

GENETICS

Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing

Igor Stevanovski^{1†}, Sanjog R. Chintalaphani^{1,2,3†}, Hasindu Gamaarachchi^{1,4}, James M. Ferguson¹, Sandy S. Pineda^{5,6}, Carolin K. Scriba^{7,8}, Michel Tchan⁹, Victor Fung⁹, Karl Ng¹⁰, Andrea Cortese^{11,12}, Henry Houlden^{11,12}, Carol Dobson-Stone⁶, Lauren Fitzpatrick⁶, Glenda Halliday⁶, Gianina Ravenscroft⁷, Mark R. Davis⁷, Nigel G. Laing^{7,8}, Avi Fellner^{1,13,14}, Marina Kennerson^{15,16,17}, Kishore R. Kumar^{1,17,18}, Ira W. Deveson^{1,3*}

More than 50 neurological and neuromuscular diseases are caused by short tandem repeat (STR) expansions, with 37 different genes implicated to date. We describe the use of programmable targeted long-read sequencing with Oxford Nanopore's ReadUntil function for parallel genotyping of all known neuropathogenic STRs in a single assay. Our approach enables accurate, haplotype-resolved assembly and DNA methylation profiling of STR sites, from a list of predetermined candidates. This correctly diagnoses all individuals in a small cohort ($n = 37$) including patients with various neurogenetic diseases ($n = 25$). Targeted long-read sequencing solves large and complex STR expansions that confound established molecular tests and short-read sequencing and identifies noncanonical STR motif conformations and internal sequence interruptions. We observe a diversity of STR alleles of known and unknown pathogenicity, suggesting that long-read sequencing will redefine the genetic landscape of repeat disorders. Last, we show how the inclusion of pharmacogenomic genes as secondary ReadUntil targets can further inform patient care.

INTRODUCTION

A short tandem repeat (STR) is a short DNA sequence motif, typically 2 to 6 base pairs (bp), repeated consecutively at a given position in the genome. STRs make up ~7% of the human genome sequence and are highly polymorphic, commonly varying in length between unrelated individuals (1, 2).

Unusually long or "expanded" STR alleles are an important class of pathogenic variants in human populations. To date, STR expansions in more than 40 genes have been shown to cause heritable disorders, with the majority of these exhibiting primary neurological or neuromuscular presentations (3, 4). These include Huntington's disease (HD; *HTT*), fragile X syndrome (FXS; *FMR1*), the hereditary cerebellar

ataxias (*RFC1*, *FXN*, and others), the myotonic dystrophies (*DMPK* and *CNBP*), the myoclonic epilepsies (*CSTB*, *SAMD12*, *STARD7*, and others), and *C9orf72*-related frontotemporal dementia and amyotrophic lateral sclerosis (ALS) (3, 4). With each of the >40 STR-associated neurogenetic diseases estimated to affect ~1 to 10 individuals per 100,000, their collective prevalence is high (5–8). Moreover, the list of disorders in which STR expansions are implicated continues to grow, and many pathogenic STR genes have been described recently (5–10).

Given (i) the wide variety and collectively high prevalence of STR expansion disorders, (ii) the large number of genes involved, (iii) the frequent identification of new genes, (iv) the diversity in size and sequence conformation of pathogenic STR expansion alleles, and (v) the many gaps in our understanding of their basic biology, there is a growing need for improved methods for the molecular characterization of STRs. Established molecular techniques [e.g., Southern blot and repeat-primed polymerase chain reaction (RP-PCR)] are relatively slow, labor-intensive, and imprecise and require a separate assay with specific primers/probes for every different STR (11). This is problematic when multiple different STR expansions can manifest in a similar phenotype (locus heterogeneity) (3, 4) and is a major barrier to the implementation of tests for newly identified STR genes. Next-generation sequencing (NGS) has some utility for the analysis of STR expansions (12, 13). However, the large size, low sequence complexity, and high GC content of many pathogenic STR expansions make them refractory to analysis by short-read NGS platforms (e.g., Illumina) (11).

Long-read sequencing technologies from Oxford Nanopore Technologies (ONT) and Pacific Biosciences can be used to genotype large and complex STR expansions, even within challenging contexts such as mobile elements (5–8, 14–17). Simultaneous profiling of DNA methylation at repeat sites is another advantage of these technologies (15, 17). However, whole-genome analysis

¹Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia. ²School of Medicine, University of New South Wales, Sydney, NSW, Australia. ³St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia. ⁴School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia. ⁵Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia. ⁶The University of Sydney, Brain and Mind Centre and School of Medical Sciences, Faculty of Medicine and Health, Camperdown, NSW, Australia. ⁷Harry Perkins Institute of Medical Research, University of Western Australia, Nedlands, WA, Australia. ⁸Diagnostic Genomics, PathWest Laboratory Medicine WA, Nedlands, WA, Australia. ⁹Westmead Hospital, Westmead, NSW, Australia and Sydney Medical School, The University of Sydney, NSW, Australia. ¹⁰Department of Neurology, Royal North Shore Hospital and The University of Sydney, Sydney, NSW, Australia. ¹¹Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK. ¹²The National Hospital for Neurology and Neurosurgery, London, UK. ¹³Raphael Recanati Genetics Institute, Rabin Medical Center, Beilinson Hospital, Petah Tikva, Israel. ¹⁴The Neurology Department, Rabin Medical Center, Beilinson Hospital, Petah Tikva, Israel. ¹⁵Northcott Neuroscience Laboratory, ANZAC Research Institute, Sydney, NSW, Australia. ¹⁶Faculty of Health and Medicine, University of Sydney, Camperdown, NSW, Australia. ¹⁷Molecular Medicine Laboratory, Concord Hospital, Concord, NSW, Australia. ¹⁸Neurology Department, Central Clinical School, Concord Repatriation General Hospital, University of Sydney, Concord, NSW, Australia.

*Corresponding author. Email: i.deveson@garvan.org.au

†These authors contributed equally to this work.

remains prohibitively expensive on either platform. A cas9-based approach for targeted enrichment of STR loci and long-read sequencing was recently developed (15, 18). However, this suffers from the same limitation as existing molecular techniques, in that a unique set of cas9 guide RNAs is needed for every different STR, requiring careful design and experimental optimization.

An alternative approach to targeted long-read sequencing is ONT's "ReadUntil" functionality, whereby an ONT sequencing device can be programmed to recognize and accept/reject specific DNA sequence fragments during a sequencing experiment (19, 20). Target selection is fully flexible and requires no additional laboratory processes beyond standard library preparation. Here, we demonstrate that ONT ReadUntil can be used to achieve accurate molecular characterization of all known neuropathogenic STRs in a single assay. Using a custom panel comprising 37 STR loci associated with neurological and neuromuscular disease, we performed targeted ONT sequencing on 37 patient-derived DNA samples to identify and fully characterize a diverse range of STR expansions. Our study

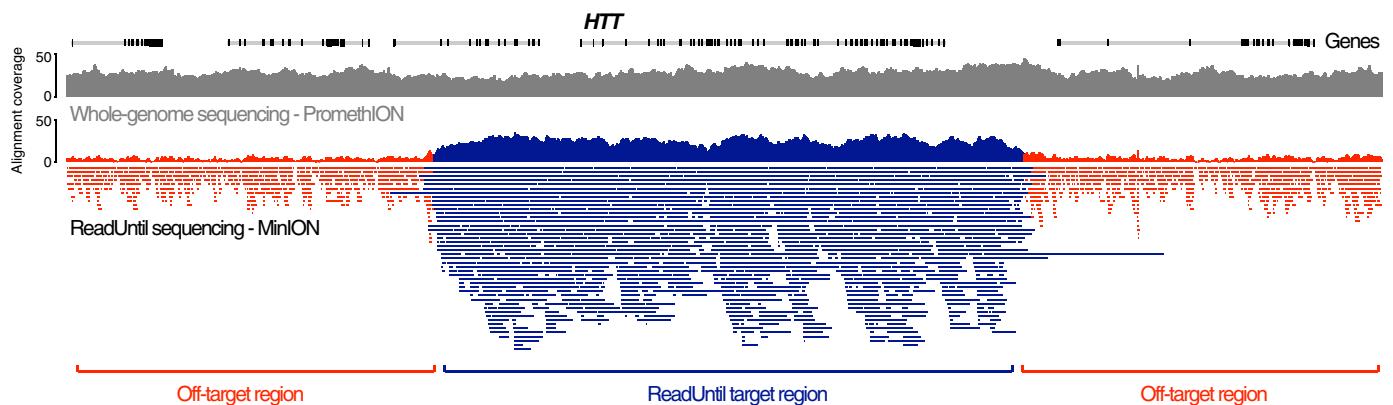
establishes the analytical validity of programmable ONT sequencing for the genetic diagnosis of STR expansion disorders and showcases the numerous advantages of this approach.

RESULTS

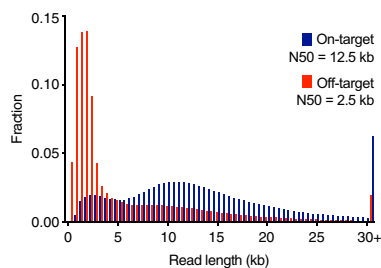
Programmable targeted nanopore sequencing of pathogenic STRs

ONT's ReadUntil function has the potential to enable simple, cost-effective sequencing of all known pathogenic STR loci but is a largely untested technology. To evaluate the use of ReadUntil for STR profiling, we designed a custom panel encompassing all genes known to harbor pathogenic STR expansions implicated in primary neurological and neuromuscular diseases ($n = 37$; table S1). For each gene, the entire locus was targeted, including 50 kb of flanking sequence in either direction (Fig. 1A). The panel included a range of additional clinically informative loci and covered ~50.5 Mb in total, equating to ~1.6% of the human reference genome (hg38; see table S2).

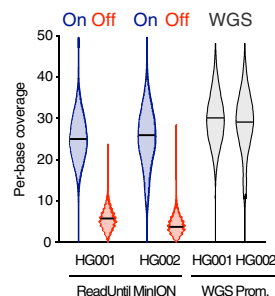
A Targeted nanopore sequencing of STR genes with ReadUntil



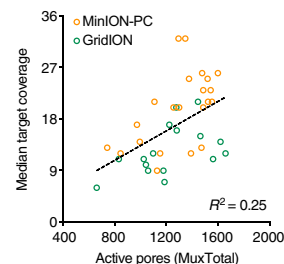
B Read length



C Target enrichment



D Flow cell quality



E Performance benchmarking

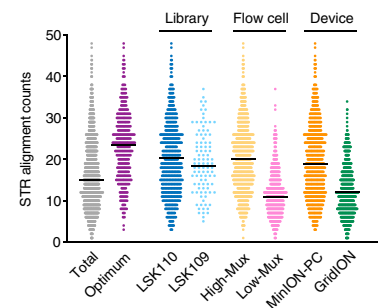


Fig. 1. Targeted sequencing of pathogenic STR sites with ONT ReadUntil. (A) Genome browser view shows sequencing alignments to the *HTT* locus and surrounding regions for a typical ONT ReadUntil experiment (lower track). Location of ReadUntil target region for *HTT* is marked below, and on-target (navy) versus off-target alignments (red) are distinguished by color. For comparison, a coverage track is also shown for a typical whole-genome ONT sequencing experiment (gray). (B) Histograms compare read length distribution for on-target (navy; N50 = 12.5 kb) versus off-target (red; N50 = 2.5 kb) alignments. Data are averaged over all ReadUntil experiments from the study ($n = 37$). (C) Violin plots show per-base coverage distributions within on-target regions (navy) versus randomly selected off-target genes (red) during ReadUntil sequencing of HG001 and HG002 reference samples, with data from whole-genome sequencing (WGS; gray) shown for comparison. (D) Scatterplot shows median coverage across on-target regions, relative to the starting number of active pores (MuxTotal) on each ONT flow cell. Colors distinguish ReadUntil experiments run on an ONT GridION (green; NVIDIA Quadro GV100 GPU; $n = 16$) or a high-specification PC workstation (orange; NVIDIA 3090 GPU; $n = 22$); see table S5 for full specifications. (E) Dot plots show the number of alignments spanning STR sites ($n = 37$) across all ONT ReadUntil experiments ($n = 37$). Colors distinguish runs performed with LSK110 (navy) versus LSK109 (blue) library preparation kit, high quality (MuxTotal > 1200 pores; yellow) versus low quality (MuxTotal < 1200 pores; pink), and GridION (green) versus MinION-PC device (orange). Data from runs with "optimum" parameters (LSK110 + MuxTotal > 1200 pores + MinION-PC device; purple) yielded a median of 24 alignments spanning target STR sites.

We used the open-source software package Readfish (20) to execute targeted sequencing on 37 genomic DNA samples obtained from reference catalogs or collected from consenting patients (table S3). Genomic DNA was sheared to ~15- to 25-kb fragments before library preparation and sequenced on an ONT MinION flow cell (see Methods). We observed a consistent reduction in read length for off-target reads (N50 = 2.5 kb) compared to on-target reads (N50 = 12.5 kb), indicating successful rejection of fragments outside the target regions (Fig. 1B). This resulted in a median 4.6-fold enrichment in sequencing depth within target regions, yielding ~9 to 40× median target coverage across the cohort (Fig. 1C and table S4). Notably, ONT ReadUntil achieved similar coverage depth and evenness to whole-genome ONT sequencing of matched samples on a high-output PromethION flow cell, performed at more than three times the cost (Fig. 1C).

To determine optimum workflow settings, we evaluated the impact of various parameters, including the choice of ONT library preparation kit, flow cell quality, and the computer used to execute ReadUntil (see Methods). This revealed (i) superior on/off-target enrichment with LSK109 library prep chemistry but greater total output and on-target coverage with LSK110 (fig. S1, A and B), (ii) the initial number of live pores on a flow cell is an important determinant of final target coverage achieved (Fig. 1D and fig. S1, C and D), and (iii) shorter average read rejection time and improved target coverage when running Readfish on a PC workstation, compared to an ONT GridION device, thanks to its superior specifications (Fig. 1D; fig. S1, E and F; and table S5). Across all sequencing experiments with optimum workflow parameters (LSK110, MinION-PC, >1200 pores), we obtained a median of 24 sequence alignments spanning targeted STR sites (Fig. 1E), demonstrating that ONT ReadUntil can be used to achieve effective targeted STR sequencing.

Validity and utility of programmable targeted STR sequencing

To establish the validity and utility of our targeted nanopore sequencing assay, we analyzed a range of patient-derived reference DNA samples and consenting patients with neurogenetic diseases ($n = 25$), including HD ($n = 5$); FXS ($n = 2$); cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS; $n = 6$); spinal and bulbar muscular atrophy of Kennedy (SBMA; $n = 1$); myotonic dystrophy 1 (DM1; $n = 5$); neuronal intranuclear inclusion disease (NIID; $n = 1$); Friedreich ataxia (FRDA; $n = 2$); ALS ($n = 1$); spinocerebellar ataxia type 1 (SCA1; $n = 1$); and oculopharyngeal muscular dystrophy (OPMD; $n = 1$), as well as known premutation carriers ($n = 6$) and unaffected individuals ($n = 6$). Samples were also subject to independent molecular testing in accredited genetic pathology laboratories or using standard approaches (see Methods). This allowed ONT sequencing data to be evaluated against the current best practices.

Huntington's disease

HD is an autosomal dominant neurodegenerative disorder caused by a polyglutamine STR expansion of ≥ 36 "CAG" motifs within the gene *HTT*, with complete penetrance at ≥ 40 copies (21, 22). STR expansion size is correlated with disease severity, as is the absence of characteristic "CAA" interrupting motifs (also coding for glutamine) within expanded alleles (23, 24). Genetic diagnosis of HD therefore requires accurate, allele-specific STR sizing and internal sequence determination.

Targeted sequencing with ReadUntil yielded a median of 18 spanning alignments at the STR site in *HTT* exon 1 (fig. S2Ai). This was sufficient to phase and assemble both STR alleles in every patient,

identifying CAG repeats ranging from 12 to 74 copies across the cohort (Fig. 2A and fig. S2Ai; see Methods). In all patients affected by HD ($n = 5$), a single expanded STR allele was detected with length in the known pathogenic range, whereas no pathogenic expansions were detected in unaffected individuals ($n = 32$). The lengths of both expanded and nonexpanded STR alleles determined by ONT sequencing were closely concordant with clinical testing ($R^2 = 0.996$; Fig. 2B). A single CAA interruption was detected within both STR alleles of all HD-affected and nonaffected individuals, with six of the latter individuals harboring a double CAA interruption (Fig. 2A and fig. S2Ai). Our assay also resolved the boundary between the disease-associated CAG polyglutamine repeat and a "CCG" polyproline repeat located immediately downstream within *HTT* exon 1, which similarly varied in size across our cohort. While polymorphism in this adjacent repeat is not considered relevant to the HD phenotype, it is an important technical variable that can confound molecular assays for sizing the disease-associated STR (25).

HD-like syndromes may be caused by other STR genes that were also included on our targeted sequencing panel (i.e., *C9orf72*, *PRNP*, *JPH3*, *TBP*, *ATXN8*, *FXN*, and *ATN1*) (3, 4). Parallel sequencing of these genes showed that all HD-affected patients harbored STR alleles within healthy ranges (fig. S2). The capacity to rule out confounding or co-occurring STR expansions in these genes, without the need for additional molecular tests, is a clear advantage of our multigene assay.

Fragile X syndrome

FXS is the most common cause of inherited intellectual disability and single-gene cause of autism spectrum disorder in males (26). FXS is caused by large (>200) "CGG" STR expansions within the chrX-linked gene *FMR1* (27). Premutation alleles of 55 to 200 CGG repeats are also associated with late-onset fragile X-associated tremor/ataxia syndrome in males and primary ovarian insufficiency in females (5). Interrupting "AGG" motifs are reported to stabilize STR alleles to protect against full expansion (28, 29). DNA methylation (5mC) is also implicated in the pathogenic mechanism of *FMR1*-related disorders, with expanded alleles typically exhibiting promoter hypermethylation and silencing of *FMR1* (30, 31). Therefore, complete genetic diagnosis of *FMR1*-related disorders requires DNA methylation profiling, in addition to STR sizing and internal sequence determination.

At the STR site in the *FMR1* 5' untranslated region (UTR; exon 1), we obtained a median of 19 spanning alignments in females and 9 in males or ~9 alignments per allele overall (fig. S2Bi). All alleles were successfully assembled into CGG STRs ranging from 20 to 654 copies across the cohort (Fig. 2A and fig. S2Bi). The length of both expanded and nonexpanded alleles was closely concordant with clinical testing ($R^2 = 0.993$; Fig. 2B). Results from ONT sequencing correctly distinguished affected male individuals ($n = 2$) and a female carrier ($n = 1$) from unaffected individuals and distinguished premutation alleles ($n = 2$) from full pathogenic STR expansions (Fig. 2A). Within both individuals that harbored a premutation allele, we detected two protective AGG motif interruptions, and similar interruptions were common across nonexpanded STR alleles in unaffected individuals (Fig. 2A and fig. S2Bi). Parallel sequencing of *FMR2* (*AFF2*), an STR expansion gene with partial phenotypic overlap to *FMR1* (32), ruled out pathogenic expansions in all individuals (fig. S2Jii).

DNA methylation profiling revealed hypermethylation of the *FMR1* promoter region in both FXS-affected males, who harbored full STR expansions (CGG₆₅₄ and CGG₆₀₆), with >75% median methylation

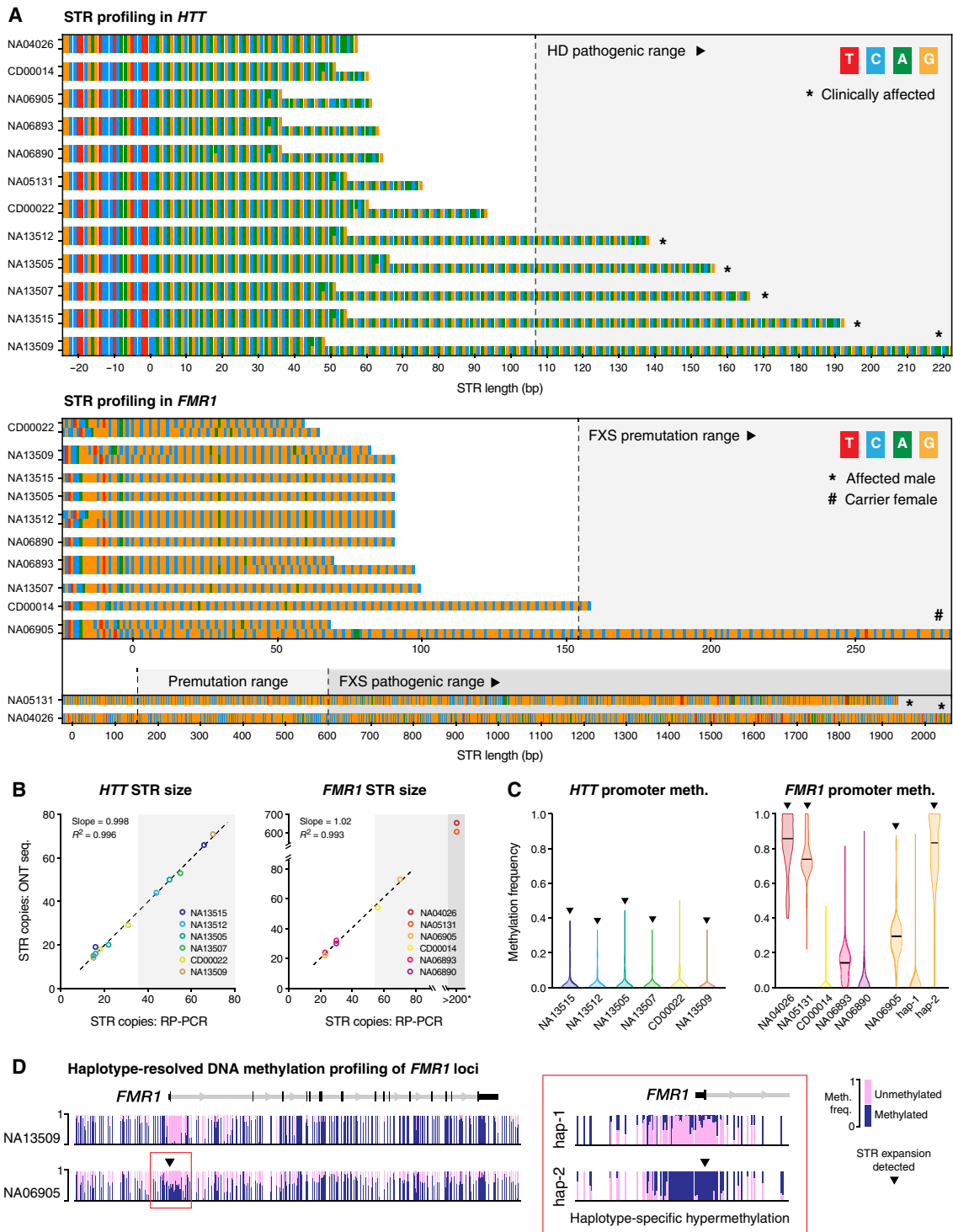


Fig. 2. Haplotype-resolved assembly and DNA methylation profiling of *HTT* and *FMR1*. (A) Sequence barcharts show *HTT* (top) and *FMR1* (bottom) STR alleles, including 25 bp of upstream flanking sequence, assembled from ONT sequencing of relevant Coriell reference DNA samples ($n = 12$; see table S3). Two alleles are shown for each individual, excepting *FMR1* for male individuals, where only one copy is present. Clinically affected and premutation carrier individuals are marked, based on clinical information from Coriell. Further details of individuals are shown in fig. S2 (Ai and Bi). (B) Scatterplots show lengths of STR alleles in *HTT* (CAG_n; left) and *FMR1* (CGG_n; right) as determined by ONT sequencing versus RP-PCR (data from Coriell). For *FMR1*, two samples exceeded the upper limit of RP-PCR genotyping (~CGG₂₀₀). (C) For the same samples, violin plots show distribution of DNA methylation frequencies recorded at CpG sites within the promoter regions of *HTT* (left) and *FMR1* (right). Triangles indicate which samples contained pathogenic STR expansions. For sample NA06905, differential methylation was observed between the two *FMR1* haplotypes; these are shown separately. (D) Genome browser view shows examples of DNA methylation profiles across the complete *FMR1* locus for two samples: NA13509 (female with no STR expansion in *FMR1*) and NA06905 (female carrier of *FMR1* premutation). Inset shows haplotype-specific promoter methylation in NA06905.

frequencies among local CpG sites (Fig. 2C; see Methods). By contrast, promoter CpG methylation frequencies were low among males with normal and premutation STR alleles (<25% median frequency; Fig. 2C). Females also showed low methylation frequencies, with the exception of the single premutation carrier (NA0695). In this individual, we observed differential methylation between the two *FMRI* haplotypes, with the premutation haplotype being predominantly methylated (CGG₇₃; 83% median frequency) but the normal haplotype unmethylated (GCC₂₃; 0% median frequency; Fig. 2, C and D). DNA hypermethylation in *FMRI* premutation alleles has been reported previously, is correlated to STR repeat size, and may account for variability in phenotypic expression in premutation carriers (33, 34). In contrast to *FMRI*, and in line with expectations, we did not observe DNA hypermethylation in the promoter of *HTT* on either expanded or nonexpanded STR alleles for any individuals, further confirming the reliability of the analysis (Fig. 2C). The capacity to obtain haplotype-resolved DNA methylation profiles, in addition to STR size and interruption status, all in a single, simple assay is a clear advantage of our approach.

Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome

CANVAS is a neurodegenerative movement disorder shown recently to be caused, in most cases, by large biallelic “AAGGG” STR expansions in the gene *RFC1* (9, 35). Expanded STR alleles in *RFC1* are relatively common and exist in several different motif conformations. The motifs “AAAAG_{exp}” and “AAAGG_{exp}” are considered nonpathogenic, regardless of size. In addition to the canonical pathogenic motif “AAGGG_{exp}”, a rare “ACAGG_{exp}” motif and mixed “AAAGG₁₀₋₂₅AAGGG_{exp}” conformation are both considered pathogenic, while the pathogenicity of various other observed conformations is currently unknown (9, 36, 37). The requirement to resolve very large STR expansions with diverse motifs, in an allele-specific fashion, makes genetic diagnosis of patients with CANVAS challenging, often yielding inconclusive results.

Targeted sequencing with ReadUntil obtained a median of 15 spanning alignments at the STR site within the second intron of *RFC1*. Haplotype-resolved STR assembly revealed a variety of different pentanucleotide repeats ranging in size from 8 to 1070 copies across the cohort, with strong concordance to molecular testing by Southern blot and/or RP-PCR ($R^2 = 0.946$; Figs. 3, A and B, and 4A). Large biallelic STR expansions of the pathogenic motif (AAGGG₄₁₀₋₁₀₇₀) were detected in five of six patients with CANVAS but not in unaffected individuals ($n = 21$; Figs. 3A and 4A).

In the single remaining patient with CANVAS [R210005, previously reported as R19955 (38)], we detected a large (~5-kb) biallelic expansion at the *RFC1* STR site. However, the pathogenic AAGGG motif was found on only one allele, with the other harboring the “AAAGG” motif that is currently considered nonpathogenic (Fig. 3A) (9). Because the two alleles are equivalent in size and the pathogenic motif is present, standard molecular testing suggested that this individual had a biallelic pathogenic expansion (Fig. 3A). Because both STR alleles are much greater than the read length of short-read NGS platforms, analysis by clinical whole-genome sequencing was also inconclusive (Fig. 3C). In contrast, our assay readily distinguished the two STR alleles, identifying distinct “AAGGG₁₀₁₀” and “AAAGG₉₆₀” conformations (Fig. 3, A and C). This highlights the utility of long-read sequencing for profiling large, complex STRs. In addition, this finding suggests potential pathogenicity of the *RFC1* AAAGG expansion at large sizes, as opposed to the current assumption that this motif is nonpathogenic, regardless of size (9).

SBMA, DM1, NIID, FRDA, ALS, SCA1, and OPMD

To further demonstrate the broad utility of our assay, we analyzed patients affected with SBMA ($n = 2$), DM1 ($n = 5$), NIID ($n = 1$), FRDA ($n = 2$), ALS ($n = 1$), SCA1 ($n = 1$), and OPMD ($n = 1$). In all cases, we were able to correctly genotype the relevant STR (fig. S2, Ci to Ii). Detailed description of the results pertaining to each disorder is provided in note S1.

In summary, the results described above demonstrate accurate, haplotype-resolved sizing, sequence determination, and DNA methylation profiling of neuropathogenic STR loci using targeted ONT sequencing. This establishes analytical validity for the genetic diagnosis of STR expansion disorders and highlights numerous advantages to this approach.

Resolving STR diversity

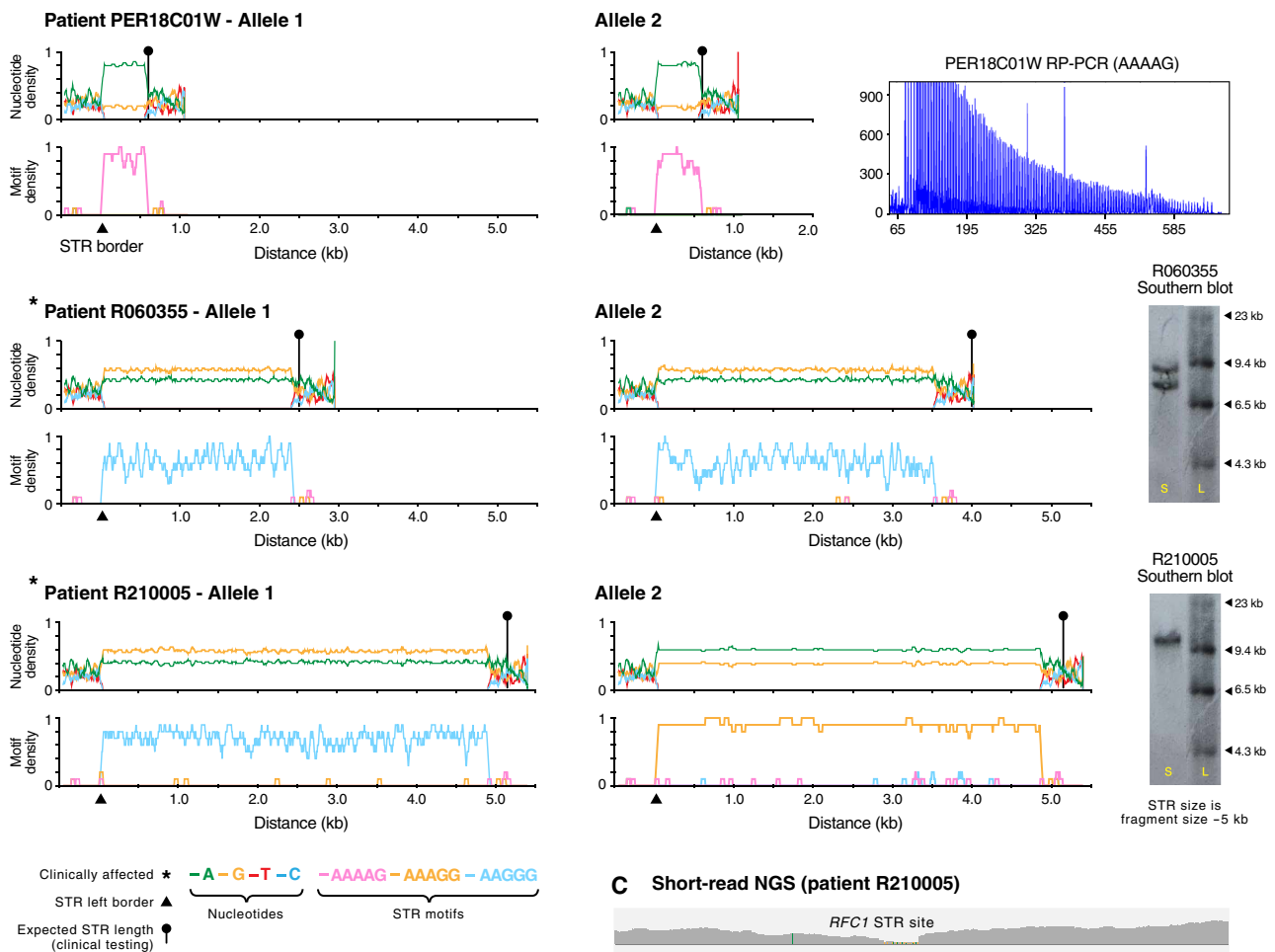
STR sequences are highly polymorphic (1, 2), yet their true diversity is likely underappreciated because of limitations in current genotyping methods (3, 4). Clinical interpretation relies on our ability to distinguish pathogenic alleles from the diversity of STRs encountered in healthy individuals. By determining the size and sequence of every allele of every disease-associated STR site in every individual tested, our targeted sequencing assay provides valuable data to help define the genetic landscape of STRs in human populations.

We observed a diverse array of STR alleles across our cohort ($n = 37$), which are visualized in full for each gene in fig. S2 (Ai to Jii) and summarized in Fig. 4A. The diversity of STR sizes among clinically nonaffected individuals was most evident in the pentanucleotide-repeat genes *RFC1* (8 to 324 copies), *DAB1* (8 to 541 copies), *BEAN1* (10 to 119 copies), *SAMD12* (13 to 113 copies), and *STARD7* (10 to 102 copies; Fig. 4, A and B). These genes also harbored multiple unique STR motif conformations. For example, among 31 individuals not affected by CANVAS, we identified three carriers of pathogenic AAGGG STR alleles in *RFC1* (<400 copies) and five carriers of noncanonical motifs “ACGGG,” “AAGAG,” and “AAAGGG” that are of unknown pathogenicity. The remaining individuals had nonpathogenic AAAGG and “AAAAG” alleles (Fig. 4B).

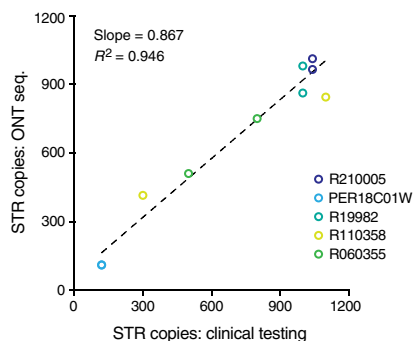
Other recently found pentanucleotide-repeat genes showed similar polymorphism. An “ATTTT” STR site within *STARD7* was recently linked to familial adult myoclonus epilepsy 2 (FAME2), with affected individuals harboring inserted “ATTTC” or “AAATG” motifs (39). While we did not genotype any patients with FAME2, we observed novel “AAACT” ($n = 4$), “AACAT” ($n = 4$), and “AAAAC” ($n = 1$) *STARD7* alleles in our cohort (fig. S2Ji). Spinocerebellar ataxia 31 (SCA31) is caused by insertion of “TGGAA” motifs immediately upstream from a “TAAAA” STR in *BEAN1* (40). It has been reported that, in a Japanese population, >99% of healthy individuals have a “TAAAA₈₋₂₀” allele at this site (41), although other motifs have been observed. Unexpectedly, of 37 individuals in our cohort, only 26 harbored two “normal” (TAAAA₈₋₂₀) alleles. Of the remaining individuals, seven harbored expanded TAAAA alleles (>20 copies), and two had expanded alleles of predominantly AACT and AACAT motifs (fig. S2Ki). The allelic diversity of these genes was in contrast to other pentanucleotide-repeat genes *ATXN10*, *MARCHF6*, *RAPGEF2*, and *TNRC6A*, where STR alleles were largely uniform in size and sequence across the cohort (Fig. 4A and fig. S2).

Our analysis also revealed consistent detection of internal STR motif interruptions (fig. S2 and table S6). In addition to “AAG” interruptions in *FMRI* and CAA interruptions in *HTT* (discussed above), we detected known “CAT” interruptions in *ATXN1*, CAA

A STR profiling in *RFC1*



B *RFC1* STR size



C Short-read NGS (patient R210005)

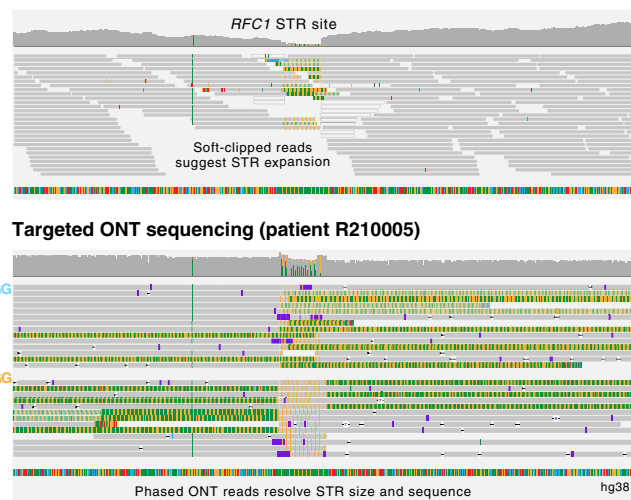


Fig. 3. Haplotype-resolved assembly of pathogenic STR site *RFC1*. (A) Line plots show nucleotide content (top) and density of pentanucleotide STR motifs (bottom) enumerated in a 50-bp sliding window across assembled STR alleles (including 1-kb up/downstream flanking sequences). Data are shown for three consenting individuals that were subjected to clinical testing for STR expansions in *RFC1*. Relevant molecular testing data (RP-PCR or Southern blot) are shown for each individual (see table S3). Asterisks indicate CANVAS-affected patients, triangles show the position of the left border of assembled STRs, and circular markers show the expected length of STR alleles, as determined by clinical testing. (B) Scatterplot shows lengths of pentanucleotide STR alleles in *RFC1* in consenting individuals ($n = 5$), as determined by ONT sequencing versus molecular testing. (C) For patient R210005, genome browser views show short-read NGS alignments (top) at the pathogenic STR site in *RFC1*. The presence of soft-clipped bases suggests that an STR expansion is present, but the size, sequence, and allelicity cannot be directly determined. Bottom panel shows phased ONT alignments from the same sample. Long reads directly measure the STR expansion size and reveal distinct motif conformations on the two *RFC1* alleles.



Fig. 4. Diversity of STR alleles across the study. (A) Dot plot shows observed sizes of STR alleles for each gene ($n = 37$) in all individuals assessed during our study ($n = 37$). Gray boxes mark expected size ranges for normal, premutation, and pathogenic STR alleles for each gene, where known. Filled circles indicate pathogenic alleles confirmed by clinical molecular testing, and empty circles were confirmed as nonpathogenic or were not tested. Full results for each individual gene are provided in fig. S2. (B) Motif barcharts show observed sizes and motif conformations of STR alleles assembled for the *RFC1* gene in each individual ($n = 37$). Red frame identifies CANVAS-affected patients, where large STR expansions in *RFC1* were detected by clinical testing and ONT sequencing.

interruptions in *ATXN2*, and CAA interruptions in *ATXN3* (fig. S2, Hi, Cii, and Dii). We also detected motif interruptions in several other genes, including *ATXN10* (“ATTGT” interruptions; $n = 8$), *TBP* (CAA interruptions; $n = 37$), *CNBP* (“GCTG/TCTG/GGCT” interruptions;

$n = 37$), *NOTCH2NLC* (“GGA” interruptions; $n = 37$), and *AR* (CAA; $n = 1$; fig. S2).

While our study does not encompass a sufficiently large and unbiased cohort to draw general conclusions, it is clear that targeted

long-read sequencing will help to describe a currently underappreciated diversity of STR alleles and likely redefine the features of nonpathogenic alleles for several genes. Moreover, it will facilitate more detailed investigation of genotype-phenotype correlations and potential disease modifiers of pathogenic repeat expansion alleles, such as interruptions.

Informative secondary targets: Pharmacogenomic genes

In addition to disease-associated STR genes, our ReadUntil sequencing panel included a range of other targets that may provide further clinical insights (see table S2). Because target selection is programmable, such secondary targets come at no additional cost and can be flexibly included/excluded on an individual basis. As an example, we analyzed 28 pharmacogenomic (PGx) genes included in the panel (table S2). PGx genotyping can anticipate an individual's capacity to metabolize specific drugs and prevent adverse drug reactions (42).

Many PGx genes, including the prototypical example *CYP2D6* (Fig. 5A), are highly polymorphic and frequently harbor structural variation and have close homologs, making them difficult to genotype using standard molecular techniques or short-read NGS (42, 43). Despite their complex architectures, ReadUntil targeted enrichment was effective within PGx genes, achieving a median 25× coverage depth (Fig. 5B). Notably, the breadth and evenness of coverage by unique ONT sequencing alignments in PGx genes were superior to those of whole-genome short-read NGS on matched samples (Fig. 5, B and C). This is best illustrated at the *CYP2D6* locus, where ONT ReadUntil achieved complete coverage and phasing of *CYP2D6* and its neighboring pseudogene *CYP2D7*, unlike short-read NGS (Fig. 5A). This emphasizes the advantages of long-read sequencing at complex genomic loci.

We analyzed the reference human DNA samples HG001 and HG002 (NA12878 and NA24385) and compared the results to high-confidence annotations from the Genome in a Bottle (GIAB) project to evaluate variant detection within PGx loci (see Methods). After parameter optimization with the software Nanopolish (44), we observed accurate detection for SNVs (Single Nucleotide Variants) within PGx regions on both samples (sensitivity, 95 and 97%; precision, 96 and 98%; Fig. 5, D and E). However, the high frequency of insertion-deletion (indel) errors in ONT reads meant that indel variants could not be reliably detected (sensitivity, 36 and 56%; precision, 56 and 58%; Fig. 5, D and E). Similar performance was also observed within STR genes and other inherited disease genes on our ReadUntil panel (table S7).

In HG001 and HG002, respectively, we detected 40 and 43 annotated PharmVar SNVs in key PGx genes *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, and *DPYD*, several of which show evidence of PGx phenotypes (45). In HG001, for example, we identified a single SNV (c.681G>A) in *CYP2C19* that is known to introduce an aberrant splice-acceptor site and truncation of the canonical open reading frame (Fig. 5F) (46). This allele (*CYP2C19*^{*2}) affects the metabolism of multiple medications and is associated with toxicity for the common antidepressants citalopram and escitalopram (level 1A evidence) (47). Given that prescription of antidepressants to patients with repeat expansion disorders such as HD is relatively common (48), this example demonstrates the potential utility of parallel PGx genotyping in patients with STR. While this is not the primary purpose of our targeted sequencing assay, secondary findings of this nature may better inform patient care at no extra cost, underscoring the appealing flexibility of programmable sequencing with ONT ReadUntil.

DISCUSSION

This study demonstrates the validity and utility of programmable targeted nanopore sequencing for the genetic diagnosis of STR expansion disorders. Unlike existing single-gene molecular techniques, our approach enabled unbiased sizing and sequence determination of all known neuropathogenic STR sites in a single targeted assay. Haplotype-resolved assembly of ONT reads was used to solve large and complex STR expansions, such as those described in *RFC1*, that eluded characterization by standard molecular testing and short-read NGS. Moreover, we identified motif interruptions within STR sequences and local DNA methylation profiles that further inform pathogenicity. ReadUntil achieves similar on-target coverage on a handheld MinION device to whole-genome nanopore sequencing on a benchtop PromethION, delivering a >3-fold price reduction per sample in addition to reduced capital costs, data storage, and computational requirements. Given these capabilities, we propose that targeted sequencing with ONT ReadUntil (19, 20) can address the pressing need for improved methods for molecular characterization of STR expansions.

Our multigene assay has the additional benefit of informing consenting patients of carrier status for pathogenic expansions not associated with their primary diagnosis (e.g., *RFC1* AAGGG pathogenic expansion in the DM1 patient NA23265). The use of long-read sequencing also enables the detection and phasing of heterozygous SNVs nearby pathogenic STR sites. Phased SNVs are useful genetic markers that distinguish pathogenic and nonpathogenic haplotypes during family genetic studies or preimplantation genetic diagnosis (49), further supporting the utility of our assay.

ONT ReadUntil permits the flexible inclusion of virtually any additional secondary targets for genetic analysis. In this study, targeted sequencing of 28 PGx genes identified clinically actionable PGx alleles that may be used to guide personalized selection of medications. Unlike a traditional targeted sequencing panel, these targets can be included/excluded from a given assay at the clinician's discretion or based on individual patient consent. The value of such secondary targets, which come at no extra cost, will continue to increase with ongoing improvement in nanopore sequencing accuracy (50), particularly if this enables reliable detection of indel variants that was beyond current capabilities.

Further optimizations to improve ReadUntil performance would also be desirable, given that the relatively modest ~5-fold median on-target coverage enrichment observed here is only sufficient to enable accurate STR profiling on a single patient sample per MinION flow cell. Our benchmarking experiments indicate that the ReadUntil rejection time is influenced by the available computer hardware and that this is a key variable in determining the overall assay performance. We anticipate that further engineering to reduce latency in the Readfish/ReadUntil software will deliver improved on-target enrichment and target coverage in the future.

While the potential benefits for genetic diagnosis of patients with STR expansion disorders are clear, targeted STR sequencing will be similarly useful as a research tool. STRs are highly polymorphic and exhibit pathogenicity through an array of different mechanisms. Much remains to be learned about the basic biology of STR expansion disorders and the distinction between a benign and pathogenic allele, particularly for recently described STR disease genes like *RFC1* (9), *GIPCI* (51), *LRP12* (52), *NOTCH2NLC* (14, 52), and *VWA1* (10). By resolving the full diversity of STR sizes and motif conformations in clinically affected and nonaffected individuals, our targeted ONT

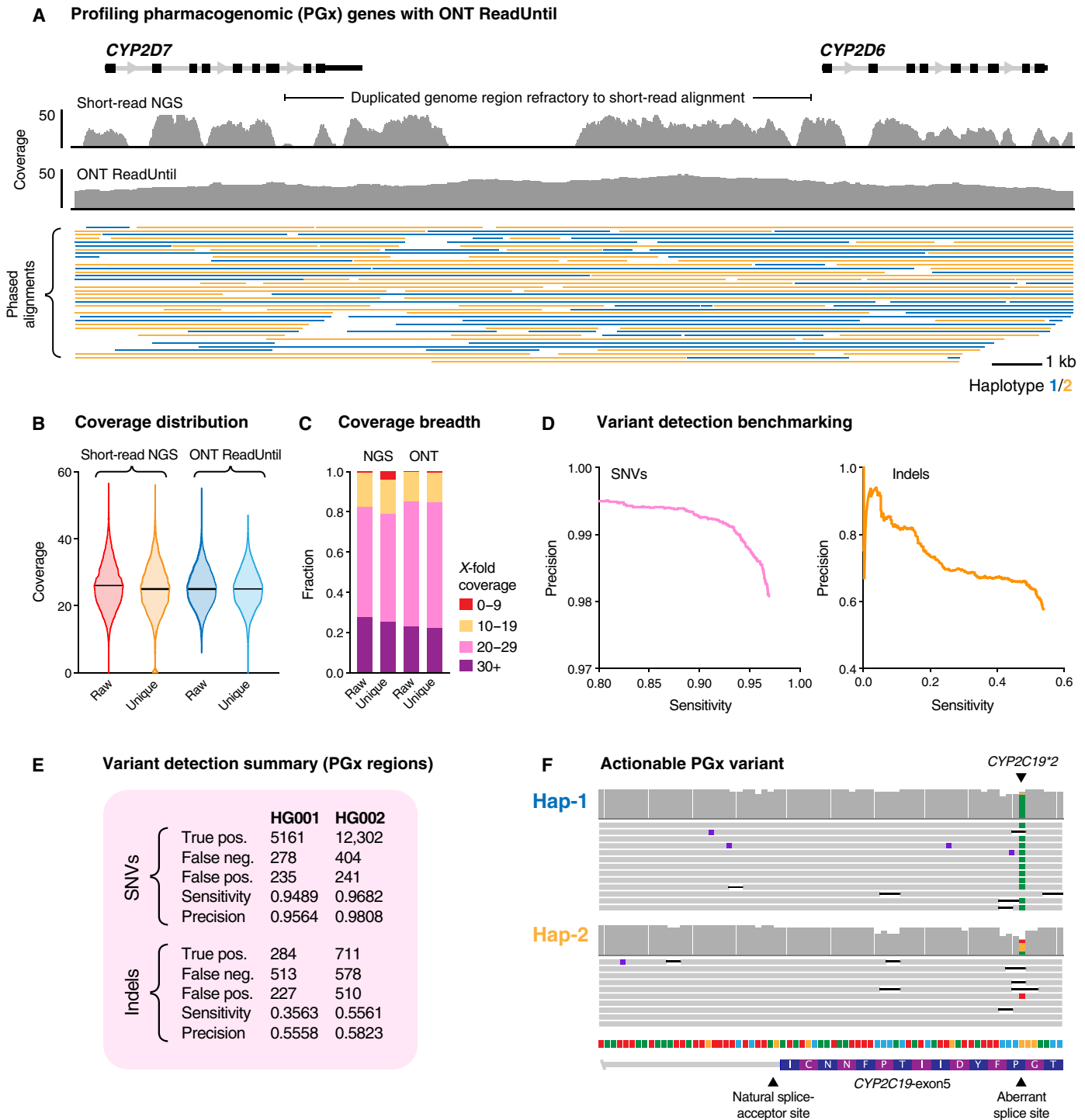


Fig. 5. Targeted ONT sequencing of PGx genes. (A) Genome browser view shows coverage distribution for uniquely aligned site reads (MapQ ≥ 30) at the *CYP2D6* gene and neighboring pseudogene *CYP2D7*. The top track shows data for short-read whole-genome sequencing (Illumina NovaSeq) of the human reference sample HG001/NA12878 (see table S3). The bottom track shows coverage and phased alignments for ONT ReadUntil targeted sequencing on the same sample. (B) Violin plots show coverage distribution across PGx gene targets ($n = 37$) with short-read NGS (left) and targeted ONT sequencing (right). For each technology, both raw alignment coverage and unique alignments (MapQ ≥ 30) are shown. (C) For the same datasets, stacked barcharts show the fraction of PGx target regions covered at different sequencing depths (red, 0 to 9x; yellow, 10 to 19x; pink, 20 to 29x; and purple, ≥ 30 x). (D) Precision-recall curves show the accuracy of variant detection within PGx gene targets for SNVs (left; pink) and indels (right; orange) using ONT ReadUntil on reference sample HG002. Precision-recall curves were used to determine optimum Nanopolish parameter settings. (E) Summary table showing final variant detection statistics within PGx targets. While indel accuracy is poor, SNVs were detected with relatively high sensitivity and precision. (F) Genome browser view showing an example of a clinically actionable PGx allele (*CYP2C19*2*) detected using ONT ReadUntil in HG001.

sequencing assay may be applied to better define pathogenic boundaries and investigate the role of phenotypic modifiers, such as internal STR interruptions. For example, in our relatively small cohort, we observed several previously unknown STR alleles in genes such as *STARD7* and *BEAN1*, the physiological relevance of which is currently unknown. We anticipate that elucidation of the full complement of STR expansion size, motif, and interruptions in health and disease will reveal previously unknown genotype-phenotype correlations, enabling better understanding of the pathomechanisms and facilitating rationale treatment design.

Haplotype-resolved DNA methylation profiling of STR expansion genes by ONT sequencing will also shed light on their epigenetic regulation and potential roles in pathogenic mechanisms. DNA methylation has been extensively studied in males with FXS, wherein expanded CGG STR alleles trigger hypermethylation of the *FMR1* promoter, resulting in gene silencing (30, 31). Consistent with this, we observed promoter hypermethylation in FXS-affected males. Furthermore, we identified haplotype-specific hypermethylation of the *FMR1* promoter in one female who was a FXS premutation carrier. While previously observed, the significance of DNA methylation in premutation carriers is not fully understood (33, 34) and can only be detected using haplotype-aware methodologies, such as ONT sequencing. DNA hypermethylation has also been described in some carriers of pathogenic STR expansions in *C9orf72*; however, the association between methylation status, repeat size, and clinical phenotype is not clear (53), warranting further interrogation with improved molecular tools. Little is known about the epigenetic regulation of most other genes on our targeted sequencing panel, highlighting the need for further investigation, with ONT sequencing being a powerful tool for this purpose.

Last, by resolving STR expansions that are not amenable to existing techniques, long-read sequencing promises to accelerate the discovery of repeat expansion genes and disorders (3, 4). While targeted sequencing is not suitable for the discovery of genes with no prior evidence, the flexible nature of ReadUntil sequencing is ideal for profiling tens/hundreds of candidate genes/regions, such as those identified by linkage mapping in affected families (14, 16, 54). Moreover, repeat expansions need not be the only pathogenic variants found by this approach, with targeted long-read sequencing also suitable for the detection of other types of structural variation (55). We anticipate that this will be a powerful approach to STR gene discovery and provide molecular diagnoses for many previously unsolved cases in the future.

METHODS

Sample collection, processing, and molecular testing

Patient-derived genomic DNA reference samples (NA* and CD* prefixes) and accompanying clinical notes were obtained from Coriell Institute (www.coriell.org/). Molecular testing results were available for STR expansions, as relevant for each individual's phenotype and/or family history. Upon receiving, genomic DNA was resuspended in nuclease-free water at ~100 ng/μl and stored at -20°C.

Patients consulting at neurology clinics in New South Wales (NSW) were consented for genomic analysis by nanopore sequencing under St Vincent's Hospital Human Research Ethics Committee protocol 2019/ETH12538. Patients in Western Australia (WA) were consented under Human Research Ethics Committee of the University of Western Australia protocol 2019/RA/4/20/1008. Deidentified patient

samples were subjected to diagnostic testing in certified clinical laboratories, according to current clinical practice, and molecular test data were provided for this study. STR sites in *RFC1* (for patients with CANVAS), *NOTCH2NLC* (for NIID), *FXN* (for FRDA), and *AR* (for SBMA) were tested by RP-PCR, and large expansions in *RFC1*, *DMPK*, and *FXN* were further analyzed by Southern blot to determine STR sizes. High-molecular weight (HMW) genomic DNA was extracted from patient blood samples using the Qiagen Genra PureGene Blood Kit (NSW) or the QIA Symphony DSP DNA Midi Kit (WA) and suspended in nuclease-free water. Deidentified samples were transferred to the Kinghorn Centre for Clinical Genomics for nanopore sequencing analysis. Full sample descriptions are provided in table S3.

ONT library preparation and sequencing

Before ONT library preparations, the DNA was sheared to ~15-kb fragment size using Covaris G-tubes and visualized, after shearing, on an Agilent TapeStation. Nanopore sequencing libraries were prepared from ~1.5 to 5 μg of HMW DNA, using native library prep kits (SQK-LSK109 or SQK-LSK110), according to the manufacturer's instructions. Each sample was loaded onto an ONT MinION flow cell (R9.4.1) and sequenced on either an ONT GridION or ONT MinION device with live target selection/rejection executed by the Readfish software package (see below) (20). Samples were run for a maximum duration of 72 hours, with nuclease flushes and library reloading performed at approximately 24- and 48-hour time points to maximize sequencing yield.

Programmable target selection

Targeted sequencing was performed using the open-source software package Readfish (20) that internally uses the ONT ReadUntil API for live rejection of off-target sequencing fragments. The targets used in this study were whole gene loci ±50 kb of flanking genome sequence to ensure coverage of promoter and UTR regions and other local regulatory elements. Our targeted sequencing panel included genes containing pathogenic STR sites associated with neurological disease ($n = 37$; see table S1), as well as other potentially informative targets, such as PGx genes exhibiting genotype-drug relationships designated as clinically actionable by the Clinical Pharmacogenetics Implementation Consortium ($n = 28$) and genes harboring clinically actionable Mendelian mutations, as designated by the American College of Medical Genetics ($n = 59$). A full description of ReadUntil targets is provided in table S2.

We installed Readfish version 0.0.5a1 inside a Python virtual environment (with system default Python version given in table S5) using `pip`. When the study was commenced, this Readfish version (0.0.5a1) was the version that supported the available MinKNOW-core and Guppy versions (major version 4). However, the dependencies (`ont-pyguppy-client-lib` and `pyguppyclient`) installed with Readfish v0.0.5a1 by default were incompatible with the Guppy versions on our machines. That is, the support for the NVIDIA 3090 GPU (Graphics Processing Unit) card on our workstation was only introduced in Guppy v4.2.2, and we did not want to interfere with the Guppy v4.2.3 already installed on the GridION. Therefore, we manually installed `ont-pyguppy-client-lib` and `pyguppyclient` (versions in table S5) with minor source code changes to accommodate the Guppy versions.

We created a new `minKNOW` configuration profile under `/opt/ont/minknow/conf/package/sequencing/sequencing_MIN106_DNA_readfish_real.toml` with equivalent configuration values except

that the value for the `break_reads_after_seconds` attribute was changed from 1.0 to 0.4, as instructed in the Readfish documentation. The reference index for Readfish was created using `minimap2` (56) with `-x map-ont` profile using the hg38 genome with alternate contigs excluded. Experiments run for this study use the high Guppy accuracy base-call configuration. Readfish parameters were configured such that (i) if the query sequence aligns to one or more locations within the target genome regions, then sequencing proceeds with no further checking (`single_on = "stop_receiving," multi_on = "stop_receiving"`); (ii) if the query sequence aligns to one or more regions not in the desired list, then the read is rejected by `ReadUntil` (`single_off = "unblock," multi_off = "unblock"`); and (iii) if the base-called sequence is unavailable or if the query sequence is unmappable, then the sequence continues, and the read is rechecked in the subsequent round (`no_seq = "proceed," no_map = "proceed"`).

Haplotype-resolved assembly and methylation profiling of STR sites

Raw ONT sequencing data were base-called using Guppy (4.4.1), and reads with mean quality < 7 were excluded from further analysis. Resulting FASTQ files were aligned to the hg38 reference genome using `minimap2` (v2.14-r883) (56).

To analyze STR sites, we retrieved all alignments within a 50-kb window centered on a given STR site. These were converted back to FASTQ format, assembled *de novo* using `Flye` (v2.8.1-b1676) (57), and polished using `Racon` (v1.4.0) (58) to generate a pseudo-haploid contig encompassing the STR region. Starting reads were realigned to this contig and then assigned to separate alleles/haplotypes via each of three methods: (i) phasing by consideration of heterozygous SNVs within the STR region using `Longshot` (v0.4.1) (59), (ii) phasing by consideration of heterozygous SVs (Structural Variants) at the STR site with respect to the assembled contig using `Sniffles` (v1.0.9) (60), and (iii) a custom all-versus-all pairwise sequence alignment and similarity clustering method that identifies allelic differences in STR size and sequence, based on read clustering, with no prior information. After phasing, the initial assembled contig was repolished separately with reads from either allele/haplotype using `Racon` (v1.4.0) (58), yielding two haploid contigs encompassing the STR site. The position of the relevant STR site was identified in each contig by mapping 150-bp unique flanking sequences (extracted from hg38) using `minimap2` (v2.14-r883) (56). STR size, motif, and summary statistics were retrieved using `Tandem Repeat Finder` (4.09) (61), followed by manual inspection and motif counting. This process was performed separately for each STR site in each individual sample.

For DNA methylation analysis, FAST5 files containing ONT signal data were converted to compressed binary SLOW5 format (62) using `slow5tools` (<https://github.com/hasindu2008/slow5tools>). DNA methylation calling was run on all on-target alignments (with respect to hg38) using `F5C call-methylation` (v0.6) (63), which is a GPU-accelerated version of the popular Nanopolish software package (44). Meth-calling results were retrieved for all phased reads at each STR site (see above) based on read IDs, and methylation frequencies were computed separately for each allele/haplotype, as well as for total unphased reads, using the `F5C meth-freq`. When enumerating the local methylation status of a given STR site, we considered all CpGs within 1500 bp with at least five meth-called reads on each allele/haplotype. Methylation frequency profiles were converted to bigwig format for visualization in IGV, using `kentutils` (<https://github.com/ENCODE-DCC/kentUtils>).

Genotyping PGx genes

Raw ONT sequencing data were base-called using Guppy (v4.4.1), and reads with mean quality < 7 were excluded from further analysis. Resulting FASTQ files were aligned to the hg38 reference genome using `minimap2` (v2.14-r883) (56) and filtered for unique alignments (`MapQ ≥ 30`). Variant calling within PGx regions ($n = 28$) was performed using `Nanopolish` (v0.13.2) (44), with candidate variants requiring a minimum read depth of 5 (`-d 5`). We compared the resulting variant candidates detected in HG001 (NA12878) and HG002 (NA24385) to high-confidence variant annotations available for these samples from the GIAB project. The comparisons were made with the RTG tools `vcfeval` utility (v3.11) with the `--squash-ploidy` parameter selected. Precision-recall curves created by RTG tools were used to determine optimum variant filtering parameters of `QUAL ≥ 30` and `BCRV ≥ 5`. Performance summary statistics were calculated after applying these filters, using the following definitions

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

To identify actionable PGx variants, we queried filtered variant candidates against PharmVar variant annotations (45).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <https://science.org/doi/10.1126/sciadv.abm5386>

[View/request a protocol for this paper from Bio-protocol.](#)

REFERENCES AND NOTES

1. H. Ellegren, Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
2. J. A. Shortt, R. P. Ruggiero, C. Cox, A. C. Wacholder, D. D. Pollock, Finding and extending ancient simple sequence repeat-derived regions in the human genome. *Mob. DNA* **11**, 11 (2020).
3. S. R. Chintalaphani, S. S. Pineda, I. W. Deveson, K. R. Kumar, An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol. Commun.* **9**, 98 (2021).
4. C. Depienne, J.-L. Mandel, 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
5. F. M. Hantash, D. M. Goos, B. Crossley, B. Anderson, K. Zhang, W. Sun, C. M. Strom, FMR1 premutation carrier frequency in patients undergoing routine population-based carrier screening: Insights into the prevalence of fragile X syndrome, fragile X-associated tremor/ataxia syndrome, and fragile X-associated primary ovarian insufficiency in the United States. *Genet. Med.* **13**, 39–45 (2011).
6. L. Ruano, C. Melo, M. C. Silva, P. Coutinho, The global epidemiology of hereditary ataxia and spastic paraplegia: A systematic review of prevalence studies. *Neuroepidemiology* **42**, 174–183 (2014).
7. T. Pringsheim, K. Wiltshire, L. Day, J. Dykeman, T. Steeves, N. Jette, The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Mov. Disord.* **27**, 1083–1091 (2012).
8. N. Vanacore, E. Rastelli, G. Antonini, M. L. E. Bianchi, A. Botta, E. Bucci, C. Casali, S. Costanzi-Porrini, M. Giacanelli, M. Gibellini, A. Modoni, G. Novelli, E. M. Pennisi, A. Petrucci, C. Piantadosi, G. Silvestri, C. Terracciano, R. Massa, An age-standardized prevalence estimate and a sex and age distribution of myotonic dystrophy types 1 and 2 in the Rome Province, Italy. *Neuroepidemiology* **46**, 191–197 (2016).
9. A. Cortese, R. Simone, R. Sullivan, J. Vandrovicova, H. Tariq, W. Y. Yau, J. Humphrey, Z. Jaunmuktane, P. Sivakumar, J. Polke, M. Ilyas, E. Tribollet, P. J. Tomaselli, G. Devigili, I. Callegari, M. Versino, V. Salpietro, S. Efthymiou, D. Kaski, N. W. Wood, N. S. Andrade, E. Buglo, A. Rebelo, A. M. Rossor, A. Bronstein, P. Fratta, W. J. Marques, S. Züchner, M. M. Reilly, H. Houlden, Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat. Genet.* **51**, 649–658 (2019).
10. A. T. Pagnamenta, R. Kaiyrzhanov, Y. Zou, S. I. Da'as, R. Maroofian, S. Donkervoort, N. Dominik, M. Lauffer, M. P. Ferla, A. Orioli, A. Giess, A. Tucci, C. Beetz, M. Sedghi, B. Ansari, R. Barresi, K. Basiri, A. Cortese, G. Elgar, M. A. Fernandez-Garcia, J. Yip, A. R. Foley,

- N. Gutowski, H. Jungbluth, S. Lassche, T. Lavin, C. Marcelis, P. Marks, C. Marini-Bettolo, L. Medne, A.-R. Moslemi, A. Sarkozy, M. M. Reilly, F. Muntoni, F. Millan, C. C. Muraresku, A. C. Need, A. H. Nemeth, S. B. Neuhaus, F. Norwood, M. O'Donnell, M. O'Driscoll, J. Rankin, S. W. Yum, Z. Zolkipli-Cunningham, I. Brusius, G. Wunderlich, G. E. R. Consortium, M. Karakaya, B. Wirth, K. A. Fakhro, H. Tajsharghi, C. G. Bönnemann, J. C. Taylor, H. Houlden, An ancestral 10-bp repeat expansion in VWA1 causes recessive hereditary motor neuropathy. *Brain* **144**, 584–600 (2021).
11. M. Bahlo, M. F. Bennett, P. Degorski, R. M. Tankard, M. B. Delatycki, P. J. Lockhart, Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res* **7**, F1000 (2018).
 12. H. Dashnow, M. Lek, B. Phipson, A. Halman, S. Sadedin, A. Lonsdale, M. Davis, P. Lamont, J. S. Clayton, N. G. Laing, D. G. MacArthur, A. Oshlack, STRetch: Detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).
 13. R. M. Tankard, M. F. Bennett, P. Degorski, M. B. Delatycki, P. J. Lockhart, M. Bahlo, Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. *Am. J. Hum. Genet.* **103**, 858–873 (2018).
 14. J. Sone, S. Mitsuhashi, A. Fujita, T. Mizuguchi, K. Hamanaka, K. Mori, H. Koike, A. Hashiguchi, H. Takashima, H. Sugiyama, Y. Kohno, Y. Takiyama, K. Maeda, H. Doi, S. Koyano, H. Takeuchi, M. Kawamoto, N. Kohara, T. Ando, T. Ieda, Y. Kita, N. Kokubun, Y. Tsuboi, K. Katoh, Y. Kino, M. Katsuno, Y. Iwasaki, M. Yoshida, F. Tanaka, I. K. Suzuki, M. C. Frith, N. Matsumoto, G. Sobue, Long-read sequencing identifies GGC repeat expansions in NOTCH2NL2 associated with neuronal intranuclear inclusion disease. *Nat. Genet.* **51**, 1215–1221 (2019).
 15. P. Giesselmann, B. Brändl, E. Raimondeau, R. Bowen, C. Rohrandt, R. Tandon, H. Kretzmer, G. Assum, C. Galonska, R. Siebert, O. Ammerpohl, A. Heron, S. A. Schneider, J. Ladewig, P. Koch, B. M. Schuldt, J. E. Graham, A. Meissner, F.-J. Müller, Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.* **37**, 1478–1481 (2019).
 16. S. Zeng, M.-Y. Zhang, X.-J. Wang, Z.-M. Hu, J.-C. Li, N. Li, J.-L. Wang, F. Liang, Q. Yang, Q. Liu, L. Fang, J.-W. Hao, F.-D. Shi, X.-B. Ding, J.-F. Teng, X.-M. Yin, H. Jiang, W.-P. Liao, J.-Y. Liu, K. Wang, K. Xia, B.-S. Tang, Long-read sequencing identified intronic repeat expansions in SAMD12 from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. *J. Med. Genet.* **56**, 265–270 (2019).
 17. A. D. Ewing, N. Smits, F. J. Sanchez-Luque, J. Faivre, P. M. Brennan, S. R. Richardson, S. W. Cheetham, G. J. Faulkner, Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell* **80**, 915–928.e5 (2020).
 18. Y.-C. Tsai, D. Greenberg, J. Powell, I. Höjjer, A. Ameur, M. Strahl, E. Ellis, I. Jonasson, R. M. Pinto, V. C. Wheeler, M. L. Smith, U. Gyllensten, R. Sebra, J. Korlach, T. A. Clark, Amplification-free, CRISPR-Cas9 targeted enrichment and SMRT sequencing of repeat-expansion disease causative genomic regions. *bioRxiv*, (2017).
 19. S. Kovaka, Y. Fan, B. Ni, W. Timp, M. C. Schatz, Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **39**, 431–441 (2021).
 20. A. Payne, N. Holmes, T. Clarke, R. Munro, B. J. Debebe, M. Loose, Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* **39**, 442–450 (2021).
 21. M. Macdonald, A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
 22. N.S. Caron, G. E. B. Wright, M. R. Hayden, Huntington Disease. 1998 Oct 23 [Updated 2020 Jun 11], GeneReviews®, M. P. Adam, H. H. Ardinger, R. A. Pagon (University of Washington, 1993–2021); www.ncbi.nlm.nih.gov/books/NBK1305/.
 23. J. F. Gusella, M. E. MacDonald, J.-M. Lee, Genetic modifiers of Huntington's disease. *Mov. Disord.* **29**, 1359–1365 (2014).
 24. G. E. B. Wright, H. F. Black, J. A. Collins, T. Gall-Duncan, N. S. Caron, C. E. Pearson, M. R. Hayden, Interrupting sequence variants and age of onset in Huntington's disease: Clinical implications and emerging therapies. *Lancet Neurol.* **19**, 930–939 (2020).
 25. S. E. Andrew, Y. P. Goldberg, J. Theilmann, J. Zeisler, M. R. Hayden, A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene: Implications for diagnostic accuracy and predictive testing. *Hum. Mol. Genet.* **3**, 65–67 (1994).
 26. R. J. Hagerman, E. Berry-Kravis, H. C. Hazlett, D. B. Bailey Jr., H. Moine, R. F. Kooy, F. Tassone, I. Gantois, N. Sonenberg, J. L. Mandel, P. J. Hagerman, Fragile X syndrome. *Nat. Rev. Dis. Primers.* **3**, 17065 (2017).
 27. A. J. Verkerk, M. Pieretti, J. S. Sutcliffe, Y. H. Fu, D. P. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. F. Victoria, F. P. Zhang, Identification of a gene (FMR-1) containing a CCG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
 28. E. E. Eichler, J. J. Holden, B. W. Popovich, A. L. Reiss, K. Snow, S. N. Thibodeau, C. S. Richards, P. A. Ward, D. L. Nelson, Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat. Genet.* **8**, 88–94 (1994).
 29. C. M. Yrigollen, L. Martorell, B. Durbin-Johnson, M. Naudou, J. Genoves, A. Murgia, R. Polli, L. Zhou, D. Barbooth, A. Rupchock, B. Finucane, G. J. Latham, A. Hadd, E. Berry-Kravis, F. Tassone, AGG interruptions and maternal age affect FMR1 CGG repeat allele stability during transmission. *J. Neurodev. Disord.* **6**, 24 (2014).
 30. B. de Vries, C. C. Jansen, A. A. Duits, C. Verheij, R. Willemsen, J. O. van Hemel, A. M. van den Ouweland, M. F. Niermeijer, B. A. Oostra, D. J. Halley, Variable FMR1 gene methylation of large expansions leads to variable phenotype in three males from one fragile X family. *J. Med. Genet.* **33**, 1007–1010 (1996).
 31. B. A. Oostra, R. Willemsen, FMR1: A gene with three faces. *Biochim. Biophys. Acta* **1790**, 467–477 (2009).
 32. Y. Gu, Y. Shen, R. A. Gibbs, D. L. Nelson, Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nat. Genet.* **13**, 109–113 (1996).
 33. D. I. Pretto, G. Mendoza-Morales, J. Lo, R. Cao, A. Hadd, G. J. Latham, B. Durbin-Johnson, R. Hagerman, F. Tassone, CGG allele size somatic mosaicism and methylation in FMR1 premutation alleles. *J. Med. Genet.* **51**, 309–318 (2014).
 34. A. G. Hadd, S. Filipovic-Sadic, L. Zhou, A. Williams, G. J. Latham, E. Berry-Kravis, D. A. Hall, A methylation PCR method determines FMR1 activation ratios and differentiates premutation allele mosaicism in carrier siblings. *Clin. Epigenetics* **8**, 130 (2016).
 35. H. Rafahi, D. J. Szmulewicz, M. F. Bennett, N. L. M. Sobreira, K. Pope, K. R. Smith, G. Gillies, P. Diakumis, E. Dolzhenko, M. A. Eberle, M. G. Barcina, D. P. Breen, A. M. Chancellor, P. D. Cremer, M. B. Delatycki, B. L. Fogel, A. Hackett, G. M. Halmagyi, S. Kapetanovic, A. Lang, S. Mossman, W. Mu, P. Patrikos, S. L. Perlman, I. Rosemergy, E. Storey, S. R. D. Watson, M. A. Wilson, D. S. Zee, D. Valle, D. J. Amor, M. Bahlo, P. J. Lockhart, Bioinformatics-based identification of expanded repeats: A non-reference intronic pentamer expansion in RFC1 causes CANVAS. *Am. J. Hum. Genet.* **105**, 151–165 (2019).
 36. S. J. Beecroft, A. Cortese, R. Sullivan, W. Y. Yau, Z. Dyer, T. Y. Wu, E. Mulroy, L. Pelosi, M. Rodrigues, R. Taylor, S. Mossman, R. Leadbetter, J. Cleland, T. Anderson, G. Ravenscroft, N. G. Laing, H. Houlden, M. M. Reilly, R. H. Roxburgh, A. Maori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. *Brain* **143**, 2673–2680 (2020).
 37. F. Akçimen, J. P. Ross, C. V. Bourassa, C. Liao, D. Rochefort, M. T. D. Gama, M.-J. Dicarie, O. G. Barsottini, B. Brais, J. L. Pedrosa, P. A. Dion, G. A. Rouleau, Investigation of the pathogenic RFC1 repeat expansion in a Canadian and a Brazilian ataxia cohort: Identification of novel conformations. *bioRxiv*, 593871 (2019).
 38. K. R. Kumar, A. Cortese, S. E. Tomlinson, S. Efthymiou, M. Ellis, D. Zhu, M. Stoll, N. Dominik, S. Tisch, M. Tchan, K. H. C. Wu, S. Devery, P. J. Spring, S. Hawke, P. Cremer, K. Ng, M. M. Reilly, G. A. Nicholson, H. Houlden, M. Kennerson, RFC1 expansions can mimic hereditary sensory neuropathy with cough and Sjögren syndrome. *Brain* **143**, e82 (2020).
 39. M. A. Corbett, T. Kroes, L. Veneziano, M. F. Bennett, R. Florian, A. L. Schneider, A. Coppola, L. Licchetta, S. Franceschetti, A. Suppa, A. Wenger, D. Mei, M. Pendziwiat, S. Kaya, M. DelleDonne, R. Straussberg, L. Xumerle, B. Regan, D. Crompton, A.-F. van Rootselaar, A. Correll, R. Catford, F. Bisulli, S. Chakraborty, S. Baldassari, P. Tinuper, K. Barton, S. Carswell, M. Smith, A. Berardelli, R. Carroll, A. Gardner, K. L. Friend, I. Blatt, M. Iacomino, C. Di Bonaventura, S. Striano, J. Buratti, B. Keren, C. Nava, S. Forlani, G. Rudolf, E. Hirsch, E. Leguern, P. Labauge, S. Balestrini, J. W. Sander, Z. Afawi, I. Helbig, H. Ishiura, S. Tsuji, S. M. Sisodiya, G. Casari, L. G. Sadleir, R. van Coller, M. A. T. Tijssen, K. M. Klein, A. M. J. M. van den Maagdenberg, F. Zara, R. Guerrini, S. F. Berkovic, T. Pippucci, L. Canafoglia, M. Bahlo, P. Striano, I. E. Scheffer, F. Brancati, C. Depienne, J. Geck, Intronic ATTC repeat expansions in STAR7 in familial adult myoclonic epilepsy linked to chromosome 2. *Nat. Commun.* **10**, 4920 (2019).
 40. N. Sato, T. Amino, K. Kobayashi, S. Asakawa, T. Ishiguro, T. Tsunemi, M. Takahashi, T. Matsuura, K. M. Flanigan, S. Iwasaki, F. Ishino, Y. Saito, S. Murayama, M. Yoshida, Y. Hashizume, Y. Takahashi, S. Tsuji, N. Shimizu, T. Toda, K. Ishikawa, H. Mizusawa, Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)_n. *Am. J. Hum. Genet.* **85**, 544–557 (2009).
 41. K. Ishikawa, Y. Nagai, Molecular mechanisms and future therapeutics for spinocerebellar ataxia type 31 (SCA31). *Neurotherapeutics* **16**, 1106–1114 (2019).
 42. M. V. Relling, W. E. Evans, Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).
 43. U. I. Schwarz, M. Gulilat, R. B. Kim, The role of next-generation sequencing in pharmacogenetics and pharmacogenomics. *Cold Spring Harb. Perspect. Med.* **9**, a033027 (2019).
 44. N. J. Loman, J. Quick, J. T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
 45. A. Gaedigk, S. T. Casey, M. Whirl-Carrillo, N. A. Miller, T. E. Klein, Pharmacogene Variation Consortium: A global resource and repository for pharmacogene variation. *Clin. Pharmacol. Ther.* **110**, 542–545 (2021).
 46. M. R. Botton, M. Whirl-Carrillo, A. L. Del Tedici, K. Sangkuhl, L. H. Cavallari, J. A. G. Agúndez, J. Duconge, M. T. M. Lee, E. L. Woodahl, K. Claudio-Campos, A. K. Daly, T. E. Klein, V. M. Pratt, S. A. Scott, A. Gaedigk, PharmVar GeneFocus: CYP2C19. *Clin. Pharmacol. Ther.* **109**, 352–366 (2021).
 47. PharmGKB; www.pharmgkb.org/variant/PA166154053/clinicalAnnotation.
 48. A.-C. Bachoud-Lévi, J. Ferreira, R. Massart, K. Youssouf, A. Rosser, M. Busse, D. Craufurd, R. Reilmann, G. De Michele, D. Rae, F. Squitieri, K. Seppi, C. Perrine, C. Scherer-Gagou,

- O. Audrey, C. Verny, J.-M. Burgunder, International guidelines for the treatment of Huntington's disease. *Front. Neurol.* **10**, 710 (2019).
49. J. K. Blacato, E. M. Wolfe, P. C. Sacks, Preimplantation genetics and other reproductive options in Huntington disease. *Handb. Clin. Neurol.* **144**, 107–111 (2017).
50. F. J. Rang, W. P. Kloosterman, J. de Ridder, From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).
51. J. Xi, X. Wang, D. Yue, T. Dou, Q. Wu, J. Lu, Y. Liu, W. Yu, K. Qiao, J. Lin, S. Luo, J. Li, A. Du, J. Dong, Y. Chen, L. Luo, J. Yang, Z. Niu, Z. Liang, C. Zhao, J. Lu, W. Zhu, Y. Zhou, 5' UTR CGG repeat expansion in GIPC1 is associated with oculopharyngodistal myopathy. *Brain* **144**, 601–614 (2021).
52. H. Ishiura, S. Shibata, J. Yoshimura, Y. Suzuki, W. Qu, K. Doi, M. A. Almansour, J. K. Kikuchi, M. Taira, J. Mitsui, Y. Takahashi, Y. Ichikawa, T. Mano, A. Iwata, Y. Harigaya, M. K. Matsukawa, T. Matsukawa, M. Tanaka, Y. Shirota, R. Ohtomo, H. Kowa, H. Date, A. Mitsue, H. Hatsuta, S. Morimoto, S. Murayama, Y. Shiio, Y. Saito, A. Mitsutake, M. Kawai, T. Sasaki, Y. Sugiyama, M. Hamada, G. Ohtomo, Y. Terao, Y. Nakazato, A. Takeda, Y. Sakiyama, Y. Umeda-Kameyama, J. Shinmi, K. Ogata, Y. Kohno, S.-Y. Lim, A. H. Tan, J. Shimizu, J. Goto, I. Nishino, T. Toda, S. Morishita, S. Tsujii, Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat. Genet.* **51**, 1222–1232 (2019).
53. J. L. Jackson, N. A. Finch, M. C. Baker, J. M. Kachergus, M. DeJesus-Hernandez, K. Pereira, E. Christopher, M. Prudencio, M. G. Heckman, E. Aubrey Thompson, D. W. Dickson, J. Shah, B. Oskarsson, L. Petrucelli, R. Rademakers, M. van Blitterswijk, Elevated methylation levels, reduced expression levels, and frequent contractions in a clinical cohort of C9orf72 expansion carriers. *Mol. Neurodegener.* **15**, 7 (2020).
54. Z. Cen, Z. Jiang, Y. Chen, X. Zheng, F. Xie, X. Yang, X. Lu, Z. Ouyang, H. Wu, S. Chen, H. Yin, X. Qiu, S. Wang, M. Ding, Y. Tang, F. Yu, C. Li, T. Wang, H. Ishiura, S. Tsuji, C. Jiao, C. Liu, J. Xiao, W. Luo, Intronic pentanucleotide TTTCA repeat insertion in the SAMD12 gene causes familial cortical myoclonic tremor with epilepsy type 1. *Brain* **141**, 2280–2288 (2018).
55. D. E. Miller, A. Sulovari, T. Wang, H. Loucks, K. Hoekzema, K. M. Munson, A. P. Lewis, E. P. A. Fuerte, C. R. Paschal, T. Walsh, J. Thies, J. T. Bennett, I. Glass, K. M. Dipple, K. Patterson, E. S. Bonkowski, Z. Nelson, A. Squire, M. Sikes, E. Beckman, R. L. Bennett, D. Earl, W. Lee, R. Allikmets, S. J. Perlman, P. Chow, A. V. Hing, T. L. Wenger, M. P. Adam, A. Sun, C. Lam, I. Chang, X. Zou, S. L. Austin, E. Huggins, A. Safi, A. K. Iyengar, T. E. Reddy, W. H. Majoros, A. S. Allen, G. E. Crawford, P. S. Kishnani; University of Washington Center for Mendelian Genomics, M.-C. King, T. Cherry, J. X. Chong, M. J. Bamshad, D. A. Nickerson, H. C. Mefford, D. Doherty, E. E. Eichler, Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).
56. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
57. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
58. R. Vaser, I. Sović, N. Nagarajan, M. Sikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
59. P. Edge, V. Bansal, Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
60. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
61. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
62. H. Gamaarachchi, H. Samarakoon, S. P. Jenner, J. M. Ferguson, T. G. Amos, J. M. Hammond, H. Saadat, M. A. Smith, S. Parameswaran, I. W. Deveson, Fast nanopore sequencing data analysis with SLOW5. *Nat. Biotechnol.* [10.1038/s41587-021-01147-4](https://doi.org/10.1038/s41587-021-01147-4) (2022).
63. H. Gamaarachchi, C. W. Lam, G. Jayatilaka, H. Samarakoon, J. T. Simpson, M. A. Smith, S. Parameswaran, GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinformatics* **21**, 343 (2020).
64. A. R. L. Spada, A. R. La Spada, E. M. Wilson, D. B. Lubahn, A. E. Harding, K. H. Fischbeck, Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
65. G. Kühlenbäumer, W. Kress, E. B. Ringelstein, F. Stögbauer, Thirty-seven CAG repeats in the androgen receptor gene in two healthy individuals. *J. Neurol.* **248**, 23–26 (2001).
66. A. La Spada, Spinal and Bulbar Muscular Atrophy. 1999 Feb 26 [Updated 2017 Jan 26]. GeneReviews®, M. P. Adam, H. H. Ardinger, R. A. Pagon (University of Washington, 1993–2021); www.ncbi.nlm.nih.gov/books/NBK1333/.
67. P. Fratta, T. Collins, S. Pemble, S. Nethisinghe, A. Devoy, P. Giunti, M. G. Sweeney, M. G. Hanna, E. M. C. Fisher, Sequencing analysis of the spinal bulbar muscular atrophy CAG expansion reveals absence of repeat interruptions. *Neurobiol. Aging* **35**, 443.e1–443.e3 (2014).
68. K. Yum, E. T. Wang, A. Kalsotra, Myotonic dystrophy: Disease repeat range, penetrance, age of onset, and relationship between repeat size and phenotypes. *Curr. Opin. Genet. Dev.* **44**, 30–37 (2017).
69. J. Deng, M. Gu, Y. Miao, S. Yao, M. Zhu, P. Fang, X. Yu, P. Li, Y. Su, J. Huang, J. Zhang, J. Yu, F. Li, J. Bai, W. Sun, Y. Huang, Y. Yuan, D. Hong, Z. Wang, Long-read sequencing identified repeat expansions in the 5'UTR of the NOTCH2NL gene from Chinese patients with neuronal intranuclear inclusion disease. *J. Med. Genet.* **56**, 758–764 (2019).
70. P. Fang, Y. Yu, S. Yao, S. Chen, M. Zhu, Y. Chen, K. Zou, L. Wang, H. Wang, L. Xin, T. Hong, D. Hong, Repeat expansion scanning of the NOTCH2NL gene in patients with multiple system atrophy. *Ann. Clin. Transl. Neurol.* **7**, 517–526 (2020).
71. Y. Yuan, Z. Liu, X. Hou, W. Li, J. Ni, L. Huang, Y. Hu, P. Liu, X. Hou, J. Xue, Q. Sun, Y. Tian, B. Jiao, R. Duan, H. Jiang, L. Shen, B. Tang, J. Wang, Identification of GGC repeat expansion in the NOTCH2NL gene in amyotrophic lateral sclerosis. *Neurology* **95**, e3394 (2020).
72. W. Y. Yau, J. Vandrovova, R. Sullivan, Z. Chen, A. Zecchinelli, R. Cilia, S. Duga, M. Murray, S. Carmona; Genomics England Research Consortium, V. Chelban, H. Ishiura, S. Tsuji, Z. Jaunmuktane, C. Turner, N. W. Wood, H. Houlden, Low prevalence of NOTCH2NL GGC repeat expansion in white patients with movement disorders. *Mov. Disord.* **36**, 251–255 (2021).
73. D. Ma, Y. J. Tan, A. S. L. Ng, H. L. Ong, W. Sim, W. K. Lim, J. X. Teo, E. Y. L. Ng, E.-C. Lim, B. W. Lim, L.-L. Chan, L. C. S. Tan, Z. Yi, E.-K. Tan, Association of NOTCH2NL repeat expansions with Parkinson disease. *JAMA Neurol.* **77**, 1559–1563 (2020).
74. V. Campuzano, L. Montermini, M. D. Moltò, L. Pianese, M. Cossée, F. Cavalcanti, E. Monros, F. Rodius, F. Duclos, A. Monticelli, F. Zara, J. Cañizares, H. Koutnikova, S. I. Bidichandani, C. Gelleria, A. Brice, P. Trouillas, G. De Michele, A. Filla, R. De Frutos, F. Palau, P. I. Patel, S. Di Donato, J. L. Mandel, S. Coccozza, M. Koenig, M. Pandolfo, Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**, 1423–1427 (1996).
75. S. I. Bidichandani, M. B. Delatycki, Friedreich Ataxia. 1998 Dec 18 [Updated 2017 Jun 1]. GeneReviews®, M. P. Adam, H. H. Ardinger, R. A. Pagon (University of Washington, 1993–2021); www.ncbi.nlm.nih.gov/books/NBK1281/.
76. M. DeJesus-Hernandez, I. R. Mackenzie, B. F. Boeve, A. L. Boxer, M. Baker, N. J. Rutherford, A. M. Nicholson, N. A. Finch, H. Flynn, J. Adamson, N. Kouri, A. Wojtas, P. Sengdy, G.-Y. R. Hsiung, A. Karydas, W. W. Seeley, K. A. Josephs, G. Coppola, D. H. Geschwind, Z. K. Wszolek, H. Feldman, D. S. Knopman, R. C. Petersen, B. L. Miller, D. W. Dickson, K. B. Boylan, N. R. Graff-Radford, R. Rademakers, Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
77. T. Bourinaris, H. Houlden, C9orf72 and its relevance in parkinsonism and movement disorders: A comprehensive review of the literature. *Mov. Disord. Clin. Pract.* **5**, 575–585 (2018).
78. C. Estevez-Fraga, F. Magrinelli, D. Hensman Moss, E. Mulroy, G. Di Lazzaro, A. Latorre, M. Mackenzie, H. Houlden, S. J. Tabrizi, K. P. Bhatia, Expanding the spectrum of movement disorders associated with C9orf72 hexanucleotide expansions. *Neurol. Genet.* **7**, e575 (2021).
79. S. A. Glasmacher, C. Wong, I. E. Pearson, S. Pal, Survival and prognostic factors in C9orf72 repeat expansion carriers: A systematic review and meta-analysis. *JAMA Neurol.* **77**, 367–376 (2020).
80. E. L. van der Ende, J. L. Jackson, A. White, H. Seelaar, M. van Blitterswijk, J. C. Van Swieten, Unravelling the clinical spectrum and the role of repeat length in C9ORF72 repeat expansions. *J. Neurol. Neurosurg. Psychiatry* **92**, 502–509 (2021).
81. H. T. Orr, M.-Y. Chung, S. Banfi, T. J. Kwiatkowski, A. Servadio, A. L. Beaudet, A. E. McCall, L. A. Duvick, L. P. W. Ranum, H. Y. Zoghbi, Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.* **4**, 221–226 (1993).
82. J. X. Lin, K. Ishikawa, M. Sakamoto, T. Tsunemi, T. Ishiguro, T. Amino, S. Toru, I. Kondo, H. Mizusawa, Direct and accurate measurement of CAG repeat configuration in the ataxin-1 (ATXN-1) gene by “dual-fluorescence labeled PCR-restriction fragment length analysis”. *J. Hum. Genet.* **53**, 287–295 (2008).
83. J. Sequeiros, S. Seneca, J. Martindale, Consensus and controversies in best practices for molecular genetic testing of spinocerebellar ataxias. *Eur. J. Hum. Genet.* **18**, 1188–1195 (2010).
84. R. P. Menon, S. Nethisinghe, S. Faggiano, T. Vannocci, H. Rezaei, S. Pemble, M. G. Sweeney, N. W. Wood, M. B. Davis, A. Pastore, P. Giunti, The role of interruptions in polyQ in the pathology of SCA1. *PLoS Genet.* **9**, e1003648 (2013).
85. J.-D. Brisson, C. Gagnon, B. Brais, I. Côté, J. Mathieu, A study of impairments in oculopharyngeal muscular dystrophy. *Muscle Nerve* **62**, 201–207 (2020).
86. A. Semmler, W. Kress, S. Vielhaber, R. Schröder, C. Kornblum, Variability of the recessive oculopharyngeal muscular dystrophy phenotype. *Muscle Nerve* **35**, 681–684 (2007).

Acknowledgments: We thank M. Halmagyi for support in patient recruitment. We thank M. Vella and NVIDIA for the donation of a powerful GPU that was used for ReadUntil experiments. We thank D. Lin and Garvan's DICE team for providing excellent HPC support.

Funding: We acknowledge the following funding support: Australian Medical Research Futures Fund (MRFF) Investigator grant MRF1173594 (to I.W.D.), MRFF Genomics Future Health Missions grant 2007681 (to N.G.L.), Australian National Health and Medical Research Council (NHMRC) Fellowships APP1122952 (to G.R.) and APP1117510 (to N.G.L.), Australian Government Research Training Program (RTP) Scholarship (to C.K.S.), philanthropic support from The Kinghorn Foundation (to I.W.D.), Margaret and Terry Orr Memorial Fund (to N.G.L.), Paul Ainsworth Family Foundation (to K.R.K.), and a Working Group Co-Lead Award from the Michael J. Fox Foundation, Aligning Science Across Parkinson's (ASAP) initiative (to K.R.K.). **Author contributions:** S.R.C., S.S.P., K.R.K., and I.W.D. conceived the project, designed the targeted sequencing panel, and planned experiments. K.N., M.T., V.F., A.C., C.K.S., M.R.D., N.G.L., G.R., M.K., S.S.P., and K.R.K. were involved in patient recruitment and clinical interpretation. C.K.S., C.D.-S., H.H., and A.C. performed diagnostic molecular assays. I.S. processed samples and prepared ONT libraries. H.G. and J.M.F. designed and built custom computer hardware for use in ONT ReadUntil experiments. I.S. and H.G. performed ReadUntil experiments. I.S., S.R.C., H.G., J.M.F., and I.W.D. performed bioinformatic analysis. I.S., S.R.C., and I.W.D. prepared the figures. I.S., S.R.C., K.R.K., and I.W.D. prepared the manuscript with

support from all authors. **Competing interests:** I.W.D. manages a fee-for-service sequencing facility at the Garvan Institute of Medical Research that is a customer of Oxford Nanopore Technologies but has no further financial relationship. H.G. and J.M.F. have received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. The authors declare that they have no other competing financial or nonfinancial interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All software used in this study is free and open source, with the exception of the proprietary ONT base-calling software Guppy. Where permitted under patient ethics protocols, raw sequencing data (FAST5 and FASTQ format) from this study have been uploaded to the NCBI Sequence Read Archive (SRA) under accession no. PRJNA786382.

Submitted 23 September 2021

Accepted 11 January 2022

Published 4 March 2022

10.1126/sciadv.abm5386

Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing

Igor StevanovskiSanjog R. ChintalaphaniHasindu GamaarachchiJames M. FergusonSandy S. PinedaCarolin K. ScribaMichel TchanVictor FungKarl NgAndrea CorteseHenry HouldenCarol Dobson-StoneLauren FitzpatrickGlenda HallidayGianina RavenscroftMark R. DavisNigel G. LaingAvi FellnerMarina KennersonKishore R. KumarIra W. Deveson

Sci. Adv., 8 (9), eabm5386. • DOI: 10.1126/sciadv.abm5386

View the article online

<https://www.science.org/doi/10.1126/sciadv.abm5386>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).