University of Glasgow | College of Medical, Veterinary & Life Sciences

Dynamic DNA and human disease: mathematical modelling and statistical inference for myotonic dystrophy type 1 and Huntington disease

by

Catherine F. Higham

Submitted for the
Degree of
Doctor of Philosophy
University of Glasgow

Date **April 2013**

Abstract

Several human genetic diseases, including myotonic dystrophy type 1 (DM1) and Huntington disease (HD), are associated with inheriting an abnormally large unstable DNA simple sequence tandem repeat. These sequences mutate, by changing the number of repeats, many times during the lifetime of those affected, with a bias towards expansion. High repeat numbers are associated with early onset and disease severity. The presence of somatic instability compromises attempts to measure intergenerational repeat dynamics and infer genotype-phenotype relationships. 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.

Dr Fernando Morales, Dr Anneli Cooper and others from the Monckton lab have characterised more than 25,000 *de novo* somatic mutations from a large cohort of DM1 patients using single-molecule polymerase chain reaction (SM-PCR). This rich dataset enables us to fully quantify levels of somatic instability across a representative DM1 population for the first time. We establish the relationship between inherited or progenitor allele length, age at sampling and levels of somatic instability using linear regression analysis. We show that the estimated progenitor allele length genotype is significantly better than modal repeat length (the current clinical standard) at predicting age of onset and this novel genotype is the major modifier of the age of onset phenotype. Further we show that somatic variation (adjusted for estimated progenitor allele length and age at sampling) is also a modifier of the age of onset phenotype. Several families form the large cohort, and we find that the level of somatic instability is highly heritable, implying a role for individual-specific *trans*-acting genetic modifiers.

We develop new mathematical models, the main focus of this thesis, by modifying a previously proposed stochastic birth process to incorporate possible contraction. A Bayesian likelihood approach is used as the basis for inference and parameter estimation. We use 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.

Mutation rates estimated under the models described above are lower than expected among individuals with inherited repeat lengths less than 100 CTGs, suggesting that these rates may be suppressed at the lower end of the disease causing range. We propose that a length-specific effect may be operating within this range and test this hypothesis by introducing such an effect into the model. To calibrate this extended model, we use blood DNA data from DM1 individuals with small alleles (inherited repeat lengths less than 100 CTGs) and buccal DNA from HD individuals who

almost always have inherited repeat lengths less than 100 CAGs. These datasets comprise single DNA molecules sized using SM-PCR. We find statistical support for a general length-specific effect which suppresses mutational rates among the smaller alleles and gives rise to a distinctive pattern in the repeat length distributions. In a novel application of this new model, fitted to a large cohort of DM1 individuals, we also show that this distinctive pattern may help identify individuals whose effective repeat length, with regards to somatic instability, is less than their actual repeat length. A plausible explanation for this distinction is that the expanded repeat tract is compromised by interruptions or other unusual features. For these individuals, we estimate the effective repeat length of their expanded repeat tracts and contribute to the on-going discussion about the effect of interruptions on phenotype.

The interpretation of the levels of somatic instability in many of the affected tissues in the triplet repeat diseases is hindered by complex cell compositions. We extend our model to two cell populations whose repeat lengths have different rates of mutation (fast and slow). Swami *et al.* have recently characterised repeat length distributions in end stage HD brain. Applying our model, we infer for each frontal cortex HD dataset the likely relative weight of these cell populations and their corresponding contribution towards somatic variation. By comparison with data from laser captured single cells we conclude that the neuronal repeat lengths most likely mutate at a higher rate than glial repeat lengths, explaining the characteristic skewed distributions observed in mixed cell tissue from the brain. We confirm that individual-specific mutation rates in neurons are, in addition to the inherited repeat length, a modifier of age of onset. Our results support a model of disease progression where individuals with the same inherited repeat length may reach age of onset, as much as 30 years earlier, because of greater somatic expansions underpinned by higher mutational rates. Therapies aimed at reducing somatic expansions would therefore have considerable benefits with regard to extending the age of onset.

Currently clinical diagnosis of DM1 is based on a measure of repeat length from blood cells, but variance in modal length only accounts for between 20 - 40% of the variance in age of onset and, therefore, is not a an accurate predictive tool. We show that in principle progenitor allele length improves the inverse correlation with age of onset over the traditional model length measure. We make use of second blood samples that are now available from 40 DM1 individuals. We show that inherited repeat length and the mutation rates underlying repeat length instability in blood, inferred from samples at two time points rather than one, are better predictors of age of onset than the traditional modal length measure. Our results are a step towards providing better prognostic information for DM1 individuals and their families. They should also lead to better predictions for drug/therapy response, which is emerging as key to successful clinical trials.

Microsatellites are another type of tandem repeat found in the genome with high levels of intergenerational and somatic mutation. Differences between individuals make microsatellites very useful biomarkers and they have many applications in forensics and medicine. As well as a general application to other expanded repeat diseases, the mathematical models developed here could be used to better understand instability at other mutational hotspots such as microsatellites.

Contents

List of Tables

List of Figures

Acknowledgement

It has been my privilege to apply mathematical techniques to molecular data collected over many years and carefully analysed in the Monckton lab. This work was made possible by the collaborative vision of my supervisors, in particular Prof. Darren Monckton, and generously funded by an interdisciplinary Lord Kelvin Scholarship from the University of Glasgow.

Concerning the provision of high quality data, I would like to thank Dr Fernando Morales and everyone in the Monckton lab, in particular, Dr Anneli Cooper and Dr Jill Couto. It has been great working with you all and I really appreciate your support and friendship. My thanks also go to Prof. Vanessa Wheeler for sharing data related to Huntington disease and offering insightful comments on Chapter 6.

The mathematical models presented here have evolved through interaction and I am grateful to Darren for the many genetic insights which inform the modelling and for giving me free rein with the mathematics. I would like to thank Prof. Dan Haydon and Dr Christina Cobbold for useful comments on Chapter 4 and our co-authored publication.

Meeting families affected by myotonic dystrophy has been inspirational and has provided much motivation for this work. I would like to thank the Myotonic Dystrophy Foundation and the Muscular Dystrophy Campaign for travel funding that has enabled me to present this work to the scientific, clinical and patient-based audiences.

Finally to my lovely family and friends who kept me smiling, I could not have done it without you and I dedicate this work to you.

Author's declaration

The research presented in this thesis is my own original work, except where stated otherwise, and has not been submitted for any other degree.

Catherine F. Higham

Chapter 1

Introduction

Over 20 genetic diseases are associated with inheriting an abnormally large number of simple sequence repeats in genomic DNA. Most of these diseases, but not all, are caused by repeat units of three nucleotide bases: CTG in myotonic dystrophy type 1, CAG in Huntington disease, and CGG in fragile X syndrome. Collectively these diseases are known as trinucleotide repeat diseases and repeats with the motif CAG·CTG comprise the largest class of repetitive elements (Gomes-Pereira & Monckton 2006). Some expanded repeat diseases are based on four or more bases. For example, the repeat unit involved in myotonic dystrophy type 2 contains four bases (CCTG) and in spinocerebellar ataxia type 10 is a repeat unit containing five bases (ATTCT) (Castel et al. 2010). Generally, the longer the inherited repeat length, the earlier symptoms appear (Gomes-Pereira & Monckton 2006).

Myotonic dystrophy type 1 (DM1) and Huntington disease (HD) are among the most common of the trinucleotide diseases. Based on clinical observations, DM1 has an incidence of around 1 in 8,000 among Europeans (Harper 1989) but is less common in some populations including African Americans and Japanese (Ashizawa & Epstein 1991). Incidences in a few other populations are much higher. In the Saguenay region of Quebec the incidence of DM1 is 1 in 500 (Mathieu et al. 1990). This is possibly due to founder effects arising through European migration (Yotova et al. 2005). The prevalence of HD is $\approx$ 1 in 10,000 people in the Americas, Europe, and Australasia (Bates et al. 2002). The highest prevalence of HD in the world is near Lake Maracaibo in Venezuela where it affects around 700 per 100,000 of the population (Wexler et al. 1987).

Inherited unstable DNA mutates by changing the number of repeats during the lifetime of the patient (Gomes-Pereira & Monckton 2006, Mirkin 2007, McMurray 2010). This happens in both the germline and soma, leading to repeat length gains between generations and variation between cells and within tissues. There are no cures for DM1 or HD although molecular therapy is currently making advances in this area (Mulders et al. 2010). The goal of treatment therefore is to help patients maintain their quality of life by preventing or reducing the severity of their symptoms. Currently, patients concerned about their own prognosis and their reproductive choices have limited information available to them about how their disease will progress. Thus, there is great potential for more sophisticated modelling and inference techniques to improve the prognostic value of genetic information. Understanding how these diseases progress in different individuals is also of critical importance in assessing treatment response and the success of clinical trials. The more we understand about the progression of disease the better we can estimate response and establish the efficacy of the treatment.

1.1 Myotonic dystrophy

1.1.1 Clinical observations

Myotonic dystrophy type 1 (DM1) is the most common form of muscular dystrophy in adults. DM1 is a multi-systemic disorder characterised by the presence of myotonia (slow relaxation of the muscles after voluntary contraction or electrical stimulation) followed by progressive weakness and wasting of distal limb and facial muscles, cardiac conduction defects, cataracts, frontal balding and testicular atrophy (Harper 1989). The way DM1 affects patients is very variable. Different sets of symptoms are observed in different patients, sometimes even within the same family. The observable characteristics of patients (or phenotype) fall into four broad clinical forms:

- mild or late onset disease: the only striking symptoms are cataracts beyond the age of 40;
- classic adult onset: patients present most symptoms by their 20s or 30s;
- juvenile onset: presence of delayed motor and growth development, myotonia and sometimes mild mental retardation occurs before 10 years;
- congenital: most severe, symptoms (breathing difficulties and poor muscle tone) are clearly present at birth. There is a high rate of neonatal mortality with babies often dying within the first few days of life.

1.1.2 DM1 mutation

DM1 is transmitted in an autosomal (from one of the 22 non-sex chromosomes) dominant fashion, meaning that the allele associated with DM1 could come from either parent and one such allele is sufficient for the child to be affected. A typical family will discover that it is affected by DM1 with the birth of a child with the congenital form the disease. Sometimes the parents of the child will have symptoms which are unrecognised, but more commonly the mother will have the classic form of the disease with a grandparent, usually the grandfather, having the mild form with just cataracts (Harper 1989).

The mutation responsible for DM1 is an expansion of an unstable CTG trinucleotide repeat located in the $3'$ untranslated region of a gene encoding a serine-threonine protein kinase, named *dystrophia myotonica* protein kinase gene (*DMPK*) and in the promoter region of the *sine oculis* homeobox homologue 5 gene (*SIX5*) located in chromosome 19q13.3 (Aslanidis et al. 1992, Brook et al. 1992, Buxton et al. 1992, Fu et al. 1992, Harley et al. 1992, Mahadevan et al. 1992). A schematic representation of the genomic organisation at the DM1 locus is shown in Figure 1.1. The CTG repeat is polymorphic (existing in many forms) in the general population, ranging from 5 to 37 repeats in healthy individuals, and from upwards of 50 to several thousand in affected DM1 patients. The gene is expressed in smooth, skeletal, and heart muscles, in brain and testis (Groenen & Wieringa 1998, Ueda et al. 2000).

The length of the inherited repeat tract correlates positively with the severity of the disease and negatively with age of onset (Ashizawa et al. 1992, Tsilfidis et al. 1992, Harley et al. 1993, Lavedan et al. 1993). Late onset cases with mild symptoms present the shortest number of repeats (usually between 50 - 150 repeats). Congenital cases with symptoms from birth show the largest number of repeats (usually more than 1,000 repeats). The repeat has been shown to be unstable in both the germline and in the soma (Harris et al. 1996).

The sex of the transmitting parent is important in determining the size of the expansion in the offspring. Congenital cases almost exclusively have an affected mother with the classic adult form of the disease (Harley et al. 1993, Lavedan et al. 1993, Redman et al. 1993, Cobo et al. 1995). However, males carrying small expansions (below 100 repeats), and associated with the late-onset form or asymptomatic, are more likely to transmit an allele associated with the adult-onset form, resulting in an excess of male carriers of small alleles present in the first generation of DM1 families (Brunner et al. 1993, Harley et al. 1993, Lavedan et al. 1993, López de Munain et al. 1995).

Although the repeat usually expands (or has the effect of expanding over time), there have been reported cases with apparent intergenerational contractions of the repeat (Ashizawa et al. 1994).

DM1-like disorders without the DM1 mutation have been recognised and described, in particular myotonic dystrophy type 2 (DM2), which is an expansion of an unstable CCTG repeat in intron 1 of the transcription factor cellular retroviral nucleic acid-binding protein 1 (Liquori et al. 2001).

Currently clinical diagnosis is based on a measure of repeat length from blood cells but variance in modal length only accounts for between 20 - 40% of the variance in age of onset (Perini et al. 1999, Marchini et al. 2000, Mladenovic et al. 2006) and, therefore, is not an accurate predictive tool. Correlations with specific symptoms are often worse, or undetectable (Merlevede et al. 2002, Modoni et al. 2004, Gharehbaghi-Schneli et al. 2008). Hence the International Myotonic Dystrophy Consortium have recommended that patients are not offered prognostic information based on the current test (Gonzalez et al. 2000).

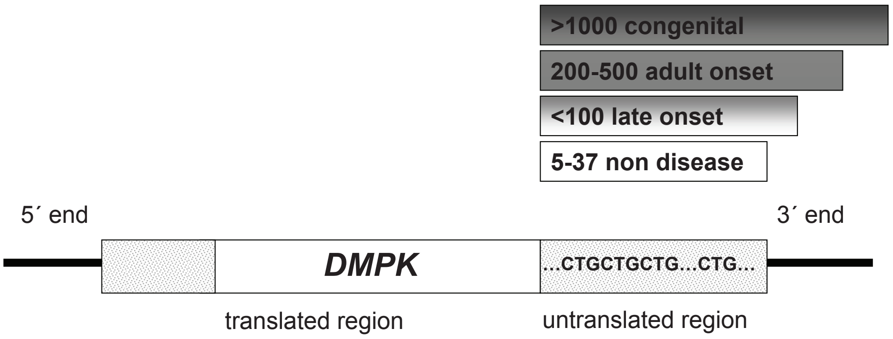


Figure 1.1: **Schematic representation of the genomic organisation at the DM1 locus.** The diagram shows the location of the CTG unstable repeat in the 3′ untranslated region of the *DMPK* gene. The non disease range and the pathological ranges of repeat lengths and associated phenotypes are shown in the boxes above.

1.1.3 Anticipation

Anticipation is defined as the occurrence of a genetic disorder at progressively earlier ages in successive generations (Harper et al. 1992). Myotonic dystrophy type 1 has been associated with the concept of anticipation since it was first described and recognised as a specific disorder (Greenfield 1911, Fleischer 1918). Until the 1980s there was controversy as to whether the phenomenon of anticipation resulted from observational and ascertainment biases (Penrose 1948) or reflected a more fundamental mechanism. Julia Bell made the first quantitative analysis of the genetic aspects of myotonic dystrophy when she assembled pedigree and clinical data on all families reported up to that time. Bell noted the extreme variability in the clinical features of myotonic dystrophy, age at onset and death, and found evidence of anticipation but, despite advances in genetics, there was not a genetic explanation as to how a gene could change down generations at this time (Bell et al. 1948).

The validation of anticipation followed the study by Höweler who provided clear genetic evidence that refuted Penrose's explanation of bias as being solely responsible (Höweler et al. 1989). He found, as Bell and others had done, that families with myotonic dystrophy showed clear intergenerational differences with both anticipation and a close correlation between disease severity and age at onset. Using segregation analysis he showed that penetrance of the gene was close to complete with 46% of offspring affected making it unlikely that there was an ascertainment bias.

The discovery of inherited unstable DNA sequences for fragile-X mental retardation (Fu et al. 1991) suggested a potential genetic mechanism for anticipation. Sutherland predicted that unstable DNA sequences might be responsible for other examples of variation in genetic disease, including anticipation in myotonic dystrophy, thus providing a spur to search for a similar genetic mechanism for this disorder and others such as HD (Sutherland et al. 1991). Independently, several groups found specific molecular abnormalities in myotonic dystrophy with a variable DNA insert of as much as 5 kb in length (Buxton et al. 1992, Harley et al. 1992, Aslanidis et al. 1992). This finding was followed, shortly after by the identification that the variable DNA insert comprised CTG repeats (Brook et al. 1992, Fu et al. 1992, Mahadevan et al. 1992).

1.2 Huntington disease

Huntington disease (HD) is an inherited neurological disorder characterised by progressive movement, psychiatric and cognitive disturbances. Neurodegenerative changes in the brain of affected individuals follow a typical pattern, with early cellular dysfunction and loss of medium spiny neurons in the striatum, followed by more generalised cell loss across the brain (Graveland et al. 1985). The cause of HD is an inherited unstable expanded CAG repeat located in exon 1 of a large gene on the small arm of chromosome 4 (The Huntington's Disease Collaborative Research Group 1993) that results in the extension of a polyglutamine tract at the N-terminus of the encoded, ubiquitously expressed protein called huntingtin. This lengthened glutamine tract is thought to confer a novel toxic property on huntingtin (Mangiarini et al. 1996) that initiates neuron loss from the striatum, in particular, and also the cortex (Vonsattel et al. 1985).

Affected individuals are seen with repeat lengths over 35 CAGs in HD (The Huntington's Disease Collaborative Research Group 1993). But whereas in DM1 inherited repeat length levels range between 50 and several thousand CTGs, in HD most individuals inherit between 40 and 50 CAGs and adult onset is the norm. There is not a known congenital form of HD.

Whilst the age of disease onset is strongly inversely correlated with the length of the expanded CAG repeat length (Andrew et al. 1993, Duyao et al. 1993, Snell et al. 1993, Stine et al. 1993, Gusella et al. 1996), with repeat length accounting for around 70% of the variability in age of onset, this reduces to less than 50% for the majority of HD patients with repeats less than 60 repeats (Myers et al. 1998, Li et al. 2003). There is evidence for heritability for the portion of age at onset not explained by CAG repeat size, which provides support, along with several studies (*e.g.* Li et al. 2003, Wexler et al. 2004), for genetic modifiers of age of onset. Measurement of biomarkers that contribute to variation in age of onset could be used to identify these genetic modifiers, which are key targets for therapies aimed at slowing or reversing the pathogenic process. These measurements could also be used to assess the relative effect of any therapy in specific individuals.

1.3 Expandable DNA repeats and human disease

Microsatellites are short DNA tandem motifs (1 to 6 base pairs in length) that comprise $\approx 3\%$ of the human genome (Lander et al. 2001). As the number of motifs at these loci is highly variable between individuals, microsatellites make very informative molecular markers with many applications in genetics, forensics and medicine. Current genome-wide association studies of single nucleotide polymorphisms have not fully detected the source of genetic variation associated with complex disease. Tandem repeats, which have been shown to affect a range of biological processes including brain function and behaviour (Fondon et al. 2008), are potential candidates for this "missing heritability" (Hannan 2010). The size of microsatellites (often much greater than 75 base pairs in length) does not make them amenable to current high throughput sequencing methods but as next generation sequencing makes it possible to sequence longer lengths, microsatellites are expected to enjoy renewed focus.

The reason that microsatellites are polymorphic is attributable to length changes which 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 (Eckert & Hile 2009). Recent work by Sun *et al.* is the largest study of new mutations to date comprising over 2,000 germ-line changes in 85,000 Icelanders at nearly 2,500 microsatellites (Sun et al. 2012). Their estimation of the mutation rate is 1.4-2.3 $\times 10^{-8}$ per base pair per generation. They observe that the ratio of paternal to maternal mutation rate is 3.3 and report a doubling in fathers from age 20 to 58. No association with age is seen in mothers. They also observe that longer alleles are more mutagenic than small alleles and tend to decrease in size. Mutation in tetra-nucleotides is mostly stepwise whereas larger gains are seen in di-nucleotides. 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 (Richards & Sutherland 1992). 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.

1.3.1 Mutation analysis by single genome PCR

Measurement of trinucleotide repeat germ-line and somatic mutations has traditionally involved polymerase chain reaction (PCR) analysis. As single genome analysis requires many PCR cycles for the detection of PCR products, there is a possibility that PCR artefacts might result in *in vitro* generated mutations. Hence specific control experiments have been designed to assess the likelihood of such artefacts (*e.g.* Cortopassi & Arnheim 1990, Zhang et al. 2002). To determine whether a single sperm mutant arose from a true germ-line event, and was not an artefact of PCR amplification, single sperm were amplified for 6 PCR cycles, after which half the PCR product was removed and saved. For the other half, the reaction continued without interruption. If a mutant was identified in this half, then the other half was also checked for mutants. Mutants arising from germ-line event should exist in each of the saved molecules. Mutants arising during the first 6 cycles of PCR would give rise to a mixture of molecules, hence allowing true mutants to be distinguished from PCR artefacts. The Arnhein lab, performing this work, reports no evidence for PCR artefacts contributing to misidentification. Another way to check for true germ-line events is to compare size distribution of sperm mutation with mutations in somatic DNA (Leeflang et al. 1995). Again, this approach led to the conclusion that the observed variation in single sperm allele sizes was due to germ-line events. Size fractionation of template DNA prior to PCR confirmed the presence of CAG repeat expansions in the striatum of mice that had inherited the *Hdh*$^{(CAG)150}$ allele (Hunter et al. 2005). This work challenges the theoretical possibility that CAG repeat expansions might occur during PCR. The correspondence of fragment size before and after amplification provides evidence that the expansions exist *in vivo*. In summary, *in vitro* PCR artefacts, discussed above, are reported to be minimal.

1.3.2 Somatic instability

DM1 and HD repeat lengths continue to evolve during the lifetime of individuals, 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 (Monckton et al. 1995, Swami et al. 2009). An increase in the number of repeats throughout the lifetime of an individual contributes toward the progressive nature of the symptoms (Morales et al. 2012) and similarly for HD (Swami et al. 2009).

Repeat length variation was first observed as a smear rather than a discrete band on a gel using polymerase chain reaction (PCR) analysis, a biochemical technique in molecular biology to amplify DNA fragments which are then loaded on to a gel dispersing DNA fragments by length (Brook et al. 1992). These results were interpreted as cells within a tissue having different repeat lengths. Later, Monckton *et al.* resolved the smear into individual alleles with heterogeneous repeat sizes by using small molecule or small pool PCR techniques (Monckton et al. 1995).

For a DM1 individual, repeat length is larger in muscle DNA than in blood DNA (Anvret et al. 1993, Ashizawa et al. 1993, Thornton et al. 1994, Monckton et al. 1995, Zatz et al. 1995). Typically, repeat length distributions for the mutant allele in DM1 blood DNA are positively skewed with a relatively sharp lower boundary below which smaller alleles are relatively rare. This lower boundary is conserved between tissues and provides an estimate for the inherited or progenitor allele length (Monckton et al. 1995).

The association of longer repeats with more severe disease and disease related tissues informs the hypothesis that the expansion-biased, age-dependent and tissue-specific nature of somatic instability contributes towards both the tissue specificity and the progressive nature of the symptoms. Up until now, there are no direct data to support this hypothesis.

The expanded CTG repeat in blood is unstable throughout the life time of the patient. Levels of somatic mosaicism in blood from DM1 patients correlates significantly with age (Monckton et al. 1995, Wong et al. 1995, Martorell 1998). These effects are also size-dependent, with larger alleles showing the most variation. These studies have also shown that somatic mosaicism in blood DNA from babies with DM1 was minimal despite the large CTG expansions associated with the congenital form of the disease.

In summary, the data discussed suggests that somatic mosaicism in DM1 is expansion-biased, age and size dependent and tissue specific in that different tissues increase at different rates, features which contribute toward the tissue specificity and progressive nature and severity of the symptoms. This suggests that individual differences in levels of somatic instability may explain why individuals inheriting the same repeat length may present symptoms with different degrees of severity.

The expanded HD CAG repeat is also somatically unstable, undergoing progressive length increases over time (Telenius et al. 1994, Kennedy et al. 2003). HD somatic instability is also tissue-specific with high levels found in striatum and cortex (Shelbourne et al. 2007) and occurs in post-mitotic neurons (Gonitel et al. 2008). Somatically expanded HD CAG repeats are transcribed and translated (Aronin et al. 1995, Wheeler et al. 2003, Gonitel et al. 2008). Evidence of somatic expansion in tissues that are the targets of pathogenesis has given rise to a hypothesis that somatic instability may itself contribute to the HD pathogenic process. Experiments in a genetically accurate Huntington disease homologue (*Hdh*) knock-in mouse model (Hdh^{Q111}), in which an early symptomatic, HD CAG length-dependent phenotype was significantly delayed in mice that lacked somatic instability as a result of the deletion of mismatch repair gene *Msh2*, supports this hypothesis (Wheeler et al. 2003).

Despite differences between DM1 and HD with respect to the repeat motif and its position, and hence differences in the tissues affected and disease pathology, the uni-modal shape of sized single molecule repeat length distributions is very similar in blood or buccal DNA (Veitch et al. 2007, Wong et al. 1995). This suggests that there may be similarities in the mechanism underlying mutation in each disease. Differences other than those linked to cell type may have a molecular basis related to flanking GC content which differ in DM1 and HD with a slightly higher percentage of GCs in HD. There is a strong correlation between the relative expandability of these repeats and the flanking GC content (Brock et al. 1999, Nestor & Monckton 2011). Varying degrees of somatic mosaicism have been reported for other related trinucleotide repeat disorders (Gomes-Pereira & Monckton 2006).

1.3.3 What are the mechanisms of repeat expansion?

The precise mechanisms that cause repeat units to become inserted or deleted from the repeat length tract are not known (Gomes-Pereira & Monckton 2006, McMurray 2010, Mirkin 2007). 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 (McMurray 2010). Two basic types of explanation have been proposed for the expansion of simple sequence repeats. The first focus on DNA replication and the second on DNA repair. There has been some debate about whether there is a single mechanism or more than one mechanism involved. It is important to note that support for the various trinucleotide repeat models has arisen from different systems and different cell types whose properties are unlikely to be the same.

DNA polymerase strand slippage has been proposed as the mechanism for instability in simple sequence repeats (Richards & Sutherland 1994). During replication, the repeats misalign, resulting in a DNA loop that if not properly repaired is either incorporated into the nascent strand leading to expansion or skipped leading to contraction of the DNA strand. However, investigation into the dynamic nature of triplet repeat sequences in mouse models, believed to provide an accurate model of somatic instability observed in man, reveals a lack of obvious correlation between levels of instability and the rates of cell turnover, with high levels of instability observed in post-mitotic tissues such as brain and muscle (Fortune et al. 2000, Seznec et al. 2000, Kennedy & Shelbourne 2000, Lia et al. 1998). Further, data from a DM1 mouse tissue culture model exhibiting expansion-biased-age-dependent somatic mosaicism found no correlation between cell proliferation rate and instability (Gomes-Pereira et al. 2001). As conceptually appealing as the simple slippage model is, these results suggest that the expansion mechanism cannot be entirely dependent on DNA replication.

A cell division-independent DNA mismatch repair (MMR) mediated mechanism has been proposed as an alternative explanation for somatic trinucleotide repeat expansion and deletion (Gomes-Pereira et al. 2004). Several components of MMR are required to generate expansions (van den Broek et al. 2002, Manley et al. 1999, Savouret et al. 2003, Kovtun & McMurray 2001, Gomes-Pereira et al. 2004) hence implicating inappropriate DNA MMR. This view hypothesises that inappropriate DNA MMR is triggered by a slipped-stranded DNA (S-DNA) structure with complementary loop-outs of 1-3 repeat units which may form when expanded repeat DNA re-anneals out of register, see Figure 1.2. These alternative DNA conformations form readily *in vitro* and are very stable (Pearson & Sinden 1996) but have not yet been reported *in vivo*. MMR proteins are then recruited and bound to each loop-out independently. The MMR machinery either incorporates the loop-out by creating a gap and filling it on the opposite strand or simply removes the loop-out. The size of a potential gap is not known but experiments using human nuclear extracts suggests that the MMR machinery may remove between 60-230 base pairs of DNA (Genschel & Modrich 2003). Incorporation of the loop-out would result in a small increase in the number of repeat units and removal of the loop-out would result in a small decrease in the number of repeat units. How the decision to incorporate or remove loop-outs is made is an unanswered question. As loop-outs re-form, this process is re-initiated. A bias towards incorporating the loops, no matter how subtle, will lead over time, through the accumulation of many small repeats, to expansion gains.

1.3.4 How do repeat expansions result in disease?

The mechanisms underlying pathology depend on where the repeat is found within the gene. The CTG repeat unit in DM1 is found in the non-coding untranslated region at the $3'$ end of the *DMPK* gene (Buxton et al. 1992, Fu et al. 1992, Brook et al. 1992). In Huntington disease the repeat unit is CAG in the coding region of the huntingtin gene (The Huntington's Disease Collaborative Research Group 1993). Repeats found in non-coding untranslated regions (*e.g.* DM1) are thought to give rise to a toxic RNA gain of function whereas repeats found in coding regions (*e.g.* HD) are often transcribed and translated, creating expanded polyglutamines and a related toxic gain of function (Castel et al. 2010).

RNA-mediated muscle disease

DM1 is hailed as the first example of an RNA-mediated disease (Wheeler & Thornton 2007). This is based on evidence that it is the RNA rather than the protein product of a disease gene that has the deleterious effect on muscle. DM1 is not explained by reduced expression of DMPK protein (Jansen et al. 1996). The RNA containing the expanded repeat forms nuclear foci in muscle cells (Taneja et al. 1995) and expression of mutant *DMPK* RNA leads to abnormal regulation of alternative splicing (Philips et al. 1998). RNA splicing is the process by which introns are removed from the RNA transcript and exons are joined together to make mRNA and is critical for regulation of gene expression. Often there are multiple introns and exons and regulated splicing decisions can yield a spectrum of alternative products for different tissues or at different stages of development. The outcome of alternative splicing is controlled by splicing regulatory proteins. One group of

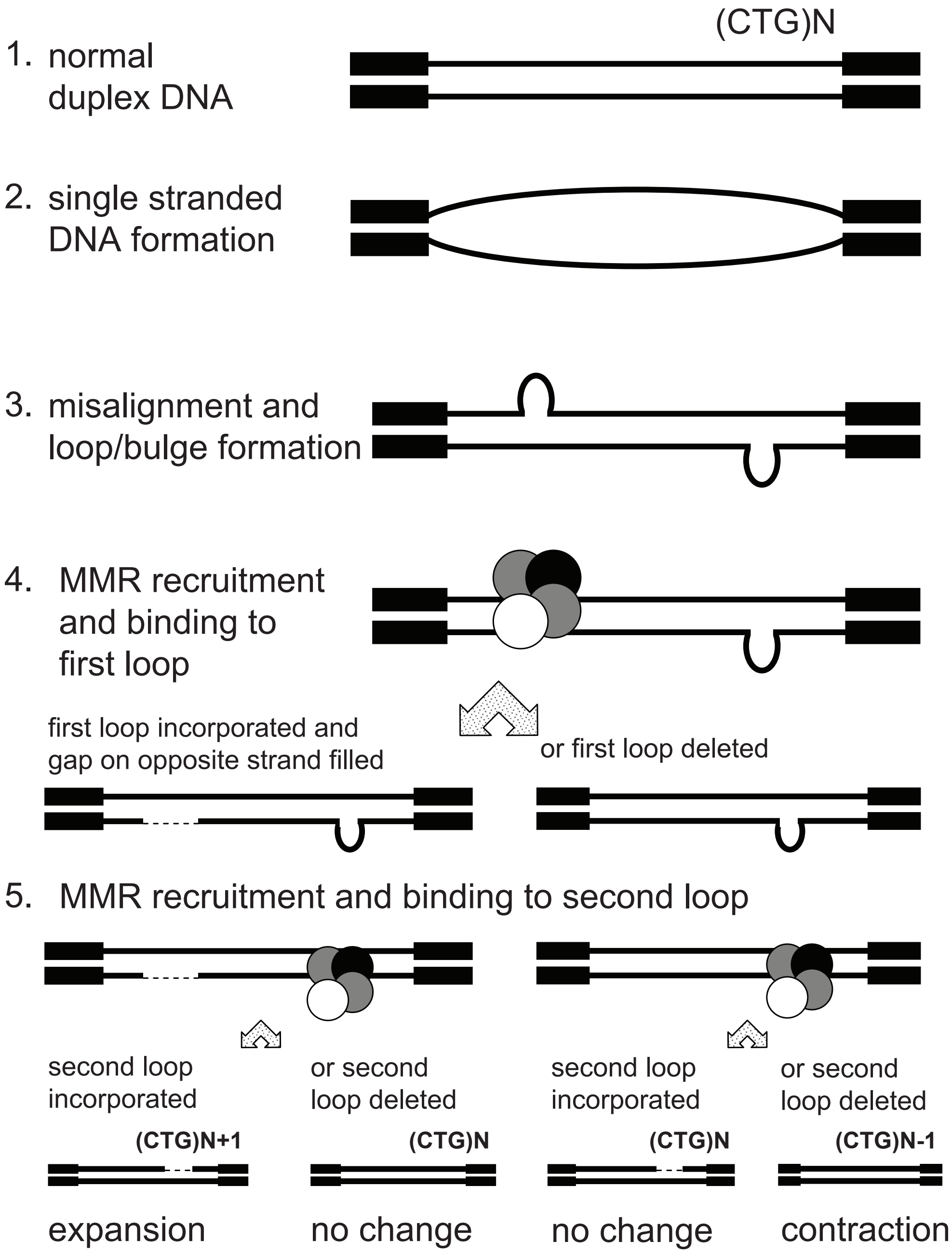

(CTG)N
1. normal duplex DNA
2. single stranded DNA formation
3. misalignment and loop/bulge formation
4. MMR recruitment and binding to first loop
first loop incorporated and gap on opposite strand filled
or first loop deleted
5. MMR recruitment and binding to second loop
second loop incorporated
or second loop deleted
second loop incorporated
or second loop deleted
(CTG)N+1
(CTG)N
(CTG)N
(CTG)N-1
expansion
no change
no change
contraction

Figure 1.2: **Hypothetical mechansism of repeat expansion based on inappropriate DNA mismatch repair (MMR).** Inappropriate MMR is triggered when expanded repeat DNA (1) re-anneals out of register (2) forming a slipped-stranded DNA (S-DNA) structure with complementary loop-outs of 1-3 repeat units (3). MMR proteins are then recruited and bound to each loop-out independently (4,5). If both events result in either the loop-out being incorporated or deleted then the net result is expansion or contraction. If the events are different then there is no change. As loop-outs re-form, this process is re-initiated. A bias towards incorporating the loop-outs, no matter how subtle, will lead over time, through the accumulation of many small repeats, to expansion gains.

RNA binding proteins implicated in myotonic dystrophy pathogenesis are splicing factors in the muscleblind-like (MBNL) family. MBNL proteins bind to *CUG* RNA *in vitro* with high affinity and are found in RNA inclusions in DM1 muscle nuclei. Considerable evidence supports the theory that sequestration of MBNL proteins is a critical step in the pathogenesis of myotonic dystrophy (Ranum & Cooper 2006). One effect of this disease process is to alter the function of alternative splicing factors and thereby perturb the regulation of RNA processing for other genes.

1.3.5 Cure and treatment

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 (Morales et al. 2012). As such, suppression of somatic expansion is expected to be therapeutically beneficial and induction of contractions potentially curative (Gomes-Pereira & Monckton 2006, Castel et al. 2010). Small molecules that may reduce somatic expansion have been identified (Gomes-Pereira & Monckton 2006) and novel technologies (Olsen et al. 2009, Mittelman et al. 2009, Aarts et al. 2009) may prove beneficial in the future. However, the feasibility of suppressing expansions/inducing contractions remains largely undetermined.

Further along the pathology cascade, defects in RNA alternative splicing are potentially reversible so there is a new focus on therapies targeted directly at reversing RNA toxicity, which are showing promise in preclinical models by correcting spliceopathy and eliminating myotonia (Wheeler 2008, Mulders et al. 2010). In particular the use of antisense oligonucleotides which target toxic RNA is a proof-of-principle therapy very effective in cell culture and mice (Lee et al. 2012). This therapeutic approach looks very promising and two major drug companies, Biogen Idec and Isis Pharmaceuticals, have recently entered a highly funded (over 50 million US Dollars) collaboration to develop and commercialise a novel antisense drug for the treatment of DM1.

Patient stories

Patient stories can be found online at support groups such as `www.myotonic.org` and `www.muscular-dystrophy.org`. These stories highlight the need for better prognostic information and display the efforts made in the hope that a cure will eventually be found.

1.4 Mathematical models

Mathematical models aim to capture and quantify key features of the biological processes of interest in order to give insights into how a system works and how it will respond to change. (These principles are described more fully in (Otto et al. 2007)). A mathematical model is developed by incorporating biological knowledge into precise mathematical language, which can then be analysed in a variety of ways. Most models require parameters, some of these are known or can be measured experimentally, but others will not be available. In the later case, modern statistical techniques exist to fit parameters to the data. Sometimes several different models may be proposed, perhaps based on competing biological hypotheses, and there is a need to compare models in terms of best fit. In the mathematical modelling community these issues are active areas of research. The challenge with a specific biological process is therefore to develop a good class of models along with methods for parameter estimation and model selection. We want the mathematical model not just to reproduce the data but to make useful hypotheses about the system that can be tested experimentally.

The biological phenomenon of interest often concerns a component or a system of interacting components and how this system changes over time. A dynamical model which aims to describe how a system changes over time can provide insights into how various forces act to change the component, which in our case is repeat length. There are two broad classifications of dynamical models: deterministic or stochastic. A deterministic model is one where the future is entirely predicted by the model whereas a stochastic model is based on the assumption that random events affect the biological system and so the model can only assign probabilities to possible outcomes. Models which are stochastic at the single cell level can often be well approximated by simpler deterministic models if there are large numbers of cells involved. Stochastic models are generally more challenging than deterministic models in terms of computational demands, analysis and data fitting. But as reductionist genetic and molecular biology produces quality time course data at single-cell resolution, the stochastic approach is needed to underpin such a process and capture complex dynamics (illustrated in (Wilkinson 2009)).

1.4.1 Models of microsatellite evolution

In the non disease case there exist models for microsatellite evolution, which are summarised in (Calabrese & Sainudiin 2005). However mutation at these sites occurs at lower rates and typically involves shorter lengths than in the pathological disease case. Also these models tend to assume that an equilibrium in the distribution of lengths has been reached in the population. In the pathological disease case the data suggests that the distribution of length is time-dependent throughout the life of a patient. This makes the analysis different as we cannot assume equilibrium status. However these models form a useful basis for our work. The earliest model for microsatellite evolution is the stepwise mutation model originally proposed by (Ohta & Kimura 1994). Kruglyak *et al.* proposed a proportional slippage model where the mutation rate increases linearly with microsatellite length (Kruglyak et al. 1998). Although most observed microsatellite mutations are by one repeat unit, not all are, so Di Rienzo *et al.* proposed a model which allows for larger mutations (Rienzo et al. 1994). We refer to (Calabrese & Sainudiin 2005) for further details.

In a different, but related context, mathematical models have been applied to the evolution of the CAG expansion in the huntingtin (HTT) gene in the general population (Falush 2009). The modelling approach (Falush et al. 2001) quantifies the rate of progression of the disease in the population by measuring the mutational flow. The model can be used to describe the repeat length change, either from parents to offspring, or during the mitotic divisions in the germ cells of a single individual, or over time in a population. These models assume stepwise mutations and incorporate an upper bound so that all repeats that reach the boundary are removed and the model is well behaved and results in dynamic equilibria. For HD, they use the upper bound to represent selection against very large repeats. Warby *et al.* looked at the haplotype background of chromosomes carrying the HD mutation and the length distribution of the CAG repeat for different haplotypes within the general population. They concluded that *cis*-elements are likely to represent a major predisposing element in HD expansion. Using evolutionary modelling of the CAG repeat length within populations, Falush *et al.* argue that the distribution of CAG repeat length and disease incidence can be explained by founder events, each of which involved expansion of repeats to lengths that are classified as normal by HD investigators (< 28 repeats). There is no need to invoke *cis*-elements as having a role in the evolution of HD chromosomes. Whilst the work by Falush *et al.* provides insights into the evolution of CAG expansion in HD, there are assumptions in the computational model which weaken their hypothesis. First, the assumption that negative selection acts strongly against chromosomes with 50 CAG repeats is unrealistic as many individuals with repeat lengths > 50 CAGs typically do not become symptomatic until their thirties. Second, the mutation rates are based on sperm typing data from 26 men in a Venezuelan HD cohort with CAG sizes ranging from 37 to 62 repeats (Leeflang et al. 1999). There are no data to validate these mutation rates for chromosomes in the intermediate allele range (27-35 CAG) or the normal range (< 27). Third, CAG dependency and upward bias of the mutation may have lower cut-off threshold dictated by Okazaki fragment length, DNA damage susceptibility, repair excision tract size and *cis*-elements (Cleary et al. 2002, Pollard et al. 2004).

These models describing the evolution of microsatellites relate to differences between generations, and to short and slowly changing repeats rather than to long and rapidly changing repeats. Pathological mutations associated with rapidly changing repeats arising during the lifetime of individuals have also been studied using a mathematical modelling framework. Leeflang *et al.* 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 (Leeflang et al. 1999). Falush *et al.* show that a simple length-dependent stepwise mitotic model can account for repeat length distribution observed in individual sperm samples, the mutation rate variation between samples with different somatic repeat lengths and the overall pattern of mutation observed in disease-chromosome transmissions (Falush et al. 2001). However they also reported discrepancies with the empirical data: underestimation of the mutation rate for female transmission; inter-individual variation; larger than stepwise changes occurring at a significant rate in sperm; and underestimation of the mutational bias in sperm samples from individuals with somatic repeat lengths > 49 CAGs. More recently Veytsman and Akhmadeyeva showed that a simple theoretical model of pathological microsatellite expansion based on hairpin formation could offer an explanation for the observed phenomena of somatic mosaicism, anticipation and rare reversions (Veytsman & Akhmadeyeva 2006). Although these models do not incorporate recent insights recognising the involvement of activities other than replication, such as repair and transcription (Castel et al. 2010), or are not based on *in vivo* data, they also form a useful reference for our work.

1.5 Statistical inference

Mathematical models have biological parameters, some of which can be measured experimentally and some of which must be inferred indirectly. Parameter estimation (recovering unknown parameters from experimental data) and model selection (rating competing models that are attempting to describe the biological processes) are important steps towards obtaining an explanatory model that can be used for simulation and prediction. Bayesian inference is being used increasingly in genetics (Beaumont & Rannala 2004) as it provides a solid foundation for parameter estimation and model selection. Model selection based on information theory is a relatively new paradigm in the biological and statistical sciences and is quite different from the usual methods based on null hypothesis testing. Model selection based on information theory is not only an intuitively attractive approach but also has philosophical and computational advantages (Burnham & Anderson 2002). There is currently much interest (Wilkinson 2011) in using statistical methods to estimate parameters of detailed mechanistic (bottom-up) biological models using quantitative time course data on the system.

1.6 Project design and aims

Currently, individuals finding out that they or their family are affected by DM1, 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 (Mladenovic et al. 2006, Perini et al. 1999, Marchini et al. 2000). Low correlation between 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.

Advances in technologies such as DNA sequencing are generating vast data sets which offer exciting opportunities for the development of quantitative methods to understand biological phenomena. Ongoing studies (Morales et al. 2012) are measuring somatic mosaicism in many DM1 families and sequencing the affected region of DNA. These recent quantitative data sets make it feasible to develop a mathematical model which aims to explore the underlying mechanism of mutation and identify the key drivers and, most importantly, have predictive power. Increasingly there is a need to combine numerical techniques with biological understanding to get the most out of the data. We seek to bridge this divide by deriving new mathematical models, using a range of deterministic and stochastic modelling techniques, for the genetic phenomenon of hypermutational DNA dynamics. This work, as well as improving prognostic information for patients, could have an important role in the design and interpretation of clinical trials. For example, by accounting for variation between patients, we should be able to exclude outliers and thus narrow the estimates for drug response.

By increasing our understanding of the mechanism underlying unstable repeats, we also expect our models to have general application to unstable microsatellites and other trinucleotide diseases.

Our extensive data arises from elaborate small pool PCR analysis of repeat length in blood cells from a cohort of 145 individuals with DM1 expansions (Morales et al. 2012). 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 (Monckton et al. 1995), 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 (Fortune et al. 2000, Martorell et al. 2000, Libby et al. 2003, Gomes-Pereira et al. 2004, Gomes-Pereira & Monckton 2004, Monckton et al. 1999, Zhang et al. 2002, Kennedy et al. 2003, Watase et al. 2003). For each individual, Morales *et al.* have used single molecule analysis to size the expanded CTG repeat tract in between 100 and 350 cells (see Figure 2.1), providing a total data set of over 25,000 observations (Morales et al. 2012). These data reveal the variation in repeat length between cells and individuals. The shapes of the distributions of repeat lengths are seen to depend on both age and typical length. 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.

The overall challenge 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 dataset. Unlike other applications where only one population may be observed over time, by sampling many cells from individuals we have many realisations of the same stochastic process at one point in time. Hence, our data provides 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.

Our model builds on Kaplan *et al.* who used a simple birth process 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 (Kaplan et al. 2007). They were able to demonstrate that somatic mosaicism contributes to disease onset and progression. However their model is concerned with only expansions. Contractions have been seen in cell models *in vitro* (Gorbunova et al. 2003, Gomes-Pereira & Monckton 2004) and mouse tissue *in vivo* (Gomes-

Pereira et al. 2004) providing a basis for the assumption that contractions occur in somatic human cells *in vivo*. There is also evidence for contractions between generations arising in the germline (Ashizawa et al. 1994, Monckton et al. 1995, Martorell et al. 2004). Thus we statistically test here the possibility that somatic variation is due to the difference between expansion and contraction mutations. We use the same stochastic modelling framework as described in (Kaplan et al. 2007), 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 (Monckton et al. 1995).

1.6.1 Experimental approaches

Our new experiments will predominantly be *in silico* using precise mathematical language and computer power to generate results. In our case the key tools required include probability theory, calculus and statistical inference.

- We use a stochastic approach to model the evolution of repeat length which assigns probabilities to the biological events of expansion and contraction. Some simple stochastic approaches can be formally analysed but others require simulation techniques to realise the model outcomes (for example Gillespie's Algorithm).

- A variety of computational techniques are required for the numerical solution of the underlying equations. This includes state-of-the-art software tools for non-linear equations and ordinary differential equations.

- In order to obtain estimates for the model parameters we will explore Bayesian techniques to calibrate the model against the biological data. This is a very modern research area with little specific existing software, and a large component of our work involves customised design and implementation of computational algorithms for our specific class of models.

- Modelling is an iterative process and the models will be subjected to tests and refinements following biological discussions and validations. This is in accord with the highly interdisciplinary nature of the project.

1.6.2 Summary of project aims

In summary, the principle aim of this work is to develop and test new mathematical models using a range of modelling techniques that capture the key features of the mutation mechanism underlying repeat length evolution. Values for the biological parameters informing the models will be inferred from the data using modern Bayesian statistical methods. A model will be developed, in the first instance, for blood DNA from DM1 affected individuals, and then extended to DNA from other tissues. We will also fit the models to DNA from HD affected individuals. We expect there to be some differences in parameter values between tissues and diseases. The calibrated models will then be employed in a number of ways:

- To investigate the possibility that expansion bias is due to the difference between expansion and contraction mutations, rather than expansion alone.
- To quantify different aspects of repeat instability such as mutation rates between tissues and diseases.
- To predict the progression of the disease in an individual and within families and hence contribute towards the development of a useful prognostic tool.
- To reduce the unaccounted for variability between patients and enable better stratification of the patient cohort in clinical trials.
- To estimate the length of the inherited allele and allow us to revisit pedigree data with a view to shedding light on important issues such as heritability.

Chapter 2

Materials and methods

2.1 Introduction

This chapter covers the construction of the mathematical models and the inference method used for fitting the models to the data. The aim of our models, in the first instance, is to describe the progression over time of the unstable repeat length found at the myotonic dystrophy type 1 locus in blood DNA. The dataset used to calibrate this first model is outlined in Section 2.2. In later chapters, we describe the extension of this first model to tissues other than blood and to Huntington disease (Chapters 5 and 6). In Chapter 7 we adapt the model to handle two DNA samples taken at different time points from one individual.

Before using individual data to infer the parameters of a model, it is informative to establish what can be inferred in the best possible scenario, when the data are generated synthetically from the appropriate model with known parameter values. Even in this idealised case, there will still be some uncertainty in the inference process due to the finiteness of the sample size and the impossibility of searching exhaustively over a high-dimensional real-valued parameter space. Hence this type of computational experiment helps to quantify the inherent uncertainty. In Section 2.7 the inference method is applied to a synthetic dataset to assess how well the method infers parameters.

2.2 Project data

The data analysed in this study comes from DNA blood samples collected from patients with myotonic dystrophy type 1 across four countries: 77 from Costa Rica provided by Dr Fernando Morales, 36 from Texas, USA provided by Prof. Tetsuo Ashizawa, 27 from the western region of Scotland provided by Dr Douglas Wilcox and Dr Alison Wilcox and 5 from Uruguay provided by Dr Claudia Braida. All the samples were purified from peripheral blood leukocytes using phenol-chloroform purification and proteinase K. The patients include affected individuals as well as asymptomatic carriers and clinical information was obtained from their hospitals. The information collected includes age of onset if applicable, age at sampling and a brief description of the main symptoms. Signed informed consent was obtained for everyone in this clinical and molecular investigation as in accordance with the relevant ethical protocols.

2.2.1 Small-pool PCR

Small-pool polymerase chain reaction (PCR) analysis was performed using oligonucleotide primers DM-C and DM-BR as previously described (Monckton et al. 1995) by Dr Fernando Morales, Berit Adams and others from the Monckton lab to estimate the progenitor allele length (the inherited repeat length) using the lower boundary of the total allele length distribution and to quantify the degree of somatic variation in 145 DM1 samples. Restriction digested genomic DNA is diluted and multiple aliquots or small pools are amplified using the PCR and primers flanking the repeat. Products are resolved by agarose gel electrophoresis and detected by Southern blot hybridization with an interval probe. The PCR products are sized using Kodak Digital Science software by lining up and comparing the bands with known molecular weight markers. To assess the variation in the data, the first issue to consider is the DNA concentration required so that individual bands can be distinguished from one another in the small pools, effectively the lanes on the gel. For samples where there is less expected variation (*i.e.* samples with short repeat lengths), the level of concentration needs to be lower (fewer bands per lane) and gels run for longer so that the bands are well dispersed and can be individually identified. Typically several gels are run at increasingly lower dilutions to determine best dilution level. As a further check that all the molecules have been accounted for, the number of molecules amplified in each reaction is expected to follow a Poisson distribution over the number of lanes. Gels that do not conform to these criteria are rejected. Further details can be found in (Morales 2006, Morales et al. 2012) and some typical output is shown in Figure 2.1. The data can be visualised as allele length frequencies in a histogram format (Figure 2.2) and the mathematical models describe these distributions in terms of the biological parameters of interest.

2.2.2 PCR artefacts and interpretation of the data

The original SP-PCR procedure (Jeffreys et al. 1994) was adapted for analysis of CTG variability by (Monckton et al. 1995). Measuring somatic instability at the expanded repeat loci is challenging and requires relatively sophisticated approaches. These methods are still heavily used today by the Monckton lab, University of Glasgow, and have been developed for HD CAG repeats, principally by the Wheeler lab, Center for Human Genetic Research, Harvard Medical School. Even between these two labs, as there are differences in typical repeat sizes for DM and HD, there are consequently differences in the equipment and products used. Although on-going comparisons between labs would provide further reassurance about the quality of the data, small-pool PCR is a well proven method (Jeffreys et al. 1994, Monckton et al. 1995) that provides a robust approach to quantification of length variation in somatic DNA. Currently, emphasis is put on checking the internal consistency of the data, discussed below, and overcoming new challenges such as characterizing interruptions in the expanded repeat (Musova et al. 2009, Braida et al. 2010). The main issue is whether the PCR products are faithful representations of alleles present in single cells. The following observations, from the data, provide positive support: variant length alleles associated with expanded allele reflect variation in an independent Southern blot analysis; the number of bands are directly proportional to quantity of input DNA; and distributions are sample-specific and not merely a reflection of allele length (two samples indistinguishable by SB have different distributions with small-pool PCR).

However PCR and other technical artefacts can confound the interpretation of the data. PCR stutter, the generation of shadow bands by products of the PCR amplification differing in length from the original allele, is a particular issue. 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 (Zhang et al. 2002). In our case, as well as minimising PCR stutter by employing fewer cycles of PCR (28), the underlying variation is typically spread over many hundreds of repeats. PCR artefacts could be included in the model likelihood as in (Leeflang et al. 1996), but 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. Hence finite sampling is of more concern than PCR artefacts. By applying our parameter estimation method to a synthetic dataset where the parameter values are known we can quantify this level of uncertainty and these results are discussed in Section 2.7.

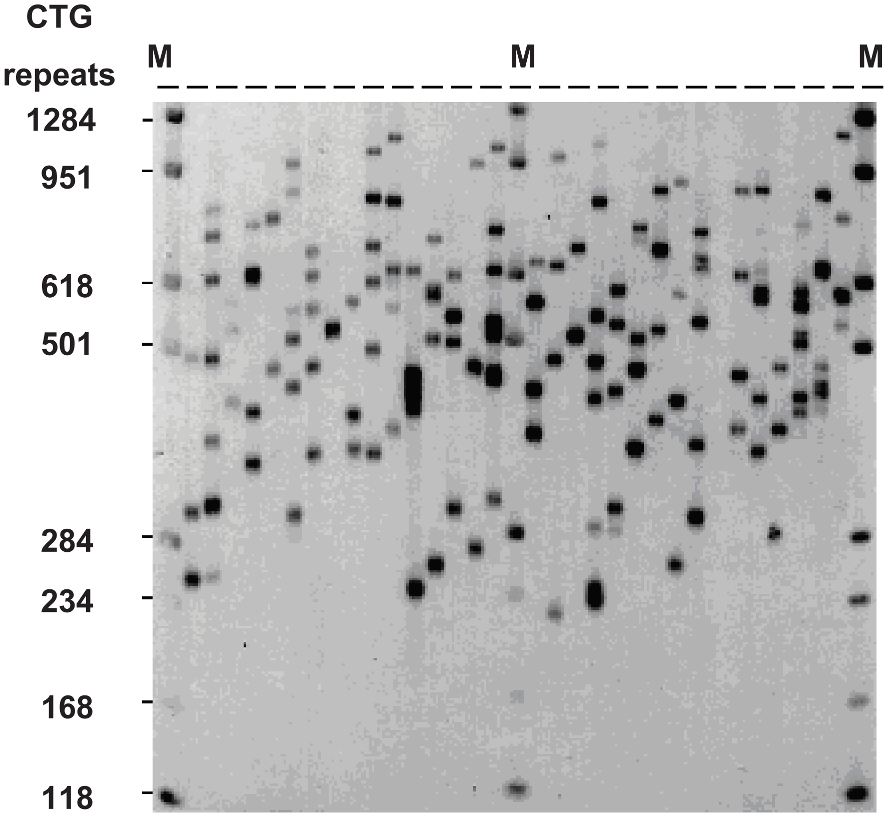


Figure 2.1: **Representative data of single molecule analysis by PCR in a DM1 sample.** The total number of molecules sized in this sample was 141 alleles. Three marker lanes were run (M) with the PCR products and the band sizes of the marker were transformed to the corresponding number of CTG repeats in the scale on the left (Morales 2006).

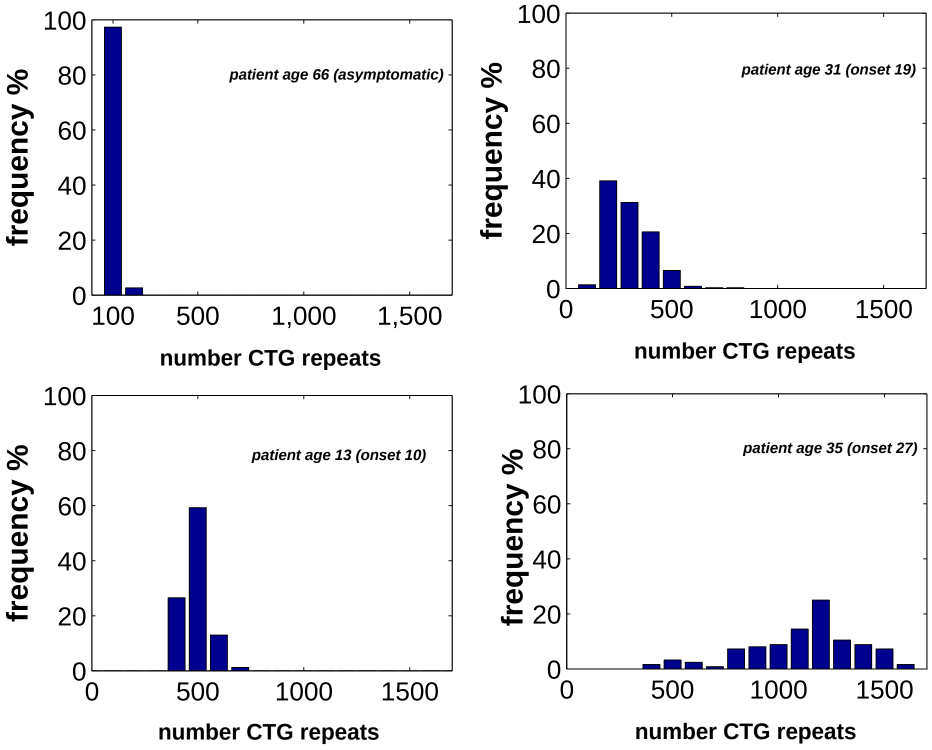


Figure 2.2: **Representative allele length distributions in DM1 patients.** The histograms represent the frequencies of allele lengths (allele sizes were collated into 80 repeat groups). As seen in the histograms, data from top right appears, although skewed to the right, to be more more normal than the data from sample top left. Data from bottom left also appears to be relatively close to a normal distribution. Data from sample bottom right shows a distribution that is highly skewed to the left, suggesting the presence of contracted alleles. The age of sampling of each patient is also shown (Morales 2006).

2.3 Modelling context

We base our model on a stochastic birth and death framework which was traditionally developed to model the growth of a population (Renshaw 1991). Birth and death models are used to count entities over time and are applied to many types of biological processes where the individuals can involve anything from molecules, cells, tissues, organisms, ecosystems or biospheres (Novozhilov et al. 2006). The entity, in our case a CTG repeat length, is treated as a discrete random variable at each continuous point in time with “birth” being the expansion of the repeat length and “death” the contraction of the repeat length. The use of random variables, whose value results from a measurement on some type of random process, means that we are concerned with how likely the events under investigation, expansion and contraction, are and assign these events a probability. We can obtain expressions for the probable CTG repeat length, given its starting length, at a point in time. We can also obtain expressions for the mean repeat length and variance in repeat length for an ensemble of alleles.

The use of a stochastic process is appropriate for our dataset as we can interpret the individual samples as resulting from many independent (see below) runs of the same process. The data that we use in our study effectively provides between 100 and 300 outcomes of an independent stochastic process in the somatic blood cells sampled at a single point in time. In total, 25,000 repeat lengths were sized, representing one of the largest databases of its kind. Of those alleles, over 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.

Our model builds on Kaplan et al. who used a simple birth process to describe repeat length evolution and derived expressions to fit basic clinical data (age at onset and modal repeat length) for a range of diseases associated with expanded repeats (Kaplan et al. 2007). They were able to demonstrate that somatic mosaicism contributes to disease onset and progression. However, because their data was 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. As mentioned above, we have information about the variation and distribution of repeat lengths.

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.

We investigate here the possibility that somatic variation is due to the difference between expansion and contraction mutations. We use the same stochastic modelling framework as Kaplan et al. but extend it to include contractions (death process) and a threshold below which expansion and contraction does not occur (Kaplan et al. 2007). Such a threshold is consistent with the relative stability of the normal allele (Monckton et al. 1995). In the context of this work, we are counting the number of CTG repeats within each cell. The mathematical model quantifies the probability of an increase or decrease in the repeat length per unit time. As circulating white blood cells typically do not replicate, we assume 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 hematopoietic stem cells is estimated between 11,000 and 22,000 (Abkowitz et al. 2002, Catlin et al. 2011). 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 (Catlin et al. 2011). 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 (Wong et al. 1995, Martorell 1998, Martorell et al. 2000), 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 (Wong et al. 1995, Martorell 1997, Wong & Ashizawa 1997, Martorell 1998). On this basis it is reasonable to assume that the stem cells effectively have independent mutational histories. Thus we interpret our samples of between 100-350 cells as a proxy representation of the 11,000 - 22,000 ultra progenitor stem cells with each sample informing us about the underlying process. Hence the stochastic process model is 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 is the number of CTGs inserted or deleted at either mutational event. Studies using microsatellite data (Weber & Wong 1993, Xu et al. 2000) found that the majority of insertions or deletions were of one CTG repeat. Data from HD individuals where the alleles are smaller and there is less variation, and where it is assumed that a similar mechanism underlies DNA instability, provide 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 (Veitch et al. 2007, Wheeler et al. 2007) 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. The same observation is made from data from DM1 individuals with small alleles (less than 100 CTGs) (Morales et al. 2012). So, in our case, it is a reasonable working assumption that the birth and death process treats one CTG repeat as the individual unit and we associate “birth” with expansion and “death” with contraction. In effect we consider whether the observed distributions from individuals could have arisen from the cumulative effect of small incremental gains and losses of one repeat length.

The overall aim of this work is to develop and test 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 realisations of the same stochastic process at one point in time. Hence, our data provides a unique opportunity to access directly the inherent fluctuations that are required to fit a stochastic process. Since a likelihood arises naturally from the stochastic process, both maximum likelihood and Bayesian methods lend themselves to fitting the data to the model. We are able to quantify several important biological parameters relating to the mechanism underlying repeat length evolution. This is an important step towards understanding pathological mutations and providing better prognostic information for individuals with diseases arising from these mutations.

2.4 Mathematical model

The mathematical modelling approach commences by quantifying the probability of an increase or decrease in the repeat length in one cell. 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 increment step size. Then at time $t + \delta t$, where δt is small:

$*$ 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 above, the increment step size s in our model is one CTG unit. However the model could readily be extended to other step sizes by appropriate adjustment to these expressions.

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:

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)(n-a)P_n(t) + \lambda(n-a-1)P_{n-1}(t) + \mu(n-a+1)P_{n+1}(t), \quad (2.1)$$

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 may then derive expressions for repeat length mean, M, and variance, V, from the first and second moments of $P_n(t)$, denoted $M(t)$ and $M_2(t)$, respectively, and defined as

$$\begin{aligned} M(t) &= \sum_{n\geq a} nP_n(t), & (2.2)\\ M_2(t) &= \sum_{n\geq a} n^2 P_n(t). & (2.3) \end{aligned}$$

Differentiating both (2.2) and (2.3) with respect to t and substituting (2.1) into the result leads, after some manipulation, to

$$\begin{aligned} \frac{dM(t)}{dt} &= (\lambda-\mu)(M(t)-a), & (2.4)\\ \frac{dM_2(t)}{dt} &= 2(\lambda-\mu)M_2(t) + [\lambda+\mu-2a(\lambda-\mu)]M(t) - a(\lambda+\mu). & (2.5) \end{aligned}$$

Solving (2.4) and (2.5) with $M(t=0) = n_0$ and $V(t=0) = 0$, where n_0 is inherited repeat length, and setting $V(t) = M_2(t) - (M(t))^2$ for the variance at time t gives the analytical expressions (2.6) and (2.7).

For completeness, we mention that this modelling approach may also be extended to a more general setting that allows a range of possible increments to be incorporated. Here, a general state-dependent function g could be supplied such that, given length n at time t, at time $t+\delta t$:

- $*$ the probability that the length is $i \approx g\left(n,i\right)\delta t$, for $i = 0, 1, 2, \ldots, n-1, n+1, n+2, \ldots$,
- $*$ the probability that the length is $n \approx 1 - \Sigma_{i=0,i\neq n}^{\text{inf}} g\left(n,i\right)\delta t$.

Here, the non-negative function g must be chosen so that $\Sigma_{i=0,i\neq n}^{\text{inf}} g\left(n,i\right)\delta t$ is finite. The form of the function g would, of course, require justification from a biological perspective, and the extra freedom of specifying a range of possible expansion and contraction increments would come at the expense of an increase in the number of unknown model parameters.

2.4.1 Analytical expressions for mean and variance

Equations (2.6) and (2.7) 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:

$$M\left(t\right) = \left(n_0 - a\right) e^{(\lambda-\mu)t} + a, \tag{2.6}$$

$$V\left(t\right) = \left(n_0 - a\right) \left(\frac{\lambda+\mu}{\lambda-\mu}\right) \left(e^{2(\lambda-\mu)t} - e^{(\lambda-\mu)t}\right), \tag{2.7}$$

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

We see from (2.6) that mean repeat length changes exponentially over time at a rate determined by the difference denoted $\phi = \lambda - \mu$. 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 and 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. However, in the next subsection, we describe a more systematic, likelihood-based approach to parameter estimation.

2.5 Model comparison and parameter estimation

2.5.1 Likelihood

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 (2.1) numerically in order to obtain the probability distribution function $P_n\left(t\right)$, which gives the probability that a repeat length is n at time t. The likelihood $L^{[i]}$ is then the product over all the data $d_j^{[i]}$, 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,

$$L^{[i]} = \prod_j P_{d_j^{[i]}}(t^{[i]};\theta^{[i]}), \tag{2.8}$$

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

$$L = \prod_i L^{[i]}. \tag{2.9}$$

The model parameters comprise the contraction rate, $\mu^{[i]}$, the expansion minus contraction rate, $\phi^{[i]}$, 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 an appropriate amount of synthetic data, generated from the underlying stochastic birth death process with known parameter values (see Section 2.7). This gives us an indication of the level of certainty arising from the inference procedure.

2.5.2 Evaluation of the likelihood

The likelihood, Equation (2.9), is calculated numerically using a computer algorithm. A representative MATLAB code with comments for evaluating the likelihood for an individual sample is provided in Appendix 1. The main numerical method used, *ode15s*, is based on a family of implicit schemes, known as backward differentiation formulae (BDF). The program implements the formulae between orders one and five, adaptively choosing both the order and the discretisation level (step size) in order to meet the specified error tolerance. The BDF family are examples of implicit linear multi-step methods, widely used because of their excellent stability properties (Shampine et al. 2003)

It is computationally very expensive to evaluate the full likelihood equation for reasons to do with the stiffness of the ODE problem. In Chapter 4, 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 (Wedderburn 1974) where the relationship between mean and variance can be used to inform a quasi-likelihood which has the required properties of a full likelihood. The full details of this approach are found in Chapter 4.

2.5.3 Model comparison

The Akaike information criterion (AIC) is used to assess the goodness of the fit of the model (Akaike 1974). AIC uses the maximised value of the likelihood of the model, L_{max}, penalised by the number of model parameters, k, to rank models thus

$$AIC = 2k - 2\log L_{max}. \tag{2.10}$$

As an alternative, the likelihood ratio test statistic can be estimated for pairs of nested models with maximised likelihoods L_{max1} and L_{max2} and number of independent parameters k_1 and k_2 respectively, as follows

$$2\left(\log L_{max2} - \log L_{max1}\right). \tag{2.11}$$

This statistic has asymptotically a $\chi^2_{k_2-k_1}$ distribution under the null hypothesis (Cox & Hinkley 1994) thus it can be established whether the difference between the two models is significant.

We obtain the maximum value of the likelihood by evaluating the likelihood over a broad parameter space, as illustrated in Tables 4.1, 5.1 and 6.1. Maximisation of the likelihood L in Equation (2.9) is equivalent to the maximisation of $L^{[i]}$, in Equation (2.8), of each dataset from an individual.

2.5.4 Bayesian parameter estimation

We use a Bayesian framework for parameter estimation. Bayes' theorem (Sivia 2006) states that the posterior distribution, π, of the parameters $\theta^{[i]}$ given the observed data $d_j^{[i]}$ is

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) = \frac{L\left(d_j^{[i]}|\theta^{[i]}\right)p\left(\theta^{[i]}\right)}{f\left(d_j^{[i]}\right)}, \tag{2.12}$$

where $L\left(d_j^{[i]}|\theta^{[i]}\right)$ is the likelihood of the data given the parameter values, $p\left(\theta^{[i]}\right)$ is the prior distribution of the parameters representing our initial beliefs about the parameter values before observing any data and $f\left(d_j^{[i]}\right)$ is the normalising constant that makes the posterior distribution a valid probability function, otherwise interpreted as the model evidence. Equation (2.12) has the important consequence

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) \propto L\left(d_j^{[i]}|\theta^{[i]}\right)p\left(\theta^{[i]}\right). \tag{2.13}$$

In the special case of a uniform prior, $p\left(\theta^{[i]}\right)$ is greater than zero only for a truncated range of $\theta^{[i]}$, see Table 4.1, and hence a constant c can be chosen so that the probabilities sum to unity and Equation (2.13) further simplifies to

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) \propto L\left(d_j^{[i]}|\theta^{[i]}\right). \tag{2.14}$$

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.

2.5.5 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

$$\pi\left(\theta^{[i]}|d_j^{[i]}\right) \propto L\left(d_j^{[i]}|\theta^{[i]}\right) p\left(\theta^{[i]}|\alpha_\mu, \beta_\mu, \alpha_\phi, \beta_\phi\right) p\left(\alpha_\mu\right) p\left(\beta_\mu\right) p\left(\alpha_\phi\right) p\left(\beta_\phi\right). \tag{2.15}$$

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

2.6 Other techniques

Sections 2.4 and 2.5 above outline the main mathematical and statistical tools used. Further techniques are introduced in context in the following chapters with their related experiments and results.

2.7 Synthetic experiments

One hundred datasets, of a comparable size to the individual data, were simulated using the Gillespie algorithm adapted for our specific stochastic process with the model parameters preassigned (Renshaw 1991, Wilkinson 2011). The inference procedure, described in detail in Chapter 4, Section 4.5.2, was then applied to infer the parameters back from the synthetic data set, as illustrated in Figure 2.3.

2.7.1 Simulation method

A pseudo code for the simulation of repeat length evolution in several cells using the Gillespie algorithm is as follows:

```
for each repeat length do
    initialise time, $t$, to 0 and repeat length, $N$, to the inherited repeat length value
    while $t$ is less than the age of the individual when the sample was taken do
        set $\lambda_N$ to $\lambda * (N - a)$ and $\mu_N$ to $\mu * (N - a)$
        choose a number $Y_1$ at random uniformly in $(0, 1)$
        if $Y_1 < \frac{\lambda_N}{\lambda_N + \mu_N}$ then
            the next event is an expansion and $N$ is updated to $N + 1$
        else
```

```
            the next event is a contraction and $N$ is updated to $N - 1$
        **end if**
        choose a number $Y_2$ at random uniformly in $(0, 1)$
        the time to the next event $s$ is $\frac{-\log(Y_2)}{\lambda_N+\mu_N}$ and $t$ is updated to $t + s$
    **end while**
**end for**
```

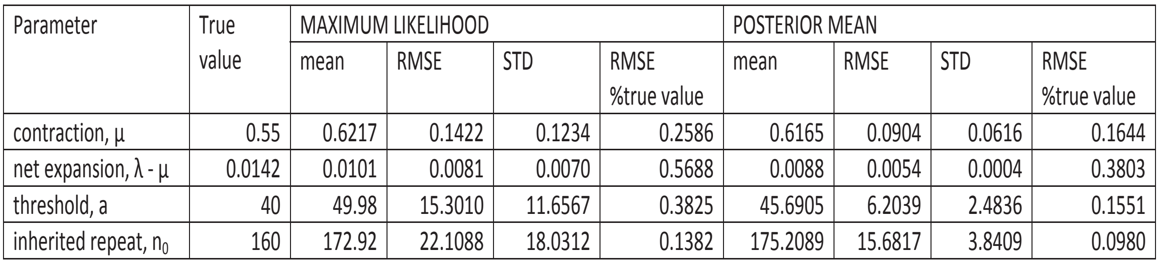
One hundred synthetic datasets were simulated from the expansion and contraction model with parameters $n_0 = 160$, $\mu = 0.55$, $\phi = 0.0142$, $a = 40$ and $t = 30$.

Root mean square error (RMSE) was calculated for the maximum likelihood (ML) solution and the posterior mean so that these potential point estimates could be assessed. Results are shown as an absolute distance and percentage difference of the underlying true parameter, see Table 2.1A. The ML estimate has smaller estimated bias than the posterior mean for inherited repeat length and net expansion, but in all cases the posterior mean has lower RMSE. To quantify the possible effect of PCR stutter (small errors in sizing alleles, discussed in Section 2.2), random unit amounts (-3,-2-1, 0, 1, 2, 3) were added t the dataset to recreate a situation where PCR stutter led to either lower or higher estimates. This does not appear to affect the accuracy of the parameter inference, see Table 2.1B. In summary, we attribute the uncertainty to finite sampling and would expect this to reduce if larger samples could be obtained.

A histogram of one illustrative synthetic dataset is shown in Figure 2.3. The same data is shown as a cumulative distribution, along with the inferred fit with the maximum likelihood value. The inferred parameter values correspond well with the actual values used to generate the synthetic data set and provide a good fit to the data.

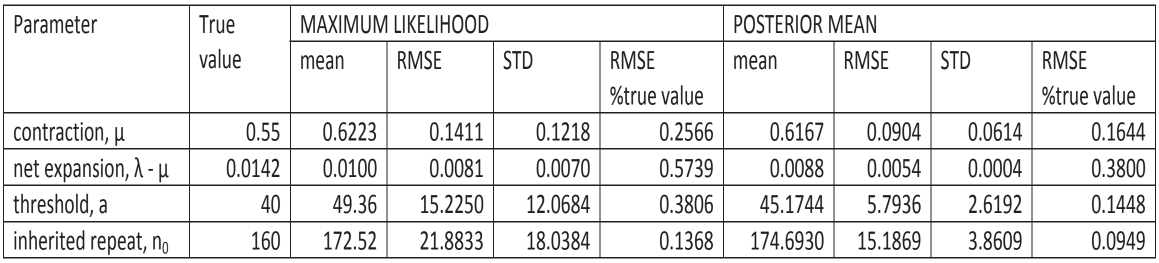
We investigated the posterior probability distributions for each parameter, marginalised by summing over all the other parameters, see Figure 2.3. The shape of the posterior probability distributions shown in Figure 2.3 convey the uncertainty in the parameter estimation. The crosses on each horizontal axis indicate the parameter value used to generated the data. Given that in this case we know the model that generated the data, the distribution reflects the stochasticity of the process and the sampling error. We see in Figure 2.3 that the credible interval for n_0 is fairly large, lying between 50 and 250 repeats, similarly for ϕ, between 0.001 and 0.06. Further analysis of these two parameters suggests that they are inversely correlated through the model. Consequently, these parameters are really only informative when considered together. This could be rectified by using prior knowledge about n_0 to improve the result for ϕ. For μ the distribution is clearly peaked, which suggests that this parameter is more well determined than the other parameters. The inference for parameter a is more clearly viewed jointly with parameter n_0 with a peak just below 50.

A. N=100 simulated datasets (t = 20 years)



Parameter	True value	MAXIMUM LIKELIHOOD				POSTERIOR MEAN			
		mean	RMSE	STD	RMSE %true value	mean	RMSE	STD	RMSE %true value
contraction, μ	0.55	0.6217	0.1422	0.1234	0.2586	0.6165	0.0904	0.0616	0.1644
net expansion, λ - μ	0.0142	0.0101	0.0081	0.0070	0.5688	0.0088	0.0054	0.0004	0.3803
threshold, a	40	49.98	15.3010	11.6567	0.3825	45.6905	6.2039	2.4836	0.1551
inherited repeat, n_0	160	172.92	22.1088	18.0312	0.1382	175.2089	15.6817	3.8409	0.0980

B. N=100 simulated datasets (t = 20 years) + random repeat unit(s) to recreate PCR stutter



Parameter	True value	MAXIMUM LIKELIHOOD				POSTERIOR MEAN			
		mean	RMSE	STD	RMSE %true value	mean	RMSE	STD	RMSE %true value
contraction, μ	0.55	0.6223	0.1411	0.1218	0.2566	0.6167	0.0904	0.0614	0.1644
net expansion, λ - μ	0.0142	0.0100	0.0081	0.0070	0.5739	0.0088	0.0054	0.0004	0.3800
threshold, a	40	49.36	15.2250	12.0684	0.3806	45.1744	5.7936	2.6192	0.1448
inherited repeat, n_0	160	172.52	21.8833	18.0384	0.1368	174.6930	15.1869	3.8609	0.0949

Table 2.1: **Analysis of the inference method.** Root mean square error (RMSE) was calculated for the maximum likelihood (ML) solution and the posterior mean so that these potential point estimates could be assessed. Results are shown as an absolute distance and percentage difference of the underlying true parameter.

The multi-modality seen originally in Figure 2.3D, still appears, when a finer grid is used, in approximately 1 in 5 cases, see the joint posterior probability distribution for the contraction rate and inherited repeat length, dataset 20, see Figure 2.4. Referring to the marginal posterior probability distribution for the contraction rate and inherited repeat separately, see Figure 2.5, an interpretation of this multi-modality is the ridge feature of the marginal posterior for inherited repeat length.

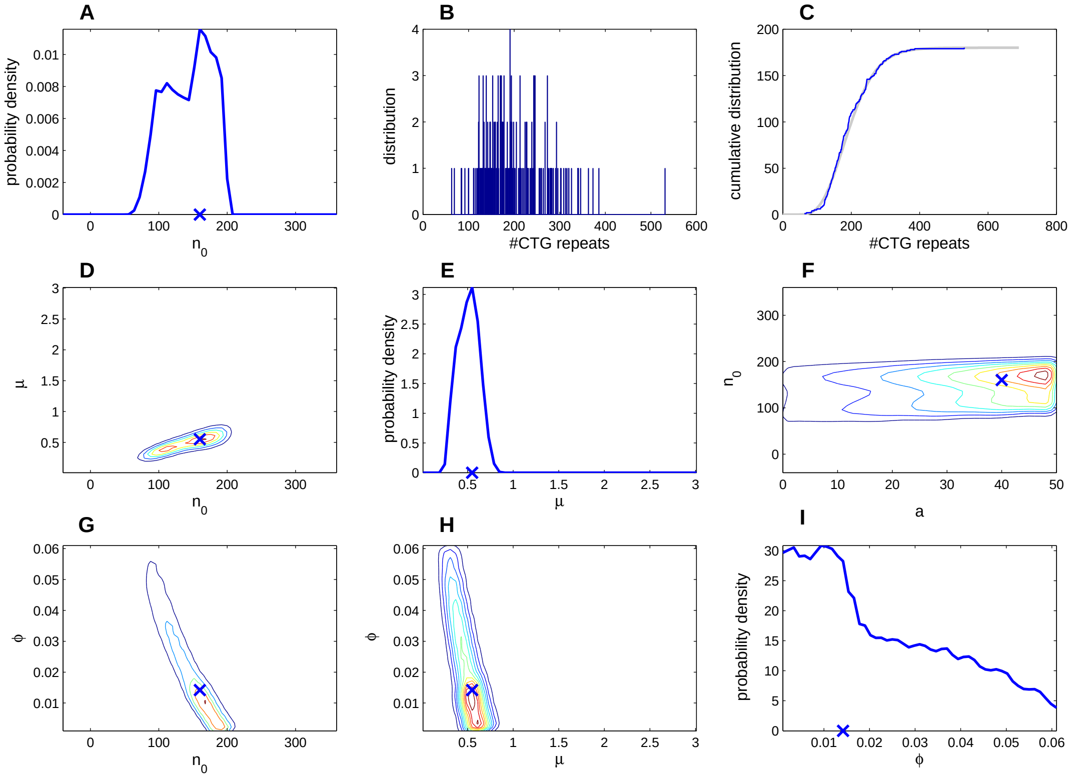


Figure 2.3: **Synthetic data and inference results from the expansion and contraction model with parameters:** $\mu = 0.55$**,** $\phi = 0.0142$**,** $n_0 = 160$**,** $a = 40$ **and** $t = 20$**.** B. The distribution of this synthetic dataset. C. The same data is shown as a cumulative distribution (dark line), along with the inferred fit with the maximum likelihood value (light line). These inferred parameter values are: $\mu = 0.61$, $\phi = 0.015$, $a = 50$ and $n_0 = 160$. The individual age, t, is taken as known and not inferred. The posterior probability density distributions for parameters n_0, the inherited repeat length, μ, the rate of contraction per CTG repeat per year, and ϕ the rate of expansion minus contraction per CTG repeat per year, marginalised for each parameter over the other parameters, are shown in panels A, E and I respectively. Marginalised joint probability distributions for parameter pairs, μ and n_0, ϕ and n_0, ϕ and μ, and n_0 and a, the threshold over which expansion and contraction occur are shown in panels D, G, H and F as contours with the dark to light direction representing increasing probability. The crosses on each horizontal axis indicate the parameter value used to generated the data. The shape of the distributions convey the uncertainty in the parameter estimation. Given that in this case we know the model that generated the data, the distribution reflects the stochasticity of the process and the sampling error.

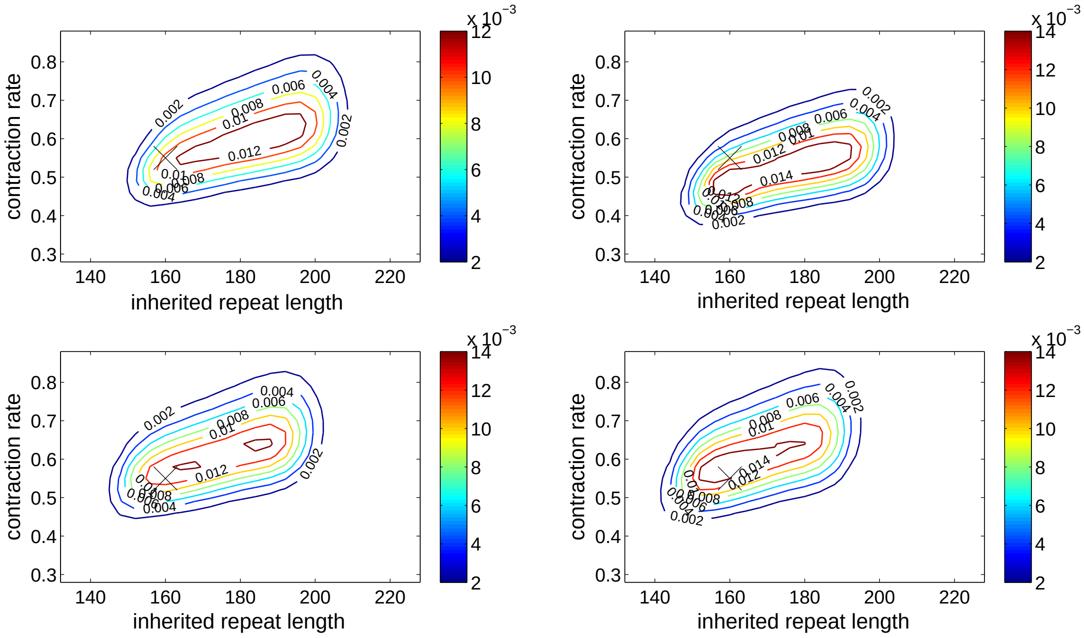


Figure 2.4: **Joint posterior distributions (μ and n_0) for four representative datasets**

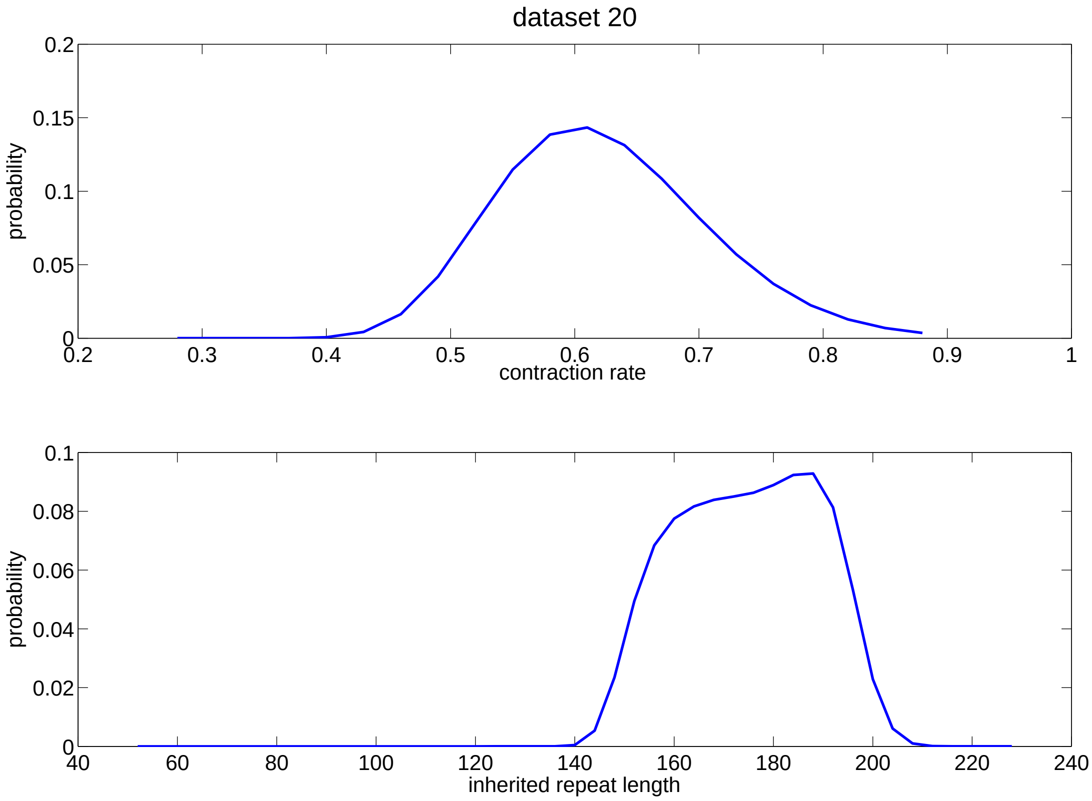


Figure 2.5: **Marginal posterior distributions (μ and n_0) for dataset 20**

Chapter 3

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

Some of the results in this Chapter (Sections 3.3.1 to 3.3.3) have been published, see Morales, F., Couto, J.M., Higham, C.F., Hogg, G., Cuenca, P., Braida, C., Wilson, R.H., Adam, B., del Valle, G., Brian, R., Sittenfeld, M., Ashizawa, T., Wilcox, A., Wilcox, D.E. and Monckton, D.G. (2012), *Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity*, Human Molecular Genetics (21), 3558 - 3567.

3.1 Abstract

The expanded CTG repeat in myotonic dystrophy type 1 (DM1) shows extremely high levels of somatic instability. These levels are age-dependent, length-dependent and tissue-specific. The presence of somatic instability compromises attempts to measure intergenerational repeat dynamics and infer genotype-phenotype relationships. Using single-molecule PCR, Morales *et al.* characterized more than 25,000 *de novo* somatic mutations from a large cohort of DM1 patients. This rich dataset enables us to fully quantify levels of somatic instability across a representative DM1 population for the first time. We establish the relationship between estimated progenitor allele length, age at sampling and levels of somatic instability using linear regression analysis. We show that the estimated progenitor allele length genotype is significantly better than modal repeat length (the current clinical standard) at predicting age of onset and this novel genotype is the major modifier of the age of onset phenotype. Further we show that somatic variation (adjusted for estimated progenitor allele length and age at sampling) is also a modifier of the age of onset phenotype. Several families form the large cohort, and we find that the level of somatic 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, ageing and inherited disease in the general population. We also investigate whether our findings can, in principle, be transferred to another dataset.

3.2 Introduction

This chapter focuses on the relationship between genotype and phenotype in myotonic dystrophy type 1 (DM1) and covers, in the first part of the chapter, the statistical analysis prepared for the publication of our findings (Morales et al. 2012). This primary dataset is used here to answer questions about the relationship between inherited repeat length (a novel genotype) and the clinical manifestation of disease (phenotype). In the second part of the chapter, we investigate transferring the findings from this dataset to another dataset to add value to the analysis. In later chapters, which form the main body of the thesis, we use this rich dataset in a different manner, to calibrate mathematical models and address a different set of questions.

As discussed in Chapter 1 (Introduction) affected DM1 individuals present with expansions from 50 CTGs to up to several thousand repeats (Brook et al. 1992). Longer alleles are associated with a more severe form of the disease and an earlier age of onset (Hunter et al. 1992, Harley et al. 1993, Redman et al. 1993). The expanded CTG repeat in DM1 shows extremely high levels of somatic instability. These levels are age-dependent, length-dependent and tissue-specific (Anvret et al. 1993, Ashizawa et al. 1993, Thornton et al. 1994, Monckton et al. 1995, Wong et al. 1995, Martorell 1998). Hence the allele lengths observed when a DNA sample is taken depend not only on the progenitor allele length but the age of the individual when, and the tissue from which, the DNA sample is taken.

Currently clinical diagnosis is based on a measure of modal repeat length from blood cells, but variance in modal length only accounts for between 20 - 40% of the variance in age of onset (Perini et al. 1999, Marchini et al. 2000, Mladenovic et al. 2006) and, therefore, is not an accurate predictive tool. Correlations with specific symptoms are often worse, or undetectable (Merlevede et al. 2002, Modoni et al. 2004, Gharehbaghi-Schneli et al. 2008). Hence the International Myotonic Dystrophy Consortium have recommended that patients are not offered prognostic information based on the current test (Gonzalez et al. 2000). We hypothesise that previous genotype-phenotype correlations have been compromised by failure to take into account the age-dependent, expansion biased nature of somatic mosaicism.

Single-molecule based small pool PCR approaches resolve the heterogeneous smear of CTG repeats into the discrete alleles present in individual cells (Monckton et al. 1995). This provides a quantitative measure of repeat length variation and reveals the underlying shape of the distribution of repeat lengths. Typically, repeat length distributions for the mutant allele in DM1 blood DNA are positively skewed with a relatively sharp lower boundary below which smaller alleles are relatively rare. This lower boundary is conserved between tissues over time and provides a useful estimate for the inherited or progenitor allele length (Monckton et al. 1995). A key aim in this chapter is to estimate progenitor allele length and measure the total level of allelic variation in blood DNA in a large number of DM1 patients using single-molecule based small pool PCR data. We quantify the effect of progenitor allele length and age at sampling on somatic variation and the relationship of these effects with disease severity. We also investigate whether variation in somatic instability is heritable.

If 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 holds then there is a wealth of data to revisit. For example, a large study at 23 neuromuscular disease clinics in the United States was initiated in April 1997 with on-going follow-up and comprises 406 DM1 affected individuals (Groh et al. 2008). At the end of this chapter, we investigate whether findings from our in-depth analysis of repeat length distribution and age of onset can, in principle, be transferred to another dataset (Groh et al. 2008).

3.3 Results

The full dataset comprises sized repeat lengths in blood DNA from 145 DM1 individuals as described in Chapter 2, Section 2.2.

In DM1 and HD, age at onset is considered to have biological and clinical relevance as it takes into account both when the disease might start and the severity of the symptoms (typically the symptoms associated with late onset are much less severe than those associated with juvenile/adult onset). This information would be useful in reproductive counselling to as it makes sense of the patterns of inheritance seen in families due to the phenomenon of anticipation. Hence efforts are directed at looking for explanatory variables for age of onset. However several issues have to be taken into account. First, modal repeat length, MA, an obvious candidate as an explanatory variable, is highly dependent on the age at which it is measured, AS. Second, as typically many individuals are recruited to studies after the symptoms have appeared in their families, there is a strong correlation between AS and age of onset. One way to remove the effect of age from this analysis is to consider the progenitor repeat length, PAL, which is the modal repeat length at birth. Another approach, would be to consider the difference between age of onset and AS, time to onset, TTO as the response. In our study and others, TTO is negative for many individuals as they are symptomatic, but this does not affect our ability to use TTO as a response indicator. Taking all these issues into account we analyse a series of models using TTO as the response dependent variable (model series 1) and age of onset, AO, as the response dependent variables. The independent explanatory variables are modal repeat length, MA, progenitor allele length, PAL, age at sampling, AS and finally, somatic instability, SI, which is defined as the difference in the number of repeats at the 10^{th} and at the 90^{th} percentile of the repeat length distribution.

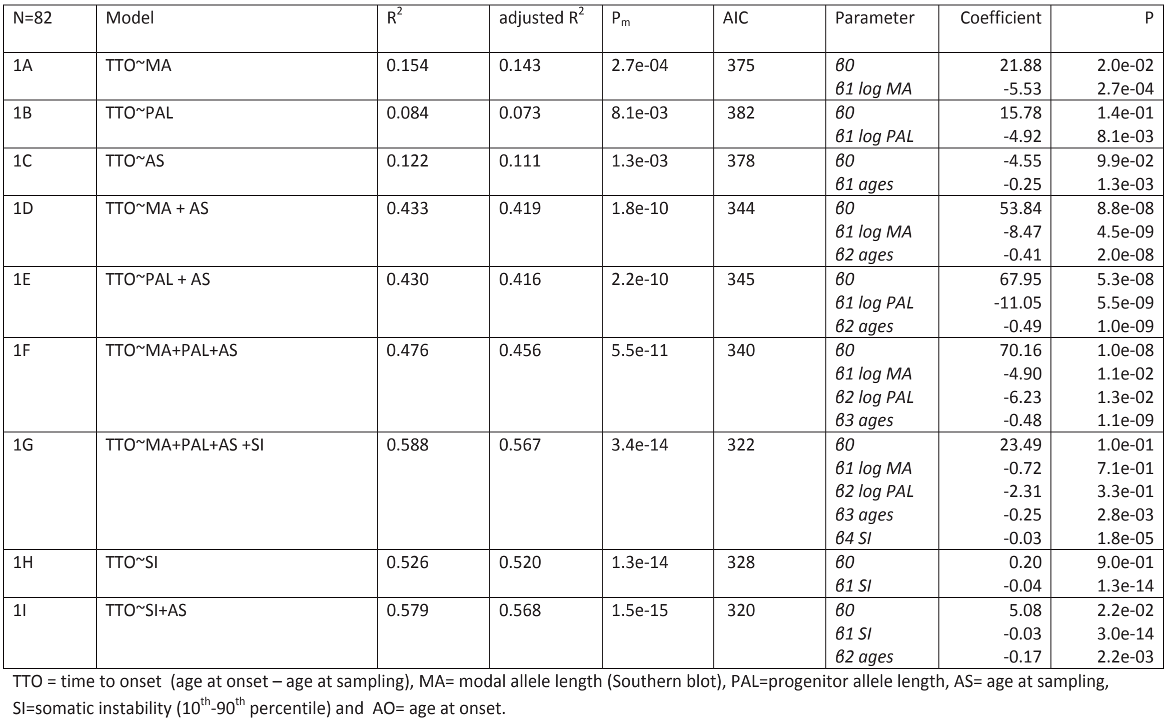
3.3.1 Progenitor allele length is a modifier of age of onset in DM1

The progenitor allele length was estimated from the lower boundary of the repeat length distribution established by small pool PCR analysis (Monckton et al. 1995) for 137 DM1 affected individuals for whom age of onset was known. Modal repeat length measured via a traditional Southern blot (SB) of restricted digested genomic DNA, currently the clinical method for establishing CTG repeat length and diagnosis of DM1, was available for a subset of 82 individuals.

Linear regression analysis was used to fit MA (Models 1A and 2A), $MA + AS$ (Models 1D and 2D), $MA + PAL + AS$ (Models 1F and 2F) and $MA + PAL + AS + SI$ (Models 1G and 2G) to TTO (Models 1A-1G) and AO (Models 2A-2G) respectively, see Tables 3.1 and 3.2. The models were compared using adjusted R^2 and AIC criteria. Concerning TTO, including AS improves the model (1D vs 1A), including PAL further improves the model (1F vs 1D) and adding SI further improves the model (1G vs 1F). These results support the basic premise that PAL and SI explain some of the variance in disease onset and progression not already explained by modal repeat length.

In terms of response variables, AO is the better response variable in terms of the adjusted R^2 criteria but comparable to TTO in terms of the AIC criteria, explained by higher correlations under AO but equivalent residuals, and hence fit between AO and TTO.

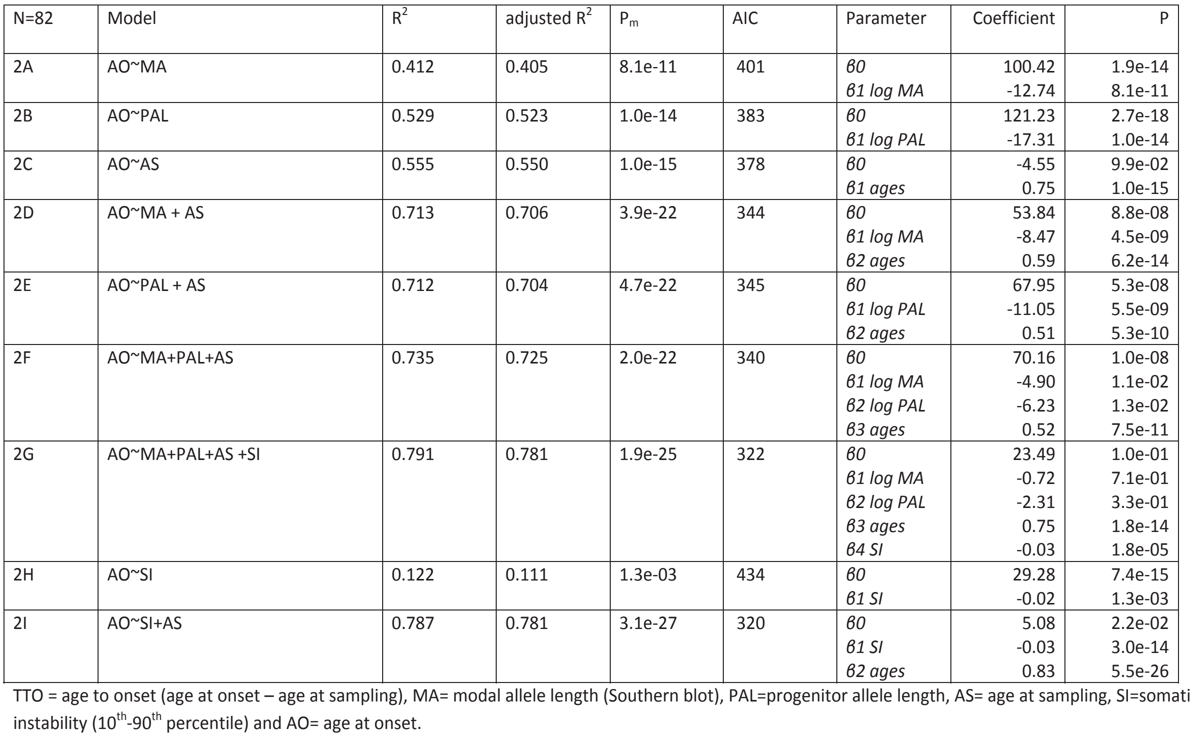
A series of models (linear, quadratic, exponential and power) were then fitted to all 137 DM1 affected individuals using least squares regression analysis with age of onset as the dependent variable and the logarithm (base 10) of the estimated progenitor allele length as the independent variable. For all these models the negative sign of the coefficient for the independent variable indicates that age at onset and estimated progenitor allele length are inversely correlated, with age of onset decreasing as progenitor allele length increases. The non-linear models (adjusted $R^2 \approx 0.7$, $P < 0.0001$) provided a better fit than the linear model (adjusted $R^2 = 0.640$, $P < 0.0001$) which suggests that



N=82	Model	R^2	adjusted R^2	P_m	AIC	Parameter	Coefficient	P
1A	TTO~MA	0.154	0.143	2.7e-04	375	*β0* *β1 log MA*	21.88 -5.53	2.0e-02 2.7e-04
1B	TTO~PAL	0.084	0.073	8.1e-03	382	*β0* *β1 log PAL*	15.78 -4.92	1.4e-01 8.1e-03
1C	TTO~AS	0.122	0.111	1.3e-03	378	*β0* *β1 ages*	-4.55 -0.25	9.9e-02 1.3e-03
1D	TTO~MA + AS	0.433	0.419	1.8e-10	344	*β0* *β1 log MA* *β2 ages*	53.84 -8.47 -0.41	8.8e-08 4.5e-09 2.0e-08
1E	TTO~PAL + AS	0.430	0.416	2.2e-10	345	*β0* *β1 log PAL* *β2 ages*	67.95 -11.05 -0.49	5.3e-08 5.5e-09 1.0e-09
1F	TTO~MA+PAL+AS	0.476	0.456	5.5e-11	340	*β0* *β1 log MA* *β2 log PAL* *β3 ages*	70.16 -4.90 -6.23 -0.48	1.0e-08 1.1e-02 1.3e-02 1.1e-09
1G	TTO~MA+PAL+AS +SI	0.588	0.567	3.4e-14	322	*β0* *β1 log MA* *β2 log PAL* *β3 ages* *β4 SI*	23.49 -0.72 -2.31 -0.25 -0.03	1.0e-01 7.1e-01 3.3e-01 2.8e-03 1.8e-05
1H	TTO~SI	0.526	0.520	1.3e-14	328	*β0* *β1 SI*	0.20 -0.04	9.0e-01 1.3e-14
1I	TTO~SI+AS	0.579	0.568	1.5e-15	320	*β0* *β1 SI* *β2 ages*	5.08 -0.03 -0.17	2.2e-02 3.0e-14 2.2e-03

TTO = time to onset (age at onset – age at sampling), MA= modal allele length (Southern blot), PAL=progenitor allele length, AS= age at sampling, SI=somatic instability (10th-90th percentile) and AO= age at onset.

Table 3.1: **Linear regression analysis to fit different models to the response variable TTO.**



N=82	Model	R^2	adjusted R^2	P_m	AIC	Parameter	Coefficient	P
2A	AO~MA	0.412	0.405	8.1e-11	401	*β0* *β1 log MA*	100.42 -12.74	1.9e-14 8.1e-11
2B	AO~PAL	0.529	0.523	1.0e-14	383	*β0* *β1 log PAL*	121.23 -17.31	2.7e-18 1.0e-14
2C	AO~AS	0.555	0.550	1.0e-15	378	*β0* *β1 ages*	-4.55 0.75	9.9e-02 1.0e-15
2D	AO~MA + AS	0.713	0.706	3.9e-22	344	*β0* *β1 log MA* *β2 ages*	53.84 -8.47 0.59	8.8e-08 4.5e-09 6.2e-14
2E	AO~PAL + AS	0.712	0.704	4.7e-22	345	*β0* *β1 log PAL* *β2 ages*	67.95 -11.05 0.51	5.3e-08 5.5e-09 5.3e-10
2F	AO~MA+PAL+AS	0.735	0.725	2.0e-22	340	*β0* *β1 log MA* *β2 log PAL* *β3 ages*	70.16 -4.90 -6.23 0.52	1.0e-08 1.1e-02 1.3e-02 7.5e-11
2G	AO~MA+PAL+AS +SI	0.791	0.781	1.9e-25	322	*β0* *β1 log MA* *β2 log PAL* *β3 ages* *β4 SI*	23.49 -0.72 -2.31 0.75 -0.03	1.0e-01 7.1e-01 3.3e-01 1.8e-14 1.8e-05
2H	AO~SI	0.122	0.111	1.3e-03	434	*β0* *β1 SI*	29.28 -0.02	7.4e-15 1.3e-03
2I	AO~SI+AS	0.787	0.781	3.1e-27	320	*β0* *β1 SI* *β2 ages*	5.08 -0.03 0.83	2.2e-02 3.0e-14 5.5e-26

TTO = age to onset (age at onset – age at sampling), MA= modal allele length (Southern blot), PAL=progenitor allele length, AS= age at sampling, SI=somati instability (10th-90th percentile) and AO= age at onset.

Table 3.2: **Linear regression analysis to fit different models to the response variable AO.**

age of onset decreases more slowly as progenitor allele length increases or equivalently, that age of onset increases more rapidly as progenitor length decreases.

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

We quantify somatic variation for an individual patient as the difference in the number of repeats at the 10^{th} and at the 90^{th} percentile of the repeat length distribution. This was recorded via single molecule PCR for 136 DM1 affected or at risk individuals in total. The group contains some asymptomatic patients whereas the group of 137 used in Section 3.3.1 did not. This measure of variation captures the repeat length range of alleles whilst eliminating outliers that are sensitive to finite sampling. Linear regression analysis was performed with a series of models that took the logarithm (base 10) of somatic variation as the dependent variable and either the logarithm (base 10) of inherited allele length, the logarithm (base 10) of age at sampling or a combination of both as the independent explanatory variables. The objective was to establish whether and to what extent age and progenitor length modify the level of somatic instability.

Alone, progenitor allele length, $\log PAL$, is positively correlated to somatic variation, $\log SI$ (adjusted $R^2 = 0.644$, $P < 0.0001$) whereas age at sampling, $\log age_s$, is not significantly correlated to $\log SI$ (adjusted $R^2 = -0.005$, $P = 0.6$). Together, $\log PAL$ and $\log age_s$ are both significantly correlated to $\log SI$ (adjusted $R^2 = 0.746$ and $P < 0.0001$) with coefficient p-values, 8.3×10^{-42} and 1.2×10^{-11}, respectively, see Table 3.3. These results suggest that progenitor allele length has a greater effect than age at sampling on levels of somatic instability. Inclusion of the interactive term, $\log PAL \times \log age_s$ ($P = 6.5 \times 10^{-3}$), and quadratic terms, $\log PAL^2$ ($P = 2.3 \times 10^{-24}$) and $\log age_s^2$ ($P = 5.6 \times 10^{-5}$), results in a better fit, allowing for the extra parameters (adjusted $R^2 = 0.890$, $P < 0.001$), and indicates that the relationship between $\log SI$, $\log PAL$ and $\log age_s$ is non-linear and complex. Analysis of the residuals, in terms of constant variance, in particular for the smaller SI values, improves and confirms the superior fit of the quadratic model, see Figure 3.1. We observe also that the trend in the plot for the linear models disappears when the quadratic term is included, justifying the need for the non-linear term.

Model	R^2	adj R^2	P_m	parameter		coeff	std error	t-stat	P
log(SI) = β_0 + $\beta_1$log(PAL)	0.646	0.644	< 0.0001	intercept	β_0	-0.84	0.21	-4.0	9.3 x 10^{-05}
				log(PAL)	β_1	1.29	0.08	15.7	4.9 x 10^{-32}
log(SI) = β_0 + $\beta_1$log(age$_s$)	0.002	-0.005	0.60	intercept	β_0	2.27	0.25	9.2	7.4 x 10^{-16}
				log(age$_s$)	β_1	0.09	0.16	0.5	6.0 x 10^{-01}
log(SI) = β_0 + $\beta_1$log(PAL) + $\beta_2$log(age$_s$)	0.750	0.746	< 0.0001	intercept	β_0	-2.24	0.26	-8.7	1.2 x 10^{-14}
				log(PAL)	β_1	1.47	0.07	20.0	8.3 x 10^{-42}
				log(age$_s$)	β_2	0.65	0.09	7.4	1.2 x 10^{-11}
log(SI) = β_0 + $\beta_1$log(PAL) + $\beta_2$log(age$_s$) + $\beta_3$log(PAL)*log(age$_s$)	0.764	0.759	< 0.0001	intercept	β_0	-0.40	0.70	-0.6	5.7 x 10^{-01}
				log(PAL)	β_1	0.79	0.25	3.2	2.1 x 10^{-03}
				log(age$_s$)	β_2	-0.56	0.44	-1.3	2.0 x 10^{-01}
				log(PAL)* log(age$_s$)	β_3	0.44	0.16	2.8	5.8 x 10^{-03}
log(SI) = β_0 + $\beta_1$log(PAL) + $\beta_2$log(age$_s$) + $\beta_3$log(PAL)*log(age$_s$) + $\beta_4$log(PAL)^2 + $\beta_5$log(age$_s$)^2	0.894	0.890	< 0.0001	intercept	β_0	-9.04	0.99	-9.1	1.1 x 10^{-15}
				log(PAL)	β_1	8.78	0.68	12.9	4.3 x 10^{-25}
				log(age$_s$)	β_2	-1.62	0.58	-2.8	6.5 x 10^{-03}
				log(PAL)* log(age$_s$)	β_3	0.40	0.15	2.7	7.7 x 10^{-03}
				log(PAL)^2	β_4	-1.67	0.13	-12.6	2.3 x 10^{-24}
				log(age$_s$)^2	β_5	0.44	0.11	4.2	5.6 x 10^{-05}

Table 3.3: **The relationship between somatic instability (SI), estimated progenitor allele length (PAL) and age at sampling (age_s), established using regression analysis.** The table shows the squared coefficient of correlation (R^2) and the statistical significance (P_m) for each relationship, and the coefficient, standard error, t-statistic and statistical significance (P) associated with each parameter in the linear regression analysis (N = 136).

3.3.3 Heritability analysis

Genetic correlations and heritability estimates were acquired using QTDT, a general test package for the association of quantitative measures in nuclear families (Abecasis et al. 2000). After correcting for the two major modifiers of somatic instability (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 may be attributable to genetic modifiers and therefore may be heritable. Of the 136 individuals with derived repeat length distributions, 89 were part of 21 families and formed 51 sibling pairs. Using QTDT (Abecasis et al. 2000), 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 heritability, $V_g = 0.42$, $V_e = 0.58$ and $V_c = 0$, establishing residual somatic instability as a heritable quantitative genetic trait.

3.4 Extension to another dataset

We hypothesise that the interpretation of SB modal repeat length is compromised by not taking progenitor allele length and age at sampling into account. Implicit in this assumption is that there is a relationship between progenitor allele length, age at sampling and SB modal repeat length. If such a relationship can be established, in a simple analytical manner, then it should be possible to deduce one of these variables from the other two. In particular, given the predictive importance of progenitor allele length, it would be useful to deduce progenitor allele length from age at sampling and SB modal repeat length. Blood DNA samples are taken in many DM1 or other related disease studies and variables such as age at sampling or SB modal length are usually known or measured, whereas progenitor allele length is not typically known or measured.

In our dataset, there are 82 individuals for whom we have age at sampling, an estimate of progenitor allele length from small pool PCR analysis and in addition, SB modal repeat length measured from traditional Southern blot of restricted digested genomic DNA. We confirmed the relationship between these variables statistically using linear regression analysis. We then projected this framework on to an American dataset (Groh et al. 2008), kindly provided by Dr William Groh and described below, to estimate progenitor allele length from SB modal repeat length and age at sampling. We therefore now investigate whether this quantity can add predictive value to the American

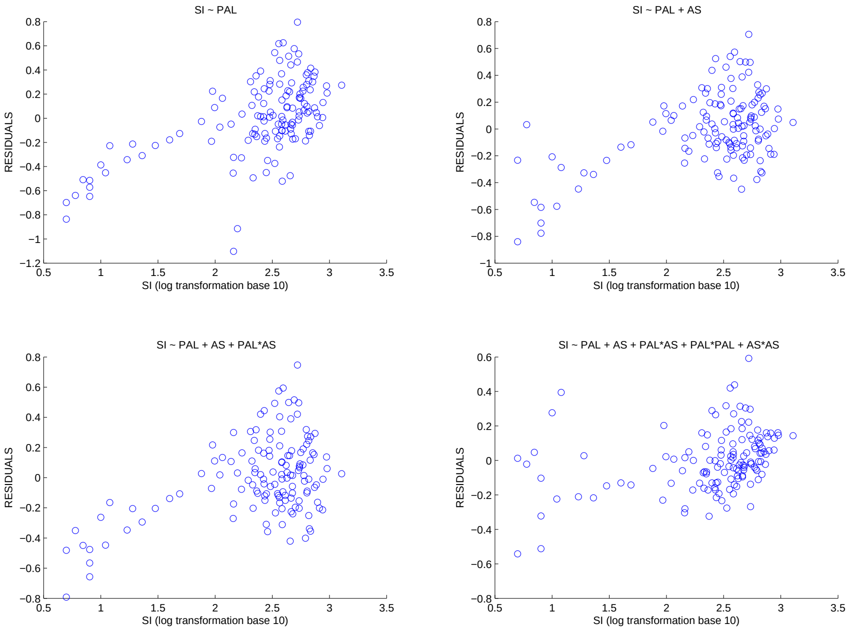


Figure 3.1: **Residual analysis for SI response models** Residuals (vertical axis) plotted against the response variable SI (log transformation base 10) for four models, see Table 3.3 for further details.

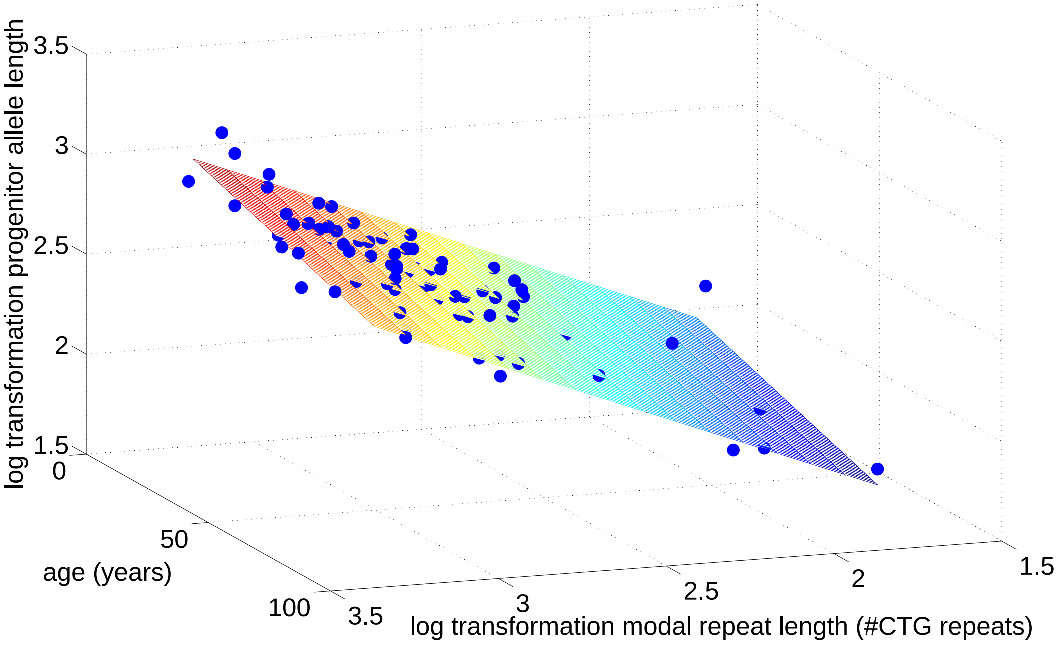


Figure 3.2: **Summary of the Glasgow data:** Log transformation of estimated progenitor allele length (vertical axis) versus log transformation of CTG modal repeat length determined by Southern blot and age at sampling (horizontal axes) for 82 patients from the Glasgow study. Surface fitted using linear regression ($R^2 = 80\%$).

data.

3.4.1 American data

A large study at 23 neuromuscular disease clinics in the United States was initiated in April 1997 with on-going follow-up. Patients comprise adults (18 years and older) with a clinical diagnosis of myotonic dystrophy and an abnormal CTG repeat sequence (one or both alleles with ≥ 38 repeats) confirmed by the traditional Southern blot of restricted digested genomic DNA (Groh et al. 2008). Also available were the ages of individuals when the blood DNA for diagnosis was taken.

3.4.2 Relationship between age of onset, SB modal repeat length, progenitor allele length and age at sampling in the University of Glasgow study

Variance of SB modal repeat length and age at sampling explains about 80% of the variance in the estimated progenitor allele length, see Figure 3.2. This finding supports the use of the fitted linear model to predict progenitor allele length for new patients given SB modal repeat length and age at sampling.

Including $MA + AS$ in the linear model to explain age of onset is mathematically equivalent to including $PAL + AS$, see Table 3.2, Models 2D and 2E. the explanation for this is that $MA + AS$ is a proxy for PAL which is the biologically meaningful parameter as opposed to AS.

3.4.3 Adding value to the American data

We analysed the American data following the approach in Section 3.3.1, see Table 3.4. Currently, SB modal repeat length is used to indicate broadly the phenotype and corresponding age of onset. In the American study, the variance of SB modal repeat length explains about 26% of the variance in age of onset (Figure 3.3, top row). This increases to 35% if the SB modal repeat length is transformed using logarithms (Figure 3.3, middle row). $MA + AS$ is treated as a proxy for PAL and this substitution improves the explained variance by a further 12 percentage points to 63%.

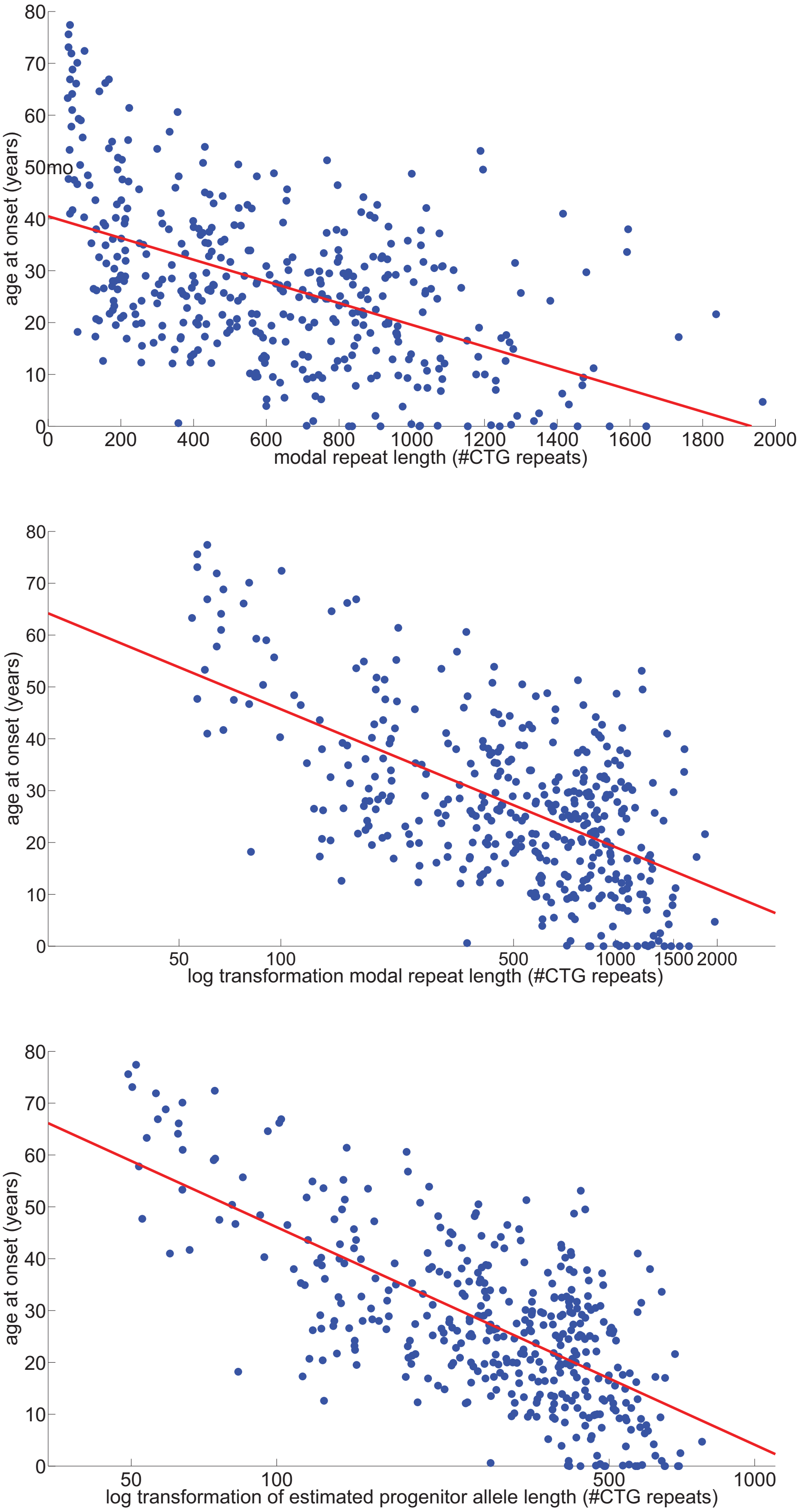


Figure 3.3: **Summary of the American data.** **Top row**: Scatter plot of SB modal repeat length (horizontal axis) and age at onset (vertical axis) for 406 US patients. Line fitted using linear regression ($R^2 = 26\%$). **Middle row**: Scatter plot of the log transformation of SB modal repeat length (horizontal axis) and age at onset (vertical axis) for the same 406 US patients ($R^2 = 35\%$). **Bottom row**: Scatter plot of the log transformation of estimated progenitor allele length (horizontal axis) and age at onset (vertical axis) for the same 406 US patients ($R^2 = 47\%$).

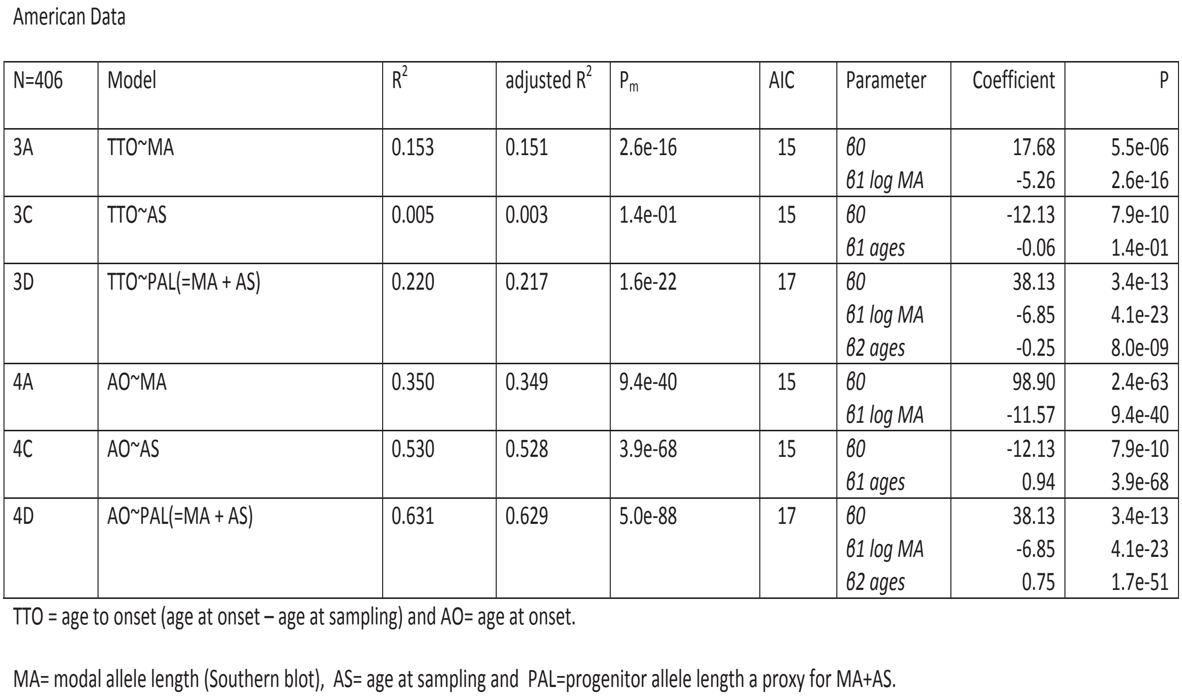
3.5 Discussion

We have shown that the progenitor allele length is better than SB modal repeat length in blood for predicting age of onset. A key factor behind this improvement is that progenitor allele length does not depend on age or tissue whereas both these factors affect and confound the interpretation of SB modal repeat length. SB modal repeat length is expected to increase with age and vary between tissues (Monckton et al. 1995) whereas the true progenitor allele length does not. Blood DNA is relatively stable compared to muscle DNA and so SB modal repeat length measured in blood may not reflect levels of instability in the disease related tissues pertinent to age of onset and disease progression. However blood is easily accessible in a large number of patients and is a tissue within which the repeat remains relatively stable. Analysing blood DNA thus gives us a good chance to estimate the progenitor allele length. Measuring instability in other tissues poses several challenges. Muscle biopsies are considered too invasive for routine testing and other tissues are only available post mortem. Complex tissues often display multi-modal distributions reflecting the presence of very different cell types within the same tissue. Although SB modal length in muscle would indicate actual levels of instability realised, in the absence of this measure the progenitor allele is closely associated with the DM1 phenotype (congenital, juvenile, adult and late adult) and is expected to be indicative of age of onset.

Some studies, such as (Groh et al. 2008), have measured the SB modal repeat length in blood and used this measure in their analysis. For those studies where age at sampling DNA is also recorded, we suggest that is possible to reinterpret the findings. In particular, we have shown that it is possible to use modal repeat length and age at sampling as a proxy for progenitor allele length. This allows re-interpretation of the relationship between genotype and phenotype using a novel genotype with more predictive power. This approach can add value to a secondary dataset and extends the range of analysis.

Basing progenitor allele length on the lower boundary is subjective and relies on the progenitor allele being sufficiently prevalent in the sample when, conceivably for advanced DM1 individuals, this may not be the case when the progenitor allele has mostly mutated and mutations have blurred the lower boundary. The main work of this thesis, described in the subsequent chapters, takes a different, more objective approach. By using a mathematical model to describe the evolution of the repeat (Chapter 4), we treat the progenitor allele length as an unknown entity and infer its value from the data. However, importantly, we have shown here that a simple estimate of progenitor allele length is available that forms a good predictor of age of onset. This type of readily accessible prognostic information is very important for families with DM1. Age to onset is an additional response variable with prognostic utility but we would argue that for DM1, it is not clear that time to onset represents a biologically relevant outcome. Thinking about how we would want to use the model in a predictive sense *e.g.* in prenatal testing then the AS versus AO ascertainment mediated correlation would on longer exist.

Overall in this chapter, evidence of somatic expansion in tissues that are the targets of pathogenesis has informed the hypothesis that somatic instability may itself contribute to the pathogenic process. Through quantification of somatic variation in a large cohort of DM1 individuals, we are able to show, for the first time, that somatic variation, adjusted for estimated progenitor allele length and age at sampling, is a modifier of the age of onset phenotype. These important results concerning somatic instability are discussed again in more detail in Chapter 5. Several families form the large cohort, and we find that the level of somatic instability is highly heritable, implying a role for individual-specific *trans*-acting genetic modifiers. Identifying these *trans*-acting genetic modifiers is a future direction for this work that is discussed in more detail in Chapter 8.



American Data

N=406	Model	R^2	adjusted R^2	P_m	AIC	Parameter	Coefficient	P
3A	TTO~MA	0.153	0.151	2.6e-16	15	*β0* *β1 log MA*	17.68 -5.26	5.5e-06 2.6e-16
3C	TTO~AS	0.005	0.003	1.4e-01	15	*β0* *β1 ages*	-12.13 -0.06	7.9e-10 1.4e-01
3D	TTO~PAL(=MA + AS)	0.220	0.217	1.6e-22	17	*β0* *β1 log MA* *β2 ages*	38.13 -6.85 -0.25	3.4e-13 4.1e-23 8.0e-09
4A	AO~MA	0.350	0.349	9.4e-40	15	*β0* *β1 log MA*	98.90 -11.57	2.4e-63 9.4e-40
4C	AO~AS	0.530	0.528	3.9e-68	15	*β0* *β1 ages*	-12.13 0.94	7.9e-10 3.9e-68
4D	AO~PAL(=MA + AS)	0.631	0.629	5.0e-88	17	*β0* *β1 log MA* *β2 ages*	38.13 -6.85 0.75	3.4e-13 4.1e-23 1.7e-51

TTO = age to onset (age at onset – age at sampling) and AO= age at onset.

MA= modal allele length (Southern blot), AS= age at sampling and PAL=progenitor allele length a proxy for MA+AS.

Table 3.4: **American Data: Linear regression analysis to fit different models to the response variables TTO and AO respectively.**

Chapter 4

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

The results in this Chapter (Sections 4.3.1 to 4.3.5) have been published, see 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,* Human Molecular Genetics (21), 2450 - 2463.

4.1 Abstract

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 toward 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 parametrise 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.

4.2 Introduction

The main 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 experimental data. This chapter focuses on the results from fitting mathematical models to the primary dataset described in Chapter 3, as published in (Higham et al. 2012). As discussed in more detail in Chapter 2, by sampling many cells from individuals we have many realisations of the same stochastic process at one point in time. Hence, our data provides 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.

4.3 Results

4.3.1 Model definitions

The features of the dynamics underlying repeat length instability are largely unknown. By fitting different models which represent different hypotheses about this mechanism to the dataset 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 hypothesised by (Kaplan et al. 2007), 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 threshold 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 M_1);
- expansion only with an individual-specific parameter for expansion and a fixed threshold (Model M_{2a});
- expansion only with individual-specific parameters for expansion and threshold (Model M_{2b});

- expansion and contraction with global parameters for expansion and contraction, and a fixed threshold (Model M_3);
- expansion and contraction with a global parameter for contraction and individual-specific parameters for expansion and threshold (Model M_4);
- expansion and contraction with a global parameter for expansion and individual-specific parameters for contraction and threshold (Model M_5);
- expansion and contraction with individual-specific parameters for expansion and contraction, and a fixed threshold (Model M_{6a}); and finally
- expansion and contraction with individual-specific parameters for expansion, contraction and threshold (Model M_{6b}).

4.3.2 Model comparison

We used model comparison methods as described in Chapter 2, Section 2.5.3, 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 maximised log-likelihood with the Akaike information criterion and the likelihood ratio test as the bases 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) was used to fit the eight models, described above, representing the different hypotheses. As detailed below in Section 4.5.1, 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, $\phi = \lambda - \mu$. These parameters were treated as unknowns and investigated over a broad range of values (Table 4.1). To formally compare the different models, we used the Akaike information criterion (AIC) (Akaike 1974, Burnham &

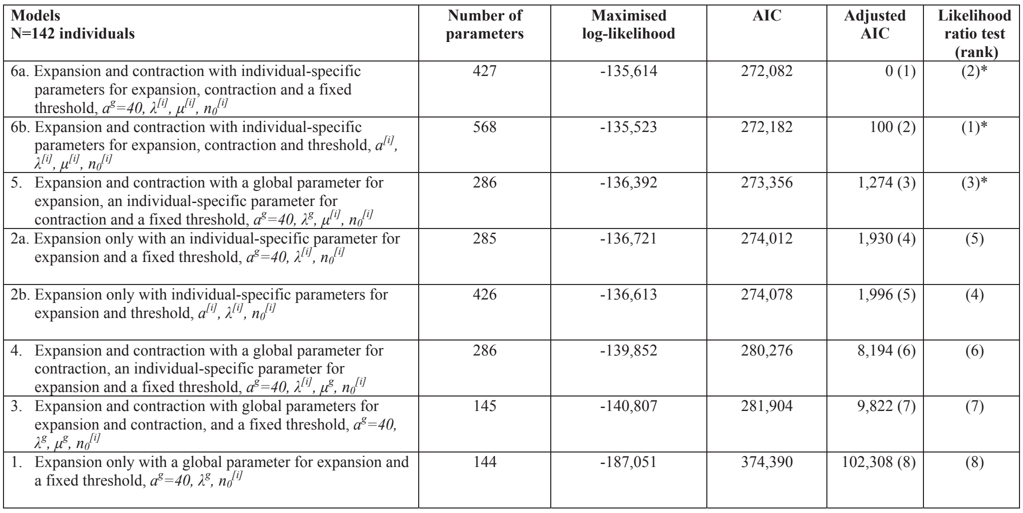
Anderson 2002) and the likelihood ratio test (Cox & Hinkley 1994) which both employ the maximised log-likelihood penalised by the number of parameters in the model, summarised in Table 4.2. The likelihood of the data given the model arose naturally from the stochastic process and we obtained the maximised log-likelihood value for each model. Further details of how the models and their likelihood were derived is found in Chapter 2, Sections 2.4 and 2.5.

Parameters	**Range for uniform prior$^{(i)}$(small alleles)**	**Increment size**
Contraction, rate per CTG repeat per year, $\boldsymbol{\mu}$	0.01 to 3.01 (0.001 to 0.011)	0.06
Expansion minus contraction, rate per CTG repeat per year, $\boldsymbol{\varphi}$	0.001 to 0.061 (0.001 to 0.026)	0.0012
Threshold, number of CTG repeats, $\boldsymbol{a}$	0 to 50	1
Inherited repeat length, number of CTG repeats, $\boldsymbol{n_0}$	82 to PAL$^{(ii)}$ + 200^{2} (51 – 81)^{1}	8 (2)

Table 4.1: **Prior ranges for parameter estimation for myotonic dystrophy type 1 blood.** Notes: (i) This range was adapted for some patients with: ^{1} small alleles in order to investigate smaller rates of contraction (see figures in parentheses); and ^{2} possibly unreliable progenitor allele length (PAL) estimates due to ambiguous or dispersed distributions. This included both extending up to the maximum possible value and down to the pathological threshold of 50 CTGs; and (ii) 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.

The very negative values of the maximised log-likelihoods, around -1.35×10^5, reflect the vast quantity of data (between 100 and 350 samples for each of the 142 individuals) and lead to correspondingly large AICs. However, 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 M_{6a} (expansion and contraction with individual-specific parameters and a fixed threshold) and model M_{6b} (expansion and contraction with individual-specific parameters and a variable threshold) with adjusted AICs of 0 and 100 respectively followed by model M_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 1,274. Expansion only models M_{2a} and M_{2b} have adjusted AICs of 1,930 and 1,996 respectively. Comparing models M_{6a} and M_{6b} using the likelihood ratio test indicates that the difference between these models is of low significance ($P = 0.01$). However comparing models M_{6a}, M_{6b} and M_5 to model M_{2a} using the likelihood ratio test gives a highly significant result ($P < 10^{-15}$). The Bonferroni correction for eight multiple tests 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 8,194, than the models with global parameters, AICs ranging from 9,822 to 102,308 (ranked 7 to 8 in Table 4.2). This suggests that there is significant parameter variation between individuals. When considering the threshold parameter, a, we observe that model M_{6a} ($a^g = 40$) provides a better fit to the data as model M_{6b} (individual values for a), providing support for the involvement of a universal threshold effect in the mechanism of repeat instability. This finding is consistent with the observed instability threshold of around 40 repeats in DM1 (Fu et al. 1992, Brook et al. 1992).



Models **N=142 individuals**	**Number of parameters**	**Maximised log-likelihood**	**AIC**	**Adjusted AIC**	**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, $a^g=40, \lambda^g, \mu^{[i]}, n_0^{[i]}$	286	-136,392	273,356	1,274 (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	1,930 (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	1,996 (5)	(4)
4. Expansion and contraction with a global parameter for contraction, an individual-specific parameter for expansion and a fixed threshold, $a^g=40, \lambda^{[i]}, \mu^g, n_0^{[i]}$	286	-139,852	280,276	8,194 (6)	(6)
3. Expansion and contraction with global parameters for expansion and contraction, and a fixed threshold, $a^g=40$, $\lambda^g, \mu^g, n_0^{[i]}$	145	-140,807	281,904	9,822 (7)	(7)
1. Expansion only with a global parameter for expansion and a fixed threshold, $a^g=40, \lambda^g, n_0^{[i]}$	144	-187,051	374,390	102,308 (8)	(8)

Table 4.2: **Model comparison summary for myotonic dystrophy type 1.** The mathematical models, listed in column 1, were ranked according to their AIC score which penalises the maximised log-likelihood by the number of parameters. Adjusted AIC (column 5) was obtained by subtracting the lowest value overall (272,082 Model M_{6a}) from the value for each model (column 4). *Significantly ($P < 10^{-15}$) better than Model M_{2a}. The models were also compared pairwise using a likelihood ratio test and ranked on this basis to provide a summary comparison to AIC (column 6).

4.3.3 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 Chapter 2, 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 (Chapter 2, Section 2.7).

The parameters ($\lambda^{[i]}$, $\mu^{[i]}$, $a^{[i]}$ and $n_0^{[i]}$ where a particular individual is denoted i for $i = 1, \ldots, 142$ 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 4.1). This approach provided not only the most probable value for each parameter but also a credible range. In some cases, there is evidence of solutions at local maxima. 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 sub-optimal 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 whilst capturing any uncertainty in the prediction. The results for individual CR35 ($i = 35$), Figure 4.1, 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 $\mu^{[35]}$ (Figure 4.1E). The peak is located at 0.25 contractions per CTG unit per year. For parameters $n_0^{[35]}$ and $\phi^{[35]}$ (Figures 4.1A and 4.1I) 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 1,300 CTGs), the posterior distribution for $a^{[35]}$ is best viewed jointly with n_0 (Figure 4.1F). The resulting contour is widely spread over the range for $a^{[35]}$ (0 to 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 (anti-correlation) concerning the best fit, as illustrated by the contour (Figure 4.1G).

The parameter values associated with the maximum likelihood for each DM1 individual form part of the supplementary files of (Higham et al. 2012), Supplementary Figure 3, available for viewing at `http://hmg.oxfordjournals.org/content/21/11/2450/suppl/DC1`. The av-

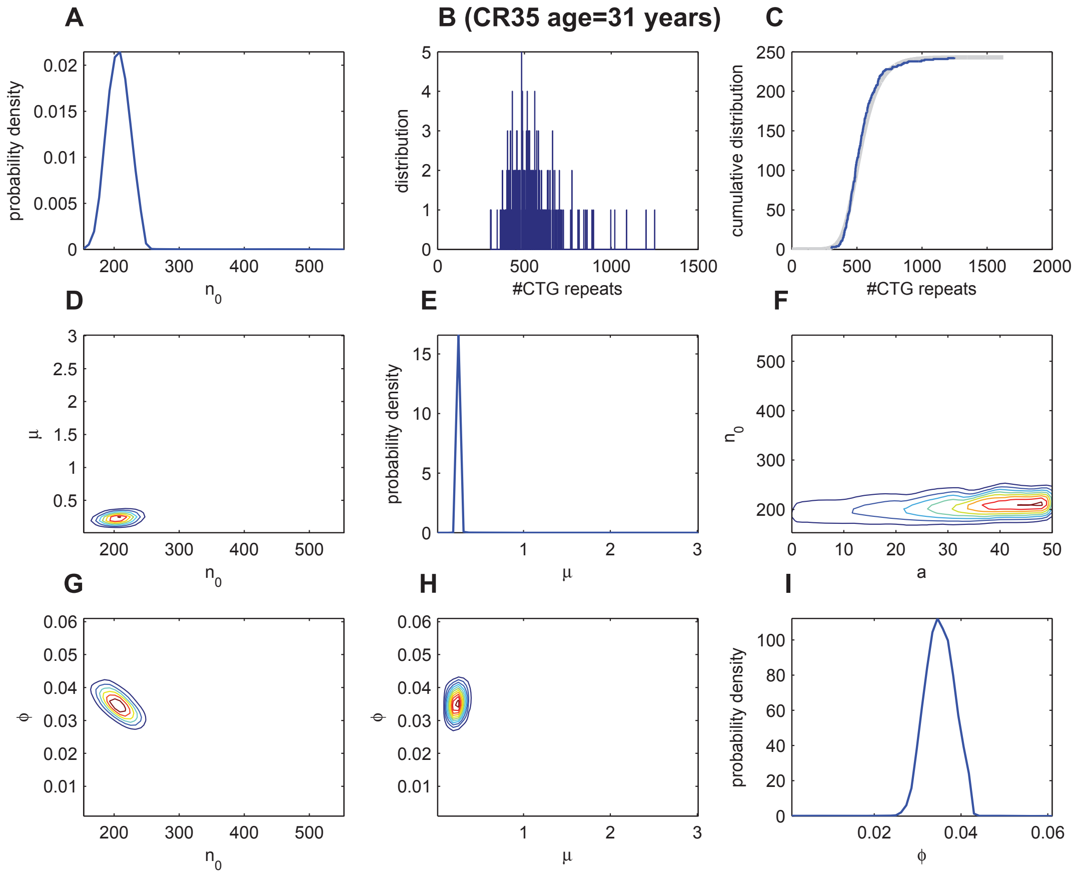


Figure 4.1: **Parameter estimation results for representative individual CR35, aged 30.** The data is presented in panel B as a histogram showing the distribution of repeat lengths for individual CR35 ($i = 35$). The posterior probability density distributions for parameters $n_0^{[35]}$, the inherited repeat length, $\mu^{[35]}$, the rate of contraction per CTG unit per year, $\phi^{[35]}$, the rate of expansion minus contraction per CTG unit per year, marginalised for each parameter over the other parameters, are shown in panels A, E and I, respectively. Marginalised joint probability distributions for parameter pairs, $\mu^{[35]}$ and $n_0^{[35]}$, $\phi^{[35]}$ and $n_0^{[35]}$, $\phi^{[35]}$ and $\mu^{[35]}$, and $n_0^{[35]}$ and $a^{[35]}$, the threshold number of repeats over which expansion and contraction occur, are shown, in panels 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 panel C, the data, shown as a cumulative distribution (jagged dark line) is compared to the inferred fit with the maximum likelihood value (light line) with associated parameter values $\mu^{[35]} = 0.25$, $\phi^{[35]} = 0.0358$, $a^{[35]} = 41$ and $n_0^{[35]} = 209$.

erage 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. Hence, 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 (correlation coefficient > 0.99).

4.3.4 Model fit

Models M_{6a} and M_{6b} fitted the data equally well but as model M_{6b} is more general concerning the threshold, we consider further the fit of model M_{6b} (expansion and contraction with individual parameters) to the data. For representative individual CR35, the maximum likelihood solution ($\mu = 0.25$, $\phi = 0.036$, $a = 41$ and $n_0 = 209$) traces closely the rising slope of the cumulative data (Figure 4.1 C) and the inferred value of μ is non-zero under the expansion and contraction model. Further to this, the maximum log-likelihood of the expansion and contraction model (-1,495) is greater than the maximum log-likelihood of the expansion only model (-1,511). 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 were simulated for 120 cells with an initial repeat length of 160 CTGs over 30 years, see Supplementary Videos 1 and 2 in (Higham et al. 2012) at `http://hmg.oxfordjournals.org/content/21/11/2450/suppl/DC1`. 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 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, see Appendix 2.

The full model M_{6b} assumed that the rates of expansion and contraction are linearly proportional to 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 individual data sets 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, see Supplementary Data, Higham et al. 2012. This excludes congential cases where repeat length is very high and asymptomatic individuals whose repeat lengths are relatively low. For low-range individuals (allele lengths less than 200 CTGs), contraction rates cluster around the low end of the parameter spectrum (Figure 4.2A). For high-range individuals (allele lengths greater than 800 CTGs), expansion minus contraction values cluster around the low end of the spectrum (Figure 4.2B). In both cases, having accounted for repeat length and age, 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. Data with which to do this is now available. In Chapter 5, we investigate small alleles through analysis of buccal cell DNA from HD individuals who, as discussed in Chapter 1, invariably have alleles in the low range as their inherited repeat lengths are much lower; typically between 40 and 50 CAGs. Over time, alleles generally expand and, in some individuals, change from the mid-range to the high-range. Hence a second blood DNA sample taken at a later point in time increases the number of long alleles in our study. We revisit this topic in Chapter 7.

4.3.5 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 summarised as the *prior* information (Table 4.3).

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 second, the subset of individuals who do not have the congenital form of the disease but do have symptoms

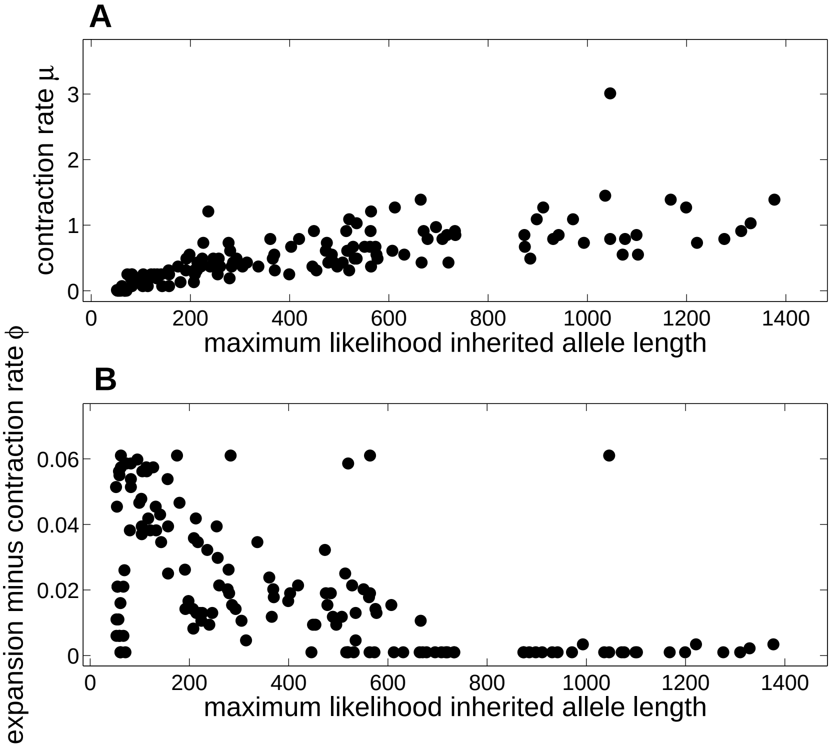


Figure 4.2: **Scatter plot of the maximum likelihood parameter values for model M_{6b}.** A. Contraction, rate per CTG unit per year, μ, on the vertical axis versus inherited allele length n_0 on the horizontal axis ($N = 142$). B. Expansion minus contraction, rate per CTG unit per year, $\phi = \lambda - \mu$.

($N = 121$). 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 (Figure 4.3A). 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 ($N = 121$) (Figure 4.3B). The credible interval (5-95 percentile) for this prediction is shown as a shaded grey area (Figure 4.3). All distributions are skewed towards the right with long tails. The lower rates when individuals with very short and very long alleles are included is another indication that there may be length effects unaccounted for in the model.

4.4 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

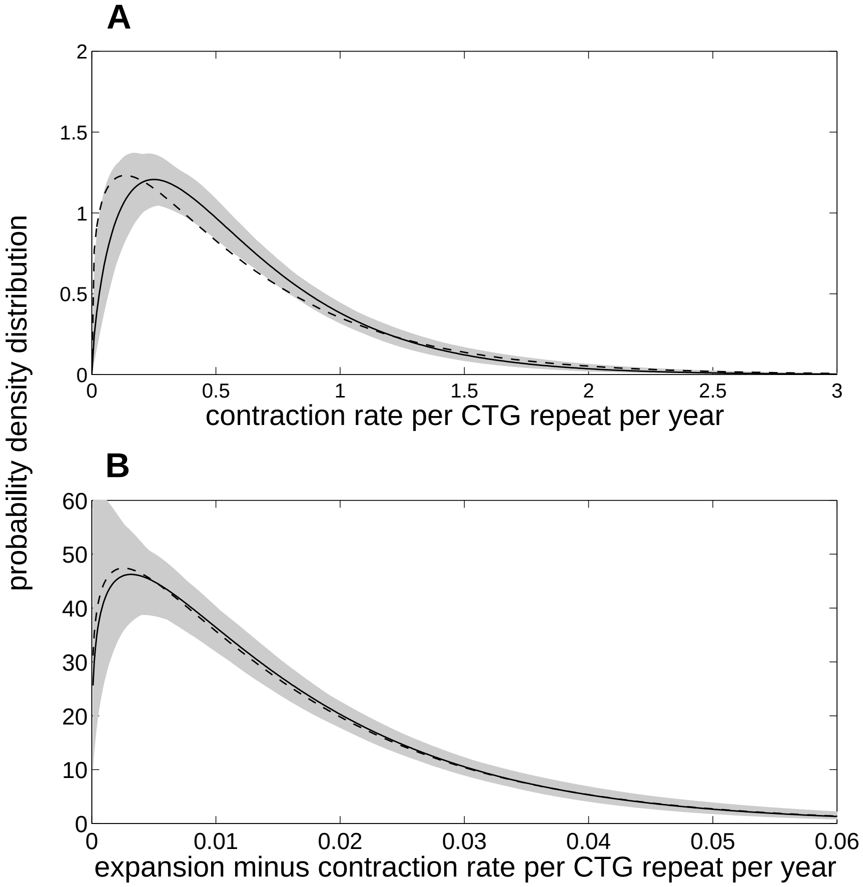


Figure 4.3: **Hierarchical Bayesian analysis results.** Panel A shows the modal distribution of the contraction rate (dark line) for all individuals except those who have had DM since birth (congenital) or who have no symptoms yet (asymptomatic), 121 individuals in total. Panel B shows the modal distribution 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 dashed line.

the blood cells of individuals with myotonic dystrophy type 1. 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 and discussed in Chapter 1, Section 1.4.1, in the larger context of models of microsatellite evolution. Leeflang *et al.* investigated germline mutation frequency in HD using a simple Okazaki fragment processing model of trinucleotide repeat instability (Leeflang et al. 1999). 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 hematopoietic stem cell population with hundreds of mutations predicted each year (see below) relative to a stem cell renewal rate of once every 40 weeks (Catlin et al. 2011). 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 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 (Veytsman & Akhmadeyeva 2006). 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.* who used a simple birth process to describe repeat length evolution (Kaplan et al. 2007). Because their data was 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. By contrast, for each DM1 individual, the data that we use in our study effectively provides 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 is whether the variation observed in these repeat lengths is solely due to expansion, as implicitly assumed in the model of Kaplan *et al.*, 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. Assuming that mutational gains and/or losses are mostly of one repeat unit, 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 hyper-mutational than we thought and is a provocative hypotheses 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. Whilst 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 (Fortune et al. 2000, Kennedy & Shelbourne 2000, Seznec et al. 2000).

The expansion and contraction rates are assumed to be constant with age. With one sample for each individual, 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 can assume that other effects are constant and quantify temporal changes. Collection of further samples is currently under way in a longitudinal study and we address these issues in Chapter 7.

For a thirty year old individual with an inherited repeat length of 200 and a net gain of two repeats per 100 expansions, the model predicts about 5,500 expansion and contraction events per cell during their life time, which is about 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 (Catlin et al. 2011). 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 base pairs of the genome, discussed in (Kunkel 1999) and (Lindahl 1993), 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 (Manley et al. 1999, van den Broek et al. 2002), 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, see Chapter 3. As such, suppression of somatic expansion is expected to be therapeutically beneficial and induction of contractions potentially curative (Gomes-Pereira & Monckton 2006, Castel et al. 2010). 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 toward expansion, net gains are the product of a very subtle bias toward expansions relative to almost equally frequent contractions. The high underlying frequency of contractions suggest therefore that a therapeutically beneficial impact may be mediated by a relatively subtle shift in the relative bias from small expansions toward 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 expansion and 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 and genetic or both, may influence instability.

Our model could also be extended to other triplet repeat expansion diseases depending on the availability of suitable datasets and we do this in Chapter 6 for Huntington disease. However, compared to DM1 the expanded repeat tract in most other triplet repeat diseases is relatively stable, particularly in blood. Other tissues such as 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 provide a deeper understanding of the underlying mechanisms and we can use a calibrated model to simulate scenarios and to make predictions. In Chapter 3, we 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 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 we investigate this issue in Chapter 7.

Further, these quantitative traits, μ and ϕ, are potential biomarkers that can be used via GWAS (genome-wide association study) to identify *trans*-acting genetic factors thought to be linked to this somatic variation, see Chapter 3. This is a future direction for our work and is discussed in our concluding remarks, Chapter 8.

4.5 Materials and methods

4.5.1 Project data

The dataset analysed in this study and described in Chapter 3 was 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 (Gibbs et al. 1993, Monckton et al. 1997) 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.

4.5.2 Other techniques

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, see Chapter 2, Equation (2.9), for reasons to do with the stiffness of the ODE 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 (Wedderburn 1974) where the relationship between mean and 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 (2.6) and (2.7) respectively, gives an expression for variance in terms of the mean adjusted for the threshold, a denoted by $\widehat{M}$:

$$\begin{aligned} \widehat{M} &= M - a, & (4.1) \\ V &= \left(\frac{\lambda+\mu}{\lambda-\mu}\right)\left(\frac{\widehat{M}^2}{n_0} - \widehat{M}\right). & (4.2) \end{aligned}$$

The equation for the variance is now a quadratic in $\widehat{M}$ 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 $\widehat{M}$ 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 $\widehat{M}$ and V:

$$\begin{aligned} p &= 1 - \frac{\widehat{M}}{V}, & (4.3) \\ r &= \frac{\widehat{M}^2}{V - \widehat{M}}. & (4.4) \end{aligned}$$

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 utilise all our data when evaluating the parameter space. Simulations with a range of individuals shows 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 over dispersion, which applies in our case as the variance exceeds the mean (Ver Hoef & Boveng 2007).

The corresponding likelihood, LNB, is

$$LNB = \prod_i \prod_j \frac{\Gamma\left(d_j^{[i]} + r^{[i]} - 1\right)}{\Gamma\left(d_j^{[i]}\right)\Gamma\left(r^{[i]} - 1\right)} \left(1 - p^{[i]}\right)^{r^{[i]}} p^{[i]^{d_j^{[i]}}}. \qquad (4.5)$$

Parameter combinations with a log likelihood value that satisfy the condition

$$\log\left(LNB\right) - \log\left(\max LNB\right) > \kappa, \tag{4.6}$$

were then subjected to the full likelihood computation. κ was chosen (typically $\kappa = -2$) to obtain computationally manageable sample sizes (about 50,000). Parameter combinations arising under these conditions generally form a cloud of values, close in distance, and are not expected to give rise to discontinuities in the likelihood at the transition.

Distribution	**Hyper parameters**	**Range for uniform prior**	**Incremental step size for parameter exploration**
$\mathit{\Gamma}_\mu \ (\alpha_\mu, \beta_\mu)$	mean $\alpha_\mu \beta_\mu$	0.3 to 0.8	0.01
	variance $\alpha_\mu \beta_\mu^2$	0.05 to 0.55	0.01
$\mathit{\Gamma}_\varphi \ (\alpha_\varphi, \beta_\varphi)$	mean $\alpha_\varphi \beta_\varphi$	0.005 to 0.03	0.0005
	variance $\alpha_\varphi \beta_\varphi^2$	0.0001 to 0.0006	0.00001

Table 4.3: **Hierarchical Bayesian analysis.** 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 repeat 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 we cover a range of possible shapes for this distribution. For ϕ, the rate of expansion minus contraction per CTG repeat per year, we also assume that the underlying distribution is a gamma distribution, Γ_ϕ, defined by a shape parameter $\alpha_\phi\beta_\phi$ and $\alpha_\phi\beta_\phi^2$.

Chapter 5

A length-specific effect is associated with less somatic instability in myotonic dystrophy type 1 and Huntington disease

5.1 Abstract

Over 20 genetic diseases, including myotonic dystrophy type 1 (DM1) and Huntington disease (HD) are caused by inheriting an unstable expanded simple sequence repeat. Repeat lengths at the lower end of the disease causing range are associated with less somatic instability, less severe symptoms and later disease onset. It was initially assumed that the repeat lengths in DM1 were pure CTG tracts, but interruptions within the repeat lengths have recently been estimated to exist in around 5% of DM1 individuals. Some of these interruptions are associated with less instability and less severe phenotypes in DM1. We have developed a mathematical model that describes changes over time in repeat length distributions in DM1 blood. This model has been fitted to sized repeat lengths from a large cohort of DM1 affected or at risk individuals with inherited repeat lengths, ranging between 50 CTGs and 1,500 CTGs, and explains differences, in repeat length distributions, across this varied group of DM1 individuals. However the mutation rates estimated under the model are lower than expected among individuals with inherited repeat lengths less than 100 CTGs, suggesting that these rates may be suppressed at the lower end of the disease causing range. We propose that a length-specific effect may be operating within this range and test this hypothesis by introducing such an effect into the model. For data, to calibrate the model, we use blood DNA from DM1 individuals with small alleles (inherited repeat lengths less than 100 CTGs) and buccal DNA from HD individuals who almost always have inherited repeat lengths less than 100 CAGs. These datasets comprise single DNA molecules sized using small-pool PCR. We find statistical support for a general length-specific effect which suppresses mutational rates among the smaller alleles giving rise to a distinctive pattern in the repeat length distributions. In a novel application of the new model, fitted to a large cohort of DM1 individuals, we also show that this distinctive pattern may help identify individuals whose effective repeat length, with regards to somatic instability, is less than their actual repeat length. A plausible explanation for this distinction is that the expanded repeat tract is compromised by interruptions or other unusual features. For these individuals, we estimate the effective repeat length of their expanded repeat tracts and contribute to the on-going discussion about the effect of interruptions on phenotype.

5.2 Introduction

In Chapter 4, we developed a mathematical model that describes changes over time in repeat length distributions in DM1 blood DNA. This model was fitted to sized repeat lengths from a large cohort of DM1 affected or at risk individuals whose estimated inherited repeat lengths range between 50 CTGs and 1,500 CTGs. This model was shown to explain satisfactorily the variable distributions of repeat lengths seen across this group of DM1 individuals. However, we observed that the variance to mean ratios of the repeat length distribution among individuals with inherited repeat lengths below 100 CTGs were very low, especially when age is also taken into account. To visualise this effect, Figure 5.1 shows the variance to mean ratios, highlighting those for the 14 individuals with estimated progenitor allele lengths below 100, along with the predicted level from model M_{6b}. In terms of the difference between the expected and the observed ratio, the fact that all 14 DM1 individuals with estimated repeat lengths less than 100 CTGs lie at the low end (15^{th} percentile) of this distribution is highly significant ($P < 10^{-5}$ using a permutation test). In addition, the estimated model parameters that quantify the rates of mutation were slightly biased, with individuals with the lowest estimated inherited repeat values having lower than expected rates of expansion and contraction (Figure 4.2). These results strongly suggest a length effect not accounted for in the model: an effect that results in proportionally less mutation within the small alleles. In this chapter, we account for a length-specific factor using first-principle mathematical modelling arguments (Section 5.5.2 and Figure 5.2) and compare the fit of this new model with the fit of the original model using the Akaike information criterion (AIC) (Akaike 1974). AIC is an appropriate choice when comparing un-nested models, as in our case. We fit this new model to blood DNA data from the 14 DM1 individuals with estimated inherited repeat lengths less than 100 CTGs (a subset of the DM1 individuals analysed in Chapter 4) and buccal DNA data from 12 HD individuals with estimated inherited repeat lengths between 39 and 48 CAGs (Veitch et al. 2007).

It was initially thought that the expanded DM1 repeats were pure, but interruptions within DM1 repeat lengths have been recently reported (Musova et al. 2009, Braida et al. 2010). These variant repeats (pure repeats containing interruptions) are now associated with less instability and less severe phenotypes in between 3 - 5% of DM1 individuals (Couto *et al.*, in preparation). As discussed in Chapter 3, inherited repeat length explains most of the variation in age of onset for many of the expanded repeat diseases and somatic instability is a candidate modifier of age of onset and disease progression. It is possible that some interruptions increase the stability of the pure repeat tract by reducing the effective length of an allele. An explanation for how variant repeats might modify

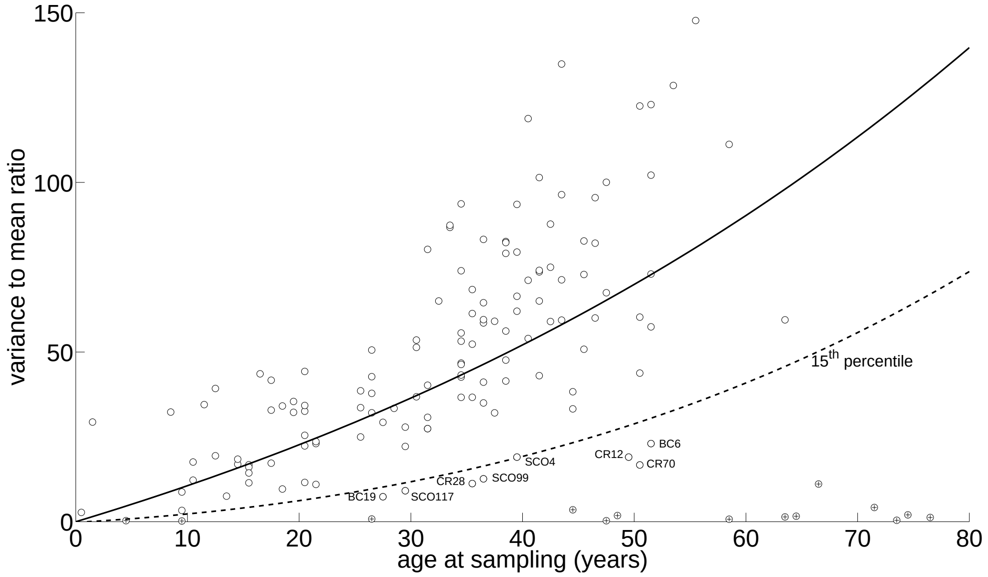


Figure 5.1: **Comparison of the actual variance over mean (adjusted for the threshold effect) by age (circles) among 142 DM1 individuals (blood DNA) and the predicted variance over mean path under model M_{6b} (line).** DM1 individuals with mean repeat lengths less than 100 are indicated by a cross within a circle. Some DM1 individuals are further identified by a code, see Table 5.3A for more details, and discussed in Section 5.3.6.

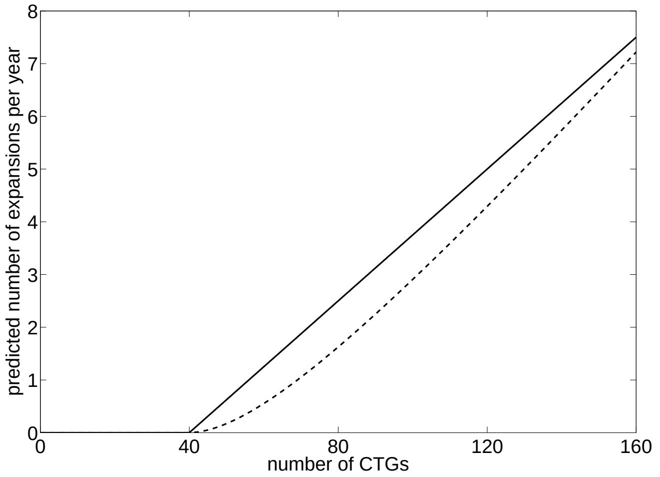


Figure 5.2: **Predicted number of expansions (of one CTG unit on the vertical axis) per year as a function of repeat length (number of CTGs on the horizontal axis) under model M_{6b} (dark straight line) and under model M_α with a length-specific effect (light curved line).**

mutational dynamics in an adjacent stretch of pure CTG repeats is outlined in (Braida et al. 2010) and illustrated in Figure 5.3. In summary, for pure expanded alleles the mutational dynamics of the CTG tract is driven by the action of a *cis*-acting modifier in the $3'$ flanking sequence. As discussed in (Braida et al. 2010) there is evidence that the content of flanking DNA has a role in repeat length stability, for example the GC content (Nestor & Monckton 2011). In the presence of variant repeats at the $3'$ end of the array, the distance between the pure CTG tract and the *cis*-acting modifier will be increased and its effect may be reduced. Alternatively, the presence of variant repeats may directly inhibit the mutability of the pure CTG tract. A novel application of our new model is to estimate the effective length, as well as the inherited length, underlying the repeat length distributions. We therefore fitted this new model to the rest of the DM1 datasets, 128 DM1 individuals with repeat lengths in the mid-to-high range, and compared the results with the original model.

The objectives of this Chapter are to develop and test a more sophisticated mathematical model based on plausible biological assumptions about small alleles. We use this model to establish whether mutational propensity is lower in the small alleles and whether reduced levels of somatic mosaicism give rise to distinctive repeat length distributions. We also introduce and quantify the concept of effective length, see Section 5.3.4. Individuals whose effective length is different from their actual length are strong candidates for further investigation, as lower than expected levels of somatic mosaicism may indicate the presence of modifiers of somatic instability such as interruptions in the repeat lengths.

5.3 Results

5.3.1 Formulation of an expansion and contraction model incorporating a length-specific effect

The framework we use to describe changes in repeat length (measured by the number of repeat units) over time in a population of cells is a stochastic birth and death process. In our context birth is the gain of one repeat unit (expansion) and death is the loss of one repeat unit (contraction) within a cell. This is a probabilistic model with probability functions defined for the mutational events of expansion and contraction. In Chapter 4, we assumed that the likelihood of a mutational event increased linearly with repeat length over a threshold number of repeats and it is this assumption and corresponding function definitions that we will refine in this chapter. Another key modelling assumption is that the cells acquire mutations independently of one another, and this was justified (see Chapter 2, Section 2.3) for DM1 blood cells. Application of the model to another disease and cell type, HD buccal cells, requires the assumption that buccal cells acquire mutations independently of one another. Buccal cells, like other external epithelium cells, are replenished from a large pool of self-renewing stem cells (Fuchs 2008) hence an assumption of independence is reasonable.

A biological explanation as to why small alleles may differ from long alleles, in terms of DNA

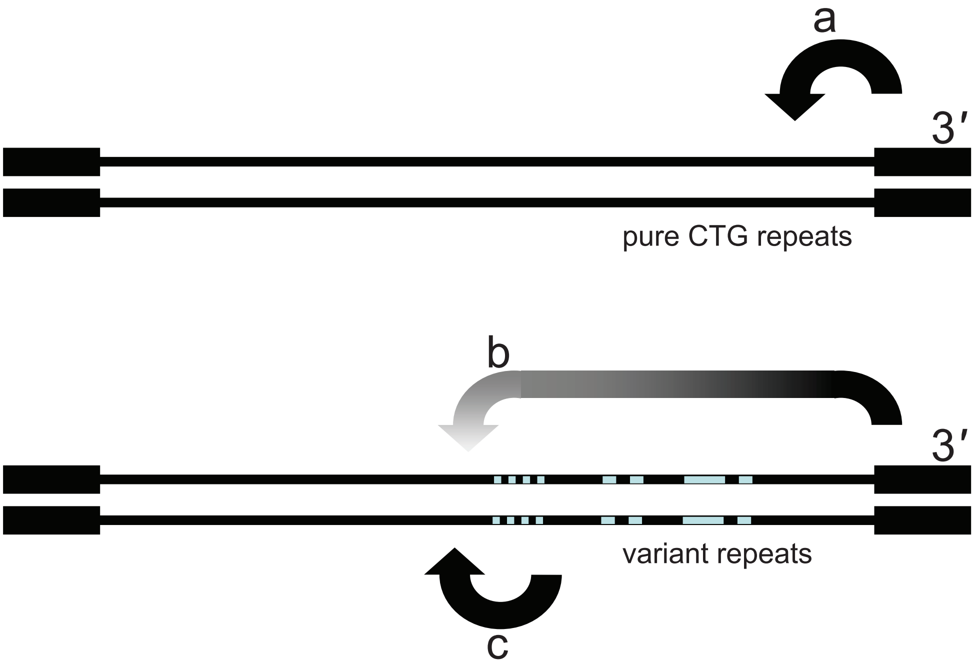


Figure 5.3: **Hypothetical explanation for how variant repeats might modify mutational dynamics in an adjacent stretch of pure CTGs.** For a pure expanded allele (top) the mutational dynamics of the CTG tract is driven by the action of a *cis*-acting modifier in the 3′ flanking sequence (a). For a CTG tract containing variant repeats at the 3′ end of the array (bottom), the distance between the pure CTG tract and the *cis*-acting modifier is increased and its effect reduced (b). Alternatively, the presence of variant repeats may directly inhibit the mutability of the pure CTG tract (c). This figure was adapted from Figure 5A in (Braida et al. 2010).

stability, can be based on the physical structure that these repeat lengths assume during the cell processes of transcription, repair, replication and recombination (Pearson & Sinden 1996). Simple repeat sequences in DNA are prone to adopt slipped strand structures comprising complementary loop-outs of one to three repeats on opposite strands (Pearson et al. 2005). One such working model for repeat instability (as illustrated in Chapter 1, Figure 1.2) is as follows. Sequences opposite the loop-outs can be excised and the gap filled by DNA mismatch repair proteins resulting in expansion. Alternatively sequences comprising the loop-outs can be excised resulting in contraction (Gomes-Pereira & Monckton 2006). Whilst loop-outs far apart may be repaired independently, it is possible that loop-outs that occur close to one another may be encompassed by the DNA repair domain and repaired together, effectively cancelling each other out and resulting in neither expansion nor contraction. The size of this domain is not known but, as discussed in Chapter 1, Section 1.3.2, may be between 60-230 base pairs of DNA (Genschel & Modrich 2003). If this is the case, then the likelihood that arbitrarily located loop-outs fall close to each other (and hence no mutation) would be higher in smaller alleles than in longer alleles, see Figure 5.4. Similarly, in support of relatively lower rates of mutation below 200 CTGs, Gellibolian *et al.* concluded from the biophysical examination of DNA mis-pairing in different CTG repeat lengths using plasmid DNA (Gellibolian et al. 1997) that for CTG repeat lengths up to 200 CTGs there is increasing mis-pairing per repeat unit (and hence increasingly more mutation) after which rates per repeat unit are constant. These investigations inform our hypothesis that a constraint on the mutational mechanism, whose effect decreases as repeat length increases, may operate at the lower end of the range of repeat lengths. We quantify this length-specific effect using a combinatorial counting method based on the length of the constraint (one interpretation being the distance between loop-outs), denoted as α, and the likelihood that mutation occurs. We introduce this effect into the probability functions for expansion and contraction, as described in Section 5.5.2.

We use statistical inference based on the available data to determine the value of α, along with the other model parameters (expansion per repeat unit per year, contraction per repeat unit per year and the inherited repeat length), and to determine whether α varies between individuals. Our approach, to quantify a length-specific effect and to determine the range over which it operates, is sufficiently general to incorporate other distance requirements, not just the distance between loop-outs, that might suppress mutation proportional to length. Such a distance requirement might be created by an interruption in the repeat length. Alternatively, α may be interpreted as the length of a DNA fragment typically processed by the DNA repair mechanism and/or DNA replication machinery.

5.3.2 Model comparison supports a role for a length-specific effect suppressing mutational rates in DM1

We introduced a length-specific effect, $R_n\left(n, \alpha\right)$, which is a function of repeat length, n, and the distance constraint, α, into model M_{6b}, as described in Section 5.5.2. We tested our hypothesis that this extended model, denoted M_α, would provide a better explanation for the mutational dynamics within the smaller alleles by fitting to sized blood DNA repeat length distributions from 14 DM1 individuals with repeat lengths at the lower end of the DM1 range. The relative goodness of fit of the original model M_{6b} and new model M_α was assessed using the Akaike information criterion (AIC) (Akaike 1974). The models have the same number of parameters and are not nested, so AIC is an appropriate method to rank the models through a relative measure of the goodness of fit. Application of AIC involved calculating the maximum likelihood value using a grid search over the parameter space, as outlined in Table 5.1 (see Chapter 2, Section 2.5, for further details). Model M_α

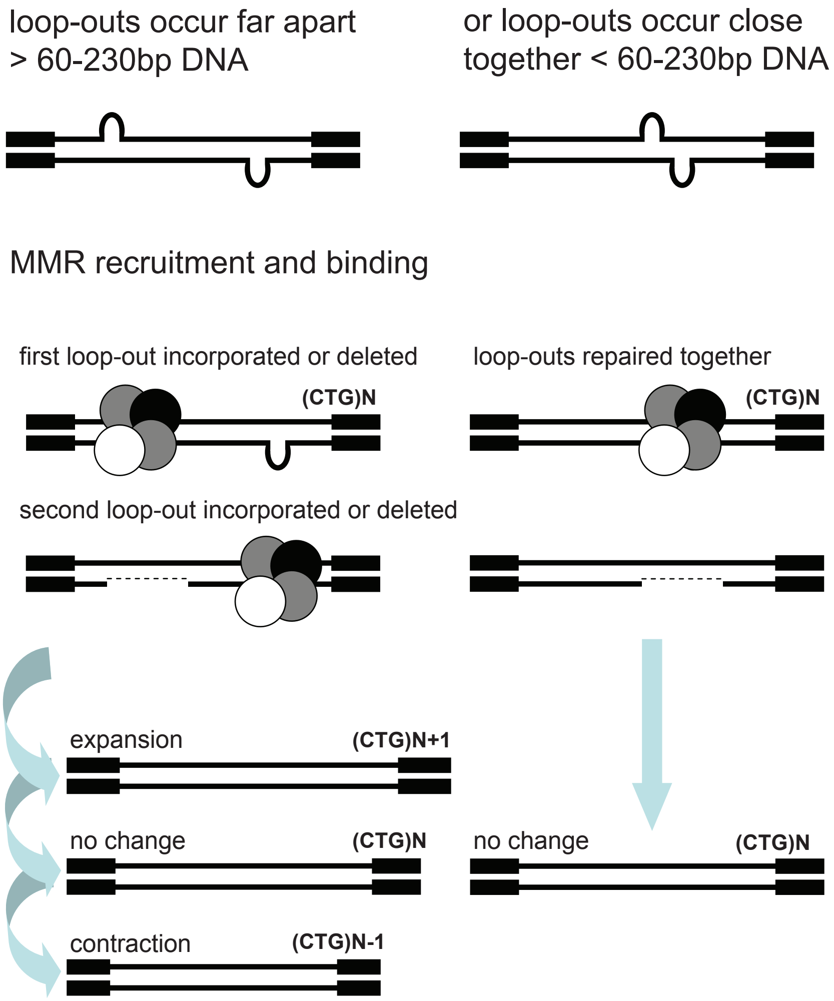

loop-outs occur far apart
> 60-230bp DNA
or loop-outs occur close
together < 60-230bp DNA
MMR recruitment and binding
first loop-out incorporated or deleted
(CTG)N
loop-outs repaired together
(CTG)N
second loop-out incorporated or deleted
expansion
(CTG)N+1
no change
(CTG)N
contraction
(CTG)N-1
no change
(CTG)N

Figure 5.4: **Hypothetical explanation for how a length-specific effect may modify mutational rates.** Loop-outs occurring far apart ($>$ 60 - 230 base pairs DNA) are repaired independently. This results in either expansion, contraction or no change in the repeat length (left column). Loop-outs occurring close together ($<$ 60 - 230 base pairs DNA) are repaired together resulting in no change in the repeat length (right column).

(maximum likelihood value = -4,779 and AIC = 9,670) ranks higher than model M_{6b} (maximum likelihood value = -4,805 and AIC = 9,721), see Table 5.2A. The difference in AIC values of 51 indicates that the relative likelihood (relative likelihood = 6.90×10^{-12}) of model M_{6b} compared with M_α is very low and so we conclude that model M_α fits the data better than model M_{6b} among individuals with repeat lengths at the lower end of the DM1 range. The model fit can be visualised as a distribution curve or a cumulative distribution curve. The fits of models M_{6b} and M_α compared with the data and each other are shown in Figures 5.5 and 5.6 for representative DM1 individuals.

Model M_α is seen to be better than model M_{6b} at tracking the initially steep ascent of the cumulative distribution typical of the distributions among these individuals.

A

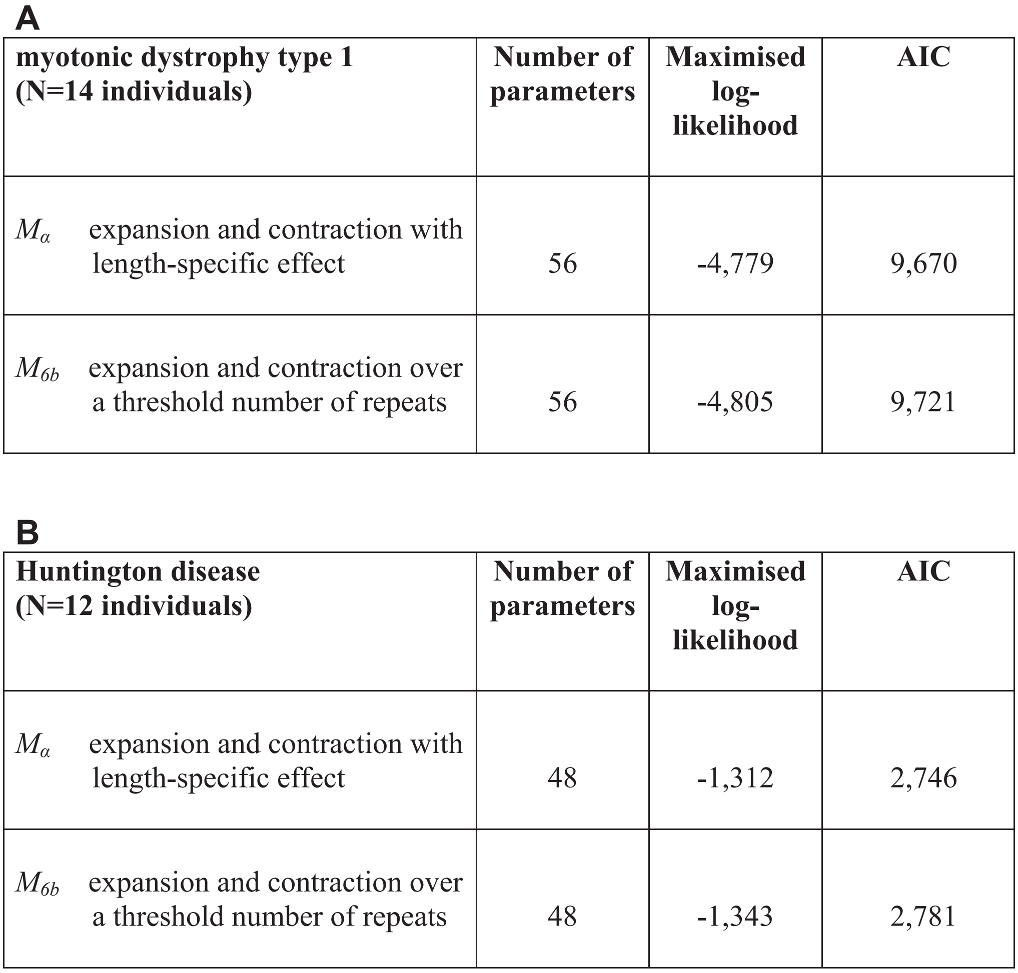
Model M_{6b} parameters	**Prior ranges DM1 blood**	**Prior ranges HD buccal**
Contraction rate per repeat unit per year (μ)	0 to 1.2	0 to 0.005
Net expansion rate per repeat unit per year (λ - μ)	0.001 to 0.5	0.0001 to 0.1
Threshold number of repeat units (a)	0 to 50	0 to 40
Inherited repeat length, number of repeat units (n_0)	51 to 100	38 to 50

B

Model M_α parameters	**Prior ranges DM1 blood**	**Prior ranges HD buccal**
Contraction rate per repeat unit per year (μ)	0 to 1.2	0 to 0.01
Net expansion rate per repeat unit per year (λ - μ)	0.001 to 0.5	0.0001 to 0.1
Length parameter number of repeat units (α)	0 to 200	0 to 40
Inherited repeat length, number of repeat units (n_0)	51 to 100	38 to 50

Table 5.1: **Prior ranges for parameter estimation for small alleles.** For model M_{6b} (A) and for model M_α (B).

The average among these 14 DM1 individuals of the maximum likelihood value of α was 51 CTGs, but there was considerable variation (standard deviation = 22 CTGs). This result places α within the DNA repair domain 60-230 bp suggested by (Genschel & Modrich 2003) and thus is consistent with a hypothesis implicating inappropriate DNA repair, as outlined in Figure 5.4. With a fixed length parameter, α = 51 CTGs, we estimate that the length-specific effect would be strongest between 51 CTGs and 173 CTGs (Figure 5.7). These results provide support for a length-specific effect operating below 200 CTGs in DM1. By suppressing the mutation rate per repeat unit, the length-specific effect makes a big difference to the shape of the repeat length distribution below 200 CTGs but increasingly less difference over 200 CTGs.



A

myotonic dystrophy type 1 (N=14 individuals)	**Number of parameters**	**Maximised log-likelihood**	**AIC**
M_α expansion and contraction with length-specific effect	56	-4,779	9,670
M_{6b} expansion and contraction over a threshold number of repeats	56	-4,805	9,721

B

Huntington disease (N=12 individuals)	**Number of parameters**	**Maximised log-likelihood**	**AIC**
M_α expansion and contraction with length-specific effect	48	-1,312	2,746
M_{6b} expansion and contraction over a threshold number of repeats	48	-1,343	2,781

Table 5.2: **Model comparison summary.** The models, listed in column 1, were compared using AIC (column 4) for myotonic dystrophy type 1 (A) and for Huntington disease (B).

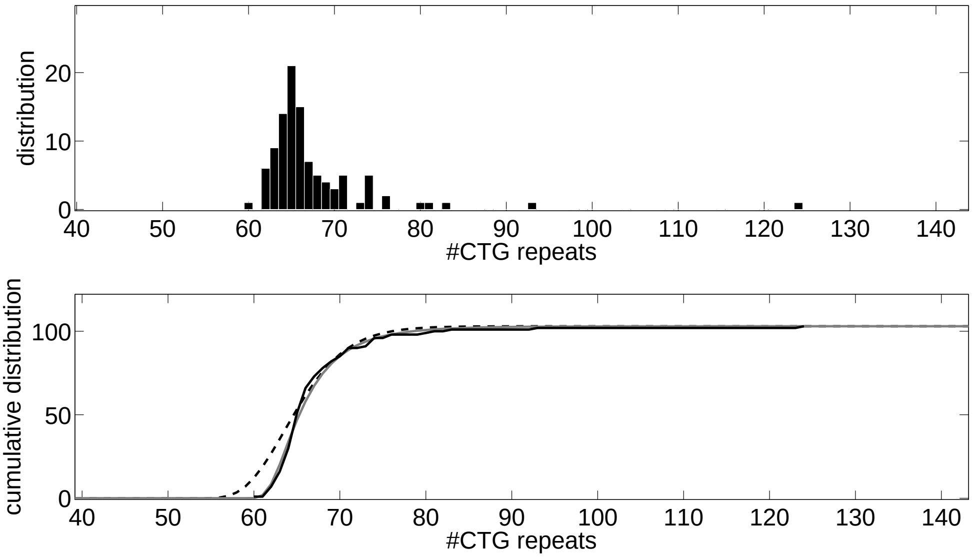


Figure 5.5: **Model fitting results for a representative DM1 individual CR8. Top:** Distribution of repeat lengths. **Bottom:** Model M_α fit (grey solid line) and model M_{6b} fit (black dashed line) compared with the cumulative distribution of repeat lengths (black jagged line).

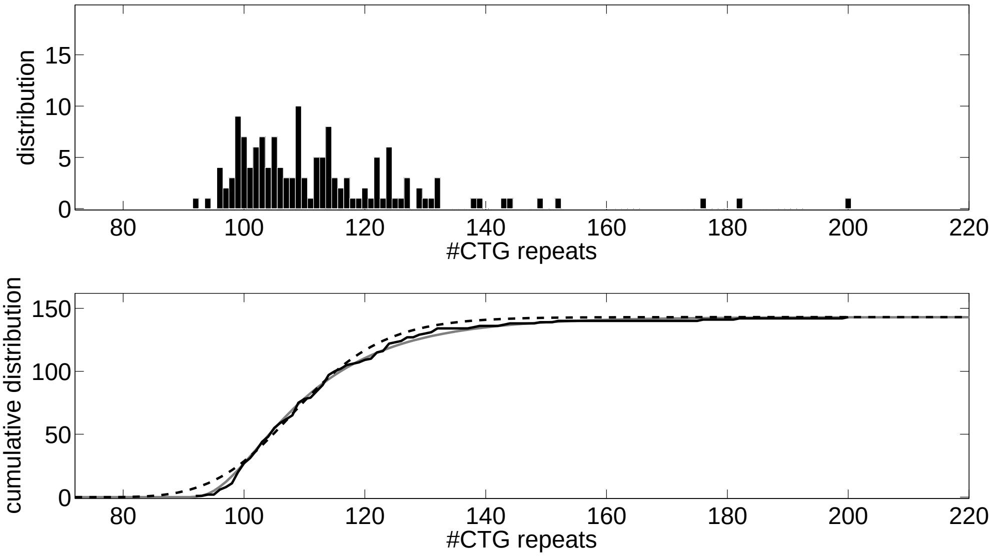


Figure 5.6: **Model fitting results for a representative DM1 individual CR27. Top:** Distribution of repeat lengths. **Bottom:** Model M_α fit (grey solid line) and model M_{6b} fit (black dashed line) compared with the cumulative distribution of repeat lengths (black jagged line).

5.3.3 Model comparison supports a role for a length-specific effect suppressing mutational rates in HD

For comparison, both models, M_{6b} and M_α, were fitted to sized distributions of buccal DNA single molecule repeat lengths from 12 unrelated HD individuals, all aged 39 years when the samples were

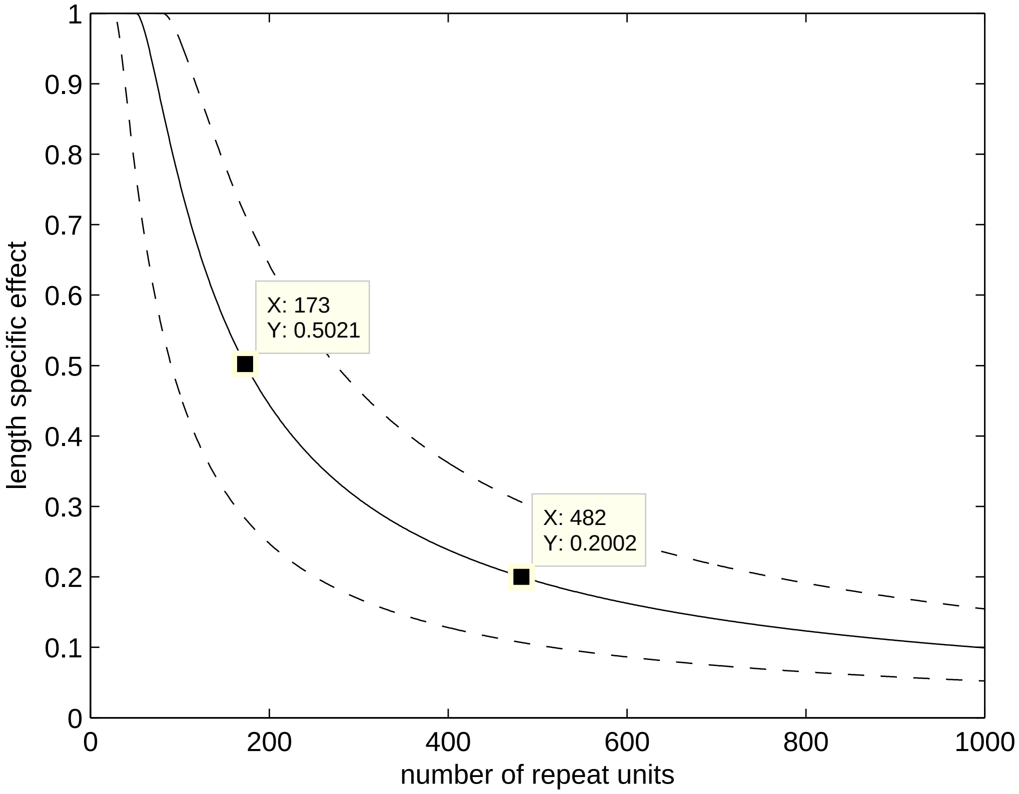


Figure 5.7: **Length-specific effect ($1-R_n$) for DM1 as a function of repeat length for an average α value 51 CTGs (solid line).** The 10^{th} and 90^{th} percentiles are indicated by the lower dashed line and upper dashed line, respectively.

taken, as collected and previously analysed by (Veitch et al. 2007). The prior ranges for Bayesian parameter estimation were chosen to represent HD buccal cells: expansion per CAG unit per year, contraction per CAG unit per year, and a threshold measured in CAG units (Table 5.1). The inherited number of CAG repeats, n_0, was treated as an unknown parameter and its value inferred from the data along with the other parameters. The maximum likelihood was calculated using a grid search over the parameter space, modified for buccal cells and HD rather than blood cells and DM1.

As for DM1, the results from AIC indicate that model M_α (maximum likelihood value = -1,312 and AIC = 2,746) ranks higher than model M_{6b} (maximum likelihood value = -1,343 and AIC = 2,781), see Table 5.2B. The difference in AIC values of 35 indicates that the relative likelihood (relative likelihood = 2.06×10^{-8}) of model M_{6b} compared with M_α is very low so we conclude that model M_α fits the buccal DNA data better than model M_{6b} among individuals with repeat lengths at the lower end of the range. As was the case for DM1, model M_α is better than model M_{6b} at tracking the initially steep ascent of the cumulative distribution, see Figure 5.8.

Among these 12 HD individuals, there were three individuals for whom the two models, M_{6b} and

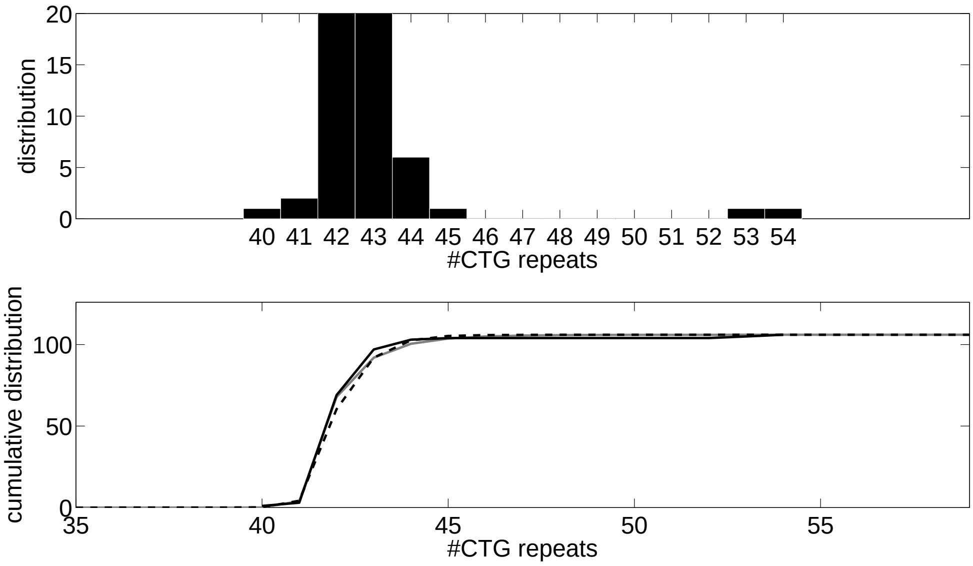


Figure 5.8: **Model fitting results for a representative HD individual HD10. Top:** Distribution of repeat lengths. **Bottom:** Model M_α fit (grey solid line) and model M_{6b} fit (black dashed line) compared with the cumulative distribution of repeat lengths (black jagged line).

M_α, were equally likely and the estimates for the fixed length parameter were below 3 CAGs. These three individuals had low levels of somatic mosaicism and hence it may not be possible to distinguish between the models and estimate the length-specific factor for this type of individual. Among the other 9 HD individuals the average value of α associated with the maximum likelihood value was 26 CAGs (standard deviation = 11 CAGs). These results provide support for a length-specific effect, suppressing the mutation rate per repeat unit, over the whole range of observed repeat lengths in this HD dataset (59 CAGs or less), see Figure 5.9.

We also considered a model with global parameters for the mutation rates and length effect and individual-specific parameters only for the inherited length (results not shown). However, as reported for DM1 in Chapter 4, global parameters did not capture the variation seen in the data, indicating that individual-specific factors play a major role in HD somatic instability. Inclusion of contraction events, *i.e.* decreases in repeat length of one CTG unit for DM1 or one CAG unit for HD, was also justified statistically, as there was no support for the contraction rates being zero.

5.3.4 Estimates of inherited repeat length under model M_α are in agreement with original predictions

In our study we treated inherited repeat length, n_0, as an unknown parameter to be inferred from the data. Our estimates of the value of n_0 are in agreement (correlation coefficient = 0.93) with

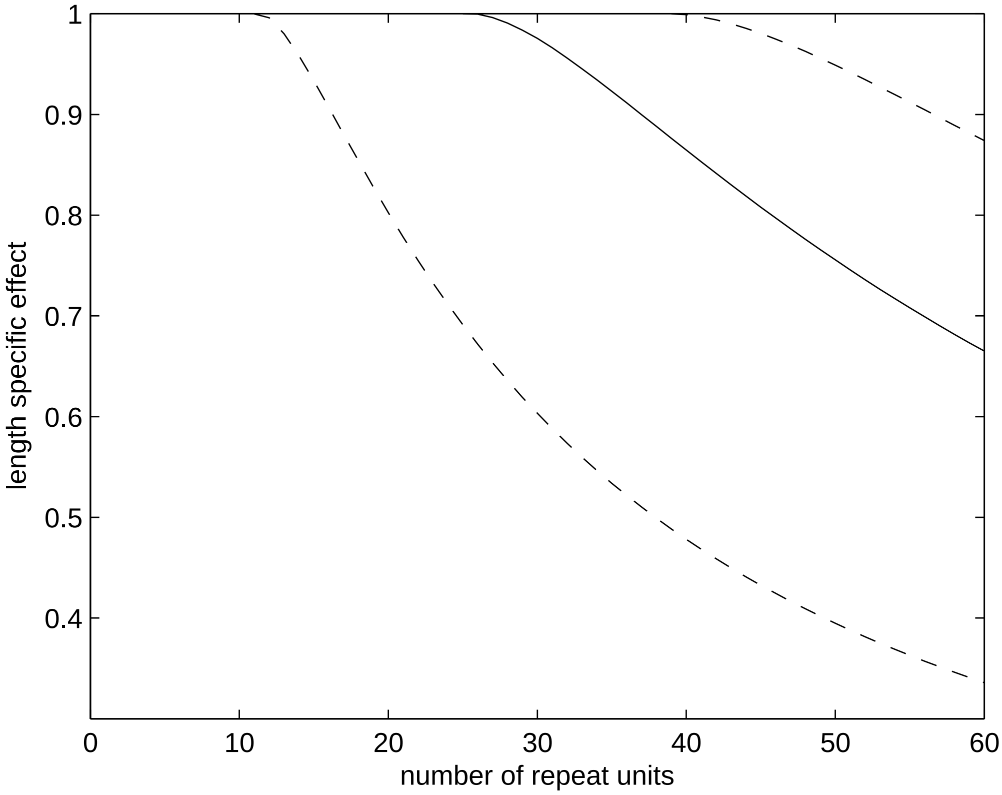


Figure 5.9: **Length-specific effect ($1 - R_n$) for HD as a function of repeat length for an average α value 26 CAGs (solid line).** The 10^{th} and 90^{th} percentiles are indicated by the lower dashed line and upper dashed line, respectively. Results are shown over the observed range of repeat lengths (less than 60 CAGs).

those estimated using the lower bound of the distribution as seen with small pool PCR, discussed in Chapter 3, Section 3.3.1. Further, our estimates of the value of n_0 are in complete agreement (correlation coefficient > 0.99) with the estimates by Veitch *et al.* which for the HD individuals in this study were based on the lower boundary of their highly skewed distributions (Veitch et al. 2007).

5.3.5 Mutational levels are higher in DM1 blood cells than in HD buccal cells, indicating differences in the overall level rather than the underlying mechanism

The parameter values associated with the maximum likelihood value provide a point estimate for mutation rates, in terms of expansion per repeat unit and contraction per repeat unit for each individual. Comparing parameter values under model M_α for DM1 blood with those for HD buccal, the median expansion rate for DM1 (9.1×10^{-2} per CTG per year) is significantly higher ($P = 8.28 \times 10^{-5}$ using the Mann Whitney U test) than for HD (8.5×10^{-4} per CAG per year).

Similarly the median contraction rate is significantly higher ($P = 1.32 \times 10^{-5}$ using the Mann Whitney U test) for DM1 (7.0×10^{-2} per CTG per year) than for HD (1.5×10^{-4} per CAG per year), see Figure 5.10 for comparison. The number of individuals is small but there appears to be a correlation between expansion and contraction rates within DM1 (correlation coefficient > 0.99) and within HD (correlation coefficient = 0.70) suggesting a link between expansion and contraction, within the mutational mechanism, in both diseases. Interestingly, the ratio of contraction to total mutation (expansion and contraction) is higher for DM1 (0.40) than for HD (0.18) and this most probably reflects biological differences between DM1 blood cells and HD buccal cells.

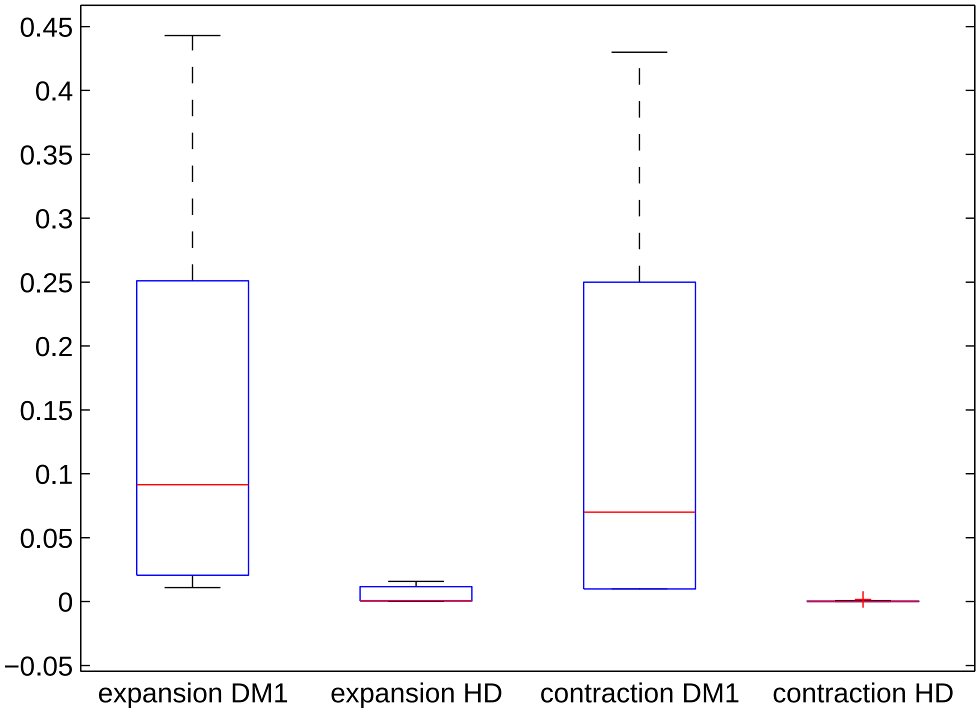


Figure 5.10: **Comparison of mutation rates between DM1 (14 individuals) and HD (12 individuals)**

5.3.6 For some DM1 individuals, effective length is lower than inherited repeat length which may have resulted from an interruption or another anomaly

We have shown that incorporating a length parameter, α, via a length-specific effect into model M_α better explains the distinctive distributions among DM1 individuals with smaller alleles than model M_{6b}. Variance to mean ratios of the repeat length distribution among individuals with inherited repeat lengths less than 100 CTGs were very low, especially when age is also taken into account.

In terms of the difference between the expected and the observed ratio, the fact that all 14 DM1 individuals lie at the low end (15^{th} percentile) of this distribution is highly significant ($P < 10^{-5}$ using a permutation test). We hypothesise that a length-specific effect may operate in other unusual individuals (particularly those with low variance to mean ratios) and that applying this new model to the data in order to infer a value for α may establish whether, and if so where, this effect operates. A further eight individuals with low variance to mean ratios (within the 15^{th} percentile) comparable to individuals with small alleles (Figure 5.1) are listed in Table 5.3A.

We now fit the new model, M_α, to distributions of repeat lengths (blood DNA) from the rest of the cohort (128 DM1 individuals with estimated inherited repeat length, n_0, over 100 CTGs including the eight DM1 individuals mentioned above). The prior range for α was chosen to be 0 to 200 CTGs (for individuals with inherited repeat length greater than 200 CTGs) and n_0 to 200 CTGs (for individuals with inherited repeat length less than 200 CTGs). Five of the 8 DM1 individuals, mentioned above, have estimates for α of 80 repeat units or more and an improvement in fit (log-likelihood gain of 2 or more). The improvement in fit can be seen by comparing the fit of both models to the data for representative individuals BC19 and SCO117 (Figures 5.11 and 5.12). As seen before (Figures 5.5 and 5.6) model M_α is better than model M_{6b} at capturing the steep rise at the beginning of the cumulative distribution. Other DM1 individuals, with estimated values for α over 100 CTGs, are listed in Table 5.3B. For most of these individuals there is an improvement in fit (likelihood gain of 2 or more).

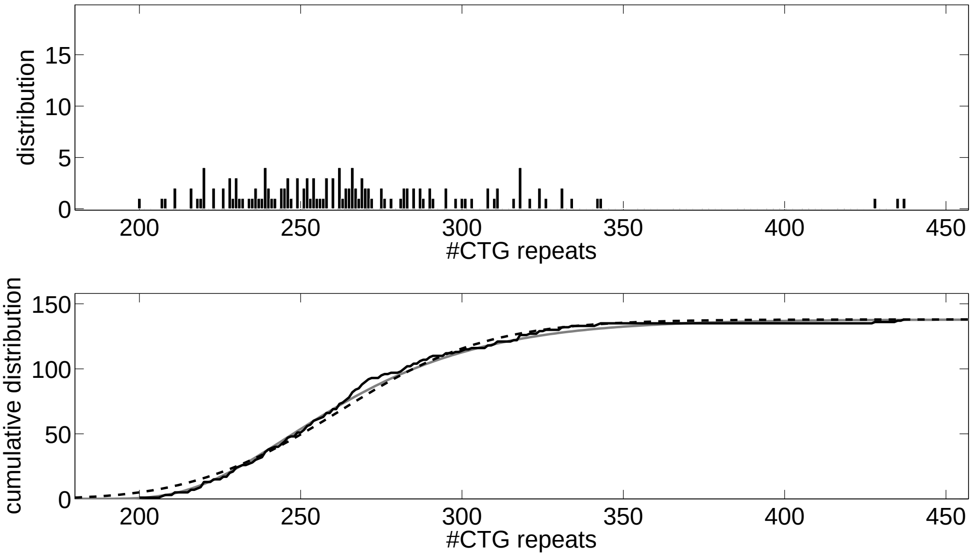
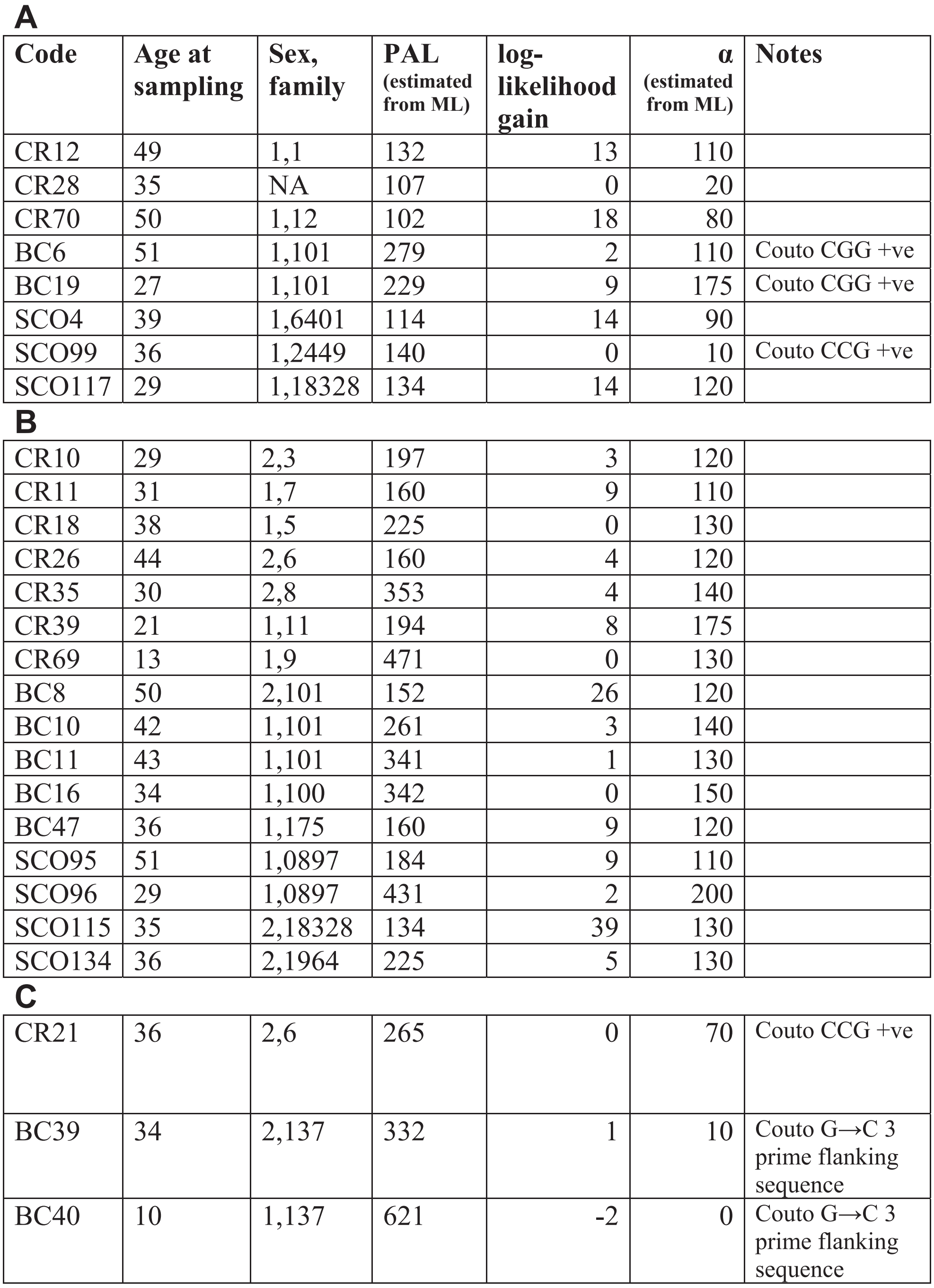


Figure 5.11: **Model fitting results for an unusual DM1 individual BC19. Top:** Distribution of repeat lengths. **Bottom:** Model M_α fit (grey solid line) and model M_{6b} fit (black dashed line) compared with the cumulative distribution of repeat lengths (black jagged line).



A

Code	Age at sampling	Sex, family	PAL (estimated from ML)	log-likelihood gain	α (estimated from ML)	Notes
CR12	49	1,1	132	13	110	
CR28	35	NA	107	0	20	
CR70	50	1,12	102	18	80	
BC6	51	1,101	279	2	110	Couto CGG +ve
BC19	27	1,101	229	9	175	Couto CGG +ve
SCO4	39	1,6401	114	14	90	
SCO99	36	1,2449	140	0	10	Couto CCG +ve
SCO117	29	1,18328	134	14	120	

B

CR10	29	2,3	197	3	120	
CR11	31	1,7	160	9	110	
CR18	38	1,5	225	0	130	
CR26	44	2,6	160	4	120	
CR35	30	2,8	353	4	140	
CR39	21	1,11	194	8	175	
CR69	13	1,9	471	0	130	
BC8	50	2,101	152	26	120	
BC10	42	1,101	261	3	140	
BC11	43	1,101	341	1	130	
BC16	34	1,100	342	0	150	
BC47	36	1,175	160	9	120	
SCO95	51	1,0897	184	9	110	
SCO96	29	1,0897	431	2	200	
SCO115	35	2,18328	134	39	130	
SCO134	36	2,1964	225	5	130	

C

CR21	36	2,6	265	0	70	Couto CCG +ve
BC39	34	2,137	332	1	10	Couto G→C 3 prime flanking sequence
BC40	10	1,137	621	-2	0	Couto G→C 3 prime flanking sequence

Table 5.3: **DM1 individuals with unusual repeat length distributions.** Individuals with low variance mean ratio (15^{th} percentile) and inherited repeat length > 100 CTGs (A). Individuals with high estimated α value (B). Individuals with unusual features (C). Notes: (i) 1=male, 2=female; (ii) Couto CGG $+$ve - these individuals tested positive for CGG interruptions in their repeat lengths; (iii) Couto CCG $+$ve - these individuals tested positive for CCG interruptions in their repeat lengths; and (iv) Couto G $\rightarrow$ C - a C instead of a G was found in the $3'$ flanking sequence.

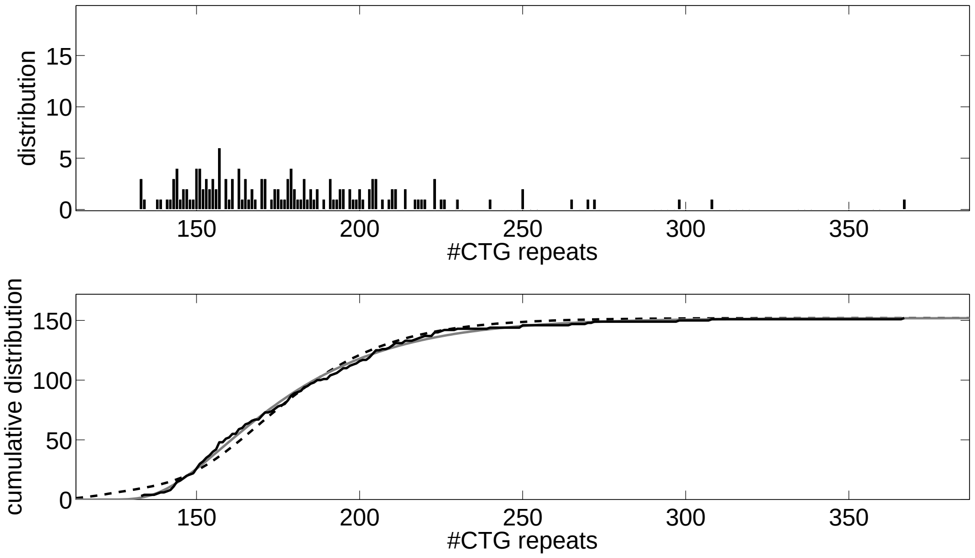


Figure 5.12: **Model fitting results for an unusual DM1 individual SCO117. Top:** Distribution of repeat lengths. **Bottom:** Model M_α fit (grey solid line) and model M_{6b} fit (black dashed line) compared with the cumulative distribution of repeat lengths (black jagged line).

In defining effective length, we use inherited repeat length as a point of reference as it is both the initial repeat length and the major modifier of age of onset. Hence we define effective length as the difference between inherited repeat and α, that is $n_0 - \alpha$. Consequently higher α values imply lower effective lengths. For example, BC19 (inherited repeat length equals 229 CTGs and α equals 175 CTGs) has an effective length of 54 CTGs. Interestingly, after a general screen for CGG and CCG interruptions in the expanded repeat lengths (Couto *et al.*, in preparation) BC19 tested positively for CGG interruptions and the father of BC19 (BC6), who has an estimated α value of 110 CTGs, also tested positive for CGG interruptions. BC19 and BC6 were previously noted for two reasons. First, they have unusually mild symptoms given their estimated inherited repeat lengths (see Figure 2 in ref. (Ashizawa et al. 1992), BC6=II.2 and BC19=III.2). Second, the germline transmission from father (BC6) to son (BC19) resulted in relatively rare apparent contractions (Ashizawa et al. 1992). Also of interest are two individuals, SCO99 and CR21, see Table 5.3A and Table 5.3C, respectively, who tested positive for CCG interruptions but did not have high values for α and two individuals who had changes in their DM1 flanking sequence, BC39 and BC40, but did not have high values for α, see Table 5.3B.

5.4 Discussion

In DM1, small inherited repeat lengths (less than 100 CTGs) are associated with late onset and less disease severity. Investigation of repeat length distributions in DM1 blood DNA among a large cohort of 145 individuals with a range of inherited repeat lengths (see Chapter 4) found that individuals with inherited repeat lengths less than 100 CTGs had very low variance to mean ratios especially when taking into account their advanced age when the blood DNA samples were taken (Figure 5.1). Correspondingly, when we estimated the rates of mutation per repeat unit per year for these 14 DM1 individuals, using the original model M_{6b}, the rates of expansion and contraction were relatively low, much lower than the rest of the cohort. Whilst this does not affect our ability to describe the changes in repeat length and hence the levels of somatic mosaicism over time, the implication that individuals with small inherited repeat lengths also have low rates of mutation does not have an obvious biological basis. It is more plausible that we did not take repeat length fully into account in our model and that there is a length effect unaccounted for. Gellibolian *et al.* concluded from the biophysical examination of DNA mis-pairing in different CTG repeat lengths, using plasmid DNA, that the number of mis-pairings per repeat unit is length dependent with relatively fewer mis-pairings per repeat unit (and hence less mutation) below 200 CTGs and reaching a constant rate over 200 CTGs (Gellibolian et al. 1997). This result supports less frequent mutation events per repeat unit in small alleles than in long alleles. To quantify this effect we introduced a length constraint into the expressions for expansion and contraction as an extension of the original mathematical model. This approach is sufficiently general to cover a wide range of possible constraints on the mutational mechanism that act by suppressing mutation rates per repeat unit proportionally less as alleles increase in size. The biological basis for mutational differences in length is very likely linked to the the mutational mechanism underlying repeat length changes. This mechanism is not fully understood, but it is thought that DNA mismatch repair plays an important role in the stability of trinucleotide repeats, see Chapter 1, Section 1.3.2 and Figure 1.2. Simple repeat sequences in DNA are prone to a slipped strand structure comprising complementary loop-outs of one to three repeats on opposite strands (Pearson et al. 2005). Loop-outs occurring during cell division would, normally, be recognised by DNA mismatch repair proteins (Wells et al. 2005) and be fixed accordingly. However it is possible that loop-outs arising in expanded repeats, independently of cell division, may be inappropriately repaired and, depending on the choice between incorporating or deleting the loop-out, become either expansions or contractions (Gomes-Pereira et al. 2004). The distance between loop-outs is one possible and highly plausible length constraint on the mutational mechanism but similarly acting constraints could depend on other features, such as the length of a DNA fragment typically processed by the DNA repair mechanism and/or DNA replication machinery. A hypothetical explanation for how a length constraint might modify mutational rates is illustrated in Figure 5.4.

In summary, we hypothesised that lower rates of mutation in small alleles are connected to the constraint that size enforces on the mutational mechanism. We extended our mathematical model to include a general, but biologically justified, length-specific effect and compared this new model, M_α, with our original model, M_{6b}, in terms of the goodness of fit. We found that under the new model, M_α, there was an improvement in the fit (Table 5.2A) which supports our conclusion that a length-specific effect acts over smaller alleles in DM1 blood.

Having fitted models M_{6b} and M_α to DM1 small alleles, these models were also adapted for HD and fitted to repeat length distributions from HD buccal cells (Veitch et al. 2007). Over all 12 HD buccal datasets, model M_α with its length-specific effect fitted better than the thresholded model, M_{6b} (Table 5.2B). This result suggests that there is also a constraint on the mutational mechanism in HD buccal cells. Estimated mutational rates, for both expansion and contraction, were significantly lower in HD buccal cells than in DM1 blood cells (Figure 5.10) and were more weighted towards expansion in HD buccal cells (82%) than in DM1 blood cells (60%). These differences have implications for the shape of the repeat length distributions (more skewed to higher repeat lengths in HD and more spread out in DM1) and hence levels of somatic mosaicism. The most likely explanation for these differences is linked to cell type rather than to disease type. Repeat length distributions measured in both blood cells and buccal cells from the same DM1 individuals (Morales *et al.*, in preparation) showed similar differences. Here, the variance to mean ratio was found to be higher in blood than in buccal cells reflecting a higher percentage of contraction and hence a lower percentage of expansion in blood than in buccal cells. Differences other than those linked to cell type may have a molecular basis related to flanking GC content which differ in DM1 and HD with a slightly higher percentage of GCs in HD. As there is a strong correlation between the relative expandability of these repeats and the flanking GC content (Brock et al. 1999, Nestor & Monckton 2011) the higher percentage of GCs in HD might explain the weighting towards expansion in HD and further illuminate a modifying role for flanking GC content.

As well as quantifying the length-specific effect, we inferred the parameter values underlying the best fit and associated with the maximum likelihood value. For model M_α, the parameters comprised expansion rate per repeat per year (λ), contraction rate per repeat per year (μ), length parameter (α) and inherited repeat length (n_0). For small alleles, in both DM1 and HD, we found statistical support for expansion and contraction and individual-specific parameters. Our estimate for the length parameter in HD (average = 26 CAGs) is lower than for DM1 (average = 51 CTGs) which may reflect differences in flanking GC content and possibly explain differences in the disease threshold, which is lower for HD (35 CAGs) than for DM1 (50 CTGs). Expansion and contraction rates were correlated in DM1 (correlation coefficient > 0.99) and in HD (correlation coefficient $= 0.70$) suggesting that expansion and contraction, in both DM1 and HD, may be different outcomes of the same underlying process or otherwise conserved components of the instability pathway. This result has direct relevance to therapies that target the mutations directly (Castel et al. 2010) in order to readdress the balance and/or reduce levels of instability, as it is very likely that both expansions and contractions will be affected by a potential therapy.

For individuals with pure repeat length tracts, without interruptions in their repeat lengths, the value of α relates directly to the size of the hypothetical length constraint on the mutational mechanism. We estimate the value of this constraint to be, on average, 26 CAGs in HD and 51 CTGs in DM1. Model M_α is sufficiently general to apply to any length constraint that acts in this manner. In addition to the 14 DM1 individuals with inherited repeat lengths less than 100 CTGs, we identified eight DM1 individuals with distributions with lower than expected (model M_{6b}) variance to mean ratios (Figure 5.1). We hypothesised that the repeat length distributions in these individuals may also have been affected by an individual-specific length constraint of biological origin. We, therefore, fitted model M_α using an extended range (0-200 CTGs) for the length constraint, α, to an additional 128 DM1 sized single molecule blood DNA datasets. We found that for six of the eight DM1 individuals, mentioned above, model M_α fitted the data better than model M_{6b} (likelihood gain greater than or equal to 2), see Table 5.3A. The explanation for this lies with high estimated values for α (80 CTGs and above) and a correspondingly better fit at the low end of the repeat length distributions in these individuals (Figures 5.11 and 5.12) whose likelihood dramatically increased. An improvement in fit (≥ 2) and a high estimated value for α (> 100 CTGs) were also observed for a further 12 individuals listed in Table 5.3B, notably individuals BC8 and SCO115. These results suggest that length-specific effect may operate in some individuals over higher repeat length ranges (greater than 200 CTGs).

As mutation rates are assumed negligible in repeat lengths less than α under model M_α, the effective length of the repeat length tract, with respect to mutation, can be considered to be the remaining number of repeats in the tract, complementary to α. We thus defined the difference between the inherited repeat length and α as the effective length of an individual. In this context, individuals with either small inherited repeat lengths (less than 100 CTGs) and/or high estimated values for α are predicted to have effective lengths much smaller than their actual lengths. The length of α may be determined by individual-specific *cis*-acting and or *trans*-acting factors. One such plausible *cis*-acting factor could be an interruption in the pure CTG tract such as CGG or CCG. A model for how variant repeats might modify mutational dynamics in an adjacent stretch of pure CTG repeats is outlined in Figure 5.3. As discussed in the introduction of this chapter, Section 5.2, in the presence of variant repeats at the $3'$ end of the array, the distance between the pure CTG tract and the *cis*-acting modifier will be increased and its effect may be reduced. In our model, high values of α suggest that pre-mutation or other mutation events (such as loop-outs or mis-pairings) either cancel one another out or do not occur over a greater distance than expected (around 50 CTGs). One possible interpretation with implications for effective length, illustrated in Figure 5.13, is that this distance or length constraint has been extended by the physical presence of an interruption. A rule where the length constraint of α applies only on one side of an interruption at position β from the other side would be entirely consistent with the uninterrupted version. We would simply infer $\alpha + \beta$ in the first instance and α in the second instance. High inferred α values correspond to low effective lengths and potentially less instability and disease.

All 142 DM1 individuals were screened for variant repeats CCG or CGG (Couto *et al.*, in preparation). BC19 and BC6 tested positive for CGG interruptions (Couto *et al.*, in preparation). These individuals (BC19 and BC6) are part of an extended family who initially came to attention (Ashizawa et al. 1992) because of the discrepancy between their clinical symptoms and the molecular diagnosis of inherited repeat length. The symptoms of BC6 are less severe than expected and the inherited repeat length of his son did not show the usual expansion effect of anticipation. Our analysis independently suggests that a length-specific effect (α values 175 CTGs and 110 CTGs, respectively) operates in these individuals and supports a role for CGG interruptions as modifying mutation rates and resulting in less somatic mosaicism. This in turn may explain reduced disease progression in this family. Interestingly, the repeat length distribution for individual SCO99, who tested positive for CCG interruptions and who also has a low variance to mean ratio, is not explained by a length constraint. We conclude that the CCG interruption in this individual does not affect mutation rates in a length dependent manner, though it may do so through another means. Other individuals (Tables 5.3A and 5.3B) with high α values not testing positive for CCG or CGG variant repeats may have other variant repeats or unusual flanking sequences which act in a length-dependent manner and are therefore candidates for further investigation.

Inherited repeat length explains a large proportion of variance in age of onset and, as shown in

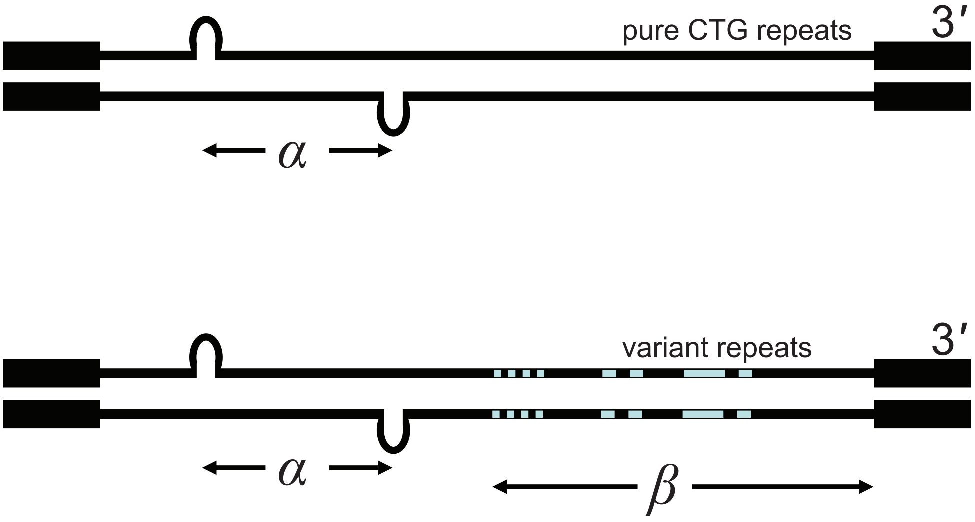


Figure 5.13: **Hypothetical explanation for how variant repeats might modify effective length.** A length constraint α may suppress mutation rates resulting in an effective length, $n_0 - \alpha$, lower than inherited repeat length, n_0 (top). The presence of variant repeats in the repeat length tract may reduce effective length further becoming $n_0 - \alpha - \beta$ (bottom).

Chapter 3, the relationship between inherited repeat length and age of onset is not straightforwardly linear. This relationship is further complicated by somatic instability, which has also been shown to modify age of onset in DM1 (Morales et al. 2012) and HD (Swami et al. 2009). In DM1, thresholds beyond which increasing allele length appears to no longer contribute toward age of onset have been reported (143 CTG (Hamshere et al. 1999) and 250 CTG (Savić et al. 2002)) but Morales *et al.* found a highly significant relationship between the logarithm of estimated inherited repeat length and variation in age of onset, both above and below the average threshold of 200 (Morales et al. 2012). The apparent threshold is likely attributed to an additional, non-linear component to the relationship between age of onset and estimated inherited repeat length. By quantifying length-specific effects we can now suggest a biologically plausible explanation for this non-linear component, namely that levels of somatic mosaicism do not progress in a linear fashion. Levels of somatic mosaicism appear to be relatively lower in small alleles than in long alleles due to the length-specific effect. This gives rise to relatively later ages of onset in small alleles than in long alleles, resulting in the observed non-linear relationship between age of onset and inherited repeat length, discussed in Chapter 3. We would expect effective length to align more closely and better predict age of onset and disease progression than inherited repeat length. To test this prediction, we obtained an estimate for effective length by adjusting inherited repeat length (by subtracting α) in the individuals with lower than expected variance to mean ratios (15^{th} percentile) and setting α equal to zero in the other individuals (16^{th} to 100^{th} percentile), 128 DM1 individuals in total. We then compared inherited repeat length and effective length, in 128 DM1 individuals, in terms of explaining age of onset using linear regression analysis. Effective length (adjusted $R^2 = 50.6\%$, $P < 10^{-15}$, $N = 128$) was better than inherited repeat length (adjusted $R^2 = 46.8\%$, $P < 10^{-15}$, $N = 128$) at explaining variance in age of onset confirming our expectation. Importantly, using model M_α, we remove some of the bias in mutation rates, mentioned above, making them less length dependent. Under model M_{6b} the mutation rates were correlated with inherited repeat length (correlation coefficient = 0.64, $P < 10^{-5}$), whereas under model M_α correlation between mutation rates and inherited repeat length was much lower (correlation coefficient = 0.30, $P < 10^{-5}$). Consequently, rates adjusted in this way will be better suited as quantitative traits to investigate *trans* or *cis*-acting modifiers of somatic mosaicism, disease onset and progression.

Our findings that mutational rates may be suppressed in the region above the disease thresholds in both HD buccal DNA (most effective up to 60 CAGs on average) and DM1 blood DNA (most effective up to 173 CTGs on average) are encouraging from a clinical perspective. Individuals with alleles in this range generally have reduced levels of somatic mosaicism, less severe phenotypes and later age of onset. 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. Suppression of somatic expansion is therefore expected to be therapeutically beneficial and induction of contractions potentially curative (Gomes-Pereira & Monckton 2006, Castel et al. 2010). However, the feasibility of suppressing expansions/inducing contractions remains largely undetermined. Our results show that, in principle, therapies aimed at reducing the length of disease DNA tracts, if successful, should result in lower levels of somatic mosaicism which should slow down disease progression. Interruptions in the disease repeat length tract have also been associated with a less severe phenotype in DM1 (Musova et al. 2009, Braida et al. 2010) and we now suggest a biological basis for this which links interruptions and the pattern of repeat length distributions to lower levels of somatic mosaicism and, in the case of one family, less severe phenotypes.

Inherited repeat length and somatic instability are emerging as key modifiers of disease onset and progression in DM1 and HD (Swami et al. 2009, Morales et al. 2012). However, the relationship between inherited repeat length, somatic instability and age of onset appears complex. Our work unravels some of this complexity through estimation of the biological parameters that drive levels of somatic mosaicism. Through quantification we can better assess the relative importance of these parameters within an individual, between individuals and between cell types and diseases. We find similarities in the underlying mechanism, as evidenced by strong correlation between expansion and contraction rates in both DM1 and HD. But we also find high levels of variation in these rates suggesting that individual-specific factors modify levels of somatic mosaicism to a large degree. Also, as illustrated here, some variant repeats or other polymorphisms may further modify repeat length distributions and disease progression. Finding factors that modify disease is an important next step that will be facilitated by the use of biologically relevant quantitative traits, such as those established here.

5.5 Material and methods

5.5.1 Project data

The data in this study comprise distributions of CTG repeat lengths sized from blood DNA from 14 DM1 affected individuals at the DM1 locus, see below, and distributions of CAG repeat lengths sized from buccal DNA from 12 HD affected individuals at the HD locus (Veitch et al. 2007). DM1 individuals with inherited repeat lengths less than 100 CTGs were selected for this study from the total cohort of 145 DM1 individuals, see Chapter 2, Section 2.2, as their repeat lengths are representative of the smallest repeat lengths seen in DM1. Out of the 14 individuals selected, 9 were asymptomatic when the blood samples were taken and 5 had late-onset with age at onset ranging from 46 years to 74 years. The 12 unrelated HD individuals (Veitch et al. 2007) had estimated inherited repeat lengths between 39 and 48 CAGs and were all 39 years old when the buccal samples were taken. The distributions were sized, in terms of the number of repeats, using single-molecule PCR assays.

5.5.2 Mathematical model with length-specific effect

As described in Chapter 2, Section 2.4, representing the expansion rate per year, the contraction rate per year and inherited repeat length by λ_n, μ_n and n_0, respectively, and letting $P_n\left(t\right)$ denote the probability that an allele has length n at time t, the rate of change of $P_n\left(t\right)$ with respect to time is governed by the master equation of the form

$$\frac{dP_n(t)}{dt} = -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t). \tag{5.1}$$

Given the allele length at time zero, n_0, we may approximate this infinite system of ordinary differential equations numerically by truncating at a suitably large value of $n = N$ and setting $P_n(t) = 0$ for all $n \geq N + 1$.

To specify the functional form of λ_n and μ_n, we departed from the traditional linear model by introducing a threshold, a, 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. Hence the definitions for λ_n and μ_n were $\lambda_n = \lambda(n - a)$ and $\mu_n = \mu(n - a)$, respectively.

To derive a new variation of this model, we formulated a length-specific factor, denoted R_n, as follows. Let the total repeat length be n. Consider now a length constraint on the mutational mechanism. A and B are locations where repair is needed (for example loop-outs). We hypothesise that subsequent mutation requires $|A-B| > \alpha$, where α is an unknown number of repeat units to be inferred from the data. The length parameter α is therefore interpreted as the minimum separation between repeats required for mutation to occur. We are interested in the likelihood that mutation occurs or the proportion of all possible distances that result in $|A-B| > \alpha$. Assuming that A and B occur at arbitrary uniformly random positions along n, and that these occurrences are independent, there are n^2 possible complementary pairs. Using combinatorial counting methods it can be shown that there are $(n-\alpha)(n-\alpha+1)$ pairs separated by distance $|A - B| > \alpha$.

Hence the ratio, R_n, of possible mutation events, is defined as

$$R_n = \frac{(n-\alpha)(n-\alpha+1)}{n^2}. \tag{5.2}$$

We note that for fixed α, $R_n \to 1$ as $n \to \infty$. This corresponds to the intuitively reasonable notion that the finite length constraint is negligible for very large repeat lengths.

R_n can be considered as the biophysical capacity of a repeat length to undergo expansion and contraction. We would expect smaller alleles to have a lower capacity than larger alleles to expand and contract. Based on these considerations, we modify our basic model, Equation (5.1), by introducing R_n as follows

$$\lambda_n = R_n \lambda n, \tag{5.3}$$

$$\mu_n = R_n \mu n, \tag{5.4}$$

for $n > \alpha$, where λ and μ are now constant rates of expansion and contraction per repeat unit per year respectively.

This introduces a nonlinearity into our equations and hence we cannot derive closed forms for the mean and variance. However, the equations can of course still be solved numerically.

5.5.3 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 (5.1) numerically in order to obtain the probability distribution function components $P_n\left(t\right)$ which give the probability that repeat length is n at time t. The likelihood $L^{[i]}$ is then the product over all the data $d_j^{[i]}$, which denotes the repeat length for the jth observation from individual i, of the probability $P_{d_j^{[i]}}(t^{[i]};\theta^{[i]})_{n\geq\alpha}$, 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,

$$L^{[i]} = \prod_j P_{d_j^{[i]}}(t^{[i]};\theta^{[i]}), \tag{5.5}$$

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

$$L = \prod_i L^{[i]}. \tag{5.6}$$

The Akaike information criterion (AIC) is used to assess the goodness of the fit of the model (Akaike 1974). AIC uses the maximised value of the likelihood of the model, L_{max}, penalised by the number of model parameters, k, to rank models thus

$$AIC = 2k - 2\log L_{max}, \tag{5.7}$$

with the model with the smallest AIC value being ranked highest.

We obtain the maximum value of the likelihood by evaluating the likelihood over a broad parameter space described in Table 5.1. Maximisation of the likelihood L in Equation (5.6) is essentially the maximisation of each $L^{[i]}$ in Equation (5.5) using each dataset from an individual. For further statistical analysis, it was useful to have point estimates for the parameters. These were taken to be the maximum likelihood values.

The relative likelihood of two models with AIC values denoted AIC_1 and AIC_2 respectively, where $AIC_1 < AIC_2$ is

$$\exp\left(\frac{AIC_1 - AIC_2}{2}\right). \tag{5.8}$$

Chapter 6

Levels of somatic instability in Huntington disease related tissue are linked to age of onset and disease progression

6.1 Abstract

Evidence of somatic expansion in tissues that are the targets of pathogenesis has given rise to the hypothesis that somatic instability may itself contribute to the pathogenic process. However the interpretation of the levels of somatic instability in many of the affected tissues in the triplet repeat diseases is hindered by complex cell compositions. It has recently been demonstrated by Swami *et al.* that larger somatic expansions and hence skewed distributions of the HD CAG repeat expansion in HD frontal cortex at end stage are significantly associated with an earlier age of disease onset, independent of any effects of inherited CAG repeat length on either somatic instability or onset age. This interesting dataset comprised post mortem end stage CAG repeat length distributions from 48 HD individuals with either an extremely young or an extremely old age of onset but matched inherited repeat lengths. We now extend our mathematical model to two cell populations whose repeat lengths have different rates of mutation (fast and slow). We infer for each frontal cortex HD dataset the likely relative weight of these cell populations and their corresponding contribution towards somatic variation. By comparison with data from laser captured single cells we conclude that the neuronal repeat lengths most likely mutate at a higher rate than glial repeat lengths, explaining the characteristic skewed distributions observed in mixed cell tissue from the brain. Derived parameter values differ significantly between the two extreme phenotypes and we show that individual-specific mutation rates in neurons are, in addition to the inherited repeat length, a modifier of age of onset. Using the parameters estimated from our HD end-stage analysis we also simulate the expected distribution of repeat lengths at age of onset. Very interestingly, the predicted repeat length distributions at disease onset in neurons are very similar between individuals, despite very different ages at onset. Our results support a model of disease progression where individuals with the same inherited repeat length may reach age of onset, as much as 30 years earlier, because of greater somatic expansions underpinned by higher mutational rates. Therapies aimed at reducing somatic expansions would therefore have considerable benefits with regard to extending the age of onset.

6.2 Introduction

Huntington disease

As discussed in Chapter 1, Huntington disease (HD) is an inherited neurological disorder characterised by progressive movement, psychiatric and cognitive disturbances. Neurodegenerative changes in the brain of affected individuals follow a typical pattern, with early cellular dysfunction and loss of medium spiny neurons in the striatum, followed by more generalised cell loss across the brain (Graveland et al. 1985). Whilst the age of disease onset is strongly inversely correlated with the length of the expanded CAG repeat length (Andrew et al. 1993, Duyao et al. 1993, Snell et al. 1993, Stine et al. 1993, Gusella et al. 1996), with repeat length accounting for around 70% of the variability in age of onset, this reduces to less than 50% for the majority of HD patients with repeat lengths below 60 CAGs (Myers et al. 1998, Li et al. 2003). There is evidence for heritability for the portion of age at onset not explained by CAG repeat size, which provides support, along with several other studies, *e.g.* (Li et al. 2003, Wexler et al. 2004), for genetic modifiers of age of onset. Measurement of biomarkers that contribute to variation in age of onset could be used to identify these genetic modifiers, which are key targets for therapies aimed at slowing or reversing the pathogenic process.

The expanded HD CAG repeat is somatically unstable, undergoing progressive length increases over time (Telenius et al. 1994, Kennedy et al. 2003). Somatic instability is also tissue-specific with high levels found in striatum and cortex (Shelbourne et al. 2007) and occurs in post-mitotic neurons (Gonitel et al. 2008). Somatically expanded HD CAG repeats are transcribed and translated (Aronin et al. 1995, Wheeler et al. 2003, Gonitel et al. 2008). Evidence of somatic expansion in tissues that are the targets of pathogenesis has given rise to a hypothesis that somatic instability may itself contribute to the HD pathogenic process. Experiments in a genetically accurate Huntington disease homologue (*Hdh*) knock-in mouse model (Hdh^{Q111}), in which an early symptomatic, HD CAG length-dependent phenotype was significantly delayed in mice that lacked somatic instability as a result of the deletion of mismatch repair genes *Msh2*, supports this hypothesis (Wheeler et al. 2003).

Different cell types in the brain (principally neurons and glia) show different levels of instability, with higher levels seen in neurons (Shelbourne et al. 2007, Gonitel et al. 2008). This provides a straightforward explanation for the multi-modal and skewed shape of the distributions of repeat lengths in brain tissue. Jung and Bonini proposed that CAG instability is linked to pathogenesis as seen in a *Drosophila* model (Jung & Bonini 2007), but the conclusions of Gonitel *et al.* and the experiments of Wheeler *et al.* (specifically showing that the rate of somatic instability correlates with huntingtin accumulation in neuronal nuclei) would contradict this suggestion and place somatic instability as a significant disease modifier with other reported factors (Lloret et al. 2006) contributing to the process.

Recently, Swami *et al.* investigated the potentially modifying role of somatic instability on the age of onset phenotype (Swami et al. 2009). Their study design chose HD individuals with extreme phenotypes for young and old age of onset. These were the individuals whose age of onset deviated the most from what would be predicted by their inherited repeat length alone. They identified 48 individuals, 24 with an extremely young age of onset and 24 with an extremely old age of onset, matched for their mutant and normal inherited repeat lengths (as established by analysis of cerebellar DNA which is somatically stable (Kennedy et al. 2003)) but with mean age of onset differing by approximately 30 years. The frontal cortex was chosen for examination as it has been shown to retain relatively high levels of mosaicism at the end stage of the disease compared to the striatum where reduced levels of variation are observed at end stage most likely due to disease related cell loss (Shelbourne et al. 2007). The alleles were sized using small-pool PCR and provide suitable distributions for quantitative analysis using a mathematical model as previously described for CTG repeat lengths in myotonic dystrophy type 1 blood (DM1) in Chapter 4 and now modified for CAG repeat lengths in HD frontal cortex. The key finding of the Swami *et al.* study was that repeat length distributions are biased towards longer alleles in individuals with earlier disease onset (Swami et al. 2009). There was a significant difference between the two groups of extreme phenotypes concerning the maximum expansion, and more robustly (as maximum expansion is only one observation), skewness (a measurement of the degree of symmetry of a distribution) of the samples with the extremely old age of onset having lower skewness than the extremely young age of onset. Their results demonstrated that larger somatic expansions of the HD CAG repeat expansion in HD patient cortex are significantly associated with an earlier age of disease onset, independent of any effects of inherited CAG repeat length on either somatic instability or onset age. A mechanism for age of onset has been proposed by Kaplan *et al.* for trinucleotide diseases in general where disease onset is triggered when a percentage of disease related cells (arbitrarily chosen to be 20%) cross a critical threshold in terms of expanded repeat length (Kaplan et al. 2007) . They suggest that this critical threshold for HD is 115 CAG repeat units. The results of Swami *et al.* are consistent with the expectation that individuals starting at the same inherited repeat length who have more somatic expansion would reach the disease threshold earlier than individuals with less somatic expansion, but do not provide further quantitative or descriptive features of this threshold.

Study aims

As the different cell types in the brain (principally neurons and glia) show different levels of instability, with higher levels seen in neurons than in glia (Shelbourne et al. 2007, Gonitel et al. 2008), it follows that the overall distribution of repeat lengths in a complex tissue sample comprising one or more cell types may be skewed or contain more than one mode. Interpretation of the components of somatic mosaicism in the frontal cortex, a tissue directly involved in the pathology of HD, is hindered by the complex cell composition of this tissue. In this study we extend our mathematical model to monitor changes over time in repeat length in two cell types: cells in which the repeat lengths expand more rapidly (fast) and cells in which the repeat lengths expand less rapidly (slow). It is predicted that the cells with the fast expanding repeats will be neurons and the cells with the slow expanding repeats will be glia. We fit this extended model to the datasets described in (Shelbourne et al. 2007, Swami et al. 2009) and infer the relative composition of these two cell types and their respective rates of mutation. We hypothesise that the derived parameters will explain some of the variability in age of onset not explained by the inherited repeat length. To further investigate the role of somatic mosaicism within the frontal cortex and its relationship to phenotype, we simulate the expected distribution of mutant alleles at age of onset. This provides further qualitative and quantitative indication of the role somatic instability plays in age of onset and disease progression.

6.3 Results

6.3.1 Assumptions underlying the mathematical models

To clarify the presentation and discussion of our results, we begin by stating and justifying our key assumptions about how the data arose. We assume that the CAG repeat lengths can undergo expansion (increase in length) and contraction (decrease in length). Our mathematical model quantifies the probability of both expansion events and contraction events in the repeat length in frontal cortex cells. The human brain contains around 160 billion cells (Azevedo et al. 2009) split roughly into equal numbers of neurons, the basic building blocks of the nervous system, and glial cells, the non-neuronal cells that provide support and protection for neurons. The ratio of glia to neurons differs from one part of the brain to another. This ratio is thought to be as high as 3.72 in the cerebral cortex, 1.48 in the cerebral cortex gray matter and as low as 0.23 in the cerebellum (Azevedo et al. 2009). In the frontal cortex the ratio is expected to be around 2.

There are differences between neurons and non-neuronal cells in terms of their generation. It is thought that virtually all neurons ($> 99\%$) are generated prenatally and retained for our lifespan, whereas there may be constant production of new glial cells in adults (Bhardwaj et al. 2006). That large repeat length changes occur in terminally differentiated, post-mitotic neurons was confirmed by (Shelbourne et al. 2007, Gonitel et al. 2008). In formulating our model for brain cells, we assume that the sampled cells, comprising both cell types, have had independent mutational histories (from their prenatal generation and differentiation from stem cells, onwards). For neurons, which are virtually all non-dividing cells, this is clearly the case. For glial cells, even if they are under production in adults as indicated by (Bhardwaj et al. 2006), the sheer number of cells makes it extremely unlikely that two sampled cells arose from the same glial stem cell and so the assumption of independent mutational histories remains valid.

The inherited repeat length was estimated for each individual from cerebellar DNA by Swami *et al.* but by treating it here as an unknown parameter we can establish whether in principle, it is possible to infer the inherited repeat length from a pathological tissue. A positive result would be useful for future work exploring somatic mosaicism where there is more instability and hence less certainty where the inherited repeat length lies, e.g. muscle in DM1. The changes in repeat length are age-dependent and as the samples were end stage taken at autopsy, the appropriate input for age or time in our model is age at death. Information about age at death was available for 38 samples (22 young age of onset and 16 old age of onset). Comparison between different sized groups, as in this case, is not an issue for our modelling or inference methods. The other additional statistical test used and comparisons with the full dataset take this difference in size into account.

Our probabilistic model assumes that the probability of mutational events occurring increases as the repeat lengths get longer. As previous work, see Chapter 5, reported that small alleles with repeat lengths less than 200 repeat units long may have less capacity to mutate than longer alleles, we have introduced into the model a further parameter, α, quantifying a length effect. This addition is particularly pertinent for HD as allele lengths are generally much lower than those found in myotonic dystrophy type 1.

The mutational gains and losses are assumed to be of one CAG unit. Such small gains and losses have been observed in several studies (*e.g.* Veitch et al. 2007). It has been suggested that the mutations occur in a synchronous manner (Gonitel et al. 2008) based on the observation that the distributions in genetically identical mouse models appear to follow a predictable pattern. We argue that a predictable pattern can also be the result of a stochastic process, where many cells are involved, as the aggregate behaviour of many cells in the stochastic processes will appear to be deterministic.

We assume that the cell types contained within the frontal cortex can be differentiated by rates of mutation in their repeat lengths: fast and slow. Hence we fit two allele distribution curves: one for cells with fast mutating repeat lengths and one for cells with slow mutating repeat lengths, with a range of weights for each type of cell (Table 6.1). We obtain the maximum likelihood value and the associated parameter values using a grid search of the parameter space (Table 6.1). This likelihood method quantifies which parameters provide the best fit to the data. The fit can be visualised as a distribution curve or a cumulative distribution curve and can be compared with the data, as in Figures 6.1, 6.2 and 6.3. As this is a computationally demanding task, involving many parameters, consideration was given to the design and organisation of the implementation to minimise the computational cost.

Model parameters	**Prior ranges**
Contraction rate per CAG unit per year in cells with fast mutating repeat lengths (μ^f)	0 – 0.12
Net expansion rate per CAG unit per year in cells with fast mutating repeat lengths (λ^f - μ^f)	0.01 – 0.031
Length parameter in cells with fast mutating repeat lengths, number of CAG units (α^f)	10 – 30
Contraction rate per CAG unit per year in cells with slow mutating repeat lengths (μ^s)	0 – 0.02
Net expansion rate per CAG unit per year in cells with slow mutating repeat lengths ($\lambda^s - \mu^s$)	0.0001 – 0.0014
Length parameter in cells with slow mutating repeat lengths, number of CAG units (α^s)	10 – 30
Percentage of cells with fast mutating repeat lengths (w^f)	0% – 100%
Inherited repeat length, number of CAG units (n_0)	38 – 50

Table 6.1: **Prior ranges for parameter estimation for Huntington disease brain.**

Another important issue is that neurons with the longest repeat expansions may be preferentially

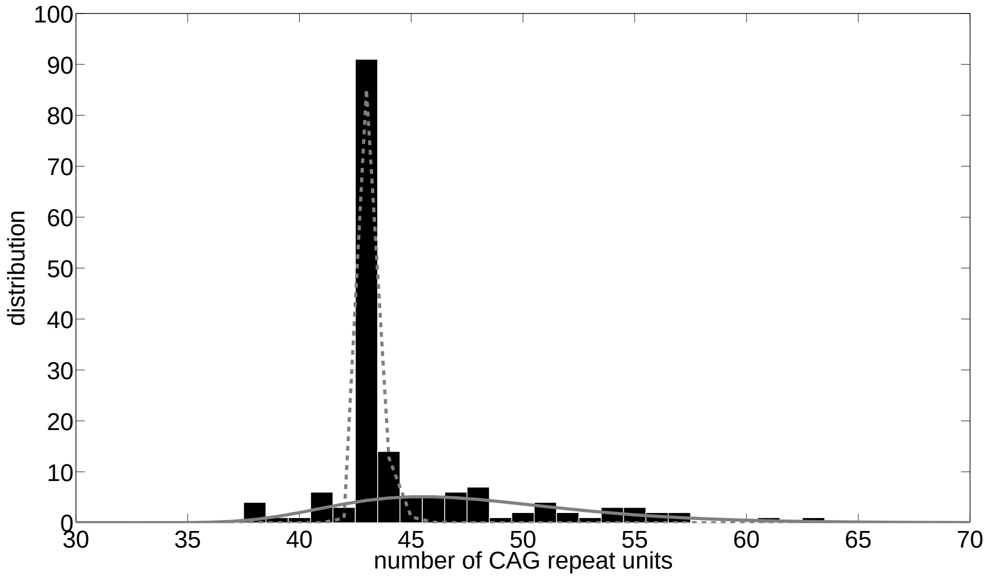


Figure 6.1: **Histogram showing the distribution of CAG repeat lengths in frontal cortex from a representative HD individual (sample 11) compared with the fitted probability distribution for fast mutating cells (solid line) and the fitted distribution for slow mutating cells (dashed line).**

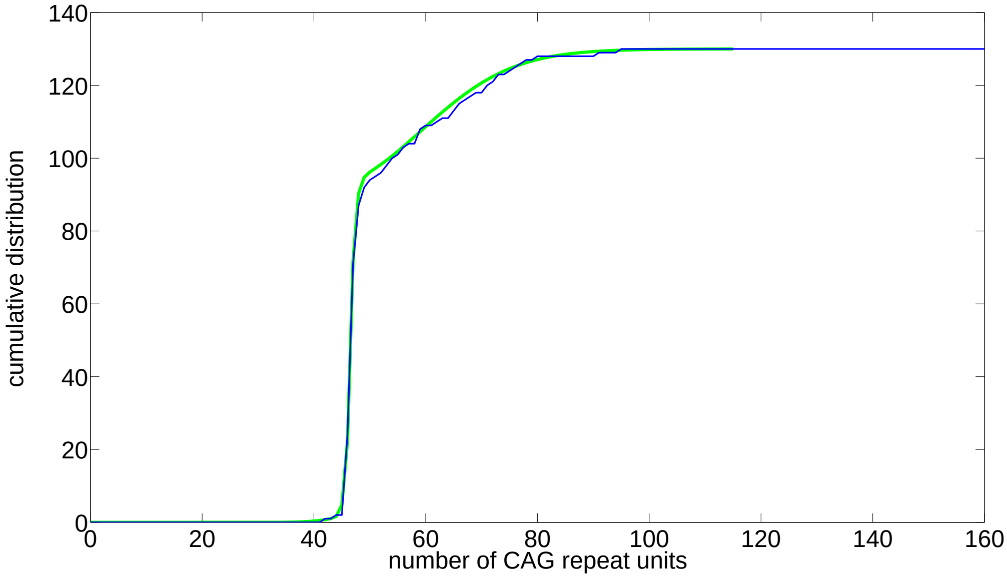


Figure 6.2: **Cumulative distribution of CAG repeat lengths in frontal cortex (jagged line) from a representative HD individual (sample 11) compared with the fitted probability distribution (smooth line).**

lost during the disease course (Kennedy et al. 2003) and we consider the importance of such an effect through a model with a truncated distribution, described more fully below. Preferential loss of neurons with long repeat lengths during the disease course and their consequential absence from the data may result in an underestimation of the mutation rates. The highest observed repeat length in this study in frontal cortex was 116 CAGs. To address this issue, we also fit allele distributions truncated at a range of repeat lengths (100 CAGs to 145 CAGs) including the length 115 CAG units which has been proposed as critical to disease onset by (Kaplan et al. 2007). These truncated distributions will predict higher rates of mutation if the distribution of the other alleles support this. We use a model comparison method (AIC) to formally compare these fits.

6.3.2 Model comparison supports two cell types and a minor role for truncation

We tested our hypothesis that cells with two different rates of mutation were responsible for the skewed, multi-modal shape of the allele distribution by fitting a mixed distribution to the data. This hypothesis was tested against the null hypothesis that one cell type would explain the changes in repeat length using the likelihood ratio test. We also fitted a truncated distribution to the data to test the extent to which preferential cell loss may have a role in shaping the repeat length distributions. In summary, the following models, with parameters specific to each HD individual, were fitted to the data:

- Model $F1$: Full distribution one mutating repeat length cell type (4 parameters per 38 HD individuals, 152 in total)
- Model $F2$: Full distribution two (fast and slow) mutating repeat length cell types (8 parameters per 38 HD individuals, 304 in total)
- Model $T1_{100}$: Truncated distribution (100 CAGs) one mutating repeat length cell type (152 parameters)
- Models $T2_{100}$, $T2_{115}$, $T2_{130}$ and $T2_{145}$: Truncated distribution (100 CAGs, 115 CAGs, 130 CAGs and 145 CAGs respectively) two (fast and slow) mutating repeat length cell types (304 parameters)

The models were compared and ranked using AIC, see Table 6.2 and Section 2.5.3 for further details about AIC. This involved calculating the maximum likelihood value using a grid search over the parameter space as outlined in Table 6.1 (see Chapter 2, Section 2.5, for further details).

Despite having more parameters, Model $F2$ is ranked higher than model $F1$ (maximum likelihoods -11,098 (304 parameters) and -13,739 (152 parameters) respectively) and similarly model $T2_{100}$ is ranked higher than model $T1_{100}$ (maximum likelihoods -11,093 (304 parameters) and -13,735 (152 parameters) respectively) providing strong support for two cell types. The other truncated models ($T2_{115}$, $T2_{130}$ and $T2_{145}$) had similar maximum likelihood values, -11,097 for Model $T2_{115}$ and -11,098 for Models $T2_{130}$ and $T2_{145}$. Model $F2$ and the truncated models are significantly different from the next ranked model, Model $T1$, with a maximum likelihood value of -13,735.

Models **(N=38 individuals)**	**Number of parameters**	**Maximised log-likelihood**	**Adjusted AIC**	**AIC rank**
Truncated Distribution for a mixture of cells with fast and slow mutating repeat lengths				
$T2_{100}$ – 100 CAGs	304	-11,093	0	1
$T2_{115}$ – 115 CAGs	304	-11,097	8	2=
$T2_{130}$ – 130 CAGs	304	-11,098	10	3=
$T2_{145}$ – 145 CAGs	304	-11,098	10	3=
$F2$ Full Distribution for a mixture of cells with fast and slow mutating repeat lengths	304	-11,098	10	3=
$T1$ Truncated Distribution for one type of cell	152	-13,735	4980	4
$F1$ Full Distribution for one type of cell	152	-13,739	4988	5

Table 6.2: **Model comparison summary for Huntington disease brain.** The models, listed in column 1, are ranked using AIC which has been adjusted by subtracting the lowest overall value (Model $T2_{100}$) from the other models (column 4).

We introduced a model with a truncated distribution to assess whether cell loss was having a significant effect on the repeat length distributions and hence whether ignoring this effect would lead to an underestimation of the rates of mutation. This does not appear to be the case as the estimated rates of mutation are virtually identical for models $F2$ and $T2_{100}$ with only two individuals having very slightly higher rates of mutation under model $T2_{100}$. The estimated percentage of fast cells (35%) was also the same for models $F2$ and $T2_{100}$. In order to obtain the parameter estimates, the truncated model used the simplifying assumption that all cells with repeat lengths over 100 CAGs would be lost. This is clearly not the case as two alleles in the total sample were over 100 (103 CAGs and 116 CAGs). So although model $T2_{100}$ with a maximum likelihood value of -11,093 was ranked higher than model F2 suggesting that the addition of a parameter for truncation is justified and was a useful hypothesis for checking the mutation rates, overall, this model requires further work to include the probability of cell death. Taking these factors into account, especially the fact that the parameter estimates are not affected, we consider the full distribution model, $F2$, is an appropriate model for further analysis.

We also considered a model with global parameters for the mutation rates and length effect and individual-specific parameters only for the inherited length and percentage of fast cells (data not shown). However, as reported for DM1 (Higham et al. 2012), global parameters did not capture the variation seen in the data which indicates that individual-specific factors play a major role in somatic instability. Inclusion of contraction events, *i.e.* decreases in repeat length of one CAG unit, was also justified as there was no statistical support for the contraction rates being zero. Model $F2$ is significantly better ($P < 10^{-15}$) than an expansion only model even taking into account the reduced number of parameters (228).

The model fit can be further examined by comparing the expected distribution, associated with the maximum likelihood value, to the data (Figures 6.1, 6.2 and 6.3). Visually, small deviances from the expected distribution are observed (*e.g.* Figure 6.2). To investigate whether these deviances could arise from the model, or whether another explanation is required, we simulated distributions for each of the 38 HD individuals. Using a two-sample Kolmogorov-Smirnov test we show that the observed distributions are very similar to the simulated distribution and hence very likely do arise from the model. An alternative explanation, such as that the small deviances are sub-populations arising from alleles that have had large contractions is not required. In summary, we regard model $F2$ as being the most successful among those considered in explaining how the data has arisen.

6.3.3 Mutation rates are predictive of onset age

Through fitting model $F2$ we obtained the following parameter estimates for 38 HD individuals: the expansion rate for fast and slow mutating cells per CAG unit per year, λ^f and λ^s, respectively, the contraction rate for fast and slow mutating cells per CAG unit per year, μ^f and μ^s, respectively, the length parameter measured in number of CAGs for fast and slow mutating cells, α^f and α^s, respectively, the inherited repeat length, n_0, and the percentage of fast cells, w^f. Our estimated values for n_0 were in close agreement (correlation coefficient $>$ 0.99) with those determined by (Swami et al. 2009) from cerebellar DNA. We hypothesised that the estimated parameter values may differ between the two extreme phenotypes and hence provide a molecular explanation for the underlying mechanism of disease progression. The individual-specific parameters associated with the maximum likelihood value provide an explanation under the full distribution model (Model $F2$) of how each individual has obtained the distribution of repeat lengths observed in frontal cortex DNA through mutational gains and losses of CAGs over their lifetime from their initial inherited length. We tested whether the parameters associated with mutation (λ^f, λ^s, μ^f, μ^s, α^f and α^s)

differ significantly between the extreme phenotypes, by comparing the multivariate means of the parameter values for each group (extremely young age of onset and extremely old age of onset) and performing a one-way multivariate analysis of variance (MANOVA). A statistically significant MANOVA effect was obtained ($P = 0.0063$) indicating that one or more of the parameters do differ significantly between the two groups. The multivariate effect size was estimated at 0.42, which implies that 42% of the variance in the dependent variables was accounted for by phenotype group. The median mutation rates for repeat lengths within both fast cells and slow cells are significantly higher for the young age of onset phenotype group than for the old age of onset phenotype group (see Figure 6.4 and Table 6.3). Concerning the length parameters, α^f and α^s, the median values were not significantly different for each phenotype group (30 CAGs and 20 CAGs respectively).

	Median young age of onset	**Median old age of onset**	**Mann Whitney U test *P-value***	**R^2 explained variance dependent variable age of onset**
Expansion rate per CAG unit per year in cells with fast mutating repeat lengths	0.055	0.031	0.0039	21.6%
Contraction rate per CAG unit per year in cells with fast mutating repeat lengths	0.038	0.018	0.0088	19.1%
Net expansion rate per CAG unit per year in cells with fast mutating repeat lengths	0.017	0.009	0.1003	1.4%
Expansion rate per CAG unit per year in cells with slow mutating repeat lengths	0.00085	0.00055	0.0179	28.0%
Contraction rate per CAG unit per year in cells with slow mutating repeat lengths	0.000075	0.00005	0.6605	11.2%

Table 6.3: **Parameter comparison between extreme phenotypes in Huntington disease brain.** Notes: (i) Bonferroni corrected significance level (5 tests) = 0.01.

Some parameter pairs are highly correlated: expansion and contraction rates for the fast mutating cells (correlation coefficient $r = 0.974$), expansion and contraction rates for the slow mutating cells (correlation coefficient $r = 0.849$). There is also positive correlation between the expansion rates for fast mutating cells and slow mutating cells (correlation coefficient $r = 0.553$) as illustrated in Figure 6.5. The expected number of expansions per year per cell type depends on the model parameters, λ^f, α^f, λ^s, α^s, but also on the current number of CAG units. In summary, our results show

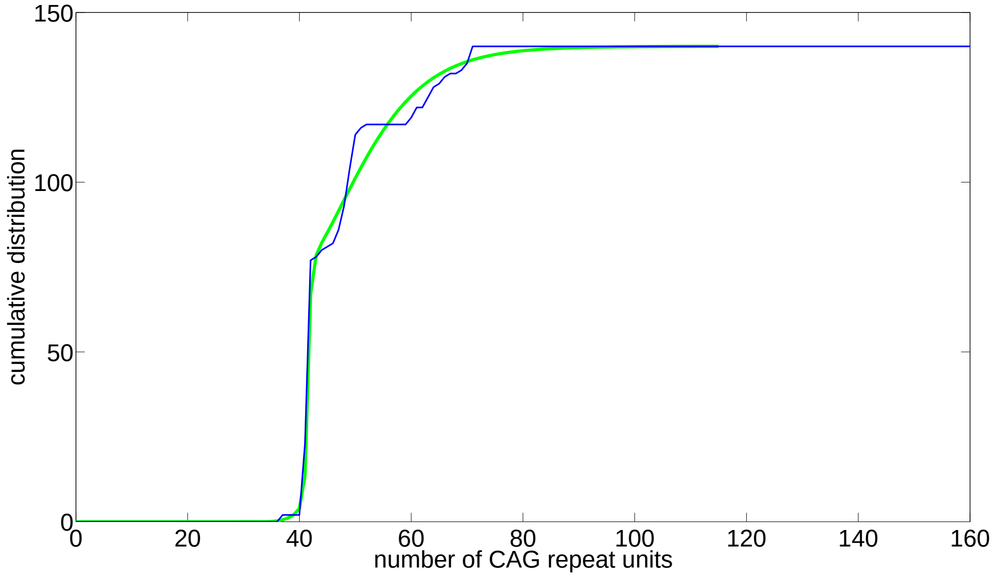


Figure 6.3: **Cumulative distribution of CAG repeat lengths in frontal cortex (jagged line) from a representative HD individual (sample 29) compared to the fitted probability distribution (smooth line).**

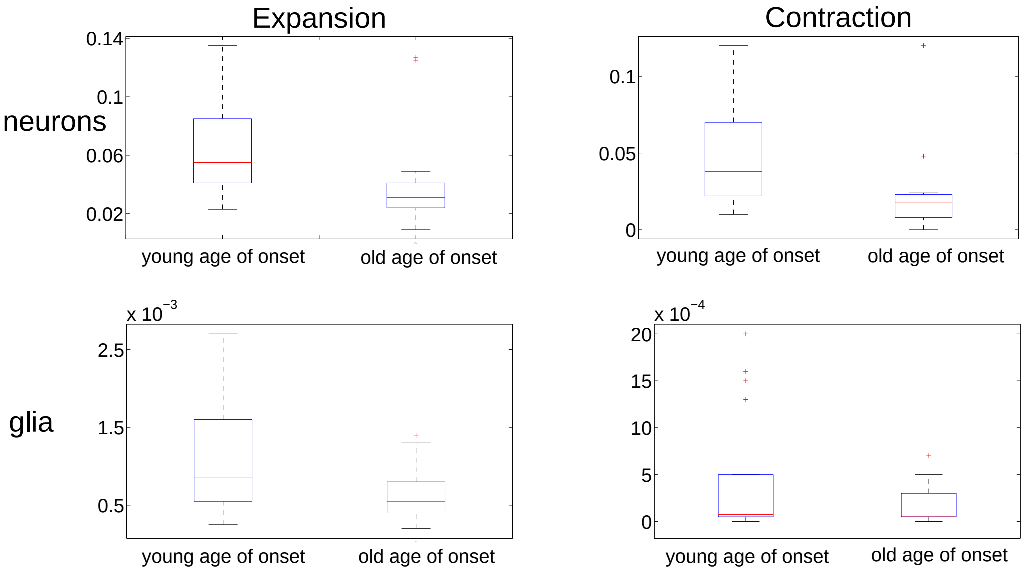


Figure 6.4: **Comparison of expansion rates in neurons between phenotype group with young age of onset (N=16) on left and old age of onset (N=22) on right (top left). Comparison of contraction rates in neurons between phenotype group with young age of onset (N=16) on left and old age of onset (N=22) on right (top right).** Comparison of expansion rates in glia between phenotype group with young age of onset (N=16) on left and old age of onset (N=22) on right (bottom left). Comparison of contraction rates in glia between phenotype group with young age of onset (N=16) on left and old age of onset (N=22) on right (bottom right).

that the median expected number of expansions per year, as a function of length, is significantly different between the extreme phenotypes, young age of onset and old age of onset, with higher levels of mutations in the young age of onset group (Figure 6.6).

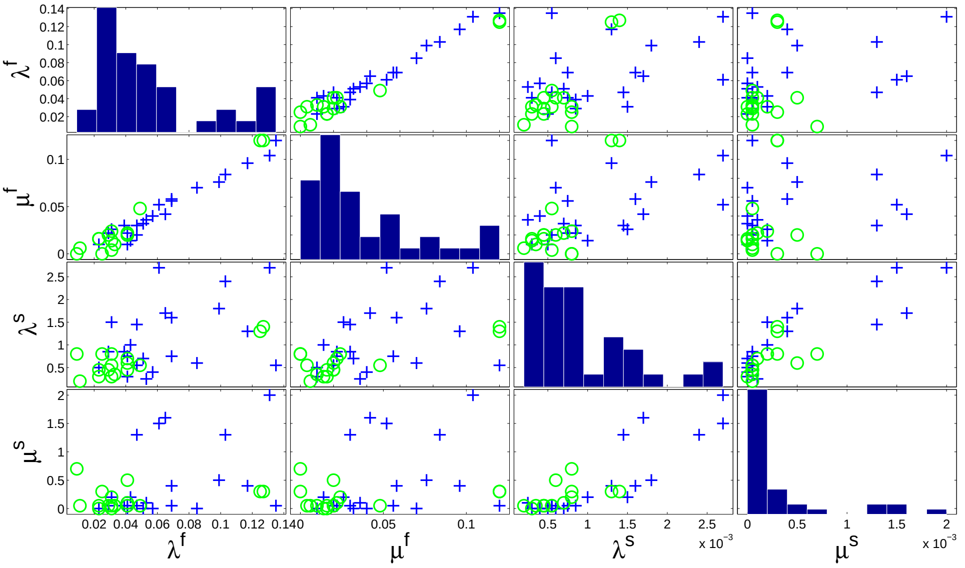


Figure 6.5: **Matrix of scatter plots for pairwise comparison of the model parameters λ^f, μ^f, λ^s, μ^s between phenotypes (young age of onset denoted by '+' and old age of onset denoted by 'o') and a histogram showing the distribution of each parameter along the diagonal.**

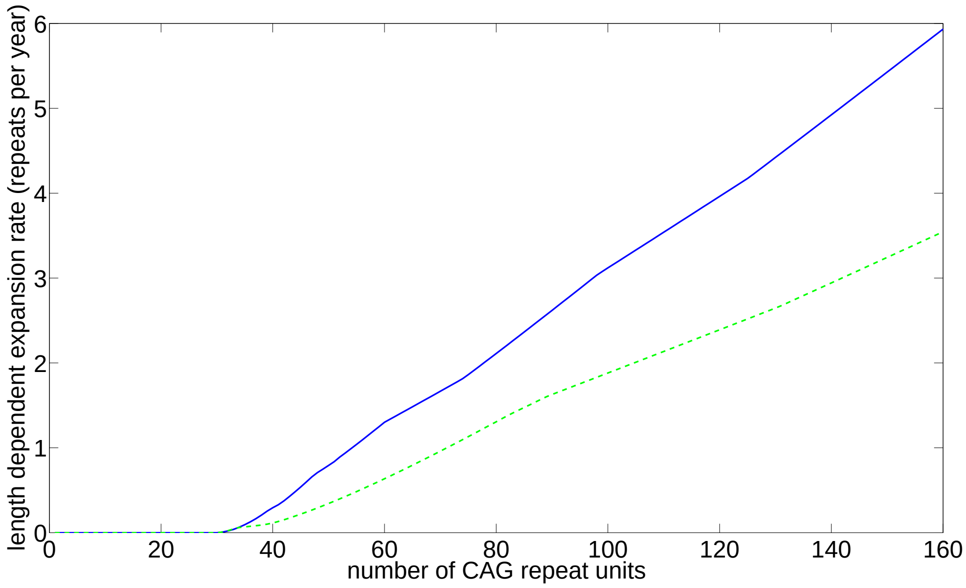


Figure 6.6: **The median expected number of expansions per year, as a function of length, for phenotype group early age of onset (dark solid line) and late age of onset (light dashed line).**

In terms of age of onset, the expansion rate for fast expanding cells explains 21.6% of the variance in age of onset (Tables 6.3 and 6.4). Together, in a linear model, inherited repeat length and expansion explain 57.8% of the variance in age of onset rising to 69.7% in an interaction model. This result indicates that both the expansion rate and the inherited repeat length contribute to age of onset in a complex, non-linear manner, consistent with the fitted model, $F2$, where the number of expansions (and contractions) is dependent on CAG repeat length.

Independent variables	**R^2 explained variance dependent variable age of onset**	**Adjusted R^2**	***P*-value**
Inherited length	52.7%	51.4%	2.4659e-07
Expansion rate per CAG unit per year in cells with fast mutating repeat lengths	21.6%	19.4%	0.0033
Inherited length and expansion rate	57.8%	55.4%	2.7843e-07
Inherited length, expansion rate and interaction, inherited length*expansion rate	69.7%	67.0%	6.2139e-09

Table 6.4: **The relationship between age of onset, inherited repeat length and expansion rate for fast expanding cells, established using linear regression analysis.**

So far we have not identified which cells belong to the fast and slow groups, the obvious distinction being neurons versus glia. Shelbourne *et al.* sized repeat lengths in laser captured single cells in different human brain tissues, caudate nucleus, accumbens, putamen region (CAP), temporal pole of the cortex (TP) and hippocampal formation (HF), separated into grey matter (neuron rich) and white matter (glia rich) (Shelbourne et al. 2007). We fitted our model to each of these datasets and found that expansion rates are higher in grey matter than in white matter ($P = 0.0019$) and that the percentage of fast cells is higher in grey matter than in white matter ($P = 0.024$). This confirms that neurons are most likely the fast mutating cells and glia most likely the slow mutating cells.

Parameter estimates for the percentage of fast mutating cells in each sample reveal that the relative proportions of the different cell types (fast and slow) vary between individuals (10% to 70%) but that the difference between the groups (mean = 30.5% for young age of onset and mean = 38.6% for old age of onset) is not significantly different. The range is greater than that which could be reasonably attributed to sampling from a mixed two cell tissue assuming a ratio of two glia for each neuron (estimated to be $35\% \pm 10\%$ with 95% confidence). As the ratio of neurons to glia varies even within sections of the brain, these individual-specific differences are most likely attributable to differences in the actual sample taken such as position or depth rather than indicating individual differences in the neuron to glia ratio.

6.3.4 Simulation of allele distributions at age of onset

We have inferred the mutation rate values that underlie the measured repeat length distributions at end stage and shown that these values are significantly higher in the young age of onset group. The estimated number of mutational events per year for fast/slow mutating cells is on average 1.241/0.024 for the young age of onset phenotype and 0.591/0.013 for the old age of onset phenotype with, in all cases, high variance. From these results, we predict that the distribution of repeat lengths will spread out over time, from the inherited repeat length, mostly towards higher repeat lengths but also slightly towards lower repeat lengths. In the young age of onset group the distribution of repeat lengths will spread out further and quicker than in the old age of onset group. The difference in rates between cells will give rise to skewed and multi-modal distributions of repeat lengths.

Using the estimated parameter values associated with the maximum likelihood value of model $F2$, the expected distribution of mutant alleles for each cell type was determined by simulation at precisely the age of onset for each HD individual (N=38). Time dependent distributions for alleles are generated under the models using an adapted Gillespie Algorithm (details given in Chapter 2, Section 2.7.1) and preassigned parameter values.

The expected mutant allele distributions, at age of onset, for extremely young onset age and extremely old onset age, further split by cell type, were compared in terms of their percentile median values (Figure 6.7). The Mann Whitney U test was used to determine the significance of any differences between the two extreme phenotypes. For the fast cells (35%), there was no significant difference between the distributions for extremely young onset age and extremely old onset age with the 70^{th} to 90^{th} percentile being highly similar. For the slow cells (65%), all differences in percentile means were significant except for the maximum repeat length at the 10^{th} percentile. These results, specifically the commonality between extreme phenotypes at age of onset, suggest that the distribution of the fast cells defines age of onset rather than the distribution of the slow cells.

Time dependent simulations of repeat length distributions were also generated for all 38 HD individuals 15 years prior to onset, 10 years prior to onset and 5 years post onset (see Figure 6.8 for a summary of the repeat length differences as disease progresses). The greatest differences during the 15 years prior to onset occur among the largest repeat lengths (95^{th} to 100^{th} percentile) where repeat length differences are between 5 and 15 CAGs.

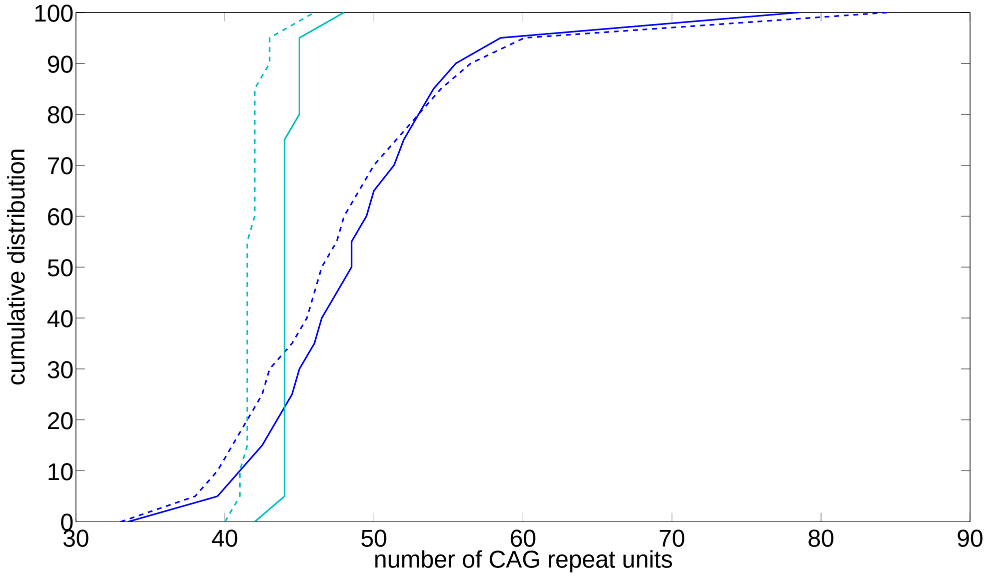


Figure 6.7: **Simulated median cumulative repeat length distributions at age of onset for fast mutating cells (young age of onset, dark solid line and old age of onset, dark dashed line) and for slow mutating cells (young age of onset, light solid line and old age of onset, light dashed line).**

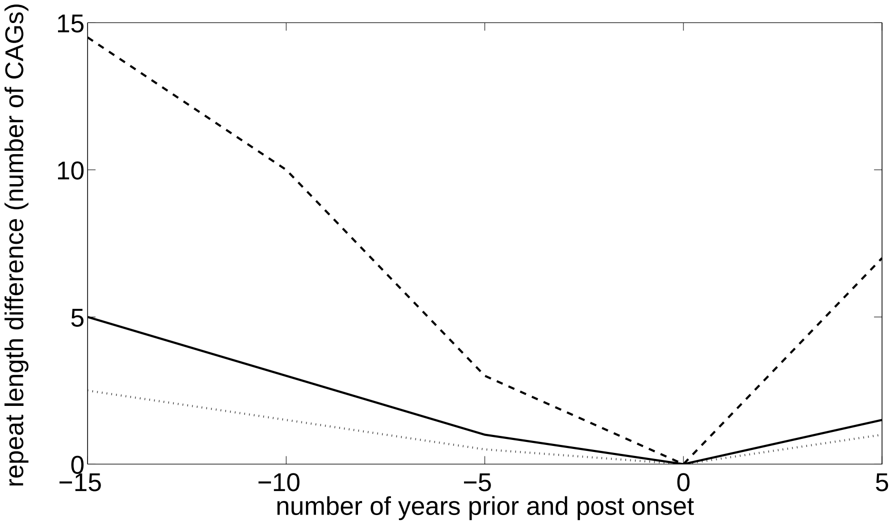


Figure 6.8: **Simulated median repeat length differences among 38 HD individuals by percentile 15 years prior to onset to 5 years post onset.** 100^{th} percentile = long dashed line, 95^{th} percentile = solid line and 70^{th} percentile = short dashed line.

6.4 Discussion

We have extended the model developed in Chapters 4 and 5 to HD frontal cortex. We hypothesised that the repeat lengths in two different cell types (neurons and glia) differ with respect to their mutation rates. This hypothesis was tested against the null hypothesis that there is no difference between the mutation rates in repeat lengths in neurons and glia. We found significant statistical support ($P < 10^{-15}$) for heterogeneous repeat length mutation rates in frontal cortex. Statistical comparison with sized alleles in grey and white brain matter (Shelbourne et al. 2007) suggests that the fast mutating repeat lengths are most likely found in neurons and the slow mutating repeat lengths are most likely found in glia.

The pathology of Huntington disease involves neuronal loss (Vonsattel et al. 1985). If neurons are lost, proportional to inherited repeat length in an age-dependent manner as suggested by (Aylward et al. 1997), we would expect the shape of the repeat length distribution to reflect this. We tested this hypothesis by fitting truncated repeat length distributions to the data. Substantial loss of neurons and hence missing data might lead to an underestimation of expansion rates and so truncated models were also considered in order to assess the effect neuron loss might have on our estimation of the mutation rates. Although there was evidence that some neurons with repeat lengths greater than 100 CAGs may have been lost, this consideration did not affect our estimates of the mutation rates. Hence we concluded that neuronal loss in frontal cortex can have only a minor truncating effect on the repeat length distributions in HD individuals with inherited repeat lengths between 40 - 50 CAGs.

Swami *et al.* defined two phenotype groups: HD individuals with a relatively young age of onset (average = 29 years) taking into account inherited repeat length and HD individuals with a relatively old age of onset (average = 61 years) taking into account inherited repeat length (Swami et al. 2009). We found that individuals with young onset age have significantly higher mutation rates (both expansion and contraction) than those individuals with old onset age ($P = 0.0063$ using MANOVA). This partly explains why individuals with similar inherited repeat lengths can differ in terms of onset age (sometimes by as much as 30 years). Our results are consistent with the finding of (Swami et al. 2009) that the larger somatic expansions of the HD CAG repeat expansion in HD frontal cortex are significantly associated with an earlier age of disease onset. By quantifying the mutation rates in neurons and glia we show further that, in particular, the expansion rate in neurons explains some of the variance in age of onset not already explained by inherited repeat length. Together, inherited repeat length and the expansion rate in neurons explain almost 70% of the variance in age of onset. We note that the results concerning explained variance in age of onset due to expansion rates apply to extreme phenotypes, chosen to have large differences in age of onset and hence the effect may be exaggerated. However these results would hold in the general HD population if expansion rates correlate with age of onset in a similar manner.

We also show that the rates of expansion and contraction are highly correlated in cells with fast mutating repeat lengths (correlation coefficient $r = 0.974$) and in cells with slow mutating repeat lengths (correlation coefficient $r = 0.849$). These results suggest that expansion and contraction events are mutationally linked and could be considered as alternative outcomes of the same process.

As a potential modifier of age of onset, it would be highly informative to have a profile of somatic mosaicism and genotype at age of onset in the pathology related tissues, as it should be highly predictive of the contribution somatic variation makes towards age of onset and disease progression. However it is not feasible to obtain brain tissue for HD individuals at age of onset. We argue here that a viable alternative is mathematical simulation of the changes in repeat length from the inherited repeat length at birth to the age of onset using as input the parameters estimated from end stage samples. The similarity between the profiles at age of onset suggests that disease onset is strongly characterised by the repeat length distribution.

We next looked at the differences between the percentile levels prior to onset (between 5 and 15 years) and onset (summarised in Figure 6.8). The greatest difference during the 15 years prior to onset occurs among the very largest repeat lengths (95^{th} to 100^{th} percentile) which expand on average from 70 CAGs to 85 CAGs. This observation strongly suggests that the larger repeat lengths (over 80 CAGs) drive disease onset (degeneration and associated symptoms) in HD frontal cortex. Our findings support the scenario proposed by (Kaplan et al. 2007) whereby disease is triggered once a percentage of pathology related cells have expanded over a disease-specific threshold. We have shown that mutation rates are higher in individuals with young age of onset than in individuals with old age of onset. In the context of the age of onset model outlined above, these mutational differences provide an explanation as to why individuals with similar inherited repeat lengths can differ in terms of onset age (sometimes by as much as 30 years). Consequently, therapies aimed at keeping CAGs below 80 CAGs would be predicted to delay the onset of symptoms in the frontal cortex. The rate of expansion is around 40% lower among the old age of onset phenotype group than among the young age of onset phenotype group. If expansion rates could be knocked down by 40% in the young age of onset phenotype group then our model would predict that the onset of symptoms in the frontal cortex would be delayed by up to 30 years.

In the Kaplan *et al.* study, the disease-specific threshold was predicted to be 115 CAGs for Huntington disease based on clinical data for age of onset and repeat length. We found the threshold to be lower than this (80 CAGs) in frontal cortex. This most likely reflects differences in mutational rates within the brain (*i.e.* higher levels in striatum than in cortex) as shown by (Shelbourne et al. 2007). Kaplan *et al.* relied on available clinical data that typically sizes repeat length by the modal repeat length. The datasets we employ provide a fuller picture of the repeat length distribution by sizing single molecules within a sample of cells. Using this data we can estimate the total mutational dispersion (expansion plus contraction) as well as the mutational drift (expansion minus contraction) which provides much more information about the underlying process. Interestingly, mutational dispersion in the cells with fast mutating repeat lengths is more important, in terms of explaining age of onset ($R^2 = 21.6\%$), than mutational drift ($R^2 = 1.4\%$), see Table 6.3. Also, mutational dispersion is more individual-specific (0.095 ± 0.070 per CAG unit per year) than mutational drift (0.016 ± 0.008 per CAG unit per year). As well as quantifying underlying dynamics of age of onset, these measurements combined with inherited repeat length improve the predictive power of the age of onset model ($R^2 = 69.7\%$), see Table 6.4.

Higher levels of somatic instability are seen in the major targets of the pathogenic process, namely the striatum and cortex regions of the brain, making these important tissues for investigation. Lee *et al.* recently investigated tissue-specific trinucleotide repeat instability and demonstrated that multiple tissue factors reflect the level of somatic instability in different tissues (Lee et al. 2010). But interpretation of somatic mosaicism at the tissue level is hindered for some tissues by a complex cell composition which, we show, can result in overlapping profiles of repeat lengths and hence skewed, multi-modal distributions. Tissues, such as the brain, comprise different cells and sometimes different cell type ratios across the tissue. Cell activities other than replication are implicated in mutation and somatic instability. Therefore, tissue differences with respect to instability must, to some extent, be due to cell differences. Hence understanding somatic instability at the cell level, in particularly *in vivo*, is fundamental to understanding somatic instability at the tissue level and its potential role in modifying the age of onset. It has been shown that the level of HD gene expression is higher in neurons than in glia (Landwehrmeyer et al. 1995). Gonitel *et al.* demonstrated that neurons are distinguished from non-neuronal cells in both mice and humans by the expression of MSH3 (Gonitel et al. 2008) which, given its requirement for instability *in vivo* (van den Broek et al. 2002), would provide the environment for greater instability in neurons independent of pathology. High rates of L1 transposition in neurons reported by (Singer et al. 2010) also suggest differences in DNA repair and damage between cell types in the human brain.

Our approach to the quantification of the mutation rates underlying somatic mosaicism did not require the cells to be physically separated (Shelbourne et al. 2007, Lee et al. 2011), a task which may be infeasible or very time-consuming in some tissues. Estimation of mutation rates for fast and slow mutating cells established the greater role of neurons in disease onset and the pathogenic process. We also showed that mutation rates are individual-specific and explain some of the variance seen in age of onset not already explained by the inherited repeat length. Further, we revealed that the mutation rates for the different cell types are correlated within individuals, implicating an influence across cells which is also individual-specific. The ability to quantify rates of mutation in complex pathologically relevant tissues answers specific questions about the contribution of cell types towards somatic instability but, equally importantly, enables comparison between HD individuals in terms of individual-specific differences and the contribution of these differences towards disease onset and progression. These quantitative traits have applications for use with genome-wide studies to find the genetic factors (and environmental factors) that contribute towards disease. Furthermore, they have applications for use in evaluating therapies or drugs. The extent of the variation in the rates of mutation in individuals makes it highly likely that therapies/drugs targeting either the DNA or the RNA will also have variable rates of success. In future trials for HD and other triplet repeat diseases, the ability, through quantification, to benchmark individuals with respect to individual-specific factors would appear key to the evaluation and successful development of therapies and drugs.

6.5 Materials and methods

6.5.1 Project data

The data analysed in this study ((Swami et al. 2009)) was derived from a cohort of 48 individuals with inherited HD expansions between 40 and 48 CAGs, determined from cerebellar DNA which has been shown to be somatically stable (Kennedy et al. 2003). Swami *et al.* identified 24 individuals with an extremely young age of onset and 24 with an extremely old age of onset, matched for their mutant and normal inherited repeat lengths but with mean age of onset differing by approximately 30 years.

Small pool-PCR analysis was used to amplify the genomic DNA isolated from frontal cortex, dissected from brains obtained at autopsy of the 48 individuals, using HD gene CAG repeat-specific primers to obtain a profile of HD CAG repeat lengths (visualised in a histogram format in Figure 6.1). For each sample, the length of the HD CAG repeat of 100 or more mutant alleles was determined. There was no significant difference in the number of normal and mutant HD alleles amplified. This indicates the absence of bias in the size of allele amplified and confirms that the targets were single molecules. As further demonstrated by (Gonitel et al. 2008), through replicated experiments, samples of this size can be considered sufficient to reflect the total population of mutant alleles.

6.5.2 Mathematical model for complex tissue

We hypothesise that the distribution of CAG repeat lengths seen in frontal cortex in end-stage HD individuals arises from two cell types, one with repeat lengths that mutate faster than the other. We therefore obtain the probability distribution function for a mixed cell sample, P_n^{f+s}, by combining a distribution function for cells with fast mutating repeat lengths, $P_n^f\left(t,\theta^f,n_0\right)$, with a distribution function for cells with slow mutating repeat lengths, $P_n^s\left(t,\theta^s,n_0\right)$. Thus

$$P_n^{f+s} = w^f P_n^f + \left(1-w^f\right) P_n^s, \tag{6.1}$$

where w^f is the unknown percentage of cells with fast mutating repeat lengths to be inferred from the data, t is the age when the sample was taken, θ^f and θ^s are the model parameters described below and n_0 is the inherited repeat length.

The corresponding master equations (see Chapter 2, Section 2.4, for an explanation of how master equation are derived) are

$$\begin{aligned}
\dot{P}_n^f\left(t\right) &= -\left(\lambda^f R_n + \mu^f R_n\right) P_n^f\left(t\right) + \lambda^f R_{n-1} P_{n-1}^f\left(t\right) + \mu^f R_{n+1} P_{n+1}^f\left(t\right), && (6.2)\\
\dot{P}_n^s\left(t\right) &= -\left(\lambda^s R_n + \mu^s R_n\right) P_n^s\left(t\right) + \lambda^s R_{n-1} P_{n-1}^s\left(t\right) + \mu^s R_{n+1} P_{n+1}^s\left(t\right), && (6.3)
\end{aligned}$$

where λ^f, λ^s are the expansion rates per CAG repeat unit per year for fast mutating repeat lengths and slow mutation repeat lengths, respectively, and μ^f, μ^s are the contraction rates per CAG repeat unit per year for fast mutating repeat lengths and slow mutation repeat lengths, respectively. R_n is the length specific factor defined in Equation (5.2), Section 5.5.2, Chapter 5, and can be considered as the biophysical capacity of a repeat length to undergo mutation. We showed in Chapter 5 that this factor influences small alleles. Hence it is pertinent to data arising from HD individuals.

Parameter estimation

We use likelihood methods to carry out parameter estimation, see Chapter 2, Section 2.5.4, for further details. The likelihood is defined to be the probability that a repeat length (from either cell type) has reached the length observed given the model and its parameters. We can solve Equations (6.2) and (6.3) numerically in order to obtain the probability distribution function components $P_n^f(t)$ and $P_n^s(t)$, which are the respective probabilities that a fast mutating repeat length is n at time t and a slow mutating repeat length is n at time t. The likelihood $L^{[i]}$ is then the product over all the data $d_j^{[i]}$, which denotes the repeat length for the jth observation from individual i, of the probability function$P_{d_j^{[i]}}^{f+s}$. This gives the likelihood for individual i,

$$L^{[i]} = \prod_j P^{f+s}{}_{d_j^{[i]}}, \qquad (6.4)$$

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

$$L = \prod_i L^{[i]}. \qquad (6.5)$$

We obtain the maximum value of the likelihood by evaluating the likelihood over a broad parameter space described in Table 6.1. For further statistical analysis, it was useful to have point estimates for the parameters. These were taken to be the maximum likelihood values. The complexity of the computation was reduced by creating libraries for the calculations common to each dataset (P_n^f and P_n^s for all parameter values) and calling these over the cell weights to finalise the likelihood calculation.

6.5.3 Simulations deriving from the parameter estimations

The maximum likelihood approach provides point estimates of the parameter values which best fit the data. We use these parameter estimates for cells with fast mutating repeat lengths and for cells with slow mutating repeat lengths, along with the inherited repeat length, to simulate, using the Gillespie algorithm adapted for our specific stochastic process (Renshaw 1991), the time dependent distribution for 100,000 cells under the full distribution model at age of onset. We assume that 35% of the cells have fast mutating repeat lengths and 65% of the cells have slow mutating repeat lengths.

Chapter 7

Availability of two DNA samples from the same individual at different points in time is better for predicting age of onset and validates the model

7.1 Abstract

Myotonic dystrophy type 1 (DM1) is a multisystemic disorder characterised by the presence of myotonia (slow relaxation of the muscles after voluntary contraction or electrical stimulation) followed by progressive weakness and wasting of distal limb and facial muscles, cardiac conduction defects, cataracts, frontal balding and testicular atrophy. The observable characteristics of patients (or phenotype) fall into four broad clinical forms: mild or late onset disease; classic adult onset; juvenile onset and congenital (onset at birth). Currently clinical diagnosis is based on a measure of repeat length from blood cells but variance in modal length only accounts for between 20 - 40% of the variance in age of onset and, therefore, is not predictive. Hence the International Myotonic Dystrophy Consortium have recommended that patients are not offered prognostic information based on the current test. Recently, Morales *et al.* showed that progenitor allele length, estimated using the lower bound of the distribution as seen with small pool PCR, significantly improves the inverse correlation with age of onset over the traditional modal length measure. Higham *et al.* have developed a mathematical approach to inferring inherited repeat length from blood DNA samples. However the estimates of inherited repeat length have wide credible intervals. New data now provides blood samples for 40 DM1 individuals at two time points. Using a mathematical approach we infer inherited repeat length from the combined blood samples. We show that inherited repeat length and the mutation rates underlying repeat length instability in blood, inferred from two samples rather than one, are better predictors of age of onset. These results support other findings that inherited repeat length and somatic instability are modifiers of disease onset and progression. Our results are a step towards providing better prognostic information for DM1 individuals and their families. They may also lead to better predictions for drug/therapy response which is emerging as key to successful clinical trials.

7.2 Introduction

So far, the mathematical models developed for DM1 (Chapters 4 and 5) have been fitted to DNA blood samples taken from DM1 affected or at risk individuals on one occasion. This occasion was most likely initiated by disease onset either in that individual or in a related individual. The DNA blood samples assessed cover a wide range of inherited repeat lengths and ages at sampling and capture the variation seen in repeat length distributions across a large cohort. We attribute this variation to differences between individuals in terms of inherited repeat length, age at sampling and individual-specific rates of mutation. Using a Bayesian context (Chapters 4 and 5) we inferred the value and credible interval of the model parameters (inherited repeat length and rates of mutation) for each DM1 individual (see posterior probability distributions in Figure 4.3). In Chapter 2, Section 2.7, we applied the inference method to a synthetic dataset (simulated from known parameter values) to assess how well the method inferred parameter values. The slanted shape and size of the credible interval for inherited repeat length and rates of mutation (Figures 2.3D and 2.3G) suggests that one sample (between 100 and 200 cells) does not provide enough information about the underlying process to distinguish clearly between inherited repeat length and rates of mutation. The intuitive explanation is that similar repeat length distributions arise from different scenarios. For example, the repeat length distribution from an individual with an inherited repeat length of 100 CTGs and a high rate of mutation may closely resemble the repeat length distribution from an individual with an inherited repeat length of 150 CTGs and a lower rate of mutation. One way to distinguish between inherited repeat length and rates of mutation would be to observe the mutational process at another point in time. Observing the process at an earlier point in time should provide more information about the inherited repeat length and observing the process again, at a later point in time, should provide more information about rates of mutation over time. In particular, it is not known whether rates of mutation are constant throughout the lifetime of an individual or whether they vary over time. Combining both observations would be expected to provide more information about the trajectory of the repeat length distribution and hence overall more information about the inherited repeat length and the rates of mutation.

Further blood samples are now available from a subset (25) of the original cohort of 142 DM1 affected or at risk individuals. Pairs of blood samples, taken from an individual at different points in time, are also available from 15 Scottish individuals recruited to a new study to investigate genetic variation. These pairs of samples (40 in total) provide an opportunity to assess repeat length changes within an individual over time. As discussed above, two samples should improve our ability to distinguish between the contribution of inherited repeat length, age and individual-specific rates of mutation. Hence we hypothesise that two samples will provide more information about the underlying mechanism, reduce the level of uncertainty in the parameter estimation and improve the ability of these parameters to predict disease onset and progression. We address these hypotheses by fitting models M_{6b} (as described in Chapter 4) and M_α with length-specific effect (as described in Chapter 5) to the data from the earlier time point, the data from the later time point and the combined data from both time points. We recall that model M_α was found to be better than model M_{6b} at describing the small repeat lengths (under 200 CTGs) but the same as model M_{6b} at describing repeat lengths above 200 CTGs. As we are now considering the evolution of repeat length over several years, small repeat lengths will be relevant to many of these individuals.

7.3 Results

7.3.1 First and second time point samples are consistent with samples from the large cohort study (142 DM1 individuals)

As expected from our analysis of repeat length distribution in a large cohort of DM1 individuals with different inherited repeat lengths and ages at sampling, the individual repeat length distributions disperse over time with an expansion bias. Figures 7.1, 7.2 and 7.3 show repeat length distributions at different time points for representative individuals. Levels of expansion depend on repeat length and age, so clear differences are seen among individuals who very likely inherited long repeat lengths (*e.g.* Figure 7.3) but also when the time between sample collection is high (*e.g.* 19 years, see Figure 7.2).

In Chapter 5, we compared the variance to mean ratio in 142 DM1 affected or at risk individuals, taking into account age at sampling, with the predicted variance to mean ratio under the thresholded model M_{6b} (see Figure 5.1). This analysis provided a useful summary of the data, highlighting the highly individual nature of somatic variation. A subset of outliers (15th percentile) was significantly dominated by individuals with small inherited repeat lengths (less than 100 CTGs). The analysis enabled identification of other individuals whose effective length is less than their actual length, possibly due to anomalies in their repeat length tracts. Returning to this type of analysis, we find that both time points broadly follow the expected variance-to-mean trajectory, see Figure 7.4. There are eight samples from four individuals who lie in the 15th percentile along with the individuals discussed above. If the samples are consistent with the model, having corrected for repeat length and

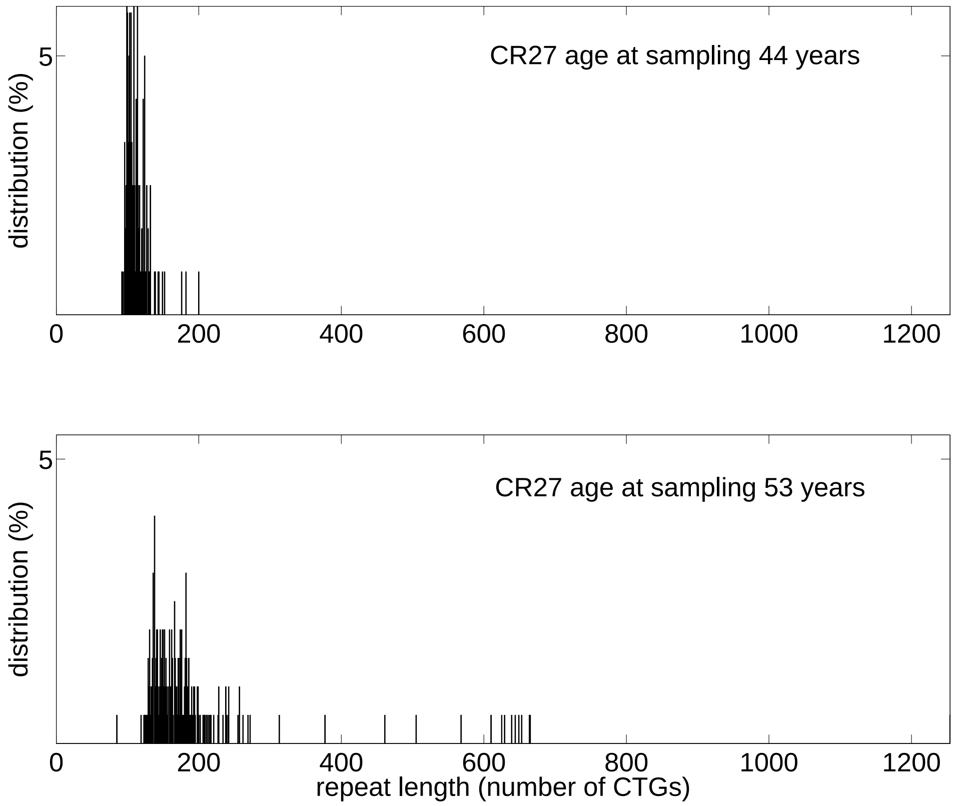


Figure 7.1: **Comparison between repeat length distribution at the first time point (top) and at the second time point (bottom) for individual CR27.**

age, we would expect the residuals (actual results minus expected results) to be normally distributed. Examination of the residual variance-to-mean ratio among the 36 individuals with two samples (excluding the four outliers) show that there is no significant correlation between this residual and age when the sample was taken ($R^2 = 0.6\%$, $P = 0.52$), but that there is a significant correlation between this residual and mean repeat length ($R^2 = 17.75\%$, $P = 2.28 \times 10^{-4}$), see Figure 7.5. Individual differences in the variance to mean ratio over time are mostly (the exceptions are discussed below) consistent with increases in line with expectations (Figure 7.4). This suggests that, under the model, the parameters may change slightly as repeat length increases with the effect of reducing the variance to mean ratio. Indeed, we observed that the variance-to-mean ratio went down between the first and second time points for some individuals with repeat lengths over 1500 CTGs, see Figure 7.4. However as this observation may be the result of missing data, either because the sample did not capture relatively rare long repeats or because the experimental technique was not able to amplify or measure the long repeats, these samples will be reinvestigated in the laboratory.

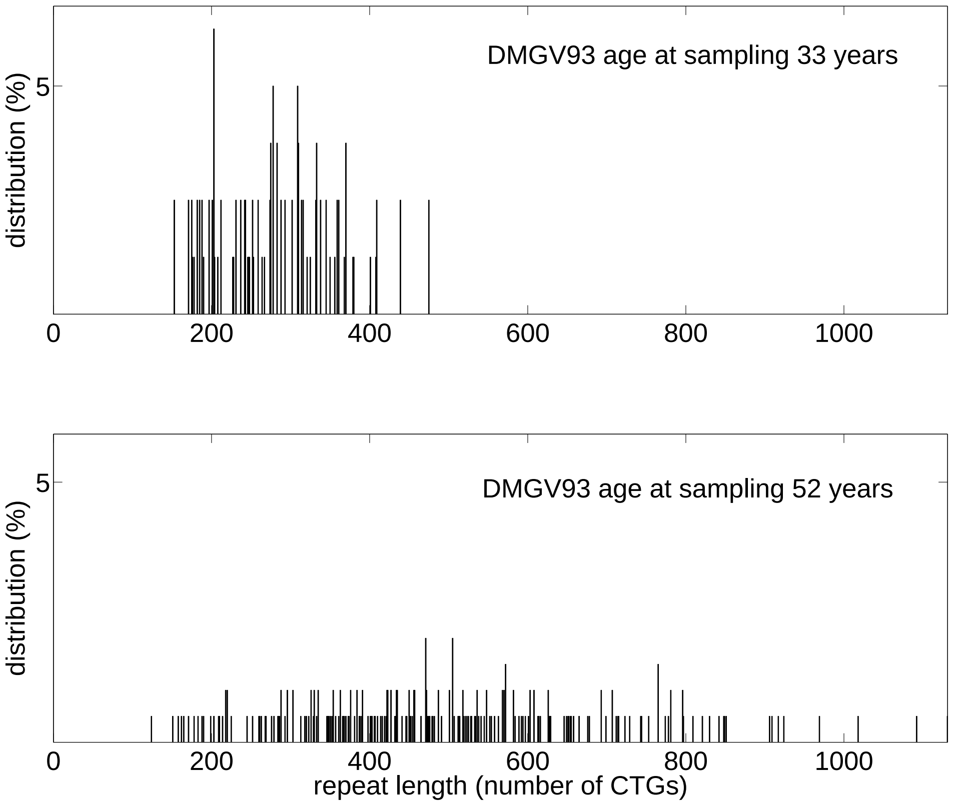


Figure 7.2: **Comparison between repeat length distribution at the first time point (top) and at the second time point (bottom) for individual DMGV93.**

7.3.2 Combining first and second time points reduces uncertainty in the parameter estimation

The model parameters, under model M_α, are expansion per CTG unit per year, λ, contraction per CTG unit per year, μ, the length parameter measured in CTG units, α, and the inherited repeat length, n_0. Model M_α was fitted to the data and the parameter values were estimated using a Bayesian inference approach (see Section 7.5.2) which involved choosing prior ranges for the model parameters (Table 7.1). As this inference approach is based at the cell level, the samples can be treated separately or combined (see Section 7.5.2). Hence it was possible to obtain posterior probability distributions for each parameter for the first earlier sample, the second later sample and the combined sample. The peak of the posterior probability distribution indicates the most likely parameter value and the spread of the distribution defines a credible interval associated with this estimate. Comparing the posterior probability distributions for the first sample, the second sample and the combined sample, the credible interval tends to be wider for the second time point sample than the first time point sample but narrower than either for the combined sample, see represen-

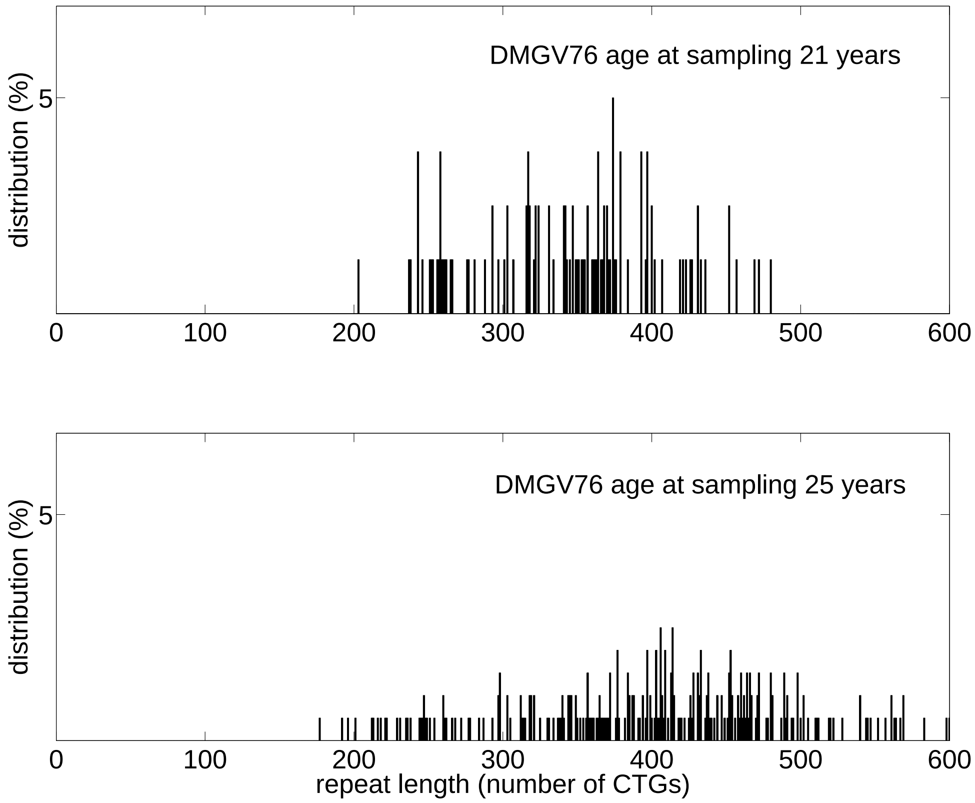


Figure 7.3: **Comparison between repeat length distribution at the first time point (top) and at the second time point (bottom) for individual DMGV76.**

tative individuals in Figures 7.6 and 7.7. This suggests that there is more information about the underlying mutational process and hence model parameters in the earlier first sample than the later second sample. This finding is consistent with a basic feature of the computational approach that is independent of the parameter values – the variance in the length distribution increases over time, and hence the accuracy in recovering the mean from a fixed number of sample points degrades. We can partially compensate for this by increasing the sample size to capture increased variance at the second time point. Our finding that combining samples from two time points further increases the information, reducing the uncertainty in the parameter estimation, provides strong evidence that the model is capturing time-dependent effects. This effect, of reducing uncertainty, is more clearly seen by comparing the joint posterior probabilities for the first sample, second sample and combined sample, representative results are given in Figures 7.8 and 7.9.

There is virtually no instability in blood at birth, even in those with the congenital form of the disease (Wong et al. 1995, Wong & Ashizawa 1997, Martorell 1997, 1998), so mean repeat length is expected to be the inherited repeat length. It is useful to visualise the expected path of mean repeat length over time as it passes from the inherited repeat length through the mean repeat length of

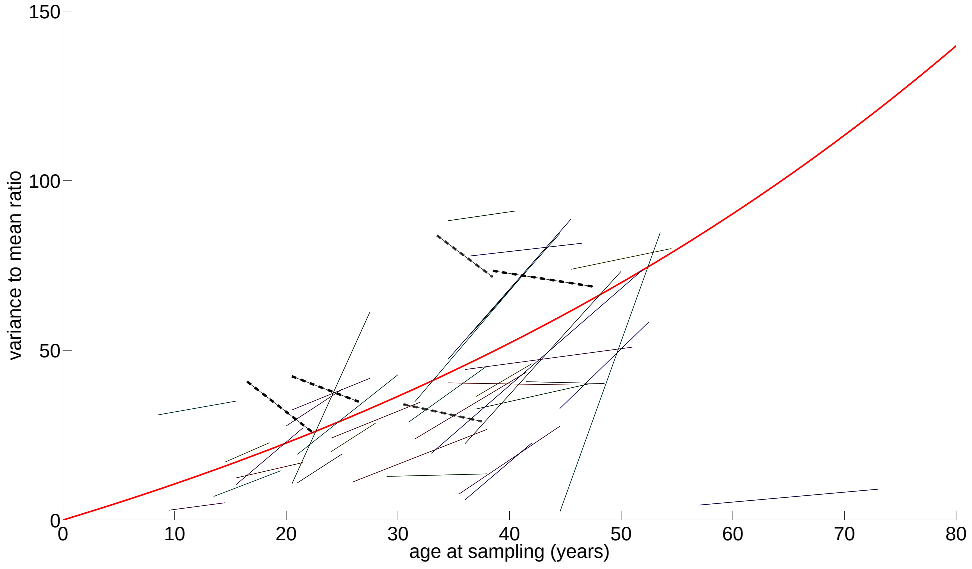


Figure 7.4: **Individual variance to mean ratio trajectories (short solid lines).** The short lines join the variance to mean ratio at time point 1 (left end) to the variance to mean ratio at time point 2 (right end). There are five individuals with variance to mean ratios that decrease over time (short dashed lines). The trajectories of the other 35 individuals broadly follow the expected trajectory under model M_{6b} (long solid curve).

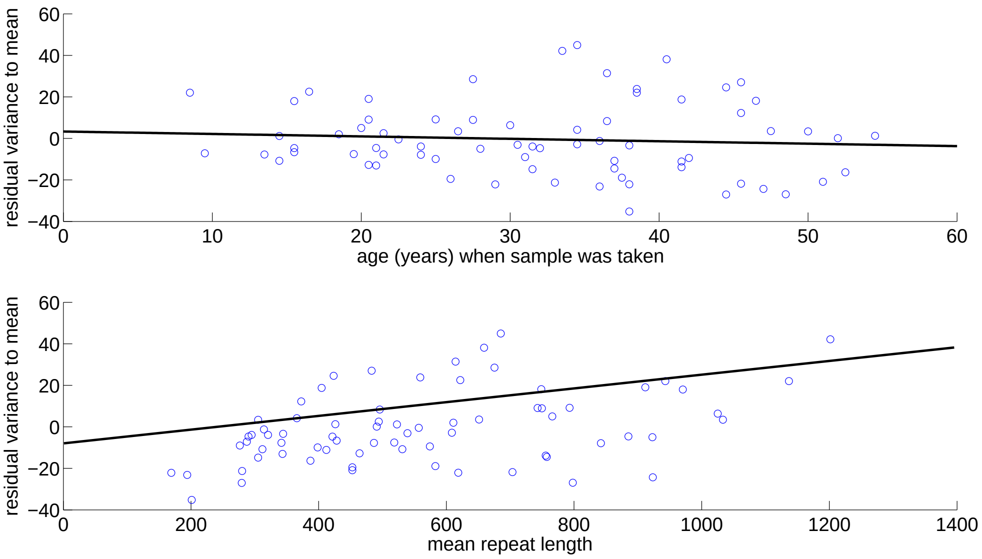


Figure 7.5: **Residual variance over mean (72 samples)** : **Upper panel:** by age when sample was taken, linear regression fit $R^2 = 0.6\%$, $P = 0.5168$; and **Lower panel:** by mean repeat length when sample was taken, linear regression fit $R^2 = 17.75\%$, $P = 2.2782 \times 10^{-4}$.

the first time sample and the second time sample, see Figures 7.10 and 7.11 for two representative individuals. Comparing the expected mean repeat length path based on the first sample, the second sample and the combined sample, the credible interval (5^{th} to 95^{th} percentile) based on the combined sample is generally narrower than either that based on the first or second samples. Generally,

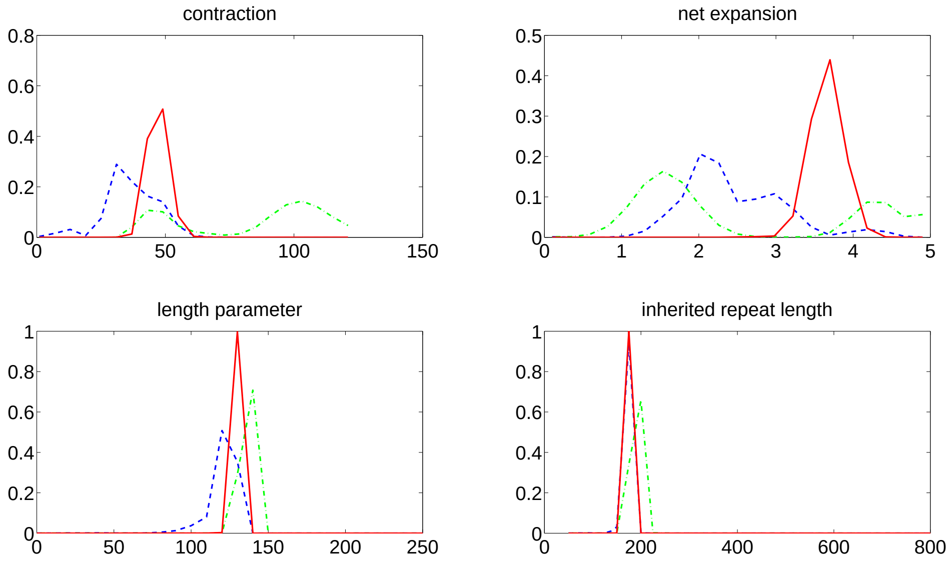


Figure 7.6: **Model parameter posterior probability distributions for representative individual DMGV4. Upper left panel:** contraction rate per repeat unit per year, **Upper right panel:** net expansion rate per repeat per year, **Lower left panel:** length parameter (number of repeat units) and **Lower right panel:** inherited repeat length (number of repeat units): based on first time point (dashed line), based on second time point (dash dot line) and combined samples (solid line).

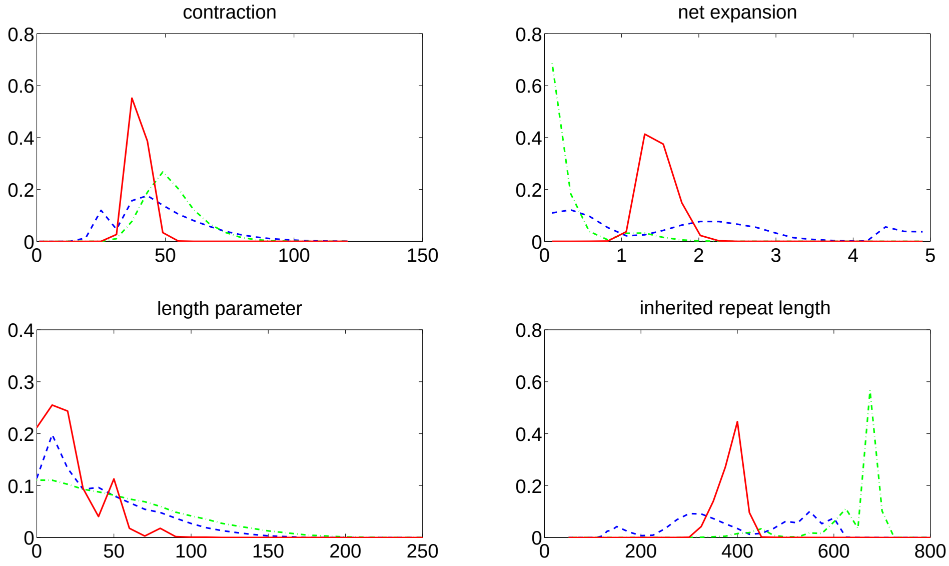


Figure 7.7: **Model parameter posterior probability distributions for representative individual CR022. Upper left panel:** contraction rate per repeat unit per year, **Upper right panel:** net expansion rate per repeat per year, **Lower left panel:** length parameter (number of repeat units) and **Lower right panel:** inherited repeat length (number of repeat units): based on first time point (dashed line), based on second time point (dash dot line) and combined samples (solid line).

the credible interval for the first sample is narrower than the credible interval for the second sample.

These results suggest that the prediction for inherited repeat length is more robust when based on

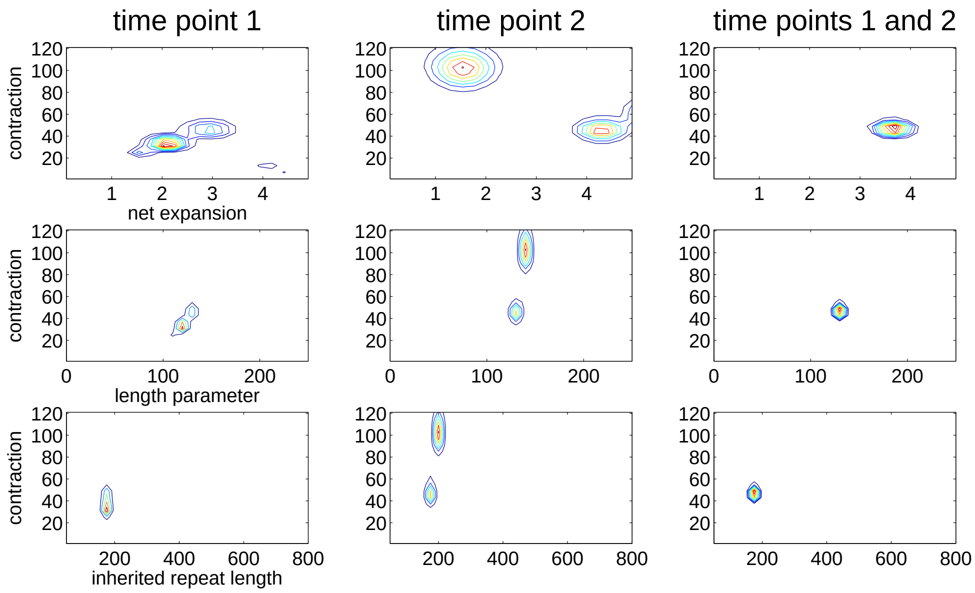


Figure 7.8: **Joint posterior probability distributions for representative individual DMGV4. Upper row:** contraction rate per 100 CTG repeat units per year (horizontal axis) and net expansion rate per 100 repeat units per year (vertical axis), **Middle row:** contraction rate per 100 CTG repeat units per year (horizontal axis) and length parameter (number of repeat units), **Lower row:** contraction rate per 100 CTG repeat units per year (horizontal axis) and inherited repeat length (number of repeat units): based on first time point (**Left column**), based on second time point (**Middle column**) and combined samples (**Right column**).

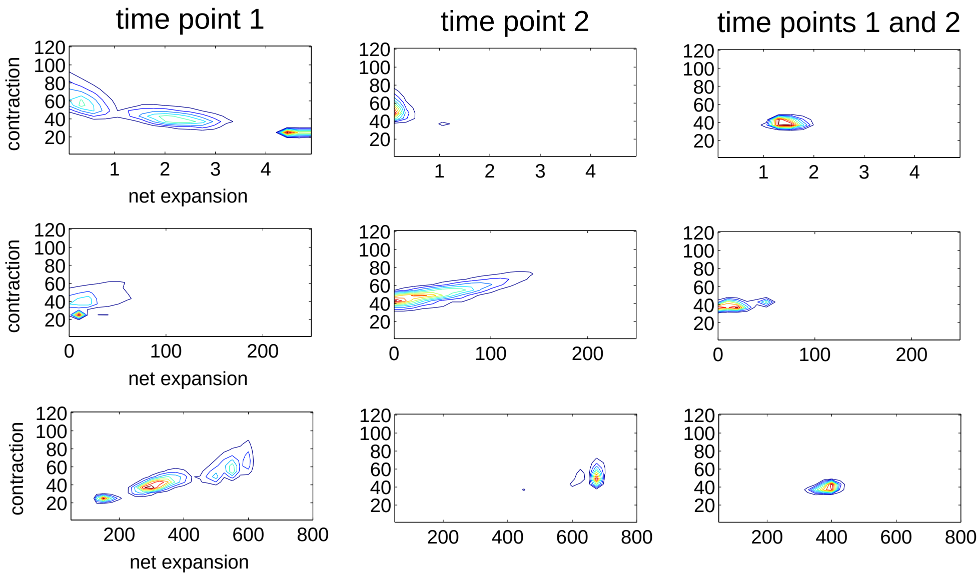


Figure 7.9: **Joint posterior probability distributions for representative individual CR022. Upper row:** contraction rate per 100 CTG repeat units per year (horizontal axis) and net expansion rate per 100 repeat units per year (vertical axis), **Middle row:** contraction rate per 100 CTG repeat units per year (horizontal axis) and length parameter (number of repeat units), **Lower row:** contraction rate per 100 CTG repeat units per year (horizontal axis) and inherited repeat length (number of repeat units): based on first time point (**Left column**), based on second time point (**Middle column**) and combined samples (**Right column**).

two samples rather than one and that the prediction for inherited repeat length is likely to be better the earlier samples are taken.

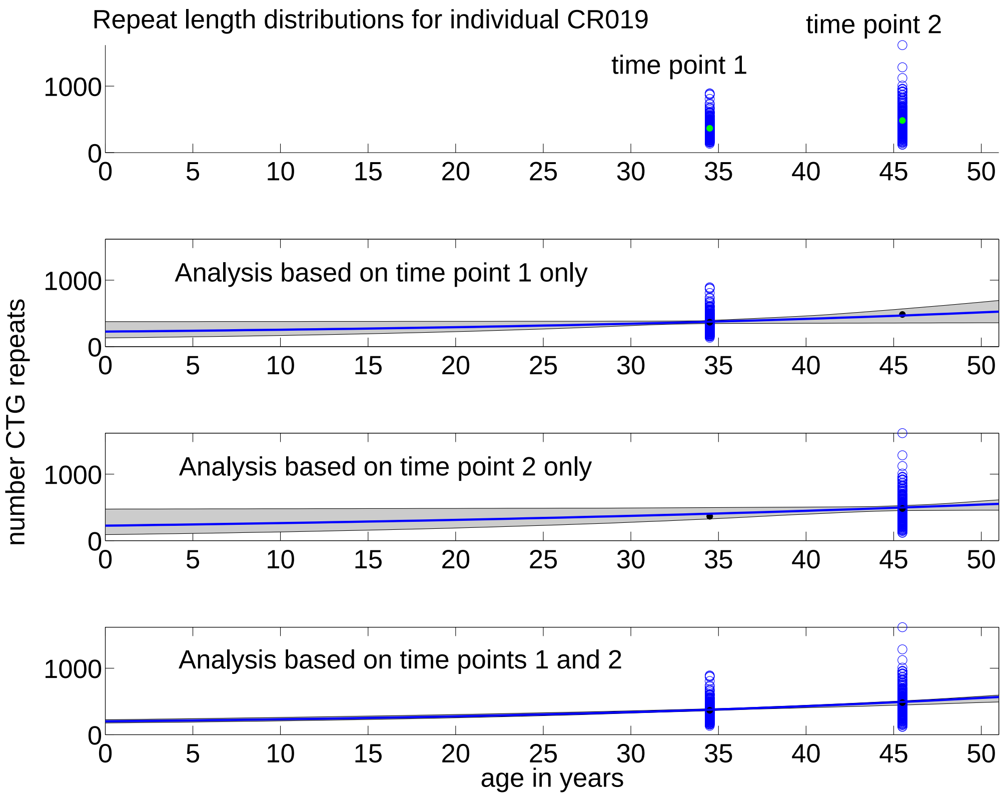


Figure 7.10: **Expected mean repeat length over time fitted to two samples from representative individual CR019. Upper panel:** repeat length distribution in first sample at time taken and repeat length distribution in second sample at time taken with mean repeat length (light grey). **Second panel:** maximum likelihood mean repeat length over time based on first sample with $5-95^{th}$ percentile credible level shaded in grey. **Third panel:** maximum likelihood mean repeat length over time based on first sample with $5-95^{th}$ percentile credible level shaded in grey. **Lower panel:** maximum likelihood mean repeat length over time based on combined sample with $5-95^{th}$ percentile credible level shaded in grey. The credible bands for the expected mean repeat length were computed from the posterior probability distributions at each time point.

7.3.3 Model parameters estimated from two time points are better predictors of age of onset

As well as inherited repeat length, (Morales et al. 2012) showed that somatic variation (adjusted for age at sampling and inherited repeat length) also explained some of the variance in age at onset not already explained by inherited repeat length. This result suggests that somatic instability, along with inherited repeat length, are modifiers of disease severity. Based on this finding, we hypothesise that the model parameters will also explain some of the variance in age at onset. We obtained point

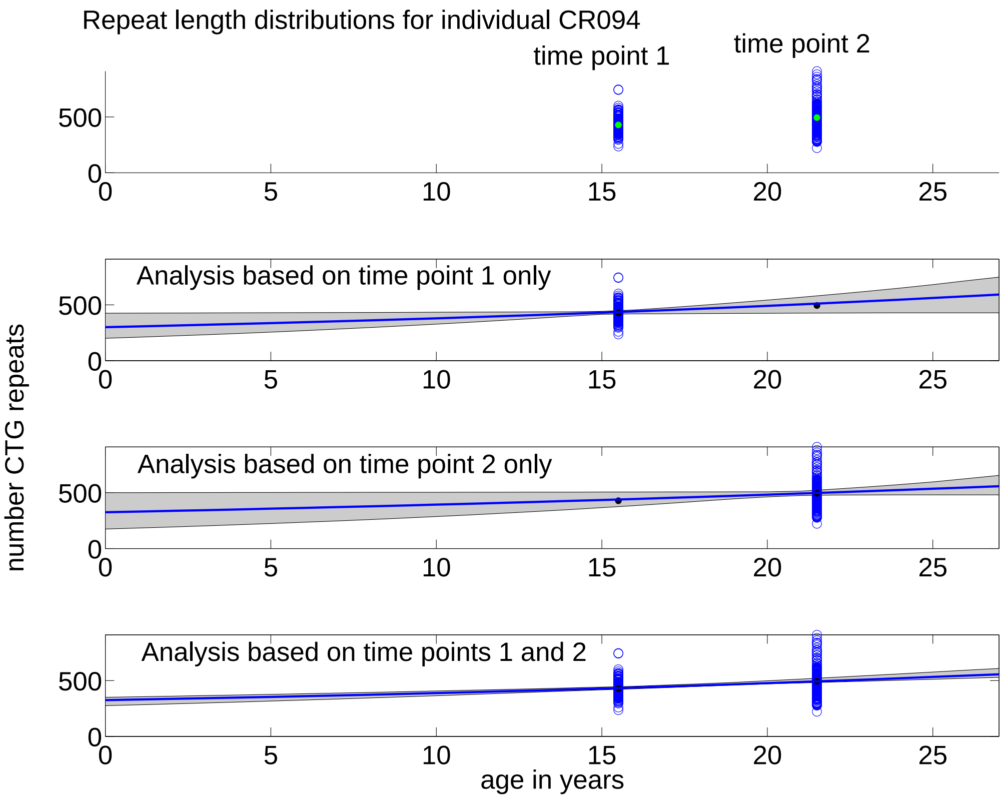


Figure 7.11: **Expected mean repeat length over time fitted to two samples from representative individual CR094. Upper panel:** repeat length distribution in first sample at time taken and repeat length distribution in second sample at time taken with mean repeat length (light grey). **Second panel:** maximum likelihood mean repeat length over time based on first sample with $5-95^{th}$ percentile credible level shaded in grey. **Third panel:** maximum likelihood mean repeat length over time based on first sample with $5-95^{th}$ percentile credible level shaded in grey. **Lower panel:** maximum likelihood mean repeat length over time based on combined sample with $5-95^{th}$ percentile credible level shaded in grey.

estimates for the model parameters from the maximum likelihood solution for the first sample, the second sample and the combined sample and tested this hypothesis on all DM1 individuals with two samples for whom age at onset was available ($N = 36$) for several parameter combinations using linear regression analysis, see Table 7.2. The results based on the first time point sample, the second time point sample and the combined sample were compared using the adjusted R^2 statistic which takes into account the number of parameters and hence allows comparison between predictive models with different numbers of parameters. In all cases, the adjusted R^2 statistic is higher for the combined sample than for the first sample or the second sample. Generally (for the results where $P < 0.004$) the adjusted R^2 statistic is higher for the first sample than the second sample. These results confirm our hypothesis that there is more information in the combined example resulting in better predictors of age of onset.

In terms of which parameters or combination of parameters are best at predicting age of onset,

Model parameters	**Prior ranges**
Contraction rate per CTG unit per year (μ)	0 – 0.12
Net expansion rate per CTG unit per year ($\varphi = \lambda - \mu$)	0.01 – 0.05
Length parameter, number of CTG units (α)	0 – 250 (or n_0 if n_0 is less than 250)
Inherited repeat length, number of CTG units (n_0)	50 – 800

Table 7.1: **Prior ranges for parameter estimation for analysis with two time points.**

Age at onset versus **N=36**	**Sample earlier time point t1**	**Sample later time point t2**	**Combined Sample**
Inherited repeat length	adj R^2 = 25.0% (*P=0.0011*)	adj R^2 = 14.0% (*P=0.0142*)	adj R^2 = 34.7% (*P=9.3963E-005*)
Expansion	adj R^2 = -2.4% (*P=0.6786*)	adj R^2 = 1.9% (*P=0.2061*)	adj R^2 = 14.2% (*P=0.0135*)
Contraction	adj R^2 = -2.4% (*P=0.6649*)	adj R^2 = 1.9% (*P=0.2053*)	adj R^2 = 14.1% (*P=0.0138*)
Net expansion	adj R^2 = -1.9% (*P=0.5659*)	adj R^2 = -2.0% (*P=0.5761*)	adj R^2 = 1.9% (*P=0.2040*)
Expansion, inherited repeat length	adj R^2 = 28.4% (*P=0.0015*)	adj R^2 = 11.5% (*P=0.0508*)	adj R^2 = 32.9% (*P=5.2203E-004*)
Contraction, inherited repeat length	adj R^2 = 28.8% (*P=0.0014*)	adj R^2 = 11.4% (*P=0.0514*)	adj R^2= 33.1% (*P=4.9559E-004*)
Expansion, contraction, inherited repeat length	adj R^2 = 29.4% (*P=0.0026*)	adj R^2 = 18.4% (*P=0.0234*)	adj R^2 = 48.2% (*P=2.2070E-005*)
Net expansion, inherited repeat length	adj R^2 = 28.0% (*P=0.0017*)	adj R^2 = 20.4% (*P=0.0087*)	adj R^2 = 49.6% (*P=4.7200E-006*)

Table 7.2: **Comparison of the relationship between age of onset, inherited repeat length and mutation rates for myotonic dystrophy type 1.** Under three scenarios: 1. when only the first time point sample was available (column 2), 2. when only the second time point was available (column 3) and 3. when both time points were available (column 4).

inherited repeat length and expansion are inversely correlated with age at onset whilst contraction is positively correlated with age of onset. Consequently, age of onset is best explained (adjusted $R^2 = 49.6\%$, $P = 4.72\times10^{-6}$) by a linear model with inherited repeat length and the net expansion rate (expansion rate minus contraction rate), see Table 7.2. Under this model, age of onset would be expected to decrease as inherited repeat length increases. In individuals with the same inherited repeat length, age of onset would be expected to be lower in the individual with the lower net expansion rate than in the individual with the higher net expansion rate. Over the prior parameter values, inherited repeat length is expected to have more impact on age of onset (between 0 and 40 years) than net expansion rate (between 0 and 25 years) but both are considerable (Figure 7.12).

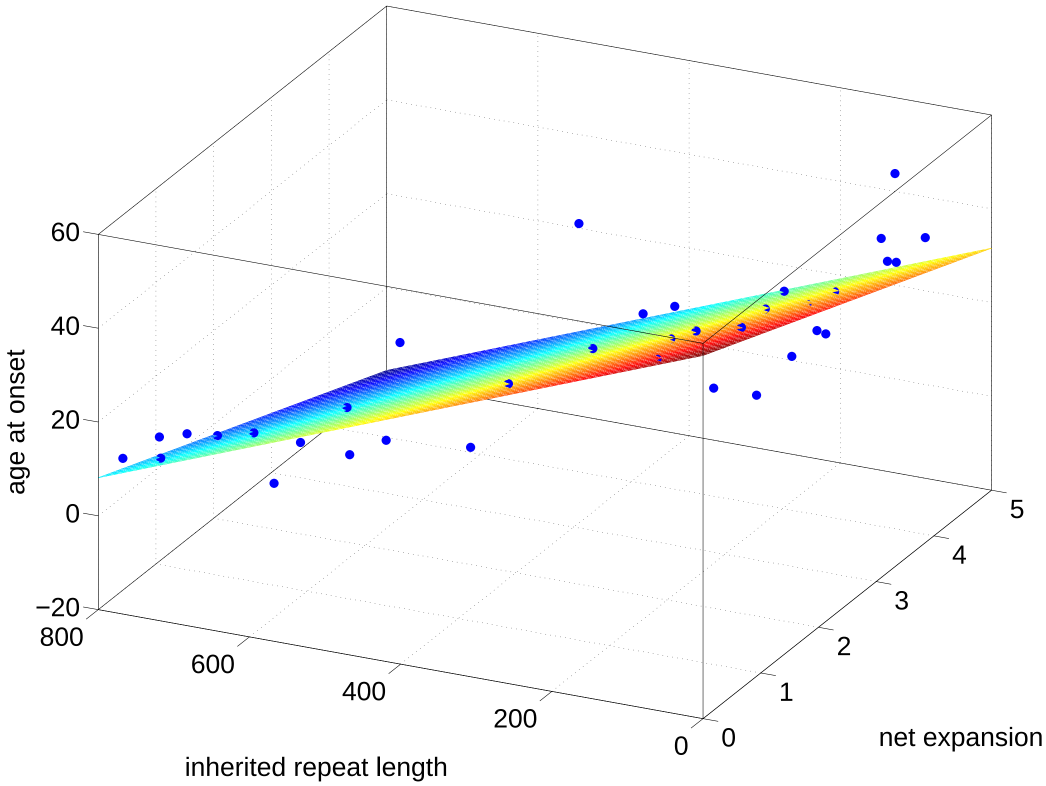


Figure 7.12: **The estimated inherited repeat length and the net expansion rate are modifiers of age of onset.** The relationship between inherited repeat length (number of CTG repeat units) and net expansion rate per CTG unit per year and age of onset (years) ($N = 36$). The surface has been fitted to the data using linear regression analysis (Table 7.2).

Having shown that inherited repeat length and both expansion and contraction are considerations for predicting age of onset, we tested whether the summary statistic given by the variance-to-mean ratio at age of sampling, which captures the effect of inherited repeat length as well as expansion and contraction, is also a potential indicator of age of onset. The results were positive (adjusted $R^2 = 55\%$, $P = 3.14 \times 10^{-13}$, $N = 72$).

7.4 Discussion

We now have second DNA blood samples taken at a later point in time for 25 DM1 individuals from the original cohort of 145 DM1 individuals. We also have pairs of DNA blood samples from 15 Scottish DM1 individuals recruited for a DM1 genetic variation study. All these samples have been sized using small pool PCR and provide an excellent opportunity to investigate repeat length changes over time within an individual. We compared these samples, in terms of their variance-to-mean ratio taking into account age at sampling, with the original samples. Analysis of the residual variance-to-mean for these samples suggests that time point one and time point two are consistent with the previous samples and that our assumption that the model parameters are fixed for individuals over time is reasonable.

The model parameters were estimated under three scenarios: 1. when only the first time point sample was available, 2. when only the second time point was available and 3. when both time points were available. The results showed that the credible intervals for the parameter values are narrower for scenario 1. than scenario 2. but narrower again for scenario 3. These results support the intuitive notion that there is more information about the model parameters in samples taken at earlier time points, when there is less deviation from the inherited repeat length, than in samples taken at later time points. Two samples tracking repeat length distribution over time in an individual provided even more information about the underlying process. The improved parameter fit over two time points also provides further validation for the models as a quantitative description of the underlying biological mechanisms.

We hypothesized that the model parameters inferred from the data would explain some of the variance seen in age of onset not already explained by inherited repeat length. As discussed in Chapter 5, blood is not the tissue where DM1 manifests itself, but under the assumption that levels of instability in blood may be correlated to levels of instability in muscle and other tissues where DM1 does manifest itself, the model parameters inferred from blood DNA may still explain age of onset. As instability is easier to measure in blood than muscle due to the lower repeat length levels present, this result would support a prognostic role for blood DNA. Our results showed that inherited repeat length, expansion rate and contraction rate, inferred from two blood DNA samples were predictive of age of onset (adjusted $R^2 = 49.6\%$, $P = 4.72 \times 10^{-6}$).

Analysis of the relationship between age of onset and the model parameters inferred from blood DNA suggests that differences in the estimated net expansion rate in blood could explain, on average, up to 25 years difference in age of onset between individuals with the same inherited repeat length. This result supports previous findings in DM1 (Morales et al. 2012) and Huntington disease (HD) (Swami et al. 2009) that levels of somatic instability modify disease onset and progression. However, for several reasons, we cannot propose a mechanistic cell based age of onset model for DM1 based on blood as we did for HD brain (Chapter 5). We do not know the the mutation rates (and the relative importance of expansion and contraction) in DM1 disease related tissues. However it is very likely that rates of mutation are correlated between cells and tissues within an individual – we have shown this to be the case for blood and buccal cells in DM1 (Chapter 4) and neurons and glia in HD brain (Chapter 5). The power of the parameters inferred from blood to predict age of onset also suggests that this may be the case. However as earlier work shows (Chapters 4 and 5) there are disease and cell differences between the proportion of expansion events and contraction events. In HD brain, we suggested that onset in the frontal cortex may be triggered by the neurons with longer repeats, the important factor being the expansion rate in neurons rather than the net expansion rate in neurons or glia. It is not known whether onset in other diseases or tissues is triggered in a similar manner. In DM1 muscle, it is unlikely that onset is triggered by a threshold. Very large repeats are seen in muscle before onset and so the threshold would have to be unrealistically high. It is more likely that onset is dependent on average repeat length and this is supported by our finding in this chapter that net expansion rather than expansion explain more of the variance in age of onset. If mutation rates in tissues directly involved in the DM1 pathology were available, future work could involve simulation of repeat length distribution at age of onset hence providing an approach to quantify the pathological drivers of disease onset and progression.

As discussed in Chapter 1, myotonic dystrophy type 1 is a multisystemic disorder characterised by the presence of myotonia. The observable characteristics of patients (or phenotype) fall into four broad clinical forms: mild or late onset disease; classic adult onset; juvenile onset and congenital (Harper 1989). Variance in modal length only accounts for between 20 - 40% of the variance in age of onset (Mladenovic et al. 2006, Perini et al. 1999, Marchini et al. 2000) and, therefore, is not an accurate predictive tool. The improvement in the predictive power of the estimated parameters, inherited repeat length and net expansion, based on the combined samples (adjusted $R^2 = 49.6\%$, $P = 4.72 \times 10^{-6}$) compared with those based on the first sample (adjusted $R^2 = 28.0\%$, $P = 0.0017$) or on the second sample (adjusted $R^2 = 20.4\%$, $P = 0.0087$), along with narrower credible intervals, suggests that the parameters based on two samples are more robust and more useful for potentially providing patients with better prognostic information. These estimates are also potential biomarkers for onset and progression and could be used in a clinical context to assess treatment response which, given the variable nature of DM1 and HD, is predicted to be also highly variable. Accurately assessing treatment response is an important factor when selecting patients for drug and therapy trials and also when deciding how long the trials should last. These considerations are emerging as critical to the success of drug and therapy trials (McGoldrick et al. 2006). Biomarkers such as inherited repeat length and mutation rates which modify age of onset and disease progression can provide trials with a better basis for assessing treatment response.

Obtaining two blood samples suitably far apart and then individually sizing the cells to obtain repeat length distributions may not be a feasible strategy for prognostic testing for patients. However, from a research perspective, this rich data is allowing us to calibrate and further validate our models and assess the levels of variation seen in the DM1 or HD population. Through model comparison, we have established the importance of contraction and individual variation. We now have a better understanding of how key summary statistics of the repeat length distributions, such as mean and variation, contribute to the underlying mutational process.

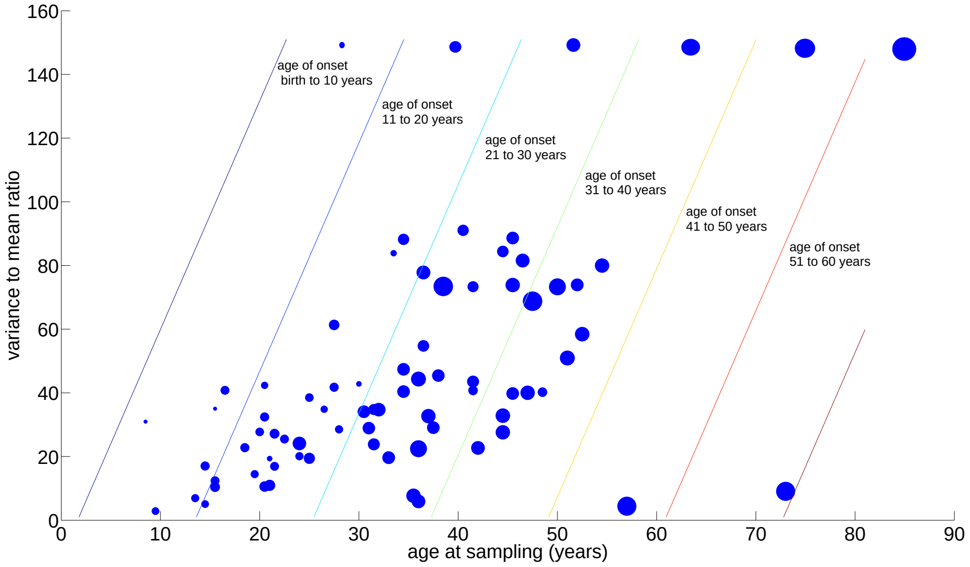


Figure 7.13: **Summary statistic variance-to-mean ratio at age of sampling is a potential indicator of age of onset.** The relationship between the variance-to-mean ratio at age of sampling and age of onset (72 samples: first and second samples combined). Actual age of onset is indicated by the size of the circle for each sample, with the largest circle indicating an age of onset of around 60 years and the smallest circle indicating an age of onset around birth. The size of circle corresponding to each age group is shown at the top of each age group. The predictive lines and associated predicted age of onsets are derived from the data using linear regression analysis (Table 7.2).

7.5 Material and Methods

7.5.1 Project data

The data used in this study comprises a further blood sample from 25 Costa Rican DM1 individuals originally in the large cohort study (discussed in Chapters 3 and 4) and two blood samples taken at different times from 15 Scottish DM1 individuals recently recruited for the DM1 genome variation study. Collecting DNA samples from blood and other tissues is ongoing in this longitudinal study. Access to previously taken diagnostic samples is obtained from the individuals recruited to the study with informed consent. The study is ethically approved by relevant local committees. Repeat lengths in these blood samples were sized using small pool PCR by Dr Fernando Morales, Dr Anneli Cooper and others from the Monckton lab.

7.5.2 Modelling approach for two samples taken at different points in time

Likelihood and maximum likelihood calculation

As in previous chapters, we represent the expansion rate per year, the contraction rate per year and inherited repeat length by λ_n, μ_n and n_0, respectively, and let $P_n(t)$ denote the probability that a cell (from either sample) has length n at time t. We know that the rate of change of $P_n(t)$ with respect to time is governed by the master equation

$$\frac{dP_n(t)}{dt} = -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t). \tag{7.1}$$

Given the allele length at time zero, n_0, we may approximate this infinite system of ordinary differential equations numerically by truncating at a suitably large value of $n = N$ and setting $P_n(t) = 0$ for all $n \geq N + 1$.

We use likelihood methods for model fitting and parameter estimation. We recall that likelihood is defined to be the probability that a repeat length has reached the length observed given the model and its parameters. For each individual in this study there are now two data samples, d_{j1} and d_{j2}, which denote the repeat length for the $j1$th observation taken at time $t1$ and the repeat length for the $j2$th observation taken at time $t2$. We fit the model first to each dataset separately and second to the combined dataset. The likelihood of observing the range of repeat lengths in the first sample, denoted L_1, is the product, over d_{j1}, of the probability of observing each repeat length, denoted $P_{d_{j1}}(t1;\theta)$, where θ are the model parameters. Similarly, the likelihood of observing the range of repeat lengths in the second sample, denoted L_2, is the product, over d_{j2}, of the probability $P_{d_{j2}}(t2;\theta)$. This gives the likelihoods for data samples, d_{j1} and d_{j2} respectively,

$$L_1 = \prod_{j1} P_{d_{j1}}(t1;\theta), \tag{7.2}$$

$$L_2 = \prod_{j2} P_{d_{j2}}(t2;\theta). \tag{7.3}$$

As the data samples are derived independently of each other, the likelihood of observing both data samples, L, is the product of each likelihood,

$$L = L_1 L_2. \tag{7.4}$$

We obtain the maximum value of each likelihood (L_1, L_2 and L) by evaluating over a broad parameter space, described in Table 7.1. For statistical analysis, it was useful to have point estimates for the parameters for each individual based on the first sample, the second sample and the combined sample. In each case, these were taken to be the maximum likelihood values.

Bayesian parameter estimation

For further statistical analysis, we obtained posterior probability distributions for each parameter and each individual, for each dataset separately and combined, by employing the likelihood in a Bayesian inference framework, see Chapter 4 for further details of this approach. In summary, the posterior distributions, under the special case of uniform priors, for each dataset separately and combined, π_1, π_2 and π, respectively, are

$$\pi_1(\theta|d_{j1}) \quad \propto \quad L_1(d_{j1}|\theta), \tag{7.5}$$

$$\pi_2(\theta|d_{j2}) \quad \propto \quad L_2(d_{j2}|\theta), \tag{7.6}$$

$$\pi(\theta|d_{j1}, d_{j2}) \quad \propto \quad L(d_{j1}, d_{j2}|\theta). \tag{7.7}$$

Chapter 8

Conclusions and future directions

We show that it is feasible to develop new mathematical models for dynamic DNA and use modern techniques from statistical inference on the latest datasets to calibrate and compare hypotheses and gain biological insights. By fitting mathematical models to extensive somatic mutation datasets arising from relevant individuals, we identify and quantify, for the first time, important features of the underlying mutational mechanism. This is the first time that large-scale populations of dynamic DNA data have been tackled in a systematic modelling framework and the results have significant implications for future work.

Individual differences in mutation rates and levels of somatic instability partially explain why individuals with the same inherited repeat lengths can have widely varying symptoms and disease onset, see Chapters 3, 4, 6 and 7. We interpret this variation as implicating *trans*-acting factors, either genetic or environmental in DM1 and HD. Having quantified several phenotype related traits, a future direction for this work is to use these traits to identify corresponding genetic factors. The availability of high-throughput genotyping technologies make it possible to survey the entire genome and uncover genetic influences using a genome wide association study (GWAS). An important consideration for the feasibility of a GWAS is the number of individuals required to ensure that the associated genes are identified. Recent traditional case-control studies, where disease individuals (case) are compared to non-disease individuals (control), have required large numbers of individuals (up to 10,000 or more), primarily because, for many complex traits, the effect of genes on the phenotype is low, less than 1.5 fold increase in risk (Hindorff et al. 2009). However we would expect gene variants that impact the observed phenotypes to explain more of the variability of the trait and consequently have more effect. Hence combining quantitative traits (QTs) such as somatic variation with a GWAS study rather than the traditional case-control approach should provide more power to find these genes and increase our understanding of the underlying mechanism (Potkin et al. 2009). Most GWAS studies are conducted using single-nucleotide polymorphisms (SNPs) instead of the whole genome. Another factor that impacts on power is the frequency of the gene variant in the population. In summary, the sample size needed to detect a related locus with a QT phenotype depends on the amount of variance explained by the QT and on the SNP allele frequencies. With conservative estimates for effect size (10%) and SNP allele frequency (between 10% and 20%), a sample size of between 500 and 800 should provide the required 80% power for a phenotype to be detected (Potkin et al. 2009, Purcell et al. 2003). Obtaining this number of affected individuals is clearly feasible as cohorts and patient registries of this size and greater have been recruited for DM1 and HD. An alternative approach would be to choose a subset of candidate genes, such as genes relating the the DNA mismatch repair mechanism, rather than the whole genome. This reduces the power requirement and hence the sample size, but could result in previously unconsidered effects being missed.

Quantifying the somatic variation phenotype for this size of group (500-800 individuals) would be challenging with existing low throughput methods, but some next generation sequencing technologies such as PacBio are claiming to be able to sequence longer lengths, 3,000 base pairs on average, than the standard 150 base pairs (Illumina). Taking into account flanking regions, we would require a technology that could accurately handle at least 7,000 base pairs for DM1 and 500 base pairs for HD, which although not currently possible, in particular for DM1, will soon be achievable. Alternatively, current technologies could be used, in combination with our models, to estimate average frequencies and mutation rates at shorter microsatellites which are potential proxies for expanded repeat loci for investigating instability across the genome. The creation of datasets that combine quantitative phenotypes with genome wide data pave the way for multivariate analysis that could uncover complex gene reactions involved in the somatic instability.

In Chapter 4, we challenge the widely held assumption that somatic DNA instability is dominated by expansion and reveal, surprisingly, that the observed expansion bias is the cumulative effect of very many expansion and contraction events. There have been no previous estimates of how often the repeat units are inserted in the repeat length tract *in vivo*. Our results suggest that mutational events happen every other day and that roughly 100 expansions and 98 contractions give rise to two overall repeat length gains. This suggests a link with regular DNA activities, such as DNA repair and transcription, rather than DNA replication. This is an example of how computational analysis can generate provocative hypotheses and drive future experimental work. Given the dependency of instability on repeat length and age, low levels of somatic instability are expected in individuals with small inherited repeat lengths, in particular very young individuals. Follow-up work could include looking at samples from these individuals which might allow us to observe small changes that could confirm our assumption that changes are typically one repeat unit, and that there are contractions as well as expansions.

Our hypothetical explanation for the repeat length instability involves DNA secondary structures and inappropriate DNA mismatch repair. However, how often the repeat lengths mutate and what determines the decision to expand or contract are unanswered questions. A cell system to study this issue could be devised involving synthesised DNA molecules with adopted structures such as loop-outs and expanded repeats. By exposing these DNA substrates to the DNA repair mechanism, extracted from cell culture, we could determine the rate of change *in vitro*. Given our results, we would expect that a high percentage of the changes were contractions. Such a system could also be used to assess potential therapies directed at reducing instability.

We show in Chapter 3 that the relationship between repeat length and levels of somatic variation is non-linear and complex. Concerning smaller alleles (less than 100 repeat units) found in late onset DM1 individuals and the majority of HD individuals, in Chapter 5 we find statistical support for a length-specific effect which suppresses mutational rates among the smaller alleles giving rise to a distinctive pattern in the repeat length distributions. In a novel application we also show that this distinctive pattern may help identify individuals whose effective repeat length, with regards to somatic instability, is less than their actual repeat length. A plausible explanation for this distinction is that the expanded repeat tract is compromised by interruptions or other unusual features. For these individuals, we are able to estimate the effective repeat length of their expanded repeat tracts and thereby contribute to the on-going discussion about the effect of interruptions on phenotype. The biochemical experiment discussed above could be extended to DNA structures containing interruptions and we could thereby consider the effect of different types of interruptions on instability. Some interruptions, such as CGG at the $3'$ end, appear to reduce instability and have *cis*-acting effects. Other individuals with reduced instability do not appear to have interruptions but instability here could be due to mutations in the *cis*-flanking region or a *trans*-acting effect due to mutations in genes on other chromosomes. Individuals, such as these, with extreme phenotypes could form part of a study, such as exome sequencing, to find the associated genes. Recently, exome sequencing with extreme phenotypes has been successful in identifying modifiers of disease (Emond et al. 2012). Enriched frequency of the gene associated trait in the extreme phenotype group improved the power of the study to find the modifiers in a moderate number of individuals (less than 100). This type of approach based on in-depth analysis of extreme phenotypes would be very applicable to DM1 and HD.

DM1 is a multi-systemic 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. Blood is also 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 a good chance to estimate the progenitor allele length, which is most indicative of age of onset. 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 also determine stability in other tissues. As complex tissues often display multi-modal distributions, likely reflecting the presence of very different cell types, dissecting the relative contribution of different cell types with different mutational dynamics is challenging. We approach this for end stage HD brain in Chapter 6 with promising results. The very large expansions observed in most other tissues of DM1 patients pose technical challenges but methods to size these repeat lengths are currently being investigated.

One of the aims of this work and ongoing work is to improve prognostic information for DM1 affected or at risk individuals. In Chapter 3, we show that progenitor allele length is the major modifier of age of onset. To measure progenitor allele length accurately, a blood sample could be recorded at birth. However this type of information is not typically available. Instead we use a novel computational approach to quantify objectively the progenitor or inherited repeat length. With blood DNA from one time point we can obtain a useful estimate and in Chapter 7 we show that the availability of two blood samples, from the same individual taken at different points in time, improves the estimates for both inherited repeat length and the rates of mutation. The resulting estimates have very promising predictive power in terms of age of onset. We also note that the success of clinical trials depends on setting targets in terms of patient response to drugs (McGoldrick et al. 2006). A better understanding of how an individual's disease is likely to progress therefore helps to set realistic targets and better evaluate potential therapies. Individuals recruited for clinical trials are quite likely to have been diagnosed previously and hence to have diagnostic samples which could retrospectively be analysed and provide the relevant information for response assessment. Extra samples will not typically be available for individuals being diagnosed for the first time but one option would be to take samples from different tissues, the easiest and least intrusive being blood and buccal. Our work shows that mutation rates correlate within HD individuals between neurons and glia so, in principle, assuming that this extends across tissues, two samples provide more information than one about inherited repeat length and individual specific levels of mutation.

We have taken steps towards establishing this principle in Chapter 6 where we adapt our instability model to infer mutation rates within two different types of cell in Huntington disease. As we build up a clearer picture of instability across disease and tissues, synergies make it possible to transfer learnings between datasets, as seen in Chapter 3, which extend the value of the analysis.

In summary, by developing and applying new mathematical models, we have added value to experimental data and provided novel and important insights into somatic instability at both the DM1 and HD loci. These computational studies have generated provocative hypotheses for directing experimental research. Our results have important implications for future therapies directed at reducing somatic variation, which in principle could cure or slow down disease progression. As well as improved prognostic information for patients and their families our models can also be used to provide better predictions for therapeutic response within clinical trials.

Bibliography

Aarts, M., Dekker, M., Dekker, R., de Vries, S., van der Wal, A., Wielders, E. & Riele, H. T. (2009), ‘Gene modification in embryonic stem cells by single-stranded DNA oligonucleotides’, *Methods in Molecular Biology (Clifton, N.J.)* **530**, 79–99.

Abecasis, G. R., Cardon, L. R. & Cookson, W. O. (2000), ‘A general test of association for quantitative traits in nuclear families’, *American Journal of Human Genetics* **66**, 279–292.

Abkowitz, J. L., Catlin, S. N., McCallie, M. T. & Guttorp, P. (2002), ‘Evidence that the number of hematopoietic stem cells per animal is conserved in mammals’, *Blood* **100**, 2665–2667.

Akaike, H. (1974), ‘A new look at the statistical model identification’, *Automatic Control, IEEE Transactions on* **19**, 716–723.

Andrew, S. E., Goldberg, Y. P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B. & Kalchman, M. A. (1993), ‘The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington’s disease’, *Nature Genetics* **4**, 398–403.

Anvret, M., Ahlberg, G., Grandell, U., Hedberg, B., Johnson, K. & Edstr¯om, L. (1993), ‘Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy’, *Human Molecular Genetics* **2**, 1397–1400.

Aronin, N., Chase, K., Young, C., Sapp, E., Schwarz, C., Matta, N., Kornreich, R., Lanwehrmeyer, B., Bird, E., Beal, M. F., Vonsattel, J., Smith, T., Carraway, R., Boyce, F. M., Young, A. B., Penney, J. B. & DiFiglia, M. (1995), ‘CAG expansion affects the expression of mutant huntingtin in the Huntington’s disease brain’, *Neuron* **15**, 1193–1201.

Ashizawa, T., Dubel, J. R., Dunne, P. W., Dunne, C. J., PhD, Y. H., MD, A. P., Caskey, C. T., PhD, E. B., 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.

Ashizawa, T., Dubel, J. R. & Harati, Y. (1993), ‘Somatic instability of CTG repeat in myotonic dystrophy’, *Neurology* **43**, 2674–2678.

Ashizawa, T., Dunne, P. W., Ward, P. A., Seltzer, W. K. & 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.

Ashizawa, T. & Epstein, H. (1991), ‘Ethnic distribution of myotonic dystrophy gene’, *The Lancet* **338**, 642–643.

Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman, J., Wormskamp, N. G. & Vooijs, M. (1992), ‘Cloning of the essential myotonic dystrophy region and mapping of the putative defect’, *Nature* **355**, 548–551.

Aylward, E. H., Li, Q., Stine, O. C., Ranen, N., Sherr, M., Barta, P. E., Bylsma, F. W., Pearlson, G. D. & Ross, C. A. (1997), ‘Longitudinal change in basal ganglia volume in patients with Huntington’s disease’, *Neurology* **48**, 394–399.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. J., Lent, R. & HerculanoHouzel, S. (2009), 'Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaledup primate brain', *The Journal of Comparative Neurology* **513**, 532–541.

Bates, G., P., H. & L., J. (2002), *Huntington's Disease*, Oxford University Press.

Beaumont, M. A. & Rannala, B. (2004), 'The Bayesian revolution in genetics', *Nature Reviews Genetics* **5**, 251–261.

Bell, J., Fisher, R. A., Penrose, L. & for National Eugenics, G. L. (1948), *The treasury of human inheritance*, Vol. 4, part 5, Cambridge University Press London.

Bhardwaj, R. D., Curtis, M. A., Spalding, K. L., Buchholz, B. A., Fink, D., Bjrk-Eriksson, T., Nordborg, C., Gage, F. H., Druid, H., Eriksson, P. S. & Frisn, J. (2006), 'Neocortical neurogenesis in humans is restricted to development', *Proceedings of the National Academy of Sciences* **103**, 12564–12568.

Braida, C., Stefanatos, R. K., Adam, B., Mahajan, N., Smeets, H. J., Niel, F., Goizet, C., Arveiler, B., Koenig, M., Lagier-Tourenne, C., Mandel, J., Faber, C. G., de Die-Smulders, C. E., Spaans, F. & Monckton, D. G. (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', *Human Molecular Genetics* **19**, 1399–1412.

Brock, G. J., Anderson, N. H. & Monckton, D. G. (1999), '*Cis*-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands', *Human Molecular Genetics* **8**, 1061–1067.

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. (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.

Brunner, H. G., Brüggenwirth, H. T., Nillesen, W., Jansen, G., Hamel, B. C. J., Hoppe, R. L. E., de Die, C. E. M., Höweler, C. J., van Oost, B. A., Wieringa, B., Ropers, H. H. & Smeets, H. J. M. (1993), 'Influence of sex of the transmitting parent as well as of parental allele site on the CTG expansion in myotonic dystrophy (DM)', *American Journal of Human Genetics* **53**, 1016–1023.

Burnham, K. P. & Anderson, D. R. (2002), *Model selection and multimodel inference: a practical information-theoretic approach*, Springer.

Buxton, J., Shelbourne, P., Davies, J., Jones, C., Tongeren, T. V., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M. & Riley, B. (1992), 'Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy', *Nature* **355**, 547–548.

Calabrese, P. & Sainudiin, R. (2005), Springer.

Castel, A. L., Cleary, J. D. & Pearson, C. E. (2010), 'Repeat instability as the basis for human diseases and as a potential target for therapy', *Nature Reviews Molecular Cell Biology* **11**, 165–170.

Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. (2011), 'The replication rate of human hematopoietic stem cells *in vivo*', *Blood* **117**, 4460–4466.

Cleary, J. D., Nichol, K., Wang, Y.-H. & Pearson, C. E. (2002), 'Evidence of *cis*-acting factors in replication-mediated trinucleotide repeat instability in primate cells', *Nature Genetics* **31**, 37–46.

Cobo, A. M., Poza, J. J., Martorell, L., Lpez de Munain, A., Emparanza, J. I. & Baiget, M. (1995), 'Contribution of molecular analyses to the estimation of the risk of congenital myotonic dystrophy', *Journal of Medical Genetics* **32**, 105–108.

Cortopassi, G. A. & Arnheim, N. (1990), ‘Detection of a specific mitochondrial DNA deletion in tissues of older humans.’, *Nucleic Acids Research* **18**, 6927–6933.

Cox, D. & Hinkley, D. (1994), *Theoretical statistics*, Chapman & Hall.

Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M. & Abbott, M. (1993), ‘Trinucleotide repeat length instability and age of onset in Huntington’s disease’, *Nature Genetics* **4**, 387–392.

Eckert, K. A. & Hile, S. E. (2009), ‘Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome’, *Molecular Carcinogenesis* **48**, 379–388.

Emond, M. J., Louie, T., Emerson, J., Zhao, W., Mathias, R. A., Knowles, M. R., Wright, F. A., Rieder, M. J., Tabor, H. K., Nickerson, D. A., Barnes, K. C., Gibson, R. L. & Bamshad, M. J. (2012), ‘Exome sequencing of extreme phenotypes identifies *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis’, *Nature Genetics* **44**, 886–889.

Falush, D. (2009), ‘Haplotype background, repeat length evolution, and Huntington’s disease’, *American Journal of Human Genetics* **85**, 939–942.

Falush, D., Almqvist, E. W., Brinkmann, R. R., Iwasa, Y. & Hayden, M. R. (2001), ‘Measurement of mutational flow implies both a high new-mutation rate for Huntington disease and substantial underascertainment of late-onset cases’, *The American Journal of Human Genetics* **68**, 373–385.

Fleischer, B. (1918), ‘Uber myotonische dystrophie mit katarakt’, *Graefe’s Archive for Clinical and Experimental Ophthalmology* **96**, 91–133.

Fondon, John W, r., Hammock, E. A. D., Hannan, A. J. & King, D. G. (2008), ‘Simple sequence repeats: genetic modulators of brain function and behavior’, *Trends in Neurosciences* **31**, 328–334.

Fortune, M. T., Vassilopoulos, C., Coolbaugh, M. I., Siciliano, M. J. & Monckton, D. G. (2000), ‘Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability’, *Human Molecular Genetics* **9**, 439–445.

Fu, Y. H., Kuhl, D. P., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S., Verkerk, A. J., Holden, J. J., Fenwick, R. G. & Warren, S. T. (1991), ‘Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox’, *Cell* **67**, 1047–1058.

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. (1992), ‘An unstable triplet repeat in a gene related to myotonic muscular dystrophy’, *Science* **255**, 1256–1258.

Fuchs, E. (2008), ‘Skin stem cells: rising to the surface’, *The Journal of Cell Biology* **180**, 273–284.

Gellibolian, R., Bacolla, A. & Wells, R. D. (1997), ‘Triplet repeat instability and DNA topology: An expansion model based on statistical mechanics’, *Journal of Biological Chemistry* **272**, 16793–16797.

Genschel, J. & Modrich, P. (2003), ‘Mechanism of $5'$ -directed excision in human mismatch repair’, *Molecular Cell* **12**, 1077–1086.

Gharehbaghi-Schneli, E. B., Finsterei, J., Korschineck, I., Mamoli, B. & Binder, B. R. (2008), ‘Genotype -phenotype correlation in myotonic dystrophy’, *Clinical Genetics* **53**, 20–26.

Gibbs, M., Collick, A., Kelly, R. G. & Jeffreys, A. J. (1993), ‘A tetranucleotide repeat mouse minisatellite displaying substantial somatic instability during early preimplantation development’, *Genomics* **17**, 121–128.

Gomes-Pereira, M., Fortune, M. T., Ingram, L., McAbney, J. P. & Monckton, D. G. (2004), 'Pms2 is a genetic enhancer of trinucleotide CAG.CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion', *Human Molecular Genetics* **13**, 1815–1825.

Gomes-Pereira, M., Fortune, M. T. & Monckton, D. G. (2001), 'Mouse tissue culture models of unstable triplet repeats: *in vitro* selection for larger alleles, mutational expansion bias and tissue specificity, but no association with cell division rates', *Human Molecular Genetics* **10**, 845–854.

Gomes-Pereira, M. & Monckton, D. G. (2004), 'Chemically induced increases and decreases in the rate of expansion of a CAG*CTG triplet repeat', *Nucleic Acids Research* **32**, 2865–2872.

Gomes-Pereira, M. & Monckton, D. G. (2006), 'Chemical modifiers of unstable expanded simple sequence repeats: What goes up, could come down', *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **598**, 15–34.

Gonitel, R., Moffitt, H., Sathasivam, K., Woodman, B., Detloff, P. J., Faull, R. L. M. & Bates, G. P. (2008), 'DNA instability in postmitotic neurons', *Proceedings of the National Academy of Sciences* **105**, 3467–3472.

Gonzalez, Ohsawa, N., Singer, R. H., Devillers, M., Ashizawa, T., Balasubramanyam, A., Cooper, T. A., Khajavi, M., Lia-Baldini, A. S., Miller, G. et al. (2000), 'New nomenclature and DNA testing guidelines for myotonic dystrophy type 1(DM1)', *Neurology* **54**, 1218–1221.

Gorbunova, V., Seluanov, A., Dion, V., Sandor, Z., Meservy, J. L. & Wilson, J. H. (2003), 'Selectable system for monitoring the instability of CTG/CAG triplet repeats in mammalian cells', *Molecular and Cellular Biology* **23**, 4485–4493.

Graveland, G. A., Williams, R. S. & DiFiglia, M. (1985), 'Evidence for degenerative and regenerative changes in neostriatal spiny neurons in Huntington's disease', *Science* **227**, 770–773.

Greenfield, J. G. (1911), 'Notes on a family of myotonia atrophica and early cataract with a report of an additional case of myotonia atrophica', *Review of Neurology and Psychiatry* **9**, 169–181.

Groenen, P. & Wieringa, B. (1998), 'Expanding complexity in myotonic dystrophy', *Bioessays* **20**.

Groh, W. J., Groh, M. R., Saha, C., Kincaid, J. C., Simmons, Z., Ciafaloni, E., Pourmand, R., Otten, R. F., Bhakta, D., Nair, G. V., Marashdeh, M. M., Zipes, D. P. & Pascuzzi, R. M. (2008), 'Electrocardiographic abnormalities and sudden death in myotonic dystrophy type 1', *The New England Journal of Medicine* **358**, 2688–2697.

Gusella, J. F., McNeil, S., Persichetti, F., Srinidhi, J., Novelletto, A., Bird, E., Faber, P., Vonsattel, J. P., Myers, R. H. & MacDonald, M. E. (1996), 'Huntington's disease', *Cold Spring Harbor Symposia on Quantitative Biology* **61**, 615–626.

Hamshere, M., Harley, H., Harper, P., Brook, J. & Brookfield, J. (1999), 'Myotonic dystrophy: the correlation of (CTG) repeat length in leucocytes with age at onset is significant only for patients with small expansions', *Journal of Medical Genetics* **36**, 59–61.

Hannan, A. J. (2010), 'Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for missing heritability', *Trends in Genetics* **26**, 59–65.

Harley, H. G., Brook, J. D., Rundle, S. A., Crow, S., Reardon, W., Buckler, A. J., Harper, P. S., Housman, D. E. & Shaw, D. J. (1992), 'Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy', *Nature* **355**, 545–546.

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

Harper, P. S. (1989), *Myotonic Dystrophy*, W B Saunders Company.

Harper, P. S., Harley, H. G., Reardon, W. & Shaw, D. J. (1992), ‘Anticipation in myotonic dystrophy: new light on an old problem’, *American Journal of Human Genetics* **51**, 10–16.

Harris, S., Moncrieff, C. & Johnson, K. (1996), ‘Myotonic dystrophy: will the real gene please step forward!’, *Human Molecular Genetics* **5**, 1417–1423.

Higham, C. F., Morales, F., Cobbold, C. A., Haydon, D. T. & 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’, *Human Molecular Genetics* **21**, 2450–2463.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. (2009), ‘Potential etiologic and functional implications of genome-wide association loci for human diseases and traits’, *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362–9367.

Höweler, C. J., Busch, H. F., Geraedts, J. P., Niermeijer, M. F. & Staal, A. (1989), ‘Anticipation in myotonic dystrophy: fact or fiction?’, *Brain: A Journal of Neurology* **112**, 779–797.

Hunter, A., Tsilfidis, C., Mettler, G., Jacob, P., Mahadevan, M., Surh, L. & Korneluk, R. (1992), ‘The correlation of age of onset with CTG trinucleotide repeat amplification in myotonic dystrophy’, *Journal of Medical Genetics* **29**, 774–779.

Hunter, J. M., Crouse, A. B., Lesort, M., Johnson, G. V. W. & Detloff, P. J. (2005), ‘Verification of somatic CAG repeat expansion by pre-PCR fractionation’, *Journal of Neuroscience Methods* **144**, 11–17.

Jansen, G., Groenen, P. J., Bächner, D., Jap, P. H., Coerwinkel, M., Oerlemans, F., van den Broek, W., Gohlsch, B., Pette, D., Plomp, J. J., Molenaar, P. C., Nederhoff, M. G., van Echteld, C. J., Dekker, M., Berns, A., Hameister, H. & Wieringa, B. (1996), ‘Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice’, *Nature Genetics* **13**, 316–324.

Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L. & Armour, J. A. L. (1994), ‘Complex gene conversion events in germline mutation at human minisatellites’, *Nature Genetics* **6**, 136–145.

Jung, J. & Bonini, N. (2007), ‘CREB-Binding protein modulates repeat instability in a drosophila model for PolyQ disease’, *Science* **315**, 1857–1859.

Kaplan, S., Itzkovitz, S. & Shapiro, E. (2007), ‘A universal mechanism ties genotype to phenotype in trinucleotide diseases’, *PLoS Computational Biology* **3**, e235.

Kennedy, L., Evans, E., Chen, C., Craven, L., Detloff, P. J., Ennis, M. & Shelbourne, P. F. (2003), ‘Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis’, *Human Molecular Genetics* **12**, 3359–3367.

Kennedy, L. & Shelbourne, P. F. (2000), ‘Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington’s disease?’, *Human Molecular Genetics* **9**, 2539–2544.

Kovtun, I. V. & McMurray, C. T. (2001), ‘Trinucleotide expansion in haploid germ cells by gap repair’, *Nature Genetics* **27**, 407–411.

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. (1998), ‘Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations’, **95**, 10774–10778.

Kunkel, T. (1999), ‘The high cost of living’, *Trends in Genetics* **15**, 93–94.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature* **409**, 860–921.

Landwehrmeyer, G. B., McNeil, S. M., Dure, L S, t., Ge, P., Aizawa, H., Huang, Q., Ambrose, C. M., Duyao, M. P., Bird, E. D. & Bonilla, E. (1995), 'Huntington's disease gene: regional and cellular expression in brain of normal and affected individuals', *Annals of Neurology* **37**, 218–230.

Lavedan, C., Hofmann-Radvanyi, H., Shelbourne, P., Rabes, J. P., Duros, C., Savoy, D., Dehaupas, I., Luce, S., Johnson, K. & Junien, C. (1993), 'Myotonic dystrophy: size-and sex-dependent dynamics of CTG meiotic instability, and somatic mosaicism', *American Journal of Human Genetics* **52**, 875.

Lee, J. E., Bennett, C. F. & Cooper, T. A. (2012), 'RNase h-mediated degradation of toxic RNA in myotonic dystrophy type 1', *Proceedings of the National Academy of Sciences* .

Lee, J., Pinto, R. M., Gillis, T., St. Claire, J. C. & Wheeler, V. C. (2011), 'Quantification of Age-Dependent somatic CAG repeat instability in *Hdh* CAG Knock-In mice reveals different expansion dynamics in striatum and liver', *PLoS ONE* **6**, e23647.

Lee, J., Zhang, J., Su, A., Walker, J., Wiltshire, T., Kang, K., Dragileva, E., Gillis, T., Lopez, E., Boily, M., Cyr, M., Kohane, I., Gusella, J., MacDonald, M. & Wheeler, V. (2010), 'A novel approach to investigate tissue-specific trinucleotide repeat instability', *BMC Systems Biology* **4**, 29.

Leeflang, E. P., McPeek, M. S. & Arnheim, N. (1996), 'Analysis of meiotic segregation, using single sperm typing meiotic drive at the myotonic dystrophy locus', *American Journal of Human Genetics* **59**, 896–904.

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. & Arnheim, N. (1999), 'Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism', *Human Molecular Genetics* **8**, 173–183.

Leeflang, E. P., Zhang, L., Tavar, S., Hubert, R., Srinidhi, J., MacDonald, M. E., Myers, R. H., de Young, M., Wexler, N. S. & Gusella, J. F. (1995), 'Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum', *Human Molecular Genetics* **4**, 1519–1526.

Li, J., Hayden, M. R., Almqvist, E. W., Brinkman, R. R., Durr, A., Dod, C., Morrison, P. J., Suchowersky, O., Ross, C. A., Margolis, R. L. et al. (2003), 'A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS study', *American Journal of Human Genetics* **73**, 682–687.

Lia, A. S., Seznec, H., Hofmann-Radvanyi, H., Radvanyi, F., Duros, C., Saquet, C., Blanche, M., Junien, C. & Gourdon, G. (1998), 'Somatic instability of the CTG repeat in mice transgenic for the myotonic dystrophy region is age dependent but not correlated to the relative intertissue transcription levels and proliferative capacities', *Human Molecular Genetics* **7**, 1285–1291.

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. & La Spada, A. R. (2003), 'Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice', *Human Molecular Genetics* **12**, 41–50.

Lindahl, T. (1993), 'Instability and decay of the primary structure of DNA', *Nature* **362**, 709–715.

Liquori, C. L., Ricker, K., Moseley, M. L., Jacobsen, J. F., Kress, W., Naylor, S. L., Day, J. W. & Ranum, L. P. W. (2001), 'Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of *ZNF9*', *Science* **293**, 864–867.

Lloret, A., Dragileva, E., Teed, A., Espinola, J., Fossale, E., Gillis, T., Lopez, E., Myers, R. H., MacDonald, M. E. & Wheeler, V. C. (2006), ‘Genetic background modifies nuclear mutant huntingtin accumulation and *HD* CAG repeat instability in Huntington’s disease knock-in mice’, *Human Molecular Genetics* **15**, 2015–2024.

López de Munain, A., Cobo, A. M., Poza, J. J., Navarrete, D., Martorell, L., Palau, F., Emparanza, J. I. & Baiget, M. (1995), ‘Influence of the sex of the transmitting grandparent in congenital myotonic dystrophy’, *Journal of Medical Genetics* **32**, 689–691.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J. & O’hoy, K. (1992), ‘Myotonic dystrophy mutation: an unstable CTG repeat in the 3’ untranslated region of the gene’, *Science* **255**, 1253–1255.

Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S. W. & Bates, G. P. (1996), ‘Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice’, *Cell* **87**, 493–506.

Manley, K., Shirley, T. L., Flaherty, L. & Messer, A. (1999), ‘*Msh2* deficiency prevents *in vivo* somatic instability of the CAG repeat in Huntington disease transgenic mice’, *Nature Genetics* **23**, 471–473.

Marchini, C., Lonigro, R., Verriello, L., Pellizzari, L., Bergonzi, P. & Damante, G. (2000), ‘Correlations between individual clinical manifestations and CTG repeat amplification in myotonic dystrophy’, *Clinical Genetics* **57**, 74–82.

Martorell, L. (1997), ‘Somatic instability of the myotonic dystrophy (CTG) n repeat during human fetal development’, *Human Molecular Genetics* **6**, 877–880.

Martorell, L. (1998), ‘Progression of somatic CTG repeat length heterogeneity in the blood cells of myotonic dystrophy patients’, *Human Molecular Genetics* **7**, 307–312.

Martorell, L., Gamez, J., Cayuela, M. L., Gould, F. K., McAbney, J. P., Ashizawa, T., Monckton, D. G. & Baiget, M. (2004), ‘Germline mutational dynamics in myotonic dystrophy type 1 males: allele length and age effects’, *Neurology* **62**, 269–274.

Martorell, L., Monckton, D., Gamez, J. & Baiget, M. (2000), ‘Complex patterns of male germline instability and somatic mosaicism in myotonic dystrophy type 1’, *European Journal of Human Genetics* **8**, 423–430.

Mathieu, J., De Braekeleer, M. & Prvost, C. (1990), ‘Genealogical reconstruction of myotonic dystrophy in the Saguenay-Lac-Saint-Jean area (Quebec, Canada)’, *Neurology* **40**, 839–842.

McGoldrick, S., Duffy, F. & Bennett, C. (2006), ‘Making headway in Huntington’s’, *Good Clinical Practice Journal* **13**, 31–36.

McMurray, C. T. (2010), ‘Mechanisms of trinucleotide repeat instability during human development’, *Nature Reviews Genetics* **11**, 786–799.

Merlevede, K., Vermander, D., Theys, P., Legius, E., Ector, H. & Robberecht, W. (2002), ‘Cardiac involvement and CTG expansion in myotonic dystrophy’, *Journal of Neurology* **249**, 693–698.

Mirkin, S. M. (2007), ‘Expandable DNA repeats and human disease’, *Nature* **447**, 932–940.

Mittelman, D., Moye, C., Morton, J., Sykoudis, K., Lin, Y., Carroll, D. & Wilson, J. H. (2009), ‘Zinc-finger directed double-strand breaks within CAG repeat tracts promote repeat instability in human cells’, *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9607–9612.

Mladenovic, J., Pekmezovic, T., Todorovic, S., Rakocevic-Stojanovic, V., Savic, D., Romac, S. & Apostolski, S. (2006), ‘Survival and mortality of myotonic dystrophy type 1 (Steinert’s disease) in the population of Belgrade’, *European Journal of Neurology* **13**, 451–454.

Modoni, A., Silvestri, G., Grazia Pomponi, M., Mangiola, F., Tonali, P. A. & Marra, C. (2004), ‘Characterization of the pattern of cognitive impairment in myotonic dystrophy type 1’, *Archives of Neurology* **61**, 1943–1947.

Monckton, D. G., Cayuela, M. L., Gould, F. K., Brock, G. J., de Silva, R. & Ashizawa, T. (1999), ‘Very large (CAG)n DNA repeat expansions in the sperm of two spinocerebellar ataxia type 7 males’, *Human Molecular Genetics* **8**, 2473–2478.

Monckton, D. G., Coolbaugh, M. I., Ashizawa, K. T., Siciliano, M. J. & Caskey, C. T. (1997), ‘Hypermutable myotonic dystrophy CTG repeats in transgenic mice’, *Nature Genetics* pp. 193–196.

Monckton, D. G., Wong, L. J., Ashizawa, T. & Caskey, C. T. (1995), ‘Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses’, *Human Molecular Genetics* **4**, 1–8.

Morales, F. A. (2006), Somatic mosaicism and genotype-phenotype correlations in myotonic dystrophy type 1, PhD thesis, University of Glasgow.

Morales, F., Couto, J. M., Higham, C. F., Hogg, G., Cuenca, P., Braida, C., Wilson, R. H., Adam, B., del Valle, G., Brian, R., Sittenfeld, M., Ashizawa, T., Wilcox, A., Wilcox, D. E. & Monckton, D. G. (2012), ‘Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity’, *Human Molecular Genetics* **21**, 3558–3567.

Mulders, S. A. M., van Engelen, B. G. M., Wieringa, B. & Wansink, D. G. (2010), ‘Molecular therapy in myotonic dystrophy: focus on RNA gain-of-function’, *Human Molecular Genetics* **19**, R90–R97.

Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., Prochazka, T., Koukal, P., Marikova, T., Kraus, J., Havlovicova, M. & Sedlacek, Z. (2009), ‘Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene’, *American Journal of Medical Genetics. Part A* **149A**, 1365–1374.

Myers, R., Marans, K. S. & MacDonald, M. (1998), Academic Press, San Diego, CA, USA.

Nestor, C. E. & Monckton, D. G. (2011), ‘Correlation of Inter-Locus polyglutamine toxicity with CAGCTG triplet repeat expandability and flanking genomic DNA GC content’, *PLoS ONE* **6**, e28260.

Novozhilov, A. S., Karev, G. P. & Koonin, E. V. (2006), ‘Biological applications of the theory of birth-and-death processes’, *Briefings in Bioinformics* **7**, 70–85.

Ohta, T. & Kimura, M. (1994), ‘A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population’, *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers* .

Olsen, P. A., Solhaug, A., Booth, J. A., Gelazauskaite, M. & Krauss, S. (2009), ‘Cellular responses to targeted genomic sequence modification using single-stranded oligonucleotides and zinc-finger nucleases’, *DNA repair* **8**, 298–308.

Otto, S. P., Day, T. & Day, T. (2007), *A Biologist’s Guide to Mathematical Modeling in Ecology and Evolution*, Princeton University Press.

Pearson, C. E., Edamura, K. N. & Cleary, J. D. (2005), ‘Repeat instability: mechanisms of dynamic mutations’, *Nature Reviews Genetics* **6**, 729–742.

Pearson, C. E. & 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.

Penrose, L. S. C. (1948), ‘The problem of anticipation in pedigrees of dystrophia myotonica’, *Annals of Eugenics* **14**, 125–132.

Perini, G. I., Menegazzo, E., Ermani, M., Zara, M., Gemma, A., Ferruzza, E., Gennarelli, M. & Angelini, C. (1999), ‘Cognitive impairment and (CTG)n expansion in myotonic dystrophy patients’, *Biological Psychiatry* **46**, 425–431.

Philips, A. V., Timchenko, L. T. & Cooper, T. A. (1998), ‘Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy’, *Science* **280**, 737–741.

Pollard, L. M., Sharma, R., Gómez, M., Shah, S., Delatycki, M. B., Pianese, L., Monticelli, A., Keats, B. J. B. & Bidichandani, S. I. (2004), ‘Replication-mediated instability of the GAA triplet repeat mutation in Friedreich ataxia’, *Nucleic Acids Research* **32**, 5962–5971.

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Torri, F., Keator, D. B. & Macciardi, F. (2009), ‘Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations’, *Cognitive Neuropsychiatry* **14**, 391–418.

Purcell, S., Cherny, S. S. & Sham, P. C. (2003), ‘Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits’, *Bioinformatics* **19**, 149–150.

Ranum, L. P. W. & Cooper, T. A. (2006), ‘RNA-mediated neuromuscular disorders’, *Annual Review of Neuroscience* **29**, 259–277.

Redman, J. B., Fenwick, R. G., Fu, Y. H., Pizzuti, A. & Caskey, C. T. (1993), ‘Relationship between parental trinucleotide GCT repeat length and severity of myotonic dystrophy in offspring’, *JAMA: The Journal of the American Medical Association* **269**, 1960–1965.

Renshaw, E. (1991), *Modelling Biological Populations in Space and Time*, Cambridge University Press.

Richards, R. I. & Sutherland, G. R. (1992), ‘Dynamic mutations: a new class of mutations causing human disease’, *Cell* **70**, 709–712.

Richards, R. I. & Sutherland, G. R. (1994), ‘Simple repeat DNA is not replicated simply’, *Nature Genetics* **6**, 114–116.

Rienzo, A. D., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M. & Freimer, N. B. (1994), ‘Mutational processes of simple-sequence repeat loci in human populations’, *Proceedings of the National Academy of Sciences of the United States of America* **91**, 3166–3170.

Savić, D., Rakočvić-Stojanović, V., Keckarević, D., Čuljković, B., Stojković, O., Mladenović, J., Todorović, S., Apostolski, S. & Romac, S. (2002), ‘250 CTG repeats in DMPK is a threshold for correlation of expansion size and age at onset of juvenile-adult DM1’, *Human Mutation* **19**, 131–139.

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

Seznec, H., Lia-Baldini, A., Duros, C., Fouquet, C., Lacroix, C., Hofmann-Radvanyi, H., Junien, C. & Gourdon, G. (2000), ‘Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability’, *Human Molecular Genetics* **9**, 1185 –1194.

Shampine, L. F., Gladwell, I. & Thompson, S. (2003), *Solving ODEs with MATLAB*, Cambridge University Press.

Shelbourne, P. F., Keller-McGandy, C., Bi, W. L., Yoon, S., Dubeau, L., Veitch, N. J., Vonsattel, J. P., Wexler, N. S., Arnheim, N. & Augood, S. J. (2007), 'Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain', *Human Molecular Genetics* **16**, 1133–1142.

Singer, T., McConnell, M. J., Marchetto, M. C., Coufal, N. G. & Gage, F. H. (2010), 'LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes?', *Trends in Neurosciences* **33**, 345–354.

Sivia, D. S. (2006), *Data analysis: A Bayesian tutorial*, Oxford University Press, USA.

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. & Shaw, D. J. (1993), 'Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease', *Nature Genetics* **4**, 393–397.

Stine, O. C., Pleasant, N., Franz, M. L., Abbott, M. H., Folstein, S. E. & Ross, C. A. (1993), 'Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15', *Human Molecular Genetics* **2**, 1547–1549.

Sun, J. X., Helgason, A., Masson, G., Ebenesersdóttir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D. & Stefansson, K. (2012), 'A direct characterization of human mutation based on microsatellites', *Nature Genetics* **44**, 1161–1165.

Sutherland, G. R., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Richards, R. I. & Haan, E. A. (1991), 'Hereditary unstable DNA: a new explanation for some old genetic questions?', *The Lancet* **338**, 289–292.

Swami, M., Hendricks, A. E., Gillis, T., Massood, T., Mysore, J., Myers, R. H. & 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', *Human Molecular Genetics* **18**, 3039–3047.

Taneja, K. L., McCurrach, M., Schalling, M., Housman, D. & Singer, R. H. (1995), 'Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues', *Journal of Cell Biology* **128**, 995–1002.

Telenius, H., Kremer, B., Goldberg, Y. P., Theilmann, J., Andrew, S. E., Zeisler, J., Adam, S., Greenberg, C., Ives, E. J. & Clarke, L. A. (1994), 'Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm', *Nature Genetics* **6**, 409–414.

The Huntington's Disease Collaborative Research Group (1993), 'A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes', *Cell* **72**, 971–983.

Thornton, C. A., Johnson, K. & Moxley, R. T. (1994), 'Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes', *Annals of Neurology* **35**, 104–107.

Tsilfidis, C., MacKenzie, A. E., Mettler, G., Barcel, J. & Korneluk, R. G. (1992), 'Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy', *Nature Genetics* **1**, 192–195.

Ueda, H., Ohno, S. & Kobayashi, T. (2000), 'Myotonic dystrophy and myotonic dystrophy protein kinase', *Progress in Histochemistry and Cytochemistry* **35**, 187–251.

van den Broek, W. J. A. A., Nelen, M. R., Wansink, D. G., Coerwinkel, M. M., te Riele, H., Groenen, P. J. T. A. & 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', *Human Molecular Genetics* **11**, 191 –198.

Veitch, N. J., Ennis, M., McAbney, J. P., Shelbourne, P. F. & Monckton, D. G. (2007), 'Inherited CAGCTG allele length is a major modifier of somatic mutation length variability in Huntington disease', *DNA Repair* **6**, 789–796.

Ver Hoef, J. M. & Boveng, P. L. (2007), 'Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?', *Ecology* **88**, 2766–2772.

Veytsman, B. & Akhmadeyeva, L. (2006), 'Simple mathematical model of pathologic microsatellite expansions: When self-reparation does not work', *Journal of Theoretical Biology* **242**, 401–408.

Vonsattel, J. P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D. & Richardson, E P, J. (1985), 'Neuropathological classification of Huntington's disease', *Journal of Neuropathology and Experimental Neurology* **44**, 559–577.

Watase, K., Venken, K. J. T., Sun, Y., Orr, H. T. & 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', *Human Molecular Genetics* **12**, 2789–2795.

Weber, J. L. & Wong, C. (1993), 'Mutation of human short tandem repeats', *Human Molecular Genetics* **2**, 1123–1128.

Wedderburn, R. W. M. (1974), 'Quasi-Likelihood functions, generalized linear models, and the Gauss-Newton method', *Biometrika* **61**, 439–447.

Wells, R. D., Dere, R., Hebert, M. L., Napierala, M. & Son, L. S. (2005), 'Advances in mechanisms of genetic instability related to hereditary neurological diseases', *Nucleic Acids Research* **33**, 3785.

Wexler, N. S., Lorimer, J., Porter, J., Gomez, F., Moskowitz, C., Shackell, E., Marder, K., Penchaszadeh, G., Roberts, S. A., Gayán, J. et al. (2004), 'Venezuelan kindreds reveal that genetic and environmental factors modulate huntington's disease age of onset', *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3498–3503.

Wexler, N. S., Young, A. B., Tanzi, R. E., Travers, H., Starosta-Rubinstein, S., Penney, J. B., Snodgrass, S. R., Shoulson, I., Gomez, F., Arroyo, M. A. R., Penchaszadeh, G. K., Moreno, H., Gibbons, K., Faryniarz, A., Hobbs, W., Anderson, M. A., Bonilla, E., Conneally, P. M. & Gusella, J. F. (1987), 'Homozygotes for Huntington's disease', *Nature* **326**, 194–197.

Wheeler, T. M. (2008), 'Myotonic dystrophy: Therapeutic strategies for the future', *Neurotherapeutics* **5**, 592–600.

Wheeler, T. M. & Thornton, C. A. (2007), 'Myotonic dystrophy: RNA-mediated muscle disease', *Current Opinion in Neurology* **20**, 572–6.

Wheeler, V. C., Lebel, L., Vrbanac, V., Teed, A., Te Riele, H. & MacDonald, M. E. (2003), 'Mismatch repair gene *Msh2* modifies the timing of early disease in Hdh^{Q111} striatum', *Human Molecular Genetics* **12**, 273–281.

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. & Group, T. U. C. R. (2007), 'Factors associated with *HD* CAG repeat instability in Huntington disease', *Journal of Medical Genetics* **44**, 695–701.

Wilkinson, D. J. (2009), 'Stochastic modelling for quantitative description of heterogeneous biological systems', *Nature Reviews Genetics* **10**, 122–133.

Wilkinson, D. J. (2011), *Stochastic modelling for systems biology*, Vol. 44, CRC press.

Wong, L. J. & Ashizawa, T. (1997), 'Instability of the (CTG)n repeat in congenital myotonic dystrophy', *The American Journal of Human Genetics* **61**, 1445–1448.

Wong, L. J., Ashizawa, T., Monckton, D. G., Caskey, C. T. & Richards, C. S. (1995), 'Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent', *American Journal of Human Genetics* **56**, 114–122.

Xu, X., Peng, M., Fang, Z. & Xu, X. (2000), ‘The direction of microsatellite mutations is dependent upon allele length’, *Nature Genetics* **24**, 396–399.

Yotova, V., Labuda, D., Zietkiewicz, E., Gehl, D., Lovell, A., Lefebvre, J.-F., Bourgeois, S., Lemieux-Blanchard, E., Labuda, M., Vézina, H., Houde, L., Tremblay, M., Toupance, B., Heyer, E., Hudson, T. J. & Laberge, C. (2005), ‘Anatomy of a founder effect: myotonic dystrophy in Northeastern Quebec’, *Human Genetics* **117**, 177–187.

Zatz, M., Passos-Bueno, M. R., Cerqueira, A., Marie, S. K., Vainzof, M. & Pavanello, R. C. (1995), ‘Analysis of the CTG repeat in skeletal muscle of young and adult myotonic dystrophy patients: when does the expansion occur?’, *Human Molecular Genetics* **4**, 401–406.

Zhang, Y., Monckton, D. G., Siciliano, M. J., Connor, T. H. & Meistrich, M. L. (2002), ‘Age and insertion site dependence of repeat number instability of a human DM1 transgene in individual mouse sperm’, *Human Molecular Genetics* **11**, 791–798.

Appendix 1

Representative examples of MATLAB programs:

1. To calculate the probability that a cell has repeat length n at time t under model M_α;
2. To calculate the log likelihood.

```
% returns Pn and T - the probability that a repeat length is length n
% at time T given the parameter combination (lam, mu, alpha, n0) and a
% finite limit for age at sampling (maxt) and n (maxn)

function [Pn T]=Pnmat_sol(lam, mu, alpha, n0, maxt, maxn)

        N=0:maxn;
        %Define matrix A where row N is the master equation dPN/dt
        A=zeros(length(N),length(N));

        %calculate the length-specific effect (RN) for each N
        RN=(max(N-alpha,0).*max(N-alpha+1,0))./max(N,1).^2;
        %1.calculate diagonal in matrix A
        vx=-(lam*RN.*N)-(mu*RN.*N);

        %calculate the length-specific effect (RNa) for each Na
        Na=(N(1:end-1)+1);
        RNa=(max(Na-alpha,0).*max(Na-alpha+1,0))./max(Na,1).^2;
        %2.calculate diagonal above in matrix A
        vxa=mu*RNa.*Na;

        %calculate the length-specific effect (RNb) for each Nb
        Nb=(N(2:end)-1);
        RNb=(max(Nb-alpha,0).*max(Nb-alpha+1,0))./max(Nb,1).^2;
        %3.calculate diagonal below in matrix A
        vxb=lam*RNb.*Nb;

        A=diag(vx)+diag(vxa,1)+diag(vxb,-1);
        A=sparse(A);

% define the range for age at sampling
tspan=[0:maxt];

%calculate the row position of n0
y0=zeros(length(N),1);
posn0=n0+1;y0(posn0)=1;
y0=y0(:);

% set options for solving the differential equations
options=odeset('Jacobian',@jacobian);
options=odeset(options,'RelTol',1e-3,'AbsTol',1e-3);

% call ordinary differential equation solver ode15s which returns
% Pn and T given dPN/dt (see function dydt below), tspan, y0 and options
[T,Pn]=ode15s(@f,tspan,y0,options);

function dydt = f(t,y)
dydt=A*y;
end

function dfdy = jacobian(t,y)
dfdy=A;
end

end
```

```
% program to calculate log likelihood over parameter grid for two time
% points for 40 DM1 individuals

clear all

% parameter grid
% set grid search values for contraction (mu)
it1=1:6:121;
M1=size(it1,2);

% set grid search values for net expansion (expansion minus contraction)
% (phi)
it2=0.1:0.24:5.1;
M2=size(it2,2);

% set grid search values for length parameter (alpha)
it4=[0:10:150,175:25:250];
M4=size(it4,2);

% set grid search values for inherited repeat length (n0)
it6=[50:10:150,175:25:800];
M6=size(it6,2);

% calculate the number of parameter combinations
% nct=M1*M2*M4*M6-parameter combinations where n0<alpha

% define mu, phi, alpha, n0 for each possible parameter combination
% initialise nmuct (contraction) nlamct (expansion) nalphct (length
% parameter) nn0ct (inherited repeat length)
nmuct=zeros(nct,1);nlamct=zeros(nct,1);nalphact=zeros(nct,1);nn0ct=zeros(nct,1);
ct=0;
for jb1=1:M1
    for jb2=1:M2
             for jb4=1:M4
                 for jb6=1:M6
                    if i6(jb6)>i4(jb4)
                    ct=ct+1;
                    nmuct(ct)=i1(jb1);
                    nlmmct(ct)=i2(jb2);
                    nlamct(ct)=i2(jb2)+i1(jb1);
                    nalphact(ct)=i4(jb4);
                    nn0ct(ct)=i6(jb6);
                    else
                    end
                end
            end
       end
end

% define LP (log likelihood) for each sample (80) and each parameter
% combination
LP=-1e32*ones(80,ct);

% calculate LP for each sample (80) and each parameter combination
% lam (expansion), mu (contraction), alpha (length parameter), n0
% (inherited repeat length)
```

```
for jb1=1:ct
    lam=nlamct(jb1)/100;
    mu=nmuct(jb1)/100;
    alpha=nalphact(jb1);
    n0=nn0ct(jb1);


% call Pnmat_sol which returns Pn (the probability that a repeat length is
% length n at time T given the parameter combination) and a finite limit for
% T (74 years) and n (3000 repeat units)

[T Pn]=Pnmat_sol(lam, mu, alpha, n0, 74, 3000);
% Assign very small probability to exceptional parameter combinations that
% return negative Pn values
Pn=max(F,1e-32);

% calculate LP (log likelihood) for each sample (80)
% LP = sum log Pn (x,t) where x is a vector of the sample lengths
% and t is age at sampling
% x and t are returned by calling xptall for each sample in turn
    for jn=1:80
            [x t]=xptall(jn);
            LP(jn,jb1)=sum(log(Pn((floor(t)+1),x+1)));
    end

end

% save definitions and log likelihood
save LPdata nlamct nmuct nlmmct nalphact nn0ct LP
```

Appendix 2

Comparison of simulated cell data with actual cell data for six DM1 individuals.

Simulations deriving from the parameter estimations

The maximum likelihood approach provided point estimates of the parameter values which best fit the data. Here we use these parameter estimates (n_0, λ, μ and a and age at sampling t) to simulate the time dependent distribution under model M_{6b} using the simulation method outlined above, for six DM1 individuals with different ranges of allele lengths. We then compare the simulated distribution of CTG repeat lengths (measured in CTG units) at age of sampling to the autoradiographs for each DM1 individual. This provides a visual representation of the model fit.

Figure A1.1

Comparison of simulated cell data (A) with actual cell data (B)

Panel A. A sample computed from the calibrated mathematical model, showing the distribution of CTG repeat units (105 cells) for DM1 individual SCO132 (aged 18 years when the sample was taken) using parameter estimates (λ = 0.94 CTGs per CTG unit per year, μ = 0.91 CTGs per CTG unit per year, a = 2 CTGs and n_0 = 514 CTGs and t = 18 years) associated with the maximum likelihood value.

Panel B. Small pool PCR analysis of repeat length variation at the expanded DM1 CTG repeat in the blood DNA of individual SCO132 at age 18 years. The scale on the right shows the length of the fragments in CTG repeats.

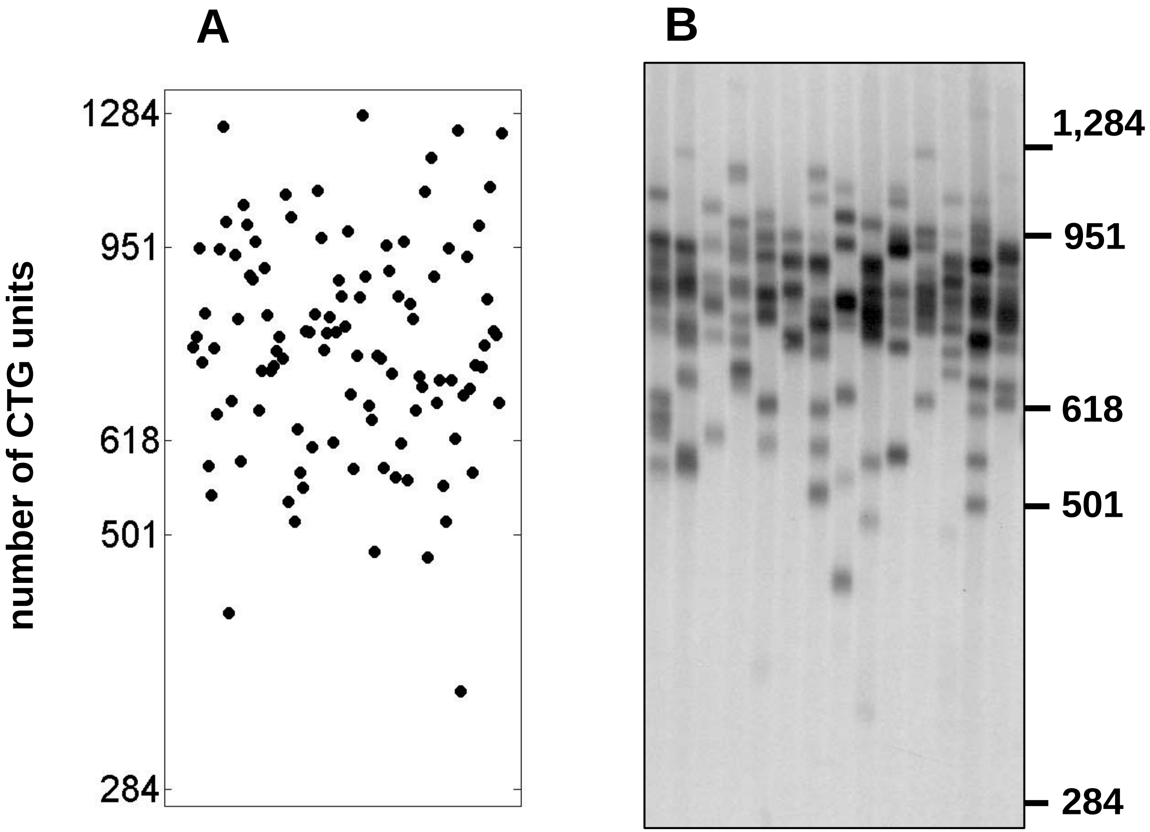


Figure A1.2

Comparison of simulated cell data (A) with actual cell data (B)

Panel A. A sample computed from the calibrated mathematical model, showing the distribution of CTG repeat units (105 cells) for DM1 individual CR69 (aged 14 years when the sample was taken) using parameter estimates (λ = 0.27 CTGs per CTG unit per year, μ = 0.25 CTGs per CTG unit per year, a = 50 CTGs and n_0 = 399 CTGs and t = 14 years) associated with the maximum likelihood value.

Panel B. Small pool PCR analysis of repeat length variation at the expanded DM1 CTG repeat in the blood DNA of individual CR69 at age 14 years. The scale on the right shows the length of the fragments in CTG repeats.

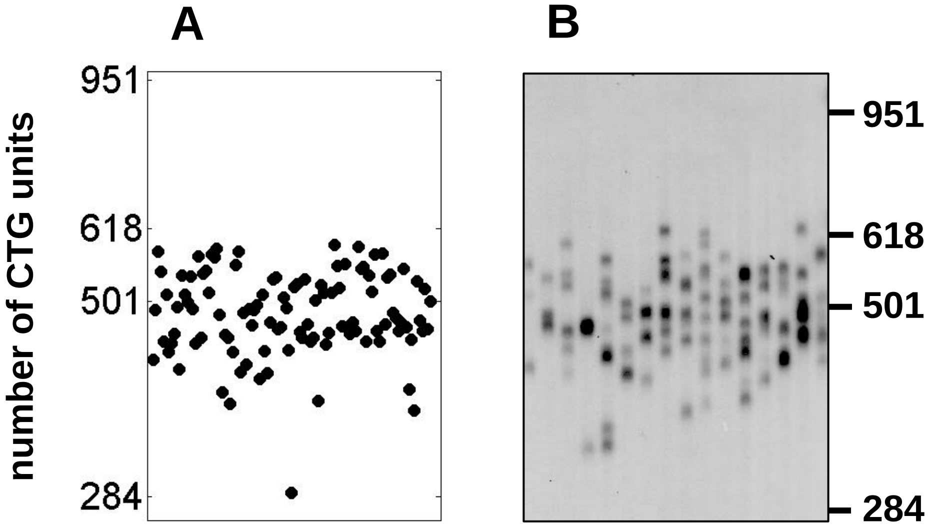


Figure A1.3

Comparison of simulated cell data (A) with actual cell data (B)

Panel A. A sample computed from the calibrated mathematical model, showing the distribution of CTGs (105 cells) for DM1 individual SCO107 (aged 43 years when the sample was taken) using parameter estimates (λ = 0.24 CTGs per CTG unit per year, μ = 0.19 CTGs per CTG unit per year, a = 48 CTGs and n_0 = 103 CTGs and t = 43 years) associated with the maximum likelihood value.

Panel B. Small pool PCR analysis of repeat length variation at the expanded DM1 CTG repeat in the blood DNA of individual SCO107 at age 43 years. The scale on the right shows the length of the fragments in CTG repeats.

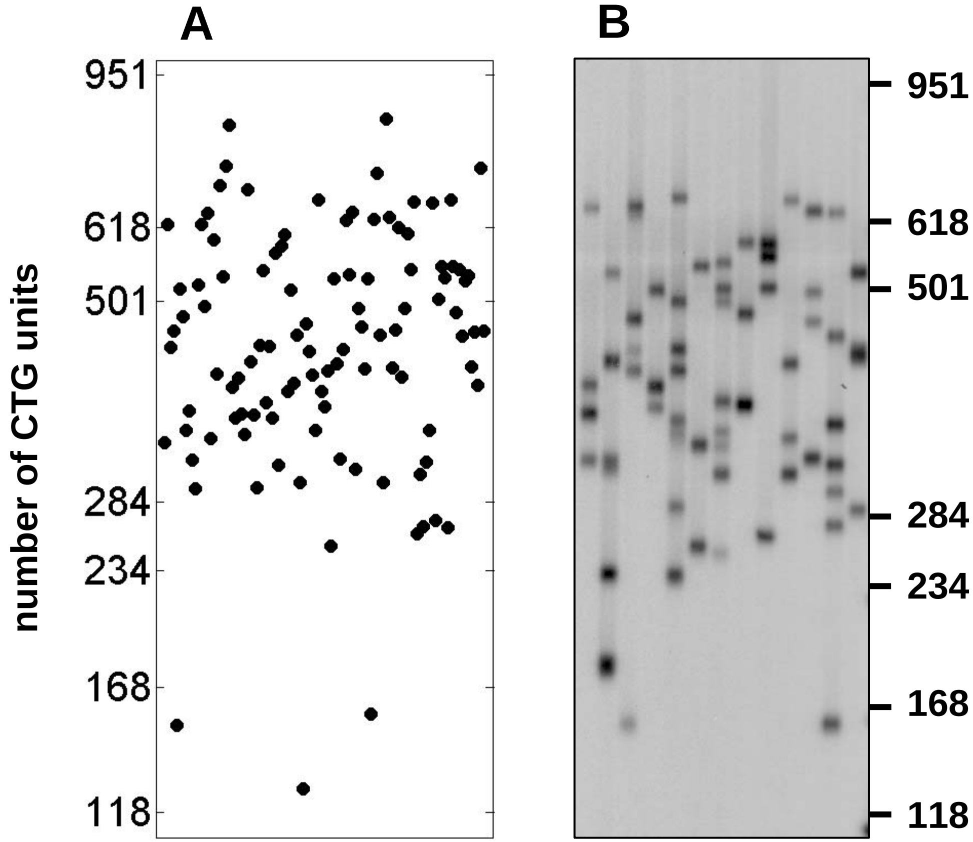


Figure A1.4

Comparison of simulated cell data (A) with actual cell data (B)

Panel A. A sample computed from the calibrated mathematical model, showing the distribution of CTGs (105 cells) for DM1 individual SCO95 (aged 52 years when the sample was taken) using parameter estimates (λ = 0.50 CTGs per CTG unit per year, μ = 0.49 CTGs per CTG unit per year, a = 50 CTGs and n_0 = 192 CTGs and t = 52 years) associated with the maximum likelihood value.

Panel B. Small pool PCR analysis of repeat length variation at the expanded DM1 CTG repeat in the blood DNA of individual SCO95 at age 52 years. The scale on the right shows the length of the fragments in CTG repeats.

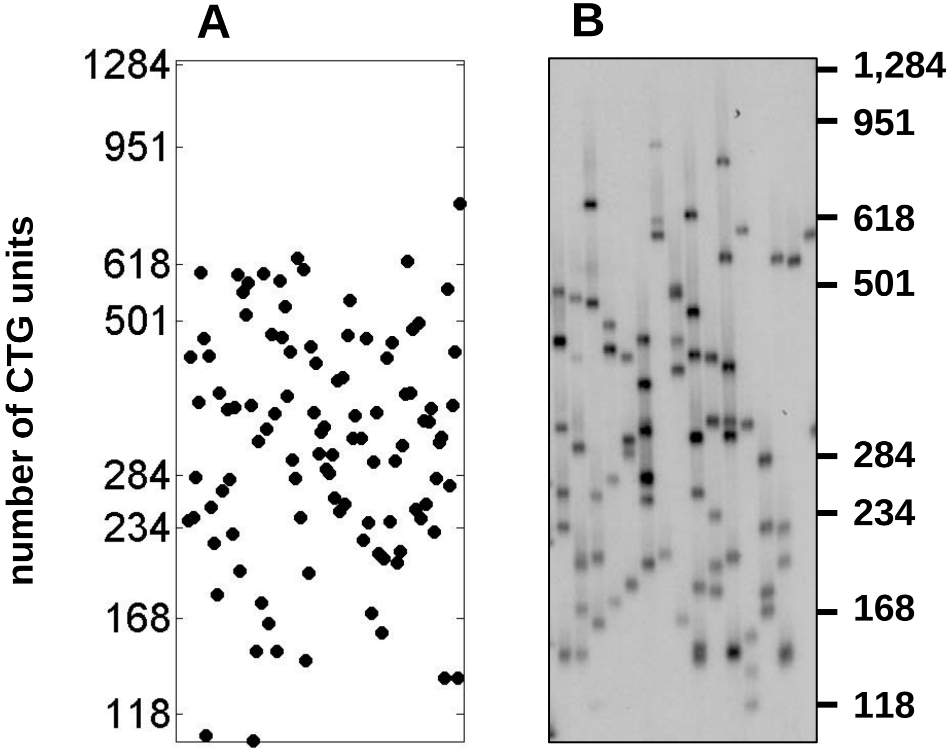


Figure A1.5

Comparison of simulated cell data (A) with actual cell data (B)

Panel A. A sample computed from the calibrated mathematical model, showing the distribution of CTGs (105 cells) for DM1 individual CR94 (aged 16 years when the sample was taken) using parameter estimates (λ = 0.29 CTGs per CTG unit per year, μ = 0.25 CTGs per CTG unit per year, a = 49 CTGs and n_0 = 255 CTGs and t = 16 years) associated with the maximum likelihood value.

Panel B. Small pool PCR analysis of repeat length variation at the expanded DM1 CTG repeat in the blood DNA of individual CR94 at age 16 years. The scale on the right shows the length of the fragments in CTG repeats.

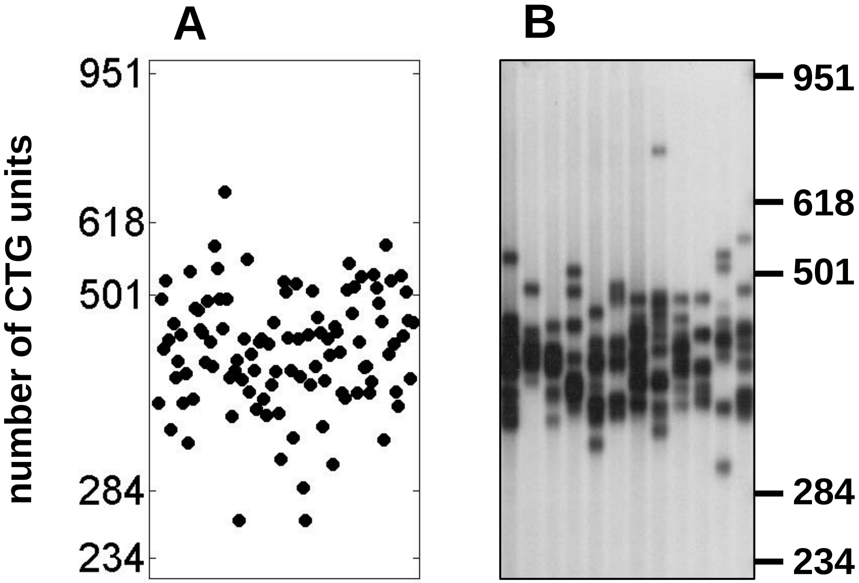
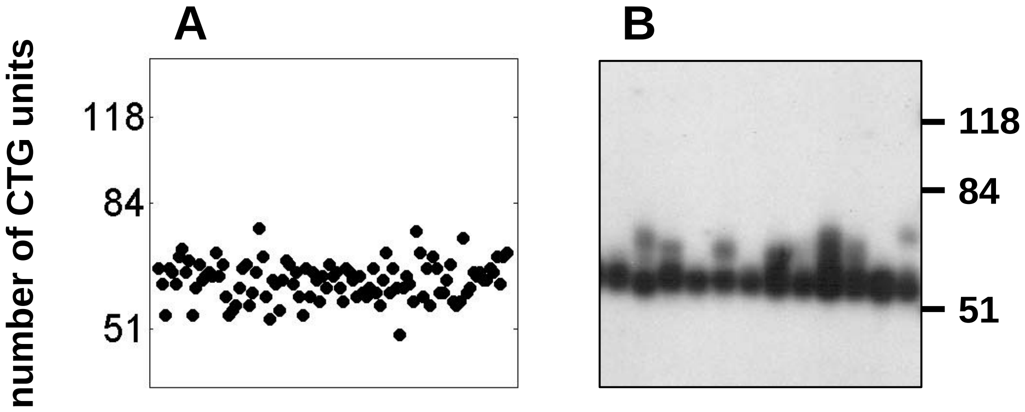


Figure A1.6

Comparison of simulated cell data (A) with actual cell data (B)

Panel A. A sample computed from the calibrated mathematical model, showing the distribution of CTGs (105 cells) for DM1 individual CR118 (aged 65 years when the sample was taken) using parameter estimates (λ = 0.014 CTGs per CTG unit per year, μ = 0.003 CTGs per CTG unit per year, a = 45 CTGs and n_0 = 53 CTGs and t = 65 years) associated with the maximum likelihood value.

Panel B. Small pool PCR analysis of repeat length variation at the expanded DM1 CTG repeat in the blood DNA of individual CR118 at age 65 years. The scale on the right shows the length of the fragments in CTG repeats.


A
B
number of CTG units