frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 25 January 2024
DOI 10.3389/fgene.2023.1258083

OPEN ACCESS

EDITED BY
Silvia Bottini,
Université Côte d'Azur, France

REVIEWED BY
Amy Brower,
American College of Medical Genetics and Genomics (ACMG), United States
Bahareh Rabbani,
Tehran University of Medical Sciences, Iran
Ammar Husami,
Cincinnati Children's Hospital Medical Center, United States

*CORRESPONDENCE
Mohd Saberi Mohamad,
saberi@uaeu.ac.ae

RECEIVED 03 August 2023
ACCEPTED 24 November 2023
PUBLISHED 25 January 2024

CITATION
Choon YW, Choon YF, Nasarudin NA, Al Jasmi F, Remli MA, Alkayali MH and Mohamad MS (2024), Artificial intelligence and database for NGS-based diagnosis in rare disease.
Front. Genet. 14:1258083.
doi: 10.3389/fgene.2023.1258083

Artificial intelligence and database for NGS-based diagnosis in rare disease

Yee Wen Choon[1,2], Yee Fan Choon[3], Nurul Athirah Nasarudin[4], Fatma Al Jasmi[4], Muhamad Akmal Remli[1,2], Mohammed Hassan Alkayali[5] and Mohd Saberi Mohamad[4]*

[1]Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, Kota Bharu, Kelantan, Malaysia, [2]Faculty of Data Science and Informatics, Universiti Malaysia Kelantan, Kota Bharu, Kelantan, Malaysia, [3]Faculty of Dentistry, Lincoln University College, Petaling Jaya, Selangor, Malaysia, [4]Health Data Science Lab, Department of Genetics and Genomics, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates, [5]School of Postgraduate Studies, United Arab Emirates University, Al Ain, United Arab Emirates

Rare diseases (RDs) are rare complex genetic diseases affecting a conservative estimate of 300 million people worldwide. Recent Next-Generation Sequencing (NGS) studies are unraveling the underlying genetic heterogeneity of this group of diseases. NGS-based methods used in RDs studies have improved the diagnosis and management of RDs. Concomitantly, a suite of bioinformatics tools has been developed to sort through big data generated by NGS to understand RDs better. However, there are concerns regarding the lack of consistency among different methods, primarily linked to factors such as the lack of uniformity in input and output formats, the absence of a standardized measure for predictive accuracy, and the regularity of updates to the annotation database. Today, artificial intelligence (AI), particularly deep learning, is widely used in a variety of biological contexts, changing the healthcare system. AI has demonstrated promising capabilities in boosting variant calling precision, refining variant prediction, and enhancing the user-friendliness of electronic health record (EHR) systems in NGS-based diagnostics. This paper reviews the state of the art of AI in NGS-based genetics, and its future directions and challenges. It also compare several rare disease databases.

KEYWORDS

rare disease, diagnosis, next-generation sequencing, artificial intelligence, machine learning, data science

1 Introduction

Collectively, rare diseases (RDs) are a diverse group of heterogeneous diseases with approximately 7,000 distinct clinical entities. These diseases are commonly a result of genetic aberrations with early onset in children (Amberger et al., 2015; Wright et al., 2018; Tatiana and Tarailo-Graovac, 2019). Despite their rarity, RDs are emerging as a priority in global public health policy. An estimated 3.5%–5.9% of the world's population (263–446 million persons) is burdened by RDs (Taruscio et al., 2010; Khosla and Valdez, 2018; Nguengang Wakap et al., 2020). RDs collectively affect a significant number of people worldwide. While each individual rare disease may impact only a small number of patients, when considered as a group, rare diseases have a substantial impact on public health. Furthermore, patients with RDs' are challenged by: 1) the struggle to locate knowledgeable clinicians to diagnose and

manage their conditions, resulting in delay-, under-or misdiagnosis, 2) costly disease-specific medications, 3) the struggle faced by clinicians to improve their competencies in managing RDs, which depends proportionately on the availability of the cases, and 4) difficulties in assembling cohorts of patients for clinical study, availability of drugs or devices, and a lack of funding to understand RDs better. Nevertheless, the emergence of various advocacy organizations and emerging genomics technologies have sped up the efforts to find cures and amelioration for this group of diseases (Elliott and Zurynski, 2015; Austin et al., 2018; Stoller, 2018; Liu et al., 2019a; Maroilley and Tarailo-Graovac, 2019; Baynam et al., 2020).

Rare diseases are inherently uncommon, there are typically severe constraints on available knowledge, research, medical expertise, and treatment options for each specific rare disease. Sharing clinical and genetic data on rare diseases can be challenging due to concerns about patient privacy and data security. Moreover, the rarity of the diseases causes the data available for each specific condition is limited. This scarcity of data makes it challenging to develop comprehensive databases and reference datasets. Rare diseases, by definition, have low prevalence. This means there is often a lack of reference data and comprehensive databases specific to these conditions. Consequently, it can be difficult to assess whether a specific genetic variant is pathogenic or benign. Variants of unknown clinical significance are common in rare diseases. These are genetic variations that are not clearly associated with disease or health. Interpreting VUS accurately is crucial for making informed clinical decisions and research advancements. As technology progresses, both public and scientific awareness has been increasing, and the accumulation, combination, and sharing of extensive data are set to greatly enhance our understanding of rare diseases (Hartley et al., 2020).

High throughput sequencing technologies are becoming an armamentarium for clinicians and researchers in modern medicine, especially in RDs (Grosse et al., 2010; Soon et al., 2013; Frésard and Montgomery, 2018; Amorim et al., 2019; Nguyen, 2019; Field, 2021). Next-generation sequencing (NGS) has been instrumental in discovering many underlying genetic aberrations of RDs. Such understanding has greatly improved the diagnosis and management of RDs (Jia and Shi, 2017; Fernandez-Marmiesse et al., 2018; Liu et al., 2019b; Rey et al., 2019; Vinkšel et al., 2021). Three NGS-based methods have exponentially identified disease-associated genes in the last 10 years, for example, the discoveries of novel genetic variants associated with age-related hearing loss (ARHL) (Girotto et al., 2019), Ménière's disease (MD) (Gallego-Martinez et al., 2019) and severe congenital myasthenic syndrome with episodic apnea (CMS-EA) (Liu et al., 2019a) by targeted sequencing. It is becoming clear that genetic defects defining RDs are as heterogeneous as the disease (Liu et al., 2019b; Posey, 2019). Furthermore, the rapid accumulation of NGS-generated genomic data would challenge traditional sampling-based statistical methods' ability to identify genetic pattern. Hence, more advanced computational techniques are in order, and artificial intelligence (AI) is fast becoming a method of choice (Cai et al., 2020). This paper summarizes the current uses of AI in NGS-based genetics and its future directions and challenges.

1.1 Targeted sequencing panels

Gene panels are used to anticipate the presence of pathogenic mutations associated with specific illnesses or disease groups by identifying specific genes or coding regions within genes (Rehm, 2013). Sequences can be sequenced to deeper levels than WES and WGS using targeted panels at a lower cost. In contrast to WES and WGS, detected variants are limited to a limited set of genes. And produce a minimal amount of data; as a result, the interpretation workload is reduced, and there is much less concern about incidental findings. However, panels need to be updated regularly in light of new knowledge and gene discoveries. WES and targeted panels have limitations in identifying structural variants, repetitive elements, and mitochondrial genetic variations (Miller et al., 2017).

1.2 Whole exome sequencing

The whole-exome sequence examines protein-coding regions of the genome, the regions of the genome that account for 1%–2% of the whole genome and are responsible for 95% of all diseases. It allows for identifying variants in genes that have not yet been linked to human genes (Rabbani, Tekin, and Mahdieh, 2014). An interpretation of WES can be provided with a preselected panel or a specific set of genes. Using bioinformatics panels, the laboratory can choose from gene lists associated with phenotypes of patients. It is also possible to compare the phenotype associated with these genes with the patient's phenotype by looking at all rare and potentially damaging variants, (Yang et al., 2013). This approach enables the discovery of novel genes (novel gene association) by detecting previously undiscovered variants. Among WES's limitations are the insufficient coverage of different regions, the limited ability to detect variations in repetitive elements, and variants in cases of somatic mosaicism. Further limitations include structural and deep intronic variants. Despite this, technology has continued to advance, enabling the method to cover exons more accurately and all disease-causing intronic variants, (Vinkšel et al., 2021).

1.3 Whole-genome sequencing

Human genomes can be largely mapped using whole-genome sequencing. The information obtained through genome sequencing promotes the discovery of new genes associated with diseases and gene modifiers that helps to answer complicated genetic inheritance questions (van El et al., 2013). Through this powerful tool, the genetic cause of many diseases can be discovered with only one test, which means it may become the most preferred genetic test in the future. WGS can detect several categories of genetic variation, including single-nucleotide variations (SNVs), insertions and deletions (indels), copy number variations (CNVs) and translocations (TLs) (Vinkšel et al., 2021). The potential benefits are unfortunately limited by the genome's inaccessibility, cost, and complexity, as well as the current limitations of bioinformatics for interpreting non-coding genomic variants (Ormond et al., 2010). The WGS and WES methodologies have great potential for diagnosing rare diseases. They can analyze multiple genes in a

10.3389/fgene.2023.1258083

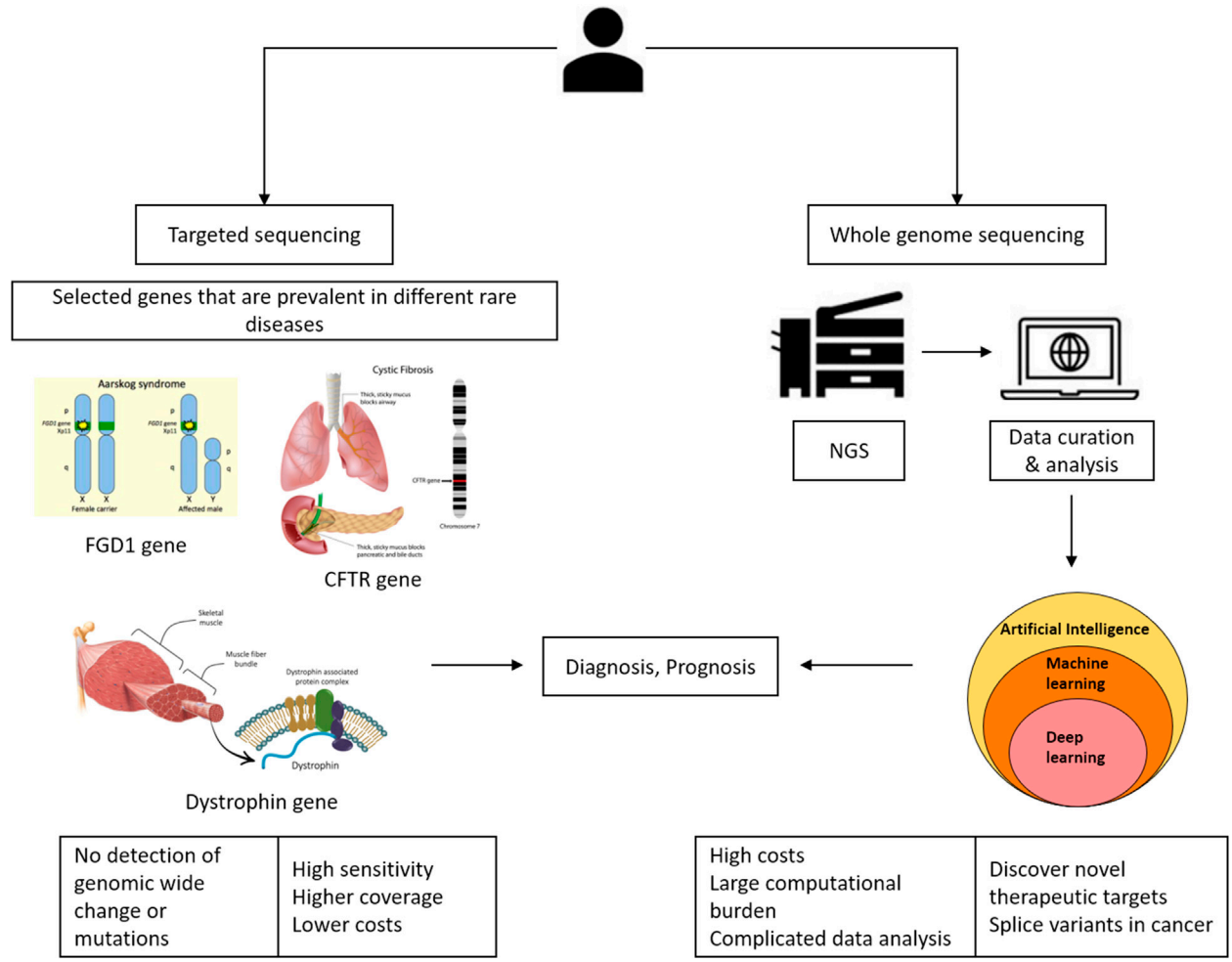

FIGURE 1
The use of sequencing in precision medicine.

single test while producing variants of unknown significance (VUS) and incidental findings. Hence, they pose additional challenges to clinicians and patients (Vinkšel et al., 2021).

2 NGS-based genetic diagnosis: challenges and opportunities

NGS offers several advantages in the clinical setting for elucidating predictive or prognostic biomarkers. NGS has advanced significantly over the last decade, with considerable improvements in throughput, quality, cost, and sequencing time. State-of-the-art algorithms, along with their capacity to process vast and intricate datasets, present novel possibilities for precision medicine treatments. As depicted in Figure 1, sequencing plays a significant role in precision medicine. At present, targeted sequencing stands as the preferred approach for clinical applications due to its advantages, such as increased sensitivity, broader coverage, and cost-effectiveness. However, it has limitations, such as the inability to identify significant genomic rearrangements or potentially pathogenic mutations in non-targeted genes. The benefit of whole-genome sequencing is that it allows for mutations and alterations throughout the genome (Huang et al., 2019).

3 Artificial intelligence for enhancing NGS-based diagnosis

Ng et al. (2009) first demonstrated the use of NGS-based methods in RDs as a proof-of-concept that WES could identify candidate genes responsible for monogenic disorders like Freeman-Sheldon syndrome (FSS). Comparing their results to WGS, the group demonstrated high concordance, low false discovery rate, and equivalent sensitivity for cSNP detection of WES. In research related to rare diseases (RDs), WES has become the preferred method due to its cost-effectiveness and efficiency in collecting and analyzing genomic data compared to WGS and its superior ability to detect novel disease-causing genes than targeting sequencing. As the number of genes that NGS can sequence

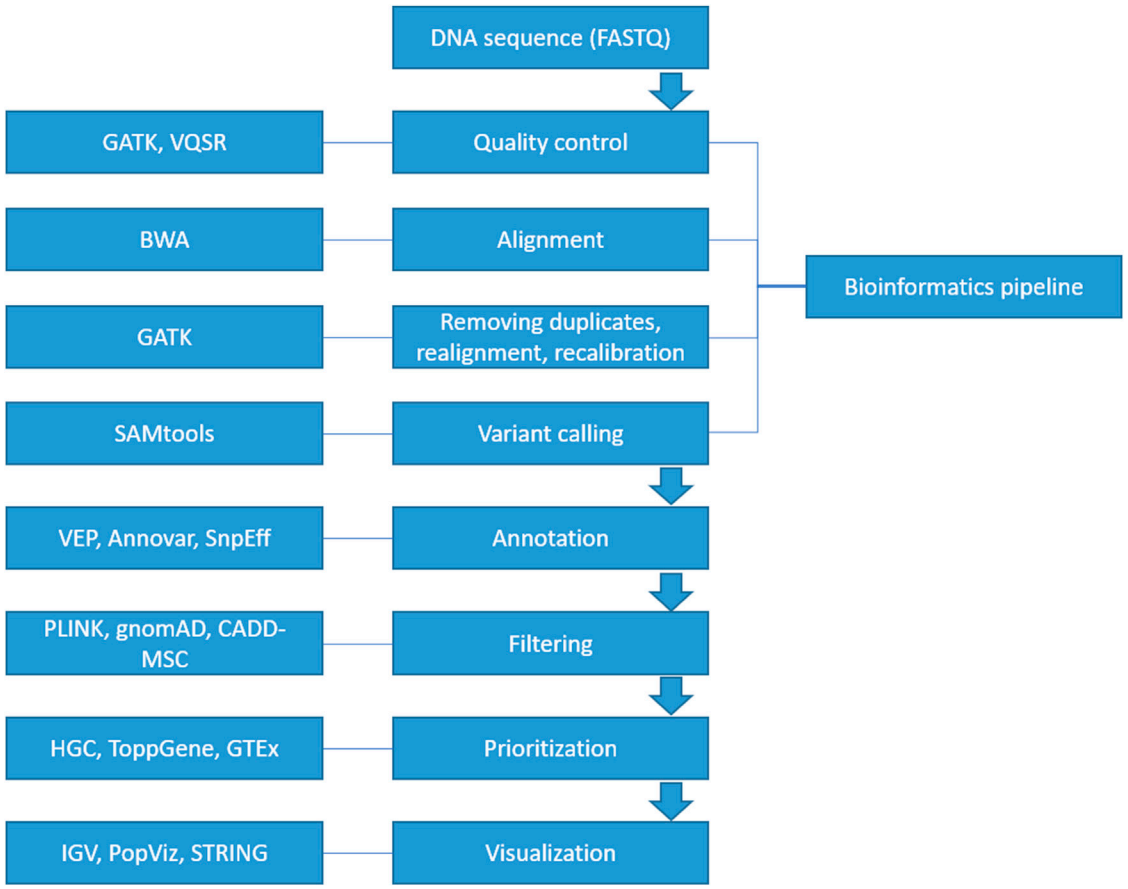

FIGURE 2
A framework of WES/WGS data analysis from individual patients with rare diseases. (GATK-Genome Analysis Toolkit, VQSR-Variant Quality Score Recalibration, BWA-Burrows-Wheeler Alignment, SAMtools-Sequence Alignment/Map tools, VEP-Ensembl Variant Effect Predictor, HGC-Hierarchical Graph-based Clustering, GTEx-Genotype Tissue Expression, IGV-Integrative Genomics Viewer).

increases, more candidate genes will likely be found. One of the challenges faced by the increasing number of associated RDs genes is the bioinformatic tools currently used in the alignment, variant calling, and annotation of NGS-generated genomic data. The use of various software packages will yield distinct final interpretations, different statistical significance thresholds, and variant calling, ultimately resulting in a diverse final list of potential genes (Fernandez-Marmiesse et al., 2018).

A suite of computational software is currently available for each step in identifying a diseasing-causing mutation in patients' genomes. The use of bioinformatics in NGS-based genetic testing is essential. There are five key stages in the NGS bioinformatics pipeline that must be completed before suitable analyses can be performed. Figure 2 illustrates a framework of WES/WGS data analysis from individual patients with rare diseases, while Figure 3 illustrates the workflow for NGS data analysis. Recently, GIAB, together with the Global Alliance for Genomics and Health (GA4GH), has been actively creating benchmarking data to set a standard reference for adopting the most effective methods for NGS data analysis (Krusche et al., 2019; Zook et al., 2019).

Artificial intelligence (AI) has a worldwide and interdisciplinary influence. Today, AI, particularly deep learning, is widely used in various biological contexts, changing the healthcare system and other disciplines outside the scope of this paper. AI has significantly contributed to the analysis of next-generation sequencing (NGS) data. AI algorithms play a crucial role in automating and enhancing various facets of NGS data analysis, thereby increasing efficiency and precision. One prominent application of AI in NGS data analysis involves the alignment of sequences to a known reference genome. Alignment, which entails matching NGS-generated sequences to a reference genome, is a critical step in detecting genome variations and mutations. AI algorithms excel at streamlining this process by identifying the most suitable matching sequences and compensating for data errors or variations. AI also plays important role in the development of novel NGS data analysis tools and methodologies. For instance, AI can be harnessed to create algorithms capable of predicting the performance of various NGS assays or to discover innovative approaches to NGS data analysis that enhance accuracy and efficiency. The substantial role of AI in NGS data analysis lies in its capacity to automate and optimize numerous aspects of the process, ultimately rendering it more efficient and precise. The ability of AI algorithms to swiftly and accurately process vast quantities of data positions them as indispensable tools in the field of NGS data analysis.

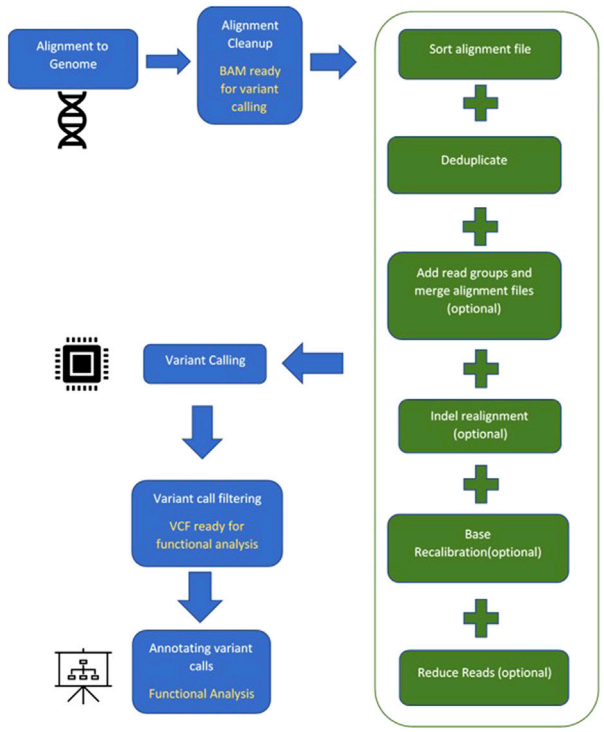

FIGURE 3
General workflow for NGS data analysis.

Machine learning, a subfield of artificial intelligence (AI) and computer science, revolves around leveraging data and algorithms to emulate human learning and continuously enhance its accuracy. This technology holds the potential to revolutionize disease identification and treatment, significantly impacting clinical decision-making. As genomic data grows exponentially, conventional statistical sampling-based approaches face difficulties in identifying genetic patterns. This is where advanced algorithms like deep learning and AI become highly advantageous. By utilizing deep neural networks as an end-to-end method, complex feature patterns can be automatically extracted, and prediction models can be built with minimal manual feature engineering. Table 1 shows the advantages and disadvantages of clinical NGS analysis. Table 2 summarises the recent studies that use machine learing algorithms in NGS data analysis.

3.1 Variant calling

The task of detecting variants from sequencing data is referred to as variant calling. Despite the existence of several variant calling algorithms, many of them still require improvement, especially in clinical settings. Machine learning-based algorithms offer an alternative approach for variant calling.

3.2 Variant prediction

The clinical implementation of NGS-based diagnosis faces a hurdle in distinguishing pathogenic mutations from benign genetic variations. Despite the creation of various variation effect prediction tools to bridge this gap, it still constitutes a limiting factor that necessitates further validation in the decision-making process (Xu et al., 2019).

3.3 EHR

Connecting genetic testing to EHR systems is essential to integrating genomics into clinical practice (Abul-Husn and Kenny, 2019). Meanwhile, the electronic health record (EHR) system has served as a centralized platform for integrating diverse digital health data, leading to improved clinical decision-making and precision medicine. The difficulty lies in integrating data profiles of different complexities within the EHR system to

TABLE 1 Advantages and disadvantages of clinical NGS analysis.

Type analysis	Advantages	Disadvantages
Variant calling	- Essential for identifying genetic variants associated with diseases	- Can be complex, involving the analysis of large volumes of data generated by next-generation sequencing technologies
	- A valuable tool for studying population genetics	
Variant Filtering	- Allows researchers to focus on the most relevant and high-confidence variants	- Risk of Excluding True Positives
	- Can reduce the number of false-positive variants	- May inadvertently filter out variants of interest, leading to potential data loss
	- Making the subsequent steps of analysis faster and more manageable	
Variant Annotation and Prioritization	- Provides detailed information about the functional consequences of variants	- Variant annotation and prioritization can be complex
	- Helps researchers or clinicians focus on the most biologically relevant variants	- Require substantial computational resources
Phenotype-genotype association	- Can capture data from all over the genome, providing a comprehensive view of genetic variations	- Require large sample sizes for robust associations
	- Enables the detection of rare and novel variants	

TABLE 2 Summarises the recent studies that use machine learning algorithms in NGS data analysis.

Models	Algorithms	Notes	Refs
Variant Calling			
Deep Variant	Deep convolutional neural network (CNN)	The process of variant calling through short-read sequencing involves creating a representation of DNA alignments in the form of an image	Poplin et al. (2018)
Clairvoyante	Deep Convolutional neural network (CNN)	A CNN model with multitasking capabilities and adaptable for long sequencing data	Luo et al. (2018)
DeepNano	Deep recurrent neural network (RNN)	An advanced RNN for conduct base calling on MinION nanopore reads, yielding outcomes comparable to the performance achieved by Oxford Nanopore's Nanonet base caller	Boža et al. (2017)
N/A	Logistic regression model	The creation of a deterministic machine-learning-based model aimed at distinguishing between two types of variant calls	van den Akker et al. (2018)
NeoMutate	Bayesian classifier, Bayesian model of admixture Heuristic methodology	Use of seven supervised machine learning algorithms, leveraging the advantages of various variant callers and integrating a unique collection of biological and sequence characteristics	Anzar et al. (2019)
GATK HaplotypeCaller algorithm	N/A	A comprehensive pipeline designed to identify the optimal method for processing NGS data to accurately call variants for subsequent analyses with confidence	Pirooznia et al. (2014)
N/A	Multivariate linear regression Random forest regression Neural network regression	A machine learning method used to predict the quality scores of variant calls obtained from BWA + GATK	Cosgun and Oh (2020)
N/A	Random forests, adaptive boosting, k-nearest neighbors, naive Bayes, support vector machines	By combining multiple supervised machine learning techniques, the prediction of phenotype group associations significantly improves when relying on observed genotypes compared to using random permutations of the exomic sequences	Kringel et al. (2018)
Variant Filtering			
SNooPer	Random forest	A machine learning approach to call somatic variants in low-depth sequencing data	Spinella et al. (2016)
GARFIELD-NGS	N/A	A tool designed to distinguish between false and true variants in exome sequencing experiments	Ravasio et al. (2018)
Intelli-NGS	Deep neural network (DNN)	A tool based on deep neural networks that assists in minimizing false positive and false negative rates while maintaining high recall performance	Singh and Bhatia (2019)
DeepSVFilter	Convolutional neural network (CNN)	A deep learning-based approach for filtering structural variants in short genome sequencing data	Liu et al. (2021)
iEVA	N/A	A tool that amplifies informative features derived from NGS data and utilizes them in a filtering process employing a Machine Learning algorithm (ML)	Urtis et al. (2019)
DOMINO	linear discriminant analysis	A tool that evaluates the probability of a gene containing dominant alterations	Quinodoz et al. (2017)
Variant Annotation and Prioritization			
Skyhawk	Deep neural network (DNN)	An artificial neural networks that imitators the expert review process to detect clinically relevant genomic variants	Luo et al. (2018)
DANN	Deep neural network (DNN)	A DNN algorithm that surpasses state-of-the-art methods like support vector machine in predicting the deleterious annotation of genetic variants	Quang et al. (2015)
DeepSEA	Deep Convolutional neural network (CNN)	A deep CNN model utilized to predict the effects of noncoding variants directly from the sequence data and subsequently applied to forecast the functional impact of variants related to autism spectrum disorder	Zhou and Troyanskaya (2015)
eDiva	N/A	The framework integrates NGS data analysis, via functional annotation, and optimized causal variant prioritization	Bosio et al. (2019)

(Continued on following page)

TABLE 2 (*Continued*) Summarises the recent studies that use machine learning algorithms in NGS data analysis.

Models	Algorithms	Notes	Refs
RENOVO	Random forest	An algorithm for reclassification of germline variants of unknown significance	Favalli et al. (2021)
Phenotype-genotype association			
DeepGestalt	Deep Convolutional neural network (CNN)	A sophisticated convolutional neural network model can distinguish rare diseases by analyzing patient face images and it can effectively discriminate different genetic subtypes	Gurovich et al. (2019)
DeepPVP	Deep neural network (DNN)	A Deep Neural Network model used for prioritizing variants by incorporating patients' phenotype information	Boudellioua et al. (2019)
Xrare	N/A	A method to prioritize causative gene variants in the diagnosis of rare diseases	Li et al. (2019)
SQUIRLS	Random forest	An algorithm in classifying splice variants	Danis et al. (2021)

N/A represents that the information is not reported in the paper.

enhance clinical diagnosis. AI advancements offer a potential solution to this challenge.

3.4 Phenotypes and genetic testing association

The main objective of a genetic association study is to investigate whether a particular sequence, such as a chromosomal region, haplotype, gene, or allele, plays a role in determining specific traits, metabolic pathways, or diseases. Deep learning has been widely used to improve diagnosis performance in medical image diagnostic systems, outperforming radiologists and pathologists (Yu et al., 2018). For example, DeepGestalt proposed by Gurovich et al. (2019) included over 17,000 pictures for over 200 rare diseases and reached 91% accuracy.

4 Databases for rare diseases

AI and NGS complement each other exceptionally well since AI thrives on extensive data while NGS generates vast amounts of data. Alongside the massive NGS data, other diagnosis-related testing data is also being produced, presenting the challenge of adequate data storage. To securely manage this data, a sophisticated informatics infrastructure is necessary. Measures have been taken to ensure that cloud-based services adhere to health privacy regulations, allowing for the secure storage of NGS data and the establishment of standardized data privacy practices among various stakeholders (Langmead and Nellore, 2018).

Although AI holds promise for improving clinical diagnosis in rare diseases, its effectiveness can be hindered by the intricate and diverse profiles of clinical data. Constructing an AI model for diagnosing rare diseases requires a substantial training dataset comprising patients with documented clinical outcomes. This paper reviews a few currently available databases for rare disease diagnosis. Table 3 summarises the available databases for rare diseases. Table 4 shows comparison between available databases for rare diseases.

4.1 National organization for rare disorders (NORD) rare disease database

Since its inception in the early 1980s, coinciding with the implementation of the Orphan Drug Act, the National Organization for Rare Disorders (NORD) has been functioning as a support and advocacy organization for those individuals impacted by rare diseases. The database subscribers are granted entry to extensive monographs containing detailed information about the causes, symptoms, standard and investigational treatments, as well as support organizations related to various rare diseases. The level of detail offered in these monographs exceeds that of other resources, making it highly valued by patients and their families.

The Rare Diseases Database presently comprises data on over 1,200 diseases, Organized in alphabetical sequence or capable of being searched by disease name or synonym. It is important to note that NORD clarifies this database is not exhaustive, given that there are nearly 7,000 acknowledged rare diseases. As a non-profit advocacy organization, NORD's resources for this informational database are limited, and it chooses to rely on volunteer specialists to contribute material.

4.2 NIH genetic and rare diseases information center (GARD)

The NORD Rare Diseases Database has a limited scope, so the website provides links to additional resources, especially the NIH Genetic and Rare Diseases (GARD) Information Center. The main objective of GARD is to provide up-to-date, precise, and easily understandable information regarding rare or genetic diseases in both English and Spanish. The GARD Information Center database contains approximately 6,700 specific diseases, and the data is generated by "information experts" with genetics degrees, according to the GARD Operations Manager (Hogan Smith, 2017).

Some information on the listed diseases is sourced from external databases like Orphanet, a European rare disease database. While GARD covers more rare diseases than the NORD Database, some entries require additional information.

TABLE 3 Summarises the available databases.

Database	Description	URL	Reference
NORD	An organization that support individuals affected by rare diseases and the entities that offer them assistance	https://rarediseases.org/for-patients-and-families/information-resources/rare-disease-information/	NORD (2021)
GARD	Offers the general public reliable, up-to-date, and user-friendly information about rare or genetic diseases	https://rarediseases.info.nih.gov/diseases	GARD (2021)
Orphanet	The goal is to gather limited information about rare diseases in order to improve the diagnosis, care, and treatment of patients afflicted by these conditions	https://www.orpha.net/consor/cgi-bin/Disease.php?lng=EN	Orphanet (2021)
OMIM	An ever-evolving repository of human genes, genetic disorders, and traits, with investigating the relationship between genes and phenotypes	https://www.omim.org/	Amberger et al. (2015)
LORIS MyeliNeuroGene	Natural history studies and clinical trial readiness	https://myelineurogene-stg.loris.ca/	Spahr et al. (2021)

Note: N/A represents unavailable information.

TABLE 4 Comparison of available databases for rare disease.

Database	Services	Advantage	Disadvantage
NORD	Offers detailed information on rare diseases, patient advocacy, support groups, and patient assistance programs	- Highly patient-centered and offers extensive support, advocacy, and information for individuals and families affected by rare diseases	- The focus is mainly on rare diseases, and it may not be as comprehensive in terms of genetic and molecular information
GARD	Offers information specialists for personalized assistance, educational materials, and government-funded resources	- Freely accessible to the public - Comprehensive information on genetic and rare diseases, including disease descriptions, research, clinical trials, and expert guidance	- While it provides extensive information, it may not have the same level of patient support and advocacy as NORD
Orphanet	Offers information for both healthcare professionals and the general public	- Provides information on rare diseases, orphan drugs, expert centers, and research projects	- Primary focus is on Europe, and some information may be less relevant for non-European users
OMIM	Offers extensive genetic and molecular information, including genetic mutations and associated clinical features	- Specializes in the genetic and molecular basis of human diseases and disorders - Freely accessible to the public	- Primarily focuses on monogenic disorders and may not provide comprehensive information on complex genetic traits or disorders influenced by multiple genes and environmental factors
LORIS MyeliNeuroGene	Offers information on rare neurological conditions, clinical trials, and genetic research	- Focuses on rare neurological diseases and disorders, particularly those affecting the central nervous system	- Limited to rare neurological diseases, so it may not be relevant for individuals seeking information on other types of rare diseases - Funding sources may not be as transparent as those of larger, more established resources

GARD also allows users to ask questions to a GARD information professional. The responses are akin to a librarian's helpful response to consumer health information queries, often pointing to general material available on the site rather than addressing individual users' specific circumstances. Since its establishment in February 2002, GARD has answered over 22,000 requests about 6,000 rare and genetic diseases, as reported by the NIH.

4.3 Orphanet

Orphanet is a European platform dedicated to rare diseases and orphan drugs, led by the Institut National de la Santé et de la Recherche Médicale (INSERM) in collaboration with various countries and organizations, primarily within the European Union. The main objective of Orphanet is to provide high-quality information about rare diseases and ensure that all stakeholders have equitable access to knowledge. The platform also publishes a series of widely downloaded publications that present aggregated data on topics relevant to all rare diseases.

The inventory of rare diseases on Orphanet can be searched using disease names, gene names, symbols, or the disease's "functional consequences" (disabilities), as well as other identifying numbers like the Online Mendelian Inheritance in Man (OMIM) number. A beta tool called PhenomizerOrphanet is also available to assist in clinical differential diagnosis through controlled vocabulary searches. Orphanet offers an "Encyclopedia for Patients," an "Encyclopaedia for Professionals," and "Emergency Guidelines" for healthcare professionals. However, it should be noted that the quantity of diseases addressed in the articles within the Encyclopedias. Is generally limited. The site's content is accessible in multiple European languages and includes information on 6,172 diseases and 5,835 genes (Orphanet, 2021).

As stated on the website, all disease entries are written by specialists and undergo evaluation by peers. However, it's important to acknowledge that the mentioned therapies may not

be evidence-based due to the limited number of cases available for gathering evidence for or against a particular treatment.

4.4 Online Mendelian Inheritance in Man (OMIM)

Online Mendelian Inheritance in Man (OMIM) is an authoritative and freely accessible database containing comprehensive information about human genes and genetic traits, which is updated on a daily basis. The comprehensive summaries in OMIM include information about all identified Mendelian diseases and over 16,000 genes. The database focuses on establishing the connection between phenotype and genotype, and its articles are regularly updated, providing numerous links to additional genetics resources.

In the early 1960s, Dr. Victor A. McKusick launched the database known as Mendelian Inheritance in Man (MIM), originally intended as a catalog of Mendelian traits and disorders. This catalog was published in twelve book versions from 1966 to 1998. Subsequently, in 1985, an online version called OMIM was developed through a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It became widely available on the Internet in 1987. Subsequently, in 1995, the National Center for Biotechnology Information (NCBI) created the World Wide Web version of OMIM. Dr. Ada Hamosh leads the McKusick-Nathans Institute of Genetic Medicine at Johns Hopkins University School of Medicine, where OMIM is authored and edited (OMIM, 2021).

Unlike primary data databases, OMIM aggregates and summarizes essential information derived from expert reviews of the biomedical literature. Consequently, OMIM has played a pioneering role in naming and classifying genetic phenotypes (Amberger et al., 2015). A simple search in the OMIM database reveals numerous genes associated with various diseases, some of which exhibit multiple inheritance patterns.

4.5 LORIS MyeliNeuroGene rare disease database

In 2021, Spahr et al. (2021) introduced the LORIS MyeliNeuroGene rare disease database for conducting natural history studies and preparing for clinical trials. This online database for rare disease and needs subscription, it is not free access like OMIM or orphanet or GARD. Employing FDA-compliant databases for developing clinical trials with historical control data could significantly impact patients and families.

Spahr et al. (2021) created an accessible multi-modal database accessible via a web browser, which included genetics, imaging, behavioral, and patient-reported outcomes. The main goals were to increase the size of cohorts, identify surrogate markers, and foster international collaborations. The database contained a comprehensive range of information, such as family, perinatal, and developmental history, clinical examinations, diagnostic investigations, neurological evaluations (e.g., spasticity, dystonia, ataxia, etc.), disability measures, parental stress, and quality of life data.

Spahr et al. (2021) highlighted that their manuscript is the first to outline the requirements for adhering to Title 21 Code of Federal Regulations Part 11 Compliance. Subsequent studies will employ the tools developed in this project to characterize the natural progression of diverse rare diseases, with the goal of providing valuable insights to clinicians and researchers globally.

In summary, the choice of resource depends on specific research needs and interests. Each of these databases serves a unique purpose. NORD and GARD are more patient-focused, while Orphanet provides comprehensive European coverage. OMIM offers specialized genetic information for professionals, and LORIS MyeliNeuroGene is niche-focused on neurological diseases.

5 Conclusion and future perspectives

Genetic testing is becoming increasingly popular and accessible for both individuals and clinicians in today's world. While challenges and obstacles persist, NGS technologies hold significant promise as the initial stage in genetic testing for rare disease diagnoses.

This paper focuses solely on certain aspects of NGS-based genetic testing in clinical implementation and omits other vital factors. These include genetic counseling to improve the patient-physician relationship, addressing ethnic considerations in the adoption and delivery of genetic testing, and educational initiatives aimed at promoting the acceptance of genetic testing in clinical settings.

The challenge of data interpretation remains a significant obstacle when employing routine clinical NGS for diagnosis. Dealing with large datasets and interpreting them requires substantial resources and expertise from bioinformaticians. These datasets contain information on variations that need to be classified for accurate diagnosis. Although AI shows great potential in healthcare, it faces challenges, including the increasing data volume and associated costs from automated computing. AI systems demand specialized computational resources for swift data processing, making them expensive. Additionally, AI-based solutions require proper training and understanding by intended users before being integrated into routine clinical practice. Addressing ethical concerns regarding patient data use is critical, necessitating ethical standards and procedures to ensure patient safety and privacy.

AI is beginning to tap into its potential to enhance clinical usefulness and diagnostic capabilities by supplementing phenome-wide and genome-wide data profiles. iBoth government agencies and professional communities are actively supporting and initiating efforts to standardize regulations for NGS-based testing and AI applications. When dealing with rare diseases, further research is needed as traditional monogenic models may not be sufficient. Exploring the digenic/oligogenic model and investigating polygenic causes for undiagnosed cases could provide valuable insights (Katsanis et al., 2001; Hoefele et al., 2007; Boisson-Dupuis et al., 2018; Posey, 2019).

Author contributions

YWC: Writing–original draft. YFC: Validation, Writing–original draft. NN: Writing–review and editing. FA:

Conceptualization, Validation and editing. MR: Conceptualization, Writing–review and editing. MA: Investigation, Writing–review and editing. MM: Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was sponsored by the ASPIRE, the technology program management pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), via the ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) award grant number VRI-20-10. The United Arab Emirates University also supported this work through the Research Start-up Program (Grant # 12M109).

Conflict of interest

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

Publisher's note

References

Abul-Husn, N. S., and Kenny, E. E. (2019). Personalized medicine and the power of electronic health records. *Cell* 177 (1), 58–69. doi:10.1016/j.cell.2019.02.039

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids Res.* 43 (D1), D789–D798. doi:10.1093/nar/gku1205

Amorim, B. R., Santos, P. A. C. D., Lima, C. L. D., Andia, D. C., Mazzeu, J. F., and Acevedo, A. C. (2019). "Protocols for genetic and epigenetic studies of rare diseases affecting dental tissues," in *Odontogenesis* (New York, NY: Humana Press), 453–492.

Anzar, I., Sverchkova, A., Stratford, R., and Clancy, T. (2019). NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. genomics* 12 (1), 63–14. doi:10.1186/s12920-019-0508-5

Austin, C. P., Cutillo, C. M., Lau, L. P., Jonker, A. H., Rath, A., Julkowska, D., et al. (2018). Future of rare diseases research 2017–2027: an IRDiRC perspective. *Clin. Transl. Sci.* 11 (1), 21–27. doi:10.1111/cts.12500

Baynam, G. S., Groft, S., van der Westhuizen, F. H., Gassman, S. D., du Plessis, K., Coles, E. P., et al. (2020). A call for global action for rare diseases in Africa. *Nat. Genet.* 52 (1), 21–26. doi:10.1038/s41588-019-0552-2

Boisson-Dupuis, S., Ramirez-Alejo, N., Li, Z., Patin, E., Rao, G., Kerner, G., et al. (2018). Tuberculosis and impaired IL-23–dependent IFN-γ immunity in humans homozygous for a common TYK2 missense variant. *Sci. Immunol.* 3 (30), eaau8714. doi:10.1126/sciimmunol.aau8714

Bosio, M., Drechsel, O., Rahman, R., Muyas, F., Rabionet, R., Bezdan, D., et al. (2019). eDiVA—classification and prioritization of pathogenic variants for clinical diagnostics. *Hum. Mutat.* 40 (7), 865–878. doi:10.1002/humu.23772

Boudellioua, I., Kulmanov, M., Schofield, P. N., Gkoutos, G. V., and Hoehndorf, R. (2019). DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinforma.* 20 (1), 65–68. doi:10.1186/s12859-019-2633-8

Boža, V., Brejová, B., and Vinař, T. (2017). DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PloS one* 12 (6), e0178751. doi:10.1371/journal.pone.0178751

Cai, Y., Huang, T., and Jia, P. (2020). Editorial: advanced interpretable machine learning methods for clinical NGS big data of complex hereditary diseases. *Front. Genet.* 11, 600902. doi:10.3389/fgene.2020.600902

Cosgun, E., and Oh, M. (2020). Exploring the consistency of the quality scores with machine learning for next-generation sequencing experiments. *BioMed Res. Int.* 2020, 8531502. doi:10.1155/2020/8531502

Danis, D., Jacobsen, J. O., Carmody, L., Gargano, M. A., McMurry, J. A., Hegde, A., et al. (2021). Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *bioRxiv* 108, 2205. doi:10.1016/j.ajhg.2021.09.014

Elliott, E. J., and Zurynski, Y. A. (2015). Rare diseases are a'common'problem for clinicians. *Aust. Fam. physician* 44 (9), 630–633.

Favalli, V., Tini, G., Bonetti, E., Vozza, G., Guida, A., Gandini, S., et al. (2021). Machine learning-based reclassification of germline variants of unknown significance: the RENOVO algorithm. *Am. J. Hum. Genet.* 108 (4), 682–695. doi:10.1016/j.ajhg.2021.03.010

Fernandez-Marmiesse, A., Gouveia, S., and Couce, M. L. (2018). NGS technologies as a turning point in rare disease research, diagnosis and treatment. *Curr. Med. Chem.* 25 (3), 404–432. doi:10.2174/0929867324666170718101946

Field, M. A. (2021). Detecting pathogenic variants in autoimmune diseases using high-throughput sequencing. *Immunol. Cell Biol.* 99 (2), 146–156. doi:10.1111/imcb.12372

Frésard, L., and Montgomery, S. B. (2018). Diagnosing rare diseases after the exome. *Mol. Case Stud.* 4 (6), a003392. doi:10.1101/mcs.a003392

Gallego-Martinez, A., Requena, T., Roman-Naranjo, P., and Lopez-Escamez, J. A. (2019). Excess of rare missense variants in hearing loss genes in sporadic Meniere disease. *Front. Genet.* 10, 76. doi:10.3389/fgene.2019.00076

Girotto, G., Morgan, A., Krishnamoorthy, N., Cocca, M., Brumat, M., Bassani, S., et al. (2019). Next generation sequencing and animal models reveal SLC9A3R1 as a new gene involved in human age-related hearing loss. *Front. Genet.* 10, 142. doi:10.3389/fgene.2019.00142

Grosse, S. D., Kalman, L., and Khoury, M. J. (2010). Evaluation of the validity and utility of genetic testing for rare diseases. *Rare Dis. Epidemiol.* 686, 115–131. doi:10.1007/978-90-481-9485-8_8

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., et al. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* 25 (1), 60–64. doi:10.1038/s41591-018-0279-0

Hartley, T., Lemire, G., Kernohan, K. D., Howley, H. E., Adams, D. R., and Boycott, K. M. (2020). New diagnostic approaches for undiagnosed rare genetic diseases. *Annu. Rev. genomics Hum. Genet.* 21, 351–372. doi:10.1146/annurev-genom-083118-015345

Hoefele, J., Wolf, M. T., O'Toole, J. F., Otto, E. A., Schultheiss, U., Dêschenes, G., et al. (2007). Evidence of oligogenic inheritance in nephronophthisis. *J. Am. Soc. Nephrol.* 18 (10), 2789–2795. doi:10.1681/ASN.2007020243

Hogan Smith, K. (2017). Review of rare diseases resources: national organization for rare disorders (NORD) rare disease database, NIH genetic and rare diseases information center, and Orphanet. *J. Consumer Health Internet* 21 (2), 216–225. doi:10.1080/15398285.2017.1311613

Huang, J., Qian, Z., Gong, Y., Wang, Y., Guan, Y., Han, Y., et al. (2019). Comprehensive genomic variation profiling of cervical intraepithelial neoplasia and cervical cancer identifies potential targets for cervical cancer early warning. *J. Med. Genet.* 56 (3), 186–194. doi:10.1136/jmedgenet-2018-105745

Jia, J., and Shi, T. (2017). Towards efficiency in rare disease research: what is distinctive and important? *Sci. China Life Sci.* 60 (7), 686–691. doi:10.1007/s11427-017-9099-3

Katsanis, N., Ansley, S. J., Badano, J. L., Eichers, E. R., Lewis, R. A., Hoskins, B. E., et al. (2001). Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* 293 (5538), 2256–2259. doi:10.1126/science.1063525

Khosla, N., and Valdez, R. (2018). A compilation of national plans, policies and government actions for rare diseases in 23 countries. *Intractable rare Dis. Res.* 7 (4), 213–222. doi:10.5582/irdr.2018.01085

Kringel, D., Geisslinger, G., Resch, E., Oertel, B. G., Thrun, M. C., Heinemann, S., et al. (2018). Machine-learned analysis of the association of next-generation sequencing–based human TRPV1 and TRPA1 genotypes with the sensitivity to heat stimuli and topically applied capsaicin. *Pain* 159 (7), 1366–1381. doi:10.1097/j.pain.0000000000001222

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* 37, 555–560. doi:10.1038/s41587-019-0054-x

Langmead, B., and Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* 19 (4), 208–219. doi:10.1038/nrg.2017.113

Li, Q., Zhao, K., Bustamante, C. D., Ma, X., and Wong, W. H. (2019). Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet. Med.* 21 (9), 2126–2134. doi:10.1038/s41436-019-0439-8

Liu, Y., Huang, Y., Wang, G., and Wang, Y. (2021). A deep learning approach for filtering structural variants in short read sequencing data. *Briefings Bioinforma.* 22 (4), bbaa370. doi:10.1093/bib/bbaa370

Liu, Z., Zhang, L., Shen, D., Ding, C., Yang, X., Zhang, W., et al. (2019a). Compound heterozygous CHAT gene mutations of a large deletion and a missense variant in a Chinese patient with severe Congenital Myasthenic Syndrome with Episodic Apnea. *Front. Pharmacol.* 10, 259. doi:10.3389/fphar.2019.00259

Liu, Z., Zhu, L., Roberts, R., and Tong, W. (2019b). Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: where are we? *Trends Genet.* 35 (11), 852–867. doi:10.1016/j.tig.2019.08.006

Luo, R., Sedlazeck, F. J., Lam, T. W., and Schatz, M. C. (2018). Clairvoyante: a multi-task convolutional deep neural network for variant calling in single molecule sequencing. *bioRxiv*, 310458.

Maroilley, T., and Tarailo-Graovac, M. (2019). Uncovering missing heritability in rare diseases. *Genes* 10 (4), 275. doi:10.3390/genes10040275

Miller, E. M., Patterson, N. E., Zechmeister, J. M., Bejerano-Sagie, M., Delio, M., Patel, K., et al. (2017). Development and validation of a targeted next generation DNA sequencing panel outperforming whole exome sequencing for the identification of clinically relevant genetic variants. *Oncotarget* 8 (60), 102033–102045. doi:10.18632/oncotarget.22116

National Organization for Rare Disorders (NORD) Rare Disease Database (2021). Rarediseases. Available at: https://rarediseases.org/(Accessed September 13, 2021).

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461 (7261), 272–276. doi:10.1038/nature08250

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., et al. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* 28, 165–173. doi:10.1038/s41431-019-0508-0

Nguyen, K. V. (2019). Potential epigenomic co-management in rare diseases and epigenetic therapy. *Nucleosides, Nucleotides Nucleic Acids* 38 (10), 752–780. doi:10.1080/15257770.2019.1594893

NIH (United States National Institutes of Health), National human genome research Institute. 2015. "NIH's genetic and rare diseases information center offers new, web-based search feature." Accessed September 13, 2021. Available at: https://www.genome.gov/10506225/.

OMIM - Online Mendelian Inheritance in Man (2021). Omim. Available at: https://www.omim.org (Accessed September 13, 2021).

Ormond, K. E., Wheeler, M. T., Hudgins, L., Klein, T. E., Butte, A. J., Altman, R. B., et al. (2010). Challenges in the clinical application of whole-genome sequencing. *Lancet* 375 (9727), 1749–1751. doi:10.1016/S0140-6736(10)60599-5

Orphanet (2021). Orphanet in numbers. Available at: http://www.orpha.net/consor/cgi-bin/index.php?lng=EN (Accessed September 9, 2021).

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., et al. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. genomics* 8 (1), 14–10. doi:10.1186/1479-7364-8-14

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

Posey, J. E. (2019). Genome sequencing and implications for rare disorders. *Orphanet J. rare Dis.* 14 (1), 153–210. doi:10.1186/s13023-019-1127-0

Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31 (5), 761–763. doi:10.1093/bioinformatics/btu703

Quinodoz, M., Royer-Bertrand, B., Cisarova, K., Di Gioia, S. A., Superti-Furga, A., and Rivolta, C. (2017). DOMINO: using machine learning to predict genes associated with dominant disorders. *Am. J. Hum. Genet.* 101 (4), 623–629. doi:10.1016/j.ajhg.2017.09.001

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* 59 (1), 5–15. doi:10.1038/jhg.2013.114

Ravasio, V., Ritelli, M., Legati, A., and Giacopuzzi, E. (2018). Garfield-ngs: genomic variants filtering by deep learning models in NGS. *Bioinformatics* 34 (17), 3038–3040. doi:10.1093/bioinformatics/bty303

Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* 14 (4), 295–300. doi:10.1038/nrg3463

Rey, T., Tarabeux, J., Gerard, B., Delbarre, M., Béchec, A. L., Stoetzel, C., et al. (2019). "Protocol GenoDENT: implementation of a new NGS panel for molecular diagnosis of genetic disorders with orodental involvement," in *Odontogenesis* (New York, NY: Humana Press), 407–452.

Singh, A., and Bhatia, P. (2019). Intelli-NGS: intelligent NGS, a deep neural network-based artificial intelligence to delineate good and bad variant calls from IonTorrent sequencer data. *bioRxiv*.

Soon, W. W., Hariharan, M., and Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Mol. Syst. Biol.* 9 (1), 640. doi:10.1038/msb.2012.61

Spahr, A., Rosli, Z., Legault, M., Tran, L. T., Fournier, S., Toutounchi, H., et al. (2021). The LORIS MyeliNeuroGene rare disease database for natural history studies and clinical trial readiness. *Orphanet J. Rare Dis.* 16 (1), 328–410. doi:10.1186/s13023-021-01953-8

Spinella, J. F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., et al. (2016). SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC genomics* 17 (1), 912–1011. doi:10.1186/s12864-016-3281-2

Stoller, J. K. (2018). The challenge of rare diseases. *Chest* 153 (6), 1309–1314. doi:10.1016/j.chest.2017.12.018

Taruscio, D., Vittozzi, L., and Stefanov, R. (2010). National plans and strategies on rare diseases in Europe. *Rare Dis. Epidemiol.* 686, 475–491. doi:10.1007/978-90-481-9485-8_26

Tatiana, M., and Tarailo- Graovac, M. (2019). Uncovering missing heritability in rare diseases. *GenesGenes 10* 4, 275.

Urtis, M., Smirnova, A., Di Toro, A., Giuliani, L., Pilotto, A., Di Giovannantonio, M., et al. (2019). P5723 IEVA: integration and extraction of variant attributes in NGS analysis. *Eur. Heart J.* 40 (Suppl. ment_1), ehz746–0663. doi:10.1093/eurheartj/ehz746.0663

van den Akker, J., Mishne, G., Zimmer, A. D., and Zhou, A. Y. (2018). A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. *BMC genomics* 19 (1), 263–269. doi:10.1186/s12864-018-4659-0

Van El, C. G., Cornel, M. C., Borry, P., Hastings, R. J., Fellmann, F., Hodgson, S. V., et al. (2013). Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. *Eur. J. Hum. Genet.* 21 (6), 580–584. doi:10.1038/ejhg.2013.46

Vinkšel, M., Writzl, K., Maver, A., and Peterlin, B. (2021). Improving diagnostics of rare genetic diseases with NGS approaches. *J. Community Genet.* 12 (2), 247–256. doi:10.1007/s12687-020-00500-5

Wright, C., FitzPatrick, D., and Firth, H. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* 19, 253–268. doi:10.1038/nrg.2017.116

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., et al. (2019). Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Hum. Genet.* 138, 109–124. doi:10.1007/s00439-019-01970-5

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* 369 (16), 1502–1511. doi:10.1056/NEJMoa1306555

Yu, K. H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nat. Biomed. Eng.* 2 (10), 719–731. doi:10.1038/s41551-018-0305-z

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. methods* 12 (10), 931–934. doi:10.1038/nmeth.3547

Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., et al. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* 37, 561–566. doi:10.1038/s41587-019-0074-6