**Editorial summary**

**Expansions of short tandem repeats linked to neuropsychiatric disorders are measured by single-molecule sequencing.**

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

Pay Giesselmann§1, Björn Brändl§1,2, Etienne Raimondeau3, Rebecca Bowen3, Christian Rohrandt4, Rashmi Tandon2, Helene Kretzmer1, Günter Assum5, Christina Galonska1, Reiner Siebert5, Ole Ammerpohl5, Andrew Heron3, Susanne A. Schneider6, Julia Ladewig7,8,9,10, Philipp Koch7,8,9, Bernhard M. Schuldt2, James E. Graham3, Alexander Meissner1,11 and Franz-Josef Müller\*1,2

**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 (München, 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 (DKFZ) (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, USA)

§ Pay Giesselmann and Björn Brändl contributed equally to this work.

\* corresponding author

**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 & evaluation, integrates conventional sequence mapping of nanopore reads with a 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 [(G4C2)n] within the C9orf72 gene is the most frequent monogenic cause of Frontotemporal Dementia and Amyotrophic Lateral Sclerosis c9FTD/ALS (c9FTD/ALS; OMIM: # 105550).6 Similarly, accumulation of a CGG motif in the FMR1 gene underlies the Fragile X Syndrome (FXS; OMIM # 300624), currently one of the most common identifiable genetic causes of mental retardation and autism.7 In both prototypical repeat expansion disorders (Supplementary Fig. 1, 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 mega-base long nucleotide sequences.11,12 Recent development efforts have focused on increasing throughput, read length13 and the reliable detection of epigenetic modifications from the nanopore raw signal.14,15

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

To enable a robust repeat analysis, we developed a general-purpose signal processing algorithm for the exact quantification of STR numbers in raw nanopore signals (STRique: **S**hort **T**andem **R**epeat **i**dentification, **qu**antification & **e**valuation, Fig. 1a, 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 (G4C2)n-repeat lengths.16 Current (May 2019) production grade (Guppy v3.0.3) software 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 basecalling algorithms cannot satisfactorily resolve expanded STR sequences (Supplementary Fig. 4). For our purpose we systematically combined outputs from three ONT basecaller generations (Albacore, Flappie, Guppy) and different parameter sets with two current sequence based STR quantification approaches (Decoy Alignment18 and RepeatHMM19, Fig. 1b). Albacore performed best with increased window size for the decoy alignment strategy while the high-accuracy model of Guppy provided the best sequence-derived results in combination with the RepeatHMM algorithm (Supplementary Fig. 4,5). 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 basecallers, while enabling improved single read base quality on the genomic level,20 become unreliable for more than 32 (G4C2)n-repeats.

For overcoming these issues with our STRique signal analysis software (see Online Methods for details), first reads spanning a STR location are identified by aligning the conventionally base-called sequences to a reference.21 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 any given STR sequence with a Hidden Markov Model (HMM, Supplementary Fig. 3).22 Aggregated STRique repeat counts matched closely gel electrophoresis profiles (Bioanalyzer) from our synthetic repeat constructs and could be confirmed on the single molecule level by manually counting repeat patterns in raw signal traces (Fig. 1b,c, Supplementary Fig. 6,4).

Previously, repeat instability had been noted in Bacterial Artificial Chromosomes (BAC) containing expanded C9orf72 (G4C2)n-repeats (Online Methods).17 Analysing BAC clone 239 from a c9FTD/ALS patient (G4C2)~800 (Ref. 17) with STRique, we observed STR contractions in many reads and a secondary peak at 800 repeats (Fig 1d, Online Methods), while all evaluated sequence space based methods failed to mirror previously published Southern blot results (Fig.1d, Supplementary Fig. 5). 17

Next, to establish a baseline reference data set, we performed nanopore sequencing of a whole genome library from c9FTD/ALS patient-derived DNA yielding a total of 29 Gbp from a single MinION flow cell. Consistent with approximately 10-fold genome-wide coverage, 10 reads covered the C9orf72 target region. To improve the coverage of any predetermined STR, but particularly the (G4C2)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 The Cas12a-RNP was first applied to selectively target DNA sequences from a patient-derived induced pluripotent stem cell line (24/5#2) adjacent to the (G4C2)n-repeat resulting in unique 4 bp overhangs as molecular tags amenable to ligation of a linker oligo and subsequent attachment of the nanopore sequencing adapter (Fig. 2a: Workflow I, Online Methods). To further improve enrichment results we replaced the oligo adapter 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 adapters. In this enrichment protocol variant, 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 adapters. 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 (G4C2)n-repeat including 40 reads from the expanded allele from a single MinION flow cell (Supplementary Fig. 7, Supplementary Table 3, Supplementary Note). Strikingly, consistent with Southern blot results from the same cell line (Supplementary Fig. 8 a,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, SC105iPS7) carrying distinct FMR1-repeat expansions.24 Employing an 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, Supplementary Fig. 8 c,d).

Since other CRISPR/Cas-nucleases also generate phosphorylated 5’-ends after DNA cleavage,25 we explored, if nucleases such as Cas9 might enable additional improvements of the enrichment results. Therefore, we prepared libraries in parallel with Cas12a- and Cas9-RNPs targeting both FMR1 and C9orf72 regions. Remarkably Cas9-targeting results in an additional increase in sequencing depth in the order of a magnitude for both targeted regions concomitant with a notable reduction in off-target reads (Fig. 2d, Supplementary Fig. 7c). To understand, if the number of reads on target can be further improved by exposing the same Cas12a- or Cas9- enrichment library to an increased number of pores, we subjected equimolar aliquots from the same pooled library preparations from 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).

The epigenetic modification of C9orf72 and FMR1 loci have been correlated with STR expansion status and patient characteristics in both disorders, however without quantification at the single molecule level so far.10,26 Therefore we integrated single read CpG methylation analysis of regions adjacent to both STRs using nanopolish14 with our STRique results (Fig. 3a). We found that in the 24/5#2 line all reads with STR expansions > 750 repeats showed a significantly increased methylation level at the promoter CpG island. In contrast all wild type reads and those with ~450 repeats were not or only partially methylated (Two sided Wilcoxon rank sum test p < 0.001, Fig. 3b, Supplementary Fig. 9, Supplementary Note).

Additionally, in c9FTD/ALS patients pervasive CpG methylation of the (G4C2)n-repeat itself has been reported.27 Assessed with a strictly qualitative assay, the expanded STR itself 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 a CGG repeat expansion at the FMR1-locus beyond a threshold of > 200 repeats leads in most cases to the silencing of the entire FMR1-gene through CpG-methylation.28

Due to the intrinsic heterogeneity in STR length, especially reference genome based methods such as nanopolish14 cannot be used to determine CpG methylation on the repeat expansion itself. To detect 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 then can be summarized into the mean repeat methylation level over the whole repetitive sequence.

When applying the methylation-aware STRique, all expanded FMR1-STRs in nanopore reads from patient SC105 are found to be highly methylated (Supplementary Fig. 10a), consistent with previous analyses29 and our Southern blot results (Supplementary Fig. 8). We next evaluated this approach on plasmids containing n=76 synthetic (G4C2)n and n=99 CGG-repeats (Addgene, #63089), which were covalently modified with the methyltransferase M.SssI (Supplementary Fig. 10).14 In addition we tested the algorithm on (G4C2)n-containing reads (online methods) from patient-derived DNA, which had been modified with M.SssI *in vitro*. In summary, we found that STRique can determine the repeat CpG methylation state correctly in all positive and negative controls evaluated.

Surprisingly though, 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 multi-layered molecular characterization of pathological short tandem repeat expansions. We have increased the enrichment for regions of interest on the background of the human genome approximately two orders of magnitude without any target amplification by using selective, multiplexed CRISPR/Cas-nuclease-based chemical tagging of DNA fragments. Importantly, our method does not require any additional instruments in contrast to other previously reported30 enrichment strategies and enables reporting the DNA methylation status of the same alleles. The CRISPR/Cas-nuclease-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 as of yet unsequenceable genomic regions in human health and disease. This type of analysis improves diagnostic workflows in regard to accuracy and resolution of unstable repeat expansion while enabling efforts to gain mechanistic insights into effects on differentiation, aging and future therapeutic agents that modify DNA methylation.

## Acknowledgements

We are deeply thankful for the invaluable support by c9FTD/ALS and FXS patients and their families who donated biomaterials for this study. The C9orf72 BAC was generously provided by Robert Baloh and Shaughn Bell (Cedars Sinai Medical Center, Los Angeles, CA, USA). We are grateful to Jeanne Loring and Ai Zhang (The Scripps Research Institute, La Jolla, CA, USA) for providing us with hiPSC lines from a Fragile X patient (Supported by NIH R33MH087925-03). We thank Philip van Damme and Wim Robberecht (Laboratory for Neurobiology; VIB-KU Leuven Center for Brain & Disease Research, Leuven, Belgium) for providing the c9FTD/ALS patient-derived fibroblasts used for reprogramming. We acknowledge the expert assistance of the technical staff of the molecular genetics laboratory of the Institute of Human Genetics, Ulm. PK and JL acknowledge financial support by the Hector Stiftung II gGmbH. FJM and RT received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany`s Excellence Strategy – EXC 22167-390884018. FJM and BMS were supported by the BMBF (PluriTest2, 13GW0128A). This work was supported by the Max Planck Society.

Ethics Statement

This work was overseen and approved by the Ethics Committee of the Christian-Albrechts-University, Kiel, Germany (reference number A 145/11). Informed consent was obtained from all donors of cells and tissues for the generation of hiPSC lines and their subsequent genetic and epigenetic analysis. All materials were donated gratuitously by our patients.

## Author Contributions

PG, BMS and FJM conceived the project. BB and RT performed cell culture, plasmid and BAC expansion and extraction. PG wrote the STRique pipeline. PG, BMS, CR, HK conducted additional bioinformatic analyses. PK and JL reprogrammed the c9FTD/ALS hiPSC from patient fibroblasts used in this study. ER, RB, AH, JEG developed the Cas12a- and Cas9- protocols. BB further developed the Cas12a- and Cas9- protocols with c9FTD/ALS and FXS patient-derived DNA and performed nanopore library preparation and nanopore sequencing for the results presented in this manuscript, RT and CG worked on optimization of aspects of the enrichment protocol. GA and RS conducted diagnostic testing of the repeat expansions by Southern Blot and PCR analyses, SS, RS, OA and GA provided clinical and diagnostic advice. PG, BB, BMS, AM and FJM wrote the manuscript. FJM oversaw the study. All authors contributed to the editing and completion of the manuscript.

## Competing Financial Interests Statement

ER, RB, AH and JEG are employees of Oxford Nanopore Technologies Ltd (ONT). CR was reimbursed for travel costs for an invited talk at the Nanopore Days 2018 in Heidelberg (Germany) by ONT. PG 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 and writing of the manuscript.

## 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* 1–7 (2019). doi:10.1080/21678421.2019.1588904

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. *Handbook of Clinical Neurology* **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 G4C2 Repeat in ALS with a C9orf72 Expansion. *The American Journal of Human Genetics* **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. *Nature Biotechnology* **34,** 518–524 (2016).

12. Brown, C. G. & Clarke, J. Nanopore development at Oxford Nanopore. *Nature Biotechnology* **34,** 810–811 (2016).

13. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36,** 338–345 (2018).

14. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* (2017). doi:10.1038/nmeth.4184

15. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* (2017). doi:10.1038/nmeth.4189

16. Mizielinska, S. *et al.* C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. *Science* **345,** 1192–1194 (2014).

17. O'Rourke, J. G. *et al.* C9orf72 BAC Transgenic Mice Display Typical Pathologic Features of ALS/FTD. *Neuron* **88,** 892–901 (2015).

18. Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol* **19,** 121 (2018).

19. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the ‘unsequenceable’ genomic trinucleotide repeat disorders by long-read sequencing. *Genome Medicine 2017 9:1* **9,** 65 (2017).

20. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *bioRxiv* 543439 (2019). doi:10.1101/543439

21. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **3,** 321 (2018).

22. Schreiber, J. & Karplus, K. Analysis of nanopore data using hidden Markov models. *Bioinformatics* **31,** 1897–1903 (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. *Nature Biotechnology* **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. *The Journal of Molecular Diagnostics* **17,** 463–471 (2015).

29. Hansen, R. S., Gartler, S. M., Scott, C. R., Chen, S.-H. & xLaird, 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., Sharim, H., Michaeli, Y. & Ebenstein, Y. Cas9-Assisted Targeting of CHromosome segments (CATCH) for targeted nanopore sequencing and optical genome mapping. *bioRxiv* 110163 (2017). doi:10.1101/110163

## Figure legends

**Figure1** 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 signal of interest. **b)** BioAnalyzer electropherogram, decoy alignment, RepeatHMM and STRique counts of synthetic (G4C2)n repeats (10k random reads per barcode, +/- 10 % intervals around expected repeat length). **c)** Manual confirmation of detected repeat counts in synthetic repeats (n=15, 49, 45, 48, 47, Pearson correlation). **d)** Nanopore sequencing and analysis of BAC clone 239 from a c9ALS/FTD patient compared to cropped corresponding lane from O'Rourke et al. 2015.17 for illustration purpose.

**Figure 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 on c9orf72 locus revealing two distinct repeat bands of ~450 and ~750 (G4C2)n-repeats (n=1810,738,363 evaluated reads with difference in repeat length 392 [95% CI: 383:400, p<2.2e-16]). Coloured points indicate reads used in Fig. 3b. **c)** Repeat quantification of SC105iPS6/iPS7 sample on FMR1 locus (n=174,168 evaluated reads with difference in repeat length -343 [95% CI: -361:-325, p<2.2e-16]). P-values in (**b-c**) from two sided wilcoxon rank sum test illustrated as \* 0.05 - 0.01; \*\* 0.01 - 0.001; \*\*\* < 0.001, data presented as boxplots (centerline: median; box limits: first and third quartiles; whiskers: 1.5x interquartile range) **d)** Mean coverage on target per MinION flow cell (FAK68900, FAK67802 and FAK67994) compared to genome wide means of 100k tiles for WGS, Cas12 and Cas9 enrichment (boxplots for tiles [n=30971] as above, outlier not shown)

**Figure 3** Methylation state analyses on single read level.

**a)** C9orf72 methylation status in Hues64 as measured by WGBS, wild type (blue) allele and expanded allele (~450 (G4C2)n-repeats orange, ~750 (G4C2)n-repeats red) of patient 24/5#2 as measured by nanopore. **b)** Single read nanopore methylation of c9orf72 covering reads from minus strand (n=259,100,43) sorted by detected repeat length. (row: single read, column: single CpG, log-p > 2.5: methylated, log-p < -2.5: unmethylated, two sided Wilcoxon rank sum test on mean promoter CGI methylation, median methylation difference [95% CI] wt-ex450: 3.9e-5 [4.8e-6 - 3.4e-2, p=5.3e-9], wt-ex750: 0.56 [0.46 - 0.64, p<2.2e-16], ex450-ex750: 0.53 [0.40 - 0.64, p<2.2e-16], p-values illustrated as \* 0.05 - 0.01; \*\* 0.01 - 0.001; \*\*\* < 0.001)

## Online Methods

### Raw nanopore signal processing

We propose a raw nanopore signal based pipeline to detect arbitrary repeat patterns based on signal alignment and HMM driven quantification. We adapted the SeqAn2 C++ library31 (<https://github.com/seqan/seqan>) to support semi global alignment of generic signals using the score function:

The distance score in the dynamic programming matrix of two signal values xi, yi 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 negative ones. The negative score is capped at a constant threshold. In contrast to the regular dynamic time warping (DTW) algorithm we applied affine gap costs without penalizing the first and last gap of the alignment (semi-global), allowing us to search for patterns of interest in long nanopore raw signals (Fig. 1a, Supplementary Fig. 3a-b).

Given the prior knowledge of the prefix and suffix sequence of the targeted repeat we simulate these sequences according to the pore model and use the signal alignment to map prefix and suffix in the signal space. A compound profile HMM of prefix, repeat and suffix signal states quantifies 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 is built of generic profile HMM blocks similar to the architecture proposed in Schreiber et al. 2015.22 For a given profile sequence, the expected nanopore signal values are extracted from a pore model. The emission probabilities of the match states are then expressed by normal distributions around these values. Insertion states with an uniform distribution between model minimum and maximum allow to compensate for intermediate state and noise measurements. Silent deletion states enable the model to dodge states without observations. Beside the repeat count STRique provides the alignment scores of prefix and suffix signal. In this study we used a threshold of 3.8 for human and 4.0 for plasmid and BAC samples in order to discard low quality counts.

### MinION base-calling and alignment

All MinION runs were processed using the Nanopype (v0.7.0) pipeline (https://github.com/giesselmann/nanopype).32 The included basecaller versions are Albacore (v2.3.3), Flappie (master, 9ef4edf) and Guppy (v3.0.3). Quality filtering was disabled for any basecalling. Alignments were made against human genome hg19 using minimap2 (v2.14) to identify reads spanning any region of interest and to determine strand orientation respectively.21

### Repeat simulation with Nanopore SimulatION

For development and in silico experimental setup validation repeats were simulated with our Nanopore SimulatION tool (https://github.com/crohrandt/nanopore\_simulation).33 It 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 simulation time. The parameters include read length distribution, signal capture characteristics and experiment metadata. Furthermore the .ini file is a mean of controlling simulation parameters 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 pre derived model of all 6-mers is used. Nanopore SimulatION provides fast5 files that are compatible to the standard software pipeline used for basecalling.

### RepeatHMM repeat detection

RepeatHMM was used as first instance to quantify our synthetic (G4C2)n-repeat plasmid data.19 The software, initially validated with data only from trinucleotide (SCA3/ATXN3) and pentanucleotide (SCA10/ATXN10) tandem repeat expansions, was extended by user defined repeat parameters in this study. Using this feature it was possible to detect the G4C2-hexanucleotide. Furthermore, the software was modified to also report a per read repeat count as it originally only outputs 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 naïve repeat detection approach by base calling reads and aligning against a set of ‘decoy’ references with each possible repeat length in a reasonable range, inspired by the STRetch method proposed by Dashnow and colleagues.18 The repeat count is obtained from the best matching reference without allowing multiple alignments.

We note that RepeatHMM was only developed and validated for trinucleotide (SCA3/ATXN3) and pentanucleotide (SCA10/ATXN10) 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 after repeat numbers increase from ~250 to up to 800 (Supplementary Fig. 5). Interestingly, a subgroup of repeat expansion counts remains in agreement between STR decoy alignment and STRique.

Notably only STRique replicates the peak around ~800 repeats that mirrors the expected repeat expansion maximum seen in Southern blots from the c9FTD/ALS patient and the BAC derived from the patient which we used for this study (see for comparison Fig. 1C Patient 29 in Sareen et al. 2013 and Fig. 1S lane 24 in O’Rourke et al. 2015).9,10

### Methylation detection on expanded reads

Methylation tracks were generated by masking the repeat signal section of expanded reads in the raw fast5 (script fast5Masker.py), re-basecalling and alignment with minimap2 (v2.14) and finally nanopolish (v0.11.0) methylation calling each with default parameters.14 The raw nanopore methylation log-likelihood ratio per CpG is interpreted as methylated for values greater 2.5 and unmethylated for values less than -2.5. Intermediate values are used in the single read heat maps but discarded in the region’s methylation tracks.

The methylation state of the repeat expansion cannot be detected using an established reference genome based workflow of basecalling, alignment and nanopolish. Since the observed repeat length heterogeneity requires a non-reference based CpG methylation detection algorithms 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 is duplicated, allowing the model to switch between native and modified sequence for each repeat iteration. The output is a character string of zero and one allowing the computation of a mean repeat modification level per read.

The repeat methylation detection was evaluated on 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 found to be highly methylated in the samples from SC105iPS6/iPS7.24 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 (G4C2)-repeats by sequencing a M.SssI treated library from our c9FTD/ALS hiPSC line 24/5#2 and detected as expected both, wild type and expanded repeats fully methylated (Supplementary Fig. 10c).

The detection of 5mC from raw nanopore data depends on the signal difference in the CpG-kmer context.14 For the analysed 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 are for the (G4C2)-repeat template 2.75pA against complement 6.88pA and for the CGG-repeat template 5.62pA against complement 11.71pA. We therefore only report repeat methylation readouts from the minus strand in this study.

### Bioinformatics

Plots were generated with R34 (v3.6.0) using GViz (v1.28.0) for region methylation tracks and Complex Heatmaps35 (v2.1.0) with circlize36 (v0.4.6) for single read methylation heat maps.

### Preparation of high molecular weight DNA from cultured c9FTD/ALS patients derived cells

High Molecular Weight DNA (HMW DNA) was prepared with a modified phenol-chloroform extraction method.37 Briefly 2x107 of undifferentiated hiPSC were detached with TrypLE Select (Thermo Fisher Scientific) for 3 min at 37 °C. Enzymatic reaction was stopped with DMEM/F-12 (Thermo Fisher Scientific) and the cell suspension was centrifuged for 3 min at 260 x g. Supernatant was discarded and cells were resuspended in 100 µl of 1X 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 (w/v) and 20 µg/mL RNase A (Qiagen) for 1 h at 37 °C. Proteins were subsequently digested at 50 °C for 3 h using 50 µl Proteinase K (> 600 mAU/mL, 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 ultra-pure saturated phenol (Thermo Fisher Scientific) was added. Samples were placed onto a rotator at 40 rpm for 10 min until a fine emulsion had formed and a phase separation was performed by centrifugation at 2.800 x 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 ultra-pure saturated phenol and 5 mL 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 gentle inversion for 10 times. Precipitated DNA was spooled out of solution using a glass rod and carefully submerged in 80 % ethanol. Washed HMW DNA were transferred into a 1,5 mL DNA LoBind tube (Eppendorf) containing 1 mL of 80 % ethanol and centrifuged at 16.000 x g for 10 min. Supernatant was removed and the DNA pellet was dried at 40 °C for 5-10 min. Rehydration of DNA was done 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 days and DNA was further homogenized on a rotator at 37 °C and 20 rpm overnight.

### Selective ligation of genomic fragments with CRISPR-Cas12a-RNPs mediated DNA cleavage.

We designed CRISPR-Cas12a crRNAs (IDT) targeting genomic regions adjacent to the C9orf72 (G4C2)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 had been introduced into the literature as Cpf1 (**C**RISPR from **P**revotella and **F**ranciscella**1**) 23 but is now more frequently referred to as Cas12a.23,40 Next we designed unique ssDNA oligomere (“bottom strand”, IDT) in parts complementary to a universal ssDNA barcode (NB01) referred to as “top strand” based on the ONT Native Barcoding Kit 1D (EXP-NBD103). Oligomere 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 were 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 mixture was cooled down to 25°C over 70 min with a cooling rate of approximately -1 °C per minute. Successfully annealed barcodes serve as a linker that provide complementary overhangs to both, the respective Cas12a cut site and the BAM 1D adapter, thus allowing a sticky-end ligation.

Preparation of programmed Cas12a nucleoprotein complex (Cas12a-RNP) was performed by combining the EnGen Lba Cas12a (NEB) with IDT Alt-R™ Cpf1 crRNA that consists of a target specific 21 bp protospacer domain and a constant 20 bp loop domain. Briefly 2 nmol of lyophilized crRNA were reconstituted in 20 µl of 10 mM nuclease-free TE buffer pH 7.5 to obtain a 100 µM solution. Next, crRNA were pooled and adjusted to 10 µM using nuclease-free water. Formation of secondary RNA structure were done in 1X 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 assembled complex was stored on ice until use.

HMW DNA were digested with 10 µl of FastDigest BamHI (Thermo Fisher Scientific) in 1X NEB CutSmart Buffer for 12 h at 37 °C and subsequently heat inactivated at 80 °C for 5 min. As a next step 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. Sample were equilibrated to room temperature prior Cas12a-RNP incubation. To allow sufficient Cas12a cleavage, dephosphorylated DNA were supplemented with 10 µl of Cas12a-RNPs and incubated for 1 h at 37 °C. Cleavage reaction was continued at 4 °C overnight.

Next day the cleaved DNA samples were carefully cleaned and concentrated by using 1X volume of Agencourt AMPure XP beads (Beckman Coulter). DNA binding to AMPure XP beads was done for 10 min at RT followed by quick pulse centrifugation. Next, beads were immobilized at a DynaMag Spin magnet (Thermo Fisher Scientific) for 5 min, and the 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 sec. AMPure XP beads were carefully resuspended in 40 µl of pre-warmed (37 °C) 0,1X TE buffer using large bore tips, followed by a 10 min incubation step at 37 °C. Sample tube was placed on the magnet and eluted DNA were used for barcode adapter ligation.

Annealed barcode adapter (see above) were pooled and adjusted to 1 µM working concentration. A total of 0,2 µl pooled barcode mix were added to the cleaved DNA. Sample 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 adapter from the ONT Native Barcoding Kit 1D (EXP-NBD103) were added to the above mix and subsequently ligated at RT for 30 min.

The sequencing library were purified again by adding 0.4X volume of Agencourt AMPure XP beads for 10 min at RT. Excess of BAM 1D were washed away with adapter bead binding (ABB) solution according to the manufacturer’s protocol. Final sequencing library were 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 eluted library were transferred to a fresh DNA LoBind tube at 4 °C. Sample was then quantified using a Qubit fluorometer together with the Qubit dsDNA BR Assay Kit (both Thermo Fisher Scientific) and a total of 17,5 µl of library were combined with 35 µl of running buffer (RBF) and 22,5 µl of library loading beads (LLB) from the EXP-LLB001 kit (ONT). After removal of the AMPure XP beads, a ONT SpotON Flow Cell R9.4 (FLO-MIN106) was primed and loaded following the manufacturer’s instructions with no modifications to the procedure. Supplementary tables 3-6 give an overview of the flow cells, DNA samples, yields and run characteristics used for this study.

Further optimization of the enrichment results were achieved with a modified Cas12a protocol that contained a dA-Tailing reaction resulting in substantially higher amount of on target reads and took advantage of an improved sequencing kit (SQK-LSK109). Initial BamHI digestion, the use of barcode adapter and overnight incubation with RNP was no longer required. Instead HMW DNA was diluted to a final concentration of 200 ng/µl and a total of 5 µg HMW DNA were dephosporylated with 2 µl Quick CIP in 1X CutSmart buffer (NEB) with same parameters as described above. Assembled RNP complex were 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 were adjusted to 49 µl using nuclease-free water. Briefly 5 U of NEBs Klenow Fragment (3’→5’ exo-) were added to the sample and both dA-tailing as well as RNP cleavage were done 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 adapter the dA-tailed library were incubated with an adapter ligation mix composed of 25 µl ligation buffer (LNB), 10 µl Quick T4 DNA Ligase (NEB), 10 µl nuclease-free water and 5 µl of adapter mix (AMX). Adapter ligation was done for 20 min at RT and mixture was subsequently diluted with 1X volume of 10 mM nuclease-free Tris-Cl pH 8. Following ligation sample were incubated with 0,3X volume of AMPure XP beads for 10 min at RT and washed twice with 250 µl long fragment buffer (LFB) following the manufacturer’s instructions. Washed library were eluted with 16 µl 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 on the manufacturer’s protocol with minor modification. Prior sample loading a SpotON priming mix composed of 20 µl sequencing buffer (SQB), 0,4 µl sequencing tether (SQT) and 19,6 µl nuclease-free water were added dropwise to the SpotON port. The eluted DNA library were supplemented with 25 µl SQB together with 10 µl loading beads (LB) and immediately loaded as described earlier.

### Selective ligation of genomic fragments with CRISPR-Cas9-RNPs mediated DNA cleavage

Cas9-based target enrichment experiments were performed similar to the Cas12a enrichment setup. Briefly we designed CRISPR-Cas9 crRNAs (IDT) targeting genomic regions adjacent to the C9orf72 (G4C2)n- and FMR1 CGG-repeats with the ChopChop online tool (Supplementary Table 2). Preparation of Cas9 nucleoprotein complex (Cas9-RNP) was performed as follows. Lyophilized Cas9 crRNA XT and tracrRNA (both 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 crRNA were pooled to obtain a 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. Sample was incubated for 5 min at 95 °C in a thermal cycler (BioRad) and allowed to anneal at RT. A 10X mastermix 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 10X NEB CutSmart buffer (NEB) and 79,2 µl nuclease-free H2O. Cas9-RNP were formed for 30 min at RT and 10 µl was subsequently used for incubation of 5 µg dephoshporylated HMW DNA (see above) together with 1 µl of 10 mM dATP and 5 U of Taq polymerase (NEB). Reaction mixture was incubated for 30 min at 37 °C on a thermo block and Cas9 enzyme was heat inactivated for 5 min at 72 °C thereafter. All subsequent steps were performed as described in the CRISPR-Cas12a-RNP section.

For methylation calling on C9orf72 (G4C2)n- and FMR1 CGG-repeats from patient 24/5#2 we included a methylation step with M.SssI prior to library preparation. Briefly 4 µg of HMW DNA was supplemented with 10X CutSmart buffer (1X final), a final concentration of 640 µM S-adenosylmethionine (SAM) and subsequently treated with 20 U of M.SssI CpG Methyltransferase (both NEB) at 37 °C for 3 h followed by additional 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 from our C9orf72 patient sample (#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 DNA at 7200 rpm using a Covaris g-Tube (Covaris). A 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 recommendation. Eluted DNA was pooled and bound to 0,7X AMPure beads for 10 min at RT followed by two washing steps with 80 % ethanol. Two aliquots of 3,5 µg of genomic DNA were used for NEBNext FFPE Repair mix (NEB) and dA-tailing with the NEBNext Ultra II End-prep enzyme mix (NEB) in a single reaction and incubated at 20 °C for 5 min. Enzymatic reaction was stopped at 65 °C for 5 min and DNA was washed using 0,7X AMPure beads. A total of 4 µg end-repaired library was split into two aliquots for the adapter ligation followed by two washing steps with ONT long fragment buffer (LFB). The resulting library was eluted with ONT elution buffer (EB) and 600 ng DNA was loaded onto a MinION flow cell. The sequencing library was refreshed every 22 h by using a nuclease flush protocol. Briefly 40 U of DNase I (NEB) was supplemented to 380 µl wash solution A (EXP-WSH002) and was 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 subsequently loading 600 ng of fresh library onto the flow cell.

For sequencing our synthetic (G4C2)n-repeats 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 clean-up following the manufacturer’s instructions. Linearized plasmid DNA, was further processed with the ONT Ligation Sequencing kits 1D (SQK-LSK108) following the manufacturer’s instructions without modifications. In some instances, the plasmids were barcoded using adapters 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 99x CGG FMR1, Addgene #63089) that contained a stretch of 99 CGG-repeats as a kind gift from Nicolas Charlet-Berguerand. A bacterial stab culture was used to streak *E. coli Stbl3* onto selective Luria Broth Agar plates containing 100 µg/mL of ampicillin. Agar plates were incubated for 48h at RT and individual colonies were picked and transferred in Luria Broth medium supplemented with 100 µg/mL of ampicillin. Inoculated samples were incubated at RT for 20 h and 200 rpm in a bacterial shaker (N-Biotek) and FMR1 plasmid was isolated using the Qiagen Plasmid Plus Midi Kit (Qiagen) following the manufacturer’s recommendations. Integrity of the FMR1 plasmid was confirmed by a restriction digest with 20 U of ScaI , 20 U of EcoRV and 20 U of SpeI (all NEB). Prior to 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 digest was stopped at 80 °C for 20 min followed by a SPRI clean up. 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 final library was loaded onto a MIN-FLO106 flow cell.

DNA obtained from pCC1-BAC clone 239 17 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) prior sample loading on the flow cells.

### Synthetic (G4C2)n-repeat cloning, modification and visualisation

For cloning (G4C2)n fragments carrying a defined repeat number we used recursive directional ligation as described in Mizielinska et al. 2014.16 Briefly, complementary DNA oligonucleotides (Supplementary Table 7) containing three or four G4C2 -repeats flanked 5’ by BamHI and BspQI (GCTCTTCC\*GGCC) recognition sites and 3’ 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), 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/Maxi kits (Qiagen), following the manufacturer’s instructions. Constructs were screened using standard restriction enzyme digest and agarose gel electrophoresis (2 % or 4 %). Next we sequence-verified (GENterprise Genomics, Mainz, Germany) plasmids with up to 32 repeats, to confirm repeat size and lack of interruptions (data not shown). In case of the 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/NotI. For exact repeat length visualization and comparison with nanopore results from the same plasmids, we followed procedures outlined by Kwok and colleagues for the detection of the FMR1-repeat expansionwith 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 (G4C2)n-plasmids with restriction enzymes. We analyzed repeat inserts from different repeat lengths (8, 32, 50, 56, 76), which were excised with BamHI/NotI or NdeI/ScaI from the pCR Script Amp-BE plasmids and analyzed with a 2100 Agilent Bioanalyzer on a 1000 DNA gel cassette following the manufacturer's instructions (Agilent, Böblingen, Germany). Bioanalyzer raw data was 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 bacterial artificial chromosome (pCC1-BAC clone 239)17 containing a (G4C2)-repeat expansion from a c9FTD/ALS patient 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 Chloramphenicol at temperatures < 30 °C as higher temperatures lead to repeat contraction and/or loss of the BAC. Extraction of BAC DNA was done using the Qiagen 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 high17, with rates between 20 % and up to 80 % depending on bacterial media and growth temperature conditions with richer media and faster/denser growth rates (Shaughn Bell, Cedar Sinai Medical Center, Los Angeles, CA, USA, personal communication). Therefore attention was paid to grow the cells in relative low density and at lower temperature (~ 27 °C).

### Generation and culture of hiPSC lines from c9FTD/ALS patients

Human induced stem cells from a c9FTD/ALS patient 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 post transduction clones were manually picked and clonally expanded as hiPS cell lines on feeder cells. Established hiPSC cell colonies show the typical morphology of human pluripotent stem cells, stain positive for alkaline phosphatase (AP) and express the pluripotency-associated surface proteins TRA-1-60 and TRA-1-80. High resolution SNP-karyotyping was performed to exclude major karyotypic abnormalities induced by the reprogramming or culturing process (Supplementary Fig. 11 a-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 days as clumps with 0,5 mM EDTA in 1X PBS.42 Detection of AP was performed using the VECTOR Blue Alkaline Phosphatase (Blue AP) Substrate Kit (Vector Laboratories) following the manufacturer's instructions. Staining against TRA-1-60 and TRA-1-81 was done as described below. Briefly cells were rinsed with 1X PBS followed by a paraformaldehyde (4 %) fixation for 10 min at RT. Samples were washed twice with 1X PBS prior incubation with primary antibodies against Tra1-60 and Tra1-81 (both 1:500 diluted, Thermo Fisher Scientific) for 2h at RT. Cells were washed with 1X PBS, incubated with Alexa 488 anti-ms secondary antibody (1:1000, 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 c9FTD/ALS patients were screened by Southern blot analysis (see below), and the line 24/5#2 (Supplementary Fig. 11 a-e) was selected for further nanopore sequencing analysis as it displayed at least two distinct maxima of the expanded allele around 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 FraX patient with 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 instruction.24 Our Southern blot results showed peaks at approximately 750 (SC105iPS6) and 500 (SC105iPS7) CGG repeats ([Supplementary Fig.](https://docs.google.com/document/d/1ZYvzmagspZy13ZuJBBXpunfrlePCeDwxBCoKnXizyTk/edit?usp=sharing) 8) in agreement with the nanopore sequencing results from the same DNA preparations (see main text). We note that marked instability of FMR1-CGG STR expansions in other hiPSC and human embryonic stem cell lines through *in vitro* culture has been described before.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 means to estimate the repeat expansions present in our hiPSC lines from patients ([Supplementary Fig.](https://docs.google.com/document/d/1ZYvzmagspZy13ZuJBBXpunfrlePCeDwxBCoKnXizyTk/edit?usp=sharing) 8). The samples were processed and analysed in the routine clinical diagnostic workflow for FXS and c9ALS/FTD (certified according to DIN EN ISO 15189:2014) at the Department for Human Genetics at Ulm University (Ulm, Germany). The diagnosticians were blinded for the experimental nature of the hiPSC samples and the DNA was processed in parallel with actual clinical samples.

Briefly, for the Southern blot determination of repeat expansion length of the (G4C2)n- and (CGG)n-repeats in our hiPSC lines 10 µg HMW DNA was digested overnight with HindIII (20 U) and XbaI [20 U, for (G4C2)n] and EcoRI (20 U) and NruI [20 U, for (CGG)n] respectively prior to electrophoresis. The fragmented gDNA was separated on a 0.8 % Agarose gel for 20 min with 180 V followed by 65 V overnight (C9orf72) or 20 min at 180 V followed by 60 V overnight (FMR1) respectively. The resulting gel was imaged with an ethidium bromide fluorescent stain and a copy of the image 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 hr.

The hybridization probes were either a 210 bp PCR-fragment located upstream of the G4C2-repeat in the *C9orf72*-gene or a 480 bp PCR-fragment located downstream of the CGG-Repeat in the *FMRI*-gene respectively. Probes (100 ng/ filter) were labelled with 50 mCi α32PdCTP 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 wildtype *C9orf72*-alleles are approximately 2.3 kb in length. Wildtype *FMRI*-fragments from active, unmethylated X-chromosomes are approximately 2.9 kb, from inactivated, methylated X-chromosomes 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 using Pearson correlation (*r*) as indicated in the respective figures. The significance of repeat count differences was determined using two sided Wilcoxon rank sum tests (R: wilcox.test). The significance of the difference in mean methylation levels of the promoter CpG island between repeat count clusters (Fig. 3b) and between unmethylated BAC and cell line 24/5#2 (Supplementary Fig. 9c) was determined using two sided Wilcoxon rank sum tests. These p-values were corrected for multiple testing according to Holm (R: p.adjust).

The experiments to generate a reference repeat count data set (plasmid for 8, 32, 50, 56 and 76 repeats, BAC for ~800 repeats) were performed once. The characterization of the 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/(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.

## Code availability

All custom code developed for this study is under MIT license available at:

<https://github.com/giesselmann/STRique>

The RepeatHMM package was forked and is modified available at:

<https://github.com/giesselmann/RepeatHMM>

## Methods-only 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 1–8 (IEEE, 2019). doi:10.1109/BIBM.2018.8621253

34. R Core Team. R: A language and environment for statistical computing. *R Project* (2019).

35. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps 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. *Nucl. 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. *Nucl. 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. *Clinica Chimica Acta* **456,** 137–143 (2016).

42. Chen, G. *et al.* Chemically defined conditions for human iPSC derivation and culture. *Nat Methods* (2011). doi:10.1038/nmeth.1593

43. Mertens, J. *et al.* APP processing in human pluripotent stem cell-derived neurons is resistant to NSAID-based γ-secretase modulation. *Stem Cell Reports* **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. *Molecular Autism* **7,** 165 (2016).