Benchmarking tools for matched germline-tumor variant calling in circulating tumor DNA using semi-synthetic data

Boudewijn Bartelds S4489543
Supervisors: Prof. dr. Anke van den Berg, Nick Veltmaat
Course: Research Project Biomedical Sciences WBBY902-10

1 Abstract

Next-generation sequencing (NGS) of circulating tumor DNA (ctDNA) within cell-free DNA (cfDNA) shows promise for cancer detection and characterization. However, ctDNA is only a small portion of cfDNA, thereby needing highly sensitive bioinformatics tools known as variant callers to identify mutations. This study benchmarks six variant callers (VarDict, LoFreq, VarScan2, MuTect2, Strelka2 and Octopus) using semi-synthetic NGS data with known spiked variants to create a germline-tumor matched variant calling pipeline. Variant callers were evaluated for sensitivity and precision when detecting single nucleotide variants (SNVs) and insertions/deletions (indels) at varying allele frequencies (VAFs). Results indicated that VarDict demonstrated the highest overall sensitivity, while LoFreq provided the best balance between sensitivity and precision, as reflected in its highest F1 score of 0.81. VarScan2 also performed well in both metrics. Contrarily, Mutect2, Octopus, and Strelka2 showed lower sensitivity and precision. The study concludes that LoFreq and VarDict are particularly effective for detecting low-frequency variants in cfDNA, highlighting their potential for clinical applications for cancer genomics. These results suggest that a sensitive method for matching germline and tumor variant calling is feasible , which could help improve the current ctDNA-based diagnostics.

2 Introduction

Next generation sequencing (NGS) of cell-free DNA (cfDNA) is a promising method for cancer detection and characterization[1] [2]. cfDNA is DNA that is taken up and present in body fluids such as blood. In the case of a tumor being present, circulating tumor DNA (ctDNA) can also be present in the blood and be a part of the cfDNA[3]. Identifying this ctDNA can be troublesome, because the ctDNA is only a small portion of the cfDNA[4]. For this reason a highly sensitive method is needed to identify the ctDNA. In this study we use variant callers to do this. Variant callers are bioinformatic tools that play an important role in genomics research by identifying mutations that may be associated with diseases such as cancer[5]. Amongst other types of genomic variations, they can identify single nucleotide variants (SNVs) from sequencing reads[6][7]. The variant allele frequency (VAF) of these SNVs are also identified by the variant callers. The VAF is the amount of sequence reads that support a specific variant divided by the total reads[8].

For this study variant callers were benchmarked to create a variant calling pipeline that uses germline-tumor matched DNA. The benefits of using germline-tumor matched DNA instead of only tumor DNA is that personal variants and sequence errors are filtered out because the germline-tumor matched are from the same patient[9][10].

Advances in NGS allow for variant calling with a high sensitivity[6]. However most variant callers can handle standard DNA sequencing data, but not all of them are optimized for analyzing cfDNA. The extent to which a variant caller can handle cfDNA data depends on its sensitivity and precision in detecting low-frequency variants, because of the cfDNA samples containing a mixture of normal DNA fragments and ctDNA at varying concentrations. Some variant callers may be specifically de-

signed to analyze cfDNA data, while others may require additional preprocessing or calibration to effectively analyze such samples[11]. In order to create a pipeline with variant callers capable of matched germline-tumor variant calling, it is crucial to benchmark them based on sensitivity and precision and compare these benchmarking results.

The aim of this study was to benchmark six different variant callers using NGS data from cfDNA samples, spiked with known variants in-silico, for a germline-tumor matched variant calling pipeline.

Next generation sequencing of circulating-tumor DNA (ctDNA) is a promising method for cancer identification and characterization. Variant callers are bio-informatics tools that play an important role in genomics research by identifying mutations that may be associated with diseases such as cancer. They can call single nucleotide variants (SNVs) and insertions and deletions (indels) from sequencing reads. These tools identify these genetic variants by analyzing sequencing data to detect differences between a given reference genome and the patient's genome that is sequenced [12]. The subject genome is sequenced from the circulating tumor DNA (ctDNA) that is part of the circulating free DNA (cfDNA). However, a large portion of the cfDNA is secreted from normal cells [13]. For that reason, somatic mutations are hard to distinguish from technical artifacts like PCR and sequencing errors [6]. Therefore, a sensitive detection method is needed to identify somatic DNA mutations in cfDNA.

Advances in targeted next-generation sequencing (NGS) allows for variant calling with a high sensitivity. Variant callers identify mutations by detecting differences between a reference genome and the target genome across all reads for each position in the sequencing data [6]. However most variant callers can handle standard DNA sequencing data, but not all of them are optimized for analyzing cfDNA. The extent to which a variant caller can handle cfDNA data depends on its sensitivity and specificity in detecting low-frequency variants, because of the cfDNA samples containing a mixture of DNA fragments from different tissues at varying concentrations. Some variant callers may be specifically designed to analyze cfDNA data, while others may require additional preprocessing or calibration to effectively analyze such samples [3]. In a previous study, our group developed a pipeline specifically designed to analyse tumor only samples, comparing it to a panel of normals (PoN). To our knowledge, a pipeline comparing ctDNA with matched germline DNA has not yet been established in a similar fashion. In order to create a new pipeline for this purpose, variant callers capable of matched germline-tumor variant calling are necessary. To find fitting variant callers for this new pipeline it is crucial to benchmark different variant callers based on sensitivity, specificity and processing speed. By comparing the benchmark results, a suitable variant caller can be identified that is the most accurate and reliable for detecting the somatic mutations in the new pipeline [14].

For this benchmarking project, the sensitivity of variant callers will be tested on semi-synthetic data. Clinical data is spiked with known variants at different variant allele frequency (VAF) levels. This provides a known reference to perform the benchmark against. Due to the structure of the utilised pipeline, only variant callers with the ability of paired tumor-control data will be considered, and only paired tumor-control testing will be performed. Furthermore, the to-be-used variant callers need to have the ability of identifying SNVs as well as indels. Variant callers also need to be sensitive enough to detect low-frequency variants. Variant callers derived from our previous

pipeline which are known to perform well on cfDNA include MuTect2, VarDict, and LoFreq [15, 16]. Based on previous research, variant callers that fit this criteria (and are not yet present in the pipeline) include Strelka, VarScan2, and Octopus [17, 18, 19, 20, 21].

3 Methods

3.1 Generating dummy variants in real data

cfDNA NGS data from plasma of PTLD patients was aligned to the human genome (hg38) using Burrows-Wheeler Aligner (BWA) [22]. Subsequently, aligned reads were processed to mark duplicate reads using the MarkDuplicates tool from the Picard suite, followed by Base Quality Score Recalibration (BQSR) using the Genome Analysis Toolkit (GATK) [22]. These files were saved in the .bam format. For this study, two .bam files were used of which both contained sequenced data from tumor samples.

Dummy mutations were spiked in the .bam files in selected regions of 1000bp using BamSurgeon, each being at least 150bp apart [23]. In short, mutations were spiked by selecting reads covering a specific nucleotide within the original .bam file and then changing the base of all overlapping reads. These altered reads were placed back into the original .bam file. Mutations were spiked in the following VAFs: 0.0005, 0.001, 0.005, 0.01, 0.02 and 0.05. Mutations with each VAF were spiked 42 times for a total of 252 spiked mutations.

The genomic coordinates and VAF level of the spiked mutations were saved to a .tsv file for later use, referred to as the list or set of spiked mutations.

3.2 Benchmarking prerequisites

In order to select the optimal tools for a tumor-normal matched variant calling pipeline, Mutect2, VarDict, LoFreq, Strelka, VarScan2, and Octopus were used.

For the benchmarking, one of the spiked .bam files was regarded as the normal file and the other as the tumor file. Furthermore, a reference genome (hg38) .fasta file was supplied. Lastly, a .bed file was supplied containing the spiked regions in order to limit the considered regions to just the spiked regions while running the variant calling tools. All results were output as a .vcf file, except for VarScan2 which was output as a .snp file.

3.3 Variant calling

The semi-synthetic data containing dummy mutations was analyzed by all variant callers, mostly using default settings. Exceptions include the minimum allele frequency which was set to 0.0001, the maximum variant size which was set to 10 and the minimum supporting reads which was set to 1. Furthermore, if applicable, annotation options used to write the VAF into the output file were enabled and default filters were disabled.

3.4 Statistics

Statistical analysis was performed per variant caller. The list of spiked mutations was used as reference to identify true positives (TP) and false negatives (FN). Spiked mutations that were identified by a variant caller were marked as TP. Spiked mutations that were not identified were marked as FN.

Per variant caller, the total number of identified spiked mutations (recall) was measured by comparing the genomic coordinates of all identified mutations with the genomic coordinates in the list of spiked mutations. The total number of matching genomic coordinates was used as the recall value. This same method was used to measure the recall per VAF per variant caller. Furthermore, uniquely called spiked variants were measured by comparing the recalled variants of a specific caller to the results of all other variant callers. Every recalled variant that had a match with another variant caller was subsequently discarded, resulting in a list of called variants specific to each variant caller.

The absolute percentual difference per VAF was calculated per variant caller in order to assert that the spiked mutations' VAFs were identified at a comparable VAF level. This was calculated as $\frac{|i-j|}{(i+j)/2} * 100.0$, in which i is the spiked VAF and j is the identified VAF.

As it is hard to obtain false positive (FP) values from tumor samples, likely containing somatic mutations along with the spiked dummy mutations, a consensus calling approach was used. Non-spiked identified mutations were marked as false positive if the mutation was called by only one variant caller. Non-spiked mutations called by more than one variant caller were considered as a somatic mutation. Somatic mutations were subsequently not counted as a FP in further analyses. Furthermore, true negatives (TN) were not taken into consideration for the analyses as this would include every single base pair that wasn't mutated.

Per variant caller, the sensitivity and precision were calculated. Sensitivity was calculated as: $Sensitivity = \frac{TP}{TP+FN}$. Precision was calculated as $Precision = \frac{TP}{TP+FP}$. Sensitivity values were calculated overall and per VAF in order to identify differences and overlap in sensitivity at specific VAF levels between variant callers. Overall sensitivity was calculated as $\frac{TP}{252}$ and sensitivity per VAF was calculated as $\frac{TP}{42}$. The precision could only be calculated as overall precision, as the FP mutations could not be categorized per VAF.

In order to symmetrically represent the precision and sensitivity metrics in one value, the harmonic mean (F1) value was calculated per variant caller. The F1 value was calculated as $F1 = 2 * \frac{Precision*Sensitivity}{Precision+Sensitivity}$.

3.4.1 Visualization and analysis

Data visualization and statistical calculations were performed using Python 3.9.5 using custom plotting and statistical analyses scripts. Packages used were Matplotlib [**Hunter:2007**], seaborn [**seaborn**] and scikit-learn [**scikit-learn**].

4 Results

4.1 Metric Selection

To select the tools we used a metric with relevant information regarding their capabilities (Figure 1 .

Table 1: An overview of variant caller information

Variant caller	Release date	Last updated	Citations	Hits with 'cfDNA'	Main language	Version
VarDict	2016	2020	774	173	Perl	1.8.2
LoFreq	2012	2020	1197	68	C	2.1.5
VarScan2	2010	2016	4846	343	Java	2.4.4
Mutect2	2018	2024	5444	383	Java	2.2
Octopus	2015	2021	73	0	C++	0.7.4
Strelka2	2016	2018	1041	80	C++	2.9.9

4.2 Sensititvity

4.2.1 Overall Sensitivity

To determine the sensitivity of the variant callers, the overall recall and thus the true positives need to be determined first. The spiked mutations that were called per variant caller are shown below in Figure 1. VarDict has the highest recall which makes it the most sensitive. LoFreq follows with the second highest recall. VarScan2 also performed fairly well, however Mutect2 did not, showing a low sensitivity. Octopus and Strelka did not perform well at all, having almost no recall. When looking at the venn diagram 35, also shown below in Figure 1, spiked mutations can be seen that are called by VarDict and no other variant caller. VarDict is the most sensitive and calls almost all spiked mutations (235/250).

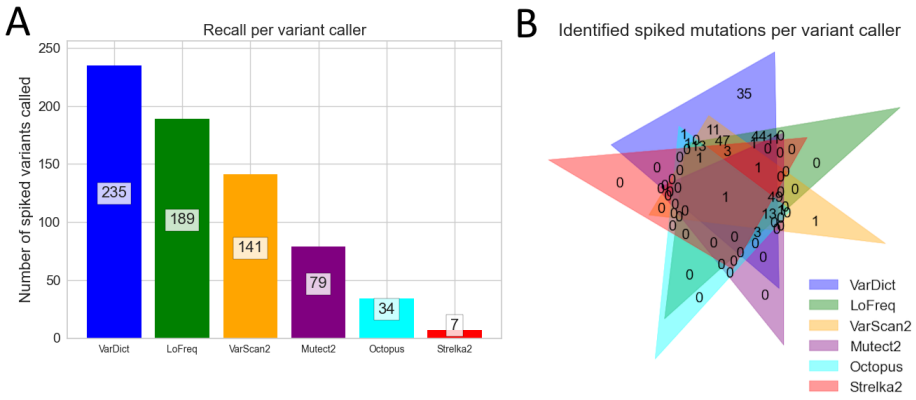

Figure 1: A) Bar graph showing the recall per variant caller. VarDict is shown in blue, LoFreq in green, VarScan2 in yellow, Mutect2 in purple, Octopus in cyan and Strelka2 in red. **B)** Venn diagram showing the overlap of identified spiked mutations per variant caller. The colors of the planes represent the six different variant callers, as depicted in the bottom right. The values of the non-overlapping planes represent the amount of spiked mutations identified by their representative variant caller. Values of overlapping planes indicate the amount of spiked mutations identified by that combination of variant callers.

4.2.2 VAF specific Sensitivity

The mutations were spiked at known VAFs, these specific VAFs are then called by the variant callers. To see which variant caller worked best at which VAF, a bar graph was made to illustrate this. as shown below in Figure 2. It is shown that VarDict is sensitive across all VAFs, as for the other variant callers this is not the case. A rising line can be observed from the low VAFs to the high VAFs. This is because the lower a VAF is the less chance there is for a mutation to be found in the reads. As for Octopus and Strelka2, the results are disappointing and there is a very low sensitvity observed.

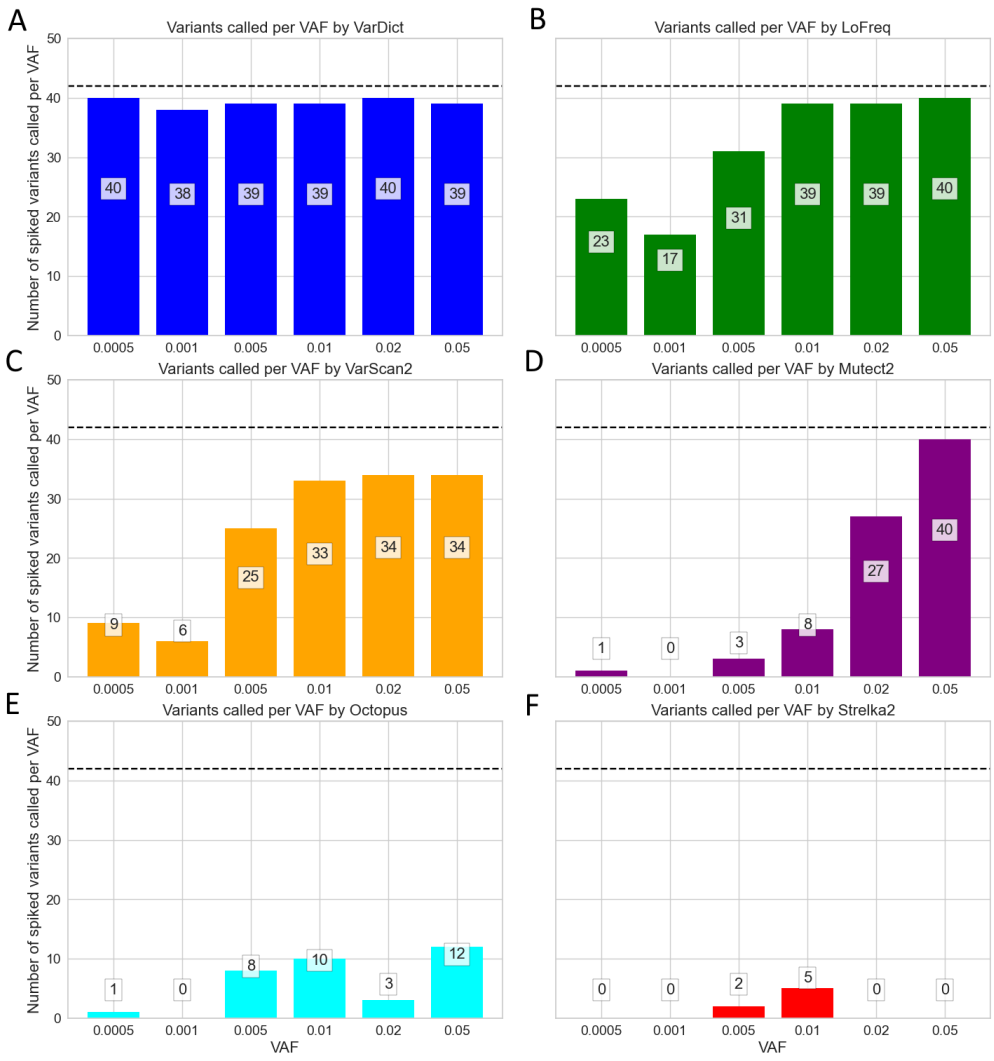

Figure 2: Bar graph of the number of spiked mutations found per VAF for **A)** VarDict, **B)** LoFreq, **C)** VarScan2, **D)** Mutect2, **E)** Octopus, **F)** Strelka2. The black striped line is set at y = 42, indicating the maximum amount of spiked mutations per VAF.

Venn diagrams were also made to illustrate the overlapping called variants at specific VAFs, As shown below in Figure 3. VarDict shows unique calls for lowest VAFs. Meaning that only VarDict can call some of the variants that are spikedin at the lowest VAFs.

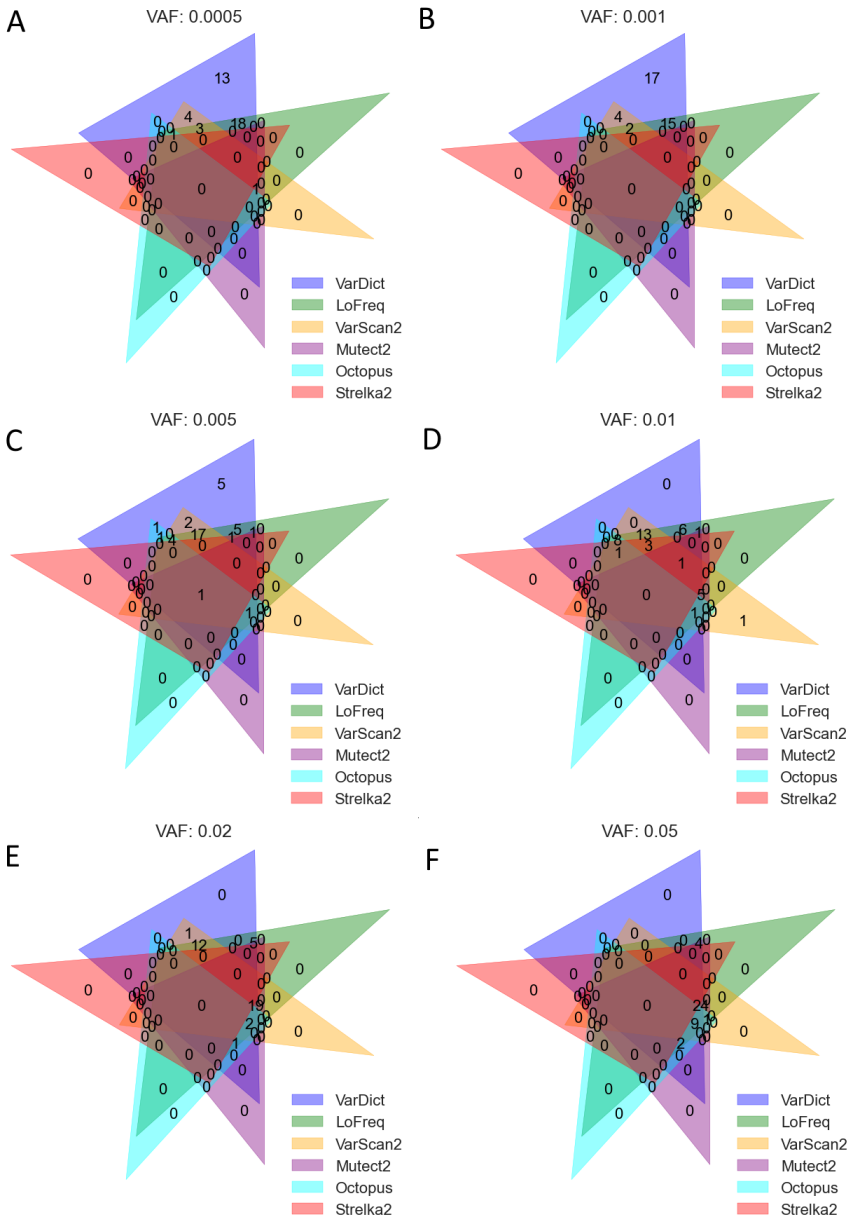

Figure 3: Venn diagram showing the overlap of identified spiked mutations per variant caller with a VAF of **A)** 0.0005, **B)** 0.001, **C)** 0.005, **D)** 0.01, **E)** 0.02, **F)** 0.05. The colors of the planes represent the six different variant callers, as depicted in the bottom right. The values of the non-overlapping planes represent the amount of spiked mutations identified by their representative variant caller. Values of overlapping planes indicate the amount of spiked mutations identified by that combination of variant callers

To ensure the spiked mutations were identified at the specific VAF levels, the absolute percentage difference between the spiked VAF level and the identified VAF level was calculated and visualised in violin plots, as shown below in Figure 4. As the VAF levels go higher, the measured VAF levels are more similar to the known VAF levels at which the variants were spiked in, because the absolute percentual difference drops. This is true for VarDict, LoFreq and VarScan2 however, for Mutect2, Octopus and Strelka2 the results were contradictory due to their low sample size.

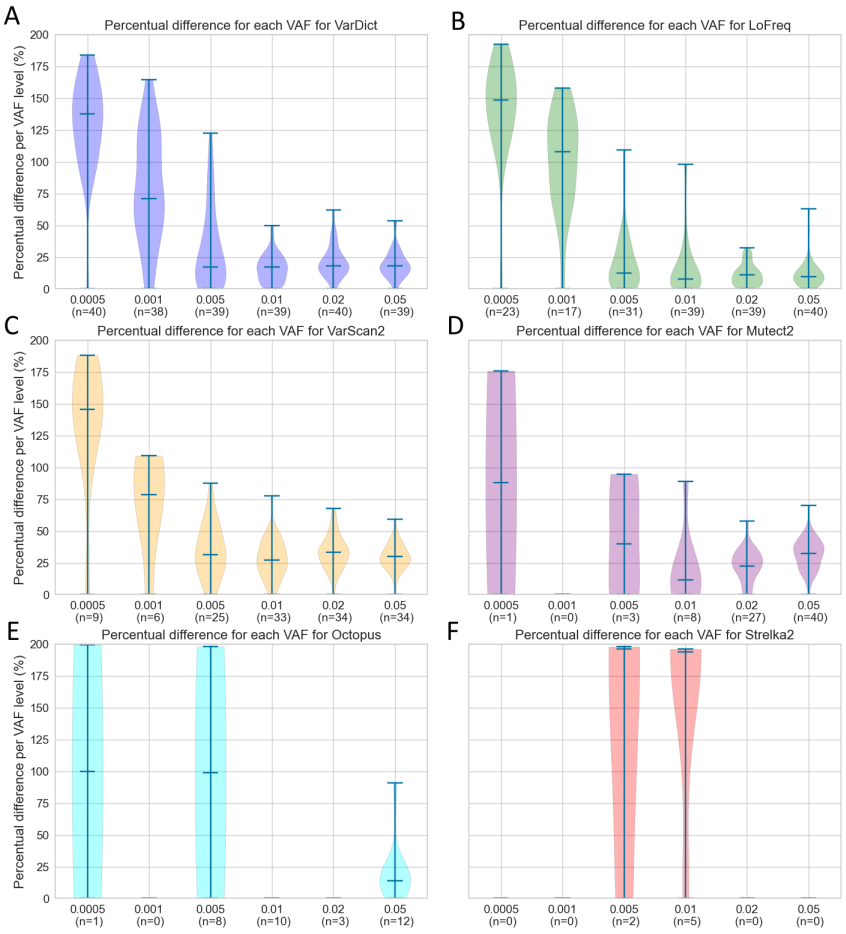

Figure 4: Violin plot showing the absolute percentual difference between spiked and identified VAFs for **A)** VarDict, **B)** LoFreq, **C)** VarScan2, **D)** Mutect2, **E)** Octopus, **F)** Strelka2.

4.2.3 False positives

For approximating the false positives from the non-spiked mutations, A consensus calling approach was used. The total of 322 non-spiked mutations were identified across all of the six variant callers. Of these 110 mutations were identified by at least two vari-

ant callers. what can be observed is that as the minimum number of tools required to call a specific non-spiked mutation increased, the number of called mutations gradually decreased. It is also important to look at the overlap of mutations called by multiple tools, there it can be observed that VarDict called a large portion of the non-spiked mutations, as shown in Figure 5. Based on these results, we classified any non-spiked mutation called by more than one variant caller as a somatic mutation, thus excluding them from the false positives. Therefore we classified the 112 non-spiked mutations identified by only one variant caller as false positives.

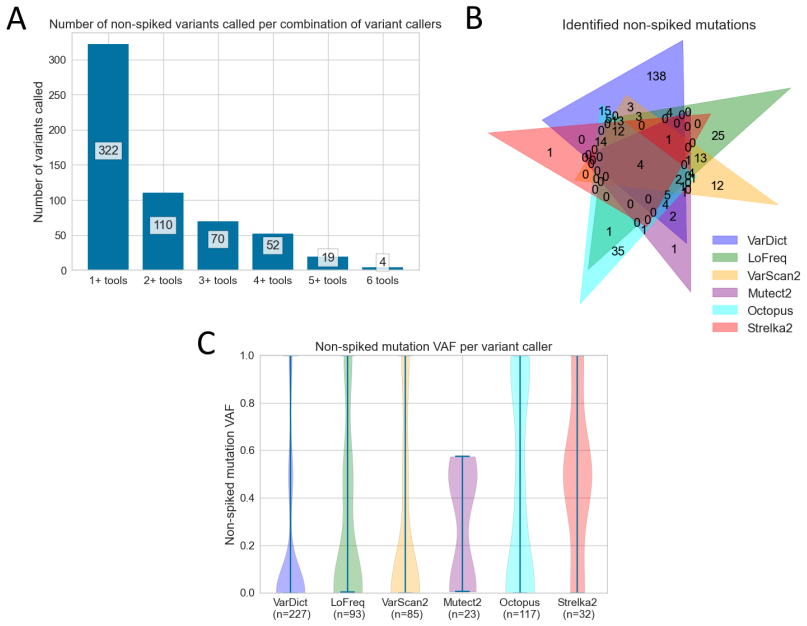

Figure 5: A) Bar graph of the number of non-spiked mutations identified by a certain amount of variant callers. The x-axis describes the minimum amount of tools that called the same non-spiked mutation. **B)** Venn diagram showing the overlap of identified non-spiked mutations per VAF per variant caller. The colors of the planes represent the six different variant callers, as depicted in the bottom right. The values of the non-overlapping planes represent the amount of non-spiked mutations identified by their representative variant caller. Values of overlapping planes indicate the amount of spiked mutations identified by that combination of variant callers. **C)** Violin plot of the identified VAFs from the non-spiked mutations per variant caller. The variant callers, along with their respective amount of non-spiked mutations identified are shown on the x-axis.

4.3 Sensitivity and Precision

The overall performance of the six variant callers can be estimated by combining the sensitivity with the precision. A confusion matrix is made to illustrate the proportions between true positives (TP), false positives (FP) and false negatives (FN). Where the

'true' labels represent the list of spiked variants and the 'predicted' labels as the list of the variants found by the variant callers. VarDict shows the highest amount of true positives but also shows a high amount of false positives. LoFreq and VarScan2 both show high true positives while having a much lower amount of false positives. Mutect2, Octopus and Strelka2 all three show low true positives and false positives (Figure 6).

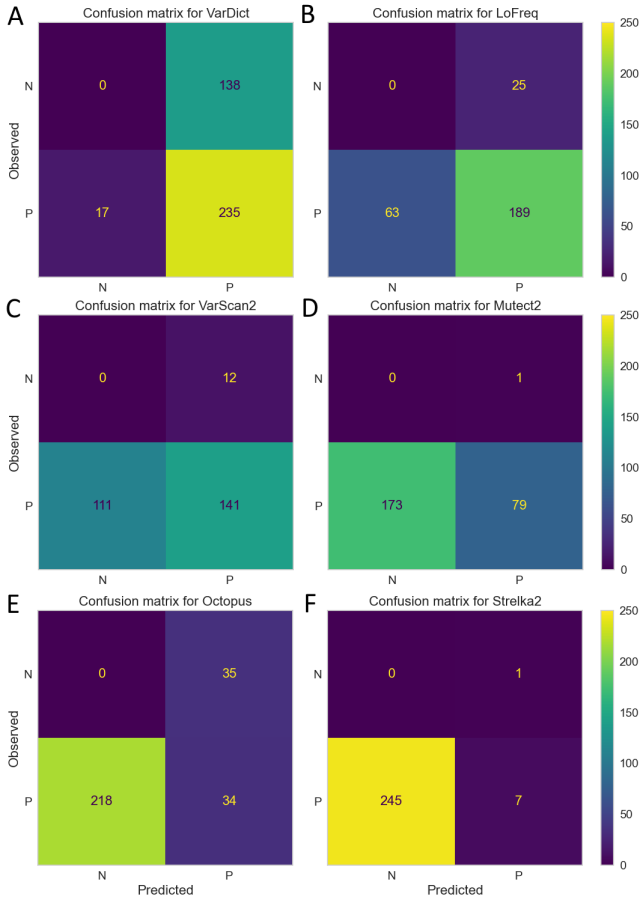

Figure 6: Confusion matrices per variant caller. P = positive, N = negative. Top left = TN, top right = FP, bottom left = FN, bottom right = TP. The color of the squares corresponds to the amount of TN, FP, FN or TP as indicated to the right of the matrices.

The sensitvity per VAF of each variant caller was plotted in a line graph as wel as a scatterplot (Figure 7). Here it can be seen that VarDict as shown in previous recall graphs has the highest sensitivity at each VAF, however LoFreq can be seen as the second best as it has the second highest sensitivity at low VAFs and almost the same sensitivity as VarDict at the higher VAFs. The other four are underperfroming at every VAF. Furthermore the precision is also plotted in a bar graph (Figure 7). Here Mutect2 is shown to have the highest precision. Varscan2 and LoFreq also have performed well

in precision. Vardict on the other hand has one of the lowest precision. Strelka2 also has performed well in the precision. Octopus has performed poorly as expected.

To get a more conclusive result a test was done to evaluate the sensitivity and precision at the same time. This is illustrated in scatterplot (Figure 7). In this scatterplot it can be seen that VarDict and LoFreq are the two variant callers that perform well in this test, with LoFreq showing the best sensitivity/precision ratio.

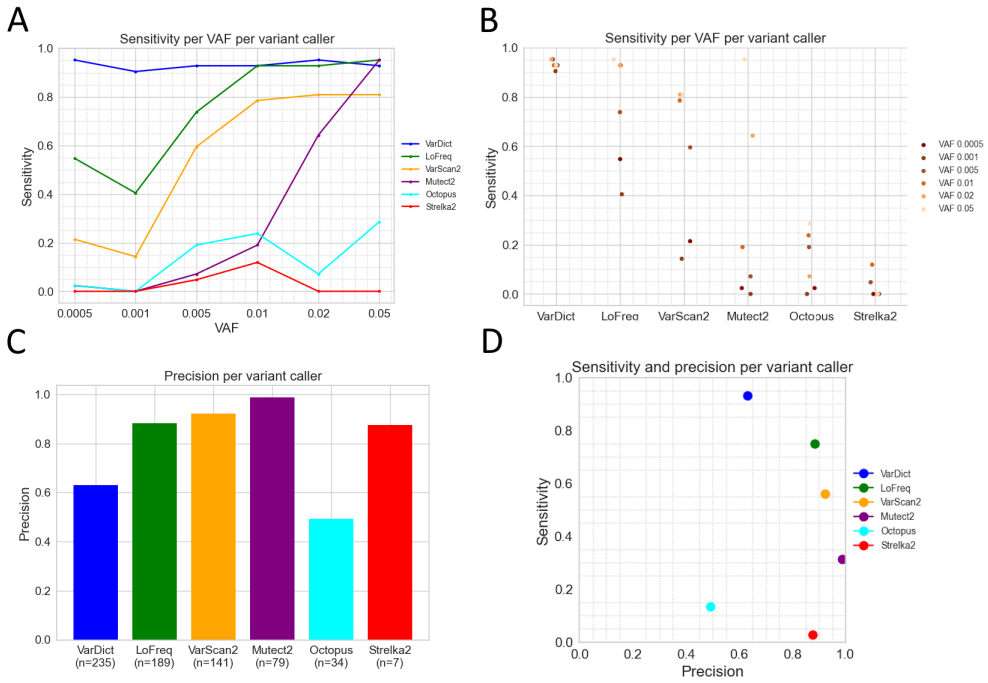

Figure 7: A) Line graph of the sensitivity per VAF per variant caller. The colors represent the variant callers as depicted on the right. **B)** Scatterplot of the sensitivity per VAF per variant caller. VAF levels are represented by different colors. A brighter color indicates a higher VAF level, a darker color a lower VAF level, as depicted on the right. **C)** Bar graph of the overall precision per variant caller. The x-axis shows the variant caller, along with their respective number of identified spiked mutations. **D)** Scatterplot of the precision and sensitivity for each variant caller. Each variant caller is represented by a color as depicted by the legend on the right.

This ratio is calculated and defined in a F1 score (Table 2). LoFreq having the highest score followed by VarDict and VarScan2. Mutect2 Octopus and Strelka2 did not perform well.

Table 2: F1 values per variant caller. The highest value is highlighted in bold.

Tool	F1 value
LoFreq	**0.81**
VarDict	0.75
VarScan2	0.70
Mutect2	0.48
Octopus	0.21
Strelka2	0.05

4.3.1 Paired sensitivity and precision

To test the effectiveness of combining variant callers, paired test were done to determine which pair performed the best.

The sensitivity of paired variant callers per VAF is shown in a line graph (Figure 8). The combination of VarDict and LoFreq performed the best and all of the other pairs that come close to this combination have either VarDict or LoFreq in them. The overall precision per pair is also illustrated in a bar graph (Figure 8). This graph shows that the pairs containing a combination of LoFreq, Varscan2 and Mutect2 have the best precision. In a scatterplot where the overall sensitvity and precision of each variant pair is illustrated (Figure 8), the combination of LoFreq and VarDict performed the best in this test. The pair is the closest to the top right. Thus having relatively the highest combination of precision and sensitivity.

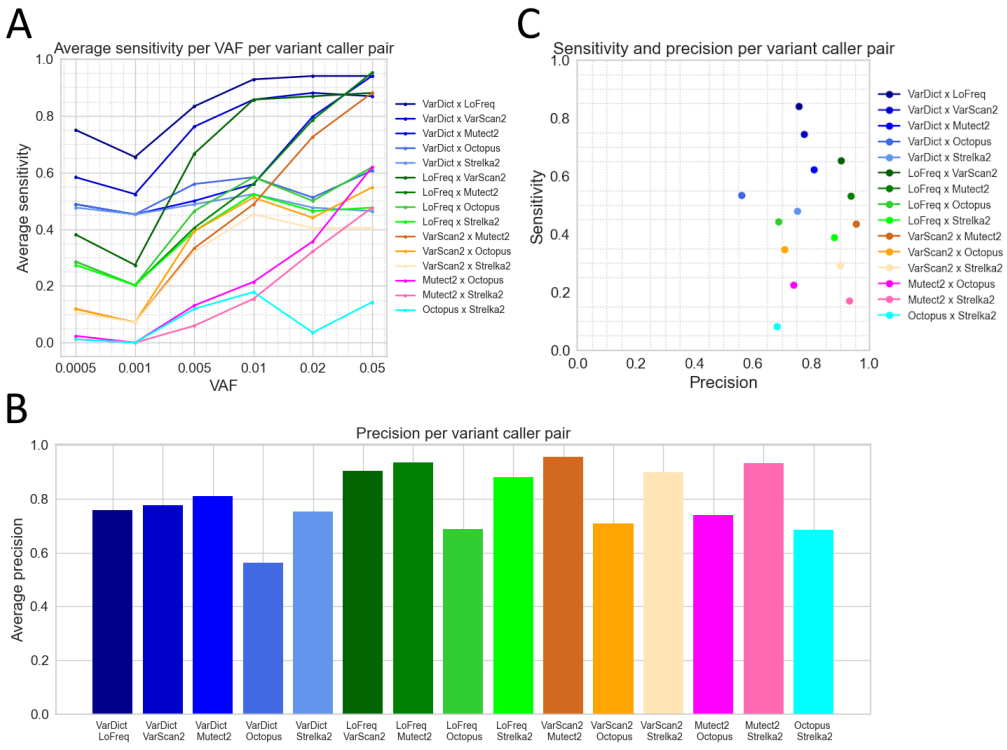

Figure 8: A) Line graph of the sensitivity per VAF per variant caller pair. The colors represent the variant caller pairs as depicted on the right. **B)** Bar graph of the overall precision per variant caller pair. The x-axis shows the variant caller pairs, along with their respective number of identified spiked mutations. **C)** Scatterplot of the precision and sensitivity for each variant caller pair. Each variant caller pair is represented by a color as depicted by the legend on the right.

The results of the sensitivity and precision per pair are again used to calculate the F1 score for the pairs (Figure 3). Where the pair of the combination of VarDict and LoFreq has the highest score.

Table 3: F1 values per variant caller pair. The highest value is highlighted in bold.

Pair	F1 value
VarDict x LoFreq	**0.80**
VarDict x VarScan2	0.76
LoFreq x VarScan2	0.76
VarDict x Mutect2	0.70
LoFreq x Mutect2	0.68
VarScan2 x Mutect2	0.60
VarDict x Strelka2	0.59
VarDict x Octopus	0.55
LoFreq x Octopus	0.54
LoFreq x Strelka2	0.54
VarScan2 x Octopus	0.47
VarScan2 x Strelka2	0.44
Mutect2 x Octopus	0.34
Mutect2 x Strelka2	0.29
Octopus x Strelka2	0.14

5 Discussion

The goal of this benchmark study was to find the most suitable variant callers for a new pipeline that can detect somatic mutations most accurately at low variant allele frequencies in ctDNA. After benchmarking, results showed that all of the variant callers (Mutect2, VarDict, LoFreq, Strelka, VarScan2, and Octopus) have unique strengths and weaknesses. VarDict and LoFreq showing very promising results. Octopus and Strelka2 however, proved to be very difficult to work with and thereby showing very little or disappointing results.

Sensitivity, measured by the recall rate of spiked-in mutations at various VAFs, varied noticeably among the tools. LoFreq and VarDict demonstrated the highest sensitivity, particularly at lower VAFs; all variant callers performed significantly worse than Vardict. However VarDict also had the most false positive calls because of its sensitivity, yet VarDict was matched in sensitivity by LoFreq at higher VAFs, making them both suitable for detecting rare variants in ctDNA. Strelka and Octopus did not perform well and were significantly less sensitive compared to the other variant callers. Mutect2 and VarScan2, while effective, showed reduced sensitivity compared to the others, especially in the lower VAF range.

Precision was a very important parameter. The consensus calling approach, which required a mutation to be called by at least 2 out of the six variant callers to be considered probable, helped filter out false positives. This method showed that while some tools were highly sensitive, they also called more false positives. Precision was calculated as the portion of called spiked mutations relative to all spiked mutations, gave a well balanced view of each tool's performance. It showed that LoFreq together with VarScan2, Mutect2 and Strelka2 had the best precision, indicating a good balance be-

tween sensitivity and specificity. VarDict, whilst being highly sensitive, was slightly lower in precision due to its higher false positive rate. Octopus showed some disappointing results with its low precision and had an overall low sensitivity.

When the sensitivity and precision were both taken in consideration altogether, the F1 score showed that VarDict and LoFreq scored the highest, both solo as well as paired together. To achieve a high sensitivity and precision, pairs can be made between the variant callers to balance out the strengths and weaknesses of each variant caller. With this method it was observed that VarDict and LoFreq are the pair that scored the highest in sensitivity. They were also the highest scoring pair in the test where the sensitivity and precision were both compared to each other.

Looking back at these results it can be concluded that VarDict paired with LoFreq prove to be the best combination for detecting the spiked somatic mutations in the semi-synthetic data. Where LoFreq is the most precise and VarDict is the most sensitive especially at the lower VAFs. Making it the best pair for the pipeline.

The semi-synthetic data that was used for this benchmark study, did not include the germline-tumor matched DNA that will be used in the new pipeline. In this study a second tumor file was used as the germline matched DNA file. This was done because of the unavailability of a germline matched DNA, due to the novelty of the pipeline.

Some variant callers performed as expected but some did not. VarDict, LoFreq and VarScan2 performed as expected when looking at results of previous research. VarDict is known for calling many false positives and also here it finds many false positives. LoFreq and VarScan2 showed high precision in identifying variants again in accordance with previous studies. The variant callers that did not perform as expected are Mutect2, Octopus and Strelka2. Because these variant callers performed poorly whilst they performed well in other studies. It could be explained by the command-line settings for these tools because they might not have been set correctly or optimized for tumor-matched variant calling. However for Octopus almost all options for relevant settings were exhausted. Together with the fact that some tools that are sensitive to low quality reads may underperform, when the mutations are in such reads.

These results are limited by using only one set of semi-synthetic data, so no cross comparisons were possible. Future studies should address this by using multiple sets of ground truth and semi-synthetic data. This would reduce the impact of one low-quality set and make the findings more reliable.

Despite its limitations, this study can serve as a basic guide for tumor-normal matched variant calling and help develop new variant calling methods.

References

[1] Riziero Esposito Abate et al. “Next generation sequencing-based profiling of cell free DNA in patients with advanced non-small cell lung cancer: Advantages and pitfalls”. en. In: *Cancers (Basel)* 12.12 (Dec. 2020), p. 3804.

[2] Qiang Gao et al. “Circulating cell-free DNA for cancer early detection”. en. In: *Innovation (Camb.)* 3.4 (July 2022), p. 100259.

[3] Elodie Bohers, Pierre-Julien Viailly, and Fabrice Jardin. “CfDNA sequencing: Technological approaches and bioinformatic issues”. en. In: *Pharmaceuticals (Basel)* 14.6 (June 2021), p. 596.

[4] Stanislav Volik et al. “Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies”. In: *Molecular Cancer Research* 14.10 (Oct. 2016), pp. 898–908. ISSN: 1541-7786. DOI: 10.1158/1541-7786.MCR-16-0044. eprint: https://aacrjournals.org/mcr/article-pdf/14/10/898/2181746/898.pdf. URL: https://doi.org/10.1158/1541-7786.MCR-16-0044.

[5] Regina Bohnert, Sonia Vivas, and Gunther Jansen. “Comprehensive benchmarking of SNV callers for highly admixed tumor data”. en. In: *PLoS One* 12.10 (Oct. 2017), e0186175.

[6] Zixi Chen et al. “Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency”. en. In: *Sci. Rep.* 10.1 (Feb. 2020), p. 3501.

[7] Chang Xu. “A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data”. en. In: *Comput. Struct. Biotechnol. J.* 16 (Feb. 2018), pp. 15–24.

[8] Luca Boscolo Bielo et al. “Variant allele frequency: a decision-making tool in precision oncology?” en. In: *Trends Cancer* 9.12 (Dec. 2023), pp. 1058–1068.

[9] Diana Mandelker and Ozge Ceyhan-Birsoy. “Evolving significance of tumor-normal sequencing in cancer care”. en. In: *Trends Cancer* 6.1 (Jan. 2020), pp. 31–39.

[10] Stephen E Lincoln et al. “Yield and utility of germline testing following tumor sequencing in patients with cancer”. en. In: *JAMA Netw. Open* 3.10 (Oct. 2020), e2019452.

[11] Elisabeth M Jongbloed et al. “Machine learning-based somatic variant calling in cell-free DNA of metastatic breast cancer patients using large NGS panels”. en. In: *Sci. Rep.* 13.1 (June 2023), p. 10424.

[12] Mark A DePristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. en. In: *Nat. Genet.* 43.5 (May 2011), pp. 491–498.

[13] Maria Maco et al. “Circulating tumor DNA in Hodgkin lymphoma”. en. In: *Ann. Hematol.* 101.11 (Nov. 2022), pp. 2393–2403.

[14] Nathan D Olson et al. “Best practices for evaluating single nucleotide variant calling methods for microbial genomics”. en. In: *Front. Genet.* 6 (July 2015), p. 235.

[15] Zhongwu Lai et al. “VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research”. en. In: *Nucleic Acids Res.* 44.11 (June 2016), e108.

[16] Andreas Wilm et al. “LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets”. en. In: *Nucleic Acids Res.* 40.22 (Dec. 2012), pp. 11189–11201.

[17] Nancy F Hansen et al. “Shimmer: detection of genetic alterations in tumors using next-generation sequence data”. en. In: *Bioinformatics* 29.12 (June 2013), pp. 1498–1503.

[18] Christopher T Saunders et al. “Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs”. en. In: *Bioinformatics* 28.14 (July 2012), pp. 1811–1817.

[19] Daniel C Koboldt et al. “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing”. en. In: *Genome Res.* 22.3 (Mar. 2012), pp. 568–576.

[20] Marth G. Garrison E. “Haplotype-based variant detection from short-read sequencing”. In: (2012).

[21] Daniel P. Cooke, David C. Wedge, and Gerton Lunter. “A unified haplotype-based method for accurate and comprehensive variant calling”. In: *Nature Biotechnology* (2021). ISSN: 1546-1696. DOI: 10.1038/s41587-021-00861-3. URL: https://doi.org/10.1038/s41587-021-00861-3%20https://www.nature.com/articles/s41587-021-00861-3.pdf.

[22] Nick Veltmaat et al. “Genomic profiling of post-transplant lymphoproliferative disorders using cell-free DNA”. en. In: *J. Hematol. Oncol.* 16.1 (Sept. 2023), p. 104.

[23] Adam D Ewing et al. “Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection”. en. In: *Nat. Methods* 12.7 (July 2015), pp. 623–630.