QUALITY CONTROL OF FASTQ FILES

Jean Silva de Souza Resende[1]

[1]Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba, 81520-260, Brazil.

E-mails: jean.souza@edu.unipar.br

Contents

1. INTRODUCTION

The first step of an RNA-seq pipeline is to assess the quality of the sequenced reads. FASTQ is the standard format used for sequences generated from state-of-the-art sequencing technologies. This type of data is an evolution of FASTA, which is formed by the sum of the biological sequence with the quality information of each sequenced base. The FASTQ of a single reading record consists of four lines:

- **1st line:** starts with '@' and follows with reading information.
- **2nd row:** nucleotide sequence sequenced.
- **3rd line:** starts with '+' and follows with information from line 1, or leaves only the '+'.

- **4th line:** string representing the quality scores of each sequenced nucleotide.

There are different ranges of quality encoding - which only differs in displacement in the ASCII table. Below are the quality scores for Phred-33.

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

The quality score represents the probability that the sequenced nucleotide is incorrect. The calculation of this score uses the logarithmic base and is calculated as follows:

```
Q = -10 x log10(P)
```

P = probability that the base is wrong.

Below is another way of interpreting the data:

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%

2. HIGH-THROUGHPUT SEQUENCING QUALITY CONTROL WITH FastQC

FastQC provides a simple way to do some quality control checks on high-throughput sequencing data. It accepts BAM, SAM or FastQ files as input. Output is permanent HTML with summary graphs and tables.

2.1. FastQC Installation

```
sudo apt install fastqc
```

2.2. Run FastQC on samples

```
fastqc Examples/*fastq.gz -o Results_FastQC
```

In this code FastQC will analyze all files ending with 'fastq.gz' and the result will be saved in the folder Rsults_FastQC.

2.3. FastQC Results

FastQC has a page documented with more details about the results obtained, however, I present below a brief explanation of each module.

2.3.1. Basic Statistics

The first module presents the basic statistics of the analyzed file. For the 'SRR453566_1.fastq.gz' file in this first module, we can see that the read length is 101 base pairs and the GC content is 41%.

Measure	Value
Filename	SRR453566_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	0
Sequence length	101
%GC	41

- **Filename:** the name of the parsed file.
- **File type:** it says if it contains real bases or if it needed to convert the data.
- **Enconding:** which ASCII encoding was found.
- **Total Sequences:** count of processed sequences.
- **Sequences flagged as poor quality:** sequences removed with poor quality.
- **Sequence Length:** shorter and longer length of the set of reads.
- **%GC:** GC content in all sequences.

2.3.2. Per Base Sequence Quality

In this step, an overview of the range of quality values in all bases in each position in the file is presented. In the file we smoothed (SRR453566_1.fastq.gz), we have some base pairs that didn't get great scores at the end of some reads, this caused the quality at the end of the reads to drop. However, most of the base pairs were in the region of great score, pointing to a good sequencing.

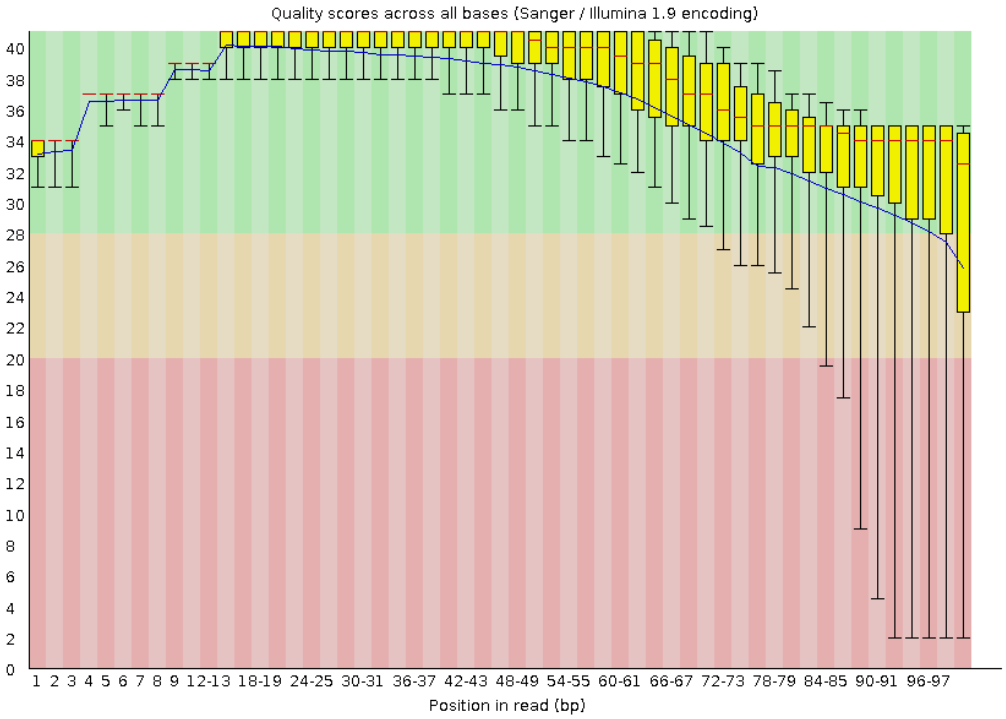

Figure 1: Quality of basis by sequence. X axis indicates nucleotide position and Y axis points to quality score. The red region (0-20) indicates that you have bases with poor quality scores and you are advised to remove these bases. The orange region (21-28) represents the acceptable region. The green region (>28) represents great score. For each position there is a BoxWhisker graph, the red line is the median value and the blue line represents the average quality.

2.3.3. Per Tile Sequence Quality

This chart is suitable for data that used Illumina library, which retains its original sequence identifiers and allows the observation of quality scores of each block, allowing the observation of quality control by the flow cell. We saw in the file we submitted for analysis that the blocks had no sequencing problems.

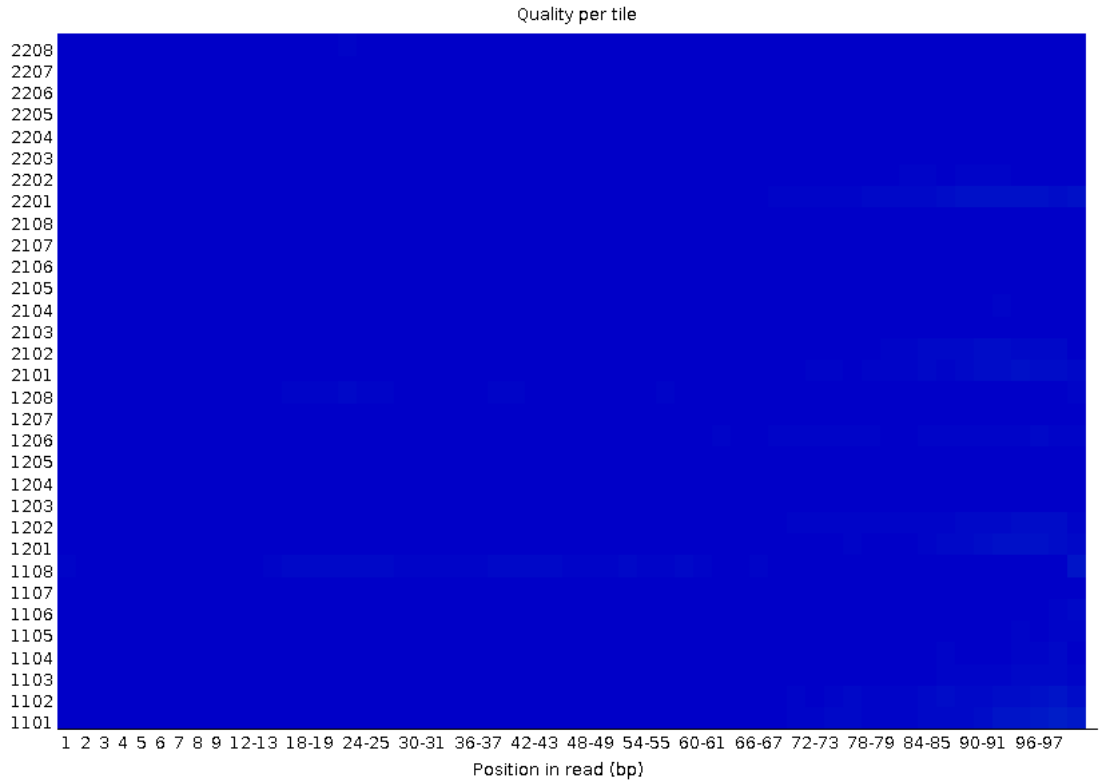

Figure 2: Sequence quality per block. Colors are on a scale from cold to warm, with cold colors being positions where the quality was at or above average for that base in the run, and warmer colors indicate that one tile had worse qualities than other tiles for that base. The X axis has the nucleotide position and the Y axis has the blocks.

2.3.4. Per Sequence Quality Scores

Quality Score by String report, allowing you to see which quality score the data is most focused on. We saw in example file R1 that the base pairs had quality scores concentrating between 25 to 39. But, they also had base pairs with low quality scores, for example 11 to 20.

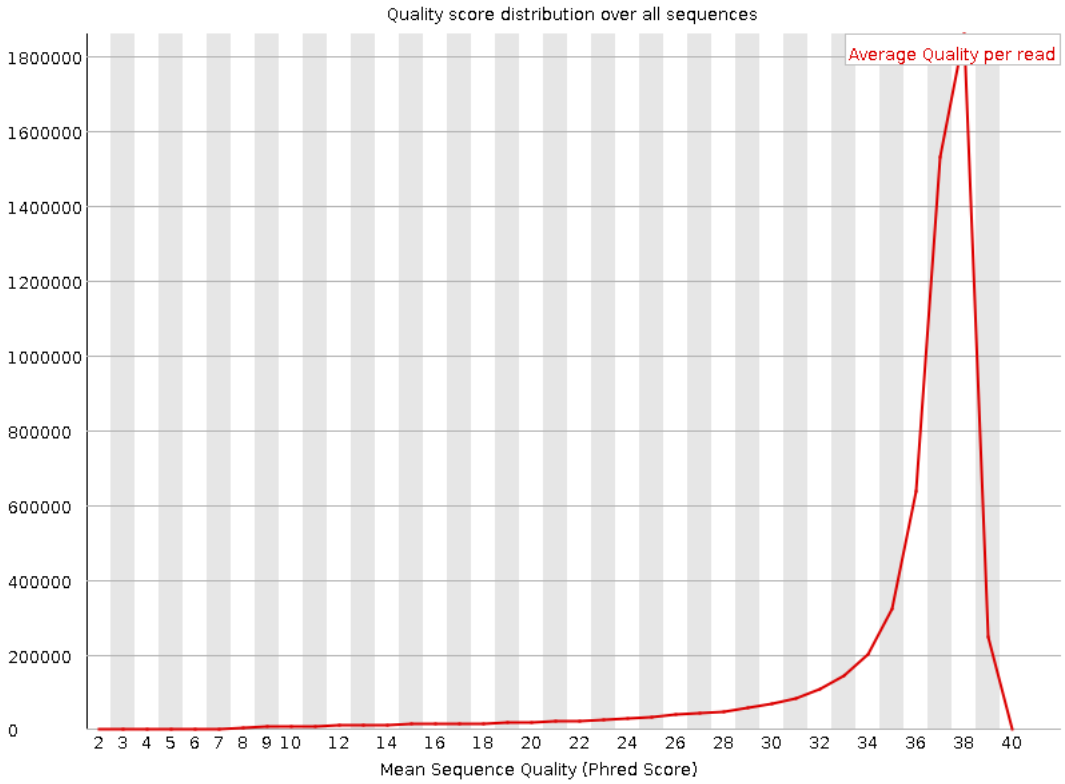

Figure 3: Quality Score by String. On the X axis the position of the nucleotide and on the Y axis the number of nucleotides.

2.3.5. Per Base Sequencec Content

In this analysis, the proportion of each nucleotide in the file is seen. Between base 1 to +-13 the proportion of base pairs varies a lot, however this result is expected for RNA-seq data (this is the case in this example) due to the hexamers used in the construction of the library.

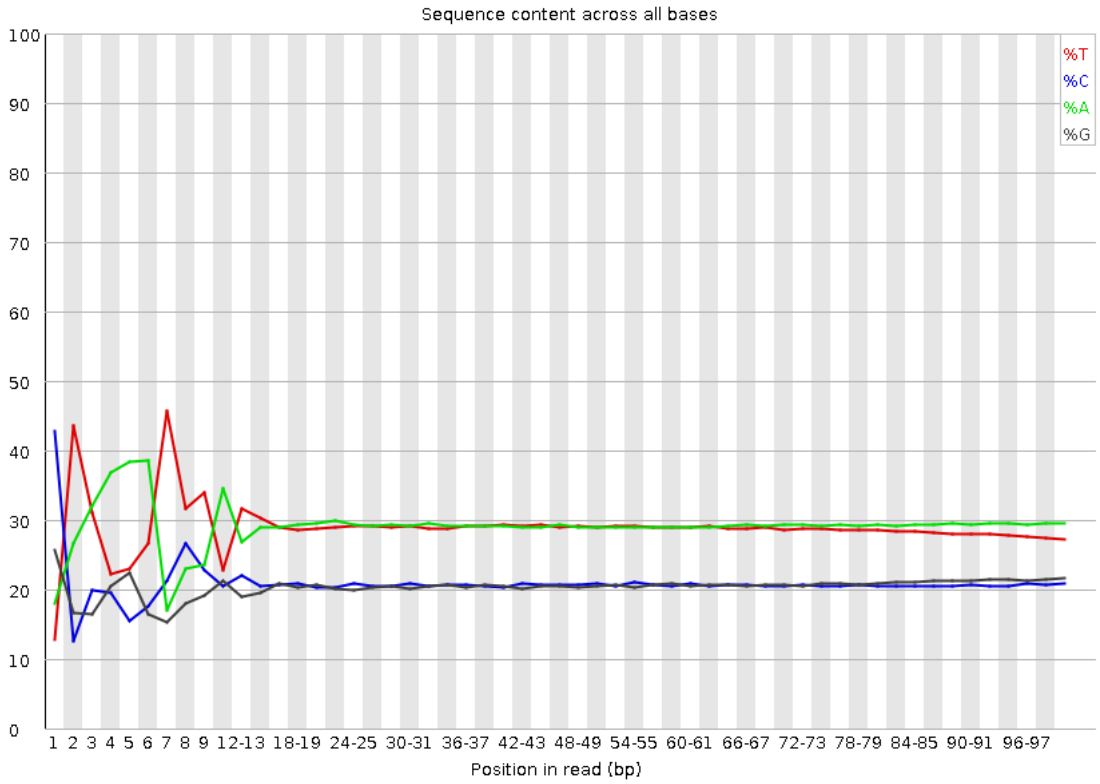

Figure 4: Base content by string. On the X axis is the position of the nucleotide and on the Y axis is the content. Each line points to a nucleotide.

2.3.6. Per Sequence GC Content

In this topic, the GC content in each sequence is evaluated and compared to a modeled normal distribution of the GC content. If the distribution format is unusual, the library may be contaminated. Higher peaks can indicate contamination by specific contaminants - adapters, for example. Wider peaks may indicate contamination from another species. The file that we sub-targeted in the analysis showed that the GC content found was close to what was expected.

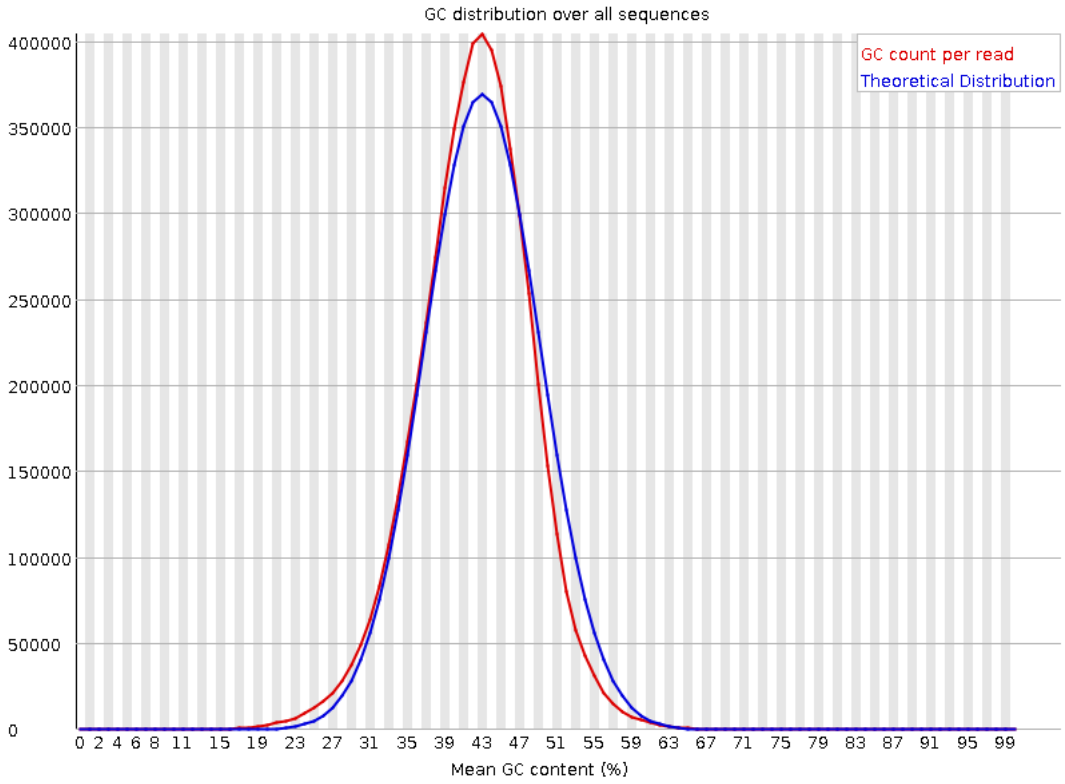

Figure 5: GC content by sequence. X axis has the GC mean and the Y axis has the GC content.

2.3.7. Per Base N Content

When a sequencer cannot correctly identify a base, it puts it as 'N'. In this graph we have the percentage of N in each position of the reads. In the file submitted for analysis, no high N contents were found in specific regions of the reads.

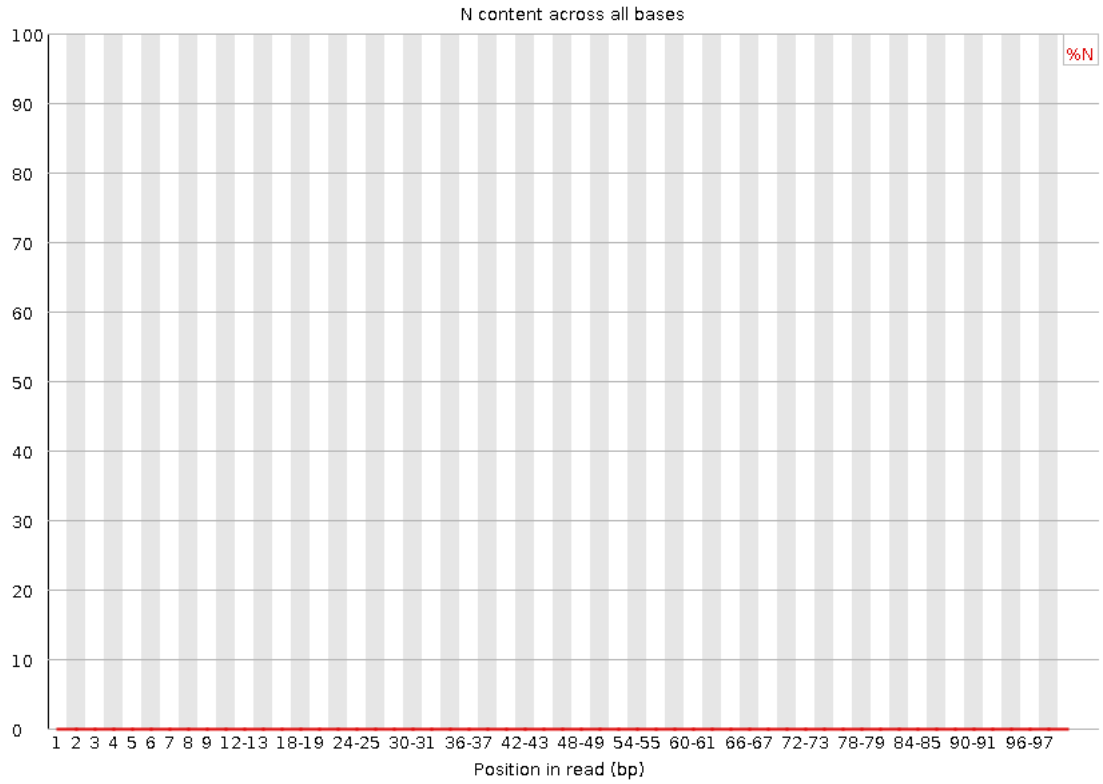

Figure 6: Content of N per base. The X axis contains the nucleotide position and the Y axis has the N content in percent.

2.3.8. Sequence Length Distribution

In this graph we see the length of the reads, some sequencers work with reads with uniform size, but after some filtering the reads can have different lengths. Our results showed that the reads are 101 base pairs long.

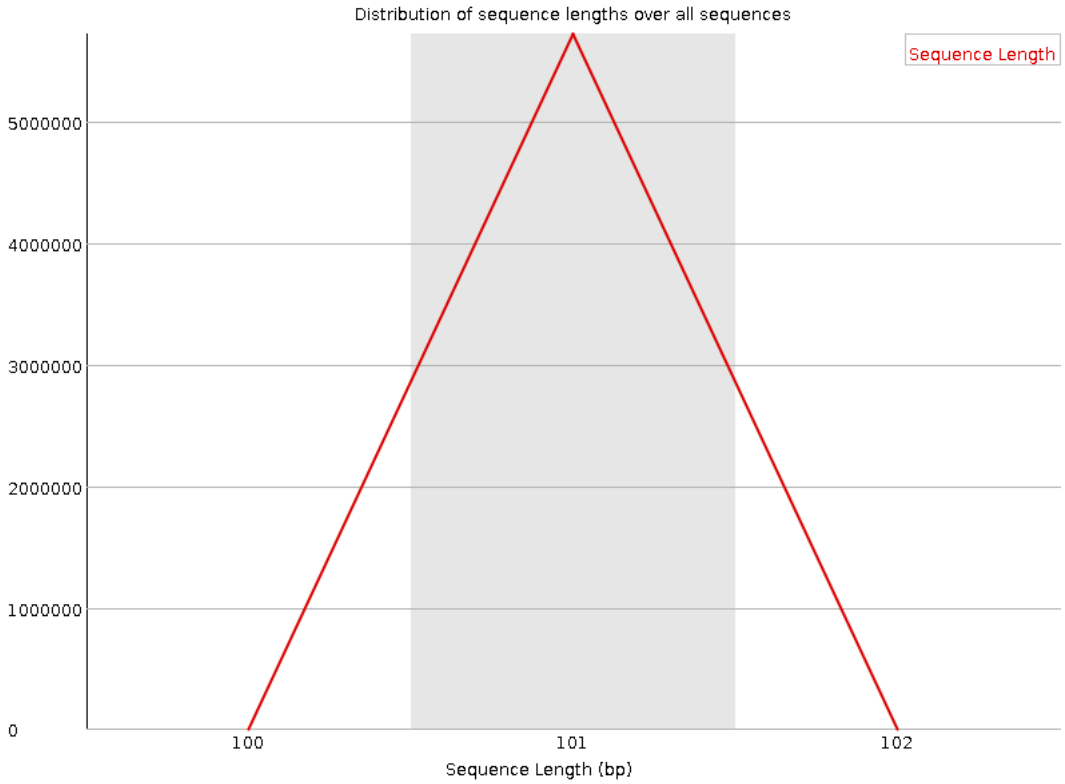

Figure 7: String length distribution. On the X axis we have the length of the read and on the Y axis we have the number of reads.

2.3.9. Duplicates Sequences

In this topic the level of duplication is seen, in the case of many duplications it can indicate a very high level of coverage of the sequence, or it can indicate a type of enrichment bias. In the results of our example we see that it has a peak with >10, >50 and >100 doublings.

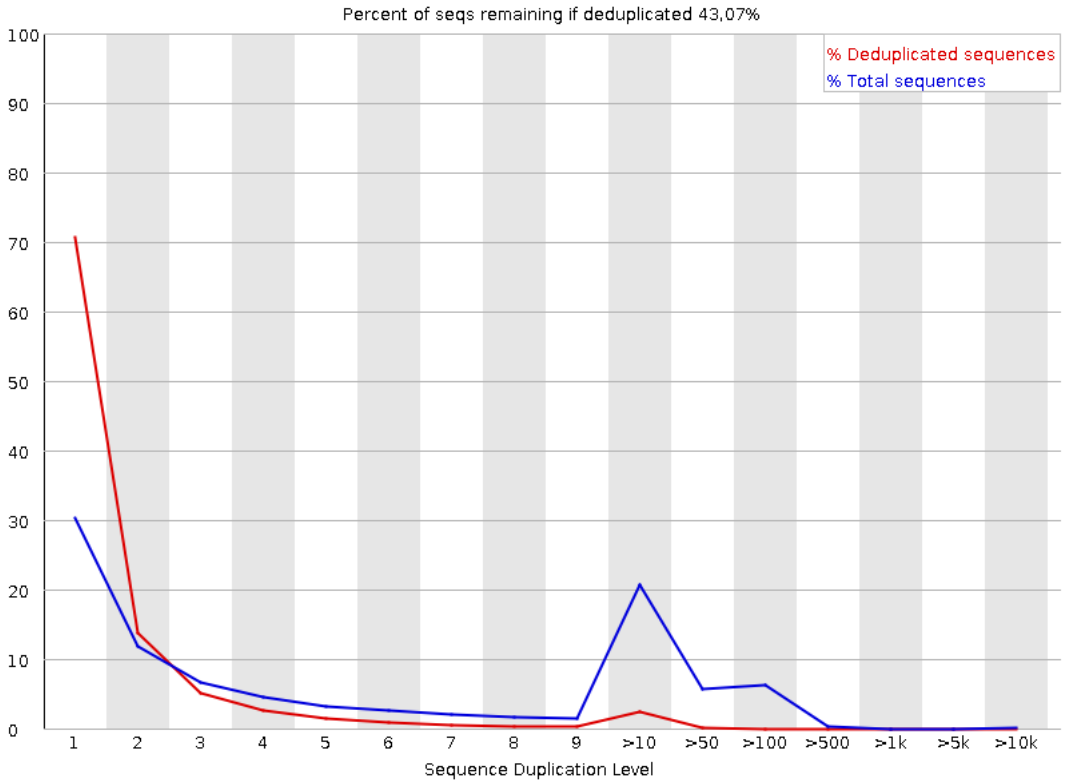

Figure 8: Duplicate sequences. X axis contains the level of duplication, Y axis contains the percentage of duplication.

2.3.10. Overrepresented Sequences

Shown here is the over-representation of sequences, the discovery of over-represented sequences can mean a high significance from a biological point of view or it can indicate a contamination of the library, or even show that the data are so diverse. In our results we see an over-represented sequence and this sequence is from an adapter.

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGC	13810	0.24119195281649677	TruSeq Adapter, Index 2 (100% over 50bp)

2.3.11. Adapter Content

In this step, the content of adapters present in the dataset is analyzed. This step comes in handy for targeting the next scans, if you have a significant amount of adapters that need to be pulled out or not. Our results show that there is adapter contamination in our data and that this adapter is an Illumina universal adapter.

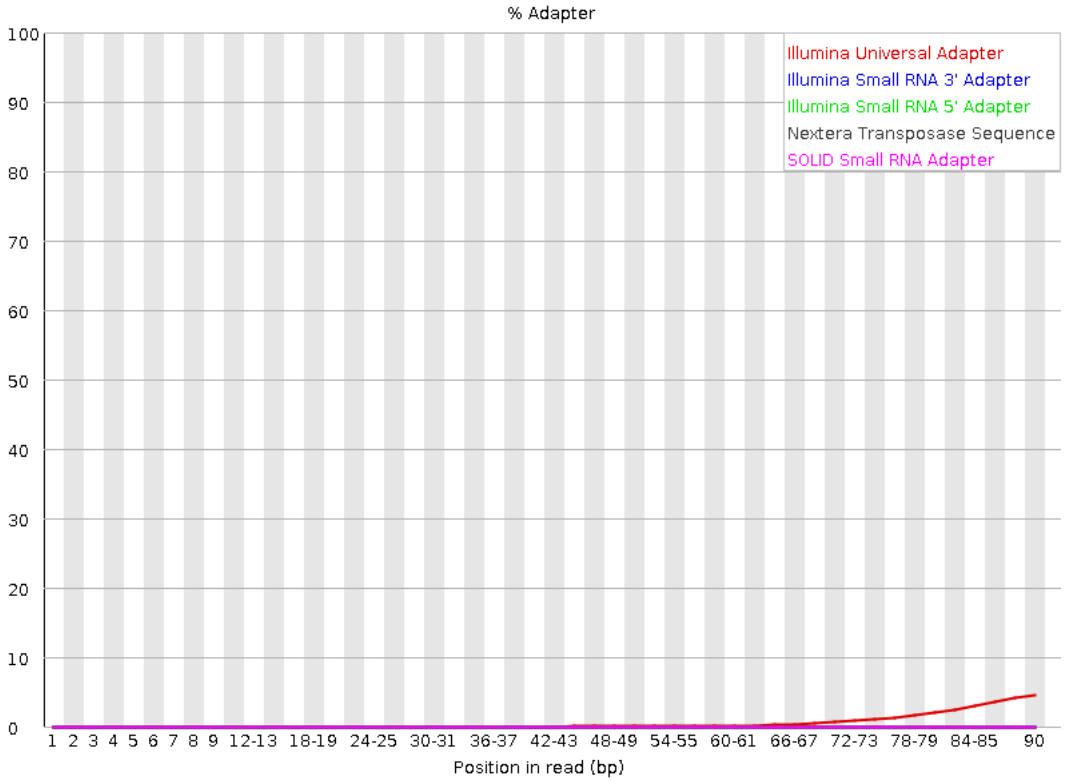

Figure 9: Adapter content. On the X axis is the position of the nucleotide in read, on the Y axis is the percentage of content.