MHC Genomics Analysis Project

---

1. Project Overview and Biological Background

What is the MHC and why do we study it?

The **Major Histocompatibility Complex (MHC)** is a genomic region that contains some of the most important genes in the human immune system. These genes encode proteins that present antigens to immune cells and play a central role in:

- Immune recognition
- Response to infections
- Autoimmune diseases
- Transplant compatibility
- Cancer immunology

The MHC region (on **chromosome 6**) is:

- Highly **polymorphic**
- Gene-dense
- Technically challenging to analyze using short-read sequencing

Because of this, it is an excellent real-world example for learning **advanced genomics and bioinformatics analysis**.

---

2. Project Goal

The goal of this project is to **analyze next-generation sequencing (NGS) data** from the MHC region in multiple samples and:

- Characterize **genetic diversity** in immune-related genes

- Understand how sequencing coverage varies across MHC genes
- Identify candidate **HLA alleles**
- Practice building and interpreting a complete bioinformatics pipeline

By the end of this project, you should be able to go from **raw alignment files (BAM)** to **biologically meaningful conclusions**.

---

3. What You Are Expected to Learn

Technical Skills You Will Develop

By completing this project, you will learn how to:

- Work with **BAM files** using command-line tools
- Perform **quality control (QC)** on sequencing data
- Annotate genomic reads with **gene information**
- Focus analyses on specific gene families:
    - MHC Class I (e.g., HLA-A, HLA-B, HLA-C)
    - MHC Class II
    - MIC genes
- Perform **sequence alignment** against reference databases
- Create **clear visualizations** to summarize genomic data

These are core skills used in real genomics research labs.

---

4. Data and Inputs

You will be provided with:

- BAM files for **80 samples**

- Gene annotation files (e.g., GTF/GFF)

You are expected to work **programmatically and reproducibly**, not manually.

---

5. Analysis Tasks and Deliverables

Part 1: Quality Assessment (QC)

Goal: Determine whether the sequencing and alignments are reliable before biological interpretation.

What you must do:

- Assess **mapping quality** across all samples
- Extract basic read statistics:
    - Total reads
    - Mapped reads
    - Properly paired reads
- Identify:
    - Low-quality samples
    - Unusual patterns (e.g., poor mapping in MHC)

Deliverables:

- Mapping quality distribution plots
- A summary table of read statistics per sample
- A short written interpretation:
    - Are there problematic samples?
    - What might cause these issues?

Part 2: Gene Annotation Analysis

Goal: Understand which genes are actually covered by the sequencing data.

What you must do:

- Annotate aligned reads with gene information
- Compute **gene-level coverage** for each sample
- Identify the **top 10 most covered genes** per sample
- Compare gene coverage across samples

Key questions to answer:

- Are the same genes highly covered in all samples?
- Are there sample-specific differences?
- Are MHC genes among the most covered?

Deliverables:

- Gene annotation tables
- Top-10 gene lists per sample
- Comparative plots or heatmaps across samples

Part 3: MHC-Specific Coverage Analysis

Goal: Focus specifically on immune-related genes.

Gene groups to analyze:

- **MHC Class I** (e.g., HLA-A, HLA-B, HLA-C)
- **MHC Class II**
- **MIC genes**

What you must do:

- Calculate **coverage depth** for each MHC gene
- Compare:
    - HLA-A vs HLA-B vs HLA-C
- Identify:
    - Coverage gaps
    - Coverage hotspots
    - Systematic biases

Deliverables:

- Coverage depth plots (bar plots, heatmaps)
- Tables summarizing coverage per gene
- Written interpretation of observed patterns

Part 4: Allele Typing (HLA Analysis)

Goal: Identify candidate HLA alleles from sequencing data.

What you must do:

- Align reads to HLA reference sequences
- Identify candidate alleles for each sample
- Assign **confidence scores** to allele calls

Comparative analysis:

- Compare allele frequencies across all samples
- Identify:
    - Common alleles
    - Rare alleles

Deliverables:

- Allele typing tables per sample
- Summary of allele frequencies
- Confidence assessment and limitations

Part 5: Biological Interpretation

Goal: Move beyond computation and interpret results biologically.

You must discuss:

- Patterns of **MHC diversity**
- Differences between samples
- Potential implications for:
    - Immune function
    - Disease susceptibility
- **Limitations** of:
    - Short-read sequencing
    - MHC region analysis
    - Allele typing accuracy

This section is critical and will be graded heavily.

---

6. Final Report Structure

Your final report must follow **standard scientific format**:

1. **Introduction**
    - MHC biology

    - Importance of genetic diversity
    - Project objectives
2. **Methods**
    - Tools used
    - Parameters
    - Analysis pipeline
    - Justification of choices
3. **Results**
    - QC results
    - Gene annotation findings
    - MHC coverage analysis
    - Allele typing outcomes
    - Figures and tables
4. **Discussion**
    - Biological interpretation
    - Comparison to known literature
    - Limitations
5. **Conclusion**
    - Key takeaways
    - Future improvements or analyses

---

7. Required Submission Components

You must submit:

- All **code and scripts**
- QC metrics for each sample
- Gene annotation summary tables
- MHC coverage visualizations
- Allele typing results with confidence
- Comparative analysis across all samples

Reproducibility matters.

---

8. Evaluation Criteria

You will be graded on:

- Correctness of bioinformatics methods
- Completeness of required analyses
- Clarity and quality of figures
- Depth of biological insight
- Professional scientific writing

---

9. Learning Objectives Assessment

By completing this project, you should be able to:

1. Convert raw sequencing data into interpretable results
2. Select appropriate tools for specific genomics tasks
3. Identify technical artifacts vs real biological signals
4. Relate genomic variation to immunological function

5. Communicate complex analyses clearly and professionally

---

10. Final Outcome

The final product should resemble a **conference-ready genomics study**, demonstrating that you can:

- Analyze complex genomic regions
- Interpret immune gene variation
- Present results at a professional scientific level

If you want, next I can:

- Convert this into a **student handout (PDF/Word)**
- Add **exact commands/tools** (samtools, bedtools, etc.)
- Create a **grading rubric**
- Simplify it further for undergraduate vs graduate level

look wirte in task orianted way . firs task is they have to anoted the bam files with gtf files . make csv of gene name, statrt , end , sequence seoncd taskis that y have to plot quality disturbuion of rrad fro each bam files . third task is thay have ro extract the mhc genes class 1 , 2, 3, and by pari wise algimnet they have to type (allel of genes) those genes

Got it 👍
Below is a **clear, task-oriented version**, written **step by step**, with **concrete outputs** for each task.
This is suitable to give directly to students as an instruction sheet.

MHC Genomics Analysis Project

Task-Oriented Assignment Guide

This project is organized into **three sequential tasks**.
Each task builds on the previous one and produces specific files that will be used in later steps.

Task 1: Annotate BAM Files Using Gene Annotations

Objective

Annotate aligned sequencing reads with gene information and generate a **gene-level annotation table** for each sample.

Input

- BAM files (aligned sequencing data)
- Gene annotation file (**GTF format**)

What you must do

1. Use the provided **GTF file** to annotate each BAM file.
2. For every gene that has at least one aligned read:
    - Identify genomic coordinates
    - Extract basic gene-level information
3. Generate a **CSV file per sample** containing gene annotations.

Required output (CSV format)

Each CSV file must contain **at least** the following columns:

Column name	Description
gene_name	Official gene symbol

chromosome	Chromosome name
start	Gene start position
end	Gene end position
strand	+ or −
sequence	gene sequence

📌 One CSV file **per BAM file**.

Key questions to think about

- Which genes are covered by the sequencing data?
- Are there genes with unusually high or low coverage?

Task 2: Read Quality and Mapping Quality Analysis

Objective

Assess the **quality of sequencing reads** and alignment reliability for each BAM file.

Input

- BAM files (same as Task 1)

What you must do

1. Extract **mapping quality (MAPQ)** values from each BAM file.
2. For each sample:

    - Compute the distribution of read mapping quality

3. Visualize read quality using plots.

Required plots

For **each BAM file**, generate:

- A **mapping quality distribution plot** (histogram or density plot)

📌 All plots must:

- Have labeled axes
- Clearly indicate the sample name
- Use consistent binning and scale across samples

Required written output

A short paragraph answering:

- Are the reads generally high quality?
- Are there samples with poor mapping quality?
- How might this affect downstream MHC analysis?

Task 3: MHC Gene Extraction and Allele Typing

Objective

Focus specifically on the **MHC region** and perform **allele typing** for immune genes using sequence alignment.

Input

- Annotated gene tables from Task 1
- BAM files

Step 1: Extract MHC Genes

From your annotated gene tables, extract genes belonging to:

- **MHC Class I** (e.g., HLA-A, HLA-B, HLA-C)
- **MHC Class II**
- **MHC Class III**

📌 Create a filtered table containing **only MHC genes**.

Step 2: Extract Gene Sequences

For each MHC gene:

- Extract the **gene sequence**
- Store sequences in FASTA format

Step 3: Pairwise Alignment and Allele Typing

1. Perform **pairwise sequence alignment** between:
    - Sample gene sequences
    - Reference allele sequences
2. Identify the **best-matching allele** for each gene based on:
    - Alignment score
    - Sequence similarity

Required outputs

For each sample:

1. **MHC gene table** containing:
    - Gene name
    - MHC class (I / II / III)
    - Best-matching allele
    - Alignment score
    - Percent identity
2. A short summary describing:
    - Which alleles are most common
    - Differences between samples
    - Confidence and limitations of allele calls

Final Submission Checklist

You must submit:

- ✅ Annotated gene CSV files (Task 1)
- ✅ Read quality distribution plots (Task 2)
- ✅ MHC-only gene tables (Task 3)
- ✅ Allele typing results with alignment scores
- ✅ Short written interpretation for each task

Learning Outcomes

After completing this project, you should be able to:

- Annotate sequencing data using gene models
- Evaluate sequencing and alignment quality
- Extract biologically meaningful subsets of genes
- Apply pairwise alignment for allele typing
- Interpret immune gene variation scientifically