Контроль качества. Fastq-формат

Прикладная биоинформатика: NGS-технологии и Omics-анализ

Риды (прочтения, reads)

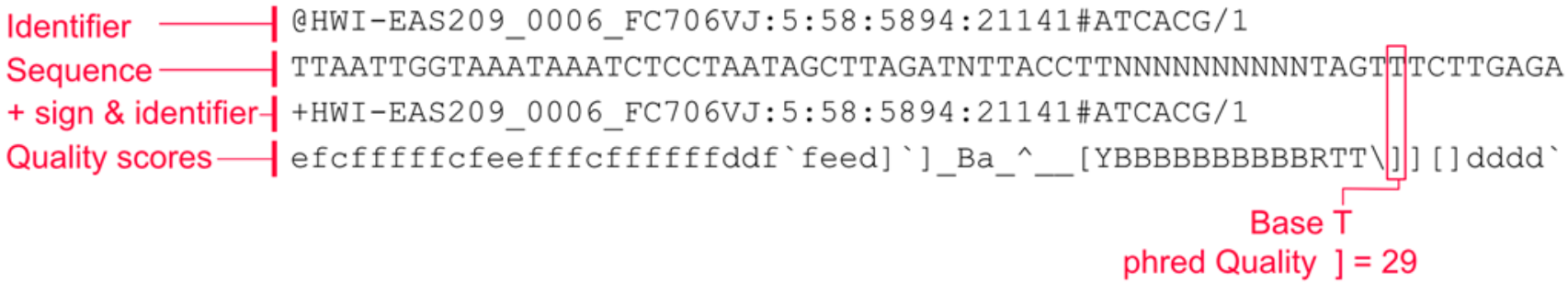
- Риды – это последовательности нуклеотидов, которые соответствуют участку на фрагменте секвенированной ДНК (или всему фрагменту целиком).
- Первое поколение секвенирования (Сэнгер) продуцирует риды длиной до 1500 нуклеотидов, второе – до 300, третье – более 2 млн.

Способ записи нуклеотидных последовательностей

- Один из наиболее распространённых способов записи нуклеотидных и аминокислотных последовательностей – формат FASTA.
- Этот формат подразумевает, что у каждой последовательности есть своё название, которое отражается строкой выше, начинающейся со знака “>”.

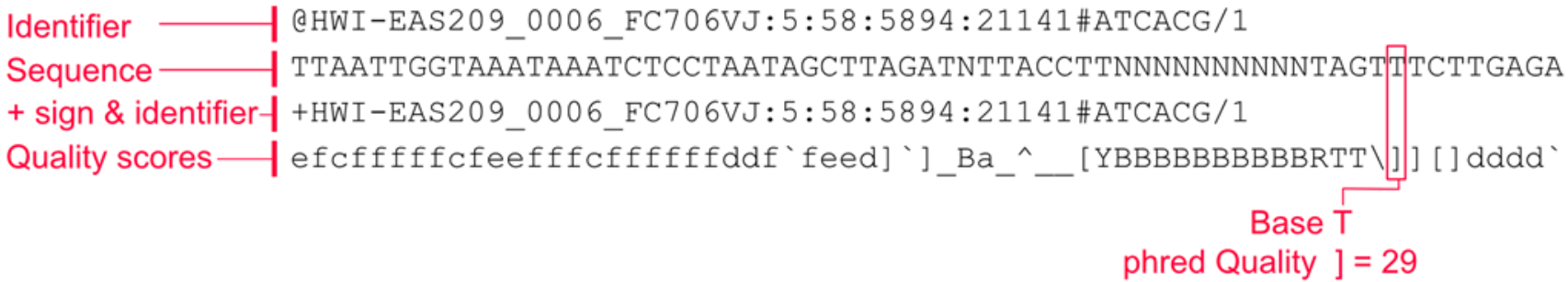
Формат FASTQ

- FASTQ формат является расширением FASTA, которое используется для хранения результатов секвенирования.
- Каждая последовательность в формате FASTQ занимает ровно 4 строки любой длины.



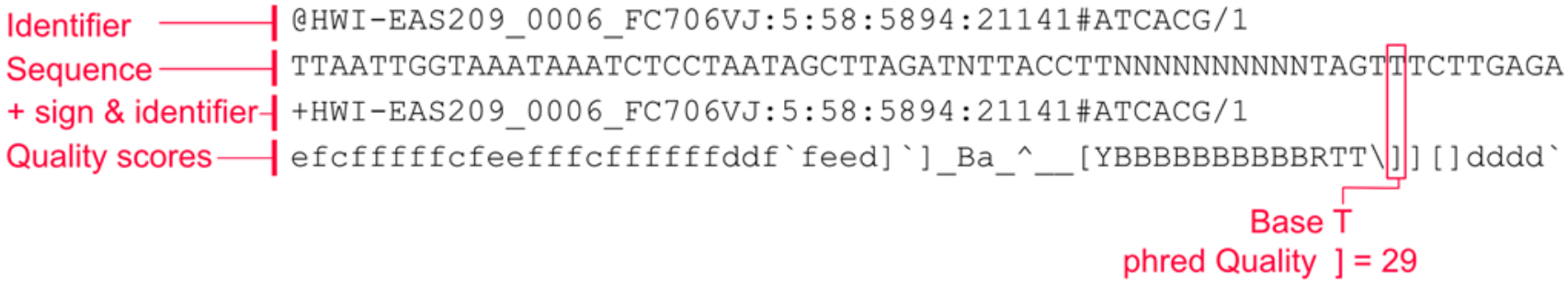
Формат FASTQ: идентификатор

- Первая строка начинается с “@” и является идентификатором, в котором может быть записана техническая информация об идентификаторе (координатах) ячейки, номере лейна и т.д.



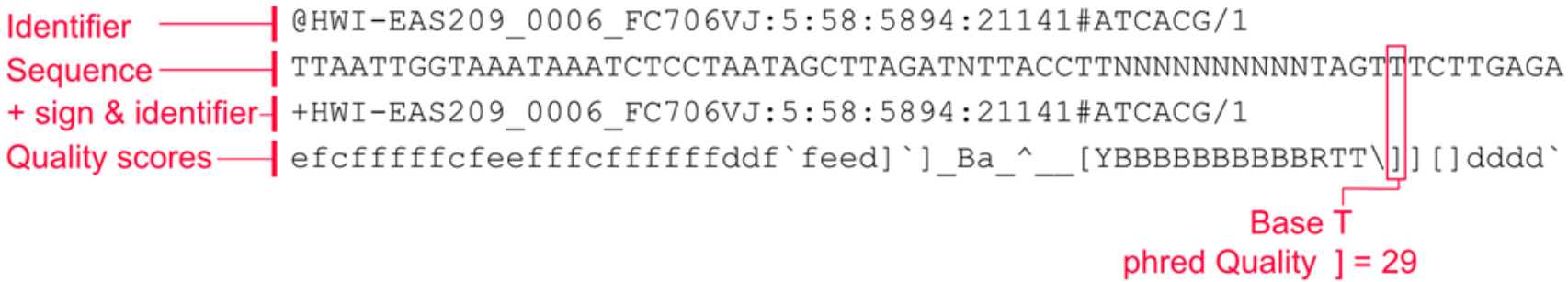
Формат FASTQ: последовательность и третья строка

- Вторая строка содержит последовательность символов, как правило, [ATGCN], N – неопределённый нуклеотид.
- Третья строка начинается со знака “+” нужна для маркировки окончания последовательности и может не содержать никакой информации. Часто она состоит из “+” и переноса строки.



Формат FASTQ: значения качества

- Четвёртая строка кодирует значения качества для каждого символа последовательности во второй строке.
- Для краткости записи значения качества по шкале PHRED представлены символами в кодировке ASCII.



Шкала Phred

- Шкала Phred – это мера качества идентификации нуклеотидов секвенаторами.
- Значение по шкале Phred (Q) логарифмически связано с вероятностью ошибки при секвенировании (p):

$$Q_{Phred} = -10 \log_{10} p$$

Значение по шкале Phred	Вероятность ошибки	Точность секвенирования
10	1 из 10	90%
20	1 из 100	99%
30	1 из 1000	99,9%
40	1 из 10000	99,99%

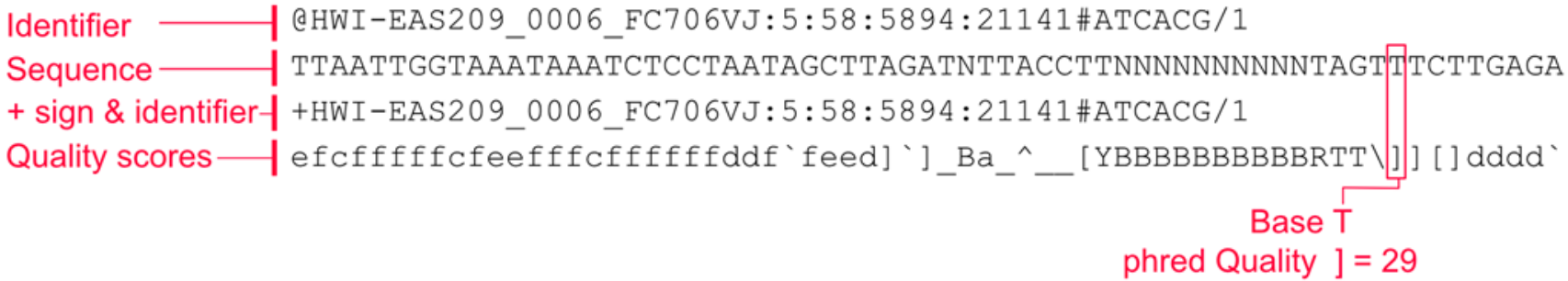
ASCII кодировка

- Полностью писать очки качества в четвёртой строке FASTQ файла неудобно, т.к. это приводит к неизбежному удлинению её по сравнению с первой строкой.
- Таким образом, числовые значения шкалы кодируются в виде ASCII символов.

Letter	Quality value	Estimated error probability
(	40	20%
7	55	0.6%
F	70	0.02%
U	85	0.0006%
d	100	0.00002%

Формат FASTQ: ASCII

- Можно достаточно быстро определить качество секвенирования при взгляде на четвёртую строку: если в ней преимущественно встречаются буквы, то строка “хорошего” качества, если иные символы – “плохого”.
- NB! Качество нуклеотидов, обозначенных как N, определять не вполне корректно – они уже не могут быть определены (обычно их Q < 2).



Какие FASTQ файлы в итоге получаются

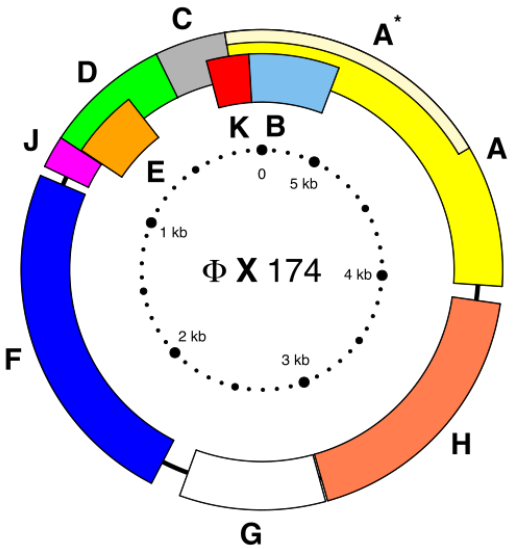
- Современные приборы способны автоматически определять, к какой пробе (sample) относится тот или иной рид, и группируют их по соответствующим файлам. Этот процесс называется демультиплексированием.
- Однако, не все риды удаётся распределить таким образом: ошибки и внедрение PhiX генома в эксперимент может приводить к появлению неопределённых ридов (Undetermined.fastq)

Run Statistics

Sample Name	Total # of Single-End Reads	Total # PF Reads	% Undetermined	% PhiX Aligned
1.0	23,251,484	21,809,529	10.4	2.5

PhiX

- PhiX – это бактериофаг с одноцепочечной ДНК и геномом из 5386 п.о. Геном PhiX используется для контроля качества генерации кластеров, секвенирования и выравнивания.
- PhiX добавляется в библиотеку в низкой известной концентрации вместе с пробами.
- Обычно риды с PhiX помещаются в Undetermined.fastq, поэтому беспокоиться о присутствии их в файлах с пробами в общем случае не стоит.



Контроль качества ридов

- Современные секвенаторы вырабатывают десятки миллионов последовательностей за пробег (run). Прежде, чем анализировать такие массивы данных следует убедиться, что Ваши сырые данные соответствуют некоторому уровню качества, необходимого для получения адекватных результатов.
- Для контроля качества ридов используют программу FastQC или производные от неё.
- FastQC генерирует отчёт из нескольких разделов, каждый из которых, кроме первого, может иметь несколько состояний: нормальное (зелёная галка), требующее внимания (warning, жёлтый “!”) и аномальное (failure, красный “x”).

Basic statistics

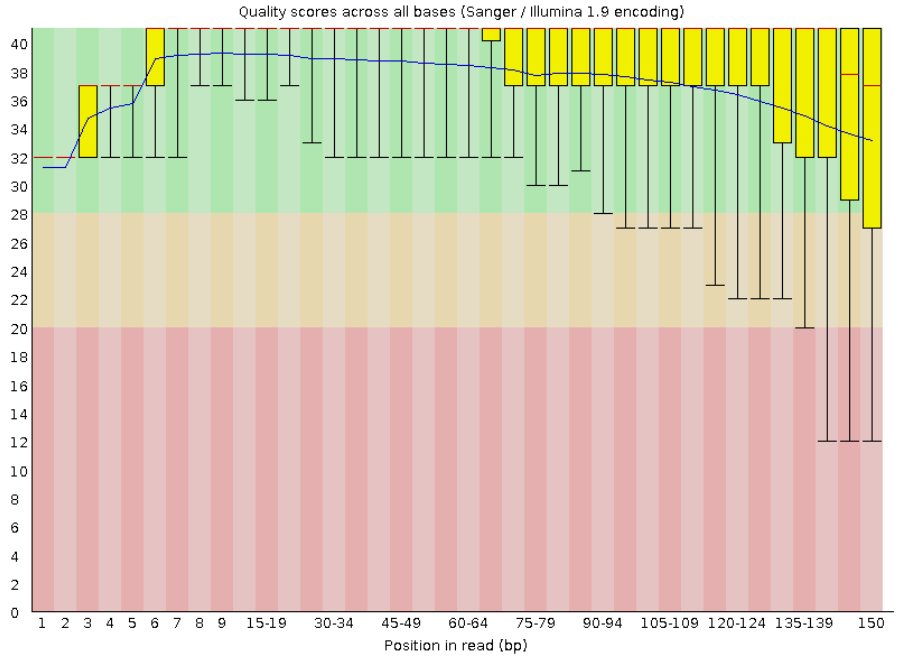
- File type – файлы бывают двух типов: готовые для анализа риды (fastq) или технические данные, которые необходимо перевести в риды.
- Encoding – кодировка Phred шкалы.
- Total Sequences – количество прочитанных последовательностей.

Basic Statistics

Measure	Value
Filename	Mov10_oe_1.subset.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	305900
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

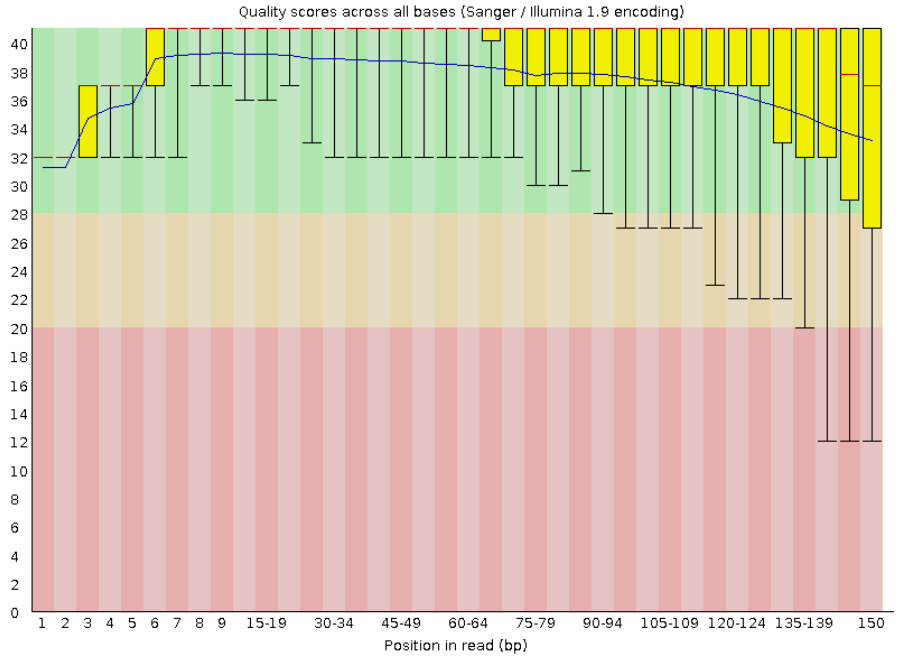
Base quality graph

- Второй раздел показывает, как распределено качество прочтения каждого нуклеотида по всем ридам.
- Синяя линия – средние значения.
- Красная полоса – медиана.
- Жёлтый ящик – 25й и 75й перцентили.
- Усы – 10й и 90й перцентили.



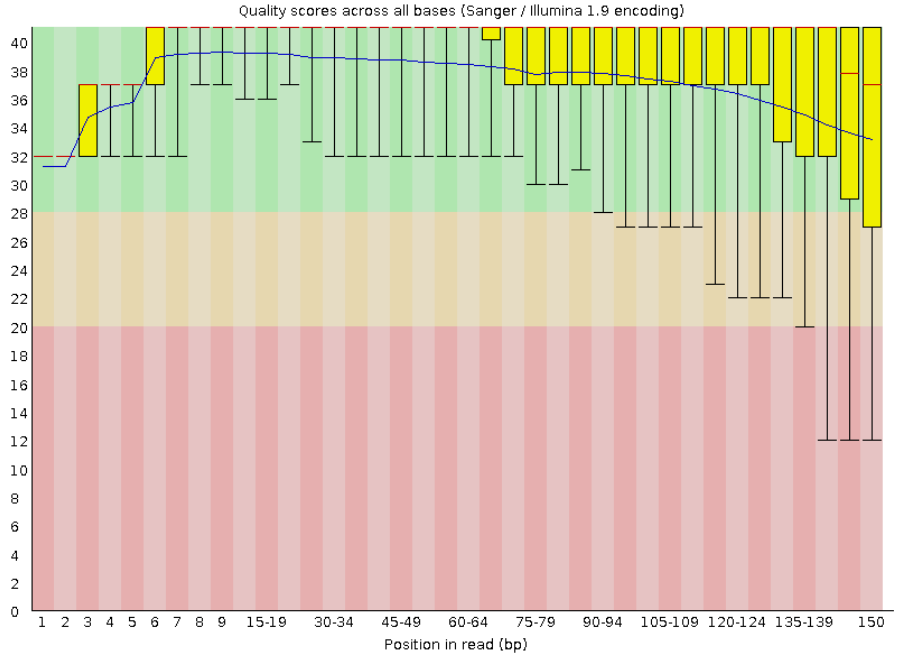
Base quality graph

- Оцениваются Phred-значения 25го квартиля и медианы по всем позициям.
- Warning:
  - 25й квартиль меньше 10, но больше 5;
  - Медиана меньше 25, но больше 20.
- Failure:
  - 25й квартиль меньше 5;
  - Медиана меньше 20.



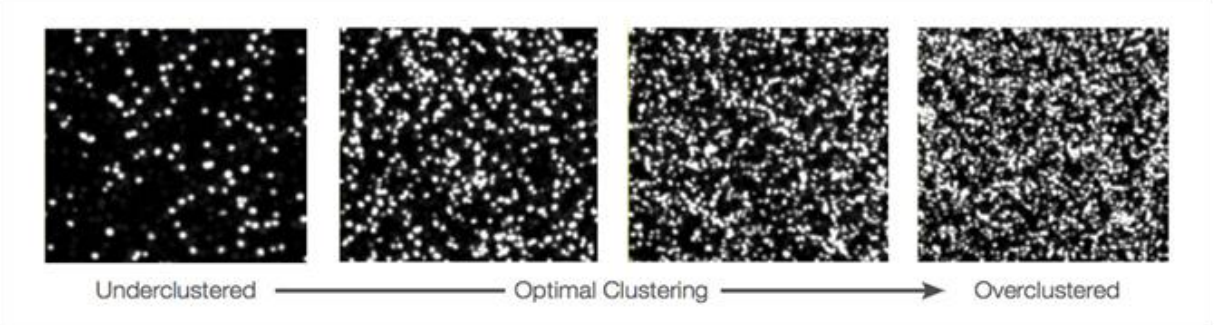
Почему качество падает “в конце”?

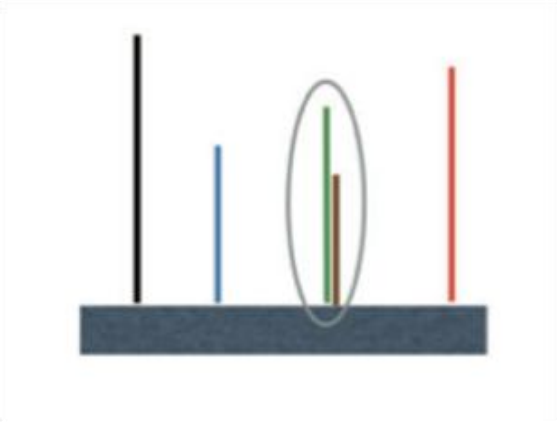
- К концу рида качество обычно падает. Это нормально:
  - Деградирование флуорофоров;
  - Часть ампликонов не элонгируется, они остаются короткими и не дают сигнала;
  - Кластер рассинхронизируется, т.е. каждый рид в нём начинает читаться немного в другом месте, чем большинство, прибавляя “шум” от кластера.



По каким причинам может падать общее качество?

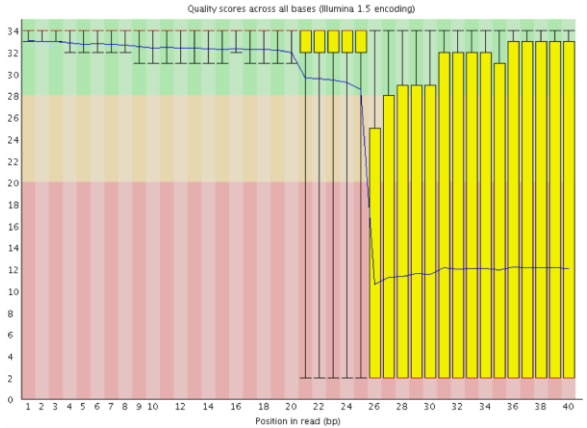
- Перегруз кластерами: когда кластеров слишком много, сигналы от соседних кластеров могут интерферировать друг с другом. Камера будет воспринимать их как один кластер, дающий смешанный, низкокачественный сигнал.
- В таком случае падение качества будет наблюдаться вдоль всего рида.

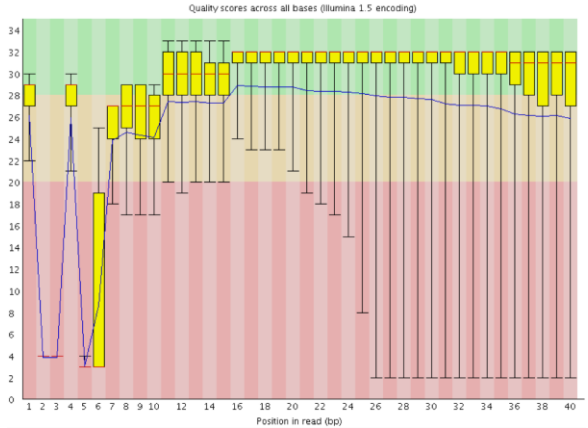


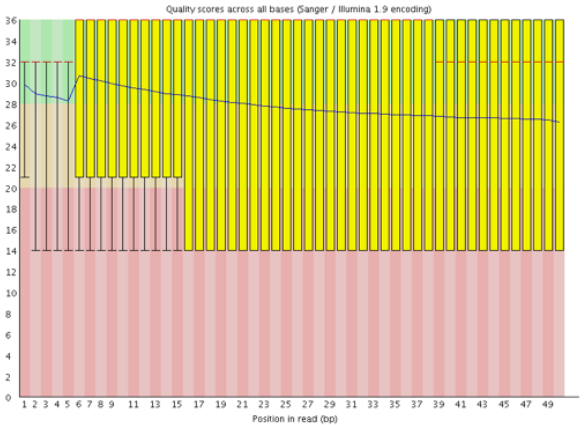


Поломка инструмента

- Возможна поломка секвенатора, тогда качество будет низким не вдоль всего рида, а фрагментарно:

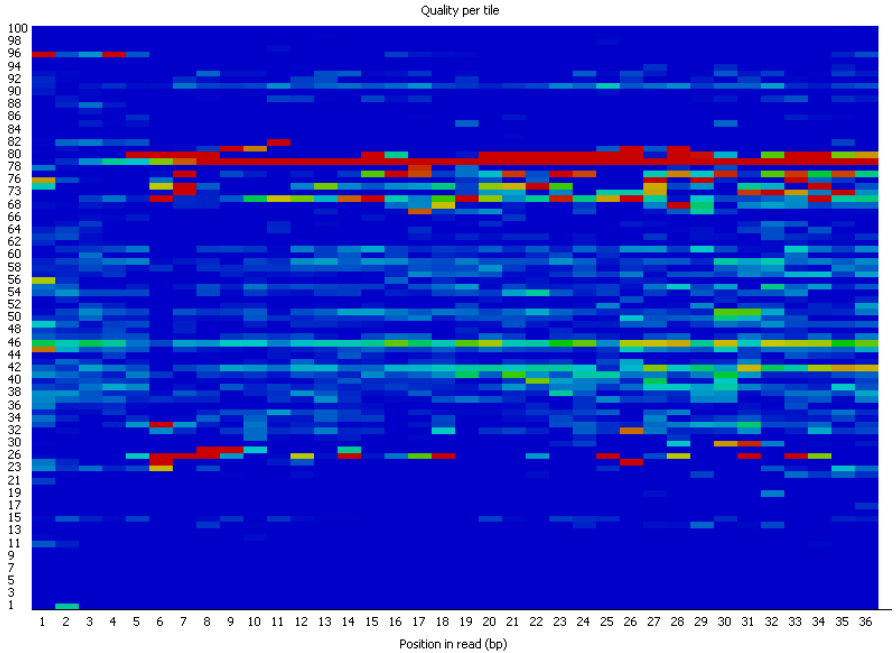






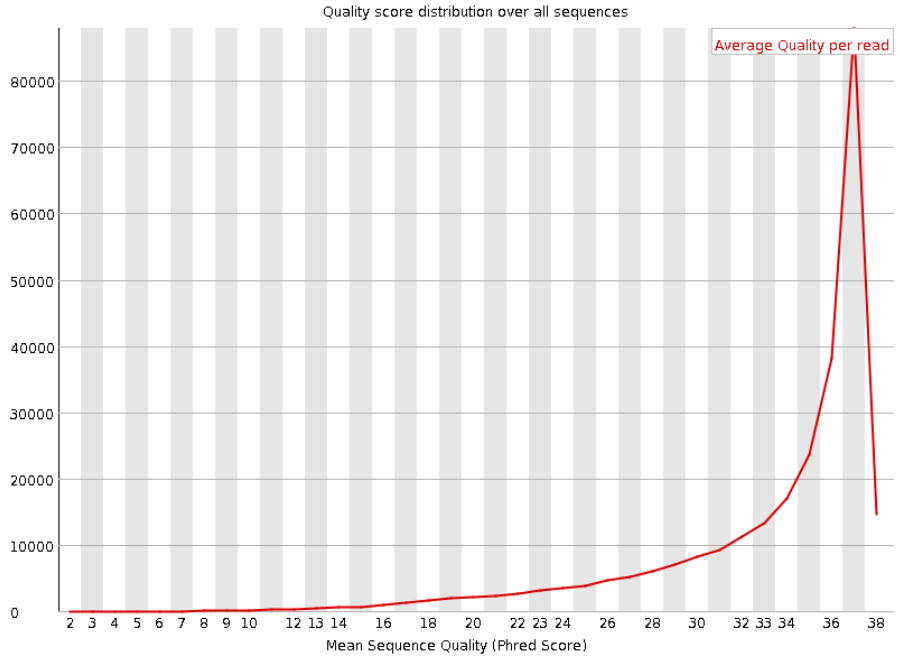
Per Tile Sequence Quality

- При помощи этого графика можно локализовать повреждение или перегрузку картриджа.
- Warning: Phred по любому tile в некоторой позиции отличается больше, чем на 2, но меньше, чем на 5 от среднего Phred по позиции.
- Failure: Phred в некотором tile в хоть одной позиции отличается больше чем на 5



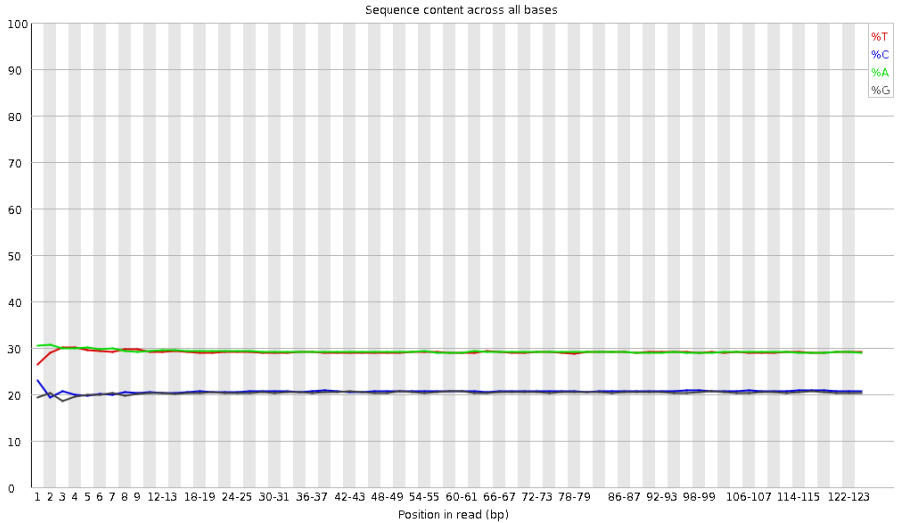
Per Sequence Quality Scores

- График распределения ридов по качеству: значительное число ридов с худшим, чем у большинства, качеством, может свидетельствовать о локальном повреждении ячейки, например.
- Warning: мода среднего качества прочтения нуклеотидов в риде менее 27, но более 20 (частота ошибки от 0,2% до 1%).
- Failure: мода менее 20 (1%).



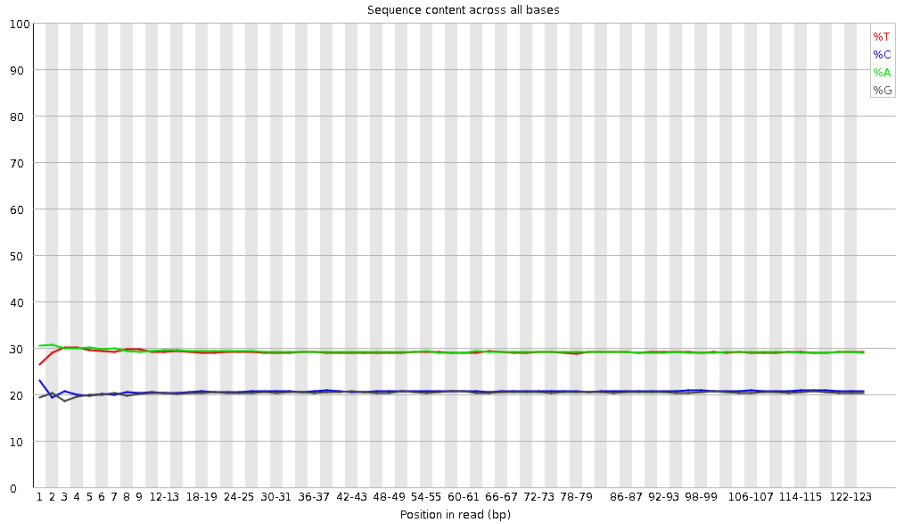
Per Base Sequence Content

- На графике отражена доля каждого нуклеотида (ATGC) в каждой позиции по всем ридам.
- Превалирование одних оснований над другими говорит о том, что, возможно, в библиотеке есть большое количество копий одного и того же участка (overrepresented sequences).
- Однако, это может быть обусловлено специфичной фрагментацией (фермент ищет определённые локусы).



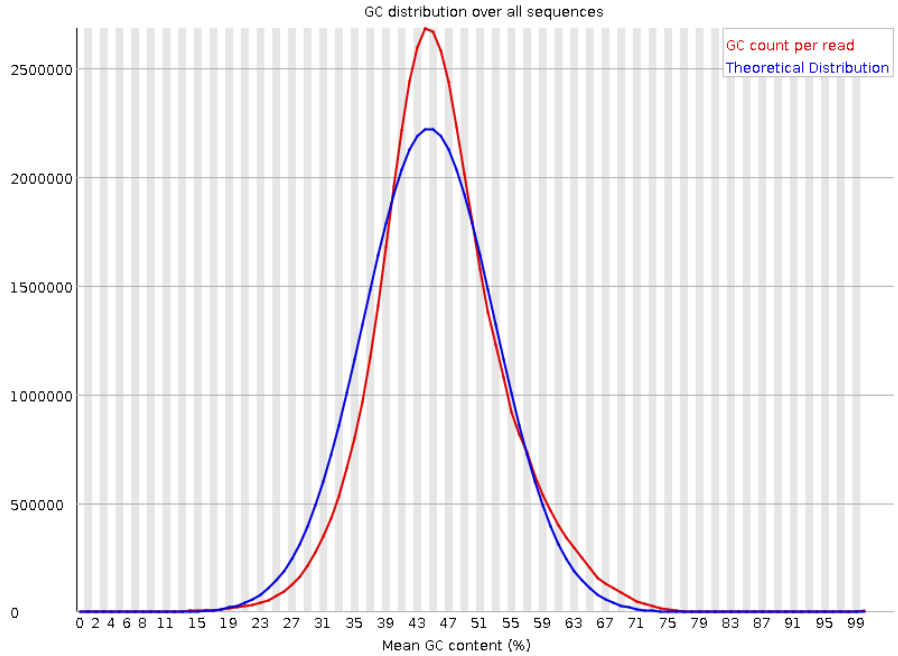
Per Base Sequence Content

- Warning: разница между долями A и T или G и C составляет от 10 до 20% в любой позиции.
- Failure: разница между долями A и T или G и C превышает 20% в любой позиции.



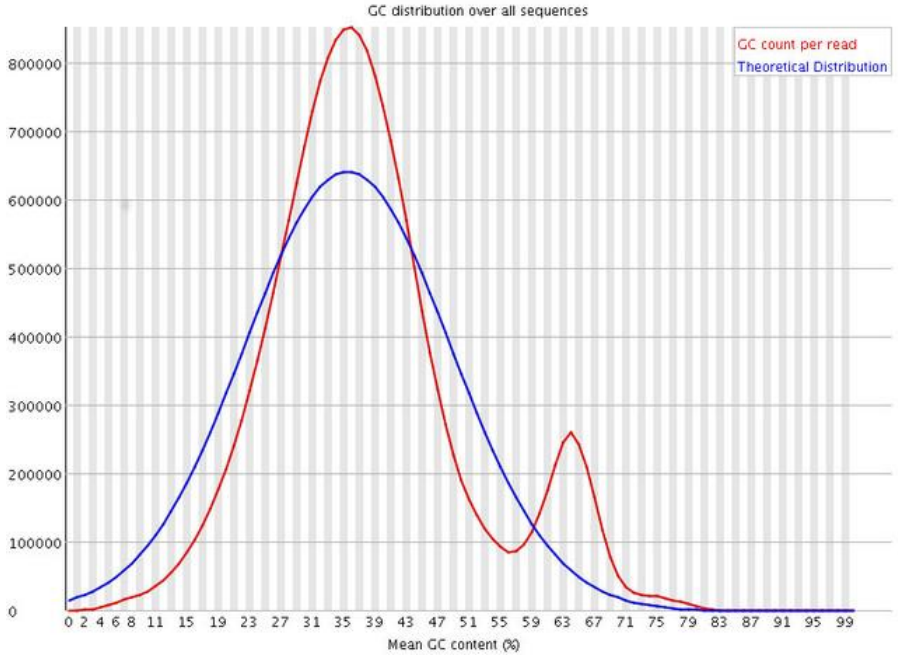
Per Sequence GC Content

- Распределение ридов по их GC-составу в сравнении с нормой.
- Для того, чтобы анализировать этот график, нужно понимать, каковы свойства секвенируемого Вами (участка или целиком) генома: является ли нормальным наблюдаемое распределение для изучаемого объекта?
- Наличие двух и более пиков может свидетельствовать о контаминации.



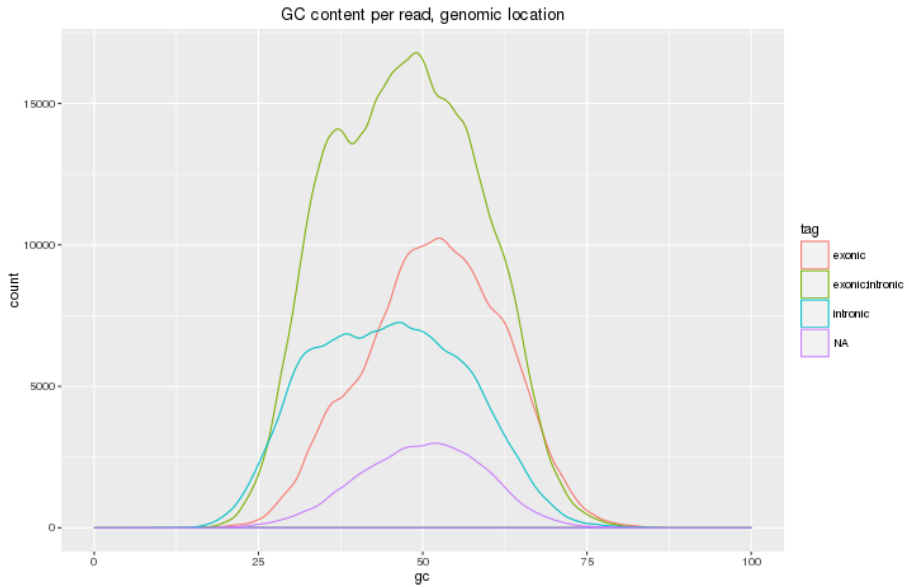
Per Sequence GC Content

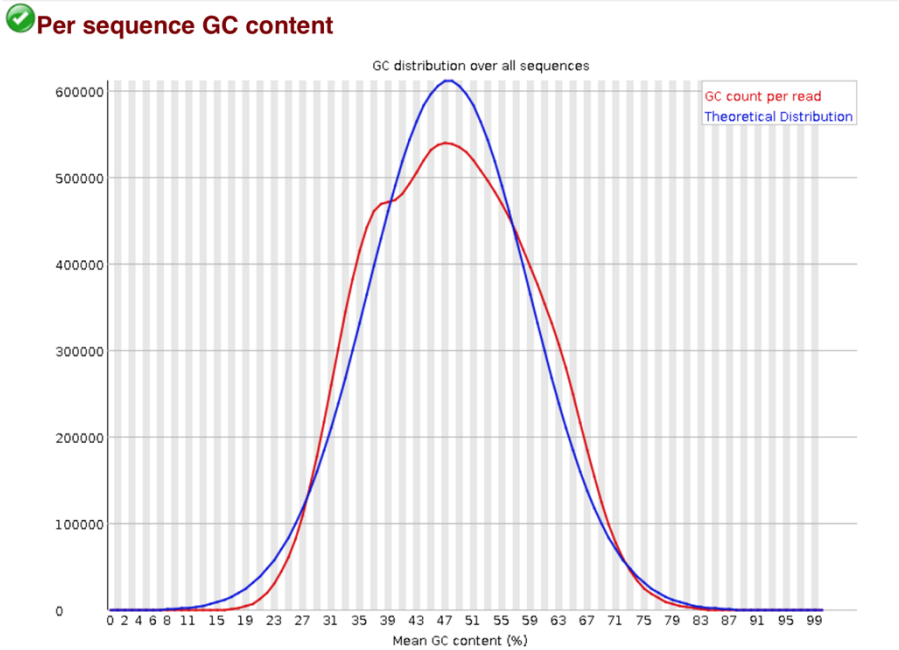
- Контаминация



Per Sequence GC Content

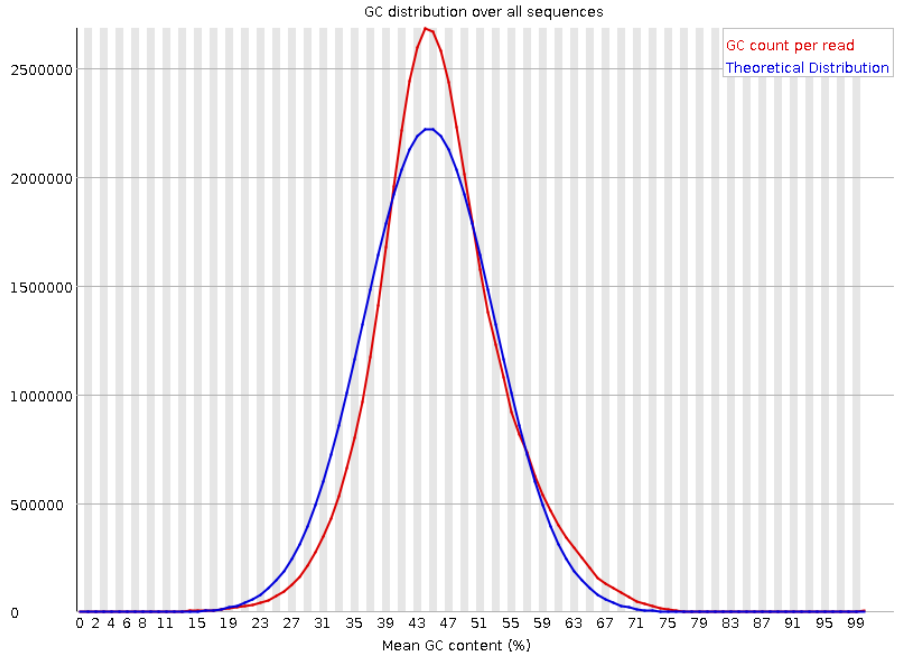
- Экзоны и интроны





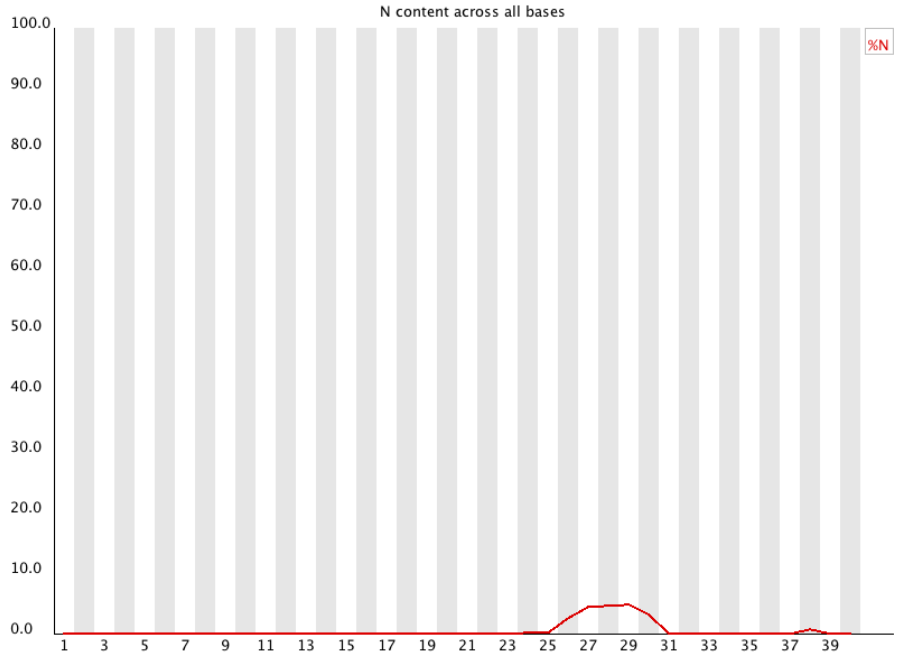
Per Sequence GC Content

- Warning: доля отклонившихся от нормального распределения ридов лежит в пределах от 15 до 30%.
- Failure: доля отклонившихся от нормального распределения ридов превышает 30%.



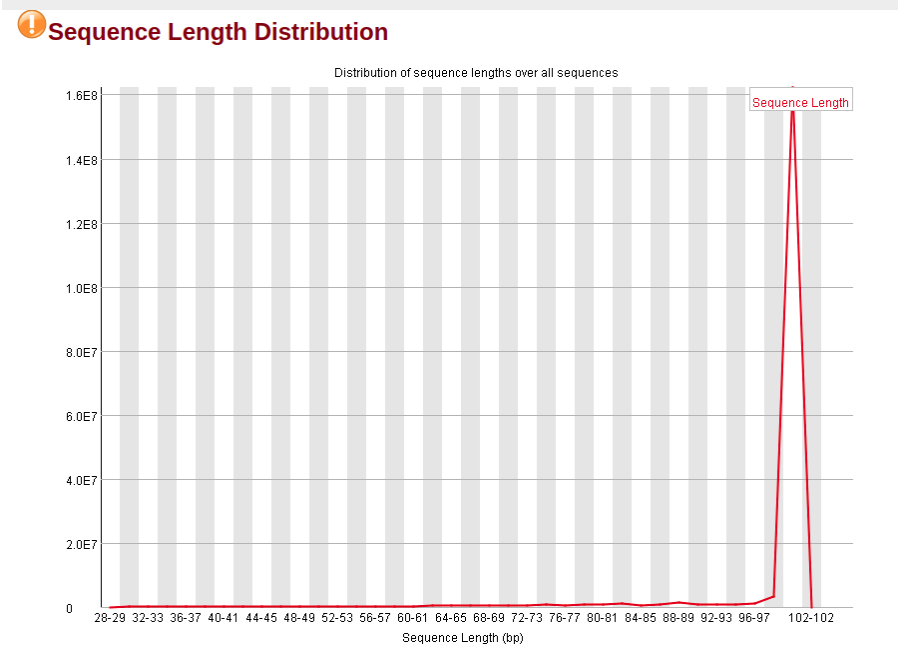
Per Base N Content

- Если секвенатор не может установить нуклеотид, вместо него ставится N.
- Доля N в позициях по всем ридам отражена на графике.
- Warning: 5%<N<=20%, N - доля N в любой позиции.
- Failure: доля N > 20%.



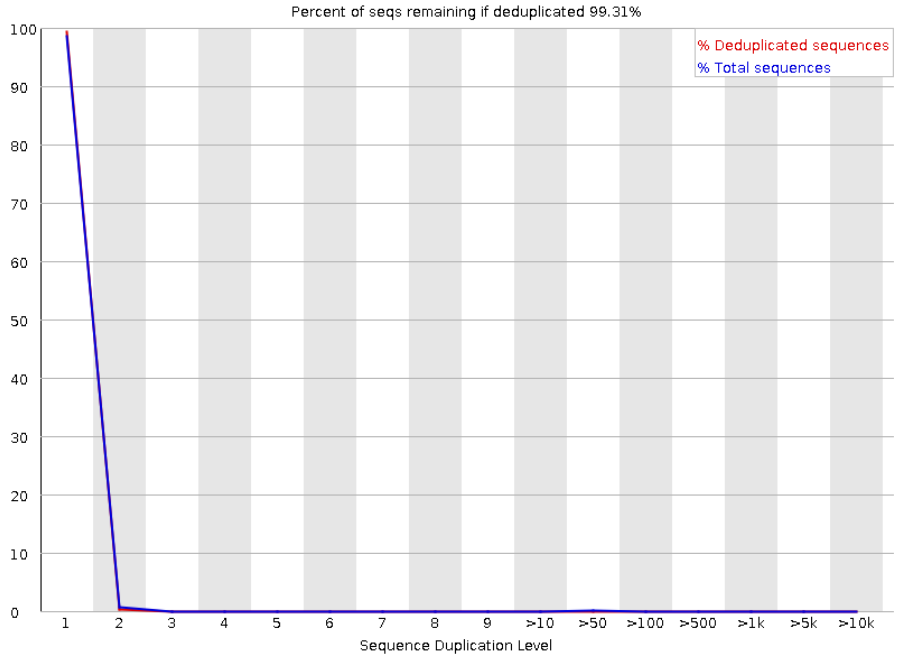
Sequence Length Distribution

- Распределение ридов по длине может быть различным – это зависит от того, какая система секвенирования применялась, а также осуществлялась ли предварительная обработка (обрезка) ридов.
- Warning: не все риды одинаковой длины.
- Failure: есть риды с длиной 0.



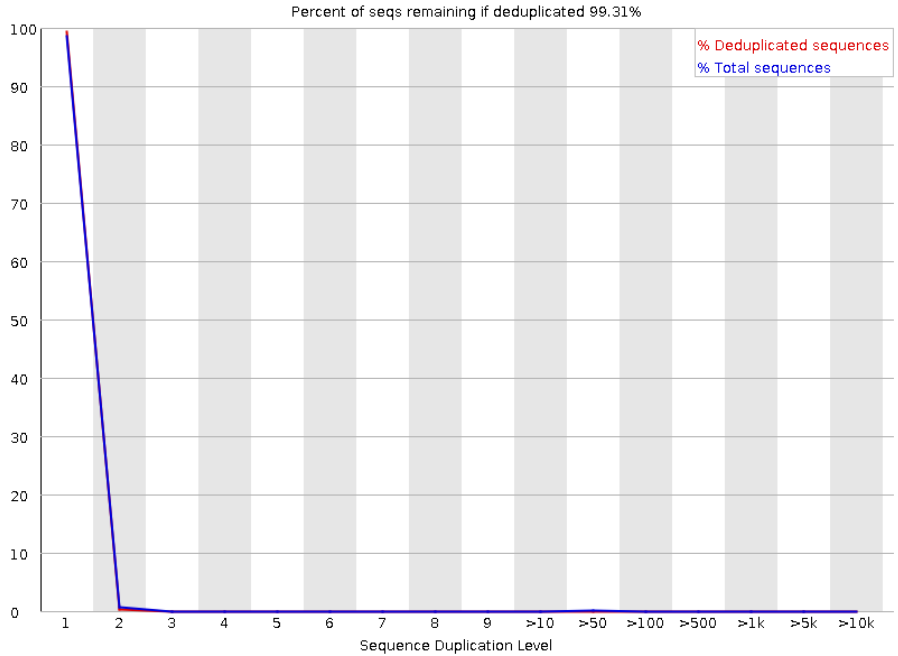
Sequence Duplication Levels

- График отражает доли ридов по количеству их дупликатов (ПЦР или оптические).
- Может выглядеть очень плохо, если в ходе приготовления библиотеки использовалась наработка ампликонов при помощи ПЦР (вместо гибридизационного захвата).



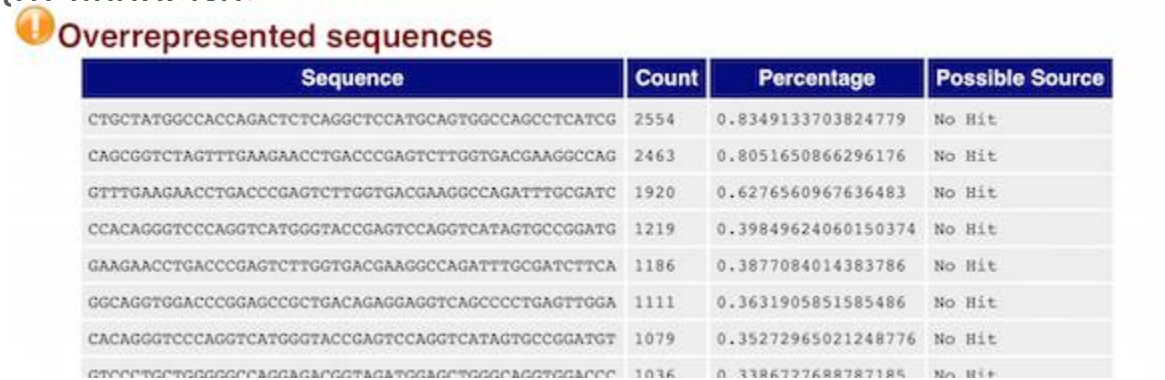
Sequence Duplication Levels

- Warning: дупликаты составляют более 20, но менее 50% от всех ридов.
- Failure: более 50% дупликатов.



Overrepresented Sequences

- Таблица, показывающая (суб)последовательности (от 20 п.о.), которые встречаются в более чем 0,1% ридов.
- Например, если в библиотеку были занесены ампликоны из другого организма, то вместе с аномальным GC-составом, данная таблица может указать, какой вид явился источником контаминации (нужно провести BLAST последовательности)

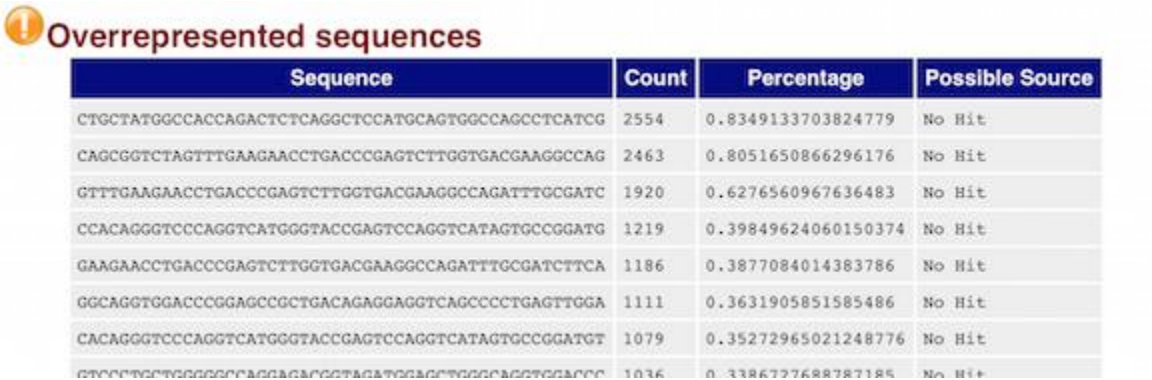


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit

Overrepresented Sequences

- Warning: в таблице есть хоть один hit, при этом все они встречаются в менее 1% ридов.
- Failure: в таблице есть последовательности, встречающиеся в более 1% ридов.

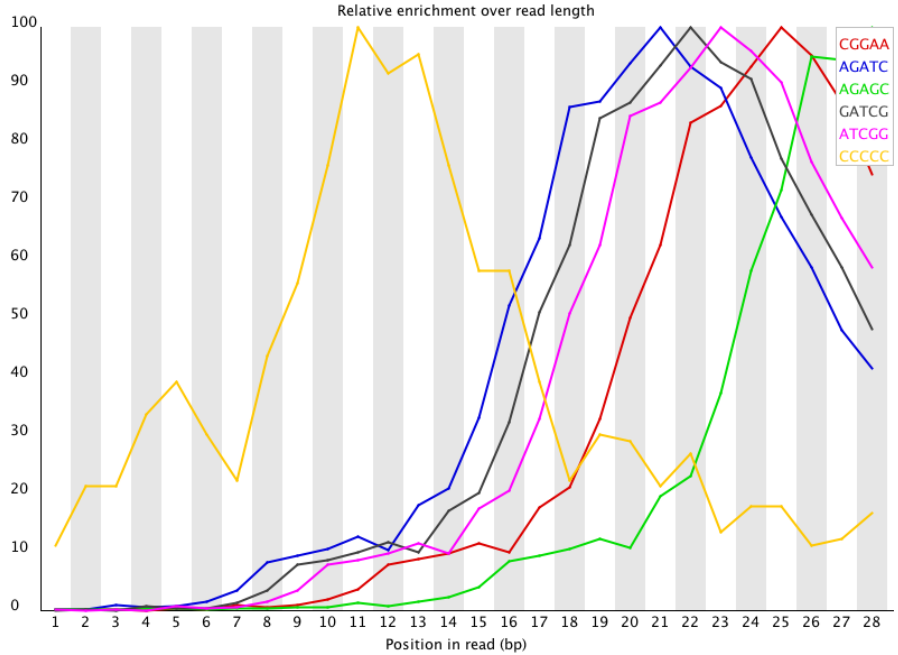


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit

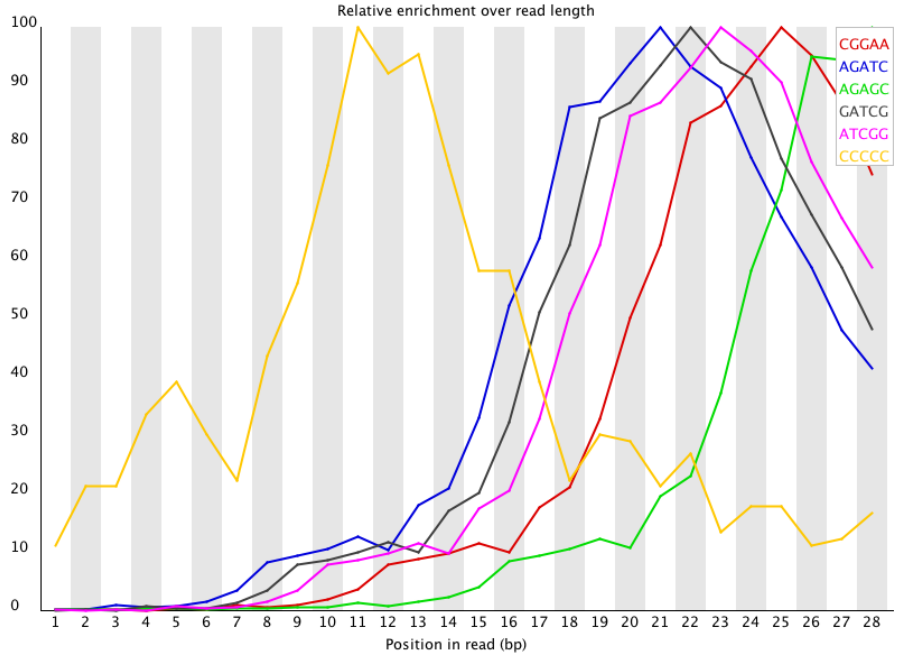
Kmer Content

- В случае, если реальные дупликаты, в силу низкого качества прочтения, опознаются как различные последовательности, или если они находятся в разных частях ридов, обнаружить контаминацию можно при помощи меньших, чем 20 п.о., субпоследовательностей – k-меров.
- Это может свидетельствовать о контаминации димерами адаптеров (когда в риде нет insert).



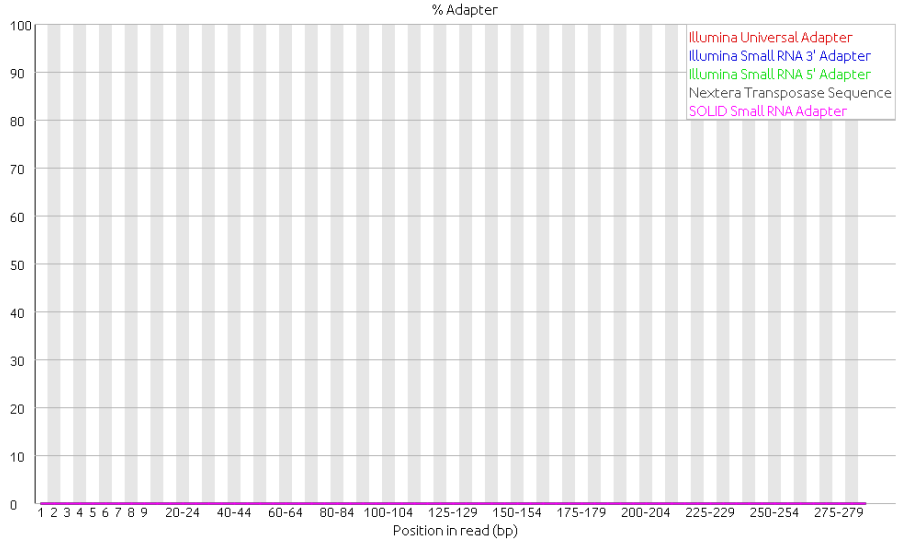
Kmer Content

- На данном графике не должно быть преимущественного отбора одних k-меров над другими, они должны быть равномерно распределены.
- Warning: несбалансированный k-мер с p-значением от 10-5 до 0,01.
- Failure: p-значения менее 10-5.



Adapter Content

- График, показывающий распределение адаптеров по позициям ридов.
- Warning: последовательность адаптера в более 5, но менее 10% ридов.
- Failure: адаптер в более 10% ридов.
- Неудовлетворительный результат говорит только о том, что нужно осуществить тримминг ридов.



Обрезка ридов

- Некоторые аспекты “неправильности” fastq файлов можно нивелировать при помощи:
  - Удаления дупликатов на дальнейших этапах работы с выравниванием (bam-файлом) – об этом позднее;
  - Тримминга адаптеров;
  - Обрезки ридов по качеству.
- Для последних двух действий существует масса программ, которые можно использовать (см. ДЗ).