Workshop 6 Sequence quality

Introduction

Sequence quality

The fastq format was developed to provide a convenient way of storing the sequence and the quality scores in the same file. These are text files and they look like:

```
@seq_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
@seq_2
ATCGTAGTCTAGTCTATGCTAGTGCGATGCTAGTGCTAGTCGTATGCATGGCTATGTGTG
+
208DA8308AD8SF83FH0SD8F08APFIDJFN34JW830UDS8UFDSADPFIJ3N8DAA
```

In this file every sequence has 4 lines. In the first line we get the sequence name after the symbol `@` and, optionally, the description. The second line has the sequence and the fourth line has the quality scores encoded as letters.

Quality coding (modified from wikipedia).

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
|                         |    |        |                              |
33                        59   64       73                           104

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina      Phred+64,  raw reads typically (0, 40)
```

Illumina Q Score

The sequencing quality score of a given base, `Q`, is defined by the following equation:

```
Q = -10 x log10(e)
```

where `e` is the estimated probability of the base call being wrong.

Higher Q scores indicate a smaller probability of error. Lower Q scores can result in a significant portion of the reads being unusable. They may also lead to increased false-positive variant calls, resulting in inaccurate conclusions.

A quality score of **20** represents an error rate of **1 in 100**, with a corresponding call accuracy of **99%**.

Relationship Between Sequencing Quality Score and Base Call Accuracy

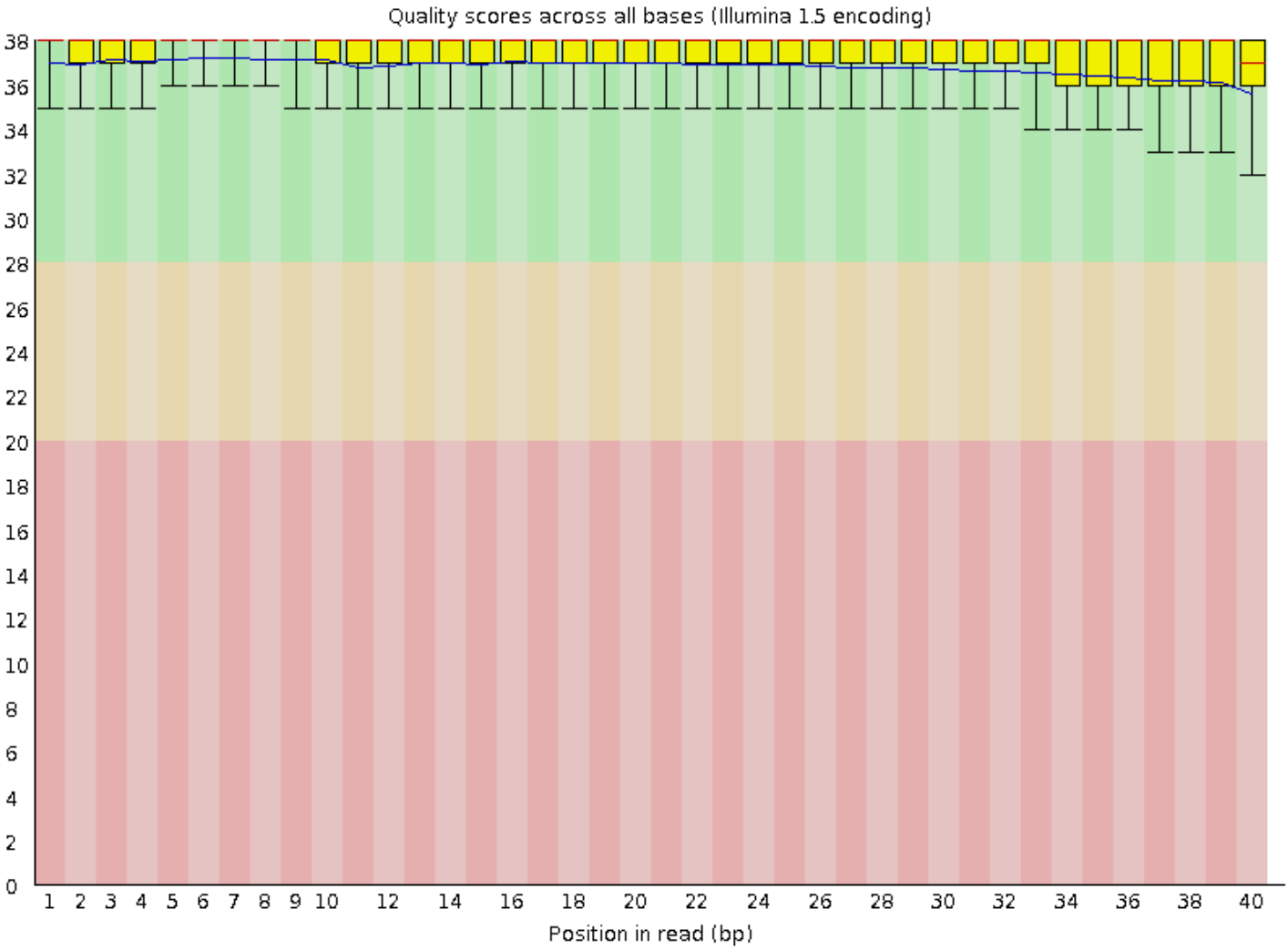
Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

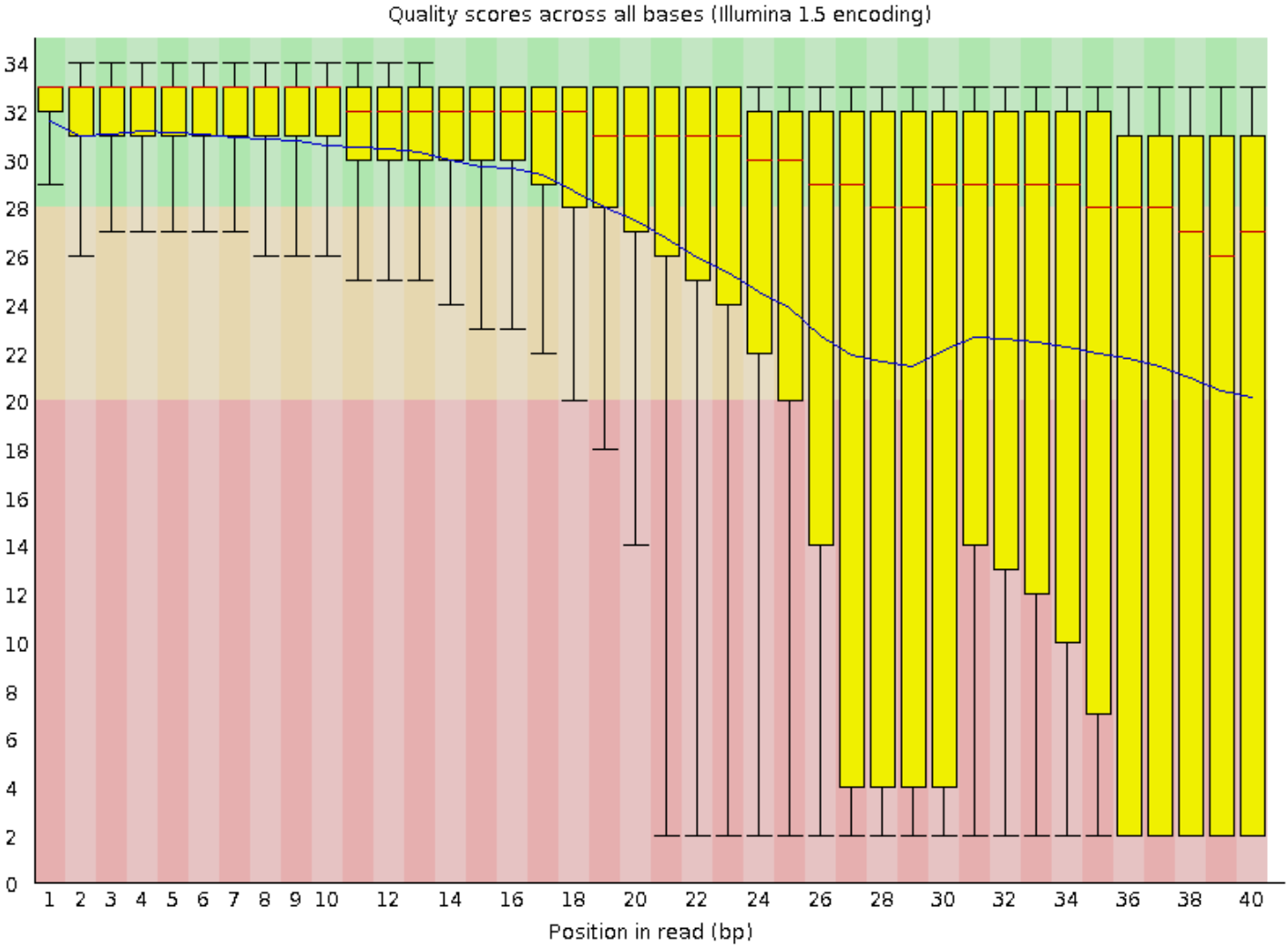
Reference: Illumina SBS Technology

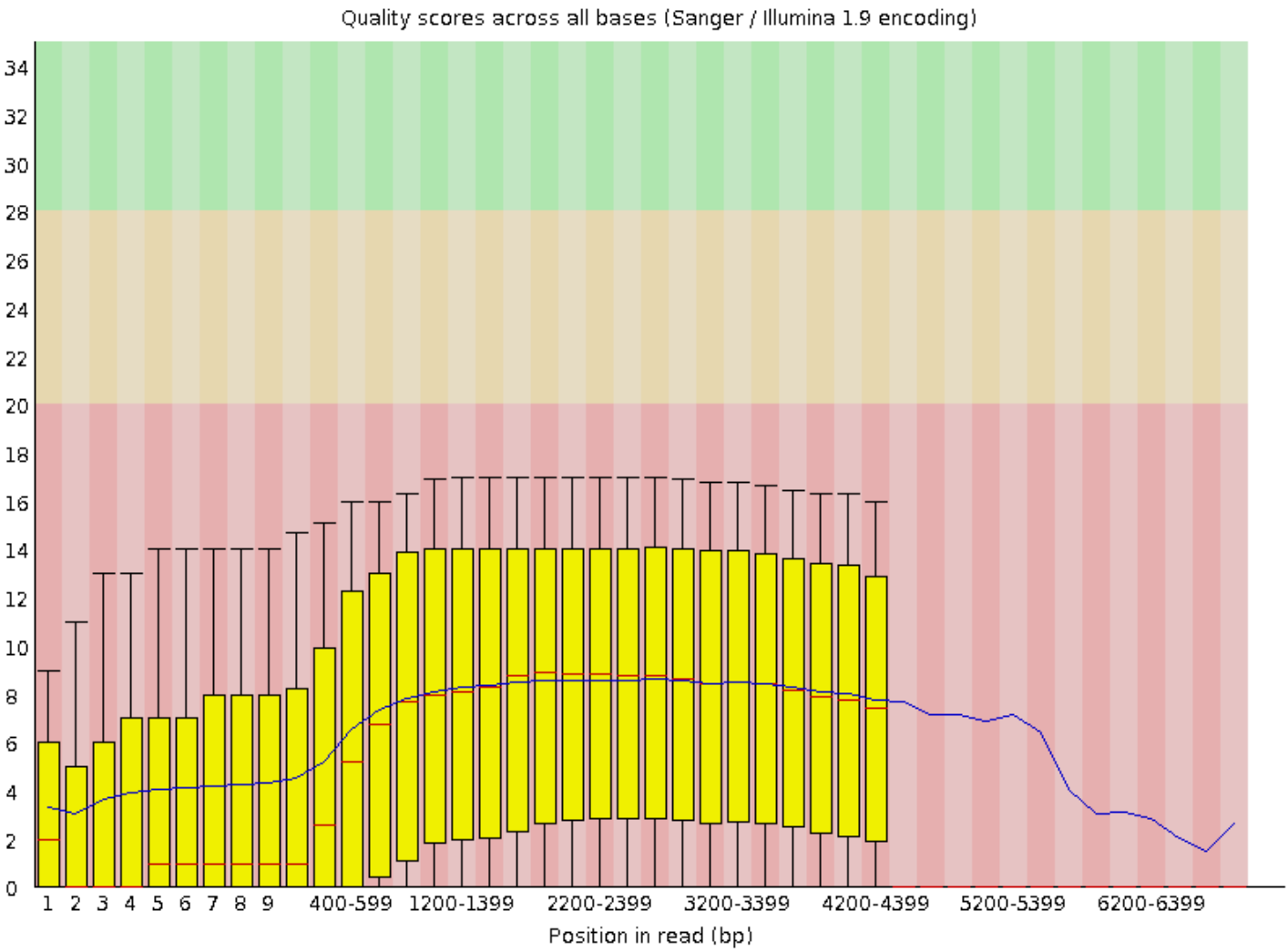
Testing quality

FASTQ

FastQC provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing platforms. It provides a graphic interface for local run (on your laptops) and command line interface for pipelines or remote computing.


Quality scores across all bases (Illumina 1.5 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)


Quality scores across all bases (Illumina 1.5 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)



Download fastqc

Try Fastq on the command line...

```
fastqc --outdir out /media/aquaexcel/<...>.fq.gz
```

Improving the quality of sequences

- Filtering of sequences
  - with small mean quality score
  - too small
  - with too many N bases
  - based on their GC content
  - ...
- Cutting/Trimming sequences
  - from low quality score parts
  - tails

    - ...

```
trim_galore -q 20 /media/aquaexcel/<...>.fq.gz
```

Key points

- Run quality control on every sequencing dataset before any other analyses
- Choose QC parameters carefully
- Re-run FastQC to check the impact of the quality control