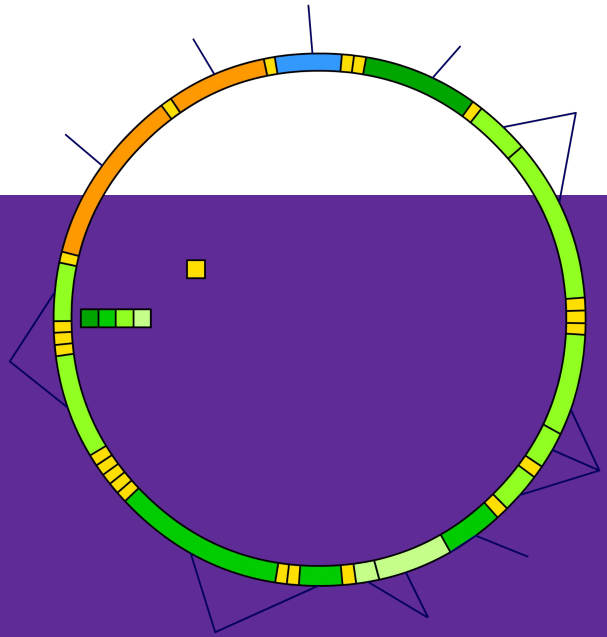
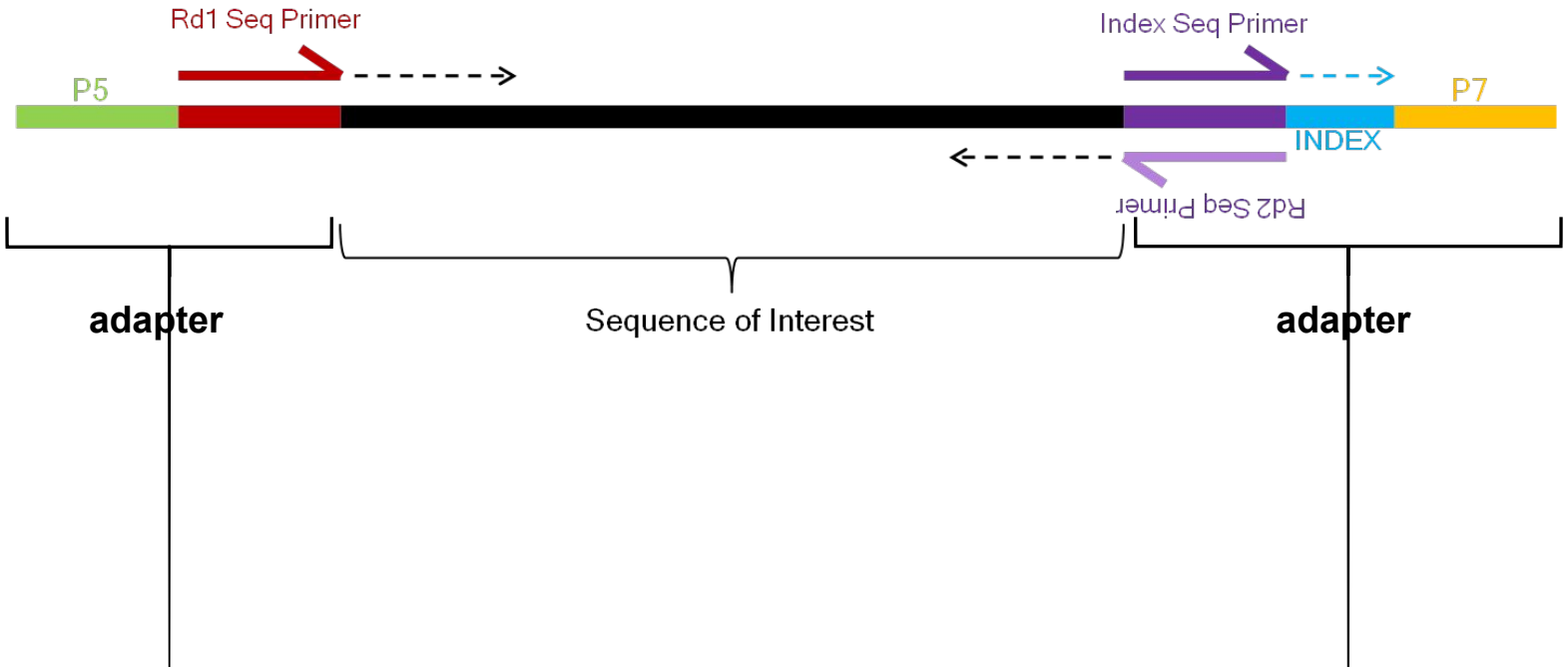
Первичная обработка геномных данных.

Сириус
Образовательный центр



Повторение: Прочтения

Чтения (риды) - **фрагменты ДНК**, полученные при **секвенировании**.



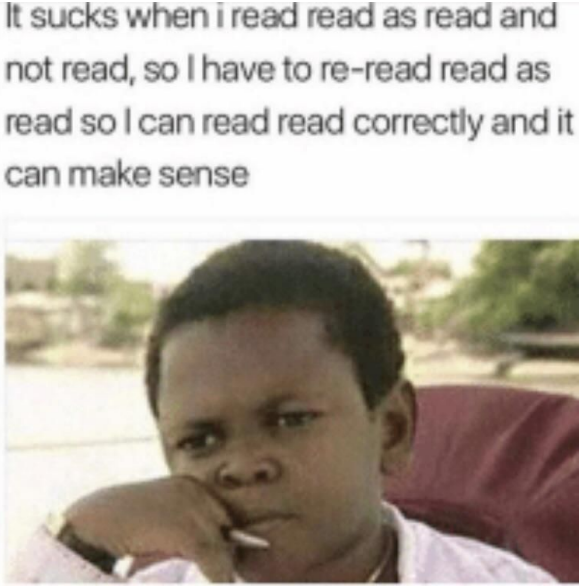
Повторение: Прочтения

- **Нормальный рид**

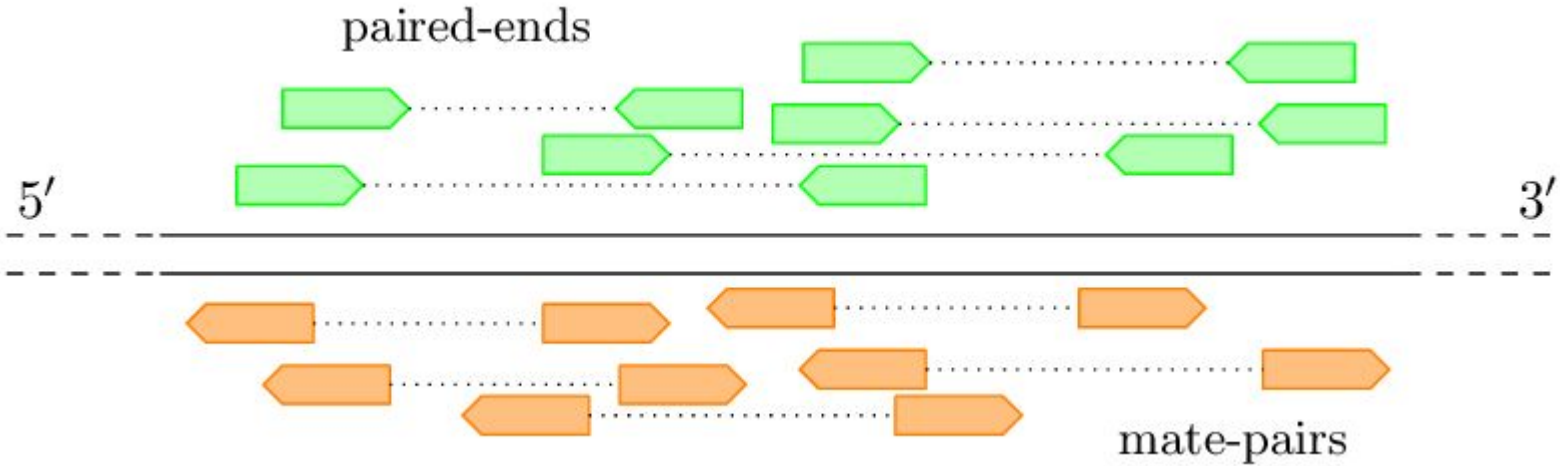
- **Димер адаптеров**

- **Сквозное прочтение**

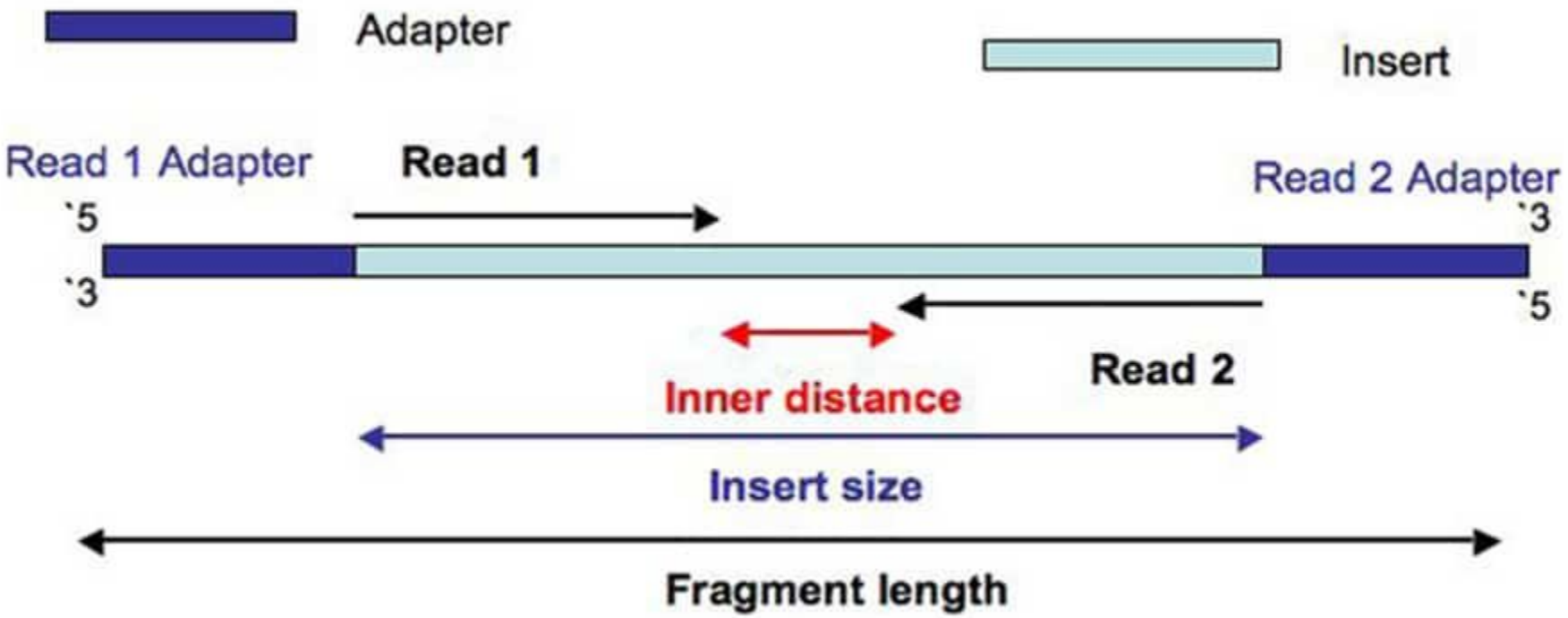
Фрагмент образца ДНК короче длины рида.



Повторение: Парные и непарные прочтения.



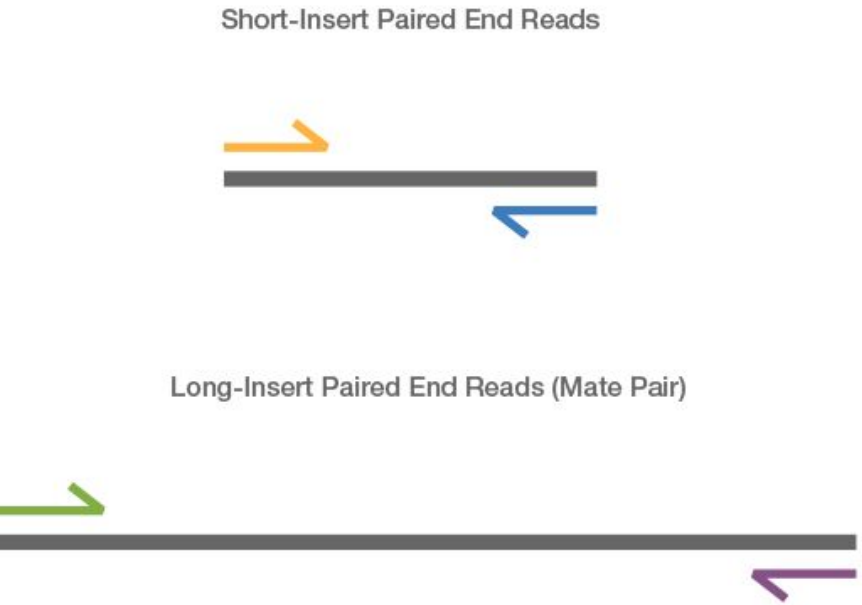
Повторение: Парные и непарные прочтения.

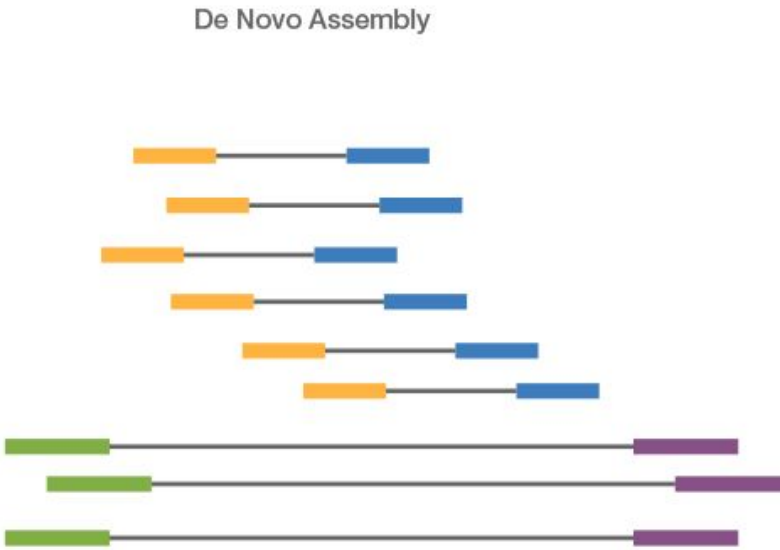


Повторение: Парные и непарные прочтения.

При **секвенировании парных прочтений** к концам ДНК пришивают **два вида адаптеров**, так чтобы к разным концам одного фрагмента ДНК были присоединены разные адаптеры. После **мостиковой амплификации** противоположно ориентированные копии исходного фрагмента не удаляются.

Повторение: Прочтения со вставками.





Повторение: Прочтения со вставками.

Секвенирование спаренных концов (англ. **Mate Pair Sequencing**).

Позволяет секвенировать две **последовательности**, изначально располагающиеся в геноме **на расстоянии до 5000 нуклеотидов** друг от друга, **как единое целое**. Такой подход может быть полезен при **de novo секвенировании**, при **поиске мутаций**, для корректной **сборки генома**.

FASTA

текстовый формат для **нуклеотидных** или полипептидных последовательностей, в котором нуклеотиды или аминокислоты обозначаются при помощи **однобуквенных кодов**

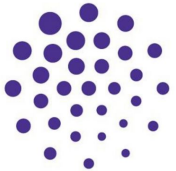
Последовательности в формате FASTA начинаются с **однострочного описания**, за которым следуют строки, содержащие собственно последовательность. **Описание отмечается символом «больше» («>»)** в первой колонке. Слово за этим символом и до первого пробела является идентификатором последовательности

В один файл могут быть записаны несколько последовательностей, таким образом получается **мульти-FASTA** файл, однако перед каждой последовательностью должен стоять свой идентификатор.

Расширение: `.fas, .fasta, .fna, .ffn, .faa, .frn, .afa, .mfa`

Узнать больше: https://ru.wikipedia.org/wiki/FASTA

FASTA



Порядок: от 5'- к 3'-концу для нуклеиновых кислот, от N- к C-концу для аминокислот

Допускаются пробелы, оба регистра.

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

FASTQ

(FASTA + Quality) текстовый формат данных, используемый для представления биологической **последовательности** (обычно нуклеотидной) и **показателей качества каждого элемента** последовательности.

4 строки на каждую последовательность.

Строка 1: начинается с символа «@», за ней следует **идентификатор последовательности** и описание.

Строка 2: **символы последовательности**.

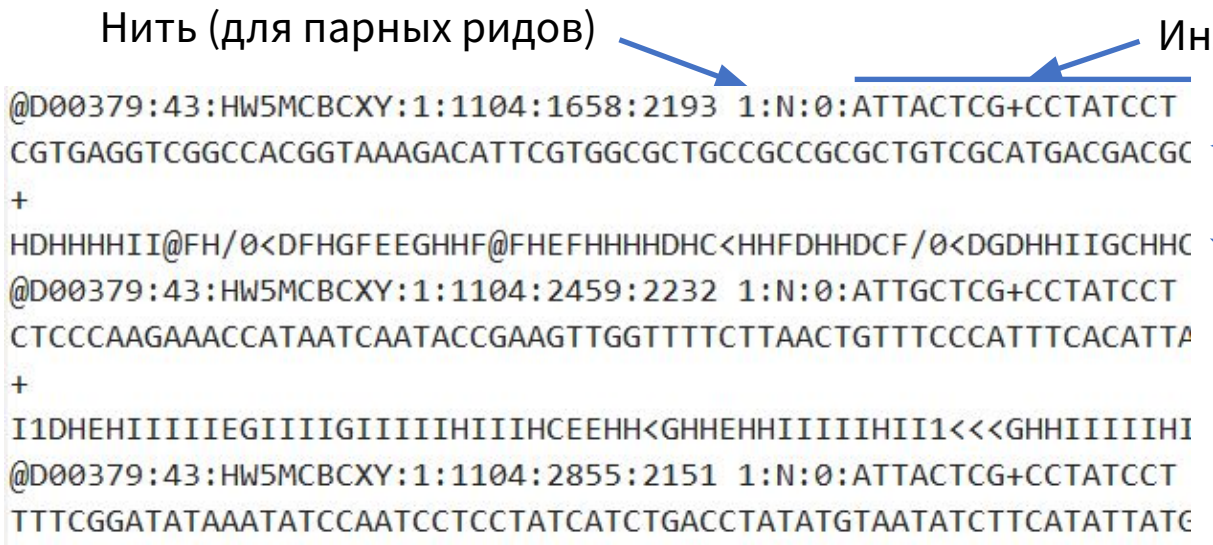
Строка 3: начинается с символа «+» и является необязательной.

Строка 4: **значения качества** для последовательности в строке 2, должна содержать то же количество символов.

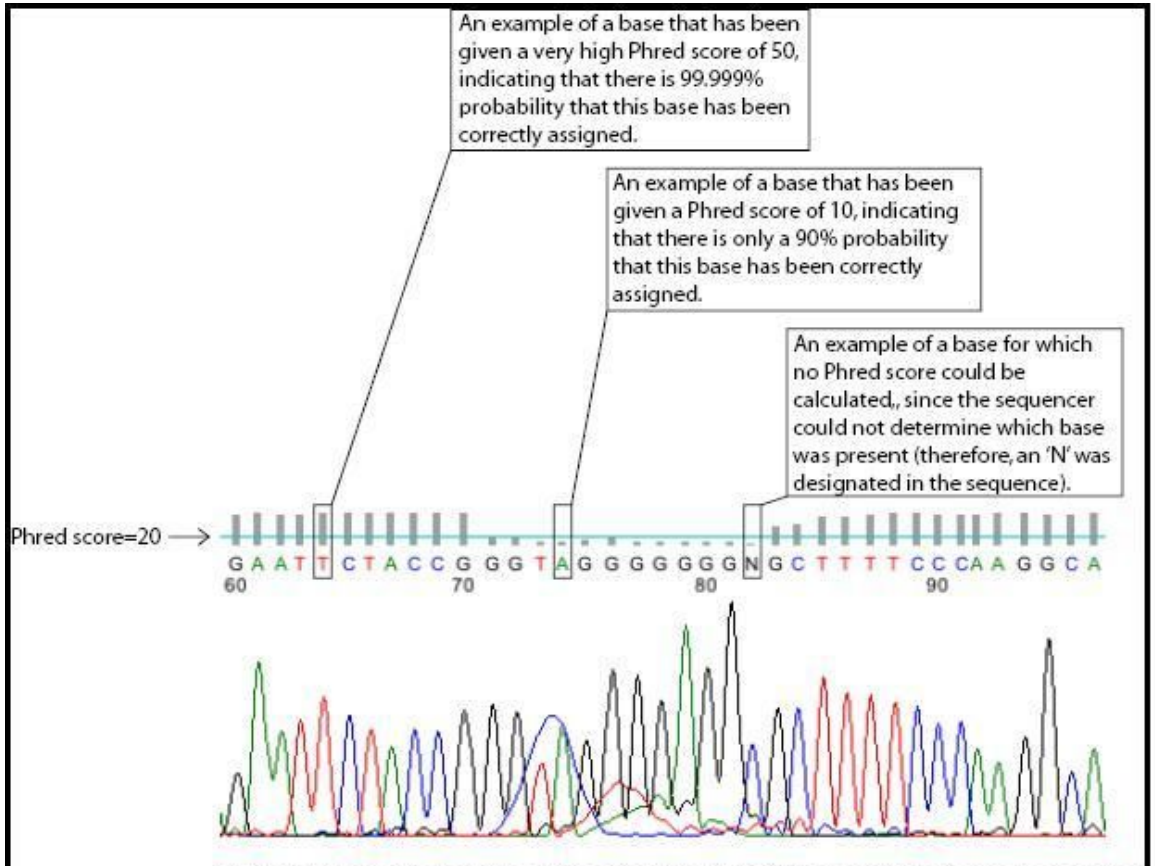
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

https://ru.wikipedia.org/wiki/FASTQ

FASTQ


Нить (для парных ридов)
Индекс
@D00379:43:HW5MCBCXY:1:1104:1658:2193 1:N:0:ATTACTCG+CCTATCCT
CGTGAGGTCGGCCACGGTAAAGACATTCGTGGCGCTGCCGCCGCGCTGTCGCATGACGACGC
Последовательность
+
HDHHHHII@FH/0<DFHGFEEGHHF@FHEFHHHHDHC<HHFDHHDCF/0<DGDHHIIGCHHC
@D00379:43:HW5MCBCXY:1:1104:2459:2232 1:N:0:ATTGCTCG+CCTATCCT
CTCCCAAGAAACCATAATCAATACCGAAGTTGGTTTTCTTAACTGTTTCCCATTTCACATTA
Качество
+
I1DHEHIIIIIEGIIIIGIIIIIHIIIHCEEHH<GHHEHHIIIIIHII1<<<GHHIIIIIHI
@D00379:43:HW5MCBCXY:1:1104:2855:2151 1:N:0:ATTACTCG+CCTATCCT
TTTCGGATATAAATATCCAATCCTCCTATCATCTGACCTATATGTAATATCTTCATATTATG

Качество прочтения



Phred Score

$$Q = -10 \lg (P)$$

P - вероятность ошибки, вычисленная по форме пика.

Хороший показатель для Illumina:

95-75% нуклеотидов с Q > 30

Q-value	Вероятность ошибки
30	**0.001** (99.9% точность)
20	**0.01** (99% точность)
10	**0.1** (90% точность)

Phred Score

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

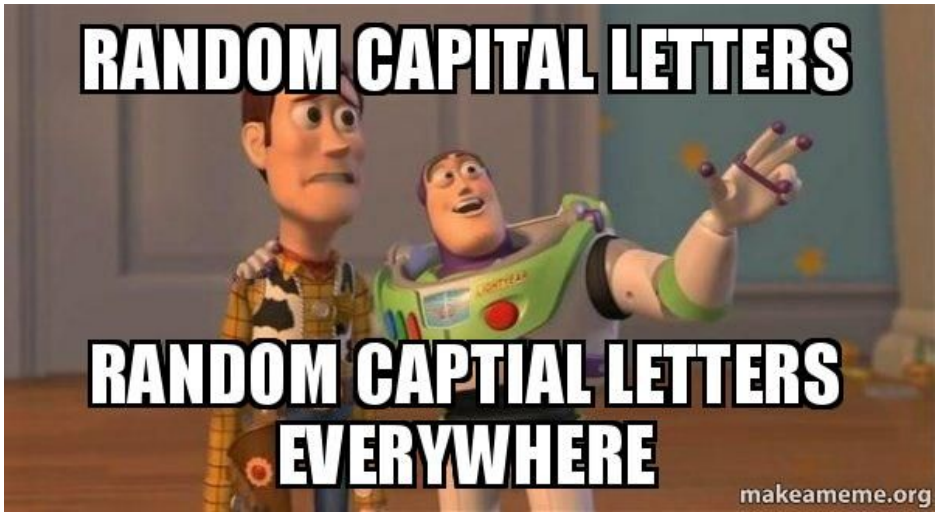
Phred Score

Rule of thumb: у хорошего рида большинство позиций имеет качество прописной буквой

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
                  |         |         |         |         |
 Quality score:    01........11........21........31........41
```

Header Sequence Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
```



FASTQC: что такое хорошо и что такое плохо

Программа для **анализа качества ридов**

Риды

FASTQC

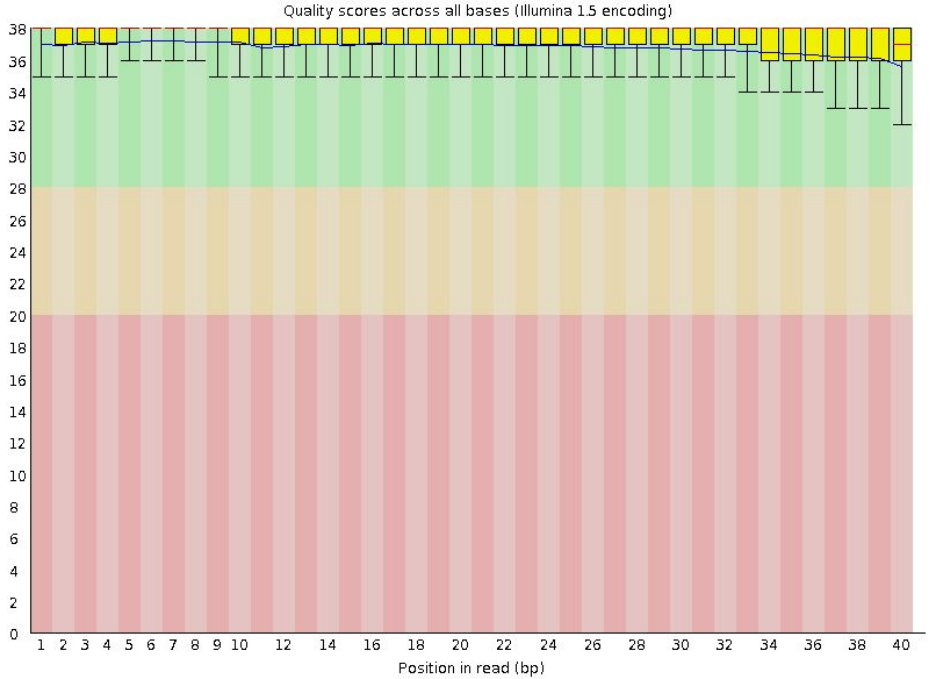
Отчёт

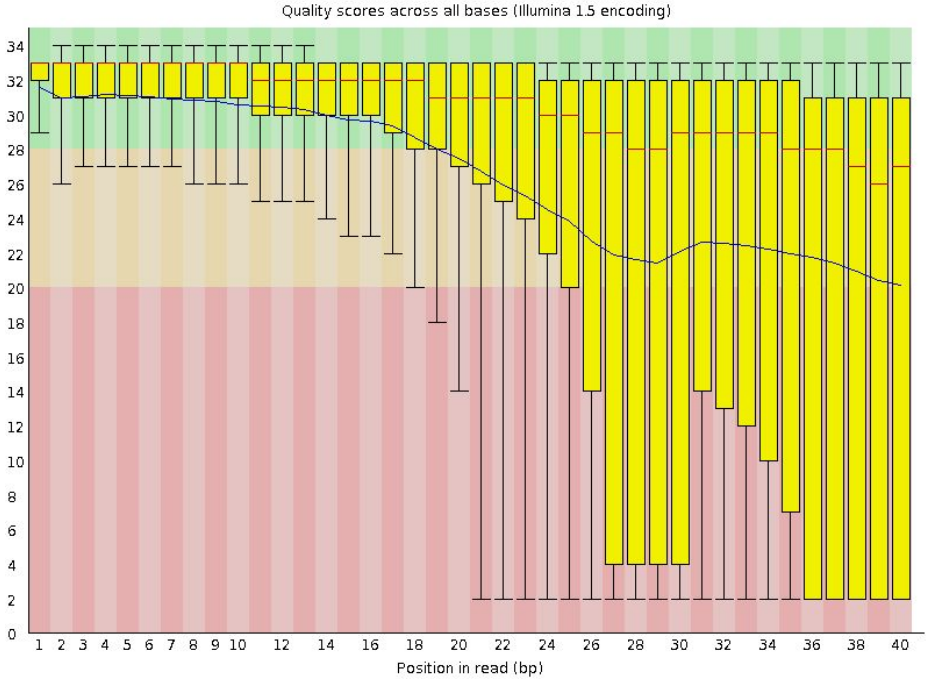
Примеры отчётов:
Хороший:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
Плохой:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FASTQC: Basic Statistics

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

FASTQC: Per base sequence quality





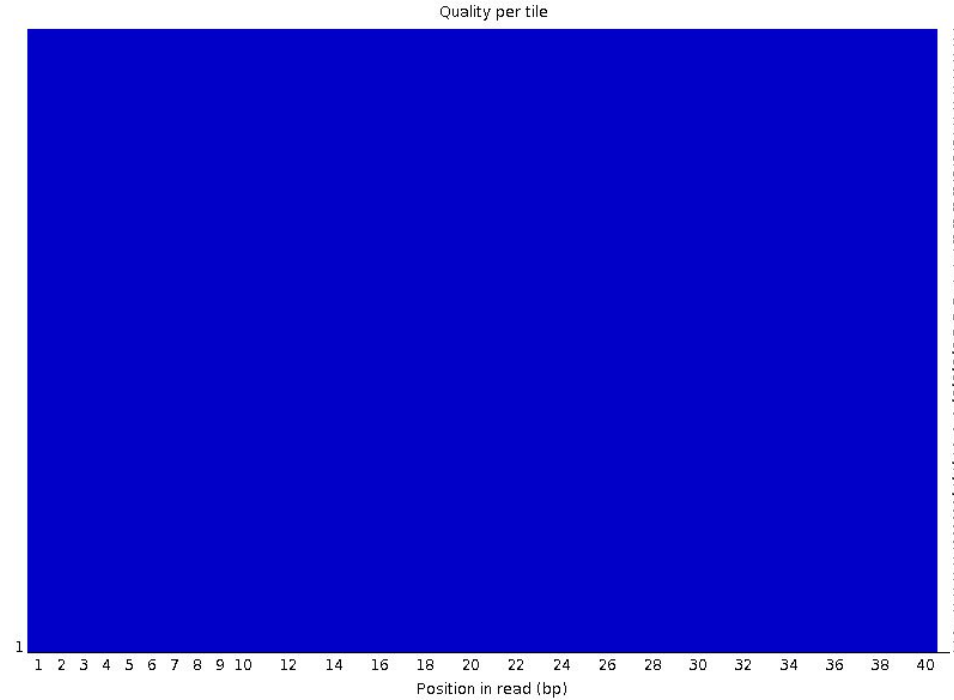
FASTQC: Per base sequence quality

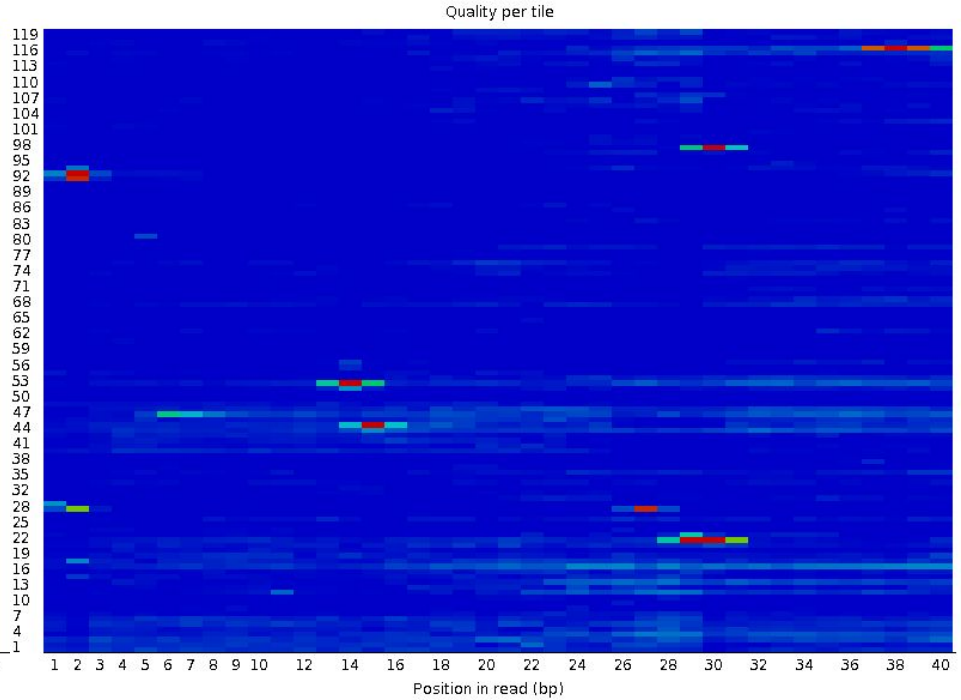
По оси **X** откладывается **позиция нуклеотида** в риде, по оси **Y**- его **качество**.

Анализируются суммарные данные по нуклеотидам в определённых позициях у всех ридов. В зелёную область попадают нуклеотиды с наилучшим качеством, в персиковую - с приемлемым, в красную - с плохим.

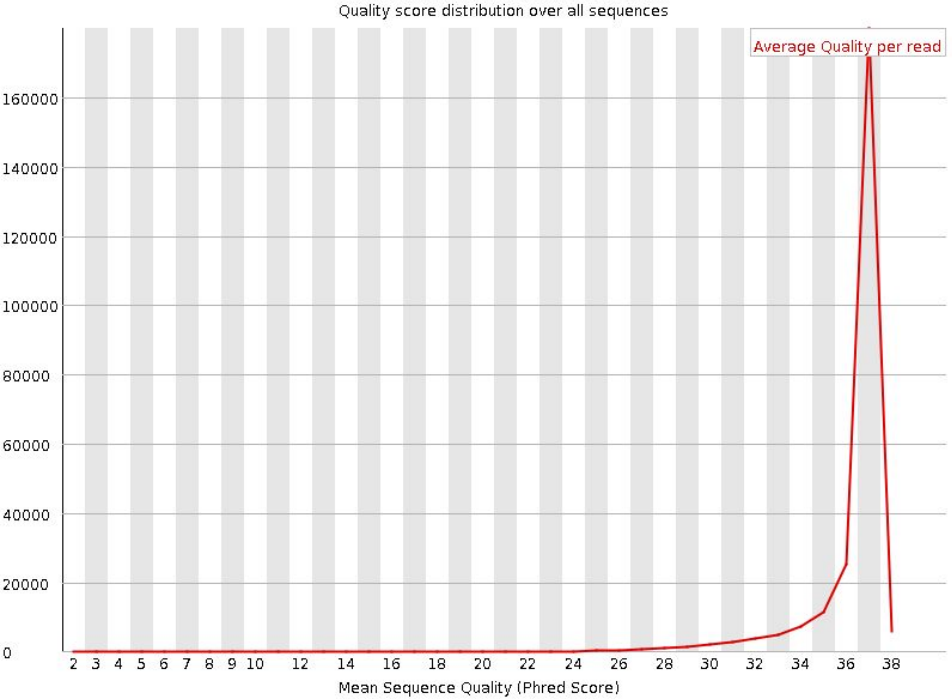
Синяя линия показывает среднее значение качества, красная линия внутри каждого бокса - медиану.

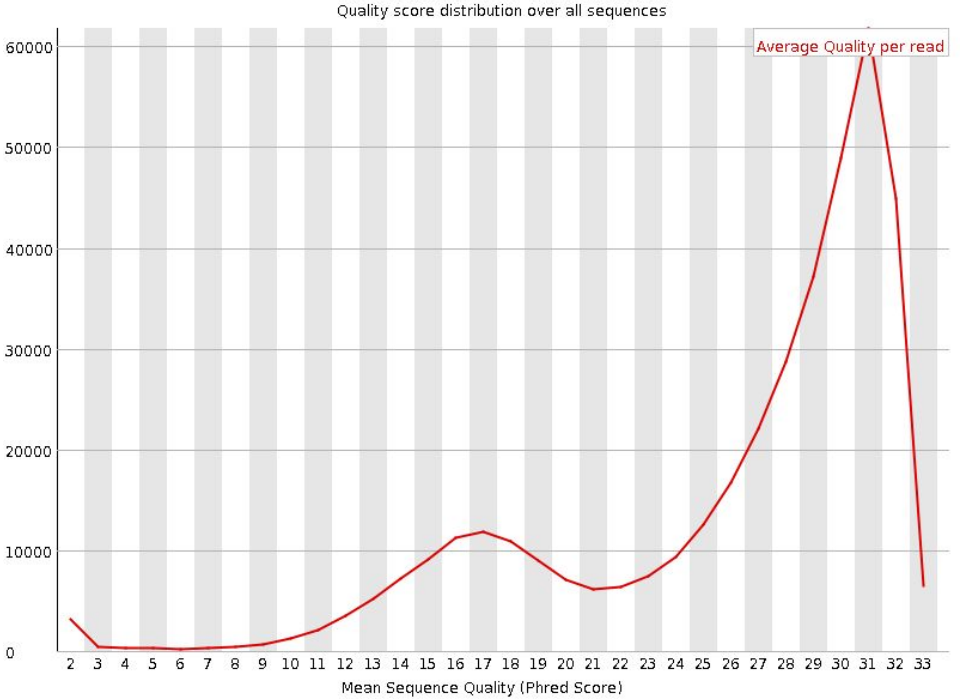
FASTQC: Per tile sequence quality





FASTQC: Per sequence quality scores



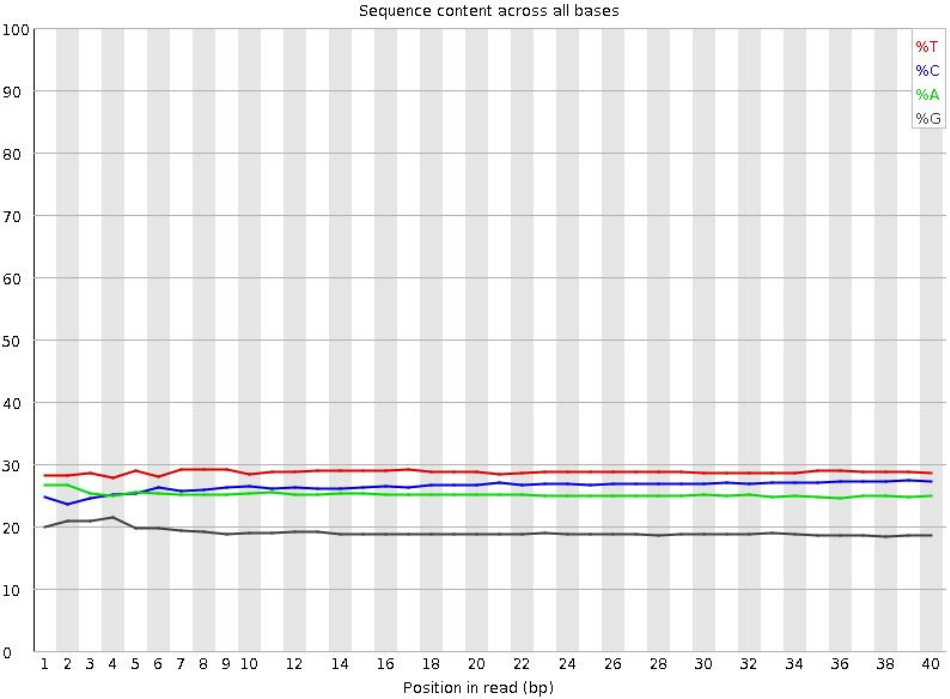


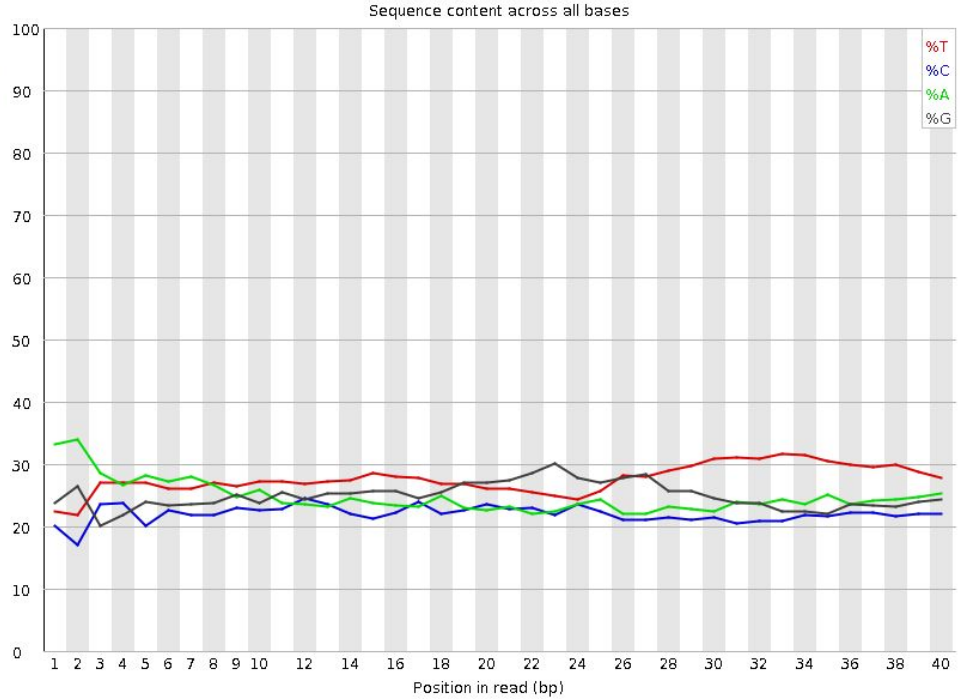
FASTQC: Per sequence quality scores

График показывает **среднее качество ридов**.

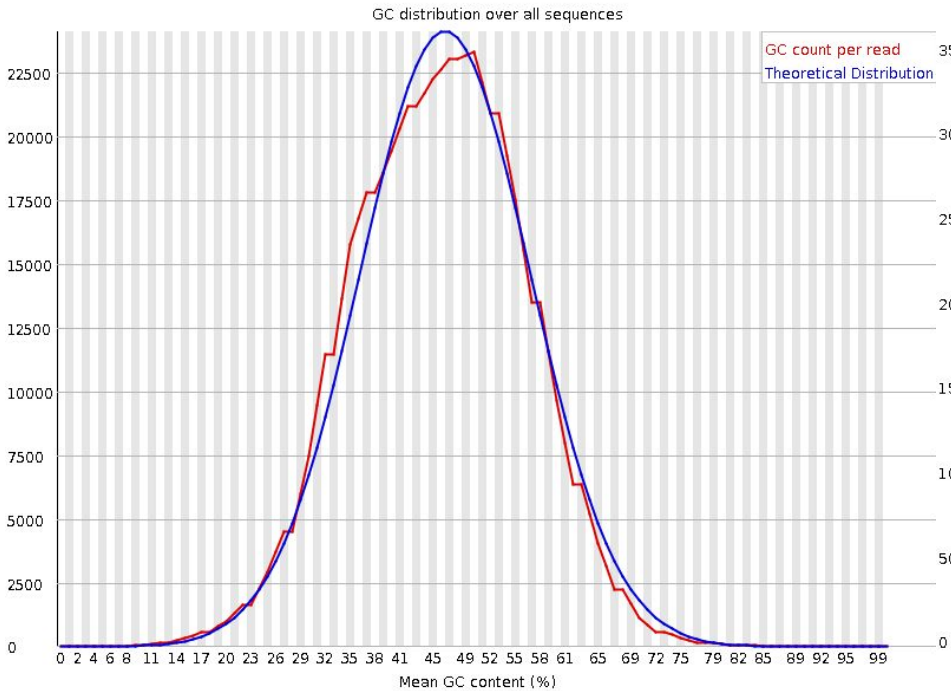
Если пик сдвинут в область с высокими показателями качества, то это значит, что риды хорошие и вполне достоверные.

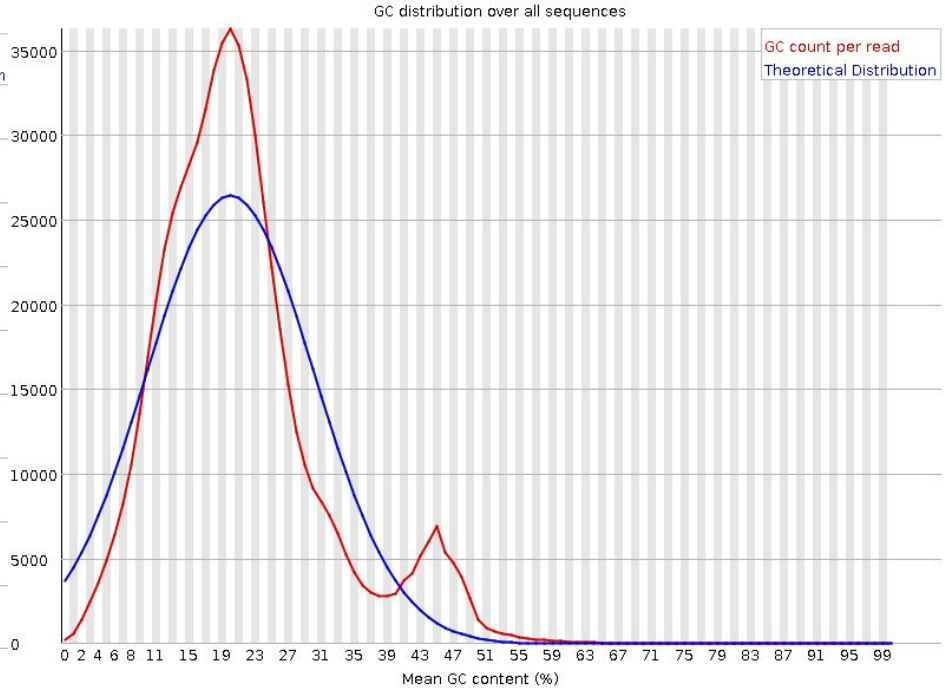
FASTQC: Per base sequence content





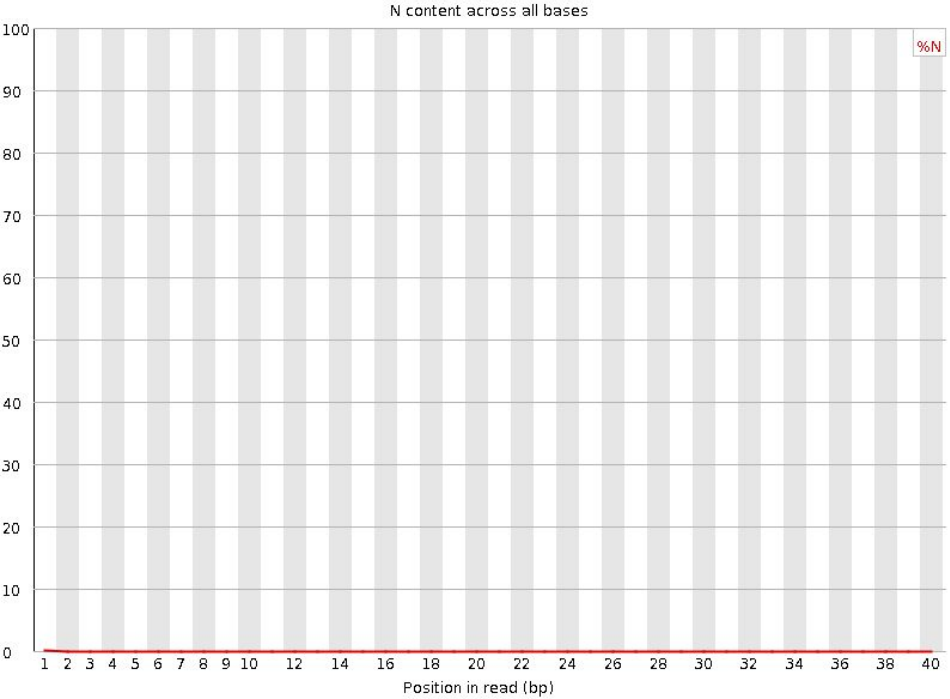
FASTQC: Per sequence GC content

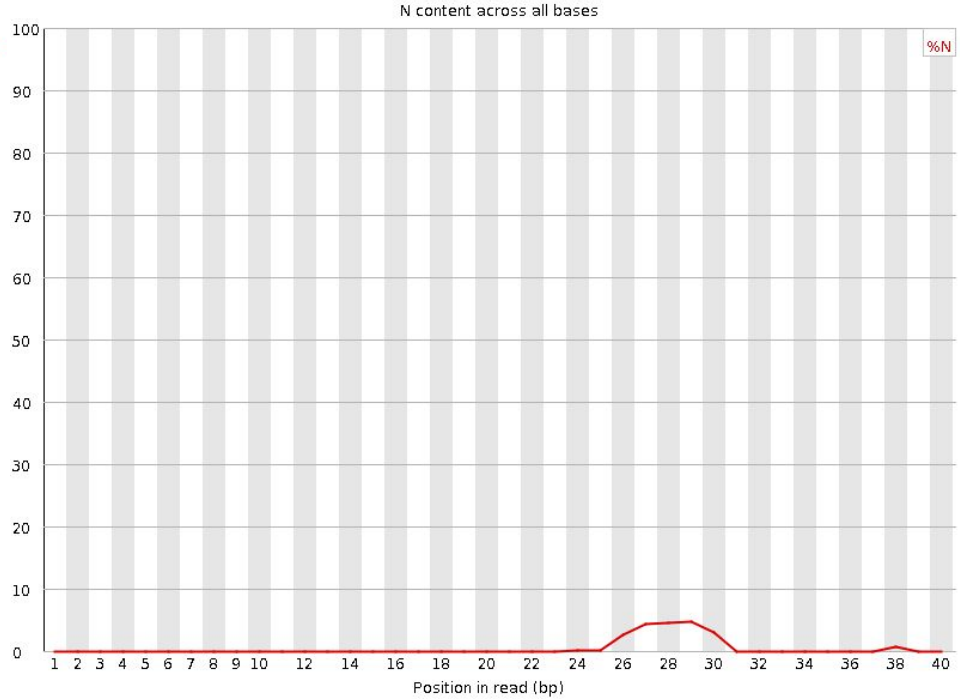




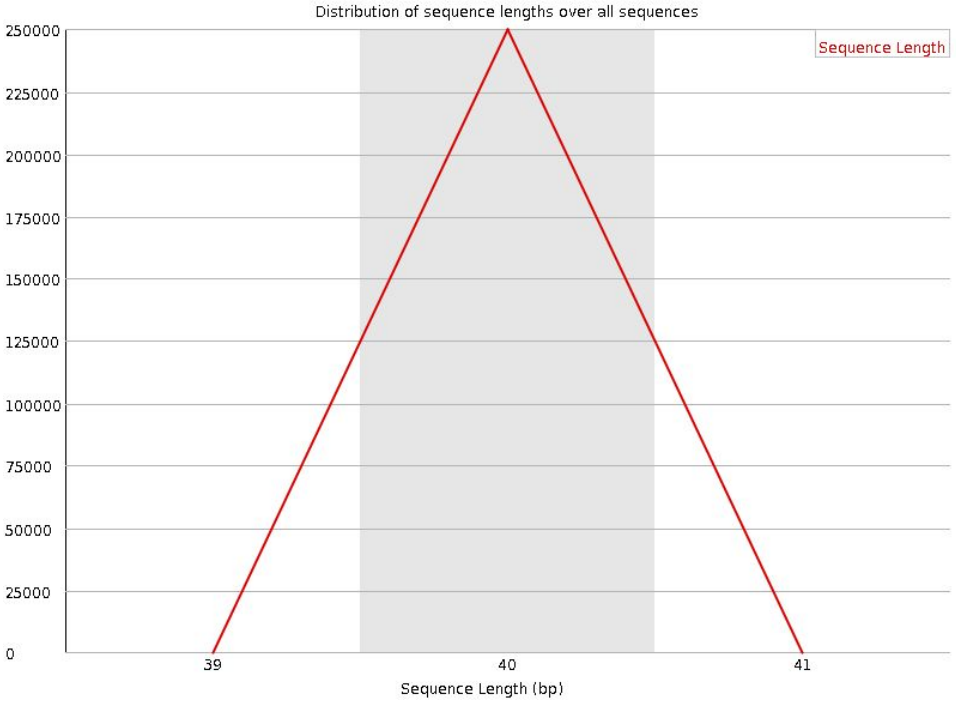
Дополнительные пики - признаки контаминации

FASTQC: Per base N content

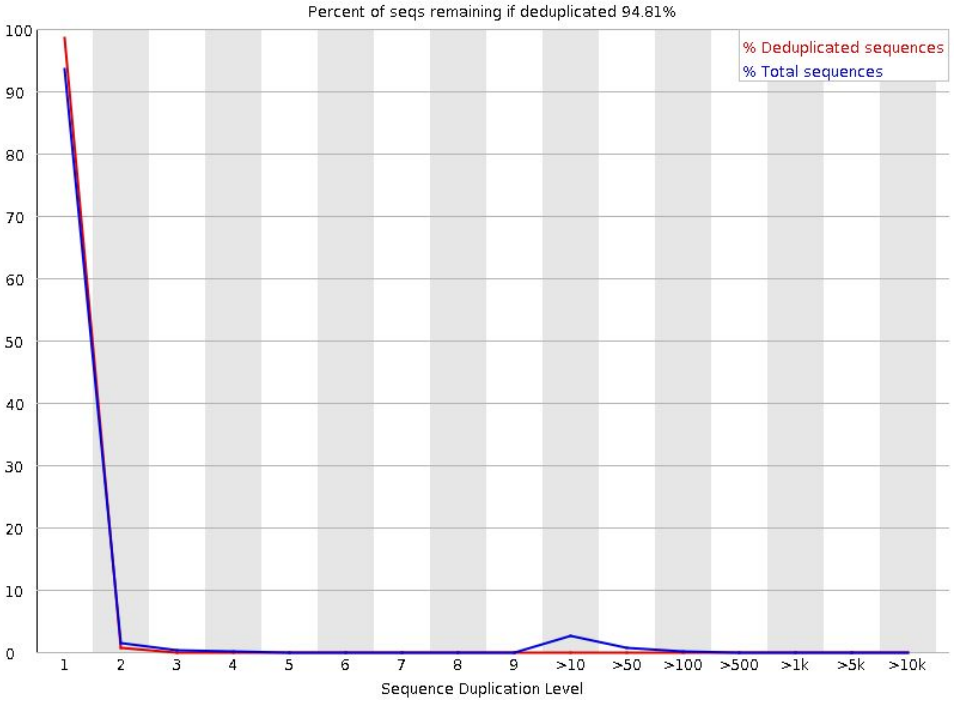


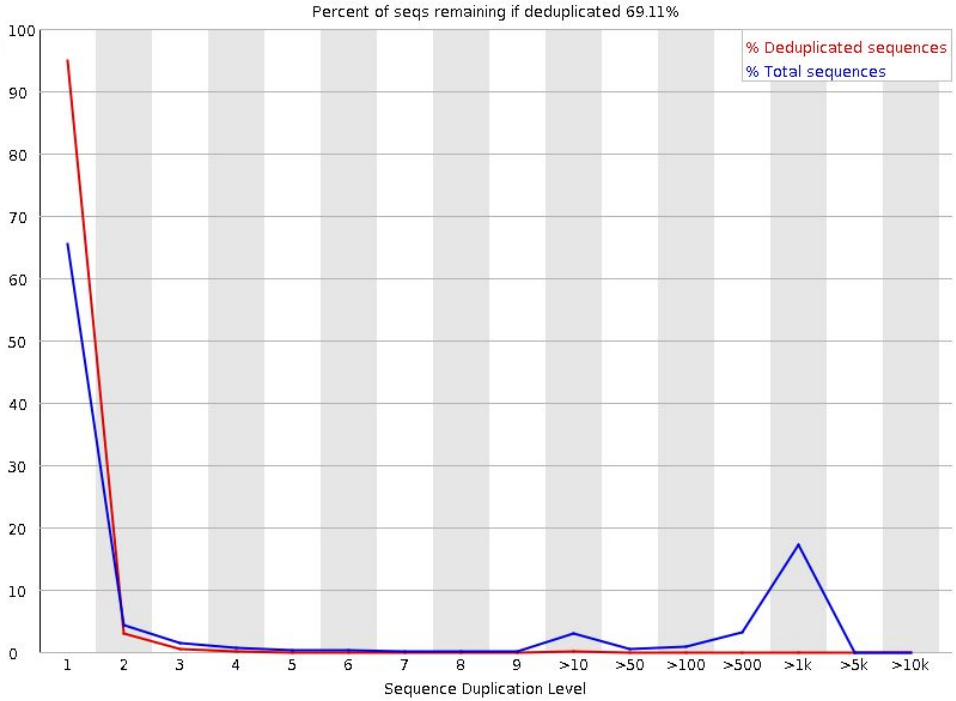


FASTQC: Sequence Length Distribution



FASTQC: Sequence Duplication Levels

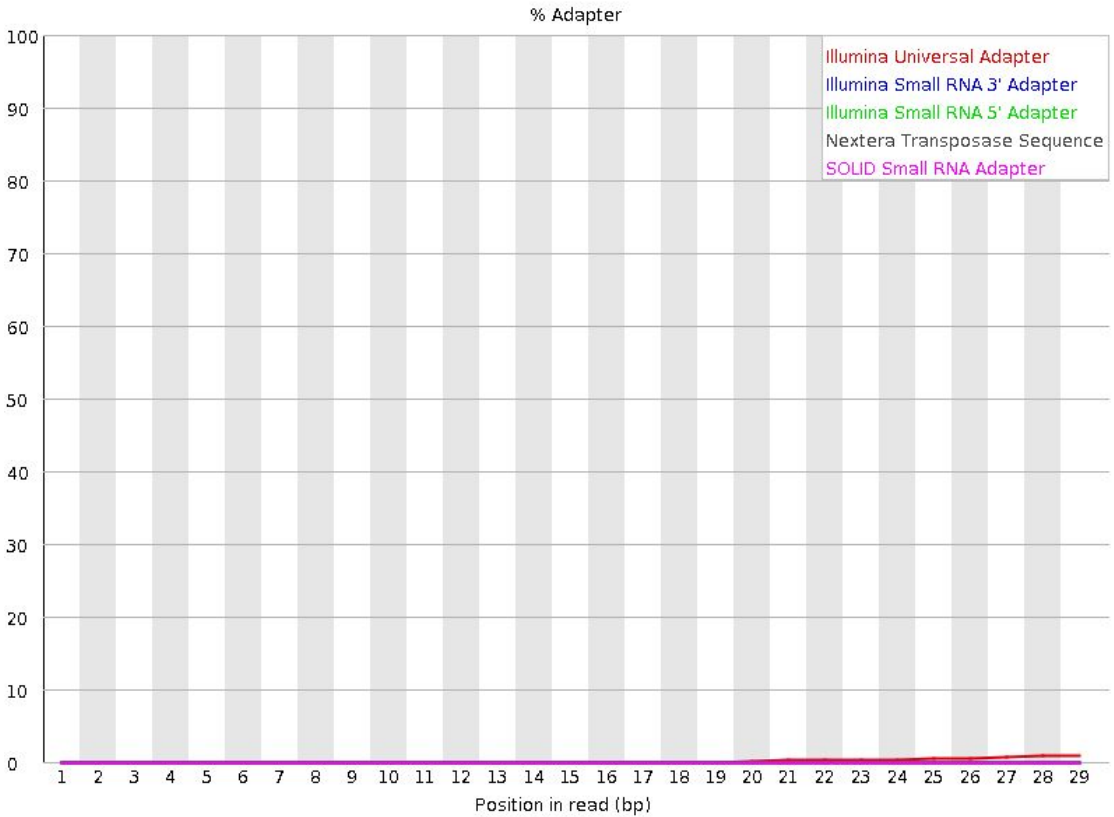




FASTQC: Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	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
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	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
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	No Hit

FASTQC: Adapter Content



Тримминг


Тримминг
RAW FastQ read
5' Adapter
Processed sRNA read
3' Adapter
До
После

Тримминг

Процедура **улучшения качества** прочтений:

Удаляет адаптеры;

Позволяет **убрать** риды, которые **короче определённой длины** (короткие риды не несут полезной информации);

Позволяет **фильтровать** риды **по качеству**.

Программы для Тримминга

Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)

FastX (http://hannonlab.cshl.edu/fastx_toolkit/)

Деконтаминация: bowtie2 + геномные датабазы
(http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)

Для секвенирования спаренных концов с NexTera toolkit– NxTrim
(https://github.com/sequencing/NxTrim)

Trimmomatic: основные опции

ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.

SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5‟ end and clips the read once the average quality within the window falls below a threshold.

LEADING: Cut bases off the start of a read, if below a threshold quality

TRAILING: Cut bases off the end of a read, if below a threshold quality

CROP: Cut the read to a specified length by removing bases from the end

HEADCROP: Cut the specified number of bases from the start of the read

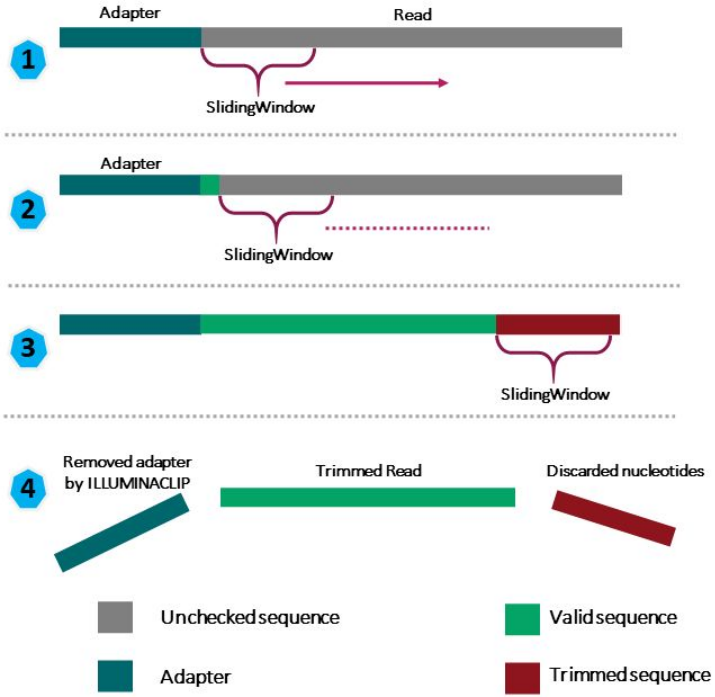
MINLEN: Drop the read if it is below a specified length

AVGQUAL: Drop the read if the average quality is below the specified level

Никто:
Мануалы по биоинформатике:

Trimmomatic

Последовательное выполнение команд. Аккуратнее с **CROP, HEADCROP, MINLEN**!



Практическая часть