Dukda *et al. Human Genomics* (2025) 19:97
https://doi.org/10.1186/s40246-025-00811-z

Human Genomics

REVIEW Open Access

Increasing pathogenic germline variant diagnosis rates in precision medicine: current best practices and future opportunities

Sonam Dukda[1], Manoharan Kumar[1], Andrew Calcino[1], Ulf Schmitz[1,2,3] and Matt A. Field[1,4,5*]

Abstract

The accurate diagnosis of pathogenic variants is essential for effective clinical decision making within precision medicine programs. Despite significant advances in both the quality and quantity of molecular patient data, diagnostic rates remain suboptimal for many inherited diseases. As such, prioritisation and identification of pathogenic disease-causing variants remains a complex and rapidly evolving field. This review explores the latest technological and computational options being used to increase genetic diagnosis rates in precision medicine programs.

While interpreting genetic variation via standards such as ACMG guidelines is increasingly being recognized as a gold standard approach, the underlying datasets and algorithms recommended are often slow to incorporate additional data types and methodologies. For example, new technological developments, particularly in single-cell and long-read sequencing, offer great opportunity to improve genetic diagnosis rates, however, how to best interpret and integrate increasingly complex multi-omics patient data remains unclear. Further, advances in artificial intelligence and machine learning applications in biomedical research offer enormous potential, however they require careful consideration and benchmarking given the clinical nature of the data. This review covers the current state of the art in available sequencing technologies, software methodologies for variant annotation/prioritisation, pedigree-based strategies and the potential role of machine learning applications. We describe a key set of design principles required for a modern multi-omic precision medicine framework that is robust, modular, secure, flexible, and scalable. Creating a next generation framework will ensure we realise the full potential of precision medicine into the future.

*Correspondence:
Matt A. Field
matt.field@jcu.edu.au
[1]Centre for Tropical Bioinformatics and Molecular Biology, College Science and Engineering, James Cook University, Cairns, QLD, Australia
[2]Centenary Institute, The University of Sydney, Camperdown, Australia
[3]Computational Biomedicine Lab, James Cook University, Townsville, Australia
[4]Immunogenomics Lab, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia
[5]Menzies School of Health Research, Charles Darwin University, Darwin, Australia

Introduction

Identifying targetable disease-causing genetic variants lies at the heart of advancing precision medicine, improving clinical diagnostics, and enhancing our understanding of genetic contributions to diseases [1]. Precision medicine entails a healthcare delivery model that relies extensively on patient specific data points to guide the development of customised therapies [2]. One key driver of such initiatives is the ability to pinpoint pathogenic variants that greatly improves our mechanistic understanding of the disease process [3]. This progress has been largely enabled by the increasing affordability and accessibility of high-quality sequence data. Despite this progress, conclusively linking genetic variants with disease remains resource-intensive and time-consuming [4]. The most significant challenge remains differentiating the key genetic drivers from the large volumes of background genetic variation naturally present in every person. Further challenges exist with variant detection, annotation and prioritisation methods with the lack of global standards resulting in an over reliance on variable bespoke in-house solutions [5]. Groups such as the American College of Medical Genetics and Genomics (ACMG) are addressing this by providing guidelines on variant detection and interpretation; for example they propose guidelines to establish consistent cataloguing of genetic variants, classifying variants into five categories based on the strength of the evidence for disease causation. Despite these standards, many variants remain annotated as variants of uncertain significance (VUS), lacking sufficient functional assay data required for reliable classification [6]. Additionally, recurrent false positive variants can be included in clinical databases [7].

There is also an increasing recognition of the role of large and repetitive genetic variants in driving disease, however these remain challenging to detect with current short-read sequencing technologies and are better suited to more expensive long-read sequencing approaches [8]. Additionally, non-coding variants are being recognised in driving disease including intronic variants which create cryptic splice sites as well as variants modifying regulatory elements such as enhancers and promotors [9]. Another challenge specific to complex disease is the often-little understood interactions between genetic and environmental factors. Factors such as lifestyle and surrounding environment can contribute significantly to disease development and progression; however, our understanding of these processes is limited [10].

Segregation analysis of germline variants within families plays a critical role in precision medicine by enabling the accurate interpretation of genetic findings in the context of inherited disease risk [11]. By studying how a specific variant co-segregates with a disease phenotype across multiple family members, clinicians and researchers can distinguish pathogenic mutations from benign polymorphisms, thus improving diagnostic accuracy [12]. This analysis not only helps validate the clinical relevance of a variant but also informs risk assessment, surveillance strategies, and therapeutic decisions for both affected individuals and at-risk relatives. In precision medicine, where individualized care hinges on the precise understanding of genetic contributions to disease, segregation analysis remains a cornerstone for translating genomic data into meaningful clinical outcomes. Overall family data provides evidence needed to reclassify variants, either supporting pathogenicity through clear segregation or suggesting benign status through inconsistent patterns. This improved classification directly impacts clinical decision-making, from diagnosis and treatment selection to family planning [13]. Family analysis enables confident clinical recommendations, reduces uncertainty in genetic counselling, and identifies at-risk family members who would benefit from testing or enhanced surveillance, exemplifying the promise of precision medicine.

Despite steady progress, genetic diagnostic discovery rates for many complex diseases using traditional approaches remain low and capturing "missing heritability" requires multi-pronged approaches. These include a variety of sequencing-based (e.g. long-read, single-cell sequencing) and computer-based approaches (e.g. machine learning, multi-omic workflows). Robust frameworks capable of integrating huge volumes of complex patient data, genetic variant and annotation information are urgently needed [14].

Sequencing technology

A diverse selection of sequencing technologies are now available for identifying genetic variants [15]. DNA-based solutions include whole genome sequencing (WGS), whole exome sequencing (WES), and targeted gene panels, while genetic variants can also be detected in cDNA used for RNA sequencing (RNA-Seq) [16]. Newer approaches include long-read sequencing, suitable for identifying more complex, larger genetic variants [17], and single-cell sequencing to identify rare or cell-type-specific variants [18, 19].

Current DNA-based sequencing options

Targeted gene panel sequencing is appropriate when driver genes are largely known for a disease. In such cases, gene panels can obtain high diagnostic rates and the simple deployment, interpretation, and lower costs, offers an attractive alternative to WES/WGS [16, 20, 21]. Targeted panels are typically sequenced at a high depth to identify rare variants, a process which can be combined with unique molecular identifier (UMI)-based approaches to further increase resolution [22]. There are limitations of this approach however such as their

inability to identify novel variants and large genetic variants [16, 23].

WES is effective in finding both known [24] and novel driver mutations [25] in coding regions of the genome. WES employs a targeted approach via a capture array containing most known coding exons, thus covering the majority of the coding regions [3]. Despite only accounting for ~1% of the genome, an estimated 85% of mutations responsible for diseases are thought to occur within exons [26–28]. WES is attractive with regard to price and sequence depth relative to WGS [29, 30].

WGS is an unbiased method that provides sequence data across the entire genome [31]. WGS is increasingly becoming the first choice for patient sequencing due to advantages including the ability to detect small and large genetic variants as well as achieving relatively even sequence coverage [32–34]. The diagnostic superiority of WGS to chromosomal microarray (CMA), karyotyping, targeted sequencing assays and WES [35–40] has been demonstrated. Accordingly, precision medicine programs are employing WGS as the first option resulting in the development of increasingly standardised methodologies [41, 42].

Current RNA-based sequencing options

Bulk RNA-seq is a high-throughput method used to examine the complete set of RNA transcripts within a biological sample [6]. Bulk RNA-Seq typically obtains sequence data from a mixed heterogenous population of cells in contrast to tagged individual cells as is done in single-cell RNA-Seq (scRNA-Seq) [43]. The clinical utility has been demonstrated largely for the ability to identify dysregulated genes that warrant further investigation within the genome [44]. Additionally, RNA-Seq allows the identification of aberrant splicing events such as retained introns or skipped exons and gene fusions.

Single-cell sequencing

In contrast to traditional bulk sequencing methods, single-cell technologies incorporate cell-specific barcodes to obtain per-cell sequence information for thousands of cells simultaneously. To date, most single-cell platforms utilise RNA as input (scRNA-Seq) however a growing number of platforms offer single-cell DNA sequencing (scDNA-Seq).

Single-cell RNA sequencing is an advanced technology able to evaluate transcriptional similarities and variances within a population of cells, revealing cell-type-specific levels of heterogeneity previously undetectable by bulk sequencing methods [45–47]. Nonetheless, scRNA-Seq remains technically challenging with limitations including generation of doublets and dead cells, lower sequencing depth per cell, data sparsity, high input cell requirements, and high cost. Despite these challenges, scRNA-Seq offers an unprecedented opportunity to track disease progression within heterogenous cell populations.

Single-cell DNA sequencing allows the detection of per-cell or cell-type-specific rare genetic variants from mixed heterogenous input samples [48]. Many of the same limitations are shared with scRNA-Seq, however variant detection is feasible for targeted gene panels using technologies such as Mission Bio's Tapestri platform. Additionally, significant amplification is typically required [49], a process known to introduce errors and uneven coverage resulting in challenges in downstream data analysis [50].

Long-read sequencing

Third generation single molecule long-read sequencing overcomes many of the limitations of short-read technologies [51]. Initially plagued by high error rates, continual improvements are producing progressively longer and higher quality reads, with lengths of up to 2 Megabase pairs now possible [52, 53]. The third-generation sequencing market is primarily dominated by two technologies: (i) Pacific Biosciences and (ii) Oxford Nanopore Technologies (ONT) [54], both of which offer DNA or RNA based sequencing. Long-read DNA sequencing is critical if driver variants are repetitive or complex in nature (e.g. long tandem repeats, copy number variants) or occur in repetitive gene families, GC-rich regions or pseudogenes. Long-read RNA sequencing captures full-length isoforms and can identify novel transcripts, skipped exons, retained introns and gene fusion products [55].

Prioritisation strategies

All sequencing approaches generate large numbers of genetic variants, the majority of which are not relevant for the underlying disease. Reducing the genetic search space for causal variants can be done through a combination of careful patient selection, variant annotation (at the level of variant, gene and gene network), and software development and optimisation (Fig. 1).

Sample selection strategies

Sample selection strategies are critical to increasing the likelihood of identifying disease causing variants. Common strategies include grouping unrelated individuals by phenotype, selecting patients with early onset and/or extreme phenotype and pedigree sequencing for families.

For single or unrelated individuals, strategies include sequencing samples with early onset or extreme phenotypes as well grouping patients with similar phenotypes that potentially share an underlying genetic cause [56]. Choosing unrelated individuals with a shared well-characterised phenotype requires annotation with

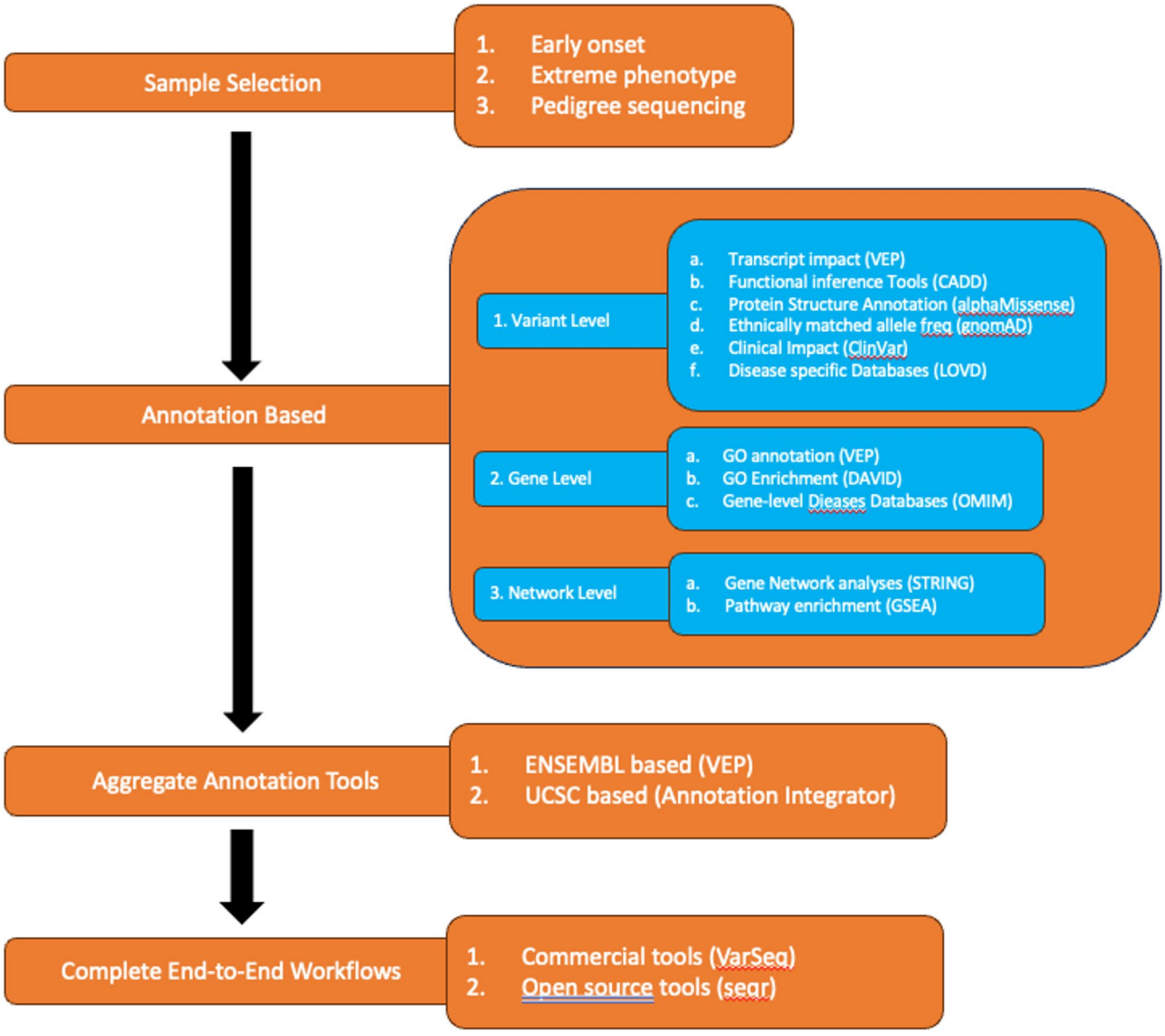

Fig. 1 Variant prioritisation strategies

standardised human phenotype ontology (HPO) terms [57]. Further defining inclusion and exclusion criteria based on clinical features - disease progression, or other relevant parameters – allows the generation of a study cohort that is relatively homogenous and well suited for detecting shared driver variants (or pathways) linked to the observed phenotype [58]. The significance of careful patient selection is highlighted in numerous studies as an effective way to increase diagnosis rates [59, 60].

Pedigree-based strategies

For related individuals, utilization of pedigree sequencing is an extremely effective strategy for reducing the genomic search space for causal variants. Pedigree sequencing is particularly useful for the identification of rare familial variants which segregate with the phenotype of interest [61, 62]. This approach yields additional information including inheritance modes and can track the segregation of variants within families, however custom software is required [12]. Sequencing of a proband child with healthy parents is very successful for rare, early onset diseases by focusing on a small number *de novo* variants. Similarly, consanguineous pedigree sequencing reduces the search space to homozygous variants [63].

Segregation analysis plays a vital role in interpreting and prioritising genetic variants by examining how a variant is inherited within a family and whether it consistently appears in affected individuals while being absent in those unaffected. Such co-segregation patterns strengthens the case for pathogenicity, particularly when it is aligned with the expected pattern of inheritance [64, 65]. In the line with ACMG/AMP guidelines, segregation

data can count towards variant classification as 'supporting evidence (PP1)', with increasing weight assigned as more segregation evidence is observed [66]. Recognition of segregating *de novo* or very rare variants in dominant conditions further strengthens the case for pathogenicity (PS2) [65]. In contrast, variants that do not segregate with disease phenotype in affected family members may indicate benign status, reduced penetrance, or the presence of phenocopies—individuals who exhibit symptoms due to non-genetic factors or different underlying genetic causes [67]. Such scenarios may point to potential locus heterogeneity, where similar phenotypes arise from mutations in different genes [68, 69]. The identification of such cases is vital for accurate genetic diagnosis. When causal variants are present in asymptomatic carriers, we can estimate penetrance and variability in disease severity thus enabling personalized monitoring strategies and more accurate prognoses for at-risk relatives [70].

Annotation-based strategies – variant level

Variants are first annotated based on the exact genomic coordinate and the observed nucleotide change. This includes assessing impact on transcripts, estimating impact on protein function, comparing to population variant levels and overlapping with clinical variant databases.

Variants are typically overlapped to a transcript model and stratified according to their impact using popular tools such as SnpEff, VEP or ANNOVAR [71–73]. High level genomic overlaps are first considered (e.g. intergenic, intronic, exonic, etc.) and further refined by transcript effect when appropriate (e.g. exonic SNVs classified as synonymous or non-synonymous, exonic indels classified as frameshift or non-frameshift). Next, variants are run through functional inference tools to predict potential functional consequence of genetic variants. These tools utilise information including evolutionary conservation patterns, protein annotations and structural information to make predictions about the functional impact of genetic variants. Many tools exist including PolyPhen2 and SIFT [74, 75] for missense mutations and CADD for all variant types [76]. While these tools are useful, their results should be combined with other types of evidence as they are known to suffer from high false positive rates, particularly for variant subclasses such as pharmacogenetic variants [77, 78].

Until recently, a significant limitation in the annotation of protein features was the lack of comprehensive protein structures for all human proteins. AlphaFold successfully generated extremely accurate protein structures for all human proteins offering new opportunities for variant prioritisation [79]. The accuracy of functional impact tools can be increased with AlphaFold integration by considering the structural context of genetic variants [79]. AlphaMissense was subsequently developed based on AlphaFold2 predictions and fine-tuned using human and primate variant frequency databases [80]. Over time AlphaMissense (and subsequent developments) are likely to play a big role in improving functional impact metrics [81]. Overall there are many options for variant functional annotation (e.g. ENSEMBL Variant Effect Predictor (VEP), ANNOVAR, SnpEff) and functional inference prediction (Polyphen2, SIFT, CADD, AlphaMissense). Table 1 summarises several options including their relative strengths and weaknesses.

Population level databases of variant frequency are a powerful tool for understanding potential biological impact of genetic variation both globally and within matched ethnicities. Assigning variant allele frequencies enables the identification of rare or novel variants, a group enriched for pathogenic variants [82]. Variant databases have grown progressively larger over time, however historically most variants were of European origin [82]. Databases such as dbSNP and 1000 Genomes are two of the earliest databases and have proved invaluable for assigning variant frequencies [83, 84]. The Genome Aggregation Database, commonly known as gnomAD, is a more recent entry providing a comprehensive and publicly accessible repository that aggregates genomic data from a diverse range of populations [85]. It provides a wealth of information regarding the frequency and distribution of genetic variants across the human genome [86]. Critically, population databases have recognised the importance of incorporating non-European individuals, however many groups remain underrepresented [87].

Disease focused variant databases are another critical tool for variant prioritisation. While population level databases serve to filter out large numbers of common variants, disease variant databases help identify candidate pathogenic variants. These databases contribute significantly to the interpretation of genetic variants in a clinical context, aiding in the identification of variants associated with diseases thus informing clinical decision-making. Some of the larger databases include ClinVar [88], Human Gene Mutation Database (HGMD) [89] and Leiden Open Variation Database (LOVD) [90]. ClinVar is one of the largest clinical genomic databases, serving as a repository for variant data from clinical laboratories, clinicians, expert groups, patients, researchers, and other databases; it is a freely accessible, publicly curated database maintained by the National Centre for Biotechnology Information (NCBI) [91]. ClinVar ranks variants based on evidence such as functional assays providing a consistent scoring system across all potential clinically relevant variants. Similarly, the Human Gene Mutation Database (HGMD) aims to catalogue all mutations associated with inherited diseases [92, 93]. The mutation data in HGMD are sourced solely from scientific literature

Table 1 Variant annotation tools

Software (Class)	Strengths	Weaknesses
VEP (Functional annotation)	Open source; supports web, CLI, and API usage; annotates coding/non-coding variants via Ensembl & RefSeq models; integrates allele frequencies, pathogenicity scores & phenotype databases; customizable output & filtering options; regularly updated.	Complex output with multiple transcript annotations; requires filtering to simplify; non-coding annotations require additional configuration; plugin setup can be complex; slower performance on large datasets without caching.
ANNOVAR (Functional annotation)	Broad annotation support (e.g., RefSeq, gnomAD, CADD); flexible framework; efficient variant filtering; fast runtime.	Lower HGVS accuracy (93.3% concordance); limited support for complex/structural variants; requires manual database updates; not optimised for polygenic traits; collapses transcript isoforms.
SnpEff (Functional annotation)	Fast annotation for high-throughput pipelines; supports multiple genomes & transcript models; Coding annotation accuracy (~89.8%) comparable to VEP.	Lower concordance for indels & frameshifts (< 75%); protein syntax often mismatches references;
Polyphen2 (Functional inference prediction)	Predicts impact of missense mutations based on protein structure and evolutionary conservation.	High false-positive rate for certain variant classes, does not handle non-missense variants.
SIFT (Functional inference prediction)	Missense variants prediction based on sequence conservation.	Lower accuracy for less conserved regions of proteins.
CADD (Functional inference prediction)	Combines 60 + annotations into a single impact score; ranks deleteriousness across coding & non-coding variants; Incorporates both simulated & observed variants for robust training; Machine learning (SVM) framework improves generalizability and prioritization.	Lacks variant-type specificity (e.g. splicing vs. missense); less precise for rare or population-specific variants ; computationally demanding for non-pre-computed variants; may inflate scores for non-coding variants.
AlphaMissense (Functional inference prediction)	Combines structure & conservation; high concordance with REVEL/CADD; effective for prioritizing pathogenic missense variants.	Still emerging; may inflate pathogenicity scores in some domains; needs further validation; accuracy varies across genes/proteins classes.

Table 2 Gene model options

Gene Model	Strengths	Weaknesses
ENSEMBL	Reliable cross-species gene annotation via combined manual & automated methods; supports transcript diversity & comparative genomics; regularly updated with VEP & BioMart links; core genome browser support.	Annotation varies by species; complex transcripts could be inconsistently modelled; dependent on quality of assembly genome
GENCODE	High-quality gene annotation for human/mouse; includes lncRNAs, pseudogenes, & transcripts; integrates manual curation & automation; captures transcript diversity; used within Ensembl, RefSeq, & UCSC	Incomplete experimental support for all transcripts; redundancy & unclear function in many lncRNAs / pseudogenes; manual curation limits scalability; interversion differences may affect coordinate tracking.
RefSeq	High-quality, curated reference sequences for genomic, transcript, & protein data; consistent across species annotations; integrate manual curation with scalable automation; widely adopted in tools like VEP, ANNOVAR, GATK.	RefSeq tends to be conservative and includes fewer transcript isoforms; RefSeq updates less frequently than other models; RefSeq is centrally managed by NCBI and no community input
UCSC	Curated gene models from mRNA/protein alignments enhances RNA-seq quantification; emphasizes reliable transcripts & simplifies isoform sets for reproducible gene counts; integrated with UCSC Genome Browser.	Limited transcript diversity & isoforms & non-coding RNAs. Fewer splice junctions reduce RNA-seq accuracy; biased toward canonical genes
Uniprot	Provides detailed protein-level annotation including domains, function, and subcellular localization	Does not directly annotate variants or regulatory elements; protein-focused

and undergo rigorous manual curation using manual screening of journals and automated text mining [93]. Finally, LOVD is a freely available web-based platform for the collection, display, and curation of DNA variants in locus-specific databases (LSDBs) [90]. The design of LOVD system includes flexibility and seamless integration with other locus-specific LOVD instances, and introduced the "custom column" feature, enabling curators to tailor field setups according to their needs [94].

Annotation based strategies – gene level

A critical decision in any annotation workflow is the selection of the gene model. There are many efforts to standardise both gene sets and naming conventions however many challenges persist. Some of the popular models include ENSEMBL [95], GENCODE [96], RefSeq [97], UCSC [98] and UniProt [99]. Table 2 highlights the models' strengths and weaknesses.

In addition to variant-specific annotations, gene-level annotations are subsequently applied. In many cases a specific variant may not have been explicitly linked to disease pathogenesis, however the role the gene plays

in driving the disease is well characterised. Gene level annotations largely consist of disease databases and gene ontology (GO) term annotation and enrichment analysis. Databases are typically curated repositories of gene-disease associations and help identify whether gene dysfunction has previously been implicated in similar diseases. There are many such databases, the largest being the Online Mendelian Inheritance in Man (OMIM) database [100].

If, however, a gene has not been directly implicated in causing the disease, gene ontology (GO) is able to identify the function for each gene of interest potentially linking GO terms to the observed disease phenotype. GO annotation is standardised through large international efforts such the Gene Ontology Consortium (http://www.geneontology.org) which enable quick and easy GO annotation of structured domain-specific ontologies [101]. A common application using GO terms is enrichment analysis, which aims to identify over-represented biological process, molecular function or cellular component shared by genes implicated in driving a polygenic disease.

Annotation based strategies – gene network level

Beyond gene-level annotations, the gene's role within larger biological networks can be examined. To do this there are a variety of gene network analysis tools designed to analyse and interpret biological data, particularly gene expression data, in the context of biological networks. These tools aim to uncover relationships and interactions between genes to gain insights into the underlying biological processes. They contribute significantly to the understanding of the complex relationships within biological systems, helping researchers unravel the functional implications of gene interactions. Several popular tools in this space are STRING and Ingenuity Pathway Analysis (IPA). STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is an online platform created to identify protein-protein interactions (PPIs) and functional associations [102]. The STRING database, available at https://string-db.org/, systematically compiles and integrates protein–protein interactions, encompassing both physical interactions and functional associations. The database is populated from various sources, including automated text mining of scientific literature, computational predictions based on co-expression and conserved genomic context, databases containing interaction experiments, and curated sources describing known complexes and pathways [103]. IPA is a proprietary tool developed by QIAGEN with similar functionality that is employed for applications including biomarker discovery, metabolomics, microRNA research, next-generation sequencing data analysis, proteomics, toxicogenomics, and transcriptomics [104]. While most current gene network tools incorporate well-characterised large, curated datasets new tools that construct custom networks from genome/transcriptome patient data for a single patient offer potential for custom treatments [105].

Comprehensive annotation tools

While many pipelines overlap annotation datasets consecutively in series, aggregate tools are becoming increasing popular to manage the increasing number of disparate annotation resources. With continuously updated databases, it is increasingly important to apply consistent, up to date annotations [106]. ENSEMBL's Variant Effect Predictor (VEP) is a prominent aggregate tool in the landscape of functional annotation. VEP provides comprehensive annotations for genetic variants, including their functional consequences, conservation scores, and potential associations with known diseases [107]. The tool is known for its user-friendly interface and frequent updates, ensuring that researchers have access to the latest genomic information by linking results to ENSEMBL's latest gene model. VEP's ability to handle diverse types of genomic variants and its integration with various databases make it a valuable resource in annotation improvement efforts [108]. Additional tools like ANNOVAR perform a similar role [109].

Complete end-to-end workflows

Beyond aggregate annotation tools, there are an increasing number of complete end-to-end variant prioritisation workflows. These tools typically integrate functionalities including variant annotation, filtering, and interpretation aiming to identify potentially pathogenic variants directly from input variant lists with little to no manual interpretation required. Example include VarSeq, a commercial variant analysis software tool that is designed to streamline the entire workflow, from variant discovery to interpretation for gene panels, exomes or genomes [110]. Non-commercial options include WANNOVAR, a web-based tool designed for annotation and functional prediction of genetic variants and VariantDB which is designed for the annotation, prioritisation and analysis of genetic variants [111]. Seqr is an increasingly popular option developed by the Broad Institute [112].

While this next generation of tools are promising, challenges with installation and lack of configurability hamper their widespread uptake with researchers and clinicians often favouring to combine multiple tools to perform concurrent steps in bespoke workflows [38].

Machine learning applications

Addressing the significant challenges associated with pathogenic variant identification requires a multi-faceted approach. One potential solution being considered is the development of advanced machine learning (ML) algorithms. By leveraging ML, algorithms can potentially

improve diagnosis rates by identifying underlying complex relationships within biological systems [113]. ML algorithms are gaining traction in life sciences due to the capacity to deal efficiently with complex genomic data patterns [114]. Early works suggest that ML algorithms have the potential to learn from, and act upon complex heterogeneous datasets by identifying new biological patterns that increase diagnosis accuracy [115–118]. It is likely that ML algorithms will play an increasingly important role in detecting pathogenic variants. Some recently published ML algorithms for precision medicine are listed in Table 3 .

Variant pathogenicity

One of the most common current applications of ML is predicting variant pathogenicity, with numerous computational tools available [119–129]. ML models are typically trained on both known pathogenic and benign variants and deployed to predict the functional consequences of genetic variants on both protein structure and function [130]. Depending on the composition of training data, these approaches can be divided into genome-wide, disease-specific, or even gene-specific categories [131]. Popular tools offering genome-wide data predictions include Rare Exome Variant Ensemble Learner (REVEL) [124], BayesDel [132], ClinPred [133] and AlphaMissense [52].

Variant prioritisation

Other common ML applications include variant detection, prioritisation and feature discovery. DeepVariant is an increasingly popular variant calling workflow compatible with Illumina, PacBio HiFi, and Oxford Nanopore sequence data [134]. DeepTrio is built upon DeepVariant and uses neural networks to identify variants specifically in pedigree of two or three members. M-CAP is a prioritisation tool that eliminates uncertain variants and reports 95% sensitivity levels [130]. MLVar is a another variant prioritisation workflow that follows ACMG guidelines and uses variant annotation features to predict probabilistic pathogenicity scores [135]. MAVERICK uses a neural network approach to predict pathogenic variants for Mendelian monogenic diseases [136]. Identifying cryptic biological features is another important ML application, for example the prediction and recognition of transcription start sites (TSSs) [137], splice sites [138], promoters [139], enhancers [140], and nucleosome position [141]. Larger end-to-end workflows employing ML are an active area of development.

Genome-wide association studies (GWAS) are another active area of ML algorithm development. Tools such as genomic best linear unbiased prediction (gBLUP) [142], support vector machine (SVM) [143], xGBoost [144], and random forest (RF) [145] are widely used to identify relevant traits in GWAS, with the large, well annotated GWAS data suitable for training purposes. Similarly, large datasets divided into disease and controls serve as suitable training data for ML algorithms that are able to predict traits and identify enriched variants [146]. Many tools however fail to generalise to specific diseases. To address this, studies often run multiple ML methods using a consensus-based approach.

Large Language model

Large language models (LLM) show promise in a variety of precision medicine applications such as reducing literature search time for variant classification and interpretation. Microsoft developed the generative AI tool the EvAgg, which reports improvements in the sensitivity and specificity of pathogenic variant identification [147]. EvAgg reduced the manual curation time by 34% and increased the number of papers, variants, and cases evaluated per unit time. Such works have led to papers concluding that variant classification will be standardized in the near future, however achieving this requires overcoming significant challenges [148]. A recent benchmarking study considered four LLMs across ten fictional oncology patients and encouragingly found that LLMs were able to identify several important treatment strategies and provide some reasonable suggestions not easily found by experts [149]. However, they concluded that LLMs are not yet applicable for routine clinical analysis

Table 3 ML applications in precision medicine

Software	Function	ML approach	Website
AlphaMissense	Functional inference	Deep learning	https://alphamissense.hegelab.org/
DeepVariant	Variant detection	CNN	https://github.com/google/deepvariant
M-CAP	Variant prioritisation	Supervised ML	http://bejerano.stanford.edu/MCAP/
MLVar	Variant prioritisation	Method/Pipeline using ML	https://github.com/GiovannaNicora/MLVar
REVEL	Functional inference	Random Forest (RF)	https://sites.google.com/site/revelgenomics/
BayesDel/ PEARCH	Functional inference	Likelihood Based approach	https://fenglab.chpc.utah.edu/BayesDel.html
ClinPred	Functional inference	Random forest and Gradient boosting models	https://sites.google.com/site/clinpred/
MAVERICK	Functional inference	Neural network-based	https://github.com/ZuchnerLab/Maverick
EvAgg	Clinical curation	LLM model	https://github.com/microsoft/healthfutures-evagg

as an aid for clinical decision-making in oncology. Analysing complex germline disease represents an even bigger challenge.

While ML holds great promise for precision medicine, it is not without challenges due to the complexity and uniqueness of an individual's genome highlighting the need for accurate, robust and interpretable models. Challenges include issues with the accuracy of existing variant classifications as well as the rising number of variants of uncertain significance (VUS) [150]. Building robust ML models requires large, high-quality data that has been extensively benchmarked using both simulated and established reference data sets [151]. Generating such data sets is compounded by the inherent complexity of the human genome, with numerous non-genetic factors contributing to complex disease. While ML algorithms are increasingly able to process complex genomic information to identify novel patterns and associations relevant to variant interpretation, patient datasets are increasingly heterogeneous in terms of data type and source [152]. Increasingly, data sets contain a number of data modalities (e.g. transcriptomics, epigenomics, and clinical data) requiring updates and changes to existing ML models trained exclusively on other types of data [153]. ML algorithms face additional challenges with their need for large diverse and homogenous training datasets without potential biases, a known challenge with many complex molecular datasets [154]. Overall, the most significant barriers to widespread clinical adoption of ML approaches are model interpretability, model validation and data harmonisation.

Model interpretability remains one of the most significant barriers to clinical uptake as any critical treatment decisions require a clear understanding of how models arrive at recommendations. Deep learning models, while highly effective at pattern recognition in genomic data and medical imaging, often function as "black boxes" where the decision-making process remains opaque. This is particularly problematic in precision medicine, where treatment decisions involve weighing complex risk-benefit profiles for individual patients. For example, a neural network might accurately predict cancer treatment response, but if clinicians cannot understand the biological rationale, they will be reluctant to act on it. New tools such as StratoMod [155] are using interpretable ML to predict variant calling and sequencing errors however without additional orthogonal validation data, clinical uptake remains unlikely. To address 'black box' challenges in security sensitive applications new tools like EnEXP are using interpretable ensemble tree approach to achieve a global interpretation of the entire dataset through the aggregation of individual sample insights [156]. Recent advances in explainable AI, including LIME and SHAP, offer promising approaches, but these post-hoc explanations may not accurately reflect the model's true decision-making process [157].

Data validation is limiting clinical uptake as it presents unique challenges beyond traditional evaluation metrics. While models might demonstrate excellent performance on test sets, real-world clinical translation requires additional considerations. Temporal validation represents a particular challenge as medical practice and treatment protocols evolve continuously [158], for example models trained on historical data may not perform reliably on current patients in rapidly advancing fields like oncology where new biomarkers are regularly discovered. External validation across different healthcare systems and patient populations is also essential but difficult to achieve in practice. Models developed at academic centres may not generalize to community hospitals with different demographics. Overall, the validation process must account for clinical decision-making's dynamic nature, where model predictions influence subsequent patient management, creating complex feedback loops difficult to capture in traditional frameworks.

Data harmonization is another significant barrier to clinical uptake. Genomic data harmonization involves reconciling different sequencing platforms, analytical pipelines, and annotation standards which can introduce systematic biases if not properly addressed [159]. Clinical data harmonization faces complexities from varying electronic health record systems, coding standards, and documentation practices. Laboratory records may use different units, clinical observations different terminologies, and treatment protocols may vary significantly across institutions. Temporal alignment of multi-modal data presents another significant challenge as genomic data is collected at specific time points while clinical data accumulates continuously, making it difficult to create coherent longitudinal patient profiles for ML training. Patient data is often siloed within healthcare systems due to privacy regulations and competitive concerns. Federated learning approaches offer potential solutions but introduce additional technical complexities [160]. Applying consistent guidelines and principles for data structure harmonization are critical; for example Findable, Accessible, Interoperable, and Reusable (FAIR) principles for data sharing [161] are gaining popularity. There have also been some advances in scalable approaches [162] and guidelines on using ML ethically [163]. However, these need more empirical validation before implementing within health care systems.

Core principles for best practices

The state of art for best practices in pathogenic variant detection is a rapidly moving target, however core design principles are key in creating a robust and flexible framework able to integrate new modalities as they gain

traction in precision medicine. Here we describe several core principles needed to develop a system appropriate for both the current and future needs in precision medicine.

For general design considerations, it is critical to design a workflow that is modular, scalable, parallelisable, secure, reproducible and flexible. A modular design that can incorporate new data types and algorithms as needed is a key requirement when working in this rapidly evolving space. Design flexibility is also key to ensure longevity. For example the ability to run multiple tools and employ a consensus based approach is increasingly being recognised in a variety of applications including variant detection and RNA-Seq data analysis [164]. It is also preferable to incorporate well-tested standardised tools when available to avoid introducing unintentional errors arising from less well tested internally developed software. Another key consideration is the ability to parallelise and scale analysis components, important for reducing turnaround time for individual patients and for handling an increasing volume of samples. Consistency regarding input format requirements and outputs is also critical in ensuring backwards compatibility and the ability for re-analysis.

More specific considerations for precision medicine are around security, interactivity and the ability to generate concise clinical reports. Handling patient data securely is critical for many reasons. Patient data that includes genomic information, medical records and family histories requires the utmost care and sensitivity to meet patient expectations. Misuse of such data may lead to discrimination and potential legal ramifications. De-identification is a common approach however it must be done properly to not allow re-identification via cross referencing of metadata or other public datasets. Ideally de-identification capability needs to be managed via an additional layer of access control. Security needs to be balanced with the development of interactive systems for clinicians who interrogate the data to identify pathogenic variants. Such web-based tools are critical however they need to ensure data is secure and protected throughout. The interface needs to support variant filtering and display variant summaries with the information needed to assess pathogenicity. The interface should support free text entry where clinicians record their determination and describe the evidence justifying the classification. A final consideration is the ability to develop robust clinical reports. Developing user friendly reports requires many iterations with clinicians to determine both the filters to employ and the level of detail to include for each candidate driver. The report design requires flexibility to incorporate addition, often disease specific, information as needed.

Arguably the most critical component in precision medicine system is reproducibility. There are many practical considerations needed to ensure complete reproducibility, which can be achieved through a combination of code versioning, log file generation, robust testing suites and employing a software pipeline manager. Versioning all code and config files during development coupled with thorough logging of all commands lays the foundation for reproducible workflows. Extensive documentation of code and protocols is also critical particularly when multiple team members are involved. The development of a robust testing suite with gold standard datasets will help ensure any changes will not generate unintended downstream consequences. Finally, it is recommended to employ a modern pipeline manager tool such as NextFlow or Snakemake to facilitate running the workflow in a variety of hardware infrastructures including local infrastructure, HPC or cloud. Collectively following these design principles will drastically increase the reproducibility and longevity of the workflow.

Discussion

While precision medicine has increased diagnosis rates around the world, current challenges exist including handling sensitive patient data and the lack of disease specific annotation, data standards and ethnically matched variant annotations. Future challenges include the need of integration of multi-omic data and the incorporation of new NGS data types (Fig. 2).

Current challenges

The systematic annotation of genetic variants within precision medicine programs has led to the discovery of large numbers of pathogenic variants however it has become clear no single strategy is universally effective for all diseases. Success within a disease class depends on multiple factors including sequence technology, patient selection and annotation strategy. For example, pedigree sequencing has been critical in identifying disease-causing *de novo* mutations driving neurodevelopmental disorders [165] and autism spectrum disorders [166] and in tracing inheritance patterns across multiple generations in condition like Huntington's disease [167]. In autoimmune disease research, B-cell or T-cell receptor repertoire sequencing is increasingly utilised to identify pathogenic clonal lineages [168]. Additionally, certain mutation types are strongly associated with a specific disease class such as copy number variations (CNVs) in neurological and autoimmune disorders [19, 169] and splicing variants in Duchenne muscular dystrophy [170]. Careful patient selection based on genomic profiling has been instrumental in matching patients with targeted therapies as demonstrated by studies such as the NCI-MATCH (Molecular Analysis for Therapy Choice) trial

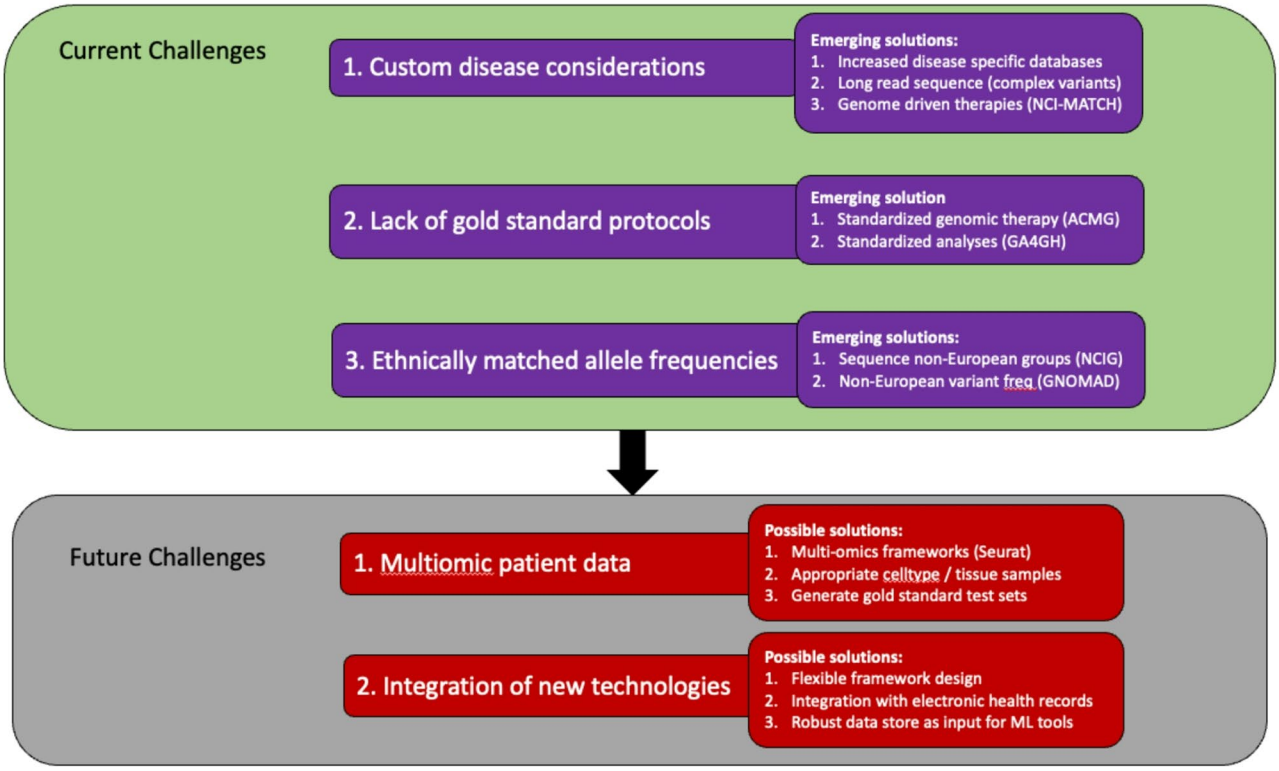

Fig. 2 Current and future challenges

[171]. Collectively these findings underscore the need for additional disease-specific considerations to improve genetic diagnosis yields.

A major obstacle to incorporating genomic data into clinical practice is the lack of standard procedures for both analysing NGS data and summarising the relevant information into clinical reports. While progress has been made, consistent and reproducible methods remains essential to ensure the reliability of genomic results [172]. While it is widely acknowledged that ACMG standards and guidelines serve as the de facto gold standard for genetic therapy, their recommended datasets and workflows are often distributed across different online platforms and databases [173]. Further, laboratories often employ different tools and cutoffs, resulting in discrepancies in variant classification with such inconsistencies impacting patient care [174].

Harmonizing interpretation guidelines is crucial for providing clinicians with reliable genomic information to guide personalized medical interventions [175]. However, any standardization must remain flexible given the fast pace of technological advancement and evolving data types [176]. Establishing global collaborations and encouraging data sharing initiatives can contribute to the development of comprehensive, standardized guidelines such as the Global Alliance for Genomics and Health (GA4GH) [177]. Standardized guidelines enhance consistency across laboratories, improve accuracy in variant classification, and ultimately contribute to the reliability of genomic data.

The global genetic landscape is shaped by ethnic diversity and influenced by historical, geographical, and demographic factors [178]. To account for this, it is important to establish population-specific databases to capture the unique genetic variations across diverse ethnic groups. GNOMAD, for instance, has been instrumental in cataloguing genetic variations across varied populations, offering a valuable resource for ethnicity level variant frequency estimations [179]. This information can influence not only disease susceptibility but treatment response; for example, certain pharmacogenetic variants affect drug metabolism differently in various populations [180]. However establishing population-specific databases comes with challenges, including ethical considerations, data privacy, and ensuring adequate representation [181]. Initiatives such as 'All of Us Research Program' aims to address these issues by building inclusive, large-scale dataset that reflects the global genetic diversity. Such efforts are crucial in unravelling the complexities of genetic variants across distinct ethnic groups.

Future challenges

The rise of affordable sequencing technology has led to a growing reliance on data generated across various

biological levels [182]. For example, the microbiome is increasingly being linked to human health outcomes with composition shifts observed during the onset of many diseases such as type II diabetes [183–186]. Integrating metagenomic and other multi-omic patient data with clinical information has the potential to improve prognostics and predictive accuracy of disease phenotypes ultimately leading to better treatment and prevention strategies [187, 188]. However, the analysis of these complex datasets remains a challenge due to the inherent heterogeneity in individual omics datasets and the computational resources required for analysis and integration [182]. The rapid pace of change within the multi-omics space means benchmarking studies are essential to ensure appropriate tools are chosen to address specific biological questions [189–191]. Future developments should prioritise reducing complexity, enhancing interoperability, and creating user-friendly frameworks to consolidate multi-omics data.

The effective application of precision medicine depends on precise, evidence-driven interpretation of genetic data, ensuring proper clinical management and care, while avoiding flawed conclusions that could lead to harm [192]. Genomic data is inherently dynamic and influenced by fast moving technological advancements meaning variants that were once classified as benign may need re-evaluation as new evidence emerges [193]. Advanced informatics and ML algorithms are poised to enable real-time data integration of diverse datasets, identifying clinically-relevant patterns that contribute to the continuous refinement of variant interpretation [194]. Such systems will empower clinicians with the most up-to-date variant interpretations, enhancing the precision and effectiveness of personalized healthcare. Central to this vision is working with Electronic Health Records (EHRs), which serve as comprehensive repositories of patient-specific data, encompassing medical histories, treatment responses, and other relevant information. Incorporating EHR data into the genetic variant prioritisation process enables a holistic view of the patient's health journey [195]. Embracing a patient-centric framework will allow healthcare providers to tailor genetic interpretations that align with individual needs, ultimately realising the potential of personalised patient care.

Conclusion

Prioritising disease-causing genetic variants is fundamental for progressing personalized medicine, improving clinical diagnostics, and understanding genetic contributions to diseases. The adoption of precision medicine programs underscores the importance of prioritizing genetic variants, tailoring patient care based on genetic makeup and individual characteristics. While the affordability of quality sequence data has improved, substantiating the link between genes and diseases remains resource intensive. Accurate identification of disease-causing variants enhances diagnostic precision, aiding in early detection and targeted interventions for genetic disorders. Addressing current challenges today will ensure better precision medicine in the future.

Author contributions

M.A.F and A.C conceived the review. S.D and M.A.F wrote the main text with contributions from M.K, U.S. and A.C. All authors reviewed the manuscript.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 7 July 2025 / Accepted: 6 August 2025
Published online: 22 August 2025

References

1. Yang Y, Lyu J, Wang R, Wen Q, Zhao L, Chen W, et al. A digital mask to safeguard patient privacy. Nat Med. 2022;28(9):1883–92.
2. Ginsburg GS, Phillips KA. Precision medicine: from science to value. Health Aff (Millwood). 2018;37(5):694–701.
3. Petersen B-S, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 2017;18(1):14.
4. Magger O, Waldman YY, Ruppin E, Sharan R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. PLoS Comput Biol. 2012;8(9):e1002690.
5. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies. PLoS ONE. 2015;10(11):e0143199.
6. Berger SM, Appelbaum PS, Siegel K, Wynn J, Saami AM, Brokamp E, et al. Challenges of variant reinterpretation: opinions of stakeholders and need for guidelines. Genet Med. 2022;24(9):1878–87.
7. Field MA, Burgio G, Chuah A, Al Shekaili J, Hassan B, Al Sukaiti N, et al. Recurrent miscalling of missense variation from short-read genome sequence data. BMC Genomics. 2019;20(Suppl 8):546.
8. Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep. 2019;9(1):9345.
9. Murphy DA, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. Elife. 2023;12.
10. Virolainen SJ, VonHandorf A, Viel K, Weirauch MT, Kottyan LC. Gene-environment interactions and their impact on human health. Genes Immun. 2023;24(1):1–11.
11. Jeffreys AJ, Wilson V, Thein SL, Weatherall DJ, Ponder BA. DNA fingerprints and segregation analysis of multiple markers in human pedigrees. Am J Hum Genet. 1986;39(1):11–24.
12. Field MA, Cho V, Cook MC, Enders A, Vinuesa C, Whittle B et al. Reducing the search space for causal genetic variants with VASP: variant analysis of sequenced pedigrees. Bioinformatics. 2015.
13. Aureliano W. Difficult decisions and possible choices: rare diseases, genetic inheritance and reproduction of the family. Soc Sci Med. 2024;363:117380.
14. Strianese O, Rizzo F, Ciccarelli M, Galasso G, D'Agostino Y, Salvati A et al. Precision and Personalized Medicine: How Genomic Approach Improves the Management of Cardiovascular and Neurodegenerative Disease. Genes (Basel). 2020;11(7).
15. Rehm HL, Fowler DM. Keeping up with the genomes: scaling genomic variant interpretation. Genome Med. 2019;12(1):5.

16. Pei XM, Yeung MHY, Wong ANN, Tsang HF, Yu ACS, Yim AKY et al. Targeted sequencing approach and its clinical applications for the molecular diagnosis of human diseases. Cells. 2023;12(3).
17. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med. 2018;20(1):159–63.
18. Young C, Singh M, Jackson KJL, Field MA, Peters TJ, Angioletti-Uberti S et al. A triad of somatic mutagenesis converges in self-reactive B cells to cause a virus-induced autoimmune disease. Immunity. 2025.
19. Singh M, Louie RHY, Samir J, Field MA, Milthorpe C, Adikari T, et al. Expanded T cell clones with lymphoma driver somatic mutations accumulate in refractory Celiac disease. Sci Transl Med. 2025;17(798):eadp6812.
20. McCabe MJ, Gauthier M-EA, Chan C-L, Thompson TJ, De Sousa SMC, Puttick C, et al. Development and validation of a targeted gene sequencing panel for application to disparate cancers. Sci Rep. 2019;9(1):17052.
21. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013;14(4):295–300.
22. Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ. 2016;4:e2074.
23. Yu H, Yu H, Zhang R, Peng D, Yan D, Gu Y et al. Targeted gene panel provides advantages over whole-exome sequencing for diagnosing obesity and diabetes mellitus. J Mol Cell Biol. 2023.
24. Johar AS, Mastronardi C, Rojas-Villarraga A, Patel HR, Chuah A, Peng K, et al. Novel and rare functional genomic variants in multiple autoimmune syndrome and sjogren's syndrome. J Transl Med. 2015;13:173.
25. Dunkerton S, Field M, Cho V, Bertram E, Whittle B, Groves A et al. A de Novo mutation in KMT2A (MLL) in monozygotic twins with Wiedemann-Steiner syndrome. Am J Med Genet A. 2015.
26. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745–55.
27. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl):228–37.
28. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011;48(9):580–9.
29. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010;42(1):30–5.
30. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59(1):5–15.
31. Nakagawa H, Wardell CP, Furuta M, Taniguchi H, Fujimoto A. Cancer whole-genome sequencing: present and future. Oncogene. 2015;34(49):5943–50.
32. Scocchia A, Wigby KM, Masser-Frye D, Del Campo M, Galarreta CI, Thorpe E, et al. Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. NPJ Genom Med. 2019;4:5.
33. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci USA. 2015;112(17):5473–8.
34. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of exome and genome sequencing technologies for the complete capture of Protein-Coding regions. Hum Mutat. 2015;36(8):815–22.
35. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med. 2018;20(4):435–43.
36. Bertoli-Avella AM, Beetz C, Ameziane N, Rocha ME, Guatibonza P, Pereira C, et al. Successful application of genome sequencing in a diagnostic setting: 1007 index cases from a clinically heterogeneous cohort. Eur J Hum Genet. 2021;29(1):141–53.
37. Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, et al. Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. NPJ Genom Med. 2016;1:15012.
38. Willig LK, Petrikin JE, Smith LD, Saunders CJ, Thiffault I, Miller NA, et al. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. Lancet Respir Med. 2015;3(5):377–87.
39. Ostrander BEP, Butterfield RJ, Pedersen BS, Farrell AJ, Layer RM, Ward A, et al. Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. Npj Genomic Med. 2018;3(1):22.
40. Rajagopalan R, Gilbert MA, McEldrew DA, Nassur JA, Loomes KM, Piccoli DA, et al. Genome sequencing increases diagnostic yield in clinically diagnosed Alagille syndrome patients with previously negative test results. Genet Sci. 2021;23(2):323–30.
41. Austin-Tse CA, Jobanputra V, Perry DL, Bick D, Taft RJ, Venner E, et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. Npj Genomic Med. 2022;7(1):27.
42. Hamzeh AR, Andrews TD, Field MA. Detecting causal variants in Mendelian disorders using Whole-Genome sequencing. Methods Mol Biol. 2021;2243:1–25.
43. Hegenbarth J-C, Lezzoche G, De Windt LJ, Stoll M. Perspectives on Bulk-Tissue RNA sequencing and Single-Cell RNA sequencing for cardiac transcriptomics. Front Mol Med. 2022;2.
44. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386).
45. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343(6167):193–6.
46. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–9.
47. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20(9):1131–9.
48. Goldman SL, MacKay M, Afshinnekoo E, Melnick AM, Wu S, Mason CE. The impact of heterogeneity on Single-Cell sequencing. Front Genet. 2019;10:8.
49. Brittain HK, Scott R, Thomas E. The rise of the genome and personalised medicine. Clin Med (Lond). 2017;17(6):545–51.
50. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175–88.
51. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S et al. Next-Generation sequencing technology: current trends and advancements. Biology (Basel). 2023;12(7).
52. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019;35(13):2193–8.
53. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.
54. Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. Third-Generation sequencing: the spearhead towards the radical transformation of modern genomics. Life (Basel). 2021;12(1).
55. Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021;31(3):448–60.
56. Johar AS, Anaya JM, Andrews D, Patel HR, Field M, Goodnow C et al. Candidate gene discovery in autoimmunity by using extreme phenotypes, next generation sequencing and whole exome capture. Autoimmun Rev. 2014.
57. Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):D1018–27.
58. Patino CM, Ferreira JC. Inclusion and exclusion criteria in research studies: definitions and why they matter. J Bras Pneumol. 2018;44(2):84.
59. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.
60. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
61. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64–9.
62. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet. 2015;97(2):199–215.
63. Al Sukaiti N, AbdelRahman K, AlShekaili J, Al Oraimi S, Al Sinani A, Al Rahbi N, et al. Agammaglobulinaemia despite terminal B-cell differentiation in a patient with a novel LRBA mutation. Clin Translational Immunol. 2017;6(5):e144.
64. Jarvik GPBB. Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants. 2016.
65. Richards SAN, Bale S et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American

college of medical genetics and genomics and the association for molecular pathology. 2015.
66. Biesecker LG, Byrne AB, Harrison SM, Pesaran T, Schaffer AA, Shirts BH, et al. ClinGen guidance for use of the PP1/BS4 co-segregation and PP4 phenotype specificity criteria for sequence variant pathogenicity classification. Am J Hum Genet. 2024;111(1):24–38.
67. MacArthur DGMT, Dimmock DP et al. Guidelines for investigating causality of sequence variants in human disease. 2014.
68. Biesecker LGGR. Diagnostic clinical genome and exome sequencing. 2014.
69. Ng KPB, Kiat Puar TH et al. COVID-19 and the Risk to Health Care Workers: A Case Report. 2020.
70. Cooper CSN, Livingston G et al. Ethnic inequalities in the use of health services for common mental disorders in England. 2013.
71. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
72. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.
73. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
74. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
75. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
76. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
77. McConnell H, Andrews TD, Field MA. Efficacy of computational predictions of the functional effect of idiosyncratic Pharmacogenetic variants. PeerJ. 2021;9:e11774.
78. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, et al. Comparison of predicted and actual consequences of missense mutations. Proceedings of the National Academy of Sciences of the United States of America; 2015.
79. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–10.
80. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583–9.
81. Schmidt A, Röner S, Mai K, Klinkhammer H, Kircher M, Ludwig KU. Predicting the pathogenicity of missense variants using features derived from AlphaFold2. Bioinformatics. 2023;39(5).
82. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: lessons from GnomAD. Hum Mutat. 2022;43(8):1012–30.
83. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
84. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. DbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
85. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv. 2019:531210.
86. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
87. Calcino A, Cooke I, Cowman P, Higgie M, Massault C, Schmitz U, et al. Harnessing genomic technologies for one health solutions in the tropics. Global Health. 2024;20(1):78.
88. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–5.
89. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human gene mutation database (HGMD): 2003 update. Hum Mutat. 2003;21(6):577–81.
90. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011;32(5):557–63.
91. Shabani M, Dyke SOM, Marelli L, Borry P. Variant data sharing by clinical laboratories through public databases: consent, privacy and further contact for research policies. Genet Sci. 2019;21(5):1031–7.
92. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
93. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136(6):665–77.
94. Fokkema IFAC, Kroon M, López Hernández JA, Asscheman D, Lugtenburg I, Hoogenboom J, et al. The LOVD3 platform: efficient genome-wide sharing of genetic variants. Eur J Hum Genet. 2021;29(12):1796–803.
95. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S et al. The Ensembl gene annotation system. Database (Oxford). 2016;2016.
96. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
97. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33(Database issue):D501–4.
98. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics. 2006;22(9):1036–46.
99. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699.
100. Hamosh A, Amberger JS, Bocchini C, Scott AF, Rasmussen SA. Online Mendelian inheritance in man (OMIM®): victor mckusick's magnum opus. Am J Med Genet A. 2021;185(11):3259–65.
101. Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, et al. Gene ontology annotations and resources. Nucleic Acids Res. 2013;41(Database issue):D530–5.
102. Zhuang Y, Xing F, Ghosh D, Banaei-Kashani F, Bowler RP, Kechris K. An augmented High-Dimensional graphical Lasso method to incorporate prior biological knowledge for global network learning. Front Genet. 2022;12.
103. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–46.
104. Kramer A, Green J, Pollard J Jr., Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics. 2014;30(4):523–30.
105. Gillman R, Field MA, Schmitz U, Karamatic R, Hebbard L. Identifying cancer driver genes in individual tumours. Comput Struct Biotechnol J. 2023;21:5028–38.
106. American Academy of Microbiology Colloquia Reports. In: Washington DC, editor. An experimental approach to genome annotation: this report is based on a colloquium sponsored by the American academy of microbiology held July 19–20, 2004. Washington (DC): American Society for Microbiology Copyright 2004 American Academy of Microbiology; 2004.
107. Tuteja S, Kadri S, Yap KL. A performance evaluation study: variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing. J Pathol Inf. 2022;13:100130.
108. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
109. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164–e.
110. Pranckėnienė L, Jakaitienė A, Ambrozaitytė L, Kavaliauskienė I, Kučinskas V. Insights into de Novo mutation variation in Lithuanian exome. Front Genet. 2018;9:315.
111. Vandeweyer G, Van Laer L, Loeys B, Van den Bulcke T, Kooy RF. VariantDB: a flexible annotation and filtering portal for next generation sequencing data. Genome Med. 2014;6(10):74.
112. Pais LS, Snow H, Weisburd B, Zhang S, Baxter SM, DiTroia S, et al. Seqr: A web-based analysis and collaboration tool for rare disease genomics. Hum Mutat. 2022;43(6):698–707.
113. Wang R, Jiang Y, Jin J, Yin C, Yu H, Wang F, et al. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Res. 2023;51(7):3017–29.
114. Routhier E, Mozziconacci J. Genomics enters the deep learning era. PeerJ. 2022;10:e13613.
115. Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. The use of machine learning in rare diseases: a scoping review. Orphanet J Rare Dis. 2020;15:1–10.

116. Setty ST, Scott-Boyer M-P, Cuppens T, Droit A. New developments and possibilities in reanalysis and reinterpretation of whole exome sequencing datasets for unsolved rare diseases using machine learning approaches. Int J Mol Sci. 2022;23(12):6792.
117. Cohen AS, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, et al. Genomic answers for children: dynamic analyses of > 1000 pediatric rare disease genomes. Genet Sci. 2022;24(6):1336–48.
118. Okazaki A, Ott J. Machine learning approaches to explore digenic inheritance. Trends Genet. 2022;38(10):1013–8.
119. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266(5182):66–71.
120. Aljarf R, Shen M, Pires DEV, Ascher DB. Understanding and predicting the functional consequences of missense mutations in BRCA1 and BRCA2. Sci Rep. 2022;12(1):10458.
121. Crockett DK, Piccolo SR, Ridge PG, Margraf RL, Lyon E, Williams MS, et al. Predicting phenotypic severity of uncertain gene variants in the RET proto-oncogene. PLoS ONE. 2011;6(3):e18380.
122. Evans P, Wu C, Lindy A, McKnight DA, Lebo M, Sarmady M, et al. Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets. Genome Res. 2019;29(7):1144–51.
123. Hart SN, Polley EC, Shimelis H, Yadav S, Couch FJ. Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. NPJ Breast Cancer. 2020;6:13.
124. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85.
125. Lai C, Zimmer AD, O'Connor R, Kim S, Chan R, van den Akker J, et al. LEAP: using machine learning to support variant classification in a clinical setting. Hum Mutat. 2020;41(6):1079–90.
126. Crockett DK, Lyon E, Williams MS, Narus SP, Facelli JC, Mitchell JA. Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. J Am Med Inf Assoc. 2012;19(2):207–11.
127. Karalidou V, Kalfakakou D, Papathanasiou A, Fostira F, Matsopoulos GK. MARGINAL: an automatic classification of variants in BRCA1 and BRCA2 genes using a machine learning model. Biomolecules. 2022;12(11).
128. Khandakji MN, Mifsud B. Gene-specific machine learning model to predict the pathogenicity of BRCA2 variants. Front Genet. 2022;13:982930.
129. Padilla N, Moles-Fernández A, Riera C, Montalban G, Özkan S, Ootes L, et al. BRCA1- and BRCA2-specific in Silico tools for variant interpretation in the CAGI 5 ENIGMA challenge. Hum Mutat. 2019;40(9):1593–611.
130. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–6.
131. Kang M, Kim S, Lee D-B, Hong C, Hwang K-B. Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants. Sci Rep. 2023;13(1):10478.
132. Feng BJ. PERCH: A unified framework for disease gene prioritization. Hum Mutat. 2017;38(3):243–51.
133. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify Disease-Relevant nonsynonymous Single-Nucleotide variants. Am J Hum Genet. 2018;103(4):474–83.
134. Poplin R, Chang P-C, 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.
135. Nicora G, Zucca S, Limongelli I, Bellazzi R, Magni P. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. Sci Rep. 2022;12(1):2517.
136. Danzi MC, Dohrn MF, Fazal S, Beijer D, Rebelo AP, Cintra V, et al. Deep structured learning for variant prioritization in Mendelian diseases. Nat Commun. 2023;14(1):4167.
137. Ohler U, Liao G-c, Niemann H, Rubin GM. Computational analysis of core promoters in the drosophila genome. Genome Biol. 2002;3(12):research00871.
138. Degroeve S, De Baets B, Van de Peer Y, Rouzé P. Feature subset selection for splice site prediction. Bioinformatics. 2002;18(Suppl 2):S75–83.
139. Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol. 1990;212(4):563–78.
140. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311–8.
141. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature. 2006;442(7104):772–8.
142. Akbarzadeh M, Dehkordi SR, Roudbar MA, Sargolzaei M, Guity K, Sedaghati-khayat B, et al. GWAS findings improved genomic prediction accuracy of lipid profile traits: Tehran cardiometabolic genetic study. Sci Rep. 2021;11(1):5780.
143. Mittag F, Büchel F, Saad M, Jahn A, Schulte C, Bochdanovits Z, et al. Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. Hum Mutat. 2012;33(12):1708–18.
144. Guo Y, Wu C, Yuan Z, Wang Y, Liang Z, Wang Y, et al. Gene-Based testing of interactions using XGBoost in Genome-Wide association studies. Front Cell Dev Biol. 2021;9:801113.
145. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93(2):278–88.
146. Badré A, Zhang L, Muchero W, Reynolds JC, Pan C. Deep neural network improves the Estimation of polygenic risk scores for breast cancer. J Hum Genet. 2021;66(4):359–69.
147. Twede H, Conard AM, Pais L, Bryen S, O'Heir E, Smith G et al. Evidence Aggregator: AI reasoning applied to rare disease diagnostics. bioRxiv. 2025:2025.03.10.642480.
148. Fowler DM, Rehm HL. Will variants of uncertain significance still exist in 2030? Am J Hum Genet. 2024;111(1):5–10.
149. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large Language models for decision support in personalized oncology. JAMA Netw Open. 2023;6(11):e2343689.
150. McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet. 2021;108(4):535–48.
151. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the End-Game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet. 2020;11:350.
152. Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. Electronics. 2023;12(8):1789.
153. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine learning and integrative analysis of biomedical big data. Genes (Basel). 2019;10(2).
154. Wang R, Chaudhari P, Davatzikos C. Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies. Proc Natl Acad Sci. 2023;120(6):e2211613120.
155. Dwarshuis N, Tonner P, Olson ND, Sedlazeck FJ, Wagner J, Zook JM. StratoMod: predicting sequencing and variant calling errors with interpretable machine learning. Commun Biology. 2024;7(1):1316.
156. Lee Y-S, Yen S-J, Jiang W, Chen J, Chang C-Y. Illuminating the black box: an interpretable machine learning based on ensemble trees. Expert Syst Appl. 2025;272:126720.
157. Karim MR, Islam T, Shajalal M, Beyan O, Lange C, Cochez M et al. Explainable AI for bioinformatics: methods, tools and applications. Brief Bioinform. 2023;24(5).
158. Patharkar A, Cai F, Al-Hindawi F, Wu T. Predictive modeling of biomedical Temporal data in healthcare applications: review and future directions. Front Physiol. 2024;15:1386760.
159. Field MA. Bioinformatic challenges detecting genetic variation in precision medicine programs. Front Med. 2022;9.
160. von Gerich H, Chomutare T, Peltonen LM. Building bridges for federated learning in healthcare: review on approaches for common data model development. Stud Health Technol Inf. 2024;315:711–2.
161. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018.
162. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. Npj Digit Med. 2018;1(1):18.
163. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123–44.
164. Waardenberg AJ, Field MA. ConsensusDE: an R package for assessing consensus of multiple RNA-seq algorithms with RUV correction. PeerJ. 2019;7:e8206.
165. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de Novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216–21.
166. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De Novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237–41.

167. CAG Repeat Not. Polyglutamine length determines timing of huntington's disease onset. Cell. 2019;178(4):887–e90014.
168. Field MA. Detecting pathogenic variants in autoimmune diseases using high-throughput sequencing. Immunol Cell Biol. 2020.
169. Jiang SH, Mercan S, Papa I, Moldovan M, Walters GD, Koina M, et al. Deletions in VANGL1 are a risk factor for antibody-mediated kidney disease. Cell Rep Med. 2021;2(12):100475.
170. Aartsma-Rus A, Krieg AM. FDA approves Eteplirsen for Duchenne muscular dystrophy: the next chapter in the Eteplirsen Saga. Nucleic Acid Ther. 2017;27(1):1–3.
171. Tsimberidou AM, Iskander NG, Hong DS, Wheler JJ, Falchook GS, Fu S, et al. Personalized medicine in a phase I clinical trials program: the MD Anderson cancer center initiative. Clin Cancer Res. 2012;18(22):6373–83.
172. Carter TC, He MM. Challenges of Identifying Clinically Actionable Genetic Variants for Precision Medicine. J Healthc Eng. 2016;2016.
173. Wang Z, Zhao G, Zhu Z, Wang Y, Xiang X, Zhang S, et al. VarCards2: an integrated genetic and clinical database for ACMG-AMP variant-interpretation guidelines in the human whole genome. Nucleic Acids Res. 2024;52(D1):D1478–89.
174. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. Genet Med. 2015;17(5):405–24.
175. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al. Performance of ACMG-AMP Variant-Interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. Am J Hum Genet. 2016;98(6):1067–76.
176. Harrison SM, Riggs ER, Maglott DR, Lee JM, Azzariti DR, Niehaus A et al. Using ClinVar as a Resource to Support Variant Interpretation. Curr Protoc Hum Genet. 2016;89:8.16.1-8.23.
177. Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, Babb LJ et al. GA4GH: international policies and standards for data sharing across genomic research and healthcare. Cell Genom. 2021;1(2).
178. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008;9:403–33.
179. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
180. Rosas LG, Nasrallah C, Park VT, Vasquez JJ, Duron Y, Garrick O, et al. Perspectives on precision health among racial/ethnic minority communities and the physicians that serve them. Ethn Dis. 2020;30(Suppl 1):137–48.
181. Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. PLoS ONE. 2018;13(10):e0204937.
182. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.
183. Thang MWC, Chua XY, Price G, Gorse D, Field MA. MetaDEGalaxy: galaxy workflow for differential abundance analysis of 16s metagenomic data. F1000Res. 2019;8:726.
184. Khudhair Z, Alhallaf R, Eichenberger RM, Field M, Krause L, Sotillo J et al. Administration of hookworm excretory/secretory proteins improves glucose tolerance in a mouse model of type 2 diabetes. Biomolecules. 2022;12(5).
185. Pierce DR, McDonald M, Merone L, Becker L, Thompson F, Lewis C, et al. Effect of experimental hookworm infection on insulin resistance in people at risk of type 2 diabetes. Nat Commun. 2023;14(1):4503.
186. Khudhair Z, Alhallaf R, Eichenberger RM, Whan J, Kupz A, Field M, et al. Gastrointestinal helminth infection improves insulin sensitivity, decreases systemic inflammation, and alters the composition of gut microbiota in distinct mouse models of type 2 diabetes. Front Endocrinol (Lausanne). 2020;11:606530.
187. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform. 2018;19(6):1370–81.
188. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
189. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–62.
190. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration-a comparison of unsupervised clustering methodologies. Brief Bioinform. 2019;20(4):1269–79.
191. Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform. 2020;21(2):541–52.
192. Zouk H, Yu W, Oza A, Hawley M, Vijay Kumar PK, Koch C, et al. Reanalysis of eMERGE phase III sequence variants in 10,500 participants and infrastructure to support the automated return of knowledge updates. Genet Med. 2022;24(2):454–62.
193. Manolio TA, Abramowicz M, Al-Mulla F, Anderson W, Balling R, Berger AC, et al. Global implementation of genomic medicine: we are not alone. Sci Transl Med. 2015;7(290):290ps13.
194. Wang Y, Liu L, Wang C. Trends in using deep learning algorithms in biomedical prediction systems. Front Neurosci. 2023;17:1256351.
195. Adler-Milstein J, Raphael K, Bonner A, Pelton L, Fulmer T. Hospital adoption of electronic health record functions to support age-friendly care: results from a National survey. J Am Med Inf Assoc. 2020;27(8):1206–13.

Publisher's note