LETTERS
https://doi.org/10.1038/s41587-019-0293-x
nature biotechnology

Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing

Pay Giesselmann[1,12], Björn Brändl[1,2,12], Etienne Raimondeau[3], Rebecca Bowen[3], Christian Rohrandt[4], Rashmi Tandon[2], Helene Kretzmer[1], Günter Assum[5], Christina Galonska[1], Reiner Siebert[5], Ole Ammerpohl[5], Andrew Heron[3], Susanne A. Schneider[6], Julia Ladewig[7,8,9,10], Philipp Koch[7,8,9,10], Bernhard M. Schuldt[2], James E. Graham[3], Alexander Meissner[1,11] and Franz-Josef Müller[1,2*]

Expansions of short tandem repeats are genetic variants that have been implicated in several neuropsychiatric and other disorders, but their assessment remains challenging with current polymerase-based methods[1–4]. Here we introduce a CRISPR–Cas-based enrichment strategy for nanopore sequencing combined with an algorithm for raw signal analysis. Our method, termed STRique for short tandem repeat identification, quantification and evaluation, integrates conventional sequence mapping of nanopore reads with raw signal alignment for the localization of repeat boundaries and a hidden Markov model-based repeat counting mechanism. We demonstrate the precise quantification of repeat numbers in conjunction with the determination of CpG methylation states in the repeat expansion and in adjacent regions at the single-molecule level without amplification. Our method enables the study of previously inaccessible genomic regions and their epigenetic marks.

The expansion of unstable genomic short tandem repeats (STRs) causes more than 30 Mendelian human disorders[5]. An extended GGGGCC repeat $(G_4C_2)_n$, within the *C9orf72* gene is the most frequent monogenic cause of frontotemporal dementia and amyotrophic lateral sclerosis c9FTD/ALS (c9FTD/ALS; MIM 105550)[6]. Similarly, accumulation of a CGG motif in the *FMR1* gene underlies the fragile X syndrome (FXS; MIM 300624) and is currently one of the most common identifiable genetic causes of mental retardation and autism[7]. In both prototypical repeat expansion disorders (Supplementary Fig. 1 and Supplementary Note), recent evidence has suggested pronounced inter- and intraindividual repeat variability as well as focal changes in DNA methylation to modulate the disease phenotype[8–10].

Nanopore sequencing is an evolving technology for direct sensing of up to megabase-long nucleotide sequences[11,12]. Recent development efforts have focused on increasing throughput and read length[13] and on the reliable detection of epigenetic modifications from the nanopore raw signal[14,15].

To overcome the current difficulties in characterizing expanded STRs (Supplementary Fig. 2 and Supplementary Note), we focused on three areas: (1) optimization of nanopore sequencing and signal processing to capture STRs; (2) development and implementation of a target enrichment strategy to increase efficiency; and (3) integration of expansion measurements with CpG methylation at the single-molecule level.

To enable robust repeat analysis, we developed a general-purpose signal processing algorithm for the exact quantification of STR numbers in raw nanopore signals (STRique: short tandem repeat identification, quantification and evaluation; Fig. 1a and Supplementary Fig. 3; https://github.com/giesselmann/STRique).

To first benchmark existing repeat expansion counting methods, we constructed, verified and nanopore sequenced plasmids with several synthetic $(G_4C_2)_n$-repeat lengths[16]. Current (as of May 2019) production-grade software (Guppy v.3.0.3) developed by Oxford Nanopore Technologies (ONT) was used to translate the nanopore raw signal into base-space representations.

The analysis results revealed that the current generation of general-purpose base-calling algorithms cannot satisfactorily resolve expanded STR sequences (Supplementary Fig. 4). For our purpose, we systematically combined outputs from three ONT base-caller generations (Albacore, Flappie and Guppy) and different parameter sets with two current sequence-based STR quantification approaches (Decoy Alignment[17] and RepeatHMM[18]; Fig. 1b). Albacore performed the best, with increased window size for the decoy alignment strategy, while the high-accuracy Guppy model provided the best sequence-derived results in combination with the RepeatHMM algorithm (Supplementary Fig. 4 and 5). Notably, we observed a systematic sequence strand bias resulting in more accurate counts for the GGGGCC sequence compared to the complementary strand (GGCCCC; Supplementary Fig. 5). We conclude, that the mentioned neural network base-callers, while enabling improved single-read base quality at the genomic level[19], become unreliable for more than 32 G_4C_2 repeats.

To overcome these issues with our STRique signal analysis software (see Methods), the reads spanning an STR location are first identified by aligning the conventional base-called sequences to a reference[20]. Next, STRique maps the upstream and downstream boundaries of each repeat more precisely with a signal alignment algorithm and, as a third step, quantifies the number of repeats

[1]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Universitätsklinikum Schleswig-Holstein Campus Kiel, Zentrum für Integrative Psychiatrie gGmbH, Kiel, Germany. [3]Oxford Nanopore Technologies, Oxford, UK. [4]Kiel University of Applied Sciences, Institute for Communications Technologies and Embedded Systems, Kiel, Germany. [5]Institute for Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany. [6]Department of Neurology, Ludwig-Maximilians-Universität, Munich, Germany. [7]Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany. [8]HITBR Hector Institute for Translational Brain Research gGmbH, Mannheim, Germany. [9]German Cancer Research Center, Heidelberg, Germany. [10]Institute of Reconstructive Neurobiology, University of Bonn School of Medicine & University Hospital Bonn, Bonn, Germany. [11]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [12]These authors contributed equally: Pay Giesselmann, Björn Brändl. *e-mail: franz-josef.mueller@uksh.de

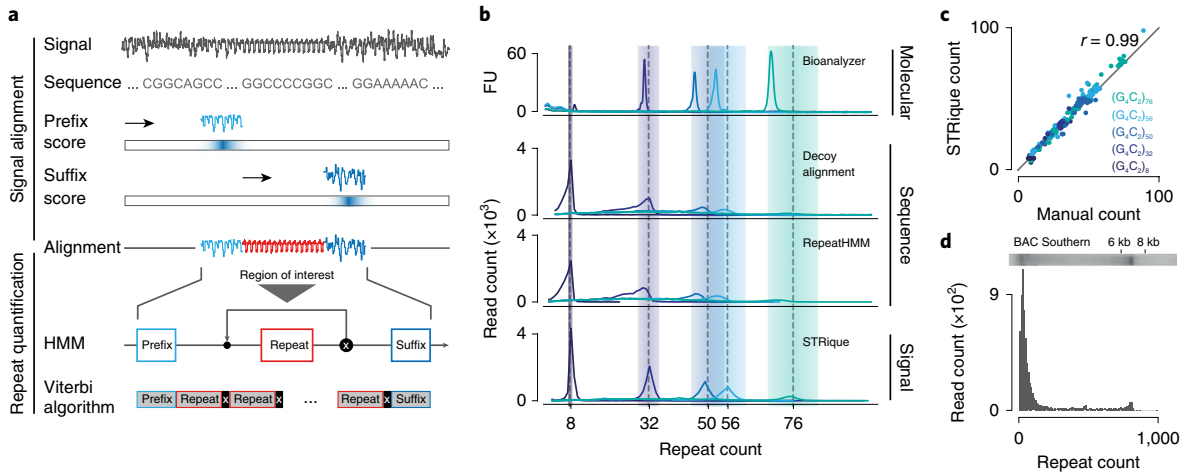

Fig. 1 | STRique: generic repeat detection pipeline on raw nanopore signals. a, Repeat quantification enabled by raw signal alignment of flanking prefix and suffix regions and HMM-based count on the signal of interest. **b**, Bioanalyzer electropherogram, decoy alignment, RepeatHMM and STRique counts of synthetic $(G_4C_2)_n$ repeats (10,000 random reads per barcode, ±10% intervals around expected repeat length). FU, fluorescence units computed from Bioanalyzer raw output. **c**, Manual confirmation of detected repeat counts in synthetic repeats ($[G_4C_2]_8$ plasmid n_{reads} = 15, $[G_4C_2]_{32}$ plasmid n_{reads} = 49, $[G_4C_2]_{50}$ plasmid n_{reads} = 45, $[G_4C_2]_{56}$ plasmid n_{reads} = 48 and $[G_4C_2]_{76}$ plasmid n_{reads} = 47; Pearson correlation). **d**, Nanopore sequencing and analysis of BAC clone 239 from a patient with c9FTD/ALS compared to a cropped corresponding lane from O'Rourke et al.[22] for the purpose of illustration.

of any given STR sequence with a hidden Markov model (HMM; Supplementary Fig. 3)[21]. Aggregated STRique repeat counts closely matched the gel electrophoresis profiles (Bioanalyzer) from our synthetic repeat constructs and were confirmed at the single-molecule level by manually counting repeat patterns in raw signal traces (Fig. 1b,c and Supplementary Figs. 4 and 6).

Previously, repeat instability had been noted in bacterial artificial chromosomes (BACs) containing expanded *C9orf72* $(G_4C_2)_n$ repeats (see Methods)[22]. Through analyzing BAC clone 239 from a patient with c9FTD/ALS ($(G_4C_2)_{\sim 800}$; ref. [22]) using STRique, we observed STR contractions in many reads and a secondary peak at 800 repeats (Fig. 1d; see Methods), whereas all evaluated sequence-space-based methods failed to mirror previously published Southern blot results (Fig. 1d and Supplementary Fig. 5)[22].

Next, to establish a baseline reference dataset, we performed nanopore sequencing of a whole-genome library using DNA obtained from patients with c9FTD/ALS, yielding a total of 29 Gb from a single MinION flow cell. Consistent with approximately tenfold genome-wide coverage, ten reads covered the *C9orf72* target region. To improve the coverage of any predetermined STR, but particularly the $(G_4C_2)_n$ region in our proof-of-concept study, we took advantage of the programmable CRISPR–Cas12a ribonucleoprotein (Cas12a RNP), which cleaves DNA via staggered double-strand breaks[23]. Cas12a RNP was first applied to selectively target DNA sequences from a patient-derived human induced pluripotent stem cell (hiPSC) line (24/5#2) adjacent to the $(G_4C_2)_n$ repeat, resulting in unique 4-bp overhangs as molecular tags amenable to ligation of a linker oligonucleotide and subsequent attachment of the nanopore sequencing adaptor (Fig. 2a, workflow I; see Methods). To further improve enrichment results we replaced the oligonucleotide–adaptor ligation step by adding Klenow fragment to fill in the Cas12a overhangs. The resulting dA-tailed DNA ends enabled even more efficient ligation of the sequencing adaptors. In this variant of the enrichment protocol, the phosphorylated 5′ ends generated by Cas-nuclease-mediated cleavage provide the molecular tag for selectively ligating the nanopore sequencing adaptors to the targeted DNA fragment (Fig. 2a, workflow II).

Additional dephosphorylation of all 5′ ends before Cas12a RNP digestion chemically protects DNA 'background' fragments from being ligated to sequencing adaptors. Consequently, only those fragments cut by Cas12a RNPs are capable of being sequenced by this procedure (Fig. 2a). As a result, we were able to obtain up to 82 reads covering the $(G_4C_2)_n$ repeat, including 40 reads for the expanded allele from a single MinION flow cell (Supplementary Fig. 7, Supplementary Table 3 and Supplementary Note). Strikingly, consistent with Southern blot results from the same cell line (Supplementary Fig. 8a,b), we found two distinct repeat expansion distributions (Fig. 2b). To further explore the general applicability of our enrichment, sequencing and signal processing protocol to other repeat expansion disorders, we tested two isogenic patient-derived cell lines (SC105iPS6 and SC105iPS7) carrying distinct *FMR1*-repeat expansions[24]. Employing a new set of *FMR1*-targeting Cas12a RNPs (Supplementary Table 1), we found two different repeat expansion distributions as predicted by Southern blot analysis (Fig. 2c and Supplementary Fig. 8c,d).

Because other CRISPR–Cas nucleases also generate phosphorylated 5′ ends after DNA cleavage[25], we explored whether nucleases such as Cas9 might enable additional improvements of the enrichment results. Therefore, we prepared libraries in parallel with Cas12a RNPs and Cas9 RNPs, targeting both the *FMR1* and *C9orf72* regions. Remarkably, Cas9 targeting resulted in an additional increase in sequencing depth of one order of magnitude for both targeted regions, concomitant with a notable reduction in off-target reads (Fig. 2d and Supplementary Fig. 7c). To understand whether the number of reads on target could be further improved by exposing the same Cas12a or Cas9 enrichment library to an increased number of pores, we subjected equimolar aliquots from the pooled library preparations in Fig. 2b to nanopore sequencing on PromethION flow cells, which contain on average six times as many nanopores. However, we did not observe a gain in reads on target with the larger flow cells (Supplementary Fig. 7c).

Changes in DNA methylation at the *C9orf72* and *FMR1* loci have been correlated with STR expansion status and patient characteristics in both disorders, but have not been quantified at the single-molecule level so far[10,26]. Therefore, we integrated single-read CpG methylation analysis of regions adjacent to both STRs using Nanopolish[14] with our STRique results (Fig. 3a). We found that, in the 24/5#2 line, all reads with STR expansions of >750 repeats showed a significantly increased methylation level at the promoter CpG island (CGI). In contrast, all wild-type reads and those with

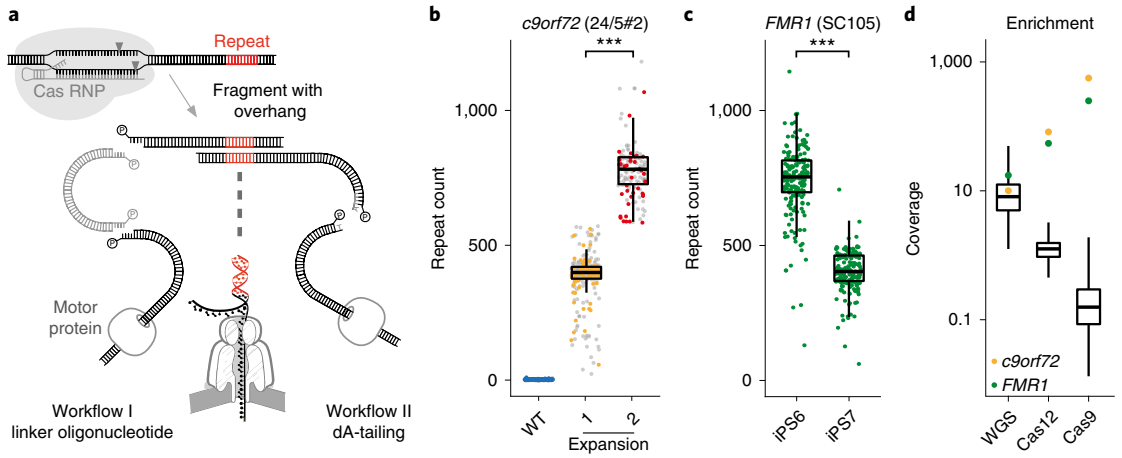

Fig. 2 | Targeted enrichment and nanopore sequencing with CRISPR–Cas. a, Illustration of the CRISPR–Cas target enrichment procedure. **b**, Repeat quantification of sample 24/5#2 at the *C9orf72* locus, revealing two distinct repeat bands of ~450 and ~750 $(G_4C_2)_n$ repeats (n = 1,810, 738 and 363 evaluated reads with a difference in repeat length of 392 (95% confidence interval (CI): 383 to 400), $P < 2.2 \times 10^{-16}$). Colored points indicate reads used in Fig. 3b. WT, wild type. **c**, Repeat quantification of the SC105iPS6 and SC105iPS7 samples at the *FMR1* locus (n = 174 and 168 evaluated reads with a difference in repeat length of −343 (95% CI: −361 to −325), $P < 2.2 \times 10^{-16}$). P values in **b** and **c** were obtained by a two-sided Wilcoxon rank-sum test; ***$P < 0.001$. Data are presented as boxplots (centerline, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range). **d**, Mean coverage on target per MinION flow cell (FAK68900, FAK67802 and FAK67994) compared to genome-wide means of 100,000 tiles for whole-genome sequencing (WGS) and Cas12 and Cas9 enrichment (boxplots for tiles (n = 30,971) are as indicated above; outliers not shown).

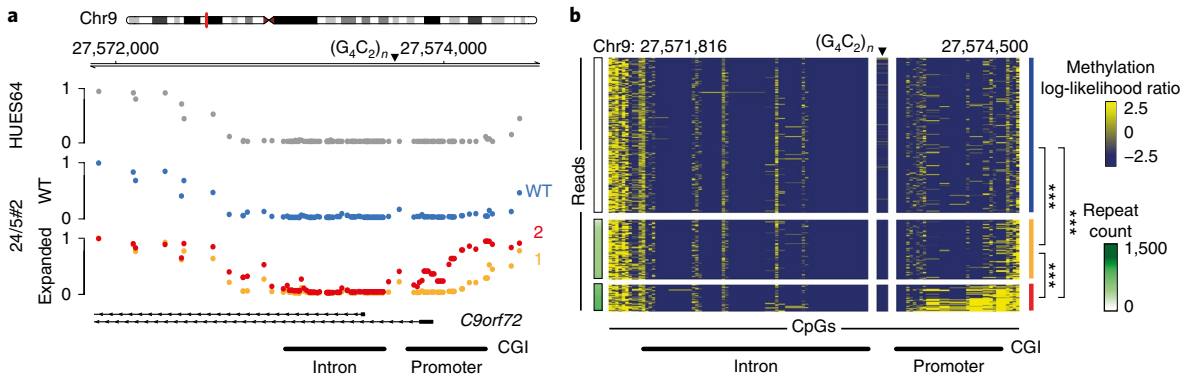

Fig. 3 | Methylation state analyses at the single-read level. a, *C9orf72* methylation status in HUES64 as measured by whole-genome bisulfite sequencing. This is a technique used to determine methylation across the genome. The wild-type (blue) allele and expanded (ex; orange) alleles (with ~450 and ~750 $(G_4C_2)_n$ repeats (red), respectively) are shown for patient 24/5#2, as measured by nanopore sequencing. **b**, Single-read nanopore methylation of *C9orf72* covering reads from the minus strand (n = 259, 100 and 43 rows per block) sorted by detected repeat length (rows, single read; columns, single CpGs). CpGs with log$P$ ratio > 2.5 are considered methylated, while those with log$P$ ratio < −2.5 are considered unmethylated. The median methylation difference (95% CI) and P value (determined by two-sided Wilcoxon rank-sum test on mean promoter CGI methylation) for comparisons were as follows: WT–ex450: 3.9×10^{-5} (4.8×10^{-6} to 3.4×10^{-2}), $P = 5.3 \times 10^{-9}$; WT–ex750: 0.56 (0.46–0.64), $P < 2.2 \times 10^{-16}$; ex450–ex750: 0.53 (0.40–0.64), $P < 2.2 \times 10^{-16}$; ***$P < 0.001$.

~450 repeats were not (or were only partially) methylated (two-sided Wilcoxon rank-sum test, $P < 0.001$; Fig. 3b, Supplementary Fig. 9 and Supplementary Note).

Additionally, in patients with c9FTD/ALS, pervasive CpG methylation of the $(G_4C_2)_n$ repeat itself has been reported[27]. Assessed with a strictly qualitative assay, the expanded STR was reported to be methylated in the majority of cases examined[27]. A similar observation has been directly implicated in the pathogenesis of FXS, where CGG-repeat expansion at the *FMR1* locus beyond a threshold of >200 repeats, leads in most cases to silencing of the entire *FMR1* gene through CpG methylation[28].

Owing to the intrinsic heterogeneity in STR length, reference-genome-based methods, such as Nanopolish[14], cannot be used to determine CpG methylation on the repeat expansion itself. To detect 5-methylcytosine (5mC) modifications on STRs, we extended STRique by employing a parallel HMM with unmodified and 5mC paths. This single-read analysis returns a methylation state for each tandem repeat, which can then be summarized into the mean repeat methylation level over the whole repetitive sequence.

When applying methylation-aware STRique, all expanded *FMR1* STRs in nanopore reads from patient SC105 were found to be highly methylated (Supplementary Fig. 10a), consistent with previous analyses[29] and our Southern blot results (Supplementary Fig. 8). We next evaluated this approach on plasmids (Addgene, cat. no. 63089) containing n = 76 synthetic G_4C_2 repeats and n = 99 CGG repeats that were covalently modified with the methyltransferase M.SssI (Supplementary Fig. 10)[14]. In addition, we tested the algorithm on $(G_4C_2)_n$-containing reads (see Methods) from patient-derived DNA

that had been modified with M.SssI in vitro. In summary, we found that STRique determined the repeat CpG methylation state correctly in all positive and negative controls evaluated.

Surprisingly, all reads covering the *C9orf72* STR from our patient-derived samples showed little to no CpG methylation, independently of the repeat expansion length or methylation status of the promoter CGI (Fig. 3b).

Our results demonstrate the precise and multilayered molecular characterization of pathological STR expansions. We increased the enrichment for regions of interest on the background of the human genome by approximately two orders of magnitude without any target amplification by using selective, multiplexed CRISPR–Cas nuclease-based chemical tagging of DNA fragments. Notably, our method does not require any additional instruments in contrast to other previously reported enrichment strategies[30], and enables reporting of the DNA methylation status of the same alleles. The CRISPR–Cas nuclease-based-target enrichment and STRique can be rapidly adapted to any other genomic region of interest, ensuring broad applicability to overcome challenges associated with the single-molecule analysis. This allows for immediate integration of genetic and epigenetic signals associated with unstable repeat expansions or any other currently unsequenceable genomic regions in human health and disease. This type of analysis will improve diagnostic workflows with regard to the accuracy and resolution at which unstable repeat expansions can be characterized, while enabling efforts to gain mechanistic insights into the effects of differentiation, aging and future therapeutic agents on STR expansions and their associated DNA methylation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0293-x.

Received: 20 October 2018; Accepted: 18 September 2019;
Published online: 18 November 2019

References

1. DeJesus-Hernandez, M. et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of *C9ORF72* causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
2. Renton, A. E. et al. A hexanucleotide repeat expansion in *C9ORF72* is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
3. Crook, A. et al. The *C9orf72* hexanucleotide repeat expansion presents a challenge for testing laboratories and genetic counseling. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **20**, 310–316 (2019).
4. Klepek, H., Goutman, S. A., Quick, A., Kolb, S. J. & Roggenbuck, J. Variable reporting of *C9orf72* and a high rate of uncertain results in ALS genetic testing. *Neurol. Genet.* **5**, e301 (2019).
5. Gatchel, J. R. & Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.* **6**, 743–755 (2005).
6. Paulson, H. Repeat expansion diseases. *Handb. Clin. Neurol.* **147**, 105–123 (2018).
7. Verkerk, A. J. et al. Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
8. van Blitterswijk, M. et al. Association between repeat sizes and clinical and pathological characteristics in carriers of *C9ORF72* repeat expansions (Xpansize-72): a cross-sectional cohort study. *Lancet Neurol.* **12**, 978–988 (2013).
9. Xi, Z. et al. Hypermethylation of the CpG island near the G_4C_2 repeat in ALS with a *C9orf72* expansion. *Am. J. Hum. Genet.* **92**, 981–989 (2013).
10. Russ, J. et al. Hypermethylation of repeat expanded *C9orf72* is a clinical and molecular disease modifier. *Acta Neuropathol.* **129**, 39–52 (2015).
11. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
12. Brown, C. G. & Clarke, J. Nanopore development at Oxford Nanopore. *Nat. Biotechnol.* **34**, 810–811 (2016).
13. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
14. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
15. Rand, A. C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).
16. Mizielinska, S. et al. *C9orf72* repeat expansions cause neurodegeneration in *Drosophila* through arginine-rich proteins. *Science* **345**, 1192–1194 (2014).
17. Dashnow, H. et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).
18. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the 'unsequenceable' genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.* **9**, 65 (2017).
19. Wick, R.R., Judd, L.M. & Holt, K.E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
20. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **3**, 321 (2018).
21. Schreiber, J. & Karplus, K. Analysis of nanopore data using hidden Markov models. *Bioinformatics* **31**, 1897–1903 (2015).
22. O'Rourke, J. G. et al. *C9orf72* BAC transgenic mice display typical pathologic features of ALS/FTD. *Neuron* **88**, 892–901 (2015).
23. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system. *Cell* **163**, 759–771 (2015).
24. Boland, M. J. et al. Molecular analyses of neurogenic defects in a human pluripotent stem cell model of fragile X syndrome. *Brain* **140**, 582–598 (2017).
25. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
26. Hornstra, L. K., Nelson, D. L., Warren, S. T. & Yang, T. P. High resolution methylation analysis of the *FMR1* gene trinucleotide repeat region in fragile X syndrome. *Hum. Mol. Genet.* **2**, 1659–1665 (1993).
27. Xi, Z. et al. The *C9orf72* repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathol.* **129**, 715–727 (2015).
28. Lyons, J. I., Kerr, G. R. & Mueller, P. W. Fragile X syndrome: scientific background and screening technologies. *J. Mol. Diagn.* **17**, 463–471 (2015).
29. Hansen, R. S., Gartler, S. M., Scott, C. R., Chen, S.-H. & Laird, C. M. Methylation analysis of CGG sites in the CpG island of the human *FMR1* gene. *Hum. Mol. Genet.* **1**, 571–578 (1992).
30. Gabrieli, T. et al. Selective nanopore sequencing of human *BRCA1* by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Res.* **46**, e87 (2018).

Methods

Raw nanopore signal processing. We proposed a raw nanopore signal-based pipeline to detect arbitrary repeat patterns on the basis of signal alignment and HMM-driven quantification. We adapted the SeqAn2 C++ library[31] (https://github.com/seqan/seqan) to support semiglobal alignment of generic signals using the score function:

$$s_{i,j} = max\begin{cases} c - |x_i - y_i| \\ -c \end{cases}$$

The distance score in the dynamic programming matrix of two signal values x_i and y_i is computed as their absolute difference subtracted from a constant offset to map more similar values to a positive score and less similar values to a negative score. The negative score is capped at a constant threshold. In contrast to the regular dynamic time warping algorithm, we applied affine gap costs without penalizing the first and last gaps of the alignment (semiglobal), allowing us to search for patterns of interest in long nanopore raw signals (Fig. 1a and Supplementary Fig. 3a,b).

Given prior knowledge of the prefix and suffix sequences of the targeted repeat, we simulated these sequences according to the pore model and used the signal alignment to map the prefix and suffix in the signal space. A compound profile HMM of prefix, repeat and suffix signal states quantified the repeat by counting the iterations in the repeat part of the model after computing the Viterbi path of the entire model given the observed signal (Supplementary Fig. 3a).

The compound HMM was built of generic profile HMM blocks, similarly to the architecture proposed by Schreiber et al.[21]. For a given profile sequence, the expected nanopore signal values were extracted from a pore model. The emission probabilities of the matched states were expressed by normal distributions around these values. Insertion states, with a uniform distribution between the model minimum and maximum, provided compensation for the intermediate state and noise measurements. Silent deletion states enabled the model to avoid states without observations. In addition to repeat count, STRique provided the alignment scores of the prefix and suffix signal. In this study we used a threshold of 3.8 for human samples and 4.0 for plasmid and BAC samples to discard low-quality counts.

MinION base calling and alignment. All MinION runs were processed using the Nanopype (v.0.7.0) pipeline (https://github.com/giesselmann/nanopype)[32]. The included base-caller versions were Albacore (v.2.3.3), Flappie (master, 9ef4edf) and Guppy (v.3.0.3). Quality filtering was disabled for any base calling. Alignments were made against human genome hg19, using Minimap2 (v.2.14) to identify reads spanning any region of interest and to determine strand orientation[20].

Repeat simulation with Nanopore SimulatION. For development and validation of the in silico experimental setup, repeats were simulated with the Nanopore SimulatION tool (https://github.com/crohrandt/nanopore_simulation)[33]. This tool takes a reference genome fasta file and a configuration (.sic file) from a previous real nanopore experiment as input for a realistic raw signal simulation of nanopore sequencing data. The configuration may be modified within an .ini file at the time of simulation. The parameters included read-length distribution, signal-capture characteristics and experiment metadata. Furthermore, the .ini file is a means of controlling simulation parameters, such as the induced signal error, the simulation of only specific sequences in full length or embedding a ground-truth sequence into the output. For simulation metrics, a predefined model of all hexanucleotides was used. Nanopore SimulatION provides fast5 files that are compatible with the standard software pipeline used for base calling.

RepeatHMM repeat detection. RepeatHMM was used in the first instance to quantify the synthetic $(G_4C_2)_n$ repeats in plasmids[18]. The software, initially validated with data only from trinucleotide (Machado-Joseph disease; MJD; MIM 109150) and pentanucleotide (Spinocerebellar ataxia 10; SCA10; MIM 603516) tandem repeat expansions, was extended by user-defined repeat parameters in this study. Using this feature, it was possible to detect the G_4C_2 hexanucleotide. Furthermore, the software was modified to also report a per-read repeat count, as it originally only provided a distribution over the whole number of reads passed. The modified version is available at https://github.com/giesselmann/RepeatHMM.

Alignment-based repeat detection. We next implemented a naive repeat-detection approach by base calling reads and aligning them against a set of decoy references, with each possible repeat length in a reasonable range, inspired by the STRetch method, proposed by Dashnow et al.[17]. The repeat count was obtained from the best matching reference without allowing multiple alignments.

We noted that RepeatHMM was only developed and validated for trinucleotide (MJD) and pentanucleotide (SCA10) repeat expansions, yet its parameter and output set could be adapted to also accommodate hexanucleotide-repeat expansions. As with the synthetic repeat sequences, STR decoy alignment assigns, in general, lower repeat counts to most sequences in comparison to STRique, with the divergence becoming more pronounced when repeat numbers increase from ~250 to up to 800 (Supplementary Fig. 5). Interestingly, a subgroup of repeat expansion counts remain in agreement between STR decoy alignment and STRique.

Notably, only STRique replicated the peak at around ~800 repeats that mirrored the expected repeat expansion maximum seen in Southern blots from the patient with c9FTD/ALS and the BAC derived from the patient used for this study (see for comparison Fig. 1s, lane 24 in O'Rourke et al.)[22].

Methylation detection on expanded reads. Methylation tracks were generated by masking the repeat signal section of expanded reads in the raw fast5 signal (script fast5Masker.py), repeating base calling and alignment with Minimap2 (v.2.14) and finally using Nanopolish (v.0.11.0) for methylation calling, each with default parameters[14]. The raw nanopore methylation log-likelihood ratio per CpG was interpreted as methylated for values >2.5 and unmethylated for values <−2.5. Intermediate values were used in the single-read heat maps but discarded in the region's methylation tracks.

The methylation state of the repeat expansion could not be detected using an established reference-genome-based workflow of base calling, alignment and Nanopolish. Because the observed heterogeneity in repeat length required a non-reference-based CpG methylation detection algorithm for the repetitive segment of each read, we extended the STRique package to evaluate multiple pore models on the repeat signal in parallel. By specifying a base and a modification model for 5mC, the repeat part of the compound HMM was duplicated, allowing the model to switch between a native and modified sequence for each repeat iteration. The output was a character string of zeros and ones, allowing the computation of a mean repeat modification level per read.

Repeat methylation detection was evaluated on untreated or M.SssI-treated plasmid DNA containing a synthetic repeat expansion for both GGGGCC in the *C9orf72* context and CGG repeats in the *FMR1* context (Supplementary Fig. 10). For the *FMR1* repeat expansion, the surrounding region and the repeat were highly methylated in the samples from SC105iPS6 and SC105iPS7 (ref. [24]). While the promoter CGI of the *C9orf72* samples showed increased methylation in Nanopolish analyses, we could not find any evidence for 5mC modifications on the repeat expansion with our STRique methylation-calling method. We then tested the ability of STRique to detect 5mC in expanded G_4C_2 repeats by sequencing an M.SssI-treated library from our c9FTD/ALS hiPSC line 24/5#2 and detected, as expected, both wild-type and expanded repeats that were fully methylated (Supplementary Fig. 10c).

The detection of 5mC from raw nanopore data depends on the signal difference in the CpG *k*-mer context[14]. For the analyzed repeat sequences, we computed the absolute signal difference per strand and observed for both repeats, stronger and thus more reliable methylation calls on the minus strand. The expected ionic current differences were as follows: G_4C_2 repeat plus strand, 2.75 pA; minus strand, 6.88 pA; CGG repeat plus strand, 5.62 pA; minus strand, 11.71 pA. We therefore only report repeat methylation readouts from the minus strand in this study.

Bioinformatics. Plots were generated with R[34] (v.3.6.0) using GViz (v.1.28.0) for regional methylation tracks and Complex Heat maps[35] (v.2.1.0) with Circlize[36] (v.0.4.6) for single-read methylation heat maps.

Preparation of high-molecular-weight DNA from cultured cells derived from patients with c9FTD/ALS. High-molecular-weight (HMW) DNA was prepared with a modified phenol–chloroform extraction method[37]. Briefly, 2×10^7 undifferentiated hiPSCs were detached with TrypLE Select (Thermo Fisher Scientific) for 3 min at 37 °C. The enzymatic reaction was stopped with DMEM/F-12 (Thermo Fisher Scientific), and the cell suspension was centrifuged for 3 min at 260*g*. Supernatant was discarded, and cells were resuspended in 100 µl of 1× PBS. Cells were lysed by adding 10 ml of TLB solution composed of 10 mM Tris-Cl (pH 8), 25 mM EDTA (pH 8), 0.5 % SDS (wt/vol) and 20 µg ml^{-1} RNase A (Qiagen) for 1 h at 37 °C. Proteins were subsequently digested at 50 °C for 3 h using 50 µl of proteinase K (>600 mAU ml^{-1}; Qiagen). The viscous solution was transferred into a 50-ml Falcon tube containing 5 g of phase-lock gel (High Vacuum Grease, Dow Corning), and 10 ml of ultrapure saturated phenol (Thermo Fisher Scientific) was added. Samples were placed onto a rotator at 40 r.p.m. for 10 min until a fine emulsion had formed, and phase separation was performed by centrifugation at 2,800*g* for 10 min. The aqueous phase was carefully poured into a fresh 50-ml Falcon tube containing 5 g of phase-lock gel followed by a second phase separation using 5 ml of ultrapure saturated phenol and 5 ml of chloroform (Merck). Samples were mixed and centrifuged as described above. The aqueous phase was poured into a fresh 50-ml Falcon tube, and the genomic DNA was precipitated using 4 ml of 5 M ammonium acetate together with 30 ml of ice-cold ethanol (absolute) and gently inverted ten times. Precipitated DNA was spooled out of solution using a glass rod and carefully submerged in 80% ethanol. Washed HMW DNA was transferred into a 1.5-ml DNA LoBind tube (Eppendorf) containing 1 ml of 80% ethanol and centrifuged at 16,000*g* for 10 min. Supernatant was removed and the DNA pellet was dried at 40 °C for 5–10 min. Rehydration of DNA was performed at 50 °C for 1–2 h using 100 µl of 10 mM Tris-HCl (pH 8). Samples were stored at 4 °C for 2 d and DNA was further homogenized on a rotator at 37 °C and 20 r.p.m. overnight.

Selective ligation of genomic fragments with CRISPR–Cas12a RNP-mediated DNA cleavage. We designed CRISPR–Cas12a crRNAs (IDT) targeting genomic regions adjacent to the *C9orf72* $(G_4C_2)_n$ and *FMR1* CGG repeats with the ChopChop online tool (http://chopchop.cbu.uib.no; Supplementary Table 1)[38,39], creating staggered 4-bp overhangs at the 5′ end. We note that Cas12a was introduced into the literature as Cpf1 (CRISPR from Prevotella and Franciscella1)[23] but is now more frequently referred to as Cas12a[23,40]. Next we designed a unique single-stranded DNA (ssDNA) oligomer ('bottom strand'; IDT) in parts complementary to a universal ssDNA barcode (NB01) referred to as the 'top strand', on the basis of the ONT Native Barcoding Kit 1D (EXP-NBD103). Oligomers were resuspended in 10 mM nuclease-free Tris-Cl (pH 8) to a final concentration of 100 µM and subsequently used for barcode annealing. Briefly 44 µl of top strand was combined with 40 µl of the respective bottom strand and 16 µl of nuclease-free duplex buffer (IDT). Samples were heated for 2 min at 95 °C in a thermocycler and the mixture was cooled to 25 °C over 70 min with a cooling rate of approximately −1 °C per minute. Successfully annealed barcodes served as a linker that provided complementary overhangs to the respective Cas12a cut site and the BAM 1D adaptor, thus allowing a sticky-end ligation.

Preparation of programmed Cas12a nucleoprotein complex (Cas12a RNP) was performed by combining EnGen Lba Cas12a (NEB) with IDT Alt-R Cpf1 crRNA, which consists of a target-specific 21-bp protospacer domain and a constant 20-bp loop domain. Briefly, 2 nmol of lyophilized crRNA was reconstituted in 20 µl of 10 mM nuclease-free TE buffer (pH 7.5) to obtain a 100 µM solution. Next, crRNA was pooled and adjusted to 10 µM using nuclease-free water. Formation of secondary RNA structure was performed in 1× NEB CutSmart buffer with a final crRNA concentration of 500 nM. The sample was heated for 6 min at 90 °C using a thermomixer and snap-cooled on wet ice before adding 0.25 µl of Cas12a enzyme (100 µM stock) to obtain a final concentration of 500 nM. Cas12a RNPs were formed for 20 min at room temperature (RT), and the assembled complex was stored on ice until use.

HMW DNA was digested with 10 µl of FastDigest BamHI (Thermo Fisher Scientific) in 1× NEB CutSmart buffer for 12 h at 37 °C and subsequently heat-inactivated at 80 °C for 5 min. Then, the 5′ ends were dephosphorylated for 30 min at 37 °C using 4 µl of quick calf intestinal alkaline phosphatase (CIP; NEB) followed by a heat-inactivation step for 5 min at 80 °C. The samples were equilibrated to RT before Cas12a RNP incubation. To allow sufficient Cas12a cleavage, dephosphorylated DNA was supplemented with 10 µl of Cas12a RNPs and incubated for 1 h at 37 °C. The cleavage reaction was continued at 4 °C, overnight.

The next day, the cleaved DNA samples were carefully cleaned and concentrated by using a 1 volume of Agencourt AMPure XP beads (Beckman Coulter). DNA binding to AMPure XP beads was carried out for 10 min at RT followed by quick-pulse centrifugation. Next, beads were immobilized on a DynaMag Spin magnet (Thermo Fisher Scientific) for 5 min and samples were washed with 80% ethanol according to the manufacturer's protocol. Residual ethanol was completely removed by pipetting, and beads were air-dried for 60 s. AMPure XP beads were carefully resuspended in 40 µl of prewarmed (37 °C) 0.1× TE buffer using large-bore tips, followed by a 10-min incubation step at 37 °C. The sample tube was placed on the magnet and the eluted DNA was used for barcode adaptor ligation.

Annealed barcode adaptors (see above) were pooled and adjusted to a 1 µM working concentration. A total of 0.2 µl of pooled barcode mix was added to the cleaved DNA. The samples were incubated at 65 °C for 5 min in a thermocycler and equilibrated to RT before adding 60 µl of NEB Blunt/TA Ligase Master Mix (NEB) followed by an additional incubation for 30 min at RT. Next, 20 µl of BAM 1D sequencing adaptor from the ONT Native Barcoding Kit 1D (EXP-NBD103) was added to the mix described above and ligation was subsequently performed at RT for 30 min.

The sequencing library was purified again by adding 0.4 volumes of Agencourt AMPure XP beads for 10 min at RT. Excess BAM 1D was washed away with adaptor bead binding solution according to the manufacturer's protocol. The final sequencing library was eluted from the magnetic beads using 18.5 µl of prewarmed elution buffer (ELB) from the ONT Ligation Sequencing Kit 1D (SQK-LSK108) at 37 °C for 20 min. The sample tube was placed on the magnet, and the eluted library was transferred to a fresh DNA LoBind tube at 4 °C. Samples were then quantified using a Qubit fluorometer, together with the Qubit dsDNA BR assay kit (Thermo Fisher Scientific), and a total of 17.5 µl of the library was combined with 35 µl of running buffer (RBF) and 22.5 µl of library loading beads from the EXP-LLB001 kit (ONT). After removal of the AMPure XP beads, an ONT SpotON Flow Cell R9.4 (FLO-MIN106) was primed and loaded following the manufacturer's instructions with no modifications. Supplementary Tables 3–6 give an overview of the flow cells, DNA samples, yields and run characteristics for this study.

Further optimization of the enrichment results was achieved with a modified Cas12a protocol that contained a dA-tailing reaction resulting in a substantially higher amount of on-target reads and took advantage of an improved sequencing kit (SQK-LSK109). Initial BamHI digestion, the use of a barcode adaptor and overnight incubation with RNP were no longer required. Instead, HMW DNA was diluted to a final concentration of 200 ng µl^{-1} and a total of 5 µg of HMW DNA was dephosporylated with 2 µl of Quick CIP in 1× CutSmart buffer (NEB) using the same parameters as described above. The assembled RNP complex was added to the dephosphorylated DNA together with 1 µl of 1 mM dNTP (NEB) mix and 0.5 µl of 10 mM dATP (NEB). The final reaction volume was adjusted to 49 µl using nuclease-free water. Briefly, 5 U of NEB Klenow Fragment (3′→5′ exo–) was added to the sample, and dA-tailing and RNP cleavage were performed in parallel while incubating the sample at 37 °C for 15 min, followed by a second incubation at 65 °C for 7 min. Instead of using the BAM 1D adaptor, the dA-tailed library was incubated with an adaptor ligation mix composed of 25 µl of ligation buffer (LNB), 10 µl of Quick T4 DNA Ligase (NEB), 10 µl of nuclease-free water and 5 µl of adaptor mix (AMX). Adaptor ligation was performed for 20 min at RT, and the mixture was subsequently diluted with 1 volume of 10 mM nuclease-free Tris-Cl (pH 8). Following ligation, the samples were incubated with 0.3 volumes of AMPure XP beads for 10 min at RT and washed twice with 250 µl of long fragment buffer (LFB) according to the manufacturer's instructions. The washed library was eluted with 16 µl of elution buffer (EB), and 1 µl was used for Qubit quantification, as described previously. Priming of the ONT SpotON Flow Cell R9.4 (FLO-MIN106) was performed according to the manufacturer's protocol with a minor modification. Before loading samples, a SpotON priming mix composed of 20 µl of sequencing buffer (SQB), 0.4 µl of sequencing tether (SQT) and 19.6 µl of nuclease-free water was added dropwise to the SpotON port. The eluted DNA library was supplemented with 25 µl of SQB together with 10 µl of loading beads (LBs) and immediately loaded, as described earlier.

Selective ligation of genomic fragments with CRISPR–Cas9 RNP-mediated DNA cleavage. Cas9-based target enrichment experiments were performed similarly to the Cas12a enrichment setup. Briefly, we designed CRISPR–Cas9 crRNAs (IDT) targeting the genomic regions adjacent to the *C9orf72* $(G_4C_2)_n$ and *FMR1* CGG repeats with the ChopChop online tool (Supplementary Table 2). Preparation of the Cas9 nucleoprotein complex (Cas9 RNP) was performed as follows. Lyophilized Cas9 crRNA XT and tracrRNA (IDT) were reconstituted with nuclease-free TE buffer (pH 8) (IDT) to yield a stock concentration of 100 µM. For Cas9 multiplexing the *FMR1* and *C9orf72* crRNAs were pooled to obtain an equimolar solution. A crRNA–tracrRNA duplex was formed by diluting the equimolar crRNA solution and tracrRNA in nuclease-free duplex buffer (IDT) to a final concentration of 10 µM. The sample was incubated for 5 min at 95 °C in a thermal cycler (Bio-Rad) and allowed to anneal at RT. A 10× master mix was generated by combining 0.8 µl of Alt-R S.p. HiFi Cas9 Nuclease V3 (IDT) with 10 µl of crRNA–tracrRNA duplex, 10 µl of 10× NEB CutSmart buffer (NEB) and 79.2 µl of nuclease-free water. Cas9 RNP was formed for 30 min at RT and 10 µl was subsequently used for incubation with 5 µg of dephoshporylated HMW DNA (see above) together with 1 µl of 10 mM dATP and 5 U of Taq polymerase (NEB). The reaction mixture was incubated for 30 min at 37 °C on a thermo block, and Cas9 enzyme was then heat-inactivated for 5 min at 72 °C. All subsequent steps were performed as described in the CRISPR–Cas12a RNP section.

For methylation calling on *C9orf72* $(G_4C_2)_n$ and *FMR1* CGG repeats from patient 24/5#2, we included a methylation step with M.SssI before library preparation. Briefly, 4 µg of HMW DNA was supplemented with 10× CutSmart buffer (1× final) and *S*-adenosylmethionine (SAM) at a final concentration of 640 µM and was subsequently treated with 20 U of M.SssI CpG Methyltransferase (NEB) at 37 °C for 3 h, followed by substitution of fresh SAM and further incubation for 4 h.

Nanopore whole-genome, plasmid and BAC sequencing. For whole-genome sequencing, we used HMW DNA for our sample from a patient with a *C9orf72* repeat (24/5#2) in combination with the ONT 1D Ligation Sequencing kit (SQK-LSK109). DNA was sheared to an average fragment size of 20 kb by centrifugation of 30 µg of DNA at 7,200 r.p.m. using a g-Tube (Covaris). Size selection on the High Pass Plus 0.75% Agarose Cassette (Sage Bioscience) was performed on the Blue Pippin instrument (Sage Bioscience) using 6 µg of fragmented DNA, following the manufacturer's recommendations. Eluted DNA was pooled and bound to 0.7 volumes of AMPure beads for 10 min at RT followed by two washing steps with 80% ethanol. Two aliquots of 3.5 µg genomic DNA were used for NEBNext FFPE Repair mix (NEB) and dA-tailing with NEBNext Ultra II End-prep enzyme mix (NEB) in a single reaction and incubated at 20 °C for 5 min. The enzymatic reaction was stopped at 65 °C for 5 min and DNA was washed using 0.7 volumes of AMPure beads. A total of 4 µg of the end-repaired library was split into two aliquots for adaptor ligation, followed by two washing steps with ONT LFB. The resulting library was eluted with ONT EB, and 600 ng of DNA was loaded onto a MinION flow cell. The sequencing library was refreshed every 22 h using a nuclease flush protocol. Briefly, 40 U of DNase I (NEB) was supplemented with 380 µl of wash solution A (EXP-WSH002) and loaded onto the flow cell. Nuclease solution was incubated for 30 min and the flow cell priming procedure was performed following the manufacturer's recommendation, with subsequent loading of 600 ng of fresh library onto the flow cell.

For sequencing our synthetic $(G_4C_2)_n$ repeat-containing plasmids, we first linearized the plasmid DNA with FastDigest NdeI (pcDNA3.1(+)) or FastDigest ScaI (pCR Script Amp-BE), heat-inactivated the restriction enzymes and cleaned up the samples using the NucleoSpin Gel and PCR cleanup kit following the manufacturer's instructions. Linearized plasmid DNA was further processed with the ONT Ligation Sequencing Kit 1D (SQK-LSK108) following the manufacturer's

instructions without modifications. In some instances, the plasmids were barcoded using adaptors included in the 1D native barcoding kit (EXP-NBD103) following the manufacturer's instructions without modifications before the 1D ligation sequencing kit was applied. For optimizing methylation calling on *FMR1* repeats, we used a control plasmid from Addgene (5′ UTR 99× CGG *FMR1*; Addgene, cat. no. 63089) that contained a stretch of 99 CGG repeats as a kind gift from N. Charlet-Berguerand (Institute of Genetics and Molecular and Cellular Biology, Illkirch, France). A bacterial stab culture was used to streak Stbl3 *Escherichia coli* onto selective LB agar plates containing 100 µg ml^{-1} ampicillin. Agar plates were incubated for 48 h at RT and individual colonies were picked and transferred to LB medium supplemented with 100 µg ml^{-1} ampicillin. Inoculated samples were incubated at RT for 20 h and 200 r.p.m. in a bacterial shaker (N-Biotek), and *FMR1* plasmid was isolated using the Plasmid Plus Midi Kit (Qiagen) following the manufacturer's recommendations. Integrity of the *FMR1* plasmid was confirmed by restriction digestion with 20 U of ScaI, 20 U of EcoRV and 20 U of SpeI (all NEB). Before the library preparation for nanopore sequencing, 2 µg of *FMR1* plasmid DNA was treated with 20 U of M.SssI CpG methyltransferase, as described earlier. Methylated as well as unmethylated plasmid DNA was linearized with 20 U of ScaI for 1 h at 37 °C, and restriction digestion was stopped at 80 °C for 20 min followed by SPRI cleanup. Purified and linearized DNA was used for preparation of the sequencing library (SQK-LSK108) together with the 1D native barcoding kit (EXP-NBD103) following the manufacturer's recommendations, and 260 ng of the final library was loaded onto a MIN-FLO106 flow cell.

DNA obtained from pCC1-BAC clone 239 (ref. [22]) was sequenced using the ONT 1D rapid sequencing kit (SQK-RAD002) following the manufacturer's instructions without modifications. All samples were supplemented with beads from the library loading bead kit (EXP-LLB001) before loading samples onto the flow cells.

Synthetic $(G_4C_2)_n$-repeat cloning, modification and visualization. For cloning $(G_4C_2)_n$ fragments carrying a defined repeat number, we used recursive directional ligation as described by Mizielinska et al.[16]. Briefly, complementary DNA oligonucleotides (Supplementary Table 7) containing three or four G_4C_2 repeats flanked at the 5′ end by BamHI and BspQI (GCTCTTCC*GGCC) recognition sites and at the 3′ end by EcoO109I (GG*GGCCT) and NotI recognition sites were annealed and ligated into the vector pCR Script Amp-BE using the BamHI and NotI sites. Inserts were excised by digestion of the vector with BspQI (LguI) and EcoO109I (Thermo Fisher Scientific), electrophoretically separated, purified from agarose gels (Qiagen MinElute Gel Extraction Kit) and subsequently re-ligated into the BspQI (LguI)-linearized pCR Script Amp-BE vector. Six cycles of recursive directional ligation gave rise to increasing insert sizes of up to 100 repeat elements. Transformations were performed using recombination-deficient Stbl3 *E. coli* (Thermo Fisher Scientific) at 30 °C for longer repeats to minimize retraction of repeats. DNA was extracted using Plasmid Mini and Maxi kits (Qiagen), following the manufacturer's instructions. Constructs were screened using standard restriction enzyme digestion and agarose gel electrophoresis (2% or 4%). Next we sequence-verified (GENterprise Genomics) plasmids with up to 32 repeats, to confirm the repeat size and lack of interruptions (data not shown). In the case of the plasmids with 7, 8, 32, 76 and 100 repeats, the appropriate fragment length was excised from pCR Script Amp-BE and re-ligated into pcDNA3.1(+) via BamHI and NotI. For exact repeat length visualization and comparison with nanopore results from the same plasmids, we followed the procedures outlined by Kwok et al. for detection of the *FMR1* repeat expansion with one relevant modification[41]. Instead of using PCR-amplified fragments from the repeat region, we used fragments cut out of our synthetic pCR Script Amp-BE $(G_4C_2)_n$ plasmids with restriction enzymes. We analyzed repeat inserts with different repeat lengths (8, 32, 50, 56 and 76), which were excised with BamHI and NotI or with NdeI and ScaI from the pCR Script Amp-BE plasmids and analyzed with a 2100 Agilent Bioanalyzer on a 1,000 DNA gel cassette following the manufacturer's instructions (Agilent). Bioanalyzer raw data were normalized and plotted using an in-house script following Agilent's analysis steps (script bioa.py).

***C9orf72* BAC expansion and DNA extraction.** A 174-kb BAC (pCC1-BAC clone 239)[22] containing a $(G_4C_2)_n$-repeat expansion from a patient with c9FTD/ALS was amplified as previously described. Briefly, transfected DH10B T1 cells (Thermo Fisher Scientific) were grown on agar plates (LB broth with agar) and in LB broth containing 12.5 ng µl^{-1} chloramphenicol at temperatures <30 °C, as higher temperatures lead to repeat contraction and/or loss of BACs. Extraction of BAC DNA was performed using the Large-Construct kit (Qiagen) including an ATP-dependent exonuclease step for sufficient removal of genomic DNA, following the manufacturer's instructions. The contraction rate of the BAC was previously reported to be high[22] at rates between 20% and 80%, depending on bacterial media and growth temperature conditions, with richer media and faster/denser growth resulting in faster contraction (S. Bell, personal communication). Therefore, the cells were kept at a relatively low density and at lower temperature (~27 °C).

Generation and culture of hiPSC lines from patients with c9FTD/ALS. hiPSCs from a patient with c9FTD/ALS were generated by transducing patient-derived fibroblasts with non-integrating viral vectors (CytoTune 1.0 iPS Sendai Reprogramming Kit; Life Technologies) expressing the reprogramming factors Oct4, Sox2, Klf4 and c-Myc. Four weeks after transduction, clones were manually picked and clonally expanded as hiPSC lines on feeder cells. Established hiPSC colonies showed the typical morphology of human pluripotent stem cells, stained positive for alkaline phosphatase (AP) and expressed the pluripotency-associated surface proteins TRA-1-60 and TRA-1-81. High-resolution SNP karyotyping was performed to exclude major karyotypic abnormalities induced by the reprogramming or culturing process (Supplementary Fig. 11a–e).

For the study, the cultures were adapted to feeder-free culture conditions (mTeSR-E8, Stem Cell Technologies and BD-Matrigel, BD Biosciences) and passaged every 3–4 d as clumps with 0.5 mM EDTA in 1× PBS[42]. Detection of AP was performed using the Blue AP Substrate Kit (Vector Laboratories) following the manufacturer's instructions. Staining against TRA-1-60 and TRA-1-81 was performed as described below. Briefly, cells were rinsed with 1× PBS followed by a paraformaldehyde (4%) fixation for 10 min at RT. Samples were washed twice with 1× PBS before incubation with primary antibodies against Tra1-60 and Tra1-81 (both diluted 1:500; Thermo Fisher Scientific) for 2 h at RT. Cells were washed with 1× PBS, incubated with AlexaFluor488-conjugated anti-mouse secondary antibody (diluted 1:1,000; Thermo Fisher Scientific) and counterstained with DAPI (Sigma) for 45 min at RT in the dark. Samples were mounted in Mowiol 4-88 mounting solution (Carl Roth) for improved long-term stability. High-resolution SNP karyotyping was performed as described previously[43].

Four hiPSC cell lines from three patients with c9FTD/ALS were screened by Southern blot analysis (see below), and line 24/5#2 (Supplementary Fig. 11a–e) was selected for further nanopore sequencing analysis, as it displayed at least two distinct maxima for the expanded allele around an estimated 350 and 800 repeats in our Southern blot analyses (Supplementary Fig. 8).

Similarly, we adapted the previously characterized and described cell lines SC105iPS6 and SC105iPS7 from a patient with FXS who had concomitant autism spectrum disorder to the same feeder-free culture and passaging conditions[24]. The cell lines were previously reported to carry approximately 380 (SC105iPS6) and 335 (SC105iPS7) CGG repeats using the AmplideX FMR1 PCR Kit (Asuragen) according to the manufacturer's instructions[24]. Our Southern blot results showed peaks at approximately 750 (SC105iPS6) and 500 (SC105iPS7) CGG repeats (Supplementary Fig. 8) in agreement with the nanopore sequencing results from the same DNA preparations. We note that marked instability of *FMR1* CGG STR expansions has previously been described in other hiPSC and human embryonic stem cell lines through in vitro culture[44].

Clinical-grade Southern blot-based determination of STR expansion size in DNA from C9FTD/ALS and *FMR1* repeat-expansion-carrying hiPSC lines. We used Southern blotting as a means to estimate the repeat expansions present in our hiPSC lines from patients (Supplementary Fig. 8). The samples were processed and analyzed using the routine clinical diagnostic workflow for FXS and c9FTD/ALS (certified according to DIN EN ISO 15189:2014) at the Department for Human Genetics at Ulm University (Ulm, Germany). The diagnosticians were blinded to the experimental nature of the hiPSC samples, and the DNA was processed in parallel with actual clinical samples.

Briefly, for Southern blot determination of repeat expansion length for the $(G_4C_2)_n$ and CGG repeats in our hiPSC lines, 10 µg of HMW DNA was digested overnight with HindIII (20 U) and XbaI (20 U) and with EcoRI (20 U) and NruI (20 U), respectively, before electrophoresis. The fragmented genomic DNA was separated on a 0.8% agarose gel for 20 min at 180 V followed by 65 V overnight (*C9orf72*) or 20 min at 180 V followed by 60 V overnight (*FMR1*). The resulting gel was imaged with an ethidium bromide fluorescent stain and a copy of the image was used for determination of a migration distance/DNA ladder standard curve.

DNA was transferred to a positively charged nylon membrane (Roche Applied Science) by capillary blotting and was baked at 80 °C for 2 h.

The hybridization probes were either a 210-bp PCR fragment corresponding to a sequence upstream of the $(G_4C_2)_n$repeat in the *C9orf72* gene or a 480-bp PCR fragment corresponding to a sequence downstream of the CGG repeat in the *FMR1* gene. Probes (100 ng per filter) were labeled with 50 mCi of [α-^{32}P] dCTP and hybridized to the filters at 71 °C overnight. After washing, X-ray films were exposed to the filters. The lengths of hybridizing fragments were calculated in relation to λ-DNA digested with BstEII. Fragments derived from wild-type *C9orf72* alleles were approximately 2.3 kb in length. Wild-type *FMR1* fragments from active, unmethylated X chromosomes were approximately 2.9 kb, while those from inactivated, methylated X chromosomes were 5.2 kb in length, as the NruI restriction enzyme used in this assay cuts only 5′-TCGCGA-3′ sequences in which the CpGs are unmethylated.

Statistics and reproducibility. Linear correlations of manual and automated repeat counts were determined by Pearson correlation (r) as indicated in the respective figures. The significance of repeat count differences was determined by two-sided Wilcoxon rank-sum tests (R: wilcox.test). The significance of differences in mean methylation levels of the promoter CGI between repeat count clusters (Fig. 3b) and between unmethylated BACs and cell line 24/5#2 (Supplementary Fig. 9c) was determined by two-sided Wilcoxon rank-sum test. P values were corrected for multiple testing according to Holm (R: p.adjust).

The experiments to generate a reference repeat count dataset (plasmids for 8, 32, 50, 56 and 76 repeats and BACs for ~800 repeats) were performed once. The characterization of patient-derived hiPSC line 24/5#2 was performed once.

Enrichment protocols were tested on different cell lines and targets, as summarized in Supplementary Fig. 12, and with different crRNA combinations, as summarized in Supplementary Tables 4–6.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing data generated in this study and utilized for the determination of *FMR1* CGG_n- and *C9orf72* $(G_4C_2)_n$-repeat expansion lengths and methylation status in plasmids, BACs and patient DNA are available in a Figshare repository with identifier 7205666.

Whole-genome sequencing data and associated uncropped Southern blot images, size-marker standard curves and ethidium bromide imaging data from patient-derived cell lines are available from the corresponding author upon reasonable request through a material transfer agreement protecting the participants' genomic privacy.

Code availability

All custom code developed for this study is under MIT license and is available at https://github.com/giesselmann/STRique.

The RepeatHMM package was forked and modified and is available at https://github.com/giesselmann/RepeatHMM.

References

31. Reinert, K. et al. The SeqAn C++ template library for efficient sequence analysis: a resource for programmers. *J. Biotechnol.* **261**, 157–168 (2017).
32. Giesselmann, P., Hetzel, S., Müller, F.-J., Meissner, A. & Kretzmer, H. Nanopype: a modular and scalable nanopore data processing pipeline. *Bioinformatics* **26**, 2204 (2019).
33. Rohrandt, C. et al. Nanopore SimulatION–a raw data simulator for nanopore sequencing. in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1–8 (IEEE, 2019).
34. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
35. Gu, Z., Eils, R. & Schlesner, M. Complex heat maps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
36. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
37. Sambrook, J. & Maniatis, T. *Molecular Cloning* (Cold Spring Harbor Laboratory Press, 1989).
38. Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44**, W272–W276 (2016).
39. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).
40. Chen, J. S. et al. CRISPR–Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).
41. Kwok, Y. K. et al. Validation of a robust PCR-based assay for quantifying fragile X CGG repeats. *Clin. Chim. Acta* **456**, 137–143 (2016).
42. Chen, G. et al. Chemically defined conditions for human iPSC derivation and culture. *Nat. Methods* **8**, 424–429 (2011).
43. Mertens, J. et al. APP processing in human pluripotent stem cell-derived neurons is resistant to NSAID-based γ-secretase modulation. *Stem Cell Rep.* **1**, 491–498 (2013).
44. Zhou, Y., Kumari, D., Sciascia, N. & Usdin, K. CGG-repeat dynamics and *FMR1* gene silencing in fragile X syndrome stem cells and stem cell-derived neurons. *Mol. Autism* **7**, 165 (2016).

Acknowledgements

We are deeply thankful for the invaluable support by the patients with c9FTD/ALS and FXS and their families who donated biomaterials for this study. The *C9orf72* BAC was generously provided by R. Baloh and S. Bell (Cedars Sinai Medical Center, Los Angeles, CA, USA). We are grateful to J. Loring and A. Zhang (Scripps Research Institute, La Jolla, CA, USA) for providing us with hiPSC lines from a patient with FXS (supported by NIH R33MH087925-03). We thank P. van Damme and W. Robberecht (Laboratory for Neurobiology; VIB-KU Leuven Center for Brain & Disease Research, Belgium) for providing the fibroblasts derived from patients with c9FTD/ALS used for reprogramming. We acknowledge the expert assistance of the technical staff of the Molecular Genetics Laboratory of the Institute of Human Genetics (Ulm, Germany). P.K. and J.L. acknowledge financial support by the Hector Stiftung II gGmbH. F.J.M. and R.T. received funding from the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy–EXC 22167-390884018. F.J.M. and B.M.S. were supported by the BMBF (PluriTest2, 13GW0128A). This work was supported by the Max Planck Society. This work was overseen and approved by the Ethics Committee of the Christian-Albrechts-University (Kiel, Germany; reference no. A 145/11). Informed consent was obtained from all donors of cells and tissues used for the generation of hiPSC lines and their subsequent genetic and epigenetic analysis. All materials were donated graciously by our patients.

Author contributions

P.G., B.M.S. and F.J.M. conceived the project. B.B. and R.T. performed cell culture as well as plasmid and BAC expansion and extraction. P.G. wrote the STRique pipeline. P.G., B.M.S., C.R. and H.K. conducted additional bioinformatic analyses. P.K. and J.L. reprogrammed the c9FTD/ALS hiPSCs from patient fibroblasts used in this study. E.R., R.B., A.H. and J.E.G. developed the Cas12a and Cas9 protocols. B.B. further developed the Cas12a and Cas9 protocols with DNA from patients with c9FTD/ALS and FXS and performed nanopore library preparation and nanopore sequencing for the results presented in this manuscript. R.T. and C.G. worked on the optimization of aspects of the enrichment protocol. G.A. and R.S. conducted diagnostic testing of the repeat expansions by Southern blot and PCR analyses. S.S., R.S., O.A. and G.A. provided clinical and diagnostic advice. P.G., B.B., B.M.S., A.M. and F.J.M. wrote the manuscript. F.J.M. oversaw the study. All authors contributed to the editing and completion of the manuscript.

Competing interests

E.R., R.B., A.H. and J.E.G. are employees of ONT. C.R. was reimbursed for travel costs for an invited talk at the Nanopore Days 2018 conference in Heidelberg (Germany) by ONT. P.G. was reimbursed for travel costs for an invited talk at the London Calling 2019 conference. ONT had no role in the study design, interpretation of results or writing of the manuscript.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0293-x.

Correspondence and requests for materials should be addressed to F.-J.M.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Franz-Josef Müller

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Nanopore sequencing data was collected on the MinION and PromethION platforms using the most recent available version of the MinKNOW software (access through the ONT community). The version per experiment is logged in the raw fast5 output and ranges from v1.4.2 (protocol v1.4.2) to v3.3.2 (protocol v4.0.5) for the MinION and v1.14.2 (protocol 0.0) to v3.3.2 (protocol v4.0.5) for the PromethION.
Data analysis	This study used previously published software: basecalling was performed with ONT Albacore (2.3.3), Flappie (master, 9ef4edf) and Guppy (v3.0.3), alignment was performed with minimap2 (v2.14) against hg19, methylation analysis was performed with minimap2 (v2.14) and nanopolish (v0.11.0). For raw nanopore signal processing SeqAn (v2.4.0) was used. All custom code developed for this study is available at: https://github.com/giesselmann/STRique and https://github.com/giesselmann/RepeatHMM

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data generated in this study and utilized for the determination of FMR1/(CGG)n- and C9orf72/(G4C2)n- repeat expansion lengths and methylation status in plasmids, BACs and patient DNA are available in a Figshare repository with identifier doi:10.6084/m9.figshare.7205666.
Whole genome sequencing data, associated uncropped Southern Blot images, associated size marker standard curves and associated ethidium bromide imaging data from patient derived cell lines are available from the corresponding author upon reasonable request through a Material Transfer Agreement protecting the participants' genomic privacy.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	5 plasmids with synthetic CCCCGG repeats (8, 32, 50, 56, 76), 1 x Bacterial Artificial Chromosome from a c9FTD/ALS patient, 1 x c9FTD/ALS hiPSC line, 2 x FXS line, whole genome bisulfite sequencing data set from 1 x hESC cell line Publicly availiable data published by Jain et. al 2018 in Nature Biotechnology from the NA 12878 cell line was used for vizualisation of the unaffected C9orf72 locus. This cell line was otherwise not used nor cultured in our laboratory for this study.
Data exclusions	No data was excluded from the analysis.
Replication	during Cas12a-RNP protocol optimization, several experiments (crRNA combinations) were repeated. For a detailed list of all flow cells / Cas12a-RNP combinations/ conditions see Suppl. Table 5 All replication experiments were successful.
Randomization	No randomization on the subject/sample level was attempted, as the genetic properties of each subject were known a priory and the raw data generated from the samples/patient cells were used to develop a novel STR analysis algorithm. Instead, appropiate controls and orthogonal analysis methods- where applicable - were used to control for potential biases, e.g. the use of plasmids with variable STR lengths in combination of several other base calling and repeat counting methods.
Blinding	The DNA samples used for the Southern blot analysis were submitted to a diagnostic laboratory at the time of submission and analysis unaware of the experimental nature of the samples. Blinding was otherwise not relevant to this study, as nanopore raw data was generated for the development of an raw signal analysis algorithm and its function was validated with appropiate a priori defined positive and negative controls. In several instances of the plasmid experiments, barcoded samples were analyzed on the same flow cell and each barcode related to a unique experimental condition e.g. STR length or CpG methylation state, thus allowing the determination of potential biases in the algorithms tested with a ground truth that can be uniquely identified through an orthogonal method within the same experiment.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☐	☒ Human research participants

Antibodies

Antibodies used	anti TRA-1-60 (cat. # MA1-023, Thermo Fisher Scientific) , anti TRA-1-81 (cat. # 41-1100, Thermo Fisher Scientific)
Validation	Both monoclonal antibodies have been validated extensively in recent publications (Andrews et al. Hybridoma 1984 and Natunen

et al. Glycobiology 2011). Binding specificity in ICC experiments was confirmed by using negative (fibroblast) and positive (hPSC) controls (data not shown).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	c9FTD/ALS hiPSC cell lines were provided by Philipp Koch and Julia Laedewig (Life & Brain Instute, University of Bonn, Germany, now at the Zentralinstitut für seelische Gesundheit, Mannheim, Germany). FXS hiPSC cell lines were provided by Jeanne Loring and Ai Zhang (The Scripps Research Institute La Jolla).
Authentication	c9FTD/ALS cell lines were authenticated by a in house, PCR-based STR analysis by Philipp Koch and Julia Laedewig. FXS cell lines were authenticated by SNP genotype analysis (from 450k Illumina bead arrays) by Michael Boland and Jeanne Loring for the publictaion describing the cell lines in: Boland et al. Brain 2018 The cell lines do possess distinct very rare genomic features not present in most other cell lines (specifically tandem repeat expansions at the C9orf72 or FMR1 locus, respectively). All orthogonal tests for these rare genomic features (nanopore sequencing, Southern blots) yielded consistent results and corroborated the authenticity of the cell lines used for this study. Cell line NA 12878 was not cultured in our laboratory, rather publicly aviliable data from Jain et al Nature Biotech 2018 was downloaded and re-analzed. Therefore we have not otherwise autheticated this cell line and have used the raw data solely for vizualisation purposes in a supplementary figure.
Mycoplasma contamination	Monthly Mycoplasma PCR tests (ATCC Universal Mycoplasma Detection Kit (ATCC® 30-1012K™) have been conducted. If a cell culture was detected to be contaminated, it was immediately discarded. No samples with mycoplasma contamination were used for DNA extraction.
Commonly misidentified lines (See ICLAC register)	not applicable, as none of the ICLAC registered lines were used in this study

Human research participants

Policy information about studies involving human research participants

Population characteristics	The human subjects who donated biomaterials for the generation of the hiPSC lines used in this study were solely included because of the presence of pathologic repeat expansions in the C9orf72 or FMR1-genes respectively. No other characteristics (genotype, phenotype etc.) are relevant to this study and thus no other patient characteristics will be reported. The research was overseen and approved by the Ethics commitee of the Christian-Albrechts University (Reference number [Aktenzeichen]: A 145/11)

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging