Genomics & Informatics

REVIEW Open Access

Rare disease genomics and precision medicine

Juhyeon Hong[1†], Dajun Lee[1†], Ayoung Hwang[1], Taekeun Kim[1], Hong-Yeoul Ryu[2] and Jungmin Choi[1*]

Abstract

Rare diseases, though individually uncommon, collectively affect millions worldwide. Genomic technologies and big data analytics have revolutionized diagnosing and understanding these conditions. This review explores the role of genomics in rare disease research, the impact of large consortium initiatives, advancements in extensive data analysis, the integration of artificial intelligence (AI) and machine learning (ML), and the therapeutic implications in precision medicine. We also discuss the challenges of data sharing and privacy concerns, emphasizing the need for collaborative efforts and secure data practices to advance rare disease research.

Keywords Rare disease genomics, Big data analytics, Precision medicine

1 Introduction

Rare diseases pose significant challenges in diagnosis and treatment due to their low prevalence and diverse presentations. In the USA, a disease is considered rare if it affects fewer than 200,000 people, while in Europe, it is classified as rare if it affects fewer than 1 in 2000 individuals [1]. Despite their rarity, over 10,000 distinct types of rare and genetic diseases collectively affect around 400 million people globally [1]. Approximately, 80% of rare diseases are attributed to genetic causes, highlighting the importance of genetic testing for accurate diagnosis [1]. Understanding their genetic interplay has driven the development of targeted treatments such as gene therapy, gene editing, and personalized medicine.

Although big data has been integrated in rare disease genomics, major barriers still need to be addressed, including difficulties in identifying causal variants and translating findings into clinical practice. Various large consortia have increasingly emerged in response to these challenges, such as the National Biobank of Korea [2, 3], which contains the latest established large Korean rare disease cohort. The integration of AI and ML in rare disease research has improved the identification of disease-causing variants and enhanced diagnostic accuracy [4]. These technologies are driving advancements in precision medicine, enabling more personalized and effective treatments through gene-targeted therapies [5]. Data privacy concerns are inevitable in handling genomics data, and several efforts have been made to prevent the exposure of patient information, which will be explored further in this review. Hence, this paper aims to comprehensively review genomics techniques and tools used in rare disease research alongside therapeutic applications. Unlike previous review articles that have dealt with certain topics (e.g., deep learning or public health) [6, 7], this review will offer insights into the broader landscape of rare disease genomics and therapeutic medicine.

1.1 Advancements in genomic technologies for rare disease diagnosis

Diagnosing rare diseases has been historically challenging. In the late twentieth century, Sanger sequencing was

†Juhyeon Hong and Dajun Lee contributed equally to this work.

*Correspondence:
Jungmin Choi
jungminchoi@korea.ac.kr
[1] Department of Biomedical Sciences, Korea University College of Medicine, Seoul 02841, Republic of Korea
[2] School of Life Sciences, BK21 FOUR KNU Creative BioResearch Group, College of Natural Sciences, Kyungpook National University, Daegu 41566, Republic of Korea

the most commonly used technique for about 25 years [8]. However, it could only analyze one gene at a time, making it time-consuming and costly, especially in cases involving genetic heterogeneity or unclear clinical manifestations [9–12].

The advent of next-generation sequencing (NGS) about a decade ago revolutionized the diagnostic workflow. Short-read sequencing (SRS) technologies, such as exome and genome sequencing, became incorporated into routine diagnostic procedures for rare diseases [13–16]. Whole exome sequencing (WES) has been applied to patients suspected of rare diseases with unusual phenotypic characteristics (e.g., cerebellar hypoplasia, epilepsy, or global developmental delay), leading to a definitive diagnosis for 28.3% of the patients [17]. However, due to the complicated genetic underpinnings of rare diseases, NGS-based methods had a detection rate of only 25–50% in undiagnosed patients [15].

To mitigate these limitations, long-read sequencing (LRS) has emerged as a promising tool, allowing for more accurate detection of complex genetic variants such as short tandem repeats (STRs), copy number variations (CNVs), and structural variants (SVs). Two primary LRS technologies have gained prominence: Oxford Nanopore Technologies' nanopore sequencing and Pacific Biosciences' (PacBio) single-molecule, real-time (SMRT) sequencing [18]. Both technologies offer advantages in detecting complex genetic variants but differ in approach and output characteristics.

Specifically, LRS has proven successful in diagnosing previously undiagnosed rare disease patients. For instance, nanopore LRS facilitated the detection of deep intronic variants in the TSC1 and TSC2 genes, leading to the identification of aberrant splicing events and a confirmed diagnosis of tuberous sclerosis [19]. Similarly, LRS enabled the diagnosis of patients with Cornelia de Lange syndrome (CDLS) by identifying a complex chromothripsis event affecting the NIPBL gene, which had been undetectable by SRS [20]. Furthermore, PacBio HiFi reads revealed a repeat expansion in the DAB1 gene, associated with spinocerebellar ataxia 37 (SCA37), in a family exhibiting autosomal dominant ataxia [21].

These cases demonstrate the utility of LRS in resolving diagnostically challenging genetic variants, particularly complex structural variants and intronic mutations, contributing significantly to the diagnosis of rare diseases.

1.2 Collaborative efforts through large consortia

The establishment of large consortia for rare diseases addresses the need for coordinated research efforts [22]. Despite initiatives like the Rare Disease Clinical Research Network (RDCRN), rare disease research often remains siloed, focusing on single conditions [23]. In response, diverse collaborations have been launched to unite researchers and foster collaborative efforts across multiple rare diseases (Table 1).

For example, task forces (TFs) [33], adopted by the International Rare Diseases Research Consortium (IRDiRC), have addressed actionable subjects such as reducing the duration of the diagnostic process [34]. The Matchmaker Exchange (MME) TF devised a federated platform to expedite gene discovery for rare diseases by matchmaking patients with similar phenotypes. Six novel candidate genes associated with rare diseases, including armfield X-linked intellectual disability (XLID) syndrome [35], neurodevelopmental disorder [36], polyneuropathy [37], and ZNFX1 deficiency [38], were identified from undiagnosed patients enrolled in Care4Rare Canada [39] through the application of MME. These consortia function as hubs for data exchange among researchers studying rare diseases.

1.3 Big data analytics in rare disease genomics

Due to the implementation of large consortia, a lot of data, so-called big data, is accumulated, emphasizing the necessity of implementing a big data-based analysis pipeline. Processing big data presents impediments, including storage limitations, computational power requirements, and data security concerns [40]. Cloud platforms offer a scalable solution, enabling researchers to store and analyze large datasets efficiently [41]. Cloud platforms facilitate data sharing and collaboration without geographic constraints.

Researchers have increasingly utilized cloud platforms to analyze big data in rare diseases. For example, the All of Us Research Program utilizes a cloud-based Researcher Workbench built on Google Cloud through Terra, which provides secure computational power for analysis [42]. The Genome Analysis Toolkit (GATK) team recommended running GATK across various cloud platforms, particularly Terra, for its user-friendly graphical interface [43]. Amazon Web Services (AWS) hosts large public datasets, such as Genome Aggregation Database (gnomAD) [44], UK Biobank [45–47], and 100,000 Genomes Project (100KGP) [48] allowing users to analyze data and build services using a broad range of data analytics products.

DRAGEN is now widely available on platforms like Illumina Connected Analytics (ICA) and AWS Marketplace. It offers faster analysis times, requires fewer computational resources, and accurately detects various variants [49, 50]. For instance, while using BWA and HaplotypeCaller for variant calling requires 32 h, leveraging DRAGEN can significantly reduce this time to just 37 min [51]. Both methods show comparable accuracy

Table 1 Overview of large consortia/initiatives for rare diseases

Consortia/initiatives	Description	Data availability	URL	Accession number	References
Global scale					
European Joint Programme on Rare Disease (EJP RD)	Europe-wide initiative with the aim of improving diagnosis and treatment of rare diseases	Data available within the website	https://resourcemap.ejprarediseases.org/	None	[24]
International Rare Diseases Research Consortium (IRDiRC)	Global Consortium that coordinates research efforts to develop 1000 new therapies for rare diseases by 2027	Data available within the website	https://irdirc.org/resources-2/irdirc-recognized-resources/	None	[25]
National Organization for Rare Disorders (NORD)	Advocacy organization in the USA that provides support for patients and advocates for rare disease research	Data available on request from the team	https://rarediseases.org/resource-library/	None	[26]
Rare Disease Clinical Research Network (RDCRN)	International collaboration designed to develop medical research on rare diseases with increased support for clinical studies	Data available on request from the team	https://www.rarediseasesnetwork.org/research/data-sharing-and-standards/data-sharing-resources	None	[27]
Undiagnosed Diseases Network International (UDNI)	Global network focused on enhancing the understanding of diagnosis of previously undiagnosed diseases	Data not publicly available	None	None	[28]
National scale					
Canadian Organization for Rare Disorders (CORD)	Advocacy organization in Canada focused on reinforcing public policy and support for the well-being of patients with rare diseases	Data not publicly available	None	None	[29]
CIHR Rare Disease Research Initiative	Canadian program under the Canadian Institutes of Health Research found to increase collaboration across the rare disease community	Data not publicly available	None	None	[30]
Initiative on Rare and Undiagnosed Diseases in Japan	National program in Japan dedicated to advancing research and healthcare strategies for rare diseases	Data not publicly available	None	None	[31]
Korean Undiagnosed Diseases Program (KUDP)	National program in South Korea aimed at diagnosing undiagnosed patients and building long-term research infrastructure	Data available on request from the authors	None	None	[32]

in variant calling, with DRAGEN achieving 99.07% for single-nucleotide polymorphisms (SNPs) and 88.39% for insertions and deletions (indels), while Burrows-Wheeler Aligner (BWA) combined with HaplotypeCaller reaches 98.68% for SNPs and 89.45% for indels [51]. When analyzing large-scale data, cloud platforms and pipelines should be tailored to fit the user's specific data and cost requirements [52].

1.4 Artificial intelligence (AI) and machine learning (ML) in rare disease analysis

Patients with rare diseases often face challenges such as diagnostic delay and misdiagnosis, and more than 90% of rare diseases lack effective treatments [53–55]. AI and ML technologies contribute to rare disease research by assisting the analysis of vast amounts of genomic and clinical data to identify disease patterns, predict treatment outcomes, and develop personalized therapies, ultimately improving diagnostic accuracy and advancing drug development [56].

In the variant calling stage, deep learning models such as DeepVariant [57] and Clairvoyante [58] transform sequencing data into an image-like format and use convolutional neural networks (CNNs) to interpret DNA alignments as visual patterns for detecting genetic variants. Tools like NeoMutate [59], which utilize Bayesian classifiers and supervised learning algorithms, further integrate multiple methods to improve variant detection. These tools allow researchers to identify genetic variations with increased accuracy. DeepSVFilter [60], a CNN-based tool, filters SVs from genome sequencing data in the variant filtering stage. Tools like Intelli-NGS [61] use deep neural networks (DNNs) to minimize false-positive and false-negative rates, significantly improving the filtering process.

Once variants are identified, AI-driven tools aid in its annotation and prioritization. MetaSVM [62] and MetaLR [62] provide ensemble predictions for deleterious effects, while combined annotation-dependent depletion (CADD) [63] combines functional annotations and evolutionary conservation. Sorting Intolerant From Tolerant (SIFT) [64] and Polymorphism Phenotyping v2 (PolyPhen-2) [65] assess sequence homology and structural features, respectively. Variant Effect Scoring Tool (VEST3) [66] and Protein Variation Effect Analyzer (PROVEAN) [67] score the functional impact of missense mutations, and MutationTaster2 [68] incorporates evolutionary conservation and disease associations. Mendelian Clinically Applicable Pathogenicity (M-CAP) [69] classifies rare variants; Missense badness, PolyPhen-2, and Constraint (MPC) [70, 71] enhance predictions using constraint metrics, Functional Analysis through Hidden Markov Models with an eXtended Feature set (FATHMM-XF) [72], and Missense Variant Pathogenicity prediction (MVP) [73] focuses on potentially pathogenic variants. Additional tools include Skyhawk [74], DANN [75], DeepSEA [76], exome Disease Variant Analysis (eDiva) [77], and RENOVO [78], utilizing neural networks and random forest to prioritize clinically relevant variants and assess noncoding or germline variants.

AI has significantly advanced the field of phenotype-genotype association, particularly in diagnosing rare diseases. DeepGestalt [79], which employs a deep CNN, analyzes facial images to distinguish between genetic subtypes, offering powerful diagnostic support. Deep PhenomeNET Variant Predictor (DeepPVP) [80], modeled by adopting DNN, prioritizes variants by integrating patient phenotype information, enhancing the identification of disease-causing variants. Xrare [81] focuses on prioritizing causative gene variants in rare diseases by utilizing phenotype-genotype association methods, providing clinicians with a streamlined approach to diagnosis. Additionally, Super-quick Information content Random Forest Learning of Splice Variants (SQUIRLS) [82], which uses a random forest algorithm, classifies splice variants, further improving the genotype–phenotype correlation by assessing the impact of genetic variants on splicing mechanisms. These tools collectively enhance the accuracy and efficiency of rare disease diagnosis by linking phenotypic features with underlying genetic data. The integration of AI technologies with biomarker discovery from genomics data and advanced imaging diagnostics offers a promising approach to accelerating the diagnosis and treatment of rare diseases and reducing patients' diagnostic odyssey. Additionally, the widespread implementation of AI-driven tools increases accessibility. It provides more comprehensive, data-driven insights, empowering clinicians and nonspecialists to make more informed decisions in managing rare genetic diseases.

1.5 Expanding genomic research: perspectives from Korean Bio-Big Data

Despite representing about 22% of the global population, East Asians are under-represented in genetic research and are often missing from control databases. To address this imbalance, initiatives have been promoted to create a comprehensive Korean control database and to analyze the Korean Reference Genome.

Existing Korean databases include the Korean National Standard Reference Variome (KoVariome) [83], the Korean Reference Genome Database (KRGDB) [84], KOVA 2 [85, 86], the Korean Reference Genome (KRG), the Korean Genetic Diagnosis Program for Rare Diseases (KGDP), Korea4K [87], and National Biobank of Korea (Table 2) [2, 3]. KoVariome offers a comprehensive

Table 2 Major databases of Korean Bio Big Data

Database	No. of individuals	Technology	Sample type	Published year	Data availability	URL	Accession number	Reference
KoVariome	50	WGS	Healthy individuals	2018	Data available within the article or its supplementary materials	https://koreangenome.org/The_Korean_Reference_Variome:_KoVariome	None (uses FTP server)	Kim et al. [83]
KRGDB	1722	WGS	Integrated	2020	Data available on request from the authors	None	None	Jung et al. [84]
KOVA 2	1896 3409	WGS WES	Healthy individuals	2022	Data available within the article or its supplementary materials	https://www.kobic.re.kr/kova/downloads	None (uses FTP server)	Lee et al. [86]
KRG (pilot phase)	1490	WGS	Healthy individuals	2022	Data available on request from the authors	None	None	Hwang et al. [88]
KGDP (Phase II)	1890	Multi-omics	Rare disease patients	2023	Data available on request from the authors	None	None	Kim et al. [89]
Korea4K	4157	WGS	Integrated	2024	Data available on request from the authors	https://ega-archive.org/studies/EGAS00001007580	EGAD00001015348	Jeon et al. [87]
National Biobank of Korea (pilot phase)	772,319 14,905	Multi-omics WGS	Integrated Rare diseases patients and their relatives	2024	Data generated at a National Biobank of Korea, available upon request	https://biobank.nih.go.kr/cadaver/EgovPageLink.do?menuNo=34&link=eng%2Fmain%2Fcontent%2FBiobankingActives%2FControlPage	None	[2, 3]

As of September 2024

catalogue of genetic variations, including novel variants, enhancing the accuracy of identifying pathogenic genetic variants specific to the Korean population [83]. KRGDB contains genomic variant data, including frequency information, functional annotations, and genome-wide association studies (GWAS) results for common diseases [84]. The KOVA 2, built on the earlier KOVA dataset [85], offers critical insights into population-specific genetic variants and loci under selection [85, 86].

KRG project aims to identify the genome architecture of the Korean population and develop Korean-specific genomic resources, intending to include 20,000 participants [88]. KGDP Phase II enhances diagnostic capabilities through collaboration with the Korean Undiagnosed Diseases Program (KUDP) [89, 90]. In 2024, Jeon et al. presented the second phase of the Korean Genome Project (KGP), known as Korea4K, to build a comprehensive reference dataset [87]. Korea4K provides a valuable large-scale genome-phenome variome database for the Korean population and detailed information on various clinical traits, representing the most extensive genomic and phenomic data resources [87]. Beyond control databases, the rare disease cohort in the National Biobank of Korea Project includes whole genome sequencing (WGS) data from 14,905 patients in a pilot study, aiming to expand to a cohort of 400,000 by 2028. A pilot study on this rare disease cohort enables estimation of neuronal intranuclear inclusion disease (NIID) prevalence in the Korean population [91].

1.6 Strategies for identifying and characterizing pathogenic variants

The process of data acquisition, identifying and characterizing genetic variants, followed by clinical application,

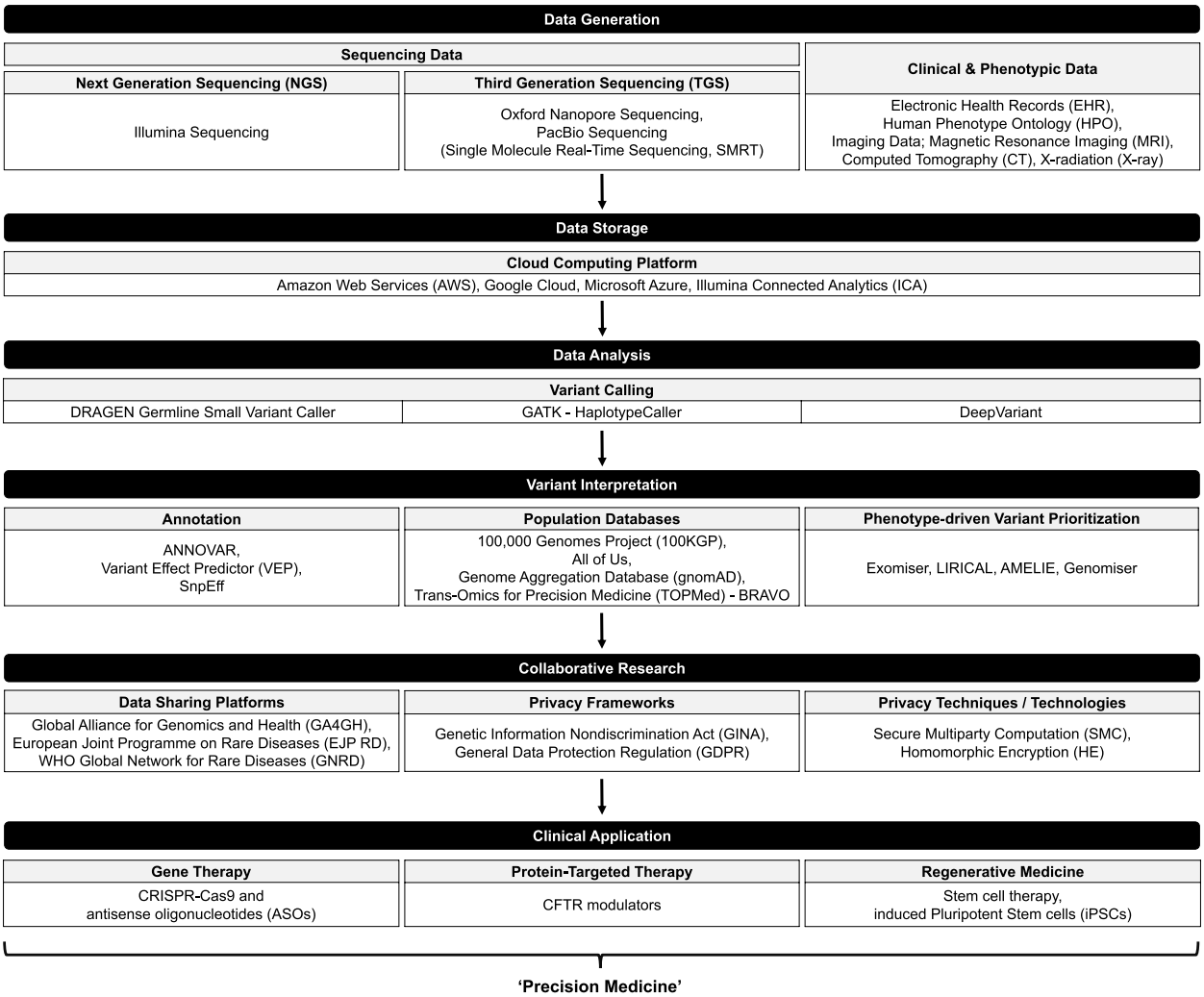

Fig. 1 Integrated workflow for rare disease diagnosis and research

involves multiple steps (Fig. 1). While single-nucleotide variant and small insertion and deletion variant calling has been robust along with the development of variant calling tools like GATK, DRAGEN, and DeepVariant, interpreting variants' pathogenicity and their relevance to specific phenotypes remains challenging [57]. Annotation databases such as ANNOVAR, Variant Effect Predictor (VEP), and SnpEff [92] are publicly available for research. Still, the sheer volume of data and variability in clinical significance complicate the interpretation process [93, 94].

At the variant level, despite the availability of numerous tools for predicting the pathogenicity of missense variants [95], accurately determining the clinical significance of these variants remains a significant challenge in genomic interpretation [72, 96–98]. Deep learning models like AlphaMissense and PrimateAI-3D have recently been developed to predict variants' pathogenicity [99, 100]. AlphaMissense utilizes AlphaFold's structural predictions and evolutionary conservation to achieve 90% precision on the ClinVar dataset [101], excelling in identifying deleterious variants in conserved regions and correlating well with multiplexed assays of variant effect (MAVEs) data [99, 102]. PrimateAI-3D outperforms AlphaMissense in real-world cohorts, including rare disease patients with clinical characteristics, including developmental disorders (DDD), autism spectrum disorders (ASD), and congenital heart disorders (CHD). It shows superior predictive power in biobank phenotypes and proteomics [103].

Another essential aspect of variant characterization and interpretation is the frequency of variants. Large population databases such as the gnomAD and NHLBI's Trans-Omics for Precision Medicine (TOPMed)-BRAVO help researchers determine how rare a variant is [70, 104]. Rare variants are frequently linked to rare diseases

due to their potential to disrupt critical biological functions or pathways essential for health. Their low frequency in the general population often reflects negative selection effects, as highly pathogenic variants tend to be eliminated from the gene pool over time due to their detrimental impact on reproductive fitness.

For instance, the identification of a novel variant in the NSD1 gene, which has been reported to occur at a low allele frequency (MAF = 0.006%, 7/114,570) in gnomAD3.1.1, has provided valuable insights into its potential pathogenic role in patients with Sotos syndrome [105]. However, common variants also play a role in rare disease etiology as genetic modifiers influencing disease onset, progression, or severity [106]. In this context, polygenic risk scores (PRS), which aggregate the effects of many common variants, are increasingly being explored in rare disease genetics to help explain variable expressivity and incomplete penetrance and to potentially improve diagnostic and prognostic accuracy in conjunction with rare variant analysis [107]. For instance, the study examined 2759 cases with developmental and epileptic encephalopathies (DEEs) or epilepsy with intellectual disability (ID) and 447,760 population-matched controls to explore the relevance of PRS [108]. It found that even in cases with a known deleterious variant, common genetic variation contributes significantly to the risk, explaining between 0.08 and 3.3% of the phenotypic variance across epilepsy subtypes [108].

The landscape of rare disease genetics has evolved significantly with the advent of WGS. While historically focused on exonic mutations, research now recognizes the importance of noncoding regions in harboring disease-causing variants [109, 110]. However, accurately classifying these noncoding variants remains challenging. Current guidelines, such as those from American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP), primarily address coding variants [111], leaving a gap in interpreting noncoding variants [112]. To address this, new recommendations have emerged, focusing on defining regulatory regions, filtering clinically relevant variants, incorporating functional evidence (e.g., RNA sequencing, chromatin interaction assays), and applying bioinformatics tools like SpliceAI [113], MotifbreakR [114], and UTRannotator [115] to assess their pathogenicity [112]. These approaches aim to provide a more comprehensive framework for evaluating variants across the entire genome, potentially enhancing rare disease diagnosis and understanding.

Finding the causal variant of rare diseases necessitates precise evaluation and prioritization of genetic variants. Previous prioritization methods have primarily focused on in silico assessments of variant pathogenicity, resulting in decreased sensitivity and difficulties in understanding the results. While valuable, manual curation of genetic variants is limited by human error, subjectivity, and the overwhelming volume of data produced by NGS technologies. These biases can lead to missed or incorrectly prioritized variants, particularly in noncoding regions or when dealing with novel variants lacking extensive annotation. Automated gene/variant prioritization tools such as Exomiser [116], MAVERICK [117], LIkelihood Ratio Interpretation of Clinical AbnormaLities (LIRICAL) [118], Automatic Mendelian Literature Evaluation (AMELIE) [119], and Genomiser [120] significantly reduce manual curation efforts and minimize human bias in rare disease diagnosis. These tools integrate diverse information sources to generate a ranked list of candidate causal genes or variants, including phenotypic data encoded as Human Phenotype Ontology (HPO) terms, known disease associations, and functional predictions [118, 119]. By systematically and exhaustively analyzing vast amounts of data, these resources provide a comprehensive and unbiased approach to variant interpretation, surpassing the limitations of manual literature searches. This automation improves diagnostic precision and efficiency and enables more consistent and reproducible results across different clinical settings. Consequently, these tools enhance treatment strategies and patient outcomes in precision medicine, offering a scalable solution to the growing complexity of genomic interpretation in rare disease diagnostics.

1.7 Therapeutic innovations and precision medicine approaches

Therapeutic implications and precision medicine for rare diseases increasingly rely on advanced genomic technologies like WES and WGS. These tools enable the identification of pathogenic variants, allowing for tailored treatment strategies. Gene therapies, such as clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 and antisense oligonucleotides (ASOs), are at the forefront of this approach. For example, onasemnogene abeparvovec (Zolgensma) treats spinal muscular atrophy (SMA) by delivering a functional SMN1 gene [121], while nusinersen (Spinraza) modifies SMN2 splicing to enhance functional protein levels [122].

Protein-targeted therapies, like CFTR modulators for cystic fibrosis, improve defective protein function directly [123]. Recent advancements in regenerative medicine, including stem cell therapy and induced pluripotent stem cells (iPSCs), also offer promising avenues for repairing damaged tissues [124]. Together, these innovative strategies enhance patient outcomes and demonstrate the potential of precision medicine in rare disease treatment.

1.8 Challenges in data sharing and privacy concerns

Data sharing between researchers is essential in advancing rare disease research, as it increases diagnostic yield and unravels the underlying disease mechanisms. For instance, the German TRANSLATE-NAMSE project found that interdisciplinary case conferences led to definitive diagnoses for 32% of pediatric and 26% of adult patients previously undiagnosed [6, 125]. The Global Alliance for Genomics and Health (GA4GH), an international coalition with members from over 90 countries, was established to facilitate sharing of genomic and clinical data and promote interoperability among institutions.

The European Joint Programme on Rare Diseases (EJP RD), one of 24 'Driver Projects' of GA4GH, maintains repositories containing more than 130,000 WES and WGS datasets across multiple resources including the European Genome-Phenome Archive (EGA), DECIPHER, and the RD-Connect Genome-Phenome Analysis Platform (GPAP) [126]. In 2023, EJP RD launched a Virtual Platform, a public portal that provides access to Findable, Accessible, Interoperable, and Reusable (FAIR)-compliant resources, streamlining data searching while safeguarding patient confidentiality [127]. International data exchange brings significant benefits.

However, data privacy remains a critical challenge, particularly for genomic and clinical data. Data misuse can violate the privacy of individuals and their biological relatives. Individual patients can be uniquely identified through distinctive genetic markers, such as rare single-nucleotide variants (SNVs) specific to their genome [128, 129].

To tackle this privacy concern, frameworks such as the Genetic Information Nondiscrimination Act of 2008 (GINA) and the General Data Protection Regulation (GDPR) [130, 131] have introduced frameworks ensuring data security. Despite these efforts, legal protections remain inconsistent, especially in the USA, where federal laws like HIPAA provide limited protection, particularly once data has been anonymized, as this anonymized data can be reidentified using several techniques, such as surname inference [132]. Some participants in the 100KGP were reidentified as their surnames could be inferred by analyzing Y-chromosome STRs and cross-referencing with genealogy databases [133].

There is an unavoidable trade-off between data privacy concerns and the societal benefits of data sharing. An approach to mitigate the risk of reidentification includes employing cryptographic methods, such as secure multiparty computation (SMC), to secure genomic data sharing and allow computations without exposing raw data. SMC enables multiple parties to jointly compute GWAS statistics, such as minor allele frequency, without sharing their raw data [134]. Ultimately, privacy-preserving strategies should be prioritized to ensure the benefits of data sharing in rare disease research do not come at the cost of individual privacy.

2 Conclusion

The field of rare disease research has undergone significant advancements, driven by technological innovations in genomic sequencing, big data analytics, and AI. LRS technologies, cloud computing platforms, and AI/ML-driven tools have greatly enhanced our ability to detect complex genetic variants and interpret their clinical significance. Large-scale collaborative efforts and the establishment of comprehensive genomic databases have expanded our knowledge of rare diseases.

Although significant progress has been made, challenges continue to arise. The complexity of variant interpretation calls for advanced prediction tools and automated systems for prioritization. Additionally, while sharing data is crucial for further research, it introduces privacy concerns that must be addressed through robust legal frameworks and advanced privacy-preserving technologies.

The integration of multi-omics data, the refinement of AI models, and the expansion of diverse population databases will be vital in advancing the diagnosis and treatment of rare diseases. The emergence of precision medicine, mainly through gene and protein-targeted therapies, highlights its potential in rare disease management. As the field continues to balance collaborative data sharing with stringent privacy protections, significant progress is expected in understanding, diagnosing, and treating rare diseases, ultimately enhancing the lives of millions of affected individuals worldwide.

Acknowledgements
This work was supported by Korea University College of Medicine

Authors' contributions
J.H., D.L., A.H., and T.K. contributed to the original draft and figure/table preparation. H-.Y.R., and J.C. designed and edited the manuscript.

Funding
No fundings.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Received: 17 September 2024 Accepted: 16 November 2024
Published online: 03 December 2024

References

1. RARE disease facts 2018 [Available from: https://globalgenes.org/rare-disease-facts/.
2. CODA. Public Resources 2024 [cited 2024 Aug 31]. Available from: https://coda.nih.go.kr/stats/selectRegList.do.
3. The National Project of Bio Big Data [cited 2024 Aug 31]. Available from: https://bighug.nih.go.kr/bigdata/.
4. Abdallah S, Sharifa M, MK IKA, Khawar MM Sr, Shaikh U, Balabel KM, et al. The impact of artificial intelligence on optimizing diagnosis and treatment plans for rare genetic disorders. Cureus. 2023;15(10):e46860.
5. Yu TW, Kingsmore SF, Green RC, MacKenzie T, Wasserstein M, Caggana M, et al. Are we prepared to deliver gene-targeted therapies for rare diseases? Am J Med Genet C Semin Med Genet. 2023;193(1):7–12.
6. Taruscio D, Gahl WA. Rare diseases: challenges and opportunities for research and public health. Nat Rev Dis Primers. 2024;10(1):13.
7. Lee J, Liu C, Kim J, Chen Z, Sun Y, Rogers JR, et al. Deep learning for rare disease: a scoping review. J Biomed Inform. 2022;135:104227.
8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.
9. Vinksel M, Writzl K, Maver A, Peterlin B. Improving diagnostics of rare genetic diseases with NGS approaches. J Community Genet. 2021;12(2):247–56.
10. Jamuar SS, Tan EC. Clinical application of next-generation sequencing for Mendelian diseases. Hum Genomics. 2015;9(1):10.
11. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97.
12. Payne K, Gavan SP, Wright SJ, Thompson AJ. Cost-effectiveness analyses of genetic and genomic diagnostic tests. Nat Rev Genet. 2018;19(4):235–46.
13. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009;106(45):19096–101.
14. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: from raw data to diagnosis with lifetime value. Clin Genet. 2018;93(3):508–19.
15. Sullivan JA, Schoch K, Spillmann RC, Shashi V. Exome/genome sequencing in undiagnosed syndromes. Annu Rev Med. 2023;74:489–502.
16. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19(5):325.
17. Lee JY, Oh SH, Keum C, Lee BL, Chung WY. Clinical application of prospective whole-exome sequencing in the diagnosis of genetic disease: experience of a regional disease center in South Korea. Ann Hum Genet. 2024;88(2):101–12.
18. Udaondo Z, Sittikankaew K, Uengwetwanit T, Wongsurawat T, Sonthirod C, Jenjaroenpun P, et al. Comparative analysis of PacBio and Oxford Nanopore Sequencing Technologies for transcriptomic landscape identification of Penaeus monodon. Life (Basel). 2021;11(8):862.
19. Ura H, Togi S, Niida Y. Target-capture full-length double-stranded cDNA long-read sequencing through nanopore revealed novel intron retention in patient with tuberous sclerosis complex. Front Genet. 2023;14:1256064.
20. Bestetti I, Crippa M, Sironi A, Bellini M, Tumiatti F, Ballabio S, et al. Long-read sequencing reveals chromothripsis in a molecularly unsolved case of Cornelia de Lange syndrome. Front Genet. 2024;15:1358334.
21. Steyaert W, Sagath L, Demidov G, et al. Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. Preprint. medRxiv. 2024;2024.05.03.24305331.https://doi.org/10.1101/2024.05.03.24305331.
22. Berry SA, Coughlin CR 2nd, McCandless S, McCarter R, Seminara J, Yudkoff M, et al. Developing interactions with industry in rare diseases: lessons learned and continuing challenges. Genet Med. 2020;22(1):219–26.
23. Genes G, Alliance G. Rare diseases, common challenges. Nat Genet. 2022;54(3):215.
24. The European Joint Programme on Rare Diseases (EJP RD) 2021. Available from: https://www.ejprarediseases.org/.
25. International Rare Diseases Research Consortium (IRDiRC) 2022. Available from: https://irdirc.org/.
26. NORD. Barriers to rare disease diagnosis, care and treatment in the US2020. Available from: https://rarediseases.org/wp-content/uploads/2020/11/NRD-2088-Barriers-30-Yr-Survey-Report_FNL-2.pdf.
27. Rare Disease Clinical Research Network (RDNRN) 2002. Available from: https://www.rarediseasesnetwork.org/.
28. UDNI - Undiagnosed Diseases Network International. Available from: https://www.udninternational.org/.
29. Canadian Organization for Rare Disorders 2024. Available from: https://www.raredisorders.ca/.
30. CIHR Rare Disease Research Initiative 2023. Available from: https://cihr-irsc.gc.ca/.
31. Takahashi Y, Mizusawa H. Initiative on rare and undiagnosed disease in Japan. JMA J. 2021;4(2):112–8.
32. Kim SY, Lee S, Woo H, Han J, Ko YJ, Shim Y, et al. The Korean undiagnosed diseases program phase I: expansion of the nationwide network and the development of long-term infrastructure. Orphanet J Rare Dis. 2022;17(1):372.
33. PLUTO PROJECT – Disregarded Rare Diseases. Available from: https://irdirc.org/pluto-project-disregarded-rare-diseases/.
34. Monaco L, Zanello G, Baynam G, Jonker AH, Julkowska D, Hartman AL, et al. Research on rare diseases: ten years of progress and challenges at IRDiRC. Nat Rev Drug Discov. 2022;21(5):319–20.
35. Lee YR, Khan K, Armfield-Uhas K, Srikanth S, Thompson NA, Pardo M, et al. Mutations in FAM50A suggest that Armfield XLID syndrome is a spliceosomopathy. Nat Commun. 2020;11(1):3698.
36. Salpietro V, Dixon CL, Guo H, Bello OD, Vandrovcova J, Efthymiou S, et al. AMPA receptor GluA2 subunit defects are a cause of neurodevelopmental disorders. Nat Commun. 2019;10(1):3094.
37. Chelban V, Wilson MP, Warman Chardon J, Vandrovcova J, Zanetti MN, Zamba-Papanicolaou E, et al. PDXK mutations cause polyneuropathy responsive to pyridoxal 5'-phosphate supplementation. Ann Neurol. 2019;86(2):225–40.
38. Vavassori S, Chou J, Faletti LE, Haunerdinger V, Opitz L, Joset P, et al. Multisystem inflammation and susceptibility to viral infections in human ZNFX1 deficiency. J Allergy Clin Immunol. 2021;148(2):381–93.
39. Osmond M, Hartley T, Dyment DA, Kernohan KD, Brudno M, Buske OJ, et al. Outcome of over 1500 matches through the Matchmaker Exchange for rare disease gene discovery: the 2-year experience of Care4Rare Canada. Genet Med. 2022;24(1):100–8.
40. Sitalakshmi Venkatraman RV. Big data security challenges and strategies. AIMS Press. 2019.
41. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018;19(4):208–19.
42. Ramirez AH, Sulieman L, Schlueter DJ, Halvorson A, Qian J, Ratsimbazafy F, et al. The All of Us Research Program: data quality, utility, and diversity. Patterns (N Y). 2022;3(8):100570.
43. Running GATK on the cloud (Overview) 2024. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360057893792-Running-GATK-on-the-cloud-Overview.
44. Broad Institute gnomAD Dataset on AWS 2017. Available from: https://registry.opendata.aws/broad-gnomad/.
45. Analytic and Translational Genetics Unit MGHatBI. Pan UK-Biobank: Pan-ancestry genetic analysis of the UK Biobank [Available from: https://pan.ukbb.broadinstitute.org/downloads. Accessed 22 Nov 2024.
46. Analytic and Translational Genetics Unit MGHatBI. UK Biobank Pan-Ancestry Summary Statistics [Available from: https://registry.opendata.aws/broad-pan-ukb. Accessed 22 Nov 2024.
47. UK Biobank Linkage Disequilibrium Matrices. 2020. Available from: https://registry.opendata.aws/ukbb-ld/.
48. Weissbrod O. UK Biobank linkage disequilibrium matrices [Available from: https://registry.opendata.aws/ukbb-ld. Accessed 22 Nov 2024.
49. Betschart RO, Thiery A, Aguilera-Garcia D, Zoche M, Moch H, Twerenbold R, et al. Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. Sci Rep. 2022;12(1):21502.
50. Behera S, Catreux S, Rossi M, et al. Comprehensive and accurate genome analysis at scale using DRAGEN accelerated algorithms. Preprint. bioRxiv. 2024;2024.01.02.573821. https://doi.org/10.1101/2024.01.02.573821.
51. Goyal A, Kwon H, Lee K, Garg R, Yun S, Hee Kim Y, Lee S, Seob Lee M. Ultra-fast next generation human genome sequencing data processing using DRAGENTM Bio-IT processor for precision medicine. Open J Genet. 2017;7:9–19. https://doi.org/10.4236/ojgen.2017.71002.

52. Banimfreg BH. A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics. Healthcare Analytics. 2023;3.
53. Faye F, Crocione C, Anido de Pena R, Bellagambi S, Escati Penaloza L, Hunter A, et al. Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a rare barometer retrospective patient survey. Eur J Hum Genet. 2024;32(9):1116–26.
54. Wojtara M, Rana E, Rahman T, Khanna P, Singh H. Artificial intelligence in rare disease diagnosis and treatment. Clin Transl Sci. 2023;16(11):2106–11.
55. Kaufmann P, Pariser AR, Austin C. From scientific discovery to treatments for rare diseases - the view from the National Center for Advancing Translational Sciences - Office of Rare Diseases Research. Orphanet J Rare Dis. 2018;13(1):196.
56. Visibelli A, Roncaglia B, Spiga O, Santucci A. The impact of artificial intelligence in the odyssey of rare diseases. Biomedicines. 2023;11(3):887.
57. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7.
58. Luo R, Sedlazeck FJ, Lam TW, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10(1):998.
59. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med Genomics. 2019;12(1):63.
60. Liu Y, Huang Y, Wang G, Wang Y. A deep learning approach for filtering structural variants in short read sequencing data. Brief Bioinform. 2021;22(4):bbaa370.
61. Park Y, Heider D, Hauschild AC. Integrative analysis of next-generation sequencing for next-generation cancer research toward artificial intelligence. Cancers (Basel). 2021;13(13):3148.
62. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37.
63. 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.
64. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
65. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.
66. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3.
67. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745–7.
68. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361–2.
69. 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.
70. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
71. Bu F, Zhong M, Chen Q, Wang Y, Zhao X, Zhang Q, et al. DVPred: a disease-specific prediction tool for variant pathogenicity classification for hearing loss. Hum Genet. 2022;141(3–4):401–11.
72. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34(3):511–3.
73. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12(1):510.
74. Ruibang Luo T-WL, Michael C. Schatz. Skyhawk: an artificial neural network-based discriminator for reviewing clinically significant genomic variants. Int J Comput Biol Drug Des. 2021;13:5–6.
75. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761–3.
76. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4.
77. Bosio M, Drechsel O, Rahman R, Muyas F, Rabionet R, Bezdan D, et al. eDiVA-classification and prioritization of pathogenic variants for clinical diagnostics. Hum Mutat. 2019;40(7):865–78.
78. Favalli V, Tini G, Bonetti E, Vozza G, Guida A, Gandini S, et al. Machine learning-based reclassification of germline variants of unknown significance: the RENOVO algorithm. Am J Hum Genet. 2021;108(4):682–95.
79. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60–4.
80. Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: phenotype-based prioritization of causative variants using deep learning. BMC Bioinformatics. 2019;20(1):65.
81. Li Q, Zhao K, Bustamante CD, Ma X, Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med. 2019;21(9):2126–34.
82. Danis D, Jacobsen JOB, Carmody LC, Gargano MA, McMurry JA, Hegde A, et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. Am J Hum Genet. 2021;108(9):1564–77.
83. Kim J, Weber JA, Jho S, Jang J, Jun J, Cho YS, et al. KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. Sci Rep. 2018;8(1):5677.
84. Jung KS, Hong KW, Jo HY, Choi J, Ban HJ, Cho SB, et al. KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. Database (Oxford). 2020;2020:baz146.
85. Lee S, Seo J, Park J, Nam JY, Choi A, Ignatius JS, et al. Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. Sci Rep. 2017;7(1):4287.
86. Lee J, Lee J, Jeon S, Lee J, Jang I, Yang JO, et al. A database of 5305 healthy Korean individuals reveals genetic and clinical implications for an East Asian population. Exp Mol Med. 2022;54(11):1862–71.
87. Jeon S, Choi H, Jeon Y, Choi WH, Choi H, An K, et al. Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups. Gigascience. 2024;13:giae014.
88. Hwang MY, Choi NH, Won HH, Kim BJ, Kim YJ. Analyzing the Korean reference genome with meta-imputation increased the imputation accuracy and spectrum of rare variants in the Korean population. Front Genet. 2022;13:1008646.
89. Kim MJ, Kim B, Lee H, Lee JS, Chae SW, Shin HS, et al. The Korean Genetic Diagnosis Program for Rare Disease Phase II: outcomes of a 6-year national project. Eur J Hum Genet. 2023;31(10):1147–53.
90. Kim SY, Lim BC, Lee JS, Kim WJ, Kim H, Ko JM, et al. The Korean undiagnosed diseases program: lessons from a one-year pilot project. Orphanet J Rare Dis. 2019;14(1):68.
91. Lee S, Yoon JG, Hong J, Kim T, Kim N, Vandrovcova J, et al. Prevalence and characterization of NOTCH2NLC GGC repeat expansions in Koreans: from a hospital cohort analysis to a population-wide study. Neurol Genet. 2024;10(3): e200147.
92. Cingolani P, Platts A, le Wang 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.
93. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16).
94. 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.
95. Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36(5):513–23.
96. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat. 2016;37(3):235–41.

97. 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.
98. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.
99. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
100. Gao H, Hamp T, Ede J, Schraiber JG, McRae J, Singer-Berk M, et al. The landscape of tolerated genetic variation in humans and primates. Science. 2023;380(6648):eabn8153.
101. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91–5.
102. Minton K. Predicting variant pathogenicity with AlphaMissense. Nat Rev Genet. 2023;24(12):804.
103. David A. Parry TB, Tobias Hamp, Petko P. Fiziev, Abhishek Sharma, Irfahan Kassam, Jeremy McRae, View ORCID ProfileKyle Kai-How Farh. PrimateAI-3D outperforms AlphaMissense in real-world cohorts. 2024.
104. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. Nature. 2021;590(7845):290–9.
105. 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.
106. Timberlake AT, Choi J, Zaidi S, Lu Q, Nelson-Williams C, Brooks ED, et al. Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles. Elife. 2016;5:e20125.
107. Wang Z, Choi SW, Chami N, Boerwinkle E, Fornage M, Redline S, et al. The value of rare genetic variation in the prediction of common obesity in European ancestry populations. Front Endocrinol (Lausanne). 2022;13:863893.
108. Campbell C, Leu C, Feng YA, Wolking S, Moreau C, Ellis C, et al. The role of common genetic variation in presumed monogenic epilepsies. EBioMedicine. 2022;81:104098.
109. Shin T, Song JHT, Kosicki M, Kenny C, Beck SG, Kelley L, et al. Rare variation in non-coding regions with evolutionary signatures contributes to autism spectrum disorder risk. Cell Genom. 2024;4(8):100609.
110. Chen Y, Dawes R, Kim HC, Ljungdahl A, Stenton SL, Walker S, et al. De novo variants in the RNU4-2 snRNA cause a frequent neurodevelopmental syndrome. Nature. 2024;632(8026):832–40.
111. 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.
112. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, Campbell C, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med. 2022;14(1):73.
113. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using pangolin. Genome Biol. 2022;23(1):103.
114. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics. 2015;31(23):3847–9.
115. Zhang X, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5′untranslated region variants with the UTRannotator. Bioinformatics. 2021;37(8):1171–3.
116. Robinson PN, Kohler S, Oellrich A, Sanger Mouse Genetics P, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8.
117. 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.
118. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Am J Hum Genet. 2020;107(3):403–17.
119. Birgmeier J, Haeussler M, Deisseroth CA, Steinberg EH, Jagadeesh KA, Ratner AJ, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020;12(544):eaau9113.
120. Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet. 2016;99(3):595–606.
121. Strauss KA, Farrar MA, Muntoni F, Saito K, Mendell JR, Servais L, et al. Onasemnogene abeparvovec for presymptomatic infants with two copies of SMN2 at risk for spinal muscular atrophy type 1: the phase III SPR1NT trial. Nat Med. 2022;28(7):1381–9.
122. Neil EE, Bisaccia EK. Nusinersen: a novel antisense oligonucleotide for the treatment of spinal muscular atrophy. J Pediatr Pharmacol Ther. 2019;24(3):194–203.
123. Burgener EB, Moss RB. Cystic fibrosis transmembrane conductance regulator modulators: precision medicine in cystic fibrosis. Curr Opin Pediatr. 2018;30(3):372–7.
124. Anderson RH, Francis KR. Modeling rare diseases with induced pluripotent stem cell technology. Mol Cell Probes. 2018;40:52–9.
125. Rillig F, Gruters A, Schramm C, Krude H. The interdisciplinary diagnosis of rare diseases. Dtsch Arztebl Int. 2022;119(27–28):469–75.
126. 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):100029.
127. Yanis Mimouni JH, Yanna Petton, Pauline Adam, Clément Moreau, Ana Rath, Roseline Favresse, Birute Tumiene, Daria Julkowska. The European Joint Programme on Rare Diseases: building the rare diseases research ecosystem. Rare Dis Orphan Drugs J. 2024;3(3):N-A.
128. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet. 2020;52(7):646–54.
129. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science. 2004;305(5681):183.
130. Commission USEEO. Genetic Information Nondiscrimination Act of 2008. Available from: https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008.
131. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 20162016. Available from: https://dvbi.ru/Portals/0/DOCUMENTS_SHARE/RISK_MANAGEMENT/EBA/GDPR_eng_rus.pdf.
132. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet. 2022;23(7):429–45.
133. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321–4.
134. Xie W, Kantarcioglu M, Bush WS, Crawford D, Denny JC, Heatherly R, et al. SecureMA: protecting participant privacy in genetic association meta-analysis. Bioinformatics. 2014;30(23):3334–41.

Publisher's Note