Практикумы 11–14

Бахмарин Степан, группа 202

Практикум 11

Подготовка референса

Индексация для hisat2

Внутри своей директории для выполнения этого практикума (/mnt/scratch/NGS/bakhsv) создал директорию ref, в нее скопировал файл с хромосомой 3. Запустил программу для индексирования hisat2:

```
hisat2-build Homo_sapiens.GRCh38.dna.chromosome.3.fa chr3
```

Первый аргумент (Homo_sapiens.GRCh38.dna.chromosome.3.fa) — FASTA-файл с третьей хромосомой, второй аргумент (chr3) — префикс, который будет использоваться в названиях файлов индекса. Программа создала восемь файлов в той же директории ref (chr3.n.ht2, n — от одного до восьми). Это бинарные файлы, заглянуть в них не получилось.

Индексация samtools

В той же директории запустил программу для индексации samtools:

```
samtools faidx Homo_sapiens.GRCh38.dna.chromosome.3.fa
```

Получившийся файл Homo_sapiens.GRCh38.dna.chromosome.3.fa.fai содержит пять чисел, разделенных табуляцией:

3	198295559	56	60	61

“3” — это имя последовательности в FASTA-файле, 198 295 559 — длина последовательности, 56 — номер байта, начиная с нуля, с которого начинается последовательность в файле, 60 — количество букв в одной строке файла, 61 — количество байт в одной строке (включая символ переноса строки.

Чтения ДНК

Описание образца

Создал директорию reads/original, положил в нее нужные файлы с прямыми и обратными чтениями. В табл. 1 содержится описание этих чтений.

Табл. 1. Описание образца

SRR ID	SRR10720402
Информация об образце на сайте NCBI. Первая ссылка — страница, посвященная образцу, вторая — эксперименту. В них содержится похожая информация.	https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR10720402&display=metadata https://www.ncbi.nlm.nih.gov/sra/SRX7397223
Прибор	Illumina Genome Analyzer IIx
Организм	*Homo sapiens*
Стратегия секвенирования	Экзомное
Парно-/одноконцевые	Парноконцевые
Сколько чтений ожидается	28 966 798

Чтобы оценить качество чтений, запустил программу FastQC на прямых и обратных чтениях:

```
fastqc reads/SRR10720402_1.fastq.gz
fastqc reads/SRR10720402_2.fastq.gz
```

Получившиеся в результате работы программы файлы переложил в директорию fastqc_results.

Получилось 28 966 798 прямых и столько же обратных чтений, т. е. столько же, сколько ожидалось. Иллюстрации per base sequence quality приведены на рис. 1 и 2.

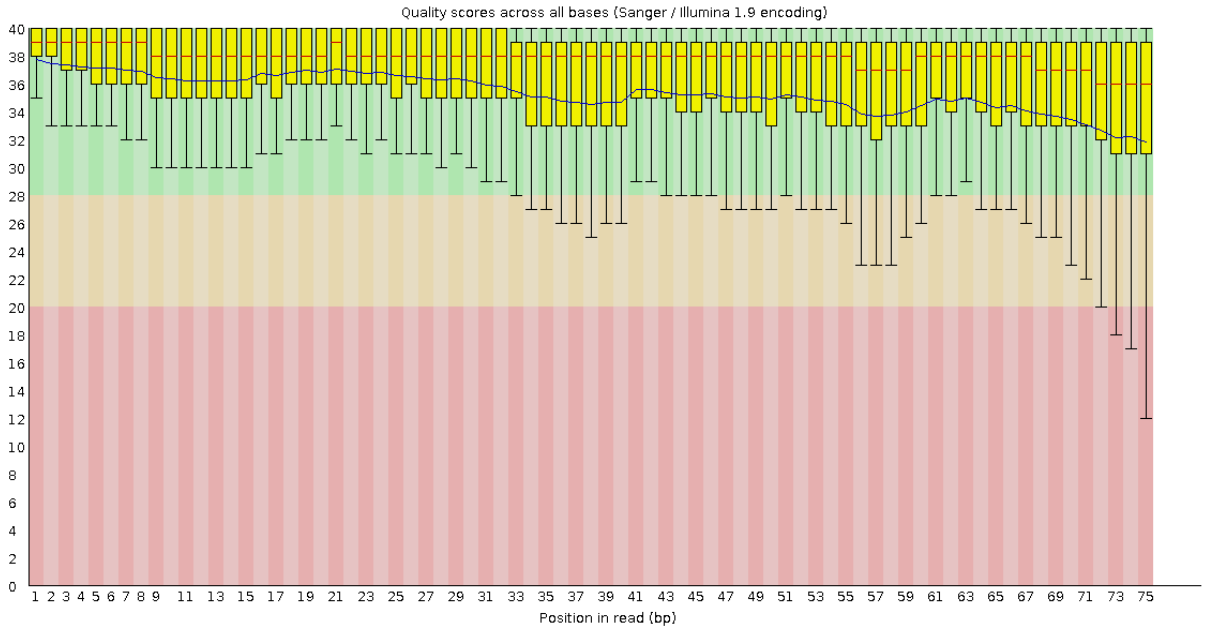

Рис. 1. Прямые чтения

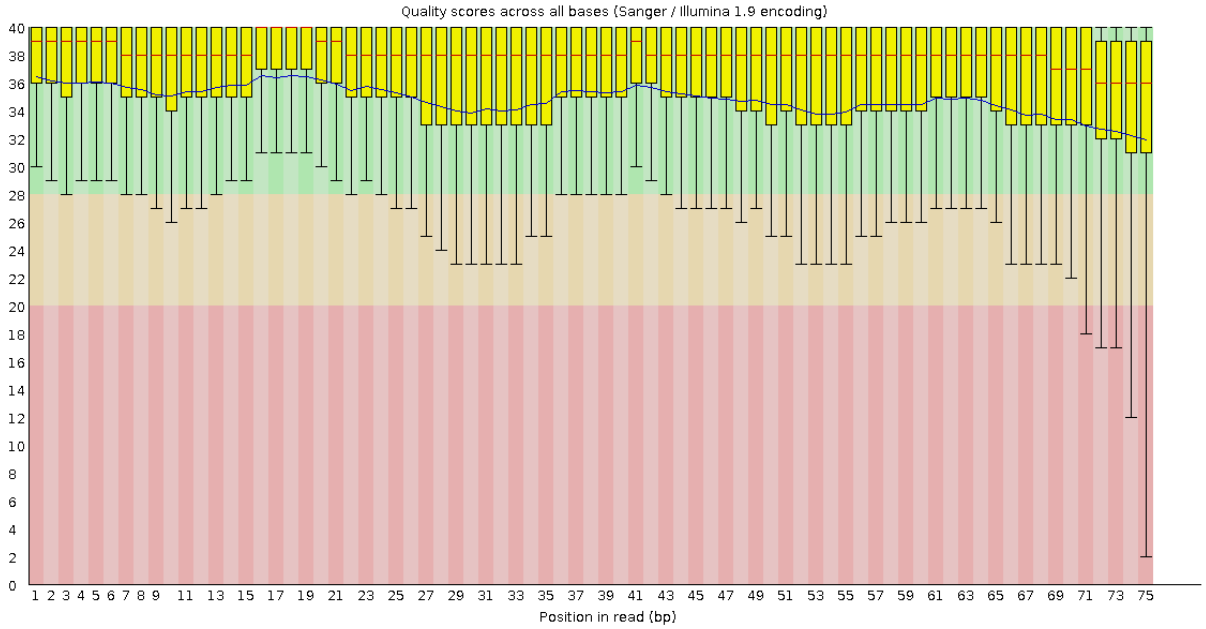

Рис. 2. Обратные чтения

Качество чтений немного падает к концу, после ≈70 нуклеотида появляется некоторая доля нуклеотидов с совсем низким качеством. В обратных чтениях 5-й процентиль пониже, чем в прямых, т.е. худшие обратные чтения хуже худших прямых, но верхний квартиль во второй половине чтения наоборот выше. В общем, я не сказал бы, что обратные чтения получились хуже прямых.
FastQC отметила этот параметр зеленой галочкой, т.е. нижний квартиль для любой позиции не меньше 10 и медиана для любой позиции не меньше 25 (на самом деле нижний квартиль для любой позиции не меньше даже 30).
Судя по этому параметру, чтения качественные.

Иллюстрации распределения длины чтений приведены на рис. 3 и 4.

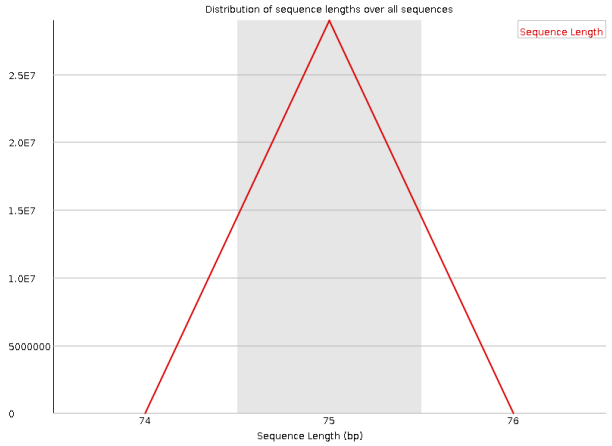

Рис. 3. Прямые чтения

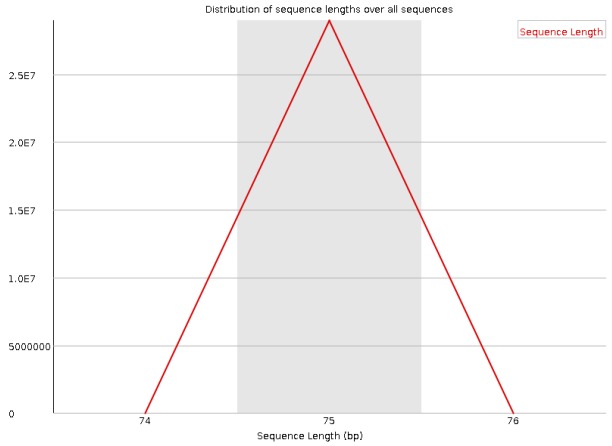

Рис. 4. Обратные чтения

Все чтения имеют длину 75.

Фильтрация чтений

Чтобы повысить качество чтений перед дальнейшими манипуляциями, их нужно отфильтровать. В задании требовалось удалить с конца чтений нуклеотиды с качеством ниже 20 и оставить чтения длиннее 50 нуклеотидов. Для этого запустил программу TrimmomaticPE:

```
TrimmomaticPE -phred33 -trimlog trimlog.txt
reads/original/SRR10720402_1.fastq.gz reads/original/SRR10720402_2.fastq.gz
reads/trimmed/fw_paired.fastq.gz reads/trimmed/fw_unpaired.fastq.gz
reads/trimmed/rev_paired.fastq.gz reads/trimmed/rev_unpaired.fastq.gz
TRAILING:20 MINLEN:50
```

Опция `-phred33` указывает, что качество записано в кодировке +33 (по умолчанию это +64). Опция `-trimlog` указывает адрес файла, в который будет записан лог. Затем шесть аргументов этой программы — это имена входных и выходных файлов. TRAIILING:20 указывает, что нужно удалить с конца нуклеотиды с качеством ниже 20, а MINLEN:50 — отставить после этого чтения не короче 50 нуклеотидов

(триммиорование происходит в том порядке, в каком шаги указаны при запуске программы).

Программа возвращает четыре файла, потому что одно чтение из пары может быть отброшено после триммирования (т.е. оказаться короче 50 нуклеотидов после удаления некачественных, в данном случае), и тогда второе окажется без пары. Программа выводит такие “распаренные” чтения в отдельные файлы.

Проверка качества триммированных чтений

Запустил FastQC на триммированных чтениях (в директории reads/trimmed):

```
fastqc * -o ../../fastqc_results/
```

Опция `-o` указывает, в какую директорию надо вывести результат.
Осталось 27 172 718 пар чтений (93,8 % от исходного количества).
Неспаренных прямых чтений оказалось 1 197 393, а неспаренных обратных — 473 184, т. е. из общего числа отфильтрованных пар в 66,7 % случаев программа убрала обратное чтение, в 26,4 % — прямое, а оставшихся 6,9 % — оба.

Если (абстрактный) фактор, ухудшающий качество чтения, действует на чтения независимо, т.е. вероятность, что данное чтение будет отброшено, не зависит от того, отброшено ли парное ему, то мы ожидаем, что качество неспаренных чтений будет таким же, как у парных, а если зависимо — то хуже. Я посчитал соответствующий коэффициент корреляции, он равен 0,11. Я ожидал, что будет он больше, но, кажется, на такой выборке он должен быть значимым (я не ожидал), так что мы ожидаем, что качество неспаренных чтений будет хуже.

Per base quality показано на рис. 5–8.

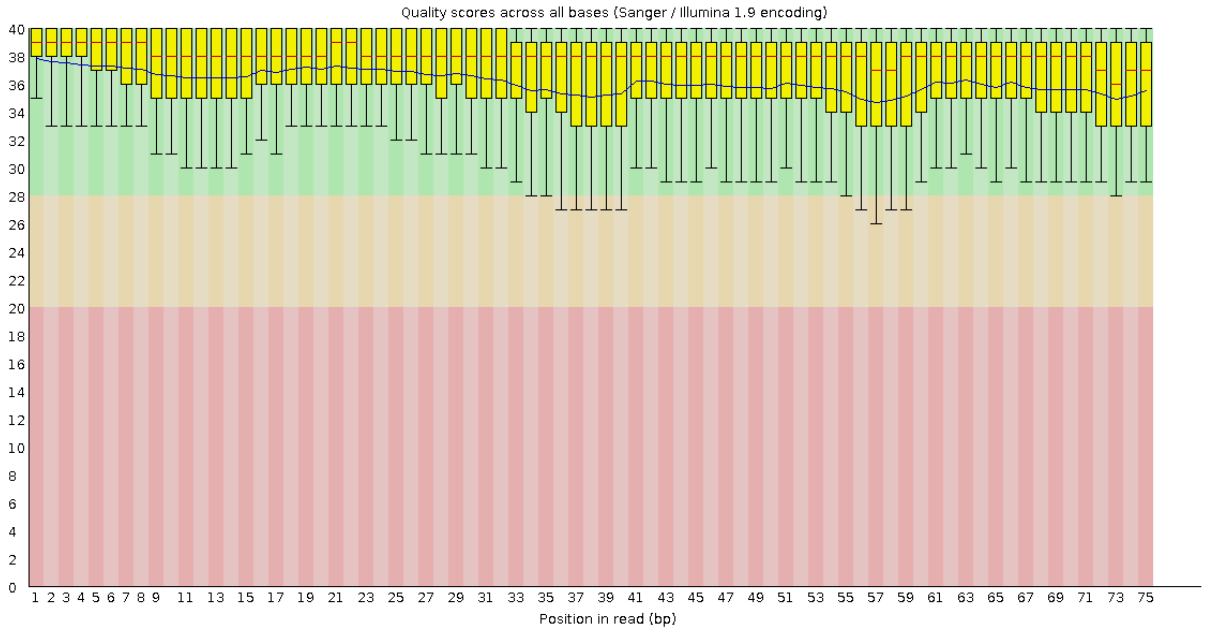

Рис. 5. Прямые парные чтения

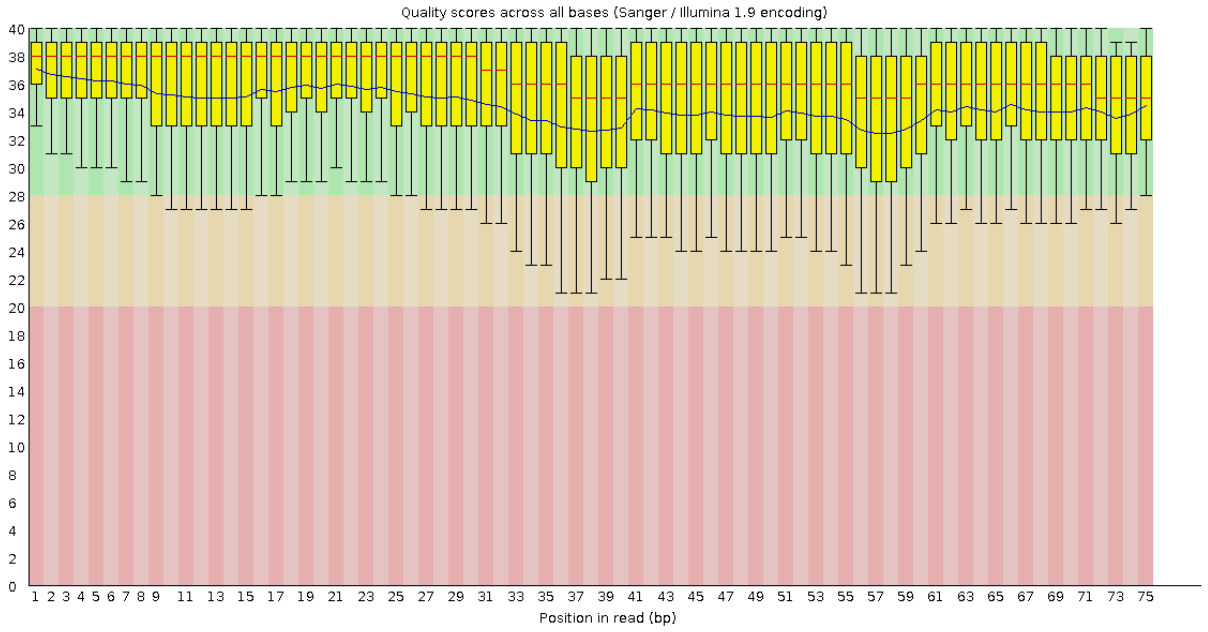

Рис. 6. Прямые неспаренные чтения

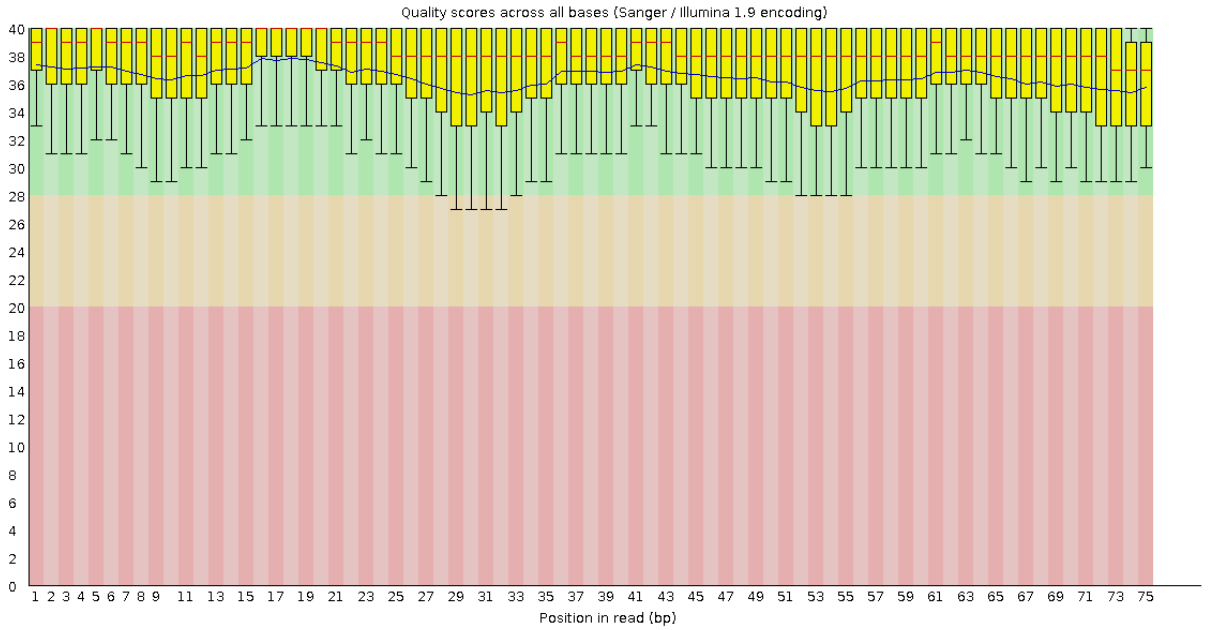

Рис. 7. Обратные парные чтения

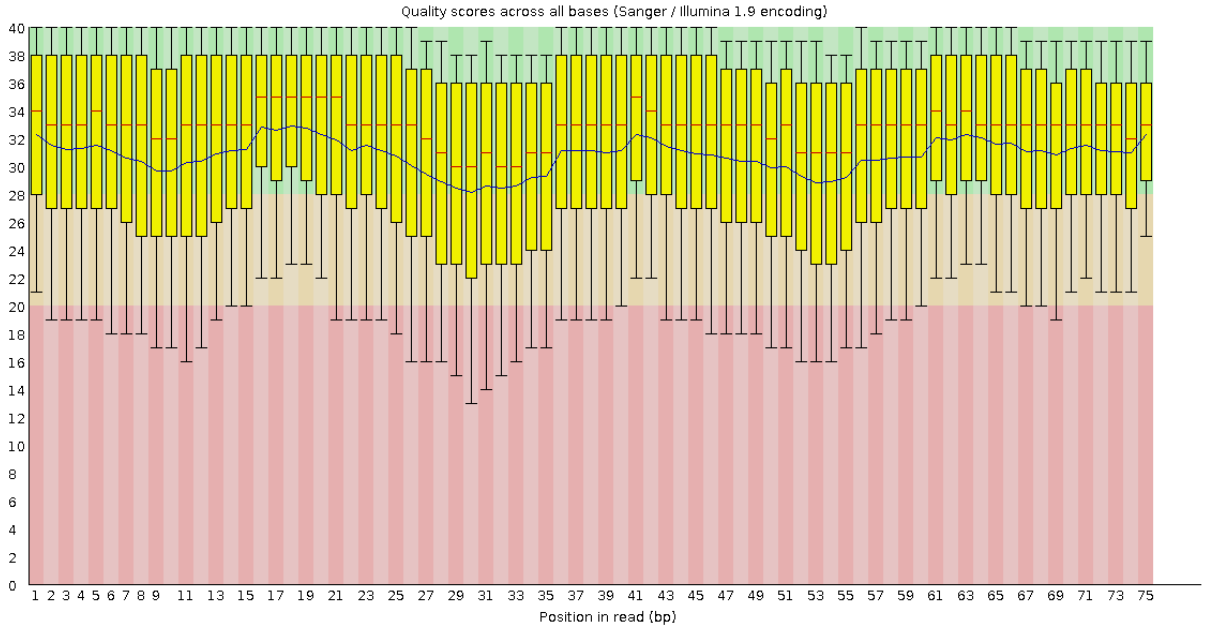

Рис. 8. Обратные неспаренные чтения

Видно, что обратные неспаренные чтения выглядят хуже всего, т.е. если прямое чтение было некачественным (настолько, что оказалось отброшено), то и обратное будет скорее некачественным. Прямые неспаренные чтения выглядят лучше, т. е. обратное верно в меньшей степени.

Парные чтения после триммирования выглядят совсем хорошо — исчезла “борода” в конце.

Распределение длин парных чтений после триммирования показано на рис. 9–10.

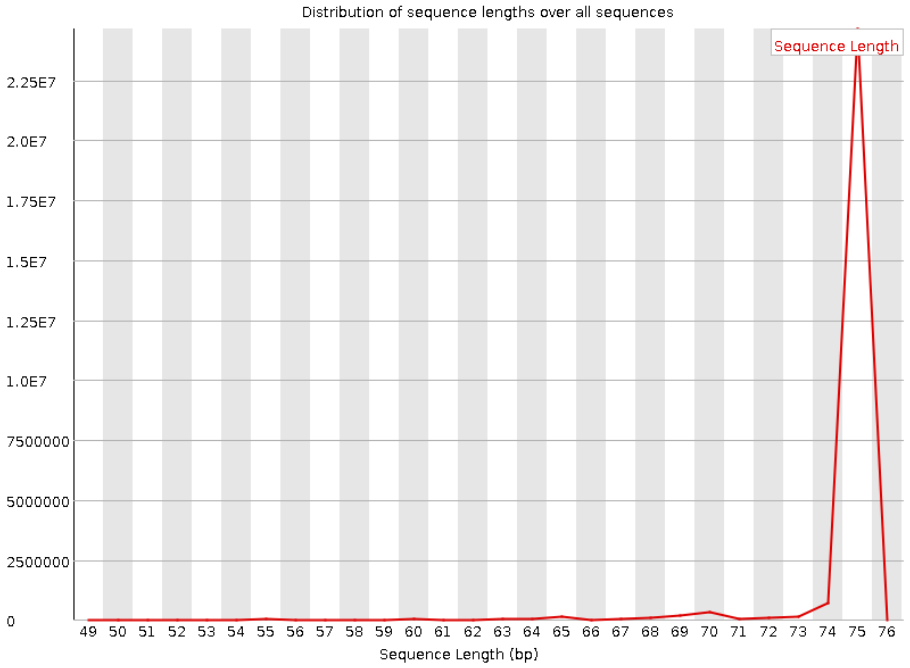

Рис. 9. Прямые чтения

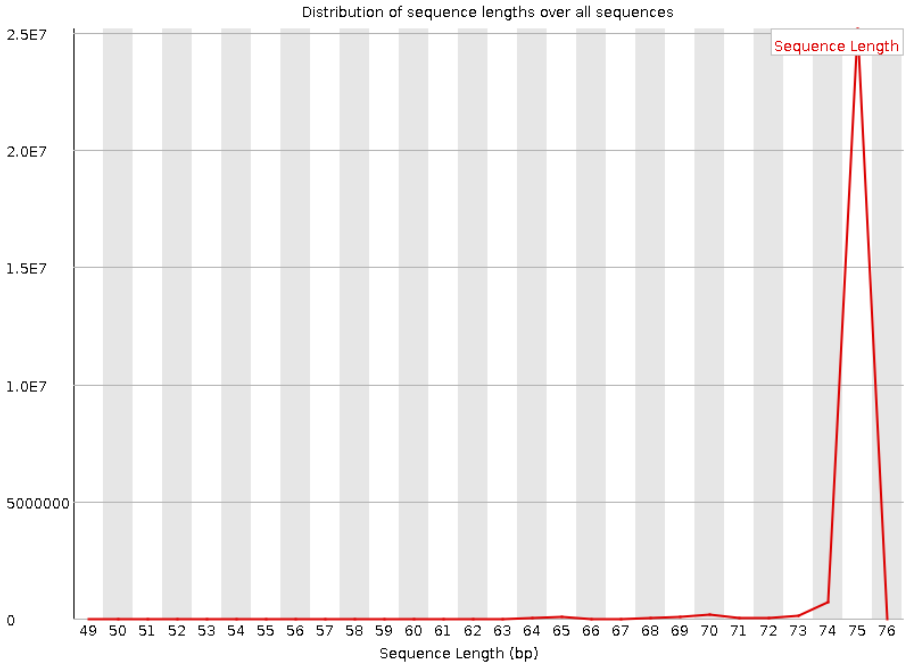

Рис. 10. Обратные чтения

После триммирования появилась небольшая доля чтений длиной 50–74.

Практикум 12

Картирование чтений на референс

Дальше нужно выровнять чтения на референсный геном. Для этого запустил картировщик:

```
hisat2 -x chr3 -1 ../reads/trimmed/fw_paired.fastq.gz -2
../reads/trimmed/rev_paired.fastq.gz -p 4 --no-spliced-alignment -S
../chr3.sam -t 2> ../logs_etc/hisat_log.txt
```

-x указывает префикс, использованный при индексации, -1 и -2 — прямые и обратные чтения, -p — количество потоков, на которые программа распараллеливает вычисления (в чате Иван Сергеевич писал, что больше 4-х ядер система не выделит, поэтому и указал 4), --no-spliced-alignment запрещает сплайсовать чтения, -S — адрес выходного файла, -t говорит программе печатать в stderr время, за которое были выполнены разные операции. В задании Вы просили собрать логи; я не уверен, что верно понял, что Вы имели в виду. Я сохранил все, что программа печатала в stderr, в файл hisat_log.txt.

Анализ и фильтрация

Получившийся SAM весит 11 ГБ. Конвертировал его в BAM, SAM удалил.

```
samtools sort -o chr3.bam chr3.sam
```

Получившийся BAM весит 3,2 Гб. Проиндексировал его, чтобы дальше программы из samtools могли с ним работать:

```
samtools index chr3.bam
```

Чтобы понять, все ли прошло нормально, и что записалось в получившийся BAM-файл, запустил программу samtools flagstat:

```
samtools flagstat chr3.bam > ch3_flagstat.txt
```

4 021 282 чтения (7,36 % от оставшихся после триммирования) картировано. 3 234 406 (5,95 %) картировано в правильных парах.

Посмотрел в Ensembl, какую долю генома составляет третья хромосома. Это 6,4 %. Значит, если здесь нет каких-то искажений (biases … как перевести это на русский?), то третья хромосома чуть насыщеннее генами, чем “средняя” хромосома. Проверим это. Оказалось, нет, и даже наоборот — на третьей хромосоме находится 5,45 % кодирующих генов и 5,37 % “некодирующих генов”. Возможно, большая доля чтений

картировалась из-за повторов, присутствующих на разных хромосомах, и, если картировать на весь геном, то сумма по отдельным хромосомам будет превышать 100%. К сожалению, сейчас нет времени пытаться выяснить подробнее. Затем оставил только чтения, картированные на третью хромосому (удалил некартировавшиеся):

```
samtools view -b -h chr3.bam 3 > only_chr3.bam
```

Опция -b указывает, что выдача нужна в формате BAM, -h добавляет заголовок. В задании Вы приводили пример, в котором еще была указана опция -S, но в той версии samtools, которая стоит на kodomo (1.17) эта опция игнорируется, т.к. программа сама определяет формат входного файла (http://www.htslib.org/doc/1.17/samtools-view.html). Здесь “3” после имени файла — это имя хромосомы.

Оставил только чтения, картированные в правильных парах:

```
samtools view -f 0x2 -b only_chr3.bam > properly_paired_only_chr3.bam
```

Опция -f указывает, чтения с какими sam flags надо оставить. 0x… — число в шестнадцатеричном формате. -f 0x2 значит, что нужно оставить только те чтения, у которых в flag входит flag 2 — “read mapped in proper pair”.

Собрал статистику по этому файлу с помощью samtools flagstat:

```
samtools flagstat properly_paired_only_chr3.bam >
flagstat_properly_paired_only_chr3.txt
```

Всего осталось картировано на референс 3 234 406 чтения. В задании Вы просите указать, сколько чтений картировано в правильных парах от общего числа чтений (если я верно понял, в файле, получившемся после всех этих манипуляций). Так как в предыдущем пункте я оставлял только чтения, картированные в правильных парах, то, естественно, 100 %.

Практикум 13

Получение вариантов

Оставшиеся правильно картикрованные чтения нужно превратить в таблицу вариантов, отличающихся от референсного генома (в гомозиготном или гетерозиготном варианте).

Запустил bcftools:

```
bcftools mpileup -f ../ref/Homo_sapiens.GRCh38.dna.chromosome.3.fa
properly_paired_only_chr3.bam | bcftools call -mv -o chr3_vars.vcf
```

Первая программа, bcftools mpileup, определяет вероятности генотипов для каждой позиции в референсном геноме, bcftools call с опцией -v оставляет из них только вариабельные сайы. -m указывает, какой алгоритм для этого использовать (не разбирался в деталях), -o — адрес выходного файла.

Вот здесь: https://samtools.github.io/bcftools/howtos/variant-calling.html еще рекомендуют в таком пайплайне указать у bcftools mpileup опцию -Ou, чтобы избежать бессмысленной конвертации из бинарного формата в текстовый и обратно. При написании сценария укажу эту опцию.

Посмотрел статистику по получившемуся файлу:

```
bcftools stats  -F ../ref/Homo_sapiens.GRCh38.dna.chromosome.3.fa
chr3_vars.vcf
```

-F — референсный геном.
Получилось 76 354 варианта, из них 74 187 SNP и 2167 инделей (т. е. все варианты — либо SNP, либо индели).

Фильтрация вариантов

Дальше из полученных вариантов необходимо оставить только те, в которых мы достаточно уверены. Для этого отфильтровал варианты с помощью bcftools filter:

```
bcftools filter -i'%QUAL>30 && DP>50' chr3_vars.vcf > chr3_vars_filtered.vcf
```

Такой запуск программы оставит только варианты, вероятность неправильной интерпретации которых меньше 10^{-30} (QUAL — вероятность, записанная в стиле Phred score) и глубина покрытия которых не меньше 50.

Посмотрел статистику по оставшимся после фильтрации вариантам:

```
bcftools stats -F ../ref/Homo_sapiens.GRCh38.dna.chromosome.3.fa
chr3_vars_filtered.vcf > filtered_stats.vcf
```

Осталось 1541 SNP и 34 инделя. Кажется, это более реалистичное количество при сравнении двух человеческих экзомов, чем 75 000, которые были до фильтрауции.

Аннотация вариантов

Аннотировал варианты с помощью сервиса VEP. Вот ссылка на результат: https://www.ensembl.org/Homo_sapiens/Tools/VEP/Results?tl=N3NqhbiB3nvr9Pbh-9765927 . Я не нашел быстро у них на странице информацию, сколько он хранится, но, наверное, будет доступен по ссылке несколько дней.

В табл. 2 и на рис. 11 и 12 представлена основная статистика по полученной аннотации.

Табл. 2.

Variants processed	1575
Variants filtered out	0
Novel / existing variants	397 (25.2) / 1178 (74.8)
Overlapped genes	632
Overlapped transcripts	3378
Overlapped regulatory features	137

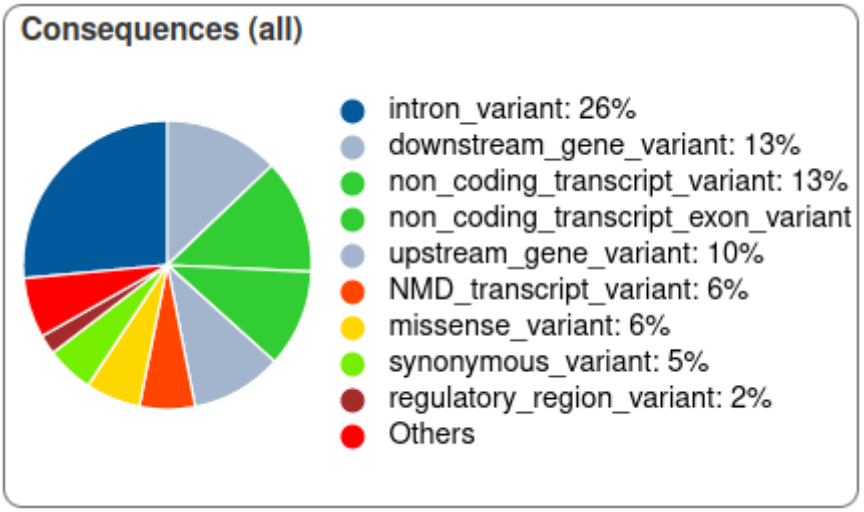

Рис 11. Распределение полученных вариантов по элементам генома

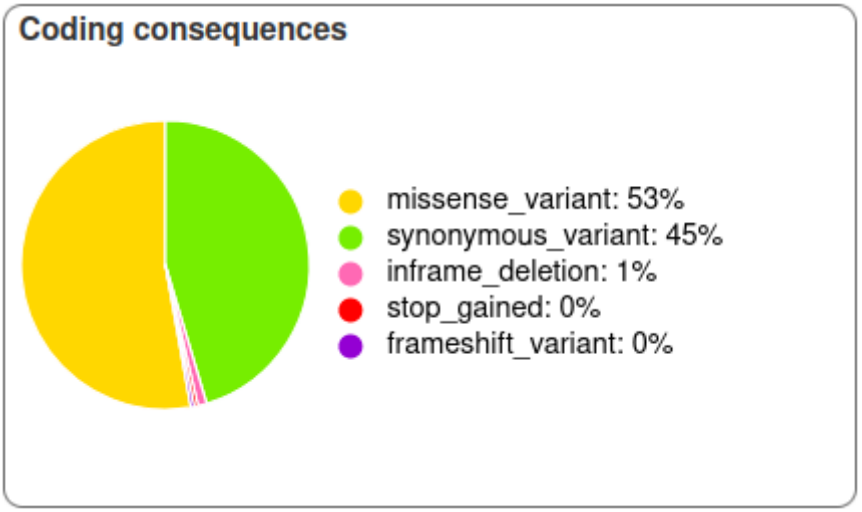

Рис. 12. Варианты в кодирующих последовательностях

Вариантов с IMPACT HIGH получилось 35. Среди них 7 вариантов в трех генах, приводящих к сдвигу рамки считывания, и 8 вариантов в трех генах, приводящих к появлению стоп-кодона в кодирующей последовательности (два гена перекрывается). Это гены PDCD6IP, ULK4, MST1, FRG2C. Очевидно, эти четыре гена не функционируют в секвенированном образце (как я понял, секвенировали раковую опухоль легкого). Остальные варианты с IMPACT HIGH — варианты сплайсинга.

Практикум 14

Описание образца

Описание образца приведено в табл. 3.

Табл. 3

ID	ENCFF038OLY
Ссылки на информацию об образце	https://www.encodeproject.org/experiments/ENCSR096USV/ https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1101698
Организм и ткань	*Homo sapiens*, мышца ноги
Стратегия секвенирования	Полиаденилированная РНК
Парно-/одноконцевые	Одноконцевые
Цепь-специфичность	Нет

Проверка качества чтений

Запустил FastQC, чтобы посмотреть на качество чтений:

```
fastqc reads/ENCFF038OLY.fastq.gz
```

FastQC сказала, что чтения провалили проверку per base sequence content и Sequence duplication levels и выдала предупреждения для per tile sequence quality и overrepresented sequences. На рис. 13–17 и в табл. 4 показана выдача FastQC.

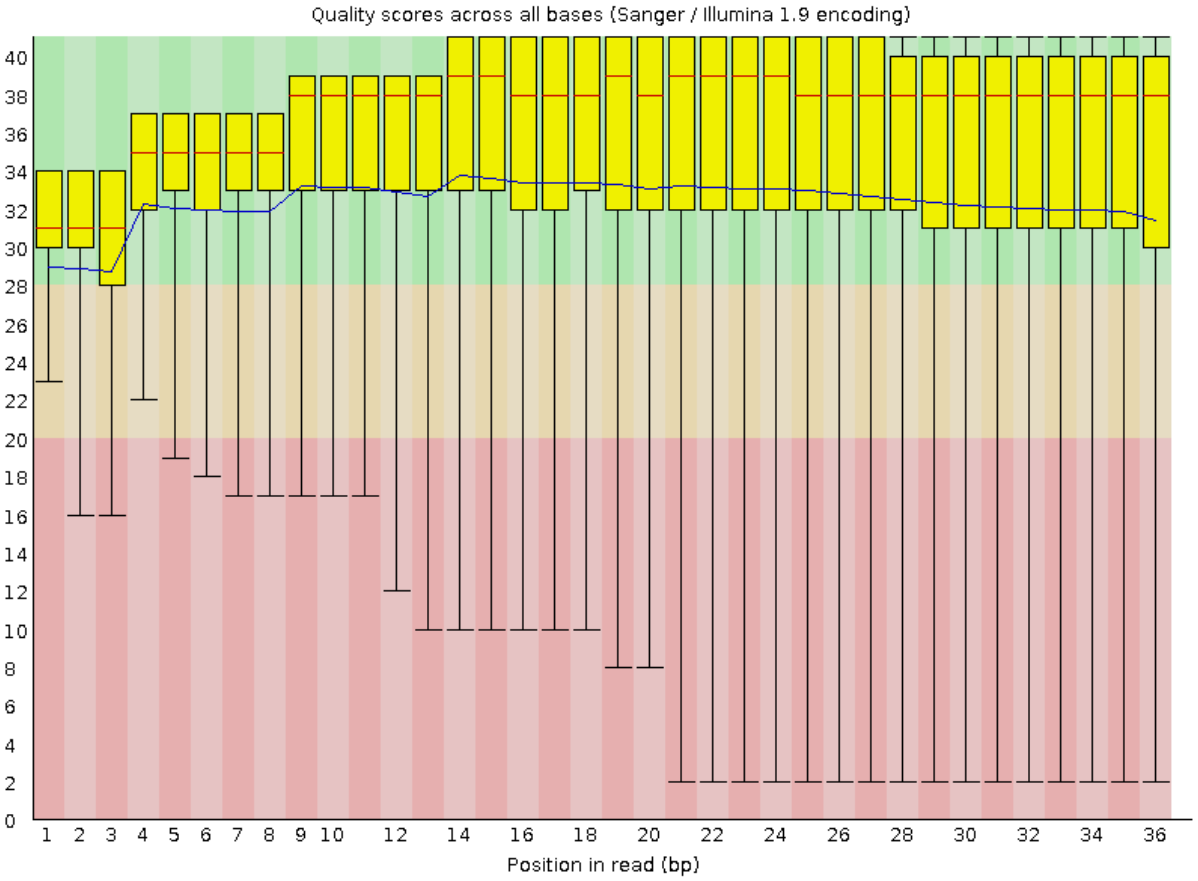
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Position in read (bp)

Рис. 13

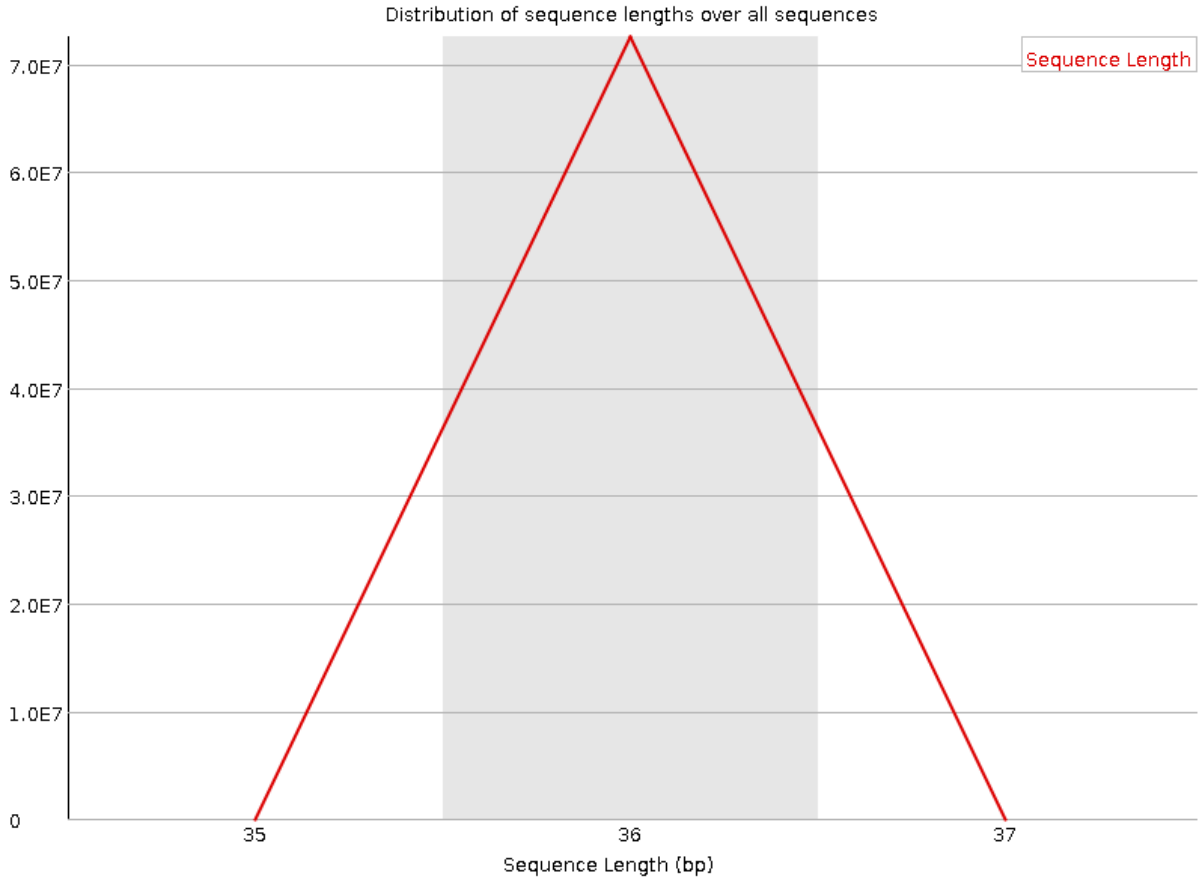

Рис. 14

Все чтения длины 36. Это хорошо.

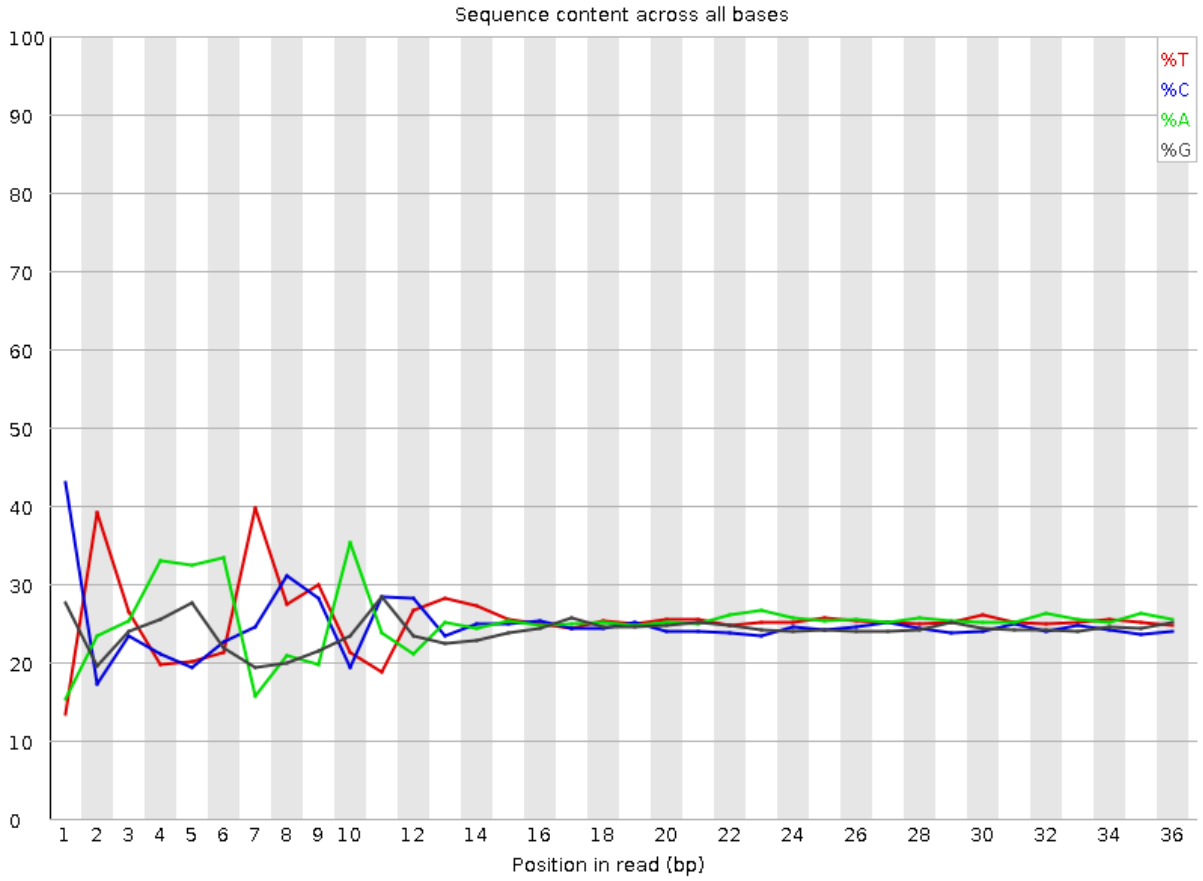
Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Position in read (bp)

Рис. 15

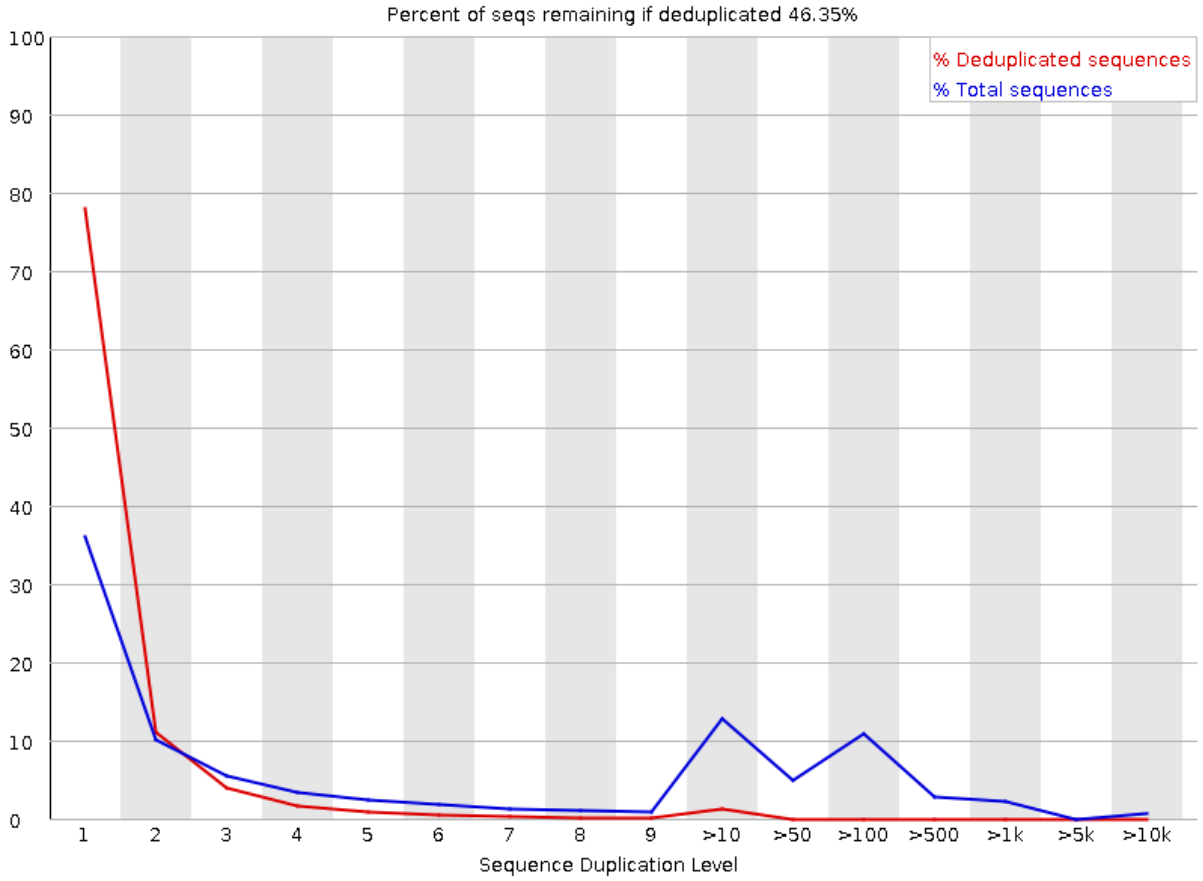
Percent of seqs remaining if deduplicated 46.35%
% Deduplicated sequences
% Total sequences
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k
Sequence Duplication Level

Рис. 16

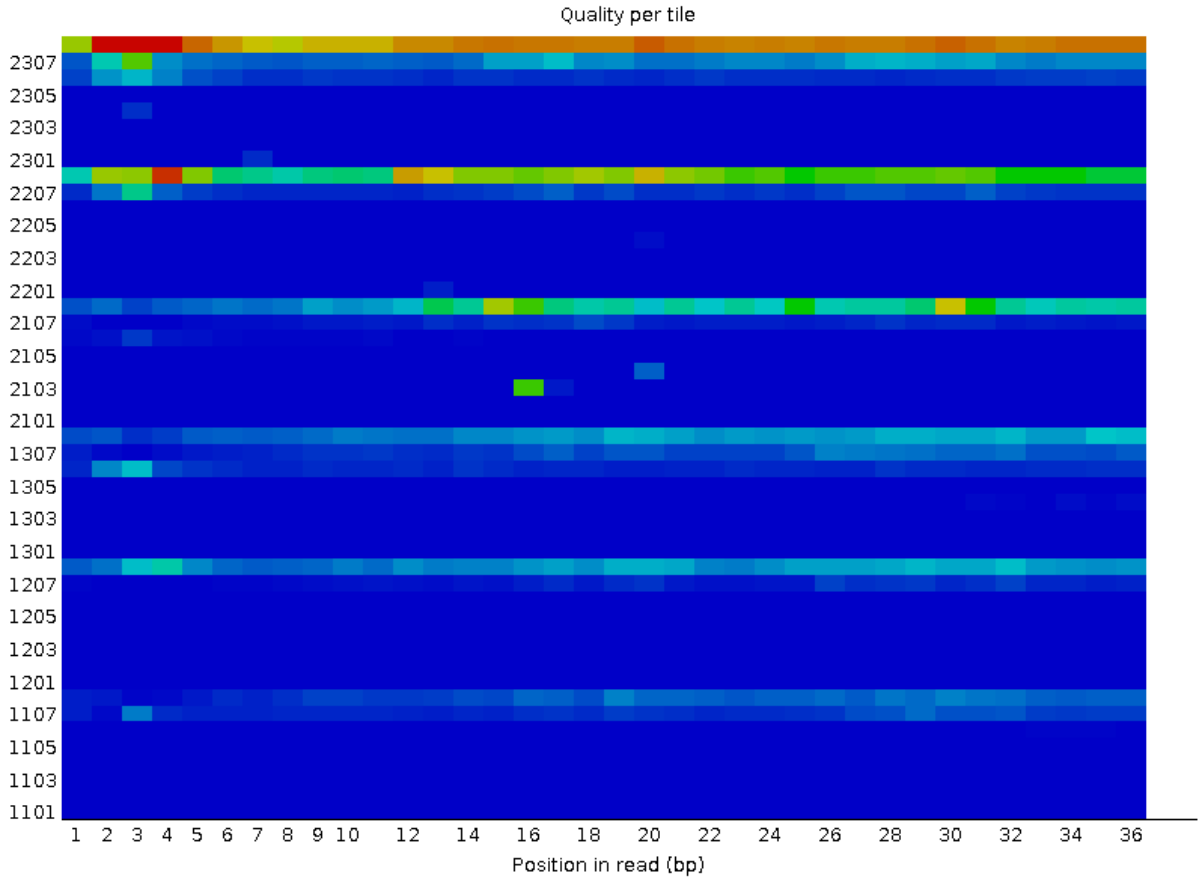

Рис. 18

Табл. 4. Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAG	459179	0.6331960720632148	TruSeq Adapter, Index 7 (97% over 36bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTA	145109	0.2001015917997579	TruSeq Adapter, Index 7 (97% over 35bp)

По-моему, в этих чтениях есть три проблемы.
Во-первых, проблема с частью тайлов (рис. 18). Выглядит так, будто с какой-то частью ячейки что-то было не в порядке. Эту проблему нельзя решить дистанционно, надо разбираться на месте.
Во-вторых, в чтениях есть адаптеры (табл. 4 и, видимо рис. 16 — duplication level). Это можно решить триммированием.
В-третьих, непонятная для меня проблема в начале чтений. В начале ниже качество (рис. 13) и нарушено соотношение нуклеотидов (рис. 15). Рис. 15 на первый взгляд выглядит так, будто в начале просто меньше данных, поэтому они шумнее, но этого, очевидно, не может быть. У всех последовательностей, очевидно, есть первые нуклеотиды (в данном случае, последние тоже, так как все чтения одинаковой длины).

Значит, единственное объяснение рис. 15 — в начале есть повторяющаяся последовательность у многих чтений. Но это не адаптер, так как a) адаптер не может прочитаться в начале, если я верно понимаю технологию секвенирования Illumina и b) выходит, что это последовательность CTNAAATYYA, а ее нет в адаптерах.

Я не понял до конца, что именно произошло с этими чтениями, но, я думаю, если бы было больше времени, можно было бы разобраться. В задании практикума триммировать чтения не требуется, хотя по-хорошему их, конечно, надо триммироват.

Картирование чтений на референс

Чтобы картировать и отфильтровать чтения, поменял сценарий, написанный для практикумов 12–13. Программы из сценария выглядят так:

```
hisat2 -x ${prefix} -p 4 -k 3 -U ${1} -S all_reads.sam -t
samtools sort -o  all_reads.bam all_reads.sam
samtools index all_reads.bam
samtools view -b -h all_reads.bam ${chr_name} > proper_chr.bam
```

hisat2 запускается для одноконцевых чтений (-U вместо -1 и -2), -k указывает максимальное количество первичных (самых качественных) выравниваний, которое ищет картировщик.

На хромосому картировалось 5 844 703 чтения.

Поиск экспрессирующихся генов

Файл с генной разметкой — это таблица из девяти колонок: имя последовательности, источник аннотации, тип элемента (gene, CDS и т.п.), начало, конец, вес (score), цепь (+ или -), рамка считывания и тэги с дополнительной информацией.

Чтобы составить профиль экспрессии, запустил htseq-count:

```
htseq-count -f bam -s no -m union -t gene bam/rna.bam
../DATA/genes/Homo_sapiens.GRCh38.110.chr.gtf > expression.txt
```

-f — формат входного файла, -s — были ли чтения цепь-специфичными, -m указывает, что делать, если чтение перекрывается сразу с несколькими генами (в данном случае — засчитать в оба), -t указывает тип элементов, перекрывание с которыми будет считать программа.

Получившийся файл — таблица из двух колонок, где первая колонка — идентификатор гена, вторая — количество попавших в него чтений.

В конце файла написано, что мимо генов попало 293 702 чтения. Соответственно, остальные 5 551 001 чтение попали в границы генов.