BLG555:

Bioinformatics & Computational Biology

Spring 2024

Instructor: Mehmet Baysan

Term Project Report

Halil Berkay Çelik

Pelin Gelmez

Abstract

In this project, we benchmarked various bioinformatics tools (COSAP, Galaxy, and Command Line) to identify Single Nucleotide Polymorphisms (SNPs) and insertions/deletions (indels). By analyzing their performance on a standardized dataset, we assessed precision, recall, F1-score, and accuracy. Our results, including Principal Component Analysis (PCA) and heatmap visualizations, reveal significant performance disparities among tools, emphasizing the need for careful selection based on specific research goals. Execution times were also recorded to evaluate computational efficiency. These findings provide a foundation for future improvements and standardization in genomic variant detection.

Table of Contents

List of Figures

List of Tables

1. Introduction

The accurate identification of genetic variations, such as SNPs and indels, is critical for advancing genomic research and clinical applications[1]. DNA sequencing technologies have evolved, enabling the detailed study of these variants[2]. However, the performance of bioinformatics tools used to detect these variations can vary significantly.

The human genome contains numerous genetic variations and somatic mutations[3]. Several studies showed that somatic mutations are known to cause cancer and rare diseases[4]. Identifying these variants and mutations is a crucial aspect of human genetics[5]. Over the past decade, next-generation sequencing technologies and advanced analysis algorithms have successfully identified genetic variations and somatic mutations[6]. However, there is still a need for standardized variant calling pipelines for clinicians and researchers.

In this project, we evaluated several algorithms and variant calling pipelines via using Next Generation Sequencing (NGS) data. The pipeline creation tools analyzed include COSAP[7] (COmparative Sequencing Analysis Platform), Galaxy, and Command Line. Our benchmarking was performed on computers with the following specifications:

	Pelin Gelmez	Halil Berkay Çelik
PC Model	Acer Aspire A315-42G	Acer Aspire A515-57
OS	Ubuntu 22.04 LTS	Ubuntu 22.04 LTS
CPU	AMD Ryzen 5 3500U, 8 cores	I5-1235U, 12 cores
RAM	16 GB	20 GB
Storage	256 GB	512 GB

Table 1: Features of the computers used in the project

These specifications ensured sufficient computational power for running multiple pipelines. However, for handling large datasets, especially BAM (Binary Alignment Map) files, Pelin faced some serious issues on the storage.

In order to detect SNPs and indels, we needed to download these datasets:

1. SRR7890850 1.fastq.gz & SRR7890850 2.fastq.gz (Tumor)[8]
2. SRR7890851 1.fastq.gz & SRR7890851 2.fastq.gz (Normal)[9]
3. Homo sapiens assembly38.fasta
4. 1000G phase1.snps.high confidence.hg38.vcf.gz
5. Mills and 1000G gold standard.indels.hg38.vcf.gz[10]
6. Bowtie[11] and BWA[12]mapper indexes
7. High confidence VCF files[13]
8. Exome regions BED files[14]

In order to variant calling, we needed:

1. COSAP
2. Galaxy[15]
3. Fastp[16]
4. Bwa-mem[17]
5. Bowtie2[18]
6. Gatk[19]
7. Strelka[20]
8. Varscan somatic[21]
9. Samtools[22]

After creating the environments, we are ready to perform variant calling steps.

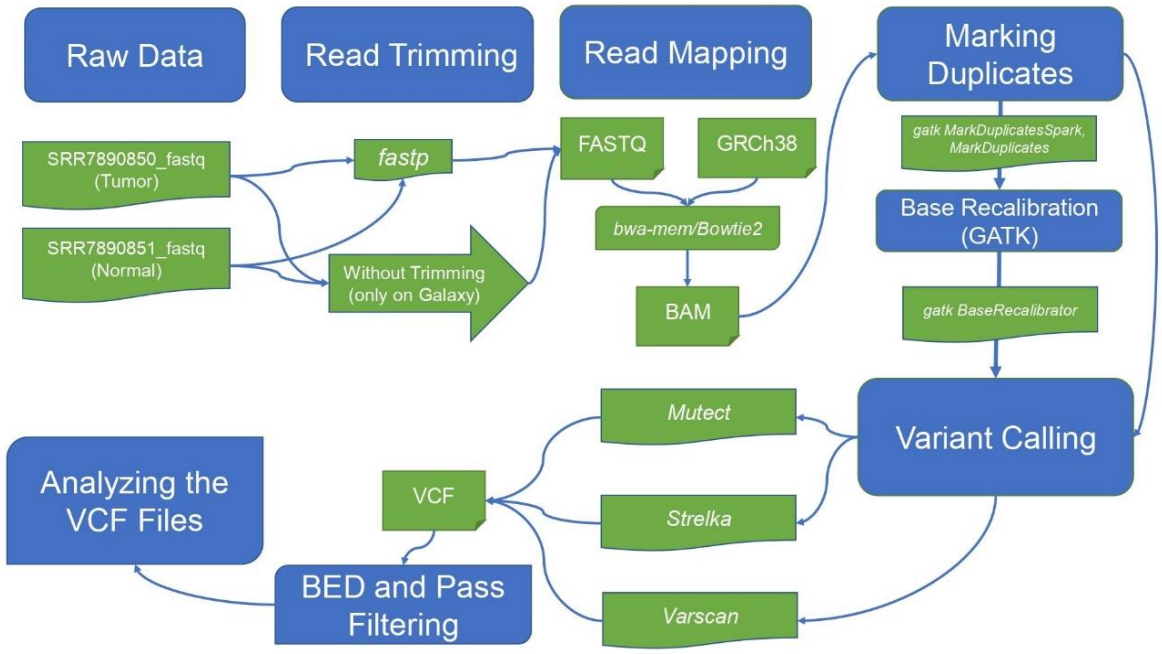

Figure 1: Diagram of the project workflow

2. Results & Discussion

After calling the variants and applying BED filtering to the resulting VCFs, we applied PASS filtering and compared the number of resulting variants in the presence and absence of this filter for all 18 methods (Fig. 2).

This comparison showed that PASS filtering significantly decreased the number of variants for most VCFs. However, as can be seen in the figure, our command line and Galaxy Mutect2 VCFs didn't have PASS information, resulting in no variants in the PASS filtered condition. And some other VCFs had more variants than anticipated when compared to ground truth VCFs and we couldn't pin down the reason of this deviation. Hence, after this step, we decided to continue our analysis by swapping the VCFs that had more than 5000 variants for SNPs and more than 500 for INDELs with the VCFs provided by the course assistant (labeled as Emul et al., 2024).

Between the number of variants obtained with different platforms, aligners or callers no observable pattern was found (Fig. 3).

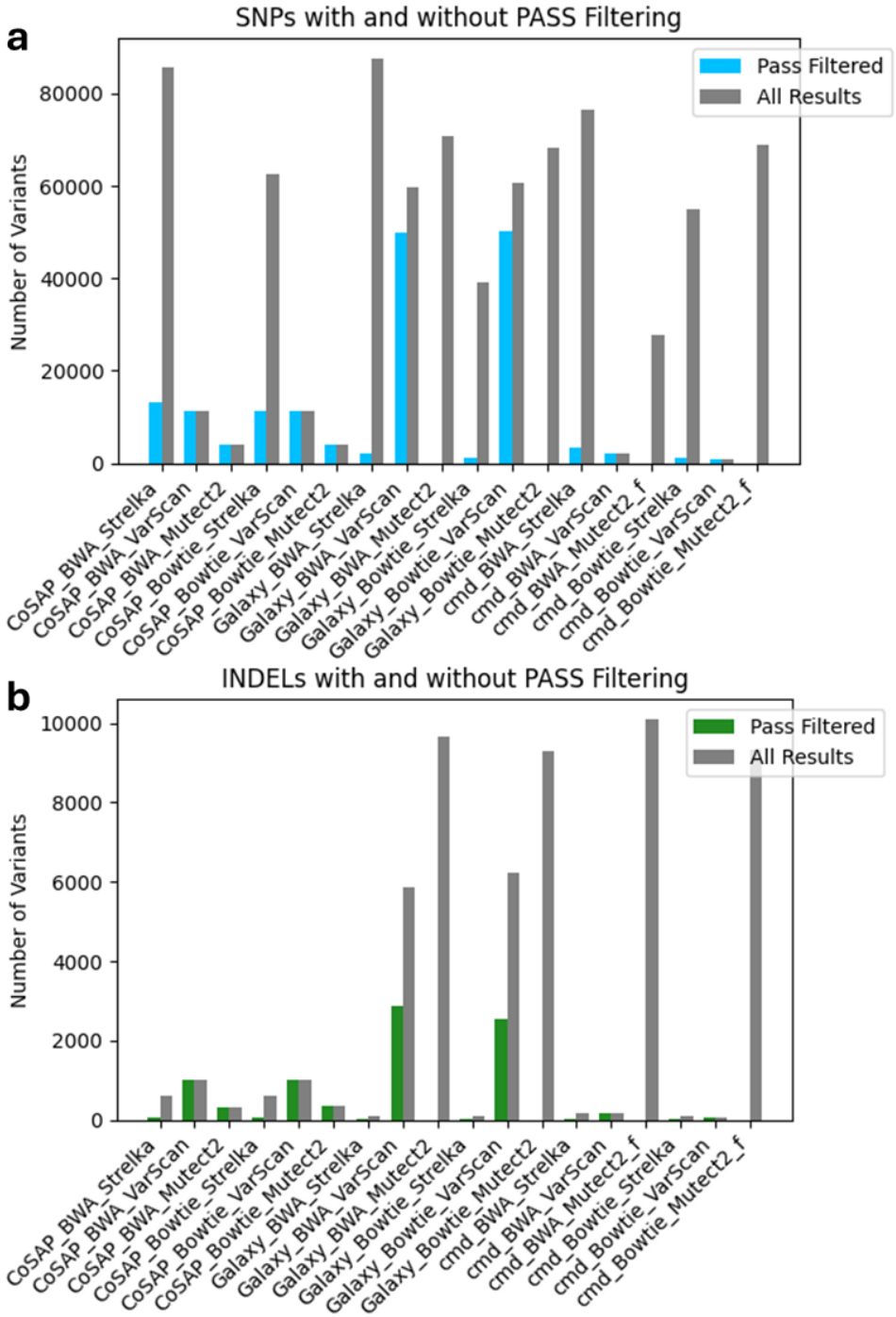

Figure 2: Number of SNPs (a) and INDELs (b) respectively, for each method in the absence (grey) and presence (green) of PASS filtering.

In order to investigate the relation of the parameters, we conducted Venn diagram analysis (Fig. 4 and Fig. 5). Our results showed that CoSAP has the highest number of unique SNPs (Fig. 4a). For the caller algorithms, Mutect2 has the highest detected unique SNPs (Fig. 4e) compared to other algorithms. These enormous gaps between the parameters indicate that there will be distinct results for the CoSAP and Mutect2.

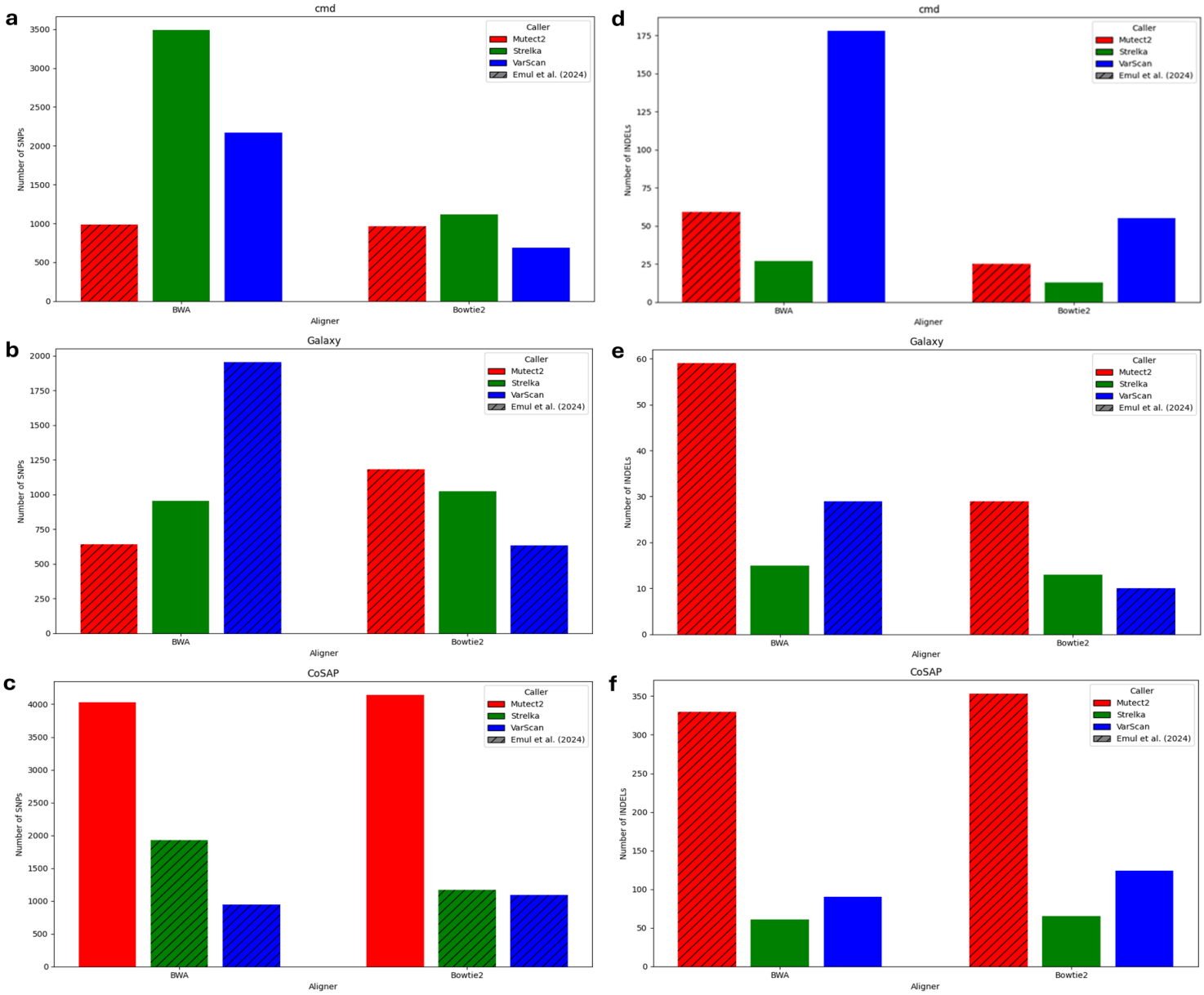

Figure 3: Number of SNPs (a-c) and INDELs (d-f) according to each platform, aligner and caller. Aligners are labeled in the x-axis and callers are shown as bar colors.

Figure 4: Venn diagrams of the unions (a,c,e) and intersections (b,d,f) of the SNPs.

Figure 5*: Venn diagrams of the unions (a,c,e) and intersections (b,d,f) of the INDELs.*

Next, we utilized clustered heatmaps comparing our methods to see if their results were concordant with each other and how much. We also added ground truth VCFs labeled as HC to see with which methods it clustered close with. As shown in Fig. 6a and 6e, SNP variants showed more concordance among methods than INDEL variants. Also, in both variant types, the parameter that tends to cluster together were callers (Fig. 6b and 6f), meaning that mostly each caller found similar variants even when the other parameters changed. We also observed which variants were found by which methods (Fig. 7). In both SNPs and INDELs CoSAP Mutect2 and command line BWA VarScan found a lot of variants that did not overlap with each other or the rest of the methods. The rest of the methods were more concordant with each other, especially in SNP variants.

For the precision-recall and F1 scores of our VCFs, our results indicate distinct F1 evaluation scores across the files, especially INDELs, as shown in Fig. 8 and Fig. 9, respectively. As we are producing the VCF files, we didn't encounter major errors. However, we can't explain these results.

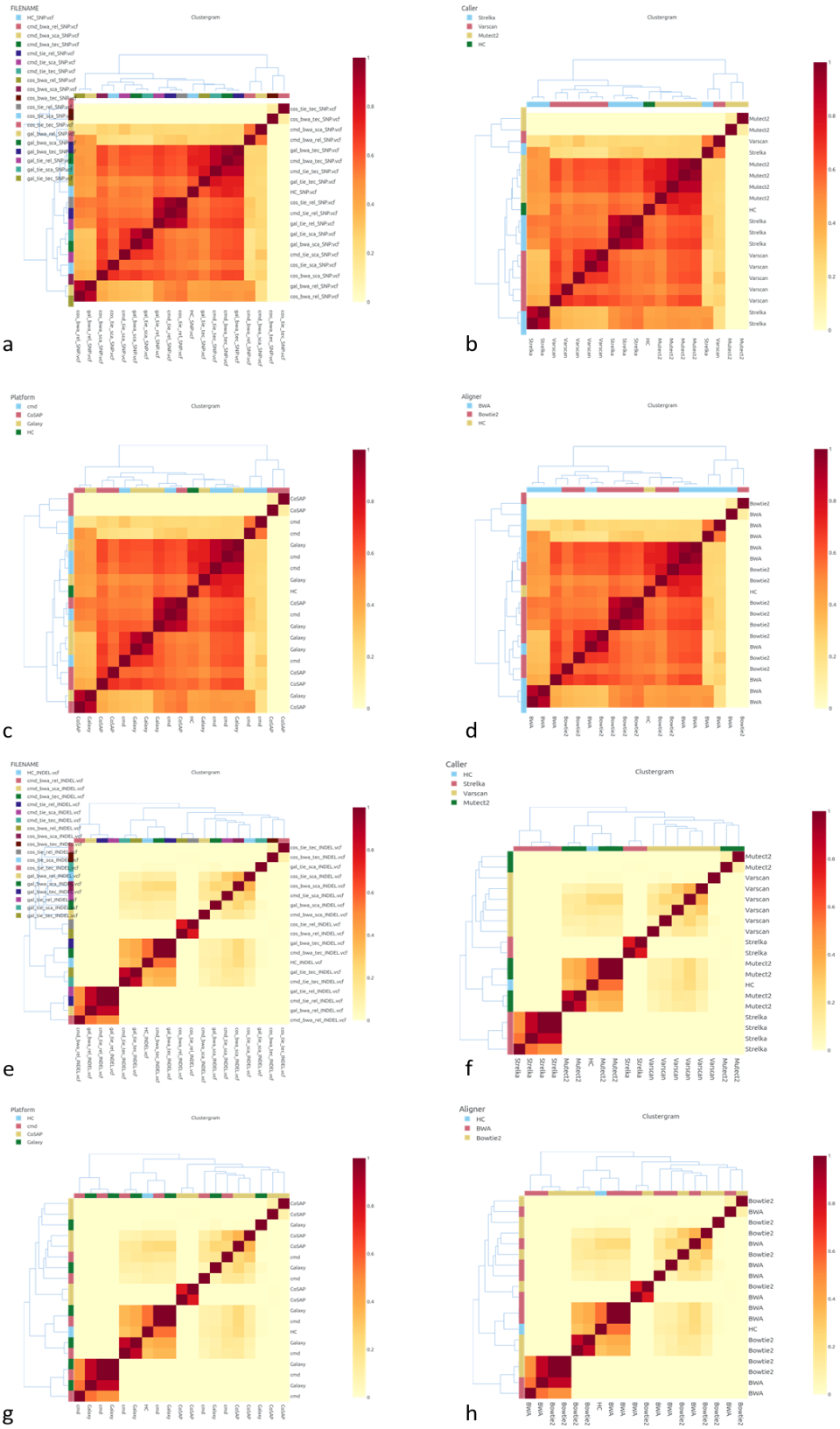

Figure 6: Heatmaps of concordance between methods. Clustered by all parameters; caller, platform and aligner respectively for SNPs (a-d) and INDELs (e-h).

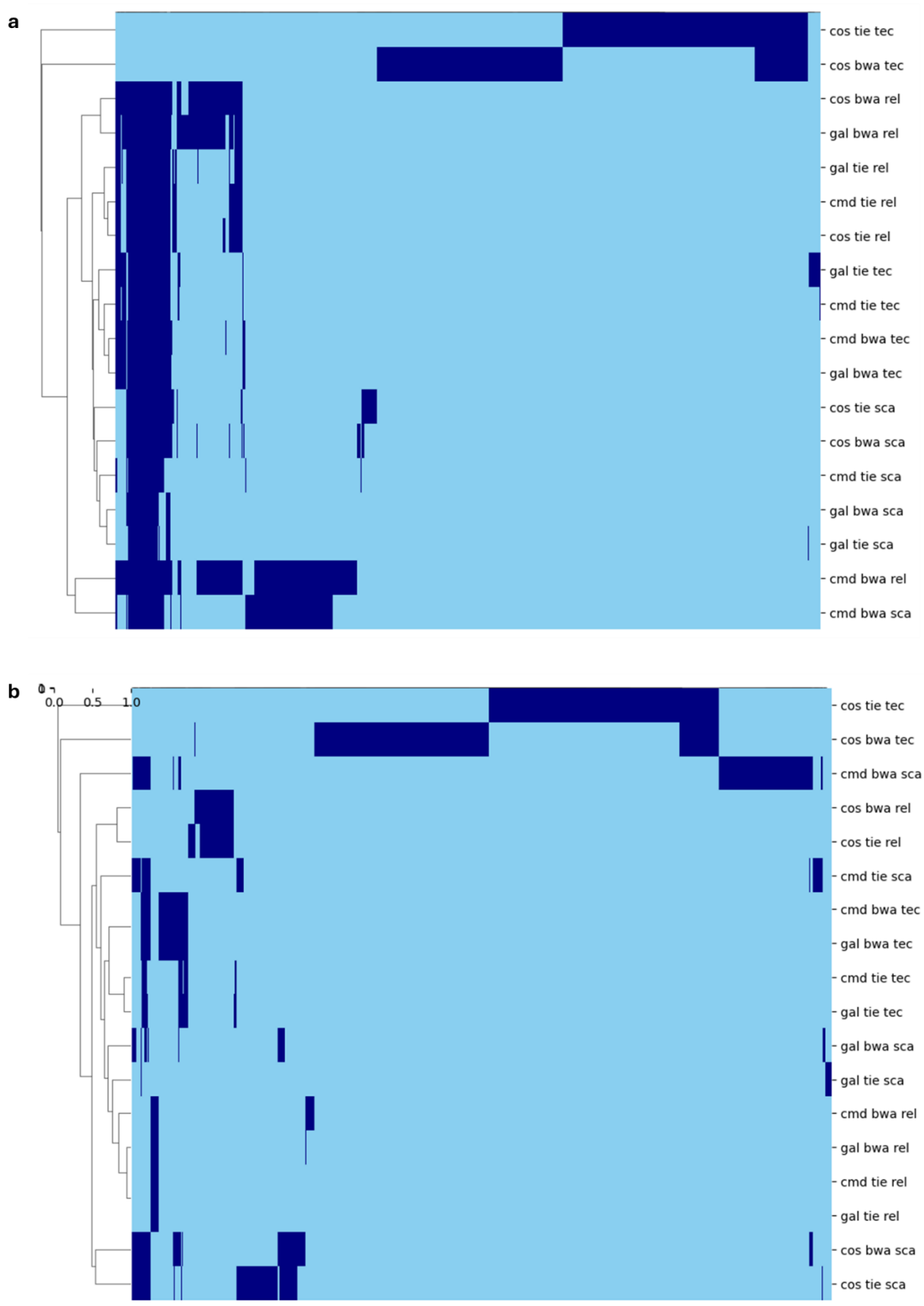

Figure 7: Clustered heatmaps of SNPs (a) and INDELs (b) found by corresponding methods.

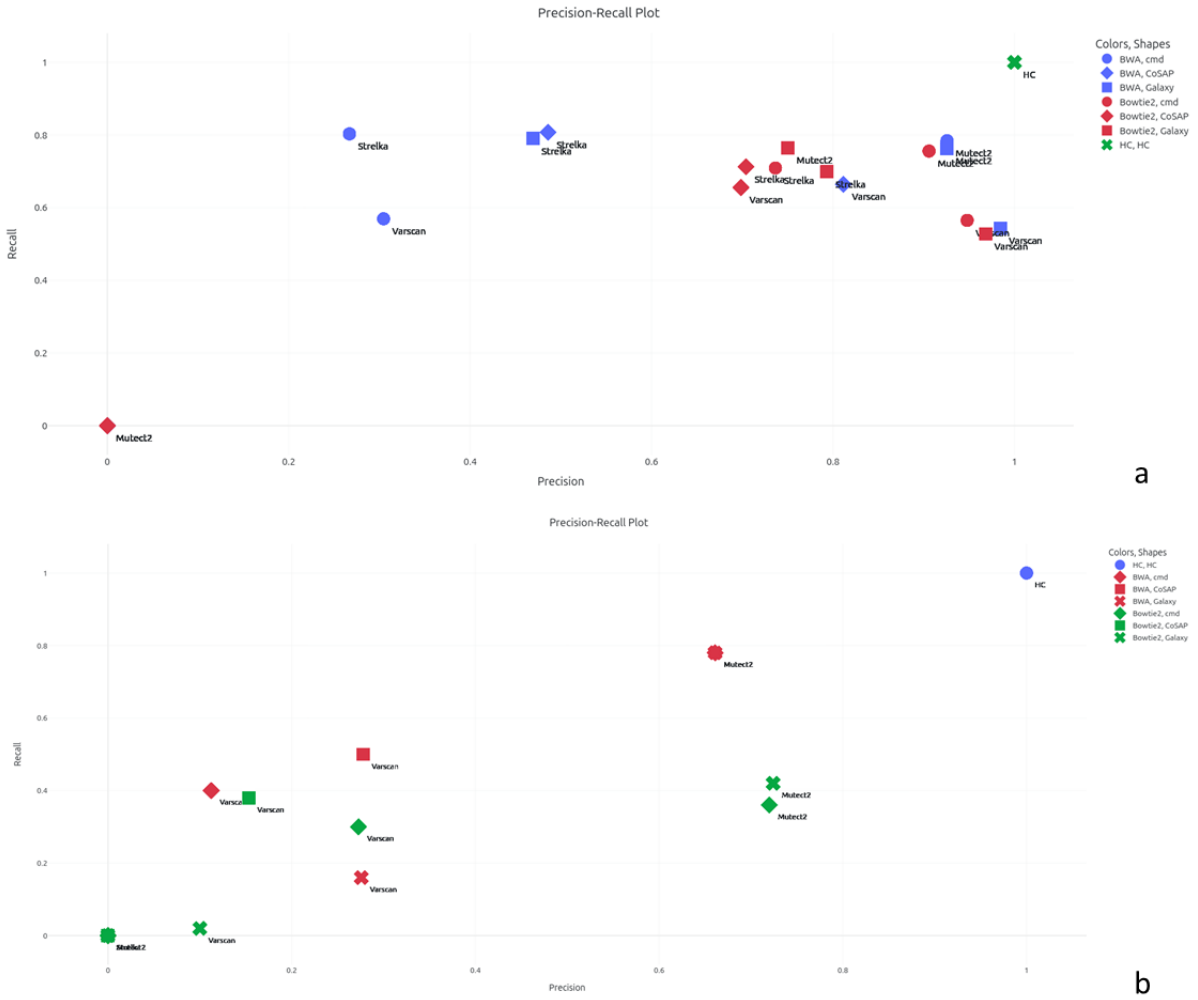

Figure 8: Precision-Recall Plot of SNPs (a) and INDELs (b).

To investigate the correlation between the platforms, callers, and aligners, we applied Principal Component Analysis (PCA). Our findings suggested that detected variants are quite similar between the pipelines, except for CoSAP-BWA-Mutect2 and CoSAP-Bowtie2-Mutect2 (Fig. 10).

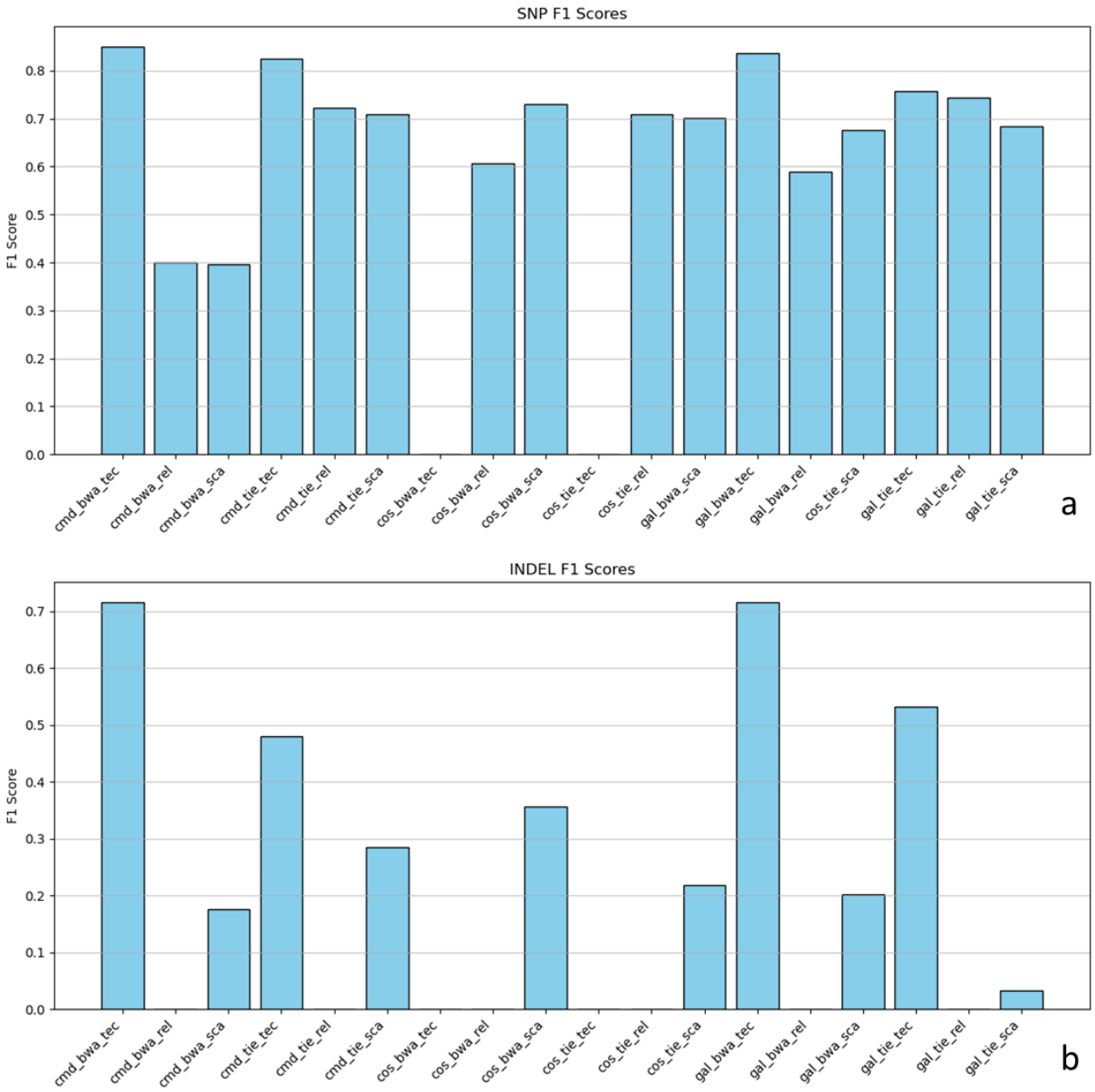

Figure 9: F1 Scores of SNPs (a) and INDELs (b).

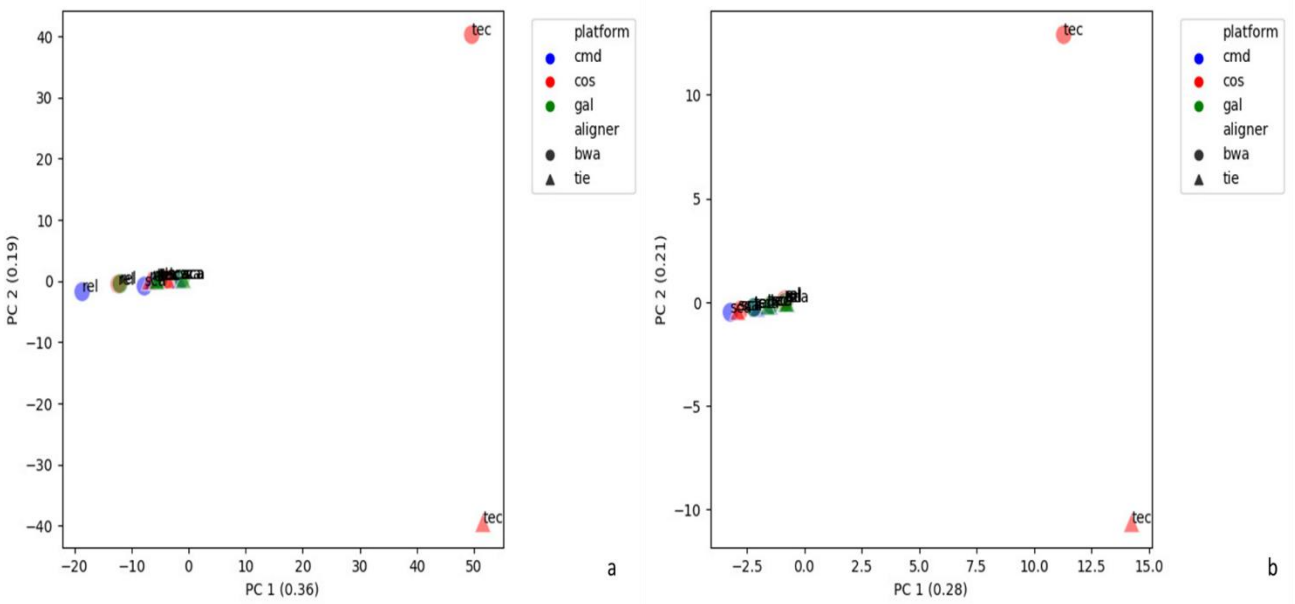

Figure 100: PCA Plot of SNPs (a) and INDELs (b).

Additionally, we conducted bonus research to compare the differences between the with/without Base Recalibration and Trimming steps, just for the Galaxy platform. For the detecting variants, there weren't significant changes for both SNPs and INDELs (data not shown). Clustered heatmap showed Mutect2 and VarScan exhibited dissimilarity compared to others. PCA indicates that same methods showed similar clusters except for Mutect2 (both in SNPs and INDELs). Lastly, for the F1 scores, Mutect2 and VarScan have significantly less scores compared to others for the SNPs. As for the INDELs, the scores are on the floor except BWA-BowTie2/Mutect2 and BWA/VarScan. These results indicate that Base Recalibration and Trimming steps are not so important but if the researchers want to be sure of their study, it is worth applying these steps.

3. Conclusion

Our results didn't show the expected concordance among different parameters and correlation with the verified ground truth VCF. This implies that all of these parameters play a significant role in the resulting variants and that the results will differ from user to user and user error in different steps of variant calling might have a major impact. All of this shows that WGS sequencing analysis might produce meaningless results in the wrong hands.

References

1. Majidian S, Agustinho DP, Chin CS, Sedlazeck FJ, Mahmoud M. Genomic variant benchmark: if you cannot measure it, you cannot improve it. Genome Biol. 2023 Oct 5;24(1):221. doi: 10.1186/s13059-023-03061-1. PMID: 37798733; PMCID: PMC10552390.
2. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, Malonia SK. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 2023 Jul 13;12(7):997. doi: 10.3390/biology12070997. PMID: 37508427; PMCID: PMC10376292.
3. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20:467–84.
4. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, Gotoh K, Ariizumi SI, Wardell CP, Hayami S, Nakamura T, Aikata H, Arihiro K, Boroevich KA, Abe T, Nakano K, Maejima K, Sasaki-Oku A, Ohsawa A, Shibuya T, Nakamura H, Hama N, Hosoda F, Arai Y, Ohashi S, Urushidate T, Nagae G, Yamamoto S, Ueda H, Tatsuno K, Ojima H, Hiraoka N, Okusaka T, Kubo M, Marubashi S, Yamada T, Hirano S, Yamamoto M, Ohdan H, Shimada K, Ishikawa O, Yamaue H, Chayama K, Miyano S, Aburatani H, Shibata T, Nakagawa H. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016;48(5):500–9. https://doi.org/10.1038/ng.3547.
5. Fujimoto, A., Wong, J.H., Yoshii, Y. *et al*. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med* **13**, 65 (2021). https://doi.org/10.1186/s13073-021-00883-1
6. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome sequencing. Cell. 2019;177(1):70–84.
7. "GitHub - MBaysanLab/cosap: Comparative Sequencing Analysis Platform (CoSAP)." Accessed: May 13, 2024. [Online]. Available: https://github.com/MBaysanLab/cosap
8. European Bioinformatics Institute. (n.d.). ENA Browser - SRR7890850. [Online]. Available: https://www.ebi.ac.uk/ena/browser/view/SRR7890850
9. European Bioinformatics Institute. (n.d.). ENA Browser - SRR7890851. [Online]. Available: https://www.ebi.ac.uk/ena/browser/view/SRR7890851
10. Google Cloud Platform. (n.d.). Genomics Public Data: hg38. [Online]. Available: https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0;tab=objects?prefix=&forceOnObjectsSortingFiltering=true

11. Google Drive. (n.d.). Folder: 1Vwtf5Cr1fj0_dO0KFSTp7z2hjCzSrKVy. [Online]. Available: https://drive.google.com/drive/folders/1Vwtf5Cr1fj0_dO0KFSTp7z2hjCzSrKVy
12. Google Drive. (n.d.). Folder: 1rIr0LYk6JkH5uKIp36P3AdOlwNsEd_tr. [Online]. Available: https://drive.google.com/drive/folders/1rIr0LYk6JkH5uKIp36P3AdOlwNsEd_tr
13. NCBI. (n.d.). Somatic Mutation Working Group - Release v1.2.1. [Online]. Available: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/release/v1.2.1/
14. NCBI. (n.d.). Exome Target Regions for Reference Samples. [Online]. Available: https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/technical/reference_genome/Exome_Target_bed/
15. Galaxy Project. (n.d.). Galaxy. [Online]. Available: https://usegalaxy.org/
16. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2020). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 36(18), 548-549. [Online]. Available: https://manpages.debian.org/testing/fastp/fastp.1.en.html
17. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. [Online]. Available: https://bio-bwa.sourceforge.net/bwa.shtml
18. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357-359. [Online]. Available: https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#command-line
19. GATK. (n.d.). GATK | Docs. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us
20. Illumina. (n.d.). Strelka2 User Guide. [Online]. Available: https://github.com/Illumina/strelka/tree/v2.9.x/docs/userGuide#configuration
21. VarScan. (n.d.). Using VarScan 2. [Online]. Available: https://varscan.sourceforge.net/using-varscan.html#v2.3_somatic
22. Twelve years of SAMtools and BCFtools, Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li,*GigaScience*, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008