Пояснение к домашней работе №3 по биоинформатике

Данные были загружены отсюда:
https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR24651073&display=metadata
run SRR24651073
experiment SRX20431105

Данные были загружены из командной строки:
```
fasterq-dump SRR24651073
```

Работа с Kedro:

1. **Создание виртуальной среды:**

```
mkdir kedro-environment && cd kedro-environment
apt install python3.10-venv
python3 -m venv .venv
source .venv/bin/activate
```

2. **Установка Kedro**

```
pip install kedro
```

3. **Создание нового проекта**

```
kedro new
cd <project name>
pip install -r src/requirements.txt
```

4. **Конфигурация пайплайна**

Для запуска пайплайна, нужно указать расположение входных файлов в системе. В директории *…/kedro-environment/helloworld/data/01_raw* содержатся конфигурационные файлы, указывающие на расположение входных данных. В файле ./bwa_in/refseq_fasta_path.txt нужно указать путь до референсной последовательности, а в ./fastqc_in/ пути до двух входных .fastq файлов соответственно.

5. **Использование написанного пайплайна:**

1. В папке kedro-environment активировать виртуальную среду
```
source .venv/bin/activate
```
2. `cd helloworld`
3. `kedro run –pipeline genomic_pipeline`

Для визуализации:
pip install kedro-viz

Запуск:
В папке kedro-environment/helloworld:
```
kedro viz
```

В целом, схема выполнения пайплайна получилась схожей, помимо того, что она содержит некоторые вспомогательные переменные, отвечающие за выполнение/невыполнение команд и тех, которые не используются в пайплайне, но должны быть указаны согласно принципу работы фреймворка.

Также, фреймворк заточен под ML, поэтому не удалось осуществить передачу файлов напрямую из одного блока пайплайна в следующий, это пришлось осуществить путем передачи путей до нужных файлов.