NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Quality Control of NGS data

FASTQ files

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

Line 1: An “@” followed by a sequence identifier (machine ID, x-y coordinates, ...)
Line 2: The nucleotide sequence
Line 3: A “+” optionally followed by the sequence identifier again
Line 4: Encoded quality scores, one for each base in the read

- Bad nucleotides are sometimes written as N (“unknown”, “any”)
- Character “!” is quality zero
- If you see an uppercase character, the quality is ok

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%

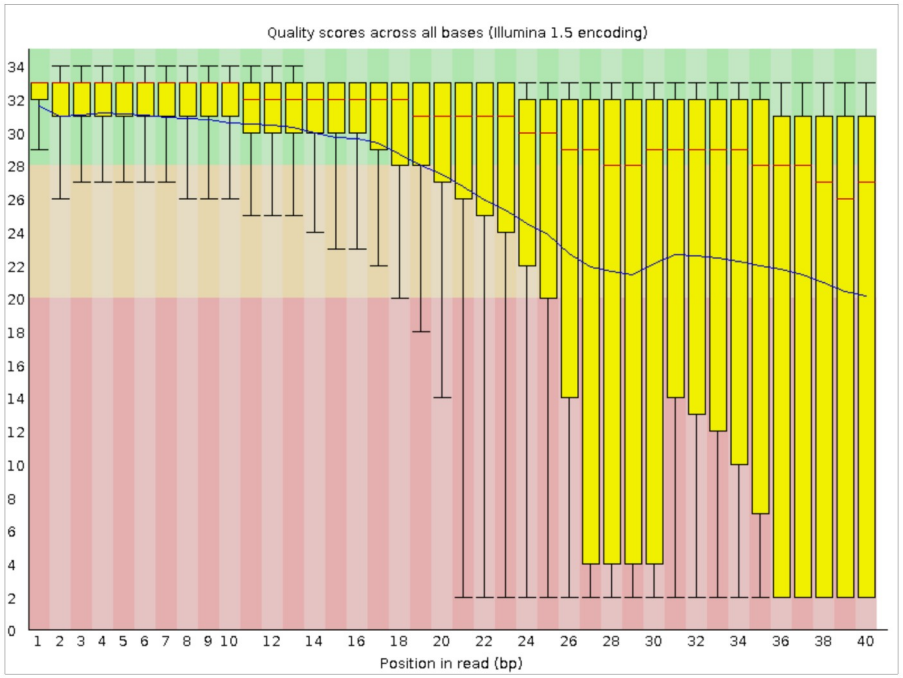
FastQC

Babraham Bioinformatics

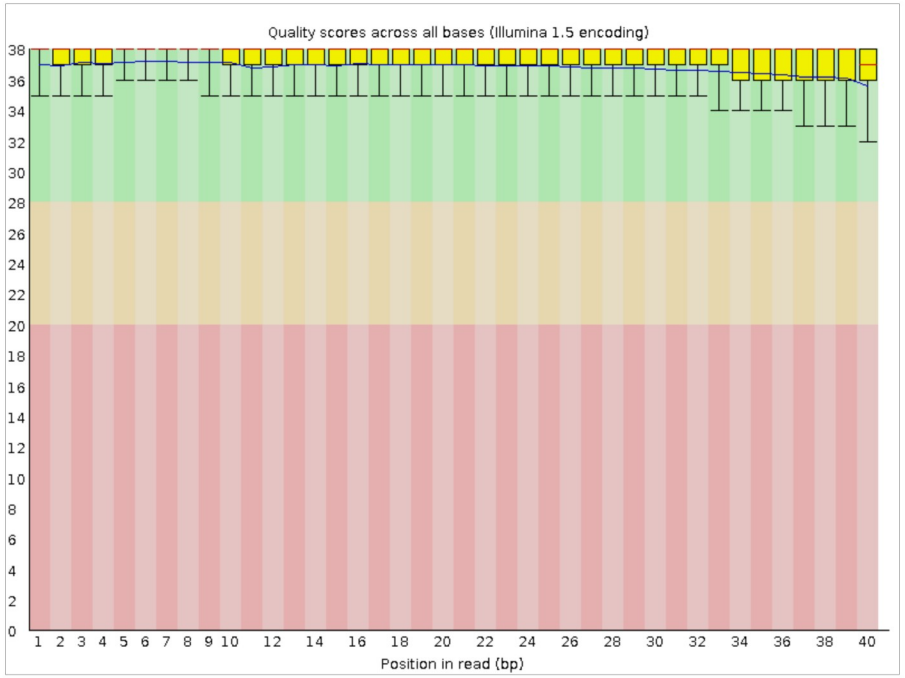
About | People | Services | Projects | Training | Publications

FastQC

Bad qualities:



Good qualities:



Quality trimming

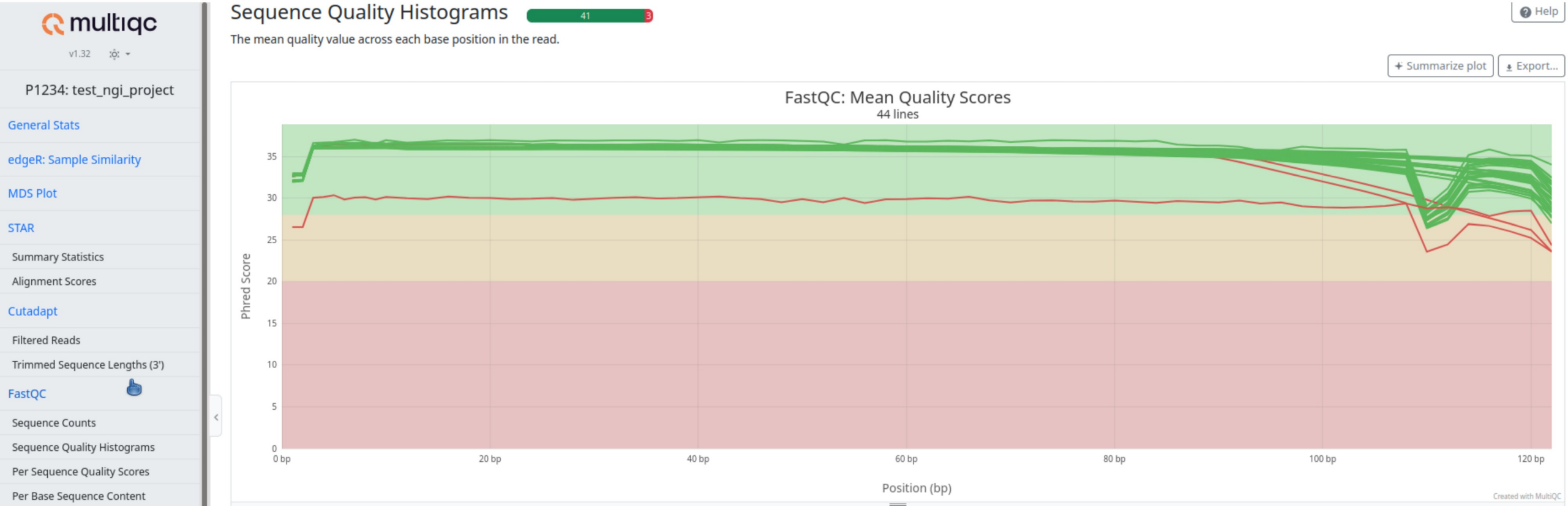
- Low-quality read ends may introduce spurious variant calls or other problems
- A tool to trim these is Cutadapt. For 3' trimming, run:

```
cutadapt -q 20 -o output.fastq.gz input.fastq.gz
```

- Removes ends with quality less than 20 (you choose)
- Often not needed because read aligners should handle it
- (Cutadapt can also remove adapters and primers)

MultiQC

- Gathers output from other bioinformatics tools and generates a single report for all samples
- Example NGI report:



What is QC?

- Different NGS application have their own problem areas and require 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

Links

- FastQC homepage
- Explanation of FastQC’s sequence quality plot
- MultiQC homepage
- MultiQC example report