GENERAL ARTICLE

Longitudinal increases in somatic mosaicism of the expanded CTG repeat in myotonic dystrophy type 1 are associated with variation in age-at-onset

Fernando Morales[1,*], Melissa Vásquez[1], Eyleen Corrales[1], Rebeca Vindas-Smith[1], Carolina Santamaría-Ulloa[1], Baili Zhang[2], Mario Sirito[2], Marcos R. Estecio[3], Ralf Krahe[2] and Darren G. Monckton[4]

[1]Instituto de Investigaciones en Salud (INISA), Universidad de Costa Rica, San José, Costa Rica, [2]Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, TX, USA, [3]Department of Epigenetics & Molecular Carcinogenesis, University of Texas MD Anderson Cancer Center, Houston, TX, USA and [4]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

*To whom correspondence should be addressed at: Ciudad Universitaria Rodrigo Facio, Instituto de Investigaciones en Salud, Universidad de Costa Rica, 2060, San Pedro de Montes de Oca, San José, Costa Rica, Tel: +506 2511 2124; Fax: +506 2511 5130; Email: fernando.moralesmontero@ucr.ac.cr

Abstract

In myotonic dystrophy type 1 (DM1), somatic mosaicism of the $(CTG)_n$ repeat expansion is age-dependent, tissue-specific and expansion-biased. These features contribute toward variation in disease severity and confound genotype-to-phenotype analyses. To investigate how the $(CTG)_n$ repeat expansion changes over time, we collected three longitudinal blood DNA samples separated by 8–15 years and used small pool and single-molecule PCR in 43 DM1 patients. We used the lower boundary of the allele length distribution as the best estimate for the inherited progenitor allele length (ePAL), which is itself the best predictor of disease severity. Although in most patients the lower boundary of the allele length distribution was conserved over time, in many this estimate also increased with age, suggesting samples for research studies and clinical trials should be obtained as early as possible. As expected, the modal allele length increased over time, driven primarily by ePAL, age-at-sampling and the time interval. As expected, small expansions <100 repeats did not expand as rapidly as larger alleles. However, the rate of expansion of very large alleles was not obviously proportionally higher. This may, at least in part, be a result of the allele length-dependent increase in large contractions that we also observed. We also determined that individual-specific variation in the increase of modal allele length over time not accounted for by ePAL, age-at-sampling and time was inversely associated with individual-specific variation in age-at-onset not accounted for by ePAL, further highlighting somatic expansion as a therapeutic target in DM1.

Introduction

Myotonic dystrophy type 1 (DM1) (OMIM #160900) is the most common hereditary muscular dystrophy in adults. It shows a highly variable phenotype mainly characterized by myotonia, muscle weakness and wasting, cardiac defects, cataracts and neurological manifestations (1,2). DM1 is a progressive, multi-system autosomal dominant disorder that affects individuals of all ages and both sexes (2). The DM1 mutation was identified in 1992 as an expansion of an unstable CTG trinucleotide repeat located in the 3′-untranslated region of the DM1 protein kinase (*DMPK*) gene (3–8). The repeat is polymorphic in the general population ranging from 5 to ∼37 repeats, while affected patients carry alleles from ∼50 up to several thousand repeats (4,6,8,9). Expansion size correlates positively with disease severity and negatively with the age-at-onset, but these correlations are relatively poor with little predictive value (10–13). Since its discovery, DM1 mutation has been shown to be highly unstable and biased toward expansion in the germline, providing the molecular basis for anticipation (i.e. the increase in the severity and decrease in the age-at-onset through successive generations) (14,15). Although parents of both sexes can transmit DM1, the sex and age of the parent are important in determining intergenerational amplifications of the mutation (16).

The repeat is also somatically unstable in an expansion-biased, age-dependent and tissue-specific manner (17–20), with repeat lengths being much larger in DNA of skeletal muscle tissue than blood (19,21–24). Molecular diagnosis of DM1 has traditionally been carried out using southern blot analysis of restriction-enzyme-digested genomic blood DNA, where the expanded allele usually presents as a broad smear instead of a discrete band. It is common practice to report the midpoint of the smear as the allele size of the patient (4,6,8,10,11,13,25–30). However, given that somatic mosaicism is expansion-biased and age-dependent, the number of repeats measured will be dependent on the age at which the DNA sample was obtained, confounding genotype-to-phenotype associations (9,31). By using small amounts of input DNA in small pool-PCR (SP-PCR), it has been possible to resolve the diffuse smear into its component alleles (19) and perform more detailed analysis of somatic mosaicism in DM1. In particular, by estimating the inherited or progenitor allele length (ePAL), it is possible to account for the confounding effects of somatic mosaicism and substantially improve genotype-to-phenotype correlations (9,31–33). Using SP-PCR data, it has been possible to link individual-specific variation in the rate of somatic expansion in blood DNA (after correcting for age-at-sampling and ePAL) with individual-specific variation in disease severity (after correcting for ePAL) (9,31).

These data strongly implicate somatic expansion in the pathogenesis, a notion further supported by the repeat-stabilizing and disease-moderating impact of variant repeat interruptions within the *DMPK* CTG repeat tract (31,33–37) as well as polymorphisms in the *MSH3* DNA mismatch repair gene (32,38). Given the recent similar findings in Huntington disease (HD), somatic expansion is being increasingly recognized as a novel therapeutic target in the repeat expansion disorders (38–43). As such, detailed analyses of the *DMPK* CTG repeat expansion in blood DNA are clearly warranted, in order to better understand the mutational dynamics and evaluate blood as a potential surrogate biomarker tissue. Although cross-sectional studies have been important in defining the basic features of somatic mosaicism, there are only limited data on the longitudinal progression of somatic mosaicism over time within patients (18,44). Here, we sought to provide new insights into the somatic mutational dynamics of the CTG repeat expansion during the lifetime of DM1 patients over an 8- to 15-year period.

Results

In order to further investigate the mutational dynamics of the expanded CTG repeat, we performed SP-PCR on three longitudinal blood DNA samples (time points 1, 2 and 3 (TP1, TP2 and TP3), obtained eight to 15 years apart) in 43 DM1 patients (Fig. 1). All participants involved in this study were tested for the presence of AciI-sensitive variant repeats within the CTG repeat tract, and none were identified (data not shown). As expected, we observed that both the range and modal allele length increased over time (Figs 1 and 2A). We have previously used the approximate lower boundary of the repeat length distribution to estimate the inherited progenitor allele length (ePAL), which has proven to be a highly accurate predictor of age-at-onset (9,31). After independently processing all DNA samples, we were able to determine that the lower boundary observed remained unchanged in 22 patients (51%, 22 of 43) over all three time points (Fig. 1A and B). In contrast, 21 patients (49%, 21 of 43) showed complex patterns of allele length distributions over time. Specifically, 11 patients (26%, 11 of 43) showed no obvious change between TP1 and TP2 but an apparent change between TP2 and TP3 (Fig. 1C). In three samples (7%, 3 of 43), the lower boundary was not obviously conserved between TP1 and TP2 but was between TP2 and TP3 (Fig. 1D). In seven patients (16%, 7 of 43), the lower boundary observed at TP1 was different at TP2 and at TP3 (Fig. 1E). These data suggest that estimating the PAL is best determined using the earliest possible blood DNA sample.

In order to further analyze the mutational dynamics of each DM1 sample over time, we used single-molecule SP-PCR to size at least 40 *de novo* expanded alleles per sample per time point (in total, 23 225 single-molecule products were sized) and performed a detailed quantitative and qualitative analysis of the somatic mosaicism over time in all 43 samples. Each sample could be broadly classified into three groups according to the type of allele length distributions: group 1, small expansions with clear expansion bias, but without obvious increase in modal allele over time (Supplementary Material, Fig. S1A and B); group 2, moderately sized expansions with clear expansion bias, with increasing modal allele length but decreasing frequency of the modal alleles over time (Supplementary Material, Fig. S1C and D); and group 3, large expansions, where the increase of the modal allele length was not as evident and the repeat seemed to be stabilized and/or there was a high frequency of contractions (Supplementary Material, Fig. S1E and F).

Previously, we have used single-molecule data to quantify the amount of somatic mosaicism by calculating the degree of somatic instability (SI) as the range between the 10th and 90th percentile of the total allele length distribution (9,32,45). This measure provides an estimate of the degree of variance of the allele length distribution. An alternative measure of somatic mosaicism that we have also used is to determine the change in modal allele length relative to inherited ePAL DM1 (31,32,38). This measure provides an estimate of the average increase in repeat length over a given time period. Here, we determined both measures of somatic mosaicism at each time point (TP1 to TP3). Although measuring two slightly different aspects of somatic mosaicism, these two measures of somatic mosaicism are nonetheless highly correlated with each other at all three time points ($r^2 = 0.78$, 0.77 and 0.75, $P < 0.0001$ at TP1, TP2 and
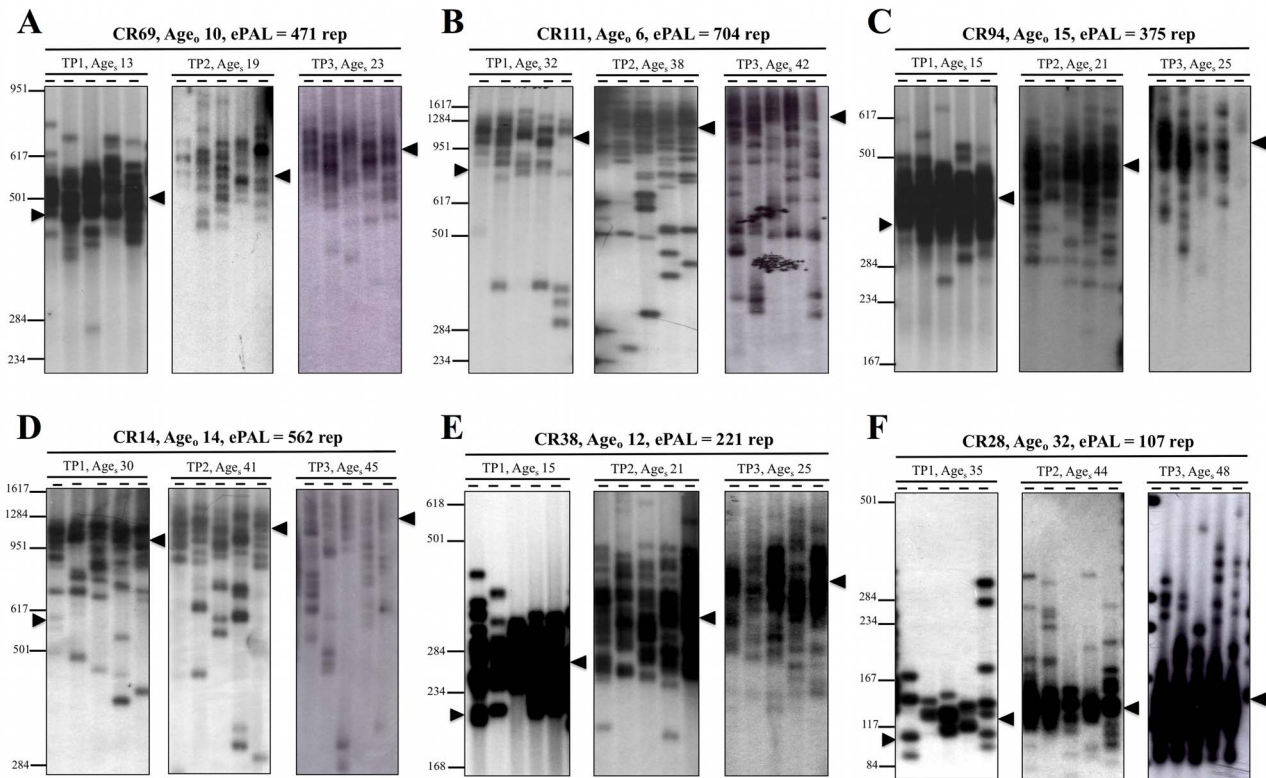
Figure 1. Representative small pool-PCR analyses of six Costa Rican DM1 patients sampled longitudinally at three different time points (TP1, TP2 and TP3 indicated by age in years). Since the longitudinal samples were collected and processed at different time points, samples were analyzed individually, and each autoradiograph represents the result at each time point analyzed. For comparisons, individual autoradiographs of each patient were size-matched according to the molecular weight marker. The approximate lower boundary of the allele length distribution at TP1 was used to estimate the progenitor allele length in DM1 blood DNA, indicated with an arrowhead at the left-hand side of the gel image. The lower boundary can be conserved through all three time points (**A** and **B**), only between two time points (**C** (TP1 to TP2) and **D** (TP2 to TP3)) or not obviously conserved through all three time points (**E**). (**F**) The change of the total allele length distribution from a skewed, to a more normal distribution with time. The arrowhead on the right-hand side of each gel image indicates the modal allele length, which as can be seen, increases with time. Modal allele length was estimated as the densest point of the allele length distribution across all of the replicate SP-PCRs for each participant at each time point. Some samples with large ePALs (e.g. CR111) showed a high frequency of contractions. The sizes of the molecular weight standard converted into the number of CTG triplet repeats is shown on the left. The identifier, ePAL size, age-at-onset and the age-at-sampling are shown for each sample. For each sample of the same DM1 patient, five replicate PCRs were performed with approximately 200–300 pg of DNA.
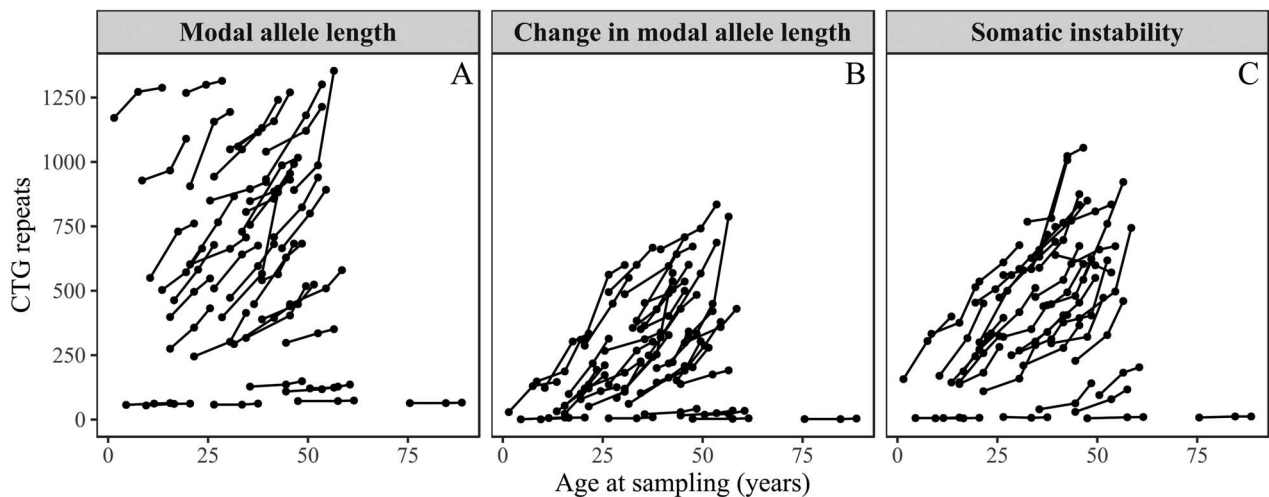
Figure 2. 'Spaghetti' plots for the modal allele length (**A**), change in modal allele length (**B**) and the degree of somatic instability (**C**). The plots show the data for all samples (each line represents a single DM1 patient) for the modal allele length, change in modal allele length and SI with age. The extremes of each line are time points 1 and 3 (TP1, TP3); the inflection point represents the data at time point 2 (TP2). The three variables increase over time.

Table 1. Changes in somatic mosaicism over time

Variables	Time points			Mean increment between time points		Paired *t*-test		
	TP1	TP2	TP3	TP2–TP1 (TG1)	TP3–TP2 (TG2)	TP1 vs TP2	TP2 vs TP3	TG1 vs TG2
Mean modal allele length change	181 (SD = 164)	268 (SD = 202)	342 (SD = 237)	87 (SD = 69)	73 (SD = 73)	$P = 2 \times 10^{-10}$	$P = 6 \times 10^{-8}$	$P = 0.279$
Mean degree of somatic instability	324 (SD = 225)	406 (SD = 256)	494 (SD = 280)	82 (SD = 75)	88 (SD = 63)	$P = 9 \times 10^{-9}$	$P = 2 \times 10^{-11}$	$P = 0.633$
Mean time difference between TPs	TP1 to TP2 (TG1) 7.72 years (SD = 1.76, range = 4–11 years)			TP2 to TP3 (TG2) 4.05 years (SD = 0.31, range = 4–6 years)		Paired *t*-test TG1 to TG2 $P = 2 \times 10^{-13}$		

TP = time point; TG = time gap; SD = standard deviation.

TP3, respectively) (Supplementary Material, Fig. S2). As expected, we also observed significant increases in mean modal length change between TP1 and TP2 (paired *t*-test, $P = 2 \times 10^{-10}$) and between TP2 and TP3 (paired *t*-test, $P = 2 \times 10^{-8}$) and mean SI between TP1 and TP2 (paired *t*-test, $P = 9 \times 10^{-9}$) and between TP2 and TP3 (paired *t*-test, $P = 2 \times 10^{-11}$) (Table 1, Fig. 2B and C). However, despite the fact that time gap 1 (TG1, TP1 to TP2) was significantly longer than time gap 2 (TG2, TP2 to TP3) (paired *t*-test, $P = 2 \times 10^{-13}$) (Table 1), there was no detectable difference between the change in modal allele length or SI between TG1 and TG2 (for modal allele length change, $P = 0.279$; for SI, $P = 0.633$) (Table 1).

In order to better understand the dynamics of somatic mosaicism over the short-to-medium period of time, we calculated the difference in the modal allele length between TP1 and TP3 and applied the previously derived multivariate regression model including age-at-sampling and ePAL (9,32,45), with the addition of the time interval (TG3 = TP3–TP1). These analyses revealed that approximately 89% of the variation in modal allele length change over time could be explained by an interaction between age-at-sampling, the time interval and ePAL and sex (Model 1a, Table 2). These results are consistent with our previous observation of a subtle effect of sex on somatic mutational dynamics (32). When we split the regression model between males and females, both models remained significant, with a slightly better fit for males ($r^2 = 0.90$, Model 1b, Table 2) than females ($r^2 = 0.87$, Model 1c, Table 2).

Interestingly, in some samples with large expansions, we observed an apparent accumulation of large contractions (e.g. Fig. 1B and D). In order to quantify these effects, we used the single-molecule data to determine the frequency of large contractions (defined as reductions in length of >10% relative to ePAL). Despite the high level of sampling error for any one estimate, the proportion of contractions detected was highly dependent on ePAL (TP1 $r^2 = 0.25$, $P = 0.0003$, Fig. 3A; TP2 $r^2 = 0.38$, $P = 7 \times 10^{-6}$, Fig. 3B; TP3 $r^2 = 0.42$, $P = 2 \times 10^{-6}$, Fig. 3C).

Next, we sought to determine if ongoing expansion of the CTG repeat could explain some of the variation in age-at-onset in DM1. To this end, we first performed a linear regression analysis to determine the relationship between ePAL and age-at-onset in DM1 using the previously described statistical model (9,32). The linear regression model indicated that the log transformation of ePAL explained approximately 74% (in a quadratic model)

of the variation in age-at-onset (Model 2a, Table 3). Previously, we demonstrated that the residual variation in SI was inversely correlated with the residual variation in the age-at-onset, that is, DM1 patients with higher-than-average SI had an earlier age-at-onset than average (9,32). In order to confirm if variation in the somatic mutational dynamics over time was related with variation in age-at-onset in DM1 patients, we performed a multivariate regression analysis that included the residual variation in modal allele length change over time as an independent variable. The statistical model showed an improvement of the fit of the model ($r^2 = 0.78$, with all independent variables being significant), suggesting a modifying role of the somatic mutational dynamics over time on the age-at-onset in DM1 (Model 2b, Table 3).

Finally, we considered if methylation flanking the CTG repeat expansion (previously implicated in the congenital form of DM1, CDM (46–49)) could be acting as a modifier of the somatic mutational dynamics of the CTG repeat. We performed a pyrosequencing-based methylation analysis (PMA) to interrogate multiple CpG sites upstream (11 CpG sites spanning CTCF-1) or downstream (six CpG sites spanning CTCF-2) of the CTG repeat (Fig. 4A), in order to determine the methylation levels and status in PBL DNA samples over time from all 43 DM1 patients. Samples were considered to be hypermethylated if the mean methylation levels in either the region upstream or downstream of the CTG repeat were >10% (45,50). After processing all samples, we found that the samples of only three DM1 patients across all three time points (9 of 129 (3 TPs × 43 DM1 patients), 7%) were hypermethylated upstream of the CTG repeat (Fig. 4B, Supplementary Material, Fig. S3). Hypermethylation downstream of the CTG repeat spanning CTCF-2 was only observed in the samples of the same three DM1 patients across almost all three time points (one sample was not hypermethylated at TP3) (Fig. 4B, Supplementary Material, Fig. S4). Since most of the samples were not hypermethylated downstream of the CTG repeat at TP1 or TP2 and hypermethylation has previously mainly been observed upstream of the CTG repeat (45,46,49), we only analyzed downstream methylation at TP3 in a subset number of DM1 samples (Fig. 4B, Supplementary Fig. S4). The three DM1 patients displaying hypermethylation were two CDM and one pediatric case (onset around 2 years of age). In the three patients with hypermethylation, the absolute levels of methylation upstream were not statistically different from those

Table 2. Multivariate regression models for the relationship between estimated progenitor allele length (ePAL), age-at-sampling (age$_{s\_TP1}$), time (TP1–TP3), sex and modal allele length change over time (MALC, mode TP3–mode TP1)

Model	Adjusted r^2	P	Parameter		Coefficient	Standard error	z-statistic	P
Model 1a: $\log(MALC) = \beta_0 + \beta_1$	0.887	< 0.001	Intercept	β_0	-10.37	1.98	-5.2	< 0.001
$\log(ePAL) + \beta_2 \log(ePAL)^2 + \beta_3$			$\log(ePAL)$	β_1	11.48	1.36	8.5	< 0.001
$\log(age_s) + \beta_4$ time $+ \beta_5$			$\log(ePAL)^2$	β_2	-2.30	0.28	-8.4	< 0.001
$(\log(ePAL){*}\log(Age){*}time) + \beta_6$			$\log(age_s)$	β_3	-1.35	0.41	-3.3	0.002
sex			Time	β_4	-0.14	0.05	-2.5	0.015
all individuals			Interaction	β_5	0.04	0.01	2.8	0.009
$n = 43$ DM1 patients			Sex	β_6	0.16	0.07	2.4	0.023
Model 1b: $\log(MALC) = \beta_0 + \beta_1$	0.907	< 0.001	Intercept	β_0	-5.62	2.96	-1.9	0.075
$\log(ePAL) + \beta_2 \log(ePAL)^2 + \beta_3$			$\log(ePAL)$	β_1	10.7	1.73	6.1	< 0.001
$\log(age_s) + \beta_4$ time $+ \beta_5$			$\log(ePAL)^2$	β_2	-2.35	0.36	-6.6	< 0.001
$(\log(ePAL){*}\log(Age){*}time)$			$\log(age_s)$	β_3	-2.59	0.69	-3.7	0.002
males only			Time	β_4	-0.41	0.12	-3.4	0.004
$n = 22$ DM1 patients			Interaction	β_5	0.10	0.03	3.3	0.004
Model 1c: $\log(MALC) = \beta_0 + \beta_1$	0.865	< 0.001	Intercept	β_0	-12.49	3.45	-3.6	0.003
$\log(ePAL) + \beta_2 \log(ePAL)^2 + \beta_3$			$\log(ePAL)$	β_1	12.02	2.01	5.9	< 0.001
$\log(age_s) + \beta_4$ time $+ \beta_5$			$\log(ePAL)^2$	β_2	-2.34	0.42	-5.5	< 0.001
$(\log(ePAL){*}\log(Age){*}time)$			$\log(age_s)$	β_3	-0.58	1.06	-0.6	0.590
females only			Time	β_4	-0.02	0.11	-0.2	0.848
$n = 21$ DM1 patients			Interaction	β_5	0.01	0.03	0.4	0.676

The table shows the squared coefficient of correlation (r^2) and statistical significance (P) for each model, and the coefficient, standard error, z-statistic and statistical significance (P), associated with each parameter in the model. The number of individuals used in each analyses is indicated (n). In the model, age at TP1 was used, and time was the difference between age at TP1 and TP3.
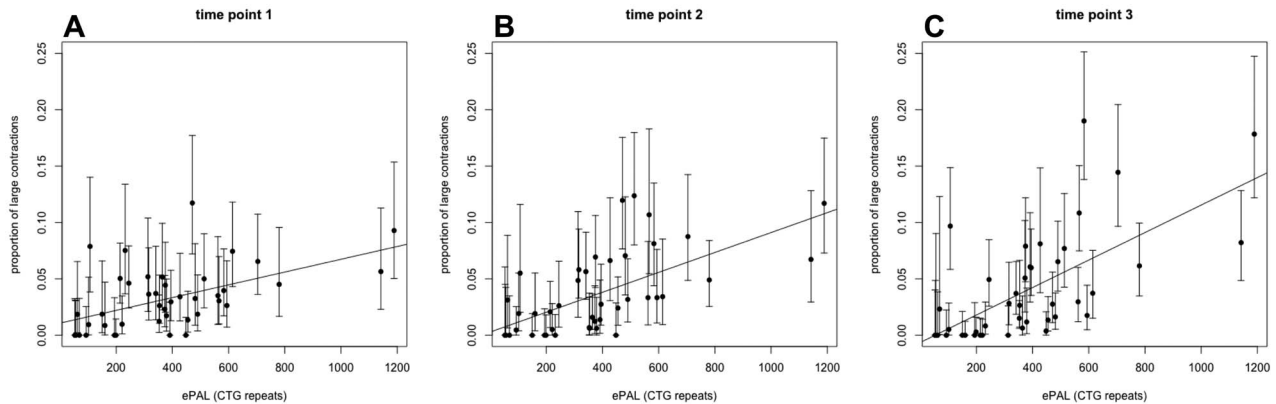

Figure 3. The proportion of large contractions increases with estimated progenitor allele length. The scatter plots show the proportion of large contractions (defined as contractions >10% the length of estimated progenitor allele length (ePAL)), plotted against the ePAL at each time point. The error bars are the 95% confidence intervals for each point estimate. The line represents the line of best fit for a regression analysis for the proportion of large contractions dependent on ePAL: time point 1 (**A**) $r^2 = 0.25$, $P = 0.0003$; time point 2 (**B**) $r^2 = 0.38$, $P = 7 \times 10^{-6}$ and time point 3 (**C**) $r^2 = 0.42$, $P = 2 \times 10^{-6}$.

downstream of the CTG repeat expansion (median$_{upstream}$ = 25.6 versus median$_{downstream}$ = 15.105, Mann–Whitney U rank sum test, $P = 0.178$, $n = 17$). In those individuals in which we observed hypermethylation, there was an apparent negative relationship between age-at-sampling and hypermethylation levels upstream and downstream of the CTG repeat (Fig. 4B, Supplementary Material, Fig. S5).

Discussion

It has been well documented that somatic mosaicism of the DM1 $(CTG)_n$ repeat expansion mutation is tissue-specific, expansion-biased and size- and age-dependent (9,17–21,23,51), all features that contribute to the observed variation in age-at-onset and the progressive nature of the disease (9,31,32). However, there is only limited data regarding the somatic mutational dynamics of the repeat expansion over time in the soma of DM1 patients. Here, by using SP-PCR we have performed a detailed qualitative and quantitative longitudinal analysis of the unstable CTG repeat over three points in 43 DM1 patients. As previously observed, we have confirmed that the modal allele length and range of alleles observed increases over time (18,20,31,44).

Previously, we have also shown that by defining the approximate lower boundary of the allele length distributions, we can estimate the inherited progenitor allele length (ePAL), and that by accounting for the confounding effects of age-at-sampling, this measure is the best predictor of disease severity (9,31–33). Here, we observed that this lower boundary was conserved

Table 3. Linear regression models of the relationship between estimated progenitor allele length (ePAL), the standardized residuals and age-at-onset (age_o).

Model	Adjusted r^2	P	Parameter	Coefficient	Standard error	t-statistic	P
Model 2a: $age_o = \beta_0 + \beta_1 \log(ePAL) + \beta_2 \log(ePAL)^2$ all individuals $n = 39$	0.736	< 0.001	Intercept β_0	255.97	62.22	4.1	< 0.001
			$\log(ePAL)$ β_1	-151.22	51.01	-2.9	0.005
			$\log(ePAL)^2$ β_2	22.60	10.38	2.2	0.036
Model 2b: $age_o = \beta_0 + \beta_1 \log(ePAL) + \beta_2 \log(ePAL)^2 + \beta_3$ standardized residuals (MALC model 1a) all individuals $n = 39$	0.772	< 0.001	Intercept β_0	259.05	57.80	4.5	< 0.001
			$\log(ePAL)$ β_1	-153.33	47.39	-3.2	0.003
			$\log(ePAL)^2$ β_2	22.95	9.65	2.4	0.023
			Standardized residuals β_3	-2.98	1.15	-2.6	0.014

The table shows the squared coefficient of correlation (r^2) and statistical significance (P) for each model, and the coefficient, standard error, t-statistic and statistical significance (P), associated with each parameter in the model. The number of individuals used in each analyses is indicated (n). Note that four asymptomatic patients were excluded from these age-at-onset analyses.

over time in about 50% of the patients analyzed (Fig. 1). In the remaining DM1 patients analyzed here, the lower boundary also increased over time concomitant with an increase in modal allele length. These results highlight that age-at-sampling in DM1 patients affects PAL estimation and confirms the value of sampling patients as early as possible before somatic expansion drives nearly all alleles beyond the inherited allele length (19,31). For the purposes of prospective research studies and clinical trials, it may in some cases prove possible to obtain historical blood DNA samples taken many years previously for diagnostic purposes (e.g. (31)).

In order to analyze the dynamics of somatic mosaicism further, we performed single-molecule SP-PCR to quantify the degree of somatic instability (range between the 10th and 90th percentile of the allele length frequency distribution thus derived (9,32,45)) at all three time points. The results confirm the step-wise gain of small numbers of repeats over time, exemplified by the shift to the right for allele length distributions (Supplementary Material, Fig. S1), the increase in modal allele length and the extent of somatic instability over time (18–20,31) (Figs 1 and 2, Supplementary Material, Fig. S1). However, contrary to the previous report by Martorell *et al.* (18), we observed that all patients with ePALs between 90 and 200 CTG repeats showed an increase in the modal allele length and the degree of somatic instability, even over a period of time as short as 4 years between samplings. This difference is very likely a consequence of the more sensitive detection method for measuring the degree of somatic instability and estimation of the PAL. In DM1 patients with ePALs < 90 CTG repeats ($n = 5$), the repeat remained very stable, with no obvious increases in modal allele length or the degree of somatic instability larger than 7 CTG repeats over a period of 10 years. These results indicate that the threshold over which changes in somatic mosaicism in blood DNA start to become readily detectable and lead to obvious increases in modal allele length using agarose gel-based SP-PCR in DM1 patients is between ~100 and 150 CTG repeats. However, alleles < 90 repeats are nonetheless detectably unstable, and it is very likely that using methods more sensitive to very small length changes, such as high-throughput amplicon sequencing (40), longitudinal changes in smaller expansions will be detectable.

In each case the modal CTG repeat length in blood DNA increased over time, indicating a clear net expansion bias. A sub-set of cells nonetheless acquired contractions (7.2%, 1691/23225). Many such contractions (49.6%, 839/1691) were relatively small, within ≤ 10% of ePAL. Such small contractions may well have arisen via small length change events that represent the reciprocal products of the small expansions that are assumed to drive net expansions over time (52) and are dependent on DNA mismatch repair (MMR) proteins (42,43,53). However, over 50% of the observed contractions were relatively large, > 10% of ePAL. Such large contractions were clearly much more frequent in larger alleles with higher ePALs (Fig. 3). Interestingly, a length-dependent increase in the frequency of somatic contractions has been observed in a transgenic HD mouse model carrying relatively large expansions (>300 CAG·CTG repeats) (54). In animal models with moderately sized CAG·CTG repeat expansions (<200 repeats), the loss of function of a variety of DNA MMR genes results in complete stabilization of the repeat (55–60). However, in studies using DM1 transgenes with 300–400 CTG repeats, MMR gene knockouts showed an apparent increase in the frequency of repeat contractions (55,58). Likewise, in both human fibrosarcoma and induced pluripotent stem cells, knock-down of MMR genes reduced repeat instability and appeared to

A

```
                                                         CTCF1-FS
                                                     CTCF1-F
CGCCTGCCAGTTCACAACCGCTCCGAGCGTGGGTCTCCGCCCAGCTCCAGTCCTGTGATCCG

GGCCCGCCCCCTAGCGGCCGGGGAGGGAGGGGCCGGGTCCGCGGCCGGCGAACGGGGCTCGA
       CTCF1-R
AGGGTCCTTGTAGCCGGGAATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGC

TGCTGCTGCTGCTGCTGCTGGGGGGATCACAGACCATTTCTTTCTTTCGGCCAGGCTGAGGC

          CTCF2-F
CCTGACGTGGATGGGCAAACTGCAGGCCTGGGAAGGCAGCAAGCCGGGCCGTCCGTGTTCCA
          CTCF2-FS
TCCTCCACGCACCCCCACCTATCGTTGGTTCGCAAAGTGCAAAGCTTTCTTGTGCATGACGC
                                                        CTCF2-R
CCTGCTCTGGGGAGCGTCTGGCGCGATCTCTGCCTGCTTACTCGGGAAATTTGCTTTTGCCA

AACCCGCTTTTTCGGGGATCCCGCGCCCCCCTCCTCACTTGCGCTGCTCTCGGAGCCCCAGC
```

B

Figure 4. DNA methylation analysis of CpGs flanking the DM1 $(CTG)_n$ repeat. (**A**) Location of the CTCF-binding sites and CpG sites interrogated flanking the DM1 $(CTG)_n$ expansion. The figure shows part of the sequence of exon 15 of the *DMPK* gene. There are two CTCF-binding sites (gray) flanking the DM1 CTG repeat (boxed). In bold and italics, the CpG sites interrogated by PMA at both CTCF sites. In all, we analyzed 11 sites upstream and six downstream of the CTG repeat. The location of the primers used to analyze both regions is also indicated. (**B**) Heatmap of methylation levels upstream and downstream of the CTG repeat expansion. The heatmap shows that hypermethylation was only detected in three patients (with very early age-at-onset) who carry the largest CTG expansions. Methylation levels, measured in 43 DM1 patients, sampled longitudinally three times through their lifetime (TP1, TP2 and TP3), are displayed in a white to red scale (0–100%). The levels shown correspond to the average methylation from 11 (for CTCF-1) and 6 (for CTCF-2) CpG sites analyzed by pyrosequencing-based methylation analysis (PMA). Gray boxes indicate that PMA was not carried out for the downstream region at TP3 in most participants. Each column of the heatmap represents the data from a single patient, who have been arranged in increasing order according to the size of the ePAL. The top-row annotation shows the color-encoded clinical subtype of each patient; the bottom-row annotation displays the size of the ePAL (53–1189 CTG repeats).

increase the frequency of contractions (61,62). These data indicate that natural DNA repair pathways in mammalian cells can mediate repeat contractions, appearing to be more important for larger alleles, and are likely to be at least partially DNA MMR independent.

Our data on modifiers of somatic mosaicism over time have confirmed the complex relationship between age and ePAL, in addition to the time interval between sampling, these factors overall accounting for ~89% of the observed variance in the increase in modal allele length between TP1 and TP3 (Table 2, Model 1a), very similar to the variance explained in our previous cross-sectional analyses (9,32). In addition, these data suggest subtle differences in the dynamics of somatic mosaicism between males and females (Table 2, Models 1b, c). Such an observation is consistent with our previous findings in the Costa Rican population (32) but is at odds with recent data in the European OPTIMISTIC cohort where no effect of sex on the

lifetime change in modal allele length was detected (31). However, these cohorts have notably different recruitment strategies. The Costa Rican cohort was family based, with all family members with a positive diagnosis offered the opportunity to participate, including those with very mild symptoms, and severely affected children. In contrast, OPTIMISTIC was restricted to individuals with moderate symptoms, and very mild, and very severe, cases were excluded (31). Because smaller expansions are more unstable and liable to expand in the male germline, ascertained DM1 families present with an excess of mildly affected transmitting grandfathers with relatively modest expansions (16,29,63). Indeed, although not statistically significant, in the Costa Rican DM1 cohort analyzed here, there is an excess of males relative to females (6:2) with ePALs <130 repeats. Additional studies will thus be needed to determine if apparent sex differences underlying somatic mosaicism in DM1 represent genuine population- or sex-specific effects or

the possible confounding effects of the age-at-sampling and ePAL biases that are inherent in DM1 families (16,29,63).

Although overall changes in modal allele length over time were well explained by age, ePAL and the time between samples, visual inspection of the data (e.g. Fig. 2A and B) suggests that the relative role of ePAL is not a simple linear correlation. Clearly, small alleles with ePALs <100 repeats show low or only very modest detectable increases in modal allele length. Alleles with ePALs >150 CTG repeats show clear increases in modal allele length over time. However, the rate of change of modal allele length with time does not obviously increase notably beyond alleles with ePALs >300 repeats (see the broadly parallel slopes for most of the larger alleles in Fig. 2B). Indeed, when considered in the context of total allele length, those alleles with the largest ePALs (top left in Fig. 2A) appear to have a slower relative rate of increase. These observations were derived by quantifying somatic instability using SP-PCR, which is ultimately limited by PCR efficiency (19). It is therefore possible that this is, at least in part, an artifact of amplification failure of very large expanded alleles. In order to confirm this observation, it would be necessary to use a more robust approach to measure very large expanded alleles, such as PCR-free PacBio sequencing (64) or optical mapping (65). Nonetheless, a reduction in the relative rate of expansion for larger alleles could be consistent with the allele length-dependent increase in contractions that we also observed and parallel similar non-linear length dependencies of the rate of expansion in a transgenic HD mouse model (54).

As demonstrated previously in cross-sectional studies, individual-specific differences in the rate of somatic expansion are associated with variation in DM1 disease severity not accounted for by ePAL (9,31,32). Here, we provide evidence that individual-specific variation in the longitudinal accumulation of expansion is also associated with variation in age-at-onset of DM1 (Table 3, Models 2a and 2b), confirming the observation that DM1 patients with higher-than-average somatic expansion rates have an earlier age-at-onset than average. Unfortunately, we do not have detailed longitudinal clinical data for the patients analyzed in this study; thus, we cannot test directly for an association between longitudinal changes in disease severity and the accumulation of expansions. Previously though, it has been demonstrated in cross-sectional studies that DM1 phenotypes that progressively decline such as muscle strength (e.g. shoulder abductor strength) and cognitive ability (e.g. trail making test) are best explained by an age/ePAL interaction (31,33) and that many progressive phenotypes are also associated with individual-specific variation in somatic expansion rates (31). Given the attractiveness of suppressing somatic expansion as a therapeutic target (41–43), additional longitudinal data relating somatic expansion rates and disease progression over time are clearly warranted, in order to further evaluate the best time windows for intervention and predict likely efficacy.

It is apparent that individual-specific variation in the rate of somatic expansion in DM1 is driven in part by *cis*-acting modifiers such as the presence of variant repeats within the *DMPK* CTG array (31,33–37) and by *trans*-acting factors such as polymorphisms in DNA repair genes (32,38). To investigate if somatic expansion rates might also be modified by *cis*-acting epigenetic factors, we determined if CpG site hypermethylation in two CTCF binding sites flanking the CTG repeat might contribute to longitudinal changes in somatic mosaicism over time. Previous studies on methylation of the flanking DNA in DM1 PBL have determined that hypermethylation is present mainly in DM1 congenital cases (47,48). We also recently showed that hypermethylation was also present in saliva DNA from congenital and very early-onset pediatric cases (45). In agreement with these studies, we observed hypermethylation in only two congenital and one pediatric case (with very early age-at-onset). The other 40 DM1 patients analyzed, all adult-onset cases, did not show hypermethylation at either TP1 or TP2 (methylation levels in these patients were <10%). Interestingly, in the three cases with hypermethylation, we observed an apparent age-dependent decrease in hypermethylation levels flanking the CTG repeat (Supplementary Material, Fig. S5). It is well established that some loci show age-dependent methylation changes, both increasing or decreasing depending on the specific locus (66–68). Further longitudinal studies involving more congenital and pediatric cases, and using more sensitive methods capable of quantifying low levels of methylation, will be required to confirm age-dependent changes in methylation status, better evaluate any role for methylation in modifying somatic mosaicism, and its potential pathogenic significance.

In conclusion, using a longitudinal analysis, we have confirmed that the mutant DM1 CTG repeat continues to expand throughout the lifetime of an individual. A more accurate prediction of the behavior of the CTG mutational dynamics over time will require a larger sample size and the development of more objective methods for determining the PAL, such as might be achieved using computational modeling (e.g. (52)) and/or access to very early blood DNA samples (e.g. samples taken at birth for population screening programs). A greater understanding of the genotype–phenotype relationship in DM1 should facilitate the stratification of patients based on the mutational behavior of their CTG repeat expansion in future clinical trials.

Materials and Methods

Patient population

The ethics committee of the Universidad de Costa Rica approved the project. All patients signed the informed consent in accordance with the ethical protocols approved by the Ethical Scientific Committee of the Universidad de Costa Rica. Previously, we had determined the size of the estimated progenitor allele length (ePAL, the size of the expanded allele transmitted from the parent to their offspring), the presence/absence of AciI-sensitive variant repeats, the modal allele length, the degree of somatic instability, age-at-sampling and age-at-onset in a large cohort of DM1 patients (9,32), defined here as time point one (TP1). Subsequently, we obtained additional blood DNA samples from 43 of these DM1 patients (22 males and 21 females) taken at two additional time points, i.e. time point two (TP2) and time point three (TP3), separated from 8 to 15 years in total (4 to 11 years from TP1 to TP2, and 4 to 6 years from TP2 to TP3) (Table 1). Here, we have now determined the modal allele length and the degree of somatic instability (SI) at the additional two time points TP2 and TP3. All samples were previously tested for the presence of AciI-sensitive CCG/CGG variant repeats, and none were detected.

Estimation of the modal allele length and quantification of somatic instability

By using SP-PCR as previously described (9,19), we were able to analyze a second and a third blood DNA sample from 43 Costa Rican DM1 patients. SP-PCR was carried out using 200–300 pg of input DNA per reaction. These data were used to estimate the lower boundary of the allele length distribution and estimate the modal allele length as the most dense point

of the allele length distribution at each time point, as previously described (9,19,31,33). In order to perform a detailed qualitative and quantitative analysis of the somatic mutational dynamics over time, we also used single-molecule SP-PCR (using 6–50 pg of input DNA per reaction) to determine the degree of somatic instability (difference between the 10th and 90th percentiles of the allele length distribution as described previously (9)) on each sample at TP2 and TP3. For this, we sized at least 40 single molecules per patient per time point (at least 40 single molecules for patients carrying an ePAL < 70 CTG repeats and at least 100 single molecules for patients carrying an ePAL ≥ 71 CTG repeats) (23 225 single molecules sized in total). We used UVI BANDMAP (UVITEC, Cambridge, UK) software to size the SP-PCR products. Supplementary Material, Table S1 contains clinical data and the molecular data for all patients at the three longitudinal time points.

Methylation analysis of CTCF-binding sites by pyrosequencing methylation analysis (PMA)

Analysis of methylation levels and status of the region flanking the DM1 $(CTG)_n$ repeat, including two CTCF-binding sites (CTCF-1, upstream and CTCF-2, downstream), was carried out by pyrosequencing-based methylation analysis (PMA) as previously described (45,50). Briefly, 300 ng of DNA from each DNA sample was used for bisulfite conversion using the EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions. For each sample, 15 ng of bisulfite-treated DNA was amplified using 100 pmoles of forward and reverse primer in a final volume of 25 µL containing 1X Hot StarTaq Master Mix (Qiagen, Hilden, Germany). PMA was carried out using the PyroMark Q96 MD platform (Qiagen) and PyroMark Gold Q96 reagents (Qiagen) according to the manufacturer's instructions, using the respective sequencing primer at 0.4 µM (Fig. 4A). PCR and PMA conditions have been published previously (45). In total, we interrogated 11 CpG sites upstream of the CTG repeat, of which seven are located within the CTCF1-binding site, and six downstream, three of which are located within the CTCF2-binding site (Fig. 4A). The results were analyzed with the PyroMark CpG 1.0.11 software (Qiagen), which calculates the mean methylation percentage (mC/(mC + C)) for each CpG site, allowing quantitative comparisons. The methylation index (MI) was calculated as the average value of mC/(mC + C) upstream (11 CpGs) or downstream (six CpGs) the CTG repeat.

Statistical analysis

Paired *t*-tests, Mann–Whitney *U* test, multivariate and linear regression analyses were carried out using the SPSS software package (v20), STATA (v13) or R (3.6.3) and R Studio (v1.2.5033). In the regression models (Tables 2 and 3), logarithmic (base 10) transformations and quadratic component of ePAL were included to normalize the distributions and reflect non-linear components of the associations, as described previously (9,32). Values of $P < 0.05$ were considered significant.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We would like to thank all DM1 patients for their participation in this study. We would also like to thank William Araya for his technical support during the execution of this project, and the DGM group at the University of Glasgow for the support provided in this project.

Conflict of Interest Statement. D.G.M. has been a scientific consultant and/or received an honoraria or stock options from Biogen Idec, AMO Pharma, Charles River, Vertex Pharmaceuticals, Triplet Therapeutics, LoQus23 and Small Molecule RNA and has had research contracts with AMO Pharma and Vertex Pharmaceuticals. The other authors declare no conflict of interest with the publication of this article.

References

1. Harper, P.S. (2001) *Myotonic Dystrophy*. Harcourt Publishers Ltd, London.
2. Harper, P.S. (1989) *Myotonic Dystrophy*. W B Saunders Company.
3. Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman, J., Wormskamp, N.G.M., Vooijs, M. *et al.* (1992) Cloning of the essential myotonic dystrophy region and mapping of the putative defect. *Nature*, **355**, 548–551.
4. Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J.P., Hudson, T. *et al.* (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell*, **68**, 799–808.
5. Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M., Riley, B. *et al.* (1992) Detection of an unstable fragment of DNA specific to individuals with Myotonic dystrophy. *Nature*, **355**, 547–548.
6. Fu, Y.H., Pizzuti, A., Fenwick, R.G., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T., de Jong, P. *et al.* (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, **255**, 1256–1258.
7. Harley, H.G., Brook, J.D., Rundle, S.A., Crow, S., Reardon, W., Buckler, A.J., Harper, P.S., Housman, D.E. and Shaw, D.J. (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature*, **355**, 545–546.
8. Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'hoy, K. *et al.* (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. *Science*, **255**, 1253–1255.
9. Morales, F., Couto, J.M., Higham, C.F., Hogg, G., Cuenca, P., Braida, C., Wilson, R.H., Adam, B., Del Valle, G., Brian, R. *et al.* (2012) Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum. Mol. Genet.*, **21**, 3558–3567.
10. Ashizawa, T., Dubel, J.R., Dunne, P.W., Dunne, C.J., Fu, Y.H., Pizzuti, A., Caskey, C.T., Boerwinkle, E., Perryman, M.B., Epstein,

H.F. *et al.* (1992) Anticipation in Myotonic dystrophy. 2. Complex relationships between clinical findings and structure of the GCT repeat. *Neurology*, **42**, 1877–1883.

11. Harley, H.G., Rundle, S.A., MacMillan, J.C., Myring, J., Brook, J.D., Crow, S., Reardon, W., Fenton, I., Shaw, D.J. and Harper, P.S. (1993) Size of the unstable CTG repeat sequence in relation to phenotype and parental transmission in myotonic dystrophy. *Am. J. Hum. Genet.*, **52**, 1164–1174.

12. Hunter, A., Tsilfidis, C., Mettler, G., Jacob, P., Mahadevan, M., Surh, L. and Korneluk, R. (1992) The correlation of age of onset with CTG Trinucleotide repeat amplification in Myotonic dystrophy. *J. Med. Genet.*, **29**, 774–779.

13. Redman, J.B., Fenwick, R.G., Fu, Y.H., Pizzuti, A. and Caskey, C.T. (1993) Relationship between parental trinucleotide GCT repeat length and severity of myotonic dystrophy in offspring. *JAMA*, **269**, 1960–1965.

14. Howeler, C.J., Busch, H.F.M., Geraedts, J.P.M., Neirmeijer, M.F. and Staal, A. (1989) Anticipation in myotonic dystrophy: fact or fiction? *Brain*, **112**, 779–797.

15. Ashizawa, T., Dunne, C.J., Dubel, J.R., Perryman, M.B., Epstein, H.F., Boerwinkle, E. and Hejtmancik, J.F. (1992) Anticipation in Myotonic dystrophy. 1. Statistical verification based on clinical and haplotype findings. *Neurology*, **42**, 1871–1877.

16. Morales, F., Vasquez, M., Cuenca, P., Campos, D., Santamaria, C., Del Valle, G., Brian, R., Sittenfeld, M. and Monckton, D.G. (2015) Parental age effects, but no evidence for an intrauterine effect in the transmission of myotonic dystrophy type 1. *Eur. J. Hum. Genet.*, **23**, 646–653.

17. Jansen, G., Willems, P., Coerwinkel, M., Nillesen, W., Smeets, H., Vits, L., Howeler, C., Brunner, H. and Wieringa, B. (1994) Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in (CTG)n repeat variation and selection against extreme expansion in sperm. *Am. J. Hum. Genet.*, **54**, 575–585.

18. Martorell, L., Monckton, D.G., Gamez, J., Johnson, K.J., Gich, I., de Munain, A.L. and Baiget, M. (1998) Progression of somatic CTG repeat length heterogeneity in the blood cells of myotonic dystrophy patients. *Hum. Mol. Genet.*, **7**, 307–312.

19. Monckton, D.G., Wong, L.J.C., Ashizawa, T. and Caskey, C.T. (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet.*, **4**, 1–8.

20. Wong, L.-J.C., Ashizawa, T., Monckton, D.G., Caskey, C.T. and Richards, C.S. (1995) Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *Am. J. Hum. Genet.*, **56**, 114–122.

21. Anvret, M., Ahlberg, G., Grandell, U., Hedberg, B., Johnson, K. and Edstrom, L. (1993) Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum. Mol. Genet.*, **2**, 1397–1400.

22. Ashizawa, T., Dubel, J.R., Tran, L., Harati, Y., Perryman, M.B. and Epstein, H.F. (1993) Muscle DNA shows different triplet expansion size from peripheral blood leukocyte DNA in myotonic dystrophy. *Neurology*, **43**, A 413–A 413.

23. Thornton, C.A., Johnson, K. and Moxley, R.T. (1994) Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Ann. Neurol.*, **35**, 104–107.

24. Zatz, M., Passos-Bueno, M.R., Cerqueira, A., Marie, S.K., Vainzof, M. and Pavanello, R.C.M. (1995) Analysis of the CTG repeat in skeletal muscle of young and adult myotonic dystrophy patients: when does the expansion occur. *Hum. Mol. Genet.*, **4**, 401–406.

25. Ashizawa, T., Anvret, M., Baiget, M., Barcelo, J.M., Brunner, H., Cobo, A.M., Dallapiccola, B., Fenwick, R.G., Grandell, U., Harley, H. *et al.* (1994) Characteristics of intergenerational contractions of the CTG repeat in myotonic dystrophy. *Am. J. Hum. Genet.*, **54**, 414–423.

26. Ashizawa, T., Dunne, P.W., Ward, P.A., Seltzer, W.K. and Richards, C.S. (1994) Effects of the sex of myotonic dystrophy patients on the unstable triplet repeat in their affected offspring. *Neurology*, **44**, 120–122.

27. Brunner, H.G., Bruggenwirth, H.T., Nillesen, W., Jansen, G., Hamel, B.C.J., Hoppe, R.L.E., de Die, C.E.M., Howeler, C.J., van Oost, B.A., Wieringa, B. *et al.* (1993) Influence of sex of the transmitting parent as well as of parental allele size on the CTG expansion in myotonic dystrophy (DM). *Am. J. Hum. Genet.*, **53**, 1016–1023.

28. Cobo, A.M., Poza, J.J., Martorell, L., de Munain, A.L., Emparanza, J.I. and Baiget, M. (1995) Contribution of molecular analyses to the estimation of the risk of congenital myotonic dystrophy. *J. Med. Genet.*, **32**, 105–108.

29. Lavedan, C., Hofmann-Radvanyi, H., Shelbourne, P., Rabes, J.-P., Duros, C., Savoy, D., Dehaupas, I., Luce, S., Johnson, K. and Junien, C. (1993) Myotonic dystrophy: size and sex dependent dynamics of CTG meiotic instability, and somatic mosaicism. *Am. J. Hum. Genet.*, **52**, 875–883.

30. Lopez de Munain, A., Cobo, A.M., Huguet, E., Marti Masso, J.F., Johnson, K. and Baiget, M. (1994) CTG trinucleotide repeat variability in identical twins with myotonic dystrophy. *Ann. Neurol.*, **35**, 374–375.

31. Cumming, S.A., Jimenez-Moreno, C., Okkersen, K., Wenninger, S., Daidj, F., Hogarth, F., Littleford, R., Gorman, G., Bassez, G., Schoser, B. *et al.* (2019) Genetic determinants of disease severity in the myotonic dystrophy type 1 optimistic cohort. *Neurology*, **93**, e995–e1009.

32. Morales, F., Vasquez, M., Santamaria, C., Cuenca, P., Corrales, E. and Monckton, D.G. (2016) A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair (Amst)*, **40**, 57–66.

33. Overend, G., Legare, C., Mathieu, J., Bouchard, L., Gagnon, C. and Monckton, D.G. (2019) Allele length of the DMPK CTG repeat is a predictor of progressive myotonic dystrophy type 1 phenotypes. *Hum. Mol. Genet.*, **28**, 2245–2254.

34. Braida, C., Stefanatos, R.K., Adam, B., Mahajan, N., Smeets, H.J., Niel, F., Goizet, C., Arveiler, B., Koenig, M., Lagier-Tourenne, C. *et al.* (2010) Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum. Mol. Genet.*, **19**, 1399–1412.

35. Cumming, S.A., Hamilton, M.J., Robb, Y., Gregory, H., McWilliam, C., Cooper, A., Adam, B., McGhie, J., Hamilton, G., Herzyk, P. *et al.* (2018) De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. *Eur. J. Hum. Genet.*, **26**, 1635–1647.

36. Pesovic, J., Peric, S., Brkusanin, M., Brajuskovic, G., Rakocevic-Stojanovic, V. and Savic-Pavicevic, D. (2018) Repeat interruptions modify age at onset in myotonic dystrophy type 1 by stabilizing DMPK expansions in somatic cells. *Front. Genet.*, **9**, 601.

37. Tome, S., Dandelot, E., Dogan, C., Bertrand, A., Genevieve, D., Pereon, Y., group, D.M.c.s, Simon, M., Bonnefont, J.P., Bassez,

G. *et al.* (2018) Unusual association of a unique CAG interruption in 5′ of DM1 CTG repeats with intergenerational contractions and low somatic mosaicism. *Hum. Mutat.*, **39**, 970–982.

38. Flower, M., Lomeikaite, V., Ciosi, M., Cumming, S., Morales, F., Lo, K., Hensman Moss, D., Jones, L., Holmans, P., Investigators, T.-H. *et al.* (2019) MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain*, **142**, 1876–1886.

39. Genetic Modifiers of Huntington's Disease Consortium. Electronic address, g.h.m.h.e. and Genetic Modifiers of Huntington's Disease, C (2019) CAG repeat not polyglutamine length determines timing of Huntington's disease onset. *Cell*, **178**, 887–900 e814.

40. Ciosi, M., Maxwell, A., Cumming, S.A., Hensman Moss, D.J., Alshammari, A.M., Flower, M.D., Durr, A., Leavitt, B.R., Roos, R.A.C., team, T.-H. *et al.* (2019) A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine*, **48**, 568–580.

41. Gomes-Pereira, M. and Monckton, D.G. (2006) Chemical modifiers of unstable expanded simple sequence repeats: what goes up, could come down. *Mutat. Res.*, **598**, 15–34.

42. Khristich, A.N. and Mirkin, S.M. (2020) On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.*, **295**, 4134–4170.

43. Massey, T.H. and Jones, L. (2018) The central role of DNA damage and repair in CAG repeat diseases. *Dis. Model. Mech.*, **11**, 1–14.

44. Martorell, L., Martinez, J.M., Carey, N., Johnson, K. and Baiget, M. (1995) Comparison of CTG repeat length expansion and clinical progression of myotonic dystrophy over a five year period. *J. Med. Genet.*, **32**, 593–596.

45. Corrales, E., Vasquez, M., Zhang, B., Santamaria-Ulloa, C., Cuenca, P., Krahe, R., Monckton, D.G. and Morales, F. (2019) Analysis of mutational dynamics at the DMPK (CTG)n locus identifies saliva as a suitable DNA sample source for genetic analysis in myotonic dystrophy type 1. *PLoS One*, **14**, e0216407.

46. Barbe, L., Lanni, S., Lopez-Castel, A., Franck, S., Spits, C., Keymolen, K., Seneca, S., Tome, S., Miron, I., Letourneau, J. *et al.* (2017) CpG methylation, a parent-of-origin effect for maternal-biased transmission of congenital Myotonic dystrophy. *Am. J. Hum. Genet.*, **100**, 488–505.

47. Filippova, G.N., Thienes, C.P., Penn, B.H., Cho, D.H., Hu, Y.J., Moore, J.M., Klesert, T.R., Lobanenkov, V.V. and Tapscott, S.J. (2001) CTCF-binding sites flank CTG/CAG repeats and form a methylation- sensitive insulator at the DM1 locus. *Nat. Genet.*, **28**, 335–343.

48. Steinbach, P., Glaser, D., Vogel, W., Wolf, M. and Schwemmle, S. (1998) The DMPK gene of severely affected myotonic dystrophy patients is hypermethylated proximal to the largely expanded CTG repeat. *Am. J. Hum. Genet.*, **62**, 278–285.

49. Lopez Castel, A., Nakamori, M., Tome, S., Chitayat, D., Gourdon, G., Thornton, C.A. and Pearson, C.E. (2011) Expanded CTG repeat demarcates a boundary for abnormal CpG methylation in myotonic dystrophy patient tissues. *Hum. Mol. Genet.*, **20**, 1–15.

50. Colella, S., Shen, L., Baggerly, K.A., Issa, J.P. and Krahe, R. (2003) Sensitive and quantitative universal pyrosequencing methylation analysis of CpG sites. *BioTechniques*, **35**, 146–150.

51. Martorell, L., Johnson, K., Boucher, C.A. and Baiget, M. (1997) Somatic instability of the myotonic dystrophy (CTG)n repeat during human fetal development. *Hum. Mol. Genet.*, **6**, 877–880.

52. Higham, C.F., Morales, F., Cobbold, C.A., Haydon, D.T. and Monckton, D.G. (2012) High levels of somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-frequent expansion and contraction mutations. *Hum. Mol. Genet.*, **21**, 2450–2463.

53. Gomes-Pereira, M., Fortune, M.T., Ingram, L., McAbney, J.P. and Monckton, D.G. (2004) Pms2 is a genetic enhancer of trinucleotide CAG.CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Hum. Mol. Genet.*, **13**, 1815–1825.

54. Larson, E., Fyfe, I., Morton, A.J. and Monckton, D.G. (2015) Age-, tissue- and length-dependent bidirectional somatic CAG*CTG repeat instability in an allelic series of R6/2 Huntington disease mice. *Neurobiol. Dis.*, **76**, 98–111.

55. Foiry, L., Dong, L., Savouret, C., Hubert, L., Riele, H.T., Junien, C. and Gourdon, G. (2006) Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Hum. Genet.*, **119**, 520–526.

56. Manley, K., Shirley, T.L., Flaherty, L. and Messer, A. (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat. Genet.*, **23**, 471–473.

57. Pinto, R.M., Dragileva, E., Kirby, A., Lloret, A., Lopez, E., St Claire, J., Panigrahi, G.B., Hou, C., Holloway, K., Gillis, T. *et al.* (2013) Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS Genet.*, **9**, e1003930.

58. Savouret, C., Brisson, E., Essers, J., Kanaar, R., Pastink, A., te Riele, H., Junien, C. and Gourdon, G. (2003) CTG repeat instability and size variation timing in DNA repair-deficient mice. *EMBO J.*, **22**, 2264–2273.

59. Wheeler, V.C., Lebel, L.A., Vrbanac, V., Teed, A., te Riele, H. and MacDonald, M.E. (2003) Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. *Hum. Mol. Genet.*, **12**, 273–281.

60. Dragileva, E., Hendricks, A., Teed, A., Gillis, T., Lopez, E.T., Friedberg, E.C., Kucherlapati, R., Edelmann, W., Lunetta, K.L., MacDonald, M.E. *et al.* (2009) Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol. Dis.*, **33**, 37–47.

61. Du, J., Campau, E., Soragni, E., Jespersen, C. and Gottesfeld, J.M. (2013) Length-dependent CTG.CAG triplet-repeat expansion in myotonic dystrophy patient-derived induced pluripotent stem cells. *Hum. Mol. Genet.*, **22**, 5276–5287.

62. Nakatani, R., Nakamori, M., Fujimura, H., Mochizuki, H. and Takahashi, M.P. (2015) Large expansion of CTG*CAG repeats is exacerbated by MutSbeta in human cells. *Sci. Rep.*, **5**, 11020.

63. Lopez de Munain, A., Cobo, A.M., Poza, J.J., Navarrete, D., Martorell, L., Palau, F., Emparanza, J.I. and Baiget, M. (1995) Influence of the sex of the transmitting grandparent in congenital myotonic dystrophy. *J. Med. Genet.*, **32**, 689–691.

64. Hafford-Tear, N.J., Tsai, Y.C., Sadan, A.N., Sanchez-Pintado, B., Zarouchlioti, C., Maher, G.J., Liskova, P., Tuft, S.J., Hardcastle, A.J., Clark, T.A. *et al.* (2019) CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat. *Genet. Med.*, **21**, 2092–2102.

65. Bocklandt, S., Hastie, A. and Cao, H. (2019) Bionano genome mapping: high-throughput, ultra-long molecule genome analysis system for precision genome assembly and

haploid-resolved structural variation discovery. *Adv. Exp. Med. Biol.*, **1129**, 97–118.

66. Bjornsson, H.T., Sigurdsson, M.I., Fallin, M.D., Irizarry, R.A., Aspelund, T., Cui, H., Yu, W., Rongione, M.A., Ekstrom, T.J., Harris, T.B. *et al.* (2008) Intra-individual change over time in DNA methylation with familial clustering. *JAMA*, **299**, 2877–2883.

67. Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J. *et al.* (2012) Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 10522–10527.

68. Jung, M. and Pfeifer, G.P. (2015) Aging and DNA methylation. *BMC Biol.*, **13**, 7.