Repeat Detector: accurate, efficient, and flexible sizing of expanded CAG/CTG repeats from targeted DNA sequencing

Alysha S. Taylor[1*], Dinis Barros[2*], Nastassia Gobet[2*], Thierry Schuepbach[3*], Branduff McAllister[4,5], Lorene Aeschbach[2], Emma L. Randall[1], Evgeniya Trofimenko[2,6], Eleanor R. Heuchan[1], Paula Barszcz[2], Marc Ciosi[7], Joanne Morgan[4], Nathaniel J. Hafford-Tear[8], Alice E. Davidson[8], Thomas H. Massey[4], Darren G. Monckton[7], Lesley Jones[4], REGISTRY Investigators of the European Huntington's disease network[9], Ioannis Xenarios[2,10], and Vincent Dion[1,11].

Affiliations

1. UK Dementia Research Institute at 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. MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK
5. Current address: Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA.
6. Current address: Sorbonne Université, École normale supérieure, PSL University, CNRS, Laboratoire des biomolécules, LBM, 75005 Paris, France.
7. Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, Davidson Building, University of Glasgow, Glasgow, G12 8QQ, UK
8. UCL Institute of Ophthalmology, 11-43 Bath Street, London, EC1V 9EL UK
9. http://www.ehdn.org/wp-content/uploads/REGISTRY-contributors-full-list.pdf
10. Health2030 Genome Center, Ch des Mines 14, 1202 Genève, Switzerland
11. Correspondence: dionv@cardiff.ac.uk

*: These authors contributed equally to this work.

Keywords: Expanded repeat disorders, long-read sequencing, repeat instability, Huntington's disease, Fuchs endothelial corneal dystrophy.

Abstract:

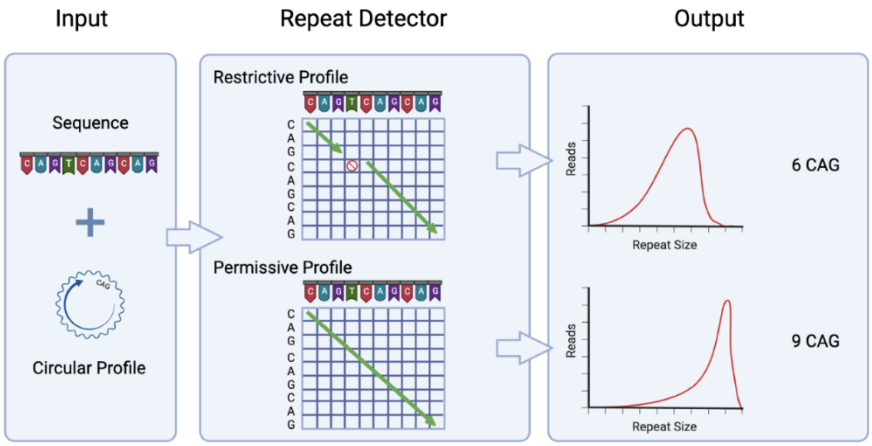
Targeted DNA sequencing approaches will improve how the size of short tandem repeats is measured for diagnostic tests and pre-clinical studies. The expansion of these sequences causes dozens of disorders, with longer tracts generally leading to a more severe disease. In addition, interruptions are sometimes present within repeats and can alter disease manifestation. Despite advances in methodologies, determining repeat size and identifying interruptions in targeted sequencing datasets remains a major challenge. This is because standard alignment tools are ill-suited for the repetitive nature of these 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 (HD) 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 HD blood-derived samples and did not require prior knowledge of the flanking sequences or their polymorphisms within the patient population. We demonstrate that RD can be used to identify individuals with repeat interruptions and may 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 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 characterised 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 importance of other factors contributing to disease pathology. One such factor is likely to be the level of repeat expansion that occurs in affected somatic tissues [5]. This is underscored by the number of genes implicated in repeat instability that also appear to modify age at disease onset [6,7]. Moreover, when repeats are interrupted, the repeat tract is stabilised and correlates with a later appearance of disease symptoms [6,8]. Thus, if ongoing repeat expansion in somatic tissues contributes to disease phenotypes, gains or losses of interruptions should lead to changes in the age at disease onset. 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 [9]. Alleles without this CAACAG interruption are associated with an earlier onset than predicted based on their repeat size [6,9–12], whereas those with two CAACAG were either found to have no effect on the age at disease onset [6,9] or were associated with a later onset [11].

Both somatic expansion and repeat interruptions appear to influence disease outcome in other expanded repeat disorders [13]. For example, somatic expansion is seen in affected tissues in myotonic dystrophy [14–17]. Moreover, some of the genetic modifiers of HD implicated in repeat expansion may also modify disease onset in other repeat disorders [18]. 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 [16,17,27–29,19–26].

Interruptions are difficult to find using current PCR-based diagnostic tools [24,30], whereas repeat instability is not typically measured. The advent of high-throughput sequencing offers an opportunity to improve the throughput and the accuracy of repeat sizing as well as the identification of interrupted alleles. Targeted sequencing of expanded repeats has been achieved with MiSeq [9,11,31], Single-Molecule, Real Time (SMRT) [24,27,30–37], and Oxford Nanopore Technology MinION [33,38,39], revealing sequence details within patient populations. One remaining bottleneck 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 either aim to correct for systematic errors generated by the sequencing platform used [40,41] or rely on intensive alignments to a large number of reference sequences with a variety of repeat sizes and flanking sequences, which are thus limited to a single locus [9,34,42]. Here we have developed Repeat Detector (RD), a fast



algorithm applicable to targeted sequencing data from multiple loci and sequencing platforms.

***Fig. 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.*

Implementation:

Repeat Detector

RD (Fig. 1) is based on the deterministic profile weighting algorithm, pfsearchV3.0, which has so far only been used for protein motifs and domain detection [43,44]. The concept of RD is similar to that of algorithms that align sequences to a library of references with multiple repeat sizes. However, given the repetitive nature of the sequence, there is redundancy in the alignment. RD takes advantage of this redundancy by aligning each sequence to a circular profile of the repeated motif of interest. Thus, instead of making pairwise comparisons to every reference sequence in the library, RD uses the circular representation of the repeat tract to account for the score of the previous alignments. 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.

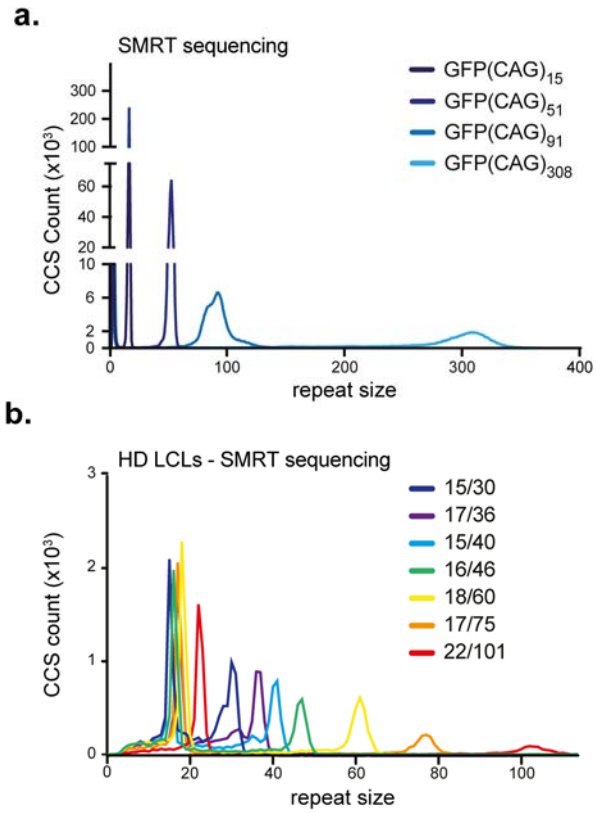
Results:

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 [45]. Our first dataset consisted of CCSs from HEK293-derived cell lines with 15, 51, 91, and 308 CAG/CTG repeats inserted at a hemizygous ectopic GFP reporter on chromosome 12 [46–48]. 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 characterised GFP$(CAG)_{101}$ line (see methods and [48]). The second dataset was composed of 21 DNA samples and lymphoblastoid cell (LBC) lines from HD individuals obtained from the Coriell BioRepository with repeats ranging from 15 to 750 units (Table S1). Taking both datasets together, we recovered the expected repeat sizes based on Coriell's data or our prior work [49], except for one sample (Fig. 2ab). Only the sample with the longest repeat tract, GM14044, which we have shown to contain 750 repeats [49], 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 shorter than 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, the sequencing reaction, 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 [30,31].

***Fig. 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 Fig. S1. Read depth and mapping metrics for all datasets can be found in Table S3.*



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 [11]. 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 (Table S2). For 42 individuals, there are data for both blood and LBC. Hereafter, we refer to this dataset as the HD MiSeq dataset since it was generated using Illumina MiSeq technology [11]. This dataset was originally analysed for modal repeat size and flanking sequences using ScaleHD [50]. 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. Of the 649 samples, we analysed 609 with both algorithms, totalling 1218 germline alleles (Fig. 3a). For the shorter alleles, the modal repeat size was determined to be the same with both software (Fig. 3b). Of the longer alleles, 599 out of 609 (98.3%) had the same modal allele size (Fig. 3b). Of the remaining 10 alleles, 9 differed by one CAG and one allele by two (Fig. S2). One of these differences came from a homozygous individual with 2 alleles of 15 repeats. 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 (Fig. 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 (Fig. 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 in this sample because they did not align well with the library of reference sequences and were discarded. It is unclear why this occurred. RD does not rely on an alignment to the locus of interest and thus counted both alleles accurately (Fig. S3b). These data highlight the accuracy of RD and shows that it is comparable to ScaleHD for the *HTT* locus.

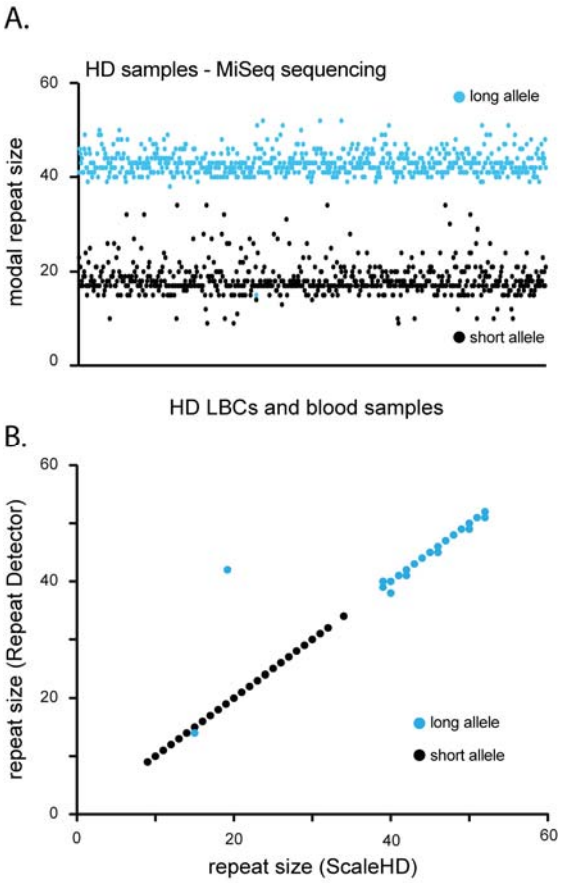


Fig. 3: ***Repeat Detector is highly accurate on HD samples.*** *A) Modal repeat size in the HD MiSeq dataset determined by RD using the restrictive parameters. Each dot is an allele. Blue dots are the longer of the two alleles in a sample, whereas black dots are the shorter alleles. Read depth and mapping metrics for all datasets can be found in Table S3. B) modal repeat size in the HD MiSeq samples comparing ScaleHD and Repeat detector.*

RD applied to multiple different repeat motifs

To test the applicability of RD to other repeat compositions, we analysed publicly available datasets generated using PCR-free libraries for SMRT [34,35] and MinION [38,51] 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 (Fig. S4). However, with the MinION sequencing data containing expanded GGGGCC repeats [38], RD dramatically underestimated the repeat size (Fig. S5a). Upon visual inspection of the MinION sequencing reads, we found that the expected repeat motif was too often mutated to be reliably detected (Fig. S5b). This is consistent with Ebert *et al.* [33], who found that when generating whole genome sequences using MinION there was no read aligning to the GGGGCC repeat at the *C9orf72* locus. We conclude that RD is limited by the error rate of the sequencing method used, but is otherwise applicable to datasets regardless of the library preparation protocol, the locus sequenced, or the repeated motif.

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 larger repeat tracts showing more size heterogeneity (Fig. 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 [9,31]. 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) [52] 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 (Fig. 4) [34]. We found that we could reproduce, for all samples, the modal repeat size determined previously (Table 1). In addition, repeat instability was obvious with expansion-biased mosaicism, especially for longer alleles (Table 1, Fig. S6). We found that RD was largely in agreement with previous studies by Hafford-Tear *et al.* [34] in determining the

largest repeat tract present in a sample. In one case, however, RD found a repeat size nearly 1300 units larger than previously identified. This may be because RD does not require an alignment prior to counting repeats and larger repeat sizes are expected to align less well to a reference sequence. Together these results suggest that RD may be used to determine the frequency of repeat instability, in addition to modal repeat size.

***Fig. 4:* RD exposes repeat instability in amplification-free datasets.** *Repeat size distribution of the CTG repeat found within a FECD patient cohort from reference* [34] *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 Fig. S6. Read depth and mapping metrics for all datasets can be found in Table S3.*

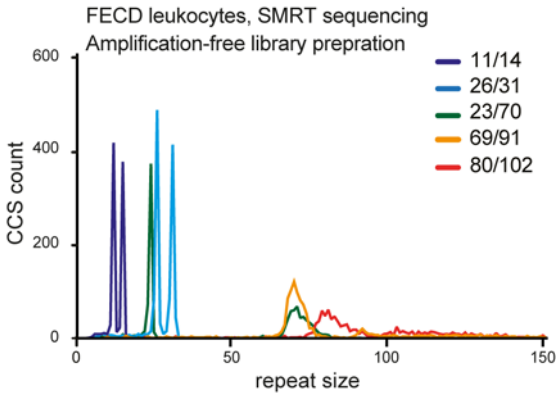


Table 1: *Comparison between Repeat Detector and previously published data for the FECD SMRT dataset.*

	Modal Repeat Size*		Largest Repeat Tract	
Sample	**Shorter allele**	**Larger allele**	**Repeat Detector§**	**Hafford-Tear *et al.*** [34]
1	11	14	15	Not determined
2	25	30	37	37
3	23	70	90	90
4	23	73	115	115
5	11	80	169	169
6	32	110	1875	566
7	9	131	1381	1361
8	80	102	498	498
9	72	118	1285	1593
10	69	91	1047	1014
11	79	141	1581	Not determined

**Modal Repeat Size was determined using the restrictive parameters of RD.*
§The restrictive parameters were used to determined repeat size.

Identifying individuals with interruptions using RD

When optimising 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 ones 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 3 repeats between the permissive and restrictive parameters (Fig. 5ab). By contrast, alleles without the interruption yield only one repeat difference between the two profiles (Fig. 5ab) and the ones with a duplicated CAACAG motif show 5 units differences (Fig. 5ac). The differences can be used to identify samples with repeat interruptions or unusual allele structures and narrow down which samples need to be inspected manually. Using this approach, we could accurately identify the sole sample in the HD MiSeq dataset with a CAC interruption within its CAG repeat (Fig. 5d) and the interrupted non-pathogenic allele in Sample 7 of the FECD dataset (Fig. 5e). We could also identify a previously unknown 111bp insertion in the GFP$(CAG)_{308}$ cell line (Fig. 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.

Discussion:

Here we applied RD, which accurately detects and counts short tandem repeats in targeted sequencing data. 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.

RD could detect repeat instability in HD and FECD blood-derived samples prepared with a PCR-based or amplification-free protocol, respectively. At the *TCF4* locus where PCR artefacts 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 [53]. Our data, together with that of a recent pre-print on DM1 [30], 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.

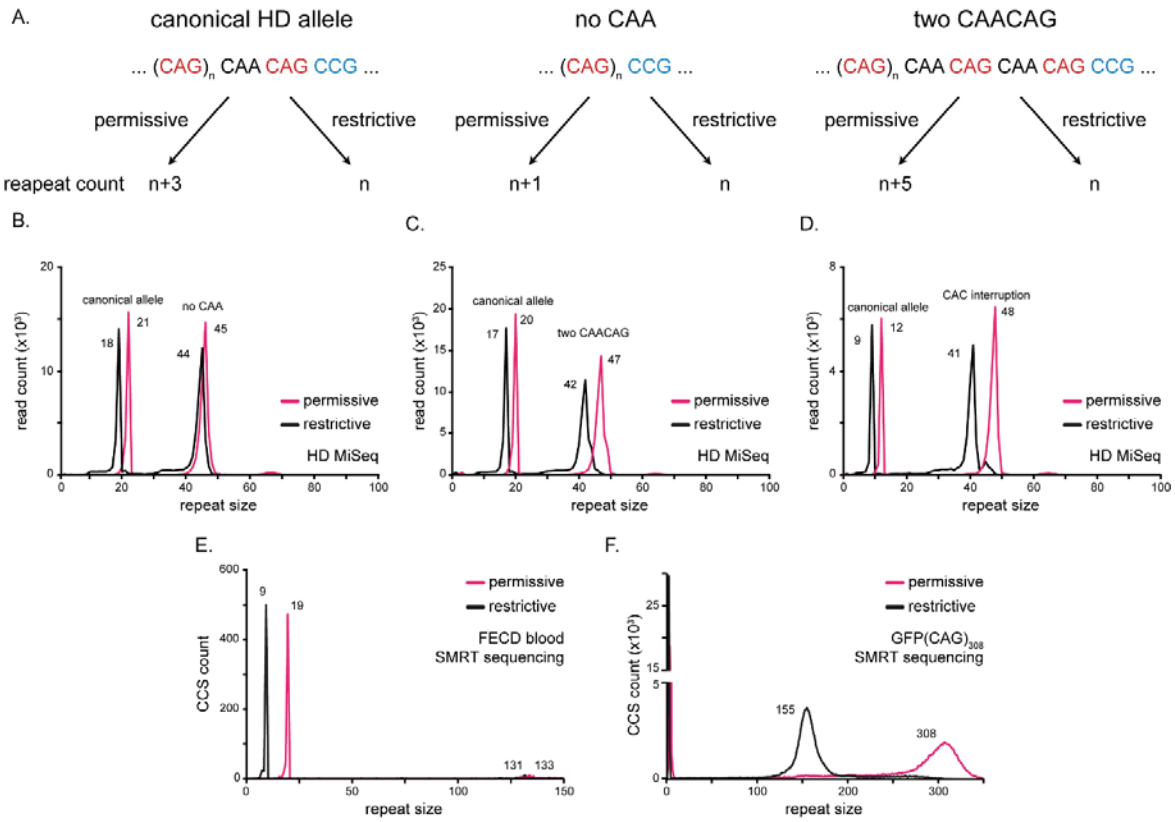


Fig. 5: ***Identifying samples with interruptions using RD.*** *A) Interruptions at the 3' end of HD alleles can be distinguished using the difference in repeat size between RD's permissive and restrictive parameters. For instance, the most common allele (left), containing a CAA interruption will return 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 samples 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* [34] *was correctly identified. F) Our GFP(CAG)$_{308}$ line was found to have an insertion of 111bp after 155 CAG repeats.*

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 [24,54]. This

leads to a pattern on capillary electrophoresis with a periodicity the size of the repeated unit and of decaying intensity. Interruptions in the intensity traces appear 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. 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 requires user intervention to identify the nature of the interruption detected in a sample and cannot discriminate between a single or 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 [24,30]. 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 size of the flanking repeated motif. 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. Finally, RD ignores flanking sequences and thus would miss, for example, the significant polymorphism found downstream of the HD allele [9]. 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.

Conclusion:

RD accurately determines repeat size from targeted sequencing data from SMRT, MiSeq, and potentially MinION sequencing platforms. It is not limited by 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.

Methods:

Cell culture and cell lines

The GFP(CAG)$_x$ cell lines were cultured as described before [48,55]. 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 LBCs or their DNA used for SMRT sequencing were obtained from the Coriell BioRepository (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)$_{101}$ [48] 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 [48] 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 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 Table S4) 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 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 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 [56] (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). Subjects were selected as in [11].

HD MiSeq dataset

A total of 652 DNA samples were sequenced, with the majority of these being immortalised 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 [31,50]. Of note, the method includes a size selection step that biases towards longer alleles. 649 of the original 652 samples were successfully sequenced (>99%). Table S2 describes the numbers of each sample as well as the numbers of each DNA type that was successfully sequenced. For comparisons between ScaleHD and RD presented in Fig. 3B, Fig. S2, and Fig. 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 differ between reads from the same sample. These are due to PCR and sequencing errors. The ScaleHD parameters are as done previously [11].

FECD SMRT dataset

The FECD SMRT dataset is a amplification-free SMRT sequencing dataset from blood samples of FECD patients published previously [34].

Repeat Detector

Repeat Detector source code and dependencies will be found here upon publication: https://github.com/DionLab/RepeatDetector. To determine repeat sizes for GFP$(CAG)_x$, FECD [34], HD MiSeq [11] and c9orf72 loci [38] datasets, unaligned reads were assessed using permissive and restrictive profiles with a range of [0-1000]. For each analysis, the --with-revcomp option was enabled and data was output to a density plot (-o histogram option).

Density plots obtained were graphed using GraphPad PRISM version 9. Observed repeat distribution for *C9ORF72* locus data sequenced using Oxford Nanopore MinION technology were compared against an expected distribution modelled with the PertDist sciPy package version 0.1.1 (https://github.com/rpouillot/PertDist), using the range [0-1000] given to Repeat Detector and the repeat size obtained by Giesselmann et al [38] as the mode.

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.

Conflict of interests:

LJ is 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.

Data availability:

The source code and dependencies are available here: https://github.com/DionLab/RepeatDetector. The HD SMRT and GFP$(CAG)_x$ SMRT datasets will be publicly available by the time of publication. The GFP$(CAG)_x$ lines are available upon request from the corresponding author.

Author contributions:

AST 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 VD, DB, and EH. DB optimised 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. NG tested the beta version of RD and optimised the parameters used here. TS created and programmed RD under the supervision of IX. LA, ER, ET, and PB optimised and produced PacBio sequencing libraries. JM advised on library preparation, trained ER, ran the sequencing, and generated the CCSs. EH analysed the GFP$(CAG)_x$ SMRT datasets. BM generated the MiSeq sequences together with MC under the supervision of DGM, TM, and LJ. NJHT provided raw data for the FECD PCR-free dataset under the supervision of AED. VD, AST, DB, NG, and IX designed the experiments. VD wrote the manuscript and all the co-authors provided input and feedback.

Funding:

The Dion lab is funded by a Professorship from the Academy of Medical Sciences (AMSPR1\1014) and by the 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. BM received a Cardiff University School of Medicine Studentship. LJ was supported by an MRC Centre grant (MR/L010305/1); DGM, MC, LJ and THM by CHDI. THM was supported by a Welsh Clinical Academic Track Fellowship, an MRC Clinical Training Fellowship (MR/P001629/1) and a Patrick Berthoud Charitable Trust Fellowship through the Association of British Neurologists. BM, TM and LJ were supported by the Brain Research Trust (201617-06).

References:

1. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J. Biol. Chem. American Society for Biochemistry and Molecular Biology Inc.; 2020. p. 4134–70.

2. Bates GP, Dorsey R, Gusella JF, Hayden MR, Kay C, Leavitt BR, et al. Huntington disease. Nat Rev Dis Prim. 2015;1:15005.

3. Holmans PA, Massey TH, Jones L. Genetic modifiers of Mendelian disease: Huntington's disease and the trinucleotide repeat disorders. Hum. Mol. Genet. 2017. p. R83–90.

4. McAllister B, Gusella JF, Landwehrmeyer GB, Lee J-M, MacDonald ME, Orth M, et al. Timing and Impact of Psychiatric, Cognitive, and Motor Abnormalities in Huntington Disease. Neurology. Neurology; 2021;10.1212/WNL.0000000000011893.

5. Monckton DG. The Contribution of Somatic Expansion of the CAG Repeat to Symptomatic Development in Huntington's Disease: A Historical Perspective. J Huntingtons Dis. IOS Press; 2021;10:7–33.

6. Lee JM, Correia K, Loupe J, Kim KH, Barker D, Hong EP, et al. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell. Cell Press; 2019;178:887-900.e14.

7. Lee JM, Wheeler VC, Chao MJ, Vonsattel JPG, Pinto RM, Lucente D, et al. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell. 2015;162:516–26.

8. Gu X, Richman J, Langfelder P, Wang N, Zhang S, Bañez-Coronel M, et al. Uninterrupted CAG repeat drives striatum-selective transcriptionopathy and nuclear pathogenesis in human Huntingtin BAC mice. Neuron. 2022;

9. Ciosi M, Maxwell A, Cumming SA, Hensman Moss DJ, Alshammari AM, Flower MD, et al. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. EBioMedicine. Elsevier B.V.; 2019;48:568–80.

10. Wright GEB, Collins JA, Kay C, McDonald C, Dolzhenko E, Xia Q, et al. Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. Am J Hum Genet. Cell Press; 2019;104:1116–26.

11. McAllister B, Donaldson J, Binda CS, Powell S, Chughtai U, Edwards G, et al. FAN1 nuclease activity affects CAG expansion and age at onset of Huntington's disease. bioRxiv. Cold Spring Harbor Laboratory; 2021;2021.04.13.439716.

12. Findlay Black H, Wright GEBB, Collins JA, Caron N, Kay C, Xia Q, et al. Frequency of the loss of CAA interruption in the HTT CAG tract and implications for Huntington disease in the reduced penetrance range. Genet Med. Springer Nature; 2020;22:2108–13.

13. Wheeler VC, Dion V. Modifiers of CAG/CTG Repeat Instability: Insights from Mammalian Models. Jones L, Pearson CE, Wheeler V, editors. J Huntingtons Dis. IOS Press; 2021;10:123–48.

14. Ashizawa T, Dubel JR, Harati Y. Somatic instability of CTG repeat in myotonic dystrophy. Neurology. United States; 1993;43:2674–8.

15. Anvret M, Ahlberg G, Grandell U, Hedberg B, Johnson K, Edstrōm L. Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. Hum Mol Genet. 1993;2:1397–400.

16. Overend G, Légaré C, Mathieu J, Bouchard L, Gagnon C, Monckton DG. Allele length of the DMPK CTG repeat is a predictor of progressive myotonic dystrophy type 1 phenotypes. Hum Mol Genet. 2019;28:2245–54.

17. Cumming SA, Jimenez-Moreno C, Okkersen K, Wenninger S, Daidj F, Hogarth F, et al. Genetic determinants of disease severity in the myotonic dystrophy type 1 OPTIMISTIC cohort. Neurology. 2019;93:e995 LP-e1009.

18. Bettencourt C, Hensman-Moss D, Flower M, Wiethoff S, Brice A, Goizet C, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol. John Wiley and Sons Inc.; 2016;79:983–90.

19. Pešović J, Perić S, Brkušanin M, Brajušković G, Rakočević-Stojanović V, Savić-Pavićević D.

Repeat Interruptions Modify Age at Onset in Myotonic Dystrophy Type 1 by Stabilizing DMPK Expansions in Somatic Cells. Front Genet. 2018;9.

20. Hayward BE, Kumari D, Usdin K. Recent advances in assays for the fragile X-related disorders. Hum. Genet. Springer Verlag; 2017. p. 1313–27.

21. Kraus-Perrotta C, Lagalwar S. Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. Cerebellum & Ataxias. Springer Nature; 2016;3:20.

22. Sobczak K, Krzyzosiak WJ. Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability. Hum Mutat. Hum Mutat; 2004;24:236–47.

23. Chung MY, Ranum LPW, Duvick LA, Servadio A, Zoghbi HY, Orr HT. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nat Genet. Nat Genet; 1993;5:254–8.

24. Mangin A, de Pontual L, Tsai YC, Monteil L, Nizon M, Boisseau P, et al. Robust detection of somatic mosaicism and repeat interruptions by long-read targeted sequencing in myotonic dystrophy type 1. Int J Mol Sci. MDPI AG; 2021;22:1–24.

25. 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. Blackwell Publishing Ltd; 2017. p. 355–64.

26. Tomé S, Gourdon G. DM1 Phenotype Variability and Triplet Repeat Instability: Challenges in the Development of New Therapies. Int J Mol Sci. MDPI AG; 2020;21:457.

27. Cumming SA, Hamilton MJ, Robb Y, Gregory H, McWilliam C, Cooper A, et al. 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. Nature Publishing Group; 2018;26:1635–47.

28. Ballester-Lopez A, Koehorst E, Almendrote M, Martínez-Piñeiro A, Lucente G, Linares-Pardo I, et al. A DM1 family with interruptions associated with atypical symptoms and late onset but

not with a milder phenotype. Hum Mutat. John Wiley and Sons Inc.; 2020;41:420–31.

29. Tomé S, Dandelot E, Dogan C, Bertrand A, Geneviève D, Péréon Y, et al. Unusual association of a unique CAG interruption in 5′ of DM1 CTG repeats with intergenerational contractions and low somatic mosaicism. Hum Mutat. John Wiley and Sons Inc.; 2018;39:970–82.

30. Tsai Y-C, de Pontual L, Heiner C, Stojkovic T, Furling D, Bassez G, et al. Identification of a CCG-enriched expanded allele in DM1 patients using Amplification-free long-read sequencing. bioRxiv. 2022;2022.02.22.481438.

31. Ciosi M, Cumming SA, Chatzi A, Larson E, Tottey W, Lomeikaite V, et al. Approaches to Sequence the HTT CAG Repeat Expansion and Quantify Repeat Length Variation. Jones L, Pearson CE, Wheeler V, editors. J Huntingtons Dis. 2021;10:53–74.

32. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Res. Genome Res; 2013;23:121–8.

33. Ebbert MTW, Farrugia SL, Sens JP, Jansen-West K, Gendron TF, Prudencio M, et al. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. Mol Neurodegener. BioMed Central Ltd.; 2018;13:46.

34. Hafford-Tear NJ, Tsai YC, Sadan AN, Sanchez-Pintado B, Zarouchlioti C, Maher GJ, et al. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. Genet Med. Nature Publishing Group; 2019;21:2092–102.

35. Höijer I, Tsai YC, Clark TA, Kotturi P, Dahl N, Stattin EL, et al. Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. Hum Mutat. John Wiley and Sons Inc.; 2018;39:1262–72.

36. McFarland KN, Liu J, Landrian I, Godiska R, Shanker S, Yu F, et al. SMRT Sequencing of Long Tandem Nucleotide Repeats in SCA10 Reveals Unique Insight of Repeat Expansion Structure. Boden M, editor. PLoS One. Public Library of Science; 2015;10:e0135906.

37. Wieben ED, Aleff RA, Basu S, Sarangi V, Bowman B, McLaughlin IJ, et al. Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy. Lewin AS, editor. PLoS One. Public Library of Science; 2019;14:e0219446.

38. Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol. Nature Research; 2019;37:1478–81.

39. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. United States; 2022;8:eabm5386.

40. Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. Genome Med. BioMed Central Ltd.; 2017;9:65.

41. Ummat A, Bashir A. Resolving complex tandem repeats with long reads. Bioinformatics. Oxford University Press; 2014;30:3491–8.

42. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019;20:58.

43. Lüthy R, Xenarios I, Bucher P. Improving the sensitivity of the sequence profile method. Protein Sci. 1994;3:139–46.

44. Schuepbach T, Pagni M, Bridge A, Bougueleret L, Xenarios I, Cerutti L. pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. Bioinformatics. 2013;29:1215–7.

45. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. American Association for the Advancement of Science; 2009;323:133–8.

46. Santillan BA, Moye C, Mittelman D, Wilson JH. GFP-Based Fluorescence Assay for CAG Repeat Instability in Cultured Human Cells. Magdinier F, editor. PLoS One. Public Library of

Science; 2014;9:e113952.

47. Ruiz Buendía GA, Leleu M, Marzetta F, Vanzan L, Tan JY, Ythier V, et al. Three-dimensional chromatin interactions remain stable upon CAG/CTG repeat expansion. Sci Adv. American Association for the Advancement of Science; 2020;6:eaaz4012.

48. Cinesi C, Aeschbach L, Yang B, Dion V. Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. Nat Commun. Nature Publishing Group; 2016;7:13272.

49. Malbec R, Chami B, Aeschbach L, Ruiz Buendía GA, Socol M, Joseph P, et al. µLAS: Sizing of expanded trinucleotide repeats with femtomolar sensitivity in less than 5 minutes. Sci Rep. Nature Publishing Group; 2019;9:23.

50. Ciosi M, Cumming S, Alshammari A, Symeonidi E, Herzyk P, McGuinness D, et al. 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. 2020;

51. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. Nature Research; 2020;38:433–8.

52. Wieben ED, Aleff RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards AO, et al. A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. Lewin A, editor. PLoS One. Public Library of Science; 2012;7:e49083.

53. Monckton DG, Wong LJC, Ashizawa T, Caskey CT. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. Hum Mol Genet. England; 1995;4:1–8.

54. Tomé S, Gourdon G. Fast Assays to Detect Interruptions in CTG.CAG Repeat Expansions. Methods Mol Biol. United States; 2020;2056:11–23.

55. Cinesi C, Yang B, Dion V. GFP reporters to monitor instability and expression of expanded CAG/CTG repeats. Methods Mol Biol. United States: Humana Press Inc.; 2020;2056:255–68.

56. Orth M, Handley OJ, Schwenke C, Dunnett SB, Craufurd D, Ho AK, et al. Observing Huntington's Disease: the European Huntington's Disease Network's REGISTRY. PLoS Curr. United States; 2010;2:RRN1184.

Repeat Detector: accurate, efficient, and flexible sizing of expanded CAG/CTG repeats from targeted DNA sequencing

Alysha S. Taylor[1*], Dinis Barros[2*], Nastassia Gobet[2*], Thierry Schuepbach[3*], Branduff McAllister[4,5], Lorene Aeschbach[2], Emma L. Randall[1], Evgeniya Trofimenko[2,6], Eleanor R. Heuchan[1], Paula Barszcz[2], Marc Ciosi[7], Joanne Morgan[4], Nathaniel J. Hafford-Tear[8], Alice E. Davidson[8], Thomas H. Massey[4], Darren G. Monckton[7], Lesley Jones[4], REGISTRY Investigators of the European Huntington's disease network[9], Ioannis Xenarios[2,10], and Vincent Dion[1,11].

Affiliations

1. UK Dementia Research Institute at 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. MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK
5. Current address: Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA.
6. Current address: Sorbonne Université, École normale supérieure, PSL University, CNRS, Laboratoire des biomolécules, LBM, 75005 Paris, France.
7. Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, Davidson Building, University of Glasgow, Glasgow, G12 8QQ, UK
8. UCL Institute of Ophthalmology, 11-43 Bath Street, London, EC1V 9EL UK
9. http://www.ehdn.org/wp-content/uploads/REGISTRY-contributors-full-list.pdf
10. Health2030 Genome Center, Ch des Mines 14, 1202 Genève, Switzerland
11. Correspondence: dionv@cardiff.ac.uk

*: These authors contributed equally to this work.

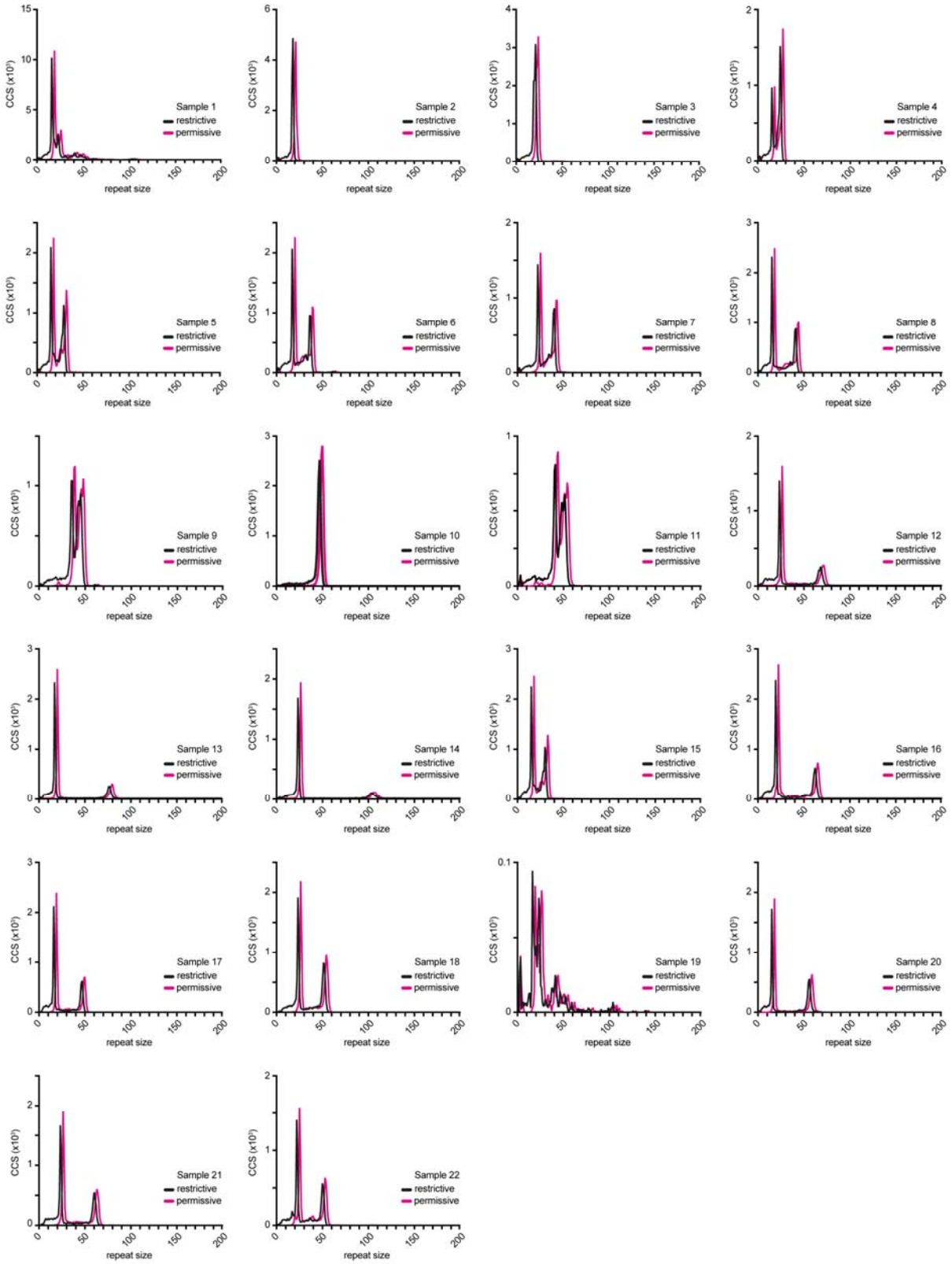


Fig. S1: *RD density plots for all the samples in the HD SMRT dataset. Sample information can be found in Table S1.*

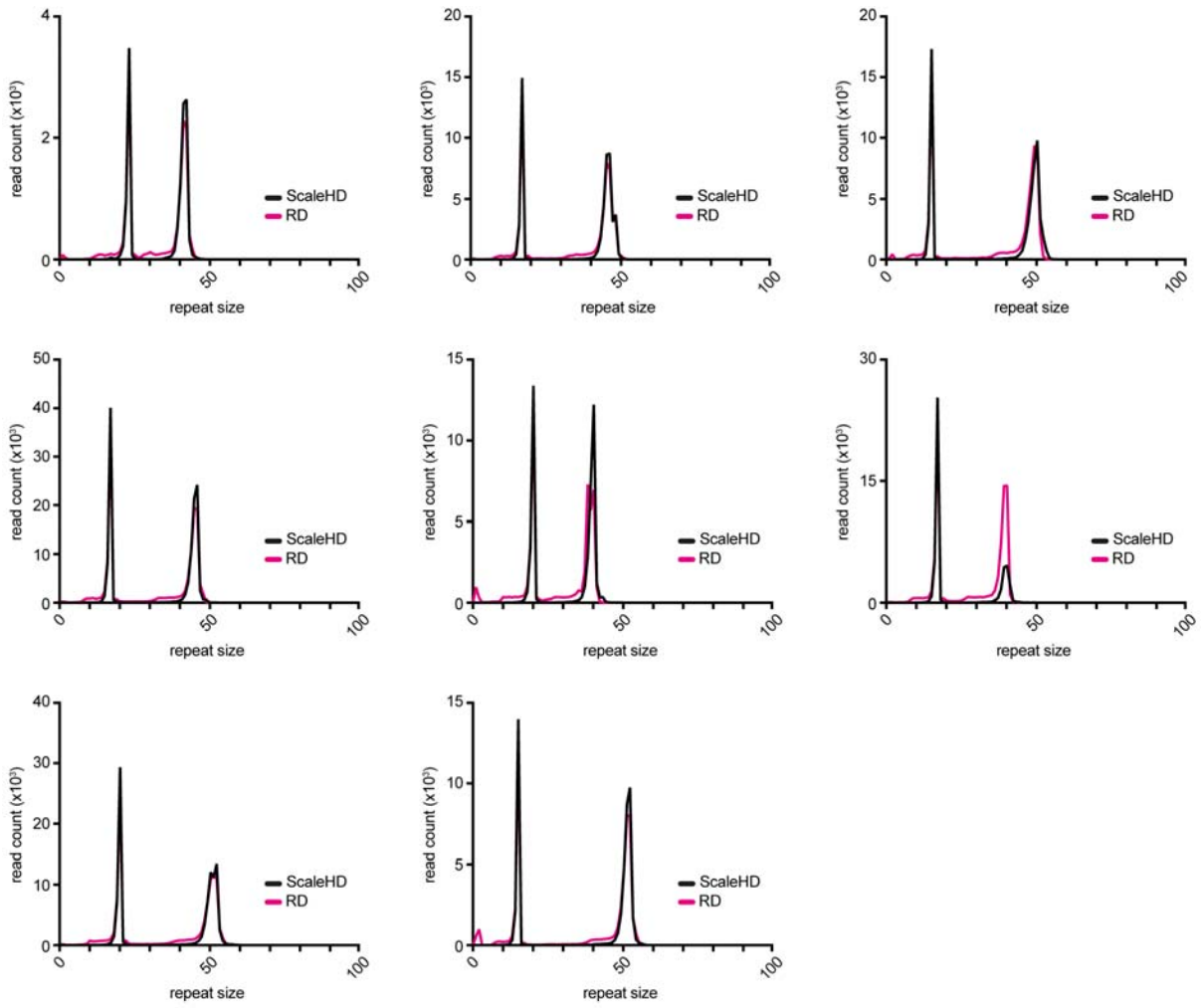


Fig. S2: *Comparison of RD and ScaleHD on 9 samples that differed by one or two repeats. For RD, we used the restrictive parameters.*

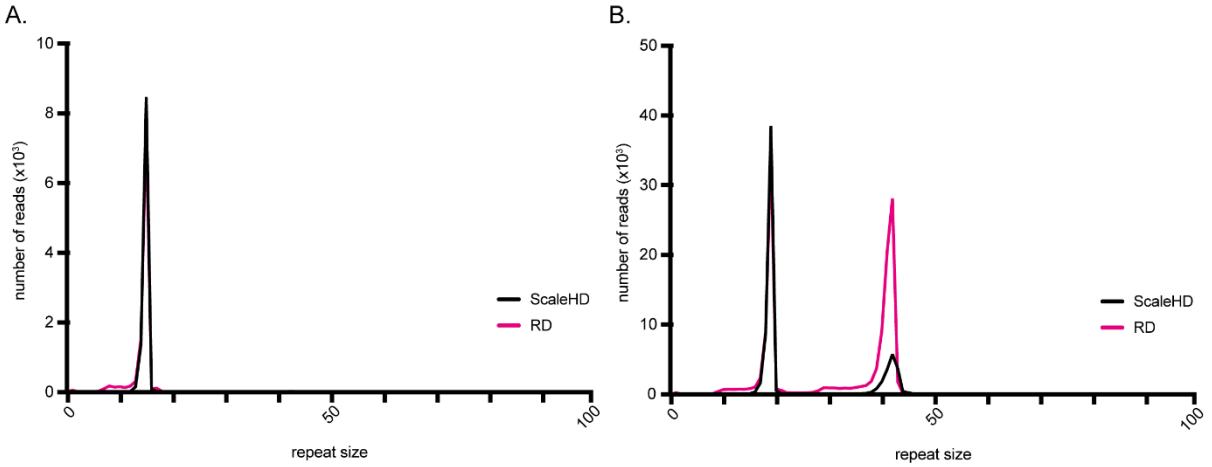


Fig. S3: *Density plot of samples that differed in their modal repeat count between ScaleHD (black) and Repeat Detector (restrictive parameters - pink).*

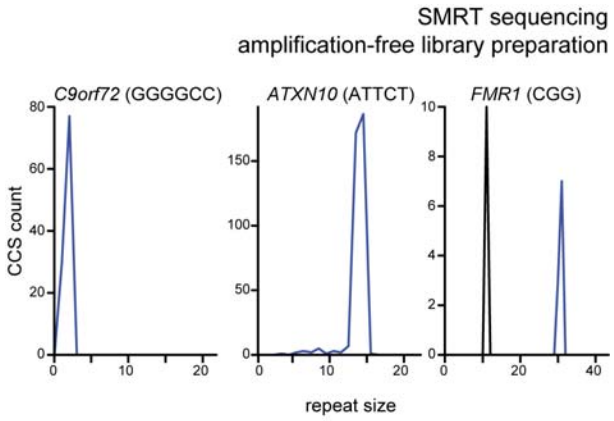


Fig. S4: *Representative RD density plots for the non-pathogenic alleles at three non-CAG loci in an HD individual. The sequencing data is from* [35].

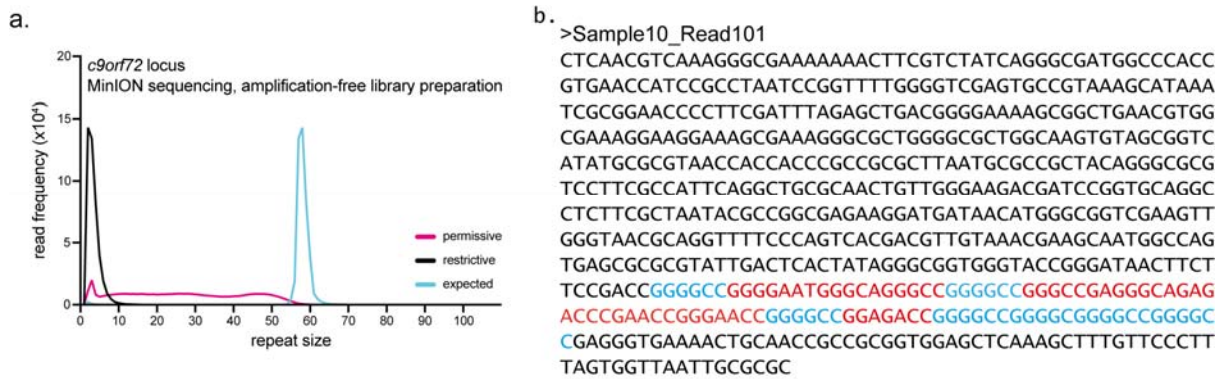


Fig. S5: *RD returns short repeat sizes on MinION sequencing data due to the large number of errors found in the repeat tract. A) Repeat size distribution analysed using RD's restrictive and permissive parameters for the* C9ORF72 *locus for Sample 10 from* [51]. *B) A representative read from Sample 10 showing that the G_4C_2 repeat (in blue) is peppered with sequencing mistakes that disrupt the sequence motif (red), which RD depends on for determining repeat size.*

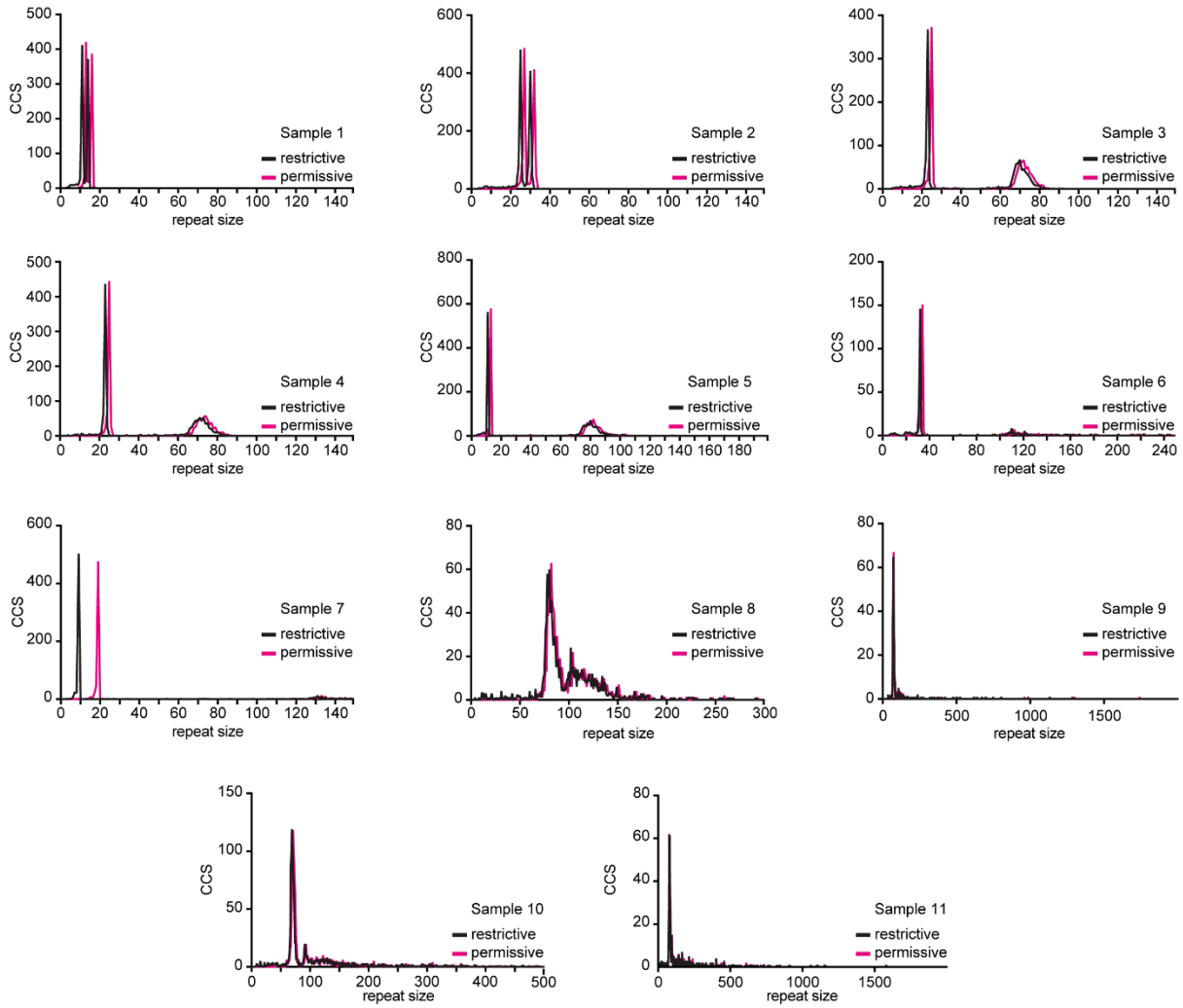


Fig. S6: *RD density plots of the FECD SMRT dataset. Sample numbering is the same as in Hafford-Tear et al* [34].

Table S1: HD SMRT sequencing samples, barcodes and primers used.

Flowcell	Sample*	Barcode (on forward primer)	Forward primer	Reverse primer	Forward Primer name	Reverse primer name
1	NA20245	TCAGACGATGCGTCAT	GCCTCCCTTACCATGCAGT	GCCCTCTTCCCTCTCAGACT	oVIN2492	oVIN1566
1	NA20206	CTATACATGACTCTGC			oVIN2493	
1	NA20207	TACTAGAGTAGCACTC			oVIN2494	
1	NA20246	TGTGTATCAGTACATG			oVIN2495	
1	NA20247	ACACGCATGACACACT			oVIN2496	
1	NA20248	GATCTCTACTATATGC			oVIN2497	
1	NA20249	ACAGTCTATACTGCTG			oVIN2498	
1	NA20250	ATGATGTGCTACATCT			oVIN2499	
1	NA20208	CTGCGTGCTCTACGAC			oVIN2500	
1	NA20209	GCGCGATACGATGACT			oVIN2501	
1	NA20251	CGCGCTCAGCTGATCG			oVIN2502	
1	NA20252	GCGCACGCACTACAGA			oVIN2503	
1	NA20210	ACACTGACGTCGCGAC			oVIN2504	
1	NA20253	CGTCTATATACGTATA			oVIN2505	
1	GM02168	ATAGAGACTCAGAGCT			oVIN2506	
1	GM03620	TAGATGCGAGAGTAGA			oVIN2507	
1	GM13504	CATAGCGACTATCGTG			oVIN2508	
1	GM13505	CATCACTACGCTAGAT			oVIN2509	
1	GM13506	CGCATCTGTGCATGCA			oVIN2510	
1	GM13507	TATGTGATCGTCTCTC			oVIN2511	
1	GM13508	GTACACGCTGTGACTA			oVIN2512	
1	GM14044	CGTGTCGCGCATATCT			oVIN2513	
2	GFP(CAG)$_{15}$	CACACGCGCGTGCTCG	AAGAGCTTCCCTTTACACAACG	TACTTGTACAGCTCGTCCATGC	oVIN0881	oVIN2514
2	GFP(CAG)$_{51}$	TCTCTCACAGTCGAGC			oVIN0887	
2	GFP(CAG)$_{91}$	CGCGCGTGTGTGCGTG			oVIN0879	
2	GFP(CAG)$_{308}$	GTCATCACACATCTCT			oVIN0889	
3	GFP(CAG)$_{15}$	CACACGCGCGTGCTCG			oVIN0881	
4	GFP(CAG)$_{51}$	TCTCTCACAGTCGAGC			oVIN0887	
5	GFP(CAG)$_{91}$	CGCGCGTGTGTGCGTG			oVIN0879	
6	GFP(CAG)$_{308}$	GTCATCACACATCTCT			oVIN0889	

*: Samples starting with NA were DNA samples obtained directly from Coriell. GM samples were LBCs obtained from Coriell and grown in our laboratory prior to DNA isolation.

Table S2: Selected samples for sequencing (total N=652).

DNA type	Number sequenced*	Number successfully sequenced
LBCs (Early samples)	250	249
LBCs (Late samples)	250	248
Time course LBCs	47	47
Blood DNA	98 (49)	98
Positive controls	7 (1)	7

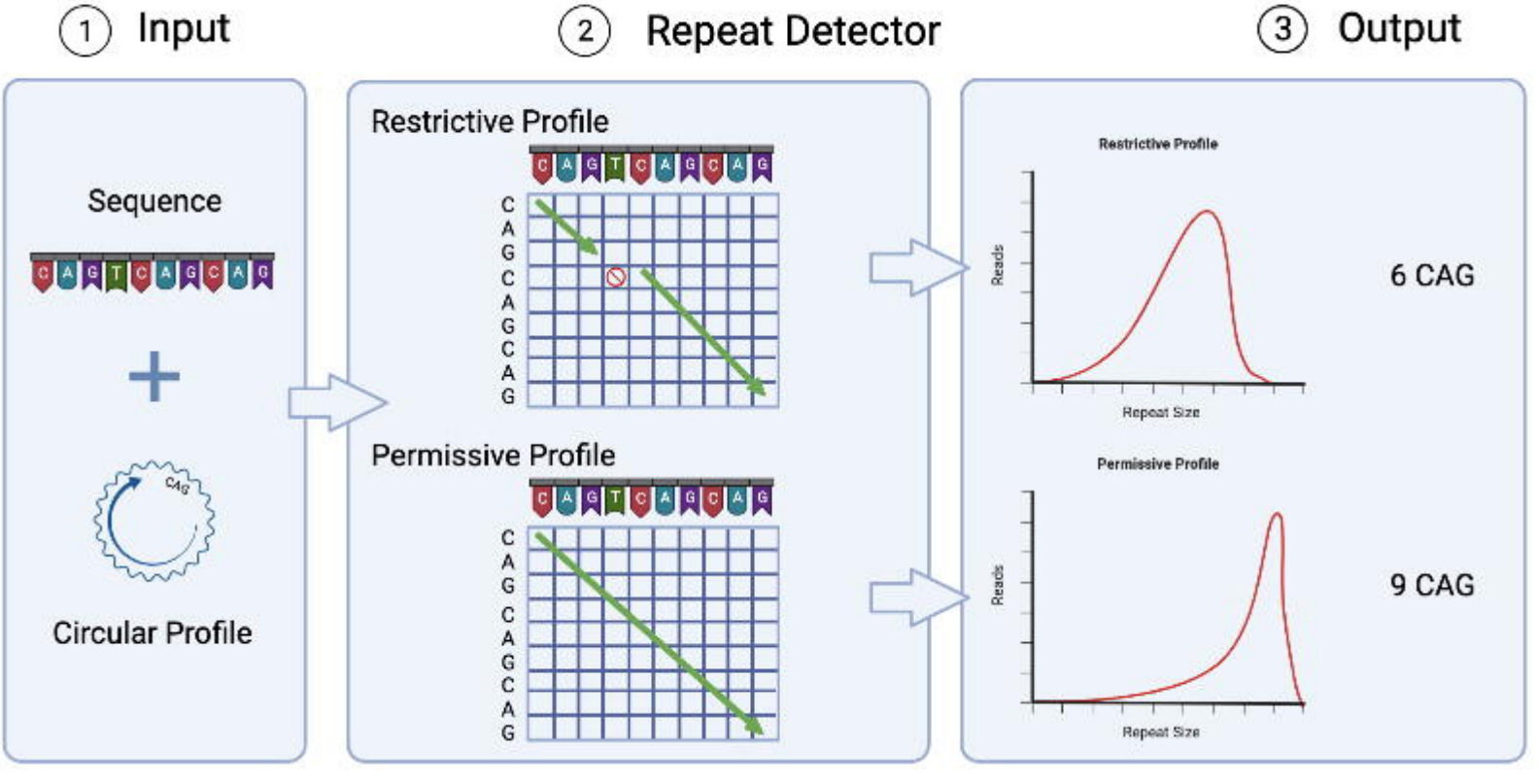
*Numbers in brackets indicate the actual number of individuals for each group; blood DNA were all sequenced twice, and the single positive control was sequenced on each plate for a total of 7 times. Blood DNA consisted of the 7 expected onset individuals as well as 26 early and 16 late.

Table S3: Sequencing metrics for all the datasets analysed

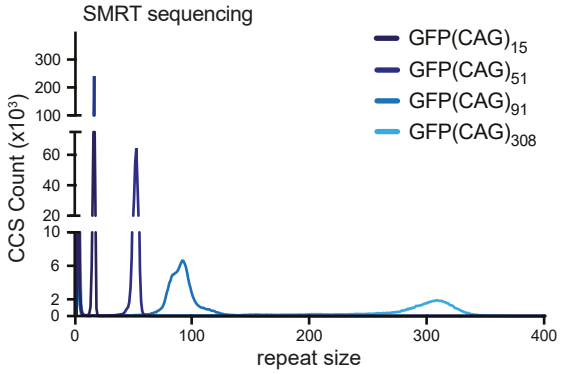
See TableS3_SequencingMetrics.xlsx attached.

Table S4: Primers used to sequence the repeat tract in GFP$(CAG)_x$ cells and to verify the presence of the duplication in GFP$(CAG)_{308}$.

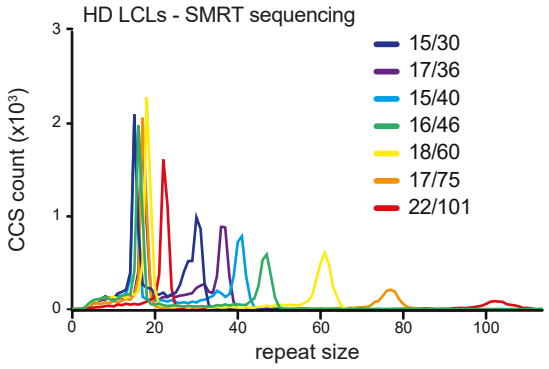
Primer	Sequence	Reference
oVIN-459	AAGAGCTTCCCTTTACACAACG	[48]
oVIN-460	TCTGCAAATTCAGTGATGC	[48]


1 Input
2 Repeat Detector
3 Output
Sequence
C A G T C A G C A G
CAG
Circular Profile
Restrictive Profile
Permissive Profile
Reads
Repeat Size
6 CAG
9 CAG

a.

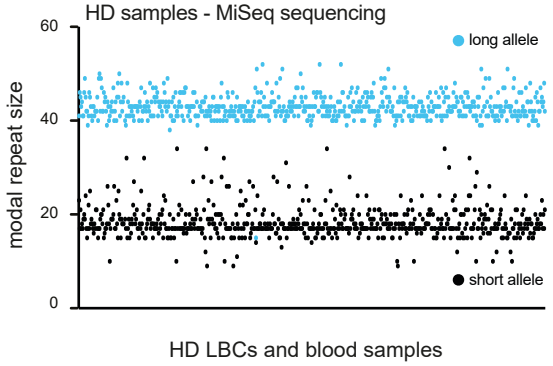


b.

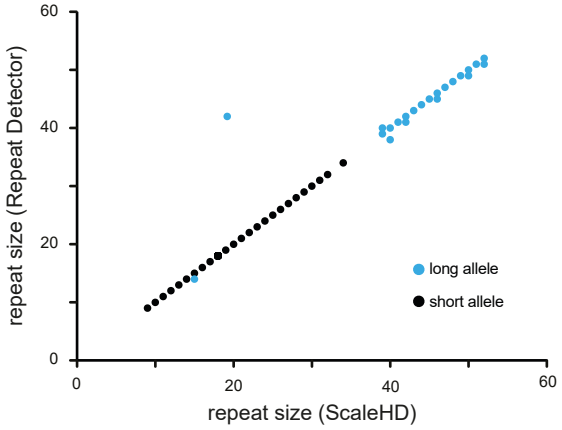


Taylor et al Fig. 2

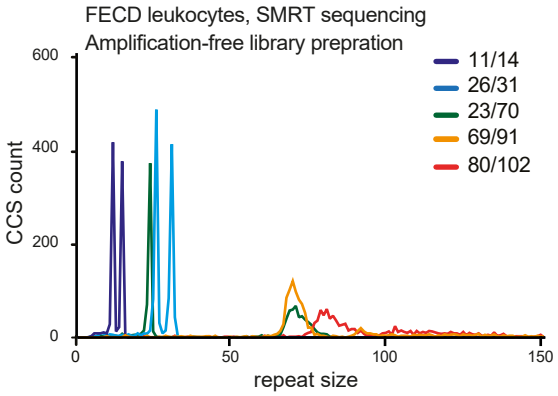
A.



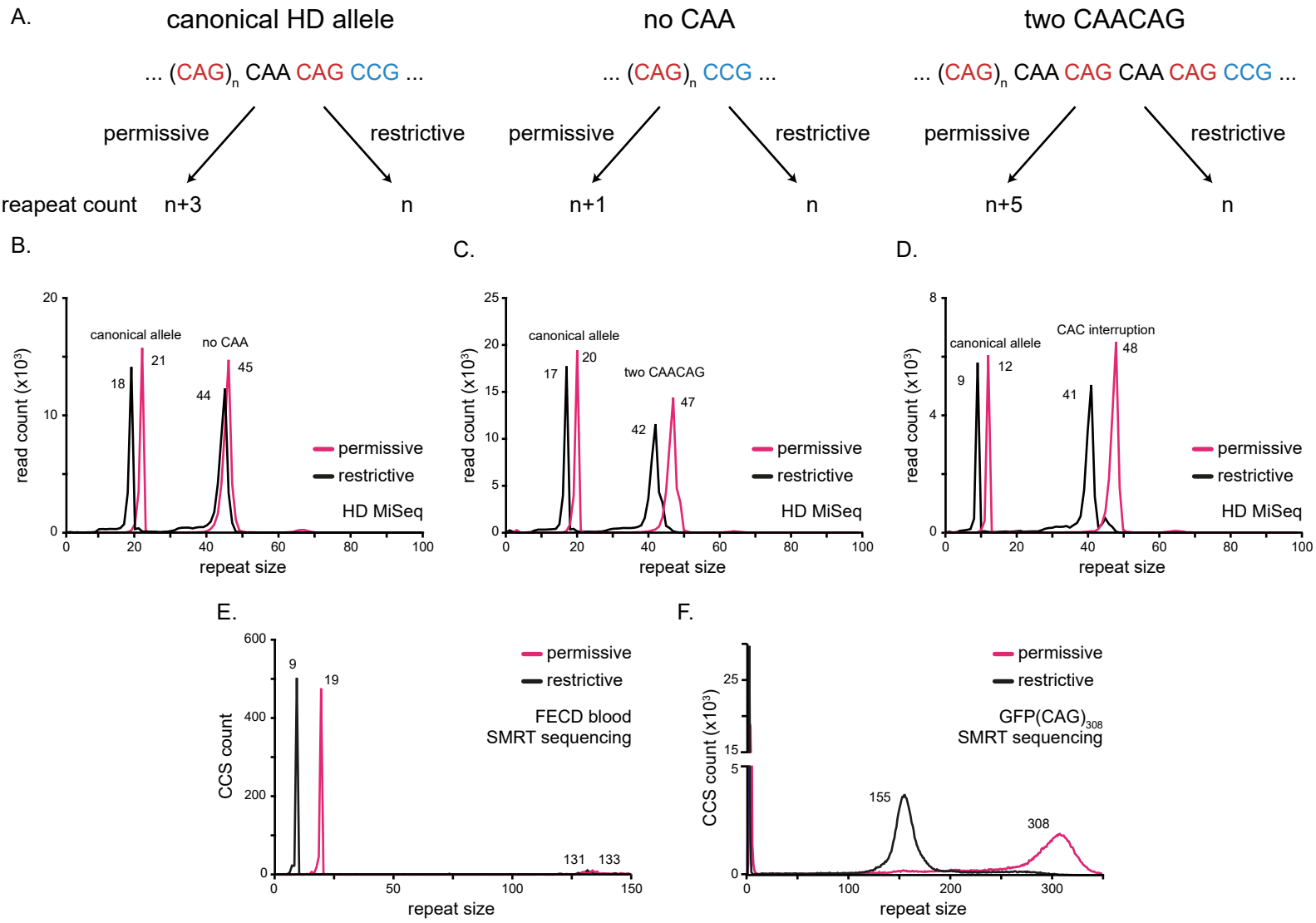
B.



Taylor et al Fig. 3



Taylor et al. Figure 4



Taylor et al Fig. 5