ISTANBUL TECHNICAL UNIVERSITY

COMPUTER ENGINEERING DEPARTMENT

BLG 348E
INTRODUCTION TO BIOINFORMATICS PROJECT REPORT

GROUP MEMBERS:

150190802 : ÖZCAN ANBALAY

150210921 : DUC QUANG NGUYEN

FALL 2023

Abstract

Advancements in Next-Generation Sequencing (NGS) technologies have ushered genomics into the big data era, revolutionizing computational biology. Nevertheless, accurate variant detection is still an ongoing challenge in the field. This study undertakes a comprehensive evaluation of 12 variant calling pipelines, considering the interplay of two aligners (Bowtie and BWA), three variant callers (Mutect, Strelka, and Somatic Sniper), and the incorporation of base recalibration. Leveraging two sequenced genomes (SRR7890850 and SRR7890851), we compare pipelines against a high-confidence variant call set and examine outcomes with respect to precision, recall, and F1 Score. We observe remarkable divergences in pipeline performance, with Mutect-running pipelines exhibiting robustness and Somatic Sniper-running pipelines demonstrating instability. Aligner choice further influences outcomes, with BWA mapping displaying a more aggressive strategy in variant detection than Bowtie. In-depth analysis unveils associations between read depth and accuracy, thus indicating read depth is a factor contributing to discordant variant calls. The report concludes by acknowledging limitations, urging caution in generalizing findings, and speculating on future research directions.

Contents

1 INTRODUCTION

Advancements in high-throughput sequencing technologies have revolutionized genomics research, enabling the analysis of entire genomes to uncover genetic variations associated with various phenotypes. In this context, the accurate identification of genetic variants through variant calling pipelines is crucial for understanding the molecular basis of diseases, including cancer. The aim of this project is to comprehensively evaluate 12 distinct variant calling (VC) pipelines, considering the interplay of two widely used short-read aligners—Bowtie and BWA—three variant calling methods—Mutect, Strelka, and Somatic Sniper—and the incorporation of base recalibration.

Our investigation focuses on two sequenced genomes, denoted as SRR7890850 and SRR7890851, sourced from the breast cancer cell line HCC1395 and its B lymphocyte-derived normal counterpart HCC1395BL, respectively (Fang et al., 2021). These genomes, obtained in the form of FASTQ files from the United States National Library of Medicine's repository via the SRA Toolkit, represent a comprehensive dataset for our comparative analysis.

The variant calling process adheres to a typical pipeline, encompassing Read Trimming, Read Mapping, Duplicate Removal, Base Recalibration, Variant Calling, and Variant Annotation. While the Read Trimming and Duplicate Removal stages remain constant, we introduce variability by combining different mapping and variant calling methods, resulting in 12 distinct pipelines. The outcome of this process yields 12 Variant Call Format (VCF) files, each corresponding to a unique pipeline configuration.

Subsequently, the generated VCF files undergo an evaluation against a high-confidence variant list for both SRR7890850 and SRR7890851, which serves as the ground truth for our study. The results obtained from this comparative analysis form the basis for interpretation and discussion. This report details our methodology, findings, and insights derived from the comparative evaluation of variant calling pipelines.

The computational aspects of our study were executed on a Windows 11 platform utilizing Docker and Windows Subsystem for Linux (WSL), with 32GB of RAM and more than 200GB of storage capacity. The computational workload required approximately 5 days for completion.

2 RESULTS

2.1 CONVERGENCE OF PIPELINES

The high-confidence VCF file for the SRR7890850 and SRR7890851 genome sets returns 1161 somatic variants, all of which are of single-nucleotide polymorphism (SNP) type. The pipelines that we ran recognized 3242 more variants. In total, 13 pipelines return a total of 4403 variants. The number of variants detected by each pipeline varies widely (Figure 1, Supplementary Table 1). The highest variant returns of a pipeline (Somatic Sniper with BWA mapper, no base recalibration) are around 3.19 times more than its lowest counterpart. On average, Somatic Sniper returns more variants than Strelka, which returns more variants than Mutect. Pipelines running base recalibration return fewer variants than their counterparts not running base recalibration. Pipelines running Bowtie mapper also appear to return fewer variants than pipelines running BWA mapper, though this trend is not observed in Mutect-running pipelines.

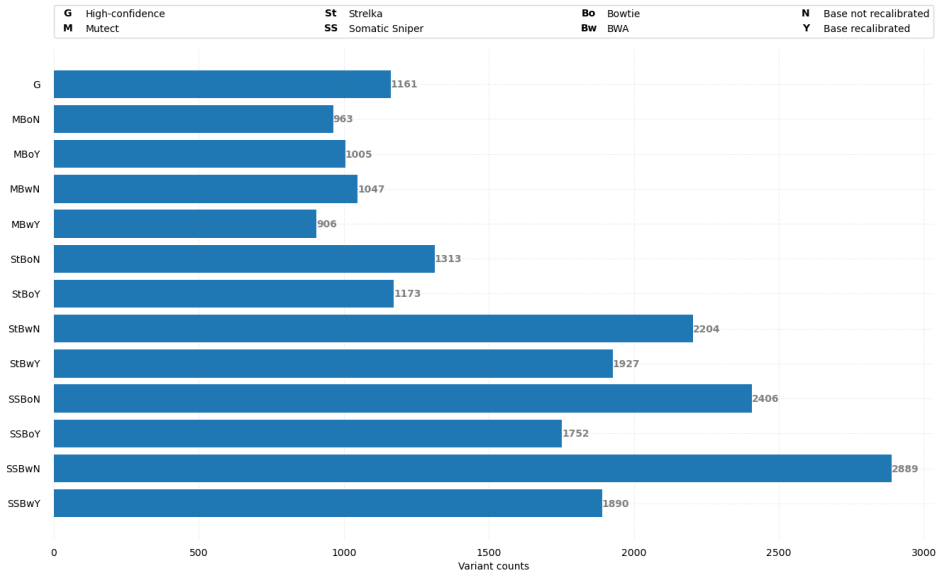

Figure 1: Variant counts of 12 pipelines and the ground truth

To compare the convergence of VC pipelines with respect to which variants are identified, we compute the Jaccard distance between every two pipelines, taking also into account the high-confidence pipeline (Figure 2). Results show that variant call sets depend on the choice of VC algorithms, mapping methods and the incorporation of base recalibration in the pipeline. For each of Mutect, Strelka, and Somatic Sniper algorithms, the resulting variant call sets cluster tightly together, shown by three darker square areas along the heatmap diagonal. Among these three VC algorithms, the variant call sets of

Mutect-running pipelines are closest in distance from the ground truth variant set. In addition, the choice of mappers also affects the variant call sets, particularly in Strelka-running pipelines. Two Strelka-running pipelines with Bowtie mapper identify variants quite different from their counterparts with BWA mapper, depicted by the constrasting colors in the Strelka square region. A similar trend is somewhat observed in Somatic Sniper-running pipelines. On the other hand, the effect of mappers is not seen in Mutect-running pipelines: all of their variant call sets are close in Jaccard distance. Finally, base recalibration also influences the output of VC, though not as significantly as variant callers and mappers. In conclusion, variant call sets appear to be most influenced by the choice of variant callers. The difference between variant call sets returned by Mutect and Somatic Sniper can be as high as 70%.

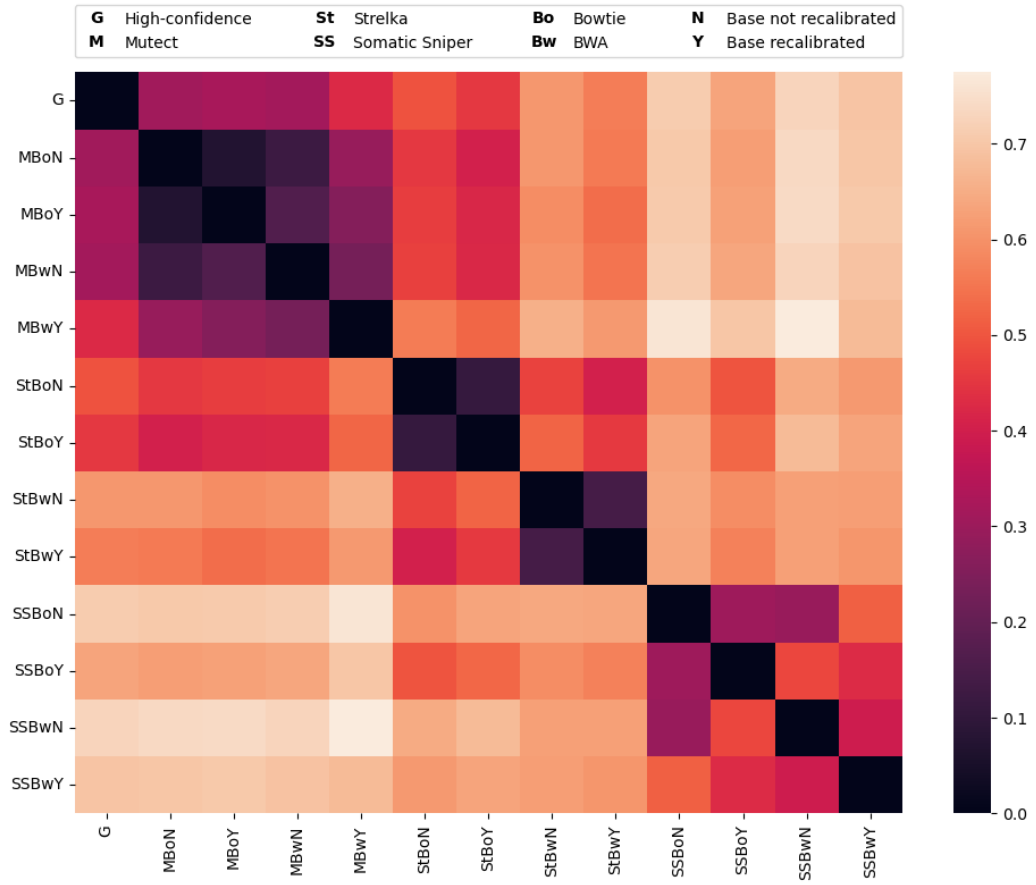

Figure 2: Convergence of VC pipelines with respect to their variant call sets, represented by Jaccard distances between any two pipelines.

We then conduct PCA on the variant call sets detected by 12 pipelines and the high-confidence pipeline to see the relationship among the pipelines in terms of variant detection (Figure 3). The result confirms our observation above. The pipelines cluster according to their variant callers: the Mutect-running pipelines form the tightest cluster, including

also the high-confidence pipeline. The Strelka cluster is divided into two smaller clusters based on mappers, with the two pipelines running Bowtie mapper being much similar to the high-confidence pipeline in variant call sets than the two pipelines running BWA mapper. The Somatic Sniper cluster, on the contrary, appears to be separated more by base recalibration than by the choice of mappers. The variant call sets of two pipelines which perform base recalibration are more similar to each other than to those which do not.

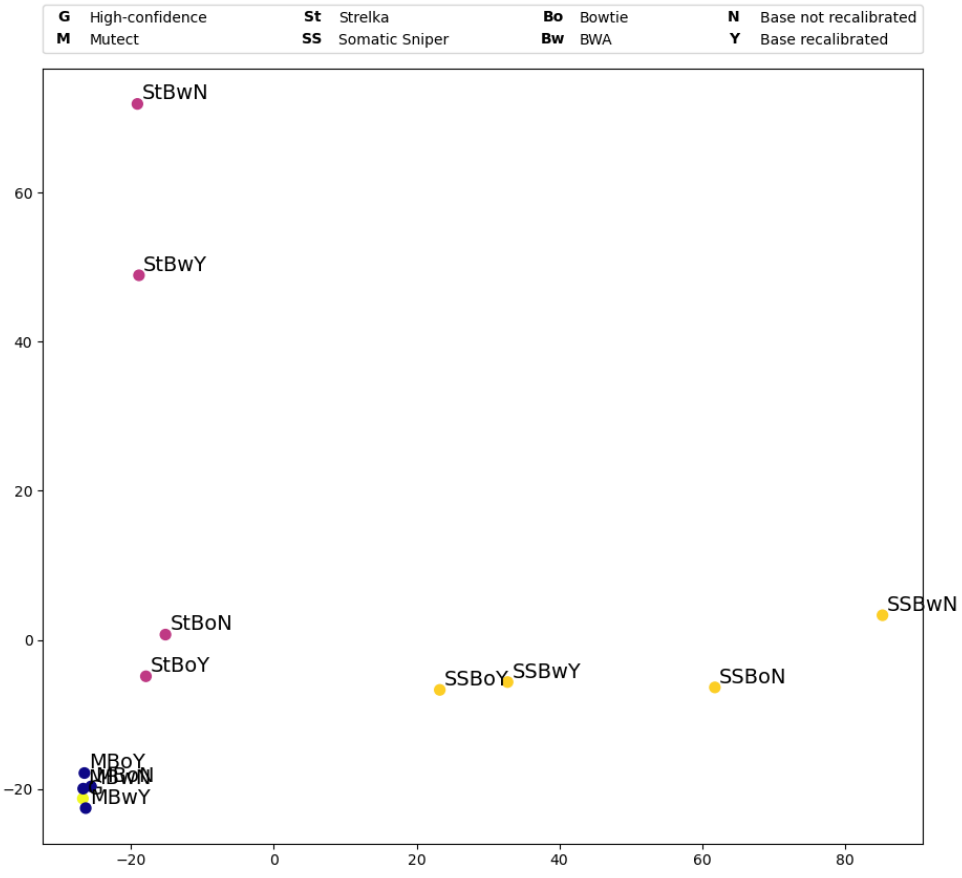

Figure 3: PCA of all pipelines with respect to variant calls

2.2 PRECISION, RECALL, AND F1 SCORE

After determining the convergence of 13 pipelines in terms of their variant call sets, we examine the legitimacy of these variant call sets with respect to our gold standard (high-confidence variant call sets). The legitimacy of variant call sets is measured in three categories: precision, recall, and F1 Score (Figure 4). We decide to exclude accuracy from our evaluation categories due to the fact that genome datasets are naturally unbalanced: the rate of variants in a human genome is only about 0.4% (Genome.gov, n.d.), meaning any VC pipelines can achieve a very high number of true negatives and inflate accuracy as a result.

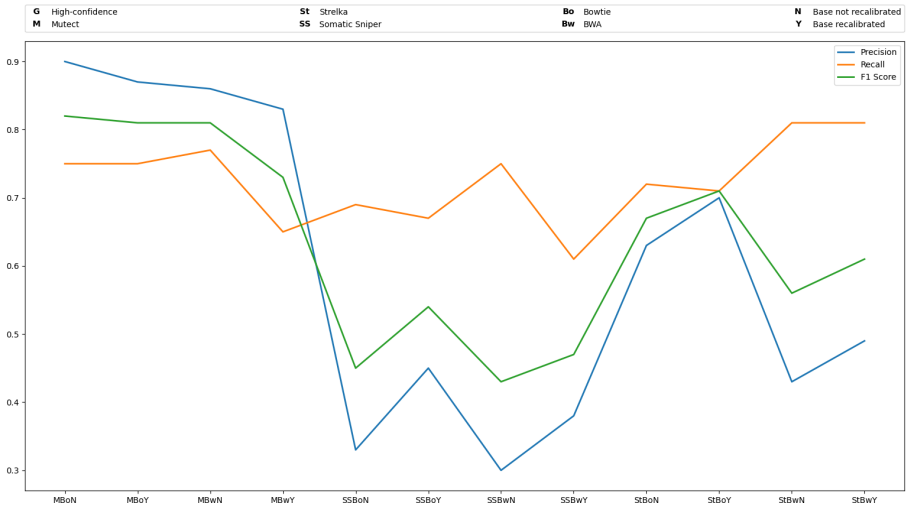

Figure 4: Performance of 12 pipelines in terms of precision, recall, and F1 Score

Once again, there is a clear distinction with respect to variant callers. Among the three variant callers, Somatic Sniper-running pipelines seem the most aggressive, highly favoring recall over precision. Strelka-running pipelines are divided: the two pipelines using Bowtie mapper are moderate, balancing its result of precision and recall, while the two pipelines using BWA mapper are much more aggressive, sacrificing precision to achieve the highest recall among all the pipelines (0.81, Supplementary Table). On the other hand, Mutect-running pipelines seem conservative, having high precision and slightly lower recall.

The choice of either Bowtie or BWA mapper appears to be a trade-off between precision and recall. Pipelines with Bowtie mapper often receive higher precision, but lower recall than pipelines with BWA mapper (Figure 5 (left)). Whether to include base recalibration in the pipelines or not is also a precision-recall trade-off for pipelines running Somatic Sniper caller; including base recalibration improves precision but reduces recall (Figure

5 (right)). Nevertheless, neither mapper nor base recalibration seems to affect F1 Score to any significant amount; F1 Score clearly depends on the variant caller of the running pipelines (Figure 4). Overall, out of three variant callers, Mutect seems to achieve the highest performance.

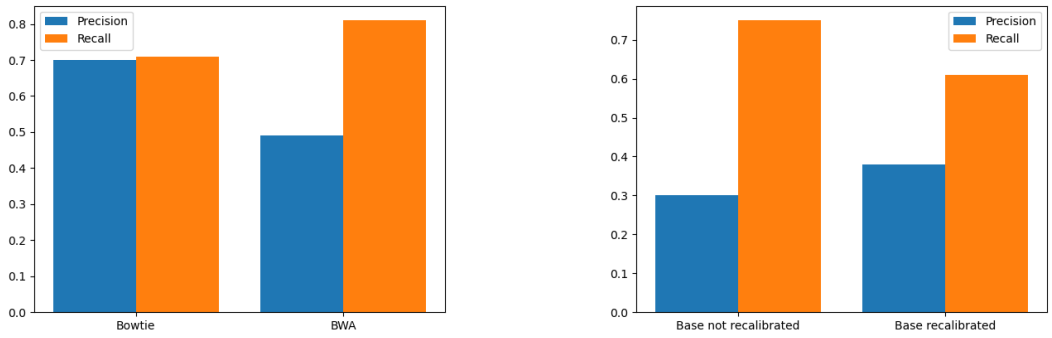

Figure 5: (left) Precision and recall of two base recalibrated Strelka pipelines with Bowtie and BWA mapping - (right) Precision and recall of two BWA-Somatic Sniper pipelines without and with base recalibration

In fact, we can explore the precision - recall trade-off to the difference in the number of variants outputted by each pipeline (Figure 6). In general, as a pipeline outputs more variants, its precision quickly deteriorates as it produces more false positive variants. In exchange for that, it gains a number of new true positives, thus raising its recall value. Nevertheless, the increasing rate of recall seems much slower than the decreasing rate of precision as variant counts rise.

Figure 6: Precision - Recall curve with respect to the number of variant outputs

2.3 FALSE POSITIVES - FALSE NEGATIVES

Call concordance refers to the difference in the number of pipelines calling a variant or variants of interest (Hwang et al., 2019). The assessment of call concordance among the 12 variant calling pipelines reveals intriguing patterns in the identification of variants. Out of the 1161 variants present in the high-confidence variant call sets, over half of these variants are consistently identified by all 12 pipelines, underscoring a robust concordance in their detection capabilities (Figure 7). Conversely, a distinct subset, constituting around 15% of the variants, eludes detection by any of the pipelines, emphasizing the existence of challenging genomic regions or potential limitations in the employed methodologies. This dichotomy in variant detection is noteworthy, illustrating that the majority of variants fall into one of two categories: they are either universally identified across almost all pipelines or remain undetected by any pipeline.

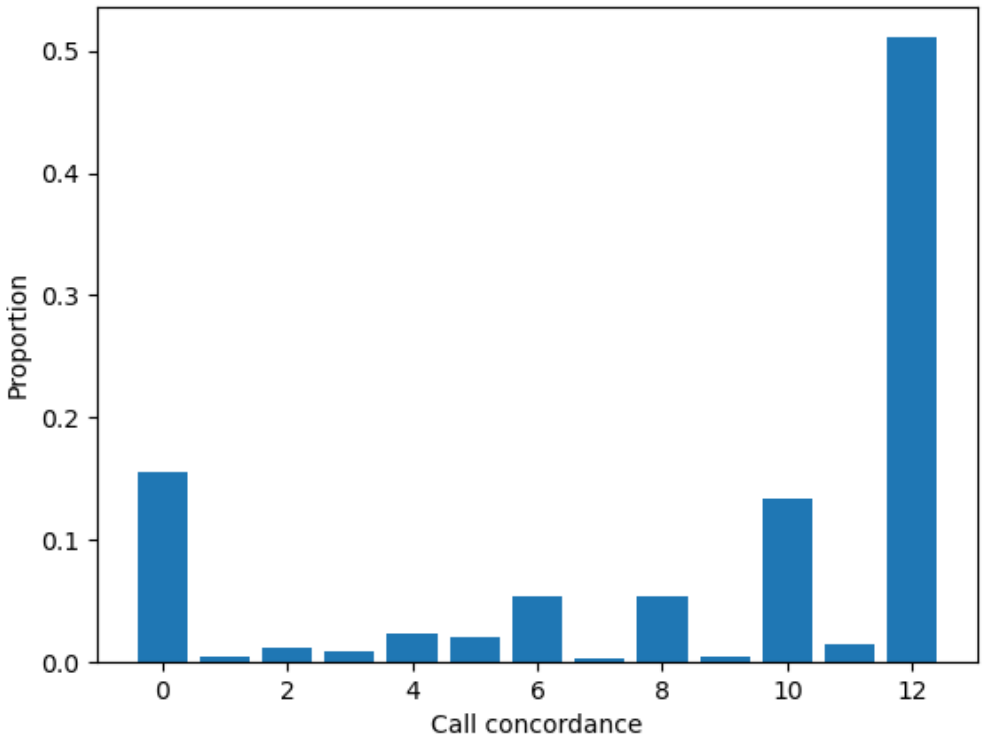

Figure 7: Call concordance of pipelines

To delve deeper into the intricacies of variant calling pipeline performance, we define a subset of variants as false negatives—those missed by at least one pipeline, excluding variants undetected by any pipeline for a more targeted analysis. This subset represents a particularly challenging region for detection and serves as a litmus test for the robustness of each pipeline in challenging genomic contexts. Conducting PCA on this variant subset

yields several following insights (Figure 8). Notably, high-performing Mutect pipelines remain tightly clustered, closely aligning with the high-confidence pipeline. However, an outlier surfaces within the Mutect category, where the Mutect pipeline with BWA alignment and base recalibration deviates from the otherwise cohesive Mutect cluster. In contrast, among the Somatic Sniper pipelines, all remain distanced from the high-confidence pipeline. Of particular interest are two Strelka pipelines with BWA mapping, which form a closely-knit cluster resembling the high-confidence pipeline. These pipelines, known for their aggressive approach prioritizing recall over precision, exhibit resilience in capturing true positives within this challenging region, thus showcasing heightened robustness compared to their counterparts.

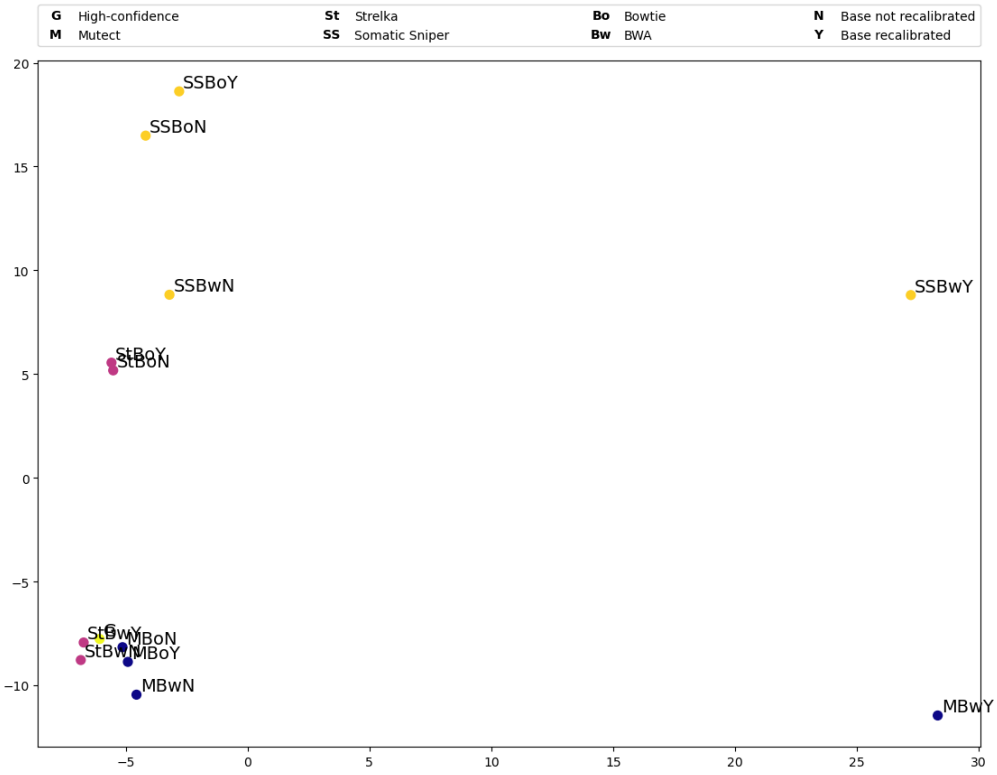

Figure 8: PCA of false negative variant calls

Turning our attention to false positive variants, defined as those erroneously called by at least one variant caller, we conduct a PCA analysis tailored to this specific subset (Figure 9). The outcomes of this analysis are in line with expectations, as all high-performing pipelines form a cohesive cluster closely aligned with the ground truth. The two Strelka-running pipelines with BWA mapping deviate significantly from the ground truth cluster this time. So while maximizing recall, their adoption of an aggressive strategy concurrently results in the inclusion of a substantial number of false positive variants.

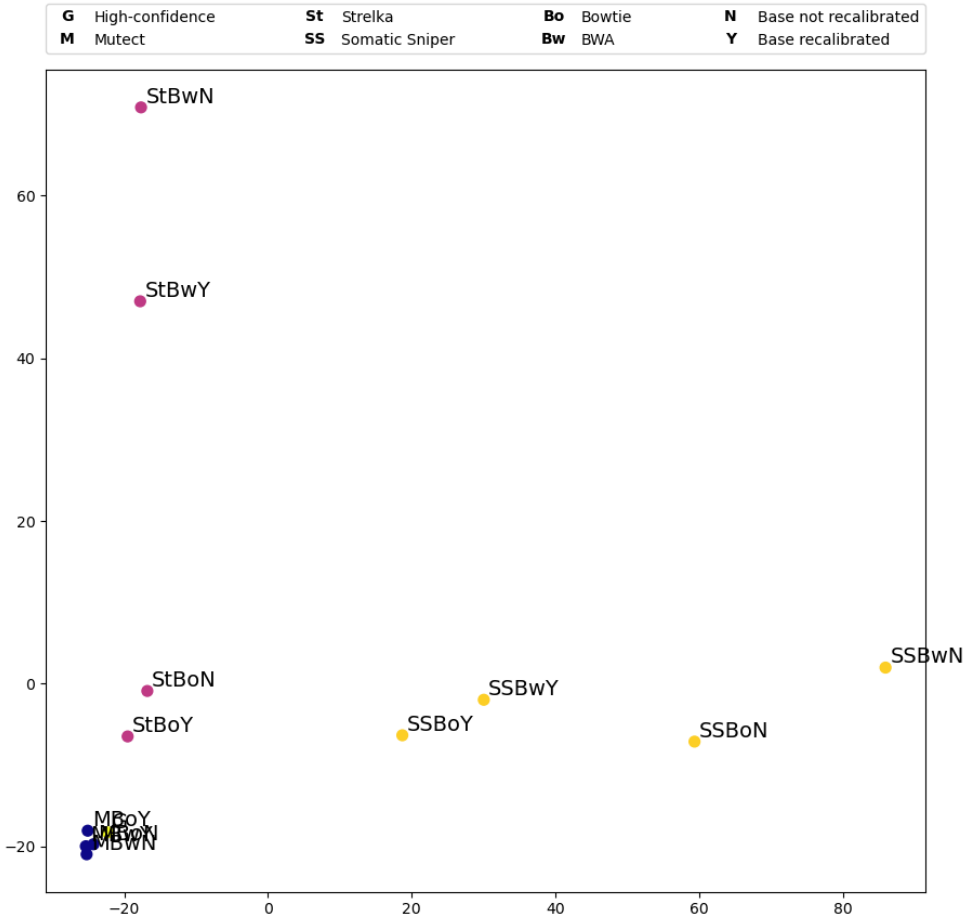

Figure 9: PCA of false positive variant calls

All in all, exploring false negative and false positive variants has shed light on the trade-offs associated with different pipeline configurations and underscores the importance of striking a balance between sensitivity and precision in variant calling analyses.

2.4 READ DEPTH AND DISCORDANT VARIANT CALLS

Discordant variant calls, encompassing both false positive and false negative calls as defined earlier, have been subject to extensive scrutiny in existing literature. Wall et al. (2014) identified read depth and allelic imbalance as frequent contributors to discordant variant calls. Building on this foundation, Hwang et al. (2019) expanded the scope by pinpointing six factors associated with discordance, including read depth, GC content, mapping quality, and, notably, minor allele frequency. Fang et al. (2021) added another layer to our understanding, emphasizing that false positives tend to be concentrated in genomic regions presenting alignment challenges. Specifically, regions characterized by high GC content or low-complexity pose difficulties for current short-read technologies, leading to inadequate coverage. In the context of our project's limited scope, we choose to narrow our focus to explore read depth as a potential factor linked to false positive variant calls. This selective approach allows us to delve into specific aspects within the broader landscape of factors influencing discordant variant calls.

We will examine the read depth dynamics within three pipelines—each utilizing Bowtie for mapping and omitting base recalibration—while differing in their variant callers (Mutect, Strelka, Somatic Sniper). Focusing on both true positive and false positive calls made by these pipelines, we count the occurrences, derive descriptive statistics, and present the findings through side-by-side box and whisker plots (Figure 10).

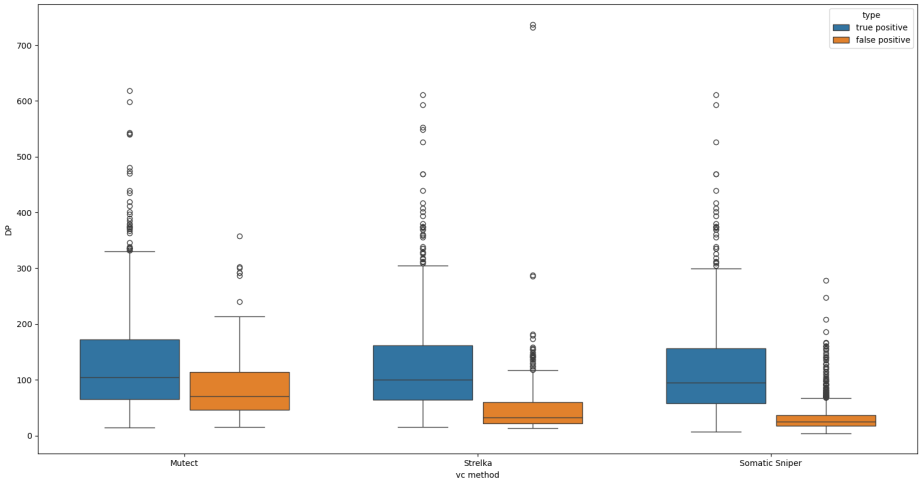

Figure 10: Comparing read depth between true positives and false positives across three pipelines with Bowtie mapping and no base recalibration

Across the three pipelines, the read depth statistics for true positive calls exhibit remarkable consistency. The median read depth hovers around 100, with mean values

ranging from 117 to 130. Additional descriptive statistics, such as the 1st quantile, 3rd quantile, and the presence of outliers, align consistently across these pipelines. In contrast, the read depth statistics for false positive calls showcase greater variability.

In addition to that, on average, false positive variants exhibit significantly lower coverage depth compared to their true positive counterparts (Welch t-test, $p<0.0001$). This observation suggests a positive association between read depth and the accuracy of variant calls—variants with higher coverage are more likely to be true positives.

Nevertheless, further exploration reveals that elevated coverage depth is not always synonymous with call quality. To underscore this point, we turn our attention to the depth of coverage within two pipelines utilizing Somatic Sniper with BWA mapping—one incorporating base recalibration and the other not (Figure 11).

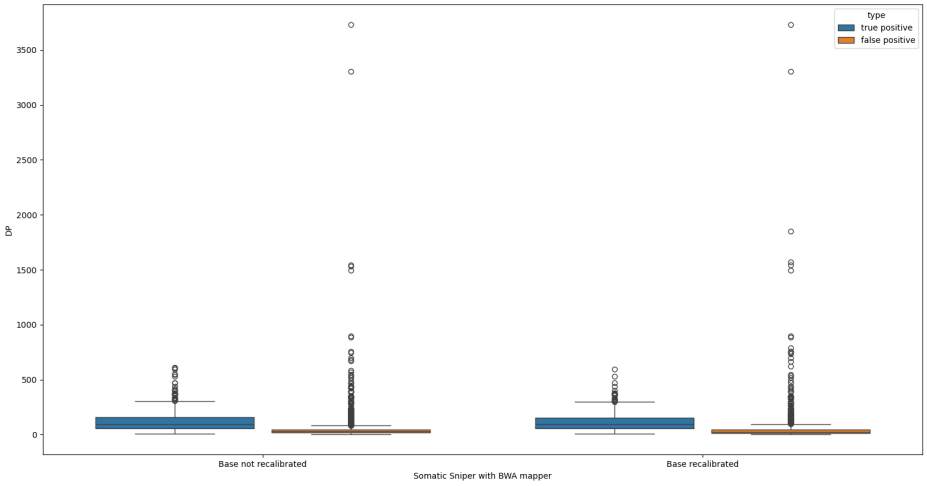

Figure 11: Comparing read depth between true positives and false positives across 2 Somatic Sniper-running pipelines with BWA mapping

This plot remarkably reveals the presence of exceptionally high-covered outliers within false positive variant calls. In both pipelines, a subset of false positive variants exhibits read depths surpassing 1000, with two outliers even exceeding the remarkable threshold of 3000. Intriguingly, these variants, despite their exceptionally high depth coverage, are false variants. This observation raises questions regarding the reliability of variants experiencing exceedingly high read depths and suggests the presence of a potential cap on read depth for optimal variant detection.

3 DISCUSSION

The advent of Next-Generation Sequencing (NGS) has revolutionized genomics, propelling the field into the realm of big data. This technological leap enables the sequencing of genomes on an unprecedented scale, ushering in the era of computational biology. However, the effective application of high-throughput sequencing, particularly in critical areas like clinical settings, hinges on the ability to comprehensively and accurately detect genetic variants—a challenge that the field continues to grapple with. Numerous studies have called into questşon the varied quality and reproducibility of sequencing and variant calling pipelines, influenced by diverse short-read technologies and variant calling algorithms (Fang et al., 2021; Hwang et al., 2019; Barbitoff et al., 2022). In this project, our objective is to conduct a comparative evaluation of 12 variant calling pipelines against a high-confidence variant call set. Through this analysis, we aim to illuminate the strengths and weaknesses inherent in each technology. Additionally, we delve into the association between sequencing read depth and the accuracy of variant detection, shedding light on the nuanced interplay between technology choices and the reliability of variant calls.

Wide variations in output variant calls and in evaluation criteria such as precision, recall, and F1 Score are first observed between variant callers. Mutect exhibits a notably conservative approach, characterized by a restrained variant output, indicating a prioritization of precision to avoid false positives. In contrast, Strelka and Somatic Sniper pursue a more aggressive strategy, aiming for an expansive list of true positives, albeit at the expense of an increased false positive rate. Overall, pipelines employing Mutect emerge with the most robust performance, showcasing a balanced trade-off between precision and recall. Conversely, Somatic Sniper-running pipelines appear to be more susceptible to variability, displaying a greater range of instability. The difference between the most precise pipeline and the least precise one reaches a notable margin of 0.6, emphasizing the considerable diversity in performance among the evaluated pipelines.

A similar observation surfaces as we examine the influence of aligner choice, specifically between Bowtie and BWA, on variant calls. Notably, in pipelines utilizing Strelka or Somatic Sniper, those employing BWA mapping consistently yield a higher number of variants compared to their Bowtie-mapped counterparts. This divergence, coupled with our analysis of precision and recall, also unveils that BWA mapping leans towards a more aggressive strategy, prioritizing higher recall at the expense of diminished precision, while Bowtie mapping can be characterized as more conservative.

However, this observed trend does not hold true for pipelines utilizing Mutect. The discrepancy prompts a hypothesis that either Mutect's inherently conservative approach to variant calling or its high performance with our specific dataset (SRR7890850 and SRR7890851) mitigates the discernible effects of Bowtie and BWA mapping. Further

in-depth analysis is needed to understand the underlying cause here.

We also observe a precision - recall trade-off in terms of the number of variants each pipeline produces. However, the rate of trade-off between precision and recall is not equal: recall rises much more slowly than precision falls as variant counts increase. Further scrutiny of call concordance within variant calling pipelines unravels a distinctive pattern: the majority of variants are either collectively identified by most pipelines or absent across nearly all pipelines. In fact, around 15% of gold standard variants are not detected by any of our 12 pipelines, indicating the presence of hard-to-read genomic regions. This bimodal distribution aligns with findings from a parallel study by Hwang et al. (2019). This pattern presumably explains the unequal rate of trade-off between precision and recall. Even though a pipeline pursues a more aggressive strategy and outputs more variants, most of these variants are likely false positives since undetected gold standard variants lie in more challenging genomic regions.

In exploring potential linkers to the underperformance of variant calling pipelines in discordant calls, our focus narrows onto one key factor: read depth. Our investigation reveals a positive association between a pipeline's read depth and the number of true positive variants it successfully identifies. True positive variants exhibit both higher mean and median read depth compared to false positives. However, the wide variances in read depth, wherein the read depth of true positives often intersects with that of false positives, underscore the complexity of relying solely on read depth as an indicator of accuracy. It becomes evident that multiple factors must be considered to comprehensively gauge the accuracy of variant calls.

Another revelation surfaces as we scrutinize extreme cases: several variant calls exhibiting exceptionally high read depth, exceeding 1000, are identified as false positives. This observation implies a potential adverse impact of extremely high read depth on variant calling accuracy, prompting contemplation on the imposition of a read depth cap for optimal results.

Several limitations inherently impact the scope and generalizability of our project. Firstly, our reliance on a single package variant calling tool, COSAP, restricts our ability to optimize parameters for individual aligners and variant callers. Fine-tuning each mapper or caller independently could potentially yield diverse results and conclusions. Secondly, our analysis lacks hypothesis testing for the observed variations among different pipelines. The absence of statistical validation suggests caution in interpreting the significance of observed differences, urging for further investigations to establish robust conclusions. Furthermore, when interpreting precision and recall, we assume the independence of true positives, false positives, and false negatives, potentially overlooking the interdependence of these variables. Lastly, our project's conclusions are inherently tied

to the specific genome sets (SRR7890850 and SRR7890851) under examination. The lack of broader genomic diversity in our analysis prompts caution in generalizing our findings to other genome sets.

Building on the insights gained from our project, we suggest several directions for future research endeavors. First, a more exhaustive exploration of factors contributing to false positive variants is warranted. Beyond read depth, investigating additional parameters and their impact on false positives could uncover valuable nuances in variant calling pipeline performance. Second, finding the causes of false negative variants is another important direction. Particular attention should be directed towards understanding the subset of high-confidence variants that eluded detection by any pipeline. Unraveling the intricacies surrounding these variants could shed light on gaps in current variant calling methodologies. Third, the potential integration of an ensemble approach, where several mappers and variant callers are employed in a single pipeline, could be explored. This approach aims to harness the individual strengths of each tool while mitigating their respective weaknesses, thus enhancing accuracy and robustness in variant calling pipelines. Finally, a comprehensive study aggregating major papers comparing variant calling pipelines could serve as a cornerstone for future research. Synthesizing existing literature and distilling guidelines for benchmarking and evaluating pipeline performance would provide a valuable resource for researchers navigating the complexities of variant calling analysis. This meta-analysis could also offer a roadmap for standardized methodologies, fostering consistency and comparability across diverse studies in the field.

4 REFERENCES

Barbitoff, Y.A., Abasov, R., Tvorogova, V.E. et al. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics 23, 155 (2022). https://doi.org/10.1186/s12864-022-08365-3

Fang, L.T., Zhu, B., Zhao, Y. et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol 39, 1151–1160 (2021). https://doi.org/10.1038/s41587-021-00993-6

Human genomic variation. Genome.gov. (n.d.). https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation

Hwang, KB., Lee, IH., Li, H. et al. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Sci Rep 9, 3219 (2019). https://doi.org/10.1038/s41598-019-39108-2

Xiao, W., Ren, L., Chen, Z. et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol 39, 1141–1150 (2021). https://doi.org/10.1038/s41587-021-00994-5

5 SUPPLEMENTARY TABLE

		Base cal?	Variant counts	Precision	Recall	F1-score
	Gold standard		1161			
Mutect	Bowtie	No	963	0.9	0.75	0.82
		Yes	1005	0.87	0.75	0.81
	BWA	No	1047	0.86	0.77	0.81
		Yes	906	0.83	0.65	0.73
Somatic Sniper	Bowtie	No	2406	0.33	0.69	0.45
		Yes	1752	0.45	0.67	0.54
	BWA	No	2889	0.3	0.75	0.43
		Yes	1890	0.38	0.61	0.47
Strelka	Bowtie	No	1313	0.63	0.72	0.67
		Yes	1173	0.7	0.71	0.71
	BWA	No	2204	0.43	0.81	0.56
		Yes	1927	0.49	0.81	0.61