Assessment of MuTect2 and VarScan2 for somatic mutation detection in exome sequencing

Carmen Alves Sabin, Juliane Weller

SUMMARY

Next generation sequencing is generally performed to identify somatic mutations in cancer, with increasing use in, not only research, but also for diagnosis of clinical oncological patients to personalize and improve treatments. Somatic variant callers need two sets of sequencing data, one from cancer tissue and its normal tissue counterpart, to compare and detect somatic mutations. There are many somatic variant callers to choose from, but few comparison papers have been published, and therefore it is pivotal to find an efficient way of comparison between these tools, as there is no standard for detection of somatic mutations. An assessment of two somatic variant callers, MuTect2 and VarScan2, was performed on two matching data samples, tumoral and non-tumoral, acquired from the publication “SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach.” by Wang et al. (1). We hypothesized that MuTect2 would perform better with cancer samples, as it employs the probabilistic framework of Bayesian statistics, used by most existing variant callers. Their performance was analyzed in both synthesized and real cancer samples. Both variant callers performed similarly in different samples, although VarScan2 usually surpassed MuTect2 when mutation frequency was high, MuTect2 was more consistent throughout all mutation frequencies. We found out that VarScan2 has a higher number of concordant mutations at high frequencies but, when they drop below 20%, MuTect2 performs better identifying up to 4000 mutations to VarScan2’s 1000. Similarly, at frequencies over 40%, VarScan2 has a lower rate of missing mutations than MuTect2. Also, VarScan2 had a higher recall and higher precision than MuTect2. However, through the measuring of the F1-Score, MuTect2 proved to cover a wider range of accuracy for different mutation frequencies. MuTect2 outperforms VarScan2 in the synthetic data, as well as most of the data acquired from cancer patients.

INTRODUCTION

Mutations, changes in the DNA sequence of a cell, can result from errors in DNA replication during cell division, exposure to mutagens or viral infections. In contrast to germline mutations that occur in reproductive cells and passed onto an offspring, somatic mutations may arise during replication of cells during a lifetime or through DNA damage, caused for example by ultraviolet light (UV). Somatic mutations are widely connected to diseases, especially cancer, such as skin cancer or lung cancer, or others, like Sturge-Weber syndrome. (2) Types of DNA damage include mutations by substitution, deletion, or insertion of base pairs. Often, mutations are harmless but if a coding sequence is affected, they might lead to malfunctioning of the protein, which can lead to diseases, including cancer.(3)

To analyze these small changes in the DNA Sequence and how they affect cells as a whole, new techniques have been developed, the most advanced being Next-Generation Sequencing (NGS). It has revolutionized the way genetic laboratories and research groups operate and perform their genomic analyses (4). The lower costs and higher throughput allowed genetic testing of patients to move from single gene examinations to assessing all exons or even the whole genome for mutations. Human whole genome sequencing (WGS) allows detection of disease-causing variants in both protein encoding- and non-coding regions of the genome, with the prospect of being gradually implemented as a major tool in precision medicine. Whole-exome sequencing in contrast is focusing on the protein-coding regions of the genome. Since cancer mutations are directly linked to alterations in the coding part of the DNA, whole-exome sequencing has provided a new edge to analyze the human exome. As the human exome represents less than 2% of the genome but contains ~85% of known disease-related variants, it makes this method a cost-effective alternative to whole-genome sequencing. (5)

After sequencing a DNA sample, the collected genomic data is compared to a reference genome to identify somatic mutations. For whole-exome sequencing, a large variety of tools, called Variant Callers, have been implemented to find somatic mutations in a sample, each having different strengths. Examples include MuTect2 (6), VarScan2 (7), Strelka2 (8), or Virmid (9). Choosing the best tool impacts the analysis outcome and cost of analysis. In this paper, we will explore and compare the following bioinformatic tools: VarScan2 and MuTect2. Both VarScan2 and Mutect2 call Single Nucleotide Variants (SNV), but, unlike Strelka2 and Virmid, they allow for single-sample input. These two were chosen to compare since they use different core algorithms: Mutect2 uses allele frequency analysis and VarScan2 uses the heuristic threshold approach.

The general workflow for variant callers seen in Figure 1 can be summarized in the following steps: quality control, sequence alignment, post-alignment processing, variant calling, and downstream analysis. (10)

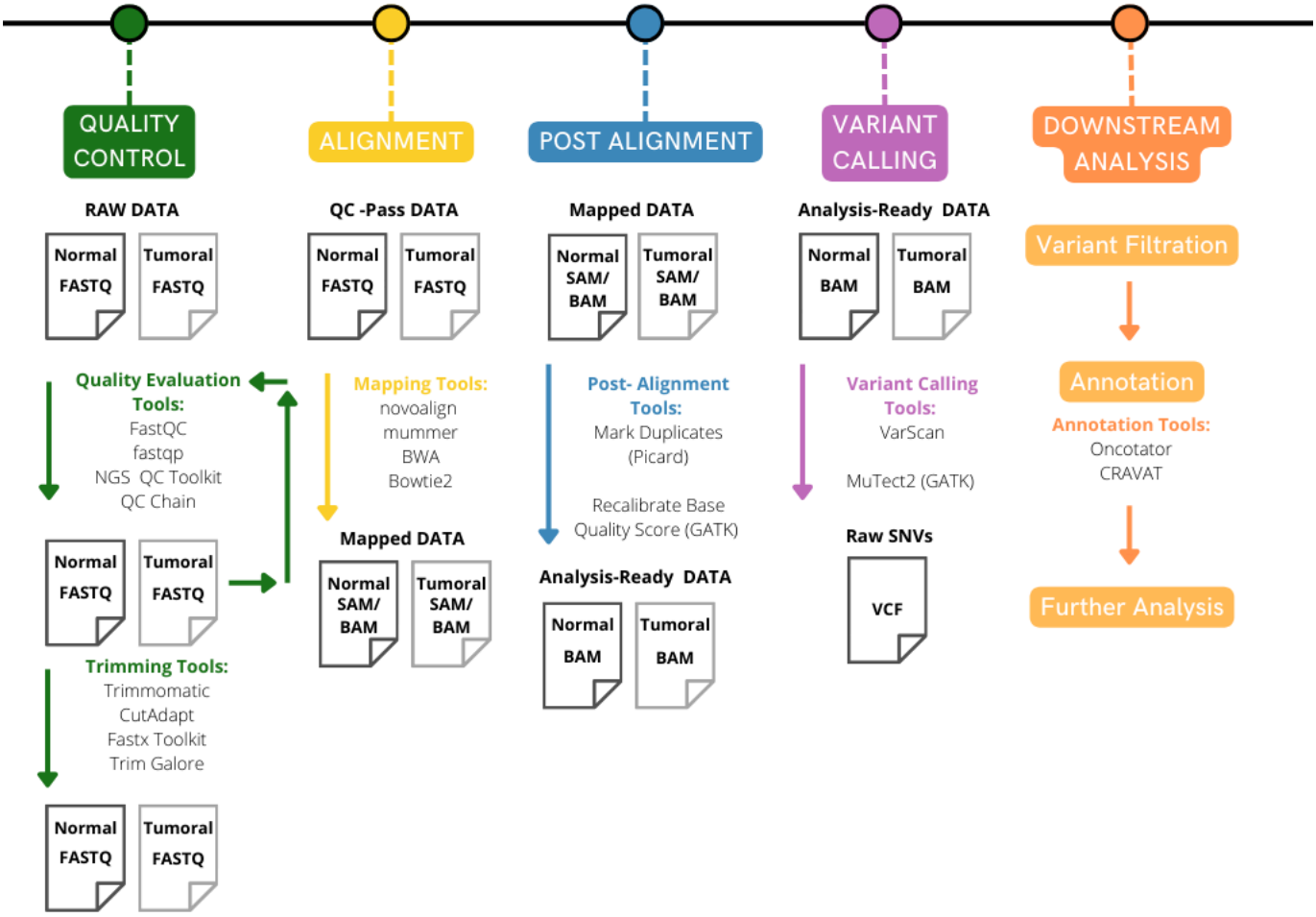

Figure 1. *A somatic variant workflow. In each step of the process, the most widely used tools are also stated. The processing steps are carried out for both the normal and tumoral samples.*

In quality control, the raw data from a sequencing machine are most widely provided as FASTQ files, which include sequence information. Usually, this raw data is not immediately ready to be used for variant calling. The first step of Whole Exome Sequencing is the quality control (QC) step to reduce the chances of encountering bias or missing data. The QC process is cyclical, where (i) the quality is evaluated, (ii) QC is stopped if the quality is adequate, and (iii) a data altering step, such as trimming low-quality reads is performed, and then the QC is repeated beginning from step (i). The most used tool for evaluating the quality of FASTQ files is FastQC (36), and therefore used for our data, but others include, fastqp (37), NGS QC Toolkit (38), PRINSEQ (39), and QC-Chain (40) (10).

To find the exact locations of reads in the alignment step, each must be aligned to a reference genome. Efficiency and accuracy are crucial in this step because large quantities of reads could take days to align, and a low-accuracy alignment would cause inadequate analyses. For humans, the most current and widely used reference sequences are GRCh37 (hg19) and GRCh38 (hg38). There are multiple tools for aligning sequences to the reference genome, to name a few, BWA (11), Bowtie2 (12), novoalign (13), and mummer (14). To keep our analysis tools consistent and minimize false callings, Novoalign Software was used in the aligning step. After aligning, a Sequence Alignment Map (SAM) file is produced, containing the reads aligned to the reference. The binary version of a SAM file is called a Binary Alignment Map (BAM) file and is used for random access purposes. The SAM/BAM file consists of a header and an alignment section. The header section contains contigs of aligned reference sequence, read groups, and sometimes data processing tools applied to the reads. The alignment section includes information on the alignments of reads. (15) (16)

The next step is post-alignment data processing, which produces analysis-ready BAM files. This step includes data clean-up operations, such as marking duplicates, and recalibrating base quality scores. Duplicates arise in PCR amplification of fragments, and since they share the same sequence and alignment position, they can lead to problems in variant detection. To overcome this, PCR duplicates are marked with a tagging algorithm, MarkDuplicates in the tool Picard (17), also used in our data. Next-Generation Sequencing tools provide quality scores of each base measured with the Phred Score: A Phred Score of 10 represents 90% accuracy, 20 equals 99%, 30 equals 99.9%, and so on. The raw scores produced by the sequencing machine are prone to over- or underestimating base quality scores, Base Quality Score Recalibration (BQSR) approaches, such as those provided by the Genome Analysis Toolkit (GATK) (18)(19), the one applied to our data, readjust the scores. This recalibration improves the reliability of the downstream steps in further analyses.

Following data processing steps, the reads are ready for downstream analyses, and the following step is usually variant calling. Variant calling is the process of identifying differences between the sequencing reads, resulting from NGS experiments and a reference genome. Methods for detecting short variants can be broadly categorized into probabilistic methods and heuristic-based algorithms. In the variant calling step for somatic mutations, both the tumor and normal processed read data are used to identify somatic Single Nucleotide Variant (SNV)/indels, i.e., short variants that are present in the tumor but not in the normal. Several tools exist for tumor-normal somatic variant calling, such as MuTect2 (6), VarScan2 (20), Strelka (8), SomaticSniper (21), SAMtools (22) and SomaticSeq (23). After the raw SNVs have been identified by the somatic variant calling, the program outputs a VCF (Variant Call Format) file.

In the downstream analysis step, the VCF file with the raw SNVs is filtered and annotated for interpretation. Different approaches exist for filtering raw somatic variants, where tumor-specific

metrics most frequently include the estimation of tumor heterogeneity and cross-sample contamination. In terms of the annotation tools used, several of them can provide cancer-specific annotations in addition to the general annotation available, for example are Oncotator (24) and CRAVAT (25). (10)

VarScan is a platform-independent mutation caller for targeted, exome, and whole-genome sequencing data generated on Illumina, SOLiD, Life/PGM, Roche/454, and similar instruments. The newest version, VarScan2, is written in Java, so it runs on most operating systems. It can be used to detect different types of variation, such as germline variants (SNPs and indels) in individual samples or pools of samples. Most of the published variant callers for next-generation sequencing data employ a probabilistic framework, such as Bayesian statistics, to detect variants and assess confidence in them. These approaches generally work quite well but can be confounded by numerous factors such as extreme read depth, pooled samples, and contaminated or impure samples. In contrast, VarScan2 applies heuristic methods and a statistical test to detect variants that meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance. (7)

MuTect2 (6) is a method developed at the Broad Institute for the reliable and accurate identification of somatic mutations (expected to occur at a rate of ~1 in a Mb) in next generation sequencing data of cancer genomes. MuTect2 consists of three steps. Firstly, the aligned reads in the tumor and normal sequencing data are preprocessed, ignoring reads with too many mismatches or very low-quality scores. Then, a statistical analysis that identifies sites that are likely to carry somatic mutations with high confidence by using two Bayesian classifiers – it detects whether the tumor is non-reference at a given site and verifies that the normal does not carry the variant allele. In practice the classification is performed by calculating a LOD score (log odds) and comparing it to a cutoff determined by the log ratio of prior probabilities of the considered events. Lastly, the candidates of somatic mutations are post-processed to eliminate artifacts of next-generation sequencing, short read alignment and hybrid capture. During variant calling MuTect2 also generates a coverage file, which indicates for every base whether it is sufficiently covered in the tumor and normal to be sensitive enough to call mutations. Mutect2 currently uses cutoffs of at least 14 reads in the tumor and at least 8 in the normal (these cutoffs are applied after removing noisy reads in the preprocessing step). In addition, wiggle files can also be generated of the observed depth in the tumor and in the normal. (6) (26)

With the ever-changing world of bioinformatics, there are multiple different tools that can be used for somatic variant calling and it is at times difficult to see which one is most efficient in each scenario. The aim of this project will be to compare the different variant callers - VarScan2 and MuTect2 and determine the better tool in different situations when analyzing somatic mutations of cancer genomes. Both VarScan2 and Mutect2 are popular tools when calling Single Nucleotide Variants (SNV). However, these two variant callers employ two different core

algorithms: Mutect2 uses allele frequency analysis and VarScan2 uses the heuristic threshold approach, so wanted to compare the different mathematics behind these widely-used variant callers.

We measured the effectiveness of the variant callers at different mutation frequencies by considering the following parameters: the SNV Count (the number of SNVs in the sample reported by the variant caller), Concordance Count (the number of SNVs reported by the variant caller that were actually SNVs), Missing Count (the number of SNVs that were in the sample that the variant caller failed to report), Precision rates (the ratio between Concordance Count and SNV Count), Recall rates (the ratio between the Concordance count and the number of SNVs that were in the sample), F1 Scores, predicted positives, predicted negatives, true negatives and true positives. These mutation frequencies and rates provide information about how often a mutation may occur in a sample. Very frequent mutations are often associated with driver mutations in cancer, while less frequent mutations give indications on genetic diversity and cancer evolution as cancer cells accumulate mutations over time (34).

We hypothesized that MuTect2 would perform better with cancer samples, as it employs the probabilistic framework of Bayesian statistics that most existing variant callers use. We found out that VarScan2 generally detects more mutations and has a higher number of concordant mutations at high frequencies but when they drop below 20%, MuTect2 performs better identifying up to 4000 mutations in comparison to VarScan2's 1000 mutations. Similarly, at high mutation frequencies over 40%, VarScan2 has a lower rate of missing mutations than MuTect2. Also, VarScan2 had a higher recall (low false positive rate) and high precision (low false negative rate) than MuTect2. However, through the measuring of the F1-Score, MuTect2 proved to cover a wider range of accuracy for different mutation frequencies, since MuTect2 outperformed VarScan2 for most of the mutation frequencies. MuTect2 outperforms VarScan2 in the synthetic data, as well as most of the data acquired from cancer patients. Though VarScan2 performs better at very high mutation frequencies (60% to 100%), Mutect2 tends to be more consistent, especially for low mutation frequencies.

RESULTS

Here, the performance of two variant callers, VarScan2 and MuTect2, are compared for a range of mutation frequencies and on different synthetic samples and cancer samples. To measure the effectiveness of the variant callers at different mutation frequencies, the SNV Count, Concordance Count, Missing Count, Precision rates, Recall rates, F1 Scores, predicted positives, predicted negatives, true negatives and true positives were considered.

Whilst the SNV count is interesting, when paired with concordance count it becomes more relevant, since the number of true mutations called out of all mutations called by the variant caller can be seen. The performance of the somatic mutation callers on datasets with different

frequencies of computationally introduced mutations were compared. At high mutation frequencies, VarScan2 detects more mutations and has a higher number of concordant mutations (Figure 2). For frequencies below 20%, MuTect2 performs better at identifying up to 4000 mutations in comparison to VarScan2 that identifies up to 1000 mutations (Figure 2).

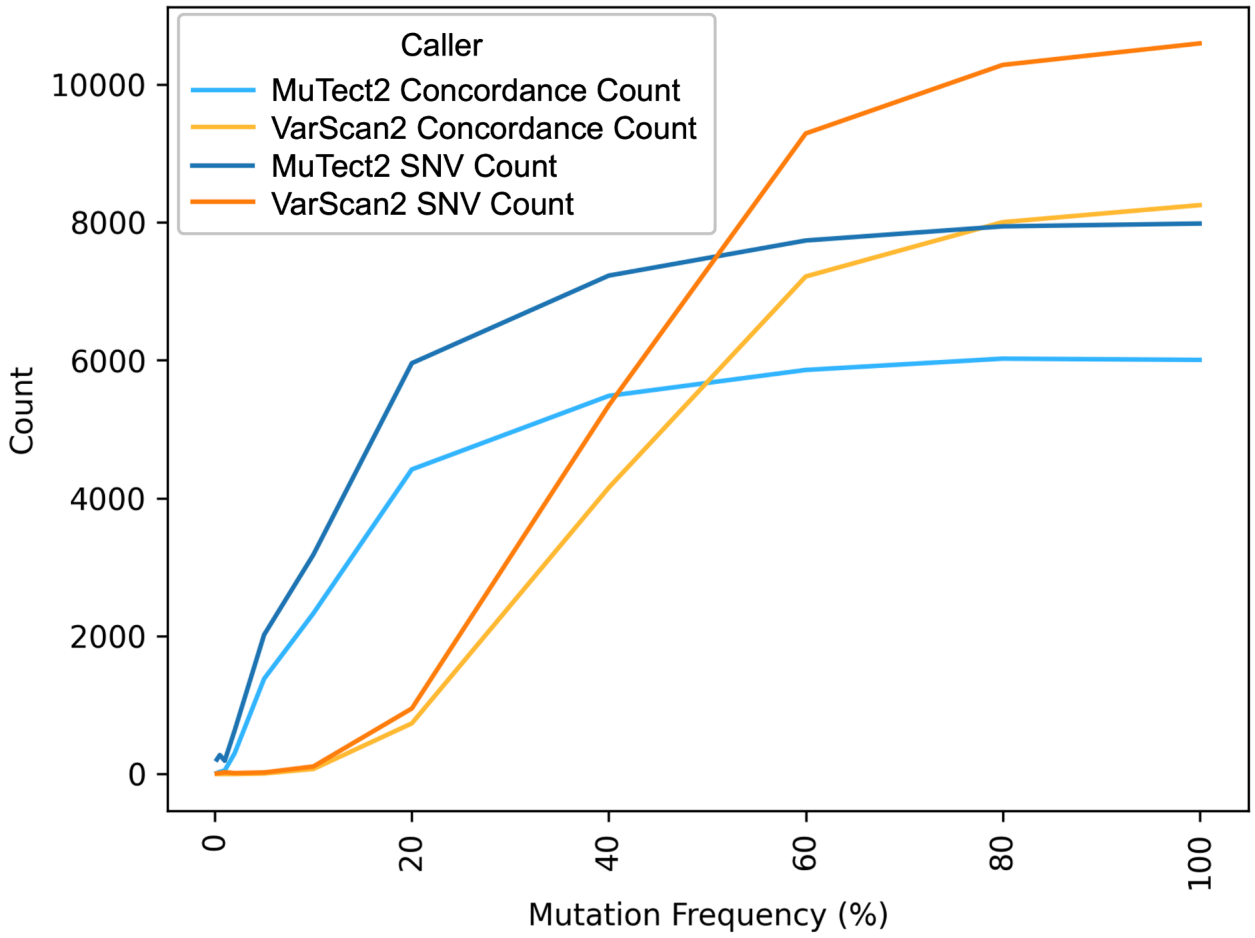

Figure 2. *SNV and Concordance Count of Variant Callers on different datasets. Curves indicating the total SNV count and the concordance counts for MuTect2 and VarScan2 for different mutation frequencies ranging from 0.2% to 100%. (light blue line: concordance MuTect2, dark blue line: SNV counts MuTect2, yellow line: concordance VarScan2, orange line: SNV count VarScan2). At mutation frequencies higher than 40%, VarScan2 performs better than MuTect2, since it has both a higher SNV and Concordance count.*

For appropriate treatment, identifying all somatic mutations in a sample is important. When comparing the missing counts for both somatic mutation callers, at high mutation frequencies, VarScan2 has a lower rate of missing mutations than MuTect2 (Figure 3). For frequencies below

40%, MuTect2 performs better since it misses fewer mutations than VarScan2 (Figure 3). This result is consistent with the previously observed concordance count (Figure 2).

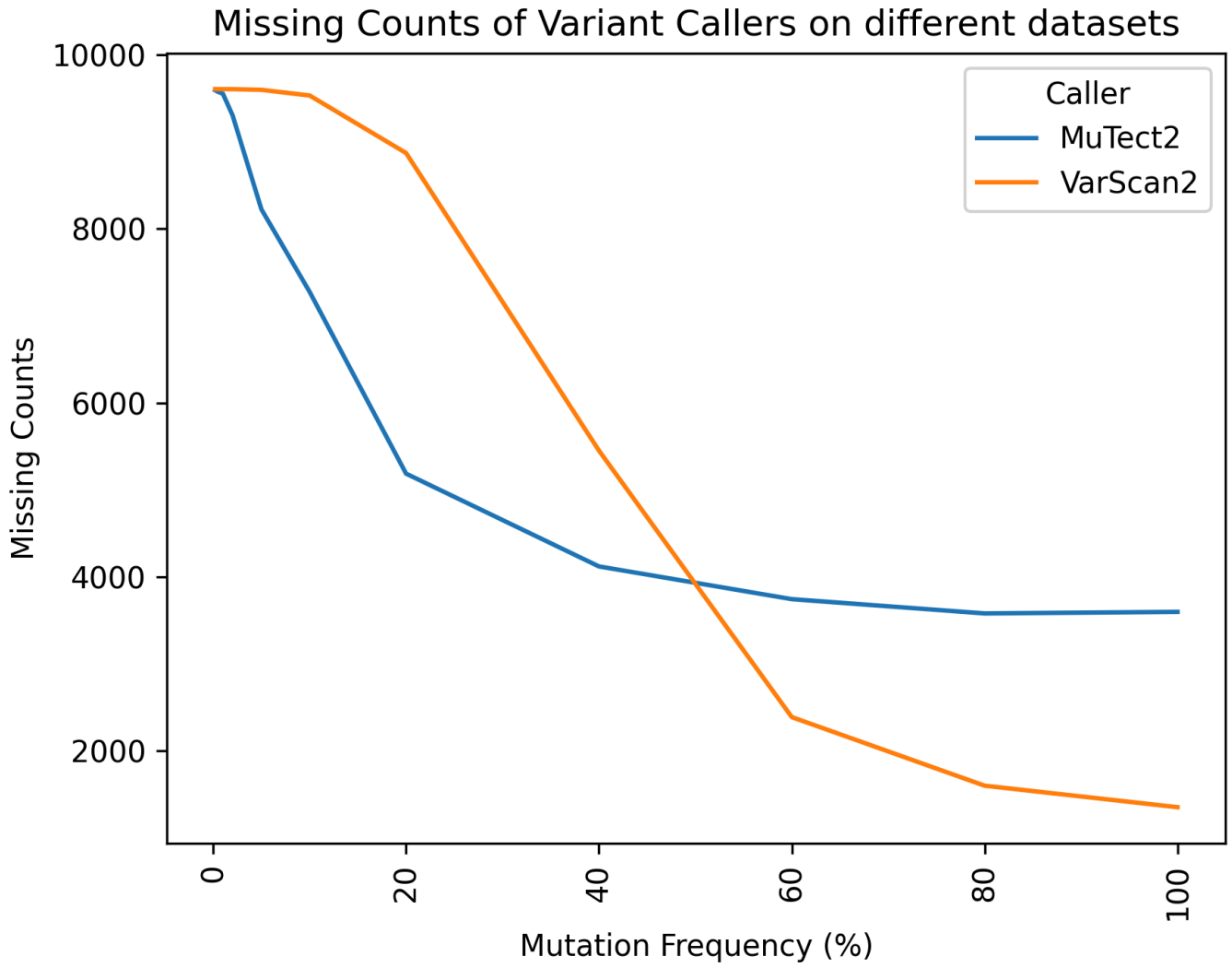

***Figure 3**. Counts of missed SNV. Curves depicting the missing counts for VarScan2 and MuTect2 in different mutation frequencies ranging from 0.2% to 100%. Blue line: MuTect2, orange line: VarScan2. For mutation frequencies higher than 50%, VarScan2 outperforms MuTect2, since it misses less SNVs present in the sample.*

For SNVs, VarScan2 had a higher concordance and SNV count at high frequency mutation samples than MuTect2. If the mutation frequency drops below 20%, VarScan2 calls approximately 700 mutation calls at 20% frequency, whilst MuTect2 has approximately 4000 calls at 20% mutation frequency (Figure 2). In relation to missing counts of the total, for

VarScan2, approximately 10% - 20% of the counts are missing for frequencies ranging from 60% to 100%, which increases noticeably to 90% for mutation frequencies less than 20% (Figure 3). For MuTect2, in the range of frequencies from 20% to 100%, the missing count rate is around 35%, increasing substantially to 90%, if the frequency is lowered from 5% (Figure 3).

Another way to measure the effectiveness of the variant callers is comparing the recall rate to their precision rate, obtained from different mutation frequencies. The precision is useful when measuring the accuracy of the tool, whilst the recall is interesting when investigating whether the variant caller returns most of the positive results. VarScan2 overall performed better than MuTect2, since it has a larger area under the curve, meaning that VarScan2 had a higher recall (low false positive rate) and high precision (low false negative rate), returning accurate amount of most positive calls present in the samples (Figure 4).

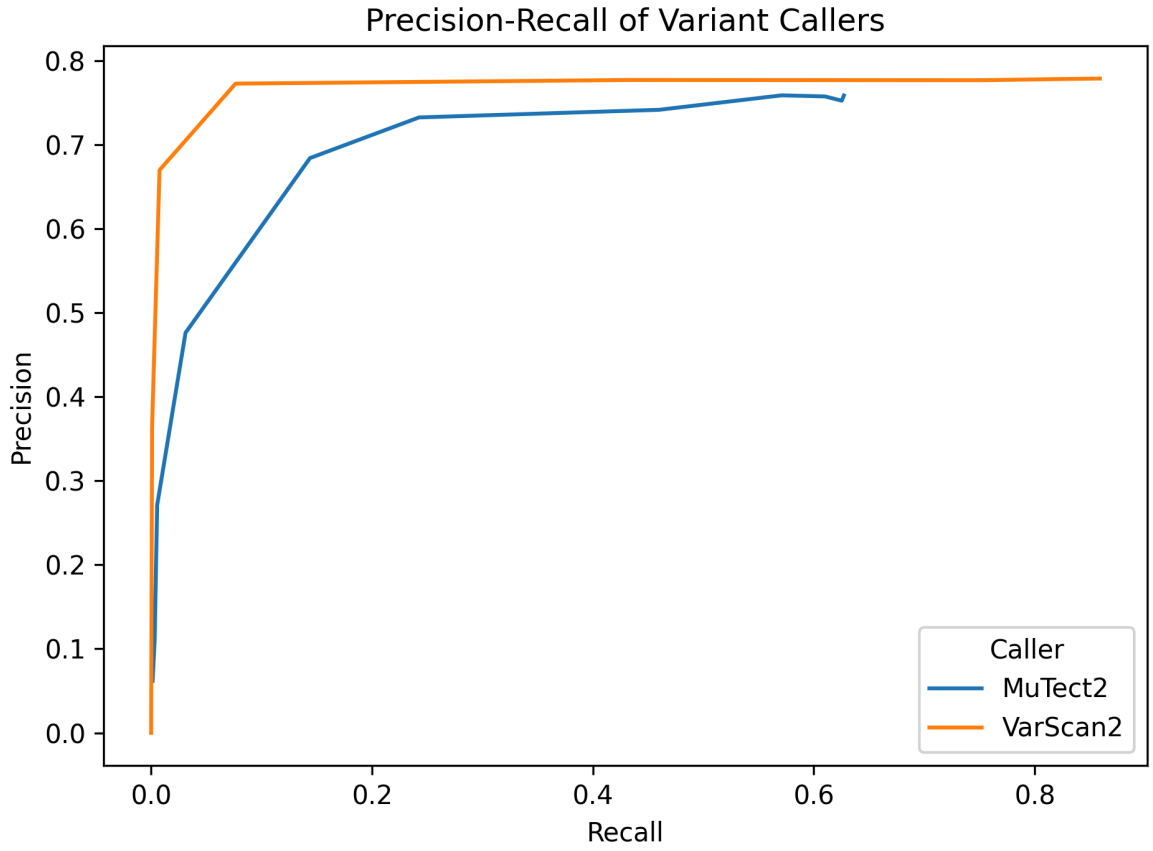

Figure 4. *Precision-Recall curves for MuTect2 and VarScan2 for different mutation frequencies ranging from 0.2% to 100% [0.2%, 0.5%, 1%, 2%, 5%, 10%, 20%, 40%, 60%, 80%, 100%]. VarScan2 overall performed better than MuTect2, since it has a larger area under the curve, meaning that VarScan2 had a higher recall (low false positive rate) and high precision (low false negative rate), returning an accurate amount of most positive calls present in the samples.*

The accuracy of the somatic variant calling in at different mutation frequencies was also measured through the F1-Score, the harmonic mean between recall and precision. Again, VarScan2 performed better for high mutation frequencies, slightly higher than 0.8 (1.0 being perfect recall and precision), than MuTect2 (Figure 5). However, MuTect2 proved to cover a wider range of accuracy for different mutation frequencies, since MuTect2 outperformed VarScan2 for low mutation frequencies and similarly for high mutation frequencies (Figure 5).

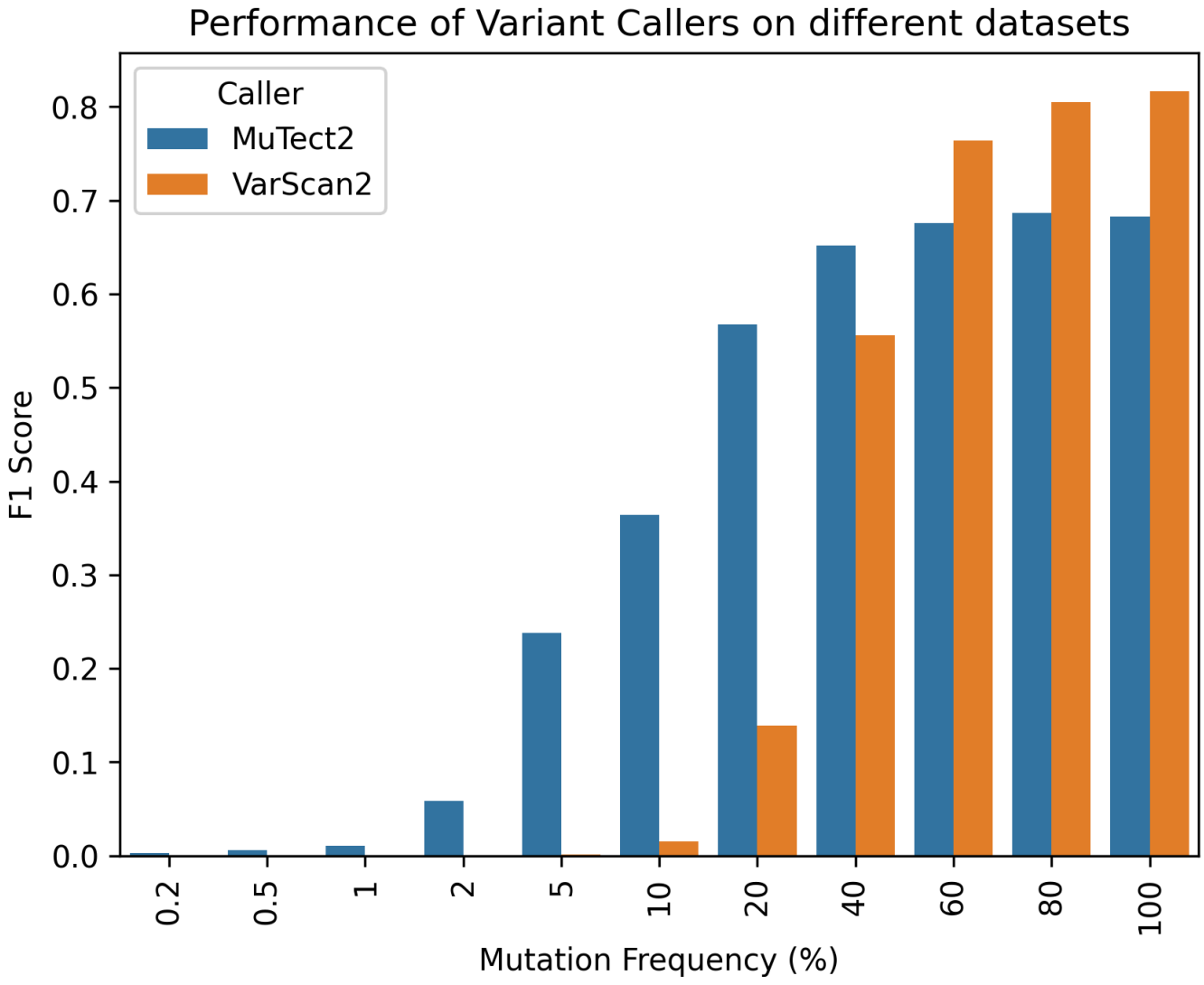

Figure 5. *F1-Score for VarScan2 and MuTect2 on different datasets. Blue bars: Mutect2, orange bars: VarScan2. VarScan2 performed better for high mutation frequencies, slightly higher than 0.8 (1.0 being perfect recall and precision), than MuTect2. However, MuTect2 proved to cover a wider range of accuracy for different mutation frequencies, since MuTect2 outperformed VarScan2 for most of the mutation frequencies.*

A confusion matrix was plotted for both VarScan2 and MuTect2, investigating their effectiveness at 50% sequencing coverage. In a confusion matrix, the relationship between the number of true negatives, true positives, predicted negatives and predicted positives can be observed. Whilst VarScan2 had a higher rate of predicted negatives turning out to be true positives (with a mean of 364.5 predicted negative calls turning out to be true), MuTect2 had a higher rate of predicted positives being actual true positives (with a mean of 326.5 predicted positive calls turning out to be true) (Figure 6). Also, whilst VarScan2 had no predicted positives that turned out to be true negatives, MuTect2 had a rate of 12.2. This means that when VarScan2 predicts a mutation to be positive, it has a higher chance of it being true than a predicted positive by MuTect2 (VarScan2 had an 100% accuracy in this case). However, when the variant caller predicts a mutation to be negative, VarScan2 turns out to be false more times than MuTect2 (Figure 6). The rate of predicted negatives being actual true negatives is unknown.

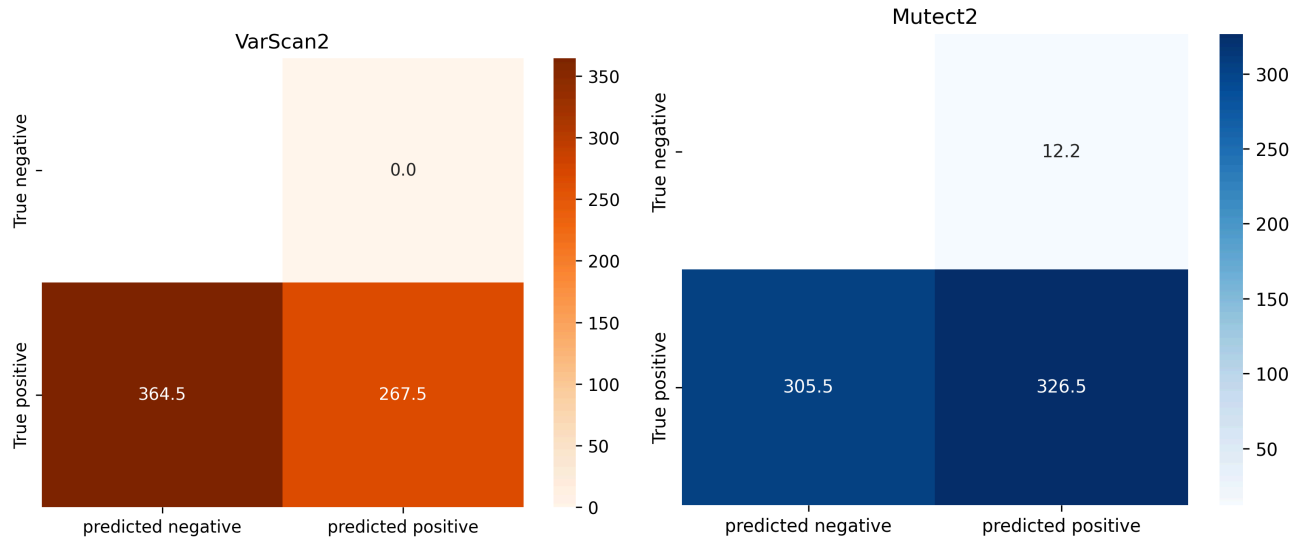

Figure 6. *Confusion matrix depicting the true positives, predicted negatives and predicted positives for VarScan2 and MuTect2 with 50% sequencing coverage. The vertical axis represents the number of SNV counts in each category. Whilst VarScan2 had a higher rate of predicted negatives turning out to be true positives (with a mean of 364.5 calls), MuTect2 had a higher rate of predicted positives being actual true positives (with a mean of 326.5 calls).*

Finally, the F1 Score was also analyzed to evaluate the performance of the variant callers with datasets containing different types of mutations, both synthesized data and clinical data. MuTect2 outperforms VarScan2 in the synthetic data, as well as the data acquired from cancer patients with chronic lymphocytic leukemia (CLL) and metastatic melanoma cell line (COLO), whilst VarScan2 performed better for acute myeloid leukemia (AML) and a malignant pediatric brain tumor in the cerebellum (MB) (Figure 7). VarScan2 and MuTect2 performed very similarly for MB tumor and COLO829. However, in the other two samples (AML and CLL), there was an

approximately 0.1 difference in the F1-Score for acute myeloid leukemia (AML), in favor of VarScan2, and a 0.2 difference for chronic lymphocytic leukemia (CLL), in favor of MuTect2 (Figure 7).

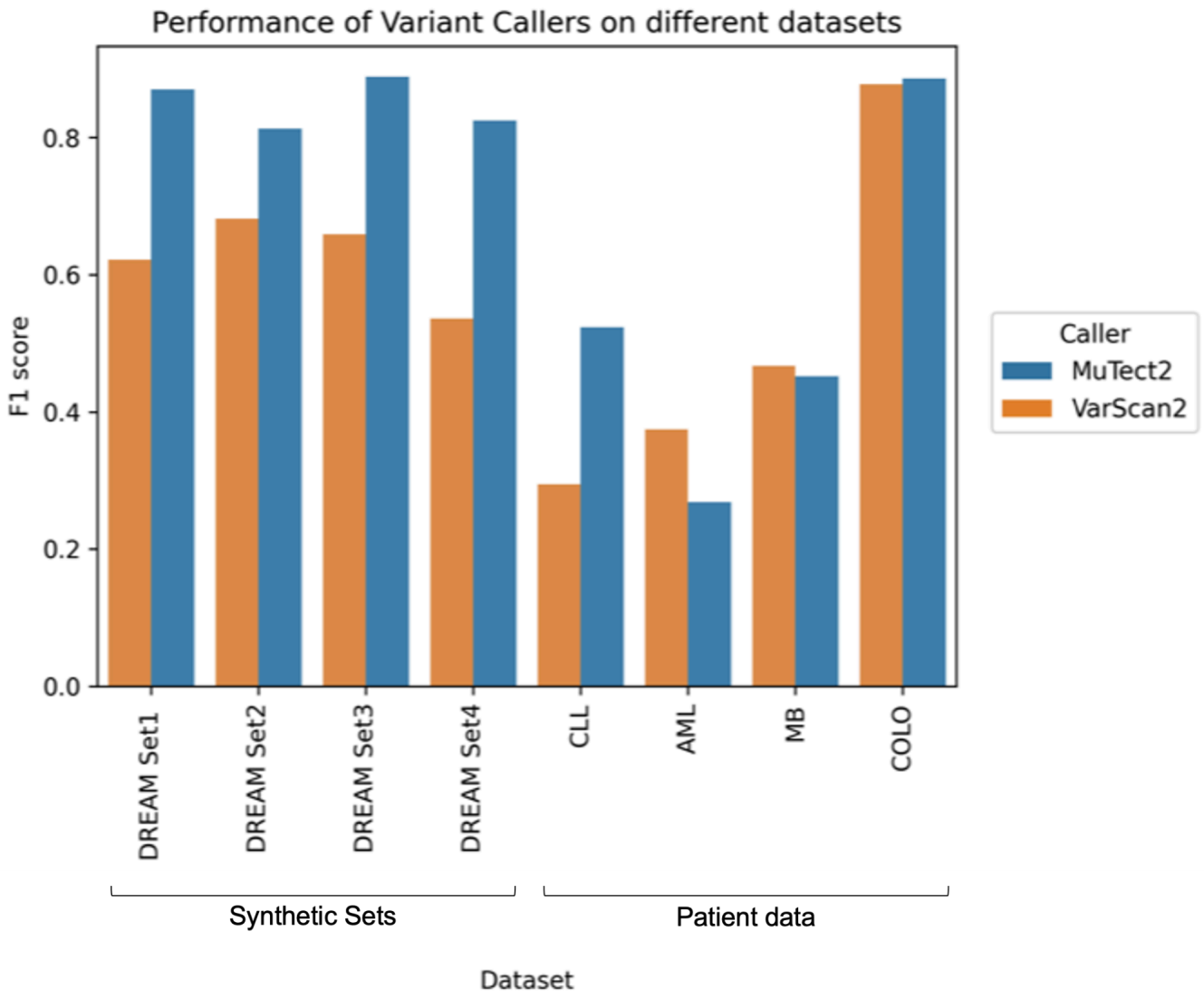

Figure 7. *F1-Score for VarScan2 and MuTect2 for datasets with different mutations, from synthesized sets (DREAM Set1, DREAM Set2, DREAM Set3,DREAM Set4) to real sets (CLL, AML, MB and COLO). Blue bars: Mutect2, orange bars: VarScan2. MuTect2 outperforms VarScan2 in the synthetic data, as well as the data acquired from cancer patients with chronic lymphocytic leukemia (CLL) and metastatic melanoma cell line (COLO), whilst VarScan2 performed better for acute myeloid leukemia (AML) and a malignant pediatric brain tumor in the cerebellum (MB).*

DISCUSSION

Somatic variant callers need to balance between detecting true non-repeating somatic mutations and fine tuning the calling procedure to reduce the number of false positive calls. Each variant caller algorithm has its own version of this balancing act, although usually the performance is dependent on the nature of the cancer samples (35). Here, we compared the performance of MuTect2 and VarScan2 at different mutation frequencies from NA11840 and NA12878, two cell lines from the GIAB project, and viewed the performance on synthetic mutations of DNA and a variety of mutations of real cancers (CLL, AML, MB and COLO). (1)

Overall, an increase in mutation frequency results in a higher level of variant calling for the two different callers. For SNVs, VarScan2 has been shown to be more sensitive to high frequency mutation samples. However, after the mutation frequency drops below 20% mutation frequency, VarScan2 performs poorly, whilst MuTect2 performs much more consistently throughout the range of mutation frequencies (Figure 2). When the mutation frequency of the sample is decreased, the percentage of missing counts of the total increases (Figure 3). For VarScan2, as discussed above in the results, approximately 10% - 20% of the counts are missing for frequencies above 60% to 100%, which increases noticeably to 90% for mutation frequencies less than 20%. For MuTect2, however, the missing count rate is around 35% in frequencies above 20%, only increasing substantially to 90%, if the frequency is lower than 5% (Figure 3). This suggests that these algorithms are reliable in both medium and high mutation frequencies, where VarScan2 is seen to recall mutation better for high frequencies, reporting a high fraction of mutations better. Therefore, they can be categorized as performing in a more robust, coverage-independent manner, in other words, the results are independent of the number of times the sample has been sequenced. No variant caller detects all the mutations but by using a combination of somatic variant callers this may be achieved, which might improve sensitivity (25).

Controlling the false positives is a major challenge in somatic variant calling. Usually, higher content of tumor cells means that true somatic mutations are more easily detected, which usually reduces the false positive rate. At 50%, VarScan2 performed better than MuTect2 in terms of false positive rate since it had a rate of 0 mutations to be false positives. However, when predicted negative by the variant caller, MuTect2 is more likely not to miss a true positive, since when predicted positive, MuTect2 has a higher rate (with a mean of 326.5 calls that were predicted positive turned out to be true positives as opposed to the 267.5), and a lower for the missing true positives (a mean of 305.5 calls were predicted negative that were actually true) than VarScan2 (Figure 6). Therefore, although VarScan2 presented a slightly lower false positive rate, the likelihood of MuTect2 missing a true positive is less than for VarScan2 (Figure 6). However, having a high confidence that the calls are true, can be interesting, when analyzing specific mutations in the tumor sample to create targeted treatment for the specific mutation.

Also, the two somatic variant callers performed differently in different types of cancer tissue and synthetic tissue. VarScan2 performed better in the synthetic samples (DREAM sets), useful when investigating in the laboratory the impact of different mutations in the function of cells (Figure 7). However, when the variant callers analyzed the different real cancer samples, VarScan2 and MuTect2 performed very similarly for a malignant pediatric brain tumor in the cerebellum (MB) and a metastatic melanoma cell line (COLO829). However, in the other two samples (AML and CLL), VarScan2 had a slightly higher F1 Score for acute myeloid leukemia (AML), whilst it was higher for MuTect2 for chronic lymphocytic leukemia (CLL) (Figure 7). Although the use of real biological data has some advantages, a major disadvantage to using it is the difficulty of defining the true number of somatic mutations in the cancer samples, which may cause the calculations of true false positive and false negative rates, and therefore the F1 score, to be inaccurate.

The treatment of the sample in the variant calling process, especially the alignment of the sequenced reads, can influence substantially the subsequent variant calling. Since mostly only two aligner tools, NovoAlign (13) and MarkDuplicates (17) were used, it might negatively impact the results, and a comparison of aligned data with different aligner tools might improve the veracity of the data. However, both NovoAlign (13) and MarkDuplicates (17) have repeatedly been shown to be one of the best performing alignment tools available, as it is highly accurate.

It was hypothesized that MuTect2 would perform consistently better than VarScan2 across cancer samples, since it uses the most common probabilistic framework, that of Bayesian statistics. However, it can be concluded that generally, VarScan2 performs better than MuTect2 only when there is a high mutation frequency, whilst MuTect2 is more consistent in a larger range of mutation frequencies. In relation to the performance of the variant callers in real cancer samples, there is no clear preference, since there is not a distinguishing pattern of which one is more effective, although they usually perform similarly.

Despite there having been previous studies on multiple variant callers, such as the publication “SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach.” by Wang et al. (1), none of them exclusively compare VarScan2 and MuTect2 directly. Usually, but not exclusively, in early stages of cancer, there is a lower mutation frequency, so therefore these samples would be preferable to be analyzed with MuTect2, since MuTect2 performed better in low frequencies than VarScan2. However, if the mutation frequency is known to be very high (70% to 100% of the sample), VarScan2 would be recommended. If the mutation frequency is unknown, MuTect2 is recommended, since it is more reliable over a larger range of mutation frequencies. It is difficult to select which one of these two variant callers would work best, and further investigation would be required to determine how different types of cancer affect the variant calling of mutations.

MATERIALS AND METHODS

Acquisition of Data

The datasets used in the results were downloaded from “SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach.” by Wang et al. (1) based on BAM files for synthetic data set1, set2, set3, set4 and the somatic truth sets from the ICGC-TCGA DREAM Mutation Calling challenge (27), tumor samples from chronic lymphocytic leukemia (CLL) and a malignant pediatric brain tumor in the cerebellum (MB) (28), acute myeloid leukemia (AML) (29), and metastatic melanoma cell line (COLO829) (30) and raw FASTQ files of NA12878-NA11840 dilution WES series datasets (31). Wang et al. trimmed the FASTQ files using the Trimmomatic program (v0.36) (33) to 36 bp and then aligned reads to the hg19 reference genome using the Novoalign software (v3.00.05) (13). Duplicate reads were removed using the MarkDuplicates module of the Picard software (v1.126; 17). Additionally, their analysis uses only properly aligned read pairs, in the sense that the two ends of each pair must be mapped to the reference genome in complementary directions and must reflect a reasonable fragment length (300 ± 100 bp). The high-quality alignments for each sample were further refined according to a local realignment strategy around known and novel sites of insertion and deletion polymorphisms using the RealignerTargetCreator, IndelRealigner and BQSR modules from the Genome Analysis Toolkit (GATK v3.8.1). Thus, the BAMs files were ready for somatic calling. For the DREAM, AML and COLO datasets, they used the downloaded BAMs for somatic calling. (1) They then performed somatic variant calling with Mutect2 and VarScan2 on these datasets.

Data Analysis

The data was analyzed using python (python3), numpy (numpy 1.24.0) pandas (1.5.2), seaborn (0.12.1) and matplotlib (3.6.2). The graphs were created using the program seaborn with data imported through pandas from an excel sheet that contained the data acquired from the publication “SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach.” by Wang et al. (1). For the confusion matrix, the true positives were calculated as the SNV count subtracted by the extra count and the false negative rate was calculated as missing counts divided by the sum of true positives and missing count.

REFERENCES

1. Wang M, et al. “SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach.” Nature Scientific Reports, no.12898, Jul. 2020, doi:10.1038/s41598-020-69772-8.
2. Fjaer R, et al. “A novel somatic mutation in GNB2 provides new insights to the pathogenesis of Sturge-Weber Syndrome” Hum Mol Genet., Oct 2022, doi: 10.1093/hmg/ddab144.
3. Wang Q, et al. “Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data.” BMC Med Genomics, published online, Dec. 2019, doi:10.1186/s12920-019-0636-y.
4. Rabbani B, Tekin M, Mahdieh N. “The promise of whole-exome sequencing in medical genetics.” Journal of Human Genetics, vol. 59; 5-15, Jan 2014, doi:10.1038/jhg.2013.114.
5. van Dijk E L, et al. “Ten Years of Next-Generation Sequencing Technology.” Trends in Genetics, vol. 20, 9, 2014, doi:10.1016/j.tig.2014.07.001.
6. “Mutect2”. GATK. gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2. Accessed 9 Nov 2022.
7. Koboldt DC, et al. “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.” Genome Research, vol. 22:568–76, Feb. 2012, doi:10.1101/gr.129684.111.
8. Kim S, et al. “Strelka2: fast and accurate calling of germline and somatic variants.” Nature Methods, vol. 15:591–4, July 2018, doi:10.1038/s41592-018-0051-x.
9. Kim S, et al. “Virmid: accurate detection of somatic mutations with sample impurity inference.” Genome Biol, vol. 14: R90, Aug 2013, doi:10.1186/gb-2013-14-8-r90.
10. Uğur Sezerman O, et al. “Bioinformatics Workflows for Genomic Variant Discovery, Interpretation and Prioritization.” In book: Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations. Published by Intech Open, Jun 2019, doi: 10.5772/intechopen.85524.
11. Li H, Durbin R. “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics*, Volume 25, Issue 14, July 2009, doi:10.1093/bioinformatics/btp324.
12. Langmead B, Salzberg SL. “Fast gapped-read alignment with Bowtie 2.” Nature Methods, vol. 9; 357–9, Mar 2012, doi:10.1038/nmeth.1923.
13. “Novoalign: Powerful tool designed for mapping of short reads onto a reference genome from Illumina, Ion Torrent, and 454 NGS platforms.” Novocraft. www.novocraft.com/products/novoalign/. Accessed 1 Dec 2022.
14. Marçais G, et al. “MUMmer4: A fast and versatile genome alignment system.” PLoS Comput Biol, vol. 14:e1005944, Jan 2018, doi:10.1371/journal.pcbi.1005944
15. “SAM file format”. Metagenomics. www.metagenomics.wiki/tools/samtools/bam-sam-file-format. Accessed 1 Dec 2022.

16. “BAM File Format.” Illumina. support.illumina.com/help/BS_App_RNASeq_Alignment_OLH_1000000006112/Content/Source/Informatics/BAM-Format.htm. Accessed 1 Dec 2022.
17. “Picard.” Broadinstitute Github. broadinstitute.github.io/picard/. Accessed 1 Dec 2022.
18. “GATK.” Broadinstitute. gatk.broadinstitute.org/hc/en-us. Accessed 1 Dec 2022.
19. McKenna A, et al. “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.” Genome Research. vol 20:1297–303, Jul 2010, doi:10.1101/gr.107524.110.
20. “VarScan.” VarsScan Sourceforge. varscan.sourceforge.net/. Accessed 1 Dec 2022.
21. Larson DE, et al. “SomaticSniper: identification of somatic point mutations in whole genome sequencing data.” Bioinformatics, vol 28:311–7, Feb 2012, doi: 10.1093/bioinformatics/btr665.
22. Li H, et al. “The Sequence Alignment/Map format and SAMtools.” Bioinformatics, vol 25:2078–9, Aug 2009, doi: 10.1093/bioinformatics/btp352.
23. Fang LT, et al. “An ensemble approach to accurately detect somatic mutations using SomaticSeq.” Genome Biol, vol. 16:197, 2015, doi: 10.1186/s13059-015-0758-2.
24. Ramos AH, et al. “Oncotator: cancer variant annotation tool.” Human Mutation, vol. 36: E2423–9, Apr 2015, doi: 10.1002/humu.22771.
25. Douville C, et al. “CRAVAT: cancer-related analysis of variants toolkit.” Bioinformatics, vol. 29:647–8, Mar 2013, doi: 10.1093/bioinformatics/btt017.
26. Benjamin DI, et al. “Calling Somatic SNVs and Indels with Mutect2.” Cold Spring Harbor Laboratory. Dec 2019, doi: 10.1101/861054.
27. “ICGC-TCGA DREAM Mutation Calling challenge.” Bionetworks S. www.synapse.org/#!Synapse:syn312572/wiki/58893. Accessed 20 Dec 2022.
28. Alioto TS, et al. “A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing.” Nature Communications, vol. 6:1000, Dec 2015, doi: 10.1038/ncomms10001.
29. Griffith M, et al. “Optimizing cancer genome sequencing and analysis.” Cell Systems, vol. 1:210–23, Sep 2015, doi:10.1016/j.cels.2015.08.015.
30. Craig DW, et al. “A somatic reference standard for cancer genome sequencing.” Scientific Reports Nature, vol. 6:24607, Apr 2016, doi: 10.1038/srep24607.
31. “Overview – precisionFDA.” Precision FDA. precision.fda.gov/. Accessed 20 Dec 2022.
32. Yu G, et al. “Whole-Exome Sequencing of Nasopharyngeal Carcinoma Families Reveals Novel Variants Potentially Involved in Nasopharyngeal Carcinoma.” Scientific Reports Nature, vol. 9:9916, Jul 2019, doi: 10.1038/s41598-019-46137-4.
33. Bolger AM, Lohse M, Usadel B. “Trimmomatic: a flexible trimmer for Illumina sequence data.” Bioinformatics, vol. 30, no. 2114–20, Aug 2014, doi:10.1093/bioinformatics/btu170.
34. Lähnemann D, et al. “Accurate and scalable variant calling from single cell DNA sequencing data with ProSolo.” Nature Communications, vol. 12, no. 6744, Nov 2021, doi: 10.1038/s41467-021-26938-w.

35. Krøigård AB, et al. “Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data.” PLoS One, vol. 11:e0151664, Mar 2016, doi: 10.1371/journal.pone.0151664.
36. “FastQC A Quality Control tool for High Throughput Sequence Data.” *Babraham Bioinformatics*. www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 18 Feb 2023.
37. Shirley M. “fastqp: Simple FASTQ quality assessment using Python.” *Github*. github.com/mdshw5/fastqp. Accessed 18 Feb 2023.
38. “NGS QC Toolkit: a toolkit for quality check and filtering of next generation sequencing data of Roche and Illumina technology”. *NGSQCToolkit*. www.nipgr.res.in/ngsqctoolkit.html. Accessed 18 Feb 2023.
39. Tammi MT. “PRINSEQ”. *Bioinformatics*. bioinformaticshome.com/tools/rna-seq/descriptions/PRINSEQ.html#gsc.tab=0. Accessed 18 Feb 2023.
40. Zhou Q, et al. “QC-Chain: fast and holistic quality control method for next-generation sequencing data.” *PLoS One*, vol. 8:e60234, Apr 2013, doi:10.1371/journal.pone.0060234.