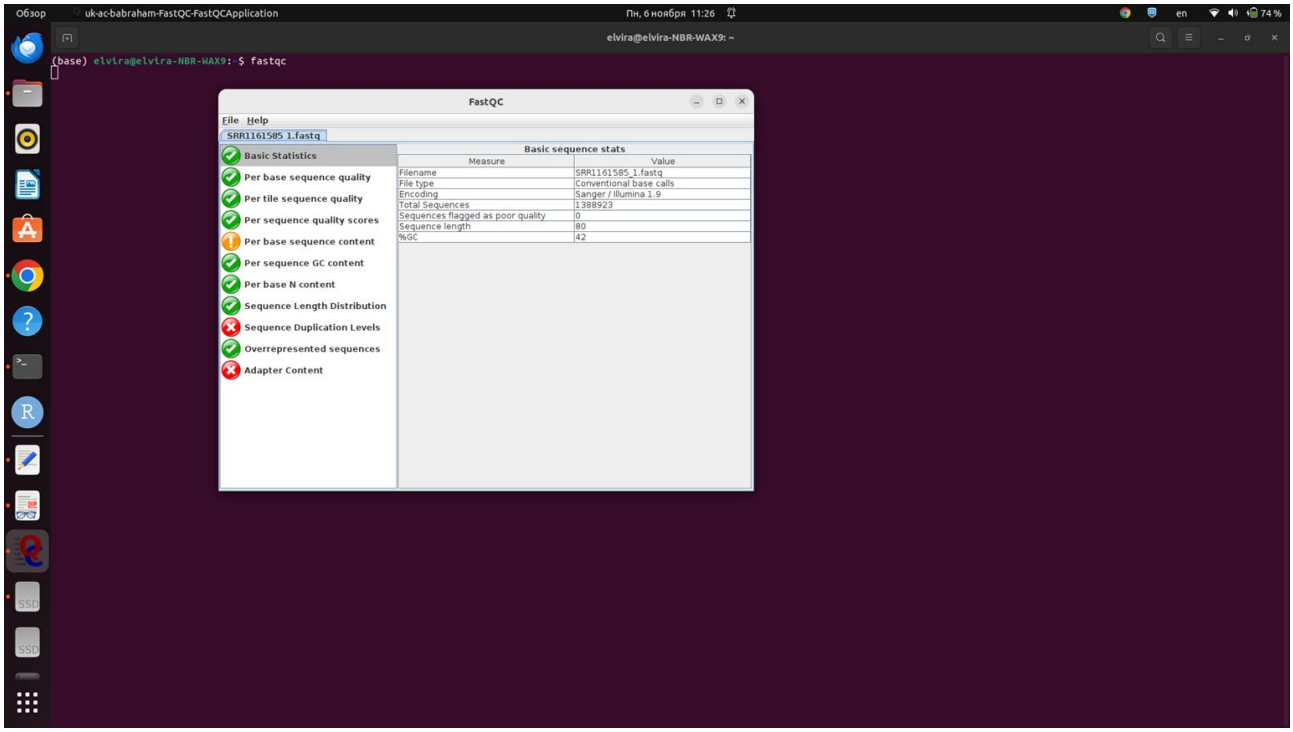
Quality control for NGS data

Step 1. Learn about the experiment and get a file for analysis

Organism: Mammuthus columbi (Columbian mammoth)
Sequencing Platform: ILLUMINA
Library Layout: PAIRED

Step 2. Install the FASTQC program



Step 3. Analyze the file

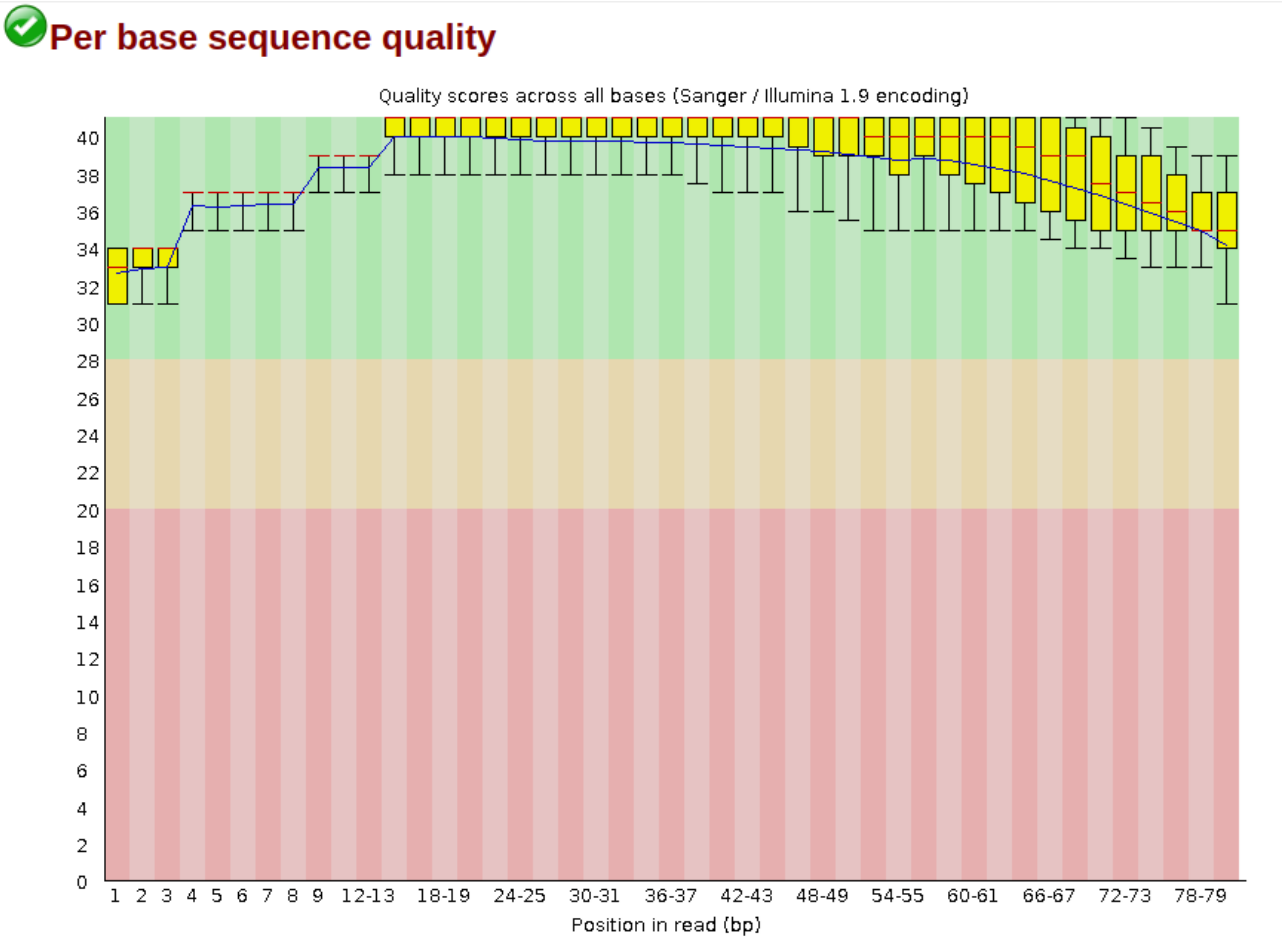
- **General statistics: number of reads and their length**

Basic Statistics

Measure	Value
Filename	SRR1161585_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1388923
Sequences flagged as poor quality	0
Sequence length	80
%GC	42

Total sequences — 1388923
Sequence length — 80

● **Quality of individual nucleotides and average quality of reads**



At the beginning and at the end of the reads, the quality is lower compared to the middle. However, the overall read quality is high throughout the entire length of the reads.

The average quality score of the reads is a very high.

Possible reasons for the decrease in quality at the beginning of the reads:

1. Illumina Sequencing: In Illumina sequencing technology, sequencing signals can be less accurate during the initial cycles, which may reduce the quality of nucleotides at the beginning of the read.
2. Adapters and Other Structural Elements: The presence of adapters and other structural elements at the beginning of the read can cause issues with signal quality.
3. Clustering Cycles: There may be lower clustering cycles at the beginning of sequencing, which can impact the quality.

Possible reasons for the decrease in quality at the end of the reads:

Illumina is known to have some systematic errors.

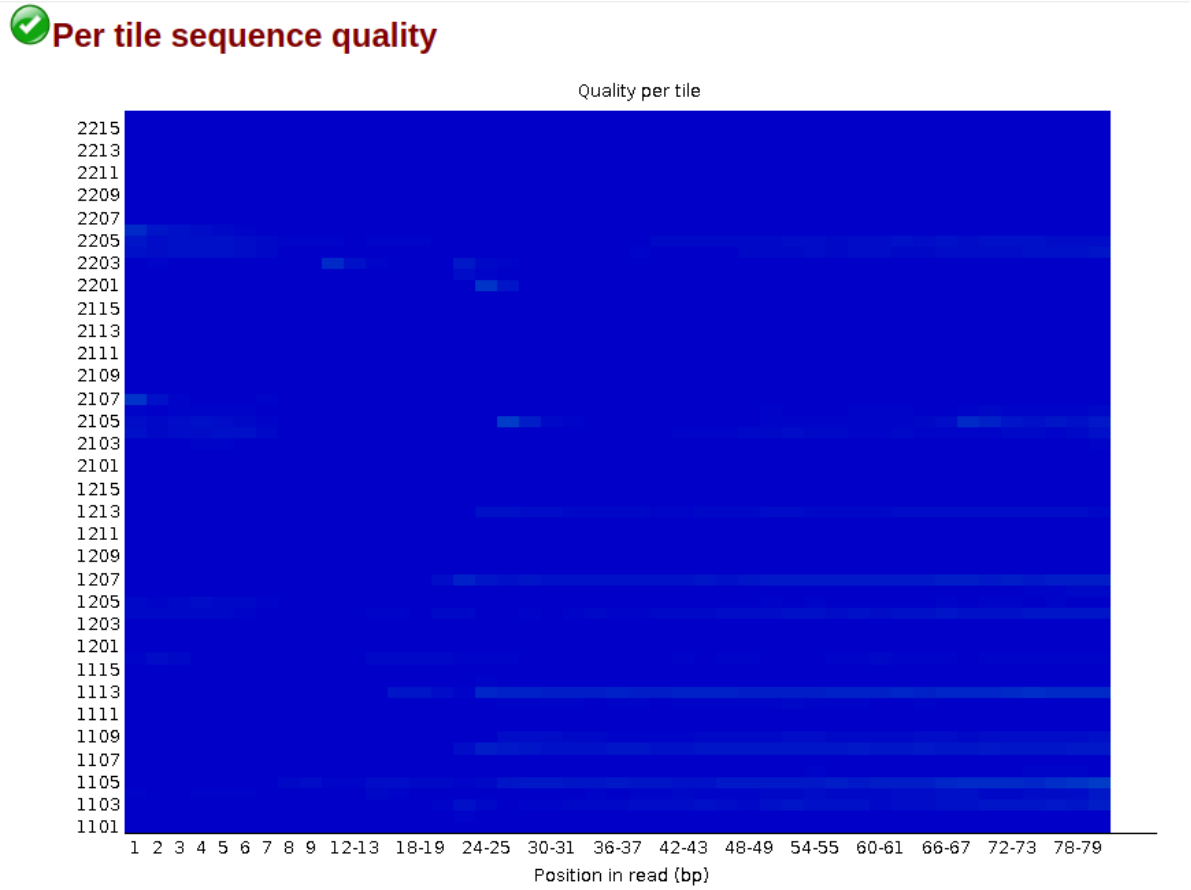
1. The primary reason is cluster desynchronization (phasing/pre-phasing). Ideally, all molecules in a cluster should be identical, but individual molecules may lag behind or advance ahead of the majority. This leads to reduced cluster synchrony and a decrease in

the signal-to-noise ratio. The longer the read length, the more delayed and advanced molecules there will be in the cluster.

2. The second possible reason is the defocusing of the sequencer's optical system and cluster burnout. Lower cluster intensity results in lower read quality.

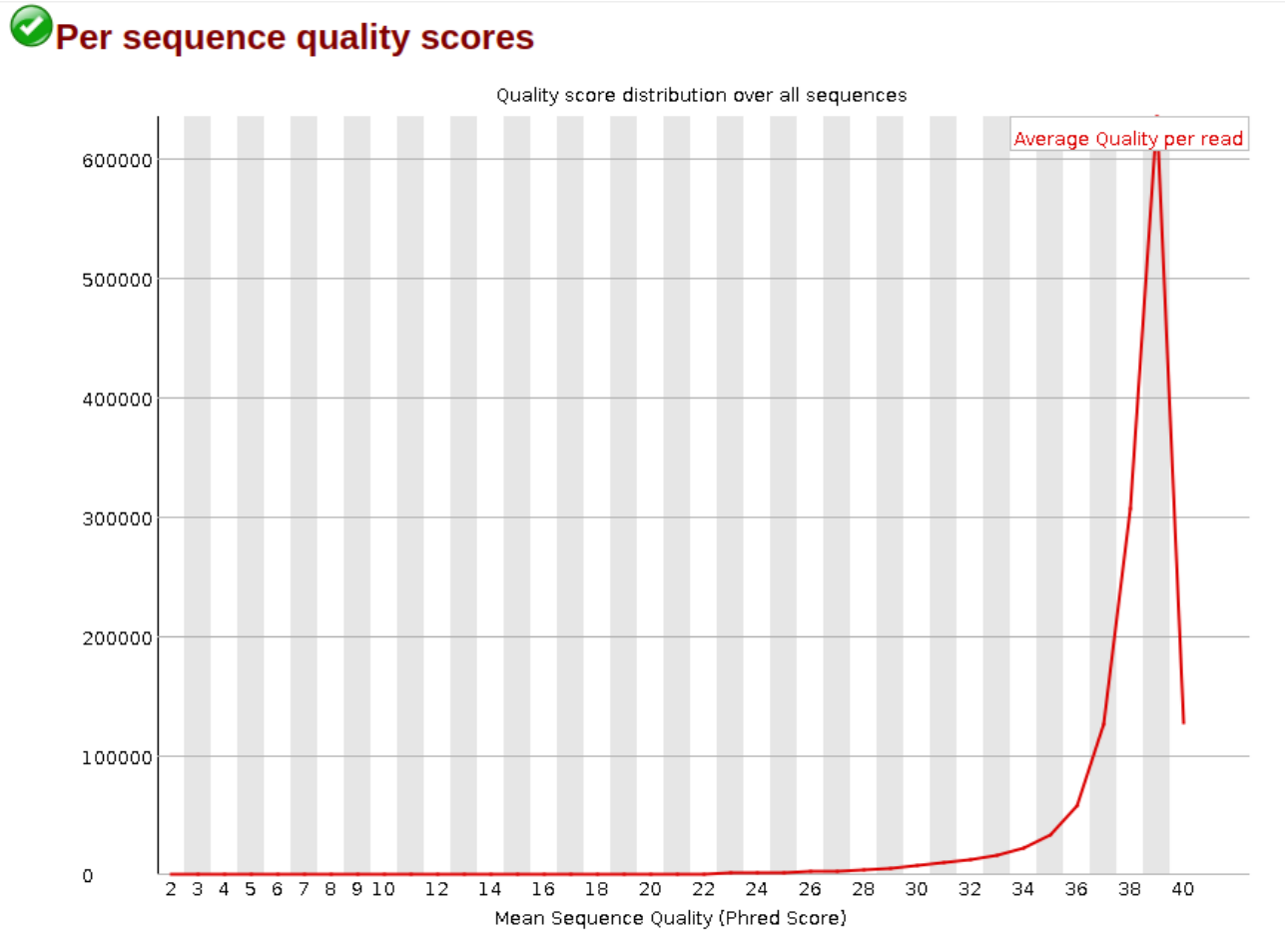
3. The third possible reason is a higher number of errors associated with GGC and/or inverted repeats.

● **Per Tile Sequence Quality**



The sequence quality is high on each tile.

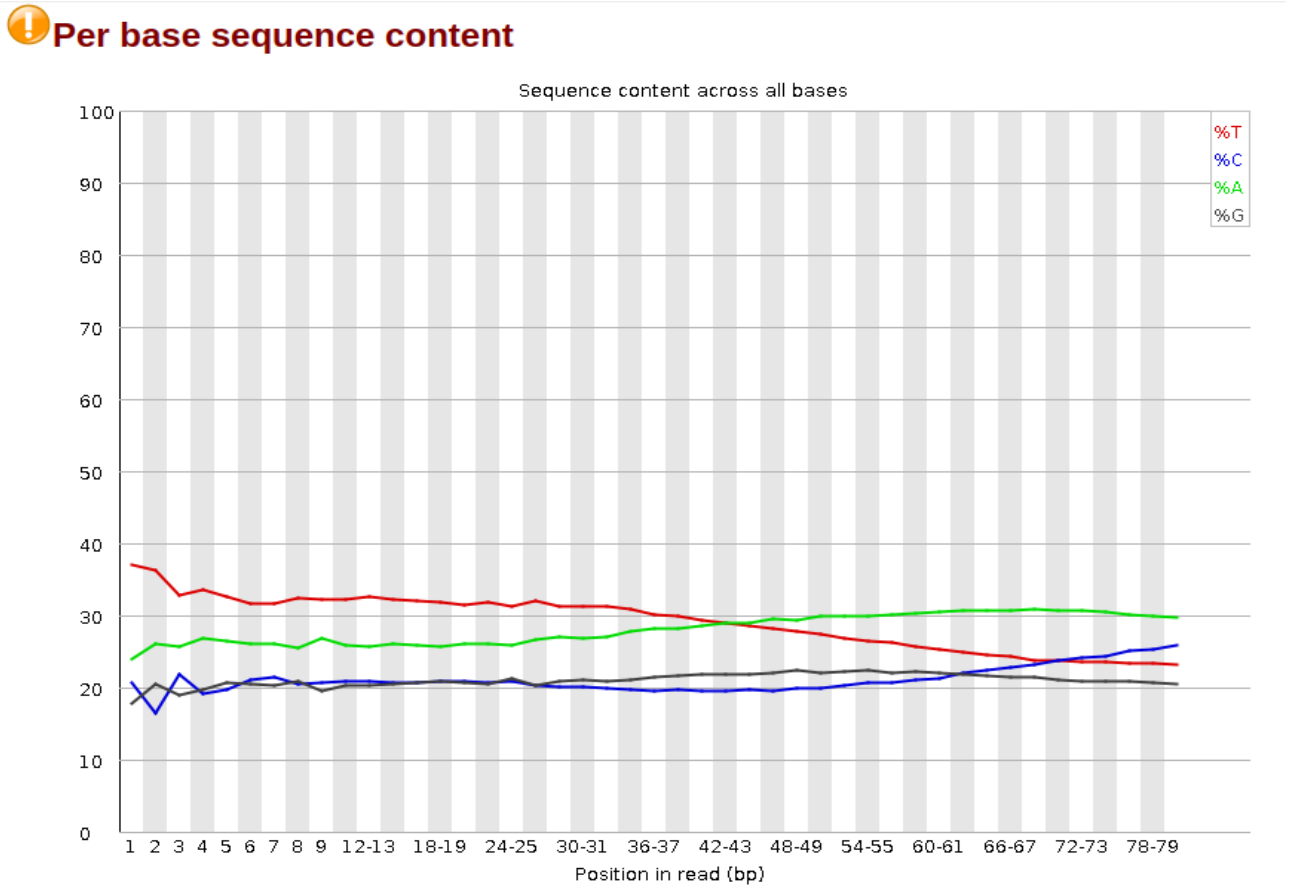
● **Per Sequence Quality Scores**



All the sequences in the dataset exhibit very high quality. The maximum quality value is 39, which corresponds to very high quality of sequenced nucleotides.

A single peak with the maximum quality value indicates uniform and high data quality. This suggests that the majority of sequences in the sample have reliable quality and can be used for analysis.

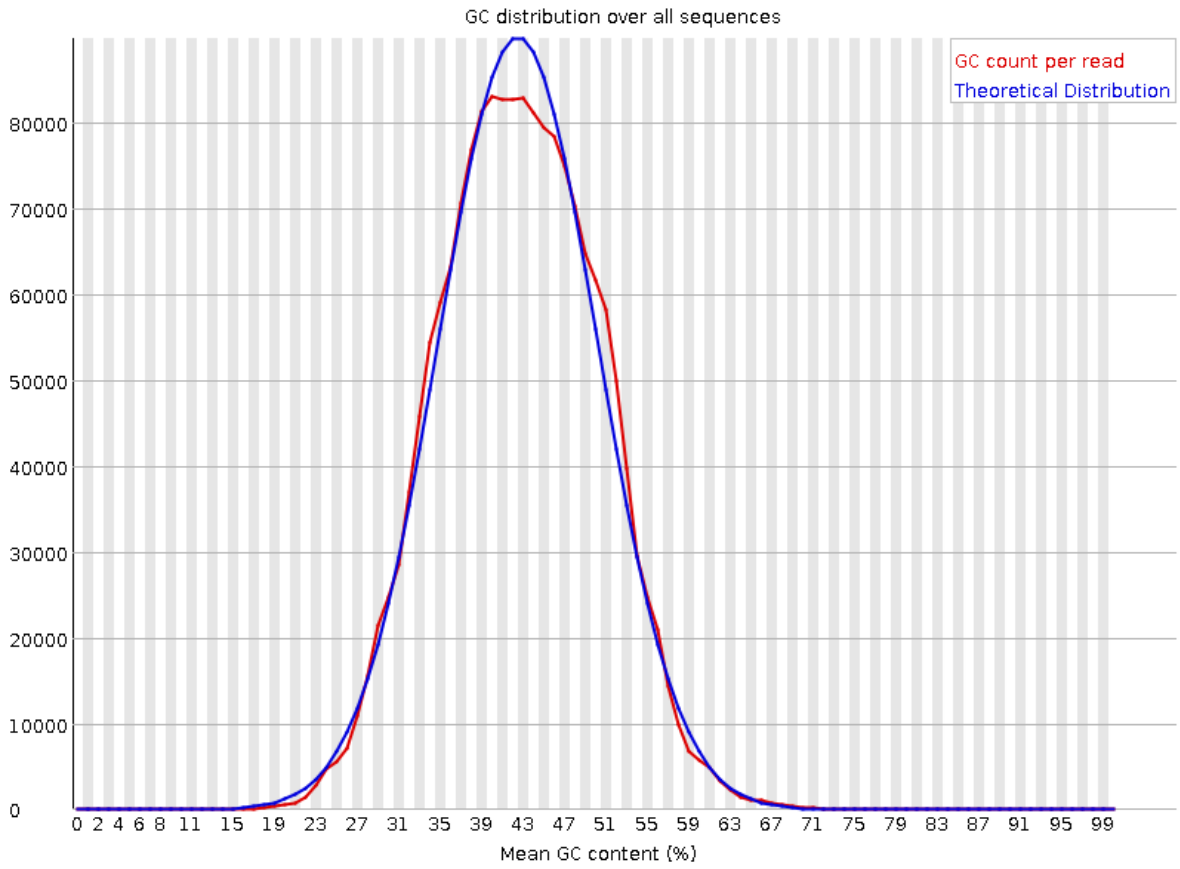
● **Per Base Sequence Content**



Deviation is detected.
At the beginning of the reads, there is an elevated content of T (~38%), but towards the end of the reads, the T content decreases (~22%).
The A content increases towards the end of the reads (from ~23% to 30%).
There is also a noticeable increase in the C content towards the end of the reads (from ~20% to 27%).
The G content remains relatively stable throughout the reads (~20%).

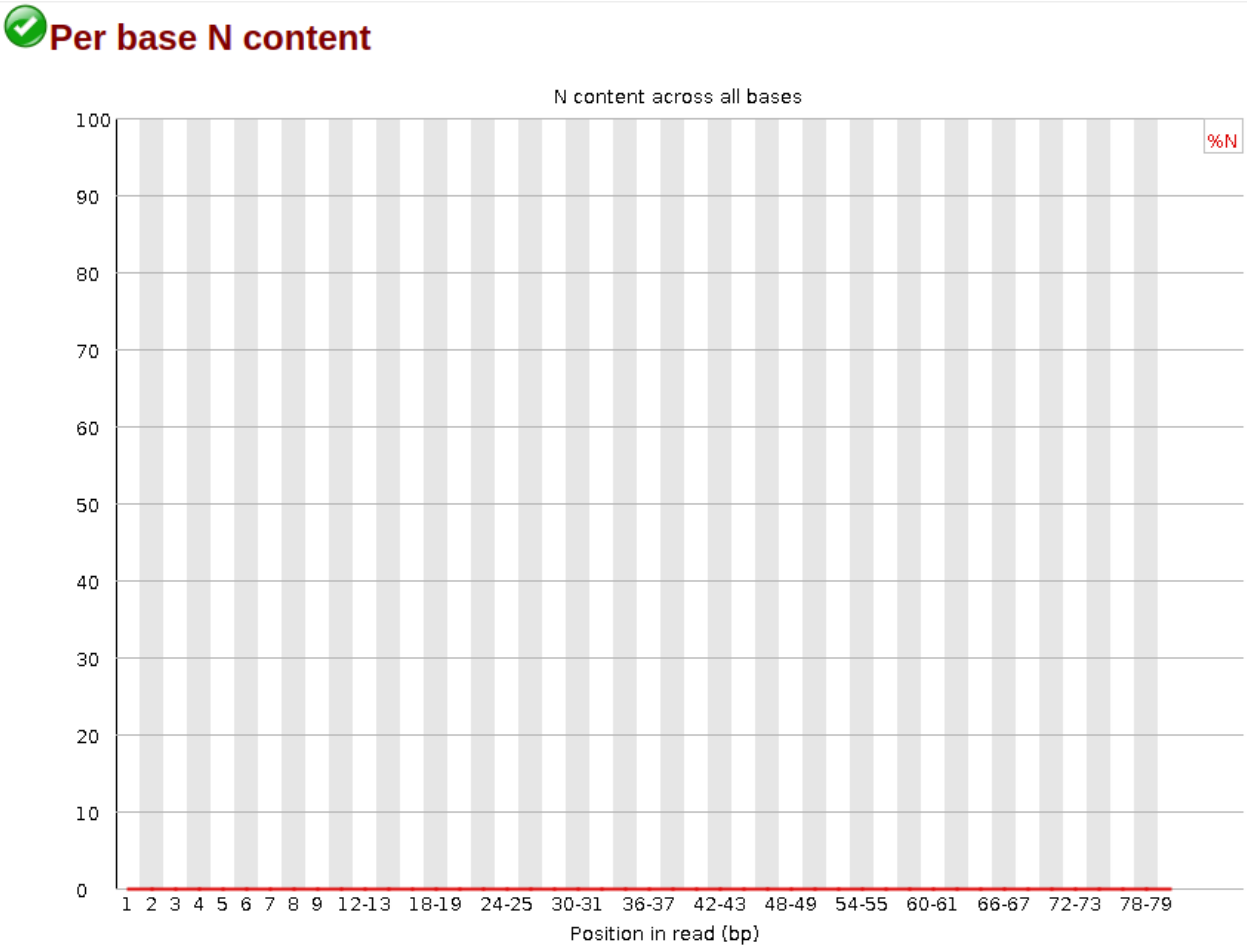
● **Per Sequence GC Content, what distribution do you see?**

Per sequence GC content



The GC content in the sequences matches the theoretical distribution. The central peak corresponds to the overall GC content. The distribution is normal.

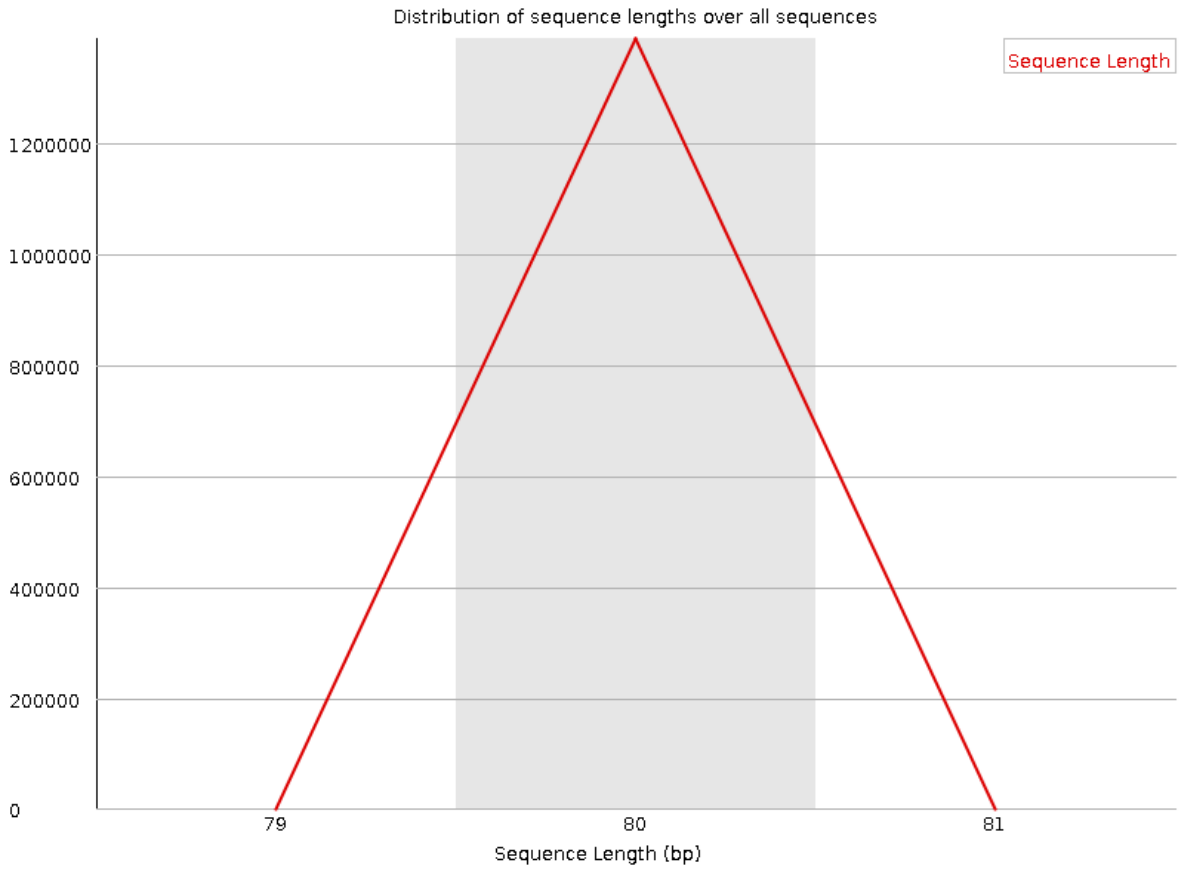
● **Per Base N Content**



The nucleotide content of N in each position is 0%. Thus, all bases were identified by the sequencer.

- **Sequence Length Distribution, are there sequences that differ in length?**

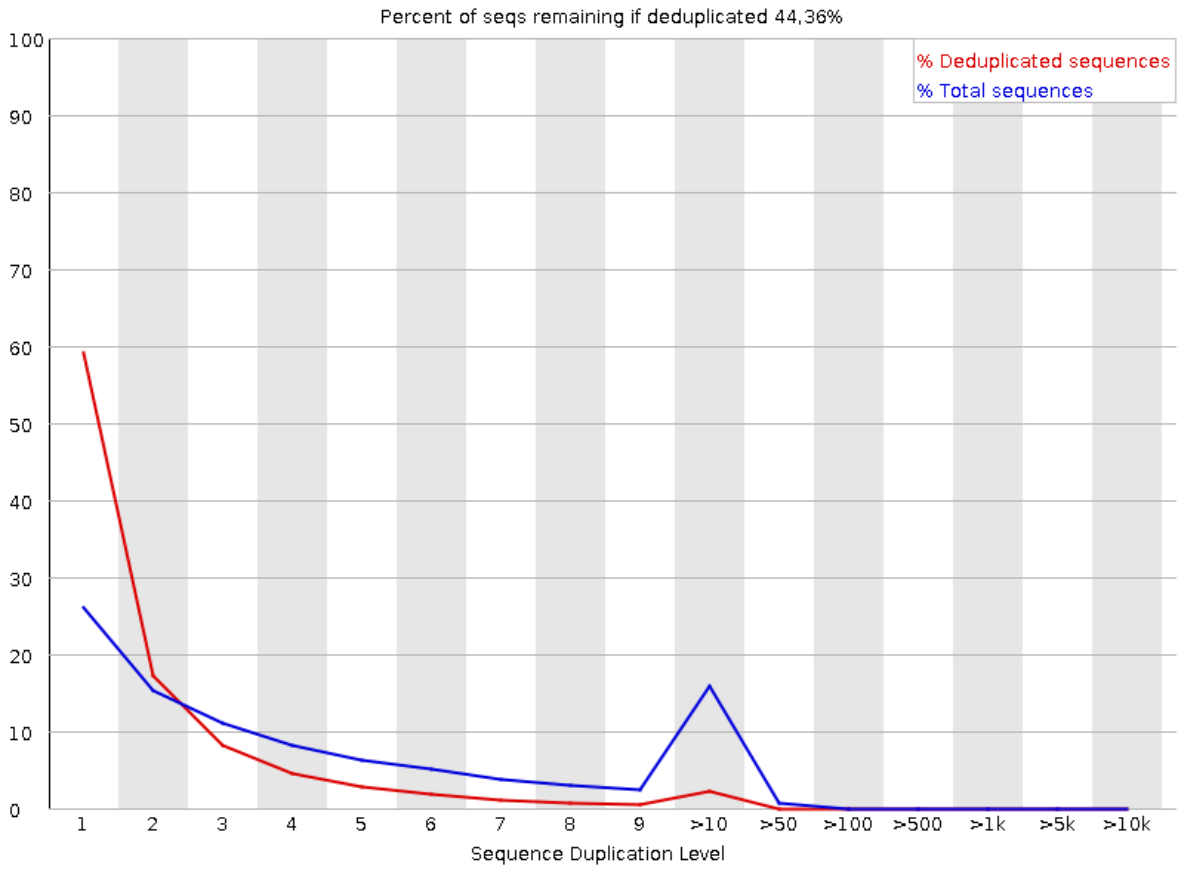
Sequence Length Distribution



There is a clear single peak at 80 bp, and there are no differences in read lengths.

● **Duplicate Sequences, low or high duplication?**

Sequence Duplication Levels

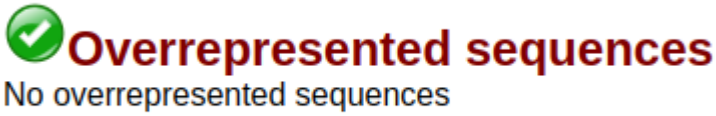


There is a high level of duplication. This likely indicates some systematic enrichment error, such as PCR amplification.

Percent of sequences remaining if deduplicated 44.36%.

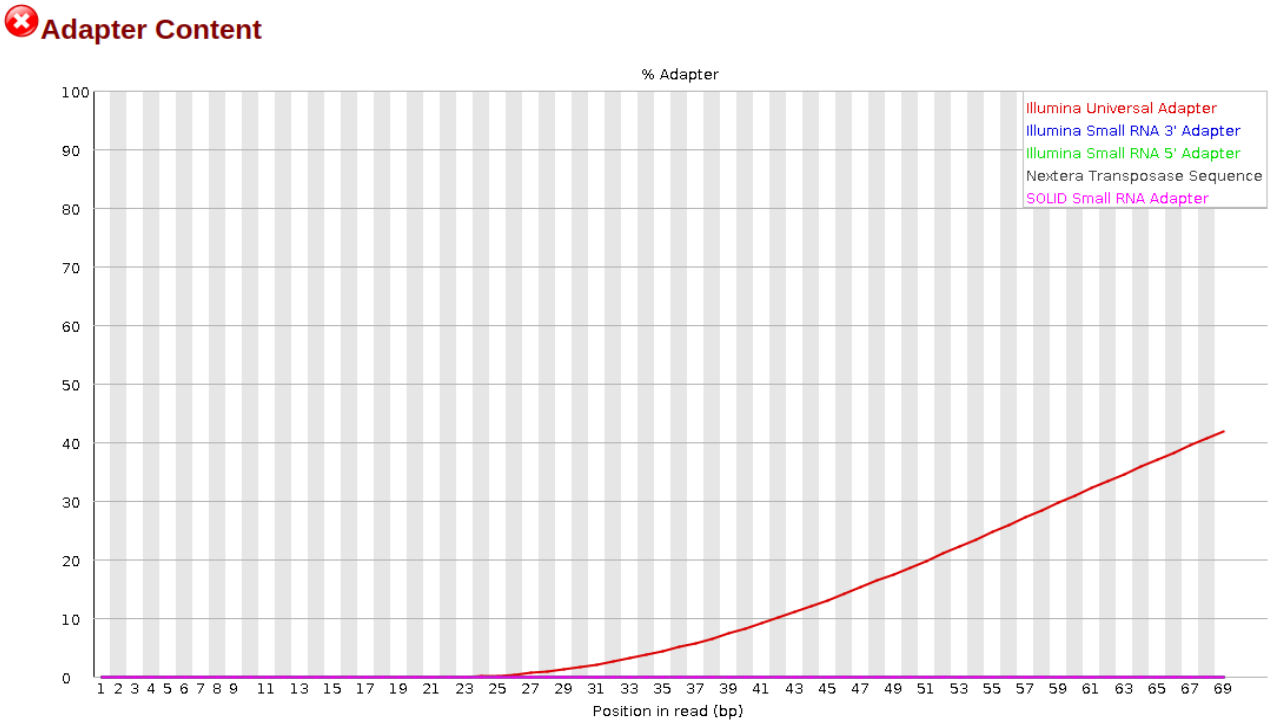
Some software (e.g., SAMtools) can be used to detect and remove duplicates from sequenced data. It will help to reduce the impact of PCR amplification on the analysis results.

● **Overrepresented Sequences, do they exist? if yes, is anything known about them?**



There are no overrepresented sequences.

● **Availability of Adapter Content, if it will**



The reads contain the Illumina universal adapter.
Trimming tools such as Trimmomatic or Cutadapt can be used to remove the adapters at the end of the reads.

Summary

Quality control for NGS data of Mammuthus columbi (Columbian mammoth) was performed with FastQC program.
Some problems were detected:

1. High level of duplication.

This likely indicates some systematic enrichment error, such as PCR amplification.
Percent of sequences remaining if deduplicated 44.36%.
Some software (e.g., SAMtools) can be used to detect and remove duplicates from sequenced data. It will help to reduce the impact of PCR amplification on the analysis results.

2. The reads contain the Illumina universal adapter.

Trimming tools such as Trimmomatic or Cutadapt can be used to remove the adapters at the end of the reads.