Nepal J Biotechnol. 2026 Jul; 14 (1): 85-98 Review Article DOI: https://doi.org/10.54796/njb.v14i1.481

Nepal Journal of Biotechnology
Publisher: Biotechnology Society of Nepal
Journal Homepage: https://nepjb.com/index.php/NJB
ISSN (Online): 2467-9313
ISSN (Print): 2091-1130

SNP Detection Strategies in Genomic Research: A Comparative Review of Major Tools, Algorithms, Challenges and Applications.

Shikhi Baruri[1], Sunita Khanal[2]

[1]Center for predictive medicine, University of Louisville, Louisville, KY, USA

[2]Casey Eye Institute, Oregon Health & Science University, Portland, OR, USA

Received: 08 Feb 2026; Revised: 08 Jul 2026; Accepted: 12 Jul 2026; Published online: 31 Jul 2026

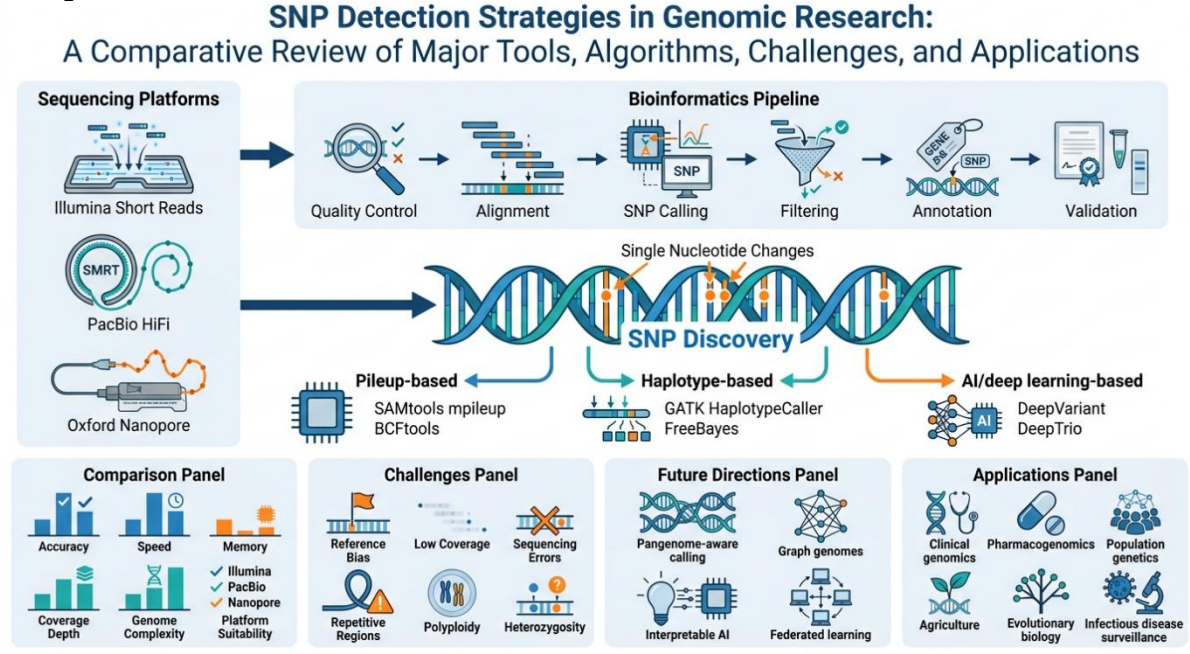
Abstract

Single nucleotide polymorphisms (SNPs), representing the most frequent form of genetic variation, serve as essential biomarkers for mapping complex traits, tracing evolutionary lineages, and identifying the genetic basis of disease susceptibility. Next-generation sequencing (NGS) enables large-scale SNP discovery across diverse organisms, yet accurate detection remains challenging due to sequencing errors, genome complexity, reference bias, and coverage depth. This narrative/comparative review synthesizes primary tool publications, benchmarking studies, and recent literature identified from PubMed, Google Scholar, Web of Science, and official software documentation. This review compares SNP detection programs such as GATK, BCFtools, FreeBayes, SAMtools, and DeepVariant and their algorithmic structures, namely pileup-based, haplotype-based, and machine-learning approaches. In general, the comparison suggests that no single tool is the best, as the performance of these tools largely depends on the organism type and genome complexity, sequencing platform, sequencing depth, type of variant, and computational resources. GATK is sensitive and precise among the detection tools reviewed, yet computationally intensive; BCFtools is fast and versatile for non-human datasets; FreeBayes is a high-precision tool for haplotype-based and multiallelic variants; SAMtools is a powerful and stable tool for low-coverage data; and DeepVariant achieves high precision through deep learning architectures but at a high computational cost. We address important hurdles in SNP detection, such as polyploidy, heterozygosity, and repetitive regions, alongside the emerging necessity of Pangenomic Equity, defined here as the use of diverse pangenome references to reduce ancestry-related reference bias in SNP discovery. Lastly, we discuss emerging developments, including artificial intelligence-based variant calling, transformer-based models, graph-based reference genomes, and pangenome-aware methods that help to overcome the limitations of linear genomic models. Therefore, this review summarizes the key considerations for tool selection and outlines future directions for robust and inclusive SNP discovery in genomic research.

Keywords: SNP Discovery, NGS, Reference bias, Pangenomics, Haplotype-based calling, Deep learning variant calling.

Corresponding author, email**:** s0baru04@louisville.edu, shikhibaruri18@gmail.com, sunitakhanal7@gmail.com

Graphical abstract



Introduction

Single Nucleotide Polymorphisms (SNPs) are common genetic variations found in any species, that take place when there is a variation in a single base substitution [1]. These variations can occur in non-coding regions or within coding sequences as either synonymous (no change in amino acid) or nonsynonymous (change in amino acid) [2]. Approximately one percent of the 3 billion base pairs of the human DNA sequence is variable between any two individuals, resulting in tens of millions of SNPs across the genome [3]. While earlier research relied heavily on the GRCh38 reference genome, the transition to Telomere-to-Telomere (T2T) gapless assemblies has fundamentally expanded the scope of SNP discovery to previously inaccessible regions, including centromeres and telomeres [4]. Due to abundance and uniqueness within individuals of the same species or related species, SNPs serve as essential genetic markers for studying complex genetic traits and genome evolution [2, 5, 6].

With recent advancements in next-generation sequencing (NGS) technologies, SNP discovery has become a rapid, high-throughput, and cost-effective approach for whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted sequencing. These techniques are standard in population genomics, cancer research, clinical diagnosis, microbial genomics, and evolutionary research. However, the accuracy of SNP detection depends on complex computational algorithms and strong statistical models required to translate raw sequencing reads into high-quality and accurate SNP calls. Furthermore, the choice of SNP detection tools is critical because variant calls vary with the underlying algorithm, sequencing depth, genome quality, filtering strategy, and complexity of the biological system [7, 8]. These differences may impact downstream interpretation, along with false-positive and false-negative SNP calls, reproducibility across studies, and the selection of candidate variants for biological or clinical validation [7, 8].

Beyond drug response, SNP tools enable the development of diagnostic and prognostic algorithms that integrate multi-omics and clinical information to identify biomarkers and guide precision therapies for complex diseases, including cancer and cardiovascular disorders [9-12]. Therefore, the development of these diverse SNP discovery tools has assisted researchers in transitioning from basic genetic methods to advanced personalized healthcare, enabling the prediction, prevention, and treatment of various diseases across multiple biological domains (**Figure 1**) [12-14].

To our knowledge, few recent comparative reviews have (i) compared Pileup-based, Haplotype-based, and Machine learning-based approaches across various organisms and sequencing platforms and translated these comparisons into practical tool-selection recommendations; (ii) presented the concept of Pangenomic Equity as a framework for dealing with the reference bias related to ancestry; and (iii) integrated the new AI-driven and graph-based methods in a coherent framework [15-17]. Despite these advancements, the detection process is frequently hindered by sequencing errors, misaligned sequences, repetitive genomic sequences, reference genome bias, and biological complexities such as polyploidy and heterogeneity. This review evaluates commonly used SNP detection tools and their underlying algorithms, specifically GATK, BCFtools, FreeBayes, SAMtools, and DeepVariant, assessing their computational challenges for performance trade-offs among biological complexities. By synthesizing the strengths and limitations of these tools, we provide a comparative framework to assist researchers in selecting and optimizing the most effective pipelines for diverse genomic applications. Furthermore, this review examines the potential of new bioinformatics techniques, such as pangenome-aware variant calling and the use of artificial intelligence, to address persistent challenges in genomic studies, including reference bias and low-coverage data, and offers a technical and ethical roadmap for the future of inclusive research.

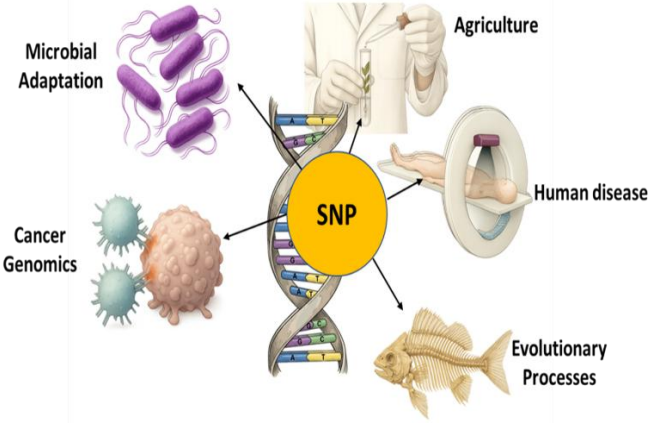


Figure 1: Application of single nucleotide polymorphism (SNPs) detection in biological and translational research. The diagram illustrates the role of SNP discovery across various fields, including microbial adaptation, agricultural improvement, human disease diagnostics, evolutionary biology, and cancer genomics.

Methodology

This review aims to provide a narrative and comparative overview of SNP detection strategies, tools, algorithms, challenges, and applications in genomic research. The literature was searched using PubMed, Google Scholar, Web of Science, and official software documentation. This review focuses on publications from the past decade (2015–2025) to reflect recent developments in sequencing technology, variant calling algorithms, machine learning, and pangenome-based methods. Keywords included "SNP detection," "variant calling", "GATK", "BCFtools", "FreeBayes", "SAMtools", "DeepVariant", "short-read sequencing", "long-read sequencing", "pangenome variant calling", "graph genome", "machine learning variant calling", and "deep learning SNP detection". Articles were prioritized if they focused on primary tools for SNP calling, benchmarking, comparisons of sequencing platforms, pangenome and/or graph-based variant detection, or major biological and computational challenges in SNP discovery. Foundational publications outside this period were also cited when necessary to describe original tools or core methodological concepts.

Pipeline for single nucleotide polymorphism (SNP) detection

A standard bioinformatics workflow for detecting SNPs from sequencing data is based on several fundamental steps designed to accurately determine and confirm the existence of genomic variants (Figure 2). The SNP detection process begins with raw sequencing data generated using NGS platforms, such as Illumina or Oxford Nanopore Technologies (ONT). Quality control tools, such as FastQC and Trimmomatic, screen reads for base quality scores, adapter contamination, and read length. Filtered reads are then aligned to a reference genome using high-performance aligners, such as BWA, Bowtie2, or STAR, producing an alignment file (e.g., BAM format) that indicates the correspondence of sequencing reads to the reference genome. During post-alignment processing, the alignments are further refined to ensure precise variant calling by identifying duplicates (reads created by PCR amplification) and recalibrating base quality scores to eliminate systematic errors. Notably, while earlier pipelines required a separate local realignment for insertions/deletions (indels), modern tools, such as GATK pipelines (*v3.6+* and *v4.0*), have integrated this IndelRealigner step into the variant calling or HaplotypeCaller algorithm [18], streamlining the workflow. Subsequently, variant callers, such as GATK HaplotypeCaller, FreeBayes, and SAMtools mpileup, identify SNPs by detecting mismatches against the reference genome. To minimize false-positives, raw SNP calls are filtered by depth coverage, allele frequency and strand bias, followed by functional annotation using SnpEff or ANNOVAR. Final SNP validation can be displayed in a genome browser (IGV or UCSC) or cross-confirmed through complementary laboratory procedures such as DNA melting analysis or allele-specific PCR [19]. In a specific study using SGSautoSNP for complex genomes, an SNP detection accuracy greater than 93% was reported; therefore, this value should be interpreted as study- and tool-specific rather than as a general accuracy estimate for all SNP detection pipelines. This automated, less expensive pipeline has been successfully applied to large and complex genomes, including polyploid organisms, and supports a variety of output formats, such as VCF and BCF formats [19, 20]. Wet-lab approaches, such as DNA ligation and rolling circle amplification, may serve as orthogonal validation or genotyping methods, but they are not part of the standard bioinformatics SNP-calling pipeline [21]. Furthermore, the use of containerization and standardization of tools increases reproducibility and pipeline robustness [15, 22].

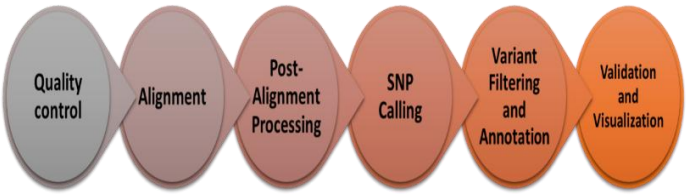


Figure 2: Standard bioinformatics pipeline for SNP detection from NGS data. The workflow illustrates the sequential stages from raw data quality control, sequence alignment, and post-alignment refinement to variant calling, functional annotation, and final validation.

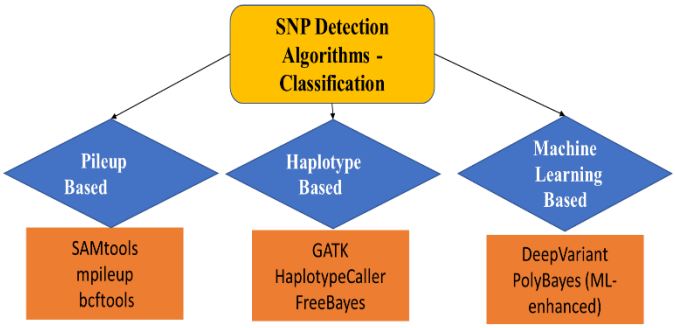


Figure 3: Classification of SNP detection algorithms based on methodological approaches. The diagram outlines the three primary categories of variant calling strategies: pileup-based, haplotype-based, and Machine Learning-based methods.

SNP detection algorithm classification

SNP detection algorithms are categorized into three methodological frameworks: pileup-based, haplotype-based, and machine learning-based methods. Figure 3

Table 1 Classification of SNP Detection Algorithms.

Algorithm	Core Principle	Representative Tools	Strengths	Limitations	Applications	Sources
Haplotype-based	Performs local assembly to infer haplotypes explaining read data	GATK Haplotype Caller, FreeBayes	High precision and sensitivity; improve indel and complex variant detection	Computationally intensive; sensitive to preprocessing quality	Clinical genomics, WGS/WES, cohort studies	[26, 27, 37]
Machine learning–based	Learn patterns of true variants from training data using statistical or deep learning models	DeepVariant, PolyBayes (ML-enhanced)	High precision and sensitivity; reduce false- positives; adaptable across platforms	High computational requirements; dependent on training data quality	Clinical-grade variant calling, large-scale genomics	[36, 38, 39]
Pileup-based	Evaluate base frequencies and qualities at each genomic position relative to a reference	SAMtools mpileup, BCFtools	Fast, simple, low computational cost; effective at moderate to high coverage	Reduced precision and sensitivity in repetitive regions and low- coverage; limited haplotype resolution	Population genomics, non-model organisms, low-cost analyses	[23-25]

provides a conceptual overview of these three major classes, and **Table 1** summarizes their defining features, representative tools, strengths, limitations, and applications.

Pileup-Based Methods

Pileup-based methods compare aligned sequence reads (pileups) at specific genomic positions against a reference to identify SNPs by evaluating the frequency and quality of nucleotide changes at detection sites. This method is often applied in next-generation sequencing data processing when base calls from multiple reads are accumulated to distinguish true variants from sequencing noise. Pileup-based approaches are simple and commonly used, and are integrated into variant callers, such as SAMtools mpileup and BCFtools. In contrast, the GATK HaplotypeCaller is better classified as a haplotype-based method because it uses local haplotype assembly. Although pileup-based methods are computationally efficient and highly sensitive at high coverage, they may struggle in complex genomic regions or with low-coverage data. Conversely, multi-sample calling strategies using pileup data can improve variant calling accuracy [23-25]. The core concept of the pileup-based approach is to decide: "Is the observed base different from the reference often enough to be real?"

Haplotype-Based Methods

These techniques examine clusters of co-segregating SNPs as haplotype blocks, rather than isolated sites. In variant calling, these methods perform local assembly or haplotype reconstruction to determine which candidate haplotypes best explain the observed sequencing reads. Representative tools include the GATK HaplotypeCaller and FreeBayes [26-28]. This approach informs the imputation of missing genotypes and determines historical genetic connections within a population. Haplotype block structures and tagging strategies are utilized in software tools to select representative SNPs for association studies [29, 30]. The primary goal of the haplotype-based approach is to understand which haplotype best explains the sequence reads.

Haplotype analyses have elucidated ethnic variations in genes, such as *GRK4*, related to hypertension, revealing population-specific haplotype structures and frequencies relevant for association interpretations [31]. They also enhance the fine mapping of disease loci, complementing single SNP analyses with a richer genetic context. Emerging advances in high-throughput genotyping, statistical modeling, and population-specific haplotype maps are progressively improving haplotype inference and interpretation, enabling integration with functional genomics for a deeper understanding of complex disease genetics [32, 33]. Thus, haplotype-based methods remain pivotal for dissecting genetic contributions to phenotypes beyond single SNP analysis.

Machine Learning-Based Methods

Machine learning (ML) techniques, including the random forest (RF) algorithm and support vector machines (SVM), contribute to SNP detection by identifying complex patterns in high-dimensional data that traditional statistics may miss. These methods improve classification and phenotype-prediction performance in genome-wide association studies (GWAS) and population genetics by effectively handling such datasets and capturing nonlinear relationships [34, 35]. Recent research indicates that using machine learning models,

Table 2: Technical comparison of SNP Detection Tools.

Tool	Best Application	Algorithm / main feature	Key Strength	Limitations	Practical recommendation
GATK	Human WGS/WES, clinical research, and cohort-level studies.	Haplotype-based local assembly and probabilistic variant calling.	Widely used standard for human WGS/WES; supports joint genotyping; and has strong documentation and community support.	Computationally intensive and requires preprocessing and filtering less convenient for small or resource-limited projects.	It is recommended for human WGS/WES and cohort studies when high accuracy, reproducibility, and standardized workflows are priorities. Use variant quality score recalibration (VQSR) when suitable training resources are available; otherwise, apply carefully optimized hard filters.
BCFtools	Non-human species, large cohorts, and population genomics.	Pileup- and genotype-likelihood-based variant calling and filtering.	Fast, flexible, open-source, and computationally efficient; useful for large cohorts and non-model organisms.	Reduced sensitivity to complex indels and structural variants.	It is recommended when speed, flexibility, and low computational cost are important, especially for non-model organisms and large population datasets. Strict filtering is required for low-depth or repetitive regions.
FreeBayes	Highly polymorphic regions in diverse organisms.	Bayesian haplotype-based variant detection.	Detects multi-allelic variants and can model complex allele structures; does not require known variant databases.	Slower than pileup-based callers; requires careful parameter optimization; reduced sensitivity at low coverage.	It is recommended for haplotype-aware SNP discovery, multi-allelic sites, highly polymorphic regions, and polyploid or mixed samples, especially when sufficient sequencing depth is available.
SAMtools mpileup - BCFtools call	Rapid exploratory SNP discovery, low-cost workflows, and low-coverage screening.	Pileup-based evidence from aligned reads combined with downstream variant calling in BCFtools	It is stable, widely used, lightweight, and useful for quick preliminary variant discovery.	Requires substantial downstream filtering; less powerful for complex variants; RNA-seq variant calling can be affected by splicing, allele-specific expression, and RNA editing.	It is recommended for exploratory SNP discovery and preliminary screening. The final interpretation should use strict filtering, validation, or comparison with another caller.
DeepVariant	High-accuracy SNP and small indel calling; high-quality WGS; PacBio HiFi or ONT data when platform-matched models are available.	Deep-learning-based variant calling from read pile-up images.	Low error rates in benchmarked datasets, strong performance in difficult genomic regions, and reduced manual filtering.	High computational cost (often requiring GPU acceleration), limited structural variant detection and dependency on platform-specific training data, and limited interpretability.	This is recommended when maximum SNP/small indel accuracy is required and sufficient computational resources are available. Training data bias and platform compatibility should be considered, especially for underrepresented populations and non-model organisms.

such as RF for feature selection and SVM for classification, can help predict an individual's likelihood of developing asthma based on SNP data [34]. Machine learning methods effectively automate the filtering process and minimize manual validation. For example, an earlier study combining PolyBayes with an ML classifier (C4.5) reported an increase in precision, also known as positive predictive value (PPV), from 7.8% to

84.8%; however, this benchmark should be interpreted as an older example of ML-assisted SNP discovery rather than as representative of current deep learning-based variant callers. While both PolyBayes and the ML classifier identified a similar number of true positives, the ML classifier greatly reduced false-positive SNP calls, generating only 249 false-positives compared to 16,955 for PolyBayes. This justifies the inclusion of ML as an auxiliary component of SNP discovery to improve filtering efficiency processes [36]. The fundamental philosophy of a machine learning-based approach is "Let the model learn how real SNPs look."

Collectively, these three categories of SNP detection methods–pileup-based algorithms for raw variant calling, haplotype-based approaches leveraging population-level genomic structure, and machine learning techniques for complex pattern recognition–offer complementary strengths that researchers utilize depending on the data type, study aims, and available computational resources.

Comparative evaluation of SNP detection tools

This section provides a detailed assessment of the most widely utilized SNP discovery tools, highlighting their algorithmic foundations and performance trade-offs (Table 2). To provide a more structured comparison, the tools were evaluated according to sequencing platform, biological application, genome complexity, coverage depth and computational requirements. GATK and DeepVariant are typically more appropriate for short-read Illumina WGS/WES in human or clinical studies, as they are more accurate, have a more standardized workflow, and are more effective at detecting SNPs and small indels than other tools. In case of non-human species, microbial genomics, and large population datasets, BCFtools and SAMtools are sometimes more convenient due to their speed, flexibility, and low memory usage. FreeBayes is also useful when the region is highly polymorphic, the site is multiallelic, or the sample is polyploid or mixed; however, it requires adequate sequencing depth and careful parameter optimization. DeepVariant and other long-read-aware callers can be used for long-read Pacific Biosciences (PacBio) or ONT datasets to yield better results in challenging regions of the genome but should be used with platform-specific error profiles and computational costs in mind[24, 25, 38, 39].

The ranked performance of GATK and DeepVariant differs across studies, as both depend on coverage depth, evaluation metrics, read mappers, and whether they are measuring single-nucleotide variants (SNVs) or indels. At 30×, GATK and DeepVariant achieved almost equal SNV F1-scores in one NA12878 WGS study, with DeepVariant outperforming GATK in the case of indels [28]. A trio exome study demonstrated that DeepVariant achieved a lower Mendelian error rate than GATK (3.09% vs. 5.25%) [40]. A systematic benchmark also revealed that the ranking of GATK was sensitive to its filtering strategy (with hard filtering, it was a close runner-up to DeepVariant, while with 2D-CNN filtering, it ranked worst on SNPs), and that the choice of aligner had a significant impact on the other callers as well [41].

GATK

The Genome Analysis Toolkit (GATK) is highly accurate and sensitive in SNP detection, utilizing advanced algorithms such as HaplotypeCaller, which performs local *de novo* assembly and probabilistic modelling to minimize false- positives. GATK aligns well with joint genotyping workflows, allowing cohort-level variant calling that improves the accuracy of homozygous reference genotypes while remaining versatile across multiple data types, such as whole-genome, exome, and RNA-seq data. A significant advantage of GATK is its large community adoption, frequent updates, and extensive documentation, which support its reproducibility and extensive use [27]. However, GATK requires substantial computational resources and user expertise to optimize complex parameters and filtering processes. The precision of variant calls depends on preprocessing, including indel realignment and base quality score recalibration, which means that poorly prepared data can negatively affect results and may produce false- positives in complex genomic regions such as repetitive or structured areas. Additionally, although GATK HaplotypeCaller can detect SNPs and small indels, larger indels, copy-number variants, and structural variants require specialized workflows or complementary tools, such as GATK-SV, GATK germline copy-number variant tools, Manta, and DELLY [42-44].

BCFtools

BCFtools is recognized for speed, efficiency, and flexibility in processing and analyzing high-throughput sequencing data in population genomics and non-human datasets. The moderate-coverage performance claim is supported by empirical benchmarking in non-human systems, including chicken NGS datasets evaluated across approximately 5X, 10X, 20X, and 30X coverage levels, where BCFtools-multiple showed high sensitivity across this range, with performance generally improving as sequencing depth increased. It provides a

comprehensive suite of tools for variant calling, filtering, and annotation and is freely available under an open-source license [23, 24, 45]. Additionally, it supports efficient genotype likelihood modeling and includes extensions such as BCFtools/RoH for detecting runs of homozygosity [46]. Although computationally efficient, BCFtools requires careful parameter tuning for complex variant types, including indels and structural variants, compared with larger toolkits, such as GATK. Because BCFtools relies on pileup-based evidence and genotype likelihoods, its variant calls can be affected by low read depth, poor mapping quality, multi-mapping reads, indel-associated artifacts and repetitive genomic regions. These conditions may contribute to missed variants or false-positive calls if filtering parameters are not optimized [45, 47].

FreeBayes

FreeBayes is a Bayesian genetic variant detector that performs haplotype-based variant calling and is optimized for highly polymorphic regions, multi-allelic sites, and small indels, where haplotype-aware calling can provide advantages over SNP-only approaches. When used in combination with suitable data processing strategies, its probabilistic model has been adapted for specialized data types, such as bisulfite sequencing; however, this requires additional modification, such as base-quality score manipulation, and should not be interpreted as having broad default applicability to all difficult data types. FreeBayes has been benchmarked as being competitive in sensitivity and precision with other popular callers, such as GATK and SAMtools; however, these performance comparisons are context-dependent and have been reported in specific systems, including plant and chicken genome datasets, rather than representing universal performance across all organisms or sequencing designs [25, 45, 48]. However, FreeBayes sensitivity can drop at low allele frequencies or low-coverage, and its processing speed is generally slower than some specialized callers for particular data types or complex genomic datasets [23, 45].

SAMtools mpileup

SAMtools mpileup offers a robust and practical technique for SNP finding and variant calling in large-scale sequencing data, particularly when paired with widely used mapping software such as BWA-MEM [49]. In current workflows, SAMtools mpileup is commonly used together with BCFtools, as variant-calling functionality was separated from SAMtools into BCFtools after the SAMtools v1.0 workflow transition. Benchmarking in large plant genome resequencing showed that SAMtools mpileup achieved performance comparable to GATK under specific mapping and filtering conditions, including a strong overall AUC performance when raw reads were mapped with BWA-MEM [7, 8]. Therefore, SAMtools mpileup may be useful for exploratory SNP discovery in low-coverage or complex non-model datasets, but low-coverage remains a general limitation for all SNP callers and requires careful filtering, validation, and interpretation. However, SAMtools mpileup may not be as powerful in identifying complex variants or low-frequency mutations without additional filtering steps. Similarly, the alignment quality and chosen mapper can affect its performance [8, 25, 50].

DeepVariant

DeepVariant utilizes deep learning neural network models to provide an automated pipeline that minimizes human error and scales efficiently to large datasets, delivering high accuracy and often surpassing traditional SNP callers across various sequencing platforms and genomic datasets [39]. For example, Hall et al. benchmarked deep learning-based variant callers on bacterial nanopore sequencing data and reported a strong performance for DeepVariant in that context [38]. Additionally, it is an open-source tool, which allows it to be transparent, contribute to the community, and continue to improve. Nonetheless, the substantial high-cost computational power, potential training data/modeling biases, and high-skilled operation requirements remain the primary limitations of DeepVariant for broad adoption in resource-constrained environments. The recent development of deep learning-based variant calling has also expanded beyond single-sample analysis; for example, DeepTrio extends the DeepVariant framework to child-parent trio or duo sequencing data and improves variant calling by jointly using familial sequence information [51]. Additionally, despite its high accuracy, false-positive or false-negative results may still occur, particularly in complicated or substandard genomic areas in complex, low-quality, or underrepresented genomic regions [39].

Advancements in bioinformatics enhancing SNP discovery from next-generation sequencing data

Computational Acceleration and Workflow Optimization

Recent developments in bioinformatics have vastly improved SNP discovery using next-generation sequencing (NGS) data, leading to increased processing efficiency, accuracy, error reduction, and analytical capacity. A key development involves optimized

algorithms and computing frameworks, such as enhanced versions of the Genome Analysis Toolkit (GATK), including Parabricks and Sentieon, which dramatically reduce the processing time for large human genomic datasets while maintaining high concordance with standard GATK outputs rather than representing absolute variant-calling accuracy against ground-truth datasets [52-54]. These improvements allow for the rapid, high-throughput analysis required to handle the massive data volumes generated by NGS. Additionally, refined error-correction algorithms across biological domains, including immunogenomics and virology, help minimize sequencing errors, although researchers must still carefully balance sensitivity and precision based on the data type [55].

Error Correction, Alignment, and Variant-Calling Improvements

Modern tools now integrate sophisticated sequence alignment and SNP-calling algorithms, employing suffix array-based aligners and highly sensitive hash-based aligners. Additionally, SNP and genotype calling algorithms utilize heuristic and probabilistic methods to effectively discern true polymorphisms from sequencing noise and errors [16].

Feature Selection and Dimensionality Reduction

To manage the computational burden and overfitting with high-dimensional NGS data, feature selection and extraction approaches, such as statistical, machine, and deep learning techniques, have improved dimensionality reduction, model performance, interpretability, and computational efficiency. Such innovations enhance SNP identification and derive biological information from complex datasets [35, 36].

Whole-Genome Resequencing and Plant Genomics

The implementation of improved bioinformatics pipelines and whole-genome resequencing (WGR) systems allows SNP detection, validation, and genotyping on a genome-wide scale in plants. WGR coupled with bulked segregant analysis can be used to increase the speed of causal mutation and quantitative trait loci (QTL) mapping. Also, long-read sequencing platforms, accompanied by more advanced bioinformatics, are likely to bring major advantages of accuracy and utility due to the fact that NGS is inherently limited [56].

SNP detection across sequencing platforms

SNP detection is currently dominated by short-read sequencing platforms, such as Illumina, which are characterized by low error rates and high throughput. Typically providing reads of 50-300 base pairs, these platforms provide comprehensive coverage, acceptable base quality, low duplication rates, and high accuracy for determining SNPs and minimal insertions or deletions (indels), with wide applications in population genetics and disease association studies [57, 58]. However, the limitations of short-read platforms are most evident in genomic regions where short sequences cannot be uniquely mapped. This mapping ambiguity, mainly in regions such as centromeres, telomeres, and segmentally duplicated areas, results in an incomplete representation of the SNP landscape associated with complex neurodevelopmental, diabetes, and cardiovascular traits [59, 60].

In contrast, long-read sequencing platforms, including ONT and PacBio, produce reads spanning thousands of base pairs, enabling superior resolution of complex genomic genotypes, repetitive sequences, structural variants, and haplotype phasing, addressing several drawbacks inherent to short-read technology. Although long-read technologies have historically suffered from higher per-base error rates, recent improvements, such as ONT R10.4.1/Q20+ chemistry, have been reported to achieve greater than 99% raw-read accuracy, improving their usefulness for SNP detection at sufficient coverage depths [61]. PacBio HiFi reads, generated through circular consensus sequencing (CCS), have also been reported with approximately 99.8% average read accuracy and an average read length of 13.5 kb, while broader HiFi datasets commonly report read accuracies greater than 99.5% with read lengths of approximately 10–25 kb [62, 63]. Long-read sequencing excels at detecting variations in repetitive or challenging regions that cannot be effectively resolved by short reads. This capability is particularly valuable in applications such as antimicrobial resistance gene identification and comprehensive genome assembly [57–58, 64, 65]. This advancement was instrumental in the first complete Telomere-to-Telomere (T2T-CHM13) human genome assembly, which uncovered over 2 million previously inaccessible variants in highly repetitive heterochromatic regions [4]. Consequently, the integration of multi-platform pipelines—combining the raw depth of short reads with the structural clarity of HiFi and ONT—is becoming the gold standard for exhaustive SNP discovery across all genomic domains [4, 61-63, 66].

SNP detection applications

SNP technology is extensively used in biomedical research, clinical diagnostics and personalized medicine. It is vital for identifying genetic variations linked to diseases and allows accurate molecular characterization

across diverse fields. SNP microarrays are widely applied to tumor genomes to identify heterozygotic loss, copy number alterations, and cancer predisposition markers, such as DNA methylation, genetic mutations in tumor suppressor genes, or inherited gene changes, informing diagnosis and prognosis [67]. Another critical field where SNP detection plays an important role is pharmacogenomics. Mapping SNPs related to drug response, toxicity, or metabolism allows clinicians to customize drug treatments based on personal genetic data, enhancing effectiveness and minimizing adverse effects. For example, SNPs in pharmacogenomic genes such as COMT, HTR2A, CYP2D6, CYP2C19, and ABCB1 have been associated with variations in drug response, analgesic effect, and antidepressant dosing, supporting their relevance to personalized treatment strategies [68-70]. The development of large SNP databases in combination with clinical data is ongoing for personalized medicine strategies [71]. In evolutionary and population genomic analyses, tools such as SNPGenie can use pooled NGS-derived SNP data to estimate nucleotide diversity, codon-level variation, and gene- or genome-wide signals of natural selection, supporting the functional interpretation of SNP datasets [72]. Infectious disease surveillance and variant detection technologies also rely on SNP detection methods. For example, nucleic acid probe-based multiplexed electrical detection assays facilitate the use of single-nucleotide specificity and speed in the rapid discrimination of viral subtypes, such as SARS-CoV-2 variants, and facilitate large-scale pandemic screening [73]. Other applications include forensic analysis, gene expression profiling, complicated disease research, and the development of portable real-time biosensors for genetic diagnostics. Complementary wet-lab and hardware-based platforms, including microarray-based, biosensor-based, and electrochemical SNP detection methods, can expand targeted SNP genotyping options; however, these approaches should be interpreted as supporting technologies rather than replacements for computational SNP-calling workflows [74].

Challenges in SNP detection

SNP detection faces multiple technical and biological challenges despite significant advances in sequencing technologies and bioinformatics. Reference genome bias remains a primary concern. The reliance on a single reference genome is not enough to capture the global genetic diversity present in the populations [75, 76]. When the linear reference genome lacks the ancestral diversity of the sample being sequenced, reads containing non-reference alleles are often misaligned or discarded as sequencing noise. Additionally, using a distantly related species or an incomplete genome assembly can lead to incorrect mapping, biased estimates of genetic diversity, inaccurate detection of SNPs, and misinterpretation of structural genomic features, such as chromosomal rearrangements, heterozygosity, and nucleotide diversity [77]. Sequencing errors, including base-calling mistakes and amplification artifacts, can introduce false variants, a major issue for true variant calls in low-coverage datasets (e.g., ≤15×), where supporting reads are insufficient for confident heterozygous genotype calling [78, 79].

Genome complexity also poses significant mapping difficulties. Repetitive genomic regions lead to multi-mapping or misalignment, false-positive calls, or missed polymorphisms, especially in repeat-rich and large genomes, such as those of many plant species, complicating SNP discovery and reducing the reliability of variant calls [80, 81]. Furthermore, polyploidy and heterozygosity complicate variant calling by requiring specialized algorithms to distinguish between homeologous variants and true SNPs. Effectively managing these challenges requires greater coverage, rigorous preprocessing, and refined algorithms for reliable SNP discovery, particularly in non-model or recently hybridized species [82-84].

Some of the challenges are not merely data-collection issues but algorithmic ones. For example, pileup and haplotype-based callers show false-negative rates in populations with a high genetic-distance from the reference genome resulting in genomic exclusion which leads to the misclassification of pathogenic variants or failure to identify pharmacogenomic markers [85].

To minimize these biases, the field is transitioning toward Pangenomic Equity. The 2023 release of the first draft human pangenome reference by the Human Pangenome Reference Consortium (HPRC) provided 47 phased diploid assemblies from genetically diverse individuals and added approximately 119 million base pairs of euchromatic polymorphic sequences relative to GRCh38 [86]. However, the growth of pangenomics demands high computational resources, particularly extensive memory for processing complete graphs, specialized visualization tools and sophisticated indexing methods. These requirements ultimately create challenges in clinical interpretation, as well as a trade-off between comprehensiveness and usability [87].

While AI-driven variant calling improves efficiency and scalability, it also introduces significant risks of training

bias. If deep learning-based variant callers, such as DeepVariant for SNP/small-indel detection, DeepTrio for family based variant calling, Clair for long-read germline variant calling, or Medaka for Nanopore consensus and variant calling, are trained predominantly on Genome in a Bottle (GIAB) datasets, the model performance may be limited by the diversity of the training data and may overfit to the error profiles of specific genomes or sequencing platforms. Additionally, because the interpretation of AI models is limited by their trained datasets, their reproducibility and user confidence remain questionable, especially in diagnostic and clinical workflows [17].

In general, published benchmarking studies indicate that the performance of SNP calling is context-specific and not consistent across all tools. There are several reasons why different studies may have different tool rankings, including variations in the sequencing platform, coverage depth, reference genome quality, mapper used, filtering thresholds, organism type, and validation strategy. Thus, there is limited or inconsistent evidence for low-coverage datasets, repetitive genomic regions, polyploid or highly heterozygous genomes, non-model organisms, long-read platforms, AI-assisted callers, and pangenome-aware workflows [7, 8, 25, 38]. These conflicting rankings across benchmarking studies are largely driven by differences in datasets, sequencing platforms, filtering strategies, and specific evaluation metrics.

Emerging trends and future directions

Emerging trends in SNP detection are increasingly driven by a shift from linear, reference-based models towards multi-modal artificial intelligence (AI), deep learning, and novel genomic frameworks, such as graph-based reference genomes and pangenome-aware variant calling, along with improved tools tailored for non-model organisms.

AI and deep learning are increasingly improving SNP detection by enabling highly accurate and scalable variant calling from next-generation sequencing (NGS) data. Deep convolutional neural networks, exemplified by tools such as DeepVariant, outperform traditional statistical models by learning complex patterns in sequencing read pileups to identify genetic variants with high precision [39, 88]. Furthermore, emerging shifts toward explainable AI (XAI) allow researchers to use interpretability approaches to better understand model-based predictions [89]. Therefore, future directions should emphasize interpretable AI models and integrative frameworks that combine multi-omics data to enhance precision medicine applications and clinical trust [12, 13, 17]. However, AI-powered variant calling is still not perfect. It relies on the quality, diversity, and platform relevance of training datasets. AI callers may exhibit lower accuracy in underrepresented populations, non-model organisms, repetitive regions, low-quality samples, or sequencing platforms with varying error profiles [17, 38, 39]. Another limitation is the high computational cost, need for GPUs, lack of interpretability, and lack of benchmarking in a variety of organisms, which are important obstacles to the routine use of the approach [17, 39].

Graph-based reference genomes and pangenome models are emerging as transformative frameworks that overcome the limitations of linear reference genomes by simultaneously representing multiple haplotypes and diverse genomic variants within a graph structure. Pangenome graphs improve read mapping and variant detection accuracy and can represent complex genomic variations at multiple scales [90]. Tools such as the Minigraph-Cactus pipeline allow the construction of pangenomes directly from whole-genome alignments and support the analysis of large collections of related genomes, including non-human data [91]. Additionally, PanGenie uses k-mer-based genotyping with a haplotype-resolved pangenome reference to genotype SNPs, indels, and structural variants. The reported greater than fourfold speed improvement at 30-fold coverage should be attributed specifically to PanGenie benchmarking rather than Minigraph-Cactus pangenome construction [91, 92]. However, pangenome-based approaches remain constrained by high memory usage and computational costs, complex graph building, underdeveloped visualization tools, and the absence of standardized clinical interpretation tools. In contrast to traditional linear reference-based pipelines, variant annotation and reporting are more challenging in graph-based systems [87, 90, 93, 94].

Future SNP detection methods should also prioritize pangenome-aware variant calling, wherein variant discovery pipelines incorporate the complex variation encoded in pangenome graphs rather than relying on a single linear reference. This approach enhances sensitivity and accuracy, particularly in structurally diverse regions of the genome and in populations that are inadequately represented by current reference sequences [90]. Additionally, improved computational tools are being developed for SNP detection in non-model organisms, which have traditionally lacked rich genomic resources. Some AI-enabled variant calling models have

shown potential beyond human datasets, but their performance in non-human species still requires careful benchmarking and validation. Pangenome construction pipelines have also demonstrated applicability beyond humans, such as in Drosophila melanogaster and plant species, such as rice, broadening the scope of genetic variant discovery in ecological and agricultural research [39, 91, 95].

It is also crucial to differentiate between well-developed technologies and those still in development. Many pipelines and variant callers for short-read Illumina WGS/WES are already well established for rapid SNP discovery in human, microbial, plant, population-genomic, and clinical studies [23, 26, 27, 39, 58]. However, variant callers based on transformers, explainable AI frameworks, federated learning models, graph-based clinical pipelines, and routine pangenome-aware clinical interpretation are still under development [89-92]. These new methods offer the potential for more efficient SNP discovery in various populations and challenging genome regions, but need to be further benchmarked, standardized, computationally optimized, and better interpreted [87, 90, 93, 94]. Ultimately, the clinical implementation of these AI-based and pangenome-aware workflows still requires large-scale validation before widespread adoption.

Conclusion

SNP detection continues to be a cornerstone of genomic research, supporting an extensive range of scientific and clinical applications worldwide. This review demonstrates that although numerous detection tools are available, the choice depends on several factors, such as the sequencing platform, organism under study, coverage depth, and specific research objectives. Each tool, such as GATK, BCFtools, FreeBayes, SAMtools mpileup, and DeepVariant, offers specific benefits and drawbacks regarding computational requirements, ability to handle polymorphic site complexity, and sensitivity to complex genomic architectures. For low-coverage non-model organism datasets, BCFtools and SAMtools may be more practical than GATK due to their lower computational overhead and flexible parameter control; however, their performance depends on sequencing depth, mapping quality, reference genome quality, and filtering strategy. DeepVariant can provide high accuracy but requires substantial computational resources, preferably GPU acceleration and high-quality training data.

A limitation of this review is that it is a narrative/comparative review rather than a uniform benchmarking study; therefore, tool performance was interpreted from published studies and reported use cases rather than from the reanalysis of the same datasets. Despite technological advancements, the identification of single nucleotide polymorphisms remains challenging because of the influence of linear reference genome bias, sequencing errors, and complex genome structures. Furthermore, issues such as low or uneven data coverage and biological factors, such as polyploidy and genetic diversity (heterozygosity), play an important role. Overcoming these challenges to correctly identify SNP sites requires a combination of rigorous preprocessing, parameter optimization, and robust downstream analyses to generate biologically significant results. Future studies should focus on standardized benchmarking rather than broad claims of immediate clinical implementation. Major SNP callers should be compared with each other in the same datasets, on the same sequencing platforms, with the same coverage, genome complexity, organism type, and pangenome-aware workflows, using common parameters such as sensitivity, precision, false-positive rate, runtime, memory use, and reproducibility. These targeted comparisons will help direct the choice of tools for context and more accurately identify single SNPs.

Author's contribution

SB conceptualized the review article, conducted the literature analysis, compared the bioinformatic tools, and drafted the original manuscript. SK performed a critical revision of the manuscript for refined scientific arguments and handled technical editing and formatting to meet journal guidelines.

Funding

There is no funding to report.

Conflicts of interest

The authors report no conflict of interest.

Data and code availability

The software tools and algorithms reviewed in this study are all open-source and publicly available. No original datasets were generated in this review. All comparative data and benchmarks discussed are derived from the cited literature and are available through the respective publications.

Ethical approvals

This study does not involve experiments on animals or human subjects.

Acknowledgments

The authors gratefully acknowledge the Editor and the anonymous Reviewers for their thorough reviews and constructive feedback, which improved the quality of this article.

Declaration of generative ai use

The graphical abstract was created using Figurelabs.ai based on author-provided prompts and subsequently reviewed and edited by the authors.

Reference

1. Brookes AJ. The essence of SNPs. Gene. 1999;234(2):177-86. https://doi.org/10.1016/s0378-1119(99)00219-x
2. Shastry BS. SNPs: impact on gene function and phenotype. Methods Mol Biol. 2009;578:3-22. https://doi.org/10.1007/978-1-60327-411-1_1
3. Cotton CM, Murray JC. The Human Genome and Neonatal Care. In: Gleason CA, Juul SE, editors. Avery's Diseases of the Newborn (Tenth Edition). Philadelphia: Elsevier; 2018. p. 180-9.e2.
4. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53. https://doi.org/10.1126/science.abj6987
5. Syvanen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature reviews Genetics. 2001;2(12):930-42. https://doi.org/10.1038/35103535
6. Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics. 2006;7(1):438. https://doi.org/10.1186/1471-2105-7-438
7. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013;5(3):28. https://doi.org/10.1186/gm432
8. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. PLoS One. 2013;8(9):e75619. https://doi.org/10.1371/journal.pone.0075619
9. Lee MS, Flammer AJ, Lerman LO, Lerman A. Personalized medicine in cardiovascular diseases. Korean Circ J. 2012;42(9):583-91. https://doi.org/10.4070/kcj.2012.42.9.583
10. Saeed RF, Awan UA, Saeed S, Mumtaz S, Akhtar N, Aslam S. Targeted Therapy and Personalized Medicine. Cancer Treat Res. 2023;185:177-205. https://doi.org/10.1007/978-3-031-27156-4_10
11. De La Vega FM, Dailey D, Ziegle J, Williams J, Madden D, Gilbert DA. New generation pharmacogenomic tools: a SNP linkage disequilibrium Map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. Biotechniques. 2002;Suppl(sup6):48-50, 2, 4. https://www.ncbi.nlm.nih.gov/pubmed/12083398
12. Baruri S, Batra L, Adhikari S, El-Baz A. Integrative Epigenomics: Bioinformatics Strategies for Multi-Omics Data Analysis in Health and Disease. 2026. https://www.preprints.org/manuscript/202605.0428
13. Akter S. AI-driven precision medicine: transforming personalized cancer treatment. Journal of AI-Powered Medical Innovations. 2024;2(1):10-21. https://doi.org/10.60087/vol2iisue1.p21
14. Zambare V. The Transformative Role of Artificial Intelligence in Shaping Science and Technology. Nepal Journal of Biotechnology. 2023;11(2):35-6. https://doi.org/10.54796/njb.v11i2.289
15. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet. 2012;131(10):1541-54. https://doi.org/10.1007/s00439-012-1213-z
16. Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet. 2016;57(1):71-9. https://doi.org/10.1007/s13353-015-0292-7
17. Abdelwahab O, Torkamaneh D. Artificial intelligence in variant calling: a review. Front Bioinform. 2025;5:1574359. https://doi.org/10.3389/fbinf.2025.1574359
18. 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):4038. https://doi.org/10.1038/s41467-018-06159-4
19. Lorenc MT, Hayashi S, Stiller J, Lee H, Manoli S, Ruperao P, et al. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. Biology (Basel). 2012;1(2):370-82. https://doi.org/10.3390/biology1020370
20. Kwok PY, Chen X. Detection of single nucleotide polymorphisms. Curr Issues Mol Biol. 2003;5(2):43-60. https://www.ncbi.nlm.nih.gov/pubmed/12793528
21. Pickering J, Bamford A, Godbole V, Briggs J, Scozzafava G, Roe P, et al. Integration of DNA ligation and rolling circle amplification for the homogeneous, end-point detection of single nucleotide polymorphisms. Nucleic Acids Res. 2002;30(12):e60. https://doi.org/10.1093/nar/gnf060
22. Gruening B, Sallou O, Moreno P, da Veiga Leprevost F, Menager H, Sondergaard D, et al. Recommendations for the packaging and containerizing of bioinformatics software. F1000Res. 2018;7:ISCB Comm J-742. https://doi.org/10.12688/f1000research.15140.2
23. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008
24. Lefouili M, Nam K. The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. Sci Rep. 2022;12(1):11331. https://doi.org/10.1038/s41598-022-15563-2
25. Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, et al. Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics. 2020;21(1):360. https://doi.org/10.1186/s12859-020-03704-1
26. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012. https://doi.org/10.48550/arXiv.1207.3907
27. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11 0 1- 0 33. https://doi.org/10.1002/0471250953.bi1110s43
28. Supernat A, Vidarsson OV, Steen VM, Stokowy T. Comparison of three variant callers for human whole genome sequencing. Scientific Reports. 2018;8(1):17851. https://doi.org/10.1038/s41598-018-36177-7
29. Stram DO. Software for tag single nucleotide polymorphism selection. Hum Genomics. 2005;2(2):144-51. https://doi.org/10.1186/1479-7364-2-2-144
30. Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, et al. Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. Plant Biotechnol J. 2014;12(6):787-96. https://doi.org/10.1111/pbi.12183
31. Lohmueller KE, Wong LJ, Mauney MM, Jiang L, Felder RA, Jose PA, et al. Patterns of genetic variation in the hypertension candidate gene GRK4: ethnic variation and haplotype structure. Ann Hum Genet. 2006;70(Pt 1):27-41. https://doi.org/10.1111/j.1529-8817.2005.00197.x
32. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225-9. https://doi.org/10.1126/science.1069424
33. Stram DO. Multi-SNP Haplotype Analysis Methods for Association Analysis. Methods Mol Biol. 2017;1666:485-504. https://doi.org/10.1007/978-1-4939-7274-6_24
34. Gaudillo J, Rodriguez JJR, Nazareno A, Baltazar LR, Vilela J, Bulalacao R, et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction. PLoS One. 2019;14(12):e0225574. https://doi.org/10.1371/journal.pone.0225574
35. Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, et al. Machine learning in genome-wide association studies. Genet Epidemiol. 2009;33 Suppl 1(S1):S51-7. https://doi.org/10.1002/gepi.20473
36. Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP. Application of machine learning in SNP discovery. BMC Bioinformatics. 2006;7(1):4. https://doi.org/10.1186/1471-2105-7-4
37. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. biorxiv. 2018:201178. https://doi.org/10.1101/201178

38. Hall MB, Wick RR, Judd LM, Nguyen AN, Steinig EJ, Xie O, et al. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. Elife. 2024;13:RP98300. https://doi.org/10.7554/eLife.98300
39. 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):983-7. https://doi.org/10.1038/nbt.4235
40. Lin YL, Chang PC, Hsu C, Hung MZ, Chien YH, Hwu WL, et al. Comparison of GATK and DeepVariant by trio sequencing. Sci Rep. 2022;12(1):1809. https://doi.org/10.1038/s41598-022-05833-4
41. 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 Genomics. 2022;23(1):155. https://doi.org/10.1186/s12864-022-08365-3
42. Warden CD, Adamson AW, Neuhausen SL, Wu X. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. PeerJ. 2014;2:e600. https://doi.org/10.7717/peerj.600
43. Brouard J-S, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. Journal of animal science and biotechnology. 2019;10(1):44. https://doi.org/10.1186/s40104-019-0359-0
44. do Valle IF, Giampieri E, Simonetti G, Padella A, Manfrini M, Ferrari A, et al. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinformatics. 2016;17(Suppl 12):341. https://doi.org/10.1186/s12859-016-1190-7
45. Liu J, Shen Q, Bao H. Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. PLoS One. 2022;17(1):e0262574. https://doi.org/10.1371/journal.pone.0262574
46. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32(11):1749-51. https://doi.org/10.1093/bioinformatics/btw044
47. Gaston JM, Alm EJ, Zhang AN. Fast and accurate variant identification tool for sequencing-based studies. BMC Biol. 2024;22(1):90. https://doi.org/10.1186/s12915-024-01891-4
48. Nunn A, Otto C, Fasold M, Stadler PF, Langenberger D. Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional bayesian approaches. BMC Genomics. 2022;23(1):477. https://doi.org/10.1186/s12864-022-08691-6
49. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. https://doi.org/10.1093/bioinformatics/btp324
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(16):2078-9. https://doi.org/10.1093/bioinformatics/btp352
51. Brambrink L, Kolesnikov A, Goel S, Nattestad M, Yun T, Baid G, et al. DeepTrio: variant calling in families using deep learning. BioRxiv. 2021:2021.04. 05.438434. https://doi.org/10.1101/2021.04.05.438434
52. Franke KR, Crowgey EL. Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. Genomics Inform. 2020;18(1):e10. https://doi.org/10.5808/GI.2020.18.1.e10
53. 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. https://doi.org/10.1101/115717
54. Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, et al. Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. Front Genet. 2019;10:736. https://doi.org/10.3389/fgene.2019.00736
55. Mitchell K, Brito JJ, Mandric I, Wu Q, Knyazev S, Chang S, et al. Benchmarking of computational error-correction methods for next-generation sequencing data. Genome Biol. 2020;21(1):71. https://doi.org/10.1186/s13059-020-01988-3
56. Xu X, Bai G. Whole-genome resequencing: changing the paradigms of SNP detection, molecular mapping and gene discovery. Molecular Breeding. 2015;35(1):33.
57. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019;10(1):4660. https://doi.org/10.1038/s41467-019-12493-y
58. Kim HM, Jeon S, Chung O, Jun JH, Kim HS, Blazyte A, et al. Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. Gigascience. 2021;10(3):giab014. https://doi.org/10.1093/gigascience/giab014
59. Barroso I. The importance of increasing population diversity in genetic studies of type 2 diabetes and related glycaemic traits. Diabetologia. 2021;64(12):2653-64. https://doi.org/10.1007/s00125-021-05575-4
60. Ding W, Li X, Zhang J, Ji M, Zhang M, Zhong X, et al. Adaptive functions of structural variants in human brain development. Sci Adv. 2024;10(14):eadl4600. https://doi.org/10.1126/sciadv.adl4600
61. Zhang T, Li H, Ma S, Cao J, Liao H, Huang Q, et al. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. Appl Environ Microbiol. 2023;89(10):e0060523. https://doi.org/10.1128/aem.00605-23
62. Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020;7(1):399. https://doi.org/10.1038/s41597-020-00743-4
63. 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(10):1155-62. https://doi.org/10.1038/s41587-019-0217-9
64. Bogaerts B, Van den Bossche A, Verhaegen B, Delbrassinne L, Mattheus W, Nouws S, et al. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. J Clin Microbiol. 2024;62(5):e0157623. https://doi.org/10.1128/jcm.01576-23
65. Sierra R, Roch M, Moraz M, Prados J, Vuilleumier N, Emonet S, et al. Contributions of Long-Read Sequencing for the Detection of Antimicrobial Resistance. Pathogens. 2024;13(9):730. https://doi.org/10.3390/pathogens13090730
66. Mostafa HH. An evolution of Nanopore next-generation sequencing technology: implications for medical microbiology and public health. J Clin Microbiol. 2024;62(5):e0024624. https://doi.org/10.1128/jcm.00246-24
67. Beroukhim R, Lin M, Park Y, Hao K, Zhao X, Garraway LA, et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. PLoS Comput Biol. 2006;2(5):e41. https://doi.org/10.1371/journal.pcbi.0020041
68. Amare AT, Schubert KO, Baune BT. Pharmacogenomics in the treatment of mood disorders: Strategies and Opportunities for personalized psychiatry. EPMA J. 2017;8(3):211-27. https://doi.org/10.1007/s13167-017-0112-8
69. Wang H, An J, Zhong S, Qiao S, Zhang L, Zhao L, et al. The influence of COMT and ABCB1 gene polymorphisms on sufentanil analgesic effect for postoperative pain in children with fracture. Medicine. 2024;103(17):e37980. https://doi.org/10.1097/MD.0000000000037980
70. Hicks JK, Sangkuhl K, Swen JJ, Ellingrod VL, Muller DJ, Shimoda K, et al. Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. Clinical pharmacology and therapeutics. 2017;102(1):37-44. https://doi.org/10.1002/cpt.597
71. Ahmed S, Zhou Z, Zhou J, Chen SQ. Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. Genomics Proteomics Bioinformatics. 2016;14(5):298-313. https://doi.org/10.1016/j.gpb.2016.03.008
72. Nelson CW, Moncla LH, Hughes AL. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. Bioinformatics. 2015;31(22):3709-11. https://doi.org/10.1093/bioinformatics/btv449
73. Vogels CBF, Breban MI, Ott IM, Alpert T, Petrone ME, Watkins AE, et al. Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. PLoS Biol. 2021;19(5):e3001236. https://doi.org/10.1371/journal.pbio.3001236
74. Knez K, Spasic D, Janssen KP, Lammertyn J. Emerging technologies for hybridization based single nucleotide polymorphism detection. Analyst. 2014;139(2):353-70. https://doi.org/10.1039/c3an01436c

75. Rech GE, Radio S, Guirao-Rico S, Aguilera L, Horvath V, Green L, et al. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. Nat Commun. 2022;13(1):1948. https://doi.org/10.1038/s41467-022-29518-8
76. Matalon DR, Zepeda-Mendoza CJ, Aarabi M, Brown K, Fullerton SM, Kaur S, et al. Clinical, technical, and environmental biases influencing equitable access to clinical genetics/genomics testing: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2023;25(6):100812. https://doi.org/10.1016/j.gim.2023.100812
77. Maurstad MF, Hoff SNK, Cerca J, Ravinet M, Bradbury I, Jakobsen KS, et al. Reference genome bias in light of species-specific chromosomal reorganization and translocations. Genome Biol. 2025;26(1):355. https://doi.org/10.1186/s13059-025-03761-w
78. Kousathanas A, Leuenberger C, Link V, Sell C, Burger J, Wegmann D. Inferring Heterozygosity from Ancient and Low Coverage Genomes. Genetics. 2017;205(1):317-32. https://doi.org/10.1534/genetics.116.189985
79. Song K, Li L, Zhang G. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. Sci Rep. 2016;6(1):35736. https://doi.org/10.1038/srep35736
80. Qi J, Chen Y, Copenhaver GP, Ma H. Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. Proc Natl Acad Sci U S A. 2014;111(27):10007-12. https://doi.org/10.1073/pnas.1321897111
81. You FM, Huo N, Deal KR, Gu YQ, Luo MC, McGuire PE, et al. Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. BMC Genomics. 2011;12(1):59. https://doi.org/10.1186/1471-2164-12-59
82. Bundock PC, Casu RE, Henry RJ. Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. Plant Biotechnol J. 2012;10(6):657-67. https://doi.org/10.1111/j.1467-7652.2012.00707.x
83. Karbstein K, Tomasello S, Hodac L, Wagner N, Marincek P, Barke BH, et al. Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large Ranunculus auricomus species complex. New Phytol. 2022;235(5):2081-98. https://doi.org/10.1111/nph.18284
84. Tiwari VK, Heesacker A, Riera-Lizarazu O, Gunn H, Wang S, Wang Y, et al. A whole-genome, radiation hybrid mapping resource of hexaploid wheat. Plant J. 2016;86(2):195-207. https://doi.org/10.1111/tpj.13153
85. Samwald M, Blagec K, Hofer S, Freimuth RR. Analyzing the potential for incorrect haplotype calls with different pharmacogenomic assays in different populations: a simulation based on 1000 Genomes data. Pharmacogenomics. 2015;16(15):1713-21. https://doi.org/10.2217/pgs.15.108
86. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312-24. https://doi.org/10.1038/s41586-023-05896-x
87. Gong Y, Li Y, Liu X, Ma Y, Jiang L. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? J Anim Sci Biotechnol. 2023;14(1):73. https://doi.org/10.1186/s40104-023-00860-1
88. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11(1):70. https://doi.org/10.1186/s13073-019-0689-8
89. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel, Switzerland). 2020;23(1). https://doi.org/10.3390/e23010018
90. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020;21(1):139-62. https://doi.org/10.1146/annurev-genom-120219-080406
91. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol. 2024;42(4):663-73. https://doi.org/10.1038/s41587-023-01793-w
92. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet. 2022;54(4):518-25. https://doi.org/10.1038/s41588-022-01043-w
93. Kamaraj V, Gupta A, Raman K, Narayanan M, Sinha H. GViNC: an innovative framework for genome graph comparison reveals hidden patterns in the genetic diversity of human populations. NAR Genom Bioinform. 2025;7(3):lqaf121. https://doi.org/10.1093/nargab/lqaf121
94. Nyaga DM, Zaied RE, Silander OK, Black MA, O'Sullivan JM. Beyond single references: pangenome graphs and the future of genomic medicine. Front Genet. 2025;16:1679660. https://doi.org/10.3389/fgene.2025.1679660
95. Wang J, Yang W, Zhang S, Hu H, Yuan Y, Dong J, et al. A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biol. 2023;24(1):19. https://doi.org/10.1186/s13059-023-02861-9