

One read is easy. What about millions?

Assessing sequencing data with FastQC

FastQC summarizes quality patterns across an entire FASTQ file so you can decide what needs a closer look before downstream analysis.

FASTQ file

Millions of reads



FastQC

Summarizes patterns



HTML report

Plots + flags

KEY IDEA

FastQC inspects the data. It does not trim, filter, or fix reads.

What questions is FastQC helping us ask?

Think of the report as a set of clues about the dataset.

1

Base-call quality

Does confidence stay high across read positions?

2

Sequence composition

Do base and GC patterns look plausible for this library?

3

Technical patterns

Are adapters, duplication, read-length changes, or other unusual patterns visible?

IMPORTANT

A FastQC warning or failure means a pattern is unusual by its default thresholds. It does not automatically mean the sequencing data are unusable.

Install FastQC in Ubuntu

We will run it from the same WSL2 environment used in Parts 1-3.

1. Update the package list

```
sudo apt update
```

2. Install FastQC

```
sudo apt install fastqc -y
```

3. Check that it works

```
fastqc --version
```

READY?

If a FastQC version is printed, the command-line installation is ready.
You only need to install it once.

Use two small FASTQ datasets

The sequences are the same. Only the quality profile is different.

fastqc_good.fastq.gz

5,000 synthetic reads

High quality across the full 100-base read

fastqc_tail_drop.fastq.gz

5,000 synthetic reads

Quality becomes lower toward the read end

After downloading them into your training folder:

```
cd ~/bioinformatics_training  
ls -lh fastqc_*.fastq.gz
```

These are teaching files, not biological sequencing data.

Generate the quality-control reports

FastQC can analyse one or many FASTQ files. We will keep the commands easy to read.

1. Run both files

```
fastqc fastqc_good.fastq.gz  
fastqc fastqc_tail_drop.fastq.gz
```

2. Look at the output files

```
ls -lh *fastqc*
```

OUTPUT

Each FASTQ file produces an HTML report for viewing and a ZIP archive containing the report data.

3. Open the folder in Windows

```
explorer.exe .
```

Double-click each FastQC HTML report to open it in your browser.

Start with these three sections

Do not try to interpret every FastQC module at once.

1

Basic Statistics

How many reads? What read length? What GC content?

2

Per base sequence quality

How does the Phred quality change from the start to the end of the read?

3

Adapter Content

Does known adapter sequence become more common at particular read positions?

PART 3 CONNECTION

The per-base quality plot is simply the Phred-score idea from Part 3 summarized across thousands or millions of reads.

Green, yellow, and red are clues

FastQC uses thresholds to flag patterns for your attention.

PASS

Within the module's default expectations

WARN

Unusual enough to inspect

FAIL

Strongly unusual by that module's threshold

WET-LAB CHECK

Before deciding that a warning is a problem, ask whether it fits the sample type, library design, sequencing platform, and expected biology. A flag tells you where to look; it does not tell you the exact cause.

Your turn.

Run FastQC and compare the two practice datasets.

- 1 Run FastQC on both FASTQ files.
- 2 Open both HTML reports.
- 3 Which dataset shows lower quality toward the end of the read?
- 4 Which FastQC module makes that easiest to see?
- 5 Does the report tell you the exact wet-lab or sequencing cause? Why not?

TO REMEMBER

FastQC detects patterns across a dataset. Interpretation still requires experimental and biological context.

NEXT

Part 5: Which FastQC warnings matter, and what should you do next?