1. Подготовка референса

1) Получение референса

Нам досталась девятая хромосома человека. Никаких особо знаменитых генов она не содержит, разве что гликозилтрансферазы систем групп крови ABO. Ну и пусть.

Для дальнейшей работы надо проиндексировать хромосому, чтобы программа картирования лучше ориентировалась в файле и в последовательности. Проиндексируем для hisat2:

```
hisat2-build Homo_sapiens.GRCh38.dna.chromosome.9.fa chr9_indexed
```

На выходе 8 файлов вида: chr9_indexed.?.ht2

2) Индексация samtools

Индексация - создание вспомогательного файла-индекса для быстрого доступа к данным. Проиндексировали samtools:

```
samtools faidx Homo_sapiens.GRCh38.dna.chromosome.9.fa
```

Выдача — файл с данными:

Таблица 1.

<NAME>	<LENGTH>	<OFFSET>	<LINEBASES>	<LINEWIDTH>
<название последовательности>	<длина>	<смещение до первой буквы>	<длина строки в нуклеотидах>	<длина строки в байтах>
9	138394717	56	60	61

OFFSET = 56, то есть первые 56 байт, это название последовательности:
">9 dna:chromosome chromosome:GRCh38:9:1:138394717:1 REF", а дальше уже идут буковки последовательности. Последняя строка может быть короче, поэтому параметры LINEBASES и LINEWIDTH отличаются.

2. Чтение ДНК

1) Описание образца по странице NCBI

Таблица 2.

SRR	SRR10720404
Sample (образец):	SAMN13614131
Прибор для секвенирования	Illumina Genome Analyzer IIx
Организм	Homo sapiens
Стратегия секвенирования (полногеномное, экзомное, таргетная панель)	OTHER
Парноконцевые или одноконцевые чтения	PAIRED (парноконцевые)
Сколько чтений ожидается (spots)	38'518'929

2) Проверка качества исходных чтений

Проверяем качество чтений:

```
fastqc SRR10720404_1.fastq.gz
fastqc SRR10720404_2.fastq.gz
```

На выходе 2 файла SRR10720404_1_fastqc.html и SRR10720404_2_fastqc.html.

Total Sequences (количество пар чтений) — 38518929 и для прямых, и для обратных чтений

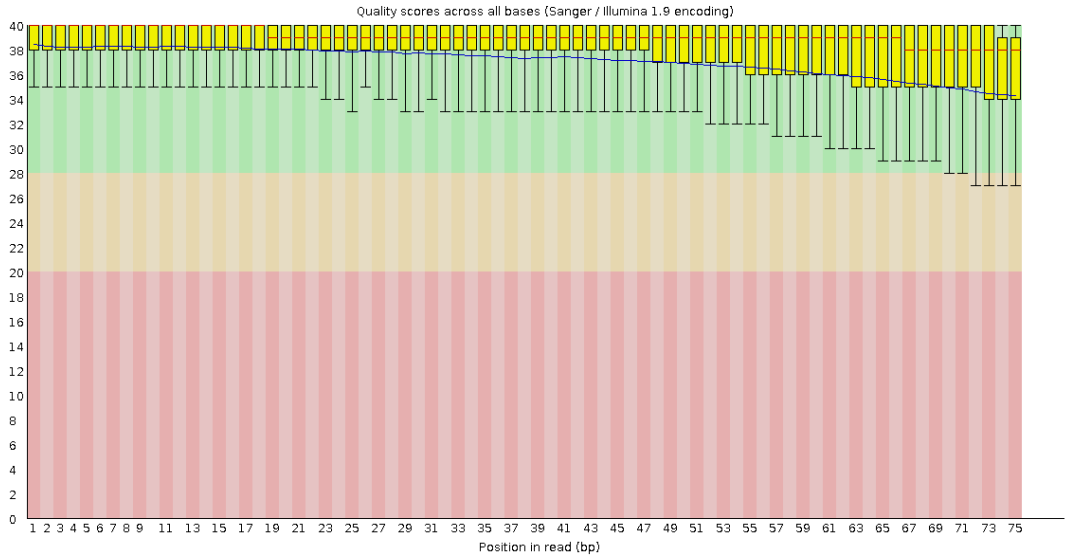


Рис. 1 Per base sequence quality для “прямых” чтений

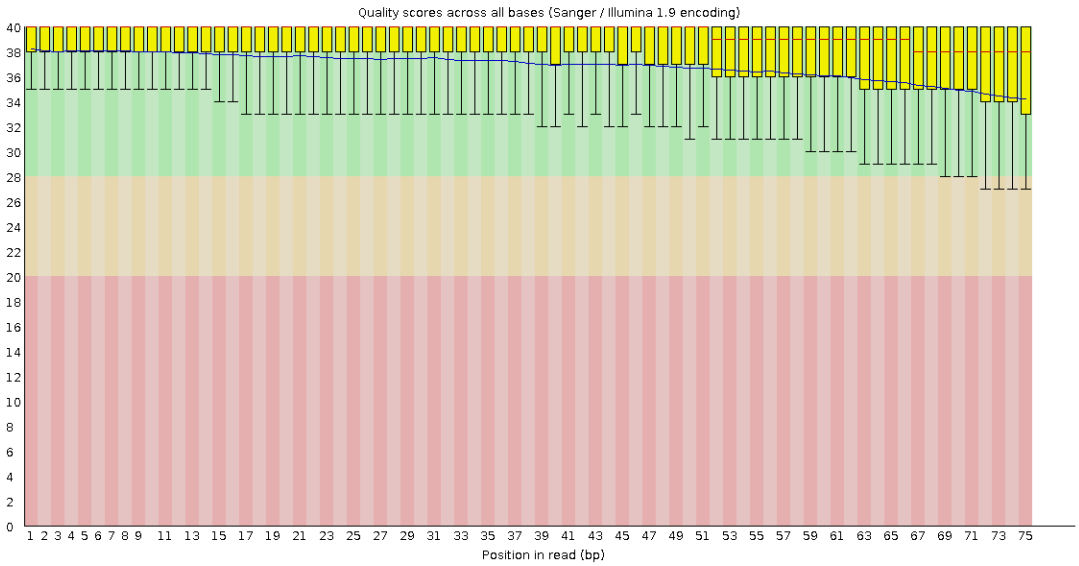


Рис. 2 Per base sequence quality для “обратных” чтений

В целом качество очень хорошее, только боксплоты четырёх последних позиций выходят из зелёной зоны, и то незначительно.

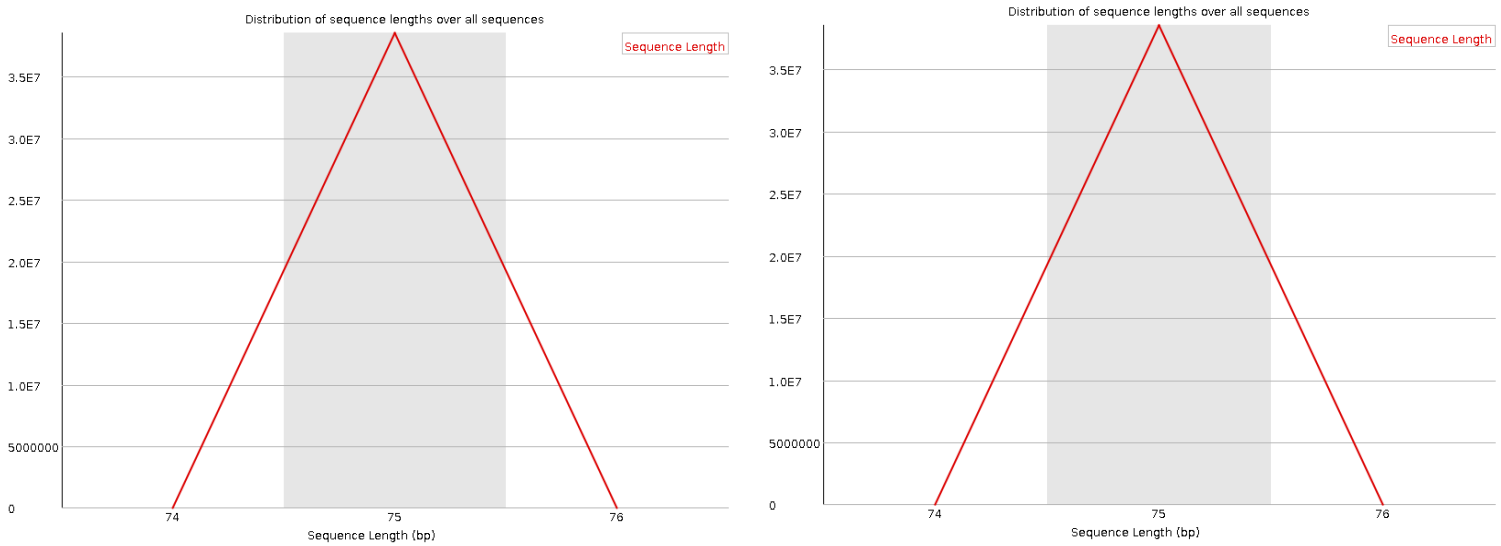


Рис. 3. Sequence Length Distribution (слева - для прямых чтений, справа - для обратных). Все чтения длиной 75 нуклеотидов.

3) Фильтрация чтений

Программа Trimmonatic выполняет пошаговую обрезку и фильтрацию парных чтений (R1 и R2), удаляя низкокачественные нуклеотиды, адаптеры и технические последовательности, слишком короткие чтения. Она анализирует чтения из архива.

PE — paired ends, режим для парно-концевых данных;

-phred33 — шкала качества Phred+33 (стандарт для Illumina). Указали входные файлы и по два выходных для каждого: оба чтения прошли фильтрацию и когда второе чтение не прошло.

TRAILING:20 — обрезка с конца, где среднее качество <20;

MINLEN:40 — оставить только чтения длиной ≥40 нуклеотидов.

```
TrimmomaticPE -phred33 \
  SRR10720404_1.fastq.gz SRR10720404_2.fastq.gz \
  SRR10720404_1_paired.fastq.gz   SRR10720404_1_unpaired.fastq.gz \
  SRR10720404_2_paired.fastq.gz SRR10720404_2_unpaired.fastq.gz \
  TRAILING:20 MINLEN:40
```

4) Проверка качества триммированных чтений

```
fastqc SRR10720404_1_paired.fastq.gz
fastqc SRR10720404_1_unpaired.fastq.gz
fastqc SRR10720404_2_paired.fastq.gz
fastqc SRR10720404_2_unpaired.fastq.gz
```

На выходе 4 html файла.

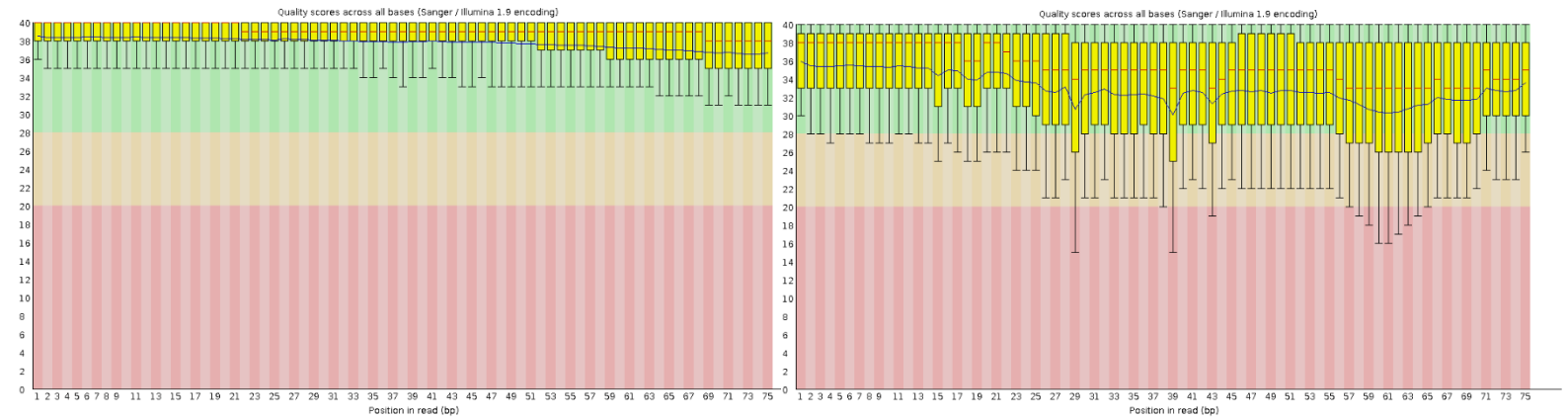


Рис. 4. Прямые чтения, слева paired, справа unpaired

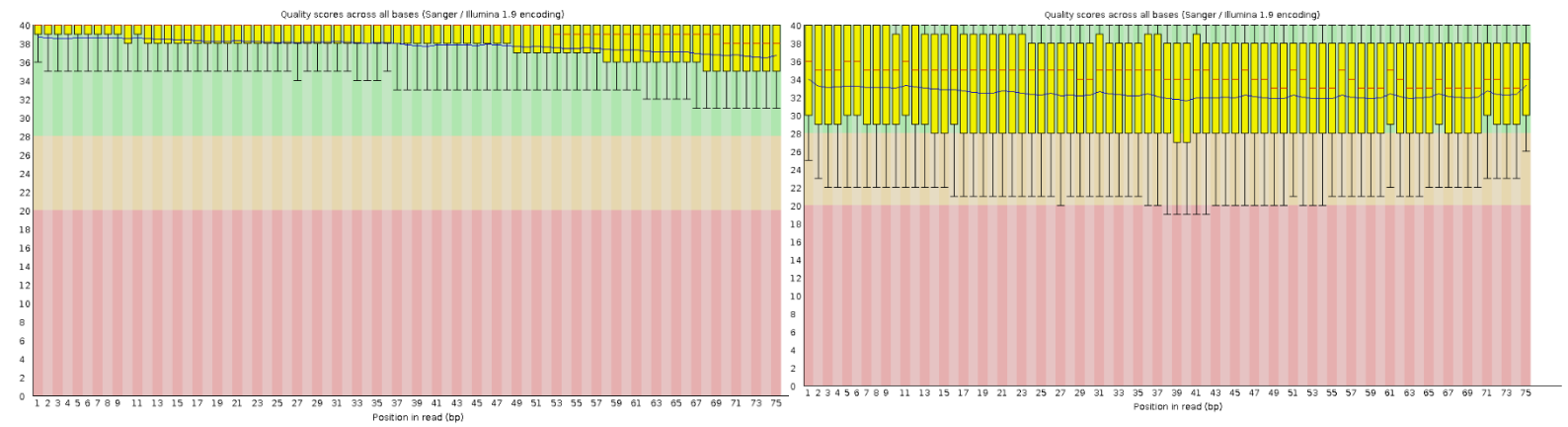


Рис. 5. Обратные чтения, слева paired, справа unpaired

Видно, что unpaired чтения не особо получились. Все позиции плохого качества, некоторые заходят в красную зону, но все имеют качество больше 20, это подтверждает нашу фильтрацию по качеству триммером. Непарные чтения (unpaired) — это чтения, для которых один член пары не прошёл фильтрацию, поэтому «пара» потеряна. Закономерно, что у таких чтений будет хуже качество, если их пары не смогли пройти фильтрацию.

Paired чтения стали лучше, последние позиции тоже находятся в зелёной зоне.

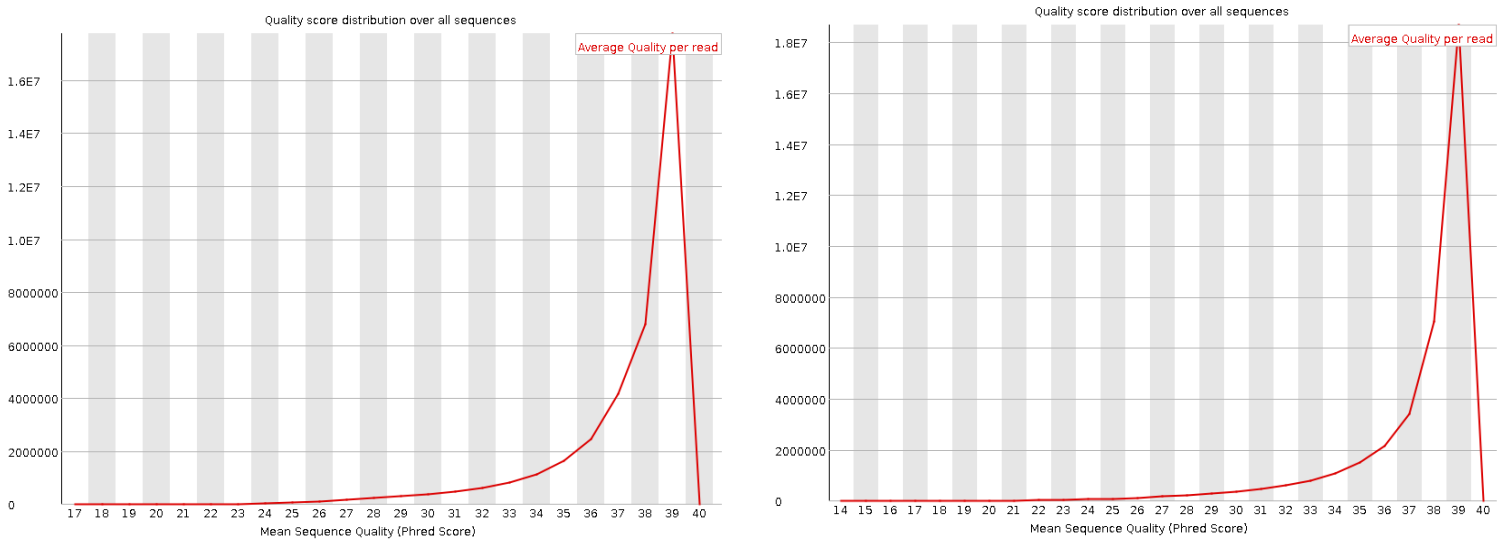


Рис. 6. Sequence Length Distribution слева - для парных чтений, справа - для непарных)

Средняя длина прочтения уменьшилась – с 75 до 39 позиций. Если на **Рис. 3** мы видели ровное распределение длин, то тут оно стало более неоднородным.

Часть II

1. Картирование чтений на референсный геном

Картирование это процесс определения, где именно на геномной последовательности расположены сиквенсные прочтения.

Для картирования используем программу hisat2, для которой мы индексировали нашу хромосому ранее.

-x HISAT2 автоматически ищет все файлы с шаблоном `chr9_indexed.*.ht2` в текущей директории (или по указанному пути)

-1 и **-2** Файл с парными прямыми и обратными чтениями

-p указать чтобы потоки поиска шли параллельно на разных ядрах процессора

--no-splice-alignment отключить сплайсированные выравнивания, чтения не игнорируют интроны

-S mapped.sam направляет выдачу в файл **.sam**, мы в явном виде указываем имя файла

2> hisat2.log направляет ошибки и логи из STDERR в такой файл

```
hisat2 \
  -x chr9_indexed \
  -1 SRR10720404_1_paired.fastq.gz \
  -2 SRR10720404_2_paired.fastq.gz \
  -p 8 \
  --no-spliced-alignment \
  -S mapped.sam \
  2> hisat2.log
```

На выходе получили соответственно два файла — mapped.sam это основной файл выдачи и hisat2.log с обзором процесса картирования.

2. Конвертация sam в bam

Измерим размер mapped.sam:

```
du -h mapped.sam
```

Получилось 15Gb. Очень много. В квоту бы не уложились.

Шапка файла выглядит так:

@HD - версия формата и сортировка

@SQ - информация о референсных последовательностях: название последовательности — хромосома 9 и LN — её длина

@RG - информация о read group (о программе)

CL – инфа про командную строку

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:9    LN:138394717
@PG     ID:hisat2       PN:hisat2       VN:2.2.1
CL:"/usr/bin/hisat2-align-s --wrapper basic-0 -x chr9_indexed -p 8 --no-spliced-alignment -S
 mapped.sam -1 /tmp/250949.inpipe1 -2 /tmp/250949.inpipe2"
```

В общем выдача SAM выглядит так:

Рис. 7. Выдача .sam

Конвертируем этого тяжеловеса в формат полегче — **.bam**:

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

Аналогично взвесим bam:

```
du -h mapped.bam
```

— 4Gb значительно лучше

Ну и индексируем бам файл:

```
samtools index maped.bam
```

В выдаче файл mapped.bam.bai

3. Анализ bam файла

Проанализируем bam файл с помощью команды samtools:

```
samtools flagstat mapped.bam
```

В первой строке выходных данных указано общее количество операций чтения, которые прошли проверку качества или завершились неудачей. Всего 76’279’686 прочтений, все успешные (#PASS + #FAIL). Процент картированных чтений низкий (7.18%), потому что наши прочтения всего генома выравнивались не ко всему экзому, а на отдельно взятую девятую хромосому.

```
76279686 + 0 in total (QC-passed reads + QC-failed reads)
74921916 + 0 primary
1357770 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
5475440 + 0 mapped (7.18% : N/A)
4117670 + 0 primary mapped (5.50% : N/A)
74921916 + 0 paired in sequencing
37460958 + 0 read1
37460958 + 0 read2
3473420 + 0 properly paired (4.64% : N/A)
3565024 + 0 with itself and mate mapped
552646 + 0 singletons (0.74% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

1) Что значит число в поле «in total»?
   Это общее количество прочтений, которые присутствуют в BAM-файле : QC-passed + QC-failed. Это сумма прочтений, прошедших и не прошедших контроль качества.
2) Сколько чтений (не пар!) поступило на картирование?
   Чтений поступило 74921916, некоторые из них выровнялись к нескольким местам, и программа выровняла их повторно — это строка 1357770 + 0 secondary
3) Сколько чтений картировано на референс в корректных парах в штуках?
   Строка 3473420 + 0 properly paired (4.64% : N/A)
4) Сколько чтений картировано на референс в корректных парах в процентах относительно потупивших на картирование?
   Из той же строки — 4.64%

4. Получение чтений, картированных на вашу хромосому

Для этого воспользуемся командой samtools и именем хромосомы, которое нашли ранее:

```
samtools view -h -bS mapped.bam 9 > 9.chr.bam
```

-h указывает что мы извлекаем данные для определенной области, например, для хромосомы 9.

По дефолту на выходе получается **sam** файл, но указав **-bS** программа сразу конвертирует его и выдаст на выходе **bam**.

Аналогично п.3 применим команду flagstat и получим вот такое:

```
6028086 + 0 in total (QC-passed reads + QC-failed reads)
4670316 + 0 primary
1357770 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
5475440 + 0 mapped (90.83% : N/A)
4117670 + 0 primary mapped (88.17% : N/A)
4670316 + 0 paired in sequencing
2335158 + 0 read1
2335158 + 0 read2
3473420 + 0 properly paired (74.37% : N/A)
3565024 + 0 with itself and mate mapped
552646 + 0 singletons (11.83% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Ключевые изменения:

Таблица 2.

Параметр	Картирование на весь геном	Картирование на нашу хромосому
Общее количество прочтений	76,279,686	6,028,086
Процент выравнивания	7.18%	90.83%
Properly paired	4.64%	74.37%
Singletons	0.74%	11.83%

5. Получение только правильно картированных пар чтений

```
samtools view -f 2 -bS 9.chr.bam > correct_al_9.bam
```

-f - фильтр: оставить только риды с указанными флагами

2 = битовая маска 0x2 - "each segment properly aligned". То есть мы указывает оставлять только корректные выравнивания:

- Оба рида выровнены
- Правильная ориентация: → ← (forward-reverse)
- Ожидаемое расстояние между ридами
- Одна хромосома для обоих ридов
- POS read1 < POS read2 (для forward-reverse ориентации)

Проанализируем с помощью **samtools flagstat**:

```
samtools flagstat correct_al_9.bam > info_correct_ali_9.out
```

В полученном файле вот такая выдача:

```
4395284 + 0 in total (QC-passed reads + QC-failed reads)
3473420 + 0 primary
921864 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
4395284 + 0 mapped (100.00% : N/A)
3473420 + 0 primary mapped (100.00% : N/A)
3473420 + 0 paired in sequencing
1736710 + 0 read1
1736710 + 0 read2
3473420 + 0 properly paired (100.00% : N/A)
3473420 + 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)
```

Количество чтений ещё уменьшилось с 6 млн до 4,4 млн

mapped теперь 100%, то есть все прочтения выровнены, и все они корректные — properly paired (100.00%)

Проиндексируем файл с хорошо картированными прочтениями:

```
samtools index correct_al_9.bam
```

6. Получение чтений, картированных только в границы экзома

Удалим чтения, которые содержат интроны. Сделаем это с помощью bedtools intersect. Он берет координаты каждого чтения из BAM, сравнивает с координатами из BED файла и оставляет только те чтения, которые пересекаются с BED регионами.

```
bedtools intersect -a correct_al_9.bam -b
/mnt/scratch/NGS/DATA/genes/seqcap_hg38.bed  > exom_9_chr.bam


samtools flagstat exom_9_chr.bam
```

Вновь читаем с samtools flagstat

```
2515277 + 0 in total (QC-passed reads + QC-failed reads)
2027892 + 0 primary
487385 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2515277 + 0 mapped (100.00% : N/A)
2027892 + 0 primary mapped (100.00% : N/A)
2027892 + 0 paired in sequencing
1012582 + 0 read1
1015310 + 0 read2
2027892 + 0 properly paired (100.00% : N/A)
2027892 + 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)
```

От всех правильно картированных прочтений для девятой хромосомы, экзомных только 2,5 млн или 57%.

Все чтения выровнены и все правильно спарены, ну это закономерно следует из фильтрации в пункте 5.

7. Получение чтений, картированных в границы расширенного экзома

Повторим пункт 6, но в качестве разметки экзома возьмём рисширенный файл. Судя по названию это экзоны с окрестностью в 5 нуклеотидов. То есть при таком считывании мы ещё можем узнаем последовательности сайтов сплайсинга, мутации в промоторах, изменения в энхансерах или сделать предположения насчёт альтернативного сплайсинга.

```
bedtools intersect -a correct_al_9.bam -b
/mnt/scratch/NGS/DATA/genes/seqcap_hg38_50.bed  >
exom_extended_9_chr.bam

samtools flagstat exom_extended_9_chr.bam
```

Выдача samtools falgstat:

```
2731685 + 0 in total (QC-passed reads + QC-failed reads)
2207166 + 0 primary
524519 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2731685 + 0 mapped (100.00% : N/A)
2207166 + 0 primary mapped (100.00% : N/A)
2207166 + 0 paired in sequencing
1103091 + 0 read1
1104075 + 0 read2
2207166 + 0 properly paired (100.00% : N/A)
2207166 + 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)
```

Расширенный экзом захватил ещё ~216 тысяч (8.6%) ридов дополнительно, при тех же 100% качества выравнивания.

Практикум 12

1. Получение вариантов

bcftools mpileup вычисляет вероятность нахождения той или иной буковки на определённой позиции. На выходе будет pileup файл с подсчётом нуклеотидов в каждой позиции. Программы работают со стандартными потоками STDINN/STDOUT, поэтому можем напрямую направлять выход из одной в другую через пайп.

bcftools call на основе вычисленных вероятностей определяет, является ли наблюдаемое отличие от референса настоящим генетическим вариантом или это ошибка прочтения или выравнивания.

```
bcftools mpileup -f \
Homo_sapiens.GRCh38.dna.chromosome.9.fa correct_al_9.bam | \
bcftools call -mv -o variants.vcf
```

-f —указание на последовательность референсной последовательности, которую мы индексировали в пункте 2

-m — multiallelic caller рассматривает все возможные аллели в позиции одновременно.

-v —выводить только вариабельные сайты, те которые совпадают с референсом, не выводить в файл.

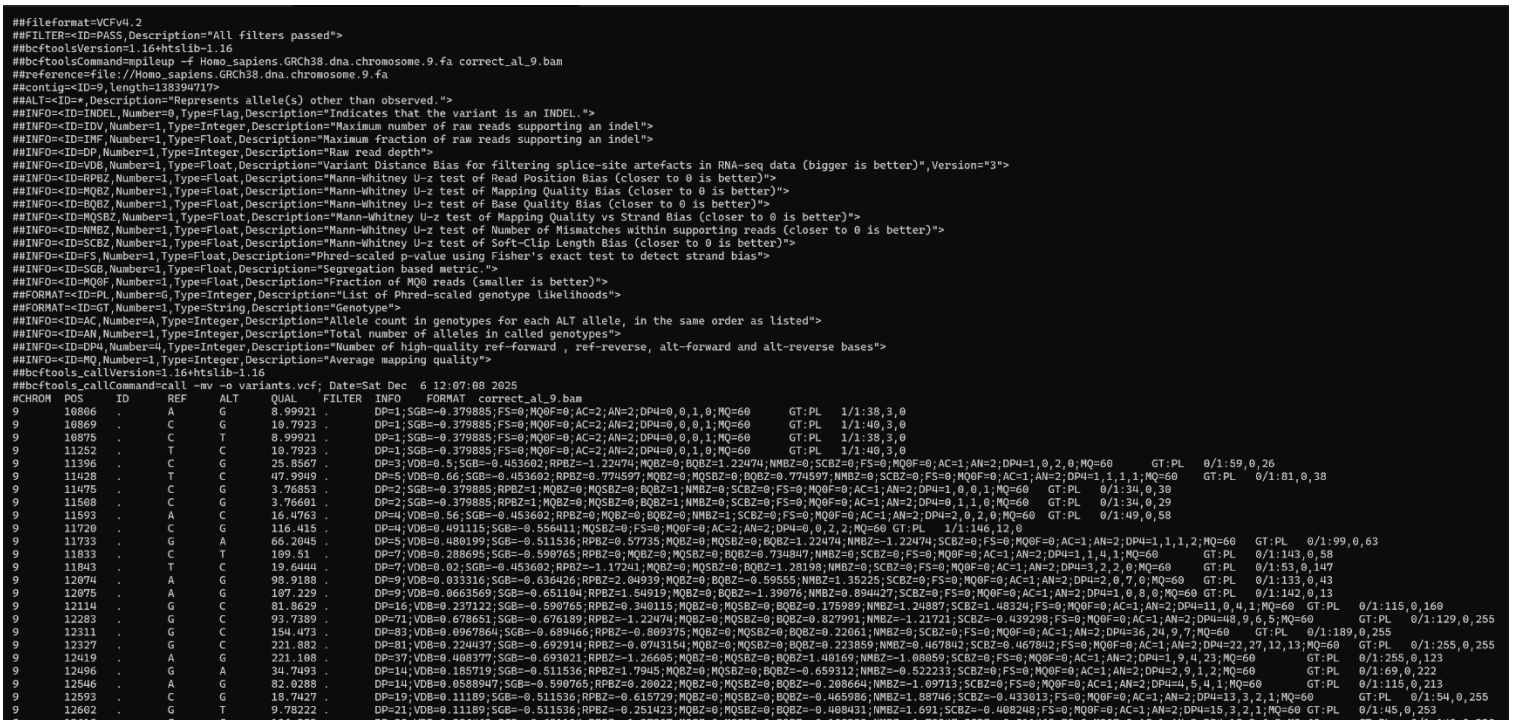
-o – указание output file

Структура и анализ VCF файла

- ## Шапка с метаданными, указанием референсной последовательности, её длиной и тд.
- # заголовки столбцов:

CHROM	Название нашей девятой хроомосомы
POS	Позиция варианта
ID	Здесь предполагается информация но у нас везде (.)
REF	Буква позиции из референсной последовательности
ALT	Альтернативная аллель
QUAL	Качество варианта (Phred-шкала)
FILTER	Может предполагаться фильтрация
INFO	Сведения о позиции и варианте

FORMAT	формат данных для каждого образца
correct_al_9.bam	фактические данные для каждого образца в формате



```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.16+htslib-1.16
##bcftoolsCommand=mpileup -f Homo_sapiens.GRCh38.dna.chromosome.9.fa correct_al_9.bam
##reference=file://Homo_sapiens.GRCh38.dna.chromosome.9.fa
##contig=<ID=9,length=138394717>
##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=Sat Dec  6 12:07:08 2025
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  correct_al_9.bam
9       10806   .       A       G       8.99921 .       DP=1;SGB=-0.379885;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60      GT:PL   1/1:38,3,0
9       10869   .       C       G       10.7923 .       DP=1;SGB=-0.379885;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60      GT:PL   1/1:40,3,0
9       10875   .       C       T       8.99921 .       DP=1;SGB=-0.379885;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60      GT:PL   1/1:38,3,0
9       11252   .       T       C       10.7923 .       DP=1;SGB=-0.379885;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60      GT:PL   1/1:40,3,0
9       11396   .       C       G       25.8567 .       DP=3;VDB=0.5;SGB=-0.453602;RPBZ=-1.22474;MQBZ=0;BQBZ=1.22474;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,0,2,0;MQ=60      GT:PL   0/1:59,0,26
9       11428   .       T       C       47.9949 .       DP=5;VDB=0.66;SGB=-0.453602;RPBZ=0.774597;MQBZ=0;MQSBZ=0;BQBZ=0.774597;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,1,1,1;MQ=60    GT:PL   0/1:81,0,38
9       11475   .       C       G       3.76853 .       DP=2;SGB=-0.379885;RPBZ=1;MQBZ=0;MQSBZ=0;BQBZ=1;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,0,0,1;MQ=60   GT:PL   0/1:34,0,30
9       11508   .       C       G       3.76601 .       DP=2;SGB=-0.379885;RPBZ=1;MQBZ=0;MQSBZ=0;BQBZ=1;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=0,1,1,0;MQ=60   GT:PL   0/1:34,0,29
9       11593   .       A       C       16.4763 .       DP=4;VDB=0.56;SGB=-0.453602;RPBZ=0;MQBZ=0;BQBZ=0;NMBZ=1;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=2,0,2,0;MQ=60  GT:PL   0/1:49,0,58
9       11720   .       C       G       116.415 .       DP=4;VDB=0.491115;SGB=-0.556411;MQSBZ=0;FS=0;MQ0F=0;AC=2;AN=2;DP4=0,0,2,2;MQ=60 GT:PL   1/1:146,12,0
9       11733   .       G       A       66.2045 .       DP=5;VDB=0.480199;SGB=-0.511536;RPBZ=0.57735;MQBZ=0;MQSBZ=0;BQBZ=1.22474;NMBZ=-1.22474;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,1,1,2;MQ=60   GT:PL   0/1:99,0,63
9       11833   .       C       T       109.51  .       DP=7;VDB=0.288695;SGB=-0.590765;RPBZ=0;MQBZ=0;MQSBZ=0;BQBZ=0.734847;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,1,4,1;MQ=60      GT:PL   0/1:143,0,58
9       11843   .       T       C       19.6444 .       DP=7;VDB=0.02;SGB=-0.453602;RPBZ=-1.17241;MQBZ=0;MQSBZ=0;BQBZ=1.28198;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=3,2,2,0;MQ=60     GT:PL   0/1:53,0,147
9       12074   .       A       G       98.9188 .       DP=9;VDB=0.033316;SGB=-0.636426;RPBZ=2.04939;MQBZ=0;BQBZ=-0.59555;NMBZ=1.35225;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=2,0,7,0;MQ=60   GT:PL   0/1:133,0,43
9       12075   .       A       G       107.229 .       DP=9;VDB=0.0663569;SGB=-0.651104;RPBZ=1.54919;MQBZ=0;BQBZ=-1.39076;NMBZ=0.894427;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,0,8,0;MQ=60 GT:PL   0/1:142,0,13
9       12114   .       G       C       81.8629 .       DP=16;VDB=0.237122;SGB=-0.590765;RPBZ=0.340115;MQBZ=0;MQSBZ=0;BQBZ=0.175989;NMBZ=1.24887;SCBZ=1.48324;FS=0;MQ0F=0;AC=1;AN=2;DP4=11,0,4,1;MQ=60  GT:PL   0/1:115,0,160
9       12283   .       G       C       93.7389 .       DP=71;VDB=0.678651;SGB=-0.676189;RPBZ=-1.22474;MQBZ=0;MQSBZ=0;BQBZ=0.827991;NMBZ=-1.21721;SCBZ=-0.439298;FS=0;MQ0F=0;AC=1;AN=2;DP4=48,9,6,5;MQ=60       GT:PL   0/1:129,0,255
9       12311   .       G       C       154.473 .       DP=83;VDB=0.0967864;SGB=-0.689466;RPBZ=-0.809375;MQBZ=0;MQSBZ=0;BQBZ=0.22061;NMBZ=0;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=36,24,9,7;MQ=60    GT:PL   0/1:189,0,255
9       12327   .       G       C       221.882 .       DP=81;VDB=0.224437;SGB=-0.692914;RPBZ=-0.0743154;MQBZ=0;MQSBZ=0;BQBZ=0.223859;NMBZ=0.467842;SCBZ=0.467842;FS=0;MQ0F=0;AC=1;AN=2;DP4=22,27,12,13;MQ=60   GT:PL   0/1:255,0,255
9       12419   .       A       G       221.108 .       DP=37;VDB=0.408377;SGB=-0.693021;RPBZ=-1.26605;MQBZ=0;MQSBZ=0;BQBZ=1.40169;NMBZ=-1.08059;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=1,9,4,23;MQ=60        GT:PL   0/1:255,0,123
9       12496   .       G       A       34.7493 .       DP=14;VDB=0.185719;SGB=-0.511536;RPBZ=1.7945;MQBZ=0;MQSBZ=0;BQBZ=-0.659312;NMBZ=-0.522233;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=2,9,1,2;MQ=60        GT:PL   0/1:69,0,222
9       12546   .       A       G       82.0288 .       DP=14;VDB=0.0588947;SGB=-0.590765;RPBZ=0.20022;MQBZ=0;MQSBZ=0;BQBZ=-0.208664;NMBZ=-1.09713;SCBZ=0;FS=0;MQ0F=0;AC=1;AN=2;DP4=4,5,4,1;MQ=60       GT:PL   0/1:115,0,213
9       12593   .       C       G       18.7427 .       DP=19;VDB=0.11189;SGB=-0.511536;RPBZ=-0.615729;MQBZ=0;MQSBZ=0;BQBZ=-0.465986;NMBZ=1.88746;SCBZ=-0.433013;FS=0;MQ0F=0;AC=1;AN=2;DP4=13,3,2,1;MQ=60       GT:PL   0/1:54,0,255
9       12602   .       G       T       9.78222 .       DP=21;VDB=0.11189;SGB=-0.511536;RPBZ=-0.251423;MQBZ=0;MQSBZ=0;BQBZ=-0.408431;NMBZ=1.691;SCBZ=-0.408248;FS=0;MQ0F=0;AC=1;AN=2;DP4=15,3,2,1;MQ=60 GT:PL   0/1:45,0,253
```

Рис. 8. VCF файл

Поле формат требует пояснений. Они есть [здесь](#).

Обычно **FORMAT** содержит данные типо GT:AD:DP:GQ:PL или подобное. Это список метрик через двоеточие.

У нас указано GT:PL. Это значит, что что для образца будут указаны два показателя: генотип (GT - Genotype) и вероятности генотипов (PL - Phred-scaled Genotype Likelihoods) — три числа, соответствующие трем возможным генотипам (0/0, 0/1 и 1/1).

Например: 0/1:129,0,255 , то есть 0/1 наиболее вероятный, чем меньше число, тем выше вероятность

```
- 0/0 : the sample is homozygous reference
- 0/1 : the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles
- 1/1 : the sample is homozygous alternate
```

Рис. 9. Значения GT

Проанализируем полученный vcf файл с помощью команды **bcftools stat**:

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

На выходе файл с разными статистиками.

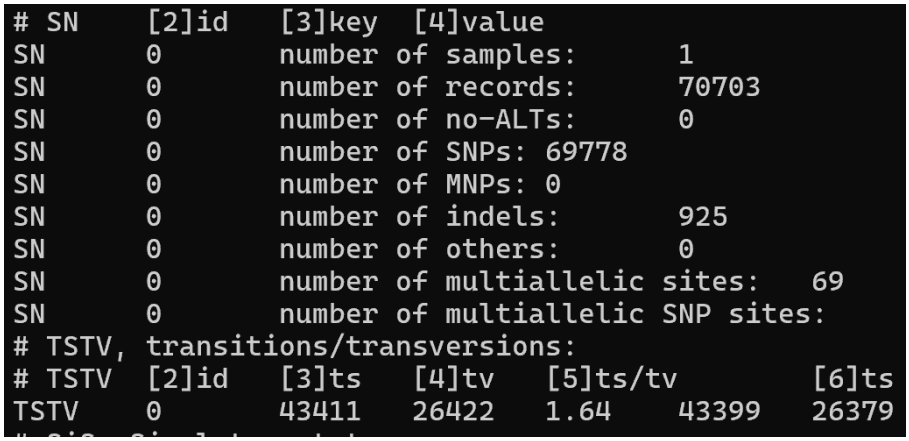


Рис. 10. Фрагмент выдачи bcftools stats

a) Сколько получилось вариантов?

— number of records: 70703

b) Сколько из полученных вариантов являются однонуклеотидными заменами?

— number of SNPs: 69778

c) Сколько получилось коротких вставок и делеций?

—number of indels: 925

d) Посмотрим подробнее на вывод команды bcftools mpileup:

```
bcftools mpileup -f Homo_sapiens.GRCh38.dna.chromosome.9.fa
correct_al_9.bam > whole_variants.vcf
```

-f так же указывает референсную последовательность

То есть программа выдаст информацию по вообще всем позициям последовательности хромосомы, и вариабельным и неизменным.

Тут для большинства позиций в поле ALT стоит <*>, что, в общем-то логично. Нет альтернативных позиций. Ну и файл заметно тяжелее — 4.1G против 8.5M отфильтрованных вариантов.

2. Фильтрация вариантов.

Отфильтруем варианты относительно покрытия и качества:

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

-i отобирает сайты, у которых quality больше 30 и глубина прочтений, то есть количество ридов, которые прошли фильтрацию и соответствуют одной из аллелей, больше 50

-o указание выходного файла

Проанализируем полученный vcf файл с помощью команды bcftools stat:

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

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      1
SN      0       number of records:      2081
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 2007
SN      0       number of MNPs: 0
SN      0       number of indels:       74
SN      0       number of others:       0
SN      0       number of multiallelic sites:   5
SN      0       number of multiallelic SNP sites:       2
# TSTV, transitions/transversions:
# TSTV  [2]id   [3]ts   [4]tv   [5]ts/tv        [6]ts (1st ALT)
TSTV    0       1376    633     2.17    1376    631     2.18
```

Рис. 11. Фрагмент выдачи bcftools stats после фильтрации

a) Сколько осталось вариантов после фильтрации (в штуках и в процентах)?

number of records: 2081 (2,84%)

b) Сколько осталось однонуклеотидных замен (в штуках и в процентах)?

number of SNPs: 2007 (2,88%)

c) Сколько осталось коротких вставок и делеций (в штуках и в процентах)?

number of indels: 74 (8%)

3. Аннотация вариантов

Воспользуемся Variant Effect Predictor , его web версией

Таблица 3

Category	Count
Variants processed	2081
Variants filtered out	0
Novel / existing variants	425 (20.4) / 1656 (79.6)
Overlapped genes	764
Overlapped transcripts	7123
Overlapped regulatory features	23

Всего обработано 2081 вариант, ровно столько мы и нафильтровали в предыдущем пункте. Других вариантов VEP не отфильтровал.

Новых вариантов (то есть тех, которых нет в Ensembl Variation database) - 425 (20,4%) Описанных - 1656 (76,9%)

Количество генов, на которые попадают варианты – 764. На 7123 транскрипта и 23 регулирующие области.

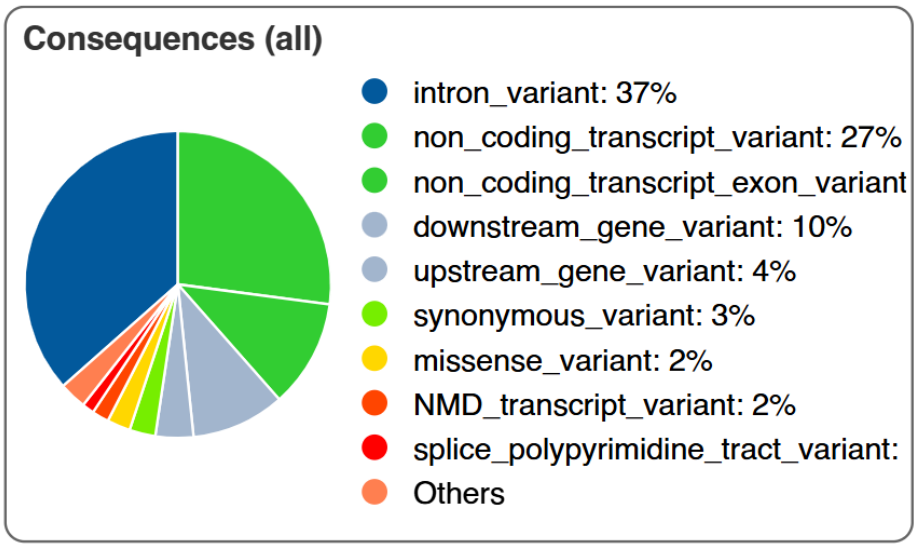


Рис. 12. Распределение попаданий вариантов на всю последовательность.

Видим, что большая часть попала на интроны и некодирующие участки. Что логично, ведь их совокупная длина больше таковой у кодирующих областей.

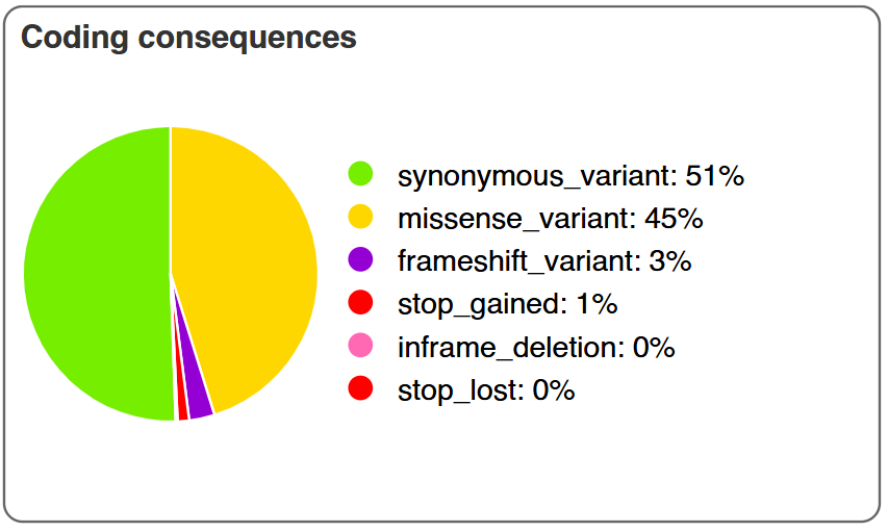


Рис. 13. Распределение попаданий вариантов на кодирующие участки

Большинство вариантов содержат синонимичные мутации, то есть которые не приводят к изменению структуры белка, либо миссенс мутации, которые могут привести к изменению структуры белочка и обеспечить мутационную изменчивость организма.

Установили фильтр `Impact is HIGH`, чтобы посмотреть сколько и какие генетические варианты имеют шанс на функциональное воздействие на белок или ген.

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon
.	9:16967-16967	G	splice_donor_variant, non_coding_transcript_variant	HIGH	WASHC1	ENSG00000181404	Transcript	ENST00000696150.1	protein_coding_CDS_not_defined	-
.	9:17476-17476	C	stop_gained	HIGH	WASHC1	ENSG00000181404	Transcript	ENST00000442898.5	protein_coding	6/11
.	9:17476-17476	C	stop_gained	HIGH	WASHC1	ENSG00000181404	Transcript	ENST00000696149.1	protein_coding	6/11
.	9:117253-117257	G	frameshift_variant	HIGH	FOXD4	ENSG00000170122	Transcript	ENST00000382500.4	protein_coding	1/1
.	9:6587205-6587205	A	stop_gained	HIGH	GLDC	ENSG00000178445	Transcript	ENST00000321612.8	protein_coding	15/25
.	9:6587205-6587205	A	stop_gained	HIGH	GLDC	ENSG00000178445	Transcript	ENST00000884130.1	protein_coding	15/24
.	9:6587205-6587205	A	stop_gained	HIGH	GLDC	ENSG00000178445	Transcript	ENST00000920236.1	protein_coding	15/25
.	9:6587205-6587205	A	stop_gained	HIGH	GLDC	ENSG00000178445	Transcript	ENST00000920237.1	protein_coding	15/24
.	9:6587205-6587205	A	stop_gained	HIGH	GLDC	ENSG00000178445	Transcript	ENST00000953081.1	protein_coding	15/26
.	9:15017482-15017482	C	splice_donor_variant, non_coding_transcript_variant	HIGH	-	ENSG00000298992	Transcript	ENST00000759707.1	lncRNA	-

Рис. 14. Варианты, сильно влияющие на структуру продукта

Всего их 58. Видим, что это в основном варианты (мутации) приводящие к преждевременной остановке транскрипции из-за образования стоп-кодона на месте кодирующего триплета, либо приводящие к сдвигу рамки считывания.

Также много вариантов отмеченных как splice_donor_variant,/ non_coding_transcript_variant, то есть мутации в сайте сплайсинга или в некодирующих зонах гена (вариант попадает в транскрипт, который не кодирует белок).

Чаще всего такие “опасные” вариации встречаются в 9 генах:

`WASHC1, FOXD4, GLDC, ADAMTSL1, HAUS6, IFNA10, IFNE, PGM5P2, FLJ43315,`

`WASHC1`, например, обеспечивает активность связывания альфа-тубулина и убиквитин-протеинлигазы.

`HAUS6` имеет отношение к формированию кинетохора и центрального веретена.

Эти гены важны для клетки, поломка в них скорее всего нарушит процесс деления из-за нарушений в механизмах образования и взаимодействий микротрубочек. Такая вариативность этих генов для меня стало неожиданным и показалось загадочным.

Практикум 13

Задача практикума: построить экспрессионный профиль на основании данных секвенирования РНК.

1. Описание образца

Таблица 4

ID образца РНК-чтений	ENCFF763QQV
Ссылка на информацию об образце	ENCSR816HLU
Организм и ткань (если есть)	Homo sapiens left lung tissue male adult (40 years)
Стратегия секвенирования (тотальная РНК, малые РНК, ...)	RNA-seq (total RNA-seq)
Парноконцевые или одноконцевые чтения	single-ended 100nt
цепь-специфичность	Strand-specific (reverse)

2. Проверка качества исходных чтений

Проверим как и раньше с помощью fastqc:

```
fastqc ENCFF763QQV.fastq.gz
```

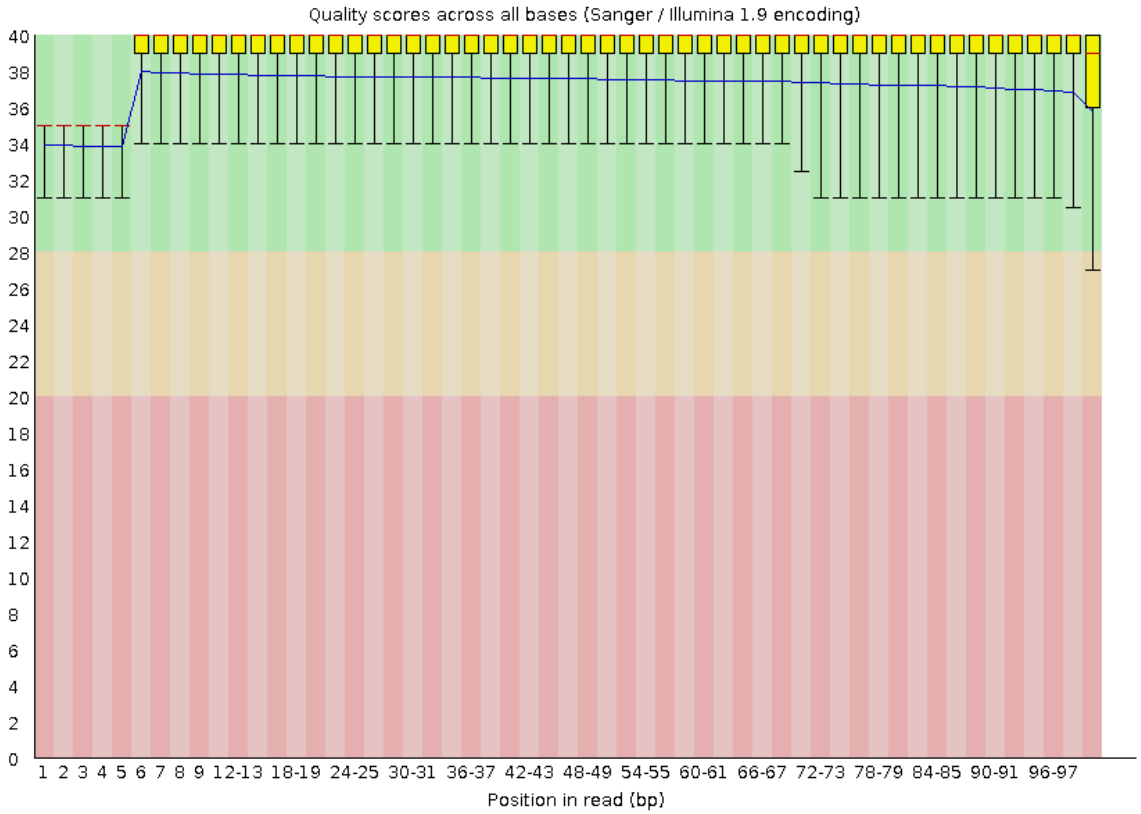


Рис. 15. Per base sequence quality

Всего чтений 50552179. Видно, что начала чтений, первые пять, плохие и некачественные. Так же самое последнее основание из чтений тоже неудовлетворительного качества.

Все чтения длины 100:

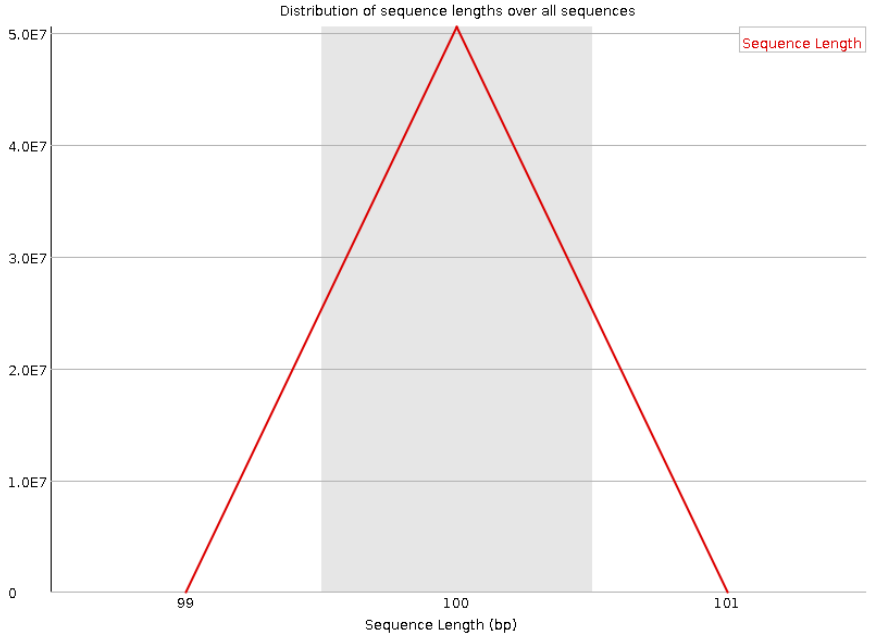


Рис. 16. Sequence Length Distribution

Также интерес вызывает GC-состав.

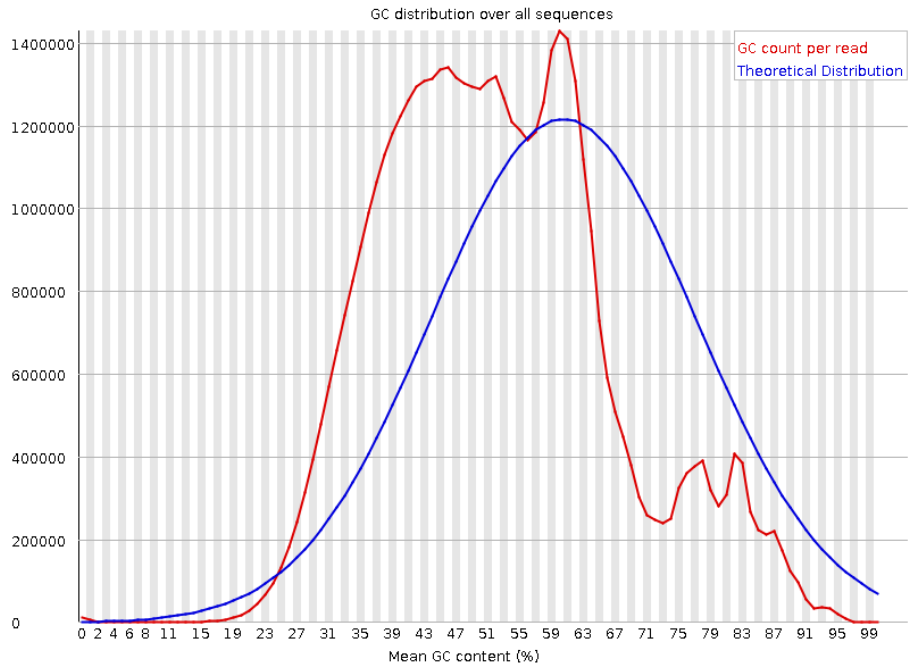


Рис. 17. Per sequence GC content

Раньше мы наблюдали распределение, напоминающее нормальное, а это какое-то кривоватое. Иногда происходит всплеск GC и они составляют 70-80% от последовательности.

3. Картирование чтений на референс

Картируем с помощью hisat2 на уже индексированную хромомсому

```
hisat2 -x /mnt/scratch/NGS/belll-3/chromosome/chr9_indexed -k 3 -U
ENCFF763QQV.fastq.gz -S rna_cart.sam 2> rna_hisat2.log
```

-x указание пути к префиксу индексированной хромосомы

-k 3 ищет до трёх выравниваний на чтение

-U указание на .gz файл с чтениями

-S выход в sam файл

2> выводить логи в отдельный файл

Содержимое лог файла:

```
50552179 reads; of these:
  50552179 (100.00%) were unpaired; of these:
    47191783 (93.35%) aligned 0 times
    3221149 (6.37%) aligned exactly 1 time
    139247 (0.28%) aligned >1 times
```

Из него можем понять, что все риды одинарные. Выровнялось на нашу 9 хромосому только 6,37 + 0,28 = 6,65% от всех (50552179) прочтений. Получилось довольно много, ведь образец из человека и картируем мы на тот же организм.

Переводим sam в bam

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

Индексируем bam

```
samtools index rna_cart.bam
```

И заглянем внутрь

```
samtools flagstat rna_cart.bam
```

Выдача ниже. Видим, что поступило 50’552’179 прочтений, 226’684 выровнялись повторно. В properly paired стоит N/A т.к. у нас одноконцевые чтения

```
50778863 + 0 in total (QC-passed reads + QC-failed reads)
50552179 + 0 primary
226684 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3587080 + 0 mapped (7.06% : N/A)
```

```
3360396 + 0 primary mapped (6.65% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Дальше отберем только те чтения, которые картировались на хромосому:

```
samtools view -h -bS rna_cart.bam 9 > rna_cart_9.bam
```

-h указывает что мы извлекаем данные для определенной области, например, для хромосомы 9.

По дефолту на выходе получается sam файл, но указав -bS программа сразу конвертирует его и выдаст на выходе bam.

Что же там внутри?!?!???!

```
samtools flagstat rna_cart_9.bam
```

```
3587080 + 0 in total (QC-passed reads + QC-failed reads)
3360396 + 0 primary
226684 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3587080 + 0 mapped (100.00% : N/A)
3360396 + 0 primary mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

То есть всего картировано 3’360’396 чтений ~6,7% от всех.

4. Поиск экспрессирующихся генов

Копировали файл с геномной разметкой себе в папочку

```
cp /mnt/scratch/NGS/DATA/genes/Homo_sapiens.GRCh38.110.chr.gtf .
```

Вот так он выглядит через less:

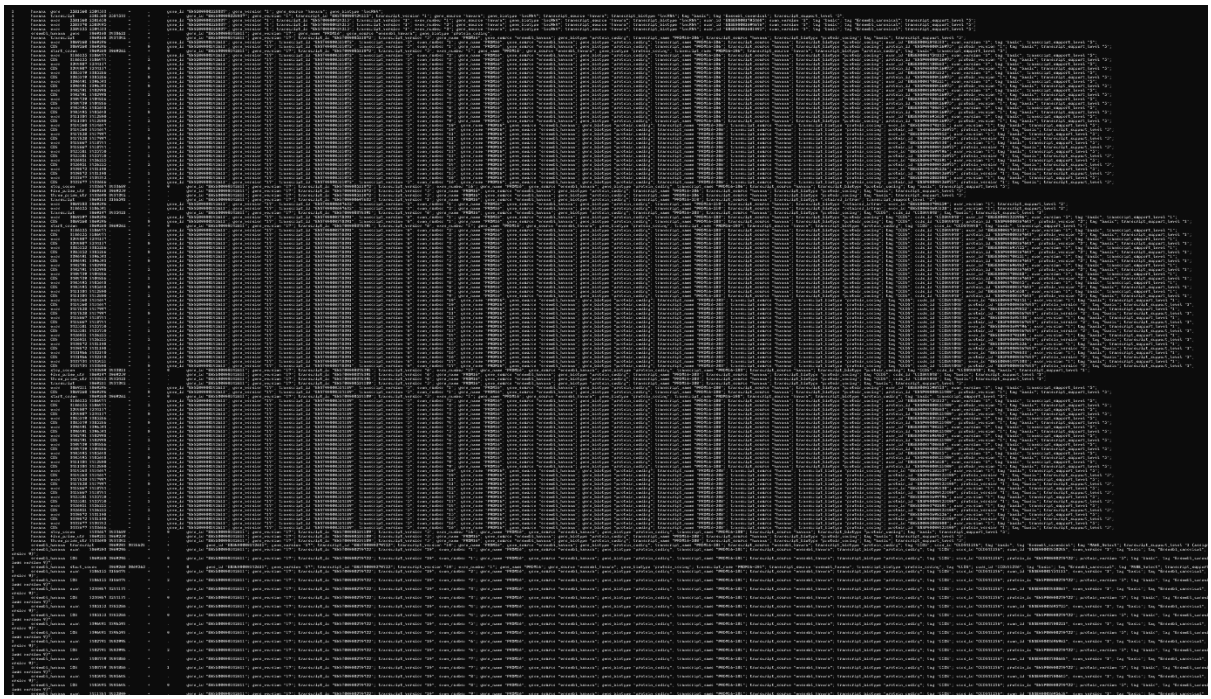


Рис. 18. Вид издалека на GTF файл геномной разметки

Gene Transfer Format — это табличный формат для аннотации геномных элементов. Каждая строка описывает один элемент (ген, экзон, транскрипт и т.д.).

Шапка:

```
#!genome-build GRCh38.p14   версия разметки
#!genome-version GRCh3814   версия генома
#!genome-date 2013-1238     дата публикации
#!genome-build-accession GCA_000001405.29  AC разметки
 #!genebuild-last-updated 2023-03001405.29 последнее обновление
```

Содержит следующие поля:

1) seqname – название хромосомы или скэффолда, на которых аннотирован ген. Тривиальное типо chr9 или ID Ensembl
2) source - название программы, создавшей этот элемент, или источника данных (название базы данных или проекта)
3) feature -название элемента, например Gene, Variation, Similarity
4) start – стартовая позиция, начиная с 1.

5) end – последняя позиция элемента
6) score -значение плавающей запятой (оценка).
7) strand - + или - цепь
8) frame - Одно из значений "0", "1" или "2". "0" означает, что первое основание признака является первым основанием кодона, "1" означает, что второе основание является первым основанием кодона, и так далее.
9) attribute - дополнительные атрибуты в формате "ключ значение;"

В поле source у нас указано `havana`

Грепнем чтобы посчитать количество генов:

```
grep '^9' Homo_sapiens.GRCh38.110.chr.gtf | cut -f 3 | \
grep 'gene' | wc -l
```

Получилось 2417. Википедия гласит: "По разным оценкам, 9-я хромосома содержит от 800 до 1200 генов". То есть наша оценка заметно завышено. Возможно это из-за того что некоторые гены аннотированы в этой графе как gene хотя таковыми уже не являются.

Вот с помощью такого конвейера легко убедиться в нашей гипотезе:

```
grep '^9' Homo_sapiens.GRCh38.110.chr.gtf | cut -f 3,6,9 | grep
'^gene' | less
```

В выдаче полно строчек типо:

```
gene    .       gene_id "ENSG00000270683"; gene_version "1";
gene_name "GARIN3P1"; gene_source "havana"; gene_biotype
"processed_pseudogene";
```

Processed pseudogene (обработанный псевдоген) — это нефункциональная копия гена, созданная через обратную транскрипцию мРНК и встройку в геном.

Далее для каждого гена из разметки посчитаем количество картированных конкретно на этот ген прочтений с помощью программы htseq-count:

```
htseq-count -f bam -s reverse -m union -t gene  rna_cart_9.bam \
Homo_sapiens.GRCh38.110.chr.gtf    1> gene_counts.txt  2> htseq.log
```

-f формат входных данных sam или bam

-s strandedness (ориентация ридов)

-m режим подсчёта пересечений (по дефолту union)

-t тип элемента для подсчёта

1>стандартный вывод

2>ошибки и предупреждения

Чтобы помсотреть на сколько генов лягли наши прочтения, можно посмотреть конец файла

```
tail 5 genes_count.txt
```

```
__no_feature    242620
__ambiguous     118053
__too_low_aQual 0
__not_aligned   0
__alignment_not_unique  139247
```

И из общего количества чтений отобранных в п.3 вычтем неспецифику (no_feature, ambiguous, alignment_not_unique):

3’360’396 - 242’620 - 118’053 - 139’247 = 2’860’476

Аннотация, кстати, значит:

- no_feature – то есть прочтение не легло ни на один элемент, который мы выбрали параметром **-t,** ни на один ген в нашем случае.
- ambiguous – двусмысленные, которые одновременно перекрываются с несколькими аннотированными объектами
- too_low_aQual - отклонённые из-за слишком низкого качества выравнивания
- not_aligned - риды, которые не удалось выровнять ни к одной позиции в референсном геноме
- alignment_not_unique – прочтения, которые выровнялись более ччем в одно место. Это могут быть повторы в геноме, паралоги или проблемы качества ридов.

Либо то же самое можно сделать вот таким конвейером:

```
grep '^ENSG' genes_count.txt | \
awk '$2 > 0 {sum += $2} END {print sum}'
2860476
```

Выдачи согласуются. Это 85,1% от общего числа прочтений.

Соответственно не на гены попало 499,920 ридов (14,9%)

Вот такой конвейер покажет на сколько генов легло не 0 чтений:

```
grep '^ENGS' genes_count.txt | awk '$2 > 0' | wc -l
1424
```

Это уже больше напоминает оценку количества генов на 9 хромосоме из Википедии.

5. Аннотация высоко экспрессируемых генов

Хотим узнать что-нибудь интересное про гены, на которые картировалось наибольшее количество прочтений. Так как у нас рнк-секвениирование, можем сделать вывод, что это наиболее экспрессируемые гены:

```
grep '^ENSG' gene_counts.txt | sort -k2,2nr | head -n 10
```

sort -k2,nr – то что мы сортируем по второму столбцу, в численном формате и по убыванию

Выдача:

```
ENSG00000265735 1047109
ENSG00000196205 140577
ENSG00000277027 105076
ENSG00000137076 30004
ENSG00000041982 24375
ENSG00000044574 22573
ENSG00000130635 19051
ENSG00000223551 18527
ENSG00000231991 16555
ENSG00000148303 16319
```

Первый и третий это RNA component of mitochondrial RNA processing endoribonuclease, второй не найдет (?), вот четвёртый кодирует белочек talin-1.

Ген кодирует белок цитоскелета, который концентрируется в областях контакта клетка-субстрат и клетка-клетка. Кодируемый белок играет важную роль в сборке актиновых волокон и в распространении и миграции различных типов клеток, включая фибробласты и остеокласты.

Можем вспомнить, что наши образцы взяты из ткани лёгкого. Возможно, клетки там плотно не зафиксированы по отношению друг к другу, поэтому в них увеличивается экспрессия белка, отвечающего за адгезию к другим клеткам.

Что интересно, TLN (talin-1) связан с фибробластами. Мутации в гене RNA component of mitochondrial RNA processing endoribonuclease, наиболее экспрессируемые в нашей выдаче, вызывают хондроэктодермальную дисплазию (синдром хряща-волоса, Cartilage-Hair Hypoplasia, CHH)[1]. Я вижу здесь занимательную связь.

Можно срастить это в теорию: Talin-1 обеспечивает связь клетки с матриксом, активирует сигнальные пути (например, FAK, PI3K/AKT). RMRP регулирует продуктивность митохондрий, поставляет энергию для энергозатратных процессов (миграция, адгезия).

Гипотеза: в лёгочной ткани высокая экспрессия RMRP может поддерживать энергозатраты клеток, активно использующих talin-1 (например, фибробластов при росте ткани). Может подопытный мужчина курильщик и клетки лёгких активно делятся и держатся друг за друга чтобы не погибать. Впрочем, это фантазии.

Пятый в выдаче tenascin C: белок внеклеточного матрикса. Он участвует в управлении мигрирующими нейронами, а также аксонами в процессе развития, синаптической пластичности и регенерации нейронов. Тоже интересная находка.

Визуализируем talin-1 с помощью геномного браузера в NCBI:

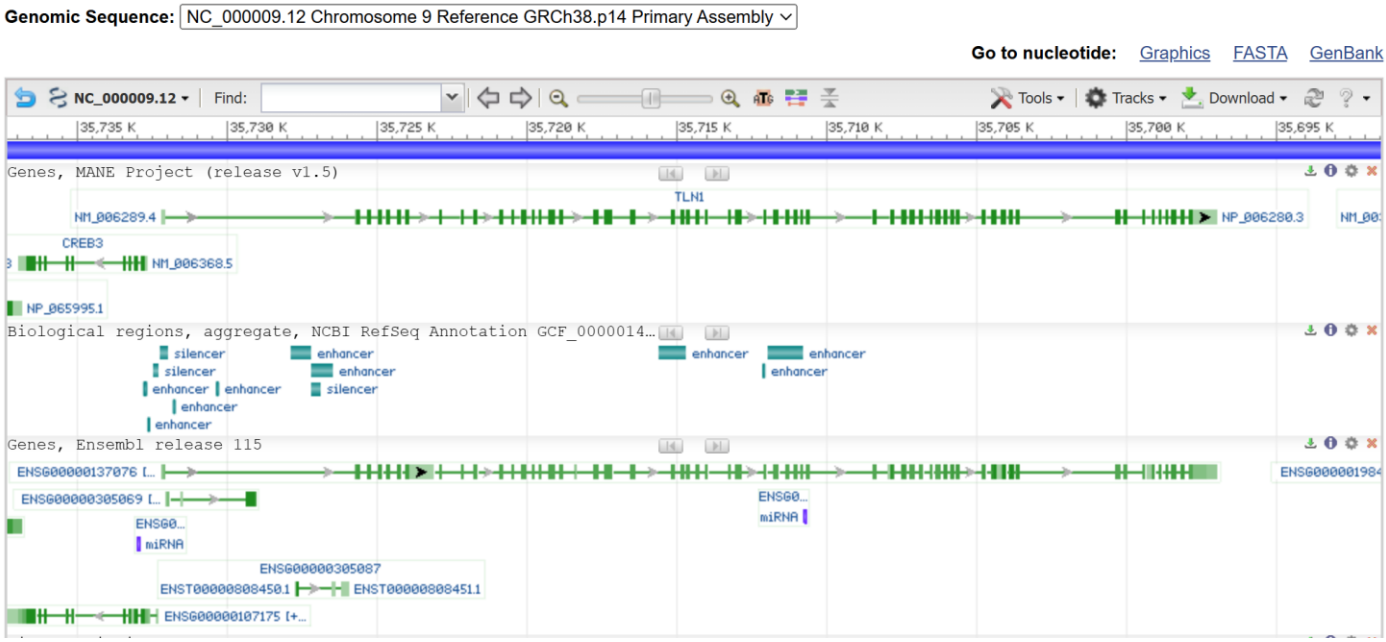


Рис. 19. Экзон-интронная структура гена ENSG00000137076

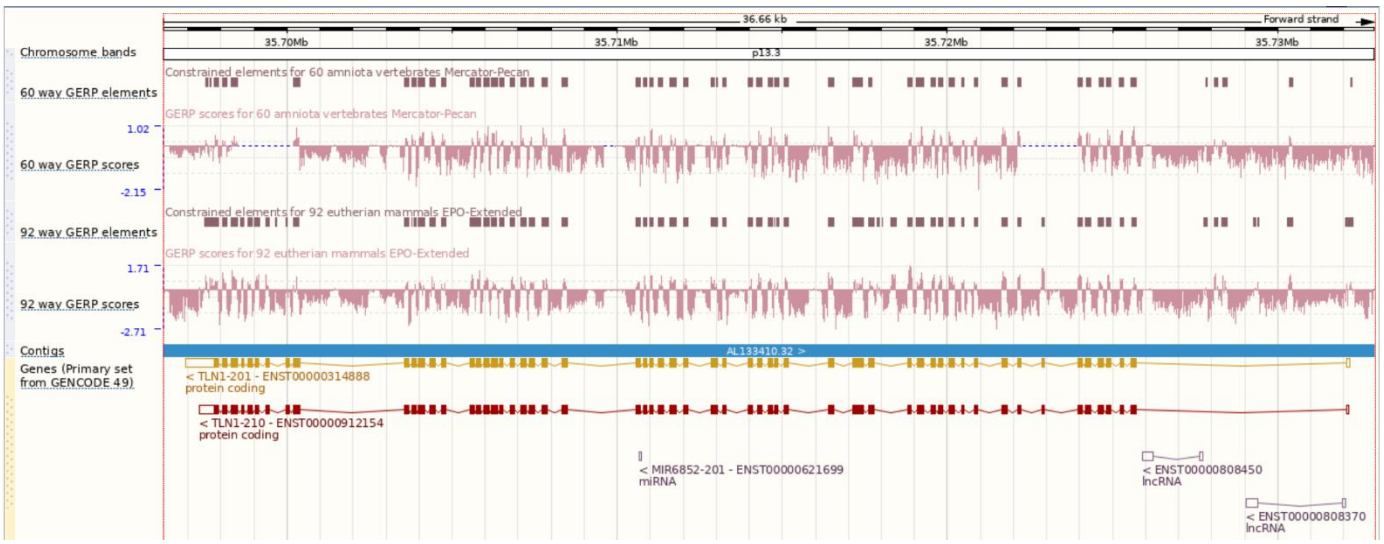


Рис. 20. Треки консервативности

Видим, что в гене очень много интронов, экзоны узкие, как будто сгруппированы в ~7 групп.

Из треков консервативности видно, что четкие консервативные пики соответствуют экзонам. В начале и в конце есть голубая пунктирная линия. Она соответствует двум длинным интронам в начале и в конце гена (Рис. 19.)