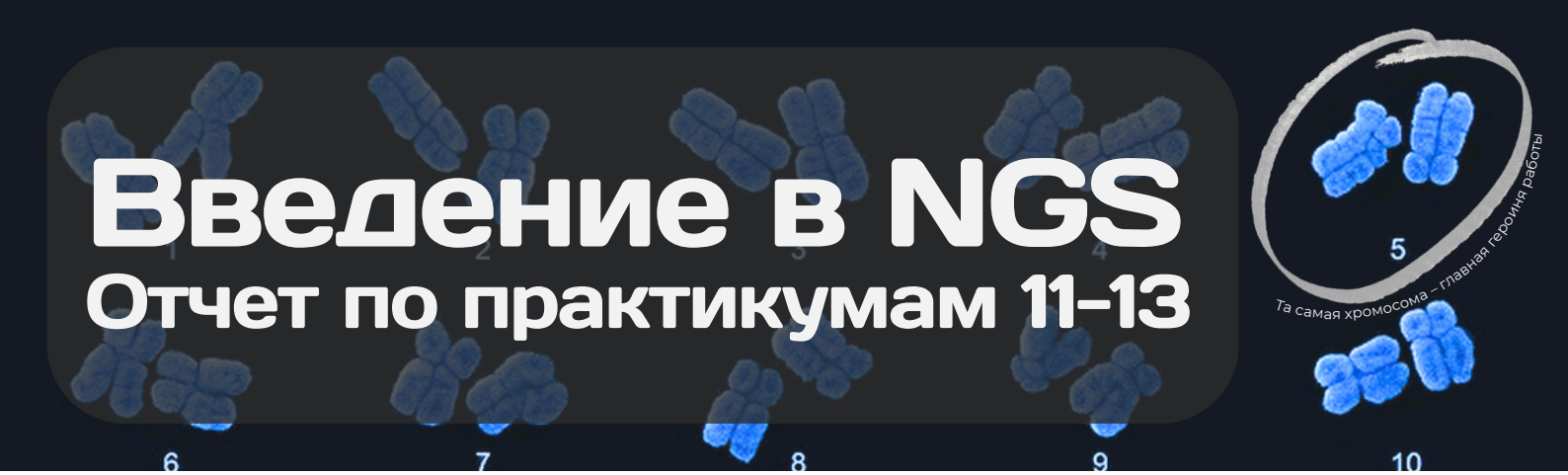


Автор : Долматова Алиса, alicedol Хромосома: 5 ДНК-образец: SRR10720402 РНК-образец: ENCFF763QQV

Практикум 11

Часть 1: подготовка референса

1.1 Получение референса

Выполненные команды:

```
cp /mnt/scratch/NGS/DATA/hg38/Homo_sapiens.GRCh38.dna.chromosome.5.fa
00_raw_data/dna/reference/
```

1.2 Индексация референса

Выполненные команды:

```
# Индексация для hisat2
hisat2-build Homo_sapiens.GRCh38.dna.chromosome.5.fa chr5_index

# Индексация для samtools
samtools faidx Homo_sapiens.GRCh38.dna.chromosome.5.fa
```

Результаты:

1. Созданы 8 индексных файлов для hisat2 (формат .ht2)
2. Создан индексный файл для samtools (формат .fai)

Информация из .fai файла:

```
5    181538259    6    60    61
```

Расшифровка:

- *Имя хромосомы: 5*
- *Длина хромосомы: 181,538,259 нуклеотидов*
- *Смещение в файле: 6 байт*
- *Баз в строке: 60*
- *Байт в строке: 61*

Часть 2: чтения ДНК

2.1 Описание образца

a) **SRR ID образца**: SRR10720402
b) **Ссылка на информацию**: https://www.ncbi.nlm.nih.gov/sra/?term=SRR10720402
c) **Прибор для секвенирования**: Illumina Genome Analyzer IIx
d) **Организм**: Homo sapiens
e) **Стратегия секвенирования**: Экзомное секвенирование
f) **Тип чтений**: Парноконцевые
g) **Spots**: 28,966,798

***Комментарий**: Образец представляет собой данные экзомного секвенирования человека, полученные на приборе Illumina Genome Analyzer IIx.*

2.2 Проверка качества исходных чтений

Выполненные команды:

```
fastqc SRR10720402_1.fastq.gz -o 01_preprocessing/dna/fastqc_raw

fastqc SRR10720402_2.fastq.gz -o 01_preprocessing/dna/fastqc_raw
```

a) **Количество пар чтений:** 28,966,798 пар.

b) **Совпадение количества прямых и обратных чтений:** Да, полностью совпадает (по 28,966,798 чтений в каждом файле).

c) **Качество чтений (Per base sequence quality):**

- **Прямые чтения:** Качество высокое на всём протяжении чтения. Медианное качество: позиции 1-32 — Phred 35-40, позиции 33-72 — Phred 32-40. На позициях 73-75 минимальные значения (нижние усы) опускаются до Phred 12-32.
- **Обратные чтения:** Качество также высокое, но снижение к концу более выражено: позиции 1-27 — Phred 35-40, позиции 27-71 — Phred 33-40, позиции 70-75 — минимальные значения до Phred 2-32 (на последней позиции до Phred 2).
- ***Комментарий:*** *Качество секвенирования отличное.*

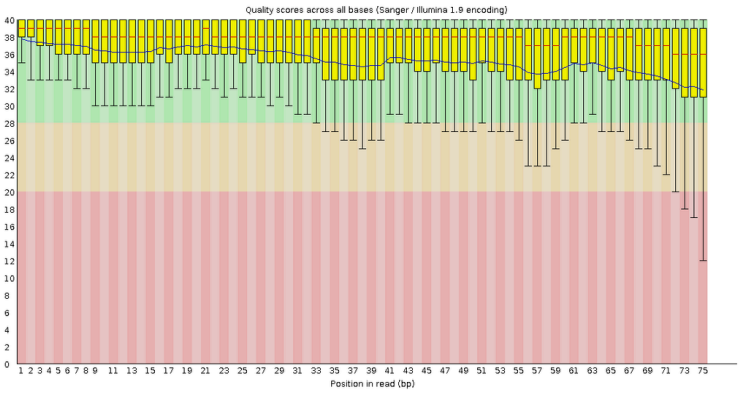


Рис. 1 Качество прямых чтений

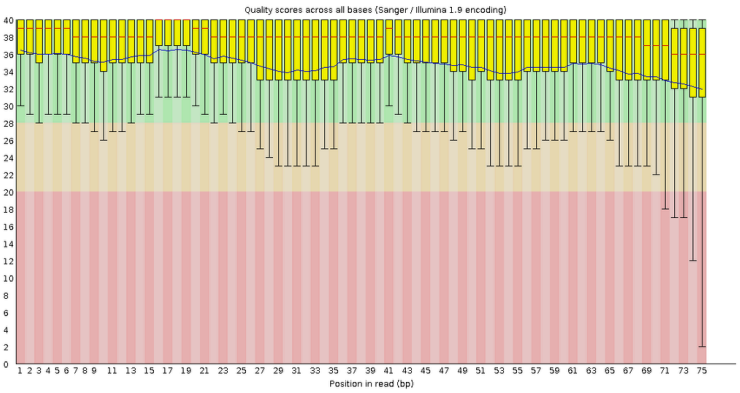


Рис. 2 Качество обратных чтений

d) **Длина чтений (Sequence Length Distribution):**
Распределение длины представлено симметричным треугольником с пиком на 75bp в обоих файлах. Это означает, что все чтения имеют строго одинаковую длину 75 нуклеотидов. Обрезанных или фрагментированных чтений не обнаружено.

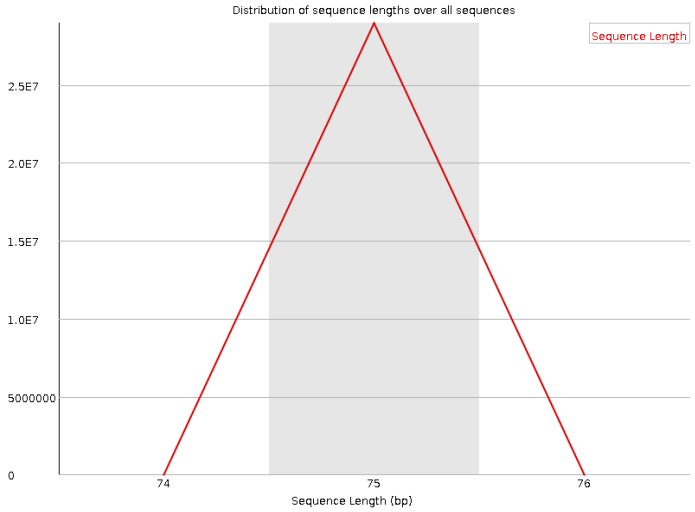


Рис. 3. Длина чтений прямых и обратных чтений до фильтрации

e) **Другие наблюдения:**

1. **Per base sequence content:** Отклонение на начальных позициях (до 9-го нуклеотида для A/T и до 7-го для C/G), скорее всего из-за праймирования.
2. **Adapter Content:** 0% адаптерного загрязнения в обоих файлах.
3. **Overrepresented sequences:** Отсутствуют.
4. ***Комментарий:*** *Данные высокого качества по всем ключевым параметрам, пригодны для дальнейшего анализа*

2.3 Фильтрация чтений

Выполненные команды:

```
TrimmomaticPE -phred33 \
  00_raw_data/dna/reads/SRR10720402_1.fastq.gz \
  00_raw_data/dna/reads/SRR10720402_2.fastq.gz \
  01_preprocessing/dna/trimmed_reads/SRR10720402_1_paired.fastq.gz \
  01_preprocessing/dna/trimmed_reads/SRR10720402_1_unpaired.fastq.gz \
  01_preprocessing/dna/trimmed_reads/SRR10720402_2_paired.fastq.gz \
  01_preprocessing/dna/trimmed_reads/SRR10720402_2_unpaired.fastq.gz \
  TRAILING:20 MINLEN:40
```

Результаты фильтрации:

- **Исходное количество пар чтений:** 28,966,798 пар
- **Парных чтений осталось:** 27,509,530 пар (94.97% от исходных)
- **Непарных чтений (только прямые):** 1,077,588 (3.72%)
- **Непарных чтений (только обратные):** 305,300 (1.05%)
- **Полностью отброшено:** 74,380 пар (0.26%)

Объяснение 4 выходных файлов:

1. `SRR10720402_1_paired.fastq.gz` — прямые чтения, у которых сохранилась парная обратная.
2. `SRR10720402_2_paired.fastq.gz` — обратные чтения, у которых сохранилась парная прямая.
3. `SRR10720402_1_unpaired.fastq.gz` — прямые чтения, чьи парные обратные были отброшены.
4. `SRR10720402_2_unpaired.fastq.gz` — обратные чтения, чьи парные прямые были отброшены.

Причина получения 4 файлов: В результате получилось 4 файла. Это связано с тем, что Trimmomatic разделяет чтения на парные и непарные: если оба чтения проходят фильтрацию — они остаются в paired, если только одно — оно уходит в unpaired.

2.4 Проверка качества триммированных чтений

Выполненные команды:

```
fastqc SRR10720402_1_paired.fastq.gz -o 01_preprocessing/dna/fastqc_trimmed

fastqc SRR10720402_2_paired.fastq.gz -o 01_preprocessing/dna/fastqc_trimmed

fastqc SRR10720402_1_unpaired.fastq.gz -o 01_preprocessing/dna/fastqc_trimmed

fastqc SRR10720402_2_unpaired.fastq.gz -o 01_preprocessing/dna/fastqc_trimmed
```

a) **Количество пар чтений после фильтрации:** 27,509,530 пар.

b) **Процент сохранённых пар чтений:** 94,97% от исходного количества.

c) **Сравнение качества paired vs unpaired:**

- **Парные чтения:** Высокое и стабильное качество. Основная масса чтений находится в зелёной зоне (Phred ≥28) на всех позициях. Лишь незначительное количество позиций имеют нижние усы, опускающиеся до Phred 26-27.
- **Непарные чтения:** Качество снижено. Особенно выражено у обратных непарных чтений, где боксы начинаются в жёлтой зоне (Phred 23-26), а усики опускаются до Phred 12.
- ***Комментарий:*** *Чтения, потерявшие пары в процессе фильтрации, имеют статистически более низкое качество, что объясняет их "одиночный" статус.*

d) **Сравнение качества до и после триммирования (только paired):**

- **До:** Фиксированная длина 75 bp, выраженное терминальное снижение качества (нижние усы до Phred 2-32).
- **После:** Переменная длина 40-75 bp, значительное улучшение качества на терминальных позициях. Исчезли "красные усы", все позиции имеют медианное качество в зелёной зоне.
- ***Комментарий:*** *Фильтрация эффективно устранила низкокачественные участки на концах чтений.*

Таблица 1. Сравнение качества чтений до и после триммирования

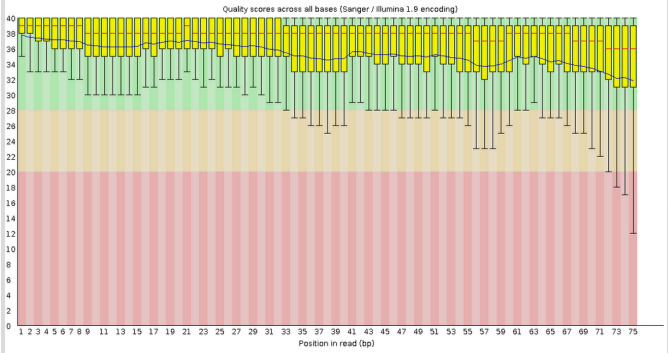


Рис. 4. Качество прямых чтений до фильтрации

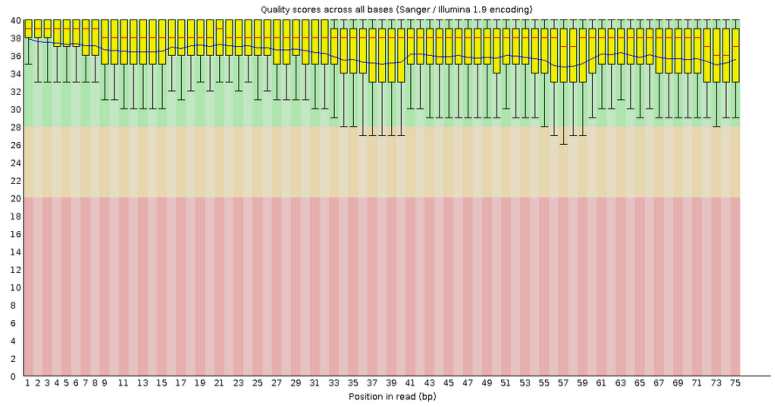


Рис. 5. Качество прямых парных чтений

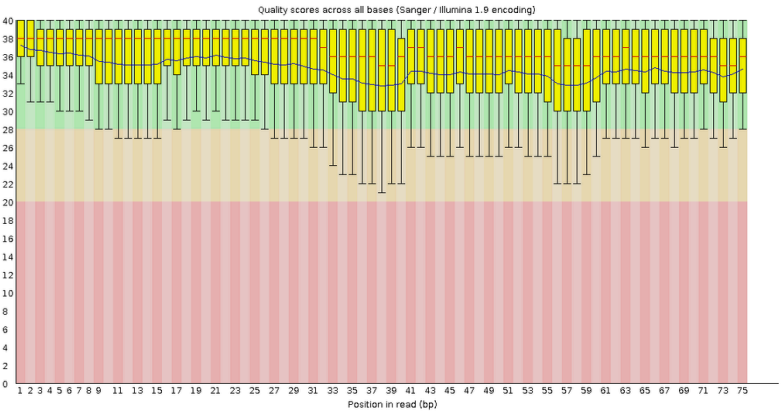


Рис. 6. Качество прямых непарных чтений

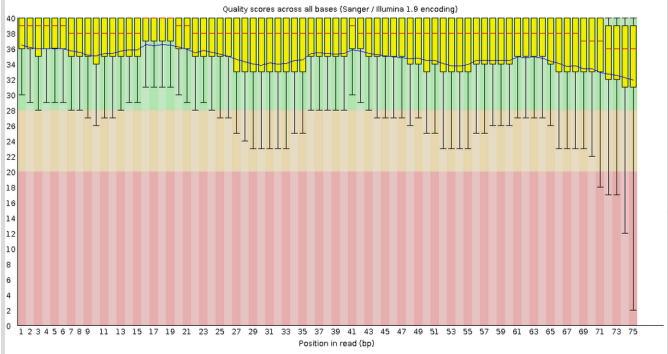


Рис. 7. Качество обратных чтений до фильтрации

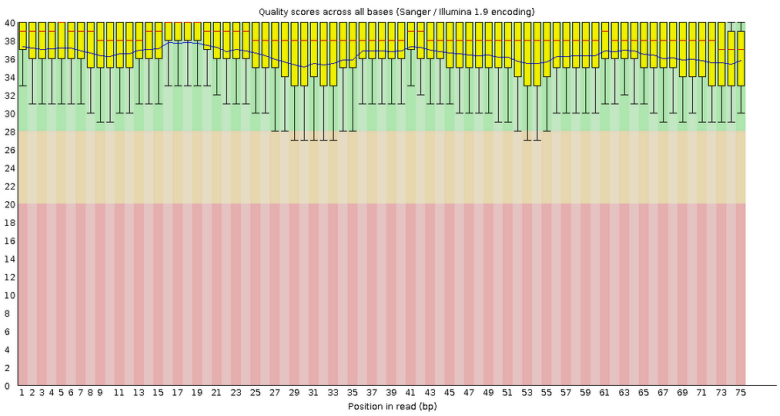


Рис. 8. Качество обратных парных чтений

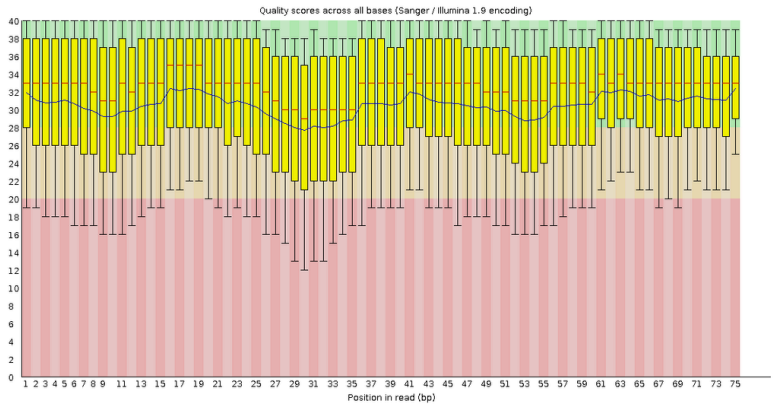


Рис. 9. Качество обратных непарных чтений

e) **Изменение длины чтений:**
Распределение длины изменилось с фиксированного (строго 75 bp) на переменное (40-75 bp). Это прямое следствие обрезки низкокачественных концов (`TRAILING:20`) и отсеивания коротких чтений (`MINLEN:40`).

Таблица 2. Сравнение изменения длины чтений до и после триммирования

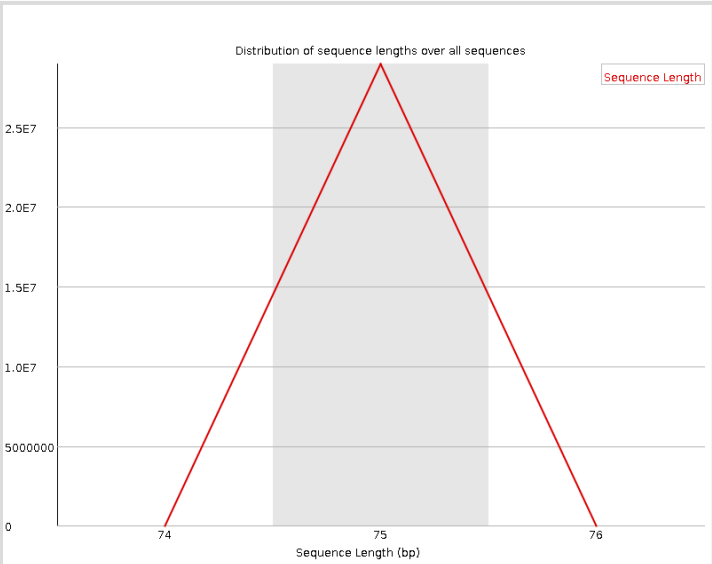


Рис. 10. Длина чтений (прямых) до фильтрации

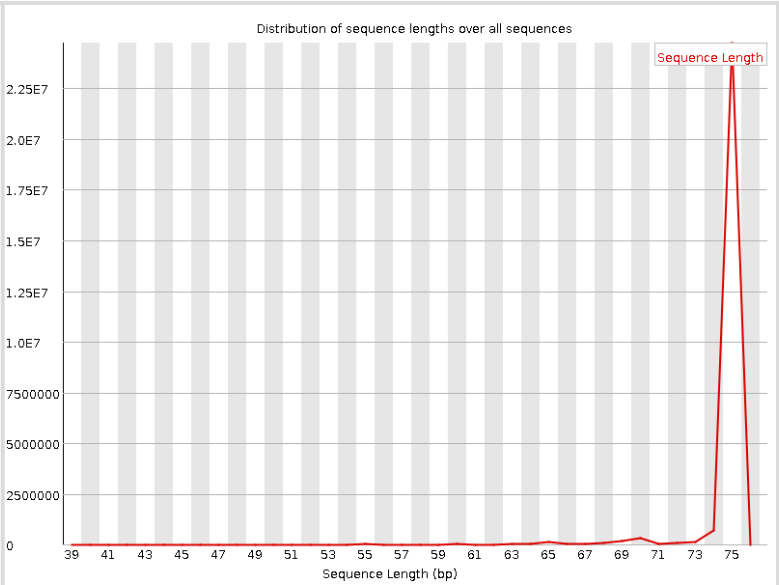


Рис. 11. Длина чтений (прямых парных) после фильтрации

Часть 3: картирование чтений на референсный геном и отбор чтений

3.1 Картирование чтений на референсный геном

Выполненные команды:

```
hisat2 -x 00_raw_data/dna/reference/chr5_index \
  -1 01_preprocessing/dna/trimmed_reads/SRR10720402_1_paired.fastq.gz \
  -2 01_preprocessing/dna/trimmed_reads/SRR10720402_2_paired.fastq.gz \
  -S 02_mapping/dna/SRR10720402.sam \
  --no-spliced-alignment \
  -p 4
```

Параметры hisat2:

- `-x`: путь к индексу референса (хромосома 5)
- `-1`: прямые чтения хорошего качества (после фильтрации)
- `-2`: обратные чтения хорошего качества (после фильтрации)
- `-S`: выходной файл в формате SAM
- `--no-spliced-alignment`: **запрет сплайсинга** (так как это ДНК, а не РНК)
- `-p 4`: использование 4 процессорных ядер для ускорения

Результаты картирования:

- **Общее количество пар чтений:** 27,509,530 пар
- **Процент общего картирования:** 5.99%
- **Конкордантно картированы 0 раз:** 26,075,510 пар (94.79%)
- **Конкордантно картированы 1 раз:** 1,342,931 пар (4.88%)
- **Конкордантно картированы >1 раз:** 91,089 пар (0.33%)
- **Дискордантно картированы:** 7,117 пар (0.03% от неконкордантных)

Размеры файлов:

- SAM файл: 10.52 ГБ
- BAM файл: 3.24 ГБ (после конвертации и сжатия)

Комментарий: *Низкий процент картирования (5.99%) ожидаем, так как мы картировали экзомные чтения (которые происходят из всех хромосом) только на одну хромосому (хромосому 5)*

3.2 Конвертация SAM в BAM

Выполненные команды:

```
samtools sort -@ 4 -o 02_mapping/dna/SRR10720402.bam 02_mapping/dna/SRR10720402.sam

samtools index 02_mapping/dna/SRR10720402.bam
```

Результаты конвертации:

- **Размер SAM файла:** 10.52 ГБ
- **Размер BAM файла:** 3.24 ГБ (сжатие примерно в 3.3 раза)

3.3 Анализ BAM файла

Выполненные команды:

```
samtools flagstat 02_mapping/dna/SRR10720402.bam
```

Результаты `samtools flagstat`:

```
55678383 + 0 in total (QC-passed reads + QC-failed reads)
55019060 + 0 primary
659323 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3952561 + 0 mapped (7.10% : N/A)
3293238 + 0 primary mapped (5.99% : N/A)
55019060 + 0 paired in sequencing
27509530 + 0 read1
27509530 + 0 read2
2868040 + 0 properly paired (5.21% : N/A)
2932898 + 0 with itself and mate mapped
360340 + 0 singletons (0.65% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Ответы на вопросы:

1. **Что значит число в поле «in total»?**
   Это общее количество записей в BAM файле (55 678 383), включая primary (55 019 060) и secondary (659 323) выравнивания. Secondary выравнивания возникают, когда чтение картируется в несколько мест с меньшим качеством.
2. **Сколько чтений (не пар!) поступило на картирование?**
   55 019 060 чтений (primary alignments), что соответствует 27 509 530 парам.
3. **Сколько чтений картировано на референс в корректных парах в штуках?**
   2 868 040 чтений картированы в корректных парах.
4. **Сколько чтений картировано на референс в корректных парах в процентах относительно поступивших на картирование?**
   5.21% чтений картированы в корректных парах.

Комментарий: *Низкий процент картирования (5.21%) ожидаем, так как мы картировали экзомные чтения (которые происходят из всех хромосом) только на одну хромосому. Из всех прочитанных экзомных фрагментов лишь небольшая часть принадлежит хромосоме 5.*

3.4 Получение чтений, картированных на хромосому

Выполненные команды:

```
# Индексация исходного BAM файла
samtools index 02_mapping/dna/SRR10720402.bam
# Отбор чтений, картированных на хромосому 5
samtools view -h -b 02_mapping/dna/SRR10720402.bam 5 >
02_mapping/dna/SRR10720402_chr5.bam
# Индексация нового файла
samtools index 02_mapping/dna/SRR10720402_chr5.bam
# Получение статистики
samtools flagstat 02_mapping/dna/SRR10720402_chr5.bam
```

Результаты `samtools flagstat` для файла с хромосомой 5:

```
4312901 + 0 in total (QC-passed reads + QC-failed reads)
3653578 + 0 primary
659323 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3952561 + 0 mapped (91.65% : N/A)
3293238 + 0 primary mapped (90.14% : N/A)
3653578 + 0 paired in sequencing
1826789 + 0 read1
1826789 + 0 read2
2868040 + 0 properly paired (78.50% : N/A)
2932898 + 0 with itself and mate mapped
360340 + 0 singletons (9.86% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Чем этот файл отличается от аналогичного файла из п.3?

1. **Резкое уменьшение общего числа записей:** с 55,678,383 до 4,312,901 (в ~12.9 раз). Это ожидаемо, так как мы оставили только чтения, картированные на хромосому 5.
2. **Увеличение процента картированных чтений:** с 7.10% до 91.65%. Теперь большинство чтений в файле картированы.
3. **Увеличение процента правильно спаренных:** с 5.21% до 78.50% от общего числа чтений в файле.
4. **Появление вторичных выравниваний (secondary alignments)** в статистике: 659,323 чтений, которые картировались в несколько мест на хромосоме 5.

Комментарий: *После фильтрации остались только чтения, относящиеся к хромосоме 5: риды, которые картировались на другие хромосомы или вообще не выровнялись, были удалены. В результате в данных осталась только та часть, которая действительно нужна для анализа именно 5 хромосомы.*

3.5 Получение только правильно картированных пар чтений

Выполненные команды:

```
# Отбор правильно спаренных чтений (флаг 2)
samtools view -h -b -f 2 02_mapping/dna/SRR10720402_chr5.bam > 02_mapping/dna/SRR10720402_chr5_proper_pairs.bam
# Индексация файла
samtools index 02_mapping/dna/SRR10720402_chr5_proper_pairs.bam
# Получение статистики
samtools flagstat 02_mapping/dna/SRR10720402_chr5_proper_pairs.bam
```

Что указано в качестве значений для параметра -f?
Флаг `-f 2` отбирает чтения, у которых установлен бит "properly paired" (правильно спаренные). В двоичном представлении флаг 2 соответствует `00000010`, что означает:

- Оба чтения в паре картированы на референс
- Они картированы в правильной ориентации (чтение 1 ⇄ чтение 2)
- Расстояние между чтениями соответствует ожидаемому размеру фрагмента библиотеки
- Оба чтения картированы на одну и ту же хромосому

Результаты `samtools flagstat` для файла с правильно спаренными чтениями:

```
3283334 + 0 in total (QC-passed reads + QC-failed reads)
2868040 + 0 primary
415294 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3283334 + 0 mapped (100.00% : N/A)
2868040 + 0 primary mapped (100.00% : N/A)
2868040 + 0 paired in sequencing
1434020 + 0 read1
1434020 + 0 read2
2868040 + 0 properly paired (100.00% : N/A)
2868040 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Чем этот файл отличается от аналогичного файла из п.4?

1. **Удаление синглтонов:** полностью отсутствуют чтения без пары (0 singletons против 360,340 в предыдущем файле).
2. **100% правильно спаренных:** все чтения в файле теперь являются правильно спаренными.
3. **Снижение количества вторичных выравниваний:** с 659,323 до 415,294, так как некоторые вторичные выравнивания могли принадлежать синглтонам.
4. **100% картированных чтений:** все чтения в файле картированы на референс.
5. **Только парные чтения:** все чтения имеют пару (2,868,040 primary paired reads).

Комментарии: *Фильтрация по флагу 2 обеспечила получение высококачественных данных — только правильно спаренные чтения, картированные на хромосому 5.*

3.6 Получение чтений, картированных только в границы экзома

Выполненные команды:

```
# Отбор чтений в границах экзома

bedtools intersect -abam 02_mapping/dna/SRR10720402_chr5_proper_pairs.bam -b
00_raw_data/dna/reference/seqcap_hg38.bed >
02_mapping/dna/SRR10720402_chr5_proper_pairs_exome.bam

# Отбор чтений в границах расширенного экзома (+50 bp)

bedtools intersect -abam 02_mapping/dna/SRR10720402_chr5_proper_pairs.bam -b
00_raw_data/dna/reference/seqcap_hg38_50.bed >
02_mapping/dna/SRR10720402_chr5_proper_pairs_exome_extended.bam
```

Результаты для экзомных чтений (исходный BED):

```
1944621 + 0 in total (QC-passed reads + QC-failed reads)
1707246 + 0 primary
237375 + 0 secondary
1707246 + 0 properly paired (100.00% : N/A)
```

Результаты для расширенного экзома (+50 bp):

```
2114490 + 0 in total (QC-passed reads + QC-failed reads)
1850817 + 0 primary
263673 + 0 secondary
1850817 + 0 properly paired (100.00% : N/A)
```

Сравнение:

- **Чтений в точном экзоме:** 1,707,246 primary парных чтений (59.5% от чтений на хромосоме 5).
- **Чтений в расширенном экзоме:** 1,850,817 primary парных чтений (64.5% от чтений на хромосоме 5).
- **Увеличение при расширении:** +143,571 чтений (+8.4%), что соответствует чтениям, картированным вблизи границ экзонов.

Комментарий: *Большинство правильно спаренных чтений на хромосоме 5 попадают в экзомные регионы. Расширение экзома на 50 bp увеличивает охват, включая чтения, картированные вблизи сплайс-сайтов или содержащие небольшие индели у границ экзонов.*

Практикум 12

Часть 1: Получение вариантов

Выполненные команды:

```
# Вызов вариантов с помощью bcftools

bcftools mpileup -f
../00_raw_data/dna/reference/Homo_sapiens.GRCh38.dna.chromosome.5.fa
../02_mapping/dna/SRR10720402_chr5_proper_pairs.bam | bcftools call -mv -o
SRR10720402_chr5.vcf

# Получение статистики по VCF файлу

bcftools stats SRR10720402_chr5.vcf > SRR10720402_chr5_stats.txt
```

Как устроен полученный VCF файл?

VCF (Variant Call Format) - текстовый формат для хранения геномных вариантов.

Файл состоит из:

- **Заголовка** (строки, начинающиеся с `##`): содержит метаинформацию (версия формата, описание полей, параметры запуска, ссылка на референс)
- **Строки с колонками** (10 обязательных колонок):
  1. `CHROM` - хромосома (в нашем случае 5)
  2. `POS` - позиция на хромосоме
  3. `ID` - идентификатор варианта в базах данных (`.` если неизвестен)
  4. `REF` - референсный аллель
  5. `ALT` - альтернативный аллель
  6. `QUAL` - качество варианта в формате Phred-score
  7. `FILTER` - статус фильтрации (`.` если не фильтровался)
  8. `INFO` - дополнительная информация (глубина покрытия, количество аллелей и т.д.)
  9. `FORMAT` - формат данных для образцов
  10. Данные для образца (генотип, likelihoods)

Пример строки:

```
5       12114   .       C       G       5.Ø4598 .
DP=1;SGB=-Ø.379885;FS=Ø;MQØF=Ø;AC=2;AN=2;DP4=Ø,Ø,1,Ø;MQ=6Ø      GT:PL
1/1:33,3,Ø
```

Расшифровка:

1. `5` — `CHROM` *(хромосома): вариант расположен на хромосоме 5*
2. `12114` — `POS` *(позиция): находится на 12,114-й нуклеотидной позиции хромосомы 5 (1-индексированная система)*
3. `.` — `ID` *(идентификатор): нет известного идентификатора в базах данных (rsID)*
4. `C` — `REF` *(референсный аллель): в референсном геноме человека в этой позиции находится цитозин (C)*
5. `G` — `ALT` *(альтернативный аллель): в анализируемом образце обнаружен гуанин (G)*
6. `5.Ø4598` — `QUAL` *(качество варианта): Phred-score ≈ 5.05*
7. `.` — `FILTER` *(статус фильтрации): не фильтровался*
8. `DP=1;SGB=-Ø.379885;FS=Ø;MQØF=Ø;AC=2;AN=2;DP4=Ø,Ø,1,Ø;MQ=6Ø` – `INFO`
   - `DP=1` — `Depth` *(глубина покрытия): всего 1 чтение покрывает данную позицию*
   - `SGB=-Ø.379885` — `Segregation based metric`*: статистика сегрегации (отрицательное значение указывает на вероятность ошибки)*
   - `FS=Ø` — `Fisher Strand bias`*: тест Фишера на страновое смещение = 0 (нет смещения)*
   - `MQØF=Ø` — `Fraction of MQØ reads`*: доля чтений с качеством картирования 0 = 0%*
   - `AC=2` — `Allele count`*: количество альтернативных аллелей в выборке = 2 (гомозиготный альтернативный)*
   - `AN=2` — `Allele number`*: общее число аллелей в выборке* `= 2` *(диплоидный геном)*
   - `DP4=Ø,Ø,1,Ø` — `Depth per strand`:
     - *Первые 2 числа: чтения, поддерживающие REF на прямой и обратной цепях (*`Ø` *и* `Ø`*)*
     - *Последние 2 числа: чтения, поддерживающие ALT на прямой и обратной цепях (*`1` *и* `Ø`*)*
     - *Итог: REF не поддерживается, ALT поддерживается 1 чтением с прямой цепи*
   - `MQ=6Ø` — `Mapping quality`*: среднее качество картирования = 60 (максимальное значение)*
9. `GT:PL` – `FORMAT`
   a. `GT` — `Genotype` *(генотип)*
   b. `PL` — `Phred-scaled genotype likelihoods` *(вероятности генотипов)*
10. `1/1:33,3,Ø` – *Данные образца*
    - `1/1` — *Генотип: гомозиготный по альтернативному аллелю (оба аллеля — G)*
    - `33,3,Ø` — *Вероятности генотипов (Phred-scaled, чем меньше — тем вероятнее):*
      - `33` — *для генотипа 0/0 (REF/REF)*
      - `3` — *для генотипа 0/1 (REF/ALT)*
      - `Ø` — *для генотипа 1/1 (ALT/ALT)*

Результаты анализа bcftools stats:

a) **Всего вариантов:** 67 898

b) **Однонуклеотидных замен (SNP):** 67 065

c) **Коротких вставок и делеций (indels):** 833

- Вставки: 428
- Делеции: 428

d) **Описание выхода bcftools mpileup:**
Команда `bcftools mpileup` анализирует BAM файл и для каждой позиции референса собирает:

- Глубину покрытия (DP)
- Качество оснований в поддерживающих чтениях
- Распределение аллелей по прямым и обратным цепям
- Статистику выравнивания
  Эти данные в двоичном формате BCF передаются в `bcftools call`, который на основе вероятностной модели вызывает варианты (SNP и индели).

Часть 2: Фильтрация вариантов

Выполненные команды:

```
# Фильтрация вариантов по качеству >30 и глубине покрытия >50

bcftools filter -i'QUAL>30 && DP>50' SRR10720402_chr5.vcf -o
SRR10720402_chr5_filtered.vcf

# Статистика по отфильтрованному файлу

bcftools stats SRR10720402_chr5_filtered.vcf >
SRR10720402_chr5_filtered_stats.txt
```

Результаты фильтрации:

a) **Вариантов после фильтрации:** 1 286 вариантов (1.89% от исходных 67 898)

b) **Однонуклеотидных замен после фильтрации:** 1 235 SNP (1.84% от исходных 67 065)

c) **Инделей после фильтрации:** 52 индели (6.24% от исходных 833)

- Вставки: 25
- Делеции: 27

Комментарий: *Фильтрация по строгим критериям (QUAL>30 и DP>50) отсеяла ~98% вариантов, что указывает на преобладание низкокачественных вариантов с малой глубиной покрытия в исходных данных. Индели сохранились лучше, чем SNP (6.12% против 1.84%), возможно, из-за их большей консервативности или особенностей вызова вариантов. Оставшиеся варианты имеют высокую достоверность и могут использоваться для дальнейшего анализа.*

Часть 3: Аннотация вариантов

Подготовка файла для VEP:

```
# Сжатие и индексация отфильтрованного VCF файла

cd /mnt/scratch/NGS/alicedol/04_annotation

bgzip -c ../03_variant_calling/SRR10720402_chr5_filtered.vcf >
SRR10720402_chr5_filtered.vcf.gz

tabix -p vcf SRR10720402_chr5_filtered.vcf.gz
```

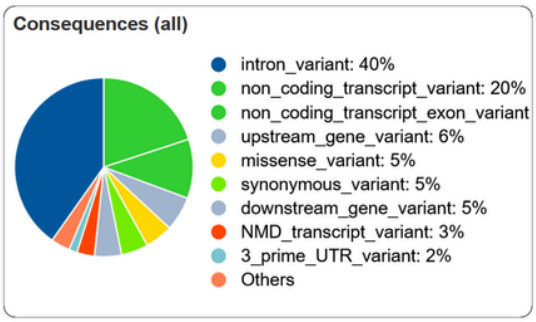
Выполнено: VCF файл `SRR10720402_chr5_filtered.vcf` загружен в онлайн-сервис VEP (Ensembl Variant Effect Predictor) с настройками для человека (GRCh38).

Результат: sep2025.archive.ensembl.org

3.1 Summary statistics из VEP

Summary statistics

Category	Count
Variants processed	1286
Variants filtered out	0
Novel / existing variants	364 (28.3) / 922 (71.7)
Overlapped genes	580
Overlapped transcripts	5224
Overlapped regulatory features	12



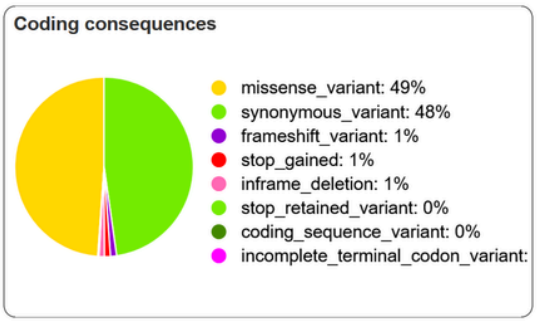


Рис. 12. Summary statistics

Таблица 3. Расшифровка Summary statistics

Полученная статистика VEP:	**Распределение по типам последствий:**	**Кодирующие последствия:**
• Обработано вариантов: 1286 • Отфильтровано VEP: 0 (все прошли) • Новые варианты: 364 (28.3%) • Известные варианты: 922 (71.7%) • Затронуто генов: 580 • Затронуто транскриптов: 5224 • Затронуто регуляторных элементов: 12	• `intron_variant`: 40% (варианты в интронах) • `non_coding_transcript_variant`: 20% (в некодирующих транскриптах) • `non_coding_transcript_exon_variant`: 5% • `upstream_gene_variant`: 6% • `missense_variant`: 5% • `synonymous_variant`: 5% • `downstream_gene_variant`: 5% • `NMD_transcript_variant`: 3% • `3_prime_UTR_variant`: 2% • Остальные: 9%	• `missense_variant`: 49% от кодирующих • `synonymous_variant`: 48% от кодирующих • `frameshift_variant`: 1% от кодирующих • `stop_gained`: 1% от кодирующих • `inframe_deletion`: 1% от кодирующих

Комментарии:

1. *Большинство обнаруженных вариантов (около **85%**) находятся в **некодирующих регионах**:*
   - `intron_variant` *(интронные варианты) — 40%*
   - `non_coding_transcript_variant` *— 20%*
   - `upstream/downstream_gene_variant` *— 11%*
   - `3_prime_UTR_variant` *и другие — около 9%*
2. *Среди кодирующих вариантов преобладают синонимичные и миссенс-замены*
   - *В кодирующих регионах:*
     - `missense_variant` *(изменение аминокислоты) — 49%*
     - `synonymous_variant` *(синонимичная замена) — 48%*
   - *Скорее всего, большинство кодирующих вариантов могут быть нейтральными или иметь умеренный функциональный эффект*
3. *Серьёзные функциональные последствия встречаются редко*
   - *Варианты с потенциально сильным эффектом (*`frameshift_variant`*,* `stop_gained`*,* `inframe_deletion`*) составляют всего ~3% от кодирующих.*
   - *Высокопатогенные варианты редки в анализируемой выборке.*
4. *Высокий процент известных вариантов*
   - *71,7% вариантов уже известны в базах данных, может говорить о распространённости в популяции*
   - *28,3% — новые варианты, могут представлять исследовательский интерес*

3.2 Варианты с IMPACT HIGH

Выполненные команды:

```
# Анализ аннотированного VCF файла для подсчета HIGH impact вариантов

grep -v "^#" vep_output.vcf | awk -F'\t' '{print $8}' | grep -o "CSQ=[^;]*" |
cut -d= -f2 | tr ',' '\n' | awk -F'|' '$3 == "HIGH"' | wc -l

# Получение детальной информации о HIGH impact вариантах

grep -v "^#" vep_output.vcf | awk -F'\t' '{print $8}' | grep -o "CSQ=[^;]*" |
cut -d= -f2 | tr ',' '\n' | awk -F'|' '$3 == "HIGH" {print $2}' | sort | uniq
-c
```

Результат: 73 варианта с IMPACT HIGH.

Распределение по типам последствий среди HIGH impact:

- `frameshift_variant`: 29
- `stop_gained`: 28
- `splice_acceptor_variant`: 9
- `splice_acceptor_variant&NMD_transcript_variant`: 3
- `splice_acceptor_variant&non_coding_transcript_variant`: 4

3.3 Характеристика вариантов с IMPACT HIGH

Таблица 4. Характеристика вариантов с IMPACT HIGH

Тип последствия	Количество	Биологическое значение
`frameshift_variant`	29	Сдвиг рамки считывания, обычно приводит к преждевременному стоп-кодону и деградации мРНК или синтезу нефункционального белка
`stop_gained`	28	Появление преждевременного стоп-кодона, приводит к неправильно сформированному нестабильному белку
`splice_acceptor_variant`	17 (включая комбинированные)	Нарушение сплайсинга, может приводить к пропуску экзона, включению интрона или изменению рамки считывания

Практикум 13

Часть 1: Описание образца

a. ID образца РНК-чтений: ENCFF763QQV

b. ссылку на информацию об образце:
www.encodeproject.org/experiments/ENCSR816HLU/

c. организм и ткань (если есть): *Homo sapiens, ткань левого лёгкого (left lung tissue), мужчина, 40 лет*

d. стратегию секвенирования (тотальная РНК, малые РНК, ...): тотальная РНК (total RNA-seq)

e. парноконцевые или одноконцевые чтения: одноконцевые

f. цепь-специфичность: Strand-specific (reverse)

Часть 2: Проверка качества исходных чтений

a. Количество чтений: 50 552 179

b. Качество чтений: Отличное. На графике "Per base sequence quality" все боксплоты находятся в зелёной зоне (Phred score > 30 по всем позициям), что свидетельствует о высоком качестве прочтения всех нуклеотидов.

c. Длина чтений: Все чтения имеют длину 100 нуклеотидов (график "Sequence Length Distribution" показывает 100% чтений длиной 100 bp).

d. Другие наблюдения:

- Адаптеры не обнаружены (график "Adapter Content" показывает 0%)
- Распределение нуклеотидов равномерное без аномалий

Вывод: Качество чтений RNA-seq высокое, фильтрация не требуется.

Часть 3: Картирование чтений на референс

Выполненная команда:

```
hisat2 -x 00_raw_data/dna/reference/chr5_index -k 3 -U
00_raw_data/rna/reads/ENCFF763QQV.fastq.gz -S 02_mapping/rna/ENCFF763QQV.sam
```

Результаты картирования:

- Всего чтений: 50,552,179
- Общий процент картирования: 5.35%
- Закартировалось ровно 1 раз: 2,550,811 (5.05%)
- Закартировалось >1 раз: 153,198 (0.30%)
- Не закартировалось: 47,848,170 (94.65%)

Преобразование SAM → BAM, выполненные команды:

```
samtools sort -o 02_mapping/rna/ENCFF763QQV.bam 02_mapping/rna/ENCFF763QQV.sam

samtools index 02_mapping/rna/ENCFF763QQV.bam
```

Отбор чтений для хромосомы 5, выполненные команды:

```
samtools view -h -b 02_mapping/rna/ENCFF763QQV.bam 5 >
02_mapping/rna/ENCFF763QQV_chr5.bam

samtools index 02_mapping/rna/ENCFF763QQV_chr5.bam
```

а) Сколько чтений закартировалось на вашу хромосому?
На хромосому 5 закартировалось 2,952,020 чтений (из 50,552,179, что составляет 5.35%). Из них:

- Primary alignments: 2,704,009 (91.6%)
- Secondary alignments: 248,011 (8.4%)

Комментарий: *Низкий процент картирования ожидаем, так как мы картировали тотальную РНК со всего генома только на одну хромосому. В реальном RNA-seq анализе картирование выполняется на весь геном, и процент картирования был бы значительно выше (в районе 70-90%).*

Часть 4: Поиск экспрессирующихся генов

4.1. Изучение файла с генной разметкой (GTF)

Файл с аннотацией генов:
`00_raw_data/rna/annotation/Homo_sapiens.GRCh38.110.chr.gtf`
Этот файл содержит информацию о всех генах, их транскриптах, экзонах и других элементах для сборки генома GRCh38 (Ensembl release 110).

Формат GTF (Gene Transfer Format) — это табулированный текст, где каждая строка описывает один геномный признак (gene, transcript, exon, CDS и т.д.) и содержит 9 обязательных полей:

1. `seqid` — идентификатор хромосомы (например, 5).
2. `source` — источник аннотации (например, ensembl_havana).
3. `type` — тип признака (gene, transcript, exon, CDS, start_codon, stop_codon).
4. `start` — стартовая позиция (1-индексация).
5. `end` — конечная позиция.
6. `score` — оценка достоверности (обычно . для отсутствия значения).
7. `strand` — цепь (+ или -).
8. `phase` — рамка считывания для CDS (0,1,2 или . для неприменимых типов).
9. `attributes` — атрибуты в формате key "value", разделенные точкой с запятой. Содержат идентификаторы (gene_id, transcript_id), имя гена (gene_name), биотип (gene_biotype) и др.

Пример первых строк файла:

```
#!genome-build GRCh38.p14

#!genome-version GRCh38

#!genome-date 2013-12

#!genome-build-accession GCA_000001405.29

#!genebuild-last-updated 2023-03

1       havana  gene    182696  184174  .       +       .       gene_id
"ENSG00000279928"; gene_version "2"; gene_name "DDX11L17"; gene_source
"havana"; gene_biotype "unprocessed_pseudogene";

1       havana  transcript      182696  184174  .       +       .
gene_id "ENSG00000279928"; gene_version "2"; transcript_id "ENST00000624431";
transcript_version "2"; gene_name "DDX11L17"; gene_source "havana";
gene_biotype "unprocessed_pseudogene"; transcript_name "DDX11L17-201";
transcript_source "havana"; transcript_biotype "unprocessed_pseudogene"; tag
"basic"; tag "Ensembl_canonical"; transcript_support_level "NA";

1       havana  exon    182696  182746  .       +       .       gene_id
"ENSG00000279928"; gene_version "2"; transcript_id "ENST00000624431";
transcript_version "2"; exon_number "1"; gene_name "DDX11L17"; gene_source
"havana"; gene_biotype "unprocessed_pseudogene"; transcript_name "DDX11L17-
201"; transcript_source "havana"; transcript_biotype "unprocessed_pseudogene";
exon_id "ENSE00003759020"; exon_version "2"; tag "basic"; tag
"Ensembl_canonical"; transcript_support_level "NA";

1       havana  exon    183132  183216  .       +       .       gene_id
"ENSG00000279928"; gene_version "2"; transcript_id "ENST00000624431";
transcript_version "2"; exon_number "2"; gene_name "DDX11L17"; gene_source
"havana"; gene_biotype "unprocessed_pseudogene"; transcript_name "DDX11L17-
201"; transcript_source "havana"; transcript_biotype "unprocessed_pseudogene";
exon_id "ENSE00003759581"; exon_version "2"; tag "basic"; tag
"Ensembl_canonical"; transcript_support_level "NA";

1       havana  exon    183494  183571  .       +       .       gene_id
"ENSG00000279928"; gene_version "2"; transcript_id "ENST00000624431";
transcript_version "2"; exon_number "3"; gene_name "DDX11L17"; gene_source
"havana"; gene_biotype "unprocessed_pseudogene"; transcript_name "DDX11L17-
201"; transcript_source "havana"; transcript_biotype "unprocessed_pseudogene";
exon_id "ENSE00003804405"; exon_version "1"; tag "basic"; tag
"Ensembl_canonical"; transcript_support_level "NA";
```

Таким образом, файл позволяет сопоставить каждое прочтение с конкретным геном.

(*) Сколько на вашей хромосоме аннотировано генов?

Чтобы узнать общее число генов на хромосоме 5, можно выполнить:

```
grep -w "gene" 00_raw_data/rna/annotation/Homo_sapiens.GRCh38.110.chr.gtf |
cut -f9 | grep -o 'gene_id "[^"]*"' | sort -u | wc -l
```

В нашем файле аннотировано **62 700** генов (это включает различные биотипы: белок-кодирующие, псевдогены, длинные некодирующие РНК и т.д.).

4.2. Подсчёт числа чтений на гены с помощью htseq-count

Для подсчёта количества прочтений, попавших в каждый ген, использована программа htseq-count (входит в пакет HTSeq). Запуск производился следующей командой:

```
htseq-count -f bam -s reverse -m union -t gene \

    02_mapping/rna/ENCFF763QQV_chr5.bam \

    00_raw_data/rna/annotation/Homo_sapiens.GRCh38.110.chr.gtf \

    > 05_rna_analysis/counts/ENCFF763QQV_chr5_gene_counts.txt \

    2> logs/htseq_count.log
```

Объяснение параметров:

- `-f bam` — входной файл в формате BAM (бинарный вариант SAM).
- `-s reverse` — режим учёта цепи. Поскольку эксперимент цепь-специфичный (strand-specific) и протокол указан как "reverse", чтения соответствуют обратной цепи. Параметр reverse означает, что если чтение картировано на цепь +, то оно считается принадлежащим гену на цепи - и наоборот. (Если бы эксперимент был неспецифичным, использовали бы `-s no`.)
- `-m union` — режим разрешения неоднозначностей. `union` означает, что чтение засчитывается гену, если оно полностью попадает в его границы; если чтение перекрывает несколько генов (например, перекрывающиеся гены), оно попадает в категорию `__ambiguous.`
- `-t gene` — тип признака, по которому происходит подсчёт. В GTF-файле строки с типом gene определяют границы гена. Можно было бы считать по экзонам (exon), но задание требует подсчёт на уровне генов.
- *Последние два аргумента* — путь к BAM-файлу и к файлу аннотации.
- Результат перенаправлен в текстовый файл ****_gene_counts.txt*, сообщения об ошибках и прогрессе — в лог-файл.

Работа программы заняла около 10–15 минут. Полученный файл содержит две колонки: идентификатор гена (ENSEMBL) и количество чтений, попавших в этот ген. В конце файла добавлены специальные строки, начинающиеся с символа подчёркивания, которые суммируют чтения, не попавшие однозначно ни в один ген.

4.3. Анализ результатов

Вывод первых и последних строк:

```
$ head -20 05_rna_analysis/counts/ENCFF763QQV_chr5_gene_counts.txt

ENSG00000000003 0

ENSG00000000005 0

ENSG00000000419 0

ENSG00000000457 0

ENSG00000000460 0

ENSG00000000938 0

ENSG00000000971 0

ENSG00000001036 0

ENSG00000001084 0

ENSG00000001167 0

$ tail -10 05_rna_analysis/counts/ENCFF763QQV_chr5_gene_counts.txt

ENSG00000292369 0

ENSG00000292370 0

ENSG00000292371 0

ENSG00000292372 0

ENSG00000292373 0

__no_feature    272633

__ambiguous     161394

__too_low_aQual 0

__not_aligned   0

__alignment_not_unique  153198
```

а) Сколько чтений попало в границы генов?
Суммируем все значения во втором столбце, исключая строки, начинающиеся с __:

```
$ grep -v "^__" ENCFF763QQV_chr5_gene_counts.txt | awk '{sum+=$2} END {print
sum}'

2116784
```

Ответ: 2 116 784 чтений (из числа **primary alignments**, т.е. основных картирований на хромосому 5) попало в границы аннотированных генов. Это примерно 78,3% от всех primary alignments (2 704 009) и 71,7% от всех чтений, закартировавшихся на хромосому 5 (2 952 020). Большая часть чтений, картировавшихся на хромосому 5, действительно происходит из транскрибируемых областей.

б) Сколько чтений попало мимо границ генов?
Из специальных строк берём значения:

- `__no_feature` — 272 633 чтения не перекрываются ни с одним аннотированным геном. Это чтения из межгенных регионов или участков, не аннотированных в текущей версии GTF.
- `__ambiguous` — 161 394 чтения перекрывают несколько генов одновременно (например, перекрывающиеся гены на противоположных цепях или вложенные гены). Такие чтения исключены из подсчёта, поскольку нельзя определить, к какому именно гену они относятся.
- `__alignment_not_unique` — 153 198 чтений имеют множественные картирования (с одинаковым лучшим качеством) и поэтому не учитываются.

Итого мимо однозначно определённых генов (т.е. чтения, которые не попали в категорию однозначного попадания в один ген) оказалось:
272 633 + 161 394 + 153 198 = **587 225** чтений. Это около 21,7% от primary alignments.

Строки `__too_low_aQual` и `__not_aligned` равны 0, что ожидаемо, поскольку мы уже отфильтровали чтения по качеству картирования (входной BAM содержит только успешно закартированные чтения).

(*) **Объяснение строк аннотационного файла, начинающихся с “_”:**

Каждая такая строка подводит итог по определённой категории проблемных чтений:

- `__no_feature` — чтения, которые не перекрываются ни с одним признаком типа gene (т.е. не попали ни в один аннотированный ген). Они могут располагаться в межгенных областях или в участках, для которых в GTF нет записи типа gene (например, в интронах, если считать только гены целиком, а не экзоны).
- `__ambiguous` — чтения, перекрывающие несколько генов одновременно. Программа не может решить, к какому гену их отнести, поэтому они помечаются как неоднозначные.

- `__too_low_aQual` — чтения с низким качеством картирования (обычно MAPQ < порогового значения, которое по умолчанию равно 0, поэтому такие чтения игнорируются). В нашем случае значение 0.
- `__not_aligned` — чтения, которые не были картированы (в нашем BAM таких нет, так как мы отобрали только закартированные на хромосому 5).
- `__alignment_not_unique` — чтения, которые картировались в несколько мест генома с одинаково высоким качеством (multi-mappers). Программа не может выбрать единственное лучшее расположение, поэтому такие чтения исключаются.

Вывод: Результаты htseq-count показывают, что примерно 78% чтений, картировавшихся на хромосому 5, скорее всего, могут быть отнесены к конкретным генам. Остальные чтения либо находятся вне аннотированных генов, либо имеют неоднозначное картирование. Это нормальная ситуация для RNA-seq данных: часть транскриптов может происходить из слабо аннотированных регионов, часть чтений — повторяющиеся элементы (откуда возникают мульти-маппинги), а перекрывающиеся гены часто встречаются в геноме.

Часть 5: Аннотация высоко экспрессируемых генов

5.1. Топ-10 самых высоко экспрессируемых генов

Файл с экспрессионным профилем (`ENCFF763QQV_chr5_gene_counts.txt`) был отсортирован по убыванию количества чтений (второй столбец), строки, соответствующие аннотированным генам, отобраны исключением специальных категорий (начинающихся с "__").

```
grep -v "^__" 05_rna_analysis/counts/ENCFF763QQV_chr5_gene_counts.txt | sort -k2,2nr > 05_rna_analysis/counts/ENCFF763QQV_chr5_gene_counts_sorted.txt
```

Дополнительно из GTF-файла были извлечены названия генов (gene_name) и биотипы (gene_biotype) для каждого идентификатора.

Таблица 5. Топ-10 самых высоко экспрессируемых генов на хромосоме 5 в образце лёгкого

Gene ID	Counts	Gene name	Gene biotype
ENSG00000250182	101 137	EEF1A1P13	processed_pseudogene
ENSG00000223361	41 494	FTH1P10	transcribed_processed_pseudogene
ENSG00000153395	29 537	LPCAT1	protein_coding
ENSG00000127022	28 859	CANX	protein_coding
ENSG00000113448	28 608	PDE4D	protein_coding
ENSG00000134352	27 565	IL6ST	protein_coding
ENSG00000247627	27 504	MTND4P12	processed_pseudogene
ENSG00000019582	27 288	CD74	protein_coding
ENSG00000038427	27 200	VCAN	protein_coding
ENSG00000113140	24 943	SPARC	protein_coding

Комментарий. *Среди наиболее экспрессирующихся генов присутствуют как функциональные белок-кодирующие гены, так и псевдогены. Псевдогены часто имеют высокие значения каунтов из-за гомологии с функциональными генами, что может приводить к неоднозначному картированию чтений. Для дальнейшего анализа выбран белок-кодирующий ген* ***LPCAT1****, занимающий третье место по уровню экспрессии и имеющий прямое отношение к ткани образца (лёгкое).*

5.2. Визуализация гена LPCAT1

Для визуализации был использован геномный браузер Ensembl (сборка GRCh38.p14)

Рис. 13. Регион гена LPCAT1 в браузере Ensembl.

На рисунке видны экзон-интронная структура гена (трек GENCODE 49) и трек консервативности GERP. Экзоны выделены утолщёнными блоками, интроны — линиями. Пики консервативности совпадают с экзонами, особенно в белок-кодирующих участках.

*Наличие трека консервативности позволяет оценить, насколько разные участки гена сохранялись в ходе эволюции. В случае LPCAT1 видно, что экзоны обладают высокой консервативностью — это обычно указывает на их функциональную значимость. Вероятно, это связано с тем, что белок участвует в синтезе лёгочного сурфактанта, то есть выполняет важную для организма функцию.

5.3. Функция гена LPCAT1 (по данным NCBI Gene)

LPCAT1 (lysophosphatidylcholine acyltransferase 1) — белок-кодирующий ген, расположенный на хромосоме 5 (5p15.33). Кодируемый фермент принадлежит к семейству ацилтрансфераз и катализирует превращение лизофосфатидилхолина в фосфатидилхолин в присутствии ацил-КоА. Этот процесс критически важен для:

- синтеза лёгочного сурфактанта — вещества, снижающего поверхностное натяжение в альвеолах и предотвращающего их спадение;
- образования фактора активации тромбоцитов (PAF), участвующего в воспалительных и аллергических реакциях.

Согласно данным NCBI Gene, LPCAT1 наиболее активно экспрессируется в лёгких (RPKM 73.0), селезёнке (RPKM 32.6) и других тканях.