Сборка и анализ геномов

hisat2-build ../chr5.fa prefix
Команда создает 8 файлов с индексами генома.
samtools faidx chr5.fa
Создаёт файл chr5.fa.fai из одной строки:
5 181538259 56 60 61
где 5 - номер хромосомы,
181538259 - её длина
56 - номер байта, с которого начинается последовательность
60 - число нуклеотидов в каждой строке
61 - число байтов в каждой строке

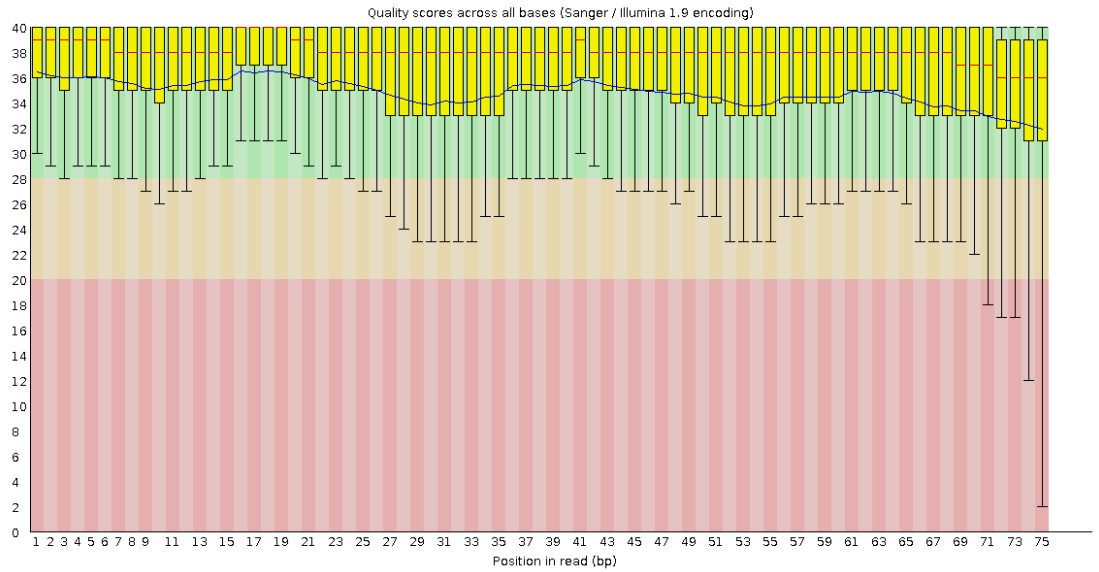
Описание образца

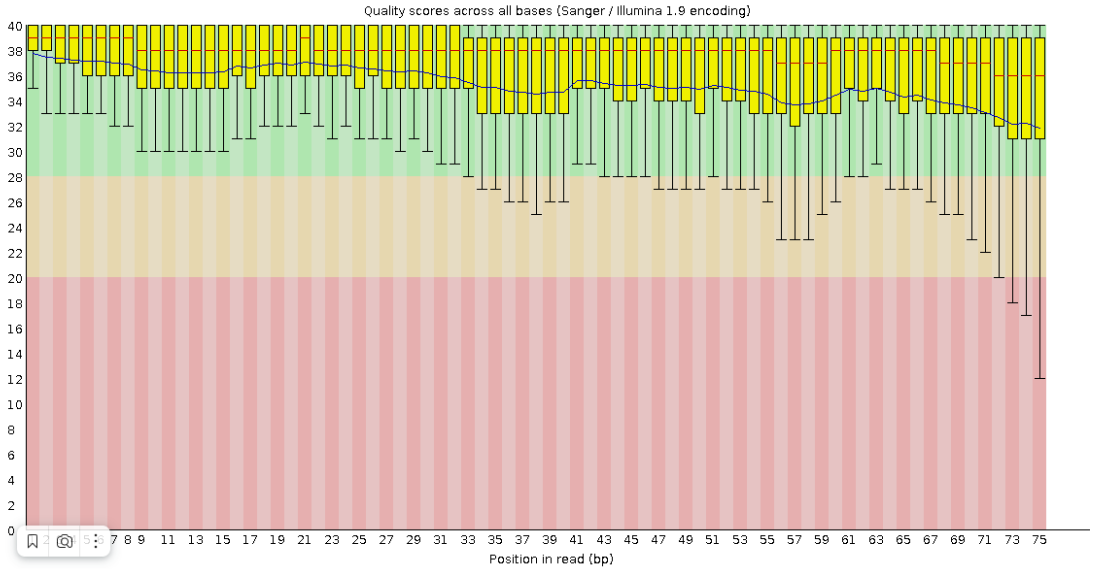
SRR ID: 10720402
Ссылка на NCBI: https://www.ncbi.nlm.nih.gov/sra/SRR10720402
Прибор для секвенирования: ILLUMINA (Illumina Genome Analyzer IIx)
Организм: Homo sapiens
Стратегия секвенирования: Экзомное секвенирование
Тип чтений: парноконцевые
Ожидаемое число чтений: 28966798

Проверка качества исходных чтений

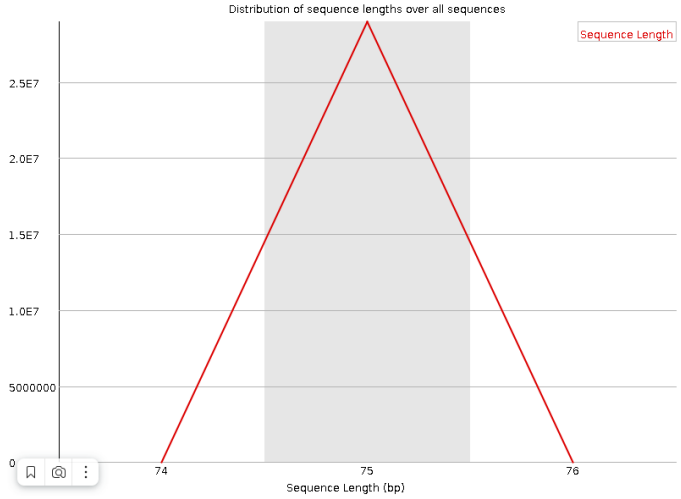
fastqc SRR10720402_1.fastq.gz
fastqc SRR10720402_2.fastq.gz

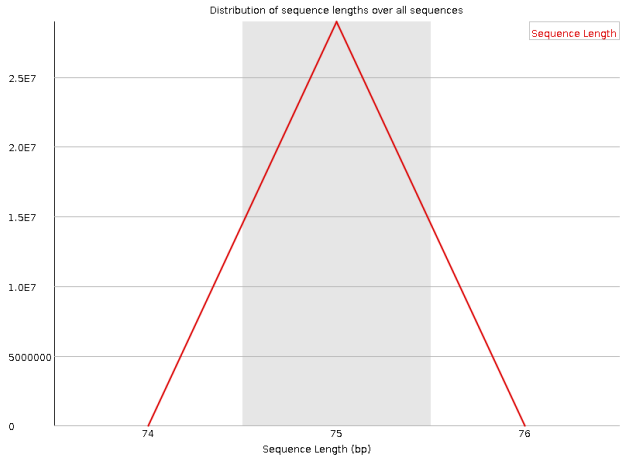
```
Всего 28966798 чтений, число прямых и обратных чтений совпадает
      Качество пар чтений
```



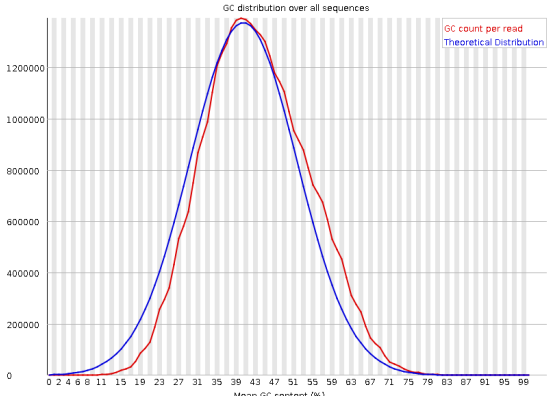


У качества большой разброс, но среднее качество и медиана довольно постоянны. На концах качество значительно падает и уходит красную зону.
Длина чтений

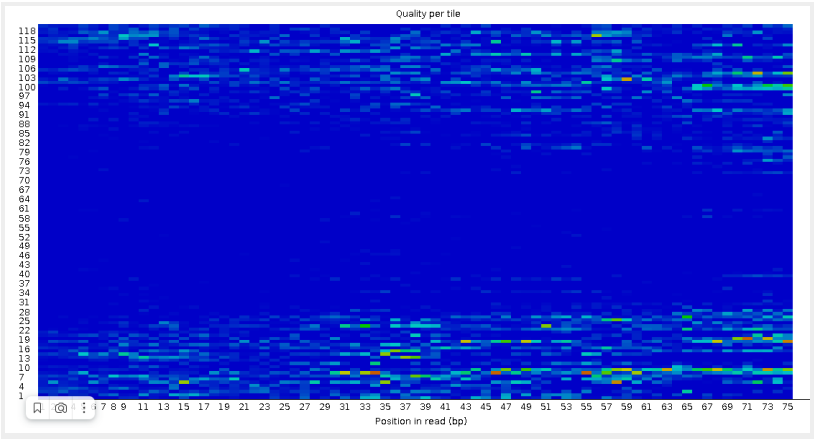


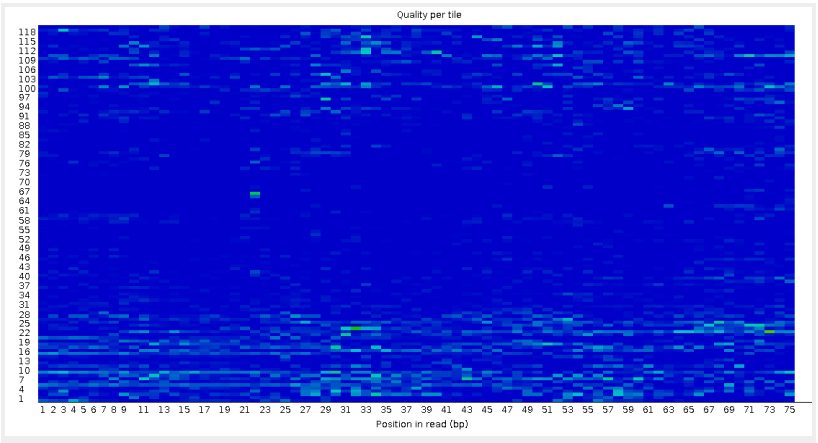


Картинки абсолютно идентичные - все чтения имеют длину 75.
Разброс по длине минимален, что хорошо.
Пункты, которые программа воспринимает как требующие внимания
В прямых чтениях программе не нравится Per sequence GC content и Per title sequence quality.



В целом, распределение GC состава выглядит близким к нормальному, поэтому на предупреждение программы можно не обращать внимания.
В обратных ридах программе тоже не нравится Per title sequence quality.





В Per title sequence quality по x отложен номер нуклеотида, по y - номер ячейки. Цвет соответствует качеству - от синего(высокое качество) до красного(низкое качество).
В обратных чтениях ситуация в целом приемлемая, в прямых - хуже.

Фильтрация чтений

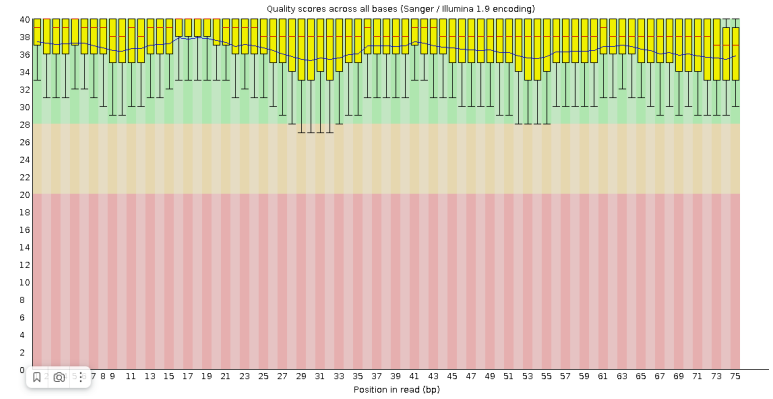
TrimmomaticPE -phred33 SRR10720402_1.fastq.gz SRR10720402_2.fastq.gz trimmed_forward_paired.fastq.gz trimmed_reverse_paired.fastq.gz trimmed_forward_unpaired.fastq.gz trimmed_reverse_unpaired.fastq.gz TRAILING:20 MINLEN:50

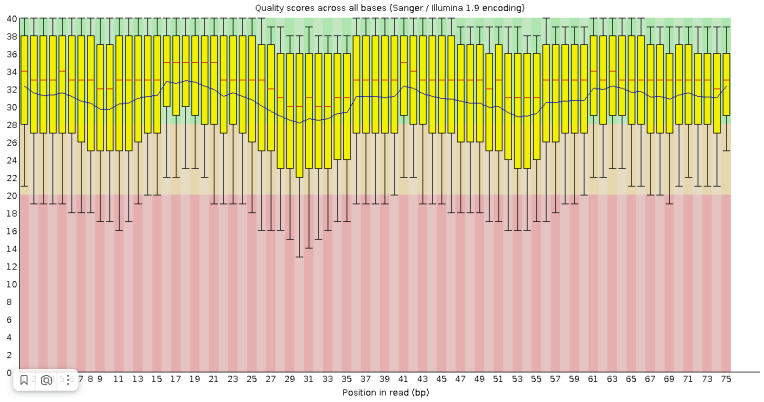
Команда обрезает 20 нуклеотидов с конца ридов и удаляет чтения длины меньшей, чем 50 нуклеотидов. В паре чтений - прямом и обратном - одно из чтений может быть удалено. Такие чтения записываются в файлы с неспаренными чтениями. Парные чтения, в которых осталось и прямое чтение, и обратное, записываются в 2 других файла.

Проверка качества триммированных чтений

Парных чтений осталось 27172718 - это 93% от исходного числа парных чтений.

- Сравнение качества чтений после триммирования: парные и непарные чтения

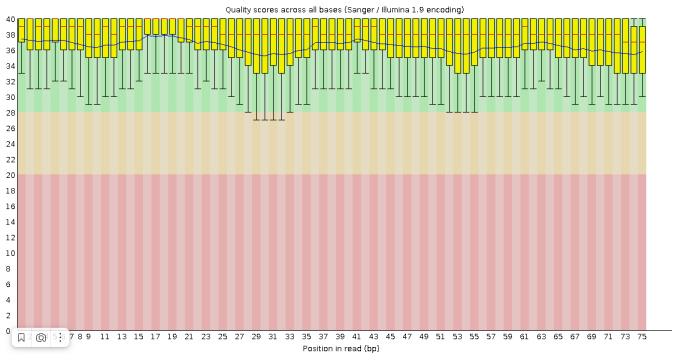




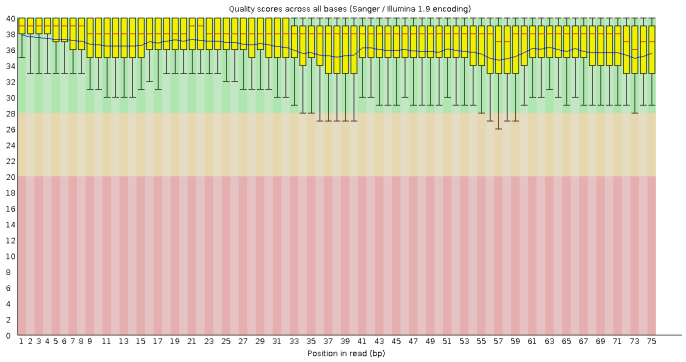
Первая картинка - парные прямые обратные риды, вторая - непарные обратные риды(с прямыми чтениями ситуация аналогичная, хоть и менее выражена). Видно, что качество нуклеотидов в непарных ридах значительно ниже, чем в парных, и даже хуже, чем в исходных чтениях. Поэтому дальше мы с непарными чтениями не работаем.

Сравнение качества парных чтений до и после триммирования

Прямые риды после тримммирования

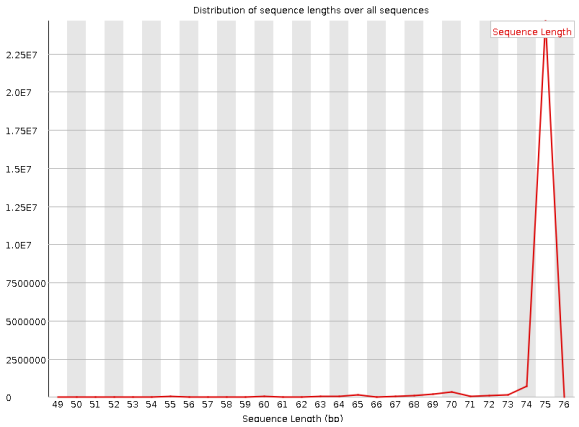


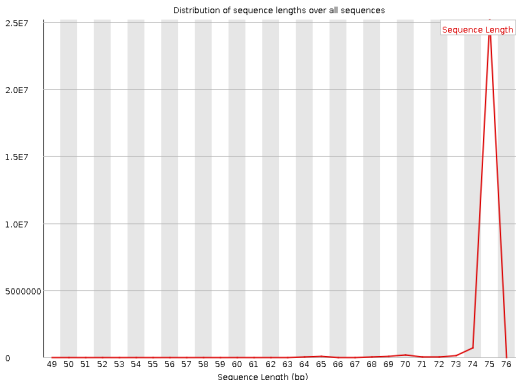
Обратные риды после триммирования



Видно, что качество чтений после триммирования значительно увеличилось.

- Длина чтений после триммирования





Мода длин чтений осталась той же, но появился “хвост” из чтений меньшей длины.

Картирование чтений на референсный геном

hisat2 -p 10 -x genome/hisat2_index/prefix -1 reads/trimmed_forward_paired.fastq.gz -2 reads/trimmed_reverse_paired.fastq.gz --no-spliced-alignment -S chr5.sam 2>hisat2.log

-p -число потоков
-x -префикс индексных файлов
-1 -прямые чтения
-2 обратны чтения

--no-spliced-alignment -запрет сплайсинга
-S - сохранить в файл с заданным названием
Конвертация sam в bam:
samtools sort -o chr5.bam chr5.sam
Sam-файл весит 11 ГБ, а bam-файл -3,3 ГБ, что сильно меньше.
Индексирование bam-файла:
samtools index chr5.bam
Анализируем bam-файл:
samtools flagstat chr5.bam
На выходе получился файл со статистикой по bam-файлу.
На референс картировалось 3879651 чтений - это 7.06% триммированных чтений. Корректными парами картировалось 3879651 чтений - это 5.23% от чтений.

Получение только правильно картированных пар чтений

samtools view -h -bS chr5.bam 5 > mapped.bam
Фильтрует чтения, картированные на 5 хромосому.
-h - вывод с заголовком, -b - вывод в bam-файл, -S - определять тип файла
samtools view -f 0x2 -bS mapped.bam > cor_mapped.bam
-f 0x2 - фильтрует чтения, выровнявшиеся корректно
samtools flagstat cor_mapped.bam > cor_mapped.flagstat
- собираем статистику по файлу с отобранными картированными чтениями

Корректно картировалось 3253816 чтений - это 100% чтений. В парах корректно картировалось 3253816 чтений - это тоже 100%. Значит, мы отобрали только чтения, картировавшиеся корректно, что радует.

Получение вариантов

bcftools mpileup -f ../genome/chr5.fa cor_mapped.bam | bcftools call -mv -o chr5.vcf
Команда ищет различия между картированными чтениями и референсом.
bcftools stats chr5.vcf > chr5.vcf.stat - собирает статистеку из длинного vcf-файла
Всего различных вариантов нашла программа 66982, из них SNP -65061(подавляющее большинство), инделей - 1921.
Отфильтруем варианты с хорошим покрытием и качеством:
bcftools filter -i'%QUAL>30 && DP>50' chr5.vcf > filtered_chr5.vcf
И соберем по нему статистику
bcftools stats filtered_chr5.vcf > filtered_chr5.vcf.stat
После фильтрации осталось 1252 варианта(1,86% от исходного числа), из них 1217 - SNP(1,87%), и 35 - индели(1,82%).
Аннотация вариантов

Variants processed	1252
Variants filtered out	0

Novel / existing variants	429 (34.3) / 823 (65.7)
Overlapped genes	498
Overlapped transcripts	2413
Overlapped regulatory features	113

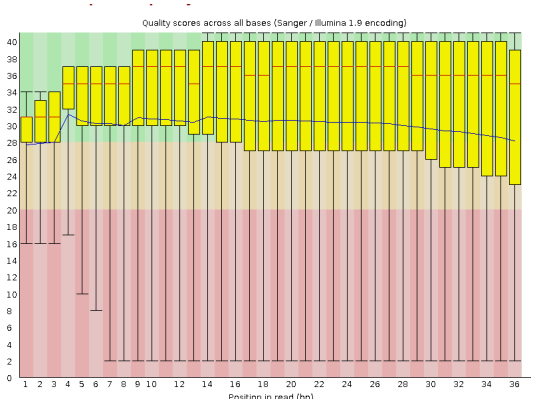
vcf-файл был загружен в VEP. Impact HIGH имеет 28 вариантов. Этот параметр указывает степень влияния на продудкт гена. Все из отобранных вариантов попали в гены. Из них 12 попали в интроны, а 16 - в экзоны. Все варианты среди попавших в интроны - это splice_acceptor_variant, и 4 среди них - в некодирующих транскриптах РНК(non_coding_transcript_variant). Среди попавших в экзоны 12 - это преждевременная остановка трансляции в результате однонуклеотидной замены и 4 - в результате сдвига рамки считывания.

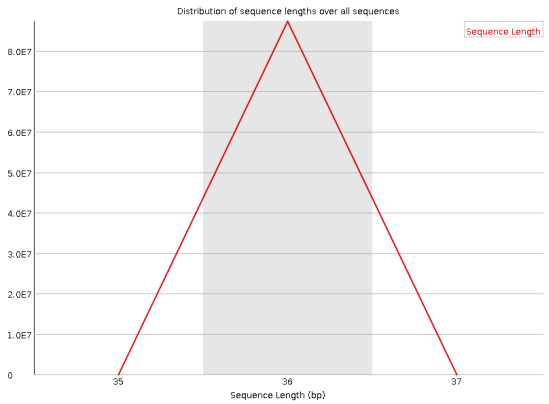
Анализ РНК-чтений

Описание образца

a. ID образца: ENCSR047LLJ
b. Ссылка: https://www.encodeproject.org/report/?type=Experiment&control_type!=*&status=released&perturbed=false&searchTerm=ENCFF975AUW
c. Организм и ткань: Homo sapiens heart tissue male embryo (120 days)
d. Стратегия секвенирования: polyA plus RNA-seq (то есть мРНК)
e. Тип чтений: одноконцевые
f. Цепь-специфичность: нет

Проверка качества исходных чтений





Всего чтений 87265266, у качества очень большой разброс. Длина меньше, чем была у ДНК-ридов, у всех чтений длина одинаковая. Картируем чтения

hisat2 -x ../genome/hisat2_index/prefix -k 3 -U rna_reads.fastq.gz -S rna_reads.sam 2>hisat_rna.log
Большая часть чтений - 78854766 (90.36%) не выровнялась(тк секвенировалась вся мРНК, а выравнивалась только на 1 хромосому), 7411691 (8.49%) чтение выровнялась 1 раз, 998809 (1.14%) - более раза.

Конвертирование в bam:
samtools sort -o rna_reads.bam rna_reads.sam
Индексирование:
samtools index rna_reads.bam
Отбор картированных на 5 хромосому:
samtools view -h -bS rna_reads.bam 5 > rna_map.bam
Проверка отбора:
samtools flagstat rna_map.bam > rna_map.flagstat
После отбора остались только чтения, картированные на 5 хромосому(8410500 шт, 100%)

Поиск экспрессирующихся генов

Устройство gff-файла
Файл несет информацию о расположении генов
Содержит столбцы:
seqname - номер хромосомы
source - программа-источник информации
feature - что находится
start, end - координаты
score - “достоверность” нахождения
strand - цепь + и -
frame - рамка считывания: 0, 1 и 2
attribute - дополнительная информация

Подсчет числа генов на хромосоме
grep "^9" gen_location.gtf | cut -f 3 | grep -c '^gene'
2417

Подсчет чтений, картированных на ген
htseq-count -s no -f bam -t gene -m union -o mapping_rid.sam rna_map.bam gen_location.gtf > gen_cnt.txt 2>count_log.tx
-f - формат ввода
-s - специфичность цепи
-t - использовать только аннотацию определенного элемента
-o - вывод информации о выравниваниях
-m - способ определения участка картирования
В файле gen_cnt.txt последние строчки такие:
__no_feature   337470 - чтения, которые не попали в ген
__ambiguous   2839198 - не ясные
__too_low_aQual 0

__not_aligned   0
__alignment_not_unique   998809 - попали в несколько позиций
Всего попало на 5 хромосому 8410500
Попали на гены на 5 хромосоме = Всего на 5 хромосоме -
__no_feature - __ambiguous - __alignment_not_unique = 8410500 -
2839198 - 998809 = 4572493 чтения попали в гены