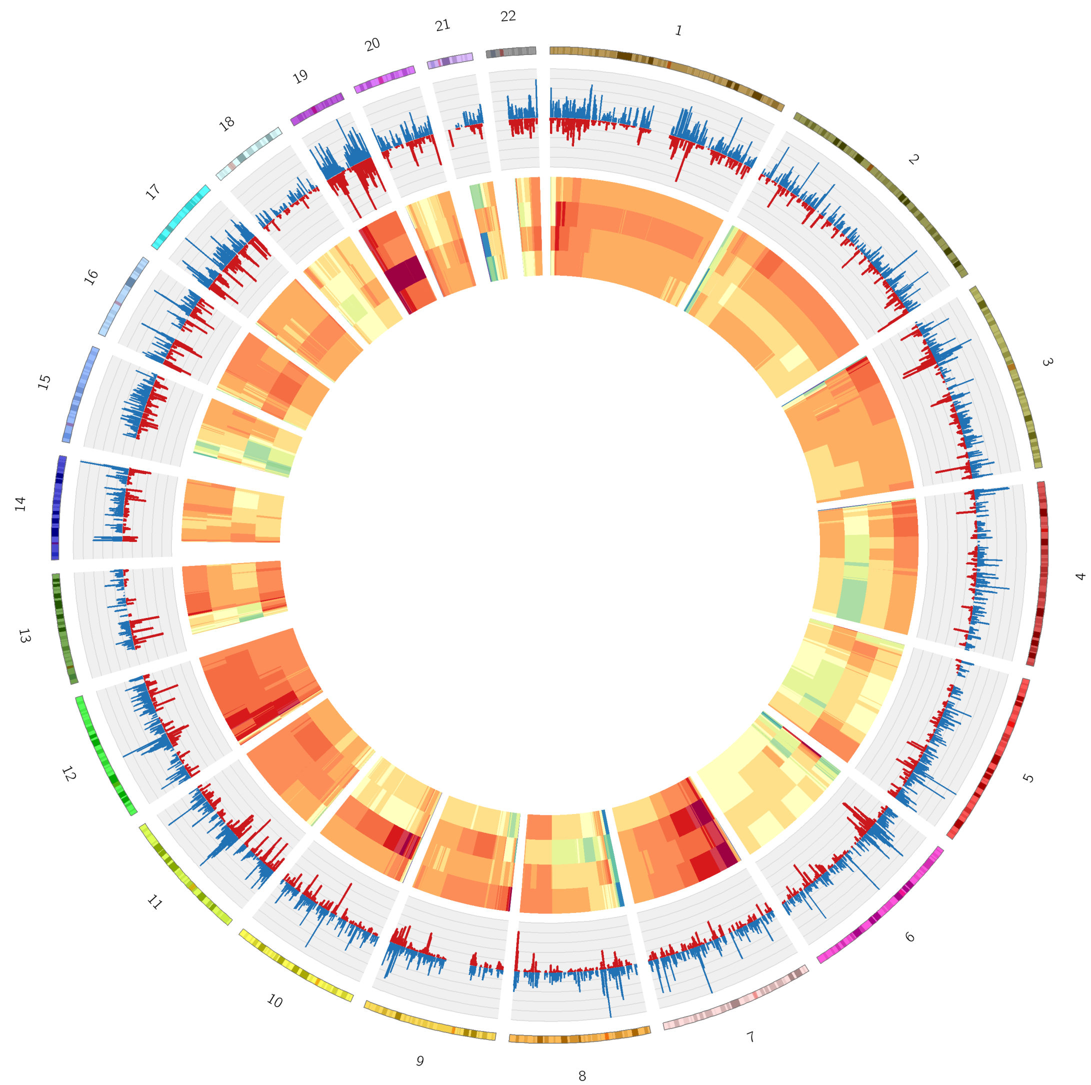


Quality control and artefact removal

Joanna Krupka

CRUK Summer School in Bioinformatics

Cambridge, July 2020

CANCER RESEARCH UK

MRC Cancer Unit

UNIVERSITY OF CAMBRIDGE

Why do we need quality control?

... Because sometimes things can go wrong

NGS sequencing generates highly accurate data, but can have few types of errors:

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

...

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
- Quick look at the potential problems with your experiment

Unaligned sequence: FASTQ

Quality scores come after the "+" line

Quality Q is proportional to -log10 probability of sequence base being wrong e

$$Q = -10 \cdot log_{10}(e)$$

```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAAGAATG
+
AAAFFJJJJJFJJ<J<FJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
```

Encoded in ASCII to save space:

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

Used in quality assessment and downstream analysis

Probability of incorrect base calls

Quality scores come after the "+" line

Quality Q is proportional to -log10 probability of sequence base being wrong e

$$Q = -10 \cdot log_{10}(e)$$

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%

https://hbctraining.github.io/Intro-to-rnaseq-hpc-orchestra/lessons/06_assessing_quality.html

FastQC - 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

Simple information about input FASTQ file: its name, type of quality score encoding, total number of reads, read length and GC content.

FastQC - summary

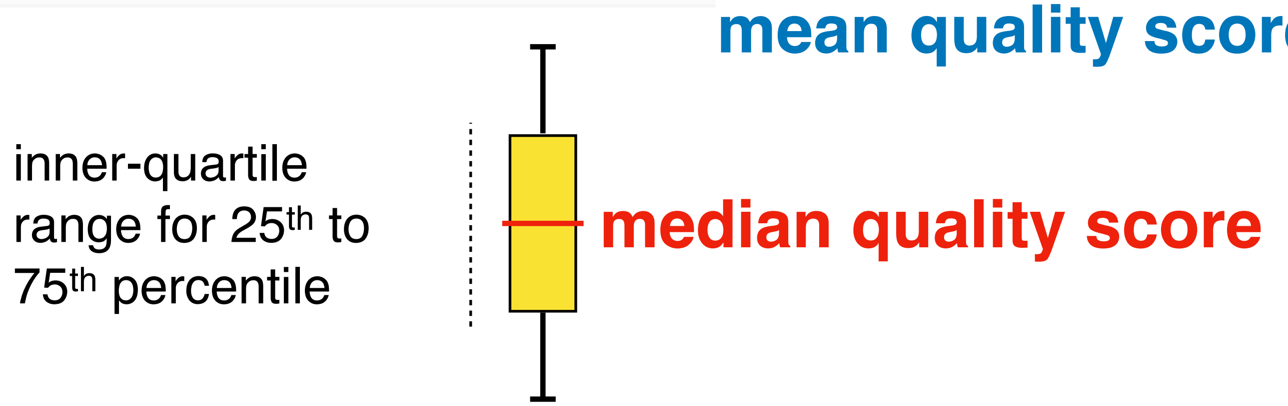
Summary

- 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

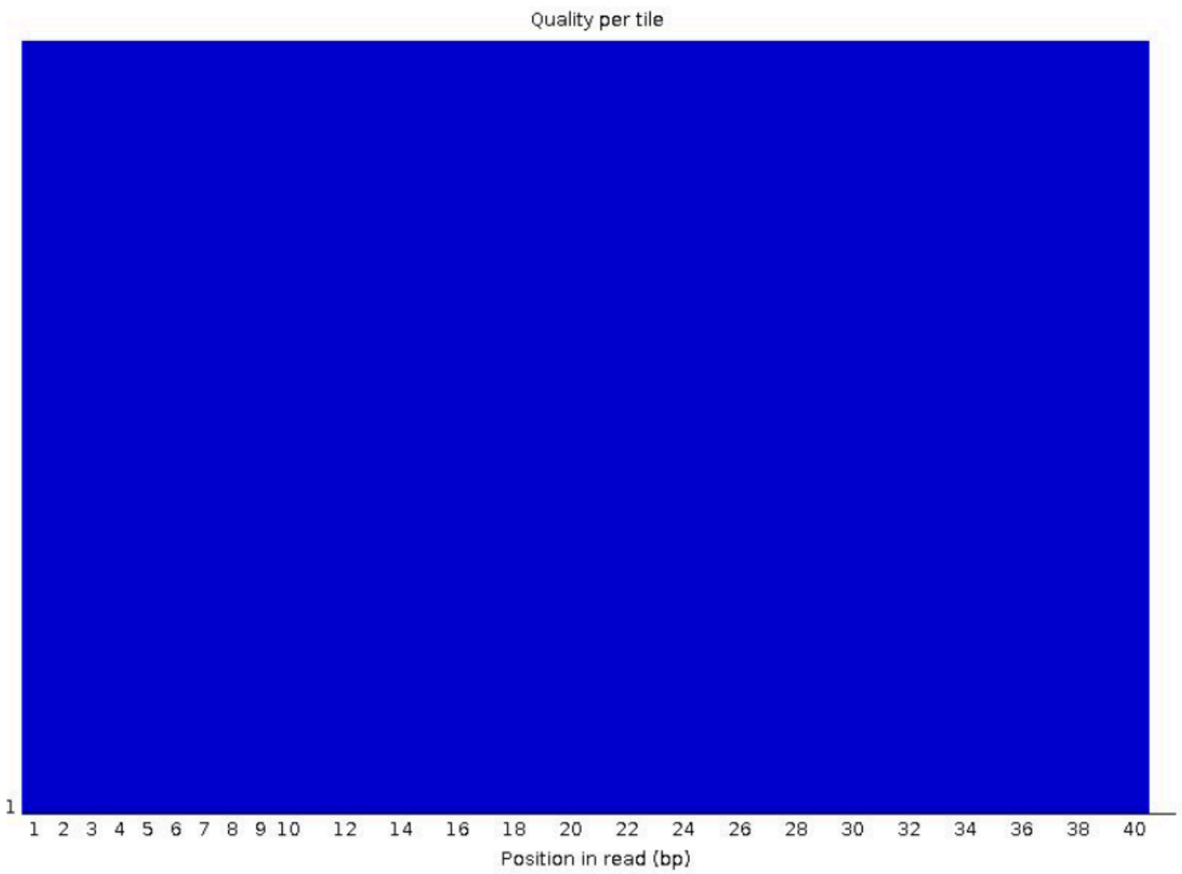
Summary

- 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

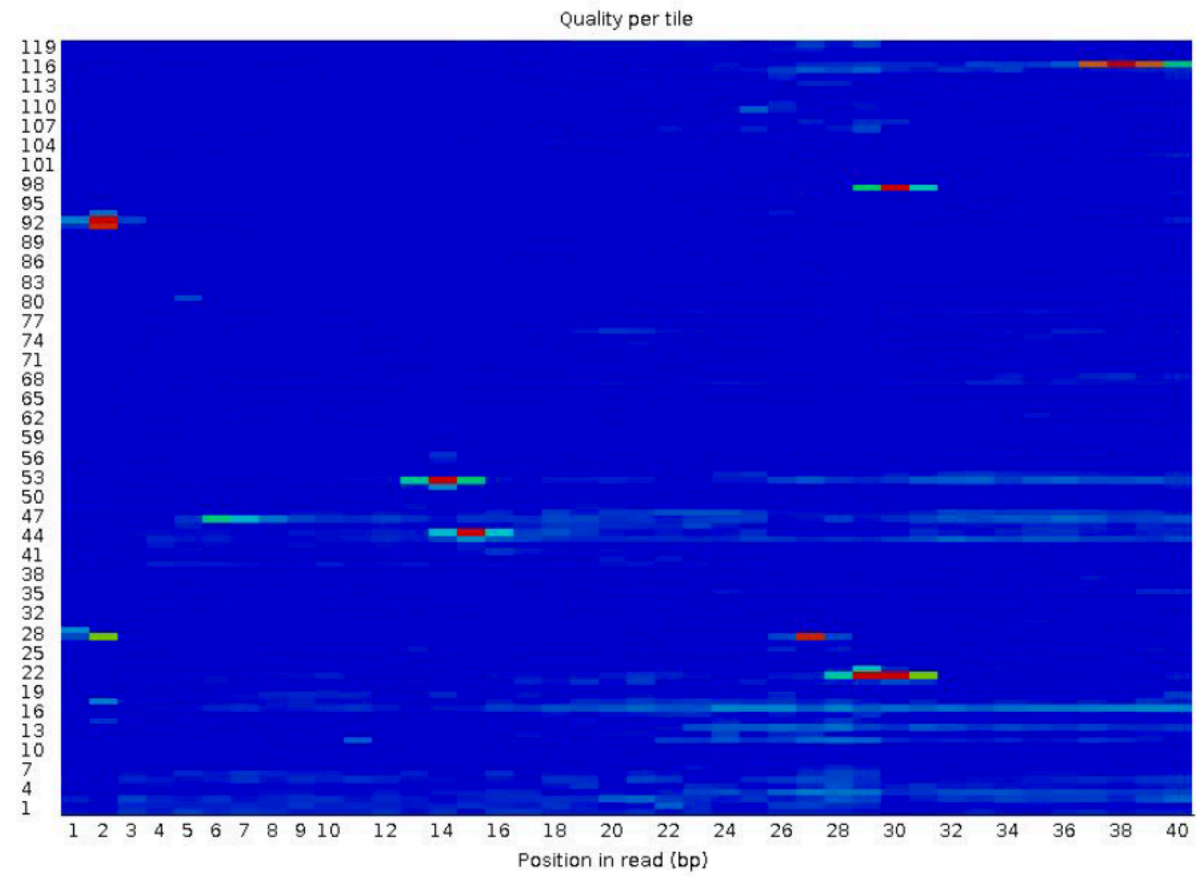
Per base sequence quality



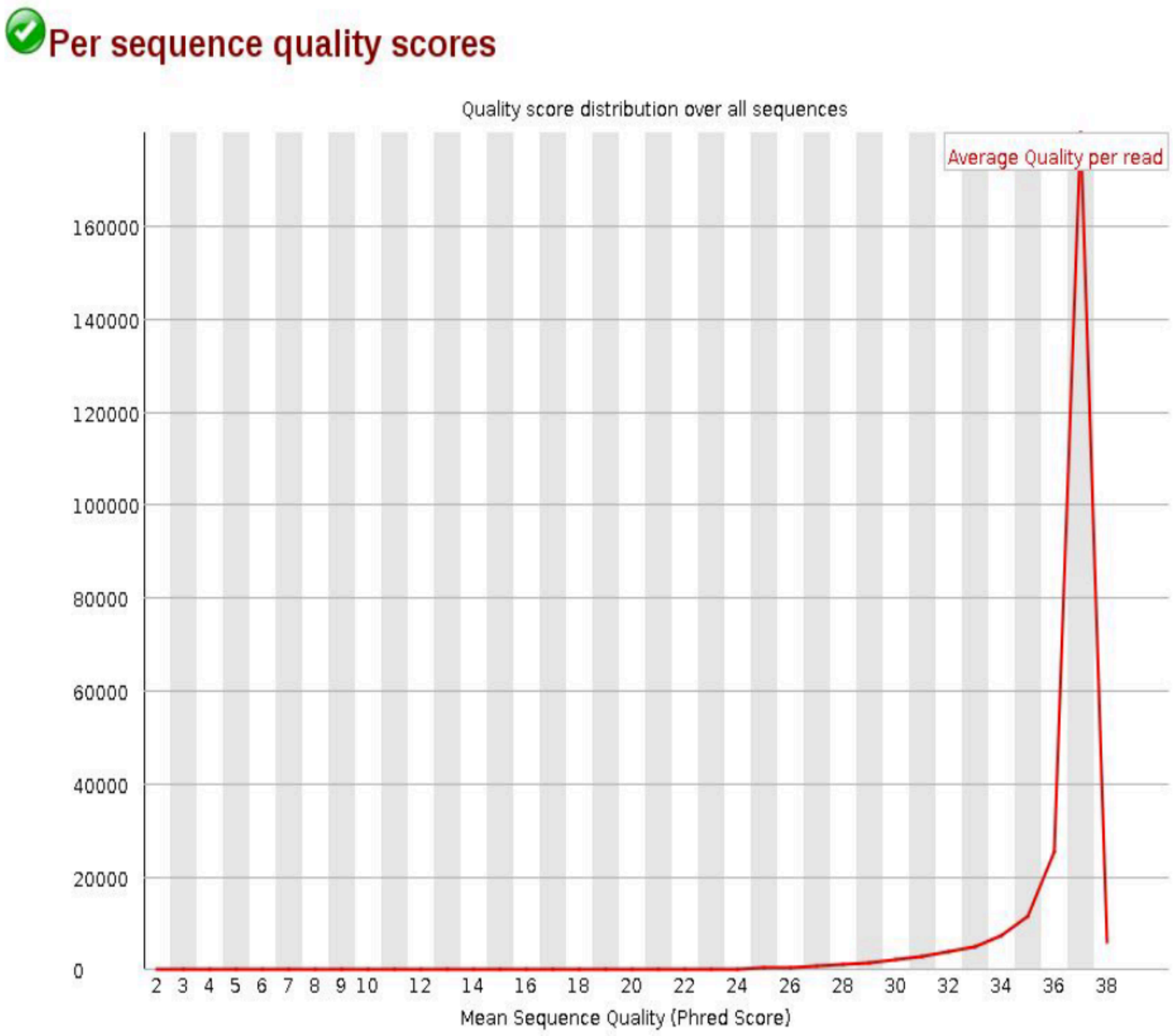
Per tile sequence quality

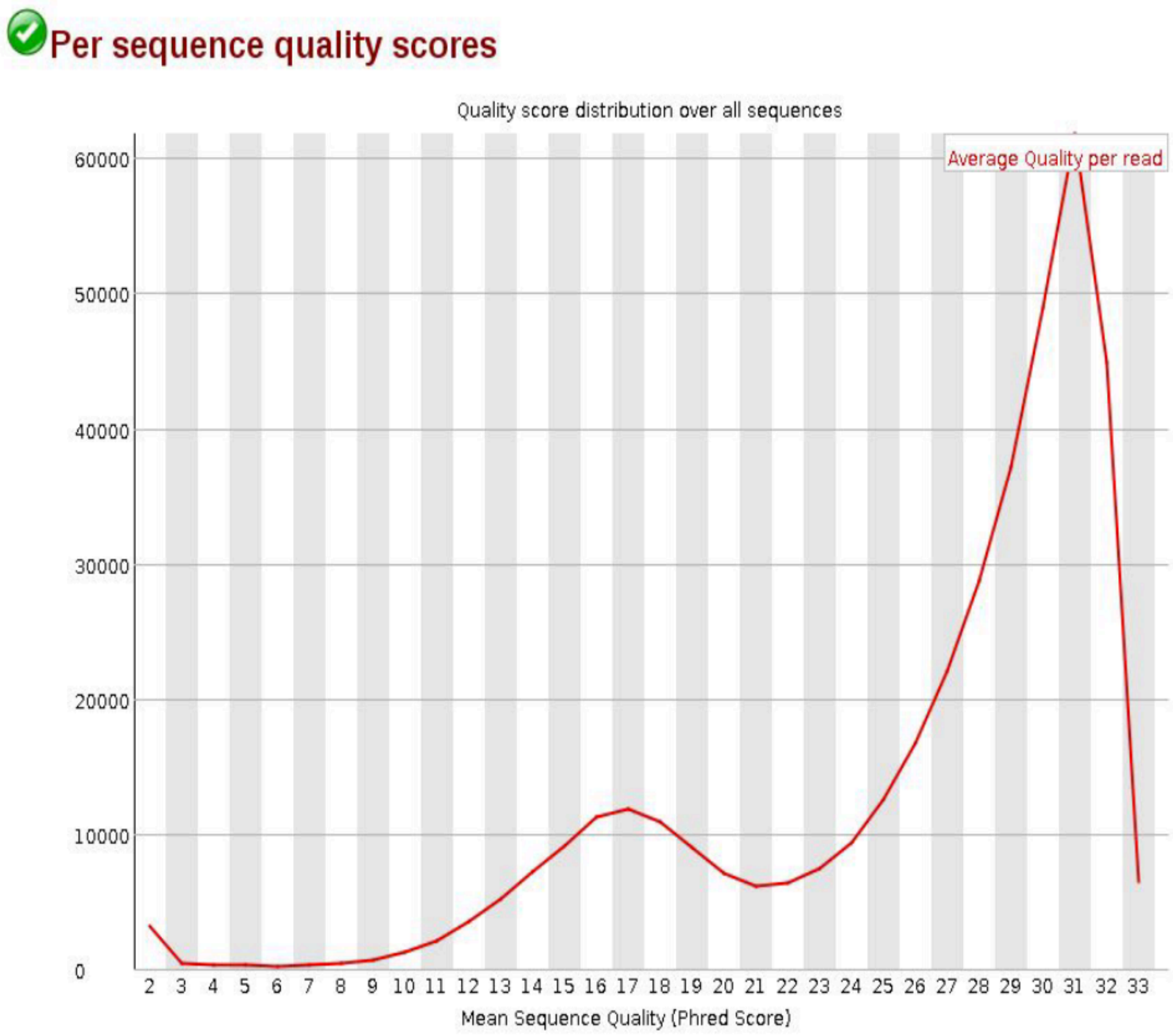


Per tile sequence quality



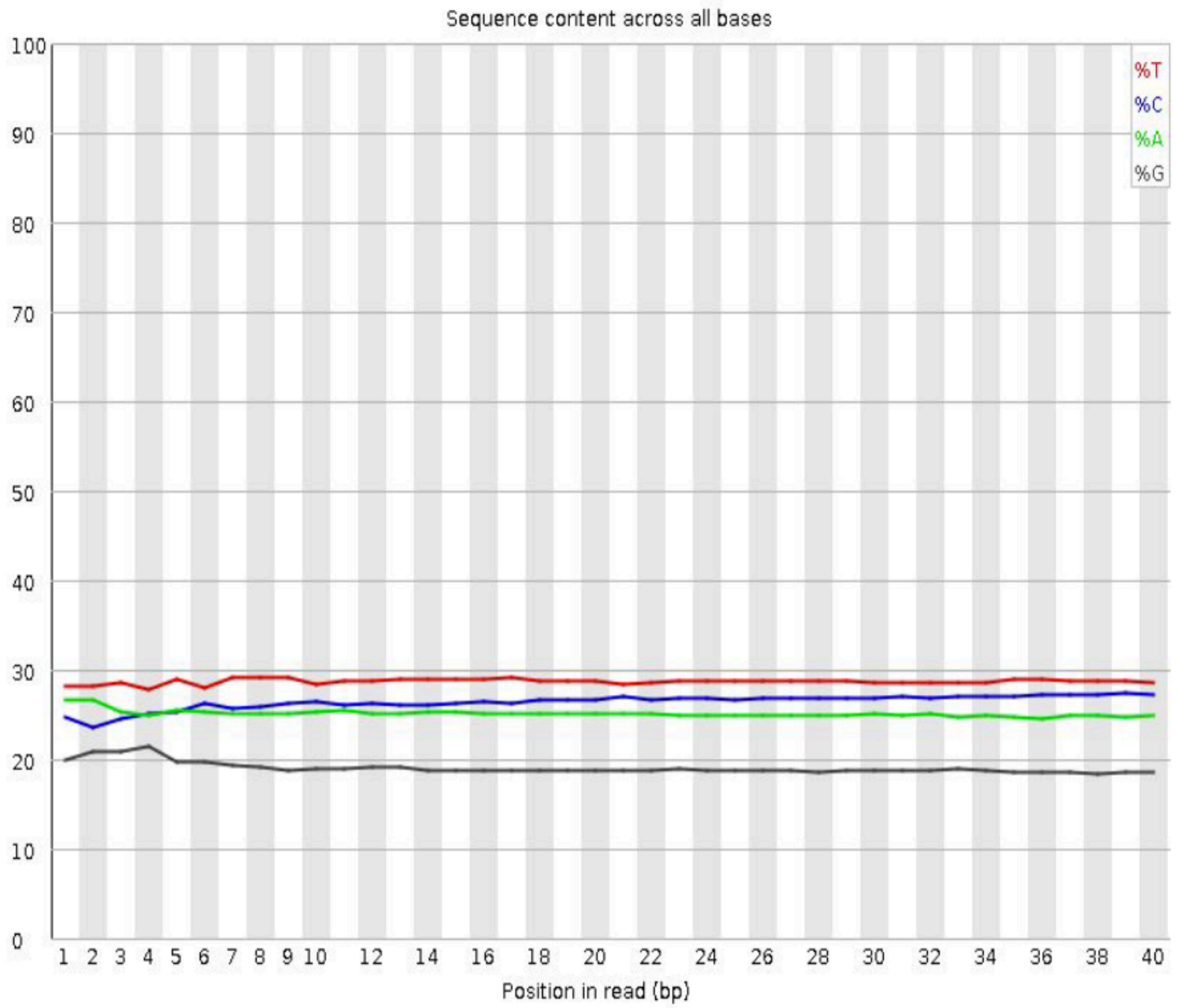
Per sequence quality scores



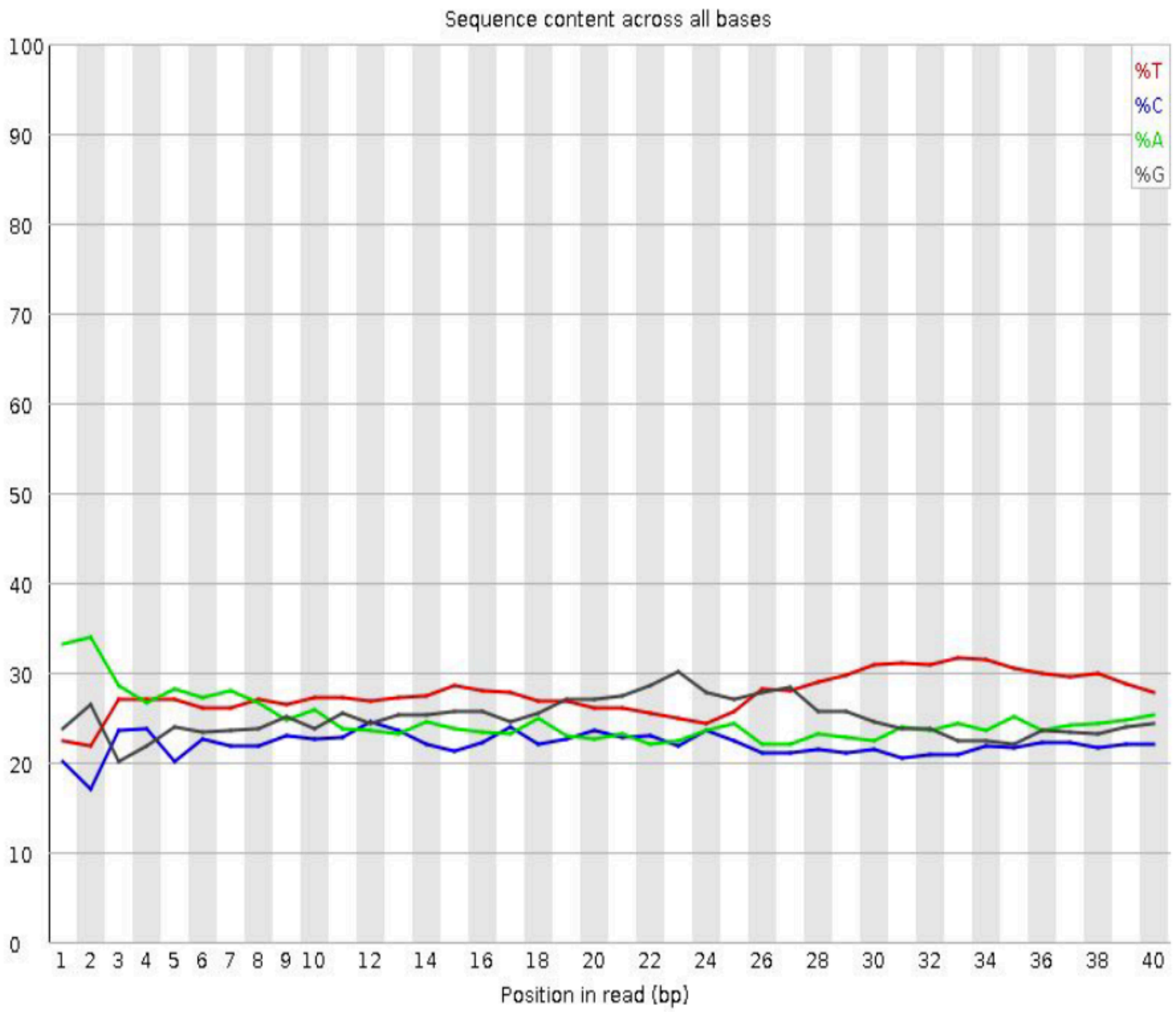


Per sequence content

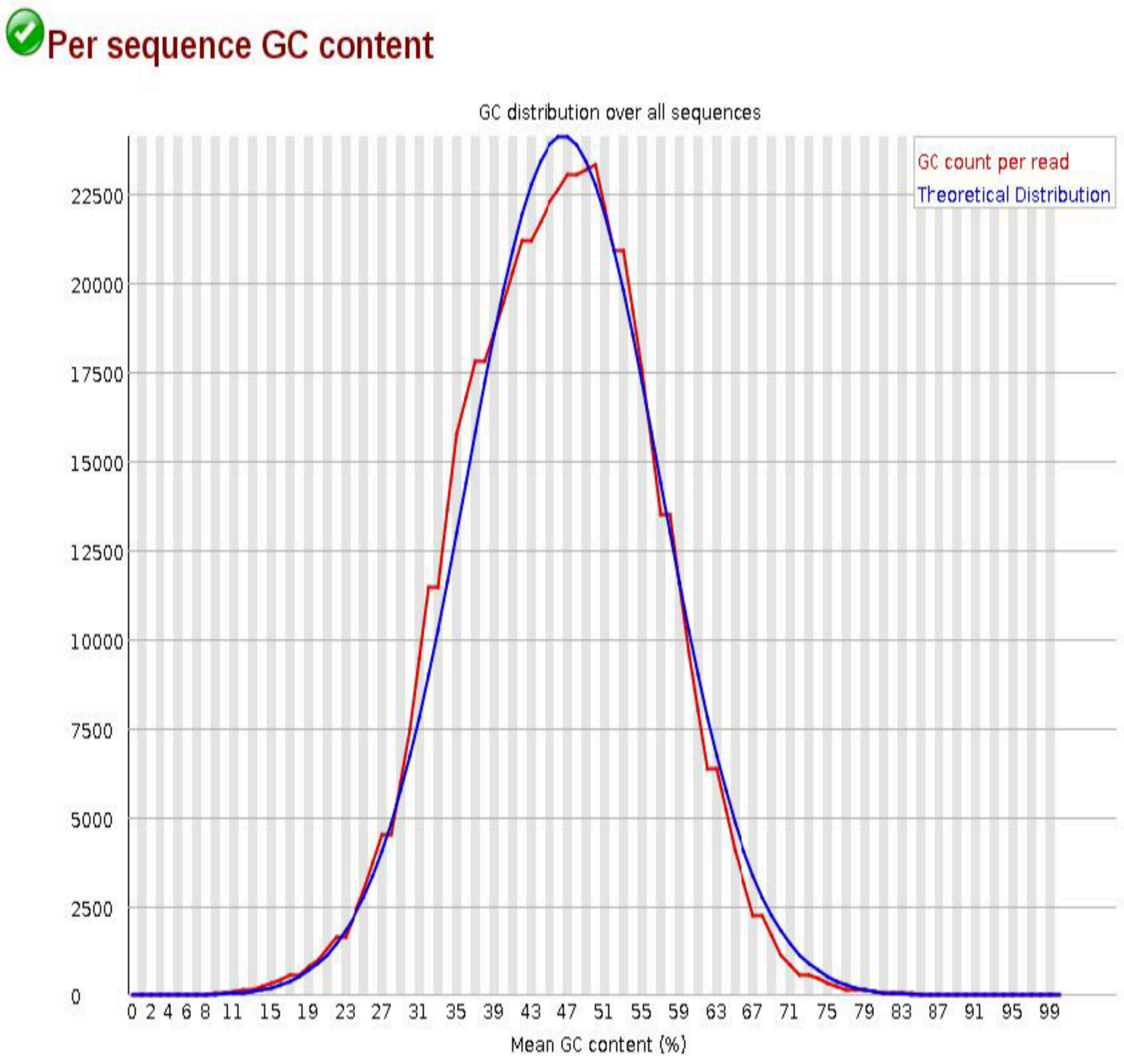
Per base sequence content

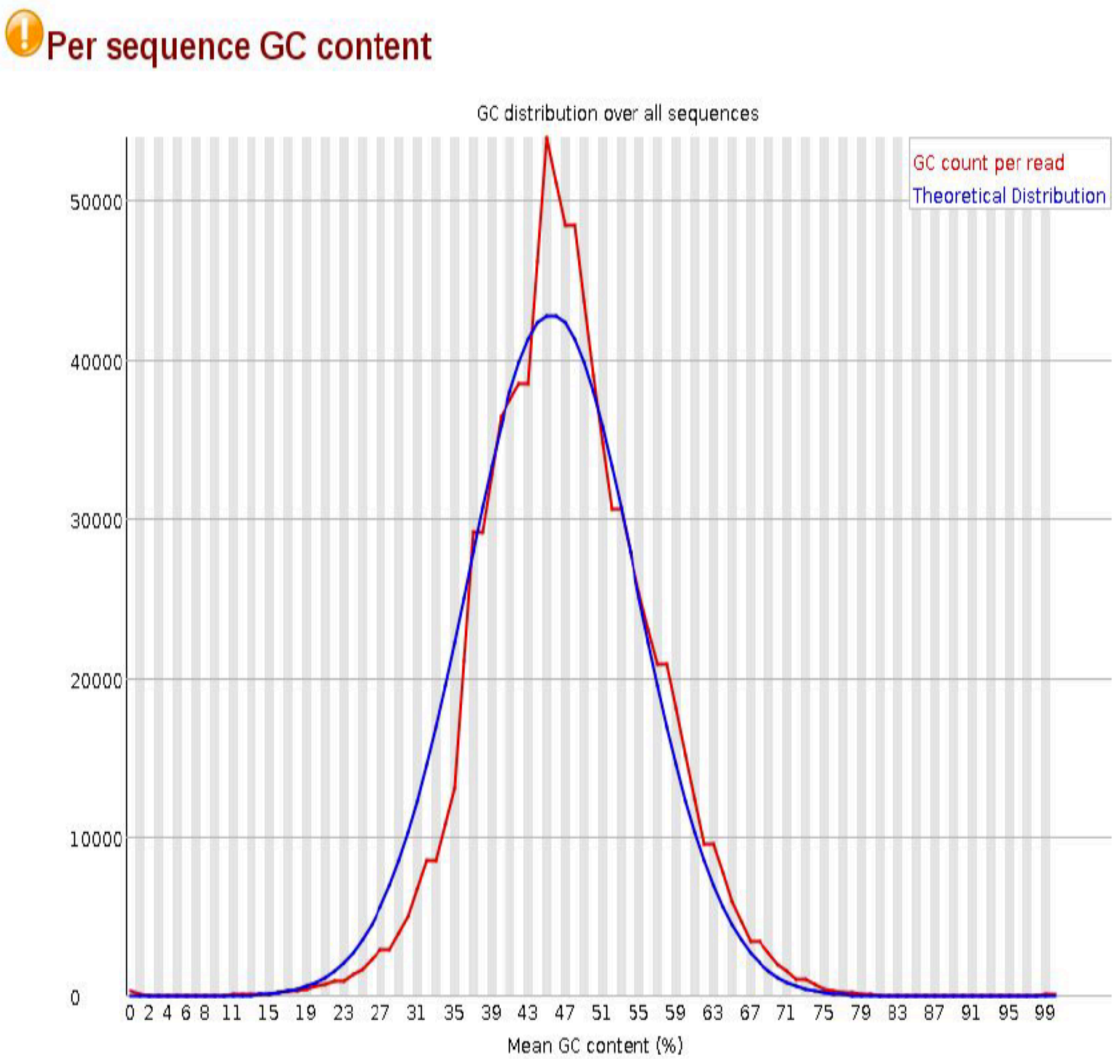


Per base sequence content



% of bases called for each of the four nucleotides at each position across all reads in the file.



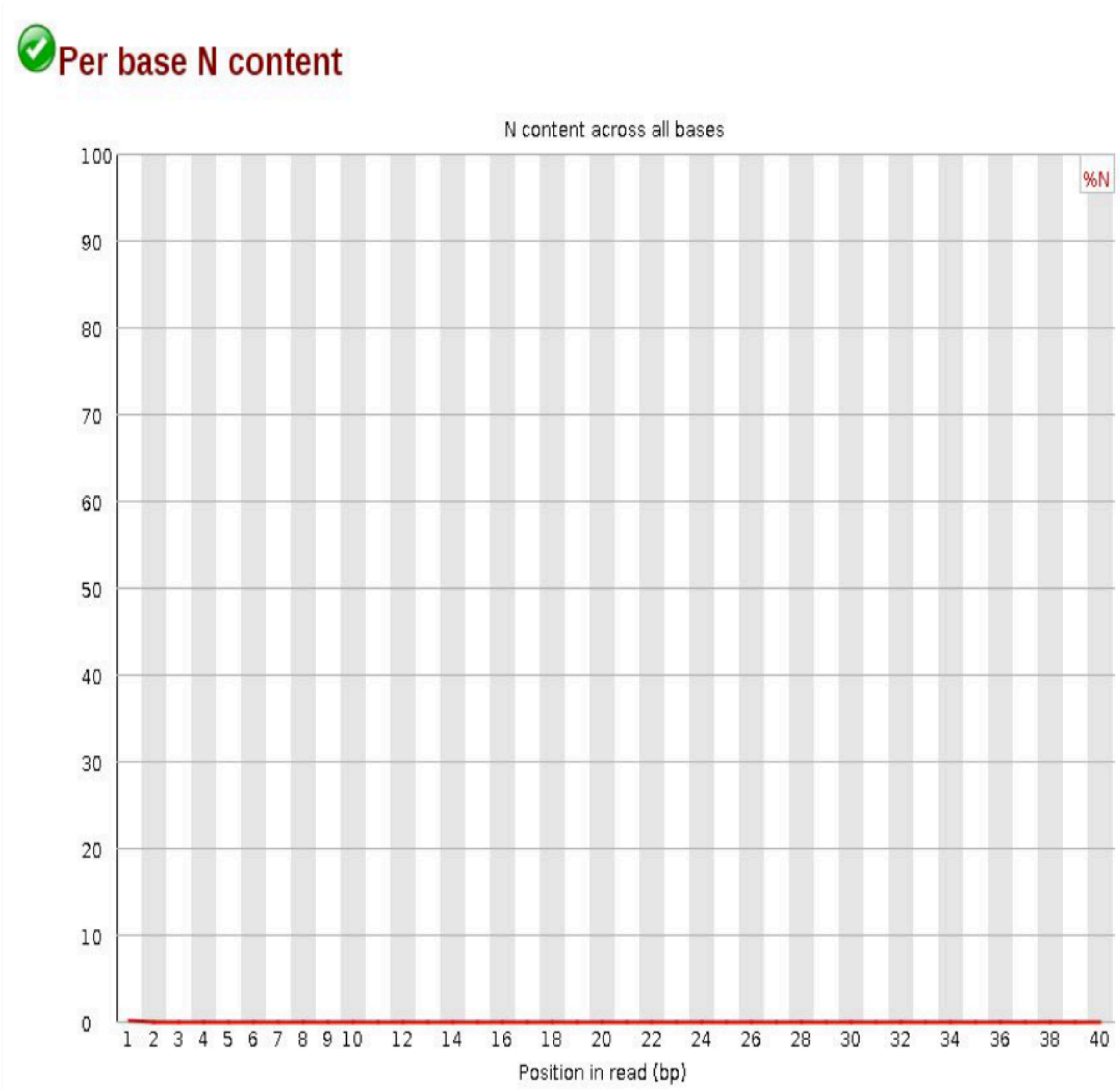


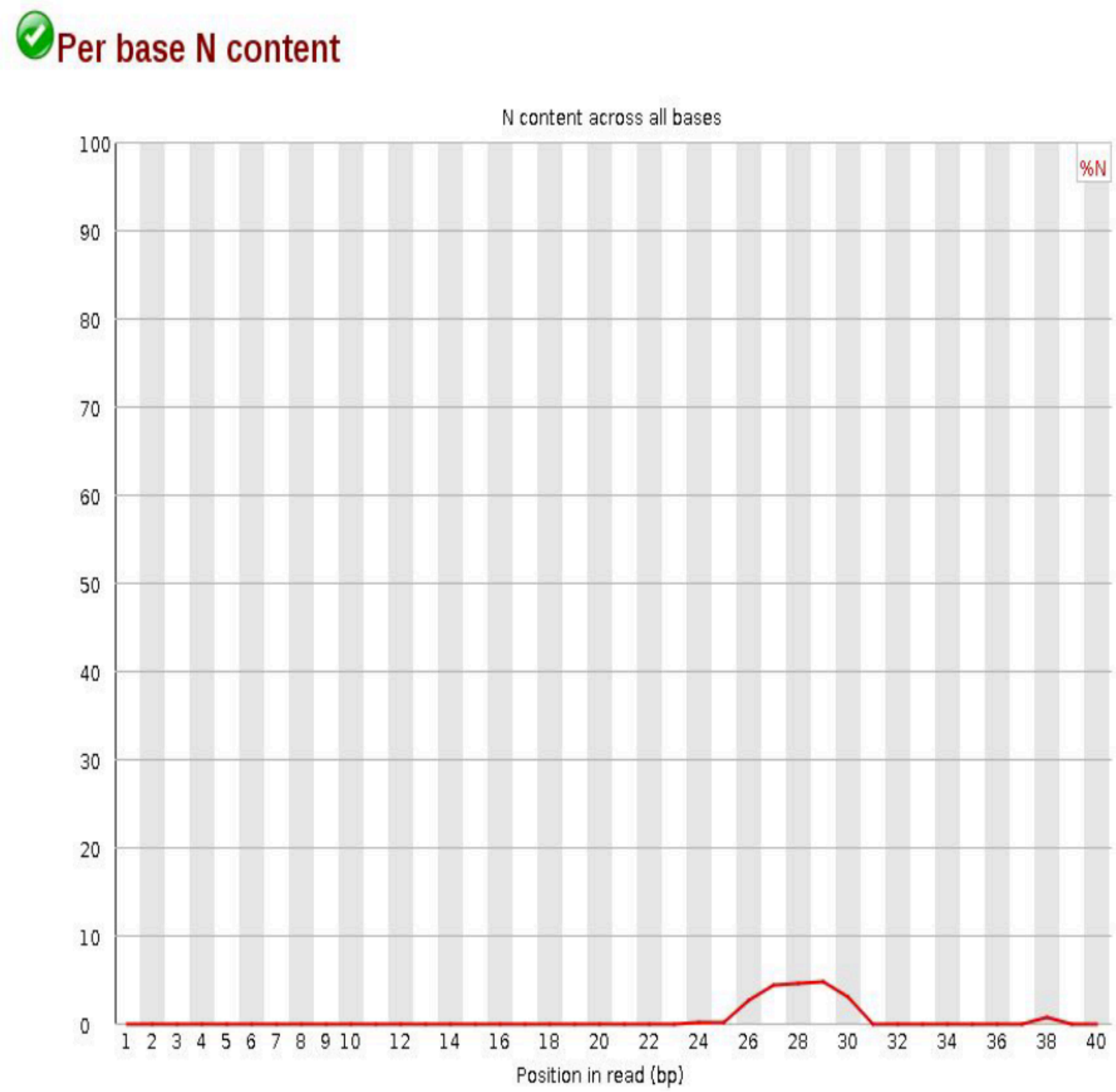
Theoretical distribution

Data distribution

Plot of the number of reads vs. GC% per read.

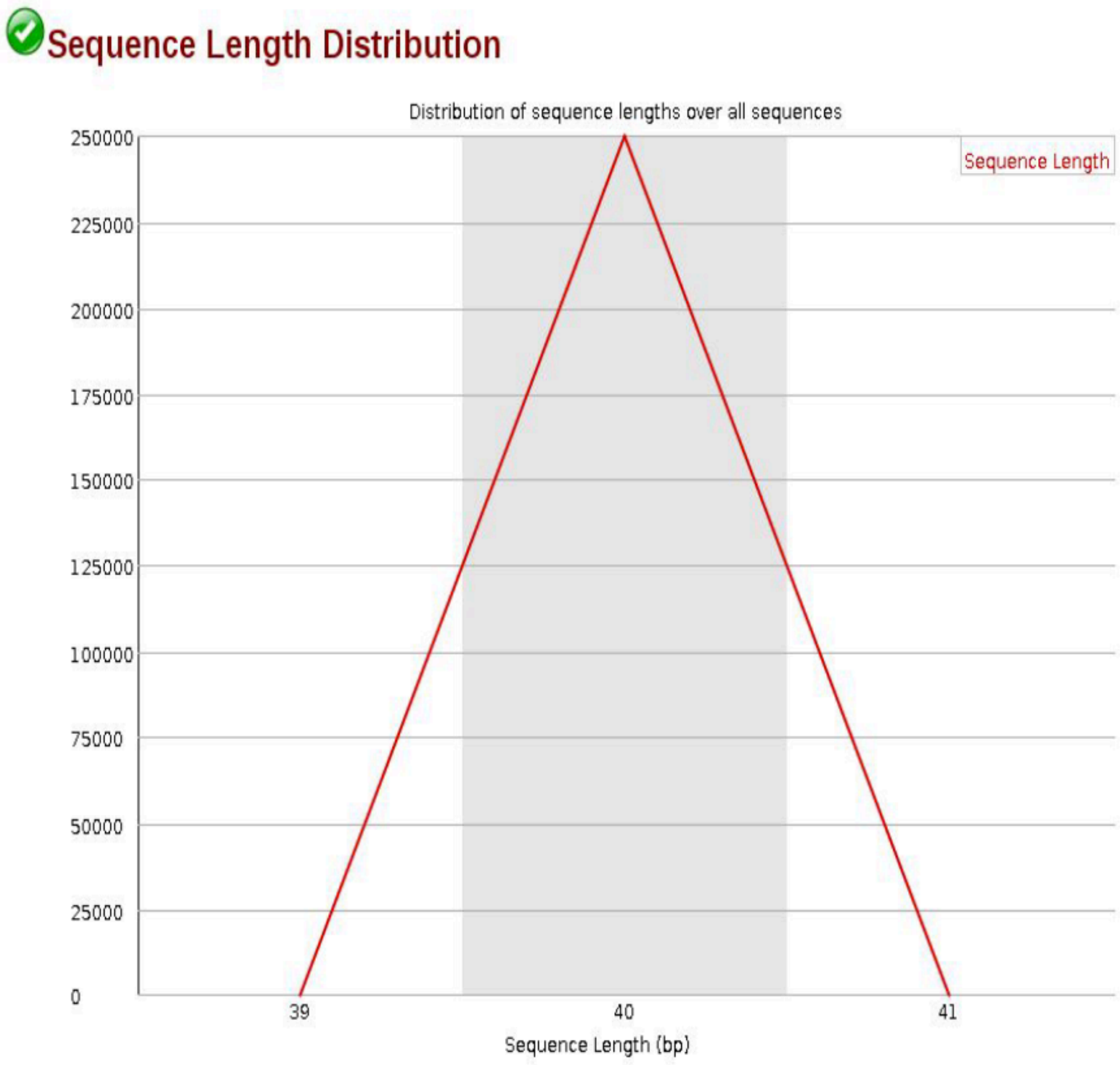
Per base N content

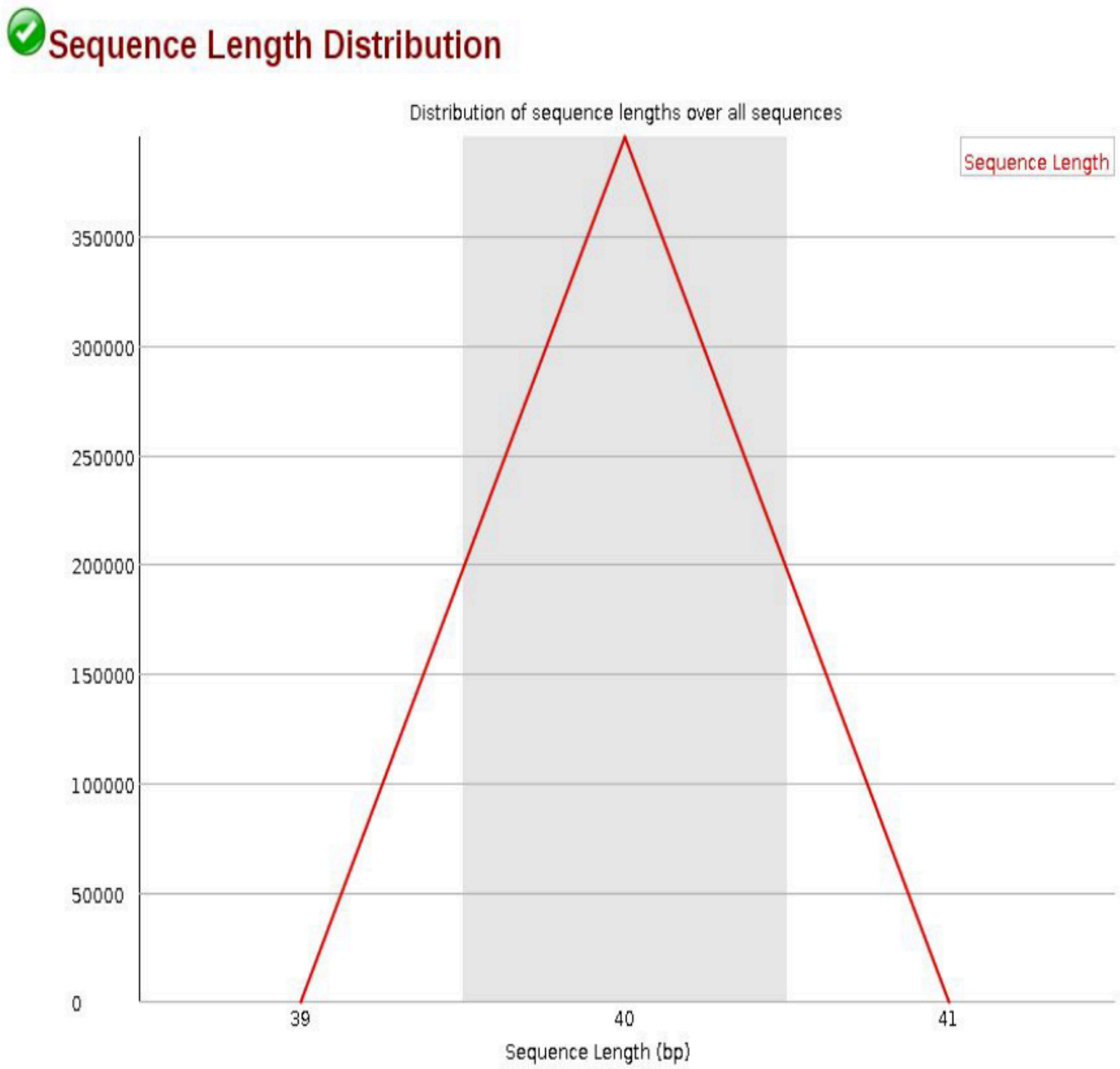




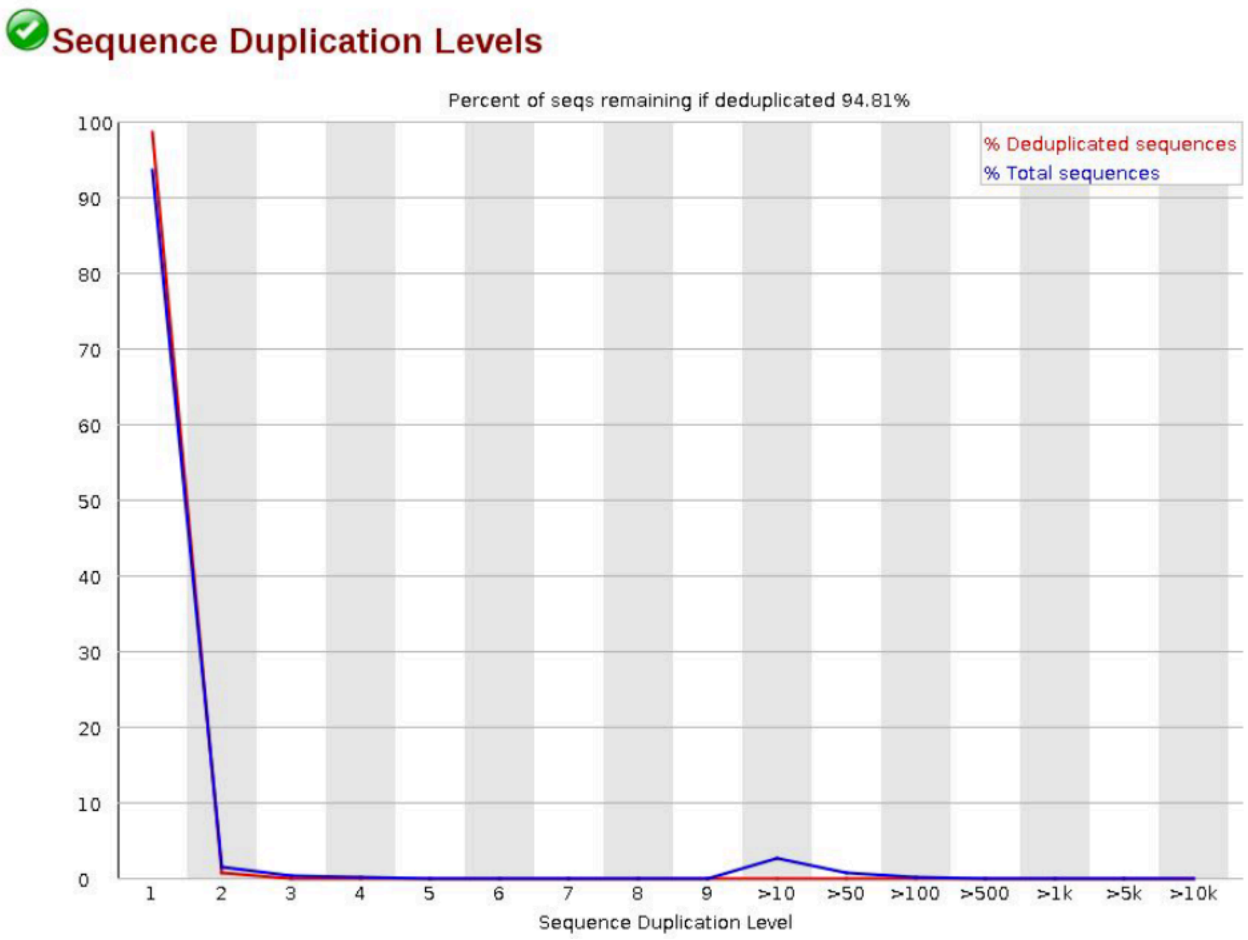
Percent of bases at each position or bin with no base call, i.e. ‘N’.

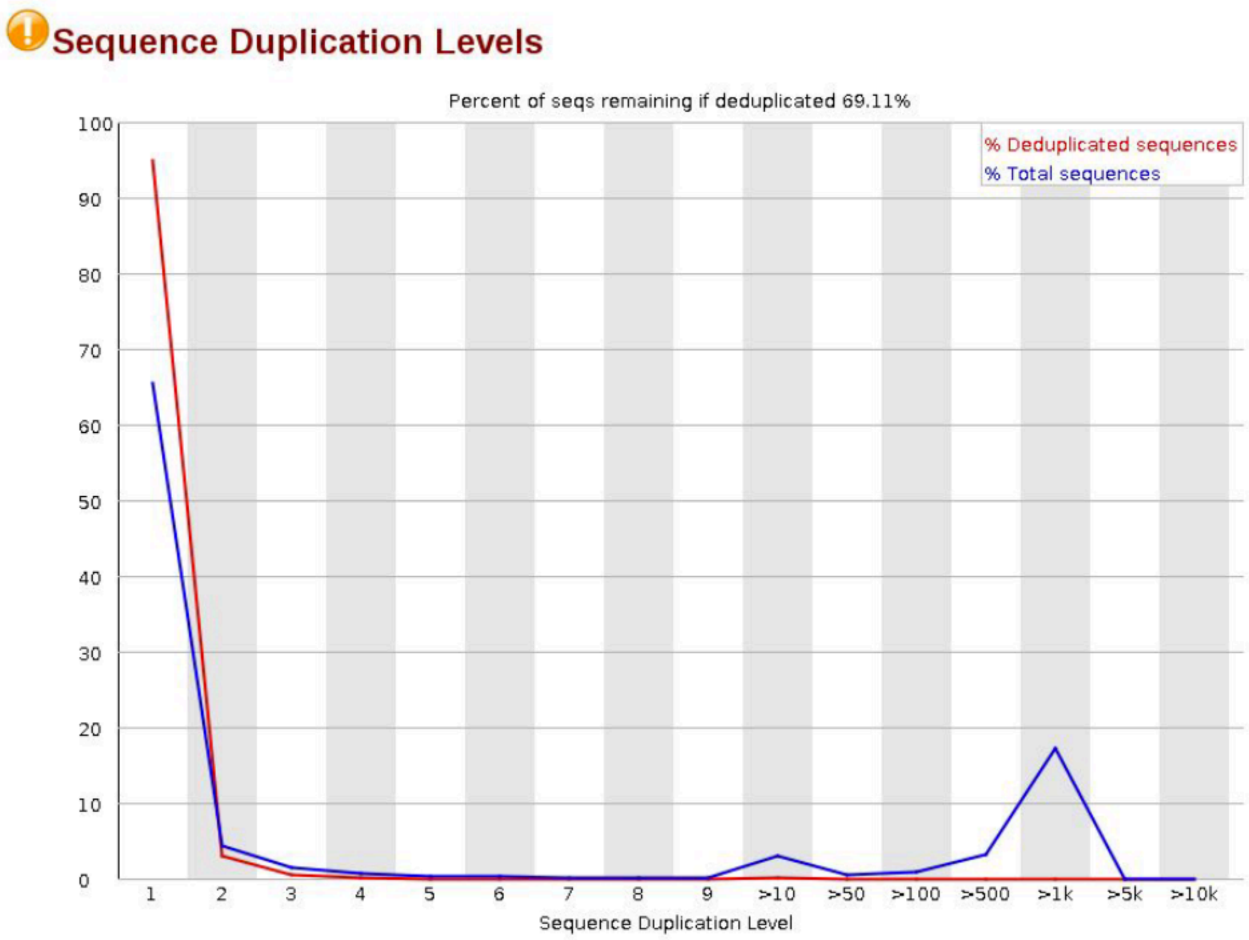
Sequence length distribution





Sequence duplication level





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

Overrepresented sequences

Overrepresented sequences

No overrepresented sequences

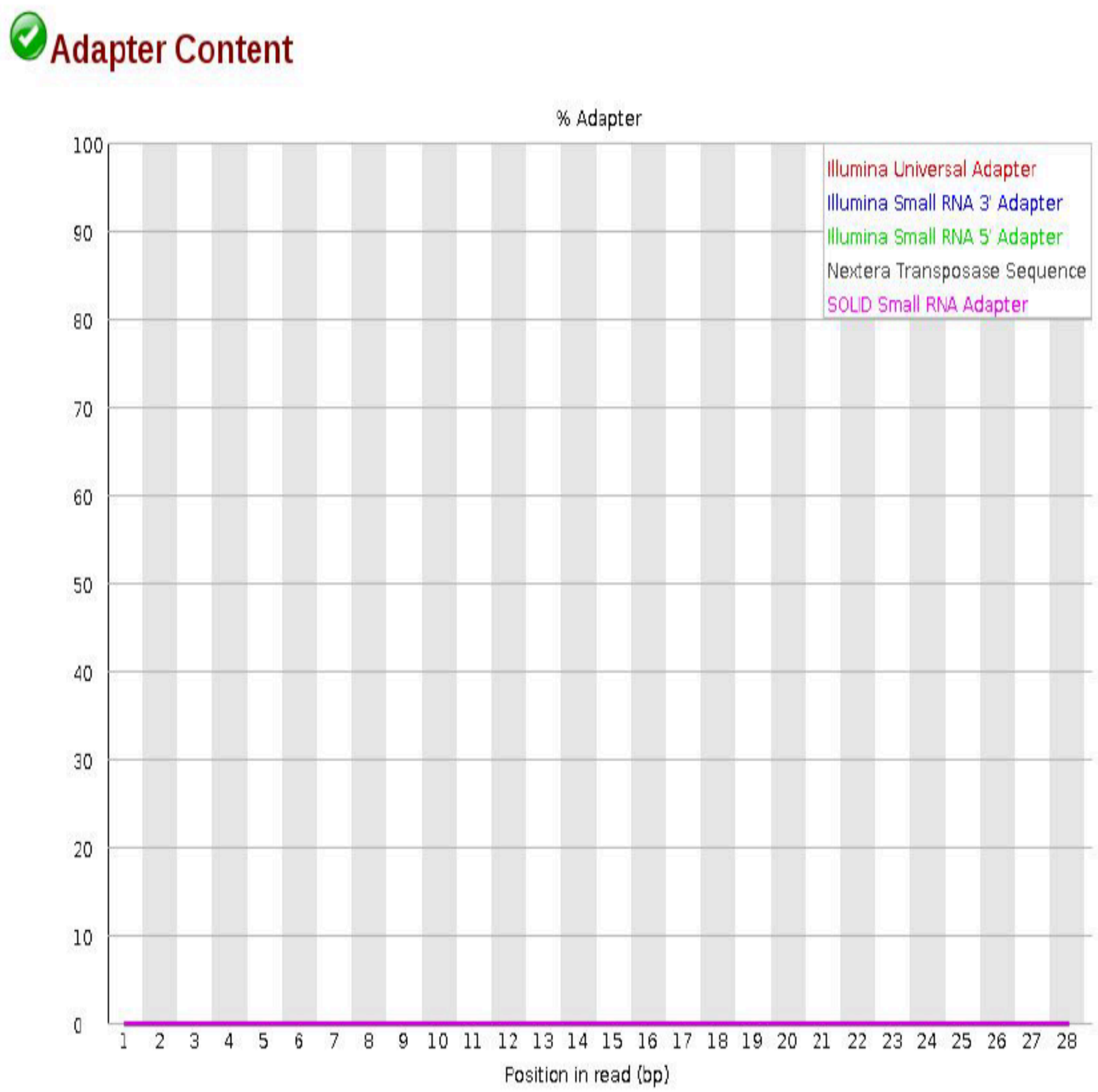
Overrepresented sequences

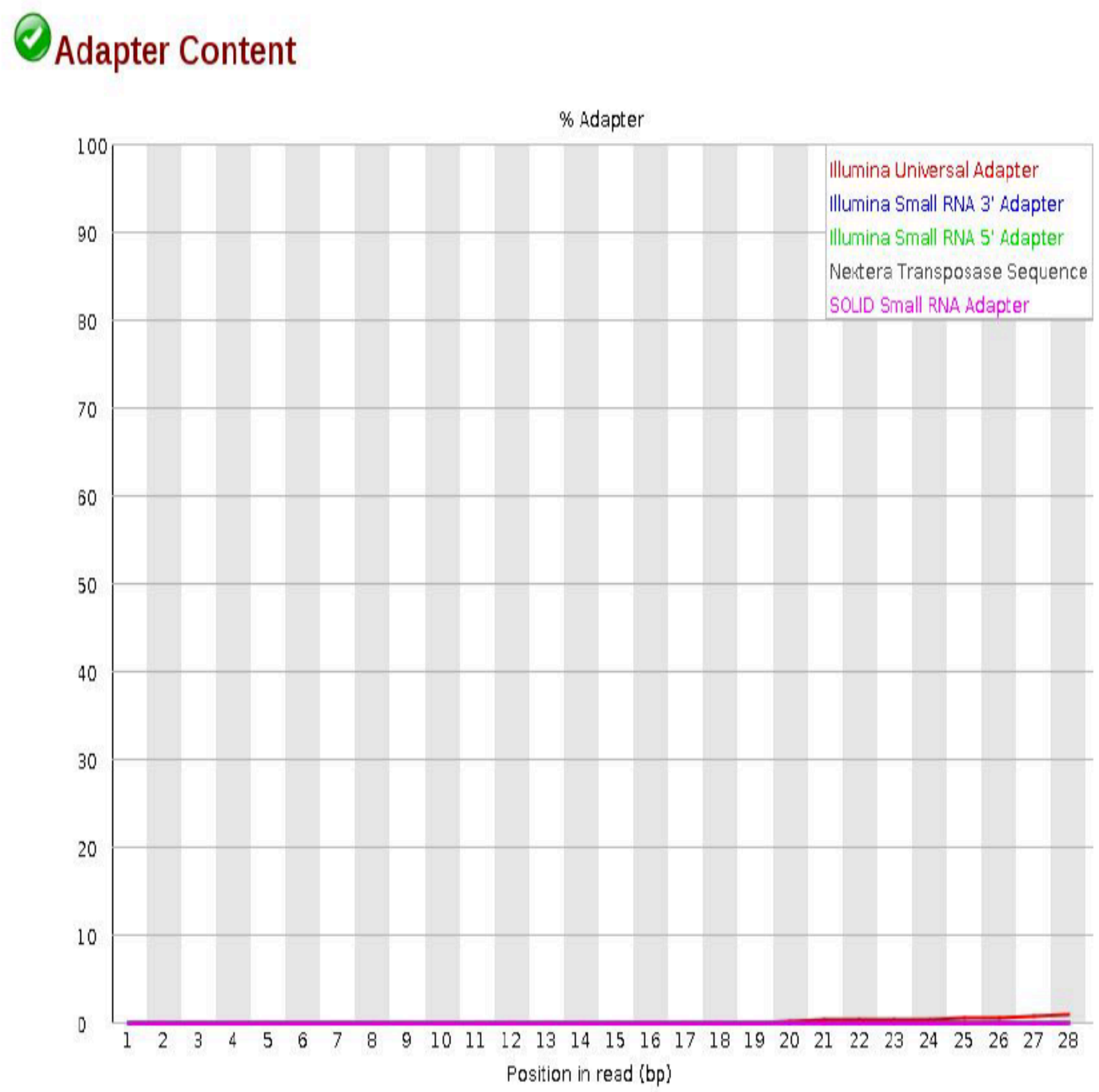
Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAGTTATATGGCTGGGGGTTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGC	532	0.1345854162028698	No Hit
CTGCGTCATGGAAGCGATAAAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCGC	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAGTTATATGGCTGGGGGTTTTTTTG	411	0.10397482341988626	No Hit

- List of sequences which appear more than expected in the file.
- Only the first 50bp are considered.
- A sequence is considered overrepresented if it accounts for ≥ 0.1% of the total reads.

https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

Adapter content

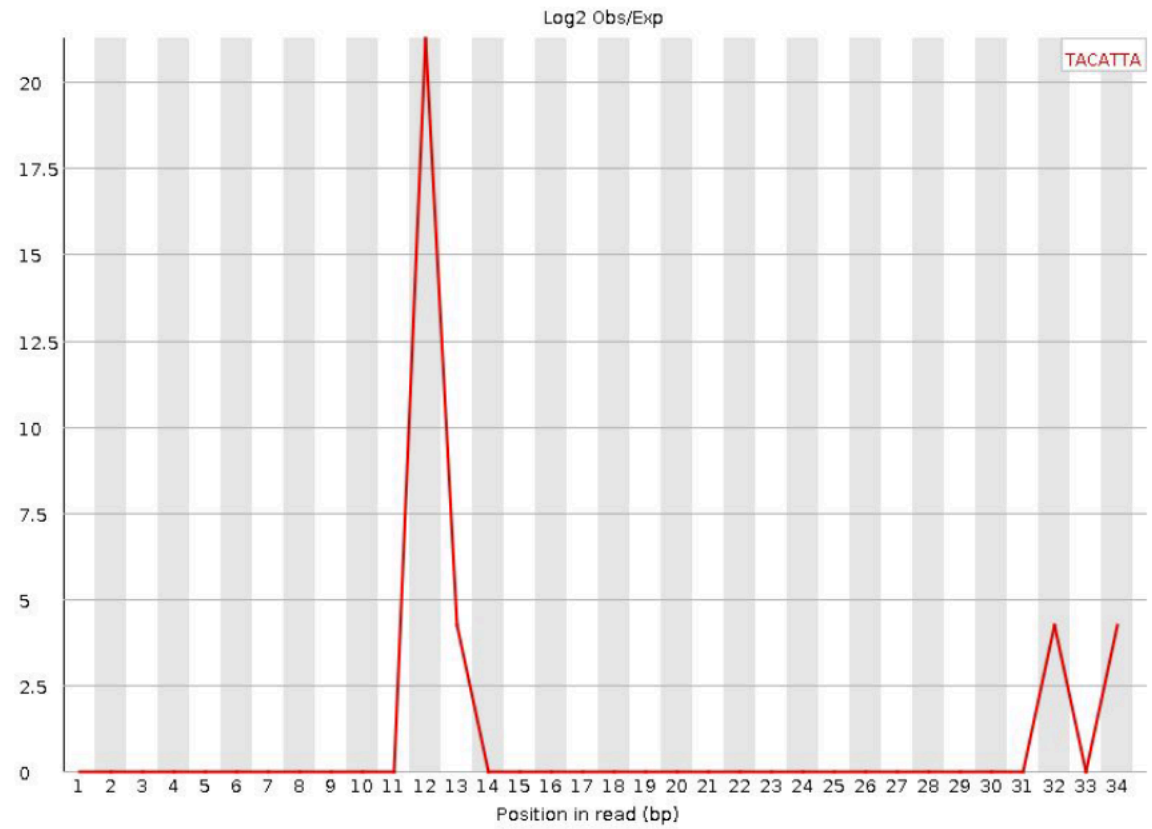




Cumulative plot of the fraction of reads where the sequence library adapter sequence is identified at the indicated base position.

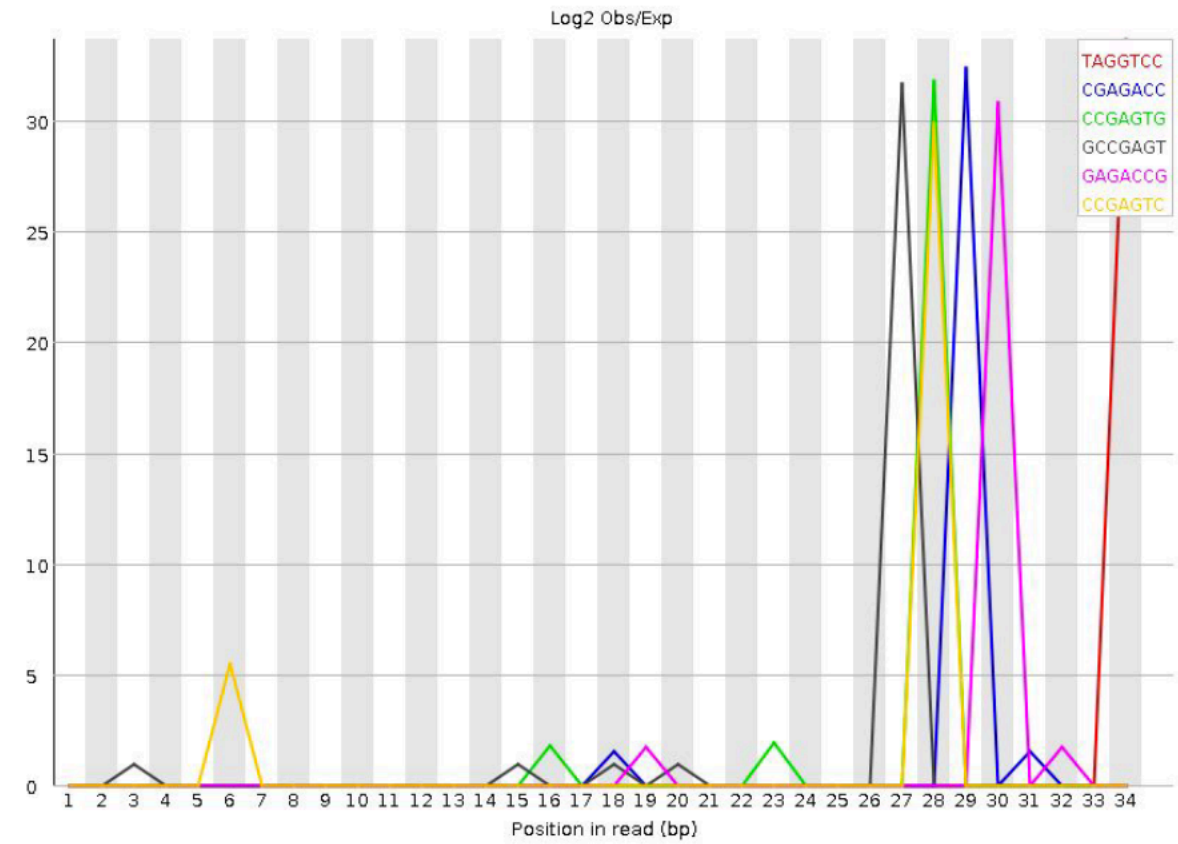
Kmer content

Kmer Content



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

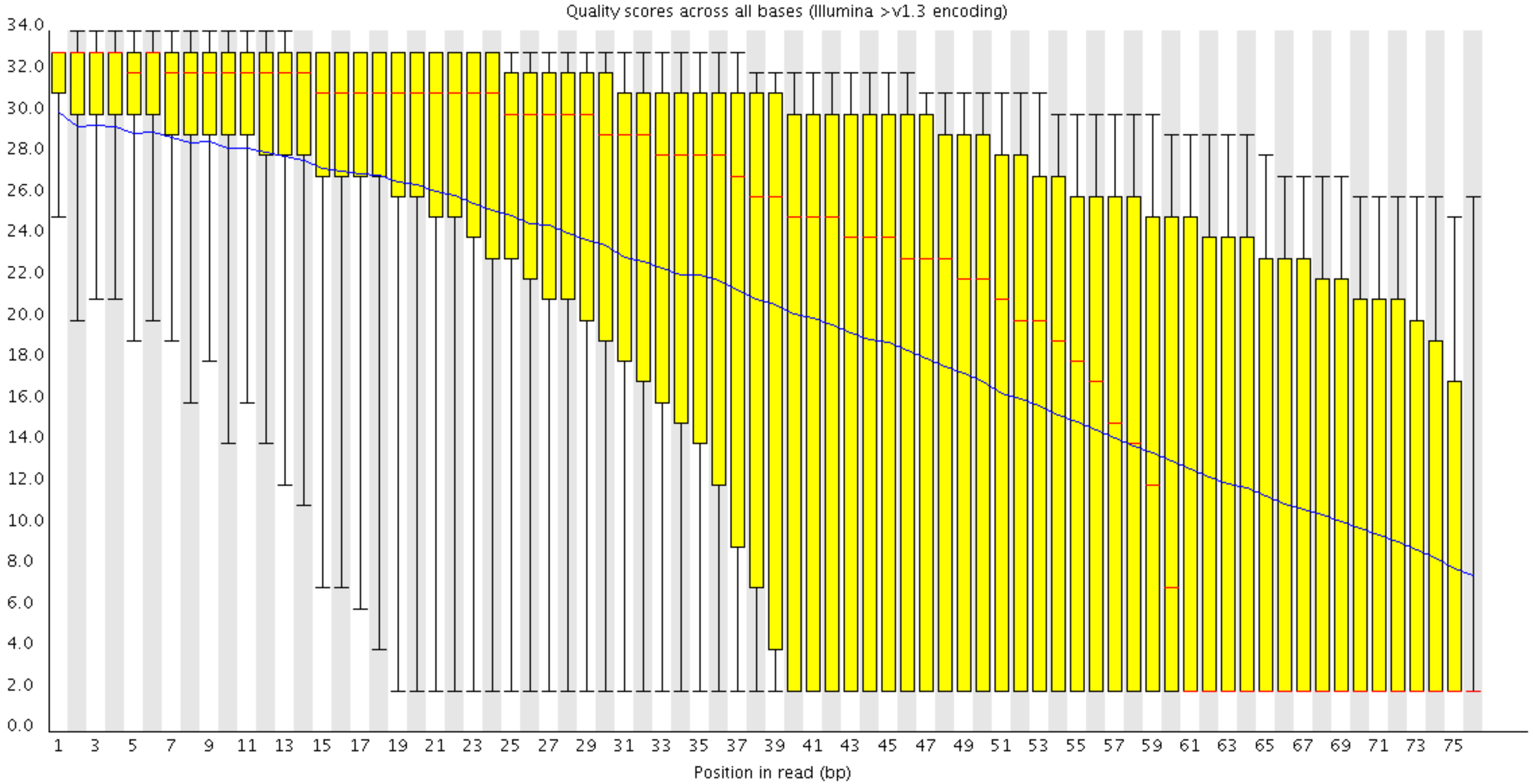
Kmer Content



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

Measures the count of each short nucleotide of length k (default = 7) starting at each positon along the read.

Common problems with quality



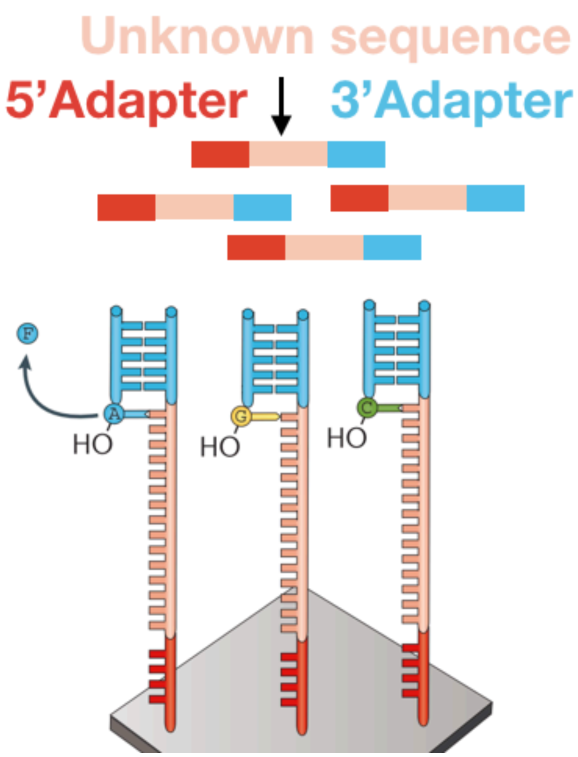
Drop in sequence quality towards 3'end of a read

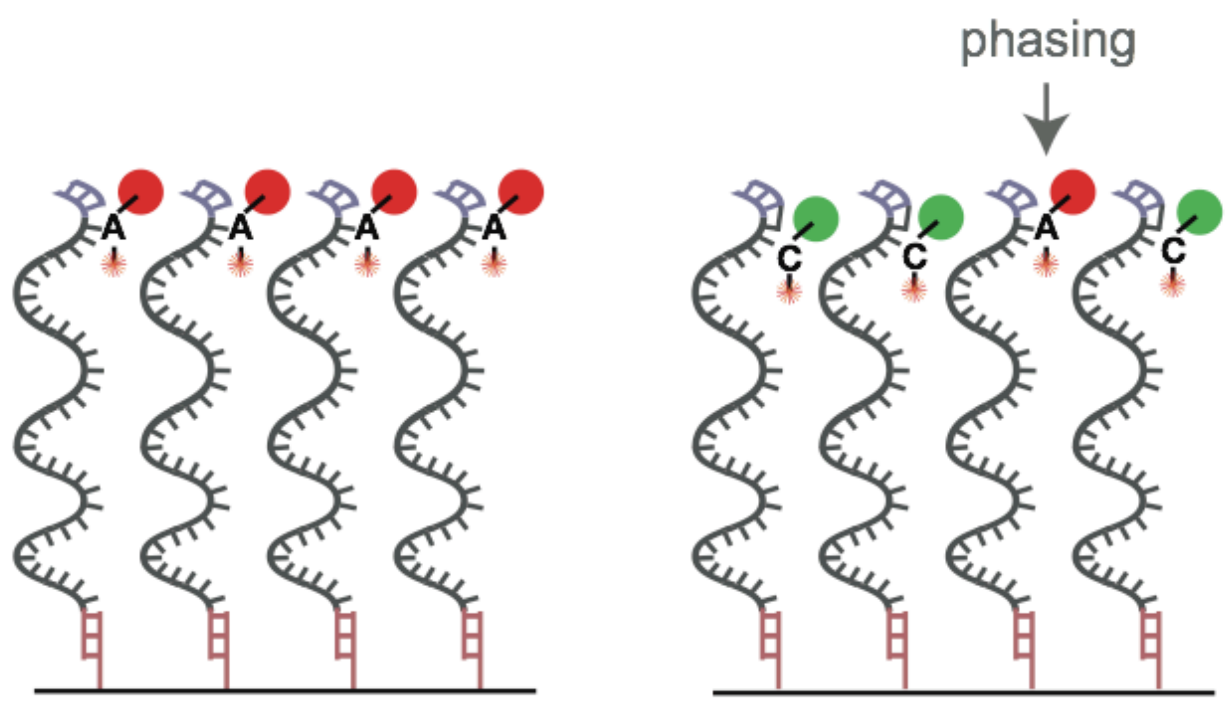
Common problems with quality

Phasing

the blocker of a nucleotide is not correctly removed after signal detection. In the next cycle no new nucleotide can bind on this DNA fragment and the old nucleotide is detected one more time.

From now on this DNA fragment will be 1 cycle behind the rest (out of phase), polluting the light signal that the sequencer's camera has to read.

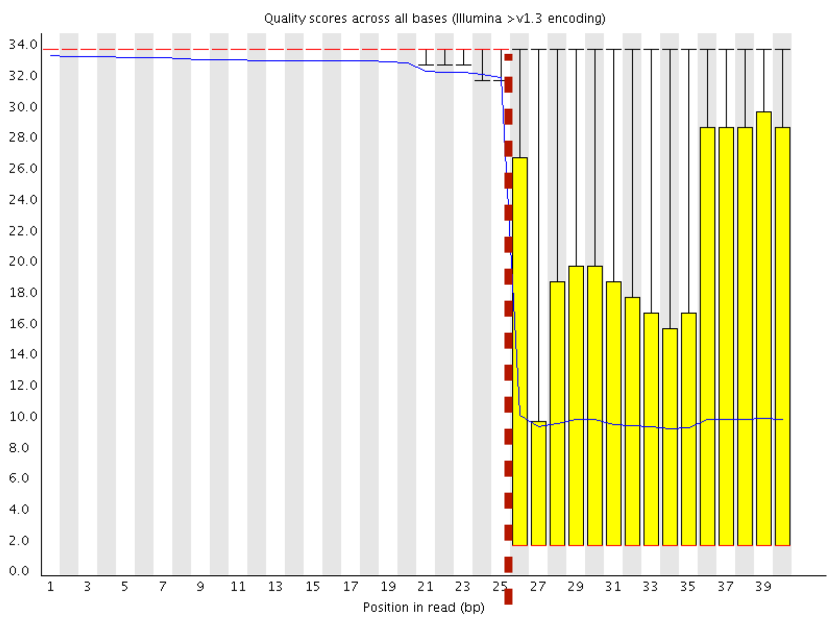




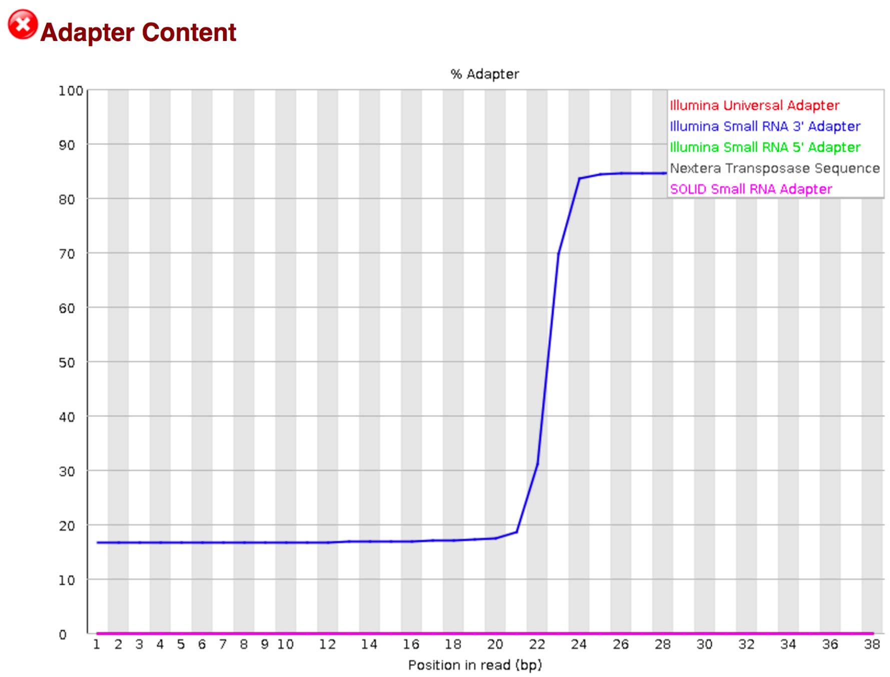
https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

Artefact removal: when the quality needs to be increased

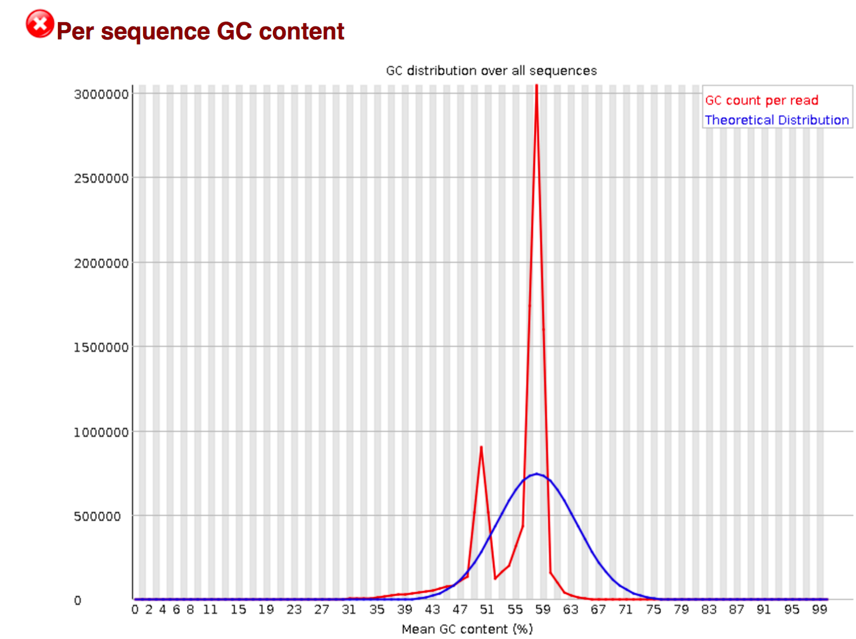
If we want to accurately align as many reads as possible, we may remove unwanted/noisy information from our data, eg:



Poor quality bases at read ends



Leftover adapter sequences



Known contaminants (strings of As/Ts, other sequences)

Today we will use **Cutadapt** to perform quality trimming of our sample dataset.

Sequencing data repositories

dbGaP

SRA

GEO Gene Expression Omnibus

ENA European Nucleotide Archive

ArrayExpress

EUROPEAN GENOME-PHENOME ARCHIVE

Example data sets

Study type	Recommended submissions route(s)	Data repository/ies	Recommended retrieval route(s)
Array-based mouse genotyping	MAGE-Tab	ArrayExpress	ArrayExpress
Small-scale sequence-based mouse genotyping	MAGE-Tab SRA-Webin	SRA	ArrayExpress SRA
Human (restricted access) genotyping	EGA	EGA	EGA

More about recommended data repositories: https://www.nature.com/sdata/policies/repositories

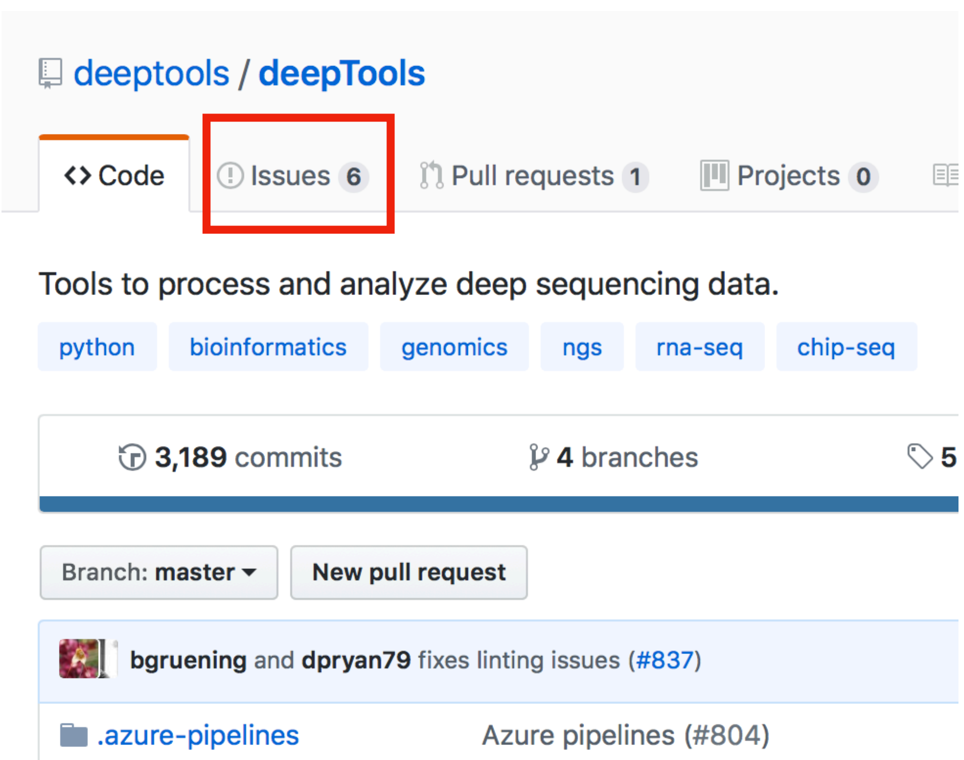
Data downloading: https://www.ebi.ac.uk/ena/browse/read-download

https://sites.psu.edu/yuka/2016/04/07/how-to-use-sra-toolkit/

Still lost?

Google!

Package manual, GitHub



Bioinformatics forums and discussion groups:

https://www.biostars.org

https://support.bioconductor.org

http://seqanswers.com

Let's practice!