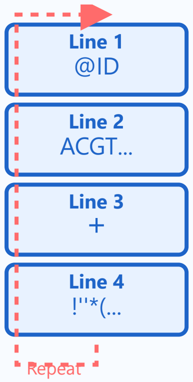
FASTQ Format

FASTQ File Structure



```
@SEQ_ID (Sequence identifier)
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+ (Separator)
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Line 1: @Identifier

Unique read ID with instrument and run information

Line 2: Sequence

Raw nucleotide sequence (A, T, C, G, N)

Line 3: +

Separator (sometimes repeats identifier)

Line 4: Quality Scores

Phred quality scores (ASCII encoded)

Phred Score: $Q = -10 \times \log_{10}(P)$ | Q30 = 99.9% accuracy, Q40 = 99.99%

Detailed Component Breakdown

Line 1: Sequence Identifier

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

Components:

- **@** - Indicates start of FASTQ record
- **HWUSI-EAS100R** - Instrument name
- **6** - Flow cell lane
- **73** - Tile number within lane
- **941** - X-coordinate on tile
- **1973** - Y-coordinate on tile
- **#0** - Index sequence (for multiplexing)
- **/1** - Read number (paired-end: /1 or /2)

Line 2: Nucleotide Sequence

```
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

Details:

- Raw base calls from sequencing instrument
- Standard nucleotides: **A** (Adenine), **T** (Thymine), **C** (Cytosine), **G** (Guanine)
- **N** represents ambiguous base call
- Length varies by sequencing platform (typically 50-300 bp)
- Read direction: 5' → 3'

Line 3: Separator Line

```
+
```

Purpose:

- Always begins with **+** symbol
- Separates sequence from quality scores
- May optionally repeat the identifier from Line 1
- Modern FASTQ files typically use just "+" for efficiency

Line 4: Quality Scores

```
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Encoding System:

- ASCII characters represent Phred quality scores
- Each character corresponds to one base in Line 2
- **Must be same length as sequence**

Quality Score Examples:

- **!** (ASCII 33) = Q0 = 0% accuracy
- ***** (ASCII 42) = Q9 = 87.4% accuracy
- **5** (ASCII 53) = Q20 = 99% accuracy
- **?** (ASCII 63) = Q30 = 99.9% accuracy
- **I** (ASCII 73) = Q40 = 99.99% accuracy

Calculation: Quality (Q) = ASCII value - 33

Phred Quality Score Visual Guide

ASCII Character to Quality Score Mapping

Low Quality (Q0-Q20)

Characters: ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4

Medium Quality (Q20-Q30)

Characters: 5 6 7 8 9 : ; < = >

High Quality (Q30+)

Characters: ? @ A B C D E F G H I J K

Phred Quality Score Formula

$Q = -10 \times \log_{10}(P)$ | P = 10^(-Q/10) | Where P = probability of incorrect base call

Complete FASTQ Record Example

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Interpretation:

- Sequence length: 36 bases
- Quality scores: Mostly "I" characters (Q40 = 99.99% accuracy)

- Last few bases show slightly lower quality: "9" (Q24 = 99.4%), "I" (Q40), "G" (Q38), "9" (Q24), "I" (Q40), "C" (Q34)
- Overall: High-quality read suitable for downstream analysis