Comparative analysis of CRISPR/Cas9-targeted nanopore sequencing approaches in repeat expansion disorders

Louise Benarroch[1,*], Pierre-Yves Boëlle[2,*], Hélène Madry[3], Badreddine Mohand Oumoussa[3], Noboyuki Eura[4], Ichizo Nishino[4], Karim Labrèche[2], Guillaume Bassez[1,5], Tanya Stojkovic[1,5], Geneviève Gourdon[1], Gisèle Bonne[1], Stéphanie Tomé[1]

Affiliations

[1]Sorbonne Université, Inserm, Institut de Myologie, Centre de Recherche en myologie, Paris, France ; louise.benarroch@inserm.fr, guillaume.bassez@aphp.fr, stojkovic.tanya@aphp.fr, genevieve.gourdon@inserm.fr, g.bonne@institut-myologie.org, stephanie.tome@inserm.fr

[2]Sorbonne Université, UMS Pass, Center for Informatics-Bioanalysis Sorbonne Université (CINBIOS), Paris, France ; pierre-yves.boelle@upmc.fr, karim.labreche@upmc.fr

[3]Sorbonne Université, Inserm, UMS Pass, Plateforme Post-génomique de la Pitié-Salpêtrière (P3S), Paris, France ; badreddine.mohand_oumoussa@sorbonne-universite.fr, helene.madry@sorbonne-universite.fr

[4] Department of Neuromuscular Research, National Center of Neurology and Psychiatry (NCNP), Tokyo, Japan; neura@naramed-u.ac.jp, nishino@ncnp.go.jp

[5] Centre de Référence des Maladies Neuromusculaires Nord-Est/Île-de-France, Institut de Myologie, GHU Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Paris, France.

*Equally contributed

Corresponding authors: Louise Benarroch and Stéphanie Tomé

ABSTRACT

More than 50 repeat expansion disorders have been identified, with long-read sequencing marking a new milestone in the diagnosis of these disorders. Despite these major achievements, the

comprehensive characterization of short tandem repeats in a pathological context remains challenging, primarily due to their inherent characteristics such as motif complexity, high GC content, and variable length.

In this study, our aim was to thoroughly characterize repeat expansions in two neuromuscular diseases: myotonic dystrophy type 1 (DM1) and oculopharyngodistal myopathy (OPDM) using CRISPR/Cas9-targeted long-read sequencing (Oxford Nanopore Technologies, ONT). We conducted precise analyses of the DM1 and OPDM loci, determining repeat size, repeat length distribution, expansion architecture and DNA methylation, using three different basecallers (Guppy, Bonito and Dorado). We demonstrated the importance of the basecalling strategy in repeat expansion characterization. We proposed guidelines to perform CRISPR-Cas9 targeted long-read sequencing (no longer supported by ONT), from library preparation to bioinformatical analyses. Finally, we showed, for the first time, somatic mosaicism, hypermethylation of *LRP12* loci in symptomatic patients and changes in the repeat tract structure of OPDM patients.

We propose a strategy based on CRISPR/Cas9-enrichment long-read sequencing for repeat expansion diseases, which could be readily applicable in research but also in diagnostic settings.

Keywords

Myotonic dystrophy, oculopharyngodistal myopathy, repeat expansion, long-read sequencing, Oxford Nanopore Technology, basecallers

ACKNOWLEDGEMENTS

This work was funded by Association Institut de Myologie (AIM), INSERM and Sorbonne Université. L Benarroch was supported by Solve-RD project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257 and MDUK grant No. 18GROI-PG24-0140-2. This work was also supported by a grant from Fondation Maladies Rares N°GenOmics(2023)_041704

53

54 **INTRODUCTION**

55 Over the last 30 years, more than 50 repeat expansion disorders (REDs) have been identified. Fragile X
56 syndrome (FXS, OMIM #300624) and spinal and bulbar muscular atrophy (SBMA, OMIM #313200) were
57 the first REDs identified caused by a CGG expansion in the 5'-untranslated region (5'-UTR) of *FMR1*
58 gene [1–4] and a CAG expansion in the coding region of *AR* gene [5], respectively. Since then, the
59 exponential development of genetic tools and next generation sequencing technologies such as long-
60 read sequencing (LRS), mark a new milestone in RED diagnosis with the discovery of new REDs over
61 the last decade [6]. However, fully characterizing these expansions (number of motifs, expansion
62 architecture, methylation status, somatic mosaicism, etc.) remains challenging due to their inherent
63 complexity. In this study, we focused on two disorders, myotonic dystrophy type 1 (DM1) and
64 oculopharyngodistal myopathy (OPDM). DM1 is the most common adult muscular dystrophy,
65 characterized by multisystemic symptoms among which muscle weakness, myotonia, cataracts,
66 cardiac conduction defects, respiratory distress and neurological impairments. DM1 is caused by a
67 highly unstable CTG expansion in the 3'-UTR of *DMPK* gene and is one of the most complex REDs due
68 to genetic anticipation, age-dependent and tissue-specific somatic instability [7]. OPDM is a rare
69 neuromuscular disorder characterized by progressive ptosis, ophthalmoplegia, dysphagia and facial
70 and distal limb weakness. OPDM is due to a CGG repeat expansion in the 5'-UTR of five different genes
71 identified so far: *LRP12* (OPDM1) [8], *GIPC1* (OPDM2) [9], *NOTCH2NLC* (OPDM3) [10], *RILPL1* (OPDM4)
72 [11], and *ABCD3* gene (OPDM5) [12].

73 Despite the great effort to decipher the genetic cause of rare genetic disorders, only one third of them
74 have been solved using short-read sequencing. The low diagnostic rate can be explained in part by the
75 mutation localization within complex regions, inaccessible by short-read sequencing approaches such
76 as short tandem repeat regions [13]. LRS, now, give access to those regions, opening a new era for the
77 diagnosis of rare genetic diseases, especially for REDs. Two LRS approaches are available provided by
78 Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) which are slowly introduced into

79 the diagnostic field. Here, we used ONT's long-read approaches, particularly the targeted sequencing
80 approach using CRISPR/Cas9 strategy, to assess repeat expansions at DM1 and OPDM loci. Despite
81 available protocols provided by ONT and open access bioinformatic pipelines, we realized that
82 precisely characterizing expansion in our cases remained challenging. Indeed, inconsistency in
83 sequencing metrics (i.e. read length distribution, coverage, etc) as well as sequencing errors were
84 prone to appear depending on the library preparation protocol used and/or the basecalling strategies
85 apply to our raw data. The choice of the basecaller, to translate the changes of the raw electrical signal
86 into a nucleotidic sequence appeared to be critical when investigating tandem repeats. Therefore, we
87 tested 3 protocols (two from ONT and one homemade) to perform CRISPR/Cas9-enriched LRS in 4 DM1
88 patients and 4 OPDM patients. We then tested the performance of three different basecallers, Guppy,
89 Dorado and Bonito. And finally, we investigated the architecture of the repeat expansion by assessing
90 their structure and their methylation levels.
91
92 **RESULTS**
93 **CRISPR/Cas9 long-read sequencing metrics of DM1 and OPDM loci**
94 We performed long-read sequencing using ONT technologies on DNA extracted from 8 patients, 4 DM1
95 patients and 4 OPDM patients (Table 1). We compared the 2 available chemistries at the time, ONT
96 R9.4 and ONT R10.4 to perform CRISPR/Cas9-based long-read sequencing to target specifically the
97 DM1 locus (*DMPK* gene) and 3 OPDM loci (*LRP12*, *GIPC1* and *NOTCH2NLC* genes). Basecalling was
98 performed using Guppy basecaller, implemented in PromethION 24 sequencer. Coverage was
99 estimated using the number of total reads and also reads with a phred score equal or above 20. Phred
100 score is widely used to estimate the accuracy of the calling for each base, and we can interpret this
101 score as an estimation of error. For instance, a phred score of 20 (Q20) estimates 1% chance of error,
102 so 99% of accuracy. In sequencing data, Q20 or above is considered sufficient and we have used that
103 score as a quality cutoff throughout this study.
104

Patient ID		Mutated gene / Targeted loci	Repeat size (median)	DNA input	Coverage			Avg Q
					#reads	#reads Q20	%Q20	
Chemistry R9.4								
DM1	**EG**	*DMPK*	201	4µg	22	11	50%	20.4
	DV	*DMPK*	545	2µg	83	23	27.7%	17.3
	N18	*DMPK*	501	5.6µg	37	14	37.8%	18.25
	A2	*DMPK*	141	2µg	1161	560	48.2%	20.5
OPDM	**L1**	*LRP12*	127	2µg	285	107	37.5%	18.5
	G1	*GIPC1*	109	2µg	119	37	22.7	17.9
	N1	*NOTCH2NLC*	105	2µg	133	57	42.8%	18.9
Chemistry R10.4								
DM1	**A2 [ONT]**	*DMPK*	130	2µg	418	305	72.9%	30.7
	A2 [HM]	*DMPK*	129	2µg	398	277	69.6%	31.3
OPDM	**L2 [ONT]**	*LRP12*	130	2µg	144	131	91%	33.9
	G1 [ONT]	*GIPC1*	106	2µg	46	46	100%	38.1
	N1 [ONT]	*NOTCH2NLC*	104	2µg	142	132	92.9%	34.2

Table 1. *Long-read sequencing metrics for DM1 and OPDM patients. Coverage was determined by the number of total reads (#reads), number of reads with a mapping quality above 20 (#reads Q20) and the percentage of reads with a mapping quality above 20 (%Q20). Average quality (Avg Q) was assessed for all reads using a phred-scale quality score. In chemistry R10.4, three different protocols were used, ONT protocol [ONT] and home-made protocol [HM].*

Here, we showed that using R9.4 chemistry, less than 50% of reads reached Q20 cutoff with a mean score of 18.82 [17.3-20.5], regardless of the disease investigated. Chemistry R10.4 significantly improved reads quality, as expected, with more than 70% of total reads reached Q20 score with a mean quality score of 33.70 [30.7 – 38.1] (Mann-Whitney test, average quality [9.4] vs [10.4] p-value = 0.0012) (Table 1). Interestingly, the allele carrying the repeat expansion displayed a lower average quality than the normal allele with R9.4 chemistry, a difference that is lost with the R10.4 chemistry for OPDM loci but not for DM1 locus (Table 2). Indeed, even with the increase in the overall score

above 20, the DM1 expanded allele average score remained lower than the normal allele for A2 patient, suggesting an impact of the CTG repeat expansion on basecalling performance. Comparing patient A2 (DM1) and L2 patient (OPDM) carrying a comparable expanded allele size (130 repeats), a difference in average quality is observed (phred score 25.8 vs 33.2, respectively) suggesting that the type of expanded codon (CTG vs GCC) has an impact on sequencing quality, more than just the presence of the expansion.

Unfortunately, CRISRP/Cas9-enrichement protocol is not developed anymore by ONT. Therefore, we decided to set up a 'home-made' protocol using similar existing products combined with the adapter ligation steps of ONT ligation sequencing kit (see Material and Methods section). We tested our protocol on DM1 patient A2 and reached similar score than the ONT protocol, regarding coverage, the percentage of Q20 reads and their average quality (Table 1-2).

Patient ID		**Mutated gene**	**Normal allele**			**Expanded allele**		
			#reads	**%Q20**	**Avg Q**	**#reads**	**%Q20**	**Avg Q**
Chemistry R9.4								
DM1	**EG**	*DMPK*	13	61.5%	22.9	9	33.3%	16.9
	DV	*DMPK*	49	47%	21.1	34	0%	11.8
	N18	*DMPK*	22	63.6%	21.9	15	0%	12.9
	A2	*DMPK*	586	59.6%	22.4	575	36.7%	18.5
OPDM	**L1**	*LRP12*	162	52.5%	20.1	123	17.9%	16.3
	G1	*GIPC1*	57	50.9%	19.8	62	25%	16.2
	N1	*NOTCH2NLC*	75	41.3%	19	57	45.6%	18.7
Chemistry R10.4								
DM1	**A2 [ONT]**	*DMPK*	219	94.5%	36	199	49.2%	24.8
	A2 [HM]	*DMPK*	195	92.3%	36.9	203	47.8%	25.8
OPDM	**L2**	*LRP12*	74	89.2%	34.6	70	92.8%	33.2
	G1	*GIPC1*	28	100%	39.2	18	100%	36.3
	N1	*NOTCH2NLC*	72	95.8%	35.5	70	90%	32.9

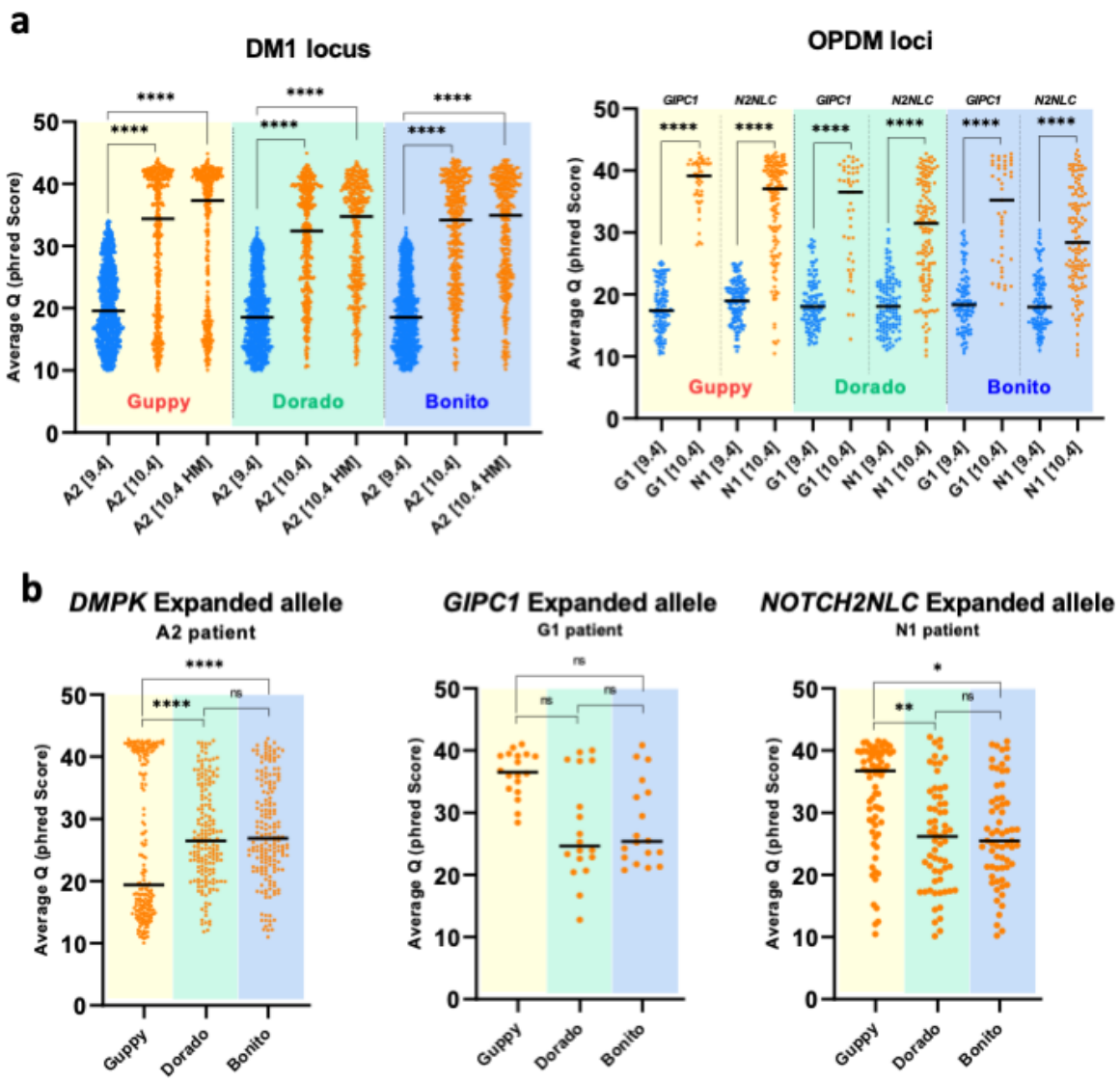
***Table 2**. LRS metrics on the normal and expanded alleles in DM1 and OPDM patients. For both allele, coverage was determined by the total number of reads (#reads) and the percentage of reads with a mapping quality above*

20 (%Q20). Average quality (Avg Q.) was assessed by calculating the mean quality for all reads specific of each allele. [ONT] ONT protocol and [HM] home-made protocol.

Basecaller comparison on average quality of Q20 reads of the expanded allele

Over the course of our study, ONT have replaced Guppy basecaller by Dorado, another open-sourced basecaller. We decided to compare the accuracy of these two basecallers, as well as a third open-sourced basecaller available called Bonito. We used data from patients analyzed with R10.4 chemistry as we saw a significant increase in reads' quality compared to R9.4 chemistry, independently of the basecaller and the pathology (Fig. 1a). By focusing on A2 DM1 patient, we showed a significant difference in the average quality of reads when comparing the three basecallers for the expanded allele where Dorado and Bonito basecallers showed the best average quality (Fig. 1b). Using the R10.4 HM protocol, A2 patient's reads of the expanded allele displayed an average quality score of 25.8 when using Guppy, that reached 29.8 and 30.4 by using Dorado and Bonito basecallers, respectively (Kruskal-Wallis test, [10.4 HM] Guppy *vs* Dorado, p-value < 0.0001; Guppy *vs* Bonito, p-value < 0.0001; Dorado *vs* Bonito, p-value > 0.99) (Fig. 1b and Table 3). No significant difference was observed for the normal allele (Supp Fig. 1a). Unlike DM1, Guppy displayed a better average quality score (phred score > 30) for both OPDM patients, G1 and N1 patients (Fig. 1b and Table 3). For G1 patient, reads targeting the expanded allele of *GIPC1* gene showed a higher average score with Guppy but the differences were not significant when comparing to Dorado and Bonito quality scores (Fig. 1b and Table 3). This result could be explained by the lower coverage (less than 50 reads) obtain for this patient (Table 1). For N1 patient, reads targeting the expanded allele of *NOTCH2NLC* gene displayed a significantly higher average quality score with Guppy than with Dorado and Bonito (Kruskal-Wallis test, [10.4] Guppy *vs* Dorado, p-value = 0.0033; Guppy *vs* Bonito, p-value = 0.0205; Dorado *vs* Bonito, p-value > 0.99) (Fig. 1b). No significant difference was observed between basecallers for the normal allele for *GIPC1* and *NOTCH2NLC* genes (Supp Fig. 1b-c).

159 Based on our results, we recommend R10.4 chemistry when investigating expansion diseases as well
160 as Dorado basecaller to investigate CTG repeat expansion and Guppy basecaller to investigate CGG
161 repeat expansion.
162



163

164 ***Figure 1.*** *Basecallers comparison on reads' average quality at DM1 and OPDM loci, with R10.4 chemistry.* ***(a)***
165 *Comparison of average quality of reads between R9.4 and R10.4 chemistry at DM1 (left panel) and OPDM (right*
166 *panel) loci. For all graphs, ONT R9.4 chemistry is represented with blue dots, ONT R10.4 chemistry (home-made*
167 *'HM') with orange dots. Average quality was estimated for the three basecallers: Guppy (yellow frame), Dorado*
168 *(green frame) and Bonito (blue frame). [N2NLC – NOTCH2NLC]* ***(b)*** *Expanded allele average quality using R10.4*

169 *chemistry. Kruskal-Wallis test, ns – not significant; [*] p-value < 0.05; [**] p-value < 0.01; [***] p-value < 0.001*
170 *and [****] p-value < 0.0001*
171

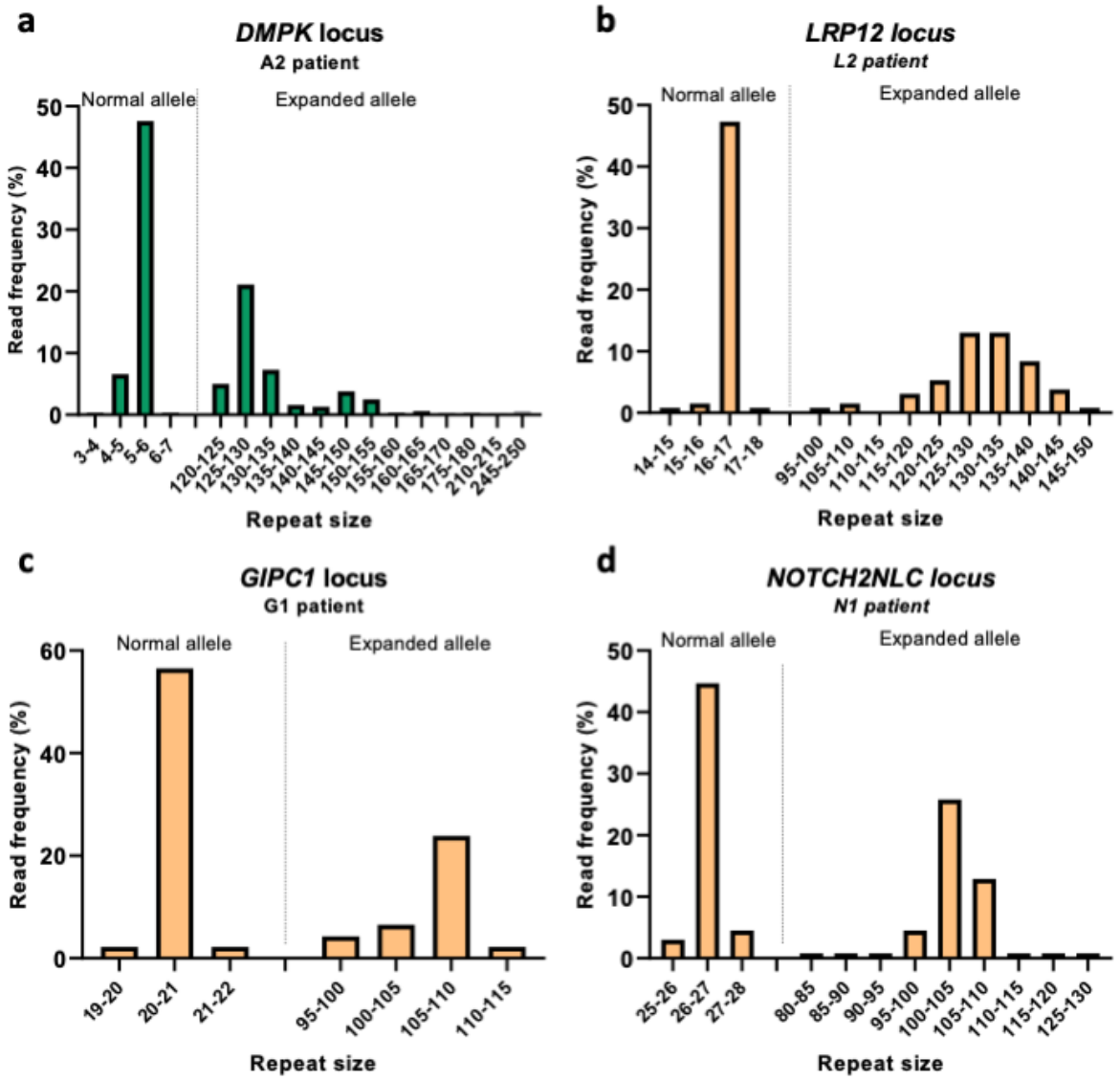
Expanded allele	Patient ID	Guppy		Dorado		Bonito	
		Avg Q	%Q20	Avg Q	%Q20	Avg Q	%Q20
DMPK	**A2 [ONT]**	24.8	49.2%	29.4	80.4%	30.2	86.7%
	A2 [HM]	25.8	47.8%	29.8	81.7%	30.4	84.5%
GIPC1	**G1**	36.3	100%	29.1	88.9%	28.1	100%
NOTCH2NLC	**N1**	32.9	90%	29.9	73.8%	29.6	78.1%

172
173 ***Table 3.*** *Basecaller comparison on average quality of Q20 reads of the expanded allele. Three base-callers were*
174 *tested: Guppy (yellow), Dorado (green) and Bonito (blue). Avg Q - Average quality (phred score); %Q20 -*
175 *percentage of Q20 reads on total expanded allele reads.*
176

177 **Estimation of repeat expansions length and somatic mosaicism**

178 DM1 A2 patient and other family members have been extensively investigated and well characterized
179 using different sequencing techniques: classical PCR, 3' triplet-primed-PCR, Single-Molecule Real Time
180 (SMRT) long read sequencing (Pacific Bioscience) and now ONT long-read sequencing [14–16]. From
181 here on out, A2 patient will be used as an experimental control to determine repeat expansion length,
182 somatic mosaicism, and the presence of interruptions. The repeat size frequency distribution showed
183 that this patient carried 5 CTG repeats (± 1 tract due to sequencing error) on the normal allele and
184 between 120 and 170 CTG motifs in the expanded allele, with a peak at 130, highlighting the somatic
185 mosaicism already described in this patient (Fig. 2a). For OPDM patients, similarly to the DM1 patient,
186 the repeat size frequency distribution of the normal allele peaks at 16, 20 and 26 for *LRP12*, *GIPC1* and
187 *NOTCH2NLC* genes (± 1 tract due to sequencing error), respectively (Fig. 2b-d). Regarding the expanded
188 allele, we obtained peaks at 130, 105 and 104 for *LRP12*, *GIPC1* and *NOTCH2NLC* genes, respectively,
189 ranging from 80 to 150 triplet repeats, highlighting for the first time somatic mosaicism in OPDM,

190 independent of the locus analysis (Fig. 2b-d). However; further investigations are required to fully
191 attest the somatic mosaicism, which is not the focus of this paper.
192



193
194 ***Figure 2.*** *Read length distribution at* ***(a)*** *DM1 and* ***(b-d)*** *OPDM loci. The x-axis displays the length of the repeats*
195 *detected (number of repeats, 3bp per repeat) for each locus:* ***(a)*** *DMPK,* ***(b)*** *LRP12,* ***(c)*** *GIPC1, and* ***(d)*** *NOTCH2NLC.*
196 *The normal and expanded alleles are separated by a striated line. The y-axis displays read frequency calculated*
197 *as a ratio of number of Q20 reads on total number of reads.*
198
199 **Changes in expansion sequence's architecture in OPDM patients**

200 Using CRISPR/Cas9-targeted LRS (R10.4 chemistry), a single CAG interruption in the expanded allele of
201 A2 patient was clearly identified as previously described [14] (Fig. 3a). This result validates ONT long-
202 read sequencing suitable for clearly determining sequence architecture changes in trinucleotide repeat
203 diseases, particularly in OPDM. At *LRP12* loci (patient L2), reads were aligned using a 'TAG' motif
204 located right before the start of the 'CCG' repeat tract in the reference sequence (Additional file 1). In
205 the normal allele, the repeat sequence is framed by two motifs, 5'-$(ACG)_2$-CCG-ACG-3' at the 5' end
206 and 5'-AGC-CAC-CGG-3' at the 3' end of the repeats. In the expanded allele, those two motifs are lost,
207 replaced by only one 'ACG' codon at the 5' end of the expansion and a single 'G' at the 3' end of the
208 expansion (Fig. 3b and Additional file 1). The loss of those motifs was also found in another *LRP12*
209 mutated patient (L1 patient, Additional file 1), supporting the fact that changes in *LRP12* sequence
210 structure is a consequence of the presence of the expansion. At *GIPC1* locus (G1 patient), reads were
211 aligned using 'CCC' motif, upstream of the 'GCC' expansion. In the normal allele, the repeat sequence
212 started with 5'-TCC-GCC-TCC-3' motif in the reference sequence that is lost in the expanded allele and
213 replaced by 5'-$(TCC)_4$-GCC-$(GCG)_3$-3' interruption (Fig. 3c and Additional file 1). Finally, *NOTCH2NLC*
214 (N1 patient) reads were aligned using 'CCA' motif, upstream of the 'GGC' expansion. In the normal
215 allele, 'GGC' repeats ended with 5'-$(GGA)_2$-$(GGC)_2$-3' motif whereas in the expanded allele, additional
216 'GGC' are inserted between the two 'GGA', creating a 5'-GGA-$(GGC)_3$-GGA-$(GGC)_5$-3' motif (Fig. 3d and
217 Additional file 1).
218 Thus, CRISPR/Cas9-targeted sequencing allowed us to highlight somatic mosaicism in OPDM but also
219 to identify changes in the structure of repeated tracks in the three OPDM genes, never described
220 before : *i)* in *LRP12* expansion, loss of two complex motifs associated with the expansion; *ii)* in *GIPC1*
221 expansion, presence of a 5'-$(TCC)_4$-GCC-$(GCG)_3$-3' interruption; and *iii)* in *NOTCH2NLC* expansion,
222 presence of 5'-GGA-$(GGC)_3$-GGA-$(GGC)_5$-3' interruptions.
223

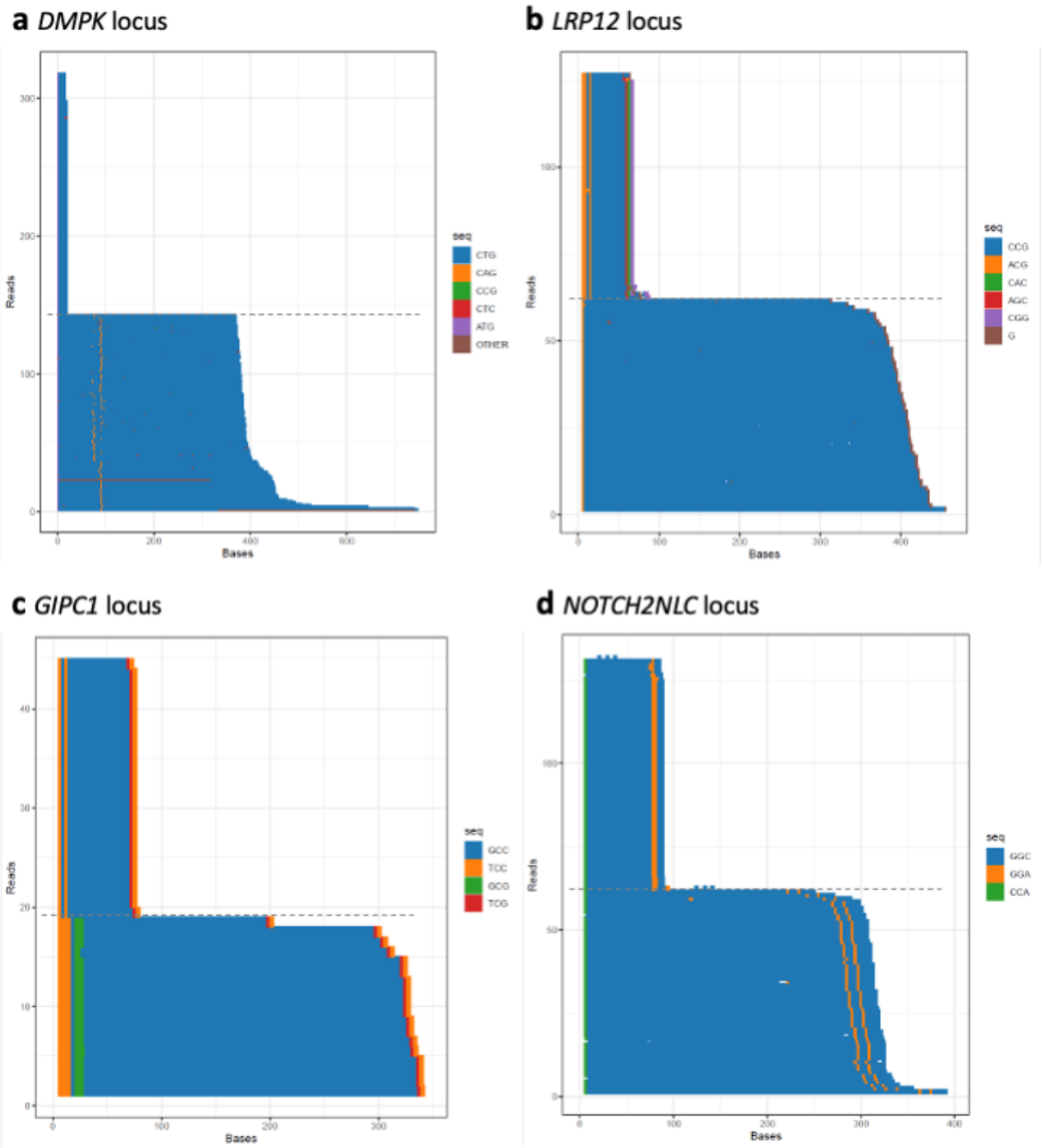
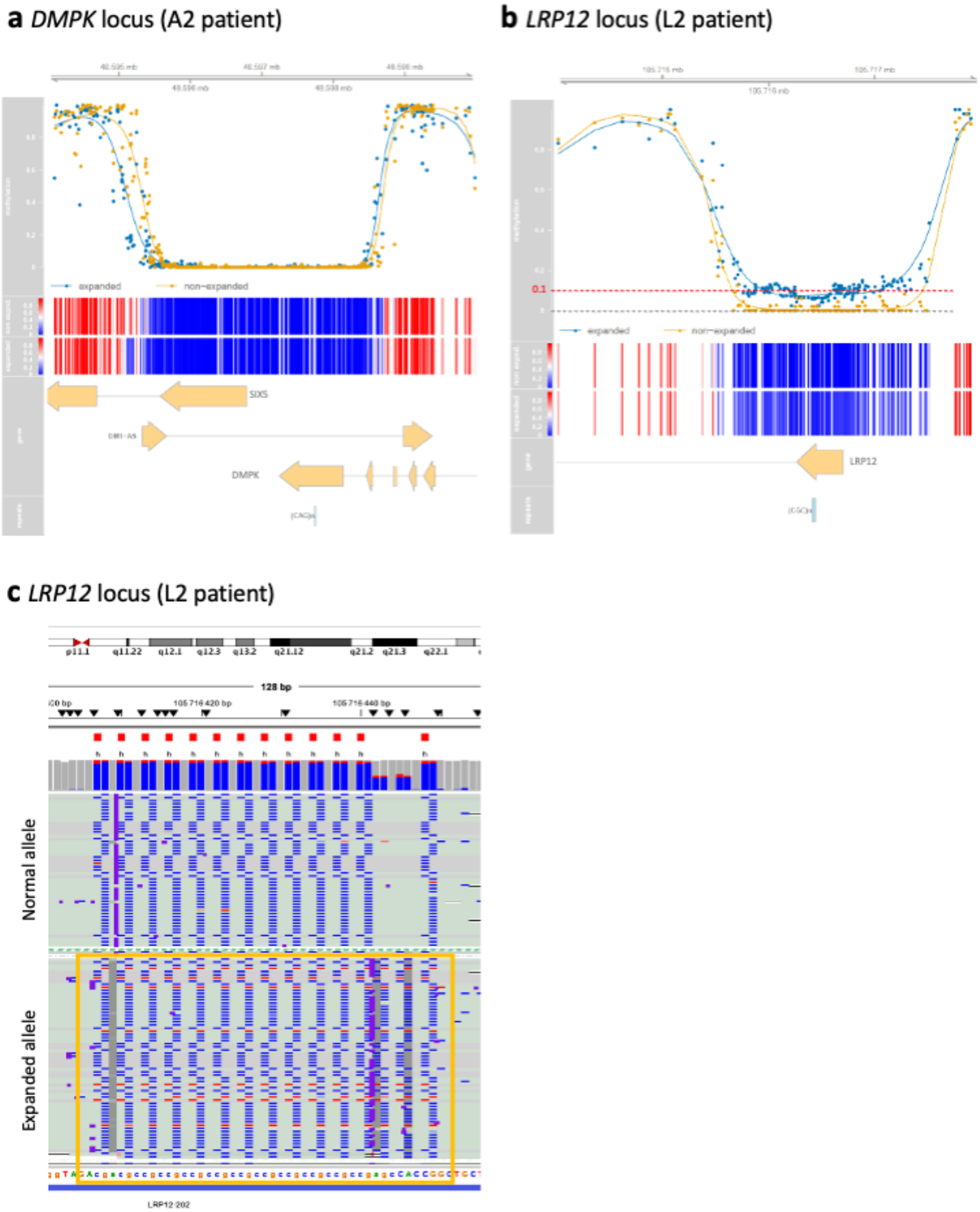


Figure 3. *Waterfall plots highlighting the expansion structures at* ***(a)*** *DM1 locus and* ***(b-d)*** *OPDM loci. For each waterfall, the x-axis shows the repeat expansion length in base pairs. The y-axis displays all the reads obtained after sequencing. The normal allele is above the striated line whereas the expanded allele is under the striated lines. The reference codon* ***(a)*** *CTG,* ***(b)*** *CCG,* ***(c)*** *GCC or* ***(d)*** *GGC is represented in blue; different colors are used to highlight codons within the repeat sequence that are not the reference codon, within the expansion.*

231 **DNA methylation profiling at DM1 and OPDM loci**

232 To assess DNA methylation pattern, we used Modkit, a methylation-calling pipeline to detect DNA
233 modifications not only within the flanking region but also within the expanded region. Several DNA
234 modifications can be identified using ONT-LRS such as 5-methylcytosine (5mC), 5-
235 hydroxymethylcytosine (5hmC) and N6-methyladenine (6mA), where 5mC is the most common
236 methylation mark. We did not see any changes in 5mC methylation profile for the DM1 A2 patient (Fig.
237 4a), and OPDM patients, G1 and N1 (Supp Fig. 2), either within the expanded region or in the flanking
238 regions. However, we observed hypermethylation in the *LRP12* locus (Fig. 4b) Indeed, we detected an
239 increase in methylation frequency in *LRP12* 5'UTR region, reaching approximatively 10%
240 hypermethylation that was present in the flanking region but also within the expanded region (Fig. 4c).
241 To our knowledge it is the first report of an existing hypermethylation in a symptomatic OPDM patient,
242 whatever the causing gene.

243



244

245 ***Figure 4.*** *5-methylcytosine (5mC) methylation levels at* ***(a)*** *DMPK and* ***(b)*** *LRP12 loci. Each graph displays, from*

246 *top to bottom, the genomic localization; the methylation levels where each dot represent one read on the normal*

247 *allele (yellow) or the expanded allele (blue); heatmap of methylation levels; gene(s) located in the region of*

248 *interest; and localization of the repeats.* ***(c)*** *IGV screenshot of the LRP12 expanded region, displaying the*

249 *methylation status of each read. The normal allele (top) and the expanded allele (bottom) are separated by a*

250 *dotted line. Each line represents one read. Unmethylated sites are colored in blue, methylated sites are colored*
251 *in red. The expanded region is surrounded by a yellow frame.*
252
253 The 5hmC modification is a product of the demethylation of 5mC-methylated site but also a novel
254 epigenetic regulator, however its functions have yet to be clearly established. It has been described to
255 play a role in several diseases but has never been studied in DM1 and OPDM. In this study we observed
256 that 5hmC profiles are similar between normal and expanded alleles at the DM1 and OPDM loci (Supp
257 Fig.3).
258
259 For the first time, we were able to detect 6mA modifications for the DM1 and OPDM loci. In all patient
260 tested, we observed a high variability in 6mA frequency compared to the 5mC frequency which is more
261 homogeneous. (Fig. 5a-d). The methylation frequency appeared to be high around the 5'-UTR region
262 of the three OPDM genes, *LRP12* (Fig. 5b), *GIPC1* (Fig. 5c) and *NOTCH2NLC* (Fig. 5d) and quite extensive
263 around the 3'-UTR region of *DMPK* (Fig. 5a). Although some variations are observed between normal
264 and expanded alleles for *LRP12* and *GIPC1*, it is difficult to conclude due to the high variability of the
265 6mA profiles.
266

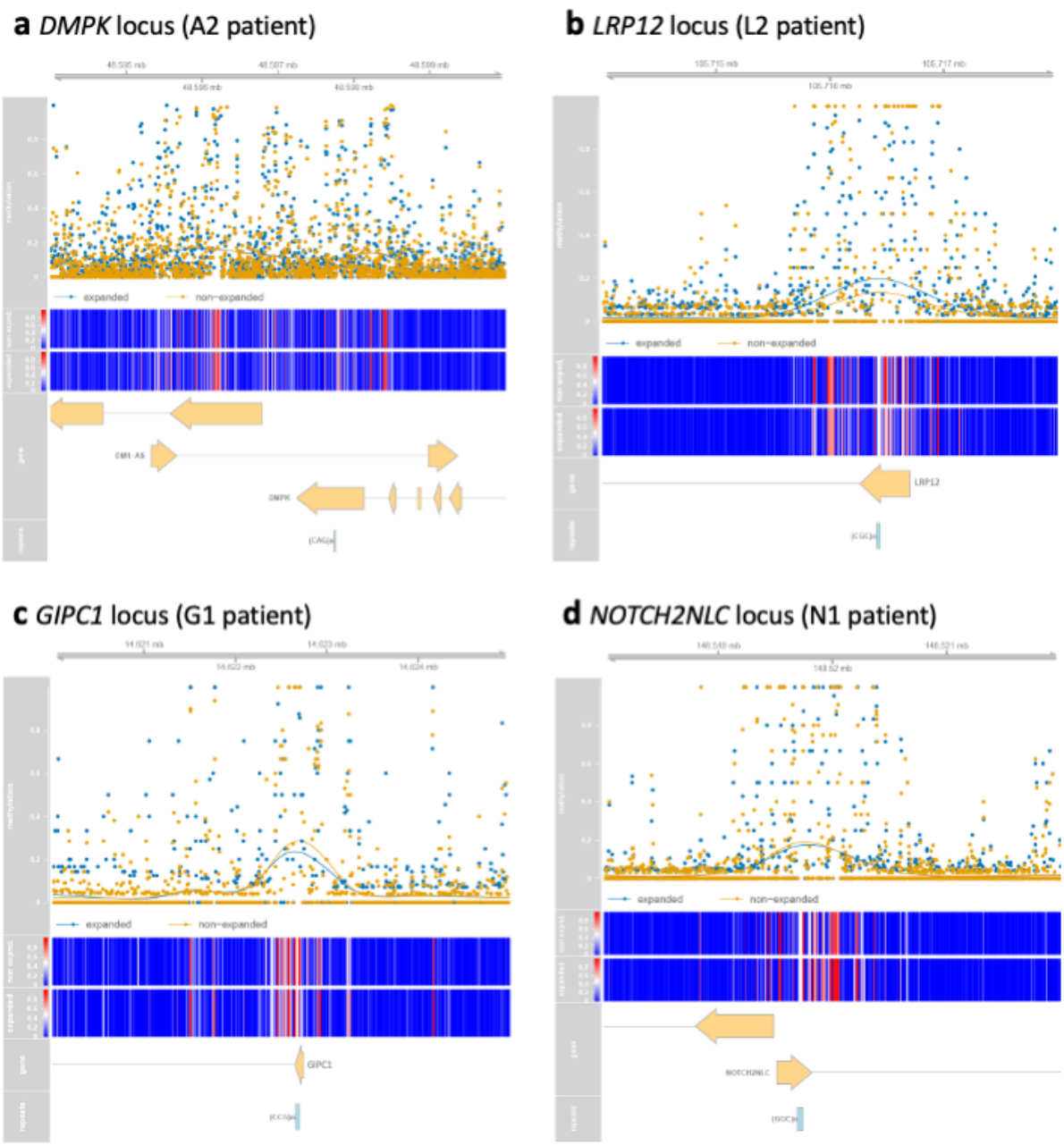


Figure 5. *N6-methyldeoxyadenosine (6mA) methylation levels at* ***(a)*** *DM1 and* ***(b-d)*** *OPDM loci. Each graph displays, from top to bottom, the genomic localization; the methylation levels where each dot represent one read on the normal allele (yellow) or the expanded allele (blue); heatmap of methylation levels; gene(s) located in the region of interest; and localization of the repeats.*

DISCUSSION

274 In this study we show that LRS is a suitable strategy to investigate REDs but requires some adjustments
275 depending on the disease of interest. We provided guidelines for DM1 and OPDM expanded loci
276 characterization and described the architecture of OPDM repeat tracts (presence of interruptions,
277 methylation status). As expected, the R10.4 chemistry significantly improved sequencing metrics on
278 expanded regions for both DM1 and OPDM patients. We compared two protocols for CRISPR/Cas9
279 library preparation (ONT provided and home-made) and we did not see any significant differences
280 between them. So, we propose here a suitable alternative to continue CRISPR/Cas9-enrichment using
281 the R10.4 chemistry, in a research setting. While ONT long-read sequencing shows considerable
282 promise in addressing the diagnostic challenges associated with genetic diseases [17], its application
283 on REDs in a diagnostic setting still requires improvements. Particularly, there are notable variations
284 observed between flowcells, impacting pore availability, coverage or multiplexing (data not shown)
285 and the bioinformatics tools. These inconsistencies represent significant obstacles to ensure the
286 reliability and reproducibility of LRS-based diagnostics for REDs. Basecalling is a critical step in the
287 sequencing workflow as it translates the electrical signal received by the pore into the DNA sequence,
288 hence all downstream investigations depend on it. ONT used 3 basecallers: Guppy, the basecaller
289 implemented to ONT sequencers, Dorado that replaced Guppy since 2023 in ONT sequencers and
290 finally Bonito, an open source basecaller written in Python similar to Dorado. As mentioned above, we
291 observed some discrepancies among the three basecallers' performances depending on the locus
292 investigated, especially between Guppy and Dorado. We concluded that Dorado or Bonito was better
293 suited for DM1 investigation while Guppy was better suited for OPDM investigation. These
294 discrepancies may be due to differences in each basecaller's computational processes to translate the
295 electrical signal of 'CTG' codon versus 'GGC' codon. To precisely determine the triplet repeat number,
296 the architecture of the expanded allele and flanking regions as well as the presence of somatic
297 mosaicism, we show here that the choice of basecaller is crucial and the combination of several of
298 them may also be needed in some cases [18]. Therefore, when investigating REDs, we highly
299 recommend evaluating several basecaller performances before making any conclusion [19,20].

300 ONT technologies allowed us to determine not only the size but also the structure of the repeat tracts
301 which can both have a huge impact on patients' prognosis and the clinical variability of the disease.
302 For instance, the presence of trinucleotidic interruption within the repeat tract can cause different
303 pathophysiological mechanisms underlying disease development and progression. In different REDs,
304 interruptions have been thoroughly investigated and these studies showed the biological impact of
305 these interruptions, either protective or aggravating, in disease clinical manifestations or progression.
306 In fragile X syndrome (FXS), 'AGG' codons that commonly interrupt the 'CGG' repeat expansion in
307 normal individuals, are less frequent in FXS patients causing transmission instability, hence increasing
308 the expansion risk. The presence of AGG interruption, within the expansion, act as a stabilizing factor
309 preventing instability of the repeats. Knowing the 'AGG' interruption size can help estimate the risk for
310 FXS intermediate individuals (45 to 54 CGG repeats) or pre-mutated individual (55 to 200 repeats) to
311 develop a full-mutation expansion (over 200 repeats) [21]. In DM1 patients, the presence of CCG, CTC
312 or CGG interruptions within the 'CTG' expansion can either be associated with additional atypical
313 neurological symptoms or can prevent or delay severe onset of the disease [22]. These interruptions
314 as well as an unique CAG interruption are been associated with the stabilization of the repeat through
315 unknown mechanisms [14,22–29]. In OPDM patients, changes in the expansion sequence have never
316 been described before, to our knowledge. However, it has been shown that in other *NOTCH2NLC*-
317 associated diseases, 'GGA' interruption within the 'GGC' repeat tract can cause different phenotypes
318 based on the number of repeats, going from Parkinson disease (less than 100 repeats) to dominant
319 muscle weakness (over 200 repeats) [30]. In our *NOTCH2NLC*-mutated patient (N1 patient), we did not
320 find any 'GGA' expansion as in the other *NOTCH2NLC*-associated disease but a change in the sequence
321 motif with the insertion of 'GGC' codons between the two 'GGA'. However, more *NOTCH2NLC*-
322 associated OPDM patients would be required to confirm this motif, and to attest if it specific to OPDM
323 patients carrying *NOTCH2NLC* mutation and investigate its potential role in the pathophysiological
324 mechanisms of the disease. Regarding *LRP12* and *GIPC1* expansions, no loss of motif or presence of

325 interruption have ever been reported, but as for *NOTCH2NLC* patients, additional patients would be
326 required to confirm these observations.
327 One of ONT technologies' strength is the direct detection of DNA or RNA modifications, during the
328 translocation of the molecule through the pore. Already available with R9.4 chemistry, R10.4 chemistry
329 allowed us to detect three type of modification: 5-methylcytosine (5mC), its oxidative counterpart 5-
330 hydroxymethylcytosine (5hmC) and N6-methyldeoxyadening (6mA), in the flanking regions as well as
331 within the expanded region. The 5mC mark is the most abundant, representing 70-80% of cytosine
332 preceding a guanidine (CpG) mainly located in promoter regions, into CpG islands [31]. DNA
333 methylation changes were observed in many diseases, among which OPDM. Interestingly in OPDM,
334 the epigenetic changes were only described in asymptomatic patients carrying longer expansion than
335 the disease-causing range, in *GIPC1*, *NOTCH2NLC, RILPL1 and ABCD3* genes [9,11,12,32,33]. As
336 expected, we did not see any changes in affected OPDM patients carrying either a *GIPC1* or
337 *NOTCH2NLC* expansion. However, we notice an increase in 5mC methylation levels (around 10%) in
338 *LRP12* patients that has never been reported before. We hypothesized that due to the expansion of a
339 'CCG' codon in OPDM patients, it may increase the methylation levels of the expanded region in
340 affected individuals but not sufficiently enough to have a protective impact, as the one observed in
341 asymptomatic carriers. Similar increase was observed in affected *NOTCH2NLC* patients (around 30% of
342 methylation) but the difference with healthy individuals was not significant and was associated with
343 an increase mRNA level in muscle but not in blood, supporting our hypothesis [34]. To go further we
344 investigated the 5hmC modification, a product of the demethylation of 5mC-methylated. It is
345 considered as a novel epigenetic regulator, but its functions have yet to be clearly established [35]. We
346 did not see any changes in 5hmC levels in any patients which was not surprising as it is lowly found in
347 mammalian tissues. Indeed, 5hmC content reach approximatively 0.1%, with the highest being in the
348 brain with 1% [35]. Finally, to our knowledge it is the first time that 6mA methylation have been
349 investigated in REDs. Preferentially found in prokaryote, 6mA modification represent 0.05% of the total
350 adenines in the human genome and are found preferentially in the brain, as 5hmC modifications. This

351 mark is enriched in coding region of transcriptionally active genes and mostly occur on [G/C]AGG[C/T]
352 motif [36]. In our study, we observed great variability of the methylation frequency at the different
353 loci, but the frequency did not seem significantly impacted by the presence of the expansion. Contrary
354 to what is described in the literature [36], the frequency of 6mA modifications seems higher in the 5'-
355 UTR region that in the intronic region, especially for OPDM loci. However, so little is known about 6mA
356 modification that is still too early to make any definite conclusions regarding the role of these
357 modifications in OPDM and DM1.
358
359 **CONCLUSION**
360 In this study we showed the importance of sequencing and bioinformatical strategies when studying
361 REDs. We provide guidelines and recommendation to investigate DM1 and OPDM loci that could
362 hopefully help in the diagnostic and the care of patients.
363
364 **METHODS**
365
366 **DNA samples**.
367 High molecular weight DNA from DM1 patients' blood or lymphoblastoid cell lines were extracted
368 either using Nanobing kit (PacBio) or using Chemagic 360 system (Revvity chemagen Technologie
369 Gmbh).
370 DNA samples from OPDM patients were kindly provided by Pr. Ichizo Nishino's group at the National
371 Center of Neurology and Psychiatry (Tokyo, Japan). DNA were extracted from blood samples from four
372 patients carrying expansion either in *LRP12* gene (L1 and L2 patients), *GIPC1* gene (G1 patient) or
373 *NOTCH2NLC* gene (N1 patient). For all DNA samples, quality and quantity were assessed using Qubit
374 fluorometer (Fisher Scientific) and the TapeStation System (Agilent).
375
376 **Cas9-enrichement long-read sequencing by Oxford Nanopore technologies**

377 Targeted long-read sequencing was performed on DNA samples from both DM1 and OPDM patients,
378 in a singleplex and/or multiplex setting. Three different protocols were tested, two from Oxford
379 Nanopore Technologies: Ligation Sequencing gDNA-Cas9 enrichment protocol (Chemistry R9.4; SQK-
380 CS9109 version: CAS_9106_v109_revD_16Sep2020), the Ligation sequencing kit V14 (Chemistry R10.4;
381 SQK-LSK114 version: GDE_9161_v114_revT_29Jun2022) and one "home-made" protocol (Chemistry
382 R10.4). For all three protocols, dephosphorylation of genomic DNA (gDNA), Cas9 cleavage and dA-
383 tailing were performed following the Cas9 enrichment protocol (SQK-CS9109, discontinued in March
384 2024, ONT). In brief, 2-5μg of gDNA was dephosphorylated, to prevent Cas9 off-target. In parallel,
385 Cas9-ribonucleoprotein complexes (Cas9-RNPs) were formed by pooling 100μM of each Alt-R crRNA
386 (Integrated DNA Technologies - IDT) specific of each targeted loci (*LRP12*-F: 5'-
387 TCTTTACCTCTTACGGCAAG-3'; *LRP12*-R: 5'-TCTGGTCCCGTAGAGTCCAC-3'; *GIPC1*-F: 5'-
388 GACGCTGTTGTCATTACTAT-3'; *GIPC1*-R: 5'-ACATCTAAGCTGCTCACCAT-3'; *NOTCH2NLC*-F: 5'-
389 GGAGACATCTTGGTGCATAC-3'; *NOTCH2NLC*-R: 5'-AAGCGCATTTGTACATGAGA-3'; *HTT*-F1: 5'-
390 TTTGCCCATTGGTTAGAAGC-3'; *HTT*-R1: 5'-GGACAAAGTTAGGTACTCAG-3'; *HTT*-F2: 5'-
391 TCTTATGAGTCTGCCCACTG-3'; *HTT*-R2: 5'-CTAGACTCTTAACTCGCTTG-3'; DMPK-5mc-US1: 5'-
392 ACCCAAGGCTCGCCCAATA-3'; DMPK-5mc-US2: 5'-GGGGAGAGCGGTACCACTTG-3'; DMPK-5mc-DS1: 5'-
393 GACGAGGTTACTTCAGACAT-3'; DMPK-5mc-DS2 [37]: 5'-GACCTGCGAGTCACACAAC-3'), 50 μM
394 transactivation crRNA (tracrRNA, IDT) and 62μM Alt-R S.p. HiFi Cas9 Nuclease V3 (IDT). The
395 dephosphorylated gDNA was then incubated with Cas9-RNPs, allowing Cas9 to cleave at the different
396 loci. Non-templated A nucleotides (dATP) were added to gDNA blunt ends formed by Cas9 cleavage
397 using TAQ polymerase, adding a dA-tail to prepare DNA for adapter ligation and sequencing. The
398 adapter ligation steps differ slightly when using the Cas9 sequencing kit V109 or the ligation sequencing
399 kit V114, the main difference being the adapter mix used, Adapter Mix (AMX) or Ligation Adapter (LA),
400 respectively. Adapters were added to the DNA thanks to the T4 ligase, either provided by the ligation
401 sequencing kit or by New England Biolabs (NEB). End-prepped DNA were purified using AMPure XP
402 Beads (Beckman Coulter) and eluted using the elution buffer provided by the Ligation sequencing kit.

403 Regarding multiplexing experiments, we follow ONT recommendation and used two protocols: Native
404 barcoding expansion 1-12 and 13-24 kit (EXP-NBD104) and Native Barcoding lot 24 V14 (SQK-
405 NBD114.24). The first steps are similar than a singleplex experiment, from gDNA dephosphorylation
406 step to Cas9 cleavage and dA-tailing step. Native barcodes were then ligated to the dA-tailed DNA,
407 using blunt/TA ligase Master Mix (NEB), then followed by the adapter ligation step. As mentioned
408 above, depending on which sequencing kit was used (V109 or V114), different adapter mix were used,
409 either "AMII" buffer for V109 kit or "NA" buffer for V114 kit.
410 Cas9-targeted sequencing protocol will be discontinued by ONT in 2024. Therefore, we adapted their
411 protocol using similar existing products. For the first steps (gDNA dephosphorylation to dA-tailing),
412 each ONT's buffers and enzymes can be replaced by NEB's products (Table 4). The adapter ligation step
413 was performed following the Ligation Sequencing kit V114 protocol.
414 Library quantification and quality controls were performed using the Qubit fluorometer (Fisher
415 Scientific) and the TapeStation System (Agilent). Libraries were sequenced using FLO-MIN106D (R9.4)
416 or FLO-MIN114 (R10.4) flow cell using PromehtION sequencer.
417

Steps	**Cas9-squencing kit (SQK-CS9109)**		**Ligation sequencing kit (SQK-LSK114)**		**Homemade**
	no barcoding	with barcoding	no barcoding	with barcoding	no barcoding
Cas9-RNPs complex	S. pyogenes Cas9 Alt-R® crRNAs (IDT) S. pyogenes Cas9 Alt-R® tracrRNAs (IDT) Alt-R® S.p. HiFi Cas9 Nuclease V3 (IDT) Nuclease-free duplex buffer (IDT)				
	RB (ONT)	RB (ONT)	RB (ONT)	RB (ONT)	NEB CutSmart Buffer (NEB)
gDNA dephosphorylation	RB (ONT)	RB (ONT)	RB (ONT)	RB (ONT)	NEB CutSmart Buffer (NEB)
	PHOS (ONT)	PHOS (ONT)	PHOS (ONT)	PHOS (ONT)	CIP (NEB)
Cas9 cleavage and dA-tailing	TAQ (ONT)	TAQ (ONT)	TAQ (ONT)	TAQ (ONT)	Taq polymerase (NEB)

	dATP (ONT)	dATP (ONT)	dATP (ONT)	dATP (ONT)	dATP (NEB)
Native barcoding ligation	-	NEB Blunt/TA ligase Master Mix (NEB)	-	NEB Blunt/TA ligase Master Mix (NEB)	-
	-	Fresh 70% Ethanol	-	Fresh 80% Ethanol	-
	-	Native Barcoding Expansion 1-12	-	Native Barcoding Kit 24 (V14)	-
	-	-	-	EDTA	-
Adapter ligation	LB (ONT)	LB (ONT)	LB (ONT)	LB (ONT)	LB (ONT)
	Adapter mix AMX (ONT)	Adapter mix AMII (ONT)	Adapter mix LA (ONT)	Adapter mix NA (ONT)	Adapter mix LA (ONT)
	T4 ligase	T4 ligase	NEBNext Quick T4 DNA Ligase (NEB)	NEBNext Quick T4 DNA Ligase (NEB)	NEBNext Quick T4 DNA Ligase (NEB)
	SFB (ONT)	SFB (ONT)	SFB (ONT)	SFB (ONT)	SFB (ONT)
	SPRI Dilution Buffer	SPRI Dilution Buffer	-		-
Elution	Elution Buffer	Elution Buffer	Elution Buffer	Elution Buffer	Elution Buffer

418

419 ***Table 4****. ONT long-read sequencing protocols. All reagents available using ONT ligation sequencing kit are*
420 *referred with (ONT); other reagents are obtained through New England Biolabs (NEB) or Integrated DNA*
421 *Technologies (IDT). RB – Reaction Buffer; PHOS – phosphatase; TAQ – Taq polymerase; LB – ligation buffer; SFB*
422 *– Short fragment buffer*

423

424 **Bioinformatic analyses**

425 FAST5 files were processed using Guppy, Dorado (v5.3), and Bonito (v7.0) employing the HAC model.
426 Then, FASTQ files were aligned to the T2T genome utilizing Minimap2. Reads overlapping genes of
427 interest were extracted via Samtools. To identify reads spanning the repeat region, local alignment
428 was performed with adjacent regions, each spanning 100 bases before and after the repeat site. All
429 retained reads were oriented in the same direction. The repeat region was precisely located by
430 identifying the two 100-base regions flanking the repeat, as taken from the T2T consensus sequences.

431 Reads exhibiting over 90% identity with the consensus on each flanking region and having average
432 quality larger than 20 were retained. The count of repeats was determined by dividing the total
433 number of bases in between the 2 flanking regions by the length of the repeat motif. We also used
434 straglr on the same set of reads. Methylation status was assessed using Modkit on the same reads
435 after transformation to POD5. Reads were stratified into 'expanded' and 'non-expanded' subgroups
436 based on the largest gap in length observed in the sorted list of read lengths. Visualization used R and
437 Gviz. Files are available on github (https://github.com/PYB-SU/STR).
438

439 **ABBREVIATIONS**

440 5hmc – 5-hydroxymethylcytosine; 5mc – 5-methylcytosine; 6mA – N6-methyladenine; Bp – base pair;
441 CpG – Cytosine preceding a guanidine; DM1 – Myotonic dystrophy; FXS – Fragile X syndrome; HM –
442 Homemade; LB – Ligation Buffer; LRS – Long-read sequencing; PacBio – Pacific Bioscience; ONT –
443 Oxford Nanopore Technologies; OPDM – Oculopharyngodistal myopathy; PHOS – Phosphatase; RB –
444 Reaction Buffer; RED – Repeat expansion disease; RNP – ribonucleoprotein complexes; SBMA – Spinal
445 and bulbar muscular atrophy; SFB – Short fragment buffer; TAQ – Taq polymerase; UTR – Untranslated
446 region
447

448 **ACKNOWLEDGEMENTS**

449 This work was funded by Association Institut de Myologie (AIM), INSERM and Sorbonne Université. L
450 Benarroch was supported by Solve-RD project which has received funding from the European Union's
451 Horizon 2020 research and innovation programme under grant agreement No 779257 and MDUK grant
452 No. 18GROI-PG24-0140-2. This work was also supported by a grant from Fondation Maladies Rares
453 N°GenOmics(2023)_041704
454

455 **REFERENCES**

456 1. Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, et al. Variation of the CGG repeat at
457 the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell. 1991;67:1047–
458 58.

459 2. Oberlé I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, et al. Instability of a 550-base pair DNA
460 segment and abnormal methylation in fragile X syndrome. Science. 1991;252:1097–102.

461 3. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, et al. Identification of a gene (FMR-1)
462 containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in
463 fragile X syndrome. Cell. 1991;65:905–14.

464 4. Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, et al. Mapping of DNA instability at the
465 fragile X to a trinucleotide repeat sequence p(CCG)n. Science. 1991;252:1711–4.

466 5. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations
467 in X-linked spinal and bulbar muscular atrophy. Nature. 1991;352:77–9.

468 6. Depienne C, Mandel J-L. 30 years of repeat expansion disorders: What have we learned and what
469 are the remaining challenges? Am J Hum Genet. 2021;108:764–85.

470 7. Tomé S, Gourdon G. DM1 Phenotype Variability and Triplet Repeat Instability: Challenges in the
471 Development of New Therapies. Int J Mol Sci. 2020;21:457.

472 8. Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, et al. Noncoding CGG repeat expansions in
473 neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease.
474 Nat Genet. 2019;51:1222–32.

475 9. Deng J, Yu J, Li P, Luan X, Cao L, Zhao J, et al. Expansion of GGC Repeat in GIPC1 Is Associated with
476 Oculopharyngodistal Myopathy. Am J Hum Genet. 2020;106:793–804.

477 10. Ogasawara M, Iida A, Kumutpongpanich T, Ozaki A, Oya Y, Konishi H, et al. CGG expansion in
478 NOTCH2NLC is associated with oculopharyngodistal myopathy with neurological manifestations. Acta
479 Neuropathol Commun. 2020;8:204.

480 11. Yu J, Shan J, Yu M, Di L, Xie Z, Zhang W, et al. The CGG repeat expansion in RILPL1 is associated
481 with oculopharyngodistal myopathy type 4. Am J Hum Genet. 2022;109:533–41.

482 12. Cortese A, Beecroft SJ, Facchini S, Curro R, Cabrera-Serrano M, Stevanovski I, et al. A CCG
483 expansion in ABCD3 causes oculopharyngodistal myopathy in individuals of European ancestry
484 [Internet]. medRxiv; 2023 [cited 2024 Jan 10]. p. 2023.10.09.23296582. Available from:
485 https://www.medrxiv.org/content/10.1101/2023.10.09.23296582v1

486 13. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. J Hum
487 Genet. 2020;65:11–9.

488 14. Tomé S, Dandelot E, Dogan C, Bertrand A, Geneviève D, Péréon Y, et al. Unusual association of a
489 unique CAG interruption in 5' of DM1 CTG repeats with intergenerational contractions and low
490 somatic mosaicism. Hum Mutat. 2018;39:970–82.

491 15. Mangin A, de Pontual L, Tsai Y-C, Monteil L, Nizon M, Boisseau P, et al. Robust Detection of
492 Somatic Mosaicism and Repeat Interruptions by Long-Read Targeted Sequencing in Myotonic
493 Dystrophy Type 1. Int J Mol Sci. 2021;22:2616.

494 16. Tsai Y-C, de Pontual L, Heiner C, Stojkovic T, Furling D, Bassez G, et al. Identification of a CCG-
495 Enriched Expanded Allele in Patients with Myotonic Dystrophy Type 1 Using Amplification-Free Long-
496 Read Sequencing. J Mol Diagn. 2022;24:1143–54.

497 17. Olivucci G, Iovino E, Innella G, Turchetti D, Pippucci T, Magini P. Long read sequencing on its way
498 to the routine diagnostics of genetic diseases. Front Genet. 2024;15:1374860.

499 18. LoTempio J, Delot E, Vilain E. Benchmarking long-read genome sequence alignment tools for
500 human genomics applications. PeerJ. 2023;11:e16515.

501 19. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore
502 sequencing. Genome Biol. 2019;20:129.

503 20. Pagès-Gallego M, de Ridder J. Comprehensive benchmark and architectural analysis of deep
504 learning models for nanopore sequencing basecalling. Genome Biology. 2023;24:71.

505 21. Latham GJ, Coppinger J, Hadd AG, Nolin SL. The role of AGG interruptions in fragile X repeat
506 expansions: a twenty-year perspective. Front Genet. 2014;5:244.

507 22. Santoro M, Masciullo M, Silvestri G, Novelli G, Botta A. Myotonic dystrophy type 1: role of CCG,
508 CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular diagnosis. Clin
509 Genet. 2017;92:355–64.

510 23. Musova Z, Mazanec R, Krepelova A, Ehler E, Vales J, Jaklova R, et al. Highly unstable sequence
511 interruptions of the CTG repeat in the myotonic dystrophy gene. Am J Med Genet A.
512 2009;149A:1365–74.

513 24. Braida C, Stefanatos RKA, Adam B, Mahajan N, Smeets HJM, Niel F, et al. Variant CCG and GGC
514 repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute
515 toward unusual symptoms in some myotonic dystrophy type 1 patients. Hum Mol Genet.
516 2010;19:1399–412.

517 25. Santoro M, Fontana L, Masciullo M, Bianchi MLE, Rossi S, Leoncini E, et al. Expansion size and
518 presence of CCG/CTC/CGG sequence interruptions in the expanded CTG array are independently
519 associated to hypermethylation at the DMPK locus in myotonic dystrophy type 1 (DM1). Biochim
520 Biophys Acta. 2015;1852:2645–52.

521 26. Botta A, Rossi G, Marcaurelio M, Fontana L, D'Apice MR, Brancati F, et al. Identification and
522 characterization of 5' CCG interruptions in complex DMPK expanded alleles. Eur J Hum Genet.
523 2017;25:257–61.

524 27. Cumming SA, Hamilton MJ, Robb Y, Gregory H, McWilliam C, Cooper A, et al. De novo repeat
525 interruptions are associated with reduced somatic instability and mild or absent clinical features in
526 myotonic dystrophy type 1. Eur J Hum Genet. 2018;26:1635–47.

527 28. Pešović J, Perić S, Brkušanin M, Brajušković G, Rakočević-Stojanović V, Savić-Pavićević D. Repeat
528 Interruptions Modify Age at Onset in Myotonic Dystrophy Type 1 by Stabilizing DMPK Expansions in
529 Somatic Cells. Front Genet. 2018;9:601.

530 29. Fontana L, Santoro M, D'Apice MR, Peluso F, Gori G, Morrone A, et al. Identification, molecular
531 characterization and segregation analysis of a variant DMPK pre-mutation allele in a three-generation
532 Italian family. Acta Myol. 2020;39:13–8.

533 30. Huang X-R, Tang B-S, Jin P, Guo J-F. The Phenotypes and Mechanisms of NOTCH2NLC-Related
534 GGC Repeat Expansion Disorders: a Comprehensive Review. Mol Neurobiol. 2022;59:523–34.

535 31. Angeloni A, Bogdanovic O. Enhancer DNA methylation: implications for gene regulation. Essays
536 Biochem. 2019;63:707–15.

537 32. Fukuda H, Yamaguchi D, Nyquist K, Yabuki Y, Miyatake S, Uchiyama Y, et al. Father-to-offspring
538 transmission of extremely long NOTCH2NLC repeat expansions with contractions: genetic and
539 epigenetic profiling with long-read sequencing. Clin Epigenetics. 2021;13:204.

540 33. Zeng Y-H, Yang K, Du G-Q, Chen Y-K, Cao C-Y, Qiu Y-S, et al. GGC Repeat Expansion of RILPL1 is
541 Associated with Oculopharyngodistal Myopathy. Ann Neurol. 2022;92:512–26.

542 34. Yu J, Deng J, Guo X, Shan J, Luan X, Cao L, et al. The GGC repeat expansion in NOTCH2NLC is
543 associated with oculopharyngodistal myopathy type 3. Brain. 2021;144:1819–32.

544 35. Kumar S, Chinnusamy V, Mohapatra T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and
545 Beyond. Front Genet. 2018;9:640.

546 36. Xiao C-L, Zhu S, He M, Chen D, Zhang Q, Chen Y, et al. N6-Methyladenine DNA Modification in the
547 Human Genome. Molecular Cell. 2018;71:306-318.e7.

548 37. Rasmussen A, Hildonen M, Vissing J, Duno M, Tümer Z, Birkedal U. High Resolution Analysis of
549 DMPK Hypermethylation and Repeat Interruptions in Myotonic Dystrophy Type 1. Genes (Basel).
550 2022;13:970.

551