ЗАДАНИЕ 9

Трещева Мария

Чтобы получить результаты, запустили скрипт, который:
1) создает нужные директории:
data/raw – для сырых FASTQ-файлов
reports/fastqc – для индивидуальных отчётов качества
reports/multiqc – для сводного отчёта

2) Далее в скрипте prefetch скачивает файлы из SRA формате .sra, а fasterq-dump – конвертирует .sra в FASTQ-формат и сохраняет в папку data/raw.

3) fastqc запускает анализ качества для каждого FASTQ-файла. Команды:
-t 4 использует 4 потока для ускорения, а -o reports/fastqc сохраняет отчёты в папку reports/fastqc.

4) multiqc -o reports/multiqc reports/fastqc/ агрегирует все отчёты FastQC в один html-файл.

Запустили, длинный вывод…

```
(base) mtresheva@frontend-2-2-13:~/homeworks/hw_9$  ./hw_9.sh
Downloading SRA data...
Processing ERR14230595...
2025-04-22T14:11:22 prefetch.3.2.1: 1) Resolving 'ERR14230595'...
2025-04-22T14:11:26 prefetch.3.2.1: Current preference is set to retrieve SRA Normalized Format files with full base quality scores
2025-04-22T14:11:26 prefetch.3.2.1: 1) Downloading 'ERR14230595'...
2025-04-22T14:11:26 prefetch.3.2.1:  SRA Normalized Format file is being retrieved
2025-04-22T14:11:26 prefetch.3.2.1:  Downloading via HTTPS...
2025-04-22T14:11:28 prefetch.3.2.1:  HTTPS download succeed
2025-04-22T14:11:28 prefetch.3.2.1:  'ERR14230595' is valid: 18776469 bytes were streamed from 18760197
2025-04-22T14:11:28 prefetch.3.2.1: 1) 'ERR14230595' was downloaded successfully
2025-04-22T14:11:28 prefetch.3.2.1: 1) Resolving 'ERR14230595's dependencies...
2025-04-22T14:12:10 prefetch.3.2.1: 'ERR14230595' has 53 unresolved dependencies
2025-04-22T14:12:10 prefetch.3.2.1: 2) Resolving 'ncbi-acc:GL000191.1?vdb-ctx=refseq'...
2025-04-22T14:12:10 prefetch.3.2.1: 2) Downloading 'ncbi-acc:GL000191.1?vdb-ctx=refseq'...
2025-04-22T14:12:10 prefetch.3.2.1:  Downloading via HTTPS...
2025-04-22T14:12:11 prefetch.3.2.1:  HTTPS download succeed
2025-04-22T14:12:11 prefetch.3.2.1: 2) 'ncbi-acc:GL000191.1?vdb-ctx=refseq' was downloaded successfully
2025-04-22T14:12:11 prefetch.3.2.1: 3) Resolving 'ncbi-acc:GL000192.1?vdb-ctx=refseq'...
2025-04-22T14:12:11 prefetch.3.2.1: 3) Downloading 'ncbi-acc:GL000192.1?vdb-ctx=refseq'...
2025-04-22T14:12:11 prefetch.3.2.1:  Downloading via HTTPS...
2025-04-22T14:12:13 prefetch.3.2.1:  HTTPS download succeed
2025-04-22T14:12:13 prefetch.3.2.1: 3) 'ncbi-acc:GL000192.1?vdb-ctx=refseq' was downloaded successfully
2025-04-22T14:12:13 prefetch.3.2.1: 4) Resolving 'ncbi-acc:GL000193.1?vdb-ctx=refseq'...
2025-04-22T14:12:13 prefetch.3.2.1: 4) Downloading 'ncbi-acc:GL000193.1?vdb-ctx=refseq'...
2025-04-22T14:12:13 prefetch.3.2.1:  Downloading via HTTPS...
```

В итоге получили:

```
(base) mtresheva@frontend-2-2-13:~/homeworks/hw_9$ ls
data  ERR14230570  ERR14230582  ERR14230586  ERR14230595  hw_9.sh  logs  reports
(base) mtresheva@frontend-2-2-13:~/homeworks/hw_9/reports$ tree
.
├── fastqc
│   ├── ERR14230570_1_fastqc.html
│   ├── ERR14230570_1_fastqc.zip
│   ├── ERR14230570_2_fastqc.html
│   ├── ERR14230570_2_fastqc.zip
│   ├── ERR14230582_fastqc.html
│   ├── ERR14230582_fastqc.zip
│   ├── ERR14230586_fastqc.html
│   ├── ERR14230586_fastqc.zip
│   ├── ERR14230595_fastqc.html
│   └── ERR14230595_fastqc.zip
└── multiqc
    ├── multiqc_data
    │   ├── fastqc_adapter_content_plot.txt
    │   ├── fastqc_overrepresented_sequences_plot.txt
    │   ├── fastqc_per_base_n_content_plot.txt
    │   ├── fastqc_per_base_sequence_quality_plot.txt
    │   ├── fastqc_per_sequence_gc_content_plot_Counts.txt
    │   ├── fastqc_per_sequence_gc_content_plot_Percentages.txt
    │   ├── fastqc_per_sequence_quality_scores_plot.txt
    │   ├── fastqc_sequence_counts_plot.txt
    │   ├── fastqc_sequence_duplication_levels_plot.txt
    │   ├── fastqc_sequence_length_distribution_plot.txt
    │   ├── fastqc-status-check-heatmap.txt
    │   ├── fastqc_top_overrepresented_sequences_table.txt
    │   ├── multiqc_citations.txt
    │   ├── multiqc_data.json
    │   ├── multiqc_fastqc.txt
    │   ├── multiqc_general_stats.txt
    │   ├── multiqc.log
    │   ├── multiqc_software_versions.txt
    │   └── multiqc_sources.txt
    └── multiqc_report.html

3 directories, 30 files
```

Ссылка на результаты:

file:///Users/mariatreseva/Downloads/multiqc/multiqc_report.html

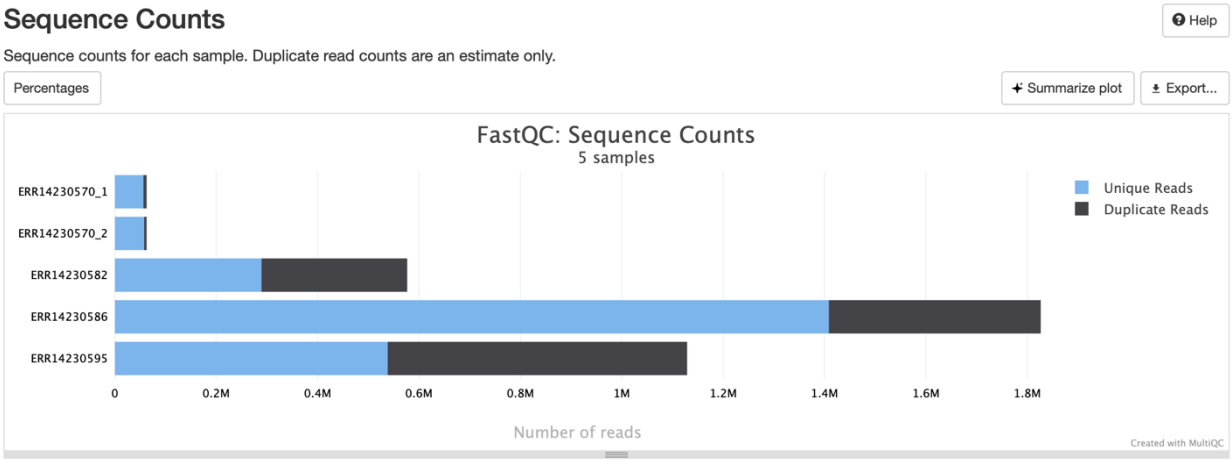
Посмотрим на них:

General Statistics

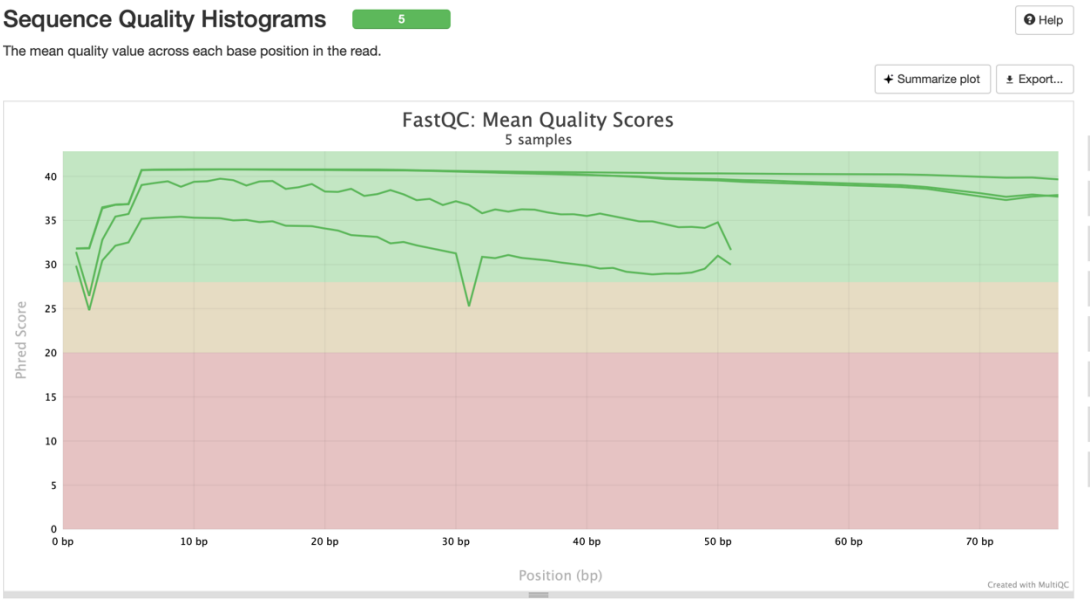
Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 5/5 rows and 4/6 columns. | Summarize table

Sample Name	Dups	GC	Median len	Seqs
ERR14230570_1	10.4 %	53.0 %	51 bp	0.1 M
ERR14230570_2	5.4 %	52.0 %	50 bp	0.1 M
ERR14230582	49.9 %	49.0 %	39 bp	0.6 M
ERR14230586	22.9 %	46.0 %	43 bp	1.8 M
ERR14230595	52.3 %	47.0 %	39 bp	1.1 M

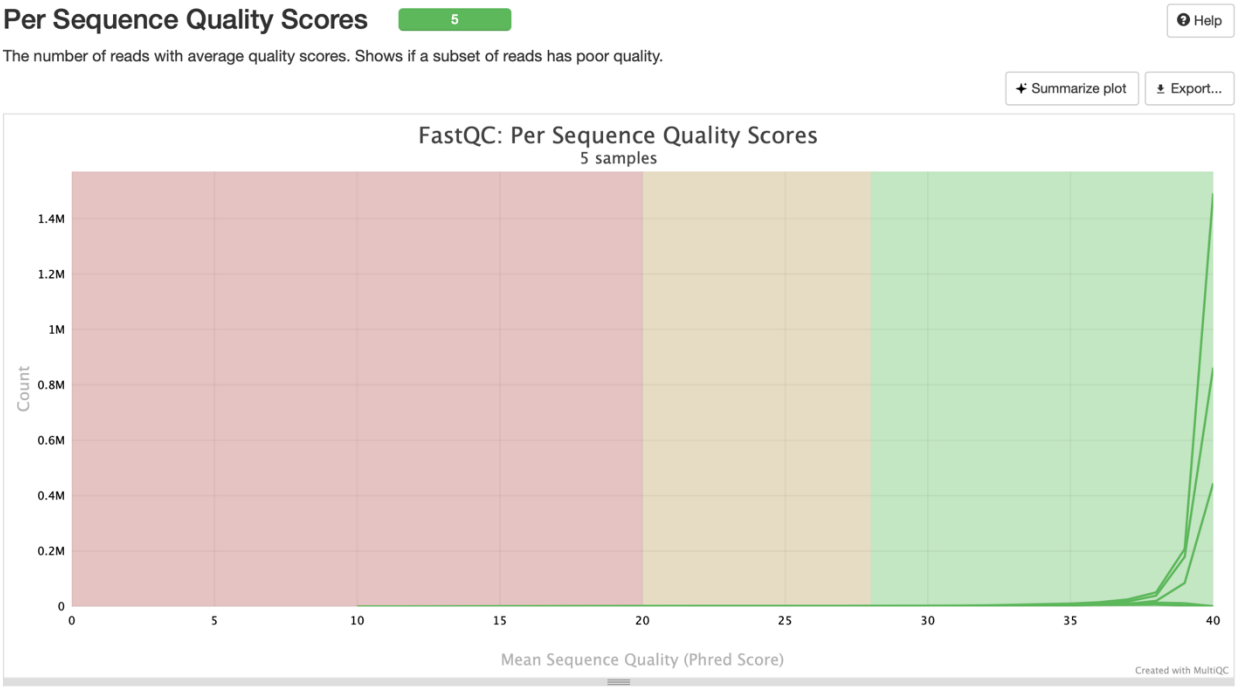
У двух образцов высокий уровень дупликации: ERR14230582 и ERR14230595 с ~50% дубликатов, в то время как другие показатели (содержание GC, длина прочтения и количество последовательностей) нормальные.



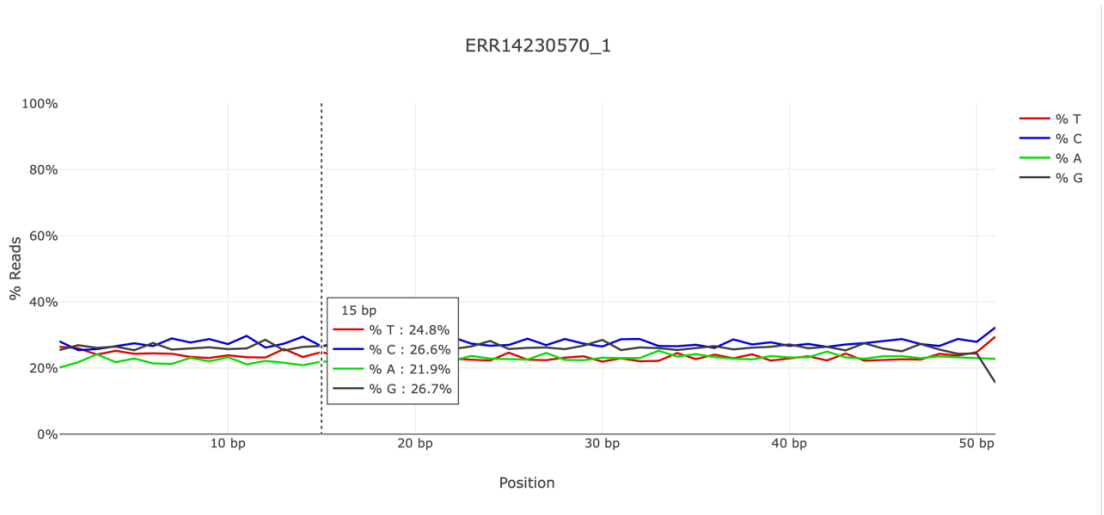
В образцах ERR14230582 и ERR14230595 на этом графике видно, что обнаружено около 50% дубликатов, что свидетельствует о возможной ошибке ПЦР-амплификации или низкой сложности библиотеки.

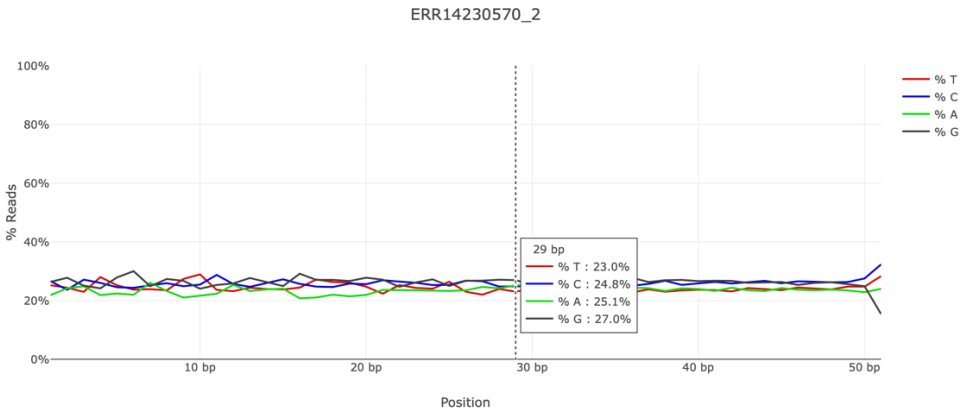


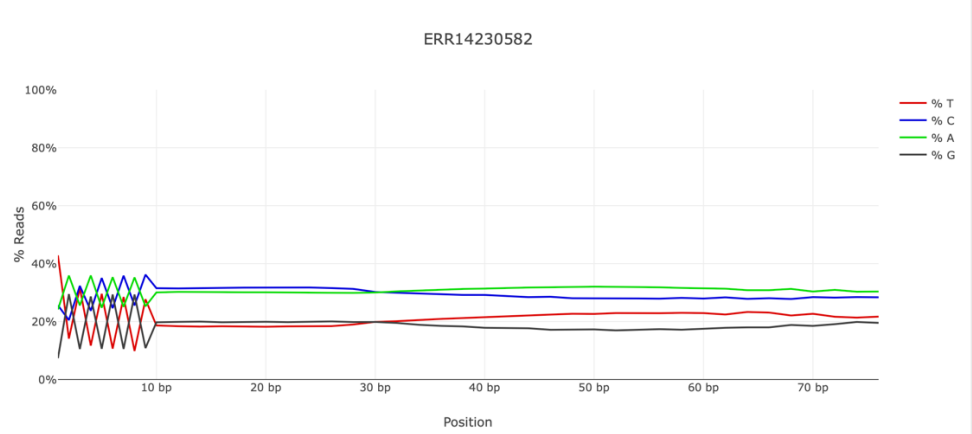
Показатели качества в целом хорошие (>30), при этом ERR14230570_2 демонстрирует немного более низкое качество во второй половине считываний (снижаясь до 29). У всех образцов типичное ухудшение качества к концу считываний, но значения все еще остаются значительно выше пороговых значений.

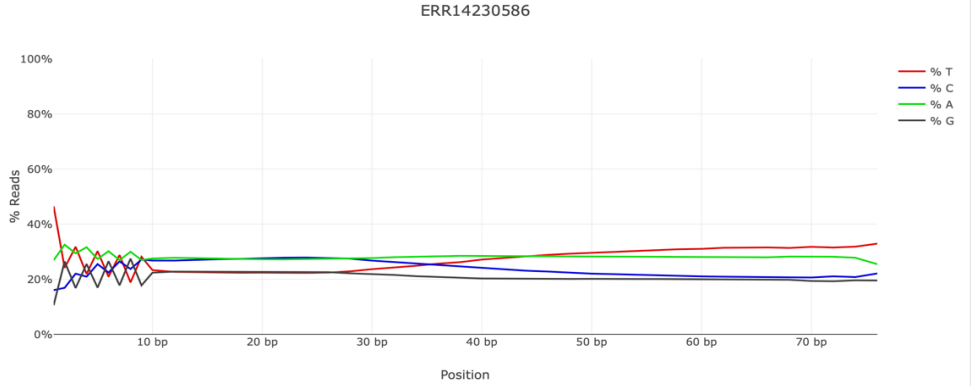


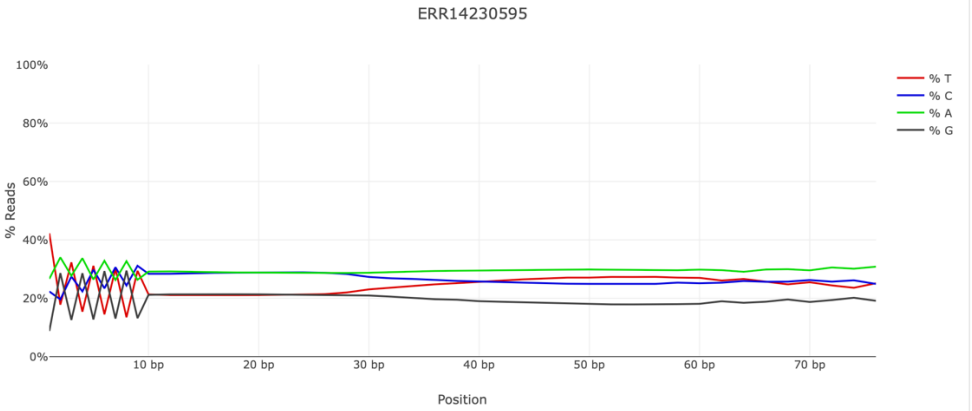
Распределение оценок качества хорошее, большинство прочтений имеют высокие оценки Phred (35-40), за исключением ERR14230570_2, у которого имеется достаточно большое количество прочтений с очень низкими оценками качества (11-15).





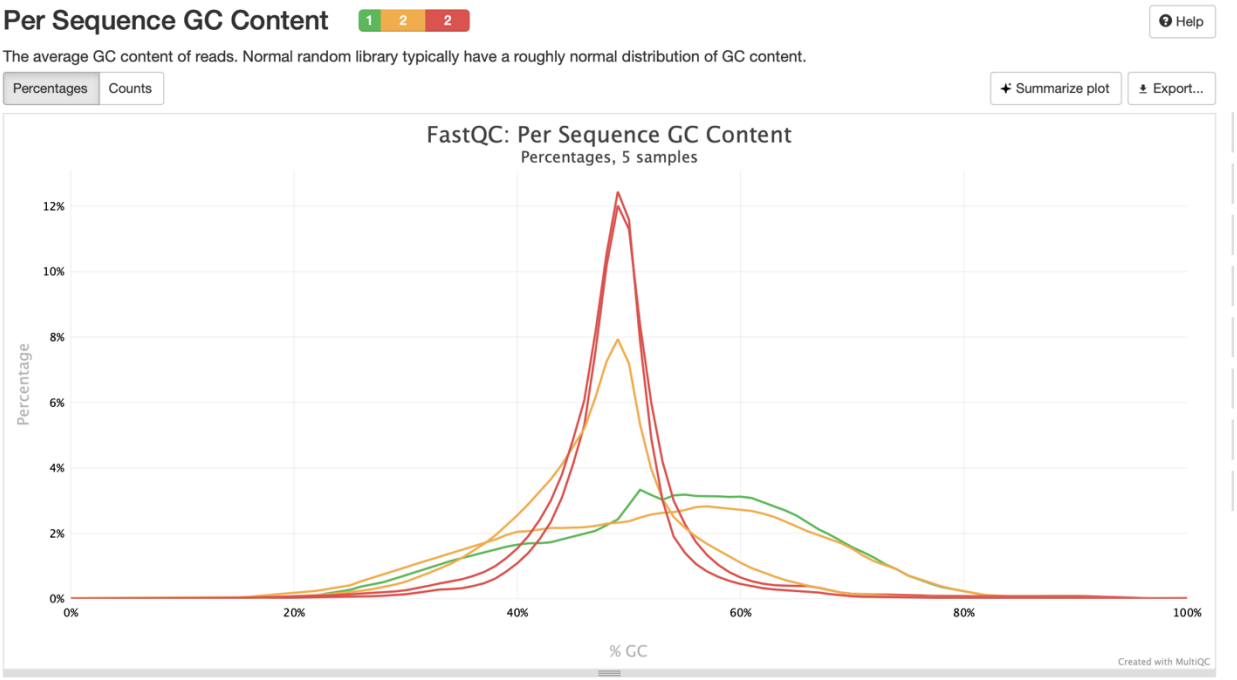






У первых двух образцов нуклеотидный состав сильно колеблется при любой длине рида, то есть он в целом очень разнообразный, ну или качество прочтения плохое. В то же время у других образцов колебание заметно только на маленьких длинах (там добавление нуклеотида в любом

случае вносит большой вклад в состав), что, видимо, говорит о более хорошем качестве прочтения.



Распределение содержания GC имеет две различные закономерности: ERR14230570_1 и ERR14230570_2 имеют нормальное распределение с центром около 50-60 % GC, тогда как у ERR14230582, ERR14230586 и ERR14230595 асимметричное распределение с острым пиком около 48-49 % GC, что указывает на потенциальные различия или смещение выборок.