Mini-Review

The additional diagnostic yield of long-read sequencing in undiagnosed rare diseases

Giulia F. Del Gobbo[1] and Kym M. Boycott[1,2]

[1] *Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Ontario, Canada K1H 5B2;* [2] *Department of Genetics, Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada K1H 8L1*

Long-read sequencing (LRS) is a promising technology positioned to study the significant proportion of rare diseases (RDs) that remain undiagnosed as it addresses many of the limitations of short-read sequencing, detecting and clarifying additional disease-associated variants that may be missed by the current standard diagnostic workflow for RDs. Some key areas where additional diagnostic yields may be realized include: (1) detection and resolution of structural variants (SVs); (2) detection and characterization of tandem repeat expansions; (3) coverage of regions of high sequence similarity; (4) variant phasing; (5) the use of de novo genome assemblies for reference-based or graph genome variant detection; and (6) epigenetic and transcriptomic evaluations. Examples from over 50 studies support that the main areas of added diagnostic yield currently lie in SV detection and characterization, repeat expansion assessment, and phasing (with or without DNA methylation information). Several emerging studies applying LRS in cohorts of undiagnosed RDs also demonstrate that LRS can boost diagnostic yields following negative standard-of-care clinical testing and provide an added yield of 7%–17% following negative short-read genome sequencing. With this evidence of improved diagnostic yield, we discuss the incorporation of LRS into the diagnostic care pathway for undiagnosed RDs, including current challenges and considerations, with the ultimate goal of ending the diagnostic odyssey for countless individuals with RDs.

Rare diseases (RDs) encompass a diverse group of disorders that are individually rare in the population yet represent a significant burden to global health. RDs are conditions that affect fewer than 1 in 2500 individuals (Ferreira 2019), however, with several thousand recognized RDs, they collectively affect up to 1.5%–6.2% of the population globally (Ferreira 2019; Nguengang Wakap et al. 2020). About 70% of RDs are childhood-onset (Nguengang Wakap et al. 2020) and up to 65% are associated with a reduced life span, with about a quarter being potentially life-limiting by 5 years of age (Ferreira 2019). It is estimated that ∼70% of RDs have an underlying genetic etiology (Nguengang Wakap et al. 2020). Currently, the Online Mendelian Inheritance in Man database reports over 6400 phenotypes for which a molecular basis is known and over 4500 genes with phenotype-causing variants (https://www.omim.org/statistics/geneMap), with more disease-gene associations continually discovered. With such large numbers, rarity, and heterogeneity between and within RDs, molecular tools for diagnoses have been of utmost importance. Identifying the molecular cause of RDs provides affected individuals and their families with improved access to support services and potential treatments, information on prognosis and management of the disorder, options for further testing for family planning, and ends often long diagnostic odysseys. Despite the clear importance of reaching a diagnosis, more than half of RDs may remain undiagnosed following standard-of-care clinical genetic testing (Shashi et al. 2014).

Current genetic diagnostic workflows for RDs incorporate the patient's clinical presentation and the suspected mechanism of their rare genetic disease and may use various techniques to achieve a molecular diagnosis (Conlin et al. 2022; Kernohan and Boycott 2024). Over the last decade, short-read sequencing (SRS) of fragments of DNA 50–300 bp has been increasingly used in clinical settings for RD diagnosis, providing sequencing of targeted regions, the protein-coding exome (SR-ES), or nearly the entire genome (SR-GS). SRS enables a high-throughput method to accurately assess sequence variants (single nucleotide variants [SNVs] or insertions/deletions <50 bp), with the added ability to detect certain copy number variants and some structural variants (SVs, genomic alterations >50 bp in size). Since its more widespread adoption, genome-wide SRS, primarily SR-ES, has emerged as an effective first-tier test for indications such as neurodevelopmental disorders (NDDs) or multiple congenital anomalies (Srivastava et al. 2019; Manickam et al. 2021), with a diagnostic yield of ∼30%–35% depending on the indication and previous testing history (Clark et al. 2018; Splinter et al. 2018; Shickh et al. 2021; Chung et al. 2023; Hartley et al. 2024). While this has been a transformative technology, this still leaves nearly two-thirds of patients undiagnosed following SR-ES. Expanding beyond the protein-coding portion of the genome using SR-GS improves the assessment of noncoding sequence variants and detection of copy number variants and SVs; however, to date, the evidence for incremental diagnostic yield over SR-ES has been limited and may only be up to 10% (Ewans et al. 2022).

Several factors may contribute to the high rate of undiagnosed RDs following genome-wide SRS, including interpretation challenges for variants of uncertain significance (VUSs), variants residing in novel disease genes, and complex genetic and/or environmental causes of disease. Additionally, inherent limitations to SRS technology may be a significant contributor. Given the nature of the short reads, SRS struggles in alignment of nonunique sequences such as regions of the genome that are highly repetitive

Corresponding author: kboycott@cheo.on.ca

Article published online before print. Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279970.124. Freely available online through the *Genome Research* Open Access option.

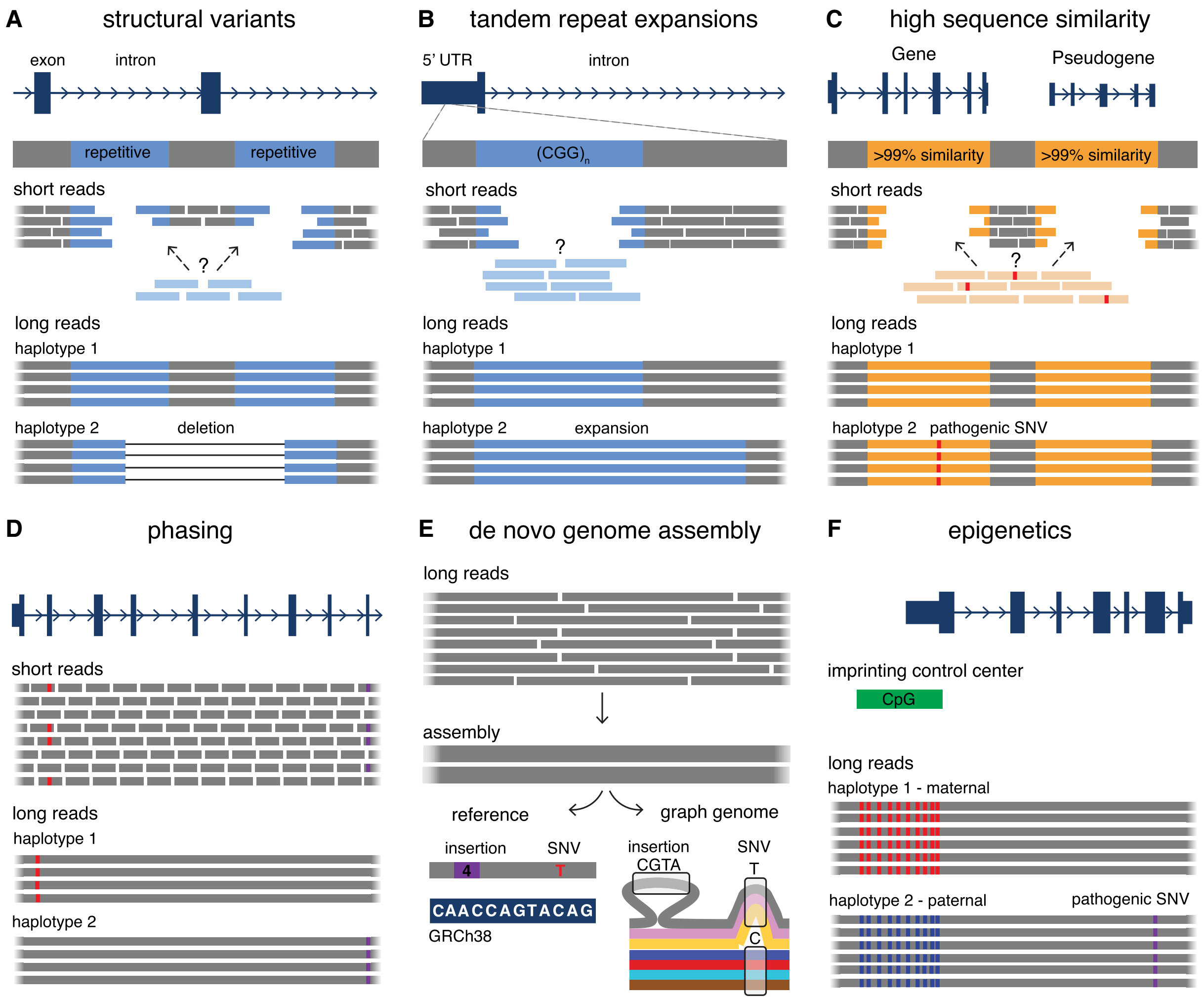

Figure 1. Summary of the utility of LRS over SRS in undiagnosed RDs. (*A*) LRS has an improved ability to detect SVs compared to SRS, especially in challenging regions for SRS such as repetitive DNA (blue), which often mediates the formation of SVs. (*B*) LRS enables improved sequencing and alignment of short tandem repeat sequences (blue) compared to SRS, enabling the accurate detection of tandem repeat expansions. Examples of genes with disease-associated short tandem repeat expansions include *FMR1* (Fragile X syndrome), *HTT* (Huntington disease), and several genes associated with cerebellar ataxias (*ATXN3*, *FGF14*, etc.). (*C*) LRS allows for improved mapping and coverage of regions of high sequence similarity in the genome (orange), which are challenging for SRS. This enables the differentiation of sequences between genes and their pseudogenes, and therefore detection of variation in these challenging genes. Examples of disease-associated genes in regions of high sequence similarity include *PKD1* (polycystic kidney disease), *IKBKG* (X-linked immunodeficiencies), and *SMN1* (spinal muscular atrophy). (*D*) LRS enables haplotype phasing over long ranges, which is helpful to confirm compound heterozygosity of variants (red, purple) without the requirement of parental samples for segregation, or in scenarios where one or more variants are de novo. (*E*) Long reads derived from LRS can be used to build high quality and highly contiguous de novo genome assemblies without requiring alignment to a reference genome. These assemblies can either be compared to a linear reference genome (*bottom left*) to detect variants (a 4 bp insertion compared to the reference, purple box; and a C > T SNV, red) or they can be compared to de novo assemblies derived from other individuals (*bottom right*, colored lines indicate assemblies from different individuals) for the generation of graph-genomes that describe genetic variation among individuals (4 bp insertion seen only in one individual; T/C SNV that is more common in the group). (*F*) Sequencing of native DNA strands in LRS enables concomitant assessment of base modifications, such as differentiating between methylated cytosines (red) and unmethylated cytosines (blue) in cytosine-guanine dinucleotides (CpGs). This can be used in combination with phasing information to investigate imprinted loci that have parent-of-origin-specific DNA methylation patterns. Examples of disease-associated imprinted genes include *H19*/*IGF2* (Silver–Russell syndrome), *UBE3A* (Angelman syndrome), and *PLAGL1* (transient neonatal diabetes mellitus).

or have high sequence similarity, and in the detection and characterization of SVs (Fig. 1A–C). The short reads are also difficult to use to piece together haplotypes or generate de novo assemblies, which make read-based variant phasing a challenge and limit the reconstruction of complex genomic rearrangements (CGRs) or use of reference-free methods for variant discovery (Fig. 1D,E). Additionally, the polymerase chain reaction (PCR) amplification step in SR-ES can introduce biases and have trouble amplifying regions that are GC-rich. SRS also does not allow for concurrent detection of modifications to the native DNA strand, such as methylation (Fig. 1F). Overall, these limitations may leave many disease-causing variants undiscoverable or uninterpretable in individuals with undiagnosed RDs following SRS. To address many of these limitations, long-read sequencing (LRS) technologies were developed, enabling the genome-wide sequencing of native DNA fragments at multiple orders of magnitude larger than those in SR-GS, over 10 kb and up to megabases in size. In this mini-review, we review the existing evidence for LRS to increase diagnostic yields in undiagnosed RDs by highlighting variant types and situations in which LRS may be a particularly useful approach, emerging evidence from the application of LRS in cohorts of undiagnosed RDs, and considerations for its incorporation into clinical diagnostic pathways for RDs.

Long-read sequencing technologies

LRS technologies can include either "true" LRS or synthetic LRS. In true LRS technologies, long fragments of nucleic acid are directly sequenced. In synthetic methods, such as synthetic long reads (Peters et al. 2012; Li et al. 2015; Bankevich and Pevzner 2016) or linked-reads (Zheng et al. 2016; Marks et al. 2019; Wang et al. 2019; Chen et al. 2020), subfragments of long molecules of nucleic acids are co-barcoded to "link" or tag fragments from the same original long molecule. These short fragments are sequenced by SRS and then synthetically reconstructed into the original long fragment by bioinformatic methods. Synthetic LRS provides improvements over SRS in haplotype phasing, de novo genome assembly, and SV discovery (Bankevich and Pevzner 2016; Zheng et al. 2016; Elyanow et al. 2018; Chaisson et al. 2019; Marks et al. 2019; Wang et al. 2019). However, because the base sequencing unit is still a short read, some limitations of SRS may remain (e.g., poor coverage in low-complexity regions, poor SV reconstruction, GC-biases), and they have historically been inferior to true LRS (Chaisson et al. 2019; Ebbert et al. 2019). As a result, this has likely impacted their more limited application in the field of RDs compared to true LRS. Further development of synthetic LRS technologies is ongoing, with the recently released complete long-read sequencing (CLR) method by Illumina showing particular promise (Gorzynski et al. 2024); however, these are not the primary focus of this mini-review.

In terms of true LRS technologies, two platforms have dominated the market since their introduction: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio high-fidelity (HiFi) sequencing uses a sequencing-by-synthesis method. Double-stranded, high-molecular-weight DNA from a size-selected library of ∼15 kb is first circularized by ligating adaptors to the end of the fragments, then each circular DNA molecule undergoes multiple rounds of sequencing by a polymerase incorporating fluorescently labeled nucleotides to allow for real-time sequence determination. The output of these multiple sequencing passes (subreads) are merged to generate a consensus sequence, a HiFi read, with a very high per-base accuracy and typically between 10 and 30 kb in size (Wenger et al. 2019; Vollger et al. 2020). ONT uses nanopores embedded in electro-resistant membranes through which single-stranded DNA molecules are fed using a motor protein. The disruption of the electric current as bases pass through the nanopore is read in real-time and translated into base sequences using base-calling algorithms. ONT can theoretically be used to sequence fragments of DNA of any size depending on the sample input and library preparation; however, they are typically >10 kb. Library preparation kits for generating ultra-long reads >50 kb are available (https://store.nanoporetech.com/ultra-long-dna-sequencing-kit-v14.html), and studies have demonstrated abilities to generate long reads up to megabases in size (Logsdon et al. 2020). Although initial iterations of both technologies had lower per-base accuracies than common Illumina SRS platforms, these have improved substantially over the years to now be highly competitive with SRS (Logsdon et al. 2020; Damaraju et al. 2024; Kosugi and Terao 2024; Mahmoud et al. 2024). At ∼30× coverage, accuracies (F1 scores) can be up to 98.5%–99.9% for SNVs and 84.9%–99.4% for indels by PacBio HiFi LRS and up to 98.1%–99.7% for SNVs and 69.7%–84.1% for indels by ONT LRS, depending on the chemistry, variant caller, or benchmark sample used (Pei et al. 2021; Harvey et al. 2023; Kolesnikov et al. 2024). Performance for ONT has further improved recently with duplex sequencing using the latest R10.4 chemistry (Kolesnikov et al. 2024). Additional detailed descriptions and comparisons between the two technologies have been summarized elsewhere (Logsdon et al. 2020; Harvey et al. 2023; Mastrorosa et al. 2023; Oehler et al. 2023; van Dijk et al. 2023) and are beyond the scope of this mini-review.

Additional diagnostic yield provided by long-read sequencing

Detection and resolution of structural variants

SVs, genomic alterations >50 bp in size including insertions, deletions, duplications, inversions, translocations, and CGRs, account for the largest amount of sequence variation between individual genomes (Sudmant et al. 2015b; Chaisson et al. 2019) and are thus crucial to comprehensively assess in RD diagnosis. It has been long recognized that SVs contribute to genetic disorders (Stankiewicz and Lupski 2010); however, no single previous clinical genomic testing technology has been able to accurately assess SVs across the full spectrum of sizes and genomic context (Conlin et al. 2022; Kernohan and Boycott 2024). Despite the many available tools for SV detection from SR-GS (Kosugi et al. 2019), with significantly longer reads to better span SVs and resolve repetitive and nonunique regions that are enriched for SVs (Sudmant et al. 2015a), LRS consistently outperforms SR-GS in the detection of SVs in a wide range of sizes. In head-to-head comparisons, LRS can detect at least three to five times more SVs than SRS (Huddleston et al. 2017; Chaisson et al. 2019; Ebert et al. 2021). So far, evidence for gains in diagnostic yields by detection of SVs previously missed by SR-GS is primarily in CGRs or SVs involving challenging regions/sequences for SRS. For example, a CGR involving several breakpoints in Chromosomes 7 and 9 with an insertional translocation, inversion, and deletion was identified by PacBio LRS in a proband with a complex NDD that was previously missed by SR-GS analysis (Hiatt et al. 2021). Additionally, examples of insertions involving transposable elements, highly repetitive DNA sequences that can insert themselves into new places in the genome and potentially disrupt coding sequences, splicing, or gene regulation, have been reported. These include SINE-VNTR-*Alu* (SVA) element insertions in introns of disease genes *SMARCB1* and *NR5A1* (Sabatella et al. 2021; Del Gobbo et al. 2024) and a LINE-1-mediated insertion that resulted in a single-exon duplication of *CDKL5* (Hiatt et al. 2021). There is also ample evidence of LRS successfully identifying disease-causing SVs such as deletions, insertions, inversions, or more complex rearrangements that were missed by previous standard-of-care genetic testing in undiagnosed RD patients (Mizuguchi et al. 2019, 2021; Xie et al. 2020; de la Morena-Barrio et al. 2022; Daida et al. 2023; Damián et al. 2023; Yanagi et al. 2023). As SR-GS was not performed in these examples, the added benefit of LRS over SR-GS was not demonstrated; however, they confirm the utility of LRS as a tool to detect disease-associated SVs following typical diagnostic testing.

Not only can LRS identify SVs that were missed by previous genomic testing in patients with undiagnosed RDs, but it can also clarify known SVs by fine-mapping breakpoints and/or revealing additional SV complexity. In eight individuals with previously identified CGRs, Miller et al. (2021) demonstrated that all rearrangements could be identified by targeted ONT LRS, and additional information about the CGR was gained for all individuals, including the precise resolution of breakpoints, determination of orientation of alterations, or identification of additional events in the CGRs that were not previously detected. This has the potential to boost diagnoses, as highlighted in an affected individual with a previously nondiagnostic balanced translocation t(8;18)(q22;q21) where ONT LRS revealed a chromothripsis-like CGR at the translocation site involving 19 rearranged fragments, including a deletion impacting disease-associated genes *RAD21* and *EXT1* that explained the patient's presentation (Lei et al. 2020). Additionally, in a study assessing SR-GS for the diagnosis of NDDs, LRS was necessary to fully resolve complex SVs in four patients, supporting its specific added yield over SRS (Sanchis-Juan et al. 2023). Additional examples of LRS revealing greater complexity of SVs and/or fine-mapping breakpoints to clarify and confirm the pathogenicity of variants highlight the utility of this added information to upgrade previously nondiagnostic variants and provide answers to individuals with undiagnosed RDs (Dutta et al. 2019; Schieffer et al. 2021; Sund et al. 2024). This may be particularly useful in reconstructing and mapping breakpoints of CGRs that occur in challenging regions such as segmental duplications or repetitive elements that commonly mediate complex SV formation (Schuy et al. 2022). This was recently demonstrated in a study by Grochowski et al. (2024), in which LRS aided to resolve and fine-map breakpoints of complex duplication–triplication/inverted-duplications associated with *MECP2* duplication syndrome that are mediated by nearby inverted low-copy repeats.

Detection and characterization of tandem repeat expansions and contractions

Tandem repeat expansions or contractions are another key class of SV for which LRS can aid in the diagnosis and discovery of genetic causes of undiagnosed RDs. Tandem repeats are regions in the genome in which sequences of DNA are repeated in numerous copies next to one another, often distinguished as short tandem repeats (STRs), repeated motifs of 1–6 bp, or variable number tandem repeats (VNTRs) with repeated motifs of >7 bp. There are an estimated over 1.7 million tandem repeat loci in the human genome, together accounting for ∼8% of our genome (English et al. 2024). Tandem repeats are highly mutable and prone to expansion or contraction due to DNA replication errors. STR expansions in particular contribute to numerous RDs, especially adult-onset neurological disorders including several forms of spinocerebellar ataxia, Huntington disease, and amyotrophic lateral sclerosis (Chintalaphani et al. 2021; Depienne and Mandel 2021). Given their highly repetitive nature and sizes that are often orders of magnitude larger than can be captured in a single SRS read, STR expansions have been notoriously challenging to assess by SRS technologies. In clinical settings, these require targeted approaches including Southern blot or repeat-primed PCR for molecular diagnosis. LRS, therefore, holds promise for the accurate sizing and characterization of sequence composition and epigenetic modifications of STR expansions on a genome-wide scale.

Studies benchmarking LRS in patients with a variety of known pathogenic repeat expansions have consistently demonstrated that LRS can effectively recapitulate molecular diagnoses in a single comprehensive assay by identifying pathogenic expansions and providing information on the sequence composition and single base-resolved DNA methylation at STRs (Höijer et al. 2018; Giesselmann et al. 2019; Miyatake et al. 2022; Stevanovski et al. 2022; Erdmann et al. 2023; Dolzhenko et al. 2024). These provide important proof-of-principle for the ability of LRS to detect pathogenic expansions; however, few studies to date have systematically assessed the concordance of sizes of expansions (especially large expansions) detected by LRS compared to standard technologies. Stevanovski et al. (2022) reported an $R^2 = 0.996$ for lengths of normal and expanded repeats at the *HTT* locus, 0.993 for *FMR1*, and 0.946 for *RFC1* using targeted ONT compared to repeat-primed PCR or Southern blot. A similar 100% concordance between repeat sizes from targeted PacBio LRS and PCR fragment analysis at the *HTT* locus was also observed in 11 patients with Huntington disease (Höijer et al. 2018). Using Cas9-targeted ONT at 10 STR loci associated with ataxia, Erdmann et al. (2023) reported that the majority of loci agreed with PCR fragment analysis within ±3 repeat units for unexpanded or short expansion (<100 bp) alleles in their method validation cohort, within the expected error rate for PCR fragment analysis. In 28 expansion-positive individuals from their cohort of patients with adult-onset ataxia, LRS-predicted sizes were within ±4 repeats of estimates from PCR or were within the range of sizes determined by PCR (Erdmann et al. 2023). In addition to sizing, the accurate assessment of the sequence composition of STRs by LRS can also be helpful. Sequence interruptions or noncanonical repeat motifs may alter the pathogenicity, severity, age of onset, or stability of an expanded repeat, and are, therefore, important for informing the diagnosis and prognosis of RDs caused by STR expansions (Rajan-Babu et al. 2024). This has been demonstrated in sequencing $(CTG)_n$ repeat expansions in *DMPK* associated with myotonic dystrophy type 1, where PacBio LRS enabled accurate detection of sizes, sequence interruptions, and somatic mosaicism in individuals with >1000 repeats, and the finding that CCG interruptions near the 3′ end of the STR are associated with increased somatic stability of the repeat and milder phenotypes (Cumming et al. 2018; Mangin et al. 2021). Additional studies of larger cohorts of patients with various known expansions are necessary to evaluate the accuracy of LRS in comparison to current diagnostic standards and to facilitate its incorporation as a comprehensive molecular tool for STR expansion disorders.

LRS has also been instrumental in the discovery of novel STR expansion disorders in undiagnosed RDs, contributing to the increase in discoveries over the past several years. Some early examples include the discovery of an intronic $(TTTCA)_n$ or $(TTTTA)_n$ expansion in *SAMD12* associated with autosomal dominant (AD) benign familial adult myoclonic epilepsy (Ishiura et al. 2018; Zeng et al. 2019) and a $(GGC)_n$ expansion in the 5′ UTR of *NOTCH2NLC* associated with neuronal intranuclear inclusion disease (Sone et al. 2019; Tian et al. 2019). Since then, LRS has aided in the discovery or further characterization of several novel STR disorders, reviewed elsewhere (Chintalaphani et al. 2021; Depienne and Mandel 2021; Gall-Duncan et al. 2022). This includes recent discoveries of novel expansions in *ZFX3* and *THAP11* associated with AD forms of spinocerebellar ataxia (Tan et al. 2023; Chen et al. 2024c), and a 27 bp duplication in a polyalanine tract in *HOXD13* associated with synpolydactyly 1 (Melas et al. 2022). Additionally, LRS was also necessary to fully characterize the size and sequence composition of the recently discovered intronic $(GAA)_n$ expansion in *FGF14* associated with AD late-onset spinocerebellar ataxia type 27B (Pellerin et al. 2023; Rafehi et al. 2023). This STR expansion has since proven to account for a considerable proportion of undiagnosed patients with ataxia, particularly in individuals of European descent and especially in French

Canadians (Hengel et al. 2023; Novis et al. 2023; Pellerin et al. 2023; Rafehi et al. 2023; Méreaux et al. 2024). The ability to accurately sequence the expanded repeats >1000 bp supported that only pure $(GAA)_n$ expansions are associated with disease, as large expansions in unaffected individuals were found to contain different GA-rich motifs than the expanded $(GAA)_n$ motif observed in all affected individuals (Pellerin et al. 2023). Given the limited availability of STR genotype data from population cohorts sequenced by LRS and tools for analysis, the discovery of novel STR expansions from genome-wide LRS is challenging. As such, novel STR expansion discoveries have been powered by large pedigrees, often guided by linkage studies identifying candidate regions for targeted sequencing or analysis of LRS data. With recent improvements in tools for STR genotyping and discovery and a growing number of population cohorts, genome-wide analyses will soon be more feasible and will help to power additional discoveries of novel STR expansions underlying undiagnosed RDs.

Variant discovery in regions of high sequence similarity

Genomic regions with high sequence similarity are another problem area for SRS technologies as the difficulty to uniquely map short reads between two or more highly similar regions contributes to regions with poor or no sequencing coverage in SRS. This is relevant for undiagnosed RDs because many disease-associated genes reside in such regions, having either one or more pseudogenes (genomic regions with high sequence similarity to known genes that do not generate functional protein products), high sequence similarity to other functional genes, or multiple regions within the gene itself that are highly similar (Mandelker et al. 2016). Examples include *PKD1*, *HYDIN*, *IKBKG*, and *SMN1*, all of which have historically required multiple targeted molecular technologies to comprehensively assess disease-associated variation. In a study published in 2019, Ebbert et al. identified 36,794 regions in gene bodies, including 2855 in coding sequences, that they termed as SR-GS "dark regions": regions with either no/low sequencing coverage or low mapping quality of sequencing reads because of difficulties in adequately assembling or aligning SR-GS reads. Using earlier iterations of LRS technologies, they demonstrated that PacBio and ONT LRS significantly improved coverage in 88% and 95% of the regions in coding sequences, respectively (Ebbert et al. 2019). Furthermore, Wenger et al. (2019) found that 152/193 (79%) of medically relevant genes with at least one exon in an SRS dark region were fully mappable using PacBio CCS. Recently, Sanford Kobayashi et al. (2022) found that PacBio HiFi LRS successfully covered 98% of annotated SRS dark regions genome-wide in their cohort of 30 participants. This facilitated the identification of a pathogenic variant in *IKBKG* in a participant with a previously undiagnosed immunological disorder, demonstrating the successful use of LRS to uncover new diagnoses in these challenging regions for SRS (Sanford Kobayashi et al. 2022). Additional studies of specific genes also support this utility. Borràs et al. (2017) found that targeted PacBio LRS of *PKD1* and *PKD2* in a cohort of patients with AD polycystic kidney disease identified all previously known pathogenic variants with high sensitivity and specificity, and also identified additional variants that were missed by previous testing. Additionally, Fleming et al. (2024) demonstrated that ONT LRS with a modified SRS bioinformatic pipeline aided in differentiating variation in *HYDIN* from its pseudogene *HYDIN2* and supported disease-associated variant discovery in a cohort of patients with primary ciliary dyskinesia. Recently, Chen et al. (2023) developed a tool, Paraphase, that accurately differentiates full-length haplotypes in *SMN1* and its paralog *SMN2* to facilitate variant discovery and diagnosis from PacBio LRS data. This tool can now be applied to 160 long (>10 kb) segmental duplication regions with >99% sequence similarity, encompassing 316 genes, including 11 medically relevant genes (Chen et al. 2024b). With the human genome having over 6000 genes with SRS dark regions (Ebbert et al. 2019), it is likely that yet undiscovered pathogenic variation in these regions may underlie undiagnosed RDs, and LRS holds promise to bring them to light.

Improved read-based phasing

Variant phasing, in which variants are assigned to either the maternal or paternal chromosome, is often an important step in RD diagnostics. Particularly for individuals with compound heterozygous variants in genes associated with recessive disease, determining whether the variants are in *cis* (on the same parental chromosome) or in *trans* (on different parental chromosomes) is crucial for variant interpretation. Phasing directly from SRS reads requires variants to be near enough to one another to either both be captured within a single short-read or paired-end read or have nearby heterozygous variants that can act as proxies. Alternatively, phase can be determined for inherited variants when genotypes from both parents or other informative family members are available, by statistical methods using genotypes from large population data sets to infer phase (typically requiring genome-sequencing data), or by more laborious methodologies that physically separate chromosomes or selectively amplify one allele before sequencing. Statistical phasing methods can be applied to SRS data with good success; however, these methods typically have increased error rates for rare variants that are not commonly feasible in clinical RD diagnostics. Laboratories primarily rely on familial genotyping or direct read-based phasing if possible.

With significantly longer read lengths, LRS outperforms SRS in read-based phasing, increasing phase block N50s (largest haplotype block length such that 50% of all heterozygous sites are contained in haplotype blocks of equal or greater size) by at least 10-fold from ∼1 kb to over 100 kb for PacBio and up to megabases in size for ONT (Choi et al. 2018; Chaisson et al. 2019; Majidian and Sedlazeck 2020). LRS read-based phasing has further improved with the latest sequencing platforms and phasing tools. Using haplotype-aware variant calling with PEPPER-Margin-DeepVariant, highly accurate haplotype blocks with N50s of 0.24 Mb from 35× PacBio HiFi or 2–6 Mb from 25–75× ONT LRS have been achieved (Shafin et al. 2021). This method enabled up to 66% or 93% of annotated genes to be fully captured within a haplotype block using 35× PacBio HiFi or 75× ONT LRS, respectively (Shafin et al. 2021). Incorporating other variant types such as SVs or STRs in addition to small variants further improves phasing, as demonstrated by a new tool HiPhase which generated haplotype block N50s of 0.48 Mb and fully phased 88% of annotated genes from PacBio HiFi data (Holt et al. 2024). When informative family members are unavailable for testing, one or more variants are de novo, or distances between variants are too great than can be phased by standard clinical methodologies, LRS, therefore, has the potential to provide a resolution for RD diagnostics. Several examples of LRS successfully phasing compound heterozygous variants in genes associated with autosomal recessive (AR) disease have been reported. For example, a LINE-1 insertion in exon 7 was confirmed in *trans* with a maternally inherited coding variant in exon

36 of *CC2D2A* in two siblings with a clinical diagnosis of Joubert syndrome, even when paternal DNA was unavailable (Yanagi et al. 2023). The readily available phasing information from LRS also improves the analysis and interpretation of singleton data, especially when a candidate gene or region may be targeted for sequencing or for analysis. This was demonstrated by Miller et al. (2022), who used targeted ONT LRS to identify missing second variants in *trans* in 8 of 9 individuals with a clinical diagnosis of AR Werner syndrome.

De novo genome assembly and pangenome approaches for variant discovery

Another compelling benefit of LRS over SRS is the improved feasibility of generating de novo genome assemblies and using pangenome approaches for variant discovery. Standard reference-based variant detection methods rely on the alignment of sequencing reads to the reference genome and then using bioinformatic tools to identify variants compared to the reference. However, until the most recent CHM13-T2T reference genome (Nurk et al. 2022), previous reference genomes remained incomplete, leaving regions of unknown sequence (gaps) distributed throughout the genome which hinder read alignment and variant discovery (Schneider et al. 2017; Nurk et al. 2022). Additionally, these references are representative consensus calls from only a small number of human genomes, which can introduce biases in the calling of nonreference sequences (Miga and Wang 2021). LRS allows the generation of de novo genome assemblies, where individual genomes are assembled into long contiguous haplotype-resolved sequences directly from the long reads instead of first aligning these to a reference genome. Variant calling from LRS assemblies against the reference genome can improve the precision and recall of SVs and indels compared to standard reference-based variant detection (Ebert et al. 2021; Harvey et al. 2023). Alternatively, these assemblies can be further leveraged to discover variants without relying on the reference at all by comparing variation among multiple assembled genomes as a pangenome graph. This can enhance variant discovery, particularly for SVs (Liao et al. 2023). This approach has not yet been thoroughly explored in the RD space; however, a recent study demonstrated that the generation of a graph pangenome of 574 assemblies from a pediatric RD cohort and 94 control assemblies improved the reproducibility of SVs compared to standard reference-based approaches (Groza et al. 2024). It also improved the prioritization of rare, potentially disease-associated SVs, leading to the discovery of a novel diagnostic SV in *KMT2E* in a patient with a previously undiagnosed RD (Groza et al. 2024). Several limitations to pangenome approaches such as high computational burden and the need for additional tools tailored for variant discovery mean that the current utility of these methods in RD diagnostics is primarily in improved variant calling accuracy and genotyping, especially for complex SVs or variants in complex loci (Taylor et al. 2024). With the increasing use of LRS and continued developments in pangenome references and analytic tools (Liao et al. 2023; Taylor et al. 2024), this emerging field may prove useful to improve yields in undiagnosed RDs in the future.

Utility beyond the DNA sequence

Because LRS technologies directly sequence unamplified DNA, modifications to DNA bases are preserved and can be detected by both ONT and PacBio LRS either based on unique disruptions to the electrical current or alterations in polymerase kinetics, respectively (Flusberg et al. 2010; Rand et al. 2017). Thus, LRS enables the assessment of epigenetic modifications at a base pair resolution. This has been useful in delineating pathogenic mechanisms of newly discovered RDs such as demonstrating 5-methylcytosine (5mC) hypermethylation of expanded $(CGG)_n$ alleles in *NOTCH2NLC* associated with neuronal intranuclear inclusion disorder (Ishiura et al. 2019). It also has significant clinical utility by supporting molecular diagnoses for RDs in which 5mC is altered. This includes imprinting disorders, where disrupted parent-of-origin-specific DNA methylation patterns at imprinted loci may be detected with the use of haplotype-phased reads and 5mC information (Cheung et al. 2023; Yamada et al. 2023; Bækgaard et al. 2024), or STR expansion disorders such as Fragile X syndrome, Friedreich's ataxia, or myotonic dystrophy type 1 in which pathogenic expanded STRs are hypermethylated (Giesselmann et al. 2019; Stevanovski et al. 2022; Cheung et al. 2023; Erdmann et al. 2023; Dolzhenko et al. 2024). Methods and tools are also being developed for the assessment of genome-wide DNA methylation outliers to aid in improving yields in undiagnosed RDs. In a recent proof-of-principle study, Cheung et al. (2023) analyzed rare 5mC hypermethylation events in PacBio LRS data from a cohort of 276 individuals from 152 families with undiagnosed pediatric RDs and identified hypermethylation associated with a repeat expansion in *DIP2B* in a patient with global developmental delay. This added information on epigenetic modifications provided along with DNA sequence in LRS, therefore, has the potential to aid in identifying novel candidates and diagnoses in undiagnosed RDs.

Beyond DNA sequencing, the benefits of longer reads in LRS also extend to RNA sequencing (RNA-seq). RNA-seq can increase diagnostic yields in RDs through the assessment of gene expression alterations, splicing changes, and allele-specific expression related to disease-associated DNA variants (Cummings et al. 2017; Kremer et al. 2017; Frésard et al. 2019; Gonorazky et al. 2019; Lee et al. 2020; Murdock et al. 2021; Yépez et al. 2022). However, previous studies have relied on SRS, which has limited ability to reconstruct full-length mRNA transcripts, making it challenging to interpret the exact impact of splice-altering variants or fully resolve novel gene fusions. Long-read RNA-seq in contrast, enables full-length isoform sequencing and quantification. This has powered the discovery of thousands of novel transcript isoforms in human tissues (Glinos et al. 2022). It has also helped clarify the impact of splice-altering VUSs in undiagnosed RDs, such as a homozygous intronic c.600-31T>C in *MFN2* that created five novel isoforms that all disrupted the reading frame and resulted in nonsense-mediated mRNA decay (Stergachis et al. 2023) and an intronic c.1079-23T>A in *CLPB* that created a novel isoform with a new splice site causing the insertion of 7 amino acids in the conserved P-loop of *CLPB* (Farrow et al. 2023). With the increasing interest in using RNA-seq as a second-tier test to clarify transcriptomic alterations of noncoding SNVs and SVs, it will be important to consider the added strengths of long-read RNA-seq to further boost diagnoses in the future.

Emerging evidence of diagnostic yields in cohorts of undiagnosed RDs

Much of the existing literature applying LRS in undiagnosed RDs to improve diagnostic yields is from individual case reports or proof-of-concept studies sequencing known positive controls, making it difficult to accurately assess the increased diagnostic yield provided by LRS. Increasingly, however, studies are emerging applying LRS to cohorts of individuals with undiagnosed RDs, providing some of the first demonstrations of the true potential added yield of LRS. The exact indications and previous genetic testing in each study are variable and therefore diagnostic yields vary accordingly. Several cohort studies have applied LRS directly following standard clinical testing using SRS gene panels or SR-ES. This includes a cohort of 34 families with various RDs of suspected AR inheritance that remained undiagnosed following SR-ES in which analysis of regions of homozygosity in PacBio LRS data identified diagnostic variants in 13 families, 8 of which (23.5% of total) were not detectable by SR-ES (AlAbdi et al. 2023). Additional yields associated with LRS reported in more specific RD cohorts undiagnosed following SRS panels or SR-ES are highly variable. These include 18% in a cohort of 11 families with antithrombin deficiency with previous negative analyses of *SERPINC1* (de la Morena-Barrio et al. 2022), 44% in a cohort of nine families with muscular dystrophy (Bruels et al. 2022), 50% in a cohort of 26 patients with clinical diagnoses of tuberous sclerosis complex (Duan et al. 2024), and up to 100% in a small cohort of five families with undiagnosed hereditary spastic paraplegia (Fukuda et al. 2023). Finally, Miller et al. (2021) demonstrated a 60% yield of targeted ONT LRS at candidate genes to identify a missing pathogenic or likely pathogenic variant in a cohort of 10 individuals with various undiagnosed RDs for which either a single pathogenic variant in an AR disease gene, or no pathogenic variant for a specific suspected AD or X-linked disorder had been identified by previous clinical testing. Unfortunately, these studies do not allow us to assess the yield of LRS over a more comparable SRS technology, SR-GS. Indeed, some of the identified diagnostic variants could have been detected by SR-GS such as some larger deletions and splice-altering intronic SNVs. However, these studies nonetheless support an increased diagnostic yield of LRS following typical standard-of-care testing and may help in the future when determining where to place LRS in the diagnostic care pathway.

Only a few reported studies have used LRS in cohorts of undiagnosed RDs following negative SR-GS. These studies are crucial to determine the additional yield of LRS over currently available SRS technologies. Some of the earliest reports are from small cohorts (<10) of patients with undiagnosed NDDs. While Pauper et al. (2021) did not identify any disease-associated candidates in five probands with undiagnosed NDDs presenting with intellectual disability and other features following trio PacBio LRS, Hiatt et al. (2021) identified a likely diagnostic variant in 2 of 6 (33%) probands also using trio PacBio LRS, including a de novo CGR and a de novo LINE-1-mediated insertion, both impacting known disease genes. Notably, both studies only focused on assessing de novo variants, which commonly underlie NDDs (Sebat et al. 2007; Vissers et al. 2010). Further larger studies in heterogeneous cohorts of undiagnosed RDs have also supported incremental yields over SR-GS. In a cohort of 30 patients from 26 families with undiagnosed pediatric RDs, Sanford Kobayashi et al. (2022) found a modest increased yield by PacBio LRS using a singleton approach with one additional diagnosis that was missed by SR-GS, a known likely pathogenic hemizygous stop-loss variant in *IKBKG*. Another larger effort from the Genomic Answers for Kids project sequenced 256 affected participants with diverse pediatric disorders, many of which remained undiagnosed following SR-GS (Cohen et al. 2022). Although specific diagnostic yields were not reported, at least five examples of diagnoses made by LRS were provided as proof-of-principle for the technology, including previously unidentified pathogenic repeat expansions, a CGR, and the use of phasing of compound heterozygous variants to support diagnoses (Cohen et al. 2022).

Driven by increased throughput, decreased costs, and improved analysis capabilities of LRS, recent comprehensive genome-wide analyses of LRS data in larger cohorts of undiagnosed RDs have begun to shed more light on the increased diagnostic yields LRS may provide over SR-GS. Two studies of cohorts of ~100 patients with various undiagnosed RDs suggest a specific increased yield of LRS over SR-GS between 7% and 17% (Hiatt et al. 2024; Steyaert et al. 2024). Building on their 2021 study (Hiatt et al. 2021), Hiatt et al. (2024) applied PacBio LRS in 96 probands with undiagnosed RDs presenting with NDD, multiple congenital anomalies, or a suspected congenital myopathy. They found new disease-relevant or potentially disease-associated variants in 16 probands, noting that 7 of these (7.3% of total) were exclusively identifiable by LRS (Hiatt et al. 2024). Additionally, in a cohort of 232 individuals from 93 families with undiagnosed RDs presenting with neurological, neuromuscular, or epilepsy phenotypes, Steyaert et al. (2024) identified 13 novel diagnoses (13%) and four additional compelling candidate disease-associated SVs (4.3%) using PacBio LRS. In addition to this undiagnosed cohort, they also studied a small cohort of 21 families with rare clinically recognizable unsolved syndromes (Aicardi, Hallermann–Streiff, Gomez–Lopez–Hernandez, Oculo-auriculo-vertebral spectrum disorders), but did not identify candidate genes or loci shared among affected individuals for any of these syndromes (Steyaert et al. 2024).

Together, evidence from these cohort studies is promising for a modest but significant increase in diagnostic yield from LRS over SRS technologies, ultimately providing diagnoses for families with RDs that would not have otherwise been possible. It is also worth noting that most of these studies have focused primarily on variants impacting known disease-associated loci, de novo variants, or large, exon-overlapping SVs. As analysis methods and control cohorts continue to grow and improve, these yields may further increase as the community is better able to harness the full potential of LRS in undiagnosed RDs.

Considerations for incorporating long-read sequencing in the RD diagnostic workflow

Although the surge in studies demonstrating the utility of LRS in undiagnosed RDs shows its promise to improve diagnostic yields, several factors from sample input through to analysis still limit the widespread incorporation of this new technology as part of the RD diagnostic workflow. Firstly, current library preparation procedures require relatively large amounts of high-quality, high-molecular-weight DNA to achieve long-read lengths for genome-wide LRS. This necessitates starting biological materials and DNA extraction protocols that preserve the integrity of these large DNA molecules and may be a limitation when minimal amounts of samples are available or when collection of blood or other invasive samples may not be possible. Secondly, the higher cost and lower throughput compared to SRS have been limitations to its widespread adoption. Although recent developments such as the PacBio Revio and ONT PromethION systems have made significant improvements, bringing material costs per 30× genome down to ~$720–$1000 USD and sequencing up to 1300 (Revio) to nearly 5000 (PromethION) genomes per year (https://www.pacb.com/revio/; https://nanoporetech.com/products/sequence/promethion), this is still not nearly comparable to the latest developments in SRS technology such as Illumina's NovaSeq X Series which proposes to be able to generate more than 20,000 genomes per year at as low as $200 USD per sample in material costs (https://www.illumina.com/systems/sequencing-platforms/novaseq-x-plus/applications/transition.html). It is worth noting, however, that many additional factors that contribute to the cost of a clinical test (e.g., sample collection and handling, laboratory overhead costs, analysis and reporting) may be more similar between the two methods; therefore, this cost differential between SRS and LRS testing is likely less significant overall than suggested simply by the cost of sequencing. Thirdly, the historically higher costs and lower throughput of LRS also contributed to a limited availability of control data sets for variant allele frequency annotation. Given the greater sensitivity for the detection of SVs and coverage of challenging SRS regions by LRS, databases of allele frequencies derived from population cohorts also sequenced by LRS are essential for rare variant analyses in RDs. Unfortunately, the number of publicly available LRS genomes currently pales in comparison to those from SRS data sets such as gnomAD (Chen et al. 2024a). Promisingly, several recent efforts have begun to make headway. This includes CoLoRSdb (https://colorsdb.org/), a single resource of compiled data from over 1400 PacBio LRS genomes from several cohorts, including the Human Pangenome Reference Consortium (Liao et al. 2023), the Human Genome Structural Variant Consortium (Ebert et al. 2021), and RD cohorts such as the Genomic Answers for Kids project (Cohen et al. 2022). Additionally, a data set of 1019 samples from the 1000 Genomes Project sequenced at intermediate coverage (16.9×) by ONT LRS suitable for SV analysis (Schloissnig et al. 2024) and the first 100 samples sequenced at a minimum 30× depth from the 1000 Genomes Project ONT Sequencing Consortium were recently released (Gustafson et al. 2024). As more human genomes are sequenced by LRS, these allele frequency databases will continue to grow and improve. Indeed, the *All of Us* initiative recently performed a feasibility study, establishing a method for LRS using PacBio at scale for accurate small variant and SV discovery at the

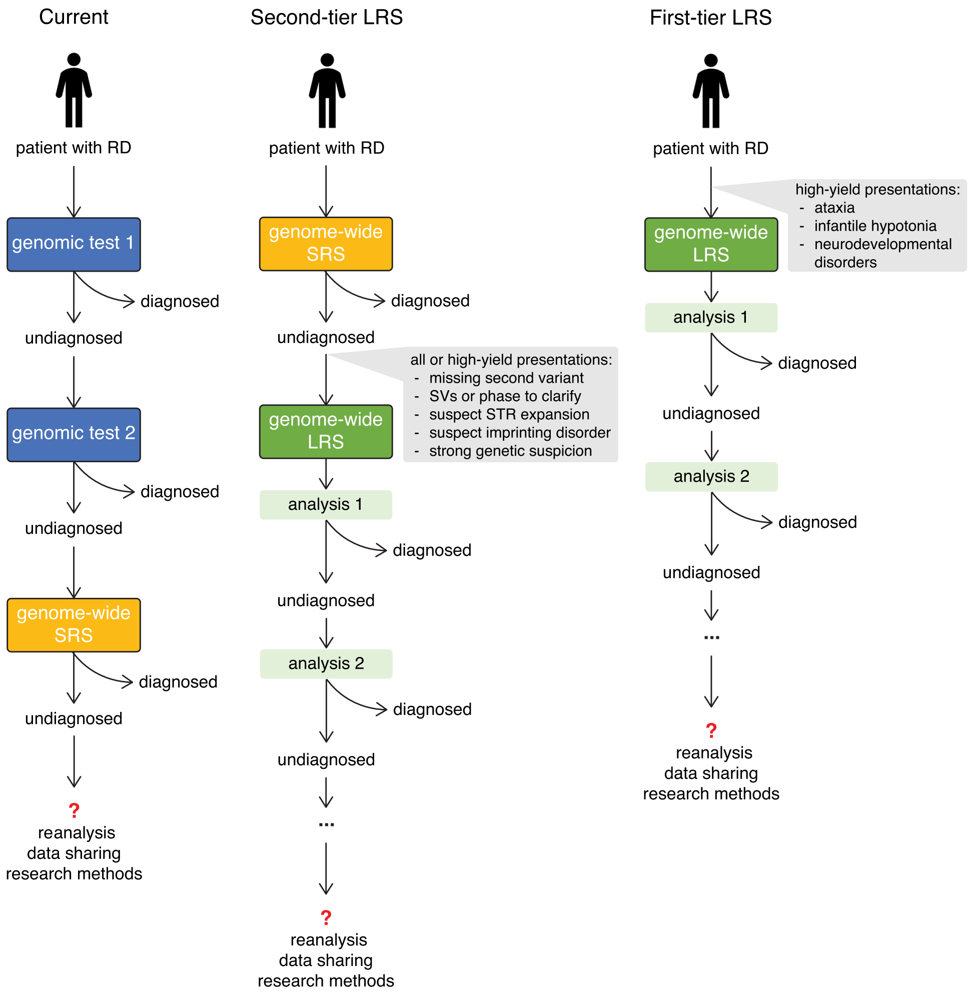

Figure 2. Hypothetical incorporation of LRS into the undiagnosed RD care pathway. (*Left*) The current care pathway in which patients with RDs may undergo long diagnostic odysseys, receiving numerous different consecutive genomic tests depending on indications, including genome-wide SRS. (*Middle*) Proposed incorporation of LRS as a second-tier test following nondiagnostic genome-wide SRS testing. Studies support an increased diagnostic yield when incorporating LRS following genome-wide SRS; therefore, this pathway has the potential to reduce the number of patients with undiagnosed RDs. (*Right*) Proposal for incorporating LRS as a first-tier test in the future, which would be primarily useful for RDs in which LRS first may be most cost-effective. This reduces the step-wise diagnostic pathway but allows for consecutive analyses of LRS data (e.g., coding variation, repeat expansions, SVs, methylation).

population-level in preparation for its planned population-scale LRS effort (Mahmoud et al. 2024). Data from the first 1000 participants are now available by registered access through their Researcher Workbench (https://www.researchallofus.org/). Finally, a fourth key factor still limiting the incorporation of LRS into RD diagnostic workflows centers around data analysis and infrastructure. While many tools for LRS data analysis exist to support initial base calling, alignment, assembly, phasing, variant calling, and more (Amarasinghe et al. 2020), best practices and standards for bioinformatic and analytical pipelines have yet to be established. These are necessary to ensure consistency of results among and within clinical laboratories. There is also a need to consider the increased infrastructure required for the computation and storage of this genome-wide data. All these limitations are active areas of development and have improved significantly even in the past few years, making incorporating LRS into clinical workflows more and more feasible soon.

As we consider incorporating LRS into clinical testing for RDs, we face a major consideration of where to place it in the RD diagnostic workflow. Until many of the aforementioned limitations have been addressed, the current utility of LRS clinically lies in its application as a second-tier test following nondiagnostic clinical genome-wide SRS (Fig. 2). Focusing LRS in undiagnosed RDs in situations where LRS is especially powerful over SRS may provide the highest yields. For example, in patients with missing second mutations for AR conditions, to clarify and fine-map SVs identified by karyotyping or microarray, when phasing would clarify molecular diagnoses, in undiagnosed RDs suspected to be caused by tandem repeat expansions (e.g., neurodegenerative conditions, AD inheritance, demonstrate anticipation), suspected imprinting disorders, or other RDs for which family history or phenotypic presentation supports a high likelihood of an underlying monogenic disorder (Fig. 2). However, given that LRS can identify nearly the full spectrum of variant types and can assess epigenetic modifications that historically have all required multiple different and often consecutive molecular diagnostic methods, it is also promising to implement as a first-tier test (Fig. 2). Using LRS near the beginning of the RD diagnostic workflow would streamline clinical genetic testing and improve access to comprehensive genetic testing for all individuals with RDs, ultimately reducing the diagnostic odyssey by identifying molecular diagnoses faster and for more individuals than current clinical standards (Conlin et al. 2022; Damaraju et al. 2024). When we also consider the collective costs and burden of multiple different sample collection and handling procedures, specific training required, analyses and reporting, and other overhead costs associated with each individual test in the current standard of cascade testing, using LRS as a single first-tier test could even prove more efficient and cost-effective in certain cases (Damaraju et al. 2024). This has not yet been thoroughly demonstrated, therefore, studies evaluating diagnostic yields, clinical utility, and economic analyses of LRS as a first-tier test compared to standard clinical testing workflows will be crucial. As a starting point, considering LRS as a first-tier test for scenarios where rapid and thorough diagnostics are needed or for presentations in which numerous step-wise and laborious clinical tests could be replaced by a single LRS test would aid in demonstrating this utility (Fig. 2). For example, LRS as a first-tier test in critically ill patients for rapid diagnostics, or as a single test to thoroughly assess all known disease-causing repeat expansions in individuals presenting with ataxia. Additionally, first-tier LRS is compelling for infants presenting with hypotonia, where current standards may include many different tests including karyotyping, microarray, and targeted assessments of several genes depending on clinical suspicion, including challenging genes *SMN1* and *SMN2* for spinal muscular atrophy, STR expansions in *DMPK* for myotonic dystrophy, and/or methylation and copy number testing at 15q11.2 for Prader–Willi syndrome (Fig. 2; Sharma et al. 2021). So far, studies have demonstrated that LRS is feasible as a first-tier test for comprehensive assessment of STRs associated with ataxia (Stevanovski et al. 2022) and to provide ultrarapid diagnoses in a pediatric critical care setting (Gorzynski et al. 2022; Zalusky et al. 2024). As LRS costs continue to decrease, throughput increases, and analyses mature, we anticipate a more rapid incorporation of this technology into diagnostic care pathways.

Conclusions and future prospects

A key approach to alleviating some of the global burden of RDs is to provide accurate diagnoses. These direct care, management, counseling, and access to resources for families, in addition to improving our understanding of the causes of RDs to improve targeted therapies and disease management. With about two-thirds of RDs remaining undiagnosed following genome-wide SRS, the need for additional strategies to find answers for these remaining families is of high importance. LRS is a key technology to incorporate into the undiagnosed RD care pathway to tackle this challenge. LRS has demonstrated the capability to identify challenging variants for SRS, resolve known VUSs, and provide additional supportive information such as epigenetic alterations to boost diagnostic yields in undiagnosed RDs. LRS is also an attractive technology to provide streamlined comprehensive genomic testing in the future that could improve timeliness to diagnoses and access to comprehensive genomic testing for more individuals with undiagnosed RDs. Additional studies of LRS that clearly define diagnostic utility will further solidify the need for this technology, and studies of outcomes and costs in comparison to standard-of-care testing and/or genome-wide SRS will aid in navigating where LRS fits best in the undiagnosed RD care pathway. Further developments to improve throughput, reduce costs, increase available control/population cohort data for allele frequency annotation, and standardize analysis methods will support the incorporation of this technology in clinical diagnostic laboratories. Given the significant advancements that have been made in the field of LRS to date, we believe the time for more widespread use of LRS to tackle the remaining undiagnosed RDs is here, and we are optimistic that this technology will help take the field closer to the ultimate goal of accurate and timely diagnoses for all RDs.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was performed under the Care4Rare Canada Consortium funded by Genome Canada and the Ontario Genomics Institute (OGI-147), the Canadian Institutes of Health Research, Ontario Research Fund, Genome Alberta, Genome British Columbia, Génome Québec, and Children's Hospital of Eastern Ontario Foundation. G.F.D.G. is supported by a CIHR Fellowship award (MFE-491710) and K.M.B. is supported by a CIHR Foundation Grant (FDN-154279) and a Tier 1 Canada Research Chair in Rare Disease Precision Health.

References

AlAbdi L, Shamseldin HE, Khouj E, Helaby R, Aljamal B, Alqahtani M, Almulhim A, Hamid H, Hashem MO, Abdulwahab F, et al. 2023. Beyond the exome: utility of long-read whole genome sequencing in exome-negative autosomal recessive diseases. *Genome Med* **15:** 114. doi:10.1186/s13073-023-01270-8

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21:** 30. doi:10.1186/s13059-020-1935-5

Bækgaard CH, Lester EB, Møller-Larsen S, Lauridsen MF, Larsen MJ. 2024. Nanoimprint: a DNA methylation tool for clinical interpretation and diagnosis of common imprinting disorders using nanopore long-read sequencing. *Ann Hum Genet* **88:** 392–398. doi:10.1111/ahg.12556

Bankevich A, Pevzner PA. 2016. TruSPAdes: barcode assembly of TruSeq synthetic long reads. *Nat Methods* **13:** 248–250. doi:10.1038/nmeth.3737

Borràs DM, Vossen RHAM, Liem M, Buermans HPJ, Dauwerse H, van Heusden D, Gansevoort RT, den Dunnen JT, Janssen B, Peters DJM, et al. 2017. Detecting *PKD1* variants in polycystic kidney disease patients by single-molecule long-read sequencing. *Hum Mutat* **38:** 870–879. doi:10.1002/humu.23223

Bruels CC, Littel HR, Daugherty AL, Stafki S, Estrella EA, McGaughy ES, Truong D, Badalamenti JP, Pais L, Ganesh VS, et al. 2022. Diagnostic capabilities of nanopore long-read sequencing in muscular dystrophy. *Ann Clin Transl Neurol* **9:** 1302–1309. doi:10.1002/acn3.51612

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10:** 1784. doi:10.1038/s41467-018-08148-z

Chen Z, Pham L, Wu T-C, Mo G, Xia Y, Chang PL, Porter D, Phan T, Che H, Tran H, et al. 2020. 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* **30:** 898–909. doi:10.1101/gr.260380.119

Chen X, Harting J, Farrow E, Thiffault I, Kasperaviciute D, Genomics England Research Consortium, Hoischen A, Gilissen C, Pastinen T, Eberle MA. 2023. Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am J Hum Genet* **110:** 240–250. doi:10.1016/j.ajhg.2023.01.001

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, et al. 2024a. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625:** 92–100. doi:10.1038/s41586-023-06045-0

Chen X, Baker D, Dolzhenko E, Devaney JM, Noya J, Berlyoung AS, Brandon R, Hruska KS, Lochovsky L, Kruszka P, et al. 2024b. Genome-wide profiling of highly similar paralogous genes using HiFi sequencing. bioRxiv doi:10.1101/2024.04.19.590294

Chen Z, Gustavsson EK, Macpherson H, Anderson C, Clarkson C, Rocca C, Self E, Alvarez Jerez P, Scardamaglia A, Pellerin D, et al. 2024c. Adaptive long-read sequencing reveals GGC repeat expansion in *ZFX3* associated with spinocerebellar ataxia type 4. *Mov Disord* **39:** 486–497. doi:10.1002/mds.29704

Cheung WA, Johnson AF, Rowell WJ, Farrow E, Hall R, Cohen ASA, Means JC, Zion TN, Portik DM, Saunders CT, et al. 2023. Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort. *Nat Commun* **14:** 3090. doi:10.1038/s41467-023-38782-1

Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. 2021. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun* **9:** 98. doi:10.1186/s40478-021-01201-x

Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. 2018. Comparison of phasing strategies for whole human genomes. *PLoS Genet* **14:** e1007308. doi:10.1371/journal.pgen.1007308

Chung CCY, Hue SPY, Ng NYT, Doong PHL, Chu ATW, Chung BHY. 2023. Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations. *Genet Med* **25:** 100896. doi:10.1016/j.gim.2023.100896

Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, Kingsmore SF. 2018. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genomic Med* **3:** 16. doi:10.1038/s41525-018-0053-8

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B, et al. 2022. Genomic answers for children: dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* **24:** 1336–1348. doi:10.1016/j.gim.2022.02.007

Conlin LK, Aref-Eshghi E, McEldrew DA, Luo M, Rajagopalan R. 2022. Long-read sequencing for molecular diagnostics in constitutional genetic disorders. *Hum Mutat* **43:** 1531–1544. doi:10.1002/humu.24465

Cumming SA, Hamilton MJ, Robb Y, Gregory H, McWilliam C, Cooper A, Adam B, McGhie J, Hamilton G, Herzyk P, et al. 2018. De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. *Eur J Hum Genet* **26:** 1635–1647. doi:10.1038/s41431-018-0156-9

Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, Bolduc V, Waddell LB, Sandaradura SA, O'Grady GL, et al. 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* **9:** eaal5209. doi:10.1126/scitranslmed.aal5209

Daida K, Funayama M, Billingsley KJ, Malik L, Miano-Burkhardt A, Leonard HL, Makarious MB, Iwaki H, Ding J, Gibbs JR, et al. 2023. Long-read sequencing resolves a complex structural variant in *PRKN* Parkinson's disease. *Mov Disord* **38:** 2249–2257. doi:10.1002/mds.29610

Damaraju N, Miller AL, Miller DE. 2024. Long-read DNA and RNA sequencing to streamline clinical genetic testing and reduce barriers to comprehensive genetic testing. *J Appl Lab Med* **9:** 138–150. doi:10.1093/jalm/jfad107

Damián A, Núñez-Moreno G, Jubin C, Tamayo A, de Alba MR, Villaverde C, Fund C, Delépine M, Leduc A, Deleuze JF, et al. 2023. Long-read genome sequencing identifies cryptic structural variants in congenital aniridia cases. *Hum Genomics* **17:** 45. doi:10.1186/s40246-023-00490-8

de la Morena-Barrio B, Orlando C, Sanchis-Juan A, García JL, Padilla J, de la Morena-Barrio ME, Puruunen M, Stouffs K, Cifuentes R, Borràs N, et al. 2022. Molecular dissection of structural variations involved in antithrombin deficiency. *J Mol Diagn* **24:** 462–475. doi:10.1016/j.jmoldx.2022.01.009

Del Gobbo GF, Wang X, Couse M, Mackay L, Goldsmith C, Marshall AE, Liang Y, Lambert C, Zhang S, Dhillon H, et al. 2024. Long-read genome sequencing reveals a novel intronic retroelement insertion in *NR5A1* associated with 46, XY differences of sexual development. *Am J Med Genet A* **194:** e63522. doi:10.1002/ajmg.a.63522

Depienne C, Mandel J-L. 2021. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am J Hum Genet* **108:** 764–785. doi:10.1016/j.ajhg.2021.03.011

Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* **42:** 1606–1614. doi:10.1038/s41587-023-02057-3

Duan J, Pan S, Ye Y, Hu Z, Chen L, Liang D, Fu T, Zhan L, Li Z, Liao J, et al. 2024. Uncovering hidden genetic variations: long-read sequencing reveals new insights into tuberous sclerosis complex. *Front Cell Dev Biol* **12:** 1415258. doi:10.3389/fcell.2024.1415258

Dutta UR, Rao SN, Pidugu VK, Vineeth VS, Bhattacherjee A, Bhowmik AD, Ramaswamy SK, Singh KG, Dalal A. 2019. Breakpoint mapping of a novel de novo translocation t(X;20)(q11.1;p13) by positional cloning and long read sequencing. *Genomics* **111:** 1108–1114. doi:10.1016/j.ygeno.2018.07.005

Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, Kauwe JSK, Belzil V, Pregent L, Carrasquillo MM, et al. 2019. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol* **20:** 97. doi:10.1186/s13059-019-1707-2

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117

Elyanow R, Wu H-T, Raphael BJ. 2018. Identifying structural variants using linked-read sequencing data. *Bioinformatics* **34:** 353–360. doi:10.1093/bioinformatics/btx712

English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, Park J, Gu B, Wagner J, Eberle MA, et al. 2024. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol* doi:10.1038/s41587-024-02225-z

Erdmann H, Schöberl F, Giurgiu M, Leal Silva RM, Scholz V, Scharf F, Wendlandt M, Kleinle S, Deschauer M, Nübling G, et al. 2023. Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain* **146:** 1831–1843. doi:10.1093/brain/awac377

Ewans LJ, Minoche AE, Schofield D, Shrestha R, Puttick C, Zhu Y, Drew A, Gayevskiy V, Elakis G, Walsh C, et al. 2022. Whole exome and genome sequencing in Mendelian disorders: a diagnostic and health economic analysis. *Eur J Hum Genet* **30:** 1121–1131. doi:10.1038/s41431-022-01162-2

Farrow E, Jay A, Means J, Younger S, Biswell R, Koseva B, Thiffault I, Pastinen T, Pappas K, Toriello H. 2023. Case of *CLPB* deficiency solved by HiFi long read genome sequencing and RNAseq. *Am J Med Genet A* **191:** 2908–2912. doi:10.1002/ajmg.a.63365

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

Fleming A, Galey M, Briggs L, Edwards M, Hogg C, John S, Wilkinson S, Quinn E, Rai R, Burgoyne T, et al. 2024. Combined approaches, including long-read sequencing, address the diagnostic challenge of *HYDIN* in primary ciliary dyskinesia. *Eur J Hum Genet* **32:** 1074–1085. doi:10.1038/s41431-024-01599-7

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7:** 461–465. doi:10.1038/nmeth.1459

Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, Bonner D, Kernohan KD, Marwaha S, Zappala Z, et al. 2019. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25:** 911–919. doi:10.1038/s41591-019-0457-8

Fukuda H, Mizuguchi T, Doi H, Kameyama S, Kunii M, Joki H, Takahashi T, Komiya H, Sasaki M, Miyaji Y, et al. 2023. Long-read sequencing revealing intragenic deletions in exome-negative spastic paraplegias. *J Hum Genet* **68:** 689–697. doi:10.1038/s10038-023-01170-0

Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. 2022. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Res* **32:** 1–27. doi:10.1101/gr.269530.120

Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, Kretzmer H, Assum G, Galonska C, Siebert R, et al. 2019. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol* **37:** 1478–1481. doi:10.1038/s41587-019-0293-x

Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608:** 353–359. doi:10.1038/s41586-022-05035-y

Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, Kao D, Ohri K, Viththiyapaskaran S, Tarnopolsky MA, et al. 2019. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *Am J Hum Genet* **104:** 466–483. doi:10.1016/j.ajhg.2019.01.012

Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, Spiteri E, Pesout T, Monlong J, Baid G, et al. 2022. Ultrarapid nanopore genome sequencing in a critical care setting. *N Engl J Med* **386:** 700–702. doi:10.1056/NEJMc2112090

Gorzynski JE, Marwaha S, Reuter C, Jensen TD, Ferrasse A, Raja AN, Fernandez L, Kravets E, Carter J, Bonner D, et al. 2024. Clinical application of Complete Long Read genome sequencing identifies a 16kb intragenic duplication in EHMT1 in a patient with suspected Kleefstra syndrome. medRxiv doi:10.1101/2024.03.28.24304304

Grochowski CM, Bengtsson JD, Du H, Gandhi M, Lun MY, Mehaffey MG, Park K, Höps W, Benito E, Hasenfeld P, et al. 2024. Inverted triplications formed by iterative template switches generate structural variant diversity at genomic disorder loci. *Cell Genomics* **4:** 100590. doi:10.1016/j.xgen.2024.100590

Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, Rizzo WB, Evrony G, Curran T, Bourque G, et al. 2024. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun* **15:** 657. doi:10.1038/s41467-024-44980-2

Gustafson JA, Gibson SB, Damaraju N, Zalusky MP, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34:** 2061–2073. doi:10.1101/gr.279273.124

Hartley T, Marshall D, Acker M, Fooks K, Gillespie MK, Price EM, Graham ID, White-Brown A, MacKay L, Macdonald SK, et al. 2024. Evaluation of the diagnostic accuracy of exome sequencing and its impact on diagnostic thinking for patients with rare disease in a publicly funded health care system: a prospective cohort study. *Genet Med* **26:** 101012. doi:10.1016/j.gim.2023.101012

Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, Porubsky D, Beck CR, Marschall T, Garimella K, et al. 2023. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. *Genome Res* **33:** 2029–2040. doi:10.1101/gr.278070.123

Hengel H, Pellerin D, Wilke C, Fleszar Z, Brais B, Haack T, Traschütz A, Schöls L, Synofzik M. 2023. As frequent as polyglutamine spinocerebellar ataxias: SCA27B in a large German autosomal dominant ataxia cohort. *Mov Disord* **38:** 1557–1558. doi:10.1002/mds.29559

Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, Boston LB, Williams M, Plott CB, Jenkins J, et al. 2021. Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *Hum Genet Genomics Adv* **2:** 100023. doi:10.1016/j.xhgg.2021.100023

Hiatt SM, Lawlor JMJ, Handley LH, Latner DR, Bonnstetter ZT, Finnila CR, Thompson ML, Boston LB, Williams M, Nunez IR, et al. 2024. Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. *Genome Res* **34:** 1747–1762. doi:10.1101/gr.279227.124

Höijer I, Tsai Y-C, Clark TA, Kotturi P, Dahl N, Stattin E-L, Bondeson M-L, Feuk L, Gyllensten U, Ameur A. 2018. Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing. *Hum Mutat* **39:** 1262–1272. doi:10.1002/humu.23580

Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, Eberle M. 2024. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40:** btae042. doi:10.1093/bioinformatics/btae042

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* **27:** 677–685. doi:10.1101/gr.214007.116

Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, Toyoshima Y, Kakita A, Takahashi H, Suzuki Y, et al. 2018. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet* **50:** 581–590. doi:10.1038/s41588-018-0067-2

Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, Almansour MA, Kikuchi JK, Taira M, Mitsui J, et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat Genet* **51:** 1222–1232. doi:10.1038/s41588-019-0458-z

Kernohan KD, Boycott KM. 2024. The expanding diagnostic toolbox for rare genetic diseases. *Nat Rev Genet* **25:** 401–415. doi:10.1038/s41576-023-00683-w

Kolesnikov A, Cook D, Nattestad M, Brambrink L, McNulty B, Gorzynski J, Goenka S, Ashley EA, Jain M, Miga KH, et al. 2024. Local read haplotagging enables accurate long-read small variant calling. *Nat Commun* **15:** 5907. doi:10.1038/s41467-024-50079-5

Kosugi S, Terao C. 2024. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var* **11:** 18. doi:10.1038/s41439-024-00276-x

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20:** 117. doi:10.1186/s13059-019-1720-5

Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, Haack TB, Graf E, Schwarzmayr T, Terrile C, et al. 2017. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun* **8:** 15824. doi:10.1038/ncomms15824

Lee H, Huang AY, Wang L, Yoon AJ, Renteria G, Eskin A, Signer RH, Dorrani N, Nieves-Rodriguez S, Wan J, et al. 2020. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet Med* **22:** 490–499. doi:10.1038/s41436-019-0672-1

Lei M, Liang D, Yang Y, Mitsuhashi S, Katoh K, Miyake N, Frith MC, Wu L, Matsumoto N. 2020. Long-read DNA sequencing fully characterized chromothripsis in a patient with Langer-Giedion syndrome and Cornelia de Lange syndrome-4. *J Hum Genet* **65:** 667–674. doi:10.1038/s10038-020-0754-6

Li R, Hsieh C-L, Young A, Zhang Z, Ren X, Zhao Z. 2015. Illumina synthetic long read sequencing allows recovery of missing sequences even in the "Finished" *C. elegans* genome. *Sci Rep* **5:** 10814. doi:10.1038/srep10814

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617:** 312–324. doi:10.1038/s41586-023-05896-x

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21:** 597–614. doi:10.1038/s41576-020-0236-x

Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S, Pionzio A, Schatz MC, et al. 2024. Utility of long-read sequencing for All of Us. *Nat Commun* **15:** 837. doi:10.1038/s41467-024-44804-3

Majidian S, Sedlazeck FJ. 2020. PhaseME: automatic rapid assessment of phasing quality and phasing improvement. *Gigascience* **9:** giaa078. doi:10.1093/gigascience/giaa078

Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, Duffy E, Hegde M, Santani A, Lebo M, et al. 2016. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med* **18:** 1282–1289. doi:10.1038/gim.2016.58

Mangin A, de Pontual L, Tsai Y-C, Monteil L, Nizon M, Boisseau P, Mercier S, Ziegle J, Harting J, Heiner C, et al. 2021. Robust detection of somatic mosaicism and repeat interruptions by long-read targeted sequencing in myotonic dystrophy type 1. *Int J Mol Sci* **22:** 2616. doi:10.3390/ijms22052616

Manickam K, McClain MR, Demmer LA, Biswas S, Kearney HM, Malinowski J, Massingham LJ, Miller D, Yu TW, Hisama FM. 2021. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the

American College of Medical Genetics and Genomics (ACMG). *Genet Med* **23:** 2029–2037. doi:10.1038/s41436-021-01242-6

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. 2019. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29:** 635–645. doi:10.1101/gr.234443.118

Mastrorosa FK, Miller DE, Eichler EE. 2023. Applications of long-read sequencing to Mendelian genetics. *Genome Med* **15:** 42. doi:10.1186/s13073-023-01194-3

Melas M, Kautto EA, Franklin SJ, Mori M, McBride KL, Mosher TM, Pfau RB, Hernandez-Gonzalez ME, McGrath SD, Magrini VJ, et al. 2022. Long-read whole genome sequencing reveals HOXD13 alterations in synpolydactyly. *Hum Mutat* **43:** 189–199. doi:10.1002/humu.24304

Méreaux J-L, Davoine C-S, Pellerin D, Coarelli G, Coutelier M, Ewenczyk C, Monin M-L, Anheim M, Le Ber I, Thobois S, et al. 2024. Clinical and genetic keys to cerebellar ataxia due to *FGF14* GAA expansions. *EBioMedicine* **99:** 104931. doi:10.1016/j.ebiom.2023.104931

Miga KH, Wang T. 2021. The need for a human pangenome reference sequence. *Annu Rev Genomics Hum Genet* **22:** 81–102. doi:10.1146/annurev-genom-120120-081921

Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, Lewis AP, Fuerte EPA, Paschal CR, Walsh T, et al. 2021. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet* **108:** 1436–1449. doi:10.1016/j.ajhg.2021.06.006

Miller DE, Lee L, Galey M, Kandhaya-Pillai R, Tischkowitz M, Amalnath D, Vithlani A, Yokote K, Kato H, Maezawa Y, et al. 2022. Targeted long-read sequencing identifies missing pathogenic variants in unsolved Werner syndrome cases. *J Med Genet* **59:** 1087–1094. doi:10.1136/jmedgenet-2022-108485

Miyatake S, Koshimizu E, Fujita A, Doi H, Okubo M, Wada T, Hamanaka K, Ueda N, Kishida H, Minase G, et al. 2022. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *NPJ Genomic Med* **7:** 62. doi:10.1038/s41525-022-00331-y

Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, Nakamura M, Nagasaki M, Kinoshita K, Okamura Y, et al. 2019. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J Hum Genet* **64:** 359–368. doi:10.1038/s10038-019-0569-5

Mizuguchi T, Okamoto N, Yanagihara K, Miyatake S, Uchiyama Y, Tsuchida N, Hamanaka K, Fujita A, Miyake N, Matsumoto N. 2021. Pathogenic 12-kb copy-neutral inversion in syndromic intellectual disability identified by high-fidelity long-read sequencing. *Genomics* **113:** 1044–1053. doi:10.1016/j.ygeno.2020.10.038

Murdock DR, Dai H, Burrage LC, Rosenfeld JA, Ketkar S, Müller MF, Yépez VA, Gagneur J, Liu P, Chen S, et al. 2021. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *J Clin Invest* **131:** e141500. doi:10.1172/JCI141500

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

Novis LE, Frezatti RS, Pellerin D, Tomaselli PJ, Alavi S, Della Coleta MV, Spitz M, Dicaire M-J, Iruzubieta P, Pedroso JL, et al. 2023. Frequency of GAA-*FGF14* ataxia in a large cohort of Brazilian patients with unsolved adult-onset cerebellar ataxia. *Neurol Genet* **9:** e200094. doi:10.1212/NXG.0000000000200094

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Oehler JB, Wright H, Stark Z, Mallett AJ, Schmitz U. 2023. The application of long-read sequencing in clinical settings. *Hum Genomics* **17:** 73. doi:10.1186/s40246-023-00522-3

Pauper M, Kucuk E, Wenger AM, Chakraborty S, Baybayan P, Kwint M, van der Sanden B, Nelen MR, Derks R, Brunner HG, et al. 2021. Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur J Hum Genet* **29:** 637–648. doi:10.1038/s41431-020-00770-0

Pei S, Liu T, Ren X, Li W, Chen C, Xie Z. 2021. Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief Bioinform* **22:** bbaa148. doi:10.1093/bib/bbaa148

Pellerin D, Danzi MC, Wilke C, Renaud M, Fazal S, Dicaire M-J, Scriba CK, Ashton C, Yanick C, Beijer D, et al. 2023. Deep intronic *FGF14* GAA repeat expansion in late-onset cerebellar ataxia. *N Engl J Med* **388:** 128–141. doi:10.1056/NEJMoa2207406

Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487:** 190–195. doi:10.1038/nature11236

Rafehi H, Read J, Szmulewicz DJ, Davies KC, Snell P, Fearnley LG, Scott L, Thomsen M, Gillies G, Pope K, et al. 2023. An intronic GAA repeat expansion in *FGF14* causes the autosomal-dominant adult-onset ataxia SCA50/ATX-FGF14. *Am J Hum Genet* **110:** 105–119. doi:10.1016/j.ajhg.2022.11.015

Rajan-Babu I-S, Dolzhenko E, Eberle MA, Friedman JM. 2024. Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nat Rev Genet* **25:** 476–499. doi:10.1038/s41576-024-00696-z

Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* **14:** 411–413. doi:10.1038/nmeth.4189

Sabatella M, Mantere T, Waanders E, Neveling K, Mensenkamp AR, van Dijk F, Hehir-Kwa JY, Derks R, Kwint M, O'Gorman L, et al. 2021. Optical genome mapping identifies a germline retrotransposon insertion in *SMARCB1* in two siblings with atypical teratoid rhabdoid tumors. *J Pathol* **255:** 202–211. doi:10.1002/path.5755

Sanchis-Juan A, Megy K, Stephens J, Armirola Ricaurte C, Dewhurst E, Low K, French CE, Grozeva D, Stirrups K, Erwood M, et al. 2023. Genome sequencing and comprehensive rare-variant analysis of 465 families with neurodevelopmental disorders. *Am J Hum Genet* **110:** 1343–1355. doi:10.1016/j.ajhg.2023.07.007

Sanford Kobayashi E, Batalov S, Wenger AM, Lambert C, Dhillon H, Hall RJ, Baybayan P, Ding Y, Rego S, Wigby K, et al. 2022. Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. *Sci Rep* **12:** 16945. doi:10.1038/s41598-022-20113-x

Schieffer KM, Feldman AZ, Kautto EA, McGrath S, Miller AR, Hernandez-Gonzalez ME, LaHaye S, Miller KE, Koboldt DC, Brennan P, et al. 2021. Molecular classification of a complex structural rearrangement of the *RB1* locus in an infant with sporadic, isolated, intracranial, sellar region retinoblastoma. *Acta Neuropathol Commun* **9:** 61. doi:10.1186/s40478-021-01164-z

Schloissnig S, Pani S, Rodriguez-Martin B, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, et al. 2024. Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. bioRxiv doi:10.1101/2024.04.18.590093

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27:** 849–864. doi:10.1101/gr.213611.116

Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. 2022. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends Genet* **38:** 1134–1146. doi:10.1016/j.tig.2022.06.003

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445–449. doi:10.1126/science.1138659

Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, Baid G, Kolmogorov M, Eizenga JM, Miga KH, et al. 2021. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods* **18:** 1322–1332. doi:10.1038/s41592-021-01299-w

Sharma S, Repnikova E, Noel-MacDonnell JR, LePichon J. 2021. Diagnostic yield of genetic testing in 324 infants with hypotonia. *Clin Genet* **100:** 752–757. doi:10.1111/cge.14057

Shashi V, McConkie-Rosell A, Rosell B, Schoch K, Vellore K, McDonald M, Jiang Y-H, Xie P, Need A, Goldstein DB. 2014. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet Med* **16:** 176–182. doi:10.1038/gim.2013.99

Shickh S, Mighton C, Uleryk E, Pechlivanoglou P, Bombard Y. 2021. The clinical utility of exome and genome sequencing across clinical indications: a systematic review. *Hum Genet* **140:** 1403–1416. doi:10.1007/s00439-021-02331-x

Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, et al. 2019. Long-read sequencing identifies GGC repeat expansions in *NOTCH2NLC* associated with neuronal intranuclear inclusion disease. *Nat Genet* **51:** 1215–1221. doi:10.1038/s41588-019-0459-y

Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, Eng CM, Esteves C, Gahl WA, Hamid R, et al. 2018. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N Engl J Med* **379:** 2131–2139. doi:10.1056/NEJMoa1714458

Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, Firth HV, Frazier T, Hansen RL, Prock L, et al. 2019. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med* **21:** 2413–2421. doi:10.1038/s41436-019-0554-6

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61:** 437–455. doi:10.1146/annurev-med-100708-204735

Stergachis AB, Blue EE, Gillentine MA, Wang L-K, Schwarze U, Cortés AS, Ranchalis J, Allworth A, Bland AE, Chanprasert S, et al. 2023. Full-length isoform sequencing for resolving the molecular basis of Charcot-Marie-Tooth 2A. *Neurol Genet* **9:** e200090. doi:10.1212/NXG.0000000000200090

Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, Tchan M, Fung V, Ng K, Cortese A, et al. 2022. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv* **8:** eabm5386. doi:10.1126/sciadv.abm5386

Steyaert W, Sagath L, Demidov G, Yépez VA, Esteve-Codina A, Gagneur J, Ellwanger K, Derks R, Weiss M, den Ouden A, et al. 2024. Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. medRxiv doi:10.1101/2024.05.03.24305331

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349:** aab3761. doi:10.1126/science.aab3761

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Sund KL, Liu J, Lee J, Garbe J, Abdelhamed Z, Maag C, Hallinan B, Wu SW, Sperry E, Deshpande A, et al. 2024. Long-read sequencing and optical genome mapping identify causative gene disruptions in noncoding sequence in two patients with neurologic disease and known chromosome abnormalities. *Am J Med Genet A* **194:** e63818. doi:10.1002/ajmg.a.63818

Tan D, Wei C, Chen Z, Huang Y, Deng J, Li J, Liu Y, Bao X, Xu J, Hu Z, et al. 2023. CAG repeat expansion in *THAP11* is associated with a novel spinocerebellar ataxia. *Mov Disord* **38:** 1282–1293. doi:10.1002/mds.29412

Taylor DJ, Eizenga JM, Li Q, Das A, Jenike KM, Kenny EE, Miga KH, Monlong J, McCoy RC, Paten B, et al. 2024. Beyond the human genome project: the age of complete human genome sequences and pangenome references. *Annu Rev Genomics Hum Genet* **25:** 77–104. doi:10.1146/annurev-genom-021623-081639

Tian Y, Wang J-L, Huang W, Zeng S, Jiao B, Liu Z, Chen Z, Li Y, Wang Y, Min H-X, et al. 2019. Expansion of human-specific GGC repeat in neuronal intranuclear inclusion disease-related disorders. *Am J Hum Genet* **105:** 166–176. doi:10.1016/j.ajhg.2019.05.013

van Dijk EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, Hernandez C. 2023. Genomics in the long-read sequencing era. *Trends Genet* **39:** 649–671. doi:10.1016/j.tig.2023.04.006

Vissers LELM, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, et al. 2010. A de novo paradigm for mental retardation. *Nat Genet* **42:** 1109–1112. doi:10.1038/ng.712

Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, Wenger AM, Concepcion GT, Kronenberg ZN, Munson KM, et al. 2020. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* **84:** 125–140. doi:10.1111/ahg.12364

Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, Sun Y, Anderson E, Lam HK, Chen D, et al. 2019. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res* **29:** 798–808. doi:10.1101/gr.245126.118

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Xie Z, Sun C, Zhang S, Liu Y, Yu M, Zheng Y, Meng L, Acharya A, Cornejo-Sanchez DM, Wang G, et al. 2020. Long-read whole-genome sequencing for the genetic diagnosis of dystrophinopathies. *Ann Clin Transl Neurol* **7:** 2041–2046. doi:10.1002/acn3.51201

Yamada M, Okuno H, Okamoto N, Suzuki H, Miya F, Takenouchi T, Kosaki K. 2023. Diagnosis of Prader-Willi syndrome and Angelman syndrome by targeted nanopore long-read sequencing. *Eur J Med Genet* **66:** 104690. doi:10.1016/j.ejmg.2022.104690

Yanagi K, Coker J, Miyana K, Aso S, Kobayashi N, Satou K, Richman A, Indupuru S, Matsubara Y, Kaname T. 2023. Biallelic CC2D2A variants, SNV and LINE-1 insertion simultaneously identified in siblings using long-read whole-genome sequencing and haplotype phasing. *J Hum Genet* **68:** 431–435. doi:10.1038/s10038-023-01130-8

Yépez VA, Gusic M, Kopajtich R, Mertes C, Smith NH, Alston CL, Ban R, Beblo S, Berutti R, Blessing H, et al. 2022. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med* **14:** 38. doi:10.1186/s13073-022-01019-9

Zalusky MPG, Gustafson JA, Bohaczuk SC, Mallory B, Reed P, Wenger T, Beckman E, Chang IJ, Paschal CR, Buchan JG, et al. 2024. 3-hour genome sequencing and targeted analysis to rapidly assess genetic risk. *Genet Med Open* **2:** 101833. doi:10.1016/j.gimo.2024.101833

Zeng S, Zhang M-Y, Wang X-J, Hu Z-M, Li J-C, Li N, Wang J-L, Liang F, Yang Q, Liu Q, et al. 2019. Long-read sequencing identified intronic repeat expansions in *SAMD12* from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. *J Med Genet* **56:** 265–270. doi:10.1136/jmedgenet-2018-105484

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34:** 303–311. doi:10.1038/nbt.3432

The additional diagnostic yield of long-read sequencing in undiagnosed rare diseases

Giulia F. Del Gobbo and Kym M. Boycott

Genome Res. 2025 35: 559-571 originally published online February 3, 2025
Access the most recent version at doi:10.1101/gr.279970.124