



# Strategies to Uplift Novel Mendelian Gene Discovery for Improved Clinical Outcomes

Eleanor G. Seaby<sup>1,2,3,4\*</sup>, Heidi L. Rehm<sup>1,3</sup> and Anne O'Donnell-Luria<sup>1,3,4,5</sup>

<sup>1</sup> Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, United States,

<sup>2</sup> Genomic Informatics Group, University Hospital Southampton, Southampton, United Kingdom, <sup>3</sup> Center for Genomic Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, United States, <sup>4</sup> Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, United States, <sup>5</sup> Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, United States

Rare genetic disorders, while individually rare, are collectively common. They represent some of the most severe disorders affecting patients worldwide with significant morbidity and mortality. Over the last decade, advances in genomic methods have significantly uplifted diagnostic rates for patients and facilitated novel and targeted therapies. However, many patients with rare genetic disorders still remain undiagnosed as the genetic etiology of only a proportion of Mendelian conditions has been discovered to date. This article explores existing strategies to identify novel Mendelian genes and how these discoveries impact clinical care and therapeutics. We discuss the importance of data sharing, phenotype-driven approaches, patient-led approaches, utilization of large-scale genomic sequencing projects, constraint-based methods, integration of multi-omics data, and gene-to-patient methods. We further consider the health economic advantages of novel gene discovery and speculate on potential future methods for improved clinical outcomes.

**Keywords:** Mendelian, novel gene discovery, disease–gene relationships, rare genetic disorders, genomics, rare disease, genetics

## OPEN ACCESS

### Edited by:

Gavin R. Oliver,  
Mayo Clinic, United States

### Reviewed by:

Jamie Ellingford,  
The University of Manchester,  
United Kingdom  
Lisa Bastarache,  
Vanderbilt University Medical Center,  
United States

### \*Correspondence:

Eleanor G. Seaby  
eseaby@broadinstitute.org

### Specialty section:

This article was submitted to  
Human and Medical Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 01 March 2021

**Accepted:** 12 May 2021

**Published:** 17 June 2021

### Citation:

Seaby EG, Rehm HL and  
O'Donnell-Luria A (2021) Strategies  
to Uplift Novel Mendelian Gene  
Discovery for Improved Clinical  
Outcomes. *Front. Genet.* 12:674295.  
doi: 10.3389/fgene.2021.674295

## INTRODUCTION

Rare genetic disorders affect 1-in-17 individuals in their lifetime (Turnbull et al., 2018). They encompass some of the most severe disorders affecting patients worldwide, including childhood cancers, neurodevelopmental disorders, and muscle diseases to name a few (Simpson, 2016). Many are severe and life limiting, with significant morbidity and mortality. Indeed, 30% of children with rare diseases die before their fifth birthdays (Rode, 2005). Many affected patients are wheelchair bound and require respiratory support, feeding support, specialized community services, and significant hospitalizations (Yoon et al., 1997; Dodge et al., 2011). This not only impacts the patients involved but their caregivers and families as well.

Approximately 80% of rare diseases have a genetic basis, yet many of the underlying genes have not yet been identified, nor has the wide spectrum of pathogenic variation been delineated for each gene (Yoon et al., 1997; Dodge et al., 2011; Wright et al., 2018). As such, on average across all specialties, the causal variant(s) are only identified in ~30–40% of rare disease patients, leaving the majority of patients and their families without a reliable prognosis, rendering medical care largely

supportive and palliative (Retterer et al., 2016; Adams and Eng, 2018; Clark et al., 2018; Srivastava et al., 2019).

One of the biggest challenges in reaching a molecular diagnosis is the paucity of scientific knowledge into the biological function of all ~20,000 human genes. Indeed, the disease phenotypes have yet to be discovered for ~50% of those genes with a genomic signature suggesting haploinsufficiency (Seaby and Ennis, 2020). Therefore, diagnosing rare diseases is extremely challenging without a prior correlation between a clinical phenotype and causative gene. New gene disorders preclude detection when for every genome sequenced, millions of variants of uncertain significance reside in genes of unknown function (Smedley et al., 2015; Karczewski et al., 2020). Even the best computational methods available at present will typically overlook a gene of undetermined biological significance when analyzing a patient's exome (restricted to protein-coding regions of DNA) or genome (all DNA regions). Therefore, new rare genetic diseases will be overlooked until further studies are undertaken or new methods are developed to uplift novel gene discovery (Seaby and Ennis, 2020).

## WHY IS NOVEL MENDELIAN GENE DISCOVERY IMPORTANT?

The significance of uplifting novel Mendelian gene discovery is not to be underestimated. Every new disease gene discovered goes toward ending the notorious “diagnostic odyssey” of rare disease. This pertains to rare disease patients who move between specialties and undergo myriad diagnostic tests in search for a unifying genetic explanation (Thevenon et al., 2016). For most, these often expensive evaluations only elucidate the clinical phenotype and seldom aid in diagnosis. In the United Kingdom, over a 10-year period, undiagnosed rare diseases have cost NHS England an average of £13,064 (US\$18,279) per patient and in excess of £3.4 billion (US\$4.8 billion) in total (Imperial College Health Partners, 2018). In Australia, the cost per diagnosis using standard care is AU\$27,050 (US\$21,241), and in the United States, the same cost basis was calculated at US\$19,100 (Soden et al., 2014; Stark et al., 2017). While these figures all showcase the cost burden of rare diseases, it is ill advised to compare cost evaluations between countries due to differing healthcare systems.

Novel discoveries directly impact diagnostic potential. Diagnoses not only provide answers for patients and families but have far-reaching clinical impact, including but not limited to guiding personalized treatments; offering patient support networks; collecting and gaining knowledge on disease trajectory and prognosis; enabling participation in research studies; informing reproductive choices; and impacting the health of relatives. Even when little can be done therapeutically following diagnosis, the importance of that diagnosis to patients and families should not be overlooked; when a cause is identified, this often alleviates guilt and blame felt by patients and families who believe a given rare disease is their fault (Muir, 2016).

Novel gene discovery is critical in the research space to expand biological understanding of human genes and variation and to

identify therapeutic drug targets that may lead to successful and life-altering therapies (Legendre et al., 2013; Ribeil et al., 2017; Mirmiran et al., 2019). Gene augmentation therapies have been developed for a number of conditions, for example, subretinal injection of adeno-associated virus vectors to deliver *RPE65* cDNA to treat Leber congenital amaurosis (MIM: 204100) (Maguire et al., 2009; Pierce and Bennett, 2015) and the FDA-approved one-time intravenous administration of *SMN* cDNA to treat spinal muscular atrophy type 1 (MIM: 253300) (Mendell et al., 2017; Beck et al., 2020). Small molecular therapies for cystic fibrosis (MIM: 602421) are well studied and include ivacaftor, which increases the time fraction that the cystic fibrosis transmembrane conductance regulator (CFTR) channel remains open, and lumacaftor, which increases the amount of CFTR that reaches the cell surface (Wainwright et al., 2015; Habib et al., 2019). Development of antisense oligonucleotides is proving effective in preclinical and clinical studies to treat neurodegenerative diseases (Smith et al., 2006; Ly and Miller, 2018); and it is hoped that identification of novel disease genes may guide further protein targets.

In contrast to the development of new gene therapies, it is not uncommon for existing therapies to be repurposed when knowledge of a given gene and biological pathway is implicated in diseases. For example, in 2011, autosomal recessive variants in *MTHFD1* (a gene involved in folate metabolism) were found to cause combined immunodeficiency and megaloblastic anemia with or without hyperhomocysteinemia (MIM: 617780) (Watkins et al., 2011). Simple folic acid has proven life changing for patients with recessive mutations in *MTHFD1* (Burda et al., 2015; Ramakrishnan et al., 2016).

## THE PRIOR DECADE OF NOVEL GENE DISCOVERY

Since the advent of next-generation sequencing (NGS) technologies, there has been a stepwise acceleration in novel gene discovery leading to uplifted diagnostic rates for rare disease patients (Chong et al., 2015; Posey et al., 2019). Between 2005 and 2009, there were ~170 novel discoveries per year. This is compared to ~240 per year between 2010 and 2014 when NGS became widely adopted (Chong et al., 2015). In the history of disease–gene relationship discovery, NGS approaches are responsible for ~36% of all reported Mendelian disease genes. Their contribution to novel gene discoveries is accelerating, with 87% of new gene disorders now discovered using NGS approaches (Bamshad et al., 2019). Novel discoveries are still progressing, although the pace of discovery appears to have reached a steady state that balances the time required to build international cohorts, undertake functional experiments, and publish findings (Posey et al., 2019). Despite this, approximately 250 new genes are added to the literature annually, and a recent review predicted that more than 6,000 Mendelian conditions remain to be discovered (Bamshad et al., 2019). Therefore, with thousands of monogenic disease–gene relationships yet to be elucidated, there is clear evidence that the recognition of disease-causing variation in the exome is far

from saturated (Posey et al., 2019). For more data on the pace of novel discoveries, we recommend the excellent review by Bamshad et al. (2019).

## TRADITIONAL FAMILY-BASED APPROACHES TO DIAGNOSIS AND NOVEL GENE DISCOVERY

At present, most exome and genome analyses are conducted on a “per family” basis, that is, to say on a small number of related individuals, most commonly a trio (parents and child). Analyzing NGS data is labor-intensive, taking up to 20–40 h to compile a report (Dewey et al., 2014). The challenge of handling vast quantities of genomic data has improved with advancing methods; however, for each exome or genome sequenced (and depending on whether segregation analysis is available using family studies), there are anywhere from tens to thousands of plausible candidate variants (Adams and Eng, 2018). If a variant is found in a known disease gene, a rapid diagnosis can often be made with rigorous variant curation against laboratory standard guidelines (Richards et al., 2015; Amendola et al., 2016). However, for ~60% of rare disease patients who undergo clinical exome or genome sequencing, their sequencing report is non-diagnostic, despite the fact that for many, the causal variant is present but unrecognized in their sequencing results. New diagnoses therefore cannot be made until new disease–gene relationships are discovered and the full spectra of pathogenic or disease-causing variants in known disease genes are elucidated (Seaby and Ennis, 2020).

It is worth noting that there is a difference between clinical testing and research-led sequencing studies. Clinical testing typically focuses on variants in established or known disease genes, whereas research studies have scope to evaluate variants in genes of unknown function or not currently linked to disease. Unless patients undergoing diagnostic testing are additionally recruited into research studies, opportunities to evaluate variants in new disease genes are limited. That said, there is increasing involvement of diagnostic centers with research laboratories and affiliated universities; indeed, many families are now concomitantly offered diagnostic testing and recruitment to further research studies. However, for these diagnostic labs capable of bridging the gap between clinical testing and research, many are ill-equipped to investigate the plethora of plausible disease candidates remaining after variant filtration and prioritization. Theoretically, the only way to establish new diagnoses when novel genes are identified in a research setting is to conduct functional experiments on all potential candidates. This approach is proving a major bottleneck as most research laboratories do not have the finances, nor the time or resources, to support functionally validating a large number of candidate variants without any guarantee that the selected variants are disease causing. Few laboratories will invest resources into a particular variant for one patient without additional kindreds with overlapping phenotypes or prior published studies on the gene’s function. Many current practices in gene discovery are still hindered by a “bottom-up” approach that takes single-patient

cases that brings the phenotype to a novel gene; a piecemeal approach limited by intensive functional experiments on genes of equally predicted causality. Therefore, efforts in recent years have focused on strategies to prioritize the best candidates for functional follow-up.

## STRATEGIES TO UPLIFT NOVEL GENE DISCOVERY

### Collaborative, Data-Sharing Approaches

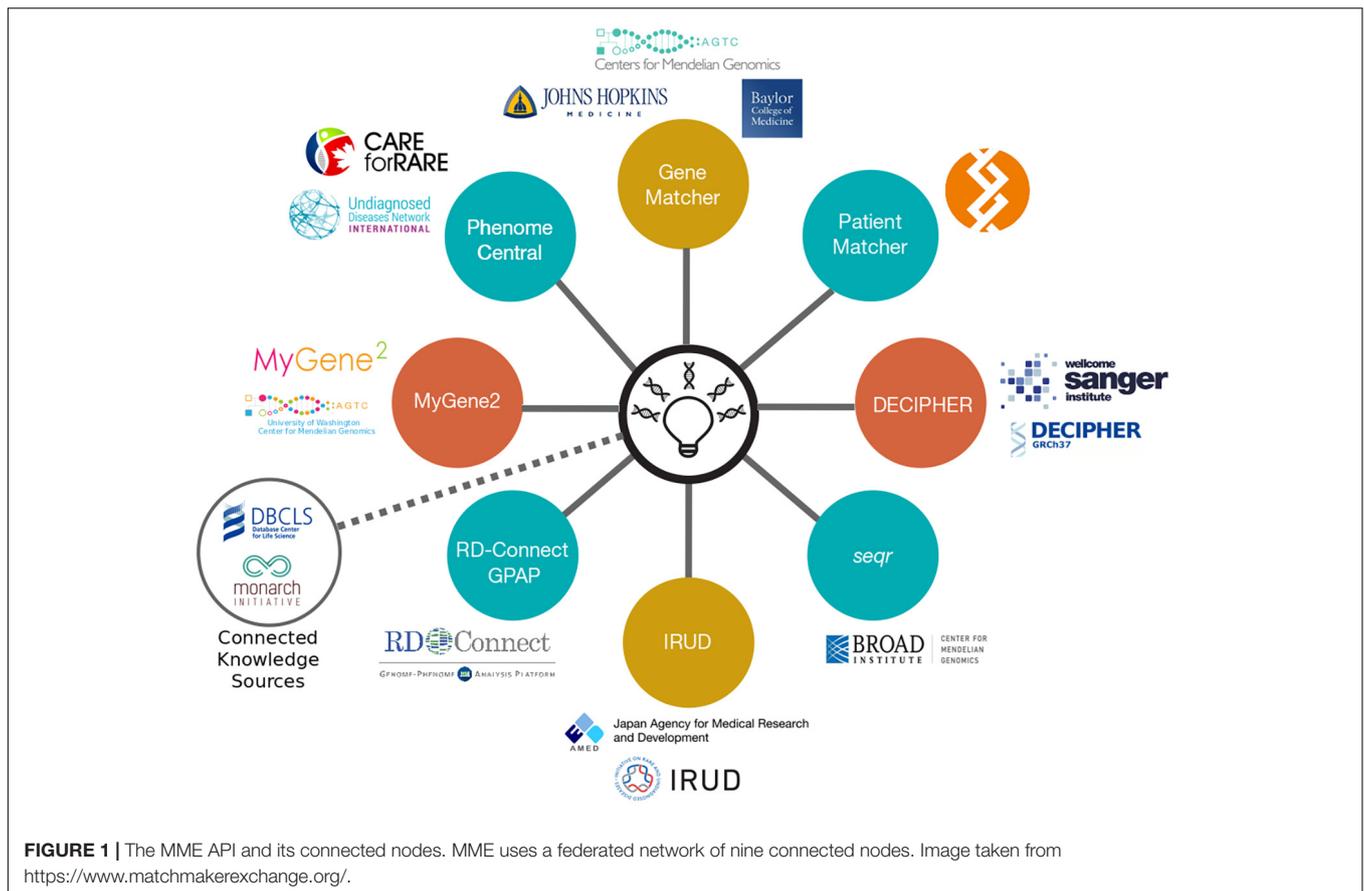
Collaborative projects, data sharing, and building disease cohorts have proven invaluable in genomics. In 2010, *MLL2* (*KMT2D*) was discovered as the cause of the Kabuki syndrome (MIM: 147920). Ten unrelated patients with the same characteristic clinical phenotype underwent exome sequencing. Seven of the 10 individuals were found to have loss-of-function variants in *MLL2*, which led to its disease association (Ng et al., 2010). Historically, the approach of building a case series of affected individuals has been a rate-limiting step, relying on local connections or collaborations built through conferences or publications. Given the rarity of monogenic disorders, it can take many years to accrue sufficiently sized cohorts with similar clinical features and genotypes. This method is therefore inefficient and inadequate to rapidly support novel gene discovery (Azzariti and Hamosh, 2020).

In 2017, the directors’ board of the American College of Medical Genetics and Genomics released a position statement on how genomic data sharing is critical to improving genetic healthcare (ACMG Board of Directors, 2017). With an ever more connected world, global efforts to share genotype and phenotype data have proven essential in the endeavor of novel gene discovery. Improved data governance, drives for open data science, and advancing informatics methods have since led to the practice of genomic matchmaking, facilitating researchers and clinicians from across the globe to share phenotype/genotype data for accelerated discoveries (Azzariti and Hamosh, 2020).

### Matchmaker Exchange

In 2015, the Matchmaker Exchange (MME) was launched, providing a systematic and robust approach to novel Mendelian gene discovery by facilitating a mechanism for matching patients across genomic centers, research laboratories, diagnostic laboratories, and physicians through a federated network (**Figure 1**) (Philippakis et al., 2015).

MME builds on the success of earlier genomic matchmaking platforms by connecting datasets through an application programming interface (API) enabling searches of multiple databases with a single query. The advantage of using a federated network enables individual submitters to maintain control and autonomy over their data and keep the content up to date, while ensuring compliance with their local and national data-sharing policies (Azzariti and Hamosh, 2020). By identifying additional affected kindreds with overlapping phenotypes, the best candidate variants and genes can be targeted for functional validation. The MME API has been widely adopted by scientists and clinicians globally and has led to numerous international



collaborations and publications. One such example is the discovery of a *KMT2E*-related neurodevelopmental disorder, the O'Donnell-Luria-Rodan syndrome (MIM: 618512), following identification of 38 individuals from 36 families, of which 28 were ascertained using MME. This discovery goes beyond elucidating disease–gene etiology and has identified a potential therapy already widely used in healthcare that could be evaluated for this syndrome (O'Donnell-Luria et al., 2019).

## Patient-Led Approaches

It could be argued that no one is more invested in ending the diagnostic odyssey of rare disease than affected individuals and their families. In an era when patients are actively involved in research studies as participants (Kaye et al., 2012), it is unsurprising that patients and caregivers are also invested in genomic matchmaking efforts (Lambertson et al., 2015). Patients and families are beginning to take control of their own data and utilize open data sharing and social media in an effort to discover new genetic disorders. Embedded within the MME API is a family-facing platform called MyGene2, which gives patients and caregivers autonomy over their data, facilitating direct data sharing when desired, while still enabling scientists and clinicians to access these shared anonymized data (Azzariti and Hamosh, 2020).

Social networking sites such as Facebook, Twitter, and Instagram are also proving popular with patients/caregivers as

a matchmaking resource (Macnamara et al., 2019). In 2014, Matthew and Cristina Might harnessed the power of social media to identify additional cases of *NGLY1* deficiency, leading to identification of a new gene disorder (Might and Wilsey, 2014; Might and Might, 2017). Following their son's diagnosis, the Might family explored options for conceiving a child unaffected by the same condition. Their son's diagnosis facilitated not only conception of a healthy sibling but a pathway for other affected families to conceive healthy children using preimplantation genetic testing or non-invasive prenatal diagnostic testing (Might and Might, 2017). The Might family created a legacy for others to follow, having built a global community of families providing mutual support, in addition to facilitating research and international *NGLY1* meetings (Might and Wilsey, 2014).

Inspired by the success of the Might family, families across the globe have harnessed the networking potential of social media to match with other affected kindreds with similar phenotypes and genotypes. Indeed, social media additionally facilitated the identification of three children with variants of uncertain significance in *KDM1A*, leading to discovery of another novel gene disorder (MIM: 616728) (Chong et al., 2016). The success of such endeavors has now inspired the Undiagnosed Diseases Network (UDN), started at the National Institute for Health in 2008, with 11 additional clinical sites across the United States, to use social media in a similar way. With appropriate consent, webpages are created for individual participants, showcasing the

clinical phenotype, significant variants, and candidate genes. This approach has proven successful in identifying additional affected patients with variants in *NACCI*, leading to the discovery of its associated phenotype (Macnamara et al., 2019).

## Large-Scale Programs for Novel Gene Discovery

Genomic sequencing has become increasingly affordable and possible in recent years. However, exome and genome sequencing are seldom first-line investigations, with many healthcare systems and health insurance policies not covering the cost. This has perhaps inspired the creation of large-scale international sequencing programs, often with government funding, offering exome and/or genome sequencing to thousands of rare disease patients and their families. These projects benefit from pooled resources and focus on diagnosing patients who were undiagnosed following conventional clinical testing, in addition to better elucidating the underlying mechanisms of Mendelian diseases. Such examples include the United Kingdom's 100,000 Genomes Project (100KGP) (Turnbull et al., 2018) and Deciphering Developmental Disorders study (Firth et al., 2011) as well as the United States' Centers for Mendelian Genomics (Bamshad et al., 2012). These programs benefit from sequencing large numbers of patients with improved power to match patients with similar genotypes and phenotypes, both internally and through the MME (Philippakis et al., 2015). Furthermore, most of these programs recruit patients for both clinical diagnostics and follow-on research, meaning that where possible, novel discoveries and variants of uncertain significance can be investigated further; this has previously been a limitation of clinical diagnostic studies (Seaby and Ennis, 2020). Consequently, thousands of novel gene discoveries have been identified through these projects. Indeed, by 2018, the Centers of Mendelian Genomics had reported >3,500 disease gene–phenotype pairs, expanding both known and novel disease gene associations (Posey et al., 2019). And in 2020, 28 new genetic disorders were discovered by leveraging data from the Deciphering Developmental Disorders study (Kaplanis et al., 2020).

## Phenotype-Driven Approaches

In recent years, novel gene discovery has shifted from phenotype-driven methods to genotype-driven approaches, i.e., taking genotype data and matching phenotypes to that genotype through matchmaking efforts, though both remain important (Bamshad et al., 2019). Efforts to standardize phenotype terms through the Human Phenotype Ontology (HPO) database have aided comparative statistics using a universal library of agreed clinical terms involved in disease (Robinson et al., 2008). This has paved the way for computational phenotype analyses that can assess a candidate gene's relevance to phenotype data observed in patient(s). A number of tools (Table 1) have been developed that estimate the similarity between HPO terms in an individual and those representing disease in a database. By incorporating phenotype ontology data across species, these tools are capable of prioritizing candidate genes without known disease association (Köhler et al., 2009; Schulz et al., 2011; Bauer et al., 2012). Similar

approaches have been commercialized, taking advantage of advanced artificial intelligence to identify and rank potential disease-causing variants following a multidimensional analysis; examples include Fabric Genomics<sup>1</sup> and Emedgene<sup>2</sup>. More experience and data are needed to understand the strengths and limitations of these tools.

One of the challenges for novel gene discovery is the requirement for accurate and deep phenotyping. Optimally, this should be collected longitudinally (Seaby and Ennis, 2020). While HPO terms do help to standardize the recording of phenotype information and indeed are used universally in many databases including those connected through MME (Philippakis et al., 2015), they are often only collected at a point in time and may lack the “full narrative” of the clinical history. This can be problematic when assessing new genotype–phenotype correlations, since for many neurodevelopmental disorders, phenotypes can significantly overlap. It can also be difficult to weight the severity or prominence of clinical features as the conversion to a list of terms tends to weight all the features similarly.

## Constraint-Based Approaches

Given the mutation rate and the Earth's current population size, we expect to observe every variant compatible with life in a living human. Indeed, the aggregation of large population datasets has begun to reveal the spectrum of damaging variants across the human genome (Lek et al., 2016; Karczewski et al., 2020). It is typical to observe approximately 100 loss-of-function variants per genome with ~20 genes completely inactivated (knockouts) even in perfectly healthy individuals from the general population (MacArthur et al., 2012). Population data can be utilized to evaluate the strength of natural selection at the gene level and to differentiate rare from common loss-of-function variants. As deleterious variants are purged from human populations through natural selection, there are opportunities to identify genes and regions that are constrained for variation compared to expected mutation rates, revealing which genes are most intolerant to inactivation of one (haploinsufficient) or both (knockout) copies (Samocha et al., 2014; Karczewski et al., 2020).

## Loss-of-Function Constraint

Lek et al. (2016) (Kosmicki et al., 2017) defined a set of genes with high probability of intolerance to heterozygous predicted loss-of-function variation (pLI) modeled on ~60,000 exomes from the Exome Aggregation Consortium (ExAC) population database (Lek et al., 2016). This pLI score can be used to identify candidate haploinsufficient disease genes constrained for loss of function in a dichotomous way; i.e., a gene is predicted to be haploinsufficient (pLI > 0.9) or not (Figure 2).

Karczewski et al. (2020) refined the model and regenerated pLI scores utilizing a larger dataset of ~141,000 exomes and genomes from the Genome Aggregation Database (gnomAD)<sup>3</sup>. The

<sup>1</sup><https://fabricgenomics.com/>

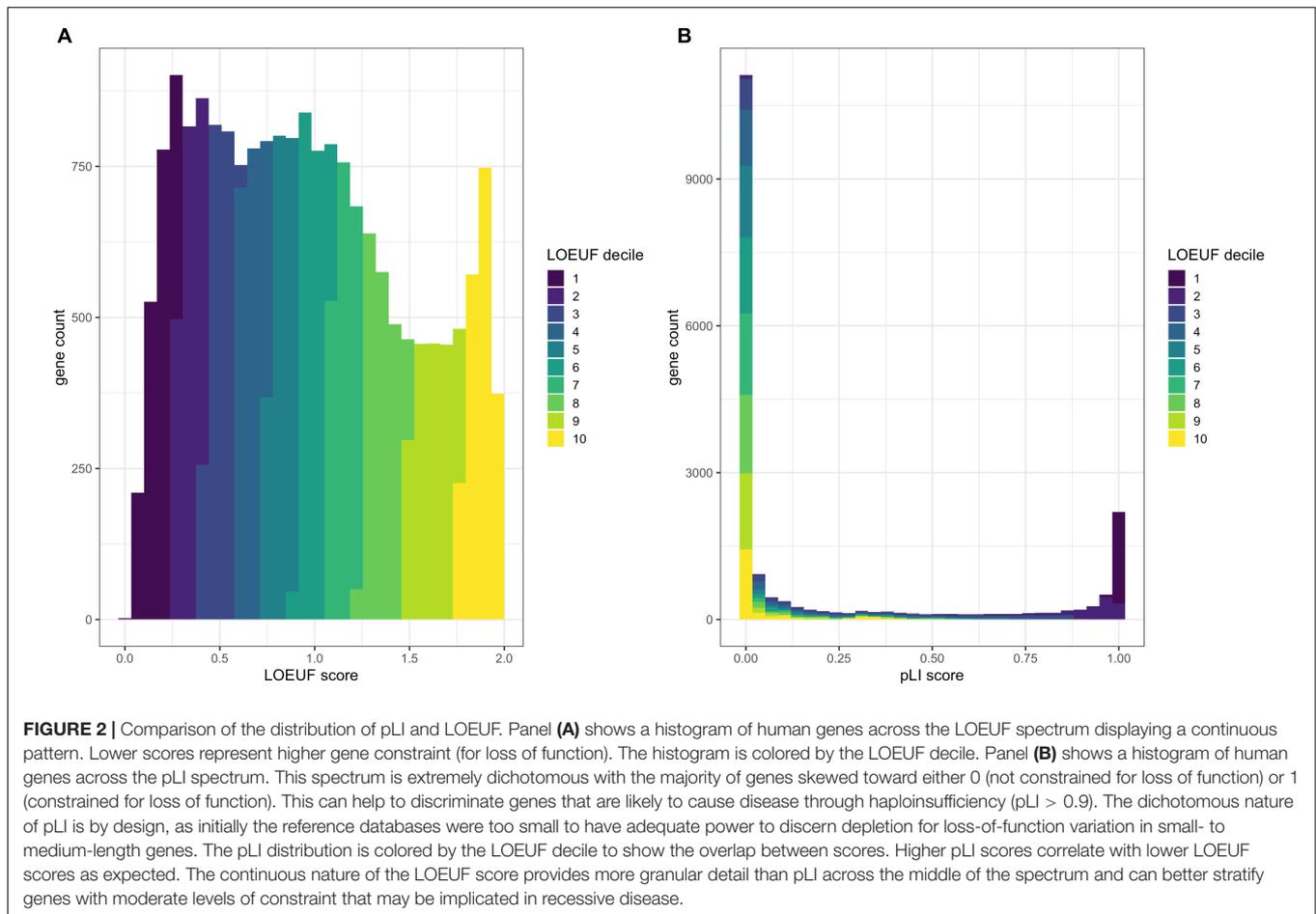
<sup>2</sup><https://www.emedgene.com/>

<sup>3</sup><https://gnomad.broadinstitute.org>

**TABLE 1** | Four phenotype-driven tools for prioritization of known and novel disease genes.

Tool	Principle	Application	Access
Exomiser (Smedley et al., 2015)	Uses random-walk analysis of protein–protein interaction networks, cross-species phenotype comparisons, and a wide range of additional filters that consider prediction models, disease segregation, and allele frequency	Focused on identifying novel and known disease genes	<a href="http://www.sanger.ac.uk/science/tools/exomiser">http://www.sanger.ac.uk/science/tools/exomiser</a>
eXtasy (Sifrim et al., 2013)	Prioritizes non-synonymous variants predicted to be pathogenic using a fusion methodology that integrates multiple strategies in a phenotype-specific manner	Focused on identifying candidates in novel and known disease genes	<a href="http://extasy.esat.kuleuven.be/">http://extasy.esat.kuleuven.be/</a>
Phevor (Singleton et al., 2014)	Combines outputs of multiple biomedical ontologies and propagates patient phenotype information across and between ontologies for improved variant interpretation	Focused on identifying candidates in novel and known disease genes	<a href="http://weatherby.genetics.utah.edu/cgi-bin/Phevor/PhevorWeb.html">http://weatherby.genetics.utah.edu/cgi-bin/Phevor/PhevorWeb.html</a>
Phen-Gen (Javed et al., 2014)	Uses a systematic Bayesian framework which combines patient sequencing data with phenotype information for improved rare disease variant analysis of both coding and non-coding variation	Focused on identifying candidates in novel and known disease genes	<a href="http://phen-gen.org/">http://phen-gen.org/</a>

*Bioinformatic phenotype-driven tools for the detection and prioritization of variants in novel and known disease genes using the integration of patient genotype and phenotype data.*



authors also developed the Lower Observed/Expected Upper-bound Fraction (LOEUF) score, a continuous metric which places >19,000 human genes on a spectrum of intolerance to

inactivation (**Figure 2**). Genes with the lowest LOEUF scores, i.e., the fewest loss-of-function variants compared to an expectation, are the most constrained for loss of function, highlighting

their potential biological essentiality. Both LOEUF and pLI were validated by comparison to several orthogonal indicators of constraint and shown to be accurate at discriminating haploinsufficient disease genes from autosomal recessive and polymorphic (unconstrained) genes (Karczewski et al., 2020). A companion paper by Collins et al. (2020) additionally showed that structural variants share the same pattern of constraint as LOEUF and are responsible for about a quarter of all rare loss-of-function events per genome.

As LOEUF identifies genes constrained for loss-of-function variation, we expect these genes to be enriched for dominant disease genes and to lesser extent recessive disease genes. As of January 2021, 65% of genes in the lowest LOEUF decile are yet to have an OMIM disease association (calculated using data from <https://omim.org>), highlighting thousands of high-probability candidate disease genes awaiting discovery of the associated phenotypes.

## Missense Constraint

The majority of coding variants of uncertain clinical significance are missense variants, as reported in ClinVar, a public database where diagnostic laboratories and researchers share variant classifications (i.e., pathogenic, benign, and uncertain significance) (Pérez-Palma et al., 2019). Similar to methods for assessing loss-of-function constraint, methods to identify missense constraint have emerged by comparing the observed vs. expected numbers of missense variants modeled on population data (Lek et al., 2016; Samocha et al., 2017; Havrilla et al., 2019; Hayeck et al., 2019; Pérez-Palma et al., 2020). However, missense constraint varies across a gene; for example, unstructured regions are often less constrained than important functional domains, which has necessitated the development of regional missense constraint models (Havrilla et al., 2019; Abramovs et al., 2020). Furthermore, clustering patterns of pathogenic missense variants vary depending on the inheritance pattern. Turner et al. (2015) showed that dominant missense variants cluster more than recessive variants. Therefore, testing for non-random clustering patterns may identify novel regions of interest across large sample sizes (Turner et al., 2015). The application of these metrics has aided in the discovery of new gene disorders, including a *de novo* missense variant in a constrained region of *GABRA2* responsible for an early-onset epileptic encephalopathy (MIM: 618557) (Orenstein et al., 2018).

## Model Organism Databases

Identifying the phenotypic effects of gene disruption may be possible using model organisms when there is enough conserved evolutionary function of the pathway/organ/system involving the gene of interest. Where the functional consequences of most human gene variants are yet to be established, model organism databases serve as a useful resource. Indeed, 58% of human genes have orthologs with disease-associated phenotypes reported in at least one model organism (Mungall et al., 2017). Although non-human models are not necessarily perfect proxies for human diseases, they can still serve as important biological models, particularly when data are aggregated across species.

## Monarch Initiative

The Monarch Initiative<sup>4</sup> is an open-science, collaborative project that aims to integrate phenotype–genotype data from a variety of species and sources (Mungall et al., 2017). Its user-friendly web portal promotes rapid assessment of phenotypes of orthologs in organisms and other species. Researchers can query genes, phenotypes, and diseases to identify candidate disease genes. Exomiser (Smedley et al., 2015) and Genomiser (Smedley et al., 2016) have utilized the Monarch Initiative in their gene prediction algorithms, which have led to diagnoses in participants in the UDN including the aforementioned discovery that the disruption of *STIM1* results in the York platelet syndrome (MIM: 805070) (Bone et al., 2016).

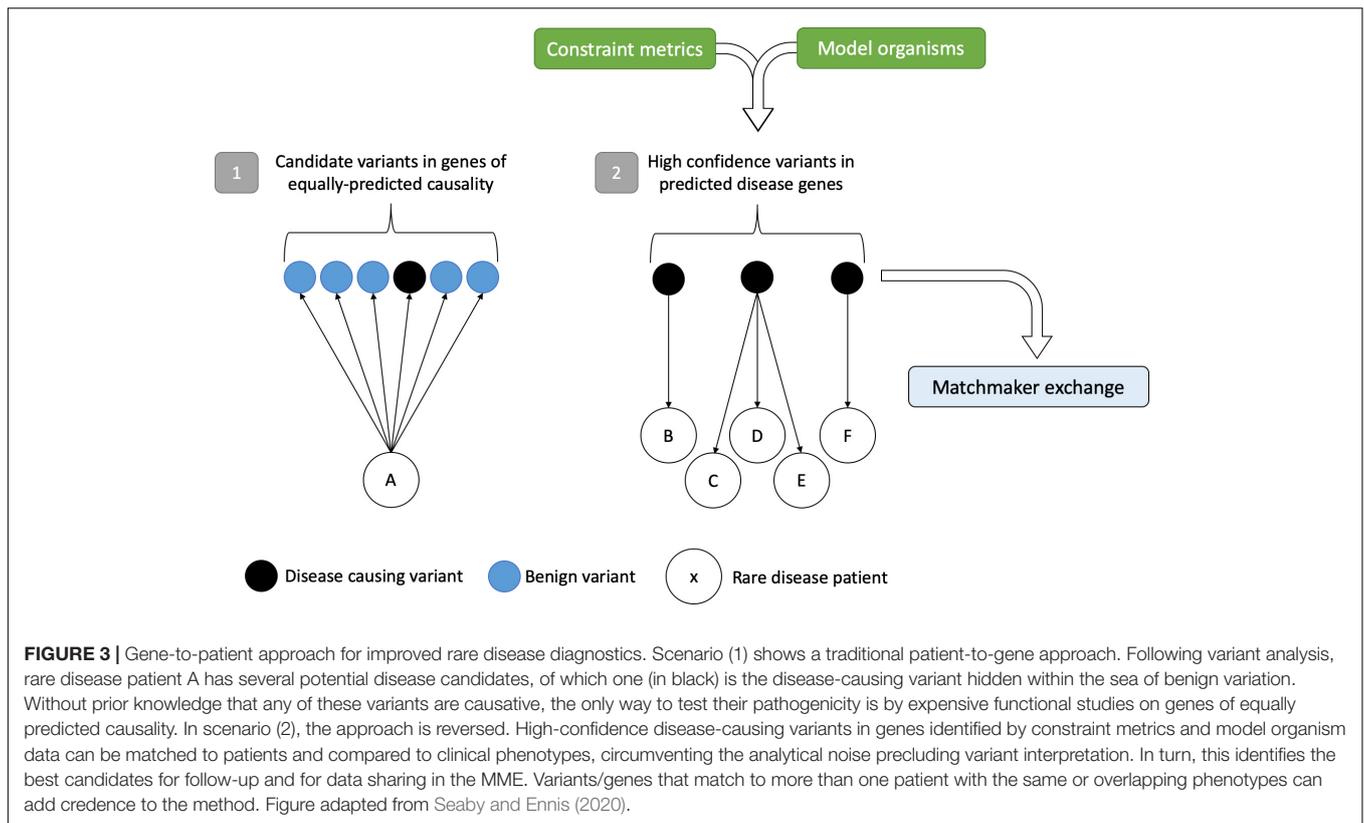
## Mouse Knockout Databases

Several mouse model organism databases exist including the Mouse Genome Database (MGD) (Smith et al., 2017; Bult et al., 2019), the Knockout Mouse Project (KOMP) (Austin et al., 2004), and the International Mouse Phenotyping Consortium (IMPC) (Meehan et al., 2017; Muñoz-Fuentes et al., 2018). These projects are building comprehensive catalogs of mammalian gene function, genotype–phenotype associations, and detailed phenotype data from mouse knockouts of every protein-coding gene (Muñoz-Fuentes et al., 2018; Bult et al., 2019). By 2019, the IMPC has fully or partially phenotyped 5,861 mouse genes, a third of which are non-viable (Muñoz-Fuentes et al., 2018; Cacheiro et al., 2019). Data from IMPC have aided the discovery of many novel Mendelian phenotypes (Bowl et al., 2017; Moore et al., 2018; Rozman et al., 2018). That said, there is still much more to be gleaned from mouse data; of the >10,000 mouse genes linked to at least one non-lethal phenotype in a mutant strain in MGD, Bamshad et al. (2019) showed that human orthologs for 72% of those genes are yet to be associated with a Mendelian disorder, providing another rich data source for candidate genes awaiting discovery of the human Mendelian phenotype.

## Incomplete Penetrance

Identifying novel disease genes can be challenged by incomplete penetrance, that is to say, when a disease-causing variant does not always result in any clinical expression of the disease. If a novel candidate gene has been associated with a given phenotype, yet some or all alleles are incompletely penetrant, then it can be difficult to gather sufficient evidence for a new disease–gene association using traditional genetic evidence such as case observations and familial segregation. To mitigate this, larger cohorts that can support statistical association studies must be pursued. Furthermore, researchers are exploring how combinations of genomic variants such as oligogenic models or co-inherited protective alleles, environmental exposures, and mosaicism may impact the onset of Mendelian disorders (Gruber and Bogunovic, 2020). One such approach is to specifically identify individuals that are resilient to rare disease, despite harboring pathogenic variants (Chen et al., 2016). Another area of interest is how *cis*-regulatory variation may modify the penetrance of coding variants (Castel et al., 2018).

<sup>4</sup><https://monarchinitiative.org>



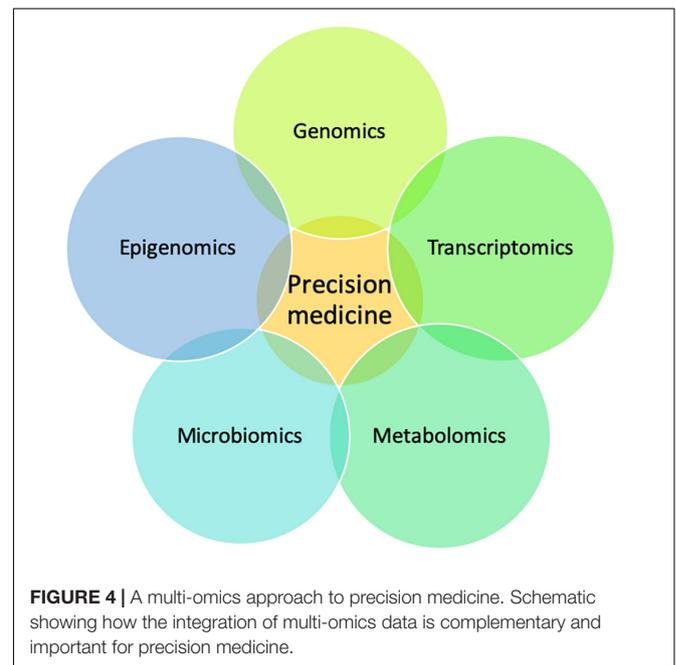
## Gene-to-Patient Approaches

In recent years, government funding has invested in national sequencing projects for rare disease (Turnbull et al., 2018; Posey et al., 2019). Despite 100,000 individuals with rare disease being sequenced in the United Kingdom as part of the Genomics England 100KGP, the diagnostic rates are similar to those reported elsewhere in the literature (Houses of Parliament, 2015; Turnbull et al., 2018). However, the scale of such datasets welcomes opportunities for new approaches to novel gene discovery.

With increasing data available on genetic variation from a variety of sources including gene constraint, mouse models, and phenotype-driven methods, there is scope to utilize the power of large cohort sizes for novel gene discovery. Instead of bringing a patient to a gene, there are opportunities, with large enough sample sizes, to be sufficiently powered to detect rare variation and bring candidate genes to large genomic datasets from patients (Figure 3). These “gene-to-patient” approaches are already being applied to accelerate novel gene discovery and prioritize genes for functional studies (Seaby and Ennis, 2020).

## Integrating Multi-Omics Data

Multiple omics technologies such as epigenomics, transcriptomics, metabolomics, microbiomics, and proteomics are being adopted as approaches in the effort to delineate the functional impact of genetic variation (Figure 4) (Hasin et al., 2017). These integrative approaches can complement genomic data and aid in the validation and discovery of novel genes.



The Genotype-Tissue Expression (GTEx) project<sup>5</sup> (Lonsdale et al., 2013) provides a public repository of tissue-specific gene expression and a multi-tissue reference for identifying

<sup>5</sup><https://gtexportal.org/home/>

variants associated with changes in gene expression or expression quantitative trait loci (eQTL) (Stranger et al., 2017). The GTEx consortium recently published results on their v8 release, providing insights into functional mechanisms and the architecture of genetic regulation (GTEx Consortium, 2020). The integration of transcriptome sequencing (RNA-seq) has led to improved diagnostic rates for Mendelian diseases (Kremer et al., 2017; Wai et al., 2020), even when no strong candidate variants were identified from exome or genome data (Cummings et al., 2017). Expression outliers, altered splicing, and allelic imbalance in the transcriptome due to non-sense-mediated decay can all be clues to candidate genes worth closer scrutiny in the exome or genome data (Brechtmann et al., 2018; Green et al., 2018; Mertens et al., 2021). Large-scale transcriptome data can also be used in network analysis (Deelen et al., 2019). One caveat with RNA-seq is that splicing aberrations and differential gene expression are best assessed by sampling disease-relevant tissues. However, these may not always be clinically accessible, e.g., brain tissue in neurodevelopmental disorders. Aicher et al. (2020) showed that many splicing events in non-clinically accessible tissues are lowly expressed and poorly evaluated from more commonly accessible tissues such as skin and blood. The authors developed a tool (MAJIQ-CAT) which allows researchers to explore potentially accessible tissues that best represent splicing in genes of interest (Aicher et al., 2020). A recent preprint in 2021 describes an alternative approach that has advantages over expression-based methods. Minimum Required Sequencing Depth (MRSD) informs biosample selection (whole blood, lymphoblastic cell lines, and skeletal muscle) by estimating the minimum sequencing depth required from RNA-sequencing to achieve desired coverage across a given gene or gene panel. The authors reported high precision, and their results suggest that lymphoblastic cell lines may be suitable for ~70% of established disease gene panels (Rowlands et al., 2021).

The Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects have been instrumental in the generation of human reference epigenomes and epigenome maps, mainly from cell lines (Bernstein et al., 2010; Dunham et al., 2012; Satterlee et al., 2019). These data have successfully been used to conduct research on how the epigenome contributes to human development, environmental factors, and disease mechanisms (Dunham et al., 2012; Satterlee et al., 2019). More specifically, one of most commonly studied epigenetic phenomena, DNA methylation, is aiding diagnosis and gene discovery. Alterations in DNA methylation patterns are implicated in imprinting disorders and diseases of short tandem repeat (STR) expansions. The application of DNA methylation analyses has been successful in identifying molecular diagnoses in neurodevelopmental disorders where clinical microarray and other conventional genetic testing have been non-diagnostic (Aref-Eshghi et al., 2019; Turinsky et al., 2020). LaCroix et al. (2019) investigated cases of Baratela-Scott syndrome (BSS) (MIM: 615777) and identified hypermethylation of exon 1 of *XYLT1* associated with a GGC expansion and gene silencing. This not only confirmed BSS as a trinucleotide repeat expansion disorder but also highlighted the relative prevalence of methylation abnormalities in the disease pathogenesis of BSS.

The hypermethylated allele accounted for 50% of the pathogenic alleles in their cohort, showcasing the importance of investigating epigenetic changes in disease cohorts with missing heritability (LaCroix et al., 2019).

National biobanks such as the United Kingdom Biobank (Bycroft et al., 2018), United States All of Us Research Program (Sankar and Parker, 2017), and Finland Biobank (FinnGen) provide opportunities to study genomic data and phenotype data alongside associated molecular markers from electronic medical records. While their data are best studied in the context of complex disease, they are also important in rare disease by providing population-level allele frequencies, biomarker results, and phenotypic information for comparative analyses. Unlu et al. (2020) utilized Vanderbilt's biobank BioVU to identify a phenotypic profile that aided in the identification of a novel Mendelian syndrome CATIFA (cleft lip, cataract, tooth abnormality, intellectual disability, facial dysmorphism, attention-deficit hyperactivity disorder) that is due to loss of function of *RIC1* (MIM: 618761).

The emerging application of metabolomics with exome/genome sequencing is helping to improve diagnostic rates in rare disease. Targeted and untargeted metabolomics are proving successful in validating variants of uncertain significance in inborn errors of metabolism (Graham et al., 2018; Almontashiri et al., 2020). It is hoped that with increasing research, metabolomics will continue to complement rich genomic data and aid in discovery of novel genes.

These aforementioned approaches often applied in combination have been pivotal both in clinical diagnostics and in identification of novel candidate disease genes. For example, a study in 2015 using epigenomics, comparative genomics, and genome editing identified a pathway for adipocyte thermogenesis regulation involving *IRX3*, *IRX5*, and *ARID5B* in obesity (Claussnitzer et al., 2015), and in 2017, the complex I assembly factor *TIMMDC1* was established as a novel mitochondrial disease-gene by utilizing genomic and transcriptomic sequencing (Kremer et al., 2017).

## DISCUSSION

The NGS era has undoubtedly accelerated novel Mendelian gene discovery for significant patient benefit. Although in recent years there has been increasing interest in the non-coding space, there is still much to be gleaned from the human exome (Bamshad et al., 2019; Posey et al., 2019). Furthermore, disease-gene associations are complex; phenotypic and functional consequences of variation across a gene are highly variable and influenced by variant type, inheritance pattern, and gene position. Following identification of the disease-gene relationship, characterization of the full allelic series is needed.

With many strategies now available for novel gene discovery, best practices are likely to benefit from the aggregation of methods. Additionally, as drives for data uniformity are developed and adopted (e.g., use of HPO terms, standardized disease nomenclature, functional equivalence variant calling pipelines, and joint calling data) (Regier et al., 2018), it will

become far easier to automate bioinformatics pipelines that are capable of processing and integrating data from a variety of rich datasets to increase power for diagnosis and gene discovery.

## Role of Data Sharing

It is unquestionable that open science and data sharing have been pivotal in uplifting diagnoses and advancing the field of genomic medicine. International collaborations are now commonplace in matching patients across the globe with specific genotypes, leading to high-impact publications on novel gene discoveries (Philippakis et al., 2015; Azzariti and Hamosh, 2020). Furthermore, the use of variant databases such as ClinVar<sup>6</sup> have proven invaluable in providing the scientific community with a repository of variants, classified by pathogenicity, that can be applied to variant analysis for diagnostic interpretation (Landrum et al., 2014).

The success of open data science has been further driven by cloud computing. The presence of large datasets on cloud platforms can facilitate the access of desired data within a secure data-sharing platform. Examples include NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab (AnVIL) Space<sup>7</sup> where rare disease data from the Centers for Mendelian Genomics along with data from additional projects such as 1,000 Genomes, Centers for Common Disease Genomics, and GTEx can be accessed after application in dbGaP. Other trusted research environments include NHLBI's BioData Catalyst cloud platform with TOPMed data; Genomics England Research environment where 100 KGP data are stored and accessed; and RD-Connect with rare disease genomic data from various European sources. In these trusted research environments, increasingly large amounts of data can be aggregated; researchers can bring tools directly to the data and share these analysis workflows, saving time, expense, and security risks of moving and maintaining local copies of large genomic datasets. However, data sharing presents its challenges. There is still urgent need for an international code of conduct that provides clear, unified data-sharing rules across jurisdictions that comply with regional laws such as the European General Data Protection Regulation (GDPR) and the United States' Health Insurance Portability and Accountability Act (Phillips et al., 2020).

## Addressing the Translational Gap

Our understanding of the genetic basis of rare disease is constantly changing with new genes and variation being linked to disease at a rapid pace. Given the direct application of these discoveries to the clinical diagnosis of rare disease in patients, guidance is needed for understanding what information is ready to be incorporated into clinical care and what mechanisms are needed to quickly translate that information into medical practice. The Clinical Genome Resource (ClinGen) has developed a systematic framework for evaluating genetic and functional evidence for disease–gene relationships, enabling their classification as definitive, strong, moderate limited, no human evidence, disputed, or refuted with respect to their reported

role in disease (Strande et al., 2017). ClinGen supports Gene Curation Expert Panels that bring together international groups of disease and curation experts to evaluate gene–disease claims in their respective fields<sup>8</sup>. ClinGen's efforts are combined with other public and private gene curation efforts and are accessible within the Gene Curation Coalition database<sup>9</sup>. Currently, it is recommended that a gene–disease relationship reach moderate classification before it is included on predefined diagnostic gene panels for specific conditions (Bean et al., 2020). However, when performing exome and genome approaches on individuals with rare disease, variation can be detected in genes that have not yet been linked to disease but may be strong candidates. Although practices vary between laboratories and countries, some professional standards recommend reporting these findings back to patients when there is a reasonable chance that new evidence may evolve over time to strengthen the gene–disease relationship, similar to the return of variants of uncertain significance in genes already linked to the patient's condition (Richards et al., 2015; Rehder et al., 2021). This approach also allows patients to be partners in solving the causes of rare disease (Mnookin, 2014). It is hoped that such a framework will achieve global recognition and be universally adopted to ensure consistency in translating research findings into the clinic.

## Clinical Impact of Novel Mendelian Conditions

While novel gene discoveries widen the known functional repertoire of disease genes, the focus and drive are ultimately uplifting diagnosis rates and improving patient outcomes. There have been thousands of pivotal Mendelian discoveries throughout history, and each one is no more important than another, at least not for the families involved.

Since the discovery of the *CFTR* gene in 1989 (Kerem et al., 1989), we are now able to diagnose cystic fibrosis (MIM: 602421) rapidly, predict pancreatic functional status, and plan preventative care with modulator therapy (Ramsey et al., 2011; Farrell et al., 2020). In 2004, the discovery that hypermorphic or gain-of-function variants of *PCSK9* cause familial hypercholesterolemia type III (MIM: 603776) (Timms et al., 2004) has led to the successful development and FDA approval of monoclonal antibodies against PCSK9, which are also used to treat non-familial forms of hypercholesterolemia (Blom et al., 2014; Roth et al., 2014; Cannon et al., 2015; Kereiakes et al., 2015). As more collaborative, cohort-based studies have emerged in the NGS era, many candidate genes have been discovered that have directly impacted treatment and clinical outcomes. In one study on neurometabolic disorders, whole-exome sequencing diagnosed 68% of patients and identified 11 novel candidate genes, leading to a targeted intervention in 44% of patients (Tarailo-Graovac et al., 2016).

Diagnosing Mendelian disorders as a direct result of novel gene discovery not only impacts the primary patient involved but their families and caregivers. Families of children with rare genetic diseases are adversely impacted by lack of peer

<sup>6</sup><https://www.ncbi.nlm.nih.gov/clinvar/>

<sup>7</sup><https://anvilproject.org/data/>

<sup>8</sup><https://clinicalgenome.org/affiliation>

<sup>9</sup><https://thegenc.org>

support groups and psychological support as well as delays in diagnosis (Anderson et al., 2013). Parents of children with rare disorders have called for better education, reduction in avoidable diagnostic delays, and early access to interventions and treatments (Zurynski et al., 2017). For many, a genetic diagnosis can be life changing, even in the absence of a therapeutic option (Lingen et al., 2016). Following diagnosis, quality of life is often improved by participation in support groups that can provide longitudinal prognostic information, genetic counseling, and informed reproductive decisions with opportunities for pre-implantation genetic diagnosis or prenatal testing particularly in the case of inherited variants where there is a sizable recurrence risk (Sexton et al., 2008; Wojcik et al., 2020).

## Financial Impact of Novel Discoveries

Undiagnosed rare diseases are hugely expensive. A typical patient's diagnostic odyssey lasts an average of 8 years and costs a total of \$5,000,000 throughout a patient's lifetime (Chong et al., 2015). Two prospective Australian studies have shown that early exome sequencing is making significant headway as a cost-saving diagnostic approach. Stark et al. (2017) showed that integrating whole-exome sequencing as a first-line test had an incremental cost saving per additional diagnosis of (converted to United States dollars) \$1,543 (95% CI: \$92–4,143). The cost per diagnosis was \$4,248 (95% CI: \$3,425–5,588), \$14,893 less than standard diagnostic care (Stark et al., 2017). Tan et al. (2017) concluded that whole-exome sequencing performed at initial presentation to tertiary care resulted in an incremental cost saving of (converted to United States dollars) \$6,383 per additional diagnosis (95% CI: \$3,045–10,900) compared with standard diagnostic care. However, cost savings are only possible when sequencing can identify the causal variant. Therefore, every new genetic disorder identified, published, and shared in publicly available databases will have wide-reaching diagnostic and cost-saving potential. Taking the average of costs saved per additional diagnosis from the two studies (\$3,964) and extrapolating this on 100,000 patients could save an estimated US\$400 million.

## LOOKING TO THE FUTURE

It is estimated that by 2025, 60 million patients will have their genome sequenced in a research or healthcare setting (Birney et al., 2017). While the sheer volume of data poses computational challenges, it also provides opportunities to learn more about the genetic architecture of health and disease. However, this necessitates improved methods for interpreting the spectrum of functional variation across all genes and particularly in the interpretation of non-coding variation, an area of investigation still in its infancy but beginning to make headway. Indeed,

disruption of non-coding topologically associated domains have been associated with limb malformations (Lupiáñez et al., 2015; Spielmann et al., 2018), and non-coding variants upstream of *PRDM13* and *CCNC* have been linked to North Carolina macular dystrophy (Small et al., 2016; Green et al., 2021). While efforts like the Atlas of Variant Effect Alliance are working toward achieving the mammoth goal of interpreting the impact of all genomic variation, there is still a long way to go (Matreyek et al., 2018; Jepsen et al., 2020). It is expected that as data pour in across a variety of species and sources, more and more methods will adopt machine learning and deep learning techniques to find patterns and disease associations, but the utility of these approaches is limited by the quality of the training data and other factors influencing data interpretation. For novel gene discovery, perhaps one of the most powerful resources would be to build a publicly available human knockout database that links naturally occurring null variants in genes and supportive functional evidence to shared human phenotype data. This is an exciting time for novel gene discovery—the end is by no means in sight.

## AUTHOR CONTRIBUTIONS

ES wrote the first draft of the manuscript, performed full literature review, and conceived idea for review. HR made changes to subsequent drafts of the manuscript and conceived idea for review. AO'D-L made changes to subsequent drafts of the manuscript, contributed to literature review, and conceived idea for review. All authors contributed to the article and approved the submitted version.

## FUNDING

All authors were supported by the National Human Genome Research Institute (NHGRI), the National Eye Institute, and the National Heart, Lung and Blood Institute grant UM1 HG008900. ES was supported by the Kerkut Charitable Trust and University of Southampton's Presidential Scholarship Award; HR by the NHGRI U24 HG011450 and U41 HG006834; and AO'D-L by the National Institute of Mental Health U01 MH119689 and Manton Center for Orphan Disease Research Scholar Award.

## ACKNOWLEDGMENTS

The authors would like to thank the members of the Centers for Mendelian Genomics and the Genome Aggregation Database teams for helpful discussions in the development of approaches to novel gene discovery.

## REFERENCES

- Abramovs, N., Brass, A., and Tassabehji, M. (2020). GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Nat. Genet.* 52, 35–39. doi: 10.1038/s41588-019-0560-2
- ACMG Board of Directors (2017). Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. *Genet. Med.* 19, 721–722. doi: 10.1038/gim.2016.196

- Adams, D. R., and Eng, C. M. (2018). Next-generation sequencing to diagnose suspected genetic disorders. *N. Engl. J. Med.* 379, 1353–1362. doi: 10.1056/nejmra1711801
- Aicher, J. K., Jewell, P., Vaquero-Garcia, J., Barash, Y., and Bhoj, E. J. (2020). Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet. Med.* 22, 1181–1190. doi: 10.1038/s41436-020-0780-y
- Almontashiri, N. A. M., Zha, L., Young, K., Law, T., Kellogg, M. D., Bodamer, O. A., et al. (2020). Clinical validation of targeted and untargeted metabolomics testing for genetic disorders: a 3 year comparative study. *Sci. Rep.* 10:9382. doi: 10.1038/s41598-020-66401-2
- Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., et al. (2016). Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. *Am. J. Hum. Genet.* 98, 1067–1076. doi: 10.1016/j.ajhg.2016.03.024
- Anderson, M., Elliott, E. J., and Zuryski, Y. A. (2013). Australian families living with rare disease: experiences of diagnosis, health services use and needs for psychosocial support. *Orphanet J. Rare Dis.* 8:22. doi: 10.1186/1750-1172-8-22
- Aref-Eshghi, E., Bend, E. G., Colaiacovo, S., Caudle, M., Chakrabarti, R., Napier, M., et al. (2019). Diagnostic utility of genome-wide DNA methylation testing in genetically unsolved individuals with suspected hereditary conditions. *Am. J. Hum. Genet.* 104, 685–700. doi: 10.1016/j.ajhg.2019.03.008
- Austin, C. P., Battey, J. F., Bradley, A., Bucan, M., Capecchi, M., Collins, F. S., et al. (2004). The knockout mouse project. *Nat. Genet.* 36, 921–924. doi: 10.1038/ng0904-921
- 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
- Bamshad, M. J., Nickerson, D. A., and Chong, J. X. (2019). Mendelian gene discovery: fast and furious with no end in sight. *Am. J. Hum. Genet.* 105, 448–455. doi: 10.1016/j.ajhg.2019.07.011
- Bamshad, M. J., Shendure, J. A., Valle, D., Hamosh, A., Lupski, J. R., Gibbs, R. A., et al. (2012). The Centers for Mendelian genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. *Am. J. Med. Genet. A* 158A, 1523–1525. doi: 10.1002/ajmg.a.35470
- Bauer, S., Köhler, S., Schulz, M. H., and Robinson, P. N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics* 28, 2502–2508. doi: 10.1093/bioinformatics/bts471
- Bean, L. J. H., Funke, B., Carlston, C. M., Gannon, J. L., Kantarci, S., Krock, B. L., et al. (2020). Diagnostic gene sequencing panels: from design to report—a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* 22, 453–461. doi: 10.1038/s41436-019-0666-z
- Beck, D. B., Petracovici, A., He, C., Moore, H. W., Louie, R. J., Ansar, M., et al. (2020). Delineation of a human Mendelian disorder of the DNA demethylation machinery: TET3 deficiency. *Am. J. Hum. Genet.* 106, 234–245. doi: 10.1016/j.ajhg.2019.12.007
- Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048. doi: 10.1038/nbt1010-1045
- Birney, E., Vamathevan, J., and Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. *bioRxiv* [Preprint] 203554. doi: 10.1101/203554
- Blom, D. J., Hala, T., Bolognese, M., Lillestol, M. J., Toth, P. D., Burgess, L., et al. (2014). A 52-week placebo-controlled trial of evolocumab in hyperlipidemia. *N. Engl. J. Med.* 370, 1809–1819. doi: 10.1056/nejmoa1316222
- Bone, W. P., Washington, N. L., Buske, O. J., Adams, D. R., Davis, J., Draper, D., et al. (2016). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.* 18, 608–617. doi: 10.1038/gim.2015.137
- Bowl, M. R., Simon, M. M., Ingham, N. J., Greenaway, S., Santos, L., Cater, H., et al. (2017). A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. *Nat. Commun.* 8:886. doi: 10.1038/s41467-017-00595-4
- Brechtmann, F., Mertes, C., Matuszevičiūtė, A., Yépez, V. A., Avsec, Ž, Herzog, M., et al. (2018). OUTFRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *Am. J. Hum. Genet.* 103, 907–917. doi: 10.1016/j.ajhg.2018.10.025
- Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., and Richardson, J. E. (2019). Mouse genome database (MGD) 2019. *Nucleic Acids Res.* 47, D801–D806.
- Burda, P., Kuster, A., Hjalmarson, O., Suormala, T., Burer, C., Lutz, S., et al. (2015). Characterization and review of MTHFD1 deficiency: four new patients, cellular delineation and response to folic and folinic acid treatment. *J. Inher. Metab. Dis.* 38, 863–872. doi: 10.1007/s10545-015-9810-3
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z
- Cacheiro, P., Haendel, M. A., Smedley, D., Meehan, T., Mason, J., Mashhadi, H. H., et al. (2019). New models for human disease from the International Mouse Phenotyping Consortium. *Mamm. Genome* 30, 143–150. doi: 10.1007/s00335-019-09804-5
- Cannon, C. P., Cariou, B., Blom, D., McKenney, J. M., Lorenzato, C., Pordy, R., et al. (2015). Efficacy and safety of alirocumab in high cardiovascular risk patients with inadequately controlled hypercholesterolaemia on maximally tolerated doses of statins: the ODYSSEY COMBO II randomized controlled trial. *Eur. Heart J.* 36, 1186–1194. doi: 10.1093/eurheartj/ehv028
- Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., et al. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* 50, 1327–1334. doi: 10.1038/s41588-018-0192-y
- Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* 34, 531–538. doi: 10.1038/nbt.3514
- Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., et al. (2015). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* 97, 199–215.
- Chong, J. X., Yu, J.-H., Lorentzen, P., Park, K. M., Jamal, S. M., Tabor, H. K., et al. (2016). Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. *Genet. Med.* 18, 788–795. doi: 10.1038/gim.2015.161
- Clark, M., Stark, Z., Farnaes, L., Tan, T., White, S., Dimmock, D., et al. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom. Med.* 3:16. doi: 10.1038/s41525-018-0053-8
- Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373, 895–907. doi: 10.1056/nejmoa1502214
- Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., et al. (2020). A structural variation reference for medical and population genetics. *Nature* 581, 444–451. doi: 10.1038/s41586-020-2287-8
- Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* 9:eal5209.
- Deelen, P., van Dam, S., Herkert, J. C., Karjalainen, J. M., Brugge, H., Abbott, K. M., et al. (2019). Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nat. Commun.* 10:2837.
- Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., et al. (2014). Clinical interpretation and implications of whole-genome sequencing. *JAMA* 311, 1035–1045.
- Dodge, J. A., Chigladze, T., Donadiou, J., Grossman, Z., Ramos, F., Serlicorni, A., et al. (2011). The importance of rare diseases: from the gene to society. *Arch. Dis. Child.* 96, 791–792. doi: 10.1136/adc.2010.193664
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247
- Farrell, P. M., Rock, M. J., and Baker, M. W. (2020). The impact of the CFTR gene discovery on cystic fibrosis diagnosis, counseling, and preventive therapy. *Genes* 11:401. doi: 10.3390/genes11040401
- Firth, H. V., Wright, C. F., and Study, D. (2011). The deciphering developmental disorders (DDD) study. *Dev. Med. Child Neurol.* 53, 702–703. doi: 10.1111/j.1469-8749.2011.04032.x
- Graham, E., Lee, J., Price, M., Tarailo-Graovac, M., Matthews, A., Engelke, U., et al. (2018). Integration of genomics and metabolomics for prioritization of rare

- disease variants: a 2018 literature review. *J. Inherit. Metab. Dis.* 41, 435–445. doi: 10.1007/s10545-018-0139-6
- Green, C. J., Gazzara, M. R., and Barash, Y. (2018). MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. *Bioinformatics* 34, 300–302. doi: 10.1093/bioinformatics/btx565
- Green, D. J., Lenassi, E., Manning, C. S., McGaughey, D., Sharma, V., Black, G. C., et al. (2021). North Carolina macular dystrophy: phenotypic variability and computational analysis of disease-implicated non-coding variants. *medRxiv* [Preprint] 2021.03.05.21252975. doi: 10.1101/2021.03.05.21252975
- Gruber, C., and Bogunovic, D. (2020). Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. *Hum. Genet.* 139, 745–757. doi: 10.1007/s00439-020-02131-9
- GTEX Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi: 10.1126/science.aaz1776
- Habib, A.-R. R., Kajbafzadeh, M., Desai, S., Yang, C. L., Skolnik, K., and Quon, B. S. (2019). A systematic review of the clinical efficacy and safety of CFTR modulators in cystic fibrosis. *Sci. Rep.* 9:7234.
- Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83. doi: 10.1186/s13059-017-1215-1
- Havrilla, J. M., Pedersen, B. S., Layer, R. M., and Quinlan, A. R. (2019). A map of constrained coding regions in the human genome. *Nat. Genet.* 51, 88–95. doi: 10.1038/s41588-018-0294-6
- Hayeck, T. J., Stong, N., Wolock, C. J., Copeland, B., Kamalakaran, S., Goldstein, D. B., et al. (2019). Improved pathogenic variant localization via a hierarchical model of sub-regional intolerance. *Am. J. Hum. Genet.* 104, 299–309. doi: 10.1016/j.ajhg.2018.12.020
- Houses of Parliament (2015). *The 100,000 Genomes Project*. POSTNOTE: Parliamentary Office of Science and Technology. London: Houses of Parliament.
- Imperial College Health Partners (2018). *A Preliminary Assessment of the Potential Impact of Rare Diseases on the NHS*. London: Imperial College Health Partners.
- Javed, A., Agrawal, S., and Ng, P. C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. Methods* 11, 935–937. doi: 10.1038/nmeth.3046
- Jepsen, M. M., Fowler, D. M., Hartmann-Petersen, R., Stein, A., and Lindorff-Larsen, K. (2020). “Classifying disease-associated variants using measures of protein activity and stability,” in *Protein Homeostasis Diseases*, ed. A. L. Pey (Amsterdam: Elsevier), 91–107. doi: 10.1016/b978-0-12-819132-3.00005-1
- Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 586, 757–762. doi: 10.1038/s41586-020-2832-5
- 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
- Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S. M., Kanellopoulou, N., et al. (2012). From patients to partners: participant-centric initiatives in biomedical research. *Nat. Rev. Genet.* 13, 371–376. doi: 10.1038/nrg3218
- Kereiakes, D. J., Robinson, J. G., Cannon, C. P., Lorenzato, C., Pordy, R., Chaudhari, U., et al. (2015). Efficacy and safety of the proprotein convertase subtilisin/kexin type 9 inhibitor alirocumab among high cardiovascular risk patients on maximally tolerated statin therapy: the ODYSSEY COMBO I study. *Am. Heart J.* 169, 906–915.e13.
- Kerem, B.-S., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., et al. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science* 245, 1073–1080. doi: 10.1126/science.2570460
- Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457–464. doi: 10.1016/j.ajhg.2009.09.003
- Kosmicki, J. A., Samocha, K. E., Howrigan, D. P., Sanders, S. J., Slowikowski, K., Lek, M., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* 49, 504–510. doi: 10.1038/ng.3789
- Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* 8:15824.
- LaCroix, A. J., Stabley, D., Sahraoui, R., Adam, M. P., Mehaffey, M., Kernan, K., et al. (2019). GGC repeat expansion and exon 1 methylation of XYLT1 is a common pathogenic variant in Baratela-Scott syndrome. *Am. J. Hum. Genet.* 104, 35–44. doi: 10.1016/j.ajhg.2018.11.005
- Lambertson, K. F., Damiani, S. A., Might, M., Shelton, R., and Terry, S. F. (2015). Participant-driven matchmaking in the genomic era. *Hum. Mutat.* 36, 965–973. doi: 10.1002/humu.22852
- Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985.
- Legendre, C. M., Licht, C., Muus, P., Greenbaum, L., Babu, S., Bedrosian, C., et al. (2013). Terminal complement inhibitor eculizumab in atypical hemolytic-uremic syndrome. *N. Engl. J. Med.* 368, 2169–2181.
- 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.
- Lingen, M., Albers, L., Borchers, M., Haass, S., Gärtner, J., Schroeder, S., et al. (2016). Obtaining a genetic diagnosis in a child with disability: impact on parental quality of life. *Clin. Genet.* 89, 258–266. doi: 10.1111/cge.12629
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585.
- Lupiañez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025. doi: 10.1016/j.cell.2015.04.004
- Ly, C. V., and Miller, T. M. (2018). Emerging antisense oligonucleotide and viral therapies for ALS. *Curr. Opin. Neurol.* 31, 648–654. doi: 10.1097/wco.0000000000000594
- MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823–828.
- Macnamara, E. F., D’Souza, P., Undiagnosed Diseases Network, and Tiffit, C. J. (2019). The undiagnosed diseases program: approach to diagnosis. *Transl. Sci. Rare Dis.* 4, 179–188. doi: 10.3233/TRD-190045
- Maguire, A. M., High, K. A., Auricchio, A., Wright, J. F., Pierce, E. A., Testa, F., et al. (2009). Age-dependent effects of RPE65 gene therapy for Leber’s congenital amaurosis: a phase 1 dose-escalation trial. *Lancet* 374, 1597–1605.
- Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50, 874–882. doi: 10.1038/s41588-018-0122-z
- Meehan, T. F., Conte, N., West, D. B., Jacobsen, J. O., Mason, J., Warren, J., et al. (2017). Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat. Genet.* 49, 1231–1238.
- Mendell, J. R., Al-Zaidy, S., Shell, R., Arnold, W. D., Rodino-Klapac, L. R., Prior, T. W., et al. (2017). Single-dose gene-replacement therapy for spinal muscular atrophy. *N. Engl. J. Med.* 377, 1713–1722.
- Mertes, C., Scheller, I. F., Yépez, V. A., Çelik, M. H., Liang, Y., Kremer, L. S., et al. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* 12:529. doi: 10.1038/s41467-020-20573-7
- Might, M., and Might, C. C. (2017). What happens when N= 1 and you want plus 1? *Prenat. Diagn.* 37, 70–72. doi: 10.1002/pd.4975
- Might, M., and Wilsey, M. (2014). The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated. *Genet. Med.* 16, 736–737. doi: 10.1038/gim.2014.23
- Mirmiran, A., Schmitt, C., Lefebvre, T., Manceau, H., Daher, R., Oustric, V., et al. (2019). Erythroid-progenitor-targeted gene therapy using bifunctional TFR1 ligand-peptides in human erythropoietic protoporphyria. *Am. J. Hum. Genet.* 104, 341–347. doi: 10.1016/j.ajhg.2018.12.021
- Mnookin, S. (2014). *One of a Kind*. New York, NY: The New Yorker.
- Moore, B. A., Leonard, B. C., Sebbag, L., Edwards, S. G., Cooper, A., Imai, D. M., et al. (2018). Identification of genes required for eye development by high-throughput screening of mouse knockouts. *Commun. Biol.* 1:236.
- Muir, E. (2016). *The Rare Reality-An Insight into the Patient and Family Experience of Rare Disease*. London: Rare Disease UK.
- Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., et al. (2017). The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 45, D712–D722.

- Muñoz-Fuentes, V., Cacheiro, P., Meehan, T. F., Aguilar-Pimentel, J. A., Brown, S. D., Flenniken, A. M., et al. (2018). The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conserv. Genet.* 19, 995–1005. doi: 10.1007/s10592-018-1072-9
- Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. L., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42, 790–793. doi: 10.1038/ng.646
- O'Donnell-Luria, A. H., Pais, L. S., Faundes, V., Wood, J. C., Sveden, A., Luria, V., et al. (2019). Heterozygous variants in KMT2E cause a spectrum of neurodevelopmental disorders and epilepsy. *Am. J. Hum. Genet.* 104, 1210–1222. doi: 10.1016/j.ajhg.2019.03.021
- Orenstein, N., Goldberg-Stern, H., Straussberg, R., Bazak, L., Hushman, M. W., Kropach, N., et al. (2018). A de novo GABRA2 missense mutation in severe early-onset epileptic encephalopathy with a choreiform movement disorder. *Eur. J. Paediatr. Neurol.* 22, 516–524. doi: 10.1016/j.ejpn.2017.12.017
- Pérez-Palma, E., Gramm, M., Nürnberg, P., May, P., and Lal, D. (2019). Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. *Nucleic Acids Res.* 47, W99–W105. doi: 10.1093/nar/gkz411
- Pérez-Palma, E., May, P., Iqbal, S., Niestroj, L. M., Du, J., Heyne, H. O., et al. (2020). Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res.* 30, 62–71. doi: 10.1101/gr.252601.119
- Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., et al. (2015). The matchmaker exchange: a platform for rare disease gene discovery. *Hum. Mutat.* 36, 915–921.
- Phillips, M., Molnár-Gábor, F., Korbel, J. O., Thorogood, A., Joly, Y., Chalmers, D., et al. (2020). Genomics: data sharing needs an international code of conduct. *Nature* 578, 31–33. doi: 10.1038/d41586-020-00082-9
- Pierce, E. A., and Bennett, J. (2015). The status of RPE65 gene therapy trials: safety and efficacy. *Cold Spring Harb. Perspect. Med.* 5:a017285. doi: 10.1101/cshperspect.a017285
- Posey, J. E., O'Donnell-Luria, A. H., Chong, J. X., Harel, T., Jhangiani, S. N., Akdemir, Z. H. C., et al. (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet. Med.* 21, 798–812.
- Ramakrishnan, K. A., Pengelly, R. J., Gao, Y., Morgan, M., Patel, S. V., Davies, E. G., et al. (2016). Precision molecular diagnosis defines specific therapy in combined immunodeficiency with megaloblastic anemia secondary to MTHFD1 deficiency. *J. Allergy Clin. Immunol. Pract.* 4, 1160–1166.e10.
- Ramsey, B. W., Davies, J., McElvaney, N. G., Tullis, E., Bell, S. C., Dřevínek, P., et al. (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *N. Engl. J. Med.* 365, 1663–1672.
- Regier, A. A., Farjoun, Y., Larson, D. E., Krashenina, O., Kang, H. M., Howrigan, D. P., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* 9:4038. doi: 10.1038/s41467-018-06159-4
- Rehder, C., Bean, L. J. H., Bick, D., Chao, E., Chung, W., Das, S., et al. (2021). Next-generation sequencing for constitutional variants in the clinical laboratory, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* 1–17. doi: 10.1038/s41436-021-01139-4
- Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* 18, 696–704.
- Ribeil, J.-A., Hacein-Bey-Abina, S., Payen, E., Magnani, A., Semeraro, M., Magrin, E., et al. (2017). Gene therapy in a patient with sickle cell disease. *N. Engl. J. Med.* 376, 848–855.
- 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.* 17, 405–424. doi: 10.1038/gim.2015.30
- Robinson, P. N., Köhler, 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
- Rode, J. (2005). *Rare Diseases: Understanding this Public Health Priority*. Paris: EURORDIS.
- Roth, E. M., Taskinen, M.-R., Ginsberg, H. N., Kastelein, J. J., Colhoun, H. M., Robinson, J. G., et al. (2014). Monotherapy with the PCSK9 inhibitor alirocumab versus ezetimibe in patients with hypercholesterolemia: results of a 24 week, double-blind, randomized Phase 3 trial. *Int. J. Cardiol.* 176, 55–61. doi: 10.1016/j.ijcard.2014.06.049
- Rowlands, C. F., Taylor, A., Rice, G., Whiffin, N., Hall, H. N., Newman, W. G., et al. (2021). MRSD: a novel quantitative approach for assessing suitability of RNA-seq in the clinical investigation of mis-splicing in Mendelian disease. *medRxiv* [Preprint] 2021.03.19.21253973. doi: 10.1101/2021.03.19.21253973
- Rozman, J., Rathkolb, B., Oestereicher, M. A., Schütt, C., Ravindranath, A. C., Leuchtenberger, S., et al. (2018). Identification of genetic elements in metabolism by high-throughput mouse phenotyping. *Nat. Commun.* 9:288.
- Samocho, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., et al. (2017). Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* [Preprint] 148353. doi: 10.1101/148353
- Samocho, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46, 944–950. doi: 10.1038/ng.3050
- Sankar, P. L., and Parker, L. S. (2017). The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. *Genet. Med.* 19, 743–750. doi: 10.1038/gim.2016.183
- Satterlee, J. S., Chadwick, L. H., Tyson, F. L., McAllister, K., Beaver, J., Birnbaum, L., et al. (2019). The NIH common fund/roadmap epigenomics program: successes of a comprehensive consortium. *Sci. Adv.* 5:eaw6507. doi: 10.1126/sciadv.aaw6507
- Schulz, M. H., Köhler, S., Bauer, S., and Robinson, P. N. (2011). Exact score distribution computation for ontological similarity searches. *BMC Bioinformatics* 12:441. doi: 10.1186/1471-2105-12-441
- Seaby, E. G., and Ennis, S. (2020). Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Brief. Funct. Genomics* 19, 243–258. doi: 10.1093/bfpg/ela009
- Sexton, A., Sahhar, M., Thorburn, D., and Metcalfe, S. (2008). Impact of a genetic diagnosis of a mitochondrial disorder 5–17 years after the death of an affected child. *J. Genet. Couns.* 17, 261–273. doi: 10.1007/s10897-007-9145-9
- Sifrim, A., Popovic, D., Tranchevent, L.-C., Ardehshirdavani, A., Sakai, R., Konings, P., et al. (2013). eXtasy: variant prioritization by genomic data fusion. *Nat. Methods* 10, 1083–1084.
- Simpson, A. (2016). *The Hidden Costs of Rare Diseases: A Feasibility Study*. London: Genetic Alliance UK.
- Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* 94, 599–610. doi: 10.1016/j.ajhg.2014.03.010
- Small, K. W., DeLuca, A. P., Whitmore, S. S., Rosenberg, T., Silva-Garcia, R., Udar, N., et al. (2016). North Carolina macular dystrophy is caused by dysregulation of the retinal transcription factor PRDM13. *Ophthalmology* 123, 9–18. doi: 10.1016/j.ophtha.2015.10.006
- Smedley, D., Jacobsen, J. O., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* 10, 2004–2015. doi: 10.1038/nprot.2015.124
- Smedley, D., Schubach, M., Jacobsen, J. O., 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
- Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., Bult, C. J., and Mouse Genome Database Group (2017). Mouse genome database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.* 46, D836–D842. doi: 10.1093/nar/gkx1006
- Smith, R. A., Miller, T. M., Yamanaka, K., Monia, B. P., Condon, T. P., Hung, G., et al. (2006). Antisense oligonucleotide therapy for neurodegenerative disease. *J. Clin. Invest.* 116, 2290–2296. doi: 10.1172/jci25424
- Soden, S. E., Saunders, C. J., Willig, L. K., Farrow, E. G., Smith, L. D., Petrikin, J. E., et al. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci. Transl. Med.* 6:265ra168. doi: 10.1126/scitranslmed.3010076

- Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nat. Rev. Genet.* 19, 453–467. doi: 10.1038/s41576-018-0007-0
- Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet. Med.* 21, 2413–2421. doi: 10.1038/s41436-019-0554-6
- Stark, Z., Schofield, D., Alam, K., Wilson, W., Mupfeki, N., Macciocca, I., et al. (2017). Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. *Genet. Med.* 19, 867–874. doi: 10.1038/gim.2016.221
- 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
- Stranger, B. E., Bringham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., et al. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease The eGTEx Project. *Nat. Genet.* 49, 1664–1670. doi: 10.1038/ng.3969
- Tan, T. Y., Dillon, O. J., Stark, Z., Schofield, D., Alam, K., Shrestha, R., et al. (2017). Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. *JAMA Pediatr.* 171, 855–862. doi: 10.1001/jamapediatrics.2017.1755
- Tarailo-Graovac, M., Shyr, C., Ross, C. J., Horvath, G. A., Salvarinova, R., Ye, X. C., et al. (2016). Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* 374, 2246–2255.
- Thevenon, J., Duffourd, Y., Masurel-Paulet, A., Lefebvre, M., Feillet, F., El Chehadeh-Djebbar, S., et al. (2016). Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. *Clin. Genet.* 89, 700–707. doi: 10.1111/cge.12732
- Timms, K. M., Wagner, S., Samuels, M. E., Forbey, K., Goldfine, H., Jammulapati, S., et al. (2004). A mutation in PCSK9 causing autosomal-dominant hypercholesterolemia in a Utah pedigree. *Hum. Genet.* 114, 349–353. doi: 10.1007/s00439-003-1071-9
- Turinsky, A. L., Choufani, S., Lu, K., Liu, D., Mashouri, P., Min, D., et al. (2020). EpigenCentral: portal for DNA methylation data analysis and classification in rare diseases. *Hum. Mutat.* 41, 1722–1733. doi: 10.1002/humu.24076
- 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
- Turner, T. N., Douville, C., Kim, D., Stenson, P. D., Cooper, D. N., Chakravarti, A., et al. (2015). Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Hum. Mol. Genet.* 24, 5995–6002. doi: 10.1093/hmg/ddv309
- Unlu, G., Qi, X., Gamazon, E. R., Melville, D. B., Patel, N., Rushing, A. R., et al. (2020). Phenome-based approach identifies RIC1-linked Mendelian syndrome through zebrafish models, biobank associations and clinical studies. *Nat. Med.* 26, 98–109. doi: 10.1038/s41591-019-0705-y
- Wai, H. A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., et al. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet. Med.* 22, 1005–1014. doi: 10.1038/s41436-020-0766-9
- Wainwright, C. E., Elborn, J. S., Ramsey, B. W., Marigowda, G., Huang, X., Cipolli, M., et al. (2015). Lumacaftor–ivacaftor in patients with cystic fibrosis homozygous for Phe508del CFTR. *N. Engl. J. Med.* 373, 220–231.
- Watkins, D., Schwartzentruber, J. A., Ganesh, J., Orange, J. S., Kaplan, B. S., Nunez, L. D., et al. (2011). Novel inborn error of folate metabolism: identification by exome capture and sequencing of mutations in the MTHFD1 gene in a single proband. *J. Med. Genet.* 48, 590–592. doi: 10.1136/jmedgenet-2011-100286
- Wojcik, M. H., Stewart, J. E., Waisbren, S. E., and Litt, J. S. (2020). Developmental support for infants with genetic disorders. *Pediatrics* 145:e20190629. doi: 10.1542/peds.2019-0629
- Wright, C. F., FitzPatrick, D. R., and Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* 19, 253–268. doi: 10.1038/nrg.2017.116
- Yoon, P. W., Olney, R. S., Khoury, M. J., Sappenfield, W. M., Chavez, G. F., and Taylor, D. (1997). Contribution of birth defects and genetic diseases to pediatric hospitalizations: a population-based study. *Arch. Pediatr. Adolesc. Med.* 151, 1096–1103. doi: 10.1001/archpedi.1997.02170480026004
- Zurynski, Y., Deverell, M., Dalkeith, T., Johnson, S., Christodoulou, J., Leonard, H., et al. (2017). Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. *Orphanet J. Rare Dis.* 12:68.

**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.

Copyright © 2021 Seaby, Rehm and O'Donnell-Luria. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.