Answers: Quality Control and Trimming Practical

# 📁 File Management & Initial Questions

**Where does the filename come from?**

The filename `SRR957824\_1.fastq.gz` comes from the ENA (European Nucleotide Archive) accession number. It identifies the sequencing run and read pair.

**Why are there \_1 and \_2 in the file names?**

`\_1` and `\_2` indicate paired-end sequencing: `\_1` is the forward read, and `\_2` is the reverse read.

# 🔍 FastQC (Pre-Trimming)

**What should you pay attention to in the FastQC report?**

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

**Which file is of better quality?**

`SRR957824\_1.fastq.gz` has slightly better quality. The second read (`\_2`) often shows lower quality toward the 3' end. This is common in Illumina sequencing because the second read is generated after the first, using a second round of synthesis. By that stage, the sequencing chemistry and reagents may have degraded slightly, or the instrument may introduce more errors in later cycles, especially near the 3' end of read 2. The second read (\_2) often shows lower quality toward the 3' end.

# ✂️ Trimmomatic Adapter Trimming

**What adapters were used and how does Trimmomatic identify them?**

The adapters were downloaded from https://osf.io/v24pt/download and passed via the `ILLUMINACLIP` parameter. Trimmomatic scans for matches (full or partial) to these sequences and clips them out based on quality thresholds.

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

# If one read in a pair is too short or low quality after trimming, it is excluded from the paired set and saved separately. This allows the remaining good-quality read to still be used in analyses that can handle unpaired reads. It helps preserve data rather than discarding both reads in a pair when only one is problematic.

# 🔄 FastQC (Post-Trimming)

**What improvements are visible after trimming?**

- Better per-base sequence quality (especially at the 3' end)  
- Reduced adapter content  
- Fewer overrepresented sequences

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

- Average quality improved.  
- Read lengths became more variable due to trimming of poor-quality ends.

# 📊 Additional Questions about FastQC

**Which FastQC modules showed the most improvement after trimming?**

- Per-base sequence quality  
- Adapter content  
- Overrepresented sequences

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

Yes, adapter content decreased significantly. This is shown in the 'Adapter Content' module of the FastQC report (green or yellow instead of red).

**How does sequence length distribution look after trimming compared to before?**

- Before: Fixed-length reads (e.g. 150bp)  
- After: Mixed lengths, with many shorter reads due to trimming.

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

Not always. It depends on the sequencing quality, platform, and whether adapter removal was performed during preprocessing.

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

- False variant calls  
- Failed alignments or assemblies  
- Misinterpretation of diagnostic results  
- Waste of computational resources