frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED XX XX 2023
DOI 10.3389/fmolb.2023.1169109

# Resources and tools for rare disease variant interpretation

Luana Licata[1†‡], Allegra Via[2†‡], Paola Turina[3*‡], Giulia Babbi[3‡], Silvia Benevenuta[4‡], Claudio Carta[5‡], Rita Casadio[3‡], Andrea Cicconardi[6,7], Angelo Facchiano[8‡], Piero Fariselli[4‡], Deborah Giordano[8‡], Federica Isidori[9‡], Anna Marabotti[10‡], Pier Luigi Martelli[3‡], Stefano Pascarella[2‡], Michele Pinelli[11‡], Tommaso Pippucci[9‡], Roberta Russo[11,12‡], Castrense Savojardo[3‡], Bernardina Scafuri[10‡], Lucrezia Valeriani[13] and Emidio Capriotti[3‡]

[1]Department of Biology, University of Rome Tor Vergata, Roma, Italy, [2]Department of Biochemical Sciences "A. Rossi Fanelli", University of Rome "La Sapienza", Roma, Italy, [3]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [4]Department of Medical Sciences, University of Torino, Torino, Italy, [5]National Centre for Rare Diseases, Istituto Superiore di Sanità, Roma, Italy, [6]Department of Physics, University of Genova, Genova, Italy, [7]Italiano di Tecnologia—IIT, Genova, Italy, [8]National Research Council, Institute of Food Science, Avellino, Italy, [9]Medical Genetics Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy, [10]Department of Chemistry and Biology "A. Zambelli", University of Salerno, Fisciano, SA, Italy, [11]Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Napoli, Italy, [12]CEINGE Biotecnologie Avanzate Franco Salvatore, Napoli, Italy, [13]Center for Technology and Innovation, Trieste, Italy

Collectively, rare genetic disorders affect a substantial portion of the world's population. In most cases, those affected face difficulties in receiving a clinical diagnosis and genetic characterization. The understanding of the molecular mechanisms of these diseases and the development of therapeutic treatments for patients are also challenging. However, the application of recent advancements in genome sequencing/analysis technologies and computer-aided tools for predicting phenotype-genotype associations can bring significant benefits to this field. In this review, we highlight the most relevant online resources and computational tools for genome interpretation that can enhance the diagnosis, clinical management, and development of treatments for rare disorders. Our focus is on resources for interpreting single nucleotide variants. Additionally, we present use cases for interpreting genetic variants in clinical settings and review the limitations of these results and prediction tools. Finally, we have compiled a curated set of core resources and tools for analyzing rare disease genomes. Such resources and tools can be utilized to develop standardized protocols that will enhance the accuracy and effectiveness of rare disease diagnosis.

KEYWORDS

**rare disease, genetic disorder, single nucleotide variant (SNV), genome interpretation, precision medicine, genotype-phenotype association, machine learning**

OPEN ACCESS

EDITED BY
Barbara Cellini,
University of Perugia, Italy

REVIEWED BY
Zailong Qin,
Maternal and Child Health Hospital of Guangxi Zhuang Autonomous Region, China
Andrea Mozzarelli,
University of Parma, Italy

*CORRESPONDENCE
Paola Turina,
paola.turina@unibo.it

‡ELIXIR Italy Rare Disease Community

†These authors have contributed equally to this work and share first authorship

RECEIVED 18 February 2023
ACCEPTED 25 April 2023
PUBLISHED XX XX 2023

CITATION
Licata L, Via A, Turina P, Babbi G, Benevenuta S, Carta C, Casadio R, Cicconardi A, Facchiano A, Fariselli P, Giordano D, Isidori F, Marabotti A, Martelli PL, Pascarella S, Pinelli M, Pippucci T, Russo R, Savojardo C, Scafuri B, Valeriani L and Capriotti E (2023), Resources and tools for rare disease variant interpretation. *Front. Mol. Biosci.* 10:1169109. doi: 10.3389/fmolb.2023.1169109

## 1 Introduction

The recent major advances in genome sequencing and analysis technology have opened the road to exome and genome sequencing (ES/GS) as a common diagnostic tool for individual patients (Turro et al., 2020; 100,000 Genomes Project Pilot Investigators et al., 2021). Especially in the field of rare genetic diseases, the use of ES/GS has brought unprecedented progress, and holds the potential for further large-scale impact in the clinical setting, allowing early diagnosis and early, precisely tuned, treatment (Pogue et al., 2018; Liu et al., 2019; Claussnitzer et al., 2020; Bonne, 2021). Presently, the definition of a rare disease (RD) varies among different regions. In Europe, it is defined as a condition affecting not more than 1 person per 2,000 in the European population (Regulation Orphan Medicinal Product, 2000). In the United States, it is defined as a condition that affects less than 200,000 people in the country (U.S. Food and Drug Administration, 2022), while in Japan it is defined as affecting fewer than 50,000 people, or one in 2,500 (Hayashi and Umeda, 2008). Collectively, RDs represent a significant burden to health and society, as their estimated prevalence is approximately 3.5%–5.9% of the worldwide population, resulting in about 30 million people affected in Europe and 300 million worldwide (Nguengang Wakap et al., 2020). Approximately 7,000 different RDs have been identified to date, even though the exact number is debated (Hartley et al., 2018; Ferreira, 2019; Haendel et al., 2020), of which an estimated 70% are genetic (with 4,418 involved genes identified so far, November 2022), whilst the remaining are the results of infections, allergies and environmental causes. Most likely, the number of involved genes is bound to increase, as rapidly increasing quantities of exomic data are analyzed in the clinic (Boycott et al., 2018; 2019). From 2010 to 2020, the diagnosis of RDs saw a remarkable increase, with 886 new RDs being identified. During this period, the total number of genes associated with RDs grew from approximately 2,400 to over 4,000, and the number of new orphan drugs approved by the US and/or the European Union rose to 438 (Monaco et al., 2022).

Due to the very status of being rare, knowledge, research, medical expertise, and therapeutic opportunities for each particular RD are often extremely limited, and geographically sparse. Along with technological advances, the public and scientific awareness has been growing, and the knowledge on RDs is going to massively benefit from large scale data collection, integration, and sharing (Hartley et al., 2020). Many international initiatives and consortia (Gainotti et al., 2018; Azzariti and Hamosh, 2020; Bonne, 2021; Baxter et al., 2022; Laurie et al., 2022; Monaco et al., 2022) aim to significantly increase the overall percentage of RD patients with a confirmed (molecular) diagnosis, estimating that thousands of RD genes and disease mechanisms still remain undiscovered (Frésard and Montgomery, 2018; Boycott et al., 2019; Hartin et al., 2020). Exome Sequencing (ES) has been the most significant technology driving progress in the discovery and diagnosis of RDs over the past decade. While some RD diagnoses may require the integration of multiple omics data (Frésard and Montgomery, 2018; Marwaha et al., 2022), it is expected that ES will continue to play a crucial role in future efforts (Boycott et al., 2019).

The sheer re-analysis of exomic data after 1–3 years updating of the major disease variants and disease-gene association databases is reported to have increased the diagnosed cases by over 10% (Wenger et al., 2017; Setty et al., 2022). Remarkably, a further improvement in the yields could be obtained by reanalysing the data in collaboration with the clinician who made the diagnosis (Basel-Salmon et al., 2019). The contribution of research laboratories has provided an additional increase, aided by the application of novel computational and analysis tools (Eldomery et al., 2017). Thus, the fundamental step in ES data processing is the interpretation of the identified variations, i.e., the estimate of their likelihood of having a causative role in contributing to the disease. Indeed, RD-affected individuals often carry multiple variations in the gene(s) associated with the disease, with only a fraction of them being actually pathogenic (Summers, 1996). Criteria for the objective classification of variants into a five-tier system (pathogenic/likely pathogenic/uncertain significance/likely benign/benign) have been provided to the biomedical community, together with scoring rules that weight each criterion used to classify the variants. In this context, computational tools have a role in supporting the evidence framework for a benign or a pathogenic assertion (Richards et al., 2015).

This paper aims to provide an updated overview of the most frequently adopted and publicly accessible online resources and computational tools for predicting genotype-phenotype associations in RDs. In the first part of this review, we focus on the main databases collecting genes and variants associated with RDs. In addition, we describe the most popular computational methods for gene and variant prioritization, showing how information derived from molecular databases and tools can improve the diagnosis of RDs in clinical settings. Finally, we discuss the central role of FAIR data sharing in boosting research and diagnosis in the field and provide future perspectives.

## 2 Online resources and databases for rare diseases

Large-scale sequencing efforts on healthy individuals and patients allowed the collection of large databases of genetic variants and their association with human phenotypes. Based on their content and purposes, two groups of online resources for RDs can be identified: one group includes databases that define phenotype ontologies and controlled vocabularies for the description and classification of human diseases and phenotypes; the second group includes databases collecting the frequency of variants in the human population and their relationship with genetic disorders. Here, we summarize the most popular resources for medical diagnosis, focusing specifically on those related to RDs.

### 2.1 Disease and phenotype classification databases and ontologies

Nowadays, different resources for the classification of RDs are available. In particular, specific ontologies based on controlled vocabularies are defined for the description of human disorders. This enables a standardized description and classification of RDs, thereby enhancing and supporting data sharing. A standardized medical terminology was defined for developing the Medical Subject Headings (MeSH), an organized collection of hierarchical trees with increasing specificity of the downstream terms (Rogers, 1963). Later, ontologies based on diagnostic terms were created. Among them, the International Classification of Diseases (ICD), which represents the healthcare classification system maintained by the World Health Organization (World Health Organization, 2019), and the Systematized Nomenclature of Medicine (SNOMED), which implements a directed acyclic graph architecture for the

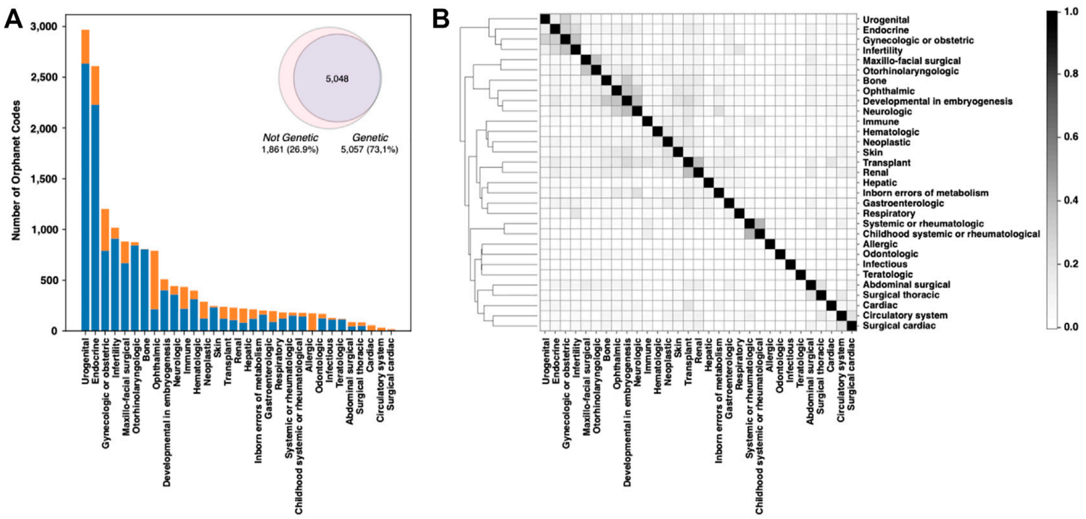

FIGURE 1
Analysis of the Orphanet database composition. **(A)** Fraction of genetic and nongenetic RDs in the different classes. **(B)** Plot showing the fraction of genes shared by each RD-class pair. Genes and Orphanet codes can be found to be associated with multiple RD classes.

automatic exploration of relationships among terms. In the 80s, the US National Library of Medicine created the Unified Medical Language System (UMLS) to harmonize the various classification systems (Bodenreider, 2004). ULMS, with its well-defined semantic relationships, is widely recognized as one of the most comprehensive resources for determining disease similarity and for the harmonization of RD data (Zhu et al., 2020). The increasing popularity of controlled vocabularies for the classification of human disorders further stimulated the creation of disease- and phenotype-oriented ontologies. The Human Phenotype Ontology (HPO) is a standardized vocabulary describing phenotypic abnormalities (Robinson et al., 2008). It is structured as a directed acyclic graph, in which a child node corresponds to a more specialized term with respect to its parent. Currently, HPO contains over 13,000 terms and over 156,000 annotations to hereditary diseases. The Disease Ontology (DO) is an open source ontology for the integration of biomedical data associated with human disease. The DO integrates concept terms from SNOMED, ICD, MeSH, and UMLS, using various semantic similarity measures (Schriml et al., 2012). The current version of DO (August 2022) collects more than 11,000 disease terms divided in 6 major classes. Mondo is the disease ontology of the Monarch Initiative (Shefchek et al., 2020) which integrates genotype-phenotype data across different species. The Mondo ontology is an open, community-driven resource which currently collects ~44,650 terms divided in three main categories (disease characteristic, disease or disorder, disease susceptibility). In Mondo, human diseases are grouped in 36 classes. Biomedical ontologies serve various purposes, such as: 1) systematizing the description of biomedical concepts for literature and clinical data recording (e.g., MeSH), 2) capturing individual clinical phenotypes, even in the absence of a recognized disease, and providing a corresponding classification in animal models (e.g., HPO), and 3) categorizing nosological entities for epidemiological and clinical management purposes (e.g., ICD). There is often overlap between these classifications, with some incorporating features from others. These ontologies of concepts are also utilized to annotate molecular data databases for the purpose of storing, analyzing, and exploring genotype-phenotype relationships.

The Online Mendelian Inheritance in Man (OMIM) database was created in the 60s by McKusick to systematically identify the relationship between disease and genetic components (Amberger et al., 2009). In November 2022, OMIM collected 7,301 phenotypes, ~85% of which were associated with at least one of the 4,674 listed genes. Focusing on the classification of rare human disorders, Orphanet is a unique resource that provides high-quality information for defining a specific nomenclature for RD (Rath et al., 2012). The description of RDs in Orphanet is based on the Orphanet Rare Disease Ontology (ORDO), a structured vocabulary capturing relationships between diseases, genes, and other relevant features. The November 2022 version of the Orphanet database collects 6,918 RDs classified in 33 major groups. All groups include genetic-caused RD, except for the "toxic effects" group. The fraction of rare genetic diseases across the remaining 32 major classes ranges from 99.6% of the *"rare inborn errors of metabolism"* to the lowest percentage of *"infectious"* diseases. The most abundant types of rare genetic disorders are those having *"developmental"* and *"neurologic"* effects (Figure 1A). Overall, RDs with genetic origin represent ~73% of the total (Figure 1A, inset). In terms of RD-associated genes, Orphanet collects more than 4,400 genes. Several of those genes are found to be associated with more than one RD class. An index of similarity (Jaccard index), based on the fraction of shared genes, has been calculated between each pair of RD groups, and is plotted in Figure 1B. The groups of "neurological" and "developmental" RD are sharing the highest number of disease-associated genes, with a Jaccard index ~0.37. The full list of the fraction of genetic RD and associated genes is reported in Supplementary Table S1.

In terms of enzymatic function, out of 5,057 genetic RDs reported in Orphanet (Supplementary Table S1), 1,596 (31.6%) are associated with enzymes, distributed among all the seven major enzyme classes (Table 1). The most represented enzyme classes are Transferases,

**TABLE 1 For each enzyme class, the table lists the number of enzymes associated with Orphanet RDs, the number of the corresponding Orphanet diseases, and the number of the corresponding Reactome roots and pathways. The data were derived from DAR database (Savojardo et al., 2022) that integrates gene-disease associations reported in UniProt, Monarch, and ClinVar.**

| Enzyme class | Enzymes[a] | Orphanet diseases[b] | Reactome roots | Reactome pathways |
|---|---|---|---|---|
| All classes | 1,218 | 1,596 | 27 | 1,098 |
| EC 1: Oxidoreductases | 186 | 259 | 20 | 209 |
| EC 2: Transferases | 474 | 738 | 26 | 799 |
| EC 3: Hydrolases | 401 | 611 | 27 | 592 |
| EC 4: Lyases | 63 | 81 | 15 | 93 |
| EC 5: Isomerases | 40 | 62 | 17 | 78 |
| EC 6: Ligases | 58 | 76 | 7 | 28 |
| EC 7: Translocases | 44 | 77 | 11 | 36 |

[a]In the distribution among classes, multiclass enzymes are counted multiple times.

[b]In the distribution among classes, diseases associated with enzymes from different classes are counted multiple times.

Hydrolases, and Oxidoreductases. Orphanet RDs can be linked to their corresponding enzyme metabolic pathways through the Reactome database, a comprehensive resource that catalogs all human metabolic reactions in 2,580 hierarchically organized pathways, with 27 main roots. Table 1 shows, for each enzyme class, the number of enzymes involved in Orphanet RDs, the number of Orphanet diseases that involve those enzymes, their Reactome roots and pathways.

The Disease And Reactome (DAR) database (Savojardo et al., 2022) provides a wealth of information on enzymes, including their relationships with Reactome pathways, molecular interactions within the pathways, and tissue expression levels as recorded in the Human Protein Atlas (Uhlén et al., 2015).

In general, the evaluation of the evidence supporting gene-disease relationships is a critical factor for an accurate diagnosis (Strande et al., 2017). To prevent mistakes in the diagnostic process, the curators of the Clinical Genome Resource (ClinGen) (Rehm et al., 2015) defined evidence-based Standard Operating Procedures for the classification of clinically relevant genes based on the presence of pathogenic variants (Section 1.3). Gene-disease relationships are classified in six groups that qualitatively describe the strength of the supporting evidence. The default class assigned to genes without any detected disease-causing variants is *"No Reported Evidence"*. Supporting evidence for gene-disease relationships is classified into four categories: *"Limited"*, *"Moderate"*, *"Strong"* and *"Definitive"*. When both supporting and conflicting evidence are present, the gene-disease relationship is classified as *"Contradictory"*. Within the ClinGen framework, the systematic review of genetic, clinical and experimental evidence, reported in databases such as OMIM and Orphanet, is used to assign one of the categories mentioned above to the reported gene-disease relationship.

## 2.2 Gene and protein network databases

A single gene defect is the most common origin of rare genetic diseases collected in the databases mentioned above. However, to investigate the molecular mechanisms underlying a RD, it is fundamental to understand and contextualize the resulting phenotype. At the protein level, defining the macromolecular complexes and pathways perturbed by the defective gene can be a useful strategy to understand the pathology itself and to intervene to restore the healthy phenotype.

A genetic variant can impact protein function and, depending on the central or marginal role of the mutated node inside a protein-protein interactions network, also the capability of the network to find alternative paths in the edges map. Changes in specific interactions can drastically perturb cellular networks and generate disease phenotypes (Barabasi et al., 2011; Menche et al., 2015).

Molecular interactions, mostly protein-protein interactions (PPIs), are annotated and archived, in structured formats, into several public resources. The major public databases collecting molecular interaction data can be divided into primary, predictive and meta-databases. Primary databases collect only manually curated molecular interactions, extracted from peer-reviewed journals, such as the IMEx Consortium resources (MINT (Calderone et al., 2020), IntAct (Del Toro et al., 2022), DIP (Salwinski et al., 2004), MatrixDB (Clerc et al., 2019)), and BioGRID (Oughtred et al., 2021). Meta-databases integrate data coming from primary databases, such as HiPPIE (Alanis-Lobato et al., 2017) and mentha (Calderone et al., 2013). Predictive databases use computational methods to predict PPIs (De Las Rivas and Fontanillo, 2012), such as STRING (Szklarczyk et al., 2021), IID (Pastrello et al., 2020) or ProfPPIdb (Tran et al., 2018).

In the panorama of molecular interaction resources, only the IMEx Consortium databases annotate interaction associated features, such as binding sites involved in the interaction or mutation effects (Porras et al., 2020). In particular, the IMEx mutation dataset contains annotations of experimental evidence where mutations have been shown to affect a protein interaction (~75,000 records) (IMEx Consortium Curators et al., 2019). The dataset can be used to map selected pathogenic variants to manually curated PPIs and to understand the effect of a specific variant on the interactions at protein-protein interface. Moreover, from the IntAct datasets, it is possible to download a RD specific dataset of molecular interactions extracted from literature. The dataset is enriched with experimentally proven impact of clinical mutations on interactions, and also with the non-clinical mutations which are found to impact protein functionality. So far, the dataset contains over 7,900 interactions involving about 2,500 interactors. The dataset can be visualized and filtered in the IntAct result page, or in Cytoscape (Shannon et al., 2003), using the IntAct App (Ragueneau et al., 2021).

Disease specific biological networks can also be constructed or integrated with data coming from signaling pathways databases such as Signor (Lo Surdo et al., 2023), WIKIPathway (Martens et al., 2021) or OmniPath (Türei et al., 2016). They can then be imported into Cytoscape by using resource specific CytoscapeApps (Kutmon et al., 2014; Ceccarelli et al., 2020; De Marinis et al., 2021), to gain more insight into the molecular mechanisms involved in the disease. Moreover, pathway resources such as KEGG (Kanehisa et al., 2017) and Reactome (Jassal et al., 2020) databases are very important to discover whether some disease-associated subnetworks are enriched for a particular functional pathway.

By the combination of PPI with genotype-phenotype relationships, functional similarities have been used to generate specific disease networks defining similarity across different human disorders (Goh et al., 2007; Menche et al., 2015; Buphamalai et al., 2021). Such networks have been shown to be useful for studying the biological mechanisms of diseases and for the development of gene prioritization tools (Zhang and Itan, 2019). Some examples of gene prioritization tools, specific for RDs, will be discussed in Section 2.2.

## 2.3 Databases of variants

The Human Genome Variation Society (HGVS) maintains comprehensive lists of databases focused on variations, from locus-specific mutation databases to SNP databases, to chromosomal variations, to other mutation databases, including nonhuman and artificial mutations. However, given the high number of resources, it is nearly impossible to perform an exhaustive description of all those that are available. We will therefore focus on selected, curated and widely used resources. None of them is specifically dedicated to RDs; however, it is possible to collect data and information on RD-associated variations.

In general, variant databases can be divided into two groups, according to whether they focus on the variant's frequency across the human populations or on their pathogenic effect.

The variant's frequency can be derived from sequencing experiments on a large set of individuals. For example, the 1,000 Genome project, started in 2008, collected and sequenced the genomes of 2,504 individuals from 26 populations worldwide, characterizing more than 88 million variants, including >99% of SNP variants with a frequency higher than 1% (1000 Genomes Project Consortium et al., 2015). The datasets and the related analyses have been freely shared with the scientific community by setting up the International Genome Sample Resource (IGSR) (Fairley et al., 2020) to ensure their future usability and accessibility. Data about these variants can be explored through the Ensembl Variation database (Hunt et al., 2018), a project aimed at automatically annotating the genomes, integrating biological data and making all information accessible via a website. Those variants were grouped into subsets, based on the origin of the individual and on the frequency of occurrence. In the same period, the UK10K project (UK10K Consortium et al., 2015) sequenced the whole genomes of about 10,000 individuals, characterizing over 24 million novel sequence variants. That information was made available via a dedicated website and via the European Genome-phenome Archive (EGA) (Freeberg et al., 2022), a resource for permanently archiving and sharing personally identified genetic, phenotypic and clinical data, obtained by biomedical research projects. Another analogous study is the "All of Us" research program, funded by NIH, sequencing 100,000 genomes from ethnic groups underrepresented in previous projects (All of Us Research Program Investigators et al., 2019). While the "All of Us" project was of broader scope, the 100,000 Genomes Project, focused on patients with an RD (161 disorders covering a broad spectrum of RDs were present) or with one among 20 different common cancer types (Turnbull et al., 2018). A pilot study, conducted on the genomes of 4,660 people, increased the diagnosis number for 25% of participants. Among them, 14% of the cases were new diagnoses based on variants found in regions usually missed in conventional, non-whole genomic tests (100,000 Genomes Project Pilot Investigators et al., 2021).

A widely used database collecting variant frequency data is the Genome Aggregation Database (gnomAD). The gnomeAD is the successor to the Exome Aggregation Consortium (ExAC), a project that was launched to aggregate and harmonize exome and genome sequencing data from a variety of large-scale sequencing projects (Karczewski et al., 2020). The National Center for Biotechnology Information (NCBI) at the NIH hosts several resources for investigating and understanding human variations. dbSNP and dbVar are two freely available databases, the former hosting a broad collection of small genetic polymorphisms (SNP, deletion/insertion polymorphisms, etc.), and the latter hosting a broad collection of large variants (>50 bp) (Lappalainen et al., 2013).

The second class of variant databases collect information about their clinical significance and their association with human disorders. To this purpose, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP), developed specific guidelines, where variants are classified into five types: *"pathogenic"*, *"likely pathogenic"*, "uncertain significance", *"likely benign"*, and *"benign"* (Richards et al., 2015). On the one hand, *"pathogenic"* and *"likely pathogenic"*, variants are classified by using multiple criteria grouped in four weighted categories: *"very strong"*, *"strong"*, *"moderate"*, and *"supporting"*. On the other hand, *"likely benign"*, and *"benign"* variants are classified using a combination of rules grouped in three weighted categories: *"stand-alone"*, *"strong"* and *"supporting"*. All previous criteria are based on eight categories of information, including among them data from population studies and computational predictions. If a variant does not meet any criteria or the evidence for benign and pathogenic is conflicting, the class assigned by default is *"uncertain significance"*. As recently shown (Tavtigian et al., 2020), the ACMG/AMP guidelines are compatible with a quantitative Bayesian formulation, whose scaling as odd of pathogenicity allows an empirical calibration of the strength of the reported evidence.

A reference database, collecting annotated genetic variants by adopting the ACMG/AMP guidelines, is ClinVar (Landrum et al., 2020) which represents one of the main sources of information for gene classification in the ClinGen database (Section 1.1). ClinVar is a freely accessible, public archive, collecting reports of variants found in patient samples, assertions about their clinical significance, and other data, including the availability of supporting evidence. ClinVar thus allows users to infer relationships between human variations and the health status of the patients. Each variation has its own accession number, and, if multiple submitted records about the same variation/condition pair are present, they are aggregated under a single accession number. The adoption of a single variant identifier allows users to review all data submitted for a single variant, regardless of the condition for which it was interpreted. In fact, ClinVar neither curates content nor modifies interpretations associated with a single record. The alleles are reported according to the HGVS standards. Focusing on protein variants, the UniProt consortium is releasing a curated file reporting a list of protein variants, grouped in 3 classes: "*Disease*", "*Polymorphism*" and "*Unclassified*" (humsavar UniProt, 2023). In the *humsavar* file, an OMIM identifier is associated with each pathogenic variant. Alternatively, the Human Gene Mutation Database (HGMD) is a proprietary database of mutations in human genes, associated with inherited diseases, which contains both inherited and somatic mutations (Stenson et al., 2020). The GWAS Catalog, supported by a collaborative initiative between the National Human Genome Research Institute and the EMBL-EBI, is another popular freely available database of SNP-trait associations, which can be easily integrated with other resources (Sollis et al., 2023).

The collection and curation of several variant databases is supported by ELIXIR, an intergovernmental organization that brings together bioinformatics resources for life sciences from across Europe (https://elixir-europe.org/). For example, the European Variation Archive (EVA) (Cezard et al., 2022) is an open-access database of all types of genetic variations (SNP, large structural variants, observed in germline or somatic sources) from all species. Submitted variants (in Variant Call Format, VCF) are merged, normalized and annotated for functional consequences and to determine allele frequency values in a study-specific manner. Human variants are also exchanged with dbSNP and other NCBI resources. DisGeNet (Piñero et al., 2020) is another database that integrates information on human gene-disease associations and variant-disease associations from different repositories. The data are annotated with controlled vocabularies and community-driven ontologies, and several original metrics are provided to assist the prioritization of genotype–phenotype relationships. Another ELIXIR core resource collecting information about variation is Ensembl Variation (Hunt et al., 2018), a resource linked to the Ensembl Genome Browser. It stores variants found in many species (including human) and, where available, associated diseases and phenotype information. Variant data are imported from a variety of sources (e.g., dbSNP) and subjected to a quality control process. They are then classified and their consequences predicted. Moreover, variant sets are created to help people retrieve a specific group of variants from a particular dataset. For human data, the linkage disequilibrium is also calculated for each variant, by population. A list of resources and databases for RD cited in this paragraph is reported in Supplementary Table S2.

# 3 Tools for rare disease genome interpretation

## 3.1 Automatic variant calling pipelines

The analysis of next-generation sequencing (NGS) experiments requires substantial bioinformatics resources. During the last years, a variety of analytical tools have been developed for the detection of genetic variants. Such tools assist all steps of the variant calling process, including quality control and trimming, alignment to the reference genome, identification, and annotation of SNVs and short indels. Although the Genome Analysis ToolKit (GATK) Best Practices guidelines define standard procedures for setting up a variant analysis pipeline, selecting the best approach among the variety of tools for NGS data processing can still be challenging. To overcome the issue and simplify the variant calling process, several "ready-to-use" bioinformatics pipelines to process ES and GS data have been made available. Some of them include: fastq2vcf (Gao et al., 2015), SeqMule (Guo et al., 2015), ExScalibur (Bao et al., 2015), Appreci8 (Sandmann et al., 2018), JWES (Ahmed et al., 2021), OVarFlow (Bathke and Lühken, 2021) and the recent DeepVariant (Poplin et al., 2018) that integrates a deep-learning-based variant caller. Most of those pipelines integrate many variant calling tools to increase sensitivity, but they are command-line applications to be installed on local servers. Alternatively, web-based options are available, e.g., Maser (Kinjo et al., 2018), CSI NGS Portal (An et al., 2020), and the most popular Galaxy (Afgan et al., 2018). Recently, *seqr*, a web tool for the analysis of rare disease genomes, has been made available by the Broad Institute (Pais et al., 2022). They are open-source platforms that provide a user-friendly graphical interface, improving the accessibility to computation analyses of genomic data. In particular, Galaxy users can freely create custom workflows or find already existing workflows, available on Galaxy Toolshed (Blankenberg et al., 2014), which can be run on public Galaxy servers. The disadvantages of using the web-based options are the limited amount of data that can be uploaded, the CPUs time, and the limitations on some tools on the public Galaxy platforms. However, Galaxy pipelines can also be run on a local Galaxy installation, or on a paid cloud infrastructure, e.g., Amazon cloud (AWS), using CloudMan (Afgan et al., 2010). Terra is another example of a web- and cloud-based platform, providing a compute environment to run optimized pipelines on Google Cloud. Galaxy, Terra, and other analysis components are integrated in a unified environment for data analysis and management, AnVIL (Schatz et al., 2022), designed to manage and store genomics data, enable population-scale analysis, and facilitate collaborative large-scale research projects. Nevertheless, "best practices" for variant calling in clinical settings, should be considered before choosing the most appropriate sequencing strategy, and the most reliable combination of tools for read alignment/preprocessing, variant calling and filtering (Koboldt, 2020).

Furthermore, to ensure the reproducibility of complex bioinformatics analysis, different workflow languages have been used to develop specific data analysis pipelines. The NextFlow core community (Ewels et al., 2020) collected a curated set of optimized procedures for the analysis of genomic data specific for rare disease. Similar projects include Dockstore (O'Connor et al., 2017), which provides containerized tools and workflows, currently supporting 4 different languages: the Workflow Description Language (WDL), Common Workflow Language (CWL), Nextflow, and Galaxy Workflows (GWs). Moreover, several workflows accessible on Dockstore can be easily launched in web-based platforms, such as Terra. These workflow languages are designed to handle some aspects of computational workflows, such as resources, software, and execution of analysis steps. Among those, Snakemake (Köster and Rahmann, 2012) and Nextflow (Di Tommaso et al., 2017) are commonly used for developing new research pipelines, while WDL and CWL workflows are preferred for large-scale projects (Reiter et al., 2021). Recently, a specific pipeline for the analysis of rare disease genome has been made available in NextFlow (Ewels et al., 2020).

Most of the above semi-automatic pipelines help streamline the generation of variant lists (in vcf format), but lack the downstream annotation and filtering steps that are necessary to identify disease-causing variants. To this end, different data-warehousing solutions to store genomic variants, along with the relevant genomic annotations, were deployed to allow a flexible and efficient data exploration. An example is GEMINI (Paila et al., 2013) and OpenCGA that supply the platform and the analysis framework to build customized genomic databases, to efficiently store data to be queried and visualized. A list of tools for variant calling and annotation is reported in Supplementary Table S3.

## 3.2 Gene prioritization tools

The objective of gene prioritization is to rank a large list of potential candidate genes based on their relevance for a disease. The prioritization algorithms identify the most promising genes, as to their association to the molecular basis of a given disease and/or a specific phenotype, for defining a therapeutic and/or diagnostic procedure. From the experimental point of view, the high-throughput techniques reduced the costs for generating a high amount of information about gene mutations. On the other hand, the identification of real links between genes and diseases is still a time- and money-consuming task. Therefore, the help of computational tools to reduce the number of genes to be investigated is strongly needed. Beyond the assumption that one gene codes for one function, the possibility that defects of one gene may be related to multiple diseases is now taken into account. At the same time, more genes can be involved in a given disease. In fact, a given metabolic pathway is composed of several protein functions, a defect in any of which may result in the pathway failure. Computational tools for gene prioritization use different sources of information to rank the candidate genes. Possible features are direct experimental data on gene sequences, mutations, expression (co-expression), gene-gene and protein-protein interactions, as well as more indirect evidence as ontologies, literature, information derived by model organisms. Different types of tools may differ by the focusing level (e.g., disease-specific or not), by the applied methodology (e.g., text-mining, similarity profiling, network analysis), by the approach to select the best candidate genes (e.g., ranking or filtering into smaller subsets), by the assumptions (i.e., genes may be directly or indirectly associated with a disease), or simply by the type of experimental evidence used for the analysis. Several works list the available tools on the basis of the state-of-the-art and classification applied (Moreau and Tranchevent, 2012; Piro and Di Cunto, 2012; Gill et al., 2014; Zolotareva and Kleine, 2019; Cabrera-Andrade et al., 2020; Jacobsen et al., 2022; Yuan et al., 2022). For instance, Jacobsen et al. applied phenotype-driven methods to improve diagnostic yields for RD, and listed 16 freely available tools (Jacobsen et al., 2022). Zolotareva and Kleine listed 14 tools, classifying them according to strategies, approach types, interfaces, input, and the types of evidence sources (Zolotareva and Kleine, 2019). Smedley and Robinson compared 7 tools and summarized their features in terms of exome input, types of variants analyzed, and approach (Smedley and Robinson, 2015). Problems related to long-term maintenance of academic software are very common (Jacobsen et al., 2022) and solutions have been proposed (Rother et al., 2012). A list of tools from the cited literature is reported in Supplementary Table S4. Among all gene prioritization methods, for instance, VarElect and ToppGene are part of standard diagnostic pipelines in the clinical settings. In particular, VarElect (Stelzer et al., 2016a) is a comprehensive, phenotype-dependent, variant/gene prioritization tool, based on the GeneCards suite (Stelzer et al., 2016b). The input of VarElect is a gene list together with a free-text phenotype description, such as disease and symptom terms, which represents a useful interface for non-skilled users. The tool prioritizes the genes on the basis of scores for the associated terms, computed on the appearance frequency in the entire GeneCards knowledgebase. The latter includes also the human disease database MalaCards (Rappaport et al., 2017), the human biological pathways of Pathcards (Belinky et al., 2015), and the gene expression information in cells and tissues of LifeMap Discovery (Edgar et al., 2013), for a total of 120 sources. The results of VarElect are displayed as a table of genes with decreasing phenotype relation scores. Alternatively, the gene prioritization task can be performed by ToppGene (Chen et al., 2007; Chen et al., 2009a; Chen et al., 2009b), a suite including tools for gene list functional enrichment, candidate gene prioritization, and identification and prioritization of novel disease candidate genes in the interactome. ToppGene selects genes in the training set on the basis of their association with disease, pathway, GO term, phenotype. The test set can be given by candidate genes from linkage analysis studies, differential expression in a particular disease or phenotype, interactome knowledge. The enrichment step is based on a variety of data sources that cover 14 types of annotation. For each type of annotation of each test gene, a similarity score is generated, by comparison to the enriched terms in the training set. The prioritized gene list is ranked on the aggregated values of the 14 similarity scores.

Finally, specific algorithms for the prioritization of RD-associated genes were recently developed (Zhu et al., 2012; Liu et al., 2017; Buphamalai et al., 2021; de la Fuente et al., 2023). Among them, for instance, an algorithm was developed, based on the calculation of a vertex-similarity score between each pair of genes, that was tested on a set of ~1,600 known orphan disease-causing genes associated with 172 RDs (Zhu et al., 2012). Another method, which computes the topological similarity between genes connected in a PPI network, ranks the candidate genes combining two scores reflecting the local and global connectivity of the network (Liu et al., 2017). The success rate of this method can reach 50%–75% on a set of ~1,200 genes collected from the Orphanet database. A more comprehensive approach evaluates the impact of rare gene defects, building a multiplex network with more than 20 million gene relationships organized into 46 network layers (Buphamalai et al., 2021). The analysis of 3,771 RDs reveals distinct phenotypic modules that can be used to accurately predict RD gene candidates. A recent tool (GLOWgenes), based on 33 functional networks classified in 13 knowledge categories, was able to recover genes associated with 91 genetic diseases classified into 20 families (de la Fuente et al., 2023). When applied to 15 unsolved cases, GLOWgenes was able to identify three new genes potentially associated with syndromic inherited retinal dystrophies.

## 3.3 Variant interpretation methods

Variant interpretation tools are *in silico* predictive programs that can help researchers in establishing the pathogenicity of the variations identified in the gene(s) of interest. Many approaches have been developed to perform these predictions, and their number has grown very rapidly in the last years. They mainly focus on predicting the impact of a missense variation on the structure and function of the associated protein, or on predicting effects on RNA splicing.

More recently, programs addressing more general noncoding variants have also been developed (Özkan et al., 2021). Researchers and clinicians tend to use variant interpretation tools in combination, as also suggested by the ACMG/AMP guidelines (Section 1.3). Nevertheless, their concordance in asserting the variant effects (especially of the predicted benign ones) has been rather low until present. More recently, however, newly developed algorithms have shown good performance in many types of genes and mutation mechanisms. Furthermore, by using gene-specific algorithms, and by calibrating them with well-characterized sets of benign and pathogenic variants, better results may be reached, than with general use algorithms (Ghosh et al., 2017).

In the last two decades, an impressive number of methods and algorithms for single amino acid substitution (SAS) have been devised to predict the variant effect on protein structure, function and interactions, to eventually identify those involved in molecular pathogenicity. As a matter of fact, SASs represent more than 40% of the unique variants found in the Exome Aggregation Consortium (Lek et al., 2016). Those methods are obviously not specific to RDs and have a broad range of applications (Capriotti et al., 2019; Katsonis et al., 2022; Pancotti et al., 2022). A selection of the most recent methods and resources is reported in Supplementary Table S5. Many of the early methods were based on the prediction of the effect of a single mutation on the protein thermodynamic stability, as destabilization is one of the key factors in pathogenesis (Capriotti et al., 2008; Dehouck et al., 2011; Worth et al., 2011; Fariselli et al., 2015; Laimer et al., 2015; Quan et al., 2016; Savojardo et al., 2016; Yang et al., 2018; Marabotti et al., 2020; Pires et al., 2020; Montanucci et al., 2022). Subsequent efforts and developments in the field produced last-generation methods, using one of three general strategies: i) prediction of the likelihood of a missense mutation for causing pathogenic changes in a protein (Sim et al., 2012; Adzhubei et al., 2013; Carter et al., 2013; Katsonis et al., 2014; Niroula et al., 2015; Capriotti et al., 2017; Raimondi et al., 2017; Rentzsch et al., 2019; Pejaver et al., 2020; Manfredi et al., 2022; Quinodoz et al., 2022); ii) evolutionary conservation analysis of the mutated sites; iii) methods combining different strategies (Stein et al., 2019; Petrosino et al., 2021). More recently, several methods have been developed to also predict the impact of variants in noncoding regions (Rojano et al., 2019; Katsonis et al., 2022; Tabarini et al., 2022). These methods include generic tools, which predict single-nucleotide pathogenic variants across the entire genome (Quang et al., 2015; Shihab et al., 2015; Zhou and Troyanskaya, 2015; Capriotti and Fariselli, 2017; Rentzsch et al., 2019) and more specific algorithms, which predict the impact of splicing variants (Desmet et al., 2009; Cheng et al., 2019; Jaganathan et al., 2019; Rentzsch et al., 2021). In particular, splicing-affecting variants are established contributors to RD, of which they may modulate the phenotypic outcome (Li et al., 2016; Scotti and Swanson, 2016).

To assess the performance of the available variant interpretation algorithms on the variants specifically associated with RDs, we collected a dataset of SAS from ClinVar (March 2022). Such a dataset (sas-rd-202203 in Supplementary Materials) is composed of ~27,600 SAS in genes associated with rare genetic disorders from different RD classes. From RD-associated ClinVar genes, we selected 16,012 variants classified as *Pathogenic* and 11,633 *Benign*. The results of our analysis, scoring the performance of 4 state-of-the-art variant interpretation tools (CADD, FATHMM, PhD-SNP$^g$ and VEST4), show that the selected methods reach on average 83% overall accuracy (Q2), 0.65 Matthews correlation coefficient (MCC), and >0.90 area under the ROC curve (AUC) (Supplementary Table S6). These results are in the same range of those reported in previous works, not limited to RDs (Capriotti and Fariselli, 2017; Benevenuta et al., 2021). A chromatic representation of the performance of the methods (Figure 2) reveals that VEST4 reaches the highest AUC (0.96) while FATHMM the lowest (0.83). Taking into account some possible data overfitting, we expect that the resulting measures of performance might be overestimated by no more than 2%–5% (Capriotti and Fariselli, 2017). The results of the four tools in predicting the effect of different RD classes exhibit some variation. Specifically, for the *Ophthalmic* RD class, with 7,889 variants (28.5%), the performance of the methods is slightly higher than average, reaching 85% overall accuracy, 0.69 Matthews correlation coefficient and 0.92 AUC. Conversely, the lowest performance was observed in predicting the impact of 2,152 variants associated with the *Cardiac* RD class (~7.8%), with an overall accuracy below 80%, a Matthews correlation coefficient of 0.58, and an AUC of 0.87. Although our analysis shows that state-of-the-art methods for the prioritization of causative variants in RD-associated genes result in a high-performance level, further work is needed for improving the tools' reliability, in view of the residual ~10% of misclassified variants. In this regard, it appears that an important aspect to be considered for improving the predictions reliability is the conservation level at the variation site (Capriotti and Fariselli, 2022).

## 3.4 Genotype/phenotype association methods

Despite the progress in our capacity to prioritize disease-causing genotypes in clinical exomes and genomes, the large number of variants that remain to be evaluated for the diagnosis-making

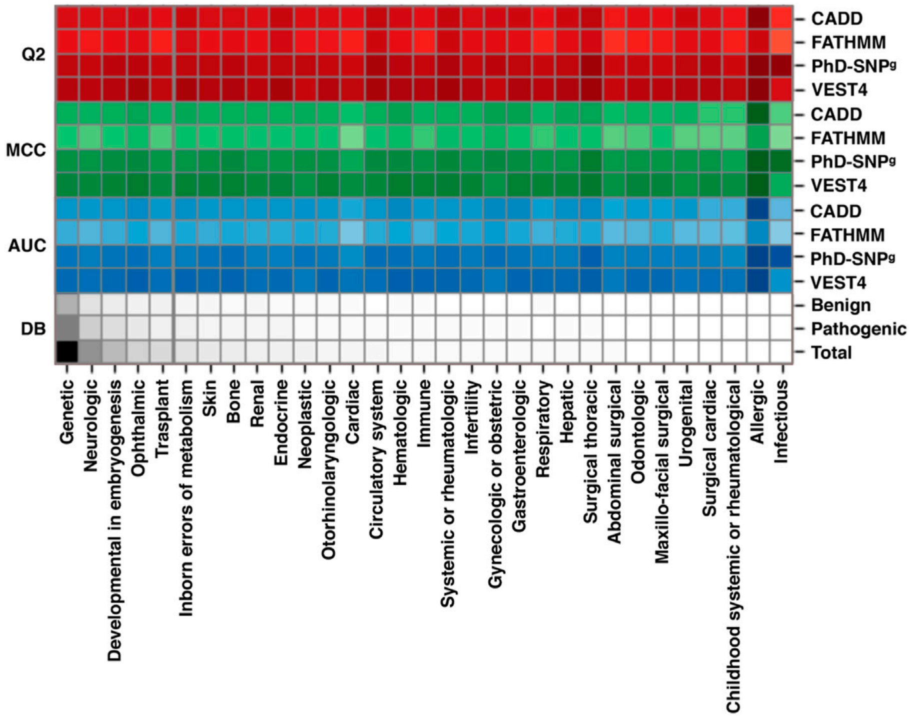

FIGURE 2
Performance of four state-of-the-art methods (CADD, FATHMM, PhD-SNP$^g$ and VEST4) on a dataset of RD-associated variants from ClinVar database, featuring at least one annotation as Pathogenic or Benign. The scores are calculated for the different classes of RDs. All 27,648 variants (16,012 Pathogenic and 11,633 Benign) in our dataset are in the Genetic class. According to the Clinvar annotation, each variant can be classified in multiple RD groups. Performance parameters shown are: Q2, Overall Accuracy; MCC, Matthews Correlation Coefficient; AUC, Area Under the receiver operator characteristic Curve. DB indicates the fraction of each RD group in the dataset. The performance of CADD was calculated considering a Phred-like score threshold of 20. The color darkness in the drawing is proportional to the numerical values, which are reported in Supplementary Table S6. The predictions of the four methods are reported in Supplementary Materials.

process is still a challenge. Computational analysis of phenotypes, in addition to genotypes, has proven powerful to improve the standardization and automation of NGS diagnostic pipelines from raw sequences to prioritized variants. The general principle, followed by such analyses, is to compute measures of similarity between the clinical manifestation in a patient and the description of disease(s) linked to a gene. Gene and/or variant prioritization tools measure ontological similarity between a set of query terms, representing the compendium of the patient's clinical phenotypes, and the set of terms that are associated with any disease-gene (Smedley and Robinson, 2015). Algorithms underlying such tools have been developed, exploiting standardized collections of clinical terminologies, the most widely adopted of which is the Human Phenotype Ontology (HPO). The latter is used to assist clinical scientists and researchers in clustering and comparing phenotypes of patients with shared molecular background, with the aim to improve genetic diagnosis and genotype-to-phenotype correlations. Many tools that exploit the knowledge of known phenotypes of disease genes in humans and animal models have been developed. Such tools can be broadly categorized into two groups: those that take both phenotype and genotype data (VCF + HPO) as input, and those that only accept phenotype data (HPO only). These tools have been thoroughly reviewed and evaluated in a recent publication (Yuan et al., 2022). As an example, one of the earliest and most used tools is Exomiser (Robinson et al., 2014). Exomiser combines the most popular strategies for variant filtration with HPO to prioritize data in a VCF file. Despite its name, the Exomiser analysis framework is not limited to the exome but incorporates the Regulatory Mendelian Mutation (ReMM) score for relevance prediction of non-coding variations (SNVs and small InDels) (Smedley et al., 2016). In their review, Yuan et al. (2022) identified the two best performers in HPO-based gene prioritization to be LIRICAL (Robinson et al., 2020) and AMELIE (Birgmeier et al., 2020). Both of those recently published tools propose innovative and interesting analysis approaches. LIRICAL aims to overcome simple gene or variant ranking based on semantic similarity as a prioritization scheme, by introducing a likelihood-ratio test to provide an estimate of the post-test probability of candidate diagnoses. AMELIE, conversely, consists in an end-to-end machine learning approach with web interface, that finds relevant literature supporting the disease causality of genetic variants and their association with different clinical presentations. In the benchmark from Yuan et al. (2022), the two methods often resulted in quite different predictions of highly ranked causal genes, and such a complementarity suggests a possible integrative approach to further enhance the diagnostic efficiency. In a recent work, genotype/phenotype association methods were tested on a set of 4,877 molecularly diagnosed cases, affected by RDs, from the 100,000 Genome Project (Jacobsen et al., 2022). On this set, Exomiser was able to recall 82% and 92% of the diagnosed cases as

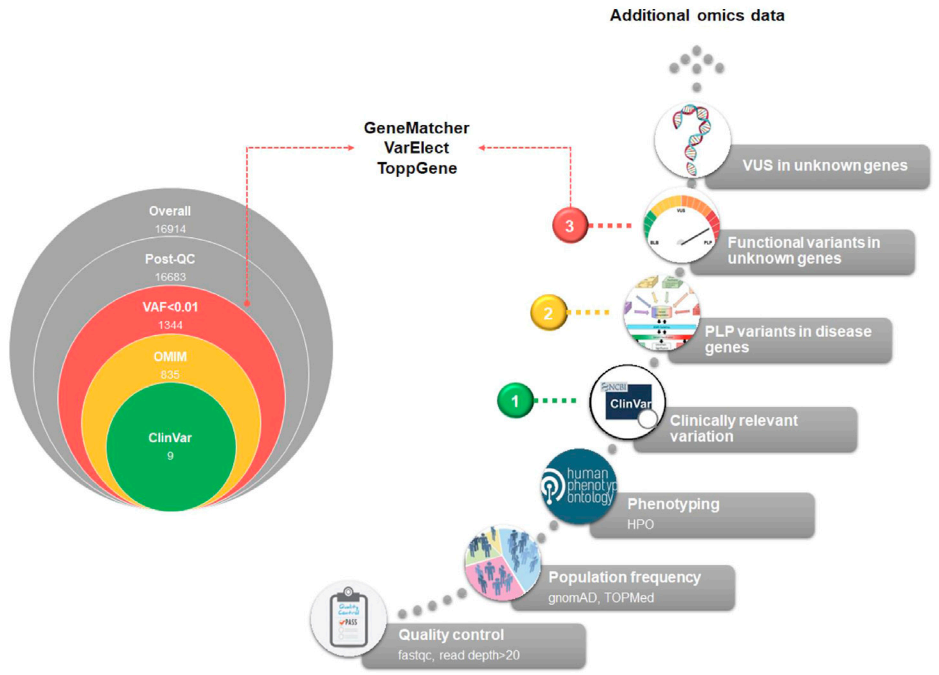

FIGURE 3
Exome analysis flowchart. A diagram of the main steps of NGS data analysis is shown. On the left, the progressive reduction by filtering in the number of likely disease-causing variants is shown, for a general patient case. The reported numbers are from a typical single patient case. On the right, the filtering process is detailed. Based on the identified variants, we can recognize three different diagnostic situations: (1, green dot) identification of P/LP variants with well-established association to RD phenotype; (2, yellow dot) identification of new P/LP variants in genes with known association to the phenotype; (3, red dot) identification of functional variants in genes with unknown association with the phenotype. A fourth case should be considered, i.e., the identification of VUS variants in genes with unknown association with the phenotype. In this case, complementing different approaches, such as short-read genome sequencing with RNA sequencing, and methyl profiling, should be considered to elucidate the molecular mechanism of the disease and improve the diagnostic yield.

the top hit, and within the top 5 scores, respectively. These positive results are going to render phenotype-genotype association tools essential for RD diagnosis in the clinical routine.

# 4 Use cases on rare disease genome interpretation

The diagnostic workflow of NGS genetic testing is composed of three levels of data analysis: i) quality control of raw data, ii) variant annotation, and iii) variant filtering. On the basis of the annotation level of the variants detected after the variant calling procedure, we can identify three main steps of analysis. In Figure 3 we summarized the main filtering steps, including the approximate number of unique variants that can be identified in a single subject after a clinical exome. In order to efficiently prioritize clinically relevant variations among all types of captured variants described in Figure 4, we need to adopt different analytical strategies. Several resources, including databases of genomic variation and phenotypes, population frequency data and *in silico* prediction approaches, can be used for the interpretation of each type of variant (Supplementary Table S5).

According to the variant types, distinguished by their gene location and the currently available resources for variant interpretation, we can define the three following possible cases.

## 4.1 Identification of pathogenic variants associated with RD phenotypes

In this case, the analysis workflow is well-defined and relatively easy. In the hypothetical case shown in Figure 4, after the application of the filters of variant allele frequency (VAF) and phenotyping (see below), known clinically relevant variants in disease genes that fit with the phenotype are selected to be reported.

The VAF is retrieved from population databases (Section 1.3), using resources that aggregate exome and genome sequencing data from a variety of large-scale sequencing projects, and make summary data available for the wider scientific community. They include sequencing data of both affected and unaffected subjects or different populations (Gudmundsson et al., 2022). In current diagnostic settings, ultra-rare and rare variants (VAF < 0.001 and VAF < 0.01, respectively), as well as private variants (not annotated in population databases) are selected. Of course, this primary filter can be modified according to the prevalence in the population of a specific disorder. Thus, in some diagnostic settings, also low-frequency variants (VAF < 0.05) can be selected (Andolfo et al., 2021).

The exact characterization of the phenotype (*"phenotyping"*) is one of the most relevant aspects of NGS genetic testing, and it is often considered a major challenge for the NGS-based genetic diagnosis. Generally, *phenotyping* is obtained using a standardized vocabulary of phenotypic abnormalities encountered

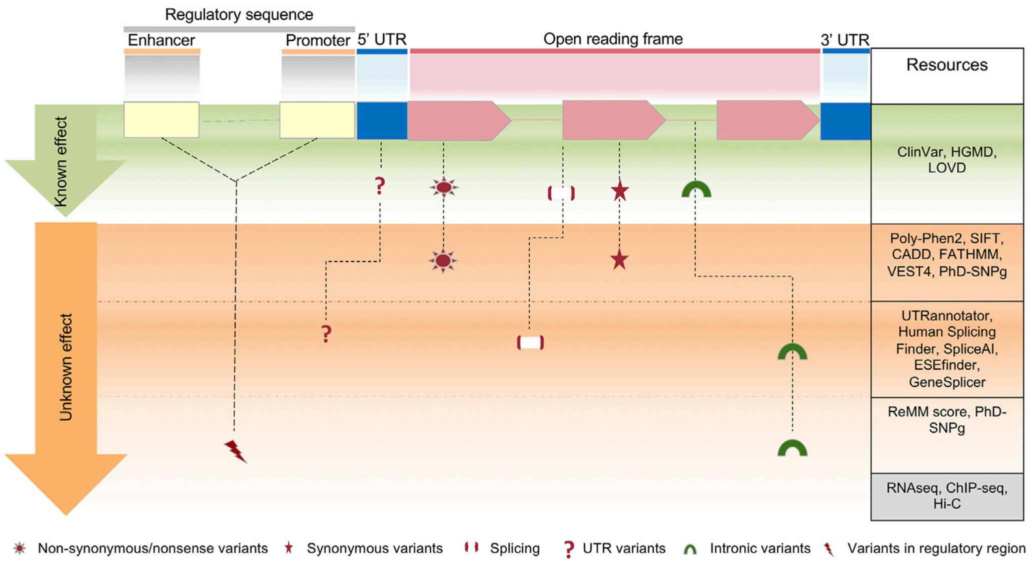

FIGURE 4
Schematic view of the clinical variant interpretation process. In a human protein-coding gene, a variant in the exons of an open reading frame can result in synonymous or nonsynonymous changes, while a variant in other areas (splice or intronic regions) can impact on splicing regulation. Changes within regulatory sequences (yellow and blue) can affect transcription and translation regulation of gene expression. On the right column, a selection of the most commonly used resources for variant interpretation is reported, distinguished by their gene location. Several methods are currently available to predict the effect of coding variants, however the interpretation of variants in deep-intronic regions or in regulatory elements is still challenging, due to the limited number of *in silico* prediction approaches. Such shortcomings can be overcome by parallel sequence analysis of the whole exome/genome together with multi-omics technologies, including RNA sequencing (transcriptome analysis), ChIP-seq (chromatin immunoprecipitation assay) and HiC (high-throughput chromosome conformation capture).

in human disease, such as that provided by HPO database (Section 2.4). Clinically relevant variants can be prioritized using public repositories reporting correlation between genetic variants and phenotypes, such as ClinVar and HGMD (Section 1.3; Supplementary Table S2).

## 4.2 Identification of pathogenic variants in RD associated genes

Herein, after the application of the aforementioned filters of variant frequency, phenotyping, and clinically relevant variants in disease genes, no variants are prioritized. In this case, to prioritize pathogenic/likely pathogenic (PLP) variants associated with the phenotype, ACMG/AMP guidelines for variant interpretation (Section 1.3) are used.

According to those guidelines, the pathogenicity of each variant is evaluated by gathering evidence from various sources: population data, computational and predictive data, functional data, and segregation data. Computational and predictive data are obtained by using several *in silico* prediction programs described in Section 2.3. Those tools are mainly devoted to the evaluation of the missense variants, which constitute a major set of VUS (Variant of Unknown Significance). The ACMG/AMP guidelines recommend complete concordance of predictions among all *in silico* algorithms used, without specifying the number or types of algorithms. However, many studies have provided rules for the classification of non-synonymous variants based on the integration of different prediction tools (Ghosh et al., 2017; Li et al., 2019; Nicora et al., 2022).

For phenotype characterization, the analysis of splicing variants is also relevant. The prioritization of splice site variants can be performed by web server tools, such as Human Splicing Finder (Desmet et al., 2009), MMSplice (Cheng et al., 2019), SpliceAI (Jaganathan et al., 2019) and CADD-splice (Rentzsch et al., 2021), that can highlight potential splicing-affecting variants outside the canonical splicing sites. ACMG/AMP variant classification can be achieved in such cases by using InterVar or wInterVar (Li and Wang, 2017), a web server that enables user-friendly variant interpretation with both an automated interpretation step and a manual adjustment step. Functional data that supports the pathogenic effect of newly discovered variants is not typically included in the standard diagnostic process of NGS genetic testing. Nevertheless, laboratories with an extensive experience in a specific disease area, can provide additional functional evaluation for new variants as part of their diagnostic protocols (Thouvenot et al., 2016; Ellingford et al., 2022).

Finally, segregation and allele data are fundamental to correctly assess the pathogenicity of variants. For this reason, in diagnostic settings the trio analysis, i.e., the combined genomic analysis of patient and parents, is strongly recommended (Alfares et al., 2020; French et al., 2022; Gabriel et al., 2022).

## 4.3 Identification of functional variants in genes with unknown RD association

Currently, the diagnostic process reaches a definitive diagnosis only in about 50% of the cases, leaving many patients with strongly-suspected genetic diseases without molecular explanations. In such cases, all variants with potential functional effects on any gene must be considered, under the hypothesis that the pathogenic role of the causative gene is still unknown. The initial filtering steps, similar to the previous scenarios, consist of removing all variants unlikely to be implicated in the disease, either because of low quality in exome or high frequency in population. Very stringent frequency thresholds are used, since it is likely that the considered disease is extremely rare. Then, the pedigree is analyzed to maintain only the variants that co-segregate with the phenotype according to any Mendelian transmission model. The variant-affected genes are prioritized to remove those that show a high grade of variability in the general population and to highlight those with a plausible biological role in the disease phenotype. The resulting set of genes with functional variants, poor population variability and biological compatibility is released in gene matching tools to search for other patients who are affected by alterations in the same genes (Section 2.2). Once a 'match' occurs, the researchers are connected through the system and can share molecular and clinical details of the patients, potentially concluding that they are both affected by the same disease, caused by the matched genes.

An example of successful gene-matching regards a 19-years-old girl seen at Federico II University Hospital, Naples, Italy. The girl was affected by a severe clinical picture, composed of complex brain malformations, extraocular muscle anomalies, severe intellectual disability, and drug-refractory epilepsy. Despite the presentation strongly suggesting an underlying genetic cause, thorough molecular and metabolic investigation failed to yield any plausible explanation. The patient was, then, enrolled in the Telethon Undiagnosed Diseases Program (TUDP) and underwent patient-parent trio ES. Variant filtering and manual revision did not find causative variants but highlighted those in four non-disease genes (PLEKHN1, NR5A2, TMEM89, DHX37). The patient's clinical and molecular descriptions were released in PhenomeCentral for gene-matching (Buske et al., 2015; Sobreira et al., 2017; Osmond et al., 2022), where only for DHX37 a consistent match with other patients with syndromic intellectual disability was found (Paine et al., 2019).

However, depending on disease type and patient selection, exome sequencing has been estimated to lead to a diagnosis in 30%–50% of rare Mendelian diseases (Frésard and Montgomery, 2018). A recent analysis shows that 14% of the recent diagnoses could be successfully performed by the combination of automatic and research approaches, looking for variants occurring in genomic regions poorly covered by exome sequencing (100,000 Genomes Project Pilot Investigators et al., 2021). Thus, the whole genome sequencing approach is becoming more relevant for the diagnosis of rare disorders (Turro et al., 2020). Accordingly, a large variety of computational approaches have been recently developed to score the impact of variants in noncoding regions (Shihab et al., 2015; Zhou and Troyanskaya, 2015; Ioannidis et al., 2016; Ionita-Laza et al., 2016; Capriotti and Fariselli, 2017; Rentzsch et al., 2019; Wells et al., 2019). In addition, for the interpretation of these potentially regulatory variants, the simultaneous and integrated use of multiple layers of omics technologies, such as whole-genome and transcriptome sequencing, is also increasingly being considered (Hasin et al., 2017; Kerr et al., 2020).

We expect that such methods will soon become the reference diagnostic tools in clinical settings. In this direction, a recent work describes approaches and discusses strategies for the diagnosis of rare and undiagnosed diseases, based on the analysis of the whole genome (Marwaha et al., 2022).

## 5 Data sharing and FAIRification

In the context of RDs, data sharing between institutions and across countries is crucial for maximising the potential of the generated genomic data (Saunders et al., 2019). It allows for the recruitment of larger cohorts of patients, thereby increasing statistical, and diagnostic, power. Sensitive RD patient data are collected by multiple institutions, whose registries are always difficult to aggregate. Sharing such data is essential for the development and maintenance of large databases, which are essential for federated analysis and discovery. In this context, the guiding principles of Findable, Accessible, Interoperable and Reusable (FAIR) data for humans and computers (Wilkinson et al., 2016) were developed, to ensure responsible sharing of health data and safeguarding of subjects. Since 2014, when "FAIR" acronym was first coined, and, because of their potential, FAIR principles have been widely endorsed by the RD community, the International Rare Diseases Research Consortium (IRDiRC) and the ELIXIR research infrastructure. In fact, adopting FAIR principles allows researchers and clinicians to integrate data from different resources in compliance with the restrictions of data accessibility, and thus answer questions involving multiple resources. For example, many types of genomic data, including features linked to the genomic coordinates of a reference genome, are always difficult to locate and access. A recent application of the FAIR principles to genomic data allowed the development of a track search service, which integrates metadata from various hubs, by adopting a set of recommendations for genomic data sharing (Gundersen et al., 2021). In addition, tools and pipelines developed for the analysis of genomic data, such as those described in this review, undoubtedly fall in the category of "research software", which is now considered part of FAIR by the European Commission. Indeed, FAIR principles can be applied not only to data, but to research software as well (Jiménez et al., 2017; Lamprecht et al., 2020).

A recent initiative, aiming at making FAIR ('FAIRification') 24 ERN (European Reference Networks) registries of RD patients, allowed collecting ninety-eight critical FAIRification challenges and proposing solutions to address them (Dos Santos Vieira et al., 2022). Awareness of the FAIRification challenges learned from initiatives like this one, which are strongly supported by the ELIXIR community, plays an important role in identifying solutions aimed at harmonizing RD data. Nevertheless, most resources collect unique data and there are wide differences in content, format, and language across them. This heterogeneity makes it virtually impossible to harmonize data from different resources, wasting the time and effort of data analysts and compromising any large-scale project aimed at improving RD research and supporting RD patients. It is therefore critical to put effort in the FAIRification process, both for humans and machines, so that data (including registries) can be queried in an unambiguous, global and federated way.

Inline with this need, the ELIXIR bio. tools portal (Ison et al., 2019) provides a comprehensive registry of software and databases that facilitates the search, understanding, use, and recognition of biomedical scientific resources. Among the 27,471 tools available on the portal, we identified 303 tools that are part of the *"Rare Diseases"* collection, domain, or topic and refer to a total of 165 functions described with the semantic terms of the EDAM ontology (Ison et al., 2013). After reviewing a list of 303 RD tools, we integrated them with other bio. tools methods, to develop a curated set of core resources for analyzing rare disease genomes.

The resources and tools collected in Supplementary Tables S2–S5 have been evaluated according to five criteria, related to their accessibility, update status, number of citations, and development stage (reported with "mature" tag in bio. tools). This type of evaluation, which assigns a score ranging from 1 to 5, represents a step toward the establishment of a standardized protocol for their clinical application.

## 6 Conclusions and future directions

Quick and accurate diagnosis are key issues for public health in general and for RDs in particular. The diagnostic delay for many RDs may at present reach up to decades (Molster et al., 2016; Heuyer et al., 2017), with an average time of about 4–5 years (Yan et al., 2020). In the journey towards diagnosis (also named the "*diagnostic odyssey*") patients may receive misdiagnosis and consequent inappropriate treatments and care. Diagnostic delay and misdiagnosis are due to many factors: RDs are infrequent, thus it is difficult to achieve a critical mass of data; data are sensitive, heterogeneous (clinical data, patient registries, variants, etc.) and usually fragmented (different communities and efforts collect data on specific RDs of interest using different formats, schemas, etc.) with poor interoperability, and a single, comprehensive repository for RDs does not exist.

In recent years, the development of new tools and resources, and the advances in data sharing practices and integrated analyses have allowed to reach an appropriate diagnosis for a sizable proportion of patients (Marwaha et al., 2022). Indeed, combining data from different sources, and using computational tools to analyze them in an integrated manner, is crucial to validate candidate variants, identify disease causative genes, perform genotype-phenotype associations, and elucidate the underlying molecular mechanism of a disease.

However, RD patients and expertise are still very scattered from each other, and knowledge and data sparsity, fragmentation, heterogeneity and poor interoperability often make integration and sharing of information extremely difficult if not impossible. Moreover, RD data are sensitive and recent technologies and practices gave rise to the further challenge of reconciling the benefits of data sharing and integration with privacy protection and ethical issues. Indeed, one of the major challenges nowadays consist in the implementation of reliable procedures for improving data sharing and the development of standardized tools and pipelines to enable reproducible research, while at the same time guaranteeing privacy rights.

To address these challenges, many international consortia have been established to create and integrate global infrastructure for RD research. At the European level, Solve-RD (solving the unsolved RDs, (Zurek et al., 2021), and RD-Connect (Lochmüller et al., 2018) enabled the creation of interdisciplinary teams to actively share and jointly analyze existing patient's data. These initiatives leverage existing computational infrastructures to share registries and standardize data among clinicians and scientists. In particular, the RD-Connect consortium promoted the development of the Genome-Phenome Analysis Platform (GPAP) (Laurie et al., 2022), and its integration with the PhenomeCentral (Osmond et al., 2022) and DECIPHER (Foreman et al., 2022). The GPAP platform facilitates the collation, discovery, sharing, and analysis of standardized genome-phenome data within a collaborative environment.

In this context, the implementation of a FAIR ecosystem of federated resources is essential for boosting research and diagnosis by decreasing RD data fragmentation and increasing data quality, with great advantages also in terms of time saving and sustainability. Although the developers and maintainers of the major RD resources and tools are already moving in the direction of FAIR data and software sharing, much still remains to be done to achieve the systematic application of FAIR principles by all players of the ecosystem, including data providers, data stewards and managers, software developers, researchers and clinicians, patients associations, research institutions, hospitals, and infrastructures. The transparent access to data and tools by the scientific community is recognized nowadays as one of the major challenges for improving RD diagnosis.

## Author contributions

EC, CC, RC, AF, PF, LL, AM, PM, SP, MP, TP, CS, PT, and AV contributed to conception and design of the study. GB, SB, CS, EC, and LV performed the statistical analysis. EC wrote the first draft of the manuscript. GB, EC, CC, RC, AF, FI, LL, AM, PM, SP, MP, TP, RR, CS, PT, AV wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## Funding

This work was supported by the following PRIN projects of the Ministero Istruzione, Università e Ricerca: "Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation" (PRIN201744NR8S), "Protein Bioinformatics for Human Health" (PRIN2017483NH8). DG is supported within the framework of CIR01_00017–"CNRBiOmics–Centro Nazionale di Ricerca in Bioinformatica per le Scienze Omiche"–Rafforzamento del capitale umano–CUP B56J20000960001.

## Acknowledgments

The authors acknowledge financial support of the Rare Disease implementation study of the European Network for Life Science ELIXIR (IT-UNIBO-2022-Strategic 202201217). AM and BS also acknowledge financial support from Banca D'Italia.

## 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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1169109/full#supplementary-material

**SUPPLEMENTARY TABLE S1**
Analysis of the disease types in the Orphanet database.

**SUPPLEMENTARY TABLE S2**
Databases of ontologies, gene/protein networks and variants.

**SUPPLEMENTARY TABLE S3**
Tools for variant calling and annotation.

**SUPPLEMENTARY TABLE S4**
Resources for gene prioritization.

**SUPPLEMENTARY TABLE S5**
Resources and tools for variant interpretation.

**SUPPLEMENTARY TABLE S6**
Performance of 4 methods in the prediction of rare disease associated variants.

**SUPPLEMENTARY DATA**
Predictions of CADD, FATHMM, PhD-SNP$^g$ and VEST4 on the dataset of 27,468 rare disease associated variants (sas-rd-202203) from 2,697 genes.

## References

100,000 Genomes Project Pilot Investigators et al., 2021 100,000 Genomes Project Pilot InvestigatorsSmedley D., Smith K. R., Martin A., Thomas E. A., McDonagh E. M., et al. (2021). 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N. Engl. J. Med.* 385, 1868–1880. doi:10.1056/NEJMoa2035790

1000 Genomes Project Consortium et al., 2015 1000 Genomes Project ConsortiumAuton A., Brooks L. D., Durbin R. M., Garrison E. P., Kang H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit7.20. doi:10.1002/0471142905.hg0720s76

Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., and Taylor, J. (2010). Galaxy CloudMan: Delivering cloud compute clusters. *BMC Bioinforma.* 11, S4. doi:10.1186/1471-2105-11-S12-S4

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi:10.1093/nar/gky379

Ahmed, Z., Renart, E. G., Mishra, D., and Zeeshan, S. (2021). JWES: A new pipeline for whole genome/exome sequence data processing, management, and gene-variant discovery, annotation, prediction, and genotyping. *FEBS Open Bio* 11, 2441–2452. doi:10.1002/2211-5463.13261

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi:10.1093/nar/gkw985

Alfares, A., Alsubaie, L., Aloraini, T., Alaskar, A., Althagafi, A., Alahmad, A., et al. (2020). What is the right sequencing approach? Solo VS extended family analysis in consanguineous populations. *BMC Med. Genomics* 13, 103. doi:10.1186/s12920-020-00743-8

All of Us Research Program InvestigatorsDenny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., et al. (2019). The "all of us" research program. *N. Engl. J. Med.* 381, 668–676. doi:10.1056/NEJMsr1809937

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's online mendelian inheritance in man (OMIM). *Nucleic Acids Res.* 37, D793–D796. doi:10.1093/nar/gkn665

An, O., Tan, K.-T., Li, Y., Li, J., Wu, C.-S., Zhang, B., et al. (2020). CSI NGS portal: An online platform for automated NGS data analysis and sharing. *Int. J. Mol. Sci.* 21, E3828. doi:10.3390/ijms21113828

Andolfo, I., Martone, S., Rosato, B. E., Marra, R., Gambale, A., Forni, G. L., et al. (2021). Complex modes of inheritance in hereditary red blood cell disorders: A case series study of 155 patients. *Genes* 12, 958. doi:10.3390/genes12070958

Azzariti, D. R., and Hamosh, A. (2020). Genomic data sharing for novel mendelian disease gene discovery: The matchmaker exchange. *Annu. Rev. Genomics Hum. Genet.* 21, 305–326. doi:10.1146/annurev-genom-083118-014915

Bao, R., Hernandez, K., Huang, L., Kang, W., Bartom, E., Onel, K., et al. (2015). ExScalibur: A high-performance cloud-enabled suite for whole exome germline and somatic mutation identification. *PloS One* 10, e0135800. doi:10.1371/journal.pone.0135800

Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi:10.1038/nrg2918

Basel-Salmon, L., Orenstein, N., Markus-Bustani, K., Ruhrman-Shahar, N., Kilim, Y., Magal, N., et al. (2019). Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 21, 1443–1451. doi:10.1038/s41436-018-0343-7

Bathke, J., and Lühken, G. (2021). OVarFlow: A resource optimized GATK 4 based open source variant calling workFlow. *BMC Bioinforma.* 22, 402. doi:10.1186/s12859-021-04317-y

Baxter, S. M., Posey, J. E., Lake, N. J., Sobreira, N., Chong, J. X., Buyske, S., et al. (2022). Centers for mendelian genomics: A decade of facilitating gene discovery. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 24, 784–797. doi:10.1016/j.gim.2021.12.005

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., et al. (2015). PathCards: Multi-source consolidation of human biological pathways. *Database J. Biol. Databases Curation* 2015, bav006. doi:10.1093/database/bav006

Benevenuta, S., Capriotti, E., and Fariselli, P. (2021). Calibrating variant-scoring methods for clinical decision making. *Bioinforma. Oxf. Engl.* 36, 5709–5711. doi:10.1093/bioinformatics/btaa943

Birgmeier, J., Haeussler, M., Deisseroth, C. A., Steinberg, E. H., Jagadeesh, K. A., Ratner, A. J., et al. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* 12, eaau9113. doi:10.1126/scitranslmed.aau9113

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., et al. (2014). Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* 15, 403. doi:10.1186/gb4161

Bodenreider, O. (2004). The unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270. doi:10.1093/nar/gkh061

Bonne, G. (2021). The Treatabolome, an emerging concept. *J. Neuromuscul. Dis.* 8, 337–339. doi:10.3233/JND-219003

Boycott, K. M., Dyment, D. A., and Innes, A. M. (2018). Unsolved recognizable patterns of human malformation: Challenges and opportunities. *Am. J. Med. Genet. C Semin. Med. Genet.* 178, 382–386. doi:10.1002/ajmg.c.31665

Boycott, K. M., Hartley, T., Biesecker, L. G., Gibbs, R. A., Innes, A. M., Riess, O., et al. (2019). A diagnosis for all rare genetic diseases: The horizon and the next Frontiers. *Cell* 177, 32–37. doi:10.1016/j.cell.2019.02.040

Buphamalai, P., Kokotovic, T., Nagy, V., and Menche, J. (2021). Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nat. Commun.* 12, 6306. doi:10.1038/s41467-021-26674-1

Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., et al. (2015). PhenomeCentral: A portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum. Mutat.* 36, 931–940. doi:10.1002/humu.22851

Cabrera-Andrade, A., López-Cortés, A., Jaramillo-Koupermann, G., Paz-Y-Miño, C., Pérez-Castillo, Y., Munteanu, C. R., et al. (2020). Gene prioritization through consensus strategy, enrichment methodologies analysis, and networking for osteosarcoma pathogenesis. *Int. J. Mol. Sci.* 21, E1053. doi:10.3390/ijms21031053

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). Mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* 10, 690–691. doi:10.1038/nmeth.2561

Calderone, A., Iannuccelli, M., Peluso, D., and Licata, L. (2020). Using the MINT database to search protein interactions. *Curr. Protoc. Bioinforma.* 69, e93. doi:10.1002/cpbi.93

Capriotti, E., and Fariselli, P. (2017). PhD-SNPg: A webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res.* 45, W247–W252. doi:10.1093/nar/gkx369

Capriotti, E., and Fariselli, P. (2022). Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants. *Hum. Genet.* 141, 1649–1658. doi:10.1007/s00439-021-02419-4

Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinforma.* 9, S6. doi:10.1186/1471-2105-9-S2-S6

Capriotti, E., Martelli, P. L., Fariselli, P., and Casadio, R. (2017). Blind prediction of deleterious amino acid variations with SNPs&GO. *Hum. Mutat.* 38, 1064–1071. doi:10.1002/humu.23179

Capriotti, E., Ozturk, K., and Carter, H. (2019). Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 11, e1443. doi:10.1002/wsbm.1443

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 14, S3. doi:10.1186/1471-2164-14-S3-S3

Ceccarelli, F., Turei, D., Gabor, A., and Saez-Rodriguez, J. (2020). Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinforma. Oxf. Engl.* 36, 2632–2633. doi:10.1093/bioinformatics/btz968

Cezard, T., Cunningham, F., Hunt, S. E., Koylass, B., Kumar, N., Saunders, G., et al. (2022). The European variation archive: A FAIR resource of genomic variation for all species. *Nucleic Acids Res.* 50, D1216–D1220. doi:10.1093/nar/gkab960

Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G. (2007). Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinforma.* 8, 392. doi:10.1186/1471-2105-8-392

Chen, J., Aronow, B. J., and Jegga, A. G. (2009a). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinforma.* 10, 73. doi:10.1186/1471-2105-10-73

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009b). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi:10.1093/nar/gkp427

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., et al. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* 20, 48. doi:10.1186/s13059-019-1653-z

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., et al. (2020). A brief history of human disease genetics. *Nature* 577, 179–189. doi:10.1038/s41586-019-1879-7

Clerc, O., Deniaud, M., Vallet, S. D., Naba, A., Rivet, A., Perez, S., et al. (2019). MatrixDB: Integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res.* 47, D376–D381. doi:10.1093/nar/gky1035

de la Fuente, L., Del Pozo-Valero, M., Perea-Romero, I., Blanco-Kelly, F., Fernández-Caballero, L., Cortón, M., et al. (2023). Prioritization of new candidate genes for rare genetic diseases by a disease-aware evaluation of heterogeneous molecular networks. *Int. J. Mol. Sci.* 24, 1661. doi:10.3390/ijms24021661

De Las Rivas, J., and Fontanillo, C. (2012). Protein-protein interaction networks: Unraveling the wiring of molecular machines within the cell. *Brief. Funct. Genomics* 11, 489–496. doi:10.1093/bfgp/els036

De Marinis, I., Lo Surdo, P., Cesareni, G., and Perfetto, L. (2021). SIGNORApp: A Cytoscape 3 application to access SIGNOR data. *Bioinforma. Oxf. Engl.* 38, 1764–1766. btab865. doi:10.1093/bioinformatics/btab865

Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinforma.* 12, 151. doi:10.1186/1471-2105-12-151

Del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., et al. (2022). The IntAct database: Efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* 50, D648–D653. doi:10.1093/nar/gkab1006

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human splicing finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37, e67. doi:10.1093/nar/gkp215

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi:10.1038/nbt.3820

Dos Santos Vieira, B., Bernabé, C. H., Zhang, S., Abaza, H., Benis, N., Cámara, A., et al. (2022). Towards FAIRification of sensitive and fragmented rare disease patient data: Challenges and solutions in European reference network registries. *Orphanet J. Rare Dis.* 17, 436. doi:10.1186/s13023-022-02558-5

Edgar, R., Mazor, Y., Rinon, A., Blumenthal, J., Golan, Y., Buzhor, E., et al. (2013). LifeMap Discovery$^{TM}$: The embryonic development, stem cells, and regenerative medicine research portal. *PloS One* 8, e66629. doi:10.1371/journal.pone.0066629

Eldomery, M. K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J. A., Gambin, T., Stray-Pedersen, A., et al. (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med.* 9, 26. doi:10.1186/s13073-017-0412-6

Ellingford, J. M., Ahn, J. W., Bagnall, R. D., Baralle, D., Barton, S., Campbell, C., et al. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* 14, 73. doi:10.1186/s13073-022-01073-3

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 38, 276–278. doi:10.1038/s41587-020-0439-x

Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* 48, D941–D947. doi:10.1093/nar/gkz836

Fariselli, P., Martelli, P. L., Savojardo, C., and Casadio, R. (2015). INPS: Predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinforma. Oxf. Engl.* 31, 2816–2821. doi:10.1093/bioinformatics/btv291

Ferreira, C. R. (2019). The burden of rare diseases. *Am. J. Med. Genet. A* 179, 885–892. doi:10.1002/ajmg.a.61124

Foreman, J., Brent, S., Perrett, D., Bevan, A. P., Hunt, S. E., Cunningham, F., et al. (2022). DECIPHER: Supporting the interpretation and sharing of rare disease phenotype-linked variant data to advance diagnosis and research. *Hum. Mutat.* 43, 682–697. doi:10.1002/humu.24340

Freeberg, M. A., Fromont, L. A., D'Altri, T., Romero, A. F., Ciges, J. I., Jene, A., et al. (2022). The European genome-phenome archive in 2021. *Nucleic Acids Res.* 50, D980–D987. doi:10.1093/nar/gkab1059

French, C. E., Dolling, H., Mégy, K., Sanchis-Juan, A., Kumar, A., Delon, I., et al. (2022). Refinements and considerations for trio whole-genome sequence analysis when investigating Mendelian diseases presenting in early childhood. *HGG Adv.* 3, 100113. doi:10.1016/j.xhgg.2022.100113

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

Gabriel, H., Korinth, D., Ritthaler, M., Schulte, B., Battke, F., von Kaisenberg, C., et al. (2022). Trio exome sequencing is highly relevant in prenatal diagnostics. *Prenat. Diagn.* 42, 845–851. doi:10.1002/pd.6081

Gainotti, S., Torreri, P., Wang, C. M., Reihs, R., Mueller, H., Heslop, E., et al. (2018). The RD-connect registry and biobank finder: A tool for sharing aggregated data and metadata among rare disease researchers. *Eur. J. Hum. Genet.* 26, 631–643. doi:10.1038/s41431-017-0085-z

Gao, X., Xu, J., and Starmer, J. (2015). Fastq2vcf: A concise and transparent pipeline for whole-exome sequencing data analyses. *BMC Res. Notes* 8, 72. doi:10.1186/s13104-015-1027-x

Ghosh, R., Oak, N., and Plon, S. E. (2017). Evaluation of *in silico* algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol.* 18, 225. doi:10.1186/s13059-017-1353-5

Gill, N., Singh, S., and Aseri, T. C. (2014). Computational disease gene prioritization: An appraisal. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 21, 456–465. doi:10.1089/cmb.2013.0158

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U A* 104, 8685–8690. doi:10.1073/pnas.0701361104

Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., et al. (2022). Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* 43, 1012–1030. doi:10.1002/humu.24309

Gundersen, S., Boddu, S., Capella-Gutierrez, S., Drabløs, F., Fernández, J. M., Kompova, R., et al. (2021). Recommendations for the FAIRification of genomic track metadata. *F1000Research* 10, ELIXIR–268. doi:10.12688/f1000research.28449.1

Guo, Y., Ding, X., Shen, Y., Lyon, G. J., and Wang, K. (2015). SeqMule: Automated pipeline for analysis of human exome/genome sequencing data. *Sci. Rep.* 5, 14283. doi:10.1038/srep14283

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., et al. (2020). How many rare diseases are there? *Nat. Rev. Drug Discov.* 19, 77–78. doi:10.1038/d41573-019-00180-y

Hartin, S. N., Means, J. C., Alaimo, J. T., and Younger, S. T. (2020). Expediting rare disease diagnosis: A call to bridge the gap between clinical and functional genomics. *Mol. Med. Camb. Mass* 26, 117. doi:10.1186/s10020-020-00244-5

Hartley, T., Balcı, T. B., Rojas, S. K., Eaton, A., Canada, C. R., Dyment, D. A., et al. (2018). The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM®. *Am. J. Med. Genet. C Semin. Med. Genet.* 178, 458–463. doi:10.1002/ajmg.c.31662

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

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83. doi:10.1186/s13059-017-1215-1

Hayashi, S., and Umeda, T. (2008). 35 years of Japanese policy on rare diseases. *Lancet lond. Engl.* 372, 889–890. doi:10.1016/S0140-6736(08)61393-8

Heuyer, T., Pavan, S., and Vicard, C. (2017). The health and life path of rare disease patients: Results of the 2015 French barometer. *Patient Relat. Outcome Meas.* 8, 97–110. doi:10.2147/PROM.S131033

humsavar UniProt (2023). *UniProt humsavar*. Available at: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/variants/humsavar.txt (Accessed Jan, 2023).

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., et al. (2018). Ensembl variation resources. *Database J. Biol. Databases Curation* 2018, bay119. doi:10.1093/database/bay119

IMEx Consortium CuratorsDel-Toro, N., Duesbury, M., Koch, M., Perfetto, L., Shrivastava, A., et al. (2019). Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. *Nat. Commun.* 10, 10. doi:10.1038/s41467-018-07709-6

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., et al. (2016). Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99, 877–885. doi:10.1016/j.ajhg.2016.08.016

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48, 214–220. doi:10.1038/ng.3477

Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., et al. (2013). EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinforma. Oxf. Engl.* 29, 1325–1332. doi:10.1093/bioinformatics/btt113

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., et al. (2019). The bio.tools registry of software tools and data resources for the life sciences. *Genome Biol.* 20, 164. doi:10.1186/s13059-019-1772-6

Jacobsen, J. O. B., Kelly, C., Cipriani, V., Research Consortium, G. E., Mungall, C. J., Reese, J., et al. (2022). Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. *Hum. Mutat.* 43, 1071–1081. doi:10.1002/humu.24380

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548. doi:10.1016/j.cell.2018.12.015

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi:10.1093/nar/gkz1031

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., et al. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research* 6, ELIXIR-876. doi:10.12688/f1000research.11407.1

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T.-K., Lua, R. C., Wilkins, A. D., et al. (2014). Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Sci. Publ. Protein Soc.* 23, 1650–1666. doi:10.1002/pro.2552

Katsonis, P., Wilhelm, K., Williams, A., and Lichtarge, O. (2022). Genome interpretation using *in silico* predictors of variant impact. *Hum. Genet.* 141, 1549–1577. doi:10.1007/s00439-022-02457-6

Kerr, K., McAneney, H., Smyth, L. J., Bailie, C., McKee, S., and McKnight, A. J. (2020). A scoping review and proposed workflow for multi-omic rare disease research. *Orphanet J. Rare Dis.* 15, 107. doi:10.1186/s13023-020-01376-x

Kinjo, S., Monma, N., Misu, S., Kitamura, N., Imoto, J., Yoshitake, K., et al. (2018). Maser: One-stop platform for NGS big data from analysis to visualization. *Database J. Biol. Databases Curation* 2018, bay027. doi:10.1093/database/bay027

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Med.* 12, 91. doi:10.1186/s13073-020-00791-w

Köster, J., and Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi:10.1093/bioinformatics/bts480

Kutmon, M., Lotia, S., Evelo, C. T., and Pico, A. R. (2014). WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research* 3, 152. doi:10.12688/f1000research.4254.2

Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015). MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinforma.* 16, 116. doi:10.1186/s12859-015-0548-6

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, , R., Martin Del Pico, E., et al. (2020). Towards FAIR principles for research software. *Data Sci.* 3, 37–59. doi:10.3233/DS-190026

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., et al. (2020). ClinVar: Improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844. doi:10.1093/nar/gkz972

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., et al. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res.* 41, D936–D941. doi:10.1093/nar/gks1213

Laurie, S., Piscia, D., Matalonga, L., Corvó, A., Fernández-Callejo, M., Garcia-Linares, C., et al. (2022). The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. *Hum. Mutat.* 43, 717–733. doi:10.1002/humu.24353

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi:10.1038/nature19057

Li, Q., and Wang, K. (2017). InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.* 100, 267–280. doi:10.1016/j.ajhg.2017.01.004

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604. doi:10.1126/science.aad9417

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. Off. J. Am. Coll. Med. Genet.* 21, 2126–2134. doi:10.1038/s41436-019-0439-8

Liu, X., Yang, Z., Lin, H., Simmons, M., and Lu, Z. (2017). DIGNiFI: Discovering causative genes for orphan diseases using protein-protein interaction networks. *BMC Syst. Biol.* 11, 23. doi:10.1186/s12918-017-0402-8

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

Lo Surdo, P., Iannuccelli, M., Contino, S., Castagnoli, L., Licata, L., Cesareni, G., et al. (2023). SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res.* 51, D631–D637. doi:10.1093/nar/gkac883

Lochmüller, H., Badowska, D. M., Thompson, R., Knoers, N. V., Aartsma-Rus, A., Gut, I., et al. (2018). RD-connect, NeurOmics and EURenOmics: Collaborative European initiative for rare diseases. *Eur. J. Hum. Genet. EJHG* 26, 778–785. doi:10.1038/s41431-018-0115-5

Manfredi, M., Savojardo, C., Martelli, P. L., and Casadio, R. (2022). E-SNPs&GO: Embedding of protein sequence and function improves the annotation of human pathogenic variants. *Bioinforma. Oxf. Engl.* 38, 5168–5174. doi:10.1093/bioinformatics/btac678

Marabotti, A., Scafuri, B., and Facchiano, A. (2020). Predicting the stability of mutant proteins by computational approaches: An overview. *Brief. Bioinform.* 22, bbaa074. doi:10.1093/bib/bbaa074

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., et al. (2021). WikiPathways: Connecting communities. *Nucleic Acids Res.* 49, D613–D621. doi:10.1093/nar/gkaa1024

Marwaha, S., Knowles, J. W., and Ashley, E. A. (2022). A guide for the diagnosis of rare and undiagnosed disease: Beyond the exome. *Genome Med.* 14, 23. doi:10.1186/s13073-022-01026-w

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347, 1257601. doi:10.1126/science.1257601

Molster, C., Urwin, D., Di Pietro, L., Fookes, M., Petrie, D., van der Laan, S., et al. (2016). Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet J. Rare Dis.* 11, 30. doi:10.1186/s13023-016-0409-z

Monaco, L., Zanello, G., Baynam, G., Jonker, A. H., Julkowska, D., Hartman, A. L., et al. (2022). Research on rare diseases: Ten years of progress and challenges at IRDiRC. *Nat. Rev. Drug Discov.* 21, 319–320. doi:10.1038/d41573-022-00019-z

Montanucci, L., Capriotti, E., Birolo, G., Benevenuta, S., Pancotti, C., Lal, D., et al. (2022). DDGun: An untrained predictor of protein stability changes upon amino acid variants. *Nucleic Acids Res.* 50, W222–W227. gkac325. doi:10.1093/nar/gkac325

Moreau, Y., and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat. Rev. Genet.* 13, 523–536. doi:10.1038/nrg3253

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

Nicora, G., Zucca, S., Limongelli, I., Bellazzi, R., and Magni, P. (2022). A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Sci. Rep.* 12, 2517. doi:10.1038/s41598-022-06547-3

Niroula, A., Urolagin, S., and Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PloS One* 10, e0117380. doi:10.1371/journal.pone.0117380

O'Connor, B. D., Yuen, D., Chung, V., Duncan, A. G., Liu, X. K., Patricia, J., et al. (2017). The Dockstore: Enabling modular, community-focused sharing of docker-based genomics tools and workflows. *F1000Research* 6, 52. doi:10.12688/f1000research.10137.1

Osmond, M., Hartley, T., Johnstone, B., Andjic, S., Girdea, M., Gillespie, M., et al. (2022). PhenomeCentral: 7 years of rare disease matchmaking. *Hum. Mutat.* 43, 674–681. doi:10.1002/humu.24348

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.* 30, 187–200. doi:10.1002/pro.3978

Özkan, S., Padilla, N., Moles-Fernández, A., Diez, O., Gutiérrez-Enríquez, S., and de la Cruz, X. (2021). "Chapter 6 - the computational approach to variant interpretation: Principles, results, and applicability," in *Clinical DNA variant interpretation. Translational and applied genomics*. Editors C. Lázaro, J. Lerner-Ellis, and A. Spurdle (Academic Press), 89–119. doi:10.1016/B978-0-12-820519-8.00007-7

Paila, U., Chapman, B. A., Kirchner, R., and Quinlan, A. R. (2013). GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* 9, e1003153. doi:10.1371/journal.pcbi.1003153

Paine, I., Posey, J. E., Grochowski, C. M., Jhangiani, S. N., Rosenheck, S., Kleyner, R., et al. (2019). Paralog studies augment gene discovery: DDX and DHX genes. *Am. J. Hum. Genet.* 105, 302–316. doi:10.1016/j.ajhg.2019.06.001

Pais, L. S., Snow, H., Weisburd, B., Zhang, S., Baxter, S. M., DiTroia, S., et al. (2022). seqr: A web-based analysis and collaboration tool for rare disease genomics. *Hum. Mutat.* 43, 698–707. doi:10.1002/humu.24366

Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., et al. (2022). Predicting protein stability changes upon single-point mutation: A thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* 23, bbab555. doi:10.1093/bib/bbab555

Pastrello, C., Kotlyar, M., and Jurisica, I. (2020). Informed use of protein-protein interaction data: A focus on the integrated interactions database (IID). *Methods Mol. Biol. Clifton N. J.* 2074, 125–134. doi:10.1007/978-1-4939-9873-9_10

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., et al. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* 11, 5918. doi:10.1038/s41467-020-19669-x

Petrosino, M., Novak, L., Pasquo, A., Chiaraluce, R., Turina, P., Capriotti, E., et al. (2021). Analysis and interpretation of the impact of missense variants in cancer. *Int. J. Mol. Sci.* 22, 5416. doi:10.3390/ijms22115416

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Pires, D. E. V., Rodrigues, C. H. M., and Ascher, D. B. (2020). mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res.* 48, W147–W153. doi:10.1093/nar/gkaa416

Piro, R. M., and Di Cunto, F. (2012). Computational approaches to disease-gene prediction: Rationale, classification and successes. *FEBS J.* 279, 678–696. doi:10.1111/j.1742-4658.2012.08471.x

Pogue, R. E., Cavalcanti, D. P., Shanker, S., Andrade, R. V., Aguiar, L. R., de Carvalho, J. L., et al. (2018). Rare genetic diseases: Update on diagnosis, treatment and online resources. *Drug Discov. Today* 23, 187–195. doi:10.1016/j.drudis.2017.11.002

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

Porras, P., Barrera, E., Bridge, A., Del-Toro, N., Cesareni, G., Duesbury, M., et al. (2020). Towards a unified open access dataset of molecular interactions. *Nat. Commun.* 11, 6144. doi:10.1038/s41467-020-19942-z

Quan, L., Lv, Q., and Zhang, Y. (2016). STRUM: Structure-based prediction of protein stability changes upon single-point mutation. *Bioinforma. Oxf. Engl.* 32, 2936–2946. doi:10.1093/bioinformatics/btw361

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

Quinodoz, M., Peter, V. G., Cisarova, K., Royer-Bertrand, B., Stenson, P. D., Cooper, D. N., et al. (2022). Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am. J. Hum. Genet.* 109, 457–470. doi:10.1016/j.ajhg.2022.01.006

Ragueneau, E., Shrivastava, A., Morris, J. H., Del-Toro, N., Hermjakob, H., and Porras, P. (2021). IntAct App: A Cytoscape application for molecular interaction network visualization and analysis. *Bioinforma. Oxf. Engl.* 37, 3684–3685. doi:10.1093/bioinformatics/btab319

Raimondi, D., Tanyalcin, I., Ferté, J., Gazzo, A., Orlando, G., Lenaerts, T., et al. (2017). DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* 45, W201–W206. doi:10.1093/nar/gkx390

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 45, D877–D887. doi:10.1093/nar/gkw1012

Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: The Orphanet approach to serve a wide range of end users. *Hum. Mutat.* 33, 803–808. doi:10.1002/humu.22078

Regulation Orphan Medicinal Product (2000). *Regulation (EC) No 141/2000 of the European parliament and of the council of 16 december 1999 on orphan medicinal products*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32000R0141.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., et al. (2015). ClinGen--the clinical genome resource. *N. Engl. J. Med.* 372, 2235–2242. doi:10.1056/NEJMsr1406261

Reiter, T., Brooks, P. T., Irber, L., Joslin, S. E. K., Reid, C. M., Scott, C., et al. (2021). Streamlining data-intensive biology with workflow systems. *GigaScience* 10, giaa140. doi:10.1093/gigascience/giaa140

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi:10.1093/nar/gky1016

Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* 13, 31. doi:10.1186/s13073-021-00835-9

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). 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. Off. J. Am. Coll. Med. Genet.* 17, 405–424. doi:10.1038/gim.2015.30

Robinson, P. N., Kohler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83, 610–615. doi:10.1016/j.ajhg.2008.09.017

Robinson, P. N., Kohler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C. J., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* 24, 340–348. doi:10.1101/gr.160325.113

Robinson, P. N., Ravanmehr, V., Jacobsen, J. O. B., Danis, D., Zhang, X. A., Carmody, L. C., et al. (2020). Interpretable clinical genomics with a likelihood ratio paradigm. *Am. J. Hum. Genet.* 107, 403–417. doi:10.1016/j.ajhg.2020.06.021

Rogers, F. B. (1963). Medical subject headings. *Bull. Med. Libr. Assoc.* 51, 114–116.

Rojano, E., Seoane, P., Ranea, J. A. G., and Perkins, J. R. (2019). Regulatory variants: From detection to predicting impact. *Brief. Bioinform* 20, 1639–1654. doi:10.1093/bib/bby039

Rother, K., Potrzebowski, W., Puton, T., Rother, M., Wywial, E., and Bujnicki, J. M. (2012). A toolbox for developing bioinformatics software. *Brief. Bioinform.* 13, 244–257. doi:10.1093/bib/bbr035

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi:10.1093/nar/gkh086

Sandmann, S., Karimi, M., de Graaf, A. O., Rohde, C., Göllner, S., Varghese, J., et al. (2018). appreci8: a pipeline for precise variant calling integrating 8 tools. *Bioinforma. Oxf. Engl.* 34, 4205–4212. doi:10.1093/bioinformatics/bty518

Saunders, G., Baudis, M., Becker, R., Beltran, S., Béroud, C., Birney, E., et al. (2019). Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* 20, 693–701. doi:10.1038/s41576-019-0156-9

Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). INPS-MD: A web server to predict stability of protein variants from sequence and structure. *Bioinforma. Oxf. Engl.* 32, 2542–2544. doi:10.1093/bioinformatics/btw192

Savojardo, C., Baldazzi, D., Babbi, G., Martelli, P. L., and Casadio, R. (2022). Mapping human disease-associated enzymes into Reactome allows characterization of disease groups and their interactions. *Sci. Rep.* 12, 17963. doi:10.1038/s41598-022-22818-5

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., et al. (2022). Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics lab-space. *Cell Genomics* 2, 100085. doi:10.1016/j.xgen.2021.100085

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., et al. (2012). Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Res.* 40, D940–D946. doi:10.1093/nar/gkr972

Scotti, M. M., and Swanson, M. S. (2016). RNA mis-splicing in disease. *Nat. Rev. Genet.* 17, 19–32. doi:10.1038/nrg.2015.3

Setty, S. T., Scott-Boyer, M.-P., Cuppens, T., and Droit, A. (2022). New developments and possibilities in reanalysis and reinterpretation of whole exome sequencing datasets for unsolved rare diseases using machine learning approaches. *Int. J. Mol. Sci.* 23, 6792. doi:10.3390/ijms23126792

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., et al. (2020). The Monarch initiative in 2019: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 48, D704–D715. doi:10.1093/nar/gkz997

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., et al. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31, 1536–1543. doi:10.1093/bioinformatics/btv009

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457. doi:10.1093/nar/gks539

Smedley, D., and Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* 7, 81. doi:10.1186/s13073-015-0199-2

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., et al. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *Am. J. Hum. Genet.* 99, 595–606. doi:10.1016/j.ajhg.2016.07.005

Sobreira, N. L. M., Arachchi, H., Buske, O. J., Chong, J. X., Hutton, B., Foreman, J., et al. (2017). Matchmaker exchange. *Curr. Protoc. Hum. Genet.* 95, 9.31.1–9.31.15. doi:10.1002/cphg.50

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., et al. (2023). The NHGRI-EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Res.* 51, D977–D985. doi:10.1093/nar/gkac1010

Stein, A., Fowler, D. M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and mechanistic models for disease-causing protein variants. *Trends biochem. Sci.* 44, 575–588. doi:10.1016/j.tibs.2019.01.003

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., et al. (2016a). VarElect: The phenotype-based variation prioritizer of the GeneCards suite. *BMC Genomics* 17, 444. doi:10.1186/s12864-016-2722-2

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016b). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* 54, 1.30.1–1.30.33. doi:10.1002/cpbi.5

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., et al. (2020). The human gene mutation database (HGMD$^{®}$): Optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* 139, 1197–1207. doi:10.1007/s00439-020-02199-3

Strande, N. T., Riggs, E. R., Buchanan, A. H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S. S., et al. (2017). Evaluating the clinical validity of gene-disease associations: An evidence-based framework developed by the clinical genome resource. *Am. J. Hum. Genet.* 100, 895–906. doi:10.1016/j.ajhg.2017.04.015

Summers, K. M. (1996). Relationship between genotype and phenotype in monogenic diseases: Relevance to polygenic diseases. *Hum. Mutat.* 7, 283–293. doi:10.1002/(SICI)1098-1004(1996)7:4<283::AID-HUMU1>3.0.CO;2-A

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Tabarini, N., Biagi, E., Uva, P., Iovino, E., Pippucci, T., Seri, M., et al. (2022). Exploration of tools for the interpretation of human non-coding variants. *Int. J. Mol. Sci.* 23, 12977. doi:10.3390/ijms232112977

Tavtigian, S. V., Harrison, S. M., Boucher, K. M., and Biesecker, L. G. (2020). Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum. Mutat.* 41, 1734–1737. doi:10.1002/humu.24088

Thouvenot, P., Ben Yamin, B., Fourrière, L., Lescure, A., Boudier, T., Del Nery, E., et al. (2016). Functional assessment of genetic variants with outcomes adapted to clinical decision-making. *PLoS Genet.* 12, e1006096. doi:10.1371/journal.pgen.1006096

Tran, L., Hamp, T., and Rost, B. (2018). ProfPPIdb: Pairs of physical protein-protein interactions predicted for entire proteomes. *PloS One* 13, e0199988. doi:10.1371/journal.pone.0199988

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967. doi:10.1038/nmeth.4077

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., et al. (2018). The 100 000 genomes project: Bringing whole genome sequencing to the NHS. *BMJ* 361, k1687. doi:10.1136/bmj.k1687

Turro, E., Astle, W. J., Megy, K., Gräf, S., Greene, D., Shamardina, O., et al. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* 583, 96–102. doi:10.1038/s41586-020-2434-2

U.S. Food and Drug Administration (2022). Medical products for rare diseases and conditions. Available at: https://www.fda.gov/industry/medical-products-rare-diseases-and-conditions (Accessed Jan, 2023).

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419. doi:10.1126/science.1260419

UK10K ConsortiumWalter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526, 82–90. doi:10.1038/nature14962

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., et al. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* 10, 5241. doi:10.1038/s41467-019-13212-3

Wenger, A. M., Guturu, H., Bernstein, J. A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 19, 209–214. doi:10.1038/gim.2016.88

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

World Health Organization (2019). *International classification of diseases (ICD)*. Available at: https://www.who.int/standards/classifications/classification-of-diseases (Accessed Jan, 2023).

Worth, C. L., Preissner, R., and Blundell, T. L. (2011). SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* 39, W215–W222. doi:10.1093/nar/gkr363

Yan, X., He, S., and Dong, D. (2020). Determining how far an adult rare disease patient needs to travel for a definitive diagnosis: A cross-sectional examination of the 2018 national rare disease survey in China. *Int. J. Environ. Res. Public. Health* 17, E1757. doi:10.3390/ijerph17051757

Yang, Y., Urolagin, S., Niroula, A., Ding, X., Shen, B., and Vihinen, M. (2018). PON-tstab: Protein variant stability predictor. Importance of training data quality. *Int. J. Mol. Sci.* 19, 1009. doi:10.3390/ijms19041009

Yuan, X., Wang, J., Dai, B., Sun, Y., Zhang, K., Chen, F., et al. (2022). Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief. Bioinform.* 23, bbac019. doi:10.1093/bib/bbac019

Zhang, P., and Itan, Y. (2019). Biological network approaches and applications in rare disease studies. *Genes* 10, 797. doi:10.3390/genes10100797

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

Zhu, C., Kushwaha, A., Berman, K., and Jegga, A. G. (2012). A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Syst. Biol.* 6, S8. doi:10.1186/1752-0509-6-S3-S8

Zhu, Q., Nguyen, D.-T., Sid, E., and Pariser, A. (2020). Leveraging the UMLS as a data standard for rare disease data normalization and harmonization. *Methods Inf. Med.* 59, 131–139. doi:10.1055/s-0040-1718940

Zolotareva, O., and Kleine, M. (2019). A survey of gene prioritization tools for mendelian and complex human diseases. *J. Integr. Bioinforma.* 16, 20180069. doi:10.1515/jib-2018-0069

Zurek, B., Ellwanger, K., Vissers, L. E. L. M., Schüle, R., Synofzik, M., Töpf, A., et al. (2021). Solve-RD: Systematic pan-European data sharing and collaborative analysis to solve rare diseases. *Eur. J. Hum. Genet. EJHG* 29, 1325–1331. doi:10.1038/s41431-021-00859-0