

Quality control and artefact removal

Izzy Newsham
MRC Cancer Unit
University of Cambridge

CRUK Bioinformatics Summer School 2021
21st July 2021

Many thanks to Joanna Krupka, the original creator of these slides

Why do we need quality control?

NGS generates highly accurate data, however there are still a few types of errors:

- Contamination with adapters
- Technical duplication in the library
- Failure at specific parts of the flow cell
- Amplification bias - PCR duplicates
- ...

Quality scores in fastq files

Fastq files store quality scores in ASCII characters

```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAGAAATG
+
AAAFFJJJJJJFJJ<J<FJJJJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
```

e = probability
of sequence
base being
wrong

$$\longrightarrow Q = -10 \cdot \log_{10}(e) \longrightarrow \text{Quality character} = \text{ASCII}(Q + 33)$$

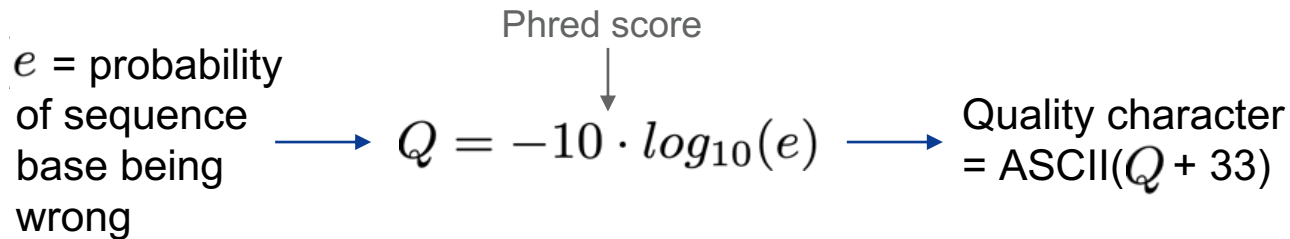
ASCII-quality score mapping:

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |           |           |           |           |
Quality score: 0.....10.....20.....30.....40
```

Quality scores in fastq files

Fastq files store quality scores in ASCII characters

```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAGAAATG
+
AAAFFJJJJJJFJJ<J<FJJJJJJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
```



ASCII-quality score mapping:

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |           |           |           |           |
Quality score: 0.....10.....20.....30.....40
```

FastQC

So we use FastQC

- A tool to generate reports based on sequencing quality information from FASTQ or SAM/BAM files
- Command line and interactive mode
- Outputs an html report and a .zip file with the raw quality data
- Enables a quick look at the potential problems with your experiment



FastQC report - summary






Good quality sequence

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Bad quality sequence

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

FastQC report – basic statistics



Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45



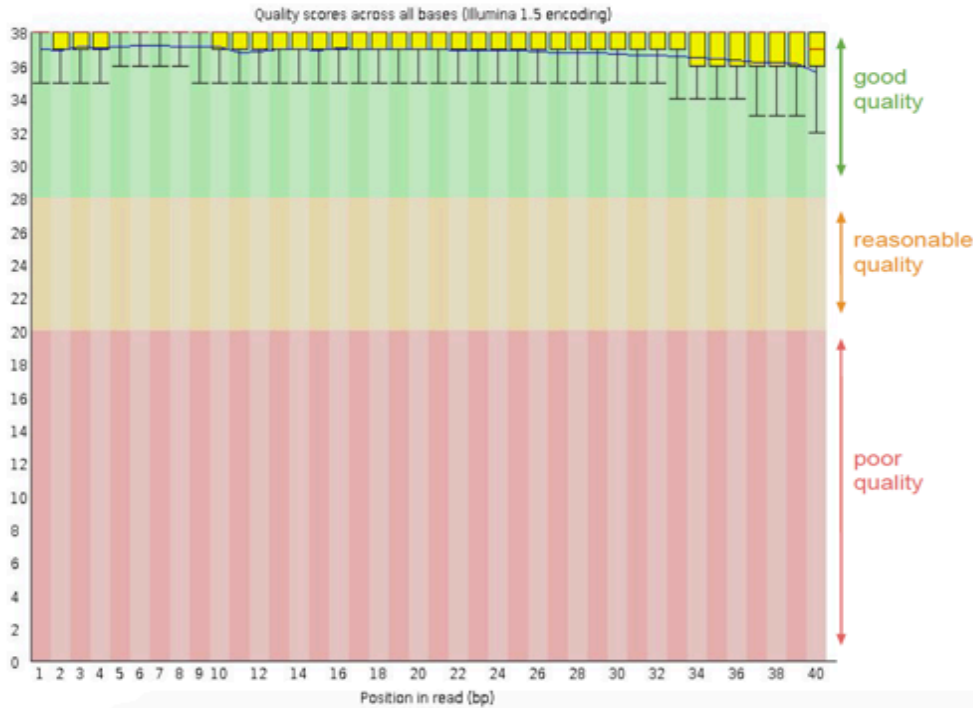
Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

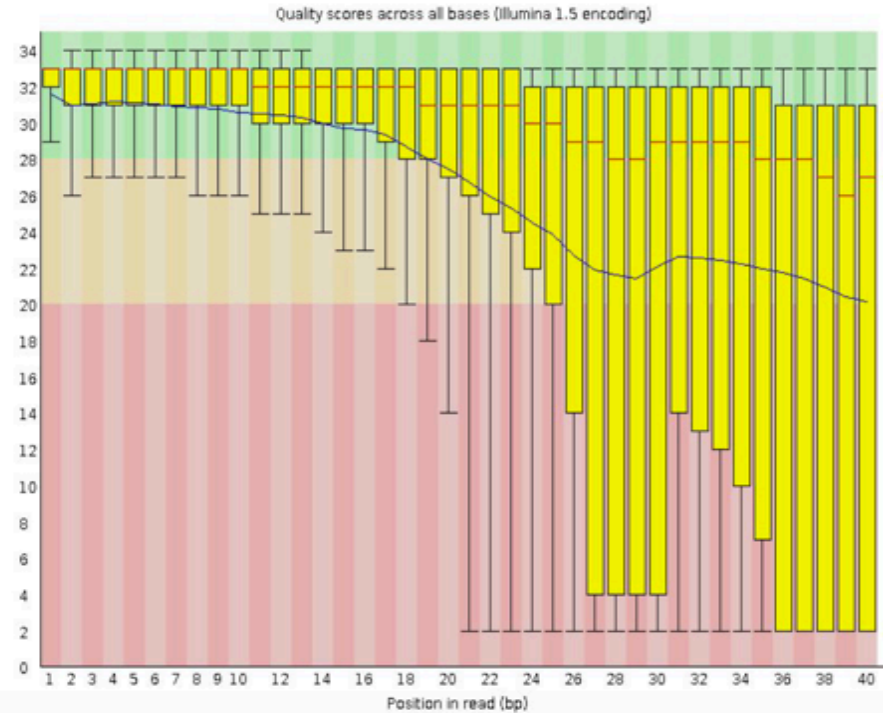
FastQC report – per base sequence quality

Examines the Phred quality scores

✔ Per base sequence quality



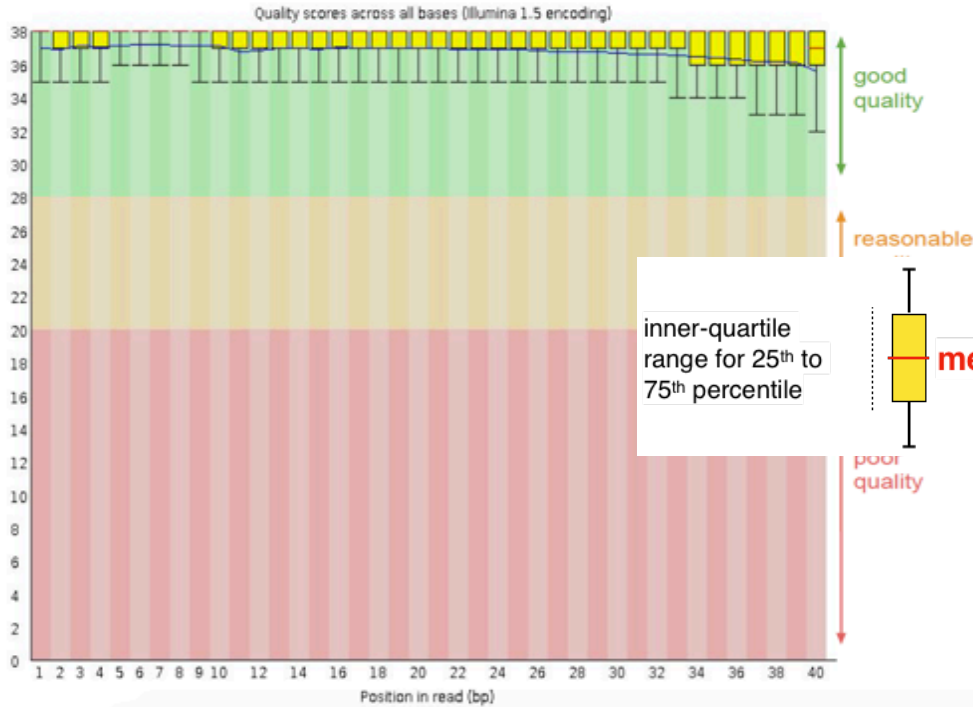
✘ Per base sequence quality



FastQC report – per base sequence quality

Examines the Phred quality scores

✔ Per base sequence quality



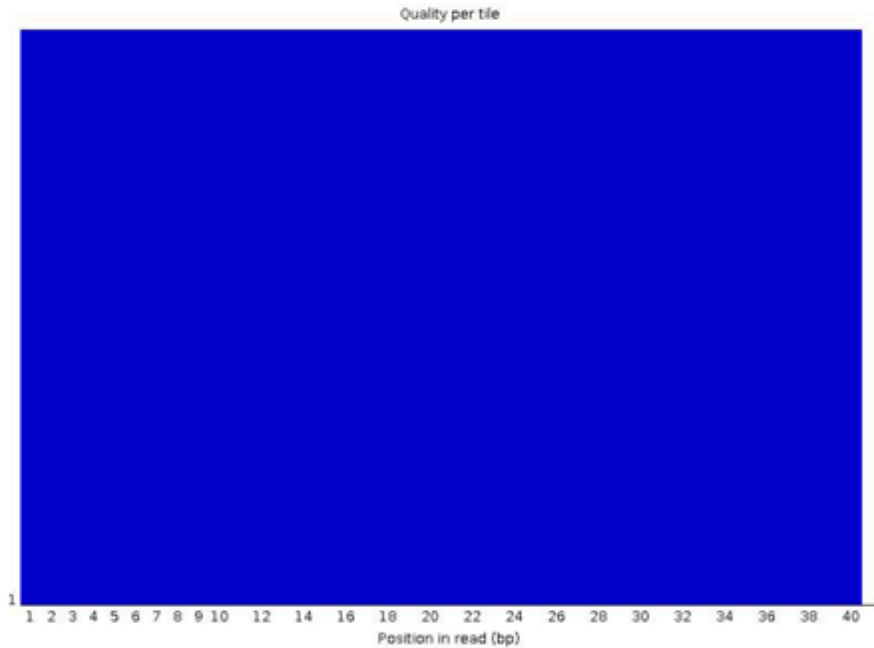
✘ Per base sequence quality



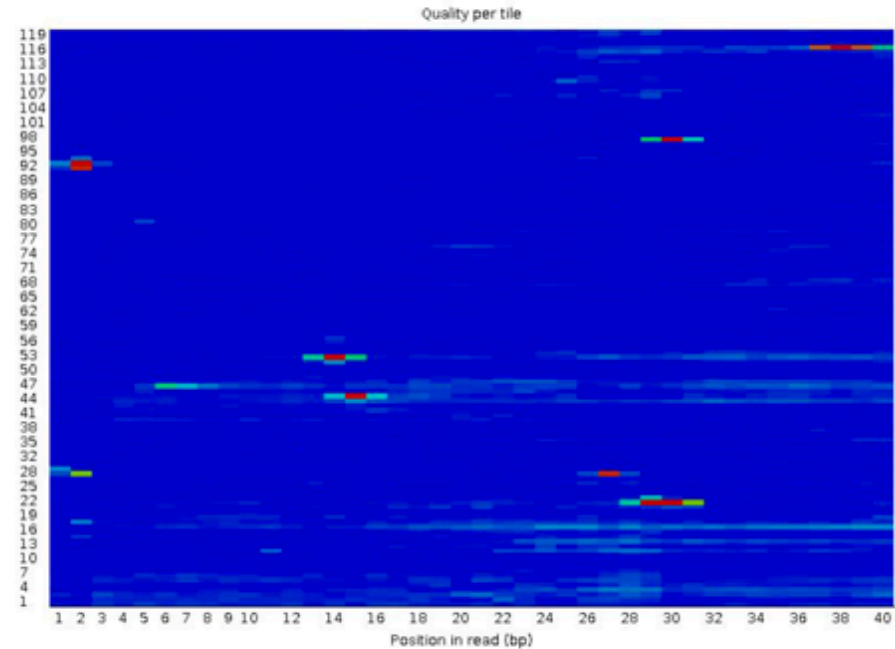
FastQC report – per tile sequence quality

Examines the Phred quality scores

✔ Per tile sequence quality



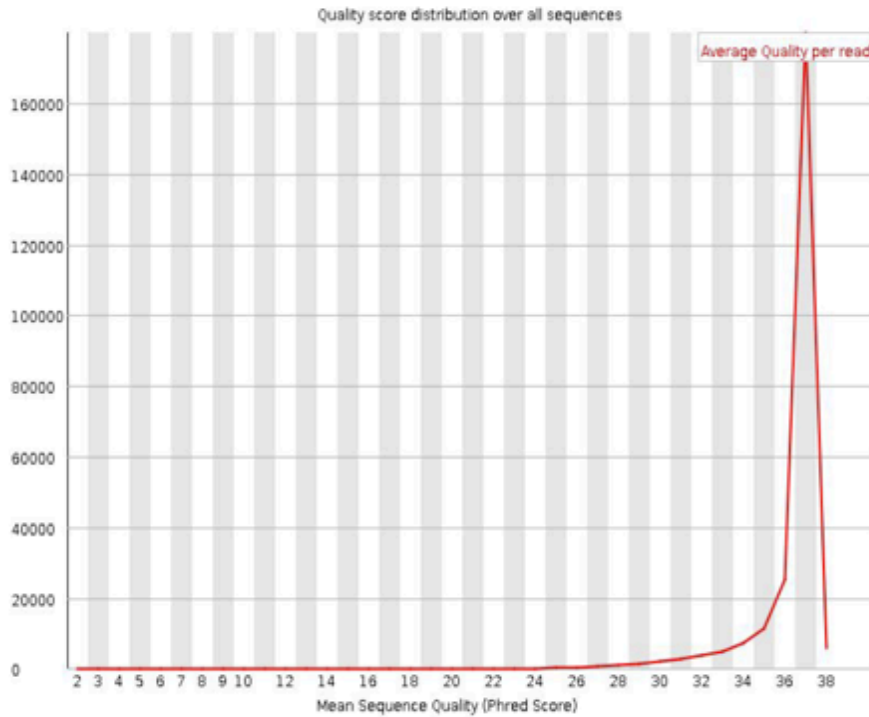
✘ Per tile sequence quality



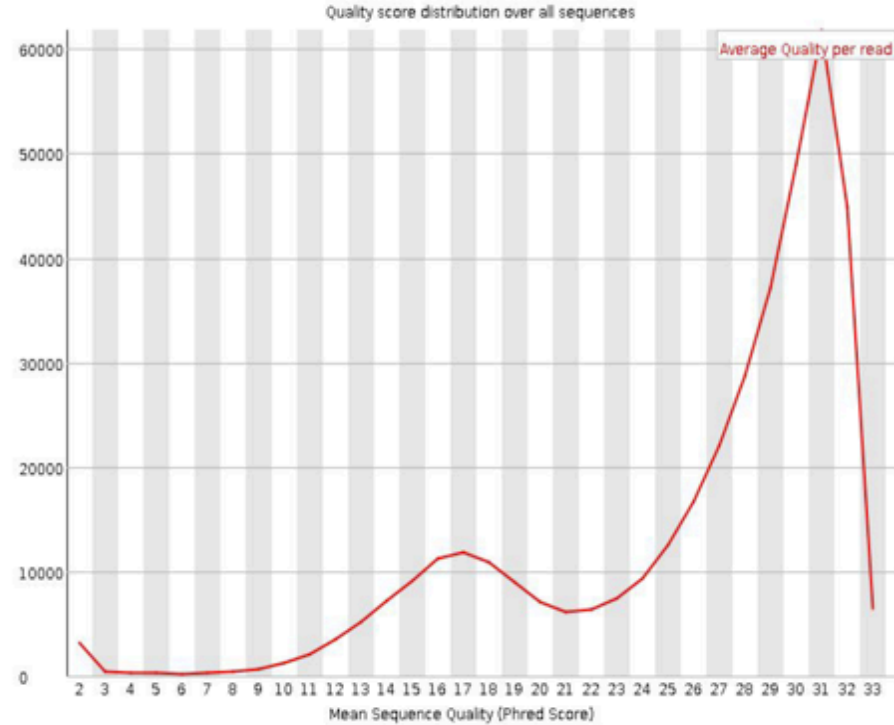
FastQC report – per sequence quality scores

Examines the Phred quality scores

✔ Per sequence quality scores



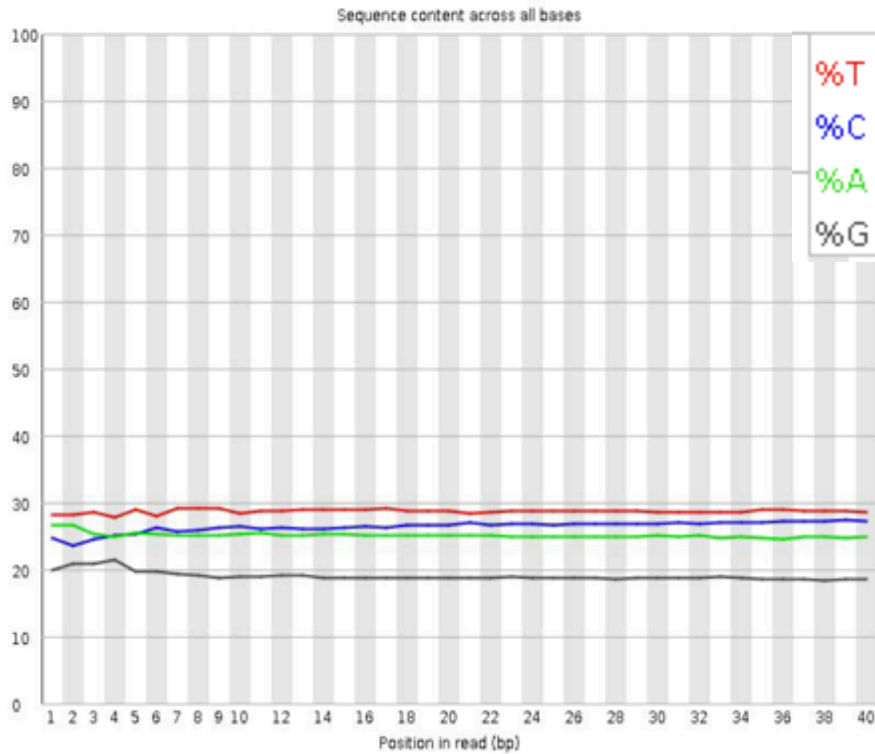
✔ Per sequence quality scores



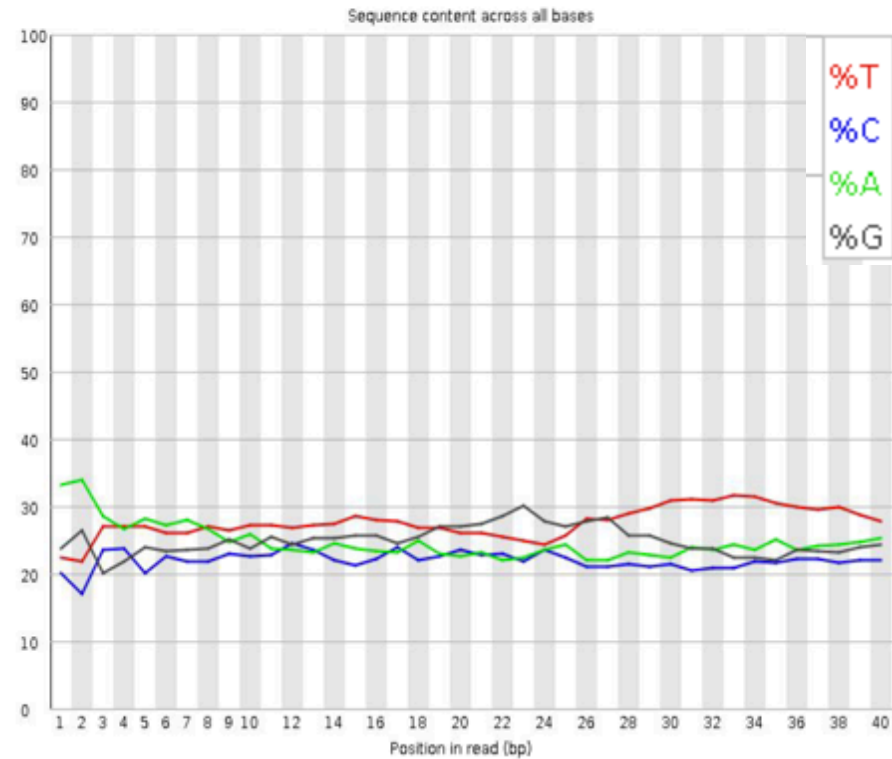
FastQC report – per base sequence content

Examines the sequence base content

✔ Per base sequence content



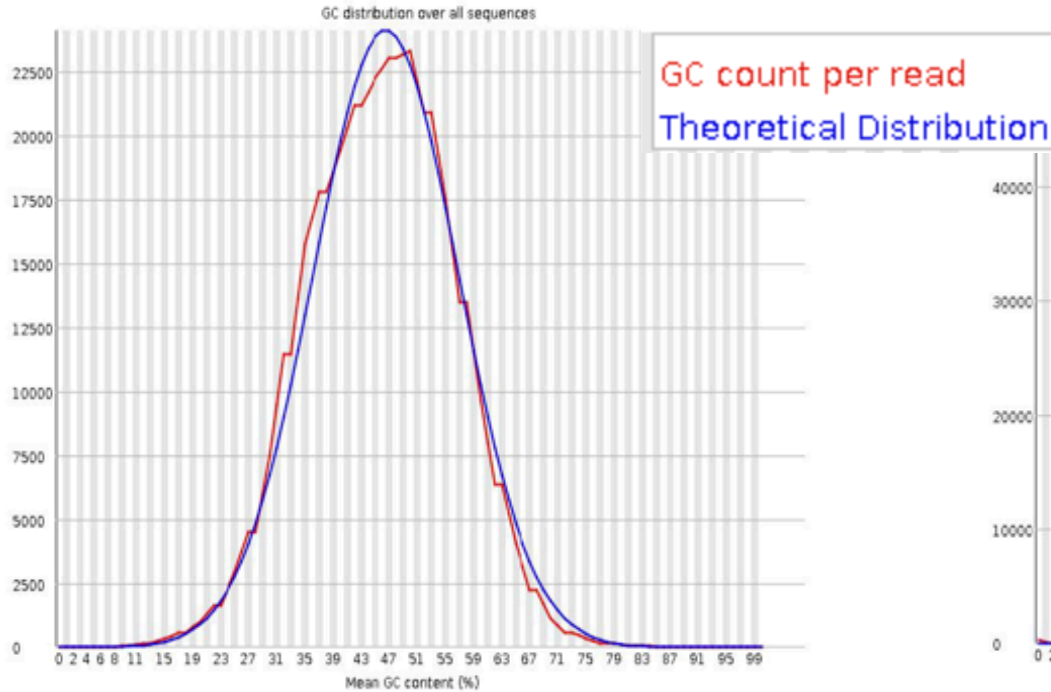
! Per base sequence content



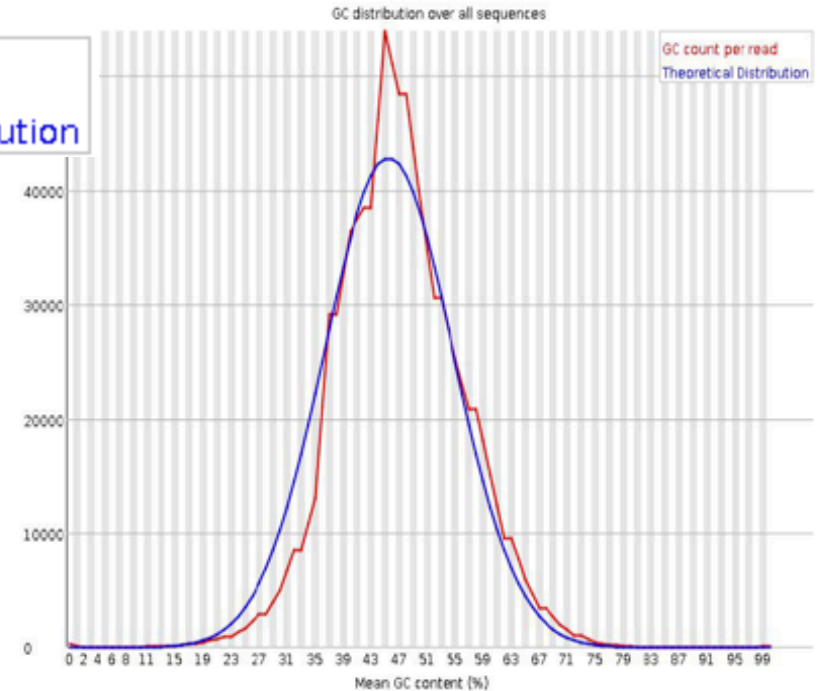
FastQC report – per sequence GC content

Examines the sequence base content

✔ Per sequence GC content



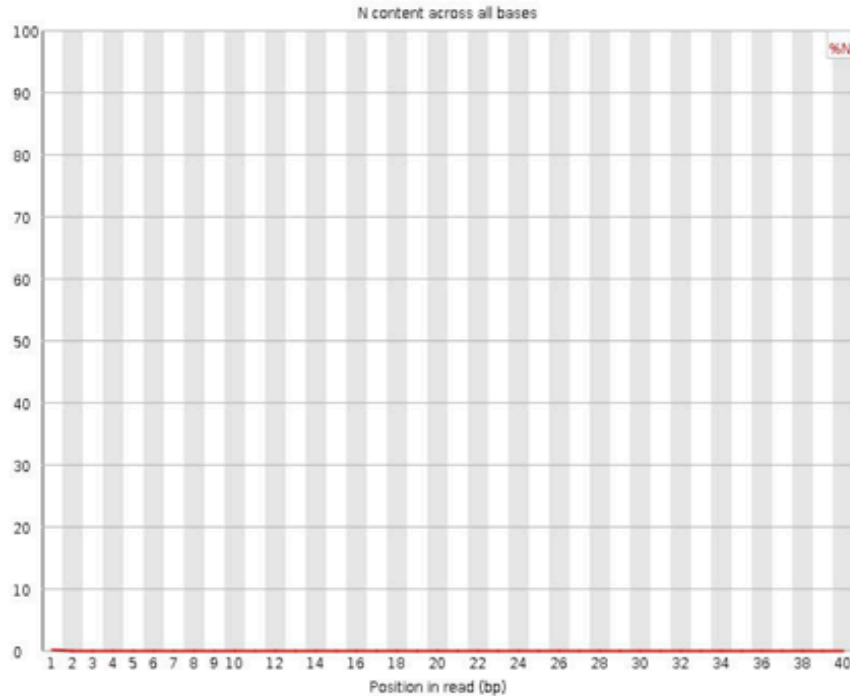
⚠ Per sequence GC content



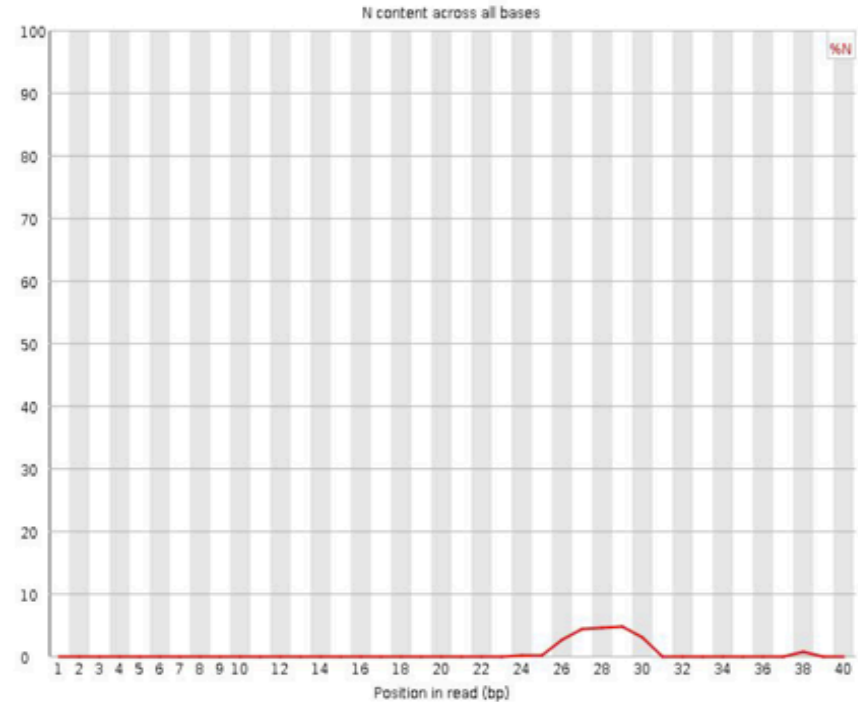
FastQC report – per base N content

Examines the sequence base content

✔ Per base N content



✔ Per base N content

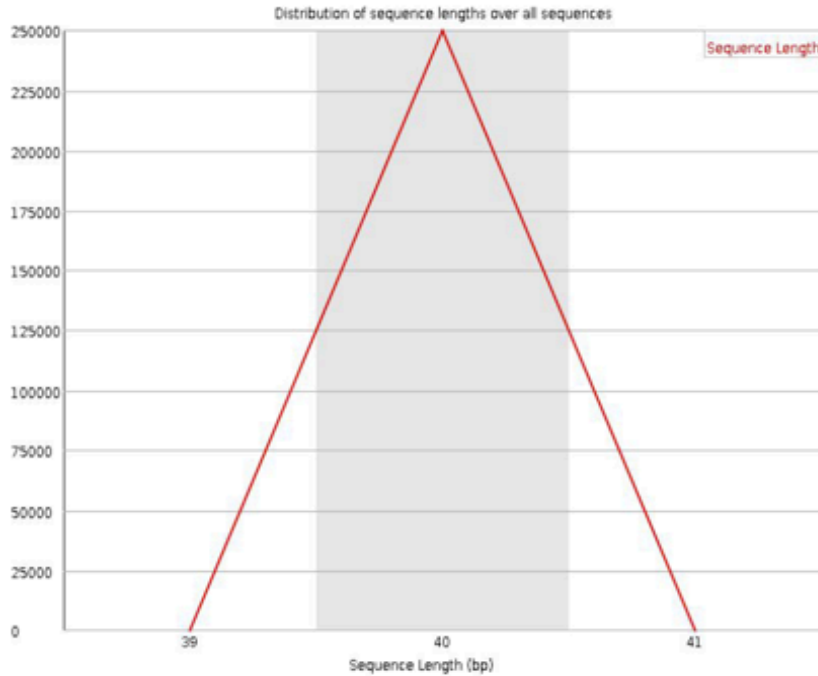


N is a base that could not be called

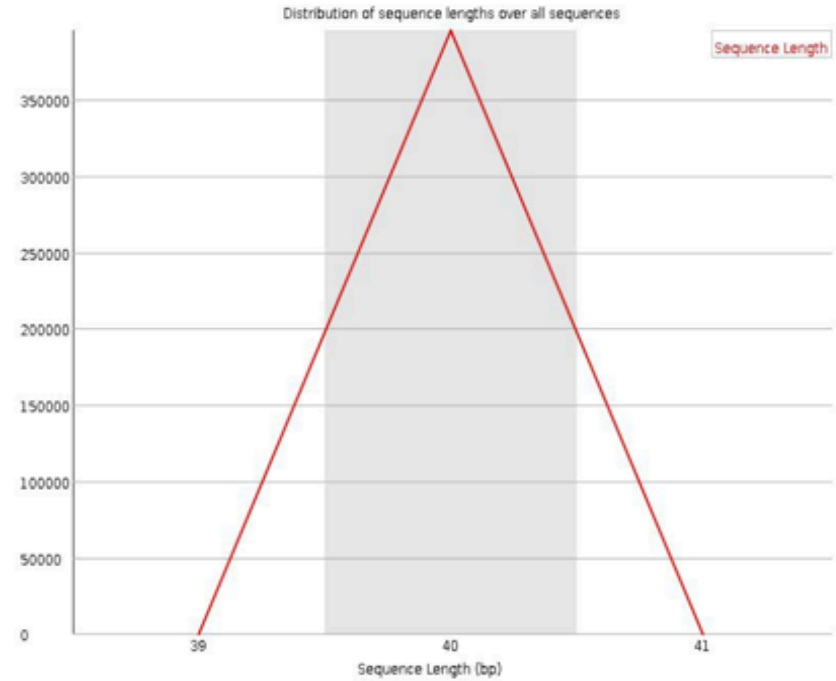
FastQC report – sequence length distribution

Examines the sequence length

✔ Sequence Length Distribution



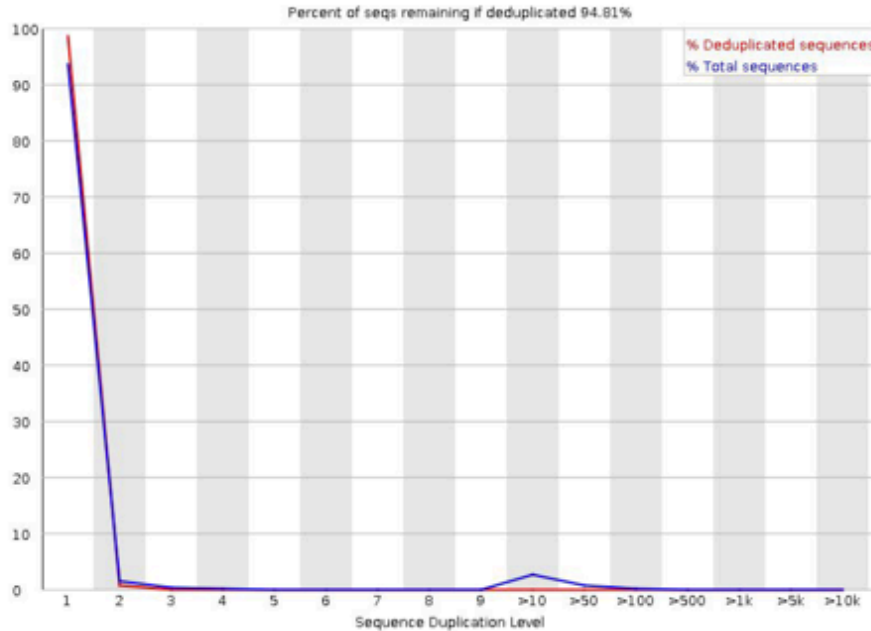
✔ Sequence Length Distribution



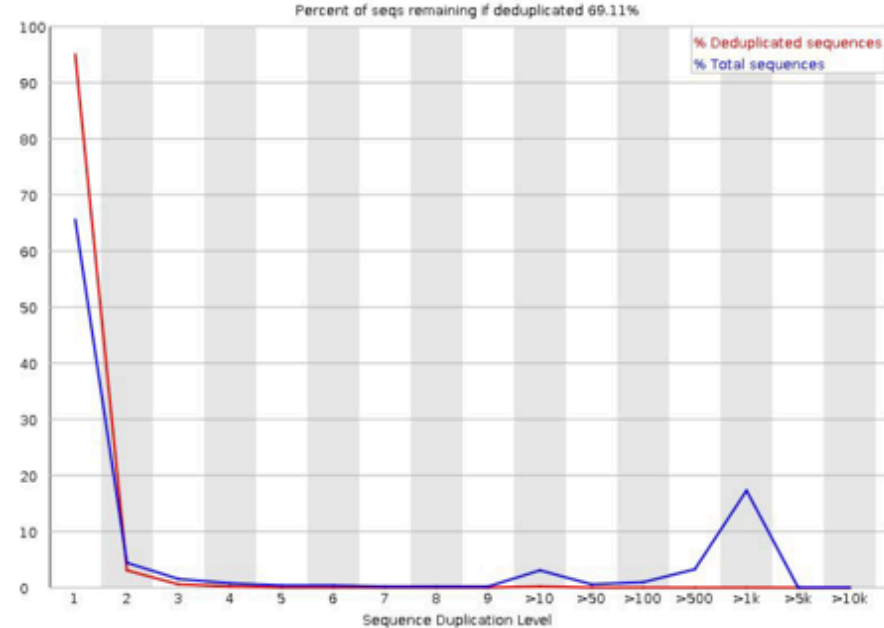
FastQC report – sequence duplication level

Examines potential unwanted sequences

✔ Sequence Duplication Levels



⚠ Sequence Duplication Levels



Percentage of reads of a given sequence which are present a given number of times in the file.

FastQC report – overrepresented sequences

 **Overrepresented sequences**
No overrepresented sequences

Examines potential unwanted sequences

 **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
AGAGTTT AT CGCTT CCAT GACGCAGAAGTT AACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGT AT CCAACCT GCAGAGTTT AT CGCTT CCAT G	2047	0.5178502762542754	No Hit
ATTGGCGT AT CCAACCT GCAGAGTTT AT CGCTT CCAT GA	2014	0.5095019327680071	No Hit
CGAT AAAAAT GATT GGCGT AT CCAACCT GCAGAGTTT AT	1913	0.4839509420979134	No Hit
GT AT CCAACCT GCAGAGTTT AT CGCTT CCAT GACGCAGA	1879	0.47534961850600066	No Hit
AAAAAT GATT GGCGT AT CCAACCT GCAGAGTTT AT CGCT	1846	0.4670012750197325	No Hit
TGATT GGCGT AT CCAACCT GCAGAGTTT AT CGCTT CCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTT AT CGCTT CCAT GACGCAGAAGTT AA	1836	0.46447147396328753	No Hit
GAT AAAAAT GATT GGCGT AT CCAACCT GCAGAGTTT AT C	1831	0.4632065734350651	No Hit
AAATGATT GGCGT AT CCAACCT GCAGAGTTT AT CGCTT C	1779	0.45005160794155147	No Hit
ATGATT GGCGT AT CCAACCT GCAGAGTTT AT CGCTT CCA	1779	0.45005160794155147	No Hit
AATGATT GGCGT AT CCAACCT GCAGAGTTT AT CGCTT CC	1760	0.4452449659343061	No Hit
AAAATGATT GGCGT AT CCAACCT GCAGAGTTT AT CGCTT	1729	0.4374026026593269	No Hit
CGT AT CCAACCT GCAGAGTTT AT CGCTT CCAT GACGCAG	1713	0.43335492096901496	No Hit
AT CCAACCT GCAGAGTTT AT CGCTT CCAT GACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTT AT CGCTT CCAT GACGCAGAAGTT AACACTT	1684	0.42601849790532476	No Hit
TGCAGAGTTT AT CGCTT CCAT GACGCAGAAGTT AACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTT AT CGCTT CCAT GACGCAGAAGTT A	1668	0.4219708162150128	No Hit
TAT CCAACCT GCAGAGTTT AT CGCTT CCAT GACGCAGAA	1630	0.4123575722005221	No Hit
CGGTT CAGCAGGAAT GCCGAGAT CGGAAGAGCGGTT CAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCCCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAAGTTATATGCTGGGGGTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCCCG	532	0.13458541620289598	No Hit
CTGCTCATGGAAGCGATAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCCCG	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAAGTTATATGCTGGGGGTTTTTTT	411	0.10397482341988626	No Hit

A sequence is considered overrepresented if it accounts for $\geq 0.1\%$ of the total reads.

FastQC report – overrepresented sequences

Examines potential unwanted sequences

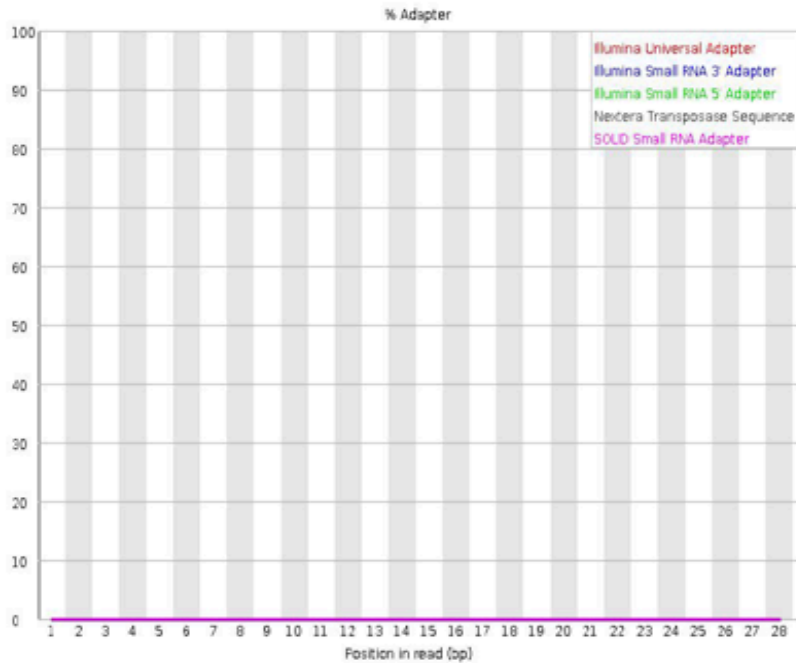
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGC	4156	0.20779999999999998	TruSeq Adapter, Index 6 (100% over 50bp)
TT	3490	0.1745	No Hit

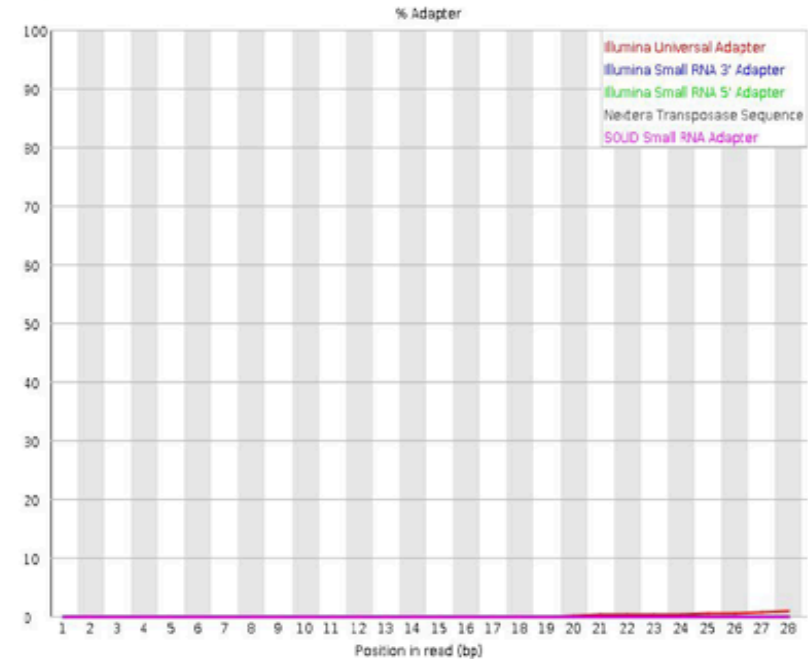
FastQC report – adapter content

Examines potential unwanted sequences

Adapter Content



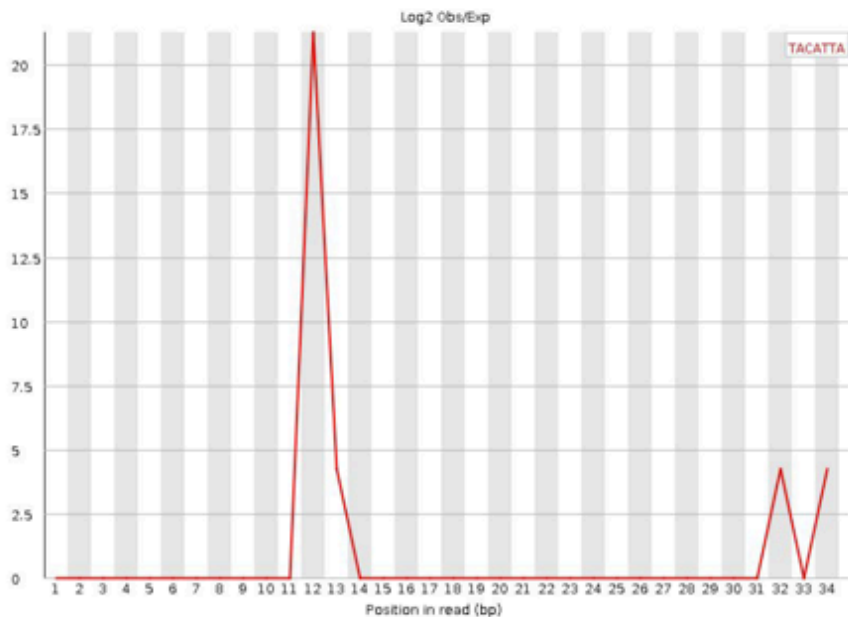
Adapter Content



Cumulative plot of the percentage of reads where an adapter sequence has been identified at the indicated base position.

FastQC report – Kmer content

⚠ Kmer Content

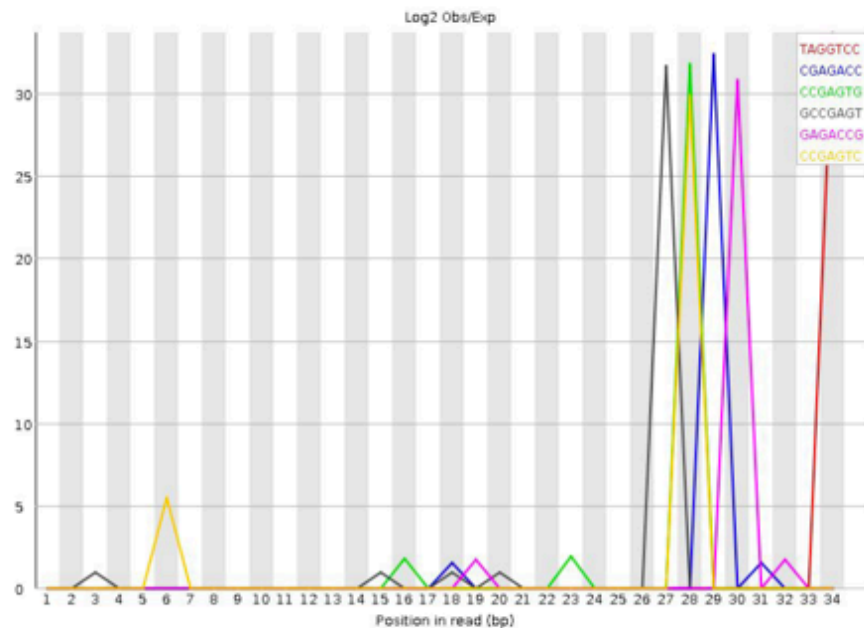


Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TACATTA	40	0.003151852	21.2465	12

By default not displayed in the report
Measures the count of each k-mer (by default k = 7) at each position along the read

⚠ Kmer Content

Examines potential unwanted sequences

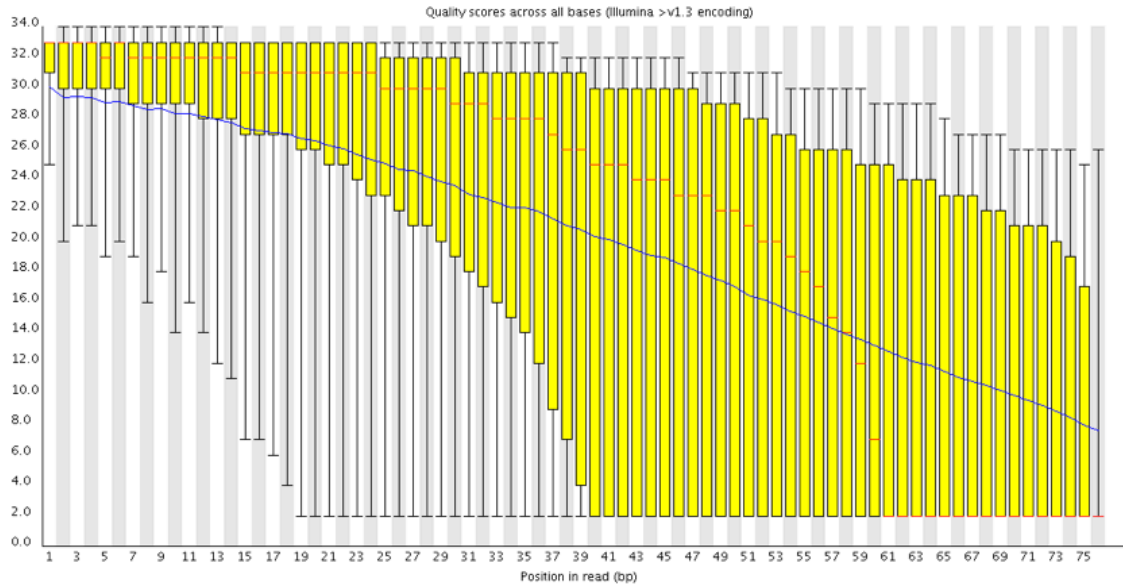


Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TAGGTCC	30	1.5992917E-5	33.6211	34
CGAGACC	105	0.0	32.37975	29
CCGAGTG	90	0.0	31.803032	28
GCCGAGT	170	0.0	31.625078	27
GAGACCG	95	0.0	30.826315	30
CCGAGTC	30	4.3762376E-4	29.815344	28

A common quality problem

A drop in sequence quality towards the 3' end of the read is common

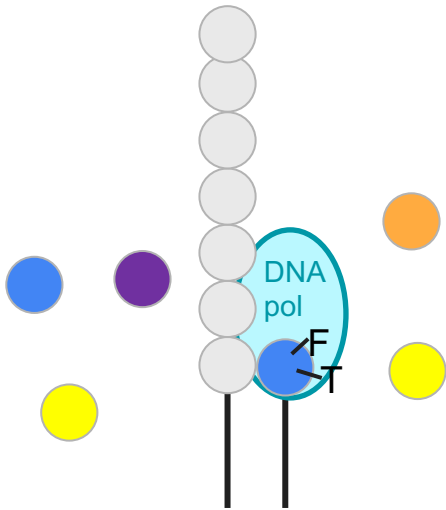
Normally, you can trim the reads so that you just keep the parts that have a Phred score > 20



A common quality problem

This drop in sequence quality is often caused by **phasing**

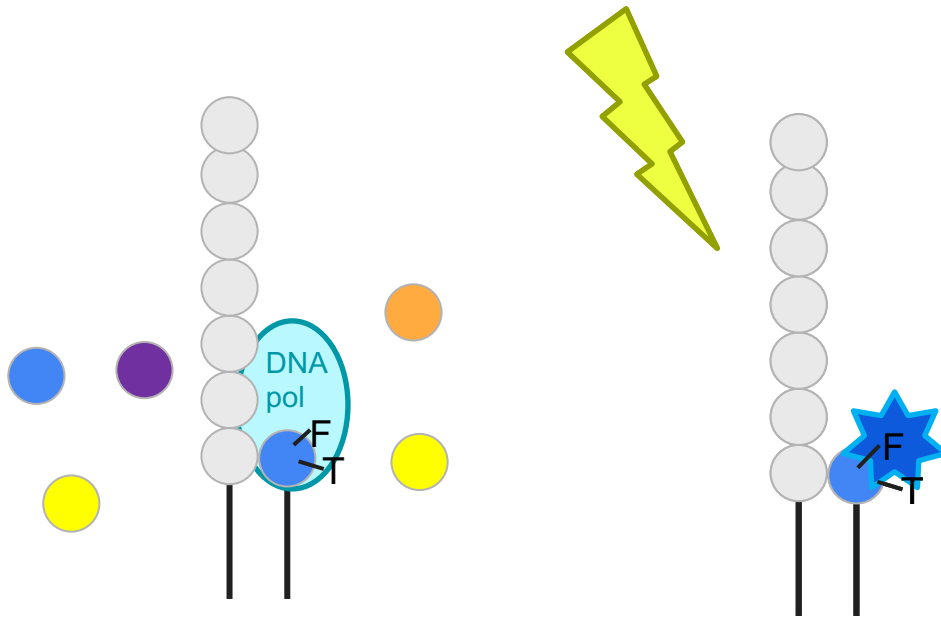
Normal Illumina sequencing by synthesis:



A common quality problem

This drop in sequence quality is often caused by **phasing**

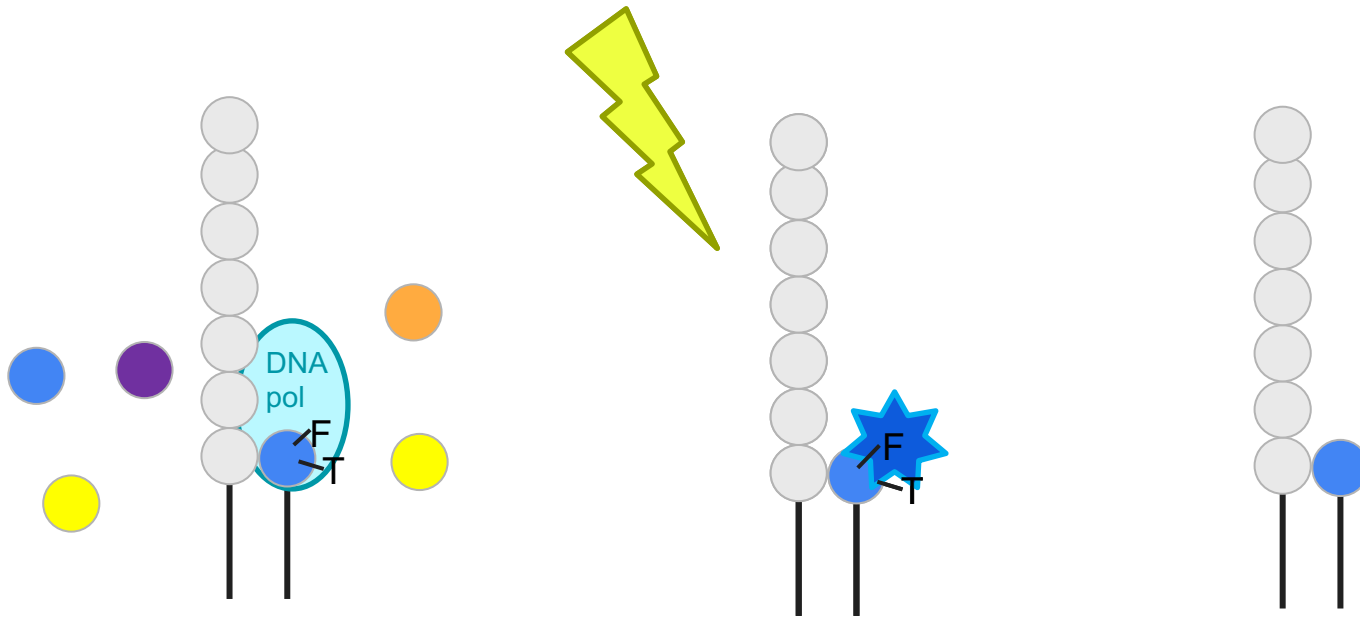
Normal Illumina sequencing by synthesis:



A common quality problem

This drop in sequence quality is often caused by **phasing**

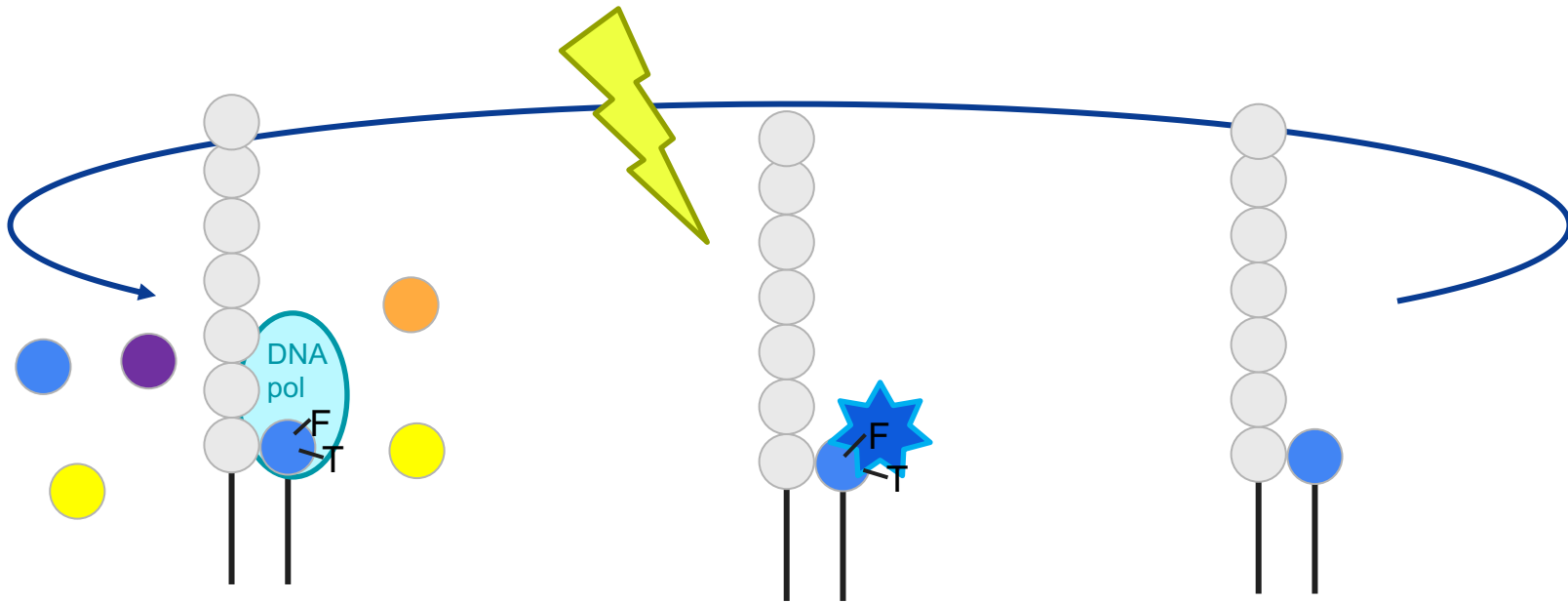
Normal Illumina sequencing by synthesis:



A common quality problem

This drop in sequence quality is often caused by **phasing**

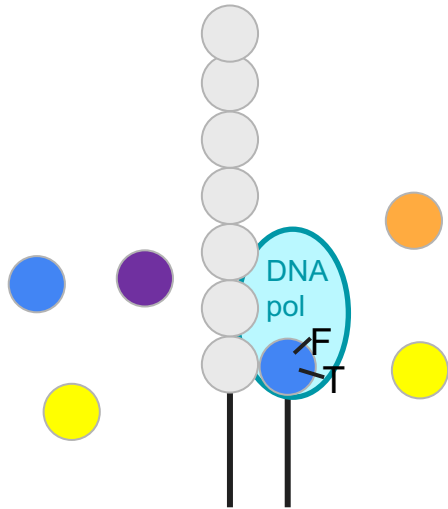
Normal Illumina sequencing by synthesis:



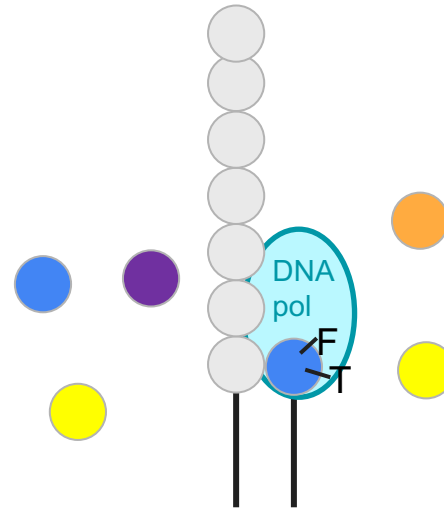
A common quality problem

This drop in sequence quality is often caused by **phasing**

Sometimes, terminators are not properly removed:



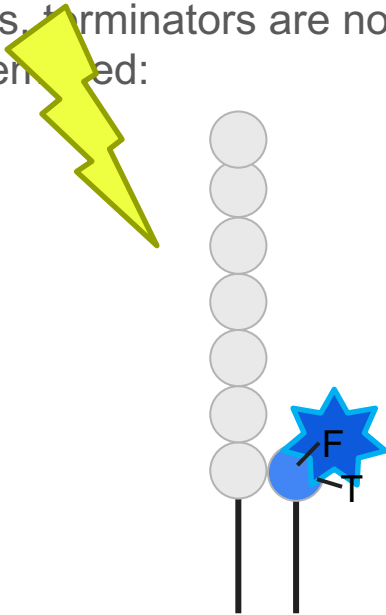
Correct removal of terminators:



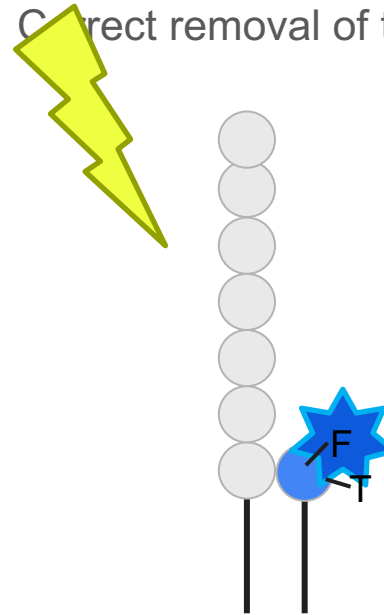
A common quality problem

This drop in sequence quality is often caused by **phasing**

Sometimes, terminators are not properly removed:



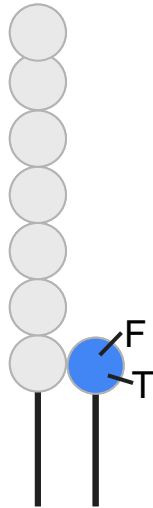
Correct removal of terminators:



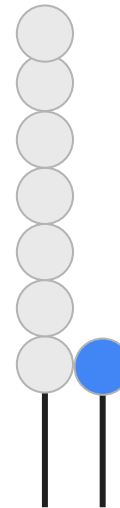
A common quality problem

This drop in sequence quality is often caused by **phasing**

Sometimes, terminators are not properly removed:



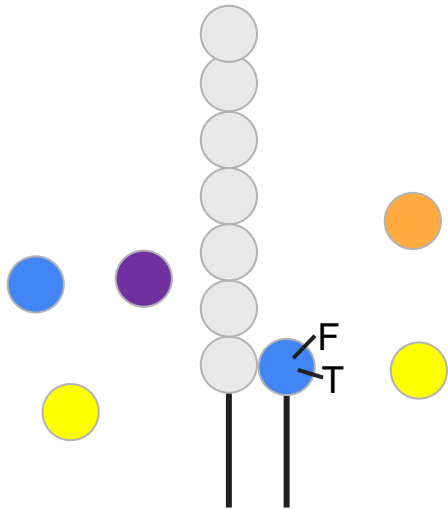
Correct removal of terminators:



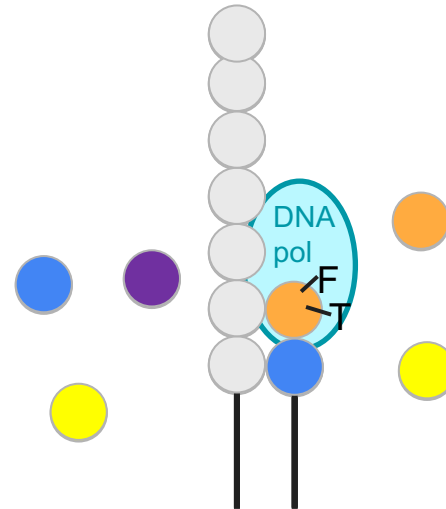
A common quality problem

This drop in sequence quality is often caused by **phasing**

Sometimes, terminators are not properly removed:



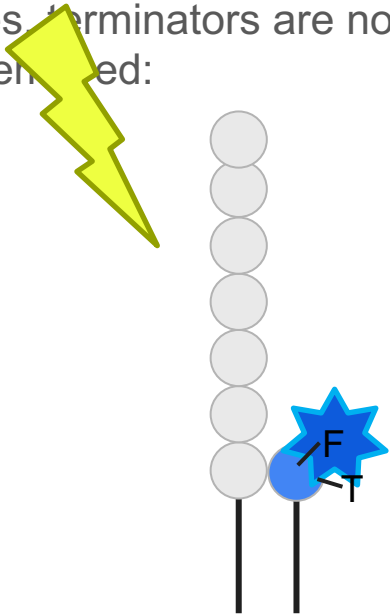
Correct removal of terminators:



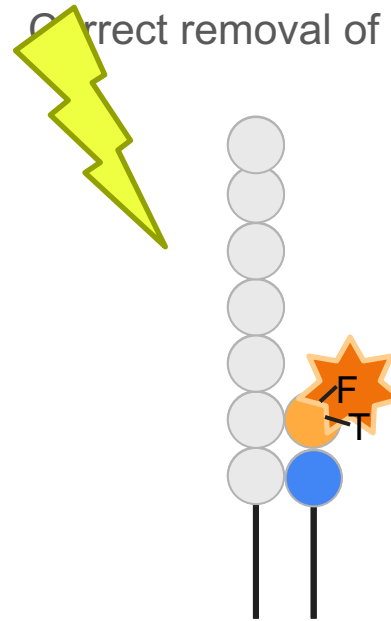
A common quality problem

This drop in sequence quality is often caused by **phasing**

Sometimes terminators are not properly removed:



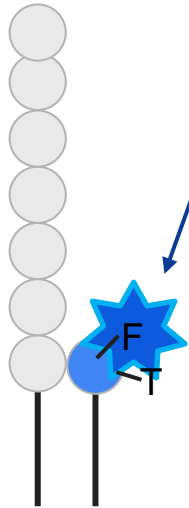
Correct removal of terminators:



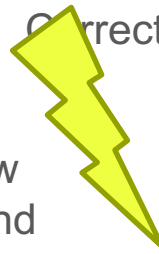
A common quality problem

This drop in sequence quality is often caused by **phasing**

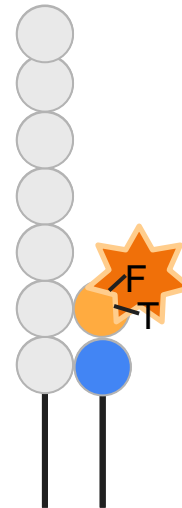
Sometimes terminators are not properly removed:



This DNA fragment is now one cycle behind the rest. It is **out of phase**



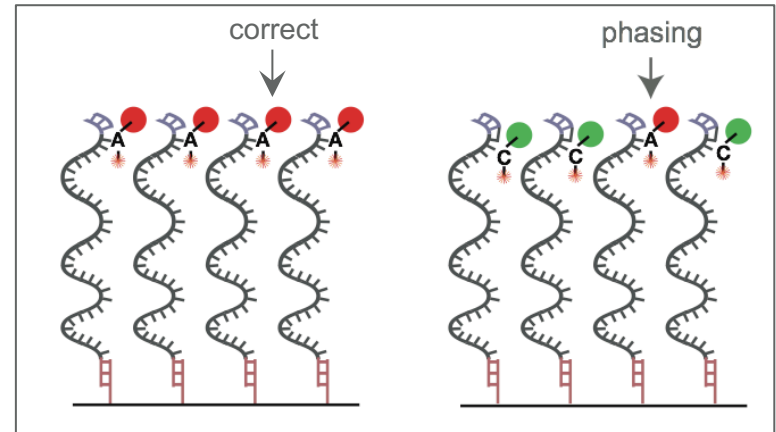
Correct removal of terminators:



A common quality problem

This drop in sequence quality is often caused by **phasing**

- This fragment will pollute the light signal that the sequencer's camera has to read
- Over time, with increasing read length, they add up and pollute the light signal more and more, leading to lower and lower quality scores
- Defect terminator caps can also cause a similar effect, where two nucleotides can bind in one cycle (called prephasing)

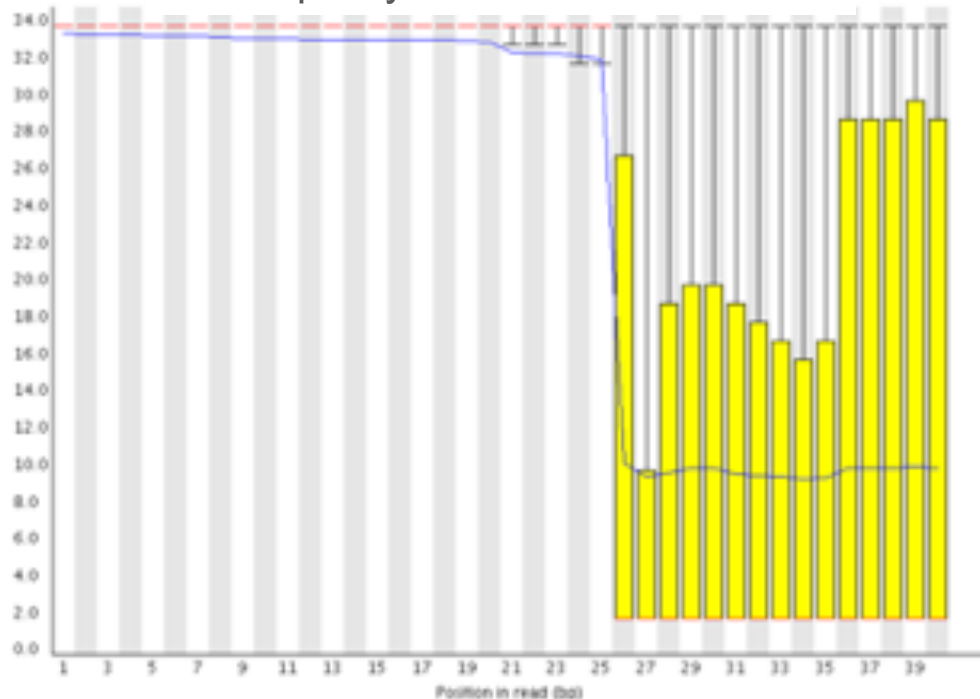


Artifact removal

So what do we do if the quality isn't good enough?

Often we need to remove the bad parts

Poor quality bases at read ends:

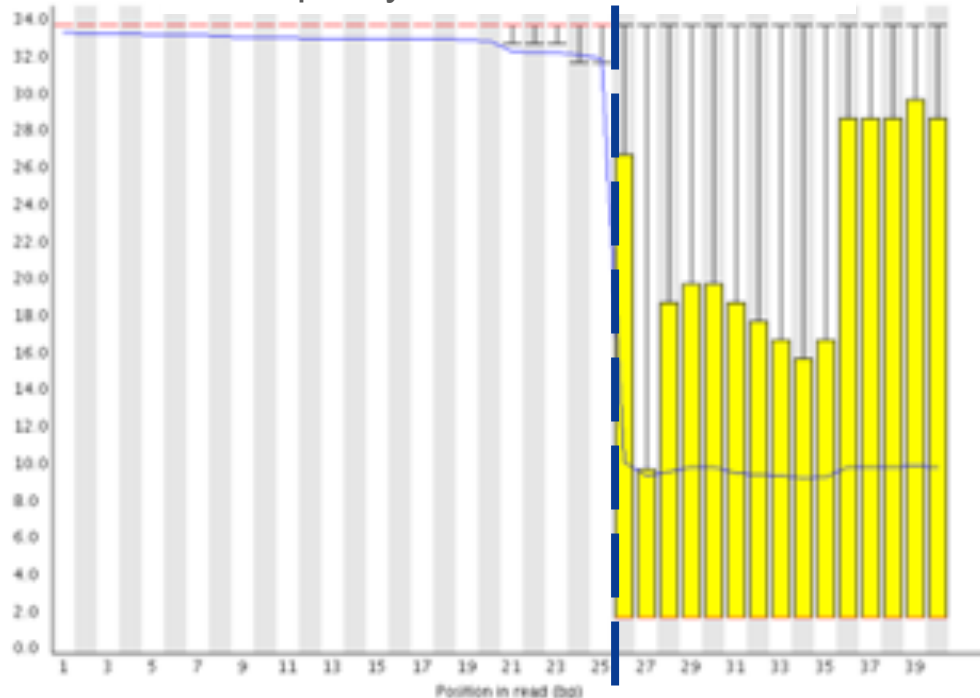


Artifact removal

So what do we do if the quality isn't good enough?

Often we need to remove the bad parts

Poor quality bases at read ends:



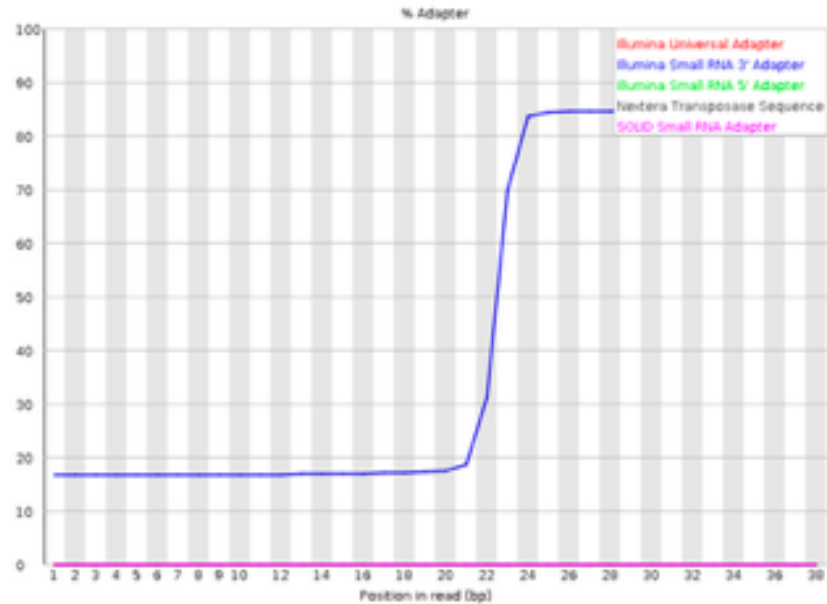
Artifact removal

So what do we do if the quality isn't good enough?

Often we need to remove the bad parts

Leftover adapter sequences:

Adapter Content

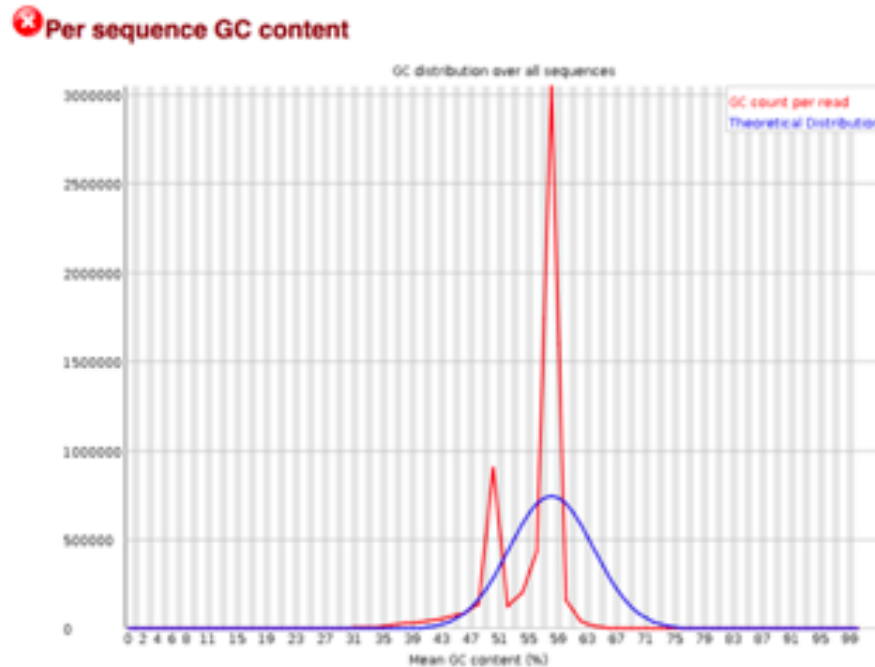


Artifact removal

So what do we do if the quality isn't good enough?

Often we need to remove the bad parts

Known contaminants:



Artifact removal

So what do we do if the quality isn't good enough?

Often we need to remove the bad parts

In the practical, we will use **Cutadapt** to perform quality trimming of our sample dataset

Practical time

Let's practice!