# Introduction to NGS data formats, quality check and analytical tools

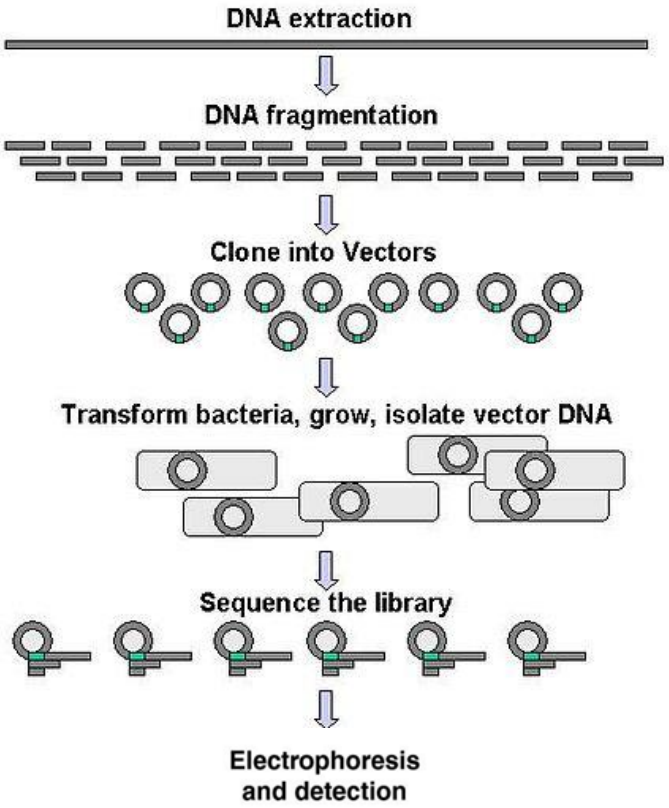
Valeria Michelacci

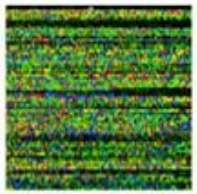
Bioinformatics training,
July 2019

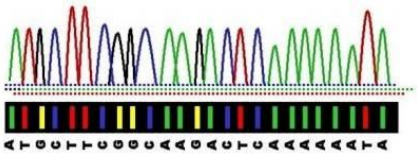
Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health
European Union and National Reference Laboratory for *E. coli*, Rome, Italy

# Conventional sequencing vs NGS

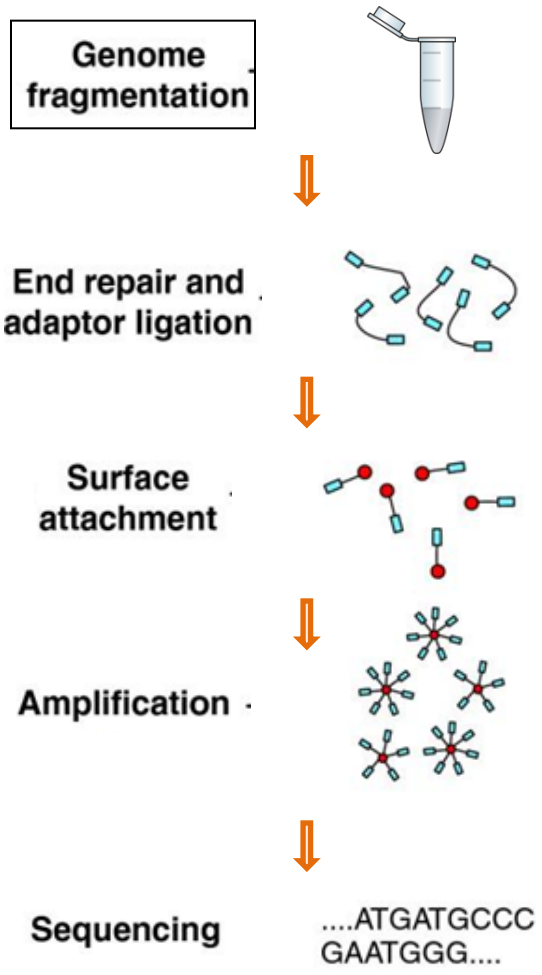
## Conventional



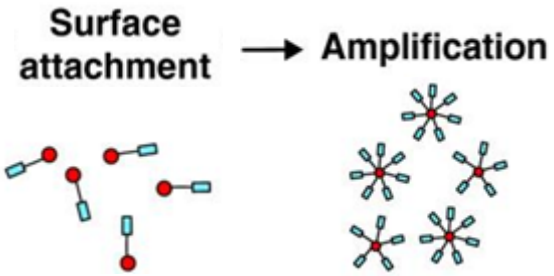





## NGS Pipeline



# Next generation sequencing




...

...

illumina MiSeq

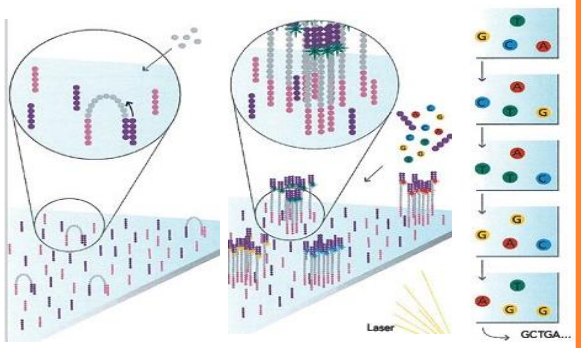
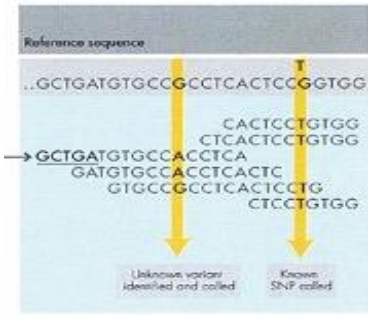
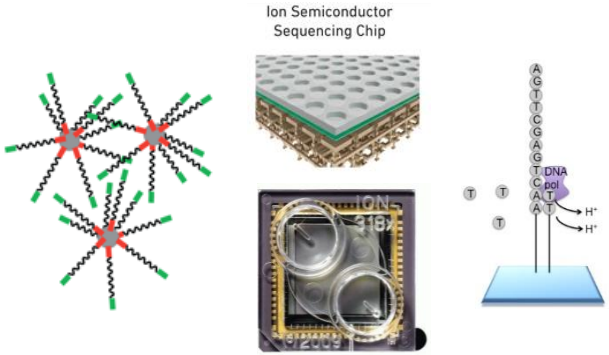



Image capture
Fluoresence detectioin

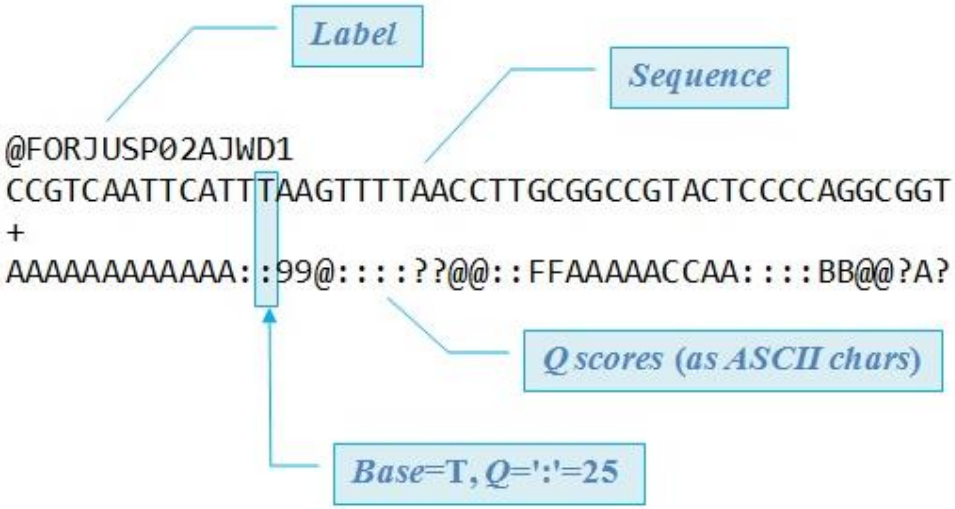



200bp-400bp short reads




pH variation when incorporating nucleotides in the growing strand

# .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/
```

@

@

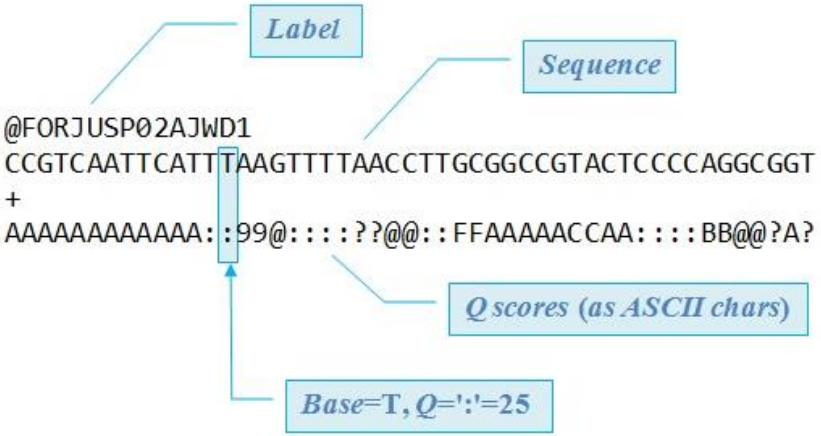
@

@

## …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 undetermined bases (N)
- Distribution of nucleotides
- Length of the reads
- Coverage

**Adoption of corrective actions is possible to minimize some of these problems**

# Coverage (depth)

## Reads mapped on a reference genome


Ref seq

COVERAGE

Mapped reads

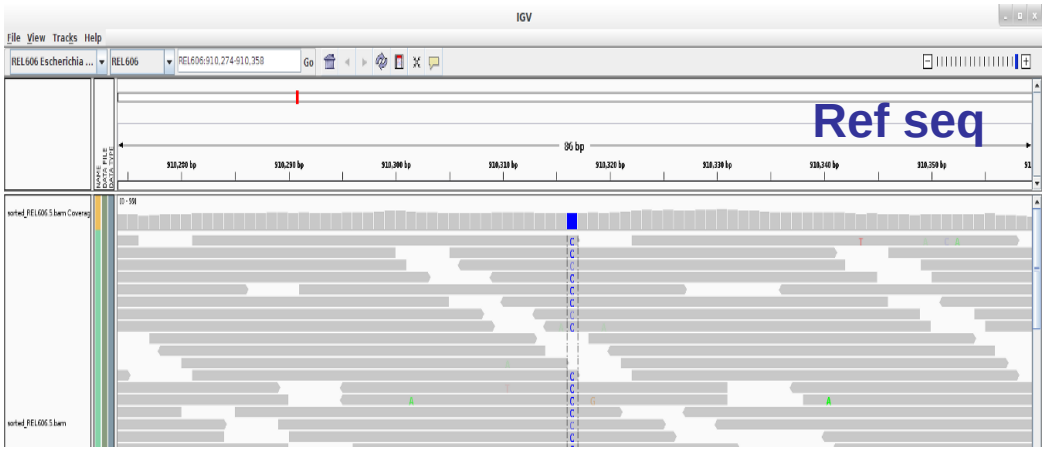
## Coverage assessment:

Total length sequenced in Mb / expected genome size (5 Mb)

Count of reads mapping on housekeeping genes (e.g. MLST)

# Alignment (mapping)

**Alignment of the sequencing reads on a reference sequence or on a database of reference sequences**




**Possibility to directly inspect the presence/absence of a target sequence and the presence of SNPs at interesting positions by opening the bam file with a graphic viewer (e.g. IGV)**

| QNAME | FLAG | RNAME | POS | MAPQ | CIGAR |
|---|---|---|---|---|---|
| ME2UT:01383:01267 | 0 | gad:3:EF547388 | 1285 | 0 | 113M18I4M |
| ME2UT:02555:01592 | 16 | gad:4:CP001925 | 1123 | 0 | 27M1I248M39I4M |
| ME2UT:02231:01820 | 0 | gad:5:CP001846 | 87 | 1 | 138M |
| ME2UT:01605:00255 | 16 | gad:5:CP001846 | 399 | 1 | 51M |
| ME2UT:01345:02031 | 16 | gad:5:CP001846 | 685 | 1 | 176M |
| ME2UT:03330:02136 | 16 | gad:5:CP001846 | 1050 | 1 | 6M1I38M |
| ME2UT:01475:02165 | 0 | gad:6:BA000007 | 1 | 0 | 3M31I47M1D130M |
| ME2UT:01488:00709 | 16 | gad:6:BA000007 | 1 | 0 | 4M32I55M1I149M |
| ME2UT:01943:01152 | 16 | gad:6:BA000007 | 13 | 1 | 196M1I50M1I10M |

**Possibility to convert the output in a sam file (tabular) to extract interesting info and sequences**

# Assembly

## Short sequencing reads

### .fastq file

```
@HWI-ST700693:238:B0224ACXX:1:1101:1218:1982
NACACTTGCTTTGGTGACAGCGGGGCATCCTCAAGC
+
#1=DDDDDHAFF?GEFGIIIIIIIIIIIIIIIIIFI
@HWI-ST700693:238:B0224ACXX:1:1101:1161:1986
NGATTTTGACCTCTCCAGTTTCCTCTTAACACTTTG
+
#1:BDFFFGHHHGJJJIIJHIJJJJJJJJJJJJJJJ
@HWI-ST700693:238:B0224ACXX:1:1101:1193:1989
NTATCCAGCCTGCGGTGCTACTTGGTGGAAGAGGAT
+
#1=DDFFFHGHGGJJFGHJJIJJJIEGECHDFHCC?
@HWI-ST700693:238:B0224ACXX:1:1101:1440:1981
NTCAAGAATCCAAGTGGGGCCAGCATAATGTACGCT
+
#1=DDFFFHGHDFDAEGIIFGIICGGHGBFGEFDHI
@HWI-ST700693:238:B0224ACXX:1:1101:1367:1983
NATTAGAACAGATCGCTACTTCGCCCGAAGATACAT
+
#4BDFFFFFHHHHHJGIJIJJJIJJJJJJJJJJJJIJ
@HWI-ST700693:238:B0224ACXX:1:1101:1395:1988
NTGGAAACGTTTTTAAACGCGGAGACAGCGTGGAGT
+
#1=DDFFFHCFFHJJJIJJIJJIJJJJGGIFHIGI7
@HWI-ST700693:238:B0224ACXX:1:1101:1285:1994
NCTTTGCTGTATTGACCGTTTGTAGATTTGAATCTT
+
#4=DDFFFHBHHHHIGIJFHIJFGGGIGIHIJIJII
@HWI-ST700693:238:B0224ACXX:1:1101:1632:1989
NTCTATGAATGTTCAAGCGGTAGCTGAGGAGAGTCC
+
```

FastqSize ≈ GenomeSize x Coverage x 2

**At least 0.5 GB per genome**

## Partially assembled genome (contigs)

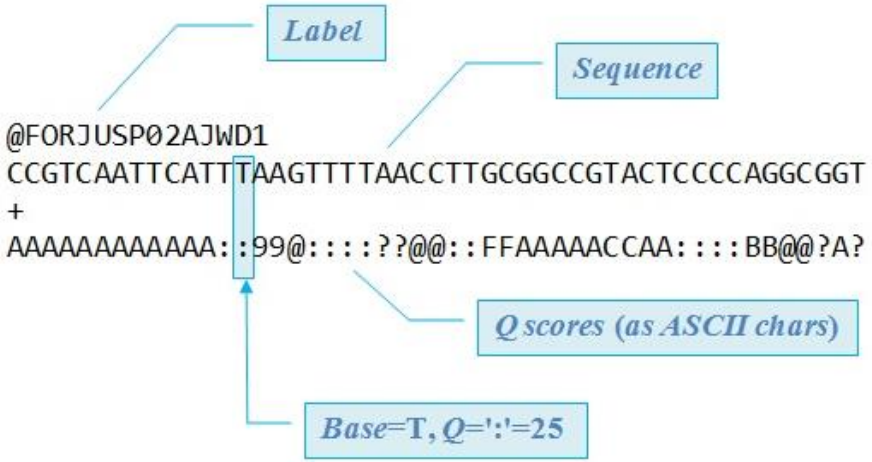
### .fasta file

```
>NODE_1_length_449_cov_4.835189
ATCTTTCGCGCCTTCCAGCTCCAGCCATTCGGAACCGTTCGCCAGAAAACGGGCGTAATC
GGGTAAGACATAGCGCGGTTTGTACGGCGCATGACCTTCAAACATATCGCAGATTACACC
TTCATCCAGCGCGCGGCGGGCTTCGGCAGGAAGCTGTGGGTAAGGCAGATTGTTTTCTGC
TTCCAGTGCCAGAAAATGGCGCTTCTGCTCCGGGCTAAGCACTGGGCTGGTGACAATTTG
CTGGCAACGTTGTTGCAGTGCATTTTCATGAGAAGTGGGCATCTTCTTTTCCTTTTATGC
CGAAGGTGATGCGCCATTGTAAGAAGTTTCGTGATGTTCACTTTGATCCTGATGCGTTTG
CCACCACTGACGCATTCATTTGAAAGTGAATTATTTGAACCAGATCGCATTACAGTGATG
CAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGG
>NODE_2_length_309_cov_4.686084
ACTGGTCAGTGCGGGTATCCTTGGACAATGGCCGATTGGACGTCTGGCGGATAAGTTTGG
TCGACTGCTGGTGTTGCGTGTTCAGGTCTTTGTCGTCATTCTCGGCAGTATCGCGATGCT
TAGCCAGGCGGCGATGGCCCCAGCGTTATTCATCCTCGGTGCCGCTGGCTTTACGCTATA
TCCGGTGGCGATGGCATGGGCTTGCGAGAAAGTTGAACATCATCAACTGGTGGCGATGAA
CCAGGCCTTACTGTTGAGCTATACTGTGGGAAGTCTGCTTGGCCCGTCATTTACCGCTAT
GCTAATGCAGAATTTCTCCGATAATTTATTGTT
>NODE_3_length_101_cov_3.346535
AGCGCATGAGCGCGCAGCGCCGCGTTACGTGGTGCATCAGCATGATGTTGGCCGGAGAG
TACAGAGACTCCCCTTCATCCATGATGCCCTCTTTCACCAGCAGTTCTTCAATCATCACC
AGACC
>NODE_4_length_311_cov_3.610933
CATCAACGCTAAAAGCCAGATGACGCAGACCGCAAGCTTCCGGTCGGCTGGGTCGTTCCG
GCGGGAACGGAAATGAGAAAAGCTCAATCACATATTGCCCATTAAGCGCCAAATCCCCTT
TCCATGAGTCGCGCGCTTCGCGATAGACTTCGCTTTGCAGCGTGAAACCAAGAATATCGC
AGTAGAAAGCTTTGCTCACCGCATAATCCGTCGCAATAATCGCAATATGGTGAACCTGTT
TTAAACCCAGCATAACGTCTCCTTTATTTGTTAACAGCACGTTACTCGCCCGGAAGCCGC
TCTGGCAAGTTATCCCGCCATTTTTAGGACTCGTA
>NODE_5_length_186_cov_4.973118
CGAAGATATAAGAAAGCGAACCAGAAAGAATGCCGGAGAACTTCATCAATTCATCACCTG
CATTGAGCAGATTTTGCAGGTTCTCAATAACCGGTAATCCAGCCCCAACGTTGGTGTCAT
AGAGGAATTTACGCCGCGATTTTTCCGCCGCATAACGCAACTGATGGTAGTAATCCATCG
ACGAGGTGTTGGCCTTTTTGTTCGGCGTGA
```

FastaSize for *E. coli* contigs

**~5.5 MB**

# 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% |

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**

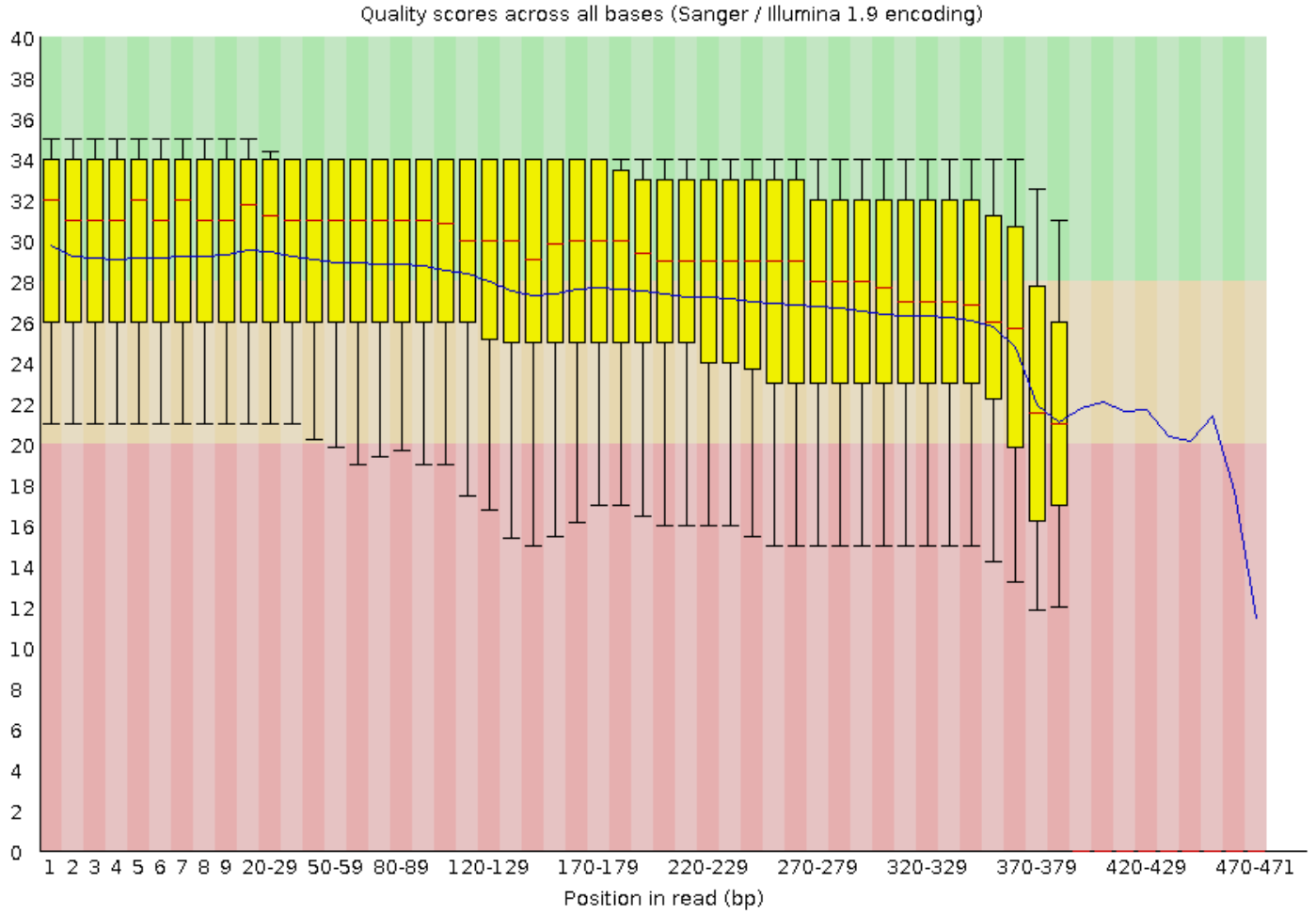
-1

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

Execute

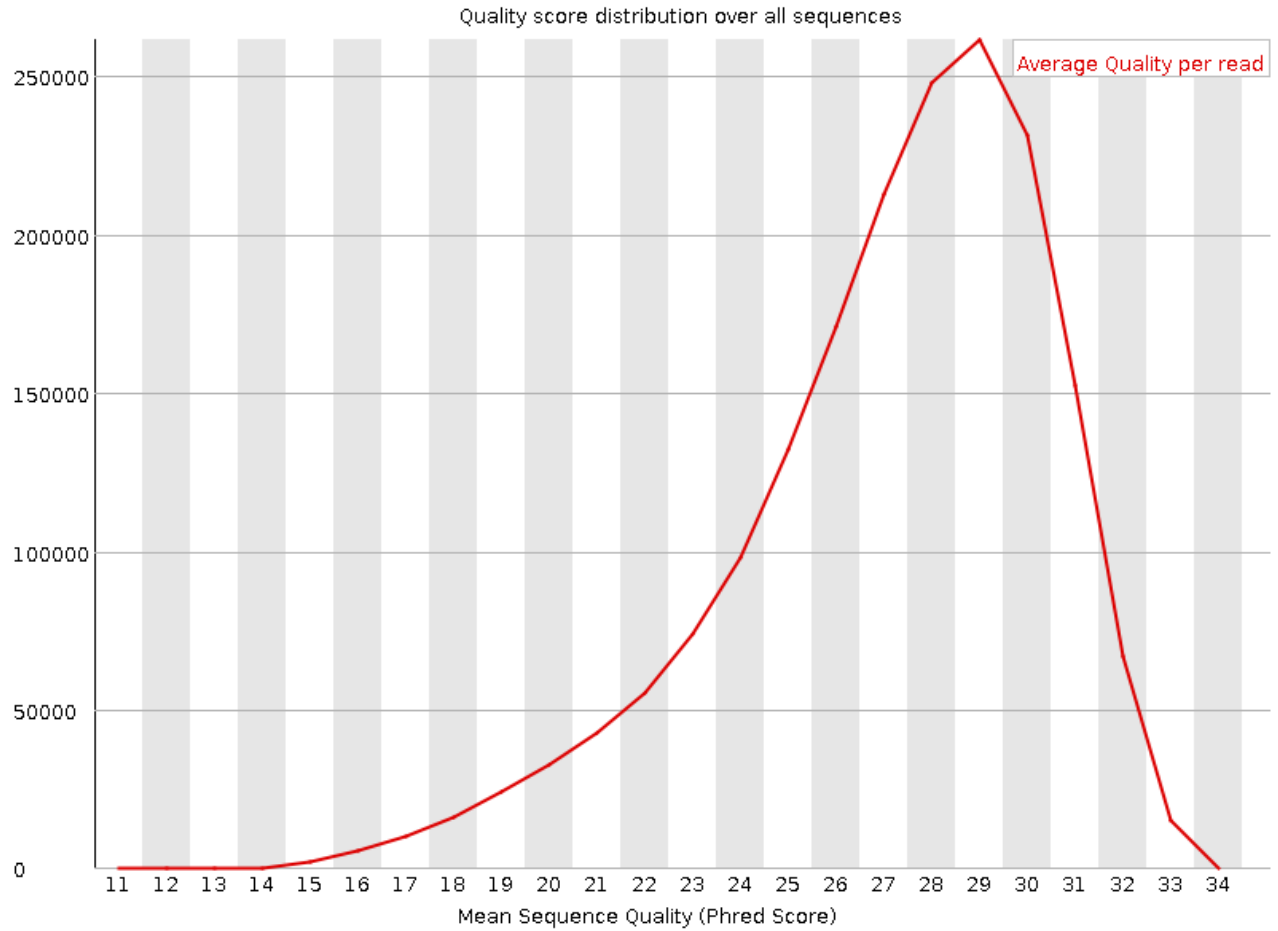
# FastQC – quality check of aw data

## Per base sequence quality



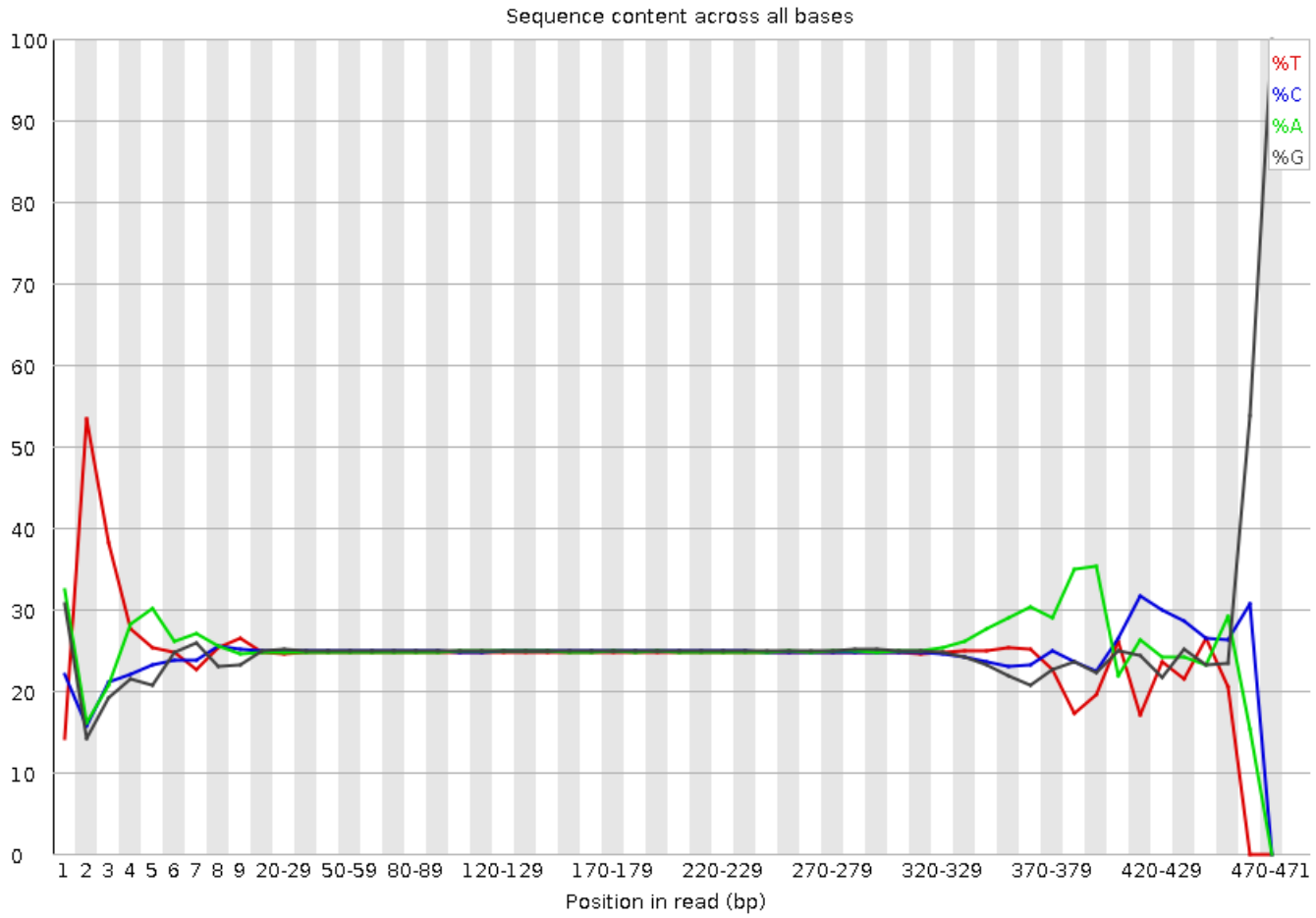
# FastQC – quality check of aw data

## Per sequence quality scores



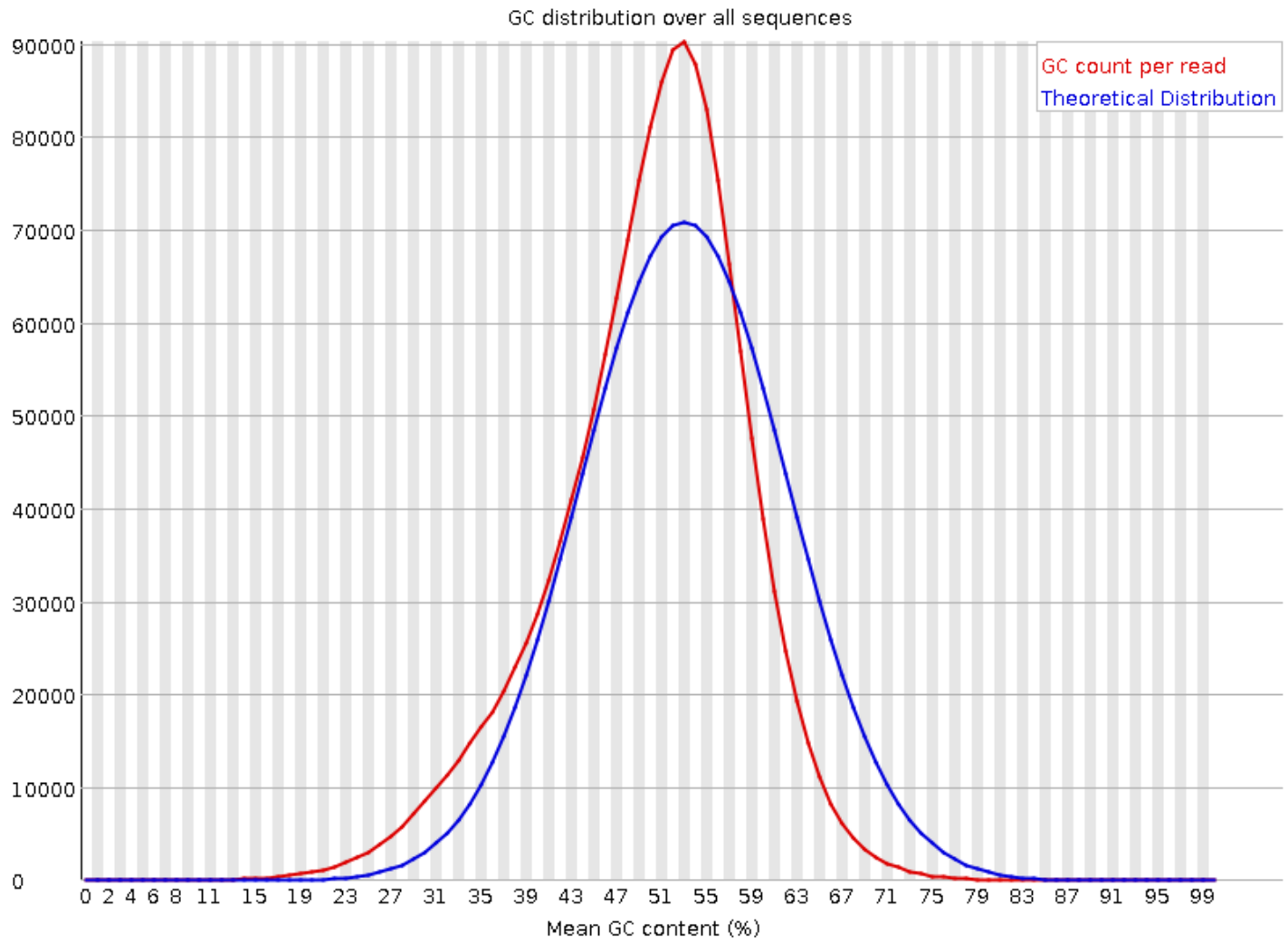
# FastQC – quality check of raw data

**Per base sequence content**



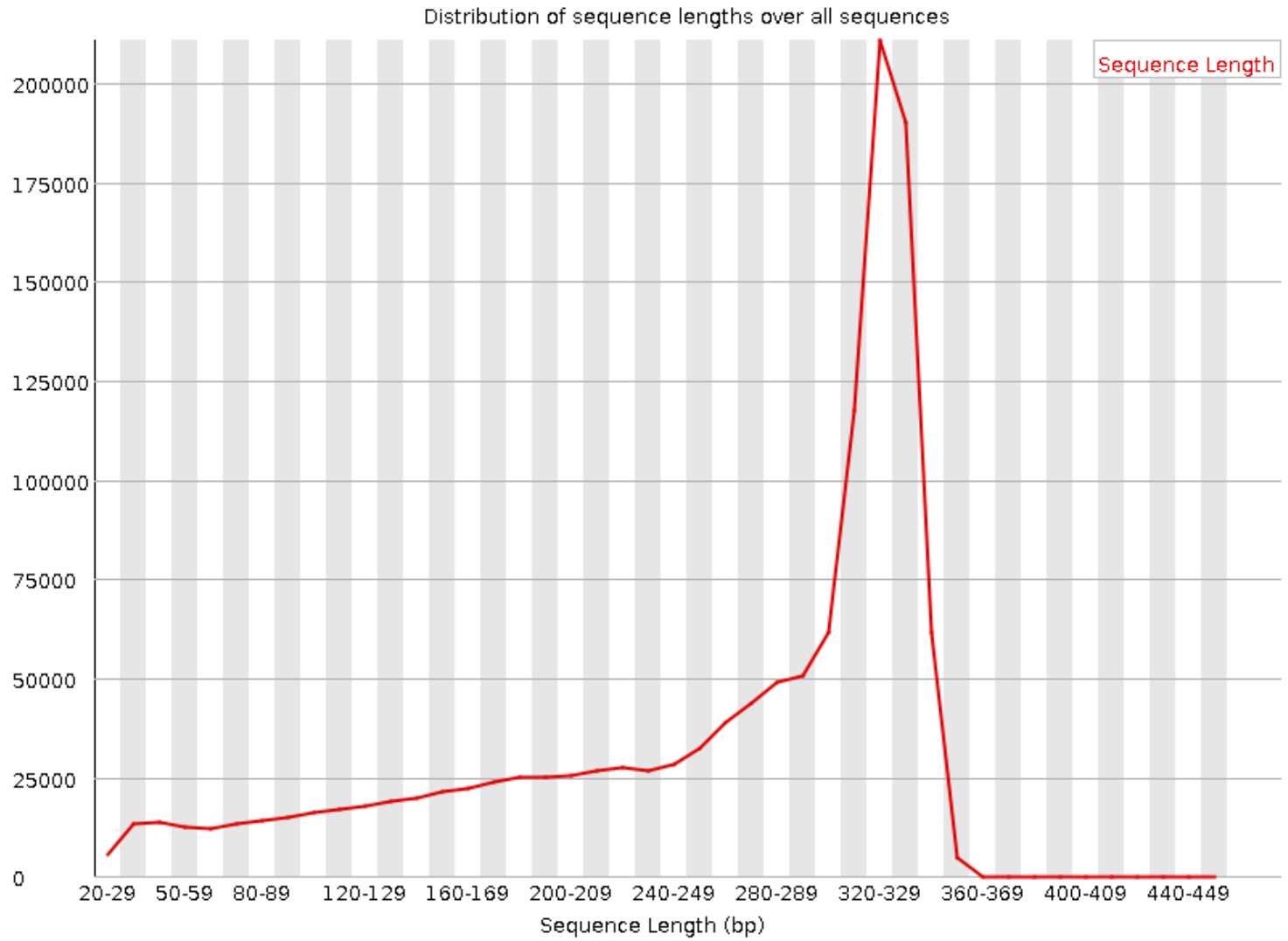
# FastQC – quality check of aw data

## Per sequence GC content



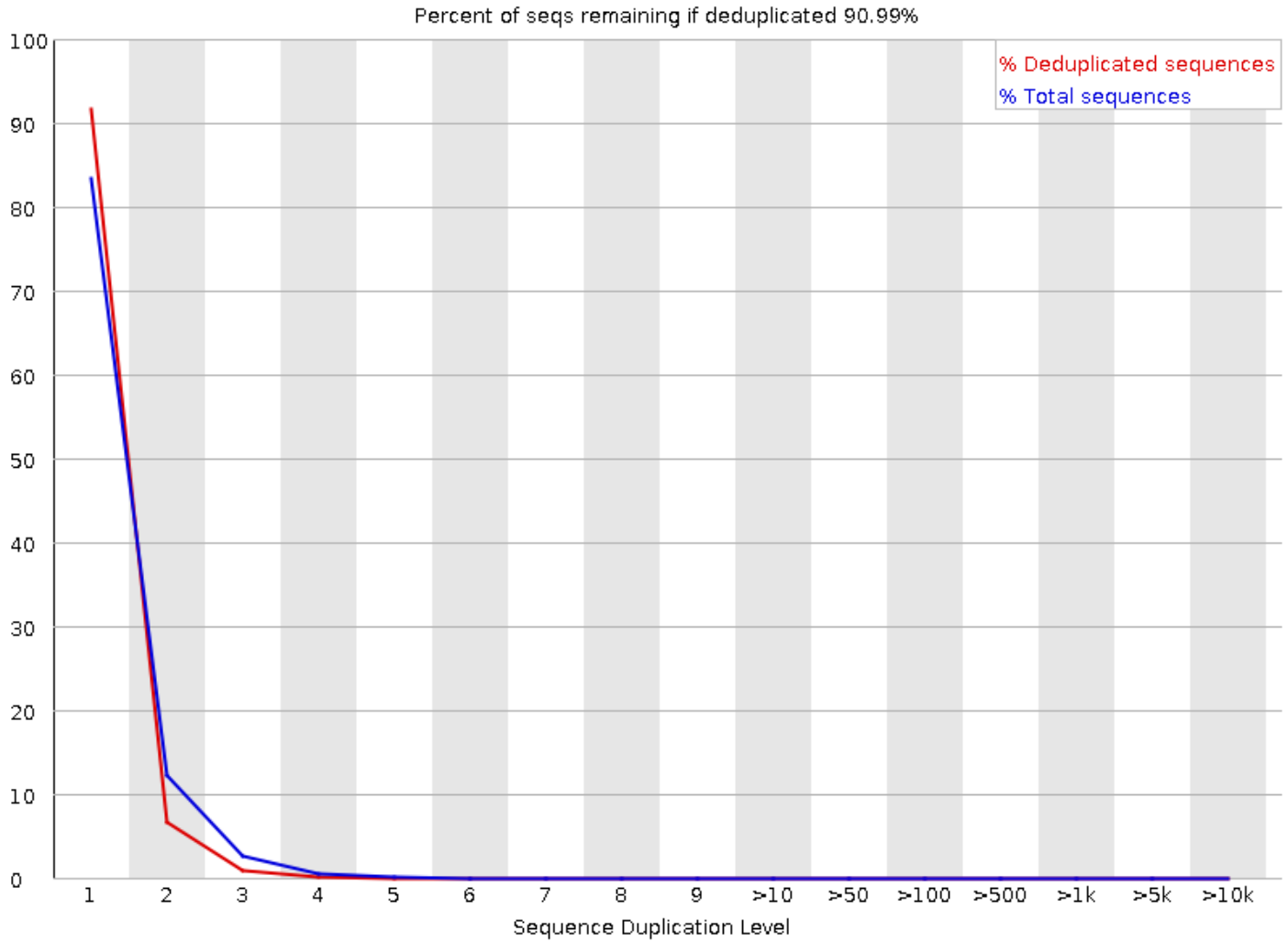
# FastQC – quality check of aw data

## Sequence Length Distribution



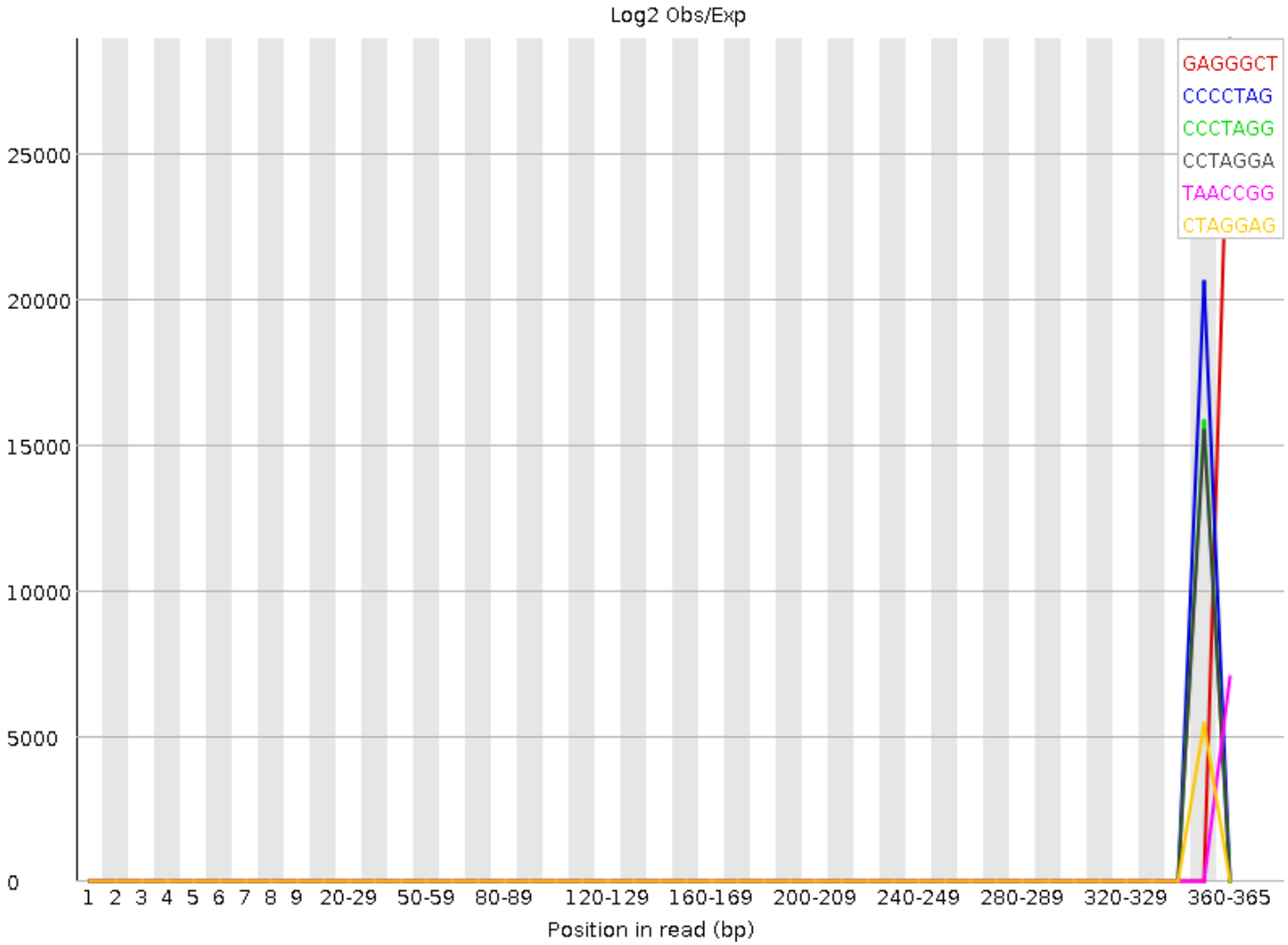
# FastQC – quality check of aw data

## Sequence Duplication Levels



# FastQC – quality check of raw 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

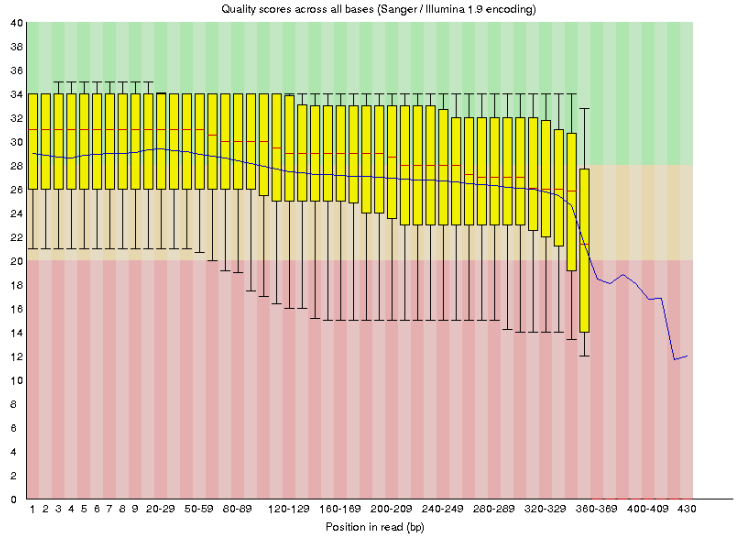
**Minimum length filtering**

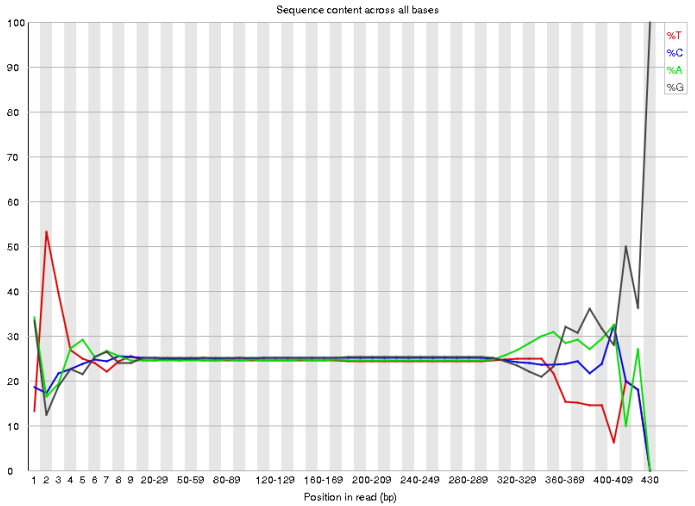
-1

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

Execute

# Before trimming




# After trimming