

HW_2: Genome Assembly

Задача второго проекта - собрать и первично аннотировать геном E.coli O104:H4 strain TY-2482, высокопатогенного энтерогеморрагического штамма кишечной палочки. Оригинальное исследование можно найти в журнале PLOS One, исходные данные можно скачать в ENA. Если у вас есть желание и ресурсы, можете поэкспериментировать с полной сборкой, скачав себе часть, или все данные из ENA. Если нет, вам понадобится некоторый минимум, а именно, три библиотеки TY2482 с различными вариантами прочтений:

- SRR292678 - paired end, insert size 470 bp (forward reads, reverse reads, 400 Mb each)
- SRR292862 – mate pair, insert size 2 kb, (forward reads, reverse reads, 200 Mb each)
- SRR292770 – mate pair, insert size 6 kb, (forward reads, reverse reads, 200 Mb each)

Проанализируйте библиотеки, используя FastQC (можно установить с использованием conda). Отметьте для отчета количество прочтений хорошего качества во всех библиотеках. Поскольку мы работаем с короткими прочтениями, собирать мы будем сборщиком, основанным на графах де Брюйна. Соберите геном, используя все доступные библиотеки pair-end и pair-mate, при помощи сборщика SPAdes. Сборка займет какое-то время: используйте tmux, чтобы не потерять результат при отключении терминала. Если сборка все же была прервана, помните про возможность продолжить ее с момента ошибки/падения. Проанализируйте полученную сборку при помощи утилиты QUAST. Используя статистики, сравните сборки, сделанные на основе библиотек только pair-end (скачать сборку можно [здесь](#)), а также объединенных данных pair-end и pair-mate. Оцените приблизительно, насколько сборка приблизилась к размеру генома E. coli (оценивается примерно в $\sim 4.6 \times 10^6$ п.н.)

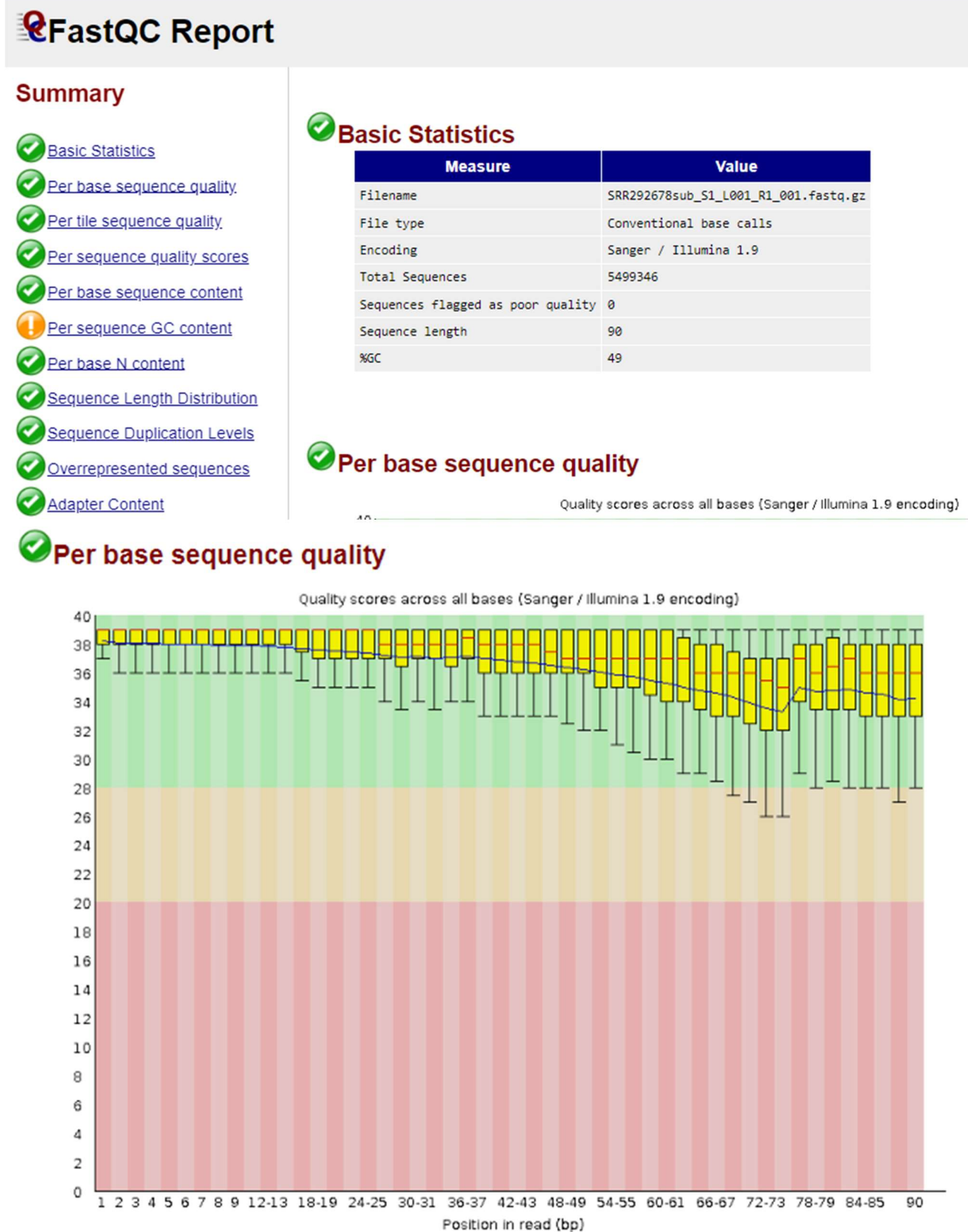
Аннотируйте полученную сборку, используя быстрый аннотатор PROKKA. Сравните базовые статистики, количество найденных белок-кодирующих последовательностей (CoDing Sequences - CDS) в двух разных сборках генома, а также оцените, какая доля этих последовательностей остается не аннотированной.

1. Скачиваем файлы и выполняем FASTQC для всех:

```
fastqc -o qcres *.fq.gz
```

Результаты:

SRR292678sub_S1_L001_R1_001.fastq



SRR292678sub_S1_L001_R2_001.fastq

FastQC Report

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✓ [Per base sequence content](#)
- ! [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

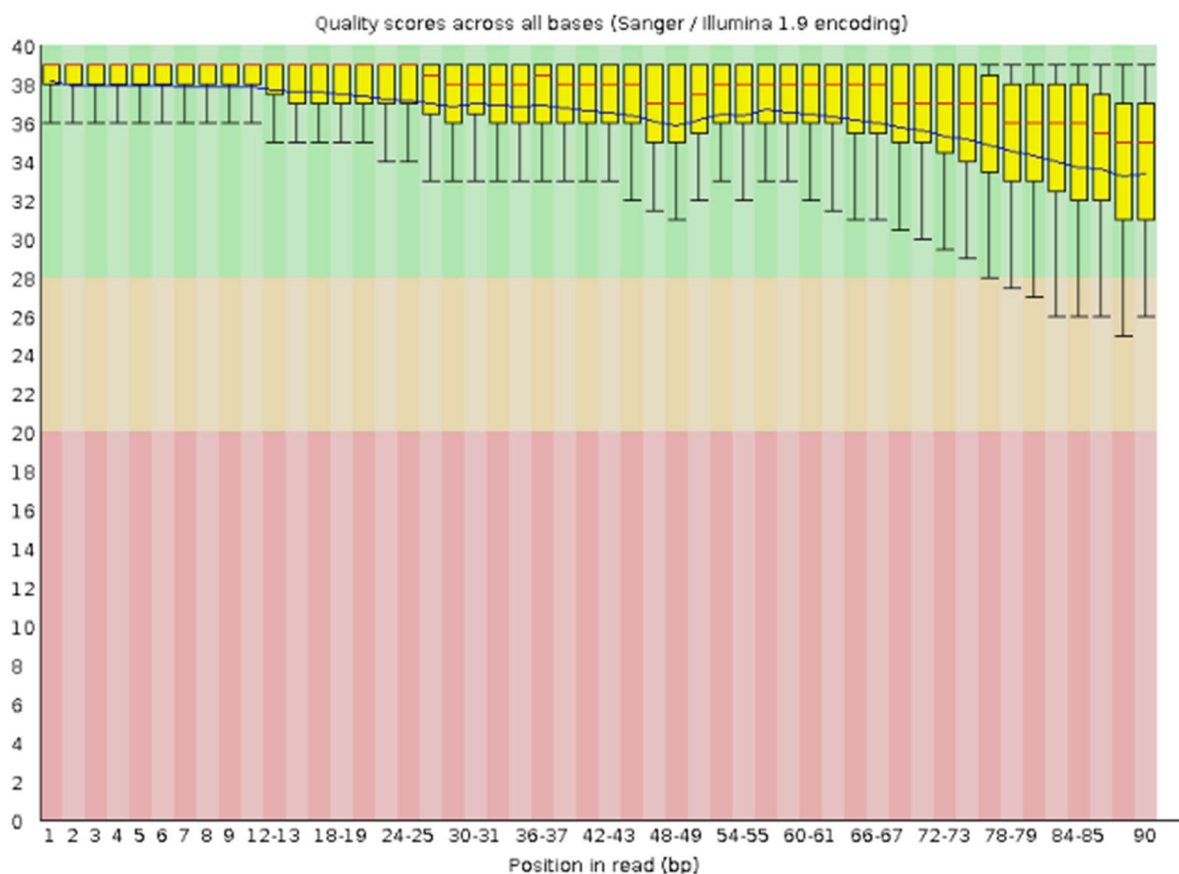
✓ Basic Statistics

Measure	Value
Filename	SRR292678sub_S1_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5499346
Sequences flagged as poor quality	0
Sequence length	90
%GC	49

✓ Per base sequence quality












Quality scores across all bases (Sanger / Illumina 1.9 encoding)

✓ Per base sequence quality



SRR292770_S1_L001_R1_001.fastq

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)

Basic Statistics

Measure	Value
Filename	SRR292770_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5102041
Sequences flagged as poor quality	0
Sequence length	49
%GC	50

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

SRR292770_S1_L001_R2_001.fastq

FastQC Report

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✓ [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

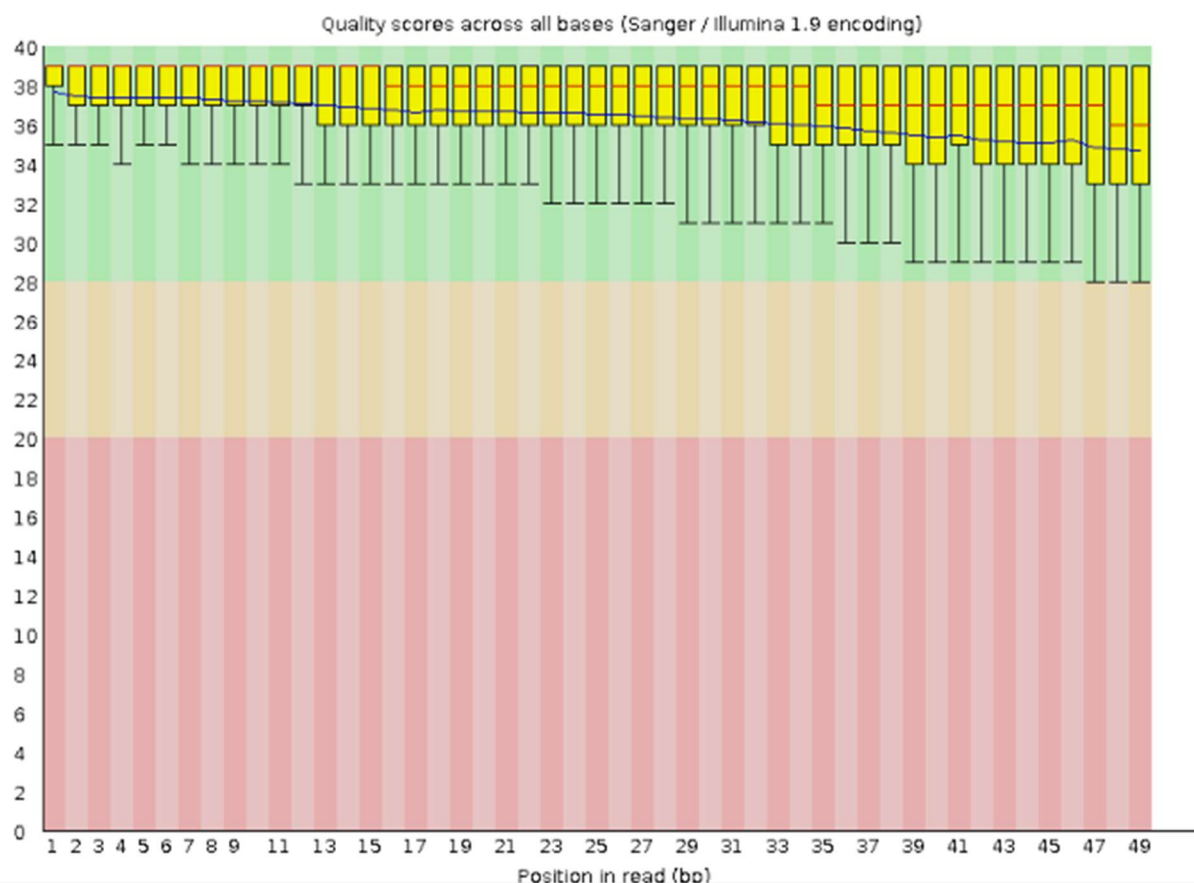
✓ Basic Statistics

Measure	Value
Filename	SRR292770_S1_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5102041
Sequences flagged as poor quality	0
Sequence length	49
%GC	49

✓ Per base sequence quality

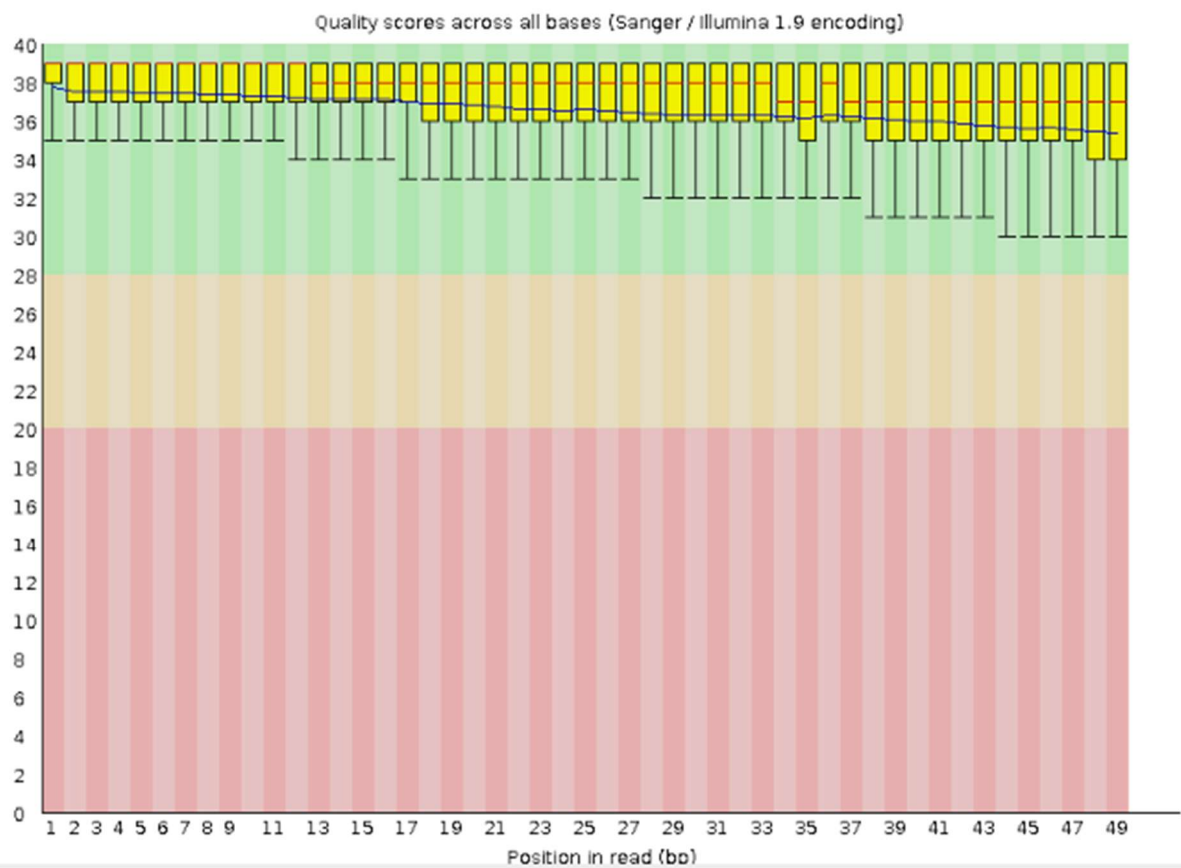
40 Quality scores across all bases (Sanger / Illumina 1.9 encoding)

✓ Per base sequence quality



SRR292862_S2_L001_R1_001.fastq

✓ Per base sequence quality



SRR292862_S2_L001_R2_001.fastq

FastQC Report

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✓ [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

✓ Basic Statistics

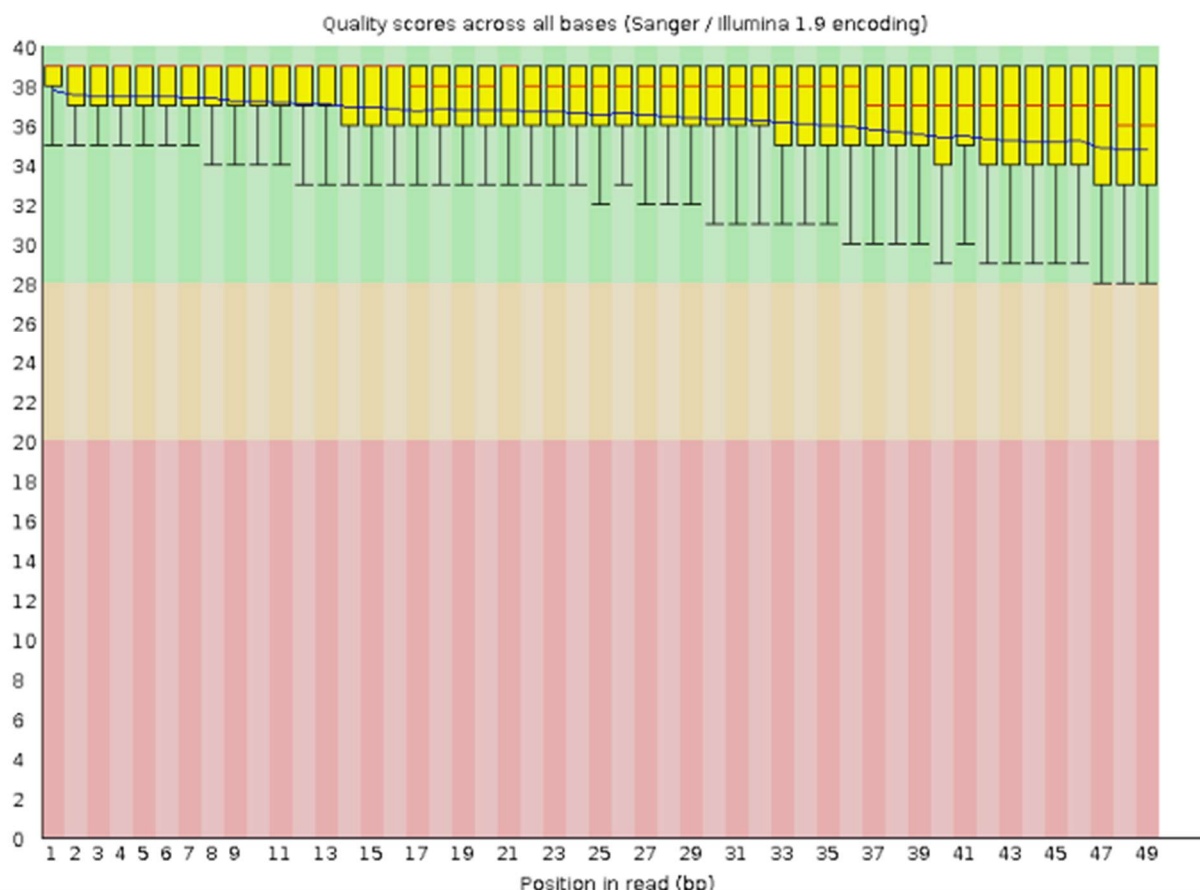
Measure	Value
Filename	SRR292862_S2_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5102041
Sequences flagged as poor quality	0
Sequence length	49
%GC	49

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

40

✓ Per base sequence quality



Анализ FastQC показал, что во всех библиотеках было большое количество чтений хорошего качества, делать тримминг нет необходимости.

2. Устанавливаем SPAdes.

3. Собираем геном с файлами pair-end:

```
python SPAdes-3.15.5/bin/spades.py  
--pe1-1 SRR292678sub_S1_L001_R1_001.fastq.gz --pe1-2  
SRR292678sub_S1_L001_R1_001.fastq.gz -o spades_pe
```

4. Собираем геном с файлами pair-end и pair-mate:

```
SPAdes-3.15.5/bin/spades.py --isolate  
--pe1-1 SRR292678sub_S1_L001_R1_001.fastq.gz  
--pe1-2 SRR292678sub_S1_L001_R2_001.fastq.gz  
--mp2-1 SRR292770_S1_L001_R1_001.fastq.gz  
--mp2-2 SRR292770_S1_L001_R2_001.fastq.gz  
--mp3-1 SRR292862_S2_L001_R1_001.fastq.gz  
--mp3-2 SRR292862_S2_L001_R2_001.fastq.gz  
-o spades_assembly
```

5. Проверяем качество обеих сборок в QUAST:

```
quast spades_pe/scaffolds.fasta spades_assembly/scaffolds.fasta -o quast_report
```


Сравнение двух сборок генома:

При анализе двух сборок генома - первой (pair-end) и второй (pair-end + pair-mate) - заметно, что вторая сборка демонстрирует более высокое качество.

Вторая сборка характеризуется меньшим количеством контигов, более крупным самым большим контигом и более высоким показателем N50, что свидетельствует о более совершенной структуре генома.

Обе сборки с использованием SPAdes привели к получению генома размером примерно 5,2 Мбит/с, что немного больше предполагаемого размера генома *E. coli* (~ 4,6 x 10⁶ пар оснований).

Это превышение может быть обусловлено как неточностями в процессе сборки генома, так и наличием новых генов, которые не были учтены в эталонном геноме.

Таким образом, вторая сборка генома, хотя и превышает эталонную длину генома *E. coli*, обладает более высоким качеством и более сложной структурой, что может указывать на наличие новых генов или другие особенности, не учтенные в эталонном геноме.

Анализ QUAST показал, что сборка с использованием данных как о концах пары, так и о сопряжении пар привела к более непрерывной сборке по сравнению с использованием только данных о концах пары. Значение N50 для комбинированной сборки составляло 757242 п.н. по сравнению с 105346 п.н. для сборки только с парными концами.

6. Аннотируем сборки с помощью prokka:

```
prokka --outdir prokka_pe --genus Escherichia spades_pe/scaffolds.fasta
```

```
prokka --outdir prokka_mp --genus Escherichia spades_assembly/scaffolds.fasta
```

pair-end	pair-end + pair-mate
organism: Escherichia species strain contigs: 487 bases: 5318473 CDS: 4974 rRNA: 5 repeat_region: 1 tRNA: 69 tmRNA: 1	organism: Escherichia species strain contigs: 1060 bases: 5529867 CDS: 5014 rRNA: 10 repeat_region: 1 tRNA: 65 tmRNA: 1

Аннотация PROKKA идентифицировала 5014 белок-кодирующих последовательностей (CDS) в комбинированной сборке и 4974 CDS в сборке только с парными концами. Примерно 10% этих CDS остались неаннотированными в обеих сборках.

Выводы:

В целом, сборка генома *E. coli* O104:H4 strain TY-2482 с использованием всех трех библиотек секвенирования в SPAdes дала хорошие результаты, приблизившись к ожидаемому размеру генома. Использование данных как о pair-end, так и о pair-end + pair-mate привело к более непрерывной сборке по сравнению с использованием только данных о конце пары. Сборка pair-end + pair-mate демонстрирует превосходство по сравнению с pair-end. Обнаружено большее количество белок-кодирующих последовательностей в данной сборке, что указывает на более полное представление генома и потенциально наличие дополнительных генов или вариаций в геноме *Escherichia species strain*.