Nucleic Acids Research, 2024, 1–17
https://doi.org/10.1093/nar/gkae1202
Genomics

Enhanced detection and genotyping of disease-associated tandem repeats using HMMSTR and targeted long-read sequencing

Kinsey Van Deynze[1,†], Camille Mumm[1,2,†], Connor J. Maltby[3], Jessica A. Switzenberg[1], Peter K. Todd[3,4] and Alan P. Boyle[1,2,*]

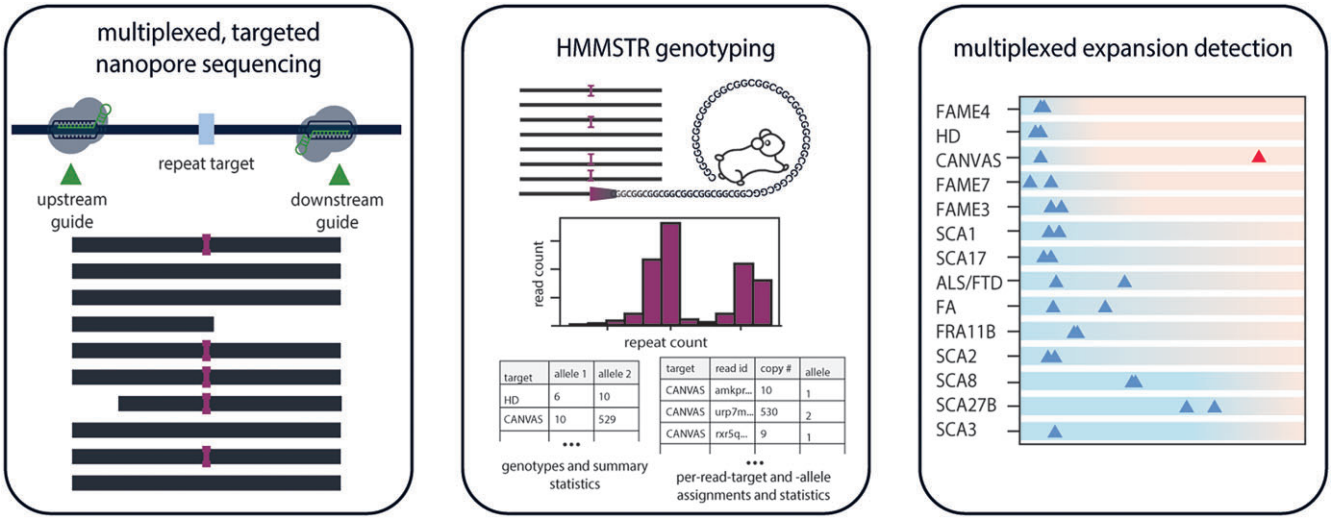
[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[3]Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA
[4]Ann Arbor Veterans Administration Healthcare, Ann Arbor, MI 48105, USA

[*]To whom correspondence should be addressed. Tel: +1 734 763 7382; Fax: +1 734 763 7382; Email: apboyle@umich.edu
[†]These authors contributed equally to this work.

Abstract

Tandem repeat sequences comprise approximately 8% of the human genome and are linked to more than 50 neurodegenerative disorders. Accurate characterization of disease-associated repeat loci remains resource intensive and often lacks high resolution genotype calls. We introduce a multiplexed, targeted nanopore sequencing panel and HMMSTR, a sequence-based tandem repeat copy number caller which outperforms current signal- and sequence-based callers relative to two assemblies and we show it performs with high accuracy in heterozygous regions and at low read coverage. The flexible panel allows us to capture disease associated regions at an average coverage of >150x. Using these tools, we successfully characterize known or suspected repeat expansions in patient derived samples. In these samples, we also identify unexpected expanded alleles at tandem repeat loci not previously associated with the underlying diagnosis. This genotyping approach for tandem repeat expansions is scalable, simple, flexible and accurate, offering significant potential for diagnostic applications and investigation of expansion co-occurrence in neurodegenerative disorders.

Graphical abstract



Introduction

Tandem repeat (TR) sequences constitute about 8% (1) of the human genome, and more than 50 neurodegenerative conditions are associated with short tandem repeat (STR) expansions. Some examples of TR expansion disorders include amyotrophic lateral sclerosis (ALS) and frontotemporal demen-

Received: May 1, 2024. **Revised:** October 16, 2024. **Editorial Decision:** November 15, 2024. **Accepted:** November 19, 2024

tia (FTD), polyglutamine-associated spinocerebellar ataxias, Huntington's disease and myotonic dystrophy (2). A number of these neurodegenerative conditions exhibit overlapping clinical symptoms. Thus, when screening for pathogenic expansions, multiple regions need to be examined. Current molecular methods for TR sizing, such as repeat primed polymerase chain reaction (PCR) and southern blotting, genotype only one locus at a time and are labor intensive. As such, current diagnostic methods remain inefficient (3).

Alternatives to these methods include high-throughput sequencing approaches. The application of short-read sequencing techniques to TR genotyping is limited, as the number of repeat motif copies needed for pathogenicity often exceeds the length of a single read. Thus, their ability to characterize large, low complexity regions has stifled their adoption in clinical settings (4–8). Long-read sequencing methods, such as PacBio and Oxford Nanopore Technologies (ONT), provide an attractive alternative that can sequence through disease associated regions despite their length and complex motif structure. However, these methods come with their own challenges, including relatively high cost and decreased accuracy in repetitive regions (9–11).

While whole-genome, long-read sequencing has been revolutionary in characterizing TRs, it remains resource intensive. Compared to PacBio, ONT sequencing benefits from being more cost effective and accessible. Additionally, multiple targeted sequencing methods have been developed to provide efficient characterization of genes and regions of interest. Targeted sequencing, or the ability to selectively sequence regions of interest, offers an advantage over whole genome sequencing (WGS) by decreasing the amount of reagents and sequencing needed to obtain high coverage over targets (12).

Recently, two techniques have been developed for targeted sequencing using ONT. The first, ReadFish, uses computational tools to select for target fragments during a live sequencing experiment (12). In 2022, this technique was used to successfully genotype disease-associated TRs using the ReadFish API. The second targeted sequencing approach, termed nanopore Cas9-Targeted sequencing (nCATs), uses CRISPR-Cas9 to selectively sequence regions of interest using RNA guides and was found to outperform computational enrichment for TR genotyping (9,13–16). However, the application of nCATS to comprehensively genotyping disease-associated TRs has yet to be extended beyond common ataxias found in European populations (15).

Targeted sequencing techniques have been used in conjunction with existing long-read sequencing bioinformatic genotyping tools and significantly improve on WGS for TR genotyping. Various bioinformatics tools have been developed to overcome sequencing errors for copy number determination from ONT long-read sequencing data. However, many are designed for WGS and prioritize the discovery of novel, large expansions rather than accurately genotyping specified targets (17–19).

Methods designed for PacBio HiFi data, such as TRGT (20), rely on the generation of consensus sequences from highly accurate circular consensus sequencing (CCS). However, comparatively high error rates in repetitive regions, which can impede accurate TR genotyping, necessitate accounting for errors in ONT reads (9). Nanopore-specific, signal-based methods have emerged as a promising approach for directly calling TR copy numbers from nanopore signal data to minimize errors introduced by basecalling (9,13,14). These methods have demonstrated success with targeted sequencing, but they suffer both runtime and storage capacity burdens due to the need for storing and processing the signal data (9,14,15,21).

In response to the current limitations and challenges of genotyping the wide range of disease-associated TRs in a cost-effective manner, our study aims to establish a scalable and easily extendable sequencing and bioinformatics workflow for accurate and efficient copy number determination capable of interrogating all known disease-associated repeat expansion loci. Our strategy combines a multiplexed nCATS approach using a guide pool to target over 50 disease-associated repeat expansion loci on a single ONT MinION flow cell, the most comprehensive panel to date (15). Alongside this, we introduce a profile Hidden-Markov Model STR (HMMSTR) copy-number caller optimized for sequence-based targeted sequencing data. HMMSTR models ONT errors and aims to combine the workflow and accuracy of signal-based copy number callers with the efficiency of sequence-based methods.

Materials and methods

The HMMSTR model

The HMMSTR model is a modified version of a profile HMM (22). The model is made up of five distinct sections: the upstream genome state, prefix states, repeat motif states, suffix states, and downstream genome state (Figure 1A). The prefix, repeat and suffix states follow a three layer (match, insertion, deletion) profile HMM structure. Match state emission probabilities are encoded with the expected base at each position in either the flanking sequence or the repeat motif while insertion state emissions follow a uniform distribution across all observed bases and the deletion states encode a silent character. Emission probabilities at match states reflect mismatch rates based on the expected base at a given position and transition probabilities encode expected rates of insertions and deletions. One copy of the expected motif is encoded in the model and edges are added between the last states in the motif to states at the beginning of the repeat section (Figure 1A). This allows the Viterbi algorithm to find paths through as many repeat motifs as are found in the given sequence. Our model allows for local alignment to the TR and the direct flanking sequence through the use of genome states with emission probabilities following a uniform distribution (default).

Model parameter estimation

Baum-Welch was run on sequences obtained by the alignment of plasmid sets to plasmid backbone sequence per expected repeat count to estimate model parameters (23). While Baum–Welch failed to converge, multiple parameter estimates at later iterations were stable and corresponded well with literature on the same sequencing chemistry for all emission probabilities and combined deletion-insertion rate (10). Notably, Baum–Welch successfully recovered previously reported bias in substitution errors. All other parameters were estimated based on literature. Model parameters can be updated as chemistries and basecallers improve or for custom use (see HMMSTR documentation: https://github.com/BoyleLab/HMMSTR).

Modification of the Viterbi algorithm

The Viterbi (24) algorithm was modified to allow for paths through deletion states without requiring labeled deletions in

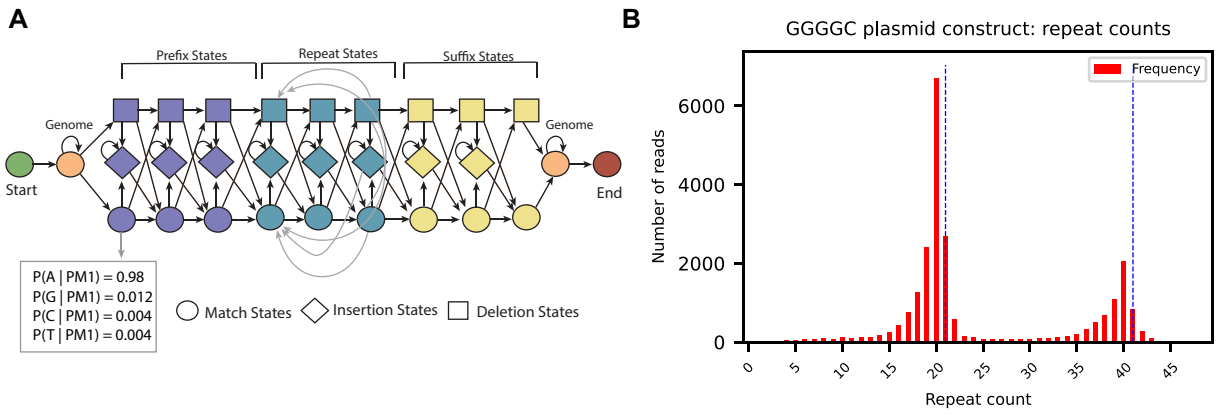


Figure 1. The HMMSTR model and concordance with ground truth plasmid sets. (**A**) A breakdown of the HMMSTR model including sample emission probabilities for a position with an expected 'A' nucleotide. (**B**) Results from GGGGC plasmid benchmarking construct with target repeat lengths 21 and 41, HMMSTR calls 20 and 40. Dashed lines represent ground truth repeat count.

the observed sequence prior to the run. Effectively, the deletion states in the HMMSTR model are treated as silent states where the emission and transition probabilities combined represent the deletion rate and transitions to deletion states do not consume an observed symbol (25). Briefly, we added a third dynamic programming trellis to keep track of when a transition through a deletion state was optimal and allowed for horizontal paths through the traceback trellis to account for these transitions. In this way, we can determine when to index horizontally instead of diagonally and add a deletion character as opposed to an observed character from the input sequence.

The HMMSTR workflow

Read processing

Input flanking sequences are aligned to each read in the sample independently using Mappy (version 2.24) (26). A given read is considered for downstream analysis if (1) it has a valid alignment to both the prefix and suffix of at least one target, (2) the prefix and suffix for at least one target are in the correct orientation with respect to the strand and (3) the prefix and suffix mapq scores exceed the score cutoff threshold (default: mapq 30). If the read satisfies these requirements, it is assigned to all qualifying targets (multiple targets may be considered).

Alternatively, a bam file may be passed as input along with target coordinates and a reference genome. In this case, reads are assigned to targets based on their alignment coordinates using pysam (https://github.com/pysam-developers/pysam).

Once a read passes the initial filtering and assignment step, its sequence is trimmed to include only the 400bp flanking the prefix and suffix alignment positions. This step is used to decrease the runtime of the Viterbi step (which scales to observation length linearly). We choose to keep the larger flanking sequence to mitigate the effect of poor alignment and ensure use of the entire directly flanking sequence.

Repeat motif counting

Repeat copy numbers were counted by taking the total length of the identified repetitive region in the labeled sequence, subtracting the number of insertions identified and dividing by the length of the given motif. Note that deletions are accounted for in the labeled sequence and are thus included in this calculation.

Identifying and filtering outlier repeat copy numbers

HMMSTR has an option to filter outlier repeat counts. This can be useful when dealing with larger datasets where there are more likely off-target reads, high coverage datasets with large tails or PCR products with amplification bias. We designate outliers using the interquartile range (IQR) of the repeat copy number data for a given target. Reads with copy numbers that are outside of this range will be filtered before peak calling is performed.

Summary statistics and peak calling

By default, HMMSTR chooses between Kernel Density Estimation (KDE) with a gaussian kernel and Gaussian Mixture Modeling (GMM) for calling genotypes from the per-read copy number data. Both methods have advantages in distinct situations depending on the distribution of the data.

KDE resolves homozygotes better than GMM in situations where the data has a narrow IQR with few outliers. In this situation, the GMM will overcall heterozygous regions while the KDE is more able to distinguish hetero- and homozygosity. However, the GMM can more accurately detect distributions with larger distance between means and is less often skewed by outliers. A third case is also considered where the quantile range of the data is narrow but there exists few outliers. In this case, a KDE is optimal with the exception of the outliers. Thus, in this case outliers are filtered and a KDE is used to call the genotype since the majority of the data remains in the IQR. This choice can also be overridden as an input along with multiple KDE parameters.

The number of alleles, or zygosity, of a locus is determined either by minimum Akaike information criterion (AIC) amongst GMMs with number of components ranging from 1 to the maximum number of alleles given or given by the number of maxima detected in the kernel density estimate. By default, HMMSTR assumes a diploid sample and the maximum number of alleles called is set to 2, however this is a customizable parameter in HMMSTR.

Plasmid constructs and benchmarking

Four sets of plasmids were constructed using a pcDNA3.1 backbone that contained distinct motifs (AAAAG, AAGGG, GGGGC, and CGG) and two to five motif copy numbers (Supplemental Table S1). The plasmids were restriction en-

zyme digested, pooled based on motif, and sequenced as described in Mumm et al (27) using ONT R9.4.1 Flongle flow cells and SQK-LSK110 kit. Next, the data was basecalled with Guppy 5.0.13. Supplemental Table S1 lists the specific motifs, copy number, and restriction enzymes used for the digestion reactions for each plasmid sequenced. HMMSTR was run on all constructs using 200bp flanking sequence from each backbone as prefix and suffix sequences, the expected number of peaks as max_peaks parameter and either GMM or KDE as the preferred peak calling method based on the noise level of each construct (e.g. KDE was used for CGG plasmid constructs).

To benchmark Straglr on the plasmid sets, all reads were aligned to the corresponding backbone sequence with the shortest repeat count (e.g. AAAAG plasmid reads were aligned to the 16 AAAAG plasmid backbone sequence) and the coordinates for the repeats were calculated from the backbone sequences. Straglr was then run with the resulting bam files with each plasmid backbone as the reference genome and max_num_clusters equal to the expected number of peaks. Both Straglr and HMMSTR output two genotype call files, one corresponding to the final genotype call for a given target and one recording genotype calls per supporting read. Here, final copy number calls used to compare HMMSTR and Straglr accuracy are with respect to the final, per-target, Straglr genotype call. Since Straglr uses TandemRepeatFinder (TRF) to call repeat copy numbers, the per-read copy numbers are reported with respect to the motif reported by TRF for a given read. This results in heterogeneous motif lengths across the reads reported. To account for this in our visualization of the per-read copy number distribution for Straglr, Supplementary Figure S2 shows the distribution of per-read repeat lengths reported by Straglr divided by the expected, underlying, motif length in the given plasmid construct.

WGS benchmarking pipeline

We started all benchmarking set construction from the TR set defined by English et al. (1) which standardizes TR coordinates in the reference genome and annotates hg38 TR calls with fields such as: 'ovl_flag', which categorizes TR annotation based on the nature of their overlapping annotations according to TR Finder; 'n_subregions', which designates how many subregions exist within a TR annotation; 'n_annos', which designates the number of annotations remained after filtering redundant annotations; and 'mu_purity', which describes the percent match between copies and reflects the degree of unbroken repeat units on average across annotations (in the case of multiple annotations for the same region) (28,29). The annotated bed file for hg38 and a full description of these fields can be on github here: https://github.com/ACEnglish/adotto/.

To decrease ambiguity in interpreting results from tools that may treat more complex TR regions differently in terms of copy number calling, we selected regions successively by the following flags that resulted in the number of regions, respectively:

(1) Full TR set: 1.78 M regions
(2) Simple TR annotations ('ovl_flag' = 0 × 1): 1.2 M regions
(3) TRs with only 1 annotation filtered (n_filtered = 1): 411 323 regions
(4) TRs with only 1 annotation ('n_annos' = 1): 325 653 regions
(5) Over 80% purity with respect to the reference genome ('mu_purity' > 80): 324 779 regions

We assess accuracy metrics based on copy numbers from a PacBio HiFi assembly from the Human Genome Structural Variation Consortium (HGSVC) (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20200417_Marschall-Eichler_NBT_hap-assm/). Assembly genotypes were found by aligning the 2000-bp flanking sequence using Mappy to each assembly haplotype followed by copy number determination using TRF. Regions were retained in the benchmark set if they met the following conditions in assessing assembly copy numbers:

- Both 2000-bp flanking sequences aligned to both assembly haplotypes with mapq score of 60 such that only unique mappings were kept.
- The alignments to a given region were in the correct orientation with respect to the assembly strand.
- The reference motif according to our starting annotations was recovered by TRF when assessing copy number from the assembly.

This procedure recovered genotypes for 187 149 regions from our starting set. To note, our assembly genotyping procedure requires all regions we benchmark against to be reference motifs in GM12878.

For our accuracy assessment, we then divided these regions to include only homozygous and heterozygous regions for independent assessment. Regions were considered homozygous if the copy numbers in the assembly were equal across haplotypes and confidently heterozygous if the region had over a two-copy difference. The final benchmarking set numbers are as follows:

- 4237 heterozygous regions.
- 168 568 total homozygous regions.

We chose to only include homozygous regions from chromosome 1 for the sake of the runtime of both HMMSTR and RepeatHMM. This resulted in 15 224 homozygous regions. The following criteria were used for inclusion of regions in accuracy assessment:

- All tools must successfully return a non-null genotype for a given region.
- All tools must call the genotype with respect to motifs with the same length as the reference genome.

Exact numbers for the number of regions called by each tool and the number of motifs called as a different length than the reference genome by Straglr can be found in Supplemental Tables S2 and S3.

GM12878 WGS benchmarking datasets

Both the GM12878 ONT WGS dataset (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md) and the PacBio CCS WGS dataset (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA540705) were downloaded. Reads overlapping the starting TR benchmarking set were then extracted using bedtools intersect.

HMMSTR, Straglr, and RepeatHMM benchmarks

HMMSTR was run on the GM12878 benchmarking sets in with –mapq_cutoff 60 and –discard_outliers on the respective bam input file and with –mode corresponding to the given dataset (ONT or PacBio dataset). Since HMMSTR returns allele copy number calls with respect to both the mode and the median of a given cluster, the median call was used for all GM12878 benchmarking analyses. Straglr 1.41 was run on the same input bam and bed files with default settings. RepeatHMM was run on a pattern (pa) file generated from the bed file used for Straglr and HMMSTR using a custom script and was run in BAMinput mode with –SeqTech corresponding to the dataset (PacBio or Nanopore).

Both the full datasets used and the regions of interest had approximately 30× coverage and the number of regions genotyped by each tool has been reported in Supplemental Tables S2 and S3 for both GM12878 benchmarks.

Assigning haplotype comparison sets

The copy number calls per tool were sorted by size and labeled as heterozygous and homozygous for all benchmarking results. For regions identified as homozygous in the assembly, all genotype calls (homozygous and heterozygous) were compared to the assembly allele. The Haplotype 2 (H2) comparison set includes all regions called homozygous as well as the heterozygous calls closest to the assembly allele. The Haplotype 1 (H1) comparison set includes the allele farthest from the assembly call for all regions that were called as heterozygous.

For regions identified as heterozygous in the assembly, all calls were sorted such that H1 contains the smaller allele call and H2 contains the larger allele call for each region. All alleles were compared in this manner. In the case of homozygous calls, the homozygous call was included in both H1 and H2.

Assessing correlation, mean absolute difference, and percent zygosity misclassification

Pearson R correlation between the assembly TRF calls and tool copy number calls was calculated for each of the sets according to the sorting described above using scipy stats pearsonr. Mean absolute difference (MD) was calculated for the same region sets as the correlation as follows:

- sum(|(H1 call) - (H1 assembly call)|)/(total number of regions in H1 set).
- sum(|(H2 call) - (H2 assembly call)|)/(total number of regions in H2 set).

The homozygous and heterozygous misclassification rates were calculated for each tool using all regions called by a given tool as follows:

- sum(# of regions called as homozygous or heterozygous)/(total # of regions in called in the set).

For misclassification rates, for a region to be assigned to ‘homozygous’, only one copy number was reported by the tool. For a region to be assigned to ‘heterozygous’, more than one copy number was reported by the tool.

CHM13 benchmarking

HMMSTR and Straglr were run on regions defined by Fang et al. (9). This set was 439 regions over 200 bp not within 500 bp of another STR. Basecalled reads (Guppy 5.0.7) were downloaded from the telomere-to-telomere (T2T) consortium https://github.com/marbl/CHM13. Reads were aligned to the CHM13 (v2.0) using minimap2 (version 2.26). Reads were then extracted from regions of interest using samtools. HMMSTR was run with –discarded outliers, –mapq_cutoff 60 and maximum peaks parameter of 1. Straglr was run with maximum peaks of 1 and default settings.

Straglr failed to recover the target motif for 11 regions, and these were discarded before the mean absolute difference calculation.

Downsampling analysis was performed by running samtools view -s with fractions of 0.5, 0.33, 0.15, 0.10 and 0.05 for each chromosome. HMMSTR was then run with the parameters stated above.

Runtime comparison analysis

HMMSTR, Straglr and RepeatHMM were run on three 30x downsampled sets of 400 100 and 10 STR targets from the CHM13 dataset. All tools were run on the same input regions using BAM or FASTQ files. HMMSTR and Straglr were run with the same parameters as the CHM13 benchmark and both HMMSTR and Straglr were run with 44 cpus; however, RepeatHMM BAMInput does not support multithreading or maximum allele number as a parameter and was thus run using default parameters. HMMSTR was additionally run with the flanking_size parameter set to 30 to decrease the size of the models for this runtime analysis. Accuracy for this model size was assessed by running HMMSTR with the same parameters and inputs as the GM12878 benchmarking with the addition of flanking_size set to 30 (Supplemental Tables S8 and S9). Runtime and peak memory usage were measured using the Linux time command on Intel(R) Xeon(R) CPU E5-2696 v4 @ 2.20GHz with 256GiB memory.

Defining disease-associated regions

The table outlining the known repeat expansions underlying neurological disorders was created through the adaptation of multiple literature sources combined with manual curation from publicly available genomic data found at NCBI (2,30–34).

Defining normal, intermediate and pathogenic ranges

Normal, intermediate and pathogenic repeat copy number ranges for each disease-associated loci were defined according to Chaisson et al. (34) or Stripy (33). The intermediate range was defined as any copy number between the upper limit of the normal range and the lower limit of the pathogenic range.

Swimlane plot genotype calls

Disease-associated regions were genotyped with HMMSTR with mapq_cutoff of 60 and default parameters with the exception of few loci which required 200-bp flanking sequence for optimal target specificity (see github). The flanking_filter flag was passed to discard reads with spurious sequence.

Estimating softclip or non-through read repeat copy numbers with HMMSTR

HMMSTR does not currently support integration of non-spanning (softclip) reads as support to its genotype calls; how-

ever, for diagnostic purposes, non-spanning reads containing expansions are helpful for determining expansion status. We estimate copy numbers from softclip reads as additional support for expansion calls, but these reads are not included in the HMMSTR estimates.

Softclip reads were identified using samtools view on sample bam files using samtools view. Reads spanning the entirety of the repeat region per target were filtered based on HMMSTR output such that only softclip reads were left. The softclip reads per target were then converted to fasta format independently. Softclip reads were kept separate per target to ensure correct target assignment in the absence of adequate flanking sequence on both sides of the given repeat. Softclip read repeat copy numbers were found with HMMSTR with the following parameters:

- –flanking_size 60
- –mode sr
- –mapq_cutoff 0
- –use_full_read

where the prefix and suffix were made up of the concatenation of the respective repeat motif. This circumvented the need for the read to contain any flanking sequence not included in the softclip reads. These parameters allow Mappy to align the short (60 bp) flanking sequence inputs to the reads successfully and ensure the full read is considered regardless of which repeat the input flanking sequence aligns to. The mapq cutoff ensures the reads will not be thrown out due to multimapping of the repetitive prefix and suffix inputs. The number of repeats found in the flanking sequence were added to the total repeat copy number reported by HMMSTR.

Calculating motif composition

We define motif composition as the underlying *k*-mer sequence making up a given TR, which may include a single or a mix of motifs of the same or different *k* length that can be decomposed into repeating units (1,20,35). uTR (36) was run on reads from each allele per target independently. Motif decompositions were then processed using a custom script, and composite motif compositions were constructed per individual allele as follows: the number of occurrences of a given motif per read was calculated; the median number of occurrences of each motif was taken across all reads; the final composition was calculated as the percent of the total median motif occurrences a given motif made up.

Cell culture

The GM12878 cell line was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. GM12878 was cultured at 37°C, 5% CO_2 in RPMI 1640 media with L-glutamine (11 875 093, ThermoFisher), and supplemented with 15% fetal bovine serum (10 437 028, ThermoFisher) and 1X antimycotic–antibiotic (15 240 112, ThermoFisher). Cells were regularly passed and the media replenished every 3 days.

Dermal fibroblasts were obtained from consenting patients clinically diagnosed with CANVAS spectrum disorder and genetically confirmed to possess biallelic *RFC1* expansions and iPSCs were cultured as described in Maltby et al.(37). ALS patient fibroblasts originated from a 66 yo male Michigan Medicine patient and obtained as skin biopsy punch under local IRB approval (HUM00030934). Tissues were obtained from UV protected areas such as behind the knee/back of the upper leg using a 5.0 mm biopsy punch. The biopsy punch was cut to remove subcutaneous fat, and then pulverized and plated in a small dish containing Fibroblast Medium (DMEM, 10% (v/v) FBS, 1% (v/v) 100x NEAA, 1% (v/v) Pen/Strep, 1.5% (v/v) 1M HEPES) at 37 °C and 5% CO_2 until fibroblasts emerged from the tissue and were maintained at low passage number prior to expanding and banking.

Five post-mortem cerebellum samples were obtained from the University of Michigan Brain Bank with informed consent of the patients or their relatives and the approval of the local institutional review boards (IRB).

Genomic DNA preparation

High molecular weight genomic DNA (HMW gDNA) was extracted from CANVAS patient derived iPSC cell pellets (~30 M cells) using the salting out method detailed in McDonald et al. ALS/FTD fibroblast and GM12878 LCL HMW gDNA was extracted using the Monarch® HMW DNA Extraction Kit for Cells and Blood (T3050L, NEB) following the manufacturers' instructions. Brain tissue HMW gDNA was extracted from 50mg sections using the Monarch® HMW DNA Tissue Extraction Kit (T3060L, NEB) following the manufacturer's protocol with the following changes in the lysis step. Around 40 µL of 10 mg/mL Proteinase K (3 115 879 001, Roche) was added to 580 µL of Tissue Lysis Buffer. The tissue was placed at 56°C for 15 min on a ThermoMixer (Eppendorf) with 2000 rpm mixing then incubated at 56°C for 30 min without agitation.

Guide selection

Guide RNAs (sgRNAs) were designed according to ONT's best practices for nCATS (https://community.nanoporetech.com/docs/plan/best_practice/targeted-amplification-free-dna-sequencing-using-crispr-cas/v/eci_s1014_v1_revf_11dec2018). In Panel 1, three guides were designed 2–5-kb upstream and downstream of 54 targets. These 20-bp sgRNAs guides were chosen using a pipeline consisting of command line ChopChop and CRISPRon (38,39). The nanopore enrichment model was used with ChopChop, where we specified hg38 regions 2–5-kb upstream and downstream of the disease-associated TR target. The candidate guides were then filtered based on strand for nCATs directionality and further scored using an in-house adaption of CRISPRon.

We then inserted the 20bp sgRNA guide into the following template for pooled amplification and transcription based on Gilpatrick et al. (40).

5′-TAATACGACTCACTATAG-***20nt-seq***- GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTT

In subsequent panel iterations (Panel 2 and 3), guides were added for additional targets and removed for high off-targeting (Supplementary Tables S11–S16).

Guide preparation

The oligo pool from TWIST Biosciences was resuspended to 1ng/µL and 1ng was used for amplification with the primers below according to manufacturer instructions with the following PCR conditions: PCR conditions: 95°C 3 min, 20 cycles of 98°C 20 s, 65°C 15 s and 72°C 15 s, 72°C 10 min and then held at 12°C.

T7 anchored fwd (5′-CGCGCGTAATACGACTCACTATAG-3′)

T7 rev (5′-AAGCACCGACTCGGTGCC-3′)

The product of this reaction (24μL) was then used as input for an 8× reaction in a second round of amplification in an eight-tube strip with the same PCR conditions.

The final amplification product was pooled from the tube strip and cleaned up using the QIAquick PCR Purification kit (28 104, Qiagen) and 500 ng was used for *in vitro* transcription using the NEB HiScribe T7 RNA synthesis kit (E2040S, NEB) according to kit instructions. The sgRNA was purified using a Trizol/chloroform clean-up followed by ethanol precipitation, as detailed in McDonald et al.

Library preparation

nCATS library preparation was performed following McDonald et al. (41) with adaptations for LSK114 and pore R10.4.1 chemistry. First, 7.5 ug (or 30 ul of gDNA at a concentration greater than 100 ng/ul) was dephosphorylated in a 40 μL reaction with 6 μL Quick CIP (M0525S, NEB) and 4 μL 10X rCutSmart buffer (B7204S, NEB). This reaction was inverted and gently tapped to mix and then incubated at 37°C for 30 min, followed by 2-min heat inactivation at 80°C.

The Cas9 ribonucleoprotein (RNP) was formed by combining 850ng of *in vitro* transcribed guide RNA pool, 1 μL of a 1:5 dilution of Alt-R *S.p.*Cas9 Nuclease V3 (1 081 058, IDT) or Alt-R *S.p.*HiFi Cas9 Nuclease V3 (1 081 060, IDT) and 1X rCutSmart buffer (B7204S, NEB) in a total of 30 μL. This reaction was incubated at room temperature for 20 min.

Next both the prepped gDNA and RNP are placed on ice before being combined. Around 1 μL of 10mMol dATP and 1.5 μL of Taq DNA Polymerase (M0273S, NEB) was added to the cut reaction and inverted and gently tapped to mix. This reaction is then incubated at 37°C for 30 min for Cas9 cutting and brought to 75°C for a-tailing. CANVAS samples were additionally treated with 2 ul of thermolabile Proteinase K (P8111S) for 15 min at 37°C, followed by heat inactivation at 55°C.

For adapter ligation, the cut reaction is transferred to a 1.5 mL tube. We then added 5 μL T4 DNA ligase (M0202M, NEB) and 5 μL ONT LSK114 Ligation Adapter (LA; SQK-LSK114, ONT). This reaction is inverted to mix and incubated at room temperature for 20 min with rotation. Following ligation, we add 1 volume of 1X TrisEDTA (TE) and invert to mix. Next 0.3X Ampure beads (SQK-LSK114, ONT) are added and incubated for 5 min with rotation followed by 5 min sitting at room temperature without rotation. The beads are then washed twice with 150 μL Long Fragment Buffer (LFB; SQK-LSK114, ONT) followed by incubation with 20–50 μL Elution Buffer (EB; SQK-LSK114, ONT) at 37°C for 30 min. Finally, we loaded the R10.4.1 MinION flow cell following the ONT protocol using 12 μL of the library and sequenced for 72 h.

nCATs data processing

ONT targeted sequencing data was basecalled and aligned using Dorado 0.6.0 using the super accuracy model with CG methylation calling.

Results

To address the current need for comprehensive screening of disease-associated TR genotypes from ONT data, we first developed a TR copy number caller, HMMSTR, which accounts for read level ONT-specific error profiles as a companion bioinformatic tool for a targeted sequencing panel targeting 60 loci known or suspected of repeat expansion. We benchmark HMMSTR against four repeat-containing plasmid constructs as well as two assemblies and compare its performance against current signal- and sequence-based methods compatible with targeted genotyping. Further, we demonstrate our sequencing strategy's performance across three panels in a control cell line and apply it to samples from nine individuals with either cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS) or ALS/FTD. Using this strategy, we are able to genotype disease-associated expansions at previously uncharacterized loci in these individuals.

A modified profile HMM for TR genotype determination from nanopore-targeted sequencing reads

HMMSTR is optimized for targeted sequencing data and, as such, it assumes a population of on-target reads which contain 2–5 kb of sequence flanking the target. HMMSTR takes basecalled read data and performs local alignment of unique flanking sequences to assign reads to the most likely target. Previous work as well as our model parameter estimation ('Materials and methods') have shown that there is a bias in substitution errors in nanopore data such that purines are more likely to be miscalled as other purines and vice versa for pyrimidines (10). Thus, because substitution errors are not symmetrical with respect to strand, HMMSTR constructs two, strand-specific, models per target that reflect this bias and passes assigned reads to the corresponding model.

Similar to previous HMM-based TR callers, such as STRique (14), which uses raw nanopore signal data (fast5), and RepeatHMM (42), which uses basecalled data, HMMSTR uses a modified profile HMM for estimating repeat lengths. In contrast to these methods, however, HMMSTR alone models the unique flanking sequence as basecalled data for a given target and a single copy of the expected repeat motif (Figure 1A, 'Materials and methods'). This model structure allows for fitting of the flanking region to the repeat sequence and has proven successful in other applications.

HMMSTR further utilizes a modified Viterbi algorithm that labels the most likely positions of insertions and deletions before calculating the copy number for a given read. Previous HMM based TR callers have addressed deletion errors in various ways, including encoding deletion states with the transition and emission probabilities of the neighboring match state (42) and the use of silent states (14). Our Viterbi implementation leverages similar logic to silent states such that the transition to a deletion state does not require an emission symbol and non-emitting state probabilities are calculated separately from emitting states ('Materials and methods' section).

Finally, once the repeat copy numbers are calculated across all reads and targets, allele copy numbers are called for each target using either a Gaussian mixture model or KDE ('Materials and methods' section).

Benchmarking HMMSTR genotype calls

To assess the accuracy of HMMSTR, we benchmark against three ground truth sets: four repeat expansion plasmid constructs with defined expansion lengths, a well-annotated PacBio HiFi diploid assembly from the HGSVC (43), and the haploid CHM13 reference genome from the T2T project (44)

We compare the HMMSTR benchmarking results to four additional TR callers compatible with targeted genotyping: Straglr (17), RepeatHMM (42), DeepRepeat (9), and Strique (14). However, we primarily focus our benchmark against Straglr as it is a comparable sequence-based long-read genotyper that is currently utilized by Oxford Nanopore in their EPI2ME STR expansion workflow (https://epi2me.nanoporetech.com/). Additionally, we include comparison to RepeatHMM in our GM12878 comparisons because it is also a sequence-based caller which shares a similar model architecture; however, we exclude it from our plasmid benchmark because it does not allow for multi-allelic genotyping. DeepRepeat and Strique, two targeted methods that call TRs directly from ONT signal data, were not included in the GM12878 benchmark due to the storage and runtime limitations of the signal data (9).

Plasmid construct benchmark

We ran HMMSTR on four plasmid constructs with variable repeat motif and copy number inserts ranging from 16 to 153 copies, including plasmids with AAAAG (16, 31 and 61 copies), AAGGG (16, 31 and 61 copies), GGGGC (21 and 41 copies) and CGG motifs (20, 39, 77, 115 and 153 copies). We show high concordance with all expected repeat copy numbers with a mean absolute difference of 1.39 copies and standard error of 0.675 per allele across all constructs (Figure 1B, Supplemental Figure S1). We observe moderate heterogeneity in constructs, which is likely due to repeat instability in bacterial culture. Because both HMMSTR and Straglr allow for genotyping of multiple alleles, we also ran Straglr on all plasmid constructs and found that it consistently underestimated the repeat insert sizes with a mean absolute difference of 4.24 copies with a standard error of 0.895 across all alleles (Supplemental Figure S2). Straglr failed to detect the largest plasmid insert in the CGG plasmid construct, and this estimate was not included in the mean absolute difference or standard error calculations. Overall, HMMSTR performs with high accuracy on nanopore sequenced plasmid constructs with variable number of repeat inserts.

GM12878 benchmarking

We selected a subset of TR regions from a gold standard set from English et al. (1) to only include simple TRs, that is, TRs with a single repeat motif that had no overlap with other annotated TRs and over 80% purity. In this benchmarking, only reference motifs were used as inputs and the reference motif must have been detected in the assembly to be included in our truth set (see 'Materials and methods' section for details).

Benchmarking homozygous genotype calls

We queried a total of 15 224 regions identified as homozygous by TRF in the assembly from chromosome 1 in the GM12878 ONT (45) and PacBio CCS WGS datasets (46). We ran HMMSTR, Straglr (17) and RepeatHMM (42) with reference motif input on all regions. HMMSTR and Straglr successfully genotyped the majority of queried regions for the ONT and PacBio datasets; however, RepeatHMM failed to call over half of these regions in both datasets, which decreased the number of total regions we compare in this analysis (Figure 2). Furthermore, although Straglr accepts target repeat motifs in its targeted mode, it can override these motifs and call copy numbers relative to the motif called by TRF, which may differ in length. This is problematic in copy number interpretation and consequently all regions where Straglr called a motif with different length than that of the target motif were discarded. In this homozygous benchmark, 477 regions were discarded from the ONT set and 1126 regions were discarded from the PacBio set due to discordant motif calls described above (Supplemental Table S2).

In homozygous regions, we observe HMMSTR performs comparably to Straglr in terms of Pearson R correlation in the ONT dataset (over 0.97) and both outperform RepeatHMM in both haplotype comparison sets (0.92 and 0.94 in H1 and H2, respectively) (Figure 2, Supplemental Figure S3). All three tools share high correlation with assembly calls in the PacBio dataset (Supplemental Figure S4, Supplemental Table S4). In terms of specificity in zygosity calls, Straglr calls the fewest number of homozygous regions heterozygous in the ONT dataset at a rate of 0.28% compared to HMMSTR and RepeatHMM which miscall homozygotes at rates of 4.4% and 17.16% respectively. Misclassification rates in the PacBio dataset follow this trend at lower rates across all tools (Supplemental Table S4). HMMSTR achieves the lowest mean absolute difference from assembly copy number calls across all haplotype comparison sets and datasets compared to Straglr and RepeatHMM (Figure 2, Supplemental Table S4).

Overall, while both HMMSTR and Straglr have comparable correlation with the assembly in homozygous regions, HMMSTR consistently shows lower mean absolute difference across the same regions in both ONT and PacBio test datasets.

Benchmarking heterozygous genotype calls

To validate HMMSTR heterozygous calls, we identified a set of 4237 heterozygous regions which have at least a three copy difference between alleles in the GM12878 PacBio HiFi assembly (Methods). All tools called over 88% of regions successfully in the ONT set (Figure 3) and over 75% in the PacBio dataset (Supplemental Table S5). Straglr called 58 and 108 regions as motifs of differing length from the input that was discarded from this analysis (Supplemental Table S3).

Figure 3 shows that in the ONT dataset, HMMSTR outperforms both Straglr and RepeatHMM in terms of correlation in the H1 comparison set (0.98 versus 0.95) and H2 comparison set (0.98 versus 0.96) respectively as well as in both mean absolute difference and heterozygous misclassification rate (Figure 3, Supplemental Figure S5). While all tools had comparable, high, correlations to the assembly in the PacBio dataset, HMMSTR maintained both lower mean absolute differences and the lowest number of regions miscalled homozygous across both sequencing technologies (Figure 3, Supplemental Table S5, Supplemental Figure S6).

Another current challenge in characterizing TRs is the ability to distinguish multiple alleles without additional genotype information. High resolution TR genotyping is invaluable for not only disease diagnosis but also for understanding disease mechanisms. The ability to accurately genotype similarly sized alleles can aid in diagnosis of repeat expansions where the threshold between normal or intermediate and pathogenic

Tool	Percent of targets successfully genotyped	Pearson R Correlation		Mean absolute difference		Number of misclassified regions	Overall miscall rate
		H1	H2	H1	H2		
HMMSTR	99.64%	0.9712	0.9833	0.6574	0.5222	591	4.4%
Straglr	98.99%	0.9735	0.9800	0.8112	0.8072	29	0.28%
RepeatHMM	49.52%	0.9242	0.9421	1.2858	1.1007	1,241	17.16%

Figure 2. Chromosome 1 homozygous benchmarking. Statistics from benchmarking of HMMSTR, Straglr and RepeatHMM against regions identified as homozygous in the GM12878 HiFi assembly. This includes the percent of queried regions (15 224 total) with non-null genotypes returned, Pearson R correlation and mean absolute difference (MD) between copy number calls from the ONT GM12878 dataset and HiFi assembly TRF estimates, as well as the number of regions incorrectly called as heterozygous with respect to regions genotyped across all three tools. The overall miscall rate is with respect to all calls per tool.

Tool	Percent of targets successfully genotyped	Pearson R Correlation		Mean absolute difference		Overall miscall rate
		H1	H2	H1	H2	
HMMSTR	98.61%	0.9753	0.9843	2.4886	2.4753	37.36%
Straglr	99.41%	0.9480	0.9917	3.8692	3.7107	96.42%
RepeatHMM	88.79%	0.9684	0.9589	3.4420	3.6056	59.89%

Figure 3. Heterozygous benchmarking. Statistics from benchmarking of HMMSTR, Straglr and RepeatHMM against regions identified as heterozygous in the GM12878 HiFi assembly, including the percent of queried regions (4237 total) with non-null genotypes returned as well as Pearson R correlation and mean absolute difference (MD) between copy number calls from the ONT GM12878 dataset and HiFi assembly TRF estimates. The overall miscall rate is with respect to all calls per tool.

length is small, as well as aid in improving ill defined thresholds that may be significant to clinical outcomes. Additionally, high resolution calls allow for more precise heritability analysis where TR haplotypes can be traced through generations and alleles that undergo copy number changes can be better differentiated from similarly sized alleles.

For these reasons, we compare which heterozygous regions are correctly called heterozygous across all tools. Of the regions successfully called by all three tools in the ONT dataset, the largest intersection was between the two HMM-based methods with 1130 regions called correctly heterozygous by both HMMSTR and RepeatHMM followed by 1110 regions called heterozygous only by HMMSTR (Figure 4A). These results are mirrored in the PacBio dataset (Figure 4B). Notebly, Chiu et al. state that Straglr resolves heterozygous TR loci when alleles differ in size by over 100bp. In this heterozygous benchmark, only 80 out of 4237 regions (1.89%) have a difference in allele size of greater than 100 bp, which may contribute to the high misclassification percentages for this tool. We also assessed heterozygous call concordance among regions with allele size difference of over and under 100 bp (Supplemental Figures S7 and S8). Indeed, regions with greater than 100-bp difference between alleles showed higher heterozygous call concordance across all three tools (Supplemental Figure S7A and B, 37.93% in ONT and 33.33% in PacBio) than regions with less than 100-bp difference between alleles (Supplemental Figure S8A and B, 1.25% in both ONT and PacBio).

Since the zygosity of a TR region should be the same across both test datasets as well as the assembly, we compared the concordance of heterozygous calls across these sets per tool for all regions called in both the ONT and PacBio datasets. We find that HMMSTR has the largest percent overlap in heterozygous calls across all three datasets at 59.64% compared to 19.12% and 2.86% overlap from regions called by RepeatHMM and Straglr, respectively (Figure 4C–E). In terms of regions with over 100 bp between alleles, HMMSTR maintained a higher overlap in regions called heterozygous across both ONT and PacBio datasets (75.34%, Supplemental Figure S7C) compared to both Straglr (66.23%, Supplemental Figure S7D) and RepeatHMM (16.67%, Supplemental Figure S7E). In contrast, of regions with allele size difference of under 100 bp, HMMSTR called 59.36%, while Straglr called 1.68%, and RepeatHMM called 19.15% of regions correctly heterozygous across both sequencing technologies (Supplemental Figure S8C–E).

Overall, HMMSTR correctly calls a greater percentage of heterozygous regions heterozygous across both ONT and PacBio datasets for regions with alleles of both small and relatively large base pair difference compared to Straglr and RepeatHMM. While all three tools tested showed high correlation to the GM12878 assembly, HMMSTR achieves more consistent correlation across both alleles in heterozygous regions and returns reference, repeat copy numbers with the lowest mean absolute difference across all sets (Supplemental Tables S6 and S7).

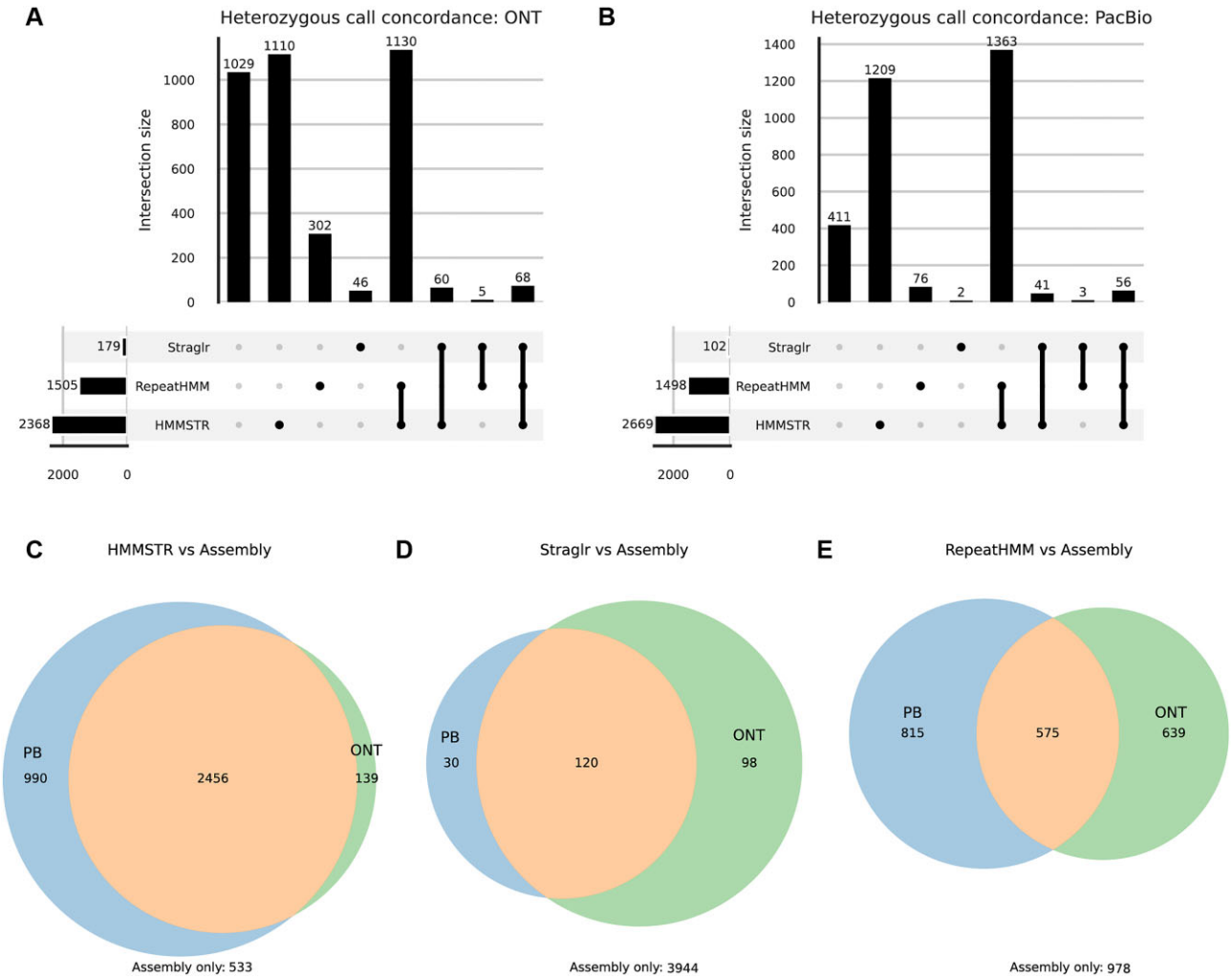


Figure 4. Heterozygous call concordance with the assembly across tools. (**A** and **B**) Upset plots of all regions successfully genotyped by HMMSTR, Straglr and RepeatHMM showing the intersection of regions called as heterozygous in the (**A**) ONT GM12878 dataset and (**B**) PacBio GM12878 dataset. (**C**, **D** and **E**) Venn diagrams of all regions called as heterozygous by each tool across both the ONT and PacBio datasets and the assembly by (**C**) HMMSTR, (**D**) Straglr and (**E**) RepeatHMM.

CHM13 large STR benchmark

Next, we benchmarked HMMSTR and Straglr against a previously published set of large, STR regions (over 200 bp in length) from the CHM13 reference genome (9). CHM13 is known to be effectively haploid (44), thus HMMSTR and Straglr were run with a maximum allele count of one. HMMSTR reports genotypes based on either the mode or median of a given per-read copy number distribution. Here we report mean absolute copy number difference results from both calls separately. Further, Straglr called 11 regions with motifs of differing length than CHM13 annotations that were thus discarded from the mean absolute difference calculation.

We show HMMSTR calls CHM13 STRs with high accuracy (1.1 mean absolute copy difference from mode call, Figure 5A, 1.45 from median call, Supplemental Figure S9A) and Straglr calls the set with a mean absolute difference of 3.77 (Supplemental Figure S9B). This is compared to the previously benchmarked callers on this dataset: DeepRepeat (9), RepeatHMM (42) and STRique (14) with 3.56, 4.47 and 11.2 average absolute differences, respectively (9). The authors of this comparison note that they did not run either STRique or DeepRepeat on the full 126× dataset due to storage limitations; however, we show that we retain lower absolute copy number difference in our downsampled set even at 5× average coverage (2.90 average absolute difference from mode call, 1.87 from median call, Figure 5B). HMMSTR not only allows for the use of basecalled read data, it outperforms both current signal-based and profile HMM-based methods at low coverage.

Comparison of runtime

To simulate targeted sequencing experiments, we ran HMMSTR, Straglr and RepeatHMM on three 30x downsampled STR sets from the CHM13 dataset to measure the effects of dataset size and target number on runtime (Supplemental Figure S10). HMMSTR and Straglr were run with 44 threads; however, RepeatHMM BAMInput does not support multithreading or maximum allele number as a parameter and was thus run using default parameters. By default, HMMSTR models are encoded with 100bp flanking each repeat of interest. This larger model size allows for fitting to TRs which may be situated within regions of similar sequence, such as various regions in our disease associated panel; however,

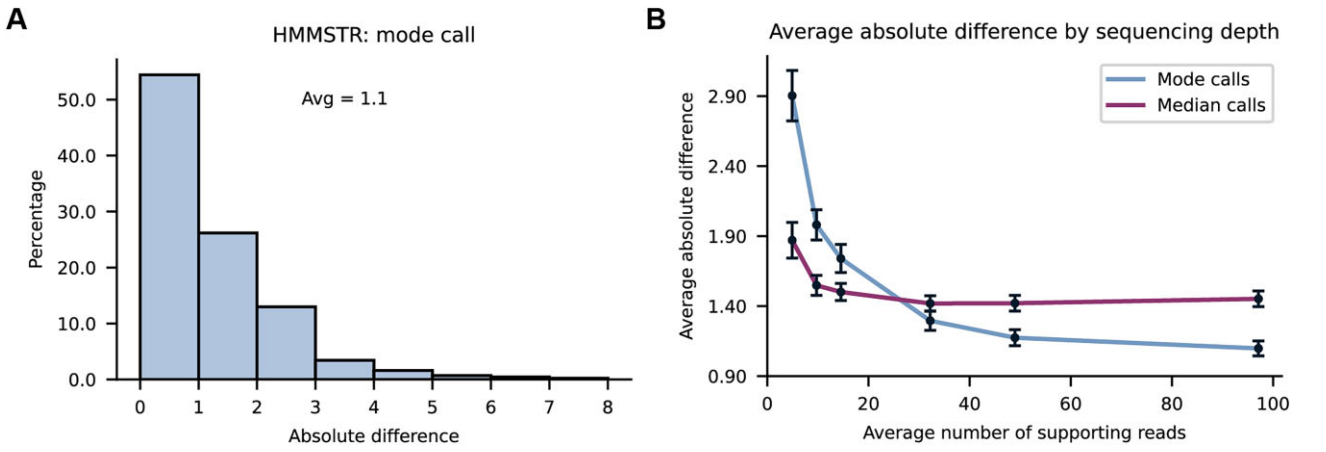


Figure 5. CHM13 HMMSTR Benchmark. (**A**) Average absolute difference between 439 long STR loci in CHM13 and HMMSTR mode calls from nanopore sequencing of CHM13. (**B**) Average absolute difference between CHM13 and HMMSTR median and mode calls from downsampled CHM13 nanopore dataset (5x-100x average coverage). Error bars show standard error per coverage tested.

the Viterbi algorithm scales quadratically with the number of states in the model and thus this model size is not ideal for genotyping at scale. Thus, here we include runtime analysis for running HMMSTR with models encoding 100 and 30 bp of flanking sequence. Further, we show that the 30-bp flanking sequence model encoding results in little to no accuracy decrease in regions not nested within additional repeats, such as those shown in the GM12878 and CHM13 benchmarking (Supplemental Tables S8 and S9).

We find that HMMSTR (∼2.5 s per region with 100-bp model, ∼0.4 s per region with 30-bp model) runs significantly faster than RepeatHMM (∼14 s per region) across all target numbers, with RepeatHMM scaling worse with increased target number while Straglr is ∼4 times faster than HMMSTR using the 30-bp model and ∼25 times faster than HMMSTR using the 100-bp model (∼0.1 s per region).

When we compare HMMSTR and Straglr runtimes on a single plasmid construct dataset, the AAAAG plasmid set, which has a very high read coverage of 137 006×, we find that the difference between Straglr and HMMSTR runtimes decreases (21 versus 25 min for Straglr and HMMSTR, respectively) and that HMMSTR run with 30-bp flanking sequence greatly outperforms Straglr (4 min). This is likely due to the differences in how Straglr and HMMSTR parallelize targets and reads due to their respective use cases: HMMSTR prioritizes multithreading over read processing because it expects input from nCATs, with high coverage and fewer targets, while Straglr is optimized accessing many more targets in WGS data.

HMMSTR processes the 400 target set with under 1.5GB peak memory usage when run with fasta file input. Tests were performed on Intel(R) Xeon(R) CPU E5-2696 v4 @ 2.20GHz with 256GiB memory.

Multiplexed TR-targeted sequencing

We next designed a set of flanking sgRNA guides targeting the majority of known disease-associated TR loci obtained through literature search (2,30–34). These guides can be amplified and transcribed in a pooled approach that offers a simple and flexible multiplexed protocol. This set of guides was then used in the nCATSs approach to perform benchmark enrichments using a control sample with non-pathogenic repeat copy numbers at all loci (see 'Materials and methods' section for details).

Using HMW genomic DNA (gDNA) from GM12878 lymphoblasts, we obtained an average coverage of 250× across 54 targets (Panel 1, Figure 6A). During guide optimization, we designed an additional two panels (panels 2 and 3), where we added disease-associated targets and switched to IDT HiFi Cas9. This improved our on-target rates, where on-target is defined as the number of reads spanning the target TR loci divided by the total number of passing (Q > 9) reads (Figure 6A, Supplemental Figures S11 and S12A, Supplemental Table S10).

Characterization of CANVAS patient-derived iPSCs

After evaluating the targeted sequencing panel in a control cell line, we applied panel 3 to three CANVAS patient-derived iPSC lines which carry large biallelic expansions in *RFC1* (PAT1, PAT3 and PAT4). The autosomal recessive CANVAS expansion in intron two of *RFC1* is highly heterogeneous and multiple pathogenic and non-pathogenic alleles have been described (20,47). While the reference, and most common allele, at this locus is made of up to 400 copies of the 'AAAAG' motif, pathogenic expansions include 'AAGGG', 'ACAGG' and 'AGGGC' motifs which can exceed 5 kb in length (20,47). Thus, long-read sequencing is particularly well suited to characterizing this locus.

Using the disease-associated panel to characterize these three lines, we obtained an average of 142×, 117× and 168× read depth at our target repeat loci (Figure 6, Supplemental Figures S12B and S13). We observe biallelic 'AAGGG' expansions at the CANVAS locus in all three samples, with PAT3 having alleles with 821 and 947 copies, PAT4 harboring 1095 and 1174 copies, and PAT1 with 375 and 1228 copies (Figure 6B). In addition to spanning reads, we can also observe the large repeat expansions directly in soft clipped alignments and detect pathogenic length expansions in these fragmented reads (Supplemental Figure S14).

In the PAT1 sample, we observe not only a biallelic expansion at *RFC1* but also a pathogenic length expansion at *FGF14* of 319 copies of the uninterrupted 'AAG' motif (Supplemental Figure S12B). Interestingly, in PAT4, we also

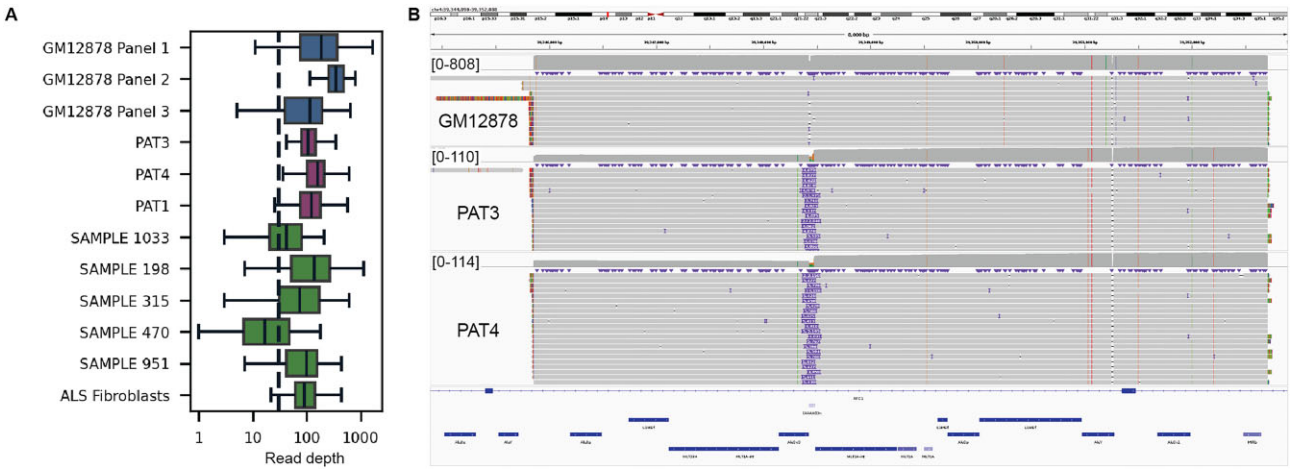


Figure 6. Targeted sequencing panel obtains high coverage over disease-associated STR. (**A**) Average read depth (log10 scale) at targets in 12 samples, including our control cell line GM12878, three CANVAS patient derived iPSC lines,five post-mortem cerebellum samples from individuals with ALS/FTD, and one ALS patient-derived fibroblast line. (**B**) Integrative Genomics Viewer image showing a 6.8-kb region excised with flanking guides at *RFC1*. We obtained 808× read depth at this locus in the GM12878 control with no repeat expansion and 110× and 114× read depth in the CANVAS PAT3 and PAT4, both with large biallelic expansions.

observe two intermediate sized *FGF14* expansions (146 and 211 copies) along with the biallelic *RFC1* expansions.

Survey of ALS/FTD tissue samples for pathogenic repeat expansions

We next applied the 60 target panel (panel 3) to six samples from individuals with neurodegenerative disease. A heterozygous 'GGGGCC' repeat expansion in *C9orf72* is the most common genetic cause of ALS/FTD cases (48). Like the CANVAS expansion, the length of the pathogenic, expanded allele can exceed 5 kb while the CG rich repeat motif can make characterization with amplification based methods more difficult (3,49). Normal copies of the hexanucleotide motif range from 3 to 24 while pathogenic expansions can exceed one thousand copies (33,34,48). In addition, recent work has uncovered additional repeat expansions in NIPA1 and ATXN2 in individuals with ALS/FTD, suggesting potential pleiotropy in pathogenesis (48).

First, we used our method to characterize an ALS patient-derived cell line with a known *C9orf72* expansion. With our targeted sequencing panel, we obtained high coverage at the locus and characterized the heterozygous expansion with HMMSTR. In this line, the pathogenic expanded allele contained 972 copies of the motif and the normal allele was nine copies (Figure 7A). In addition, in this sample we observed a heterozygous expansion in the *RFC1* intron which causes CANVAS. This expanded allele harbors 1487 copies of the pathogenic 'AAGGG' motif and indicates this individual is a carrier for CANVAS (Figure 7B).

Next, we genotyped five post-mortem cerebellum samples where we identified two heterozygous *C9orf72* expansions (Supplemental Figures S14–S16). Overall, due to HMW gDNA quality, the tissue samples yielded lower coverage than experiments using cell lines; however, we were able to successfully genotype most target loci. To investigate the possibility of other expansion loci contributing to disease, we examined the genotypes of these samples at the rest of the disease-associated TR loci. We identified two additional interesting repeat expansions in these individuals (Supplemental Figure S10). In the sample from individual 198, in addition to an expanded *C9orf72* allele, we observed a heterozygous expansion in the spinocerebellar ataxia 31 (SCA31) associated repeat in *BEAN1*. The pathogenic expansion at this locus is a non-reference, nested repeat expansion consisting of 'AAAAT', 'AAGGT' and 'AAGAT' (Supplemental Figure S15A). In this sample, we obtained an inconclusive genotype with approximately 528 copies of an 'AAAAT', 'AAGAT' or 'AATGG' motif that was inconsistent in expanded reads.

We also identified a potentially pathogenic expansion in individual 1033 at the familial adult myoclonic epilepsy 2 (FAME2) associated locus in *STARD7* (Supplemental Figure S14B). Like SCA31 there are pathogenic and non-pathogenic motifs. FAME loci are heterogeneous and several additional motifs have been identified at this locus across populations, but the pathogenicity of some of these motifs is still unclear (50). In this sample, we observed a 552 copy heterozygous expansion of the uncharacterized motif, 'AATAC'.

In addition to the expansions we identified in these samples, we also observed a number of intermediate expansions as well as rare and non-reference alleles that would have been missed using short-read sequencing. Examples include large, reference alleles at FAME loci, and a previously described rare, complex allele at *RFC1* (Supplemental Figure S17) (51).

Discussion

Accurate genotyping is essential for understanding and diagnosing disease-associated TR expansions. Here we establish a computational method designed for targeted sequencing, HMMSTR, that outperforms both signal-based and sequence-based TR copy number callers and provide a targeted sequencing panel for disease-associated TRs. To demonstrate HMMSTR's performance, we benchmark against repeat constructs and assembly data and show high concordance across all sets for both ONT and PacBio HiFi datasets.

Paired with targeted sequencing, HMMSTR's sequence-based approach is efficient and requires fewer resources for storage than WGS or signal-based genotyping. We show HMMSTR outperforms these methods with the lowest mean absolute difference from two assemblies across both ONT

A

Michigan brain bank id	*C9orf72* Genotype		Other expansions identified	Neuropathology notes
	H1	H2		
198	1211	9	*BEAN1* (SCA31)*	frontotemporal lobar dementia with motor neuron disease like inclusions
315	12	9		
470	1074	12		
951	9	9		
1033	15	9	*STARD7* (FAME2)†	motor neuron disease
Fibroblasts	972	9	*RFC1* (CANVAS), *DAB1* (SCA37)*	

B

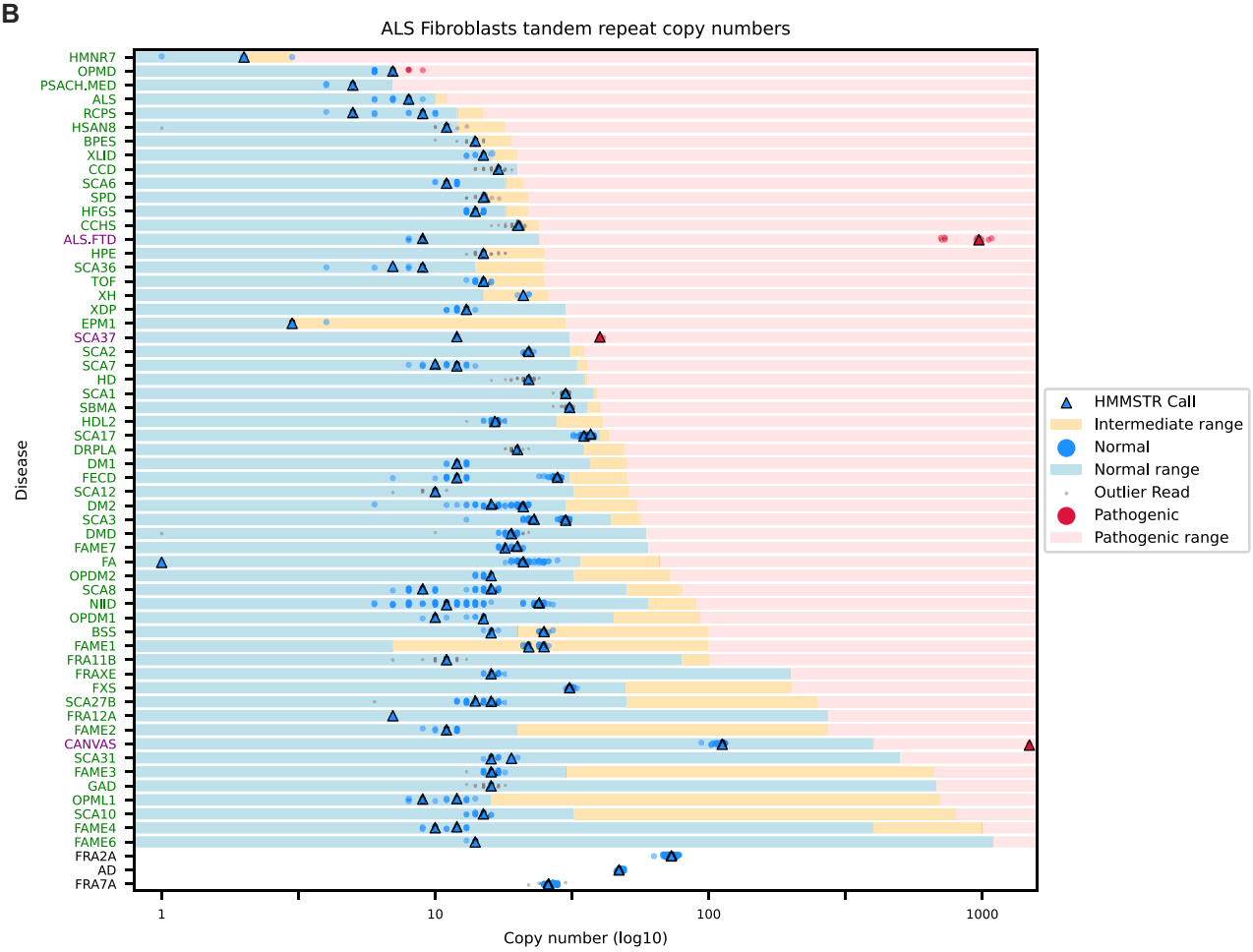


Figure 7. Application of targeted sequencing panel and genotyping with HMMSTR to ALS/FTD samples. (**A**) Sample information and *C9orf72* genotypes in five post-mortem cerebellum samples from individuals with ALS/FTD and one ALS patient-derived fibroblast line. Three samples carry one expanded *C9orf72* hexanucleotide repeat allele. We additionally identified four other disease-associated expanded alleles. Patient 198 has a 536 copy number heterozygous, non-pathogenic repeat expansion in SCA31-associated *BEAN1*. Patient 1033 carries an expanded copy of the *STARD7* with 552 non-reference motifs (Supplemental Figure S11B) and the ALS/FTD fibroblast line carries one pathogenic CANVAS allele. *non-pathogenic motif. †motif of unknown significance. (**B**) The swimlane plot shows the genotypes across 60 disease associated target loci for the ALS/FTD fibroblast line. All markers and ranges are displayed in log10 scale while the *x*-axis reflects the absolute copy number. Dots indicate repeat counts for each read and triangles show HMMSTR median calls, where blue markings correspond to normal allele sizes and red indicate pathogenic length reads and calls. Gray dots indicate outlier copy number calls. For each disease associated locus, the log10 copy number *x*-axis has been shaded to show the ranges of normal, intermediate, and pathogenic repeat copy numbers. For rows without shading, the ranges of normal and pathogenic lengths have not been described. Disease abbreviations (*y*-axis) shown in green indicate HMMSTR call was in the normal range for both alleles, purple indicates one allele was in the pathogenic range, and black indicates missing data or no ranges available (33,34).

and PacBio CCS datasets in homozygous and heterozygous regions as well as across repeat-plasmid constructs with high coverage and up to five allele lengths. Although the accuracy of ONT sequencing has improved significantly, sequencing errors in low complexity regions still impede direct genotyping, particularly at low coverage. HMMSTR models repeat errors to address this and we show it calls repeat lengths at as low as 5x coverage with high accuracy.

We further show that HMMSTR accurately discriminates between similarly sized, heterozygous repeats. Indeed, HMMSTR has the highest level of agreement with heterozygous GM12878 assembly regions in both the PacBio and ONT test datasets and the greatest agreement between the two data modalities. When we compare the regions called heterozygous across all three tools tested, we find that RepeatHMM and HMMSTR have the largest overlap and that RepeatHMM calls heterozygous regions correctly at a higher rate than Straglr. These results suggest that HMM-based methods, which attempt to model technology-specific error rates on a read level, have an increased ability to discern alleles compared to TRF-based Straglr at regions with small base pair separation. Additionally, differences in peak calling and clustering methods across the three tools may contribute to the resolution of each tool's zygosity calls. One of the main differences between how HMMSTR calls zygosity compared to other methods is that it chooses between peak calling with KDE and a Gaussian mixture model (GMM) according to the spread of the per-read repeat copy number distribution ('Materials and methods' section). This is compared to both Straglr and RepeatHMM which both use a GMM to call alleles from their repeat copy number data. This flexible zygosity call precision allows for increased genotype accuracy of both alleles in heterozygous regions (Figure 3) without incorporation of additional genotype information such as neighboring single nucleotide polymorphisms or phase information.

HMMSTR is optimized for targeted sequencing and assumes an enrichment for reads corresponding to the target loci. While HMMSTR's use of local alignment allows for reference-free analysis, it also introduces potential for misassigning reads when there are a large number of off-target reads or targets with similar flanking sequence. This limitation is particularly relevant when attempting to run HMMSTR on WGS data. To account for this and increase target specificity in the analysis of WGS data, HMMSTR also allows for target assignment directly from an aligned bam input file. As some TRs are located within low-complexity regions, by default HMMSTR constructs models encoding 100 bp of sequence flanking each target repeat allowing for high specificity for repeats located in these regions. However, model size can also be expanded or decreased to help with target specificity at the cost of runtime and vice-versa. As shown in Supplemental Tables S8 and S9 and our runtime analysis, decreasing the size of the model can result in significant runtime speedups with minimal accuracy loss, particularly in simple TRs not nested in low complexity regions.

While one advantage of HMMSTR's targeted approach is its specificity and ability to account for errors based on prior knowledge of target loci, challenges persist with targeted models when considering the variability of some disease-associated TRs. Previous studies have reported over 98% of TRs have low sequence polymorphism (20), but multiple disease-associated TRs carry non-reference motifs or complex motif composition. Models constructed on reference motifs generally obtain the same copy number as models constructed with pathogenic motifs due to the similarity in both length and base composition, as well as specificity of flanking sequences. However, we do find that, in some expansion cases with non-reference motifs, HMMSTR may underestimate the copy number using a reference motif model (data not shown). To prevent this shift in copy number due to non-reference motif composition in the expansion cases described in our analysis, HMMSTR was run with reference and pathogenic motifs when applicable for this analysis. Alternatively, HMMSTR can be run with 'N' in the place of expected nucleotides for variable positions for targets with non-reference motifs (eg 'AANNG' for CANVAS) which mitigates this issue. Motif composition can also be recovered in downstream analysis from the per-read repeat coordinates returned by HMMSTR given adequate coverage (Supplemental Figure S17). Integration of alternate motifs into target models would potentially increase the efficiency and ease-of-use of genotyping these variable regions and future iterations of HMMSTR may incorporate these optimizations (20).

Combined with HMMSTR, our pooled guide strategy for genotyping disease-associated TR expansions offers a simple, flexible and accurate screening methods for interrogating a wide array of samples and targets. In addition, this method is highly scalable, as amplification and transcription of the pool can yield sufficient sgRNA for hundreds of samples. Since the design of Panel 3, recent work has identified and described additional disease-associated repeats and we aim to iteratively optimize our panel and add targets as necessary for comprehensive screening (34,52). Moreover, as FDA approved treatments emerge for repeat expansion disorders and they become the subject of gene-targeting clinical trials, critical additional information related to polymorphic repeat structures and alterations in surrounding sequence will prove critical for accurate management of therapeutic options.

We successfully apply our targeted sequencing panel and genotyping to a variety of samples, including two patient groups. Using patient-derived lines and post-mortem cerebellum samples, we are able to genotype large, biallelic repeat expansions at the CANVAS locus in *RFC1* and large, CG rich expansions in *C9orf72*. The rate of co-occurrent repeat expansion in the samples we tested highlights the utility of this targeted panel approach. In two of the three CANVAS samples, we identified potentially pathogenic expansions in *FGF14*, another common cause of late onset cerebellar ataxia with overlapping disease presentation (53,54). Few cases of SCA27B have been reported in CANVAS carriers (55); however, to our knowledge this is the first reported biallelic *RFC1* and *FGF14* expansion to date. In PAT4, we observe two intermediate sized *FGF14* expansions with biallelic *RFC1* expansions. Recent efforts to better characterize the *FGF14* expansion associated with SCA27B have suggested that biallelic, intermediate, expansions where alleles do not exceed the full penetrance threshold of 300 copies may exhibit an additive effect to determine penetrance (53,56). While investigation into the effects of these *RFC1* and *FGF14* co-expansions exceeds the scope of this current work, future application of our panel may elucidate if these co-occurences have an effect on disease etiology or presentation in affected individuals as well as aid in differential diagnosis of these two disorders.

Multiple pathogenic and intermediate TR expansions have been identified in individuals with ALS/FTD with and without the *C9orf72* expansion (48,57). The prevalence and signifi-

cance of these co-occurrent and additional expansions is yet to be determined, highlighting the importance of screening multiple loci (48). In the ALS/FTD samples we analyzed, we identified several expanded and non-reference alleles at the panel targets including in *STARD7, BEAN1, RFC1* and *DAB1*. Recently, whole genome, long-read sequencing projects have uncovered significant heterogeneity in both motif composition and size at these pentanucleotide repeat loci, as well as other disease-associated TRs, across populations (20,58).

Using this targeted panel we are able to efficiently and accurately genotype these loci as well as rare and complex alleles with HMMSTR. This enables an increased power to investigate and characterize a wide range of variation at disease-associated TRs in both cases and controls. However, despite our ability to target our repeats of interest with high coverage, some genotyping challenges remain. Though we observe very high coverage at guide cut sites with targeted sequencing, we obtain fewer reads that span very large expansions. This is consistent with other work and may be due to fragmentation and secondary structure, particularly in samples with pathogenic motifs (15).

In conclusion, we demonstrate the utility of our combined targeted bioinformatic and sequencing strategy. By obtaining high coverage and accurate genotypes across disease-associated loci simultaneously, we are able to not only genotype normal-length repeats and confirm pathogenic expansions but also identify expansions at unexpected loci and co-expansions in samples from individuals with neurodegenerative conditions. This strategy holds promise for more economic and comprehensive diagnostics as well as further study of the diversity at previously elusive TR loci.

Data availability

Data from GM12878 and ALS/FTD Cas9 Enrichments are available at SRA bioproject PRJNA1079777. Additional sample data are available upon reasonable request.

Code availability

HMMSTR is available at https://github.com/Boyle-Lab/HMMSTR and https://doi.org/10.5281/zenodo.14181359.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank members of the Todd, Mills and Boyle labs for helpful discussions and suggestions related to his manuscript. We would like to acknowledge Mr. Matthew D. Perkins for his help with postmortem tissue from the University of Michigan Brain Bank. In addition, we would like to thank the generous patients who contributed their tissues.

Author contributions: A.P.B., P.K.T., K.V. and C.M. conceived the project. C.J.M. and J.S. established and cultured cell lines and isolated gDNA. P.K.T and C.J.M obtained patient tissue samples for culture and screening. C.M. developed the guide panels and performed Cas9 targeted enrichment and nanopore sequencing. K.V. developed and benchmarked HMMSTR as well as performed computational analysis. All authors guided the data analysis strategy. A.P.B., K.V. and C.M. wrote the manuscript. All authors edited the manuscript. All authors read and approved the final manuscript.

Funding

National Institutes of Health [P30AG072931 to the University of Michigan Brain Bank and Alzheimer's Disease Research Center]; NIH NINDS R21 [NS129096 to P.K.T.]; NIH NINDS R01 [NS099280 to P.K.T and A.B.]; NIH NHGRI R21 [HG011493 to A.B.] and NIH NIGMS R01 [GM144484 to A.B.]. A. Alfred Taubman Medical Research Institute at the University of Michigan.

Conflict of interest statement

None declared.

References

1. English,A.C., Dolzhenko,E., Ziaei Jam,H., McKenzie,S.K., Olson,N.D., De Coster,W., Park,J., Gu,B., Wagner,J., Eberle,M.A., *et al.* (2024) Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-024-02225-z.
2. Malik,I., Kelley,C.P., Wang,E.T. and Todd,P.K. (2021) Molecular mechanisms underlying nucleotide repeat expansion disorders. *Nat. Rev. Mol. Cell Biol.*, **22**, 589–607.
3. Ciotti,P., Di Maria,E., Bellone,E., Ajmar,F. and Mandich,P. (2004) Triplet repeat primed PCR (TP PCR) in molecular diagnostic testing for Friedreich Ataxia. *J. Mol. Diagn.*, **6**, 285.
4. Ibañez,K., Polke,J., Tanner Hagelstrom,R., Dolzhenko,E., Pasko,D., Thomas,E.R.A., Daugherty,L.C., Kasperaviciute,D., Smith,K.R., McDonagh,E.M., *et al.* (2022) Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.*, **21**, 234–245.
5. Dolzhenko,E., Deshpande,V., Schlesinger,F., Krusche,P., Petrovski,R., Chen,S., Emig-Agius,D., Gross,A., Narzisi,G., Bowman,B., *et al.* (2019) ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, **35**, 4754–4756.
6. Dolzhenko,E., Bennett,M.F., Richmond,P.A., Trost,B., Chen,S., van Vugt,J.J.F.A., Nguyen,C., Narzisi,G., Gainullin,V.G., Gross,A.M., *et al.* (2020) ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.*, **21**, 1–14.
7. Dashnow,H., Pedersen,B.S., Hiatt,L., Brown,J., Beecroft,S.J., Ravenscroft,G., LaCroix,A.J., Lamont,P., Roxburgh,R.H., Rodrigues,M.J., *et al.* (2022) STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol.*, **23**, 1–20.
8. Mousavi,N., Shleizer-Burko,S., Yanicky,R. and Gymrek,M. (2019) Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.*, **47**, e90.
9. Fang,L., Liu,Q., Monteys,A.M., Gonzalez-Alegre,P., Davidson,B.L. and Wang,K. (2022) DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol.*, **23**, 108.
10. Delahaye,C. and Nicolas,J. (2021) Sequencing DNA with nanopores: troubles and biases. *PLoS One*, **16**, e0257521.
11. Oehler,J.B., Wright,H., Stark,Z., Mallett,A.J. and Schmitz,U. (2023) The application of long-read sequencing in clinical settings. *Hum. Genomics*, **17**, 1–13.
12. Stevanovski,I., Chintalaphani,S.R., Gamaarachchi,H., Ferguson,J.M., Pineda,S.S., Scriba,C.K., Tchan,M., Fung,V., Ng,K., Cortese,A., *et al.* (2022) Comprehensive genetic diagnosis of

tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci. Adv.*, **8**, eabm5386.
13. Sitarčík,J., Vinař,T., Brejová,B., Krampl,W., Budiš,J., Radvánszky,J. and Lucká,M. (2023) WarpSTR: determining tandem repeat lengths using raw nanopore signals. *Bioinformatics*, **39**, btad388.
14. Giesselmann,P., Brändl,B., Raimondeau,E., Bowen,R., Rohrandt,C., Tandon,R., Kretzmer,H., Assum,G., Galonska,C., Siebert,R., *et al.* (2019) Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.*, **37**, 1478–1481.
15. Erdmann,H., Schöberl,F., Giurgiu,M., Leal Silva,R.M., Scholz,V., Scharf,F., Wendlandt,M., Kleinle,S., Deschauer,M., Nübling,G., *et al.* (2022) Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain*, **146**, 1831–1843.
16. Gilpatrick,T., Lee,I., Graham,J.E., Raimondeau,E., Bowen,R., Heron,A., Downs,B., Sukumar,S., Sedlazeck,F.J. and Timp,W. (2020) Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.*, **38**, 433–438.
17. Chiu,R., Rajan-Babu,I.-S., Friedman,J.M. and Birol,I. (2021) Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.*, **22**, 224.
18. Mitsuhashi,S., Frith,M.C., Mizuguchi,T., Miyatake,S., Toyota,T., Adachi,H., Oma,Y., Kino,Y., Mitsuhashi,H. and Matsumoto,N. (2019) Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.*, **20**, 58.
19. Bolognini,D., Magi,A., Benes,V., Korbel,J.O. and Rausch,T. (2020) TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience*, **9**, giaa101.
20. Dolzhenko,E., English,A., Dashnow,H., De Sena Brandine,G., Mokveld,T., Rowell,W.J., Karniski,C., Kronenberg,Z., Danzi,M.C., Cheung,W.A., *et al.* (2024) Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.*, **42**, 1606–1614.
21. Brais,B., Pellerin,D. and Danzi,M.C. (2023) Deep intronic FGF14 GAA repeat expansion in late-onset cerebellar ataxia. Reply. *N. Engl. J. Med.*, **388**, e70.
22. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
23. Baum,L.E., Petrie,T., Soules,G. and Weiss,N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, **41**, 164–171.
24. Forney,G.D. (1973) The viterbi algorithm. *Proc. IEEE Inst. Electr. Electron. Eng.*, **61**, 268–278.
25. Wakabayashi,K. (2019) Silent HMMs: generalized representation of hidden semi-markov models and hierarchical HMMs. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*. Association for Computational Linguistics, Dresden, Germany, pp. 98–107.
26. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
27. Mumm,C., Drexel,M.L., McDonald,T.L., Diehl,A.G., Switzenberg,J.A. and Boyle,A.P. (2023) Multiplexed long-read plasmid validation and analysis using OnRamp. *Genome Res.*, **33**, 741–749.
28. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
29. Fondon,J.W. and Garner,H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. USA*, **101**, 18058–18063.
30. Repeat expansion diseases (2018) In: *Handbook of Clinical Neurology*. Elsevier, Amsterdam, Netherlands, Vol. **147**, pp. 105–123.
31. Depienne,C. and Mandel,J.-L., 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? (2021). *Am. J. Hum. Genet.*, **108**, 764–785.
32. Chintalaphani,S.R., Pineda,S.S., Deveson,I.W. and Kumar,K.R. (2021) An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol. Commun.*, **9**, 1–20.
33. Halman,A., Dolzhenko,E. and Oshlack,A. (2022) STRipy: a graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Hum. Mutat.*, **43**, 859–868.
34. Chaisson,M.J.P., Sulovari,A., Valdmanis,P.N., Miller,D.E. and Eichler,E.E. (2023) Advances in the discovery and analyses of human tandem repeats. *Emerg. Top Life Sci.*, **7**, 361–381.
35. Lu,T.-Y., Smaruj,P.N., Fudenberg,G., Mancuso,N. and Chaisson,M.J.P. (2023) The motif composition of variable number tandem repeats impacts gene expression. *Genome Res.*, **33**, 511–524.
36. Masutani,B., Kawahara,R. and Morishita,S. (2023) Decomposing mosaic tandem repeats accurately from long reads. *Bioinformatics*, **39**, btad185.
37. Maltby,C.J., Krans,A., Grudzien,S.J., Palacios,Y., Muiños,J., Suárez,A., Asher,M., Willey,S., Van Deynze,K., Mumm,C., *et al.* (2024) AAGGG repeat expansions trigger RFC1-independent synaptic dysregulation in human CANVAS neurons. *Sci Adv*, **10**, eadn2321.
38. Labun,K., Montague,T.G., Krause,M., Torres Cleuren,Y.N., Tjeldnes,H. and Valen,E. (2019) CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.*, **47**, W171–W174.
39. Anthon,C., Corsi,G.I. and Gorodkin,J. (2022) CRISPRon/off: cRISPR/Cas9 on- and off-target gRNA design. *Bioinformatics*, **38**, 5437–5439.
40. Gilpatrick,T., Wang,J.Z., Weiss,D., Norris,A.L., Eshleman,J. and Timp,W. (2023) IVT generation of guideRNAs for Cas9-enrichment nanopore sequencing. bioRxiv doi: https://doi.org/10.1101/2023.02.07.527484, 07 February 2023, pre-print: not peer-reviewed.
41. McDonald,T.L., Zhou,W., Castro,C.P., Mumm,C., Switzenberg,J.A., Mills,R.E. and Boyle,A.P. (2021) Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat. Commun.*, **12**, 1–13.
42. Liu,Q., Zhang,P., Wang,D., Gu,W. and Wang,K. (2017) Interrogating the 'unsequenceable' genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.*, **9**, 1–16.
43. Porubsky,D., Ebert,P., Audano,P.A., Vollger,M.R., Harvey,W.T., Marijon,P., Ebler,J., Munson,K.M., Sorensen,M., Sulovari,A., *et al.* (2021) Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.*, **39**, 302–308.
44. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A., *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.
45. Jain,M., Koren,S., Miga,K.H., Quick,J., Rand,A.C., Sasani,T.A., Tyson,J.R., Beggs,A.D., Dilthey,A.T., Fiddes,I.T., *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.
46. Vollger,M.R., Dishuck,P.C., Harvey,W.T., DeWitt,W.S., Guitart,X., Goldberg,M.E., Rozanski,A.N., Lucas,J., Asri,M., Munson,K.M., *et al.* (2023) Increased mutation and gene conversion within human segmental duplications. *Nature*, **617**, 325–334.
47. Dominik,N., Magri,S., Currò,R., Abati,E., Facchini,S., Corbetta,M., Macpherson,H., Bella,D., Sarto,E., Stevanovski,I., *et al.* (2023) Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. *Brain*, **146**, 5060–5069.
48. Henden,L., Fearnley,L.G., Grima,N., McCann,E.P., Dobson-Stone,C., Fitzpatrick,L., Friend,K., Hobson,L., Chan Moi Fat,S., Rowe,D.B., *et al.* (2023) Short tandem repeat expansions in sporadic amyotrophic lateral sclerosis and frontotemporal dementia. *Sci. Adv.*, **9**, eade2044.
49. Fazal,S., Danzi,M.C., Cintra,V.P., Bis-Brewer,D.M., Dolzhenko,E., Eberle,M.A. and Zuchner,S. (2020) Large scale in silico characterization of repeat expansion variation in human genomes. *Sci. Data*, **7**, 1–14.
50. Corbett,M.A., Kroes,T., Veneziano,L., Bennett,M.F., Florian,R., Schneider,A.L., Coppola,A., Licchetta,L., Franceschetti,S.,

Suppa,A., *et al.* (2019) Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. *Nat. Commun.*, **10**, 1–10.
51. Dolzhenko,E., English,A., Dashnow,H., De Sena Brandine,G., Mokveld,T., Rowell,W.J., Karniski,C., Kronenberg,Z., Danzi,M.C., Cheung,W., *et al.* (2023) Resolving the unsolved: comprehensive assessment of tandem repeats at scale. bioRxiv doi: https://doi.org/10.1101/2023.05.12.540470, 14 May 2023, pre-print: not peer-reviewed.
52. (2024) Exonic trinucleotide repeat expansions in ZFHX3 cause spinocerebellar ataxia type 4: a poly-glycine disease. *Am. J. Hum. Genet.*, **111**, 82–95.
53. Ouyang,R., Wan,L., Pellerin,D., Long,Z., Hu,J., Jiang,Q., Wang,C., Peng,L., Peng,H., He,L., *et al.* (2024) The genetic landscape and phenotypic spectrum of GAA-FGF14 ataxia in China: a large cohort study. *Ebiomedicine*, **102**, 105077.
54. Pellerin,D., Wilke,C., Traschütz,A., Nagy,S., Currò,R., Dicaire,M.-J., Garcia-Moreno,H., Anheim,M., Wirth,T., Faber,J., *et al.* (2024) Intronic GAA repeat expansions are a common cause of ataxia syndromes with neuropathy and bilateral vestibulopathy. *J. Neurol. Neurosurg. Psychiatry*, **95**, 175–179.
55. Awad,P.S., Lohmann,K., Hirmas,Y., Hinrichs,F., Thomsen,M., Kauffman,M., Lüth,T., Trinh,J., Westenberger,A., Chaná-Cuevas,P., *et al.* (2023) Shaking up Ataxia: FGF14 and RFC1 repeat expansions in affected and unaffected members of a Chilean Family. *Mov. Disord.*, **38**, 1107–1109.
56. Mohren,L., Erdlenbruch,F., Leitão,E., Kilpert,F., Sebastian Hönes,G., Kaya,S., Schröder,C., Thieme,A., Sturm,M., Park,J., Schlüter,A., *et al.* (2024) Identification and characterisation of pathogenic and non-pathogenic FGF14 repeat expansions. *Nat Commun*, **15**, 7665.
57. Nagy,Z.F., Pál,M., Engelhardt,J.I., Molnár,M.J., Klivényi,P. and Széll,M. (2024) Beyond C9orf72: repeat expansions and copy number variations as risk factors of amyotrophic lateral sclerosis across various populations. *BMC Med. Genom.*, **17**, 1–8.
58. Gustafson,J.A., Gibson,S.B., Damaraju,N., Zalusky,M.P.G., Hoekzema,K., Twesigomwe,D., Yang,L., Snead,A.A., Richmond,P.A., De Coster,W., *et al.* (2024) High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.*, **34**, 2061–2073.

Received: May 1, 2024. **Revised:** October 16, 2024. **Editorial Decision:** November 15, 2024. **Accepted:** November 19, 2024