Practical: Quality Control and Trimming

In this practical, you will learn to import, view, and check the quality of raw high-throughput sequencing data using FastQC and Trimmomatic (via Docker).

The first dataset is from an Illumina MiSeq sequencing of *enterohaemorrhagic E. coli* (EHEC), serotype O157 — a potentially fatal gastrointestinal pathogen involved in a 2011 outbreak in St. Louis, USA. The data is paired-end 2×150 bp reads.

Working Directory

Create a working directory

```
mkdir ~/work
cd ~/work
```

Downloading the Data

In this practical, you will use the European Nucleotide Archive (ENA) to download real Illumina sequencing data for quality control and trimming exercises. ENA is a public database that stores nucleotide sequencing data, such as DNA and RNA sequences, submitted by researchers around the world.

It is part of the EMBL-EBI (European Bioinformatics Institute). ENA provides free access to raw sequencing reads, genome assemblies, and functional annotation. Accession numbers like SRR957824 are unique IDs that help you find specific datasets.

The raw data is available on ENA under accession number SRR957824.

Step 1: Download data

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR957/SRR957824/SRR957824_1.fast
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR957/SRR957824/SRR957824_2.fast
```

Step 2: Check file size

```
ls -l
```

There are 500 000 paired-end reads taken randomly from the original data

One last thing before we get to the quality control: those files are writeable. By default, UNIX makes things writeable by the file owner. This poses an issue with creating typos or errors in raw data. We fix that before going further

```
chmod u-w *
```

Step 3: Preview a FASTQ file

```
zless SRR957824_1.fastq.gz
```

Tip: Use the spacebar to scroll and `q` to exit.

The fastq format is a text-based format that represents nucleotide sequences but also contains the corresponding quality of each nucleotide. It is the standard for storing the output of high-throughput sequencing instruments such as the Illumina machines.

A fastq file uses four lines per sequence:

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence of letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

You can read more on the FASTQ format here.

Question: Where does the filename come from?

Question: Why are there `_1` and `_2` in the file names?

FastQC (Pre-Trimming Quality Check)

To check the quality of the sequence data, we will use a tool called FastQC.

FastQC has a graphical interface and can be downloaded and run on a Windows or Linux computer without installation. It is available here.

However, FastQC is also available as a command-line utility in Docker. Docker makes things simple and consistent. Instead of installing software like FastQC or Trimmomatic yourself, Docker lets you download small packages (called containers) that already have everything set up. You can then run them directly on your data using simple commands. Docker makes it easy to run bioinformatics tools without worrying about installation or compatibility.

Think of Docker like a shipping container for software. Just like shipping containers hold goods and can be transported anywhere, Docker containers hold software and can run anywhere: on your laptop, a lab server, or the cloud without needing to be unpacked or reinstalled. Everything the program needs is bundled inside.

Step 1: Pull the FastQC Docker image

```
docker pull biocontainers/fastqc:v0.11.9_cv8
```

Step 2: Run FastQC in Docker

```
docker run --rm -v "$PWD":/data biocontainers/fastqc:v0.11.9_cv8 fastqc /
```

Step 3: List output files

```
ls *fastqc*
```

For each file, FastQC has produced both a .zip archive containing all the plots and a HTML report. Download and open the HTML files with your favourite web browser.

Question: What should you pay attention to in the FastQC report?

Question: Which file is of better quality?

Pay special attention to:

- Per-base sequence quality
- Sequence length distribution
- Adapter content

Explanations for the various quality modules can be found here. Also, have a look at examples of a good and a bad illumina read set for comparison.

You will note that the reads in your uploaded dataset have fairly poor quality (<20) towards the end. There are also outlier reads that have very poor quality for most of the second half of the reads.

Adapter Trimming with Trimmomatic

Trimmomatic is a tool that removes adapter sequences and trims low-quality regions from sequencing reads. It works by:

Scanning the reads for known adapter sequences using exact or partial matches.

Removing low-quality bases from the start and end of each read.

Trimming using a sliding window that checks the average quality within a region.

Discarding reads that are too short after trimming.

It helps clean up the data so that poor-quality sequences don't interfere with downstream analysis.

Step 1: Download the adapter file

```
curl -O -J -L https://osf.io/v24pt/download -o adapters.fa
```

Step 2: Pull Trimmomatic Docker image

```
docker pull quay.io/biocontainers/trimmomatic:0.39--hdfd78af_2
```

Step 3: Run Trimmomatic (paired-end)

```
docker run --rm -v "$PWD":/data quay.io/biocontainers/trimmomatic:0.39--h
```

Question: What adapters were used, and how does Trimmomatic identify them?

Question: What are the unpaired FASTQ files, and why are they generated?

FastQC (Post-Trimming Quality Check)

Step 1: Run FastQC again on filtered reads

```
docker run --rm -v "$PWD":/data biocontainers/fastqc:v0.11.9_cv8 fastqc /
```

Step 2: List outputs

```
ls *trimmed*_fastqc.html
```

Open both `.html` files in your browser and look at the reports.

Question: What improvements are visible after trimming?

Question: How did trimming affect per-base quality and read lengths?

Additional Questions about FastQC

Question: Which FastQC modules showed the most improvement after trimming?

Question: Did the adapter content change after trimming? How can you tell?

Question: How does the sequence length distribution look after trimming compared to before?

Question: Would you expect every dataset to need trimming? Why or why not?

Question: What are the potential consequences of skipping quality control and trimming in a diagnostic laboratory?

Summary

- You downloaded and inspected paired-end Illumina reads.
- Used Docker-based FastQC to assess raw data.

- Trimmed adapters and low-quality regions using Trimmomatic.
- Verified improvements using FastQC.

This workflow ensures data quality and prepares reads for reliable downstream analysis.