Задание 11

Подготовка референса

1)Получение референса

Создала папку для последовательности генома и индекса к программе для картирования с помощью hisat2 и скопировала в нее файл с вашей хромосомой. /mnt/scratch/NGS/anastasia.l/pr_11 2 папки genes и index

2)Индексация для hisat2

Проиндексировала референсный геном. Будет создавать индексную базу данных для выравнивания ридов на геномную информацию человека, содержащуюся в файле Homo_sapiens.GRCh38.dna.chromosome.4.fa. Выходные файлы базы данных будут иметь префикс prefix

Команда:
hisat2-build Homo_sapiens.GRCh38.dna.chromosome.4.fa prefix

Выдача: 8 файлов prefix.1-8.ht2

3)Индексация samtools.

Этот индекс позволяет эффективно находить и получать доступ к конкретным участкам генома, что является важным для многих биологических и генетических исследований.
Команда: samtools faidx Homo_sapiens.GRCh38.dna.chromosome.4.fa

Из полученного Homo_sapiens.GRCh38.dna.chromosome.4.fa.fai узнаем точное имя своей хромосомы и длину вашей хромосомы в нуклеотидах:

```
4	190214555	56	60	61
```

4-номер хромосомы/точное имя хромосомы
190214555-длина хромосомы в нуклеотидах
56: номер байта, с которого начинается сама нуклеотидная последовательность в fasta
60: число нуклеотидов в каждой строке в нуклеотидах
61: число байтов в каждой из строк

Чтения ДНК

В базе NCBI (https://www.ncbi.nlm.nih.gov/) в разделе SRA ID нашего образца.
1)SRR ID образца ДНК-чтений: SRR10720407

2)Ссылку на информацию об образце из NCBI:
https://www.ncbi.nlm.nih.gov/sra/?term=SRR10720407
3)Прибор для секвенирования: ILLUMINA (Illumina Genome Analyzer IIx)
4)Организм: Homo sapiens
5)Стратегию секвенирования (полногеномное, экзомное, таргетная панель):
Whole-exome
6)Парноконцевые или одноконцевые чтения: PAIRED
7)Сколько чтений ожидается (spots) : 38,530,707

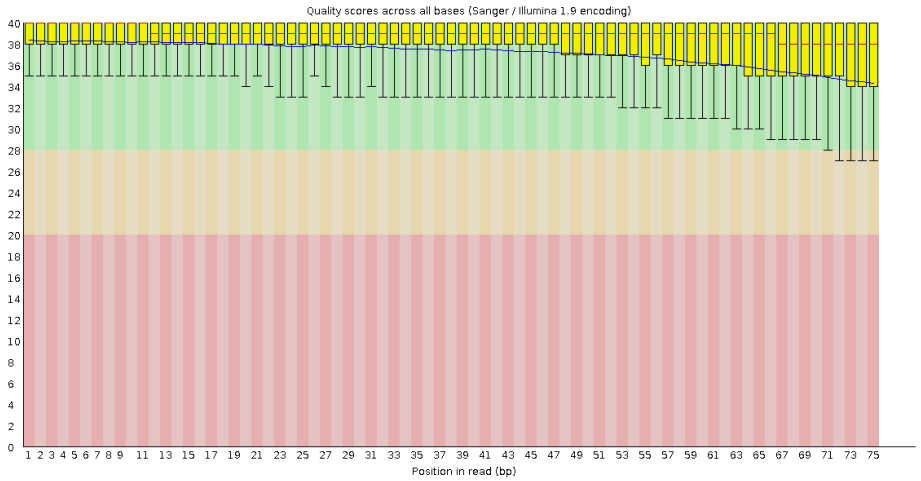
Проверка качества исходных чтений

Анализируем качество исходных чтений с помощью программы fastqc.
Запустить fastqc можно с помощью команды fastqc file.fastq.gz
Файлы папке quaity_check

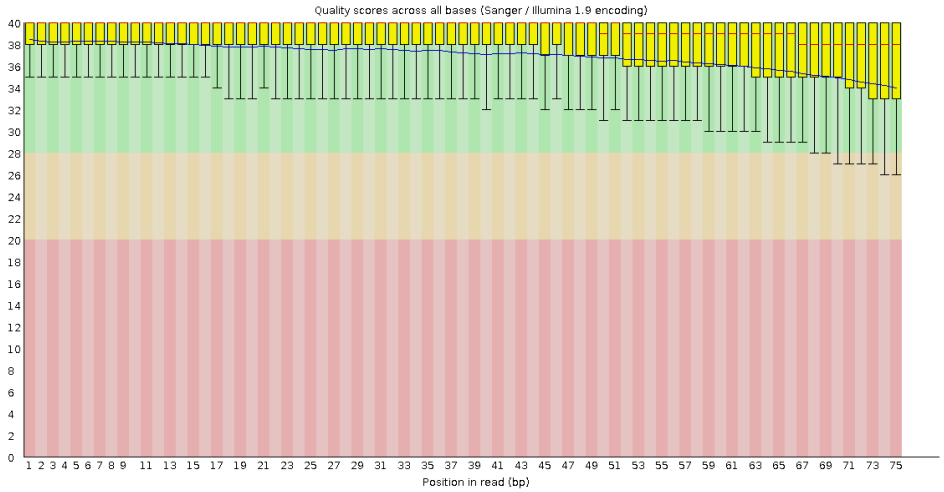
Рассматриваем файлы : SRR10720407_2_fastqc.zip и SRR10720407_1_fastqc.zip

1)Какое количество пар чтений получилось: 38530707
2)Совпадает ли количество чтений у “прямых” чтений и “обратных” чтений: да, совпадает
3)Краткий комментарий качества пар чтений по результатам fastqc: качество чтений хорошее, только к концу становится чуть меньше 30 (ориентировалась относительно медиан красных и синих средних значений)

Прямые:

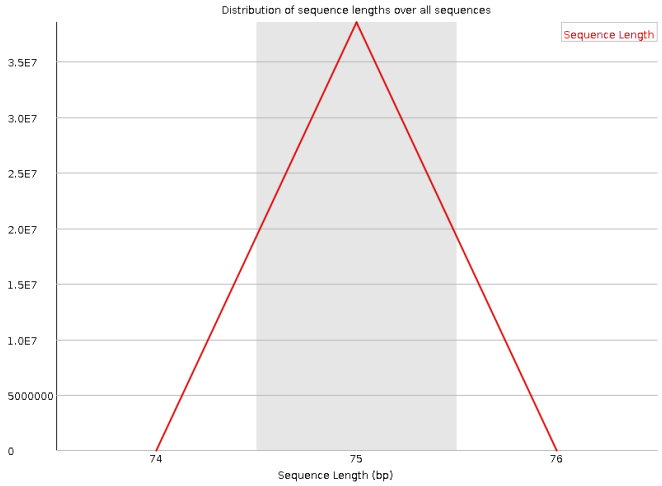


Обратные:

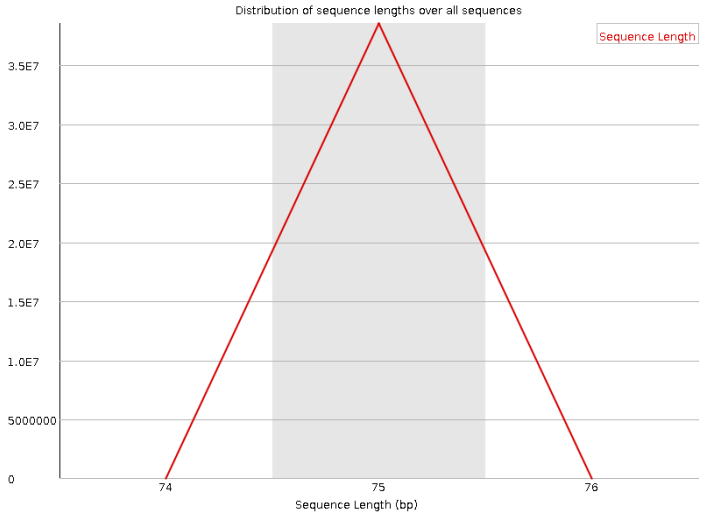


4)Краткий комментарий о длине ваших чтений по результатам fastqc- длина чтения 75

Прямые:

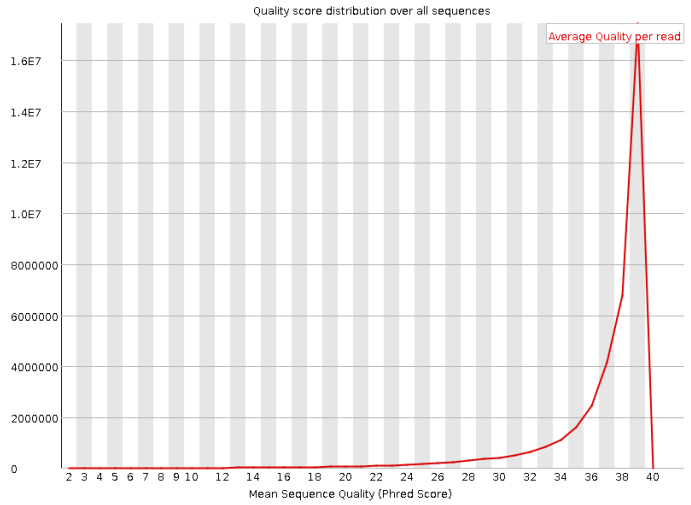


Обратные:

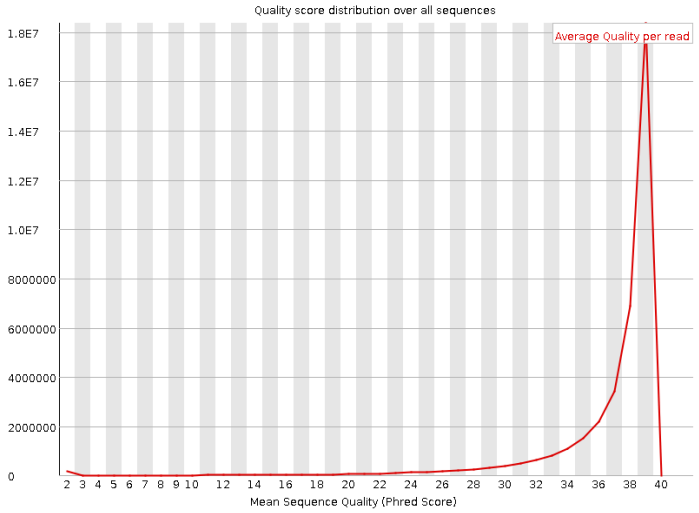


5) Краткий комментарий о качестве чтений: можем заключить , что здесь как и на первых рис в большинстве случаем качество больше 30

Прямые:



Обратные:



Фильтрация чтений

Используем команду TrimmomaticPE для обрезки и фильтрации последовательностей чтения (ридов) в парных файлах fastq.gz

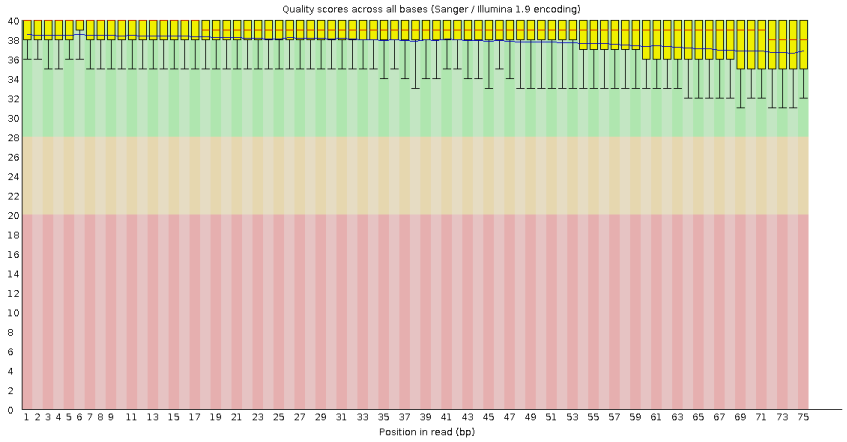
Команда :TrimmomaticPE -phred33 SRR10720407_1.fastq.gz SRR10720407_2.fastq.gz trimmed_forward_paired.fastq.gz trimmed_reverse_paired.fastq.gz trimmed_forward_unpaired.fastq.gz trimmed_reverse_unpaired.fastq.gz TRAILING:20 MINLEN:50

Аргументы:
- **phred33**:указывает, что входные файлы в формате fastq используют кодировку качества Phred 33. Такая кодировка используется в большинстве современных прочтений Illumina
- **SRR10720407_1.fastq.gz**:первый файл
- **SRR10720407_2.fastq.gz**: второй файл
- **trimmed_forward/reverse_paired/unpaired.fastq.gz**: выходной файл, в котором содержатся обрезанные и отфильтрованные прочтения прямой/обратной цепей парных и непарных прочтений
- **TRAILING:20**: если качество последних нуклеотидов в риде оказывается ниже 20, они будут удалены.
- **MINLEN:50:** Этот параметр указывает, что после всех операций обрезки и фильтрации, прочтения, имеющие длину меньше 50 нуклеотидов, будут удалены.

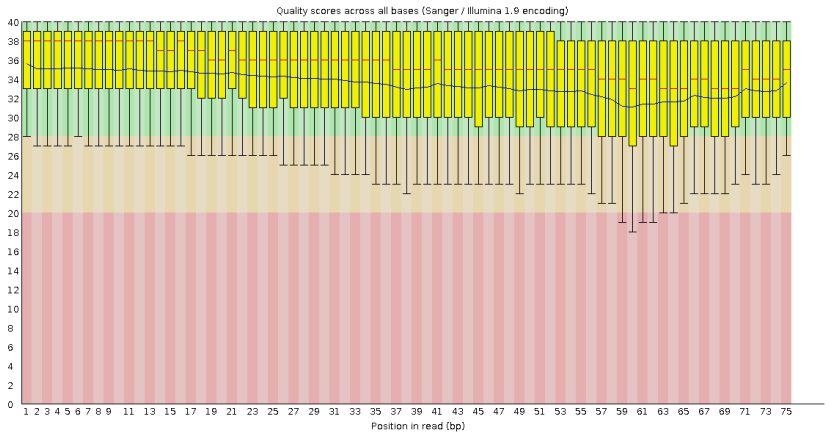
Проверка качества триммированных чтений

Анализируем качество чтений после обработки программой Trimmomatic с помощью программы fastQC.
1)Какое количество пар чтений осталось (paired) в штуках:37276728
2)Какой процент пар чтений остался (paired) (процент от исходного количества пар чтений):86.4
3)Краткий комментарий о сравнении качества чтений после триммирования: paired vs unpaired: качество непарных чтений заметно сильно хуже, чем качество парных

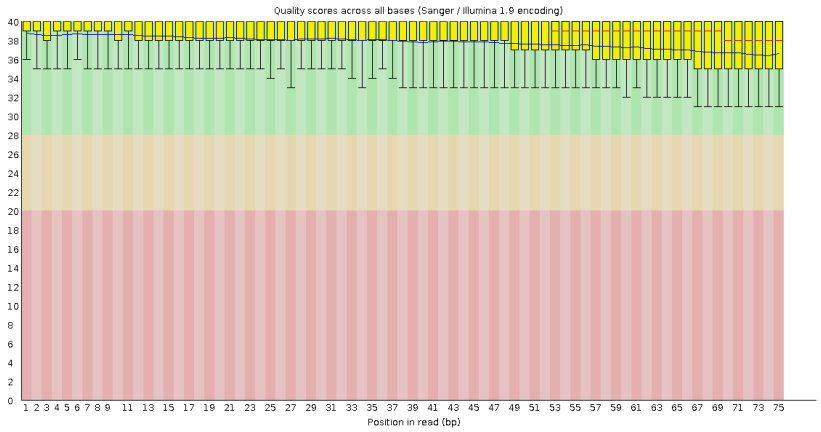
Прямые парные:



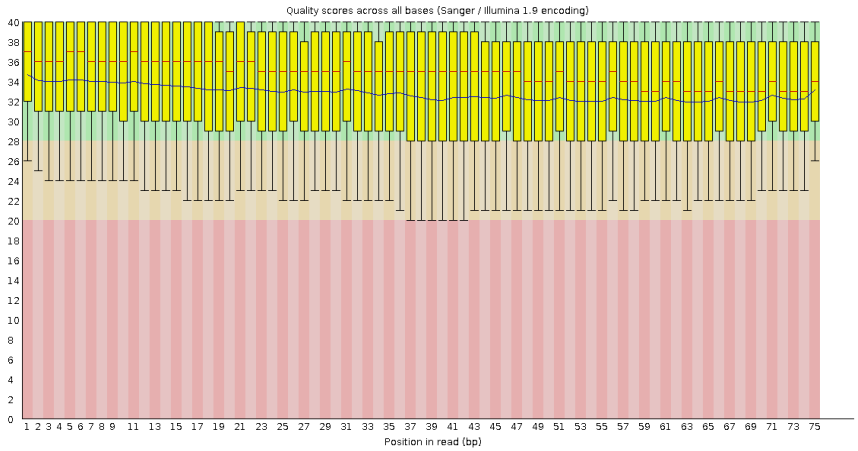
Прямые непарные:



Обратные парные:



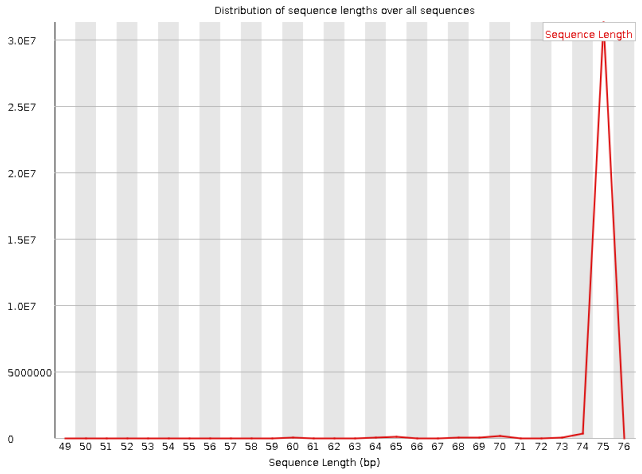
Обратные непарные:



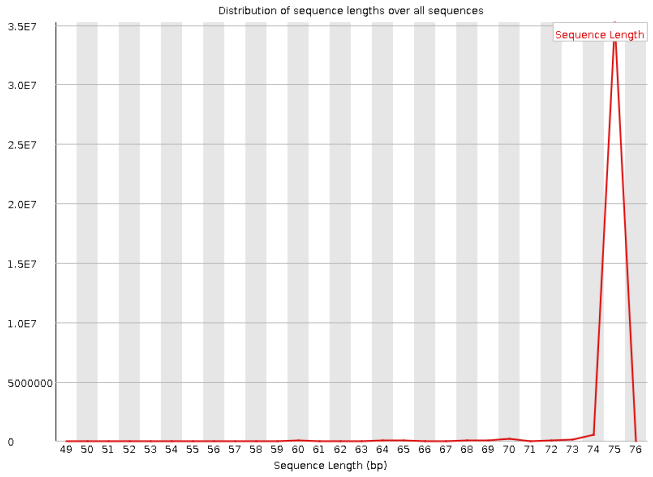
4)Краткий комментарий о сравнении качества чтений до и после триммирования (только paired) : качество особенно к концу значительно стало лучше

5)Как изменилась длина чтений после триммирования? также 75, а вот у непарных длина совсем изменилась, так как появились значительной величины пики на графиках

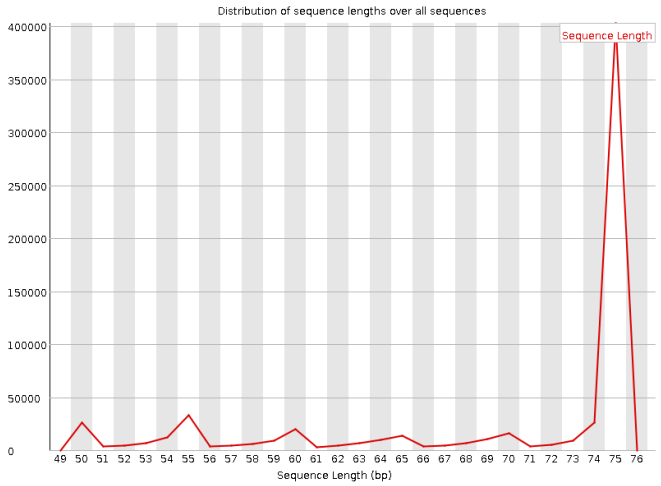
Прямые парные:



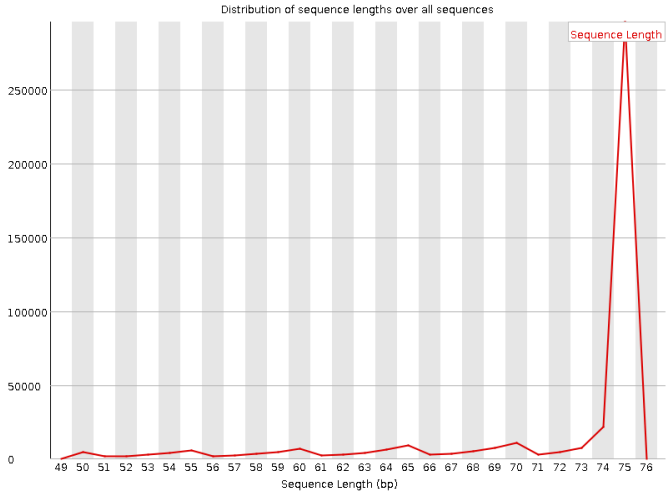
Прямые непарные:



Обратные парные:



Обратные непарные:



Задание 12

Картирование чтений на референсный геном

hisat2 -x ../reference/prefix -1../trimmed_forward_paired.fastq.gz -2 ../trimmed_reverse_paired.fastq.gz -p 10 --no-spliced-alignment > map1.sam 2> map1_log.txt

Аргументы:
-указываем путь до индекса референсного генома
-указываем путь до первого файла прямого и обратных чтений после обрезки и фильтрации ридов
-p 10: аргумент указывает количество потоков (параллельных процессов), которые будут использоваться во время выполнения сопоставления.
--no-spliced-alignment: Этот аргумент указывает HISAT2 не выполнять сопряженную сплайс-сайтовую выравнивание (splice site alignment) при поиске соответствий в геноме.(если работаем с референсными геномами без интронов или не хотим учитывать сплайс-сайты в сопоставлении)
-перенаправления стандартного вывода команды в файл **map.sam**
- 2> map_log.txt: 2> используется для перенаправления вывода ошибок (stderr) команды в файл map1_log.txt.

Конвертация sam в bam

1. Сколько весит sam файл в Гб?- 267M (du -h map.sam)

Команда конвертации: samtools sort -o map.bam map.sam

2. Сколько весит bam файл в Гб? 86M (du -h map.bam)

Команда индексации получившийся bam файла: samtools index map.bam (получаем:map.bam.bai)

Анализ bam файла

Выполняем анализ выравнивания ридов из файла map.bam с помощью инструмента SAMtools и сохраняет результат в файл analys_bam.txt
Команда: samtools flagstat map.bam > analys_bam.txt

Выдача analys_bam.txt:
1392336 + 0 in total (QC-passed reads + QC-failed reads)
1343762 + 0 primary
48574 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
142559 + 0 mapped (10.24% : N/A)
93985 + 0 primary mapped (6.99% : N/A)
1343762 + 0 paired in sequencing
671881 + 0 read1
671881 + 0 read2
8 + 0 properly paired (0.00% : N/A)
6598 + 0 with itself and mate mapped
87387 + 0 singletons (6.50% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

1) Сколько чтений картировано на референс в штуках? 142559
2) Сколько чтений картировано на референс в % от количества триммированных чтений? 10.24%
3) Сколько чтений картировано на референс в корректных парах в штуках? 8
4) Сколько чтений картировано на референс в корректных парах в % от количества триммированных чтений? 0.00%

Получение чтений, картированных на вашу хромосому

Мы извлекаем прочтения, относящиеся только к хромосоме номер 4 из файла map.bam, конвертирует его в формат BAM и сохраняет результат в файле chr4_map.bam
Команда: samtools view -h -bS map.bam 4 > 1chr4_map.bam
Аргументы:
1. **samtools view**- команда для просмотра или фильтрации данных в файле формата SAM или BAM.
2. **-h**- указывает samtools сохранить заголовок файла вместе с прочтениями. Без этого флага заголовок будет исключен из вывода.
3. **-bS**- указывает samtools конвертировать входной формат файла из SAM в BAM. Формат SAM - это текстовый формат, а формат BAM - бинарный формат, который занимает меньше места на диске и может быть быстро обработан.
4. **map.bam**- имя входного файла, который должен быть в формате SAM.
5. **4**-номер хромосомы которое указывает samtools извлечь только прочтения, относящиеся к этой конкретной хромосоме.
6. **> chr4_map.bam**- перенаправление вывода, которое указывает на сохранение результата в файл с именем chr4_map.bam.

Получение только правильно картированных пар чтений

1)Фильтруем прочтения в файле chr4_map.bam и сохраняет только те прочтения, которые являются правильно выровненными и имеют парные прочтения.

Команда: samtools view -f 0x2 -bS 1chr4_map.bam > 1true_pairs_chr4_map.bam
Аргументы:
1. **samtools view**- команда для просмотра или фильтрации данных в файле формата SAM или BAM.
2. **-f 0x2**- указывает samtools применить фильтр, чтобы только парные прочтения были включены в результат. **0x2** - соответствует "properly aligned" (правильно выравнено), который обозначает, что оба концы прочтения успешно выровнены.
3. **-bS**: это флаг, который указывает samtools конвертировать входной формат файла из SAM в BAM.
4. **chr4_map.bam**: это имя входного файла
5. **> 1true_pairs_chr4_map.bam**: это перенаправление вывода, которое указывает на сохранение результата в файл с именем true_pairs_chr4_map.bam.

2)А далее , используя команду ниже вычисляем статистические метрики для флагов прочтений в файле true_pairs_chr4_map.bam и сохраняет результаты в текстовом файле true_pairs.txt(информация о количестве прочтений, выровненных правильно, неправильно выровненных, пропущенных, дублированных и других статистических показателях качества выравнивания прочтений)

Команда:samtools flagstat true_pairs_chr4_map.bam > true_pairs.txt

Выдача true_pairs.txt:
28 + 0 in total (QC-passed reads + QC-failed reads)
8 + 0 primary
20 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
28 + 0 mapped (100.00% : N/A)
8 + 0 primary mapped (100.00% : N/A)
8 + 0 paired in sequencing
4 + 0 read1
4 + 0 read2
8 + 0 properly paired (100.00% : N/A)
8 + 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)

1. Сколько чтений картировано на референс в корректных парах в штуках?-8

2. Сколько чтений картировано на референс в корректных парах в % от общего количества картированных чтений?-100.00%

3)Теперь проиндексирую файл(Индексирование BAM-файла позволяет ускорить поиск участков генома, которые соответствуют определенным координатам, и облегчает работу с данными. Создание индекса файла позволяет быстро определить положение прочтения внутри файла и ускоряет ряд операций, таких как поиск, обход и доступ к участкам генома)

Команда: samtools index true_pairs_chr4_map.bam
Получаем:true_pairs_chr4_map.bam.ba

Задание 13

1)Создаем файл покрытия на основе выравнивания данных секвенирования в формате BAM, с помощью команды bcftools mpileup, а затем команда bcftools call выполняет вариантный вызов на основе этого файла покрытия, сохраняя результаты в формате VCF

Команда: bcftools mpileup -f Homo_sapiens.GRCh38.dna.chromosome.4.fa true_pairs_chr4_map.bam | bcftools call -mv -o var.vcf

Аргументы:
- **bcftools mpileup** - инструмент для создания файлов покрытия для множества образцов на основе их выравнивания с использованием программы BCFtools.
- **-f Homo_sapiens.GRCh38.dna.chromosome.4.fa** - указывает на используемый файл референсной последовательности .
- **true_pairs_chr4_map.bam** - файл входных данных в формате BAM, который содержит секвенированные данные.
-В результате выполнения этой части команды генерируется файл, содержащий информацию о покрытии данных секвенирования.
- **bcftools call** - инструмент для вызова гетерозиготных и гомозиготных вариантов на основе секвенированных данных.
- **-mv** - указывает на вызов только тех вариантов, которые являются наиболее вероятными, и сохранение многоаллельных вариантов.
- **-o var.vcf** - указывает имя выходного файла (var.vcf), в котором будут сохранены полученные варианты в формате VCF.

Структура файл VCF (Variant Call Format):

1) Строки, начинающиеся с ## и содержат метаданные и информацию о файле(информация о версии формата, использованных референсных последовательностях, форматах данных и других метаданных)
 2) Строка, начинающаяся с #(#CHROM: Хромосома, на которой находится вариант) и содержит 10 колонок:
- POS: Позиция варианта на хромосоме.
- ID: Идентификатор варианта (может быть пустым).
- REF: Референсная последовательность в данной позиции.
- ALT: Альтернативные последовательности (варианты) в данной позиции.
- QUAL: Качество вызова варианта.
- FILTER: Фильтр, примененный к варианту (пусто, если не было применено).

- INFO: Дополнительная информация о варианте, представленная в формате "ключ=значение"

Смотрим файл var.vcf :

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.11+htslib-1.11-4
##bcftoolsCommand=mpileup -f chr.4.fa true_pairs_chr4_map.bam
##reference=file://chr.4.fa
##contig=<ID=4,length=190214555>
##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=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##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=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">
##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">
##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=1.11+htslib-1.11-4
##bcftools_callCommand=call -mv -o var.vcf; Date=Sun Feb  4 00:03:30 2024
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT true_pairs_chr4_map.bam
4 124614845 . C A 9.88514 .
DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60 GT:PL 1/1:39,3,0
```

2)А теперь проанализируем var.vcf
С помощью утилиты bcftools результаты статистического анализа запишем в файл var_stats.txt

Команда:bcftools stats var.vcf > var_stats.txt

Смотрим файл var_stats.txt :

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

a) Сколько получилось вариантов?- 1
b) Сколько из полученных вариантов являются однонуклеотидными заменами?-1
c) Сколько получилось коротких вставок и делеций?-0

3)Фильтрация вариантов (будут оставлены только те варианты, у которых значение качества (QUAL) больше 30 и глубина покрытия (DP) больше 50)

Команда:bcftools filter -i'%QUAL>30 && DP>50' var1.vcf -o filt_variants.vcf

4)С помощью утилиты bcftools результаты статистического анализа запишем в файл filt_var_stats.txt

Команда:bcftools stats filt_variants.vcf > filt_var_stats.txt

Смотрим файл filt_var_stats.txt:

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

a) Сколько осталось вариантов после фильтрации (в штуках и в процентах)? -0
b) Сколько осталось однонуклеотидных замен (в штуках и в процентах)?-0
c) Сколько осталось коротких вставок и делеций (в штуках и в процентах)?-0

Аннотация вариантов

Проаннотируйте полученные выше профильтрованные варианты с помощью сервиса VEP:

Category	Count
Variants processed	1
Variants filtered out	0
Novel / existing variants	0 (0.0) / 1 (100.0)
Overlapped genes	0
Overlapped transcripts	0
Overlapped regulatory features	0

Allele	Consequence	Existing variant
A	intergenic_variant	rs1387911427

A sequence variant located in the intergenic region, between genes

Задание 14

Задача:проанализировать одноконцевые чтения RNA-seq и составить экспрессионный профиль данного образца

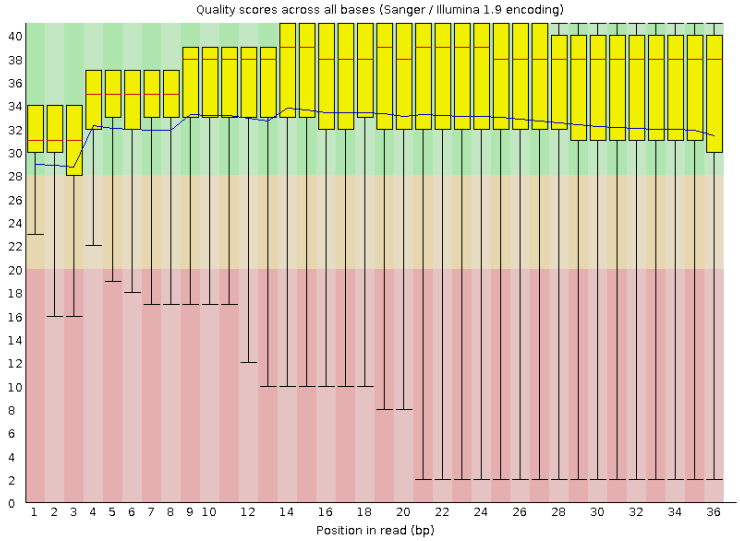
Описание образца

Найходим ID нашего образца в базе NCBI (https://www.encodeproject.org/)
Укажите:
1)ID образца РНК-чтений ENCFF038OLY
2)Ссылка на информацию об образце:
https://www.encodeproject.org/files/ENCFF038OLY/
3)Организм и ткань : Fetal Muscle, Leg primary tissue, day 127, Homo sapiens
4)Стратегия секвенирования :RNA-Seq
5)парноконцевые или одноконцевые чтения:PAIRED
6)цепь-специфичность:no

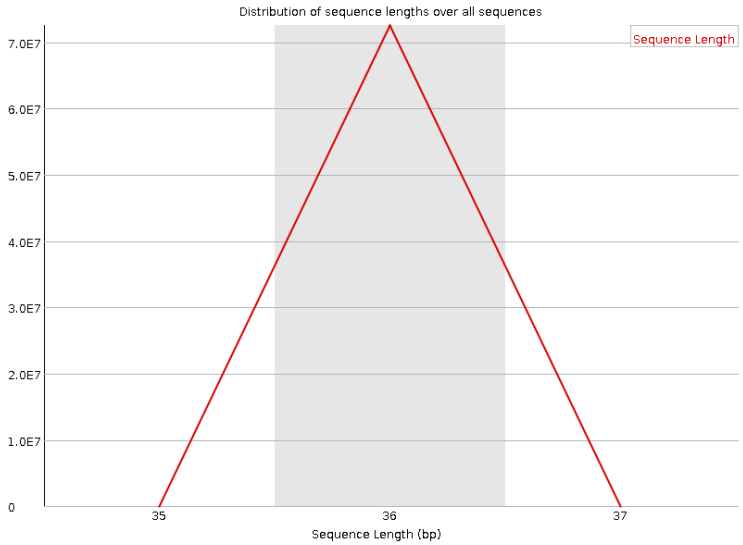
Проверка качества исходных чтений

Проанализируем качество исходных чтений с помощью программы fastqc.
В данном случае у нас только один файл с чтениям: ENCFF038OLY.fastq.gz
Команда: fastqc ENCFF038OLY.fastq.gz

Проанализируем ENCFF038OLY_fastqc.zip:
1)Количество чтений 72517664
2)Краткий комментарий качества чтений по результатам fastqc: Среднее значение, медиана, интервалы между верхними и нижними квартилями вообщем-то все в зеленой зоне , правда усы спускаются аж до 2(см. рис ниже)



3)краткий комментарий о длине ваших чтений по результатам fastqc (картинка Sequence Length Distribution) -все имеют длину 36



Картирование чтений на референс

С помощью команды HISAT2 проводим картирование коротких прочтений RNA-Seq на референсный геном.

Команда:hisat2 -x prefix -k 3 -U ENCFF038OLY.fastq.gz > rna_map.sam 2> rna_map_log.txt
Выдача файла rna_map_log.txt:
72517664 reads; of these:
 72517664 (100.00%) were unpaired; of these:
   67686406 (93.34%) aligned 0 times
   4025941 (5.55%) aligned exactly 1 time
   805317 (1.11%) aligned >1 times
6.66% overall alignment rate

а) Сколько чтений закартировалось на вашу хромосому?-Всего картировалось 4829676 чтений 6.66%

-Конвертируем sam bam:
Команда: samtools sort -o rna_map.bam rna_map.sam
-Индексируем :
Команда:samtools index rna_map.bam – rna_map.bam.bai
-Отбираем только те чтения, которые легли на хромосому:
Команда:samtools view -h -bS rna_map.bam 4 > chr4_rna_map.bam

-Изучаем файл с помощью samtools flagstat(Команда samtools flagstat используется для подсчета статистики флагов в файле BAM (Binary Alignment Map). Опция -O json указывает на формат вывода в формат JSON, который является структурированным форматом данных):
Команда:samtools flagstat -O json chr4_rna_map.bam > flagstat_chr4_rna.txt
Сколько чтений закартировалось на хромосому: "total": 6161222,

Поиск экспрессирующихся генов

Структура файл с генной разметкой в формате GFF (General Feature Format): состоит из строк, каждая из которых содержит 9 столбцов данных, плюс необязательные строки определения трека, такой формат позволяет описывать различные функции на геноме и связывать их с дополнительной информацией через атрибуты.
1. seqname - название хромосомы, названия хромосом могут быть указаны с или без префикса 'chr'.
2. source - название программы, которая сгенерировала эту функцию или источник данных (название базы данных или проекта).
3. feature - название типа функции, например, Gene, Variation, Similarity.
4. start - начальная позиция функции, начиная с 1.
5. end - конечная позиция функции, начиная с 1.
6. score - десятичное число.
7. strand - определяет, является ли цепь прямой (+) или обратной (-).
8. frame - одно из значений '0', '1' или '2'. '0' указывает, что первая база функции является первой базой кодона, '1' указывает, что вторая база является первой базой кодона и т.д.
9. attribute - точка с запятой разделяет список пар тег-значение, предоставляющих дополнительную информацию о каждой функции.

-Подсчет числа генов на хромосоме
Команда:grep '^4' *gtf | cut -f3 | grep 'gene' | wc -l
Получаем:2732

-Посчитаем для каждого гена число картированных на этот ген чтений с помощью htseq-count(htseq-count используется для подсчета количества выравниваний ридов на гены в файле BAM и ассоциировании их с аннотацией генов из файла GTF):

Команда:htseq-count -f bam -s no -m union -t gene rna_map.bam Homo_sapiens.GRCh38.110.chr.gtf> htseq-count-out.txt

Опции аргумента:

Опция -f bam указывает, что входной файл - формат BAM
Опция -s no используется для обозначения отсутствия информации о направленности чтения (неориентированные чтения).
Опция -m union указывает, что для чтений, которые перекрывают несколько аннотированных областей, будет учитываться объединение этих областей при подсчете.
Опция -t gene указывает, что будут считаться только гены (не траскрипты или экзоны, а именно гены).

Смотрим последние 5 сток файла htseq-count-out.txt (всего было обработано :72517664 чтений):

__no_feature 379496- чтения, которые не попали в ген
__ambiguous 805372-не ясные
__too_low_aQual 0 - порог качества
__not_aligned 67686406 -в SAM файле чтения без выравнивания
__alignment_not_unique 805317- попали в несколько позиций

Попали на гены на 4 хромосоме = Всего на 4 хромосоме - __no_feature - __ambiguous - __not_aligned-_alignment_not_unique = 72517664- 379496-67686406-805372= 3384390 чтения попали в гены