Initial quality control and adapter trimming

Initial quality control with FastQC
Adapter trimming with Trim Galore! and Trimmomatic

Software to install

Install fastQC

```
$ conda install -c bioconda fastqc
```

Install Trim galore!

```
$ conda install -c bioconda trim-galore
```

Install Trimmomatic

```
$ conda install -c bioconda trimmomatic
```

Quality control

After receiving the reads, we need to conduct initial assessment of their quality

Typical questions to ask at this stage:

1. What are the sizes of the sequencing libraries and are they enough to achieve the experimental objective?

2. What are the base qualities?

3. Is there anything wrong with key attributes of the library: sequence composition, GC content, length distribution, duplicated sequences, over-represented sequences?

4. Are the sequences contaminated with adapters?

5. Can we do anything to improve the quality of the sequencing libraries?

NOTE: Do not try to “save” failed libraries, it’s a waste of time.

Quality control

- Quality control helps us to diagnose problems and improve the data at the early stages of the workflow

- Any steps we undertake to improve the data will alter it, so we must be careful with this process

- Diagnostic software: FastQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- Take a look at FastQC documentation:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

- FastQC examples

Good data:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Bad data:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Quality control

- FastQC examples

Adapter dimer contamination:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/RNA-Seq_fastqc.html

Small RNA sequencing adapter read-through
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/small_rna_fastqc.html

PacBio
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/pacbio_srr075104_fastqc.html#M1

Useful website to examine failed libraries and sketchy situations that arise during sequencing
https://sequencing.qcfail.com/

QC practice

Small RNA mouse example

Create a folder **QC/** in **sandbox/** directory

QC folder will be our working directory for this practice

Download sample_mm_srna.fastq from github repository

```
$ wget https://raw.githubusercontent.com/slavailn/bioinf_training/main/sample_mm_srna.fastq
```

View ***fastqc*** help files; what options are available

```
$ fastqc -h # examine the help file
$ fastqc sample_mm_srna.fastq # Generated .html report and .zip file with the data used for report generation
```

We can unzip the data and take a look at the files it has

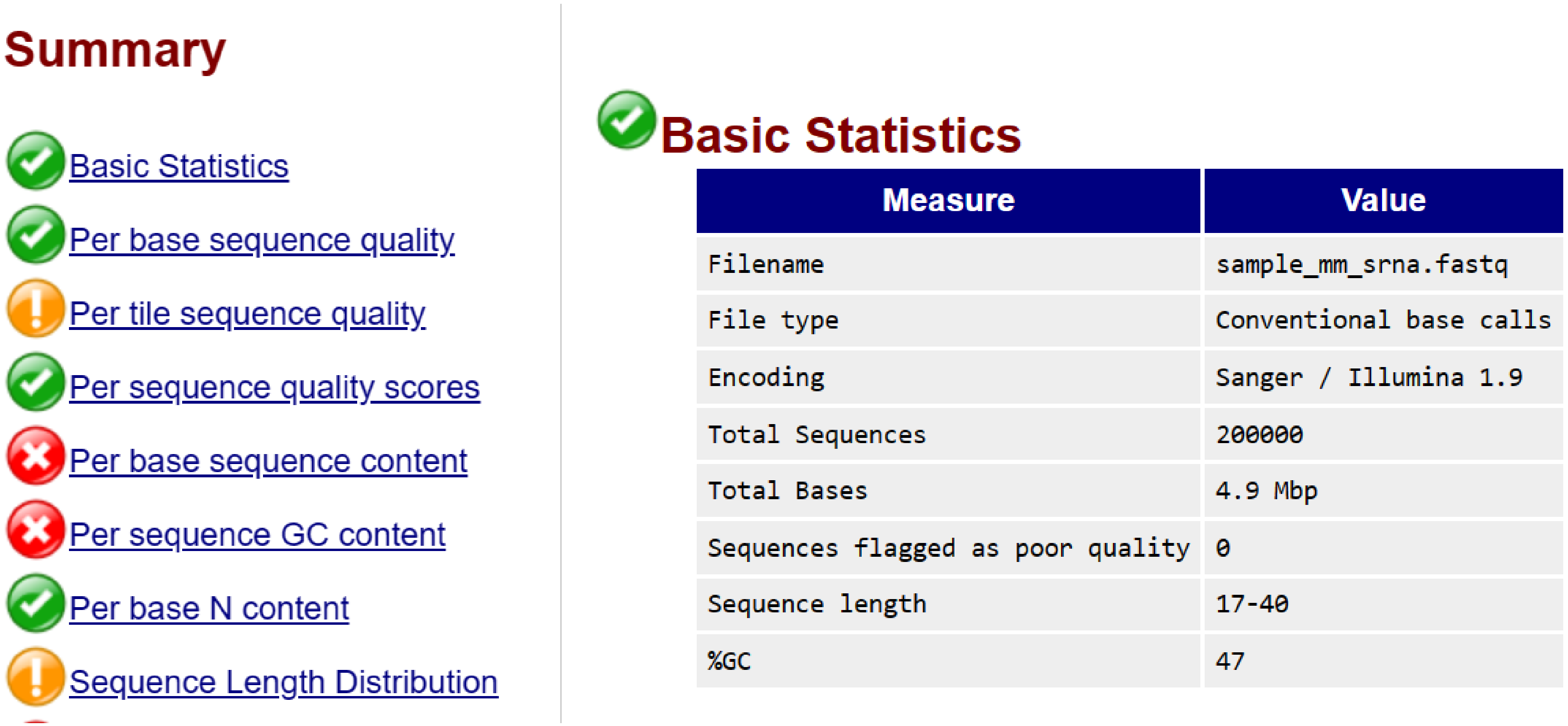
```
$ unzip sample_mm_srna_fastqc.zip
$ ls -l sample_mm_srna_fastqc/
$ less sample_mm_srna_fastqc/fastqc_data.txt # this file will contain report data as text
```

QC practice

Small RNA mouse example

View the ***html*** report

```
$ firefox sample_mm_srna_fastqc.html
```



Summary

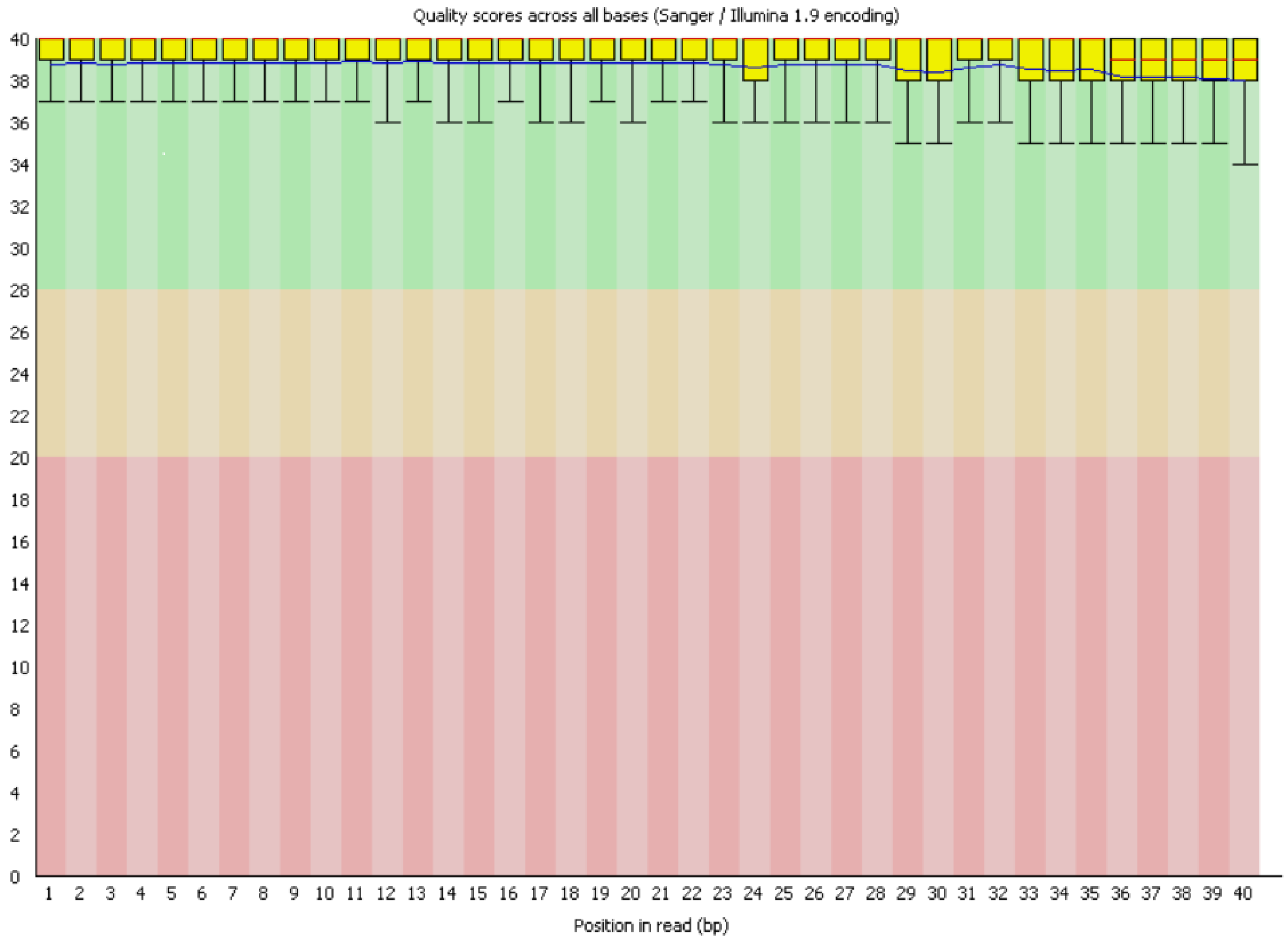
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution

✓ Basic Statistics

Measure	Value
Filename	sample_mm_srna.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	200000
Total Bases	4.9 Mbp
Sequences flagged as poor quality	0
Sequence length	17-40
%GC	47

QC practice

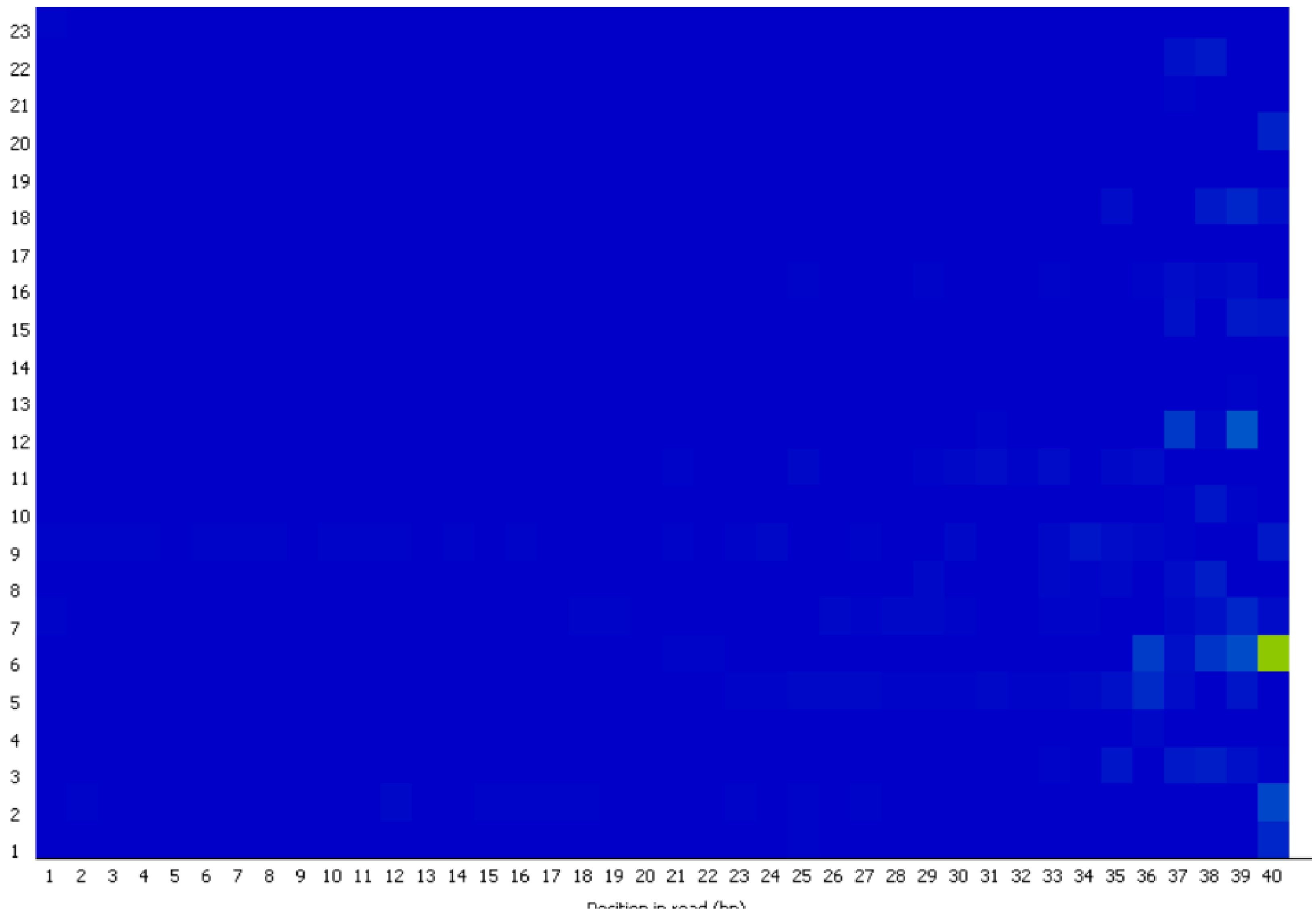
Small RNA mouse example : Per base sequence quality



- X-axis – base position
- Y-axis - qualities on Phred scale
- Quality distribution at each base are shown as box plots
- Green Q >= 30 : Good
- Orange 20 =< Q < 30 : OK
- Red Q < 20: Poor

QC practice

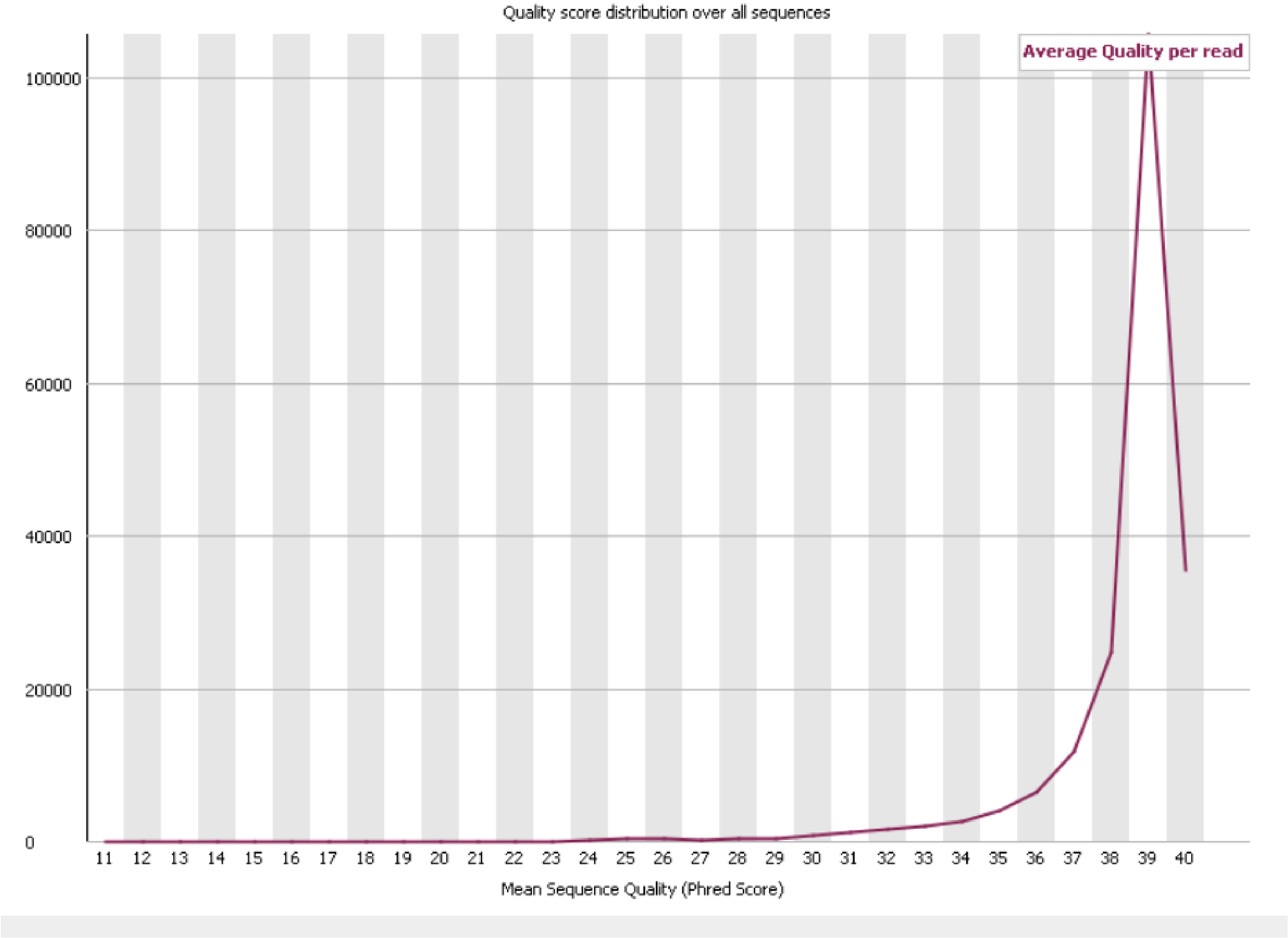
Per tile sequence quality



Average quality score per tile across all bases

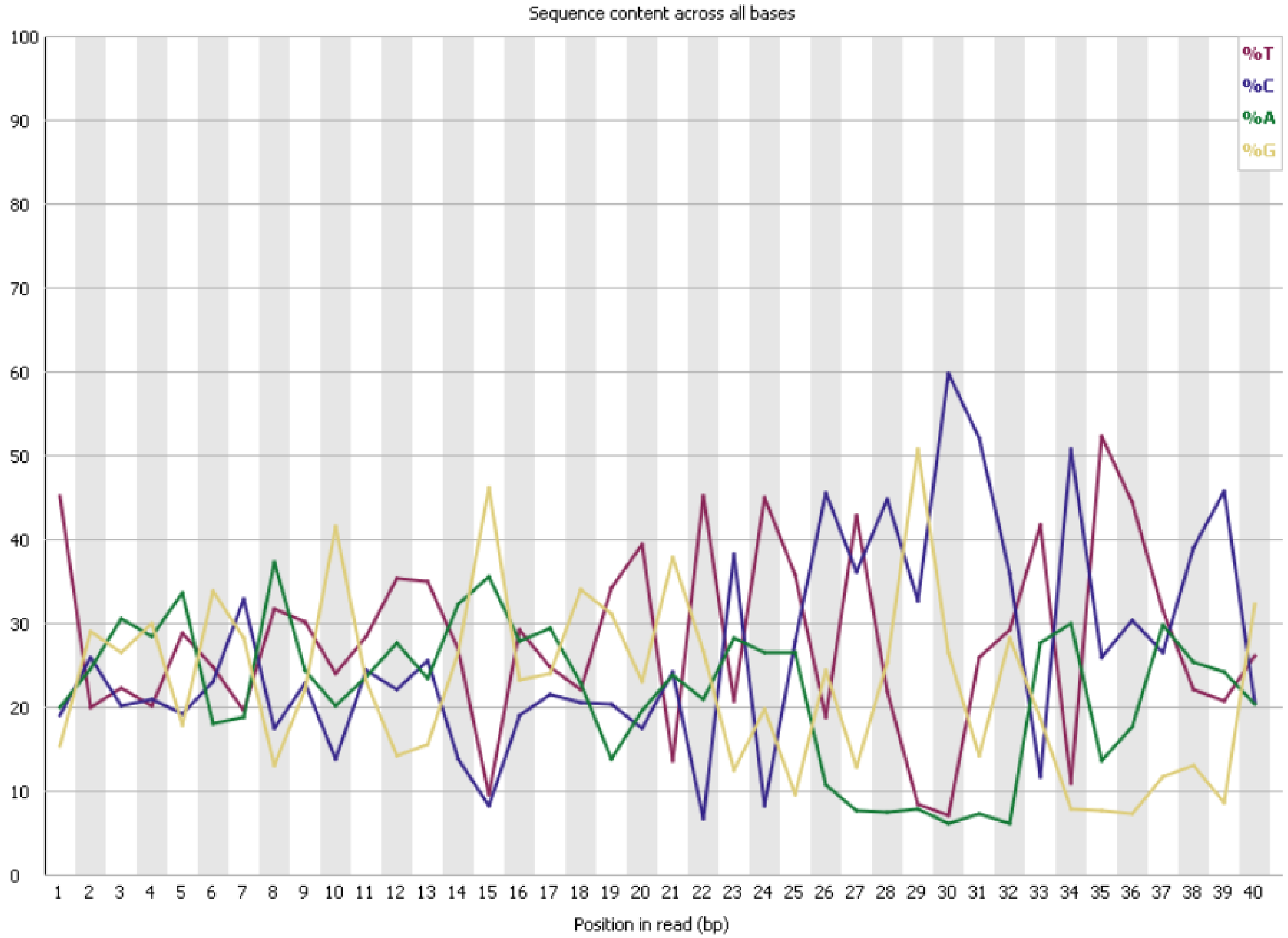
QC practice

Distribution of average quality scores over all sequences



QC practice

Base composition across the sequence

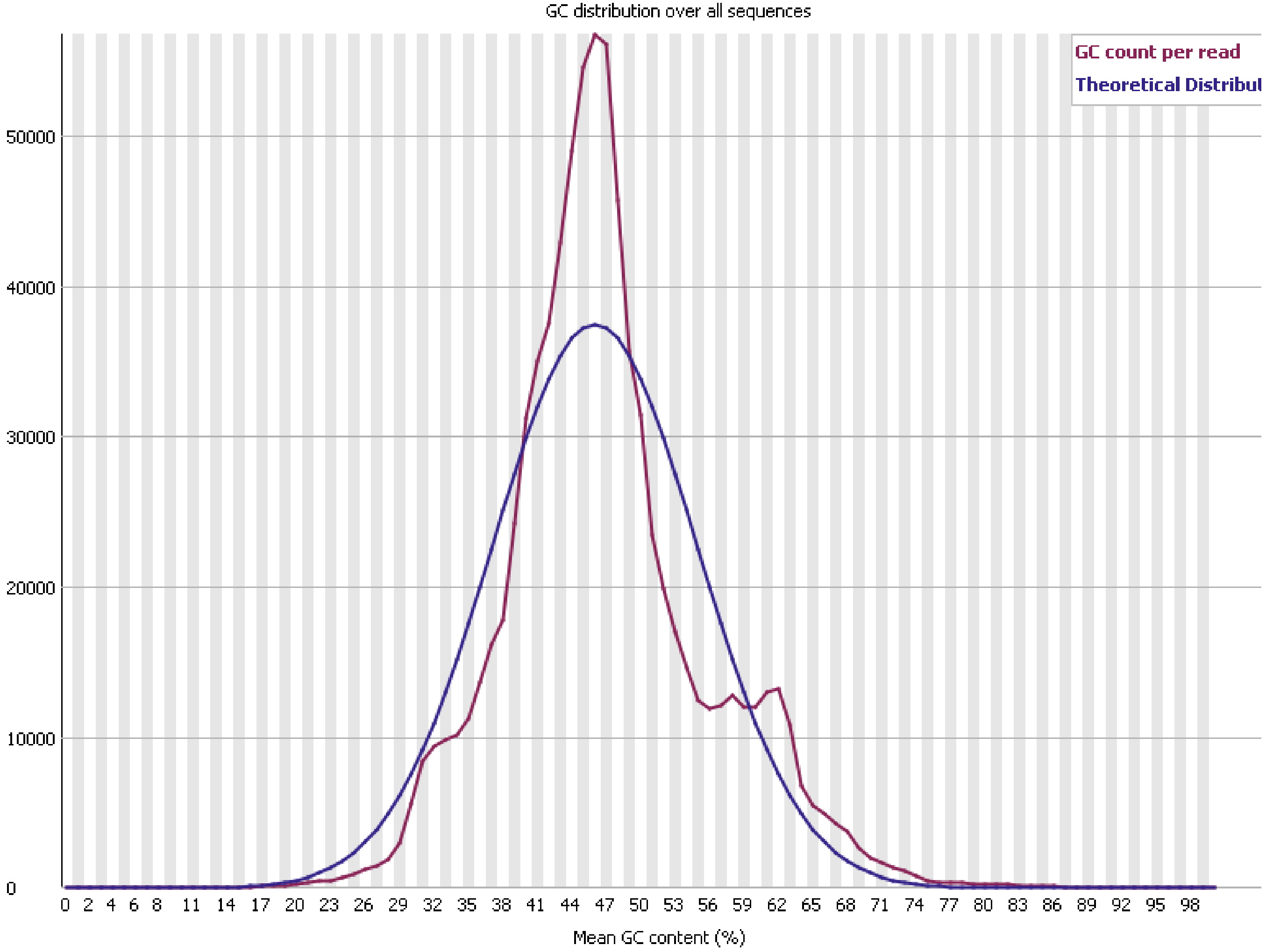


Normally we would expect a more even distribution where each of 4 bases is observed about 0.25 times

However small RNA libraries are naturally heavily imbalanced

QC practice

Distribution of mean GC content across sequences

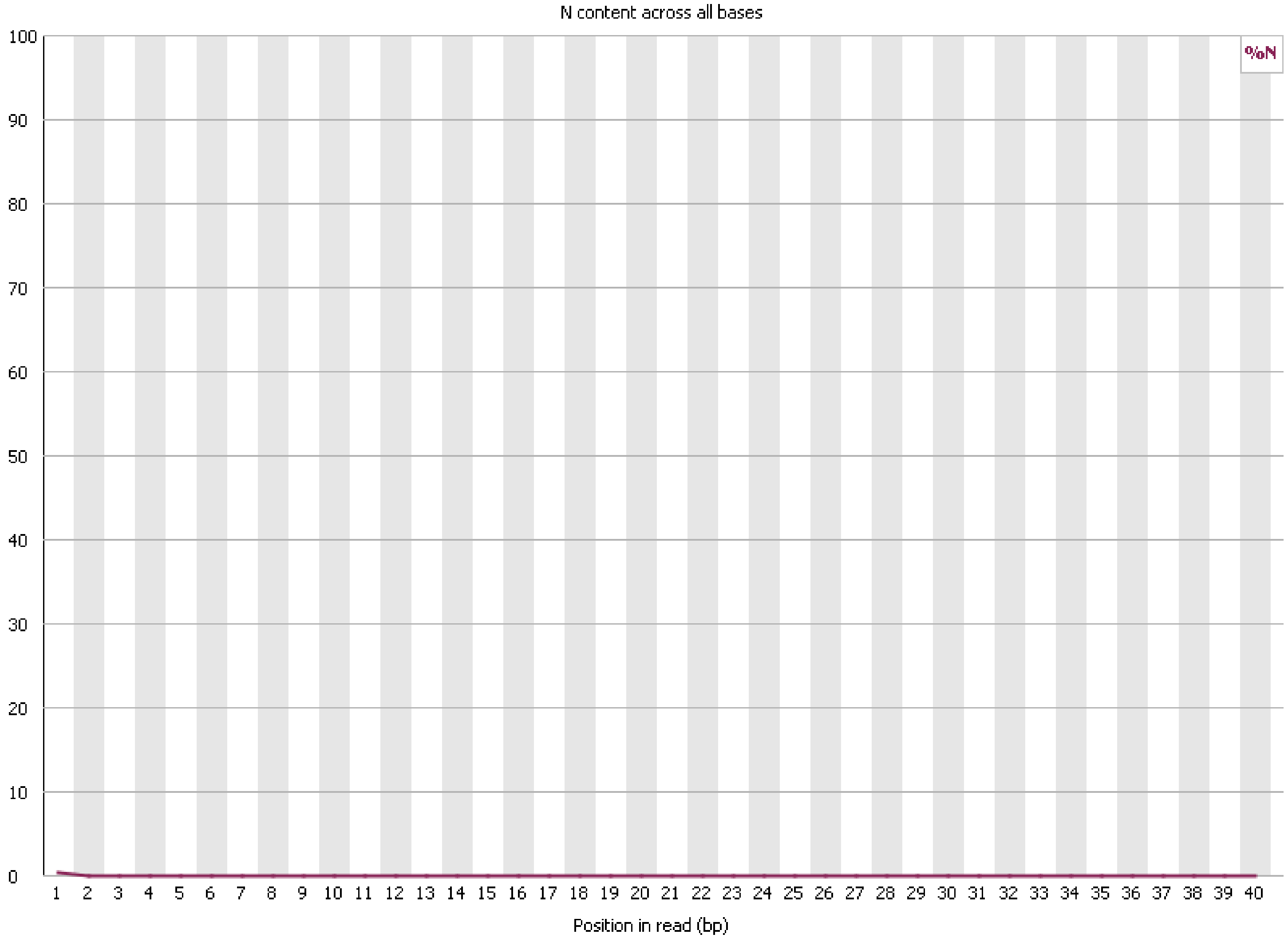


Normally we would expect observed distribution much closer to theoretical

However, small RNA libraries are a clear exception due large part of sequencing space occupied by highly expressed small RNAs and the presence of adapters in raw reads

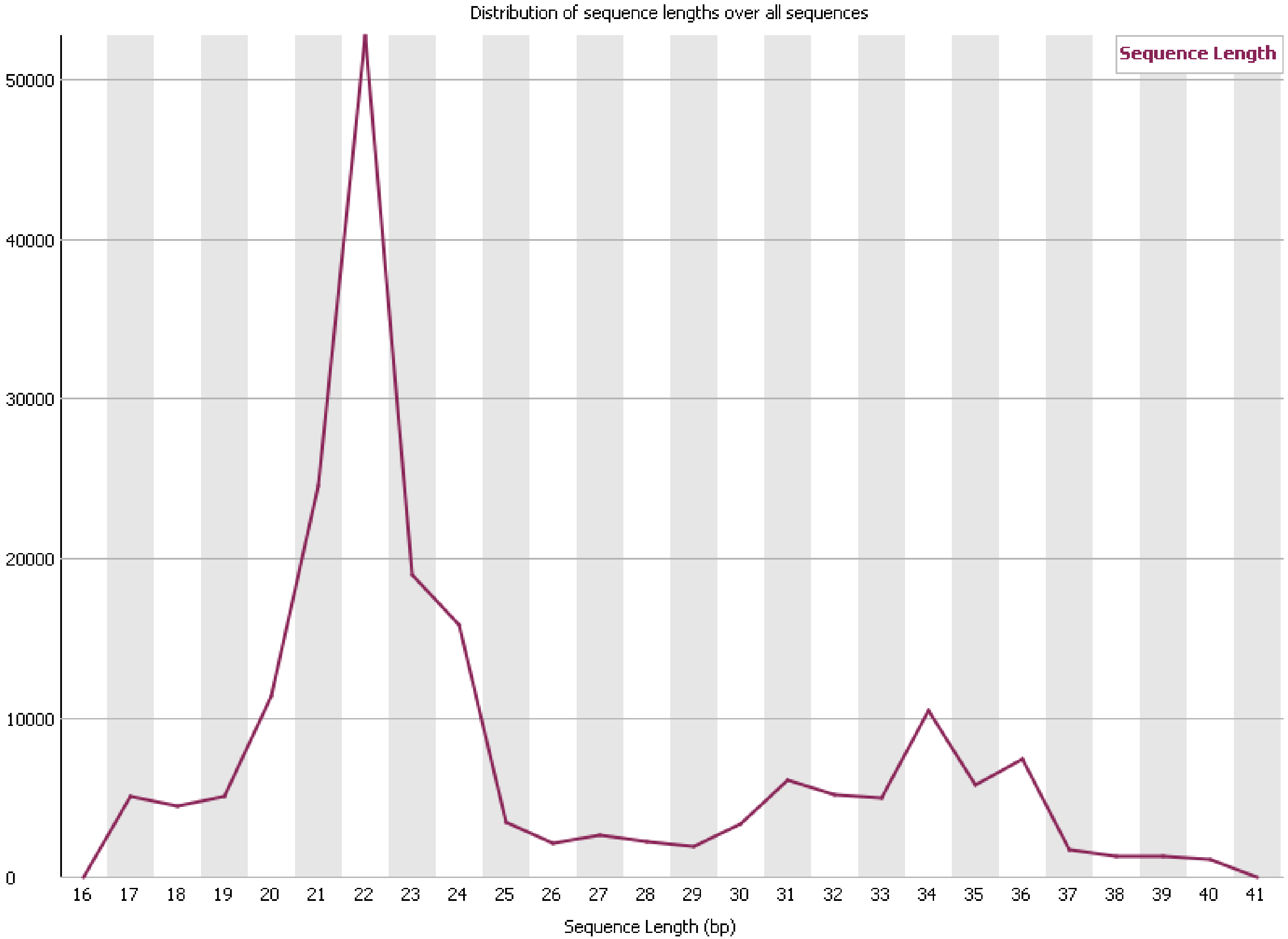
QC practice

Percentage of N bases along the read



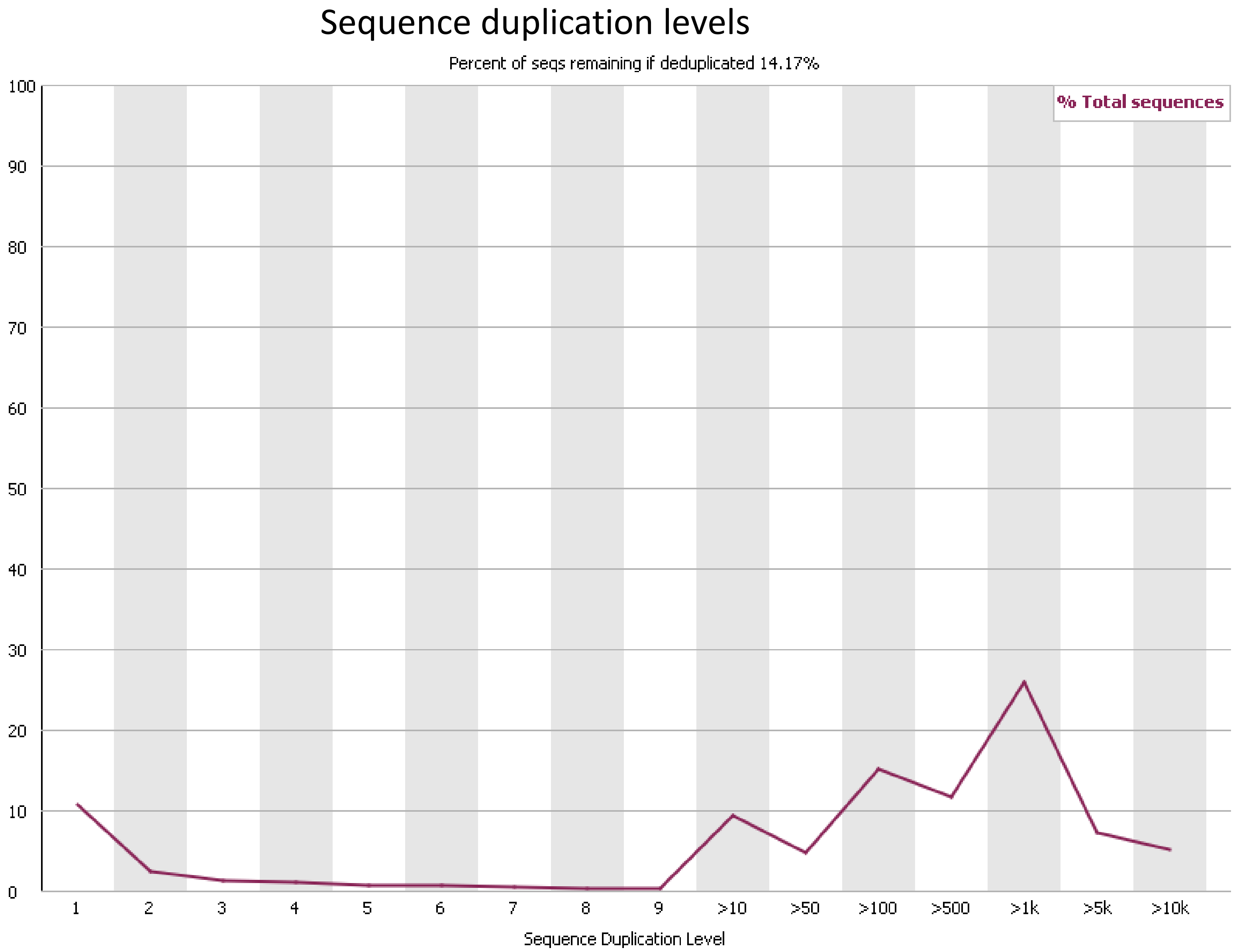
QC practice

Read length distribution



Raw reads have a single point, in this case, the sequences were already trimmed of adapters

QC practice



In the properly diverse libraries most sequences will fall to the far left of the plot

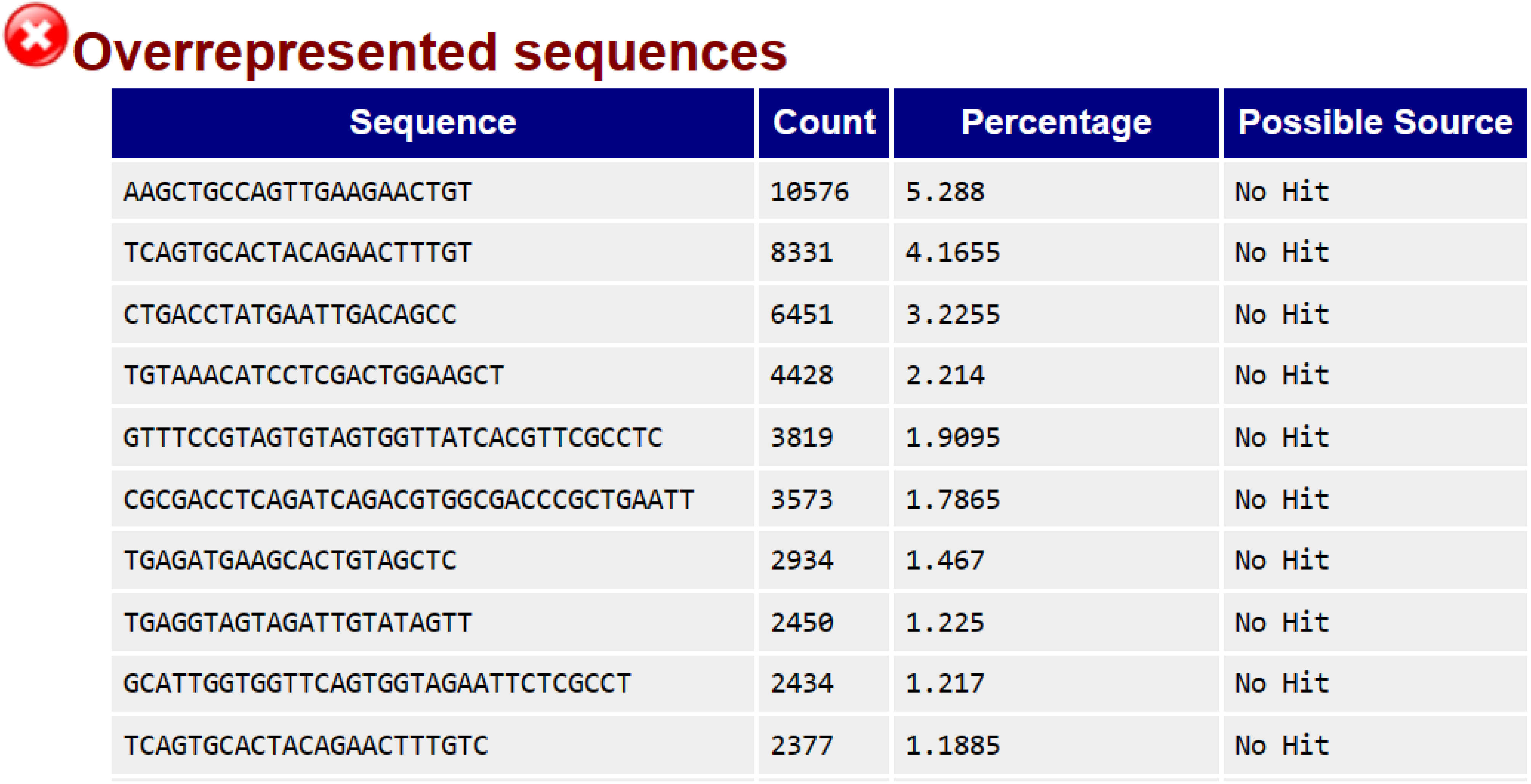
To estimate duplication the reads are trimmed to the first 50 bp and matched against each other

Only 100,000 reads are assessed

We should not rely on sequence matches to identify duplicates. They are detected as reads mapping to the same coordinates. Duplicate sequences must be removed from the analysis in variant calling

QC practice

A sequence is considered over-represented when it occupies over 0.5% of the library

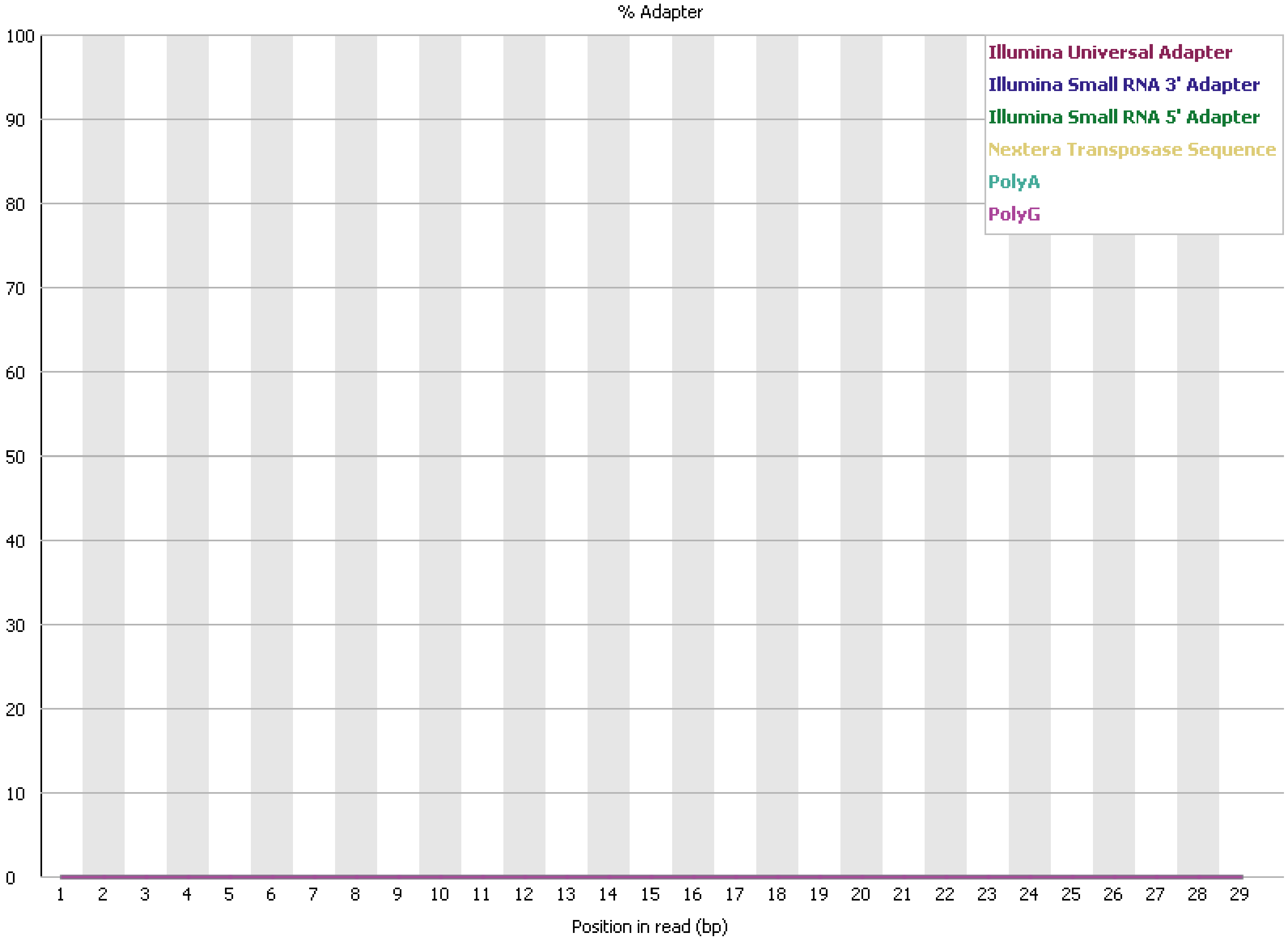


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCTGCCAGTTGAAGAACTGT	10576	5.288	No Hit
TCAGTGCACTACAGAACTTTGT	8331	4.1655	No Hit
CTGACCTATGAATTGACAGCC	6451	3.2255	No Hit
TGTAAACATCCTCGACTGGAAGCT	4428	2.214	No Hit
GTTTCCGTAGTGTAGTGGTTATCACGTTCGCCTC	3819	1.9095	No Hit
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATT	3573	1.7865	No Hit
TGAGATGAAGCACTGTAGCTC	2934	1.467	No Hit
TGAGGTAGTAGATTGTATAGTT	2450	1.225	No Hit
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCT	2434	1.217	No Hit
TCAGTGCACTACAGAACTTTGTC	2377	1.1885	No Hit

QC practice

Percentage of sequences attributed to adapters

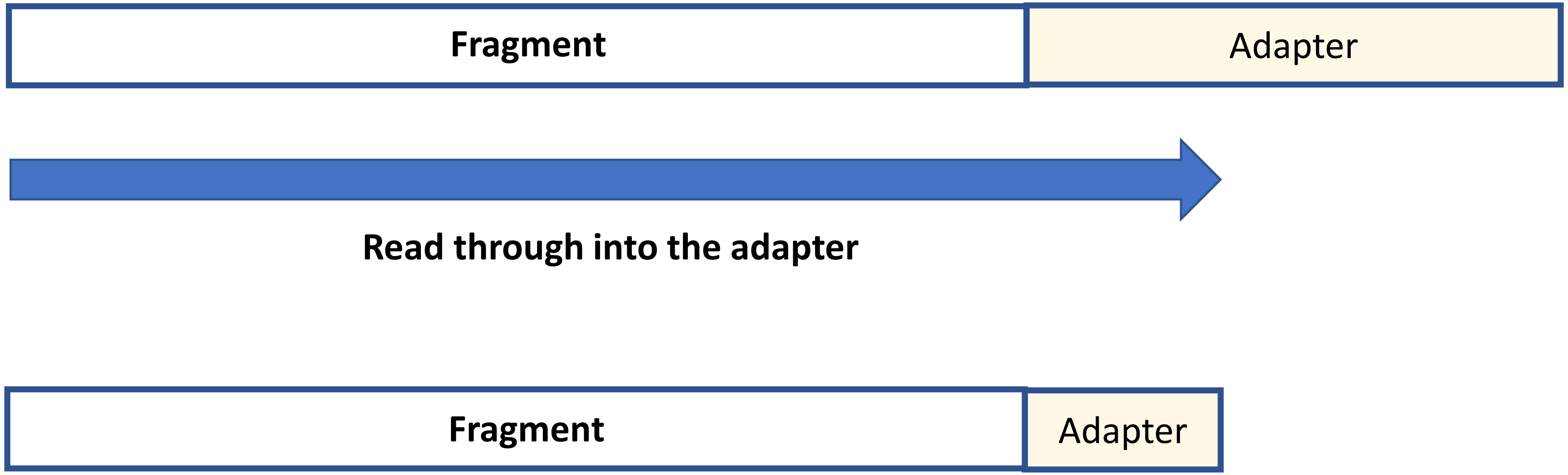


FastQC can automatically detect a number of adapters

We can also add more adapter or other contaminant sequences to **adapter.list** file in FastQC configuration directory

Adapter trimming

Our reads may contain partial adapter sequences if the number of sequencing cycles exceeds the length of the fragment



Fragment

Adapter

Read through into the adapter

Fragment

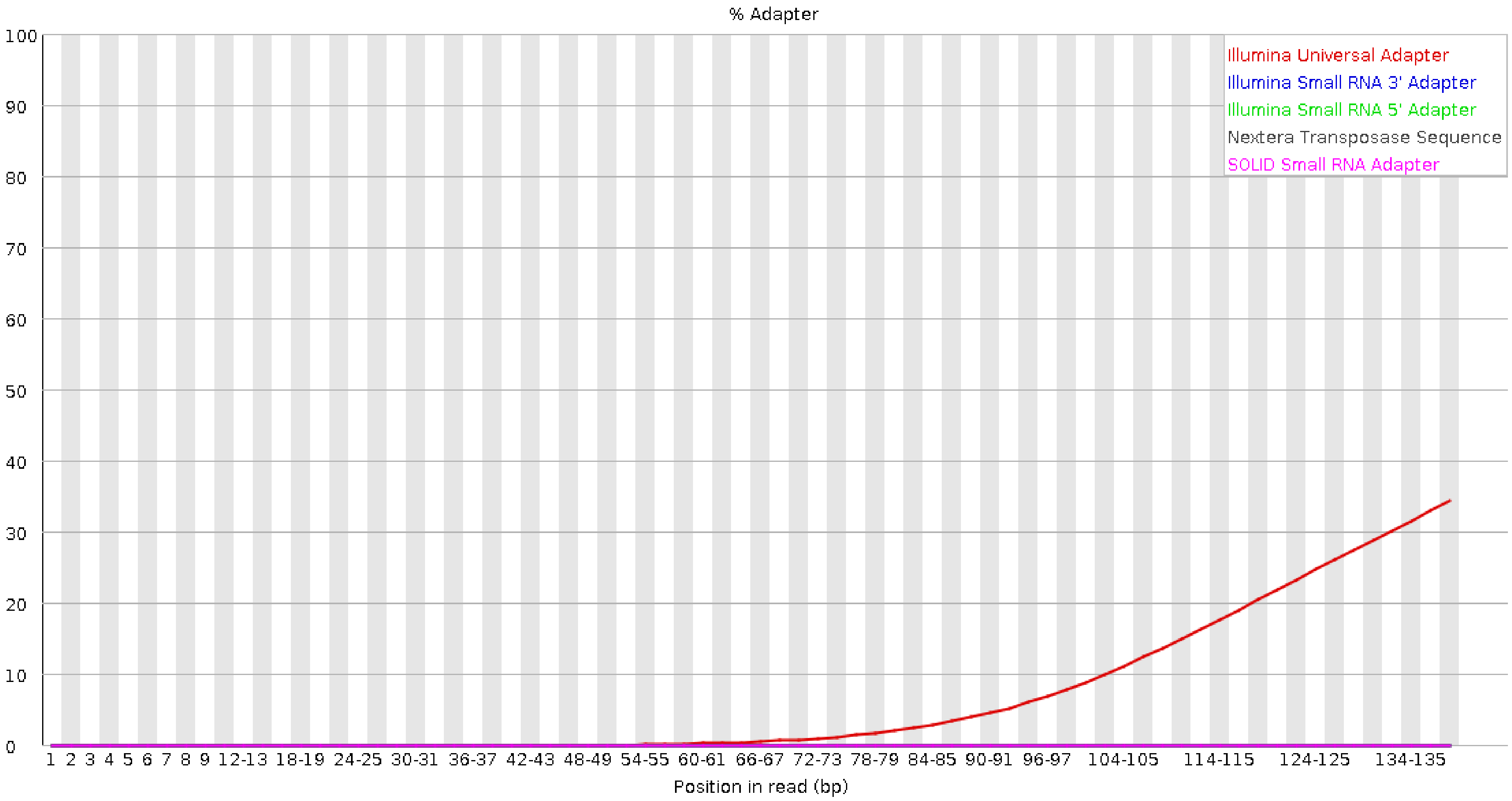
Adapter

Resulting read with partial adapter sequence

Adapter trimming

Adapter contamination

☒ Adapter Content



Adapter trimming

Adapter trimmers

Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Trimmomatic https://github.com/usadellab/Trimmomatic

Cutadapt https://cutadapt.readthedocs.io/en/stable/

Bbduk.sh https://manpages.ubuntu.com/manpages/focal/man1/bbduk.sh.1.html

and a few others

I use Trim Galore! And Trimmomatic in my projects, both have nice interface and can handle paired-end reads without breaking the order of the reads in read1 and read2 files

Adapter trimming

Download human small RNA fastq contaminated with adapters

```
$ wget https://raw.githubusercontent.com/slavailn/bioinf_training/main/smallrna_adapt_cont.fastq
```

Run fastqc

```
$ fastqc smallrna_adapt_cont.fastq
$ firefox sample_mm_srna_fastqc.html
```

Take a look at the options available with Trim Galore!

```
$ trim_galore --help
```

Run trim galore analysis, compress the output and use fastQC on the trimmed file

```
$ trim_galore --small_rna -Q 30 --gzip --fastqc smallrna_adapt_cont.fastq # this will produce trimmed reads file, trimming report and fastqc report, trimmed files will have the extension trimmed
```

Check the results

```
$ firefox smallrna_adapt_cont_trimmed_fastqc.html
```

Adapter trimming

Install Trimmomatic

Take a look at Trimmomatic help file

```
$ trimmomatic -h
```

Read the docs

https://github.com/usadellab/Trimmomatic

Download adapter sequences from Trimmomatic github page

```
$ wget https://raw.githubusercontent.com/usadellab/Trimmomatic/main/adapters/TruSeq3-SE.fa
```

Trim the reads with Trimmomatic

```
$ trimmomatic SE -phred33 smallrna_adapt_cont.fastq smallrna_trimmomatic.fastq
ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:40
```

See next page for options explanation

Adapter trimming

- Remove adapters **(ILLUMINACLIP:TruSeq3-SE.fa:2:30:10)**

- Remove leading low quality or N bases (below quality 3) **(LEADING:3)**

- Remove trailing low quality or N bases (below quality 3) **(TRAILING:3)**

- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 20 **(SLIDINGWINDOW:4:20)**

- Drop reads below the 20 bases long **(MINLEN:36)**

Check the results with fastQC

Were the adapters removed?