

# Quality Control of NGS data

Markus Mayrhofer

# FastQ files



```
@HWUSI-EAS100R:6:73:941:1973#0/1  
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT  
+  
!''*( (( (**+) ) %%%++) (%%%) .1***-+*'' ) ) **55CCF>>>>>CCCCCCC65
```

1<sup>st</sup> row: sequence identifier (machine ID, x-y coordinates, additional info)

2<sup>nd</sup> row: The actual sequence

3<sup>rd</sup> row: starts with “+” and optionally the same identifier as in the 1<sup>st</sup> row

4<sup>th</sup> row: Quality score for each base in read



# Phred Quality Scores

```
+SEQ_ID
```

```
! ' ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * *
```

A quality value  $Q$  is an integer representation of the probability  $p$  that the corresponding base call is incorrect.

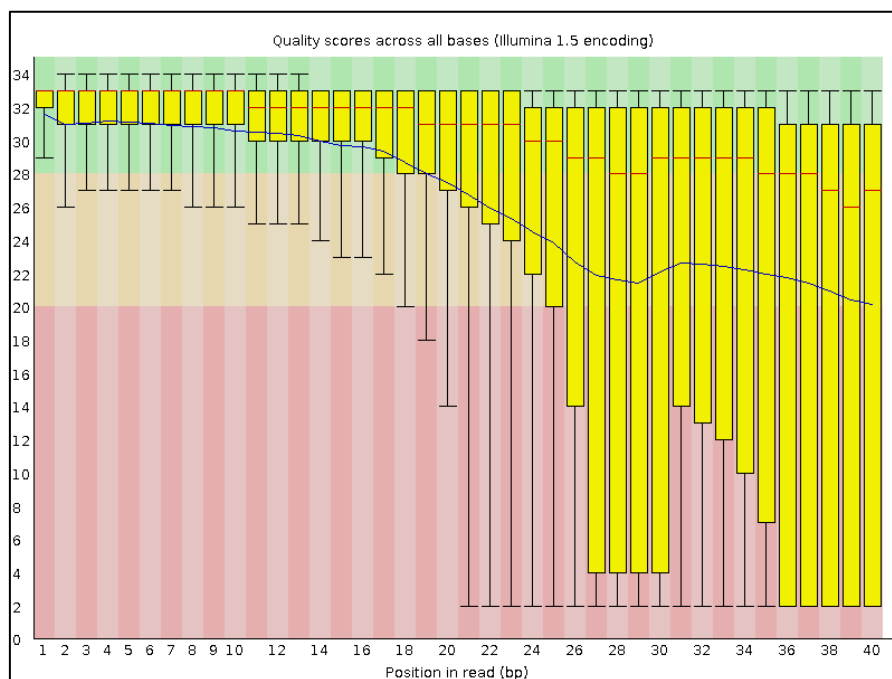
$$Q = -10 \log_{10} P \quad \longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred 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%
50	1 in 100000	99.999%

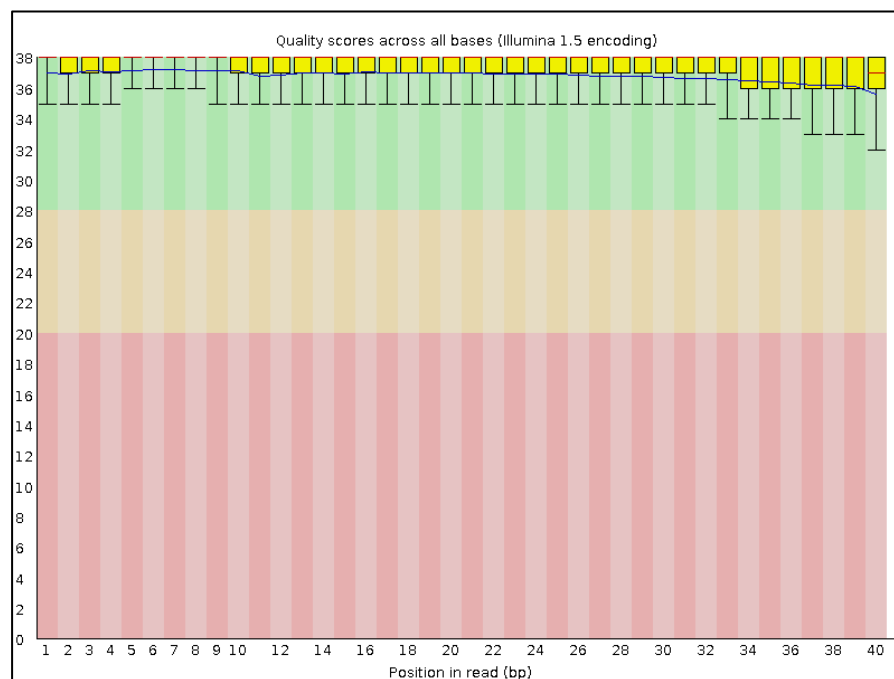


## FastQC

Bad qualities:



Good qualities:

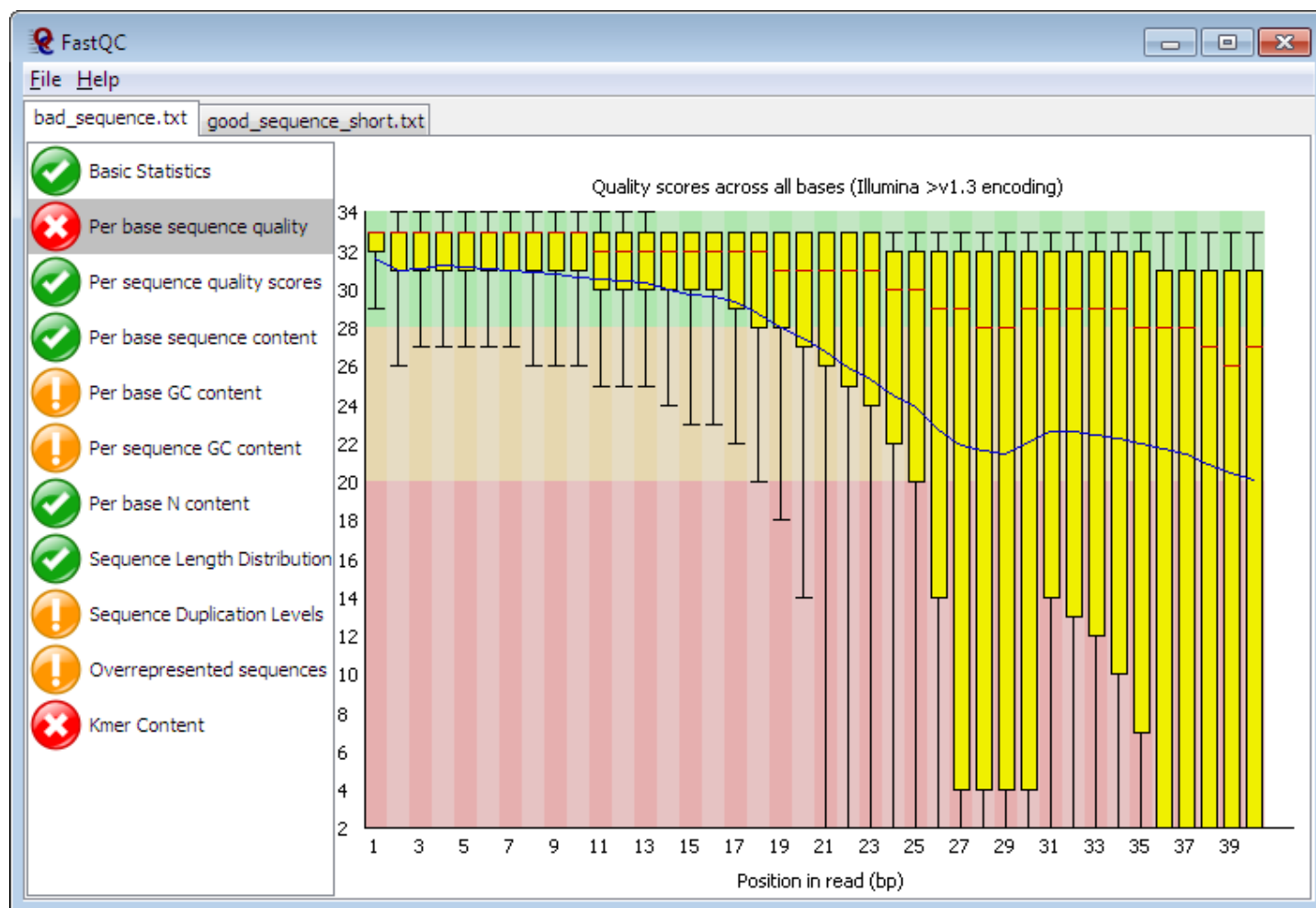


# What is QC?



- Different NGS application have their own problem areas and requires their own QC strategy
- Today: Focus on QC for whole genome sequencing
- For variant calling it is important to look at quality score distribution, sequence length distribution and duplication levels.
- Thursday: More details on QC for RNA-seq

# FastQC



# FastQC link



<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>