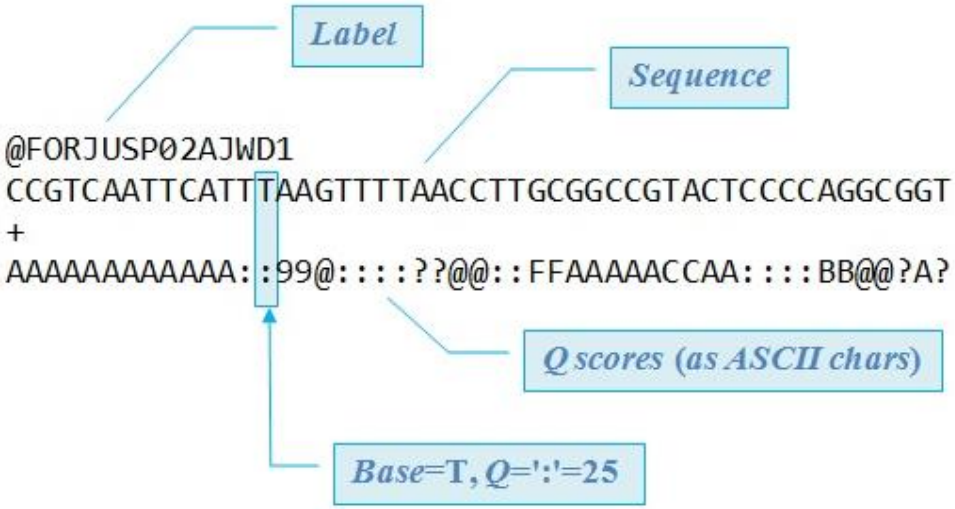
Introduction to NGS data formats, quality check and analytical tools

Valeria Michelacci

Joint Training Course on NGS
June 14th, 2022

Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health
European Union and National Reference Laboratory for *E. coli*, Rome, Italy

.fastq files



Each .fastq file covering a 5 Mb genome at 50X weights about **500 MB**

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%

Phred quality score

$$Q = -10 \log_{10} P$$

from 0 to 93 using ASCII characters 33 to 126

.fastq files

@

```
@X1L6C:01561:00672
AAATATCACCAAATAAAAAACGCCTTAGTAAGTATTTTTCAGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTG
GATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCAC
TAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCA
CCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAAGCCCGCA
CCTGACAGTGCGGGCTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCG
+
CC:9::FBC<CD7:88888(:>><C<CCCC<CCBBAAB/A@A8888,;<@;AABBB=?;B98992:B<
CGBBCGDCC??BCC;BB<ADEEED*CCCAAACCCBCABBDDBB>B??A;999;@8=>199A7>9::CBCH:B:>>>)999)
77037;<7==5=@@BBCC:C@BBB9B<E<D9>?><<6ADCBCBAABB@@@DDCCBA@@==+=.//?B<??AEB::6;DCD>
C:;;;-:9:BC<BBCCC9??<AA;AG<CB>GD@B;;;A<AE;AA<B??@9@C<BB<???BB;BBBAAAA:::BAB099/9>
@========(<<?)99997>>CCEBA>>=>2373333&3:99-33(3--717---43606704/47761
@X1L6C:01104:03031
AGAAGCTGCTATCAGACACTCTTTTTTTAATCCACACAGAGACATATTGCCCGTTGCAGTCAGAATGAAAAGCTGAAAAATA
CTTACTAAGGCGTTTTTTATTTGGTGATATTTTTTTCAATATCATGCAGCAAACGGTGCAACATTGCCGTGTCTCGTTGCTC
TAAAAGCCCCAGGCG
+
@AC=BCCC???B?@@CBB@???>>>>>*?8??>DAABEBCBABCAAA:@@>+9:8>;<;//.
98283988*44449;;9/88:?29:>>5;78333333&399298:6/./DCDDCC';>:ACBDAABB??9::+9<
1444@:?77-3<03368:8755888;:9833)3777'--'--
@X1L6C:03659:02717
GCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATA
GCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTA
CCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGACCCGCCACTGACCAGTGCG
+
???9?BB@<CAA;A8@?:?@@5::BCCCEC;C=CCC8CEJ8DE;AACF>CC?DDCCCBB:B@???9?;B=B=CAA@?;?BCG
CCCCCCBABBBBCCDDAA2:4;@???CAB@AAA9@@AB?C:;;C;CDCCC>ECCAA<AC<CB>DC<AB=CD=C9::A4::>
CC;@@@A?CI@DDAFKDDD:A@CBCDC::::99199+8;4746@CA?)<444/3:4934333-3888//
@X1L6C:02011:02071
TTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACA
CAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGTGACGCGTACAG
GAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTCGACCAAAGTAACG
+
=0>>>19;;,;;7=CCDADC;?::::::,5;;==4>273:<@BBCF=CDH;@;MMFEED@?>>>:::::*5/55<
;::@:;:BC=BCBB<B@@@D<@@B:;3:::9@<BB=BD=AC;@B:??3::CAC=CD;;;=BBAB>CC;AA;BAAAA9AD@>>
>>?955>4?9499985555558&4<>2:;661499888...88/56666666$;6/.5:8(..+'++
@X1L6C:01333:03005
GCAATGCCAGGCAGGGCATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGA
TGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACG
+
555/55/(///(////(/8:9:<=>?<?@:98A??676<:;;@:5555555554444;=4443333;383338<68>>
68=333118311111111113933644588?==<76992---2+++0/
```

@

@

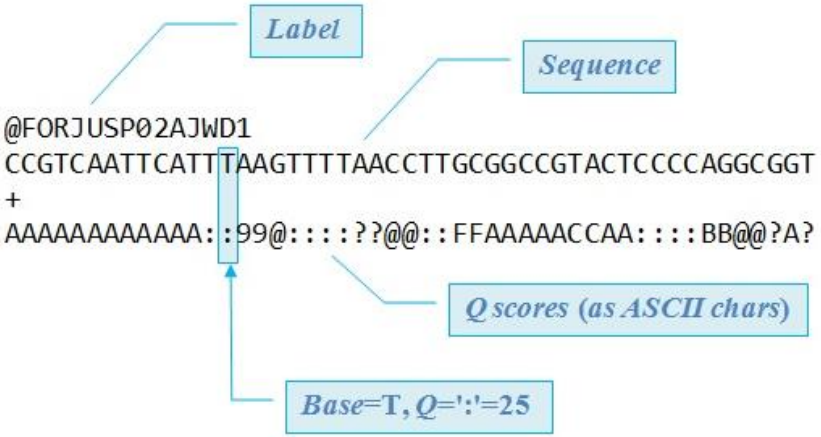
@

@

…and so on

Quality check

Output of NGS sequencers



.fastq file

Input for quality check

Sequencing errors would impact every following application

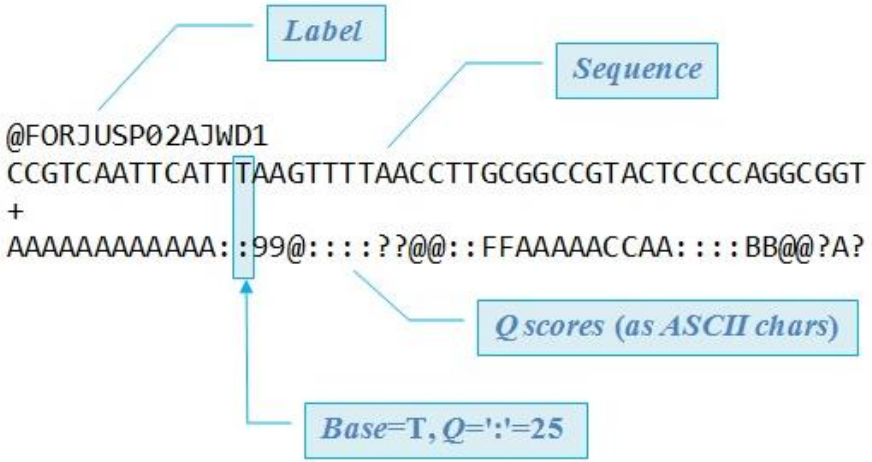
Unreliability of following results (and difficulty to detect the existence of problems!)

Parameters to control

- Phred score
- GC content distribution over all sequences
- Distribution of nucleotides
- Length of the reads
- Coverage

Adoption of corrective actions is possible to minimize these problems

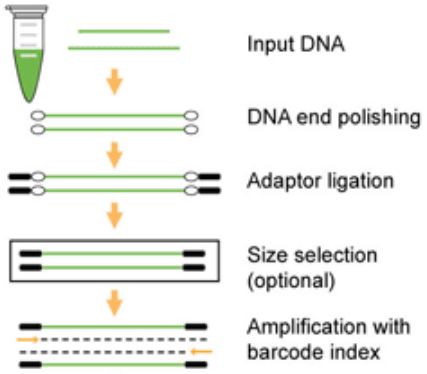
What should be trimmed out?



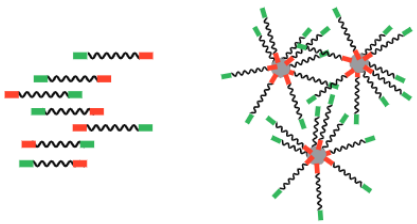
Phred quality scores are logarithmically linked to error probabilities

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%

- Adaptors and barcodes
- Low quality positions
- Very short sequencing reads (only for Ion Torrent reads)



Illumina



Ion Torrent

What should be trimmed out?

Eg: Ion Torrent reads

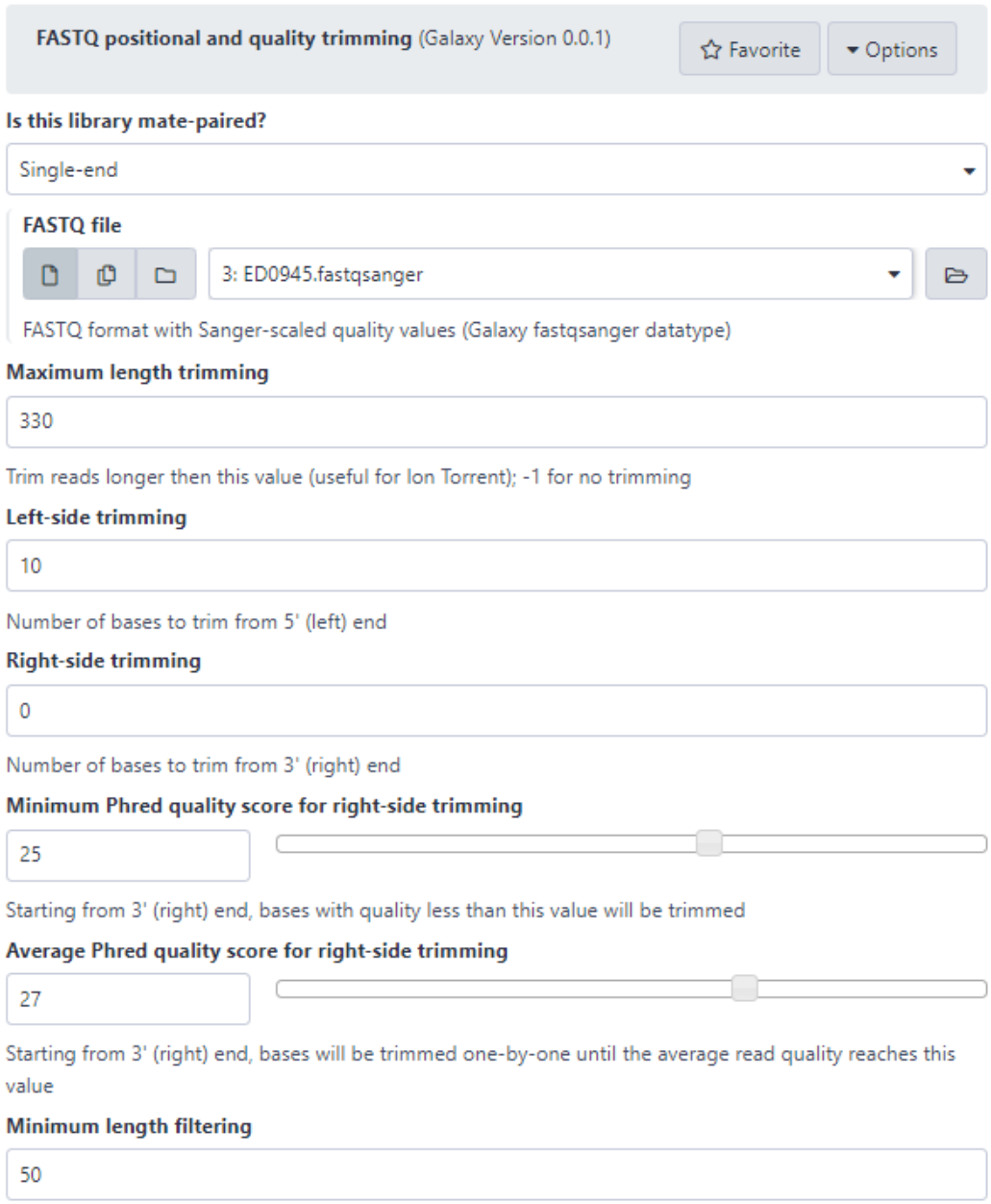
Maximum length trimming

Left-side trimming

Minimum Phred quality score for right-side trimming

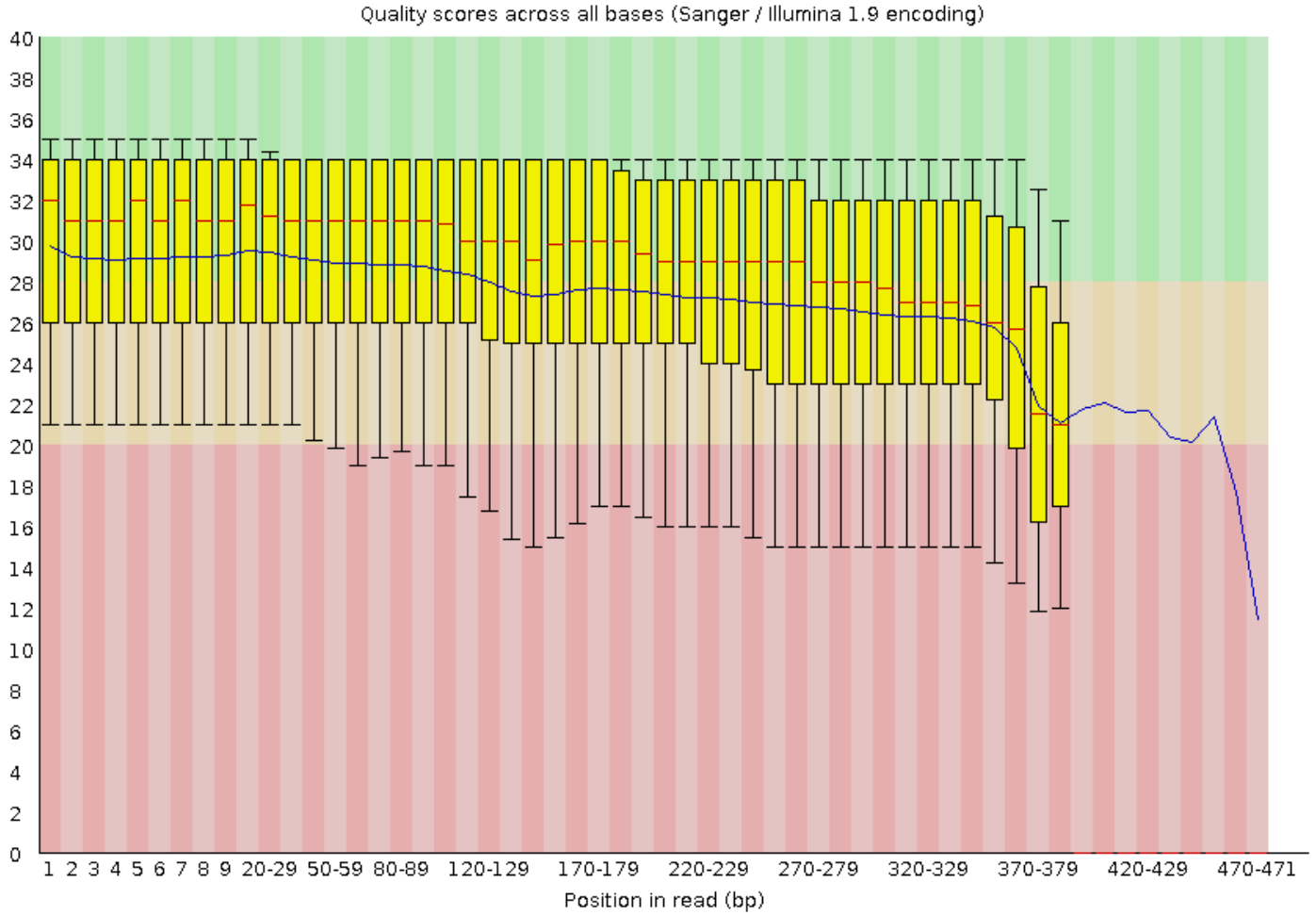
Average Phred quality score for right-side trimming

Minimum length filtering



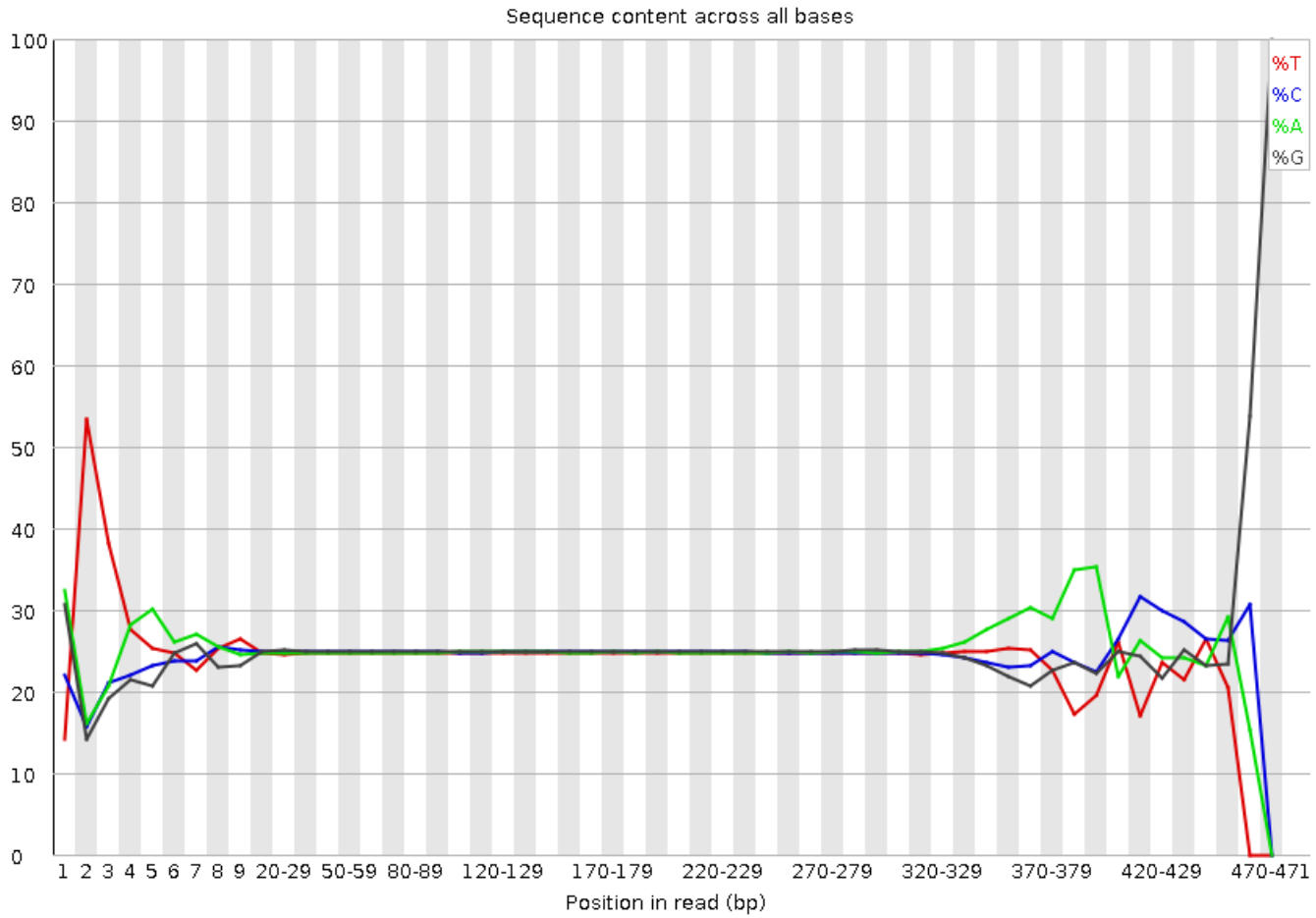
FastQC – quality check of raw data

Per base sequence quality



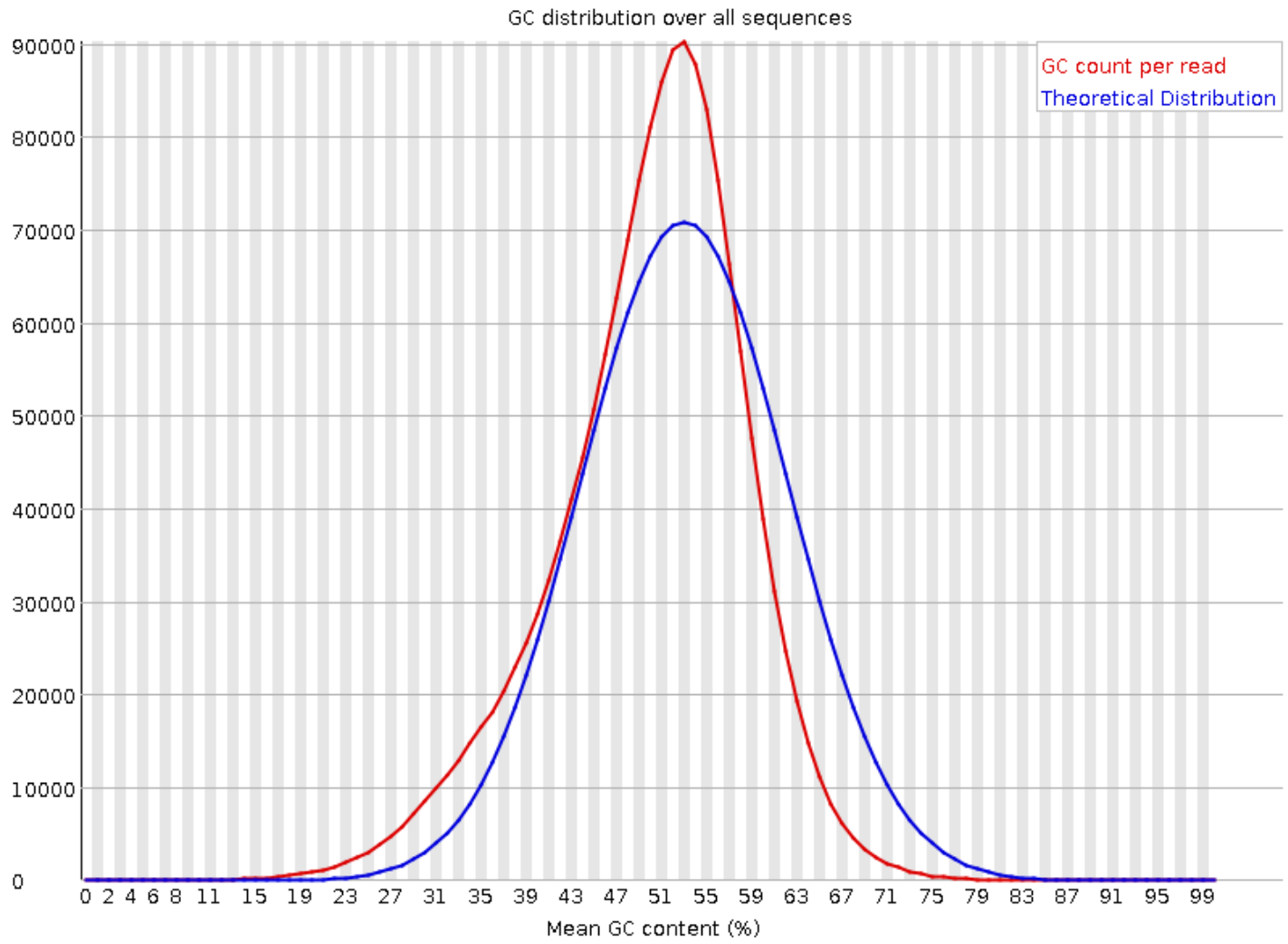
FastQC – quality check of raw data

Per base sequence content



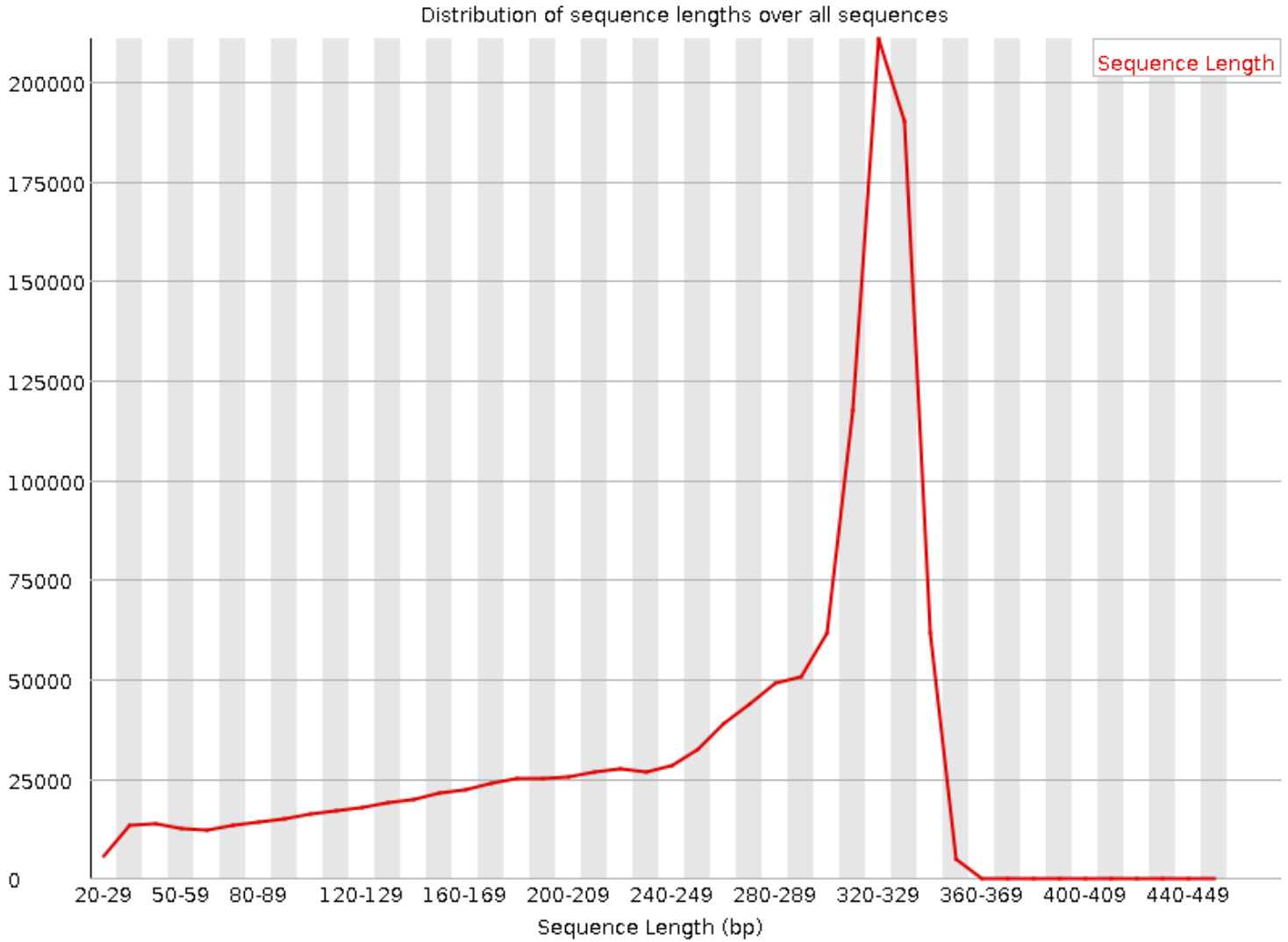
FastQC – quality check of raw data

Per sequence GC content



FastQC – quality check of raw data

Sequence Length Distribution



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

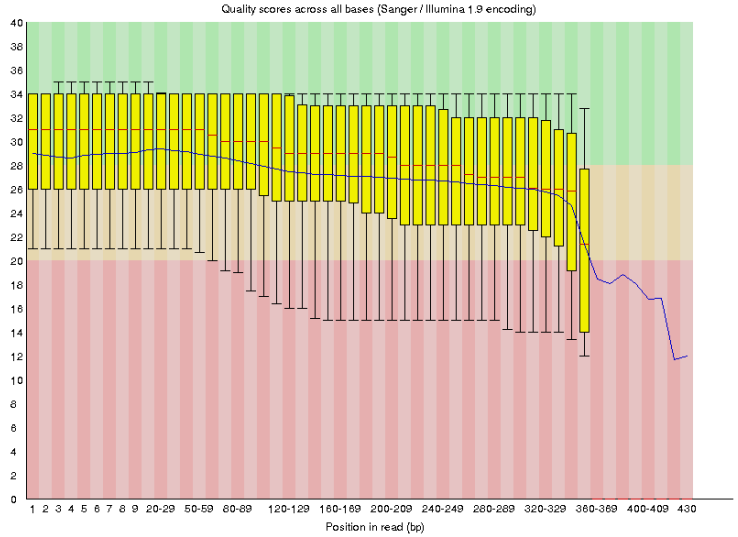
Minimum length filtering

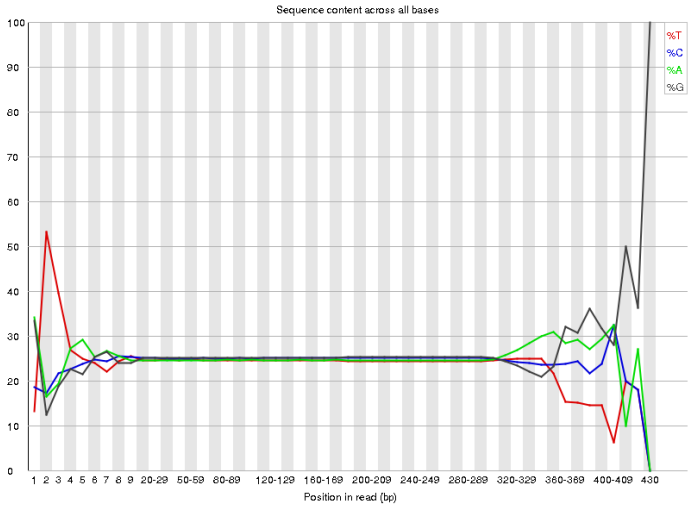
-1

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

Execute

Before trimming





After trimming