frontiers | Frontiers in Bioinformatics

TYPE Review
PUBLISHED 23 April 2025
DOI 10.3389/fbinf.2025.1574359

Check for updates

OPEN ACCESS

EDITED BY
David W. Ussery,
University of Arkansas for Medical Sciences,
United States

REVIEWED BY
Youtao Lu,
University of Pennsylvania, United States
Samir Lababidi,
United States Food and Drug Administration,
United States

*CORRESPONDENCE
Davoud Torkamaneh,
✉ davoud.torkamaneh.1@ulaval.ca

RECEIVED 10 February 2025
ACCEPTED 08 April 2025
PUBLISHED 23 April 2025

CITATION
Abdelwahab O and Torkamaneh D (2025) Artificial intelligence in variant calling: a review.
Front. Bioinform. 5:1574359.
doi: 10.3389/fbinf.2025.1574359

Artificial intelligence in variant calling: a review

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

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

Artificial intelligence (AI) has revolutionized numerous fields, including genomics, where it has significantly impacted variant calling, a crucial process in genomic analysis. Variant calling involves the detection of genetic variants such as single nucleotide polymorphisms (SNPs), insertions/deletions (InDels), and structural variants from high-throughput sequencing data. Traditionally, statistical approaches have dominated this task, but the advent of AI led to the development of sophisticated tools that promise higher accuracy, efficiency, and scalability. This review explores the state-of-the-art AI-based variant calling tools, including DeepVariant, DNAscope, DeepTrio, Clair, Clairvoyante, Medaka, and HELLO. We discuss their underlying methodologies, strengths, limitations, and performance metrics across different sequencing technologies, alongside their computational requirements, focusing primarily on SNP and InDel detection. By comparing these AI-driven techniques with conventional methods, we highlight the transformative advancements AI has introduced and its potential to further enhance genomic research.

KEYWORDS

variant calling, artificial intelligence, deep learning, genomics, machine learning (ML)

Introduction

Over the past few decades, the rapid advancement of next-generation sequencing (NGS) technologies has revolutionized the field of genomics. This technological evolution has facilitated the generation of vast amounts of data, which have greatly expanded our understanding of various biological processes, diseases, and genetic disorders (Bamshad et al., 2012; Koboldt et al., 2012; Timothy et al., 2013; Walter et al., 2015). Among the most common and impactful applications of NGS are whole genome sequencing (WGS) and whole exome sequencing (WES), which have become essential tools for dissecting genetic basis of different genetic conditions and development of diagnostic tools and guiding therapeutic decisions (Bodian et al., 2015; Willig et al., 2015; Smith et al., 2016; Schon et al., 2021). However, the sheer volume of data produced by these technologies presents significant challenges in terms of data storage, management and analysis.

To address these challenges, increasingly sophisticated computational pipelines and algorithms have been developed (McKenna et al., 2010; Garrison and Marth, 2012; Rimmer et al., 2014; Danecek et al., 2021; Helal et al., 2024), with a key focus on variant calling—the identification and characterization of genetic variants, such as single nucleotide polymorphisms (SNPs), insertions/deletions (InDels), and structural variants. Variant calling is a multi-step process in genomic analysis that involves sequencing, mapping, calling, and refinement. First, DNA or RNA is sequenced to generate raw nucleotide reads, which are then quality-checked. Next, these reads are mapped to a reference genome/transcriptome using alignment tools, producing a sequence alignment map that indicates where each read matches the reference. Variant calling tools then analyze the aligned reads to detect genetic variations which are recorded in variant call format (VCF) files. Finally, refinement steps, such as filtering to remove false positives, ensure high-confidence results for downstream analysis (Koboldt, 2020). Variant calling is a fundamental step in genomic studies, with applications ranging from population genetics to disease research and personalized medicine. It provides valuable insights into genetic diversity, disease-associated mutations, and individual genetic profiles (Maher et al., 2013; Lu et al., 2014; Witte et al., 2014; Bean and Hegde, 2016).

Historically, variant calling has relied on statistical methods to detect and interpret genomic variations that were extensively evaluated and reviewed in previous studies (Liu et al., 2013; Xu, 2018; Cameron et al., 2019; Pei et al., 2021; Helal et al., 2024; Joe et al., 2024). However, the advent of artificial intelligence (AI) has introduced a new generation of tools and pipelines that offer improved accuracy, efficiency, and scalability. In this review, we delve into the current state-of-the-art AI-based variant calling tools, examining their methodologies, strengths and limitations. We also provide a comparative analysis of their performance against conventional pipelines and discuss the broader implications of these advancements for the future of genomics.

AI-based variant callers

AI-based variant calling tools represent a cutting-edge approach in genomic research, leveraging machine learning (ML) and deep learning (DL) algorithms to improve the accuracy and efficiency of detecting genetic variants (Hall et al., 2024). Unlike traditional rule-based tools, these AI models are trained on large-scale genomic datasets to identify subtle patterns and reduce false-positive and false-negative rates. They can handle complex genomic regions, including repetitive or highly variable areas, where conventional methods often struggle (Junjun et al., 2024). Below is an overview of the most widely used AI-based variant calling tools, focusing on their unique methodologies and features.

DeepVariant

Developed by Google Health (The Value of Genomic Analysis - Google Health, 2022), DeepVariant is an open-source DL-based variant caller that uses deep convolutional neural networks (CNNs) to analyze pileup image tensors of aligned reads, outputting detected variants with high accuracy (Poplin et al., 2018). Initially designed for short-read data, it now supports long-read technologies such as PacBio High-Fidelity (HiFi) and Oxford nanopore technology (ONT) (Wenger et al., 2019; Shafin et al., 2021). One of the major strengths of DeepVariant is its ability to automatically produce filtered variants directly, eliminating the need for post-variant calling refinement. This feature, combined with its high accuracy compared to other tools (including SAMTools, Strelka, GATK, FreeBayes, and 16 GT) (Poplin et al., 2018), made it a preferred choice for large-scale genomic studies such as the UK Biobank WES consortium (500k individuals) (Szustakowski et al., 2021). However, the high computational cost associated with DeepVariant, despite its compatibility with both GPU and CPU, is a significant drawback (Abdelwahab et al., 2023).

DeepTrio

DeepTrio (Kolesnikov et al., 2021) is an advanced DL-based variant caller developed by Google Health as an extension of DeepVariant, designed specifically for analyzing genomic data from family trios, typically consisting of a child and their two parents. Building upon the foundation of DeepVariant, DeepTrio leverages deep CNNs to jointly analyze sequencing data from all three family members, enhancing the accuracy of variant detection by incorporating familial context. This approach allows DeepTrio to effectively weigh sequencing errors, mapping inaccuracies, and *de novo* mutation directly from the sequence data (Kolesnikov et al., 2021), leading to improved variant calling performance, especially in challenging genomic regions and lower sequencing coverages. DeepTrio's performance surpasses that of many traditional and non-trio variant calling methods, such as GATK (McKenna et al., 2010), Strelka (Kim et al., 2018), and FreeBayes (Garrison and Marth, 2012), by maintaining high accuracy across different sequencing technologies and coverage levels.

DNAscope

Developed by Sentieon, DNAscope is optimized for efficiency and computational speed (Freed et al., 2022a). It was originally developed for short-read sequencing and was later adapted for PacBio HiFi data for the PrecisionFDA challenge (Freed et al., 2022b; Olson et al., 2022). Recently, it has been extended to handle ONT data (DNAscope LongRead Nanopore Pipeline - Sentieon, 2025). DNAscope, which was introduced after DNAseq (Sentieon's GATK-matching germline variant calling pipeline) (Glusman et al., 2019), demonstrated very high SNP and InDel accuracy by combining GATK's HaplotypeCaller with an AI-based genotyping trained model (Freed et al., 2022a). These algorithms allow DNAscope to accurately identify and classify variations with high sensitivity and specificity. One key advantage of DNAscope is its ability to process large amounts of data quickly and accurately, without the need to set filtering thresholds manually. DNAscope achieved a significant reduction in computational cost compared to other variant callers, such as DeepVariant and GATK, by reducing the memory overhead and leveraging multi-threaded processing, which leads to faster runtimes without compromising accuracy (Freed et al., 2022a). DNAscope has demonstrated strong performance in various benchmarking studies, particularly in detecting SNPs and small InDels across diverse datasets (Pei et al., 2021; Li C. et al., 2022).

It is important to note that the methodology of DNAscope relies on ML enhancements rather than DL architectures. Even though ML is widely regarded as a subset of AI (Das et al., 2015; Chhaya et al., 2020; Helm et al., 2020), DNAscope's approach does not leverage DL techniques that are characteristic of modern AI-driven tools. Furthermore, as noted in the 'Resource Requirements' section, DNAscope does not require GPU acceleration, a common feature of DL-based AI models. These distinctions clarify DNAscope's classification as an ML-assisted AI tool rather than a DL-based AI model.

Clair, Clair3, and Clairvoyante

Clair (Luo et al., 2020) and Clair3 (Zheng et al., 2022) are DL-based variant callers that specialize in genomic variant detection from both short-read and long-read sequencing data. Developed as a successor to Clairvoyante (Luo et al., 2019), Clair builds on its predecessor's foundation, incorporating advanced neural networks to achieve improved accuracy in calling SNPs and InDels. Both Clair and Clairvoyante utilize CNNs to process genomic data, leveraging DL techniques to capture complex patterns in sequencing reads and produce accurate variant calls. Clairvoyante had some drawbacks; for instance, one major drawback was the inaccuracy in the calling of multi-allelic variants (Luo et al., 2019; Luo et al., 2020; Zheng et al., 2022). Clair extends Clairvoyante's variant calling capabilities by further optimizing the model's architecture and training it on more diverse data, improving its performance, especially on long-read sequencing data. For example, Clair3 runs faster than any of the other state-of-the-art variant callers and achieves better performance, especially at lower coverage, which are traditionally more prone to errors (Zheng et al., 2022).

Medaka

Medaka (Nanoporetech/medaka, 2018) is specifically designed for analyzing long-read sequencing data generated by ONT. It employs neural networks to perform haploid-aware variant calling, which allows it to account for the inherent error rates associated with ONT sequencing, while still producing accurate variant calls for SNPs and InDels. Medaka's variant calling pipeline typically follows after base-calling and alignment. The tool refines the mapped reads and produces high-quality consensus sequences using DL models specifically optimized for ONT data (Nanoporetech/medaka, 2018). One of Medaka's core strengths is its ability to correct sequencing errors common in long-read technologies by training on these datasets. This results in more reliable variant calls, particularly in complex genomic regions that are often challenging for short-read-based methods. While Medaka is particularly effective at detecting nucleotide variants, its accuracy for larger structural variants may not be as high as that of tools specifically designed for structural variant calling (Nanoporetech/medaka, 2018). The computational cost of running Medaka is moderate and extremely faster than other early tools tailored for ONT data (e.g., Nanopolish (Simpson et al., 2017; Jts/nanopolish, 2017)).

HELLO

Hybrid and stand-alone Estimation of smaLL genOmic variants (HELLO) (Ramachandran et al., 2021) is an open-source variant caller designed to address the challenges of detecting genomic variants of both SNPs and InDels from Illumina, PacBio, and hybrid data of both Illumina and PacBio. HELLO employs a deep neural network (DNN) architecture that explicitly models the fundamental units of sequencing data, such as reads and alleles. The DNNs used in HELLO are smaller and more efficient compared to those used in DeepVariant, allowing for faster execution and reduced computational resource requirements. It introduces operations to compare allelic evidence and uses probabilistic reasoning to produce variant calls. The pipeline involves sorting and aligning reads, with specific processes for short reads and PacBio reads. For short reads, InDel realignment is performed to ensure consistent representation of InDels, while for PacBio reads, haplotag sorting and alignment to reference and alternative alleles are conducted.

Performance comparison

AI-based variant calling tools have set a new benchmark for accuracy in detecting genetic variants across various sequencing platforms, leveraging DL architectures like CNNs and DNNs to interpret sequencing data with more nuance than traditional statistical approaches (Poplin et al., 2018; Glusman et al., 2019; Ramachandran et al., 2021; Barbitoff et al., 2022; Freed et al., 2022a; Olson et al., 2022; Wagner et al., 2022). Despite their shared goal of precision, these tools perform differently across sequencing technologies and variant types (Table 1). DeepVariant consistently delivers high SNP and InDel accuracy across Illumina and PacBio HiFi data, reaching SNP F1 scores of 99.9%, which makes it highly versatile (Poplin et al., 2018; Shafin et al., 2021; Olson et al., 2022). However, its accuracy decreases when processing ONT data, especially for InDels, due to ONT's higher base-calling error rates, with accuracy dropping to 76.8% for these variants (Shafin et al., 2021). Nonetheless, DeepVariant remains a preferred choice for large-scale projects, like the UK Biobank WES consortium, due to its capacity for accurate, filtered variant calls without post-processing (Szustakowski et al., 2021). DeepTrio extends DeepVariant's utility by integrating familial data from trios, allowing it to better detect variants, particularly *de novo* mutations and those in regions of low coverage. This family-based approach enables DeepTrio to achieve slightly better accuracy than DeepVariant, with SNP detection accuracy maintaining high levels at 99.8% for Illumina and 99.9% for PacBio HiFi, making it ideal for studies of rare genetic diseases and complex inheritance (Kolesnikov et al., 2021; Brand et al., 2024). While DNAscope, a commercial tool, also offers competitive accuracy, particularly for SNPs, it does so with significantly reduced computational overhead compared to DeepVariant and DeepTrio, making it suitable for high-throughput projects requiring rapid turnaround times (Freed et al., 2017; Freed et al., 2022a; Glusman et al., 2019). DNAscope's performance on PacBio HiFi matches other leading tools, with 99.9% SNP and 99.5% InDel accuracy, further supporting its use in time-sensitive clinical settings (Freed et al., 2022b). Clair3, the latest in the Clair series, provides robust support across all sequencing platforms, particularly excelling in long-read data with PacBio HiFi and ONT, reaching 99.9% SNP accuracy and notable InDel accuracy even at lower coverage. This, combined with moderate computational needs, makes Clair3 versatile for labs with diverse sequencing platforms (Altshuler et al., 2012; Walter et al., 2015; Luo et al., 2020; Zheng et al., 2022). Medaka, meanwhile, focuses on ONT long-read sequencing and is highly reliable for SNP detection at 99.0%, though it may face challenges with larger structural variants. Its strength lies in refining ONT-specific errors, which makes it valuable for ONT-specific projects, especially with high-coverage data (Nanoporetech/medaka, 2018). Nonetheless, Medaka recommends using Clair3 for diploid variant calling as it achieves higher accuracy and has better computational performance (Nanoporetech/medaka, 2018). However, Medaka still achieves comparable accuracies across various studies (Kuno et al., 2022; Zheng et al., 2022). Lastly, HELLO stands out for its support of both Illumina and PacBio data, excelling in hybrid settings where it reaches near-perfect SNP accuracy of 99.9% and significantly reduces InDel errors, particularly in combined Illumina-PacBio data (Ramachandran et al., 2021). This makes HELLO a strong option for studies requiring the integration of multiple sequencing technologies (Hassan et al., 2023; Zeibich et al., 2023). Moreover, it is notable to mention that HELLO outperforms DeepVariant using PacBio data using any given coverage (Ramachandran et al., 2021). In summary, while all these tools achieve high SNP detection accuracy, their InDel calling performance varies significantly for long-read technologies like ONT, with DeepVariant and Clair3 leading in multi-platform accuracy, and DNAscope and HELLO noted for efficiency and hybrid capabilities, respectively. To provide a comprehensive perspective, Table 1 also includes conventional variant calling methods, GATK, BCFTools, and FreeBayes, alongside their performance metrics, which serves as a reference to better contextualize the advancements and performance improvements introduced by the AI-based variant calling tools.

It is important to note that certain tools were not assessed for specific sequencing technologies due to their design focus and the availability of performance evaluations. DeepTrio, for example, is designed specifically for trio-based analysis and has been optimized for Illumina and PacBio HiFi data, with no published evaluations on ONT data. DNAscope, while recently adopted for ONT data, has not yet been included in comparative studies or benchmarks for this platform. HELLO is optimized for hybrid sequencing approaches (Illumina + PacBio) and currently does not support ONT-based variant calling. Conversely, Medaka is specifically designed for ONT data and is not optimized for Illumina or PacBio HiFi sequencing.

Computational resource requirements

The computational resource demands of AI-based variant callers vary considerably, influenced by the sequencing platform, dataset size, and tool architecture, with some tools necessitating substantial resources, particularly for large-scale or long-read sequencing projects (Table 2). DeepVariant is among the more resource-intensive options; for a typical 30x coverage of human whole genome sequence (WGS) using Illumina data, it requires around 5 h on a multi-core CPU system (e.g., m5.24xlarge, 96 vCPUs), but this time can be reduced to approximately 8 min when utilizing a high-performance GPU system like an NVIDIA DGX station (Deepvariant, 2023). The resource requirements escalate further when processing large population or long-read data, as DeepVariant demands substantial RAM and CPU resources, making it better suited for cloud-based or high-performance computing (HPC) environments, which may be inaccessible to smaller labs with limited infrastructure. DeepTrio, which builds on DeepVariant by integrating trio data, requires even greater computational power due to the processing of data from three individuals. When analyzing WGS data at 35x coverage of single human genome on Google Cloud, it typically necessitates a 16-thread CPU instance (Kolesnikov et al., 2021). This tool is optimal for familial studies but can be prohibitive for users without advanced computing setups.

In contrast, DNAscope offers high performance with a lower computational footprint, being designed to operate without GPU acceleration. It can efficiently process Illumina WGS data at 30x coverage in roughly 30 min on a 32-thread CPU system (Freed et al., 2022a), and PacBio HiFi data in about 3.67 h on a 32-core Intel Xeon server (Freed et al., 2022b). DNAscope's reduced resource requirements, without compromising on accuracy, make it a practical choice for labs balancing performance with computational availability. Clair3 also provides resource efficiency, typically consuming around 7 GB of RAM for Illumina and PacBio data, and requiring about 1 h for 30x coverage WGS processing on two 12-core Intel Xeon Silver 4116 processors (Luo et al., 2020). For ONT data, runtimes extend to around 5 h. Although Clair3 lacks official GPU support, its moderate RAM and CPU needs enable quick processing times, positioning it as an accessible option for labs across various sequencing platforms, especially those engaged with long-read data.

Medaka is specifically tailored for ONT long-read sequencing, and while not as fast as Clair3, it is significantly more efficient than earlier ONT tools like Nanopolish (Jts/nanopolish, 2017)—approximately 50 times faster (Nanoporetech/medaka, 2018). It supports both CPU and GPU, though its computational demands are relatively moderate, favoring ONT projects in labs that lack advanced GPU capabilities. HELLO strikes a balance between computational efficiency and robust performance, requiring around 24 CPU threads and 13 GB of RAM for standard execution, peaking at 29 threads and 18 GB of RAM in optimized flows. Although trained on GPUs, HELLO is optimized for multi-threaded CPU environments, making it accessible without GPU acceleration (Ramachandran et al., 2021). Its efficiency in handling both Illumina and PacBio data makes it suitable for hybrid sequencing projects, where it offers reliable performance without overwhelming computational demands.

In summary, while DeepVariant and DeepTrio offer the highest accuracy, they come with substantial resource requirements, particularly for long-read sequencing data. DNAscope and Clair3 present more resource-efficient alternatives, suitable for labs with limited access to HPC or cloud resources. Medaka and HELLO offer targeted advantages for ONT and hybrid datasets, respectively, while maintaining moderate computational needs, making them accessible for a wider range of research environments.

Notably, the computational resource requirements presented in Table 2 correspond to the analysis of a single human genome at 30x coverage according to literature reports. No normalization was performed to standardize CPU/GPU configurations across the original studies.

Table 3 provides a comparative overview of the AI-based variant callers highlighted in this study. It outlines their support for short-read and long-read sequencing technologies, licensing models (open-source vs. commercial), key advantages and disadvantages,

TABLE 1 F1 scores for SNPs and InDels, called from different sequencing technologies, and corresponding coverages for each variant caller.

Tool	Sequencing technology	SNP F1 score (%)	InDel F1 score (%)	Reported coverage (X)
DeepVariant	Illumina	99.9	99.6	30
	PacBio HiFi	99.9	99.2	30
	ONT	99.8	76.8	90
DeepTrio	Illumina	99.8	99.7	35
	PacBio HiFi	99.9	99.5	35
DNAscope	Illumina	99.5	>99	30
	PacBio HiFi	99.9	99.5	30
Clair	Illumina	99.9	99.6	52
	PacBio HiFi	99.9	99.3	33
Clair3	ONT	99.3	73.17	20
Medaka	ONT	99	73.2	30 (SNPs), 50 (InDel)
HELLO	Illumina	99.6	99.5	50
	PacBio	99.9	99.7	60
HELLO (Hybrid)	Illumina + PacBio	99.9	99.8	50 (Illumina), 60 (PacBio)
GATK	Illumina	99.2	97.3	26
	PacBio	99.5	77.7	40
FreeBayes	Illumina	98	92.7	26
BCFTools	Illumina	95.7	81.2	12
	PacBio HiFi	99.3	84.9	40
	ONT	90.9	0	80

Note: The InDel F1 score of 0.0 for BCFTools on ONT data reflects experimental results reported by Abdelwahab et al., where BCFTools failed to detect any InDels, resulting in both precision and recall values of zero.

and their ability to handle complex genomic regions such as repetitive or GC-rich areas. Each tool is uniquely suited to specific applications, depending on sequencing platform, computational resources, and project requirements.

Discussion

The landscape of genomic variant calling is rapidly evolving, driven by the growing adoption of AI-based tools. These tools have demonstrated substantial improvements in accuracy, scalability, and efficiency, addressing many of the challenges posed by complex sequencing technologies and diverse data types (Hall et al., 2024; Junjun et al., 2024). This section examines perspectives, challenges, and future opportunities in this field, while comparing the performance of AI-based tools with conventional methods.

Perspectives on AI in variant calling

DL models typically require large datasets for training. In the context of AI-based variant callers, all models reviewed in this study were trained using samples from the GIAB dataset. GIAB samples are widely utilized in genomics research for benchmarking and training due to their well-characterized and high-confidence variant annotations. However, the number of training samples and specific methodologies differ across pipelines.

Transfer learning capabilities

Species transferability

Among the AI-based variant callers, DeepVariant exhibits the highest potential for transfer learning. Studies have demonstrated that a DeepVariant model trained on human genomic data performs better on mouse genomic data than a model trained exclusively on mouse data, indicating effective cross-species

TABLE 2 Summary of variant caller computational requirements for a single human genome. ND: not determined.

Tool	Sequencing technology	Hardware support	System Configuration	Runtime (CPU)	Runtime (GPU)	Reported coverage (X)	References
DeepVariant	Illumina	CPU and GPU	CPU: m5.24xlarge 96 threads, GPU: NVIDIA DGX (8 x A100)	5 h	4 min	30	(Accelerate Genomic Analysis for Any Sequencer with NVIDIA Parabricks v4.2 | NVIDIA Technical Blog, 2023)
	PacBio HiFi			5 h	4 min	30	
	ONT			8 h	21 min	55	
DeepTrio	Illumina	CPU and GPU	CPU: n1-standard 16 threads	100 h	ND	35	Kolesnikov et al. (2021)
	PacBio HiFi		ND	ND	ND	ND	
DNAscope	Illumina	CPU only	CPU: 8xlarge 32 threads	30 min	ND	30	Freed et al. (2022a)
	PacBio HiFi		CPU: Intel® Xeon® server 32 threads	3.67 h	ND	30	Freed et al. (2022b)
Clair	Illumina	CPU and GPU	CPU: Intel Xeon Silver 4116 24 threads	1 h	ND	30	Luo et al. (2020)
	PacBio HiFi			1 h	ND	30	
	ONT			5 h	ND	30	
Medaka	ONT	CPU and GPU	CPU: n1-series 16 threads, GPU: 1x NVIDIA Tesla P100	95 h	40 h	50	Shafin et al. (2021)
HELLO	Illumina	CPU and GPU	ND	ND	ND	ND	Ramachandran et al. (2021)
	PacBio		CPU: Intel E5-2683 28 threads	27 min	ND	30 (chr 21)	
HELLO (Hybrid)	Illumina + PacBio	CPU and GPU	ND	ND	ND	ND	
GATK	Illumina	CPU only	CPU: Intel Xeon-P8260 16 threads	44 h	ND	12	Abdelwahab et al. (2023)
	PacBio			102 h		40	
BCFTools	Illumina	CPU only	CPU: Intel Xeon-P8260 16 threads	3 h	ND	12	Abdelwahab et al. (2023)
	PacBio HiFi			39 h		40	
	ONT			8 h		78	

TABLE 3 Comparative summary of six AI-based variant callers evaluated in this study.

Variant caller	Short-read (Illumina) support	Long-read (PacBio/ONT) support	Open-source vs. Commercial	Advantages	Disadvantages	Handling of complex genomic regions
DeepVariant	Yes	Yes (PacBio HiFi, ONT)	Open-source	High accuracy across platforms, CNN-based architecture, automated filtering	High computational cost, requires GPU for optimal performance	Performs well in repetitive and GC-rich regions
DeepTrio	Yes	Yes (PacBio HiFi)	Open-source	Optimized for trio-based analysis, improved *de novo* mutation detection	High computational cost, requires trio sequencing	Effective in low-coverage and complex regions due to familial context
DNAscope	Yes	Yes (PacBio HiFi, ONT)	Commercial (Sentieon)	High efficiency, optimized for speed and low compute cost	Not fully DL-based, less adaptable to non-human genomes	Performs well in low-depth regions but lacks full CNN adaptability
Clair/Clair3	Yes	Yes (PacBio HiFi, ONT)	Open-source	High performance for long-read sequencing, optimized for low-coverage sequencing	Lacks official GPU acceleration	Strong performance in complex regions and low-quality reads
Medaka	No	Yes (ONT)	Open-source	Optimized for ONT, accurate SNP calling	Lower InDel accuracy, requires high-coverage ONT data	Handles ONT-specific errors effectively
HELLO	Yes	Yes (PacBio)	Open-source	Hybrid caller supporting both Illumina and PacBio, efficient execution	Limited benchmarking on ONT	Performs well with hybrid sequencing data, reduces sequencing errors

generalization (Poplin et al., 2018). Additionally, DeepVariant enables users to train custom models without a gold standard set, using sites consistent with Mendelian inheritance to improve calling accuracy (DeepVariant Blog, 2018). Moreover, DeepVariant allows users to train custom models tailored to specific data types, such as BGISEQ-500 whole genome data (Deepvariant, 2024).

Locus transferability

DeepTrio enhances variant calling by leveraging familial relationships, making it more accurate in low-coverage or ambiguous regions. Clair3 and Medaka, trained on long-read data, excel in identifying variants in highly repetitive regions that typically challenge short-read-based tools. HELLO's hybrid training approach makes it particularly robust when integrating data from different sequencing technologies.

The integration of AI in variant calling brought a major shift in genome analyses. Tools such as DeepVariant, DeepTrio, and Clair3 are examples of how DL architectures can optimize the detection of SNPs and small InDels (Poplin et al., 2018; Huang et al., 2020; Yun et al., 2020; Kolesnikov et al., 2021; Zheng et al., 2022; Brand et al., 2024). These tools go beyond traditional statistical approaches by capturing patterns within sequencing data, offering an enhanced methodology for variant detection (Junjun et al., 2024). For instance, DeepVariant and HELLO consistently achieve SNP F1 scores of 99.9% on Illumina and PacBio HiFi platforms, highlighting their superior performance compared to conventional tools such as GATK (McKenna et al., 2010) and FreeBayes (Garrison and Marth, 2012).

AI-based tools have also facilitated large-scale population studies by enabling efficient data processing and analysis. Projects like the UK Biobank (Szustakowski et al., 2021) and All of Us (Mahmoud et al., 2024) research program have successfully utilized these tools to process large datasets, setting new benchmarks in genomics. Moreover, tools such as HELLO, which integrate hybrid datasets from Illumina and PacBio platforms, demonstrate how AI can address challenges in complex genomic regions by reducing sequencing errors and improving resolution (Ramachandran et al., 2021).

Challenges in implementing AI variant calling

Despite these advancements, several challenges hinder the broader adoption of AI-based tools. High computational demands are a primary obstacle, particularly for smaller research groups with limited resources. Tools like DeepVariant and DeepTrio rely heavily on GPU resources or HPC environments, and their requirements grow significantly with long-read sequencing data (Kolesnikov et al., 2021; Shafin et al., 2021; Deepvariant, 2023; Brand et al., 2024). While cloud computing offers a viable solution, it raises concerns related to data security, cost scalability, and regulatory compliance, particularly in clinical settings (Minh Dang et al., 2019; Vistro et al., 2020).

While accuracy is a concern for some long-read technologies, several tools, such as Clair3, achieve high InDel detection accuracy on ONT data when optimal sequencing coverage is maintained (Zheng et al., 2022). Coverage plays a critical role in accuracy, as lower coverage can exacerbate sequencing errors and hinder variant calling, especially for challenging genomic regions (Sims et al., 2014; Parks and Lambert, 2015). Addressing these challenges requires refining AI models to adapt to varying sequencing depths and data quality. Additionally, the interpretability of many AI models remains limited (Linardatos et al., 2020; Li X. et al., 2022), raising questions about reproducibility and user confidence, especially in regulatory and diagnostic workflows.

Superior performance of AI-based tools

Comparative analyses reveal that AI-based variant callers consistently outperform conventional tools in accuracy, sensitivity, and computational efficiency (Abdelwahab et al., 2023). For example, DeepVariant and Clair3 achieve higher detection rates for SNPs and InDels compared to widely used methods like GATK and SAMtools (Li et al., 2009; Danecek et al., 2021), particularly in challenging genomic regions. DeepTrio, with its ability to incorporate familial context, enhances the detection of *de novo* mutations (Kolesnikov et al., 2021; Brand et al., 2024), outperforming non-trio approaches such as Strelka (Saunders et al., 2012; Kim et al., 2018) and FreeBayes (Garrison and Marth, 2012). Similarly, DNAscope's optimized architecture balances speed and accuracy, making it an ideal choice for clinical applications requiring rapid and precise variant calling (Freed et al., 2022a; Freed et al., 2022b). These advancements underline the transformative potential of AI-based tools to complement or even surpass traditional methods, particularly in high-throughput and precision-driven genomic studies.

Future opportunities for AI in variant calling

The continued development of AI in variant calling offers numerous opportunities for methodological refinement and broader applications. A key area for improvement lies in the integration of interpretable AI methodologies (Watson, 2022), which would enhance model interpretability and foster greater trust among clinicians and researchers. Additionally, optimizing tools for resource-constrained environments could democratize access to advanced genomic technologies, addressing disparities in global research capabilities.

Further innovation is needed to enhance the robustness and generalizability of AI models through training on diverse datasets (Chen et al., 2023) and sequencing platforms (Yu et al., 2023). Currently, most of these tools were trained with very small and limited dataset such as Genome In A Bottle (GIAB) samples (Genome in a bottle, 2015; Zook et al., 2016). Additionally, hybrid approaches, such as those demonstrated by HELLO, combining data from short-read, long-read, and mixed sequencing platforms, hold great promise for improving variant detection accuracy while mitigating platform-specific limitations. Emerging computational frameworks, including transformer-based architectures and attention mechanisms (Choi and Lee, 2023), could further accelerate processing speeds and improve precision, driving breakthroughs in both research and clinical genomics.

Conclusion

AI-based tools have significantly advanced the field of variant calling by enhancing accuracy, scalability, and computational efficiency across diverse sequencing platforms. These tools have outperformed conventional methods in key metrics, enabling breakthroughs in both large-scale genomic research and clinical diagnostics. However, challenges such as high computational demands, limited model interpretability, and platform-specific performance constraints remain critical barriers. Addressing these issues will require further methodological innovations, including hybrid data integration and interpretable AI frameworks. Moreover, developing tools optimized for diverse sequencing technologies and resource-constrained environments will be essential for broader adoption. With continued advancements, AI-driven variant callers have the potential to redefine genomic research and enable transformative applications in personalized medicine, population studies, and beyond.

Author contributions

OA: Conceptualization, Data curation, Formal Analysis, Investigation, Validation, Writing – original draft, Writing – review and editing. DT: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was funded by Genome Canada [#6548] under Genomic Applications Partnership Program (GAPP).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. Generative AI was used to assist in drafting and refining the text, including structuring sentences, improving clarity, and ensuring coherence. However, all scientific content, analysis, and interpretations were developed and verified by the authors to ensure accuracy and originality.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdelwahab, O., Belzile, F., and Torkamaneh, D. (2023). Performance analysis of conventional and AI-based variant callers using short and long reads. *BMC Bioinforma.* 24, 472–513. doi:10.1186/s12859-023-05596-3

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi:10.1038/nature11632

Bamshad, M. J., Shendure, J. A., Valle, D., Hamosh, A., Lupski, J. R., Gibbs, R. A., et al. (2012). The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. *Am. J. Med. Genet. A* 158A, 1523–1525. doi:10.1002/AJMG.A.35470

Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S., and Predeus, A. V. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics* 23, 155–217. doi:10.1186/s12864-022-08365-3

Bean, L. J. H., and Hegde, M. R. (2016). Gene variant databases and sharing: creating a global genomic variant database for personalized medicine. *Hum. Mutat.* 37, 559–563. doi:10.1002/HUMU.22982

Bodian, D. L., Klein, E., Iyer, R. K., Wong, W. S. W., Kothiyal, P., Stauffer, D., et al. (2015). Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. *Genet. Med. 2016* 18 (3), 221–230. doi:10.1038/gim.2015.111

Brand, F., Guski, J., and Krawitz, P. (2024). Extending DeepTrio for sensitive detection of complex *de novo* mutation patterns. *Nar. Genom Bioinform* 6, lqae013. doi:10.1093/NARGAB/LQAE013

Cameron, D. L., Di Stefano, L., and Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.* 10 (1), 3240–3311. doi:10.1038/s41467-019-11146-4

Chen, N. C., Kolesnikov, A., Goel, S., Yun, T., Chang, P. C., and Carroll, A. (2023). Improving variant calling using population data and deep learning. *BMC Bioinforma.* 24, 197–215. doi:10.1186/s12859-023-05294-0

Chhaya, K., Khanzode, A., and Sarode, R. D. (2020). Advantages and disadvantages of artificial intelligence and machine learning: a literature review. *Int. J. Libr. and Inf. Sci. (IJLIS)* 9, 3. doi:10.17605/OSF.IO/GV5T4

Choi, S. R., and Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* 12, 1033. doi:10.3390/BIOLOGY12071033

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008–4. doi:10.1093/GIGASCIENCE/GIAB008

Das, S., Dey, A., Pal, A., and Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *Int. J. Comput. Appl.* 115, 31–41. doi:10.5120/20182-2402

Deepvariant (2023). NVIDIA docs. Available online at: https://docs.nvidia.com/clara/parabricks/4.1.0/documentation/tooldocs/man_deepvariant.html (Accessed January 22, 2025).

Deepvariant (2024). GitHub. Available online at: https://github.com/google/deepvariant/blob/r1.8/docs/deepvariant-training-case-study.md (Accessed March 25, 2025).

DeepVariant Blog (2018). Improved non-human variant calling using species-specific DeepVariant models. Available online at: https://google.github.io/deepvariant/posts/2018-12-05-improved-non-human-variant-calling-using-species-specific-deepvariant-models/ (Accessed March 25, 2025).

Freed, D., Aldana, R., Weber, J. A., and Edwards, J. S. (2017). The Sentieon Genomics Tools - a fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*, 115717. doi:10.1101/115717

Freed, D., Pan, R., Chen, H., Li, Z., Hu, J., and Aldana, R. (2022a). DNAscope: high accuracy small variant calling using machine learning. *bioRxiv*, 492556. doi:10.1101/2022.05.20.492556

Freed, D., Rowell, W. J., Wenger, A. M., and Li, Z. (2022b). Sentieon DNAscope LongRead – a highly accurate, fast, and efficient pipeline for germline variant calling from PacBio HiFi reads. *bioRxiv*. doi:10.1101/2022.06.01.494452

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Available online at: https://arxiv.org/abs/1207.3907v2 (Accessed September 8, 2024).

Genome in a bottle (2015). Genome in a bottle—a human DNA standard. *Nat. Biotechnol.* 33, 675. doi:10.1038/NBT0715-675A

Glusman, G., Rodrigues Alves Margarido, G., Aganezov, S., Mainzer, L. S., Kendig, K. I., Baheti, S., et al. (2019). Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.* 10, 736. doi:10.3389/fgene.2019.00736

Google Health (2022). The value of genomic analysis - Google health. Available online at: https://health.google/health-research/genomics/ (Accessed November 28, 2022).

Hall, M. B., Wick, R. R., Judd, L. M., Nguyen, A. N., Steinig, E. J., Xie, O., et al. (2024). Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. *Elife* 13. doi:10.7554/ELIFE.98300

Hassan, S., Bahar, R., Johan, M. F., Mohamed Hashim, E. K., Abdullah, W. Z., Esa, E., et al. (2023). Next-generation sequencing (NGS) and third-generation sequencing (TGS) for the diagnosis of thalassemia. *Diagnostics* 13, 373. doi:10.3390/DIAGNOSTICS13030373

Helal, A. A., Saad, B. T., Saad, M. T., Mosaad, G. S., and Aboshanab, K. M. (2024). Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Sci. Rep.* 14, 6160–6222. doi:10.1038/s41598-024-56604-2

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., et al. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr. Rev. Musculoskelet. Med.* 13, 69–76. doi:10.1007/s12178-020-09600-8

Huang, P. J., Chang, J. H., Lin, H. H., Li, Y. X., Lee, C. C., Su, C. T., et al. (2020). DeepVariant-on-Spark: small-scale genome analysis using a cloud-based computing framework. *Comput. Math. Methods Med.* 2020, 1–7. doi:10.1155/2020/7231205

Joe, S., Park, J. L., Kim, J., Kim, S., Park, J. H., Yeo, M. K., et al. (2024). Comparison of structural variant callers for massive whole-genome sequence data. *BMC Genomics* 25, 318–414. doi:10.1186/s12864-024-10239-9

Jts/nanopolish (2017). Signal-level algorithms for MinION data. Available online at: https://github.com/jts/nanopolish (Accessed September 2, 2024).

Junjun, R., Zhengqian, Z., Ying, W., Jialiang, W., and Yongzhuang, L. (2024). A comprehensive review of deep learning-based variant calling methods. *Brief. Funct. Genomics* 23, 303–313. doi:10.1093/BFGP/ELAE003

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* 15 (8), 591–594. doi:10.1038/s41592-018-0051-x

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Med.* 12, 91. doi:10.1186/s13073-020-00791-w

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490 (7418), 61–70. doi:10.1038/nature11412

Kolesnikov, A., Goel, S., Nattestad, M., Yun, T., Baid, G., Yang, H., et al. (2021). DeepTrio: variant calling in families using deep learning. *bioRxiv*. doi:10.1101/2021.04.05.438434

Kuno, A., Ikeda, Y., Ayabe, S., Kato, K., Sakamoto, K., Suzuki, S. R., et al. (2022). DAJIN enables multiplex genotyping to simultaneously validate intended and unintended target genome editing outcomes. *PLoS Biol.* 20, e3001507. doi:10.1371/JOURNAL.PBIO.3001507

Li, C., Fan, X., Guo, X., Liu, Y., Wang, M., Zhao, X. C., et al. (2022a). Accuracy benchmark of the GeneMind GenoLab M sequencing platform for WGS and WES analysis. *BMC Genomics* 23, 533–611. doi:10.1186/s12864-022-08775-3

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., et al. (2022b). Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* 64, 3197–3234. doi:10.1007/s10115-022-01756-8

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23, 18. doi:10.3390/E23010018

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B. Z. (2013). Variant callers for next-generation sequencing data: a comparison study. *PLoS One* 8, e75619. doi:10.1371/JOURNAL.PONE.0075619

Lu, Y. F., Goldstein, D. B., Angrist, M., and Cavalleri, G. (2014). Personalized medicine and human genetic diversity. *Cold Spring Harb. Perspect. Med.* 4, a008581. doi:10.1101/CSHPERSPECT.A008581

Luo, R., Sedlazeck, F. J., Lam, T. W., and Schatz, M. C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* 10 (1), 998–1011. doi:10.1038/s41467-019-09025-z

Luo, R., Wong, C. L., Wong, Y. S., Tang, C. I., Liu, C. M., Leung, C. M., et al. (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat. Mach. Intell.* 2 (4), 220–227. doi:10.1038/s42256-020-0167-4

Maher, M. C., Uricchio, L. H., Torgerson, D. G., and Hernandez, R. D. (2013). Population genetics of rare variants and complex diseases. *Hum. Hered.* 74, 118–128. doi:10.1159/000346826

Mahmoud, M., Huang, Y., Garimella, K., Audano, P. A., Wan, W., Prasad, N., et al. (2024). Utility of long-read sequencing for all of us. *Nat. Commun.* 15 (1), 837–913. doi:10.1038/s41467-024-44804-3

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/GR.107524.110

Minh Dang, L., Piran, M. J., Han, D., Min, K., and Moon, H. (2019). A survey on internet of things and cloud computing for healthcare. *Electronics* 8, 768–8. doi:10.3390/ELECTRONICS8070768

Nanoporetech/medaka (2018). Sequence correction provided by ONT research. Available online at: https://github.com/nanoporetech/medaka (Accessed September 2, 2024).

Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., et al. (2022). PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genomics* 2, 100129. doi:10.1016/J.XGEN.2022.100129

Parks, M., and Lambert, D. (2015). Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics* 16, 19–12. doi:10.1186/s12864-015-1219-8

Pei, S., Liu, T., Ren, X., Li, W., Chen, C., and Xie, Z. (2021). Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief. Bioinform* 22, bbaa148–11. doi:10.1093/BIB/BBAA148

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36 (10), 983–987. doi:10.1038/nbt.4235

Ramachandran, A., Lumetta, S. S., Klee, E. W., and Chen, D. (2021). HELLO: improved neural network architectures and methodologies for small variant calling. *BMC Bioinforma.* 22, 404–431. doi:10.1186/s12859-021-04311-4

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., et al. (2014). Integrating mapping-assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi:10.1038/ng.3036

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* 28, 1811–1817. doi:10.1093/bioinformatics/bts271

Schon, K. R., Horvath, R., Wei, W., Calabrese, C., Tucci, A., Ibañez, K., et al. (2021). Use of whole genome sequencing to determine genetic basis of suspected mitochondrial disorders: cohort study. *BMJ* 375, e066288. doi:10.1136/BMJ-2021-066288

Sentieon (2025). DNAscope LongRead nanopore pipeline - Sentieon. Available online at: https://www.sentieon.com/dnascope-nanopore/ (Accessed March 27, 2025).

Shafin, K., Pesout, T., Chang, P. C., Nattestad, M., Kolesnikov, A., Goel, S., et al. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* 18, 1322–1332. doi:10.1038/s41592-021-01299-w

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14 (4), 407–410. doi:10.1038/nmeth.4184

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15 (2), 121–132. doi:10.1038/nrg3642

Smith, L. D., Willig, L. K., and Kingsmore, S. F. (2016). Whole-exome sequencing and whole-genome sequencing in critically ill neonates suspected to have single-gene disorders. *Cold Spring Harb. Perspect. Med.* 6, a023168. doi:10.1101/CSHPERSPECT.A023168

Szustakowski, J. D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P. G., Sasson, A., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics* 53 (7), 942–948. doi:10.1038/s41588-021-00885-0

Timothy, J. L., Christopher, M., Li, D., Benjamin, J. R., Andrew, J. M., Gorden, R., et al. (2013). Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* 368, 2059–2074. doi:10.1056/NEJMOA1301689

Vistro, D., Rehman, A. U., Mahmood, S., Munawar, A., Mago Vistro, D., Rehman, A. U., et al. (2020). A literature review on security issues in cloud computing: opportunities and challenges journal of critical reviews A literature review on security issues in cloud computing: opportunities and challenges. *Article J. Crit. Rev*. doi:10.31838/jcr.07.10.282

Wagner, J., Olson, N. D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., et al. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genomics* 2, 100128. doi:10.1016/J.XGEN.2022.100128

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526 (7571), 82–90. doi:10.1038/nature14962

Watson, D. S. (2022). Interpretable machine learning for genomics. *Hum. Genet.* 141, 1499–1513. doi:10.1007/s00439-021-02387-9

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37 (10), 1155–1162. doi:10.1038/s41587-019-0217-9

Willig, L. K., Petrikin, J. E., Smith, L. D., Saunders, C. J., Thiffault, I., Miller, N. A., et al. (2015). Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. *Lancet Respir. Med.* 3, 377–387. doi:10.1016/S2213-2600(15)00139-3

Witte, J. S., Visscher, P. M., and Wray, N. R. (2014). The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* 15 (11), 765–776. doi:10.1038/nrg3786

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* 16, 15–24. doi:10.1016/J.CSBJ.2018.01.003

Yu, H., Zheng, Z., Su, J., Lam, T. W., and Luo, R. (2023). Boosting variant-calling performance with multi-platform sequencing data using Clair3-MP. *BMC Bioinforma.* 24, 308–321. doi:10.1186/s12859-023-05434-6

Yun, T., Li, H., Chang, P. C., Lin, M. F., Carroll, A., and McLean, C. Y. (2020). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* 36, 5582–5589. doi:10.1093/bioinformatics/btaa1081

Zeibich, R., Kwan, P., J. O'Brien, T., Perucca, P., Ge, Z., and Anderson, A. (2023). Applications for deep learning in epilepsy genetic research. *Int. J. Mol. Sci.* 24, 14645. doi:10.3390/IJMS241914645

Zheng, Z., Li, S., Su, J., Leung, A. W. S., Lam, T. W., and Luo, R. (2022). Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* 2 (12), 797–803. doi:10.1038/s43588-022-00387-x

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3, 160025–160026. doi:10.1038/sdata.2016.25