MHC Genomics Analysis Project - Task 2 Report

Mapping Quality (MAPQ) Analysis

Mahla Entezari*
Shahid Beheshti University, Tehran, Iran
Bioinformatics – Fall 2025

January 6, 2026

I. Introduction

In this task, I checked how reliable the read alignments are in each BAM file before doing any MHC-specific analysis. This step is important because the MHC region is difficult to map: many genes in the region are similar to each other, and the region contains repeated/homologous sequences. When a read matches multiple places in the genome, the aligner may not be able to confidently choose one location. That situation produces low mapping quality scores, which can affect coverage calculations and later allele typing.

The main value I used is **MAPQ (mapping quality)**. MAPQ is assigned by the aligner and shows how confident it is that a read is placed correctly. In most datasets, **higher MAPQ = more confident mapping**. A common practical threshold is **MAPQ** ≥ 30 to represent "confidently mapped" reads, and **MAPQ = 0** to represent reads that are ambiguous (often multi-mapped).

II. Materials and Methods

A. Input Data

The input files were BAM alignment files for multiple samples. Each BAM contains read alignments and a MAPQ value for each read.

B. MAPQ Extraction

For each BAM file, I extracted the MAPQ values for all reads using `samtools` (MAPQ is the 5th column when viewing BAM/SAM alignments). This produced one list of MAPQ values per sample. From that list I generated plots and computed summary metrics.

C. What I Computed and Why

To get a complete understanding, I used several plot types. Each plot gives a different view of the same data:

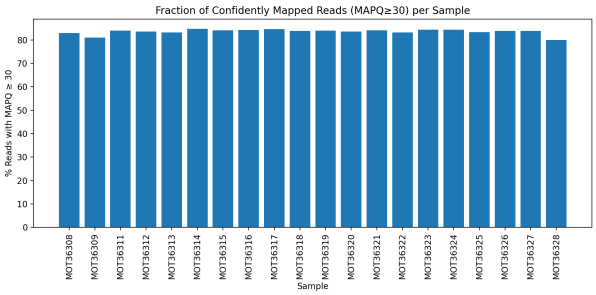
- **Histogram (counts):** shows the raw number of reads at each MAPQ value.
- **Histogram (normalized):** same histogram but converted into probabilities, so samples with different read counts can be compared fairly.
- **Density plot (KDE):** a smooth curve that summarizes the overall shape of the distribution (useful for clearly seeing modes/peaks).
- **CDF plot:** shows what fraction of reads have MAPQ $\leq x$. This is useful for quickly answering questions like "what fraction of reads are below a quality threshold?"
- **Across-sample bar plots:** show (1) the fraction of reads with MAPQ ≥ 30 and (2) the fraction with MAPQ=0 for each sample, which makes it easy to spot outliers.

III. Results

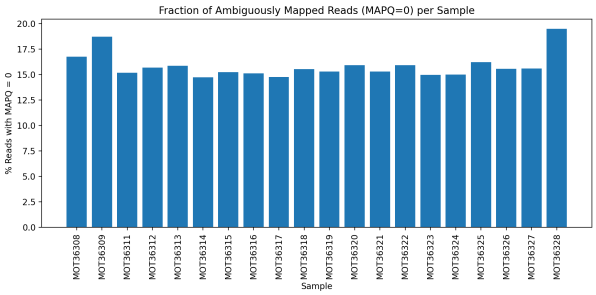
A. QC Results Across Samples

Figure I shows the overall mapping quality across all samples. Most samples have a high fraction of confidently mapped reads (MAPQ ≥ 30), typically around the low-to-mid 80% range. This indicates good overall alignment

*MahlaEntezari.sbu@gmail.com



(A) Percentage of reads with MAPQ $\geq$ 30 per sample.

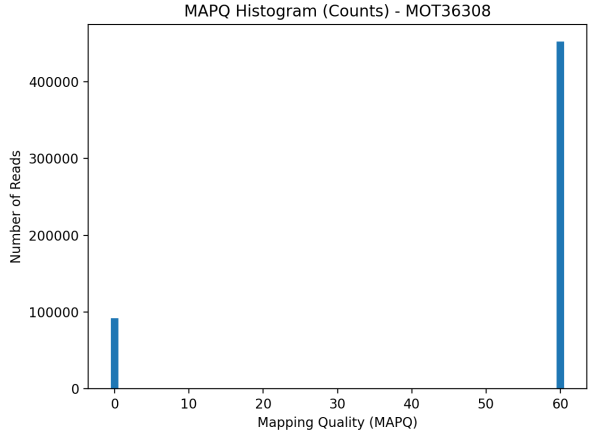


(B) Percentage of reads with MAPQ = 0 per sample.

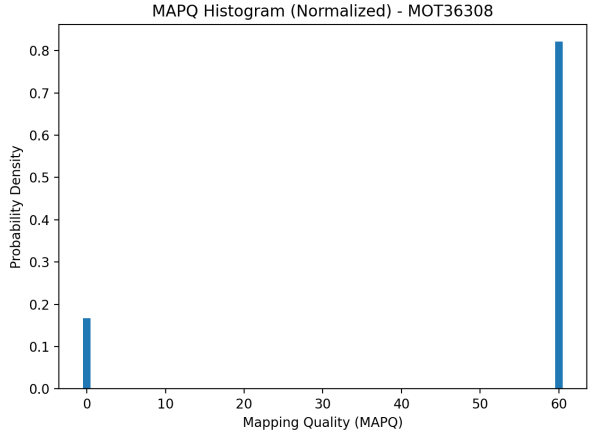
Figure I: Across-sample summary of mapping quality metrics.

quality. However, some samples show a higher fraction of MAPQ = 0 reads, suggesting increased alignment ambiguity.

B. MAPQ Distribution Shape (Representative Sample)



(A) MAPQ histogram (counts) for sample MOT36308.



(B) MAPQ histogram (normalized) for sample MOT36308.

Figure II: Histogram-based MAPQ distributions for sample MOT36308.

Figure II illustrates the distribution of MAPQ values for sample MOT36308. The raw histogram shows a strong peak near the maximum MAPQ value, indicating that most reads are confidently mapped. A smaller peak at MAPQ = 0 represents ambiguous reads. The normalized histogram confirms that most reads remain high-quality when scaled to probabilities.

C. Smooth and Cumulative Views of MAPQ

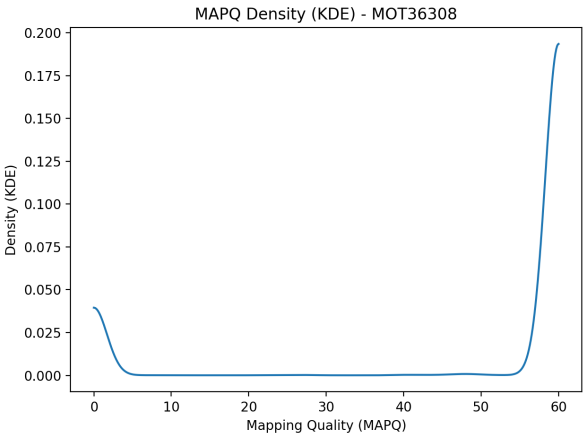
Figure III provides two complementary views of mapping quality. The KDE plot highlights two groups of reads: a small low-MAPQ group and a larger high-MAPQ group. The CDF plot shows that most reads accumulate at high MAPQ values, confirming strong alignment confidence.

IV. Discussion

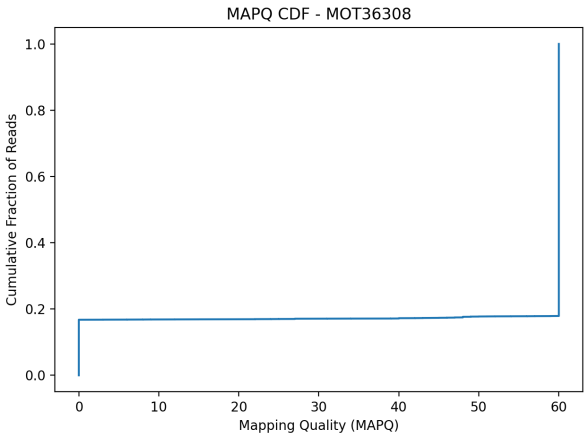
The mapping quality patterns observed in this analysis reflect the known biological properties of the MHC region. MHC genes are highly polymorphic and share strong sequence similarity with one another. As a result, some sequencing reads cannot be uniquely aligned to a single genomic location, which explains the presence of reads with low or zero mapping quality (MAPQ = 0).

Despite this expected ambiguity, most reads across the samples have high MAPQ values. This indicates that, although the MHC region is difficult to analyze, the majority of sequencing reads can still be placed confidently. Therefore, the alignment data are generally reliable for downstream biological analysis.

Similar challenges have been reported in previous studies that analyze the MHC region using short-read sequencing technologies. These studies often describe bimodal mapping quality distributions, where reads are



(A) MAPQ DENSITY (KDE) PLOT FOR SAMPLE MOT36308.



(B) MAPQ CUMULATIVE DISTRIBUTION FUNCTION (CDF) FOR SAMPLE MOT36308.

FIGURE III: SMOOTH AND CUMULATIVE REPRESENTATIONS OF MAPQ VALUES.

either confidently mapped or clearly ambiguous. The patterns observed in this analysis are consistent with those findings, suggesting that the results follow well-known characteristics of MHC sequencing data rather than being caused by technical errors.

V. LIMITATIONS

This analysis focuses only on mapping quality and does not directly evaluate whether reads are aligned to the correct allele or gene at the nucleotide level. In addition, MAPQ values depend on the alignment tool and its scoring system, meaning that MAPQ should be interpreted as a relative measure of confidence rather than an absolute indicator of correctness. Other factors, such as read length and sequencing depth, may also influence alignment quality but were not examined in this task.

VI. KEY QUESTIONS

Are the reads generally high quality? Yes. For most samples, a large fraction of reads have high mapping quality scores (MAPQ ≥ 30). This indicates that the majority of reads are confidently and uniquely aligned to the reference genome. Overall, the sequencing data appear to be of good quality and suitable for further analysis.

Are there samples with poor mapping quality? Yes. Some samples show a higher proportion of reads with MAPQ $= 0$ and a lower fraction of high-MAPQ reads compared to others. These samples contain more ambiguous alignments, which suggests increased difficulty in mapping reads uniquely. While these samples are not unusable, their results should be interpreted with greater caution in downstream analyses.

How might this affect downstream MHC analysis? Low mapping quality can affect downstream MHC analysis by reducing the accuracy of gene coverage estimates and lowering confidence in allele typing. Ambiguous reads may align to multiple similar MHC genes or alleles, which can bias comparisons and lead to incorrect biological conclusions. Therefore, samples with poorer mapping quality may require stricter MAPQ filtering or more conservative interpretation during MHC-specific analyses.

VII. CONCLUSION

In this task, mapping quality was evaluated for all BAM files as a quality control step before biological interpretation. The results show that most samples contain predominantly high-quality, confidently mapped reads, supporting their use in downstream MHC gene coverage and allele analysis. At the same time, a subset of samples exhibits increased alignment ambiguity, highlighting the importance of careful quality assessment when working with complex genomic regions such as the MHC. Overall, this task provides a strong foundation for interpreting downstream analyses within the context of alignment confidence.