

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/189174/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Benarroch, Louise, Pešović, Jovan, Rossato, Marzia, Chen, Zhongbo, Lucas, Morghan C. and Dion, Vincent 2026. Toward the clinical application of long-read sequencing in repeat-expansion disorders. *Nature Genetics* 58 , pp. 2087-2103. 10.1038/s41588-026-02694-9

Publishers page: <https://doi.org/10.1038/s41588-026-02694-9>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Towards the clinical application of long-read sequencing in repeat-expansion disorders

Louise Benarroch¹, Jovan Pešović², Marzia Rossato³, Zhongbo Chen⁴, Morghan C Lucas^{5,6,7}, Guillaume Cogan⁸, Indhu-Shree Rajan-Babu⁹, Vitus Prokosch⁵, Christina Zarouchlioti^{10,11}, Marco Carlomagno³, Ida Höijer^{12,13}, Mayra Sauer⁵, Ismail Jamail^{14,15}, Shota Shibata^{16,17,18}, Glen Lester Sequiera⁹, Benjamin P Kleinstiver^{19,20,21}, Sonia Lameiras¹⁴, Ricardo Mouro Pinto^{16,17,18}, Giuseppe Novelli²², Davina Hensman Moss^{23,24}, Natalia Dominik²⁵, Mathieu Barbier²⁶, Isabelle Le Ber^{27,28}, Guillaume Bassez^{1,29}, Denis Furling¹, Massimo Delledonne³, Karen Usdin³⁰, Sylvain Baulande¹⁴, Geneviève Gourdon¹, Birgitt Schüle³¹, Dušanka Savić-Pavićević², Annalisa Botta²², Alice E Davidson^{10,11}, Adam Ameur^{12,13}, Vincent Dion^{32,33}, Stéphanie Tomé^{1*} on behalf of LRS-RED Consortium

1. Sorbonne Université, Inserm, Institut de Myologie, Centre de Recherche en Myologie, Paris, France
2. University of Belgrade – Faculty of Biology, Center for Human Molecular Genetics, Studentski trg 16, 11000 Belgrade, Serbia
3. Department of Biotechnology, University of Verona, Strada le Grazie 15, 37134 Verona, Italy
4. Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University London, London, UK
5. MGZ – Medical Genetics Center, Munich, Germany
6. Medizinische Klinik und Poliklinik IV, Campus Innenstadt, Klinikum der Universität München, Munich, Germany
7. Friedrich-Baur-Institute, Department of Neurology, Klinikum der Universität, Ludwig-Maximilians-Universität, Munich, Germany
8. Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, INSERM U1127, CNRS UMR7225, AHP, Paris, France
9. Department of Medical Genetics, University of British Columbia, BC Children's and Women's Hospital
10. UCL Institute of Ophthalmology, London, UK
11. Moorfields Eye Hospital NHS Foundation Trust, London, UK
12. Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala
13. SciLifeLab, Uppsala University, Sweden
14. Institut Curie, PSL University, CurieCoreTech Next-Generation Sequencing (ICGex), 75005 Paris, France
15. Institut Curie, INSERM U1331, Mines ParisTech, PSL University, CurieCoreTech Bioinformatics (CUBIC), 75005 Paris, France
16. Center for Genomic Medicine, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
17. Department of Neurology, Harvard Medical School, Boston, MA, USA
18. Medical and Population Genetics Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA
19. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
20. Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA
21. Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

22. Department of Biomedicine and Prevention, University of Rome "Tor Vergata", Rome, Italy
23. St George's, University of London, London, UK
24. Huntington's Disease Centre, Department of Neurodegenerative Disease, Queen Square Institute of Neurology, UCL, London, UK
25. Department of Neuromuscular Diseases, Queen Square Institute of Neurology, UCL, London, UK
26. Sorbonne Université, Paris Brain Institute – Institut du Cerveau, ICM, Inserm U1127, CNRS UMR7225, APHP – Hôpital Pitié-Salpêtrière, Paris 75013, France
27. Paris Brain Institute-Institut du Cerveau-ICM, Sorbonne Université, Inserm U1127, CNRS UMR 7225, AP-HP-Hôpital Pitié-Salpêtrière, Paris, France
28. Reference Centre for Rare or Early-Onset Dementias, IM2A, Département de Neurologie, AP-HP-Hôpital Pitié-Salpêtrière, Paris, France
29. Centre de Référence des Maladies Neuromusculaires Nord-Est/Île-de-France, Institut de Myologie, GHU Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Paris, France.
30. Laboratory of Cell and Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA
31. Department of Pathology, Stanford University School of Medicine, 300 Pasteur Dr., Stanford, CA 94305-5324, USA
32. UK Dementia Research Institute at Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK
33. Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Maindy Road, Cardiff CF24 4HQ, UK

*Corresponding author: stephanie.tome@inserm.fr

Abstract

Repeat expansion disorders (REDs) are a mechanistically and clinically well-defined subgroup of rare diseases caused by the expansion of short tandem repeats (STRs). These expansions can exceed several kilobases and exhibit complex features, such as non-canonical secondary structures, somatic instability, repeat interruptions, and allele-specific methylation. These characteristics are highly relevant for understanding disease mechanisms, clinical variability, prognosis, and potentially therapeutic decision-making, but cannot be fully resolved using traditional diagnostic methods or short-read sequencing technologies. In contrast, long-read sequencing (LRS) enables accurate investigation of STR complexity in a single assay, facilitates the discovery of novel pathogenic repeat expansions and drives advances in diagnostics,

81 clinical and basic research, which may allow for better patient stratification in future clinical
82 trials. This Review discusses recent LRS-driven discoveries, methodological and bioinformatic
83 advances, and emerging diagnostic applications, to illustrate the potential of LRS in reshaping
84 both research and clinical practice.

85

86 **Keywords**

87 Repeat expansion disorders, long-read sequencing, clinical applications

88

Main

Short tandem repeats (STRs) are polymorphic sequences representing nearly five million loci in the human genome^{1,2}. STR motifs range from 1 to 6 bp, with tri- and hexanucleotide motifs being the most common in the coding regions³. The genome-wide mutation rate of STRs is estimated at $\sim 5.5 \times 10^{-5}$ mutations per locus per generation, higher than in non-repetitive sequences⁴⁻⁷. STRs can increase in number, i.e., expand, and cause neurodegenerative, neuromuscular and neurodevelopmental diseases, termed repeat-expansion disorders (REDs)⁸. They affect 1 in 3000 individuals worldwide⁹. Following the discovery of the first repeat expansion that causes a disorder in 1991, technological advances in the detection and sequencing of repeat expansions have led to the identification of over 50 REDs⁸ (Suppl. Table 1). Repeat expansions are found within coding exons, 5' or 3' untranslated regions (UTRs), introns, and even deep intergenic regions. They vary in motif size, composition and pathogenic threshold size¹⁰. These characteristics influence their behavior, stability and resulting pathogenicity. For instance, a biallelic expansion of a AAGGG motif in the *RFC1* gene is pathogenic, causing autosomal recessive cerebellar ataxia, neuropathy, and vestibular areflexia syndrome, but expansions of AAAAG, AAAGGG, and AAGAG at the same locus appear benign¹¹. Moreover, expanded repeat loci are highly polymorphic between populations as highlighted by studies in Caucasian¹¹ and Māori¹² populations, reflecting differences in population history and evolution, and highlighting the importance population-scale analyses.

Phenotypic variability in REDs, including age at onset, symptom severity, and penetrance, has been associated with germline and somatic instability of repeat expansions. Anticipation, i.e. the earlier and more severe appearance of clinical symptoms, is caused by germline expansions in dominantly inherited REDs⁸. This instability can be influenced by repeat

expansion size and sequence, and the sex and age of the transmitting parent. These factors complicate accurate prediction of intergenerational expansion sizes and thus the severity of the disease in subsequent generations¹³. Somatic instability, which is biased towards expansions, leads to age- and tissue-dependent somatic mosaicism and is a major contributor to disease progression in multiple REDs^{14–16}, with larger expansions associated with earlier onset and more severe phenotypes¹⁷. However, patients with similar expansion sizes can also show differences in their clinical presentation, suggesting that additional genetic, epigenetic, and environmental factors play a role. Among these, sequence variants within repeat tracts, termed repeat interruptions, are key modifiers. In Huntington disease (HD) and myotonic dystrophy type 1 (DM1), interruptions delay disease onset^{18–20}. In contrast, in spinocerebellar ataxia type 8 (SCA8)²¹ and 10 (SCA10)^{22–25}, interruptions can increase penetrance and cause an earlier onset. Interruptions can also alter symptoms presentation, for instance, in SCA 10 they raise the risk of seizures. Furthermore, DNA hypermethylation at CpG dinucleotides located within or near repeat expansions has been linked to differences in age at onset and disease severity across several REDs²⁶. Indeed, hypermethylation of the disease-causing repeats is associated with an earlier onset in HD, and with carriers in oculopharyngodistal myopathy (OPDM) (Suppl. Table 1)

Several approaches exist for detecting repeat expansions²⁷, each with distinct sensitivity and resolution for key molecular features (repeat length, sequence composition, mosaicism, and methylation status), which are critical for diagnosing REDs and establishing genotype–phenotype correlations. Conventional “gold-standard” approaches, such as repeat-primed PCR and Southern blot, are limited in their ability to analyze large and complex repeat expansions, including long repeats with interruptions or high GC content (see Box 1). In

contrast, long-read sequencing (LRS) technologies enable the accurate, simultaneous characterization of repeat length, sequence composition, mosaicism, and methylation a single molecule level with base-pair resolution (Fig. 1)^{28–30}. Here, we discuss long-read sequencing (LRS) technologies, highlighting their impact on repeat expansion discovery, as well as the development of dedicated bioinformatic approaches and their implementation in clinical diagnostics.

Brief overview over LRS technologies

Long-read sequencing platforms: Pacific Biosciences and Oxford nanopore technologies

Two LRS technologies currently dominate the field: Pacific Biosciences (PacBio), which detects nucleotide incorporations in real time through a fluorescence-based system³¹ and Oxford Nanopore Technologies (ONT), which identifies bases by measuring electrical signals as DNA passes through nanopores³². Regardless of the technology, several major library preparation methods can be distinguished such as PCR-based targeted approaches³³ or amplification-free approaches including whole-genome sequencing or targeted enrichment using CRISPR-Cas9^{34,35} (Fig. 1). More recently, ONT has developed a unique adaptive sampling method, which allows a selective sequencing of genomic DNA fragments matching target loci without molecular manipulations during library preparation³⁶. These approaches are summarized in Table 1 and their advantages and limitations discussed in Box 2. PacBio sequencing generates highly accurate HiFi reads (>99.9% accuracy, corresponding to a Phred quality score >30, where the Phred score reflects the probability of a base-calling error, i.e. an error rate <0.1% in this case)³⁷ making it particularly suitable for resolving repeat motifs and subtle sequence variations, such as single-repeat interruptions or novel motifs. However, its average read length of ~20 kb, may limit coverage for very large expansions or regions extending beyond

this range³⁸. In contrast, ONT generates reads with a more variable accuracy, typically 99–99.9% (corresponding to Phred scores of 20-30, i.e. an error rate 1% to <0.1%) depending on experimental setup³⁹, but offers virtually no upper limit with regard to read length. However, achieving ultra-long reads (>20kbp), typically requires to minimize DNA fragmentation, which can reduce sequencing throughput; therefore, read length and yield must be balanced according to the target repeat size. Both platforms can directly detect 5mC in native DNA, providing base-resolved methylation information alongside nucleotide sequence. Both methods can also detect 5-hydroxymethylcytosine (5-hmC), although this has not been robustly tested with REDs⁴⁰.

Remaining technological limitations and challenges

Sequencing ultra-long repeat expansions (>20kbp) remains difficult as coverage is often low and heavily biased towards shorter alleles, even with amplification-free approaches⁴¹. This limitation can be somewhat mitigated by optimizing DNA extraction, library preparation, sample multiplexing, as well as choosing the appropriate sequencing platform (PacBio or ONT), since all these parameters affect coverage, with PacBio and ONT differing in their typical throughput and read length distributions, which in turn can influence overall coverage (Box 2).

Motif composition markedly influences the ability to sequence a repeat, regardless of the platform. For example, at the *RFC1* and *FXN* loci, a decreased coverage is often reported, likely due to challenges in resolving specific repeat motifs or their ability to form noncanonical DNA conformations that impede sequencing^{35,42}. These issues can be exacerbated by certain library protocols or sequencing methods, and allelic dropout can serve as an indicator of a potential repeat expansion³⁵. Finally, analyzing repeat expansions from very small amounts of DNA, such as single cells, remains a major challenge. The emerging long-read single-cell DNA

methods^{43,44} still suffer from amplification bias, limited read lengths, and high cost. As an alternative, long-read single-cell RNA-sequencing^{45–47} can be used, if the repeat region is transcribed.

Impact of LRS in RED discovery

Landmark discoveries in the RED field

The introduction of LRS has fundamentally transformed the field of REDs^{48,49}. By generating reads sufficiently long to span entire repeat tracts, these technologies have enabled the discovery of over one-third of the currently known disease-associated repeat loci⁵⁰. Many loci were identified through clinical and linkage studies, but the causal repeat expansions remained undetected due to their size, non-reference motifs, or localization within intronic, highly repetitive, or GC-rich regions, inaccessible to PCR-based and short-read sequencing (SRS) methods. The REDs discovered by LRS encompass diverse repeat motifs, each linked to distinct clinical conditions, including myopathies, adult-onset ataxias, and familial epilepsies. A notable example is OPDM, a rare neuromuscular disorder caused by 5'-CGG repeat expansions⁵¹. Although described clinically since the late 1970s, its first causative gene was only identified in 2019 using short-read whole-genome sequencing (SR-WGS) and repeat primed PCR (RP-PCR)⁵¹. Subsequent long-read whole-genome sequencing (LR-WGS) revealed three additional OPDM genes (*GIPC1*, *NOTCH2NLC*, *RILPL1*), all with 5'-UTR CGG expansions^{52–54}. Another gene (*ABCD3*) was identified through linkage and SR-WGS, but only amplification-free targeted LRS (Adaptive sampling-ONT) made it possible to assess the size and methylation of expanded alleles⁵⁵. Similarly, a CGG expansion in the 5'-UTR of the *NUTM2B-AS1* gene, first linked to oculopharyngeal myopathy with leukoencephalopathy (OPML1) by using SR-WGS⁵¹, was later shown to instead underlie OPDM based on the use of LRS⁵⁶. Collectively, these

findings revealed that OPMD is driven by CGG expansions, rather than the canonical function of the host gene, although downstream pathogenic mechanism may differ across loci. Another example was the identification of a GAA expansion in intron 1 of the *FGF14* gene in two independent cohorts, establishing a new autosomal dominant late-onset cerebellar ataxia, SCA27B^{57,58}. Here, SR-WGS was combined with long-range PCR (LR-PCR) for the discovery of the underlying mutation, while targeted LRS resolved the length and structure of pathogenic alleles^{57,58}. Likewise, LR-WGS and amplification-free targeted LRS identified a pathogenic GGC expansion in exon 10 of the *ZFHX3* gene, defining SCA4^{59–62}. Familial adult myoclonic epilepsies (FAME) are autosomal dominant disorders caused by intronic TTTTA and TTCA expansions. Although the initial loci, *SAMD12*⁶³, *STARD7*⁶⁴, and *MARCH6*⁶⁵, were mapped through linkage and SR-WGS, pathogenic repeat size and configuration were only characterized in detail by using LR-WGS. LRS-WGS alone have also identified repeat expansions in *YEATS2* (underlying FAME4) and *RAI1*^{66,67}, expanding the motif-defined phenotypic spectrum of these disorders.

Importantly, repeat expansions in different loci, such as *NOTCH2NLC* and *FGF14*, have been implicated in Parkinson's disease^{68,69}, highlighting the role of LRS in addressing the hidden heritability of common neurodegenerative disorders.

Structural and mechanistic insights from full-length repeat sequencing

As LRS technologies improve and are applied to REDs, additional insights are gained. For example, a landmark HD study developed a single-cell long-read RNA sequencing approach to analyze the *HTT* (CAG)_n tract in post-mortem brain⁴⁵, suggesting that only repeats above 150 CAGs induce neuronal loss in striatum. This established a mechanistic threshold for cellular toxicity and informs therapeutic strategies to delay or halt somatic expansions⁷⁰.

LRS has also revealed ultra-long expansions beyond the reach of PCR-based or short-read sequencing approaches. For example, targeted amplification-free LRS resolved intronic CCTG repeat expansions up to ~50 kb in the *CNBP* gene, which cause myotonic dystrophy type 2^{41,71}. Beyond repeat length, this approach also uncovered the presence of TCTG blocks at the 3' end of the expansion, whose potential impact on disease phenotype remains to be determined⁴¹. Furthermore, amplification-free targeted LRS of the *DMPK* locus enabled the identification of interruptions in DM1 patients that had been previously missed by RP-PCR⁷². While RP-PCR studies found interruptions that are more frequent near the repeat tract ends⁷³, a recent DM1 study using PacBio amplification-free target enrichment with CRISPR-Cas9 revealed alleles with centrally located, complex interruption blocks that vary across generations and tissues⁷⁴. The use of LRS also allowed to clarify allele pathogenicity linked to repeat motif heterogeneity, particularly in pentanucleotide repeat-expansion loci, which often exhibit a complex mosaicism of reference and non-reference motifs⁷⁵. Using PacBio amplification-free target enrichment with CRISPR-Cas9, the long (ATTCT)_n tract at the *ATXN10* locus was fully resolved, showing that expansions containing ATTCC interruptions are associated with SCA10²², whereas uninterrupted tracts may be either clinically silent or lead to Parkinsonism²⁵. These findings underscore the notion that repeat purity, motif composition, and structure, rather than expansion size alone, determine the impact of repeat expansions.

Amplification-free LRS has enabled direct measurement of DNA methylation at the level of individual expanded alleles. This is particularly significant for REDs, in which methylation status modulates pathogenicity or contributes to clinical variability^{40,76,77}. Indeed, allele-specific 5-methylcytosine (5-mC) profiling using whole-genome or amplification-free targeted LRS has provided key mechanistic and clinical insights at loci, such as *FMR1* and *C9orf72*^{40,77}. At the *FMR1* locus, methylation mosaicism, characterized by the presence of unmethylated,

transcriptionally active alleles together with predominantly methylated full mutations, is typically associated with milder clinical presentations, making an accurate discrimination important for diagnosis and management^{78–80}. Amplification-free LRS has also enabled a clear distinction between males with Fragile X syndrome carrying fully methylated full mutations and individuals with FXTAS, who harbor unmethylated premutations or full mutations. It can also resolve mosaicism with regard to size and/or methylation, as well as skewed X-inactivation patterns in female carriers, providing important diagnostic and prognostic information^{77,81}. Amplification-free LRS also enabled the first quantitative, allele-specific methylation profiling at *C9orf72* in amyotrophic lateral sclerosis and/or frontotemporal dementia, revealing a correlation between the repeat number and the extent of DNA methylation, and increased methylation and age at sampling⁴⁰. Importantly, LRS can detect both repeat expansions and non-repeat variants, while also establishing their haplotypic phase^{76,82–85}.

Recent applications of LRS have revealed broader principles of repeat stability and mutability at both population and germline levels, extending beyond locus-specific characterization. At the *FGF14* locus, a common 5'-flanking variant, initially identified by Sanger sequencing, was later confirmed and characterized at scale using LR-WGS⁸⁶. This variant is found almost exclusively in wild-type alleles with fewer than 30 GAA repeats and is associated with increased intergenerational stability, indicating a protective effect against repeat expansion⁸⁶. Genome-wide, population-based LRS has enabled systematic profiling of repeat variation, identifying features such as the length of the longest pure motif segment and constraint metrics that help clarify the pathogenic potential of repeat loci and prioritize candidate disease-causing expansions⁸⁷.

Bioinformatics methods for repeat-expansion analysis in LRS

Repeat regions pose unique computational challenges that differ fundamentally from other genomic loci. The overarching difficulties around base-calling, alignment and genotyping of repetitive regions are dependent on the technology used and the sequence characteristics. For ONT, the choice of base-caller strongly affects genotyping and motif resolution, with some repeat expansions requiring multiple base-callers for optimal accuracy⁸⁸. This stems from the greater complexity of ONT base-calling, which must infer sequence information from raw electrical signals using computational models⁸⁹. In contrast, PacBio HiFi reads are generated directly on the instrument with fixed settings, producing highly accurate sequences without the need for downstream base-calling choices⁹⁰. After base-calling, poor-quality reads may be removed to increase the accuracy of the LRS data, although a stringent filtering may inadvertently reduce the coverage of large expansions (>5kb), which are typically more prone to sequencing errors. Several LRS repeat-expansion analysis tools have been developed that greatly increase the ability to detect and perform in-depth characterization of these sequencings (Table 2). Some of the ONT-specific methods operate directly on raw signals^{91–94}, but the majority of tools use aligned or unaligned reads and a repeat definition file as inputs^{30,95–97}. Many tools are focused on analysis of known repeat loci in targeted LRS data⁹⁸, while some have been developed for genome-wide repeat expansion discovery from WGS data^{96,99,100}. These repeat-calling tools represent a major advance by enabling the simultaneous detection of multiple repeat expansions, thereby improving diagnostic efficiency.

Repeat-length analysis, motif annotation and methylation tools

Tools that perform repeat-expansion genotyping generally operate by haplotyping or clustering reads from the repeat region, building a consensus sequence for each molecule, and then estimating repeat counts for each read³⁰ (Table 2). Targeted genotyping of RED-associated loci using these approaches offers a rapid and high-throughput means to screen a pre-defined set of loci, which can be expanded for custom or genome-wide genotyping. Repeat expansions can generally be phased using long-read sequencing data; however, challenges remain in cases where the expansion is very large and the supporting reads lack sufficient flanking sequence variation to confidently establish haplotype phase.

Tools, including Tandem Repeat Genotyping Tool (TRGT)³⁰, Straglr⁹⁶ and LongTR⁹⁵, can generate allele-specific consensus sequences or provide supporting read-level data that can be used to derive consensus sequences for downstream motif annotation, sequence composition analysis, and interpretation. In addition to sequence information, LRS data can also provide information on base modification for the analysis of DNA methylation. For ONT, methylation tools are based on raw signal calling via ONT Remora (<https://github.com/nanoporetech/remora>), a neural-network model for detecting modified bases that is integrated into the Dorado base-caller¹⁰¹. In PacBio, kinetic signals are used to detect methylation detection at the single-molecule level through on-instrument methylation calling¹⁰².

Visualization tools

Genetic variants, including point mutations, small indels, and structural variants, are typically visualized using Integrative Genomics Viewer (IGV)¹⁰³. However, it is designed for short reads and use a reference sequence, creating problems when repeats are long and/or absent from the reference genome. With LRS, reads can be phased and grouped by haplotype on IGV, but

motif composition still cannot be resolved¹⁰⁴. Specialized software is available to visualize repeats in LRS data. TRGT generates an allele plot showing alignments of HiFi reads to each allele's inferred consensus sequence, revealing repeat length, mismatches, insertions, and deletions, as well as a waterfall plot displaying the repeat segment of reads from a given locus, revealing length and mosaicism; TRGT plots can be customized to visualize motif composition and methylation states³⁰. ProleTRact, which works on TandemTwister's VCF output, generates plots showing color-coded motifs that reveal repeat architecture across individual samples and cohorts¹⁰⁵. TRMotifAnnotator¹⁰⁰ and MotifScope¹⁰⁶ can process and annotate the motif structure of any provided consensus repeat sequence, with the former generating superpopulation stratified allele composition plots for comparisons. STRkit offers a web-based visualization tool that summarizes allele genotypes and embeds IGV for read visualization at tandem repeat loci.

The need for LRS-based repeat expansion databases and benchmarking

To evaluate the potential pathogenicity of a previously uncharacterized repeat sequence, it is essential to compare it against data from healthy individuals, ideally from the same population group. In this context, population-scale LRS genome projects provide valuable reference data, and resources cataloguing repeat expansions in 1000s of samples from diverse populations and national genome initiatives are starting to emerge^{99,100,107,108}, although their availability remains limited. In the coming years, integrating data across multiple projects to build larger, unified reference databases will be critical. Achieving this will require coordinated community efforts to harmonize results between technologies, data sources, and repeat catalogues.

Although the bioinformatics methods for analysis of repeat-expansion LRS datasets are starting to mature, a comprehensive benchmarking of these tools across platforms and chemistries remains lacking. There are few studies available that compare only a handful of tools, but these do not include recently-developed methods, and often use different metrics, as well as small or non-standard reference datasets, which makes it difficult to interpret different results between studies^{95,96,109}. Therefore, we have established this consortium to drive a coordinated benchmarking framework to systematically evaluate LRS approaches and associated analytical workflows in diagnostic settings. This will include the development and distribution of community-shared reference samples with well-characterised repeat expansions spanning a range of sizes, sequence motifs, and interruption patterns. In parallel, we recommend the adoption of standardised benchmarking protocols, including clearly defined performance metrics (e.g., sizing accuracy, interruption detection, and methylation calling) and harmonised output formats for reporting these metrics. Bioinformatic methods are rapidly evolving, and emerging approaches based on graph genomes are likely to further advance our understanding of repeat expansions, including those in genomic regions that have until now been inaccessible¹¹⁰.

Implementing LRS in RED diagnostics in clinical settings

The ongoing integration of LRS into RED clinical diagnostics represents a shift from locus-specific assays toward a more comprehensive genome characterization, which requires coordinated advances across laboratory workflows, bioinformatics, and clinical interpretation.

Practical integration of LRS in diagnostic settings

Clinical implementation of LRS requires standardization of DNA extraction, library preparation, sequencing, bioinformatics, and interpretation. Each of these stages must be validated for accuracy, precision, and reproducibility to meet internationally recognized standards for medical laboratories (e.g., ISO 15189) and applicable regional or national regulatory requirements (e.g., in-vitro diagnostic regulation in the EU) (2017/, n.d.). ISO 15189 defines requirements for quality management systems and technical competence in medical laboratories and is widely adopted across healthcare systems globally (ISO 2022). Several laboratories have recently achieved ISO 15189 accreditation for LRS-based assays, demonstrating the feasibility of clinical implementation in regulated settings^{34,77,94,111,112}. Validation should be supported by laboratory quality-assurance programs, including external quality-assessment schemes and inter-laboratory comparisons using shared reference materials¹¹³, for example, through providers such as Genomic Quality Assessment (<https://genqa.org/>), to ensure consistency and reproducibility across sites.

Sample collection and DNA isolation

Blood remains the preferred tissue for diagnostic testing because it is minimally invasive and yields DNA of sufficient quality for most loci. However, obtaining high-molecular weight DNA to detect the largest expansions is prohibitively time-consuming and thus requires improved automated procedures to scale up LRS-based diagnosis from patient tissues (see Box 2). Repeat size and mosaicism can vary substantially between tissues, so the selection of the starting biological sample should consider the analysis goal. Larger expansions have been

reported in DM1 patients when muscle or brain samples are compared with blood¹¹⁴. Similarly, *TCF4* expansions measured in blood differ from those in cornea, the clinically affected tissue in Fuchs endothelial corneal dystrophy¹¹⁵. Similar discrepancies likely occur at other loci but remain underexplored, emphasizing that caution is warranted when interpreting borderline or discordant results, particularly when the tissue affected by a condition cannot be sampled.

Sequencing approaches

For now, targeted LRS approaches offer several advantages for analysis of RED loci, particularly in settings with well-defined candidate loci¹¹⁶. Targeted LRS assays based on Cas9-enrichment have already been implemented for RED diagnostics in some laboratories. The ONT-based Clin-CATS Cas9 workflow enables simultaneous multi-locus analysis, shortening turnaround and improving yield, but currently runs only one sample per flow cell, increasing cost per sample^{77,111}. Recent studies demonstrate the feasibility of LRS-based RED diagnostics^{28,34}, highlighting ONT adaptive sampling as an efficient entry point for laboratories implementing LRS workflows, as it does not require additional library preparation or specialized enrichment protocols while enabling flexible, software-driven targeting of genomic regions. Adaptive sampling performs optimally when ~0.5-10% of the genome is targeted, which readily accommodates all currently known repeat expansions which collectively span 1 to 3% of the genome¹. The PacBio PureTarget method, based on CRISPR-Cas9-mediated enrichment¹¹⁷, supports multiplexing of samples and meets clinical accuracy requirements for RED diagnostics, enabling high-capacity diagnostic workflows. Although it requires higher upfront capital investment in sequencing infrastructure, it represents the most

cost-efficient targeted enrichment strategy among current LRS approaches. Specifically, PureTarget enables multiplexing of up to 48 samples per single-molecule real time (SMRT) cell for the manual protocol and 96 samples per flow cell for the high-throughput automated version, compared to one sample per MinION flow cell for ONT Cas9-based enrichment and approximately 2–4 samples per PromethION flow cell using adaptive sampling. Due to the higher number of samples multiplexed or the possibility to use smaller flow cells, a targeted assay is generally cheaper while achieving the same yield than their WGS counterparts. Notably, the PacBio PureTarget approach has recently been implemented in a clinical diagnostic setting in the US (<https://newsroom.questdiagnostics.com/2025-07-29-Athena-Diagnostics-Improves-Ataxia-Movement-Disorder-Testing-with-Novel-Long-Read-Gene-Sequencing>)

Despite these advantages, targeted LRS has two major limitations in a clinical context (Box 3). First, among the approximately 50 known REDs, about half were identified only recently, between 2011 and 2021¹¹⁸, with several additional disorders reported since then^{55,62,86,119–121}. It is therefore likely that further pathogenic repeat expansions will be discovered. Reliance on targeted LRS necessitates iterative updates to both laboratory protocols and bioinformatic pipelines, and, importantly, it does not allow retrospective analysis of previously sequenced patients. In contrast, WG-LRS enables reanalysis of existing datasets as new disease-associated repeats are identified, without requiring additional sequencing. This comes at the cost of lower coverage and thus limited insights into somatic mosaicism. Second, even when a RED is suspected, WG-LRS offers a more comprehensive approach by enabling the detection of all variant types within a single assay, and therefore avoid the need for additional genetic testing.

In current clinical practice, the choice between targeted and genome-wide LRS is largely driven by cost, throughput, and local regulatory constraints, with targeted approaches often serving as a pragmatic interim solution where genome-wide sequencing is not yet routinely feasible. However, as sequencing costs continue to decline, WG-LRS is becoming increasingly viable in clinical settings and could replace targeted LRS. In line with this, a recent large-scale study demonstrated WG-LRS feasibility in a clinical cohort¹²².

Clinical interpretation and reporting

LRS-based RED analysis tools must be actively maintained and version-controlled, as well as integrated into pipelines that handle quality control, phasing, and annotation. To ensure compatibility with evolving sequencing chemistries, pipelines should operate on base-called data rather than raw signal. Interpretation integrates repeat length, motif composition, interruption structure, and methylation to classify alleles as either normal, premutation, or pathogenic in accordance with locus-specific guidelines. However, several factors complicate this process. Many diagnostic thresholds for repeat expansions were originally established using PCR-based or fragment analysis methods, which have their own limitations, necessitating reassessment of locus-specific cut-offs in the context of LRS^{104,123}. In addition, indels between PCR-primer-binding sites and repeat regions may affect apparent allele size, further complicating comparisons with sequencing-based measurements. Even small differences of 1 to 2 repeat units can shift genotypes across clinical thresholds, making confirmatory testing necessary for loci with strict clinical cut-offs^{124–129}. Analytical limitations further affect interpretation. Many repeat genotyping tools have been evaluated primarily against other tools rather than independent, orthogonal benchmarks, highlighting the need for systematic benchmarking using well-characterized reference materials¹³⁰. The choice of

reference genome also substantially influences repeat detection and interpretation. Current assemblies, such as GRCh38¹³¹, do not adequately capture all repeat-rich or structurally complex regions, and some pathogenic loci are missing or incompletely represented, including *XYLT1*, *RFC1*, and *SCA31*¹³². More complete assemblies, such as the telomere-to-telomere (T2T) reference, improve the representation of previously unresolved repeat-rich regions but remain linear references³⁸. More broadly, linear reference genomes represent only a limited set of haplotypes and do not capture the full genetic diversity of human populations, including population-specific repeat variation, which may affect both detection and interpretation¹³³. Emerging approaches aim to address these limitations. Reference-free (*de novo*) assembly reconstructs genomic sequences directly from reads without reliance on a reference genome, enabling detection of novel or structurally complex repeat expansions, although its clinical application remains limited. Graph-based genome representations extend linear references by modeling alternative alleles and repeat configurations as multiple paths, reducing reference bias and improving alignment in polymorphic regions. Similar principles are applied in tools such as ExpansionHunter, which use sequence graphs for targeted genotyping¹³⁴. These approaches offer potential solutions to limitations of linear reference-based genotyping but have not yet been widely applied to repeat expansion analysis.

A range of reference resources support RED interpretation, including curated clinical summaries (GeneReviews), locus- and tool-specific resources (STRchive¹²³, STRipy¹³⁵, TReXplorer², and population-scale datasets (gnomAD¹³⁶, WebSTR¹³⁷, and TRAtlas¹³⁸). Recent genome-wide tandem repeat catalogs further support locus discovery and highlight the importance of community-driven, consensus reference datasets². While these resources provide valuable information, they are largely research-focused, fragmented across platforms,

and lack standardized clinical interpretation frameworks, limiting their utility as standalone resources for diagnostic decision-making.

In contrast to single-nucleotide and structural variants, for which centralized resources such as ClinVar¹³⁹ and ClinGen¹⁴⁰ provide standardized classification frameworks, no equivalent, clinically curated infrastructure exists for repeat expansion disorders. An ideal RED-focused clinical resource would integrate key parameters required for interpretation within a standardized framework, including curated locus definitions, consensus or evidence-based thresholds and motifs for expansion categories, genotype–phenotype correlations, and population allele distributions. Achieving this will require dedicated Variant Curation Expert Panels (VCEPs)¹⁴¹ focused on REDs, as well as coordinated international efforts, standardized data models, and the systematic aggregation of clinical and genomic evidence. The growing number of accessible WG-LRS cohorts will improve our ability to interpret repeat expansions in the future^{108,122,142}.

Clinical utility of LRS beyond diagnosis

Several antisense oligonucleotide based therapies are in development for REDs, including candidates targeting *C9orf72*¹⁴³, *HTT*¹⁴⁴, *FMR1*¹⁴⁵, *ATXN2*¹⁴⁶ and *DMPK*¹⁴⁷, as well as gene-therapy approaches and small-molecule treatments. Detailed structural information from LRS could refine allele selection, improve target specificity, and support individualized therapeutic design. LRS also enables monitoring of somatic mosaicism across tissues and over time^{148–150}, providing a window into disease dynamics and potential biomarkers of therapeutic response, with the accuracy of repeat-instability quantification depending on sequencing coverage. Such monitoring could support clinical trials of repeat-modifying or gene-editing therapies by

quantifying somatic mosaicism or methylation changes as pharmacodynamic readouts. Long-read RNA-sequencing can complement these analyses by linking expanded alleles, including somatic expansion, with altered splicing or transcriptional dysregulation¹⁵¹, offering molecular endpoints for treatment development and evaluation. More importantly, long-read RNA-sequencing can detect full-length isoforms and thus help to fully appreciate the biological relevance and pathological consequences of previously uncharacterized variants.

Concluding remarks

In conclusion, by bridging molecular diagnostics with mechanistic insight, LRS provides a framework for rational clinical trial design and patient stratification, even as its current clinical implementation remains focused primarily on accurate diagnosis and classification.

Acknowledgments

The consortium thanks all colleagues who contributed to this work. The consortium was created following the second workshop on long-read sequencing and repeat expansion disorders, supported by Association Française contre les Myopathies (AFM) Telethon and the CHDI Foundation.

Funding

This work was supported by the Institut National de la Santé et de la Recherche Médicale (Inserm), Sorbonne University, the Science Fund of the Republic of Serbia, Grant number #7754217, READ-DM1 and SciLifeLab NGI Sweden technology development project funding.

This research was also funded by the European Union—Next Generation EU—NRRP M6C2—Investment 2.1 Enhancement and strengthening of biomedical research in the NHS. AED is

supported by a UKRI Future Leader Fellowship (MR/S031820/1 and MR/Y019911/1). VD is supported by the UK Dementia Research Institute (UK DRI-3204) through UK DRI Ltd, principally funded by the UK Medical Research Council.

Author contributions

L.B., J.P., M.R., Z.B., M.L., G.C., I.-S.R.-B, V.P., C.Z., M.C., I.H., M.S., I.J., S.S., G.L.S., B.S., D.S.-P., A.B., A.E.D., A.A., V.D., and S.T. contributed to the writing and editing of the manuscript. L.B., J.P., M.R., Z.B., M.L., G.C., I.-S.R.-B, V.P., C.Z., M.C., I.H., M.S., I.J., S.S., G.L.S., B.P.K, S.L., R.M.P., G.N., D.H.M., N.D., M.B., I.L.B., G.B., D.F., M.D., K.U., S.B., G.G, B.S., D.S.-P., A.B., A.E.D., A.A., V.D., and S.T. contributed to the review of the manuscript. M.R. and M.C. designed Fig. 1.

COMPETING INTERESTS

S.T. received the 2019 Targeted Sequencing SMRT Grant delivered by Pacific Biosciences. V.D. had a research contract with Pfizer Inc unrelated to this work, which ended in 2023. M.C.L. has received travel and accommodation expenses to speak at ONT conferences. All other authors declare no competing interests.

LRS-RED Consortium

Louise Benarroch¹, Jovan Pešović², Marzia Rossato³, Zhongbo Chen⁴, Morghan C Lucas^{5,6,7}, Guillaume Cogan⁸, Indhu-Shree Rajan-Babu⁹, Vitus Prokosch⁵, Christina Zarouchlioti^{10,11}, Marco Carlomagno³, Ida Höjjer^{12,13}, Mayra Sauer⁵, Ismail Jamail^{14,15}, Shota Shibata^{16,17,18}, Glen Lester Sequiera⁹, Benjamin P Kleinstiver^{19,20,21}, Sonia Lameiras¹⁴, Ricardo Mouro Pinto^{16,17,18}, Giuseppe Novelli²², Davina Hensman Moss^{23,24}, Natalia Dominik²⁵, Mathieu Barbier²⁶, Isabelle Le Ber^{27,28}, Guillaume Bassez^{1,29}, Denis Furling¹, Massimo Delledonne³, Karen Usdin³⁰, Sylvain Baulande¹⁴, Geneviève Gourdon¹, Birgitt Schüle³¹, Dušanka Savić-Pavićević², Annalisa Botta²², Alice E Davidson^{10,11}, Adam Ameur^{12,13}, Vincent Dion^{32,33}, Stéphanie Tomé^{1*} on behalf of LRS-RED Consortium

- 561 1. Sorbonne Université, Inserm, Institut de Myologie, Centre de Recherche en Myologie,
562 Paris, France
- 563 2. University of Belgrade – Faculty of Biology, Center for Human Molecular Genetics,
564 Studentski trg 16, 11000 Belgrade, Serbia
- 565 3. Department of Biotechnology, University of Verona, Strada le Grazie 15, 37134
566 Verona, Italy
- 567 4. Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen
568 Mary University London, London, UK
- 569 5. MGZ – Medical Genetics Center, Munich, Germany
- 570 6. Medizinische Klinik und Poliklinik IV, Campus Innenstadt, Klinikum der Universität
571 München, Munich, Germany
- 572 7. Friedrich-Baur-Institute, Department of Neurology, Klinikum der Universität, Ludwig-
573 Maximilians-Universität, Munich, Germany
- 574 8. Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, INSERM
575 U1127, CNRS UMR7225, APHP, Paris, France
- 576 9. Department of Medical Genetics, University of British Columbia, BC Children's and
577 Women's Hospital
- 578 10. UCL Institute of Ophthalmology, London, UK
- 579 11. Moorfields Eye Hospital NHS Foundation Trust, London, UK
- 580 12. Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala
- 581 13. SciLifeLab, Uppsala University, Sweden
- 582 14. Institut Curie, PSL University, CurieCoreTech Next-Generation Sequencing (ICGex),
583 75005 Paris, France
- 584 15. Institut Curie, INSERM U1331, Mines ParisTech, PSL University, CurieCoreTech
585 Bioinformatics (CUBIC), 75005 Paris, France
- 586 16. Center for Genomic Medicine, Department of Neurology, Massachusetts General
587 Hospital, Boston, MA, USA
- 588 17. Department of Neurology, Harvard Medical School, Boston, MA, USA
- 589 18. Medical and Population Genetics Program, The Broad Institute of MIT and Harvard,
590 Cambridge, MA, USA
- 591 19. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114,
592 USA
- 593 20. Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA
- 594 21. Department of Pathology, Harvard Medical School, Boston, MA 02115, USA
- 595 22. Department of Biomedicine and Prevention, University of Rome "Tor Vergata",
596 Rome, Italy
- 597 23. St George's, University of London, London, UK
- 598 24. Huntington's Disease Centre, Department of Neurodegenerative Disease, Queen
599 Square Institute of Neurology, UCL, London, UK
- 600 25. Department of Neuromuscular Diseases, Queen Square Institute of Neurology, UCL,
601 London, UK
- 602 26. Sorbonne Université, Paris Brain Institute – Institut du Cerveau, ICM, Inserm U1127,
603 CNRS UMR7225, APHP – Hôpital Pitié-Salpêtrière, Paris 75013, France
- 604 27. Paris Brain Institute-Institut du Cerveau-ICM, Sorbonne Université, Inserm U1127,
605 CNRS UMR 7225, AP-HP-Hôpital Pitié-Salpêtrière, Paris, France
- 606 28. Reference Centre for Rare or Early-Onset Dementias, IM2A, Département de
607 Neurologie, AP-HP-Hôpital Pitié-Salpêtrière, Paris, France

29. Centre de Référence des Maladies Neuromusculaires Nord-Est/Île-de-France, Institut de Myologie, GHU Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Paris, France.
30. Laboratory of Cell and Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA
31. Department of Pathology, Stanford University School of Medicine, 300 Pasteur Dr., Stanford, CA 94305-5324, USA
32. UK Dementia Research Institute at Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK
33. Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Maindy Road, Cardiff CF24 4HQ, UK

Figure legend

Fig. 1. Overview of long-read sequencing (LRS) workflows for the analysis of short tandem repeat (STR) expansions. Schematic representation of the main experimental approaches. Repeat expansions can be analyzed using whole-genome sequencing (WGS) or by applying enrichment strategies prior to sequencing, including long-range PCR (LR-PCR), hybridization capture, droplet-based methods, CRISPR–Cas9 enrichment, or adaptive sampling. These enable either genome-wide or targeted characterization of repeat expansions using LRS technologies, namely Pacific Biosciences (PacBio) and Oxford nanopore technology (ONT). PacBio instruments require more physical space than ONT devices, an important consideration for laboratory implementation. Downstream bioinformatic pipelines can be tailored to extract information on repeat size, motif composition, interruptions, CpG methylation status and somatic mosaicism. These features are essential for understanding disease pathogenesis and are increasingly relevant for diagnosis, prognosis and therapeutic development. Challenges for clinical implementation of LRS are highlighted and further detailed in Box 3.

Tables

Table 1 Overview over LRS approaches for repeat-expansion sequencing

Method	LRS technology	Amplification-free	DNA input	DNA fragmentation	Sample multiplexing	Multiple repeat loci analysis	On-target coverage ^a	MaximumRepeat Length	Mosaicism	Methylation
WGS	ONT ¹⁷⁹	✓	1000 ng	Recommended	Not Recommended	✓	10-30x	>20kbp	Limited by coverage	✓
	PacBio ¹⁸⁰	✓	500 ng	✓	✓	✓	10-30x	~20kbp	Limited by coverage	✓
Long range PCR	ONT ¹⁸¹	-	5 ng	-	✓	✓	100-1000x	~5kbp	Limited by PCR	-
	PacBio ¹⁴⁸	-	5 ng	-	✓	✓	100 - 1000x	~5kbp	Limited by PCR	-
Hybridization	ONT ¹⁸²	-	200-500 ng	✓	✓	✓	10-1000x	~5kbp	Limited by PCR	-
	PacBio ¹⁸³	-	200-2000 ng	✓	✓	✓	75-200x	~5kbp	Limited by PCR	-
Droplets	ONT ¹⁷⁰	-	10 ng	-	Not Recommended	✓	50-330x	~5kbp	Limited by PCR	-
	PacBio ¹⁶⁹	-	4 ng	-	Not Recommended	✓	20-300x	~5kbp	Limited by PCR	-

Cas9	ONT ⁹⁴	✓	2000-5000 ng	-	Not Recommended	✓	50-2000x	>20kbp	✓ ^b	✓
	PacBio ¹⁸⁴	✓	1000-4000 ng	-	✓	✓	50-1000x	~20kbp	✓ ^b	✓
Adaptive Sampling	ONT ⁶²	✓	2000 ng	✓	✓	✓	20-100x	~5kbp	Limited by coverage	✓

^aReported coverage refers to the total amount of sequencing data obtained for both alleles. Values may vary depending on the number of targets, multiplexing strategy, sequencing run type, and flow-cell throughput, and can also be influenced by repeat size and motif composition.

^bSomatic mosaicism analysis can be limited by sequencing coverage, depending on the repeat expansion.

Tool	Platform	Novel loci detection*	Repeat size	Repeat interruptions	Read-level evidence	Methylation	Output file	Mosaicism	Haplotype resolution	Publication year	Source code	Paper	Actively maintained	Method	Version
STRkit	ONT/PacBio	N	Y	Y	Y	N	TSV/JSON/VCF/visualisation	Y	Y	2026	https://github.com/davidloughheed/strkit	https://genome.cshlp.org/content/36/3/578	Y	Read-level STR genotyping using Gaussian mixture model (GMM) with optional SNV-informed phasing	0.2
Medaka Tandem	ONT	N	Y	Y	Y	N	VCF	Y	Y	2026	https://github.com/nanoporetech/medaka/releases	https://www.biorxiv.org/content/10.64898/2026.02.28.708646v1.full.pdf	Y	Consensus-based TR genotyping with haplotype phasing	2.2
TRsv	ONT/PacBio	N	Y	Y	Y	N	VCF	N	Y	2025	https://github.com/stat-lab/TRsv	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-025-03718-z	Y	Genotypes TR copy number and structural variants by aggregating indels within TRs, clustering alleles based on read support and motif content, and jointly calling variants across samples	1.1
ATaRVA	ONT/PacBio	N	N	Y	Y	Y	VCF	Y	Y	2025	https://github.com/Sowpatilab/ATaRVA	https://www.biorxiv.org/content/10.1101/2025.05.13.653434v1	Y	Genotypes TRs by realigning reads to flanks, clustering them into haplotypes, generating consensus alleles, and decomposing motifs	0.6
HMMSTR	ONT/PacBio	N	Y	N	Y	N	TSV	Y	Y	2024	https://github.com/Boyle-Lab/HMMSTR	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-03010-y	last updated 2 years ago	Read- and sample-level TR copy-number calling from long reads using HMM	1.0

LongTR	ONT/ PacBio	N	Y	Y	Y	N	VCF	Y	Y	2024	https://github.com/gymrek-lab/longtr	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03319-2	Y	Genotypes TRs from long reads using HMM and haplotype clustering	1.
otter and TREAT	ONT/ PacBio	N	Y	Y	Y	N	VCF/ plot	Y	Y	2024	https://github.com/holstegelab/treat	https://pubmed.ncbi.nlm.nih.gov/39406499/	2 years ago	Uses local read assembly, pairwise distance-based clustering, and pseudo-partial order alignment to reconstruct allele sequences, characterize repeat motifs, and quantify copy numbers from long-read data	1.0
STRdust	ONT/ PacBio	N	Y	Y	Y	N	VCF	Y	Y	2024	https://github.com/wdecoster/STRdust	https://genome.cshlp.org/content/early/2024/08/15/gr.279265.124	Y	Aligns reads to a repeat-free reference, extracts insertions, builds consensus, and clusters reads to call repeat alleles	0.1
TRGT	PacBio	N	Y	Y	Y	Y	VCF/ plot	Y	Y	2024	https://github.com/PacificBiosciences/trgt	https://www.nature.com/articles/s41587-023-02057-3	Y	Genotypes TRs by clustering spanning HiFi reads, deriving consensus allele sequences, annotating simple or complex repeats (using graph or HMM methods), and reporting allele sequences with methylation levels	5.0
NanoRepeat	ONT/ PacBio	N	Y	N	Y	N	VCF/ plot / BED	N	Y	2023	https://github.com/WGLab/NanoRepeat	linkinghub.elsevier.com/retrieve/pii/S266624772200063X	Y	Read-level TR genotyping and GMM clustering for allele assignment	1.8

Vamos	ONT/ PacBio	Y	N	Y	N	N	VCF	Y	Y	2023	https://github.com/ChaissonLab/vamos	https://pmc.ncbi.nlm.nih.gov/articles/PMC10373352/	Y	Consensus-based VNTR annotation for haplotype-resolved alleles.	3.0
WarpSTR**	ONT	N	Y	Y	Y	N	TSV	Y	N	2023	https://github.com/fmfi-compbio/warpstr	https://academic.oup.com/bioinformatics/article/39/6/btad388/7199589	3 years ago	Read-level STR detection from raw nanopore signals using dynamic time warping	0.3
CharONT2	ONT	N	N	Y	N	N	Table/ distribution plots	Y	Y	2022	https://github.com/MaestSi/CharONT2	https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.743230/full	1 year	Consensus-based TR genotyping with read clustering to resolve allele-specific sequences	1.1
MosaicViewer	ONT/ PacBio	N	Y	Y	Y	N	VCF	Y	N	2021	https://github.com/MaestSi/MosaicViewer_FM_R1	https://elifesciences.org/article/s80229	3 years ago	Read-level TR annotation and visualization pipeline for mosaic expansions using simplified repeat representations	1.0
Straglr	ONT/ PacBio	N	Y	N	Y	N	TSV/BED/VCF	Y	Partial	2021	https://github.com/bcgsc/straglr	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02447-3	Y	TR genotyping and <i>de novo</i> expansion detection using insertion detection and GMM clustering to infer alleles	1.5
TRiCoLOR	ONT/ PacBio	Y	Y	N	Y	N	VCF	Y	N	2020	https://github.com/davidebolio1993/Tricolor	https://academic.oup.com/giga-science/article/9/10/giaa101/5	N	<i>De novo</i> TR detection using Shannon entropy and haplotype-specific consensus sequence profiling	On on vers exi

											RiCoLO R	918863?search result=1			
Nanosatellite**	ONT	N	Y	N	Y	N	plot	Y	N	2019	https://github.com/arnederock/NanoSatellite	https://pmc.ncbi.nlm.nih.gov/articles/PMC6857246/	N	Signal-level TR analysis from raw Oxford Nanopore squiggle data using dynamic time warping	0.
STRique*	ONT	N	Y	N	Y	Y	VCF	N	N	2019	https://github.com/giesselmann/STRique	https://www.nature.com/articles/s41587-019-0293-x	N	HMM-based STR genotyping and CpG methylation analysis from nanopore reads at single-molecule resolution	0.4
Tandem Genotypes	ONT/ PacBio	N	Y	N	Y	N	bed	Y	Y	2019	https://github.com/mcfrith/tandem-genotypes	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1667-6	2 years ago	TR genotyping by estimating repeat copy number changes across reads and inferring alleles	1.9
RepeatHMM	ONT/ PacBio	Y	Y	N	Y	N	BED	Y	Partial	2017	https://github.com/giesselmann/RepeatHMM	https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0456-7	N	Read-level TR genotyping using Hidden Markov Model (HMM) with GMM peak calling to infer repeat counts	2.0
pacmonSTR	ONT/ PacBio	N	Y	N	Y	N	TSV	Y	Y	2014	https://github.com/oscarlr-trs/PacMonSTR	https://academic.oup.com/bioinformatics/article/30/24/3491/2422157	N	TR genotyping using GMM clustering to call alleles	On on vers exi

Table 2. Summary of analysis tools used for long-read sequencing.

Y and N indicate Yes and No, respectively. Tools marked with two asterisks (**) are not compatible with current ONT chemistry from R10 nanopore generation, version 4.1 chemistry (R10.4.1) and are effectively depreciated for new diagnostic pipelines. *Novel loci detection: tools restricted to a variant catalogue are denoted as not able to detect novi loci, not in the reference genome of short tandem repeats. TR = tandem repeat.

Text Boxes

Box 1- Methodological approaches for repeat expansion analysis

Traditional approaches-PCR-based methods and Southern blots

Conventional PCR-based assays, using locus-specific flanking primers, often fail to amplify medium repeat expansions (>1 kb) and miss pathogenic alleles. In contrast, repeat-primed PCR (RP-PCR) offers higher sensitivity, using a combination of a locus-specific and a repeat-annealing primer, generating heterogeneous set of amplicons and a electropherogram pattern indicative of repeat expansions¹⁵². Depending on assay design, RP-PCR can detect interruptions at the 5' or 3' ends of a repeat expansion, but it does not provide the sequence or allele-specific architecture of interrupting motifs. Heavily interrupted repeat expansions may lead to false-negatives^{153,154}. To identify non-reference repeat expansions, RP-PCR assays require prior knowledge of the repeat sequence, limiting their utility given the diversity of pathogenic motif compositions. Despite these limitations, RP-PCR is the most commonly used method in the diagnostics of REDs. However, additional steps are required to assess methylation and mosaicism, increasing turnaround time and cost^{13,124}. PCR-based assays are also prone to GC and amplification biases, often favoring shorter alleles in heterozygous samples. In contrast, Southern blotting can size large REs (>1kb) but lacks precision for resolving close-size mosaics and cannot detect motif interruptions or other sequence variations. This process remains low-throughput and labor-intensive for routine diagnostics²⁷. Nevertheless, these approaches remain the current clinical standard for analyzing repeat expansions but could be replaced in the future by long-read sequencing technologies, which would improve the resolution and characterization of expanded alleles.

Optical genome mapping

This approach analyzes long DNA fragments that are labeled at specific sites across the genome. Copy number and structural variants are identified based on altered inter-label distances and their orientations or patterns along the DNA molecule¹⁵⁵. Optical genome mapping (OGM) can also detect repeat expansions larger than ~300-500 bp and up to several kilobases, as well as reveal extreme events of repeat instability and mosaicism¹⁵⁶. OGM is a robust complementary technology that enables the sizing of large expansions and the capture of broad structural context without loading or library preparation biases. However, repeat-expansion sizes are inferred indirectly from increased distances between two labels, which introduces ambiguity, as other structural variants in the same region may produce similar signatures¹⁵⁶. Nevertheless, multiple studies report strong concordance between LRS and OGM, supporting their complementary use for sizing large expansions and capturing somatic variability²⁴. As this technology captures repeat length but not sequence-level information, it cannot fully distinguish pathogenic from non-pathogenic repeat extensions¹⁰⁴.

SRS approaches

Targeted panels, exome or genome sequencing enable screening of multiple STR loci for repeat expansions¹⁵⁷. State-of-the-art SRS analysis algorithms integrate evidence from different read types: fully spanning reads, flanking reads containing both repeat and adjacent sequences, and in-repeat reads composed entirely of repeat motifs. These tools can detect repeat expansions exceeding short-read lengths (typically ~150–250 bp for reads and ~350–650 bp for DNA fragments). However, they provide only approximate size estimates and tend to underestimate large alleles, which can significantly impact disease diagnosis, as pathogenic thresholds may be underestimated or missed and could be more accurately defined using

long-read sequencing^{158,159}. Both PCR-based and PCR-free workflows are typically only reliable for the sizing of repeats that are shorter than the read length, but they provide no information on methylation and are prone to GC bias, leading to false-negatives in GC-rich loci. Mosaicism can be detected in some loci (e.g., *HTT*), although detection sensitivity depends on sequencing chemistry, depth, read length, STR zygosity, and the degree of mosaicism¹⁶⁰. For sequence-composition analysis, different tools can identify non-reference repeats or reveal interrupting motifs within fully spanned repeats or at the 5' and 3' edges when expansions exceed read or fragment length^{30,157}.

Box 2: LRS approaches for repeat expansions

Sample preparation and DNA isolation

LRS performance varies across samples and tissues and can be influenced by DNA quality¹¹⁶. Accurate fluorometric quantification and purity assessment are recommended¹⁶¹, while organic extractions (e.g., phenol–chloroform) should be avoided, as they typically result in poor sequencing performance¹⁶². Both ONT and PacBio provide recommended DNA extraction protocols and kits for the elimination of short fragments, typically enriching for DNA fragments above ~10–20 kb. Sample source significantly influences DNA integrity and yield; blood and saliva are generally well-suited to high molecular weight (HMW) extraction, whereas challenging tissues such as brain may require specialized protocols or isotachophoresis-based approaches to achieve fragment sizes compatible with long-read sequencing. For short or medium repeat expansion (<5 kb)¹¹⁸, standard column- or bead-based extraction methods generally produce fragments of adequate size. However, for long (5–20kb) and ultra-long repeats (>20 kb), HMW DNA is preferred^{161,163}, but not essential if sufficient reads fully spanning the repeat expansion are obtained. HMW DNA can be obtained using gravity columns, nanoball or magnetic-disk systems, which are now all widely accessible. Assessing fragment length is critical to obtain reads that span the repeat. When large expansions (>5kb) are expected, pulse-field electrophoresis instruments (e.g., FemtoPulse (<https://www.agilent.com/en/product/automated-electrophoresis/femto-pulse-systems/femto-pulse-system>), LightBench Discover (<https://yourgenehealth.com/products/ranger-technology-instrumentation/lightbench-discover/>), Contour-clamped Homogeneous Electric Field mapper¹⁶⁴) allow accurate sizing. Although constant-field capillary electrophoresis systems (e.g., Fragment Analyzer, Tape Station) are limited to sizing fragments below ~50kb, they are

useful to monitor DNA integrity and detect the presence of short fragments that compete with long molecules during sequencing. Removing shorter fragments through selective precipitation or automated fractionation maximizes long-read yield, although recovery may decrease¹⁶⁵. Commercial kits and bead–buffer protocols are available¹⁶⁶.

Sequencing approaches

Established targeted sequencing methods, such as LR-PCR¹⁶⁷ and hybridization capture¹⁶⁸, can be combined with LRS to enable cost-efficient analysis of repeat regions across large cohorts. For low-input samples, alternative methods, such as droplet-based enrichment^{169,170} may be considered. However, these approaches require amplification that can introduce biases, leading to misestimation of mosaicism, allelic dropout and reduced read lengths. Moreover, methylation information will be lost. Amplification-free methods overcome these limitations but require >500 ng DNA. They allow accurate sizing of long or difficult-to-amplify repeats, determination of motif composition, methylation profiling, and detailed assessment of mosaicism^{28,171}. Whole-genome sequencing (WGS) represents the most straightforward amplification-free sequencing approach. Repeat sequences can be identified even from low-coverage WGS data⁸⁷. However, an average genome-wide sequencing depth of ~20–30x is generally required to achieve sufficient coverage of most repeat loci within a sample, and even at this depth, somatic mosaicism cannot be accurately quantified. Alternatively, amplification-free enrichment methods allow higher on-target coverage at lower cost¹⁷². Cas9-based enrichment^{173,174}, provides the highest efficiency and supports repeat-level somatic instability studies. Adaptive sampling, the software-controlled enrichment approach for ONT sequencing^{36,175,176}, is a highly flexible alternative, as it allows modification of target regions without requiring changes to library preparation or re-design of the target sequencing assay.

However, adaptive sampling typically requires DNA fragmentation (commonly 8–10 kb, though ultra-long options exist) and yields moderate on-target coverage, which may limit its utility¹⁷⁷.

Multiplexing

Both platforms support medium- to high-throughput instruments and sequencing cells (50–120 Gb), enabling simultaneous analysis of multiple loci and samples and reducing costs at the diagnostic level¹⁷⁸. The multiplexing design should take into consideration that ~10x coverage per expanded allele is typically adequate for reliable repeat characterization³⁵. Lower coverage may still be acceptable for technically challenging or ultra-long repeats, while higher depth is recommended when mosaicism is present. Multiplexing of LR-PCR amplicons for repeat analysis is straightforward on both platforms^{33,148}. For Cas9-enrichment, PacBio provides a robust workflow for 38 loci and up to 48 samples. ONT-based Cas9-multiplexing remains limited for large panels¹¹¹ but works for single-gene enrichment⁴¹, albeit with a higher percentage of off-target reads. Adaptive sampling has recently enabled multiplexing of 4–5 samples for targeted chromosomal regions, leading to the discovery of a novel repeat locus in SCA4⁶². LR-WGS can also be multiplexed. However, per-sample coverage may remain limiting unless higher-capacity flow cells become available. ONT flow cells can be reused, while PacBio has announced the possibility of reusing its sequencing cells by summer 2026, enabled by the new SPRQ-Nx chemistry (<https://www.pacb.com/blog/how-sprq-nx-enables-affordable-long-read-whole-genome-sequencing-without-added-complexity/>).

Box 3: Challenges in the clinical implementation of LRS for repeat-expansion analysis

Technical or sequencing limitations

- Limited coverage and strong dependence on DNA quality/quantity and extraction methods for ultra-long repeat expansions (>20 kb)
- Reduced sequencing performance in specific repeat contexts (e.g., *RFC1*, *FXN*)
- Requirement for high coverage in loci with pronounced somatic mosaicism
- Challenges in analyzing low-input samples
- Limited availability of commercial solutions and absence of *in vitro* diagnostic-certified kits
- Constraints on multiplexing multiple samples per run (ONT)
- High per-sample sequencing costs and substantial platform setup investment (PacBio)

Bioinformatics and data interpretation

- Lack of standardized, clinically validated bioinformatics pipelines for RE analysis
- Absence of reference samples capturing the full complexity of repeat expansions beyond size and methylation
- Sequencing errors and dependence on base-calling algorithms (particularly ONT), affecting repeat sizing and motif resolution
- Platform- and pipeline-dependent variability
- Absence of comprehensive benchmarking across tools, platforms, and chemistries
- Limited integration of analysis workflows (QC, phasing, annotation, and reporting)
- Dependence on incomplete and biased reference genomes, limiting detection of non-reference or structurally complex repeat expansions

Clinical interpretation and reporting

- Lack of clinically curated reference databases integrating locus definitions, pathogenic thresholds, population distributions, and genotype–phenotype correlations
- Need for standardized reporting formats and harmonized interpretation guidelines
- Requirement for inter-laboratory validation and robust quality assurance frameworks
- Need for clinician training to interpret and report complex repeat expansions (size, motif, interruptions, and clinical correlations)

REFERENCES

1. Wright, S. E. & Todd, P. K. Native functions of short tandem repeats. *eLife* **12**, e84043.
2. Weisburd, B. *et al.* Defining a tandem repeat catalog and variation clusters for genome-wide analyses. Preprint at <https://doi.org/doi:10.1101/2024.10.04.615514> (2026).
3. Subramanian, S., Mishra, R. K. & Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**, R13 (2003).
4. Shi, Y. *et al.* Characterization of genome-wide STR variation in 6487 human genomes. *Nat Commun* **14**, 2092 (2023).
5. Porubsky, D. *et al.* Human de novo mutation rates from a four-generation pedigree reference. *Nature* **643**, 427–436 (2025).
6. Steely, C. J., Watkins, W. S., Baird, L. & Jorde, L. B. The mutational dynamics of short tandem repeats in large, multigenerational families. *Genome Biol* **23**, 253 (2022).
7. Kristmundsdottir, S. *et al.* Sequence variants affecting the genome-wide rate of germline microsatellite mutations. *Nat Commun* **14**, 3855 (2023).
8. Chen, Z. *et al.* Repeat expansion disorders. *Pract Neurol* **25**, 204–216 (2025).
9. Ibañez, K. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol* **21**, 234–245 (2022).
10. Li, Y.-C., Korol, A. B., Fahima, T. & Nevo, E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* **21**, 991–1007 (2004).
11. Cortese, A. *et al.* Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet* **51**, 649–658 (2019).
12. Beecroft, S. J. *et al.* A Māori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. *Brain* **143**, 2673–2680 (2020).
13. Monckton, D. G., Wong, L. J., Ashizawa, T. & Caskey, C. T. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum Mol Genet* **4**, 1–8 (1995).

14. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Genetic modifiers of somatic expansion and clinical phenotypes in Huntington's disease highlight shared and tissue-specific effects. *Nat Genet* **57**, 1426–1436 (2025).
15. Kacher, R. *et al.* CAG repeat mosaicism is gene specific in spinocerebellar ataxias. *Am J Hum Genet* **111**, 913–926 (2024).
16. Morales, F. *et al.* Longitudinal increases in somatic mosaicism of the expanded CTG repeat in myotonic dystrophy type 1 are associated with variation in age-at-onset. *Hum Mol Genet* **29**, 2496–2507 (2020).
17. De Antonio, M. *et al.* Unravelling the myotonic dystrophy type 1 clinical spectrum: A systematic registry-based study with implications for disease classification. *Revue Neurologique* **172**, 572–580 (2016).
18. Wright, G. E. B. *et al.* Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. *Lancet Neurol* **19**, 930–939 (2020).
19. Santoro, M., Masciullo, M., Silvestri, G., Novelli, G. & Botta, A. Myotonic dystrophy type 1: role of CCG, CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular diagnosis. *Clin Genet* **92**, 355–364 (2017).
20. Peric, S., Pesovic, J., Savic-Pavicevic, D., Rakocovic Stojanovic, V. & Meola, G. Molecular and Clinical Implications of Variant Repeats in Myotonic Dystrophy Type 1. *IJMS* **23**, 354 (2021).
21. Perez, B. A. *et al.* CCG•CGG interruptions in high-penetrance SCA8 families increase RAN translation and protein toxicity. *EMBO Mol Med* **13**, e14095 (2021).
22. Morato Torres, C. A. *et al.* ATTCT and ATTCC repeat expansions in the ATXN10 gene affect disease penetrance of spinocerebellar ataxia type 10. *Human Genetics and Genomics Advances* **3**, 100137 (2022).
23. McFarland, K. N. *et al.* Repeat interruptions in spinocerebellar ataxia type 10 expansions are strongly associated with epileptic seizures. *Neurogenetics* **15**, 59–64 (2014).
24. Sedov, K. *et al.* Novel ATXN10 Repeat Motif Patterns in Peruvian Families Modify Disease Onset. *Neurol Genet* **11**, e200316 (2025).
25. Schüle, B. *et al.* Parkinson's disease associated with pure ATXN10 repeat expansion. *npj Parkinson's Disease* **3**, 27 (2017).
26. Barbé, L. & Finkbeiner, S. Genetic and Epigenetic Interplay Define Disease Onset and Severity in Repeat Diseases. *Front Aging Neurosci* **14**, 750629 (2022).
27. Leitão, E., Schröder, C. & Depienne, C. Identification and characterization of repeat expansions in neurological disorders: Methodologies, tools, and strategies. *Revue Neurologique* **180**, 383–392 (2024).
28. Stevanovski, I. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv* **8**, 5386 (2022).
29. Taylor, A. S. *et al.* Repeat Detector: versatile sizing of expanded tandem repeats and identification of interrupted alleles from targeted DNA sequencing. *NAR Genomics and Bioinformatics* **4**, lqac089 (2022).
30. Dolzhenko, E. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* **42**, 1606–1614 (2024).
31. Treffer, R. & Deckert, V. Recent advances in single-molecule sequencing. *Current Opinion in Biotechnology* **21**, 4–11 (2010).
32. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021).
33. Mohren, L. *et al.* Identification and characterisation of pathogenic and non-pathogenic FGF14 repeat expansions. *Nat Commun* **15**, 7665 (2024).
34. Miyatake, S. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *npj Genom. Med* **7**, 62 (2022).

35. Höps, W. *et al.* HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *The American Journal of Human Genetics* **112**, 450–456 (2025).
36. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* **39**, 442–450 (2021).
37. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**, 597–614 (2020).
38. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
39. Park, G., An, H., Luo, H. & Park, J. NanoMnT: an STR analysis tool for Oxford Nanopore sequencing data driven by a comprehensive analysis of error profile in STR regions. *GigaScience* **14**, giaf013 (2025).
40. Udine, E. *et al.* Targeted long-read sequencing to quantify methylation of the C9orf72 repeat expansion. *Mol Neurodegener* **19**, 99 (2024).
41. Alfano, M. *et al.* Characterization of full-length CNBP expanded alleles in myotonic dystrophy type 2 patients by Cas9-mediated enrichment and nanopore sequencing. *eLife* **11**, e80229 (2022).
42. Currò, R. *et al.* RFC1 expansions are a common cause of idiopathic sensory neuropathy. *Brain* **144**, 1542–1550 (2021).
43. Hård, J. *et al.* Long-read whole-genome analysis of human single cells. *Nat Commun* **14**, 5164 (2023).
44. Fan, X. *et al.* SMOOTH-seq: single-cell genome sequencing of human cells on a third-generation sequencing platform. *Genome Biol* **22**, 195 (2021).
45. Handsaker, R. E. *et al.* Long somatic DNA-repeat expansion drives neurodegeneration in Huntington’s disease. *Cell* **188**, 623–639.e19 (2025).
46. Kumari, P. *et al.* Advances in long-read single-cell transcriptomics. *Hum. Genet.* **143**, 1005–1020 (2024).
47. Wen, L. & Tang, F. Single-cell omics sequencing technologies: the long-read generation. *Trends in Genetics* S0168952525001969 (2025) doi:10.1016/j.tig.2025.07.012.
48. Cogan, G., Daida, K., Blauwendraat, C., Billingsley, K. & Brice, A. Exploration of Neurodegenerative Diseases Using Long-Read Sequencing and Optical Genome Mapping Technologies. *Movement Disorders* **40**, 996–1008 (2025).
49. Su, Y. *et al.* Deciphering Neurodegenerative Diseases Using Long-Read Sequencing. *Neurology* **97**, 423–433 (2021).
50. Tanudisastro, H. A. Sequencing and characterizing short tandem repeats in the human genome. *Nat Rev Genet* **25**, 460–475 (2024).
51. Ishiura, H. *et al.* Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat Genet* **51**, 1222–1232 (2019).
52. Deng, J. *et al.* Expansion of GGC Repeat in GIPC1 Is Associated with Oculopharyngodistal Myopathy. *Am J Hum Genet* **106**, 793–804 (2020).
53. Yu, J. *et al.* The GGC repeat expansion in NOTCH2NLC is associated with oculopharyngodistal myopathy type 3. *Brain* **144**, 1819–1832 (2021).
54. Yu, J. *et al.* The CGG repeat expansion in RILPL1 is associated with oculopharyngodistal myopathy type 4. *Am J Hum Genet* **109**, 533–541 (2022).
55. Cortese, A. *et al.* A CCG expansion in ABCD3 causes oculopharyngodistal myopathy in individuals of European ancestry. *Nat. Commun.* **15**, 6327 (2024).
56. Shi, Y. *et al.* CGG repeat expansion in LOC642361/NUTM2B-AS1 typically presents as oculopharyngodistal myopathy. *J Genet Genomics* **51**, 184–196 (2024).
57. Pellerin, D. *et al.* Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med* **388**, 128–141 (2023).

58. Rafehi, H. *et al.* An intronic GAA repeat expansion in FGF14 causes the autosomal-dominant adult-onset ataxia SCA50/ATX-FGF14. *Am J Hum Genet* **110**, 105–119 (2023).
59. Figueroa, K. P. *et al.* A GGC-repeat expansion in ZFH3 encoding polyglycine causes spinocerebellar ataxia type 4 and impairs autophagy. *Nat Genet* **56**, 1080–1089 (2024).
60. Wallenius, J. *et al.* Exonic trinucleotide repeat expansions in ZFH3 cause spinocerebellar ataxia type 4: A poly-glycine disease. *The American Journal of Human Genetics* **111**, 82–95 (2024).
61. Paucar, M. *et al.* Spinocerebellar ataxia type 4 is caused by a GGC expansion in the ZFH3 gene and is associated with prominent dysautonomia and motor neuron signs. *J Intern Med* **296**, 234–248 (2024).
62. Chen, Z. *et al.* Adaptive long-read sequencing reveals GGC repeat expansion in ZFH3 associated with spinocerebellar ataxia type 4. *Mov. Disord.* **39**, 486–497 (2024).
63. Ishiura, H. *et al.* Expansions of intronic TTTC and TTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet* **50**, 581–590 (2018).
64. Corbett, M. A. *et al.* Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. *Nat Commun* **10**, 4920 (2019).
65. Florian, R. T. *et al.* Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. *Nat Commun* **10**, 4919 (2019).
66. Yeetong, P. *et al.* Pentanucleotide Repeat Insertions in *RAI1* Cause Benign Adult Familial Myoclonic Epilepsy Type 8. *Movement Disorders* **39**, 164–172 (2024).
67. Yeetong, P. *et al.* TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4. *Brain* **142**, 3360–3366 (2019).
68. Ma, D. *et al.* Association of *NOTCH2NLC* Repeat Expansions With Parkinson Disease. *JAMA Neurol* **77**, 1559 (2020).
69. Akçimen, F. *et al.* Long-read sequencing identifies FGF14 repeat expansions in Parkinson’s disease. *Brain* **In press**, (2025).
70. Bunting, E. L. *et al.* Antisense oligonucleotide-mediated MSH3 suppression reduces somatic CAG repeat expansion in Huntington’s disease iPSC-derived striatal neurons. *Sci Transl Med* **17**, eadn4600 (2025).
71. Centofanti, F. *et al.* The novel (TCTG)_n motif in CNBP expanded alleles: composition, dynamics and genotype–phenotype correlation in Myotonic dystrophy type 2 (DM2). *Hum Genomics* <https://doi.org/10.1186/s40246-026-00954-7> (2026) doi:10.1186/s40246-026-00954-7.
72. Rasmussen, A. *et al.* High Resolution Analysis of DMPK Hypermethylation and Repeat Interruptions in Myotonic Dystrophy Type 1. *Genes (Basel)* **13**, (2022).
73. Aston, A. N. & Dion, V. Interruptions impact clinical features of repeat expansion diseases, but how are they gained and lost? *Trends in Genetics* **41**, 1056–1067 (2025).
74. Lameiras, S. *et al.* CCG interruptions are unstable and hypermethylated in DM1 patients. *bioRxiv* 2024.10.10.617614 (2024) doi:10.1101/2024.10.10.617614.
75. Xu, I. R. L., Pellerin, D. & Zuchner, S. The Genetic and Evolutionary Landscape of Pentanucleotide Tandem Repeats in Human Genomes. *bioRxiv* <https://doi.org/doi.org/10.1101/2025.05.27.656002> (2025) doi:doi.org/10.1101/2025.05.27.656002.
76. Liang, Q. *et al.* Comprehensive Analysis of Fragile X Syndrome: Full Characterization of the FMR1 Locus by Long-Read Sequencing. *Clin Chem* **68**, 1529–1540 (2022).
77. Erdmann, H. Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain* **146**, 1831–1843 (2023).
78. Jiraanont, P. *et al.* Size and methylation mosaicism in males with Fragile X syndrome. *Expert Rev Mol Diagn* **17**, 1023–1032 (2017).
79. Elhawary, N. A. *et al.* Phenotypic variability to medication management: an update on

- fragile X syndrome. *Hum Genomics* **17**, 60 (2023).
80. Rajan-Babu, I.-S. & Chong, S. S. Molecular Correlates and Recent Advancements in the Diagnosis and Screening of FMR1-Related Disorders. *Genes (Basel)* **7**, (2016).
 81. Dischler, A. *et al.* Long-read sequencing reveals extensive FMR1 somatic mosaicism in Fragile-X associated tremor/ataxia syndrome in human brain. *bioRxiv* 2025.10.07.678589 (2025) doi:10.1101/2025.10.07.678589.
 82. Hunter, J. E., Berry-Kravis, E., Hipp, H. & Todd, P. K. GeneReviews®. in *FMR1 Disorders* (eds. Adam, Margaret P., Mirzaa, Ghayda M. & Pagon, Roberta A) (University of Washington, Seattle (WA), 1998).
 83. Sitzmann, A. F. Rare FMR1 gene mutations causing fragile X syndrome: A review. *Am J Med Genet A* **176**, 11–18 (2018).
 84. Ronco, R. Truncating Variants in RFC1 in Cerebellar Ataxia, Neuropathy, and Vestibular Areflexia Syndrome. *Neurology* **100**, 543–554 (2023).
 85. Hou, F. Evaluating the clinical utility of a long-read sequencing-based approach in genetic testing of fragile-X syndrome. *Clinica Chimica Acta* **551**, 117614 (2023).
 86. Pellerin, D. *et al.* A common flanking variant is associated with enhanced stability of the FGF14-SCA27B repeat locus. *Nat Genet* **56**, 1366–1370 (2024).
 87. Danzi, M. C. *et al.* Detailed tandem repeat allele profiling in 1,027 long-read genomes reveals genome-wide patterns of pathogenicity. Preprint at <https://doi.org/10.1101/2025.01.06.631535> (2025).
 88. Benarroch, L. Comparative Analysis of CRISPR/Cas9-targeted Nanopore Sequencing Approaches in Repeat Expansion Disorders. in *Genomics Proteomics Bioinformatics DOI* (2025). doi:10.1093/gpbjnl/qzaf094.
 89. Pagès-Gallego, M. & De Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol* **24**, 71 (2023).
 90. Baid, G. *et al.* DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol* <https://doi.org/10.1038/s41587-022-01435-7> (2022) doi:10.1038/s41587-022-01435-7.
 91. Sitarčík, J. WarpSTR: determining tandem repeat lengths using raw nanopore signals. *Bioinformatics* **39**, 388 (2023).
 92. Roeck, A. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biology* **20**, 239 (2019).
 93. Fang, L. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biology* **23**, 108 (2022).
 94. Giesselmann, P. *et al.* Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol* **37**, 1478–1481 (2019).
 95. Ziaei Jam, H. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biology* **25**, 176 (2024).
 96. Chiu, R. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol* **22**, 224 (2021).
 97. Mitsuhashi, S. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biology* **20**, 58 (2019).
 98. Deynze, K. Enhanced detection and genotyping of disease-associated tandem repeats using HMMSTR and targeted long-read sequencing. *Nucleic Acids Res* **53**, 1202 (2025).
 99. Chiu, R., Rajan-Babu, I.-S., Friedman, J. M. & Birol, I. A comprehensive tandem repeat catalog of the human genome. *Nat Commun* **17**, 1106 (2026).
 100. Rajan-Babu, I.-S. *et al.* Population-scale disease-associated tandem repeat analysis reveals locus and ancestry-specific insights. *medRxiv*

- <https://doi.org/doi.org/10.1101/2025.10.11.25337795> (2025)
doi:doi.org/10.1101/2025.10.11.25337795.
101. Feng, Y. *et al.* GPU-accelerated and pipelined methylation calling. *Bioinformatics Advances* **2**, vbac088 (2022).
 102. Liu, T. & Conesa, A. Profiling the epigenome using long-read sequencing. *Nat Genet* **57**, 27–41 (2025).
 103. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
 104. Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M. A. & Friedman, J. M. Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nat Rev Genet* **25**, 476–499 (2024).
 105. Moeinzadeh, H. & Vingron, M. TandemTwister: Scalable genotyping and advanced visualization of tandem repeats.
 106. Zhang, Y. Multisample motif discovery and visualization for tandem repeats. *Genome Res* **35**, 850–862 (2025).
 107. Coster, W. Visualization and analysis of medically relevant tandem repeats in nanopore sequencing of control cohorts with pathSTR. *Genome Res* **34**, 2074–2080 (2024).
 108. Gustafson, J. A. *et al.* 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 (2024).
 109. Tesi, N. Characterizing tandem repeat complexities across long-read sequencing platforms with TREAT and otter. *Genome Res* **34**, 1942–1953 (2024).
 110. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
 111. Scholz, V. *et al.* Parallel in-depth analysis of repeat expansions: an updated Clin-CATS workflow for nanopore R10 flow cells. Preprint at <https://doi.org/10.1101/2024.11.05.622099> (2024).
 112. De Boer, E. N. *et al.* Nanopore Long-Read Sequencing as a First-Tier Diagnostic Test to Detect Repeat Expansions in Neurological Disorders. *IJMS* **26**, 2850 (2025).
 113. Montalvão, S. A. D. L., Lowe, A. & Kitchen, S. Advantages of external quality assessment-EQA programs. *Haemophilia* **28**, 679–686 (2022).
 114. Thornton, C. A., Johnson, K. & Moxley, R. T., 3rd. Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Ann. Neurol.* **35**, 104–107 (1994).
 115. Zarouchlioti, C. Tissue-specific TCF4 triplet repeat instability revealed by optical genome mapping. *EBioMedicine* **108**, 105328 (2024).
 116. Oehler, J. B., Wright, H., Stark, Z., Mallett, A. J. & Schmitz, U. The application of long-read sequencing in clinical settings. *Hum Genomics* **17**, 73 (2023).
 117. Yeom, E. *et al.* Performance Evaluation of PacBio PureTarget for Multiple Short Tandem Repeat Expansion Detection. *Ann Lab Med* <https://doi.org/10.3343/alm.2025.0414> (2026) doi:10.3343/alm.2025.0414.
 118. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics* **108**, 764–785 (2021).
 119. Zeng, Y. *et al.* GGC Repeat Expansion of *RILPL1* is Associated with Oculopharyngodistal Myopathy. *Annals of Neurology* **92**, 512–526 (2022).
 120. Tan, D. *et al.* CAG Repeat Expansion in *THAP11* Is Associated with a Novel Spinocerebellar Ataxia. *Movement Disorders* **38**, 1282–1293 (2023).
 121. De Coster, W. *et al.* A repeat expansion in GOLGA8A is a major risk factor for atypical frontotemporal lobar degeneration with ubiquitin-positive inclusions. *Nat Genet* <https://doi.org/10.1038/s41588-026-02537-7> (2026) doi:10.1038/s41588-026-02537-7.
 122. De Bitter, T. J. J. *et al.* Clinical long-read genome sequencing for rare disease

- diagnostics. Preprint at <https://doi.org/10.64898/2026.01.13.26343759> (2026).
123. Hiatt, S. M. *et al.* Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. Preprint at <https://doi.org/10.1101/2024.03.22.24304633> (2024).
 124. Spector, E. Laboratory testing for fragile X, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* **23**, 799–812 (2021).
 125. Crawford, D. C. Factors involved in the initial mutation of the fragile X CGG repeat as determined by sperm small pool PCR. *Human Molecular Genetics* **9**, 2909–2918 (2000).
 126. Fernandez-Carvajal, I. *et al.* Expansion of an FMR1 Grey-Zone Allele to a Full Mutation in Two Generations. *The Journal of Molecular Diagnostics* **11**, 306–310 (2009).
 127. Killoran, A. *et al.* Characterization of the Huntington intermediate CAG repeat expansion phenotype in PHAROS. *Neurology* **80**, 2022–2027 (2013).
 128. Kay, C. *et al.* Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
 129. Lee, J.-M. *et al.* CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* **78**, 690–695 (2012).
 130. Fotsing, S. F. *et al.* The impact of short tandem repeat variation on gene expression. *Nat Genet* **51**, 1652–1659 (2019).
 131. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
 132. Gall-Duncan, T., Sato, N., Yuen, R. K. C. & Pearson, C. E. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Res.* **32**, 1–27 (2022).
 133. Omeranovic, A., Van Long, F. N., Boubaker, A., Turgeon, A. & Nabi, H. Underrepresented populations in genomic research: a qualitative study of researchers’ perspectives. *BMC Med Genomics* **18**, 72 (2025).
 134. Dolzhenko, E. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
 135. Halman, A., Dolzhenko, E. & Oshlack, A. STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Human Mutation* **43**, 859–868 (2022).
 136. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
 137. Lundström, O. (Sachenkova) *et al.* WebSTR: A Population-wide Database of Short Tandem Repeat Variation in Humans. *Journal of Molecular Biology* **435**, 168260 (2023).
 138. Cui, Y. *et al.* A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell* **187**, 2336–2341.e5 (2024).
 139. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, D1062–D1067 (2018).
 140. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N Engl J Med* **372**, 2235–2242 (2015).
 141. Rivera-Muñoz, E. A. *et al.* ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Human Mutation* **39**, 1614–1622 (2018).
 142. Schloissnig, S. *et al.* Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).
 143. Tran, H. *et al.* Suppression of mutant C9orf72 expression by a potent mixed backbone antisense oligonucleotide. *Nat. Med.* **28**, 117–124 (2022).

144. Rook, M. E. & Southwell, A. L. Antisense oligonucleotide therapy: From design to the Huntington disease clinic. *BioDrugs* **36**, 105–119 (2022).
145. Shah, S. *et al.* Antisense oligonucleotide rescue of CGG expansion-dependent FMR1 mis-splicing in fragile X syndrome restores FMRP. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2302534120 (2023).
146. Scoles, D. R. *et al.* Antisense oligonucleotide therapy for spinocerebellar ataxia type 2. *Nature* **544**, 362–366 (2017).
147. Thornton, C. A. *et al.* Antisense oligonucleotide targeting DMPK in patients with myotonic dystrophy type 1: a multicentre, randomised, dose-escalation, placebo-controlled, phase 1/2a trial. *Lancet Neurol.* **22**, 218–228 (2023).
148. Mangin, A. *et al.* Robust Detection of Somatic Mosaicism and Repeat Interruptions by Long-Read Targeted Sequencing in Myotonic Dystrophy Type 1. *IJMS* **22**, 2616 (2021).
149. Olivucci, G. *et al.* Long read sequencing on its way to the routine diagnostics of genetic diseases. *Front. Genet.* **15**, 1374860 (2024).
150. Scahill, R. I. *et al.* Somatic CAG repeat expansion in blood associates with biomarkers of neurodegeneration in Huntington’s disease decades before clinical motor diagnosis. *Nat. Med.* **31**, 807–818 (2025).
151. Pardo-Palacios, F. J. *et al.* Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat. Methods* **21**, 1349–1363 (2024).
152. Warner, J. P. A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J Med Genet* **33**, 1022–1026 (1996).
153. Botta, A. Identification and characterization of 5’ CCG interruptions in complex DMPK expanded alleles. *Eur J Hum Genet* **25**, 257–261 (2017).
154. Rajan-Babu, I.-S., Law, H.-Y., Yoon, C.-S., Lee, C. G. & Chong, S. S. Simplified strategy for rapid first-line screening of fragile X syndrome: closed-tube triplet-primed PCR and amplicon melt peak analysis. *Expert Rev Mol Med* **17**, 7 (2015).
155. Moore, S. *et al.* Genome Mapping Nomenclature. *Cytogenet Genome Res* **163**, 236–246 (2023).
156. Sanden, B. Optical genome mapping enables accurate testing of large repeat expansions. *Genome Res* **35**, 810–823 (2025).
157. Fazal, S. A genome-wide approach for the discovery of novel repeat expansion disorders in the Undiagnosed Diseases Network cohort. *Genet Med* **27**, 101462 (2025).
158. Rajan-Babu, I.-S. *et al.* Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med* **13**, 126 (2021).
159. Yau, W. Y. Diagnostic yield and limitations of whole-genome sequencing for hereditary cerebellar ataxia. *Brain Commun.* <https://doi.org/0.1093/braincomms/fcaf188> (2025) doi:0.1093/braincomms/fcaf188.
160. Sehgal, A., Ziaei Jam, H., Shen, A. & Gymrek, M. Genome-wide detection of somatic mosaicism at short tandem repeats. *Bioinformatics* **40**, 485 (2024).
161. Devonshire, A. S. *et al.* Interlaboratory evaluation of high molecular weight DNA extraction methods for long-read sequencing and structural variant analysis. *BMC Genomics* **26**, 698 (2025).
162. Trigodet, F. *et al.* High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular Ecology Resources* **22**, 1786–1802 (2022).
163. Mastroianni, F. K., Miller, D. E. & Eichler, E. E. Applications of long-read sequencing to Mendelian genetics. *Genome Med* **15**, 42 (2023).
164. O’Brien, F. G., Udo, E. E. & Grubb, W. B. Contour-clamped homogeneous electric field electrophoresis of *Staphylococcus aureus*. *Nat Protoc* **1**, 3028–3033 (2006).
165. Maghini, D. G., Moss, E. L., Vance, S. E. & Bhatt, A. S. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human

- gut microbiome. *Nat Protoc* **16**, 458–471 (2021).
166. Jones, A. *et al.* High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PLoS ONE* **16**, e0253830 (2021).
 167. Loomis, E. W. *et al.* Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research* **23**, 121–128 (2013).
 168. Wang, M. *et al.* PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* **16**, 214 (2015).
 169. Madsen, E. B., Höijer, I., Kvist, T., Ameur, A. & Mikkelsen, M. J. Xdrop: Targeted sequencing of long DNA molecules from low input samples using droplet sorting. *Human Mutation* **41**, 1671–1679 (2020).
 170. Grosso, V. *et al.* Characterization of FMR1 Repeat Expansion and Intragenic Variants by Indirect Sequence Capture. *Front. Genet.* **12**, 743230 (2021).
 171. Tsai, Y.-, Brown, K., Bernardi, M., Harting, J. & Clelland, C. Single-Molecule Sequencing of the C9orf72 Repeat Expansion in Patient iPSCs. *BIO-PROTOCOL* **14**, (2024).
 172. Iyer, S. V., Goodwin, S. & McCombie, W. R. Leveraging the power of long reads for targeted sequencing. *Genome Res.* **34**, 1701–1718 (2024).
 173. Höijer, I. *et al.* Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing. *Human Mutation* **39**, 1262–1272 (2018).
 174. Kaplun, L. *et al.* ONT in Clinical Diagnostics of Repeat Expansion Disorders: Detection and Reporting Challenges. *IJMS* **26**, 2725 (2025).
 175. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* **39**, 431–441 (2021).
 176. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat Methods* **13**, 751–754 (2016).
 177. Munro, R. *et al.* Enhancing nanopore adaptive sampling for PromethION using readfish at scale. *Genome Res.* **35**, 877–885 (2025).
 178. Tachikawa, K. *et al.* Cost-Effective Cas9-Mediated Targeted Sequencing of Spinocerebellar Ataxia Repeat Expansions. *The Journal of Molecular Diagnostics* **26**, 85–95 (2024).
 179. Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet* **51**, 1215–1221 (2019).
 180. Perera, B. L. A. *et al.* First clinical diagnosis of FAME3 via commercial Long-Read sequencing reveals mosaic repeat expansion in MARCHF6 gene. *Neurogenetics* **26**, 61 (2025).
 181. Alayed, B. *et al.* Long-Range PCR and Nanopore Sequencing Enables High-Throughput Detection of TCF4 Trinucleotide Repeat Expansions in Fuchs Endothelial Corneal Dystrophy. *Mol Diagn Ther* **29**, 801–812 (2025).
 182. Ganesamoorthy, D. *et al.* High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing. *F1000Res* **9**, 1084 (2020).
 183. Steiert, T. A. *et al.* High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing. *NAR Genomics and Bioinformatics* **4**, lqac051 (2022).
 184. Tsai, Y.-C. *et al.* Identification of a CCG-Enriched Expanded Allele in Patients with Myotonic Dystrophy Type 1 Using Amplification-Free Long-Read Sequencing. *The Journal of Molecular Diagnostics* **24**, 1143–1154 (2022).