Human Molecular Genetics, 2012, Vol. 21, No. 16 ***3558–3567***
doi:10.1093/hmg/dds185
Advance Access published on May 16, 2012

Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity

Fernando Morales[1,3], Jillian M. Couto[1], Catherine F. Higham[1,2], Grant Hogg[1], Patricia Cuenca[3], Claudia Braida[1], Richard H. Wilson[1], Berit Adam[1], Gerardo del Valle[6], Roberto Brian[4], Mauricio Sittenfeld[5], Tetsuo Ashizawa[7], Alison Wilcox[8], Douglas E. Wilcox[8] and Darren G. Monckton[1,*]

[1]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences and [2]School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK, [3]Instituto de Investigaciones en Salud y Escuela de Medicina, [4]Servicio de Neurología, Hospital Nacional de Niños, Escuela de Medicina and [5]Servicio de Neurología, Hospital San Juan de Dios, Escuela de Medicina, Universidad de Costa Rica, San José, Costa Rica, [6]Laboratorio de Neurofisiología (Neurolab), Curridabat, San José, Costa Rica, [7]Department of Neurology, University of Florida, 100 South Newell Drive, Gainesville, Florida 32611, USA and [8]Ferguson-Smith Centre for Clinical Genetics, Yorkhill Hospital, Dalnair Street, Glasgow G3 8SJ, UK

Received April 4, 2012; Revised and Accepted May 10, 2012

Deciphering the contribution of genetic instability in somatic cells is critical to our understanding of many human disorders. Myotonic dystrophy type 1 (DM1) is one such disorder that is caused by the expansion of a CTG repeat that shows extremely high levels of somatic instability. This somatic instability has compromised attempts to measure intergenerational repeat dynamics and infer genotype–phenotype relationships. Using single-molecule PCR, we have characterized more than 17 000 *de novo* somatic mutations from a large cohort of DM1 patients. These data reveal that the estimated progenitor allele length is the major modifier of age of onset. We find no evidence for a threshold above which repeat length does not contribute toward age at onset, suggesting pathogenesis is not constrained to a simple molecular switch such as nuclear retention of the *DMPK* transcript or haploinsufficiency for *DMPK* and/or *SIX5*. Importantly, we also show that age at onset is further modified by the level of somatic instability; patients in whom the repeat expands more rapidly, develop the symptoms earlier. These data establish a primary role for somatic instability in DM1 severity, further highlighting it as a therapeutic target. In addition, we show that the level of instability is highly heritable, implying a role for individual-specific *trans*-acting genetic modifiers. Identifying these *trans*-acting genetic modifiers will facilitate the formulation of novel therapies that curtail the accumulation of somatic expansions and may provide clues to the role these factors play in the development of cancer, aging and inherited disease in the general population.

INTRODUCTION

Genetic variation contributes directly to normal variation and disease. New mutations are the ultimate source of such variation, but *de novo* mutations in humans are generally very rare and very little is known about individual-specific mutational dynamics (1). In contrast to the marked stability of most of the genome, the expanded simple sequence repeats associated with a number of inherited disorders including myotonic dystrophy and Huntington disease mutate at high frequency in both the soma and germline (2,3). The dynamic

[*]To whom correspondence should be addressed at: Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK. Tel: +44 1413306213; Email: darren.monckton@glasgow.ac.uk

nature of these loci thus provide a unique route to investigating mutational dynamics in man.

Myotonic dystrophy type 1 (DM1), a clinically highly variable autosomal dominant inherited disorder, affects individuals of both sexes and all ages (4). The mutation responsible for DM1 is the expansion of an unstable CTG trinucleotide repeat located in the 3′-untranslated region of the *DMPK* gene (5–7) and in the promoter of the downstream *SIX5* gene (8). Affected DM1 patients present with expansions from 50 to up to several thousand repeats (7). Longer alleles are associated with a more severe form of the disease and an earlier age of onset (9–11). In patients, the rate of length change mutation at this locus approaches 100% in the germline (12) with a major expansion bias (10,13,14). This causes striking anticipation, where the age of onset typically decreases by 25–35 years per generation (15).

The DM1 repeat is also somatically highly unstable in a pathway that is expansion-biased, age-dependent and tissue-specific (12,16–20). Notably, the repeat length is always much larger in muscle DNA than that observed in blood (12,18–21). Given that longer repeat lengths are associated with more severe disease, it seems rational to assume that the expansion-biased, age-dependent and tissue-specific nature of somatic instability contributes toward both the tissue specificity and progressive nature of the symptoms. Attractive and logical as this hypothesis is, there are no direct data to support it.

Traditionally, DM1 is diagnosed by sizing the CTG repeat using restriction-digested genomic blood DNA and Southern blot hybridization, or a bulk PCR, using many thousands of cells worth of DNA. Using these approaches, the expanded allele frequently presents as a broad smear from which it is possible to estimate only the modal allele length. With these methods, the measured allele length typically accounts for only 20–40% of the observed variation in age of onset (22–24). Furthermore, correlations with specific symptoms are often worse, or undetectable (25–27). Similarly, measuring the allele length in muscle provides an even worse genotype–phenotype relationship than that observed with blood DNA (21). With measured allele length accounting for <50% of the variation in age at onset, it remains possible that allele length is not the major modifier of disease severity and that another factor contributes to more of the variation in age at onset. Such is the limited prognostic accuracy of the measured allele length that the International Myotonic Dystrophy Consortium recommend that patients are offered no prognostic information based on the current test (28), limiting the ability of patients to make informed lifestyle and reproductive choices. In fact, some diagnostic laboratories now only use a repeat-primed PCR (29)-based approach that provides no size estimation at all, but rather a simple ‘present’/‘absent’ result for an expanded allele, which translates into a simple binary ‘yes’/‘no’ disease diagnosis.

One explanation proposed for the poor genotype–phenotype correlations in DM1 is that above the lower disease-causing threshold of ~50 repeats, there is another upper threshold beyond which increasing repeat length makes no additional contribution toward age at onset (30–32). Prompted by their observation that *DMPK* transcripts become trapped in the nucleus beyond 80–400 CUG repeats (33), Hamshere *et al*. (30) reinvestigated the relationship between age at onset and repeat length. Using average repeat lengths measured by traditional Southern blot of restriction-digested genomic DNA and excluding congenital cases (due to the additional constellation of symptoms observed in this group), they observed a transition point around 143 repeats, beyond which increasing repeat length appeared to make no additional contribution toward age at onset (30). Likewise, Savic *et al*. (31) revealed a similar threshold of 250 repeats (the average allele length measured using a small-pool PCR) in a cohort of patients who did not contain congenital cases—a result which appeared to be confirmed by Hsaio *et al*. (32) (the average allele length measured using PCR). These data suggest the existence of a saturable disease pathway or molecular switch such as complete nuclear retention of the mutant *DMPK* transcript and/or the loss of expression of the downstream gene *SIX5* (30).

Previously, we used single-molecule-based small-pool PCR approaches to resolve the heterogeneous smear of CTG repeats in the *DMPK* gene into the discrete alleles present in individual cells (12). This allowed us to provide a detailed quantitative measure of repeat length variation and reveal the underlying shape of the distributions. Typically, repeat length distributions for the mutant allele in blood DNA from patients present as highly positively skewed with a relatively sharp lower boundary, below which smaller alleles are relatively rare. This lower boundary is conserved between tissues and provides the best estimate for the inherited or progenitor allele length (12). It is our hypothesis that previous genotype–phenotype correlations have been compromised by failure to take into account the age-dependent, expansion-biased nature of somatic mosaicism. We hypothesized that estimating the progenitor allele length would significantly improve the inverse correlation with age of onset over the traditional modal length measure, as it greatly reduces the confounding effects of somatic instability. We also hypothesized that individual-specific rates of somatic expansion will also impact on disease severity, and that individual-specific variation in expansion rates might cluster within families and be inherited as a quantitative genetic trait. Thus, we sought to estimate the progenitor allele length and measure the total level of allelic variation in the blood DNA of large numbers of DM1 patients in order to quantify the role of the progenitor allele length and age at sampling in generating somatic variation, relate these to disease severity and investigate whether variation in somatic instability is heritable.

RESULTS

Estimated progenitor allele length is the major modifier of age of onset in DM1

In order to test the hypothesis that the progenitor allele length is the major modifier of DM1 onset, we performed five small-pool PCRs per individual in a large population of symptomatic DM1 patients ($n = 137$) with approximately 30–50 genomic equivalents of DNA per reaction. From these, we used the lower boundary of the distribution to estimate the progenitor allele length (12) (see Supplementary Material, Fig. S1), and conducted a series of regression analyses. The logarithmic

Table 1. Regression models of the relationship between age of onset (Age_o) and the estimated progenitor allele length (PAL)

	Adjusted r^2	*P*-value	Parameter		Coefficient	Standard error	*t*-Statistic	*P*-value
Model 1: $Age_o = \beta_0 + \beta_1 \log(PAL)$	0.640	<0.0001	Intercept	β_0	127	6.7	19.0	4.5×10^{-25}
			log(PAL)	β_1	−42	2.7	−15.6	2.9×10^{-19}
Model 2: $Age_o = \beta_0 + \beta_1 \log(PAL) + \beta_2 \log(PAL)^2$	0.699	<0.0001	Intercept	β_0	319	37.1	8.6	2.0×10^{-14}
			log(PAL)	β_1	−205	31.4	−6.6	1.2×10^{-9}
			$\log(PAL)^2$	β_2	34	6.5	5.2	6.1×10^{-7}
Model 3: $Age_o = \alpha e^{\beta \log(PAL)}$	0.711	<0.0001	Intercept	α	1137	216.4	5.3	2.8×10^{-7}
			log(PAL)	β	−1.62	0.1	−18.0	4.5×10^{-23}

The table shows the adjusted squared coefficient of correlation (adjusted r^2), statistical significance (*P*) for each model and the coefficient, standard error, *t*-statistic and statistical significance (*P*), associated with each parameter in the model ($n = 137$). The coefficient provides an indication of the relative weight of the contribution of each parameter to the model and its associated standard error. The *t*-statistic and the corresponding *P*-value provide an indication of the statistical significance that the parameter is adding explanatory power to the model.

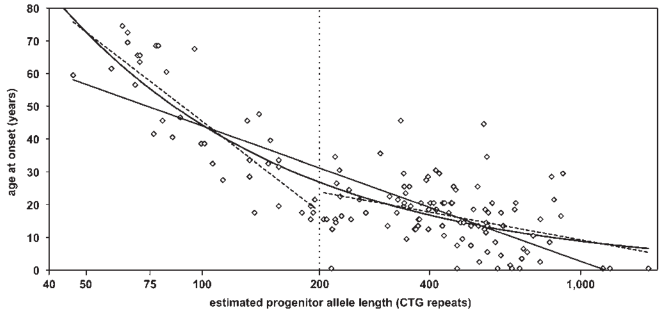

Figure 1. The estimated progenitor allele length is the major modifier of disease severity in myotonic dystrophy type 1. The graph shows the relationship between age of onset and the estimated progenitor allele length (CTG repeats) presented on a log scale. The straight solid line shows the regression line for all patients, using a linear model (model 1, Table 1, adjusted $r^2 = 0.640$, $P << 0.0001$, $n = 137$), and the curved solid line an exponential model (model 3, Table 1, adjusted $r^2 = 0.711$, $P < 0.0001$, $n = 137$). Linear regression with sub-groups of patients with the estimated progenitor allele length <200 CTG repeats (left-hand side dashed line, adjusted $r^2 = 0.737$, $P < 0.0001$, $n = 36$) and those with ≥200 CTG repeats (right hand-side dashed line, adjusted $r^2 = 0.163$, $P < 0.0001$, $n = 101$) revealed no evidence for a threshold at 200 CTG repeats (vertical dotted line) above which progenitor allele length does not contribute toward age at onset.

(base 10) transformation of the estimated progenitor allele length accounted for 64% (adjusted $r^2 = 0.640$, $P << 0.0001$) of the variation in age at onset (Table 1, model 1; Fig. 1). When compared with the modal allele length measured via a traditional Southern blot analysis of restriction-digested genomic DNA/bulk PCR analysis, available for 82 individuals in the cohort, a sample-matched analysis showed that the estimated progenitor allele length explained a greater amount of variation in age at onset (adjusted $r^2 = 0.523$, $P << 0.0001$) compared with the modal length (adjusted $r^2 = 0.405$, $P << 0.0001$) (Supplementary Material, Fig. S2). We then used the estimated progenitor allele length in our full cohort to test for the existence of a previously reported (30–32) threshold around 200 repeats [an average of the 143 repeats suggested by Hamshere *et al.* (30), and the 250 repeats suggested by Savic *et al.* (31) and Hsaio *et al.* (32)], beyond which increasing allele length appears to no longer contribute toward age at onset. We found a highly statistically significant relationship between the log-estimated progenitor allele length and variation in age of onset both below a threshold of 200 repeats (adjusted $r^2 = 0.737$, $P < 0.0001$) and above a threshold of 200 repeats (adjusted $r^2 = 0.163$, $P < 0.0001$) (Fig. 1). This result was robust to the precise position of the threshold, as correlations between age at onset and the estimated progenitor allele length remained statistically significant throughout a range of possible thresholds from 100–300 repeats (Supplementary Material, Fig. S3A and Table S1). Indeed, the absence of a detectable upper threshold remained significant (Supplementary Material, Fig. S3B and Table S1) even when excluding congenital cases as in previous analyses (30–32). Nonetheless, although the relationship remained highly significant above the putative threshold, the coefficient of correlation for longer alleles was much lower than that observed for shorter alleles and the regression line less steep (Fig. 1). Indeed, examination of the relationship between the progenitor allele length and age of onset showed that despite a logarithmic transformation, the data appeared to have non-linear components (Fig. 1). Therefore, we tested both quadratic and exponential models of the estimated progenitor allele length. Further variance was explained with these models (quadratic, model 2: adjusted $r^2 = 0.699$, $P < 0.0001$; exponential, model 3: adjusted $r^2 = 0.711$, $P < 0.0001$; Table 1, Fig. 1), indicating an additional, non-linear component to the relationship between age of onset and the estimated progenitor allele length.

Age at sampling and progenitor allele length synergistically modify the level of somatic instability

To investigate the dynamics of somatic mosaicism, the spectrum of DM1 alleles present in blood DNA was quantified using single-molecule small-pool PCR (see Supplementary Material, Fig. S4). We sized the repeat length in at least 100 expanded alleles from each of 136 individuals (14 asymptomatic and 122 symptomatic) (Supplementary Material, Table S2). The total data set comprised of 19 247 individually sized variant alleles. After accounting for the estimated progenitor allele length of the sampled individuals, and allowing for a 5% error margin, we estimated that at least 17 800 of these alleles represent *de novo* somatic mutants. For each individual, the degree of somatic instability was defined as the range between the 10th and 90th percentile of the allele length frequency distribution. This score represents a quantitative measure of somatic instability. However, given the

Table 2. Regression models of the relationship between somatic instability (SI), the estimated progenitor allele length (PAL) and age at sampling (age$_s$)

	Adjusted r^2	*P*-value	Parameter		Coefficient	Standard error	*t*-Statistic	*P*-value
Model 4: log(SI) = $\beta_0 + \beta_1$ log(PAL)	0.644	<0.0001	Intercept	β_0	−0.84	0.21	−4.0	9.3×10^{-05}
			log(PAL)	β_1	1.29	0.08	15.7	4.9×10^{-32}
Model 5: log(SI) = $\beta_0 + \beta_1$ log(age$_s$)	−0.005	0.60	Intercept	β_0	2.27	0.25	9.2	7.4×10^{-16}
			log(age$_s$)	β_1	0.09	0.16	0.5	6.0×10^{-01}
Model 6: log(SI) = $\beta_0 + \beta_1$ log(PAL) + β_2 log(age$_s$)	0.746	<0.0001	Intercept	β_0	−2.24	0.26	−8.7	1.2×10^{-14}
			log(PAL)	β_1	1.47	0.07	20.0	8.3×10^{-42}
			log(age$_s$)	β_2	0.65	0.09	7.4	1.2×10^{-11}
Model 7: log(SI) = $\beta_0 + \beta_1$ log(PAL) + β_2 log(age$_s$) + β_3 log(PAL) × log(age$_s$)	0.759	<0.0001	Intercept	β_0	−0.40	0.70	−0.6	5.7×10^{-01}
			log(PAL)	β_1	0.79	0.25	3.2	2.1×10^{-03}
			log(age$_s$)	β_2	−0.56	0.44	−1.3	2.0×10^{-01}
			log(PAL) × log(age$_s$)	β_3	0.44	0.16	2.8	5.8×10^{-03}
Model 8: log(SI) = $\beta_0 + \beta_1$ log(PAL) + β_2 log(age$_s$) + β_3 log(PAL) × log(age$_s$) + $\beta_4$log(PAL)^2+ β_5 log(age$_s$)^2	0.890	<0.0001	Intercept	β_0	−9.04	0.99	−9.1	1.1×10^{-15}
			log(PAL)	β_1	8.78	0.68	12.9	4.3×10^{-25}
			log(age$_s$)	β_2	−1.62	0.58	−2.8	6.5×10^{-03}
			log(PAL) × log(age$_s$)	β_3	0.40	0.15	2.7	7.7×10^{-03}
			log(PAL)^2	β_4	−1.67	0.13	−12.6	2.3×10^{-24}
			log(age$_s$)^2	β_5	0.44	0.11	4.2	5.6×10^{-05}

The table shows the adjusted squared coefficient of correlation (adjusted r^2) and statistical significance (*P*) for each model, and the coefficient, standard error, *t*-statistic and statistical significance (*P*) (as described in Table 1), associated with each parameter in the model ($n = 136$). Note that some parameters have coefficients of correlation that are opposite in sign to those expected. These are corrected by higher moments in the full model (e.g. log(age$_s$) in model 7, and log(age$_s$) and log(PAL)^2 in model 8) and probably reflect the non-random distribution of data due to the inherent sampling bias mediated by anticipation.

dynamic nature of this locus, it is a measurement that is expected to change throughout the lifetime of an individual and be dependent on the progenitor allele length and the age at which a patient was sampled (16,17). We tested these hypotheses using a series of regression models (Table 2). The estimated progenitor allele length explained a significant proportion of the variation in somatic instability (Table 2, model 4, adjusted $r^2 = 0.644$, $P < 0.0001$). Surprisingly, age at sampling alone was not significantly related (Table 2, model 5, adjusted $r^2 = -0.005$, $P = 0.60$), suggesting a more complicated relationship with somatic instability, potentially including the progenitor allele length. As both factors are interrelated with respect to their relationship with somatic instability (16,17), we tested additional models that included both variables. This included a model with both variables (model 6) and another with both variables and an interaction term (model 7). Both models were a marked improvement (Table 2, model 6: adjusted $r^2 = 0.746$, $P < 0.0001$; model 7: adjusted $r^2 = 0.759$, $P < 0.0001$) over the single variable models. Lastly we tested whether factors in addition to a simple linear relationship were also present in this model by adding quadratic versions of both estimated progenitor allele length and age at sampling. With this addition, the coefficients for all parameters and indeed the model were highly significant (Table 2, model 8: adjusted $r^2 = 0.890$, $P < 0.0001$; Fig. 2), accounting for 89% of the variation in somatic instability.

The level of somatic instability is an additional modifier of age of onset of DM1 symptoms

Although it is clear that the progenitor allele length is the major modifier of disease severity in DM1, it seems reasonable to assume that the subsequent rate of expansion will also impact on age at onset. To test this hypothesis, we used linear regression to investigate the relationship between the standardized residuals of age of onset not accounted for by the estimated progenitor allele length (model 2, Table 1) and the standardized residuals of somatic instability not accounted for by age at sampling and the estimated progenitor allele length (model 8, Table 2) in the 122 symptomatic individuals for whom we had acquired detailed repeat length distributions. This analysis resulted in a statistically significant negative correlation between age of onset and somatic instability (adjusted $r^2 = 0.068$, $P = 0.002$) (Fig. 3). This observation was confirmed by analyses showing that incorporating the

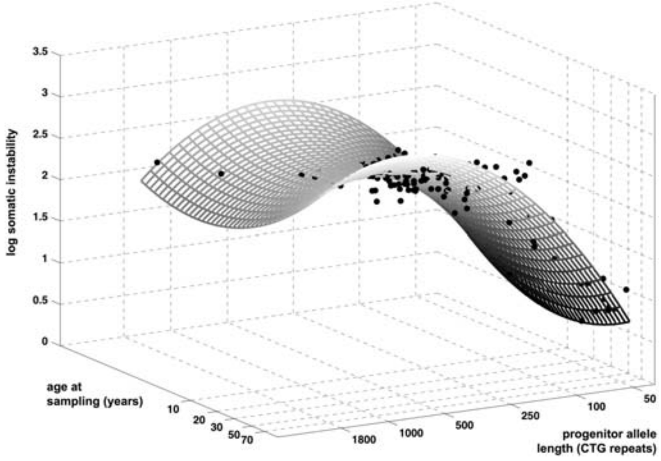

Figure 2. The estimated progenitor allele length and age at sampling are the major modifiers of somatic instability in DM1. The graph shows the relationship between the logarithmic transformation (base 10) of somatic instability, the estimated progenitor allele length (on a log scale) and the age at sampling (on a log scale) ($n = 136$). The surface has been fitted to the data using multilinear regression analysis (model 8, Table 2).

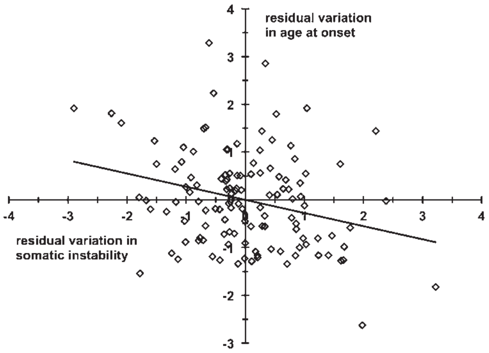

Figure 3. Residual variation in age at onset is inversely correlated with residual variation in somatic instability. The graph shows the relationship between standardized residual variation for age of onset (model 2, Table 1) and standardized residual variation data for somatic instability (model 8, Table 2) ($n = 122$). Linear regression analysis revealed a significant ($P = 0.002$) negative correlation (adjusted $r^2 = 0.068$) between the two variables.

standardized residuals of somatic instability (model 11, adjusted $r^2 = 0.574$, $P < 0.001$, Table 3) accounted for more variation in age at onset than progenitor allele length alone [models 9 (adjusted $r^2 = 0.486$, $P < 0.001$) and 10 (adjusted $r^2 = 0.543$, $P < 0.001$), Table 3]. In addition, analysis of the t-statistic for residual variation in somatic instability confirms that this parameter makes a statistically significant contribution to model 11 (Table 3, $t = -3.11$, $P = 0.0023$). Finally, both an F-nested test ($P = 0.010$) and a likelihood ratio test ($P = 0.008$) comparing the nested models 10 and 11 confirmed that model 11 incorporating the residual variation in somatic instability explains a statistically significant greater amount of the variation in age at onset than the simpler model 10. These data reveal that individuals in whom the repeat expands more rapidly than average have an age of onset earlier than average.

The residual level of somatic instability is a heritable quantitative trait

After correcting for the two major modifiers (progenitor allele length and age at sampling), the residual variation in somatic instability represents an individual-specific measure of genetic instability. Individual differences in the level of somatic instability might be attributable to genetic modifiers and might therefore be expected to be heritable. Of the 136 individuals in whom we derived detailed repeat length distributions, 89 were part of 21 families and formed 51 sibling pairs. Using the quantitative transmission disequilibrium test [QTDT (34)], we estimated the sib-pair intra-class correlations for residual somatic instability to be 0.28 ($P = 0.04$). We then used QTDT to partition this variation and yield a heritability estimate. The variance was partitioned into additive genetic (V_g), non-shared environment (V_e) and shared environment (V_c). The analysis yielded the estimates of V_g or heritability = 0.42, $V_e = 0.58$, $V_c = 0$, establishing residual somatic instability as a heritable quantitative genetic trait.

DISCUSSION

The expanded CTG repeat at the DM1 locus is one of the most unstable sequences in the human genome, with male germline length change mutation rates often in excess of 95% (12,35). The expanded DM1 repeat is also highly unstable in the soma, and mutations appear to accumulate through multiple small-length changes that are tissue-specific, age-dependent and expansion-biased (12,16–21,36). The traditional diagnostic measure of expanded allele length is derived by Southern blot analysis of restriction-digested genomic blood DNA and takes no account of the dynamic nature of CTG repeat somatic instability. Although inversely correlated with age of onset, the allele length thus measured typically accounts for only ~20–40% of the variation in age at onset (22–24), and does not provide clinically relevant prognostic information (28). However, because of the extensive somatic mosaicism, the mutant allele presents not as a discrete band, but as a heterogeneous smear from which it is possible to identify only the modal allele length and which increases in size as the patient ages (16,17). It was our hypothesis that the primary factor contributing to the poor genotype–phenotype correlations in DM1 was the failure to take into account age-dependent somatic expansion. We thus used small-pool PCR to estimate the progenitor or inherited allele length, as it is independent of age at sampling (12). These analyses revealed that the estimated progenitor allele length is indeed the major modifier of disease severity in DM1 accounting for 64% of the variation in age at onset in a simple log-linear relationship. This represents a major improvement over the traditional Southern blot analysis of restriction-digested genomic DNA. Notably, however, the relationship between the estimated progenitor allele length and age at onset appears to extend beyond the simple log-linear, as evidenced by an improved fit using either a quadratic or an exponential non-linear model. Indeed, the exponential fit accounts for 71% of the variation in age at onset, similar to the prognostic value of repeat length measures in the polyglutamine expansion disorders such as Huntington disease and spinocerebellar ataxia types 1, 2 and 7 (37–40) in which somatic instability in peripheral tissues is very low (41–43) and identification of the progenitor allele is usually less problematic. Interestingly, a similar non-linear relationship has been observed in Huntington disease, where the correlation between log repeat length and age of onset differs for juvenile and adult onset patients (44). Such non-linear components may explain why some previous DM1 studies observed an apparent threshold at around 200 repeats, above which no significant correlation between age of onset and allele length was detected (30–32). However, these studies will have also been compromised by age-dependent somatic expansion, with longer alleles more susceptible to age at sampling effects. The existence of such a threshold would have major implications for the pathogenic mechanism in DM1, implicating a saturation effect or the existence of an all-or-nothing molecular switch at the threshold (30). However, when corrected for the age at sampling bias, we found no evidence for a threshold above which repeat length has no influence on age at onset, suggesting the absence of such a simple binary switch. Nonetheless, the non-linear nature of the age at onset versus repeat length

Table 3. Regression models of the relationship between age at onset ($\mathrm{Age_o}$), the estimated progenitor allele length (PAL) and the standardized residuals of somatic instability (SI)

	Adjusted r^2	*P*-value	Parameter		Coefficient	Standard error	*t*-Statistic	*P*-value
Model 9: $\mathrm{Age_o} = \beta_0 + \beta_1 \log(\mathrm{PAL})$	0.486	<0.0001	Intercept	β_0	107	8.2	13.2	4.5×10^{-25}
			log(PAL)	β_1	−34	3.2	−10.7	2.9×10^{-19}
Model 10: $\mathrm{Age_o} = \beta_0 + \beta_1 \log(\mathrm{PAL}) + \beta_2 \log(\mathrm{PAL})^2$	0.543	<0.0001	Intercept	β_0	284	45.1	6.3	4.9×10^{-09}
			log(PAL)	β_1	−181	36.9	−4.9	3.1×10^{-06}
			$\log(\mathrm{PAL})^2$	β_2	30	7.5	4.0	1.2×10^{-04}
Model 11: $\mathrm{Age_o} = \beta_0 + \beta_1 \log(\mathrm{PAL}) + \beta_2$ Standardized residual(SI) $+ \beta_3 \log(\mathrm{PAL})^2$	0.574	<0.0001	Intercept	β_0	284	43.5	6.5	1.7×10^{-09}
			log(PAL)	β_1	−181	35.6	−5.1	1.5×10^{-06}
			Standardized residual (SI)	β_2	−3	0.8	−3.1	2.3×10^{-03}
			$\log(\mathrm{PAL})^2$	β_3	30	7.2	4.1	6.8×10^{-05}

The table shows the adjusted squared coefficient of correlation (adjusted 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 (as described in Table 1). Models 9 and 10 are equivalent to models 1 and 2 in Table 1, but the coefficients are slightly different as this data set is slightly smaller ($n = 122$) than the total data set used in Table 1 ($n = 137$). The patient data used for these analyses ($n = 122$) do not include a subset of old patients with small expansions and very young patients in whom the levels of acquired somatic mosaicism are very low. As these patients are at the extremes of the age of onset distribution, their omission serves to reduce the observed coefficients of correlation with age at onset. Nonetheless, these analyses still reveal that the inclusion of the standardized residuals of somatic instability (model 11) significantly increases the degree of variation in age at onset explained. Note that some parameters with coefficients of correlation opposite in sign to those expected are corrected by higher moments in the full model (e.g. $\log(\mathrm{PAL})^2$ in models 10 and 11) and probably reflect the non-random distribution of data due to the inherent sampling bias mediated by anticipation.

correlations does hint at a biphasic response, possibly mediated by trapping of the *DMPK* transcript within the nucleus (45) and/or suppression of *SIX5* transcription (46,47) up to the transition point of approximately 200 repeats and a CUG RNA repeat length-dependent sequestration of the MBNL family of regulators of alternative splicing (48) beyond this point. Of course, however, it is important to consider that the repeat lengths that will be present in the affected tissues at onset will be much longer than those inherited.

As discussed, genotype–phenotype relationships using the traditional approach are so poor that the International Myotonic Dystrophy Consortium has recommended that patients are simply provided with a ‘yes’ or ‘no’ diagnosis (28). Here, by estimating progenitor allele length, we have been able to provide significantly more accurate genotype–phenotype relationships. We have also shown these genotype–phenotype relationships can be even further improved by incorporating a measure of the individual-specific mutational dynamics. We hope these insights will ultimately lead to the introduction of more informative prognostic testing for patients and families. Of course, there are legitimate concerns regarding the feasibility of extending these types of analyses to the diagnostic laboratory. However, although small-pool PCR *per se* is not a standard diagnostic test, the component procedures, PCR and Southern blot hybridization currently are. The progenitor allele length can be estimated using a few replicate PCRs with 180–300 pg of DNA. Thus, estimating progenitor allele length should be within the existing expertise of many clinical genetics diagnostic laboratories and would represent no greater effort than is currently expended on genetic tests for other disorders where multiple exons are amplified and sequenced. A full-scale quantitative analysis of the repeat length distribution within a patient would represent a greater investment and is probably not currently a realistic proposition for diagnostic laboratories. Recently, alternative high-throughput approaches to quantifying somatic instability in Huntington disease transgenic mouse models have been reported (49,50). These approaches have been based on the use of fluorescently labeled primers and fragment length analysis using automated DNA-sequencing-type apparatus and the GeneMapper software. This approach can be applied to relatively small expansions, but cannot be used to quantify instability in large expanded alleles [>200 repeats (51,52)] as are frequently observed in DM1 patients. However, rapid advances in sequencing technology, in particular those based on long single-molecule reads (53,54), are likely to lead to alternative high-throughput approaches to estimating the progenitor allele and measuring somatic mosaicism that will reduce the barriers for wider implementation in the not-too-distant future.

Here, we have estimated the progenitor allele length and measured repeat dynamics in blood DNA. Blood DNA is an obvious source of DNA for establishing genotype–phenotype relationships since its acquisition is minimally invasive and blood appears to be one of the tissues in which the DM1 repeat remains most stable and hence should be a good source for estimating the progenitor allele length. Nonetheless, a considerable degree of repeat length variation is observed in blood DNA, and the lower boundary of the distribution of the DM1 repeat observed remains only an estimation of the progenitor allele length. Although biased toward net expansion overall, it is clear that somatic cells can acquire contractions (36) and it is likely that, in some younger individuals, the lower boundary of the distribution may drop below the progenitor allele. Conversely, in some older individuals with larger alleles, it is possible that the DM1 repeat in all cells may have expanded beyond the progenitor allele length. We are currently developing new objective computational modeling approaches to estimating the progenitor allele length based on repeat length distributions (36), but these are dependent on a detailed analysis of repeat length variation within an individual.

As discussed above, analysis of blood DNA remains the tissue of choice for estimating progenitor allele length.

However, it must be recognized that the hematopoietic system is not a primary target in DM1 pathogenesis and it seems unlikely that the dynamics of somatic instability in hematopoietic stem cells are having a major impact on the age at onset or disease progression. Rather, we envision a scenario where it is the somatic expansion of the CTG repeat in the primary affected tissues, such as muscle, that is crucial in mediating disease onset and subsequent progression. We assume that symptoms become detectable when a sufficient proportion of cells have acquired sufficiently long repeats that the combined level of dysfunction at the cellular level is reflected in performance at the tissue level. We therefore rationalize the demonstrated utility of the estimated progenitor allele length in predicting age at onset as a function of largely deterministic expansion dynamics in which alleles expand toward the symptomatic threshold at a predictable rate over time in the affected tissues primarily driven by repeat length, as discussed by Kaplan *et al*. (55). We interpret the role that individual-specific modifiers of somatic instability play similarly. That is, that the rate of expansion toward the disease threshold in affected tissues is modifiable by individual-specific factors, and that these factors are at least partially shared between hematopoietic stem cells and the affected tissues. To further understand these relationships it would clearly be highly desirable to monitor repeat dynamics in one of the affected tissues and relate this to dysfunction in that tissue over time. Unfortunately, there are few detailed longitudinal natural history studies in large DM1 patient populations, and repeat lengths in most tissues other than blood, and particularly in affected tissues such as muscle, are often many thousands of repeats in size (18–21), and beyond the length that can be reliably quantified using small-pool PCR approaches (approximately 1500 repeats).

Here we used single-molecule PCR techniques to derive detailed CTG repeat allele length distributions in 136 DM1 patients, characterizing over 17 000 *de novo* somatic mutations. As expected, these data show that the rate at which somatic mutations accumulate is highly dependent on allele length. This presumably reflects a greater propensity for longer alleles to adopt the slipped strand DNA structures that are the assumed mutational intermediates (56). Interestingly, simple regression analysis with age alone failed to reveal a significant correlation with the level of somatic mutation. This most likely results from the extreme sampling bias in DM1 families in which more severely affected individuals with larger alleles tend to be sampled at a much younger age than more mildly affected individuals with smaller expansions, confounding our ability to detect an age effect. Nevertheless, using multivariate analyses, age was confirmed as a major factor in modifying the level of accumulated mutations and clearly interacts synergistically with allele length. Overall, allele length and age at sampling explain 89% of the observed variation in somatic instability.

The residual variation in somatic instability, not accounted for by allele length and age at sampling, reflects individual-specific differences in the rate of accumulation of mutations. If somatic expansion contributes directly toward disease progression, then we would predict that these individual-specific differences in mutational dynamics would be reflected in disease severity. Here we have provided the first direct evidence in DM1 that this is so, establishing a statistically significant ($P = 0.002$) inverse correlation between residual variation in age of onset and residual variation in somatic instability. Put simply, these data reveal that individuals with below-average rates of expansion tend to get the disease later than expected and vice versa. These observations reveal somatic instability as an important modifier of the disease pathway and help to further compensate for the age at sampling effects compromising the utility of modal allele length in explaining variation in age at onset. However, it is important to note that the magnitude of the influence detected was relatively modest; residual somatic instability accounted for ∼7% of the residual variation in age at onset. Nonetheless, this remains a critical insight, and given the difficulties of assigning an age at onset to a patient and the confounding effects of somatic mosaicism and age at sampling on estimating the progenitor allele and quantifying somatic instability, and the measurement of somatic instability in a non-target tissue, it seems likely that this very much represents an underestimate of the true role. These observations therefore validate repeat expansion as a therapeutic target in DM1 (3). Recently, analysis of the levels of CAG repeat mosaicism in post-mortem cortical DNA samples of Huntington disease patients revealed some evidence for a similar effect: patients with extreme early onset having repeat length distributions skewed toward larger alleles (57). These observations suggest that somatic expansion likely contributes to disease progression in other disorders associated with unstable expanded microsatellite loci, as is predicted from the nature of the relationship between allele length and age at onset (55).

Individual-specific differences in the level of somatic instability are likely to be attributable to individual-specific environmental or genetic factors. Here, we have demonstrated a high level of familial clustering of residual somatic variation that indicates that some of the underlying variance can be attributed to heritable genetic variation. Our sib-pair analysis yielded evidence for familiality, and the heritability analysis suggested that ∼40% of the variance in somatic instability can be attributed to additive genetic factors. It is possible that some of the familiality is associated with *cis*-acting genetic factors. Indeed, we have previously demonstrated the stabilizing effect of CCG and CGG variant repeats within the DM1 array (58). In this study, we screened for the presence of such variant repeats and excluded individuals carrying them from the analyses. Effectively, all DM1 expansions are found on the same conserved haplotype (59), greatly reducing the likelihood that the heritability of somatic instability is mediated by flanking sequence polymorphisms. However, we cannot formally exclude the existence of unknown newly arisen sequence or epigenetic variants in the DM1 population. Nonetheless, it seems probable that most of the variation is likely to be attributable to *trans*-acting genetic factors. The most obvious candidates for *trans*-acting modifiers of instability are components of the DNA mismatch repair pathway. The *Msh2* (60), *Msh3* (61) and *Pms2* (62) mismatch repair genes have all been shown to be critical for generating somatic CTG•CAG repeat expansions in mice. Identification of the genetic modifiers of somatic mosaicism will provide insights into the molecular mechanism(s) of expansion and identify new targets for therapeutic intervention in DM1 and related unstable microsatellite disorders. More immediately, a greater understanding of the genotype–phenotype relationship in DM1 will facilitate genetic stratification and reduction of cohort variability in the clinical trials that are likely to be initiated in the not-too-distant future following on from recent successes in animal models of DM1 (63,64).

MATERIALS AND METHODS

Patient samples

Genomic DNA from all individuals in the sample was purified from peripheral blood leukocytes, using phenol–chloroform extraction and proteinase K. Because of the potential confounding effects of variant repeats within the CTG array on the disease phenotype (58,65), all patients were tested for the presence of CGG and CCG variant repeats (Couto *et al.*, unpublished observations) and individuals with variant repeats excluded. We also identified two individuals who presented three main alleles, one normal (<50 repeats) and two expanded alleles (>50 repeats). The presence of two expanded alleles in blood is assumed to reflect an early embryonic mutation event (66,67). Because of the difficulty in defining a progenitor allele length or assigning somatic variants to the appropriate allele in such individuals, these two cases were also excluded from the analyses.

Modal allele size determination

The number of CTG repeats was determined using the traditional method of sizing the most intense region of the expanded allele smear, visualized via a Southern blot hybridization of either restriction-digested genomic peripheral blood DNA or a bulk DNA PCR.

Small-pool PCR

Small-pool PCR analysis was performed using oligonucleotide primers DM-C and DM-BR as previously described (12,68). For estimating the progenitor allele length, five replicate reactions with ∼180–300 pg of genomic DNA were performed and the progenitor allele estimated from the lower boundary of the distribution (12) (Supplementary Material, Fig. S1). For detailed quantification of the degree of somatic variation, samples were amplified with 10–70 pg DNA per reaction (Supplementary Material, Fig. S4) and at least 100 single expanded alleles per individual were sized. PCR products were sized using the Kodak Molecular Imaging software 3.5.4 (Carestream Health, Inc.). The single-molecule distributions generated for each individual are provided in the Supplementary Material, Table S2.

Statistical analysis

Regression analyses were performed offline using commercial software packages (MATLAB v7.7, The MathWorks, Inc., Natick, MA, USA or SPSS Statistics, IBM, Armonk, NY, USA). For all the regression analyses, we have reported the adjusted r^2 values which incorporate a correction for the number of parameters included in the model facilitating direct comparison between the models. Genetic correlations and heritability estimates were acquired using QTDT, a general test package for the association of quantitative measures in nuclear families (34). Note that 0.5 was added to all ages at sampling and ages at onset to allow log transformation of the data.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the DGM group, Dan Haydon and Christina Cobbold for productive discussions and helpful comments on the manuscript. The authors would also like to thank the myotonic dystrophy patients and their families for their assistance.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by awards from the Lister Institute for Preventive Medicine (www.lister-institute.org.uk); the Leverhulme Trust (www.leverhulme.ac.uk); the Myotonic Dystrophy Support Group (www.myotonicdystrophysupportgroup.org); the Association Française contre les Myopathies (www.afm-telethon.fr); and the Muscular Dystrophy Campaign (www.muscular-dystrophy.org) to D.G.M. This work was also supported by awards from the University of Glasgow (www.gla.ac.uk); the Universidad de Costa Rica (www.ucr.ac.cr); the Ministerio de Ciencia y Tecnología of Costa Rica (www.micit.go.cr); and the National Council for Scientific and Technological Research in Costa Rica (CONICIT) (www.conicit.go.cr) to F.M. and a University of Glasgow Kelvin Smith PhD Scholarship awarded to C.F.H.

REFERENCES

1. Arnheim, N. and Calabrese, P. (2009) Understanding what determines the frequency and pattern of human germline mutations. *Nat. Rev. Genet.*, **10**, 478–488.
2. Brouwer, J.R., Willemsen, R. and Oostra, B.A. (2009) Microsatellite repeat instability and neurological disease. *Bioessays*, **31**, 71–83.
3. 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.
4. Harper, P.S. (2001) *Myotonic Dystrophy*, 3rd edn. WB Saunders Co., London.
5. Buxton, J., Shelbourne, P., Davies, J., Jones, C., Tongeren, T.V., 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. 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.

8. Boucher, C.A., King, S.K., Carey, N., Krahe, R., Winchester, C.L., Rahman, S., Creavin, T., Meghji, P., Bailey, M.E.S., Chartier, F.L. *et al.* (1995) A novel homeodomain-encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable $(CTG)_n$ repeat. *Hum. Mol. Genet.*, **4**, 1919–1925.
9. 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.
10. 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.
11. 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. *J. Am. Med. Assoc.*, **269**, 1960–1965.
12. Monckton, D.G., Wong, L.J., 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.
13. 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. II. Complex relationships between clinical findings and structure of the GCT repeat. *Neurology*, **42**, 1877–1883.
14. 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.
15. Höweler, C.J., Busch, H.F.M., Geraedts, J.P.M., Niermeijer, M.F. and Staal, A. (1989) Anticipation in myotonic dystrophy: fact or fiction? *Brain*, **112**, 779–797.
16. Wong, L.J., 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.
17. Martorell, L., Monckton, D.G., Gamez, J., Johnson, K.J., Gich, I., Lopez de Munain, A. 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.
18. 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.
19. Ashizawa, T., Dubel, J.R. and Harati, Y. (1993) Somatic instability of CTG repeat in myotonic dystrophy. *Neurology*, **43**, 2674–2678.
20. Thornton, C.A., Johnson, K.J. and Moxley, R.T. (1994) Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Ann. Neurol.*, **35**, 104–107.
21. 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.
22. Mladenovic, J., Pekmezovic, T., Todorovic, S., Rakocevic-Stojanovic, V., Savic, D., Romac, S. and Apostolski, S. (2006) Survival and mortality of myotonic dystrophy type 1 (Steinert's disease) in the population of Belgrade. *Eur. J. Neurol.*, **13**, 451–454.
23. Perini, G.I., Menegazzo, E., Ermani, M., Zara, M., Gemma, A., Ferruzza, E., Gennarelli, M. and Angelini, C. (1999) Cognitive impairment and (CTG)n expansion in myotonic dystrophy patients. *Biol. Psychiatry*, **46**, 425–431.
24. Marchini, C., Lonigro, R., Verriello, L., Pellizzari, L., Bergonzi, P. and Damante, G. (2000) Correlations between individual clinical manifestations and CTG repeat amplification in myotonic dystrophy. *Clin. Genet.*, **57**, 74–82.
25. Modoni, A., Silvestri, G., Pomponi, M.G., Mangiola, F., Tonali, P.A. and Marra, C. (2004) Characterization of the pattern of cognitive impairment in myotonic dystrophy type 1. *Arch. Neurol.*, **61**, 1943–1947.
26. Gharehbaghi-Schnell, E.B., Finsterer, J., Korschineck, I., Mamoli, B. and Binder, B.R. (1998) Genotype-phenotype correlation in myotonic dystrophy. *Clin. Genet.*, **53**, 20–26.
27. Merlevede, K., Vermander, D., Theys, P., Legius, E., Ector, H. and Robberecht, W. (2002) Cardiac involvement and CTG expansion in myotonic dystrophy. *J. Neurol.*, **249**, 693–698.
28. The International Myotonic Dystrophy Consortium. Gonzalez, I., Ohsawa, N., Singer, R.H., Devillers, M., Ashizawa, T., Balasubramanyam, A., Cooper, T.A., Khajavi, M., Lia-Baldini, A.S. *et al.* (2000) New nomenclature and DNA testing guidelines for myotonic dystrophy type 1 (DM1). *Neurology*, **54**, 1218–1221.
29. Warner, J.P., Barron, L.H., Goudie, D., Kelly, K., Dow, D., Fitzpatrick, D.R. and Brock, D.J. (1996) A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J. Med. Genet.*, **33**, 1022–1026.
30. Hamshere, M.G., Harley, H., Harper, P., Brook, J.D. and Brookfield, J.F. (1999) Myotonic dystrophy: the correlation of (CTG) repeat length in leucocytes with age at onset is significant only for patients with small expansions. *J. Med. Genet.*, **36**, 59–61.
31. Savic, D., Rakocvic-Stojanovic, V., Keckarevic, D., Culjkovic, B., Stojkovic, O., Mladenovic, J., Todorovic, S., Apostolski, S. and Romac, S. (2002) 250 CTG repeats in DMPK is a threshold for correlation of expansion size and age at onset of juvenile-adult DM1. *Hum. Mutat.*, **19**, 131–139.
32. Hsiao, K.M., Chen, S.S., Li, S.Y., Chiang, S.Y., Lin, H.M., Pan, H., Huang, C.C., Kuo, H.C., Jou, S.B., Su, C.C. *et al.* (2003) Epidemiological and genetic studies of myotonic dystrophy type 1 in Taiwan. *Neuroepidemiology*, **22**, 283–289.
33. Hamshere, M.G., Newman, E.E., Alwazzan, M., Athwal, B.S. and Brook, J.D. (1997) Transcriptional abnormality in myotonic dystrophy affects *DMPK* but not neighboring genes. *Proc. Natl Acad. Sci. USA*, **94**, 7394–7399.
34. Abecasis, G.R., Cardon, L.R. and Cookson, W.O. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279–292.
35. Martorell, L., Gamez, J., Cayuela, M.L., Gould, F.K., McAbney, J.P., Ashizawa, T., Monckton, D.G. and Baiget, M. (2004) Germline mutational dynamics in myotonic dystrophy type 1 males: allele length and age effects. *Neurology*, **62**, 269–274.
36. 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.
37. Snell, R.G., MacMillan, J.C., Cheadle, J.P., Fenton, I., Lazarou, L.P., Davies, P., MacDonald, M.E., Gusella, J.F., Harper, P.S. and Shaw, D.J. (1993) Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat. Genet.*, **4**, 393–397.
38. Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J.M., Weber, C., Mandel, J.L., Cancel, G., Abbas, N. *et al.* (1996) Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat. Genet.*, **14**, 285–291.
39. Ranum, L.P., Chung, M.Y., Banfi, S., Bryer, A., Schut, L.J., Ramesar, R., Duvick, L.A., McCall, A., Subramony, S.H., Goldfarb, L. *et al.* (1994) Molecular and clinical correlations in spinocerebellar ataxia type I: evidence for familial effects on the age at onset. *Am. J. Hum. Genet.*, **55**, 244–252.
40. David, G., Dürr, A., Stevanin, G., Cancel, G., Abbas, N., Benomar, A., Belal, S., Lebre, A.S., Abada-Bendib, M., Grid, D. *et al.* (1998) Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). *Hum. Mol. Genet.*, **7**, 165–170.
41. Veitch, N.J., Ennis, M., McAbney, J.P., Shelbourne, P.F. and Monckton, D.G. (2007) Inherited CAG•CTG allele length is a major modifier of somatic mutation length variability in Huntington disease. *DNA Repair*, **6**, 789–796.
42. Monckton, D.G., Cayuela, M.L., Gould, F.K., Brock, G.J.R., de Silva, R. and Ashizawa, T. (1999) Very large $(CAG)_n$ DNA repeat expansions in the sperm of two spinocerebellar ataxia type 7 males. *Hum. Mol. Genet.*, **8**, 2473–2478.
43. Chong, S.S., McCall, A.E., Cota, J., Subramony, S.H., Orr, H.T., Hughes, M.R. and Zoghbi, H.Y. (1995) Gametic and somatic tissue specific heterogeneity of the expanded *SCA1* CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, **10**, 344–350.
44. Andresen, J.M., Gayan, J., Djousse, L., Roberts, S., Brocklebank, D., Cherny, S.S., Cardon, L.R., Gusella, J.F., MacDonald, M.E., Myers, R.H. *et al.* (2007) The relationship between CAG repeat length and age of onset

differs for Huntington's disease patients with juvenile onset or adult onset. *Ann. Hum. Genet.*, **71**, 295–301.

45. Davis, B.M., McCurrach, M.E., Taneja, K.L., Singer, R.H. and Housman, D.E. (1997) Expansion of a CUG trinucleotide repeat in the 3′ untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. *Proc. Natl Acad. Sci. USA*, **94**, 7388–7393.
46. Klesert, T.R., Otten, A.D., Bird, T.D. and Tapscott, S.J. (1997) Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of *DMAHP*. *Nat. Genet.*, **16**, 402–406.
47. Thornton, C.A., Wymer, J.P., Simmons, Z., McClain, C. and Moxley, R.T. (1997) Expansion of the myotonic dystrophy CTG repeat reduces expression of the flanking *DMAHP* gene. *Nat. Genet.*, **16**, 407–409.
48. Lee, J.E. and Cooper, T.A. (2009) Pathogenic mechanisms of myotonic dystrophy. *Biochem. Soc. Trans.*, **37**, 1281–1286.
49. Lee, J.M., Zhang, J., Su, A.I., Walker, J.R., Wiltshire, T., Kang, K., Dragileva, E., Gillis, T., Lopez, E.T., Boily, M.J. *et al.* (2010) A novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC Syst. Biol.*, **4**, 29.
50. Mollersen, L., Rowe, A.D., Larsen, E., Rognes, T. and Klungland, A. (2010) Continuous and periodic expansion of CAG repeats in Huntington's disease R6/1 mice. *PLoS Genet.*, **6**, e1001242.
51. Fortune, M.T., Vassilopoulos, C., Coolbaugh, M.I., Siciliano, M.J. and Monckton, D.G. (2000) Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability. *Hum. Mol. Genet.*, **9**, 439–445.
52. Kennedy, L. and Shelbourne, P.F. (2000) Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Hum. Mol. Genet.*, **9**, 2539–2544.
53. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
54. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.
55. Kaplan, S., Itzkovitz, S. and Shapiro, E. (2007) A universal mechanism ties genotype to phenotype in trinucleotide diseases. *PLoS Comput. Biol.*, **3**, e235.
56. Pearson, C.E. and Sinden, R.R. (1996) Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*, **35**, 5041–5053.
57. Swami, M., Hendricks, A.E., Gillis, T., Massood, T., Mysore, J., Myers, R.H. and Wheeler, V.C. (2009) Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.*, **18**, 3039–3047.
58. 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.
59. Neville, C.E., Mahadevan, M.S., Barcelo, J.M. and Korneluk, R.G. (1994) High resolution genetic analysis suggests one ancestral predisposing haplotype for the origin of the myotonic dystrophy mutation. *Hum. Mol. Genet.*, **3**, 45–51.
60. 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.
61. van den Broek, W.J., Nelen, M.R., Wansink, D.G., Coerwinkel, M.M., te Riele, H., Groenen, P.J. and Wieringa, B. (2002) Somatic expansion behaviour of the $(CTG)_{(n)}$ repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum. Mol. Genet.*, **11**, 191–198.
62. 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.
63. Wheeler, T.M., Sobczak, K., Lueck, J.D., Osborne, R.J., Lin, X., Dirksen, R.T. and Thornton, C.A. (2009) Reversal of RNA dominance by displacement of protein sequestered on triplet repeat RNA. *Science*, **325**, 336–339.
64. Mulders, S.A., van den Broek, W.J., Wheeler, T.M., Croes, H.J., van Kuik-Romeijn, P., de Kimpe, S.J., Furling, D., Platenburg, G.J., Gourdon, G., Thornton, C.A. *et al.* (2009) Triplet-repeat oligonucleotide-mediated reversal of RNA toxicity in myotonic dystrophy. *Proc. Natl Acad. Sci. USA*, **106**, 13915–13920.
65. Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., Prochazka, T., Koukal, P., Marikova, T., Kraus, J. *et al.* (2009) Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. *Am. J. Med. Genet. A*, **149A**, 1365–1374.
66. Gibbs, M., Collick, A., Kelly, R.G. and Jeffreys, A.J. (1993) A tetranucleotide repeat mouse minisatellite displaying substantial somatic instability during early preimplantation development. *Genomics*, **17**, 121–128.
67. Monckton, D.G., Coolbaugh, M.I., Ashizawa, K., Siciliano, M.J. and Caskey, C.T. (1997) Hypermutable myotonic dystrophy CTG repeats in transgenic mice. *Nat. Genet.*, **15**, 193–196.
68. Gomes-Pereira, M., Bidichandani, S.I. and Monckton, D.G. (2004) Analysis of unstable triplet repeats using small-pool polymerase chain reaction. *Methods Mol. Biol.*, **277**, 61–76.