MHC Genomics Analysis Project - Report

Analysis and Discussion

Mahla Entezari*

Bioinformatics Assignment 1

January 6, 2026

1 Part 1: Quality Assessment (QC)

1.1 Objective

The goal of this step was to determine whether the sequencing data and read alignments are reliable enough for downstream biological interpretation.

1.2 Methods

Mapping quality (MAPQ) values were extracted from each BAM file using `samtools`. MAPQ represents the aligner's confidence in the correctness of each read's genomic placement. High MAPQ values indicate unique and confident alignments, while MAPQ = 0 typically indicates ambiguous or multi-mapped reads.

Several plots were generated for each sample, including histograms, kernel density estimates (KDE), and cumulative distribution functions (CDF). In addition, across-sample bar plots summarized the fraction of reads with MAPQ ≥ 30 and MAPQ = 0.

1.3 Results

Across all samples, the majority of reads showed high mapping quality, with most samples having approximately 80–85% of reads with MAPQ ≥ 30. This indicates generally reliable alignments. However, a subset of samples exhibited a higher proportion of MAPQ = 0 reads, suggesting increased alignment ambiguity.

Representative plots for sample MOT36308 showed a bimodal MAPQ distribution: a dominant peak at high MAPQ values and a smaller peak at MAPQ = 0. This pattern is consistent with known properties of the MHC region, where homologous sequences can cause ambiguous read placement.

1.4 Interpretation

Overall, the sequencing and alignment quality is sufficient for downstream analysis. Samples with higher fractions of low-MAPQ reads should be interpreted with greater caution, as ambiguous mappings may affect coverage estimates and allele typing accuracy. These observations are consistent with the detailed QC analysis reported in Task 2 [2].

2 Part 2: Gene Annotation Analysis

2.1 Objective

The goal of this step was to identify which genes are covered by the sequencing data and to compare gene-level coverage patterns across samples.

2.2 Methods

Aligned reads were annotated using a GTF gene annotation file matched to the hg19 reference genome. Gene-level read counts were computed using `featureCounts` in paired-end mode. A gene was considered covered if it had at least one overlapping fragment (count > 0).

For each sample, a gene annotation table was generated containing gene name, genomic coordinates, strand, and reference sequence. The top 10 most covered genes per sample were identified based on read counts.

2.3 Results

Each sample contained several hundred genes with nonzero coverage (for example, 734 genes in MOT36308). Many highly covered genes were shared across samples, indicating broadly consistent sequencing performance. At the same time, some sample-specific differences were observed, reflecting biological variability or differences

*mahlaentezari.sbu@gmail.com

in alignment behavior.

Several MHC-related genes were present among the covered genes, confirming that the MHC region is represented in the sequencing data.

2.4 Interpretation

The gene annotation results demonstrate that the sequencing data provide sufficient coverage across a large number of genes, including immune-related loci. Differences in coverage between samples may reflect both biological variation and technical factors such as mapping ambiguity.

3 Part 3: MHC-Specific Coverage Analysis

3.1 Objective

This step focused specifically on immune-related genes within the MHC region, including MHC Class I, Class II, and MIC genes.

3.2 Methods

For each MHC gene, coverage depth was calculated by intersecting BAM alignments with gene coordinates and computing per-base read depth. Gene-level summary statistics such as mean depth and coverage fractions were derived. Coverage depth plots (bar plots and heatmaps) were used to compare coverage patterns across genes.

3.3 Results

Coverage depth varied substantially across MHC genes. Among MHC Class I genes, HLA-A generally showed higher and more uniform coverage compared to HLA-B and HLA-C. Several coverage gaps were observed, particularly in regions known to be difficult to map. Conversely, localized coverage hotspots were detected, which may reflect repetitive elements or alignment biases.

3.4 Interpretation

Uneven coverage across the MHC region is expected due to its complex genomic structure. Systematic differences between HLA-A, HLA-B, and HLA-C highlight the importance of gene-specific analysis when interpreting immune-related sequencing data.

4 Part 4: Allele Typing (HLA Analysis)

4.1 Objective

The goal of this step was to identify candidate HLA alleles for each sample using sequence alignment.

4.2 Methods

Extracted MHC gene sequences were aligned to reference HLA allele sequences obtained from the IMGT/HLA database. For each gene, the best-matching allele was identified based on global alignment score and percent sequence identity. Confidence in allele calls was assessed using these metrics.

4.3 Results

In the analyzed sample, a high-confidence allele call was obtained for HLA-A (HLA-A*01:01:01:01), showing high alignment score and percent identity. Other HLA genes did not yield confident allele assignments and were reported as NA. This suggests either insufficient discriminatory sequence information or limitations in alignment due to sequence similarity.

Across samples, the limited number of confident allele calls indicates that only a subset of loci can be reliably typed using short-read data.

4.4 Interpretation

The allele typing results demonstrate that confident HLA allele identification is possible for some loci but remains challenging overall. Genes with clear and high-identity matches can be typed with reasonable confidence, while others require more cautious interpretation.

5 Part 5: Biological Interpretation

5.1 MHC Diversity and Sample Differences

The observed variation in coverage and allele typing success reflects the high diversity of the MHC region. Differences between samples may indicate underlying genetic variation, but technical factors such as mapping

ambiguity also play a major role.

5.2 Implications for Immune Function and Disease

Variation in HLA alleles can influence antigen presentation and immune responsiveness, potentially affecting disease susceptibility. However, given the limited resolution of short-read sequencing in the MHC region, these results should be interpreted as preliminary.

References

[1] Task 1 Report: Gene-Level Annotation from BAM using GTF and Reference Genome.

[2] Task 2 Report: Mapping Quality (MAPQ) Analysis.