Практикум 11-13

Кушнарева Дарья 202 П3

Данные: Хромосома 7, SRR10720402, ENCFF363HPJ

Работа проводилась в директории **/mnt/scratch/NGS/d.kushnareva**
Файлы:
Homo_sapiens.GRCh38.dna.chromosome.7.fa - последовательность седьмой хромосомы генома человека.
SRR10720402_1.fastq.gz - “прямые” чтения
SRR10720402_2.fastq.gz - “обратные” чтения
ENCFF363HPJ.fastq.gz - одноконцевые чтения секвенирования рнк человека

Общие задачи:
Поиск и аннотация вариантов одного человека по данным экзомного секвенирования на примере одной хромосомы (практикумы 11-12)
Построение экспрессионного профиля на основании данных секвенирования РНК (пркатикум 13)

Практикум 11

Часть 1

Задача практикума: подготовить необходимые файлы (парно-концевые прочтения и последовательность референсного генома), изучить качество предложенных чтений, проиндексировать референс.

1) **Подготовка референса**

Получение референса

```
mkdir reference
cp /mnt/scratch/NGS/DATA/hg38/Homo_sapiens.GRCh38.dna.chromosome.7.fa
reference/Homo_sapiens.GRCh38.dna.chromosome.7.fa
```

Референс и индексация будут храниться в директории **reference/**.
Проиндексируем с помощью hisat2

```
hisat2-build Homo_sapiens.GRCh38.dna.chromosome.7.fa chr7_
```

chr7_ - это префикс, с него будут начинаться полученные файлы (.ht2)
В итоге получено 8 файлов.

Индексация samtools

Проиндексируем с помощью samtools, чтобы программа могла использовать индексацию подходящую для нее.

```
samtools faidx Homo_sapiens.GRCh38.dna.chromosome.7.fa
```

В итоге получен файл Homo_sapiens.GRCh38.dna.chromosome.7.fa.fai
Его содержимое:

```
7	159345973	56	60	61
```

7 - Имя хромосомы
159345973 - Длина в нуклеотидах
56 - Смещение начала последовательности в исходном FASTA-файле (в байтах)
60 - Количество оснований в каждой строке
61 - Количество байт в каждой строке

2) **Чтения ДНК**

Описание образца

a) Образец - SRR10720402
b) Ссылка https://www.ncbi.nlm.nih.gov/sra/?term=SRR10720402
c) Прибор - Illumina Genome Analyzer IIx
d) Организм - *Homo sapiens*
e) Стратегия секвенирования - Exome (Whole Exome Sequencing)
f) Чтения - парноконцевые (PAIRED)
g) Сколько чтений ожидается - 28966798

Проверка качества исходных чтений

Все файлы связанные с чтениями будут храниться в директории **reeds/**

```
mkdir reeds
cp /mnt/scratch/NGS/DATA/dna_reads/SRR10720402_1.fastq.gz reeds/SRR10720402_1.fastq.gz
cp /mnt/scratch/NGS/DATA/dna_reads/SRR10720402_2.fastq.gz reeds/SRR10720402_2.fastq.gz
```

Проанализируем данные чтения с помощью fastqc:

```
fastqc SRR10720402_1.fastq.gz
fastqc SRR10720402_2.fastq.gz
```

Благодаря анализу было получено 2 файла: SRR10720402_1_fastqc.html и SRR10720402_2_fastqc.html. В них можно найти информацию о количестве пар чтений, а именно 28966798 для прямых чтений и столько же для обратных. Информация о качестве этих чтений представлена в следующих графиках:

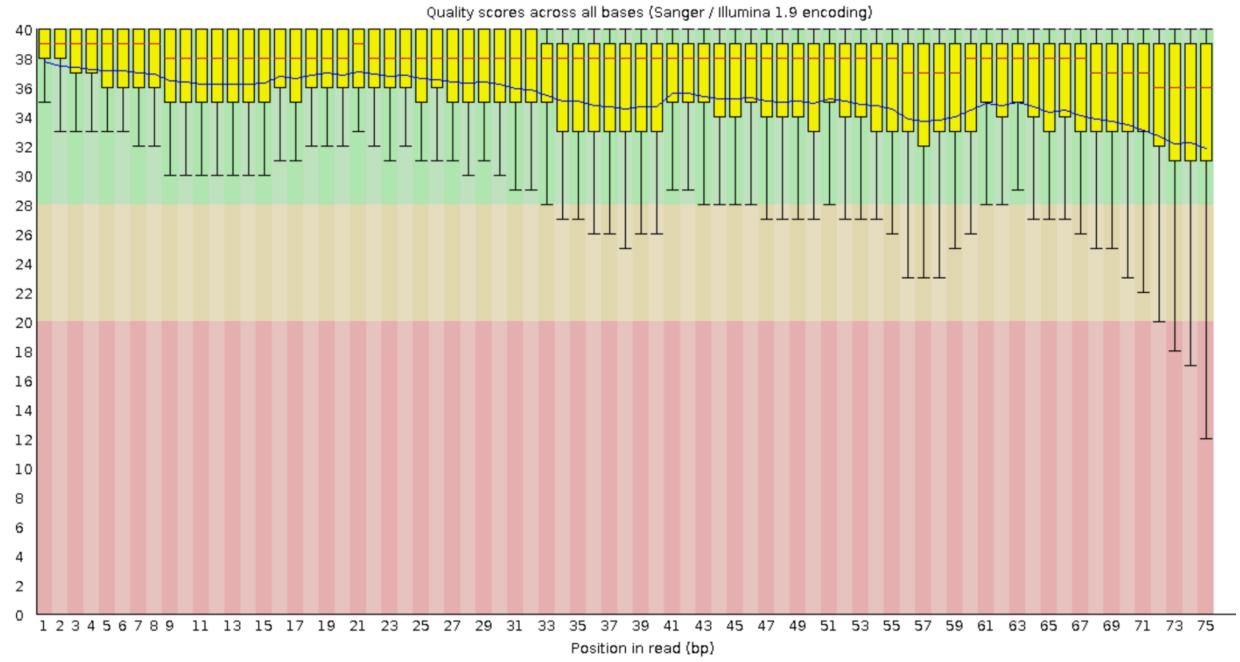

Рис.1. Качество оснований в прямых чтениях

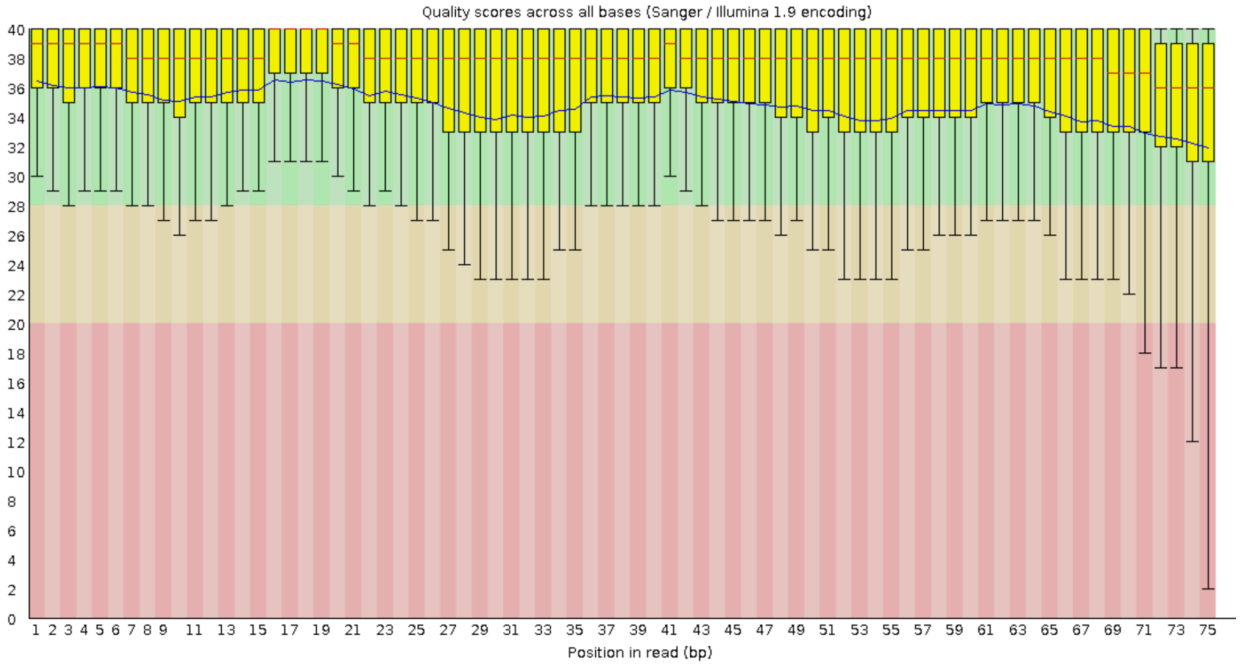

Рис.2. Качество оснований в обратных чтениях

Видно что качество чтений довольно хорошее

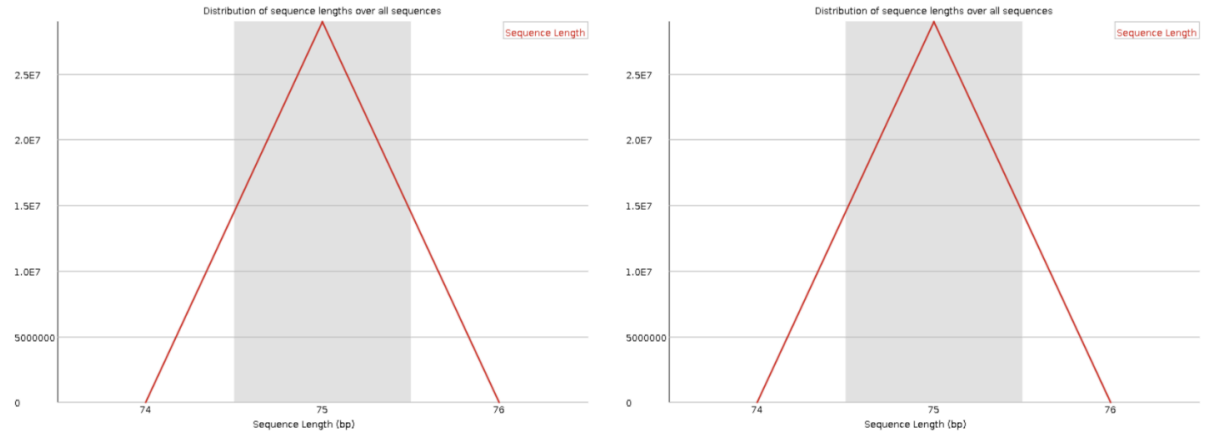

Рис. 3. Распределения длин чтений прямых (справа) и обратных (слева)

Все чтения имеют длину 75 нуклеотидов.

Фильтрация чтений

Далее чтения нужно отфильтровать, чтобы избавиться от остатков адаптеров и некачественных чтений. Это было сделано следующей командой:

```
TrimmomaticPE -phred33 SRR10720402_1.fastq.gz SRR10720402_2.fastq.gz
SRR10720402_1_paired.fastq.gz SRR10720402_1_unpaired.fastq.gz  SRR10720402_2_paired.fastq.gz
SRR10720402_2_unpaired.fastq.gz TRAILING:20 MINLEN:40
```

Опция TRAILING:20 удаляет нуклеотиды с конца с качеством ниже 20

Опция MINLEN:40 удаляет чтения с длиной меньше 40

В результате было получено 4 файла, для прямых чтений у которых сохранилась пара (SRR10720402_1_paired.fastq.gz), для прямых чтений у которых пара была откинута(fastqc SRR10720402_1_unpaired.fastq.gz), и аналогично для обратных чтений.

Проверка качества триммированных чтений

После триммирования нужно снова проанализировать чтения:

```
fastqc SRR10720402_1_paired.fastq.gz
fastqc SRR10720402_1_unpaired.fastq.gz
fastqc SRR10720402_2_paired.fastq.gz
fastqc SRR10720402_2_unpaired.fastq.gz
```

После фильтрации осталось 27509530 пар чтений (94,97% от изначального количества).

Если сравнивать качество парных и непарных чтений, то можно увидеть что парные будут более хорошего качества, а также парные чтения будут лучше чем не триммированные. А еще обратные в целом менее хорошего качества.

Длина чтений изменилась, теперь не все чтения имеют длину 75, но очень мало более коротких.

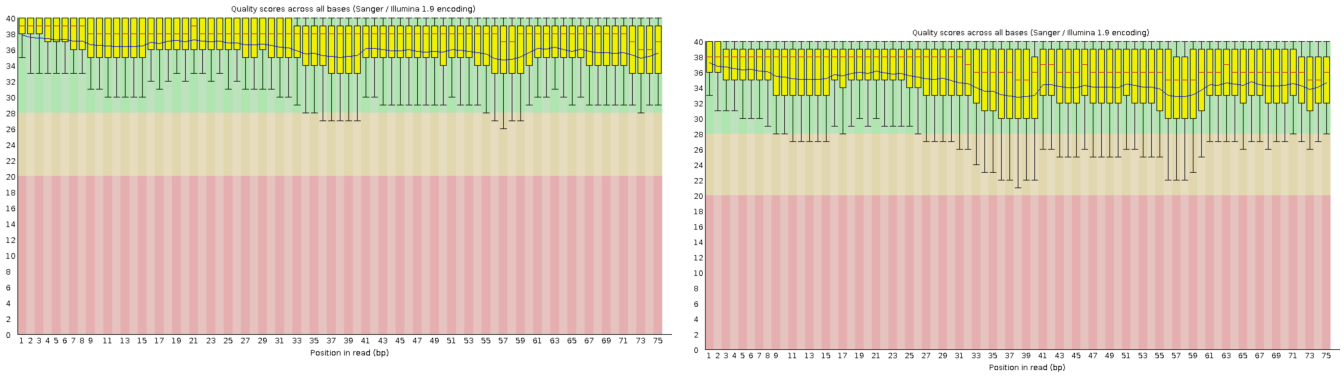

Рис 4. Качество оснований в прямых парных чтениях и в прямых непарных чтениях

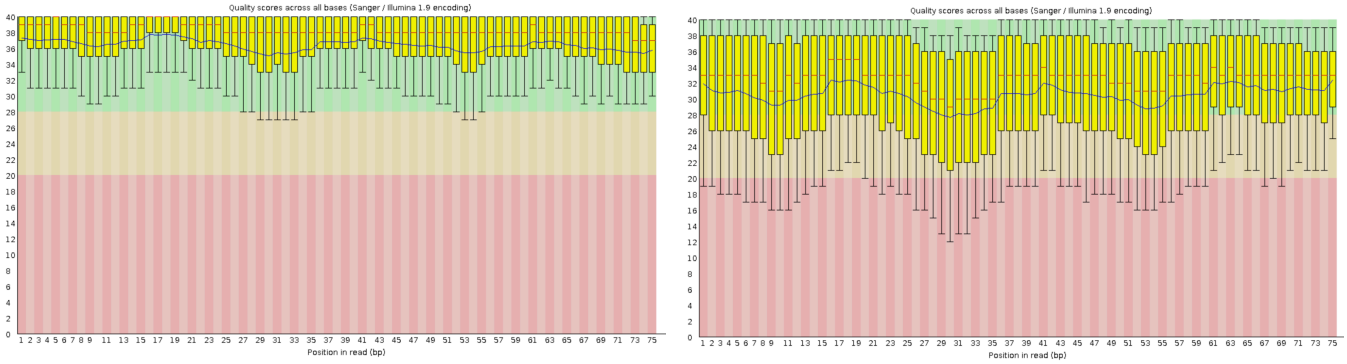

Рис 5. Качество оснований в обратных парных чтениях и в обратных непарных чтениях

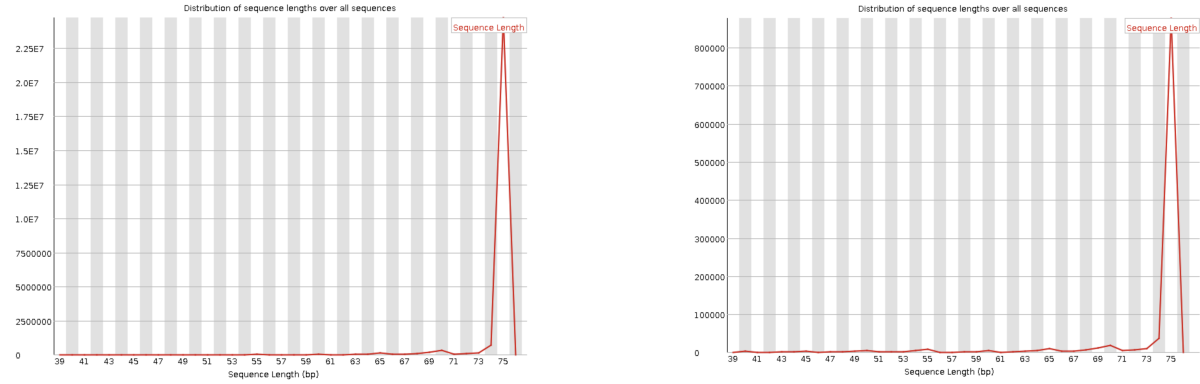

Рис 6. Распределения длин чтений прямых парных (справа) и прямых непарных (слева)

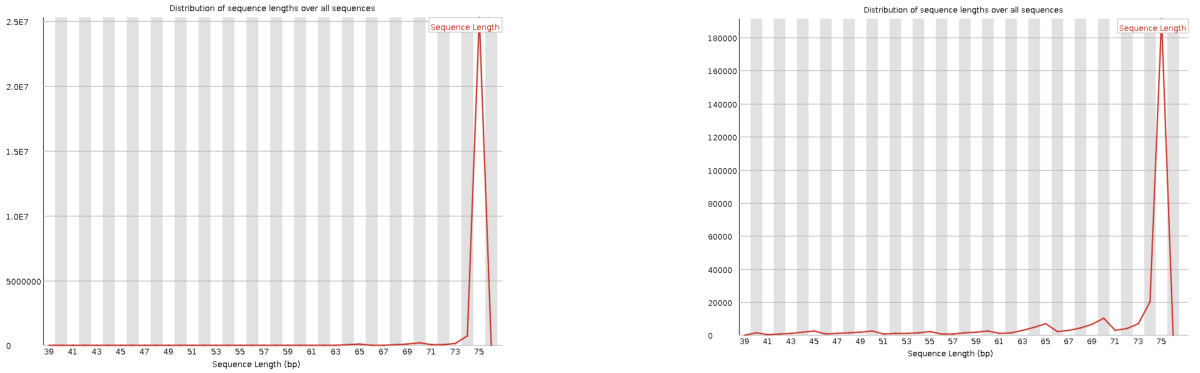

Рис 7. Распределения длин чтений обратных парных (справа) и обратных непарных (слева)

Часть 2

Задача практикума: картировать чтения хорошего качества на референсный геном и отобрать только такие чтения, которые удалось картировать в корректных парах

Картирование чтений на референсный геном

Все логи будут храниться в папке **log/**, а файлы связанные с картированием чтений в папке **map/**.

Картируем парные чтения на референс с помощью hisat2.

```
mkdir log
mkdir map
hisat2 -x reference/chr7_ -1 reeds/SRR10720402_1_paired.fastq.gz -2
reeds/SRR10720402_2_paired.fastq.gz -p 4 --no-spliced-alignment -S map/chr7_paired_map.sam 2>
log/hisat2_chr7_paired_map.log
```

Параметры запуска:

- -x путь к индексированному референсу
- -1 указание на прямое чтение (триммированое SRR10720402_1_paired.fastq.gz)
- -2 указание на обратное чтение (триммированое SRR10720402_2_paired.fastq.gz)
- -p указание на количество потоков (для ускорения работы)
- --no-spliced-alignment запрещает поиск сплайсинга
- -S имя выходного SAM-файла
- 2> log/* перенаправление stderr в лог-файл

Конвертация sam в bam

Описание sam/bam файл

.sam файлы не бинарные, благодаря этому они читабельны, но из-за этого весят сильно больше

```
ls -lh map/chr7_paired_map.sam
> -rw-r--r--. 1 d.kushnareva year-23 11G Mar 30 00:43 map/chr7_paired_map.sam
```

Он весит целых 11G. Поэтому для дальнейшей работы конвертируем его в .bam файл, а .sam удалим:

```
samtools sort -o map/chr7_paired_map.bam map/chr7_paired_map.sam
rm map/chr7_paired_map.sam
```

Проиндексируем получившийся bam файл

```
samtools index chr7_paired_map.bam
```

Анализ bam файла

Для анализа использовалась команда:

```
samtools flagstat map/chr7_paired_map.bam > chr7_paired_map_flagstat.txt
```

Содержимое файла chr7_paired_map_flagstat.txt:

```
55939226 + 0 in total (QC-passed reads + QC-failed reads)
55019060 + 0 primary
920166 + 0 secondary
0 + 0 supplementary
```

```
0 + 0 duplicates
0 + 0 primary duplicates
4438017 + 0 mapped (7.93% : N/A)
3517851 + 0 primary mapped (6.39% : N/A)
55019060 + 0 paired in sequencing
27509530 + 0 read1
27509530 + 0 read2
3038836 + 0 properly paired (5.52% : N/A)
3108542 + 0 with itself and mate mapped
409309 + 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)
```

Поле in total - это общее количество выравниваний (55939226)

primary - сколько поступило на картирование (55019060)

properly paired - количество чтений картированных на референс в корректных парах (3038836)
Это составляет 5.52% от поступивших на картирование, потому что мы картируем только на 7 хромосому.

Получение чтений, картированных на хромосому

Чтобы получить чтения картированные только на 7 хромосому, а потом проанализировать их использовались команды:

```
samtools view -h -bS map/chr7_paired_map.bam 7 > map/on_chr7_paired_map.bam
samtools flagstat map/on_chr7_paired_map.bam > on_chr7_paired_map_flagstat.txt
```

Параметры samtools view:

- -h добавляет метаданные в выдачу
- -bS файл подающийся на вход (.bam)
- 7 имя хромосомы
- >mapped_chr7.bam выходной файл

Содержимое файла on_chr7_paired_map_flagstat.txt

```
4847326 + 0 in total (QC-passed reads + QC-failed reads)
3927160 + 0 primary
920166 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
4438017 + 0 mapped (91.56% : N/A)
3517851 + 0 primary mapped (89.58% : N/A)
3927160 + 0 paired in sequencing
1963580 + 0 read1
1963580 + 0 read2
3038836 + 0 properly paired (77.38% : N/A)
3108542 + 0 with itself and mate mapped
409309 + 0 singletons (10.42% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Теперь уменьшилось общее количество выравниваний (4847326), остались только выровненные риды и синглтоны и из-за этого вырос процент картирования.

Получение только правильно картированных пар чтений

Для получения только правильно картированных пар чтений, и анализа полученных чтений использовались команды:

```
samtools view -f 2 -bS map/on_chr7_paired_map.bam > map/on_chr7_properly_paired_map.bam
samtools flagstat map/on_chr7_properly_paired_map.bam > on_chr7_properly_paired_map_flagstat.txt
```

Параметр -f 2 означает что нужно брать только правильно картированные чтения.

Содержимое файла on_chr7_properly_paired_map_flagstat.txt:

```
3701138 + 0 in total (QC-passed reads + QC-failed reads)
3038836 + 0 primary
662302 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3701138 + 0 mapped (100.00% : N/A)
3038836 + 0 primary mapped (100.00% : N/A)
3038836 + 0 paired in sequencing
1519418 + 0 read1
1519418 + 0 read2
3038836 + 0 properly paired (100.00% : N/A)
3038836 + 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)
```

В отличии от предыдущего файла теперь нет синглтонов и общее количество чтений 3701138 и они все правильно спаренные.

Далее проиндексируем:

```
samtools index on_chr7_properly_paired_map.bam
```

И получим файл on_chr7_properly_paired_map.bam.bai.

Получение чтений, картированных только в границы экзома

Для этого воспользуемся командами:

```
bedtools intersect -a on_chr7_properly_paired_map.bam -b
/mnt/scratch/NGS/DATA/genes/seqcap_hg38.bed -wa >
map/on_exome_chr7_properly_paired_map.bam
samtools flagstat on_exome_chr7_properly_paired_map.bam >
on_exome_chr7_properly_paired_map.bam_flagstat.txt
```

Параметры:

-a входной файл

-b входной файл

-wa чтения из файла a, пересекающеюся с файлами b

Содержимое файла on_exome_chr7_properly_paired_map.bam_flagstat.txt:

```
2321420 + 0 in total (QC-passed reads + QC-failed reads)
1878448 + 0 primary
442972 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2321420 + 0 mapped (100.00% : N/A)
1878448 + 0 primary mapped (100.00% : N/A)
1878448 + 0 paired in sequencing
937773 + 0 read1
940675 + 0 read2
1878448 + 0 properly paired (100.00% : N/A)
1878448 + 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)
```

В границы экзона попало 2321420 чтений (что составляет 76%).

Практикум 12

Задача практикума: получить список вариантов на основании полученного ранее bam файла и аннотировать их средствами VEP

Получение вариантов

Все файлы связанные с генетическими вариантами будут храниться в директории **var/**.

```
mkdir var
bcftools mpileup -f reference/Homo_sapiens.GRCh38.dna.chromosome.7.fa
map/on_chr7_properly_paired_map.bam | bcftools call -mv -o var/chr7_variants.vcf
bcftools stats var/chr7_variants.vcf > chr7_variants_stats.txt
```

Параметры:

-f референсный геном

-mv выводит мультиаллельные и редкие варианты

-o выходного файл

Полученный файл имеет следующее устройство:

Шапка - строки начинающиеся с #, содержат информацию о данных
Таблица с информацией о вариантах (подписи столбцов - последняя строка шапки)
CHROM - номер хромосомы
POS - позиция вакрианта
ID - обычно “.”, но может быть любая информация о варианте
REF - остаток в референсе
ALT - альтернативный вариант аллеля
QUAL - качество
INFO - характеристики варианта
FORMAT - параметры варианта

Также можем получить информацию о том что всего вариантов - 66383, 65502 нуклеотидных замен и 881 вставок и делеций.

Фильтрация вариантов

Для фильтрации генетических вариантов и их анализа использовались следующие команды:

```
bcftools filter -i'QUAL>30 && DP>50' var/chr7_variants.vcf -o var/chr7_variants_filtered.vcf
bcftools stats var/chr7_variants_filtered.vcf > chr7_variants_filtered_stats.txt
```

Варианты были отфильтрованы так чтобы качество было больше 30 (QUAL>30) и глубина покрытия была более 50 чтений (DP>50)
Из полученной статистики мы можем узнать что вариантов всего 1628 (что составляет 2.45% от исходных), в них однонуклеотидных замен 1560 (2.38%), а в вставок и делеций - 68 (7.72%).

Аннотация вариантов

С помощью сервиса VEP была сделана аннотация генетических вариантов (файл chr7_variants_filtered.vcf). Были обработаны 1628 варианта, то есть ни один не отфильтровал сервер, а значит все анализированные варианты были аннотированны. 1253 варианта уже были в базах данных, а 375, что составляет 23%, были новыми. Варианты находятся на 631 генах, 6522 транскриптонах и 15 регулирующих зонах. Это довольно хорошее покрытие, особенно любопытны регулирующие зоны.

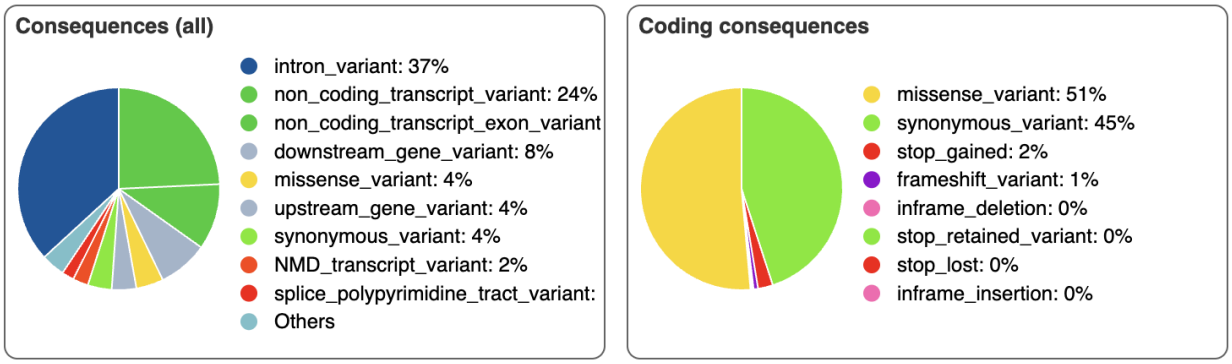

Рис 8. Статистические данные анализа вариантов на сервере VEP

На Рис 8. Видно что большая часть вариантов не кодирующие, а среди кодирующих большая часть (51%) миссенс-мутации, 45% синонимичные мутации. Миссенс могут быть интересны, но вот синонимичные мало интересны, так как не влияют практически. Еще любопытно посмотреть на нонсенс мутации (2%) и сдвигающие рамку считывания мутации (1%) потому что их хоть и мало, но они скорее всего приведут к наибольшим последствиям. Если посмотреть на IMPACT HIGH варианты (фильтр Impact is HIGH), то мы увидим 130 вариантов, большая часть из которых как раз связаны с нонсенс-мутациями.

Практикум 13

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

Описание образца

a) ID образца - ENCFF363HPJ
b) Ссылка https://www.encodeproject.org/experiments/ENCSR129VBC/
c) Организм и ткань - Homo sapiens astrocyte (человеческие астроциты)

d) Стратегия секвенирования - RNA-seq (total RNA-seq)
e) Чтения одноконцевые
f) Цепь-специфичность - Strand-specific (reverse)

Проверка качества исходных чтений

Все связанное с РНК чтениями будет находиться в папке **reedRNA/**

Проанализируем РНК чтения с помощью следующей команды:

```
mkdir reedRNA
cp /mnt/scratch/NGS/DATA/rna_reads/ENCFF363HPJ.fastq.gz reedRNA/ENCFF363HPJ.fastq.gz
fastqc ENCFF363HPJ.fastq.gz
```

Всего чтений 23387479, чтения имеют хорошее качество и все они имеют длину в 100 нуклеотидов

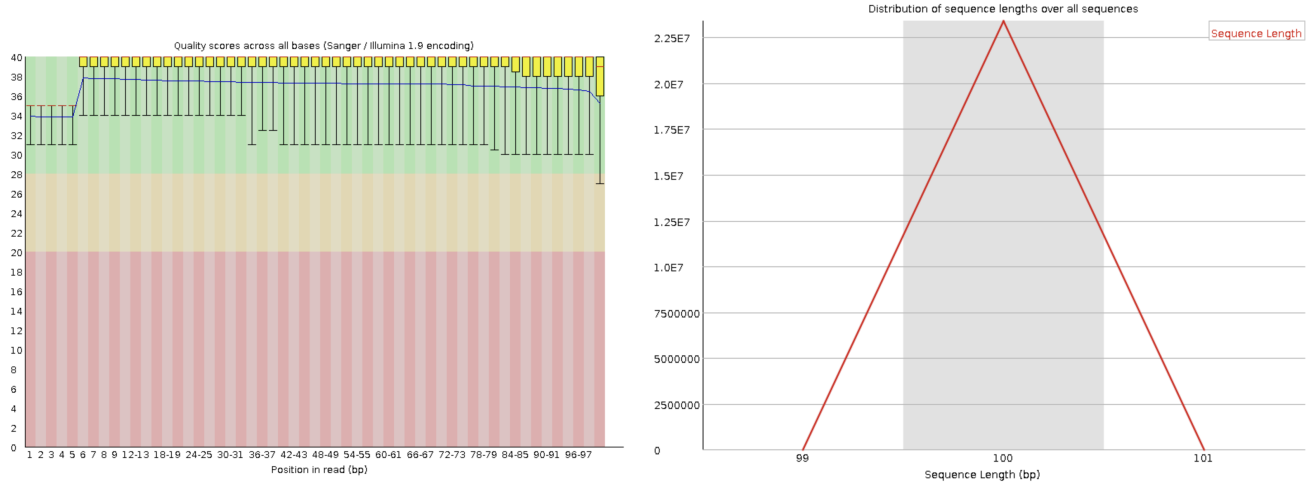

Рис.9. Качество оснований в РНК чтениях и Распределение длин чтений

Интересно как отличается качество у первых пяти нуклеотидов, возможно это связано с адаптерами, но даже так их качество довольно хорошее.

Картирование чтений на референс

Все связанное с картированием РНК чтений будет находиться в папке **mapRNA/**

Картируем РНК чтения с помощью следующей команды:

```
mkdir mapRNA
hisat2 -x reference/chr7_ -k 3 -U reedRNA/ENCFF363HPJ.fastq.gz -p 4 -S
mapRNA/mapped_reads.sam 2> log/hisat2_chr7_RNA_map.log
```

Параметры:

-k 3 необходимо находить до 3 лучших выравниваний для каждого чтения

-U входной файл с чтениями

Перевод .sam файл в сортированный .bam, индексация, удаление .sam, отбор чтений легших только на 7 хромосому проводилось с помощью следующих команд:

```
samtools sort -o mapRNA/mapped_reads.bam mapRNA/mapped_reads.sam
ls -lh mapRNA/mapped_reads.sam
>-rw-r--r--. 1 d.kushnareva 2023 5.9G Mar 30 19:45 mapRNA/mapped_reads.sam
rm mapRNA/mapped_reads.sam
samtools index mapRNA/mapped_reads.bam
samtools view -h -bS mapRNA/mapped_reads.bam 7 > mapRNA/chr7_reads_map.bam
samtools flagstat mapRNA/chr7_reads_map.bam > chr7_reads_map_flagstat.txt
samtools index mapRNA/chr7_reads_map.bam
```

Содержимое файла chr7_reads_map_flagstat.txt:

```
1350754 + 0 in total (QC-passed reads + QC-failed reads)
1276767 + 0 primary
73987 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1350754 + 0 mapped (100.00% : N/A)
1276767 + 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)
```

Всего на 7 хромосому было картированно 1350754 чтений.

Поиск экспрессирующихся генов

Для начала необходимо проанализировать .gtf файл. В нем 9 столбцов (табулированная таблица): имя хромосомы, источник аннотации, тип feature (например: “Gene”, “Variation”, “Similarity”), начальная позиция feature, конечная позиция, оценка («.» если значение отсутствует), цепь (+/-), рамка считывания,дополнительные информация (пары «ключ;значение»)

Количество картированных на гены чтений были посчитаны с помощью команды:

Теперь необходимо найти количество картированных на гены:

```
cp /mnt/scratch/NGS/DATA/genes/Homo_sapiens.GRCh38.110.chr.gtf
reference/Homo_sapiens.GRCh38.110.chr.gtf
htseq-count -f bam -s reverse -m intersection-strict -t exon mapRNA/chr7_reads_map.bam
reference/Homo_sapiens.GRCh38.110.chr.gtf  > chr7_reads_map_counts.txt
```

Параметры:

- -f - формат выходного файла
- -s - указание цепи (no - нет указания цепи)
- -m - объединение перекрывающихся чтений
- -t - тип элементов с которыми считается перекрывание

Если посмотреть выходной файл то можно узнать, что 942929 чтений попало в границы генов, а 284544 чтений попало мимо границ генов.

Аннотация высоко экспрессируемых генов

Теперь посмотрим на самые экспрессирующиеся гены, чтобы найти топ 10 генов по экспрессии воспользуемся командой:

```
grep -v "^__" chr7_reads_map_counts.txt | sort -k2,2nr | head -10 |less
```

Топ 10 генов по экспрессии:

```
ENSG00000164692 247054
ENSG00000075624 82341
ENSG00000233476 58898
ENSG00000122786 33259
ENSG00000106366 23325
ENSG00000128591 15860
ENSG00000128595 14044
ENSG00000146674 14028
ENSG00000091136 10422
ENSG00000075618 9410
```

Для аннотации я выбрала ген ENSG00000106366 или LAMB1

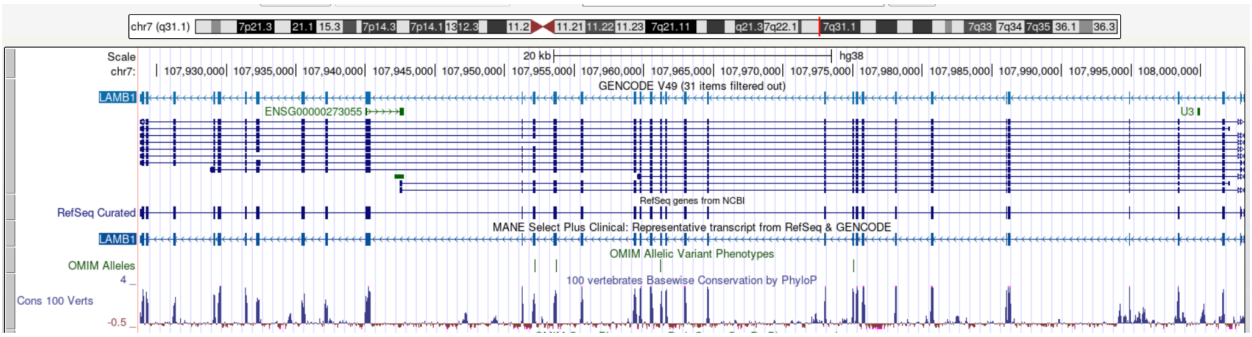

Рис.10. Визуализация гена ENSG00000106366 с помощью геномного браузера

Из Рис.10 видно, что у гена довольно крупные интроны, его устройство довольно сложное. Если смотреть на консервативность то в основном консервативность заметно выше у экзонов, нежели у интронов, но в конце гена есть несколько участков внутри интронов у которых тоже высокая консервативность, возможно это связанные с регуляцией участки.

LAMB1 - бета-1 субъединица ламинина, белка формирующего структурный каркас базальных мембран, обеспечивающего адгезию (прилипание), миграцию и дифференцировку клеток.

Скрипт для 11-12 практикума

```
#!/bin/bash

mkdir reference
cp /mnt/scratch/NGS/DATA/hg38/Homo_sapiens.GRCh38.dna.chromosome.7.fa
reference/Homo_sapiens.GRCh38.dna.chromosome.7.fa
mkdir reeds
cp /mnt/scratch/NGS/DATA/dna_reads/SRR10720402_1.fastq.gz reeds/SRR10720402_1.fastq.gz
cp /mnt/scratch/NGS/DATA/dna_reads/SRR10720402_2.fastq.gz reeds/SRR10720402_2.fastq.gz
mkdir log
mkdir map

hisat2-build reference/Homo_sapiens.GRCh38.dna.chromosome.7.fa reference/chr7_

TrimmomaticPE -phred33 SRR10720402_1.fastq.gz SRR10720402_2.fastq.gz
SRR10720402_1_paired.fastq.gz SRR10720402_1_unpaired.fastq.gz  SRR10720402_2_paired.fastq.gz
SRR10720402_2_unpaired.fastq.gz TRAILING:20 MINLEN:40

hisat2 -x reference/chr7_ -1 reeds/SRR10720402_1_paired.fastq.gz -2
reeds/SRR10720402_2_paired.fastq.gz -p 4 --no-spliced-alignment -S map/chr7_paired_map.sam 2>
log/hisat2_chr7_paired_map.log

samtools sort -o map/chr7_paired_map.bam map/chr7_paired_map.sam
rm map/chr7_paired_map.sam

samtools index map/chr7_paired_map.bam

samtools view -h -bS map/chr7_paired_map.bam 7 > map/on_chr7_paired_map.bam

samtools view -f 2 -bS map/on_chr7_paired_map.bam > map/on_chr7_properly_paired_map.bam

samtools index map/on_chr7_properly_paired_map.bam

bedtools intersect -a on_chr7_properly_paired_map.bam -b
/mnt/scratch/NGS/DATA/genes/seqcap_hg38.bed -wa >
map/on_exome_chr7_properly_paired_map.bam

mkdir var
bcftools mpileup -f reference/Homo_sapiens.GRCh38.dna.chromosome.7.fa
map/on_chr7_properly_paired_map.bam | bcftools call -mv -o var/chr7_variants.vcf

bcftools filter -i'QUAL>30 && DP>50' var/chr7_variants.vcf -o var/chr7_variants_filtered.vcf

echo “chr7_variants_filtered.vcf ”
```