Quality Control (FastQC)

FastQC Metrics

- Per base sequence quality - quality drops at read ends
- Per sequence quality scores - overall read quality distribution
- Per base sequence content - nucleotide balance
- Sequence duplication levels - PCR duplicates
- Adapter content - leftover adapter sequences
- Overrepresented sequences - contamination check

Good Quality

- Phred score >30
- Balanced GC content
- Low duplication
- No adapter contamination

Poor Quality

- Phred score <20
- GC bias
- High duplication (>50%)
- Adapter sequences present

Common Tools: FastQC, MultiQC, Trimmomatic, Cutadapt

Quality Score Principles

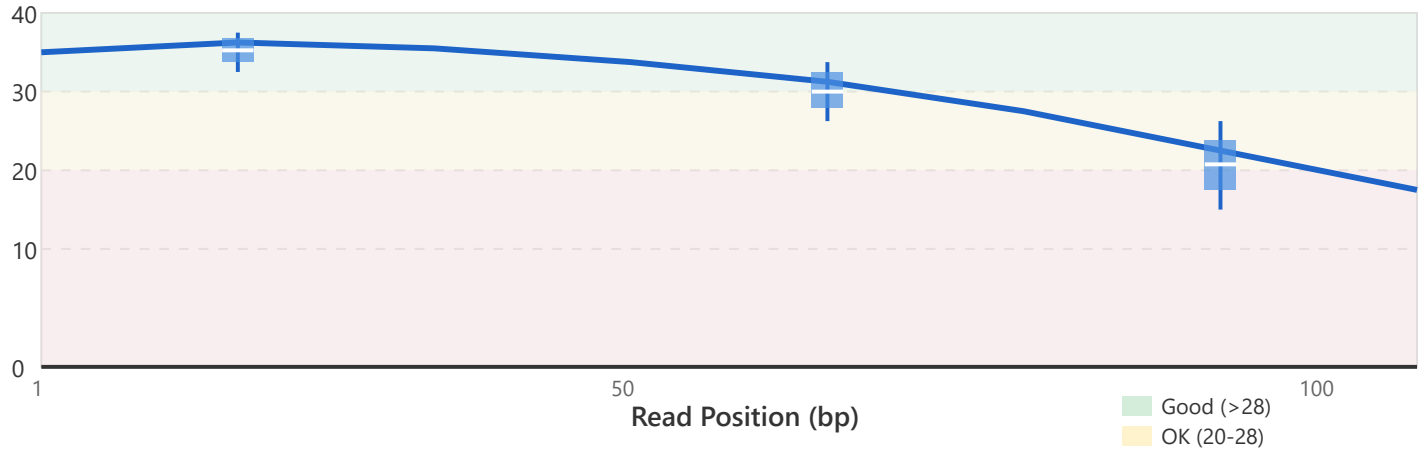
Phred Quality Score

$Q = -10 \times \log_{10}(P)$

where P is the probability of base calling error.

Phred 20	Phred 30	Phred 40
99% Accuracy	**99.9% Accuracy**	**99.99% Accuracy**
1/100 error rate	1/1,000 error rate	1/10,000 error rate

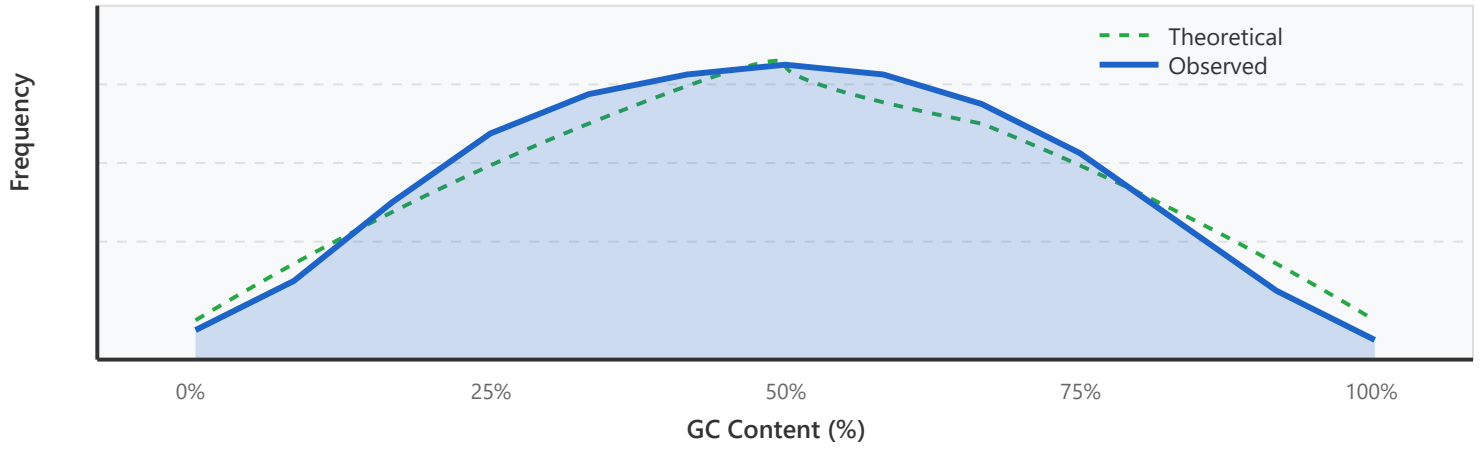
Quality Score Visualization Example



Typical pattern showing quality score decrease toward the end of reads

GC Content Analysis

GC Content Distribution



Normal samples show a normal distribution centered around the species' average GC content

Normal GC Distribution

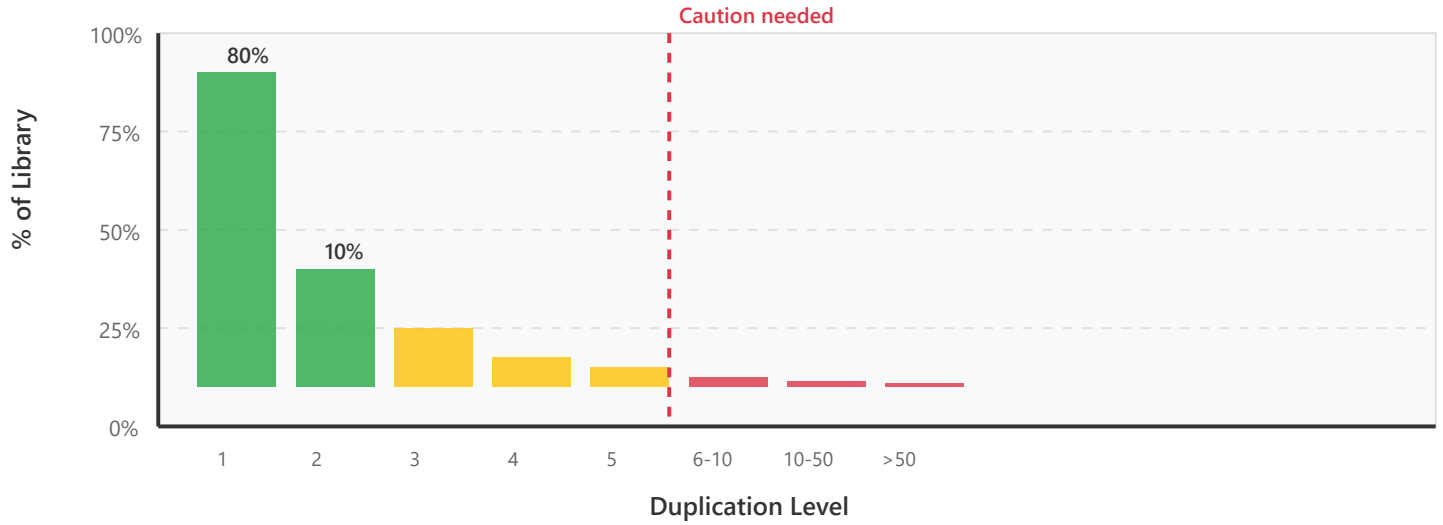
- Single peak
- Near expected GC%
- Narrow variance
- Matches theoretical distribution

Abnormal GC Distribution

- Multiple peaks (contamination)
- GC% bias
- Wide variance
- Deviates from theory

Sequence Duplication Analysis

Duplication Level Distribution



Good quality: most sequences are unique (single occurrence) / Poor quality: high duplication

- **PCR duplicates:**Caused by excessive PCR amplification during library preparation
- **Optical duplicates:**Caused by overly dense sequencing clusters
- **Biological duplicates:**Actually identical DNA fragments (highly expressed genes in RNA-seq)
- **Threshold:**Generally, >50% duplication indicates a problem

Adapter Contamination

Adapter Content Example

Read sequence:
ATCGATCGATCGATCGAGATCGGAAGAGC

Insert (original DNA):
ATCGATCGATCGATCG
+ Adapter sequence (needs removal)

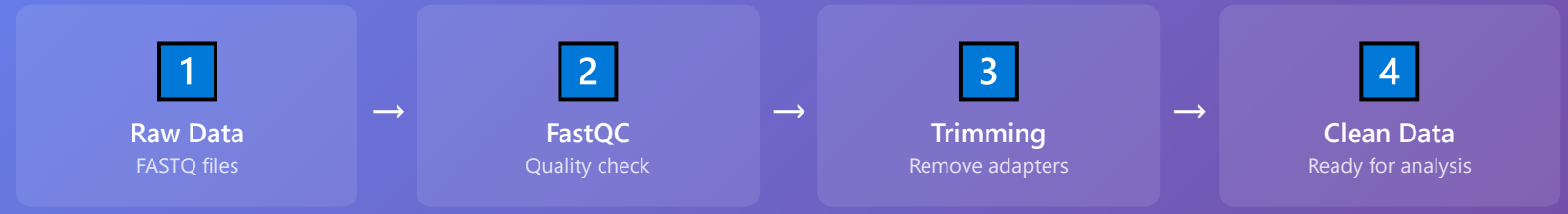
Why Adapters Remain

- When insert DNA is shorter than read length
- Sequencing reads beyond insert into adapter
- Common in small RNA-seq

Removal Methods

- Use Cutadapt, Trimmomatic
- Specify adapter sequences
- Set minimum length
- Perform quality trimming simultaneously

Quality Control Workflow



Essential first step for all NGS analysis - Quality control determines downstream analysis reliability