Практикум 11

Часть 1

1. Подготовка референса

1) Индексация референса с помощью hisat2

Для последующего картирования индексируем референсный геном, которым является 7 хромосома человека из референсной последовательности генома человека версии GRCh38.p14 (ensembl GCA_000001405.29).

```
hisat2-build chr7.fa chr7
```

Command: hisat2-build – строит индексы (в нашем случае используя 32-битные числа) референсного генома
Options: chr7 – префикс имен выходных файлов
Input: chr7.fa – имя fasta-файла с референсным геномом
Output: chr7.1.ht2, chr7.2.ht2, chr7.3.ht2, chr7.4.ht2, chr7.5.ht2, chr7.6.ht2, chr7.7.ht2, chr7.8.ht2

2) Индексация референса с помощью samtools

Особая индексация референса необходима для некоторых программ.

```
samtools faidx chr7.fa
```

Command: samtools
Options: faidx – обеспечивает доступ к fasta (и fastq) файлам в input
Input: chr7.fa - имя fasta-файла с референсным геномом
Output: chr7.fa.fai

В файле: 7 159345973 56 60 61
Точное имя хромосомы: 7
Длина хромосомы в нуклеотидах: 159345973

56: индекс байта, с которого начинается сама нуклеотидная последовательность в fasta (в input) (то есть до самой последовательности находится 55 байтов, соответствующие, имени хромосомы и другим данным в первой строке после >)
60: длина строки fasta (в input) в нуклеотидах
61: длина строки fasta (в input) в байтах (*перенос строки в Linux это 1 байт, поэтому разница 1 между длиной в нуклеотидах и байтах это символ переноса строки)

2. Чтения ДНК

1) Описание образца чтений ДНК

SRR ID образца ДНК-чтений: SRR10720413
*Ссылка на информацию об образце в NCBI SRA:
https://www.ncbi.nlm.nih.gov/sra/?term=SRR10720413

Прибор для секвенирования: Illumina Genome Analyzer IIx
Организм: Homo sapiens
Стратегия секвенирования: whole-exome and RNA sequencing (экзомное)
Какие чтения: парноконцевые
Сколько чтений ожидается (spots): 38,132,996

2) Проверка качества исходных чтений

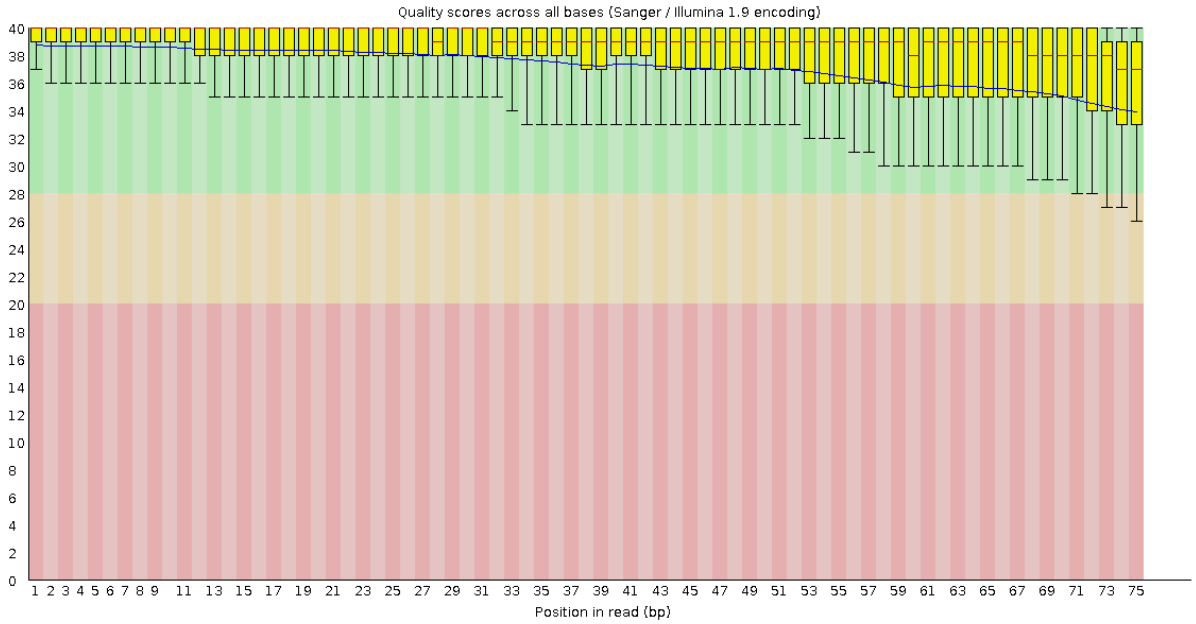
У меня есть 2 файла (так как парноконцевые чтения) - SRR10720413_1.fastq.gz (прямые) и SRR10720413_2.fastq.gz (обратные).

```
fastqc SRR10720421_1.fastq.gz SRR10720421_2.fastq.gz
```

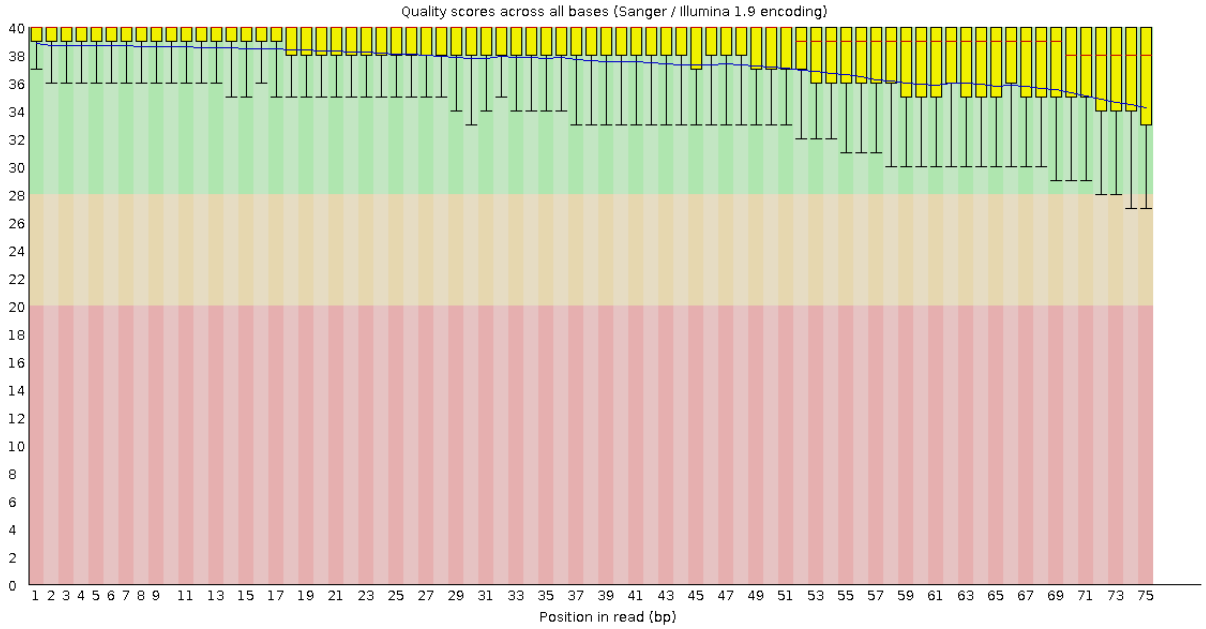
Получилось 2 html-файла, для прямых чтений (_1) и для обратных (_2).

Количество пар чтений: 38,132,996
Совпадает ли количество чтений у прямых и обратных: да, совпадает также с ожидаемым количеством

a) Анализ качества пар чтений:



a) прямые чтения



b) обратные чтения

Рис. 1. Per base sequence quality

Видим графики вида "box plot" (ящик с усами), отображающие распределение показателей качества (Quality scores) по каждой позиции в прочтении (read).

- Желтый прямоугольник («ящик»): представляет собой межквартильный размах (IQR). Он содержит 50% всех данных (от 25-го до 75-го процентиля).
  - Горизонтальная линия внутри ящика — это медиана (50-й процентиль).
- Вертикальные линии («усы»): показывают диапазон данных, не считая выбросов.
  - Верхний «ус» — это максимальное значение в пределах 1.5 * IQR от верхнего квартиля.
  - Нижний «ус» — минимальное значение в пределах 1.5 * IQR от нижнего квартиля.
- Синяя линия, проходящая через центр ящиков: это среднее значение (mean) качества по каждой позиции. Она помогает визуально отследить общую тенденцию изменения среднего качества.

Цветовые полосы на фоне графика — это качественные шкалы, соответствующие стандартной кодировке Illumina:
- Красная зона (ниже ~20): очень низкое качество. Такие базы считаются ненадежными и часто отбрасываются при обработке.
- Желтая зона (~20–30): среднее качество. Базы в этой зоне могут быть использованы, но требуют осторожности.
- Зеленая зона (выше ~30): высокое качество. Это надежные базы, которые обычно используются в анализе без фильтрации.

Общая картина:
Оба графика демонстрируют типичную для Illumina-данных динамику снижения качества к концу прочтения. Однако между ними есть небольшие, но важные различия, особенно в последних позициях.

Что общего:
- Высокое качество в начале: В обоих чтениях медиана и среднее значение на позициях 1–50 находятся в зеленой зоне (Q > 35). Это говорит о хорошем стартовом качестве.
- Снижение к концу: В обоих случаях наблюдается постепенное снижение среднего и медианного качества после 50 bp.
- Отсутствие резких аномалий: Нет внезапных падений или выбросов, что указывает на стабильность секвенирования.

Ключевые различия:
1. Качество в конце прочтения (позиции 65–75)
   a) прямое чтение (первый график):

- Медиана (горизонтальная линия в ящике) падает ниже 35 уже к позиции 60.
- К позиции 75 нижний "ус" опускается до ~26–27, что находится в желтой зоне (низкое/среднее качество).
- Синяя линия (среднее) также заметно снижается — к концу она близка к 34.

b) обратное (второй график):

- Медиана остается выше 35 почти до позиции 70.
- К позиции 75 нижний "ус" не опускается ниже 30, что еще в пределах зеленой зоны.
- Среднее значение (синяя линия) снижается медленнее — к концу оно около 36–37.

Вывод: обратное чтение сохраняет более высокое качество в конце прочтения, чем 1. Это частая ситуация при парном секвенировании — обратное чтение часто имеет лучшее качество на концах, поскольку оно начинается с "чистого" участка матрицы

2. Разброс данных

В первом разброс (высота ящика и длина усов) несколько больше, особенно в последней трети прочтения — это означает, что больше разброса.
Во втором распределение более компактное — данные более однородны.
Вывод: обратное чтение более стабильно по качеству, особенно в конце.

b) Анализ длины чтений:

вставила только 1 картинку, потому что они одинаковые для обоих типов чтений.

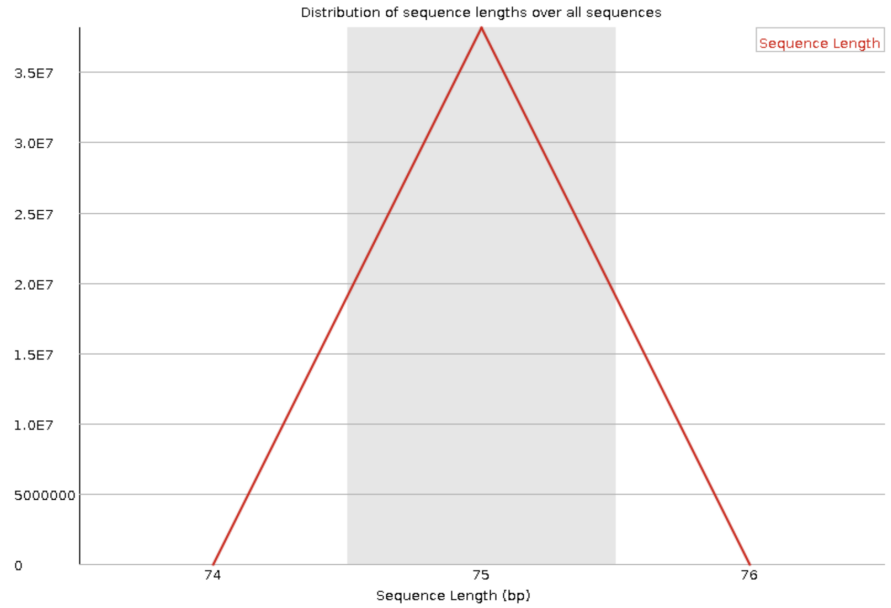


Рис. 2. Sequence length distribution

Этот график — это линейный график распределения частот (frequency distribution) длин прочтений.

Длина прочтений почти идеально однородна
График имеет острый пик на 75 bp и очень маленькое количество прочтений на 74 и 76 bp. Это говорит о том, что:
Практически все прочтения имеют длину ровно 75 bp.
Такая картина характерна для стандартных протоколов Illumina, где прочтения фиксированной длины (например, 75 bp) секвенируются без обрезки до анализа.

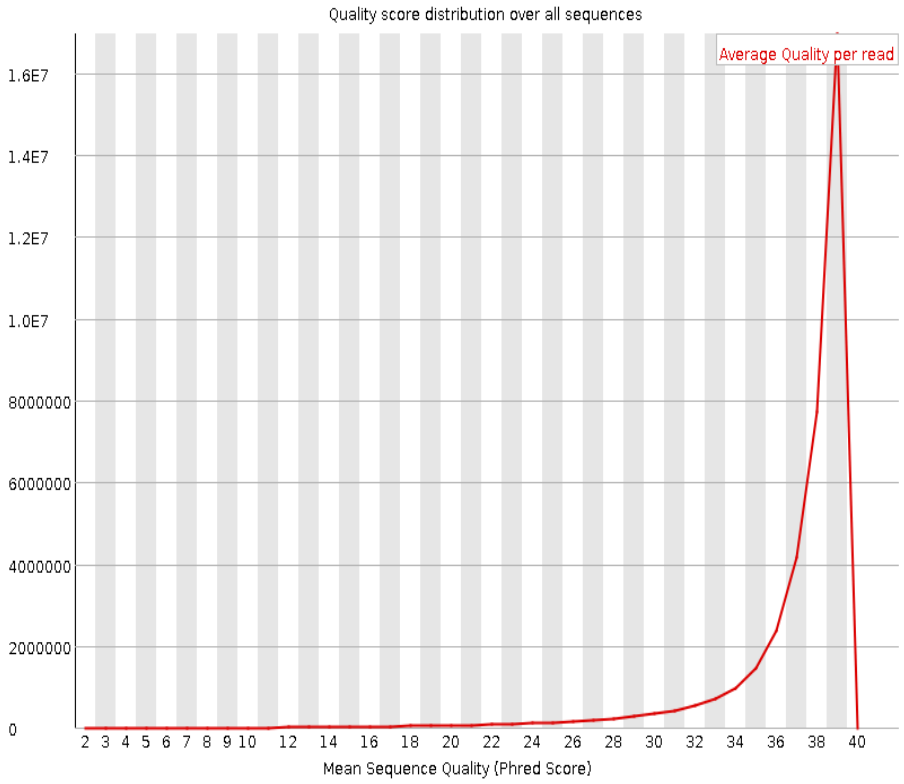
Небольшое количество коротких и длинных прочтений:
На 74 bp и 76 bp есть очень небольшой "хвост" — по нескольку миллионов прочтений.
Это может быть вызвано:
- Техническими проблемами (например, сбой в секвенаторе, ошибки при записи данных).
- Обрезкой адаптеров или качеством — если в начале или конце прочтения были низкокачественные базы, они могли быть удалены, что привело к укорочению некоторых прочтений.
- Ошибками в библиотеке — например, если часть молекул была слегка короче или длиннее из-за вариаций в подготовке.

Однако, поскольку эти значения малы по сравнению с пиком на 75 bp, их можно считать незначительными выбросами.
Из графика видно:
- Пик на 75 bp достигает ~35 миллионов прочтений — это абсолютное большинство.
- На 74 bp — около 5–10 миллионов.
- На 76 bp — еще меньше, примерно 5 миллионов.

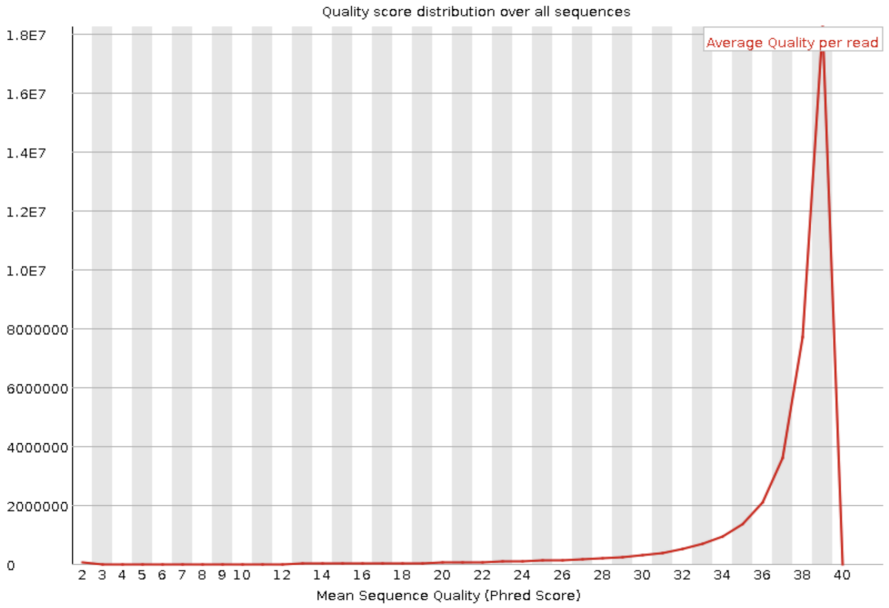
Примерное соотношение:
- 75 bp: ~85–90% всех прочтений
- 74 bp и 76 bp: ~10–15% в сумме

c) Пару слов про остальные графики:

1) Если продолжить тему качества чтений, то можно взглянуть на распределение чтений по среднему качеству.



a) прямые чтения



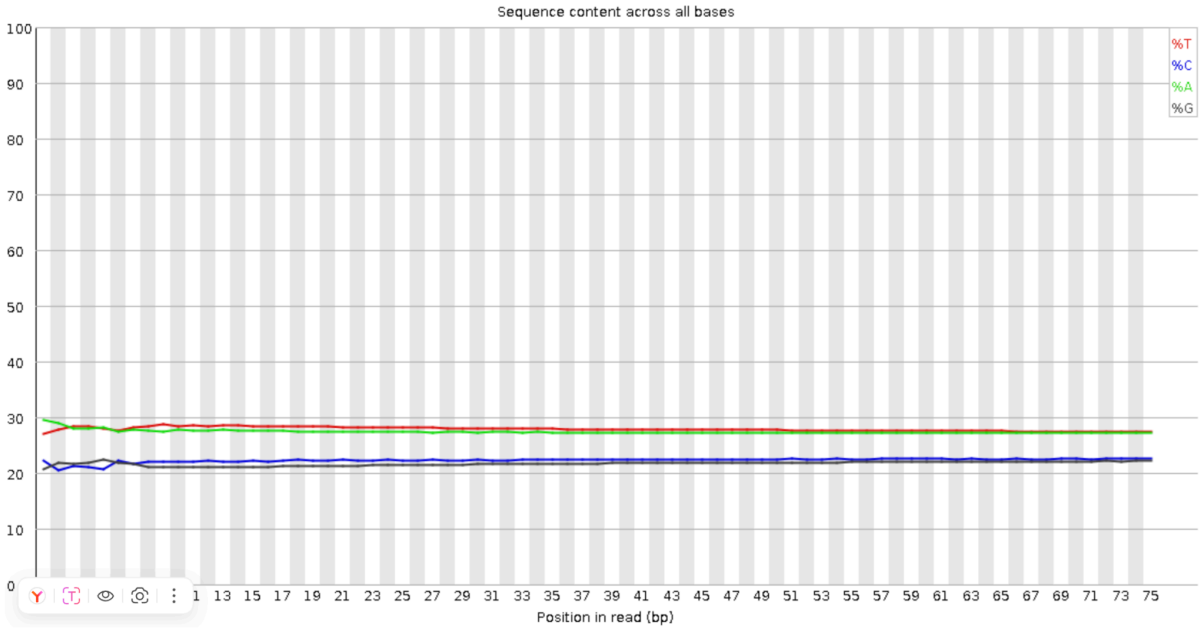
b) обратные чтения

Рис. 3. Per sequence quality scores

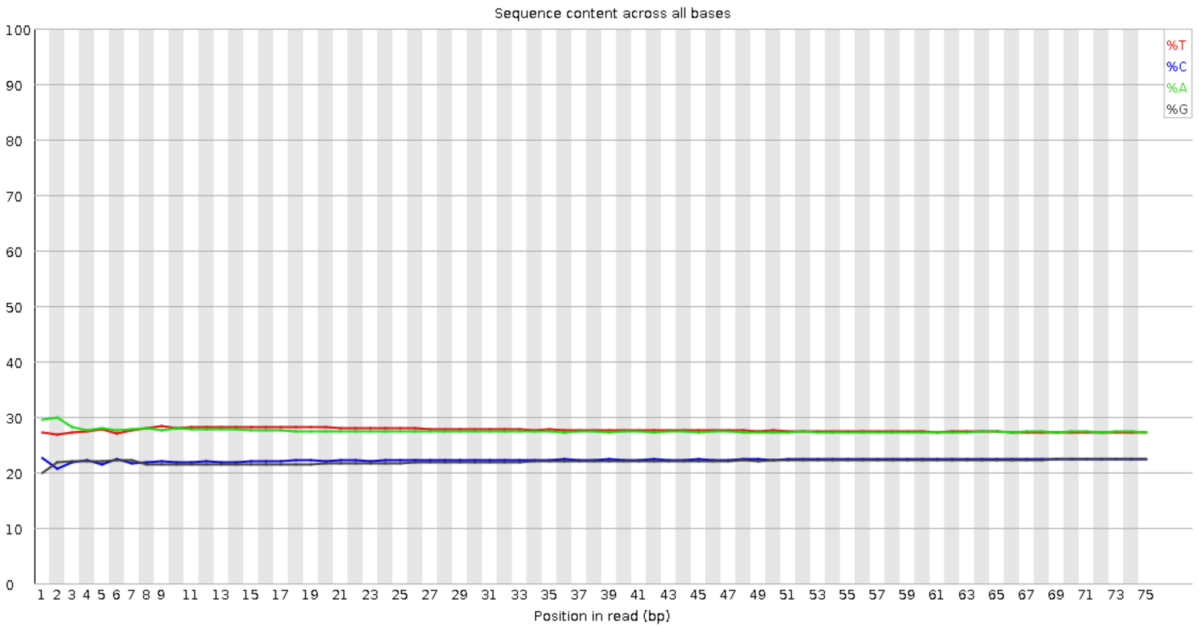
Подавляющее большинство прочтений имеют высокое-среднее качество
- Красная линия резко поднимается только в конце, начиная с ~Q36.
- Пик гистограммы (самый высокий серый столбец) находится на уровне Q38–Q40.
- Это означает, что почти все прочтения имеют среднее качество выше 35, а большинство — выше 38.

Предыдущий график (Рис.1) показывал качество по каждой позиции внутри прочтения — там мы видели снижение к концу. Этот график показывает среднее качество по всему прочтению — то есть каждое прочтение получает одну оценку (усреднённую по всем его базам).

2)По следующему графику можно судить о процентном количестве видв нуклеотидов в данной позиции. Считается, что, если в какой-то позиции разница |A-T| или |G-C| больше 10%, то это плохо (в процессе секвенирования что-то пошло не так, и получившиеся чтения лучше переделать).



а) прямые чтения

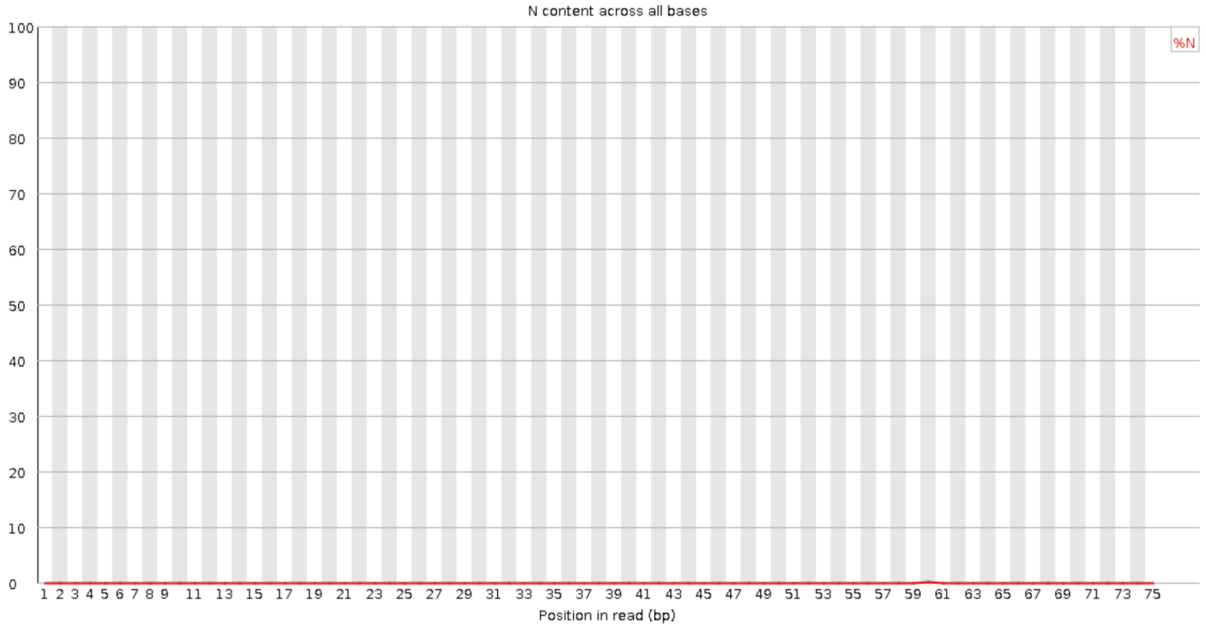


б) обратные чтения

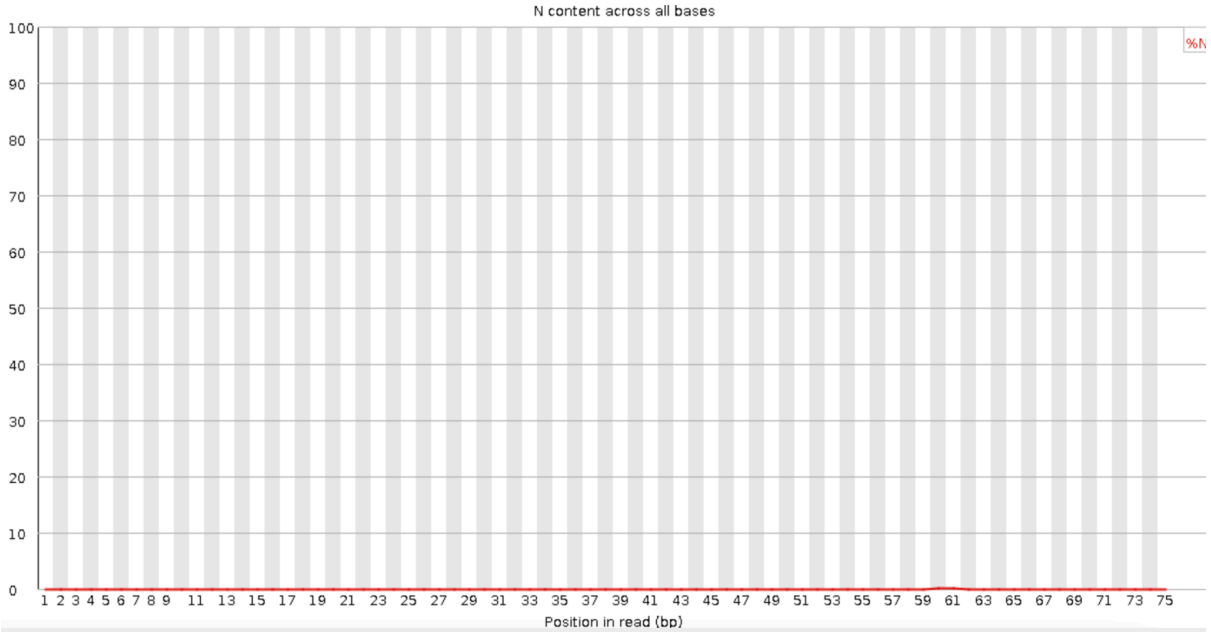
Рис. 4. Per base sequence content

На моем графике A и T практически везде совпадают — их линии почти наложены друг на друга, есть незначительные расхождения вначале. G и C также очень близки друг к другу — обе линии на уровне ~22–24.

3) Процент знака N (неопределенный нуклеотид) в позиции. Насторожить должно значение выше 5%, на моих графиках все 0%, то есть очень хорошо



а) прямые чтения



б) обратные чтения

Рис. 5. Per base N content.

Для оценки использовала мануал fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
Общая оценка качества чтений: замечательно

3) Фильтрация чтений с помощью trimmomatic

```
TrimmomaticPE   -phred33   SRR10720413_1.fastq.gz
SRR10720413_2.fastq.gz trim_1_paired.fastq.gz
trim_1_unpaired.fastq.gz    trim_2_paired.fastq.gz
trim_2_unpaired.fastq.gz   TRAILING:20   MINLEN:50
```

Command: TrimmomaticPE - запускаем trimmomatic для парноконцевых чтений (поэтому PE, SE - для одноконцевых)
Options: -phred33 - данный Quality Score
TRAILING:20 - удаляем с конца чтений нуклеотиды с качеством ниже 20
MINLEN:50 - удаляем чтения с длиной меньше 50
Input: SRR10720413_1.fastq.gz, SRR10720413_2.fastq.gz - входные файлы с чтениями
Output: trim_1_paired.fastq.gz trim_1_unpaired.fastq.gz trim_2_paired.fastq.gz trim_2_unpaired.fastq.gz - выходные файлы, _paired - парные, _unpaired - непарные

4) Анализ триммированных чтений fastqc

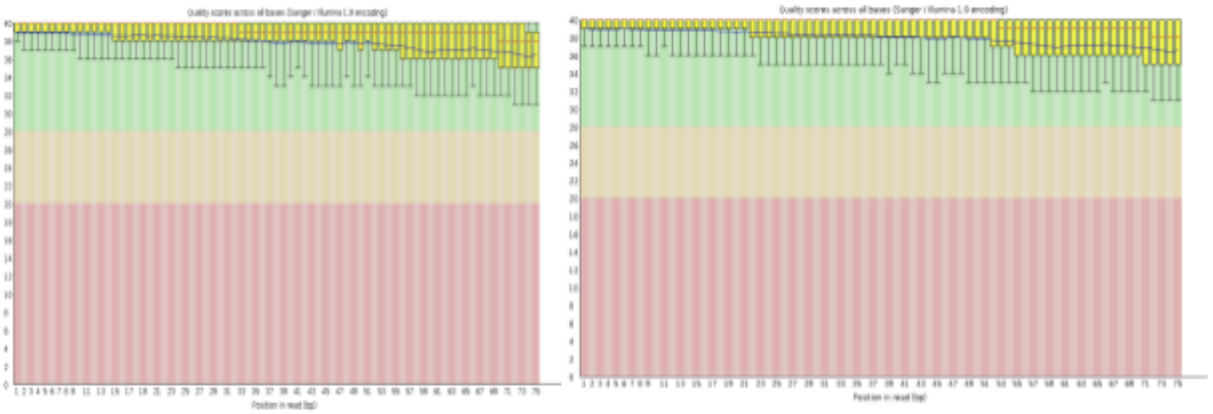
```
fastqc trim*
```

Количество пар чтений осталось: 36,950,563
Процент пар чтений осталось: 96.9%

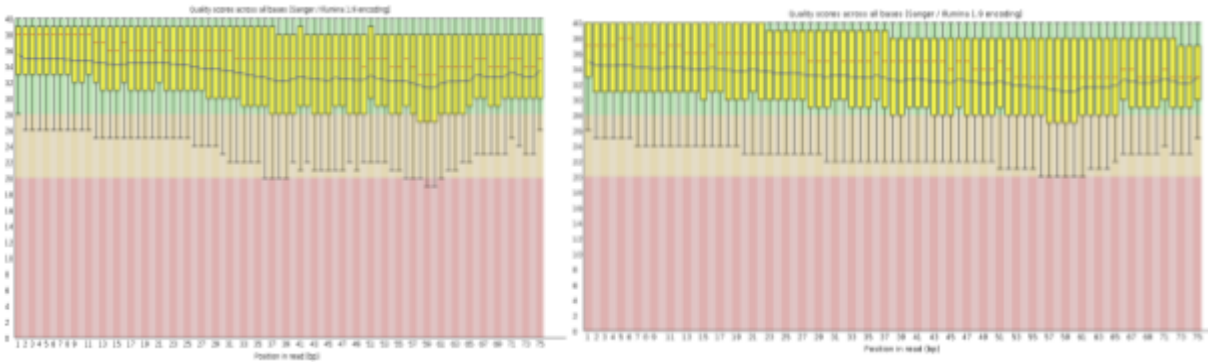
a) **Сравнение качества чтений после триммирования (paired vs unpaired):**

Качество непарных чтений заметно хуже, чем качество парных
После триммирования количество непарных чтений оказалось небольшим, и их среднее качество (Q20–Q25) было заметно ниже, чем у сохранённых парных чтений (Q30+). Это ожидаемо: непарные чтения представляют собой остатки тех пар, где один из ридов не прошёл фильтрацию по длине или качеству. Их более низкое качество отражает селективный характер предобработки данных и подтверждает эффективность триммирования.

парные чтения



прямые *обратные*



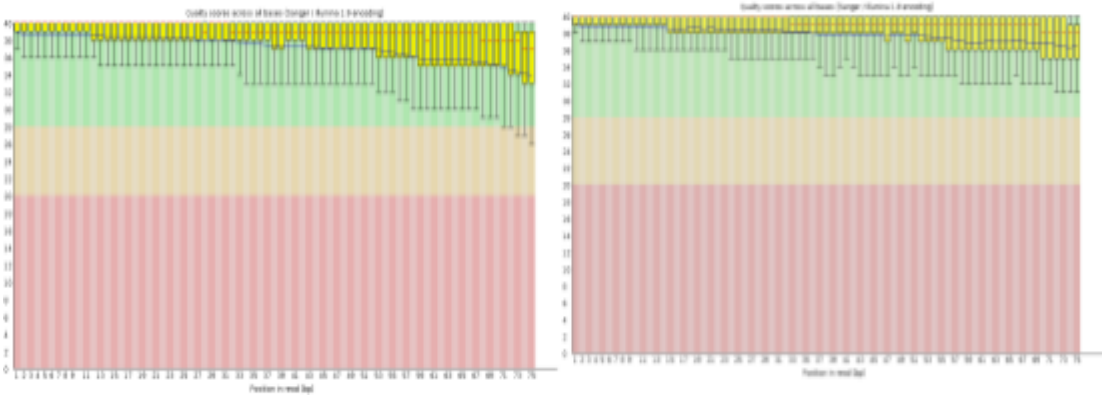
непарные

Рис. 6. Per base sequence quality после триммирования

b) С**равнение качества чтений до и после триммирования (только paired):**

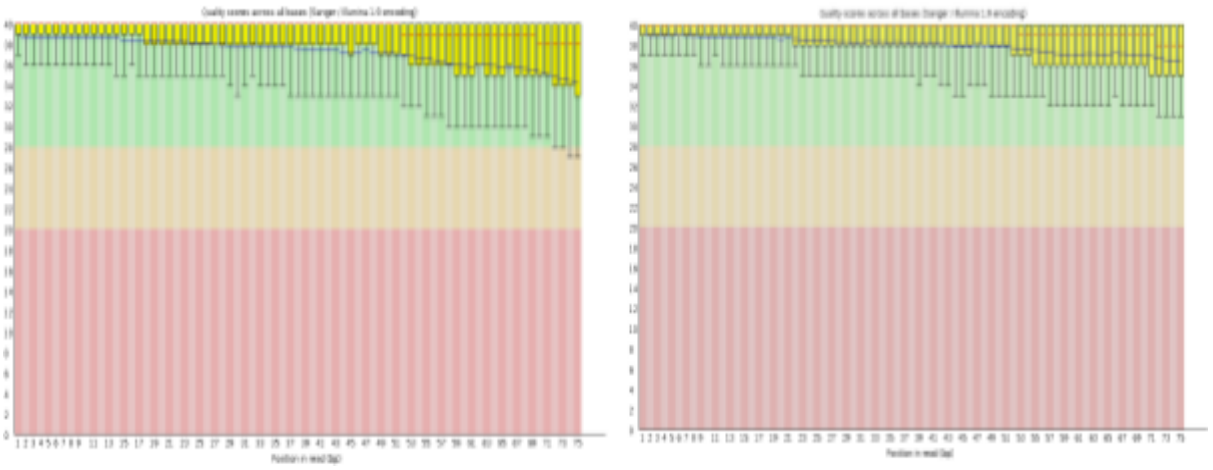
после триммирования качество чтений улучшилось

парные чтения



до триммирования *после триммирования*

Обратные чтения



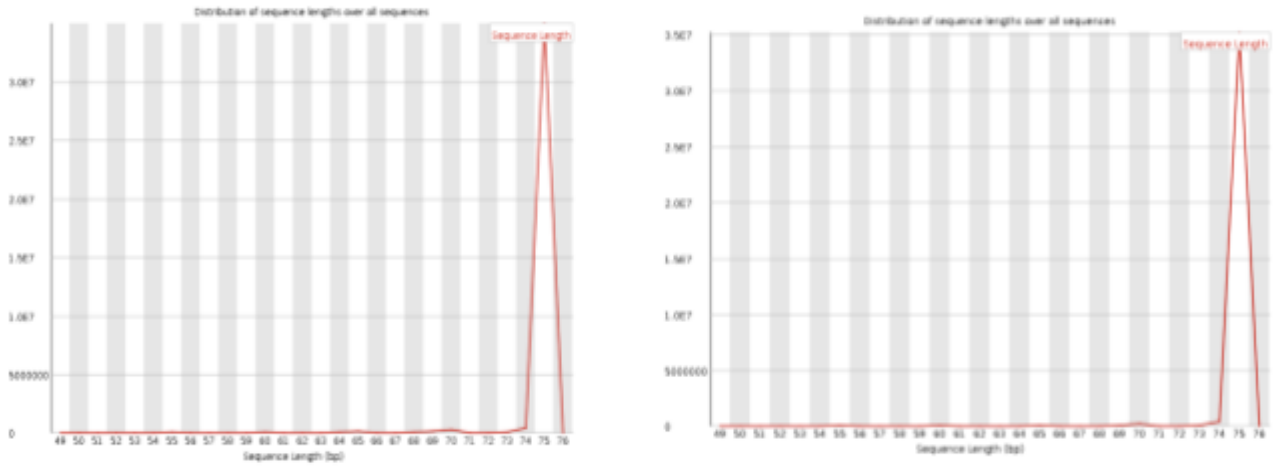
до триммирования *после триммирования*

Рис. 7. Сравнение Per base sequence quality до и после триммирования

c) Изменение длины чтений после триммирования:

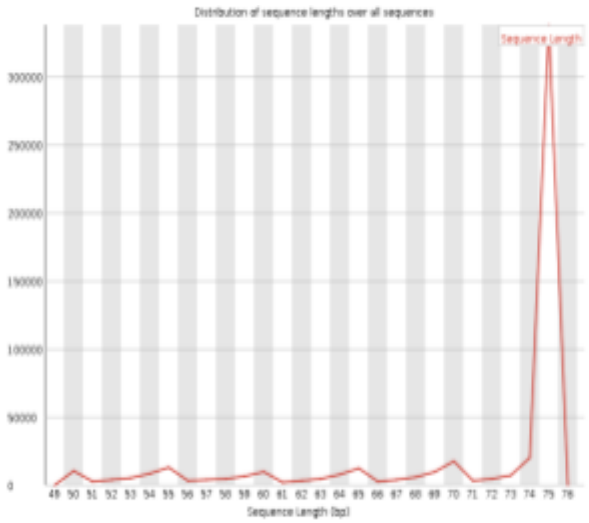
Большая часть так и осталась 75 нуклеотидов, но: в парных чтениях появилось небольшое количество длиной меньше (74, еще меньше длиной 70); а вот у непарных длина совсем изменилась, так как появились заметной величины пики на графиках

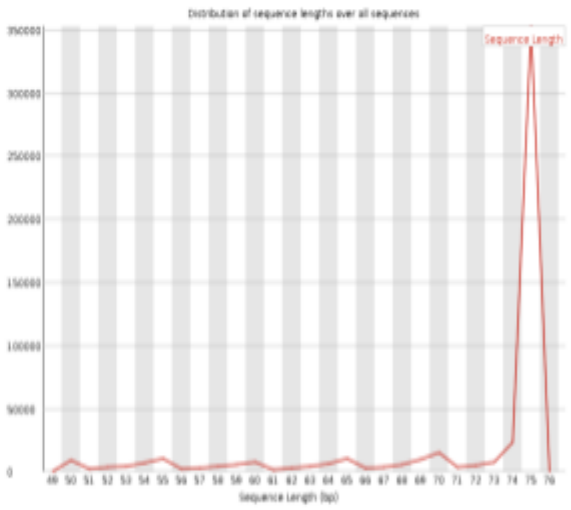
парные



прямые *обратные*

непарные





прямые *обратные*

Часть 2:

1. Картирование чтений на референсный геном

```
hisat2 -x ../chromosome/chr7 -1
../trimm/trim_1_paired.fastq.gz -2
../trimm/trim_2_paired.fastq.gz -p 10 --no-spliced-alignment >
map.sam 2> map_log.txt
```

Command: hisat2 – картирование чтений
Options: -x ../chromosome/chr7 – префикс имен файлов с индексацией референса (которые были получены после индексации референса в hisat2-build, их было 8)
-1 trim_1_paired.fastq.gz – файл с прямыми парными триммированными чтениями
-2 trim_2_paired.fastq.gz – файла с обратными парными триммированными чтениями
-p 10 – использую 10 ядер процессора, чтобы быстро посчиталось
--no-spliced-alignment - параметр, запрещающий возможность сплайсинга (то есть запрещает картирование с разрывами)
Input: ../chromosome/chr7 – файлы с индексацией референса

../trimm/trim_1_paired.fastq.gz, ../trimm/trim_2_paired.fastq.gz – парные триммированные чтения
Output: > map.sam – записываю вывод программы в файл .sam 2> map_log.txt – сохраняю логи в txt-файл

2. Конвертация sam в bam

1) Описание sam/bam файла

a) Вес sam файла – 15.34 ГБ
Конвертируем этот тяжеленный файл в его бинарный аналог – map.bam

```
samtools sort -o map.bam map.sam
```

Command: samtools sort – сортирует файл .sam, данный в input
Options: -o map.bam – вывод программы в файл map.bam
Input: map.sam – sam файл, полученный при картировании чтений
Output: map.bam – bam файл

b) Вес bam файла – 4.22 ГБ

Индексация bam файла с помощью samtools index:

```
samtools index map.bam
```

Command: samtools index – программа, которая индексирует файл
Input: map.bam – исходный файл
Output: map.bam.bai – выходной файл

Немного про файл sam:

Шапка:
@SQ SN – имя референса (7), LN – длина референса (159345973)
Еще в шапке можно найти информация о программе, которая сгенерировала этот sam файл (ID, имя и версия программы, копия полной команды для выравнивания)

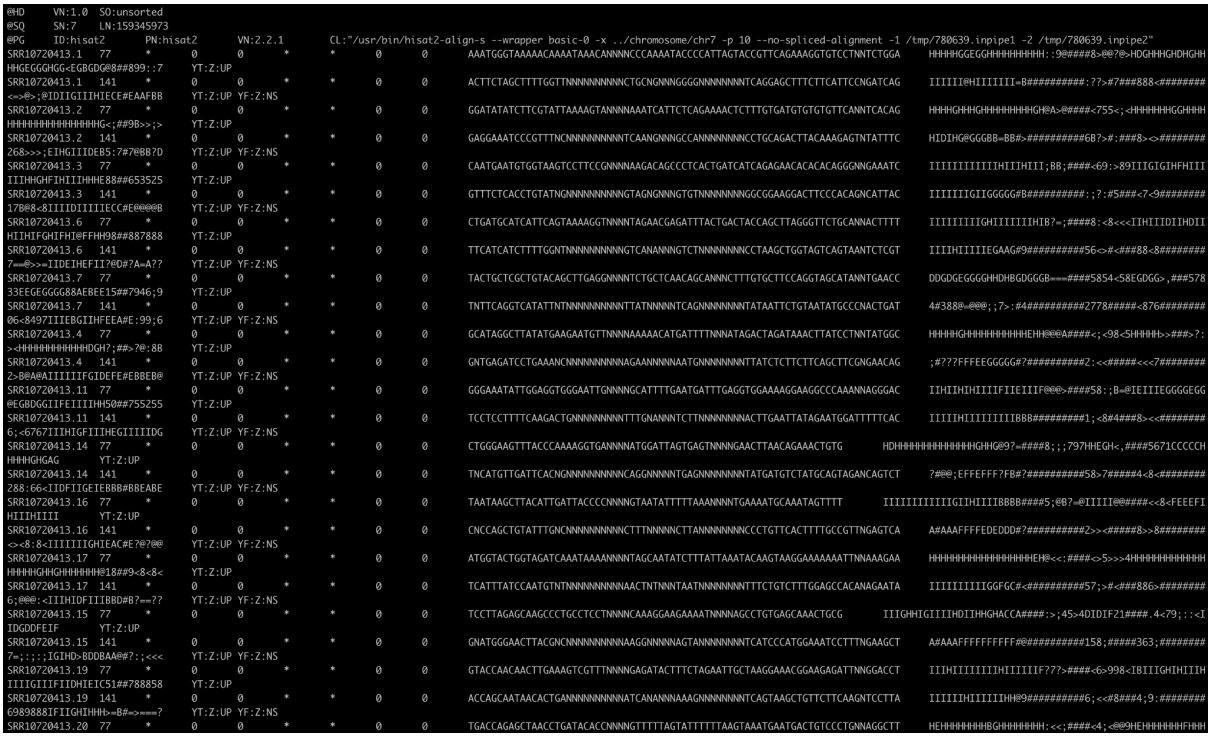


Рис. 1. Файл .sam

Содержание по столбцам:

QNAME – имя чтения

FLAG – информация о чтении (говорит нам как картировано чтение, каждой характеристике есть свой набор цифр. Нужна, когда мы переходим к файлу bam и фильтруем чтения)

RNAME – имя референсного контига (в моем случае chr7, так как она везде одинаковая, то стоит *)

POS – Позиция первого нуклеотида, куда выровнялось чтение на референс

MAPQ – Качество картирования (Q = -10lgP)

CIGAR – сколько мэтчей, количество делеций и вставок в референсе (в формате CIGAR: xMyDzI)

RNEXT – Название последовательности для парного (в случае PE) или следующего (в случае SE) чтения

PNEXT – POS для первого выравнивания парного (в случае PE) или следующего (в случае SE) чтения

TLEN – длина участка референса, к которому прикрепилось чтение (может не совпадать с длиной чтения, так как есть делеции и вставки в референсе относительно чтения)

SEQ – последовательность чтения (как в fastq)

QUAL – качество чтения для каждой позиции нуклеотида (как в fastq)
Дополнительная информация (опционально; здесь находится tag: видов тэгов много, каждому соответствует какое-то значение, например
MQ – качество картирования парного/следующего чтения,
AS – вес выравнивания, подсчитанный программой выравнивания и тд)

3. Анализ bam файла

Так как файл бинарный, просто посмотреть его (с помощью less, например) не получится. Поэтому запускаю программу для анализа этого файла:

```
samtools flagstat map.bam > analysed_bam.txt
```

Command: samtools flagstat – программа
Input: map.bam – исходный файл
Output: analysed_bam.txt – выходной файл

```
75096733 + 0 in total (QC-passed reads + QC-failed reads)
73901126 + 0 primary
1195607 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
5890778 + 0 mapped (7.84% : N/A)
4695171 + 0 primary mapped (6.35% : N/A)
73901126 + 0 paired in sequencing
36950563 + 0 read1
36950563 + 0 read2
4144524 + 0 properly paired (5.61% : N/A)
4226304 + 0 with itself and mate mapped
468867 + 0 singletons (0.63% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Все чтения прошли Quality Control (строка 1)
a) **Сколько чтений картировано на референс?** 5890778 штуки
b) **Сколько чтений картировано на референс в % от количества триммированных?** 7.84%
c) **Сколько чтений картировано на референс в корректных парах?** 4144524

d) **Сколько чтений картировано на референс в корректных парах в % от количества триммированных?** 5.61%

Так как у нас изначально чтения всего экзома, а референс у меня только 7ая хромосома, то очевидно, что не все чтения на нее картируются. Не все парные чтения корректно картировались, например, потому что парные чтения могли картироваться не по направлению друг к другу; чтение могло картироваться на референс несколько раз (и получится, что расстояние между парными чтениями будет большим); а может быть одно из чтений вообще не картировалось

4. Получение чтений, картированных на chr7

```
samtools view -h -bS map.bam 7 > 7.chr.bam
```

Command: samtools view – печатает все чтения из input картированные на референс
Options: -h – выводить в файл вместе с заголовком
-b – вывод в файл формата bam
-S – формат файла в input определить автоматически
Input: map.bam – исходный файл
7 – имя моей хромосомы
Output: 7.chr..bam – файл bam с чтениями, картированными на 7 хромосому

```
samtools flagstat 7.chr.bam > 7chranalysed_bam.txt
```

```
6359645 + 0 in total (QC-passed reads + QC-failed reads)
5164038 + 0 primary
1195607 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
5890778 + 0 mapped (92.63% : N/A)
4695171 + 0 primary mapped (90.92% : N/A)
5164038 + 0 paired in sequencing
2582019 + 0 read1
2582019 + 0 read2
4144524 + 0 properly paired (80.26% : N/A)
4226304 + 0 with itself and mate mapped
```

```
468867 + 0 singletons (9.08% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Рассмотрим ключевые изменения от аналогичного файла из п3:

Показатель	Весь BAM (все хромосомы)	BAM (только chr7)	Комментарий
Всего чтений	75 096 733	6 359 645	После фильтрации оставлены только чтения, относящиеся к chr7
Primary reads	73 901 126	5 164 038	Уменьшение связано с удалением чтений других хромосом
Secondary reads	1 195 607	1 195 607	Все secondary-выравнивания приходятся на chr7
Supplementary reads	0	0	Supplementary-выравнивания отсутствуют
Duplicates	0	0	Дедупликация не проводилась
Mapped reads	**5 890 778 (7.84%)**	**5 890 778 (92.63%)**	Абсолютное число одинаково; процент вырос из-за уменьшения общего числа чтений
Primary mapped reads	**4 695 171 (6.35%)**	**4 695 171 (90.92%)**	Почти все primary-выравнивания локализованы на chr7
Paired reads	73 901 126	5 164 038	Фильтрация по хромосоме
Properly paired reads	4 144 524 (5.61%)	4 144 524 (80.26%)	Высокая доля корректных пар после фильтрации
Read 1	36 950 563	2 582 019	Уменьшение пропорционально общему числу чтений

Read 2	36 950 563	2 582 019	Аналогично Read 1
Reads with mate mapped	4 226 304	4 226 304	Все такие пары соответствуют chr7
Singletons	468 867 (0.63%)	468 867 (9.08%)	Рост процента обусловлен уменьшением знаменателя
Mate mapped to different chr	0	0	Химерные пары не обнаружены

Таблица 1. Сравнение статистик выравнивания BAM-файла до и после фильтрации по хромосоме 7

Примечание. Фильтрация BAM-файла по chr7 не изменила абсолютное число выровненных чтений, но существенно увеличила относительные показатели выравнивания. Это указывает на то, что практически все выровненные чтения исходного BAM-файла локализованы на хромосоме 7.

5. Получение только правильно картированных пар чтений

```
samtools view -f 2 -bS 7.chr.bam > correct_7.bam
```

Options: -f 2 – оставить только те чтения, у которых в флаге выравнивания установлен бит 0x2 (то есть чтения, помеченные как properly paired)

```
5066138 + 0 in total (QC-passed reads + QC-failed reads)
4144524 + 0 primary
921614 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
5066138 + 0 mapped (100.00% : N/A)
4144524 + 0 primary mapped (100.00% : N/A)
4144524 + 0 paired in sequencing
2072262 + 0 read1
2072262 + 0 read2
4144524 + 0 properly paired (100.00% : N/A)
4144524 + 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)
```

Ключевые отличия от выдачи из пункта 4:

Показатель	BAM chr7 (все чтения)	BAM chr7 (только properly paired)	Комментарий
Всего чтений	6 359 645	5 066 138	Удалены одиночные, неправильно спаренные и не полностью выровненные чтения
Primary reads	5 164 038	4 144 524	Остались только primary properly paired чтения
Secondary reads	1 195 607	921 614	Secondary-выравнивания сохранены только для корректных пар
Supplementary reads	0	0	Supplementary-выравнивания отсутствуют
Duplicates	0	0	Дедупликация не проводилась
Mapped reads	**5 890 778 (92.63%)**	**5 066 138 (100.00%)**	После фильтрации все чтения выровнены
Primary mapped reads	**4 695 171 (90.92%)**	**4 144 524 (100.00%)**	Исключены primary чтения без корректной пары
Paired reads	5 164 038	4 144 524	Удалены некорректно спаренные пары
Properly paired reads	**4 144 524 (80.26%)**	**4 144 524 (100.00%)**	Отбор по SAM-флагу 0x2
Read 1	2 582 019	2 072 262	Пропорциональное уменьшение
Read 2	2 582 019	2 072 262	Пропорциональное уменьшение
Reads with mate mapped	4 226 304	4 144 524	Исключены некорректные пары
Singletons	468 867 (9.08%)	0 (0.00%)	Полностью исключены

Mate mapped to different chr	0	0	Межхромосомные пары отсутствуют

Таблица 2. Статистики BAM-файла chr7 до и после отбора только корректно выровненных пар чтений

Примечание. Фильтрация с использованием параметра `samtools view -f 2` приводит к формированию BAM-файла, содержащего исключительно корректно выровненные парные чтения на хромосоме 7. Это подтверждается 100% долей mapped и properly paired reads.

Теперь проиндексирую файл:

```
samtools index correct_7.bam
```

6. Получение чтений, картированных только в границы экзома

```
bedtools intersect -a correct_7.bam -b
/mnt/scratch/NGS/DATA/genes/seqcap_hg38.bed > exom_7_chr.bam
```

```
3328307 + 0 in total (QC-passed reads + QC-failed reads)
2699700 + 0 primary
628607 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3328307 + 0 mapped (100.00% : N/A)
2699700 + 0 primary mapped (100.00% : N/A)
2699700 + 0 paired in sequencing
1348451 + 0 read1
1351249 + 0 read2
2699700 + 0 properly paired (100.00% : N/A)
2699700 + 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)
```

Для выделения чтений, попадающих в экзомные регионы, был использован BED-файл с координатами экзома (seqcap_hg38.bed). Фильтрация выполнена с помощью bedtools intersect, применённого к BAM-файлу, содержащему только корректно выровненные парные чтения хромосомы 7 (correct_7.bam).

После пересечения с экзомными координатами общее число чтений составило 3 328 307, из которых 2 699 700 являются primary-выравниваниями. Все чтения в итоговом файле полностью выровнены на референсный геном, что подтверждается 100% долей mapped и properly paired reads. Одиночные чтения (singletons), некорректно спаренные пары, а также межхромосомные выравнивания отсутствуют.

Сохранение 628 607 secondary-выравниваний указывает на наличие альтернативных выравниваний для части чтений, что является ожидаемым для экзомных ‘частей’ с повторяющимися или гомологичными последовательностями. Дубликаты и supplementary-выравнивания не обнаружены.

Таким образом, итоговый BAM-файл содержит исключительно корректно выровненные парные чтения, локализованные в пределах экзомов хромосомы 7.

7. Получение чтений, картированных в границы расширенного экзома

```
bedtools intersect -a correct_7.bam -b
/mnt/scratch/NGS/DATA/genes/seqcap_hg38_50.bed >
exom_extended_7_chr.bam
```

```
3643449 + 0 in total (QC-passed reads + QC-failed reads)
2951919 + 0 primary
691530 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3643449 + 0 mapped (100.00% : N/A)
2951919 + 0 primary mapped (100.00% : N/A)

2951919 + 0 paired in sequencing
1475325 + 0 read1
1476594 + 0 read2
2951919 + 0 properly paired (100.00% : N/A)
```

```
2951919 + 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)
```

Показатель	Экзом (seqcap_hg38.bed)	Расширенный экзом (seqcap_hg38_50.bed)	Комментарий
Всего чтений	3 328 307	3 643 449	Увеличение числа чтений за счёт расширения экзомных границ
Primary reads	2 699 700	2 951 919	В расширенный экзом попали дополнительные primary-выравнивания
Secondary reads	628 607	691 530	Рост числа альтернативных выравниваний
Supplementary reads	0	0	Отсутствуют в обоих наборах
Duplicates	0	0	Дедупликация не проводилась
Mapped reads	**3 328 307 (100.00%)**	**3 643 449 (100.00%)**	Все отобранные чтения полностью выровнены
Primary mapped reads	**2 699 700 (100.00%)**	**2 951 919 (100.00%)**	100% primary выравниваний в обоих случаях
Paired reads	2 699 700	2 951 919	Все чтения являются корректно спаренными

Read 1	1 348 451	1 475 325	Пропорциональное увеличение
Read 2	1 351 249	1 476 594	Пропорциональное увеличение
Properly paired reads	**2 699 700 (100.00%)**	**2 951 919 (100.00%)**	Корректные пары сохранены полностью
Reads with mate mapped	2 699 700	2 951 919	Оба чтения пары выровнены
Singletons	0 (0.00%)	0 (0.00%)	Одиночные чтения отсутствуют
Mate mapped to different chr	0	0	Межхромосомные выравнивания отсутствуют

Таблица 3. Сравнение статистик чтений chr7, картированных в границы экзома и расширенного экзома

Примечание. Расширенный экзом (seqcap_hg38_50.bed) включает экзонные регионы с дополнительными фланкирующими участками (±50 п.н.), что приводит к увеличению числа отобранных чтений по сравнению со стандартным экзомом (seqcap_hg38.bed). При этом качество выравнивания остаётся неизменным: все чтения корректно спарены и полностью выровнены на референсный геном.

Практикум 12

1. Получение вариантов

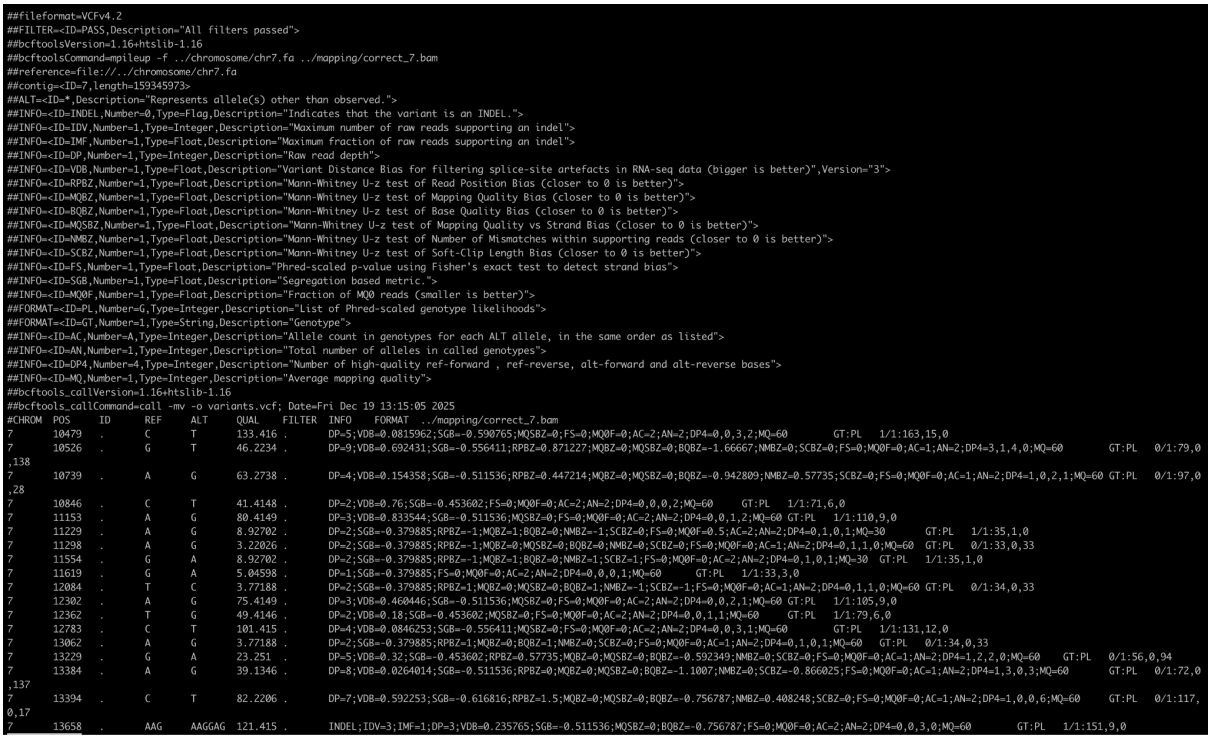
Варианты на хромосоме 7 были вызваны с использованием bcftools mpileup и bcftools call. Использован BAM-файл с корректно выровненными парными чтениями (correct_7.bam) и референс chr7.fa. Все позиции считались диплоидными, максимальная глубина на позицию ограничена 250 чтениями, вызваны только SNP и Indel

*Note: none of --samples-file, --ploidy or --ploidy-file given, assuming all sites are diploid
[mpileup] 1 samples in 1 input files
[mpileup] maximum number of reads per input file set to -d 250

```
bcftools mpileup -f ../chromosome/chr7.fa
../mapping/correct_7.bam | bcftools call -mv -o variants.vcf
```

Command: bcftools mpileup – генерирует vcf файл, в котором находятся вероятности разных вариантов (на основании выравнивания)
Options: -f ../chromosome/chr7.fa – указывают референс
Input: ../chromosome/chr7.fa – референс
../mapping/correct_7.bam – файл bam с картированными ридами
Command: bcftools call – из output (stdout) программы bcftools mpileup берет только нужные строки (характеристики указаны в options)
Options: -m – модель, которая ищет мультиаллельные и редкие варианты
-v – на выдачу попадут только варианты
-o variants.vcf – выдача в файл variants.vcf
Input: output из bcftools mpileup
Output: variants.vcf

Теперь нужно посмотреть на сам vcf файл



```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.16+htslib-1.16
##bcftoolsCommand=mpileup -f ../chromosome/chr7.fa ../mapping/correct_7.bam
##reference=file://../chromosome/chr7.fa
##contig=<ID=7,length=159345973>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of raw reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of raw reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
##INFO=<ID=RPBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Read Position Bias (closer to 0 is better)">
##INFO=<ID=MQBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Mapping Quality Bias (closer to 0 is better)">
##INFO=<ID=BQBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Base Quality Bias (closer to 0 is better)">
##INFO=<ID=MQSBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Mapping Quality vs Strand Bias (closer to 0 is better)">
##INFO=<ID=NMBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Number of Mismatches within supporting reads (closer to 0 is better)">
##INFO=<ID=SCBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Soft-Clip Length Bias (closer to 0 is better)">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.16+htslib-1.16
##bcftools_callCommand=call -mv -o variants.vcf; Date=Fri Dec 19 13:15:05 2025
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  ../mapping/correct_7.bam
7       10479   .       C       T       133.416 .       DP=5;VDB=0.0815962;SGB=-0.590765;MQSBZ=0;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,3,2;MQ=60        GT:PL   1/1:163,15,0
7       10526   .       G       T       46.2234 .       DP=9;VDB=0.692431;SGB=-0.556411;RPBZ=0.871227;MQBZ=0;MQSBZ=0;BQBZ=-1.66667;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=3,1,4,0;MQ=60        GT:PL   0/1:79,0
,138
7       10739   .       A       G       63.2738 .       DP=4;VDB=0.154358;SGB=-0.511536;RPBZ=0.447214;MQBZ=0;MQSBZ=0;BQBZ=-0.942809;NMBZ=0.57735;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,0,2,1;MQ=60 GT:PL   0/1:97,0
,28
7       10846   .       C       T       41.4148 .       DP=2;VDB=0.76;SGB=-0.453602;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,0,2;MQ=60     GT:PL   1/1:71,6,0
7       11153   .       A       G       80.4149 .       DP=3;VDB=0.833544;SGB=-0.511536;MQSBZ=0;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,1,2;MQ=60 GT:PL   1/1:110,9,0
7       11229   .       A       G       8.92702 .       DP=2;SGB=-0.379885;RPBZ=-1;MQBZ=1;BQBZ=0;NMBZ=-1;SCBZ=0;FS=0;MQ0F=0.5;AC=2;AN=2;DP4=0,1,0,1;MQ=30       GT:PL   1/1:35,1,0
7       11298   .       A       G       3.22026 .       DP=2;SGB=-0.379885;RPBZ=-1;MQBZ=0;MQSBZ=0;BQBZ=0;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=0,1,1,0;MQ=60  GT:PL   0/1:33,0,33
7       11554   .       G       A       8.92702 .       DP=2;SGB=-0.379885;RPBZ=-1;MQBZ=1;BQBZ=0;NMBZ=1;SCBZ=1;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,1,0,1;MQ=30  GT:PL   1/1:35,1,0
7       11619   .       G       A       5.04598 .       DP=1;SGB=-0.379885;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60      GT:PL   1/1:33,3,0
7       12084   .       T       C       3.77188 .       DP=2;SGB=-0.379885;RPBZ=1;MQBZ=0;MQSBZ=0;BQBZ=1;NMBZ=-1;SCBZ=-1;FS=0;MQ0F=0;AC=1;AN=2;DP4=0,1,1,0;MQ=60 GT:PL   0/1:34,0,33
7       12302   .       A       G       75.4149 .       DP=3;VDB=0.460446;SGB=-0.511536;MQSBZ=0;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,2,1;MQ=60 GT:PL   1/1:105,9,0
7       12362   .       T       G       49.4146 .       DP=2;VDB=0.18;SGB=-0.453602;MQSBZ=0;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,1,1;MQ=60     GT:PL   1/1:79,6,0
7       12783   .       C       T       101.415 .       DP=4;VDB=0.0846253;SGB=-0.556411;MQSBZ=0;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,3,1;MQ=60        GT:PL   1/1:131,12,0
7       13062   .       A       G       3.77188 .       DP=2;SGB=-0.379885;RPBZ=1;MQBZ=0;BQBZ=1;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=0,1,0,1;MQ=60   GT:PL   0/1:34,0,33
7       13229   .       G       A       23.251  .       DP=5;VDB=0.32;SGB=-0.453602;RPBZ=0.57735;MQBZ=0;MQSBZ=0;BQBZ=-0.592349;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,2,2,0;MQ=60    GT:PL   0/1:56,0,94
7       13384   .       A       G       39.1346 .       DP=8;VDB=0.0264014;SGB=-0.511536;RPBZ=0;MQBZ=0;MQSBZ=0;BQBZ=-1.1007;NMBZ=0;SCBZ=-0.866025;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,3,0,3;MQ=60       GT:PL   0/1:72,0
,137
7       13394   .       C       T       82.2206 .       DP=7;VDB=0.592253;SGB=-0.616816;RPBZ=1.5;MQBZ=0;MQSBZ=0;BQBZ=-0.756787;NMBZ=0.408248;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,0,0,6;MQ=60     GT:PL   0/1:117,
0,17
7       13658   .       AAG     AAGGAG  121.415 .       INDEL;IDV=3;IMF=1;DP=3;VDB=0.235765;SGB=-0.511536;MQSBZ=0;BQBZ=-0.756787;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,3,0;MQ=60       GT:PL   1/1:151,9,0
```

Рис. 1. Файл .vcf

Что можно сказать про файл .vcf?
Сначала идет шапка файла, каждая строка начинается с ##. Затем перед «телом файла» идет строка с заголовками столбцов, она начинается с #.

А теперь поподробнее о каждом столбце: что можно найти?
CHROM – имя хромосомы
POS – позиция варианта
ID – везде стоят «.» , но может быть любая информация о варианте
ALT – альтернативный аллель (так как у нас SNP, то здесь стоит одна буква)
QUAL – качество варианта
FILTER – везде «.» , так как ввели файл, который уже был маркирован по качеству
INFO – характеристики варианта
FORMAT – список параметров варианта (для конкретного образца)

А теперь проанализируем variants.vcf

```
bcftools stats variants.vcf > var_stats.txt
```

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      1
SN      0       number of records:      74220
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 73033
SN      0       number of MNPs: 0
SN      0       number of indels:       1187
SN      0       number of others:       0
SN      0       number of multiallelic sites:   77
SN      0       number of multiallelic SNP sites:       44
```

a) **Сколько получилось вариантов?** 74220
b) **Сколько из них являются SNP?** 73033
c) **Сколько получилось коротких вставок и делеций?** 1187 (индели)
d) **Смотрим на выход первой части команды** (bcftools mpileup)

```
bcftools mpileup -f ../chromosome/chr7.fa
../mapping/correct_7.bam > whole_variants.vcf
```

Программа выдаст информацию по всем позициям последовательности хромосомы. Тут для большинства позиций в поле ALT стоит <*>, так как нет альтернативных позиций.

2. Фильтрация вариантов

bcftools filter -i'QUAL>30 && DP>50' variants.vcf \ -o filtred_variants.vcf

Command: bcftools filter – программа, которая отфильтрует варианты из input файла по заданным параметрам
Options: -i'%QUAL>30 && DP>50' – фильтруем по параметрам: качество больше 30 и длина больше 50
-o filt_variants.vcf – вывод программы в указанный файл
Input: variants.vcf
Output: filt_variants.vcf

анализируем полученный файл

```
bcftools stats filtred_variants.vcf > filt_var_stats.txt
```

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      1
SN      0       number of records:      2340
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 2237
SN      0       number of MNPs: 0
SN      0       number of indels:       103
SN      0       number of others:       0
SN      0       number of multiallelic sites:   8
SN      0       number of multiallelic SNP sites:       2
```

a) Сколько осталось вариантов после фильтрации? 2340 штук, 3.15%
b) Сколько осталось SNP? 2237 штук, 3,06%
c) Сколько осталось коротких вставок и делеций? 103 штуки, 8,7%

3. Аннотация вариантов

Category	Count
Variants processed	2340
Variants filtered out	0
Novel / existing variants	431 (18.4) / 1909 (81.6)
Overlapped genes	986
Overlapped transcripts	10191
Overlapped regulatory features	34

Таблица 1

Всего в файле было 2340 вариантов
Из них отфильтровано 0, то есть все было проаннтированы
Новые варианты – 431 (сведения о них нет ни в каком формате в датабазах, например, в clinvar), а существующих и уже аннотированных где-то вариантов – 1909
Перекрывающихся генов – 986
Перекрывающихся транскриптов – 10191
Перекрывающихся регуляторных областей- 34

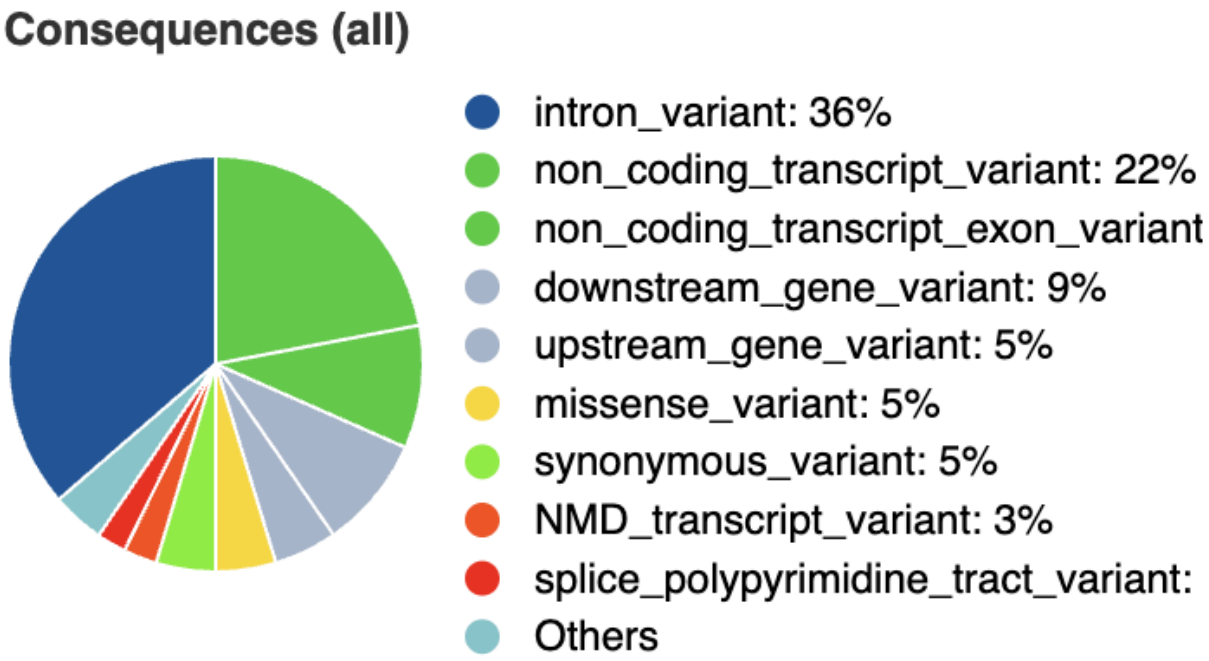


Рис. 2. Распределение эффектов (consequences) мутаций

Дадим обзорную характеристику:

Все мутации, кроме тех, которые перечислены ниже имеют влияние на транскрипт **«MODIFIER»**, то есть либо не влияют вообще, либо затрагивают некодирующие гены/области. Это и логично, ведь если бы было много мутаций, которые влияют на транскрипты,

5% мутаций относятся к synonymous variant, то есть синонимичным заменам, и имеют влияние **«LOW»**, то есть в большинстве своем не меняют функцию/поведение белка (что и понятно по определению синонимичной мутации, хотя есть случаи, когда они влияют на структуру: например, если замена синонимичная, но при этом заменить на более редкий кодон, тогда скорость трансляции уменьшается, нарушается фолдинг (который идет во время трансляции), вследствие чего структура и функции белка могут нарушиться

5% мутаций относятся к missense_mutations и имеют влияние «MODERATE». Белок транслируется, но неправильно, вследствие чего меняется его структура, а дальше либо теряется функция, либо меняется

А теперь отдельно посмотрим на кодирующие области, ведь интересующую нас патогенность или изменения вызывают, в большинстве своем, именно они

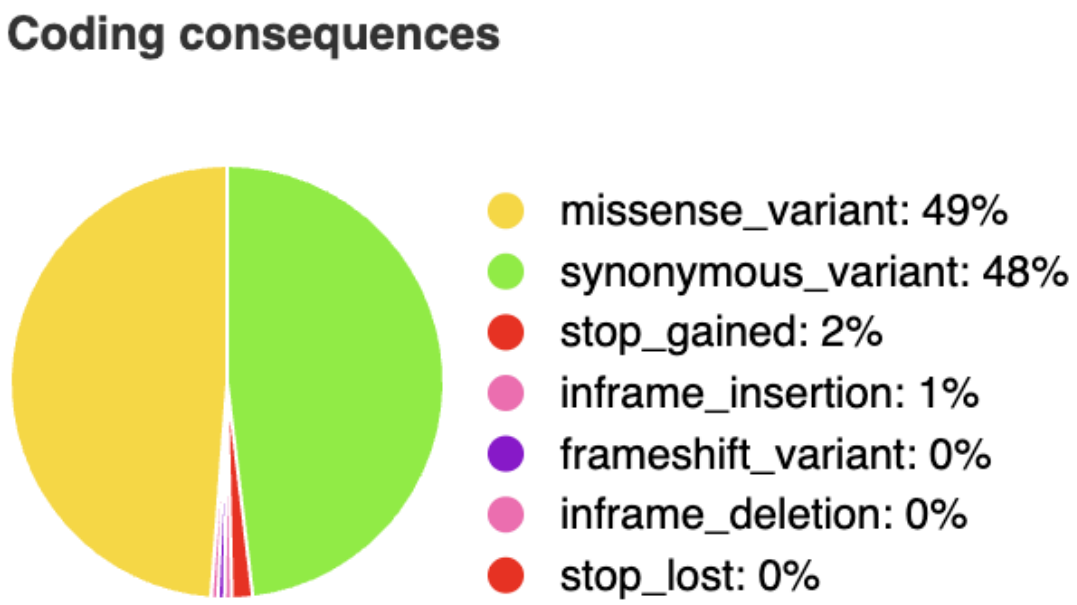


Рис. 8. Распределение эффектов (consequences) мутаций кодирующих областей

Очевидно, что большинство из них будут повторять мутации кодирующих областей из предыдущего графика.

Опасно выглядит 2% stop gained, то есть преждевременный стоп-кодон. Это приводит к укороченным транскриптам, которые могут терять/частично терять функцию. Такие мутации очень могут быть патогенными. Остальные варианты встречаются 0% (но это, видимо, просто округление?) Иначе зачем вообще вносить их в диаграмму), но в целом они с уровнями влияния на транскрипт «MODERATE» и «HIGH».

Отдельно поговорим про варианты с **HIGH IMPACT**, потому что именно они вызовут вопросы, если мы будем искать патогенные мутации. Выпишу типы мутаций из выдачи, которые характеризуются HIGH влиянием.

Подавляющее большинство мутаций с HIGH IMPACT ожидается получить в экзонах (хотя, при альтернативном сплайсинге интрон одного транскрипта может быть экзоном другого... и много разного можно придумать, но логично ожидать в экзонах!). И точно все они должны попасть в ген.

После фильтрации получилось, что **138 вариантов** с IMPACT = HIGH

Дадим характеристику вариантам:

a) Относительно генов (попали в ген или нет)

Практически все варианты попадают в гены - В колонке Gene указан Ensembl-идентификатор (ENSG...) для всех записей. Исключение составляют две строки с Gene = ENSG00000299721 и SYMBOL = - — это lncRNA (long non-coding RNA), т.е. всё равно функциональный некодирующий ген. Имеется один вариант Gene = ENSG00000273024, SYMBOL = INTS4P2 — транс-некодирующий псевдоген.

Вывод: все варианты находятся внутри аннотированных генов (кодирующих или некодирующих)

b) **Относительно структуры генов (экзон/интрон)**

Экзоны: большинство имеет непустое поле EXON (например, 8/29, 2/2, 4/4 и т.д.).
Интроны: варианты со значением в колонке INTRON отсутствуют
Сплайс-сайты: есть варианты с Consequence = splice_donor_variant или splice_acceptor_variant, которые расположены на границах экзон-интрон

Вывод: все варианты затрагивают экзоны либо сплайс-сайты, ни один не находится внутри интрона

c) **Относительно приносимых изменений (типы мутаций)**

Тип мутации	Краткое описание	Примеры в файле
stop_gained	Нонсенс-мутация (ранний стоп-кодон) → ‘усеченный’ белок	VWDE, ARL4A, NFE2L3, KMT2C, ZNF117, DNAJB6
frameshift_variant	Сдвиг рамки считывания → полная потеря функции	ARL4A, NFE2L3, ZNF727, KMT2C
stop_lost	Потеря стоп-кодона → ‘удлинённый’ белок	SEPTIN14, KMT2C

splice_donor_variant	Нарушение донорного сайта сплайсинга	KMT2C
splice_acceptor_variant	Нарушение акцепторного сайта сплайсинга	KMT2C, SEPTIN14, lncRNA
non_coding_transcript_variant	Затрагивает некодирующий транскрипт	lncRNA, псевдогены

Таблица 1.

Вывод: все варианты с IMPACT = HIGH — это мутации, серьёзно нарушающие структуру или экспрессию гена: нокдаун (stop_gained, frameshift), нарушение процессинга мРНК (splice_donor/acceptor), удлинение белка (stop_lost). Нет ни одного missense, synonymous или regulatory варианта — такие получают IMPACT = MODERATE или LOW

Практикум 13

1. Описание образца

a) ID образца РНК-чтений: **ENCFF641WPY**
b) Ссылка на информацию об образце: https://www.encodeproject.org/report/?type=Experiment&searchTerm=ENCFF641WPY
c) Организм и ткань: Mus musculus strain B6CASTF1/J heart tissue male adult (18-20 months)
d) Стратегия секвенирования: total RNA-seq
e) Тип чтений: SE (одноконцевые): SE single-ended
f) Цепь-специфичность: нет

2. Проверка качества исходных чтений

```
fastqc ENCFF641WPY.fastq.gz
```

1) **Количество чтений:** 48,692,928
2) **Качество чтений:**

Среднее качество (синяя линия) стабильно находится в зелёной зоне (>30) на протяжении всей длины прочтения.
Минимальное качество (нижняя граница ящика) также остаётся выше 30, что указывает на отсутствие значительного падения качества
В начале (позиции 1–5) среднее качество чуть ниже 35, но всё ещё в «хорошем» диапазоне, почему там нет ящичков я не знаю простите
На конце (позиция 97) — небольшое снижение, но мне кажется не критичное.
Чтения высокого качества, пригодны для последующего анализа без необходимости триммирования. Падение качества в конце и в начале минимально.

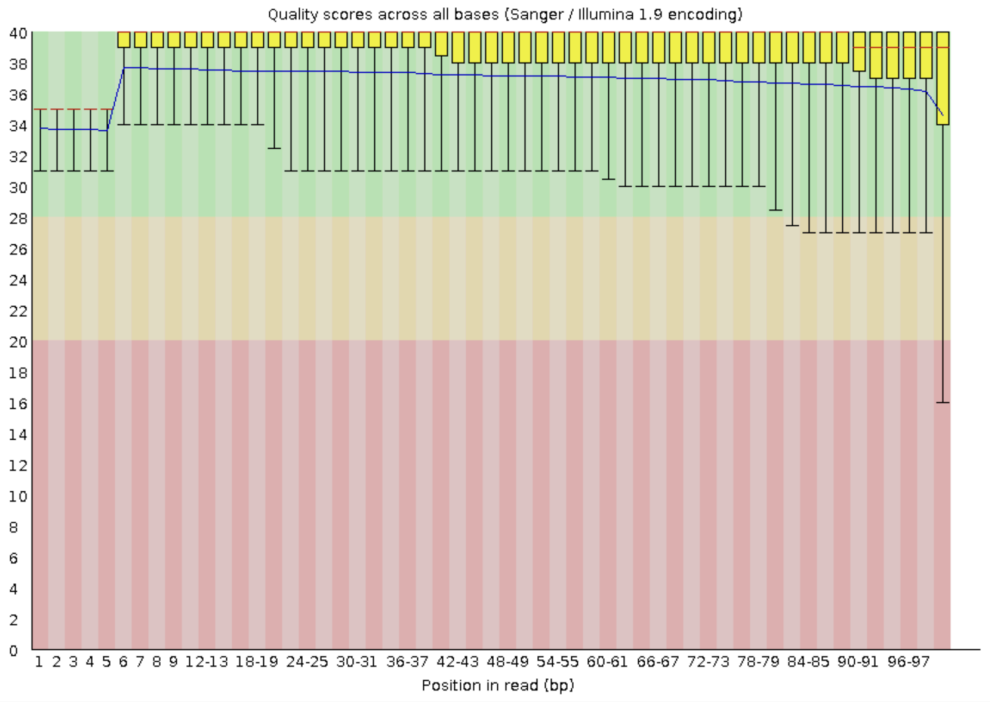


Рис. 1. Per base sequence quality

3) Длина чтений:

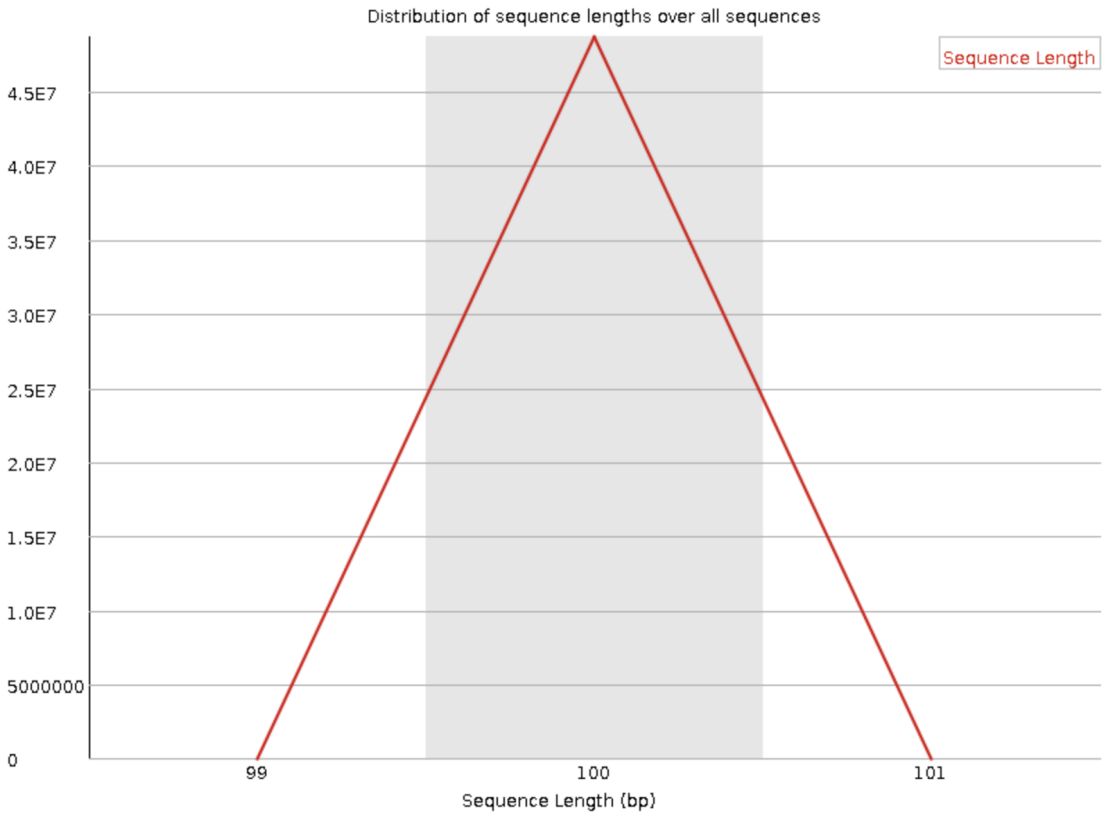


Рис. 2. Sequence Length Distribution

Длина чтений в этом образце РНК-секвенирования является строго фиксированной и составляет 100 пар оснований (bp). Количество чтений с длиной 99 bp и 101 bp пренебрежимо мало, а все остальные длины отсутствуют. Такая идеально узкая и высокая пиковая форма указывает на то, что секвенирование было выполнено с использованием строго контролируемой длины чтения, что типично для стандартных протоколов Illumina, где задается фиксированная длина цикла (например, 100 циклов).

4) Другие графики

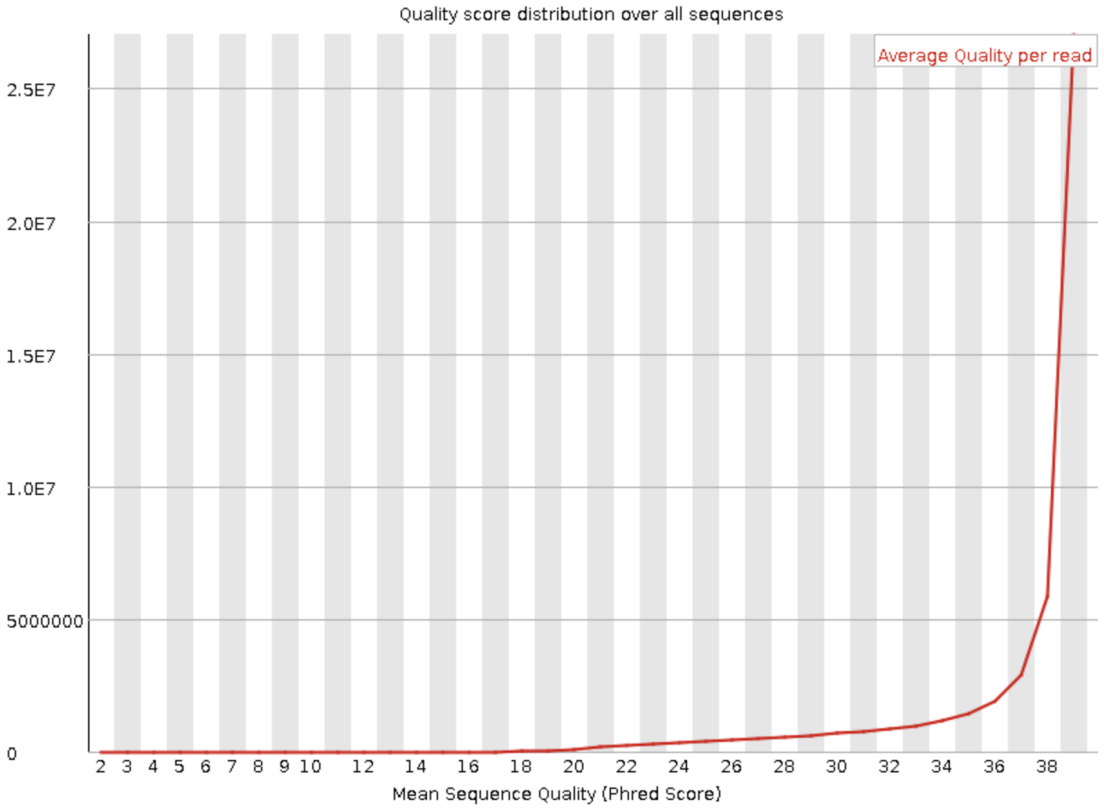


Рис. 3. Per sequence quality scores

Большинство чтений имеют средний балл качества (Phred score) выше 30. Кривая резко возрастает в правой части графика, показывая, что подавляющее большинство чтений обладают отличным качеством (в основном 35–38). Очень мало чтений имеют низкое качество (ниже 20), что свидетельствует о хорошем качестве секвенирования и отсутствии значительных технических проблем.

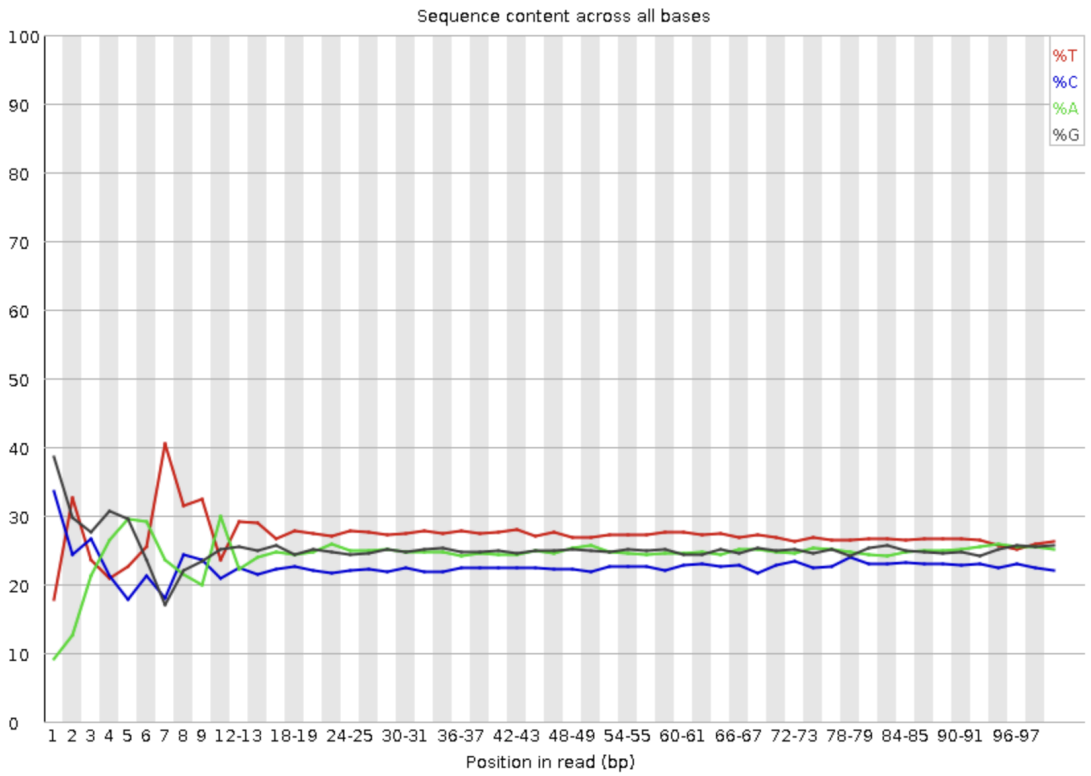


Рис. 4. Per base sequence content

В начале чтений (первые 10–15 позиций) наблюдается сильное отклонение от ожидаемого равномерного распределения нуклеотидов, что указывает на наличие артефактов — вероятно, остатков адаптеров или биологических особенностей (например, обогащение определёнными последовательностями в 5'-конце транскриптов).

Начиная примерно с 20-й позиции и до конца чтения (100 bp), содержание нуклеотидов стабилизируется и колеблется вокруг 25% для каждого из A, T, C, G — что соответствует ожидаемому поведению для случайной последовательности. Это говорит о том, что основная часть данных качественная и не содержит систематического смещения.

Такое поведение типично для RNA-seq данных, особенно при использовании стратегии total RNA-seq без отбора по поли-A хвосту: в начале могут присутствовать рибосомные РНК, некодирующие РНК или фрагменты с неравномерным составом.

3. Картирование чтений на референс

```
hisat2 -x ../chromosome/chr7 -k 3 -U
../rna/ENCFF641WPY.fastq.gz > rna_map.sam 2> rna_map_log.txt
```

Command: hisat2 – картирование чтений
Options: -x ../chromosome/chr7 – префикс имен файлов с индексацией референса (которые были получены после индексации референса в hisat2-build, их было 8)
-k 3 - максимально возможное количество выравниваний (3), чей score больше или равен score любого другого выравнивания (но hisat2 не даёт гарантий, что это лучшие выравнивания)
-U ../rna/ENCFF641WPY.fastq.gz – файл с чтениями
Input: ../chromosome/chr7 – файлы с индексацией референса
../rna/ENCFF641WPY.fastq.gz – РНК-чтения
Output: rna_map.sam – записываю вывод программы в файл .sam
rna_map_log.txt – сохраняю логи в txt-файл

Заглянем в rna_map_log.txt

```
48692928 reads; of these:
  48692928 (100.00%) were unpaired; of these:
    48667000 (99.95%) aligned 0 times
    23915 (0.05%) aligned exactly 1 time
    2013 (0.00%) aligned >1 times
0.05% overall alignment rate
```

Сколько чтений картировалось на хромосому?
выровнялись ровно 1 раз: 23915 чтений
выровнялись более 1 раза: 2013 чтений
Всего картировалось (на хромосому/хромосомы):
23915 + 2 013 = **25928** чтений
Это соответствует указанным ~0,05% от 48,692,928 чтений.
Мне казалось, что будет больше (сужу по проценту), учитывая, что качество неплохое, однако у нас весь транскриптом и только 1 хромосома

Перевод в bam файл:

```
samtools sort -o rna_map.bam rna_map.sam
```

Индексация bam файла:

```
samtools index rna_map.bam
```

Отбор только тех чтений, которые картировались на хромосому:

```
samtools view -h -bS rna_map.bam 7 > chr7_rna_map.bam
```

4. Поиск экспрессирующихся генов

Скачала файл с геномной разметкой Homo_sapiens.GRCh38.110.chr.gtf

Сначала идет шапка:
#!genome-build GRCh38.p14 - версия разметки
#!genome-version GRCh38 - версия генома, на которой строили
#!genome-date 2013-12 - дата публикации
#!genome-build-accession GCA_000001405.29- AC разметки
#!genebuild-last-updated 2023-03 - последняя дата обновления

После шапки идет тело файла. В каждой строке содержится информация о разметке, разделенная на 9 столбцов:
seqname - название последовательности, где аннотирован ген (в нашем случае имя хромосомы)
source - источник аннотации (просмотрела файл, чаще всего встречается ensembl и Havana)
feature - особенности гена
start - начало гена
end - конец гена
score – какое-то значение, вместо которого во всем файле стоит «.» (в мануале A floating point value, не особо понимаю, что это конкретно значит, но вероятно что-то о достоверности/качестве)
strand - цепь, на которой этот ген находится
frame – (сдвиг) рамка считывания
attribute - дополнительная информация

Сколько аннотированных генов на 7 хромосоме?

```
grep '^7' *gtf  | cut -f3 | grep  'gene' |  wc -l
```

Получаем ответ: **3147**
Почему-то по разным источникам их 1400-1800, а у меня получилось в 2 раза больше. Вроде по файлу нашла все логично: беру строки с 7 хромосомой, в третьем столбце (feature) считаю только те, которые описаны как ‘gene’...)

Теперь посчитаем для каждого гена из разметки число картированных на этот ген чтений:

```
htseq-count -f bam -s no -t exon -m union chr7_rna_map.bam
 ../Homo_sapiens.GRCh38.110.chr.gtf -o count_exons.sam 1>
            count_log1.txt 2> count_log2.txt
```

Command: htseq-count – картирование чтений
Options: -f bam – входной файл в формате bam
-s no – у чтений не указана цепь, поэтому они могут попадать и на прямую, и на обратную цепи
-m union – если чтения перекрываются, то их нужно объединить
-t exon – особенность гена (из 3 столбца), считаются только чтения, которые картировались на гены с этой особенностью. Так как по дефолту имеет значение exon, его и поставлю
-o count_exons.sam – выходной файл
Input: chr7_rna_map.bam – чтения, картированные на 7 хромосому
../Homo_sapiens.GRCh38.110.chr.gtf – генетическая разметка
Output: count_exons.sam – sam файл с аннотированными выравниваниями
count_log1.txt (из stdout) – сводка о работе программы
count_log2.txt (из stderr) – файл с ошибками

В файле count_log1.txt находится информация про гены:
__no_feature 2507
__ambiguous 819
__too_low_aQual 0
__not_aligned 0
__alignment_not_unique 2013

Сколько чтений попало в границы генов с (-t exon)?

```
awk '$2>0 && $1 !~ /^__/' count_log1.txt | awk '{s+=$2} END
                        {print s}'
```

20589 чтений

Сколько чтений попало мимо границ экзонов (-t exon)?

см. no feature - **2 507** чтений

Объяснение всех строк аннотационного файла:

__no_feature – чтения, которые попали вне экзонов

__ambigous – 819 чтений ассоциированы с более чем одной особенностью гена (-t, feature)

__too_low_aQual – чтения, которые были бы пропущены, если бы программе дали опцию -a (пропустить чтения с весом выравнивания ниже заданного)

__not_aligned – чтения в файле без выравнивания (таких нет, так на вход получен файл с чтениями, про которые точно известно картирование на 7 хромосому)

__alignment_not_unique – количество чтений, у которых больше одного выравнивания

5. Аннотация высоко экспрессируемых генов

сортируем по убыванию количества каунтов:

```
grep -v '^__' count_log1.txt | sort -k2,2nr >
            count_sorted_desc.txt
```

a) `head count_sorted_desc.txt`:

ENSG00000002079	3525
ENSG00000122566	2167
ENSG00000105953	2024
ENSG00000233476	1217
ENSG00000231167	843
ENSG00000122545	658
ENSG00000105835	552
ENSG00000128595	522
ENSG00000136261	415
ENSG00000168090	365

b) Для визуализации в геномном браузере выберем **один из самых** высоко экспрессируемых белок-кодирующих генов.
Из топ-10 самый явный кандидат — **ENSG00000002079**, который соответствует гену **MYH16** (myosin heavy chain 16)

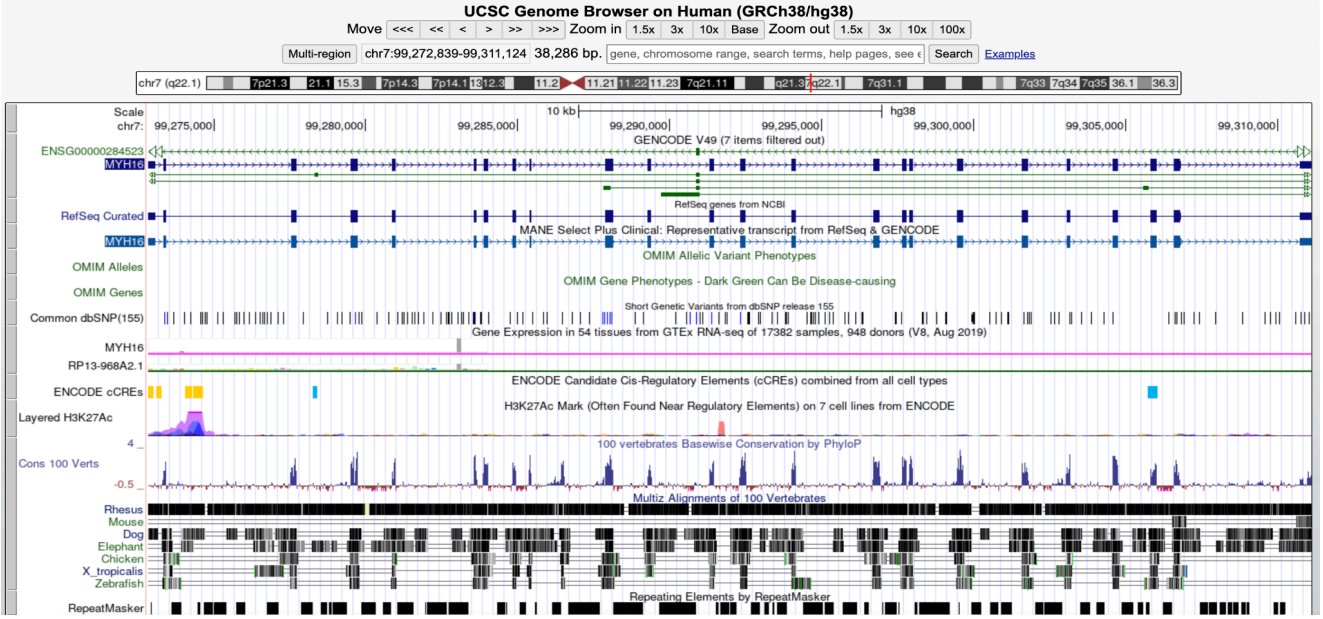


Рисунок 5. Экзон-интронная структура известных транскриптов

Визуализация получена с использованием UCSC Genome Browser (сборка GRCh38/hg38) на основе аннотаций GENCODE/Ensembl. Прямоугольниками показаны экзоны, линиями — интроны; каждая строка соответствует отдельному транскрипту гена.

c) MYH16 (myosin heavy chain 16) — ген, относящийся к семейству миозиновых тяжёлых цепей. В большинстве млекопитающих он кодирует белок миозиновой тяжёлой цепи, участвующий в сокращении мышц, особенно жевательных мышц (temporalis и masseter). Мышечные миозины — это моторные белки, которые связываются с актином и обладают активностью АТР-азы, генерируя силу и движение в мышечных волокнах. Однако в человеке MYH16 считается псевдогеном или сильно ослабленным геном, поскольку у нас есть мутация, приводящая к неработоспособному белку; это объясняет уменьшенные размеры жевательных мышц по сравнению с другими приматами. В не-человеческих приматах функциональный MYH16 обеспечивает большую жевательную силу
*Информация из GeneCards

СКРИПТ ДЛЯ ПРАКТИКУМОВ 11-12

В папке: /mnt/scratch/NGS/bratzveron/script
Запуск: `cd /mnt/scratch/NGS/bratzveron/script`
`./script.sh SRR10720421 7`

```
#!/bin/bash
#script.sh
#usage: ./script.sh sample_id chr_number
#пример: ./script.sh SRR10720421 7
#автоматизированный пайплайн:
qc->триммирование->картирование->фильтрация bam->вызов и
фильтрация vcf
#референс grch38.p14, ensembl release 110

#настройки по умолчанию
reads_dir=/mnt/scratch/NGS/DATA/dna_reads
ref_dir=/mnt/scratch/NGS/DATA/hg38
genes_gtf=/mnt/scratch/NGS/DATA/genes/Homo_sapiens.GRCh38.110.
chr.gtf
bed_file=/mnt/scratch/NGS/DATA/genes/seqcap_hg38.bed
bed_extended=/mnt/scratch/NGS/DATA/genes/seqcap_hg38_50.bed

phred=phred33
trim_trailing=20
trim_minlen=50
nthreads=10

filt_qual=30
filt_depth=50

output_dir=/mnt/scratch/NGS/bratzveron/results

#входные аргументы
ID=$1
CHR=$2

#проверка наличия аргументов
if [[ -z "$ID" || -z "$CHR" ]]; then
    echo "usage: $0 sample_id chr_number"
    exit 1
fi
```

```
echo "запуск пайплайна для образца $ID, хромосома $CHR"

#проверка наличия входных файлов
forward_reads=$reads_dir/${ID}_1.fastq.gz
reverse_reads=$reads_dir/${ID}_2.fastq.gz
ref=$ref_dir/Homo_sapiens.GRCh38.dna.chromosome.${CHR}.fa

for file in "$forward_reads" "$reverse_reads" "$ref"; do
    if [[ ! -f "$file" ]]; then
        echo "ошибка: файл не найден: $file"
        exit 1
    fi
done

#создание выходных директорий
mkdir -p $output_dir/mapping
mkdir -p $output_dir/trimmed
mkdir -p $output_dir/vcf
mkdir -p $output_dir/fastqc

#шаг 1: индексация референса
echo "[1/11] индексация референса"
hisat2-build $ref $ref_dir/chr$CHR
samtools faidx $ref

#шаг 2: контроль качества исходных чтений
echo "[2/11] fastqc исходных чтений"
fastqc $forward_reads $reverse_reads -o $output_dir/fastqc

#шаг 3: триммирование чтений
echo "[3/11] триммирование trimmomatic"
TrimmomaticPE -$phred $forward_reads $reverse_reads \
    $output_dir/trimmed/trim_1_paired.fastq.gz
$output_dir/trimmed/trim_1_unpaired.fastq.gz \
    $output_dir/trimmed/trim_2_paired.fastq.gz
$output_dir/trimmed/trim_2_unpaired.fastq.gz \
    TRAILING:$trim_trailing MINLEN:$trim_minlen

#шаг 4: контроль качества после триммирования
echo "[4/11] fastqc после триммирования"
fastqc $output_dir/trimmed/trim* -o $output_dir/fastqc

#шаг 5: картирование на референс
echo "[5/11] картирование hisat2"
hisat2 -x $ref_dir/chr$CHR \
```

```
    -1 $output_dir/trimmed/trim_1_paired.fastq.gz \
    -2 $output_dir/trimmed/trim_2_paired.fastq.gz \
    -p $nthreads --no-spliced-alignment \
    > $output_dir/mapping/${ID}_map.sam 2>
$output_dir/mapping/map_log.txt

#шаг 6: сортировка и индексирование bam
echo "[6/11] сортировка и индексирование bam"
samtools sort -o $output_dir/mapping/${ID}_map.bam
$output_dir/mapping/${ID}_map.sam
samtools index $output_dir/mapping/${ID}_map.bam
samtools flagstat $output_dir/mapping/${ID}_map.bam >
$output_dir/mapping/analysed_bam.txt

#шаг 7: отбор чтений по хромосоме
echo "[7/11] отбор чтений для хромосомы $CHR"
samtools view -h -bS $output_dir/mapping/${ID}_map.bam $CHR >
$output_dir/mapping/${CHR}.chr.bam
samtools flagstat $output_dir/mapping/${CHR}.chr.bam >
$output_dir/mapping/${CHR}_analysed_bam.txt

#шаг 8: отбор правильно парных чтений
echo "[8/11] отбор правильно картированных пар"
samtools view -f 2 -bS $output_dir/mapping/${CHR}.chr.bam >
$output_dir/mapping/correct_${CHR}.bam
samtools index $output_dir/mapping/correct_${CHR}.bam

#шаг 9: пересечение с bed
echo "[9/11] пересечение bam с bed"
bedtools intersect -a $output_dir/mapping/correct_${CHR}.bam
-b $bed_file > $output_dir/mapping/exom_${CHR}.bam
bedtools intersect -a $output_dir/mapping/correct_${CHR}.bam
-b $bed_extended >
$output_dir/mapping/exom_extended_${CHR}.bam

#шаг 10: вызов вариантов
echo "[10/11] вызов вариантов bcftools"
bcftools mpileup -f $ref
$output_dir/mapping/correct_${CHR}.bam | \
    bcftools call -mv -o $output_dir/vcf/variants.vcf
bcftools stats $output_dir/vcf/variants.vcf >
$output_dir/vcf/var_stats.txt

#шаг 11: фильтрация вариантов
echo "[11/11] фильтрация вариантов"
```

```
bcftools filter -i "QUAL>$filt_qual && DP>$filt_depth" \
    $output_dir/vcf/variants.vcf -o
$output_dir/vcf/filtred_variants.vcf
bcftools stats $output_dir/vcf/filtred_variants.vcf >
$output_dir/vcf/filt_var_stats.txt

echo "пайплайн завершён успешно"
echo "фильтрованный vcf: $output_dir/vcf/filtred_variants.vcf"
```

Output: создаются паки
/mnt/scratch/NGS/bratzveron/results/mapping
/mnt/scratch/NGS/bratzveron/results/trimmed
/mnt/scratch/NGS/bratzveron/results/vcf
/mnt/scratch/NGS/bratzveron/results/fastqc