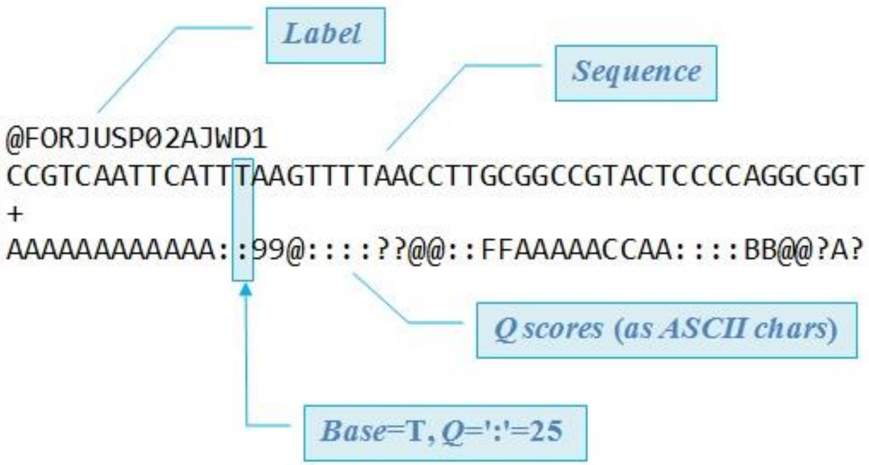
Preliminary analysis: quality check and trimming

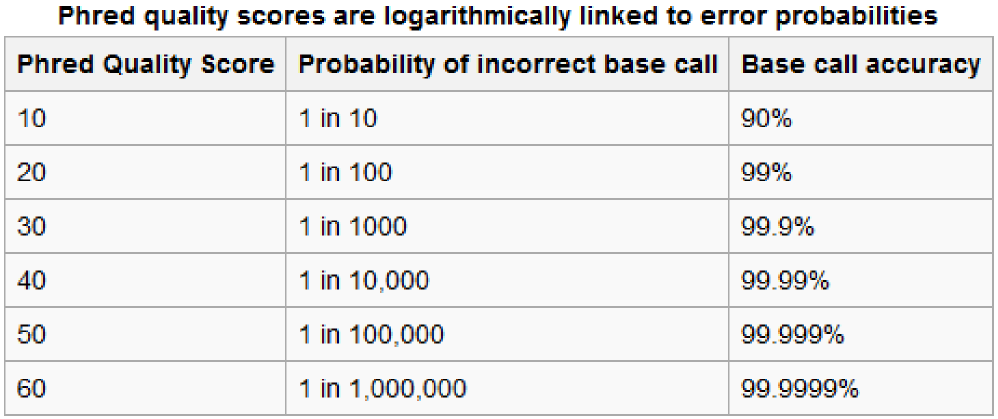
Valeria Michelacci

Bioinformatics training,
June 2018

What should be trimmed out?



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```



Phred quality scores are logarithmically linked to error probabilities

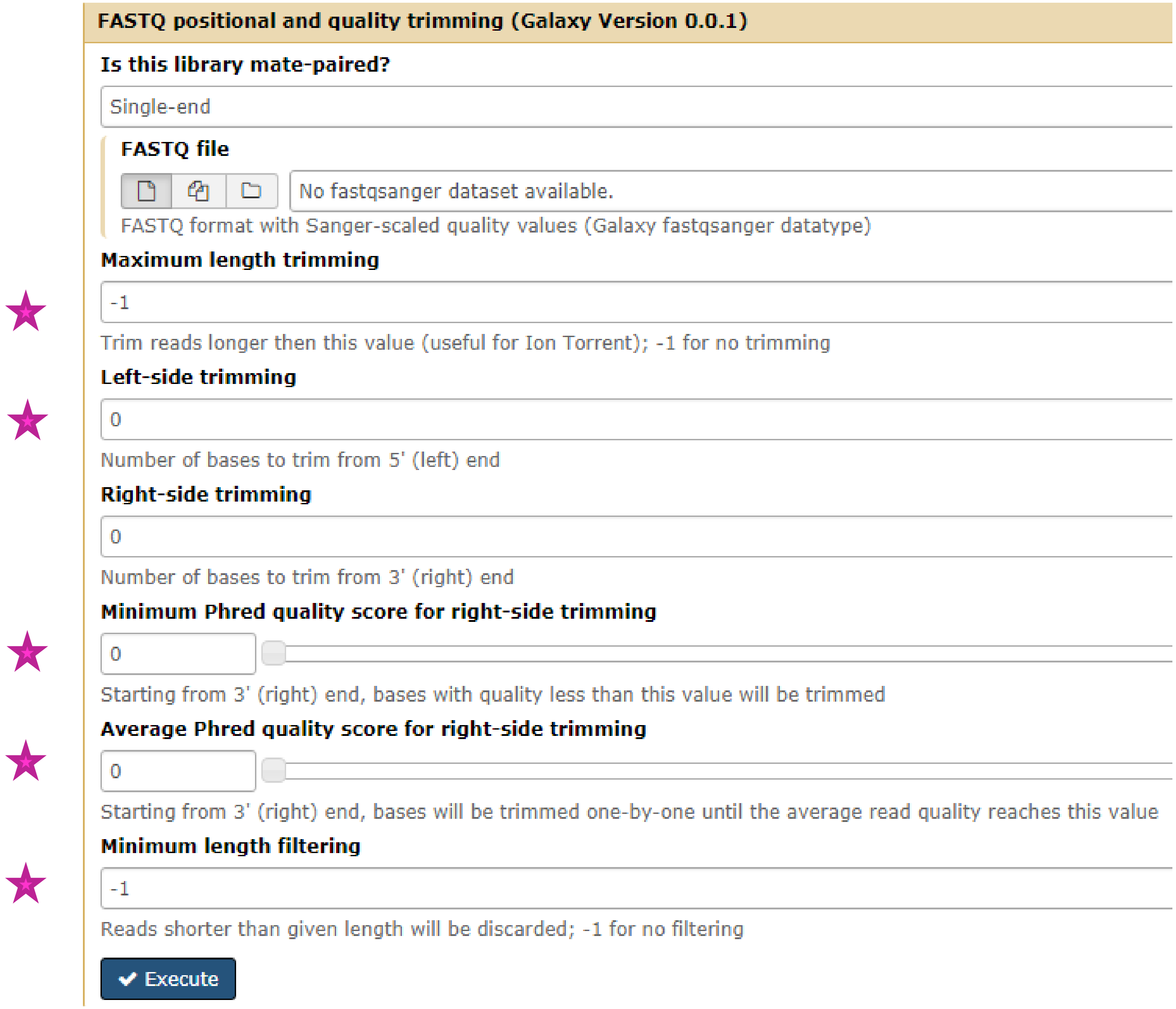
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Low quality positions

Adaptors and barcodes

Very short sequencing reads

What should be trimmed out?



FASTQ positional and quality trimming (Galaxy Version 0.0.1)

Is this library mate-paired?

Single-end

FASTQ file

No fastqsanger dataset available.

FASTQ format with Sanger-scaled quality values (Galaxy fastqsanger datatype)

★ Maximum length trimming

-1

Trim reads longer then this value (useful for Ion Torrent); -1 for no trimming

★ Left-side trimming

0

Number of bases to trim from 5' (left) end

Right-side trimming

0

Number of bases to trim from 3' (right) end

★ Minimum Phred quality score for right-side trimming

0

Starting from 3' (right) end, bases with quality less than this value will be trimmed

★ Average Phred quality score for right-side trimming

0

Starting from 3' (right) end, bases will be trimmed one-by-one until the average read quality reaches this value

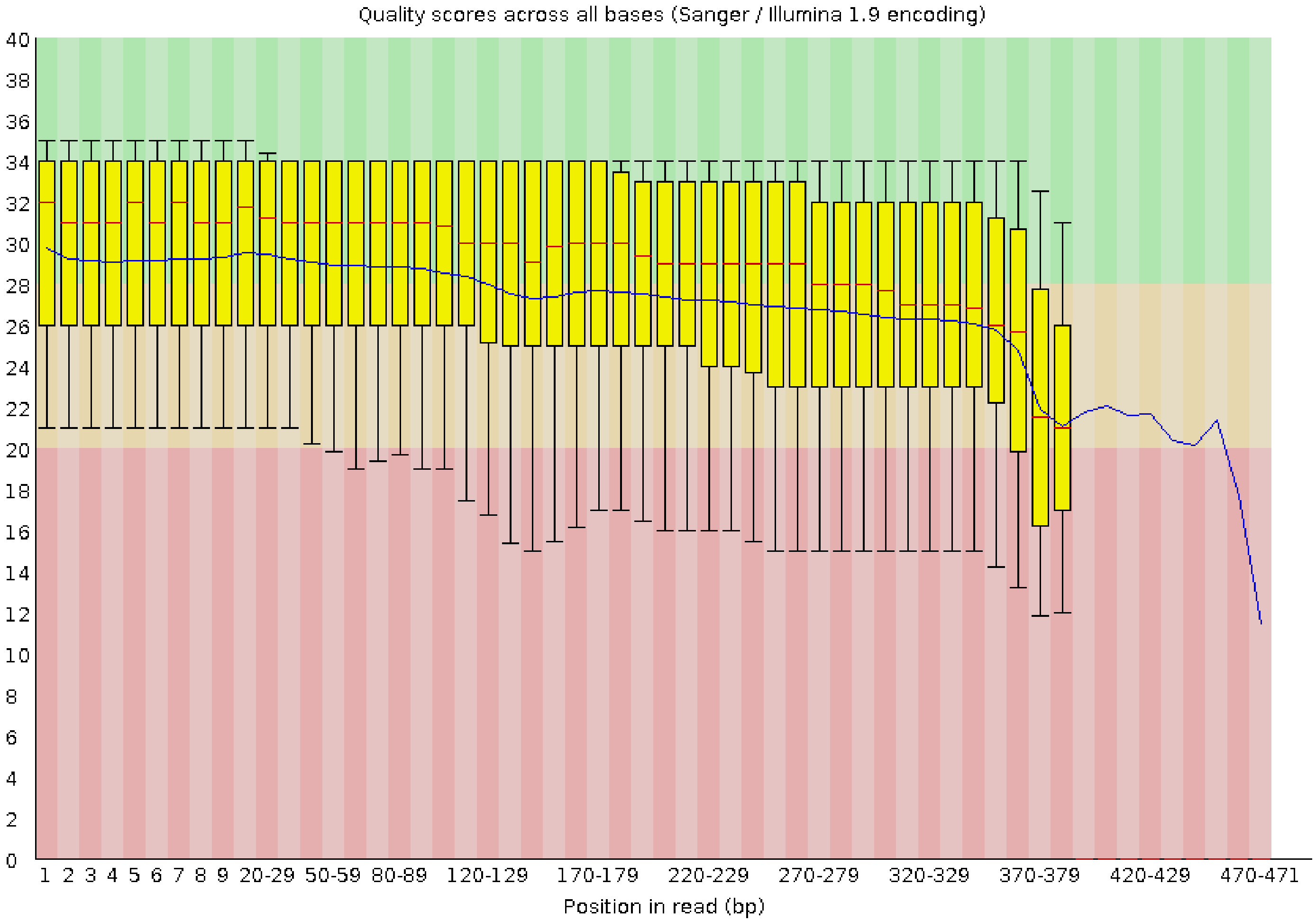
★ Minimum length filtering

-1

Reads shorter than given length will be discarded; -1 for no filtering

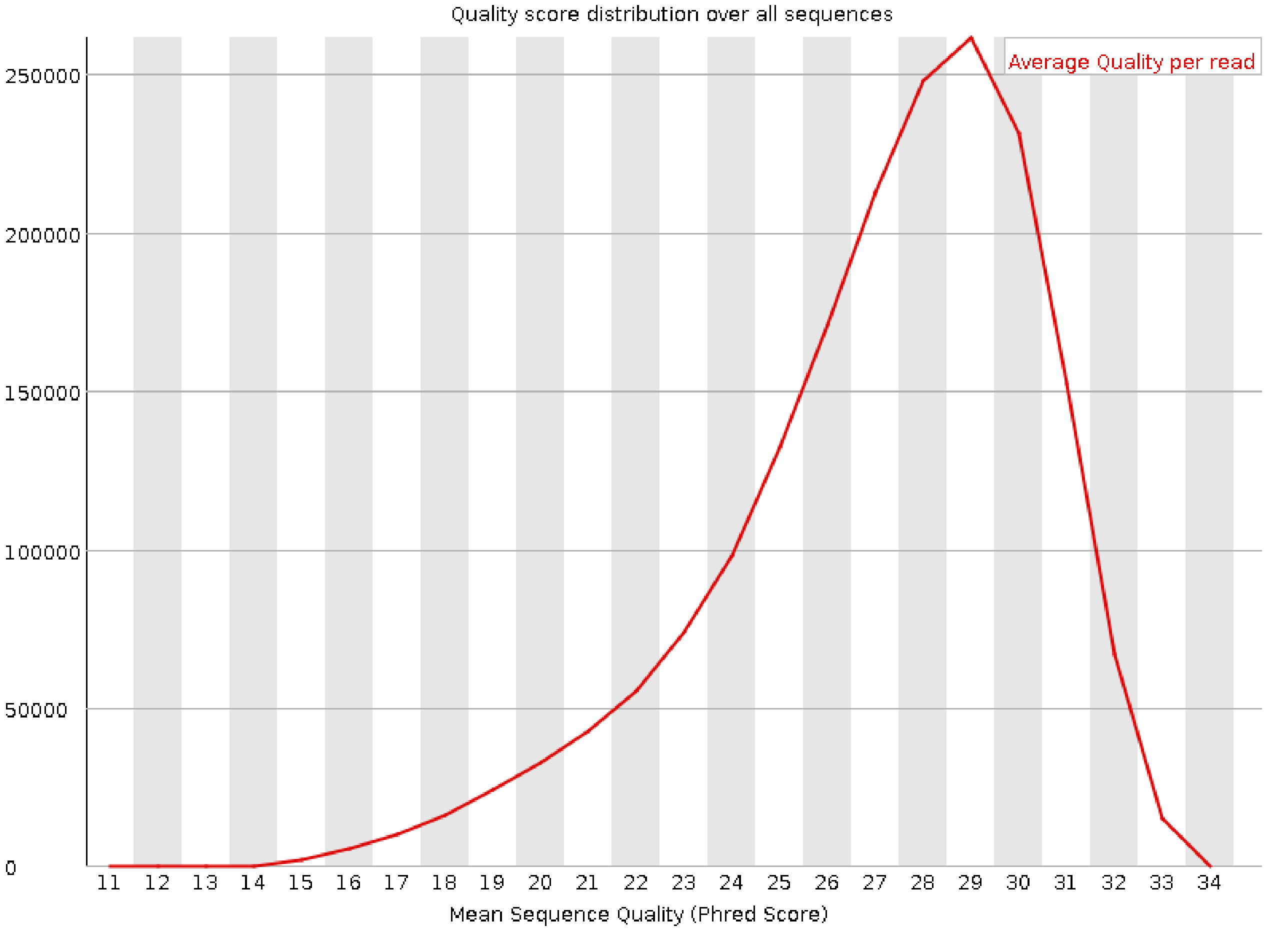
FastQC – quality check of aw data

Per base sequence quality

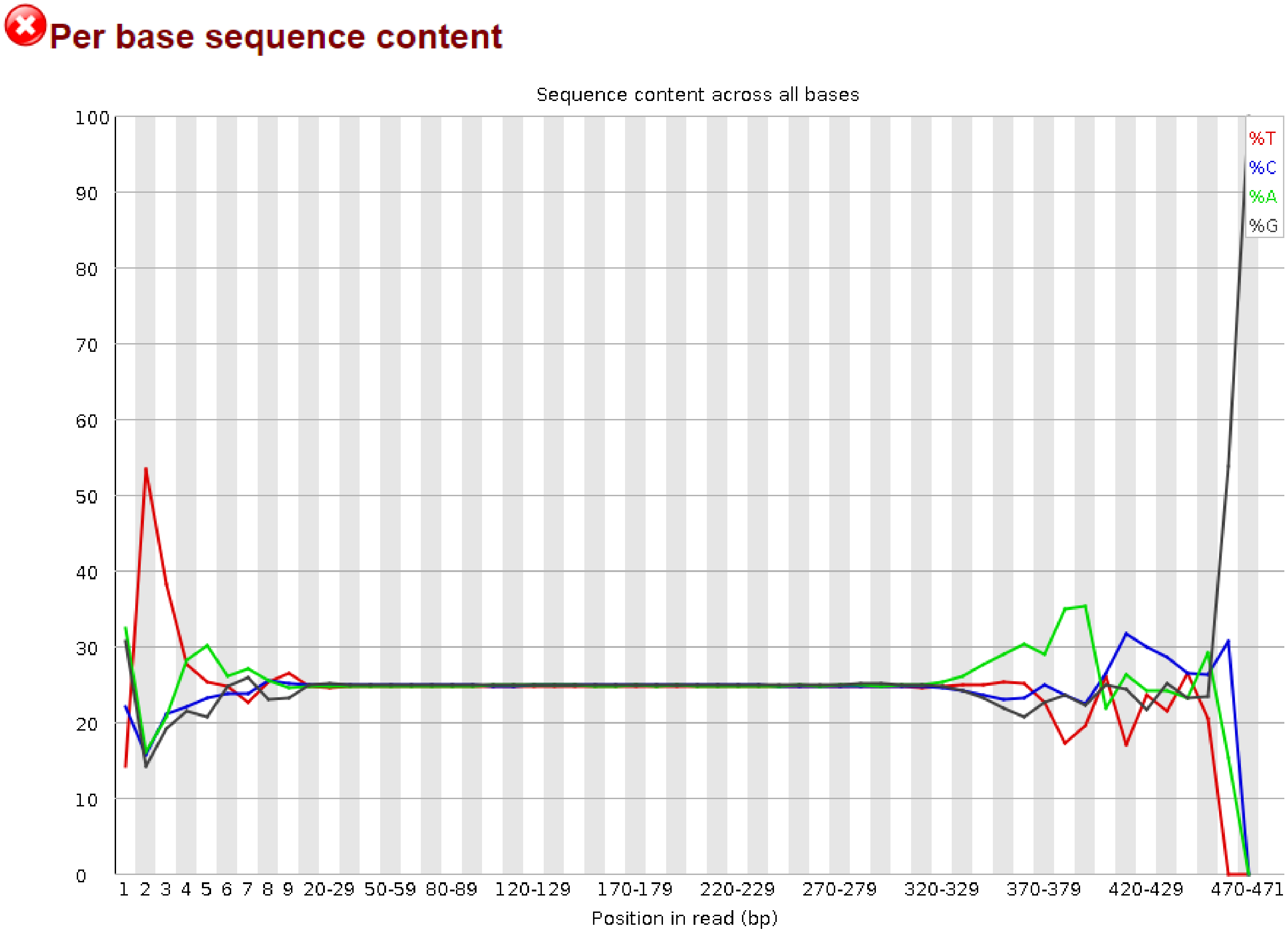


FastQC – quality check of aw data

✓ Per sequence quality scores

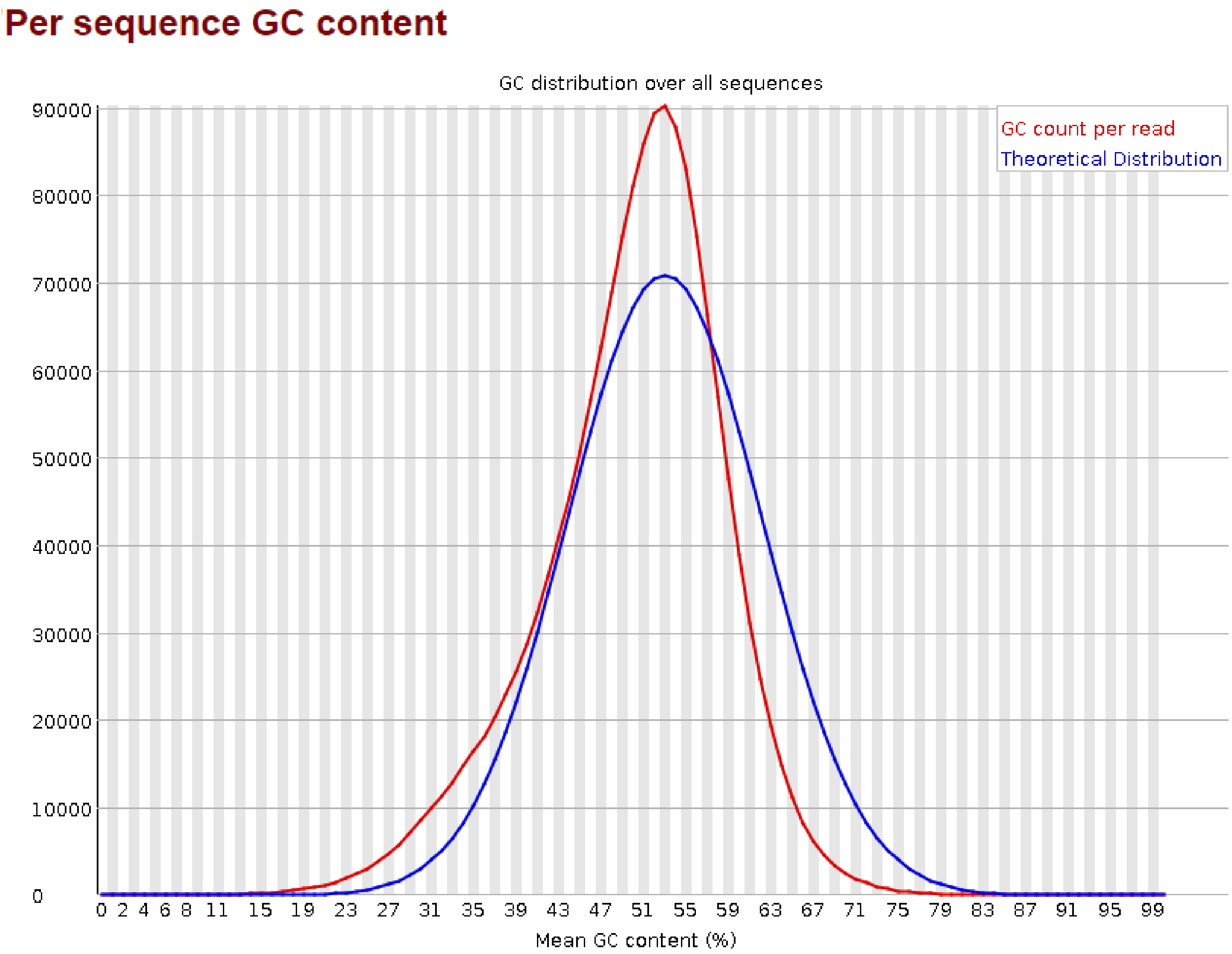


FastQC – quality check of aw data



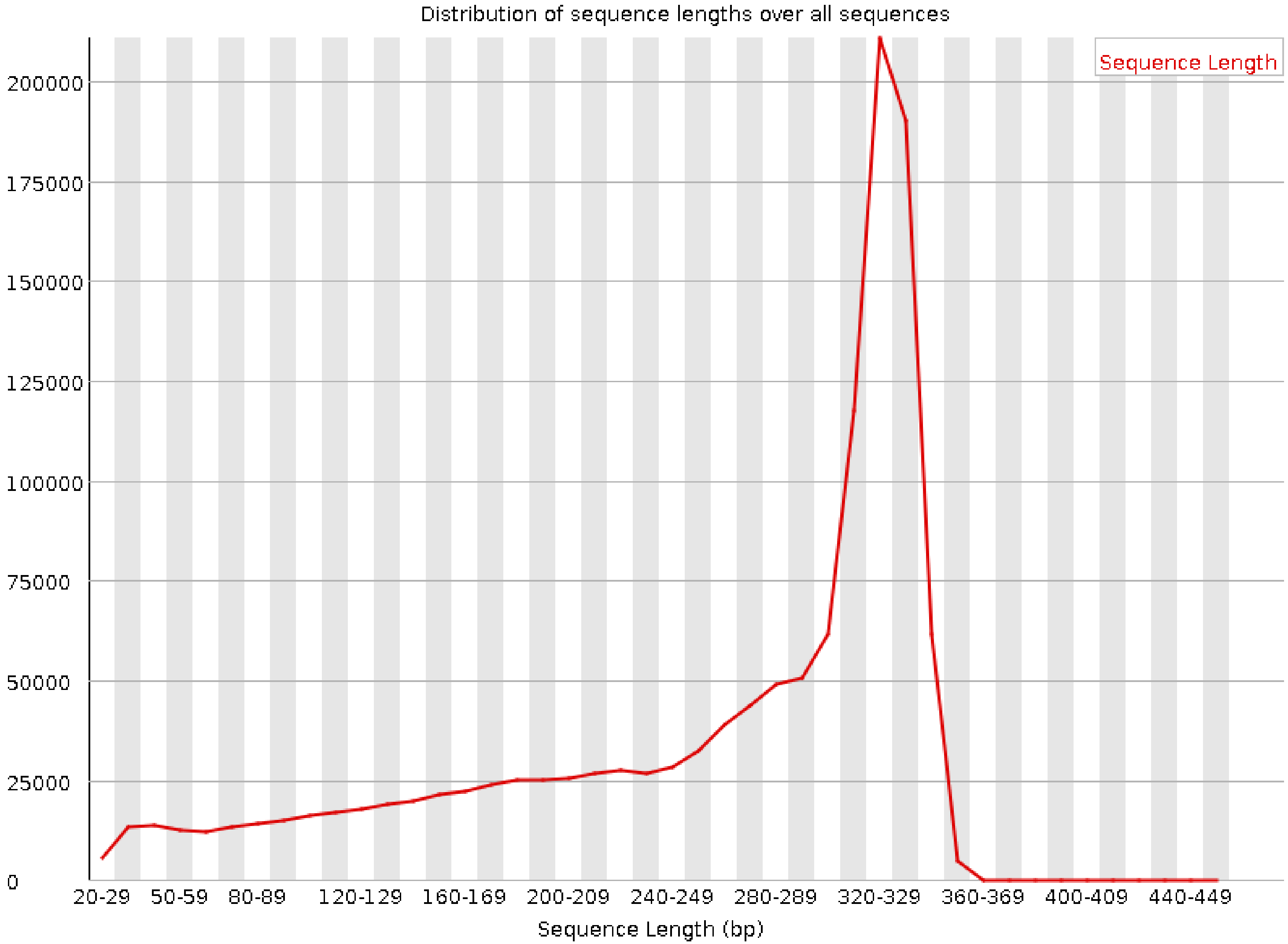
Per base sequence content

FastQC – quality check of aw data



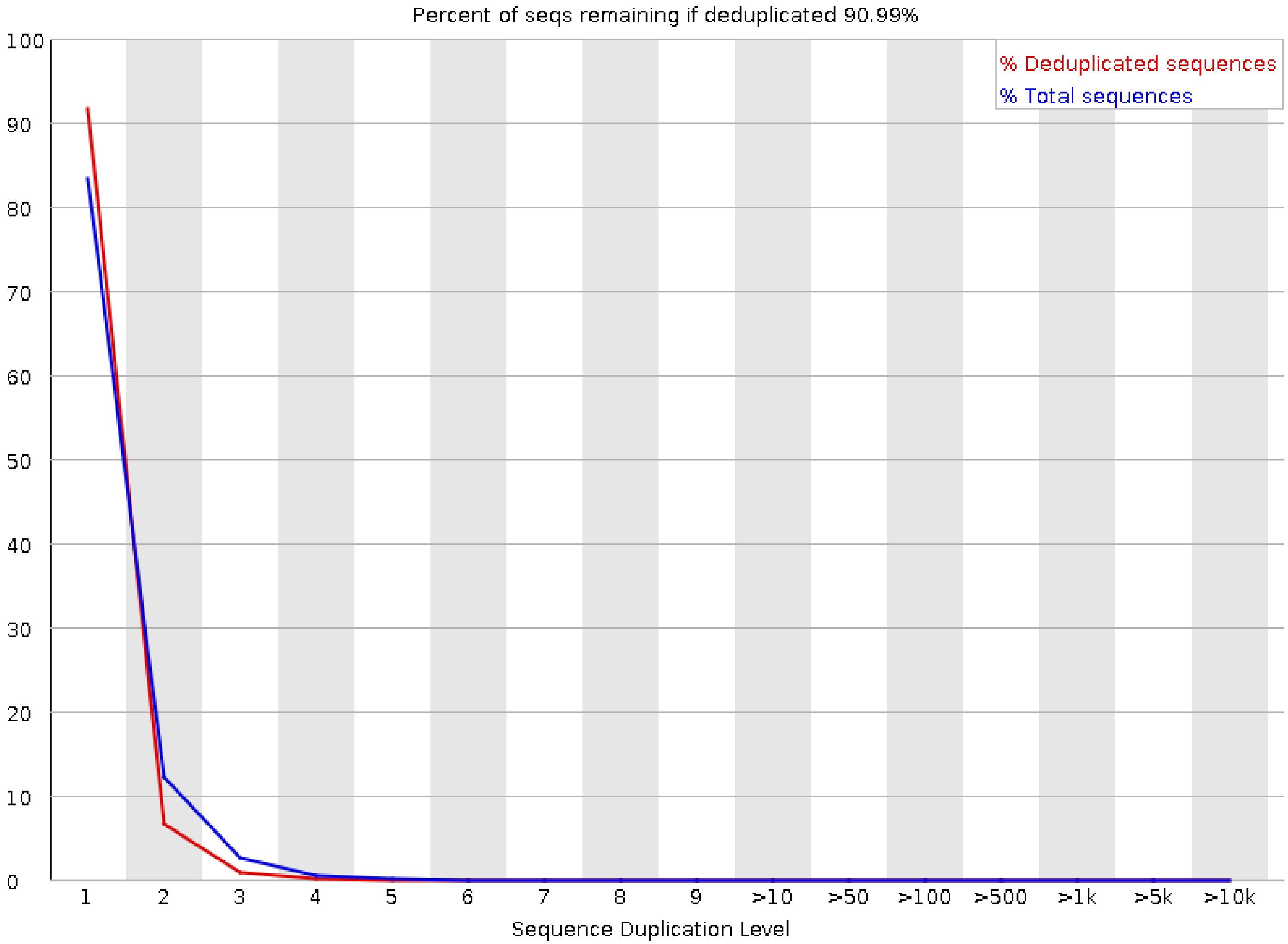
FastQC – quality check of aw data

Sequence Length Distribution



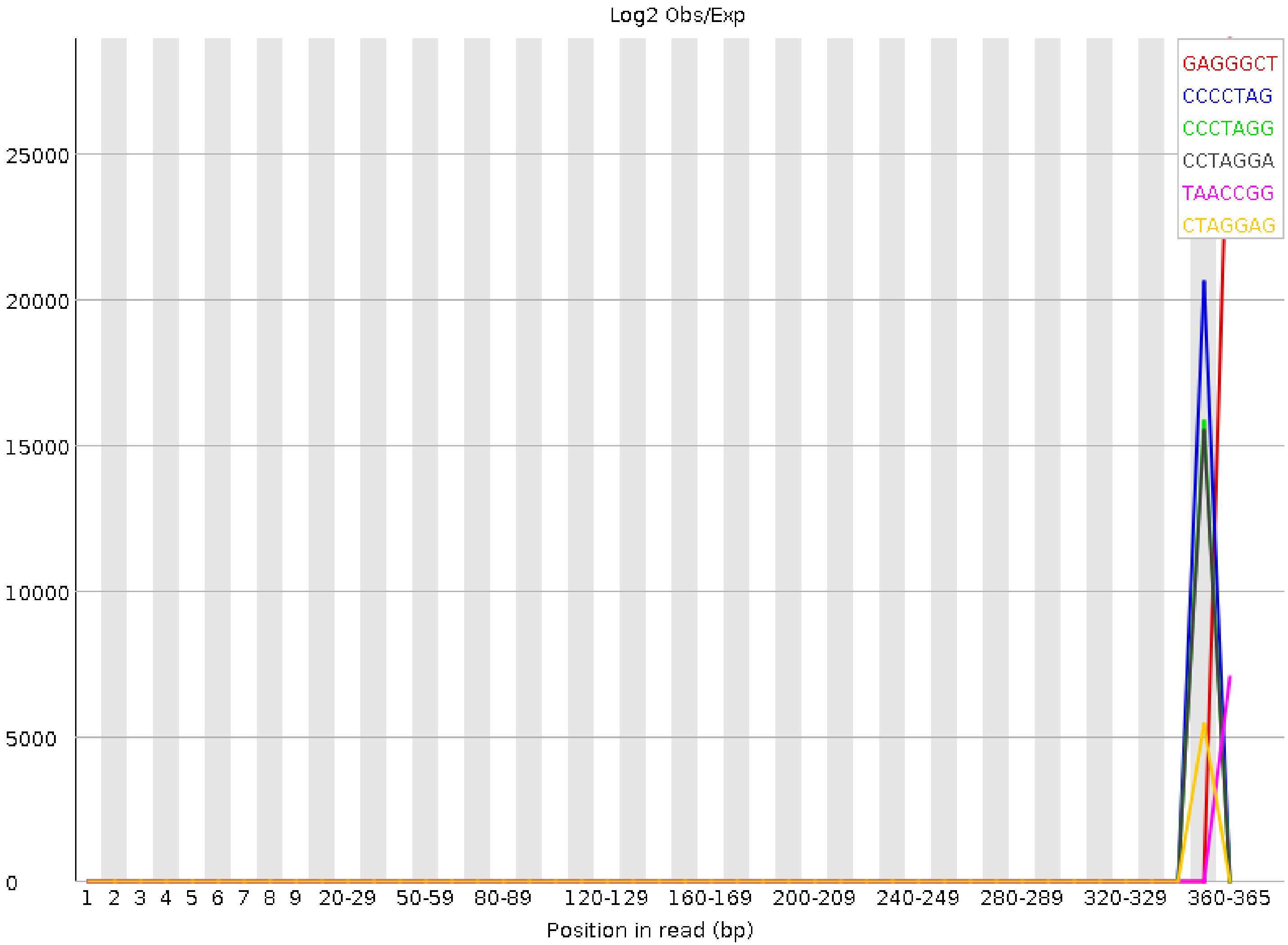
FastQC – quality check of aw data

✓ Sequence Duplication Levels

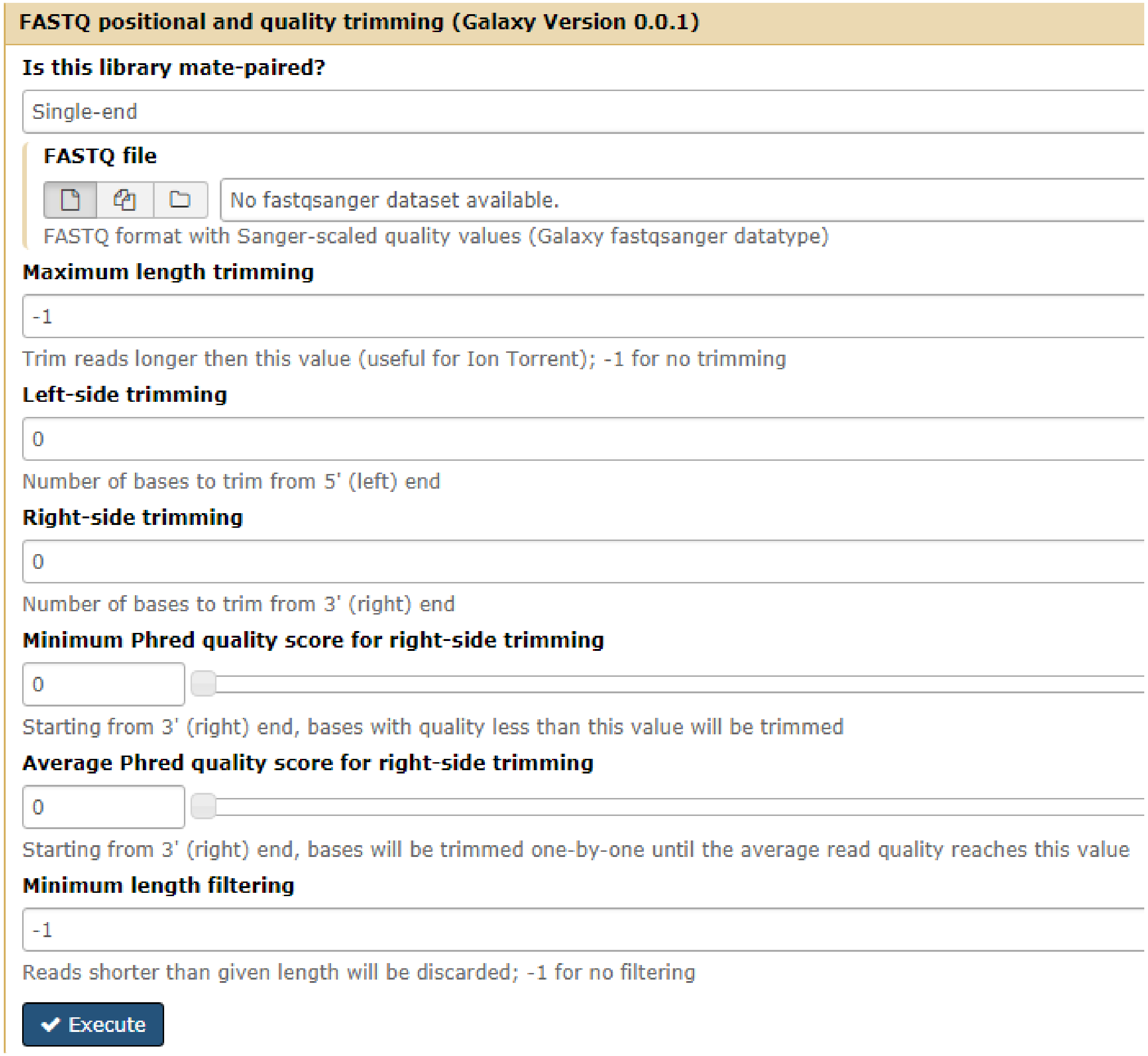


FastQC – quality check of aw data

Kmer Content



What should be trimmed out?



FASTQ positional and quality trimming (Galaxy Version 0.0.1)

Is this library mate-paired?

Single-end

FASTQ file

No fastqsanger dataset available.

FASTQ format with Sanger-scaled quality values (Galaxy fastqsanger datatype)

Maximum length trimming

★ -1

Trim reads longer then this value (useful for Ion Torrent); -1 for no trimming

Left-side trimming

★ 0

Number of bases to trim from 5' (left) end

Right-side trimming

0

Number of bases to trim from 3' (right) end

Minimum Phred quality score for right-side trimming

★ 0

Starting from 3' (right) end, bases with quality less than this value will be trimmed

Average Phred quality score for right-side trimming

★ 0

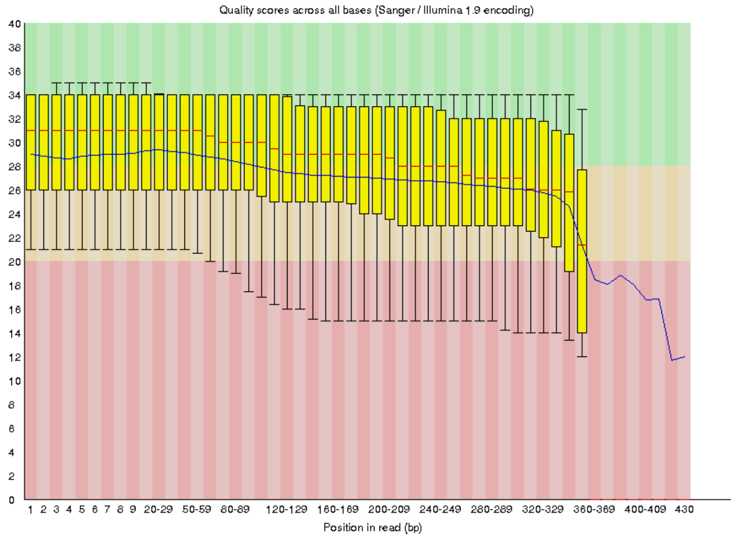
Starting from 3' (right) end, bases will be trimmed one-by-one until the average read quality reaches this value

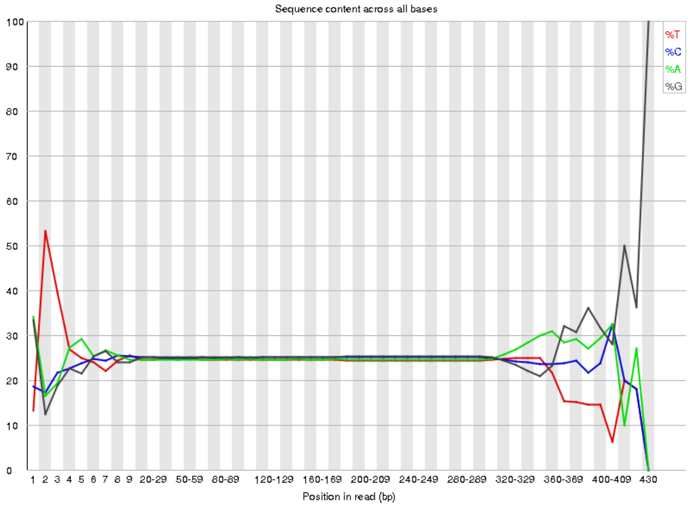
Minimum length filtering

★ -1

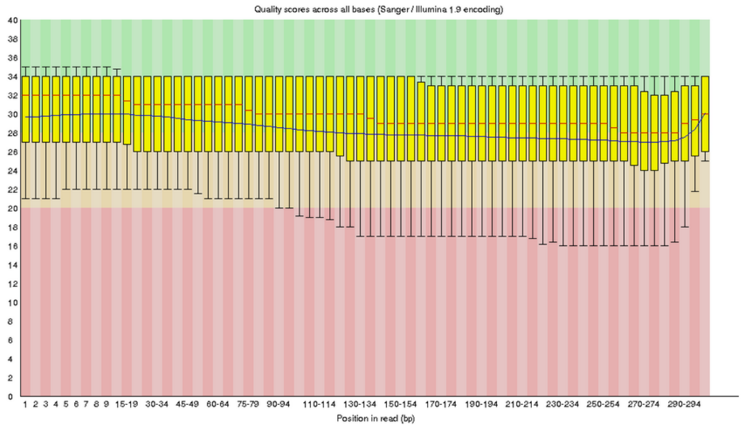
Reads shorter than given length will be discarded; -1 for no filtering

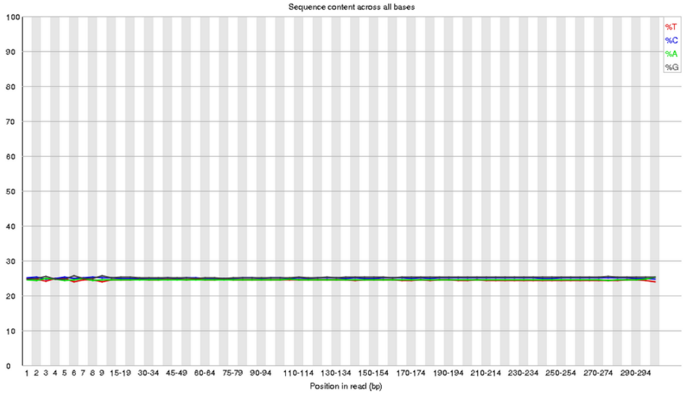
Before trimming





After trimming


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)