What is a FASTQ file?

The most common raw data output from Next-Generation Sequencing (NGS) platforms is the FASTQ file, which is a text-based format

However, depending on the specific technology used, additional files may also be generated. For instance, in PacBio sequencing, the basecall File Format (bas.h5/bax.h5) is produced, which can be visualized using HDFView. On the other hand, Illumina sequencing systems like NextSeq, HiSeq, and NovaSeq 6000 generate raw data files in binary base call (BCL) format, which can be converted to FASTQ using the bcl2fastq Conversion tool. Furthermore, for the Illumina platform, the FASTQ ORA format is available, which is a binary and compressed version of the standard FASTQ file. The fastq.ora files are considerably smaller, up to 5 times, in comparison to their corresponding fastq.gz files, making them an efficient option for data storage and management.

A FASTQ file defines each read by 4 lines:

Line 1: Read sequence identifier (encoded descriptions of instrument, lane…)

Line 2: Read sequence

Line 3: « + » sign (optional: « + » followed by seq identifier)

Line 4: a string of ASCII characters = Quality score associated to each Read

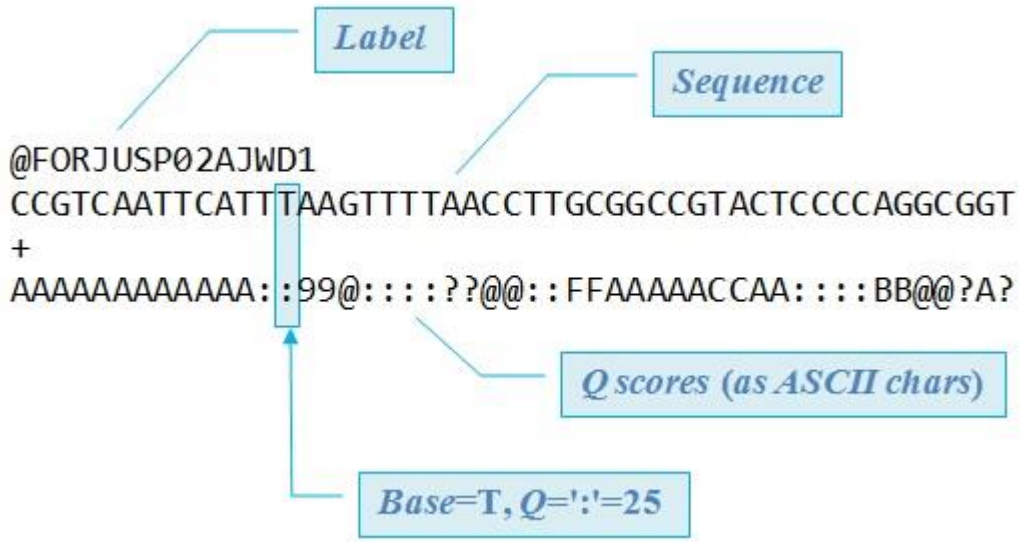


Fig 1. Diaggramatic representation of fields found in a FastQ file (https://www.drive5.com/usearch/manual/fastq_files.html)

Why do we need Quality Control?

Because of its substantial impact on the precision and dependability of biological research, quality control of sequence data is critical, particularly in fields such as genomics. This requirement is the result of numerous critical reasons. First, high-quality sequence data is critical for maintaining the accuracy and dependability of following analytical processes, as any errors or artifacts within the data can lead to misleading conclusions and interpretations.

Second, sequencing technologies can introduce a lot of technical biases and errors during the data creation phase, necessitating the use of quality control (QC) measures to detect and correct these biases, hence improving the biological accuracy of the data. Furthermore, QC allows for the early detection and deletion of low-quality data inside the analytical pipeline, which not only reduces computational burdens but also optimizes the utilization of computational resources.

Furthermore, QC contributes to cost efficiency by avoiding the need for costly re-sequencing of samples due to poor data quality. It adheres to the publication and data sharing rules established by

journals and data repositories, upholding the standards required for contributing to the scientific community.

What is FastQC?

“FastQC” aims to provide a way to do quality control checks on fastq sequence data. Quality information is contained in the fastq file that refers to the accuracy of each base call. This helps to determine any irregularities or features that may affect your results, such as adapter contamination. MultiQC is used to aggregate results from multiple FastQC runs into one single html report. A FastQC output report is described below.

1. Basic Statistics

It provides an overview of the dataset’s important statistics, enabling researchers to assess the dataset’s overall quality rapidly. The elements that can be found in a FastQC Basic Statistics report are described as follows:

- Filename: The name of the input sequencing file(s) being examined
- File Type: The input file format, such as BAM or FASTQ.
- Encoding: The method of encoding quality scores, which is frequently Illumina 1.8+ (Phred+64), Sanger (Phred+33), or another method.
- Total sequences: This value represents the size of the dataset (reads or fragments).
- Sequence Length: Details regarding the length of the dataset’s sequences, including the shortest, longest, and most frequent sequence lengths. This can confirm the sequence length that is anticipated.
- %GC: The dataset’s overall GC (guanine-cytosine) content as a percentage. Potential problems may be indicated by deviations from the intended GC content.

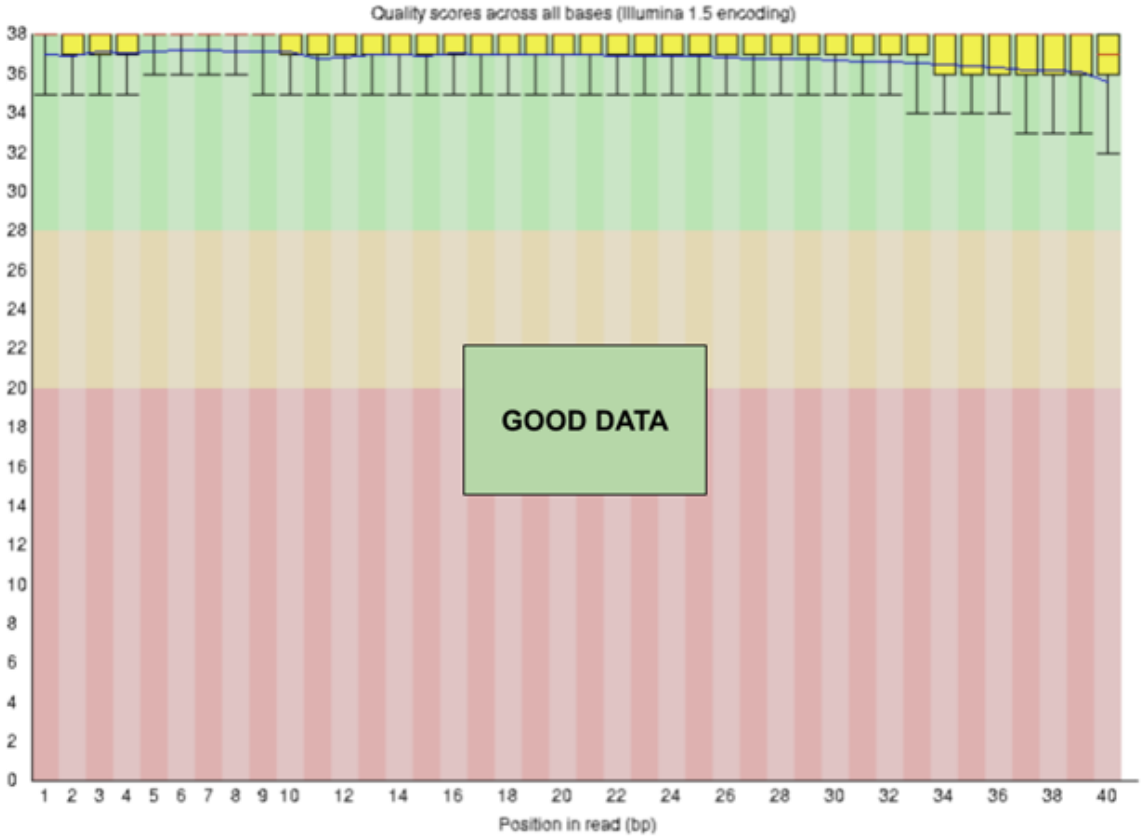
Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Total Bases	10 Mbp
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Fig 1: FASTQC Basic Statistics report

2. Per Base Sequence Quality

The graph in this section shows the quality scores for each base position in the sequencing reads. The quality score is shown by the y-axis, while the read's location is represented by the x-axis. The graph's line or bars show how the quality scores vary with the length of the reads.


Quality scores across all bases (Illumina 1.5 encoding)
GOOD DATA
Position in read (bp)

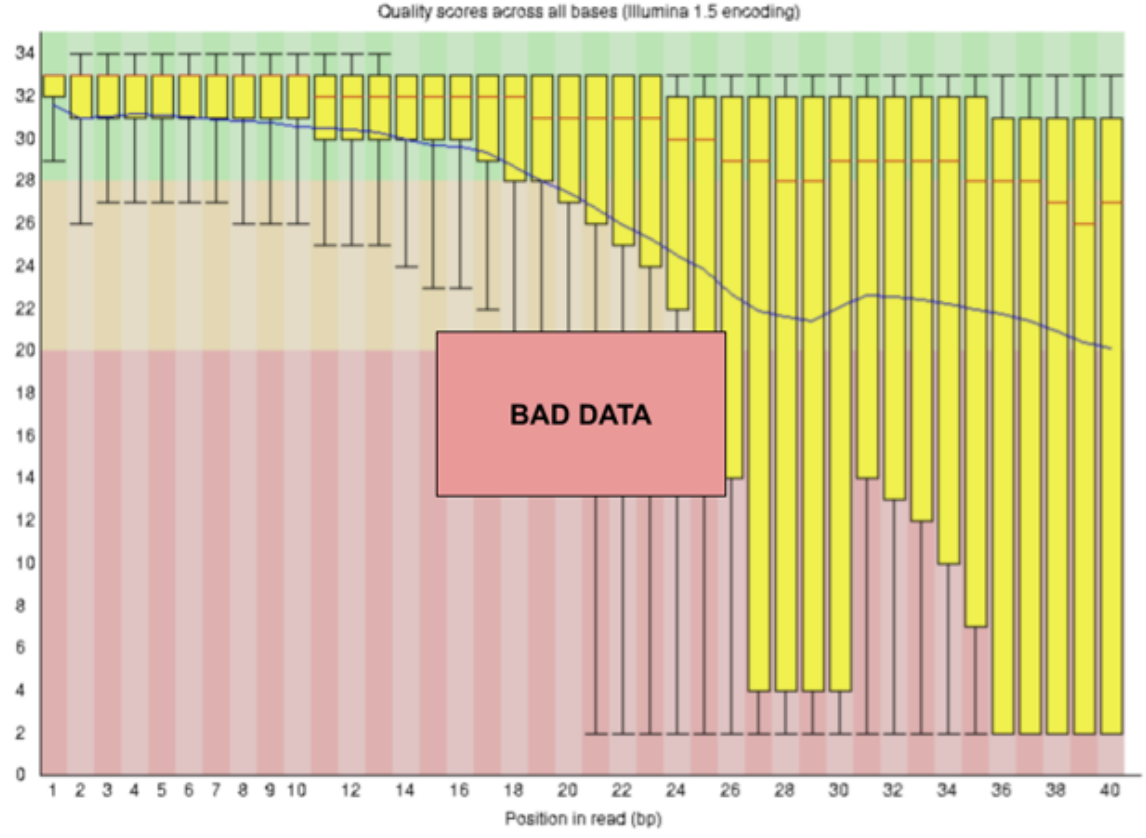

Quality scores across all bases (Illumina 1.5 encoding)
BAD DATA
Position in read (bp)

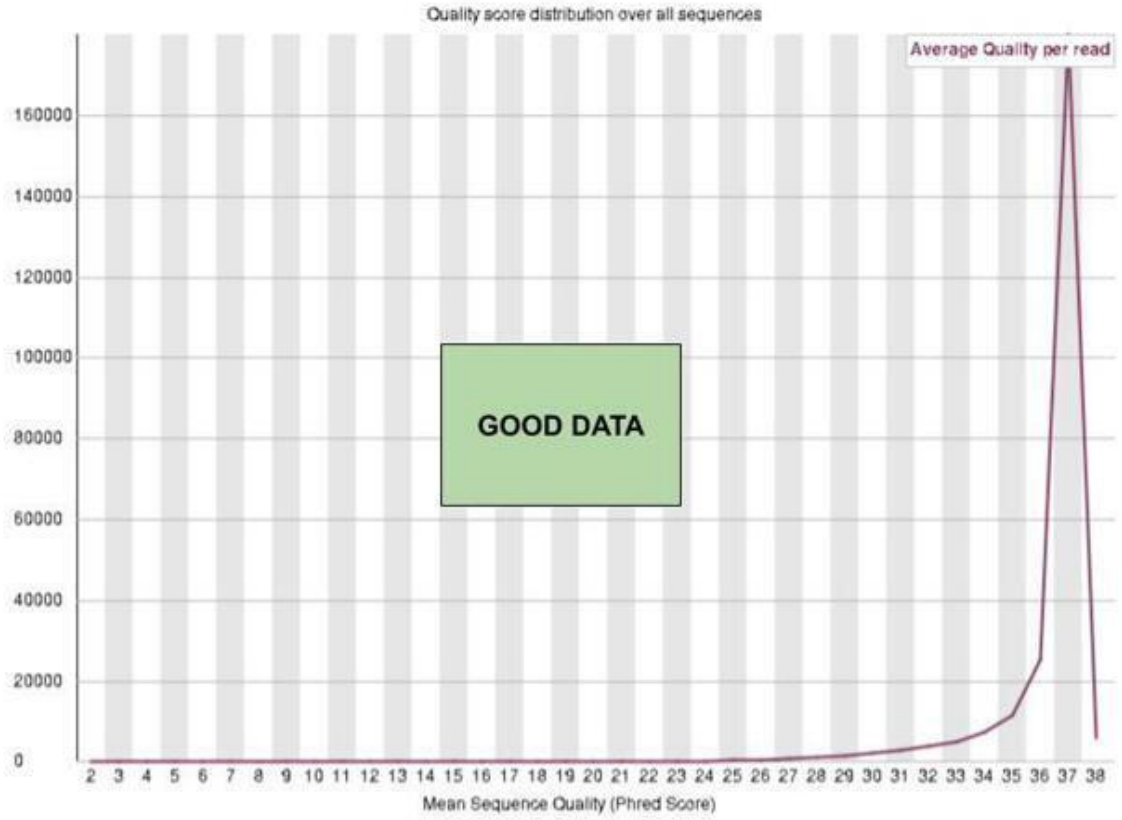
Fig 2: Plots of FASTQC Per Base Sequence Quality score with good and bad data used

3. Per Sequence Quality Scores

The graph in this section shows the average quality score distribution over all sequences. The number or percentage of sequences with that average quality score is shown on the y-axis, which is parallel to the x-axis and displays quality scores.

Warning if most frequently observed mean quality <27

Failure if most frequently observed mean quality <20



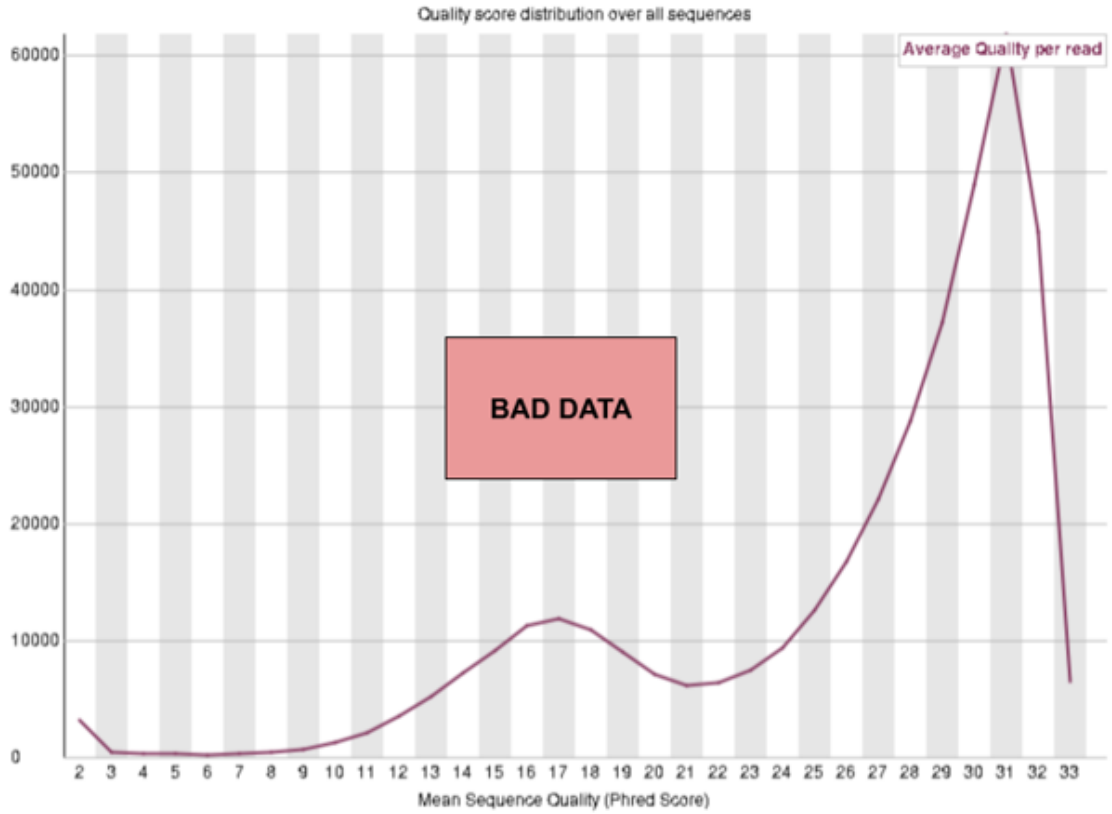
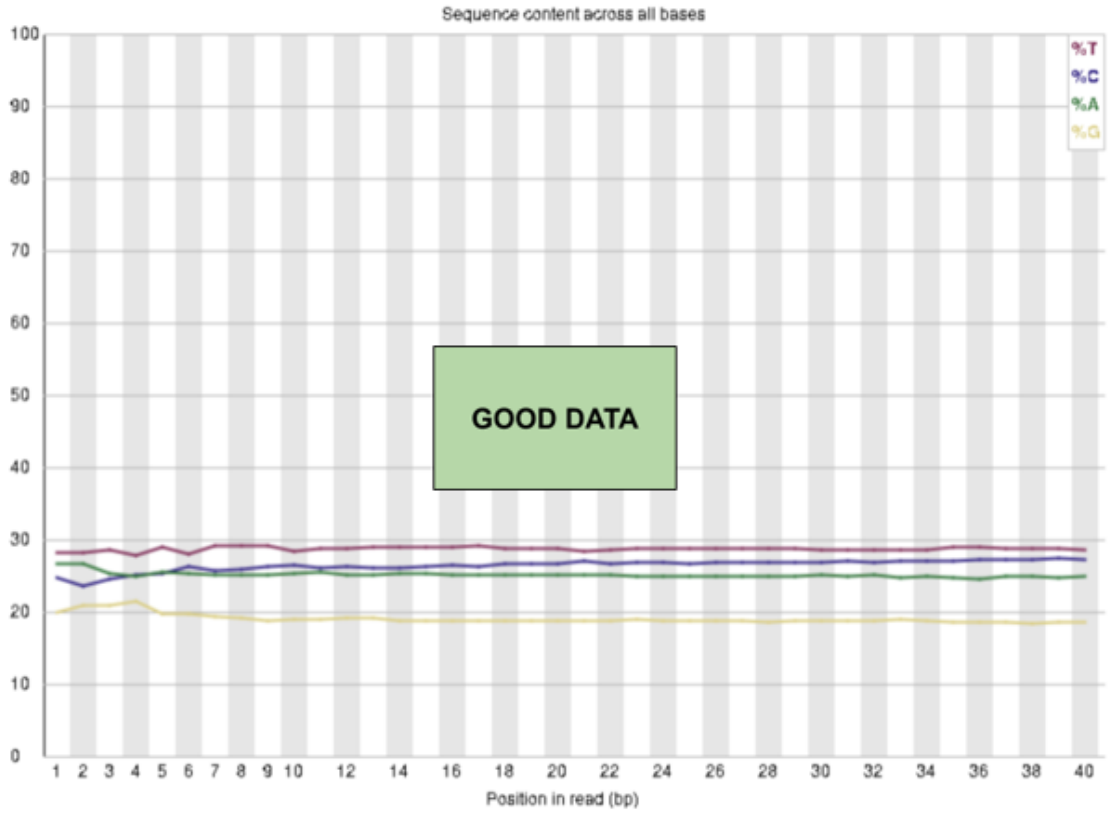


Fig 3: Graphical representation of FASTQC Per Sequence Quality score. Upper graph represents good quality data while the bottom graph shows data of low quality, and displays a bimodal or complex distribution.

4. Per Base Sequence Content

The “Per Base Sequence Content” section offers information about the distribution of nucleotides (A, T, G, and C) at each location along the sequencing runs. For a better understanding of potential sequencing biases and contaminants, this section aids in evaluating the consistency and balance of base composition across the reads.


Sequence content across all bases
%T
%C
%A
%G
GOOD DATA
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

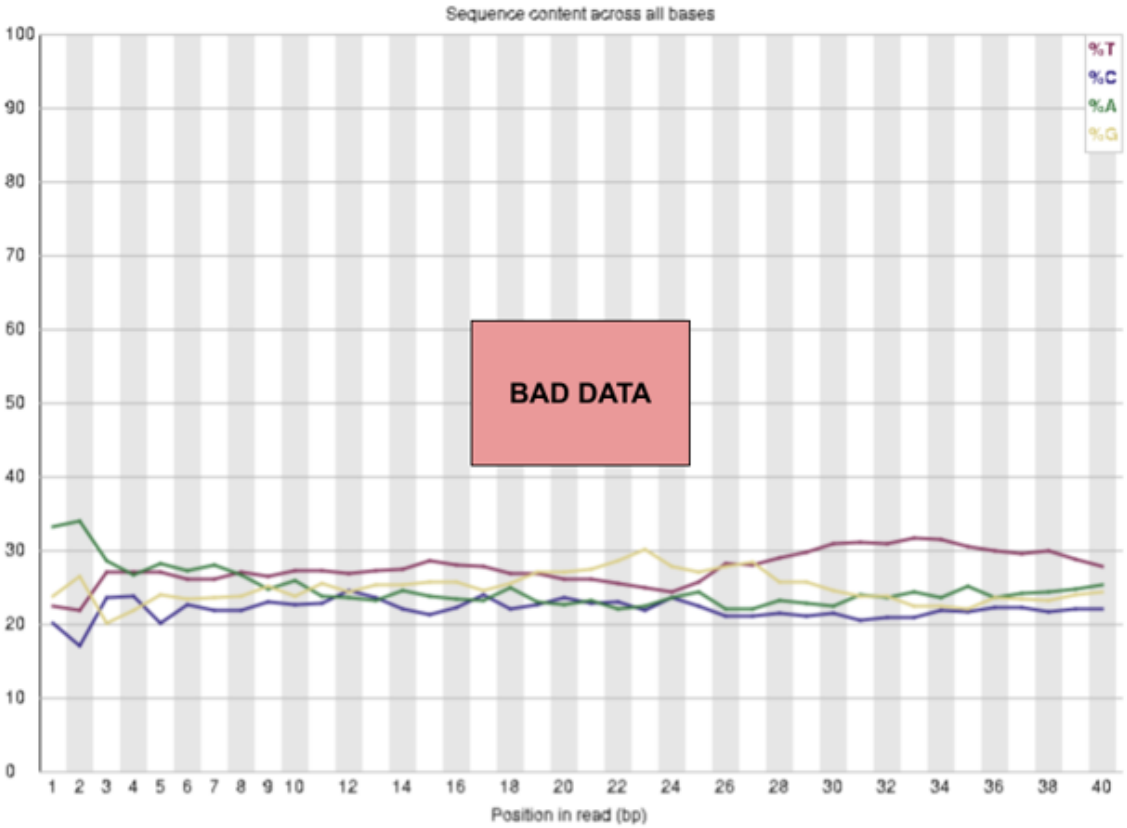
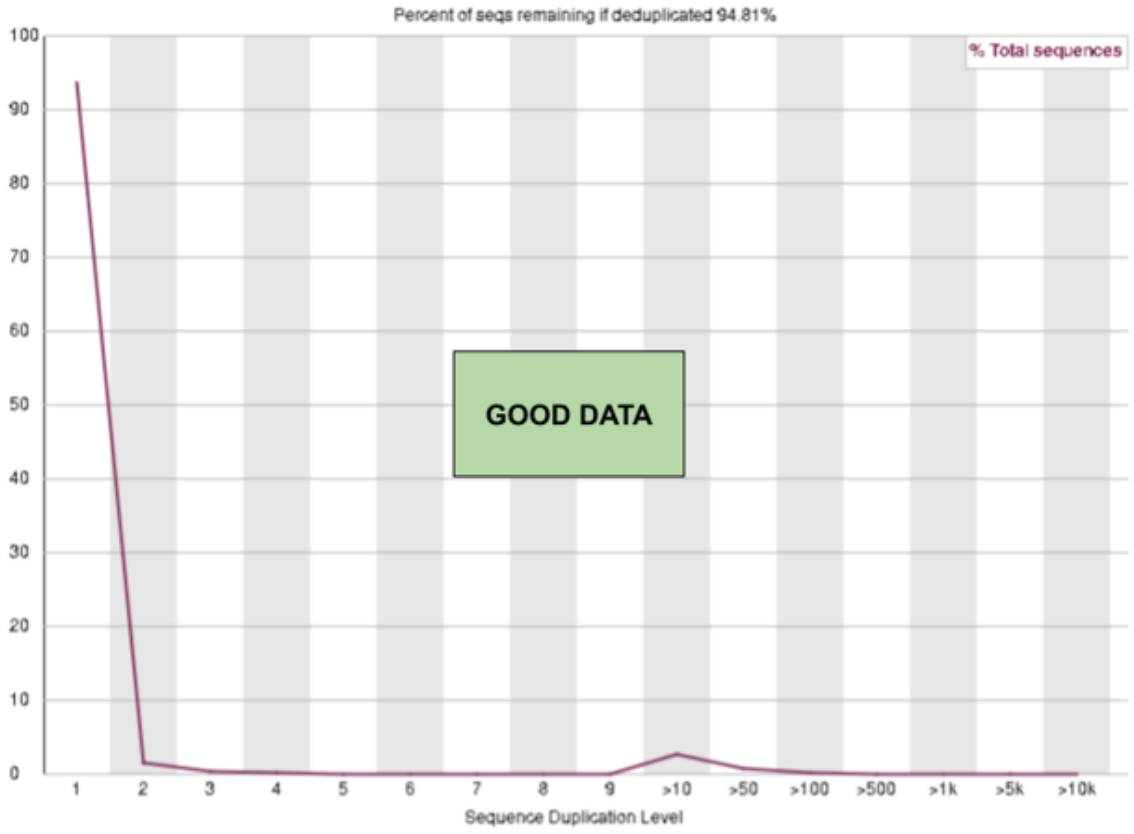

Sequence content across all bases
%T
%C
%A
%G
BAD DATA
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

Fig 4: Graphical representation of FASTQC Per Base Sequence content metric. The upper graph shows smooth calling over the read while the bottom graph shows bad data probably due to overrepresented sequences.

5. Sequence Duplication level

This metric reveals if specific sequences are overrepresented as a result of biases in the library preparation or sequencing process by giving insights into the degree of duplication within a sequencing dataset. Increased duplication levels may indicate that particular sequences were amplified or enriched more frequently than others during library preparation. The results of subsequent investigations, including variant calling or differential gene expression analysis, may become skewed as a result of biassed representation.



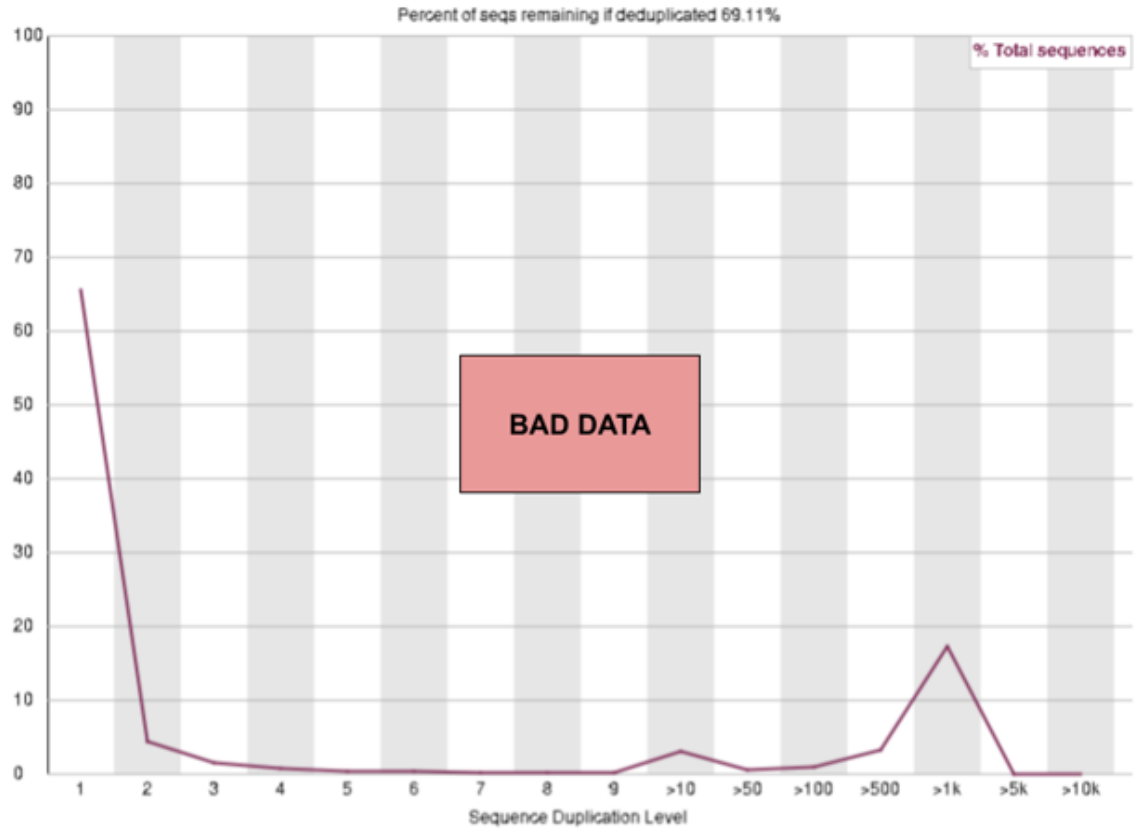


Fig 5: Graphical representation of FASTQC Sequence Duplication level metric. The upper graph represents data with low level of duplication while the bottom graph shows data with bad level of duplication.

6. Per Base N Content

This section gives details on the makeup of nitrogenous bases at each place in the sequencing reads. The balance and uniformity of the four DNA bases—adenine, thymine, guanine, and cytosine—along the length of the readings are evaluated in this section. Variations in base composition can be biologically meaningful. For instance, differing base compositions in specific genomic areas may occur naturally, and this information can be utilised to interpret the genome biologically.

There are usually several line graphs or plots in the “Per Base Nitrogen Content” section. An individual nitrogenous base (A, T, G, or C) is distributed differently along the length of each graph. The x-

axis represents the position within the read, while the y-axis shows the percentage or count of the base at that position.

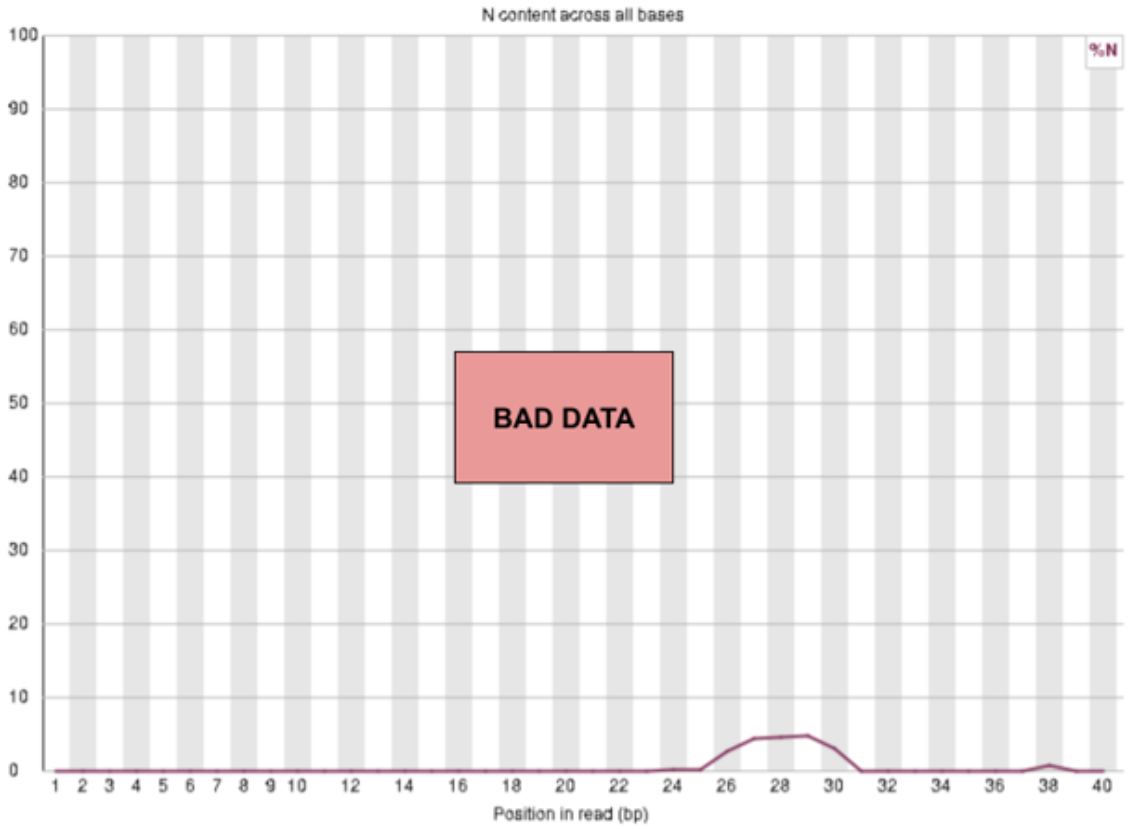


Fig 6: Graphical representation of FASTQC Per Base N Content metric. The upper graph represents data with confident base calling while the bottom graph shows data with general loss of quality.

7. Adapter content

This section gives details of the presence of sequencing adapters in raw sequencing reads. Sequencing adapters are small strands of DNA that are ligated to the ends of DNA fragments before sequencing. They provide extra sequences enabling specific primers to bind to the DNA to ensure the full DNA fragment is sequenced. High adapter content in sequencing data can indicate problems during library preparation, when not all adapters were successfully eliminated, resulting in their inclusion in the final sequenced reads. This may interfere with downstream analyses and alter the quality and accuracy of the results.

FastQC examines the distribution of adapter sequences in raw data and gives a visual depiction of their existence. This information

assists researchers in determining whether there were any issues with adapter removal during sequencing.

Warning if any sequence is present in more than 5% of all reads

Failure if any sequence is present in more than 10% of all reads

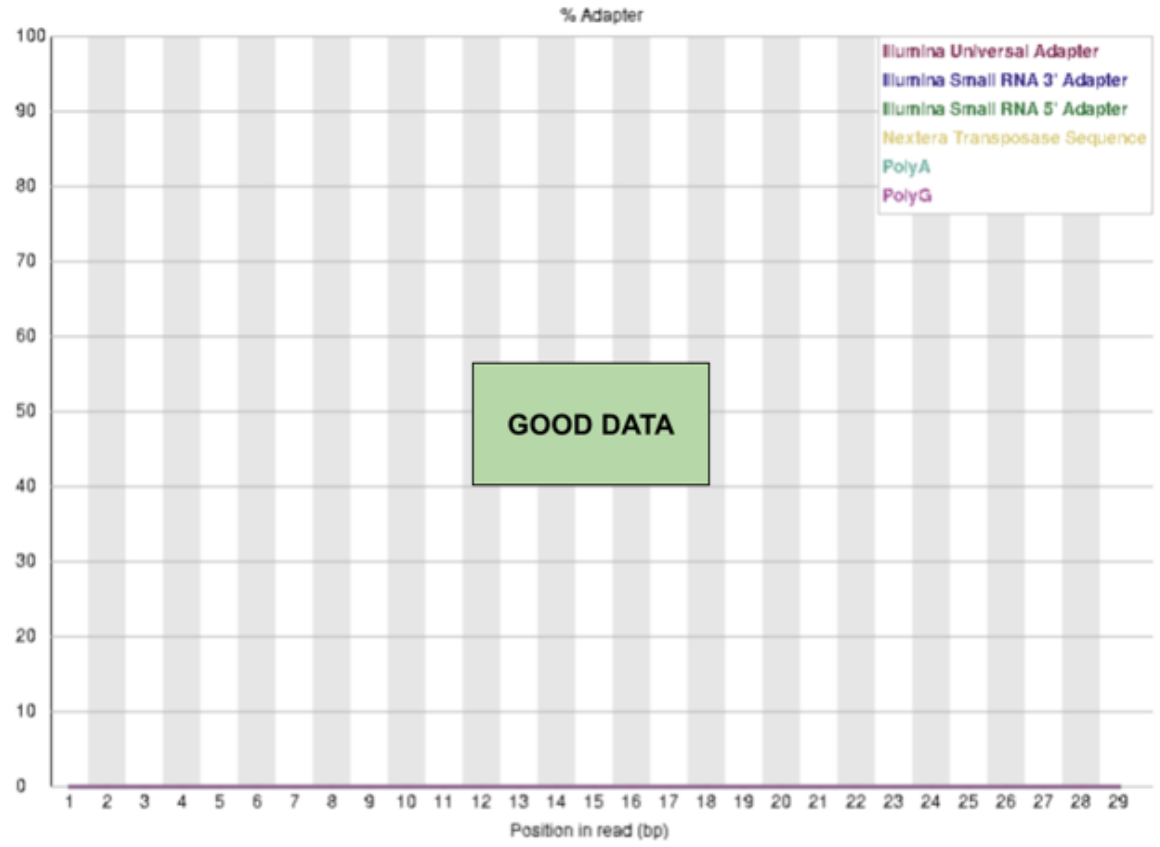
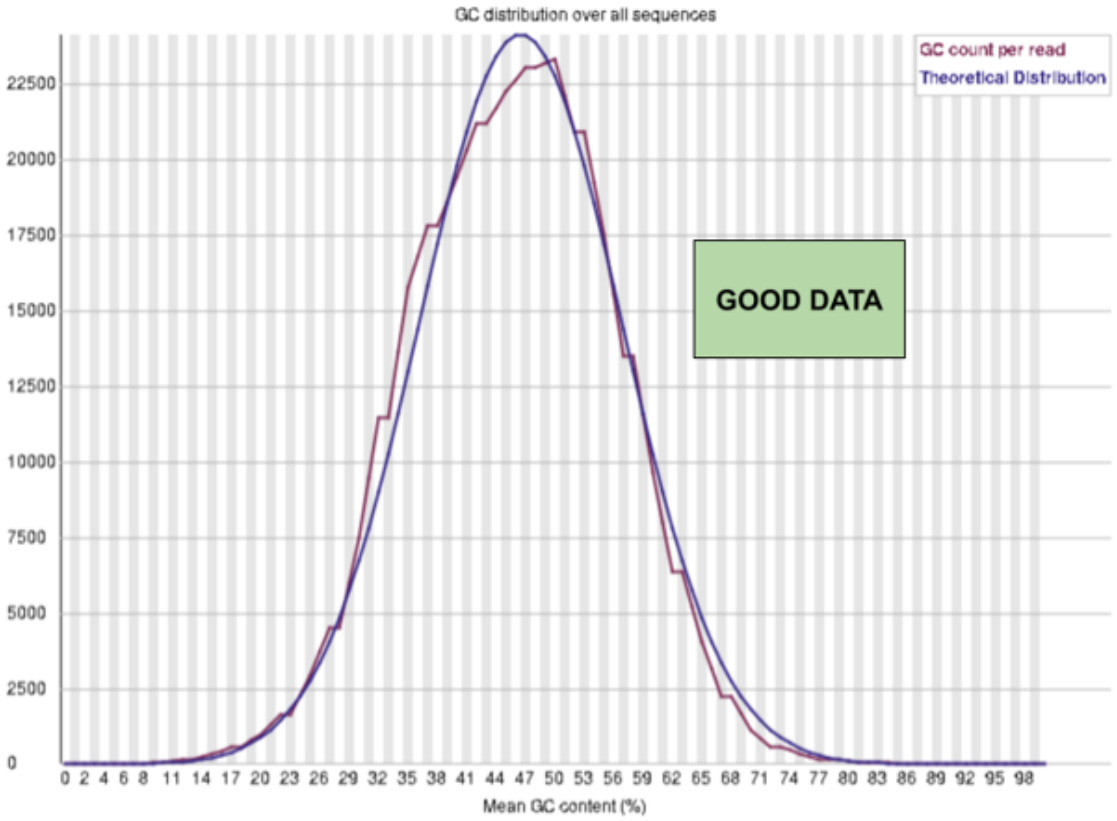


Fig 7: Graphical representation of data with no adapters, indicating they were removed successfully.

8. Per Sequence GC Content

This statistic reveals the distribution of Guanine and Cytosine content across each sequence in a dataset. The x-axis represents the GC content percentage, and the y-axis represents the number or percentage of sequences with that GC content. A reasonably consistent distribution of GC content percentages in a high-quality dataset would suggest that the sequencing procedure was unbiased and devoid of contamination.

A distribution that deviates from balance may be a sign of underlying problems. One noticeable peak, for instance, at a certain GC content percentage can indicate the presence of contaminated sequences or biases in the amplification process. Variations in GC concentration may have biological significance. On the basis of the genomic areas being sequenced, researchers may occasionally anticipate specific GC content patterns.



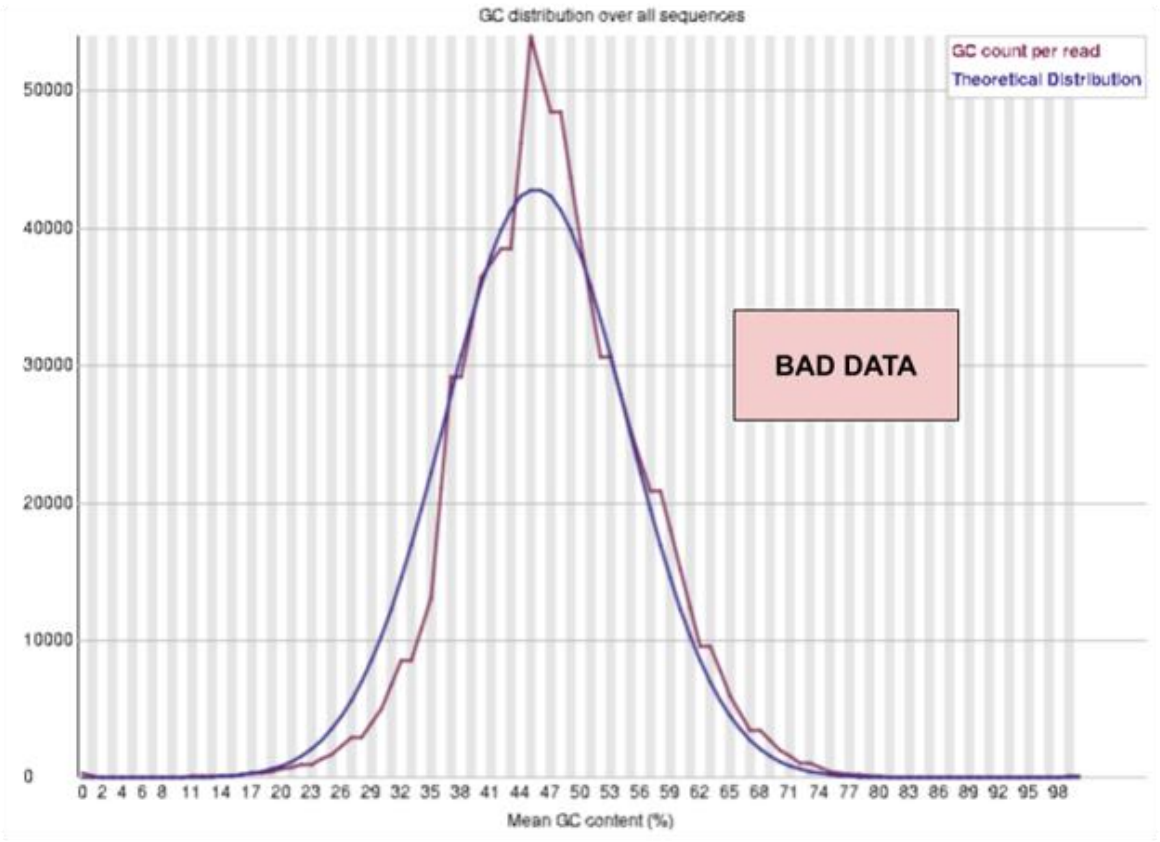


Fig 8: Graphical representation of FASTQC Per Sequence GC content. The upper graph represents a smooth normal distribution while the bottom graph shows library contamination.

9. K-mer content

Kmers are nucleotide sequences (A, T, G, and C) that are short and fixed in length that are found in DNA or RNA sequences. A 4-mer, for instance, is a sequence of four nucleotides. Kmer content analysis is used to look for anomalies or irregularities in sequencing data. Contamination, adapter sequences, or unusual sequence patterns might all be revealed.

If there are no biases or contaminations in the data, you would expect an even distribution of kmers with no one kmer predominating the plot, indicating high-quality sequencing data. Inconsistencies in the kmer plot may point to problems like contamination, overrepresented adapter sequences, or biases that are particular to certain sequence types. For instance, a prominent kmer can be a sign of contamination coming from a certain place.

Certain kmers may occasionally be biologically significant, suggesting the existence of particular motifs or genetic characteristics.

Warning if any k-mer is imbalanced with a binomial p-value <0.01

Failure if any k-mer is imbalanced with a binomial p-value < 10-5

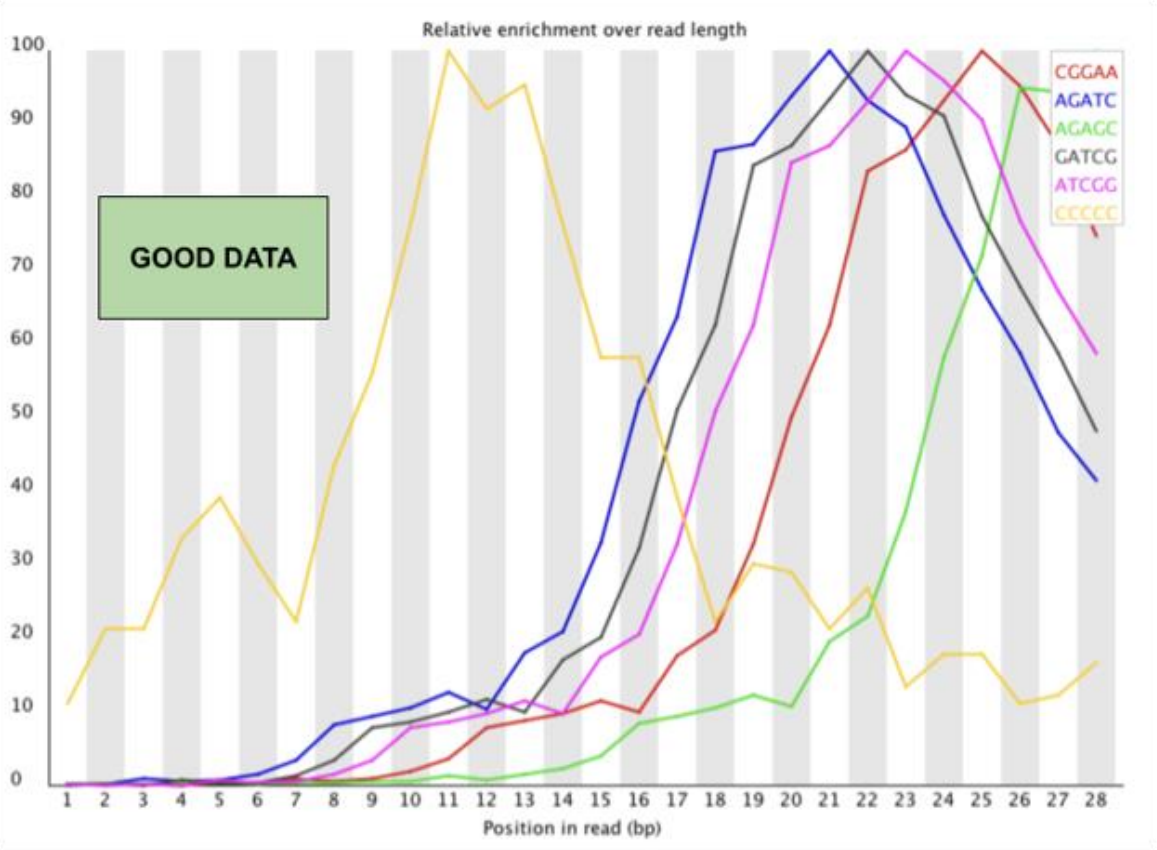


Fig 9: Graphical representation of FASTQC K-mer Content. The graph represents balanced kmer content

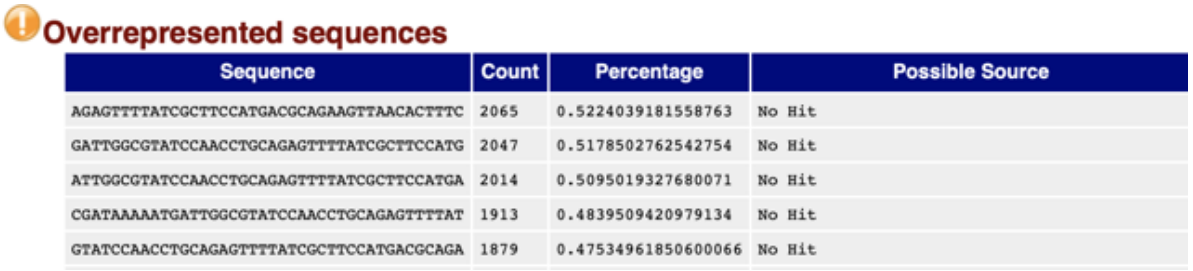
10. Overrepresented Sequences

Short nucleotide sequences that appear more frequently in your sequencing data than would be predicted by chance are known as overrepresented sequences. These include adapter sequences, primers or other widespread contaminants, which should have been eliminated during the sequencing process.

Although many overrepresented sequences are contaminants, some might be important for biology, for example over expression of certain RNA because of a disease phenotype. To distinguish between contaminants and sequences that are relevant to biology, researchers need to study the detected sequences carefully.

Warning if any sequence represent >0.1 of the total

Failure if any sequence represent >1% of the total



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

Fig 10: A list of overrepresented sequences.

Interpreting FastQC results

Although warnings or failures in NGS data analysis should be interpreted with caution, they may not necessarily have a detrimental impact on the overall analysis

In fact, certain warnings might even be anticipated in specific scenarios. For instance, when examining Sequence Duplication Levels, a low number of duplicated sequences is generally expected in genomics studies, but in transcriptomics, a higher number of duplicated sequences is often observed due to the nature of the RNA samples.

Another aspect to consider is over-represented sequences, which is particularly relevant in small RNA libraries where sequences are not randomly fragmented. In such cases, it is natural for the same sequence to be present in a significant proportion of the library, and thus, an over-representation warning should be interpreted in context.

Trimming

In cases where the read sequences exhibit poor quality for any reason, it becomes essential to consider performing sequence trimming. This process involves the removal of reads with low-quality bases to ensure the data's integrity and accuracy. By eliminating bad quality reads, the downstream analysis can be significantly improved, as it focuses on more reliable and informative data. Sequence trimming is a critical step in NGS data processing, as it helps mitigate the impact of sequencing errors and other artifacts, leading to more robust and accurate results in subsequent analyses and research interpretations. Use the link attached to read in more details on how to perform trimming on sequence data.

The figure below show sequences before and after trimming: