Отчёт по Лабораторной № 2

Выполнил студент 2 курса СПбАУ Есиков Сергей

Background and Metadata

1. Надо посчитать количество уникальных значений в стобце generation

   Ответ: 25

2. Просто посмотреть измерения таблицы

   Ответ: 62x12

3. Ставим фильтр plus на столбец Cit и считаем столбцы

   Ответ: 10

4. Ставим фильтр plus на столбец Mutator и аналогично считаем стобцы

   Ответ: 6

Assessing Read Quality

1. Загружаем архив, распаковываем, читаем хвост

```
& curl -O
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_1.fast
q.gz
& gunzip SRR2584863_1.fastq.gz
& tail -n 4 SRR2584863_1.fastq
```

Output

```
@SRR2584863.1553259 HWI-ST957:245:H73R4ADXX:2:2216:21048:100894/1
CTGCAATACCACGCTGATCTTTCACATGATGTAAGAAAAGTGGGATCAGCAAACCGGGTGCTGCTGTGGCTAGT
TGCAGCAAACCATGCAGTGAACCCGCCTGTGCTTCGCTATAGCCGTGACTGATGAGGATCGCCGGAAGCCAGCC
AA
+
CCCFFFFFHHHHGJJJJJJJJJHGIJJJIJJJJIJJJJIIIIJJJJJJJJJJJJJIIJJJHHHHHFFFFFEEEE
EDDDDDDDDDDDDDDDDDCDEDDBDBDDBDDDDDDDDDBDEEDDDD7@BDDDDDD>AA>?B?<@BDD@BDC?
BDA?
```

2. Просто нужные опции -l(long) и -h(humane)

```
& ls -lh ./
```

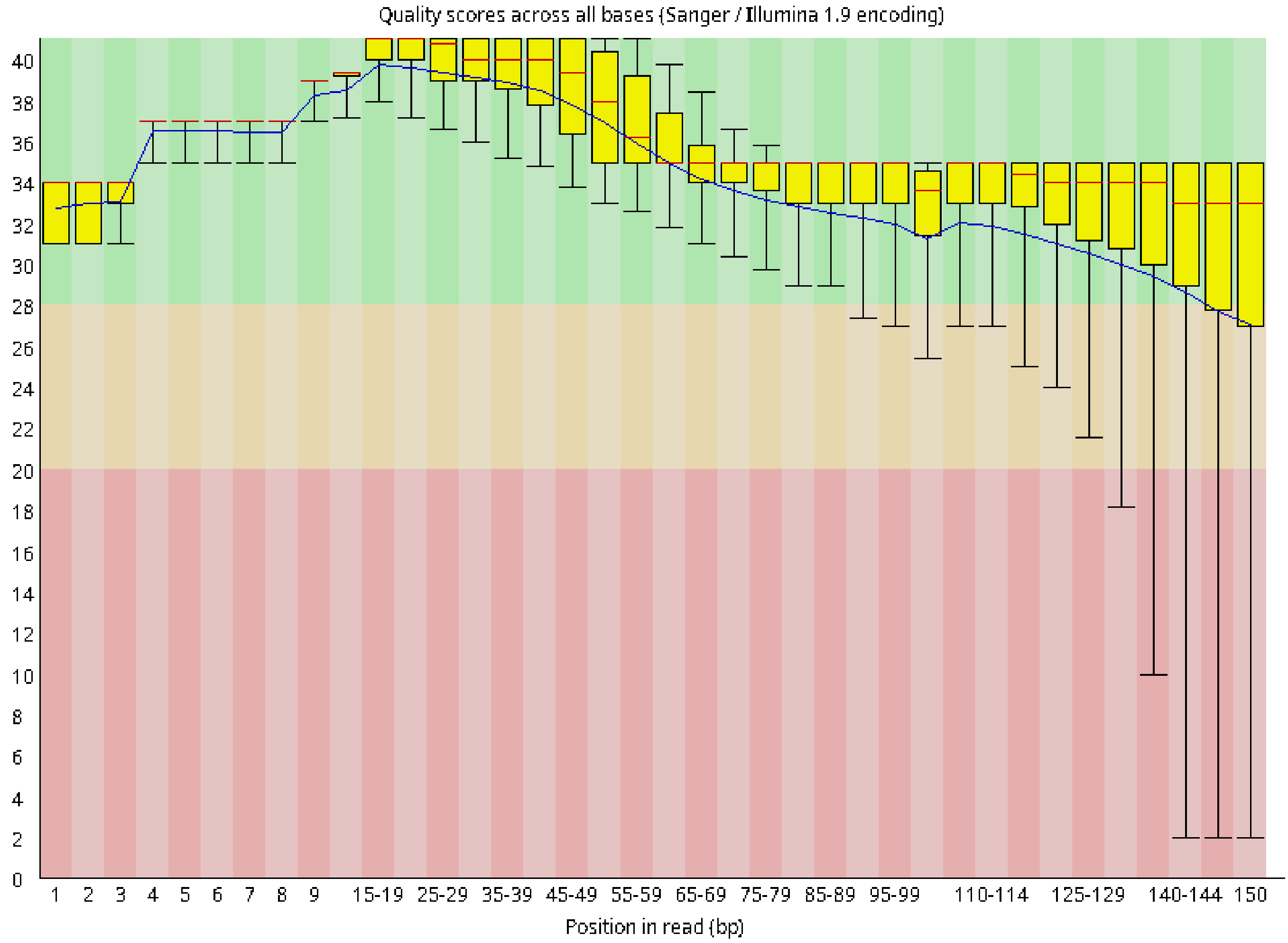
Output

```
total 1.6G
-rw-r--r-- 1 serotonin serotonin 545M Jun 10 04:37 SRR2584863_1.fastq
-rw-r--r-- 1 serotonin serotonin 183M Jun 10 04:38 SRR2584863_2.fastq.gz
-rw-r--r-- 1 serotonin serotonin 309M Jun 10 04:40 SRR2584866_1.fastq.gz
-rw-r--r-- 1 serotonin serotonin 296M Jun 10 04:41 SRR2584866_2.fastq.gz
-rw-r--r-- 1 serotonin serotonin 124M Jun 10 04:35 SRR2589044_1.fastq.gz
-rw-r--r-- 1 serotonin serotonin 128M Jun 10 04:36 SRR2589044_2.fastq.gz
```

Вот результат оценки качества

```
& fastqc *.fastq*
& ,/SRR2584863_1_fastqc.html
```

Заметно, что качество прочтения сильно портится ближе к концу



Резюме тестов:

Пройденные

```
& cat fastqc_summaries.txt | grep 'PASS'
PASS    Basic Statistics    SRR2584863_1.fastq
PASS    Per base sequence quality   SRR2584863_1.fastq
PASS    Per tile sequence quality   SRR2584863_1.fastq
PASS    Per sequence quality scores SRR2584863_1.fastq
PASS    Per base N content  SRR2584863_1.fastq
PASS    Sequence Length Distribution    SRR2584863_1.fastq
PASS    Sequence Duplication Levels SRR2584863_1.fastq
PASS    Overrepresented sequences   SRR2584863_1.fastq
PASS    Basic Statistics    SRR2584863_2.fastq.gz
PASS    Per sequence quality scores SRR2584863_2.fastq.gz
PASS    Per base N content  SRR2584863_2.fastq.gz
PASS    Sequence Length Distribution    SRR2584863_2.fastq.gz
PASS    Sequence Duplication Levels SRR2584863_2.fastq.gz
PASS    Overrepresented sequences   SRR2584863_2.fastq.gz
PASS    Basic Statistics    SRR2584866_1.fastq.gz
PASS    Per sequence quality scores SRR2584866_1.fastq.gz
PASS    Per base N content  SRR2584866_1.fastq.gz
PASS    Sequence Length Distribution    SRR2584866_1.fastq.gz
PASS    Overrepresented sequences   SRR2584866_1.fastq.gz
PASS    Basic Statistics    SRR2584866_2.fastq.gz
PASS    Per base sequence quality   SRR2584866_2.fastq.gz
PASS    Per tile sequence quality   SRR2584866_2.fastq.gz
PASS    Per sequence quality scores SRR2584866_2.fastq.gz
PASS    Per base N content  SRR2584866_2.fastq.gz
PASS    Sequence Length Distribution    SRR2584866_2.fastq.gz
PASS    Overrepresented sequences   SRR2584866_2.fastq.gz
PASS    Basic Statistics    SRR2589044_1.fastq.gz
PASS    Per base sequence quality   SRR2589044_1.fastq.gz
PASS    Per tile sequence quality   SRR2589044_1.fastq.gz
PASS    Per sequence quality scores SRR2589044_1.fastq.gz
PASS    Per base N content  SRR2589044_1.fastq.gz
PASS    Sequence Length Distribution    SRR2589044_1.fastq.gz
PASS    Sequence Duplication Levels SRR2589044_1.fastq.gz
PASS    Overrepresented sequences   SRR2589044_1.fastq.gz
PASS    Basic Statistics    SRR2589044_2.fastq.gz
PASS    Per sequence quality scores SRR2589044_2.fastq.gz
PASS    Per base N content  SRR2589044_2.fastq.gz
PASS    Sequence Length Distribution    SRR2589044_2.fastq.gz
PASS    Sequence Duplication Levels SRR2589044_2.fastq.gz
PASS    Overrepresented sequences   SRR2589044_2.fastq.gz
```

Проваленные

```
& cat fastqc_summaries.txt | grep 'FAIL'
FAIL    Per base sequence quality   SRR2584863_2.fastq.gz
FAIL    Per tile sequence quality   SRR2584863_2.fastq.gz
FAIL    Per base sequence content   SRR2584863_2.fastq.gz
FAIL    Per base sequence quality   SRR2584866_1.fastq.gz
FAIL    Per base sequence content   SRR2584866_1.fastq.gz
FAIL    Adapter Content SRR2584866_1.fastq.gz
FAIL    Adapter Content SRR2584866_2.fastq.gz
FAIL    Adapter Content SRR2589044_1.fastq.gz
FAIL    Per base sequence quality   SRR2589044_2.fastq.gz
FAIL    Per tile sequence quality   SRR2589044_2.fastq.gz
FAIL    Per base sequence content   SRR2589044_2.fastq.gz
FAIL    Adapter Content SRR2589044_2.fastq.gz
```

Вызывающие опасения

```
& cat fastqc_summaries.txt | grep 'WARN'
WARN    Per base sequence content   SRR2584863_1.fastq
WARN    Per sequence GC content SRR2584863_1.fastq
WARN    Adapter Content SRR2584863_1.fastq
WARN    Per sequence GC content SRR2584863_2.fastq.gz
WARN    Adapter Content SRR2584863_2.fastq.gz
WARN    Per tile sequence quality   SRR2584866_1.fastq.gz
WARN    Per sequence GC content SRR2584866_1.fastq.gz
WARN    Sequence Duplication Levels SRR2584866_1.fastq.gz
WARN    Per base sequence content   SRR2584866_2.fastq.gz
WARN    Per sequence GC content SRR2584866_2.fastq.gz
WARN    Sequence Duplication Levels SRR2584866_2.fastq.gz
WARN    Per base sequence content   SRR2589044_1.fastq.gz
WARN    Per sequence GC content SRR2589044_1.fastq.gz
WARN    Per sequence GC content SRR2589044_2.fastq.gz
```

Trimmomatic options

`& yay -S trimmomatic`

`& sudo find -print / -name "NexteraPE-PE.fa"`

`& cp /opt/Trimmomatic/adapters/NexteraPE-PE.fa ./`

```
& trimmomatic PE SRR2589044_1.fastq.gz SRR2589044_2.fastq.gz \
                SRR2589044_1.trim.fastq.gz SRR2589044_1un.trim.fastq.gz \
                SRR2589044_2.trim.fastq.gz SRR2589044_2un.trim.fastq.gz \
                SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-
PE.fa:2:40:15
TrimmomaticPE: Started with arguments:
 SRR2589044_1.fastq.gz SRR2589044_2.fastq.gz SRR2589044_1.trim.fastq.gz
SRR2589044_1un.trim.fastq.gz SRR2589044_2.trim.fastq.gz
SRR2589044_2un.trim.fastq.gz SLIDINGWINDOW:4:20 MINLEN:25
ILLUMINACLIP:NexteraPE-PE.fa:2:40:15
Multiple cores found: Using 4 threads
Using PrefixPair: 'AGATGTGTATAAGAGACAG' and 'AGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'CTGTCTCTTATACACATCTCCGAGCCCACGAGAC'
Using Long Clipping Sequence: 'CTGTCTCTTATACACATCTGACGCTGCCGACGA'
ILLUMINACLIP: Using 1 prefix pairs, 4 forward/reverse sequences, 0 forward
only sequences, 0 reverse only sequences
Quality encoding detected as phred33
Input Read Pairs: 1107090 Both Surviving: 885220 (79.96%) Forward Only
Surviving: 216472 (19.55%) Reverse Only Surviving: 2850 (0.26%) Dropped:
2548 (0.23%)
TrimmomaticPE: Completed successfully
```

1. 1 - 94/124 = 0.24 Ответ: 24%

```
& ls -sh | grep "RR2589044_1"
124M SRR2589044_1.fastq.gz
 94M SRR2589044_1.trim.fastq.gz
4.0K SRR2589044_1_fastqc
620K SRR2589044_1_fastqc.html
424K SRR2589044_1_fastqc.zip
 18M SRR2589044_1un.trim.fastq.gz
```

2. (94 + 91)/(128 + 124) = 0.73 Ответ: 74%

```
& ls -sh | grep "RR2589044_2"
128M SRR2589044_2.fastq.gz
 91M SRR2589044_2.trim.fastq.gz
4.0K SRR2589044_2_fastqc
624K SRR2589044_2_fastqc.html
440K SRR2589044_2_fastqc.zip
272K SRR2589044_2un.trim.fastq.gz
```

Грубый способ но быстрый и примерный

3. Просто просмотрим дирректорию с адаптерами

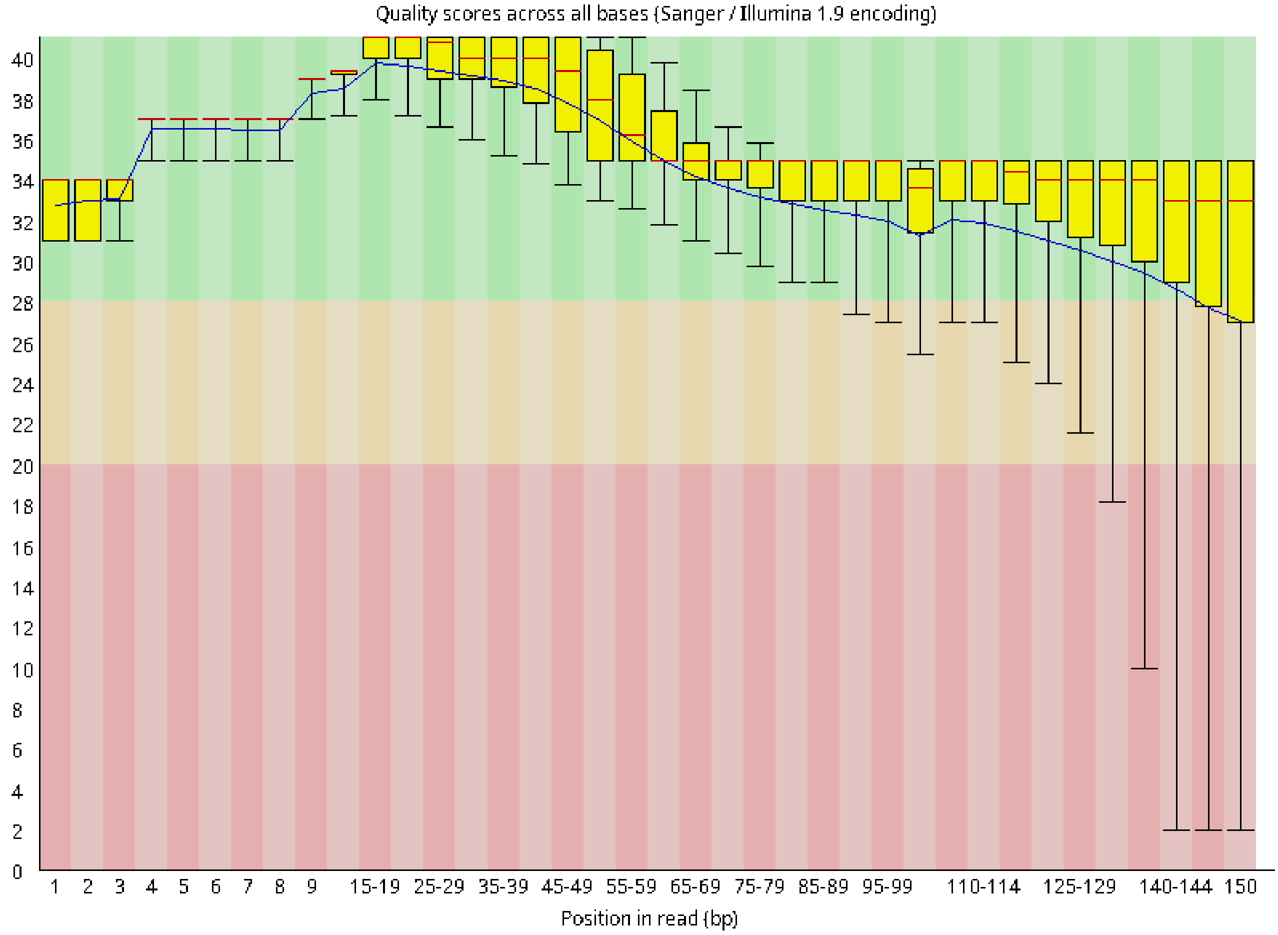
```
& ls /opt/Trimmomatic/adapters
NexteraPE-PE.fa  TruSeq2-SE.fa    TruSeq3-PE.fa
TruSeq2-PE.fa    TruSeq3-PE-2.fa  TruSeq3-SE.fa
```

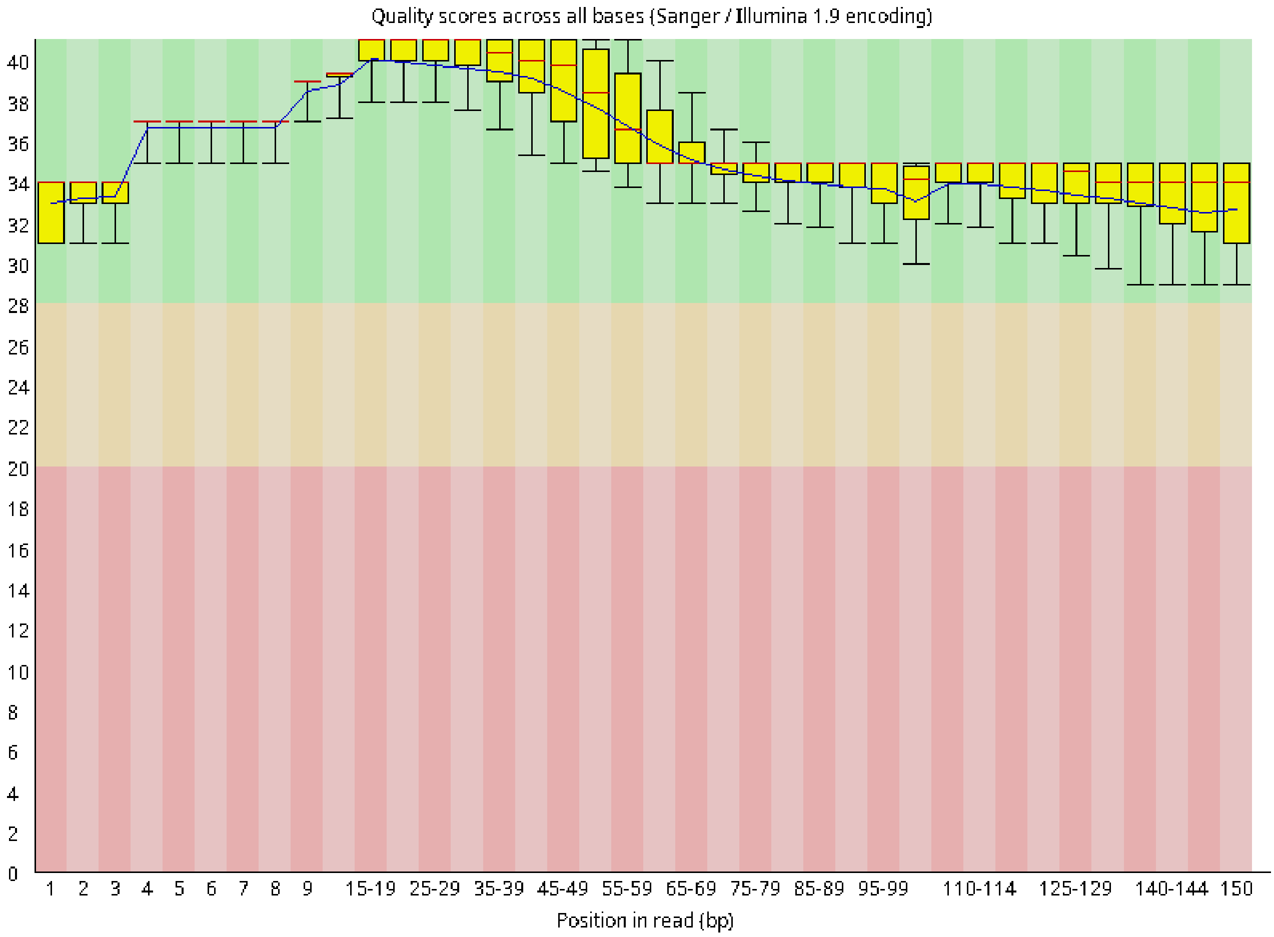
4.

```
& fastqc *.fastq*
& mkdir ../quality_check
& cp *.html ../quality_check
& cd ../quality_check/
& mkdir ../check
```

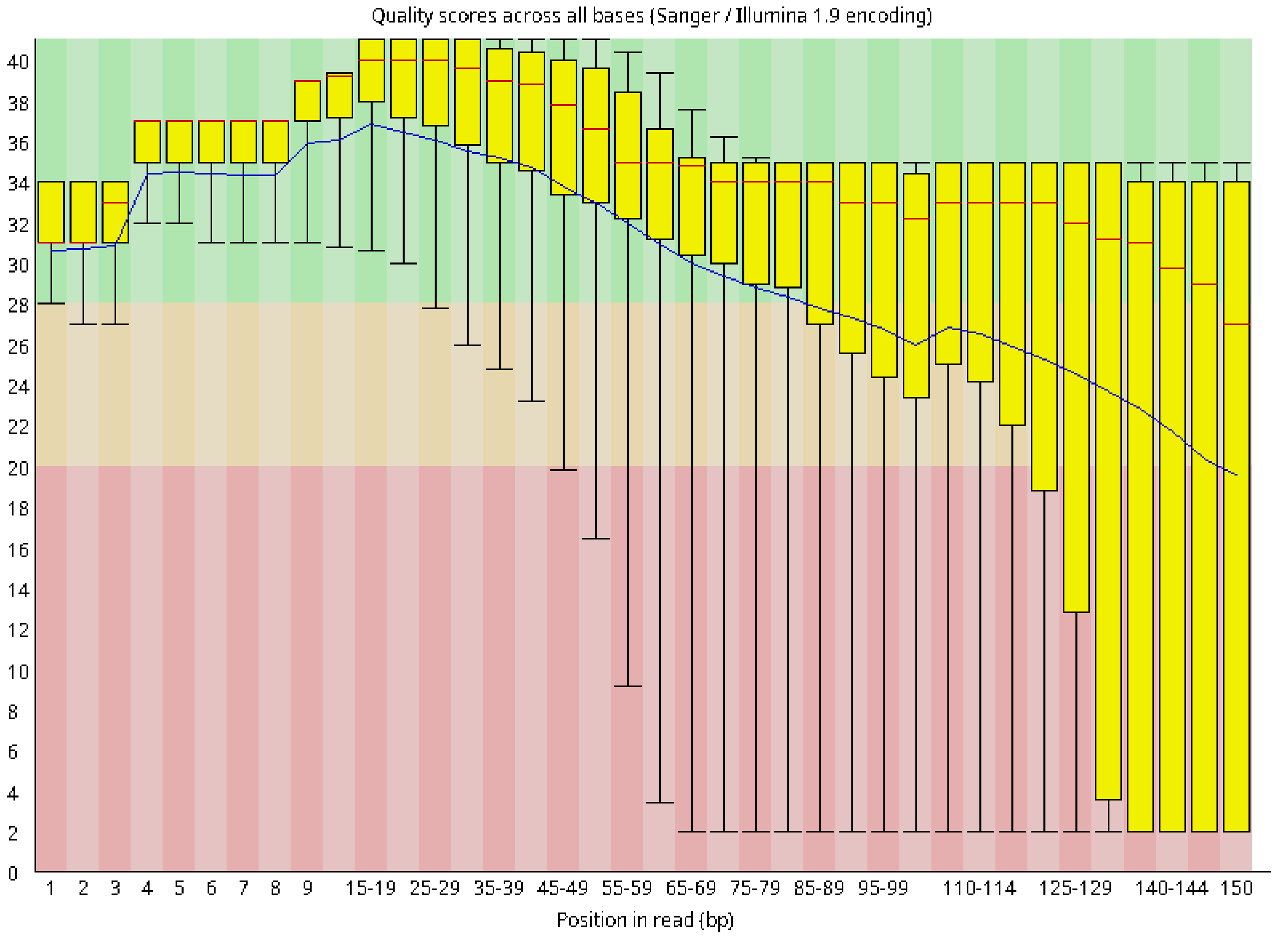
Ниже приведены сравнения показателей `Per base sequence quality` для каждой последовательности

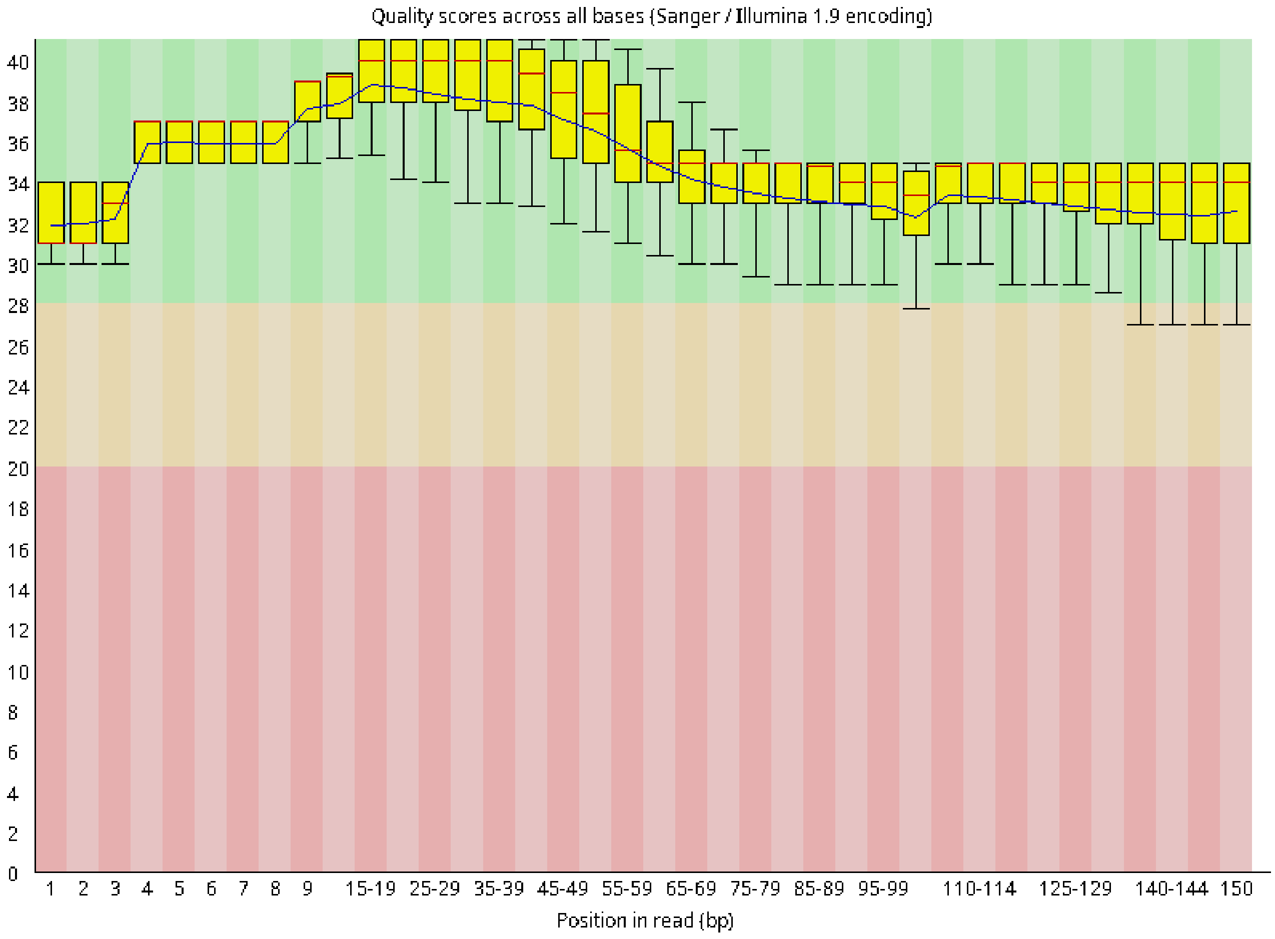
- SRR2584863_1



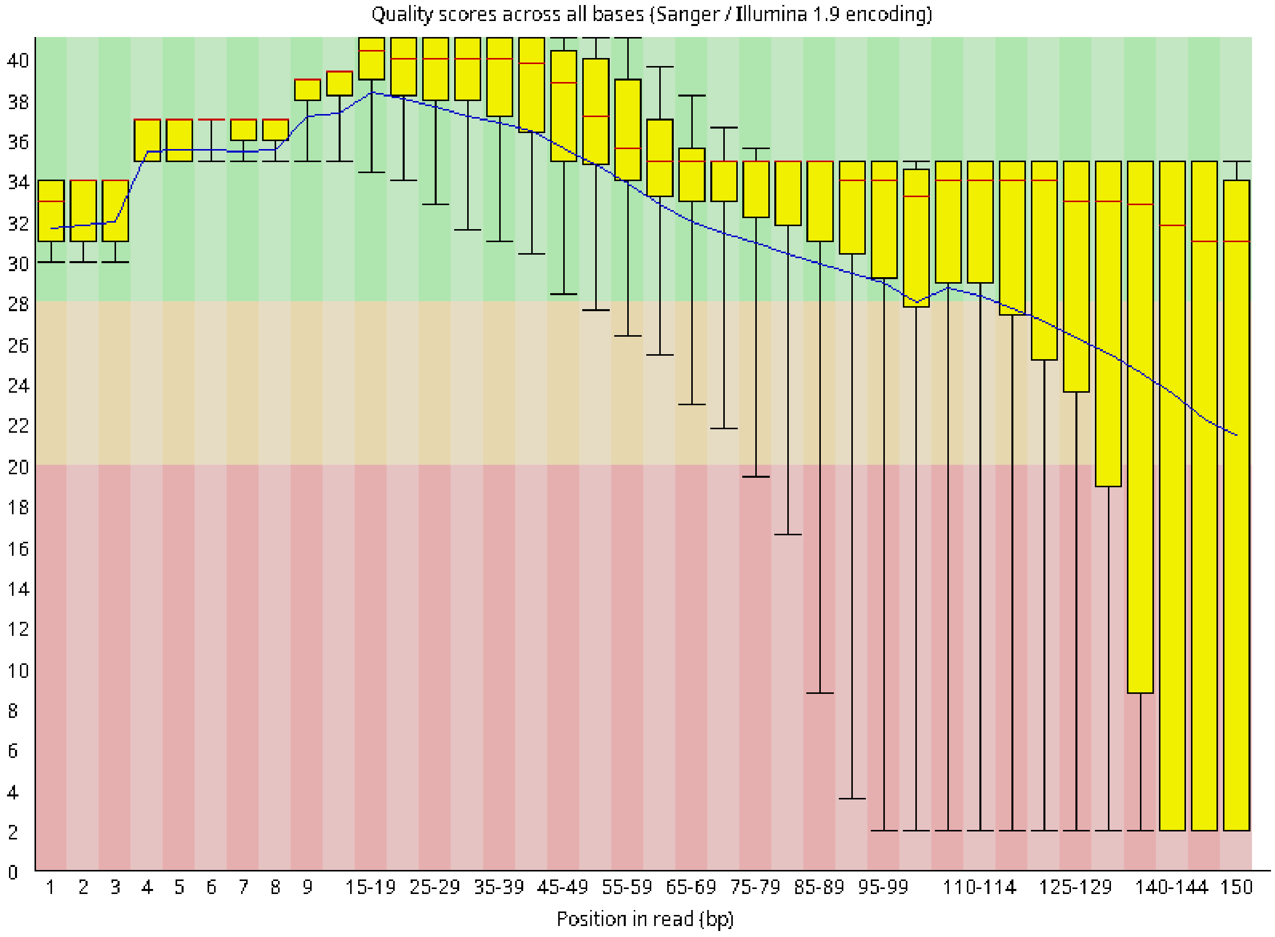


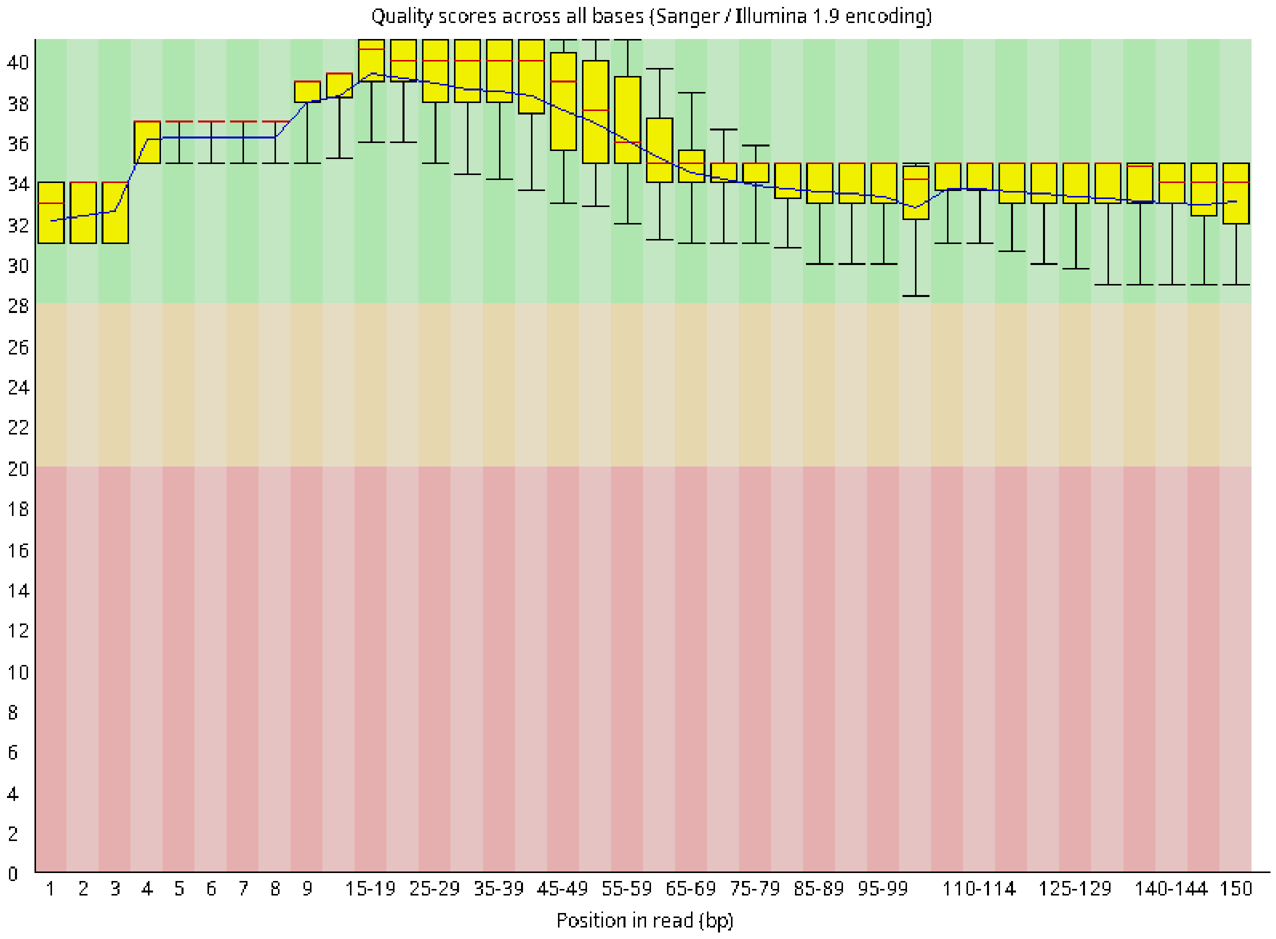
- SRR2584863_2



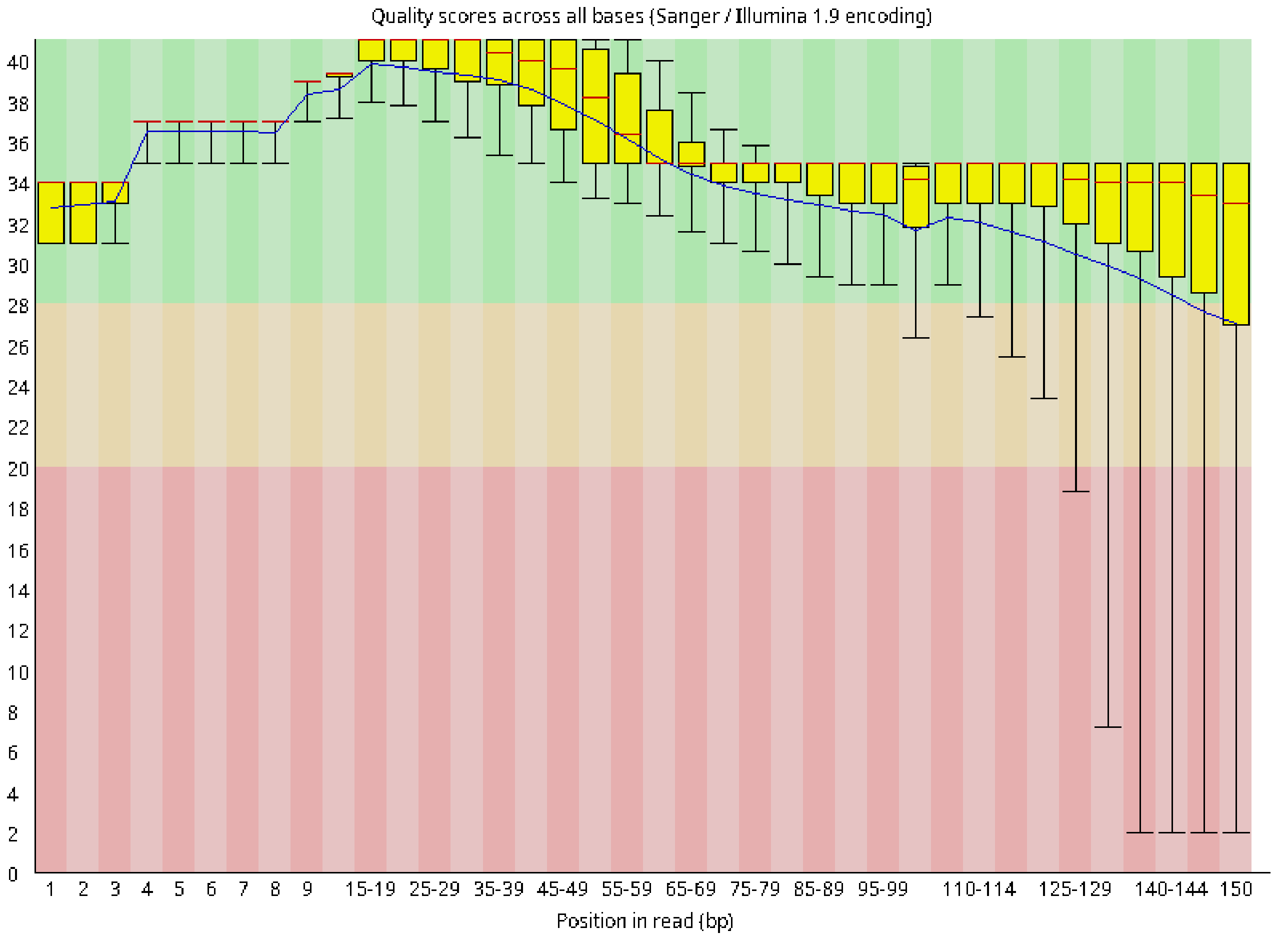


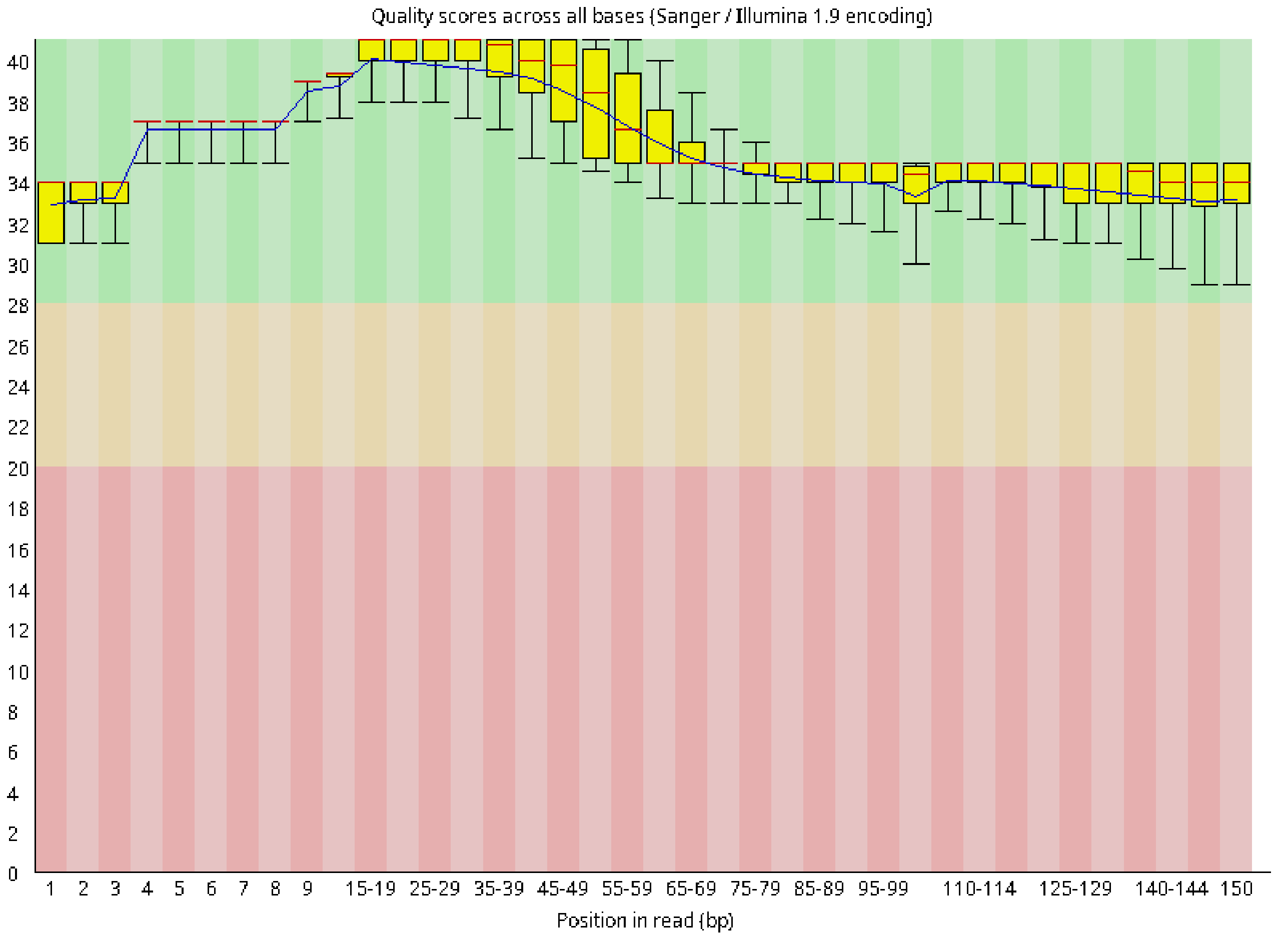
- SRR2584866_1



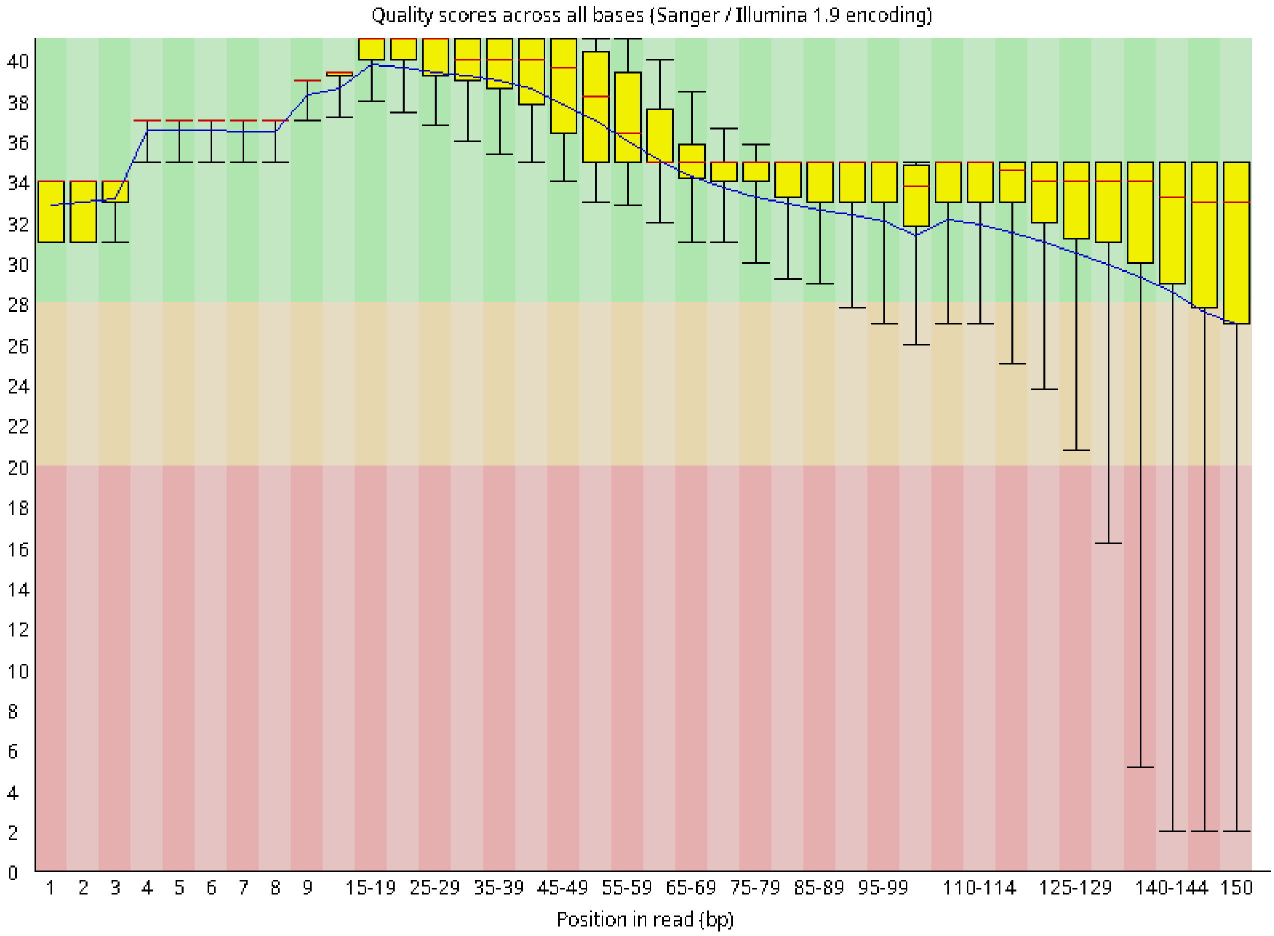


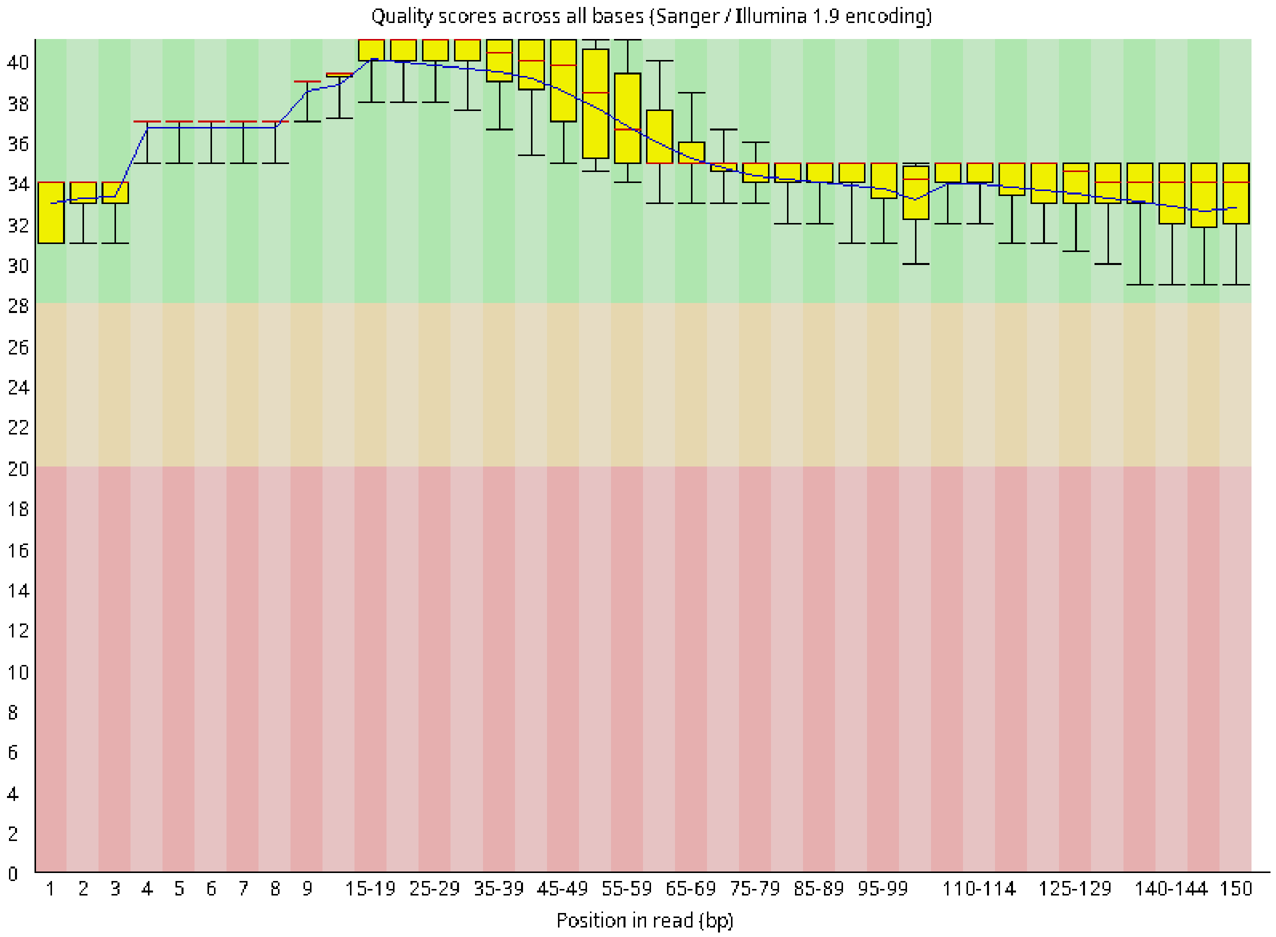
- SRR2584866_2



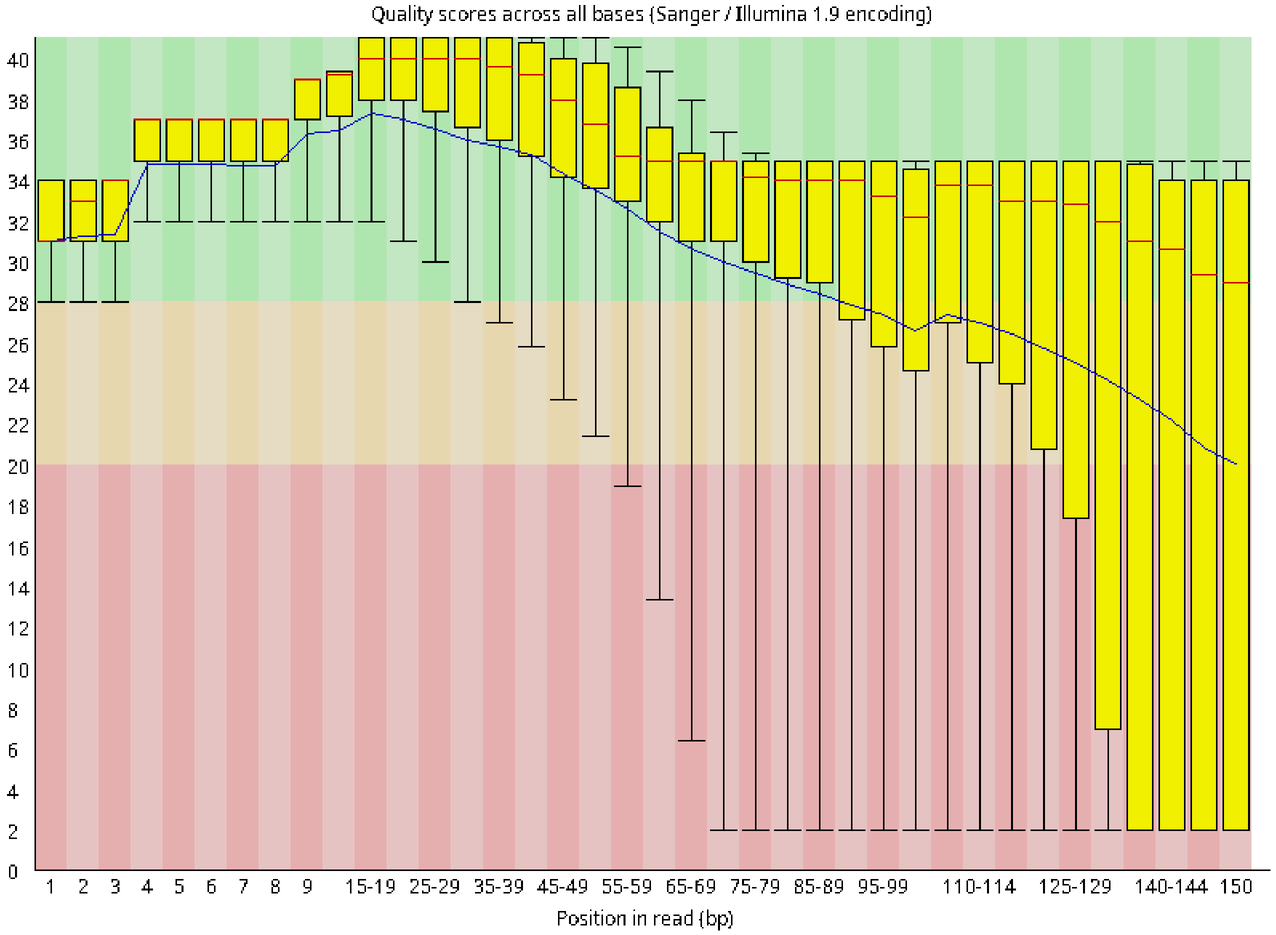


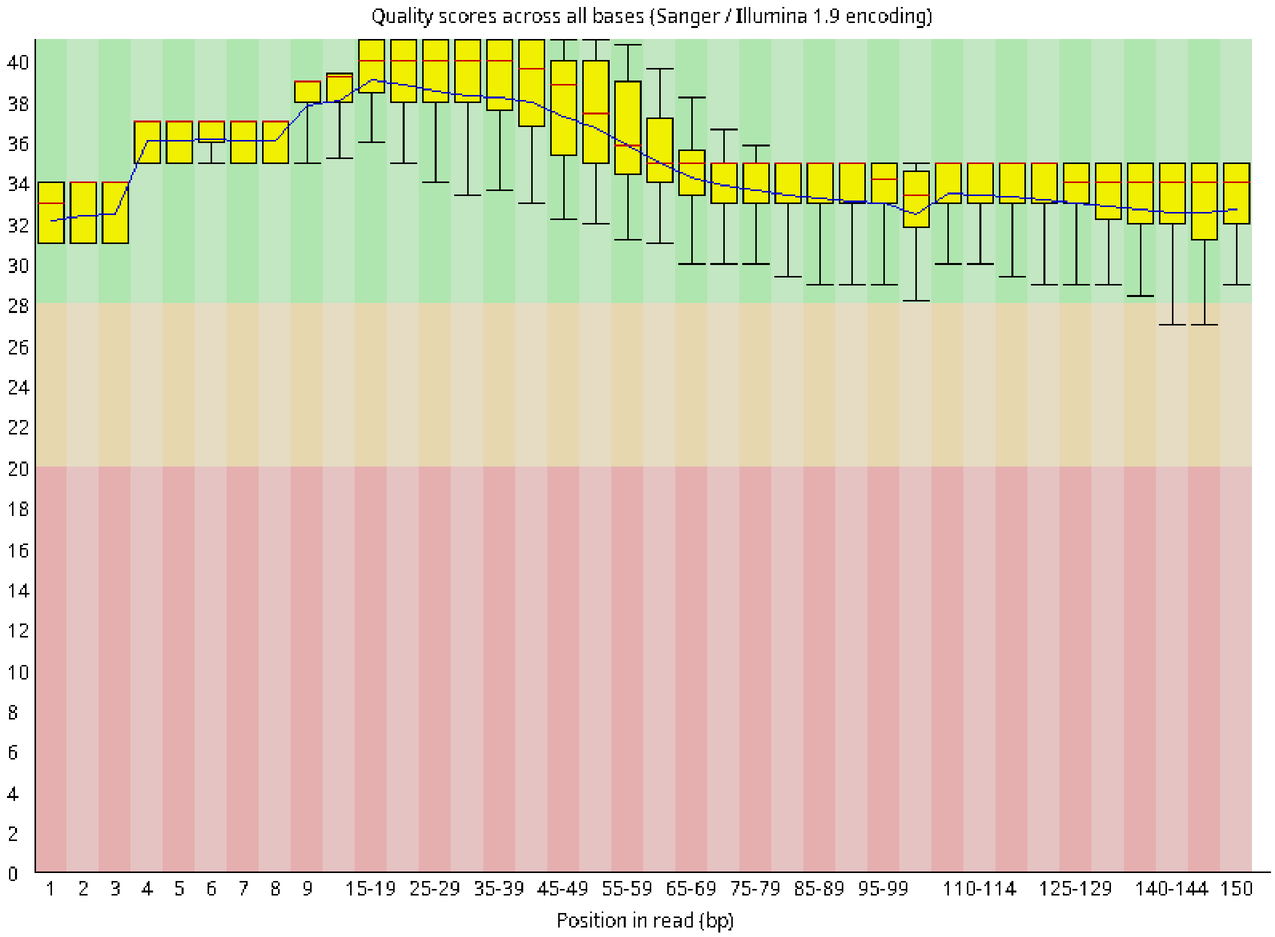
- SRR2589044_1





- SRR2589044_2





Alignment to a reference genome

1. CP000819.1

```
& head -n 1 ./data/ref_genome/ecoli_rel606.fasta
>CP000819.1 Escherichia coli B str. REL606, complete genome
```

```
$ bwa index data/ref_genome/ecoli_rel606.fasta
[bwa_index] Pack FASTA... 0.06 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 1.88 seconds elapse.
[bwa_index] Update BWT... 0.04 sec
[bwa_index] Pack forward-only FASTA... 0.04 sec
[bwa_index] Construct SA from BWT and Occ... 1.00 sec
[main] Version: 0.7.18-r1243-dirty
[main] CMD: bwa index data/ref_genome/ecoli_rel606.fasta
[main] Real time: 3.591 sec; CPU: 3.025 sec
```

А дальше тупик

```
$ bwa mem data/ref_genome/ecoli_rel606.fasta
data/trimmed_fastq_small/SRR2584866_1.trim.sub.fastq
```

```
data/trimmed_fastq_small/SRR2584866_2.trim.sub.fastq > results/sam/SRR2584866.aligned.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 77446 sequences (10000033 bp)...
[M::process] read 77296 sequences (10000182 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (48, 36728, 21, 61)
[M::mem_pestat] analyzing insert size distribution for orientation FF...
[M::mem_pestat] (25, 50, 75) percentile: (420, 660, 1774)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 4482)
[M::mem_pestat] mean and std.dev: (784.68, 700.87)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 5836)
[M::mem_pestat] analyzing insert size distribution for orientation FR...
[M::mem_pestat] (25, 50, 75) percentile: (221, 361, 576)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 1286)
[M::mem_pestat] mean and std.dev: (412.89, 227.06)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 1641)
[M::mem_pestat] analyzing insert size distribution for orientation RF...
[M::mem_pestat] (25, 50, 75) percentile: (560, 2011, 2594)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 6662)
[M::mem_pestat] mean and std.dev: (1580.30, 978.54)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 8696)
[M::mem_pestat] analyzing insert size distribution for orientation RR...
[M::mem_pestat] (25, 50, 75) percentile: (320, 549, 942)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 2186)
[M::mem_pestat] mean and std.dev: (581.31, 431.43)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 2808)
[M::mem_pestat] skip orientation FF
[M::mem_pestat] skip orientation RF
[M::mem_pestat] skip orientation RR
Segmentation fault (core dumped)
```

Вылетает Си-шная ошибка, которая часто возникает когда обращение по индексу корявое, но откуда она вылетает и в чём здесь может быть проблема, понятия не имею

Так что не знаю что делать дальше, я на это уже очень много времени потратил