Repeat expansion and methylation state analysis with nanopore sequencing

Pay Gießelmann[§1], Björn Brändl[§1,2], 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], Bernhard M. Schuldt[2], James E. Graham[3], Alexander Meissner[1,11] 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 Technology and Microelectronics (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 (Heidelberg, Germany) **9** German Cancer Research Center (DKFZ) (Heidelberg, Germany) **10** Institute of Reconstructive Neurobiology, University of Bonn Medical Center (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 neuropsychiatric and other disorders but their assessment remains challenging with current molecular methods. Here, we developed a Cas12a-based enrichment strategy for nanopore sequencing that, combined with a new algorithm for raw signal analysis, enables us to efficiently target, sequence and precisely quantify repeat numbers as well as their DNA methylation status. Taking advantage of these single molecule nanopore signals provides therefore unprecedented opportunities to study pathological repeat expansions.

The expansion of unstable genomic Short Tandem Repeats (STRs) causes more than 30 Mendelian human disorders[1]. For example, expansion of a GGGGCC-repeat $[(G_4C_2)_n]$ within the C9orf72 gene is the most frequent monogenic cause of Frontotemporal Dementia (FTD) and Amyotrophic Lateral Sclerosis (ALS; c9FTD/ALS; OMIM: # 105550)[2,3]. Similarly, accumulation of a CGG motif in the FMR1 gene underlies the Fragile X Syndrome (FXS; OMIM # 300624), currently the most common identifiable genetic cause of mental retardation and autism[4]. In both prototypical repeat expansion disorders (Suppl. Discussion 1), recent evidence has suggested pronounced inter- and intraindividual repeat variability as well as changes in DNA methylation of the respective genomic regions to modulate disease phenotype[5-8].

To overcome current difficulties in characterizing expanded STRs (Suppl. Discussion 2) most notably we focused on three areas: i) optimization of Nanopore sequencing and signal processing to capture STRs ii) development and implementation of a

43 target enrichment strategy to increase efficiency and iii) integration of expansion

44 measurements with DNA methylation of the same molecule.
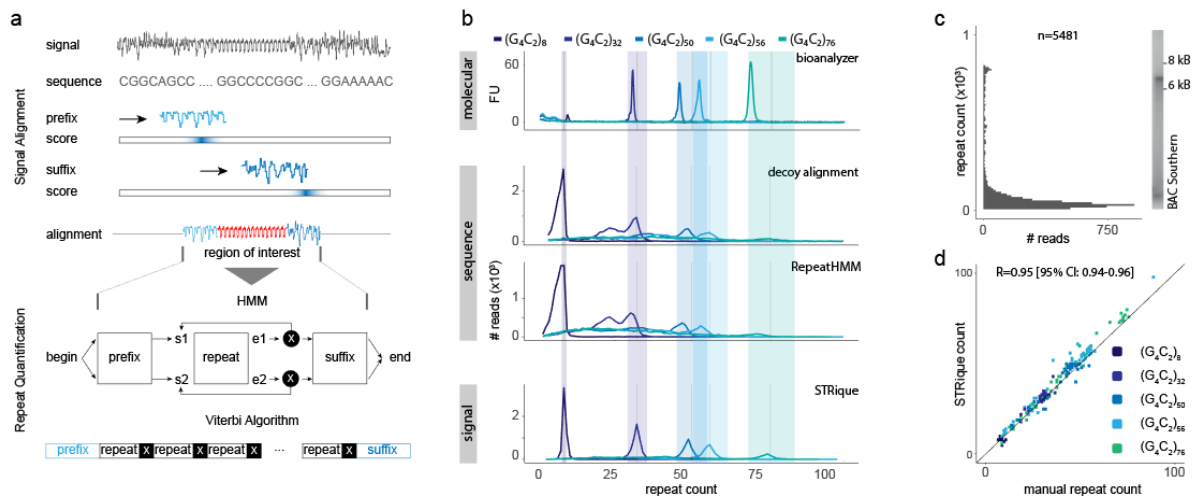
45

46 **Figure1** nanoSTRique: Generic repeat detection pipeline on raw nanopore signals.

47 **a)** Repeat quantification by signal-alignment of flanking prefix and suffix regions and HMM based
48 count on signal of interest. **b)** BioAnalyzer electropherogram, decoy alignment, RepeatHMM and
49 nanoSTRique counts of synthetic $(G_4C_2)_n$ repeats (10k random reads per barcode, +/- 10 % intervals
50 around expected repeat length). **c)** Nanopore sequencing and analysis of BAC clone 239 from a
51 c9ALS/FTD patient compared to cropped corresponding lane from Ref. 15 for illustration purpose. **d)**
52 manual confirmation of detected repeat counts in synthetic repeats (n=16, 50, 49, 49, 47).

53 First, for benchmarking repeat expansion counting methods we constructed, verified

54 and nanopore sequenced plasmids with several synthetic $(G_4C_2)_n$-repeat lengths[9].

55 We analyzed our results with currently available STR quantification pipelines[10,11] but

56 found those methods to become unreliable for more than 32 $(G_4C_2)_n$-repeats with

57 nanopore reads. To further improve the repeat analysis we developed a signal

58 processing algorithm for a more exact quantification of STR numbers in raw

59 nanopore signals (nanoSTRique: **nano**pore **S**hort **T**andem **R**epeat **i**dentification,

60 **qu**antification &**e**valuation, Fig. 1a, Suppl. Fig. 1). Briefly (see Online Methods for

61 details), reads spanning a STR location are identified by aligning the conventionally

62 base-called sequences to a reference[12]. Next, nanoSTRique maps the upstream and

63 downstream boundaries of each repeat more precisely with a signal alignment

64 algorithm using SeqAn2[13] and, as a third step, accurately quantifies the number of

65 any given STR sequence with a Hidden Markov Model (Suppl. Fig. 1)[14]. Aggregated

66 nanoSTRique repeat counts matched closely gel electrophoresis profiles from our

67 synthetic repeats and could be confirmed on the single molecule level by manually

68 counting repeat patterns in raw signal traces (Fig. 1d) (Suppl. Fig. 2, 3). Previously,

69 repeat instability had been noted in Bacterial Artificial Chromosomes (BAC)

70 containing expanded C9orf72 $(G_4C_2)_n$-repeats (Online Methods)[15]. Analysing BAC

71 clone 239 from a c9FTD/ALS patient $(G_4C_2)_{\sim 800}$[15] with nanoSTRique we observed

72 STR contractions in many reads and a secondary peak at 800 repeats (Fig 1c,

73 Online Methods), while existing methods failed to mirror previously published

74 Southern blot results (Suppl. Fig. 3).

75 Next we performed whole genome nanopore sequencing from c9FTD/ALS patient-

76 derived DNA using four MinION flow cells yielding a total of 39 GBp (8M reads, 11

77 on target, none expanded). Our and similar nanopore whole genome sequencing

78 results from others[16] suggested a relevant bias against the detection of $(G_4C_2)_n$-

79 expansions with conventional molecular and bioinformatic nanopore workflows. To

80 improve the coverage of any STR particularly the $(G_4C_2)_n$-region in our proof of

81 concept, we took advantage of the programmable CRISPR-Cas12a-

82 ribonucleoprotein (Cas12a-RNP), which cleaves DNA via staggered double-strand

83 breaks[17]. The Cas-system was applied to selectively target DNA sequences from a

84 patient-derived induced pluripotent stem cell line (24/5#2) adjacent to the $(G_4C_2)_n$-

85 repeat resulting in unique 4 bp overhangs amenable to ligation of a linker oligo and

86 subsequent attachment of the nanopore sequencing adapter (Fig. 2a, Online

87 Methods). To further improve enrichment results we replaced the oligo adapter

88 ligation step by adding Klenow fragment to fill in the Cas12a overhangs. The

4

89   resulting dA-tailed DNA ends enabled even more efficient ligation of the sequencing

90   adapters.

91
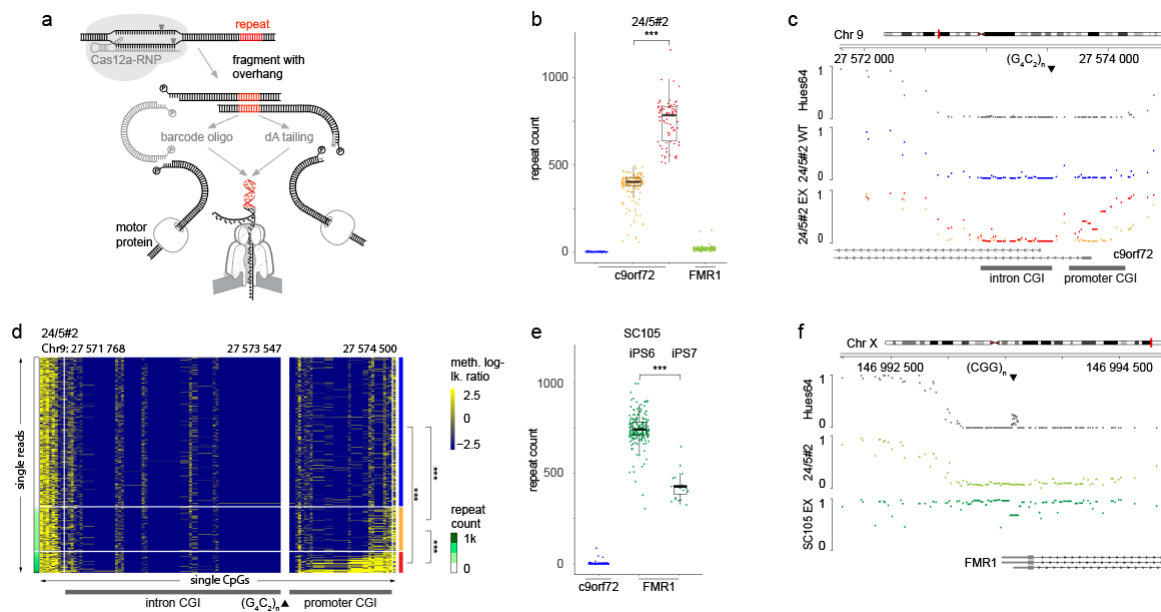
Figure 2 Targeted enrichment and nanopore sequencing with CRISPR-Cas12a.

93   **a)** Illustration of the CRISPR-Cas12a target enrichment procedure **b)** Repeat quantification of sample
94   24/5#2 on c9orf72 and FMR1 loci revealing two distinct repeat bands of ~450 and ~750 $(G_4C_2)_n$-
95   repeats. (n=442,165,78,499; difference in repeat length 385 [95% CI: 361:404]) **c)** c9orf72
96   methylation status in Hues64 (WGBS), wild type allele and expanded allele of patient 24/5#2
97   (nanopore). **d)** Single read nanopore methylation of c9orf72 covering reads (n=769, row split 30,500)
98   sorted by detected repeat length. (row:single read, column: single CpG, log-p > 2.5: methylated, log-p
99   < -2.5: unmethylated, two sided wilcoxon rank sum test on mean promoter CGI methylation, median
100   methylation difference [95% CI] wt-450: 6.9e-6 [3.5e-5 - 3.3e-2], wt-750: 0.33 [0.24 - 0.42], 450-750:
101   0.30 [0.20 - 0.42]) **e)** Repeat quantification of SC105iPS6/iPS7 sample on c9orf72 and FMR1 loci.
102   (n=850,183,23; difference in repeat length -322 [95% CI: -351:-298]) **f)** FMR1 methylation status in
103   Hues64 (WGBS), wild type allele of patient 24/5#2 and expanded alleles of samples SC105iPS6 and
104   SC105iPS7[19] (nanopore). (All box plots show median, box edges represent 1st and 3rd quartiles,
105   whiskers extend to 1.5 x IQR; two sided wilcoxon rank sum tests, p-vals: * 0.05 - 0.01; ** 0.01 - 0.001;
106   *** < 0.001)

107   Additional dephosphorylation of all 5' ends before Cas12a-RNP-digestion chemically

108   protects DNA 'background' fragments from being ligated to sequencing adapters.

109   Consequently, only those fragments cut by Cas12a-RNPs are capable of being

110   sequenced by this procedure (Fig. 2a). Using this approach we were able to obtain a

111   total of 1137 reads covering the $(G_4C_2)_n$-repeat including 442 evaluable reads from

112   expanded alleles (Suppl. Table 2, Suppl. Discussion 3). Consistent with Southern

113   blot results from the same cell line (Suppl. Fig. 6 a-d), we found two distinct repeat

5

114 expansion distributions (Fig. 2b). To explore the general applicability of our

115 enrichment, sequencing and read processing we next tested two isogenic, patient-

116 derived cell lines (SC105iPS6, SC105iPS7) carrying a FMR1-repeat expansion with

117 an additional set of FMR1-targeting Cas12a-RNPs (Suppl. Table 1) and found two

118 different repeat expansion distributions (Fig. 2e).

119 Lastly, epigenetic modification of both C9orf72 and FMR1 loci have been correlated

120 with STR expansion status and patient characteristics in both disorders[7,8]. We

121 therefore combined single molecule CpG methylation analysis using nanopolish[18]

122 with our nanoSTRique results and found that in the 24/5#2 line all reads with STR

123 expansions > 750 repeats showed significantly increased methylation at the

124 promoter CpG island while all wild type reads and those with < 500 repeats were not

125 or only partially methylated (Two sided wilcoxon rank sum test p < 0,001, Fig. 2C-D,

126 Suppl. Fig. 4, Suppl. Discussion 4). Similarly, we found all FMR1 wild type alleles

127 enriched from controls were unmethylated at a CpG island overlapping the CGG-

128 STR. Consistent with previous findings[19] and in our Southern blot analyses, all

129 expanded alleles identified by nanoSTRique from both isogenic FXS-patient derived

130 cell lines[19], were determined to be methylated (Fig. 2f, Suppl. Fig. 6f-h).

131 Our results demonstrate the power of nanopore sequencing for the precise and

132 multilayered molecular characterization of pathological short tandem repeat

133 expansions. We have increased the enrichment for regions of interest on the

134 background of the human genome approximately two to three orders of magnitude

135 without any target amplification by using selective, multiplexed Cas12a-based

136 chemical tagging of DNA fragments. Importantly, our method does not require any

137 additional instruments in contrast to other previously reported enrichment

138 strategies[20] and enables reporting the DNA methylation status of the same alleles.

139 The Cas12a-target enrichment and nanoSTRique can be rapidly adapted to any

140 other genomic region of interest, ensuring broad applicability to overcome challenges

141 associated with the single molecule analysis integrating genetic and epigenetic

142 signals associated with unstable repeat expansions or any other as of yet

143 'unsequenceable' genomic regions in human health and disease. This type of

144 analysis improves diagnostic workflows in regard to accuracy and resolution of

145 unstable repeat expansion while enabling efforts to gain mechanistic insights into

146 effects on differentiation, aging and future therapeutic agents that modify DNA

147 methylation.

148 7. Acknowledgements

161 8. Author information

162 Contributions:

163 PG, BMS and FJM conceived the project. BB and RT performed cell culture, plasmid

164 and BAC expansion and extraction. PG wrote the nanoSTRique pipeline. PG, BMS,

165 CR, HK conducted additional bioinformatic analyses. PK and JL reprogrammed the

166 c9FTD/ALS hiPSC from patient fibroblasts used in this study. ER, RB, AH, JEG

167 developed the Cas12a-RNP protocol. BB further developed the Cas12a protocol with

168 c9FTD/ALS and FXS patient-derived DNA and performed all nanopore library

169 preparation and nanopore sequencing for the results presented in this manuscript,

170 RT and CG worked on optimization of aspects of the enrichment protocol. GA and

171 RS conducted diagnostic testing of the repeat expansions by Southern Blot and PCR

172 analyses, SS, RS, OA and GA provided clinical and diagnostic advice. PG, BMS, AM

173 and FJM wrote the manuscript. FJM oversaw the study. All authors contributed to the

174 editing and completion of the manuscript.

175 Competing Interests Declaration:

176 ER, RB, AH and JEG are employees of Oxford Nanopore Technologies Ltd (ONT).

177 CR was reimbursed for travel costs for an invited talk at the Nanopore Days 2018 in

178 Heidelberg (Germany) by ONT. ONT had no role in the study design, interpretation

179 of results and writing of the manuscript.

180 9. Data availability

181 All sequencing data points generated in this study (i.e. nanopore reads, raw and

182 base called) and utilized for the determination of FMR1 and C9orf72 repeat

183 expansion lengths and methylation status will be made public upon acceptance

184 through a public repository (e.g. EGA). Whole genome sequencing data generated

185 for this study will be made accessible to researchers upon request under a Data

186 Access Agreement similar to the procedure required for managed access by the

187 European Genome-phenome Archive. (for details see:

188 https://www.ebi.ac.uk/ega/submission/data_access_committee/policy_documentation).

189

190 1. Gatchel, J. R. &Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and
191 common principles. *Nat Rev Genet***6,** 743–755 (2005).
192 2. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of
193 chromosome 9p21-linked ALS-FTD. *Neuron***72,** 257–268 (2011).
194 3. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding
195 region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron***72,** 245–256 (2011).
196 4. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a
197 breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell***65,** 905–914
198 (1991).
199 5. van Blitterswijk, M. *et al.* Association between repeat sizes and clinical and pathological
200 characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional
201 cohort study. *Lancet Neurol***12,** 978–988 (2013).
202 6. Xi, Z. *et al.* Hypermethylation of the CpG Island Near the G4C2 Repeat in ALS with a C9orf72
203 Expansion. *The American Journal of Human Genetics***92,** 981–989 (2013).
204 7. Russ, J. *et al.* Hypermethylation of repeat expanded C9orf72 is a clinical and molecular
205 disease modifier. *Acta Neuropathol.***129,** 39–52 (2015).
206 8. Hornstra, L. K., Nelson, D. L., Warren, S. T. & Yang, T. P. High resolution methylation analysis
207 of the FMR1 gene trinucleotide repeat region in fragile X syndrome. *Hum Mol Genet***2,** 1659–
208 1665 (1993).
209 9. Mizielinska, S. *et al.* C9orf72 repeat expansions cause neurodegeneration in Drosophila
210 through arginine-rich proteins. *Science***345,** 1192–1194 (2014).
211 10. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the 'unsequenceable' genomic
212 trinucleotide repeat disorders by long-read sequencing. *Genome Medicine 2017 9:1***9,** 65
213 (2017).
214 11. Dashnow, H. *et al.*STRetch: detecting and discovering pathogenic short tandem repeat
215 expansions. *Genome Biol***19,** 121 (2018).
216 12. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics***3,** 321 (2018).
217 13. Reinert, K. *et al.* The SeqAn C++ template library for efficient sequence analysis: A resource
218 for programmers. *J. Biotechnol.***261,** 157–168 (2017).
219 14. Schreiber, J. &Karplus, K. Analysis of nanopore data using hidden Markov models.
220 *Bioinformatics***31,** 1897–1903 (2015).
221 15. O'Rourke, J. G. *et al.* C9orf72 BAC Transgenic Mice Display Typical Pathologic Features of
222 ALS/FTD. *Neuron***88,** 892–901 (2015).
223 16. Ebbert, M. T. W. *et al.* Long-read sequencing across the C9orf72 'GGGGCC' repeat
224 expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol
225 Neurodegeneration***13,** 46 (2018).
226 17. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas
227 System. *Cell***163,** 759–771 (2015).
228 18. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat
229 Methods* (2017). doi:10.1038/nmeth.4184
230 19. Boland, M. J. *et al.* Molecular analyses of neurogenic defects in a human pluripotent stem cell
231 model of fragile X syndrome. *Brain***140,** 582–598 (2017).
232 20. Gabrieli, T. *et al.* Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting
233 of chromosome segments (CATCH). *Nucl. Acids Res.***46,** e87–e87 (2018).