Mitochondrial Variant Calling Project Report

# 1. Introduction

The aim of this project was to identify variants (such as single nucleotide polymorphisms and indels) in the human mitochondrial genome using next-generation sequencing (NGS) data. The raw sequencing data was aligned to the reference genome, followed by variant calling and filtering of low-quality variants. The workflow included the use of several bioinformatics tools such as Bowtie2, SAMtools, Picard, GATK, and IGV.

# 2. Materials and Methods

## 2.1 Tools Used

- Bowtie2: For aligning sequencing reads to the reference genome.  
- SAMtools: For converting and manipulating SAM/BAM files.  
- Picard: For adding read group information to the BAM file.  
- GATK (Genome Analysis Toolkit): For variant calling and filtering.  
- IGV (Integrative Genomics Viewer): For visualizing alignments and variants.

## 2.2 Dataset

- NGS Dataset: SRR30758927 (human mitochondrial genome, single-end reads) was downloaded from the Sequence Read Archive (SRA).  
- Reference Genome: NC\_012920.1 (complete human mitochondrial genome).

# 3. Workflow

The project involved preprocessing the data, aligning reads, calling variants using GATK, and filtering the variants. Below are the detailed steps followed:

## 3.1 Preprocessing of Data

Bowtie2 was used to align the reads to the reference genome, followed by conversion to BAM format and sorting:  
  
- Command:  
bowtie2 -x NC\_012920.1 -U SRR30758927\_trimmed.fastq -S aligned\_reads\_bowtie2.sam

## 3.2 Adding Read Group Information

Read group information was added using Picard to ensure compatibility with GATK for variant calling.

## 3.3 Variant Calling with GATK HaplotypeCaller

Variants were called from the aligned reads using GATK HaplotypeCaller.  
  
- Command:  
gatk HaplotypeCaller \  
 -R NC\_012920.1.fasta \  
 -I sorted\_reads\_bowtie2\_rg.bam \  
 -O output\_variants.vcf  
- Output: VCF file with called variants (output\_variants.vcf).

## 3.4 Variant Filtering

Variants were filtered using GATK VariantFiltration to remove low-quality variants.  
  
- Command:  
gatk VariantFiltration \  
 -R NC\_012920.1.fasta \  
 -V output\_variants.vcf \  
 -O filtered\_variants.vcf \  
 --filter-name "QD\_filter" --filter-expression "QD < 2.0" \  
 --filter-name "FS\_filter" --filter-expression "FS > 60.0" \  
 --filter-name "MQ\_filter" --filter-expression "MQ < 40.0"  
- Output: Filtered VCF file (filtered\_variants.vcf).

# 4. Results

The HaplotypeCaller identified several variants in the mitochondrial genome. After filtering, high-quality variants were retained, and these were stored in the filtered\_variants.vcf file. The filtered variants represent the most likely SNPs and indels in the data, which could be further analyzed.

# 5. Visualization

The aligned reads and variants were visualized using IGV. IGV allowed the confirmation of variant quality by comparing the aligned reads against the reference genome. This step confirmed that the variants were well-supported by multiple reads, and no major anomalies were detected.

# Conclusion

This project successfully identified high-confidence variants in the human mitochondrial genome using NGS data. Key steps included read alignment, variant calling with GATK HaplotypeCaller, and variant filtering. The filtered VCF file contains variants that can be used for further analysis or comparison with known mitochondrial variation databases.