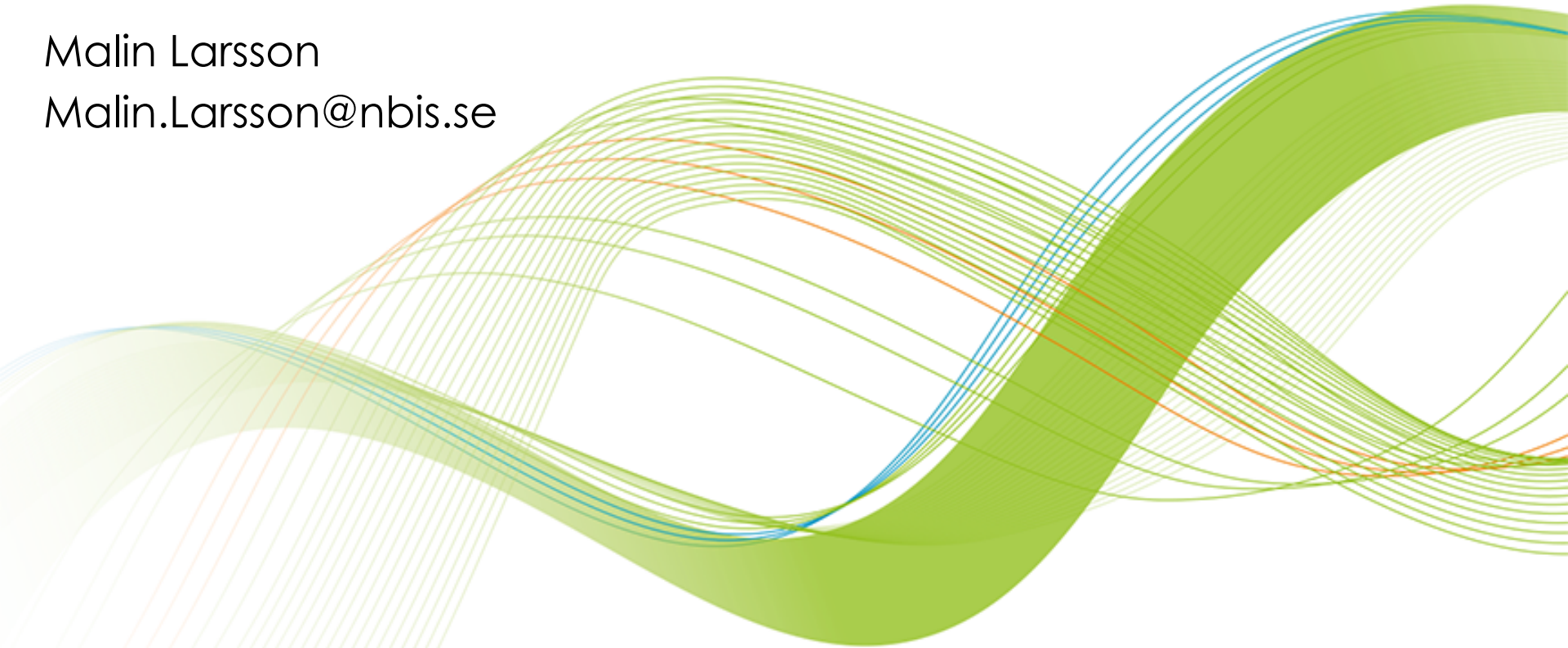
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Quality Control of NGS data

Malin Larsson
Malin.Larsson@nbis.se



FastQ files

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

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

2nd row: The actual sequence

3rd row: starts with "+" and optionally the same identifier as in the 1st row

4th 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

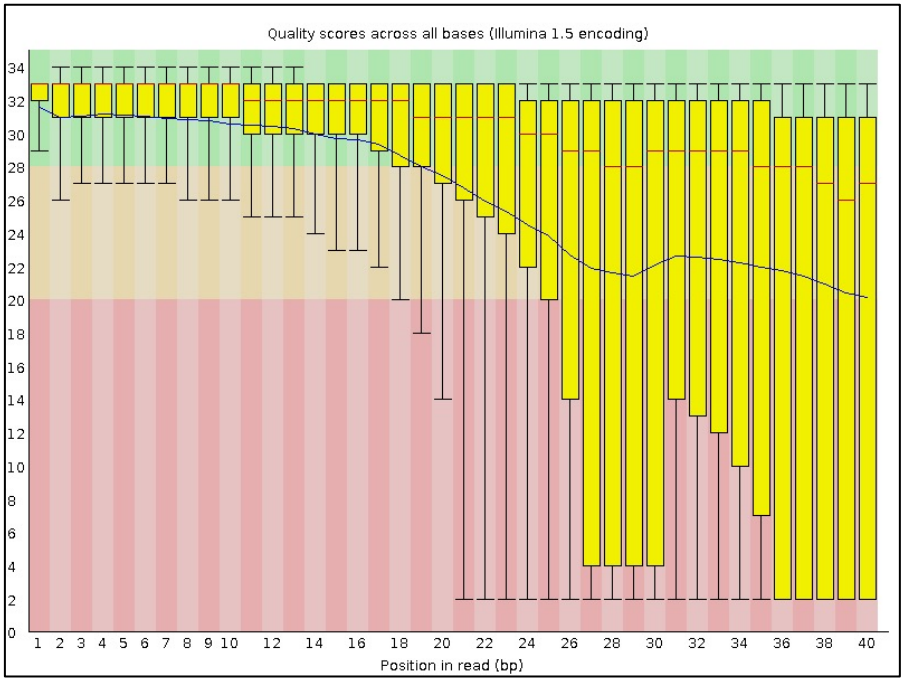
Babraham Institute

Babraham Bioinformatics

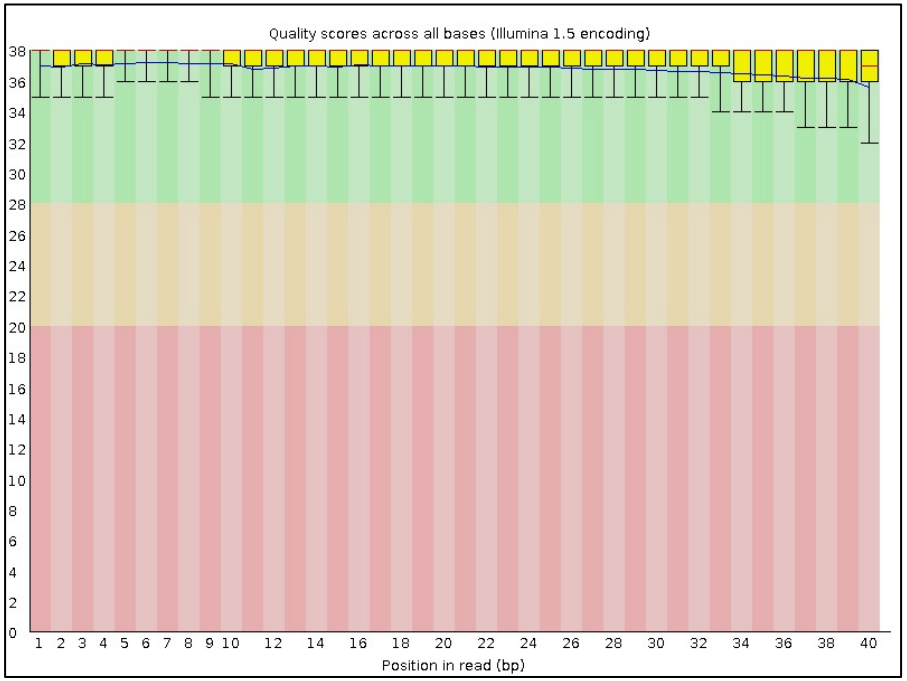
About | People | Services | Projects | Training | Publications

FastQC

Bad qualities:



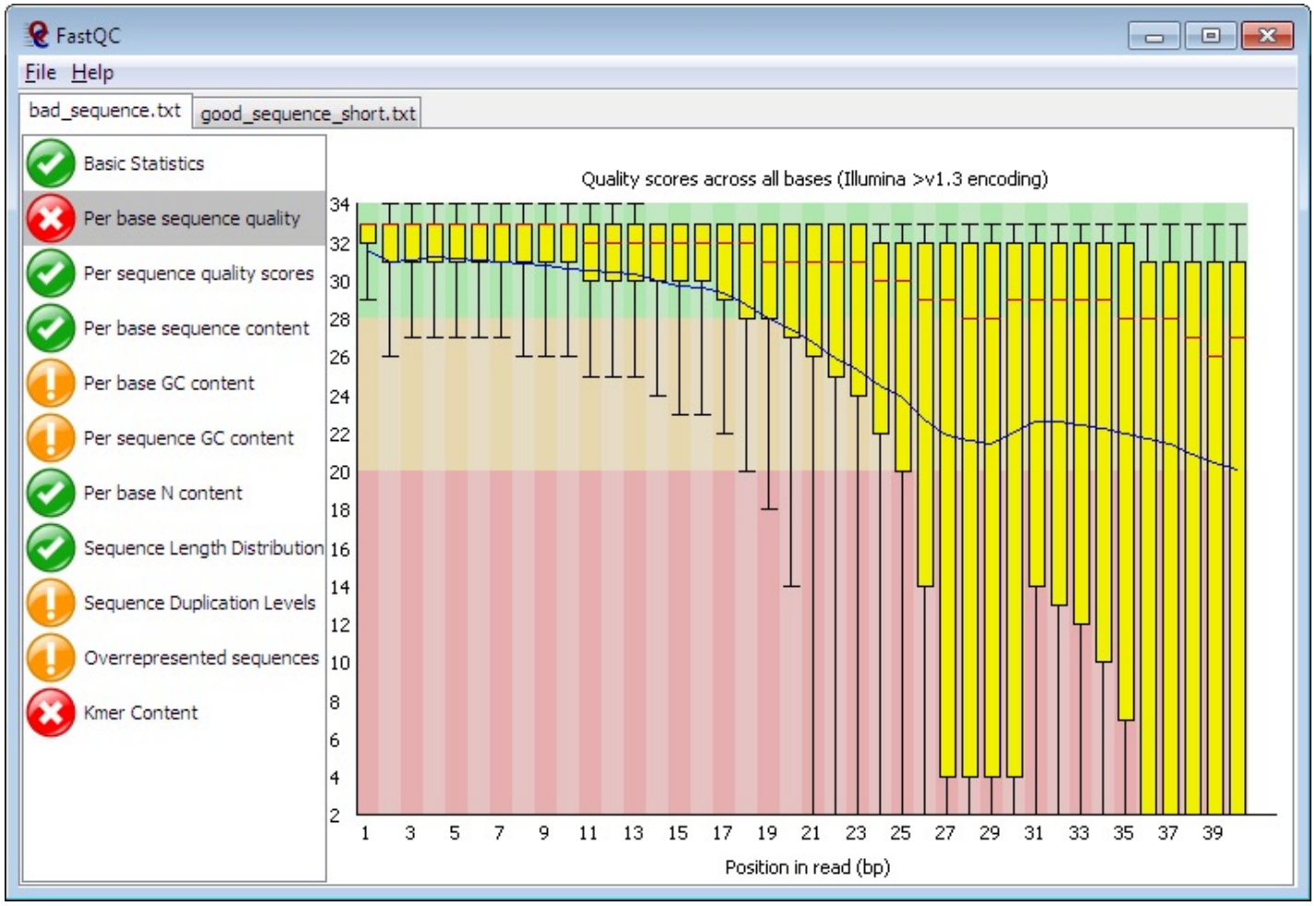
Good qualities:



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
File Help
bad_sequence.txt
good_sequence_short.txt
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

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