


REVIEW

Open Access



A guide for the diagnosis of rare and undiagnosed disease: beyond the exome

Shruti Marwaha^{1,2*} , Joshua W. Knowles^{1,3} and Euan A. Ashley^{1,2,4*}

Abstract

Rare diseases affect 30 million people in the USA and more than 300–400 million worldwide, often causing chronic illness, disability, and premature death. Traditional diagnostic techniques rely heavily on heuristic approaches, coupling clinical experience from prior rare disease presentations with the medical literature. A large number of rare disease patients remain undiagnosed for years and many even die without an accurate diagnosis. In recent years, gene panels, microarrays, and exome sequencing have helped to identify the molecular cause of such rare and undiagnosed diseases. These technologies have allowed diagnoses for a sizable proportion (25–35%) of undiagnosed patients, often with actionable findings. However, a large proportion of these patients remain undiagnosed. In this review, we focus on technologies that can be adopted if exome sequencing is unrevealing. We discuss the benefits of sequencing the whole genome and the additional benefit that may be offered by long-read technology, pan-genome reference, transcriptomics, metabolomics, proteomics, and methyl profiling. We highlight computational methods to help identify regionally distant patients with similar phenotypes or similar genetic mutations. Finally, we describe approaches to automate and accelerate genomic analysis. The strategies discussed here are intended to serve as a guide for clinicians and researchers in the next steps when encountering patients with non-diagnostic exomes.

Keywords: Diagnosis, Rare, Omics, Exome-negative, Long read

Background

Although the occurrence of individual rare diseases often seems negligible, it is estimated that 30 million people in the USA are suffering from a rare disease, affecting 1 in 10 Americans, equivalent to the prevalence of type 2 diabetes [1, 2]. About 7000 rare disorders are defined [2, 3] and many others fall under the umbrella of undiagnosed diseases. Most patients suffering from a rare or undiagnosed disease receive only symptomatic treatment. An accurate diagnosis can result in better management of the disease, identification of potential therapeutics and avoid unnecessary treatments that may have severe side effects. For inherited rare diseases, knowing the causative variant and the mode of inheritance informs patients about

the risk of passing the disease to future generations and helps evaluate alternate family planning options [4]. The diagnostic delay for rare diseases varies from months to decades, depending on the patient's phenotype, age, and available resources. The average time for accurate diagnosis of a rare disease is about 4–5 years [5–7]; in some cases, it can take over a decade [8, 9]. These patients face a diagnostic odyssey and often undergo extensive and expensive workups at several institutions. Despite this, patients often remain undiagnosed or even misdiagnosed [8], which further adds emotional distress to patients and family members.

It is estimated that 80% of rare diseases have a genetic origin [13]. Until 10 years ago, genetic testing was expensive and usually limited to a few genes at a time. The advent of next-generation sequencing technology has had a dramatic effect on the cost, accuracy, and utility of genetic testing and has supplanted older technologies.

*Correspondence: mshruti@stanford.edu; euan@stanford.edu

¹ Department of Medicine, Division of Cardiovascular Medicine, School of Medicine, Stanford University, Stanford, CA, USA

Full list of author information is available at the end of the article



Many undiagnosed diseases [14, 15] have been identified by exome sequencing (ES) that looks at the protein-coding regions, which constitute less than 2% of the genome [16]. Additionally, sequencing family members and performing segregation analysis can eliminate hundreds of non-causative variants and thus reduce the search space. In a cohort of children with undiagnosed developmental disorders ($n=989$) and unaffected parents, Wright et al. [17] observed that exome sequencing of the parent–child trios rather than singletons reduced candidate variants by ten folds. Clark et al. [18] performed a meta-analysis on five studies consisting of children with suspected genetic diseases ($n=3613$) to compare the diagnostic yield of genome sequencing (GS)/ES by individual proband and trio testing within cohorts. They found that the odds of diagnosis using trios was double that using singletons.

Programs like Care for Rare [19], Deciphering Developmental Disorders [20], Rare and Undiagnosed Diseases Diagnostic Service [21], and the Undiagnosed Diseases Network [22] have demonstrated how exome sequencing can not only end an expensive, potentially invasive and emotionally challenging journey for the patients but also help in better disease management [23, 24]. Still, a minority of patients receive a definitive molecular diagnosis [17, 25–28]. This review, aimed towards clinicians and rare disease researchers, presents the key challenges in diagnosing patients with negative exome sequencing and discusses the strategies that can potentially fill the diagnostic gap in such patients. We propose technologies that should be considered when ES is unrevealing, many of which complement each other (Fig. 1). These include sequencing the whole genome and the transcriptome. Based on the patient's phenotype, metabolomics, or proteomics or methyl profiling should be considered. In parallel, automated processes should be established for periodic re-analysis of the genomic data and identification of patients with similar phenotypes or similar genetic mutations. Finally, functional studies should be conducted to support the causality of a putative variant and understand the molecular mechanism of the rare disease.

Genome sequencing

ES can capture the protein-coding regions of the genome and in some cases also untranslated regions (UTRs) and intron-exon boundaries [29], at a low cost. In addition, augmented exome capture techniques can further improve the coverage in medically relevant genes [30]. Despite the numerous advancements of exome sequencing, it has non-uniform coverage (particularly in first exons, regions of high GC/AT, and regions of low complexity [31–33]) and is limited by the specificity of the capture probes [34]. ES has had modest success in

detecting structural variants, tandem repeats, and pathogenic variants in deep intronic regions. Some of these challenges can be addressed by genome sequencing (GS) [31, 33, 35, 36]. GS can identify canonical [37, 38] and complex structural variants [39, 40], tandem repeats [37, 38], intronic variants [37, 38], and coding variants that may not be accurately captured by ES. GS has enabled identification of the causative variants for many undiagnosed cases where prior ES was either unrevealing [38, 41] or had provided only partial diagnosis (the causative variant explained only some phenotypes of the patient) [42]. Diagnosis mediated by GS has also opened avenues for therapy in some cases by identifying the disease mechanism and potential drug targets [42–44]. In this section, we illustrate with examples how short-read genome sequencing technology can facilitate detection of structural variants and tandem repeats that are often missed by ES. We discuss different long-read sequencing platforms, its advantages over short-read, especially for detecting large, complex structural variants and methylation changes and explore the potential of pan-genome reference in aiding rare disease diagnostics.

Structural variants

Structural variants (SVs) represent a class of variants that are greater than 50 base pairs (bp) [10, 45–47] and can be as long as 3Mb [48, 49]. Structural variants include microscopic and often submicroscopic variants that comprise deletions, duplications, insertions, inversions, mobile element insertions (transposons), translocations, and complex rearrangements [10]. Since SVs often encompass several exons or genes, GS is a better tool for studying them than ES. With the advent of PCR-free library preparations, population frequency databases [47, 50, 51], benchmarking structural variant datasets [52], and recent advancements in SV detection algorithms [46, 53], many groups have implemented GS to identify pathogenic SVs in previously undiagnosed patients [39, 54–57]. In a cohort of 477 undiagnosed patients with varied phenotype, Holt et al. [55] identified molecular diagnoses for 16 cases (3.35%) by scanning for structural variants using short-read GS data. Carss et al. [57] showed that in a phenotypically heterogeneous group of 722 inherited retinal disease (IRD) patients, 33 pathogenic structural variants were responsible for the disease in 31 (4.29%) individuals. Despite the ability of GS to detect large and complex variants, their interpretation remains difficult, especially for the non-coding variants [58]. The challenges associated with identifying the causative variant from ES/GS can be broadly classified into two groups — (i) interpretation (VUS in a known disease gene, novel disease gene or non-coding variant) and (ii) detection (missing second variant in a recessive disorder, causative

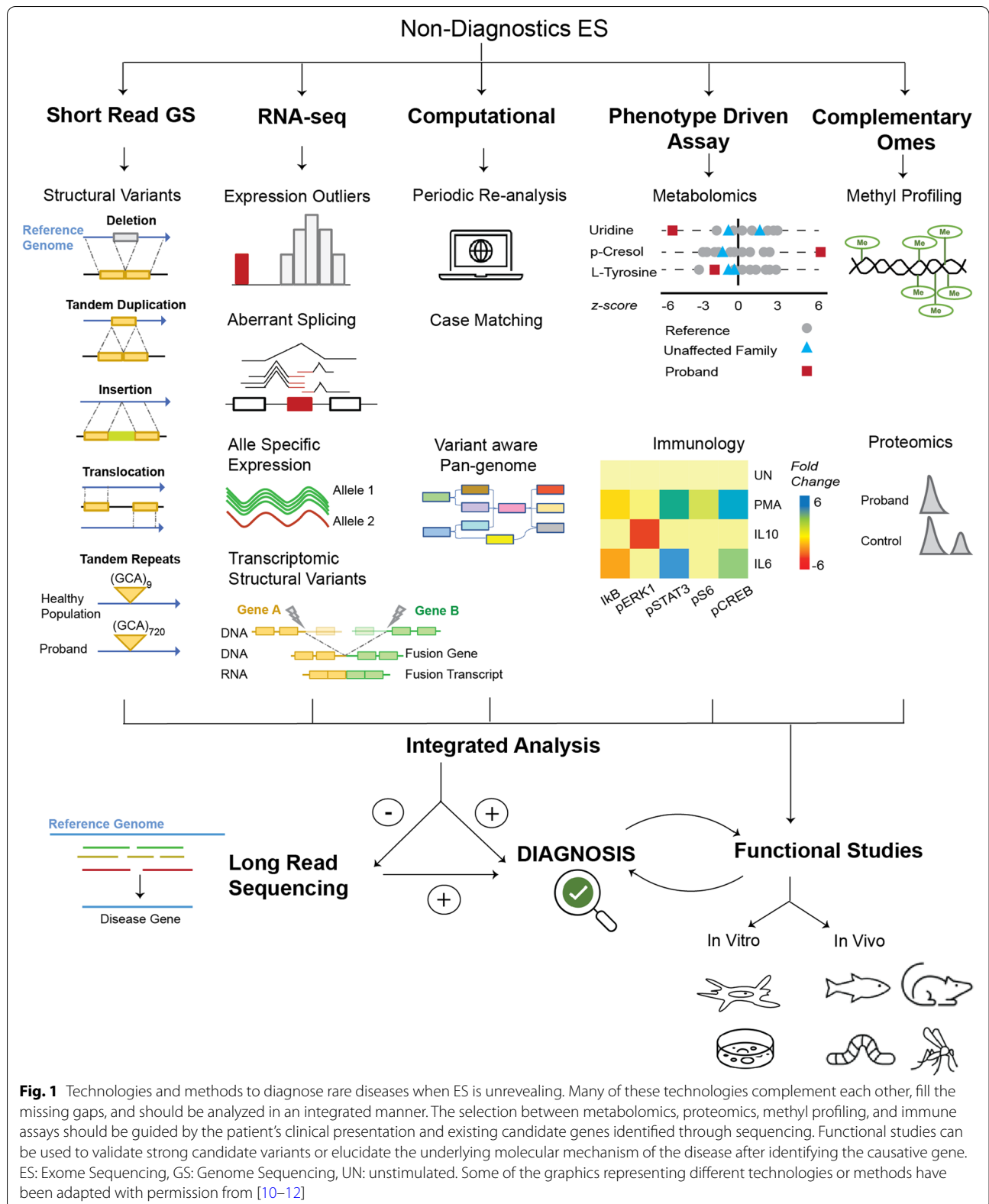


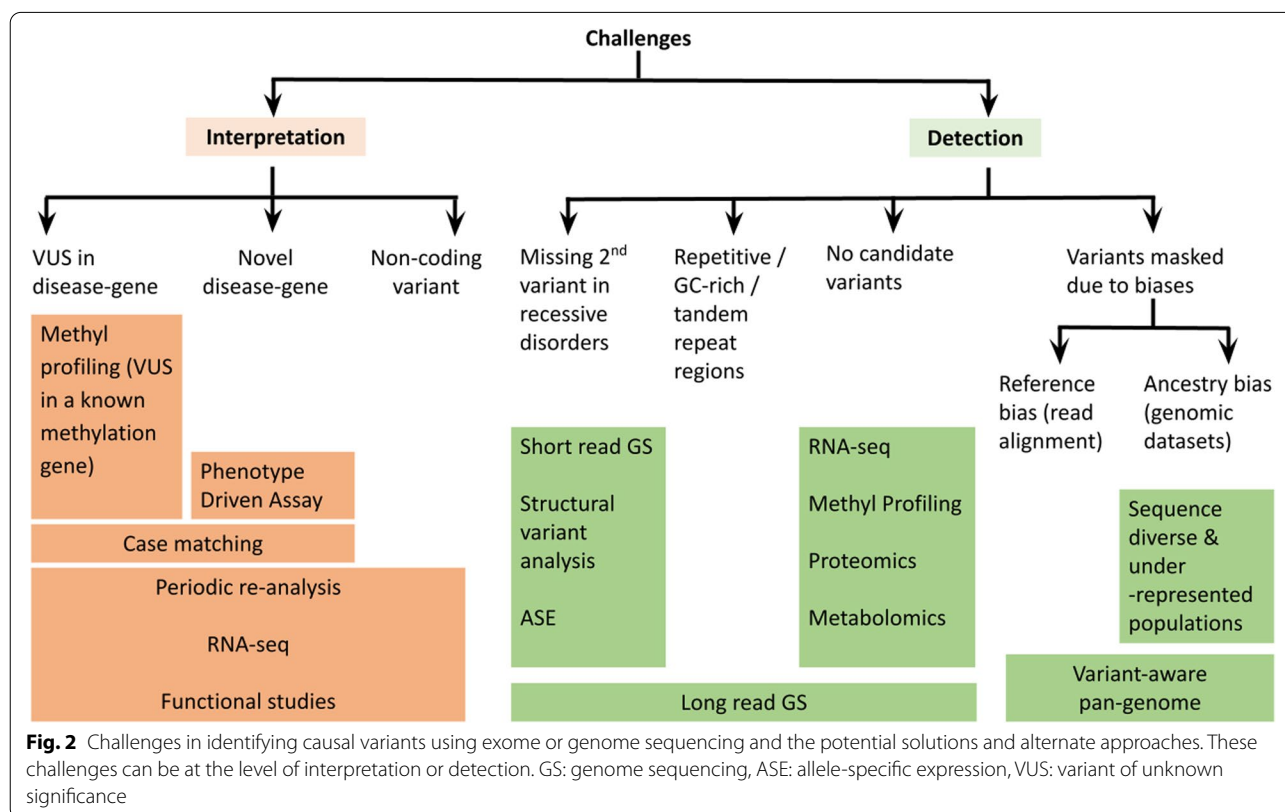
Fig. 1 Technologies and methods to diagnose rare diseases when ES is unrevealing. Many of these technologies complement each other, fill the missing gaps, and should be analyzed in an integrated manner. The selection between metabolomics, proteomics, methyl profiling, and immune assays should be guided by the patient’s clinical presentation and existing candidate genes identified through sequencing. Functional studies can be used to validate strong candidate variants or elucidate the underlying molecular mechanism of the disease after identifying the causative gene. ES: Exome Sequencing, GS: Genome Sequencing, UN: unstimulated. Some of the graphics representing different technologies or methods have been adapted with permission from [10–12]

variant lies in difficult to sequence region or is masked due to biases in reference genome and genomic datasets). In Fig. 2, we summarize these challenges and suggest alternate technologies, informatics tools, and experimental approaches that may help to reach a molecular diagnosis. With active development in variant prioritization algorithms (like genomiser [59], SpliceAI [60]), new disease-gene discoveries, and complementary omic technologies (like RNA-seq, metabolomics, and proteomics), we expect the diagnostic yield by GS will continue to improve.

Short tandem repeats (STRs) are short (1–6bp) DNA sequences repeated head-to-tail multiple times. Approximately 3% of the human genome consists of STRs [61] and 6% of human coding regions are estimated to contain STR variation [62]. Expansion of non-coding repeats can result in loss of protein function or altered RNA function while expansion of coding repeats can cause altered protein function [63]. STRs have been implicated in many neurological and genetic diseases like Friedreich’s ataxia (GAA), Huntington’s disease (CAG), Fragile X Syndrome (CGG), amyotrophic lateral sclerosis (GGGGCC), and other hereditary ataxias [63–68]. Historically, STRs were genotyped using polymerase chain reaction (PCR) and gel electrophoresis, which is time consuming, costly and limited to identifying expanded repeats in regions

previously associated with STR diseases. In past years, many bioinformatics tools like GangSTR [69], ExpansionHunter Denovo [70, 71], and STRetch [72] have been developed to predict STRs from PCR-free short-read sequencing. A recent study [73] benchmarked 8 STR prediction tools using known disease-causing full-mutation STR expansions and simulated data and found that the ensemble approach of using ExpansionHunter, STRetch, and exSTRA performed the best. These tools have enabled diagnoses of many Mendelian diseases caused by repeat expansion [74–76]. We discuss one of the cases in detail to demonstrate how STR analysis can guide diagnosis when ES is inconclusive.

Using GS and clinical and biochemical phenotyping, Kuilenburg et al. [75] identified expansion of GCA-repeat region in 5’UTR of glutaminase gene (*GLS*) in three unrelated patients with an inborn error of metabolism that resulted in reduced glutaminase activity. Initially, exome sequencing of the three patients and their families identified heterozygous, damaging variants in *GLS* gene in probands 1 and 3. Although a good phenotypic match, the ES finding was not conclusive and prompted further biochemical analysis and genome sequencing of proband 1. Expansion Hunter tool that identifies repeat expansions in a locus-specific manner, predicted a large GCA-repeat expansion in proband 1 when compared



with control population ($n=8295$). Later, large GCA-repeat expansions in the *GLS* were confirmed in all three patients using triplet repeat-primed PCR assay, facilitating molecular diagnoses for the three patients.

Although short-read sequencing can potentially identify known and novel SVs and repeat expansions across the genome, it has limited success in detecting large, complex structural variants and long tandem repeats or those which lie in highly repetitive and/ or GC-rich regions [77]. In the following section, we address how some of these limitations can be overcome by long-read sequencing.

Long-read sequencing

Reads from short-read sequencing (SRS) — typically 100–300bp long [78]— are mapped to a consensus reference during the alignment step. Nowadays, paired-end sequencing (sequencing both ends of a fragment) is often performed over single-end, allowing more accurate mapping of the reads, especially in regions with repetitive sequences. However, the alignment process is still challenging in repetitive regions of the reference genome because of the short length of reads, making it difficult to predict large variants and long tandem repeats with high certainty. In the last decade, new sequencing technologies have been developed that generate reads that typically range 10–60kb [79], with some extending to 2Mb [80]. The longer reads result in improved alignment to the reference genome and better detection of SVs, especially within repetitive elements or segmental duplications or high GC content, regions that were difficult to access using short-read technology [81].

Long-read sequencing (LRS) also allows haplotype phasing — assigning genetic variants to the homologous paternal or maternal chromosomes [82, 83]. This information helps in identifying compound heterozygous mutations and de novo autosomal dominant mutations [84]. Most variant callers developed for short-read sequencing provide unphased variants and thus require sequencing of parents to detect compound heterozygous and de novo mutations. LRS also provides precise details about the breakpoints [85], improving our understanding of the mutation and disease mechanism.

Currently, there are two main LRS platforms — single-molecule real-time (SMRT) sequencing from Pacific Biosciences [86] and nanopore-based sequencing from Oxford Nanopore Technologies (ONT) [87]. ONT can generate very long contiguous reads (2.2Mb) [80] while Circular Consensus Sequencing (by Pacific Biosciences) provides highly accurate (99.8%) high-fidelity (HiFi) reads with an average length of 13.5kb [88]. Cost-effective synthetic LRS technologies like linked reads (Transposase Enzyme Linked Long-read Sequencing (TELL-seq)

[89], single-tube long fragment read (stLFR) [90], Hi-C and chromatin cross-linking [91, 92], and optical mapping (from BioNano Genomics) [93] can provide several advantages of LRS (e.g., detection of large SV) at some additional cost. We suggest the review by Sedlazeck et al. [81] and Sakamoto et al. [94] for further in-depth comparison of these technologies.

LRS also enables direct detection of methylated nucleotides [95–97]. Among all types of methylation modifications in DNA, 5-methylcytosine (5mC) is most well studied, partly due to the advancements in bisulfite-based short-read sequencing techniques like whole genome bisulfite sequencing (WGBS) [98] and reduced representation bisulfite sequencing (RRBS) [99, 100]. Although bisulfite sequencing provides a quantitative and accurate measure of 5mC modifications at base resolution, it cannot capture other methylation changes including 6-methyladenine (6mA) and 4-methylcytosine (4mC). In contrast, long-read technology sequences native DNA and can predict the base modifications from the deviations seen in the raw signal, thus avoiding DNA amplification and bisulfite conversion steps and the biases associated with them [101]. Both nanopore technology and SMRT sequencing can capture many types of base modifications (including 5mC, 4mC, and 6mA) simultaneously. In Pacific Biosciences' SMRT sequencing, DNA polymerase adds labeled nucleotides along the template DNA, generating a succession of fluorescence pulses. Base modification can alter the kinetics of the polymerase during this process. If a nucleotide is methylated, DNA polymerase will pause before incorporating the next nucleotide. Changes in fluorescence pulses from the labeled nucleotides are used to measure the shift in polymerization speed, thus detecting the base modification. For example, the time interval between two successive fluorescence pulses called inter-pulse duration is used to detect 6mA [81, 95]. Recently, Tse OYO et al. [102] developed a method that drastically improved the detection of 5mC modifications from SMRT sequencing using sequence context, inter-pulse duration, and pulse width associated with DNA polymerase kinetics. ONT's technology identifies one of the 5 possible nucleotides based on the difference in electrical current produced when the base passes through protein nanopores embedded in a flow cell. Base modifications on the DNA or RNA cause a minor shift in current that can be detected and interpreted by algorithms [96, 103]. Comparison of performance of nanopore, SMRT, and bisulfite-based short-read sequencing on the same set of samples will inform the community about benefits and limitations of each technology and which method is most suitable under a given situation.

Using low coverage LRS, Merker et al. [56] identified a 2184-bp deletion in a patient with negative targeted clinical testing and unrevealing SRS of the genome. Maio et al. [4] demonstrated how LRS can help to solve recessive disease cases where the second pathogenic allele is missing from the ES data. Although SR GS is capable of identifying repeat expansions, it is limited in detecting undiscovered repeat diseases in long [104] and complex GC-rich regions [105]. Recent studies have proved that LRS can detect known [106, 107] and novel repeat expansions [104, 108, 109] for Mendelian diseases in which no causal variants were detected through SRS. The longer reads can encompass an entire expanded repeat or a flanking unique sequence, making long-read technology apt for analyzing tandem repeat expansions [104]. Facioscapulohumeral muscular dystrophy 1 (FSHD1) disease results from a heterozygous contraction of 3.3 kb repeat unit (referred as D4Z4) in the subtelomeric region of chromosome 4q35 and a chromosome 4 haplotype called 4qA. The D4Z4 unit varies from 11 to 100 repeats in the healthy population, but FSHD1 patients show only 1–10 repeats, hence the term contraction [110]. Conventionally Southern Blotting is used for molecular diagnosis of FSHD1 and alternative methods have been explored as Southern Blot is semi-quantitative and time consuming. However, sequencing long, repetitive subtelomeric regions of the genome is challenging for both short-read technology and Sanger sequencing. Moreover, D4Z4 has variable number of repeat units and homologous repeat array on chromosome 4 and chromosome 10. Both true LR (SMRT [111] and ONT's Minion [112]) and synthetic LR (BioNano Genomics' optical mapping) [113] technologies have shown variable degrees of success in sequencing this region, determining the repeat number and haplotype. With continuing improvements, long-read technologies can enable sequencing of more such difficult-to-sequence regions, identifying new associations between genomic regions and genetic disorders. Several groups have exemplified the clinical significance of targeted long-read sequencing using CRISPR/Cas9 [114] mediated methods [115, 116] or computational adaptive sampling [117] for enrichment of specific regions of the genome. In a small cohort of 22 patients with known canonical and complex SVs, Miller et al. [117] demonstrated that targeted LRS can not only detect all SVs previously identified with clinical testing ($n=46$) but also discovered variants ($n=41$) that were missed by the clinical test. Targeted long-read sequencing has the potential to be used clinically for patients with suspected complex SVs and tandem repeats in candidate genes.

In summary, LRS can be the single test to detect single-nucleotide variants (SNVs), insertion and deletion (INDELs), simple and complex SVs, tandem repeats,

and methylation changes and inform about phasing. Although LRS holds promise for undiagnosed genetic diseases, some challenges need to be overcome in order to bring LRS from the research setting to the clinic. The cost of ONT is now comparable to SR but Pacific Biosciences is relatively expensive. Recently, great strides have been made in sequencing technology and algorithm development to improve the accuracy of calling small variants (SNVs and INDELs) from nanopore and HiFi long-read data [88, 118]. At high coverage, both long-read sequencing platforms can outperform the short-read-based method in accurate SNV identification at whole genome scale, including segmental duplication and difficult-to-map regions. SR and HiFi have comparable performance at identifying INDELs but ONT has much lower accuracy. Identification of base modifications (epigenetic changes) by LRS is still in its infancy and suffers from low accuracy and the need for training models [101]. However, continuous improvements in sequencing technology and algorithm development are being made to further increase the accuracy and lower the cost, which will eventually enable the use of LRS routinely for clinical diagnostics. In the interim, an attractive solution is a hybrid approach combining the advantages of each in a combined assay using, for example, lower coverage (e.g., 15 \times) long-read sequencing along with higher coverage (e.g., 40 \times) short-read sequencing or using targeted LRS to evaluate candidate genes or in case of suspected tandem repeat disease or complex rearrangements.

Pan-genome reference

The current human reference genome is a linear haploid consensus sequence derived from a very small number of individuals and thus lacks genetic diversity observed across populations [119, 120]. Mapping sequencing reads to this reference genome can cause the reads to be misaligned or remain unaligned, especially in highly polymorphic or repetitive regions or regions spanning structural variant breakpoints [119, 121] or may miss a rare variant that is represented by the minor allele on the haploid reference sequence [122]. This results in a "reference bias" as non-reference alleles from a sample are difficult to align to the linear reference sequence. To overcome these limitations, many efforts have been made in the past few years to incorporate known variants in the reference in order to allow variant-aware read alignment and variant calling [119, 120, 123–125]. These efforts propose a pan-genome, which represents a collection of all genomic sequences in a population or a species or a phylogenetic clade [123, 126].

Aligning reads to a pan-genome that considers many alternate haplotypes at each locus reduces the reference bias [127], thereby improving alignment accuracy

and variant calling [119, 124]. Recently, Siren et al. [120] developed a tool called Giraffe to map short reads to pan-genome with high accuracy and speed. Giraffe detected SNVs, INDELS, and SVs more accurately when using a pan-genome than using the single reference genome, showcasing the significance and practicality of the pan-genomic approach to short-read mapping. It was able to genotype 167,000 SVs that were discovered from LR studies, in 5202 individuals from diverse populations that were sequenced by SR sequencing. Recently, precisionFDA truth challenge V2 evaluated different bioinformatics pipelines' accuracy in predicting small variants in difficult-to-map regions and Major Histocompatibility Complex using Genome In A Bottle (GIAB) benchmark data set [128]. The top performing algorithms in the short-read sequencing category used either alt-aware mapping (DRAGEN's graph mapper) or pan-genome (by Seven Bridges' GRAF pipeline). Despite the several benefits of the pan-genome, there are practical limitations associated that have hindered the community in embracing this paradigm shift. This includes high compute cost, scalability, and complexity of the tasks. Addition of variants to the existing linear reference genome is not straightforward as simply adding more variation to the reference can result in more ambiguity. Alternative methods have been proposed that aim to strike a balance between accuracy and limitations of graph-based pan-genome. Reference flow [129] involves an iterative two-step process. Reads are first aligned to the linear reference genome and the unaligned reads and the reads with low mapping-quality are then re-aligned to a set of references. Tetikol et al. [130] recommend population-specific graphs that iteratively augment tailored genome graphs for targeted populations.

By reducing the reference bias, the graph genome will be instrumental in detecting novel structural variants, large INDELS, and mutations that affect allele-specific expression [126, 131, 132]. The pan-genomic model can help to detect more accurate variants for rare disease patients from underrepresented populations [130] and even allow construction of personalized reference genome using the parent's sequencing data (if available). To the best of our knowledge, pan-genome has not yet led to a diagnosis; however, efforts are being made in this direction [133]. In the next few years, we expect further optimizations in speed and accuracy of the tools working in the pan-genome space. Since use of pan-genome and variant-aware algorithms lead to more accurate variant detection in SR sequencing, especially the structural variants, we anticipate that these approaches will benefit diagnoses of ultra-rare patients.

Transcriptomics

Although genome sequencing can theoretically capture all types of variants, prioritization and interpretation of the non-coding variants remains a big challenge. Complementing DNA sequencing with transcriptomics can help to prioritize potential disease-causing variants. In this section, we review four approaches to analyze RNA sequencing data for prioritizing candidate genes for rare diseases — expression outliers, aberrant splicing, allele-specific expression, and transcriptomic structural variants [11, 134–136]. Next, we discuss the potential of long-read sequencing to predict alternative splicing and gene fusions with high accuracy. We also highlight the potential of single-cell transcriptomics to elucidate the cellular and molecular mechanisms of rare and undiagnosed diseases that involve rare, undiscovered cell populations.

RNA sequencing can help to classify a variant of unknown significance (VUS) and provide insights into the disease mechanism or identify variant in the second allele in a recessive disease where genomic sequencing returned only one pathogenic variant [137]. In a cohort of 50 patients with rare muscle disorders, who had non-diagnostic ES and/GS, Cummings et al. [138] illustrated the utility of RNA sequencing the affected tissue (muscle), yielding a diagnostic rate of 34%.

Gene expression and mRNA isoforms can vary significantly from one tissue to another [139, 140] and so it is recommended to use the affected tissue for RNA sequencing [138, 141]. But the disease-relevant tissue is not always easily available in a non-invasive manner. Blood, fibroblasts, and induced pluripotent stem cells (iPSCs) appear to be promising alternatives [11, 134, 141, 142]. By RNA sequencing blood from 94 undiagnosed patients, representing 16 distinct disease categories, Fressard et al. [134] identified the causative variants in 7.5% of the cases, demonstrating the potential of blood transcriptome sequencing to aid the diagnoses of rare Mendelian diseases. Lee et al. [142] reported a 14.5% ($n=7$) diagnostic rate by sequencing mRNA from blood, fibroblast, and/or muscle samples from 48 genome-negative individuals, primarily affected by neurological ($n=25$) and musculoskeletal disorders ($n=12$). They identified pathogenic splicing abnormalities in seven patients with neurological or musculoskeletal diseases. They observed that fibroblast was a better tissue choice than blood for identifying the splicing defects in this cohort. Similarly, Baynam et al. reported a case of megalencephaly-capillary malformation syndrome where the causative mutation (mosaicism in PIK3CA) was detected in fibroblasts and not in blood [21].

However, many genes are expressed at very low levels in both blood and fibroblasts to be captured at high

depth by RNA sequencing. CRISPR/Cas9 technology can be used to improve coverage of low-expressed genes in a scalable manner [143]. Huang et al. [144] applied CRISPRclean method, using Cas9 nuclease and 360,000 guide RNAs to specifically remove RNA-Seq library fragments from over 4000 targeted genes and observed about a six-fold increase in coverage of untargeted genes compared to untreated RNA-Seq libraries. iPSCs are a good substitute when the candidate gene is known to be expressed at low levels in blood and fibroblast. Recently, Bonder et al. [145] unified data from five major iPSC genetic studies [146–150] to create the integrated iPSC QTL (i2QTL) consortium. They observed a fivefold enrichment of outliers in known rare disease genes as compared to non-disease genes and demonstrated detection of gene outliers in patients with Bardet-Biedl syndrome and hereditary cerebellar ataxia. Therefore, alternate tissues like fibroblasts, iPSCs, and blood should be considered carefully when the affected tissue is not available for transcriptome analysis. In the following section, we will discuss how sequencing the transcriptome can uncover pathogenic mutations, missed by studying genomic variants alone.

Expression outliers

When working with rare and undiagnosed diseases, it is assumed that most of the samples express each gene within its physiological range and the goal is to identify genes from each sample that are expressed at extremely high or low levels. This is achieved by calculating *Z*-scores, comparing each patient against others in the cohort. GTEX [140] and GEUVADIS [151] are great resources for additional control RNA-seq samples.

Caution should be exercised when applying the expression outlier approach. For example, controls should be from same tissue type as the disease samples [135]; data should be normalized for batch effect, sex, or biopsy site [11]. Typically, the *Z*-score-based approach uses an arbitrary threshold for selecting outlier genes [134, 138, 141] often followed by applying additional filters like predicted pathogenicity, minor allele frequency, and phenotypic match to further prune down the number of candidate genes [134]. Recent methods like OUTRIDER [152] and PEER [153] control for technical and biological variations among genes and the former also provides a statistical test for outlier detection in RNA-seq samples. Fresard et al. [134] demonstrated how their expression outlier pipeline prioritized a causative gene (*MECR*) within the top 15 candidate genes for two siblings with MEPAN disease. Overall, analyzing expression outliers along with genomic variants and the patient's phenotype can be a powerful strategy to identify strong candidate variants for clinical interpretation.

Aberrant splicing variants

Alternative splicing is a naturally occurring phenomenon in eukaryotes that results in a single gene coding for multiple proteins. Post transcription, non-coding sequences (introns) are removed from the pre-mRNA and some exons may be included or excluded from the final, processed mRNA [154]. Errors in this process cause several diseases including rare Mendelian diseases [155]. Splicing mutations can be broadly divided into five categories: exon skipping, inclusion of intronic pseudoexon, exon extension, exon retraction, and intron retention [142, 156]. Algorithms like LeafCutterMD [157] and Fraser [158] provide statistical frameworks that are designed for predicting splicing outliers in rare diseases.

Certain types of variants like synonymous and deep intronic variants are often filtered out by prioritization pipelines unless they have been previously associated with a disease. Such variants can lead to aberrant splicing events and it is possible to re-prioritize them using transcriptomic data [134, 158]. Lee et al. [142] have shown how RNA sequencing helped to identify the second variant in a 2-year-old girl who had an inconclusive trio ES, which reported a paternally inherited frameshift mutation in *SEPSECS* gene (OMIM 613009), associated with autosomal recessive pontocerebellar hypoplasia, type 2d [159]. GS did not reveal any pathogenic maternally inherited coding variant. However, transcriptomics data from the proband and the mother showed that half of their reads in the *SEPSECS* gene skipped exon 7, which carried a synonymous variant. This was missed earlier because typically, synonymous variants are filtered during variant prioritization of genomic data unless they are previously reported to be pathogenic.

Allele-specific expression

Allele-specific expression (ASE) is a phenomenon in diploid or polyploid genomes, where one allele has significantly higher expression than the other allele [160, 161]. When prioritizing variants from ES/GS data using recessive mode of inheritance, single heterozygous rare variants are filtered out. However, some of these heterozygous rare variants may exhibit ASE. Gonorazky et al. [141] reported that the allele imbalance approach provided diagnostic leads in three monogenic neuromuscular disorder patients, who previously had non-diagnostic ES and/or gene panel results. Kremer et al. [11] discuss how their ASE pipeline helped to establish the genetic diagnosis in a patient with mucopolisidosis, who had tested negative for the enzymatic tests available for mucopolisidosis type 1, 2, and 3 in blood leukocytes. They detected borderline non-significant low expression in an intronic variant in *MCOLNI* gene that was filtered by their ES pipeline as it was intronic. Therefore, along with

identifying expression outliers and splicing variants, ASE analysis should be performed as part of regular RNA-seq analysis, especially when genomic data identifies only one heterozygous variant for a recessive disorder.

Transcriptomic structural variants

Structural variants (SVs) like translocations, duplications, inversions, and deletions join different genomic regions together or separate one region into pieces. Transcription of such regions can result in gene fusions (exons from two or more distinct genes are transcribed together) or cause a previously non-transcribed region to be included into a gene, often leading to altered gene function in both the cases. Such modifications in the transcribed mRNA that are caused by genomic SVs are known as transcriptomic structural variants (TSVs) [162, 163].

Fusion genes are well documented in hematological and solid tissue cancers and are used as biomarkers for early diagnosis and therapeutic targets [164]. Independent case studies have reported fusion transcripts in many non-cancer diseases like brain malformation [165, 166], intellectual disability [167, 168], spastic paraplegia [169], and Gille de la Tourette Syndrome [170]. Oliver et al. [136] tailored a fusion identification pipeline for rare disease patients and applied it to a cohort of 47 individuals who previously had negative or partial diagnoses through exome sequencing. They identified eight fusion events that were confirmed using orthogonal methods, of which 2 provided clinical diagnoses for patients' phenotypes. They identified a paternally inherited pathogenic frameshift INDEL in *ATM* in an infant with T cell lymphopenia using trio exome sequencing. Pathogenic *ATM* variation causes ataxia-telangiectasia in an autosomal recessive manner but the patient's exome data did not reveal a second trans variant in *ATM*. RNA sequencing of the patient's fibroblasts identified reciprocal *ATM*-*SLC35F2* and *SLC35F2*-*ATM* fusion transcripts suggesting chromosomal inversion that was later confirmed by targeted long-read sequencing of the putatively affected introns [136]. Recently, Cmero et al. showed how using RNA sequencing alone allowed discovery of an interchromosomal translocation in the *DMD* gene in a patient with muscular dystrophy [171]. Thus, integrated analysis of transcriptomic and genomic data should be considered to detect structural variants that may result in gene fusions.

Long-read transcriptomics

Short-read RNA sequencing is a well-established and superior technique for gene expression quantification compared to microarray. However, the fragmented, short-length reads makes computational reconstruction of transcripts challenging, especially for complex genes

or gene families containing many similar isoforms [79, 81]. Long-read technology can determine the sequence of full-length RNA transcripts by sequencing the cDNA (Pacific Biosciences and ONT) or the native RNA (ONT). The longer reads can span the sequence of the entire transcript and thus determine the underlying exon combinations [79, 81]. Therefore long-read RNA sequencing can improve the analysis of alternate splicing, potentially leading to discovery of novel isoforms and novel gene fusions. Recent studies have identified many new relevant isoforms using long-read RNA sequencing in healthy [172–174] and disease states [175, 176]. Long-read RNA sequencing also allows identification of allele-specific expression through haplotype phasing [177, 178].

The high depth required for clinical long-read RNA sequencing currently makes it cost inefficient for regular genetic diagnoses. Like DNA sequencing, targeted long-read RNA sequencing is a good alternative to investigate disease-relevant genes. Dainis et al. [179] performed targeted long-read genome and transcriptome sequencing to interrogate a putative splice-site-altering mutation in *MYBPC3* gene in a hypertrophic cardiomyopathy (HCM) patient. Comparing long-read transcriptomics data for *MYBPC3* from this HCM patient to that in three additional HCM patients and six control hearts, they identified two isoforms that were exclusively seen only in the patient under question. This study exemplifies how LRS can easily characterize alternatively spliced isoforms and link the improperly spliced transcripts to variant-associated alleles.

To summarize, transcriptome-wide long-read sequencing allows detection of full-length transcripts, alternative spliced isoforms, gene fusions, transcript-based haplotype phasing, allele-specific expression, and base modifications in RNA. Although, to the best of our knowledge, long-read RNA sequencing is yet to solve an undiagnosed disease, the technology holds the promise for rare Mendelian disorders, especially when the only one heterozygous variant is identified in a recessive disease.

Single-cell transcriptomics

Although bulk RNA sequencing has the potential to identify the molecular cause and disease mechanism in rare disorders, it only captures average expression signal in the sample, which may comprise different cell types. In comparison, single-cell RNA sequencing (scRNA-seq) measures expression of genes within each cell, allowing researchers to study the sample heterogeneity and cell-to-cell variation. This enables discovery of new and rare cell types, improving our understanding of the disease mechanism. Montoro et al. used scRNA-seq on mouse tracheal epithelium to study cellular heterogeneity and identified a new and extremely rare cell

type—pulmonary ionocyte [180]. They showed that these ionocytes expressed *CFTR* gene at much higher levels than any other cell type in both mouse and human airway tissue. Mutations in *CFTR* have been extensively reported in cystic fibrosis disease, and for years, the gene was thought to be expressed at low levels in ciliated cells that are common and distributed throughout the airway.

However, like most new technologies, scRNA-seq is associated with technical challenges (low capture efficiency and extremely sparse data) and high cost as compared to bulk RNA-seq [181, 182]. scRNA-seq data is sparse, with many observed zeros, indicating that a given gene in a particular cell has no unique molecular identifiers or reads mapping to it. This could represent real biology (truly silent gene) or a technical artifact (gene is expressed but was not detected by the scRNA sequencing) [181, 183]. One alternative approach to scRNA-seq is to extrapolate cellular components of the sample from bulk RNA-seq using deconvolution methods. There are more than 50 deconvolution methods published to date, that can be broadly categorized as marker-based (uses marker gene list for deconvolution), reference-based (for the deconvolution process, it uses cell type specific gene expression profiles and list of differentially expressed genes across the cell types in the reference), and reference-free (uses reference profiles for cluster annotation after the deconvolution step) [184–187].

scRNA-seq has enabled discovery of many rare, novel cell types or sub-cell populations or markers in many different tissues—like blood [188], brain [189, 190], pancreas [191], and cancer [192, 193] to list a few, and with continued improvement in the single-cell sequencing technology and algorithms, we anticipate its application to be extended to rare disease research in future. Comprehensive characterization of transcriptome in each cell may allow discovery of new cellular and molecular components in rare disease patients' tissues and can be instrumental in elucidating the disease mechanism.

Complementary technologies

Integrating sequencing data with other technologies can also provide leads to discover the underlying mutation in undiagnosed diseases where sequencing is inconclusive [194, 195]. Here, we provide examples from metabolomics [12, 196], methyl profiling [197], proteomics [194], and immunology [198–200] that assisted in a patient's genetic diagnosis. The choice of assay is often driven by the patient's phenotype.

Methylation profiling

Epigenetic modifications like DNA methylation and histone modification have shown to have important implications in rare diseases like Immunodeficiency Centromeric

instability Facial syndrome 1, Rett syndrome, and Rubinstein-Taybi [201–203]. Methylation profiling should be considered when there is suspicion of a genomic imprinting disorder or a VUS in a known methylation gene. Genomic imprinting is a phenomenon where a subset of autosomal genes is preferentially expressed from only one of the two parental chromosomes. This results from parental-specific methylation of cytosine at CpG dinucleotides of genes during gametogenesis [204, 205]. DNA methylation defects can be divided into two groups — epi-variants [206] and epi-signatures [207, 208]. Epi-variants involve a change in DNA methylation pattern of a small number of CpGs at a specific region of the genome whereas epi-signatures are unique combinations of DNA methylation changes at multiple loci across the genome and are specific for different genetic syndromes.

Technologies like RRBS, WGBS, and long-read sequencing can be used to assess genome-wide DNA methylation. Aref-Eshghi et al. [197] developed a machine learning model using genome-wide DNA methylation data from blood to predict 14 different Mendelian syndromes with neurodevelopmental presentations and congenital anomalies (ND/CA) that are associated with epi-signature. By applying this model to a cohort of 965 ND/CA patients, who previously had unrevealing conventional genetic testing including CNV microarray or ES, they identified 15 cases with one of the 14 Mendelian syndromes. They also identified 12 patients with imprinting and trinucleotide repeat expansion disorder and 106 cases with rare epi-variants in this cohort. This work led to development of EpiSign, a clinical-grade genome-wide DNA methylation assay for patients with developmental delay or suspicion of imprinting, trinucleotide repeat expansion, or one of the 50 methylation-related disorders [209]. In a recent study, Sadikovic et al. evaluated the clinical utility of EpiSign in a cohort of 207 patients that was divided into two subgroups — a targeted cohort, which included patients with inconclusive VUS and a screening cohort that comprised of patients with clinical findings consistent with hereditary neurodevelopmental syndromes but no previous conclusive genetic findings [210]. EpiSign enabled diagnoses for 35.3% (48/136) of participants in the targeted cohort and 11.3% (8/71) of those in the screening cohort.

Clinical interpretation of rare epi-variants remains challenging, especially those in intragenic regions or in genes not yet associated with the patient's phenotype. Another limitation of this approach is the lack of large-scale databases of epi-variants in Mendelian diseases and population epigenome data that can be used as a reference and to differentiate between tolerant versus pathogenic epi-variants. Moreover, some of the disorders may

not exhibit epi-signatures or epi-variants in blood and may be tissue specific.

Metabolomics

For many rare unexplained metabolic disorders in children, where the causative variant was identified by ES, functional metabolomic studies have helped to uncover the disease mechanism [211–213] and even led to better disease management or treatment in some cases [24]. Tarailo-Graovac et al. [24] used targeted metabolomics to confirm the causality of mutations detected by ES in several individuals among a cohort of 41 patients with intellectual development disorder and unexplained metabolic phenotype. Splinter et al. [25] demonstrated how findings from metabolomics in an undiagnosed patient with multi-system disorder prompted re-analysis of exome sequencing data, followed by RNA sequencing, and led to the diagnosis. They identified consistently high levels of urinary organic acids in the patient, suggesting a deficiency in 3-hydroxy-3-methylglutaryl coenzyme A lyase (encoded by *HMGL* gene). Re-evaluation of the ES data identified a deletion in exon 1 of *HMGL*. RNA sequencing the patient's fibroblast revealed a 50% lower level of *HMGL* expression as compared with fibroblasts from eight unaffected individuals. Although metabolomics and lipidomics can potentially provide diagnostic leads, the metabolic changes in rare and undiagnosed diseases may be subtle or confounded by a patient's special diet or medication, making the analysis challenging.

Proteomics

Proteins are the final component of central dogma and the effector molecules of a cell. Proteomics has a lower throughput as compared to other 'omes, yet it can reveal impairment in protein synthesis, stability, degradation, and signaling, which may result in a disease state. Two broad categories of methods commonly used to study proteome are mass spectrometry-based and antibody-based techniques. In 2019, Grabowski et al. [194] demonstrated that mass spectrometry-based proteome analysis guided targeted genetic diagnostics and uncovered the underlying genomic mutations in two patients, which were initially missed by ES due to sequencing limitations. They studied the proteome of three rare monogenic diseases of neutrophil granulocytes — severe congenital neutropenia (SCN), leukocyte adhesion deficiency (LAD), and chronic granulomatous disease (CGD). They interrogated 4154 proteins from 16 patients with one of the three monogenic diseases of neutrophil granulocytes and 68 healthy controls. ES was unable to provide molecular diagnoses for two patients in this cohort, one with CGD and another patient with congenital neutropenia associated with albinism. For both the cases, top

10 deregulated proteins from the proteome analysis provided hints for the causative mutations — *NCF1* for CGD case and *RAB27A* for the second case with congenital neutropenia and albinism. These were missed by ES analysis because sequencing the *NCF1* gene is challenging as it shares 99% homology with two pseudogenes while re-examination of second's patients sequencing data showed that the last part of exon2 in *RAB27A* was not covered by sequencing reads.

Antibody-based cytometry techniques — flow and mass cytometry — allow to study cellular heterogeneity and phospho-signaling within each cell. Although cytometry has not yet led to diagnosis in rare disease patients, they can provide molecular clues and improve our understanding of the disease, especially for inborn errors of immunity [214, 215]. Kanolkar et al. [200] showed how findings from flow cytometry (reduced phosphorylation of STAT1 in B cells upon IFN- γ stimulation and attenuated STAT5 phosphorylation in T cells upon IL-2 stimulation) in a patient prompted genomic analysis of 132 immunologically relevant genes that revealed a compound heterozygote mutation in *IFNGR1* in the proband.

Many of the aforementioned technologies complement each other, fill the missing gaps, and better inform about the molecular pathophysiology of the disease. Although there are several successful examples of the integration of ES/GS with either transcriptomics [134, 142], metabolomics [25, 213], or proteomics [216], a single framework integrating different omics is lacking. Existing tools [217–220] for combining and analyzing multiple omics data were designed for the standard case-control studies and are not suitable for outlier-based analysis. Taking a systems biology approach by integrating results from several omics may further improve the diagnostic yield and our understanding of the disease's molecular mechanism.

Functional studies

Unraveling the molecular mechanisms of a putative disease-causing gene can help strengthen the case for causality and may provide insights for developing therapeutics. This can be achieved by modeling patients' disease-causing variant or strong candidate variants in vivo using model systems like fruit flies (*Drosophila melanogaster*), nematode worm (*Caenorhabditis elegans*), zebrafish (*Danio rerio*), mouse (*Mus musculus*) [221–224], or in vitro (disease-relevant mouse or human cell lines, primary cells or induced pluripotent stem cell models, iPSCs). Such models can be a fast and cost-effective way to mirror complex rare genetic disorders. The choice of the model system depends on the cost, time, and the ability to model and assess the patient's phenotype in the animal [223].

Several groups have used organisms like *Drosophila*, *C. elegans*, and zebrafish to model patient's mutations to (i) validate novel disease-gene associations [225], (ii) provide functional data [226], (iii) generate new biological insights [227], and (iv) even identify potential therapeutic targets [228, 229]. Splinter et al. [25] demonstrated that modeling candidate variants in *Drosophila* and zebrafish played an important role in the diagnoses of eight patients in a cohort of 382. Functional studies in *Drosophila* confirmed causation of a de novo variant in *NR5A1* gene in a patient with a 46,XX genotype and male sex characteristics [230], which later led to characterization of a new syndrome. Similarly, Kanca et al. [231] performed functional studies in *Drosophila* to establish that de novo variants in *WDR37* gene cause a novel syndromic neurological disorder and Ferreira et al. [232] used zebrafish to model variants causing Saul-Wilson syndrome.

Another useful resource to recapitulate the unique aspects of patients' disease pathology is their own cells, which can be grown into fibroblasts or induced pluripotent stem cells (iPSCs) [233]. iPSCs can be potentially differentiated into virtually any cell type with the appropriate environmental stimuli. iPSCs are especially valuable for rare disorders that affect inaccessible tissues such as neurons [234] and cardiomyocytes [235]. Modeling of disease-relevant cell types has allowed better understanding of disease pathogenesis in many rare diseases [233] like those involving neurons (ALS [236], Friedreich's ataxia [237], ataxia-telangiectasia [238]), cardiomyocytes (long QT syndrome [239], Fabry disease [240], Jervell and Lange-Nielsen syndrome [241, 242]), blood (Fanconi anemia [243], Glanzmann thrombasthenia [244]), connective tissue (Fibrodysplasia ossificans progressiva [245]), and eye (Retinitis pigmentosa [246, 247]). Yamashita et al. [248] used iPSCs to model monogenic skeletal diseases like Thanatophoric Dysplasia type 1 (TD1) and achondroplasia (ACH) and to identify clinically effective treatment for these diseases. The authors converted fibroblasts from TD1 and ACH patients into iPSCs and demonstrated that the chondrogenic differentiation of TD1 iPSCs and ACH iPSCs resulted in the formation of degraded cartilage. Next, they showed that statins, a class of drugs already approved for lowering lipids, could correct the degraded cartilage in both chondrogenically differentiated TD1 and ACH iPSCs. These results were then reproduced in mice, suggesting that statins might be an effective drug for patients with TD1 and ACH.

ES and GS yield numerous VUS, non-coding variants within functional regulatory elements and variants that disrupt splicing. Existing computational prediction algorithms have had limited success in prioritizing them.

Functional screening assays are a powerful platform to assess the impact of variants in thousands of genes in a single experiment. Such screening approaches include germline mutagenesis, CRISPR/CAS9, plasmid-based reporter assays, RNA interference, chemical screens [249], and multiplexed assays of variant effect [250, 251]. Advancements in CRISPR/Cas9 technology make it a robust tool to profile cellular phenotypes resulting from each of the thousand genetic perturbations in a high-throughput manner [252]. The underlying principle behind CRISPR screens [253, 254] is to introduce thousands of variants in a large cell population but only one gene is perturbed per cell. This results in a population of cells with a different gene disrupted in each cell. Then, sequencing is performed on the mixed population of cells to identify genetic sequences necessary for the cell's survival or a specific cellular phenotype of interest. CRISPR screens target multiple sites per gene and thus introduce random variants in the gene of interest that may not represent the exact mutation observed in the patient. However, this approach informs whether a particular genomic region may have a functional role in the disease and can narrow down to a few promising candidates for follow-up studies. CRISPR screens can be implemented to study the effect of knockout (CRISPRko), inhibition (CRISPRi), or activation (CRISPRa) of many protein-coding genes [253, 255] or non-coding regulatory elements [256, 257].

CRISPR screens have recently been used to link new genes to rare diseases [258], to understand the molecular mechanism through which different variants in a gene contribute to a disease [259] and to explore potential therapeutic targets [260]. Rao et al. [259] used a pooled CRISPR screen in human hematopoietic stem and progenitor cells (HSPCs) to study mutations in *ELANE* which is known to cause severe congenital neutropenia (SCN), a rare genetic disorder characterized by low circulating neutrophils caused by impaired neutrophil maturation. Missense and frameshift mutations in *ELANE* account for 50% of SCN cases. The authors performed a dense mutagenesis CRISPR screen in primary human HSPCs to identify *ELANE* variants associated with neutrophil maturation defects. Although CRISPR screens hold great promise for prioritization of variants identified by ES and GS, there are some practical limitations including the need to study a large number of cells (10^8) and the fact that current CRISPR screens may not model the specific prioritized variants from the patient.

Therefore, functional studies including model organisms, fibroblasts, iPSCs, and CRISPR/CAS9 screens can play a vital role in diagnosis, understanding the molecular mechanism of rare and undiagnosed diseases and exploring potential therapeutic strategies. However, each model system has its own limitations. This process is

time consuming, and moreover, none of the model systems can completely replicate human disease.

Case matching

A major challenge faced by rare disease researchers is the lack of phenotypically similar patients to establish the molecular cause of the disease and to conduct statistical analysis. Several algorithms and platforms [261–264] have been developed to discover cases with common phenotypes and disrupted genes. However, there was a lack of a federated network that would facilitate interaction between various rare disease databases in a streamlined and continuous manner. To address this, Matchmaker Exchange (MME) was launched in 2015 [265] to identify unrelated cases with a potentially pathogenic variant in the same candidate gene and overlapping phenotype. It performs genomic matching across several databases (like DECIPHER [266], GeneMatcher [267], PhenoCentral [264]) in a scalable, secure, and automated fashion through a standardized application programming interface (API). As of October 2021, MME contains information from more than 150,000 cases from 88 countries [268]. It has facilitated identification of cases with similar phenotypic and genotypic profiles for many rare diseases including 25 novel gene-disease associations and phenotype expansions. MME also allows queries against published animal models that match a patient's phenotype, connecting the clinician with model organism researcher.

Some rare mutations can cause dysmorphic and unique facial features. Integrating the patients' variant and phenotypic data with their facial features (images) can significantly narrow down the search for potential rare syndromes. Using 17,000 images representing more than 200 syndromes, Gurovich et al. [269] developed a deep neural network to classify distinctive facial features in photos of patients with congenital and neurodevelopmental disorders. Machine learning is also being applied to large electronic health record (EHR) databases to identify rare as well as common disease patients [270–272]. This can help to find patients who have similar disease trajectories (like symptoms, age, medications, labs, or procedures) but may not have a definite diagnosis.

Automated re-analysis

It is important to periodically re-analyze sequencing data with latest analytical pipelines, variant frequency databases, literature [273–275], and updates in patient's phenotype that can potentially identify recent associations between the causative gene and patient's symptoms. Several groups have demonstrated that re-evaluation of genomic data can increase the diagnostic yield: by 5–26% in case of ES [276–279] and by 4–11%

for GS [280, 281] re-analysis. Use of standard ontologies like Human Phenotype Ontology (HPO) [282] can help to prioritize candidate genes that have been previously linked to the patient's phenotype. Tools like exomiser [283], amelie [284], and Xrare [285] search for the gene-phenotype associations in human diseases (documented in databases like OMIM [286], Orphanet [287], or PubMed) and also in other models like mouse, zebrafish, and protein-protein interaction networks. Using ES from 134 diagnosed rare retinal diseases, Cipriani et al. [288] demonstrated that exomiser tool ranked the causative variant as top hit in 74% of the dataset and among top 5 in 94%. Deeply phenotyping undiagnosed patients and identifying the most relevant symptoms is critical. However, these patients have an extensive, complex medical history, and their symptoms are documented in long clinical records. Tools like ClinPhen [289] and CLiX [290, 291] can extract relevant phenotypes from clinical notes or EHR data and convert them to HPO terms, thus enabling development of an automated pipeline for phenotype-based prioritization of variants.

Along with improving the diagnostic yield for rare disease patients, clinicians and rare disease researchers would like to reduce the diagnosis time frame for all patients. It is obvious that delay in accurate diagnosis leads to inappropriate disease management and sometimes even unnecessary treatments that can have severe side effects. To scale genomic analysis and implement it at a clinical level in a secure fashion, many groups are leveraging the power of cloud computing (like Amazon Web Services [<https://aws.amazon.com/>], Google Cloud Platform [<https://cloud.google.com/>]), and even new hardware like DRAGEN (Dynamic Read Analysis for Genomics) has been designed for faster turnaround of results. DRAGEN implements FPGA (Field Programmable Gate Array) technology, an alternative to conventional CPU-based systems to expedite the execution of genome pipelines. Recently, some groups have demonstrated record-breaking fast implementation of GS using optimized SR/LR sequencing, DRAGEN/multiple cloud computing machines, and semi-automated downstream analysis to diagnose children with suspected Mendelian disorder, who were critically ill and admitted to ICU (fastest, 7 h, 18 min) [292, 293].

Challenges

Accurately diagnosing rare disease patients involves challenges at technical, financial, and policy levels. Some of the technical obstacles include interpretation of non-coding variants and VUS. This requires use of advanced technologies (like LRS, RNA-seq, epigenomics) and algorithms (like SplicAI, genomiser) to decipher the role of non-coding regions in healthy and disease conditions.

Periodic and automated re-analysis of genomic data can help to resolve some of the VUS and intronic variants as new disease-gene discoveries are being made at an accelerated pace. There is also a need for specialized methods to analyze and integrate multiomics data for rare diseases.

In 2016, Kessler et al. [294] demonstrated ancestry specific bias in genomic datasets and its impact on diagnostic accuracy and cost. Human variant databases like 1000 Genomes [295], ESP [296, 297], and Exac [298] catalogs the frequency of variants from large populations. Mutations absent or present at very low frequency in these databases are prioritized as potential causes of rare Mendelian diseases, based on the assumption that variants common in the general population are unlikely to cause a rare or undiagnosed disease. However, these databases are skewed towards European ancestry populations [299–301] which makes interpretation of ES/GS from an individual with non-European inheritance more difficult, expensive, and time consuming. Recent efforts like gnomAD [302], GenomeAsia 100K [303], and All of Us [304] have been initiated to sequence more diverse and under-represented populations. Thus, it is critical to query the allele frequencies of non-European patients from their respective ancestry and for the scientific community to fill the ethnicity gaps in the current genomic databases. A complementary approach is the use of the aforementioned pan-genome reference. Encoding the genetic diversity in the reference genome would benefit genomic analysis of non-European ethnicities by reducing the reference bias during alignment and thereby resulting in a more accurate variant calling [119, 124, 125].

One of the biggest challenges in the road to diagnosing rare disease patients is the cost. The technologies mentioned in this manuscript are often not available clinically or covered by patients' medical insurance [305] and are provided by few research programs in developed nations. Dimmock et al. [306] reported that rapid GS improved the disease management in 58 children in a cohort of 184 critically ill infants, who were admitted to ICU and reduced the hospital costs in 31 cases, by \$12,000–\$15,700 per child. Splinter et al. [25] compared the health care cost before and during the diagnosis evaluation period and found the latter to be only 6–7% of the total cost. Recently, Tisdale et al. [307] performed a pilot study on 14 rare diseases within four different healthcare system databases to estimate direct medical costs. They found that per patient direct medical costs of rare diseases are about 3–5 times higher than age matched controls, highlighting the urgent need for early and accurate diagnosis for rare disease patients that may reduce the costs associated with misdiagnosis or missed opportunities for intervention at an appropriate time. More of such

cost-effectiveness analyses are required to justify the cost of whole genome SRS or LRS to be covered by insurance and to bring a change in the policy. Also, there is a need for continuous funding to the existing research programs dedicated for diagnosing rare and yet-to-be-discovered diseases. With continued advancement in the technologies, we anticipate a decline in their cost will make them more affordable. Meanwhile, targeted sequencing and latest computational algorithms should be considered to address the challenges of detection and interpretation of genomic variants, along with machine learning approaches to identify similar patients.

Conclusions

In the past two decades, gene panels, microarrays, and ES have identified the underlying causal mutations for many rare disease patients; however, still a significant proportion of them remain undiagnosed. In this review, we summarize different approaches that can further improve the diagnostic yield and elucidate the molecular mechanism of the disease. We share examples where these technologies played a significant role in deciphering the causative mutation in undiagnosed patients.

These approaches include complementing short-read genome sequencing with RNA sequencing, metabolomics, proteomics, and methyl profiling. For patients with unrevealing short-read GS, long-read technology is a promising alternative. It is also important to functionally validate the candidate or causative variants identified through genomics using in vitro and in vivo model systems to improve our understanding of molecular mechanisms and to allow better disease management, even opening avenues towards therapeutics.

It is also critical to periodically implement fast, automated computational pipelines to identify new gene-disease associations or to find similar patients across the globe. Lately, the medical and genomics community has recognized and acknowledged the ancestry specific bias in the genomic datasets and in the haploid linear reference genome. Inclusion of diverse ethnicities in frequency databases and use of a pan-genome reference will help to improve the diagnostic accuracy for underrepresented populations.

A major bottleneck in the diagnosis of rare patients is the cost involved in the investigation: most of the assays mentioned in this review are research based and not yet available through health care systems. We anticipate that continuous improvements in accuracy and affordability of the high-throughput technologies will enable us to fill the diagnostic gap for undiagnosed patients, often with actionable findings. We envision that successful implementation of complementary multidisciplinary studies

will lead to a paradigm shift in how undiagnosed patients are diagnosed and treated.

Abbreviations

ES : Exome sequencing; GS : Genome sequencing; SVs : Structural variants; bp : Base pairs; UTRs : Untranslated regions; IRD : Inherited retinal disease; STRs : Short tandem repeats; PCR : Polymerase chain reaction; SRS : Short-read sequencing; LRS : Long-read sequencing; SMRT : Single-molecule real-time; ONT : Oxford Nanopore Technologies; HiFi : High-fidelity; WGBS : Whole genome bisulfite sequencing; RRBS : Reduced representation bisulfite sequencing; ASE : Allele-specific expression; VUS : Variant of unknown significance; TSVs : Transcriptomic structural variants; TD1 : Thanatophoric Dysplasia type 1; ACH : Achondroplasia; EHR : Electronic health record; HPO : Human Phenotype Ontology; DRAGEN: Dynamic Read Analysis for Genomics; FPGAs: Field Programmable Gate Array; iPSCs : Induced pluripotent stem cells; SNVs : Single-nucleotide variants; INDELS : Insertion and deletion; TELL-seq : Transposase Enzyme Linked Long-read Sequencing; stLFR : Single-tube long fragment read.

Authors' contributions

SM, JWK, and EAA contributed towards conception, design, and drafting of the manuscript. All authors have reviewed and approved the final version of the manuscript.

Funding

This work was supported by NIH Common Fund under Award Number U01HG010218, U41HG009649, R01 DK116750, R01 DK120565, US-ISRAEL BSF 2017265 and P30DK116074.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

EAA is co-founder to Personalis, DeepCell, Svexa; non-executive director to AstraZeneca and advisor to Genome Medical, Sequence Bio, Apple, Foresite Capital. The remaining authors declare that they have no competing interests.

Author details

¹Department of Medicine, Division of Cardiovascular Medicine, School of Medicine, Stanford University, Stanford, CA, USA. ²Stanford Center for Undiagnosed Diseases, Stanford University, Stanford, CA, USA. ³Department of Medicine, Diabetes Research Center, Cardiovascular Institute and Prevention Research Center, Stanford, CA, USA. ⁴Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA.

Received: 9 June 2021 Accepted: 10 February 2022

Published online: 28 February 2022

References

- National Diabetes Statistics Report, 2020 [Internet]. 2020 [cited 2021 May 18]. Available from: <https://www.cdc.gov/diabetes/data/statistics-report/index.html>
- RARE disease facts [Internet]. Global Genes. 2018 [cited 2021 Dec 20]. Available from: <https://globalgenes.org/rare-disease-facts/>
- Haendel M, Vasilevsky N, Unni D, Bologna C, Harris N, Rehm H, et al. How many rare diseases are there? *Nat Rev Drug Discov*. 2020;19:77–8.
- Miao H, Zhou J, Yang Q, Liang F, Wang D, Ma N, et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas*. 2018;155:32.
- Global Commission on Rare Disease [Internet]. [cited 2021 Dec 6]. Available from: <https://www.globalrare-disease-commission.com/Report>
- Accurate Diagnosis of Rare Diseases Remains Difficult Despite Strong Physician Interest - Global Genes [Internet]. Global Genes. 2014 [cited 2019 Aug 21]. Available from: <https://globalgenes.org/2014/03/06/accurate-diagnosis-of-rare-diseases-remains-difficult-despite-strong-physician-interest/>
- Yan X, He S, Dong D. 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*. 2020;17. Available from: <https://doi.org/10.3390/ijerph17051757>
- Molster C, Urwin D, Di Pietro L, Fookes M, Petrie D, van der Laan S, et al. Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet J Rare Dis*. 2016;11:30.
- Heuyer T, Pavan S, Vicard C. The health and life path of rare disease patients: results of the 2015 French barometer. *Patient Relat Outcome Meas*. 2017;8:97–110.
- Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011;12:363–76.
- Kremer LS, Bader DM, Mertes C, Kopajtic R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun*. 2017;8:15824.
- Kyle JE, Stratton KG, Zink EM, Kim Y-M, Monroe ME, et al. A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases. *Nature Scientific Data*: Bloodsworth KJ; 2021.
- RARE Facts - Global Genes [Internet]. Global Genes. [cited 2019 Aug 21]. Available from: <https://globalgenes.org/rare-facts/>
- Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu Y-F, McSweeney KM, et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med*. 2015;17:774–81.
- Pierson TM, Yuan H, Marsh ED, Fuentes-Fajardo K, Adams DR, Markello T, et al. GRIN2A mutation and early-onset epileptic encephalopathy: personalized therapy with memantine. *Ann Clin Transl Neurol*. 2014;1:190–8.
- Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*. 2009;106:19096–101.
- Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*. 2015;385:1305–14.
- Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. 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*. 2018;3:16.
- CARE for RARE [Internet]. CARE for RARE. [cited 2019 Aug 24]. Available from: <http://care4rare.ca>
- Firth HV, Wright CF, Study DDD. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol*. 2011;53:702–3.
- Baynam G, Pachter N, McKenzie F, Townshend S, Slee J, Kiraly-Borri C, et al. The rare and undiagnosed diseases diagnostic service - application of massively parallel sequencing in a state-wide clinical service. *Orphanet J Rare Dis*. 2016;11:77.
- Gahl WA, Wise AL, Ashley EA. The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. *JAMA*. 2015;314:1797–8.
- Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med*. 2011;13:255–62.
- Tarailo-Graovac M, Shyr C, Ross CJ, Horvath GA, Salvarinova R, Ye XC, et al. Exome Sequencing and the Management of Neurometabolic Disorders. *N Engl J Med*. 2016;374:2246–55.
- Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheattle-Jarvela AM, et al. Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N Engl J Med*. 2018;379:2131–9.

26. Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. *N Engl J Med*. 2017;376:21–31.
27. Balci TB, Hartley T, Xi Y, Dymont DA, Beaulieu CL, Bernier FP, et al. Debunking Occam's razor: Diagnosing multiple genetic diseases in families by whole-exome sequencing. *Clin Genet*. 2017;92:281–9.
28. Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, Tamburro C, et al. Genomic medicine for undiagnosed diseases. *Lancet*. 2019;394:533–40.
29. Guo Y, Long J, He J, Li C-I, Cai Q, Shu X-O, et al. Exome sequencing generates high quality data in non-target regions. *BMC Genomics*. 2012;13:194.
30. Patwardhan A, Harris J, Leng N, Bartha G, Church DM, Luo S, et al. Achieving high-sensitivity for clinical applications using augmented exome sequencing. *Genome Med*. 2015;7:71.
31. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17:507–22.
32. Wang Q, Shashikant CS, Jensen M, Altman NS, Girirajan S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep*. 2017;7:885.
33. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet*. 2016;135:359–62.
34. Goldfeder RL, Ashley EA. A precision metric for clinical genome sequencing [Internet]. *bioRxiv*. 2016 [cited 2021 Apr 9]. p. 051490. Available from: <https://www.biorxiv.org/content/10.1101/051490v1.abstract>
35. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. *Lancet*. 2010;375:1525–35.
36. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA*. 2014;311:1035–45.
37. Qaiser F, Sadoway T, Yin Y, Zulfiqar Ali Q, Nguyen CM, Shum N, et al. Genome sequencing identifies rare tandem repeat expansions and copy number variants in Lennox-Gastaut syndrome. *Brain Commun*. 2021;3:fab207.
38. Bergant G, Maver A, Peterlin B. Whole-Genome Sequencing in Diagnostics of Selected Slovenian Undiagnosed Patients with Rare Disorders. *Life* [Internet]. 2021;11. Available from: <https://doi.org/10.3390/life11030205>
39. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med*. 2018;10:95.
40. Palmer EE, Sachdev R, Macintosh R, Melo US, Mundlos S, Righetti S, et al. Diagnostic yield of whole genome sequencing after nondiagnostic exome sequencing or gene panel in developmental and epileptic encephalopathies. *Neurology*. 2021;96:e1770–82.
41. Zastrow DB, Kohler JN, Bonner D, Reuter CM, Fernandez L, Grove ME, et al. A toolkit for genetics providers in follow-up of patients with non-diagnostic exome sequencing. *J Genet Couns*. 2019;28:213–28.
42. Kim J, Hu C, Moufawad El Achkar C, Black LE, Douville J, Larson A, et al. Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. *N Engl J Med* [Internet]. 2019; Available from: <https://doi.org/10.1056/NEJMoa1813279>
43. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, et al. Whole-genome sequencing for optimized patient management. *Sci Transl Med*. 2011;3:87re3.
44. van Karnebeek CDM, Ramos RJ, Wen X-Y, Tarailo-Graovac M, Gleason JG, Skrypnik C, et al. Bi-allelic GOT2 Mutations Cause a Treatable Malate-Aspartate Shuttle-Related Encephalopathy. *Am J Hum Genet*. 2019;105:534–48.
45. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*. 2020;21:171–89.
46. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol*. 2019;20:246.
47. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*. 2020;583:83–9.
48. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006;7:85–97.
49. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics*. 2015;14:305–14.
50. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature*. 2020;581:444–51.
51. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. DbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Res*. 2013;41:D936–41.
52. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* [Internet]. 2020; Available from: <https://doi.org/10.1038/s41587-020-0538-8>
53. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019;20:117.
54. Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med*. 2019;21:1121–30.
55. Holt JM, Birch CL, Brown DM, Gajopathy M, Sosonkina N, Wilk B, et al. Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing [Internet]. *bioRxiv*. 2019 [cited 2019 Sep 9]. p. 627661. Available from: <https://www.biorxiv.org/content/10.1101/627661v1>
56. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med*. 2018;20:159–63.
57. Carrs KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, et al. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *Am J Hum Genet*. 2017;100:75–90.
58. Krude H, Mundlos S, Øien NC, Opitz R, Schuelke M. What can go wrong in the non-coding genome and how to interpret whole genome sequencing data. *Med Genet. De Gruyter*. 2021;33:121–31.
59. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am J Hum Genet*. 2016;99:595–606.
60. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li Yi, et al. Predicting splicing from primary sequence with deep learning. *Cell*. 2019;176:535–48.e24.
61. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860–921.
62. O'Dushlaine CT, Edwards RJ, Park SD, Shields DC. Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol*. 2005;6:R69.
63. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*. 2005;6:743–55.
64. Mirkin SM. Expandable DNA repeats and human disease. *Nature*. 2007;447:932–940.
65. Hunter J, Rivero-Arias O, Angelov A, Kim E, Fotheringham I, Leal J. Epidemiology of fragile X syndrome: a systematic review and meta-analysis. *Am J Med Genet A*. 2014;164A:1648–58.
66. Pringsheim T, Wiltshire K, Day L, Dykeman J, Steeves T, Jette N. The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov Disord*. 2012;27:1083–91.
67. Ruano L, Melo C, Silva MC, Coutinho P. The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. *Neuroepidemiology*. 2014;42:174–83.
68. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*. 2011;72:245–56.
69. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res*. 2019;47:e90.
70. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*. 2019;35:4754–6.

71. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt JJFA, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* 2020;21:102.
72. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* 2018;19:121.
73. Rajan-Babu I-S, Peng JJ, Chiu R. IMAGINE Study, CAUSES Study, Li C, et al. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med.* 2021;13:126.
74. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* 2017;27:1895–903.
75. van Kuilenburg ABP, Tarailo-Graovac M, Richmond PA, Drögemöller BI, Pouladi MA, Leen R, et al. Glutaminase deficiency caused by short tandem repeat expansion in GLS. *N Engl J Med.* 2019;380:1433–41.
76. Liu H-Y, Zhou L, Zheng M-Y, Huang J, Wan S, Zhu A, et al. Diagnostic and clinical utility of whole genome sequencing in a cohort of undiagnosed Chinese families with rare diseases. *Sci Rep.* 2019;9:19365.
77. Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun.* 2021;9:98.
78. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* 2016;17:333–51.
79. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* [Internet]. 2020; Available from: <https://doi.org/10.1038/s41576-020-0236-x>
80. Payne A, Holmes N, Rakyen V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics.* 2019;35:2193–8.
81. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet.* 2018;19:329–46.
82. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware diplotyping from noisy long reads. *Genome Biol.* 2019;20:116.
83. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet.* 2019;10:426.
84. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. *Nat Rev Genet.* 2011;12:215–23.
85. Kraft F, Wesseler K, Begemann M, Kurth I, Elbracht M, Eggermann T. Novel familial distal imprinting centre 1 (11p15.5) deletion provides further insights in imprinting regulation. *Clin. Epigenetics.* 2019;11:30.
86. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science.* 2009;323:133–8.
87. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol.* 2009;4:265–70.
88. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019;37:1155–62.
89. Chen Z, Pham L, Wu T-C, Mo G, Xia Y, Chang PL, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.* 2020;30:898–909.
90. Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, et al. Efficient and unique cobarcode of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* 2019;29:798–808.
91. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009;326:289–93.
92. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 2016;26:342–50.
93. Cao H, Hastie AR, Cao D, Lam ET, Sun Y, Huang H, et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience.* 2014;3:34.
94. Sakamoto Y, Zaha S, Suzuki Y, Seki M, Suzuki A. Application of long-read sequencing to the detection of structural variants in human cancer genomes. *Comput Struct Biotechnol J.* 2021;19:4207–16.
95. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods.* 2010;7:461–5.
96. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods.* 2017;14:411–3.
97. Gigante S, Gouil Q, Lucattini A, Keniry A, Beck T, Tinning M, et al. Using long-read sequencing to detect imprinted DNA methylation [Internet]. *Nucleic Acids Research.* 2019. p. e46–e46. Available from: <https://doi.org/10.1093/nar/gkz107>
98. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature.* 2008;452:215–9.
99. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 2005;33:5868–77.
100. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature.* 2008;454:766–70.
101. Gouil Q, Keniry A. Latest techniques to study DNA methylation. *Essays Biochem.* 2019;63:639–48.
102. Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, et al. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc Natl Acad Sci U S A* [Internet]. 2021;118. Available from: <https://doi.org/10.1073/pnas.2019768118>
103. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods.* 2017;14:407–10.
104. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. *J Hum Genet.* 2020;65:11–9.
105. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* 2013;23:121–8.
106. Zeng S, Zhang M-Y, Wang X-J, Hu Z-M, Li J-C, Li N, et al. Long-read sequencing identified intronic repeat expansions in SAMD12 from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. *J Med Genet.* 2019;56:265–70.
107. Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. *J Hum Genet.* 2019;64:191–7.
108. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, et al. Expansions of intronic TTTC A and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet.* 2018;50:581–90.
109. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NL associated with neuronal intranuclear inclusion disease. *Nat Genet.* 2019;51:1215–21.
110. Schätzl T, Kaiser L, Deigner H-P. Facioscapulohumeral muscular dystrophy: genetics, gene activation and downstream signalling with regard to recent therapeutic approaches: an update. *Orphanet J Rare Dis.* 2021;16:129.
111. Morioka MS, Kitazume M, Osaki K, Wood J, Tanaka Y. Filling in the Gap of Human Chromosome 4: single molecule real time sequencing of macrosatellite repeats in the facioscapulohumeral muscular dystrophy locus. *PLoS One.* 2016;11:e0151963.
112. Mitsuhashi S, Nakagawa S, Takahashi Ueda M, Imanishi T, Frith MC, Mitsuhashi H. Nanopore-based single molecule sequencing of the D4Z4 array responsible for facioscapulohumeral muscular dystrophy. *Sci Rep.* 2017;7:14789.
113. Dai Y, Li P, Wang Z, Liang F, Yang F, Fang L, et al. Single-molecule optical mapping enables quantitative measurement of D4Z4 repeats in facioscapulohumeral muscular dystrophy (FSHD). *J Med Genet.* 2020;57:109–20.

114. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337:816–21.
115. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38:433–8.
116. Ebbert MTW, Farrugia SL, Sens JP, Jansen-West K, Gendron TF, Prudencio M, et al. Long-read sequencing across the C9orf72 “GGGGCC” repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol Neurodegener*. 2018;13:46.
117. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet*. 2021;108:1436–49.
118. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods*. 2021;18:1322–32.
119. Rakocevic G, Semenyuk V, Lee W-P, Spencer J, Browning J, Johnson JJ, et al. Fast and accurate genomic analyses using genome graphs. *Nat Genet*. 2019;51:354–62.
120. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*. 2021;374:abg8871.
121. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nat Rev Genet*. 2020;21:243–54.
122. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet*. 2011;7:e1002280.
123. The Computational Pan-Genomics Consortium, Marschall T, Marz M, Abeel T, Dijkstra L, Dutilh BE, et al. Computational pan-genomics: status, promises and challenges. *Brief Bioinform*. Oxford Academic; 2016;19:118–135.
124. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 2018;36:875–9.
125. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37:907–15.
126. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome graphs. *Annu Rev Genomics Hum Genet*. 2020;21:139–62.
127. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. *Genome Res*. 2017;27:665–76.
128. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. precisionFDA Truth Challenge V2: calling variants from short- and long-reads in difficult-to-map regions [Internet]. *bioRxiv*. 2021 [cited 2021 Dec 20]. p. 2020.11.13.380741. Available from: <https://www.biorxiv.org/content/10.1101/2020.11.13.380741v4>
129. Chen N-C, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol*. 2021;22:8.
130. Serhat Tetikol H, Narci K, Turgut D, Budak G, Kalay O, Arslan E, et al. Population-specific genome graphs improve high-throughput sequencing data analysis: a case study on the Pan-African genome [Internet]. *bioRxiv*. 2021 [cited 2021 Dec 20]. p. 2021.03.19.436173. Available from: <https://www.biorxiv.org/content/10.1101/2021.03.19.436173v2>
131. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol*. 2020;21:35.
132. Satya RV, Zavaljevski N, Reifman J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res*. 2012;40:e127.
133. Markello C, Huang C, Rodriguez A, Carroll A, Chang P-C, Eizenga J, et al. A complete pedigree-based graph workflow for rare candidate variant analysis [Internet]. *bioRxiv*. 2021 [cited 2021 Dec 20]. p. 2021.11.24.469912. Available from: <https://www.biorxiv.org/content/10.1101/2021.11.24.469912v1>
134. Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med*. 2019;25:911–9.
135. Yépez VA, Mertes C, Mueller MF, Andrade DS, Wachutka L, Frésard L, et al. Detection of aberrant events in RNA sequencing data [Internet]. Available from: <https://doi.org/10.21203/rs.2.19080/v1>
136. Oliver GR, Tang X, Schultz-Rogers LE, Vidal-Folch N, Jenkinson WG, Schwab TL, et al. A tailored approach to fusion transcript identification increases diagnosis of rare inherited disease. *PLoS One*. 2019;14:e0223337.
137. Li D, Tian L, Hakonarson H. Increasing diagnostic yield by RNA sequencing in rare disease-bypass hurdles of interpreting intronic or splice-altering variants. *Ann Transl Med*. 2018. p. 126.
138. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* [Internet]. 2017;9. Available from: <https://doi.org/10.1126/scitranslmed.aal5209>
139. Wang ET, Sandberg R, Luo S, Khrebttukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456:470–6.
140. GTEx Consortium. Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. *Nature*. 2017;550:204–13.
141. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *Am J Hum Genet*. 2019;104:1007.
142. Lee H, Huang AY, Wang L-K, Yoon AJ, Renteria G, Eskin A, et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet Med*. 2020;22:490–9.
143. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol*. 2016;17:41.
144. EventPilot Web [Internet]. [cited 2021 Dec 14]. Available from: <https://eventpilotadmin.com/web/page.php?page=Session&project=ASHG21&id=P1342>
145. Bonder MJ, Smail C, Gloudemans MJ, Frésard L, Jakubosky D, D'Antonio M, et al. Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nat Genet*. 2021;53:313–21.
146. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, et al. Corrigendum: Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*. 2017;546:686.
147. Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, vSchuldt BM, et al. iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types [Internet]. *Stem Cell Reports*. 2017. p. 1086–100. Available from: <https://doi.org/10.1016/j.stemcr.2017.03.012>
148. Pashos EE, Park Y, Wang X, Raghavan A, Yang W, Abbey D, et al. Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci. *Cell Stem Cell*. 2017;20:558–70.e10.
149. Banovich NE, Li YI, Raj A, Ward MC, Greenside P, Calderon D, et al. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res*. 2018;28:122–31.
150. Carcamo-Orive I, Hoffman GE, Cundiff P, Beckmann ND, D'Souza SL, Knowles JW, et al. Analysis of transcriptional variability in a large human iPSC library reveals genetic and non-genetic determinants of heterogeneity. *Cell Stem Cell*. 2017;20:518–32.e9.
151. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501:506–11.
152. Brechtmann F, Mertes C, Matusевичiūtė A, Yépez VA, Avsec Ž, Herzog M, et al. OUTFRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *Am J Hum Genet*. 2018;103:907–17.
153. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*. 2012;7:500–7.
154. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*. 2003;72:291–336.

155. De Sandre-Giovannoli A, Lévy N. Altered splicing in prelamin A-associated premature aging phenotypes. *Prog Mol Subcell Biol*. 2006;44:199–232.
156. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet*. 2018;59:253–68.
157. Jenkinson G, Li Yi, Basu S, Cousin MA, Oliver GR, Klee EW. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics* [Internet]. 2020; Available from: <https://doi.org/10.1093/bioinformatics/btaa259>
158. Mertes C, Scheller I, Yépez VA, Çelik MH, Liang Y, Kremer LS, et al. Detection of aberrant splicing events in RNA-seq data with FRASER [Internet]. *bioRxiv*. 2019 [cited 2020 May 26]. p. 2019.12.18.866830. Available from: <https://www.biorxiv.org/content/10.1101/2019.12.18.866830v1.full>
159. Agamy O, Ben Zeev B, Lev D, Marcus B, Fine D, Su D, et al. Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. *Am J Hum Genet*. 2010;87:538–44.
160. Tung J, Akinyi MY, Mutura S, Altmann J, Wray GA, Alberts SC. Allele-specific gene expression in a wild nonhuman primate population. *Mol Ecol*. 2011;20:725–39.
161. Liu Z, Dong X, Li Y. A genome-wide study of allele-specific expression in colorectal cancer. *Front Genet*. 2018;9:570.
162. Ma C, Shao M, Kingsford C. SQUID: transcriptomic structural variation detection from RNA-seq. *Genome Biol*. 2018;19:52.
163. Qiu Y, Ma C, Xie H, Kingsford C. Detecting transcriptomic structural variants in heterogeneous contexts via the Multiple Compatible Arrangements Problem. *Algorithms Mol Biol*. 2020;15:9.
164. Dai X, Theobald R, Cheng H, Xing M, Zhang J. Fusion genes: A promising tool combating against cancer. *Biochim Biophys Acta Rev Cancer*. 1869;2018:149–60.
165. Nothwang HG, Kim HG, Aoki J, Geisterfer M, Kübart S, Wegner RD, et al. Functional hemizygosity of PAFAH1B3 due to a PAFAH1B3-CLK2 fusion gene in a female with mental retardation, ataxia and atrophy of the brain. *Hum Mol Genet*. 2001;10:797–806.
166. Ramocki MB, Dowling J, Grinberg I, Kimonis VE, Cardoso C, Gross A, et al. Reciprocal fusion transcripts of two novel Zn-finger genes in a female with absence of the corpus callosum, ocular colobomas and a balanced translocation between chromosomes 2p24 and 9q32. *Eur J Hum Genet*. 2003;11:527–34.
167. Yue Y, Grossmann B, Holder SE, Haaf T. De novo t(7;10)(q33;q23) translocation and closely juxtaposed microdeletion in a patient with macrocephaly and developmental delay. *Hum Genet*. 2005;117:1–8.
168. Hackmann K, Matko S, Gerlach E-M, von der Hagen M, Klink B, Schrock E, et al. Partial deletion of GILRB and GRIA2 in a patient with intellectual disability. *Eur J Hum Genet*. 2013;21:112–4.
169. Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, et al. The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am J Hum Genet*. 2014;95:143–61.
170. Bertelsen B, Melchior L, Jensen LR, Groth C, Nazaryan L, Debes NM, et al. A t(3;9)(q25.1;q34.3) translocation leading to OLFM1 fusion transcripts in Gilles de la Tourette syndrome, OCD and ADHD. *Psychiatry Res*. 2015;225:268–75.
171. Cmero M, Schmidt B, Majewski IJ, Ekert PG, Oshlack A, Davidson NM. MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data [Internet]. 2020 [cited 2020 Aug 27]. p. 2020.06.03.131532. Available from: <https://www.biorxiv.org/content/10.1101/2020.06.03.131532v1.abstract>
172. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol*. 2013;31:1009–14.
173. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells [Internet]. *Nature Communications*. 2017; Available from: <https://doi.org/10.1038/ncomms16027>.
174. Uapinyoying P, Goecks J, Knobloch SM, Panchapakesan K, Bonnemann CG, Partridge TA, et al. A long-read RNA-seq approach to identify novel transcripts of very large genes. *Genome Res*. 2020;30:885–97.
175. De Roeck A, Van den Bossche T, van der Zee J, Verheijen J, De Coster W, Van Dongen J, et al. Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. *Acta Neuropathol*. 2017;134:475–87.
176. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res*. 2018;28:1126–35.
177. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci U S A*. 2014;111:9869–74.
178. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods*. 2019;16:1297–305.
179. Dainis A, Tseng E, Clark TA, Hon T, Wheeler M, Ashley E. Targeted long-read RNA sequencing demonstrates transcriptional diversity driven by splice-site variation in MYBPC3. *Circ Genom Precis Med*. 2019;12:e002464.
180. Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature*. 2018;560:319–24.
181. Kuksin M, Morel D, Aglave M, Danlos F-X, Marabelle A, Zinoviyev A, et al. Applications of single-cell and bulk RNA sequencing in onco-immunology. *Eur J Cancer*. 2021;149:193–210.
182. Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. *Front Genet*. 2019;10:317.
183. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21:31.
184. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol*. 2021;22:102.
185. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mesdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*. 2020;11:5650.
186. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol Immunother*. 2018;67:1031–40.
187. Le T, Aronow RA, Kirshtein A, Shahriyari L. A review of digital cytometry methods: estimating the relative abundance of cell types in a bulk of cells. *Brief Bioinform* [Internet]. 2021;22. Available from: <https://doi.org/10.1093/bib/bbaa219>
188. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* [Internet]. 2017;356. Available from: <https://doi.org/10.1126/science.aah4573>
189. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. 2016;352:1586–90.
190. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*. 2018;174:1015–30.e16.
191. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst*. 2016;3:346–60.e4.
192. Tanaka N, Katayama S, Reddy A, Nishimura K, Niwa N, Hongo H, et al. Single-cell RNA-seq analysis reveals the platinum resistance gene COX7B and the surrogate marker CD63. *Cancer Med*. 2018;7:6193–204.
193. Ho Y-J, Anaparthi N, Molik D, Mathew G, Aicher T, Patel A, et al. Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome Res*. 2018;28:1353–63.
194. Grabowski P, Hesse S, Hollizeck S, Rohlfms M, Behrends U, Sherkat R, et al. Proteome analysis of human neutrophil granulocytes from patients with monogenic disease using data-independent acquisition. *Mol Cell Proteomics*. 2019;18:760–72.
195. Conboy E, Vairo F, Schultz M, Agre K, Ridsdale R, Deyle D, et al. Mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase deficiency: unique presenting laboratory values and a review of biochemical and clinical features. *JIMD Rep*. 2018;40:63–9.
196. Webb-Robertson B-JM, Stratton KG, Kyle JE, Kim Y-M, Bramer LM, Waters KM, et al. Statistically driven metabolite and lipid profiling of patients from the undiagnosed diseases network. *Anal Chem*. 2020;92:1796–803.
197. Aref-Eshghi E, Bend EG, Colaiacovo S, Caudle M, Chakrabarti R, Napier M, et al. Diagnostic utility of genome-wide DNA methylation testing in

- genetically unsolved individuals with suspected hereditary conditions. *Am J Hum Genet.* 2019;104:685–700.
198. Khanolkar A, Wilks JD, Jennings LJ, Davies JL, Zollett JA, Lott LL, et al. Signaling impairments in maternal T cells engrafted in an infant with a novel IL-2 receptor γ mutation. *J Allergy Clin Immunol.* 2015;135:1093–6.e8.
 199. Fernandez IZ, Baxter RM, Garcia-Perez JE, Vendrame E, Ranganath T, Kong DS, et al. A novel human IL2RB mutation results in T and NK cell-driven immune dysregulation. *J Exp Med.* 2019;216:1255–67.
 200. Khanolkar A, Kirschmann DA, Caparelli EA, Wilks JD, Cerullo JM, Bergerson JRE, et al. CD4 T cell-restricted IL-2 signaling defect in a patient with a novel IFNGR1 deficiency. *J Allergy Clin Immunol.* 2018;141:435–9.e7.
 201. Hansen RS, Wijmenga C, Luo P, Stanek AM, Canfield TK, Weemaes CM, et al. The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proc Natl Acad Sci U S A.* 1999;96:14412–7.
 202. Park E, Kim Y, Ryu H, Kowall NW, Lee J, Ryu H. Epigenetic mechanisms of Rubinstein–Taybi syndrome. *Neuromolecular Med.* Springer. 2014;16:16–24.
 203. Berdasco M, Esteller M. Genetic syndromes caused by mutations in epigenetic genes. *Hum Genet.* 2013;132:359–83.
 204. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat Rev Genet.* 2019;20:235–48.
 205. Falls JG, Pulford DJ, Wylie AA, Jirtle RL. Genomic imprinting: implications for human disease. *Am J Pathol.* 1999;154:635–47.
 206. Barbosa M, Joshi RS, Garg P, Martin-Trujillo A, Patel N, Jadhav B, et al. Identification of rare de novo epigenetic variations in congenital disorders. *Nat Commun.* 2018;9:2064.
 207. Aref-Eshghi E, Rodenhiser DJ, Schenkel LC, Lin H, Skinner C, Ainsworth P, et al. Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes. *Am J Hum Genet.* 2018;102:156–74.
 208. Schenkel LC, Aref-Eshghi E, Skinner C, Ainsworth P, Lin H, Paré G, et al. Peripheral blood epi-signature of Claes-Jensen syndrome enables sensitive and specific identification of patients and healthy carriers with pathogenic mutations in KDM5C. *Clin Epigenetics.* 2018;10:21.
 209. Sadikovic B, Levy MA, Aref-Eshghi E. Functional annotation of genomic variation: DNA methylation epigenatures in neurodevelopmental Mendelian disorders. *Hum Mol Genet.* 2020;29:R27–32.
 210. Sadikovic B, Levy MA, Kerkhof J, Aref-Eshghi E, Schenkel L, Stuart A, et al. Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. *Genet Med.* 2021;23:1065–74.
 211. Abela L, Simmons L, Steindl K, Schmitt B, Mastrangelo M, Joset P, et al. N(8)-acetylspermidine as a potential plasma biomarker for Snyder-Robinson syndrome identified by clinical metabolomics. *J Inherit Metab Dis.* 2016;39:131–7.
 212. Ait-El-Mkadem S, Dayem-Quere M, Gusic M, Chausseot A, Banwarth S, François B, et al. Mutations in MDH2, encoding a Krebs cycle enzyme, cause early-onset severe encephalopathy. *Am J Hum Genet.* 2017;100:151–9.
 213. Sirrs S, van Karnebeek CDM, Peng X, Shyr C, Tarailo-Graovac M, Mandal R, et al. Defects in fatty acid amide hydrolase 2 in a male with neurologic and psychiatric symptoms. *Orphanet J Rare Dis.* 2015;10:38.
 214. Solis BG, Van Den Rym A, Pérez-Caraballo JJ, Al-Ayoubi A, Lorenzo L, Cubillos-Zapata C, et al. Clinical and immunological features of human BCL10 deficiency [Internet]. Available from: <https://doi.org/10.21203/rs.3.rs-807424/v1>
 215. Cabral-Marques O, Schimke LF, de Oliveira EB Jr, El Khawanky N, Ramos RN, Al-Ramadi BK, et al. Flow cytometry contributions for the diagnosis and immunopathological characterization of primary immunodeficiency diseases with immune dysregulation. *Front Immunol.* 2019;10:2742.
 216. Crowther LM, Poms M, Plecko B. Multiomics tools for the diagnosis and treatment of rare neurological disease. *J Inherit Metab Dis.* 2018;41:425–34.
 217. Ulfenborg B. Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinformatics.* 2019;20:649.
 218. Fisch KM, Meißner T, Gioia L, Ducom J-C, Carland TM, Loguercio S, et al. Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics.* 2015;31:1724–8.
 219. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights.* 2020;14:1177932219899051.
 220. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics.* 2019;35:3055–62.
 221. Wangler MF, Yamamoto S, Chao H-T, Posey JE, Westerfield M, Postlethwait J, et al. Model organisms facilitate rare disease diagnosis and therapeutic research. *Genetics.* 2017;207:9–27.
 222. Harnish JM, Deal SL, Chao H-T, Wangler MF, Yamamoto S. In vivo functional study of disease-associated rare human variants using *Drosophila*. *J Vis Exp [Internet].* 2019; Available from: <https://doi.org/10.3791/59658>
 223. Hmeljak J, Justice MJ. From gene to treatment: supporting rare disease translational research through model systems. *Dis Model Mech [Internet].* 2019;12. Available from: <https://doi.org/10.1242/dmm.039271>
 224. Boycott KM, Campeau PM, Howley HE, Pavlidis P, Rogic S, Oriel C, et al. The Canadian Rare Diseases Models and Mechanisms (RDMM) Network: connecting understudied genes to model organisms. *Am J Hum Genet.* 2020;106:143–52.
 225. Frosk P, Arts HH, Philippe J, Gunn CS, Brown EL, Chodirker B, et al. A truncating mutation in CEP55 is the likely cause of MARCH, a novel syndrome affecting neuronal mitosis. *J Med Genet.* 2017;54:490–501.
 226. Oláhová M, Yoon WH, Thompson K, Jangam S, Fernandez L, Davidson JM, et al. Biallelic mutations in ATP5F1D, which encodes a subunit of ATP synthase, cause a metabolic disorder. *Am J Hum Genet.* 2018;102:494–504.
 227. Pena IA, Roussel Y, Daniel K, Mongeon K, Johnstone D, Weinschutz Mendes H, et al. Pyridoxine-dependent epilepsy in zebrafish caused by Aldh7a1 deficiency. *Genetics.* 2017;207:1501–18.
 228. Wen X-Y, Tarailo-Graovac M, Brand-Arzamendi K, Willems A, Rakic B, Huijben K, et al. Sialic acid catabolism by N-acetylneuraminidase pyruvate lyase is essential for muscle function. *JCI Insight [Internet].* 2018;3. Available from: <https://doi.org/10.1172/jci.insight.122373>
 229. van Karnebeek CDM, Bonafé L, Wen X-Y, Tarailo-Graovac M, Balzano S, Royer-Bertrand B, et al. NANS-mediated synthesis of sialic acid is required for brain and skeletal development. *Nat Genet.* 2016;48:777–84.
 230. Bashamboo A, Donohoue PA, Vilain E, Rojo S, Calvel P, Seneviratne SN, et al. A recurrent p.Arg92Trp variant in steroidogenic factor-1 (NR5A1) can act as a molecular switch in human sex development. *Hum Mol Genet.* 2016;25:5286.
 231. Kanca O, Andrews JC, Lee P-T, Patel C, Braddock SR, Slavotinek AM, et al. De novo variants in WDR37 are associated with epilepsy, Colobomas, Dysmorphism, developmental delay, intellectual disability, and cerebellar hypoplasia. *Am J Hum Genet.* 2019;105:672–4.
 232. Ferreira CR, Xia Z-J, Clément A, Parry DA, Davids M, Taylan F, et al. A recurrent de novo heterozygous COG4 substitution leads to Saul-Wilson Syndrome, disrupted vesicular trafficking, and altered proteoglycan glycosylation. *Am J Hum Genet.* 2018;103:553–67.
 233. Anderson RH, Francis KR. Modeling rare diseases with induced pluripotent stem cell technology. *Mol Cell Probes.* 2018;40:52–9.
 234. Li Y, Polak U, Clark AD, Bhalla AD, Chen Y-Y, Li J, et al. Establishment and maintenance of primary fibroblast repositories for rare diseases-Friedreich's Ataxia Example. *Biopreserv Biobank.* 2016;14:324–9.
 235. Sun N, Yazawa M, Liu J, Han L, Sanchez-Freire V, Abilez OJ, et al. Patient-specific induced pluripotent stem cells as a model for familial dilated cardiomyopathy. *Sci Transl Med.* 2012;4:130ra47.
 236. Dimos JT, Rodolfa KT, Niakan KK, Weisenthal LM, Mitsumoto H, Chung W, et al. Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science.* 2008;321:1218–21.
 237. Liu J, Verma PJ, Evans-Galea MV, Delatycki MB, Michalska A, Leung J, et al. Generation of induced pluripotent stem cell lines from Friedreich ataxia patients. *Stem Cell Rev Rep.* 2011;7:703–13.
 238. Carlessi L, Fusar Poli E, Bechi G, Mantegazza M, Pascucci B, Narciso L, et al. Functional and molecular defects of hiPSC-derived neurons from patients with ATM deficiency. *Cell Death Dis.* 2014;5:e1342.
 239. Malan D, Zhang M, Stallmeyer B, Müller J, Fleischmann BK, Schulze-Bahr E, et al. Human iPSC cell model of type 3 long QT syndrome recapitulates drug-based phenotype correction. *Basic Res Cardiol.* 2016;111:14.

240. Itier J-M, Ret G, Viale S, Sweet L, Bangari D, Caron A, et al. Effective clearance of GL-3 in a human iPSC-derived cardiomyocyte model of Fabry disease. *J Inherit Metab Dis*. 2014;37:1013–22.
241. Bellin M, Greber B. Human iPSC cell models of Jervell and Lange-Nielsen syndrome. *Rare Dis*. 2015;3:e1012978.
242. Zhang M, D'Aniello C, Verkerk AO, Wrobel E, Frank S, Oostwaard DW, et al. Recessive cardiac phenotypes in induced pluripotent stem cell models of Jervell and Lange-Nielsen syndrome: disease mechanisms and pharmacological rescue [Internet]. Proceedings of the National Academy of Sciences. 2014. p. E5383–92. Available from: <https://doi.org/10.1073/pnas.1419553111>
243. Raya A, Rodríguez-Pizà I, Guenechea G, Vassena R, Navarro S, Barrero MJ, et al. Disease-corrected haematopoietic progenitors from Fanconi anaemia induced pluripotent stem cells. *Nature*. 2009;460:53–9.
244. Hu L, Du L, Zhao Y, Li W, Ouyang Q, Zhou D, et al. Modeling Glanzmann thrombasthenia using patient specific iPSCs and restoring platelet aggregation function by CD41 overexpression. *Stem Cell Res*. 2017;20:14–20.
245. Cai J, Orlova VV, Cai X, Eekhoff EMW, Zhang K, Pei D, et al. Induced pluripotent stem cells to model human Fibrodysplasia Ossificans Progressiva. *Stem Cell Reports*. 2015;5:963–70.
246. Lukovic D, Artero Castro A, Delgado ABG, Bernal M de LAM, Luna Pelaez N, Díez Lloret A, et al. Human iPSC derived disease model of MERTK-associated retinitis pigmentosa. *Sci Rep*. 2015;5:12910.
247. Ramsden CM, Nommiste B, Lane AR, Carr A-JF, Powner MB, Smart MJK, et al. Rescue of the MERTK phagocytic defect in a human iPSC disease model using translational read-through inducing drugs [Internet]. Scientific Reports. 2017. Available from: <https://doi.org/10.1038/s41598-017-00142-7>
248. Yamashita A, Morioka M, Kishi H, Kimura T, Yahara Y, Okada M, et al. Statin treatment rescues FGFR3 skeletal dysplasia phenotypes. *Nature*. 2014;513:507–11.
249. Enikanolaiye A, Justice MJ. Model systems inform rare disease diagnosis, therapeutic discovery and pre-clinical efficacy. *Emerg Top Life Sci*. 2019;3:1–10.
250. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, et al. Variant interpretation: functional assays to the rescue. *Am J Hum Genet*. 2017;101:315–25.
251. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol*. 2019;20:223.
252. Hartin SN, Means JC, Alaimo JT, Younger ST. Expediting rare disease diagnosis: a call to bridge the gap between clinical and functional genomics. *Mol Med*. 2020;26:117.
253. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen T, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*. 2014;343:84–7.
254. Poirier JT. CRISPR Libraries and Screening. *Prog Mol Biol Transl Sci*. 2017;152:69–82.
255. Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat Commun*. 2018;9:5416.
256. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, et al. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol*. 2016;34:192–8.
257. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*. 2019;176:377–90.e19.
258. Breslow DK, Hoogendoorn S, Kopp AR, Morgens DW, Vu BK, Kennedy MC, et al. A CRISPR-based screen for Hedgehog signaling provides insights into ciliary function and ciliopathies. *Nat Genet*. 2018;50:460–71.
259. Rao S, Yao Y, Soares de Brito J, Yao Q, Shen AH, Watkinson RE, et al. Dissecting ELANE neutropenia pathogenicity by human HSC gene editing. *Cell Stem Cell*. 2021;28:833–45.e5.
260. Lek A, Zhang Y, Woodman KG, Huang S, DeSimone AM, Cohen J, et al. Applying genome-wide CRISPR-Cas9 screens for therapeutic discovery in facioscapulohumeral muscular dystrophy. *Sci Transl Med* [Internet]. 2020;12. Available from: <https://doi.org/10.1126/scitranslmed.aay0271>
261. Robinson PN, Köhler S, Oellrich A. Sanger Mouse Genetics Project, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014;24:340–8.
262. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*. 2009;7:e1000247.
263. Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF, et al. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet*. 2012;21:R37–44.
264. Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, et al. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum Mutat*. 2015;36:931–40.
265. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36:915–21.
266. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet*. 2009;84:524–33.
267. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*. 2015;36:928–30.
268. Matchmaker Exchange Statistics and Publications [Internet]. [cited 2021 Mar 29]. Available from: <https://www.matchmakerexchange.org/statistics.html>
269. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med*. 2019;25:60–4.
270. Colbaugh R, Glass K, Rudolf C, Tremblay Volv Global Lausanne Switzerland M. Learning to identify rare disease patients from electronic health records. *AMIA Annu Symp Proc*. 2018;2018:340–7.
271. Cohen AM, Chamberlin S, Deloughery T, Nguyen M, Bedrick S, Meninger S, et al. Detecting rare diseases in electronic health records using machine learning and knowledge engineering: case study of acute hepatic porphyria. *PLoS One*. 2020;15:e0235574.
272. Banda JM, Sarraju A, Abbasi F, Parizo J, Pariani M, Ison H, et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *NPJ Digit Med*. 2019;2:23.
273. Bruel A-L, Nambot S, Quéré V, Vitobello A, Thevenon J, Assoum M, et al. Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. *Eur J Hum Genet*. 2019;27:1519–31.
274. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet Med*. 2017;19:209–14.
275. Eldomery MK, Coban-Akdemir Z, Harel T, Rosenfeld JA, Gambin T, Stray-Pedersen A, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med*. 2017;9:26.
276. Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, et al. Reanalysis of clinical exome sequencing data. *N Engl J Med*. 2019;380:2478–80.
277. Wright CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med*. 2018;20:1216–23.
278. Ji J, Leung ML, Baker S, Deignan JL, Santani A. Clinical exome reanalysis: current practice and beyond. *Mol Diagn Ther*. 2021;25:529–36.
279. Fung JLF, Yu MHC, Huang S, Chung CCY, Chan MCY, Pajusalu S, et al. A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *NPJ Genom Med*. 2020;5:37.
280. James KN, Clark MM, Camp B, Kint C, Schols P, Batalov S, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *NPJ Genom Med*. 2020;5:33.
281. Costain G, Jobling R, Walker S, Reuter MS, Snell M, Bowdin S, et al. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. *Eur J Hum Genet*. 2018;26:740–4.
282. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gouridine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47:D1018–27.

283. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015;10:2004–15.
284. Birgmeier J, Haeussler M, Deisseroth CA, Jagadeesh KA, Ratner AJ, Guturu H, et al. AMELIE accelerates Mendelian patient diagnosis directly from the primary literature [Internet]. *bioRxiv*. 2017 [cited 2019 Aug 24]. p. 171322. Available from: <https://www.biorxiv.org/content/10.1101/171322v1>
285. Li Q, Zhao K, Bustamante CD, Ma X, Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet Med* [Internet]. 2019; Available from: <https://doi.org/10.1038/s41436-019-0439-8>
286. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43:D789–98.
287. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat*. 2012;33:803–8.
288. Cipriani V, Pontikos N, Arno G, Sergouniotis PI, Lenassi E, Thawong P, et al. An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data. *Genes* [Internet]. 2020;11. Available from: <https://doi.org/10.3390/genes11040460>
289. Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med*. 2019;21:1585–93.
290. Clinithink | OUR TECHNOLOGY [Internet]. [cited 2019 Aug 24]. Available from: <https://clinithink.com/our-technology/>
291. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med* [Internet]. 2019;11. Available from: <https://doi.org/10.1126/scitranslmed.aat6177>
292. Owen MJ, Niemi A-K, Dimmock DP, Speziale M, Nespeca M, Chau KK, et al. Rapid sequencing-based diagnosis of thiamine metabolism dysfunction syndrome. *N Engl J Med*. 2021;384:2159–61.
293. Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al. Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. *N Engl J Med* [Internet]. 2022; Available from: <https://doi.org/10.1056/NEJMc2112090>
294. Kessler MD, Yerges-Armstrong L, Taub MA, Shetty AC, Maloney K, Jeng LJB, et al. Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. *Nat Commun*. 2016;7:12521.
295. Consortium T 1000 GP, The 1000 Genomes Project Consortium. A global reference for human genetic variation [Internet]. *Nature*. 2015. p. 68–74. Available from: <https://doi.org/10.1038/nature15393>
296. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337:64–9.
297. Exome EVSNGO. Sequencing Project (ESP) Seattle. WA (URL: <http://evs.gs.washington.edu/EVS/>) [22/12/14 accessed]. 2016;
298. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285–91.
299. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol*. 2016;17:157.
300. Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, et al. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI Exome Sequencing Project. *Am J Hum Genet*. 2016;99:791–801.
301. Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorf LA, et al. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Hum Mutat*. 2018;39:1713–20.
302. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434–43.
303. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 2019;576:106–11.
304. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The “All of Us” Research Program. *N Engl J Med*. 2019;381:668–76.
305. Reuter CM, Kohler JN, Bonner D, Zastrow D, Fernandez L, Dries A, et al. Yield of whole exome sequencing in undiagnosed patients facing insurance coverage barriers to genetic testing. *J Genet Couns*. 2019;28:1107–18.
306. Dimmock D, Caylor S, Waldman B, Benson W, Ashburner C, Carmichael JL, et al. Project Baby Bear: Rapid precision care incorporating rWGS in 5 California children's hospitals demonstrates improved clinical outcomes and reduced costs of care. *Am J Hum Genet*. 2021;108:1231–8.
307. Tisdale A, Cuttillo CM, Nathan R, Russo P, Laraway B, Haendel M, et al. The IDeaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. *Orphanet J Rare Dis*. 2021;16:429.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

