

1. Контроль качества

Рисунок 1а. FastQC_report для SRR5836474_1

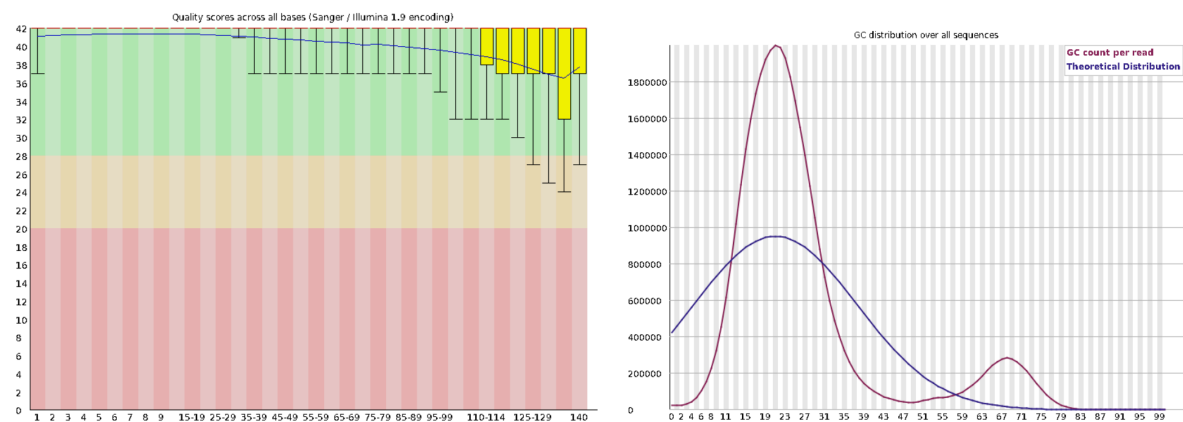
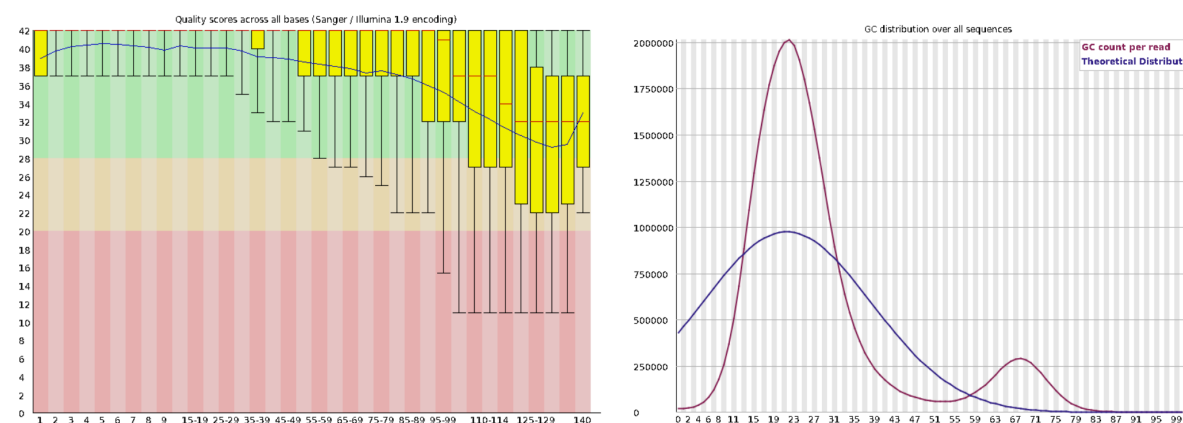


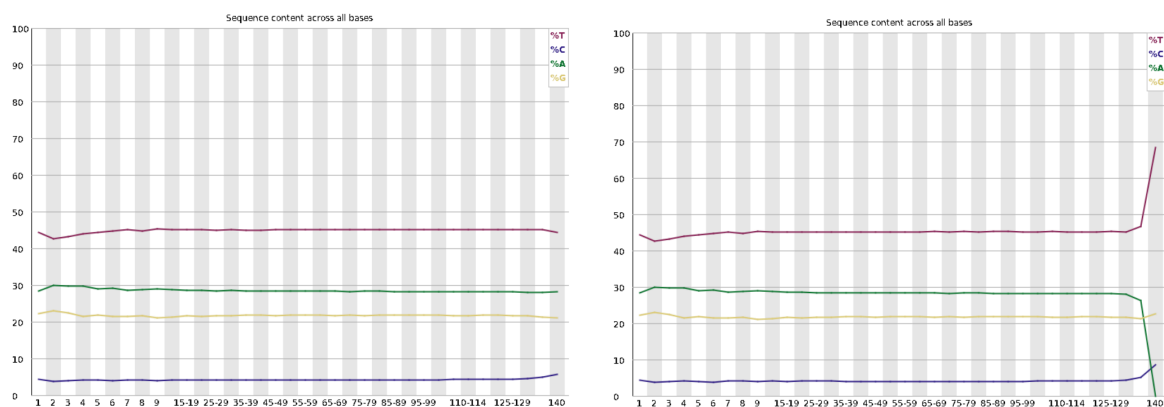
Рисунок 1б. FastQC_report для SRR5836474_2



Basic statistics:

	before trimming		after trimming	
sample	total sequences	sequence length	total sequences	sequence length
SRR5836473_1	10571668	19-140	10571203	20-140
SRR5836473_2	10571668	29-140	10571203	28-140
SRR5836474_1	39478709	19-140	39433185	20-140
SRR5836474_2	41156461	19-140	39433185	20-140
SRR5836475_1	112143129	19-140	112037168	20-140
SRR5836475_2	120436269	19-140	112037168	20-140
SRR5836476_1	102806438	19-140	102722321	20-140
SRR5836476_2	108518627	19-140	102722321	20-140

Рисунок 1в. Per base sequence content of SRR5836474_1 before/after trimming



Объяснение:

Результаты fastqc для обоих образцов показывают, что наши данные в порядке. Бимодальное распределение средних значений CG, наблюдаемое в результатах FastQC для образца SRR5836474 (рис. 1а,б), обычно указывает на наличие двух различных состояний метилирования в образце. Бимодальное распределение указывает на наличие двух различных популяций CpG-сайтов с разными уровнями метилирования. Это может свидетельствовать о наличии двух разных клеточных популяций в образце 8cell_WGBS, каждая из которых имеет свой характер метилирования.

На графиках, показанных на рис. 1а,б, не было явных различий и с после обрезки. Эта закономерность была одинаковой для всех образцов. При бисульфитном секвенировании неметилированные С преобразуются в Т, что приводит к увеличению содержания Т по сравнению с С в секвенированных фрагментах ДНК. Это смещение конверсии приводит к повышенному содержанию Т во всей последовательности. Кроме того, обработка бисульфитом не влияет на содержание А и G. Однако наблюдения (рис. 1в) указывают на резкое снижение содержания А к концу эксперимента. Это снижение может быть связано с наличием адаптерных последовательностей, богатых аденинами, которые обычно используются при составлении библиотеки для секвенирования. Последовательность адаптера может вносить искажения во время секвенирования, что приводит к различиям в содержании оснований в секвенированных фрагментах. Наблюдаемые закономерности отражают влияние бисульфитной обработки и адаптеров на состав оснований в секвенированной ДНК, и влияние наблюдалось во всех четырех образцах, что говорит о стабильности данных.

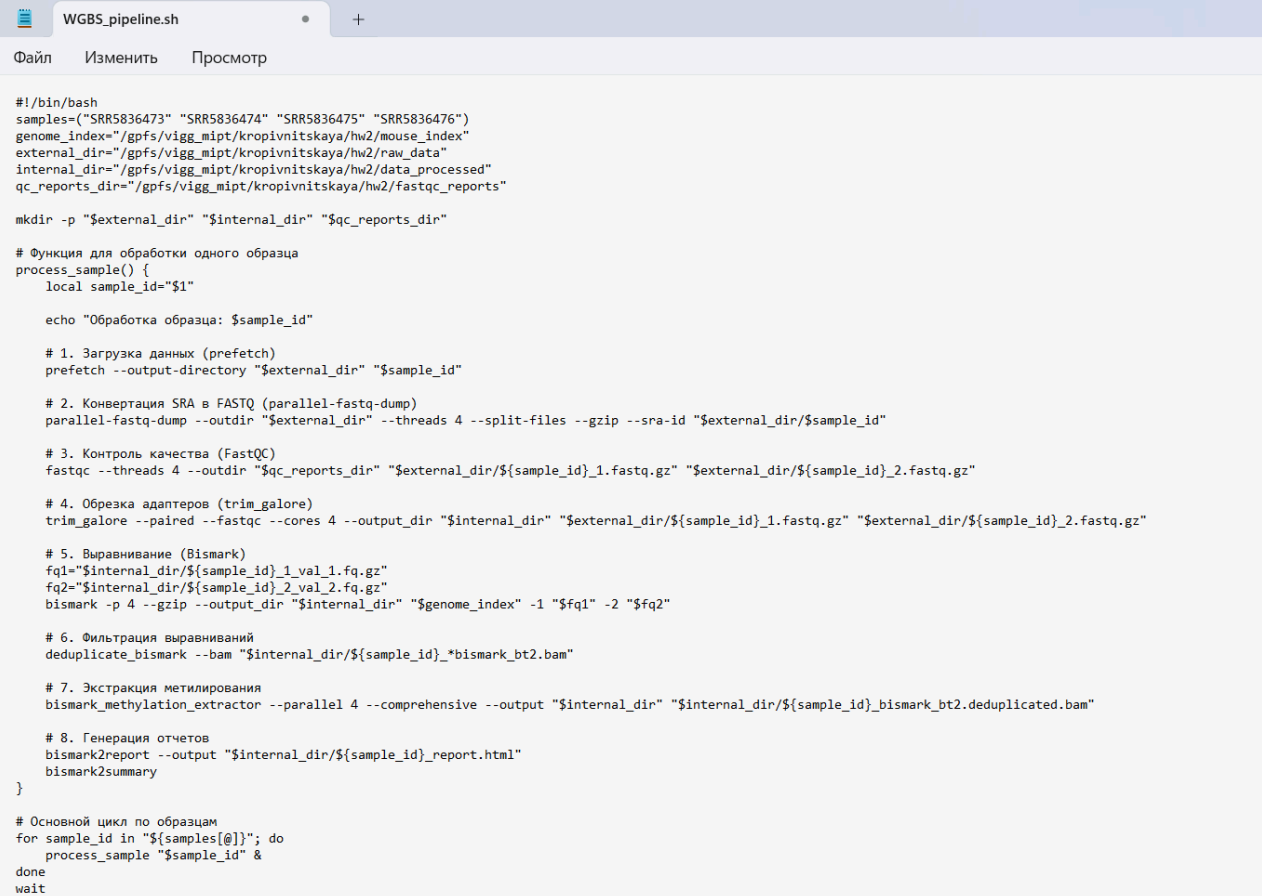
2. Выравнивание и визуализация

Пайплайн WGBS, представленный на рисунке 2а, разработан для эффективной обработки данных бисульфитного секвенирования — от получения необработанных прочтений до анализа метилирования. На первом этапе скрипт загружает данные секвенирования с помощью prefetch, после чего конвертирует их в формат FASTQ с помощью fasterq-dump. Затем проводится оценка качества прочтений с использованием FastQC. Для удаления адаптеров и обрезки низкокачественных

участков применяется Trim Galore, что обеспечивает высокое качество данных для последующего анализа.

Далее пайплайн строит индекс генома с помощью Bismark, после чего выполняется выравнивание обрезанных прочтений на бисульфитно-преобразованный референсный геном. После удаления дубликатов выравниваний осуществляется извлечение информации о метилировании с помощью Bismark, в результате чего создаются bedGraph-файлы с уровнями метилирования. На завершающем этапе пайплайн генерирует сводные отчёты и визуализации для профилирования метилирования.

Рисунок 2а. Пайплайн WholeGenomeBisulfiteSequencing



```
#!/bin/bash
samples=("SRR5836473" "SRR5836474" "SRR5836475" "SRR5836476")
genome_index="/gpfs/vigg_mipt/kropivnitskaya/hw2/mouse_index"
external_dir="/gpfs/vigg_mipt/kropivnitskaya/hw2/raw_data"
internal_dir="/gpfs/vigg_mipt/kropivnitskaya/hw2/data_processed"
qc_reports_dir="/gpfs/vigg_mipt/kropivnitskaya/hw2/fastqc_reports"

mkdir -p "$external_dir" "$internal_dir" "$qc_reports_dir"

# Функция для обработки одного образца
process_sample() {
    local sample_id="$1"

    echo "Обработка образца: $sample_id"

    # 1. Загрузка данных (prefetch)
    prefetch --output-directory "$external_dir" "$sample_id"

    # 2. Конвертация SRA в FASTQ (parallel-fastq-dump)
    parallel-fastq-dump --outdir "$external_dir" --threads 4 --split-files --gzip --sra-id "$external_dir/$sample_id"

    # 3. Контроль качества (FastQC)
    fastqc --threads 4 --outdir "$qc_reports_dir" "$external_dir/${sample_id}_1.fastq.gz" "$external_dir/${sample_id}_2.fastq.gz"

    # 4. Обрезка адаптеров (trim_galore)
    trim_galore --paired --fastqc --cores 4 --output_dir "$internal_dir" "$external_dir/${sample_id}_1.fastq.gz" "$external_dir/${sample_id}_2.fastq.gz"

    # 5. Выравнивание (Bismark)
    fq1="$internal_dir/${sample_id}_1_val_1.fq.gz"
    fq2="$internal_dir/${sample_id}_2_val_2.fq.gz"
    bismark -p 4 --gzip --output_dir "$internal_dir" "$genome_index" -1 "$fq1" -2 "$fq2"

    # 6. Фильтрация выравниваний
    deduplicate_bismark --bam "$internal_dir/${sample_id}_bismark_bt2.bam"

    # 7. Экстракция метилирования
    bismark_methylation_extractor --parallel 4 --comprehensive --output "$internal_dir" "$internal_dir/${sample_id}_bismark_bt2.deduplicated.bam"

    # 8. Генерация отчетов
    bismark2report --output "$internal_dir/${sample_id}_report.html"
    bismark2summary
}

# Основной цикл по образцам
for sample_id in "${samples[@]}; do
    process_sample "$sample_id" &
done
wait
```

Рисунок 2б. Bismark_report for SRR5836474



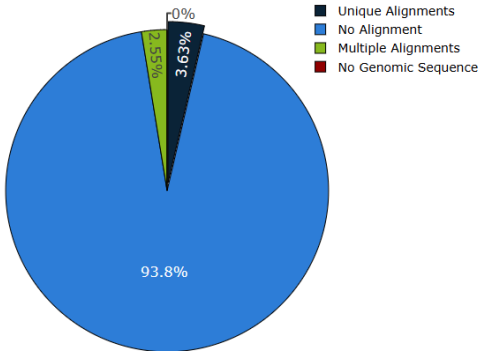
Bismark Processing Report

SRR5836474_1_val_1.fq.gz and SRR5836474_2_val_2.fq.gz

Data processed at 13:38:45 on 2025-05-08

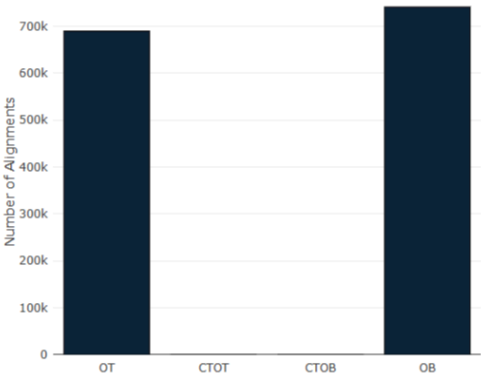
Alignment Stats

Sequence pairs analysed in total	39433185
Paired-end alignments with a unique best hit	1429855
Pairs without alignments under any condition	36998422
Pairs that did not map uniquely	1004908
Genomic sequence context not extractable (edges of chromosomes)	0



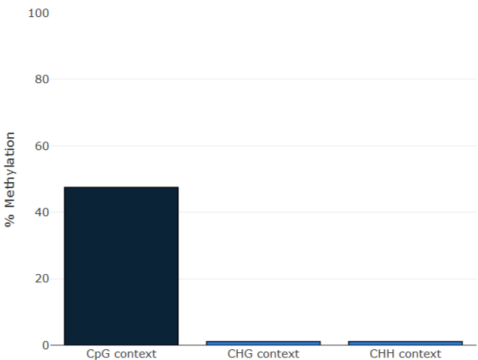
Alignment to Individual Bisulfite Strands

OT	689242	original top strand
CTOT	0	complementary to original top strand
CTOB	0	complementary to original bottom strand
OB	740613	original bottom strand



Cytosine Methylation

Total C's analysed	80722012
Methylated C's in CpG context	1802463
Methylated C's in CHG context	186194
Methylated C's in CHH context	681593
Methylated C's in Unknown context	4503
Unmethylated C's in CpG context	1988963
Unmethylated C's in CHG context	16605537
Unmethylated C's in CHH context	59457262
Unmethylated C's in Unknown context	41560
Percentage methylation (CpG context)	47.5%
Percentage methylation (CHG context)	1.1%
Percentage methylation (CHH context)	1.1%
Methylated C's in Unknown context	9.8%



На рисунке 2б показан профиль метилирования ДНК для образца SRR5836474. В общей сложности было проанализировано 39 433 185 пар прочтений, при этом эффективность картирования составила 3,63%. Из них 1 429 855 пар имели уникальное совпадение с одной областью генома (best-hit). Примечательно, что 36 998 422 пары не были сопоставлены ни при каких условиях, а 1 004 908 пары не имели уникального картирования. Из подпункта Alignment to Individual Bisulfite Strands видно, что все уникальные совпадения приходились на преобразованные цепи ДНК (верхнюю и нижнюю), а попыток сопоставить с комплементарными теоретическими цепями не было, что говорит о чистоте данных и корректности протокола обработки.

Что касается метилирования цитозинов, было проанализировано в общей сложности 80 722 012 С. Уровни метилирования варьировали в зависимости от контекста последовательности: 47,5% цитозинов были метилированы в контексте CpG, 1,1% — в контексте CHG и 1,1% — в контексте CHH. Кроме того, 9,8% цитозинов были метилированы в неизвестном контексте (CN или CHN). Эти результаты дают представление о паттернах метилирования в образце SRR5836474, подчеркивая значительное метилирование в участках CpG по сравнению с другими контекстами.

Рисунок 2г. M-bias plot for SRR5836474

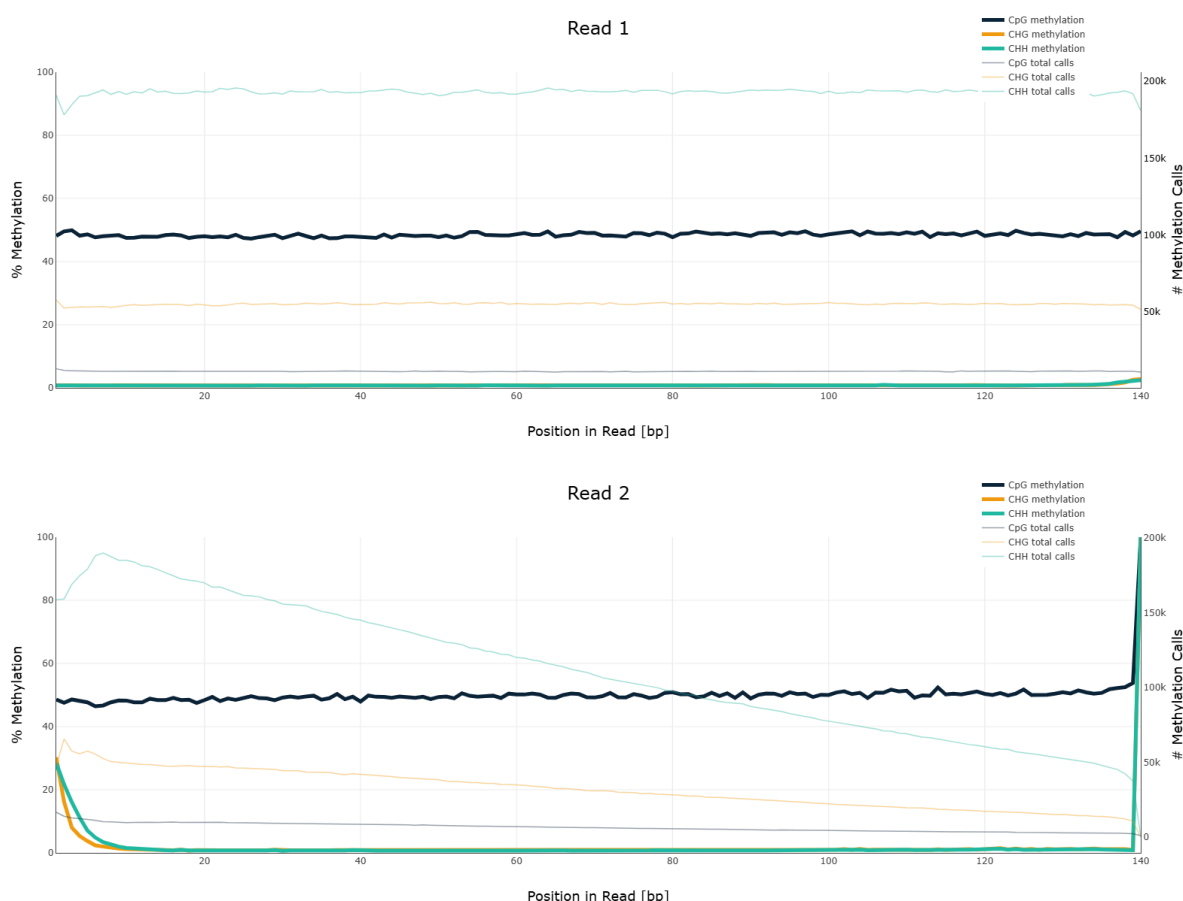


График M-Bias является показателем качества, используемым в анализе WGBS. Он позволяет оценить точность определения уровня метилирования в разных позициях внутри 2 прочтений (рис. 2г) из-за парно-концевого секвенирования. На графике:

- ось X показывает позицию в прочтении 1 или 2 (в парах оснований, bp).
- ось Y отражает уровень метилирования (процент метилированных прочтений).

Разные линии соответствуют различным типам метилирования.

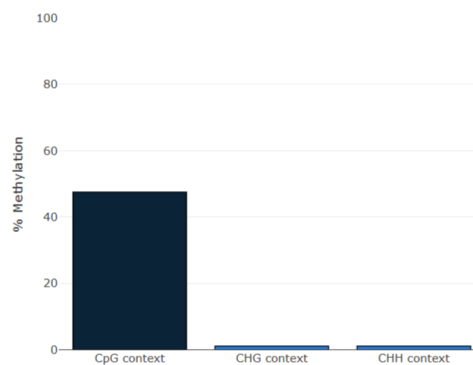
График показывает выраженный сдвиг (bias) в уровне метилирования на концах прочтений (вблизи 0 bp и 140 bp). Это связано с тем, что процесс бисульфитного превращения, который является ключевым этапом в WGBS, часто работает менее эффективно именно на концах прочтений. Такое снижение эффективности может привести к заниженной оценке уровня метилирования на этих участках.

Хотя все три типа метилирования (CpG, CHG, CHH) отражают метилирование цитозина перед другим нуклеотидом, в млекопитающих преобладает метилирование по CpG, поскольку оно играет важную роль в регуляции генов.

Рисунок 2д. Метилирование С до и после экстракции для SRR5836474

Cytosine Methylation

Total C's analysed	80722012
Methylated C's in CpG context	1802463
Methylated C's in CHG context	186194
Methylated C's in CHH context	681593
Methylated C's in Unknown context	4503
Unmethylated C's in CpG context	1988963
Unmethylated C's in CHG context	16605537
Unmethylated C's in CHH context	59457262
Unmethylated C's in Unknown context	41560
Percentage methylation (CpG context)	47.5%
Percentage methylation (CHG context)	1.1%
Percentage methylation (CHH context)	1.1%
Methylated C's in Unknown context	9.8%



Cytosine Methylation after Extraction

Total C's analysed	57279209
Methylated C's in CpG context	1230779
Methylated C's in CHG context	140371
Methylated C's in CHH context	499618
Unmethylated C's in CpG context	1290023
Unmethylated C's in CHG context	12052815
Unmethylated C's in CHH context	42065603
Percentage methylation (CpG context)	48.8%
Percentage methylation (CHG context)	1.2%
Percentage methylation (CHH context)	1.2%

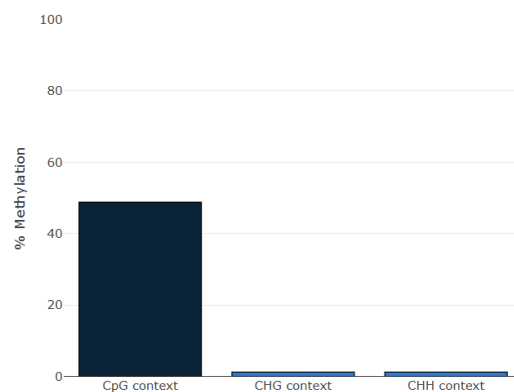
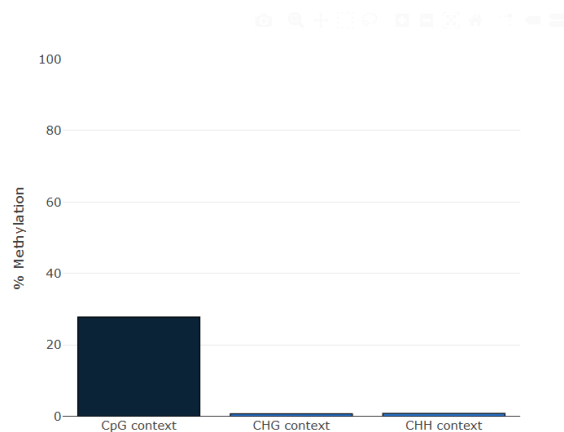


Рисунок 2е. Метилирование С до и после экстракции для SRR5836475

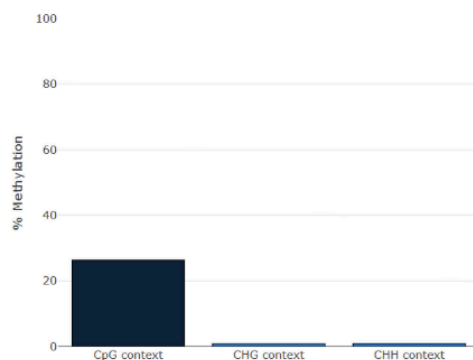
Cytosine Methylation

Total C's analysed	237577029
Methylated C's in CpG context	3093944
Methylated C's in CHG context	368745
Methylated C's in CHH context	1473210
Methylated C's in Unknown context	13372
Unmethylated C's in CpG context	8017297
Unmethylated C's in CHG context	49190489
Unmethylated C's in CHH context	175433344
Unmethylated C's in Unknown context	121269
Percentage methylation (CpG context)	27.8%
Percentage methylation (CHG context)	0.7%
Percentage methylation (CHH context)	0.8%
Methylated C's in Unknown context	9.9%



Cytosine Methylation after Extraction

Total C's analysed	161702660
Methylated C's in CpG context	1837064
Methylated C's in CHG context	266912
Methylated C's in CHH context	1029493
Unmethylated C's in CpG context	5151260
Unmethylated C's in CHG context	34303705
Unmethylated C's in CHH context	119114226
Percentage methylation (CpG context)	26.3%
Percentage methylation (CHG context)	0.8%
Percentage methylation (CHH context)	0.9%



Сравнительный анализ для образцов SRR5836474 и SRR5836475:

До экстракции оба образца показывали разные профили метилирования, при этом SRR5836474 демонстрировал значительно более высокий уровень метилирования, особенно в контексте CpG, по сравнению с SRR5836475.

После экстракции в обоих образцах наблюдалось снижение общего числа проанализированных цитозинов, однако проценты метилирования изменились минимально. При этом SRR5836474 продолжал сохранять более высокий общий уровень метилирования по сравнению с SRR5836475, что указывает на различия в динамике и стабильности метилирования между этими двумя образцами (рис. 2д,е), не связанные с технологией WGBS, а с процессами эмбрионального развития клеток.

	SRR5836473	SRR5836474	SRR5836475	SRR5836476
% (и количество) чтений, которые уникально были откартированы на геном	14,8% 1562062	3,63% 1429855	3,76% 4210090	3,95% 4055758
Глобальный процент метилирования цитозинов по	62,2%	59,5%	39,2%	36,8%

всему геному				
% метилирования цитозинов в CpG, CHG, CHH, unknown	45,3% 1,1% 1,0% 14,8%	47,5% 1,1% 1,1% 9,8%	27,8% 0,7% 0,8% 9,9%	26,7% 0,7% 0,8% 8,6%

[bismark summary report](#)

3. Анализ метилирования цитозинов в CpG контексте

Рисунок 3а. dnmtools_pipeline для нахождения сайтов метилирования

Файл Изменить Просмотр

```
#!/bin/bash

# Путь к референсной последовательности
ref_genome="/gpfs/vigg_mipt/kropivnitskaya/hw2/mouse_index/chr12.fa"

# Список идентификаторов образцов
samples=("SRR5836473" "SRR5836474" "SRR5836475" "SRR5836476")

# Количество потоков для параллельной обработки
threads=8

# Директория для выходных файлов
output_dir="dnmtools_output"
mkdir -p "$output_dir"

# Обработка каждого образца по очереди
for sample in "${samples[@]}"
do
    echo "Обработка $sample..."

    # Определяем пути к файлам для данного образца
    input_bam="${sample}_1_val_1_bismark_bt2_pe.deduplicated.bam"
    formatted_bam="${output_dir}/${sample}_format.bam"
    sorted_bam="${output_dir}/${sample}_sorted.bam"
    cpG_meth="${output_dir}/${sample}_CpG.meth"
    sym_cpG_meth="${output_dir}/${sample}_symmetric_CpG.meth"
    filtered_cpG_meth="${output_dir}/${sample}_symmetric_CpG_filtered.meth"

    # Конвертируем BAM-файл, созданный Bismark, в формат, совместимый с dnmtools
    dnmtools format -f bismark -t $threads -B "$input_bam" "$formatted_bam"

    # Сортируем полученный BAM-файл с помощью samtools
    samtools sort -o "$sorted_bam" "$formatted_bam"

    # Извлекаем информацию о метилировании цитозинов только в CpG-контексте
    dnmtools counts -cpg-only -c "$ref_genome" -o "$cpG_meth" "$sorted_bam"

    # Объединяем данные по обеим цепям ДНК (симметричное CpG-метилирование)
    dnmtools sym -o "$sym_cpG_meth" "$cpG_meth"

    # Удаляем участки, помеченные как CpGx (мутации), которые dnmtools не фильтрует автоматически
    awk ' $4 != "CpGx" ' "$sym_cpG_meth" > "$filtered_cpG_meth"

    echo "Обработка $sample завершена"
    echo "-----"
done

echo "Все образцы обработаны."
exit
```

Этот bash-скрипт (рис. 3а) описывает пайплайн для контроля качества и анализа данных бисульфитного секвенирования образцов. Процесс начинается с конвертации выровненного BAM-файла в формат DNMTTools, после чего файл сортируется. Затем рассчитываются уровни бисульфитного превращения, а также генерируются счетчики метилирования для дальнейшего анализа. После этого оцениваются уровни метилирования, с особым вниманием к сайтам CpG. Создается

профиль симметричного метилирования по CpG, а участки типа CpGx (асимметричные или артефактные) фильтруются. В целом, данный пайплайн обеспечивает полный контроль качества и комплексный анализ данных WGBS для образцов.

Рисунок 3б. Код для построения гистограмм

```

Файл  Изменить  Просмотр

# Load required library
library(ggplot2)

# Read methylation data from the .symmetric_CpGs.meth file
file_path <- "SRR5836476_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", ]

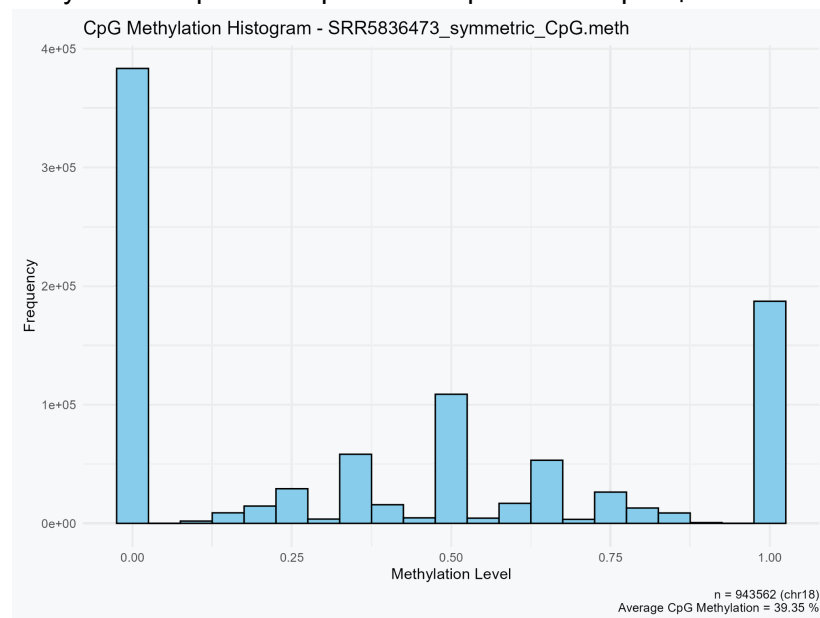
# Calculate total number of CpGs
total_cpgs <- nrow(methylation_data)

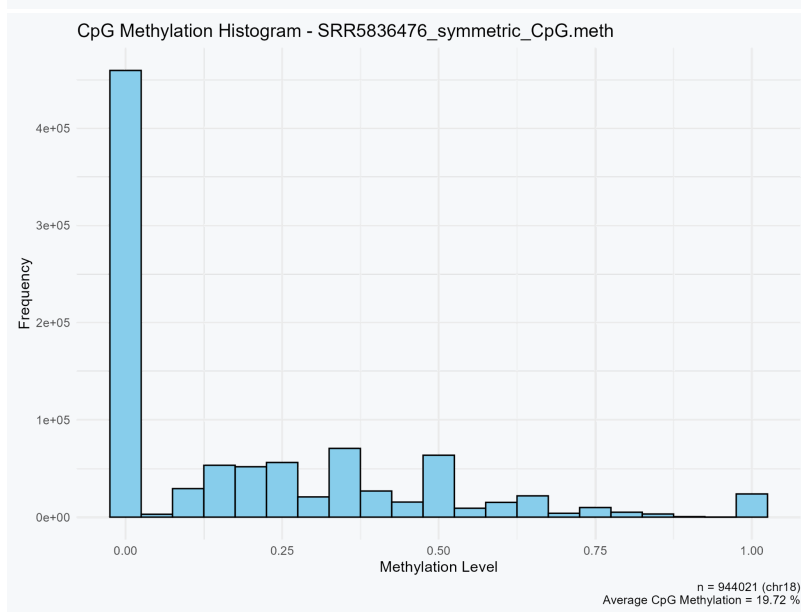
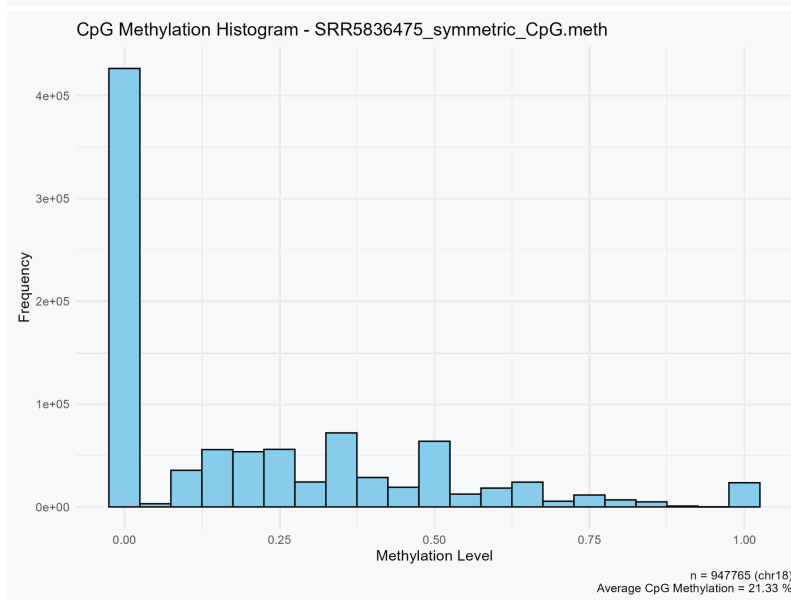
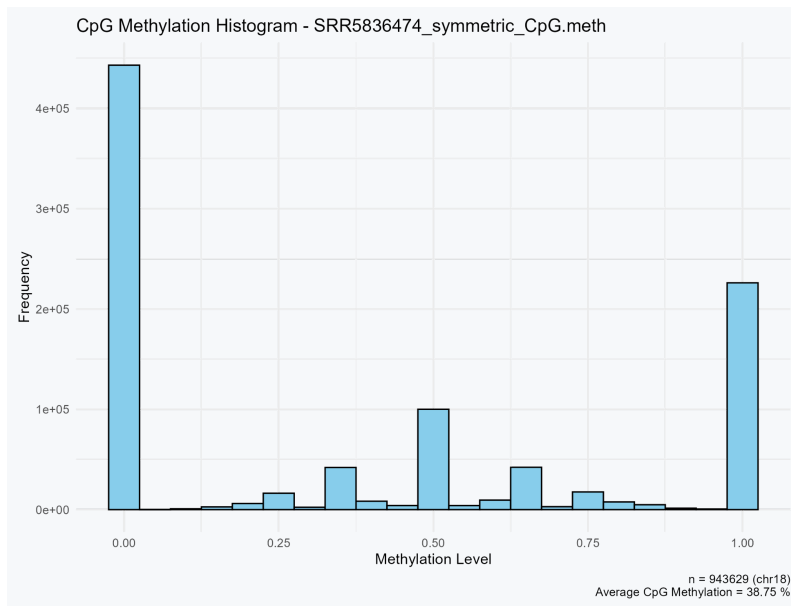
# Calculate average CpG methylation in percentage
avg_methylation <- mean(methylation_data$methylation_level, na.rm = TRUE) * 100

# 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 - ', "SRR5836476_symmetric_CpG.meth"),
       x = 'Methylation Level', y = 'Frequency',
       caption = paste('n = ', total_cpgs, '(chr18)\n',
                       'Average CpG Methylation = ', round(avg_methylation, 2), '%')) +
  theme_minimal()
p
# Save plot as PNG using ggsave
ggsave("SRR5836476_methylation_histogram.png", plot = p, width = 8, height = 6, dpi = 300)

```

Рисунок 3в. Уровень CpG метилирования образцов





Гистограммы метилирования CpG для 8cell_rep1_WGBS_symmetric_CpG.meth (SRR5836473) и 8cell_rep2_WGBS_symmetric_CpG.meth (SRR5836474) показывают похожие средние уровни метилирования: 39,35% и 38,75%, соответственно. Гистограммы показывают типичный бимодальный профиль метилирования CpG, где большинство сайтов либо полностью метилированы, либо полностью неметилированы (два выраженных типа в начале графика и в конце).

Гистограмма отображает распределение уровней метилирования на всех CpG участках в образце ICM_rep1_WGBS_symmetric_CpG.meth (SRR5836475). Средний уровень метилирования составил 21,33%, при этом большинство сайтов CpG показали уровни метилирования в диапазоне от 0% до 50%. Смещение гистограммы влево указывает на потенциальное смещение в сторону низкого уровня метилирования, что может быть связано с недостаточной эффективностью бисульфитного превращения на этапе подготовки библиотеки.

Аналогично для ICM_rep2_WGBS_symmetric_CpG.meth (SRR5836476). Средний уровень метилирования составил 19,72%, и, как и в первом репликате, большинство сайтов имели метилирование в диапазоне 0% до 50%. Гистограмма также демонстрирует левостороннее смещение, что аналогично указывает на возможные технические искажения из-за низкой эффективности бисульфитного превращения.

Наблюдаются существенные различия между 8_cell и ICM профилями метилирования. Эти различия в формах гистограмм и уровнях метилирования указывают на глобальные изменения метилирования ДНК между стадиями эмбрионального развития мыши.

3.1. Визуализация в Genome_Browser

go <https://genome.ucsc.edu/cgi-bin/hgCustom>

and do

track type=bigWig visibility=full name="SRR5836476.methylation.bigWig" color=197,172,24
maxHeightPixels=24 viewLimits=0:10 autoScale=off alwaysZero=on

bigDataUrl=<https://zenodo.org/records/15387754/files/SRR5836476.methylation.bigWig?download=1>

track type=bigWig visibility=full name="SRR5836476.coverage.bigWig" color=0,0,0
maxHeightPixels=24 viewLimits=0:20 autoScale=off alwaysZero=on

bigDataUrl=https://zenodo.org/records/15387754/files/SRR5836476_genome_coverage.bigWig?download=1

track type=bigWig visibility=full name="SRR5836475.methylation.bigWig" color=197,172,24
maxHeightPixels=24 viewLimits=0:10 autoScale=off alwaysZero=on

bigDataUrl=<https://zenodo.org/records/15387754/files/SRR5836475.methylation.bigWig?download=1>

track type=bigWig visibility=full name="SRR5836475.coverage.bigWig" color=0,0,0
maxHeightPixels=24 viewLimits=0:20 autoScale=off alwaysZero=on

bigDataUrl=https://zenodo.org/records/15387754/files/SRR5836475_genome_coverage.bigWig?download=1

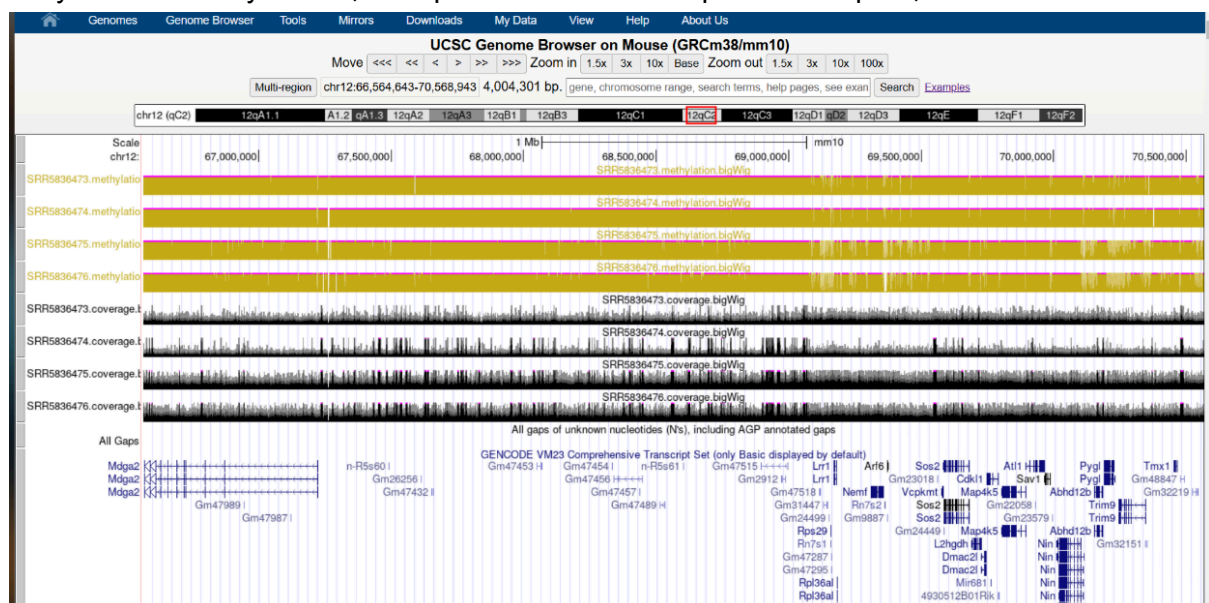
track type=bigWig visibility=full name="SRR5836474.methylation.bigWig" color=197,172,24
maxHeightPixels=24 viewLimits=0:10 autoScale=off alwaysZero=on
bigDataUrl=<https://zenodo.org/records/15387754/files/SRR5836474.methylation.bigWig?download=1>

track type=bigWig visibility=full name="SRR5836474.coverage.bigWig" color=0,0,0
maxHeightPixels=24 viewLimits=0:20 autoScale=off alwaysZero=on
bigDataUrl=https://zenodo.org/records/15387754/files/SRR5836474_genome_coverage.bigWig?download=1

track type=bigWig visibility=full name="SRR5836473.methylation.bigWig" color=197,172,24
maxHeightPixels=24 viewLimits=0:10 autoScale=off alwaysZero=on
bigDataUrl=<https://zenodo.org/records/15387754/files/SRR5836473.methylation.bigWig?download=1>

track type=bigWig visibility=full name="SRR5836473.coverage.bigWig" color=0,0,0
maxHeightPixels=24 viewLimits=0:20 autoScale=off alwaysZero=on
bigDataUrl=https://zenodo.org/records/15387754/files/SRR5836473_genome_coverage.bigWig?download=1

Рисунок 3.1а. Визуализация выравнивания метилирования образцов на chr12



[session](#)

На рисунке 3.1а представлена визуализация файлов methylation.bigWig и coverage.bigWig с использованием genome browser ucsc. Графическое представление демонстрирует все четыре образца вместе с их профилями метилирования и покрытия.

С помощью IGV можно наблюдать различные гены и хромосому 12 (chr12), что позволяет оценить распределение уровней метилирования и глубины покрытия по геномным регионам.

Эта визуализация дает исследователям возможность детально изучать паттерны метилирования и глубину покрытия в конкретных участках генома, что

облегчает выявление регионов интереса и способствует всестороннему анализу эпигенетических изменений.

4. Differential Methylation Analysis

Рисунок 4а. dmr_pipeline

```
Файл  Изменить  Просмотр

#!/bin/bash

# 1. Определяем имя образца
basename="8_cell"

# 2. Объединяем файлы метилирования репликатов
dnmttools merge -o "${basename}_WGBS.meth" ./bismark/dnmttools_output/SRR5836473_symmetric_CpG_filtered.meth ./bismark/dnmttools_output/SRR5836474_symmetric_CpG_filtered.meth
dnmttools merge -o "ICM_WGBS.meth" ./bismark/dnmttools_output/SRR5836475_symmetric_CpG_filtered.meth ./bismark/dnmttools_output/SRR5836476_symmetric_CpG_filtered.meth

# 3. Сравниваем метилирование
dnmttools diff -o "${basename}_vs_ICM.diff" "${basename}_WGBS.meth" ICM_WGBS.meth

# 4. Генерируем HMR для обеих групп
dnmttools hmr -p params.txt -o ${basename}_merge_WGBS.hmr ${basename}_WGBS.meth
dnmttools hmr -p params.txt -o ICM_merge_WGBS.hmr ICM_WGBS.meth

# 5. Определяем DMR
dnmttools dmr "${basename}_vs_ICM.diff" ${basename}_merge_WGBS.hmr ICM_merge_WGBS.hmr dmr-${basename}-lt-ICM.bed dmr-ICM_lt-${basename}.bed

# 6. Проверяем, создался ли DMR файл
if [ ! -f dmr-${basename}-lt-ICM.bed ]; then
    echo "Ошибка: DMR файл dmr-${basename}-lt-ICM.bed не найден. Прерывание."
    exit 1
fi

# 7. Пересечение DMR (опционально)
bedtools intersect -a dmr-${basename}-lt-ICM.bed -b dmr-ICM_lt-${basename}.bed > common.bed

# 8. Фильтрация DMR по количеству CpG и уровню различия
awk -F '[:\t]' ' $5 > 0' dmr-8_cell-lt-ICM.bed > dmr-8_cell-lt-ICM_nozero.bed
awk -F '[:\t]' ' $5 >= 1 && $6/$5 >= 0.05' dmr-8_cell-lt-ICM_nozero.bed > dmr-8_cell-lt-ICM-filtered.bed

# 9. Скачиваем и подготавливаем аннотацию генов, если её нет
if [ ! -f "genome.vM10.annotation.sorted.bed" ]; then
    wget https://ftp.ebi.ac.uk/pub/databases/genocode/genocode_mouse/release_M10/genocode.vM10.annotation.gtf.gz
    gunzip genocode.vM10.annotation.gtf.gz
    awk ' $3 == "gene" {match($0, /gene_id "([^"]+)"/, arr); if (arr[1] != "") print $1"\t"$4-1"\t"$5"\t"arr[1]; }' genocode.vM10.annotation.gtf > genocode.vM10.annotation.sorted.bed
    sort -k1,1 -k2,2n genocode.vM10.annotation.sorted.bed > genocode.vM10.annotation.sorted.bed
    echo "Файл успешно создан: genocode.vM10.annotation.sorted.bed"
else
    echo "Файл уже существует: genocode.vM10.annotation.sorted.bed"
fi

# 10. Поиск ближайших генов ТОЛЬКО для фильтрованных DMR
closestBed -a dmr-ICM_lt-8_cell.bed -b genocode.vM10.annotation.sorted.bed > dmr-ICM_lt-8_cell_closest_genes.txt

# 11. Отбираем нужные колонки
awk '{print $2, $3, $5, $8, $9, $10}' dmr-ICM_lt-8_cell_closest_genes.txt > dmr-ICM_lt-8_cell_gene_ids.txt

exit
```

Примечательно, что для dmr-ICM_lt-8_cell_closest_genes получились дифференциально экспрессирующиеся гены, а для dmr-8_cell-lt-ICM_closest_genes дэги отсутствовали, что свидетельствует о следующем: в процессе перехода от 8-cell к ICM клетки проходят эпигенетическую репрограммировку, включая деметилирование и активацию определённых генов, ответственных за плюрипотентность и раннюю дифференцировку. Полученные DMR (ICM_lt_8_cell) показывают участки, где метилирование снизилось в ICM, что совпадает с ожидаемой активацией генов, связанных с ICM-специфичными функциями (например, гены, связанные с развитием эмбриобласта).

После фильтрации осталось 0 генов для dmr-8_cell-lt-ICM-filtered.bed и 39 генов для dmr-ICM_lt-8_cell-filtered.bed, поэтому пункт 10 был реализован без предварительной фильтрации:

```
(hw2) [kropivnitskaya@n7 hw2]$ wc -l dmr-8_cell-lt-ICM.bed
581 dmr-8_cell-lt-ICM.bed
```

```
(hw2) [kropivnitskaya@n7 hw2]$ wc -l dmr-ICM_lt-8_cell.bed
1276 dmr-ICM_lt-8_cell.bed
```

Рисунок 4б. differential_analysis_pipeline.sh

```

# Read DMR data (assuming the file has a header row)
dmr_8_cell_lt_ICM <- read.table("C:/Users/natal/Downloads/fastqc_files/diff_analysis/dmr-ICM_lt-8_cell_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_8_cell_lt_ICM$gene_id, 1, 18)

# Get gene symbols using gene IDs (assuming unique gene symbols)
dmr_8_cell_lt_ICM$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_8_cell_lt_ICM_filtered <- dmr_8_cell_lt_ICM %>%
  distinct() %>%
  filter(!duplicated(gene_name)) %>%
  na.omit()

# Sort dmr genes based on methylation
dmr_8_cell_lt_ICM_sorted <- dmr_8_cell_lt_ICM_filtered[order(dmr_8_cell_lt_ICM_filtered$dmreads, decreasing = TRUE), ]

# Subset the first 50 genes
dmr_top50 <- dmr_8_cell_lt_ICM_sorted[1:50, ]

# Save the dmr_top20 object as a table
write.table(dmr_top50, file = "dmr_top50_table.txt", sep = "\t", quote = FALSE, row.names = FALSE)

library(ggplot2)

# Subset the first 20 genes
dmr_top20 <- dmr_8_cell_lt_ICM_sorted[1:20, ]

# Create the histogram using ggplot2
ggplot(dmr_top20, aes(x = gene_name, y = dmreads)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "white") +
  labs(title = "Histogram of Genes closest to Differentially Methylated Regions",
       x = "Gene Name", y = "dmreads") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

# Save the plot as an image
ggsave("dmr_top20.png")

```

Рисунок 4в. • Статистика по DMR

```

> summary(dmr_8_cell_lt_ICM_filtered)
start_loci_dmr      end_loci_dmr      dmreads
Min.   : 3306652    Min.   : 3308320    Min.   : 0.00
1st Qu.: 39459600   1st Qu.: 39479952   1st Qu.: 3.00
Median : 77448696   Median : 77465652   Median :11.00
Mean   : 70025430   Mean   : 70032292   Mean   :15.45
3rd Qu.:101931828   3rd Qu.:101933769   3rd Qu.:24.00
Max.   :119939360   Max.   :119948340   Max.   :92.00
start_loci_gene    end_loci_gene      gene_id
Min.   : 3247429    Min.   : 3309969    Length:540
1st Qu.: 39410905   1st Qu.: 39433213   Class :character
Median : 77526391   Median : 77645574   Mode  :character
Mean   : 70027337   Mean   : 70093369
3rd Qu.:101933168   3rd Qu.:101970183
Max.   :119945956   Max.   :120019211
gene_name
Length:540
Class :character
Mode  :character

```

```

> head(dmr_8_cell_lt_ICM_filtered)
start_loci_dmr end_loci_dmr dmreads start_loci_gene
2      3306652    3308320      6      3247429
3      3339975    3357726     31      3343815
4      3425088    3425420      1      3403877
5      3536469    3549850     25      3572380
6      3829317    3838399     34      3806006
7      3960581    3961470      6      3954950
end_loci_gene  gene_id  gene_name
2      3309969  ENSMUSG00000020671.8  Rab10
3      3357153  ENSMUSG000000097429.1  Gm26520
4      3427011  ENSMUSG000000071456.6  1110002L01Rik
5      3781796  ENSMUSG000000071454.13  Dtnb
6      3914443  ENSMUSG000000020661.15  Dnmt3a
7      3960618  ENSMUSG000000020660.5  Pomc

```

*Пометка: это для dmr_ICM_ltr_8_cell_filtered, я просто изначально код писала для dmr_8_cell_ltr_ICM_filtered и забыла переименовать.

Рисунок 4с. Гистограмма генов, расположенных близко к DMR

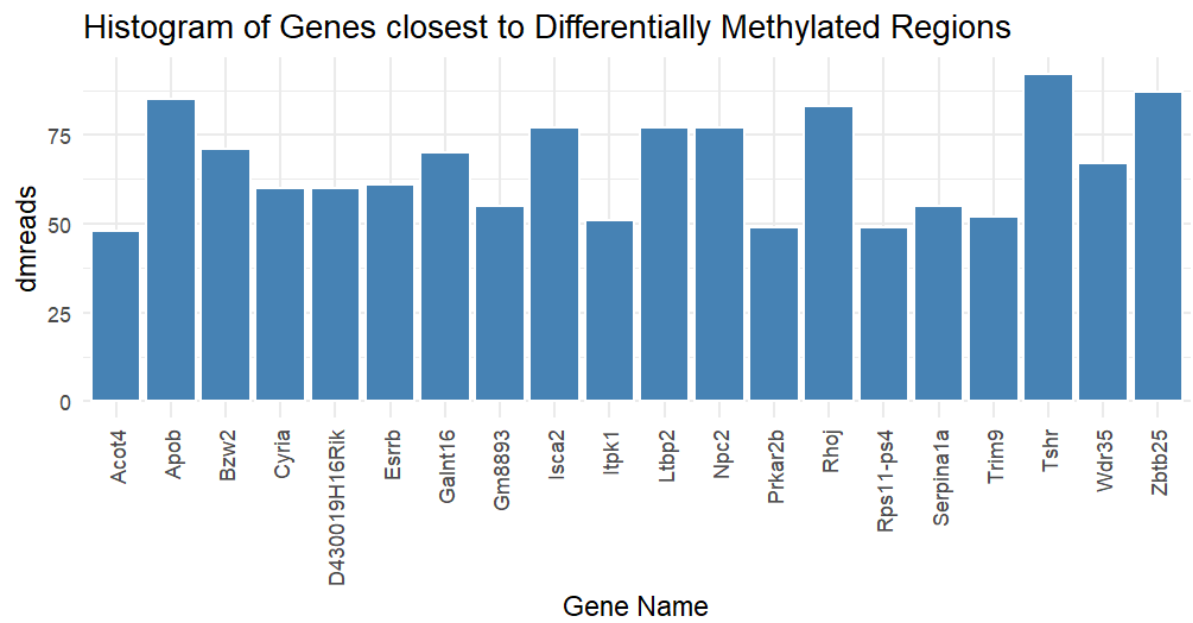


Таблица 4а. Топ 50 генов, расположенных близко к DMR:

start_loci_dmr	end_loci_dmr	distance	loci_start_gene	loci_end_gene	gene_id	
3091707	3098378	22	3082666	3087041	ENSMUSG000000092168.1	
3234392	3236616	2	3235503	3250377	ENSMUSG000000079179.9	
3306652	3308320	6	3247429	3309969	ENSMUSG000000020671.8	
3339975	3357726	31	3343815	3357153	ENSMUSG000000097429.1	
3425088	3425420	1	3403877	3427011	ENSMUSG000000071456.6	
3536469	3549850	25	3572380	3781796	ENSMUSG000000071454.13	
3736002	3752651	35	3752542	3762146	ENSMUSG000000092286.1	
3829317	3838399	34	3806006	3914443	ENSMUSG000000020661.15	
3960581	3961470	6	3954950	3960618	ENSMUSG000000020660.5	
4007345	4016364	28	3962553	4038915	ENSMUSG000000020658.9	
4080495	4084043	7	4082573	4110612	ENSMUSG000000020657.15	
4122802	4123586	1	4133102	4213525	ENSMUSG000000020654.15	
4230304	4236161	14	4196003	4234294	ENSMUSG000000020652.17	
4230304	4236161	14	4234026	4240123	ENSMUSG000000096199.1	
4466791	4468741	9	4247362	4477182	ENSMUSG000000020647.9	
4563126	4566969	11	4592825	4592927	ENSMUSG000000091619.1	
4591238	4594043	7	4593007	4713950	ENSMUSG000000020640.9	
4766988	4771978	7	4746215	4769267	ENSMUSG000000050545.8	
4766988	4771978	7	4769294	4778266	ENSMUSG000000020639.13	
4867099	4867342	4	4862439	4874359	ENSMUSG000000037336.14	
4895581	4895813	3	4879031	4907705	ENSMUSG000000020634.12	
4917375	4921837	15	4917403	5044047	ENSMUSG000000052812.4	
5061041	5070421	26	5077471	5375682	ENSMUSG000000020627.9	
5381240	5382411	4	5375869	5416631	ENSMUSG000000037735.7	
5729694	5740264	25	5843711	5843826	ENSMUSG000000088209.1	
6780340	6790878	19	7372038	7380330	ENSMUSG000000102096.1	
7772466	7781756	25	7859826	7861205	ENSMUSG000000073233.3	
8067900	8118675	85	7977647	8016835	ENSMUSG000000020609.14	
8290231	8291208	5	8208106	8285758	ENSMUSG000000037669.14	
8367589	8372116	9	8313432	8343824	ENSMUSG000000020605.8	
8460602	8461985	7	8497762	8499985	ENSMUSG000000054364.4	
8530451	8530589	0	8528482	8599066	ENSMUSG000000020600.13	
8637781	8638419	1	8674133	8752581	ENSMUSG000000020594.14	
8842345	8843012	3	8873186	8877459	ENSMUSG000000042976.5	
8919683	8925371	7	8921663	8938742	ENSMUSG000000020585.4	
8984939	8997221	67	8973891	9028847	ENSMUSG000000066643.12	
9032588	9033698	6	9029996	9036394	ENSMUSG000000066637.1	
9428132	9439757	18	9429545	9430099	ENSMUSG000000060647.1	
9508631	9520940	12	9565235	9570585	ENSMUSG000000099360.1	
10227388	10242103	22	10369972	10390173	ENSMUSG000000020622.15	
10385724	10391597	3	10390779	10395562	ENSMUSG000000020621.5	
11080245	11094767	30	11090201	11150842	ENSMUSG000000043673.10	
11187252	11187558	3	11208381	11208948	ENSMUSG000000047002.1	
11269961	11270760	3	11265885	11319785	ENSMUSG000000020608.6	
11386855	11401711	35	11325246	11436649	ENSMUSG000000054459.6	
11447219	11457922	14	11456078	11462928	ENSMUSG000000086022.1	
12216882	12254734	60	12262138	12376359	ENSMUSG000000020589.16	
12482363	12501975	32	12594822	12595352	ENSMUSG000000103645.1	
12648073	12656629	13	12682513	12682618	ENSMUSG000000098314.1	
12699598	12716878	47	12700038	12700145	ENSMUSG000000075893.1	