

Which FastQC warnings actually matter?

Flags are clues. Context determines the action.

The goal is not to make every FastQC module green. The goal is to recognize patterns that could affect your analysis and decide what to investigate next.

FLAG**A module shows WARN or FAIL****PLOT****Look at the actual pattern****CONTEXT****Ask whether it is expected for this library****ACTION****Trim, investigate, continue, or document****KEY IDEA****Do not chase green ticks. Interpret each warning in the context of the experiment.**

Ask three questions first

This prevents unnecessary trimming or filtering.

1

Is it expected?

Amplicon, targeted, RNA-seq, small-RNA, and other libraries can legitimately look unusual.

2

Could it be technical?

Adapters, poor-quality tails, primers, contamination, or run problems may create real warning signals.

3

Will it affect my next step?

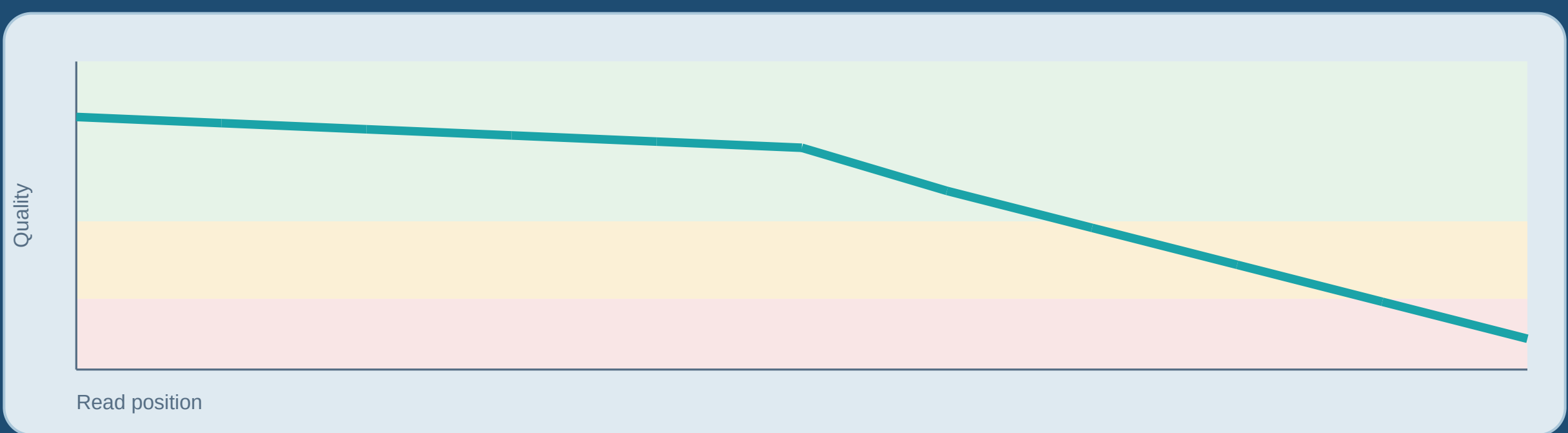
A pattern matters most when it can bias mapping, assembly, variant calling, or another downstream analysis.

PRACTICAL RULE

Inspect the plot first. Never make a processing decision from the PASS/WARN/FAIL label alone.

Per base sequence quality drops

A common pattern, especially toward the end of reads.



1

What it may mean

Later cycles have lower-confidence base calls, or a subset of reads has poor-quality tails.

2

What to do

Check how severe and widespread the drop is. Trimming may help if low-quality tails could interfere with the planned analysis.

WET-LAB LINK

Quality loss can reflect sequencing chemistry, signal quality, library quality, or platform-specific behavior. The plot shows the symptom, not the exact cause.

Adapter content or overrepresented sequences

Do not remove a sequence until you know what it is.



INSERT

ADAPTER

If the insert is shorter than the sequencing read, adapter sequence can be read.

1

What it may mean

Adapter or primer sequence is present in the reads, often because the biological insert is shorter than the read length.

2

What to do

Identify the sequence first. If it is a known adapter or unwanted primer, trim it with an appropriate tool before downstream analysis.

CAUTION

An overrepresented sequence can also be real biology, such as abundant amplicons or transcripts. Identification comes before removal.

Some warnings can be expected

The library design can explain what looks unusual.

1

Sequence duplication

High duplication may be expected in amplicon or targeted libraries. In a diverse whole-genome library, extreme duplication may deserve closer investigation.

2

GC distribution

GC content depends on the organism and library. An unexpected extra peak can suggest a mixed population or contamination, but context matters.

3

Per-base sequence content

Early-position bias can arise from primers, transposase or library-preparation chemistry. It is not automatically a failed run.

REMEMBER

FastQC was designed as a general screening tool. Specialized libraries do not always behave like random whole-genome libraries.

What should I do next?

Match the pattern to a sensible next step.

LOW-QUALITY TAIL

Assess severity. Trim or filter only if it could affect downstream analysis.

ADAPTER SIGNAL

Confirm adapter or primer identity. Then trim the unwanted sequence.

OVERREPRESENTED SEQUENCE

Identify it. Then decide whether it is technical or biological.

DUPLICATION / GC / BASE BIAS

Interpret it against the library design. Investigate only when it is unexpected.

GOOD PRACTICE

Document why you trimmed, filtered, or chose to continue. Reproducibility includes the reasoning behind preprocessing decisions.

Compare three FastQC reports

Reuse the Part 4 files and add one adapter-rich teaching dataset.

Place all three files in your training folder and run:

```
fastqc fastqc_good.fastq.gz fastqc_tail_drop.fastq.gz \  
fastqc_adapter_rich.fastq.gz
```

Then compare these modules:

1

Per base sequence quality

Which file shows the quality drop toward the read end?

2

Adapter Content

Which file shows the strongest adapter signal?

3

Overrepresented sequences

Can you identify whether the repeated sequence is technical or biological?

TEACHING DATA

These synthetic datasets are designed to make specific QC patterns easy to see. Real sequencing data are usually more complex.

Your turn.

Interpret first. Decide on an action second.

- 1 A read has a quality drop only at the final bases. What would you inspect before trimming?
- 2 FastQC reports adapter content. What should you confirm before removing sequence?
- 3 An amplicon library has high duplication. Is that automatically a problem?
- 4 A GC plot has an unexpected second peak. What possibilities would you investigate?
- 5 Why is an all-green FastQC report not the goal of quality control?

TO REMEMBER

FastQC tells you where to look. Experimental design and biological context tell you what the warning means.

NEXT

Part 6: Cleaning reads - when should we trim or filter, and when should we leave them alone?