Analysis and Comparison of Next Generation Sequencing Pipelines

Emre Çamlıca

January 3, 2024

Abstract

Sequencing of genetic data has become tenfold easier thanks to NGS techniques. However, the need for accurate detection of mutations in the sequenced data remains a problem. To overcome this, statistical and comparative analysis of NGS pipelines is needed. I statistically and comparatively analyzed 12 different pipelines, having three different options for variant callers, two different options for aligners, and the option to do base calibration or not. I used a tumor sample and a normal sample as the pipeline input. My analyses found a strong correlation between the choice of the variant caller and the statistical performance. The highest performance was observed with the Mutect variant caller, using the BWA mapper, with base recalibration applied. The other two options also had a considerable effect on performance. My analysis of the concordance between the pipelines yielded similar results, where the variant caller had the most impact on similarity. The type of aligner used also had an observable impact.

1 Introduction

DNA is the building block of every living organism on the earth. DNA sequencing refers to the process of identifying the structure of the nucleotide chains of individuals. In the past, it took decades to analyze the genome of individuals. Today, this operation can be done within a day, thanks to the Next Generation Sequencing (NGS) technology [1]. With NGS, millions of DNA parts are sequenced parallelly to obtain detailed genetic data of individuals [1]. It is crucial to obtain correct results when working with biological data, as any error might lead to wrong conclusions about an individual's health. In that manner, it is crucial to correctly detect variants and make sure of their consistency [2]. The lack of data considering the accuracy of variant calling pipelines makes it hard to evaluate their performances [2]. Hence, my project aims to contribute to the analysis of the performances of different variant calling pipelines.

I analyzed the performances of 12 different pipelines, using variant callers, short read aligners, and the option of doing base recalibration as parameters. I used 3 variant callers, namely, Mutect, Strelka, and SomaticSniper, and two aligners, BWA and Bowtie. I used the CoSAP [3] library, which is a helpful tool for creating variant calling pipelines, to run the pipelines. I used the data obtained from the 1000 Genomes Project to run and analyze the performance of the pipelines. I used the VCFtools [4] library to further filter the VCF files obtained from the Mutect and Strelka pipelines, keeping the "PASS" and "." parameters.

I did my analyses with an HP OMEN Laptop, model: 15-dc1xxx. I had 16 GB RAM, I used dual boot Linux with the Linux memory size of 192 GB and I had an Intel CORE i5, 9th Gen processor.

2 Results

2.1 Concordance Between Different Pipelines

To visualize the concordance between different pipelines, I used cluster maps. First, I recorded every different variant from every different pipeline, including the grand truth. After that, for every unique variant, I checked if that variant was identified by each pipeline. I used 1 and 0 values for whether the variant exists or not. I standardized the resulting array, and with the help of a data frame, I created the cluster map of the standardized array. Figure 1 displays the resulting cluster map. It can be observed from the resulting dendrogram of the pipelines that, the pipelines using the same Variant caller are closely related. Except for the Mutect base recalibrated pipeline using the Bowtie aligner, the option of mapper accounts for the second-highest level of similarity.

The heat map portion of the cluster map displays that the variant calls made by the same variant caller are correlated in general.

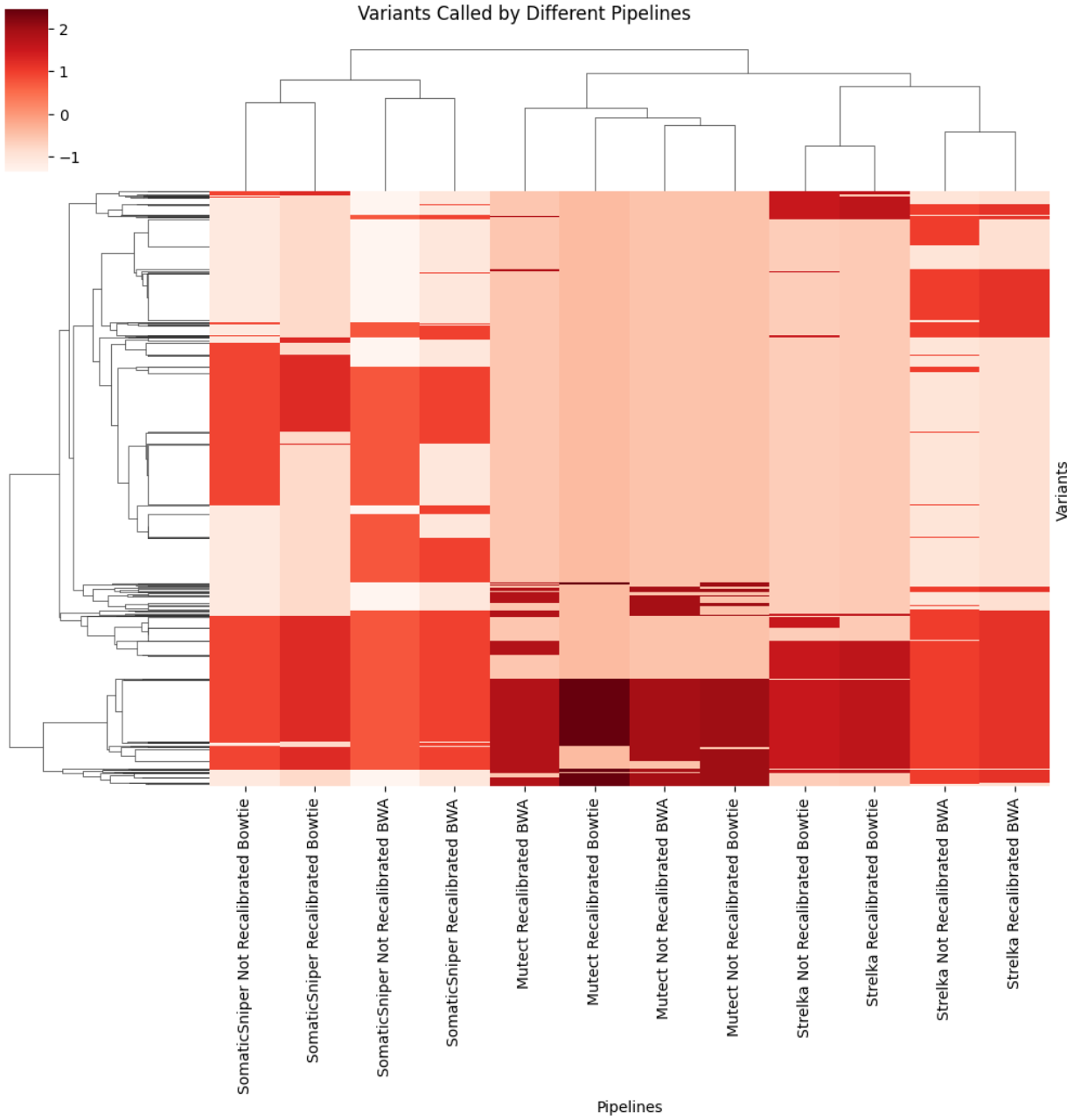

Figure 1: Cluster Map and Dendrogram of All Variants Called by Different Pipelines

Using the same array of 1 and 0's, I calculated the Jaccard distance array between each different pipeline. Figure 2 shows the resulting clustermap of the Jaccard distances. The highest similarity is obtained when using the same variant caller. Furthermore, it can be observed that the pipelines using the same aligner are slightly more similar than the ones that use different mappers. The effect of base recalibration is hard to observe from the clustermap.

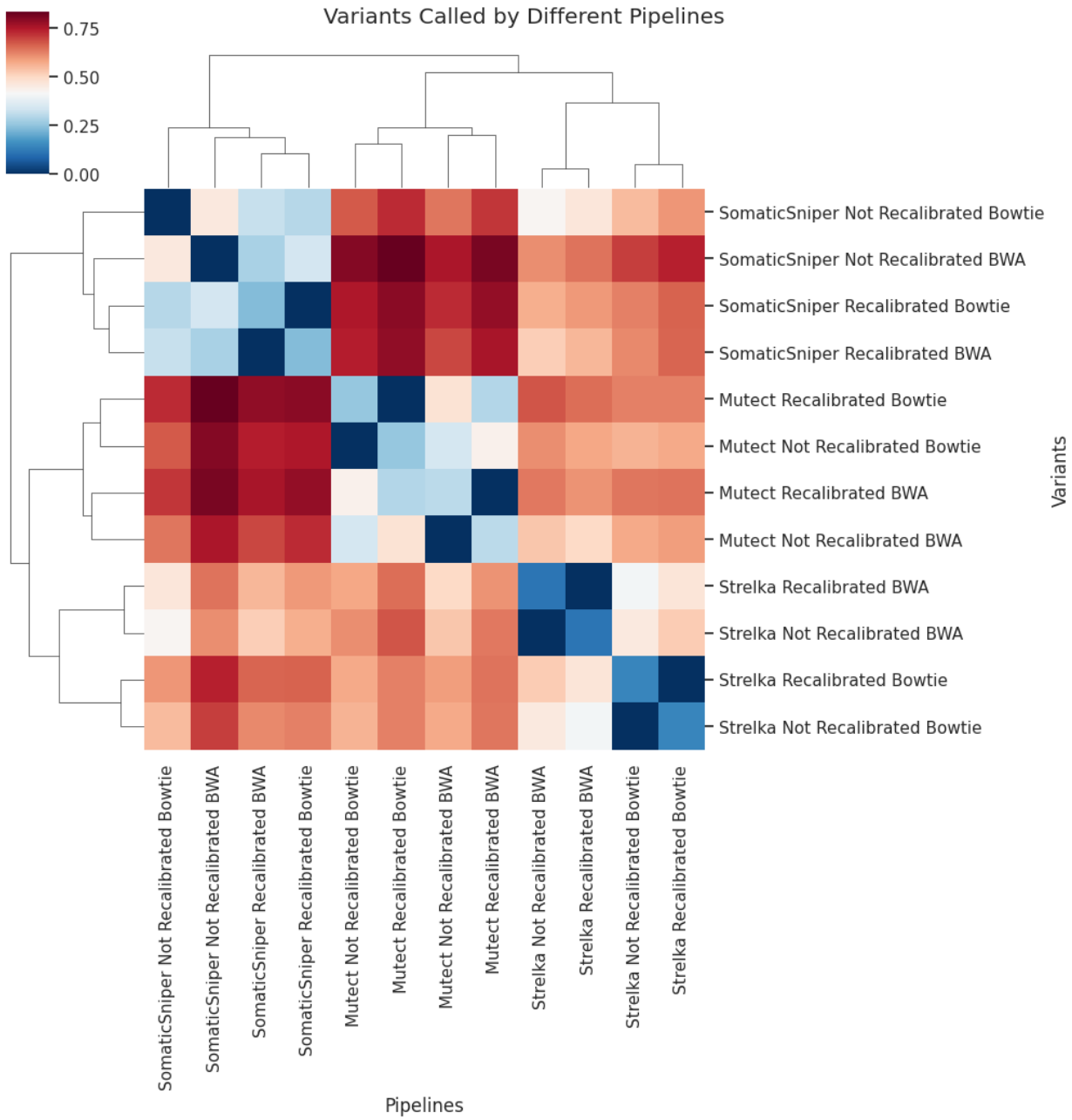

Figure 2: Clustermap Displaying the Jaccard Distance Between Different Pipelines

Figure 3 shows the result of the Principal component analysis for the array constructed by the pipeline-variant data, using two principal components. In the plot, the transparent shapes correspond to the pipelines without base recalibration applied, and the opaque shapes correspond to the base recalibrated pipelines.

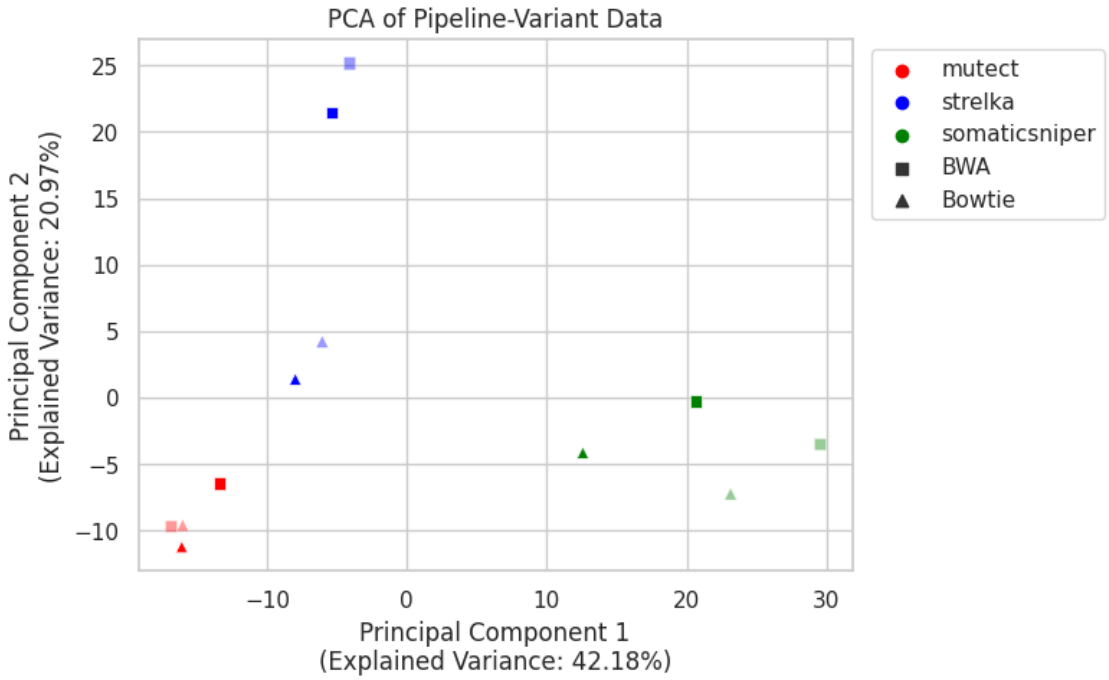

Figure 3: Results of the Principal Component Analysis

It can be observed that the pipelines having the same variant caller are closely clustered, especially the Mutect and SomaticSniper ones. It can further be observed that the type of the aligner played a role in the clustering of Strelka pipelines. Overall, two principal components, explaining a total of 63.15% variance, nicely illustrated the variability of the data.

2.2 Performance Statistics of the Pipelines

Figure 4 shows the number of variants called by each pipeline. As depicted in the figure, the Mutect variant caller called the smallest amount of variants in each pipeline variation, while the Somaticsniper called the highest amount. The type of the aligner also had an observable impact on the total number of variants called. BWA aligner called a higher number of variants than its Bowtie counterpart in all scenarios, with the difference being closer to one thousand variants in the Strelka variant caller. Applying base recalibration decreased the number of variants called in all but one case, the Mutect BWA. The pipeline using Somaticsniper variant caller and BWA aligner, without base recalibration, called the highest amount of 2889 variants, while the Mutect base recalibrated pipeline, using the Bowtie aligner called the least amount of 637 variants.

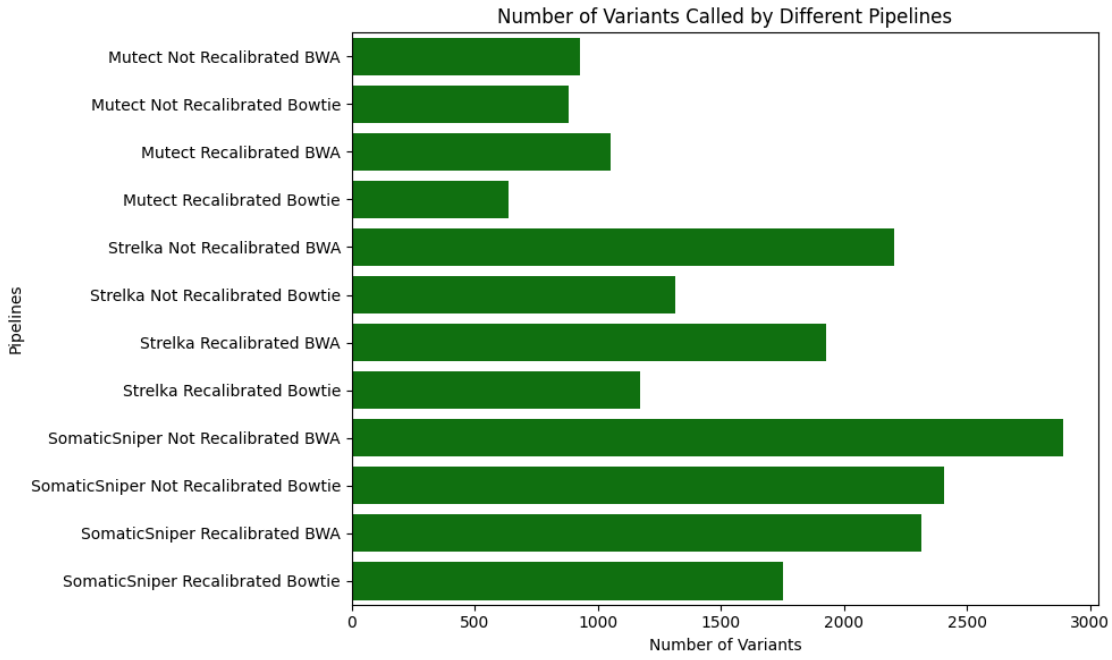

Figure 4: Number of Variants Called by Each Pipeline

I calculated the confusion matrix components for each pipeline by comparing the variants obtained by the pipelines with the ground truth VCF file. Here I did not include the true negative values, as they do not provide valuable information for my analysis. The resulting heatmaps of the confusion matrices are displayed in Figures 5, 6, and 7, for each different variant caller. The first two heatmaps from the left correspond to the pipelines without base recalibration, and the last two correspond to the pipelines with base recalibration. Similarly, the first and the third figures of each variant caller correspond to the BWA mapper, and the other two correspond to the Bowtie mapper.

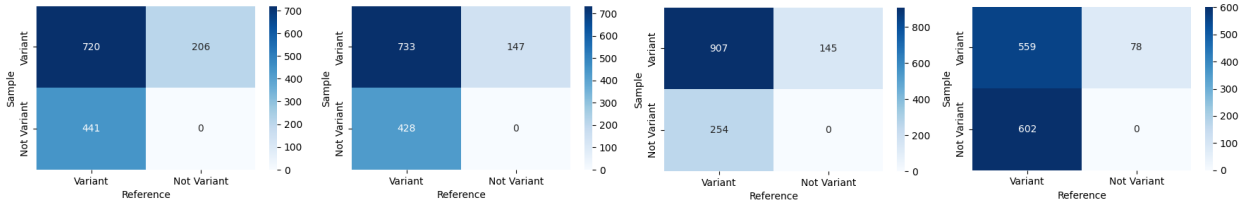

Figure 5: Heat Maps of Mutect Pipelines

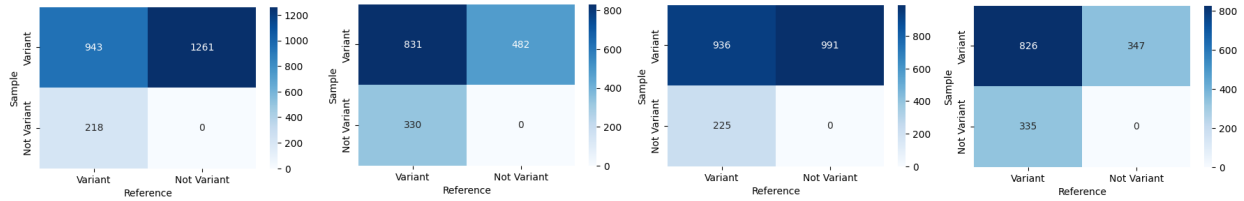

Figure 6: Heat Maps of Strelka Pipelines

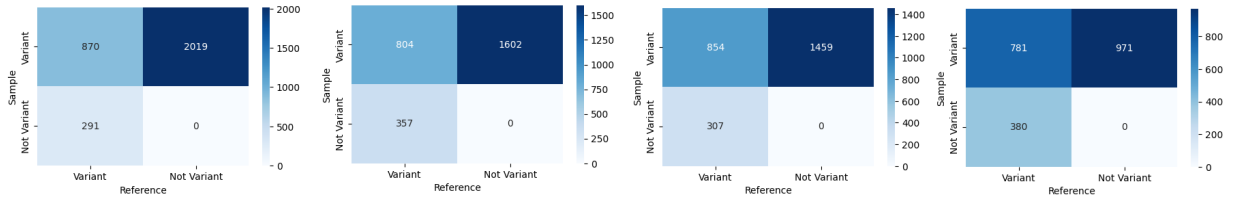

Figure 7: Heat Maps of SomaticSniper Pipelines

It can be observed from the heatmaps that there is high concordance between the pipelines of the same Variant caller, especially the Mutect and SomaticSniper ones. It can be deduced that the Mutect pipelines have the highest accuracy, while the SomaticSniper pipelines have the lowest accuracy, considering the True positive rates.

The accuracy, precision, F1 score, and recall values for all pipelines are calculated using the confusion matrix values. Figure 8 shows the performance values obtained from each pipeline.

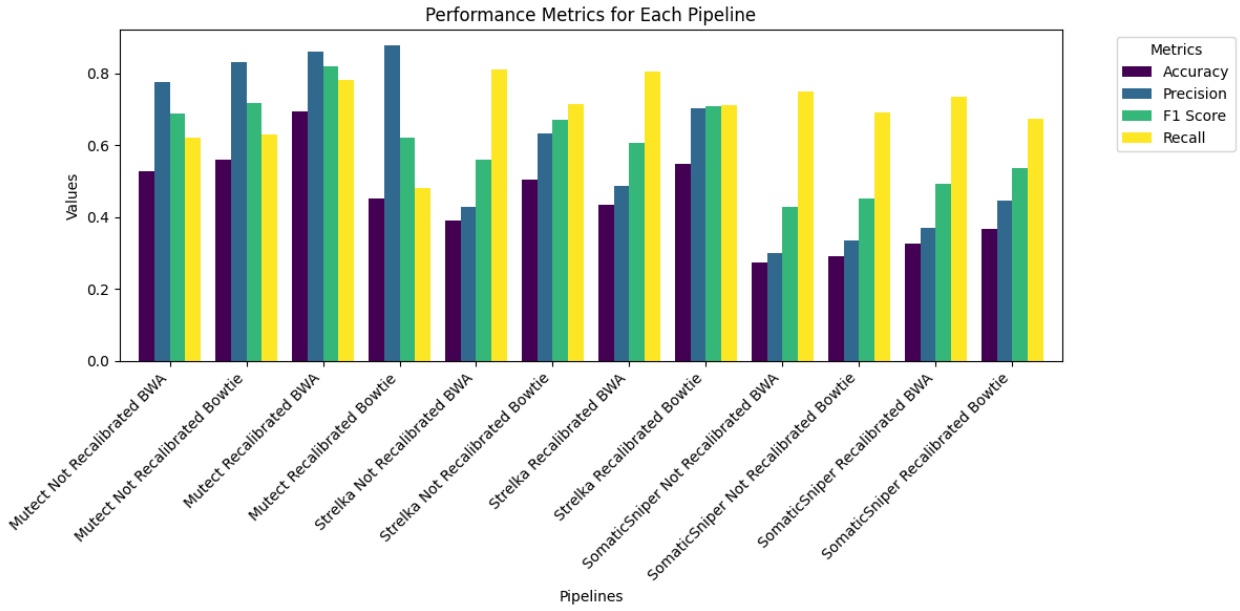

Figure 8: Performance Comparison of Different Pipelines

The pipelines using the Mutect variant caller had the highest accuracy, except the base recalibrated Bowtie pipeline. The Strelka pipelines had lower accuracy than the Mutect ones in general, but they outperformed the SomaticSniper pipelines in all cases. The base recalibrated pipelines performed better than the not recalibrated pipelines, excluding the base recalibrated Bowtie. Again, leaving out the same pipeline, the Bowtie pipelines performed slightly better than the BWA pipelines. Overall, the pipeline using the Mutect variant caller, base recalibration, and BWA variant caller had the highest accuracy score of 0.694, and the SomaticSniper pipeline without base recalibration, with BWA mapper had the lowest accuracy, having the accuracy score of 0.274.

All of these trends in accuracy were paralleled in precision values, with Mutect outperforming Strelka, which outperformed SomaticSniper. This time all the base recalibrated and Bowtie using pipelines outperformed their counterparts. The highest precision value was obtained Using Mutect Variant caller and BWA aligner, using base recalibration, while the lowest precision was obtained with the same pipeline as above. The corresponding highest and lowest values for precision were 0.878 and 0.301.

The F1 scores followed the same trend, aside from the Mutect base recalibrated Bowtie pipeline. The highest score was 0.820 and the lowest score was 0.430.

The recall values for the pipelines were similar. The Strelka variant caller performed slightly better than SomaticSniper, which performed better than Mutect. The recall value for the Mutect and Bowtie without base recalibration significantly underperformed all other pipelines, having the lowest recall value of 0.481, while the highest recall value was obtained with Strelka and BWA without base recalibration, with a value of 0.812.

Based on the results, it can be inferred that the Mutect variant caller yielded better performance results than the Strelka, which outperformed SomaticSniper. Using base recalibration slightly increased the performance in general, and using the Bowtie aligner instead of BWA considerably increased the statistical performance. The low performance of the SomaticSniper variant caller might be attributed to its tendency to call too many variants, increasing false positive variant calls.

One pipeline that deviated from the anticipated performance was the Mutect base recalibrated pipeline using the Bowtie mapper. Because it has a very high precision and a very low recall value, we can deduce that this pipeline behaved overly selectively and yielded a lower accuracy rate and f1 score than expected.

3 Discussion

My analysis of the similarity between different pipelines revealed that the most important metric for the similarity between the pipelines is the variant caller choice. This can be observed from the heat maps in the figures 1 and 2. In addition, the selection of the short read aligner is the second distinguishable metric regarding the similarity between the pipelines. These results were further supported by the clustering of the pipelines in the PCA plot.

My study of the analysis of the pipelines demonstrates that there is a strong correlation between the choice of the variant caller and the accuracy of the pipeline, with Mutect pipelines outperforming the two others, and Strelka outperforming SomaticSniper. Furthermore, applying base recalibration and using the Bowtie aligner instead of BWA increases the performance, considerably. The accuracy ratio of 2.533 between the highest and lowest performing pipelines showcases that the pipeline choice significantly affects the performance. An interesting performance result was of the "Mutect Nobase Bowtie" pipeline, which behaved similarly to the other Mutect counterparts in the heatmap, but has significantly different performance statistics. Because it was rigorously selective, it had a lower count of false positive values and, thus, high precision. However, for the same reason, the pipeline had a higher amount of false negatives, which decreased the accuracy score.

The results of this project might help enlarge the database of the performance and concordance statistics of different variant callers, which, in turn, might contribute to the advancement of NGS techniques in the long run. The project results might also be useful for the testing and improvement of the CoSAP library.

References

[1] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," Arch Dis Child Educ Pract Ed., vol. 98, no. 6, pp. 236-238, Dec. 2013. doi: 10.1136/archdischild-2013-304340. PMID: 23986538; PMCID: PMC3841808.

[2] K. B. Hwang, I. H. Lee, H. Li, et al., "Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings," Sci Rep, vol. 9, p. 3219, 2019. [Online].

[3] VCFtools is a program package designed for working with VCF files [VCFtools (https://vcftools.github.io/)].

[4] CoSAP is a pipeline creation tool for creating NGS pipelines [CoSAP (https://github.com/MBaysanLab/cosap/)].