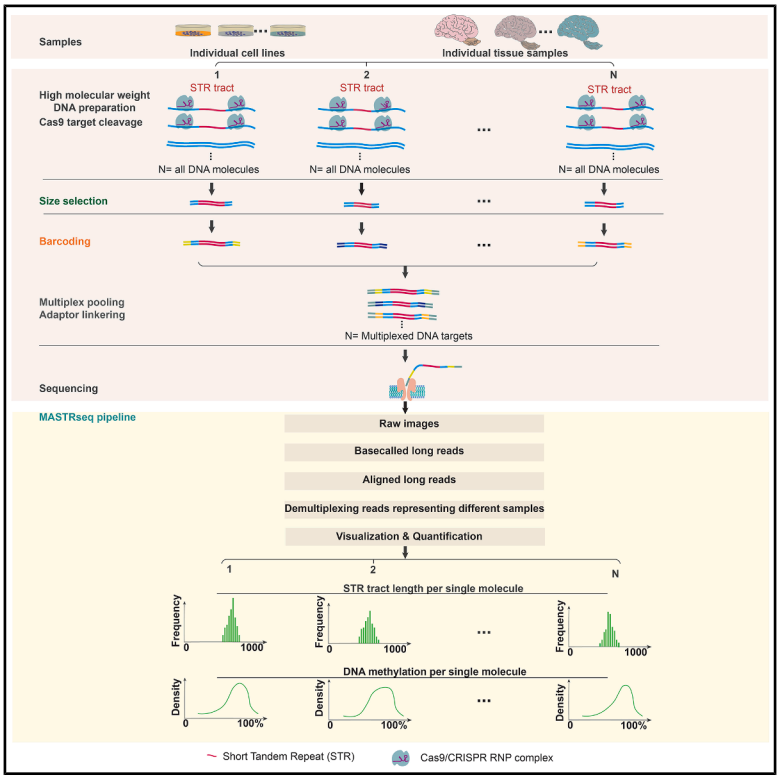
Cell Reports
Methods

Article

MASTR-seq enables multiplexed analysis of short tandem repeats with sequencing

Graphical abstract



Highlights

- We develop MASTR-seq for multiplexed analysis of short tandem repeats
- MASTR-seq employs Cas9 target enrichment and PCR-free, multiplexed nanopore sequencing
- MASTR-seq measures STR tract length and DNA methylation on the same single allele
- MASTR-seq allows for pooling of up to 8–12 samples in one Nanopore MinION flow cell

Authors

Chuanbin Su, Han-Seul Ryu, Keerthivasan Raanin Chandradoss, ..., Esteban O. Mazzoni, Kristen J. Brennand, Jennifer E. Phillips-Cremins

Correspondence

jennifer.cremins@wustl.edu

In brief

Su et al. present MASTR-seq, a cost-effective, high-throughput method for precisely measuring short tandem repeat tract length and DNA methylation at a pre-defined locus at single-molecule resolution using nanopore sequencing.

Su et al., 2026, Cell Reports Methods 6, 101341
March 23, 2026
https://doi.org/10.1016/j.crmeth.2026.101341

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

MASTR-seq enables multiplexed analysis of short tandem repeats with sequencing

Chuanbin Su,[1,2,3,4,5] Han-Seul Ryu,[1,2,3,4,5] Keerthivasan Raanin Chandradoss,[1,2,3,4,5] Thomas Malachowski,[1,2,3,4,5] Ravi Boya,[1,2,3,4,5] Linda Zhou,[1,2,3] Hoa Emma Nguyen,[6] Esteban O. Mazzoni,[6,7] Kristen J. Brennand,[8,9,10,11] and Jennifer E. Phillips-Cremins[1,2,3,4,5,12,*]

[1]Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
[5]Department of Neuroscience, Washington University School of Medicine, St. Louis, MO, USA
[6]Department of Biology, New York University, New York, NY 10012, USA
[7]Neuroscience Institute, Department of Neuroscience and Physiology, New York University School of Medicine, New York, NY 10012, USA
[8]Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, NewYork, NY 10029, USA
[9]Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[10]Friedman Brain Institute, Black Family Stem Cell Institute, Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[11]Department of Psychiatry, Yale School of Medicine, New Haven, CT 6520, USA
[12]Lead contact
*Correspondence: jennifer.cremins@wustl.edu
https://doi.org/10.1016/j.crmeth.2026.101341

MOTIVATION Short tandem repeat (STR) tracts and DNA methylation are associated with clinical phenotypes and can exhibit cell-type-specific mosaicism in several repeat expansion disorders. However, accurate genotyping of intractable repetitive sequences has thus far been limited by several technical challenges. Here, we present MASTR-seq, a Cas9-mediated, size-selection-based, PCR-free long-read sequencing method for cost-effective, high-throughput, and accurate measurement of STR genotype and DNA methylation on a single allele.

SUMMARY

More than 60 human disorders are caused by unstable expansion of short tandem repeat (STR) tracts. These can exhibit cell-type-specific mosaicism in several repeat expansion disorders and remain difficult to characterize due to technical challenges intrinsic to highly repetitive sequences. Long-read approaches can measure STR length and DNA methylation on the same single molecule but are low-throughput and cost-prohibitive across multiple experimental conditions or patient samples. Here, we present MASTR-seq, multiplexed analysis of short tandem repeats with sequencing, for cost-effective, high-throughput, accurate measurement of STR genotype and DNA methylation at single-allele resolution. MASTR-seq couples long-read sequencing, Cas9-mediated target enrichment, size selection, and PCR-free multiplexed barcoding to increase on-target read proportion for 8–12 pooled samples in a single MinION flow cell. MASTR-seq quantifies tract length and DNA methylation status for CGG, GGGGCC (G4C2), and CAG STR tracts in normal-length and mutation-length samples.

INTRODUCTION

More than 60 human diseases have been linked to the mutation-length (ML) expansion of short tandem repeat (STR) tracts.[1–6] Accurate genotyping of intractable repetitive sequences has thus far been limited by several technical challenges.[7,8] First, tandem repeat DNA sequences slip during PCR amplification, resulting in technical biases in quantifying amplicon length. Second, short-read sequencing fails to span ML STRs in many cases, thus prohibiting direct measurement of the DNA sequence and underestimating the degree of mosaicism of STR expansion in disease. Third, challenges arise in computational methods to map short reads over repeat regions, and they are often removed for downstream analysis.[7,9,10] Thus, there is a need for technology advances that allow for accurate and cost-effective approaches for

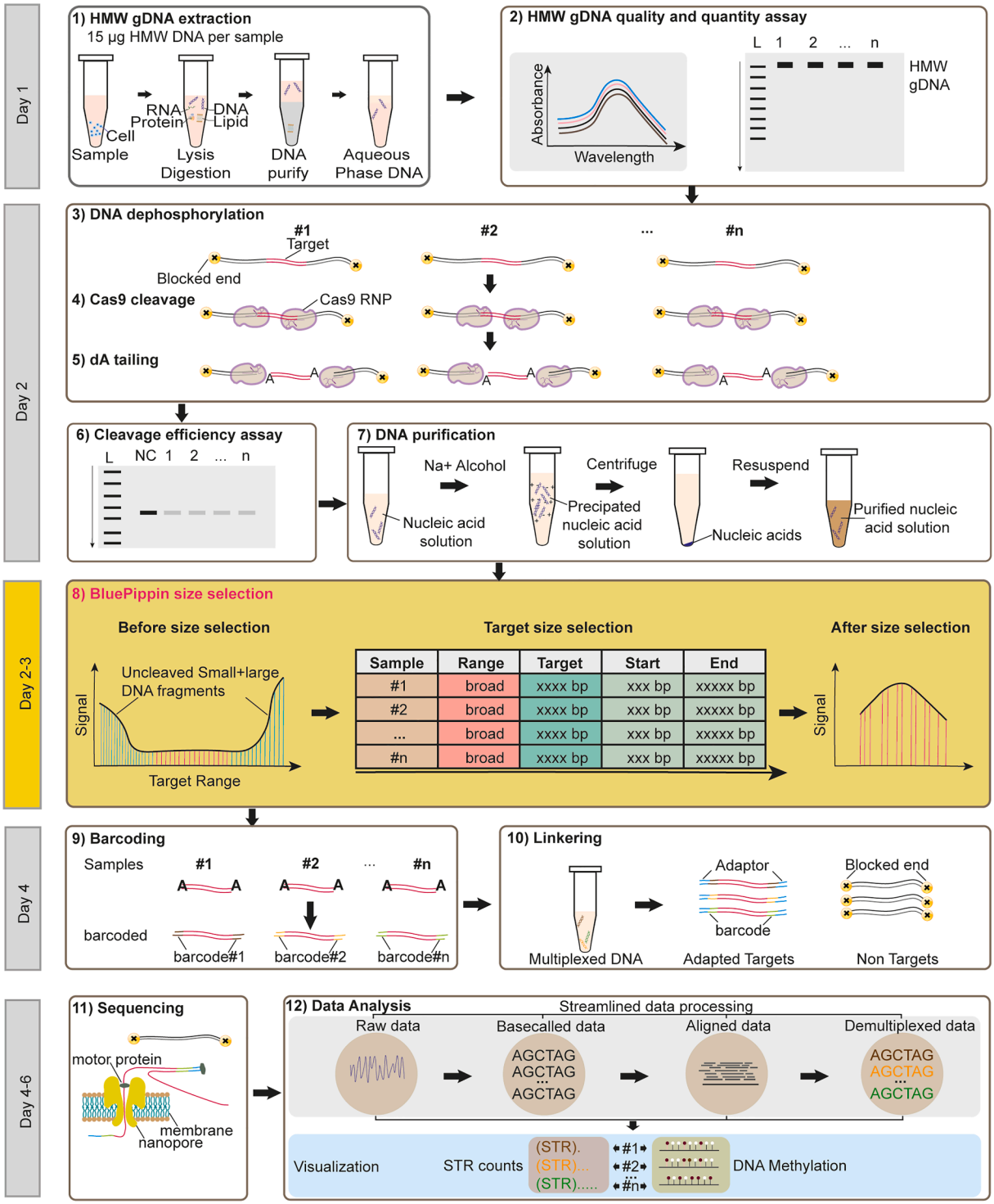


Figure 1. Schematic of MASTR-seq procedure

Day 1: (Step 1) Samples are dissociated to single-cell suspension, and cells are lysed. RNA and protein are digested by RNase A and proteinase K, respectively. DNA is subsequently purified by commercial HMW gDNA extraction kit or phenol:chloroform:isoamyl alcohol (PCIA, 25:24:1)-based method. (Step 2) HMW genomic DNA quality control assays are performed using Nanodrop for DNA quality, Qubit for quantifying DNA concentration, and gel electrophoresis for DNA integrity.

(legend continued on next page)

sequencing STR tracts in patient-derived samples with unknown genotypes.

Long-read sequencing provides a significant technical advance to overcome the challenges prohibiting direct genotyping of large copy-number variations in the repetitive genome. Whole-genome long-read sequencing was used for the accurate assembly of telomere-to-telomere reference genome maps from haploid and diploid cell lines.[11–15] It has also been effectively applied to accurately genotype a range of repetitive sequences, including STR tracts, short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), centromeres, pericentromeric regions, and telomeres.[16–27] However, for studies focused on genotyping a specific locus across a number of conditions, genome-wide long-read sequencing suffers from low sample throughput, low on-target coverage, and the prohibitively high financial burden to achieve the locus-specific read coverage required to accurately measure STR mosaics.[19,28–32] Approaches to capture a predetermined target locus and sequence with long reads have been effectively used to genotype structural variants, STR tracts, and other repetitive elements.[19,21,29,33–42] However, current targeted long-read sequencing methods remain limited by low sample throughput, high background reads, and prohibitive cost.

Recently, multiple strategies to reduce background reads in targeting approaches have been proposed. Molecular approaches include exonuclease digestion, probe-based on-target fragment enrichment, or affinity-based removal of Cas9-bound non-target DNA.[39,40,43–45] Real-time enrichment methods to circumvent a reagent-based molecular targeting step have also been developed, including Read Until,[46–48] UNCALLED,[49] and ReadFish.[50] Undesired DNA molecules are removed from the pore by reversing the current in the channel, thus allowing only targeted DNA molecules to be fully sequenced without additional specialized reagents.[46–50] Such approaches allow for 10–40× on-target read coverage but remain limited by the capacity to multiplex samples. Addressing limitations in multiplexed sample throughput, read accuracy, and low signal-to-noise read specificity would provide a high-throughput, multi-modal, on-target long-read enrichment method to accurately measure STR tract length and DNA methylation across multiple samples at single-allele resolution.

To address the above limitations, methods based on targeted Cas9-mediated capture of DNA sequences coupled with long-read sequencing have been proposed.[19,21,29,33–42,51,52] Early protocols showed proof of concept, yet readily adaptable detailed methods for ease and rigor of reproducibility and adaptability of the method for the customized purposes of individual laboratories are not available. Here, we developed MASTR-seq—multiplexed analysis of short tandem repeats with sequencing—as a Cas9-mediated, size-selection-based, PCR-free long-read sequencing method for multi-modal measurement of STR tract genotype and DNA methylation on the same single allele. MASTR-seq aims to minimize genomic background while preserving the integrity of on-target genomic fragments. Upon Cas9-mediated target enrichment with multiple guide RNAs,[34] we add a fragment size selection step and multiplexed PCR-free barcoding to increase the proportion of accurate on-target reads spanning the full ML STR, thus enabling the pooling and long-read sequencing of 8–12 independent samples in parallel. Our freely available MASTR-seq molecular protocol and code provides access to improvements in sample throughput, on-target read proportion, cost, and adaptability to custom loci that are currently not available in the literature or company kits.

OVERVIEW OF MASTR-SEQ

MASTR-seq is an increased sample throughput, on-target nanopore long-read sequencing method designed for simultaneous measurement of repeat genotype and DNA methylation in the same single allele. MASTR-seq is particularly useful for loci containing intractable ML STRs in complex mosaic-patient-derived cell lines and samples. The protocol involves five major stages (Figure 1):

(Day 1) High-molecular-weight (HMW) genomic DNA (gDNA) preparation; a minimum of 15 μg HMW gDNA is recommended.
(Day 2) Cas9-mediated target DNA cleavage and dA tailing
(Day 2–3) Target DNA fragment size selection
(Day 4) Library preparation via barcoding and adapter ligation
(Day 4–6) Sequencing and data analysis.

(Day 1) HMW gDNA preparation

1. Sample preparation: cells are harvested using line-specific single-cell dissociation reagents.
2. Cell counting normalization: cell numbers are determined using an automated Countess system or manual counting with a hemocytometer.
3. HMW gDNA extraction: HMW gDNA can be rapidly extracted using the Monarch HMW DNA Extraction Kit (1–2 h). Alternatively, use phenol:chloroform:isoamyl alcohol (PCIA, 25:24:1) extraction followed by ethanol precipitation (5–6 h).

Day 2: crRNA/tracrRNA annealing and Cas9 RNP assembling are completed (not shown in this schematic) before (step 3) HMW DNA dephosphorylation, (step 4) Cas9-mediated RNP cleavage, and (step 5) dA-tailing. (Step 6) The quality of Cas9-mediated RNP cleavage efficiency is assessed by PCR amplification of the target region using flanking primers. (Step 7) Cas9-RNP-cleaved products are purified from salts using sodium acetate-ethanol precipitation.
Days 2–3: (step 8) target DNA fragments are precisely size-selected using BluePippin.
Day 4: (step 9) samples are barcoded in a PCR-free manner and pooled for multiplexing using nanopore native barcodes. (Step 10) Adapters are ligated onto the multiplexed DNA.
Days 4–6: (step 11) DNA is loaded onto the Nanopore flow cell and sequenced for 48 h (Step 12) Raw data are processed by basecalling, aligning and demultiplexing sequentially using Dorado. Custom computational tools are employed for the analyses of single-allele DNA methylation and STR tract length.
See also Table S4.

4. HMW gDNA quality control (QC) assay: DNA quality and concentration are assessed using Nanodrop and Qubit, and DNA integrity is verified using a 0.5% agarose gel.

(Day 2) Cas9-mediated target DNA cleavage and dA tailing

1. CRISPR RNA (crRNA) annealing: equimolar crRNAs and *trans*-activating crRNA (tracrRNA) are pooled and annealed in a 1:1 molar ratio.
2. Cas9 and crRNA complex formation: annealed crRNAs and tracrRNA (1:1) are assembled with Cas9 to form Cas9 ribonucleoprotein (RNP) complexes.
3. HMW genomic DNA dephosphorylation: NEB Quick calf intestinal alkaline phosphatase (CIP) is used to dephosphorylate DNA.
4. Cas9 cleavage and dA-tailing: Cas9 RNP complex cleaves dephosphorylated DNA, followed by Taq-polymerase-mediated dA tailing to the 3′ end of dephosphorylated DNA.
5. Cas9-mediated cleavage quality control check: a PCR reaction is performed to confirm Cas9 RNP cleavage efficiency using target flanking primers.
6. DNA purification: Cas9-RNP-digested products are precipitated using ethanol. Precipitated DNA is washed with 75% ethanol and rehydrated in 10 mM Tris-HCl (pH 8.0).

(Day 2–3) Target DNA fragment size selection

1. BluePippin is utilized for size selection to remove uncleaved or non-target size DNA fragments.

(Day 4) Library preparation via multiplexed barcoding

1. Multiplexed barcoding: samples are multiplexed via PCR-free ligation of nanopore barcodes.
2. Adapter ligation: adapters are ligated to the barcoded samples.

(Day 4–6) Sequencing and data analysis

1. Loading library: library DNA is mixed with sequencing buffers and loading beads.
2. Sequencing: nanopore sequencing is set up on the lab-installed MinKnow platform.
3. Basecalling: Dorado (v.0.7.3) in SUP mode is used for basecalling to convert raw electrical signals to quality sequencing reads in.fastq format.
4. Alignment: minimap2 is used for mapping demultiplexed reads to the reference genome.
5. Demultiplexing: Dorado (v.0.7.3) demux mode is used for demultiplexing reads into unique sample.
6. Visualization: customized computational tools are employed for the visualization of STR genotype and DNA methylation.

Experimental design

crRNA design

We used the CHOPCHOP tool to design distinct crRNAs targeting sequences flanking the target locus (Figure S1). The parameters used were In: GRCh38, Target: *FMR1* or *HTT* genomic coordinates, For: nanopore enrichment, Using: CRISPR/Cas9. To increase on-target coverage, we designed at least two crRNAs upstream and downstream of the targets as reported previously.[34] We refined the selection of candidate crRNAs based on several parameters: cleavage location, (+) strand crRNAs for the region upstream of the locus, (−) strand crRNAs for the region downstream of the locus, 40%–80% GC content, self-complementary = 0, mismatches (MM0 = 0; MM1 = 0; MM2<2), and cleavage efficiency>50. Cleavage efficiency is computed by the CHOPCHOP tool (https://chopchop.cbu.uib.no/instructions) using the option "Doench et al. 2016 – only for NGG PAM," and alternate equations are available.[53,54] For cleavage location, we recommend a minimal 500 bp of flanking sequence on either side of the target STR region. For repetitive regions of low complexity, the flanking sequence is essential for downstream on-target reads alignment and STR counting. Final crRNA DNA sequences were synthesized by Integrated DNA Technologies, and the sequences are provided in Table S1.

Samples. MASTR-seq is illustrated here using data generated in patient-derived human induced pluripotent stem cells (hiPSCs), human embryonic stem cells (hESCs), and patient-derived brain tissues. It is adaptable to diverse cell types or tissues as long as high enough input can be obtained. The culture medium formulation, seeding density, growth and passage conditions, and cell or tissue dissociation methods will be specific to the cell type under investigation.

Cell number. For hiPSC or hESC samples, we recommend starting with 2.5–5 $\times$ 10^6 cells for each library to achieve at least 15 μg of HMW genomic DNA. We used TrypLE reagents to triturate hiPSCs into a single-cell suspension. We counted single iPSCs either using an automated Countess system or manually counting with a hemocytometer. We have also tested human brain tissue from a patient with fragile X syndrome (FXS) and created successful data using a 20 mg tissue sample and ~25 μg HMW gDNA. Future versions might be optimized to allow for improvements in lowering the input of DNA required to make rare samples accessible.

HMW genomic DNA preparation. DNA integrity is essential for efficient yield of Cas9 target cleavage products. DNA extracted using kits often results in fragments ranging from 10 to 100 kb in length,[20] and the phenol-chloroform extraction method gets HMW DNA exceeding 100 kb in length. In applications with 5 $\times$ 10^6 cultured cell numbers or 20 mg clinical sample, we can obtain HMW DNA using commercial kits (NEB #T3060S/3050S) and consistently achieve >20 μg high-quality HMW DNA (Table S1). When using the Monarch kit for brain tissues (NEB #T3060S), we used the extraction parameters as follows: 20 mg of brain tissue yielded approximately 25 μg of HMW gDNA using an agitation speed of 1,500 rpm at 56°C for 15 min, followed by incubation at 56°C without agitation for 30 min. To ensure efficient phase separation, after RNase A digestion, the lysate was chilled on ice for 3 min before adding the protein separation solution.

HMW DNA tends to be viscous, and the rehydration process can lead to fragmentation. After extraction, we recommend eluting HMW genomic DNA by heating at 50°C for 1 h, followed by rotation at room temperature overnight. Genomic DNA purity is assessed by calculating the A260/A280 and A260/A230 ratios

obtained through spectrophotometric analysis with Nanodrop. High-purity genomic DNA should exhibit A260/A280 and A260/A230 ratios within the ranges of 1.8–2.0 and 2.0–2.2, respectively (Figure S2A). If the ratios significantly deviate from these optimal values, superior performance can be achieved by re-purifying gDNA samples using sodium acetate DNA purification. We measure genomic DNA yield as the concentration of rehydrated DNA using the Qubit DNA Broad-Range assay (Figure S2B). We verify the integrity of HMW DNA fragments on a 0.6% 1× TAE agarose gel (Figure S2C). Using iPSC pellets consisting of 2.5–5 × 10^6 cells for each sample, we typically observe a total mass of 15–30 μg of HMW genomic DNA. Alternatively, some Cas9-targeted long-read methods employ genomic kits, such as the Gentra Puregene Cell Kit (Qiagen, 158767) for HMW DNA extraction, which also yields satisfactory results. However, in cases with limited samples, the phenol-chloroform extraction method may be preferred to maximize genomic DNA yield.

Cas9 RNP assembly

The CRISPR-Cas9 system configured as an RNP complex *in vitro* consists of a guide RNA and Cas9 nuclease. The guide RNA is comprised of two components: (1) a 17–20 nucleotide crRNA complementary to the target DNA and (2) a tracrRNA that acts as a binding scaffold for the Cas9 nuclease. To assemble Cas9 RNP complexes, we anneal crRNAs and tracrRNA in a 1:1 molar ratio. When targeting the *FMR1* gene for 10 reactions, we combine four different crRNAs, each pooled with tracrRNA (2.5 μM for each crRNA and 10 μM for tracrRNA in a 10-μL reaction volume). Next, we assemble CRISPR-Cas9 RNPs in 1× NEB rCutSmart buffer. We determine the optimal ratio of Cas9 protein to CRISPR gRNA based on *in vitro* Cas9 RNPs cleavage efficiency. We employ a 1:1 molar ratio of Cas9 protein to CRISPR gRNA to achieve efficient *FMR1* target cleavage (Figure S3A).

HMW genomic DNA dephosphorylation

Genomic DNA dephosphorylation is the initial step to reduce genomic background in MASTR-seq (Figure 1, step 3). We incubate 5 μg of genomic DNA with 3 μL Quick CIP for DNA dephosphorylation in NEB rCutSmart Buffer at 37°C for 20 min followed by inactivation at 80°C for 2 min. When optimizing for increased on-target coverage, it is recommended to divide genomic DNA into multiple reactions in the dephosphorylation step, each containing 5 μg (or less) per reaction.

Cas9 RNP cleavage and dA tailing

We incubated Cas9 RNPs with a maximum of 5 μg of dephosphorylated HMW genomic DNA at 37°C for 60 min followed by a dA-tailing of blunt ends at 72°C for 5 min. To ensure sufficient on-target coverage per sample, three technical replicates (5 μg DNA per replicate) of Cas9 target cleavage and dA tailing are performed. The three replicates per sample are then pooled together for BluePippin size selection, and the size-selected sample was used for barcoding, adaptor ligation, and sequencing. For efficient on-target cleavage, we recommend an optimization of Cas9 RNP cleavage time to increase the ratio of on-target/off-target reads. Excessive cleavage time can lead to a higher number of off-target reads. To assess Cas9 RNP cleavage, we recommend designing primers flanking the Cas9 RNP cleavage locus and performing PCR to amplify the CRISPR-targeted locus (Figure S3B). Successful Cas9 cleavage will result in little to no amplification as the target DNA is fragmented (Figure S3C). Conversely, if cleavage does not occur, PCR will produce the expected amplification product (Figure S3C).

DNA purification

We employed the sodium acetate-ethanol DNA precipitation method to purify and concentrate DNA from high-salt aqueous solutions. The precipitated DNA is rehydrated at 50°C for 1 h in 30 μL elution buffer. If DNA is not completely dissolved, an overnight rotation at room temperature is recommended.

BluePippin automated size selection

Removal of background uncleaved genomic DNA or off-target DNA fragments outside of the target locus is essential for MASTR-seq performance. We use BluePippin size selection to further enrich the target signal in our protocol. Cas9 RNP cleavage of the target (e.g., here, the *FMR1* locus) produces a 6.9-kb-sized DNA fragment in normal-length (NL) iPSCs or an 8- to 10-kb-sized DNA fragment in ML iPSCs. For potentially longer ML STR tracts (e.g., GGGGCC/G4C2 in ALS-patient-derived iPSCs), Cas9 RNP cleavage of the target *C9ORF72* locus produces a 4.2 kb size DNA fragment in NL alleles. Mutation-length alleles can be anywhere from 4.38 kb (30 repeats) to greater than 10 kb for 1,000 repeat units. To ensure the BluePippin size selection recovers the wide range of DNA fragments efficiently, we conduct size selection for large 4–15 kb fragments using the BluePippin (Sage Science) and BLF7510 cassette using the "0.75DF 3–10 kb Marker S1-Improved Recovery" cassette definition with broad range mode set at 4–15 kb (Table S2). For genotyping potentially longer ML STRs (e.g., those in genes *CNBP*, *C9ORF72*, *SAMD12*, and *ATXN10*), we recommend testing different size range modes before finalizing the size selection. To allow for size-selected HMW DNA to efficiently detach from the membrane of elution mode well, we recommend allowing samples to remain in the cassette for at least 45 min after the run is complete prior to removing eluent. Additionally increased yield can be achieved by second elution with 0.1% Tween 20 solution after removal of initial elution.

PCR-free multiplexed barcoding

To prepare the library for nanopore sequencing, we ligate size-selected DNA fragments to a unique barcode for each sample, pool multiple barcoded samples together, and ligate to an adapter for sequencing. To circumvent amplification bias over repetitive sequences, our methods are PCR free. The limitation is that the low yield of target DNA fragments makes equimolar pooling of each barcoded sample challenging. Using MASTR-seq, we can barcode each sample and pool 8–12 human-derived samples for one MinION flow cell. Library quality can be assessed by using the Agilent bioanalyzer or Tape station system (Figure S4). The MinION flow cell is limited by 512 nanopore channels for sequencing, and the PromethION flow cell contains up to 2,675 nanopore channels. For applications using the PromethION flow cell, MASTR-seq can be readily adapted to increase sample throughput and the number of target loci. In principle, one PromethION flow cell can accommodate over 50 samples.

Sequencing

To enhance the yield of on-target reads and reduce off-target reads during sequencing, we employed specific parameters such as a minimal read length of 1,000 bp, a maximum scanning

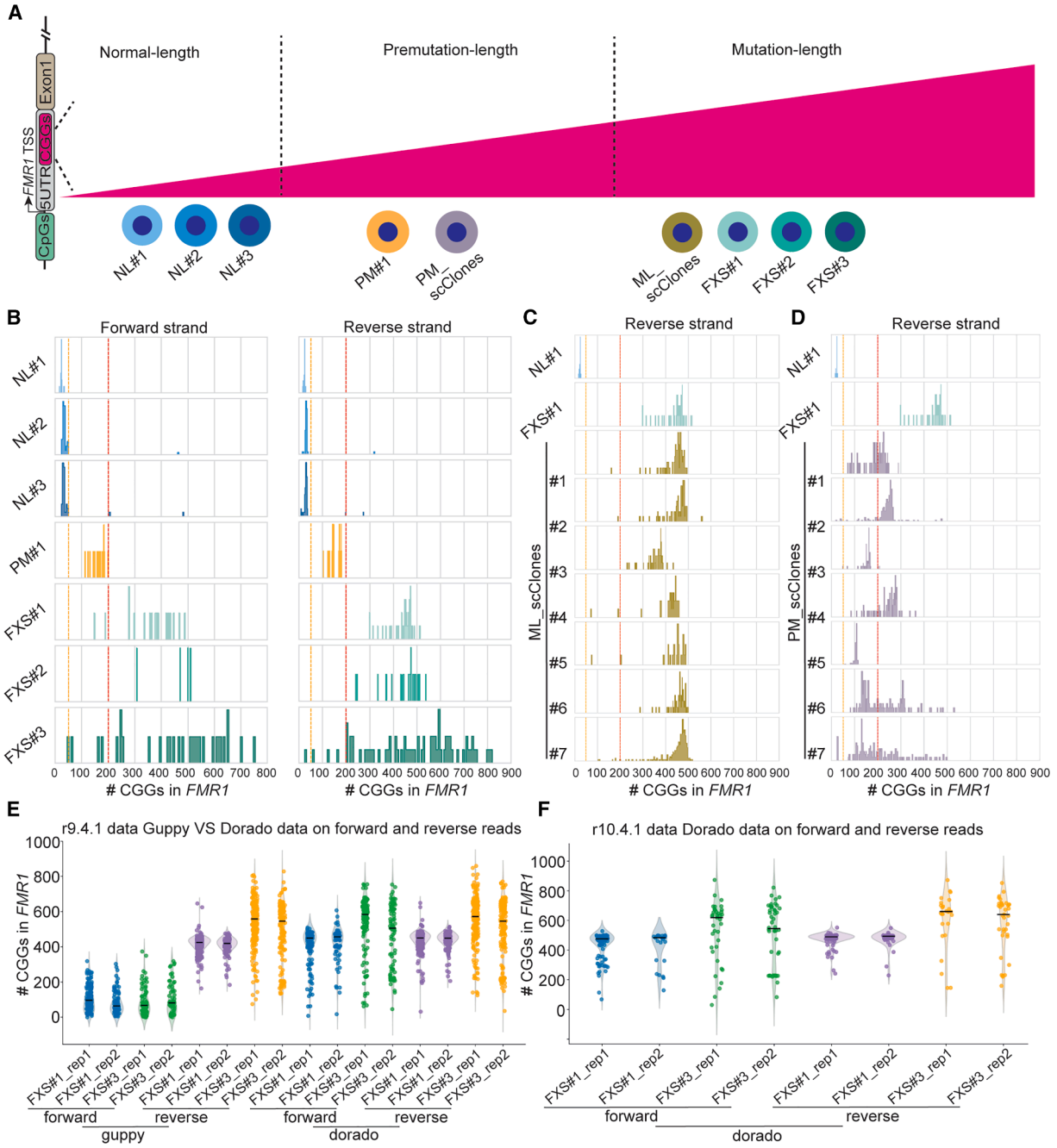


Figure 2. Precise quantification of *FMR1* CGG STR lengths in human iPSCs derived from patients with FXS using MASTR-seq

(A) Schematic representation of iPSC lines modeling normal-length (NL), premutation-length (PM), and mutation-length (ML/FXS) *FMR1* CGG STR tracts. Normal-length: 5–44 CGG repeats; intermediate-length: 45–54 CGG repeats; premutation: 55–200 repeats; full ML: >200 repeats.

(B–D) Dorado- and MASTR-seq-pipeline-processed R9.4.1 sequencing data. For the CGG size reference, orange vertical line indicates 50, and red vertical line indicates 200.

(B) Distribution of CGG STR tract length in *FMR1* 5′ UTR derived from forward and reverse nanopore long reads, shown per sample, represented as vertical lines.

(C) Number of CGG STRs in the 5′UTR of *FMR1* computed from reverse single-molecule nanopore long reads across CRISPR-engineered FXS ML single-cell iPSC clones.

(D) Number of CGG STRs in the 5′UTR of *FMR1* computed from reverse single-molecule nanopore long reads across CRISPR engineered FXS PM single-cell iPSC clones. These data are provided side by side with the ML (FXS#1) and NL#1 lines.

(legend continued on next page)

time of 1.5 h, and "pores reserve" mode (Table S2). For sequencing, we have used both the Flow Cell R9.4.1 (outdated) and R10.4.1 (FLO-MIN114) on the MinION sequencer for a duration of about 48 h.

Data processing

1. Raw data processing

Data processing steps can improve long-read basecalling and read alignment as technology platforms improve. Several studies report that higher accuracy can be achieved by using the latest Nanopore R10.4 flow cells[55–57] with improved resolution for homopolymer detection.[58,59] Simplex sequencing achieves an average read quality of 98%, and duplex reads exhibit 99% accuracy with the R10.4 flow cell.[56] After collecting data from the R10.4 flow cell, the newest basecalling algorithms can be used together with the Q20+ reagent from Nanopore library kits to achieve up to 99% base calling accuracy.[58,59]

We developed computational scripts and provided them to the user to conduct basecalling and alignment using Dorado (v.0.7.3). Using NanoStat, we generated nanopore data statistics from the bam file before demultiplexing (Table S3) and assessed data yield across sequencing time (Figures S5A and 5B). The sequencing yield accumulates much more slowly later in the run, indicating the optimal time to stop the sequencing (Figures S5A and S5B). Subsequently, we used Dorado demux (v.0.7.3) to demultiplex basecalled reads into unique samples and analyzed the distribution of reads quality and reads length across each demultiplexed sample (Figure S5C). The Phred quality scores for most reads are >30 (99.9% accuracy) across 12 samples in one nanopore run, which is consistent with data produced from flow cell R10.4.1.[58,59]

2. STR visualization

Strand-specific errors in nanopore basecalling are common in repetitive DNA sequences.[60] When assessing STR tract length, we compared the forward with reverse strands. In our previous publication using MASTR-seq data we generated with older R9 chemistry in FXS patient-derived iPSCs, we found dramatic differences in CGG STR lengths in forward and reverse reads.[61] However, here using the same published data and re-analyzing it with our upgraded dorado pipeline, we found that both forward and reverse reads provided similar STR lengths (see Figure 2). Moreover, when repeating the data collection with R10 chemistry, we found that both forward and reverse reads accurately quantify ML STR tracts (Figures 2 and 5). In this manuscript, we define *FMR1* CGG STR tract length ranges based on ACMG/EMQN clinical guidelines: NL (<45 repeats), intermediate (45–54), premutation-length (PM) (55–200), and full ML (>200). Unless otherwise stated, reads with >54 CGG repeats were used as expanded PM or ML CGG STR alleles.[62]

Recommended controls in the procedures

Cas9 RNP cleavage can be assessed as a quality control step. We recommend preparing negative control samples without Cas9 RNP cleavage to compare to cleaved samples as described in the day 2 protocol. To qualitatively evaluate if the Cas9 step was effective, we recommend *in vitro* Cas9 cleavage on a PCR amplicon or a synthesized DNA fragment containing the single-guide RNA (sgRNA) target site followed by agarose gel electrophoresis.[54] For a positive control, we recommend using cell lines with defined STR tract lengths to calibrate the method's parameters and ensure MASTR-seq is accurately measuring ground-state truth before applying it to query unknown cell lines. An evaluation of sequencing metrics can assist in assessing the quality of the sample, including the amount of loaded library, library yield, sequencing duration, total generated bases and reads, read length distribution, the proportion of unclassified and assigned reads before and after demultiplexing, and the number of reads fully spanning the STR (Table S3).

RESULTS

To assess the suitability of MASTR-seq for accurate STR length quantification, we quantified the *FMR1* CGG STR across a spectrum of tract lengths including NL, PM, and full ML hiPSCs (Figure 2A). Implementation of MASTR-seq for STR genotyping yielded an increase in on-target read proportion compared to existing publicly available published data on target capture multiplexed samples (Table S4). Using the latest Dorado method for basecalling Nanopore data, we demonstrate that forward and reverse strand reads have similar performance for genotyping the CGG STR tract length (Figures 2B, 2E, and 2F). We confirmed that normal- and PM iPSC alleles contained 23–33 and 155–194 CGG triplets, respectively, whereas all three independent iPSC lines derived from FXS patients showed averaged size of 417–478 CGG triplets (Figure 2B). Furthermore, we employed MASTR-seq to verify the length of the *FMR1* CGG STR tract in CRISPR-engineered PM clones by cutting back the ML CGG STR from a parent FXS iPSC line (Figures 2C and 2D). We compared the performance of genotyping the CGG STR using Guppy and Dorado basecallers. Forward-strand reads basecalled using Guppy showed higher error rates and poor accuracy compared to Dorado (Figure 2E). These differences reflect how recent improvements in Nanopore's wet-lab chemistry (R10.4.1 flow cell, V14 kit) have coincided with advances in computational pipelines (transition from Guppy to Dorado), making Dorado the current standard for high-accuracy STR profiling. We also demonstrate the compatibility of MASTR-seq with the R10.4.1 MinION flow cell and V14 chemistry. Using the latest chemistry and flow cell combined with Dorado basecalling resulted in significantly higher read accuracy compared to R9.4.1 flow-cell-generated data (Table S3). Both forward and reverse

(E) Comparison of Guppy- and Dorado-processed R9.4.1 sequencing data for quantification of *FMR1* CGG STR, computed from individual forward and reverse single-molecule nanopore long reads across two different FXS ML iPSC lines, each FXS iPSC line with two biological replicates (rep#1 and rep#2).
(F) Comparison of Dorado-processed R10.4.1 sequencing data for quantification of *FMR1* CGG STR, computed from individual forward and reverse single-molecule nanopore long reads across two different FXS ML iPSC lines, with each cell line having two biological replicates as in (E).
See also Tables S4 and S5.

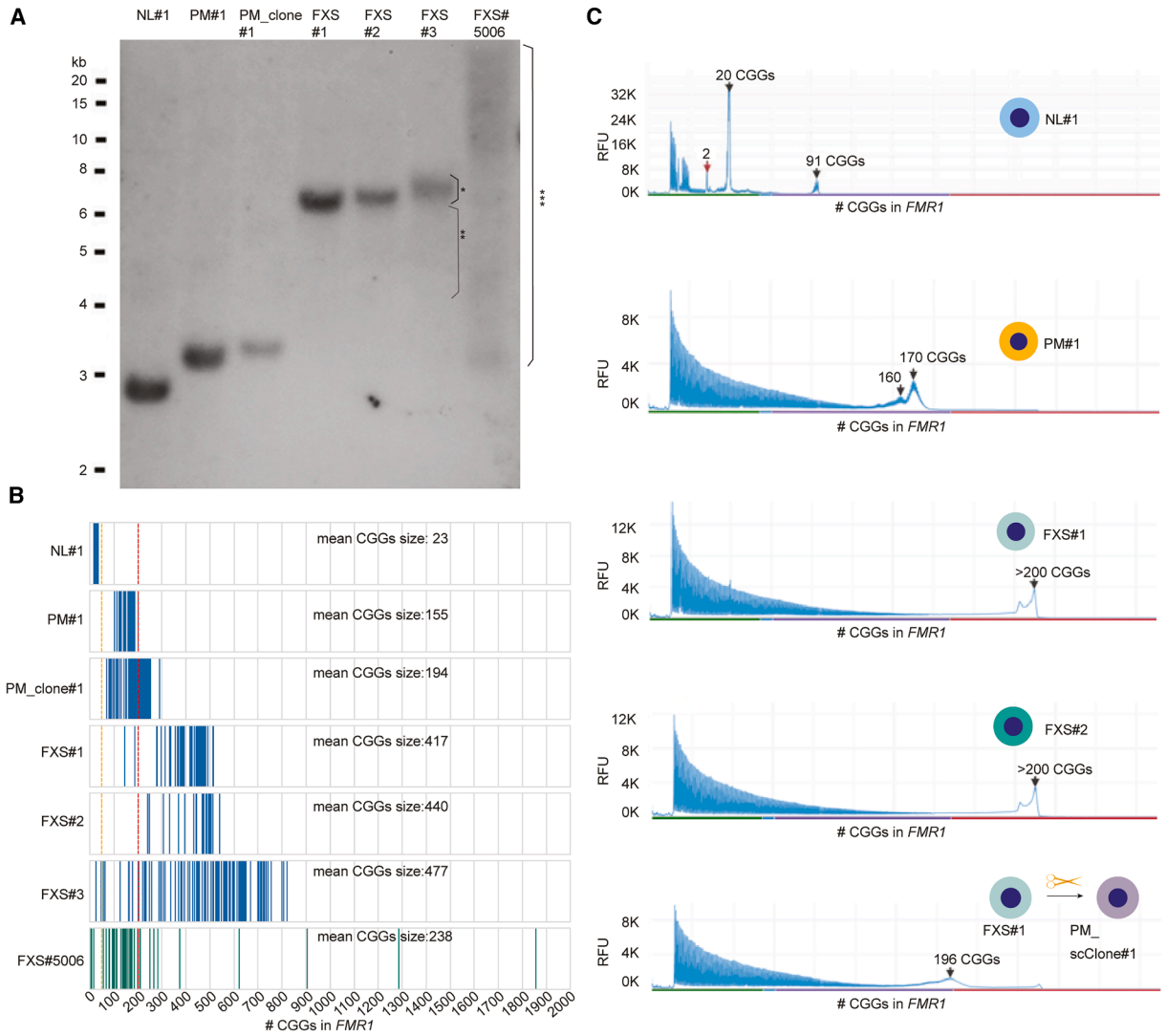


Figure 3. Validation of CGG STR genotyping in human FXS-patient-derived samples

(A) Southern blot analysis of six iPSC lines (1 WT, 2 PM, and 3 FXS) and one FXS patient brain tissue sample (FXS#5006), all males. Genomic DNA was digested with *EcoRI* and *BssHII* and hybridized with the digoxigenin-labeled GLFXDig1 probe. In males, the expected fragment size for normal alleles is ~2.7 kb. Premutation alleles appear as fragments of 2.9–3.3 kb, whereas full mutations with methylation produce fragments >5.7 kb. Mosaicism is indicated by the coexistence of premutation (2.9–3.3 kb, unmethylated) and full mutation (>5.7 kb, methylated) fragments. The major and smear bands in the FXS#3 line are marked with * and **, respectively; the smear in FXS#5006, indicative of a wide distribution of CGG repeat sizes, is marked with ***.

(B) MASTR-seq genotyped the CGG repeat length distributions using both forward and reverse reads for the same samples as (A); for the CGG size reference, orange vertical line indicates 50, and red vertical line indicates 200.

(C) Capillary gel electrophoresis traces from the AmplideX mPCR FMR1 analyses quantifying CGG repeat sizes across normal, pre-mutation, and full-mutation ranges in cell-line-derived samples.

See also Table S5.

reads can be used for more accurate STR genotyping (Figure 2F). Thus, MASTR-seq efficiently measures the CGG tract length of NL, PM, ML, and CRISPR-engineered PM cutback iPSC clones.

To independently assess the extent to which MASTR-seq accurately captures STR allele heterogeneity and size, we performed Southern blotting on six iPSC lines and one FXS-patient-derived brain sample. WT, PM, FXS#1, and FXS#2 iPSC lines showed a single band at a size consistent with the MASTR-seq-predicted STR tract length (Figures 3A and 3B). By contrast, the FXS#3 iPSC line and caudate nucleus brain tissue (FXS#5006) exhibited a broad band or smear pattern in Southern blot data, indicative of a heterogeneous CGG STR length distribution caused by somatic instability (Figures 3A and 3B). The Southern blot patterns closely matched the tract length distributions detected by MASTR-seq, suggesting that allele-specific heterogeneity reflects bona fide biological variation rather than technical noise exclusive to our method. We

further performed AmplideX mPCR on several FXS-patient-derived iPSC lines with less heterogeneity to independently quantify CGG repeat sizes across normal-, PM-, and ML tract lengths (Figure 3C). AmplideX mPCR results were consistent with Southern blot and MASTR-seq data in normal- and PM samples (Figure 3C). Moreover, consistent with known limitations, AmplideX mPCR failed to resolve the size of the ML CGG STR tract, whereas MASTR-seq uniquely quantified the alleles (Figure 3C). Together, Southern blot and AmplideX mPCR cross-validate the veracity of MASTR-seq results across normal- to ML tract lengths.

We demonstrate that nanopore long-read sequencing can accurately quantify STR length and DNA methylation at single-molecule resolution. Our method accurately confirms the known hypomethylation of CpG dinucleotides in the *FMR1* promoter in NL and PM iPSCs, as well as promoter hypermethylation in ML iPSCs (Figure 4A). Moreover, at the CGG STR tract, we observed nearly all hypomethylated alleles in all three normal length and the PM iPSC lines as well as a high level of hypermethylation in ML iPSCs (Figure 4B). Consistent with the DNA methylation data, our published *FMR1* gene expression data revealed high expression in NL and PM iPSCs and silencing in ML iPSCs (Figure 4C). In CRISPR-engineered clones, we demonstrate that five out of seven PM cutback clones exhibit depleted CpG methylation at the *FMR1* promoter and decreased CpG methylation in the CGG STR tract (Figures 4D and 4E). *FMR1* de-repression occurs in the same clones where DNA methylation is depleted compared to the parent FXS-patient-derived ML iPSC line (Figure 4F). We also demonstrated the compatibility of MASTR-seq with the R10.4.1 MinION flow cell and V14 chemistry for efficiently detecting DNA methylation on *FMR1* target regions (Figures 4G and 4H). Using the latest chemistry and R10.4.1 flow cell, together with Dorado basecalling, we observed reproducible methylated CpG island at the *FMR1* promoter region (Figures 4G and 4H) and the CGG STR tract (Figure 4I) in FXS#1 and FXS#3 cell lines (Figures 4A–4E). Together, our data demonstrate robust quantification of both DNA methylation and STR length in the same single molecule across a cohort of NL, PM, and ML FXS iPSC lines.

To validate MASTR-seq's utility beyond the CGG STR, we also employed our method to verify the length of a *HTT* CAG tract in three isogenic human embryonic stem cells (hESCs) engineered to model Huntington's disease (HD)[63] (Figure 5A). The RUES2 hESC lines have been engineered on one allele with specific, fully validated CAG tract lengths as indicated by the labels (Figure 5A). MASTR-seq accurately confirmed the presence of the unedited allele as well as the NL- or ML-edited CAG STR tract in the *HTT* gene in all isogenic hESCs (Figure 5B). Employing MASTR-seq, we accurately identified large G4C2 repeats in *C9ORF72* gene in ALS-derived hiPSCs (Figure 5C). In addition, we characterized the increased DNA methylation level at the promoter and G4C2 repeats region in ML alleles compared to NL alleles in the same sample (Figures 5D and 5E).

By adding a fragment size selection step to enrich target DNA fraction while removing large DNA fragments, the size selection effectively reduces the size of the genome from which the target DNA fragments are isolated. As a result, we achieve improved on-target read proportion in a single nanopore run (Table S4). Moreover, we employ PCR-free multiplexed barcoding to allow for the pooling of up to 12 samples in a single MinION flow cell. Altogether, MASTR-seq provides an approach for multiplexed PCR-free barcoding for high-throughput pooling at a defined target locus in multiple samples.

DISCUSSION

Advantages and applications

In our recently published work focused on FXS,[61] we applied MASTR-seq to accurately measure CGG STR length and DNA methylation for the 5′UTR and promoter of the *FMR1* gene in the same single-allele across several NL, PM, and ML patient-derived iPSC lines. By adding multiple steps to improve the proportion of long reads spanning the target locus compared to existing target capture long-read methods, we demonstrate here that MASTR-seq achieves substantial on-target read proportion at the *FMR1* CGG STR in a pool of 8–12 samples using a single MinION flow cell (Figures 1, 2, 3, 4, and 5; Table S4).[35] We show robust measurement of the CAG STR in the *HTT* gene in human embryonic stem cell (hESC) lines after knockin of defined ML CAG tracts (Figures 5A and 5B). Moreover, we demonstrate accurate measurement and deep coverage of the longer and more complex 100% CG-content GGGGCC (G4C2) tract observed in the *C9ORF72* gene in familial amyotrophic lateral sclerosis (ALS) (Figure 5C). Given that strand-specific GC content and secondary structures can affect repeat length estimation as shown in published work,[61] we have systematically evaluated forward vs. reverse strand coverage across different STR loci and observed that both can similarly estimate STRs size when using MinION flow cells (R10.4.1) together with the basecaller Dorado instead of Guppy (Figures 2 and 4). Overall, we demonstrate that MASTR-seq is effective at high-throughput measurement of varying sequence compositions and sizes across a wide range of biological contexts.

We find that MASTR-seq markedly increases on-target read mapping percentage compared to the main published method integrating nanopore sequencing, Cas9 enrichment of a single target locus, and pooling of multiple samples (Table S4).[35] Cas9 capture has also been combined with nanopore sequencing to enrich thousands of partially redundant mobile element loci across the genome and multiple samples[21] or to multiplex multiple sgRNAs.[34,64] Both designs pull down many loci simultaneously, making their read distribution and on-target proportion fundamentally different from the single-locus strategy employed in this study. Moreover, several methods coupling Nanopore or PacBio long-read sequencing with target capture for a single sample have been reported but are limited in throughput (Table S4).[29,33,34,36,38–42,51,52] Beyond published work, two commercial kits are available but have substantial limitations. In the PacBio commercial kit PureTarget, the company predetermines known STR expansion loci and allow partial customization of target panel. To our knowledge, one dataset composed of samples with expanded STRs alleles was produced and publicly available using PureTarget (repeat expansion panel Coriell samples: https://downloads.pacbcloud.com/public/dataset/PureTargetRE/Coriell/); however, the raw data were not available for comparative analyses. A kit from Nanopore also exists (https://nanoporetech.

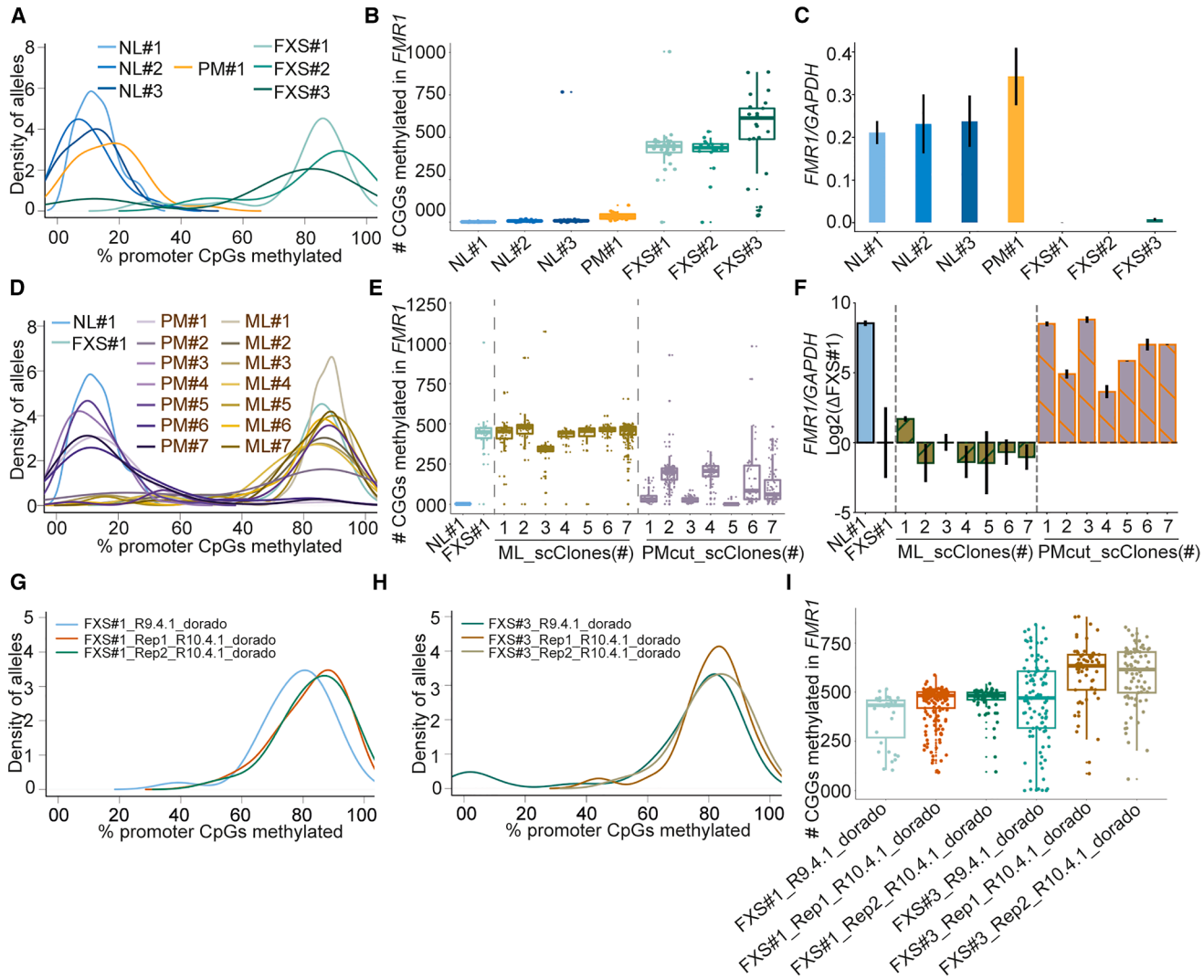


Figure 4. Precise quantification of DNA methylation and STR tract length in the same single molecule in iPSCs derived from FXS patients using MASTR-seq

(A) CpG dinucleotide DNA methylation of the *FMR1* promoter across normal-length (NL), premutation-length (PM), and full mutation-length (FXS/ML) iPSC lines. Density plot of methylated CpG islands in the *FMR1* promoter from reverse nanopore long reads computed using Nanopolish.

(B) Proportion of alleles with a methylated CGG STR tract using reverse nanopore long reads computed using STRique across NL, PM, and ML iPSC lines. Each dot denotes a single-molecule long read representing one allele.

(C) RT-qPCR measurement of *FMR1* mRNA levels normalized to *GAPDH* mRNA levels. Horizontal bar indicates the mean value from two biological replicates. Error bars represent the standard deviation between two biological replicates.

(D and E) CpG dinucleotide DNA methylation of the (D) *FMR1* promoter and (E) CGG STR tract computed from individual reverse single-molecule nanopore long reads across seven single-cell iPSC clones derived from an FXS ML line and seven PM CRISPR cutback clones also derived from the same parent FXS ML iPSC line. These data are provided side by side with the population-based ML (FXS#1) and NL (NL#1) iPSC lines.

(F) RT-qPCR measurement of *FMR1* mRNA levels normalized to *GAPDH* mRNA levels for each single-cell clone and further normalized to the values for parent line FXS#1. Horizontal bar indicates the mean value from two biological replicates. Error bars represent the standard deviation between two biological replicates.

(G and H) Density plot of methylated CpG islands within *FMR1* promoter from two biological replicates of FXS#1 and FXS#3 iPSC lines. CpG methylation from nanopore long reads was computed using Dorado.

(I) Density plot of methylated CpG islands within *FMR1* CGG STR region from two biological replicates of FXS#1 and FXS#3 iPSC lines. CpG methylation in CGG STR from nanopore long reads was computed using Dorado.

See also Table S5.

comdocument/cas-sequencing-kit); however, it lacks sample throughput and has been deprecated. Here, we provide detailed protocol to allow for reproducible experimental and computational analysis of MASTR-seq across a range of custom model systems and loci for paradigms of sample multiplexing and target capture of intractable repetitive sequences.

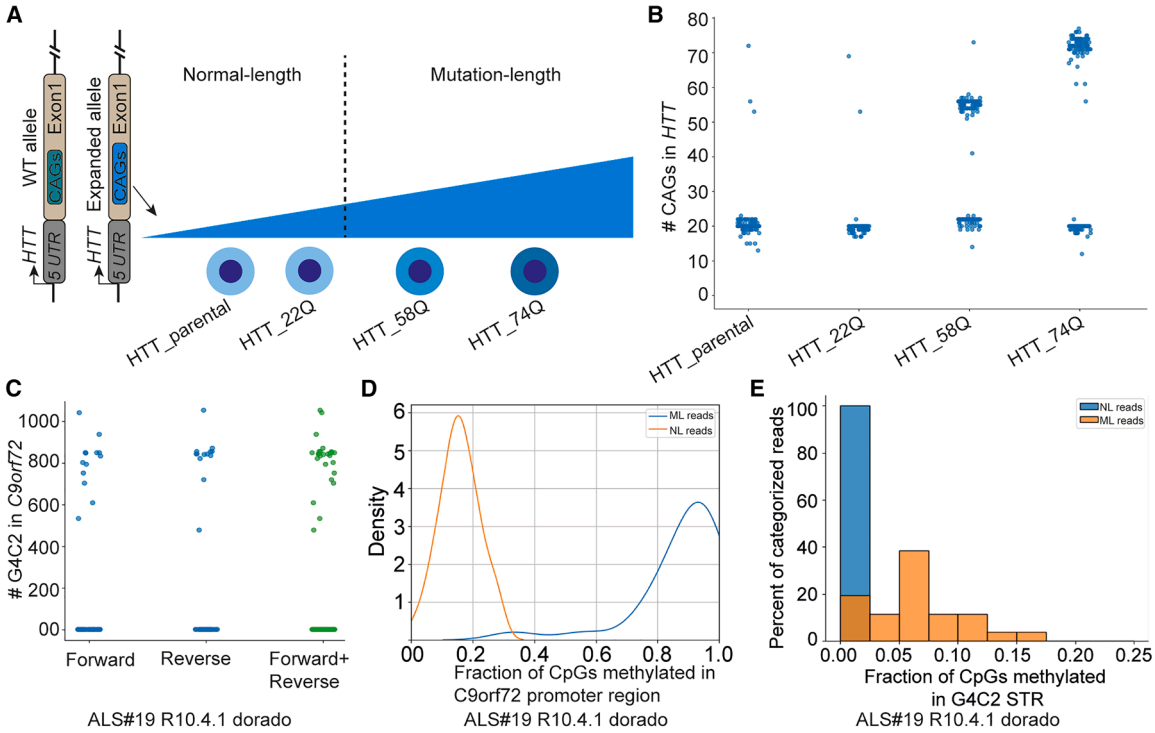


Figure 5. Precise quantification of the HTT CAG STR and C9ORF72 CCCCGG STR tracts using MASTR-seq

(A) Schematic representation of isogenic RUES2 human embryonic stem cell (hESC) lines engineered to model *HTT* CAG expansion in Huntington disease (HD). Shown here are RUES2 subclones comprising one parental line, one HTT CAG normal-length (NL) knockin line, and two *HTT* CAG mutation-length (ML) knockin lines.

(B) Distribution of CAG STR lengths in *HTT* obtained from targeted nanopore long reads.

(C) Distribution of G4C2 STR lengths in *C9orf72* obtained from targeted nanopore long reads in a C9orf72 ALS-patient-derived induced pluripotent stem cell (iPSC). Each dot signifies a single-molecule long read corresponding to one allele.

(D) Density plot of reads according to CpG dinucleotide DNA methylation within the C9orf72 promoter across NL and full ML alleles.

(E) Histogram of CpG dinucleotide DNA methylation within the G4C2 STR across NL and full ML alleles. CpG methylation was computed using Dorado-modified basecalling and custom scripts.

See also Table S5.

Our work meets an unmet need of a detailed experimental and computational protocol to conduct custom CRISPR-based target capture of the repetitive genome for multiplexed, cost-effective, multi-modal analysis of repeat genotype and DNA methylation at single-allele resolution. The ability to count and measure STR tract lengths at single-allele resolution represents a major step change improvement in sensitivity and specificity from traditional gold-standard STR genotyping methods used in the clinic such as Southern blot and triplet repeat-primed PCR (TP-PCR) (AmplideX, Asuragen). Existing target capture methods have sparse methods, poor on-target read performance, and poor multiplexing capabilities or have been commercialized into kits that lack flexibility to scale to new experimental or computational advances and the custom applications from individual laboratories. We demonstrate here that MASTR-seq protocol can be easily adapted using custom guide RNAs for a wide range of applications for multiplexed, multi-modal analysis of repeat genotype and DNA methylation at single-allele resolution. We provide a detailed computational and experimental protocol, and we upgrade MASTR-seq to improve sample analysis, read accuracy across disparate repeat sequences, and improve the proportion of on-target reads mapped to reduce cost and efficiency of sample processing.

Broader impact via future applications

MASTR-seq future applications include the following.

1. multi-modal analysis of intractable repetitive sequences for genotype and DNA methylation

Long-read sequencing has been successfully applied for the detection of copy-number variants and DNA methylation of repetitive sequences, including STRs, transposable elements,[19,21,29,33,35,61] centromeres, peri-centromeres, and telomeres[65] across a wide range of samples. Nanopore long reads can simultaneously detect mosaic DNA methylation and somatic sequence variation.[66,67] Here, we demonstrate that MASTR-seq provides a facile way to increase on-target reads to quantify repeat unit length and DNA methylation in the same single-read in pre-determined repetitive sequences across multiple samples at a single-allele resolution. MASTR-seq will broadly

scale to applications in which there is a specific repetitive sequence of interest across multiple samples to understand genotype and DNA methylation variation. The applications are not limited to STR tracts and could involve genotyping technically intractable SINE, LINE, centromere, and telomere tracts for a comparison of multiple samples in one nanopore run.

2. multiplexed PCR-free barcoding for pooling and cost-effective multi-modal measurements as a defined target locus in multiple samples

Previously published methods involving Nanopore or PacBio sequencing coupled with Cas9 enrichment highlight the power of the Cas9-mediated enrichment step in focusing reads on a target locus to save cost.[19,21,29,33–42,51,52] The Cas9-enrichment step opens the possibility of pooling multiple target loci from the same single cell line or the same locus across multiple cell lines for high-throughput comparative analysis using a single flow cell. Proof of concept of Cas9 capture long-read sequencing at single loci in single samples has been published (e.g.,[28,32,33,35,37–41,51,52]); however, early data suggest that there is a major reduction in on-target read coverage and proportion of on-target reads mapped upon pooling samples for analysis.[35] Here, by adding a large fragment size selection step, we increase on-target read proportion for a single locus compared to the other main known published Nanopore Cas9 capture method applied for sample multiplexing[35] (Table S4). To our knowledge, all other methodologies either involved multiple target loci or single samples and cannot be compared (Table S4). We further employ PCR-free barcoding to allow for the pooling of up to 8–12 different patient-derived cell line samples in a single MinION flow cell, which reduces cost and allows for high-throughput screening of more samples (Table S4). Additionally, integrating MASTR-seq's high sample throughput with a recently developed, cost-effective *in vitro* transcription (IVT) guide RNA strategy[68] offers a transformative approach for detecting genomic mutations across multiple loci in multiple samples in a single experiment.

In addition to reducing per-sample sequencing cost, sample multiplexing in MASTR-seq enables robust allele-specific comparisons across individuals or conditions within a single run, minimizing run-to-run variability. Based on our current implementation, we successfully multiplexed up to 12 samples per MinION flow cell using ONT's native barcoding kits, achieving ~260 averaged on-target reads per sample for resolving expanded STR alleles such as *FMR1*(Table S3). Given that our experiments used approximately 20% of pores in each flow cell with 12 samples pooled, we estimate that up to 48 samples could in principle be loaded in a given flow cell for simple, shorter, and less heterogeneous STR tracts. We note that simpler STR tracts with shorter ML, minimal allele heterogeneity, or lower CG content may require lower read depth for quantitative evaluation of tract length, and therefore more samples could be pooled. For complex loci, we recommend limiting multiplexing to 12 samples to ensure adequate resolution, whereas simpler loci (e.g., *HTT* CAG STR) may tolerate higher degrees of multiplexing without compromising accuracy. Future development will focus on adapting MASTR-seq to larger flow cells, such as PromethION, to further enhance target and sample throughput, solidifying its potential as a powerful tool for target sequencing and clinical diagnostics.

Experimental design considerations

We note that two types of multiplexing are available for targeted nanopore sequencing assays. One approach involves use of multiple sgRNA and a single sample to analyze multiple loci (sgRNAs multiplexing).[34,64] MASTR-seq focuses on a single locus with locus-specific sgRNAs and multiplexing up to 12 independent samples (sample multiplexing). It is noteworthy that both approaches achieve similar on-target read coverage per sample (Table S4).[64] Multiplexing loci is useful for scenarios in which the scientist desires to query multiple genomic targets from a single sample for a single flow cell run, for example, in the case of a clinical diagnostic sample requiring screening for a panel of possible STR expansion events. By contrast, multiplexing samples is useful in cases of basic science studies or large panels of human patient studies where a target locus is studied across a wide range of conditions, perturbations, or disease states. Hybrid strategies multiplexing both loci and samples could be explored in the future by merging advances between MASTR-seq and locus multiplexing methods.[34,64,68]

Comparison with other methods

Several variations on long-read sequencing have been developed to enrich for specific genomic DNA sequences. Among the most promising are CRISPR-mediated enrichment approaches that cut around a locus of interest to enrich it in the molecular sample compared to all other genomic fragments genome-wide.[18,20,28,32–41,51,52] MASTR-seq is a detailed and scalable protocol that builds on the Cas9-mediated enrichment principle by adding flexible features useful for the flexible adaptation to a myriad of basic science and clinical applications.

One major challenge with existing methods is that the on-target read mapping rate can be as low as <0.1% despite the CRISPR enrichment. Thus, money is wasted on sequencing genomic fragments irrelevant to the biological question and limiting the pooling of multiple samples for measurements. We add a fragment size selection step to increase the on-target read mapping proportion (Table S4), therefore allowing for (1) increased sampling to facilitate STR genotyping accuracy and (2) the re-distribution of read availability to allow for multiplexing multiple samples. The presence of adapted off-target DNA fragments causes competition of reads during nanopore sequencing, leading to low on-target read yield. Prior to improvements in MASTR-seq, only a limited number of samples could be sequenced with custom protocols in parallel due to severe challenges in cost and poor sensitivity in genotype measurements due to low on-target read coverage.[35] MASTR-seq's advances enable significant decreases in cost for multi-modal measurements of target regions across multiple samples.

A second challenge with existing methods is their widespread flexibility for a range of sequences and accuracy of genotype measurement in locations of the repetitive genome. For applications involving the quantification of STR tract length, it is important to address biases caused by PCR amplification. We also add PCR-free multiplexed barcoding to MASTR-seq, which

offers the advantage of reducing polymerase-based errors that occur at high frequency in repetitive regions of the genome. Both the PacBio RS II and ONT MinION platforms can sequence across challenging repetitive sequences with 100% CG-content, but the accuracy remains a hurdle[38] that future upgrades to the sequencing platforms themselves will address. Given that STR expansions, such as *C9orf72* and *FMR1*, vary across tissues within the same patient,[69,70] multiplexed, high-coverage long-read sequencing studies that accurately genotype repeat sequences are necessary to understand disease onset and progression. The design features of MASTR-seq increase accuracy of long reads at highly repetitive regions susceptible to PCR bias and also allow for the pooling of up to 8–12 samples on a single MinION flow cell.

A third major challenge is that existing methods have sparse experimental and computational details, or they are rigid kits that are inflexible to scale to researcher's custom model systems, custom target loci, and new technology advances in hardware and computational tools. Recently, older versions of MinION flow cells (R9.4.1) have been deprecated, and previously published targeted nanopore methods and kits are incompatible with the R10.4.1 flow cells. Here, we present a detailed and flexibly scalable experimental and computational pipelines that are compatible with new advances and will enable researchers to accurately analyze STR length and DNA methylation in the same single molecules with minimal new technology development.

Limitations of the study

1. High input requirement

MASTR-seq requires 5–15 μg DNA input per sample to achieve efficient on-target coverage. When using rare tissue samples, there is limited DNA amount available. The problem of high input for quality data is pervasive across nearly all genomic assays, in particular for PCR-free methods to improve read accuracy at repetitive regions, and it is not unique to MASTR-seq. As PCR-free library preparation kits advance sensitivity, the input can be reduced in the future. Investigators can continue to optimize their workflows for their customized applications by (1) optimizing sequencing adaptor ligation efficiency and (2) increasing the efficiency of the DNA extraction, Cas9 cleavage, and DNA size selection steps. Although we believe MASTR-seq will scale to clinical samples with sufficient cell numbers to reach the requirement of 5–15 μg DNA input per sample, it will be important to optimize PCR-free experimental conditions for rare patient-derived samples in the future.

2. Time-consuming size selection

The BluePippin size selection step currently takes at least 6 h, depending on sample quantity. To address this issue, it might be possible in the future to use the PippinHT System's high-throughput solution, which streamlines the process and achieves sample reproducibility with 6 h. This system has a capacity for 24 samples, significantly reducing time investment and meeting the demands of high-throughput sequencing.

3. A limited number of STRs loci tested

MASTR-seq is developed for genotyping STR loci defined by specifying the repeating motif, but it does not yet provide a comprehensive method for analyzing the full range of STRs across the genome. Expanding its compatibility with additional STRs, especially those with complex repeat structures, may require modifications such as integrating Hidden Markov Models (HMMs) to improve the identification of repeats with interrupted or variable motifs. For reference, the publicly available HMMSTR tool uses a profile HMM-based framework to estimate STR copy numbers from targeted sequencing data,[64] whereas NanoRepeat applies a Gaussian mixture model to classify reads into alleles.[71] These complementary strategies illustrate potential future directions for enhancing MASTR-seq's flexibility in integrating STR and variant analysis capabilities with established public tools. Additionally, unlike whole-genome STR detection methods, MASTR-seq requires prior knowledge of the target loci and may not be well suited for discovering novel repeat expansions.

4. Challenges with ultra-long STR expansions

The current protocol was tested on medium-length repeats (up to ~6 kb), but sequencing ultra-long expansions (>10kb) remains unknown as we have very limited samples available to tested in hand. A sequencing coverage significantly drops for these long-expanded repeats compared to those normal-length (NL) repeats (e.g., C9orf72 expansions >5kb show 26× coverage, much lower than NL repeats with 64× coverage). Future protocol optimizations should address potential biases in long-repeat sequencing.

5. Limited on-target read fraction and demultiplexing efficiency

Compared to PureTarget (>70% on-target reads), MASTR-seq exhibits a lower fraction of on-target reads. To improve on-target read fraction, guide RNA design should be optimized to minimize off-target Cas9 cleavage; stricter size selection and precise adapter concentration adjustments can further reduce non-specific fragments. Furthermore, demultiplexing efficiency remains a challenge, with only 26%–40% of reads correctly assigned to barcodes. Sequencing most unbarcoded data impacts per-sample read depth and cost-efficiency. Future optimizations in barcoding strategies, such as merging barcoding with adapter ligation into a single step, could help streamline the process and improve sequencing efficiency. Older published methods and kits have recently been deprecated due to outdated flow cells or low accuracy of older base calling methods; therefore, MASTR-seq fulfills a critical unmet need.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to lead contact, J.E. Phillips-Cremins (jennifer.cremins@wustl.edu).

Materials availability

Materials used in the current study are available from the lead contact upon request.

Data and code availability

- All MASTR-seq processed data have been provided in Table S5. New generated data in this study have been deposited at GEO accession GSE265944. Raw data have been deposited at Zenodo link: Run1&2: https://zenodo.org/records/17594975; Run3: https://zenodo.org/records/17612978; and Run4: https://doi.org/10.5281/zenodo.17618429.
- The original code for STR counting and DNA methylation assay is in https://doi.org/10.5281/zenodo.17917641 (Zenodo) and at GitHub link: https://github.com/chuanbinhub/MASTRseq.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank members of the Cremins lab for helpful discussions. H.-S.R. is the recipient of the Michael S. Brown Graduate Research Fellowship. T.M. is the recipient of the NIH F31 Fellowship. J.E.P.-C. was a New York Stem Cell Foundation Robertson Investigator and an Alfred P. Sloan Foundation Fellow during this project. This study was supported by NIH NIMH (2R01MH120269; J.E.P.-C.), NSF CAREER Award (CBE1943945; J.E.P.-C.), 4D Nucleome Common Fund (1U01DA052715; J.E.P.-C.), 4D Nucleome Common Fund (1U01DK127405; J.E.P.-C.), NIMH (1DP1MH129957; J.E.P.-C.), CZI Neurodegenerative Disease Pairs Award (DAF2022-250430; J.E.P.-C.), CZI Neurodegenerative Disease Pairs Award (DAF2022-250577; J.E.P.-C.), Cure Huntington's Disease Initiative (CHDI; ID A-18744, to J.E.P.-C.), NIH F31 Fellowship (F31NS129317; to T.M.) and NIH F30 (F300HD098015; to L.Z.).

AUTHOR CONTRIBUTIONS

Conceptualization, C.S. and J.E.P.-C; experimentation, C.S., H.-S.R., T.M., R.B., and L.Z.; computation and visualization, C.S., H.-S.R., and K.R.C.; funding, J.E.P.-C., H.-S.R., and T.M.; administration, J.E.P.-C.; writing, C.S., H.R., and J.E.P.-C.; review and editing, C.S., H.-S.R., K.R.C., T.M., R.B., and J.E.P.-C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell lines
  - Tissue samples
- METHOD DETAILS
  - Preparation of buffers and solutions
  - Cell culture and cell counting
  - Southern blotting
  - AmplideX mPCR of *FMR1* CGG STR
  - Multiplex analysis of STR by target nanopore sequencing (MASTR-seq)
  - HMW genomic DNA preparation and quality assay
  - Guide RNA design and crRNA/tracrRNA annealing
  - Cas9 RNP assembly
  - HMW genomic DNA dephosphorylation
  - Cas9 RNP cleavage, dA tailing and purification
  - BluePippin size selection
  - Library preparation and sequencing
  - STR tract length and DNA methylation analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - On-target reads calculation
  - CpG DNA methylation analysis at the *FMR1* promoter using nanopolish and data derived from an R9.4.1 flow cell
  - CpG DNA methylation analysis at the CGG STR using STRique and data derived from an R9.4.1 flow cell

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2026.101341.

Received: May 4, 2025
Revised: September 6, 2025
Accepted: February 9, 2026
Published: March 17, 2026

REFERENCES

1. La Spada, A.R., and Taylor, J.P. (2010). Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat. Rev. Genet. *11*, 247–258. https://doi.org/10.1038/nrg2748.
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. https://doi.org/10.1038/s41580-021-00382-6.
3. McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. Nat. Rev. Genet. *11*, 786–799. https://doi.org/10.1038/nrg2828.
4. Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940. https://doi.org/10.1038/nature05977.
5. Nelson, D.L., Orr, H.T., and Warren, S.T. (2013). The unstable repeats–three evolving faces of neurological disease. Neuron *77*, 825–843. https://doi.org/10.1016/j.neuron.2013.02.022.
6. Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742. https://doi.org/10.1038/nrg1689.
7. Fungtammasan, A., Ananda, G., Hile, S.E., Su, M.S.W., Sun, C., Harris, R., Medvedev, P., Eckert, K., and Makova, K.D. (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. *25*, 736–749. https://doi.org/10.1101/gr.185892.114.
8. Wagner, J., Olson, N.D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.C., Gupta, R., Wenger, A.M., Rowell, W.J., et al. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. Nat. Biotechnol. *40*, 672–680. https://doi.org/10.1038/s41587-021-01158-1.
9. Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46. https://doi.org/10.1038/nrg3117.
10. Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J., and Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am. J. Hum. Genet. *103*, 858–873. https://doi.org/10.1016/j.ajhg.2018.10.015.
11. 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. https://doi.org/10.1126/science.abj6987.
12. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84. https://doi.org/10.1038/s41586-020-2547-7.
13. 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. https://doi.org/10.1038/nbt.4060.
14. Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al. (2021). The structure,

function and evolution of a complete human chromosome 8. Nature *593*, 101–107. https://doi.org/10.1038/s41586-021-03420-7.

15. Porubsky, D., Dashnow, H., Sasani, T.A., Logsdon, G.A., Hallast, P., Noyes, M.D., Kronenberg, Z.N., Mokveld, T., Koundinya, N., Nolan, C., et al. (2025). Human de novo mutation rates from a four-generation pedigree reference. Nature *643*, 427–436. https://doi.org/10.1038/s41586-025-08922-2.
16. Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. https://doi.org/10.1038/nrg.2016.49.
17. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345–353. https://doi.org/10.1038/nature24286.
18. van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends Genet. *34*, 666–681. https://doi.org/10.1016/j.tig.2018.05.008.
19. Mitsuhashi, S., and Matsumoto, N. (2020). Long-read sequencing for rare human genetic diseases. J. Hum. Genet. *65*, 11–19. https://doi.org/10.1038/s10038-019-0671-8.
20. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nat. Rev. Genet. *21*, 597–614. https://doi.org/10.1038/s41576-020-0236-x.
21. 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*, 3586. https://doi.org/10.1038/s41467-021-23918-y.
22. Marx, V. (2023). Method of the year: long-read sequencing. Nat. Methods *20*, 6–11. https://doi.org/10.1038/s41592-022-01730-w.
23. Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., and Sedlazeck, F.J. (2019). Structural variant calling: the long and the short of it. Genome Biol. *20*, 246. https://doi.org/10.1186/s13059-019-1828-7.
24. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biol. *21*, 30. https://doi.org/10.1186/s13059-020-1935-5.
25. Zhou, Z.D., Jankovic, J., Ashizawa, T., and Tan, E.K. (2022). Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. Nat. Rev. Neurol. *18*, 145–157. https://doi.org/10.1038/s41582-021-00612-7.
26. Kernohan, K.D., and Boycott, K.M. (2024). The expanding diagnostic toolbox for rare genetic diseases. Nat. Rev. Genet. *25*, 401–415. https://doi.org/10.1038/s41576-023-00683-w.
27. Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. *27*, 677–685. https://doi.org/10.1101/gr.214007.116.
28. Merker, J.D., Wenger, A.M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K.S., et al. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet. Med. *20*, 159–163. https://doi.org/10.1038/gim.2017.86.
29. 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. https://doi.org/10.1038/s41587-019-0293-x.
30. De Roeck, A., De Coster, W., Bossaerts, L., Cacace, R., De Pooter, T., Van Dongen, J., D'Hert, S., De Rijk, P., Strazisar, M., Van Broeckhoven, C., and Sleegers, K. (2019). NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. Genome Biol. *20*, 239. https://doi.org/10.1186/s13059-019-1856-3.
31. Hiatt, S.M., Lawlor, J.M.J., Handley, L.H., Ramaker, R.C., Rogers, B.B., Partridge, E.C., Boston, L.B., Williams, M., Plott, C.B., Jenkins, J., et al. (2021). Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. HGG Adv. *2*, 100023. https://doi.org/10.1016/j.xhgg.2021.100023.
32. 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. https://doi.org/10.1186/s13059-019-1667-6.
33. Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., et al. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat. Genet. *51*, 1215–1221. https://doi.org/10.1038/s41588-019-0459-y.
34. 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. https://doi.org/10.1038/s41587-020-0407-5.
35. Alfano, M., De Antoni, L., Centofanti, F., Visconti, V.V., Maestri, S., Degli Esposti, C., Massa, R., D'Apice, M.R., Novelli, G., Delledonne, M., et al. (2022). Characterization of full-length CNBP expanded alleles in myotonic dystrophy type 2 patients by Cas9-mediated enrichment and nanopore sequencing. eLife *11*, e80229. https://doi.org/10.7554/eLife.80229.
36. Rubben, K., Tilleman, L., Deserranno, K., Tytgat, O., Deforce, D., and Van Nieuwerburgh, F. (2022). Cas9 targeted nanopore sequencing with enhanced variant calling improves CYP2D6-CYP2D7 hybrid allele genotyping. PLoS Genet. *18*, e1010176. https://doi.org/10.1371/journal.pgen.1010176.
37. Lee, I., Workman, R., Wang, J.Z., and Timp, W. (2019). Hybridization-capture for nanopore sequencing. Protocols.io. https://doi.org/10.17504/protocols.io.zxyf7pw.
38. Ebbert, M.T.W., Farrugia, S.L., Sens, J.P., Jansen-West, K., Gendron, T.F., Prudencio, M., McLaughlin, I.J., Bowman, B., Seetin, M., DeJesus-Hernandez, M., et al. (2018). Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. Mol. Neurodegener. *13*, 46. https://doi.org/10.1186/s13024-018-0274-4.
39. Stevens, R.C., Steele, J.L., Glover, W.R., Sanchez-Garcia, J.F., Simpson, S.D., O'Rourke, D., Ramsdell, J.S., MacManes, M.D., Thomas, W.K., and Shuber, A.P. (2019). A novel CRISPR/Cas9 associated technology for sequence-specific nucleic acid enrichment. PLoS One *14*, e0215441. https://doi.org/10.1371/journal.pone.0215441.
40. Wallace, A.D., Sasani, T.A., Swanier, J., Gates, B.L., Greenland, J., Pedersen, B.S., Varley, K.E., and Quinlan, A.R. (2021). CaBagE: A Cas9-based Background Elimination strategy for targeted, long-read DNA sequencing. PLoS One *16*, e0241253. https://doi.org/10.1371/journal.pone.0241253.
41. Hoijer, I., Tsai, Y.C., Clark, T.A., Kotturi, P., Dahl, N., Stattin, E.L., Bondeson, M.L., Feuk, L., Gyllensten, U., and Ameur, A. (2018). Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. Hum. Mutat. *39*, 1262–1272. https://doi.org/10.1002/humu.23580.
42. Hafford-Tear, N.J., Tsai, Y.C., Sadan, A.N., Sanchez-Pintado, B., Zarouchlioti, C., Maher, G.J., Liskova, P., Tuft, S.J., Hardcastle, A.J., Clark, T.A., and Davidson, A.E. (2019). CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat. Genet. Med. *21*, 2092–2102. https://doi.org/10.1038/s41436-019-0453-x.
43. Bethune, K., Mariac, C., Couderc, M., Scarcelli, N., Santoni, S., Ardisson, M., Martin, J.F., Montúfar, R., Klein, V., Sabot, F., et al. (2019). Long-fragment targeted capture for long-read sequencing of plastomes. Appl. Plant Sci. *7*, e1243. https://doi.org/10.1002/aps3.1243.
44. Eckert, S.E., Chan, J.Z.M., Houniet, D., The Pathseek, C., Breuer, J., and Speight, G. (2016). Enrichment by hybridisation of long DNA fragments for

Nanopore sequencing. Microb. Genom. *2*, e000087. https://doi.org/10.1099/mgen.0.000087.

45. Fisher, C.L., Dillon, R., Anguita, E., Morris-Rosendahl, D.J., and Awan, A.R. (2022). A Novel Bead-Capture Nanopore Sequencing Method for Large Structural Rearrangement Detection in Cancer. J. Mol. Diagn. *24*, 1264–1278. https://doi.org/10.1016/j.jmoldx.2022.09.006.
46. Loose, M., Malla, S., and Stout, M. (2016). Real-time selective sequencing using nanopore technology. Nat. Methods *13*, 751–754. https://doi.org/10.1038/nmeth.3930.
47. Miyatake, S., Koshimizu, E., Fujita, A., Doi, H., Okubo, M., Wada, T., Hamanaka, K., Ueda, N., Kishida, H., Minase, G., et al. (2022). Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. NPJ Genom. Med. *7*, 62. https://doi.org/10.1038/s41525-022-00331-y.
48. 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. https://doi.org/10.1126/sciadv.abm5386.
49. Kovaka, S., Fan, Y., Ni, B., Timp, W., and Schatz, M.C. (2021). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat. Biotechnol. *39*, 431–441. https://doi.org/10.1038/s41587-020-0731-9.
50. Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B.J., and Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat. Biotechnol. *39*, 442–450. https://doi.org/10.1038/s41587-020-00746-x.
51. Hoijer, I., Johansson, J., Gudmundsson, S., Chin, C.S., Bunikis, I., Haggqvist, S., Emmanouilidou, A., Wilbe, M., den Hoed, M., Bondeson, M.L., et al. (2020). Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. Genome Biol. *21*, 290. https://doi.org/10.1186/s13059-020-02206-w.
52. Watson, C.M., Crinnion, L.A., Hewitt, S., Bates, J., Robinson, R., Carr, I.M., Sheridan, E., Adlard, J., and Bonthron, D.T. (2020). Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications. Lab. Invest. *100*, 135–146. https://doi.org/10.1038/s41374-019-0283-0.
53. Gupta, D., and Beisel, C.L. (2021). Illuminating the path to DNA repair. Cell *184*, 5503–5505. https://doi.org/10.1016/j.cell.2021.10.005.
54. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191. https://doi.org/10.1038/nbt.3437.
55. Sereika, M., Kirkegaard, R.H., Karst, S.M., Michaelsen, T.Y., Sørensen, E.A., Wollenberg, R.D., and Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat. Methods *19*, 823–826. https://doi.org/10.1038/s41592-022-01539-7.
56. Ni, Y., Liu, X., Simeneh, Z.M., Yang, M., and Li, R. (2023). Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput. Struct. Biotechnol. J. *21*, 2352–2364. https://doi.org/10.1016/j.csbj.2023.03.038.
57. Sanderson, N.D., Kapel, N., Rodger, G., Webster, H., Lipworth, S., Street, T.L., Peto, T., Crook, D., and Stoesser, N. (2023). Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. Microb. Genom. *9*, mgen000910. https://doi.org/10.1099/mgen.0.000910.
58. Zhang, T., Li, H., Ma, S., Cao, J., Liao, H., Huang, Q., and Chen, W. (2023). The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. Appl. Environ. Microbiol. *89*, e0060523. https://doi.org/10.1128/aem.00605-23.
59. Deserranno, K., Tilleman, L., Rubben, K., Deforce, D., and Van Nieuwerburgh, F. (2023). Targeted haplotyping in pharmacogenomics using Oxford Nanopore Technologies' adaptive sampling. Front. Pharmacol. *14*, 1286764. https://doi.org/10.3389/fphar.2023.1286764.
60. Tan, K.T., Slevin, M.K., Meyerson, M., and Li, H. (2022). Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. Genome Biol. *23*, 180. https://doi.org/10.1186/s13059-022-02751-6.
61. Malachowski, T., Chandradoss, K.R., Boya, R., Zhou, L., Cook, A.L., Su, C., Pham, K., Haws, S.A., Kim, J.H., Ryu, H.S., et al. (2023). Spatially coordinated heterochromatinization of long synaptic genes in fragile X syndrome. Cell *186*, 5840–5858.e36. https://doi.org/10.1016/j.cell.2023.11.019.
62. Monaghan, K.G., Lyon, E., and Spector, E.B.; American College of Medical Genetics and Genomics (2013). ACMG Standards and Guidelines for fragile X testing: a revision to the disease-specific supplements to the Standards and Guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics and Genomics. Genet. Med. *15*, 575–586. https://doi.org/10.1038/gim.2013.61.
63. Ruzo, A., Croft, G.F., Metzger, J.J., Galgoczi, S., Gerber, L.J., Pellegrini, C., Wang, H., Jr., Fenner, M., Tse, S., Marks, A., et al. (2018). Chromosomal instability during neurogenesis in Huntington's disease. Development *145*, dev156844. https://doi.org/10.1242/dev.156844.
64. Van Deynze, K., Mumm, C., Maltby, C.J., Switzenberg, J.A., Todd, P.K., and Boyle, A.P. (2025). Enhanced detection and genotyping of disease-associated tandem repeats using HMMSTR and targeted long-read sequencing. Nucleic Acids Res. *53*, gkae1202. https://doi.org/10.1093/nar/gkae1202.
65. Sholes, S.L., Karimian, K., Gershman, A., Kelly, T.J., Timp, W., and Greider, C.W. (2022). Chromosome-specific telomere lengths and the minimal functional telomere revealed by nanopore sequencing. Genome Res. *32*, 616–628. https://doi.org/10.1101/gr.275868.121.
66. Gigante, S., Gouil, Q., Lucattini, A., Keniry, A., Beck, T., Tinning, M., Gordon, L., Woodruff, C., Speed, T.P., Blewitt, M.E., and Ritchie, M.E. (2019). Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Res. *47*, e46. https://doi.org/10.1093/nar/gkz107.
67. Battaglia, S., Dong, K., Wu, J., Chen, Z., Najm, F.J., Zhang, Y., Moore, M.M., Hecht, V., Shoresh, N., and Bernstein, B.E. (2022). Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. Nat. Genet. *54*, 1504–1513. https://doi.org/10.1038/s41588-022-01188-8.
68. 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. Preprint at bioRxiv. https://doi.org/10.1101/2023.02.07.527484.
69. Dols-Icardo, O., García-Redondo, A., Rojas-García, R., Sánchez-Valle, R., Noguera, A., Gómez-Tortosa, E., Pastor, P., Hernández, I., Esteban-Pérez, J., Suárez-Calvet, M., et al. (2014). Characterization of the repeat expansion size in C9orf72 in amyotrophic lateral sclerosis and frontotemporal dementia. Hum. Mol. Genet. *23*, 749–754. https://doi.org/10.1093/hmg/ddt460.
70. Nordin, A., Akimoto, C., Wuolikainen, A., Alstermark, H., Jonsson, P., Birve, A., Marklund, S.L., Graffmo, K.S., Forsberg, K., Brännström, T., and Andersen, P.M. (2015). Extensive size variability of the GGGGCC expansion in C9orf72 in both neuronal and non-neuronal tissues in 18 patients with ALS or FTD. Hum. Mol. Genet. *24*, 3133–3142. https://doi.org/10.1093/hmg/ddv064.
71. Fang, L., Monteys, A.M., Dürr, A., Keiser, M., Cheng, C., Harapanahalli, A., Gonzalez-Alegre, P., Davidson, B.L., and Wang, K. (2023). Haplotyping SNPs for allele-specific gene editing of the expanded huntingtin allele using long-read sequencing. HGG Adv. *4*, 100146. https://doi.org/10.1016/j.xhgg.2022.100146.
72. 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. https://doi.org/10.1093/nar/gkz365.

73. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
74. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
75. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
76. Molder, F., Jablonski, K.P., Letcher, B., Hall, M.B., van Dyken, P.C., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Vieira, F.G., Meesters, C., et al. (2021). Sustainable data analysis with Snakemake. F1000Res *10*, 33. https://doi.org/10.12688/f1000research.29032.3.
77. Shah, S., Sharp, K.J., Raju Ponny, S., Lee, J., Watts, J.K., Berry-Kravis, E., and Richter, J.D. (2023). Antisense oligonucleotide rescue of CGG expansion-dependent FMR1 mis-splicing in fragile X syndrome restores FMRP. Proc. Natl. Acad. Sci. USA *120*, e2302534120. https://doi.org/10.1073/pnas.2302534120.
78. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.
79. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods *14*, 407–410. https://doi.org/10.1038/nmeth.4184.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
FXS human caudate nucleus brain tissue from NIH donor 5006 (Designated as FXS#5006)	NIH NeuroBioBank	https://neurobiobank.nih.gov
FXS patient postmortem brain tissues #1031-08-GP were obtained UC-Davis (Designated as 103108GP)	UC-Davis MIND Institute	1031-08-GP
Chemicals, peptides, and recombinant proteins		
mTeSR Plus media	STEMCELL Technology	Cat# 100-0276
Matrigel hESC-Qualified Matrix	Corning	Cat# 354277
1X PBS without calcium and magnesium	Corning	Cat# 21-040-CV
Penicillin-streptomycin	Gibco	Cat# 15140122
TrypLE™ Express Enzyme (1X), phenol red	Gibco	Cat# 12605010
Versene Solution	Gibco	Cat# 15040066
10X TAE buffer	Invitrogen	Cat# 15558042
1M Tris-HCl (pH 8.0)	Invitrogen	Cat# 15568025
0.5M EDTA (pH 8.0)	Invitrogen	Cat# 15575020
20% SDS (w/v) solution	Fisher Scientific	Cat# BP1311-200
RNAse A	Sigma-Aldrich	Cat# 10109142001
Proteinase K	New England Biolabs	Cat# P8107S
UltraPure Phenol/Chloroform/Isoamyl Alcohol	Invitrogen	Cat# 15593049
5M Ammonium Acetate	Invitrogen	Cat# AM9070G
100% Ethanol	Decon Labs	Cat# 2716
0.75% Agarose Gel Cassette	Sage Science	Cat# BLF7510
S. pyogenes Cas9 tracrRNA	Integrated DNA Technologies	Cat# 1072532
Alt-R S. pyogenes HiFi Cas9 nuclease V3	Integrated DNA Technologies	Cat# 1081060
Quick Calf Intestinal Phosphatase	New England Biolabs	Cat# M0525S
rCutSmart buffer	New England Biolabs	Cat# B6004S
Nuclease-free water	ThermoFisher	Cat# AM9937
Taq DNA polymerase	New England Biolabs	Cat# M0273S
100 mM dATP solution	Thermo Scientific,	Cat# R0141
Nuclease-free duplex buffer	Integrated DNA Technologies	Cat# 11-01-03-01
Quick-Load 1 kb Extend DNA Ladder	New England Biolabs	Cat# N3239S
NEB Blunt/TA Ligase Master Mix	New England Biolabs	Cat# M0367L
T4 DNA ligase, from NEBNext Quick Ligation Module	New England Biolabs	Cat# E6056S
EcoRI	New England Biolabs	Cat# R0101S
BssH II	New England Biolabs	Cat# R0199S
AP Detection buffer 10X (Alkaline Phos. buffer)	Gene Link	Cat# 40-5031-10
AP Substrate CDP Star Tropix	Gene Link	Cat# 40-5010-10
20X SSC	Gene Link	Cat# 40-5020-25
DIG Easy Hyb	Roche	Cat# T3010S
Fragile X GeneProber™ GLFXDig1 Probe Digoxigenin labeled	Gene Link	Cat# 40-2004-41
Blocking solution 10%	Gene Link	Cat# 40-5026-10
Maleic acid buffer 10X (Buffer M 10X)	Gene Link	Cat# 40-5025-20

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
AmplideX mPCR FMR1 Kit	Asuragen	Cat# 49442
Monarch HMW DNA Extraction Kit for Tissue	New England Biolabs	Cat# T3060S
Monarch HMW DNA Extraction Kit for cell & blood	New England Biolabs	Cat# T3050S
Ligation Sequencing Kit V14	Oxford Nanopore Technologies	Cat# SQK-LSK114
Native Barcoding Kit 24 V14	Oxford Nanopore Technologies	Cat# SQK-NBD114.24
Monarch Spin gDNA Extraction Kit	New England Biolabs	Cat# T3010S
Flow Cell Priming Kit	Oxford Nanopore Technologies	Cat# EXP-FLP004
Qubit™ Broad Range dsDNA Quantitation	Invitrogen	Cat# Q32850
Qubit™ High Sensitivity dsDNA Quantitation	Invitrogen	Cat# Q32851
Bioanalyzer High Sensitivity DNA Analysis Kit	Agilent	Cat# 5067-4626
Deposited data		
NL#1(NL_18) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
NL#2(NL_27) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
NL#3(NL_25) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PM#1(PM_137) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
FXS#1(FXS_421) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
FXS#2(FXS_426) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
FXS#3(FXS_470) provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone1 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone2 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone3 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone4 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone5 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone6 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
ML_scClone7 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone1 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone2 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone3 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone4 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone5 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone6 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
PMcut_scClone7 provided in Table S5	Malachowski et al.[61]	GEO: GSE218679
FXS_#1(FXS_421) rep1	This study	GEO: GSE265944
FXS_#1(FXS_421) rep2	This study	GEO: GSE265944
FXS_#3(FXS_470) rep1	This study	GEO: GSE265944
FXS_#3(FXS_470) rep2	This study	GEO: GSE265944
HTT_parental	This study	GEO: GSE265944
HTT_Q20	This study	GEO: GSE265944
HTT_Q56	This study	GEO: GSE265944
HTT_Q72	This study	GEO: GSE265944
ALS_19	This study	GEO: GSE265944
FXS#5006	This study	GEO: GSE265944
Experimental models: Cell lines		
Human healthy iPS cell line - NL_18 (designated as NL#1 in this study)	Malachowski et al.[61]	NL_18
Human healthy iPS cell line - NL_27 (designated as NL#2 in this study)	Malachowski et al.[61]	NL_27

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human healthy iPS cell line - NL_25 (designated as NL#3 in this study)	Malachowski et al.[61]	NL_25
Human pre-mutation iPS cell line - PM_137 (designated as PM#1 in this study)	Malachowski et al.[61]	PM_137
Human FXS iPS cell line - FXS_421 (designated as FXS#1 in this study)	Malachowski et al.[61]	FXS_421
Human FXS iPS cell line - FXS_426 (designated as FXS#2 in this study)	Malachowski et al.[61]	FXS_426
Human FXS iPS cell line - FXS_470 (designated as FXS#3 in this study)	Malachowski et al.[61]	FXS_470
Human FXS iPS cell line - 135.3_CGG_034 (designated as ML_scClone1 in this study)	Malachowski et al.[61]	ML_scClone1
Human FXS iPS cell line - 4H2 (designated as ML_scClone2 in this study)	Malachowski et al.[61]	ML_scClone2
Human FXS iPS cell line - 6D12 (designated as ML_scClone3 in this study)	Malachowski et al.[61]	ML_scClone3
Human FXS iPS cell line - 135.3_CGG_116 (designated as ML_scClone4 in this study)	Malachowski et al.[61]	ML_scClone4
Human FXS iPS cell line - 135.3_CGG_125 (designated as ML_scClone5 in this study)	Malachowski et al.[61]	ML_scClone5
Human FXS iPS cell line - 135.3_CGG_128 (designated as ML_scClone6 in this study)	Malachowski et al.[61]	ML_scClone6
Human FXS iPS cell line - 135.3_CGG_131 (designated as ML_scClone7 in this study)	Malachowski et al.[61]	ML_scClone7
Human FXS iPS cell line - 4D3 (designated as PMcut_scClone1 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone1
Human pre-mutation FXS iPS cell line - 135.3_CGG_117 (designated as PMcut_scClone2 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone2
Human pre-mutation FXS iPS cell line - 135.3_CGG_187 (designated as PMcut_scClone3 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone3
Human pre-mutation FXS iPS cell line - 135.3_CGG_275 (designated as PMcut_scClone4 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone4
Human pre-mutation FXS iPS cell line - 135.3_CGG_278 (designated as PMcut_scClone5 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone5
Human pre-mutation FXS iPS cell line - 135.3_CGG_030 (designated as PMcut_scClone6 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone6
Human pre-mutation FXS iPS cell line - 135.3_CGG_313 (designated as PMcut_scClone7 in this study)	Malachowski et al.[61]	ML_CUT_PM_scClone7
HTT_parental	Ruzo et al.[63]	RUES2 hESC line (NIH0013)
HTT_Q20	Ruzo et al.[63]	RUES2-22Q
HTT_Q56	Ruzo et al.[63]	RUES2-58Q
HTT_Q72	Ruzo et al.[63]	RUES2-74Q
ALS_19	This study	N/A
Oligonucleotides		
FMR1 targeted gRNA sequences	Malachowski et al.[61]	Table S2
HTT targeted gRNA sequences	This study	see Table S1
C9orf72 targeted gRNA sequences	This study	see Table S1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
ChopChop online tool	Labun et al.[72]	https://chopchop.cbu.uib.no/
MinKNOW	Oxford Nanopore Technologies	https://community.nanoporetech.com/downloads
Dorado	Oxford Nanopore Technologies	https://github.com/nanoporetech/dorado
Nanopolish	Gilpatrick et al.[34]	https://github.com/jts/nanopolish
Minimap2	Li[73]	https://github.com/lh3/minimap2
Samtools	Danecek et al.[74]	https://www.htslib.org
Guppy	Oxford Nanopore Technologies	https:/nanoporetech.com/document/Guppy-protocol
NanoStat	De Coster et al.[75]	https://github.com/wdecoster/nanostat
STRique	Giesselmann et al.[29]	https://github.com/giesselmann/STRique
snakemake	Molder et al.[76]	https://snakemake.github.io
modkit	Oxford Nanopore Technologies	https://github.com/nanoporetech/modkit
MASTRseq	This study	https://doi.org/10.5281/zenodo.17917641 and https://github.com/chuanbinhub/MASTRseq
GeneMapper	Applied Biosystems	https://www.thermofisher.com/order/catalog/product/4366925
Other		
Agencourt AMPure XP beads	Beckman Coulter	Cat# A63881
R10.4.1 flow cell	Oxford Nanopore Technologies	FLO-MIN114
Hemocytometer	Fisher Scientific	Cat# 0267151B
BluePippin	Sage Science	Cat# BLU0001
Agilent bioanalyzer	Agilent Technologies	Cat# G2939A
MinION Mk1B Sequencing Device	Oxford Nanopore Technologies	Cat# MIN-101B
Qubit™ 2.0 Fluorometer	Invitrogen	Cat# Q32866
Nanodrop 2000 Spectrophotometer	Thermo Scientific	Cat# ND-2000
Owl™ EasyCast™ B2 Mini Gel Electrophoresis device	Thermo Scientific	Cat# 09-528118
ChemiDoc™ MP Imaging System	Bio-rad	Cat# 12003154
Magnetic rack (DynaMag-2)	Thermo Fisher Scientific	Cat# 12321D
Vortex-Genie 2 mixer	Sigma-Aldrich	Cat# Z258415
Tube Revolver Rotator	Thermo Scientific	Cat# 88881001
Applied Biosystems™ 96 Well Thermal cycler	Thermo Scientific	Cat# 4375305
ART™ Wide Bore Filtered Pipette Tips	Thermo Scientific	Cat# 2069G

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell lines

We used FXS patient-derived human iPSCs with a normal-length (NL), PM, or full mutation-length (ML) *FMR1* CGG STR tract published in our recent public work[61] (Figure 2A). We also used human ESCs engineered with defined *HTT* CAG STR tracts at normal-length and mutation-length as created by the Brivanlou lab[63] (Figure 4A). We acquired ALS patient-derived iPSC lines with a mutation-length G4C2 STR in the *C9ORF72* gene from the Mazzoni lab at New York University (Figures 4C and 4D).

Tissue samples

FXS human caudate nucleus brain tissue was from NIH donor 5006 (designated as FXS#5006), FXS sample brain tissue (103108GP) was acquired from UC Davis MIND Institute.[77]

METHOD DETAILS

Preparation of buffers and solutions

Human iPSC culture medium

Prepare mTeSR Plus media by mixing 400 mL mTeSR Plus basal medium with 100 mL 5X supplement, supplemented with 1% (v/v) penicillin-streptomycin. Store media at 4°C and bring to room temperature (RT) before use.

Cell dissociation buffer

Prepare 0.5X TrypLE cell dissociation buffer at RT by diluting 1 volume of 1X TrypLE with 1 volume of 1X PBS without calcium and magnesium.

10 mM Tris-HCl (pH 8.0) buffer

Prepare 10 mM Tris-HCl (pH 8.0) buffer by adding 0.1 mL 1 M Tris-HCl (pH 8.0) to 9 mL nuclease-free water at RT.

crRNA stock solution (100 μM)

Prepare 100 μM crRNA stock solution by resuspending 5 nmol crRNA in 50 μL TE (pH 7.5).

Store aliquots at −80°C and thaw on ice before use.

tracrRNA stock solution (100 μM)

Prepare 100 μM tracrRNA stock solution by resuspending 5 nmol tracrRNA in 50 μL TE (pH 7.5). Store aliquots at −80°C and thaw on ice before use.

Cell lysis buffer

Prepare 50 mL of fresh cell lysis buffer at RT by mixing 500 μL of 1 M Tris-HCl (pH 8.0), 400 μL of 0.5 M EDTA (pH 8.0), 1250 μL of 20% SDS (w/v) in nuclease-free water.

Preparation of the Matrigel-coated plate

Coat iPSC culture plates with 1.2% (v/v) Matrigel hESC-Qualified Matrix in DMEM/F-12 for at least 1 h at 37°C. Perform all below steps at room temperature are 24°C unless otherwise stated in the protocol.

Cell culture and cell counting

Culture human iPSC lines in mTeSR Plus media supplemented with 1% (v/v) penicillin-streptomycin on Matrigel hESC-Qualified Matrix-coated plates at 37°C and 5% CO_2. To maintain the physical separation of single colonies without merging, routinely passage cells when they reach 60–70% confluence every 4–5 days. For iPSC passaging, incubate plates in Versene solution at 37°C for 5 min, followed by deactivation of Versene with 2X volume of mTeSR Plus media, spin down cell pellets at 200 x g for 3 min. Resuspend hiPSC pellets with mTeSR Plus media and split cell suspension into multiple new Matrigel matrix-coated plates. To count cells, incubate hiPSCs with 0.5X TrypLE cell dissociation buffer on the plate for 4 min. Then gently pipette to dissociate human iPSC colonies into single cells. Take 100 μL of suspended cells into a new microcentrifuge tube and add 100 μL 0.4% Trypan Blue (final concentration 0.2%). Mix gently and incubate for 2 min at RT. Count cell numbers manually with a hemocytometer or use the Countess automated cell counter to get cell numbers. Aliquot 5–10 million cells into each 2 mL microcentrifuge tube, spin down at 300 x g for 5 min to pellet the cells, remove supernatant, and flash-freeze pellets in liquid nitrogen. Store pellets at −80°C until use.

Southern blotting

The company (Celplor, Inc.) performed Southern blotting assay according to the manufacturer's instructions. Briefly, double digest approximately 10 μg genomic DNA with 10 units of enzyme (EcoRI and BssHII) overnight at 50°C for 8 h, then ethanol-precipitate, and resuspend the digested DNA in 1× loading buffer. Separate samples on a 0.8% agarose gel (14–16 h, 45 mA), depurinate (0.25 N HCl, 10 min), denature (0.4 N NaOH/0.6 M NaCl, 30 min), and neutralize (1.5 M NaCl/0.5 M Tris-HCl pH 7.5, 30 min). Then transfer DNA to a nylon membrane using 10× SSC overnight, rinse the membrane in 2× SSC, and bake at 80°C for 2 h. Prehybridize membrane in Easy-Hyb buffer at 55°C, 3 h (Roche #11603558001) and hybridize overnight with a DIG-labeled GeneProber (Gene Link #40-2004-41). Post-hybridization, two washes in 2× SSC/0.1% SDS (RT, 5 min) and two washes in 0.5× SSC/0.1% SDS (60°C, 15 min). Block membrane in Buffer MB (Gene Link# 40-5026-10) (RT, 30 min), incubate with Anti-DIG–alkaline phosphatase (1:10,000, 30 min, RT), wash the membrane twice, 15 min per wash in 200 mL of 1x Buffer M (Gene Link# 40-5025-20) at RT, and equilibrate in detection buffer for 2 min, repeat 2 times. Detect signal with CDP-Star substrate (Gene Link #40-5010-10) and visualize on X-ray film with 1 h to overnight exposure.

AmplideX mPCR of *FMR1* CGG STR

The company (Asuragen, Inc.) performed AmplideX FMR1 PCR assay using the AmplideX mPCR FMR1 Kit (Asuragen, #49442) following the manufacturer's protocol. Briefly, 8 μL of genomic DNA were digested with restriction enzymes in parallel FAM (control) and HEX (methylation-sensitive) reactions at 37°C for 2 h followed by 4°C. Digestion products were amplified in parallel reactions using the kit-supplied mPCR GC-Rich Amp Buffer, GC-Rich Polymerase Mix, and mPCR FAM- or HEX-primers, under the following cycling conditions: 95°C for 5 min; 27 cycles of 97°C for 35 s, 62°C for 35 s, and 72°C for 4 min; final extension at 72°C for 10 min; hold at 4°C. PCR products were separated by capillary electrophoresis using Formamide/ROX 1000 Size Ladder solution. We analyzed electropherograms with GeneMapper 4.0/4.1 to determine CGG repeat length (by fragment sizing).

Multiplex analysis of STR by target nanopore sequencing (MASTR-seq)

We targeted *FMR1*, *HTT*, and *C9orf72* STR loci using MASTR-seq, which combines CRISPR-Cas9 genomic digestion, size selection with nanopore library preparation and sequencing. We performed MASTR-seq as described in our recent work[62] with several modifications.

HMW genomic DNA preparation and quality assay

We obtain HMW DNA from 5-10$\times10^6$ iPSCs or 20 mg brain tissue with commercial kits (NEB #T3050S/3060S). Use a broad range dsDNA quantitation kit to detect HMW genomic DNA concentration. Assess gDNA purity with OD260/280 by Nanodrop. Check integrity of HMW DNA on a 0.6% 1X TAE agarose gel with 1 kb Extend DNA ladder (NEB#N3239S). For unsheared high-quality HMW DNA, we expect most of the DNA to be larger than 48.5 kb.

Guide RNA design and crRNA/tracrRNA annealing

We designed 4–6 crRNAs upstream and downstream of each STR using CHOPCHOP v3.0 (hg38, CRISPR-Cas9, nanopore enrichment mode; Figure S1; Table S1). We synthesized custom lyophilized crRNAs (2 nmol each) and tracrRNA (2 nmol, IDT 1072532), resuspended them to 100 μM in 10 mM Tris-EDTA (pH 7.5), and pooled the crRNAs equally. We prepared the duplex mix by combining pooled crRNAs (10/*N* μM each, where *N* is the number of crRNAs) with 10 μM tracrRNA in Duplex Buffer (IDT). We annealed crRNAs and tracrRNA by heating at 95°C for 5 min followed by cooling to room temperature, generating crRNA·tracrRNA duplexes.

Cas9 RNP assembly

We assembled CRISPR-Cas9 ribonucleoprotein (RNP) complexes *in vitro* in a 100 μL reaction, containing 10 μL of 10 μM crRNA·-tracrRNA pool, 10 μL of 10X NEB CutSmart buffer, 78.4 μL nuclease-free water (Sigma-Aldrich#W4502), and 1.6 μL of 62 μM HiFi Cas9 (IDT#1081060). We incubated the mixture at room temperature for 30 min and then placed it on ice until use.

HMW genomic DNA dephosphorylation

For each reaction, we used 5 μg DNA and dephosphorylated it with 3 μL QuickCIP (NEB# M0525S) in 1x rCutSmart Buffer. We commonly setup 2–3 replicates for each sample. Employ a thermal cycler to incubate above samples at 37°C for 20 min, followed by 80°C for 2 min, then hold at 20°C.

Cas9 RNP cleavage, dA tailing and purification

We digested dephosphorylated genomic DNA by incubating 5 μg DNA with 10 μL Cas9 RNPs, 10 mM dATP (Thermo Scientific#R0141), and 1 μL Taq polymerase (NEB#M0273) at 37°C for 60 min, followed by dA-tailing at 72°C for 5 min. We purified Cas9-cleaved DNA by adding 16 μL 5 M ammonium acetate (Invitrogen,#AM9070G) and 126 μL cold 100% ethanol, centrifuging at 16,000 × g for 5 min. We washed the DNA pellet twice with 70% ethanol, air-dried for 5 min at room temperature, and resuspended in 20 μL 10 mM Tris-HCl (pH 8.0) at 50°C for 1 h. For replicate reactions (typically 3 replicates per sample), we pooled pellets and resuspended them in 60 μL 10 mM Tris-HCl (pH 8.0). We stored DNA at 4°C or proceeded directly to size selection.

BluePippin size selection

We size-selected Cas9-digested DNA on the BluePippin system (Sage Science) using the "*0.75DF 3–10 kb Marker S1*" cassette in 4–15 kb size range mode. Before loading, we calibrated optics, replaced cassette elution buffers with 40 μL fresh electrophoresis buffer, and sealed wells with adhesive strips. We prepared samples by mixing 60 μL DNA with 20 μL loading solution. We loaded 40 μL DNA marker S1 in lane 1 as reference lane, per sample (total volume = 80 μL), loaded 40 μL of each DNA sample across two wells. Each cassette accommodated two samples across four wells. We ran the program (∼5 h), pooled the eluted fractions (∼80 μL per sample), and quantified DNA with a Qubit fluorometer.

Library preparation and sequencing

For barcode ligation, we added 3 μL of a unique barcode from SQK-NBD114.24 kit to 47 μL of volume normalized DNA sample, then added 50 μL of Blunt/TA Ligase Master Mix (NEB) and mixed them for each sample. We incubated the mixture at RT for 10 min, spun down, and then put on a magnet. We washed the beads with 200 μL of freshly prepared 70% ethanol, without disturbing the pellet, twice, and allowed to dry for 30–60 s. We then resuspended the remaining pellet in 16 μL of nuclease-free water when processing four or fewer samples. For more than four samples, pellets from different samples were pooled and resuspended together in a total volume of 66 μL of nuclease-free water, followed by incubation at room temperature for 10 min. We placed the reaction on a magnet and transferred 65 μL of supernatant into a clean 1.5 mL Eppendorf DNA LoBind tube. We quantified samples using a Qubit fluorometer, together with the Qubit dsDNA HS assay kit (Thermo Fisher Scientific). We ligated adapters by first adding 20 μL NEBNext Quick Ligation Buffer (NEB#E6056S), 10 μL NEBNext Quick T4 DNA ligase (NEB#E6056S), and 5 μL Native Adapter (NA) at room temperature in a separate 1.5 mL Eppendorf DNA LoBind Tube. We mixed the ligation reaction thoroughly. We firstly added 20 μL of the adapter ligation reaction to the pooled native barcode-ligated samples and mixed. Immediately after mixing, we added the remaining 15 μL of the adapter ligation mix to the native barcode-ligated sample, to yield a 100 μL ligation

mix. We incubated the reaction for 10 min at room temperature. Then, we added 1 volume (100 μL) of TE (pH 8.0) to the ligation mix, followed by adding 0.4x volume (80 μL) of AMPure XP Beads. We incubated the sample for 10 min at room temperature, placed back on the magnet, and removed the supernatant. We washed the beads with 250 μL Long Fragment Buffer (LFB) twice and then air-dried for ~30 s. We eluted the library off the beads in 14 μL Elution Buffer (EB). We mixed 13 μL of the library with 37.5 μL sequencing buffer (SQB) and 25.5 μL loading beads (LIB) and loaded onto the MinION flow cell. We performed sequencing for 48 h using MinKNOW default settings.

STR tract length and DNA methylation analysis

We process raw sequencing data (pod5 files) for base-calling, alignment and sample demultiplexing using Dorado (version 0.7.3). To ensure accessibility and reproducibility, we implemented a streamlined computational workflow in snakemake (https://github.com/chuanbinhub/MASTRseq). The pipeline uses demultiplexed, sorted, and indexed BAM files together with unaligned BAM files as input, and generates visualized STR counts, promoter CpG methylation profiles, and STR loci CpG methylation. Based on STR types and loci (*FMR1*, *HTT* and *C9orf72)*, users can specify custom parameters. The workflow automatically produces STR tract length and DNA methylation plots.

QUANTIFICATION AND STATISTICAL ANALYSIS

We developed a python-based tool to quantify STR lengths from aligned and demultiplexed BAM files generated from R10.4.1 flow cell. The pipeline applies the following steps:1). Filtering: reads are retained if they (i) exceed the expected length of the Cas9-cleaved normal allele, (ii) align to the target region, and (iii) contain a user-defined conserved anchor sequence near the STR. 2). Repeat counting only if each read passing filters: (i) at least 4x STRs in tandem, (ii) exact matches to the STR motif are counted. Optionally, reads with only 2–3 repeat units are retained if expected in wild-type alleles (e.g., *C9orf72* GGCCCC). Specific filtering for R9.4.1 data: due to the lower read quality of R9.4.1 flow cells compared to R10.4.1, additional filtering is applied: (i) The two detected blocks of at least 4 tandem repeats must not be separated by more than a user-defined threshold of non-STR bases, (ii) Reads containing large consecutive "TAA", "TG", or "CAA", likely from sequencing errors, are excluded. Finally, we quantified the number of repeat units in the continuous STR tract and plotted representative nanopore long-reads for visualization.

For quantification of DNA methylation, when using data produced with the R9.4 flow cell, we employed distinct methods for calling DNA methylation at the *FMR1* promoter region (500 bp upstream to TSS) and FMR1 CGG tract, respectively. We utilized nanopolish (version 0.13.2) to identify CpG methylation within the 500 bp FMR1 promoter (hg38, chrX: 147911419–147911919). Given nanopolish's limitation in calling DNA methylation over a variable number of CGG triplets, we used STRique (version 0.4.2) to call methylation of CpGs over the CGG tract in normal-length, PM, and ML iPSCs. For the quantification of CpG methylation at the *FMR1* promoter using nanopolish, we indexed the fast5 files using the 'Index' command. We applied the 'call-methylation' command to detect CpG methylation within the specified windows of GRCh38-hg38 for *FMR1* loci (chrX:147,902,117–147,960,927). Methylated and unmethylated states were computed based on log2 likelihood values, with >0.1 considered as methylated. The proportion of methylated CpGs is computed for each single molecule read. The resulting proportions are visualized as kernel distribution estimation plots using the 'density' function in R.

To ascertain CpG methylation specifically at the CGG STR in the 5′UTR of *FMR1* using STRique, we processed the fast5 files with the 'index' command. We computed methylation status and CGG counts using the 'count' command with the respective models 'r9_4_450bps_mCpG.model' and 'r9_4_450bps.model'. We retained reads with prefix and suffix scores greater than 4 for further analyses. Those with scores <4 exhibit low-quality mapping to the upstream and downstream regions of the CGG tract. Subsequently, we calculated the number of methylated CGGs and presented them as jitter plots using the R ggplot2 package.

For data produced recently using the upgraded R10.4 flow cell, the leading nanopolish and STRique algorithms are not compatible with the reads. We developed a MASTR-seq snakemake pipeline to directly analyze Dorado basecalling results. Dorado outputs a modified bam file where the 'MM' tag of each read stores which cytosines in the read were detected as a CpG and the 'ML' tag stores methylation likelihood at every detected CpG. For analysis of CpG methylation around the STR, such as the gene promoter, we utilize the modkit extract command from the Oxford Nanopore modkit tool. The modkit extract command analyzes the 'MM' and 'ML' tags and the alignment information for each read in the modified bam file from Dorado and provides a text file that stores CpG methylation likelihood at each genomic locus in each read. Our code can be used to visualize CpG methylation around the STR. Modkit requires information about where each CpG aligns to the reference genome, which does not exist for mutation-length STRs that exceed the length represented in the reference genome. Therefore, for analysis of CpG methylation within the STR, we process the 'MM' and 'ML' tags in an unaligned Dorado-produced bam file in the snakemake pipeline. We trim the read sequence, 'MM' tag, and 'ML' tag to only contain information within the STR and plot the fraction of CpGs methylated within the STR tract.

On-target reads calculation

We calculate the average coverage of the *FMR1* target region, which includes the 500 bp upstream, the *FMR1* CGG STR, and 500 bp downstream regions. The analysis uses the hg38 reference genome at coordinates chrX:147911551–147912610. The on-target average coverage is the average number of times a nucleotide is represented by a high-quality base.[78] On-target reads are defined

as any read that contains at least 1 bp overlap with a target region. Off-target reads contain no target-overlapping bases. We obtained the average coverage of target regions and number of on-target reads using the 'samtools' functions 'coverage' with the parameters "-A -w 42 -r chrX:147911551–147912610". We calculated 'total-aligned' reads using 'samtools' command 'view'. We calculated on-target reads percentage using the following formula:

$$on\ target\ reads\ percentage = \frac{on\ target\ reads\ number}{total\ aligned\ reads\ number}100\%$$

We calculated off-target reads percentage using the following formula:

$$off\ target\ reads\ percentage = \left(1 - \frac{on\ target\ reads\ number}{total\ aligned\ reads\ number}\right)100\%$$

CpG DNA methylation analysis at the *FMR1* promoter using nanopolish and data derived from an R9.4.1 flow cell

i. Using the Nanopolish tool, we start by indexing the fast5 files using the 'Index' command.[79]

nanopolish index -d [fast5 folder directory] [demultiplexed fastq directory]/file01.fastq

ii. Subsequently, we use the 'call-methylation' command to determine CpG methylation within the specified window 'chrX:147,902,117–147,960,927' for FMR1 target loci.

nanopolish call-methylation -t 10 -r [demultiplexed fastq directory]/file01.fastq -b [aligned_sorted bam directory]/file01.sorted.bam -g [hg38 fasta reference directory]/hg38.fa -w "chrX:147902117–147960927" > [Output_folder]/file01_CpG.tsv.

iii. We determine methylated and unmethylated states based on Log2 likelihood values, considering >0.1 as methylated.

iv. Then, we calculate the proportion of methylated 19 CpGs in each single-molecule read from each sample.

v. Lastly, we use custom scripts to calculate the proportion of methylated CpGs across 19 CpG sites within *FMR1* promoter region and use the density function in R to plot the proportion of methylated CpGs, generating Kernel Density Estimation (KDE) plots for visualizing methylation distributions across samples. We used visualization code from our recently published primary research paper.[61]

CpG DNA methylation analysis at the CGG STR using STRique and data derived from an R9.4.1 flow cell

i. We index fast5 files by using the index command in the STRique tool, linking bam files to nanopore raw file(fast5) for methylation analysis.

python3 [STRique directory]/scripts/STRique.py index –recursive [fast5 directory] > [fast5 directory]/reads.fofn

ii. We compute CGG STR methylation status by running the count command with specified models, r9_4_450bps_mCpG.model and r9_4_450bps.model.

awk '($3∼"chrX")'[24] [sam file directory]/file01.sam | python3 [STRique directory]/scripts/STRique.py count [fast5 directory]/reads.fofn [STRique directory]/models/r9_4_450bps.model [STRique directory]/config/repeat_config.tsv –mod_model [STRique directory]/models/r9_4_450bps_mCpG.model > [Output directory]/file01_STR.tsv.

Note: In the config file <repeat_config.tsv>, the target STR genome coordinates are based on hg38 reference.

iii. We filter out all reads with scores <4, which indicate low-quality alignment with flanking regions of the CGG STR tract. We only retain reads with high prefix and suffix scores (>4) for downstream visualization.

iv. Subsequently, we calculate the number of methylated CGGs and generate jitter plots using.

R software. The scripts used for filtering and visualization was used and available in our recently published research paper.[61]