Published online 5 December 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 4 **1**
https://doi.org/10.1093/nargab/lqac089

Repeat Detector: versatile sizing of expanded tandem repeats and identification of interrupted alleles from targeted DNA sequencing

Alysha S. Taylor [1,†], Dinis Barros [2,†], Nastassia Gobet [2,†], Thierry Schuepbach [3,4,†], Branduff McAllister [5,6], Lorene Aeschbach [2], Emma L. Randall [1], Evgeniya Trofimenko [2,7], Eleanor R. Heuchan [1], Paula Barszcz [2], Marc Ciosi [8], Joanne Morgan [5], Nathaniel J. Hafford-Tear [9], Alice E. Davidson [9], Thomas H. Massey [5], Darren G. Monckton [8], Lesley Jones [5], REGISTRY Investigators of the European Huntington's disease network[‡], Ioannis Xenarios [2,10] and Vincent Dion [1,*]

[1]UK Dementia Research Institute, Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff, CF24 4HQ, UK, [2]Centre for Integrative Genomics, University of Lausanne, Bâtiment Génopode, 1015 Lausanne, Switzerland, [3]Vital-IT Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [4]Newbiologix, Ch. De la corniche 6-8, 1066 Epalinges, Switzerland, [5]MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK, [6]Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA, [7]Sorbonne Université, École normale supérieure, PSL University, CNRS, Laboratoire des biomolécules, LBM, 75005 Paris, France, [8]School of Molecular Biosciences, College of Medical, Veterinary and Life Sciences, Davidson Building, University of Glasgow, Glasgow, G12 8QQ, UK, [9]UCL Institute of Ophthalmology, 11-43 Bath Street, London, EC1V 9EL UK and [10]Health2030 Genome Center, Ch des Mines 14, 1202 Genève, Switzerland

Received August 02, 2022; Revised October 25, 2022; Editorial Decision November 01, 2022; Accepted November 08, 2022

ABSTRACT

Targeted DNA sequencing approaches will improve how the size of short tandem repeats is measured for diagnostic tests and preclinical studies. The expansion of these sequences causes dozens of disorders, with longer tracts generally leading to a more severe disease. Interrupted alleles are sometimes present within repeats and can alter disease manifestation. Determining repeat size mosaicism and identifying interruptions in targeted sequencing datasets remains a major challenge. This is in part because standard alignment tools are ill-suited for repetitive and unstable sequences. To address this, we have developed Repeat Detector (RD), a deterministic profile weighting algorithm for counting repeats in targeted sequencing data. We tested RD using blood-derived DNA samples from Huntington's disease and Fuchs endothelial corneal dystrophy patients sequenced using either Illumina MiSeq or Pacific Biosciences single-molecule, real-time sequencing platforms. RD was highly accurate in determining repeat sizes of 609 blood-derived samples from Huntington's disease individuals and did not require prior knowledge of the flanking sequences. Furthermore, RD can be used to identify alleles with interruptions and provide a measure of repeat instability within an individual. RD is therefore highly versatile and may find applications in the diagnosis of expanded repeat disorders and in the development of novel therapies.

INTRODUCTION

Huntington's disease (HD) is one of the best studied members of a family of disorders caused by the expansion of short tandem repeats (1). It is characterized by neurodegeneration in the striatum and cortex, leading to chorea, cognitive decline, and premature death (2). The size of the inherited CAG repeat tract at the huntingtin (*HTT*) locus accounts for about 60% of the variability in the age at motor disease onset (3, 4), with longer repeats associated with earlier onset. Consequently, it is not possible to predict HD onset solely based on *HTT* repeat size, highlighting the im-

*To whom correspondence should be addressed. Tel: +44 29 2251 0893; Email: dionv@cardiff.ac.uk
†The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.
‡http://www.ehdn.org/wp-content/uploads/REGISTRY-contributors-full-list.pdf

portance of other factors contributing to disease pathology. One such factor is likely to be somatic expansion, or the ongoing expansion of expanded repeats in affected tissues throughout an individual's lifetime (5). The contribution of somatic expansion to pathogenesis is highlighted by the number of genes implicated in repeat instability that also appear to modify age at disease onset (6, 7). It also follows that if ongoing somatic expansion contributes to disease phenotypes, gains or losses of interruptions within the repeat tract should lead to changes in the age at disease onset. When repeats are interrupted, the tract is stabilized and there correlates a later appearance of disease symptoms (6,8). About 95% of HD chromosomes have a CAACAG motif immediately 3′ to the end of the CAG repeat tract, often referred to as an interruption (8). Alleles without this CAACAG interruption are associated with an earlier onset than predicted based on their repeat size (6,8–11), whereas those with two CAACAG were either found to have no effect on the age at disease onset (6,8) or were associated with a later onset (10).

Both somatic expansion and repeat interruptions also appear to influence disease outcome in other expanded repeat disorders (12). For example, somatic expansion is seen in affected tissues in myotonic dystrophy (13–16). Moreover, some of the genetic modifiers of HD implicated in repeat expansion may also modify disease onset in other repeat disorders (17). Interruptions in SCA1, SCA2, Fragile X syndrome and myotonic dystrophy type 1 are associated with lower repeat instability, delayed symptom onset and/or modified clinical manifestations (15,16,18–28).

Despite their influence on disease outcomes, interruptions are difficult to identify using current PCR-based diagnostic tools (26,29), and repeat instability is not currently measured in the clinic. The advent of high-throughput sequencing offers an opportunity to improve diagnosis by enhancing the accuracy of repeat sizing as well as the identification of interrupted alleles. Whole genome sequencing is able to determine the modal repeat size of an allele obtained from blood samples using both long and short read sequencing [e.g. (30–32)]. However, such approaches do not allow the measurement of intra-sample repeat size mosaicism. Indeed, short reads do not span the entire length of the repeat size and instead use multiple reads that span parts of the repeat tract to obtain an estimate of the modal repeat size in the sample. This comes at the cost of describing the repeat size mosaicism within a sample. Even when using long read sequencing, the coverage from whole genome sequencing is not high enough at any one expanded repeat locus to determine repeat size distribution accurately. Therefore, targeted sequencing that spans the entirety of the repeat tract is currently the best means of obtaining an accurate measure of repeat size (30,33,34) and it has the added advantage of unveiling repeat size mosaicism (35). This makes targeted sequencing highly complementary to whole genome sequencing approaches aimed at finding novel expansions or diagnosing patients.

Targeted sequencing of expanded repeats has been achieved with Illumina MiSeq (8,10,35), Pacific BioSciences (PacBio) Single-Molecule, Real Time (SMRT) sequencing (18,26,29,35–42) and Oxford Nanopore Technology MinION (30,37,43–45). One of the remaining bottlenecks is the robustness of computational pipelines that can reliably determine repeat size and repeat interruptions at the single-molecule level in targeted sequencing datasets. Current algorithms (8,46–49) all rely on the alignment of each read to a reference sequence. The presence of a highly variable tandem repeat can result in the rejection of read from the dataset, thereby introducing biases. The alignment step also limits the application of these algorithms to specific loci or genomes. Importantly, only one currently available algorithm allows for the unsupervised identification of novel interrupted alleles, the proprietary RepeatAnalysisTools by Pacific Biosciences, but it only works on data generated using the amplification-free library preparation for SMRT sequencing (29,38). Here, we present Repeat Detector (RD), an alignment-free algorithm that accurately counts expanded repeats in targeted sequencing datasets and can identify interrupted alleles. It is versatile as it works on datasets from multiple loci, sequencing platforms and repeated motifs, making it widely applicable.

MATERIALS AND METHODS

Cell culture and cell lines

The GFP$(CAG)_x$ cell lines were cultured as described before (50,51). The culture medium used was Gibco™ Dulbecco's modified Eagle's medium (DMEM) with GlutaMAX™, 10% foetal bovine serum (FBS), 100 U ml^{-1} of penicillin/streptomycin, 15 μg ml^{-1} blasticidine and 150 μg ml^{-1} hygromycin. The HD lymphoblastoid cells (LBCs) or their DNA used for SMRT sequencing were obtained from the Coriell BioRepository (Supplementary Table S1). The LBCs were grown in Gibco™ RPMI with GlutaMAX™ supplemented with 15% Gibco™ FBS (Thermo Fisher), and 1% penicillin-streptomycin. Both the LBCs and the GFP$(CAG)_x$ cells were grown at 37°C with 5% CO_2 and tested negative for mycoplasma by Eurofins' 'Mycoplasmacheck' service. GFP$(CAG)_{91}$ is identical to the previously characterised GFP$(CAG)_{101}$ (51) but contained a contraction in the cultures used here. Similarly, GFP$(CAG)_{51}$ had a one CAG expansion compared to when it was first derived (51) and GFP$(CAG)_{308}$ had a repeat tract above 270 that we could not fully sequence with Sanger sequencing at the time. GFP$(CAG)_{15}$, GFP$(CAG)_{51}$ and GFP$(CAG)_{308}$ are all derived from GFP$(CAG)_{101}$.

Confirmation of interruption

We confirmed the presence of an interruption in GFP$(CAG)_{308}$ by first amplifying the repeat region using primers oVIN-459 and oVIN-460 (for primer sequences see Supplementary Table S2) and then Sanger sequencing using the same primers. The Sanger sequencing was done by GeneWiz.

SMRT sequencing

The HD LBCs and GFP$(CAG)_x$ datasets were generated by first isolating DNA using the Macherey-Nagel Nucleospin™ Tissue Mini kit. PCR products were generated from samples using barcoded primers as listed in Supplementary Table S1 and Thermo™ Phusion II High Fidelity polymerase. To obtain sufficient quantities of PCR product to

proceed with library preparation, multiple identical PCRs were pooled and purified using Macherey-Nagel™ Gel and PCR Clean-up kit columns. The library was generated using the SMRTbell Template Prep Kit (1.0-SPv3) according to manufacturer's instructions. Samples to be sequenced on the same flowcell were combined in equimolar pools. We loaded between 10 and 12 pM. SMRT sequencing was done using a Sequel IIe at Cardiff University School of Medicine. CCSs were generated from the resulting sequences and processed using SMRT Link.

Participants

Human subjects were selected from the European Registry-HD study (52) (N = 507) (https://www.enroll-hd.org/enrollhd_documents/2016-10-R1/registry-protocol-3.0.pdf). Ethical approval for Registry was obtained in each participating country. Participants gave written informed consent. Experiments described herein were conducted in accordance with the Declaration of Helsinki. Institutional ethical approval was gained from Cardiff University School of Medicine Research Ethics Committee (19/55). Subject selection is described in (10).

HD MiSeq dataset

A total of 652 DNA samples were sequenced, with the majority of these being immortalized lymphoblastoid (LBC) cell lines (N = 547) and a smaller number of blood DNAs (N = 49). These were sequenced using an ultra-high depth MiSeq sequencing methodology, described elsewhere (35,53). Of note, the method includes a size selection step that biases towards longer alleles. 649 of the original 652 samples were successfully sequenced (>99%). Supplementary Table S2 describes the numbers of each sample as well as the numbers of each DNA type that was successfully sequenced.

FECD SMRT dataset

The FECD SMRT dataset is a amplification-free SMRT sequencing dataset from blood samples of FECD patients published previously (38).

Repeat detector

Repeat Detector source code and dependencies are available at: https://github.com/DionLab/RepeatDetector. To determine repeat sizes for GFP$(CAG)_x$, HD SMRT, FECD SMRT (38), HD MiSeq (10), *C9ORF72* locus (43), and the HD MinION (30) datasets, unaligned reads were assessed using permissive and restrictive profiles with a repeat size range of [0–1000] in the first instance and increased if needed (e.g., with the FECD SMRT dataset). For each analysis, the –with-revcomp option was enabled and data were output to a density plot (-o histogram option). Weighting scores for the permissive and restrictive parameters can be found in Supplementary Figure S1. Density plots obtained were graphed using GraphPad PRISM version 9.

ScaleHD

The ScaleHD parameters were set as previously (10). For comparisons between ScaleHD and RD presented in Figure 3B, Supplementary Figures S2 and S3, we used the total number of reads mapping to the *HTT* locus in the R1 FASTA files, regardless of the flanking sequences that sometimes differed between reads from the same sample. These differences are due to PCR and sequencing errors.

Tandem-genotypes

The FECD SMRT dataset (38) was aligned to GRCh38.18 accessed from the Genome Reference Consortium (54) using the LAST aligner (55), as per recommendation in (48). The reference sequence was soft-masked as per LAST aligner guidelines (https://github.com/mcfrith/last-rna) and sequences were aligned using default settings as described in the wiki. Aligned sequences were examined for the FECD repeat using Tandem-Genotypes recommended settings and modal repeat sizes were extracted from the output files.

RESULTS

Repeat detector

RD (Figure 1) is based on the deterministic profile weighting algorithm, pfsearchV3.0, which was originally designed for protein motifs and domain detection (56,57). It has been adapted to use circular profiles on DNA sequences. RD is not dependent on an alignment to a specific reference sequence. Instead, the user defines the repeated motif and the weighting parameters. RD then aligns the reads to a circular representation of the motif of interest. The weights of the profile give flexibility to adapt the alignment scoring to prior knowledge, for example, about the idiosyncratic errors of a given sequencing platform or for the repeated motif of interest.

RD applied to two different loci over a wide range of repeat sizes

We first tested RD on two different datasets generated using SMRT sequencing and a standard PCR-based library preparation method. SMRT sequencing uses rolling circle replication chemistry that generates reads with multiple copies of the target sequence called subreads. A proprietary bioinformatics tool generates circular consensus sequences (CCSs) from subreads, improving base calling accuracy (58). Our first dataset consisted of CCSs from HEK293-derived cell lines with 15, 51, 91 and 308 CAG/CTG repeats inserted within a hemizygous ectopic GFP reporter on chromosome 12 (51,59,60). We refer to these cells as GFP$(CAG)_x$, with x being the number of repeats. These lines are single-cell isolates derived from the previously characterized GFP$(CAG)_{101}$ line (see Materials and Methods and (51)). The second dataset was composed of 21 DNA samples and LBCs from HD individuals obtained from the Coriell BioRepository with repeats ranging from 15 to 750 units (Supplementary Table S1). Taking both datasets together, we recovered the expected repeat sizes based on

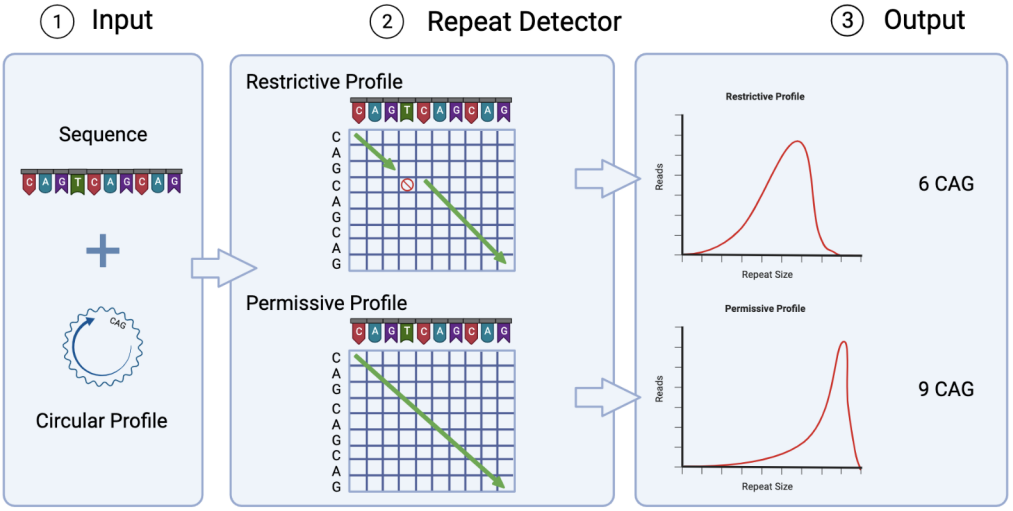

Figure 1. Repeat Detector flowchart. RD requires both the FASTA files of the DNA sequences and the circular profile of the repeating motif of interest as inputs. Using a substitution matrix, it calculates a score, taking into account matches, mismatches, gaps, and insertions. The repeat size with the largest score is deemed to be the correct one. There are two sets of parameters described in the methods. One is permissive and is lenient with non-matching nucleotides. The other is restrictive and stops counting when a mismatch, gap or insertion is encountered. RD outputs the frequencies of repeat sizes, which are then presented as density plots.

Coriell's data or our prior work (61), except for one sample (Figure 2A and B). Only the sample with the longest repeat tract, GM14044, which we have shown to contain 750 repeats (61), returned a repeat size of 50 CAGs. By inspecting reads manually, we confirmed that the sequences in the FASTA files used by RD contained repeat sizes <750 repeats, suggesting that, rather than a specific problem with RD, there was a bias against longer repeats during PCR, loading of the SMRT flowcell, sequencing, and/or the generation of CCSs. These results are in line with recent findings suggesting that up to at least 550 CAG repeats can be sequenced using SMRT sequencing (29,35).

RD is highly accurate on HD samples

We next sought to quantify the accuracy of RD in sizing clinically relevant samples. To do so, we took advantage of a previously sequenced set of 649 samples derived from 507 clinically manifesting HD individuals (10). This cohort included samples from 497 LBC lines, 49 blood samples sequenced twice, 47 LBC samples that were passaged extensively and an additional seven LBC samples from a single HD individual with a known repeat length, which ensured reproducibility (Supplementary Table S3). For 42 individuals, there are data for both blood and LBCs. Hereafter, we refer to this dataset as the HD MiSeq dataset since it was generated using Illumina MiSeq technology (10). This dataset was originally analysed for modal repeat size and flanking sequences using ScaleHD (53). This algorithm uses a library containing over four thousand reference sequences with all known flanking sequences as well as repeat sizes between 1 and 200 CAGs. This created a robust benchmark against which we could evaluate RD for its ability to determine repeat size. Of the 649 samples, we analysed 609 with both algorithms, totaling 1218 germline alleles (Figure 3A). For the shorter alleles, the modal repeat size was determined to be the same with both softwares (Figure 3B).

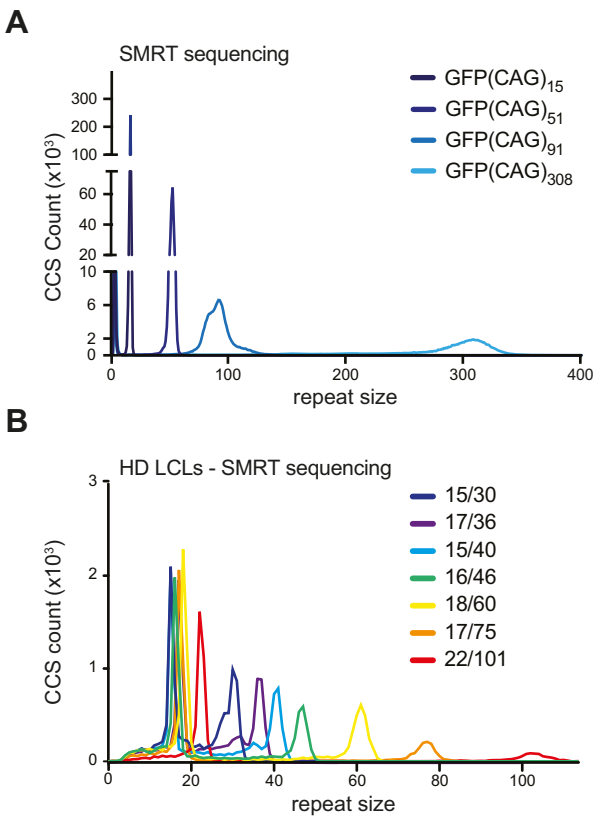

Figure 2. Repeat Detector applied to SMRT sequencing of an ectopic CAG/CTG repeat and at the *HTT* locus. (**A**) RD-generated repeat size distribution from SMRT sequenced of ectopic CAG repeats in GFP(CAG)$_x$ cell lines using a PCR-based library preparation. (**B**) Repeat size distribution of SMRT-sequenced samples of the *HTT* locus from HD-derived LBCs using a PCR-based library preparation. Only a selection of the 22 samples are shown for clarity. All samples are shown in Supplementary Figure S4. Read depth and mapping metrics for all datasets can be found in Supplementary Table S4.

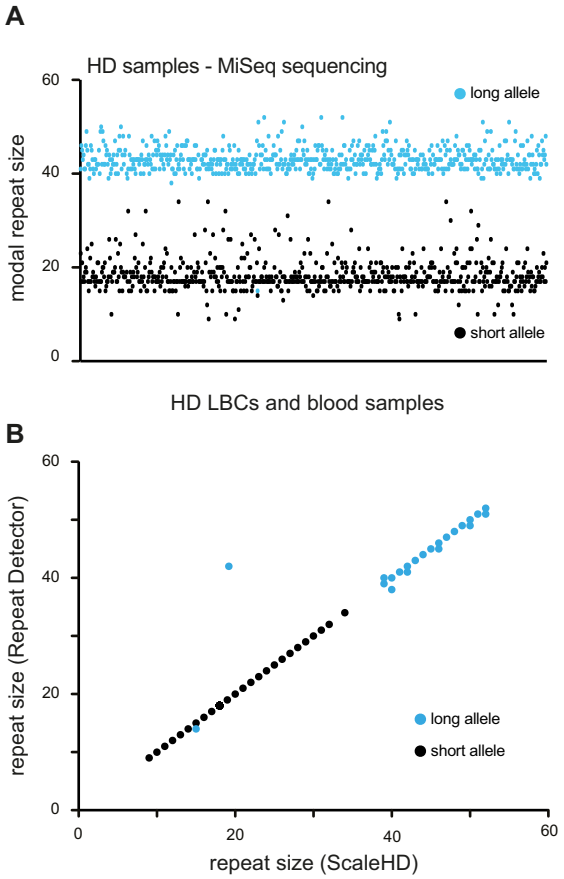

Figure 3. Repeat Detector is highly accurate on HD samples. (**A**) Modal repeat size in the HD MiSeq dataset determined by RD using the restrictive parameters. Each dot is an allele. Blue dots are the longer of the two alleles in a sample, whereas black dots are the shorter alleles. Read depth and mapping metrics for all datasets can be found in Supplementary Table S4. (**B**) Modal repeat size in the HD MiSeq samples comparing ScaleHD and Repeat detector.

Of the longer alleles, 599 out of 609 (98.3%) had the same modal allele size (Figure 3B). Of the remaining ten alleles, nine differed by one CAG and one allele by two (Supplementary Figure S2). One of these differences came from a homozygous individual with 2 alleles of 15 repeats. The script, downstream of RD, looks for the two most common allele sizes and thus determined erroneously that this sample had one allele with 15 repeats and one with 14 (Supplementary Figure S3a). The sample that differed most between ScaleHD and RD was a LBC sample derived from a confirmed HD individual. We had several samples from the same individual, yet ScaleHD determined this LBC sample to have two alleles with 19 repeats (Supplementary Figure S3b). RD, on the other hand, found one allele with 19 repeats and one with 42, in line with the other samples from this individual. The discrepancy was due to ScaleHD filtering out much of the reads containing the expanded allele. It is unclear why this occurred. RD does not rely on an alignment to the locus of interest and thus counted both alleles accurately (Supplementary Figure S3b). These data highlight the accuracy of RD and show that it is comparable to ScaleHD for the *HTT* locus.

RD is applicable on multiple repeat compositions

To test the applicability of RD to other repeat compositions, we analysed publicly available datasets generated using PCR-free libraries for SMRT (38,39) and MinION (43,62) sequencing. These datasets included expanded CAG, CTG and GGGGCC repeats, as well as short CGG, GGGGCC and ATTCT repeats. RD found the same repeat size as previously reported for every sample sequenced using SMRT technology (Supplementary Figure S5). However, with the MinION sequencing data containing expanded GGGGCC repeats (43), RD dramatically underestimated the repeat size (Supplementary Figure S6a). Upon visual inspection of the MinION sequencing reads, we found that the expected repeat motif was too often mutated to be reliably detected (Supplementary Figure S6b). This is consistent with Ebert *et al.* (37), who found that when generating whole genome sequences using MinION there was no read aligning to the GGGGCC repeat at the *C9orf72* locus. To determine whether this was indeed due to the quality of MinION sequencing rather than repeat motif composition, we used a recently published MinION dataset that included expanded CAG/CTG repeats from the *HTT* locus (30). We found that only a few sequences were accurate enough to determine repeat size. Most had a very high error rate that prevented us from obtaining accurate repeat counts in this dataset (Supplementary Figure S6cd). We conclude that RD is applicable to datasets generated with MinION, which is too error-prone to identify repeat size down to individual reads.

RD exposes repeat instability in amplification-free datasets

We next sought to determine whether we would have enough accuracy at the single CCS level to detect heterogeneity of repeat sizes within samples. This was already suggested in the previous datasets with the larger repeat tracts showing more size heterogeneity (Figure 2). However, in PCR-based library preparation methods there may be slippage errors and other PCR artefacts that may contribute to size heterogeneity, and the distribution of repeat size may not be limited to biological variation (8,35). Up to 80% of Fuchs endothelial corneal dystrophy (FECD) patients have an expansion of 50 or more CTGs in the third intron of *TCF4* (termed CTG18.1) (63). Here, we analysed a high-quality amplification-free library generated from FECD patient-derived whole blood genomic DNA samples ($n = 11$) displaying a diverse range of CTG18.1 allele lengths and zygosity status (Figure 4) (38). We found that we could reproduce, for all samples, the modal repeat size determined previously using PacBio's proprietary RepeatAnalysisTools (Table 1). In addition, repeat instability was obvious with expansion-biased mosaicism, especially for longer alleles (Table 1, Figure 4 and Supplementary Figure S7). We found that RD was largely in agreement with previous studies by Hafford-Tear *et al.* (38) in determining the largest repeat tract present in a sample. In one case, however, RD found a maximum repeat length in one of the samples to be over 1300 units larger than previously identified (566 CTGs identified using RepeatAnalysisTools versus 1875 CTGs with RD, Table 1). Tandem-Genotypes (48), by contrast, found significantly larger alleles than RD or RepeatAnalysisTools on the expanded alleles, suggesting that

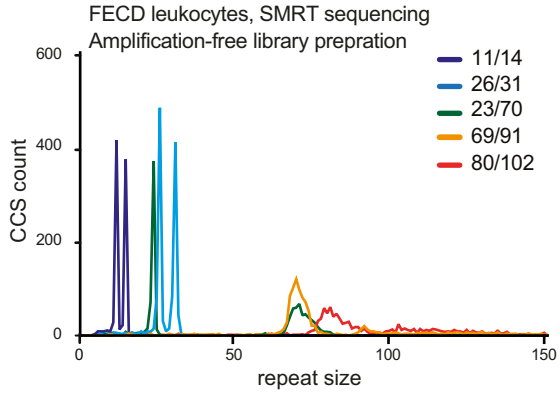

Figure 4. RD exposes repeat instability in amplification-free datasets. Repeat size distribution of the CTG repeat found within a FECD patient cohort from reference (38) prepared using an amplification-free library and SMRT sequenced. Note the wide spread of the repeat sizes on the larger alleles. Only a selection of the samples are plotted for clarity. Density plots for all the samples can be found in Supplementary Figure S7. Read depth and mapping metrics for all datasets can be found in Supplementary Table S4.

it is the more permissive algorithm. Specifically for modal repeat size, it often diverged by a few repeats compared to both RD and RepeatAnalysisTools, with the latter two being in agreement. Together these results show that RD may be used to determine the frequency of repeat instability, in addition to modal repeat size for the FECD SMRT dataset.

Identifying interrupted alleles using RD

When optimizing RD, we settled on two sets of parameters, one that allowed for the occurrence of sequencing errors (permissive) and one that did not (restrictive). The analyses presented above were conducted using the restrictive parameters. On the HD MiSeq dataset, the restrictive parameters returned the length of the pure repeat tract whereas the permissive parameters count the downstream interruption and the first triplet downstream of the repeat tract, typically CCG. Thus, alleles with the canonical CAACAG interruption will yield a difference of three repeats between the permissive and restrictive parameters (Figure 5A and B). By contrast, alleles without the interruption yield only one repeat difference between the two profiles (Figure 5A and B) and the ones with a duplicated CAACAG motif show a difference of 5 units (Figure 5A and C). The shifts can be used to identify samples with repeat interruptions or unusual allele structures and narrow down which samples need to be inspected manually.

We applied this approach to every allele in the HD MiSeq dataset to determine its accuracy (Supplementary Table S5). Of the 1286 alleles analysed, we found 6 false positives (0.5%), which had the canonical allele, but the difference between the repeat tract length from the permissive and restrictive parameters were 2 or 4, rather than the expected 3 (Supplementary Figure S8a). We identified one false negative (0.08%), which had an unusual allele of $(CAG)_{42}$CAACAACAGCCG and was expected to yield a difference of 4, but returned one of 3 instead, making it look like a canonical allele (Supplementary Figure S8b). Two more alleles (0.16%), which had rare structures with either $(CAG)_{38}$CAACAACAGCCG or $(CAG)_{48}$CAACAACAACAGCCG, were found to have differences one fewer than the expected 5 and 6, respectively. These were nonetheless flagged as non-canonical alleles. Importantly, we accurately identified the sole sample in the HD MiSeq dataset with a CAC interruption within its CAG repeat (Figure 5D). Applying it to the FECD SMRT samples, we could identify the known interruption in Sample 7 (Figure 5E). We also found a previously unknown 111bp insertion in the GFP$(CAG)_{308}$ cell line (Figure 5F), which we confirmed by PCR and Sanger sequencing, as well as in a separate flowcell. These results suggest that RD can be used to identify individual alleles with interruptions at multiple different loci with high accuracy.

DISCUSSION

Here we developed and applied RD, which detects and counts tandem repeats in targeted sequencing data. RD was as accurate as ScaleHD on the HD MiSeq dataset and as Tandem-Genotypes and RepeatAnalysisTools on the FECD SMRT dataset. RD could also identify interruptions, when present, as readily as RepeatAnalysisTools on the FECD SMRT dataset. None of the other available algorithms could be used with all of these datasets. For example, ScaleHD can identify known interruptions only at the *HTT* locus by adding them to its library of sequences whereas RepeatAnalysisTools can only be applied to amplification-free SMRT sequencing. Tandem-Genotypes could also be applied to multiple loci, but it is not designed to find interruptions. Tandem-Genotypes also requires a specific aligner, LAST (55), which does not work with artificial constructs such as our GFP reporter. Thus, the main strength of RD is its versatility: it works on multiple different sequencing platforms, multiple loci, including artificial reporters, and can identify interrupted alleles readily. Although RD allows for changing parameter scores to accommodate the systematic sequencing errors of each sequencing platform, we did not have to change the parameters when applying it to SMRT and MiSeq, or when we applied it to different loci or repeat compositions. Further optimization of the weighting profiles may help to compensate for the higher error rate of MinION sequencing datasets.

RD could detect repeat instability in HD and FECD blood-derived samples prepared with a PCR-based or amplification-free protocol, respectively. In the amplification-free TCF4 PacBio dataset where PCR biases against the longer repeats could be ruled out, some samples had large expansions with some reads having several hundreds of repeats. This is not uncommon in FECD patient-derived samples, but they are difficult to detect by any method, except perhaps for small-pool PCR followed by Southern blotting (64). Our data, together with that of a recent study on DM1 (29), suggest that it is possible to detect repeat instability as well as interruptions in PCR-free sequencing methods. More work needs to be done to validate this approach. Specifically, comparing samples with different levels of repeat instability using both small-pool PCR and amplification-free SMRT libraries will be critical. Notably, RD would not be suitable for whole

Table 1. Comparison between RepeatAnalysisTools, Repeat Detector and Tandem-Genotypes on previously published data for the FECD SMRT dataset

	Modal allele size (no. of repeat units - short/long alleles)			Largest repeat tract		
Sample	RepeatAnalysisTools[a]	Repeat Detector[b]	Tandem-Genotypes	RepeatAnalysisTools[a]	Repeat Detector[b]	Tandem-Genotypes
1	11/14	11/14	11/14	Not determined	15	18
2	25/30	25/30	25/30	37	37	37
3	23/70	23/70	23/68	90	90	90
4	23/73	23/73	23/74	115	115	116
5	11/80	11/80	11/81	169	169	170
6	32/110	32/110	31/110	566	1875	2001
7	17/131	9[c]/131	17/126	1361	1381	1393
8	80/102	80/102	80/102	498	498	506
9	72/118	72/118	73/117	1593	1285	2221
10	69/91	69/91	69/91	1014	1047	1050
11	79/141	79/141	78/140	Not determined	1581	1580

[a]Data from (38)
[b]The restrictive parameters were used to determined repeat size.
[c]This lower number is due to the presence of an interruption in this allele. This is evident when the permissive parameters are also used (see Figure 5).

genome sequencing datasets and these datasets would not be suitable to determine repeat size mosaicism.

Several datasets used Oxford Nanopore sequencing on expanded repeats (30,37,43,44), yet levels of repeat mosaicism was only reported in one study (44). This is likely because the error rate of MinION is too high to be confident about the size of the repeats in individual reads. On non-repetitive loci, this is not a problem because sequencing with a high coverage can compensate for stochastic errors in individual reads. On an unstable tandem repeat, however, this averages out the repeat size differences between reads and the distribution of the repeat size is lost. Oxford Nanopore is currently too error-prone for use to determine repeat size heterogeneity within a sample it can only be used to obtain modal repeat size. Improvements to base calling may help mitigate this issue.

Current sequencing efforts have been limited to modal repeat sizes below about 150 CAGs, with the notable exceptions of myotonic dystrophy samples (26,29). Here we could detect repeat sizes in excess of 1800 CTGs at the *TCF4* locus in individual reads. It will be interesting to test how well RD performs on datasets with longer repeats as those become available.

Interruptions within the repeat tract are classically detected using repeat-primed PCR, whereby a primer sits in the flanking sequence and another within the repeat tract itself (26,65). This leads to a pattern on capillary electrophoresis with a periodicity the size of the repeated unit and of decaying intensity as the fragments become longer. Interruptions appear in the intensity traces as gaps in places where the repeat primer could not bind. Depending on the position of the interruption within the repeat tract, these may be difficult to detect accurately, especially if they are far from the 3' or 5' ends of the repeat tracts. Once an interruption is detected, its identity and position need to be confirmed by Sanger sequencing or restriction digest. Targeted sequencing coupled with RD would identify first the presence of an interruption in the sample, and then the examination of individual reads would reveal both the position and the content of the interruption. This would dramatically speed up the process and may thereby reduce cost.

In its current version, RD has a few limitations. One is that it requires user intervention to identify the nature of the interruption detected in a sample and cannot discriminate between single and multiple interruptions in the same allele. This will be important to address as several alleles from DM1 patients, for example, with complex interruptions have been documented (26,29). In these samples, RD would return the size of the longest interruption-free repeat stretch. Moreover, the size of the interruption tolerated by the permissive parameters depends on the position of the interruption and on the number of repeated units flanking the insertion. For example, the larger interruption found in the $GFP(CAG)_{308}$ line was allowed with the permissive parameters because it was flanked by two repeat tracts of 155 and 115 repeats. Thus, in some cases, large interruptions may not be found, or the parameters may need to be adjusted. This was highlighted by the Oxford Nanopore datasets that we analysed here. RD ignores flanking sequences and thus would be blind to, for example, the significant polymorphism found in the CCG repeat downstream of the HD allele (8). To get around this, RD could be run once for the size of the CAG repeat and once for the size of the CCG repeats and its interruptions downstream of the repeat tract. Improvements to RD may also include changes to the weighting scores for improved accuracy on MinION datasets and on a wider variety of repetitive sequences (e.g. telomeres).

Some tandem repeats may not benefit from RD. For example, Variable Number Tandem Repeats (VNTRs) are not pure and often contain multiple different repeated motifs. In these cases, we would expect RD to be able to count the repeats provided that the permissive weighting scores are adjusted. The restrictive parameters would then return the longest stretch of pure repeats. Thus, for highly interrupted repeats, RD would perform similarly as on error-riddled reads.

We have shown that RD can accurately determine repeat size from targeted sequencing data from SMRT, MiSeq and MinION sequencing platforms. It is not limited by a requirement for a library of reference sequences, can be applied to a wide variety of disease loci and repeat compositions, can be used to identify alleles with interruptions, and can document repeat length mosaicism within a sample.

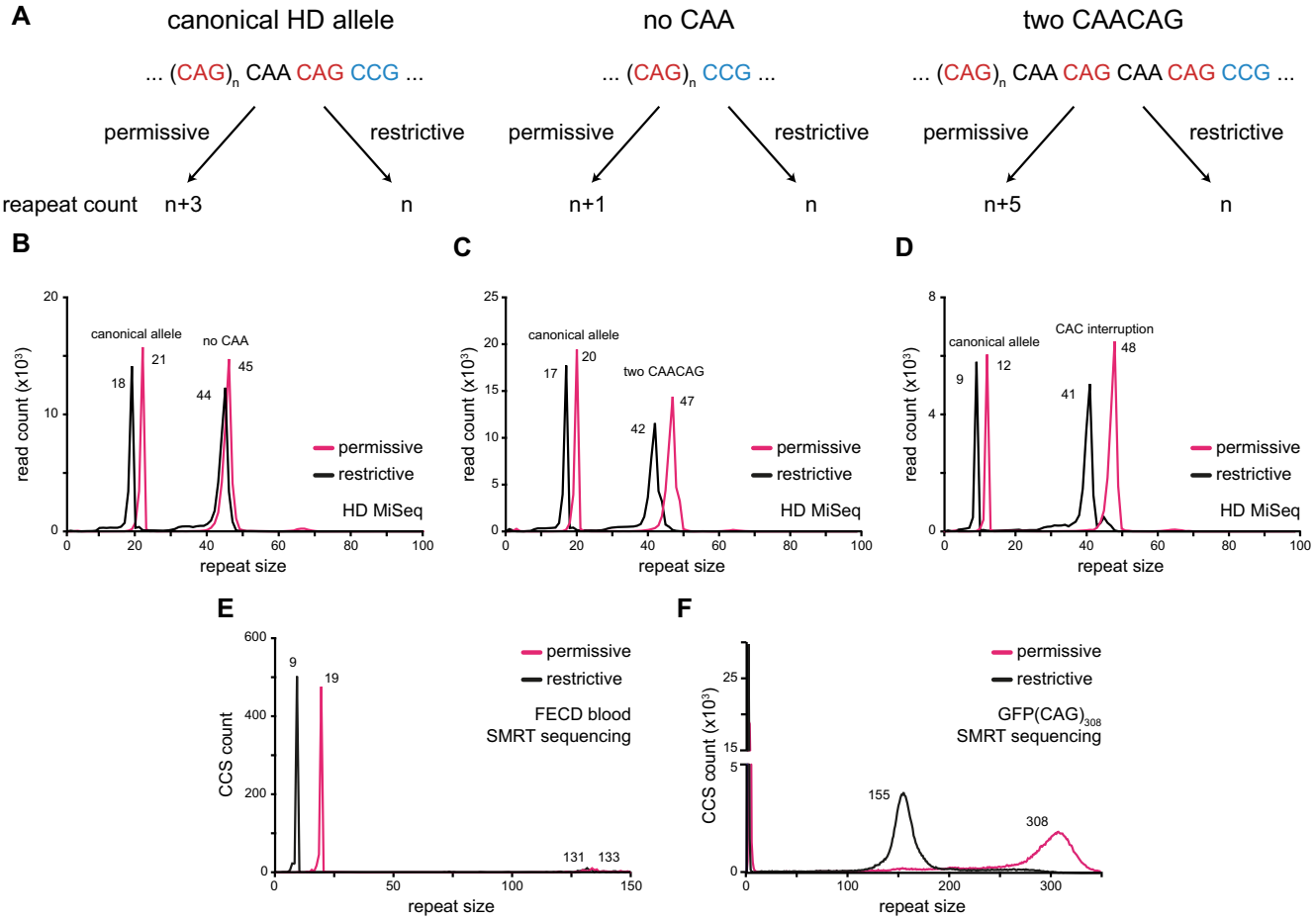

Figure 5. Identifying samples with interruptions using RD. (**A**) Interruptions at the 3' end of HD alleles can be distinguished using the difference in repeat size between RD's permissive and restrictive parameters. For instance, the most common allele (left), containing a CAA interruption will return a difference of 3 repeats between the parameter settings. By contrast, an allele without the CAA (middle) or with two CAACAG motifs return differences of 1 and 5, respectively. (**B**) Example of a sample from the HD MiSeq dataset with a canonical non-pathogenic allele, and an expanded allele without a CAA interruption. (**C**) Example for a HD MiSeq sample with a canonical short allele and an expanded allele with a duplicated CAACAG motif. (**D**) One of the samples contained a rare CAC interruption in the repeat tract that returns a difference larger than expected from the known alleles. (**E**) A previously known interrupted allele in a FECD sample (38) was correctly identified. (**F**) Our GFP(CAG)$_{308}$ line was found to have an insertion of 111bp after 155 CAG repeats.

Together, these characteristics make RD broadly applicable and capable tool for analysis of expanded tandem repeats.

DATA AVAILABILITY

The source code and dependencies are available here: https://github.com/DionLab/RepeatDetector and https://doi.org/10.5281/zenodo.7299814. The HD SMRT and GFP(CAG)$_x$ SMRT datasets are available from the Gene Expression Omnibus (GSE199005). The GFP(CAG)$_x$ lines are available upon request from the corresponding author.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank John H. Wilson, Alvaro Murillo Bartolome, Helder Ferreira, Fisun Hamaratoglu and Andrew Seeber for comments on the manuscript. Mark Ebbert kindly provided the Nanopore MinION data from ALS individuals. The Registry HD lymphoblastoid DNA and cell lines were provided by the European Huntington's Disease Network project #984. The Registry study is supported by the European Huntington's Disease Network (EHDN), funded by CHDI Foundation, Inc. The funding source had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication. We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via the Welsh Government. This work is dedicated to the memory of Lesley Jones. Her insights, wit, and generosity will be sorely missed.
Author contributions: A.S.T. analysed the interruptions and repeat size in the MiSeq HD, SMRT sequencing HD, and in the FECD SMRT sequencing datasets. She generated the tables and figures together with V.D., D.B. and E.H. D.B. optimized RD for CCS, analysed the repeat size on the HD

and GFP SMRT sequencing datasets, and initiated the analysis of repeat size in the HD MiSeq dataset. N.G. tested the beta version of RD and optimized the parameters used here. T.S. created and programmed RD under the supervision of IX. L.A., E.R., E.T. and P.B. optimized and produced PacBio sequencing libraries. J.M. advised on library preparation, trained ER, ran the sequencing, and generated the CCSs. E.H. analysed the GFP(CAG)$_x$ SMRT datasets. B.M. generated the MiSeq sequences together with M.C. under the supervision of D.G.M., T.M. and L.J., N.J.H.T. provided raw data for the FECD PCR-free dataset under the supervision of A.E.D. V.D., A.S.T., D.B., N.G. and I.X. designed the experiments. V.D. wrote the manuscript and all the co-authors provided input and feedback.

FUNDING

Academy of Medical Sciences Professorship [AMSPR1\1014 to V.D.]; UK Dementia Research Institute, which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK (to V.D.); Cardiff University School of Medicine Studentship (to B.M.); MRC Centre [MR/L010305/1 to L.J.]; CHDI (to D.G.M., L.J., T.H.M.); Welsh Clinical Academic Track Fellowship (to T.H.M.); MRC Clinical Research Training Fellowship [MR/P001629/1 to T.H.M.]; Patrick Berthoud Charitable Trust Fellowship, Association of British Neurologists (to T.H.M.); Brain Research Trust [201617–06 to B.M., T.M., L.J.]; UKRI Future Leader Fellowship [MR/S031820/1 to A.E.D.]; Moorfields Eye Charity PhD studentship (to N.J.H.T.).

Conflict of interest statement. LJ was on the scientific advisory boards of LoQus23 Therapeutics and Triplet Therapeutics and a member of the Executive Committee of the European Huntington's Disease Network. Within the last five years DGM has been a scientific consultant and/or received an honoraria/stock options/research contracts from AMO Pharma, Charles River, LoQus23, Small Molecule RNA, Triplet Therapeutics, and Vertex Pharmaceuticals. Within the last 5 years AED has been a scientific consultant for and/or received an honoraria/stock options/research contracts from Triplet Therapeutics, LoQus23 Therapeutics, Design Therapeutics, ProQR Therapeutics and Prime Medicine.

REFERENCES

1. Khristich,A.N. and Mirkin,S.M. (2020) On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.*, **295**, 4134–4170.
2. Bates,G.P., Dorsey,R., Gusella,J.F., Hayden,M.R., Kay,C., Leavitt,B.R., Nance,M., Ross,C.A., Scahill,R.I., Wetzel,R. *et al.* (2015) Huntington disease. *Nat. Rev. Dis. Prim.*, **1**, 15005.
3. Holmans,P.A., Massey,T.H. and Jones,L. (2017) Genetic modifiers of mendelian disease: Huntington's disease and the trinucleotide repeat disorders. *Hum. Mol. Genet.*, **26**, R83–R90.
4. McAllister,B., Gusella,J.F., Landwehrmeyer,G.B., Lee,J.-M., MacDonald,M.E., Orth,M., Rosser,A.E., Williams,N.M., Holmans,P., Jones,L. *et al.* (2021) Timing and impact of psychiatric, cognitive, and motor abnormalities in huntington disease. *Neurology*, **96**, e2395–e2406.
5. Monckton,D.G. (2021) The contribution of somatic expansion of the CAG repeat to symptomatic development in huntington's disease: a historical perspective. *J. Huntingtons. Dis.*, **10**, 7–33.
6. Lee,J.M., Correia,K., Loupe,J., Kim,K.H., Barker,D., Hong,E.P., Chao,M.J., Long,J.D., Lucente,D., Vonsattel,J.P.G. *et al.* (2019) CAG repeat not polyglutamine length determines timing of Huntington's disease onset. *Cell*, **178**, 887–900.
7. Lee,J.M., Wheeler,V.C., Chao,M.J., Vonsattel,J.P.G., Pinto,R.M., Lucente,D., Abu-Elneel,K., Ramos,E.M., Mysore,J.S., Gillis,T. *et al.* (2015) Identification of genetic factors that modify clinical onset of Huntington's disease. *Cell*, **162**, 516–526.
8. Ciosi,M., Maxwell,A., Cumming,S.A., Moss,Hensman, Alshammari,D.J., Flower,A.M., Durr,M.D., Leavitt,A., Roos,B.R., Holmans,R.A.C. *et al.*2019) A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine*, **48**, 568–580.
9. Wright,G.E.B., Collins,J.A., Kay,C., McDonald,C., Dolzhenko,E., Xia,Q., Bečanović,K., Drögemöller,B.I., Semaka,A., Nguyen,C.M. *et al.* (2019) Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. *Am. J. Hum. Genet.*, **104**, 1116–1126.
10. McAllister,B., Donaldson,J., Binda,C.S., Powell,S., Chughtai,U., Edwards,G., Stone,J., Lobanov,S., Elliston,L., Schuhmacher,L.-N. *et al.* (2022) Exome sequencing of individuals with Huntington's disease implicates FAN1 nuclease activity in slowing CAG expansion and disease onset. *Nat. Neurosci.*, **25**, 446–457.
11. Findlay Black,H., Wright,G.E.B.B., Collins,J.A., Caron,N., Kay,C., Xia,Q., Arning,L., Bijlsma,E.K., Squitieri,F., Nguyen,H.P. *et al.* (2020) Frequency of the loss of CAA interruption in the HTT CAG tract and implications for huntington disease in the reduced penetrance range. *Genet. Med.*, **22**, 2108–2113.
12. Wheeler,V.C. and Dion,V. (2021) Modifiers of CAG/CTG repeat instability: insights from mammalian models. *J. Huntingtons. Dis.*, **10**, 123–148.
13. Ashizawa,T., Dubel,J.R. and Harati,Y. (1993) Somatic instability of CTG repeat in myotonic dystrophy. *Neurology*, **43**, 2674–2678.
14. Anvret,M., Ahlberg,G., Grandell,U., Hedberg,B., Johnson,K. and Edstrōm,L. (1993) Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum. Mol. Genet.*, **2**, 1397–1400.
15. Overend,G., Légaré,C., Mathieu,J., Bouchard,L., Gagnon,C. and Monckton,D.G. (2019) Allele length of the DMPK CTG repeat is a predictor of progressive myotonic dystrophy type 1 phenotypes. *Hum. Mol. Genet.*, **28**, 2245–2254.
16. Cumming,S.A., Jimenez-Moreno,C., Okkersen,K., Wenninger,S., Daidj,F., Hogarth,F., Littleford,R., Gorman,G., Bassez,G., Schoser,B. *et al.* (2019) Genetic determinants of disease severity in the myotonic dystrophy type 1 OPTIMISTIC cohort. *Neurology*, **93**, e995–e1009.
17. Bettencourt,C., Hensman-Moss,D., Flower,M., Wiethoff,S., Brice,A., Goizet,C., Stevanin,G., Koutsis,G., Karadima,G., Panas,M. *et al.* (2016) DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann. Neurol.*, **79**, 983–990.
18. Cumming,S.A., Hamilton,M.J., Robb,Y., Gregory,H., McWilliam,C., Cooper,A., Adam,B., McGhie,J., Hamilton,G., Herzyk,P. *et al.* (2018) De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. *Eur. J. Hum. Genet.*, **26**, 1635–1647.
19. Ballester-Lopez,A., Koehorst,E., Almendrote,M., Martínez-Piñeiro,A., Lucente,G., Linares-Pardo,I., Núñez-Manchón,J., Guanyabens,N., Cano,A., Lucia,A. *et al.* (2020) A DM1 family with interruptions associated with atypical symptoms and late onset but not with a milder phenotype. *Hum. Mutat.*, **41**, 420–431.
20. Tomé,S., Dandelot,E., Dogan,C., Bertrand,A., Geneviève,D., Péréon,Y., Simon,M., Bonnefont,J.-P., Bassez,G. and Gourdon,G. (2018) Unusual association of a unique CAG interruption in 5′ of DM1 CTG repeats with intergenerational contractions and low somatic mosaicism. *Hum. Mutat.*, **39**, 970–982.
21. Pešović,J., Perić,S., Brkušanin,M., Brajušković,G., Rakočević-Stojanović,V. and Savić-Pavićević,D. (2018) Repeat

interruptions modify age at onset in myotonic dystrophy type 1 by stabilizing DMPK expansions in somatic cells. *Front. Genet.*, **9**, 601.
22. Hayward,B.E., Kumari,D. and Usdin,K. (2017) Recent advances in assays for the fragile X-related disorders. *Hum. Genet.*, **136**, 1313–1327.
23. Kraus-Perrotta,C. and Lagalwar,S. (2016) Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum Ataxias*, **3**, 20.
24. Sobczak,K. and Krzyzosiak,W.J. (2004) Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability. *Hum. Mutat.*, **24**, 236–247.
25. Chung,M.Y., Ranum,L.P.W., Duvick,L.A., Servadio,A., Zoghbi,H.Y. and Orr,H.T. (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat. Genet.*, **5**, 254–258.
26. Mangin,A., de Pontual,L., Tsai,Y.C., Monteil,L., Nizon,M., Boisseau,P., Mercier,S., Ziegle,J., Harting,J., Heiner,C. *et al.* (2021) Robust detection of somatic mosaicism and repeat interruptions by long-read targeted sequencing in myotonic dystrophy type 1. *Int. J. Mol. Sci.*, **22**, 2616.
27. Santoro,M., Masciullo,M., Silvestri,G., Novelli,G. and Botta,A. (2017) 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.
28. Tomé,S. and Gourdon,G. (2020) DM1 phenotype variability and triplet repeat instability: challenges in the development of new therapies. *Int. J. Mol. Sci.*, **21**, 457.
29. Tsai,Y.-C., de Pontual,L., Heiner,C., Stojkovic,T., Furling,D., Bassez,G., Gourdon,G. and Tomé,S. (2022) Identification of a CCG-Enriched expanded allele in patients with myotonic dystrophy type 1 using amplification-free long-read sequencing. *J. Mol. Diagnostics*, **24**, 1143–1154.
30. Stevanovski,I., Chintalaphani,S.R., Gamaarachchi,H., Ferguson,J.M., Pineda,S.S., Scriba,C.K., Tchan,M., Fung,V., Ng,K., Cortese,A. *et al.* (2022) Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci. Adv.*, **8**, eabm5386.
31. Ibañez,K., Polke,J., Hagelstrom,R.T., Dolzhenko,E., Pasko,D., Thomas,E.R.A., Daugherty,L.C., Kasperaviciute,D., Smith,K.R., Deans,Z.C. *et al.* (2022) 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.
32. Dolzhenko,E., van Vugt,J.J.F.A., Shaw,R.J., Bekritsky,M.A., van Blitterswijk,M., Narzisi,G., Ajay,S.S., Rajan,V., Lajoie,B.R., Johnson,N.H. *et al.* (2017) Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, **27**, 1895–1903.
33. Lockhart,P.J. (2022) Advancing the diagnosis of repeat expansion disorders. *Lancet Neurol.*, **21**, 205–207.
34. Chintalaphani,S.R., Pineda,S.S., Deveson,I.W. and Kumar,K.R. (2021) An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol. Commun.*, **9**, 98.
35. Ciosi,M., Cumming,S.A., Chatzi,A., Larson,E., Tottey,W., Lomeikaite,V., Hamilton,G., Wheeler,V.C., Pinto,R.M., Kwak,S. *et al.* (2021) Approaches to sequence the HTT CAG repeat expansion and quantify repeat length variation. *J. Huntingtons. Dis.*, **10**, 53–74.
36. Loomis,E.W., Eid,J.S., Peluso,P., Yin,J., Hickey,L., Rank,D., McCalmon,S., Hagerman,R.J., Tassone,F. and Hagerman,P.J. (2013) Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile x gene. *Genome Res.*, **23**, 121–128.
37. Ebbert,M.T.W., Farrugia,S.L., Sens,J.P., Jansen-West,K., Gendron,T.F., Prudencio,M., McLaughlin,I.J., Bowman,B., Seetin,M., DeJesus-Hernandez,M. *et al.* (2018) Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol. Neurodegener.*, **13**, 46.
38. Hafford-Tear,N.J., Tsai,Y.C., Sadan,A.N., Sanchez-Pintado,B., Zarouchlioti,C., Maher,G.J., Liskova,P., Tuft,S.J., Hardcastle,A.J., Clark,T.A. *et al.* (2019) CRISPR/Cas9-targeted enrichment and long-read sequencing of the fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. *Genet. Med.*, **21**, 2092–2102.
39. Höijer,I., Tsai,Y.C., Clark,T.A., Kotturi,P., Dahl,N., Stattin,E.L., Bondeson,M.L., Feuk,L., Gyllensten,U. and Ameur,A. (2018) Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. *Hum. Mutat.*, **39**, 1262–1272.
40. McFarland,K.N., Liu,J., Landrian,I., Godiska,R., Shanker,S., Yu,F., Farmerie,W.G. and Ashizawa,T. (2015) SMRT sequencing of long tandem nucleotide repeats in SCA10 reveals unique insight of repeat expansion structure. *PLoS One*, **10**, e0135906.
41. Wieben,E.D., Aleff,R.A., Basu,S., Sarangi,V., Bowman,B., McLaughlin,I.J., Mills,J.R., Butz,M.L., Highsmith,E.W., Ida,C.M. *et al.* (2019) Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in fuchs endothelial corneal dystrophy. *PLoS One*, **14**, e0219446.
42. Chiu,R., Rajan-Babu,I.-S., Friedman,J.M. and Birol,I. (2021) Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.*, **22**, 224.
43. Giesselmann,P., Brändl,B., Raimondeau,E., Bowen,R., Rohrandt,C., Tandon,R., Kretzmer,H., Assum,G., Galonska,C., Siebert,R. *et al.* (2019) Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.*, **37**, 1478–1481.
44. Rasmussen,A., Hildonen,M., Vissing,J., Duno,M., Tümer,Z. and Birkedal,U. (2022) High resolution analysis of DMPK hypermethylation and repeat interruptions in myotonic dystrophy type 1. *Genes*, **13**, 970.
45. Fang,L., Monteys,A.M., Dürr,A., Keiser,M., Cheng,C., Harapanahalli,A., Gonzalez-Alegre,P., Davidson,B.L. and Wang,K. (2023) Haplotyping SNPs for allele-specific gene editing of the expanded huntingtin allele using long-read sequencing. *Hum. Genet. Genomics Adv.*, **4**, 100146.
46. Liu,Q., Zhang,P., Wang,D., Gu,W. and Wang,K. (2017) Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.*, **9**, 65.
47. Ummat,A. and Bashir,A. (2014) Resolving complex tandem repeats with long reads. *Bioinformatics*, **30**, 3491–3498.
48. Mitsuhashi,S., Frith,M.C., Mizuguchi,T., Miyatake,S., Toyota,T., Adachi,H., Oma,Y., Kino,Y., Mitsuhashi,H. and Matsumoto,N. (2019) Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.*, **20**, 58.
49. DeJesus-Hernandez,M., Aleff,R.A., Jackson,J.L., Finch,N.A., Baker,M.C., Gendron,T.F., Murray,M.E., McLaughlin,I.J., Harting,J.R., Graff-Radford,N.R. *et al.* (2021) Long-read targeted sequencing uncovers clinicopathological associations for C9orf72-linked diseases. *Brain*, **144**, 1082–1088.
50. Cinesi,C., Yang,B. and Dion,V. (2020) GFP reporters to monitor instability and expression of expanded CAG/CTG repeats. *Methods Mol. Biol.*, **2056**, 255–268.
51. Cinesi,C., Aeschbach,L., Yang,B. and Dion,V. (2016) Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. *Nat. Commun.*, **7**, 13272.
52. Orth,M., Handley,O.J., Schwenke,C., Dunnett,S.B., Craufurd,D., Ho,A.K., Wild,E., Tabrizi,S.J. and Landwehrmeyer,G.B. (2010) Observing huntington's disease: the european Huntington's disease network's REGISTRY. *PLoS Curr.*, **2**, RRN1184.
53. Ciosi,M., Cumming,S., Alshammari,A., Symeonidi,E., Herzyk,P., McGuinness,D., Galbraith,J., Hamilton,G. and Monckton,D. (2020) Library preparation and miseq sequencing for the genotyping-by-sequencing of the huntington disease HTT exon one trinucleotide repeat and the quantification of somatic mosaicism. *Protoc. Exch.*, https://doi.org/10.21203/rs.2.1581/v2.
54. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.-C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
55. Hamada,M., Ono,Y., Asai,K. and Frith,M.C. (2017) Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics*, **33**, 926–928.
56. Lüthy,R., Xenarios,I. and Bucher,P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci.*, **3**, 139–146.
57. Schuepbach,T., Pagni,M., Bridge,A., Bougueleret,L., Xenarios,I. and Cerutti,L. (2013) pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. *Bioinformatics*, **29**, 1215–1217.

58. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
59. Santillan,B.A., Moye,C., Mittelman,D. and Wilson,J.H. (2014) GFP-Based fluorescence assay for CAG repeat instability in cultured human cells. *PLoS One*, **9**, e113952.
60. Ruiz Buendía,G.A., Leleu,M., Marzetta,F., Vanzan,L., Tan,J.Y., Ythier,V., Randall,E.L., Marques,A.C., Baubec,T., Murr,R. *et al.* (2020) Three-dimensional chromatin interactions remain stable upon CAG/CTG repeat expansion. *Sci. Adv.*, **6**, eaaz4012.
61. Malbec,R., Chami,B., Aeschbach,L., Ruiz Buendía,G.A., Socol,M., Joseph,P., Leïchlé,T., Trofimenko,E., Bancaud,A. and Dion,V. (2019) μLAS: sizing of expanded trinucleotide repeats with femtomolar sensitivity in less than 5 minutes. *Sci. Rep.*, **9**, 23.
62. Gilpatrick,T., Lee,I., Graham,J.E., Raimondeau,E., Bowen,R., Heron,A., Downs,B., Sukumar,S., Sedlazeck,F.J. and Timp,W. (2020) Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.*, **38**, 433–438.
63. Wieben,E.D., Aleff,R.A., Tosakulwong,N., Butz,M.L., Highsmith,W.E., Edwards,A.O. and Baratz,K.H. (2012) A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts fuchs corneal dystrophy. *PLoS One*, **7**, e49083.
64. Monckton,D.G., Wong,L.J.C., Ashizawa,T. and Caskey,C.T. (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet.*, **4**, 1–8.
65. Tomé,S. and Gourdon,G. (2020) Fast assays to detect interruptions in CTG.CAG repeat expansions. *Methods Mol. Biol.*, **2056**, 11–23.