BMC Bioinformatics

RESEARCH

Open Access

Performance analysis of conventional and AI-based variant callers using short and long reads

Omar Abdelwahab[1,2,3,4], François Belzile[1,2,3] and Davoud Torkamaneh[1,2,3,4*]

*Correspondence:
davoud.torkamaneh.1@ulaval.ca

[1] Département de Phytologie, Université Laval, Québec, Canada
[2] Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Canada
[3] Centre de recherche et d'innovation sur les végétaux (CRIV), Université Laval, Québec, Canada
[4] Institut intelligence et données (IID), Université Laval, Québec, Canada

Abstract

Background: The accurate detection of variants is essential for genomics-based studies. Currently, there are various tools designed to detect genomic variants, however, it has always been a challenge to decide which tool to use, especially when various major genome projects have chosen to use different tools. Thus far, most of the existing tools were mainly developed to work on short-read data (i.e., Illumina); however, other sequencing technologies (e.g. PacBio, and Oxford Nanopore) have recently shown that they can also be used for variant calling. In addition, with the emergence of artificial intelligence (AI)-based variant calling tools, there is a pressing need to compare these tools in terms of efficiency, accuracy, computational power, and ease of use.

Results: In this study, we evaluated five of the most widely used conventional and AI-based variant calling tools (BCFTools, GATK4, Platypus, DNAscope, and DeepVariant) in terms of accuracy and computational cost using both short-read and long-read data derived from three different sequencing technologies (Illumina, PacBio HiFi, and ONT) for the same set of samples from the Genome In A Bottle project. The analysis showed that AI-based variant calling tools supersede conventional ones for calling SNVs and INDELs using both long and short reads in most aspects. In addition, we demonstrate the advantages and drawbacks of each tool while ranking them in each aspect of these comparisons.

Conclusion: This study provides best practices for variant calling using AI-based and conventional variant callers with different types of sequencing data.

Keywords: Genomics, Sequencing, Variant calling, NGS, Artificial intelligence

Background

Genetic variations are defined as any changes in the DNA sequence of individuals within or between species [1]. Next-generation sequencing (NGS) technologies (i.e., second and third generation) have revolutionized the field of genomics by allowing researchers to decode the whole genome of many organisms and genotype very large numbers of genetic variations such as single nucleotide variants (SNVs) and short insertions/deletions (INDELs) [2]. To detect genetic variations, many computational tools, known as

variant calling tools or variant callers [3], have also been developed to efficiently identify thousands to millions of variants from sequencing reads aligned against a reference genome. This has been proven to be indispensable in the various areas of genomic research, from agriculture to environment to human health [4–6].

In the last two decades, short reads mainly derived from Illumina sequencing technologies have been the predominant data type used in various variant calling studies [7, 8]. Even though short reads provide a high base-level accuracy score, they usually fail to align unambiguously in repetitive regions [9]. While long reads can overcome the challenges posed by repetitive regions, they were not considered suitable for variant calling because of their higher rate of sequencing errors. However, in 2019, Pacific Biosciences (PacBio) introduced a single-molecule real-time (SMRT) sequencing platform that can generate high-fidelity (HiFi) long reads with an average length of 13.5 kilobases (kb) using a Circular Consensus Sequence (CCS) approach. In this approach, a single DNA molecule is circularized, and this template is sequenced multiple times. The resulting consensus provides a sequence with high base-level accuracy (~99.9%) [10]. Accordingly, HiFi data were used for the detection of genetic variants [10]. In the PrecisionFDA challenge (Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions) in 2020, HiFi technology surpassed other sequencing technologies in detecting variants in terms of both precision and recall [11].

Meanwhile, Oxford Nanopore Technologies (ONT) has changed the sequencing paradigm by introducing sequencers that are portable with real-time data delivery and are able to generate ultra-long reads [12]. This technology may look promising for variant calling due to its ability to sequence difficult-to-map regions and read-based phasing, but it has been problematic to achieve a highly accurate analysis because of the error profiles generated by the unique pore-based signal [13]. Nonetheless, recent advances in the development of variant calling tools based on artificial intelligence (AI) (e.g., PEPPER-Margin-DeepVariant) [14] demonstrate that highly accurate variant calling can be achieved from ONT data [14]. Yet, this does not necessarily mean that other variant calling approaches have the ability to detect variants using ONT data.

Over the past few years, besides the advancement of sequencing techniques, many variant calling tools have been developed and used in various genomic projects. For example, the Genome Analysis Toolkit (GATK) [8], developed by the Broad Institute, had been used to detect variants from 180 K samples in "The Trans-Omics for Precision Medicine" (TOPMed) program [15]. However, DeepVariant (an AI-based variant caller [16] developed by Google) was selected to detect variants among more than 500 K samples by the UK Biobank WES consortium [17], and DRAGEN-GATK [18] was used to genotype more than 1 million samples from the National Institutes of Health's All of Us Research Program [19]. Despite rapid advances in sequencing technologies and bioinformatics, accurately calling genetic variants from billions of short or error-prone long sequence reads remains challenging. State-of-the-art variant callers use a variety of statistical techniques to distinguish real genetic variants from errors in the reads. However, generalizing these tools to different data types derived from different sequencing technologies has proven difficult. Hence, to date, different variant callers have been used in different NGS-based studies in various species and thus far, it is still challenging to determine which variant calling tool is the best to use. Over the years, various studies

have been conducted to compare and evaluate the performance of different variant callers [20–24]. However, all these studies used only short-read sequencing data in their analyses.

In this work, we are addressing the question of whether there is an advantage of a specific variant calling tool over others using a different type of sequencing data (e.g., short vs. long reads). Most of the conventional variant calling tools have been developed and widely used for short-read analysis. However, now with the progress in generating high-quality long reads and the emergence of AI-based variant calling tools, there has been an intriguing question about their potential to supersede conventional ones for calling SNVs and INDELs using both long and short reads. Here, we used three different data types (PacBio HiFi, Illumina, and ONT) for the same set of samples from the Genome In A Bottle (GIAB) Consortium to test five variant callers, two of which are AI-based, in terms of accuracy and computational cost.

Results

Illumina variant calling performance

SNV performance

As can be seen in Fig. 1A, DNAscope achieved the highest recall performance (an average of 95.35%). This was ~2% more than its closest competitor (DeepVariant) and ~11% more than Platypus, which had the lowest recall performance (84.95%). In terms of precision, DeepVariant (98.95%), Platypus (98.49%), and BCFTools (98.83%) were almost indistinguishable, while DNAscope showed the lowest performance (94.48%). Finally, DeepVariant showed the highest F1-score (96.07%), with a very close performance of BCFTools (95.67%), while Platypus achieved the lowest performance (91.19%).

INDEL performance

DNAscope achieved the highest recall performance (83.60%) with a difference of ~6% to its closest competitor (DeepVariant) and ~22% better than Platypus, which displayed the lowest recall performance (61.17%; Fig. 1B). As for precision, Platypus achieved the highest performance (93.53%), while DNAscope had the poorest performance (44.78%), showing a significant difference. The F1-score performance was almost the same for DeepVariant (81.41%) and BCFTools (81.21%), while DNAscope showed the poorest performance (57.53%).

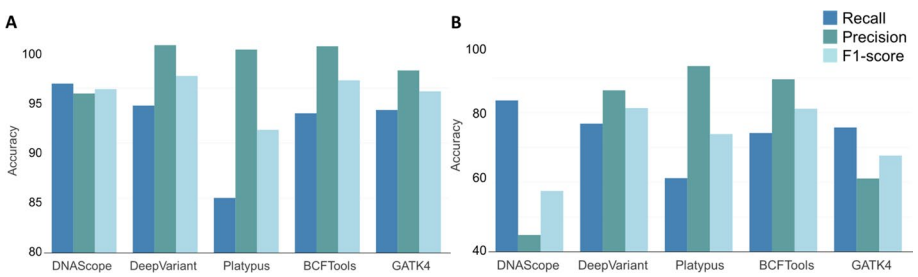

Fig. 1 Average accuracy metrics of variants (SNVs (**A**) and INDELs (**B**)) called from Illumina data using five different variant callers

PacBio HiFi variant calling performance

SNV performance

As shown in Fig. 2A, DeepVariant and DNAscope demonstrated impressive close-to-perfect performances in all accuracy metrics (> 99.9%), although all variant calling tools achieved very high performances (> 99%) in all cases. The differences among tools were almost indistinguishable as the differences were less than 1% in precision, recall, or F1-score.

INDEL performance

Similarly, DeepVariant and DNAscope achieved the highest performances (> 99.5%) in all accuracy metrics. For recall, they were ~ 9% more than the closest tool (GATK4) and ~ 10% more than BCFTools. In contrast, both BCFTools and GATK4 had a significantly lower precision (< 81%). Again, BCFTools and GATK4 saw a significant drop in F1-score, both scoring below 85% (Fig. 2B).

ONT data variant calling performance

To date, BCFTools and DeepVariant (PEPPER-Margin-DeepVariant pipeline) are the only variant callers (out of the five variant callers used in this study) that can handle ONT data. The majority of the SNVs (97.07%) were detected by both DeepVariant and BCFTools. On the other hand, BCFTools failed to detect any INDELs, while DeepVariant had 80.40% in common with the truth sets. Both tools showed a high number of private variants (variants that do not exist in the truth sets) that may be attributed to the quality of ONT sequencing data, resulting in lowering the accuracy metrics even though there is a high number of common variants. In calling SNVs and INDELs, DeepVariant showed a clear advantage over BCFTools in terms of recall, precision, and F1-score (Fig. 3).

Computational cost of variant calling

As shown in Fig. 4, Platypus, DNAscope, and BCFTools proved to be the fastest running tools among the different variant callers (0.34 h, 11.66 h, and 7.98 h, respectively) for Illumina, PacBio HiFi, and ONT, respectively, whereas GATK4 proved to be the slowest for Illumina and PacBio HiFi requiring 44.19 h, and 102.83 h, respectively, and DeepVariant was the slowest for ONT data as it required 105.22 h.

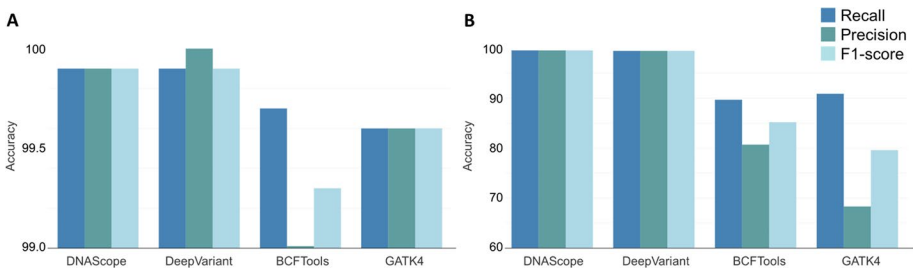

Fig. 2 Average accuracy metrics of variants (SNVs (**A**) and INDELs (**B**)) called from PacBio HiFi data using four different variant callers

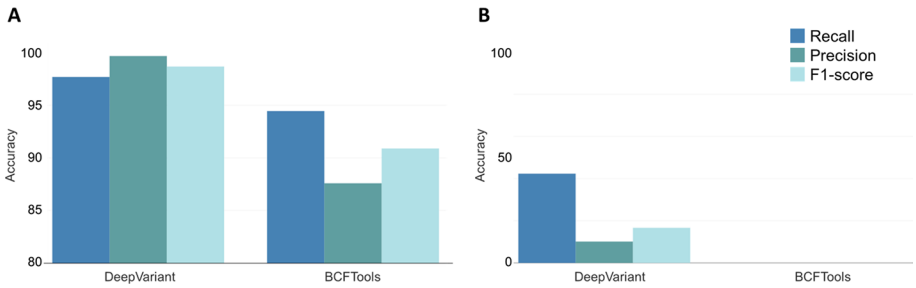

Fig. 3 Accuracy metrics of variants (SNVs (**A**) and INDELs (**B**)) called from ONT data using BCFTools and DeepVariant with the sample HG003.

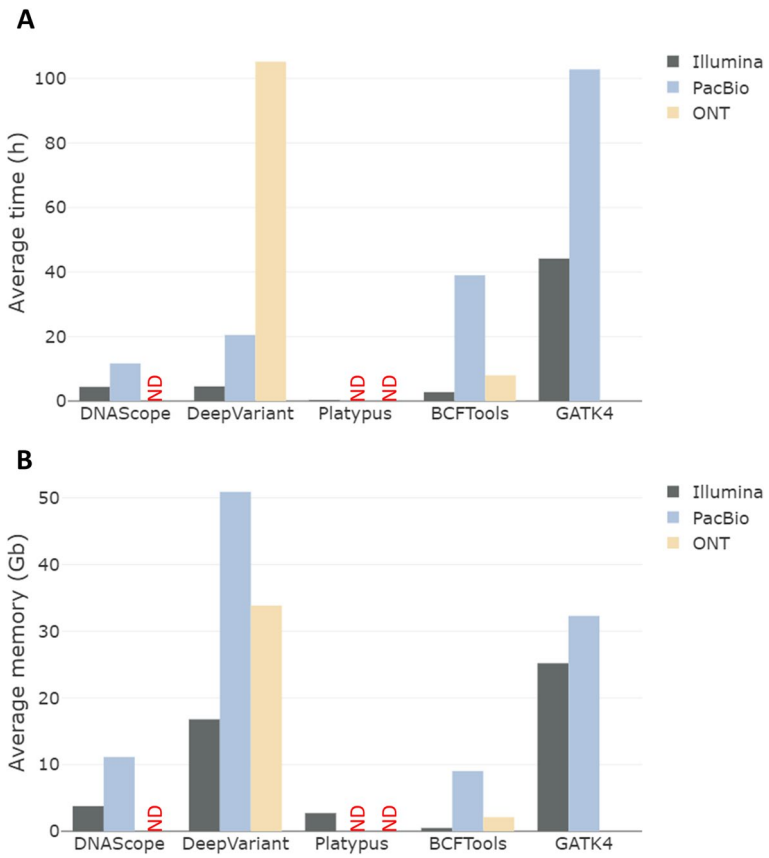

Fig. 4 Average computational cost (time (**A**) and memory (**B**)) for variant calling using three different data types: Illumina, PacBio HiFi, and ONT. ND: not determined

In terms of memory required, here also, very large differences were observed. BCFTools proved to be the most memory-efficient tool, requiring 0.49, 9.03, and 2.12 GigaBytes (Gb) to carry out the analyses using Illumina, PacBio HiFi, and ONT data, respectively. GATK4 showed the highest memory usage to process both Illumina and PacBio HiFi data, while DeepVariant was the slowest to process ONT data. However, PacBio and ONT consume much more memory, especially when using DeepVariant, where it reached 50.89 Gb and 33.85 Gb for PacBio and ONT data, respectively.

Discussion

Variant calling performance

For years, Illumina has been the benchmark sequencing technology for variant calling despite its difficulty in detecting variants from genomic regions that are considered difficult-to-map by short-read sequencing [25]. However, now with the emergence of long-read sequencing technologies (i.e., PacBio HiFi and ONT), there is a crucial inquiry of whether the conventional variant calling methods will function comparably. Furthermore, there has been a recent emergence of AI-powered variant callers, and researchers are now keen to investigate whether these tools will surpass conventional ones. This study focuses on these concerns by examining various variant callers by utilizing GIAB benchmark data obtained from diverse sequencing technologies [26].

In the context of variant caller comparisons, numerous studies have sought to assess various tools, occasionally including comparisons with AI-based tools [2, 27–32]. However, in this study, unlike previous studies that exclusively relied on Illumina data [2, 27–32], we adopted a more extensive perspective by incorporating a wide range of sequencing technologies, encompassing both short-read (Illumina) and long-read (PacBio HiFi and ONT). This comprehensive approach offers insights into the performance of variant calling tools across diverse sequencing platforms. Additionally, our study focuses on evaluating both AI-based and conventional variant calling tools allowing us to thoroughly investigate the advantages and limitations of AI-driven methods when compared to established techniques, resulting in a comprehensive assessment of variant calling strategies.

In this study, in alignment with previous comparative studies [2, 27–30], a high variant calling accuracy was observed using conventional tools with Illumina data. These studies have proven that conventional variant calling tools generate similar results with a variant concordance of 80–90% on average, where most differences correlate to variants of low coverage or low confidence. As for the AI-based tools, DNAscope proved the most powerful in correctly calling known variants found in the truth sets, but it did have the greatest tendency to call false INDELs using Illumina data resulting in a high-recall-low-precision output. The reason for this could be the insufficient coverage of the Illumina sequencing data. However, on the other hand, DeepVariant provided the most balanced calls combining the best scores for all three metrics considered, as well as the highest F1-score on both SNVs and INDELs. These results are consistent with the original publication of DeepVariant that evaluated DeepVariant's methodology against multiple conventional tools including GATK and SAMtools and mentioned that DeepVariant demonstrates > 50% fewer errors per genome [16]. This superior performance was achieved without entailing a much higher computational cost. Moreover, Olson et al. [11] documented that, with higher coverage (35X), better accuracy metrics can be achieved with DNAscope and DeepVariant, in which the harmonic mean of the F1-score for combined INDELs and SNVs can reach 0.999. Thus, the increasing popularity and use of low-coverage sequencing can pose a challenge for these tools.

PacBio HiFi has been the preferred sequencing technology by many sequencing initiative projects such as the Earth BioGenome Project [33, 34], the Vertebrate Genome Project [35], the i5K Initiative [36], and the Ag100Pest Initiative [37]. These projects, among others, have opted for PacBio HiFi technology to produce high-quality reads,

making PacBio HiFi the gold standard for generating high-confidence long-reads [38–40]. Although conventional variant callers have been competitive with AI-based tools in identifying variants using Illumina short reads, our study revealed that the AI-based tools exhibited a clear advantage in calling INDELs using PacBio HiFi data, leading to a high-recall-high-precision output. Although there were no significant differences between the tools in calling SNVs using PacBio HiFi data, the AI-based tools showed slightly better performance. Additionally, the AI-based tools outperformed the conventional ones in terms of time and memory efficiency, with DNAscope demonstrating the highest efficiency. Our findings align with the results of the "PrecisionFDA Truth Challenge V2", where the AI-based tools were the top performers in calling variants in all benchmark regions and difficult-to-map regions from PacBio HiFi data [11]. Previous studies have also shown that using PacBio HiFi data alone could yield equal or better performance to short-read sequencing in all benchmarking regions when calling variants using a single sequencing technology [10, 11]. It should be noted that the study utilized a high sequencing coverage (~40X), therefore it would be valuable to assess the effectiveness and precision of these tools when working with low-coverage data, which is becoming increasingly prevalent.

As for ONT data, previous studies [7, 14, 41, 42] have demonstrated its ability to call genomic variants. In our study, the AI-based variant caller, DeepVariant, showed better results than BCFTools in terms of SNV and INDEL performances using ONT data. However, BCFTools would be a better option in terms of time and memory efficiency when working with ONT data. As documented, it is capable of running on a low to medium-power computer. The results obtained from the variant calling with DeepVariant in this study for SNVs are consistent with the results of recent benchmark studies for ONT data [14, 41, 42]. However, on the contrary, the INDELs results of this study disagree with the original publication of DeepVariant where the authors have reported higher accuracy. This is probably due to the preprocessing step, in which they used raw ONT reads, carried out the alignment with minimap2 [43], and performed phasing and haplottaging [16]. However, here, we performed the standard procedures by running the default code on the acquired BAM files from GIAB directly for generalization between tools, and the possibility of data unavailability in some variant calling projects. A recent AI-based variant caller specifically designed for ONT data, CLAIR3 [44], has shown that it can achieve similar results to DeepVariant in calling variants. However, the "PrecisionFDA Truth Challenge V2" has mentioned DeepVariant as a top performer in calling variants from ONT data, especially from difficult-to-map regions [11]. Moreover, another study has claimed that highly accurate variants (94.25% F1-score) can be called with lower coverage in ONT data [41]. This suggests that ONT data can be used for reliable variant calling, but there is still room for improvement in the accuracy and efficiency of the tools used for this purpose.

Overall, PacBio HiFi and ONT data (long reads) have the ability to compete with Illumina (short reads) in calling genomic variations. Furthermore, utilizing AI-based variant calling tools with both short and long reads can achieve very high accuracy metrics for calling both SNVs and INDELs. Namely, DeepVariant has overall better performance with all data types even with comparatively lower coverage as in the Illumina case. Recent studies have shed light on the fact that combining sequencing technologies

produce better accuracy than any separate sequencing technology [42]. This encourages the production of more AI-based tools that can call variants from multiple technologies at the same time to achieve better results.

User experience

Although it was very smooth to set up all the tools, running them did not give the same experience. For the AI-based tools, the documentation was extremely clear and helpful. Especially when utilizing DNAscope with PacBio HiFi data, all the steps are compacted in a single command line. As a commercial tool, the DNAscope support team was very accessible and easy to reach. As for DeepVariant, it is always a singularity command that can perform all the processes of variant calling no matter the data type. Moreover, both tools do not require filtering variants or setting thresholds manually to refine the results.

On the other hand, conventional tools led to a different experience. Both BCFTools and Platypus are very easy to handle with very clear documentation. However, Platypus is still a Python 2-dependent tool, only works on short reads, and has not been updated since 2014. In contrast, BCFTools has been improved and updated regularly over the years. Platypus includes default values to filter variants, while for BCFTools and GATK4 all the filters need to be set manually. Running GATK requires an in-depth understanding of all the steps and parameters to set manually. Although this gives the user more control over the filtering process, setting thresholds for the filters might be an exhausting and time-consuming process.

The AI-based variant calling tools have an advantage in user time performance due to their automatic filtration feature which makes them less time-consuming overall. The time performance was only calculated for the variant calling step without taking into consideration setting thresholds, filtering, or any other pre-/processing steps.

Limitations

It is essential to clarify that during the timeframe of our research, higher coverage Illumina datasets were not accessible for the GIAB samples. In light of this limitation, we opted to work with the available suboptimal coverage data, which were 10.5X, 13.6X, and 12.6X. Our rationale for this choice was to evaluate the performance of these tools under realistic and potentially challenging scenarios that researchers may encounter when dealing with lower-coverage datasets. Despite the limited coverage in the sequencing data, many of the tools were able to achieve highly accurate variant calling. Nonetheless, prior studies have provided ample evidence that performing whole genome sequencing with greater coverage typically results in more accurate variant calling [45].

It is also important to take into consideration the GPU compatibility across different variant calling tools. Within tested tools, DeepVariant was the only GPU-compatible variant caller. We found that the incorporation of GPU with DeepVariant significantly decreases processing time by more than 50% when using one Tesla P100 GPU, which aligns with previous study [46]. This also indicates the potential advantages of using GPU for GPU-compatible tools and for those who have access to such resources.

Although PacBio HiFi data achieved the best results, the high coverage (~40X) might be the reason behind this advantage over Illumina data (~12X). According to [14], PEPPER-Margin-DeepVariant can achieve better results in calling INDELs when following

specific procedures that include performing phasing and haplotagging. Moreover, we compared the capability of each tool in its default form to mimic the conditions of regular users.

Finally, even with using a truth set that excluded challenging-to-map regions, we noticed variation in the number of called variants (see Additional file 2: Figures S1, S2, and S3) and their accuracy. It was challenging to investigate the causes of these erroneous calls, as there was no discernible pattern among them, and each tool produced a significant number of unique and erroneous calls. We suggest that this may be related to the specific models and algorithms implemented by each tool.

Conclusion

Currently, the long-read data show the potential to become the new standard for variant calling and genotyping. PacBio HiFi introduces low error rates with high base calling quality while having an edge in detecting repetitive regions that are difficult to handle with short-read technologies. Utilizing PacBio HiFi data is now leading to near-optimal SNV and INDEL performance competing with short-read technologies. The long reads are also the optimal technique to detect structural variants allowing now to identify all types of genetic variations with a single sequencing experiment. The only drawback of long-read technologies, which is making it behind short-read technologies, is the cost where Illumina still has the edge of being the cheapest in the market.

Combining with long reads, AI-based tools have demonstrated a clear advantage over conventional tools in calling variants, which paves the road and makes it a starting point for a new era of AI tools in the genomics field. As noticed in this article, AI-based tools do not perform in the same time frame, which might be because of the engineering design of each tool. Being said, this concludes that there is still room for improvements in AI-based tools, where they can even give better performances that might reach the gold standards in the future, achieving less computational cost and more efficacy.

Methods

Sequencing data

The sequencing data (Illumina, PacBio HiFi, and ONT) for three samples (HG003, HG006, and HG007) were obtained from the Genome in a Bottle (GIAB) Consortium [26] from the NIST GIAB FTP site: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/. Only these three samples (out of a total of seven GIAB samples) were used because the other four have been used either to evaluate or train the AI-based variant callers (DeepVariant or DNAscope) [14, 16, 47].

In summary, Illumina data for HG003, HG006, and HG007 were generated on an Illumina HiSeq2500 and resulted in a lower-than-expected sequence coverage (10.5X, 13.6X, and 12.6X, respectively) due to a large amount of PCR duplicates. The raw FASTQ files were aligned using Sentieon BWA-MEM [48, 49], against the GRCh38 reference genome (GCA_000001405.15_GRCh38_no_alt_analysis_set.fna). For the PacBio HiFi data, we downloaded the BAM files directly from the PacBio_CCS_15kb_20kb_chemistry2 directory, as these had been aligned against the same version of the reference genome. In this dataset, the depth of coverage of samples was 42.7X, 40.7X, and 37.6X for HG003, HG005, and HG007, respectively. Finally, ONT data was publicly

available only for HG003. Similarly, we downloaded the BAM file directly as it was mapped against the same reference genome. The sequencing coverage of this sample was 77.7X. In this study, we used SAMtools [50] to calculate the coverage of samples using the '-a' option to consider all positions.

Variant calling

Three conventional variant calling tools (BCFTools [50, 51], GATK v4 [8] and Platypus [52]) as well as two AI-based tools (DNAscope [47] and DeepVariant [16]) were used to call variants from the different sets of sequencing data. Variant calling was performed in the same conditions in terms of computational environment. We used default options to filter out low-quality variants. Namely, we followed the recommended parameters for each data type from the BCFTools documentation page and applied only one filter (Quality >= 20) after variant calling. For GATK, we utilized the GATK4 pipeline [53] to call variants from Illumina and PacBio data. We started from the variant calling step using the BAM files and performed two rounds of variant calling, where we recalibrated the base quality scores after the first variant calling step to produce recalibrated BAM files for the second round of variant calling. Finally, we used the default filter parameters in the pipeline. For Platypus, variant calling and filtering (default parameters) were conducted following the developers' recommendations [52]. Platypus was only able to detect variants from Illumina data. To run DNAscope with Illumina data, we started from raw FASTQ files and used the recommended pipeline from Sentieon that includes alignment and duplicate marking. Then, we used the variant calling pipeline that consists of two steps, phasing and a second pass. However, for HiFi data, the whole pipeline is wrapped in a single one-line command (*dnascope_HiFi.sh)*, as we used the HiFi BAM files directly. As for the ONT data, DNAscope does not have a pipeline for it yet. Finally, for DeepVariant, we followed the documentation on the DeepVariant GitHub repositories to run Illumina, HiFi, and ONT data using the singularity command for each data type.

Variant calling performance analysis

We identified the common variants between tools and the latest GIAB truth sets v4.2.1 [54] with each data type using the hap.py tool [55]. In this study, we used version 4.2.1 of the GIAB truth sets, excluding challenging-to-map regions. This dataset choice aimed to eliminate errors associated with genome composition and mapping in difficult-to-map regions, allowing a pure assessment of tool performance without these concerns. For evaluation, all the tools were compared in terms of precision (P), recall (R), and F1-score (F1). The equation of each accuracy metric can be represented as $P = \frac{T_p}{T_p+F_p}$, $R = \frac{T_p}{T_p+F_n}$, and $F1 = 2\frac{P\times R}{P+R,}$ where T_p, F_p, and F_n stand for true positive, false positive, and false negative, respectively. Here, we presented an average of the performance metrics, however, detailed results for each sample can be found in the Additional file 1: Table S1.

Computational resources and code availability

All the analyses were performed using a Linux system on the Valeria [56] server at Université Laval, QC, Canada. For all variant calling tools, we allowed 16 CPUs and allocated up to 200 GB of RAM to monitor the maximum RAM usage of each tool.

The custom code used for the analysis can be accessed on GitHub at https://github.com/Omar-Abd-Elwahab/Variant_Callers.

Abbreviations

NGS	Next-generation sequencing
SNVs	Single nucleotide variants
INDELs	Insertions/deletions
PacBio	Pacific Biosciences
SMRT	Single-molecule real-time
HiFi	High-fidelity
kb	Kilobases
CCS	Circular Consensus Sequence
ONT	Oxford Nanopore Technologies
AI	Artificial intelligence
GATK	Genome Analysis Toolkit
GIAB	Genome In A Bottle
Gb	GigaBytes
h	Hours

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05596-3.

Additional file 1: Table S1. Detailed results for performance analysis of calling variants from three samples using five variant calling tools with different sequencing technologies (Illuina, Pacbio HiFi, and ONT) in terms of accuracy metrics (precision, recall and F1-score), running time, and memory.

Additional file 2: Fig. S1. Venn diagrams representing the degree of overlap among variants called using five variant callers and Illumina data compared to the GIAB truth sets. A, B, and C represent SNVs, while D, E, and F represent INDELs for HG003, HG006, and HG007, respectively. **Fig. S2.** Venn diagrams representing the degree of overlap among variants called using four variant callers and PacBio HiFi data compared to the GIAB truth sets. A, B, and C represent SNVs, while D, E, and F represent INDELs for HG003, HG006, and HG007, respectively. **Fig. S3.** Venn diagrams representing the degree of overlap among variants called using two variant callers and ONT data compared to the GIAB truth sets. A represents SNVs, while B represents INDELs for HG003.

Acknowledgements
The authors wish to thank Génome Québec, Genome Canada, the government of Canada, the Ministère de l'Économie et de l'Innovation du Québec, the Canadian Field Crop Research Alliance, Semences Prograin Inc., Sollio Agriculture, Grain Farmers of Ontario, Barley Council of Canada, and Université Laval. The authors wish also to thank GIAB for providing the reference variants of the samples. Moreover, we thank Mr. Don Freed at Sentieon for his continuous fast responses and guidance regarding the best practices of the usage of DNAscope.

Author contributions
Conceptualization, OA and DT; data curation and formal analysis, OA; resources, DT; writing, review and editing, OA, DT, and FB; supervision, DT; project administration, DT; funding acquisition, DT, and FB. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by by Genome Canada [#6548] under Genomic Applications Partnership Program (GAPP).

Availability of data and materials
The samples used for the analyses in the current study have previously been made available by the GIAB consortium at the NIST GIAB FTP site: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/. The source codes are publicly available in the GitHub repository: https://github.com/Omar-Abd-Elwahab/Variant_Callers.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 August 2023 Accepted: 4 December 2023
Published online: 14 December 2023

References

1. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10:4.
2. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12:91.
3. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018;16:15–24.
4. Stange M, Barrett RDH, Hendry AP. The importance of genomic variation for biodiversity, ecosystems and people. Nat Rev Genet. 2021;22:89–105.
5. Sawyer SD, Mitchell G, Mckinley J. A role for common genomic variants in the assessment of familial breast cancer 5-fluorouracil predictive test view project psychosocial and behavioural impact of genomic testing for polygenic breast cancer risk view project. J Clin Oncol. 2012. https://doi.org/10.1200/JCO.2012.41.7469.
6. Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, et al. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. 2012;8:e1002944.
7. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1092 human genomes. Nature. 2012;491:56–65.
8. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
9. Li W, Freudenberg J. Mappability and read length. Front Genet. 2014;5:1–1.
10. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.
11. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. Cell Genom. 2022;2:100129.
12. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 2021. 2021;39:11.
13. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19:1–11.
14. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18:1322–32.
15. Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun. 2018;9:1.
16. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:10.
17. Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nature Genet. 2021;53:942–8.
18. Miller NA, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med. 2015;7:1–16.
19. All of Us Research Program | National Institutes of Health (NIH). https://allofus.nih.gov/. Accessed 11 May 2023.
20. Supernat A, Vidarsson OV, Steen VM, Stokowy T. Comparison of three variant callers for human whole genome sequencing. Sci Rep. 2018;8:1–6.
21. Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One. 2016;11:e0151664.
22. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 2013;5:91.
23. Cai L, Yuan W, Zhang Z, He L, Chou KC. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep. 2016;6:1–9.
24. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep. 2017;7:1–12.
25. Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics. 2012;28:2097–105.
26. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:1–26.
27. Sandmann S, De Graaf AO, Karimi M, Van Der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep. 2017;2017 7:1.
28. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:1–8.
29. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, et al. An analytical framework for optimizing variant discovery from personal genomes. Nat Commun. 2015;6:1–6.
30. Callari M, Sammut SJ, De Mattos-Arruda L, Bruna A, Rueda OM, Chin SF, et al. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. Genome Med. 2017;9:1–11.
31. Barbitoff YA, Abasov R, Tvorogova VE, Glotov AS, Predeus AV. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genom. 2022;23:1–17.
32. Stegemiller MR, Redden RR, Notter DR, Taylor T, Taylor JB, Cockett NE, et al. Using whole genome sequence to compare variant callers and breed differences of US sheep. Front Genet. 2023;13:1060882.
33. Stegemiller MR, Redden RR, Notter DR, Taylor T, Taylor JB, Cockett NE, et al. Using whole genome sequence to compare variant callers and breed differences of US sheep. Front Genet. 2023;13:1060882.
34. Exposito-Alonso M, Drost HG, Burbano HA, Weigel D. The Earth BioGenome project: opportunities and challenges for plant genomics and conservation. Plant J. 2020;102:222–9.

35. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:7856.
36. Evans JD, Brown SJ, Hackett KJJ, Robinson G, Richards S, Lawson D, et al. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 2013;104:595–600.
37. Childers AK, Geib SM, Sim SB, Poelchau MF, Coates BS, Simmonds TJ, et al. The usda-ars ag100pest initiative: high-quality genome assemblies for agricultural pest arthropod research. Insects. 2021;12:626.
38. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585:7823.
39. A reference standard for genome biology. Nat Biotechnol. 2018;36:1121–1121.
40. Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genom. 2022;23:1–7.
41. Yang H, Gu F, Zhang L, Hua XS. Using generative adversarial networks for genome variant calling from low depth ONT sequencing data. Sci Rep. 2022;12:1–9.
42. Luo R, Wong CL, Wong YS, Tang CI, Liu CM, Leung CM, et al. Exploring the limit of using a deep neural network on pileup data for germline variant calling. Nat Mach Intell. 2020;2:220–7.
43. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
44. Zheng Z, Li S, Su J, Leung AWS, Lam TW, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci. 2022;2:797–803.
45. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: Is WGS the better WES? Hum Genet. 2016;135:359–62.
46. Huang PJ, Chang JH, Lin HH, Li YX, Lee CC, Su CT et al. DeepVariant-on-Spark: small-scale genome analysis using a cloud-based computing framework. Comput Math Methods Med. 2020;2020.
47. Freed D, Pan R, Chen H, Li Z, Hu J, Aldana R. DNAscope: high accuracy small variant calling using machine learning. bioRxiv. 2022;2022.05.20.492556.
48. 2. Typical usage for DNAseq® — Sentieon 202112.06 documentation. https://support.sentieon.com/manual/DNAseq_usage/dnaseq/. Accessed 26 Feb 2023.
49. Freed D, Aldana R, Weber JA, Edwards JS. The sentieon genomics tools—a fast and accurate solution to variant calling from next-generation sequence data. bioRxiv. 2017;115717.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
51. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:1–4.
52. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.
53. Variant Calling Pipeline using GATK4 – Genomics Core at NYU CGSB. https://gencore.bio.nyu.edu/variant-calling-pipeline-gatk4/. Accessed 30 Nov 2022.
54. Wagner J, Olson ND, Harris L, McDaniel J, Khan Z, Farek J et al. Benchmarking challenging small variants with linked and long reads. bioRxiv. 2021;2020.07.24.212712.
55. Krusche P, Trigg L, Boutros PC, Mason CE, de La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37:555–60.
56. Plateforme de gestion de données de recherche | VALERIA. https://valeria.science/accueil. Accessed 30 Nov 2022.

Publisher's Note