ПРАКТИКУМ 11

Индексация hisat2

Команда: hisat2-build Homo_sapiens.GRCh38.dna.chromosome.8.fa chr8

Hisat2-build команда, индексирующая референсный геном. Chr8- префикс названия выходных файлов. В итоге получаем восемь файлов типа chr8.N.ht2, где N номера от 1 до 8.

Индексация samtools

Команда: samtools faidx Homo_sapiens.GRCh38.dna.chromosome.8.fa

Faidx- опция, открывающая доступ к fasta файлам. Получаем на выходе файл с расширением .fa.fai

В этом файле несколько цифр: 8 (номер хромосомы) 145138636 (длина всей последовательности в нуклеотидах) 56 (первый байт, с которого начинается последовательность) 60 (длина строки последовательности в нуклеотидах) 61 (длина строки в байтах, нт + перенос строки).

Описание образца

(Ссылка на образец в NCBI: https://www.ncbi.nlm.nih.gov/sra/?term=SRR10720421)

SRR ID: SRR10720421

Секвенирование: Illumina Genome Analyzer IIx

Стратегия: экзомное (Whole-exome and RNA sequencing)

Организм: человек (Homo sapiens)

Чтения: парноконцевые (Illumina Paired-End Multiplexed Sequencing Protocol)

Spots: 31,417,056

Проверка качества чтения:

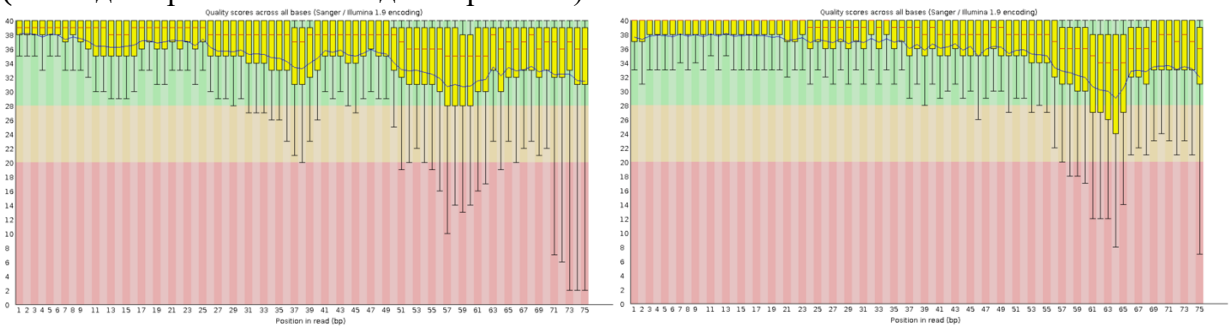
fastqc SRR10720421_1.fastq.gz SRR10720421_2.fastq.gz

В результате получились два файла в формате html (есть у меня на сайте:).

Количество пар чтений: 31,417,056

Количество чтений совпадает у прямых и обратных, и совпадает с ожидаемым количеством

Оценка качества чтений: качество чтений хорошее (все красные медианы и синее среднее значение выше 30), при этом есть участки не очень хорошего качества не в конце чтений (51-62 для прямого и 58-65 для обратного).



Прямое Обратное

Рис 1. Per base sequence quality

Оценка длины чтений: длины одинаковые для прямого и обратного: 75 нт.

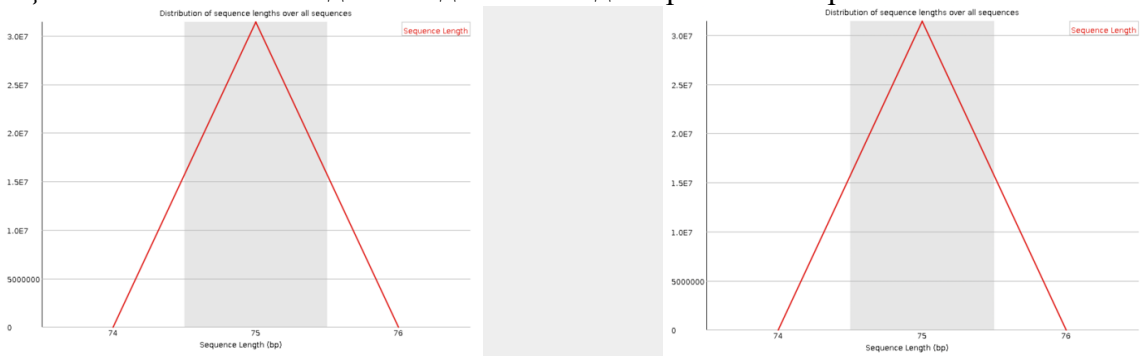


Рис 2. Длины чтений

Фильтрация чтений:

Команда: trimmomaticPE -phred33 SRR10720421_1.fastq.gz SRR10720421_2.fastq.gz trim_forward_paired.fastaq.gz trim_forward_unpaired.fastq.gz trim_reverse_paired.fastq.gz trim_reverse_unpaired.fastq.gz TRAILING:20 MINLEN:50

TrimmomaticPE- для парноконцевых чтений (от paired end), опция -phred33 определяет кодировку качества (если не задать определится автоматически), TRAILING удаляет нуклеотиды с плохим качеством (ниже 20), MINLEN удаляет короткие чтения (меньше 50). В результате получаем четыре файла: два парных (с двумя чтениями после триммирования) и два непарных (одно чтение после триммирования).

Проверка качества триммированных чтений:

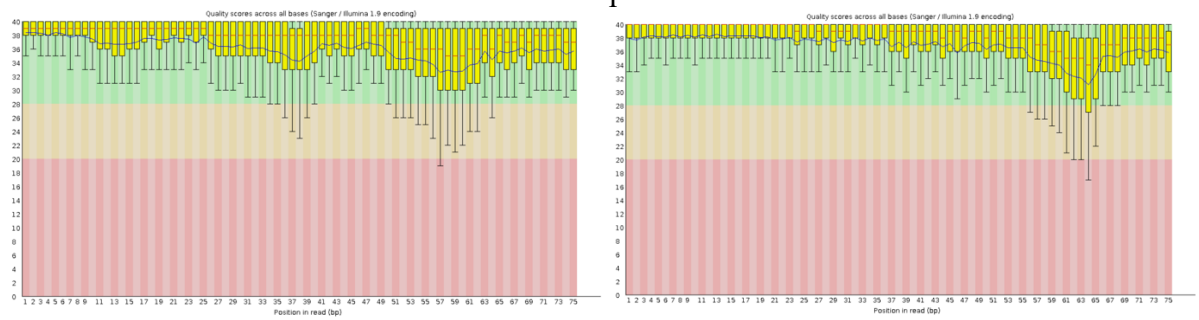
Команда: fastqc trim*

Пар после чтения: 29,626,256

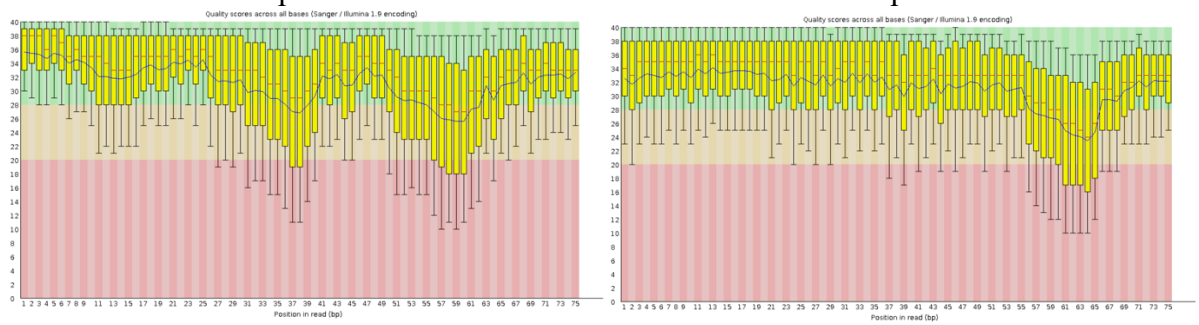
Процент: 94,3%

Сравнение парных и непарных: качество непарных существенно хуже, чем парных.

Парные



Прямое Обратное

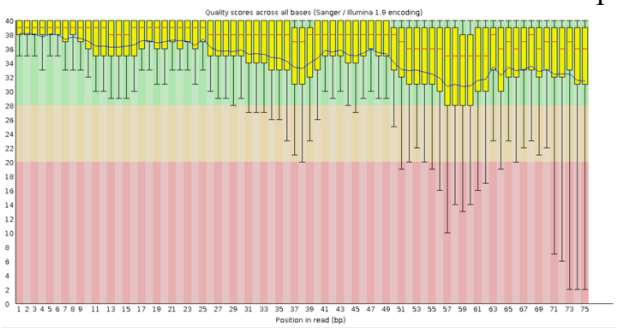


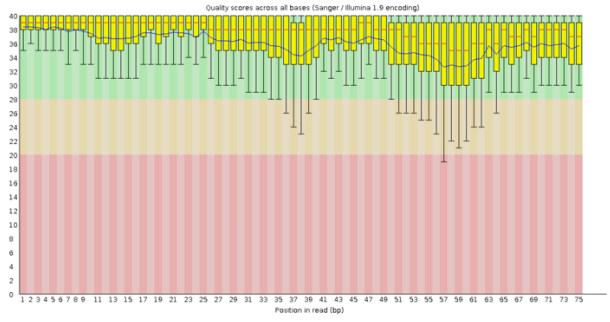
Непарные

Рис 3. Per base sequence quality после триммирования

Сравнение парных до и после триммирования: качество заметно улучшилось

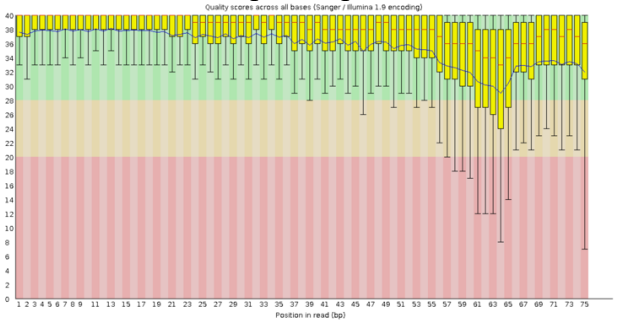
Прямое

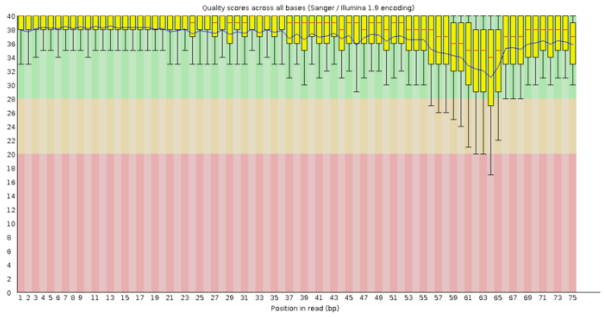




До триммирования

После триммирования

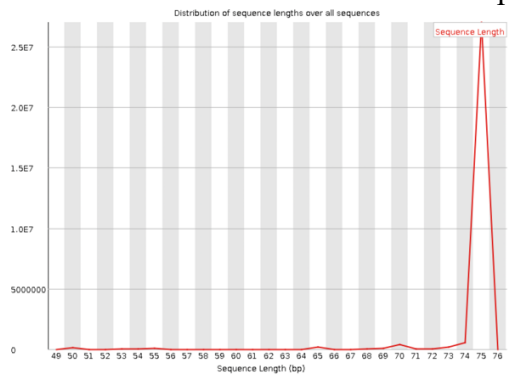


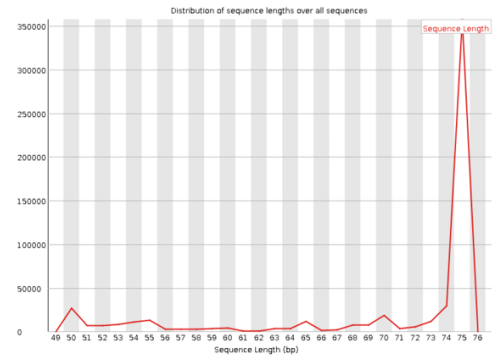


Обратное

Длина чтений: у парного появились небольшие пики на 70, а у непарного сильные пики на длине меньше 75 (основная длина в 75 нуклеотидов сохранилась). Чтения меньшей длиной после триммирования получаются из-за обрезания некачественных концов чтений, так что такой результат вполне ожидаем.

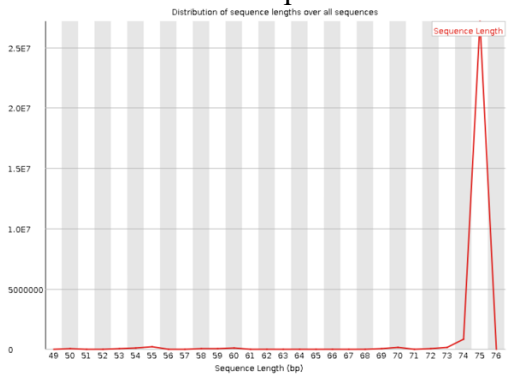
Прямое

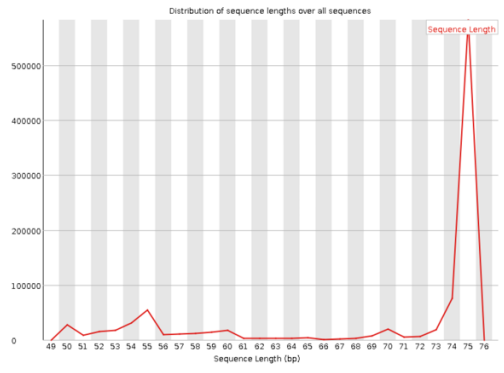




Парное

Непарное





Обратное

Рис 4. Сравнение длин после триммирования

ПРАКТИКУМ 12

Картирование чтений на референсный геном

Команда: hisat2 -x chr8 -1 trim_forward_paired.fastq.gz -2 trim_reverse_paired.fastq.gz -p 10 --no-spliced-alignment > cart.sam 2> cart.txt

-x задает префикс (chr8), полученный при индексации референса
-1 trim_forward_paired.fastq.gz файл с прямыми парными триммированными чтениями
-2 trim_reverse_paired.fastq.gz файл с прямыми парными триммированными чтениями
-p задает количество ядер процессора
--no splice-alignment параметр, запрещающий сплайсинг
> cart.sam запись выхода программы в файл sam
2> cart.txt сохранение логов

Конвертация sam в bam:

Команда: samtools sort -o cart.bam cart.sam
Samtools sort сортирует sam файл
-o вывод программы в файл bam
Вес файла sam: 12,1 Гб
Вес файла bam: 3,69 Гб
Индексируем: samtools index cart.bam
На выходе получаем файл cart.bam.bai

Анализ bam файла:

Команда: samtools flagstat cart.bam > analizedbam.txt

```
59950165 + 0 in total (QC-passed reads + QC-failed reads)
59252512 + 0 primary
697653 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
3641622 + 0 mapped (6.07% : N/A)
2943969 + 0 primary mapped (4.97% : N/A)
59252512 + 0 paired in sequencing
29626256 + 0 read1
29626256 + 0 read2
2481724 + 0 properly paired (4.19% : N/A)
2576160 + 0 with itself and mate mapped
367809 + 0 singletons (0.62% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Рис 5. Содержание файла analizedbam.txt

Чтений картировано: 3641622
В процентах от тримированных: 6.07%
Картировано в корректных парах: 2481724
В процентах от тримированных: 4.19%

Причины, почему чтения могли неправильно картироваться: адаптеры, мутации и вариабельность, высокая гомология геномов.

Получение чтений, картированных на хромосому:

Команда: samtools view -h -bS cart.bam 8 > chr8_cart.bam
Samtools view вывод всех чтений картированных на референс
-h выводит файл с заголовками
-b выводит файл с bam
-S автоматически определяет формат файла
8 имя хромомы (получила в 11 практикуме)
Получили файл chr8_cart.bam

Получение только правильно картированных пар чтений:
Команда: samtools view -f 0x2 -bS chr8_cart.bam > corpairs.bam
Samtools view вывод всех чтений картированных на референс
-f 0x2 отбор чтений по кртерию FLAG со значением 0x2, соответствующее PROPER_PAIR (только выровненные с референсом)
-b вывод в формате bam
-S автоматически определяет формат файла
Получили файл corpairs.bam, содержащий только правильно картированные попарно чтения.

Через samtools flagstat читаем его:

```
2867186 + 0 in total (QC-passed reads + QC-failed reads)
2481724 + 0 primary
385462 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2867186 + 0 mapped (100.00% : N/A)
2481724 + 0 primary mapped (100.00% : N/A)
2481724 + 0 paired in sequencing
1240862 + 0 read1
1240862 + 0 read2
2481724 + 0 properly paired (100.00% : N/A)
2481724 + 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)
```

Рис 6. Файл corpairs.bam

Картировано на референс корректных пар: 2481724
В процентах: 100%

Индексация полученного файла:
Команда: samtools index corpairs.bam
Получаем файл corpairs.bam.bai

ПРАКТИКУМ 13

Получение вариантов:

Команда: bcftools mpileup -f ../pr11/Homo_sapiens.GRCh38.dna.chromosome.8.fa corpairs.bam | bcftools call -mv -o var.vcf

Bcftools mpileup генерирует файл с вероятностями разных вариантов

-f указывает референс

Bcftools call берет из STDOUT mpileup только строки с характеристиками опций:

-m ищет редкие варианты

-v в выдачу только варианты

-o выдача файла

На выходе получаем файл var.vcf

Посмотрим этот файл:

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.11+htslib-1.11-4
##bcftoolsCommand=mpileup -f ../pr11/Homo_sapiens.GRCh38.dna.chromosome.8.fa corpairs.bam
##reference=file://../pr11/Homo_sapiens.GRCh38.dna.chromosome.8.fa
##contig=<ID=8,length=145138636>
##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=Tue Feb  6 23:13:45 2024
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  corpairs.bam
8       60168   .       C       T       24.0788 .       DP=3;VDB=0.42;SGB=-0.453602;RPB=1;MQB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,2,0;MQ=60  GT:PL   0/1:57,0,31
8       60181   .       G       A       67      .       DP=3;VDB=0.354794;SGB=-0.511536;MQ0F=0;AC=2;AN=2;DP4=0,0,3,0;MQ=60      GT:PL   1/1:97,9,0
8       60323   .       C       G       51      .       DP=5;VDB=0.28399;SGB=-0.511536;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,0,3;MQ=60        GT:PL   0/1:85,0,23
8       60345   .       A       G       69      .       DP=7;VDB=0.193383;SGB=-0.556411;RPB=0.202566;MQB=1.01283;MQSB=1.01283;BQB=0.810265;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=3,0,0,4;MQ=60     GT:PL   0/1:102,0,75
8       60408   .       G       A       50      .       DP=3;VDB=0.941105;SGB=-0.511536;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,2,1;MQ=60      GT:PL   1/1:80,9,0
8       60422   .       C       T       48.4146 .       DP=2;VDB=0.22;SGB=-0.453602;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,1,1;MQ=60   GT:PL   1/1:78,6,0
8       60565   .       C       A       35.9407 .       DP=3;VDB=0.58;SGB=-0.453602;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=0,1,1,1;MQ=60   GT:PL   0/1:69,0,31
8       61316   .       A       C       3.73859 .       DP=2;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,1,0;MQ=60   GT:PL   0/1:34,0,23
8       61373   .       A       G       10.7923 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:40,3,0
8       61501   .       C       T       7.30814 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:36,3,0
8       61780   .       T       A       3.75893 .       DP=2;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,1,0;MQ=60   GT:PL   0/1:34,0,27
8       61793   .       C       G       3.77146 .       DP=2;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,1,0;MQ=60   GT:PL   0/1:34,0,32
8       61858   .       A       G       41.4148 .       DP=2;VDB=0.26;SGB=-0.453602;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,1,1;MQ=60   GT:PL   1/1:71,6,0
8       61892   .       T       A       17.1192 .       DP=7;VDB=0.102722;SGB=-0.511536;RPB=1.01283;MQB=1.01283;MQSB=1;BQB=0.405132;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=3,1,2,1;MQ=60    GT:PL   0/1:50,0,103
8       61945   .       A       G       52      .       DP=5;VDB=0.764235;SGB=-0.511536;RPB=0.333333;MQB=1;MQSB=1;BQB=0.666667;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,1,2,1;MQ=60 GT:PL   0/1:85,0,51
8       62000   .       T       G       10.7923 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:40,3,0
8       62193   .       A       T       65      .       DP=4;VDB=0.845642;SGB=-0.511536;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=0,1,1,2;MQ=60       GT:PL   0/1:99,0,26
8       62351   .       T       G       37.8703 .       DP=3;VDB=0.52;SGB=-0.453602;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,1,1;MQ=60   GT:PL   0/1:71,0,30
8       62536   .       A       G       9.88514 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:39,3,0
8       62548   .       A       C       10.7923 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:40,3,0
8       62580   .       C       T       37.6817 .       DP=3;VDB=0.18;SGB=-0.453602;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=1,0,1,1;MQ=60   GT:PL   0/1:71,0,25
8       62737   .       G       T       26.4123 .       DP=6;VDB=0.58;SGB=-0.453602;RPB=0.666667;MQB=1;MQSB=1;BQB=0;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=2,1,0,2;MQ=60    GT:PL   0/1:59,0,60
8       63390   .       T       C       8.99921 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:38,3,0
8       63672   .       G       A       3.22451 .       DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60   GT:PL   1/1:30,3,0
```

Рис 7. Содержимое файла var.vcf

CHROM- имя хромосомы

POS- позиция варианта

ID- номер варианта

REF- референсные нуклеотиды если индель

ALT- список аллельных вариантов

QUAL- качество варианта (зависимость от частот аллелей)

FILTER- флаг, указывающий по какому фильтру прошел вариант

INFO- описание варианта

FORMAT- описание образца

С помощью bcftools stats var.vcf > varstats.txt анализируем варианты

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      1
SN      0       number of records:      58885
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 57415
SN      0       number of MNPs: 0
SN      0       number of indels:       1470
SN      0       number of others:       0
SN      0       number of multiallelic sites:   25
SN      0       number of multiallelic SNP sites:       24
```

Рис 8. Varstats.txt

Вариантов: 58885
Варианты SNP: 57415
Индели: 1470

Фильтрация вариантов:
Команда: bcftools filter -i'%QUAL>30 && DP>50' var.vcf -o filtervar.vcf
Bcftools filter фильтрует варианты по заданным параметрам
-I'%QUAL>30 && DP>50' задает параметры «качество больше 30» и «длина больше 50»
-o STDOUT в указанный файл
Получаем файл filtervar.vcf с отфильтрованными вариантами и анализируем (через stats):

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      1
SN      0       number of records:      1325
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 1302
SN      0       number of MNPs: 0
SN      0       number of indels:       23
SN      0       number of others:       0
SN      0       number of multiallelic sites:   2
SN      0       number of multiallelic SNP sites:       2
```

Рис 9. Файл filtervar.vcf

После фильтрации вариантов: 1325 (2,25%)
SNP: 1302 (2,27%)
Инделей: 23 (1,56%)

Аннотация вариантов:
Используем VEP, вносим файл из предыдущего пункта.

Category	Count
Variants processed	1325
Variants filtered out	0
Novel / existing variants	246 (18.6) / 1079 (81.4)
Overlapped genes	424
Overlapped transcripts	2156
Overlapped regulatory features	95

Рис 10. Общая характеристика

Все 1325 варианта были проаннотированы, из них 246 новые, а 1079 уже использованные где-то. 424 перекрывающих генов, 2156 перекрывающихся транскриптов, 95 перекрывающихся регулярных областей.

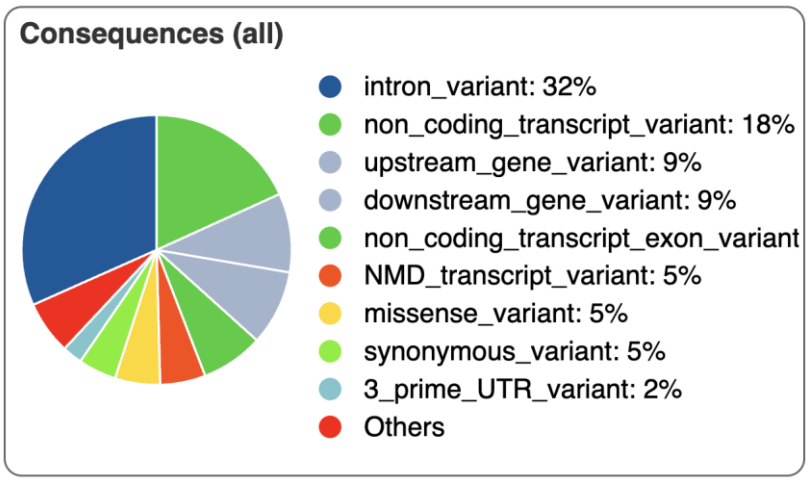


Рис 11. Распределение мутаций

Большинство мутаций, очевидно, не влияют на кодирующие области хромосом (иначе возникали бы болезни). Влияют на интроны, некодирующие транскрипты и экзоны и др.
5% вариантов вызывают точечные мутации, в результате которых изменяется кодируемая аминоксилота. Это влияет на структуру и функцию белка и может быть причиной болезней.
Еще 5% вариантов вызывают синонимичные мутации (замены нт в кодирующей части, при которых аминокислота не изменяется), в результате которых, например, может снизится скорость трансляции и из-за этого нарушится структура белка.

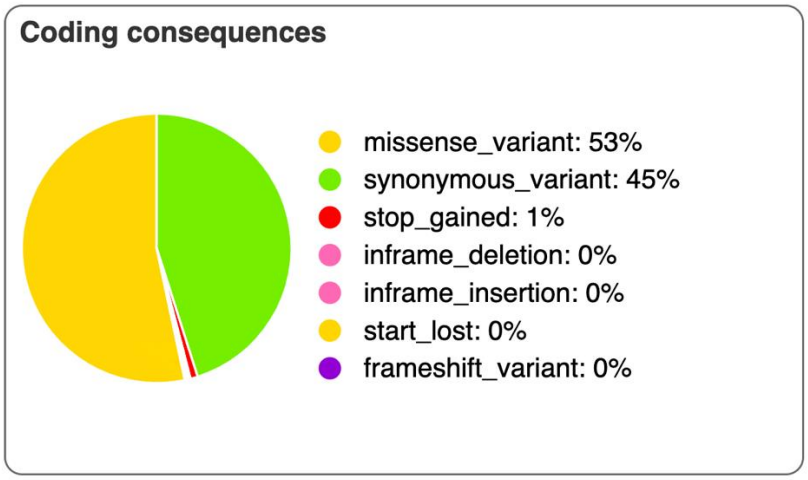


Рис 12. Мутации кодирующих последовательностей

В этом графике появляется stop gained- мутация, изменяющая основания, приводящая к образованию преждевременного стоп-кодона. Логично, это останавливает трансляцию мРНК, на которой в свою очередь собирается неправильный белок. Остальные варианты мутаций суммарно составляют 1%, поэтому их влияние (вероятно) мало.

HIGH IMPACT:
Stop_gained- 9
Stop_gained+NMD_transcript_variant 1
Frameshift_variant 1
Splice_donor_variant 2
Splice_acceptor_variant+non_coding_transcript_variant 2
Start_lost 1

https://www.ensembl.org/Homo_sapiens/Tools/VEP/Results?tl=TTcoCgTtg2PbDbHx-9929763

ПРАКТИКУМ 14

ID образца: **ENCFF038OLY**
Ссылка на информацию об образце:
https://www.encodeproject.org/search/?type=Experiment&searchTerm=ENCFF038OLY
Организм и ткань: мышечная ткань ноги эмбриона человека
Секвенирование: polyA plus RNA-seq
Тип чтений: SE
Цепь специфичность: нет

Проверка качества исходных чтений:
Команда: fastqc ENCFF038OLY.fastq.gz
На выходе получаем файл в формате html:
Количество чтений: 72517664
Качество чтений: среднее значение и медиана в зеленой зоне (хорошие), но усы сильно опущены в красную.

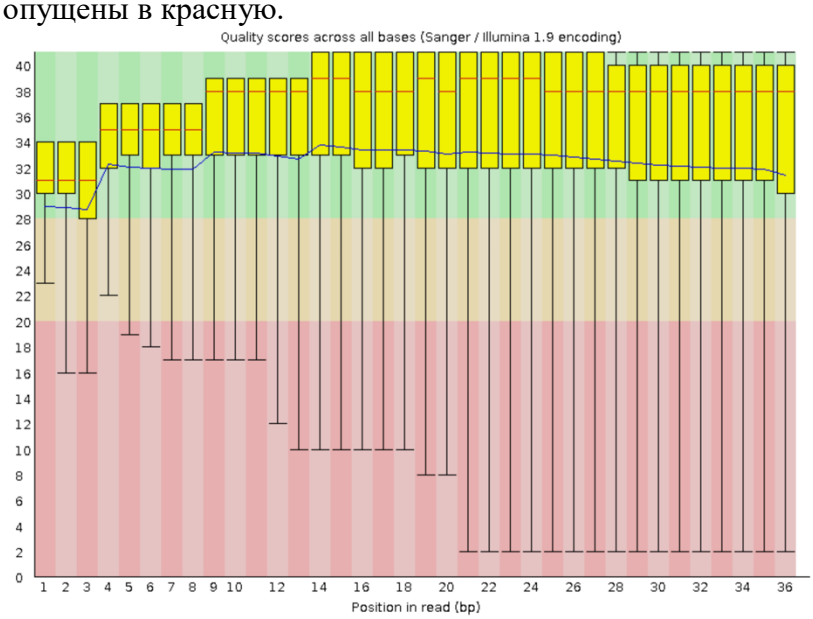


Рис 13. Quality score

Из-за большого количества усов оценим качество по последовательностям:

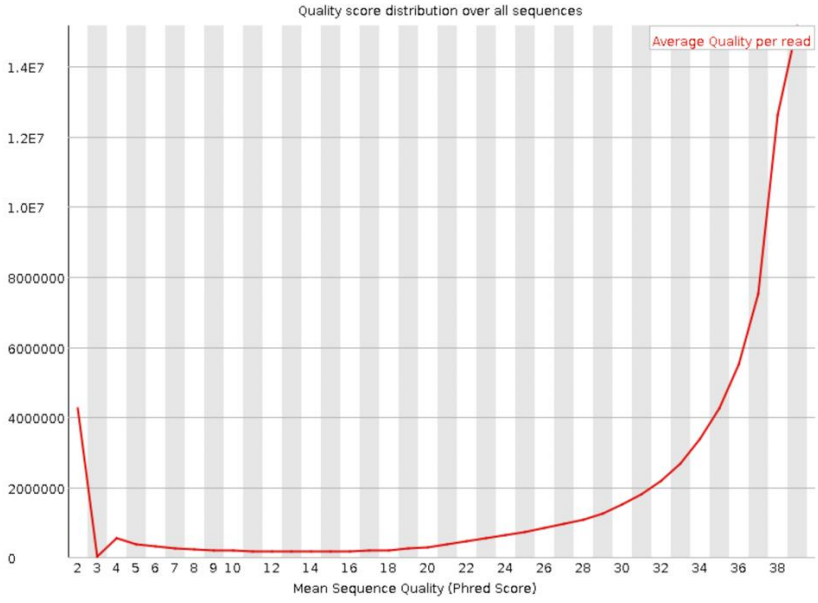


Рис 14. Quality score distribution over all sequences

Много чтений с совсем низким качеством (2), около 5,65% от всех чтений, они могут повлиять на картирование.

Длина чтений: у всех длина 36

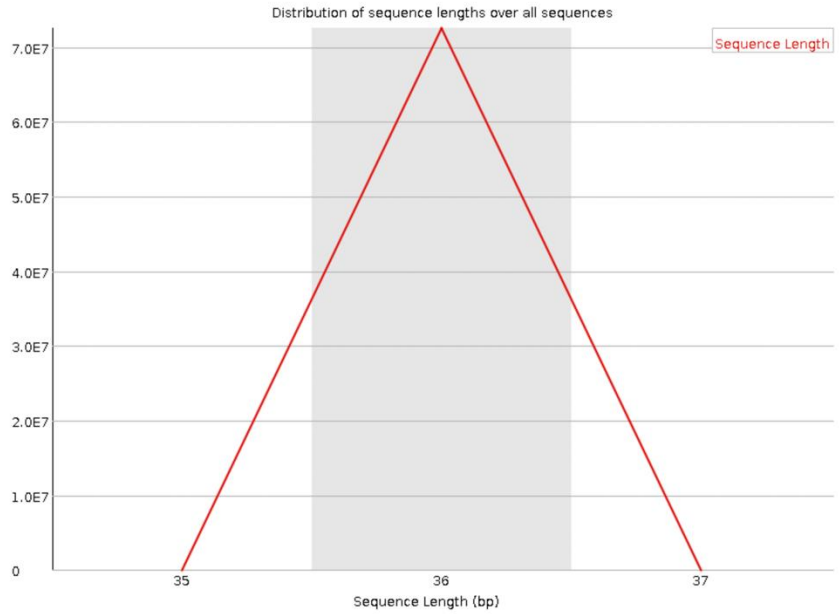


Рис 15. Длина последовательностей

Картирование чтений на референс:

Команда: hisat2 -x ../pr11/cr8 -k 3 -U ENCFF038OLY.fastq.gz > rnacart.sam 2> rnacart.txt

Hisat2- картирование чтения, -x ../pr11/chr8 для использования индексированного референса, -k 3 обеспечивает максимальное количество выравниваний (3), -U входной файл с РНК-чтениями

Получаем два файла: в .sam вывод программы, в .txt логи

Посмотрим rnacart.txt:

```
72517664 reads; of these:
  72517664 (100.00%) were unpaired; of these:
    68788319 (94.86%) aligned 0 times
    3015427 (4.16%) aligned exactly 1 time
    713918 (0.98%) aligned >1 times
5.14% overall alignment rate
```

Рис 16. Файл rnacart.txt

Картировалось 372740 чтение (5,14%)

Поиск экспрессирующихся генов:

В файле Homo_sapiens.GRCh38.110.chr.gtf с геномной разметкой:

```
#!genome-build GRCh38.p14
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession GCA_000001405.29
#!genebuild-last-updated 2023-03
```

Рис 17. Файл с референсной хромосомой

Версия разметки//версия генома//дата публикации//код доступа (AC)//последняя дата изменения

Дальше девять характеристик:

seqname – название последовательности, где аннотирован ген

source - название программы, которая аннотировала (название базы данных)

feature – вид особенности (ген или вариация)

start – начало гена (если нумерация последовательности с 1)

end – конец гена

score – оценка качества (значение с плавающей запятой)

strand – цепь, на которой находится ген (+ или -)

frame – рамка считывания, где первая- 0.

attribute – дополнительная информация о функциях

Сколько генов аннотировано на восьмую хромосому?
Команда: grep '^8' *gtf | cut -f3 | grep 'gene' | wc -l
Output: 2541

Считаем для каждого гена из разметки число картированных на этот ген чтений:
Команда: htseq-count -f bam -s no -t exon -m union -o exons.sam chrrna.bam Homo_sapiens.GRCh38.110.chr.gtf 1> exons1.txt 2> exons2.txt
Htseq-cout картирует чтения, -f bam формат входного файла, -s no дает вариабельность цепи (тк не указана), -m union объединяет перекрывшиеся чтения, -t exon только чтения которые картировались на экзоны, -o exon.sam output файл, дальше сохраняю логи 1 сводка о программе (stdout), 2 ошибки (stderr)

Exons1.txt (в конце файла):

```
__no_feature    817068
__ambiguous     116958
__too_low_aQual 0
__not_aligned   0
__alignment_not_unique  713918
```

Рис 18. Выход htseq-count

Чтений в границах: считаю с помощью скрипта на python:

```
cnt = 0
with open('/Users/habibulina/Downloads/exons1.txt', mode='r') as file:
 for line in file:
   if not line.startswith('__'):
      cnt += int(line.split('\t')[1])
print (cnt)
```

Получилось 2081401 чтений.
Чтений не в границах: 817068

СКРИПТ

script.sh

```
#!/bin/bash

# Usage script.sh config_file ID N

##############################################
# Soft
##############################################
# FastQC v0.11.9
# hisat2 version 2.2.1
# TrimmomaticPE 0.39
# multiqc version 1.15
# samtools 1.17 (using htslib 1.17)
# bcftools 1.11 (using htslib 1.11-4)
##############################################
config_file=$1
ID=$2
chrN=$3

. $config_file

hisat2-build $ref.fa $ref #индексация референса на chrN
samtools faidx $ref.fa #индексация референса
fastqc $forward_reads $reverse_reads #проверка качества исходных прямых и обратных чтений
TrimmomaticPE -$phred $forward_reads $reverse_reads $for_paired $for_unpaired $rev_paired
$rev_unpaired TRAILING:$trim_trailing MINLEN:$trim_minlen #триммирование чтений (удаляем концы
с каечством ниже trim_trailing и длиной короче trim_minlen)
fastqc $for_paired $for_unpaired $rev_paired $rev_unpaired #анализ триммированных чтений
hisat2 -x $ref -1 $for_paired -2 $rev_paired -p $nthreads --no-spliced-alignment > $map_sam 2>
$map_logs #картирование чтений на референс chrN и запрет сплайсинга
samtools sort -o $map_bam $map_sam #конвертация sam в bam
samtools index $map_bam #индексация bam файла
samtools view -h -bS $map_bam $chrN > $true_map #получение чтений, которые картировались на
референс
samtools view -f 0x2 -bS $right_paired #получение только правильно картированных
samtools index $right_paired #индексация правильные
bcftools mpileup -f $ref.fa $right_paired | bcftools call -mv -o $var_vcf #поиск вариантов
bcftools filter -i'%QUAL>{quality} && DP>{deep}' $var_vcf -o $filtervar_vcf #фильтрация по заданным
условиям (качество больше quality и глибина больше deep
```

config_file

```
ref=$chrN #референсная хромосома
forward_reads=${ID}_1.fastq.gz #прямые чтения
reverse_reads=${ID}_2.fastq.gz #обратные чтения
phred=phred33 #качество для триммирования
trim_trailing=20 #обрезание концов с качеством меньше 20
trim_minlen=50 #задание минимальной длины чтений после триммирования
for_paired=trim_1_paired.fastq.gz #триммирование прямых парных
for_unpaired=trim_1_unpaired.fastq.gz #триммирование прямых непарных
rev_paired=trim_2_paired.fastq.gz #триммирование обратных парных
rev_unpaired=trim_2_unpaired.fastq.gz #триммирование обратных непарных
nthreads=10 #ядра процессора
map_sam=map.sam #sam файл с результатом картирования
map_logs=map_logs.txt #stderr картирования
map_bam=map.bam #конвертация sam в bam
true_map=$ref_map.bam #картированные чтения
```

```
right_paired=right_pairs_$ref_map.bam #правильно картированные чтения
var_vcf=var.vcf #все варианты
quality=30 #качество вариантов
deep=50 #глубина вариантов
filtervar_vcf=filtervar.vcf #фильтрованные варианты
```