Qualitative evaluation of two Germline Variant Callers for exome sequence data

Final project report
Algorithms in Structural Biology

26/07/2023

Professor: Theodor Dalamangas

Marina Thalassini Filippidou 7115152200032

Introduction

In our study, we conducted a comprehensive review of the paper titled *"Systematic Comparison of Somatic Variant Calling Performance Among Different Sequencing Depth and Mutation Frequency.*" The paper investigated the performance of two variant calling algorithms, Mutect2 and Strelka2, by analyzing their effectiveness across various mutation frequencies and sequencing depths. The authors employed downsampling techniques to simulate different sequencing depths and mixed normal and tumor files in different proportions to replicate diverse mutation scenarios. The paper's contribution was significant as it highlighted the strengths and weaknesses of both methods, providing valuable insights into the ideal scenarios for their optimal performance.

Originally, our goal was to evaluate the results presented in the paper while expanding our assessment to include additional variant calling tools. However, we encountered obstacles that compelled us to reconsider our approach. Initially, we attempted to use the authors' provided data for somatic variant calling, but encountered persistent errors during the preprocessing steps that we were unable to resolve. Consequently, we opted to change our focus to germline variant calling instead. For this purpose, we selected 20 exome sequencing BAM samples from the 1000 Genomes project and conducted germline variant calling using GATK's Haplotype Caller and Strelka2.

In essence, our study centered on evaluating the performance of the two germline variant callers. To achieve this, we calculated concordance between the two methods and conducted a qualitative analysis of their results. While our initial goal shifted due to the challenges encountered, our research provides valuable insights into the performance of these variant calling tools in the context of germline variant calling for exome sequencing, contributing to the broader understanding of variant calling algorithms in genomic research.

Abbreviations

QUAL: Variant Quality Score. It represents the Phred-scaled quality score assigned to the variant call. The QUAL score indicates the confidence in the variant call, with higher values indicating higher confidence.

AF: Allele Frequency. It represents the frequency of the alternate allele in the population. The AF value represents the proportion of individuals carrying the alternate allele at a particular genomic position.

DP: Depth of Coverage. It represents the total depth or number of reads covering a variant position. The DP value indicates the sequencing depth or the number of times a specific position has been sequenced.

GQ: Genotype Quality. It represents the quality of the called genotype for each sample. The GQ value is assigned to each genotype call and indicates the confidence in the assigned genotype, with higher values indicating higher confidence.

QD: Quality by Depth is a variant annotation that represents the variant quality score normalized by the depth of coverage at the variant site.

Background and related work

Variant calling is a crucial bioinformatics process used to identify genetic variations or differences in DNA sequences between an individual's genome and a reference genome. These variations can include single-nucleotide polymorphisms (SNPs), insertions, deletions, and structural changes. Variant calling plays a vital role in genomic research and clinical applications as it helps scientists and clinicians understand the genetic basis of traits, diseases, and individual differences.

The distinction between somatic and germline variant calling is essential in understanding the genetic makeup of an individual. In germline variant calling, the reference genome is the standard for the species of interest, and the focus is on identifying heritable genetic variations present in the germline cells (e.g., sperm and egg cells). Most genomes are diploid, meaning individuals inherit two copies of each chromosome, and homozygous or heterozygous genotypes are expected at most loci. The exception is the sex chromosomes in male mammals, which are hemizygous.

On the other hand, somatic variant calling utilizes a related tissue from the same individual as the reference. It aims to detect genetic variations that arise after conception, occurring in somatic cells, leading to mosaicism between different cells in the same person. These somatic mutations can be crucial in understanding the development of diseases such as cancer and their potential therapeutic targets.

Given the importance of variant calling in diverse research and clinical settings, numerous variant calling tools have been developed, each with its strengths and limitations. The choice of which tool to use can significantly impact the accuracy and reliability of the results. Therefore, there is a pressing need to evaluate these variant calling tools comprehensively, considering their performance across different scenarios and datasets. Such evaluations help potential users understand the strengths and weaknesses of each tool, enabling them to make informed decisions about when and how to use specific variant calling methods for their specific research or clinical applications. By understanding the performance and limitations of these tools, researchers and clinicians can ensure the reliability of their genomic analyses and draw more accurate conclusions from the data.

Method

To streamline our analysis, we made use of GATK's bundle, which provided us with essential files including the reference genome. As part of our workflow, we initially transformed our BAM files into FASTA files. This conversion allowed us to align the reads once again with the selected reference genome. By aligning the transformed files with the reference genome, we ensured consistency and accuracy in our subsequent variant calling analysis.

To execute Haplotype Caller, we diligently followed all the steps outlined in GATK's Best Practices for Germline short variant discovery (SNPs + Indels), which are designed to optimize the variant calling process for SNP's and Indels of around 50 bases length and can be found in gatk.broadinstitute website. The workflow generally involves two to three analysis phases, starting with data pre-processing. In this initial phase, we transformed the raw sequence data (fasta) into analysis-ready BAM files by aligning them to a reference genome and performing necessary data cleanup to correct technical biases and ensure suitability for analysis. In the next phase, variant discovery phase, we utilized the analysis-ready BAM files to identify genomic variations in the individual(s). This step involved employing the HaplotypeCaller, which excels in calling both SNPs and indels simultaneously through local de-novo assembly of haplotypes in active regions.The output consisted of variant calls in VCF format. Additional steps, including filtering and annotation, were implemented to produce a callset ready for downstream genetic analysis, leveraging resources of known variation, truthsets, and metadata to enhance result accuracy and provide additional information.

We also performed variant calling using Strelka2. Strelka is a powerful variant calling tool that employs a systematic workflow for germline and somatic variation analysis. It starts by estimating genomic statistics from input alignments, followed by segmenting the genome for parallel processing. Within each segment, candidate alleles are analyzed jointly, leading to variant inferences and recalibration. Strelka's germline model uses haplotype representation for accurate variant calls and read-backed phasing. It can detect SNVs and indels up to 49 bases. Strelka's versatility and accuracy make it a valuable tool for genomic analyses. To use it for exome sequencing data we had to include the flags --exome and --callRegions known_intervals.bed.gz. Nevertheless we run the algorithm two times, one including the specific flags and once without in order to compare the difference.In somatic calling, it is recommended to provide Manta indel candidates to improve Strelka's sensitivity to call long (e.g. >20) indels but this is not necessary in germline calling.

In addition to evaluating two Germline Variant Callers for exome sequence data, we considered DeepVariant, a powerful deep learning-based variant caller. DeepVariant employs a unique approach, taking aligned reads in BAM or CRAM format and generating pileup image tensors from this data. A convolutional neural network (CNN) is then utilized to classify each tensor, enabling DeepVariant to accurately identify genetic variants. However, it is important to note that DeepVariant operates on a one-sample-per-run basis, which rendered it unsuitable for our specific multisample analysis. Despite this limitation, we acknowledged the potential value of

DeepVariant's results and opted to retain the variant outcomes for each individual sample obtained from this variant caller for all individual samples for demonstration purposes and to have the option for further research needs.

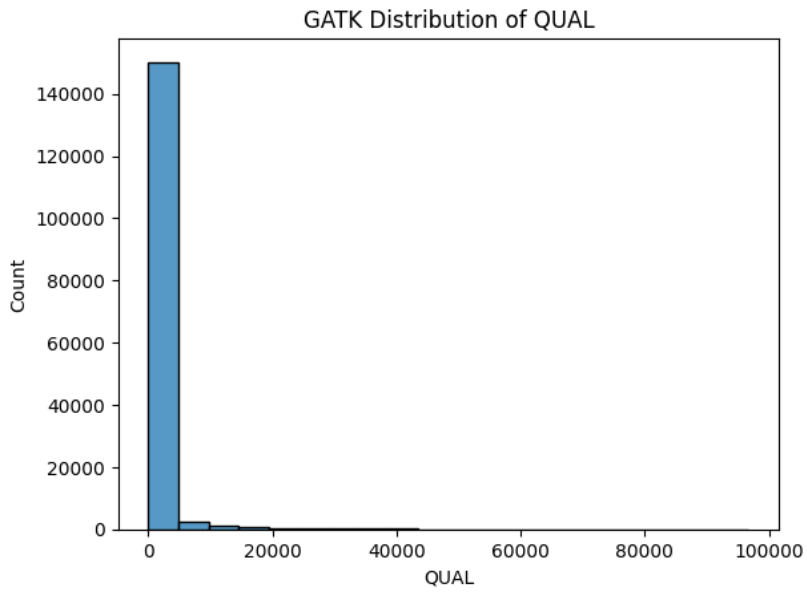
Evaluation

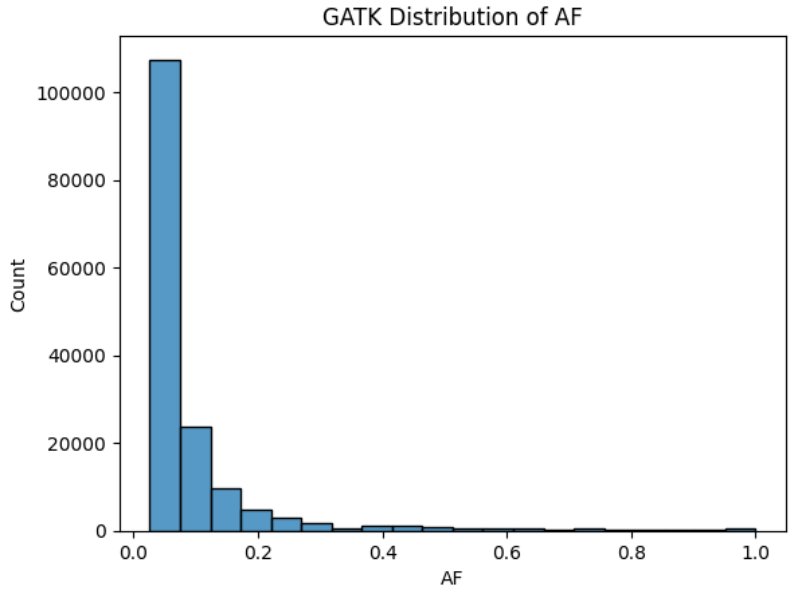
We conducted a comprehensive evaluation of the outcomes generated by the two variant callers, employing both qualitative analysis and quantitative measurements of concordance between the two methods. The table below showcases the obtained concordance values, providing a clear and concise overview of the agreement between the variant calling approaches.

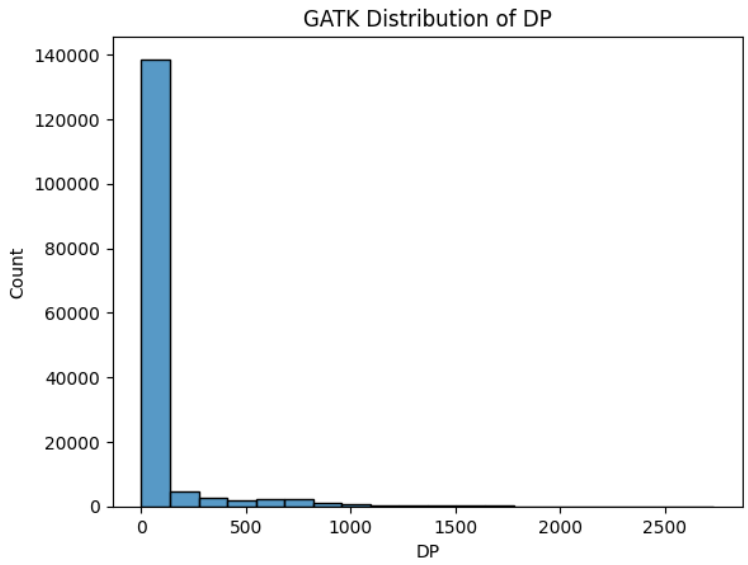
	% of intersection with Haplotype Caller
Strelka 2 (exome specific)	88%
Strelka 2	87%

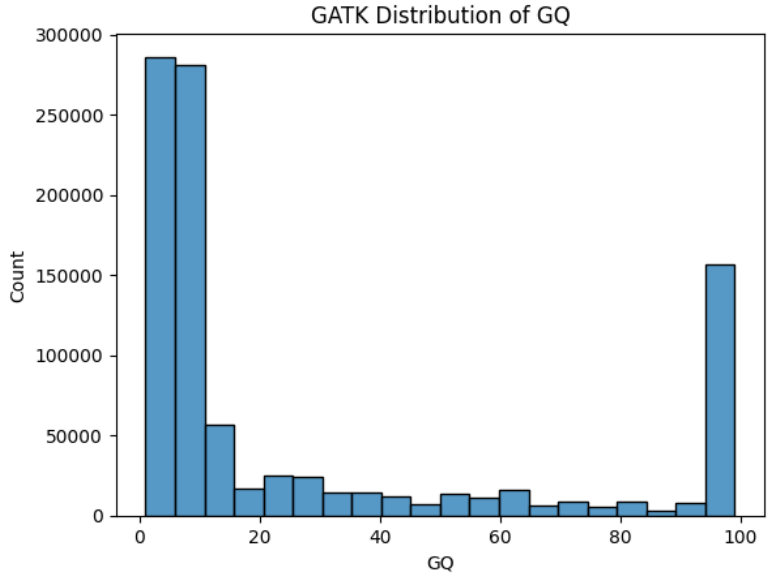
Our observations reveal a substantial and consistent percentage of intersection between Strelka and Haplotype Caller, with this overlap becoming even more pronounced when utilizing the flags specifically tailored for exome sequencing. The increased agreement between these two variant callers when applied to exome data underscores their compatibility and effectiveness in identifying shared genetic variants. This finding highlights the importance of selecting appropriate variant calling parameters, particularly when working with targeted sequencing data, to enhance the concordance and reliability of variant detection.

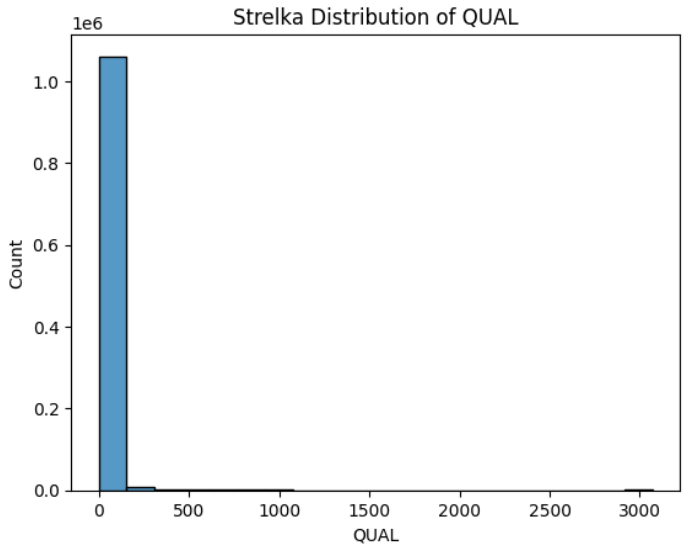
We also utilized the pyvcf library in order to visualize some of the statistics of our final vcfs. Below we can observe the graphs of the available metrics for gatk and strelka2 (exome specific):

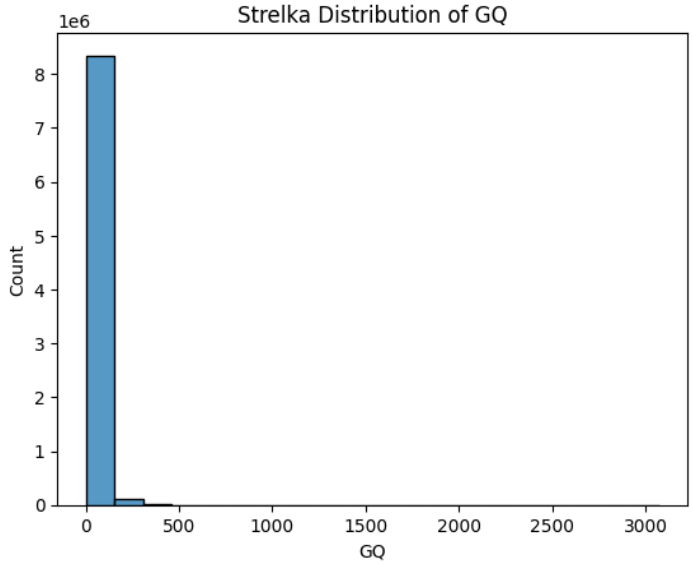

GATK Distribution of QUAL
Count
QUAL
0
20000
40000
60000
80000
100000
120000
140000


GATK Distribution of AF
Count
AF
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0


GATK Distribution of DP
Count
DP
0
20000
40000
60000
80000
100000
120000
140000
0
500
1000
1500
2000
2500


GATK Distribution of GQ
Count
GQ
0
50000
100000
150000
200000
250000
300000
0
20
40
60
80
100





An intriguing finding reveals that, even when the quality scores of the variants are equivalent between the two variant callers, Strelka consistently identifies variants with higher Genotype Quality compared to GATK.

One notable limitation of the Strelka variant calling algorithm is the absence of essential variant-level metrics, including DP (Read Depth), AF (Allele Frequency) and potentially other useful measures, in its final VCF output. As a consequence, to gain access to these critical metrics and obtain a comprehensive understanding of the called variants, additional post-processing and manipulation of the VCF data are necessary. We regard this as a significant disadvantage of the Strelka pipeline, as it adds complexity and extra effort to the analysis workflow, compared to variant callers that provide more comprehensive and informative

VCF outputs directly. Acknowledging and addressing this limitation is crucial for a thorough and fair evaluation of the variant calling methods used in our study.

Furthermore, it is important to note that Strelka does not perform variant annotation as part of its standard output. To obtain functional annotations and additional information about the variants, users must employ other annotation tools or databases to supplement the Strelka VCF output. This requirement for external annotation tools adds an additional step in the analysis process and may introduce complexities in interpreting the biological significance of the identified variants.

Taken together, researchers using Strelka should be aware of these limitations and consider incorporating supplementary tools for both variant-level metrics and variant annotation to achieve a more comprehensive analysis of the called variants in their study. Being aware of these considerations will help ensure a more robust and informed evaluation of the variant calling methods and their implications in the research.

We also evaluate the two algorithms in terms of computational time:

Algorithm	Computational time (min)
GATK	51
Strelka (-j 8)	49
Strelka (-j 24)	27

Our observations reveal that Strelka demonstrates superior computational performance compared to GATK's Haplotype Caller, particularly when scaling up the number of jobs using the -j option. As we increased the number of parallel jobs, Strelka exhibited notable advantages in terms of computational efficiency, completing the variant calling process in a significantly shorter time compared to GATK's Haplotype Caller. This finding underscores Strelka's ability to efficiently handle high-throughput data and parallel processing, making it a favorable choice for scenarios where computational speed is a critical consideration. The substantial reduction in computational time with Strelka highlights its potential to streamline large-scale variant calling tasks and expedite the overall analysis process, making it a valuable tool for researchers and practitioners dealing with extensive genomic datasets.

Conclusion

In our research, we conducted an extensive and rigorous evaluation of two germline variant calling algorithms specifically designed for exome sequencing data. The evaluation encompassed various aspects, including concordance analysis, qualitative assessment of VCF outputs, and computational efficiency.

Our analysis revealed a remarkable concordance rate of 88%, signifying a substantial level of agreement between the two variant calling methods under consideration. Notably, Strelka2 stood out as the top-performing algorithm, exhibiting a noteworthy computational advantage by completing tasks approximately 1.8 times faster than Haplotype Caller—a significant asset for large-scale data processing.

Furthermore, our analysis demonstrated that even when the quality scores of the variants were equal between the two variant callers, Strelka consistently identified variants with higher Genotype Quality compared to GATK. This finding indicates that Strelka tends to excel in identifying variants with high-quality calls, suggesting its proficiency in accurately detecting variants under conditions of elevated confidence.

Lastly, it's important to note that Strelka2 lacks some important metrics in its output and does not perform variant annotation. This limitation can add complexity to the analysis process, as researchers may need to use additional tools beyond Strelka2 to obtain the essential information required for further analysis.

In summary, our findings present valuable insights into the strengths and trade-offs of these two germline variant calling algorithms tailored for exome sequencing data analysis. The research contributes to a deeper understanding of their respective capabilities, empowering researchers to make informed decisions when selecting the most suitable variant calling tool based on the specific requirements of their studies.

References

Chen, Z., Yuan, Y., Chen, X. *et al.* Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep* **10**, 3501 (2020). https://doi.org/10.1038/s41598-020-60559-5

Wang, Q., Kotoula, V., Hsu, PC. *et al.* Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data. *BMC Med Genomics* **12** (Suppl 9), 181 (2019). https://doi.org/10.1186/s12920-019-0636-y

https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/

https://github.com/Illumina/strelka

https://www.nature.com/articles/s41592-018-0051-x
https://github.com/google/deepvariant
https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-