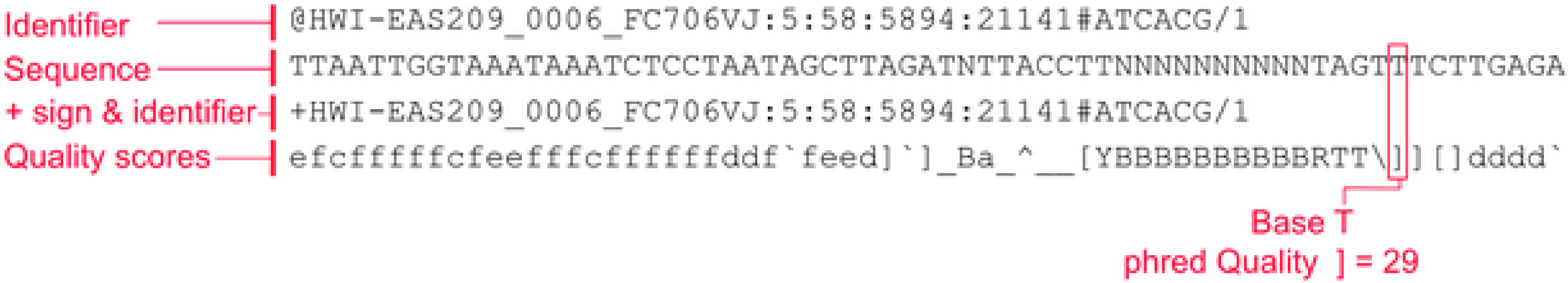
Raw sequences files and quality control

Fastq Format

This format is designed to handle base quality metrics output from sequencing machines.



```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`
```

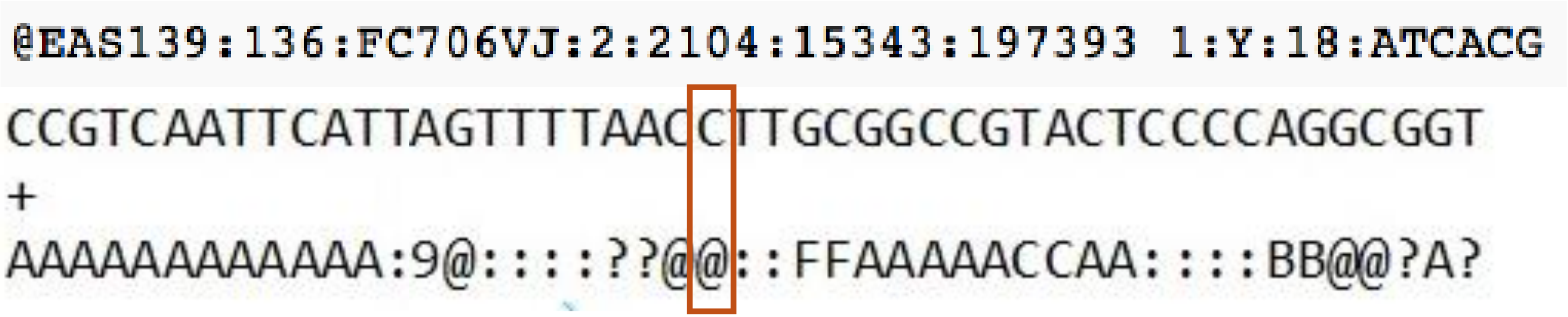
Line 1 begins with the '@' character and is followed by a sequence identifier and an optional description.

Line 2 is the sequence letters.

Line 3 begins with a '+' character; it marks the end of the sequence and is optionally followed by the same sequence identifier again in line 1.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Quality scores



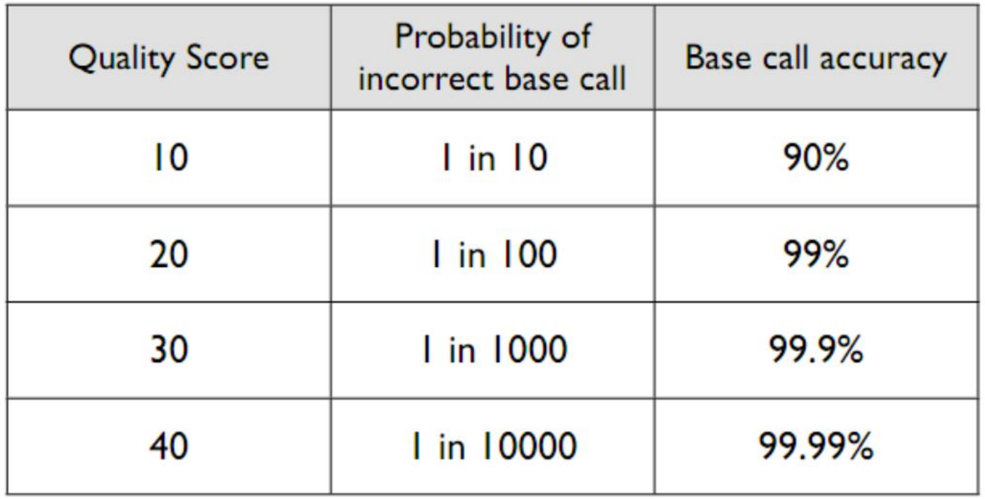
```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
CCGTCAATTCATTAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA:9@::::??@@::FFAAAAACCAA::::BB@@?A?
```

ASCII encoding

40:@	90:Z	141:a
41:A	91:[	142:b
42:B	92:\	143:c
43:C	93:]	144:d
44:D	94:^	145:e
45:E	95:_	146:f
… :…	… :…	… :…

$$\text{Phred} = -10 \log_{10} p$$

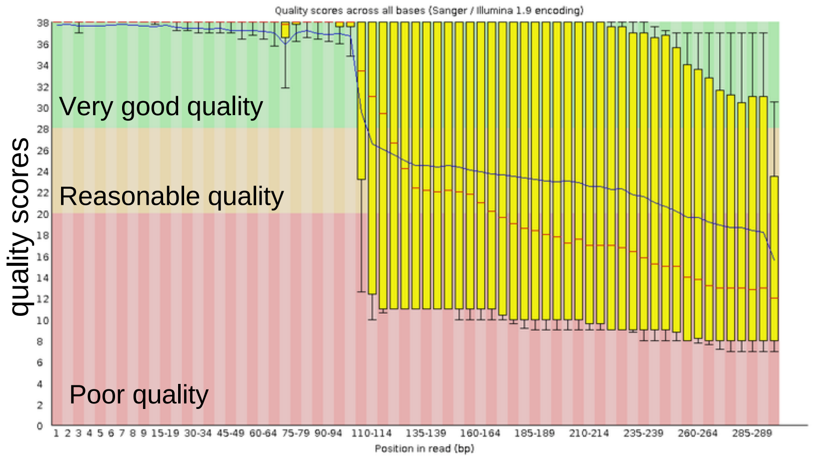
p = Probability call is incorrect



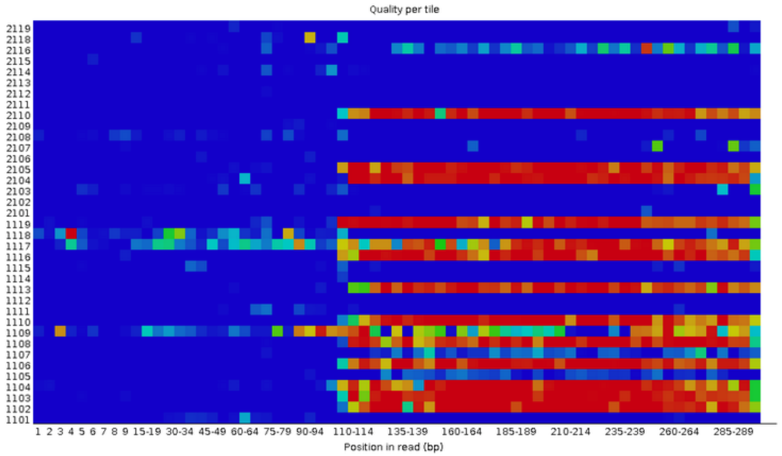
Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

Assess quality with FastQC

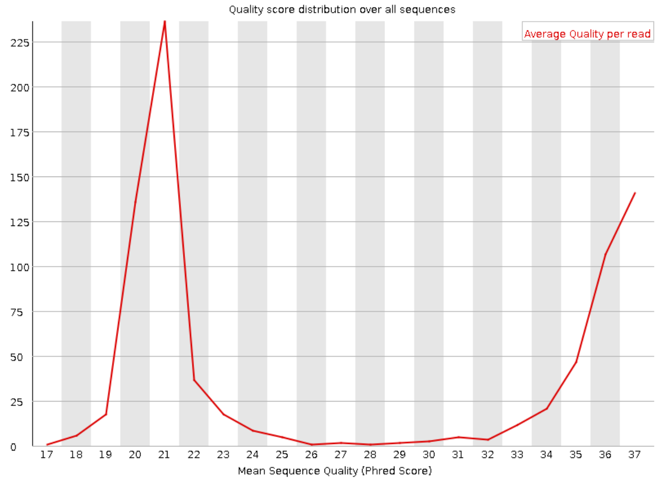
Per base sequence quality



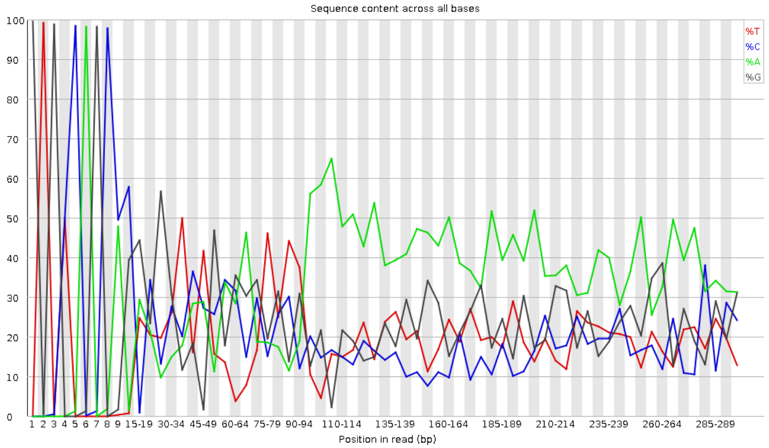
Per tile sequence quality



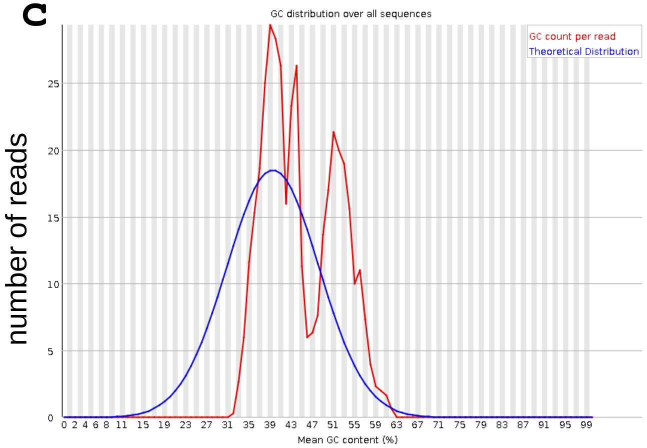
Per sequence quality scores



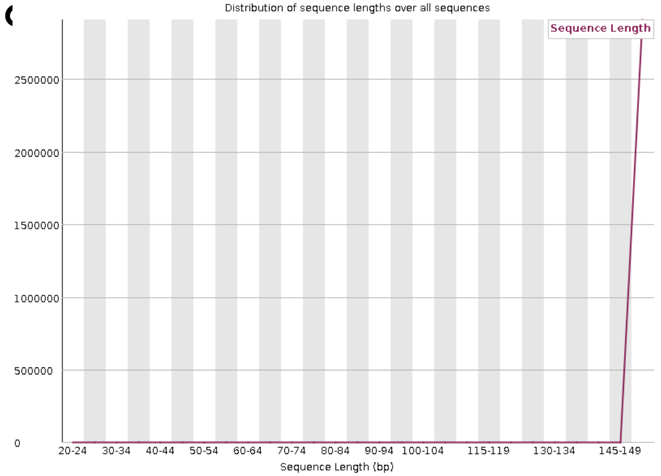
Per base sequence content



Per sequence GC c



Sequence length (



https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#sequence-length-distribution

Let’s have a look at the first few sequences and check the sequencing quality with fastqc