Bioinformatics Project Report

Ege Özgül - 150230708 İpek Güneş Aymergen - 150220728

May 15, 2025

Contents

1 Introduction

DNA sequencing has revolutionized genomics by enabling the determination of the precise order of nucleotides within a DNA molecule. This technology underpins the study of genetic variations, advancing our understanding of diseases, traits, and evolutionary biology. Among its critical applications is the identification of genomic variants, a process known as **variant calling**. Variant calling is instrumental in detecting differences between a sample genome and a reference genome, with applications in cancer research, hereditary diseases, and precision medicine.

Accurate variant calling requires robust bioinformatics pipelines integrating multiple tools and methodologies. This project focuses on evaluating variant detection pipelines for paired germline-tumor datasets. Using a combination of mappers, variant callers, and base recalibration settings, we aim to assess the pipelines' performance through metrics such as precision, recall, F1-score, and accuracy. By identifying the strengths and weaknesses of these configurations, this study provides recommendations for optimizing workflows in variant detection.

The following sections detail the objectives, datasets, tools, and methods used to evaluate variant detection pipelines, with an emphasis on their accuracy and efficiency in detecting single nucleotide polymorphisms (SNPs) and insertion-deletion mutations (Indels).

2 Objective

The primary objective of this study is to evaluate the performance of variant detection pipelines applied to germline-tumor paired datasets. This involves assessing the accuracy and reliability of various configurations by analyzing key metrics, including precision, recall, F1-score, and accuracy, for single nucleotide polymorphisms (SNPs) and insertion-deletion mutations (Indels).

The study focuses on:

- Constructing and testing bioinformatics pipelines using different combinations of mapping tools (BWA and Bowtie) and variant callers (Mutect, SomaticSniper, and Strelka).
- Exploring the impact of base recalibration on the performance of variant detection workflows.
- Comparing computational efficiency and runtime for different pipeline configurations.
- Providing recommendations for optimizing variant detection workflows based on the findings.

By systematically evaluating these pipelines, this project aims to identify configurations that maximize variant detection accuracy while maintaining computational efficiency.

3 Background

3.1 DNA Sequencing

DNA sequencing determines the exact order of nucleotides within a DNA molecule, forming the foundation of modern genomics. Advances in next-generation sequencing (NGS) have enabled rapid, high-throughput sequencing of entire genomes. This technology is pivotal for studying genetic variations, uncovering the molecular basis of diseases, and advancing fields like precision medicine and evolutionary biology.

3.2 Variant Calling and Its Importance

Variant calling identifies genetic differences between a sample genome and a reference genome, focusing on single nucleotide polymorphisms (SNPs) and insertion-deletion mutations (Indels). Accurate variant detection is crucial for understanding the genetic underpinnings of cancer, hereditary diseases, and other conditions. Reliable detection pipelines are essential for generating high-confidence variant data, which can directly influence downstream analyses, including functional annotation and clinical decision-making.

3.3 Challenges in Variant Detection

The accuracy of variant calling depends on multiple factors:

- The quality of raw sequencing data and alignment to the reference genome.
- The choice of mapping tools and variant callers, each with varying strengths and limitations.
- The inclusion of preprocessing steps, such as base recalibration, to reduce technical noise.

Despite advancements, challenges remain in achieving high recall for true variants while minimizing false positives, especially for complex mutation types like Indels and structural variants.

3.4 Study Context

This study evaluates the performance of bioinformatics pipelines for variant detection using germline-tumor paired datasets. By comparing combinations of mappers (BWA, Bowtie), variant callers (Mutect, SomaticSniper, Strelka), and recalibration settings, the project aims to address critical challenges in optimizing variant detection workflows.

4 Project Overview

4.1 Tasks

This study involved the systematic evaluation of DNA variant detection pipelines by completing the following tasks:

1. Developing pipelines using combinations of mapping tools (BWA and Bowtie) and variant callers (Mutect, SomaticSniper, and Strelka).
2. Analyzing the impact of preprocessing steps, specifically base recalibration, on variant detection performance.
3. Calculating key metrics, including precision, recall, F1-score, and accuracy, for both SNPs and Indels across different pipeline configurations.
4. Comparing computational runtimes of pipelines to assess their efficiency.
5. Visualizing performance and overlaps in variant detection using heatmaps, PCA plots, and histograms.

4.2 Tools Used

The pipelines were constructed and executed using a variety of bioinformatics tools:

- **Mapping Tools:** BWA and Bowtie for aligning sequencing reads to the *Homo sapiens* assembly38 reference genome.
- **Variant Callers:** Mutect, SomaticSniper, and Strelka for detecting genomic variants.
- **Quality Control Tools:** FastQC and MultiQC for assessing sequencing read quality.
- **Processing Tools:** SAMtools for converting, sorting, and indexing BAM files; bcftools for parsing and filtering VCF files.
- **Statistical and Visualization Tools:** Python libraries (NumPy, pandas, matplotlib, seaborn) for metric calculation and result visualization.
- **Benchmarking Resources:** High-confidence SNP and Indel datasets from the 1000 Genomes Project and dbSNP for validation.

4.3 Dataset Description

The analysis used paired-end sequencing data representing germline-tumor pairs:

- **Germline Dataset:** SRR7890850.
- **Tumor Dataset:** SRR7890851.

The datasets were aligned to the *Homo sapiens* assembly38 reference genome. Additional benchmarking datasets included:

- High-confidence SNP and Indel sets from the 1000 Genomes Project.
- dbSNP annotations for known variant validation.
- High-confidence genomic regions to focus on reliable areas for variant calling.

4.4 Pipeline Configurations

Twelve pipeline configurations were tested, combining:

- Two mapping tools (BWA, Bowtie).
- Three variant callers (Mutect, SomaticSniper, Strelka).
- Two base recalibration settings (with and without recalibration).

Each configuration was evaluated for its ability to detect variants with high precision and recall while maintaining computational efficiency.

5 Methods

5.1 FASTQ-Level Analysis

The raw sequencing data underwent quality control and preprocessing to ensure its suitability for downstream analysis.

5.1.1 Data Acquisition

The datasets used in this study included paired-end FASTQ files:

- **Germline Dataset:** SRR7890850.
- **Tumor Dataset:** SRR7890851.

These datasets were sourced from the European Nucleotide Archive (ENA) and validated using MD5 checksum to ensure data integrity.

5.1.2 Quality Control

The raw sequencing reads were analyzed using **FastQC** to assess:

- Per-base sequence quality.
- GC content distribution.

Per Base Quality Plots The Per Base Quality Plots assess the sequencing quality for each position in the reads. All FASTQ files exhibited high-quality scores, with minor declines toward the ends of the reads, a common trend in Illumina sequencing data.

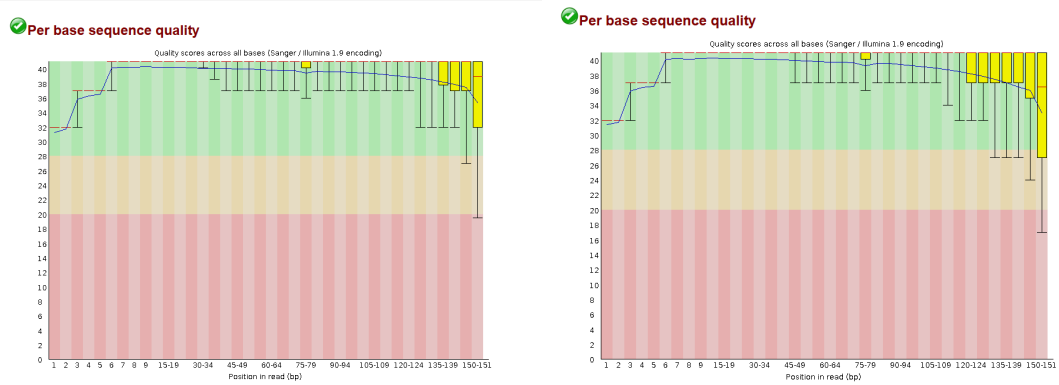

Figure 1: Per Base Quality Plots for SRR7890850 (Germline dataset). Left: Read 1. Right: Read 2. High-quality scores were observed across most bases, with minor declines toward the ends.

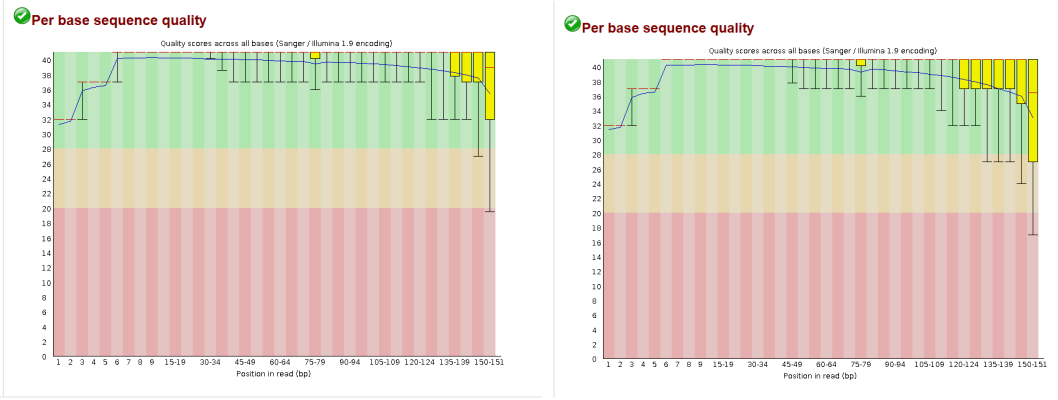

Figure 2: Per Base Quality Plots for SRR7890851 (Tumor dataset). Left: Read 1. Right: Read 2. Quality scores were consistent with germline data.

GC Content Plots The GC Content Plots show the distribution of GC content across all reads. All FASTQ files exhibited GC content distributions consistent with expectations for human genomic data, with no significant contamination.

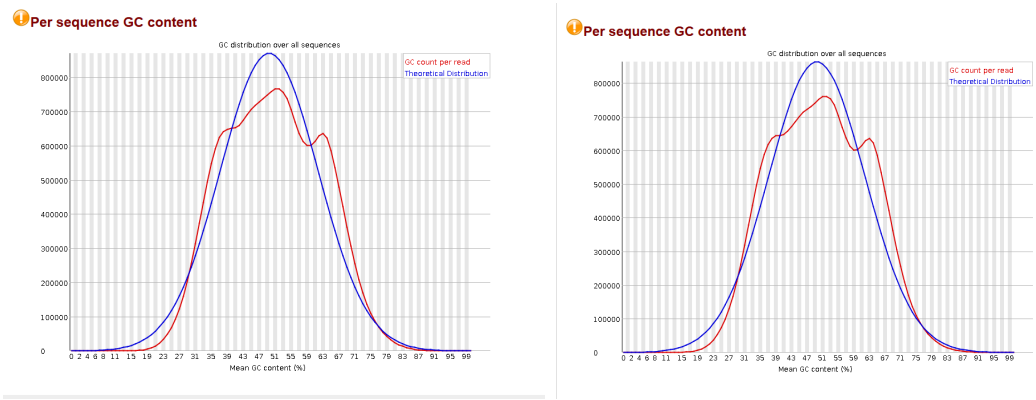

Figure 3: GC Content Plots for SRR7890850 (Germline dataset). Left: Read 1. Right: Read 2. GC content distributions align with theoretical expectations.

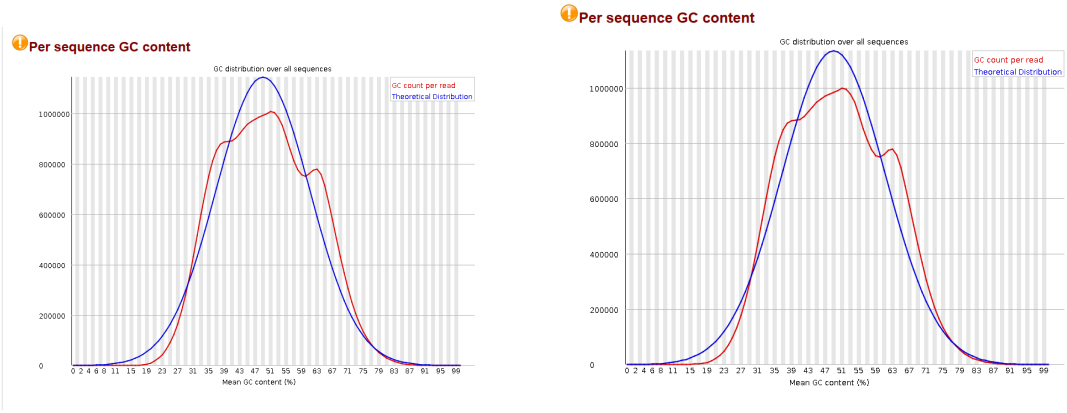

Figure 4: GC Content Plots for SRR7890851 (Tumor dataset). Left: Read 1. Right: Read 2. The GC content was consistent with germline data.

5.2 BAM-Level Analysis

The aligned sequencing data was processed to evaluate alignment quality and ensure readiness for variant calling.

5.2.1 Read Alignment

The paired-end reads were aligned to the *Homo sapiens* assembly38 reference genome using two mapping tools:

- **BWA mem:** A widely used aligner optimized for high accuracy.
- **Bowtie:** An alternative aligner for comparative performance analysis.

5.2.2 BAM Processing

The aligned SAM files were converted into BAM format and processed using **SAMtools**:

- Sorting and indexing for efficient storage and retrieval.
- Calculating alignment statistics using `flagstat`.
- Evaluating depth of coverage to ensure sufficient data for variant detection:
  - Germline dataset: 19.73x coverage.
  - Tumor dataset: 21.83x coverage.

5.2.3 Alignment Quality Assessment

Key metrics evaluated for alignment quality included:

- High mapping rates (˜99.45% for both datasets).
- Properly paired reads (˜97.88%).
- Adequate average coverage depths for reliable variant calling.

These metrics confirmed that the datasets were suitable for downstream analysis, with no significant alignment issues detected.

6 VCF-Level Analysis

The analysis of Variant Call Format (VCF) files focused on evaluating the performance and similarity of different variant detection pipelines using key metrics and visualization tools. Below, we summarize the findings:

6.1 Metrics Table

The performance metrics for each pipeline configuration, including SNP true positives (TP), false positives (FP), false negatives (FN), precision, recall, F1-score, and variant counts, are summarized in Figure 5.

File	SNP_TP	SNP_FP	SNP_FN	SNP_Precision	SNP_Recall	SNP_F1	SNP_Variant_Count
final_bwa_strelka_with_Base_onlyPass.vcf.gz	936	1875	225	33.297.758	80.620.155	47.129.908	2811
final_bwa_strelka_no_Base_onlyPass.vcf.gz	943	2228	218	29.738.252	81.223.083	43.536.471	3171
final_bwa_somaticsniper_withBase.vcf.gz	854	7435	307	1.030.281	73.557.278	18.074.072	8289
final_bwa_somaticsniper_no_Base.vcf.gz	870	8527	291	9.258.273	749.354	16.480.391	9397
final_bwa_mutect_withBase_onlyPass.vcf.gz	710	282	451	7.157.258	61.154.177	65.954.481	992
final_bwa_mutect_noBase_onlyPass.vcf.gz	699	363	462	65.819.209	60.206.718	62.887.988	1062
final_bowtie_strelka_with_Base_onlyPass.vcf.gz	826	723	335	53.324.725	71.145.564	60.959.408	1549
final_bowtie_strelka_no_Base_onlyPass.vcf.gz	831	955	330	46.528.555	71.576.227	56.396.334	1786
final_bowtie_somaticsniper_with_Base.vcf.gz	781	4740	380	14.145.988	67.269.595	23.376.233	5521
final_bowtie_somaticsniper_no_Base.vcf.gz	804	6054	357	11.723.534	69.250.645	20.052.373	6858
final_bowtie_mutect_withBase_onlyPass.vcf.gz	876	314	285	73.613.445	75.452.196	74.521.479	1190
final_bowtie_mutect_no_Base_onlyPass.vcf.gz	801	258	360	75.637.393	68.992.248	72.162.161	1059
final_galaxy_bowtie_strelka_onlyPass.vcf.gz	832	415	329	66.720.128	7.166.236	69.102.989	1247
final_galaxy_bwa_strelka_onlyPass.vcf.gz	926	1109	235	45.503.685	79.758.828	57.947.433	2035

Figure 5: Performance metrics for each pipeline configuration.

6.2 Runtime and Efficiency

The runtime comparison (Figure 16) highlighted trade-offs between computational efficiency and variant detection performance:

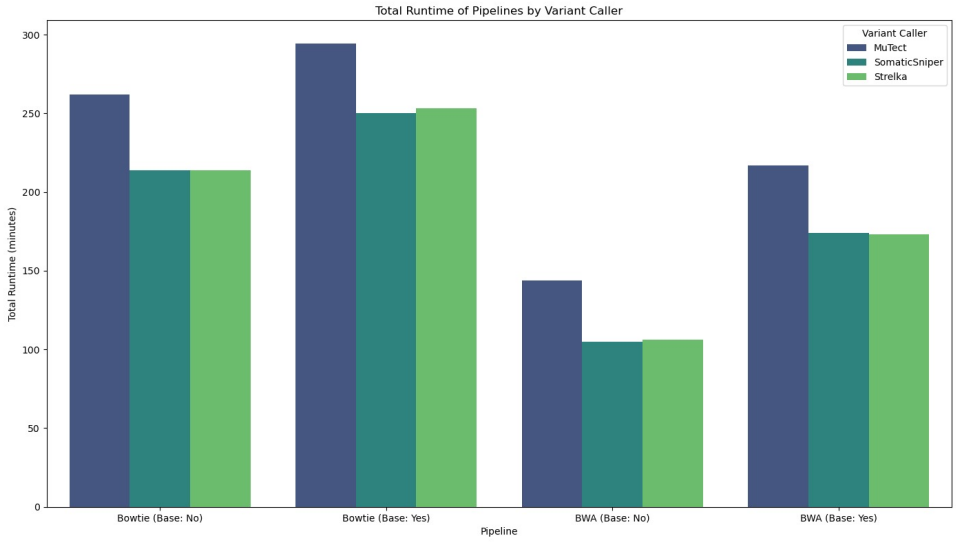

Figure 6: Runtime comparison of pipelines by variant caller and preprocessing.

- Mutect was the most computationally intensive caller, requiring up to 300 minutes for Bowtie pipelines with base recalibration.
- Strelka offered a balance of runtime efficiency and performance, while SomaticSniper demonstrated faster runtimes but poorer variant detection metrics.
- Base recalibration consistently increased runtime by approximately 20–30%, a reasonable trade-off given its impact on precision and F1-scores.

6.3 Variant Counts Across Pipelines

Figure 7 shows the total number of SNP variants detected by each pipeline. This helps compare the overall sensitivity of pipelines in identifying SNPs.

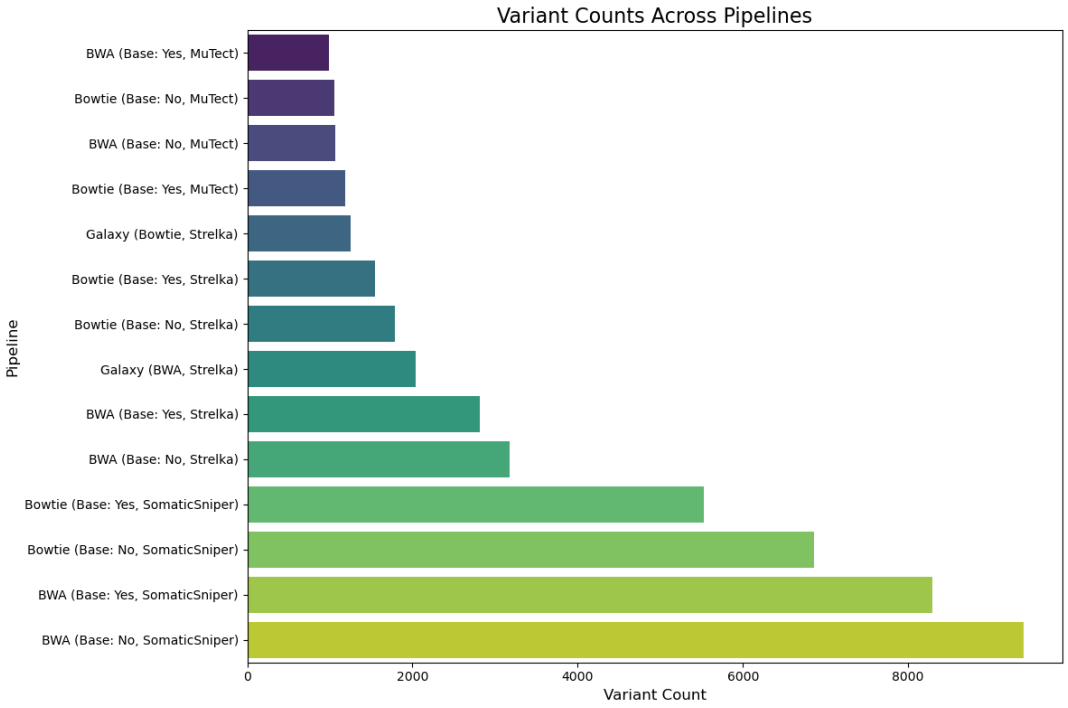

Figure 7: Total Variant Counts Across Pipelines. Each bar represents a pipeline, grouped by mapper and variant caller. Higher counts may indicate greater sensitivity but could also result in more false positives.

Key observations from the figure include:

- SomaticSniper pipelines detected the highest number of variants, indicating high sensitivity but likely contributing to high false positive rates.
- Mutect pipelines detected fewer variants, suggesting a more conservative approach with stricter thresholds.
- Strelka pipelines achieved a balance, detecting moderate variant counts while maintaining reasonable precision and recall.

6.4 Precision and Recall Across Pipelines

Figure 8 compares the precision and recall scores for each pipeline. Precision represents the proportion of correctly identified SNPs out of all SNPs detected, while recall measures the proportion of real SNPs that were successfully detected.
Key observations from the figure include:

- Pipelines using Mutect with base recalibration (e.g., BWA or Bowtie) achieved both high precision and recall, indicating balanced performance.
- SomaticSniper pipelines showed very low precision and recall, suggesting poor overall performance.

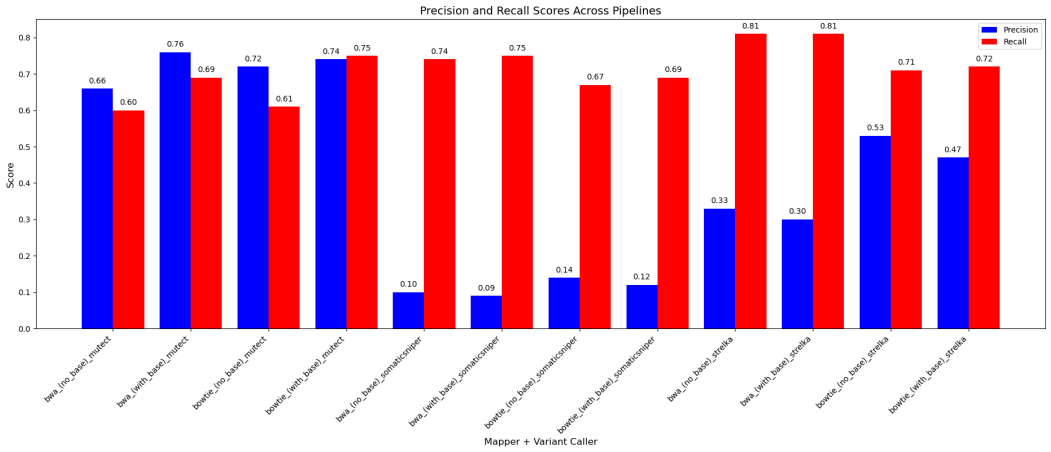

Figure 8: Precision and Recall Scores Across Pipelines. The pipelines are grouped by mapper and variant caller configurations. Higher precision indicates fewer false positives, while higher recall indicates fewer missed SNPs.

- Strelka pipelines generally achieved high recall but moderate precision, indicating a bias toward sensitivity at the cost of accuracy.

6.5 F1 Score Heatmap Across Pipelines

Figure 9 presents a heatmap comparing the F1 scores across pipelines. F1 score is the harmonic mean of precision and recall, providing a balanced measure of pipeline performance.

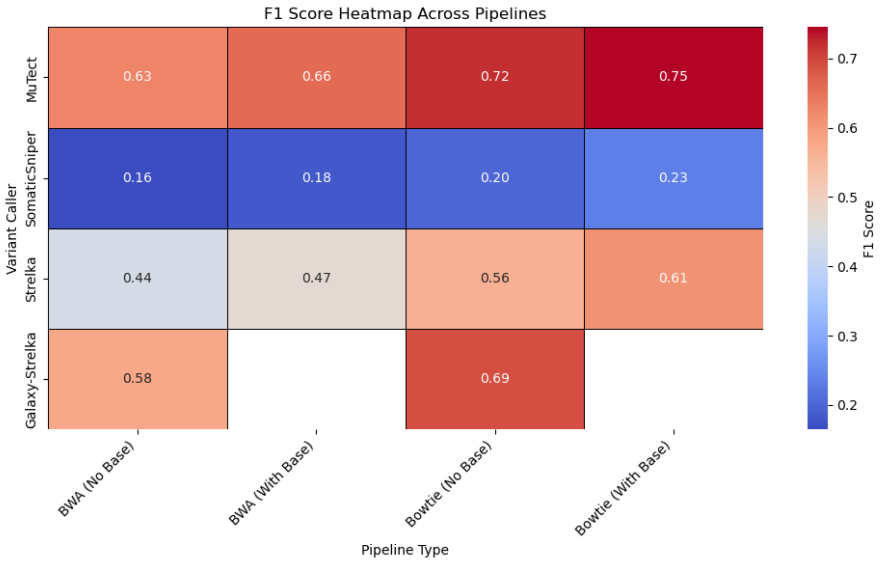

Figure 9: F1 Score Heatmap Across Pipelines. Rows represent variant callers, and columns represent pipeline configurations. Higher scores (red) indicate better performance.

Key observations from the heatmap include:

- Mutect pipelines consistently achieved the highest F1 scores, particularly with base recalibration.
- Strelka pipelines had moderate F1 scores, reflecting their high recall but only average precision.
- SomaticSniper pipelines performed poorly, with significantly lower F1 scores across all configurations.

6.6 Stacked Bar Chart of TP, FP, and FN Across Pipelines

Figure 10 presents a stacked bar chart comparing True Positives (TP), False Positives (FP), and False Negatives (FN) across various pipelines. This visualization highlights the distribution of each metric for every pipeline, helping to understand their relative performance.

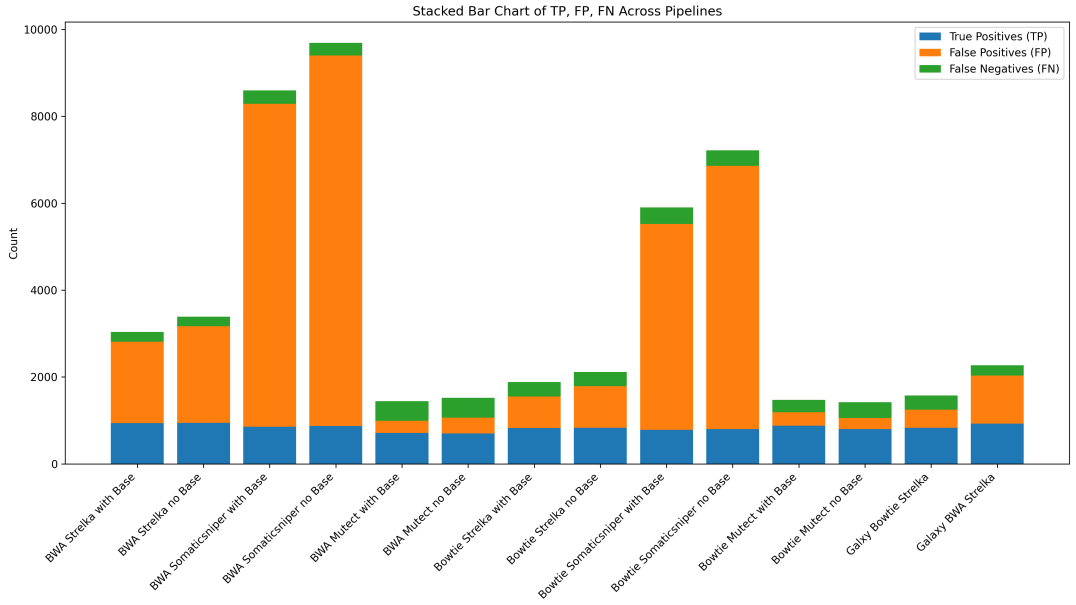

Figure 10: Stacked Bar Chart of TP, FP, and FN Across Pipelines. Each bar represents a pipeline, with the contributions of TP, FP, and FN stacked to show the total detected variants and their breakdown.

Key observations from the chart include:

- Pipelines using SomaticSniper consistently detected a high number of variants, with False Positives (orange) dominating the stack, indicating poor precision.
- Mutect pipelines detected fewer total variants but showed a better balance of True Positives (blue) with minimal False Positives (orange) and False Negatives (green), indicating strong precision and recall.
- Strelka pipelines performed moderately, with a relatively balanced distribution of TP, FP, and FN across configurations.

This chart provides a clear overview of how each pipeline's performance metrics contribute to their overall variant detection, aiding in the selection of pipelines that align with specific project goals.

6.7 Similarity Heatmap Analysis

The similarity heatmap (Figure 11) quantifies the overlap in variant calls between pipelines. Key observations include:

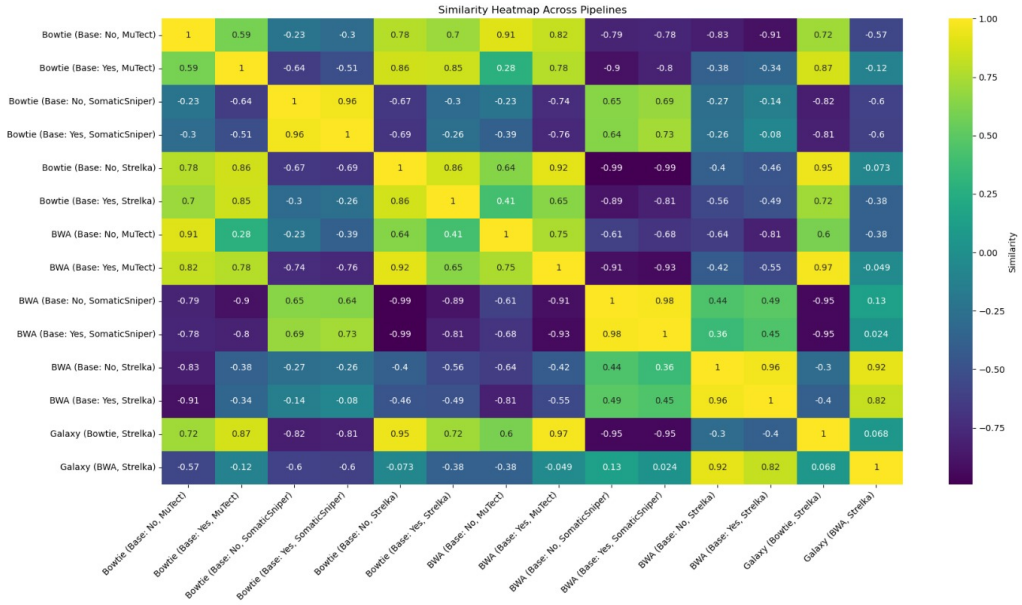

Figure 11: Similarity heatmap across pipelines based on variant calls.

- Pipelines using Mutect exhibited high similarity (≥ 0.91) when employing the same mapper (e.g., BWA with or without base recalibration).
- SomaticSniper demonstrated notable dissimilarity (≤ 0.64) with other pipelines, reflecting its reduced precision and sensitivity.
- The highest similarity was observed between BWA pipelines with recalibration and Strelka (≥ 0.97), underscoring the influence of preprocessing steps on variant overlap.

6.8 Principal Component Analysis (PCA)

The PCA plot (Figure 12) illustrates the separation and clustering of VCF pipelines based on their variant calling profiles. The first principal component (PC1) explained 57.30% of the variance, while the second principal component (PC2) accounted for 11.93%.

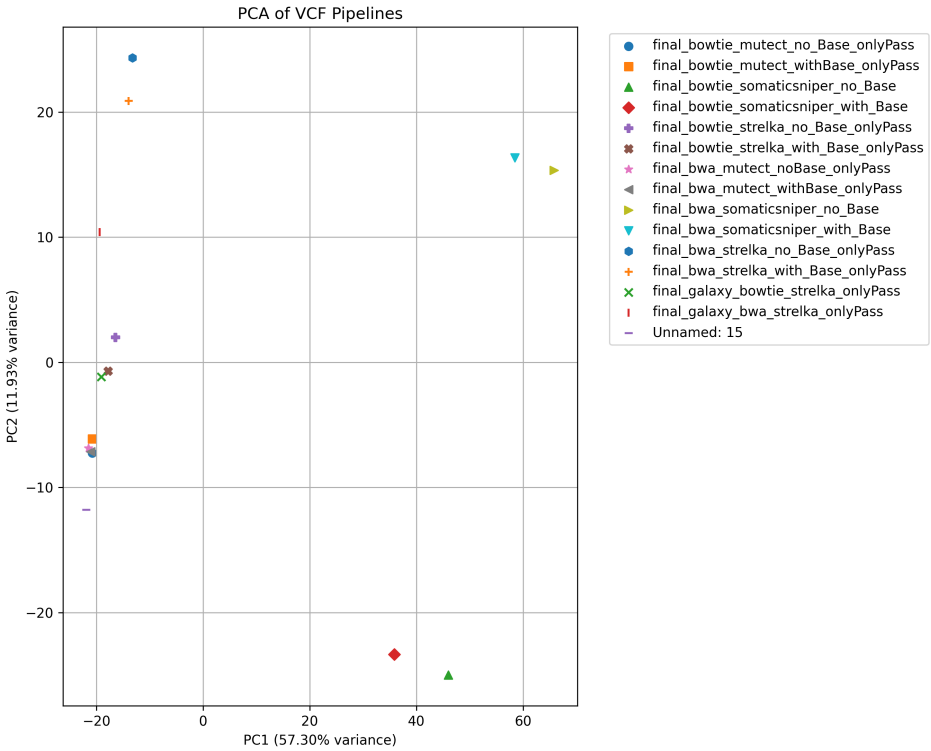

Figure 12: Principal Component Analysis (PCA) of VCF pipelines.

- Pipelines employing the same mapper (e.g., BWA or Bowtie) but different variant callers were closely clustered, highlighting their overall similarity in variant detection.
- Configurations with base recalibration demonstrated improved separation, especially for Mutect, suggesting enhanced precision and reliability in detecting true variants.
- Galaxy pipelines showed distinct clustering, indicating deviations in variant calling methodology or thresholds compared to traditional pipelines.

6.9 Variant Contributions to PCA Components

The top 20 variants contributing to PC1 and PC2 (Figures 13 and 14) highlighted key genomic regions driving pipeline differences:

- Variants on chromosomes 1, 9, and 22 were consistently influential in PC1, representing shared detection challenges across pipelines.
- PC2 contributions were more evenly distributed across chromosomes, reflecting broader differences in mapper and variant caller configurations.

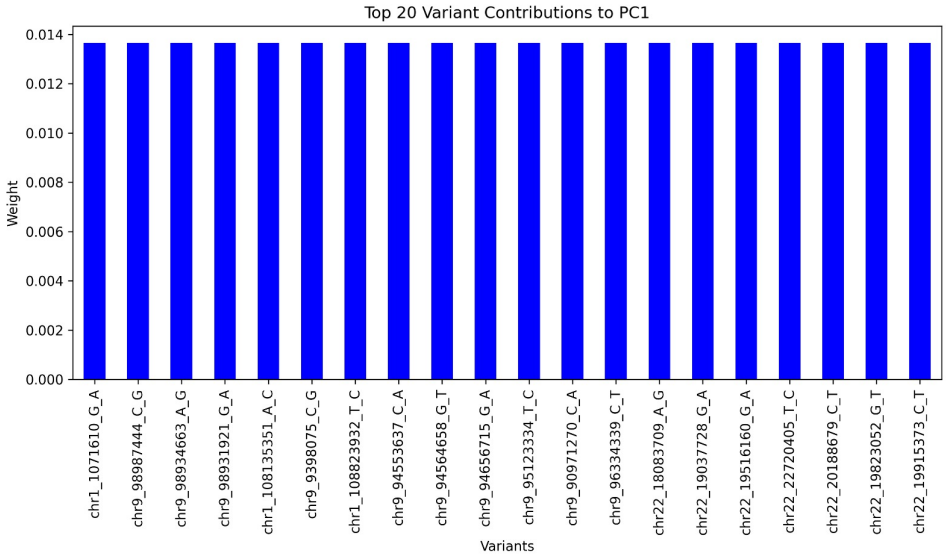

Figure 13: Top 20 variant contributions to PC1.

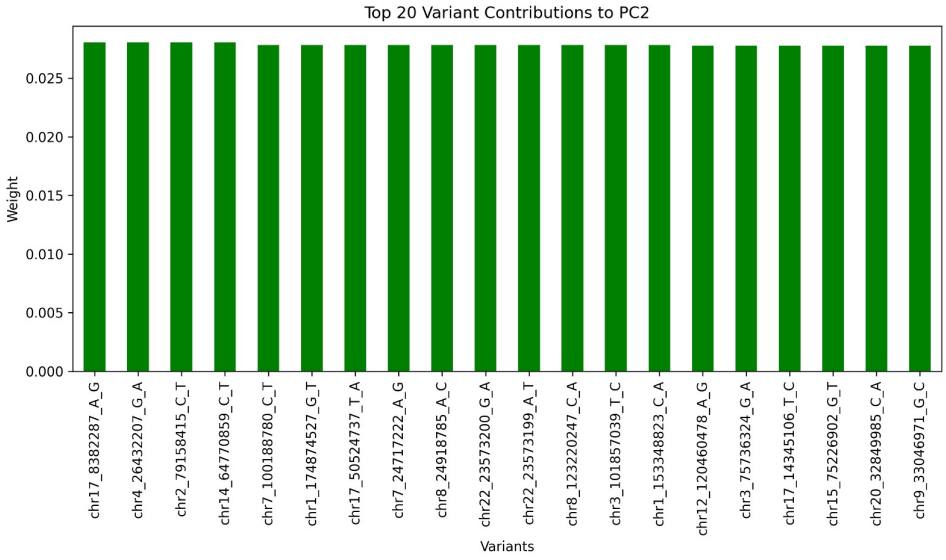

Figure 14: Top 20 variant contributions to PC2.

6.10 Clustered Intersection/Union Heatmap

Figure 15 presents a clustered heatmap showing the intersection/union ratios of variant calls across pipelines. This metric indicates the degree of overlap between the variants detected by different pipelines. The heatmap is hierarchically clustered to group pipelines with similar variant detection patterns.
Key observations from the heatmap include:

- Pipelines using the same variant caller (e.g., SomaticSniper or Mutect) tend to cluster together, indicating consistent detection patterns within the same tool.
- Pipelines with base recalibration generally show higher intersection/union ratios compared to those without recalibration, emphasizing the importance of preprocessing.

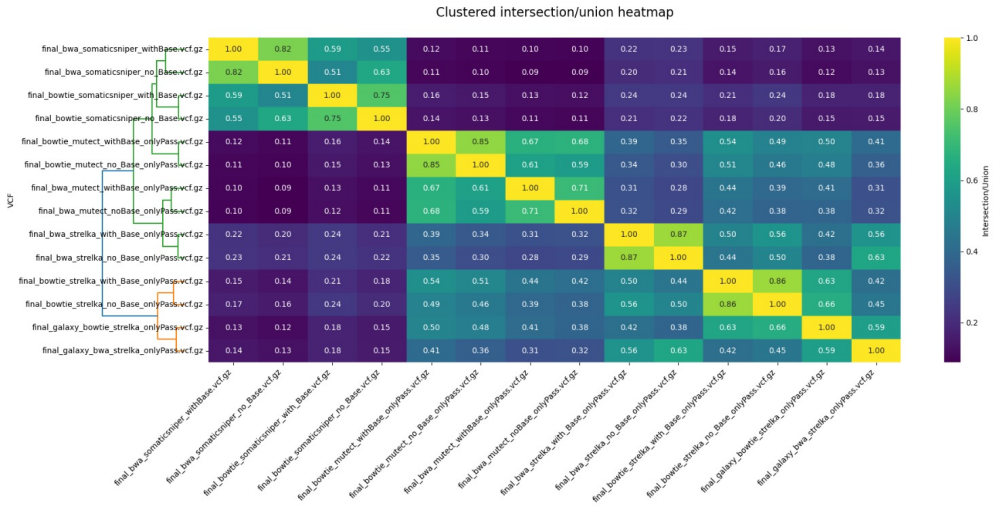

Figure 15: Clustered Intersection/Union Heatmap of Variant Calls Across Pipelines. The color scale represents the intersection/union ratio, with higher values (yellow) indicating greater overlap between pipeline variant calls and lower values (purple) indicating lesser overlap. Hierarchical clustering groups pipelines with similar detection patterns.

- Strelka pipelines exhibit moderate overlaps with other pipelines, reflecting balanced sensitivity and precision.
- Galaxy-based pipelines show distinct clustering, highlighting differences in detection methodologies compared to traditional pipelines.

This heatmap provides valuable insights into the agreement between pipelines and their underlying variant detection strategies.

Summary

The VCF-level analysis underscored the importance of pipeline configuration in variant detection. Base recalibration emerged as a critical factor for improving precision and F1-scores, particularly for Mutect and Strelka. While Strelka prioritized sensitivity, Mutect achieved a balance between precision and recall, making it suitable for high-confidence analyses. The PCA and heatmap analyses further demonstrated the distinct characteristics of pipelines, providing insights for optimizing variant detection workflows.

7 Results

7.1 Quality Assessment of Sequencing Data

The raw sequencing data quality was assessed using **FastQC**. This analysis evaluated the per-base quality scores, GC content distribution, and other key metrics to ensure the data's readiness for downstream processing.

7.1.1 Per Base Quality Analysis

The Per Base Quality Plots assess the sequencing quality for each base position. High-quality scores (Phred score ¿ Q30) were observed across most bases in all FASTQ files. Minor declines in quality toward the ends of the reads were noted, a common feature of Illumina sequencing technology.

7.1.2 GC Content Analysis

The GC Content Plots evaluate the distribution of GC content across all reads. All FASTQ files exhibited GC content distributions consistent with the human genome (40–60% GC content), with no significant deviations or contamination detected.

7.1.3 Summary of Findings

The quality assessment confirmed the high integrity of the sequencing data:

- **Per Base Quality:** All FASTQ files showed consistently high-quality scores across most bases, with only minor declines toward the ends.
- **GC Content:** GC content distributions were consistent with expected values for human genomic data, showing no contamination or abnormalities.

7.2 Performance Metrics for Variant Detection

The performance of the pipelines was evaluated using key metrics, including precision, recall, F1-score, and accuracy, for both SNPs and Indels. These metrics were calculated for each combination of mappers (BWA, Bowtie), variant callers (Mutect, SomaticSniper, Strelka), and recalibration settings (with and without base recalibration).

The results show that Mutect achieves the highest F1-scores and accuracy across configurations, particularly with base recalibration. Strelka exhibits strong recall but lower precision and accuracy. SomaticSniper consistently underperforms, with low precision and F1-scores.

7.3 Runtime Comparison

The computational efficiency of the pipelines was assessed by comparing runtimes for key steps, including mapping, base recalibration, and variant calling. The results highlight the trade-offs between runtime and performance across different configurations.
The results demonstrate that BWA is faster for mapping compared to Bowtie. Among variant callers, Mutect requires the longest runtime, while Strelka and SomaticSniper are significantly faster. Base recalibration adds noticeable overhead across all configurations.

7.4 Summary of Findings

- **Quality Assessment:** High sequencing quality and consistent GC content were observed across all datasets, confirming their suitability for downstream analysis.

Mapper	BQSR?	Mapping+MarkDup (Normal)	Mapping+MarkDup (Tumor)	BaseRecal (Normal)	BaseRecal (Tumor)	MuTect	SomaticSniper	Strelka
Bowtie	No	~1 h 38 min	~1 h 46 min	N/A	N/A	~58 min	~10 min	~10 min
Bowtie	Yes	~1 h 38 min	~1 h 46 min	~36 min 16 s	~24 min 49 s	~54 min	~10 min	~13 min
BWA	No	~53 min	~43 min	N/A	N/A	~48 min	~9 min	~10 min
BWA	Yes	~53 min	~43 min	~38 min	~29 min	~54 min	~11 min	~10 min

Figure 16: Runtime Comparison for Pipeline Configurations (in minutes).

- **Performance Metrics:** Mutect exhibited the highest precision and F1-scores, making it the most reliable variant caller in this study. Strelka demonstrated excellent recall but lower overall accuracy, while SomaticSniper consistently underperformed.
- **Runtime Comparison:** BWA outperformed Bowtie in terms of runtime for alignment. Base recalibration added computational overhead but significantly improved the precision and F1-scores for Mutect.
- **Trade-offs:** The choice of mapper and variant caller depends on the specific requirements for precision, recall, and computational efficiency in a given application.

8 Discussion

The evaluation of variant detection pipelines revealed several critical insights into the performance of mapping tools, variant callers, and preprocessing strategies:

- **Pipeline Performance:** Mutect demonstrated the most balanced performance, achieving the highest F1-scores and accuracy across configurations. This underscores its reliability for detecting both SNPs and Indels.
- **Impact of Base Recalibration:** Base recalibration significantly improved precision and F1-scores for Mutect and moderately for Strelka. However, its impact on recall was minimal, suggesting that recalibration primarily reduces false positives rather than increasing sensitivity.
- **Runtime Efficiency:** BWA outperformed Bowtie in runtime for mapping, while Bowtie required nearly double the time. Mutect was the most computationally intensive variant caller, while Strelka and SomaticSniper were faster but less precise.
- **Challenges in Indel Detection:** All pipelines struggled with detecting Indels, highlighting an area that requires further optimization in variant calling algorithms and preprocessing techniques.

Despite the strong performance of certain pipelines, challenges remain in achieving high sensitivity for complex variants and balancing computational efficiency with accuracy. These findings suggest that future research should focus on optimizing Indel detection and exploring advanced preprocessing methods.

Limitations

The study has a few notable limitations:

- **Dataset Scope:** The analysis was limited to a single germline-tumor paired dataset, which may not generalize to larger and more diverse datasets.
- **Variant Types:** Structural variants and larger mutations were not considered, restricting the study to SNPs and small Indels.
- **Functional Insights:** The lack of functional annotation for detected variants limits the biological relevance of the findings.

9 Conclusion and Recommendations

9.1 Conclusion

This study evaluated the performance of DNA variant detection pipelines using germline-tumor paired datasets. Key findings include:

- **Performance Highlights:** Mutect emerged as the most reliable variant caller, achieving the highest precision, F1-scores, and accuracy. Strelka demonstrated excellent recall but lower precision, making it suitable for sensitivity-focused applications.

- **Runtime Efficiency:** BWA was faster for mapping compared to Bowtie, and Mutect required the longest runtime for variant calling, reflecting a trade-off between accuracy and computational cost.
- **Challenges in Indel Detection:** The inability of pipelines to reliably detect Indels highlights a key area for future research and optimization.

9.2 Recommendations

Based on the findings, the following recommendations are proposed:

- **Pipeline Optimization:** Mutect with BWA and base recalibration is recommended for applications prioritizing accuracy, while Strelka is suggested for sensitivity-focused analyses.
- **Indel Detection:** Future studies should evaluate alternative tools specialized for Indel detection and explore deeper sequencing coverage to enhance sensitivity.
- **Dataset Diversity:** Expanding analyses to larger, multi-sample datasets and including structural variants will provide more comprehensive insights into pipeline performance.
- **Functional Annotation:** Integrating tools like ANNOVAR or SnpEff for functional annotation will add biological relevance to detected variants, aiding clinical and research applications.