

Мини-практика:

знакомство с данными секвенирования

*Проверка качества ридов, тримминг, визуализация выравнивания, таблица каунтов,
базы данных секвенирования*



Где искать данные?

- ✓ Большинство данных секвенирования выкладываются в открытый доступ
- ✓ В базе GEO можно найти как «сырые» данные, так и обработанные

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.



Getting Started

- [Overview](#)
- [FAQ](#)
- [About GEO DataSets](#)
- [About GEO Profiles](#)
- [About GEO2R Analysis](#)
- [How to Construct a Query](#)
- [How to Download Data](#)

Tools

- [Search for Studies at GEO DataSets](#)
- [Search for Gene Expression at GEO Profiles](#)
- [Search GEO Documentation](#)
- [Analyze a Study with GEO2R](#)
- [Studies with Genome Data Viewer Tracks](#)
- [Programmatic Access](#)
- [FTP Site](#)

Browse Content

Repository Browser	
DataSets:	4348
Series: 	129876
Platforms:	20927
Samples:	3602181

База данных GEO

Gene Expression Omnibus

[NCBI GEO database](#)

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Browse Content

Repository Browser

DataSets:	4348
-----------	------

Series: 	209928
---	--------

Platforms:	25415
------------	-------

Samples:	6717565
----------	---------

Используется для размещения данных научных исследований в открытый доступ.

Результаты array и секвенирования.

Чаще всего выкладывают fastq файлы и таблицы каунтов.

Также прикрепляют описание дизайна эксперимента и образцов исследования.

Можно искать по словам или по GEO Accession number.

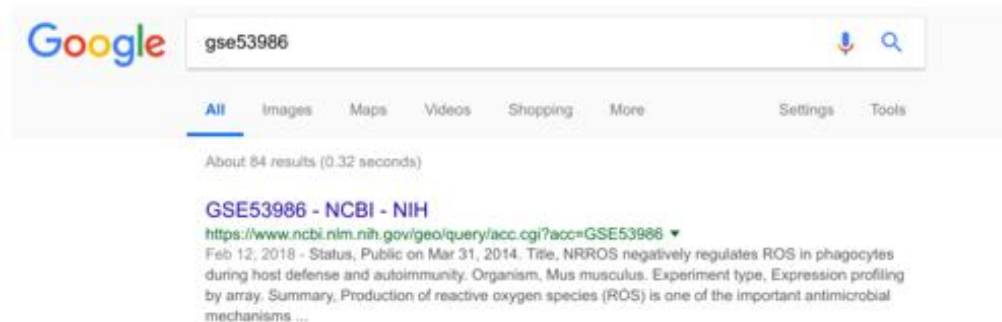
Идентификатор GSE53986



Microarray analysis. Statistical analyses of microarray data were performed using the R programming language (<http://r-project.org>). Microarray data were normalized using the RMA method²⁷. Data were prefiltered to remove probes that were not mapped to an annotated Entrez gene. We also filtered our data to retain only a single probe per gene, selecting the probe with the highest variance, if multiple probes were found for the gene²⁸. For differential expression analysis, the limma R package was used²⁹. We modelled the synergistic regulation of gene expression by the combined IFN- γ and LPS treatment as an interaction term in our linear model. This model will identify changes that are significantly different from the sum of the individual treatments. Multiple test correction was done using the method of Benjamini and Hochberg³⁰. Genes were considered significantly different if they changed more than 1.4-fold at a false discovery rate of 0.05. Genes were further filtered for immune-cell-specific expression using the gene sets defined by the Immune Response In Silico (IRIS) project³¹. As the IRIS-defined gene sets were derived from human immune cells, we mapped the human genes to mouse orthologues using the HomoloGene database³². Genes from all IRIS-defined categories were included in the analysis. Data were submitted to the NCBI (accession number GSE53986).

<http://www.nature.com/nature/journal/v509/n7499/full/nature13152.html>

Поиск



Series GSE53986		Query DataSets for GSE53986
Status	Public on Mar 31, 2014	
Title	NRROS negatively regulates ROS in phagocytes during host defense and autoimmunity	
Organism	Mus musculus	
Experiment type	Expression profiling by array	
Summary	<p>Production of reactive oxygen species (ROS) is one of the important antimicrobial mechanisms of phagocytic cells. Enhanced oxidative burst requires these cells to be primed with agents such as IFNγ and LPS with a synergistic effect of these agents on the level of the burst. However, excessive ROS generation will lead to tissue damage and has been implicated in a variety of inflammatory and autoimmune disease. Therefore, this process needs to be tightly regulated. In order to understand the genes regulating this process, we will treat bone marrow derived macrophages with above mentioned priming agents and study the gene expression.</p> <p>We used microarrays to determine the changes in gene expression that occur in bone marrow derived macrophages after treatment with IFNγ, LPS, or a combination of IFNγ and LPS</p>	
Overall design	Four condition experiment; Biological replicates: four replicates per condition	
Contributor(s)	Noubade R , Wong K , Ota N , Rutz S , Eidenschenk C , Ding J , Valdez PA , Peng I , Sebrell A , Caplazi P , DeVoss J , Soriano RH , Modrusan Z , Hackney JA , Sai T , Ouyang W	
Citation(s)	Noubade R, Wong K, Ota N, Rutz S et al. NRROS negatively regulates reactive oxygen species during host defence and autoimmunity. <i>Nature</i> 2014 May 8;509(7499):235-9. PMID: 24739962	

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53986>

Sequence Read Archive (SRA)

Чтобы скачать данные нужна специальная программа SRA Toolkit, так как данные хранятся в специальном архиве с разрешением .sra

Можно скачать по ссылке

<https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit>

SRA ▼

RNA seq[Strategy] |

Create alert Advanced

Summary ▼ 20 per page ▼

Search results

Items: 1 to 20 of 5107259

<< First < Prev Page

☐ [RNA seq of untreated HEK293 cells](#)

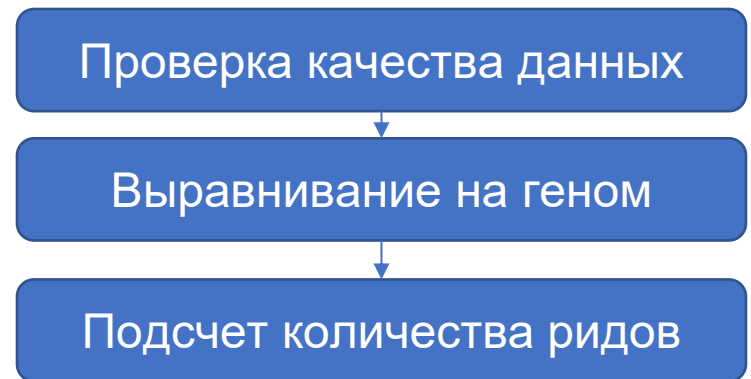
1. 1 ILLUMINA (Illumina NovaSeq X) run: 35.9M spots, 7.3G bases, 2Gb downloads
Accession: SRX26385063

☐ [RNA seq of tau expressed HEK293 cells](#)

2. 1 ILLUMINA (Illumina NovaSeq X) run: 35.7M spots, 7.2G bases, 2Gb downloads
Accession: SRX26385062

План первичного анализа

1. Данные с секвенатора – *.fastq* файлы
2. Проверка качества
3. Данные после выравнивания - *.bam* файлы
4. Визуализация выравненных ридов
5. Таблица каунтов



P.S. [Ссылка](#) на более подробный план

Для начала:

Скачиваем файлы (папка День2/toDownload):

- .fastq* (данные),
- .bam* (выравненные данные)
- .bam.bai* (индексированный файл)
- .txt* (таблица каунтов)
- *.html* (отчеты о проверке качества ридов)

*Необязательно

Формат файла *.fastq*

Файл FASTQ обычно использует четыре строки на последовательность:

- Строка 1 начинается с символа "@", за которым следует идентификатор последовательности и необязательное описание.
- Строка 2 - это нуклеотидная последовательность.
- Строка 3 - символ '+'
- Строка 4 кодирует значения качества для последовательности в строке 2

Identifier	●	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	●	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	●	+
Quality scores	●	hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier	●	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	●	GATTTGTATGAAAGTATACAACATAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	●	+
Quality scores	●	hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

Basic statistics



Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Total Bases	10 Mbp
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Всего рядов

Длина ряда

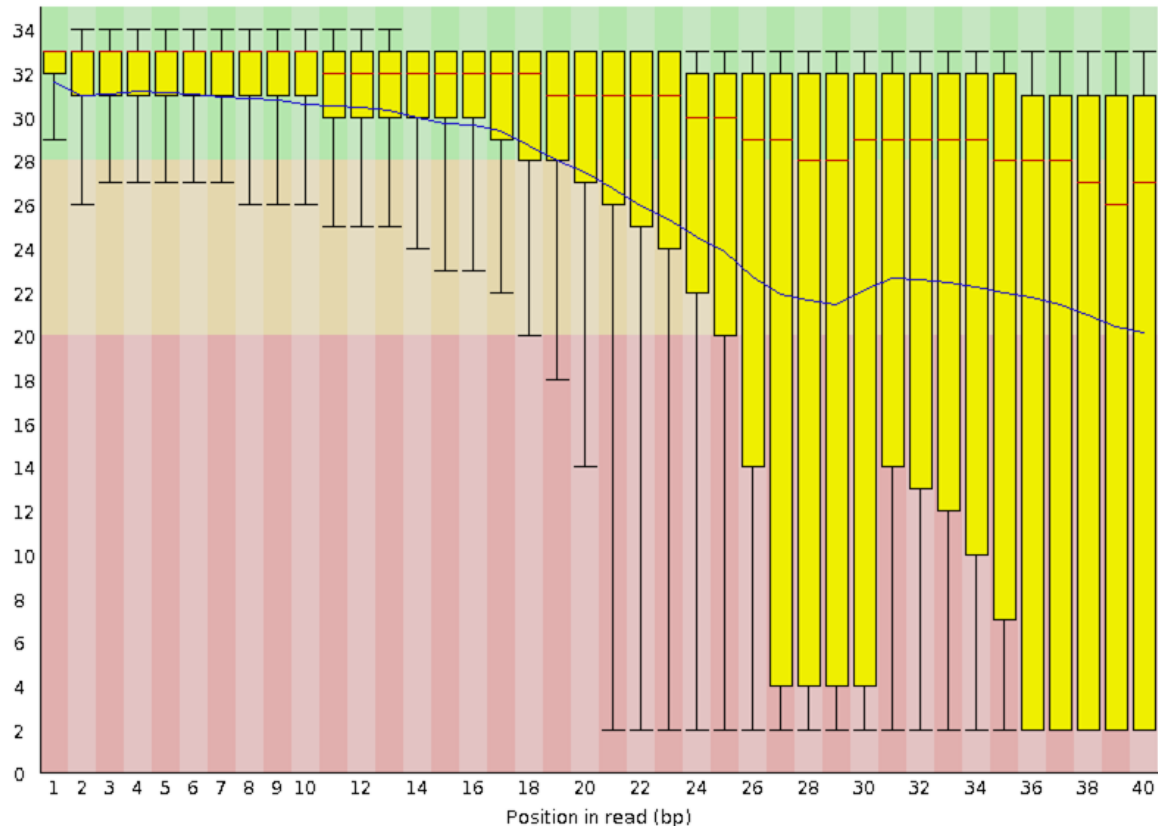
Per base sequence quality

Данный график показывает качество прочтений относительно каждого основания.

1. центральная красная линия — медиана значений.
2. желтый бокс представляет интерквартильный размах (вероятное отклонение) (25-75%).
3. верхние и нижние линии бокса показывают 10 и 90%.
4. Синяя линия представляет среднее качество.

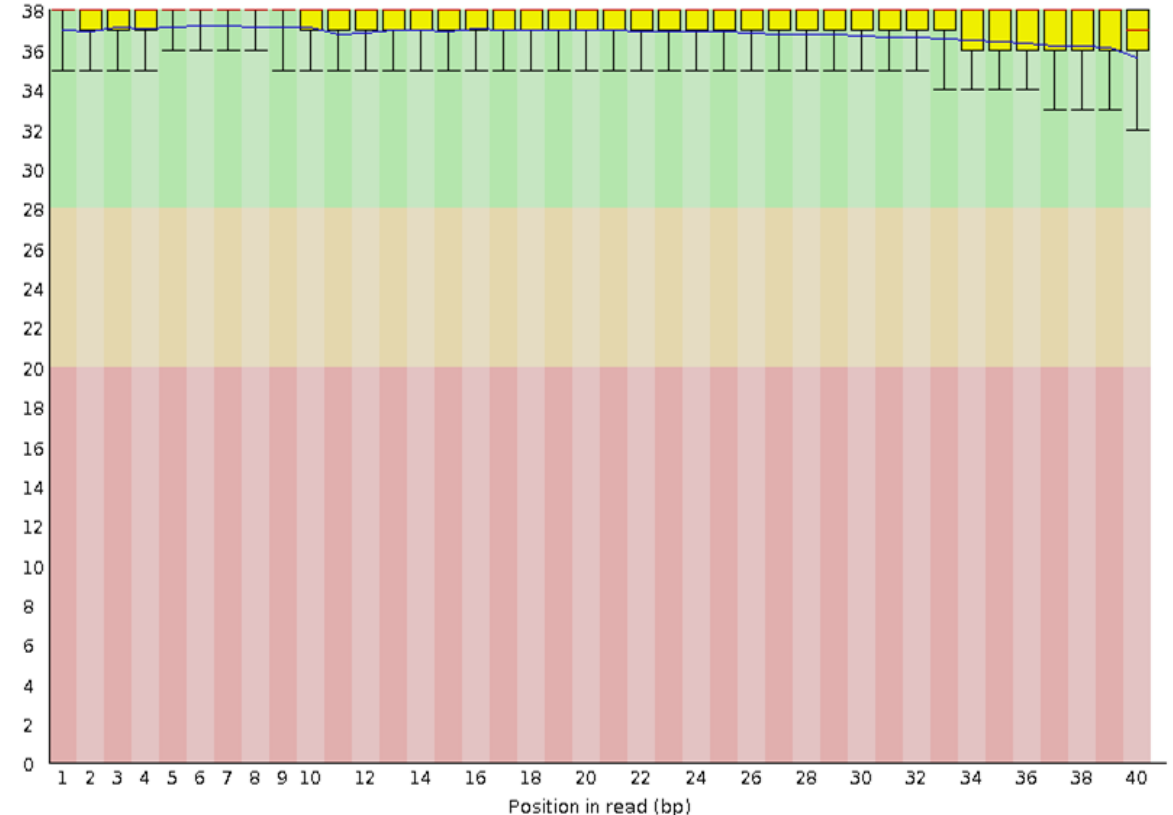
BAD

Quality scores across all bases (Illumina 1.5 encoding)



GOOD

Quality scores across all bases (Illumina 1.5 encoding)



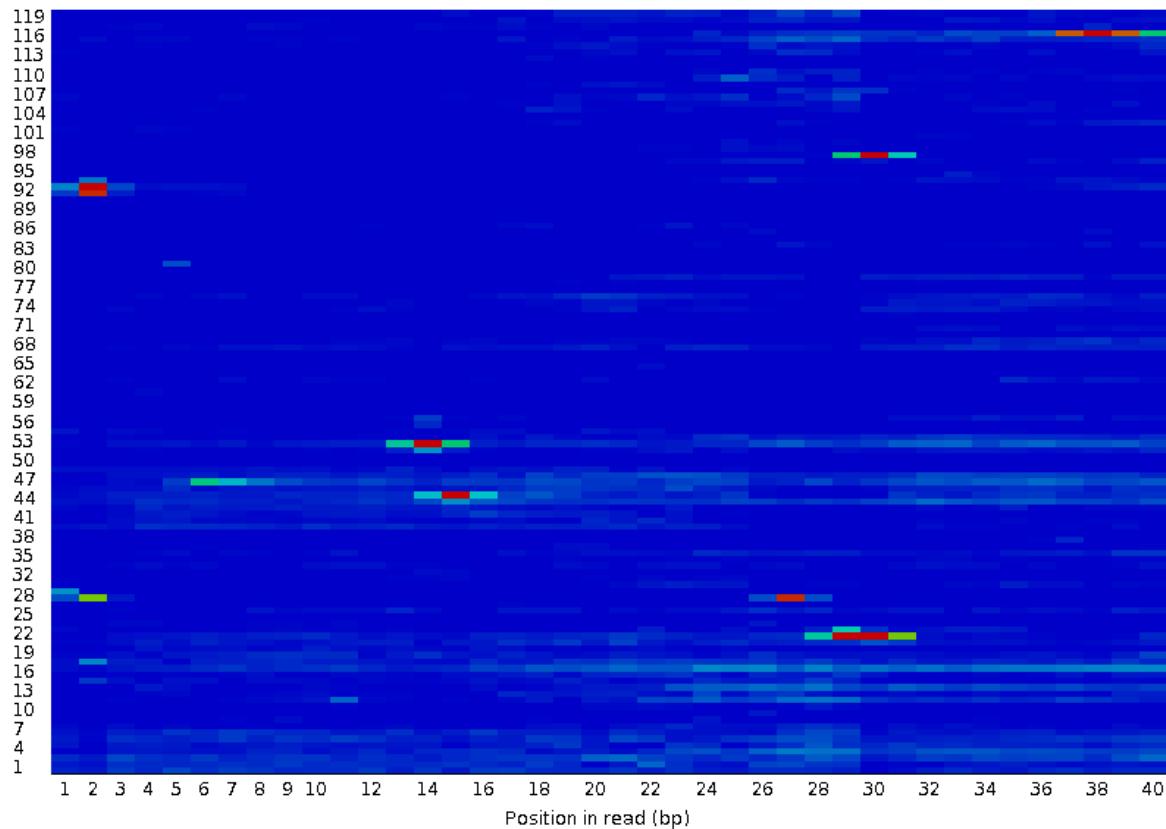
Ось X - это длина считывания, ось Y - показатель качества.

Per tile sequence quality

График показывает отклонение от среднего качества для каждого тайла. Цвета находятся в диапазоне от холодного до горячего, холодные цвета - это позиции, где качество было на уровне или выше среднего, а более теплые цвета указывают ухудшение качества.

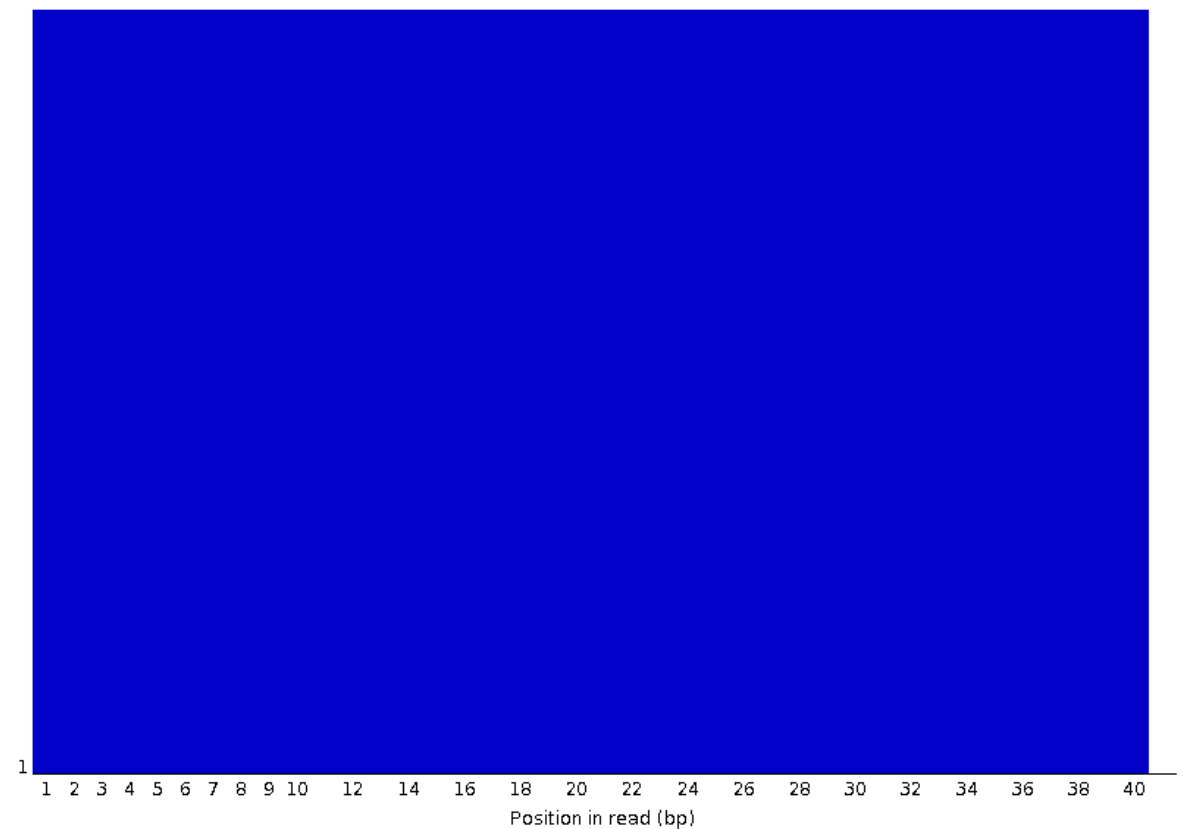
BAD

Quality per tile



GOOD

Quality per tile



Ось X - позиция на прочтении, ось Y - номер тайла.

Per tile sequence quality



Chip, slide, flow cell...



HiSeq 2500

Technology Overview - GAll

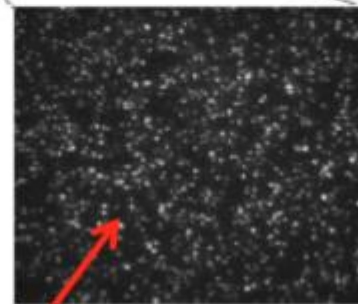
flow cell



A flow cell contains eight lanes



Tile



DNA fragment

Each lane contains two columns

Each column contains up to 50 tiles

Each tile is imaged four times per cycle – one image per base

Each image is 2.5-3.0 Mb, and ~115,000 images are produced per 36-cycle run

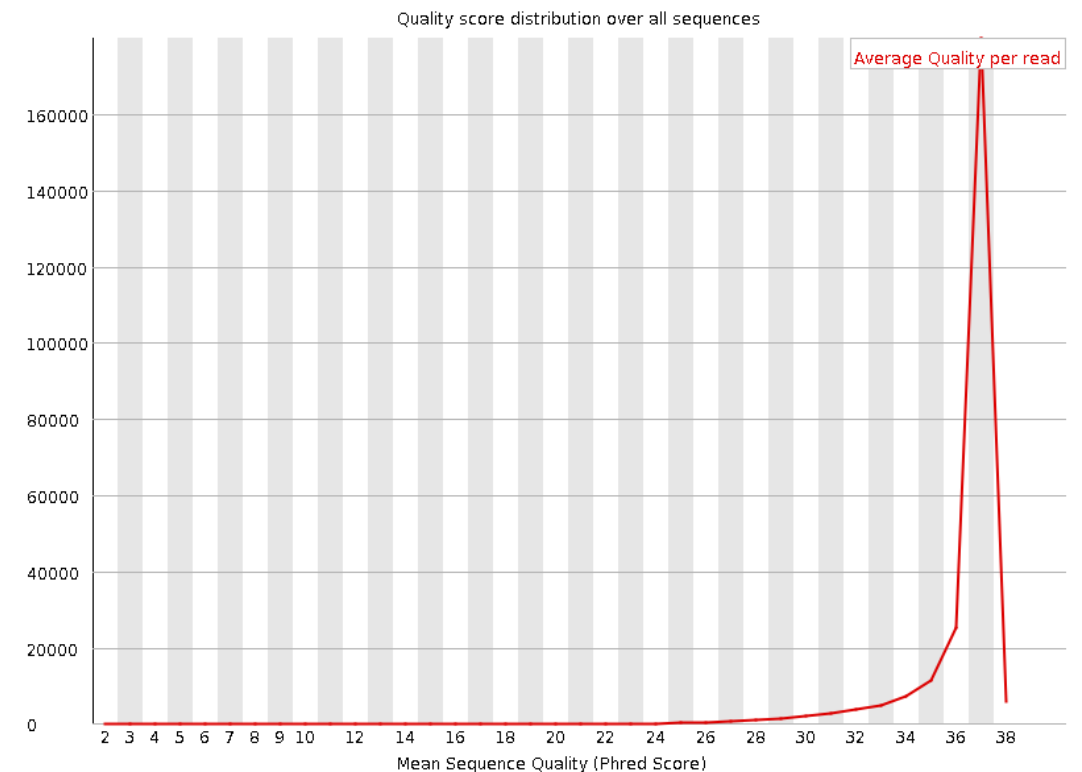
Per sequence quality scores

Когда результаты секвенирования в основном сосредоточены в высоких баллах, это доказывает, что качество секвенирования хорошее.

BAD



GOOD

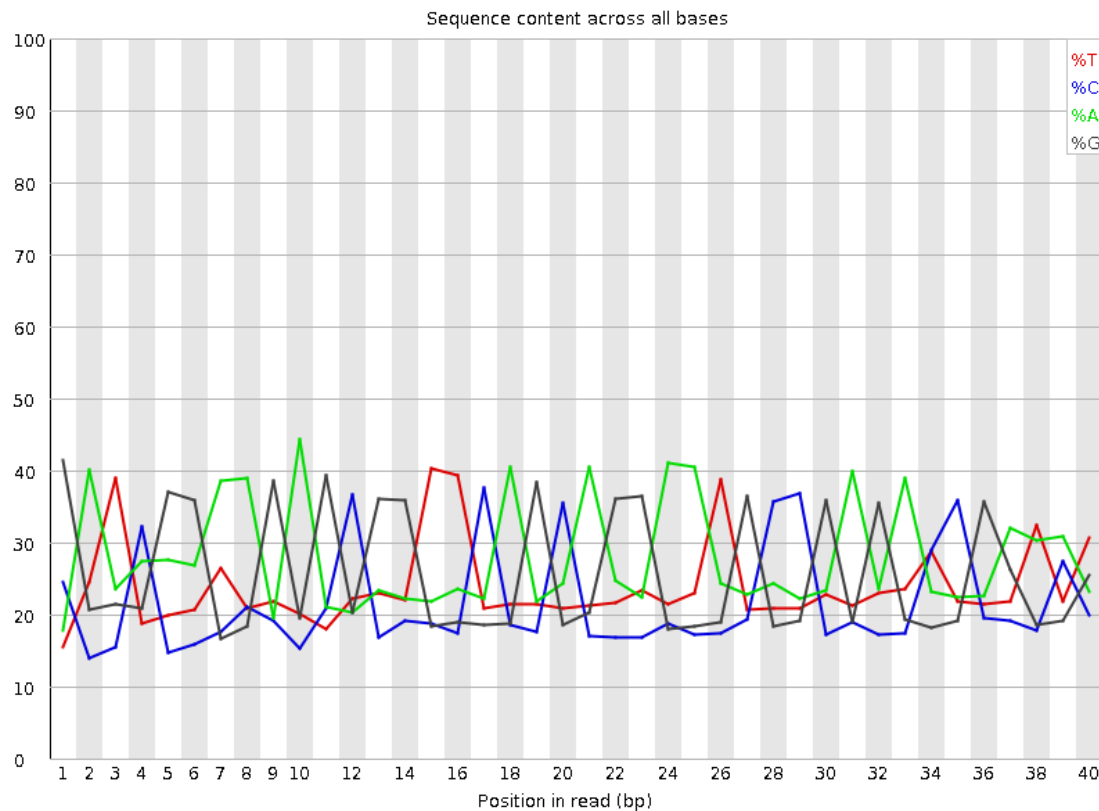


Ось X - значение Q, ось Y - количество прочтений.

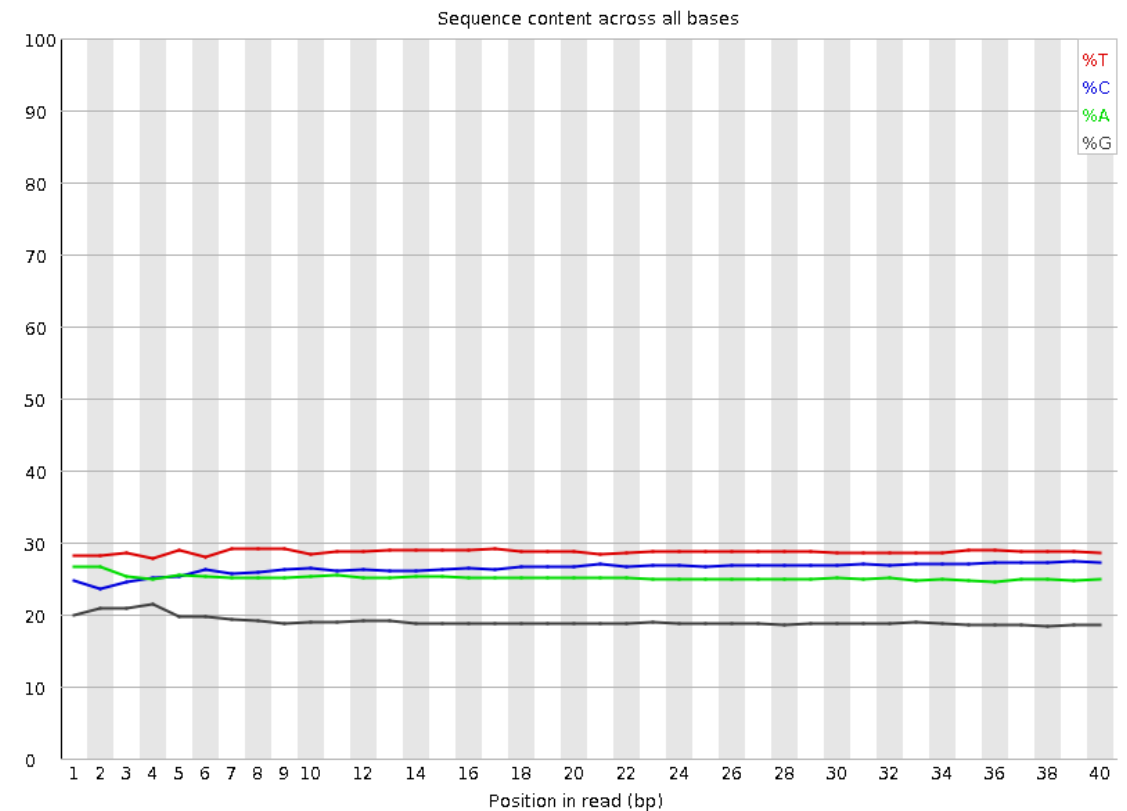
Per base sequence content

Отображает долю каждого нуклеотида во всех считываниях. Как правило, мы ожидаем увидеть примерно в 25% случаев в каждой позиции, но часто происходит сбой в начале рида из-за адаптера, который имеет не случайную последовательность.

BAD



GOOD



Ось X - позиция на прочтении, ось Y - процент оснований.

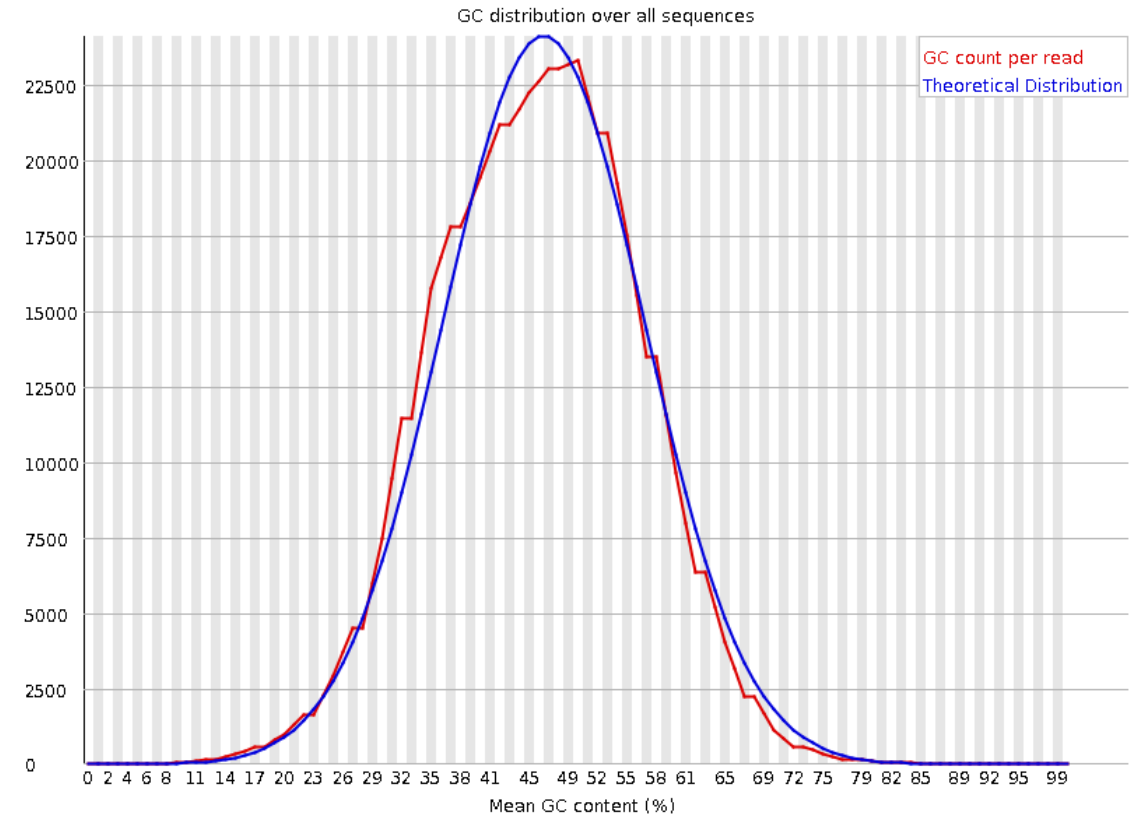
Per sequence GC content

Синий - теоретическое распределение, а красный - истинное значение. Когда появляется красный двойной пик, это означает, что последовательности ДНК загрязнены.

BAD



GOOD



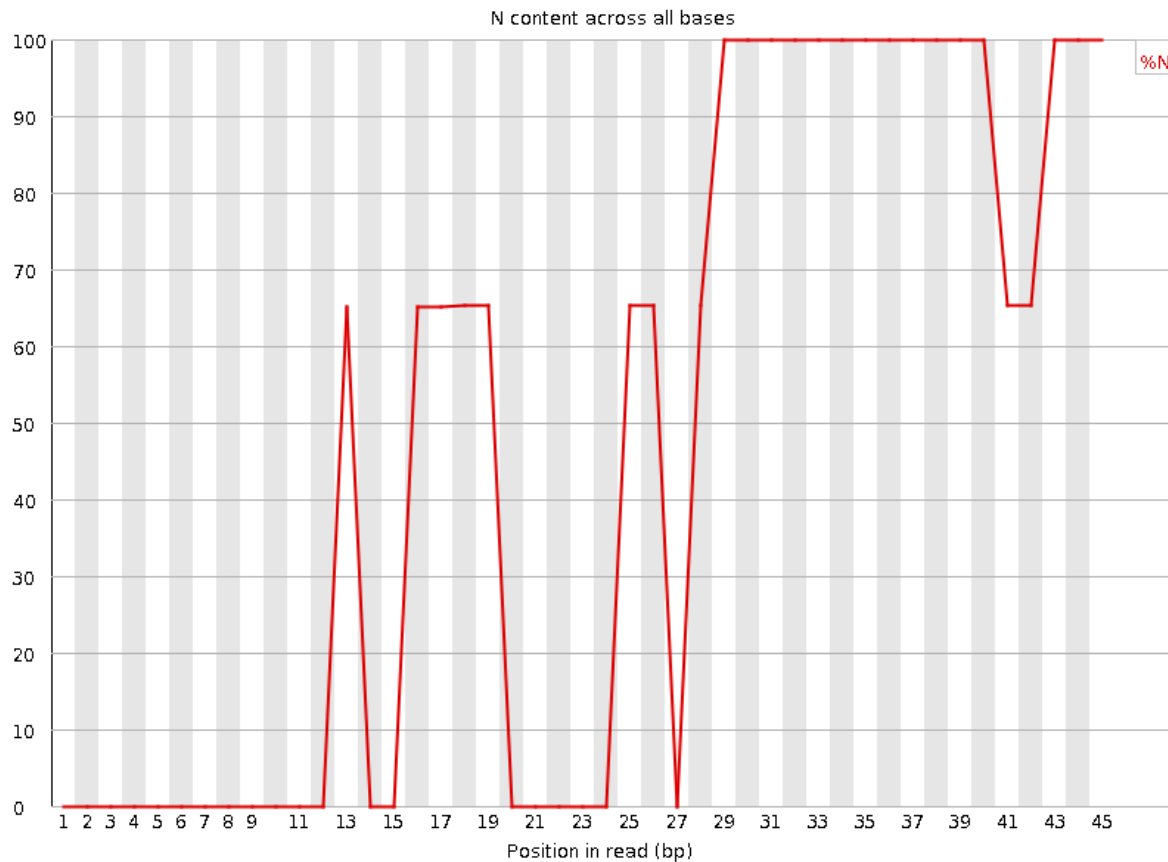
Ось X - содержание GC, ось Y - количество прочтений.

Per base N content

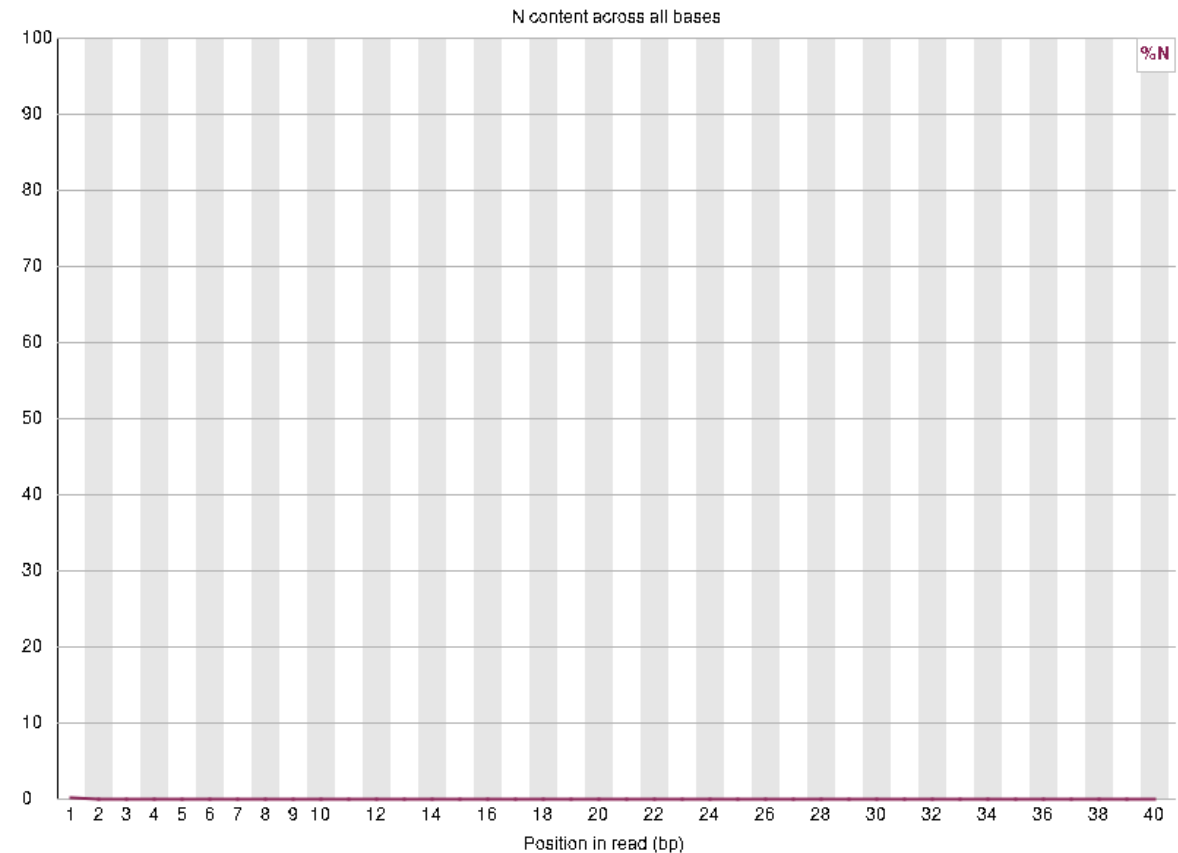
Если секвенсор не может с достаточной уверенностью распознать основание, то он обычно заменяет его на N.

График отображает процент оснований в каждой позиции, которые были определены как N.

BAD



GOOD

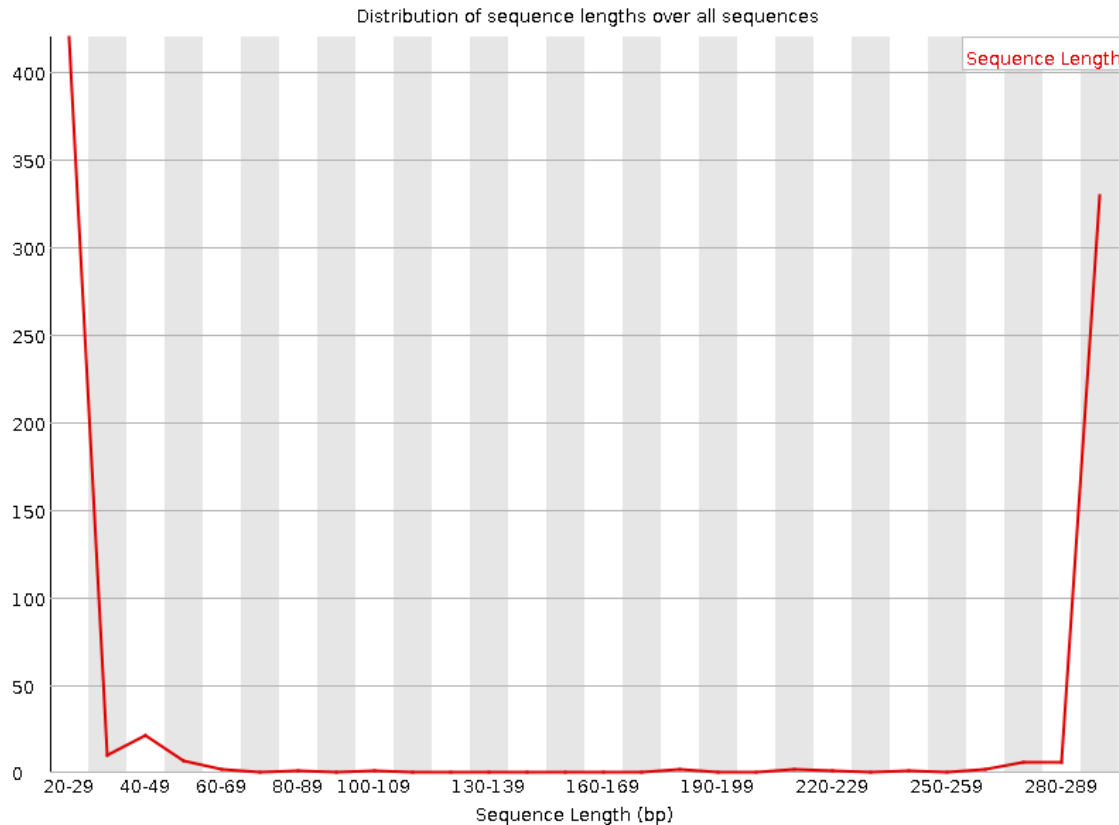


Ось X – позиция на прочтении, ось Y – процент непрочитанных нуклеотидов.

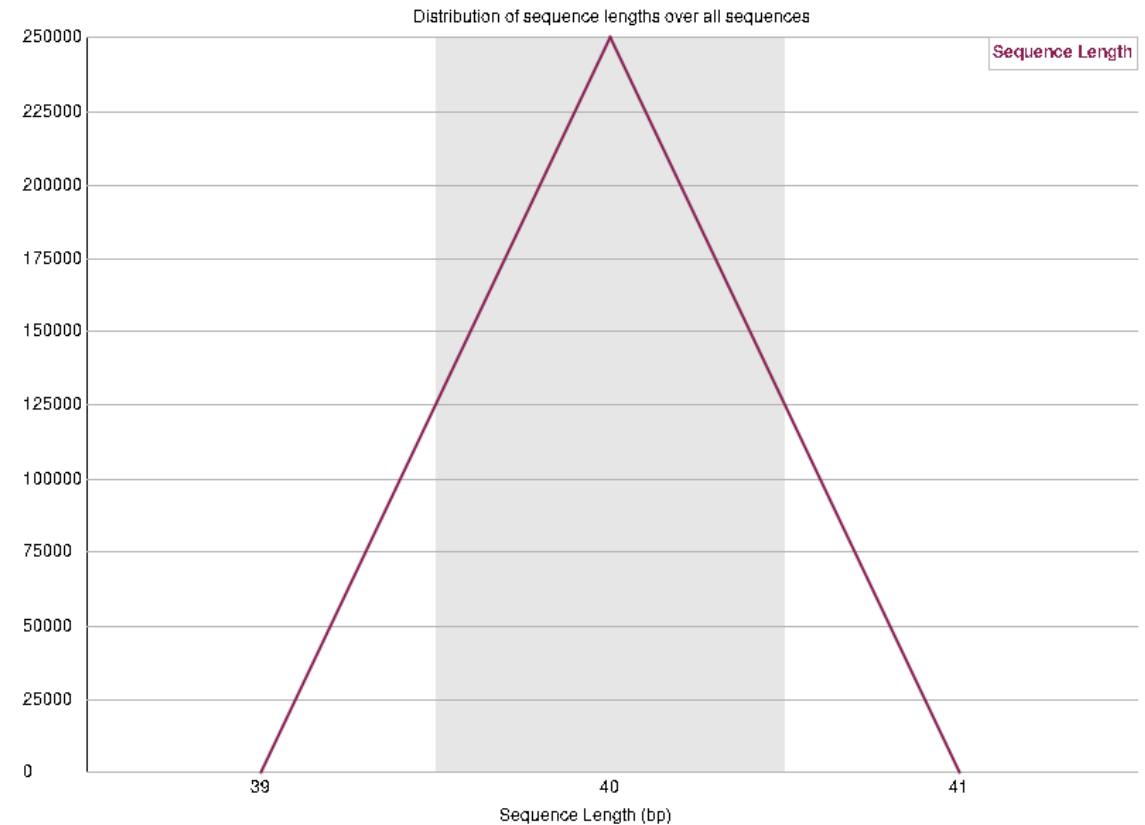
Sequence Length Distribution

Некоторые секвенаторы генерируют риды одинаковой длины, но другие могут давать риды с сильно различающейся длиной. Иногда некоторые конвейеры обрезают последовательности по качеству. Выдаст ошибку, если есть риды с нулевой длиной.

BAD



GOOD

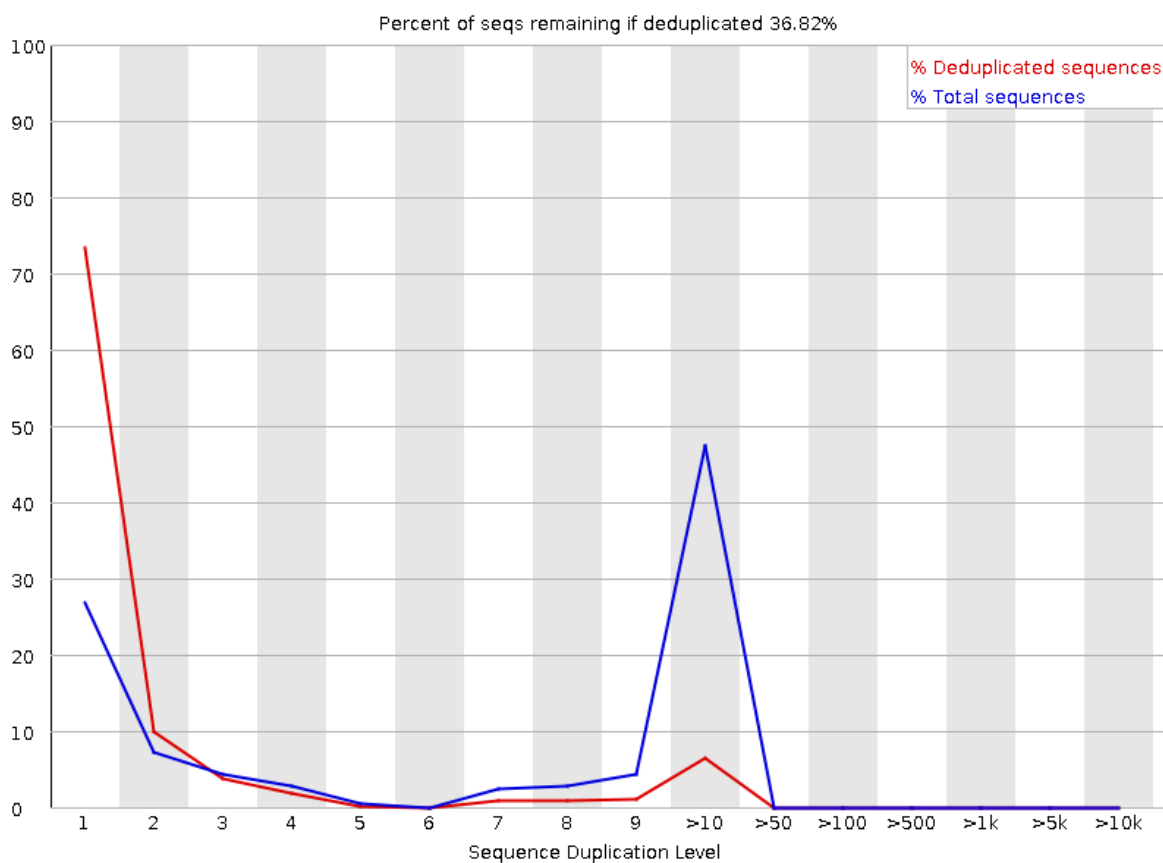


Ось X – длина прочтения, ось Y – количество прочтений.

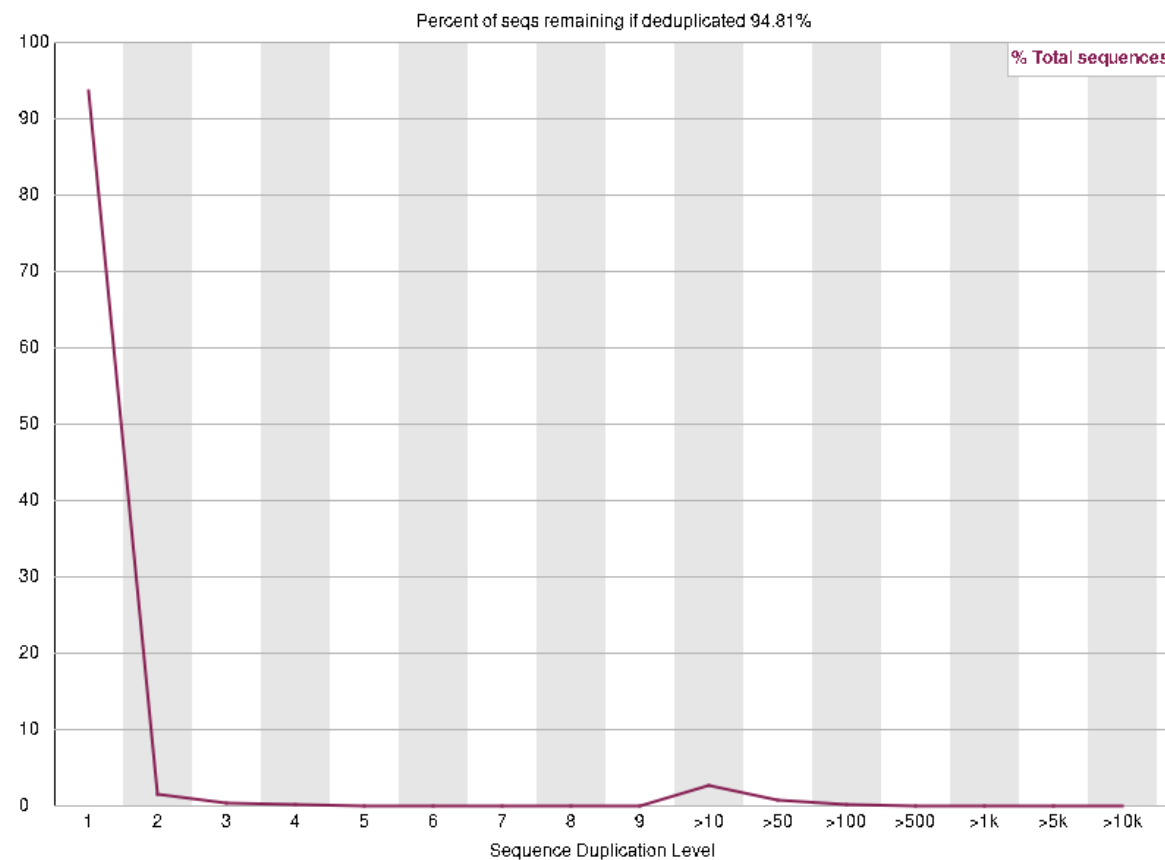
Sequence Duplication Levels

В разнообразной библиотеке большинство последовательностей будут встречаться только один раз. Но в данных РНКсека наблюдается высокий процент дублицированных последовательностей из-за особенности пробоподготовки и количества мРНК в клетке.

BAD, но не для RNAseq



GOOD



Ось X – сколько раз повторяется последовательность, ось Y – процент.

Overrepresented sequences


Если последовательность слишком широко представлена в данных, то означает либо то, что она имеет высокую биологическую значимость, либо то, что библиотека загрязнена или не так разнообразна. Перечисляется все последовательности, которые составляют более 0,1% от общего числа.

BAD

! Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit

GOOD

 **Overrepresented sequences**
No overrepresented sequences

Для каждой перепредставленной последовательности программа будет искать совпадения в базе данных распространенных загрязнителей и сообщать о лучшем найденном совпадении.

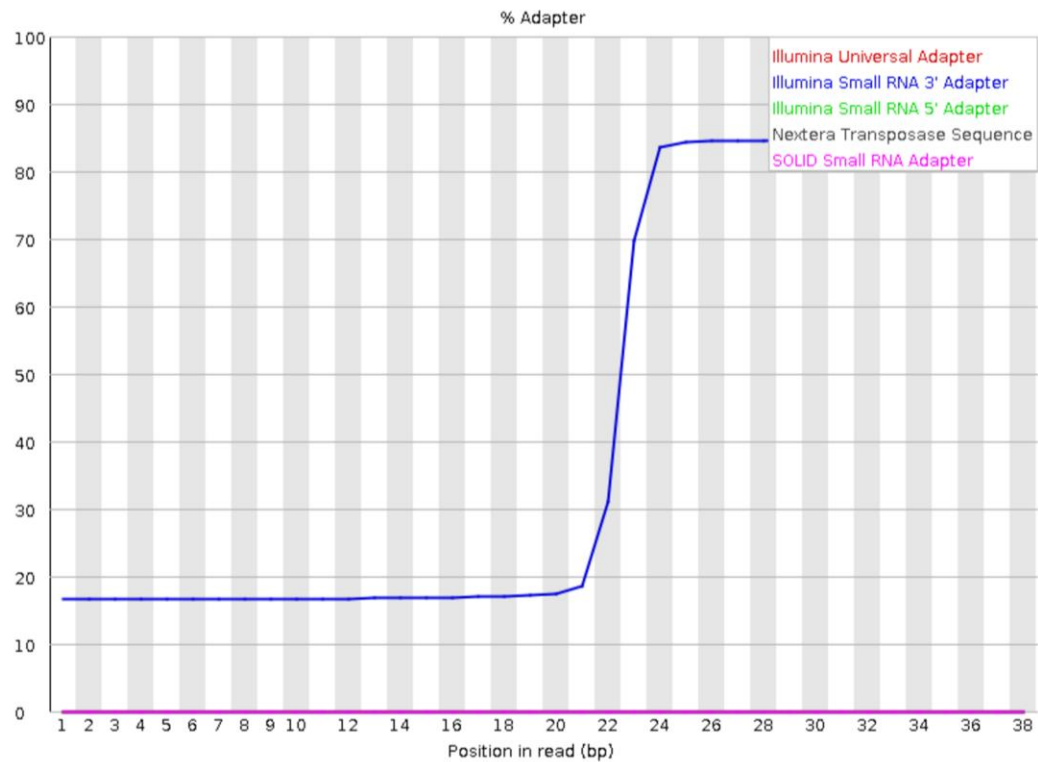
Могут «вылезти» адаптеры.

Одна и та же последовательность может естественным образом присутствовать в значительной части библиотек при smallRNAseq.

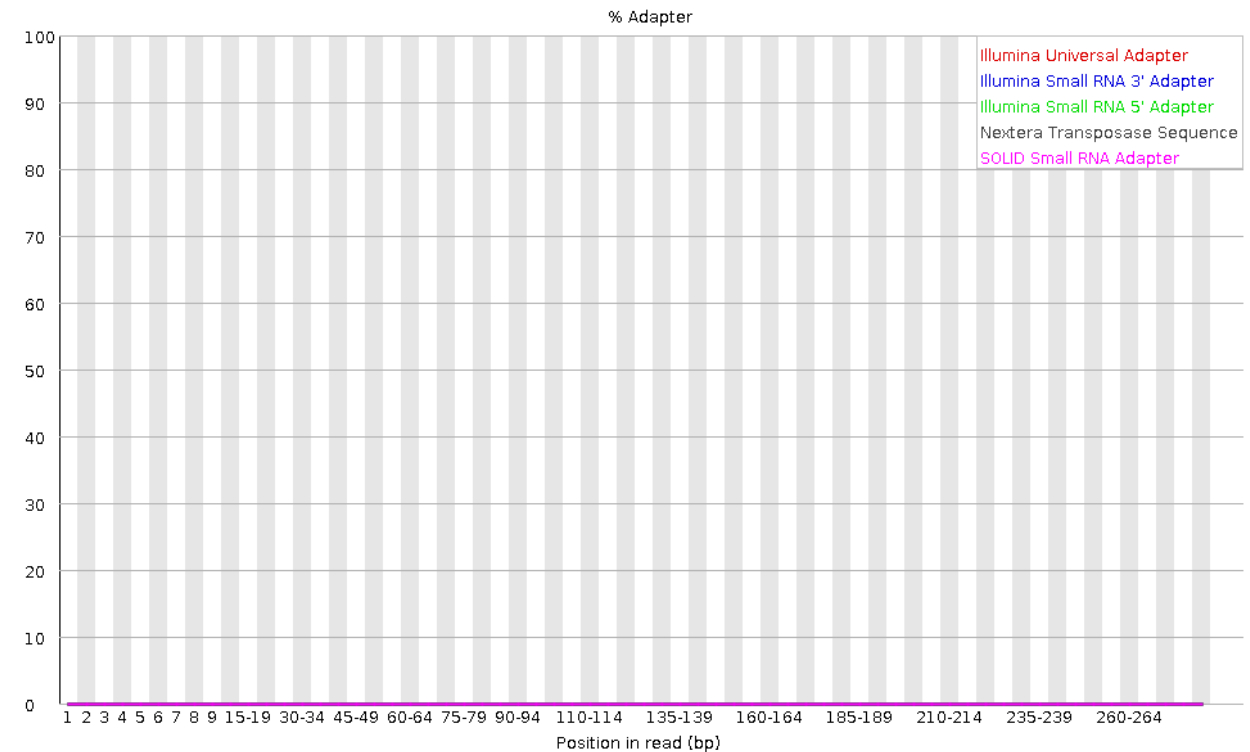
Adapter Content

Кумулятивный график доли прочтений, где последовательность адаптера найдена в указанной нуклеотидной позиции. Может быть увеличение процента на конце ряда из-за возможного маленького размера вставки.

BAD



GOOD

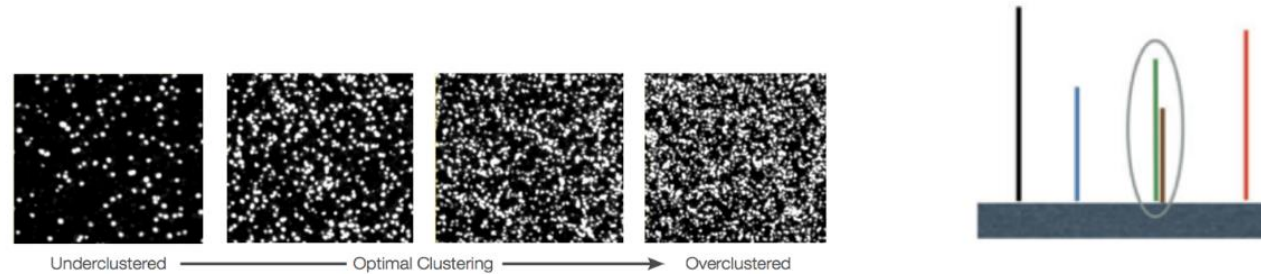


Ось X – позиция на прочтении, ось Y – процент.

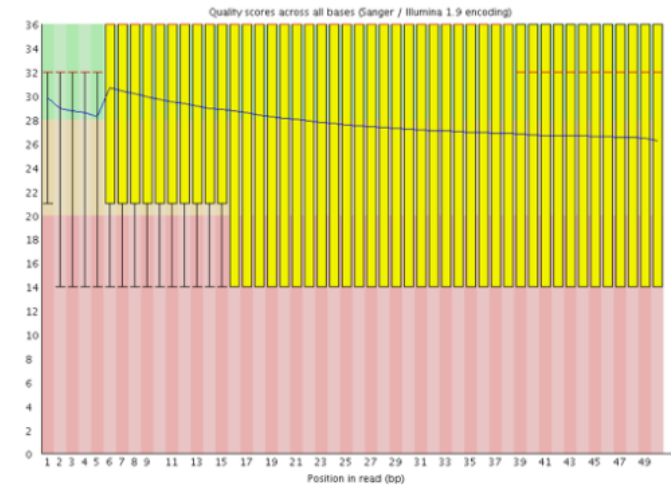
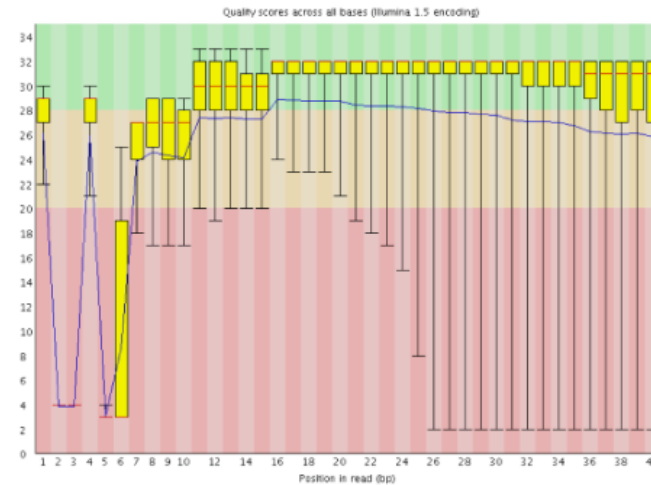
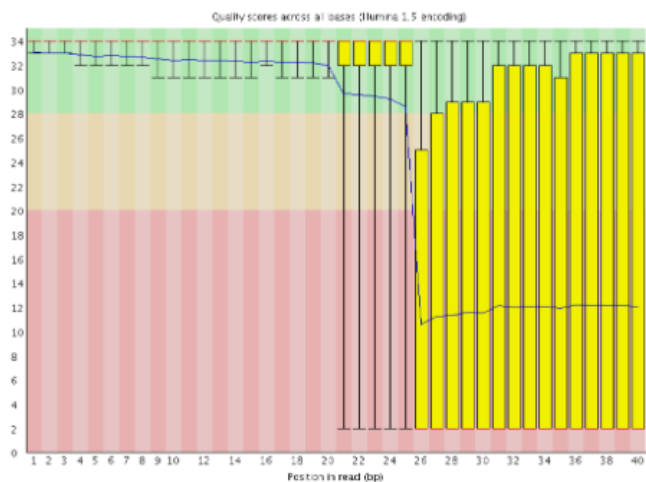
Контроль качества ридов FastQC - Профили ошибок секвенирования

Тревожные сигналы:

- 1) **Чрезмерная кластеризация:** Секвенаторы могут чрезмерно кластеризовать ячейки, что приводит к небольшим расстояниям между кластерами и перекрытию сигналов. Эти два кластера можно интерпретировать как единый кластер, в котором обнаруживаются смешанные флуоресцентные сигналы, что снижает чистоту сигнала и приводит к снижению показателей качества по всему показателю.



- 2) **Неисправность прибора:** во время работы оборудования для секвенирования иногда могут возникать проблемы с самим прибором. Любое внезапное снижение качества или большой процент некачественных считываний в процессе считывания может указывать на проблему на объекте (разрыв коллектора, потеря циклов, сбой считывания).



Контроль качества ридов FastQC - Профили ошибок секвенирования

При секвенировании Illumina качество прочтения нуклеотидных оснований связано с интенсивностью сигнала и чистотой флуоресцентного сигнала. Флуоресценция низкой интенсивности или наличие множества различных флуоресцентных сигналов может привести к снижению оценки качества, присвоенной нуклеотиду. Из-за природы секвенирования путем синтеза можно ожидать некоторого снижения качества, но другие проблемы с качеством могут указывать на проблему в установке секвенирования.

Ожидаемые ошибки

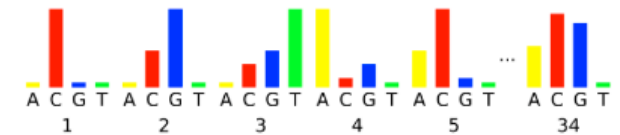
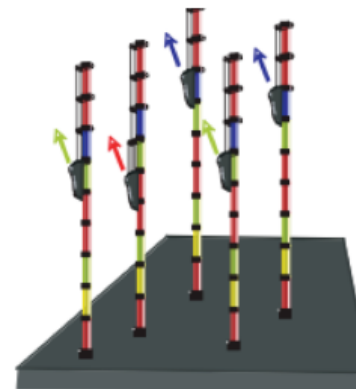
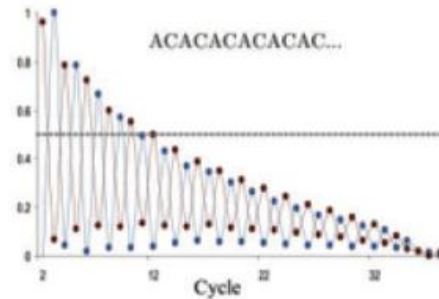
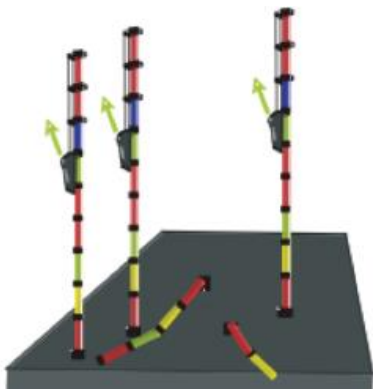
Затухание сигнала: По мере продолжения секвенирования интенсивность флуоресцентного сигнала уменьшается с каждым циклом, что приводит к снижению показателей качества в конце считывания на 3'. Это связано с:

- Разрушение флуорофоров
- Доля нитей в пучке не удлинилась

Таким образом, доля излучаемого сигнала продолжает уменьшаться с каждым циклом.

Фазирование: По мере увеличения числа циклов сигнал начинает размываться, поскольку кластер теряет синхронность, что также приводит к снижению показателей качества в конце считывания на 3'. По мере прохождения циклов в некоторых цепочках происходит случайный сбой включения нуклеотидов из-за:

- Неполное удаление 3'-терминаторов и флуорофоров
- Включение нуклеотидов без эффективных 3'-терминаторов



Контроль качества ридов FastQC

Per tile sequence quality: Стабильно низкие баллы часто обнаруживаются по краям, но низкое качество прочтения также может возникать в середине, если в какой-то момент во время запуска образовался воздушный пузырь.

Per sequence quality scores: Показатели качества для каждой последовательности - график качества для всех ридов во всех позициях (показывает, какие показатели качества являются наиболее распространенными).

Per base sequence content: отображает долю каждого нуклеотида во всех ридов. Как правило, мы ожидаем увидеть примерно в 25% случаев в каждой позиции, но часто происходит сбой в начале рида из-за адаптера, который имеет не случайную последовательность.

Per sequence GC content: график плотности среднего содержания GC в каждом из считываний.

Per base N content: процент случаев, когда 'N' встречается в позиции во всех ридов. Если в определенной позиции наблюдается увеличение, это может указывать на то, что что-то пошло не так во время секвенирования.

Sequence Length Distribution: распределение длин последовательностей всех ридов в файле. Если данные необработанные, часто наблюдается резкий пик, однако, если риды были обрезаны, может наблюдаться распределение меньшей длины.

Sequence Duplication Levels: распределение дублированных последовательностей. При секвенировании мы ожидаем, что большинство ридов будут встречаться только один раз. Если некоторые последовательности встречаются более одного раза, это может указывать на смещение обогащения (например, в результате ПЦР). Если образцы имеют высокое покрытие (или RNA-seq), это может быть неверно.

Overrepresented sequences: список последовательностей, которые встречаются чаще, чем можно было бы ожидать случайно.

Adapter Content: график, показывающий, где последовательности адаптеров встречаются при считывании.

Основные команды Linux

ls список файлов

pwd выводится полное имя текущего каталога

cd <каталог> изменить местонахождение

mkdir <имена создаваемых каталогов> создать новые каталоги

rm <имена удаляемых файлов>

rm -fr <имена удаляемых файлов или директорий>

mv старое-имя новое-имя перемещение

cp старое-имя новое-имя копирование

vim, nano консольные текстовые редакторы

man – справка по командам и программам (например, `man ls`)

less – когда в небольшом окне терминала надо просмотреть очень длинный текст

cat – читает файл и выводит на экран (stdout)

head – первые N строк в файле (например `head -10 file`)

tail – последние N строк в файле (например `tail -n 10 file`)

Контроль качества ридов

1) С помощью программы **FastQC** → Скачиваем программу ([FastQC](#))

```
sudo apt install fastqc
```

2) Открываем файл N3_S1_L001_R2_001.fastq.gz, проверяем качество ридов

Linux:

```
fastqc N3_S1_L001_R2_001.fastq.gz
```

3) После оценки качества происходит фильтрация/очистка данных с помощью программ **Trimmomatic** или **Cutadapt**

→ Устанавливаем программу [Cutadapt](#)

```
sudo apt install cutadapt
```

4) Удаляем адаптеры

```
cutadapt -a GATCGTCGGACTGTAGAACTCTGAAC -o N3_trimmed.fastq.gz N3_S1_L001_R2_001.fastq.gz
```

5) Проверяем качество полученного файла в **FastQC**

```
fastqc N3_trimmed.fastq.gz
```

Выравнивание ридов на референсный геном

- ✓ Делается +/- 2 в строчки командной строки → *Пропустим эту стадию на практике (требуем времени)*
- ✓ Программы **STAR**, **HiSat**, **Bowtie** (HiSat эффективный и быстрый, STAR - задействует много памяти)
- ✓ **Samtools** → для манипулирования выравниваниями в формате SAM, включая сортировку, объединение, индексацию и генерацию выравниваний в формате для каждой позиции.
 - .sam = текстовые файлы с разделителями табуляции, содержащие информацию для каждого отдельного выровненного рида и ее соответствие геному
 - .bam = сжатая SAM файла для уменьшения размера и обеспечения возможности индексации, что обеспечивает эффективный произвольный доступ к данным, содержащимся в файле.

Пример команды для запуска STAR

Шаг 1. Индексация генома

<code>./STAR</code>	
<code>--runThreadN 40</code>	<code>#число потоков</code>
<code>--runMode genomeGenerate</code>	<code>#создание индекса генома</code>
<code>--genomeDir ./STAR_mm10_genome_index</code>	<code>#куда сохранить индекс</code>
<code>--genomeFastaFiles ./GRCm38.primary_assembly.genome.fa</code>	<code>#путь до fasta файла генома</code>
<code>--sjdbGTFfile ./gencode.vM10.primary_assembly.annotation.gtf.gz</code>	<code>#путь до аннотации к геному</code>
<code>--sjdbOverhang 50</code>	<code>#длина ридов минус 1</code>

Индексирование генома можно объяснить аналогично индексированию книги.

Если вы хотите узнать, на какой странице появляется определенное слово или начинается глава, гораздо эффективнее/быстрее искать его в предварительно созданном индексе, чем просматривать каждую страницу книги, пока не найдете его.

То же самое касается выравниваний. Индексы позволяют выравнивателю сузить потенциальное происхождение последовательности в геноме, экономя как время, так и память.

Пример команды для запуска STAR

Шаг 2. Команда для выравнивания

./STAR

--genomeDir ./STAR_genome_index/

--readFilesCommand zcat

--runThreadN 20

--readFilesIn Read_R1.fastq.gz Read_R2.fastq.gz

--outFileNamePrefix Read_aligned

--outSAMtype BAM SortedByCoordinate

--outSAMunmapped Within

--outSAMattributes Standard

#папка с индексом генома

#риды находятся в архиве gz

#число потоков

#передний и обратный риды

#префикс в названии файла

#формат файла на выходе

#оставили в итоговом файле

только выравненные

последовательности

#использовать стандартные флаги

для оценки качества выравнивания

Недостаток STAR:

Для запуска нужно много оперативной памяти.

Скорее всего ноутбук не справится. Особенно, если это человеческий геном.

Геномный браузер

- Посмотрим на файл выровненных ридов
- Загружаем IGV – геномный браузер
<https://software.broadinstitute.org/software/igv/download>
- Открываем IGV
- В верхнем левом углу выбираем геном “Mouse (mm10)”
- File → Load from file → файл
wt_control_2_S2_L001_R1_001.bam
- Смотрим гены: *Actb*, *Lmna*, *Myh7*

Таблица каунтов

✓ Подсчет каунтов был сделан программой **featureCounts**

`./featureCounts`

`-p`

`-T 12`

`-s 2`

`-a ./genome.gtf`

`-o ./Counts.txt`

`./*.out.bam`

#путь до программы

#указывает на парные риды

#число потоков

#передний рид в паре из прямой цепи, а второй рид на обратной цепи

#путь до файла с аннотацией генома

#название таблицы каунтов

#на вход подаем файлы с выравниванием

✓ Откроем **объединенную таблицу для всех образцов**
featureCounts_LMNA.txt

Таблица каунтов

Гены



Geneid	P1_before	P2_before	P1_after	P2_after	P3_before	P3_after	P4
ENSG00000224031	55	11	50	0	5	10	
ENSG00000169962	2	0	0	0	0	0	
ENSG00000107404	212	60	190	22	22	47	
ENSG00000284372	0	0	0	0	0	0	
ENSG00000162576	60	95	134	72	33	269	
ENSG00000175756	336	143	333	73	35	105	
ENSG00000221978	23	13	27	7	5	55	
ENSG00000224870	1	0	0	0	0	0	
ENSG00000242485	34	16	40	4	5	5	
ENSG00000235098	3	0	4	0	1	8	
ENSG00000205116	0	0	0	0	0	0	
ENSG00000179403	16	5	10	4	2	15	
ENSG00000215915	2	0	1	1	0	0	

Образцы



Количество ридов для данного
образца данного гена

! Таблица экспрессии в целых числах - особенность ненормализованных данных РНК-сека (raw counts RNA-seq)

Если таблица экспрессии в действительных числах, то это либо данные микрочипов, либо таблица была нормализована (что нежелательно для реанализа)

