Human Molecular Genetics, 2012, Vol. 21, No. 11 **2450–2463**
doi:10.1093/hmg/dds059
Advance Access published on February 24, 2012

High levels of somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-frequent expansion and contraction mutations

Catherine F. Higham[1,2,*], Fernando Morales[1,5], Christina A. Cobbold[2,3], Daniel T. Haydon[2,4] and Darren G. Monckton[1,2]

[1]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, [2]Boyd Orr Centre for Population and Ecosystem Health, [3]School of Mathematics and Statistics, College of Science and Engineering and [4]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK and [5]Instituto de Investigaciones en Salud y Escuela de Medicina, Universidad de Costa Rica, San José, Costa Rica

Received November 24, 2011; Revised February 2, 2012; Accepted February 18, 2012

Several human genetic diseases are associated with inheriting an abnormally large unstable DNA simple sequence repeat. These sequences mutate, by changing the number of repeats, many times during the lifetime of those affected, with a bias towards expansion. These somatic changes lead not only to the presence of cells with different numbers of repeats in the same tissue, but also produce increasingly longer repeats, contributing towards the progressive nature of the symptoms. Modelling the progression of repeat length throughout the lifetime of individuals has potential for improving prognostic information as well as providing a deeper understanding of the underlying biological process. A large data set comprising blood DNA samples from individuals with one such disease, myotonic dystrophy type 1, provides an opportunity to parameterize a mathematical model for repeat length evolution that we can use to infer biological parameters of interest. We developed new mathematical models by modifying a proposed stochastic birth process to incorporate possible contraction. A hierarchical Bayesian approach was used as the basis for inference, and we estimated the distribution of mutation rates in the population. We used model comparison analysis to reveal, for the first time, that the expansion bias observed in the distributions of repeat lengths is likely to be the cumulative effect of many expansion and contraction events. We predict that mutation events can occur as frequently as every other day, which matches the timing of regular cell activities such as DNA repair and transcription but not DNA replication.

INTRODUCTION

Myotonic dystrophy type 1 (DM1) is one of over 20 diseases [others include Huntington's disease (HD) and fragile X syndrome] associated with inheriting an abnormally long, unstable DNA simple sequence repeat that mutates by changing the number of repeats (1–3). In the case of DM1, a CTG repeat tract is found in the non-coding region of a gene called *dystrophia myotonica* protein kinase (*DMPK*) (4–6). The number of repeats is polymorphic in the general population lying between 5 and 37, but over a pathological threshold of ~50 repeats, there is disease, as comprehensively discussed by Harper (7).

These DNA changes occur between generations, with usually a higher repeat length passed on from the parent to the child. This leads to an effect known as anticipation where the symptoms are more severe and appear ~30 years earlier in the next generation (8–10). But during the lifetime of individuals, the repeat lengths continue to evolve, with what looks like an expansion bias, leading to the presence of cells with different repeat lengths in the same tissue, known as somatic mosaicism (11). Generally, an increase in the

*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 1413306220; Fax: +44 1413304878; Email: c.higham.1@research.gla.ac.uk

number of repeats passed on from one generation to the next causes a decreasing age of disease onset (8,9), while an increase in the number of repeats throughout the lifetime of an individual contributes towards the progressive nature of the symptoms (F. Morales *et al.*, manuscript in preparation) and similarly for HD (12). Kaplan *et al*. (13) proposed a mechanism by which length-dependent somatic mosaicism drove cells across a pathological threshold which is supported by experimental and clinical data. A link between somatic mosaicism and disease onset suggests that preventing the repeat lengths from expanding might be therapeutic (1,14).

Length changes in other repetitive DNA sequences, or microsatellites, occur more commonly than other types of mutations such as individual base-pair substitutions, at between 10^{-2} and 10^{-6} per locus per generation (15), but mutation rates for DM1 are several orders of magnitude higher occurring, as described above, not just between generations, but also at a high rate during the lifetime of individuals. This has led to the introduction of the descriptive term 'dynamic' to distinguish the properties of unstable DNA sequences from other forms of mutation (16). The frequency of the mutations at the DM1 locus makes them an excellent model system. Hence, DNA samples from individuals with one of these genetic diseases provide an unusual opportunity to estimate the rates of mutation and the number of events underlying the mechanism of DNA instability.

In DM1, the precise mechanism that causes the CTG units to become inserted or deleted from the array is not known (1–3). Expansions occur at different stages of human development and within different tissues, and this instability has been linked to DNA repair, transcription and replication, but the same pathway is not necessarily at work within different tissues (2). Secondary structures, with small loop outs, formed by the expanded CTG repeat length may induce instability through their interference with the DNA replication, recombination and repair processes (17).

Currently, individuals finding out that they or their family are affected by this disease, and wanting to know more about the likely progression of the disease or their reproductive choices, have limited prognostic information available to them. This is partly because variance in modal repeat length, measured usually when the symptoms first present themselves, only accounts for around 25% of the variance in age of onset (18–20). A low correlation between the age of onset of symptoms and modal repeat length is in part due to the anticipation associated with DM1 and sampling bias caused by the tendency for people to be tested only when they or a member of their family presents with symptoms. Thus, there is great potential for more sophisticated modelling and inference techniques to improve the prognostic value of genetic information. More broadly, an accurate model for describing the mutation mechanism in DM1 is likely to give insight into DNA instability in general.

Our extensive data arises from elaborate small-pool polymerase chain reaction (PCR) analysis of repeat length in blood cells from a cohort of 145 individuals with DM1 expansions (F. Morales *et al.*, manuscript in preparation). The cohort includes affected individuals as well as asymptomatic carriers. Since the first application of small-pool PCR to quantify variation at the myotonic dystrophy locus in 1995 (11), the technique has become well established as robust and reliable and has been used to quantify triplet repeat dynamics in a wide range of scenarios and at various loci (21–29). For each individual, we have used single-molecule analysis to size the expanded CTG repeat tract in between 100 and 350 cells (Fig. 1B; Supplementary Material, Figs S2 and S3B), providing a total data set of over 25 000 observations. These data reveal the variation in repeat length between cells and individuals. The shapes of the distributions of repeat length depend on both age and the number of repeats. Older individuals with longer than average repeat lengths have broader distributions than younger subjects with similar repeat lengths, whereas older individuals with shorter repeat lengths have narrow skewed distributions. Subjects from the same family or with potentially the same inherited repeat length can have quite different distributions. These data are highly suited for quantitative treatment to develop mathematical models that capture the key features of the mutation mechanism underlying repeat length evolution.

We construct a model based on a stochastic birth and death process of the form traditionally developed to model the growth of a population (30). Because of their usefulness in counting entities, birth and death models are now applied to many other types of processes where the individuals can involve anything from molecules, cells, tissues, organisms, ecosystems or biospheres (31). Such stepwise models have been used in population genetics to describe the evolution of microsatellites (reviewed in 32) and generally focus on germline mutations. Because the germline mutation frequencies of microsatellites are very low, neutral drift can have a major impact on the relative distribution of alleles in the population. Thus, it is usual also to incorporate shared ancestry into models of microsatellite mutation based on population data. Here, however, we focus on mutations arising in the soma during the lifetime of DM1-affected individuals. When the cell mutation frequency is exceptionally high relative to the number of cell divisions, it is not usual to include shared ancestry (13,33,34). Pathological mutations associated with rapidly changing repeats arising during the lifetime of individuals have also been studied using alternative modelling frameworks. Leeflang *et al*. (33) investigated germline mutation frequency in HD using a simple Okazaki fragment processing model of trinucleotide repeat instability supporting a cell division-dependent mitotic origin for mutations in sperm. More recently, Veytsman and Akhmadeyeva (34) showed that a simple theoretical model of pathological microsatellite expansion based on hairpin formation could offer an explanation for the observed phenomena of mosaicism, anticipation and rare reversions.

Our model builds on Kaplan *et al*. (13) who used a simple birth process (expansion only) to describe repeat length evolution and derived expressions to fit basic clinical and genetic data (age at onset and modal repeat length) for a range of diseases associated with expanded repeats. They were able to demonstrate that somatic mosaicism contributes to disease onset and progression. However, their model is concerned with only expansions, while there is evidence for intergenerational contractions (11,35,36). Thus, we investigate here the possibility that somatic variation is due to the difference between expansion and contraction mutations. We use the

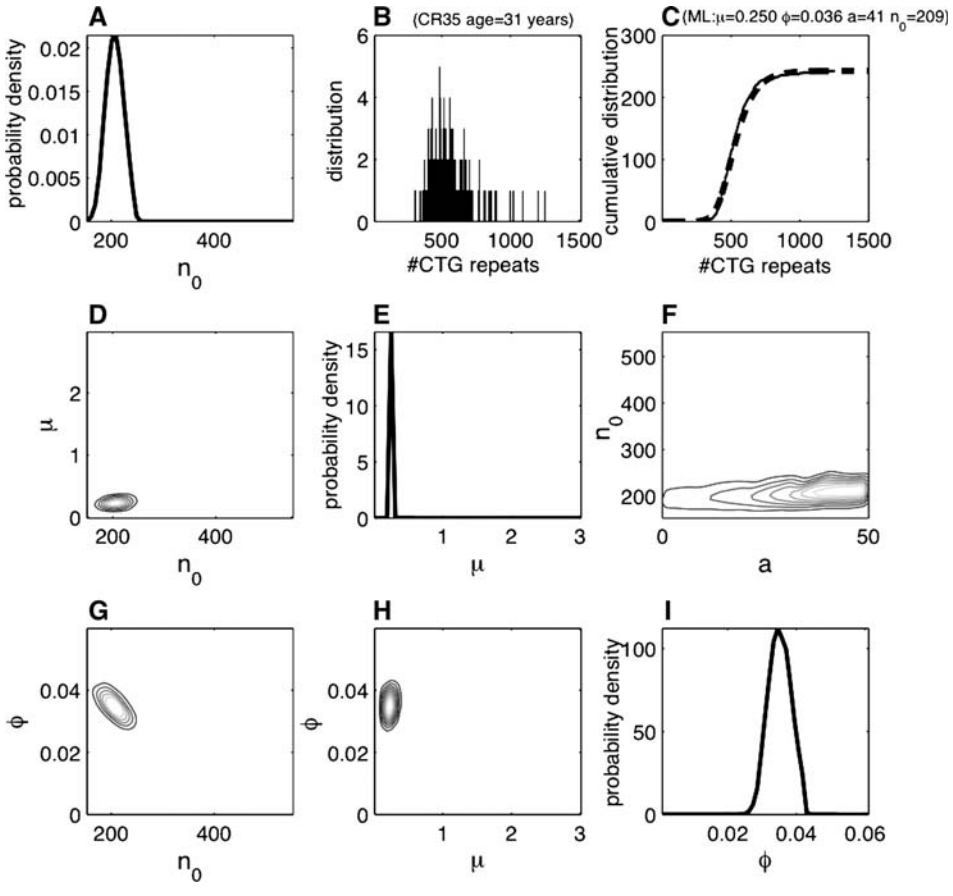

Figure 1. Parameter estimation results for representative individual CR35, aged 30. The data are presented in (**B**) as a histogram showing the distribution of repeat lengths for individual CR35 ($i = 35$). The posterior probability density distributions for parameters $n_0^{[35]}$, the inherited repeat length, $\mu^{[35]}$, the rate of contraction per CTG unit per year, and $\varphi^{[35]}$, the rate of expansion minus contraction per CTG unit per year, marginalized for each parameter over the other parameters, are shown in (**A**), (**E**) and (**I**), respectively. Marginalized joint posterior probability distributions for parameter pairs, $\mu^{[35]}$ and $n_0^{[35]}$, $\varphi^{[35]}$ and $n_0^{[35]}$, $\varphi^{[35]}$ and $\mu^{[35]}$, and $n_0^{[35]}$ and $a^{[35]}$, the threshold number of repeats over which expansion and contraction occur, are shown in (**D**), (**G**), (**H**) and (**F**), respectively, as contours with the dark to light direction representing increasing probability (the probability surface was smoothed slightly using a standard convolution filter to reduce noise). In (**C**), the data, shown as a cumulative distribution (jagged solid line), are compared with the inferred fit with the maximum-likelihood value (dashed line) with associated parameter values $\mu^{[35]} = 0.25$, $\varphi^{[35]} = 0.0358$, $a^{[35]} = 41$ and $n_0^{[35]} = 209$ and $a^{[35]}$.

same stochastic modelling framework as Kaplan *et al*. (13) but extend it to include contractions (death process) and a threshold below which expansion and contraction does not occur. Such a threshold is consistent with the relative stability of the normal allele (11). In the context of this work, we are counting the number of CTG units in the mutant repeat tract within each cell.

The overall aim of this work is to develop a mathematical model that sheds light on the underlying dynamical process of DNA mutation and calibrate it to a large data set. Unlike other applications where only one population may be observed over time, by sampling many cells from individuals, we have many realizations of the same stochastic process at one point in time. Hence, our data provide a unique opportunity to access directly the inherent fluctuations that are required to fit a stochastic process. This enables us to quantify several important biological parameters relating to the mechanism underlying repeat length evolution. This is an important step towards understanding pathological mutations and ultimately providing better prognostic information for individuals with diseases arising from these mutations.

RESULTS

Modelling context for the results

To clarify the presentation and discussion of our results, we begin by stating and justifying our key assumptions about how the data arose. Our mathematical model quantifies the probability of an increase or decrease in the repeat length in blood cells. As circulating white blood cells typically do not replicate, we assumed that the main mutational changes in DNA occur in the progenitor stem cells before cell differentiation and not in the relatively short window between cell differentiation and cell release into the bloodstream. At puberty, the steady-state number of haematopoietic stem cells is estimated between 11 000 and 22 000 (37,38). These stem cells give rise to differentiated multipotent clones that generate around 100 billion blood cells per day over a few weeks before the clone exhausts (37). These circulating blood cells, including erythrocytes and nucleated white blood cells, have lifespans typically ranging from days to weeks. As somatic mosaicism accumulates with age (22,39,40), variation must therefore be accumulating in the population of stem cells.

Stem cells replenish every 40 weeks or so, and hence, typically for the individuals in our study, many generations will have passed since the stem cells shared a common ancestor. At birth, virtually no mosaicism is seen in blood in DM1 patients, even those with the congenital form of the disease (39–42). On this basis, it was reasonable to assume that the stem cells effectively have independent mutational histories. Thus, we interpreted our samples of between 100 and 350 cells as a proxy representation of 11 000–22 000 ultra-progenitor stem cells with each sample informing us about the underlying process. Hence, the stochastic process model was derived under the assumption that the cells have independent mutational histories, and at each continuous point in time, a discrete random variable represents the repeat lengths.

Another key issue for the model formulation was the number of CTGs inserted or deleted at either mutational event. Studies using microsatellite data (43,44) found that the majority of insertions or deletions were of one repeat unit. Data from individuals with HD, where a similar mechanism underlies DNA instability but where the alleles are smaller and there is less variation, provided an opportunity to observe the possible number of repeat units that might be inserted or deleted at one mutation event. The patterns of length distribution in these data (45,46) suggest that the inserted or deleted tracts are predominantly one repeat unit long but may include occasional longer lengths up to 5–15 repeat units. So, in our case, it was a reasonable, working assumption that the birth and death process treats one CTG as the individual unit and we associate 'birth' with expansion and 'death' with contraction.

The features of the mechanism underlying repeat length instability are largely unknown. By fitting different models which represent different hypotheses about this mechanism to the data set, we can use model comparison methods to rank the hypothetical models in order of best fit. Thus, we can establish which models are more likely to explain the data than others. Is the underlying process driven by expansion only, as hypothesized by Kaplan *et al*. (13), or could it be a combination of expansion and contraction? Are the rates of expansion and contraction universal or are there significant differences between individuals indicating the influence of individual-specific factors? Is there a fixed or individual-specific number of repeats around the instability threshold of 40 CTGs?

To answer these questions, we defined the following eight models: expansion only with a global parameter for expansion and a fixed threshold (Model 1); expansion only with an individual-specific parameter for expansion and a fixed threshold (Model 2a); expansion only with individual-specific parameters for expansion and threshold (Model 2b); expansion and contraction with global parameters for expansion and contraction, and a fixed threshold (Model 3); expansion and contraction with a global parameter for contraction and individual-specific parameters for expansion and threshold (Model 4); expansion and contraction with a global parameter for expansion and individual-specific parameters for contraction and threshold (Model 5); expansion and contraction with individual-specific parameters for expansion and contraction, and a fixed threshold (Model 6a); and finally, expansion and contraction with individual-specific parameters for expansion, contraction and threshold (Model 6b).

Table 1. Parameter estimation

Parameters	Range for uniform prior[a] (small alleles)	Incremental step size for parameter exploration
Contraction, rate per CTG unit per year, μ	0.01–3.01	0.06
Expansion minus contraction, rate per CTG unit per year, φ	0.001–0.061	0.0012
Threshold, number of CTG units, a	0–50	1
Inherited repeat length, number of CTG units, n_0	82 to PAL[b] + 200[2] (51–81)[1]	8 (2)[1]

[a]This range was adapted for some patients with: [1]small alleles in order to investigate smaller rates of contraction (see figures in parentheses); [2]possibly unreliable PAL estimates due to distributions that were spread out or ambiguous in another way. This included both extending PAL + 200 up to the maximum possible value (in an expansion and contraction model, this would be the data mean) and down to the pathological disease threshold of 50.
[b]Progenitor allele length (PAL) was broadly estimated from the small-pool PCR data which resolves the cells into different lengths based on the position of the 10th percentile or a sharp lower bound if one existed. This measure can only be considered a rough estimate and the priors are set wide of this mark to eliminate any bias that this estimate could introduce into the inference procedure.

Model comparison

We used model comparison methods to evaluate several hypotheses relating to the mechanics of how the distributions of repeat lengths arise in samples of blood DNA, the shape of which can differ between individuals depending on their age when the sample was taken and the size of the repeat lengths. Since a likelihood arises naturally from the stochastic process, both Bayesian and non-Bayesian likelihood methods lend themselves to fitting the data to the model. We used the maximized log-likelihood with the Akaike information criterion (AIC) and the likelihood ratio test as the basis for model comparison. The likelihood is also employed as part of a Bayesian framework with prior information to provide parameter distributions.

Data comprising the distribution of CTG repeat lengths within a blood sample from 142 individuals (out of 145 individuals tested) were used to fit the eight models, described above, representing the different hypotheses. As detailed in the Materials and Methods sections, three individuals were excluded from this analysis. In the most general case, we had the following unknown parameters for each individual: the number of CTG units from which the process started, otherwise known as the progenitor or inherited allele length, n_0; the threshold number of CTG units over which expansion and contraction are non-negligible, a; the rates of expansion and contraction, over this threshold, per CTG unit per year, λ and μ, respectively, which define the net expansion rate, $\varphi = \lambda - \mu$. These parameters were treated as unknowns and investigated over a broad range of values (Table 1). To formally compare the different models, we used the AIC (47,48) and the likelihood ratio test (49) which both employ the maximized log-likelihood penalized by the number of parameters in the model, summarized in Table 2. The likelihood of the

Table 2. Model comparison summary

Models, $N = 142$ individuals	Parameters	Number of parameters	Maximized log-likelihood	AIC	Adjusted AIC[1] (rank)	Likelihood ratio test (rank)
6a. Expansion and contraction with individual-specific parameters for expansion, contraction and a fixed threshold	$a^g = 40$, $\lambda^{[i]}$, $\mu^{[i]}$, $n_0^{[i]}$	427	−135 614	272 082	0 (1)	(2)*
6b. Expansion and contraction with individual-specific parameters for expansion, contraction and threshold	$a^{[i]}$, $\lambda^{[i]}$, $\mu^{[i]}$, $n_0^{[i]}$	568	−135 523	272 182	100 (2)	(1)*
5. Expansion and contraction with a global parameter for expansion, an individual-specific parameter for contraction and a fixed threshold	$a^g = 40$, λ^g, $\mu^{[i]}$, $n_0^{[i]}$	286	−136 392	273 356	1274 (3)	(3)*
2a. Expansion only with an individual-specific parameter for expansion and a fixed threshold	$a^g = 40$, $\lambda^{[i]}$, $n_0^{[i]}$	285	−136 721	274 012	1930 (4)	(5)
2b. Expansion only with individual-specific parameters for expansion and threshold	$a^{[i]}$, $\lambda^{[i]}$, $n_0^{[i]}$	426	−136 613	274 078	1996 (5)	(4)
4. Expansion and contraction with a global parameter for contraction, an individual-specific parameter for expansion and a fixed threshold	$a^g = 40$, $\lambda^{[i]}$, μ^g, $n_0^{[i]}$	286	−139 852	280 276	8194 (6)	(6)
3. Expansion and contraction with global parameters for expansion and contraction, and a fixed threshold	$a^g = 40$, λ^g, μ^g, $n_0^{[i]}$	145	−140 807	281 904	9822 (7)	(7)
1. Expansion only with a global parameter for expansion and a fixed threshold	$a^g = 40$, λ^g, $n_0^{[i]}$	144	−187 051	374 390	102 308 (8)	(8)

The models, listed in column 1, were ranked according to their AIC score which penalizes the maximized log-likelihood by the number of parameters.
*Significantly ($P < 10^{-15}$) better than Model 2a.
[1]AIC adjusted by subtracting the lowest value overall (272 082 model 6a) from each model.

data given the model arose naturally from the stochastic process and we obtained the maximized log-likelihood value for each model. Further details of how the models and their likelihood were derived are found in the Materials and Methods section.

The size of the maximized log-likelihoods, around -1.35×10^5, reflects the vast quantity of data (between 100 and 350 samples for each of the 142 individuals) and leads to correspondingly large AICs. But what is important for model comparison is not the absolute value of AIC but the difference between models, with more supporting evidence for the model with the lowest value of AIC. To see this more clearly, we adjusted each AIC by subtracting the lowest value overall and ranked the models in order, with the smallest difference and hence strongest model first. We conclude that there is most support for Model 6a (expansion and contraction with individual-specific parameters and a fixed threshold) and Model 6b (expansion and contraction with individual-specific parameters and a variable threshold) with adjusted AICs of 0 and 100, respectively, followed by Model 5 (expansion and contraction with a global parameter for expansion, an individual-specific parameter for contraction and a fixed threshold) with an adjusted AIC of 1274. Expansion-only Models 2a and 2b have adjusted AICs of 1930 and 1996, respectively. Comparing Models 6a and 6b using the likelihood ratio test indicates that the difference between these models is of low significance ($P = 0.01$). However, comparing Models 6a, 6b and 5 to Model 2a using the likelihood ratio test gives a highly significant result ($P < 10^{-15}$). The Bonferroni correction for eight multiple tests, applying the standard significance level ($\alpha = 0.05$) to one test, is 0.00625. This strongly supports the hypothesis that contractions are present in the underlying process of repeat length evolution.

The models with individual-specific parameters, both with and without contractions, are better supported by the AIC evidence, ranging from 0 to 8194, than the models with global parameters, AICs ranging from 9822 to 102 308 (ranked 7–8 in Table 2). This suggests that there is significant parameter variation between individuals. When considering the threshold parameter, a, we observe that fixing the value at 40 CTGs provides as good a fit to the data as individual values, providing support for the involvement of a universal length effect in the mechanism of repeat instability. This finding is consistent with the observed instability threshold of around 40 repeats in DM1 (5,6).

Parameter estimation

The model fitting produces some evidence for individual variation in μ and φ. The maximum-likelihood approach provides point estimates of parameters, but it is also desirable to have information on the parameter distributions. We compute the parameter distributions for each individual using a Bayesian framework which fully takes into account any uncertainty arising from the finite nature of the sample for each DM1-affected individual and the PCR technique. As elaborated in the Materials and Methods section, the effect of the finite sample outweighs that of the PCR technique and simulation experiments investigating sample size show that we have enough individually sized alleles from each DM1-affected individual to satisfactorily infer the parameters of interest, namely, expansion and contraction rates, and the inherited repeat length (Supplementary Material S1).

The parameters ($\lambda^{[i]}$, $\mu^{[i]}$, $a^{[i]}$ and $n_0^{[i]}$ where a particular individual is denoted $[i]$ and $i = 1, \ldots, 142$ corresponding to the 142 individuals analysed) were treated as unknowns and their

probable values were inferred from the data using a Bayesian framework and biologically informed prior for each parameter (Table 1). This approach provided not only the most probable value for each parameter but also a credible range. In some cases, there is evidence of suboptimal solutions. By presenting the results in this way, we retain a full picture of the parameter solution space which is particularly important when the model has non-linear components causing such suboptimal solutions to arise. We report the parameter estimates as probability density functions, or posterior distributions, the peaks of which indicate the most probable parameter values while capturing any uncertainty in the prediction. The results for individual CR35 ($i = 35$), (Fig. 1), are typical of many individuals. The parameter with the highest posterior probability peak, and hence for which the data are the most informative, is the contraction rate $\mu^{[35]}$ (Fig. 1E). The peak is located at 0.25 contractions per CTG unit per year. For parameters $n_0^{[35]}$ and $\varphi^{[35]}$ (Fig. 1A and I) peaking over 209 CTGs and 0.0346 expansions minus contractions per CTG unit per year, respectively, the posterior distributions are wider than that for $\mu^{[35]}$. Given the range of repeat lengths sampled for this individual (between 300 and 1300 CTGs), the posterior distribution for $a^{[35]}$ is best viewed jointly with n_0 (Fig. 1F). The resulting contour is widely spread over the range for $a^{[35]}$ (0–50 CTGs), implying that the observed repeat lengths, for this particular individual, are not informative for this parameter. This is because the observed repeat lengths are much greater in length than the plausible range for the threshold (below 50 CTGs), and consequently, we conclude that parameter $a^{[35]}$ has little effect on the dynamics of repeat length evolution for this particular individual. Inspection of the joint probabilities for pairs of parameters can indicate interdependencies between parameters. For many individuals, there is a trade-off between φ and n_0 concerning the best fit, as illustrated by the skewed contour (Fig. 1G).

The parameter values associated with the maximum likelihood are presented for each DM1 individual (Supplementary Material, Fig. S3C). The average expansion rate is 0.53 CTGs per CTG unit per year, and the average contraction rate is 0.51 CTGs per CTG unit per year. The resulting net expansion (expansion minus contraction) is 0.02 CTGs per CTG unit per year. A relatively small gain is achieved by very many expansions and contractions. Interestingly, although there is a lot of individual-specific variation in the mutation rates, the correlation between expansion rates and contraction rates across the 142 DM1 individuals is very high ($R^2 = 0.99$, $P < 0.0001$).

Model fit

Models 6a and 6b fitted the data equally well but as Model 6b contains information about the threshold, we consider further the fit of Model 6b (expansion and contraction with individual parameters) to the data. The maximum-likelihood solution ($\mu = 0.25$, $\varphi = 0.036$, $a = 41$ and $n_0 = 209$) traces closely the rising slope of the cumulative data (Fig. 1C). As for every individual (Supplementary Material, Fig. S3), the inferred value of μ is clearly non-zero under the expansion and contraction model. Further to this, the maximum log-likelihood of the expansion and contraction model (-1495) is greater than the maximum log-likelihood of the expansion-only model (-1511); see individual CR35 Supplementary Material, Table S1. Capturing the variance seen in the data is key to fitting these models. In the expansion and contraction model, the variance seen in the data is the result of both expansion and contraction. The contraction process is playing an important role in generating the variance in the data. In the expansion-only model, the observed variance can only be explained by an inherited repeat length below the lowest observed repeat length. As well as a poorer fit, indicated by the AIC analysis, the resulting predicted inherited allele length, n_0, from the expansion-only model is also implausibly close to the range seen in the general population (5–37 CTGs) which would argue against this being a disease allele in the first place. For illustrative purposes, the time-dependent distribution generated first by the expansion and contraction model and second by the expansion-only model was simulated for 120 cells with an initial repeat length of 160 CTGs over 30 years (Supplementary Material S1 and Videos S1 and S2, respectively). In each scenario, the expansion bias was set at 0.02 CTGs per CTG unit per year. Inspection of the resulting distributions confirms that repeat length variance is much greater under the expansion and contraction model, whereas the mean repeat length is the same for each model. Under expansion only, the distribution lies above the initial repeat length. These simulations visually confirm the higher plausibility of the expansion and contraction model and support our more rigorous statistical finding that contractions underlie this mutational mechanism. Further visual evidence of the model fit is provided by comparing simulations, based on the parameter estimates for six DM1 individuals with different ranges of allele lengths, with the original autoradiographs (Supplementary Material, Supporting Text S1 and Fig. S2).

The full model (6b) assumed that the rates of expansion and contraction are linearly proportional to the repeat length beyond a threshold. Equivalently, each CTG unit beyond the threshold is equally likely to give rise to an event. The fitting of this model to the data suggests that this assumption is a good approximation for the majority of individuals (121 out of 142) whose repeat lengths lie in the mid-range. This excludes congential cases where the repeat length is very high and asymptomatic individuals whose repeat lengths are relatively low. For low-range individuals (allele lengths <200 CTGs), contraction rates cluster around the low end of the parameter spectrum (Fig. 2A). For high-range individuals (allele lengths >800 CTGs), expansion minus contraction values cluster around the low end of the spectrum (Fig. 2B). In both cases, it is reasonable to expect these rates to be randomly distributed throughout the spectrum. These results provide an indication that the overall model may be improved further by introducing a non-linear response in line with differences in the biology of small alleles or large alleles. Small alleles may have a reduced propensity to expand or contract due to possible end effects and there may be a mechanism either limiting the expansion of the large alleles or causing more contraction. To fit fully such a non-linear response requires additional analysis among low-range individuals and individuals bridging the mid-range and the high range, and will be the focus of future work if appropriate data can be collected.

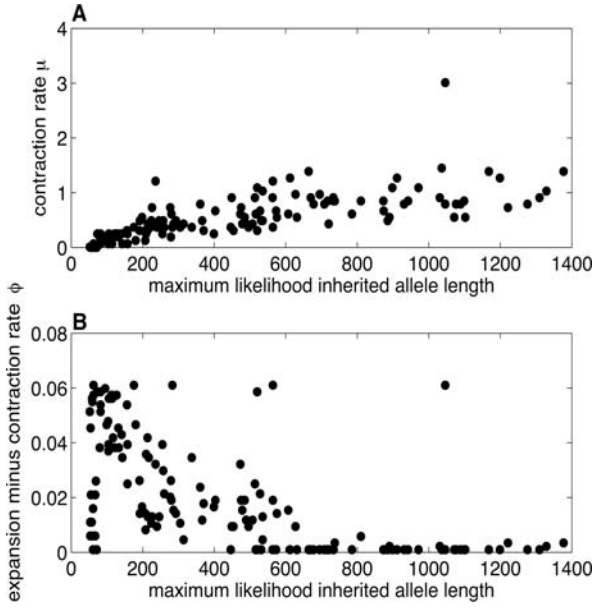

Figure 2. Scatter plot of the maximum-likelihood parameter values. (**A**) Contraction, rate per CTG unit per year, μ, on the vertical axis versus inherited allele length n_0 on the horizontal axis ($n = 142$). (**B**) Expansion minus contraction, rate per CTG unit per year, $\varphi = \lambda - \mu$.

Hierarchical Bayesian analysis

Given there is support for individual variation in μ and φ, the aim of the hierarchical Bayesian analysis was to use the data to predict the probable range and distribution of μ and φ in the general DM1 population. To do this, we make some assumptions about the shape and scale of the underlying distribution, which are summarized as the *prior* information (Table 3). This information reflects our knowledge about the mutation rates before analysing the data. In our case, the gamma distribution is a good choice as it necessarily lies over positive values and allows for the possibility that the distribution may be skewed, either towards zero or with a long tail. The shape and scale of the gamma distribution ensures that a wide range of possibilities were considered. This analysis effectively weights the probability of each parameter value of interest by the probability that it could have arisen from each of the underlying distributions under consideration. For this analysis, we considered first all our individuals together ($N = 142$) and secondly, the subset of individuals who do not have the congenital form of the disease but do have symptoms ($N = 121$). By excluding those diagnosed at birth or those asymptomatic individuals who have yet to develop symptoms, we focus on the group for whom progression of the disease is most variable and hence diagnosis most open and pertinent. The range of shared values for all 142 individuals peaks at 0.14 contractions per CTG unit per year, and the subgroup group of 121 individuals peaks at 0.25 contractions per CTG unit per year (Fig. 3A). For φ, the shared values peak at around 0.0026 expansions minus contractions per CTG unit per year ($N = 142$) and 0.0032 expansions minus contractions per CTG unit per year ($N = 121$) (Fig. 3B). The credible interval (5th–95th percentile) for this prediction is shown as a shaded grey area (Fig. 3). All distributions are skewed towards the right with long tails. The lower rates, when all individuals are included, indicate that something different is happening with the very short and very long alleles.

Table 3. Hierarchical Bayesian analysis

Distribution	Hyper parameters	Range for uniform prior	Incremental step size for parameter exploration
Γ_μ (α_μ, β_μ)	Mean $\alpha_\mu\beta_\mu$	0.3–0.8	0.01
	Variance $\alpha_\mu\beta_\mu^2$	0.05–0.55	0.01
Γ_φ (α_φ, β_φ)	Mean $\alpha_\varphi\beta_\varphi$	0.005–0.03	0.0005
	Variance $\alpha_\varphi\beta_\varphi^2$	0.0001–0.0006	0.00001

For the hierarchical Bayesian analysis, we require an assumption about the shape of the distribution underlying the model parameters of interest, μ and φ, and priors, which encapsulate any information we may have, for the parameters of that distribution. We assume that the distribution underlying μ, the rate of contraction per CTG unit per year is a gamma distribution, Γ_μ, defined by a shape parameter α_μ and a scale parameter β_μ, as the gamma distribution has many different forms over positive values. The mean and variance of this distribution are $\alpha_\mu\beta_\mu$ and $\alpha_\mu\beta_\mu^2$, respectively, and we chose, for convenience, to place our priors on the mean and variance, to ensure that we cover a range of possible shapes for this distribution. For φ, the rate of expansion minus contraction per CTG unit per year, we also assume that the underlying distribution is a gamma distribution, Γ_φ, defined by shape parameters α_φ and β_φ.

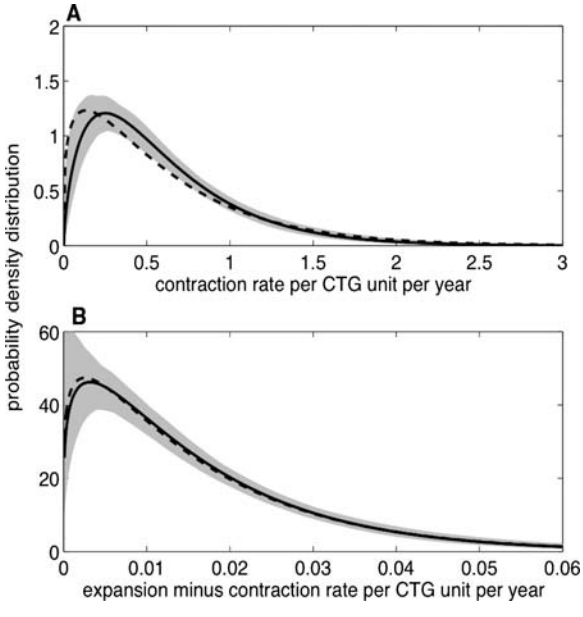

Figure 3. Hierarchical Bayesian analysis results. (**A**) The modal distribution of the contraction rate (dark line) for all individuals except those who have had DM since birth (congenital) or who have no symptoms yet (asymptomatic), 121 individuals in total. (**B**) The modal distribution of the expansion minus contraction rate (dark line) for the same 121 individuals. The shaded area, in both panels, represents the 5–95 percentile credible range. The modal distribution for all 142 individuals is shown by the dotted line.

DISCUSSION

We have shown that a thresholded stochastic birth and death process, where birth represents expansion and death

contraction, can explain a wide range of repeat length distributions arising in the blood cells of individuals with DM1. This conclusion remains valid both when individuals and the population as a whole are considered.

Alternative modelling frameworks for pathological mutations associated with rapidly changing repeats have been proposed. Leeflang *et al.* (33) investigated germline mutation frequency in HD using a simple Okazaki fragment processing model of trinucleotide repeat instability. This model could be fitted very nicely to sperm data and revealed support for a mitotic cell division-dependent mutational mechanism in the rapidly dividing spermatogonial stem cells in the male germline. In contrast, our data do not support an association with mitotic cell division in the haematopoietic stem cell population with hundreds of mutations predicted each year (see below) relative to a stem cell renewal rate of once every 40 weeks (37). Interestingly though, Leeflang *et al.*, did, as did we, reveal evidence for individual-specific mutational parameters, suggesting that both germline and somatic instability are modified by as-yet unknown genetic and/or environmental factors. More recently, Veytsman and Akhmadeyeva (34) showed that a simple theoretical model of pathological microsatellite expansion based on hairpin formation, including both expansions and contractions, could offer a qualitative explanation for the observed phenomena of mosaicism, anticipation and rare reversions. However, this model did not incorporate any *in vivo* somatic data and thus the actual parameters could not be calculated. Our model builds on Kaplan *et al.* (13) who used a simple birth process to describe repeat length evolution. Because their data were limited to modal summaries, it did not indicate any variation that might be present within an individual, making it impossible to distinguish between expansion and contraction. Hence, their work assumed that the expansion bias observed in individuals is solely due to expanding lengths. In contrast, for each DM1 individual, the data that we use in our study effectively provide between 100 and 350 outcomes of a stochastic process in the somatic blood cells sampled at a single point in time. In total, over 25 000 repeat lengths were sized representing one of the largest databases of its kind. Of those alleles, around 20 000 are estimated to be *de novo*, having arisen during the lifetime of individuals. So as well as information about the mean behaviour of this process, we also have information about the variation and distribution. This allows us to uncover more aspects of the underlying mechanism, increase the fitting capacity and obtain more information about the parameters of the biological processes involved in DM1.

The key question we posed was whether the variation observed in these repeat lengths is solely due to expansion, as implicitly assumed in the model of Kaplan *et al.* (13), or whether it is the combined result of expansions and contractions. We also wanted to establish how much variation exists between individuals. To address these questions in a rigorous, statistical way, we formulated the hypotheses as a series of models and then ranked them using AIC and the likelihood ratio test. There was most support for the expansion and contraction model with individual-specific parameters. Previously, it was thought that the expansion bias observed in individuals was mostly due to expansions with relatively rare incidences of contractions. We show that the observed expansion bias is actually the difference between expansions and contractions. Consequently, there are many more mutational events in total, comprising both expansions and contractions, than an expansion-only model would predict. Our results suggest that a relatively small net gain of two repeats may arise from 100 expansions and 98 contractions: in total 198 mutational events. This makes the DM1 locus even more hypermutational than we thought and is a provocative hypothesis for future experimental research. The closeness of the contraction and expansion rates could be experimentally verified with various model systems such as transgenic mice, assuming that the mechanisms and dynamics are accurately reflected in such models. While transgenic mouse models do not usually show large intergenerational changes, substantial expansion-biased and age-dependent somatic length changes of many hundreds of repeats are observed in some somatic tissues (but not usually in blood) (21,50,51).

The expansion and contraction rates are assumed to be constant with age. With one sample from each patient, it is not possible to distinguish clearly an age effect from another effect (genetic or environmental). Repeat samples from the same DM1 individuals at different ages would allow us to test whether the individual-specific rates of contraction and expansion vary over time. With another time sample, we could assume that other effects are constant and quantify temporal changes. Collection of further samples is currently underway in a longitudinal study which would address this issue.

For a 30-year-old individual with an inherited repeat length of 200 and a net gain of two repeats per 100 expansions, the model predicts $\sim$5500 expansion and contraction events per cell during their lifetime, which is $\sim$1 event every other day. Significantly, for establishing a causal link for instability with DNA replication, this number is not consistent with the number of stem cell divisions, once every 40 weeks (37). Rather, this number links the mutation process with the time scale of other more frequent cell activities such as DNA repair and transcription. Compared with estimates of the amount of DNA damage endured each day in a white blood cell, which is thought to be over 10^4 events and may be as many as 10^6, over the 3.2×10^9 bp of the genome, discussed in (52) and (53), mutational events at the DM1 locus are occurring between 10 and 100 times more frequently. The strong link between expansion and contraction rates within an individual may arise from similarities in the mutational mechanism, suggesting that expansions and contractions may result from the stochastic effects of one biological process rather than two. Further support for this idea is provided by studies of transgenic mice in which the expanded repeat is completely stabilized in either an *Msh2* or *Msh3* null background (54,55), implying that both the underlying expansions and contractions have been affected by loss of function of the same pathway.

Longer DM1 alleles transmitted to the next generation result in more severe symptoms and an earlier age at onset, an effect compounded by somatic expansion (F. Morales *et al.*, manuscript in preparation). As such, suppression of somatic expansion is expected to be therapeutically beneficial and induction of contractions potentially curative (1,14). However, the feasibility of suppressing expansions/inducing contractions remains largely undetermined. Our results have revealed that the

mutational pathway is even more dynamic than previously envisioned and that although overall biased towards expansion, net gains are the product of a very subtle bias towards expansions relative to almost equally frequent contractions. The high underlying frequency of contractions suggests therefore that a therapeutically beneficial impact may be mediated by a relatively subtle shift in the relative bias from small expansions towards small contractions. With the underlying expansion and contraction frequencies so closely matched, either a 3% decrease in the basal expansion frequency or a 3% increase in the basal contraction frequency would result in a net loss of repeats over time. Such a subtle intervention would appear more pharmacologically achievable than the major suppression of expansions foreseen as required in an expansion-only system.

The hierarchical analysis establishes the underlying distribution for parameters μ and φ by effectively weighting the evidence from individuals to form a population prediction. This prediction is based on individuals who have developed symptoms since birth and who represent the group for which prognosis is most variable. The results for μ suggest that population rates peak at 0.25 contractions per CTG unit per year. For φ, which represents the difference between the expansion and the contraction rate, the values peak at 0.0032 per CTG unit per year. This analysis supports the model comparison finding that individual parameters give rise to the best model fit. This indicates that individual-specific factors, either environmental or genetic or both, may influence instability.

DM1 is a multisystemic disease with even patients from the same family varying in age of onset, symptoms and the progression of the disease. Our model is calibrated to blood which although not a primary target of the disease is easily accessible in a large number of patients and is a tissue within which the repeat remains relatively stable compared with other tissues in which the main symptoms of the disorder are manifest. Analysing blood DNA thus gives us the best chance to estimate the progenitor allele length which is most indicative of age of onset (F. Morales *et al.*, manuscript in preparation). Future studies that collect data from different tissues along with more detailed information about disease progression would in theory allow us to investigate the underlying mechanism of instability in different tissues and determine stability in other tissues. However, there are several challenges to measuring instability in other tissues. Not least of these is the availability of large numbers of samples from tissues not routinely sampled during the diagnostic procedure. With the availability of genetic tests based on blood DNA, muscle biopsies are considered far too invasive for routine testing, and most other tissues are only available post-mortem. In addition, more complex tissues often display multimodal distributions (e.g. the bi-modal and tri-modal distributions typically observed in the mouse liver and kidney) (21,28), likely reflecting the presence of very different cell types within the same tissue. Dissecting the relative contribution of different cell types with different mutational dynamics would be even more challenging. In addition, the very large expansions observed in most other tissues of DM1 patients (frequently many thousands of repeats) (56,57) pose technical challenges and are not amenable to routine small-pool PCR analysis. It is not known why the lengths are so much longer in tissues such as the muscle, but our work now provides two alternative explanations. The basic mechanism may resemble that in blood, with even greater underlying expansion and contraction frequencies. Alternatively, a greater net expansion frequency may be mediated by a greater difference in the underlying expansion and contraction ratio.

Our model could also be extended to other triplet repeat expansion diseases (such as HD) depending on the availability of suitable data sets. However, compared with DM1, the expanded repeat tract in most other triplet repeat diseases is relatively stable, particularly in blood. Other tissues such as the brain are difficult to obtain and have a greater complexity than blood in terms of cell composition which would necessitate adding additional parameters partitioning mutations between cell types. If the model could be calibrated to another disease, we would expect differences in the parameter values but similarity in the underlying mechanism.

Mathematical modelling and inference of somatic DNA dynamics at the DM1 locus has enabled the estimation of biological parameters, inherited repeat length and mutation rates, which could not otherwise be obtained. The level of these measures provides a deeper understanding of the underlying mechanisms and we can use a calibrated model to answer scenarios and to make predictions. F. Morales *et al.* (manuscript in preparation) found that the inherited CTG repeat length is potentially much better than the current modal CTG repeat length measure taken during diagnosis of the expansion repeat diseases at explaining the age of onset and the progression of the disease. This is partly because the analysis of the modal repeat length is confounded by the tissue and age specificity of somatic mutations. With one blood DNA sample, our method can broadly estimate the most probable inherited repeat length. Data from another time point could in principle narrow this estimate even further, and future work will aim to establish this.

Further, these quantitative traits, μ and φ, are potential biomarkers that can be used via the genome-wide association study to identify *trans*-acting genetic factors thought to be linked to this somatic variation (F. Morales *et al.*, manuscript in preparation). Our expectation is that these *trans*-acting genetic modifiers will also apply in the general population where they will affect ageing, cancer, inherited disease and human genetic variation.

MATERIALS AND METHODS

Project data

The data analysed in this study (F. Morales *et al.*, manuscript in preparation) were derived from a large cohort of individuals with DM1 expansions ($>$50 repeats). The total cohort comprised 145 individuals. In addition to a normal allele, two individuals (CR51 and CR115) presented an expanded allele with two distinct modes. The two modes likely represent the products of an early embryonic mutation (58,59), and because of our inability to clearly apportion additional variants to either of these two progenitors, these individuals were excluded from the model comparison analysis. In addition, one other individual (CR105) who presented with very high levels of instability despite their very young age at sampling

was therefore also excluded from the model comparison analysis.

Small-pool PCR analysis was performed using oligonucleotide primers DM-C and DM-BR, as described previously (11) (Supplementary Material, Fig. S2). For the detailed quantification of the degree of somatic variation, samples were amplified with 10–70 pg DNA per reaction and at least 100 single expanded alleles per individual were sized. PCR products were sized using Kodak Molecular Imaging software 3.5.4 (Carestream Health, Inc.). Further details can be found in (F. Morales *et al.*, manuscript in preparation). The data can be visualized as allele length frequencies in a histogram format (Fig. 1B; Supplementary Material, Fig. S3B), and the mathematical models describe these distributions using the biological parameters of interest.

As discussed in the introduction, small-pool PCR is a well proven method that provides a robust approach to quantification of length variation in the blood DNA of myotonic dystrophy patients. However, PCR and other technical artefacts (particularly PCR stutter) can confound the interpretation of the data. When analysing the products of single molecules, the effect of PCR stutter is greatly reduced and has been estimated to be at most one single repeat at 35 cycles of PCR (27). In our case, as well as minimizing PCR stutter by employing fewer cycles of PCR [28 cycles], the underlying variation is typically spread over many hundreds of repeats. We consider that most of the uncertainty in our parameter estimation arises from the finite sampling of a highly diverse distribution with only a small contribution from PCR artefacts such as PCR stutter. By applying our parameter estimation method to a synthetic data set where the parameter values are known, we can quantify this level of uncertainty and these results are discussed in Supplementary Material S1.

Mathematical model

Because of the inherent stochasticity in the observed data, with individual cells evolving independently, we develop a model using the framework of birth (expansion) and death (contraction) processes. We refer to Renshaw (30) for further motivation and explanation of the basic methodology. We depart from the traditional linear model by introducing a threshold for the birth and death process. No activity takes place for repeat lengths below this threshold, and the general propensity for expansion or contraction is proportional to the excess length above the threshold, consistent with the inherent stability observed in non-diseased individuals.

Suppose that the length, defined as the number of consecutive CTG units, is n at time t. Let λ be the rate of expansion above the threshold length, a, μ the rate of contraction above a and s the step size. Then, at time $t + \delta t$:

- the probability that the length is $n + s \approx \lambda(n - a)\delta t$
- the probability that the length is $n - s \approx \mu(n - a)\delta t$
- the probability that the length is $n \approx 1 - (\lambda + \mu)(n - a)\delta t$

For reasons covered in the introduction, the step size s in our model is one CTG unit. However, the model could be extended to other step sizes by appropriate adjustment to the expressions above.

Let $P_n(t)$ denote the probability that an allele has length n at time t. Then, the rate of change of $P_n(t)$ with respect to time is governed by the master equation

$$\frac{dP_n(t)}{dt} = -(\lambda + \mu)(n - a)P_n(t) + \lambda(n - a - 1)P_{(n-1)}(t) + \mu(n - a + 1)P_{(n+1)}(t), \quad (1)$$

where $P_k(t) \equiv 0$ for all $k < a$, since $n > a$ at $t = 0$ for all individuals with the pathological condition. Given the allele length at time zero, we may solve this infinite system of ordinary differential equations numerically by truncating the system at a suitably large value of $n = N$, setting $P_n(t) = 0$ for all $n \geq N + 1$.

We can derive expressions for repeat length mean and variance from the first and second moments of $P_n(t)$, denoted $M(t)$ and $M_2(t)$, respectively:

$$M(t) = \sum_{n \geq a} nP_n(t), \quad (2)$$

$$M_2(t) = \sum_{n \geq a} n^2 P_n(t). \quad (3)$$

Differentiating both Equations (2) and (3) with respect to t and substituting Equation (1) into the result leads, after some manipulation, to

$$\frac{dM(t)}{dt} = (\lambda - \mu)(M(t) - a), \quad (4)$$

$$\frac{dM_2(t)}{dt} = 2(\lambda - \mu)M_2(t) + [\lambda + \mu - 2a(\lambda - \mu)]M(t) - a(\lambda + \mu). \quad (5)$$

Solving Equations (4) and (5) with $M(t = 0) = n_0$ and $M_2(t = 0) = 0$, where n_0 is the length of the inherited repeat length and setting $V(t) = M_2(t) - (M(t))^2$ for the variance at time t give the analytical expressions (6) and (7).

Analytical expressions for mean and variance

Equations (6) and (7) link measurable quantities of the mean and variance found in the blood DNA samples to the biological parameters which underlie the mechanism of repeat length evolution:

$$M(t) = (n_0 - a)\, e^{(\lambda - \mu)t} + a, \quad (6)$$

$$V(t) = (n_0 - a)\left(\frac{\lambda + \mu}{\lambda - \mu}\right)\left(e^{2(\lambda - \mu)t} - e^{(\lambda - \mu)t}\right) \quad (7)$$

where we recall that t is the age of the individual in years when the samples were collected, n_0 the repeat length at $t = 0$, which is referred to as the inherited repeat length, λ and μ the rates of expansion and contraction, per CTG unit per year, respectively, and a the threshold above which non-negligible expansion and contraction occurs.

We see from Equation (6) that the mean repeat length changes exponentially over time at a rate determined by the difference $\varphi = \lambda - \mu$. It follows that values for λ and μ cannot be extracted individually from the mean data alone. Only the difference can be found this way. However, the variance depends on the difference between λ and μ, but also on the sum, $\lambda + \mu$. As our data comprises many samples, resolved at the cell level, from individuals, it is possible to estimate both mean and variance making it feasible to fit $\lambda - \mu$ and $\lambda + \mu$ and hence obtain λ and μ individually.

Model comparison and parameter estimation

We use likelihood methods to carry out model comparison and parameter estimation. The likelihood is defined to be the probability that a repeat length has reached the length observed given the model and its parameters. We can solve Equation (1) numerically in order to obtain the probability distribution function $P_n(t)$ which is the probability that a repeat length is length n at time t. The likelihood $L^{[i]}$ is then the product over all the data d_j, which denotes the repeat length for the jth observation from individual i, of the probability $P_{d_j^{[i]}}(t^{[i]}; \theta^{[i]})_{n \geq a}$, where $\theta^{[i]}$ are the model parameters for that individual and $t^{[i]}$ the age of the individual when the data sample was taken. This gives the likelihood for individual i,

$$L^{[i]} = \prod_j P_{d_j^{[i]}}\left(t^{[i]}; \theta^{[i]}\right), \tag{8}$$

and the overall likelihood L is the product over all individuals in the population,

$$L = \prod_i L^{[i]}. \tag{9}$$

The model parameters comprise the contraction rate, $\mu^{[i]}$, the expansion minus contraction rate, $\varphi^{[i]}$, the threshold, $a^{[i]}$, and the inherited repeat length, $n_0^{[i]}$.

As a proof-of-principle for the inference procedure, we performed computational experiments on synthetic data, generated from the underlying stochastic birth–death process with known parameter values (Supplementary Material S1). This gives us an indication of the level of certainty available from the inference procedure.

Model comparison

The AIC is used to assess the goodness of the fit of the model (47). AIC uses the maximized value of the likelihood of the model, $L_{\max}$, penalized by the number of model parameters, k, to rank models, thus

$$\mathrm{AIC} = 2k - 2 \log L_{\max}. \tag{10}$$

To confirm these findings, the likelihood ratio test statistic can be estimated for pairs of nested models with maximized likelihoods $L_{\max 1}$ and $L_{\max 2}$ and number of independent parameters k_1 and k_2, respectively, as follows

$$2(\log L_{\max 2} - \log L_{\max 1}). \tag{11}$$

This statistic has asymptotically a χ^2 distribution with $(k_2 - k_1)$ degrees of freedom (49). Thus, the appropriate P-value can be obtained and either Model 1 accepted or rejected accordingly.

We obtain the maximum value of the likelihood by evaluating the likelihood over a broad parameter space described in Table 1. Maximization of the likelihood L in Equation (9) is essentially the maximization of $L^{[i]}$, Equation (8), of each data set from an individual.

Evaluation of the likelihood

Each individual has a unique age and inherited allele length which means that the model is fitted over a different length of time for each individual. Consequently, certain parameter combinations are less viable than others, particularly concerning n_0. It is computationally very expensive to evaluate the full likelihood equation (9) for reasons to do with the stiffness of the ordinary differential equation problem. We therefore propose a pragmatic approach, namely to approximate the likelihood function in order to explore the full parameter space and to narrow down the parameter space on which we calculate the full likelihood, thereby making the problem computationally feasible. Our approximation arises from quasi-likelihood theory (60) where the relationship between the mean and the variance can be used to inform a quasi-likelihood which has the required properties of a full likelihood. Rearranging the derived analytical expressions for mean M and variance V, Equations (6) and (7), respectively, give an expression for variance in terms of the mean adjusted for the threshold, a, denoted by

$$\hat{M} = M - a \tag{12}$$

$$V = \left(\frac{\lambda + \mu}{\lambda - \mu}\right)\left(\frac{\hat{M}^2}{n_0} - \hat{M}\right). \tag{13}$$

The equation for the variance is now a quadratic in M and the theory behind quasi-likelihood informs us that the full likelihood can be approximated by a negative binomial distribution with parameters that depend directly on M and V. We therefore approximate the full distribution, $P_n(t)$, by a negative binomial distribution with parameters P and r defined in terms of $\hat{M}$ and V:

$$p = 1 - \frac{\hat{M}}{V}, \tag{14}$$

$$r = \frac{\hat{M}^2}{V - \hat{M}}. \tag{15}$$

This approximate likelihood has the advantage of introducing the model parameters via the mean and variance into a likelihood with, by definition, the properties of a likelihood in terms of the error distribution and allows us to utilize all our data when evaluating the parameter space. Simulations with a range of individuals show this to be a good approximation, capturing both the mean and variance of the full distribution. The negative binomial distribution is also recommended for

count data when there is overdispersion which applies in our case as the variance exceeds the mean (61).

Parameter combinations with a quasi-likelihood value, L_q that satisfies the condition

$$\log(L_q) - \log(\max(L_q)) > k, \tag{16}$$

were then subjected to the full likelihood computation. κ was chosen (typically $\kappa = -2$) to obtain computationally reasonable sample sizes.

Bayesian parameter estimation

We use a Bayesian framework for parameter estimation. Bayes' theorem (62) states that the posterior distribution, π, of the parameters $\theta^{[i]}$ given the observed data $d_j^{[i]}$ is

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) = \frac{L\left(d_j^{[i]}|\theta^{[i]}\right)p\left(\theta^{[i]}\right)}{f\left(d_j^{[i]}\right)}, \tag{17}$$

where $L(d_j^{[i]}|\theta^{[i]})$ is the likelihood of the data given the parameter values, $p(\theta^{[i]})$ is the prior distribution of the parameters representing our initial beliefs about the parameter values before observing any data and $f(d_j^{[i]})$ is the normalizing constant that makes the posterior distribution a valid probability function, otherwise interpreted as the model evidence. Within a model, the normalization includes a constant and Equation (17) has the important consequence

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) \propto L\left(d_j^{[i]}|\theta^{[i]}\right)p\left(\theta^{[i]}\right). \tag{18}$$

In the special case of a uniform prior, $p(\theta^{[i]})$ is greater than zero only for a truncated range of $\theta^{[i]}$ (Table 1), and hence, a constant c can be chosen so that the probabilities sum to unity and Equation (18) further simplifies to

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) \propto L\left(d_j^{[i]}|\theta^{[i]}\right). \tag{19}$$

Note that in this case, the posterior mode of the distribution π is equal to the maximum-likelihood estimator of the parameter. Also, the posterior distribution can be said to be data-driven as the likelihood now dominates the posterior.

Hierarchical Bayes

The underlying distribution of two parameters of particular interest, μ and φ, within the population can be inferred using a hierarchical Bayesian approach. We assume that these are gamma distributions, in shape, chosen because the gamma distribution is defined by two hyper-parameters and hence offers flexibility as to the shape of this distribution. We then infer these hyper-parameters, α_μ and β_μ for parameter μ and α_φ and β_φ for parameter φ by a modification to the posterior probability distribution function

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) \propto L\left(d_i^{[i]}|\theta^{[i]}\right)p\left(\theta^{[i]}|\alpha_\mu, \beta_\mu, \alpha_\varphi, \beta_\varphi\right)p(\alpha_\mu)p(\beta_\mu)p(\alpha_\varphi)p(\beta_\varphi). \tag{20}$$

In effect, we are weighting the likelihood on the strength of the support for the parameters of interest from the underlying gamma distributions.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors would like to thank the members of their respective labs for their comments and support; in particular, they thank Dr Jillian Couto and Dr Douglas Wilcox for insightful discussion. We would also like to thank anonymous referees for their valuable feedback.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by a University of Glasgow Kelvin Smith PhD Scholarship awarded to C.F.H. 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 Tecnologнa 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. This work was further 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. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

REFERENCES

1. 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.
2. McMurray, C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.*, **11**, 786–799.
3. Mirkin, S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.
4. 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.
5. Fu, Y.H., Pizzuti, A., Fenwick, R.G., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T., Jong, P.D. *et al.* (1992) An

unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, **255**, 1256–1258.
6. 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.
7. Harper, P.S. (2001) *Myotonic Dystrophy*. WB Saunders Company, London.
8. 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.
9. 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.
10. Höweler, C.J., Busch, H.F., Geraedts, J.P., Niermeijer, M.F. and Staal, A. (1989) Anticipation in myotonic dystrophy: fact or fiction? *Brain*, **112**, 779–797.
11. 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.
12. 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.
13. Kaplan, S., Itzkovitz, S. and Shapiro, E. (2007) A universal mechanism ties genotype to phenotype in trinucleotide diseases. *PLoS Comput. Biol.*, **3**, e235.
14. Castel, A.L., Cleary, J.D. and Pearson, C.E. (2010) Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell Biol.*, **11**, 165–170.
15. Eckert, K.A. and Hile, S.E. (2009) Every microsatellite is different: intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinog.*, **48**, 379–388.
16. Richards, R.I. and Sutherland, G.R. (1992) Dynamic mutations: a new class of mutations causing human disease. *Cell*, **70**, 709–712.
17. Pearson, C.E., Edamura, K.N. and Cleary, J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.*, **6**, 729–742.
18. 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.
19. 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.
20. 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.
21. 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.
22. Martorell, L., Monckton, D., Gamez, J. and Baiget, M. (2000) Complex patterns of male germline instability and somatic mosaicism in myotonic dystrophy type 1. *Eur. J. Hum. Genet.*, **8**, 423–430.
23. Libby, R.T., Monckton, D.G., Fu, Y., Martinez, R.A., McAbney, J.P., Lau, R., Einum, D.D., Nichol, K., Ware, C.B., Ptacek, L.J., Pearson, C.E. and La Spada, A.R. (2003) Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. *Hum. Mol. Genet.*, **12**, 41–50.
24. 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.
25. Gomes-Pereira, M. and Monckton, D.G. (2004) Chemically induced increases and decreases in the rate of expansion of a CAG*CTG triplet repeat. *Nucleic Acids Res.*, **32**, 2865–2872.
26. Monckton, D.G., Cayuela, M.L., Gould, F.K., Brock, G.J., 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.
27. Zhang, Y., Monckton, D.G., Siciliano, M.J., Connor, T.H. and Meistrich, M.L. (2002) Age and insertion site dependence of repeat number instability of a human DM1 transgene in individual mouse sperm. *Hum. Mol. Genet.*, **11**, 791–798.
28. Kennedy, L., Evans, E., Chen, C., Craven, L., Detloff, P.J., Ennis, M. and Shelbourne, P.F. (2003) Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum. Mol. Genet.*, **12**, 3359–3367.
29. Watase, K., Venken, K.J.T., Sun, Y., Orr, H.T. and Zoghbi, H.Y. (2003) Regional differences of somatic CAG repeat instability do not account for selective neuronal vulnerability in a knock-in mouse model of SCA1. *Hum. Mol. Genet.*, **12**, 2789–2795.
30. Renshaw, E. (1991) *Modelling Biological Populations in Space and Time*. Cambridge University Press, Cambridge, UK.
31. Novozhilov, A.S., Karev, G.P. and Koonin, E.V. (2006) Biological applications of the theory of birth-and-death processes. *Brief. Bioinform.*, **7**, 70–85.
32. Calabrese, P. and Sainudiin, R. (2005) Models of Microsatellite Evolution. In Nielsen, R. (ed.), *Statistical Methods in Molecular Evolution*. Springer, New York, pp. 289–305.
33. Leeflang, E.P., Tavaré, S., Marjoram, P., Neal, C.O.S., Srinidhi, J., MacFarlane, H., MacDonald, M.E., Gusella, J.F., de Young, M., Wexler, N.S. and Arnheim, N. (1999) Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. *Hum. Mol. Genet.*, **8**, 173–183.
34. Veytsman, B. and Akhmadeyeva, L. (2006) Simple mathematical model of pathologic microsatellite expansions: when self-reparation does not work. *J. Theor. Biol.*, **242**, 401–408.
35. 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.
36. 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.
37. Catlin, S.N., Busque, L., Gale, R.E., Guttorp, P. and Abkowitz, J.L. (2011) The replication rate of human hematopoietic stem cells *in vivo*. *Blood*, **117**, 4460–4466.
38. Abkowitz, J.L., Catlin, S.N., McCallie, M.T. and Guttorp, P. (2002) Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. *Blood*, **100**, 2665–2667.
39. 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.
40. Martorell, L. (1998) Progression of somatic CTG repeat length heterogeneity in the blood cells of myotonic dystrophy patients. *Hum. Mol. Genet.*, **7**, 307–312.
41. Wong, L.C. and Ashizawa, T. (1997) Instability of the (CTG)n repeat in congenital myotonic dystrophy. *Am. J. Hum. Genet.*, **61**, 1445–1448.
42. Martorell, L. (1997) Somatic instability of the myotonic dystrophy (CTG)n repeat during human fetal development. *Hum. Mol. Genet.*, **6**, 877–880.
43. Xu, X., Peng, M., Fang, Z. and Xu, X. (2000) The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.*, **24**, 396–399.
44. Weber, J.L. and Wong, C. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.*, **2**, 1123–1128.
45. 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.
46. Wheeler, V.C., Persichetti, F., McNeil, S.M., Mysore, J.S., Mysore, S.S., MacDonald, M.E., Myers, R.H., Gusella, J.F., Wexler, N.S. and Group, T.U.C.R. (2007) Factors associated with HD CAG repeat instability in Huntington disease. *J. Med. Genet.*, **44**, 695–701.
47. Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **19**, 716–723.

48. Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.
49. Cox, D. and Hinkley, D. (1994) *Theoretical Statistics*. Chapman & Hall, London.
50. 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.
51. Seznec, H., Lia-Baldini, A., Duros, C., Fouquet, C., Lacroix, C., Hofmann-Radvanyi, H., Junien, C. and Gourdon, G. (2000) Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. *Hum. Mol. Genet.*, **9**, 1185–1194.
52. Kunkel, T. (1999) The high cost of living. *Trends Genet.*, **15**, 93–94.
53. Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.
54. 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.
55. van den Broek, W.J.A.A., Nelen, M.R., Wansink, D.G., Coerwinkel, M.M., te Riele, H., Groenen, P.J.T.A. 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.
56. Ashizawa, T., Dubel, J.R. and Harati, Y. (1993) Somatic instability of CTG repeat in myotonic dystrophy. *Neurology*, **43**, 2674–2678.
57. 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.
58. 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.
59. Monckton, D.G., Coolbaugh, M.I., Ashizawa, K.T., Siciliano, M.J. and Caskey, C.T. (1997) Hypermutable myotonic dystrophy CTG repeats in transgenic mice. *Nat. Genet.*, **15**, 193–196.
60. Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**, 439–447.
61. Hoef, J.M.V. and Boveng, P.L. (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, **88**, 2766–2772.
62. Sivia, D.S. (2006) *Data Analysis: A Bayesian Tutorial*. Oxford University Press, USA.