Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38442

Artificial Intelligence in Genetic Testing: From Variant Calling to Variant Interpretation

Ziyi Peng

Immersion Academy, Irvine, USA
Joey179820105131@outlook.com

Abstract. Genetic testing may help diagnose diseases by finding genetic variants, with variant calling and variant interpretation as its main steps. Variant calling compares sequencing data with a reference genome and identifies possible variants. Variant interpretation then assesses whether these variants may cause disease by using different types of evidence, such as allele frequency, phenotype matching, review status, and familial evidence. However, sequencing noise, alignment problems, and complex disease mechanisms can still affect both steps and reduce the accuracy of genetic testing. Artificial Intelligence (AI) tools can find patterns across these processes, but they currently work mainly as decision-support tools rather than independent evaluators. They also face three main problems: poor performance across different datasets, unexplainable outputs, and frequent errors in training data. New multimodal models and foundation models may help combine more types of information. Still, researchers need to confirm their safety and performance in clinical settings. This review examines the value of AI tools in genetic testing and discusses the main barriers to their independent use. It also explains why human expertise remains essential for final clinical diagnosis.

Keywords: Artificial intelligence, genetic testing, variant calling, variant interpretation

1. Introduction

Genetic testing, the process of comparing an individual's Deoxyribonucleic Acid (DNA) sequence with a reference genome to identify variants and analyzing these variants to evaluate whether they are pathogenic. This process is central to understanding disease mechanisms by allowing researchers to link genetic variants with relevant diseases. Clinically, samples of DNA are collected and sequenced to generate sequencing reads. After that, variant calling compares sequencing reads with the reference genome to find out candidate variants. Finally, these variants are classified in variant interpretation according to their clinical significance into 5 categories defined by the American College of Medical Genetics and Genomics / Association for Molecular Pathology (ACMG/AMP) standards: benign, likely benign, Variants of Uncertain Significance (VUSs), likely pathogenic, or pathogenic [1, 2]. But one individual's genome contains millions of differences compared with reference genome [3]. As a result, relying solely on manual analysis is challenging and more efficient solutions are needed.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38442

More and more Artificial Intelligence (AI) tools are starting to fill this gap: they do a good job at finding possible genetic variants and working out how likely these changes are to cause disease, based on lots of different evidence [4, 5]. A bunch of different signals like Allele Fraction (AF), Base Quality (BQ), and Mapping Quality (MQ), get put together to figure out if a candidate genetic variant is real when calling variants [4]. When interpreting these variants, AI tools figure out how likely they are to cause disease by looking at how well they match the known phenotype, what family evidence there is, and other information [5].

Even so, these tools still run into a lot of problems before they can be used by themselves [6]. One problem that makes reading alignments less dependable is that most short-read sequencers can only make short reads, normally between tens and hundreds of base pairs long [7]. Because read alignment matches up sequencing reads with a reference genome to find where they are supposed to go [8], shorter reads make an accurate comparison harder. Sequencing mistakes can also make later analysis less reliable and give low BQ values [8]. When these technical issues mess up variant calling, the wrong answers carry over into variant interpretation. That job gets even harder because how strong the evidence for benign or pathogenic variants can change greatly [2, 9]. For example, ClinVar's review status can go from one submitter to expert panels. In real clinical work, results from these tools have to get put together with functional evidence, family background, and expert judgment from clinicians [2]. Given these performance concerns, the review summarizes how AI can support genetic testing and explains why these tools cannot yet work in isolation.

This review has 6 different parts. Right after this, the next section will talk about how AI helps variant calling, and section 3 looks at the part AI tools play when in variant interpretation. Then sections 4 and 5 talk about the problems these tools are facing and where these tools might go next, in that same order. Section 6 wraps this review up with some closing thoughts.

2. Variant calling

2.1. Overview of variant calling

Variant calling shown in Figure 1 is the process of identifying candidate variants from sequencing data that contains sequencing noise. With large numbers of short reads and corresponding BQ scores for each base, the next step is alignment [10, 11]. Then, metrics such as AF, BQ, MQ, and depth are integrated by variant callers to obtain a Quality Score (QUAL), representing the confidence of the call, and to generate a Variant Call Format (VCF) output [11, 12]. Using different types of neural networks to learn complex patterns, emerging AI tools are providing new approaches for variant calling [4].

There are two main types of variants: small variants, the variant that appears on a small scale, including Single Nucleotide Variants (SNVs, changes involving one nucleotide) and InDels (insertions or deletions of less than 50 bp); and Structural Variants (SVs), which are generally variants larger than 50 bp, including Copy Number Alterations (CNAs, which involve changes in the number of copies of genomic segments relative to the reference genome) [13, 14]. Single Nucleotide Polymorphism (SNP) often refers to a common single-nucleotide variant [15].

DOI: 10.54254/2755-2721/2026.38442

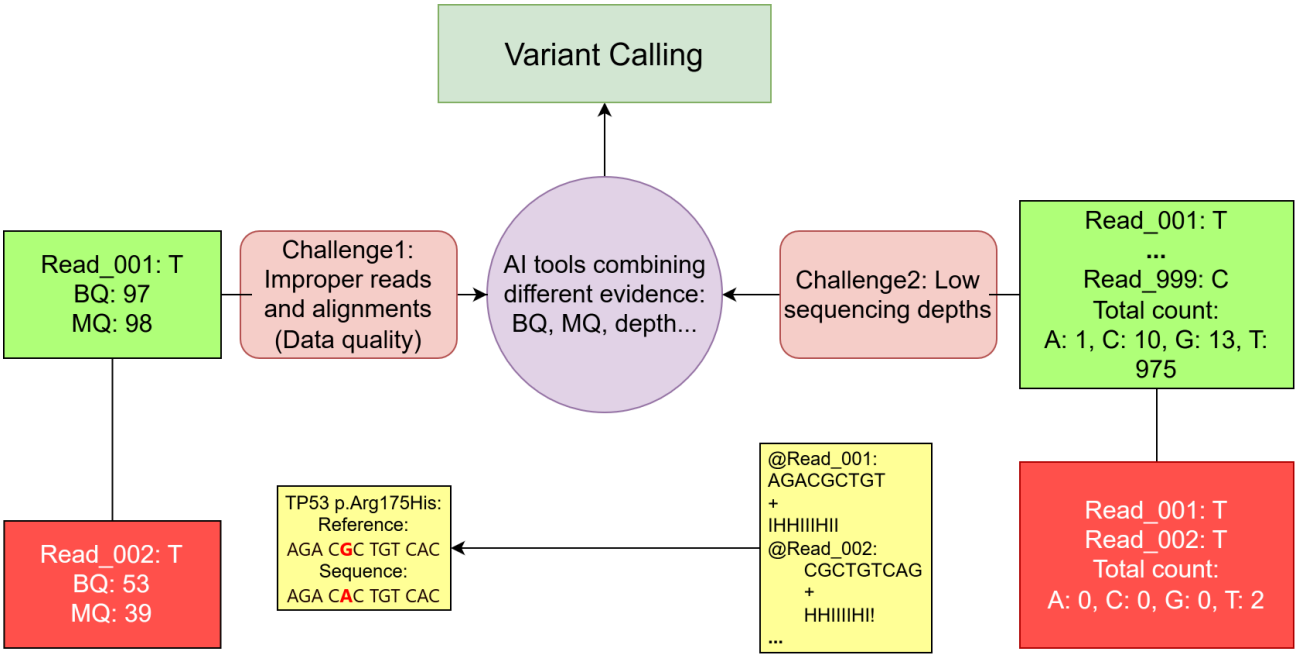


Figure 1. Overview of variant calling, including AI usage, associated challenges, and some specific examples (picture credit: original)

Table 1 shows that AI tools do not necessarily need to replace the whole pipeline; in contrast, they can focus on different stages such as refining and data denoising and imputation with different scopes to achieve promising strengths, while some risks prevent them from being used independently.

Table 1. The overview of AI tools in variant calling discussed in this review

Tool	Best for	Role in pipeline	Key strength	Main risk
DeepVariant [4]	SNPs and InDels	End-to-end variant calling pipeline	High accuracy across multiple sequencing technologies	Dependence on read alignment quality
VariantTransformer [16]	SNPs and InDels	Refining VCF	Improved performance on low-depth sequencing data	Performance gaining may decrease when sequencing depth is sufficient
scTCA [17]	Single-cell CNAs	Data denoising and imputation	Detect focal copy number alterations	Not designed for detecting SNVs and InDels

2.2. AI tools focusing on SNPs and InDels

AI tools based on different types of models can focus on SNPs and small InDels with relatively high accuracy. For example, DeepVariant is a model based on a Convolutional Neural Network (CNN) [4]. By combining aligned reads at candidate variant sites with a reference genome and converting them into pileup images (image-like representations), the model uses CNN as an innovative way to predict the probability of homozygous reference (hom-ref), heterozygous (het), homozygous alternative (hom-alt) [4]. It demonstrated competitive performance compared with contemporary models with high F1 score across multiple benchmark datasets, including Illumina whole-genome

DOI: 10.54254/2755-2721/2026.38442

sequencing, 10X Genomics whole-genome sequencing, and Ion AmpliSeq exome sequencing data [4]. This example shows that appropriate data representation and model architecture may improve performance in variant calling.

Of course, not all AI tools work in the same pipeline. VariantTransformer is such an example [16]: an AI tool that refines variant calls generated by existing variant callers, especially at sites with relatively low depth (10×–15×). With the Transformer architecture, VariantTransformer improved the accuracy of variant refinement at sites with relatively low depth [16]. It also achieved competitive performance compared with other approaches [16]. These results show that model architecture can play an important role in variant calling. Therefore, AI tools do not need to fully replace traditional variant callers to be useful. They can also work as downstream refinement modules.

In clinical practice, multiple AI tools may be used together to complement each other [4, 16]. For example, clinicians may apply VariantTransformer after DeepVariant to use advantages of both tools: sensitive initial variant detection capability of DeepVariant and refinement capability of VariantTransformer [4, 16]. VariantTransformer may help filter some potential false positive variants out after they were detected by DeepVariant [4, 16].

2.3. AI tools focusing on SVs

There is another type of AI models that focus on data denoising and imputation to facilitate downstream variant calling [17]. Unlike DeepVariant and VariantTransformer [4, 16], scTCA is an AI tool designed for denoising and imputing single-cell DNA sequencing data [17]. It takes the raw single-cell read counts matrix, which can be used to infer CNAs in cells, as input [17]. However, these data tend to be noisy and sparse due to technical limitations. CNN and Transformer were built into the model to capture local and global patterns [17]. It outputs denoised and imputed data that performs better than traditional scRNA-seq tools not designed for this task [17]. These results suggest that AI can not only improve direct variant calling approaches but also improve sequencing-derived data quality.

3. Variant interpretation

3.1. Overview of variant interpretation

Variant interpretation shown in Figure 2 integrates multiple types of evidence to evaluate whether a variant is benign or pathogenic [2]. Combining genetic, functional, population, and phenotypic evidence together is required because different types of variants can lead to different molecular consequences [2]. Population frequency can facilitate ruling out certain pathogenic hypotheses, but rarity alone is not sufficient evidence for pathogenicity [18]. Literature evidence (the label and strength of previously reported evidence) also matters [9].

AI tools based on different neural network architectures contribute to variant interpretation by integrating available features to generate pathogenicity scores [6]. This score, together with other original evidence, assists human experts in final classification of variants [2]. Computational scores play a key role in prioritization, but they should never be treated as standalone clinical classifications [2, 6].

DOI: 10.54254/2755-2721/2026.38442

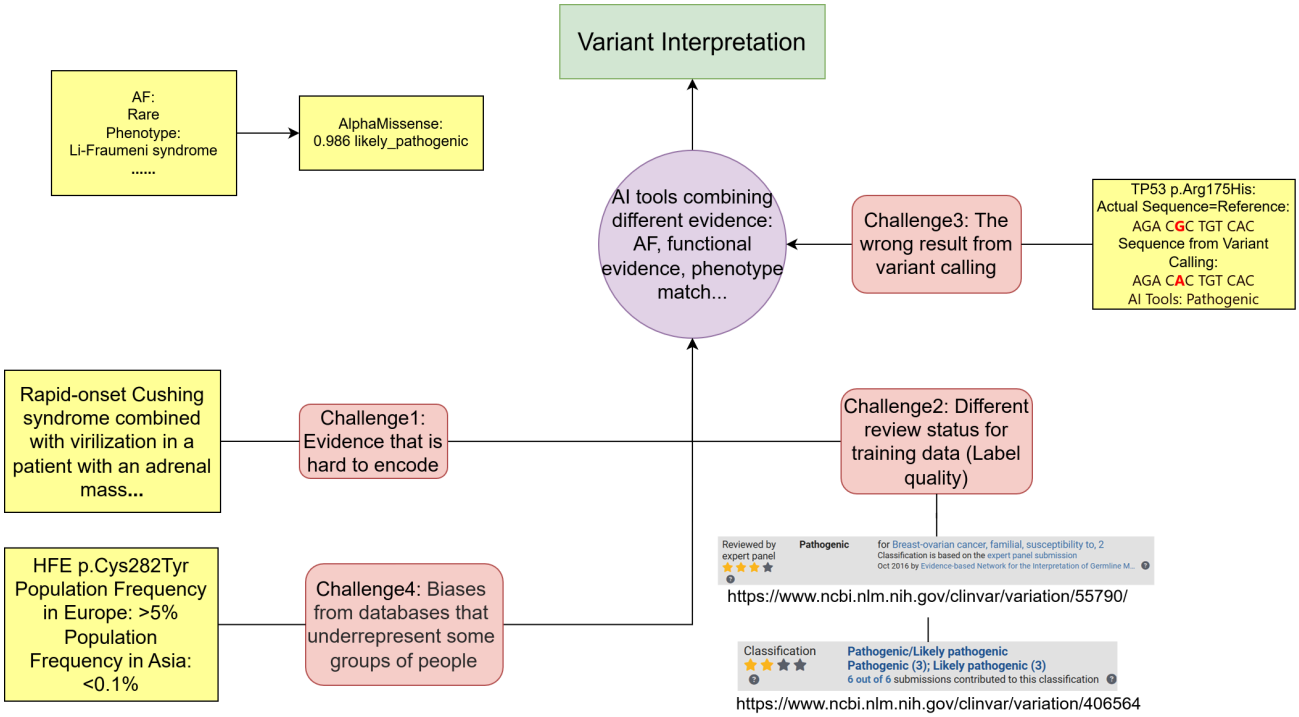


Figure 2. Overview of variant interpretation, including AI usage, associated challenges, and some specific examples (picture credit: original)

Table 2 shows that AI tools can use different input data, including DNA and protein sequence contexts and pathogenicity scores from other tools, to focus on different biological mechanisms. All these tools have limits along with their particular strengths.

Table 2. The overview of AI tools in variant interpretation discussed in this review

Tool	Best for	Input	Output	Key strength	Main risk
AlphaMissense [19]	Missense variants	Protein sequence context and evolutionary information	Pathogenicity score	High precision while maintaining broad coverage of missense variants	Limited to missense variants
Rare Exome Variant Ensemble Learner (REVEL) [20]	Rare Mendelian disease-associated missense variants	Scores from other 13 tools	Final Pathogenicity score	Improved performance compared with individual predictors	Dependent on the quality and availability of component predictors
SpliceAI [21]	Splicing variants	DNA sequence context surrounding variants	Splice site probability scores	High accuracy in predicting splice-altering variants	Requires further clinical validation
Combined Annotation Dependent Depletion (CADD) [22]	SNVs and small InDels	VCF and genomic annotations	Deleteriousness score	Genome-wide applicability	Imperfect proxy training labels

DOI: 10.54254/2755-2721/2026.38442

3.2. AI tools focusing on missense variants

In the field of variant interpretation, protein structure becomes important because changes in amino acid sequence may alter protein structure and subsequently affect protein function. Developed from Google's AlphaFold [23], AlphaMissense is an AI tool designed for missense variant interpretation [19]. It adapts the framework of AlphaFold [23] to predict the pathogenicity scores of missense variants. With its AlphaFold-derived architecture, AlphaMissense classified 89% of possible human missense variants into likely benign or likely pathogenic while keeping a 90% precision threshold [19]. In other words, it maintains a balance between high precision and broad coverage across the human missense variant landscape rather than being limited to specific genes. AlphaMissense shows that adding protein structure information may improve the performance of variant interpretation [19].

Using ensemble learning is also a nice approach to this problem. REVEL takes the predicted scores for the pathogenicity of missense variants from 13 already existing tools, uses those as features and gets a final prediction score [20]. It uses a random forest that has 1,000 trees inside, with a focus on rare missense variants [20]. This tool works better than all 13 individual predictors it combines, and it also outperforms another 7 different ensemble methods. From how well it works, it can be seen that AI tools can actually use ensemble learning to get better results when they interpret variants, instead of just depending on a single model.

3.3. AI tools focusing on splicing variants

If AI tools are matched with the biological mechanisms of variants, they work much better. A good example of this is SpliceAI [21]. This tool uses the surrounding DNA sequence to predict if a variant might break RNA splicing. To look at sequence information 10,000 base pairs around a site, SpliceAI uses a dilated CNN [21]. This particular setup is useful because it lets the tool pick up long-range information that many older methods don't catch. Because of this, SpliceAI finds more signals that are connected to splice regulation. When it was tested on different datasets, around 75% of the predicted cryptic splice variants were confirmed by RNA-seq. This shows a clear improvement over older tools like MaxEntScan, NNSplice, and GeneSplicer [21].

3.4. AI tools for all SNVs and InDels at all locations

It's trickier to figure out how non-coding variants change genetic function, because they affect genes through complicated controlling steps like transcription and gene expression [24]. CADD helps solve this problem by figuring out how harmful SNVs and InDels are, even the ones in non-coding parts, by putting together 60 different bits of information [5, 22]. These information bits fall into four groups: Conservation, Functional Prediction, Epigenetic Modification, and Genetic Context [5, 22]. From all that, researchers get scores that shows how harmful variants are. By giving every SNV and InDel a deleteriousness score, CADD lets researchers prioritize variants across the whole genome.

In clinical practice, different tools may be mixed together because each one looks at different sides of what variants do. Like, a common step-by-step process might start with CADD, which sorts out which candidate variants matter most across different possible pathways that cause disease [5, 22]. After that, clinicians can use tools that are built for specific jobs: SpliceAI for splicing variants [21], and AlphaMissense and REVEL for missense variants [19, 20].

DOI: 10.54254/2755-2721/2026.38442

4. Challenges

4.1. Base and mapping quality

In variant calling, there are some trouble regions. Low depth or reduced base or mapping quality, causing by sequencing errors, alignment difficulties, and platform-specific biases, make it hard for AI tools to find variants reliably [7, 25]. Having long read-lengths, Third-Generation Sequencing (TGS) may provide better data sources with longer reads and improved resolution than NGS. But TGS also has drawbacks: high cost and unique InDel problems when dealing with homopolymers in some sequencing platforms [7, 26]. A balance between lower cost and better base and mapping quality is needed.

4.2. Label quality

How good the labels are affects how reliable variant interpretation is. ClinVar's review status shows how strong the supporting evidence is. There are many things that decide this strength, like whether the assertion criteria are available, how detailed the evidence is documented, if experts have reviewed it, and if official practice guidelines provide support [9]. For example, an annotation provided by a single submitter has less weight than one reviewed by an expert panel. When it comes to AI tools, understanding how this evidence hierarchy works is important [27]. If researchers do not pay attention to how strong the evidence is, the model might end up learning from wrong annotations and adding noise into the training step [28].

4.3. Generalization ability

Whether AI tools work well in real world situations depends on if they can work across many different datasets and sequencing platforms. Right now, AI tools can get relatively strong performance when they are working with SNVs and small InDels across different platforms, but special training methods are usually needed to adjust for each sequencing technology [4]. Getting this kind of generalization for SVs is even harder, because different sequencing technologies give out different kinds of input signals [29].

When moving to interpreting variants, there are new problems that emerge. First off, different ClinVar review statuses have different weights, which makes it hard for AI models to keep giving consistent results [9]. Different kinds of variants also need different approaches — like missense and splicing variants, for example. That is why many AI tools only look at specific types of variants [19, 21]. Plus, some kinds of evidence, like phenotype matching and familial information, are usually written in unstructured text. This makes them really hard to organize into a standard format, because they often require integration of different types of clinical data [2, 30].

Another problem is that some ancestry groups do not get enough representation, which makes AI tools less reliable. These tools may get their bias from the databases that are used to build them. Because European genomes are more common in existing reference databases, AI tools' generalization ability is limited [31]. When population frequency is not well collected, AI models may get to the wrong conclusion. They may, for instance, treat ancestry-specific variants as potentially pathogenic [31]. There are two ways that can fix this: adjust the models for a local population before using, or train them right from the start with populations from all over the world.

DOI: 10.54254/2755-2721/2026.38442

4.4. Explainability

To cut down the bad impacts of clinical AI, researchers need to find out how algorithms get to their conclusions [32]. If these tools do not have clear explainability, doctors cannot figure out what caused a wrong prediction [33]. In the process of variant calling, DeepVariant [4] uses CNN. This model turns possible variants into image-like representations to be more transparent by providing an intermediate representation that resembles evidence considered by human experts. Even with that, the final decision from the CNN is still locked inside a computational black box [4, 33].

When people do variant interpretation, it gets even harder to make things clear. Because the input data and pathogenic mechanisms are complex, these tools usually put together high-level biological features. This is why giving clear explanations becomes harder [2, 33]. One line of work is training Large Language Models (LLMs) to explain decisions made by AI tools in natural language [34]. When given computational results from variant interpretation tools along with original evidence such as AF, these LLMs can produce explanations. Even so, users still need to validate the results, since LLMs may generate hallucinations [35].

For example, what Lu et al. did is contrasting Retrieval Augmented Generation (RAG) and fine-tuning methods on LLMs [36], leaning on LLMs' high language understanding ability to help explain pathogenic variants in particular situations. RAG method achieved better performance than fine-tuning on precision, cost, and effectiveness in this study [36]. Users can acquire data about a specific variant and can use these tools to judge whether a specific variant precipitated a specific disease [36]. That said, due to the lack of background knowledge in genomics, the inference process can be inaccurate at some cases [34, 35]. Furthermore, these tools are limited to integrating textual annotation data, while other important evidence sources remain challenging to incorporate [2].

5. Future directions

5.1. Multimodal models

Multimodal models can take diverse types of information as input to generate comprehensive analytical outputs [37]. For example, they can integrate genomic data with diverse clinical information, such as Human Phenotype Ontology (HPO) terms and Electronic Health Records (EHRs), to improve variant prioritization and generate more comprehensive interpretations [30]. Even so, how these models can integrate these diverse data still remains a major challenge, given how much they differ in format and quality [37]. Over time, these models may expand to proteomics, epigenetics, and other disciplines, getting more comprehensive in the process.

5.2. Foundation models

Human genomics can be inherently complex, and one reason is that a promoter can influence genes over long genomic distance through three-dimensional organization [38]. Foundation models in genomics are trained with massive genomic data to learn relationships in a broader pattern [39]. For instance, Enformer was designed to capture long-range regulatory interactions by using a transformer-based architecture with a larger sequence context than previous models [38]. Looking ahead, foundation models may help improve variant interpretation by finding patterns among massive genomic data and use these patterns to detect variants. This may provide additional evidence that may help human experts prioritizing VUSs that currently are hard to be categorized [39].

DOI: 10.54254/2755-2721/2026.38442

6. Conclusion

Currently, AI tools have demonstrated their potential in genetic testing. They support variant calling by improving pattern recognition in noisy sequencing data and refining uncertain candidate variants. Their tasks are not the same: imputing and denoising original data, validating outputs of traditional methods, and automating substantial parts of the process. Examples of DeepVariant, VariantTransformer, and sc-TCA suggest that AI does not necessarily need to replace the entire pipeline; it may simply make particular steps work better.

Compared with variant calling, variant interpretation is more complex because it needs to integrate diverse types of evidence due to the complexity of pathogenic mechanisms and the diversity of features that affect pathogenicity. They can study patterns from diverse biological evidence connected to different biological mechanisms to find pathogenicity scores that assist experts in evaluating variant pathogenicity, as seen in the examples of AlphaMissense, CADD, REVEL, and SpliceAI.

However, clinical interpretation should remain under qualified human oversight. Base quality, mapping quality, and label quality pose a challenge in model training, while generalization ability and explainability constrains these tools from using independently. Outputs from these tools should be interpreted as auxiliary evidence rather than direct clinical diagnosis. Human reviews are necessary to combine original evidence with AI predictions.

In the future, combining with more comprehensive datasets and improved sequencing technologies, multimodal models and foundation models are some possible ways to facilitate AI tools to become more secure, precise, explainable, generalizable and trustworthy by enabling these tools to integrate more relevant context.

References

[1] Austin-Tse, C. A., Jobanputra, V., Perry, D. L., et al. (2022). Best practices for the interpretation and reporting of clinical whole genome sequencing. *npj Genomic Medicine, 7*, 27.

[2] Richards, S., Aziz, N., Bale, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the ACMG and AMP. *Genetics in Medicine, 17*, 405–424.

[3] The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature, 526*, 68–74.

[4] Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., ... and DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology, 36*(10), 983-987.

[5] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. and Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*(D1), D886-D894.

[6] Pejaver, V., Byrne, A. B., Feng, B. J., et al. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *American Journal of Human Genetics, 109*(12), 2163–2177.

[7] Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*, 333–351.

[8] Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine, 12*, 91.

[9] Landrum, M. J., Lee, J. M., Benson, M., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research, 46*(D1), D1062–D1067.

[10] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754–1760.

[11] DePristo, M. A., Banks, E., Poplin, R., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*, 491–498.

[12] Danecek, P., Auton, A., Abecasis, G., et al. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156–2158.

DOI: 10.54254/2755-2721/2026.38442

[13] Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics, 7*, 85–97.
[14] Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics, 12*, 363–376.
[15] Brookes, A. J. (1999). The essence of SNPs. *Gene, 234*(2), 177–186.
[16] Abdelwahab, O., & Torkamaneh, D. (2026). A transformers-based framework for refinement of genetic variants. *Frontiers in Bioinformatics, 5*, 1694924.
[17] Yu, Z., Liu, F., & Li, Y. (2024). scTCA: A hybrid transformer-CNN architecture for imputation and denoising of scDNA-seq data. *Briefings in Bioinformatics, 25*(6), bbae577.
[18] Whiffin, N., Minikel, E. V., Walsh, R., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine, 19*, 1151–1158.
[19] Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., ... & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science, 381*(6664), eadg7492.
[20] Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., ... & Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics, 99*(4), 877–885.
[21] Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... & Farh, K. K. H. (2019). Predicting splicing from primary sequence with deep learning. *Cell, 176*(3), 535–548.
[22] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics, 46*, 310–315.
[23] Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589.
[24] Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., et al. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences, 111*(17), 6131–6138.
[25] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology, 14*, R51.
[26] Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in Bioinformatics, 17*(1), 154–179.
[27] Biesecker, L. G., Harrison, S. M., & ClinGen Sequence Variant Interpretation Working Group. (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genetics in Medicine, 20*, 1687–1688.
[28] Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 25*(5), 845–869.
[29] Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics, 19*, 329–346.
[30] Köhler, S., Gargano, M., Matentzoglu, N., et al. (2021). The human phenotype ontology in 2021. *Nucleic Acids Research, 49*(D1), D1207–D1217.
[31] Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell, 177*(1), 26–31.
[32] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*, 310.
[33] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*(4), e1312.
[34] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine, 29*, 1930–1940.
[35] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248, 1–38.
[36] Lu, S., & Cosgun, E. (2025). Boosting GPT models for genomics analysis: Generating trusted genetic variant annotations and interpretations through RAG and fine-tuning. *Bioinformatics Advances, 5*(1), vbaf019.
[37] Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*, 259–265.
[38] Avsec, Ž., Agarwal, V., Visentin, D., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods, 18*, 1196–1203.
[39] Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. (2025). Nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods, 22*, 287–297.