Каримова Карина 202

хромосома 6
SPA exome SRR10720404
Команды выполняются в папке /mnt/scratch/NGS/karina.kar

ЗАДАНИЕ ПР11

Подготовка референса

Получение референса

Копирую референсную хромосому в папку /mnt/scratch/NGS/karina.kar/genome:

```
mkdir genome
cp ../DATA/hg38/Homo_sapiens.GRCh38.dna.chromosome.6.fa ./genome/
```

Индексация для hisat2

Индексирую хромосому с помощью hisat2 (префикс у файлов — index) и переношу в папку /mnt/scratch/NGS/karina.kar/genome/hisat2_ind :

```
hisat2-build ./genome/Homo_sapiens.GRCh38.dna.chromosome.6.fa index
mkdir hisat2_ind
mv ./index* ./hisat2_ind/
```

Индексация samtools

Теперь индексирую хромосому для samtools:

```
samtools faidx ./genome/Homo_sapiens.GRCh38.dna.chromosome.6.fa
```

Выдача - файл ./genome/Homo_sapiens.GRCh38.dna.chromosome.6.fa.fai, в котором одна строка: 6 170805979 56 60 61

Расшифровка:
6 — имя хромосомы (6 хромосома)
170 805 979 — длина хромосомы в нуклеотидах
56 — с 56 байта в файле начинается нуклеотидная последовательность
60 — в каждой строке файла по 60 нуклеотидов
61 — в каждой строке файла 61 байт (учитывается перенос строки)

Чтения ДНК

Описание образца

Информация об образце ДНК-чтений:

- SRR ID: **SRR10720404**
- ссылка на страницу NCBI
- прибор для секвенирования: **Illumina Genome Analyzer IIx**
- организм: **Homo sapiens**
- стратегия секвенирования: **экзомное**
- чтения: **парноконцевые**
- сколько чтений ожидается: **38 518 929 spots**

Проверка качества исходных чтений

Копирую чтения к себе в папку и анализирую качество исходных чтений с помощью программы fastqc:

```
mkdir reads
cp ../DATA/dna_reads/SRR10720404_1.fastq.gz ./reads/
cp ../DATA/dna_reads/SRR10720404_2.fastq.gz ./reads/

fastqc ./reads/SRR10720404_1.fastq.gz
fastqc ./reads/SRR10720404_2.fastq.gz
```

Небольшие организационные работы:

```
mkdir fastqc_for
mv ./reads/*1_fastqc* ./fastqc_for/

mkdir fastqc_rev
mv ./reads/*2_fastqc* ./fastqc_rev/
```

Каждых чтений (и прямых и обратных) по 38518929, что совпадает с ожидаемым количеством.

Качество чтений хорошее, но немного падает к концу (рис. 1-2). Длина чтений — 75 нк (рис. 3-4).

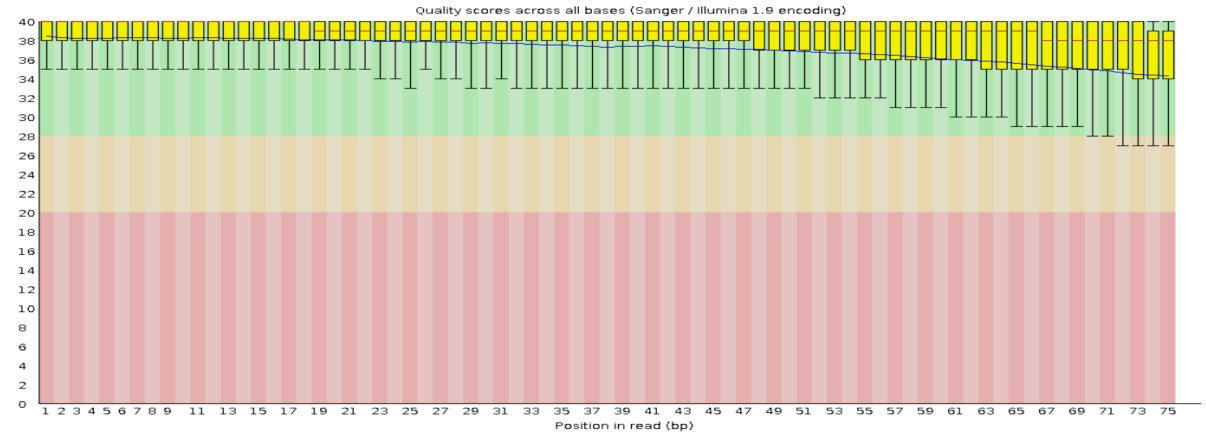


Рис .1. Per base sequence quality для прямых чтений

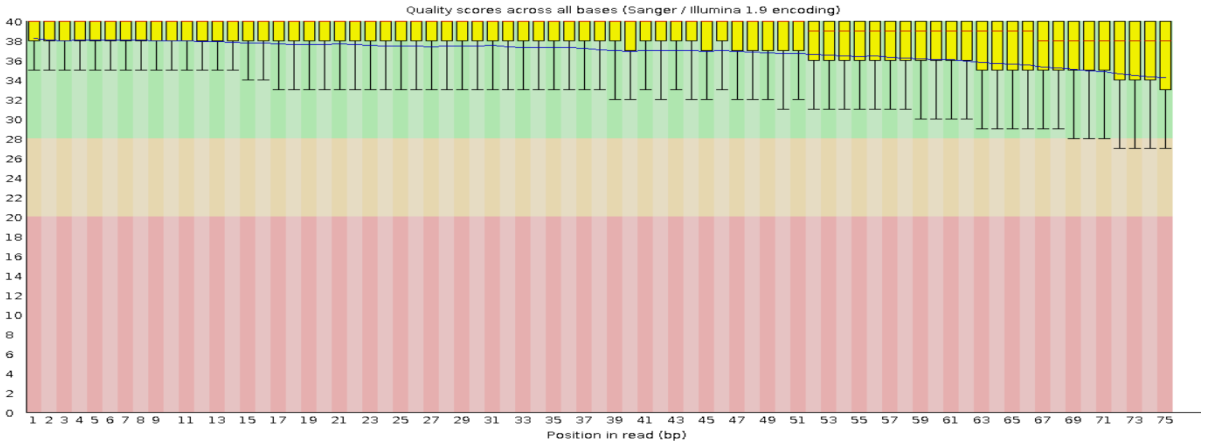


Рис .2. Per base sequence quality для обратных чтений

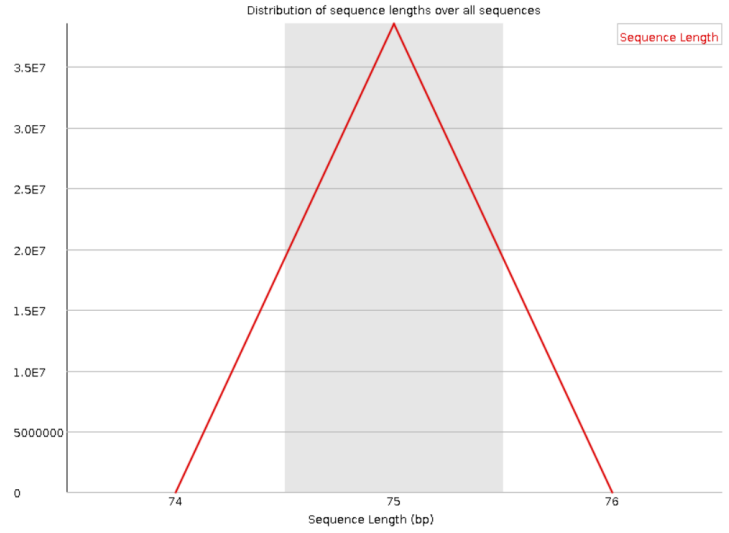


Рис .3. Sequence Length Distribution для прямых чтений

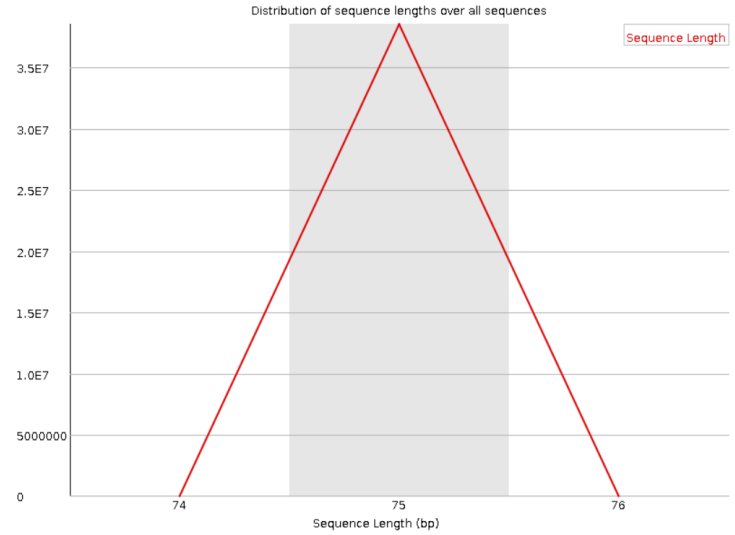


Рис .4. Sequence Length Distribution для обратных чтений

Фильтрация чтений

Далее фильтруем наши чтения (у нас парно концевые чтения, поэтому используем TrimmomaticPE). Необходимо удалить с конца чтений нуклеотиды с качеством ниже 20 (TRAILING:20) и удалить чтения длина которых ниже 50 нуклеотидов (MINLEN:50). (еще запущу это на восьми потоках и поставлю на фон знаком &)

mkdir logs
mkdir trimmed

TrimmomaticPE -threads 8 -phred33 -trimlog logs/trimmomatic.log reads/SRR10720404_1.fastq.gz reads/SRR10720404_2.fastq.gz SRR10720404_trim_forward_paired.fq.gz SRR10720404_trim_forward_unpaired.fq.gz SRR10720404_trim_reverse_paired.fq.gz SRR10720404_trim_reverse_unpaired.fq.gz TRAILING:20 MINLEN:50 &

mv ./*trim* ./trimmed/

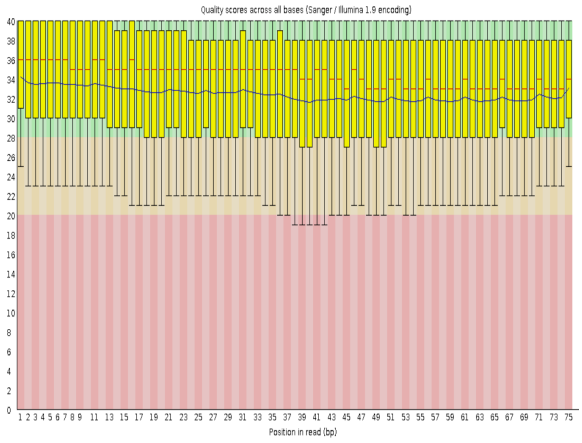
Так как у нас парно концевые чтения, то производя “чистку” если мы удалим одно чтение из пары, то оставшееся не будет иметь пары. А для последующего картирования мы будем учитывать правильно выровненные пары. Именно поэтому после работы trimmomatic получается именно 4 файла, и при картировании мы будем использовать только два (с парами).

Проверка качества триммированных чтений

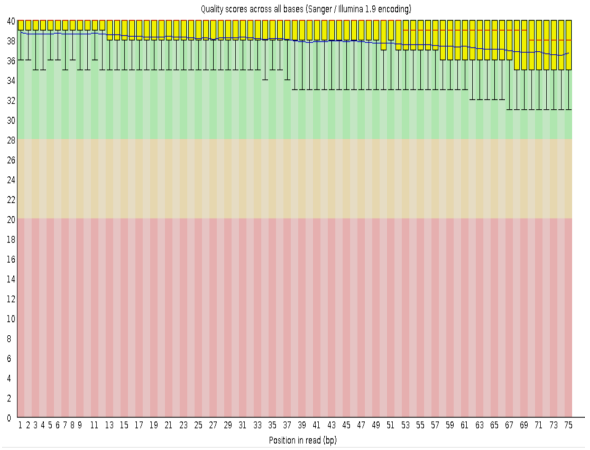
Анализируем качество чтений после обработки программой Trimmomatic с помощью программы fastQC:

```
cd trimmed
mkdir fastqc_trim
fastqc ./SRR10720404_trim_forward_paired.fq.gz
fastqc ./SRR10720404_trim_forward_unpaired.fq.gz
fastqc ./SRR10720404_trim_reverse_paired.fq.gz
fastqc ./SRR10720404_trim_reverse_unpaired.fq.gz
mv ./*fastqc* ./fastqc_trim/
cd ..
```

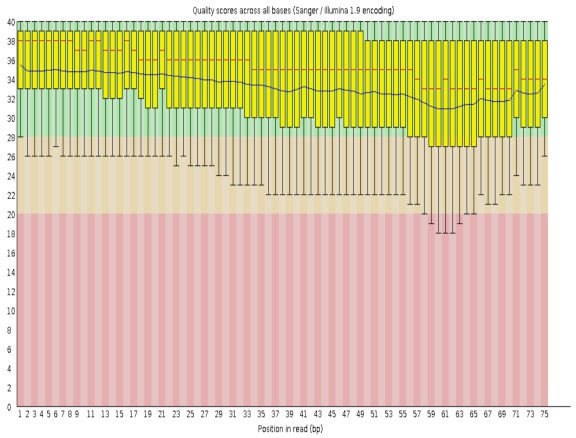
Осталось 37 136 668 (96.41%) чтений с парами (то есть осталось 18 568 334 пар чтений).



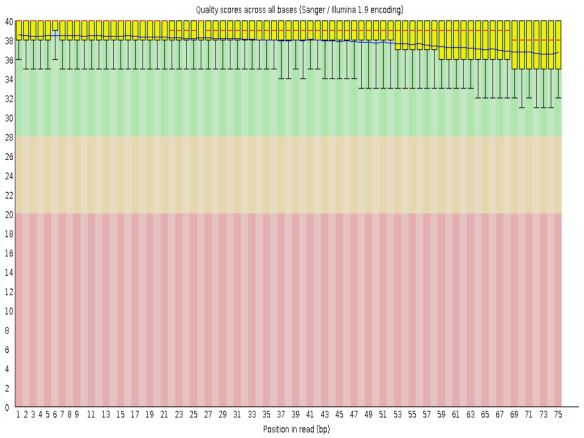
reverse_unpaired



reverse_paired



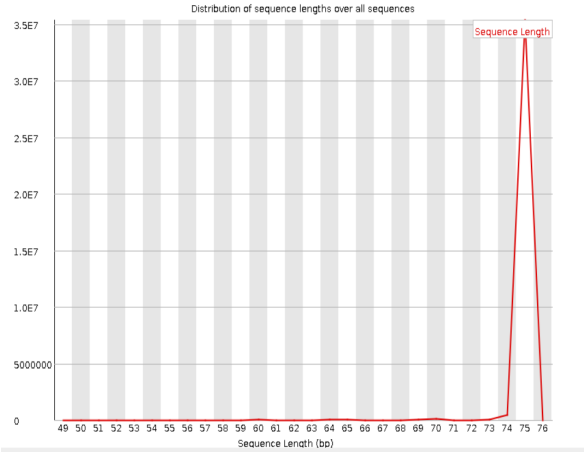
forward_unpaired



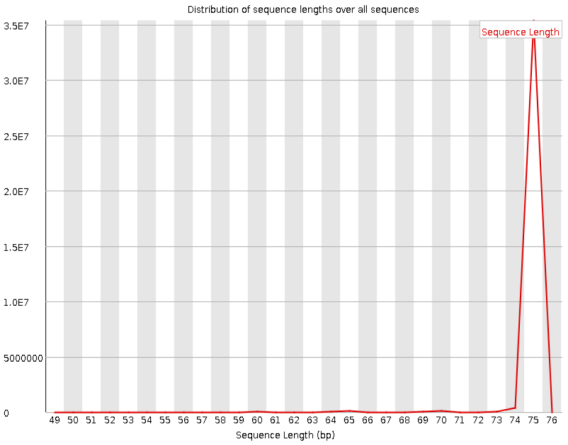
forward_paired

Качество чтений paired значительно лучше чем у unpaired.
Качество чтений paired стало лучше чем было у всех чтений до триммирования.

После триммирования появились чтения с длиной меньше 75 (что не удивительно).



reverse_paired



forward_paired

ЗАДАНИЕ ПР12

Картирование чтений на референсный геном

Картируем парные триммированные чтения на референсный геном с помощью hisat2:

mkdir mapped
cd mapped

hisat2 -x ../hisat2_ind/index -1 ../trimmed/SRR10720404_trim_forward_paired.fq.gz -2 ../trimmed/SRR10720404_trim_reverse_paired.fq.gz -p 8 --no-spliced-alignment -S SRR10720404.sam 2> hisat2.log

Параметры:
-x префикс индексного файла
-1 прямые чтения
-2 обратные чтения
-p количество потоков

--no-spliced-alignment параметр, запрещающий возможность сплайсинга

-S сохраняем в SAM файл

Конвертация sam в bam

Sam файл весит 15 Гб, переконвертируем его в сортированный bam файл:

samtools sort -@ 8 -o SRR10720404.bam SRR10720404.sam
rm SRR10720404.sam

Полученный bam файл весит 4 Гб , теперь проиндексируем его:

samtools index -@ 8 SRR10720404.bam

Анализ bam файла

Анализируем bam файл с помощью возможностей samtools:

samtools flagstat -@ 8 SRR10720404.bam > SRR10720404_flagstat.txt

Теперь напрямую заглянем в bam файл:

samtools view SRR10720404.bam | less

Из полученного flagstat файла:

На референсный геном картировалось 5 430 008 чтений (это 14.6% от триммированных), 4 122 392 в корректных парах (это 11.1% от триммированных).

Получение чтений, картированных на вашу хромосому

А теперь получим чтения, картированные только на шестую хромосому:

samtools view -@ 8 -h -bS SRR10720404.bam 6 > SRR10720404_6_Chr.bam

Получение только правильно картированных пар чтений

А теперь получим только правильно спаренные (PROPER_PAIR) картированные чтения (опция -f 0x2):

samtools view -@ 8 -f 0x2 -bS SRR10720404_6_Chr.bam > SRR10720404_6_Chr_sorted.bam

samtools flagstat -@ 8 SRR10720404_6_Chr_sorted.bam > SRR10720404_6_Chr_sorted_flagstat.txt

Чтений в корректных пар, картированных на референс, 4 122 392 (это 75.9% от общего числа картированных чтений).

Проиндексируем этот файл:

samtools index -@ 8 SRR10720404_6_Chr_sorted.bam

ЗАДАНИЕ ПР13

Выполняется в папке **/mnt/scratch/NGS/karina.kar/variants**

Получение вариантов

Получаем варианты:

bcftools mpileup -f ../genome/Homo_sapiens.GRCh38.dna.chromosome.6.fa ../mapped/SRR10720404_6_Chr_sorted.bam | bcftools call -mv -o SRR10720404_6_Chr.vcf

Анализируем полученный файл:

bcftools stats SRR10720404_6_Chr.vcf **>** SRR10720404_6_Chr_stats.txt

Всего 84 135 вариантов, из которых 81 573 — это однонуклеотидные полиморфизмы, а остальные 2 562 — это индели.

Фильтрация вариантов

Теперь фильтруем варианты по качеству (не менее 30) и глубине покрытия (не менее 50):

bcftools filter -i '%QUAL>30 && DP>50' SRR10720404_6_Chr.vcf > SSRR10720404_6_Chr_filtered.vcf

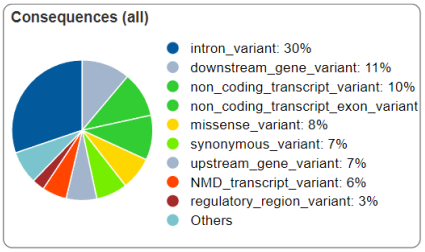
bcftools stats SSRR10720404_6_Chr_filtered.vcf > SRR10720404_6_Chr_filtered_stats.txt

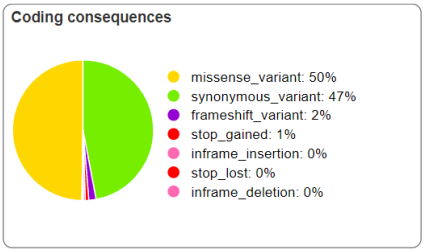
Осталось 1 945 вариантов (2.3%), из которых 1 886 (2.3%) — это однонуклеотидные полиморфизмы, а остальные 59 (2.3%) — это индели. (Проценты относительно нефильтрованных вариантов)

Аннотация вариантов

Аннотируем варианты с помощью сервиса VEP.

Category	Count
Variants processed	1945
Variants filtered out	0
Novel / existing variants	476 (24.5) / 1469 (75.5)
Overlapped genes	755
Overlapped transcripts	3542
Overlapped regulatory features	221





Вариантов с IMPACT HIGH — 68 штук.

ЗАДАНИЕ ПР14

Задание выполнялось в папке **/mnt/scratch/NGS/karina.kar/rna_seq**

Описание образца

Информация об образце RNA-seq:

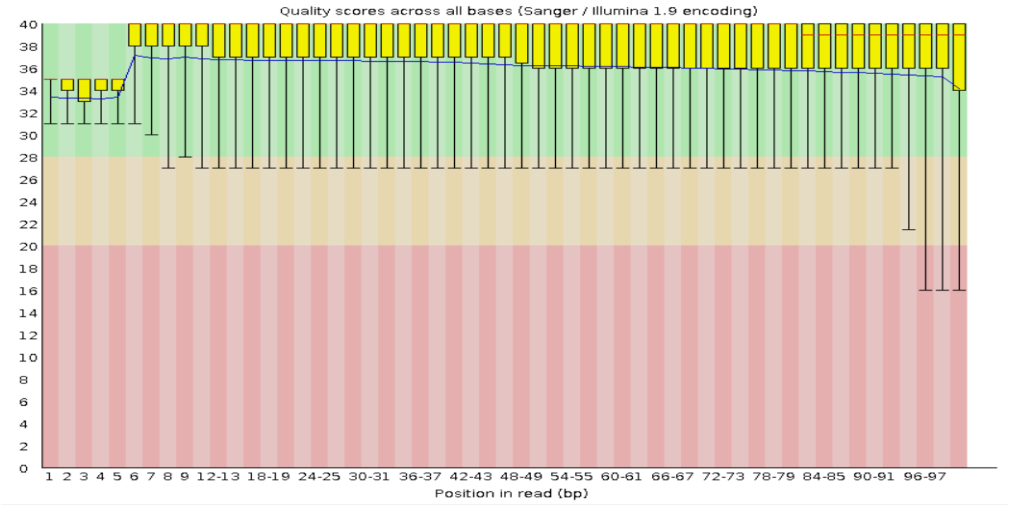
- ID: **ENCFF729YAX**
- Данный образец ENCFF729YAX взят из датасета ENCSR843RJV (эксперимента проекта ENCODE)
- клеточная линия ***Homo sapiens* GM12878** (лимфобластоидные клетки)
- стратегия секвенирования: **polyA plus RNA-seq** (то есть секвенировали полиаденилированые мРНК)
- чтения: **одноконцевые**
- цепь-специфичность: **нет**

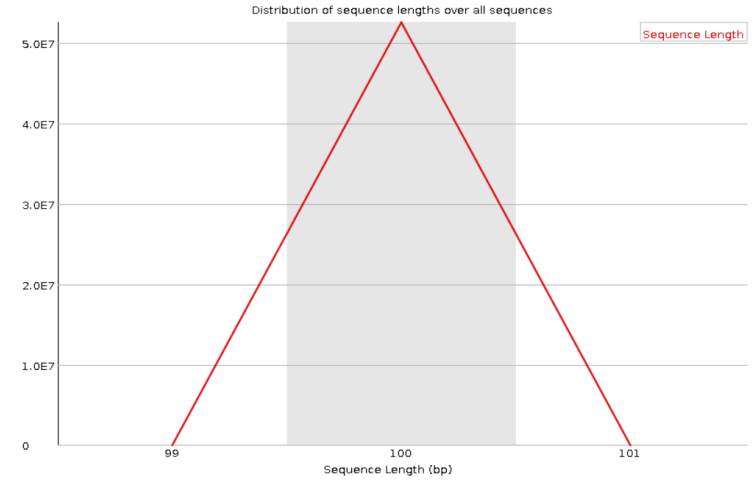
Проверка качества исходных чтений

Копирую чтения к себе в папку и анализирую качество исходных чтений с помощью программы fastqc:

```
cp ../../DATA/rna_reads/ENCFF729YAX.fastq.gz ./
fastqc ENCFF729YAX.fastq.gz
```

Количество чтений — 52 532 133. В начале чтений мы видим, что качество первых пяти нуклеотидов хуже чем последующих (но все же в зеленой зоне), к концу качество ухудшается (и разброс качества последних даже достигает красной зоны). Длина чтений — 100 нк.





Картирование чтений на референс

Картируем чтения на нашу хромосому:

hisat2 -x ../hisat2_ind/index -k 3 -U ./ENCFF729YAX.fastq.gz -S ENCFF729YAX.sam 2> hisat2.log &

Переводим sam файл в сортированный bam и индексируем:

samtools sort -@ 8 -o ENCFF729YAX.bam ENCFF729YAX.sam &
rm ENCFF729YAX.sam

samtools index -@ 8 ENCFF729YAX.bam

Отбираем чтения, легшие только на нашу хромосому:

samtools view -@ 8 -h -bS ENCFF729YAX.bam 6 > ENCFF729YAX_6_Chr.bam

samtools index -@ 8 ENCFF729YAX_6_Chr.bam
samtools flagstat -@ 8 ENCFF729YAX_6_Chr.bam > ENCFF729YAX_6_Chr_flagstat.txt

На хромосому закартировалось 4 011 887 чтений (7.6% от всех).

Поиск экспрессирующихся генов

Файл с генной разметкой имеет формат The GTF (General Transfer Format) и представляет собой таблицу где каждая строка это один feature. Написано что это за feature (ген, транскрипт, CDS и тд), расположение на хромосоме (координаты и цепь), а также комментарии.

Копирую файл с разметкой к себе:

cp ../../DATA/genes/Homo_sapiens.GRCh38.110.chr.gtf ./

Для каждого гена из разметки считаем число картированных на этот ген:

htseq-count -f bam -s=no -m union -t exon -o ENCFF729YAX_6_Chr_genes.sam ENCFF729YAX_6_Chr.bam Homo_sapiens.GRCh38.110.chr.gtf

Опции программы htseq-count:

-f формат входного файла (bam).

-s являются ли данные цепь-специфичными

-m режим для обработки считываний, перекрывающих более одного объекта (в нашем случае экзона).

-t признак из третьего столбца файла с референсным геномом. У нас РНК-секвенирование, поэтому exon.

-o записывать выход в SAM файл с указанным названием

(В получившимся файле для каждого чтения написано попал ли он в ген или нет)

Вывод программы:

...
гены и сколько на них картировалось чтений
...
__no_feature 439 465 (чтения, не попавшие в гены)
__ambiguous 292 174 (непонятно попали ли в ген/ попали в перекрывающиеся гены)
__too_low_aQual 0
__not_aligned 0
__alignment_not_unique 177 786 (попали в несколько генов)

Мимо границ генов попали 439 465 чтений, в границы генов — 3 572 422, из них однозначно в какой-то ген **3 102 462** (77.3% от картированных на хромосому).

скрипт лежит в папке /mnt/scratch/NGS/karina.kar/pr11-13