ПР11:

Задача: подготовить необходимые файлы, изучить качество предложенных чтений и проиндексировать референс.

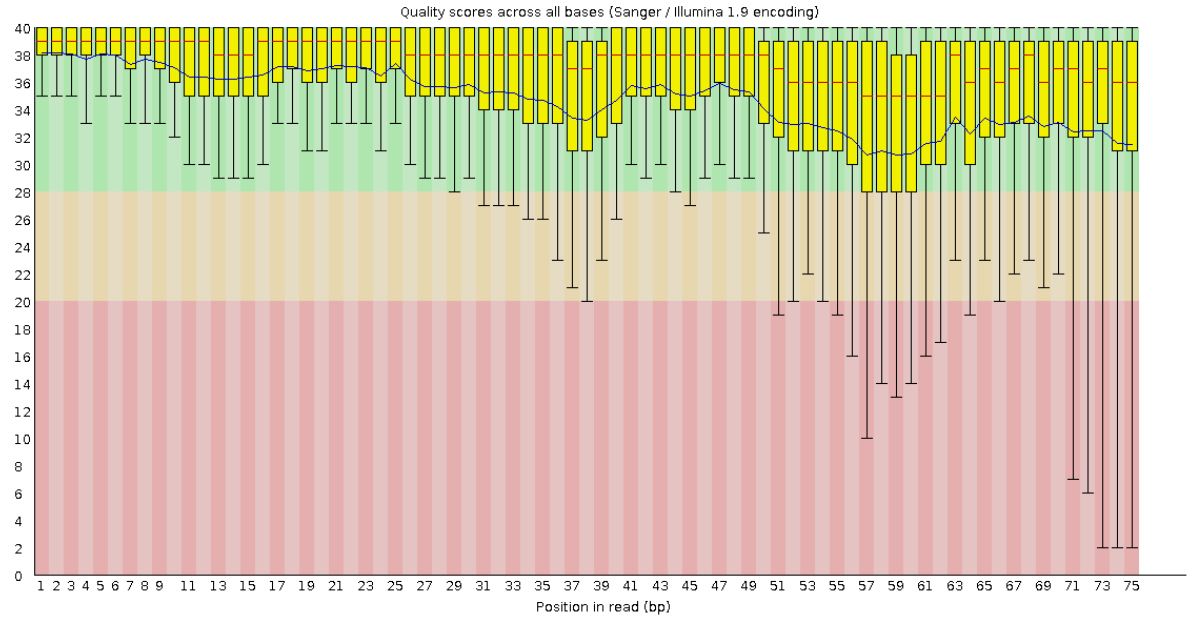
1. ***Индексация по hisat2.***
   Команда: hisat2-build Homo_sapiens.GRCh38.dna.chromosome.6.fa ref
   Получены файлы ref.N.ht2 (где N – натуральное число от 1 до 8)

2. ***Индексация Samtools.***
   Команда: samtools faidx Homo_sapiens.GRCh38.dna.chromosome.6.fa
   Получен файл Homo_sapiens.GRCh38.dna.chromosome.6.fa.fai
   Файл содержит 5 колонок:
   1) Имя референса
   2) Длина
   3) Смещение в файле FASTA, в байтах, начинающихся с нуля для 1 базы.
   4) Количество баз в каждой строке (последняя строка может быть короче)
   5) Количество байтов в каждой строке (в отличие от 4 колонки включает байты, которые кодируют разделители строк)
   Нам нужно узнать точное имя хромосомы и длину хромосомы в нуклеотидах, посмотрим на 1 и 2 колонки.
   Имя: 6
   Длина: 170805979

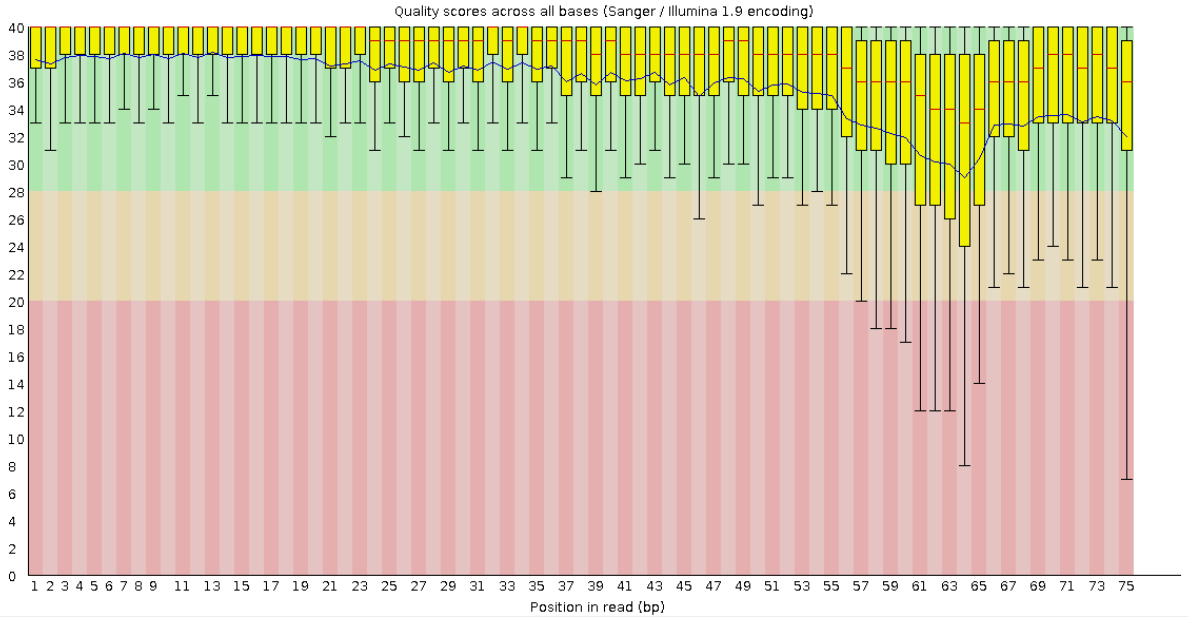
3. ***Найти описание образца в базе данных SRA NCBI.***
   Запрос: SRR10720421
   Описание образца:
   a. SRR ID: 9683501
   b. Ссылка на информацию об образце из NCBI: https://www.ncbi.nlm.nih.gov/sra/?term=9683501
   c. Прибор для секвенирования: Illumina (Illumina Genome Analyzer IIx)
   d. Организм: Homo sapiens
   e. Стратегия секвенирования: OTHER – другое (Комментарий: вот это немного не понятно, от нас прячут какую-то стратегию?)
   f. Парноконцевые или одноконцевые чтения: Парноконцевые
   g. Сколько чтений ожидается (spots): 31.4 M

4. ***Проверка качества исходных чтений с помощью программы FastQC.***
   У нас есть 2 .gz файла с «прямыми» (forward reads) и «обратными» чтениями (reverse reads) соответственно.
   Команды: fastqc SRR10720421_1.fastq.gz – для прямых чтений
   fastqc SRR10720421_1.fastq.gz – для обратных чтений
   Получены файлы: SRR10720421_1.fastqc.zip
   SRR10720421_2.fastqc.zip
   SRR10720421_1_fastqc.html
   SRR10720421_2_fastqc.html
   a. Какое количество пар чтений получилось: 31417056
   b. Совпадает ли количество чтений у «прямых» и «обратных» чтений: Совпадает
   c. Качество чтений:

«Прямые» чтения:



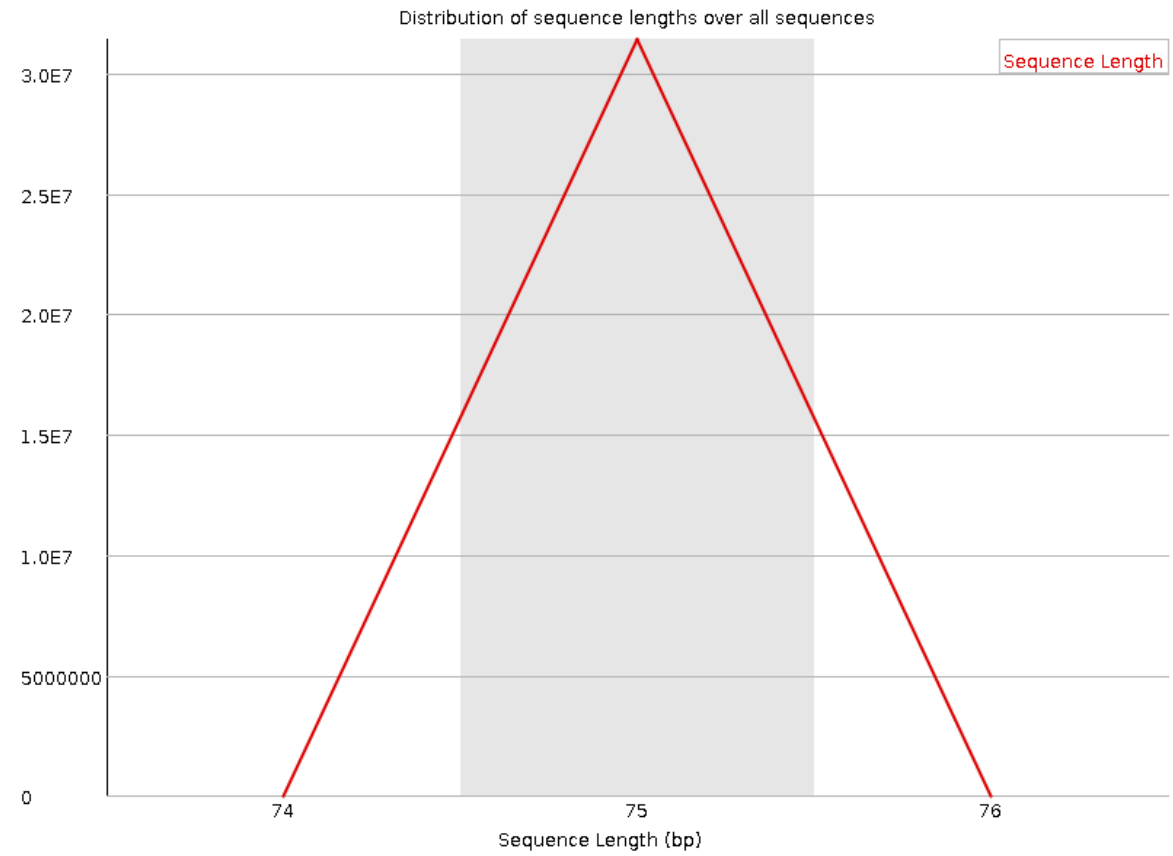
«Обратные» чтения:



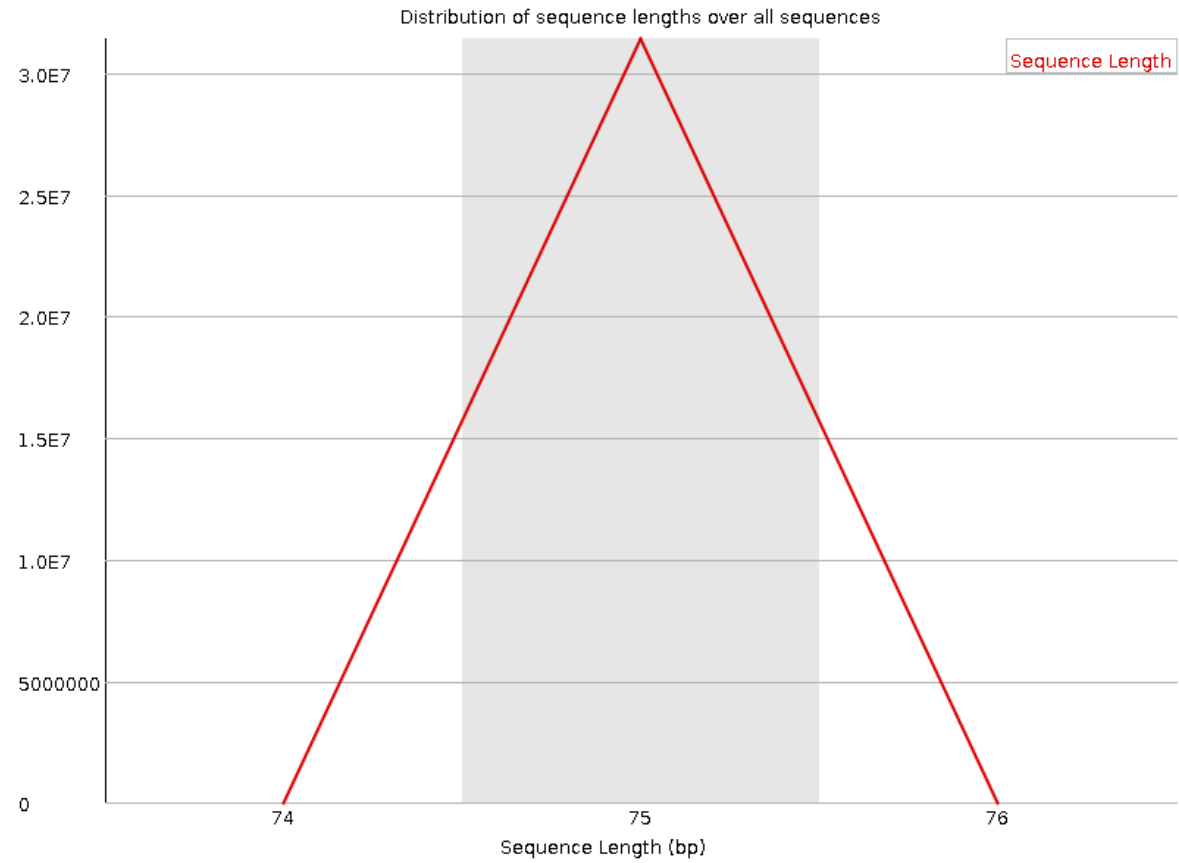
Я считаю, что качество чтений хорошее, однако концы чтений начинают иметь большой разброс в качестве.

d. Длина чтений:

«Прямые» чтения:



«Обратные» чтения:



Большинство чтений имеют длину 75

5. ***Фильтрация чтений с помощью TrimmomaticPE.***

Команда: TrimmomaticPE -phred33 SRR10720421_1.fastq.gz SRR10720421_2.fastq.gz SRR10720421_1_paired.fastq.gz SRR10720421_1_unpaired.fastq.gz SRR10720421_2_paired.fastq.gz SRR10720421_2_unpaired.fastq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 TRAILING:20 MINLEN:50

Выход: Input Read Pairs: 31417056 Both Surviving: 29620794 (94.28%) Forward Only Surviving: 567890 (1.81%) Reverse Only Surviving: 976158 (3.11%) Dropped: 252214 (0.80%) и 4 файла с соответствующими названиями, использованными при составлении команды.
Комментарий: Я думал, что убирать адаптеры надо всегда, заметил конфуз уже после выполнения практикума (вот уже его сегодня надо успеть сдать, один только VEP 3 дня ждать надо, чтобы очередь дошла). Времени теперь всё переделать уже нету, эти триммированные чтения фундаментальны для данной работы. Потерял 2711 чтений (0.01%) из-за данной ошибки (и, вероятно, испортил другие). Извините :(

Теперь проанализируем все отфильтрованные чтения с помощью FastQC.

Команды: fastqc SRR10720421_1_paired.fastq.gz
fastqc SRR10720421_1_unpaired.fastq.gz
fastqc SRR10720421_2_paired.fastq.gz
fastqc SRR10720421_2_unpaired.fastq.gz

Получены соответствующие zip и html файлы.

a. Какое количество пар чтений осталось в штуках: 29620794
b. какой процент пар чтений остался: 94.28%
c. Сравнение качеств paired и unpaired чтений после триммирования: Качество для paired намного выше, чем для unpaired.
d. Сравнение качеств до и после триммирования (paired): После триммирования качество улучшилось.
e. Как изменилась длина чтений после триммирования: Была длина 75, а стала 50-75. Видимо из-за того, что были обрезаны чтения с плохим качеством.

ПР12:

6. ***Картирование чтений на референсный геном с помощью hisat2.***

Команда: hisat2 -x ref -1 SRR10720421_1_paired.fastq.gz -2 SRR10720421_2_paired.fastq.gz --no-spliced-alignment -p 4 > mapped_SRR10720421.sam

Выход:

29620794 reads; of these:
29620794 (100.00%) were paired; of these:
27962927 (94.40%) aligned concordantly 0 times
1590328 (5.37%) aligned concordantly exactly 1 time
67539 (0.23%) aligned concordantly >1 times

27962927 pairs aligned concordantly 0 times; of these:
8184 (0.03%) aligned discordantly 1 time

27954743 pairs aligned 0 times concordantly or discordantly; of these:
55909486 mates make up the pairs; of these:
55508903 (99.28%) aligned 0 times

312690 (0.56%) aligned exactly 1 time
87893 (0.16%) aligned >1 times
6.30% overall alignment rate
А также файл mapped_SRR10720421.sam весом в 12 Гб

7. ***Конвертация SAM в BAM.***
   Команда: samtools sort -o mapped_SRR10720421.bam mapped_SRR10720421.sam

   Выход: mapped_SRR10720421.bam весом в 3.5 Гб

   Проиндексируем bam файл с помощью ***samtools index***:
   Команда: samtools index mapped_SRR10720421.bam

   Выход: mapped_SRR10720421.bam.bai

8. ***Анализ BAM файла с помощью samtools.***
   samtools flagstat – посчитает количество выравниваний для каждого типа флагов
   samtools view mapped_SRR10720421.bam | less – эта команда позволит заглянуть в BAM файл (BAM – бинарный файл)

   Команды:
   samtools flagstat -O tsv mapped_SRR10720421.bam > mapped_SRR10720421_flagstat.txt

   samtools flagstat -O json mapped_SRR10720421.bam > mapped_SRR10720421_flagstat_js.txt

   Выход: соответствующие файлы, json выглядит более удобно для работы, буду анализировать его.

   Анализ файла mapped_SRR10720421_flagstat_js.txt:

   a. Сколько чтений картировано на референс в штуках: 4256576
   b. Сколько чтений картировано на референс в % от количества триммированных чтений: 7.12%
   c. Сколько чтений картировано на референс в корректных парах в штуках: 3315734
   d. Сколько чтений картировано на референс в корректных парах в % от количества триммированных чтений: 5.60%

9. ***Получение чтений, картированных на хромосому.***
   Нам нужно вытащить bam файл с 6 хромосомой из файла полного экзона (mapped_SRR10720421.bam). Сделать это нам поможет ***samtools view***.

   Команда: samtools view -h -bS mapped_SRR10720421.bam 6 > chr6.bam

   Выход: bam файл с чтениями, картированными на 6 хромосому

10. ***Получение только правильно картированных пар чтений.***
    Будем искать по флагу. Флаг 0x2 – это правильно картированные чтения.

    Команда: samtools view -f 0x2 -bS chr6.bam > proper_mapped_chr6.bam

Выход: bam файл с только правильно картированными чтениями на 6 хромосому.

Изучим полученный файл с помощью ***samtools flagstat***, как мы делали в прошлый раз:
Команда:
samtools flagstat -O json proper_mapped_chr6.bam > flagstat_proper_mapped_chr6.txt

a. Сколько чтений картировано на референс в корректных парах в штуках: 3315734
b. Сколько чтений картировано на референс в корректных парах в % от общего количества картированных чтений: 100%

Проиндексируем этот bam файл, чтобы использовать его в будущем

Команда: samtools index proper_mapped_chr6.bam

ПР13:

11. ***Получение вариантов с помощью bcftools***.
Команда: bcftools mpileup -f Homo_sapiens.GRCh38.dna.chromosome.6.fa proper_mapped_chr6.bam | bcftools call -mv -o chr6_variants.vcf

Выход: файл VCF.

Структура файла:
1) Строки, начинающиеся с ## и содержащие информацию о файле.
2) Строка, начинающаяся с # и содержащая названия 10 колонок, разделённых табуляциями.
3) Данные, соответствующие своим колонкам.

Проанализируем файл с помощью ***bcftools stats***.
Команда: bcftools stats chr6_variants.vcf > stats_chr6_variants.txt

a) Сколько получилось вариантов: 69327
b) Сколько из полученных вариантов являются однонуклеотидными заменами: 67406
c) Сколько получилось коротких вставок и делеций: 1921

12. ***Фильтрация вариантов с помощью bcftools filter.***
Отфильтруем из изначального VCF файла варианты с хорошим качеством и большим перекрыванием.
Команда: bcftools filter -i'%QUAL>30 && DP>50' chr6_variants.vcf > filtered_chr6_variants.vcf

Проанализируем отфильтрованный VCF файл с помощью ***bcftools stats***.
Команда: bcftools stats filtered_chr6_variants.vcf > stats_filtered_chr6_variants.txt
Необходимо найти количества в штуках и в процентах, относительно нефильтрованного файла.
a) Сколько осталось вариантов после фильтрации: 1542 (~2.22%)

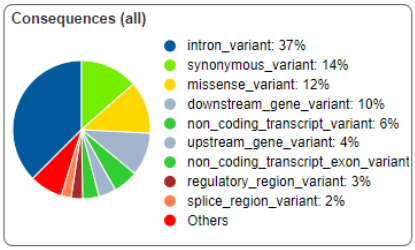
b) Сколько из полученных вариантов являются однонуклеотидными заменами: 1500 (~2.23%)
c) Сколько получилось коротких вставок и делеций: 42 (~2.19%)

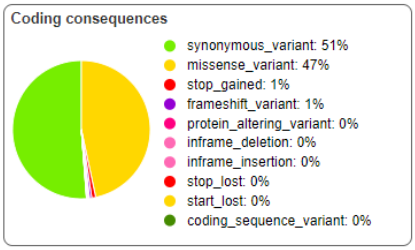
13. ***Аннотация вариантов с помощью сервиса VEP.***

Необходимо проаннотировать файл filtered_chr6_variants.vcf

Аннотация по RefSeq:

Category	Count
Variants processed	1542
Variants filtered out	0
Novel / existing variants	392 (25.4) / 1150 (74.6)
Overlapped genes	509
Overlapped transcripts	3216
Overlapped regulatory features	184

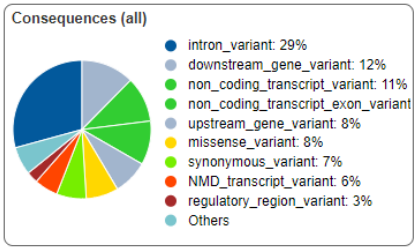


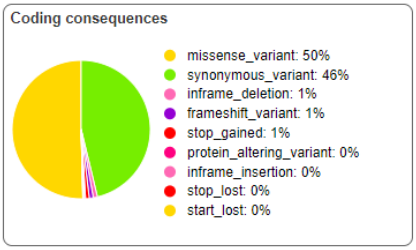


Вариантов с IMPACT HIGH: 57 штук.

Аннотация по Ensembl/GENCODE:

Category	Count
Variants processed	1542
Variants filtered out	0
Novel / existing variants	392 (25.4) / 1150 (74.6)
Overlapped genes	637
Overlapped transcripts	3133
Overlapped regulatory features	184





Вариантов с IMPACT HIGH: 39 штук.

ПР 14.

Задача - проанализируйте одноконцевые чтения RNA-seq и составьте экспрессионный профиль данного образца.

14. ***Описание образца.***

a. ID образца РНК-чтений: ENCFF729YAX
b. Ссылка на информацию об образце: https://www.encodeproject.org/files/ENCFF729YAX/
c. Организм и ткань: *Homo sapiens*, кровь. Клеточная линия GM12878.
d. Стратегия секвенирования: RNA-Seq (Полное секвенирование, тотальная РНК)
e. Парноконцевые или одноконцевые чтения: Одноконцевые
f. Цепь-специфичность: Отсутствует

15. ***Проверка качества исходных чтений с помощью FastQC.***

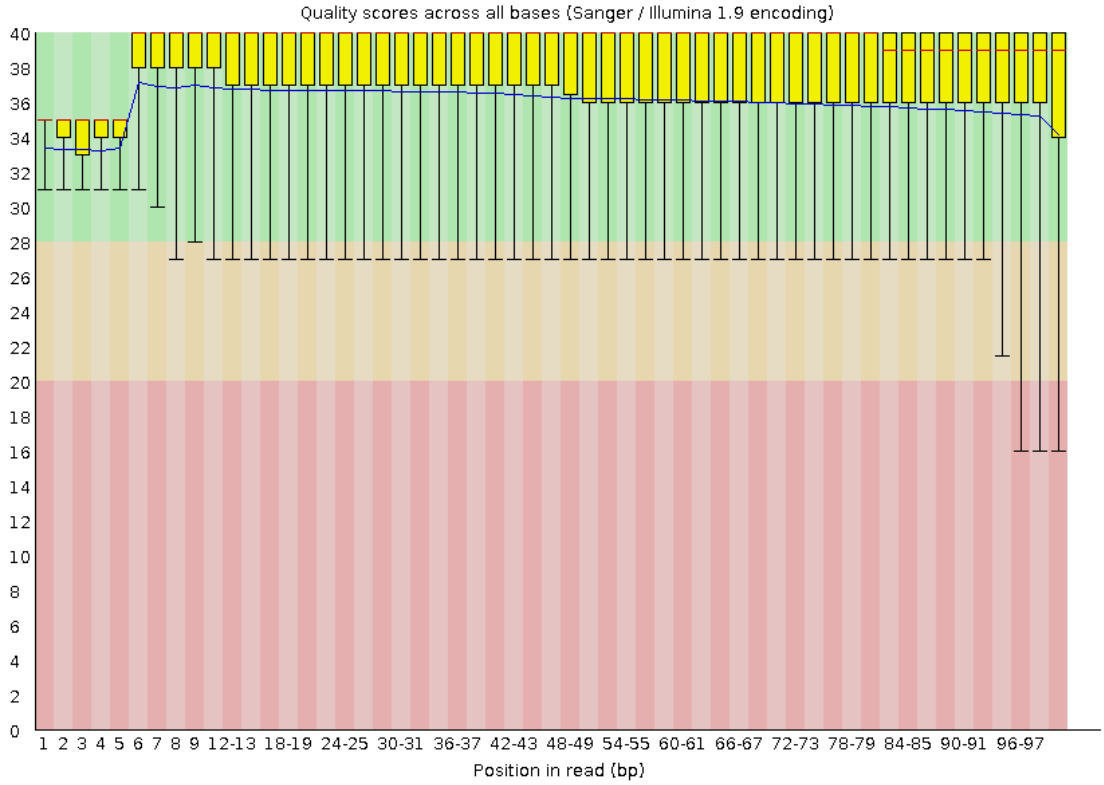
Команда: fastqc ENCFF729YAX.fastq.gz

Получены файлы: ENCFF729YAX_fastqc.zip и ENCFF729YAX_fastqc.html

Качество чтений:

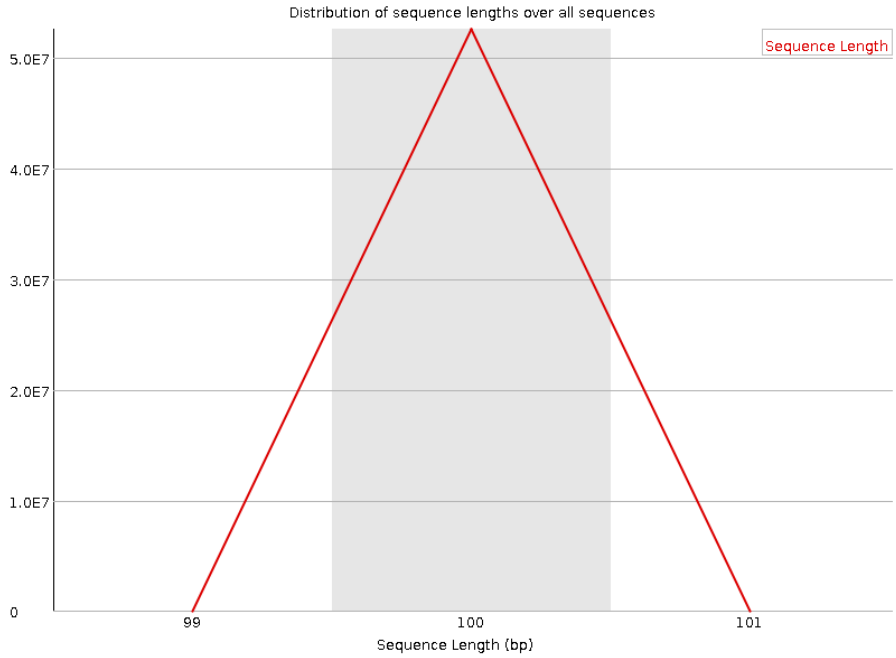
a. Количество чтений: 52532133

b. Краткий комментарий качества чтений по результатам fastqc:



Хорошее качество чтений, усы показывают стабильный разброс в качестве.

c. Краткий комментарий о длине чтений по результатам fastqc:



Большинство чтений имеют длину 100

16. ***Картирование чтений на референс с помощью hisat2:***
    Команда: hisat2 -x ref -k 3 -U ENCFF729YAX.fastq.gz > mapped_rna_ENCFF729YAX.sam

    Выход:
    52532133 reads; of these:
      52532133 (100.00%) were unpaired; of these:
        48827070 (92.95%) aligned 0 times
        3527237 (6.71%) aligned exactly 1 time
        177826 (0.34%) aligned >1 times
    7.05% overall alignment rate

    А также файл: mapped_rna_ENCFF729YAX.sam

    Конвертируем этот файл SAM в BAM:
    samtools sort -o mapped_rna_ENCFF729YAX.bam mapped_rna_ENCFF729YAX.sam

    Получен: mapped_rna_ENCFF729YAX.bam (теперь можно удалить SAM)

    Индексируем этот файл:
    samtools index mapped_rna_ENCFF729YAX.bam

    Отберём только те чтения, которые легли на нашу хромосому:
    samtools view -h -bS mapped_rna_ENCFF729YAX.bam 6 > chr6_rna.bam

    Изучим файл с помощью ***samtools flagstat:***
    samtools flagstat -O json chr6_rna.bam > flagstat_chr6_rna.txt

    Сколько чтений закартировалось на хромосому: 4011989 (100%)

17. ***Поиск экспрессирующихся генов:***

Как устроен файл с генной разметкой? (В данном случае у нас .gtf)
Каждая строка описывает один объект, строка состоит из 9 колонок, разделённых табуляциями.
Номер колонки и её назначение:
1 – Имя хромосомы (Или скэффолда)
2 – Название программы, описавшей объект (или название источника)
3 – Тип объекта (Feature type)
4 – Координаты начала объекта
5 – Координаты конца объекта
6 – Счёт (Score)
7 – Цепь-специфичность (+, -)
8 – Рамка (0, 1, 2)
9 – Атрибуты (широкая колонка с дополнительной информацией)

Посчитаем для каждого гена число картированных на этот ген чтений с помощью ***htseq-count***:
Опции для команды: -f (формат – sam/bam), -s(взята ли информация из цепь-специфичных чтений – yes/no/reverse), -m(режим – union, intersection-strict, intersection-nonempty), -t(какой тип объекта искать – из 3-ей колонки .gtf файла)
Команда:
htseq-count -f bam -s no -m union -t gene mapped_rna_ENCFF729YAX.bam Homo_sapiens.GRCh38.110.chr.gtf > htseq-count_out.txt
Выход: файл htseq-count_out.txt, каждая строка состоит из 2-х колонок: ID гена и количество картированных чтений на него. (всего было обработано 52839059 чтений)

5 последних строк:
__no_feature 193769 – не смогли назначить feature
__ambiguous 897414 – ничему не засчитались, т.к. подходили к нескольким генам (засчитались бы с опцией –nonunique all)
__too_low_aQual 0 – пропущено из-за опции -a (отвечает за порог качества по MAPQ)
__not_aligned 48827070 – в SAM файле чтения без выравнивания
__alignment_not_unique 177826 – больше 1 выравнивания в BAM/SAM, поэтому ничему не засчитало (засчитались бы с опцией –nonunique all)

Сколько чтений попало в границы генов: 2436054
Сколько чтений попало мимо границ генов: 306926
(Комментарий: можно ли в «мимо» засчитывать из 5 последних строк?)
(для счёта использовал cut -f2 htseq-count_out.txt| paste -sd+ - |bc и калькулятор)