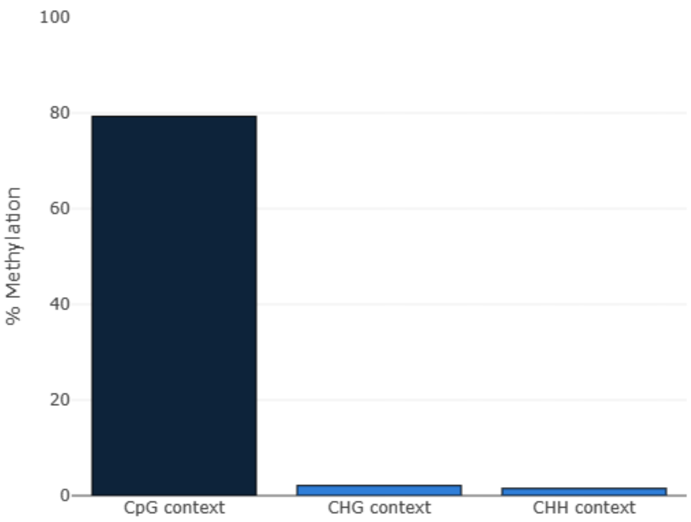


Fig 2g: Cytosine Methylation of Epiblast_rep2 before (left) and after (right) extraction



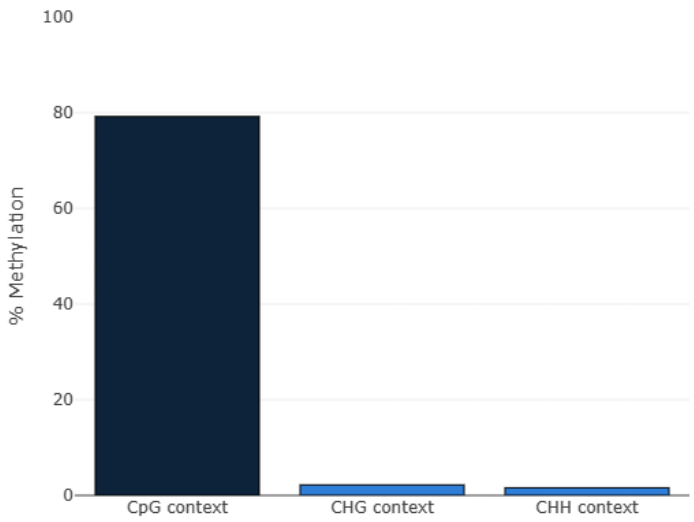
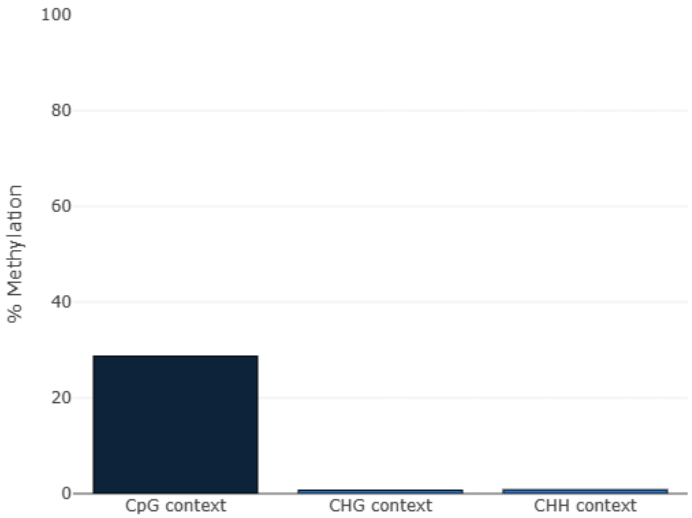
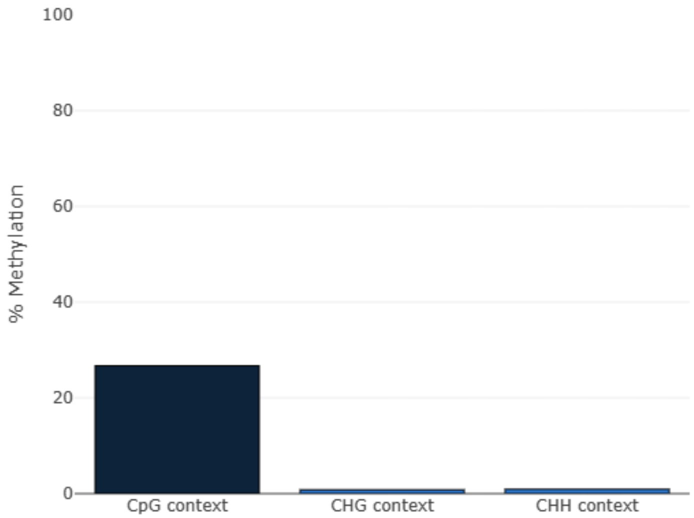


Fig 2h: Cytosine Methylation of ICM_rep1 before (left) and after (right) extraction





For ICM_rep1 (fig h), before extraction, a total of 189,298,124 Cs were analyzed, with 28.7% methylated in CpG context, 0.7% in CHG context, and 0.8% in CHH context. After extraction, the total Cs analyzed reduced to 134,018,256, and the percentage of methylation slightly decreased to 26.7% in CpG context, while the percentages in CHG and CHH contexts remained relatively stable.

For Epiblast_rep2 (fig g), before extraction, a total of 406,778,142 Cs were analyzed, with a high methylation percentage of 79.3% in CpG context, 2.1% in CHG context, and 1.5% in CHH context. After extraction, the total Cs analyzed reduced to 318,051,823, with the methylation percentages in CpG, CHG, and CHH contexts showing minimal changes.

Before extraction, both samples exhibited distinct methylation patterns, with Epiblast_rep2 showing higher methylation levels compared to ICM_rep1. After extraction, both samples experienced a reduction in the total Cs analyzed, with minimal changes in the methylation percentages. However, Epiblast_rep2 maintained a higher overall methylation level compared to ICM_rep1 even after extraction, indicating differences in the methylation dynamics between the two samples (fig g and h).

Fig 3a: WGBS Mapped QC Pipeline

janeke@medved-4u: /data/an

```bash
#!/bin/bash

# Define the basename
BASENAME="Epiblast_rep2"

# Step 1: Convert BAM file to DNMTools format
dnmtools format -f bismark -t 4 -B SRR5836479_1_val_1_1_bismark_bt2_pe.deduplicated.bam "${BASENAME}_WGBS_format.bam"

# Step 2: Sort the formatted BAM file
samtools sort -o "${BASENAME}_WGBS_sorted.bam" "${BASENAME}_WGBS_format.bam"

# Step 3: Calculate bisulfite conversion rates
dnmtools bsrate -c ~/bisulfite_WGBS/data/chr18.fa -o "${BASENAME}_WGBS_bsrate"

# Step 4: Generate methylation counts
dnmtools counts -c ~/bisulfite_WGBS/data/chr18.fa -o "${BASENAME}_WGBS_counts.meth" "${BASENAME}_WGBS_sorted.bam"

# Step 5: Perform methylation level analysis
dnmtools level -o "${BASENAME}_WGBS.levels" "${BASENAME}_WGBS_counts"

# Step 6: Generate methylation counts for CpG sites only
dnmtools counts -cpg-only -c ~/bisulfite_WGBS/data/chr18.fa -o "${BASENAME}_WGBS_CpG.meth" "${BASENAME}_WGBS_sorted.bam"

# Step 7: Generate symmetric CpG methylation
dnmtools sym -o "${BASENAME}_WGBS_symmetric_CpG.meth" "${BASENAME}_WGBS-CpG.meth"

# Step 8: Filter out CpGx sites
awk '$4 != "CpGx"' "${BASENAME}_WGBS_symmetric_CPG.meth" > "${BASENAME}_WGBS_symmetric_CPG_filtered.meth"

echo "Pipeline completed!"
"scripts/Mapped_QC-pipeline.sh" 44L, 1637C                                          1,1           Top
```

This bash script (fig 3a) outlines a pipeline for quality control and analysis of bisulfite sequencing data from an Epiblast_rep2 sample. The process begins by converting the aligned BAM file to DNMTools format, followed by sorting the formatted file. Bisulfite conversion rates are then calculated, and methylation counts are generated for subsequent analysis. Methylation levels are assessed, with a specific focus on CpG sites. Symmetric CpG methylation is generated, and CpGx sites are filtered out. Overall, this pipeline ensures thorough quality control and comprehensive analysis of the Epiblast_rep2 WGBS data, providing valuable insights into the methylation landscape of the sample.

Fig 3b: Bisulfite Conversion Rate Analysis (ICM_rep2)

ICM_rep2_WGBS_sorted.bsrate

File Edit View

OVERALL CONVERSION RATE = 0.99185
POS CONVERSION RATE = 0.991592 60276638
NEG CONVERSION RATE = 0.992109 60174966

BASE	PTOT	PCONV	PRATE	NTOT	NCONV	NRATE	BTHTOT	BTHCONV	BTHRATE	ERR	ALL	ERRRATE
1	274396	273070	0.99517	230111	165171	0.71779	504507	438241	0.86865	2675	507182	0.00527
2	252018	250818	0.99524	254725	206019	0.80879	506743	456837	0.90152	2315	509058	0.00455
3	260829	259712	0.99572	259465	225620	0.86956	520294	485332	0.93280	2140	522434	0.00410
4	268612	267492	0.99583	267029	244434	0.91538	535641	511926	0.95573	2108	537749	0.00392
5	268288	267187	0.99590	275539	260893	0.94685	543827	528080	0.97104	1800	545627	0.00330
6	270452	269371	0.99600	284416	274782	0.96613	554868	544153	0.98069	1818	556686	0.00327
7	272528	271336	0.99563	285554	278718	0.97606	558082	550054	0.98562	1872	559954	0.00334
8	270553	269408	0.99577	281003	276110	0.98259	551556	545518	0.98905	1935	553491	0.00350
9	273791	272633	0.99577	279529	276098	0.98773	553320	548731	0.99171	1893	555213	0.00341
10	272409	271278	0.99585	278946	276302	0.99052	551355	547580	0.99315	1919	553274	0.00347

Ln 4, Col 60 42,191 characters 100% Unix (LF) UTF-8

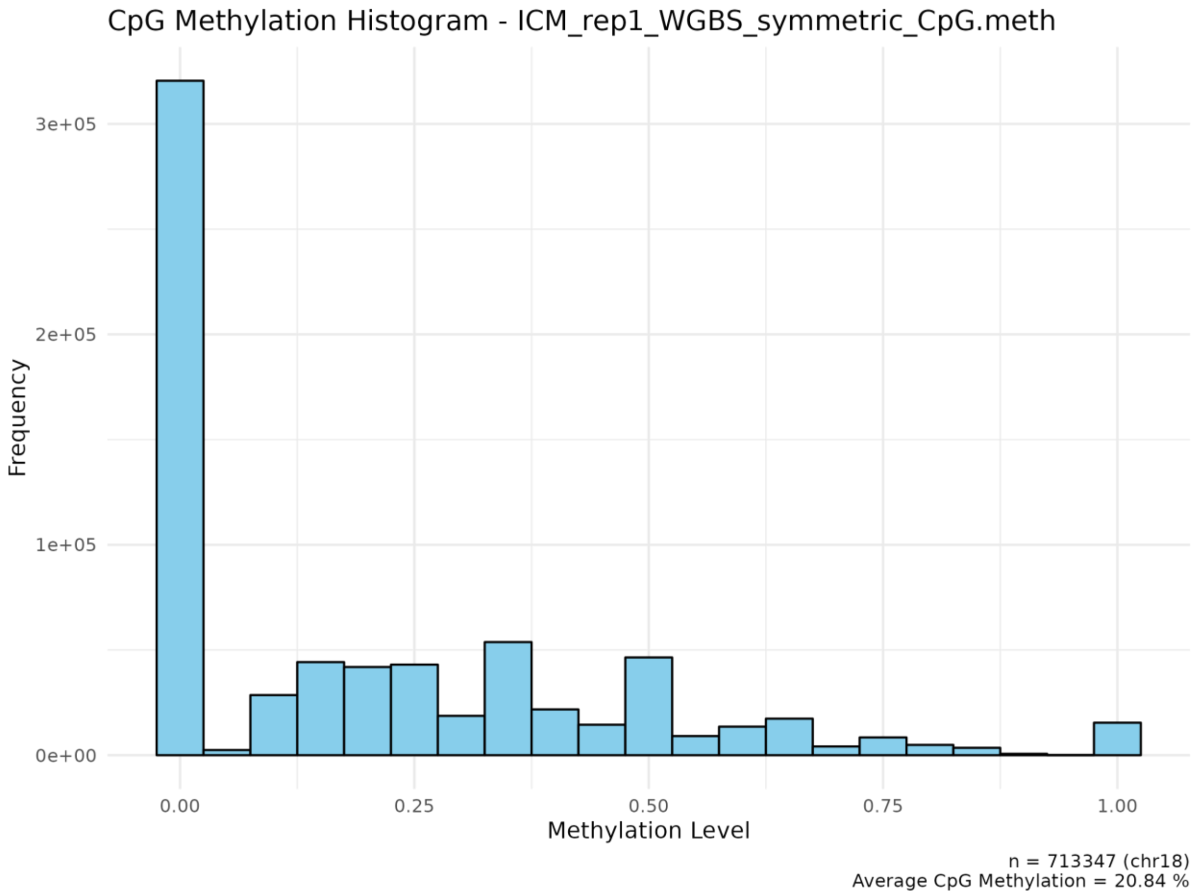
The bsrate command in dnmtools is designed to estimate the bisulfite conversion rate, which reflects the success of the bisulfite treatment in converting unmethylated cytosines to thymines. This rate is crucial for accurate DNA methylation analysis, as it ensures that unmethylated cytosines are correctly identified as converted thymines during sequencing. The output of bsrate includes positive and negative conversion rates, representing the rates of conversion on the forward and reverse DNA strands, respectively. Each line of the output corresponds to a genomic position, with columns indicating the total number of nucleotides analyzed, the number of converted cytosines, and the conversion rate. Additionally, the output provides information on sequencing error rates at each position. It's essential to note that the bisulfite conversion rate should be very high (e.g., > 0.98) for reliable methylation analysis. The labels PTOT, PCONV and PRATE give the total nucleotides used, the number converted, and the ratio of those two, for the positive-strand mappers. The corresponding numbers are also given for negative strand mappers (NTOT, NCONV, NRATE) and combined (BTH). The sequencing error rate is also shown for each position, though this is an underestimate because we assume at these genomic sites any read with either a C or a T contains no error.

All the four samples had an overall bsrate > 0.98 which shows our mapping analysis is reliable.

Fig 3c: Code for Histogram Plot

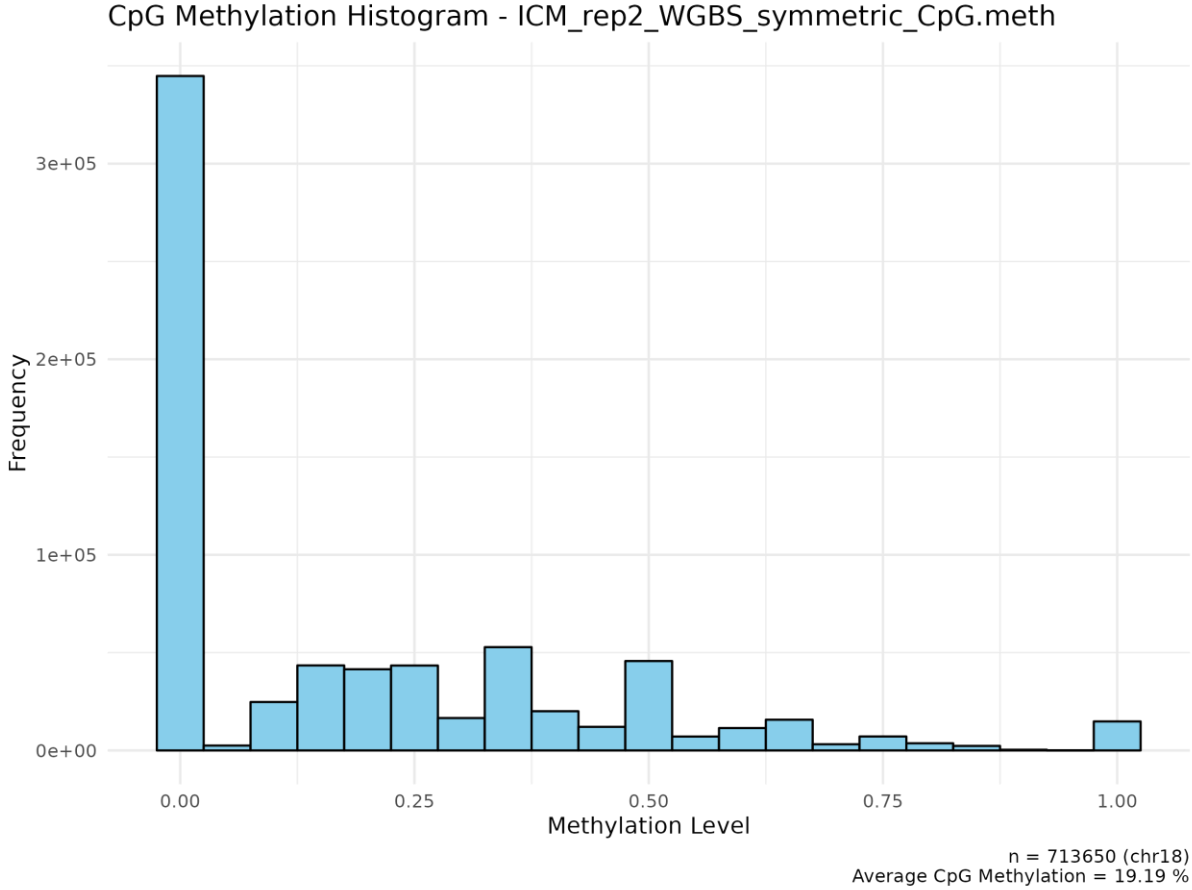
```
janeke@medved-4u: /data/ar

# Load required library
library(ggplot2)

# Read methylation data from the .symmetric_CpGs.meth file
file_path <- "SRR5836475/ICM_rep1_WGBS_symmetric_CpG_filtered.meth"
methylation_data <- read.table(file_path, header = FALSE, sep = '\t',
                               col.names = c('chromosome', 'position', 'strand',
                                             'sequence_context', 'methylation_level', 'read_count'))

# Filter out mutated CpG sites (CpGx)
methylation_data <- methylation_data[methylation_data$sequence_context != "CpGx", ]

# Calculate total number of CpGs
total_cpgs <- nrow(methylation_data)

# Calculate average CpG methylation in percentage
avg_methylation <- mean(methylation_data$methylation_level, na.rm = TRUE) * 100

# Create histogram plot
p <- ggplot(methylation_data, aes(x = methylation_level)) +
  geom_histogram(binwidth = 0.05, fill = 'skyblue', color = 'black') +
  labs(title = paste('CpG Methylation Histogram -', "ICM_rep1_WGBS_symmetric_CpG.meth"),
       x = 'Methylation Level', y = 'Frequency',
       caption = paste('n =', total_cpgs, '(chr18)\n',
                       'Average CpG Methylation =', round(avg_methylation, 2), '%')) +
  theme_minimal()

# Save plot as PNG using ggsave
ggsave("ICM_rep1_WGBS_methylation_histogram.png", plot = p, width = 8, height = 6, dpi = 300)

                                                                          30,0-1        All
```

Fig 3d: CpG Methylation Level of ICM_rep1



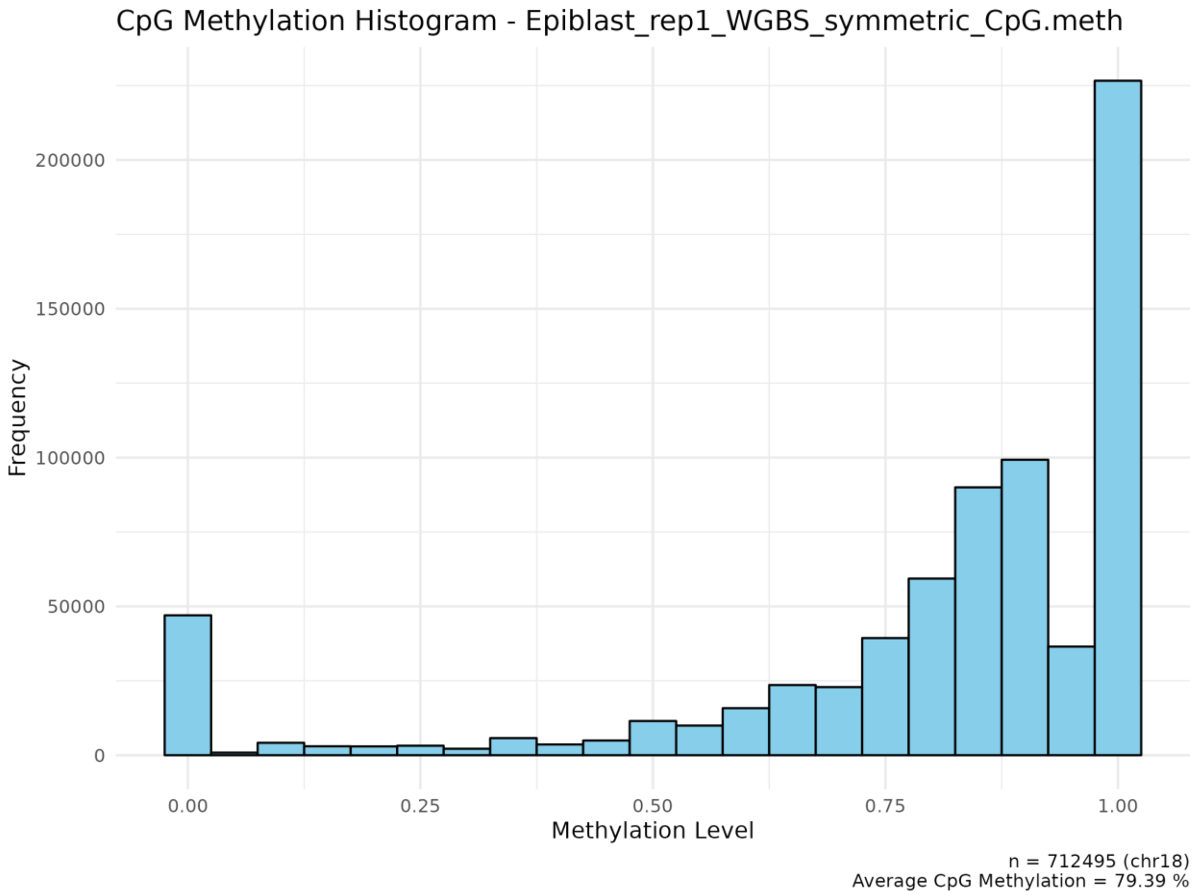
The CpG methylation histogram (fig 3d) visualizes the distribution of methylation levels across all CpG sites in the ICM_rep1_WGBS_symmetric_CpG.meth sample. The analysis reveals an average methylation level of 20.84%, with a majority of CpG sites exhibiting methylation levels between 0% and 50%. The histogram's rightward skew suggests a potential bias towards lower methylation levels, possibly due to bisulfite conversion inefficiencies during library preparation.

Fig 3e: CpG Methylation Level of ICM_rep2



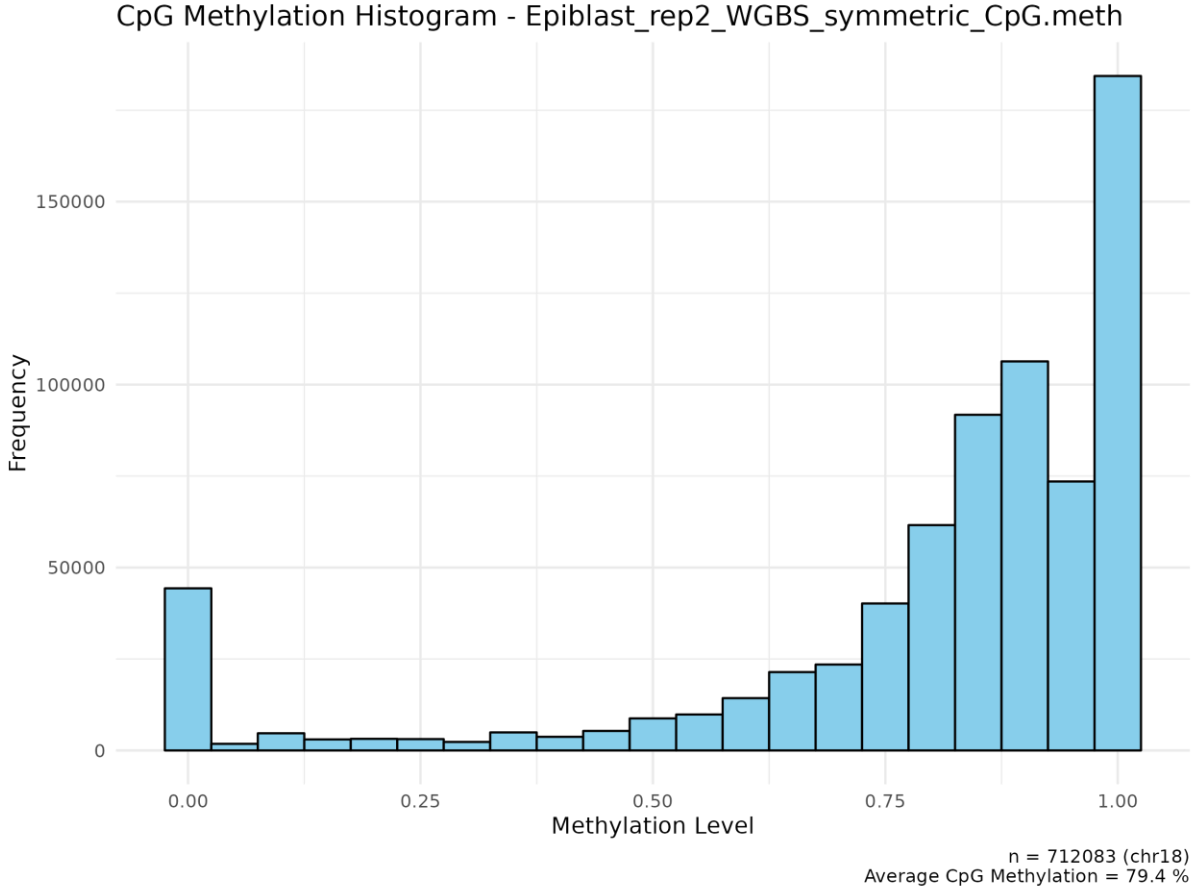
The CpG methylation histogram for ICM_rep2_WGBS_symmetric_CpG.meth shows an average methylation level of 19.19%, with a majority of CpG sites exhibiting levels between 0% and 50%. Similar to the first replicate (ICM_rep1), the histogram's rightward skew suggests a potential bias towards lower methylation levels, possibly due to bisulfite conversion inefficiencies during library preparation (fig 3e).

Fig 3f: CpG Methylation Level of Epiblast_rep1



The CpG methylation histogram for Epiblast_rep1_WGBS_symmetric_CpG.meth reveals a contrasting methylation pattern compared to the ICM samples. Here, the average methylation level is significantly higher at 79.39%, with a majority of CpG sites exhibiting methylation levels between 50% and 100% (fig 3f).

Fig 3g: CpG Methylation Level for Epiblast_rep2



The CpG methylation histogram for Epiblast_rep2_WGBS_symmetric_CpG.meth shows an average methylation level of 79.4%, with a majority of CpG sites exhibiting levels between 50% and 100%. Similar to Epiblast_rep1, the histogram's leftward skew suggests a potential enrichment for higher methylation levels compared to ICM (fig 3g).

The CpG methylation histograms reveal distinct methylation patterns between ICM and Epiblast samples. Epiblast (rep1 and rep2) exhibits a significantly higher average methylation level (around 79%) compared to ICM (around 20%) . This difference is reflected in the histogram shapes, with Epiblast enriched for sites with higher methylation levels (50% to 100%) and ICM enriched for sites with lower methylation levels (0% to 50%). These observations suggest global differences in DNA methylation between these mouse embryonic developmental stages.

Fig 4: Visualization of methylation.bigWig and coverage.bigWig using IGV

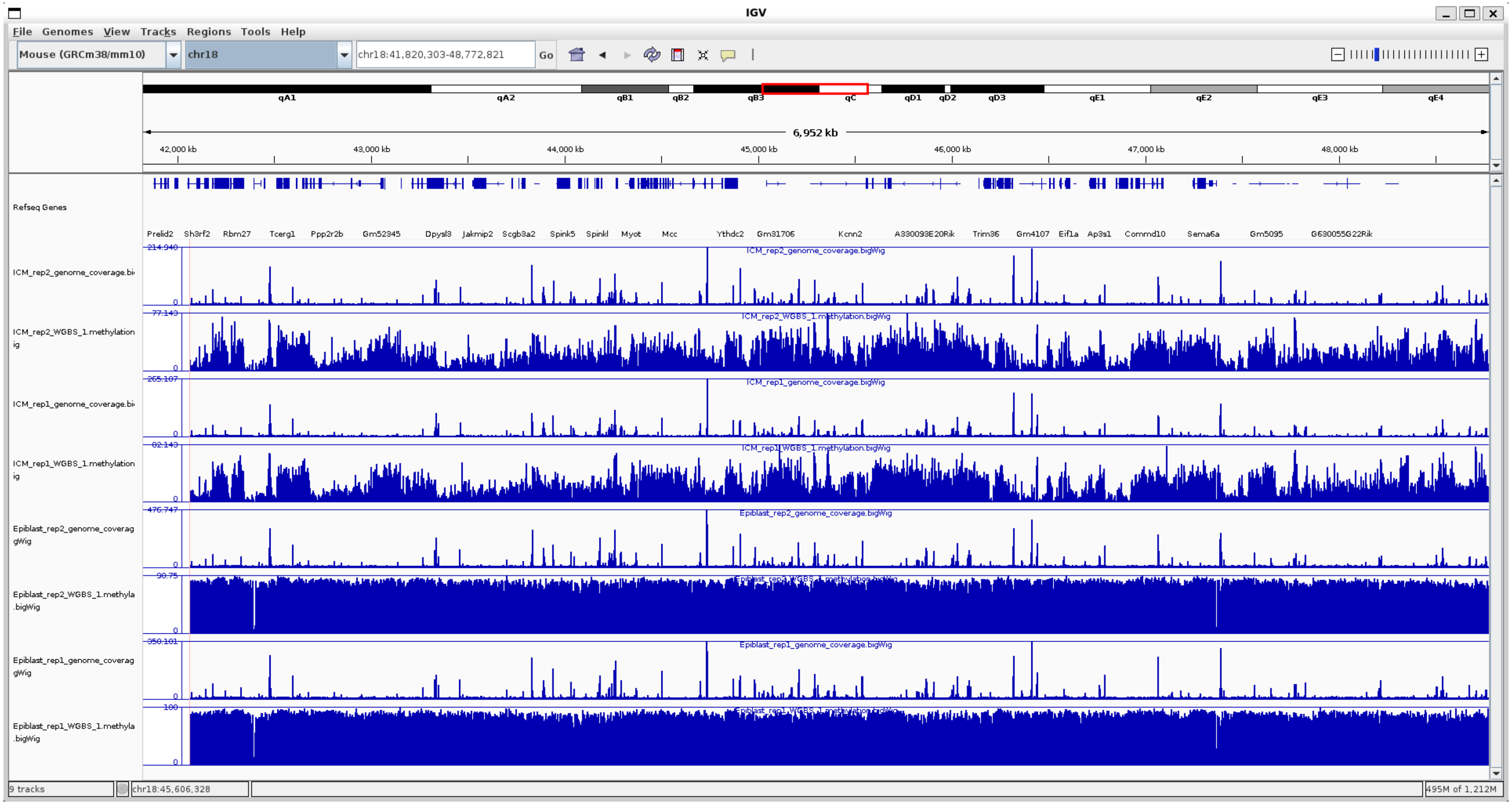


Figure 4 presents the visualization of methylation.bigWig and coverage.bigWig files using the Integrative Genomics Viewer (IGV). This graphical representation displays all four samples along with their methylation and coverage profiles. Through IGV, diverse genes and chromosome 18 (chr18) can be observed, providing insight into the distribution of methylation levels and coverage across genomic regions. The visualization enables researchers to examine methylation patterns and coverage depth in specific genomic loci, aiding in the identification of regions of interest and facilitating comprehensive analysis of epigenetic modifications.

Differential Methylation Regions (DMRs) Analysis

For the differential methylation region (DMR) analysis, dnmtools was chosen due to its specialized features tailored for DNA methylation analysis, user-friendly interface, comprehensive functionality and swift online support. This tool offers various functions, including merging methylation data, identifying DMRs, calculating methylation levels, and generating summary statistics, streamlining the analysis workflow. Additionally, dnmtools allows for customization to accommodate diverse experimental setups, and it benefits from community support and documentation. Notably, when encountering inconsistencies with the *sym* function and documentation, an instant response from the maintenance engineer was received upon sending an inquiry email, indicating robust support for troubleshooting and resolving issues promptly.

A methylation region in this experiment is called as differentially methylated if 80% of covered CpGs were significantly hypo methylated and has a methylation reads greater than 5. dnmtools dmr function works better with a coverage of 10x for mammalian HMR identification, the method can work with lower coverage. Here, the Epiblast sample (mean_depth_covered: 19.0759) has more than double the coverage compared to the ICM sample (mean_depth_covered: 8.34952).

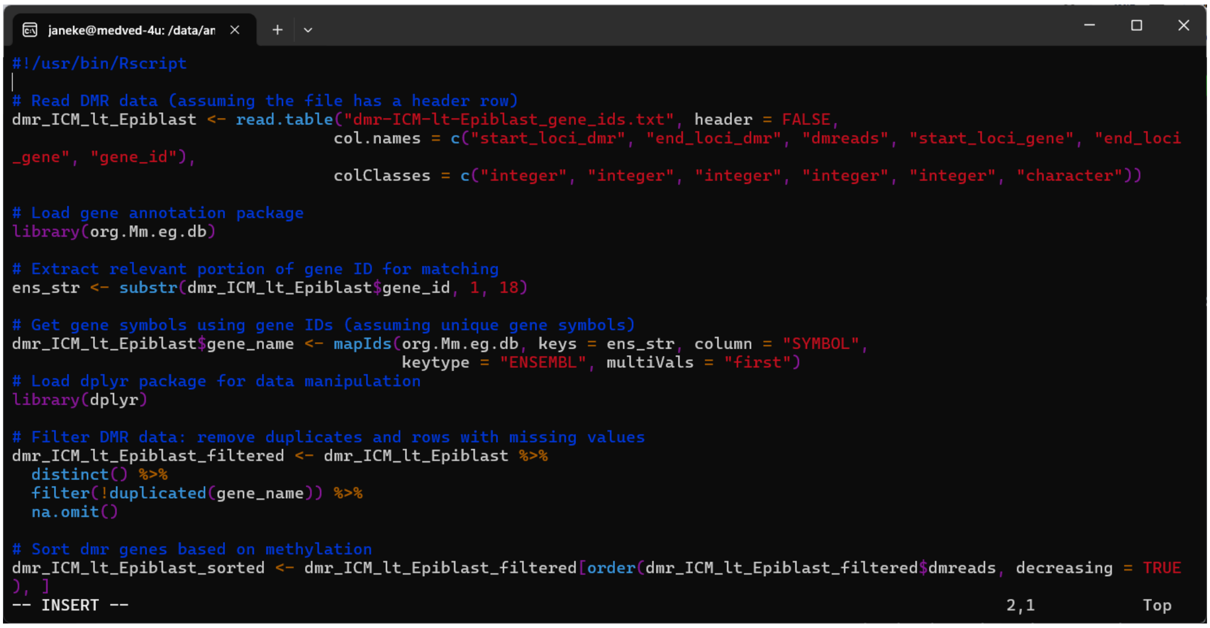
The DMR analysis (fig 5a) begins with merging the methylation data from two replicate samples, ICM_rep1 and ICM_rep2, using dnmtools merge. The resulting merged file is then compared with the methylation data from the Epiblast sample using dnmtools diff, producing a differential methylation analysis output. Subsequently, the dnmtools hmr command is utilized to identify high-methylation regions (HMRs) in the Epiblast sample based on the provided parameters in params.txt.

Following this, the differential methylation regions (DMRs) between ICM and Epiblast samples are detected using dnmtools dmr, resulting in two BED files: dmr-ICM-lt-Epiblast.bed and dmr-Epiblast_lt_ICM.bed. To verify the intersection between these DMR outputs, bedtools intersect is employed, and common.bed is generated. Further refinement is conducted by filtering DMRs based on specific criteria using awk, resulting in dmr-ICM-lt-Epiblast-filtered.bed. Finally, closestBed is utilized to determine the closest genes to the identified DMRs, with the results being stored in closest_genes.txt and then subsetted to include relevant information in dmr-ICM-lt-Epiblast_gene_ids.txt.

Fig 5a: DMR pipeline

```
janeke@medved-4u: /data/an

#!/bin/bash

# Set basename for the sample
basename="ICM"

# Merge methylation files
dnmtools merge -o "${basename}_WGBS.meth" SRR5836475/ICM_rep1_WGBS_symmetric_CpG_filtered.meth SRR5836476/ICM_rep2_WGBS_
symmetric_CpG_filtered.meth

# Perform methylation difference probabilities
dnmtools diff -o "${basename}_vs_Epiblast.diff" "${basename}_WGBS.meth" Epiblast_WGBS.meth

# Generate hypomethylated methylation region (HMR) file
dnmtools hmr -p params.txt -o Epiblast_merge_WGBS.hmr Epiblast_merge_WGBS.meth

# Perform differentially methylated regions analysis
dnmtools dmr "${basename}_vs_Epiblast.diff" "${basename}_merge_WGBS.hmr" Epiblast_merge_WGBS.hmr dmr-${basename}-lt-Epib
last.bed dmr-Epiblast_lt_${basename}.bed

# Find intersection between DMR output
bedtools intersect -a dmr-${basename}-lt-Epiblast.bed -b dmr-Epiblast_lt_${basename}.bed > common.bed

# Filter DMR BED file based on criteria
awk -F '[:\t]' '$5 >= 5 && $6/$5 >= 0.8' dmr-${basename}-lt-Epiblast.bed > dmr-${basename}-lt-Epiblast-filtered.bed

# Download gene gencode .gff, sort and convert to .bed using bedtools
sort -k1, 1-k4, 4n gencode.vM10.annotation.bed > gencode.vM10.annotation.sorted.bed

# Find closest genes to DMRs
closestBed -a dmr-${basename}-lt-Epiblast.bed -b gencode.vM10.annotation.sorted.bed > dmr-ICM-lt-Epiblast_closest_genes.
txt

                                                                                      1,4           Top
```

Fig 5b : Top genes closest to DMR pipeline



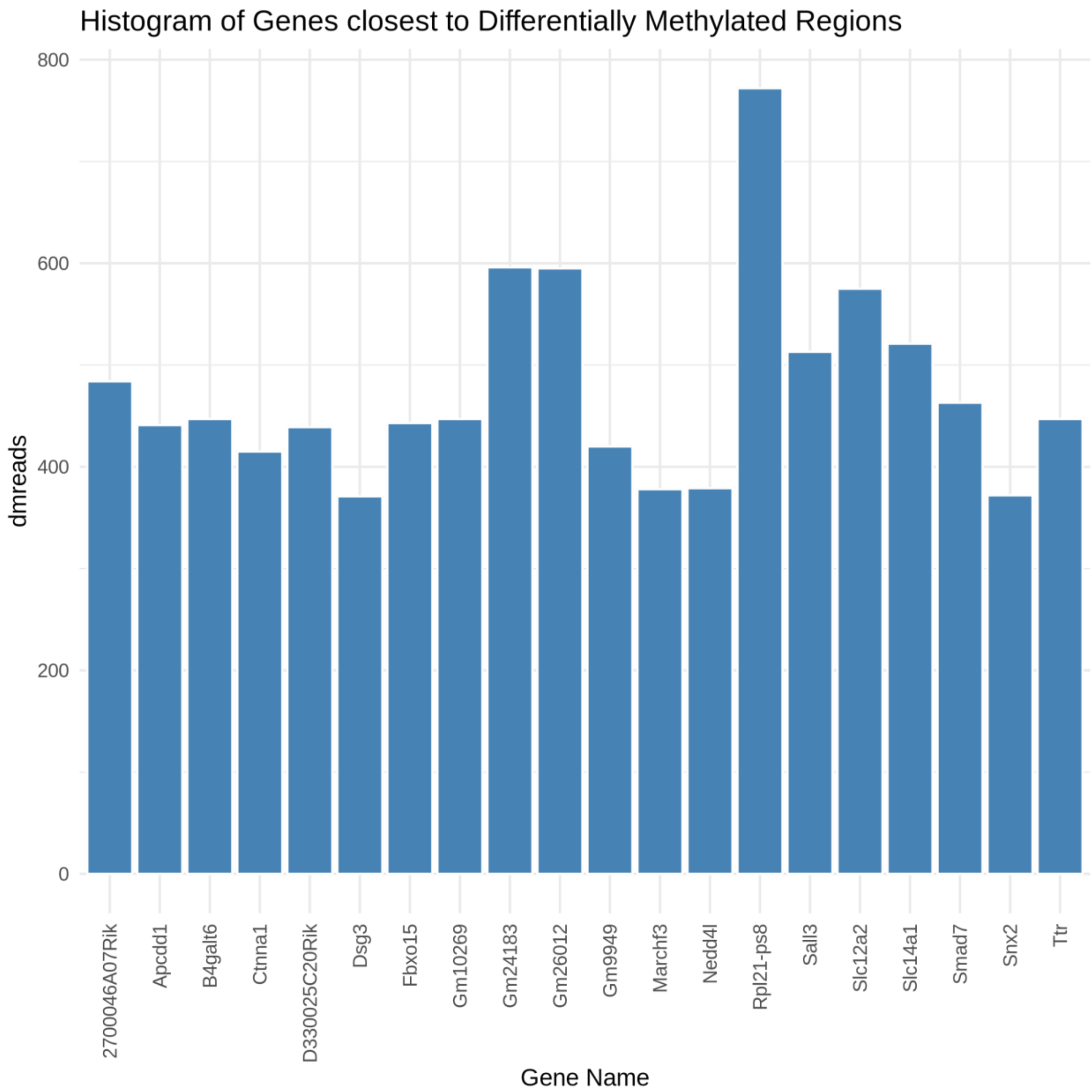
```
#!/usr/bin/Rscript

# Read DMR data (assuming the file has a header row)
dmr_ICM_lt_Epiblast <- read.table("dmr-ICM-lt-Epiblast_gene_ids.txt", header = FALSE,
                                 col.names = c("start_loci_dmr", "end_loci_dmr", "dmreads", "start_loci_gene", "end_loci
_gene", "gene_id"),
                                 colClasses = c("integer", "integer", "integer", "integer", "integer", "character"))

# Load gene annotation package
library(org.Mm.eg.db)

# Extract relevant portion of gene ID for matching
ens_str <- substr(dmr_ICM_lt_Epiblast$gene_id, 1, 18)

# Get gene symbols using gene IDs (assuming unique gene symbols)
dmr_ICM_lt_Epiblast$gene_name <- mapIds(org.Mm.eg.db, keys = ens_str, column = "SYMBOL",
                                        keytype = "ENSEMBL", multiVals = "first")
# Load dplyr package for data manipulation
library(dplyr)

# Filter DMR data: remove duplicates and rows with missing values
dmr_ICM_lt_Epiblast_filtered <- dmr_ICM_lt_Epiblast %>%
  distinct() %>%
  filter(!duplicated(gene_name)) %>%
  na.omit()

# Sort dmr genes based on methylation
dmr_ICM_lt_Epiblast_sorted <- dmr_ICM_lt_Epiblast_filtered[order(dmr_ICM_lt_Epiblast_filtered$dmreads, decreasing = TRUE
), ]
-- INSERT --                                                                                       2,1           Top
```

Fig 5b shows further analysis of genes closest to differentially methylated regions (DMRs). The DMRs data was read into dmr_ICM_lt_Epiblast, extracts relevant gene IDs, and retrieves corresponding gene symbols using the org.Mm.eg.db package. After filtering out duplicates and rows with missing values, it sorts the genes based on methylation levels and selects the top 50 genes. The script then creates a histogram using ggplot2 to visualize the distribution of methylation levels across the top 20 genes. Fig 5c shows the statistics summary of the analysis.

Fig 5c: DMR Summary Statistics

```
janeke@medved-4u: /data/an
 Mean   :46785872
 3rd Qu.:64862642
 Max.   :90504743
> summary(dmr_ICM_lt_Epiblast_filtered)
 start_loci_dmr      end_loci_dmr          dmreads        start_loci_gene
 Min.   : 3000022   Min.   : 3014770   Min.   :  4.00   Min.   : 3026900
 1st Qu.:34218305   1st Qu.:34222786   1st Qu.: 21.00   1st Qu.:34211061
 Median :43274426   Median :43295080   Median : 58.50   Median :43264337
 Mean   :46753813   Mean   :46764269   Mean   : 96.96   Mean   :46759819
 3rd Qu.:64824450   3rd Qu.:64866638   3rd Qu.:128.00   3rd Qu.:64862398
 Max.   :90410683   Max.   :90420245   Max.   :772.00   Max.   :90504614
 end_loci_gene        gene_id            gene_name
 Min.   : 3027882   Length:514         Length:514
 1st Qu.:34244706   Class :character   Class :character
 Median :43323063   Mode  :character   Mode  :character
 Mean   :46785872
 3rd Qu.:64862642
 Max.   :90504743
> head(dmr_ICM_lt_Epiblast_filtered, 2)
  start_loci_dmr end_loci_dmr dmreads start_loci_gene end_loci_gene
1        3000022      3014770     119         3026900       3027882
2        3335071      3336560       5         3266047       3337748
              gene_id   gene_name
1  ENSMUSG00000093774.1 Vmn1r-ps151
2 ENSMUSG00000063889.16        Crem
> |
```

Fig 5c: Top Genes Closest to DMR



Hypothesis Formulation

During mouse ICM and Epiblast developmental stages, differential methylation in chromosome 18 leads to the suppression of genes crucial for cellular differentiation, extracellular matrix organization, and cell signaling, thereby regulating key aspects of early embryogenesis. Specifically, genes such as Slc12a2 (Solute Carrier Family 12 Member 2), Fbn2 (Fibrillin 2) and Megf10 (Multiple EGF-like-domains 10) (highlighted in table 1), which are in proximity to DMRs, might undergo altered expression levels due to changes in DNA methylation.

These genes play essential roles in various biological processes critical for embryonic development. Slc12a2 is involved in ion transport and cellular osmoregulation, Fbn2 is a major component of the extracellular matrix involved in tissue development and maintenance, and Megf10 is implicated in cell adhesion and signaling. Differential methylation in their regulatory regions could modulate their expression, consequently influencing cell fate determination, extracellular matrix remodeling, and intercellular communication during early mouse embryonic development. Thus three null hypothesis could be formulated:

1. $\mathbf{H_O}$: There is no significant difference in the Differentially Methylated Regions (DMRs) of Inner Cell Mass (ICM) & Epiblast within chromosome 18 and the expression levels of Slc12a2 in mouse embryonic developmental stage.
2. $\mathbf{H_O}$: There is no significant difference in the Differentially Methylated Regions (DMRs) of Inner Cell Mass (ICM) & Epiblast within chromosome 18 and the expression levels of Fbn2 in mouse embryonic developmental stage.
3. $\mathbf{H_O}$: There is no significant difference in the Differentially Methylated Regions (DMRs) of Inner Cell Mass (ICM) & Epiblast within chromosome 18 and the expression levels of Megf10 in mouse embryonic developmental stage.

Table 1: 50 Top Genes closest to Loci of high DMR (50 of 514 DMR)

	Start_loci_dmr	End_loci_dmr	Dmreads	Start_loci_gene	End_loci_gene	Gene_name
1	82503225	82557376	772	82523543	82524021	Rpl21-ps8
2	55040896	55112586	596	55085564	55085657	Gm24183
3	81631636	81684106	595	81981068	81981196	Gm26012
4	57798697	57862623	575	57878677	57878806	Slc12a2
5	78073700	78123182	521	78100090	78102394	Slc14a1
6	80938289	80973688	513	80966375	80969472	Sall3
7	62739796	62789322	484	62751673	62753333	2700046A07Rik
8	75374558	75401638	463	75367528	75395935	Smad7
9	20636293	20685228	447	20665249	20665438	Ttr
10	20636293	20685228	447	20682591	20682594	Gm10269
11	20636293	20685228	447	20684598	20688320	B4galt6
12	84891284	84934278	443	84935157	84935160	Fbxo15
13	62826101	62871330	441	62922326	62922663	Apcdd1
14	80364005	80394950	439	80362780	80365826	D330025C20Rik
15	62180612	62228542	420	62180125	62184405	Gm9949
16	35057029	35106235	415	35118887	35118981	Ctnna1
17	64842324	64886005	379	64887755	64888047	Nedd4l
18	56748646	56774281	378	56761715	56762514	Marchf3
19	53074584	53117176	372	53176364	53176379	Snx2
20	20505700	20557645	371	20510303	20510374	Dsg3
21	65219051	65250730	364	65248860	65248926	Mir122
22	12268707	12305235	363	12287403	12289780	Gm16072
23	78138594	78167267	362	78146939	78146942	Slc14a2
24	61085139	61111775	359	61105571	61105684	Csf1r
25	60470670	60500785	358	60474192	60475395	Smim3
26	74155165	74182049	341	74195298	74196902	Ska1
27	74934975	74960002	324	74939321	74941316	Lipg
28	76164398	76191915	316	76170551	76170652	Mir6358
29	52558944	52607810	315	52615914	52616085	Zfp474
30	69313736	69344003	309	69343355	69343402	Tcf4
31	64770827	64808536	304	64786328	64786428	Gm24504
32	43346629	43383193	301	43320978	43438286	Dpysl3
33	63004401	63033879	300	63010212	63011547	Piezo2
34	65845849	65872950	285	65872819	65872863	Grp
35	38108420	38135452	282	38185913	38189942	Pcdh1
36	11061073	11093664	277	11052509	11085635	Gata6
37	42357758	42382291	273	42394538	42394642	Pou4f3
38	57528738	57559970	273	57533779	57533852	Ccdc192
39	68853102	68884622	273	68944632	68944705	4930546C10Rik
40	61648008	61666486	267	61649195	61649258	Mir143
41	82386823	82405195	260	82392495	82393692	Galr1
42	56652266	56675793	258	56707812	56708112	Lmnb1
43	61666730	61689737	252	61687934	61687983	Il17b
44	84722574	84742268	248	84720018	84730447	Dipk1c
45	84722574	84742268	248	84742161	84742317	Gm25005
46	57652571	57681835	247	57669452	57669558	Gm26038
47	65942769	65964869	247	65955726	65956863	Cplx4

48	53622852	53640396	245	53681723	53683232	Cep120
49	31939653	31956269	244	31942996	31946988	Gpr17
50	61015009	61034800	242	61018861	61019725	Cdx1
56	57977012	57996009	225	58008622	58010257	Fbn2
152	57055800	57070228	114	57133089	57133730	Megf10

- Promoter methylation: If the DMR overlaps with the gene's promoter region (the regulatory sequence controlling its expression), methylation can directly repress gene transcription.

However, the NKCC1 knockout mice were viable but presented with multiple debilitating phenotypes, which led the field to believe that the gene would be embryonically lethal in humans since no human with *SLC12A2* mutation had been reported until recently.

Mutations in the *SLC12A2* gene, which encodes the Na-K-2Cl cotransporter-1 (NKCC1), are linked to various conditions such as neurodevelopmental deficits, deafness, and fluid secretion in different epithelia.

https://journals.physiology.org/doi/full/10.1152/ajpcell.00238.2023

For example, your hypothesis might be something like: "During mouse ICM and Epiblast developmental stages, differential methylation in chromosome 18 leads to the suppression of genes involved in [specific pathway or process], thereby regulating [specific aspect of development]. This hypothesis could be tested by performing gene expression analysis or functional assays on the identified genes in mouse embryonic stem cells or embryos at different developmental stages, correlating their expression levels with methylation status and developmental outcomes."