

FastQC

Contents

1	Overview	2
2	Basic Statistics	2
3	Per Base Sequence Quality	3
4	Per Tile Sequence Quality	4
5	Per Sequence Quality Scores	4
6	Per Base Sequence Content	5
7	Per Sequence GC Content	5
8	Per Base N Content	6
9	Sequence Length Distribution	6
10	Sequence Duplication Levels	7
11	Overrepresented Sequences	7
12	Adapter Content	8

1 Overview

For RNA-seq data, **FastQC** is a popular non-R program for QC.

```
# S1
fastqc --noextract Data/RNAseq.fastq.gz --quiet --outdir Output/
```

```
## application/gzip
```

The output looks like this:

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](#)

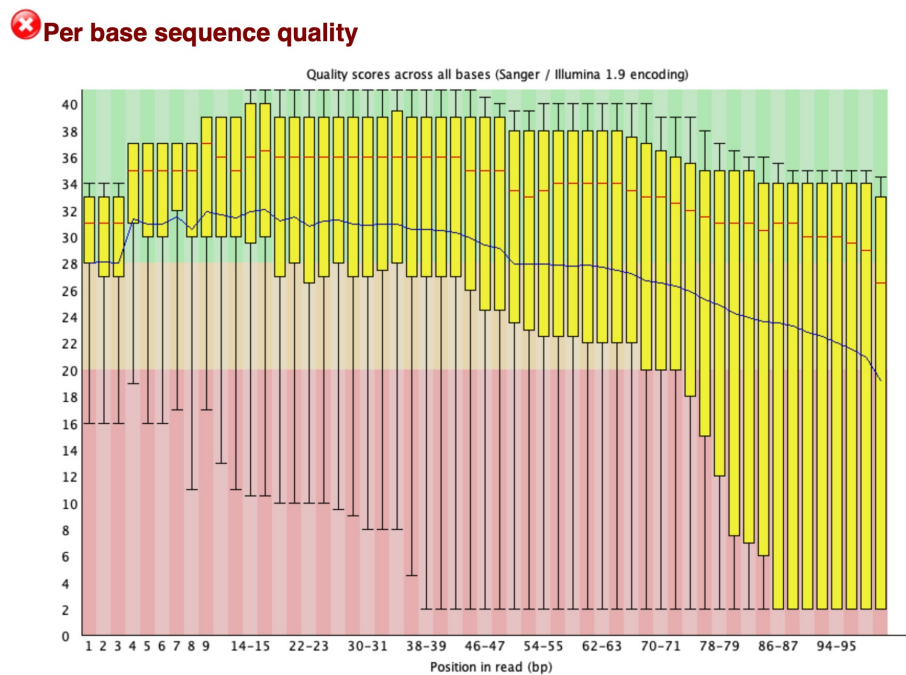


Figure 1: The summary panel of a bad example.

Sometimes we see warnings (orange triangles) and failures (red crosses) for good data and passes (green ticks) for questionable data. So, all modules should be carefully evaluated.

2 Basic Statistics

The basic statistics module contains the overview information and statistics of the reads in the input FASTQ file.

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

Figure 2: A good basic statistics report example.

Good single-cell data usually have:

- a very little number of poor quality sequences;
- most often have uniform sequence length;
- the GC content should match the overall GC content of the genome or transcriptome of the sequenced species.

3 Per Base Sequence Quality

The per base sequence quality view contains a Box Whisker type plot for each position in the read, which shows the range of quality scores across all bases at each position along all reads in the file.

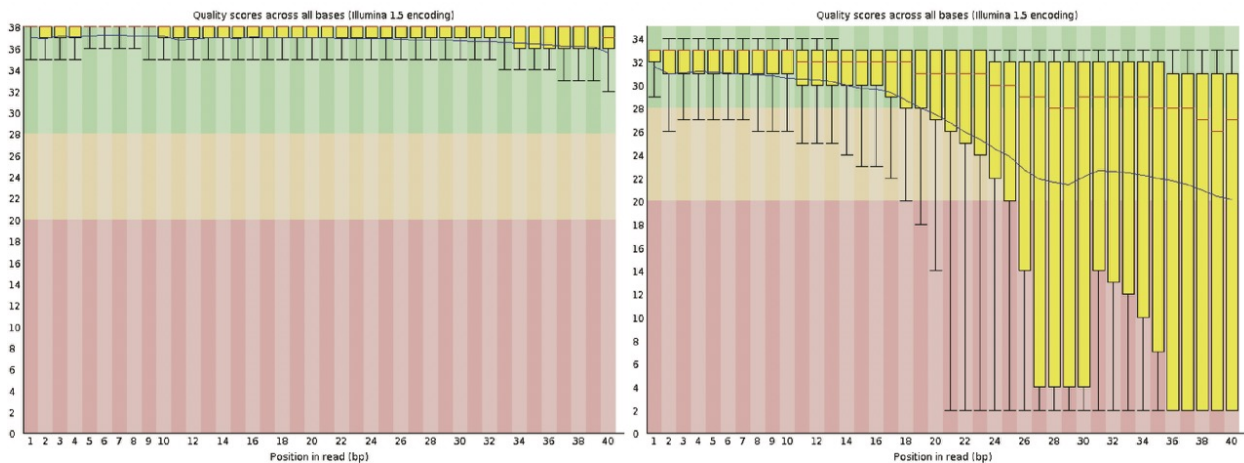


Figure 3: A good (left) and a bad (right) per-read sequence quality graph.

For single-cell data of good quality:

- all yellow boxes in the view, which represent IQR, should fall into the green (calls of good quality) area;
- so do all the whiskers, which represent the 10% and 90% of the distribution.

Focus: If poor quality calls are observed, one may consider performing quality trimming of their reads.

4 Per Tile Sequence Quality

The per tile sequence quality plot shows the deviation from the average quality for the reads in each flowcell tile that was sequenced. The “hotter” the color, the larger the deviation. For high-quality data, we should see blue over the entire plot.

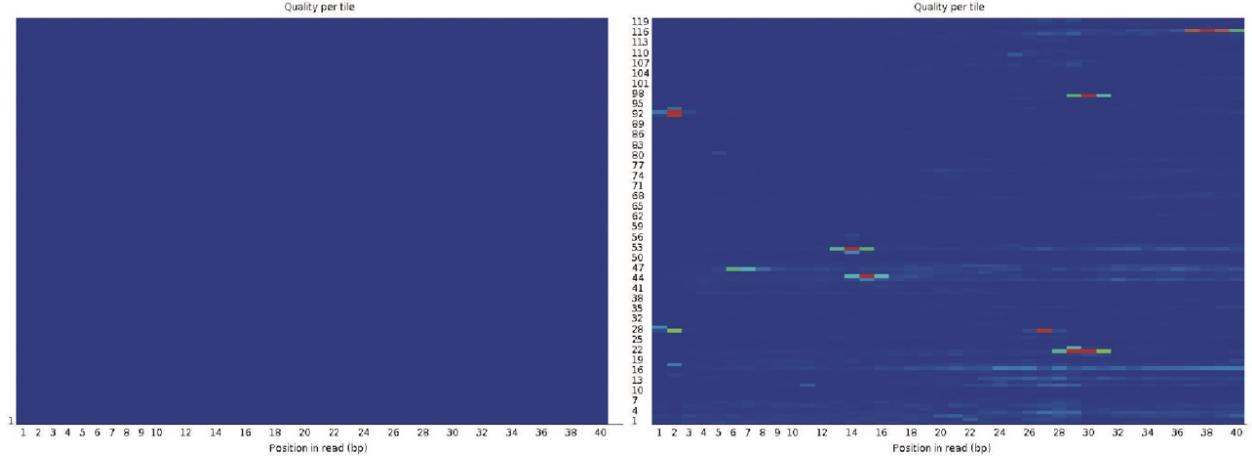


Figure 4: A good (left) and a bad (right) per tile sequence quality view.

If only part of a flowcell has poor quality, some hot colors will appear in the plot. In that case, the sequencing step might have encountered transient problems, such as bubbles going through the flowcell or smudges and debris inside the flowcell lane.

5 Per Sequence Quality Scores

The per sequence quality score plot shows the distribution of the average quality score of each read in the file. The x-axis shows the average quality scores, and the y-axis represents the frequency of each quality score.

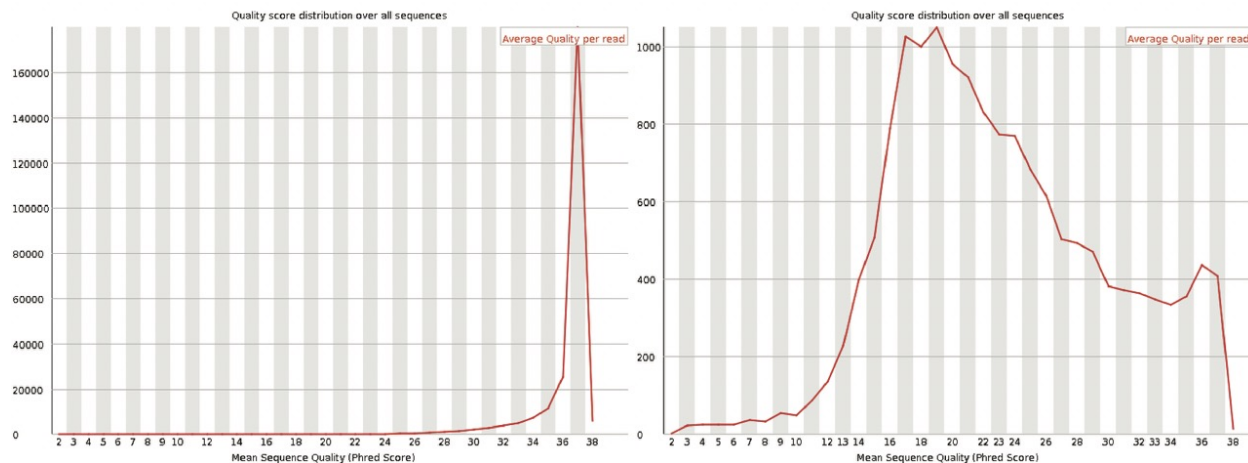


Figure 5: A good (left) and a bad (right) per sequence quality score plot.

For good data, this plot should have only one peak at the tail.

6 Per Base Sequence Content

The per base sequence content plot reports the percent of each base position of all reads in the file for which each of the four nucleotides has been called.

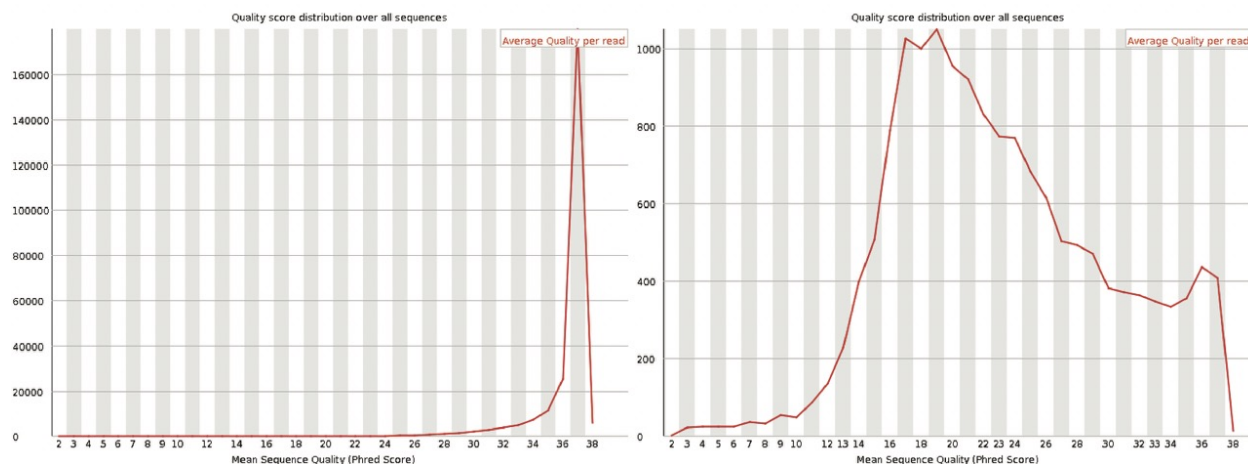


Figure 6: A good (left) and bad (right) per base sequence content plot.

For single-cell data, it is not uncommon to see fluctuations at the start of the lines because the first bases of reads represent the sequence of the priming sites, which may not be perfectly random. This happens frequently in RNA-seq libraries, even though FastQC will call a warning or failure.

7 Per Sequence GC Content

The per sequence GC content plot shows the distribution of GC content over all of the reads in red and a theoretical (expected) distribution in blue. The central peak of the observed distribution should correspond to the overall GC content of the underlying transcriptome.

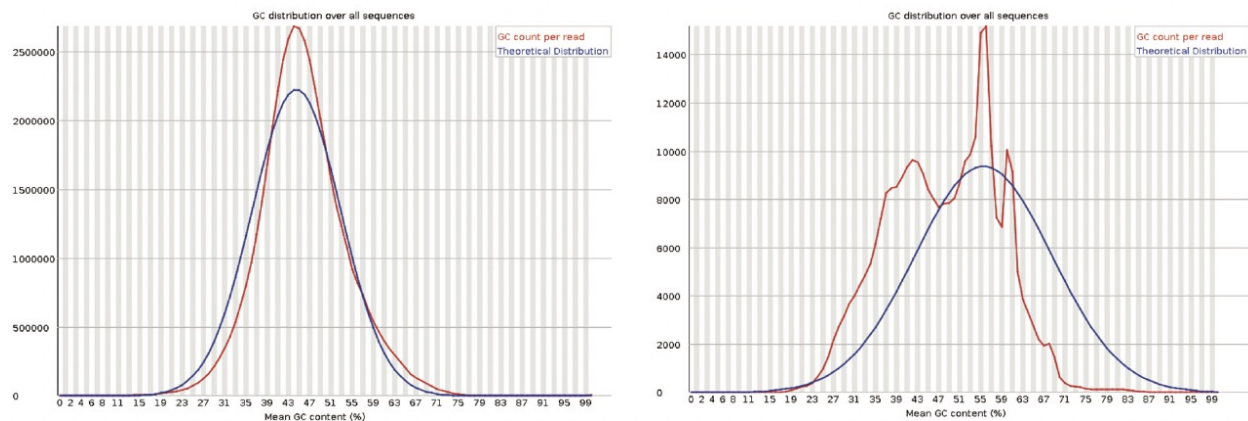


Figure 7: A good (left) and a bad (right) per sequence GC content plot.

A complex distribution usually indicates a contaminated library.

8 Per Base N Content

The per base N content plot shows the percent of bases at each position for which an N was called, which will occur when the sequencer has insufficient confidence to make a base call.

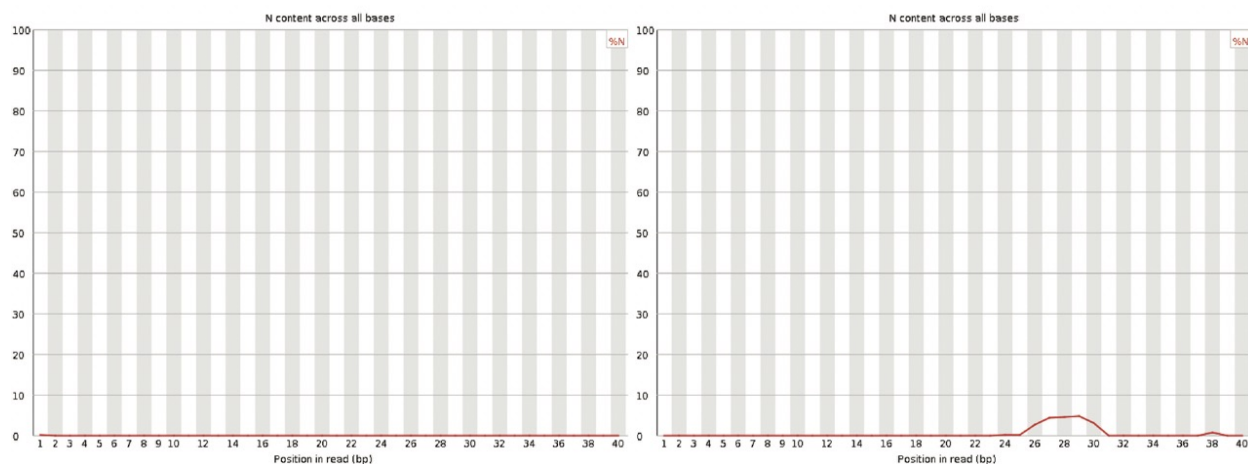


Figure 8: A good (left) and a bad (right) per base N content plot.

In a high-quality library, we should not expect any noticeable non-zero N content throughout the line.

9 Sequence Length Distribution

The Sequence length distribution graph shows the distribution of the read lengths.

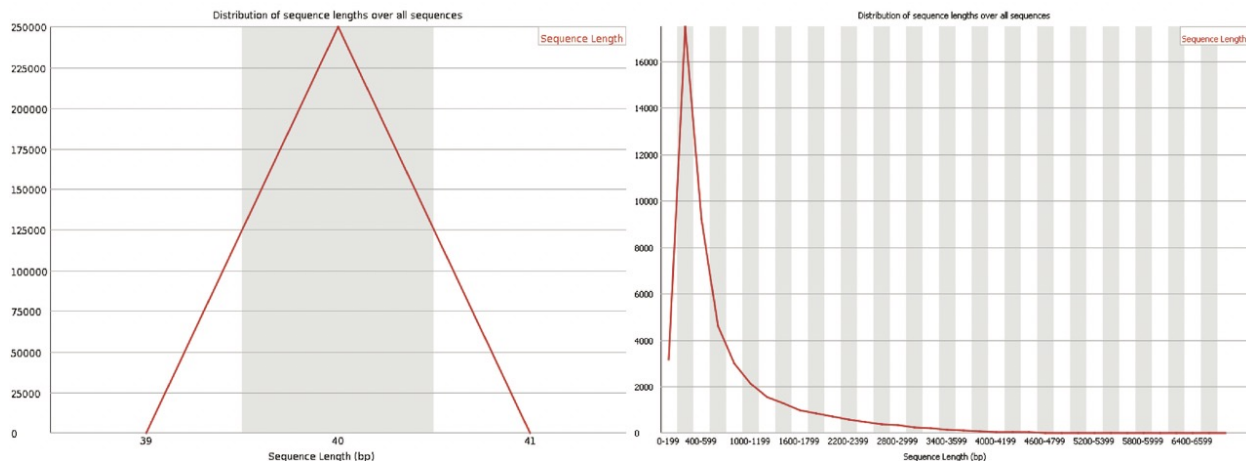


Figure 9: A good (left) and a bad (right) sequence length distribution plot.

In most single-cell chemistries, all reads will be of the same length. If trimming was performed before quality assessment, there may be some small variation in read lengths.

10 Sequence Duplication Levels

The sequence duplication level plot shows the distribution of the degree of duplication for read sequences (the blue line) and those after deduplication.

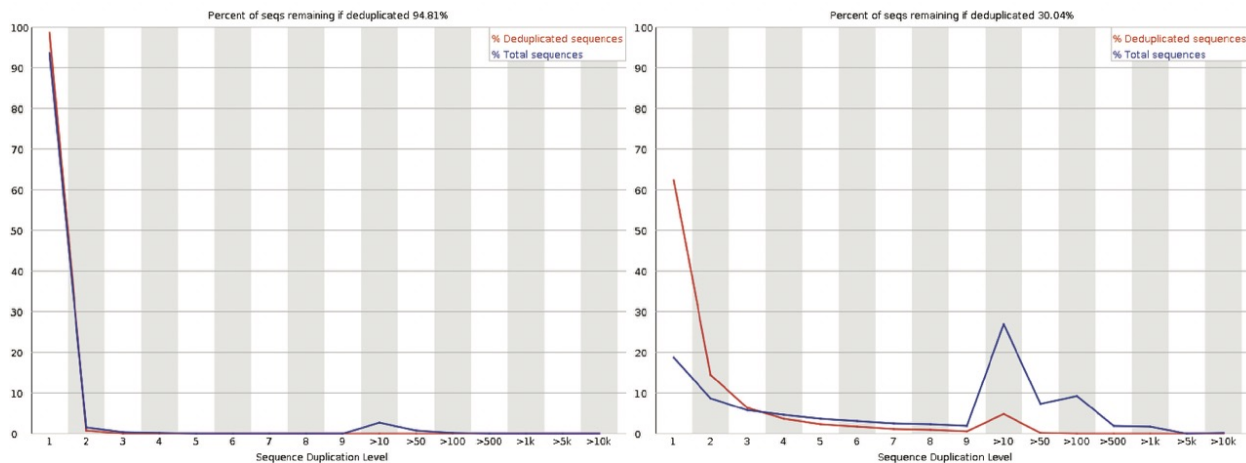


Figure 10: A good (left) and a bad (right) per sequence duplication levels plot.

As most single-cell platform requires multiple rounds of PCR, highly expressed genes usually express a large number of transcripts, and FastQC itself is not UMI aware, it is common that a small subset of sequences may have a large number of duplications. Still, the majority of sequences should have a low duplication level.

11 Overrepresented Sequences

The overrepresented sequences module lists all read sequences that make up more than 0.1% of the total number of sequences.

Sequence	Count	Percentage	Possible Source
AGAGTTTTCGCTTCCATGACGCAGAAGTTAACTTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAATGATTGGCGTATCCAACCTGCAGAGTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAATGATTGGCGTATCCAACCTGCAGAGTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit

Figure 11: An overrepresented sequence table.

We might see some overrepresented sequences from the highly expressed genes after PCR amplification, but most sequences should not be overrepresented. Note that if the possible source of the overrepresented sequences is not “No Hit”, it might indicate that the library is contaminated by the corresponding type of source.

12 Adapter Content

The adapter content module shows the cumulative percentage of the fraction of reads in which each of the adapter sequences has been observed at each base position.

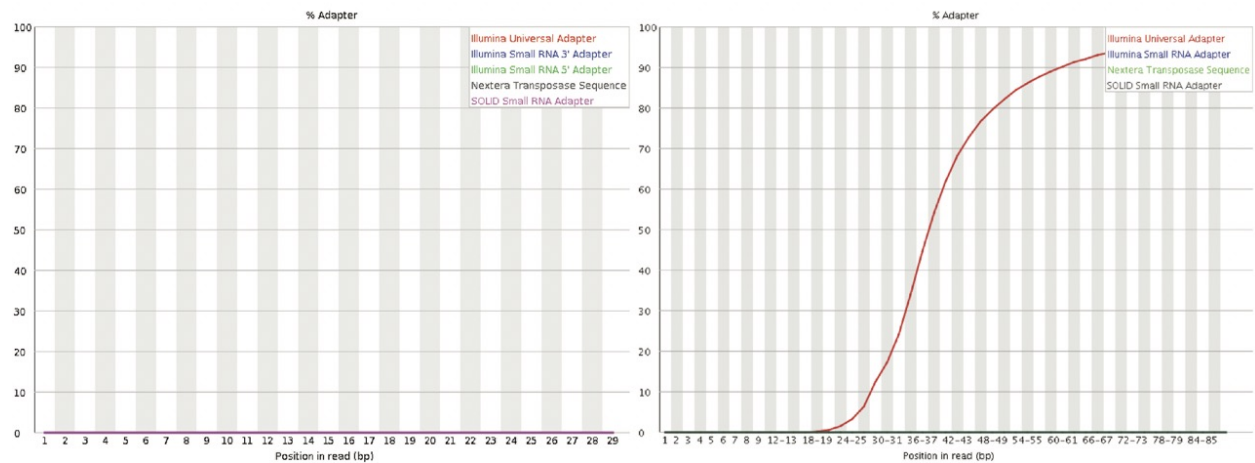


Figure 12: A good (left) and a bad (right) per sequence quality score plot.

Ideally, we should not see any abundant adapter sequences in the data.