

# Can I trust this base?

## Understanding FASTQ quality scores

A wet-lab-to-dry-lab guide to what the fourth FASTQ line actually means.

@read\_demo

ATGCGTAC

+

IIII5?I+

### Sequence

Base calls

### Quality string

Confidence per base

### KEY IDEA

Quality score = estimated confidence in each base call

# From sequencing signal to confidence

## Why does FASTQ need a quality line?

A sequencer measures a physical signal. Software then decides which base is most likely and how confident that call is.

1

### Signal measured

The instrument records fluorescence, electrical signal, or another platform-specific measurement.

2

### Base called

Base-calling software converts the measured signal into A, C, G, or T.

3

### Confidence estimated

The software assigns an estimated probability that each base call could be wrong.

4

### Stored in FASTQ

The sequence is stored together with one quality character for every base.

### IMPORTANT

A quality score tells you how confident the base call is. It does not, by itself, tell you why the quality is high or low.

# What does a Phred score mean?

Higher Q means a lower estimated error probability

FASTQ quality values are usually expressed as Phred scores. The mathematics is useful, but the interpretation matters more.

Formula

$$Q = -10 \log_{10}(P_{\text{error}})$$

| Quality | Error probability | About 1 error in | Approx. accuracy |
|---------|-------------------|------------------|------------------|
| Q10     | 0.1               | 10 bases         | 90%              |
| Q20     | 0.01              | 100 bases        | 99%              |
| Q30     | 0.001             | 1,000 bases      | 99.9%            |
| Q40     | 0.0001            | 10,000 bases     | 99.99%           |

**DO NOT MISREAD Q30**  
Q30 does not mean "30% quality." It means an estimated error probability of about 1 in 1,000 bases.

# Why are there letters and symbols?

## Modern FASTQ commonly uses Phred+33 encoding

The numerical Phred score is stored as a printable character. The character is an encoding, not a nucleotide.

+

**Q10**

about 1 in 10 incorrect

**90% accuracy**

5

**Q20**

about 1 in 100 incorrect

**99% accuracy**

?

**Q30**

about 1 in 1,000 incorrect

**99.9% accuracy**

I

**Q40**

about 1 in 10,000 incorrect

**99.99% accuracy**

### YOU DO NOT NEED TO MEMORIZE ASCII CODES

For now, remember the concept: each quality character maps to a numerical Phred score. We will let software do the decoding for real datasets.

# Decode one FASTQ read

## One quality character for every base

This eight-base read has eight quality characters. Each position can therefore be interpreted separately.

| Position     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Sequence     | A   | T   | G   | C   | G   | T   | A   | C   |
| Quality char | I   | I   | I   | I   | 5   | ?   | I   | +   |
| Phred score  | Q40 | Q40 | Q40 | Q40 | Q20 | Q30 | Q40 | Q10 |

**QUESTION** Which base has the lowest confidence?

**ANSWER** Position 8: C has "+" = Q10, the lowest score in this example.

# Why should the wet lab care?

## Quality scores connect the run back to the experiment

Low quality can arise for different reasons. The FASTQ score shows where confidence is low, but not the exact cause.

### Input material

- Degraded or low-quality nucleic acid can affect the quality of the library entering sequencing.

### Library preparation

- Suboptimal or uneven libraries can contribute to poorer or less consistent reads.

### Sequencing run

- Loading, chemistry, instrument conditions, and run performance can influence read quality.

### Read position / platform

- Quality patterns differ by sequencing technology and can change across a read.

### Base calling

- Software converts the measured signal into bases and confidence estimates.

### KEY POINT

Low quality is a symptom, not a diagnosis. Investigating the cause requires looking at the sample, library, run, and platform context.

# Create a mixed-quality FASTQ read

## No new software is needed

Use the same Ubuntu workspace from Parts 1 and 2. We will create one tiny read and inspect it.

### 1. Move into your workspace

```
cd ~/bioinformatics_training
```

### 2. Create quality\_demo.fastq

```
cat > quality_demo.fastq << 'EOF'  
@read_demo  
ATGCGTAC  
+  
IIII5?I+  
EOF
```

### 3. Inspect the file

```
cat quality_demo.fastq
```

#### YOUR TASK

Using the mapping from the previous slide, identify the lowest-quality base, the Q30 character, and why the sequence and quality strings must both contain eight characters.

# Your turn.

## Check that the quality-score idea has clicked

Try these without looking back at the previous slides.

1

What does a FASTQ quality score describe?

2

What does Q30 mean in plain language?

3

Which has lower confidence: Q10 or Q40?

4

In Phred+33, which character in our example represents Q20?

5

Why can a low quality score not identify the exact wet-lab cause by itself?

**NEXT** One read is easy. What about millions?

**Part 4: Assessing sequencing data with FastQC**