University of Glasgow
Institute of Biomedical and Life Sciences
Division of Molecular Genetics

Cis-acting modifiers of trinucleotide repeat instability

Colm Eamonn Nestor

Thesis submitted for the degree of Doctor of Philosophy

May 2008

Abstract

The dynamic expansion of CAG•CTG repeats in otherwise unrelated genes is responsible for a growing number of late-onset progressive disorders, including Huntington disease, myotonic dystrophy type 1 (DM1) and the spinocerebellar ataxias. As toxicity increases with repeat length, the intergenerational expansion of unstable CAG•CTG repeats leads to anticipation, an earlier age-at-onset in successive generations in these disorders. Crucially, disease associated alleles are also somatically unstable and continue to expand throughout the lifetime of the individual. In addition, evidence suggests that *cis*-acting elements may be major modifiers of instability.

Here it was found that the toxicity of expanded polyQ-encoding CAG•CTG tracts correlates with both the expandability of the underlying CAG•CTG repeat and the GC content of the genomic DNA flanking sequences. PolyQ toxicity does not correlate with properties of mRNA or protein sequences, or with polyQ location within the gene or protein. These data thus strongly suggest that the observed inter-locus differences in polyQ toxicity are not mediated by protein context effects, but that the rate at which somatic expansion of the DNA delivers proteins to their cytotoxic state is a critical factor in expanded polyQ-disease age-at-onset.

Using human and mouse cell lines transgenic for an expanded human *DM1* locus, it was found that an expanded CTG_{142} repeat alone is not sufficient for instability. Moreover, by generating mouse cell lines stably transfected with both a stable and unstable expanded CTG repeat, it was possible to assay the effect of *cis*-elements on these two loci in the same cell line over time. The sequences flanking the unstable repeat were hypermethylated, whereas the sequences flanking the stable transgenic repeat were unmethylated, suggesting an association between CpG methylation and repeat instability. However, methylation of the stable transgenic repeat failed to induce instability. In addition, it was revealed that transcription of an expanded repeat was not sufficient to induce instability.

Analysis of genome-wide CAG•CTG microsatellite instability revealed a significant correlation between flanking sequence GC content and microsatellite mutability. This association was most significant for short (< 7 repeats) microsatellites and for those microsatellites located within exons. However, comparison of microsatellite lengths in the human and chimpanzee genomes revealed a complex association between flanking GC content and misalignment mutations at microsatellite loci, suggesting that the modifying effect of flanking GC content on expanded repeat instability may be specific to the expanded repeat disease loci

In conclusion, this work suggests that the rate of somatic repeat expansion is a major modifier of disease progression, and that *cis*-acting elements in turn, modify repeat instability.

Table of Contents

List of Tables

List of Figures

Acknowledgements

There are many to whom I owe a great debt of gratitude for making the four years of my PhD a thoroughly rewarding and enjoyable time. First of all, I would like to thank Darren for his patience, guidance and most of all his willingness to teach me; not just about the world of tri-nucleotide repeat disorders but about the critical assessment of science in general. I could not have hoped for a better role model to learn from and aspire to. Of course, Darren also taught me that you don't have to be good at football to be good at science, thankfully! Peggy Shelbourne, Mark Bailey and Richard Wilson are also owed my sincerest thanks for their constant support, scientific input and willingness to help. I'd also like to thank all those with whom I worked in Anderson College over the last four years for their help and friendship. In particular, I'd like to thank Graham, Christine, Claudia, Fernando, John, Jill, Berit, Meera and Alison who made working in Anderson College (and drinking in Gallus) a super experience.

I thank the Wellcome Trust for funding me, and the local program coordinators, Bill, Darren, Dave and Olwyn for their continual support and encouragement.

I started my PhD along side eight others, and as classmates and friends go, I could not have been luckier. Thank you Liz, Jana, Adrienne, Theo, Alex, Christine and Mridu for all your support, and the frequent pep talks at Wednesday lunch.

There is very little that I have achieved in life that I do not owe in full to the love and devotion of my parents, and my PhD is no exception. Although it's far too modest a reward for all you have given me, this thesis is dedicated to you.

Last, but by no means least I thank Josefin. Thank you for your unswerving support and love through all my moods and bouts of PhD-induced man-flu. You're the best!

Declaration

The research reported in this thesis is my own original work, except where otherwise stated. Some preliminary analyses of the data presented in Chapter 3 were carried out as part of an MRes rotation project. The research presented here has not been presented for any other degree, except where otherwise stated.

Colm Nestor
May, 2008

Abbreviations

°C	Degrees Celsius
bp	Base pair
CAG•CTG	Trinucleotide of cytosine, adenosine and guanine
cDNA	Complementary deoxyribonucleic acid
CTCF	CCCTC-binding factor (zinc finger protein)
DEPC	Diethylpyrocarbonate
DM	Myotonic dystrophy, *dystrophia myotonica*
DMEM	Dulbecco's modified Eagle medium
DRPLA	Dentatorubral pallidoluysian atrophy
EDTA	Ethylenediaminetetracetic acid
EtBr	Ethidium bromide
FBS	Foetal bovine serum
FMR	Fragile X mental retardation
FRAXA	Fragile X syndrome
FRAXE	Fragile X site E
FRDA	Friedreich ataxia
HD	Huntington disease
HNPCC	Hereditary non-polyposis colorectal cancer
HyTK	hygromycin/thymidine kinase fusion gene
LB	Luria Bertani
M	Molar
MLH	MutL homologue
MMR	Mismatch repair
MSH	MutS homologue
neo	Neomycin
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PMS	Post-meiotic segregation
SBMA	Spinal and bulbar muscular atrophy
SCA	Spinocerebellar ataxia
S-DNA	Slipped-stranded deoxyribonucleic acid
SDS	Sodium dodecyl sulphate
SI-DNA	Slipped-stranded intermediate deoxyribonucleic acid
SP-PCR	Small pool polymerase chain reaction
SV40	Simian virus 40
UTR	Untranslated region
UV	Ultraviolet

Gene symbols

AR	*androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)*
ATN1	*atrophin 1*
ATXN1	*ataxin 1*
ATXN2	*ataxin 2*
ATXN3	*ataxin 3*
ATXN7	*ataxin 3*
CUGBP1	*CUG triplet repeat, RNA binding protein 1*
DMPK	*dystrophia myotonica-protein kinase*
FEN1	*flap structure-specific endonuclease 1*
FMR1	*fragile X mental retardation 1*
FXN	*frataxin*
HTT	*huntingtin*
MBNL1	*muscleblind-like (Drosophila)*
neo	*neomycin*
PGK1	*phosphoglycerate kinase 1*
PMS	*Post-meiotic segregation*
PPP2R2B	*protein phosphatase 2 (formerly 2A), regulatory subunit B, beta isoform*
SIX5	*SIX homeobox 5*

1. Introduction

1.1 Trinucleotide repeats and human disease

The vast majority of tandemly repeated sequence elements in the human genome are thought to be non-pathogenic. However, a growing number of human diseases have recently been associated with the expansion and instability of tandem DNA repeats, tri-nucleotide repeats (TNRs) comprising the largest class of such repetitive elements (Gomes-Pereira and Monckton, 2006). TNR instability diseases can be further categorised into two principal classes depending upon repeat position relative to the associated gene (Figure 1.1). The first class, which includes myotonic dystrophy type 1 (DM1) and fragile X syndrome (FRAXA), is defined by a repeat expansion in a non-coding region of the gene. Whereas the second class is defined by a polyglutamine (polyQ)-encoding CAG repeat. The repeats at all these loci are typically small (~5 - 30 repeats), polymorphic and stably transmitted within the general population. Disease associated alleles in patients have expanded beyond this range and typically contain at least 35 repeats. Although non-coding alleles, such as the DM1 repeat, may expand to thousands of repeats, inherited polyQ-coding alleles rarely exceed 100 repeats in humans (Gusella and MacDonald, 2000).

Expanded tri-nucleotide repeat instability is described as a 'dynamic mutation', as the frequency and magnitude of length changes vary as the repeat number changes (Richards and Sutherland, 1992). These dynamic mutations are biased towards expansion, giving rise to increases of allele length from one generation to the next. Significantly, repeat toxicity increases with length, longer repeats resulting in greater levels of cell death and dysfunction in affected tissues, and a more severe phenotype in each disorder. Therefore, intergenerational increases in expanded triplet repeat length is consistent with 'anticipation', a clinical characteristic common to these disorders, whereby an earlier age of disease onset and increased severity of symptoms is seen in successive generations (Gomes-Pereira and Monckton, 2006).

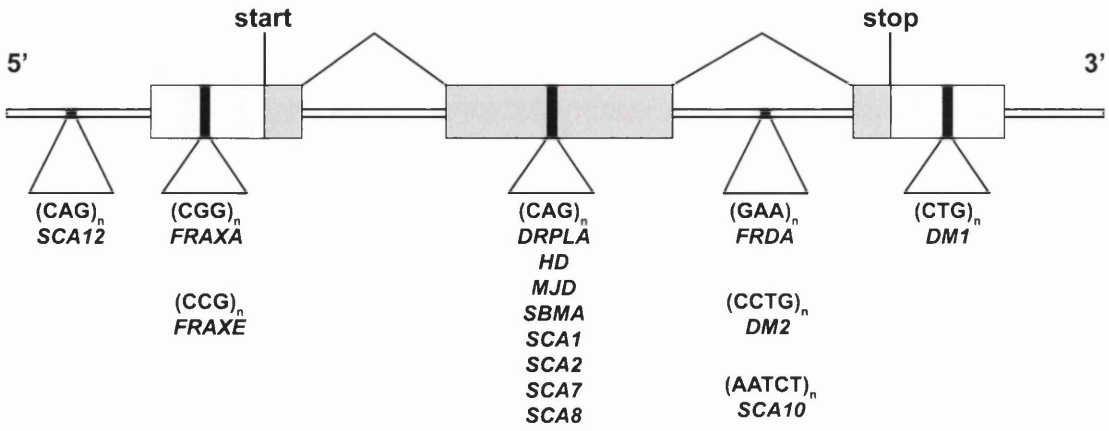


Figure 1.1 Genic location of disease associated unstable repeats. Transcription start and stop site are indicated. Exons are indicated as large boxes. Speckled boxes indicate untranslated regions, grey boxes indicate coding regions, and horizontal bar represents intergenic and intronic regions. The unstable repeat disorders shown are Huntington disease (HD), spinal-bulbar muscular atrophy (SBMA), dentatorubral-pallidoluysian atrophy (DRPLA), Machado-Joseph disease (MJD), myotonic dystrophy type 1(DM1) and 2 (DM2), Friedreich ataxia (FRDA), Fragile X syndrome (FRAXA), Fragile XE mental retardation (FRAXE) and the spinocerebellar ataxias, type 1, type 2, type 7, type 8, type 10 and type 12 (SCAs). After Gomes-Pereira and Monckton (2006).

In addition to intergenerational expansion, high levels of age-dependent, tissue-specific, expansion-biased somatic mosaicism occurs throughout the lifetime of affected individuals. Analysis of post-mortem brain tissue from Huntington disease (HD) patients found high levels of somatic mosaicism and very large expansions in the striatum, the primary affected tissue in this disorder (Kennedy *et al.*, 2003). Similarly, DM1 patients have both significantly larger absolute repeat lengths and broader ranges of expansion length in muscle compared with blood, emphasising the relationship between tissue-specific somatic mosaicism and pathogenesis (Anvret *et al.*, 1993; Ashizawa *et al.*, 1993; Thornton *et al.*, 1994). Thus, it has been proposed that whilst intergenerational repeat expansion accounts for the phenomenon of anticipation, somatic mosaicism may be a major contributing factor in disease progression and tissue specificity of symptoms (Gomes-Pereira and Monckton, 2006).

1.2 Non-coding trinucleotide repeat disorders

The non-coding repeat disorders differ in repeat sequence type, mechanisms of pathogenesis and modes of inheritance, tend to have larger repeat size ranges than the coding trinucleotide disorders and are often multi-systemic in nature. These disorders are discussed below and summarised in Table 1.1.

1.2.1 Fragile X syndrome and Fragile XE mental retardation (FRAXE)

Fragile X syndrome is the most common form of inherited mental retardation in humans, and is caused by the expansion of a CGG•CCG repeat in the 5' UTR of the *FMR1* gene (Kremer *et al.*, 1991; Yu *et al.*, 1991). The normal allele typically ranges in length from 7 to ~55 repeats, whereas the disease associated allele is typically > 230 repeats in length. Expansion into the disease-associated range is accompanied by concomitant hypermethylation of the repeat tract and surrounding CpG island, resulting in transcriptional silencing of the *FMR1* gene. Loss of the *FMR1* gene product (an RNA-binding protein) results in clinical features such as mild to severe mental retardation, facial abnormalities, macroorchidism, hyperactivity and autistic features (Debacker and Kooy, 2007; Jin and Warren, 2000). Carriers of a 'pre-mutation allele' (70 - 200 repeats) are usually unaffected, but may suffer from a poorly characterised disorder termed fragile X associated tremor ataxia syndrome (FXTAS) (Hagerman and Hagerman, 2004). The clinical features of FXTAS include progressive action tremor and cerebellar ataxia (Jacquemont *et al.*, 2003). As the pre-mutation allele is not hypermethylated and the *FMR1* gene is over-expressed (Berry-Kravis *et al.*, 2003), FXTAS is thought to be due to a toxic RNA gain-of-function (Debacker and Kooy, 2007).

Fragile XE mental retardation (FRAXE) is caused by the expansion of a GCC•GGC repeat in the 5' UTR of the *FMR2* gene, leading to hypermethylation-mediated transcriptional silencing (Knight *et al.*, 1993; Knight *et al.*, 1994). The normal and expanded repeat length ranges are similar to those of FRAXA, and affected individuals present with mild retardation and non-specific behavioural

Table 1.1 Diseases caused by non-coding expanded trinucleotide repeats

Disease	Gene Name	Locus	Inheritance	Repeat type	Genic location	Normal range	Expanded range	Phenotype	Pathogenesis
Fragile X syndrome (FRAXA)	*FMR1*	Xq27.3	X-linked recessive	CGG•CCG	5' UTR	7 - 55	> 230	Mental retardation, facial abnormalities, hyperactivity, macroorchidism	Loss of function
Fragile XE mental retardation (FRAXE)	*FMR2*	Xq28	X-linked recessive	GCC•GGC	5' UTR	4-39	> 200	Mild retardation, behavioural abnormalities	Loss of function
Friedreich ataxia (FRDA)	*FXN*	9q13-q21.1	Autosomal recessive	GAA•TCC	Intron	7 - 22	> 200	Ataxia, cardiomyopathy, motor speech disorder, diabetes	Loss of function
Spinocerebellar ataxia 12 (SCA12)	*PPP-2R2B**	5q31-33	Autosomal dominant	CTG•CAG	5' UTR	9 - 28	66 - 78	Tremors of the upper extremities, ataxia, parkinsonian symptoms	Unknown. Possibly loss of function
Myotonic dystrophy type 1 (DM1)	*DMPK*	19q13.3	Autosomal dominant	CTG•CAG	3' UTR	5 - 30	50 - >2000	Muscle hyper-excitability, progressive myopathy, cardiac conduction defects, cataracts, mental retardation, insulin resistance	(i) Loss of function (ii) Toxic RNA gain-of-function

* Association of the SCA12 repeat with *PP2A-PR55b* is not proven

abnormalities. Unlike FRAXA, however, FRAXE is extremely rare (Cummings and Zoghbi, 2000a; Debacker and Kooy, 2007).

1.2.2 Friedreich ataxia (FRDA)

Friedreich ataxia (FRDA) is caused by the expansion of a GAA•TTC repeat in an intron of the gene *frataxin (FXN)*, resulting in a reduction of the levels of the gene product, the protein frataxin (Campuzano et al., 1997; Campuzano et al., 1996). FRDA, a progressive neurodegenerative disorder is the commonest cause of inherited ataxia in Caucasians. Symptoms, which typically manifest during puberty include gait and limb ataxia, cardiomyopathy, motor speech disorder (dysarthria), and occasionally diabetes. As FRDA is autosomal recessive, rarely occurring in successive generations, anticipation has not been observed in FRDA.

1.2.3 Spinocerebellar ataxia 12 (SCA12)

Spinocerebellar ataxia 12 (SCA12) is a rare, poorly characterised, autosomal dominant neurodegenerative disorder caused by the expansion of a CAG•CTG repeat. Affected individuals present with tremors of the upper extremities, which may progress to gait and limb ataxia and subtle parkinsonian symptoms. Evidence suggests that the repeat tract may be located within the 5' UTR of the *PPP2R2B* gene, expression of which is brain specific (Holmes *et al.*, 1999). As the range of expanded pathogenic alleles identified is very narrow (66 - 78 repeats), no correlation between expansion size and age at onset of symptoms is yet evident (Holmes *et al.*, 2001).

1.2.4 Myotonic dystrophy type 1 (DM1)

Myotonic dystrophy type 1 (DM1) is an autosomal dominant neuromuscular disorder with a broad range of phenotypes but typically characterised by myotonia, muscle weakness and progressive myopathy in adult onset individuals (Wieringa, 1994). A more severe congenital (CDM) form is associated with mental retardation, developmental abnormalities and pronounced hypotonia (Roig *et al.*, 1994). All forms of DM1 are caused by a dramatic expansion in a CTG•CAG repeat in the 3' UTR of the *dystrophia myotonica-protein kinase* (*DMPK*) gene

(Brook *et al.*, 1992). The pathophysiology of DM1 is not fully understood, and several potential pathogenic mechanisms have been proposed.

It was originally proposed that repeat expansion may hinder normal transcription resulting in haploinsufficiency for *DMPK*. However, no consistent association between repeat expansion and changes in *DMPK* mRNA or protein levels has been found, and a DMPK knockout mouse model failed to recapitulate the DM1 phenotype (Jansen *et al.*, 1996; Reddy *et al.*, 1998).

Repeat expansion-mediated alterations of local chromatin structure could result in haploinsufficiency of *DMPK* or its neighbouring downstream gene, *SIX5*. Indeed sequence analysis revealed that the CTG•CAG repeat of DM1 is also located in the promoter sequence of *SIX5*. Moreover, the DM1 CTG•CAG repeat tract forms part of a CTCF-dependent insulator element, and expansion of the repeat results in ablation of CTCF binding, heterochromatin formation and the production of antisense transcripts originating from the *SIX5* promoter (Cho *et al.*, 2005; Filippova *et al.*, 2001). However, although mouse models deficient for SIX5 developed ocular cataracts, a common phenotype in human DM1 patients, they failed to exhibit any other muscle related phenotypes of DM1 (Klesert *et al.*, 2000; Sarkar *et al.*, 2000). The finding that the myotonic dystrophy type 2 (DM2) locus, a disorder with very similar phenotype to DM1, including myotonia, cataracts, distal weakness and cardiac arrhythmias, mapped to an independent locus on chromosome 3 argued against a role for *SIX5* in DM1 pathogenesis (Day *et al.*, 1999; Ranum *et al.*, 1998). Moreover, the finding that the mutation underlying DM2 was a CCTG expansion in an intron of *CCHC-type zinc finger, nucleic acid binding protein* (*CNBP*) (Liquori *et al.*, 2001), and that transgenic mice expressing an expanded CUG containing transcript in an unrelated mRNA developed myotonia and progressive myopathy (Mankodi *et al.*, 2000) strongly suggested that DM1 pathogenesis was mediated by a toxic mRNA gain-of-function. However, it is possible that the few phenotypic differences between DM1 and DM2 are mediated by altered expression of the genes proximal to each locus.

How RNAs containing CUG/CCUG expansions mediate pathogenesis is still unclear. Expanded CUG and CCUG-containing nuclear RNA foci and widespread

RNA splicing defects have been observed in cells of both DM1 and DM2 affected individuals (Liquori *et al.*, 2001; Mankodi *et al.*, 2003). It is proposed that sequestration of members of the muscleblind-like (MBNL) protein family into RNA foci causes the observed spliceopathy (Wheeler and Thornton, 2007). In support of this hypothesis MBNL1 co-localises with ribo-nuclear foci in both human DM1 and DM2 muscle and in a DM1 mouse model (Mankodi *et al.*, 2003). Moreover, a *Mbnl1* knockout mouse model lacking an expanded CUG repeat had severe myotonia and splicing defects, further implicating MBNL1 depletion in DM pathogenesis (Kanadia *et al.*, 2003). Also, the splicing defects observed in human DM1 and DM2 muscle were found to be strikingly similar to those of both a DM1 mouse model and the *Mbnl1* knockout mouse model (Lin *et al.*, 2006). Mbnl1's alternately-spliced targets include a muscle-specific chloride channel (CIC-1), the insulin receptor (IR), and cardiac tropinin T (cTnT). The functional significance of the alternate splicing of Mbnl1's targets was highlighted by a recent study of CIC-1, in which it was shown that correction of CIC-1 splicing eliminated myotonia in mouse models of DM1 (Wheeler *et al.*, 2007).

Involvement of the first identified CUG-binding protein, CUG-BP1, in myotonic dystrophy was suggested by its association with expanded CUG-containing RNA and up-regulated expression in DM1 patients and cell culture models of DM1 (Timchenko *et al.*, 2001). However, subsequent studies have found no evidence for sequestration of CUG-BP1 into RNA foci (Lin *et al.*, 1993; Mankodi *et al.*, 2003).

1.3 Coding trinucleotide repeat disorders (the polyglutamine disorders)

Unlike the more disparate phenotypes and genetic characteristics of the non-coding trinucleotide repeat disorders, the polyglutamine disorders share many features. The polyglutamine disorders including Huntington disease (HD), spinal-bulbar muscular atrophy (SBMA), dentatorubropallidoluysian atrophy (DRPLA), Machado-Joseph disease (MJD, also known as spinocerebellar ataxia 3 (SCA3)), and the spinocerebellar ataxias, type 1, type 2, and type 7 are all late-onset

neurodegenerative disorders (Gatchel and Zoghbi, 2005; Gusella and MacDonald, 2000; Manto, 2005). Each disorder is caused by the expansion of an in-frame CAG•CTG repeat tract in a coding exon of a gene, resulting in a long polyQ stretch in the mature protein. With the exception of the polyQ tract, no significant sequence, structural or functional similarity between the various expanded-polyQ proteins has been identified. However, although the expanded-polyQ genes are ubiquitously expressed, each disorder affects a specific subset of neuronal cells, suggesting a role of protein context in pathogenesis. All polyglutamine disorders show an autosomal dominant mode of inheritance, except for the X-linked disorder SBMA. The repeat tract at these loci is typically small (5 - 15 rpts), polymorphic and stably transmitted in the normal population; expansion beyond a threshold repeat length (typically 35 - 40 glutamines) initiating pathology. The polyglutamine disorders are summarised in Table 1.2.

The precise mechanism of pathogenesis in the polyglutamine disorders is unknown. The observation that all polyQ disorders display a gain-of-function toxicity upon expansion suggested a common mode of pathogenesis. Moreover, the findings that (i), expanded polyglutamine peptides alone were cytotoxic and caused neurodegeneration in *Drosophila* (Marsh *et al.*, 2000) and that (ii), an expanded CAG repeat, ectopically expressed in the hypoxanthine phosphoribosyltransferase gene (*Hprt*) resulted in a progressive late onset neurological phenotype in mice (Ordway *et al.*, 1997), indicated that the expanded polyQ tract itself was central to pathogenesis. Loss of normal protein function may contribute to pathogenesis, however, the observation that mouse knockouts of various polyglutamine disorders do not exhibit the expected phenotype, suggests that loss-of-function is not the major mediator of polyQ pathogenesis (Matilla *et al.*, 1998).

The discovery of expanded polyglutamine-containing inclusions in cells of all polyglutamine disorders suggested that aggregation of expanded polyglutamine proteins may be a common pathogenic mechanisms shared by these disorders (Michalik and Van Broeckhoven, 2003). Indeed, previous *in vitro* biochemical studies had already predicted that expanded polyQ tracts may function as polar zippers, capable of linking β-sheets together by hydrogen bonds between their

Table 1.2 Diseases caused by coding expanded CAG•CTG repeats

Disease	Gene Symbol	Locus	Inheritance	Normal range	Expanded range	Phenotype	Tissue specificity
Spinocerebellar ataxia 8 (SCA8)	*ATXN8*[a]	13q21	Autosomal dominant (reduced penetrance)	100 – 800	110 - 250	Slowly progressive ataxia, lack of gait coordination, motor speech disorder	Cerebellum, brain stem
Spinobulbar muscular atrophy (SBMA)/ Kennedy's disease	*AR*	Xq11-12	X-linked recessive	17 - 26	40 - 52	Progressive muscle weakness, atrophy, reduced fertility	Anterior horn, bulbar region, dorsal root gangion
Spinocerebellar ataxia 1 (SCA1)	*ATXN1*	6p22-p23	Autosomal dominant	6 – 40	40 - 80	Ataxia, paralysis of extraocular muscles, motor weakness	Cerebellum, brain stem, spinocerebellar tracts
Spinocerebellar ataxia 2 (SCA2)	*ATXN2*	12q24.1	Autosomal dominant	15 – 31[b]	36 - 63	Ataxia, slow eye movement, hyporeflexia, parkinsonism	Cerebellum, brain stem, Purkinje cells
Machado-Joseph Disease (MJD)/ Spinocerebellar ataxia 3 (SCA3)	*ATXN3*	14q32.1	Autosomal dominant	13 - 36	55 - 80	Ataxia, pyramidal and extra-pyramidal signs, ophthalmoplegia, bulging eyes	Basal ganglia, brain stem, spinal cord, cerebellum
Spinocerebellar ataxia 7 (SCA7)	*ATXN7*	3p12-13	Autosomal dominant	4 – 35	38 - > 130	Ataxia, loss of vision, dementia	Basal ganglia, cerebellum, inferior olive
Huntington disease (HD)	*Hdh*	4p16.3	Autosomal dominant	6 - 35	42 - > 100	Motor disturbance, cognitive loss, dystonia, personality changes	Striatum, caudate, putamen
Dentatorubro-pallidoluysian Atrophy (DRPLA)	*DRPLA*	12p13.31	Autosomal dominant	6 - 34	49 - 83	Ataxia, chorea, dementia, myoclonic epilepsy	Cerebellar cortex, striatum, cerebral cortex, basal ganglia

[a] *ATXN8* is yet to be fully characterised
[b] 99% of normal alleles are 22 – 23 repeats long (Pulst *et al.*, 1996)

main chain amides or polar side chains leading to gradual precipitation of the affected proteins (Perutz *et al.*, 1994). However, large nuclear inclusions may represent an endpoint in cytotoxicity, smaller aggregates or soluble expanded polyglutamine acting as the cytotoxic element. Indeed, it has been proposed that inclusion body formation may function as a cellular coping response, by sequestering diffuse toxic mutant huntingtin (Arrasate *et al.*, 2004).

How aggregates could mediate pathogenesis is still unresolved. The most credible current model suggests that sequestration of transcriptional regulators by polyglutamine aggregates results in widespread misregulation of gene expression and consequential cytotoxicity. This model is supported by several observations. Firstly, many models of polyglutamine disorder show widespread alterations in gene expression (Riley and Orr, 2006). Secondly, polyglutamine aggregates co-localize with other proteins; including several transcriptional regulators such as TATA-binding protein (TBP) and cAMP response element binding protein (CREB)-binding protein (CBP) (Dunah *et al.*, 2002; Nucifora *et al.*, 2001; Perez *et al.*, 1999). Finally, short polyglutamine tracts are overrepresented in transcription factors, rendering them particularly susceptible to sequestration into aggregates. However, no depletion of TBP, Sp1 or CBP was found in brains of a HD mouse model despite the presence of nuclear inclusions suggesting that the transcription factors may be binding pre-aggregate soluble mutant huntingtin (Yu *et al.*, 2002).

The observation that polyglutamine nuclear inclusions are ubiquitin positive and often contain chaperones, suggested the possible involvement of the ubiquitin-proteosome system in pathogenesis. An early study found that expanded ataxin-1 was more resistant to degradation than wild type protein *in vitro*, and a SCA1 mouse model deficient for E3-ubiquitin ligase, a core component of the ubiquitin-proteasome pathway, showed a more severe SCA1 phenotype despite the presence of fewer nuclear inclusions in their Purkinje cells (Cummings *et al.*, 1999). Employing a novel mass-spectrometry technique allowing the measurement of the by-products of the ubiquitin-proteasome system *in vivo*, it was shown that both total poly-ubiquitin chain levels and the relative proportions of various poly-ubiquitin conjugates were massively altered in the brain of HD mouse models and in the striatum and cortex of human HD post-

mortem brain samples (Bennett *et al.*, 2007). Localization of stress-response chaperones and the co-chaperone CHIP (C-terminus of Hsc70 interacting protein) to nuclear inclusions indicate that the component polyglutamines are misfolded (Al-Ramahi *et al.*, 2006; Chai *et al.*, 1999). As the ubiquitin-proteasome system and chaperones act together to rid the cell of toxic misfolded or cleaved protein fragments, their sequestration or misregulation by polyglutamine aggregates could lead to cell dysfunction and death.

1.3.1 Spinocerebellar ataxia 8 (SCA8), an atypical polyglutamine disorder?

Spinocerebellar ataxia 8 is an autosomal dominant progressive disorder showing incomplete penetrance. SCA8 is caused by a CTG•CAG expansion; affected individuals carrying 107 - 250 CTG repeats. Interestingly, carriers of shorter (71 - 110 CTG) and longer (250 - 800 CTG) seem to show reduced penetrance. As early studies failed to identify an open reading frame encompassing the repeat or evidence of transcription in the CAG orientation, SCA8 was not thought to be a polyglutamine disorder (Koob *et al.*, 1999). However, more recent work has identified anti-sense polyglutamine-encoding transcripts and the presence of polyQ-containing nuclear inclusions in both SCA8 mice and human post mortem brain samples, suggesting that SCA8 is indeed a polyglutamine disorder (Moseley *et al.*, 2006). As both CUG and CAG transcripts originate from the SCA8 repeat, it is possible that the SCA8 phenotype results from both toxic RNA and expanded polyglutamine-mediated pathogenesis (Ikeda *et al.*, 2007). However, the atypical relationship between repeat length and disease onset, as well as the presence of SCA8 expansions in unaffected individuals and in individuals with variable diseases such as bipolar disorder and schizophrenia are yet to be explained (Schols *et al.*, 2003).

1.4 Genetic instability

Genetic instability of a trinucleotide repeat is a molecular characteristic common to all unstable trinucleotide repeat disorders. Instability of the disease-associated repeats is observed in both the germline and the soma of affected

individuals, and is likely to be a major modifier of many characteristics of these disorders including tissue-specificity, age at onset of symptoms, rate of disease progression, and the degree of anticipation. As instability in all of these disorders involves a triplet repeat, is length-dependent, expansion-biased and exhibits a grossly similar threshold length for instability, it is highly likely that similar mechanisms mediate instability in all disorders. Thus, therapeutic intervention targeted at the process of repeat expansion could potentially be applied to treatment of all unstable repeat disorders, unlike therapies directed against downstream effects of repeat expansion.

1.4.1 Germline Instability

The observation that offspring tend to have repeat tracts which differ in length from the repeat tract in their parents indicates that repeat instability occurs in the germline. As longer repeats result in a more severe phenotype, intergenerational expansion of trinucleotide repeats upon transmission from one generation to the next underlies the characteristic of anticipation; whereby an earlier age at onset and more severe phenotype is typically observed in successive generations. However, both the direction (expansion or contraction) and magnitude of length changes observed upon transmission vary greatly between disorders, and exhibit a pronounced parent-of-origin effect (Brock et al., 1999; Gomes-Pereira and Monckton, 2006). Among the polyglutamine disorders SCA7 exhibits the most pronounced germline instability. Two independent studies reported repeat length changes ranging from -13 to + 85 repeats upon transmission. Significantly, 95% of observed intergenerational contractions were observed upon maternal transmission, male transmissions typically resulting in a 4-fold greater increase in repeat length than female transmissions (David *et al.*, 1998; Gouw *et al.*, 1998). In contrast, the range of intergenerational repeat length changes in MJD patients is narrower ranging from -8 to +9 repeats. And although paternal transmissions showed greater length changes than maternally transmitted alleles, no difference in the frequency of contractions or expansions were evident between paternal and maternal transmissions (Igarashi *et al.*, 1996; Maruyama *et al.*, 1995).

Of the non-coding repeat disorders, the dynamics of repeat germline instability is best characterised for myotonic dystrophy type 1 (DM1). Although DM1 families exhibit anticipation and the concomitant expansion of the underlying CTG•CAG repeat tract, DM1 repeat length changes upon transmission show a pronounced parent of origin effect. Affected individuals presenting with the congenital form of myotonic dystrophy (CDM), almost exclusively inherit the expanded allele (500 - >2000 CTG) maternally (Brunner *et al.*, 1993; Tsilfidis *et al.*, 1992). Conversely, male transmissions of shorter DM1 alleles (< 85 CTG) tend to result in greater repeat length changes than female transmissions, explaining the observed over-abundance of founder grandfathers in DM1 families (Barcelo *et al.*, 1993; Brunner *et al.*, 1993). Indeed, transmission of long repeats from fathers to offspring often appeared to result in a contraction of repeat length (Lavedan *et al.*, 1993). Analysis of sperm samples from affected fathers revealed broad distributions of repeat size ranges which extended into the upper limit of the normal repeat size range, suggesting that some of the observed intergenerational reductions in repeat size did result from contractions in the germline. However, more significantly, these studies revealed that repeat length distributions in sperm from DM1 fathers can differ markedly from the repeat length distribution in their blood; sperm samples rarely possessing alleles of > 1000 repeats (Jansen *et al.*, 1994; Martorell *et al.*, 2000; Monckton *et al.*, 1995). Thus, the majority of apparent repeat length reductions observed in paternal transmissions are likely to be artefacts of the contrasting levels of somatic mosaicism present in the germline and blood, the tissue from which most repeat size estimates are made. Thus, accurate predictions of repeat length changes from one generation to the next are affected by parental repeat length, sex of transmitting parent, age of transmission, age of sampling, and degree of somatic mosaicism, rendering genetic counselling of affected individuals difficult.

1.4.2 Somatic instability

As outlined above, genetic instability of disease-associated expanded repeats is not only present in the germline, but also occurs in somatic tissues of affected individuals. Analysis of repeat length variation in somatic tissues from DM1 and

fragile X affected individuals by genomic DNA digestion or PCR amplification followed by hybridisation to a repeat-containing probe, revealed the presence of repeat length variation of expanded mutant alleles, as evidenced by smearing of the hybridisation signal (Devys *et al.*, 1992; Lavedan *et al.*, 1993; Mahadevan *et al.*, 1992). Moreover, the levels of somatic repeat length variation observed differed between tissues (Lavedan *et al.*, 1993). Varying degrees of somatic mosaicism have since been reported for the majority of expanded trinucleotide repeat disorders (Gomes-Pereira and Monckton, 2006).

Analysis of somatic repeat length variation in DM1 individuals found that levels of somatic mosaicism in blood correlated significantly with age (Wong *et al.*, 1995). More directly, analysis of repeat length heterogeneity in blood samples obtained from the same DM1 individuals over a 1 - 7 year time period found an increase in both expansion size and the degree of heterogeneity over time (Martorell *et al.*, 1998; Wong *et al.*, 1995). Taken together, these observations indicate that somatic mosaicism is age-dependent. A study of eight somatic tissues of a DM1-affected foetus at 20 weeks found significant repeat length variation (550 - 660 CTG) of the expanded allele, suggesting that somatic expansion of mutant alleles begins early in development (Lavedan *et al.*, 1993). Interestingly, larger studies found somatic instability of expanded DM1 repeats in foetuses after 13 weeks gestational age, and no instability in foetuses before 13 weeks (Martorell *et al.*, 1997; Wohrle *et al.*, 1995). Thus, somatic mosaicism is an age-dependent process, which begins in the embryo and continues throughout the lifetime of an affected individual.

A role for somatic mosaicism in disease pathogenesis was suggested by the finding that the muscle, the primary affected tissue in DM1, exhibited greater levels of somatic mosaicism than most other tissues (Anvret *et al.*, 1993; Ashizawa *et al.*, 1993; Monckton *et al.*, 1995). However, other tissues such as heart and kidney, typically less affected in DM1, exhibited even greater levels of instability than muscle, suggesting that other factors are involved in determining tissue-specificity of pathology (Lavedan *et al.*, 1993). Similarly, a study of HD post-mortem tissues found levels of somatic mosaicism to be greatest in the brain, particularly the cerebral cortex and basal ganglia, regions of the brain associated with HD neuropathology (Telenius *et al.*, 1994). Whereas the

cerebellum, typically unaffected in HD brains, showed the lowest levels of mosaicism (Telenius *et al.*, 1994). Increased levels of somatic mosaicism were also observed in post-mortem DRPLA brains relative to other peripheral tissues (Ueno *et al.*, 1995). However, unlike HD, the region of the brain which typically exhibits the greatest pathology in DRPLA, the dentate-nucleus, did not exhibit the greatest levels of mosaicism, again suggesting the involvement of other tissue-specific factors in mediating pathogenesis (Ueno *et al.*, 1995).

Detailed quantitative analyses of somatic mosaicism were limited by poor resolution and failure to amplify and detect rare large alleles when estimating expansion size and heterogeneity from diffuse signals on autoradiographs. Use of small-pool PCR (SP-PCR) techniques allowed resolution of diffuse smears into distinct bands, permitting accurate quantification of the repeat length range present in affected tissues (see chapter 2) (Monckton *et al.*, 1995). Employing SP-PCR, it was found that HD brains of early or pre-symptomatic individuals showed dramatic expansion-biased instability in the striatum and cortex, but low levels in the cerebellum. Indeed, expanded alleles > 1000 CAG repeats, representing a 25-fold increase in size of the inherited progenitor allele, were observed in the striatum, further suggesting a role for somatic mosaicism in the tissue-specificity of this disorder (Kennedy *et al.*, 2003). Employing a combination of laser capture micro-dissection and SP-PCR to further dissect the relationship between somatic mosaicism and pathology, it was found that repeat length expansion tended to be greater in affected striatal neurons than in typically less affected striatal glia, and that neuronal repeat expansion progressed with early pathology (Shelbourne *et al.*, 2007). As the striatum typically shows the earliest pathology in HD, these results suggest that somatic mosaicism may also play a role in the progressive nature of HD, as well as the tissue-specificity.

Data supporting a role for somatic mosaicism in disease progression and tissue specificity is lacking for most other polyglutamine disorders. This is largely due to (i) a lack of affected human tissue samples and (ii), analysed tissue samples are generally end-stage, containing a skewed repeat range distribution due to loss of cells containing larger, more toxic repeats as observed with end-stage HD samples (Kennedy *et al.*, 2003).

1.5 Animal and cell models of expanded repeat instability

Our knowledge of the levels of intra-tissue and inter-tissue somatic mosaicism in humans is limited for all trinucleotide disorders. This is primarily due to the obvious difficulty in obtaining samples of affected somatic tissues from affected individuals throughout their lifetime, particularly for those disorders which show neuropathology. In order to facilitate the detailed study of somatic mosaicism several mouse models of expanded repeat instability have been generated which recapitulate many of the features of mosaicism observed in humans such as expansion-bias, tissue-specificity, repeat length-dependence, and age-dependence (Gourdon *et al.*, 1997; Libby *et al.*, 2003; Mangiarini *et al.*, 1997; Monckton *et al.*, 1997; Sato *et al.*, 1999; Shelbourne *et al.*, 1999; van den Broek *et al.*, 2002). Moreover, some models also exhibit intergenerational instability (Monckton *et al.*, 1997; Sato *et al.*, 1999) and progressive pathology (Reddy *et al.*, 1998; Seznec *et al.*, 2001) consistent with that observed in human patients .

1.5.1 Genetic instability in mouse models of polyglutamine disorders

Three out of four mouse lines transgenic for the first exon of the human HD gene, possessing a repeat tract of 114, 142, or 146 CAG repeats all exhibited both expansion-biased somatic and intergenerational instability, although size changes upon transmission were much smaller than those observed in humans (Mangiarini *et al.*, 1997). Interestingly, somatic instability was most pronounced and first observed in the CNS, the striatum and cerebral cortex showing particularly high levels of mosaicism, similar to the situation subsequently observed in human HD patients (Shelbourne *et al.*, 2007). Moreover, the cerebellum showed much less instability in two of the three mouse lines, again similar to that observed in humans, further implicating somatic mosaicism in tissue-specificity. The size of intergenerational transmissions increased with the age of the founder at conception suggesting that the somatic expansion was also age-dependent. Two HD knock-in mouse lines in which 72 or 80 CAG repeats were inserted into the endogenous mouse HD gene also exhibited both inter-

generational and somatic instability (Kennedy and Shelbourne, 2000; Shelbourne et al., 1999). Intergenerational expansion showed a parent-of-origin effect, paternal transmissions showing an expansion-bias, whereas maternal transmission resulted in small contractions (Shelbourne *et al.*, 1999). Employing SP-PCR it was found that CNS tissue showed higher levels of expansion-biased somatic instability than non-CNS tissues, that the striatum showed the highest levels of mosaicism, and that striatal instability increased in an age-dependent manner (Kennedy and Shelbourne, 2000; Shelbourne et al., 2007). Again, these features were consistent with those observed in human HD patients (Kennedy *et al.*, 2003; Shelbourne *et al.*, 2007).

A mouse model of DRPLA, harbouring and expressing a single-copy of the full-length human DRPLA gene containing a $(CAG)_{78}$ also exhibited both intergenerational and somatic repeat instability (Sato *et al.*, 1999). Somatic instability was expansion-biased, age-dependent, and tissue-specific, showing a broadly similar pattern of somatic heterogeneity as observed in human post-mortem tissues (Ueno *et al.*, 1995). The parent-of-origin effect observed upon transmission of the expanded allele was also comparable to that in humans, with paternal transmission showing a greater expansion-bias than maternal transmission, although the occurrence of contractions in the mouse model was far higher than that observed in human samples (Sato *et al.*, 1999).

Similarly, four mouse lines transgenic for a 13.5 kb human fragment of the SCA7 locus possessing a $(CAG)_{92}$ repeat also showed both intergenerational and somatic repeat instability. As with the HD, and DRPLA models expansion-biased somatic instability was particularly pronounced in the brain (Libby *et al.*, 2003).

Taken together the data from these HD mouse models recapitulate many of the characteristics of repeat instability observed in human patients and are consistent with expansion-biased, age-dependent, somatic mosaicism being a primary determinant of tissue-specificity and disease progression in the polyglutamine disorders.

1.5.2 Genetic instability in mouse models of myotonic dystrophy type 1 (DM1)

A series of DM1 mouse models harbouring a 45 kb segment of human DNA encompassing the *DMPK* gene containing 20, 55, and 300 CTG repeats have proved invaluable tools in elucidating the dynamics of expanded repeat stability in DM1 (Gourdon et al., 1997; Lia et al., 1998; Seznec et al., 2001; Seznec et al., 2000). The first mouse lines contained a $(CTG)_{55}$ repeat, and displayed moderate expansion-biased, instability (~1 CTG) in 7 % of transmissions (Gourdon *et al.*, 1997). Modest expansion-biased somatic instability (±6 CTG) was observed in most tissues, the kidney, liver, pancreas and brain exhibiting high levels of repeat length variation. Interestingly, only slight instability was observed in muscle and heart, the tissues which exhibit greatest pathology in DM1 (Gourdon *et al.*, 1997; Lia *et al.*, 1998). A similar model system carrying a larger $(CTG)_{300}$ repeat more accurately recapitulated features of DM1 instability observed in humans, although the length changes observed were smaller than those seen in humans. Approximately 95 % of transmissions were unstable with a ratio of expansions to contractions to no change of 90:5:5. In addition, as repeat size increased, the repeat length gain observed upon paternal transmission decreased, consistent with the observation in humans that congenital DM1 usually results from maternal transmission of larger alleles (Seznec *et al.*, 2000). Again, length and age-dependent somatic mosaicism was observed, with kidney, liver, and pancreas showing the highest levels of instability. Finally, it was shown the mice transgenic for the $(CTG)_{300}$ repeat had severe muscle abnormalities, myotonia, and nuclear RNA foci in their myoblasts, phenotypes observed in human DM1 patients (Seznec *et al.*, 2001).

A mouse knock-in model of DM1 containing the human *DMPK* region spanning exons 13 - 15 including a $(CTG)_{84}$ repeat in the orthologous position in the mouse *Dmpk* gene showed both age and tissue-specific somatic instability. Interestingly, it was found that the genetic background of the mouse harbouring the mutant alleles affected overall levels somatic instability, suggesting the role of *trans*-acting factors in mediating repeat stability (van den Broek *et al.*, 2002).

A third mouse model, of particular relevance to the work presented here, created to model instability in DM1, is transgenic for approximately 1.2 kb of the 3' UTR of the human *DMPK* gene. The transgene, *Dmt*, used to generate the mice consisted of a $(CTG)_{162}$ repeat flanked by 100 bp of 5' flanking sequence and 600 bp of 3' flanking sequence. Initial analysis of four such mouse lines (lines *Dmt*-B, C, D, and E) revealed modest levels of intergenerational and somatic instability (Monckton *et al.*, 1997). Using SP-PCR, a more detailed analysis of somatic repeat length heterogeneity revealed significant age-dependent, expansion-biased instability in the tissues of the mouse line *Dmt*-D, but not in the lines *Dmt*-B, -C, or -E. Instability was pronounced in kidney, a 20-month-old mouse possessing repeats which had expanded to more than 650 CTG. As reported for human DM1 patients skeletal muscle also showed greater instability than blood (Fortune *et al.*, 2000). Significantly, normalised for allele length, the levels of instability observed in this mouse model, are similar to those observed in blood of DM1 patients. Phenotypic data have yet to be reported for this mouse model.

1.5.3 The *Dmt*-D mouse cell line model

Although whole-animal models of expanded trinucleotide disorders have proven powerful tools in the analysis of expanded repeat dynamics in many disorders, their generation and subsequent manipulation is necessarily time-consuming and frequently prohibitively expensive. The generation of cell culture models would provide more easily manipulated models of expanded repeat instability and present novel experimental avenues. To this end, cell cultures were established from lung, eye, and kidney tissue of a 6-month-old *Dmt*-D transgenic mice, with a repeat length of $(CTG)_{173}$, as estimated from tail tip DNA at weaning (Gomes-Pereira *et al.*, 2001). After three-months in culture, all lines exhibited a typical fibroblast phenotype with a population doubling time of ~30 hr. Lung and eye cell cultures showed very little repeat length variation (± 5-10 repeats) even after 200 population doublings. The kidney cell lines showed highest levels of variability, expanding by ~30 - 50 repeats after just 20 population doublings. Thus, the relative levels of repeat instability observed in these tissues in the mouse model (Fortune *et al.*, 2000) are conserved between the cell lines. Cell lines derived from single cell clones of the kidney cell line also showed progressive expansion-biased instability, confirming that the principal dynamics

of expanded repeat instability observed in humans and mouse models is conserved in this mammalian cell model (Gomes-Pereira *et al.*, 2001).

1.6 Mechanisms of instability in the trinucleotide repeat disorders

The mechanism underlying expanded repeat instability is unknown. Given the striking similarities in repeat dynamics shared by all the trinucleotide disorders including expansion-biased, tissue-specific, age-dependent somatic mosaicism, and a broadly similar threshold length for instability, it is likely that the same process mediates instability in all disorders. As disease-severity increases with repeat length, the process of repeat expansion may offer a unique target for therapeutic action to all disorders.

1.6.1 Transcription

The involvement, if any, of repeat transcription in repeat instability is unclear, and largely supported by indirect evidence from mouse models. A role for transcription in repeat instability was suggested by the observation that in five HD mouse lines harbouring exon 1 of the human HD gene carrying an expanded CAG repeat (CAG_{112}-CAG_{144}), widespread somatic instability was only observed in the four mouse lines (lines R6/1, 2, 4, and 5) expressing the transgene. In contrast, the repeat was stable in the line, R6/0, in which expression of the transgene was not detected. Interestingly, the R6/0 line possessed a much longer repeat tract (142 CTG) than one of the lines, R6/1 (113 CTG), exhibiting somatic instability (Mangiarini *et al.*, 1997). Similarly, in the *Dmt* mouse model of DM1 (Monckton *et al.*, 1997), the highest levels of transgene expression were observed in line *Dmt*-D, the line exhibiting greatest intergenerational and somatic instability. However, inter-tissue levels in transgene expression level did not correlate with the levels of somatic mosaicism observed between tissues (Fortune, 2001). Similarly, seven mice transgenic for a 45 kb region of human DNA encompassing the entire *DMPK* gene and carrying an unstable $(CTG)_{55}$ repeat all showed expression of the repeat tract. However, inter-tissue transgene

expression levels did not correlate with inter-tissue levels of repeat instability. Interestingly, a recent *Drosophila* model of MJD reported no germline instability in lines in which the transgene was not expressed. In contrast, three lines showing germline transcription of the transgenic repeat, exhibited expansion-biased, intergenerational instability (Jung and Bonini, 2007). However, as reported in mouse models, levels of instability were not correlated with levels of transcription. A study of somatic mosaicism and transcription in human tissues of individuals affected with SBMA, reported a correlation between tissue instability and levels of the AR protein (Tanaka *et al.*, 1999).

Taken together these data suggest that the occurrence of repeat transcription, but not necessarily levels of repeat transcription, may modify repeat instability. Thus, the association between repeat transcription and repeat instability may not reflect causality, but result from the requirement of a common factor necessary for both processes, such as an open chromatin state. A potential mechanistic explanation of the association of transcription and instability is discussed in the context of mismatch repair in a subsequent section.

1.6.2 Replication

An apparently intuitive mechanistic model of expanded repeat instability suggests that polymerase slippage during replication causes the expansions and contractions observed in trinucleotide disorders (Cleary and Pearson, 2005; Richards and Sutherland, 1992; Ruggiero and Topal, 2004). It was proposed that Okazaki fragments, generated during the 5' to 3' synthesis of lagging strand DNA were central to repeat instability. Due to its untethered nature, an Okazaki fragment could slip or form non-B-DNA structures during polymerisation, consequently reannealing out of register with the template strand, leading to expansions or contractions upon repair or resolution (Richards and Sutherland, 1994). The observation that the threshold length for instability in trinucleotide repeat disorders (35-50 repeats; 105 - 150 bp) was not dissimilar to the length of Okazaki fragments (135 - 145 bp) seemed to lend weight to this expansion model. Subsequently, it has was shown that triplet repeats can form a variety of slipped-strand structures (S-DNA), and higher order structures *in vitro* (Pearson et al., 2002; Pearson et al., 1998b), although their existence is yet to be shown

in vivo. Several variations of this basic model have subsequently been proposed. It was suggested that either the direction of replication through the repeat, or the relative proximity of the origin of replication of the repeat might alter instability by changing the lagging strand involved in Okazaki fragment polymerisation or by altering the initiation site of the Okazaki fragments at the repeat site, respectively. These models suggest that either direction of replication or origin of replication is changed on the mutant chromosome relative to the normal chromosome (Cleary and Pearson, 2005; Pearson et al., 2002; Pearson et al., 1998b). However, no convincing evidence of involvement of direction of replication in repeat stability has been reported in mammalian systems.

Presence of the protein Flap endonuclease-1 (FEN-1), which acts in concert with other proteins to remove the RNA-primer at the 5' end of Okazaki fragments for subsequent ligation to the main lagging DNA, strand increased the stability of expanded GAA•TTC repeats *in vitro* whereas FEN-1 deficiency increased expanded CTG•CAG instability in a yeast (Freudenreich *et al.*, 1998). It was proposed that Okazaki fragments containing expanded repeats formed FEN-1 resistant structures, leading to inefficient digestion by FEN-1, and consequential expansion of the lagging strand repeat tract (Gordenin *et al.*, 1997). However, recent studies of FEN-1 deficiency found no effect of absence or haploinsufficiency of FEN-1 in a mouse model of DM1 instability, calling into question the suitability of yeast as a model system for expanded repeat instability (van den Broek *et al.*, 2006).

Despite the appeal of this conceptually simple model, mounting evidence from mouse models and human patient samples refutes a major role for replication in expanded repeat instability. Data from mouse models of trinucleotide instability in DM1 (Fortune *et al.*, 2000; Lia *et al.*, 1998; van den Broek *et al.*, 2002), HD (Kennedy and Shelbourne, 2000; Mangiarini et al., 1997), DRPLA (Sato *et al.*, 1999), and SCA7 (Libby *et al.*, 2003) have failed to observe a correlation between the proliferative status of tissues and their degree of somatic mosaicism. Moreover, cell lines generated from a DM1 mouse model exhibiting expansion-biased age-dependent somatic mosaicism found no correlation between cell proliferation rate and instability (Gomes-Pereira *et al.*, 2001).

Indeed, apicidin-mediated inhibition of cell-division and replication in these cells did not reduce the levels of instability observed (Gomes-Pereira, 2002). In addition, the high levels of instability in the terminally-differentiated striatum of human HD patients argues that replication is not required for instability (Kennedy and Shelbourne, 2000).

1.6.3 Mismatch repair and expanded trinucleotide instability

DNA mismatch repair (MMR) is an evolutionary conserved system which serves to repair single base-base mismatches and 1 - 10 base mispairs, called insertion/deletion loops (IDLs), which occur in DNA. It is widely assumed that these mismatches occur primarily during replication by processes such as nucleotide misincorporation, polymerase slippage, and erroneous re-annealing of template and daughter strands. Thus, the MMR system serves to further improve the fidelity of replication by correcting errors missed by DNA polymerase proofreading. The importance of MMR is emphasized by the mutator phenotype observed in cells deficient for MMR, as most dramatically presented in tumors of hereditary non-polyposis colorectal cancer (HNPCC) patients, which show genome-wide microsatellite instability (de la Chapelle and Peltomaki, 1995).

MMR is thoroughly reviewed elsewhere (Jiricny, 2006; Kunkel and Erie, 2005; Li, 2008). The MMR system was first identified and characterised in *E.coli*. Subsequently, several components of the mammalian MMR system were identified by homology to their prokaryotic counterparts, and named accordingly. These include the human homologues of *E.coli* MutS, human MutS homologue 1 (MSH1), MSH2, and MSH6. The primary role of these ATPases is to recognise and bind to mismatches and IDLs, in order to initiate repair. The heterodimer MutSα, consisting of one copy of MSH2 and one copy of MSH6 binds single base-base mismatches and small (1-2 nucleotides) IDLs (Li, 2008; Palombo *et al.*, 1995; Pearson *et al.*, 1997) (Figure 1.2 A). Whereas the MSH2-MSH3 heterodimer, MutSβ, binds larger (3 - 16 nucleotides) IDLs (Genschel *et al.*, 1998; Li, 2008; Palombo *et al.*, 1995) (Figure 1.2 B). Once bound, the MutS heterodimers recruit other repair proteins to the site of mismatch, most notably the MutL homologues, human MutL homologue 1 (MLH1), MLH3, postmeiotic segregation increased 1 (PMS1), and PMS2. The MutL homologues also

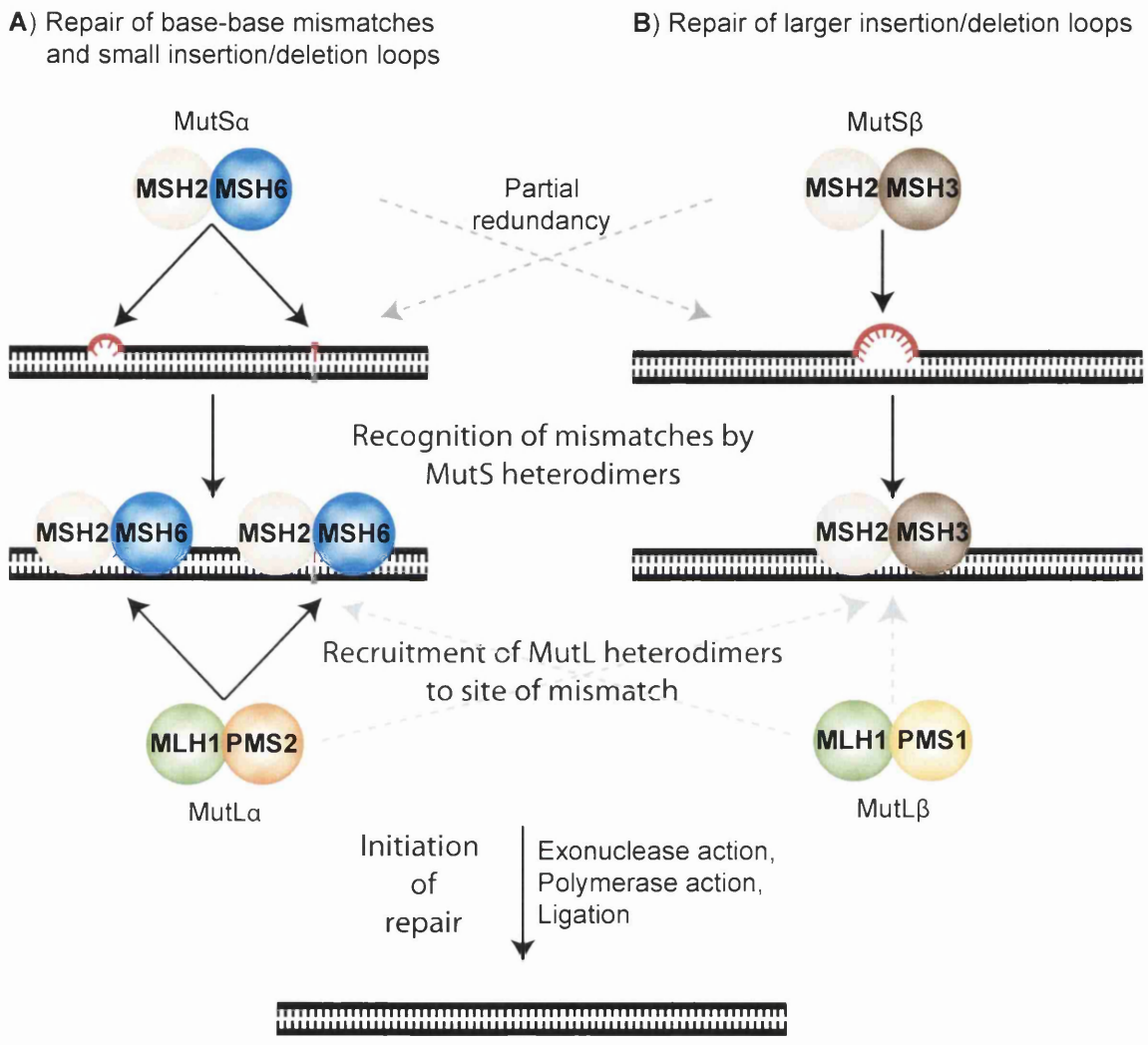


Figure 1.2. Mismatch repair of single base-base mispairs and insertion/deletion loops.
A) The heterodimer MutSα recognises both single base-base mispairs and small IDLs, whereas **B**) MutSβ binds to larger IDLs (up to 16 nucleotides). Subsequent binding of the MutL heterodimers initiates the repair process including strand discrimination, exonulcease-mediated removal of the daughter strand region, and DNA re-synthesis and re-ligation. Normal B-DNA is shown in black. Mismatches are indicated in red.

heterodimerize into MutLα (MLH1+PMS2), MutLβ (MLH1+PMS1), and MutLγ (MLH1+MLH3). Little is known of the functions of MutLβ and MutLγ. Whereas MutLα seems to play an important role in coordinating and synchronising the downstream repair processes of strand-discrimination, exonuclease removal of the daughter strand segment, DNA re-synthesis and re-ligation (Kunkel and Erie, 2005; Li, 2008) (Figure 1.2).

The observation that tumors of HNPCC patients exhibiting widespread microsatellite instability were deficient for components of the mismatch repair system, suggested that MMR could be a mediator of expanded trinucleotide repeat instability (de la Chapelle and Peltomaki, 1995; Vo et al., 2005). Indeed, instability was observed at the *DM1* locus in breast cancer tumours (Shaw *et al.*, 1996). Subsequently, it was shown that purified human MSH2 bound slipped strand structures formed in both $(CTG)_{30\text{-}50}$ and $(CAG)_{30\text{-}50}$ repeat tracts *in vitro* (Pearson *et al.*, 1997). Interestingly, the binding affinity of MSH2 increased with increasing repeat length, further suggesting a potential link between MMR and expanded repeat instability (Pearson *et al.*, 1997). However, studies using mouse models of expanded trinucleotide instability, provided the most convincing evidence for involvement of MMR in instability (Foiry *et al.*, 2006; Gomes-Pereira *et al.*, 2004; Manley *et al.*, 1999; Savouret *et al.*, 2003; Savouret *et al.*, 2004; van den Broek *et al.*, 2002; Wheeler *et al.*, 2003).

Surprisingly, when crossed with MSH2-deficient mice, a HD mouse model in which both somatic and intergenerational instability was previously observed (Mangiarini *et al.*, 1997), showed a dramatic reduction in instability at the transgenic HD locus compared to MSH2-proficient mice (Manley *et al.*, 1999). Thus MSH2, which participates in both the MutSα and MutSβ heterodimers, appeared to be required for instability, contrary to the loss of MSH2 function-induced microsatellite instability observed in HNPCC tumours. Similarly a 'humanised' mouse model of HD in which a $(CAG)_{111}$ repeat was knocked-in to the murine HD locus showed progressive somatic expansion of the transgenic repeat in MSH2-proficient ($HD^{111/+}$, $MSH2^{+/+}$) mice. Whereas, no somatic instability was observed when these mice were crossed onto an MSH2 deficient ($HD^{111/+}$, $MSH2^{-/-}$) background (Wheeler *et al.*, 2003). Moreover, the appearance

of expanded polyglutamine-staining nuclear aggregates was delayed in the MSH2-deficient mice (Wheeler *et al.*, 2003). Mice transgenic for the human *DM1* locus containing a >300 CTG repeat did not exhibit a significant change in overall mutability of the transgenic repeat when crossed onto a MSH2-deficient background ($DM1^{300/+}$, $MSH2^{-/-}$), but showed a marked shift from expansion-biased instability to a bias for contractions (Savouret *et al.*, 2003). Taken together, these data certainly indicate a requirement for MSH2 in expansion-biased somatic instability.

To further dissect the relationship between MMR and instability, mouse models of expanded repeat disorders were crossed onto backgrounds deficient for other components of the MMR system. MSH2 recognises and binds to mismatches *in vivo* only as part of the MutSα and MutSβ heterodimers, in which it is dimerized with MSH6 and MSH3, respectively. Two independent studies, in which different mouse models of DM1 repeat instability were crossed onto backgrounds lacking MSH3 ($DM1^{84\text{-}300/+}$, $MSH3^{-/-}$) showed a significant decrease in somatic instability compared to MSH3 positive mice ($DM1^{84\text{-}300/+}$, $MSH3^{+/+}$), consistent with the results observed for MSH2 deficiency (Foiry *et al.*, 2006; van den Broek *et al.*, 2002). However, when crossed onto MSH6 deficient backgrounds ($DM1^{84\text{-}300/+}$, $MSH6^{-/-}$), one study found a significant increase in the frequency of somatic expansions observed in some, but not all tissues (van den Broek *et al.*, 2002), whereas the other study reported no difference in levels of somatic instability between the MSH6-deficient ($DM1^{84\text{-}300/+}$, $MSH6^{-/-}$) mice and MSH6-proficient mice ($DM1^{84\text{-}300/+}$, $MSH6^{+/+}$) (Foiry *et al.*, 2006). The observation that MSH6-deficient mice had decreased levels of both MSH2 and MSH3 in their ovaries, suggested that MSH6 deficiency may modify instability indirectly by altering levels of MutSβ (Foiry *et al.*, 2006). Taken together, these data suggest a major role for MutSβ, responsible for recognising larger (3-15 nucleotides) IDLs, in instability.

The role of MutL proteins in instability is less well understood. Postmeiotic segregation increased 2 (PMS2) forms the MutLα dimer with MLH1. Mice, transgenic for the human DM1 CTG repeat were crossed onto a PMS2-deficient background to assess its role in somatic instability (Gomes-Pereira *et al.*, 2004). PMS2-deficient DM1 mice ($DM1^{175/+}$, $PMS2^{-/-}$) showed a ~50% reduction in the rate

of somatic expansion compared to wild-type (*DM1*$^{175/+}$, *PMS2*$^{+/+}$). However, readily detectable somatic expansion was still present in PMS2-deficient mice, and mice heterozygote for PMS2 (*DM1*$^{175/+}$, *PMS2*$^{+/-}$) showed similar levels of instability to wild-type mice, showing that PMS2 is not absolutely required for instability. The authors proposed that the observed instability in the PMS2-deficient mice could be due to partial functional redundancy between MutL heterodimers (Gomes-Pereira *et al.*, 2004).

Thus, it appears that both the MutS and MutL proteins are required for instability. The requirement of MMR for instability powerfully argues against a replication-slippage model of instability, as if somatic expansions did accrue during replication, loss of MMR would be expected to lead to an increase in the expansion rate. A cell division-independent, MMR-mediated model of somatic expansion has been proposed Figure 1.3 (Gomes-Pereira *et al.*, 2004). Complementary, non B-DNA slipped strand structures, S-DNA, form on opposite strands of a CAG•CTG repeat after DNA melting during processes such as transcription or chromatin remodelling, to which the MutSβ heterodimer binds (Figure 1.3 A-B). In the absence of a strand-discrimination signal, it was suggested that repair would be conservative, favouring incorporation of the loop-out over excision. Thus, independent repair of the complementary loop-out structures leads to expansion of the repeat by the length of one loop-out. As the IDLs bound by MutSβ are typically 3 -10 nucleotides long (though occasionally longer < 16 nucleotides), this model would lead to expansions of one to three repeats per cycle (Figure 1.3 C-H). Although this model can explain many of the characteristics of expanded repeat instability observed in humans and mouse models such as cell-division independence, requirement of a competent MMR system and expansion in small (1 - 3 repeats) units, there is no direct evidence to support some of its underlying assumption such as the formation of complementary slipped strand structures at unstable expanded loci, and preference for incorporation over excision of loop-out structures by the MMR system.

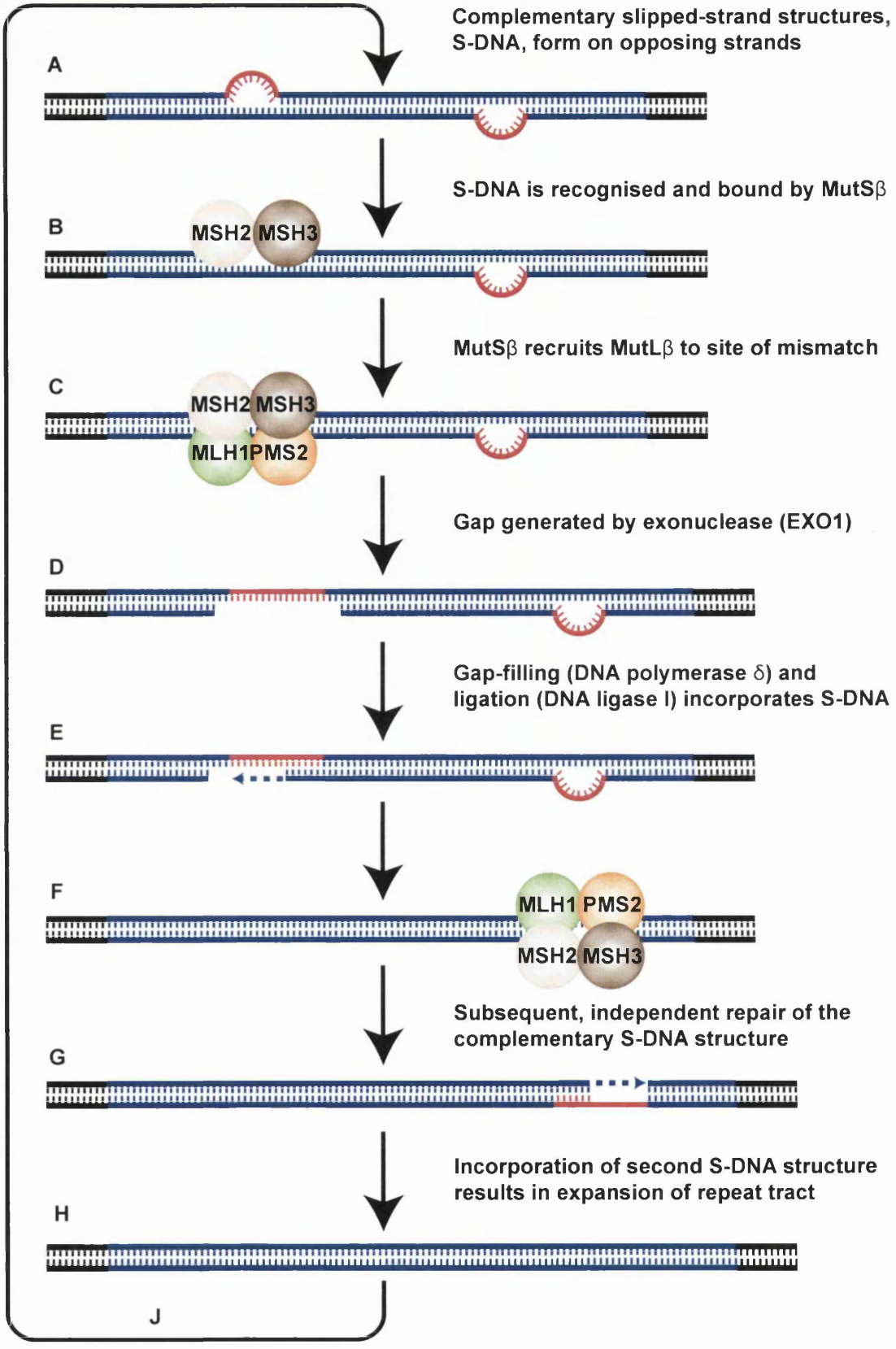


Figure 1.3 Inappropriate mismatch-repair (MMR) model of triplet repeat expansion.
A) Complimentary, small (1-3 repeats) S-DNA structures form on opposite strands of an expanded repeat tract **B**), which are recognised by the MutSβ heterodimer. **C**) MutSβ recruits a MutLβ heterodimer **D**), and subsequent exonuclease activity, possibly carried out by the protein EXO1, results in gap formation. **E**) Filling of the gap by a polymerase and subsequent ligation of the product completes repair of the S-DNA. This restores the repeat tract to its original length, that is, its length before formation of the S-DNA structures. **F-H)** Subsequent repair of the complementary S-DNA structure on the opposing strand leads to elongation of the repeat tract. Repeat DNA is indicated in blue, S-DNA in red, and non-repetitive flanking DNA in black (after Gomes-Pereira and Monckton, 2006).

The S-DNA structures central to this model have yet to be observed at expanded repeat loci *in vivo*. The instability changes observed in mouse knock-outs of MMR components may be due to other *trans*-effects induced by genome-wide mutator phenotype observed in such animals (Wheeler *et al.*, 2003). Moreover, the absence of MSH2 in one HD mouse model did not completely inhibit instability but changed the ratio of contractions to expansions observed upon transmission in a sex-of-parent dependent manner, indicating, that MSH2 independent factors are also involved in germline instability (Wheeler *et al.*, 2003).

1.7 *Cis*-acting modifiers of expanded repeat instability

1.7.1 Evidence of *cis*-acting modifiers of expanded repeat instability

In addition to the *trans*-effect of the mismatch repair system, the tissue-specificity and parent-of-origin effects on expanded trinucleotide instability suggests the involvement of many other *trans*-acting factors in repeat instability. In addition to these *trans*-acting modifiers considerable *in vivo* evidence suggests the involvement of *cis*-acting modifiers of expanded repeat instability. Although a pool of thousands of CAG•CTG microsatellites is present in the human genome, expanded repeat instability occurs at very few loci, suggesting a role for local sequence elements in facilitating instability. Moreover, despite containing the same CAG•CTG repeat configuration, many expanded repeat loci show markedly different levels of intergenerational instability when normalised for progenitor repeat length, suggesting a modifying influence of genomic location on repeat stability (Brock *et al.*, 1999). Evidence from mouse models of expanded repeat instability also suggests a role for *cis*-elements in instability. A mouse model of DM1 transgenic for a $(CTG)_{162}$ repeat and 750 bp of human *DM1* locus flanking sequence showed significantly differing levels of instability between lines, suggesting an effect of site of integration on stability (Monckton *et al.*, 1997). Interestingly, a mouse model in which mice carried a much smaller repeat (55 CTG) but were flanked by 45 kb of flanking human sequence exhibited intergenerational instability in 6 out of 7 lines (Gourdon *et al.*, 1997). Thus, shorter transgenic repeats can be rendered unstable by incorporating more

human flanking sequence. Moreover, mouse lines transgenic for a $(CTG)_{300}$ repeat tract flanked by the same 45 kb of human flanking sequence showed different patterns of somatic mosaicism between tissues (Seznec *et al.*, 2000), again implicating *cis*-sequences in modifying instability. Finally, a SCA7 mouse model in which transgenic lines carried a $(CAG)_{92}$ repeat flanked by either its full-length human cDNA sequence (cSCA7) or by 13.5 kb of its genomic sequence (gSCA7) showed strikingly different level of repeat instability (Libby *et al.*, 2003). Mice carrying the SCA7 repeat in its genomic context showed both intergenerational and somatic instability, whereas the SCA7 cDNA mice showed little instability, despite a high level of transcription. Moreover, independent mouse lines carrying the gSCA7 transgene lacking much of its original 3' sequence showed little instability, suggesting that the deleted 3' sequence contained *cis*-elements necessary for instability (Libby *et al.*, 2003). However, as few mouse lines were analysed in this study, it is possible that the observed differences were due to site-of-integration *cis*-effects, as reported in the DM mouse models.

1.7.2 Internal *cis*-acting modifiers of expanded repeat instability

The observation that all expanded trinucleotide disorders have a threshold repeat length, typically 35 - 50 repeats, below which dramatic instability is not observed, suggests that repeat-length is a major *cis*-acting modifier of repeat stability (Cummings and Zoghbi, 2000b). Moreover, alleles in the expanded disease-associated range also exhibit length-dependent somatic and intergenerational instability in both human patients (Lavedan *et al.*, 1993; Monckton *et al.*, 1995; Telenius *et al.*, 1994; Wong *et al.*, 1995) and mouse models of expanded repeat instability (Seznec *et al.*, 2000). The purity of the repeat tract at the disease loci also exhibits a powerful effect on repeat dynamics. The importance of repeat interruptions was highlighted by the observations that expanded unstable repeats at the *SCA1* and *SCA2* loci are always uninterrupted, whereas 98 % of the normal length stable alleles at these loci contain interruptions (Choudhry *et al.*, 2001; Chung *et al.*, 1993). Similarly, expanded unstable alleles at the *FMR1* locus have typically lost one or both of the AGG interruptions present in the majority of normal length alleles (Eichler *et al.*, 1994). Moreover, analysis of independent cases of expanded (> 40 CAG), but

stable, *SCA1* alleles have found the repeat tract to contain one or more interruptions (Chong *et al.*, 1995; Frontali *et al.*, 1999; Quan *et al.*, 1995). With the exception of the GAA repeat underlying Friedreich ataxia all the triplet repeat disorders are caused by an unstable repeat with the composition $(CNG)_n$, suggesting an influence of repeat sequence on instability.

How these internal cis-elements modify repeat stability is unknown. Repeat tract, length, purity and sequence may all affect the ability of the repeat to form instability mediating secondary structures. It has been shown that repeats of the type CAG•CTG have the propensity to form hairpins *in vitro*, and that the stability of these hairpins was repeat length-dependent (Gacy *et al.*, 1995). Other *in vitro* studies found that the propensity of genomic *SCA1* and *FRAXA* DNA to form S-DNA structures increased with repeat-length, whereas S-DNA structure formation was reduced by the presence of interruptions (Pearson *et al.*, 1998a). The effect of repeat characteristics on the potential to form S-DNA structures is particularly interesting in context of the S-DNA initiated MMR-mediated model of expansion outlined previously. However, it is important to emphasise that none of these repeat structures have yet been identified *in vivo*.

1.7.3 Flanking sequence composition

Normalising for progenitor allele length it was found that the intergenerational instability of expanded repeat loci differ significantly (Brock *et al.*, 1999). The same study detailed a significant positive correlation between the CG content of sequences flanking the expanded loci and their instability, suggesting an effect of flanking sequence composition on instability (Brock *et al.*, 1999). Unfortunately, data are too sparse to determine if the relationship between intergenerational instability and flanking CG content is also true for somatic instability. How flanking CG content could modify instability is not known. Flanking sequence composition could affect the ability of the repeat to form certain higher order secondary structures (Michlewski and Krzyzosiak, 2004) or affect the melting potential of the locus, and thus the potential of the repeat to form S-DNA structures. A high number of guanine nucleotides flanking the repeat may undergo oxidative damage, initiating the base excision-repair pathway, which has also been suggested as a mediator of expansion (Kovtun *et al.*, 2007).

It is possible that flanking GC content is a side-effect of higher order sequence requirements such as genic-location (i.e. proximity to the promoter region), CTCF-binding sites, nucleosome-phasing and chromatin structure, or CpG islands.

1.7.4 Epigenetic *cis*-elements

As the length threshold of repeat instability (30-50 repeats) in most expanded trinucleotide repeat disorders approximates the number of bases found in a nucleosome (146 bp), chromatin structure has been suggested has a possible modifier of repeat instability (Wang, 2007). Using electron microscopy, it was found *in vitro*, that expanded DM1 CTG repeats formed very stable nucleosomes, suggesting that the CAG•CTG repeats might profoundly alter local chromatin structure (Wang and Griffith, 1995). In contrast, a subsequent study, found that CCG repeats from Fragile X patients, displayed strong nucleosome exclusion properties in a repeat-length dependent manner (Wang *et al.*, 1996). The results of an independent study were in agreement with the previous studies and also reported that CAT or AGG interruptions within the (CAG)n and (CGG)n tracts of DM1 or fragile X syndrome, respectively, significantly reduced the nucleosome forming potential of the repeat tracts (Mulvihill *et al.*, 2005). Interestingly, in a small study employing primary fibroblast cell lines derived from DM1 individuals, loss of a DNase hypersensitive site 3' of the repeat tract in expanded alleles suggesting transition from a relaxed to more condensed chromatin formation upon repeat expansion (Otten and Tapscott, 1995). Thus, expanded repeat sequence seems to have the potential to dramatically alter local chromatin structure.

In addition to sequence composition effects, epigenetic modification of bases within and surrounding expanded loci may play a role in determining chromatin structure at disease loci. CpG methylation is involved in regulation of both DNA structure and function including X-inactivation, genomic imprinting, regulation of gene transcription and chromatin re-modelling (Bernstein et al., 2007; Takai and Jones, 2002). The majority of expanded repeat loci are located within or proximal to CpG islands, suggesting that CpG methylation of flanking sequences may be a *cis*-acting modifier of instability (Brock *et al.*, 1999). Indeed, the observation that the expanded unstable disease alleles and their flanking sequences in the *FMR1* gene are completely methylated, whereas

normal alleles possessed little or no methylation showed a direct association between CpG methylation and instability (Hornstra *et al.*, 1993). Unlike the CGG•GCC repeat at the *FMR1* locus, the CAG•CTG repeat disorders do not possess any internal CpG sites. However, hypermethylation of flanking sequences of expanded repeats was observed in congenital DM1 cases whereas shorter repeats were hypomethylated (Steinbach *et al.*, 1998). The authors suggested that methylation might ablate binding of Sp1 proximal to the repeat thereby altering chromatin structure. A methylation-mediated alteration of chromatin structure was supported by a study of epigenetic properties of the *DM1* locus (Filippova *et al.*, 2001). It was found that the DM1 repeat was flanked by two CTCF-binding sites, and that methylation of these sites resulted in ablation of CTCF-binding. CTCF is a zinc-finger DNA binding protein which can act as a barrier to the propagation of condensed chromatin along a DNA molecule, as an insulator element mediating the interaction between enhancer and promoters. It was later shown that ablation of CTCF-binding removed resulted in antisense transcription through the DM1 repeat emanating from the promoter region of the *SIX5* gene, located 3' of the *DMPK* gene (Cho *et al.*, 2005). In addition, loss of CTCF binding also affected the distribution of histone methylation across the locus, which may reflect changes in chromatin state. Significantly, CTCF-binding sites have been identified at many other expanded repeat loci (Filippova *et al.*, 2001).

However, despite the plethora of data associating epigenetic alterations with expansion of trinucleotide repeat loci, it is not clear how these changes might affect instability. Although it is possible that spread of heterochromatin at expanded disease loci affects transcription of the repeat, expression levels of expanded repeat loci do not correlate with their instability (see transcription section). Alternatively, it is possible that chromatin state modifies either the potential of the expanded repeats to form S-DNA structures or the ability of the mismatch repair machinery to correctly process S-DNA.

1.8 Project Aims

1.8.1 Hypotheses

Based on the evidence outlined heretofore, the following hypotheses were proposed:

(i) Somatic repeat instability is a major modifier of age at onset of symptoms and disease progression in expanded trinucleotide repeat disorders.
(ii) *Cis*-acting elements are major modifiers of expanded trinucleotide instability.
(iii) *Cis*-acting elements are major modifiers of microsatellite instability.

1.8.2 Aims

A considerable body of *in vitro* and *in vivo* evidence suggests a role for *cis*-acting modifiers of repeat stability. We propose to identify these elements by generating mammalian cell culture model systems of expanded repeat instability. Using published data of both locus instability and age at onset for the polyglutamine disorders, we intend to investigate the relationship between instability and disease progression. In addition, employing genome sequence data, we will attempt to identify sequence-based *cis*-acting modifiers of repeat stability using bioinformatic techniques. Similarly, we will attempt to identify *cis*-acting modifiers of genome-wide microsatellite instability by analysis of whole genome analysis of primate microsatellite flanking sequences.

It is hoped that the identification of such *cis*-elements will add to our knowledge of the process of instability in these disorders and potentially offer new routes for therapeutic intervention.

2. Materials and methods

Standard laboratory methods were performed as described in Sambrook and Russell (2001), unless stated otherwise.

2.1. Materials

The chemicals, molecular biology reagents, enzymes, kits, plastics and glassware used were obtained from suppliers such as Sigma-Aldrich Inc, New England Biolabs Ltd, Promega UK Ltd, Invitrogen Ltd and Qiagen GmbH, unless stated otherwise.

2.1.1 Cloning vectors

The vectors used to clone the transgenes generated in this project are outlined in Table 2.1.

Table 2.1: Vectors used in the course of this research project

Vector Name	Source	Use	Further Information
pIRES-EGFP	Clontech	*EGFP* cassette	
pGEM-T Easy	Promega	General cloning	www.promega.com
pBluescript II	Stratagene	General cloning	Accession #: X52328
PSURF2	D. Porteous, Edinburgh	*HyTK* cassette	(Boyd *et al.*, 1999)
pJC5-4CD	A. West, Glasgow	HS4 insulator	(Bell *et al.*, 1999)
PPNT	C. Haworth, Glasgow	*Neo* cassette	(Tybulewicz *et al.*, 1991)

2.1.2. Oligonucleotides

Custom oligonucleotides were designed with Primer3 primer design software (http://frodo.wi.mit.edu/), and were obtained from Sigma-Aldrich Inc. The primers used in the course of this research are presented in Table 2.2. Transgene binding sites are illustrated in Figure 4.2.

Table 2.2: Oligonucleotide name, sequence, melting temperature (*Tm*), and target sequence

Name	Sequence (5'-3')	Tm	Target
DM-C	AACGGGGCTCGAAGGGTCCT	72°C	*Dmt* transgene
DM-BR	CGTGGAGGATGGAACACGGAC	71°C	*Dmt* transgene
DM-H	TCTCCGCCCAGCTCCAGTCC	66°C	*Dmt* transgene
DM-F	CTGACGTGGATGGGCAAACTGC	73°C	*Dmt* transgene

DM-CR	AGGACCCTTCGAGCCCCGTTC	73°C	*Dmt* transgene
DM-ER	AAATGGTCTGTGATCCCCCCA	67°C	*Dmt* transgene
DM-PRENK	GTCCGGTACCGAATTCCGCTAGCTCCTCCCAGACCTTC	70°C	*Dmt* transgene
MDmtD-B	CACACCCTCCACTGACAGAA		Mouse genome 5' to *Dmt*D transgene
MDmtD-MR	AGCAGCTTGGATGCCTGTGGTA	60°C	Mouse genome 3' to *Dmt*D transgene
neoF1	CCTGCAGGTCAATTCTACCG	64°C	*neo* cassette
neoR1	GGGGGAACTTCCTGACTAGG	64°C	*neo* cassette
neoF4	GGCTACCCGTGATATTGCTG	64°C	*neo* cassette
T7	GTAATACGACTCACTATAGGGC	60°C	Vector sequences
T3	AATTAACCCTCACTAAAGGG	56°C	Vector sequences
M13R	GGAAACAGCTATGACCATG	54°C	Vector sequences
M13 (-21)	GTAAAACGACGGCCAGTG	58°C	Vector sequences
SP6	ATTTAGGTGACACTATAGAA	45°C	Vector sequences
CMV_F1	CAAGTCTCCACCCCATTGAC	64°C	*HyTK* cassette
Hygro_F1	GCCTGACCTATTGCATCCCC	64°C	*HyTK* cassette
TK_F1	AGAAAATGCCCACGCTACTG	64°C	*HyTK* cassette
TK_F2	TTCCGGAGGACAGACACATC	65°C	*HyTK* cassette
TK_F4	TCCTGGATTACGACCAATCG	65°C	*HyTK* cassette
TK_R1	TTGGCAAGTAGCCCGTAAAC	64°C	*HyTK* cassette
TK_R2	TCAGTTAGCCTCCCCCATC	64°C	*HyTK* cassette
mP2-1	TTCGGTGACAGATTTGTAAATG	55°C	Mouse *Pms2* gene
mP2-5	GACTTCCAAAAACCCTGGTG	63°C	Mouse *Pms2* gene

Primer extensions containing restriction enzyme sites are underlined

2.1.3 Photographic and imaging equipment

Ethidium bromide stained agarose gels were visualised on a dual intensity transilluminator, and their image recorded via a digital camera connected to a desktop computer. Digital manipulation of gel images was carried out using Adobe® Photoshop® 7.0 software.

Mammalian cells were photographed in the visual and ultra-violet range using a Canon EOS300 camera connected to a Zeiss Axiozert S100 microscope.

X-ray autoradiographs were developed using a Konica-Minolta SRX-101A tabletop film processor.

2.1.4 General solutions

Denaturing solution

0.5 M NaOH, 1.5 M NaCl in dH_2O.

Depurinating solution

0.25 M HCl in dH_2O.

Neutralising solution

1.5 M NaCl, 0.5 M Tris-HCl in dH_2O, pH 7.5.

Hybridisation solution

7% (w/v) SDS, 0.5 M sodium phosphate, 2 mM EDTA in dH_2O, pH 7.2.

20X SSC

3.0 M NaCl, 0.3 M sodium citrate in dH_20, pH 7.0.

High stringency wash solution

0.2% (w/v) SDS, 0.2 X SSC.

5X Orange G loading dye

0.06% (w/v) Orange G, 50% (v/v) glycerol in dH_2O.

1kb+ DNA ladder (Invitrogen)

60 ng/ul 1 kb ladder, 1X DNA loading dye in 1X TBE.

0.5% TBE (Tris-borate-EDTA) buffer

45 mM Tris, 45 mM Boric Acid, 1 mM EDTA in dH_2O, pH 7.0.

1× TAE

40 mM Tris•Acetate pH 8.2, 1 mM EDTA in dH_2O.

TE buffer

10 mM Tris•HCl pH 8.0, 1 mM EDTA in dH_2O.

1X Custom PCR mix (ABgene)

45 mM Tris•HCl pH 8.8, 11 mM ammonium sulphate, 4.5 mM $MgCl_2$, 6.7 mM β-mercaptoethanol, 4.4 μM EDTA, 1mM dATP, 1 mM dCTP, 1mM dGTP 1mM dTTP and 113 μg/ml BSA.

2.1.5 Tissue culture material

Tissue culture plasticware was obtained from Corning and Nalgene. Media, serum, antibiotics and reagents were obtained from Invitrogen (Gibco) and Sigma.

Dulbecco's modified Eagle medium (DMEM)

DMEM with 4500 mg/l D-Glucose, 110 mg/l Sodium Pyruvate, 862 mg/l L-Alanyl-L-Glutamine (GlutaMAX).

Foetal bovine serum (FBS)

Heat inactivated, virus and mycoplasma tested, EU origin.

Penicillin-streptomycin solution (PSS)

Stock solution: 10,000 U/ml penicillin and 10,000 μg/ml of streptomycin utilising penicillin G (sodium salt) and streptomycin sulphate: prepared in normal saline.
Working concentration: 100 U/ml penicillin, 100 μg/ml streptomycin.

Standard Growth Medium

DMEM with 10% (v/v) FBS, and 1% PSS (v/v).

Dulbecco's Phosphate-Buffered Saline (D-PBS)

200 mg/l KCl, 200 mg/l KH_2PO_4, 8,000 mg/l NaCl, and 2,160 mg/l Na_2HPO_4-$7H_2O$.

Trypsin-EDTA (10X)

Stock solution: 5 g/l Trypsin, 2 g/l EDTA•4Na, and 8.5 g/l NaCl
Working concentration: 0.5 g/l Trypsin, 0.2 g/l EDTA•4Na in D-PBS

Geneticin Liquid (G-418 Sulphate)

Stock Solution: 50 mg/ml active Geneticin® in dH_2O

Hygromycin B

Stock Solution: 50 mg/ml hygromycin B in D-PBS

2.2 Methods

2.2.1 Tissue culture methods

All tissue culture work was carried out in a dedicated tissue culture laboratory in a category 2 laminar flow cabinet. The mouse *Dmt*D kidney cell line, D2763K, hemizygous for the *Dmt*-D transgene, was supplied by Dr Mario Gomes-Pereira (Gomes-Pereira *et al.*, 2001). HeLa cells were provided by Dr Christine Haworth.

Subculturing of cell lines

Cultured cells were grown in 25 cm^2 vented flasks at 37°C and 5% CO_2. Growth media was aspirated from the culture flask taking care not to dislodge cells with the pipette tip. Cells were washed twice with 5 ml of D-PBS, after which 2.5 ml trypsin-EDTA (1 ml for HeLa cells) were added and the flask incubated at 37°C, 5% CO_2 for seven minutes (five minutes for HeLa cells). The culture was then examined under a light microscope to ascertain if the cells had successfully rounded up and lifted off the flask surface into suspension. If cells had failed to rise into suspension, the flask was gently tapped on the side and placed back in the incubator for a further 2 minutes. If the culture failed to lift from the flask surface, a sterile cell-scraper was used to dislodge cells into suspension. Trypsin was inactivated by the addition of 2.5 ml of culture medium (4 ml for HeLa cells). The suspension was repeatedly drawn into a 5 ml pipette to fragment cell clusters. An aliquot of cell suspension (typically 0.5 ml) was added to a fresh culture flask and made up to a final volume of 7 ml with fresh culture medium. Split ratios varied between 1:7 and 1:14, unless stated otherwise.

Volumes of reagents used were adjusted appropriately when using larger (75cm^2 or 125 cm^2) culture vessels.

Determination of population doubling times

Following digestion with trypsin-EDTA and neutralisation with standard culture medium, a drop of cell suspension was added to a haemocytometer. Cells were counted under a light microscope and the number of cells/ml calculated. At least 100 cells were counted when possible. Cell counts from at least 3 successive splits were used to calculate the population doubling times as follows (Martin, 1994):

$$PDT = \frac{\ln(N/N_0)}{t}$$

where PDT is the population doubling time,

ln is the natural log of the number,

N is the final cell count,

N_0 is the initial cell count,

T is the time interval between N_0 and N.

Determination of antibiotic kill curves

Cells were trypsinised and resuspended in cell culture medium as outlined above. Approximately 400 cells were then added to each well of a six-well plate containing increasing concentrations (50 μg/ml - 1000 μg/ml) of the given antibiotic. The number of colonies (>5 cells) present at each concentration of antibiotic after 7 days was counted under a light microscope. Percentage survival at each antibiotic concentration was determined by expressing the number of colonies present as a fraction of the number of colonies present in the absence of antibiotic (0 μg/ml). The minimum concentration of antibiotic required to affect 0% survival after 7 days was used to select for cells stably transfected with the resistance gene.

Transfection of mammalian cells

Mammalian cells were transfected via cationic lipid transfection using the reagent Lipofectamine™ 2000 (Invitrogen) as per the manufacturer's instructions. In brief, cells were trypsinised, resuspended in cell culture

medium, and counted as outlined above. Then, 5×10^5 cells were added to wells of a six-well plate containing 2 ml of culture medium and incubated in standard conditions for ~24 hr until the culture was at least 90 % confluent.

The cell culture medium was aspirated from each well and replaced with 0.5 ml serum-free medium. Lipid-DNA complexes were formed by mixing DNA (5 ng - 30 ng) and Lipofectamine™ 2000 (5 μl - 10 μl) in 500 μl of Opti-MEM® (Invitrogen) solution, followed by incubation at room temperature for 25 min. This solution was then added to each well and incubated in standard conditions for 4 hrs, at which time a further 1 ml DMEM (20% FBS) was then added to resulting in a final concentration of 10% FBS in each well.

Approximately 24 hr later, cells were trypsinised and resuspended in cell culture medium as described above, each transfected sample being split onto four 90 mm culture dishes containing a final volume of 8 ml of standard growth medium and the desired concentration of antibiotic.

Isolation of drug resistant clones

Cell culture dishes containing transfected cells were inspected 8-12 days post-transfection by eye and light microscopy. The position of individual colonies was marked on the base of the dish with a felt tipped pen. The culture medium was removed from the dish and cells were washed gently with 5 ml of D-PBS. A dry, sterile glass ring was placed over each colony to which one drop (from a 200 μl pipette) of trypsin was then added. The dish was incubated at 37°C for 5 min. Care was taken not to disturb the rings on transfer to the incubator. The trypsinised cells were transferred to individual wells of a 24-well plate containing 1 ml standard growth medium without antibiotic. Upon reaching confluence cells were transferred to 12-well, and subsequently 6-well plates, from which they were transferred to $25cm^2$ flasks.

Cyro-preservation of mammalian cell lines

Cells were trypsinised, resuspended in cell culture medium, and counted as outlined above. 1×10^6 cells were spun down by centrifugation at 300 g for 5 min. The supernatant was removed, taking care not to disturb the pellet, and the pellet resuspended in 1ml cell culture freezing medium (Recovery™, Invitrogen). The cell solution was transferred to a cryovial (NALGENE) and placed in a "Mr Frosty" freezing container (NALGENE) overnight at -80°C. The vials were then transferred to liquid nitrogen tanks.

Extraction of nucleic acids from mammalian cells

DNA was extracted from cell cultures using a DNeasy® Mini Kit (Qiagen) as per the manufacturers instructions. RNA was extracted using an RNeasy® Mini Kit (Qiagen) as per the manufacturers instructions.

2.2.2 Molecular cloning

Standard molecular cloning

Standard molecular cloning techniques followed the instructions set out in Sambrook and Russell (2001).

Plasmid DNA was purified from bacterial cultures using a QIAprep® Spin Miniprep Kit or EndoFree® Plasmid Maxi Kit (both Qiagen) depending upon desired final yield.

DNA fragments were purified from agarose gels using a QIAquick® Gel Extraction Kit according to the manufacturers instructions.

Plasmids were carried in Library Efficiency DH5α™ cells (Invitrogen).

Molecular cloning of expanded repeats

As long tandem repeats are unstable in bacteria undergoing frequent large deletions, bacterial colonies were screened for retention of the full-length repeat after every cloning step involving growth in bacteria. To further reduce the occurrence of deletion events MAX Efficiency® Stbl2™ Competent Cells (Invitrogen) were employed where possible.

2.2.3 Small-pool PCR

A detailed description of the theory and application of small-pool PCR (SP-PCR) is given in Gomes-Pereira *et al.* (2004), and is summarised here.

Sample preparation

The quality and quantity of sample DNA was assessed by UV spectroscopy and gel electrophoresis on a 0.8% TBE agarose gel. To reduce intra-aliquot variation 400 ng of sample DNA was digested overnight with *Hin*dIII restriction endonuclease.

Digested DNA was serially diluted with 1 × TE buffer containing 0.1 μM of primer DM-C (carrier primer). The carrier primer, which is usually the forward primer to be used in the PCR reaction, serves to protect against loss of DNA due to degradation and adsorption of DNA onto the surface of the tube. Samples were typically diluted to 1 ng/μl, 500 pg/μl, 100 pg/μl and 20 pg/μl.

PCR

All SP-PCRs presented here consisted of forward primer DM-C (0.1 μM), reverse primer DM-BR (0.1 μM), 1× PCR buffer, 0.175 U *Taq* DNA polymerase, and 0.5 μl of sample DNA made up to a final volume of 7 μl with water. Each final reaction was covered with 20 μl of mineral oil to prevent evaporation. Amplification was performed in a standard bench top thermal cycler with a heated lid (105°C). An initial denaturing step was performed at 96°C for 180 s followed by 28 cycles of denaturing at 96°C for 45 s, annealing at 68°C for 45 s, and extension at 70°C for

180 s. A final annealing step of 68°C for 60 s and extension step of 70°C for 600 s were then performed.

Gel electrophoresis

Each reaction was brought to a final volume of 10 μl by addition of 3 μl loading buffer, and 5 μl of the resulting mix was loaded onto a 1.5% (w/v) TBE agarose gel (20 × 40 cm) containing 500 nM ethidium bromide. Products were resolved in 0.5× TBE running buffer in a refrigerated room. An initial voltage of 300 V was applied for 20 min followed by 200 V for 20 hr. Progression of DNA migration through the gel was determined by visualisation of the DNA ladders on a UV transilluminator.

Southern "squash" blotting

The regions of each gel not required for blotting were removed with a scalpel. Each gel was rinsed in dH_2O and immersed in depurinating solution in a large tray for 10 min on a bench-top orbital shaker. The gel was then similarly washed with denaturing solution for 30 min and neutralising solution for 30 min, rinsing the gel with dH_2O between each step.

Gels were then transferred to a bench covered with Saran Wrap. One sheet of nylon membrane (Nytran N, Amersham) followed by two sheets of blotting paper, all of which were pre-soaked in neutralising solution, were placed atop the gel; care being taken to expel any air bubbles between the layers by gently rolling a glass pipette across the surface of the nylon membrane and subsequently the blotting paper. The blot was then topped with approximately 6 cm of paper towels and a glass plate carrying a weight of 1 kg. Transfer of DNA from the gel to the nylon membrane by capillary action was allowed for 16 h, at which time the blot was deconstructed and the membrane dried at 80°C for 2 h. DNA was fixed to the membrane by exposure to 1,200 J/m^2 of UV light.

Hybridisation of PCR products transferred to the nylon membrane

Membranes were soaked in dH_2O, rolled and placed in a hybridisation bottle, to which 5 ml hybridisation solution was added. The bottle was incubated at 65°C in a rotating hybridisation oven for at least 45 min.

Double stranded DNA (~30 ng) and 2.5 ng of DNA marker were labelled with α-^{32}P dCTP using Ready-To-Go™ Labelling Beads (Amersham) according to the manufacturers instructions. The radiolabelled probe was added to the hybridisation bottle along with 5 ml fresh hybridisation solution and incubated overnight at 65°C in a rotating oven.

The hybridisation solution was poured off and the membrane rinsed with wash solution, twice at 65°C in the hybridisation tube for 30 min and finally by shaking for 30 min in a large tray of wash solution. The membrane was dried for 2 h at 80°C and exposed to an X-ray film for 4 - 16 h.

2.2.4 RNA analysis

All RNA work was carried out in a dedicated RNA lab. All solutions were made up with 0.1% (v/v) diethyl pyrocarbonate (DEPC)-treated water.

RNA extraction

Total RNA was extracted from cultured cells using an RNeasy® Mini Kit (Qiagen) according to the manufacturers instructions. RNA concentration and quality was determined by analysing sample absorbance at wavelengths of 260 and 280 nm, using a NanoDrop® ND-1000 spectrophotometer. A ratio of sample absorbance at 260 and 280 nm of 2.0 - 2.1 indicated a pure RNA sample.

Sample quality was also analysed by gel electrophoresis of 2,500 ng of RNA on a 1.35% TAE, 0.1% SDS, agarose gel. RNA was subsequently stained with ethidium bromide and visualised with UV light.

cDNA synthesis

Prior to reverse transcription, RNA samples were treated with DNase I (RQ1 RNase-free DNase I, Promega) to remove contaminating genomic DNA, according to the manufacturers instructions.

cDNA was synthesised from RNA prepared as outlined above. First strand cDNA synthesis was performed using Superscript II™ reverse transcriptase (Invitrogen) and random hexamers (Roche) according to the manufacturers instructions. 2500ng of RNA were used in all reactions. Reactions lacking reverse transcriptase (RT) were also performed. Such RT- reactions allow for determination of the contribution of contaminating genomic DNA template to subsequent PCR amplifications from cDNA samples.

RT-PCR

Amplification of cDNA was performed using gene specific primers and standard PCR parameters.

2.2.5 Methylation assays

With the exception of Turbo™ *Nae*I (Promega), all methylation sensitive restriction enzymes and methylases were obtained from New England Biolabs, and used in accordance with the manufacturer's instructions.

2.2.6 Statistics and bioinformatics

Statistics

SPSS® and Microsoft Excel were used for small-scale statistical calculations and data manipulation. For large-scale operations, statistical methods were implemented in the Perl programming language.

Bioinformatics

All custom written software was implemented in the Perl (v5.8.4) programming language on a standard desktop PC. A MySQL relational database server (v12.22) was used for storage and manipulation of large datasets.

A computing cluster of 60 dual-processor Linux servers was used to carry out BLASTing of genome-wide microsatellite flanking sequences.

Online bioinformatics resources

The following online resources were used for data acquisition and analysis during the course of the work presented here:

Ensembl Genome Browser http://www.ensembl.org/index.html

- Genomic sequence data
- Gene sequence data

NCBI http://www.ncbi.nlm.nih.gov/

- Gene sequence data
- BLAST

ExPASy Proteomics Server http://expasy.org/

- ProtScale - protein primary sequence analysis

SAM http://www.soe.ucsc.edu/compbio/sam.html

- Secondary structure prediction

Primer3 http://frodo.wi.mit.edu/

- Primer design

3. Correlation of polyglutamine toxicity with CAG•CTG triplet repeat expandability and flanking genomic DNA GC content

This chapter has been submitted as a journal article and was under review at the time of submission. It is presented here as such, and contains minor alterations suggested by Professor Darren G Monckton.

3.1 Introduction

Expanded tri-nucleotide repeat instability is described as a 'dynamic mutation', as the frequency and magnitude of length changes vary as the repeat number changes (Richards and Sutherland, 1992). These dynamic mutations are biased towards expansion, giving rise to increases of allele length from one generation to the next. Significantly, repeat toxicity increases with length, longer repeats resulting in greater levels of cell death and dysfunction in affected tissues, and a more severe phenotype in each disorder. Therefore, intergenerational increases in expanded triplet repeat length is consistent with 'anticipation', a clinical characteristic common to these disorders, whereby an earlier age of disease onset and increased severity of symptoms is seen in successive generations (Gomes-Pereira and Monckton, 2006). In addition to intergenerational expansion, high levels of age-dependent, expansion-biased, tissue-specific somatic mosaicism are also observed. Analysis of post-mortem brain tissue from HD patients found high levels of somatic mosaicism and very large expansions in the striatum, the primary affected tissue in this disorder (Kennedy *et al.*, 2003). Similarly, DM1 patients have both significantly larger absolute repeat lengths and broader ranges of expansion length in muscle compared with blood, emphasising the relationship between tissue-specific somatic mosaicism and pathogenesis (Anvret *et al.*, 1993; Ashizawa *et al.*, 1993; Thornton *et al.*, 1994). Thus, it has been proposed that whilst intergenerational repeat expansion accounts for the phenomenon of anticipation, somatic mosaicism may be a major contributing factor in disease progression and tissue specificity of symptoms (Gomes-Pereira and Monckton, 2006).

Mutant polyQ-encoding CAG tracts also cause the atypical disorders SCA6 and SCA17. However, neither can be classified as a dynamic mutation since both are

genetically relatively stable. Nearly all expanded CAG repeat SCA17 alleles are interrupted by stabilising CAA codons (Tomiuk *et al.*, 2007), whilst even 'expanded' SCA6 alleles are still relatively small (typically 20-30 repeats) (Frontali, 2001). Moreover, it seems likely that SCA6 represents a channelopathy rather than a true polyglutamine repeat disorder since truncating mutations in the same SCA6 associated *CACNA1A* calcium channel gene cause the highly overlapping episodic ataxia type 2A phenotype (Frontali, 2001).

The precise mechanism underlying the dynamic mutation of CAG•CTG repeats is unknown. The finding that levels of somatic repeat instability are independent of the proliferative status of cells (Gomes-Pereira *et al.*, 2001) and tissues (Fortune et al., 2000; Kennedy and Shelbourne, 2000; Lia et al., 1998) argues against replication-centred models of repeat expansion . Several murine models indicate that a competent mismatch repair (MMR) system is required to affect expansion at unstable loci (Foiry *et al.*, 2006; Gomes-Pereira *et al.*, 2004; Manley *et al.*, 1999; van den Broek *et al.*, 2002; Wheeler *et al.*, 2003). It has been proposed that repeat expansion may result from the inappropriate repair of small mismatched loop-outs of 1-3 repeat units formed by the incorrect re-annealing of the DNA strands after melting during the G_0 and G_1 stages of the cell cycle (Gomes-Pereira *et al.*, 2004). The small length changes of 1 to 3 repeat units predicted by such a model are consistent with those observed *in vivo*.

In addition to obvious *trans*-acting factors involved in governing expanded repeat behaviour such as the mismatch repair system, sex of the transmitting parent and tissue type, numerous lines of evidence suggest a major role for *cis*-acting factors in CAG•CTG instability. Expanded CAG•CTG instability is locus-specific, not genome-wide indicating that factors local to the repeat influence its mutability. While repeat length is obviously a major modifier of repeat stability, using length-normalised comparisons of intergenerational repeat expansion between disorders we previously found that CAG•CTG repeat loci differed significantly from one another, suggesting the involvement of other *cis*-acting modifiers of repeat stability flanking the repeat itself (Brock *et al.*, 1999). Likewise, a growing body of evidence from murine models of CAG•CTG instability

support the involvement of *cis*-elements in determination of repeat stability (Fortune *et al.*, 2000; Libby *et al.*, 2003; Mangiarini *et al.*, 1997; Monckton *et al.*, 1997; Seznec *et al.*, 2000). It has been suggested that the ability of expanded CAG•CTG repeats to form unusual secondary structures, and the cell's attempt to process or repair these structures may be the underlying source of TNR instability (Chen *et al.*, 1995; Gacy *et al.*, 1995; Geschwind *et al.*, 1997). Furthermore, it has been found that the propensity of CAG•CTG repeats to form certain DNA structures *in vitro* is both length and flanking sequence dependent (Pearson and Sinden, 1996; Pearson et al., 1998b). We have previously reported that the GC content of sequences flanking expanded CAG•CTG repeats correlates with their intergenerational expandability (Brock *et al.*, 1999). It is possible that the GC content of CAG•CTG repeats flanking sequences affects their ability to form instability-mediating secondary structures. Alternatively, the GC content of the flanking DNA may mediate differences in downstream repair processes.

As all the dynamic repeat disorders that possess an expanded polyQ tract are dominant, display a similar inverse relationship between polyQ length and age at onset, are progressive, and lead to neuronal degeneration, it is likely that the fundamental mode of polyQ toxicity is broadly conserved between disorders. This assertion is further strengthened by the finding that cleaved fragments of polyQ-proteins are cytotoxic (Li *et al.*, 2000) and that insertion of a long polyQ tract into an unrelated protein can recapitulate many features of a polyQ-disease phenotype in mice (Ordway *et al.*, 1997). Moreover, the formation of polyQ containing aggregates and transcriptional misregulation in affected tissues are molecular abnormalities clearly shared by the affected tissues in all disorders (Riley and Orr, 2006).

Although all expanded polyQ disorders show a similar inverse relationship between polyglutamine number and age at onset of symptoms, the absolute number of polyglutamine repeats required to affect a given age-at-onset of symptoms varies considerably between the disorders (Gusella and MacDonald, 2000). For example, whereas an age-at-onset of 40 years in MJD typically requires >65 repeats, <45 repeats will effect a similar age-at-onset in SCA2 (Gusella and MacDonald, 2000). These inter-locus differences in polyQ toxicity

are widely assumed to be a consequence of the different protein contexts in which each polyQ tract is found in its host protein (de Chiara et al., 2005; La Spada and Taylor, 2003; Riley and Orr, 2006). That is, the protein sequences flanking each polyQ tract are assumed to somehow influence its cytotoxic potential, resulting in markedly different toxicity thresholds between disorders. As the size of the native expanded-polyQ containing proteins varies greatly (41 kDa - 347 kDa), their primary sequences are not similar, and the position of the tract relative to the translation start site differs, it is highly likely that the polyQ tracts have very different protein contexts. The observations that polyQ tracts alone are cytotoxic, and that each disorder affects distinct sub-sets of neuronal cells, could suggest that portions of the proteins other than the polyQ tract are involved in pathogenesis. However, as the structures of most of the expanded-polyQ proteins are unknown and show no homology to proteins of known structure, how protein context mediates polyQ dynamics and toxicity is unresolved. Recent studies in yeast showed that altering the flanking sequence of an expanded *Huntingtin* exon 1 fragment, by the simple addition of a FLAG-tag, caused a previously non-toxic fragment of the htt exon 1 to induce characteristic length-dependent polyQ toxicity (Duennwald *et al.*, 2006). In addition, it was also noted that flanking sequences affected the morphology of polyQ aggregate formation, offering a potential link between polyQ protein context and polyQ toxicity. In COS cells, deletion or replacement of the Josephin domain of expanded polyQ-containing ATXN3 significantly reduced the propensity of the protein to form aggregates (Gomes-Pereira and Monckton, 2006) as did deletion or replacement of the AXH domain of the ATXN1 protein (de Chiara *et al.*, 2005). Other findings suggest that polyQ protein context could mediate cytotoxicity by affecting the ability of the ubiquitin-proteasome system to target and clear the cell of toxic expanded proteins and aggregates (Al-Ramahi et al., 2006; Chai et al., 2004; Chai et al., 2001). However, the precise role of aggregates in cytotoxicity is unresolved, and consequently, how protein context mediated effects on cellular aggregate formation and clearance play a role in disease pathogenesis is even less clear. Although dramatic alterations of flanking sequence can have profound effects on the dynamics of an individual polyQ tract *in vitro*, how to relate such observations to actual disease-associated polyQ tracts *in vivo* is very unclear. Moreover, although each individual study has shown an effect of flanking sequence on polyQ dynamics, taken together they

offer no rationalisation of the observed inter-locus difference in toxicity observed between these disorders.

As repeat length contributes significantly to pathological load, longer repeats resulting in increased cell and tissue dysfunction, we rationalise that the rate of expansion of repeats in affected tissues is a major modifier of the age-at-onset of symptoms of a disorder. It is not suggested that CAG-expandability is the cytotoxic element in these disorders, but that the rate at which somatic expandability delivers proteins to their cytotoxic state is a critical factor in expanded polyQ-disease pathogenesis, and is the underlying cause of the observed inter-locus differences in polyQ toxicity. Consequently, any modifier of the rate of somatic expansion, although not a direct cytotoxic element, will have a profound effect on disease progression within the affected individual, and may offer novel targets for therapeutic action. Here we test this hypothesis, by quantifying the relationship between polyQ toxicity, CAG•CTG expandability and flanking DNA GC content.

3.2 Results

3.2.1 Locus toxicity correlates with repeat expandability

As a prerequisite to investigating the relationship between inter-locus polyQ toxicity and CAG•CTG expandability, we carried out a detailed statistical analysis of the nature of the relationship between inherited repeat number and age at onset both within and between the seven polyQ disorders in order to obtain a robust measure and ranking of polyQ locus 'toxicity'. The majority of individuals with these disorders first develop symptoms in adult life, with a modal age at onset of 32 years. Juvenile cases, with an age at onset under 20 years, are relatively rare, but develop an extreme phenotype that is very similar between the disorders and in which the well defined regional specificity of the adult onset neuropathology is lost (Barbeau et al., 1984; Benton et al., 1998; Cummings and Zoghbi, 2000b; Geschwind et al., 1997; Squitieri et al., 2006). Because of this differential extreme phenotype and the paucity of juvenile onset data for most of these disorders, cases with an age of onset under 20 years of age have been excluded from the analyses.

Testing a range of curve estimation regression models, an exponential decay function was found to best describe the relationship between age at onset and repeat number for all disorders (Figure 3.1). Subsequently, the repeat number corresponding to an age at onset of 32 years, the modal age at onset of all disorders, was derived from the equation of the regression analysis describing the relationship between age at onset and repeat length for each disorder (Table 3.1). We propose that repeat numbers thus obtained, are a sound quantitative measure of the relative toxicity of each locus.

Taking into account the effect of progenitor allele length, we previously quantified observed differences of intergenerational variability between expanded CAG•CTG repeat loci; calculating the relative expandability of each locus using pedigree data gleaned from the literature (Table 3.1) (Brock *et al.*, 1999). Employing these values of expandability we found that locus toxicity and locus expandability were significantly correlated using a rank order test

Table 3.1. Inter-locus polyQ toxicity and expandability of the dynamic DNA polyQ loci

Locus	r[a]	*P*-value	Toxicity [b]	Rank Toxicity	Estimated Expandability [c]	Rank Expandability
MJD	0.52	<0.001	75.4	1	0.07 (0.05-0.09)	1
DRPLA	0.21	<0.001	66.1	2	0.19 (0.14-0.24)	4
SBMA	0.39	<0.001	53.3	3	0.08 (0.00-0.22)	2
HD	0.40	<0.001	52.2	4	0.29 (0.21-0.43)	5
SCA1	0.63	<0.001	51.7	5	0.14 (0.00-0.24)	3
SCA7	0.39	<0.001	48.8	6	1.30 (0.80-1.65)	7
SCA2	0.41	<0.001	40.8	7	0.97 (0.65-1.33)	6

[a] Correlation coefficient (r) of age at onset versus repeat length obtained by fitting an exponential decay model to each dataset

[b] Repeat length corresponding to an age at onset of 32 years

[c] Intergenerational instability of each disorder as described in Brock *et al.*, 1999

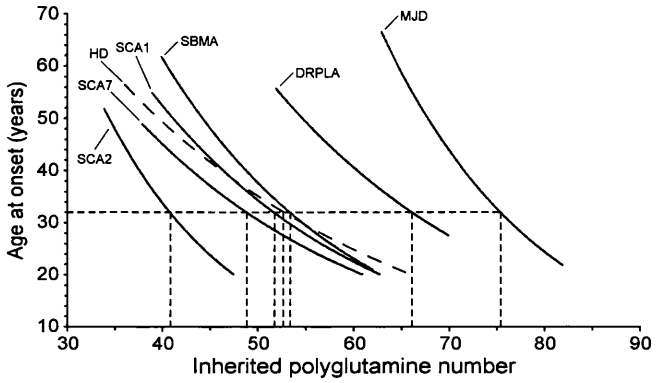


Figure 3.1. Determination of locus toxicity from regression lines describing the relationship between repeat length and age at onset for seven polyglutamine disorders. The polyQ disorders analysed were Huntington disease (HD) (dashed line), Machado-Joseph disease (MJD), spinal-bulbar muscular atrophy (SBMA), dentatorubral -pallidoluysian atrophy (DRPLA), spinocerebellar ataxia type 1 (SCA1), type 2 (SCA2) and type 7 (SCA7). Regression lines were determined using an exponential decay model. Locus toxicity was derived from the equation of the regression line of each disorder for an age at onset of 32 years (dashed line).

(Spearman's rank; *rho* = 0.79; P = 0.036; N = 7) (Figure 3.2A). Importantly, similarly significant correlations were obtained when an age at onset of 30 (Spearman's rank; *rho* = 0.78; *P* = 0.036; N = 7), 40 (Spearman's rank; *rho* = 0.86; *P* = 0.014; N = 7) or 50 (Spearman's rank; *rho* = 0.86; *P* = 0.014; N = 7) years was used to determine locus toxicity (Figure 3.2B), suggesting that locus toxicity values, as determined at 32 years age at onset, are broadly representative of the relationship between the variables throughout the dataset as a whole.

3.2.2 CTG•CAG expandability and locus toxicity correlate with flanking GC content

We previously described a significant positive correlation between repeat expandability and the GC content of flanking sequences; and postulated that flanking GC content may directly or indirectly modify repeat stability (Brock *et al.*, 1999). If repeat stability is indeed a major modifier of locus toxicity, and flanking GC content governs repeat stability, a strong association between locus toxicity and flanking GC content would be expected.

Here, employing the latest assembly of the human genome (NCBI 36) we characterise this relationship in finer detail and to a greater a distance from each locus. Employing the seven polyQ loci and two non-coding CAG•CTG loci, *DM1* and *ERDA1* (Mendlewicz *et al.*, 2004), a significant rank correlation between male germline expandability and flanking GC content was found up to a distance of 1,000 bp from the repeat when the combined flanking sequences of the loci were analysed (Table 3.2). Statistically significant correlations were also obtained when the 5' and 3' flanking sequences were analysed independently. The absence of any significant association ($P \leq 0.05$) at distances from 1 kb to 100 kb suggests that the observed correlations proximal to the repeats are not a simple function of the wider chromosomal GC content surrounding each locus (Figure 3.3). Furthermore, as all significant correlations are found proximal to the repeat, it is unlikely that the results are errors incurred by multiple testing, and correction for multiple testing (Benjamini & Hochberg false discovery rate)

Table 3.2: Correlation of flanking GC content with locus expandability of seven CAG-polyQ loci and non-coding repeats *DM1* and *ERDA1*.

Distance from repeat (bp)	5' flanking sequence			3' flanking sequence			Both flanking sequences		
	rho[a]	*P (rho)*[b]	*P (BH)*[c]	*rho*	*P*	*P (BH)*	*rho*	*P*	*P (BH)*
100,000	0.483	0.187	0.210	0.500	0.170	0.219	0.500	0.170	0.191
50,000	0.483	0.187	0.240	0.533	0.139	0.250	0.417	0.265	0.265
10,000	0.600	0.088	0.132	0.450	0.224	0.252	0.517	0.154	0.231
5,000	0.800	0.010*	0.030*	0.300	0.433	0.433	0.500	0.170	0.219
2,500	0.767	0.016*	0.036*	0.517	0.154	0.231	0.667	0.050*	0.090
1,000	0.867	0.002*	0.018*	0.600	0.088	0.198	0.850	0.004*	0.036*
750	0.850	0.004*	0.018*	0.783	0.013*	0.039*	0.833	0.005*	0.023*
500	0.700	0.036*	0.065	0.854	0.003*	0.014*	0.733	0.025*	0.056
100	0.403	0.282*	0.282	0.862	0.003*	0.027*	0.783	0.013*	0.039*

[a] Spearman's rank correlation coefficient
[b] P-value of Spearman's rank correlation coefficient
[c] P-value corrected for multiple testing
* Significant at the 0.05 level

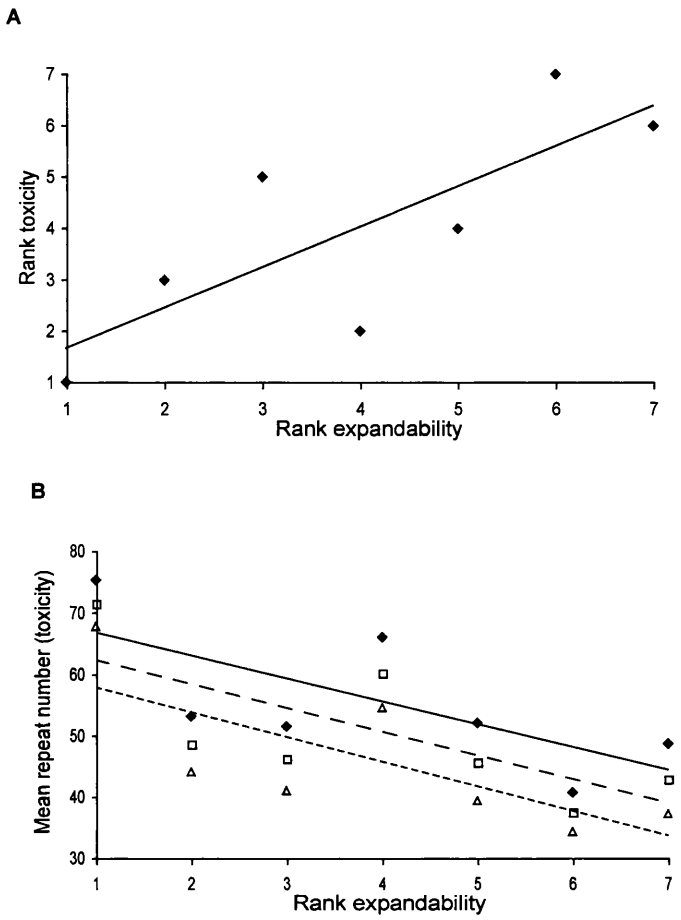


Figure 3.2. Locus expandability correlates with locus toxicity. (A) Plot of locus rank toxicity at 32 years age of onset and locus rank expandability (B) Rank correlation (Spearman's *rho*) of locus expandability and locus toxicity at an age of onset of 30 years (solid line, solid diamonds)(*rho* = 0.78; *P* = 0.036; N = 7), 40 years (dashed line, open squares)(*rho* = 0.86; *P* = 0.014; N = 7), and 50 years (dotted line, open triangles) (*rho* = 0.86; *P* = 0.014; N = 7).

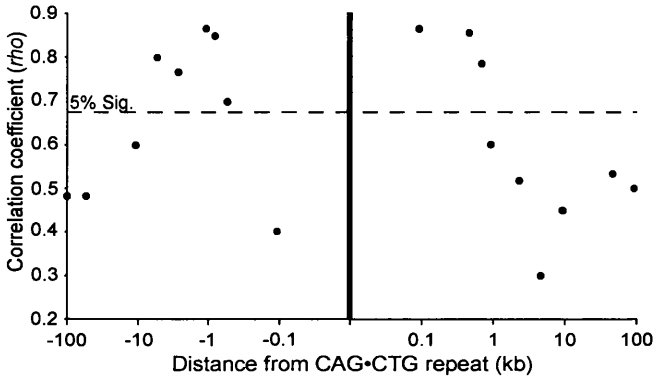


Figure 3.3. Locus expandability correlates with proximal flanking sequence GC content but not with distal flanking sequence GC content. Dashed line shows threshold for statistical significance ($P < 0.05$). Distance from CAG•CTG is plotted in log scale for clarity. The seven CAG-polyQ loci and DM1 and ERDA1 were included in this analysis.

(Benjamini and Hochberg, 1995) resulted in a broadly similar profile of significant correlations (Table 3.2).

In order to further describe the area of significant association flanking the loci we determined a continuous GC content profile of DNA flanking nine CAG•CTG loci to a distance of 5 kb from the repeat using a sliding window of 100 bp and step size of 10 bp. Subsequently, the rank correlation of GC content with the expandability of all loci (Brock *et al.*, 1999) was determined along the flanking sequences at each 10 bp interval. Interestingly, a substantial difference in the correlation profile of the 5' and 3' sequences immediately adjacent to the loci is evident. The 5' sequence shows an almost continuous significant correlation (n=7; $P < 0.05$) from a distance of 140 bp to 850 bp from the loci, whereas a more punctuated profile was found 3' of the loci (Figure 3.4A). A similar profile was obtained upon analysis of polyQ-encoding CAG repeat loci separately (Figure 3.4B).

Applying the same methodology, we analysed the association of flanking GC content with locus toxicity. As we possess reliable quantitative data for both GC content and locus toxicity a product-moment correlation (Pearson, r) was performed. A significant ($P < 0.05$) correlation between locus toxicity and flanking DNA CG content was observed from 100 bp (Figure 3.5) to approximately 400 bp flanking the repeat tract (Figure 3.6A). A similar highly significant association with flanking GC content was observed both 5' and 3' of the CAG repeat loci (Figure 3.6A).

The observed CAG repeat DNA flanking sequence GC content may reflect effects mediated at the level of the mRNA. Employing manually curated RefSeq mRNA sequences for each gene, we investigated the association between locus toxicity and mRNA sequence GC content. Firstly, no significant correlation between locus toxicity and total mRNA GC content was found (r = -0.28; P = 0.58; N = 7). Employing the sliding window approach as before, a significant correlation between flanking mRNA sequence and GC content was only found immediately proximal (< 100bp) to the repeat tract (Figure 3.6C). This region of significant correlation corresponds closely to the region of sequence defined by the 5' and 3' boundaries of the repeat-containing exons in each gene (Figure 3.6B); further

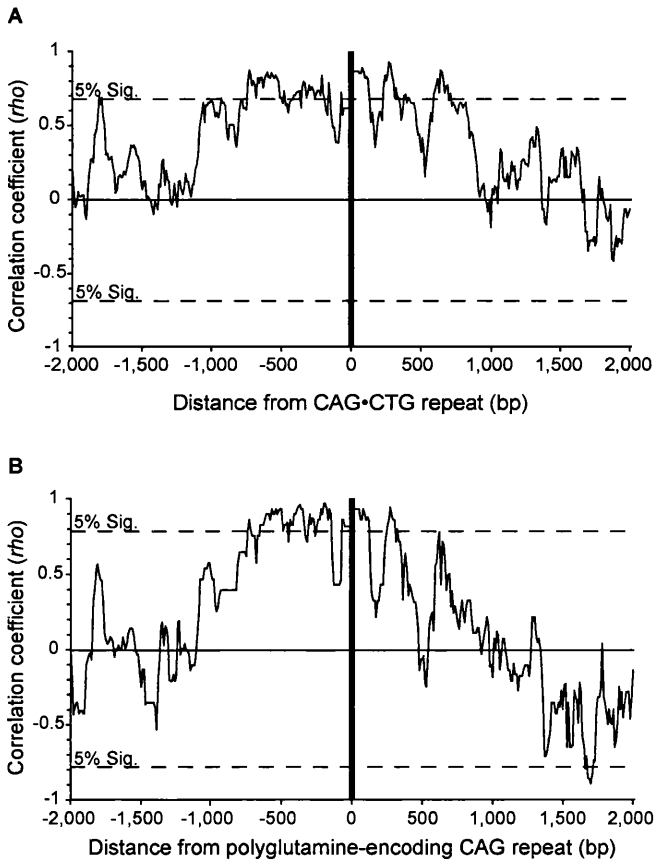


Figure 3.4. CAG•CTG locus expandability is correlated with proximal flanking GC content. (A) Plot shows correlation of flanking GC content of nine CAG•CTG disorders (including *DM1* and *ERDA1*) with locus expandability. (B) Plot shows correlation of flanking GC content of the seven CAG polyQ-encoding disorders only with locus expandability. Using a sliding window of 100 bp and step size of 10 bp, the Spearman's rank correlation (*rho*) was calculated to a distance of 2,000 bp both 5' and 3' of each repeat. The 5% statistical significance threshold (dotted lines) and the position of the repeat loci (vertical bar) are also shown.

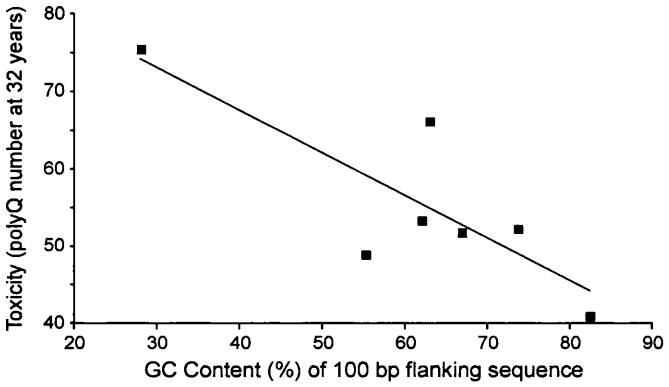


Figure 3.5. Inter-locus polyQ toxicity correlates with DNA flanking sequence GC content. Locus toxicity correlates with GC content of flanking DNA sequences at distances of 100bp (r = -0.82; P = 0.024; N =7).

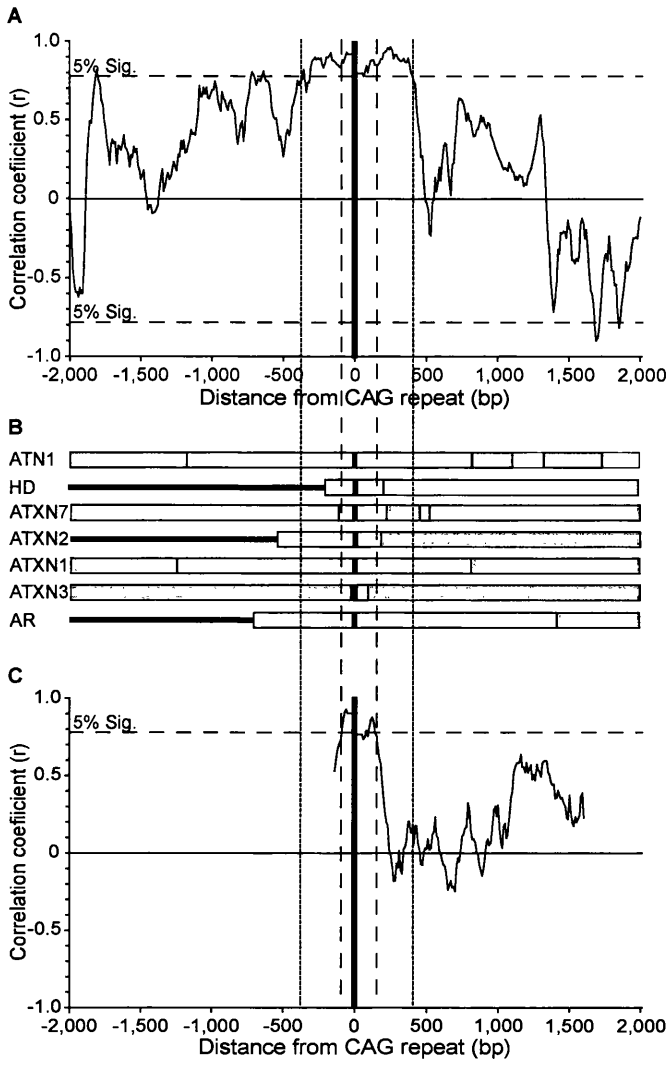


Figure 3.6. Inter-locus toxicity correlates with flanking DNA sequence GC content, but does not extend beyond the repeat containing exon in the mRNA sequence. (**A**) PolyQ loci with higher proximal flanking sequence GC content are more toxic than those with low flanking GC content. GC content was sampled using a sliding window of 100bp and a step size of 10. (**B**) Gene structure of the seven polyQ genes. All drawings are to scale. Exons (white box), introns (grey box), intergenic regions (horizontal black bar), and repeat tract (vertical black bar) are shown. (**C**) Polyglutamine locus toxicity only correlates with flanking mRNA sequence GC content to the 5' and 3' boundaries of their host exons. Vertical dashed lines define the boundary of the region of significant correlation between DNA GC content and locus toxicity.

suggesting that the correlation between flanking DNA GC content and locus toxicity does not reflect effects mediated at the level of the mRNA. In addition, locus toxicity did not correlate significantly with the distance (bp) of the repeat tract from either the transcription start site (Rank correlation; $N = 7$, $rho = 0.43$, $P = 0.3$) or translation start site (Rank correlation; $N = 7$, $rho = 0.5$, $P = 0.22$), suggesting that genic location is not a modifier of locus toxicity.

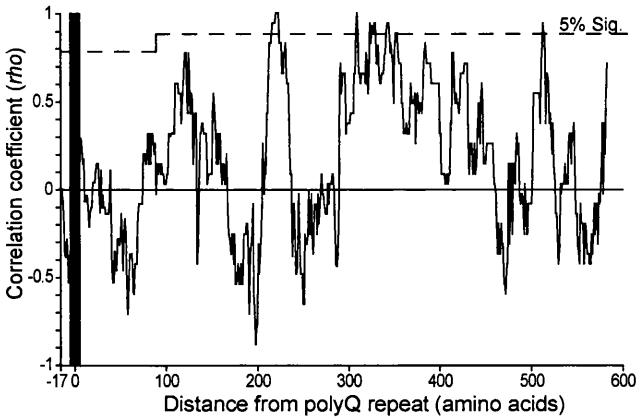
3.2.3 Locus toxicity does not correlate with flanking protein sequence properties

It is assumed that inter locus polyQ toxicity differences are mediated solely by protein context. As the structures of the proteins under investigation have not been resolved, we assessed the contribution of protein context to polyQ toxicity by carrying out *in silico* analyses of the primary sequence characteristics of each polyQ-containing protein. These analyses attempted to correlate polyQ toxicity with the physiochemical and predicted structural properties of the amino acid sequences flanking the polyQ tract. Protein properties were quantified using published, experimentally and empirically derived scales of protein physiochemical characteristics (Table 3.3). The close proximity of the polyQ tract to the N-terminus of the HD, AR and SCA7 proteins permitted only limited comparisons of amino acid sequence properties 5' of the polyQ tract. Employing these scales of predicted amino acid composition, polarity, flexibility and hydrophobicity, no correlation with locus toxicity was identified (Figure 3.7). Similarly, no correlation between the predicted secondary structure flanking the polyQ tract and locus toxicity was found (Figure 3.8). Interestingly, several secondary structure prediction algorithms failed to identify any regions of conserved structure in the sequences flanking the polyQ repeat in each protein (Figure 3.9) (Frishman and Argos, 1995; Heinig and Frishman, 2004; Kabsch and Sander, 1983; Karchin et al., 2003).

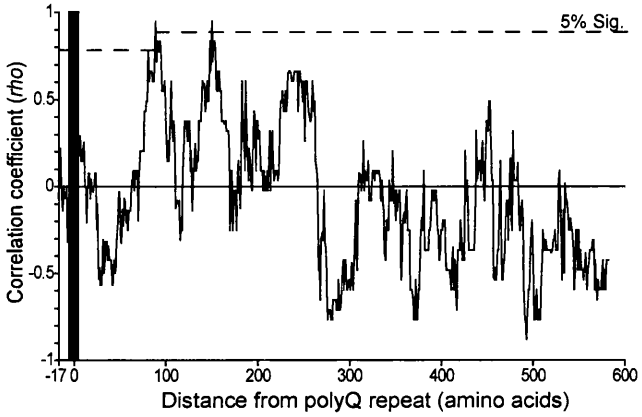
Table 3.3. Amino acid scales compared with inter-locus toxicity

Amino Acid Scale	Source
Amino Acid Composition	http://expasy.org/txt/old-rel/relnotes.51.htm#statistics
Hydrophobicity	(Eisenberg *et al.*, 1984)
Hydrophobicity	(Kyte and Doolittle, 1982)
Polarity	(Grantham, 1974)
Polarity	(Zimmerman *et al.*, 1968)
Alpha-helix	(Chou and Fasman, 1978)
Beta-turn	(Chou and Fasman, 1978)
Beta-sheet	(Chou and Fasman, 1978)
Average flexibility	(Bhaskaran and Ponnuswamy, 1984)
Coil	(Deleage and Roux, 1987)

A (Amino Acid Composition)



B (Flexibility)



C (Hydrophobicity - Eisenberg)

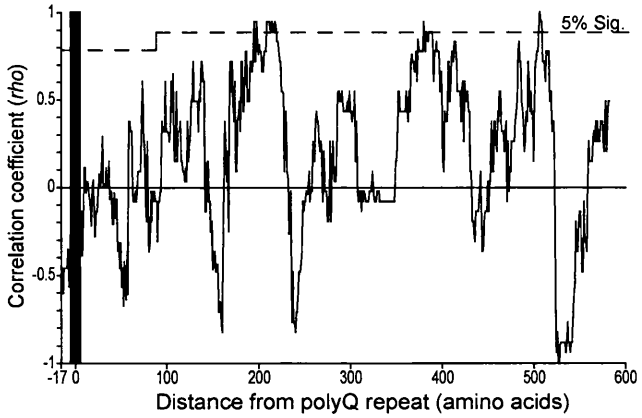
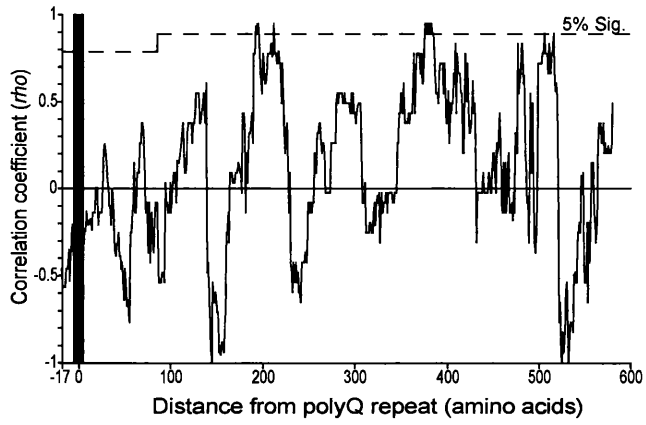
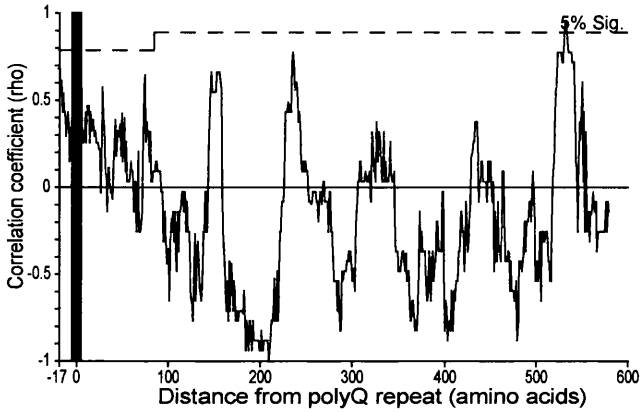


Figure 3.7. Correlation of polyglutamine tract flanking primary sequence properties with locus toxicity. Using a window size of 21 amino acids and a step size of one, locus toxicity was correlated (Spearman's rank) with various physiochemical and compositional characteristics of the primary protein sequence at every amino acid position flanking the polyglutamine repeat. Repeat size was normalised to 21 glutamines. Dotted line represents the 5% statistical significance threshold. As the 3' sequence of ATXN3 extends just 83 amino acids from the repeat, all correlations beyond this point involve the remaining six sequences with a correspondingly higher 5% significance threshold.

D (Hydrophobicity - Kyte & Doolittle)



E (Polarity - Grantham)



F (Polarity - Zimmerman)

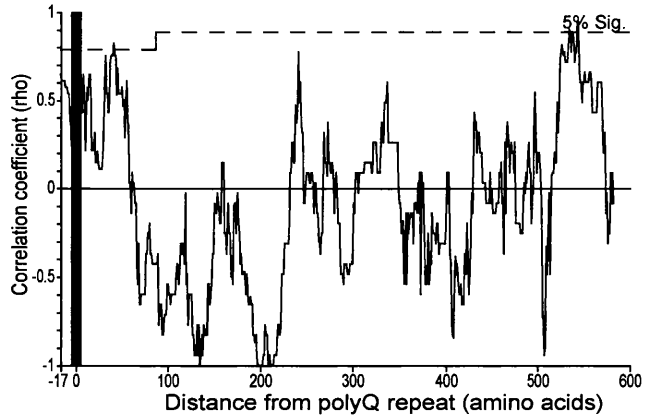
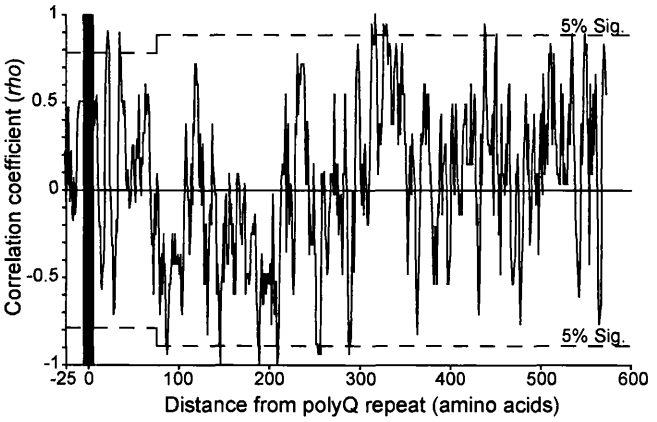


Figure 3.7 (cont). Correlation of polyglutamine tract flanking primary sequence properties with locus toxicity.

A. Alpha Helix



B. Beta Sheet

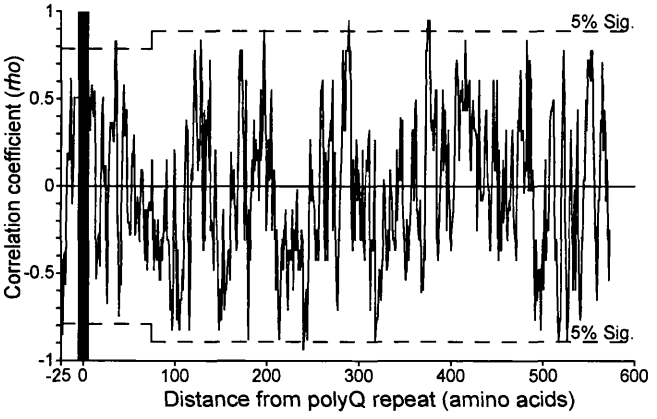
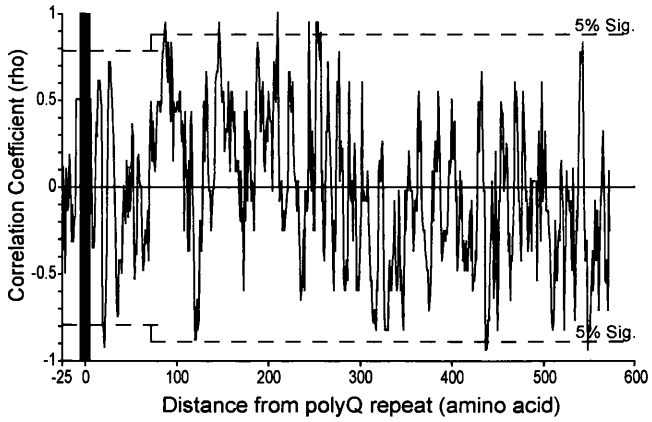


Figure 3.8. Correlation of predicted flanking secondary structure with locus toxicity. Using a window size of 4 amino acids and a step size of one, locus toxicity was correlated (Spearman's rank) with the predicted secondary structure, as determined from scales of secondary structure formation potential of the primary protein sequence at every amino acid position flanking the polyglutamine repeat. Repeat size was normalised to 21 glutamines. Dotted line represents the 5% significance threshold. As the 3' sequence of ATXN3 extends just 83 amino acids away from the repeat, all correlations beyond this point involve the remaining six sequences with a correspondingly higher 5% significance threshold.

C. Beta Turn



D. Coil

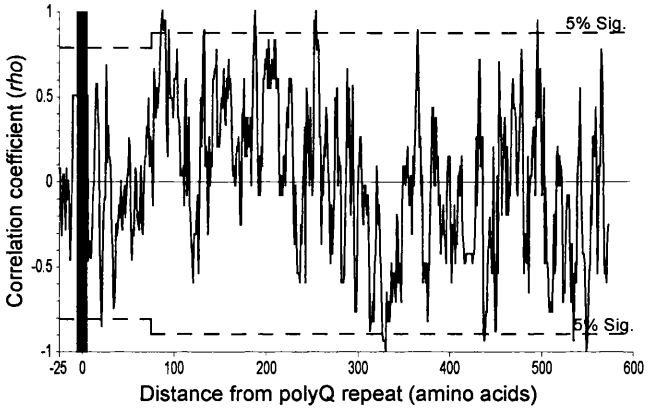


Figure 3.8 (cont.). Correlation of predicted flanking secondary structure with locus toxicity.

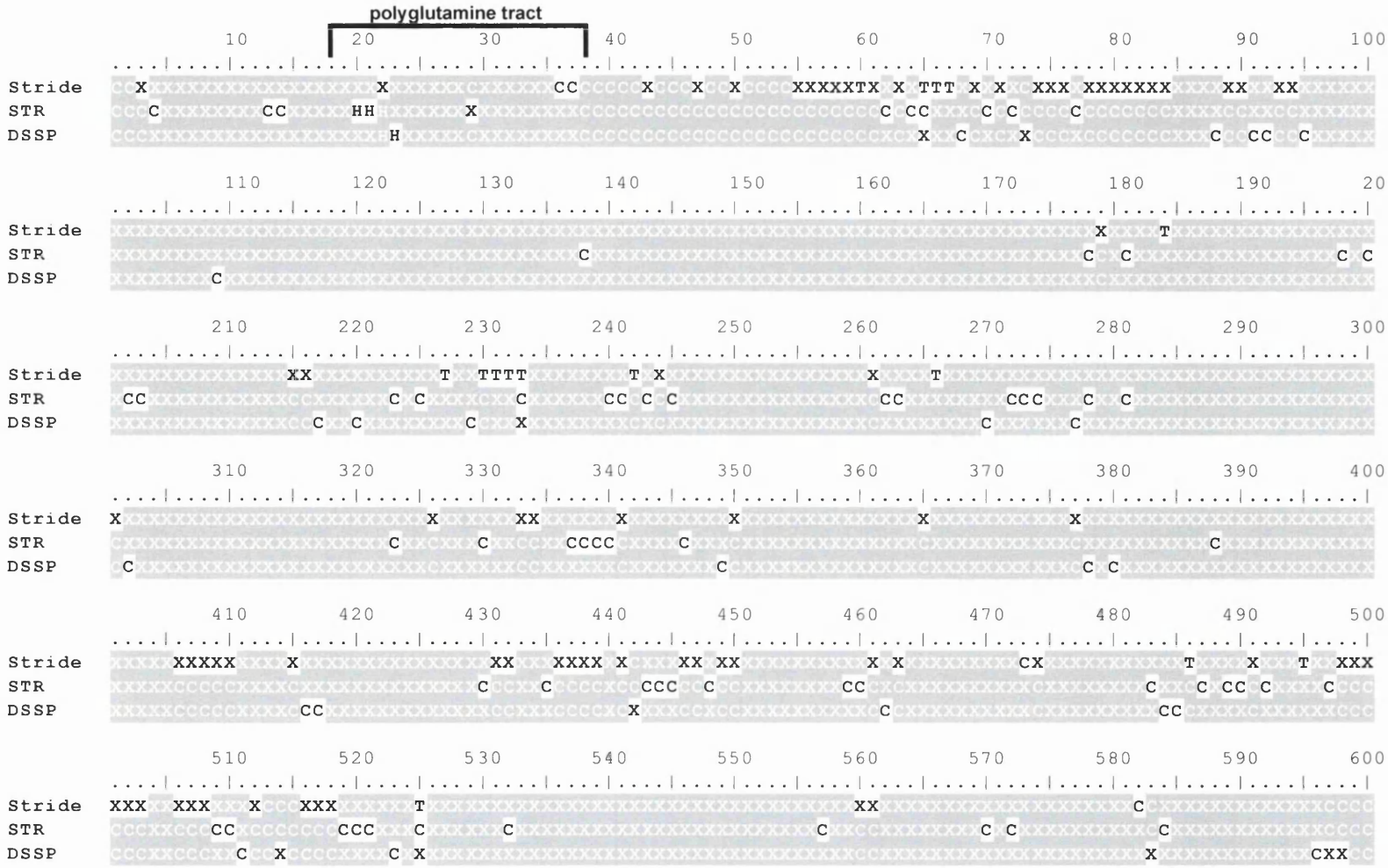


Figure 3.9. Predicted secondary structure flanking polyglutamine repeats. Secondary structure predictions were assigned by the algorithms STRIDE, STR, and DSSP. For each prediction technique, a consensus sequence for all 7 loci was derived, where identity at a given amino acid position in 5 or more sequences was required for inclusion in the consensus sequence. X = no consensus, C = coil (no predicted structure, other), H = helix, T = Hydrogen-bonded turn

3.3 Discussion

The dynamic mutation disorders such as Huntington disease and myotonic dystrophy type 1 (DM1) are caused by expanded CAG•CTG repeats in their associated genes. Repeat toxicity increases with length, longer repeats resulting in greater levels of cell death and dysfunction in affected tissues, and a more severe phenotype in each disorder. Moreover, the observation that the most affected tissues in expanded CAG•CTG repeat disorders tend to possess the longest expansions, suggests that somatic mosaicism contributes to the tissue-specificity and progressive nature of these disorders (Kennedy *et al.*, 2003; Shelbourne *et al.*, 2007).

The polyglutamine disorders, defined by a polyglutamine (polyQ)-encoding CAG repeat, are the largest class of expanded CAG•CTG repeat disorders. Although the inverse relationship between age at onset and repeat length is broadly similar in the polyQ disorders, the number of repeats required to effect a given age at onset varies markedly between disorders (Figure 3.1). It is widely assumed that this 'locus toxicity' difference is due to protein context mediated effects on polyQ tract cytotoxicity. In fact, several studies have shown that protein context can be a major modifier of polyQ tract behavior *in vitro* (de Chiara et al., 2005; Duennwald et al., 2006; La Spada and Taylor, 2003). However, no rationalisation of how protein context determines the observed order of inter-locus toxicity has yet been described. We propose that the rate at which CAG•CTG repeats expand is a major modifier of age at onset and progression of symptoms in these disorders, and can largely explain the inter-locus toxicity differences observed between the CAG•CTG disorders. Employing age at onset data for seven polyQ disorders we quantified these inter-locus polyQ toxicity differences and found that the order of polyQ toxicity correlated significantly with the underlying stability of the expanded CAG repeat tract.

We previously detailed a significant association between flanking GC content and locus expandability (Brock *et al.*, 1999). Here, analyzing this relationship in finer detail and at greater distances from each locus, we found a significant positive correlation between proximal flanking GC content and repeat instability for both

the polyQ-encoding CAG repeats and a larger sample including the non-coding repeat loci *DM1* and *ERDA1*. Not surprisingly, flanking GC content was also found to correlate with polyQ toxicity. Significantly, we then showed that toxicity is not correlated with the flanking mRNA GC content or the properties of amino acid sequence flanking the polyQ tract. In addition, we found that toxicity was not significantly correlated with the position of the repeat tract within the gene or mature protein. Thus, we propose that the significant correlation between flanking DNA GC content and inter-locus polyQ toxicity is a consequence of flanking GC content effects on DNA repeat stability.

Germline instability data from affected families are abundant, whereas sound quantitative somatic instability data are very sparse. Therefore, we used previously published estimates of intergenerational locus instability as approximations of the relative level of somatic instability between disorders (Brock *et al.*, 1999). Examining data from a published study of somatic mosaicism in post-mortem brain tissue of SCA1 and MJD affected individuals (Maciel *et al.*, 1997), we found that the repeat-length normalised levels of somatic mosaicism in SCA1 were approximately double the levels found in MJD in both cerebral cortex (N_{MJD} = 11; N_{SCA1} = 7; Mann-Whitney U = 0; $P < 0.0001$) and cerebral white matter (N_{MJD} = 9; N_{SCA1} = 6; Mann-Whitney U = 0; $P < 0.001$) samples (Tables 3.4a and 3.4b); similar to the relative levels of germline instability observed in these disorders (Table 3.1). Furthermore, in mouse models of various trinucleotide disorders, those lines showing greatest intergenerational instability tend also to exhibit higher levels of somatic instability (Fortune *et al.*, 2000; Mangiarini *et al.*, 1997; Seznec *et al.*, 2000).

Several clinical observations suggest that somatic expansion of repeats contributes towards age at onset of symptoms and disease progression. Firstly, individuals with expanded yet stable *SCA1* loci exhibit significantly delayed onset of symptoms (Matsuyama *et al.*, 1999; Quan *et al.*, 1995) or remain asymptomatic (Frontali *et al.*, 1999). These individuals contain histidine-encoding CAT interruptions in the expanded CAG repeat. CAG•CTG repeats containing interruptions tend not to expand, whereas loss of repeat interruptions is associated with repeat expansion (Choudhry *et al.*, 2001; Chung *et al.*, 1993). However, the aggregation dynamics of these histidine-containing tracts are

Table 3.4a. Age at death and repeat length of 11 MJD patients[1]

Patient	Age at Death	Normal Allele	Expanded Allele[A]	Mean no of bands in Cortex[B]	Mean no of Bands in White Matter[C]	Cortex[D] (length adjusted)	White Matter[E] (length adjusted)
90-274	63	12	69	10	9	0.14	0.13
94-547	63	20	74	10	12	0.14	0.16
M-448	62	12	71	10	10	0.14	0.14
93-423	58	20	75	10	11	0.13	0.15
94-453	56	24	73	12	10	0.16	0.14
M-268	46	18	75	12	11	0.16	0.15
M-318	44	20	77	10	10	0.13	0.13
M-420	43	20	78	10	10	0.13	0.13
88-196	27	20	80	9	10	0.11	0.13
2303	56	12	75	10	...	0.13	
1965	52	24	74	9	...	0.12	

1. Table from Maciel et al. (1997)
B. Average of 3 experiments
C. Average of 3 experiments
D. Length adjusted somatic mosaicism (B/A)
E. Length adjusted somatic mosaicism (C/A)

Table 3.4b. Age at death and repeat length of 7 SCA1 patients[1]

Patient	Age at Death	Normal Allele	Expanded Allele[A]	Mean no of bands in Cortex[B]	Mean no of Bands in White Matter[C]	Cortex[D] (length adjusted)	White Matter[E] (length adjusted)
94-538	77	31	44	11	...	0.25	
M-652	68	30	45	11	13	0.24	0.29
M-378	65	29	45	11	12	0.24	0.27
91-288	57	30	54	15	15	0.28	0.28
93-441	53	28	54	16	15	0.30	0.28
90-276	45	30	56	11	14	0.20	0.25
92-271	38	31	56	13	14	0.23	0.25

1. Table from Maciel et al. (1997)
B. Average of 3 experiments
C. Average of 3 experiments
D. Length adjusted somatic mosaicism (B/A)
E. Length adjusted somatic mosaicism (C/A)

Table 3.4c. Mann-Whitney U-Test of length adjusted mosaicism in MJD and SCA1

	MJD (sample size)	SCA1 (sample size)	U	P-value
Cortex	11	7	0	6.30E-05
White Matter	9	6	0	0.0003996

similar to those of uninterrupted polyQ tracts (Calabresi *et al.*, 2001), suggesting that delayed onset is due to increased locus stability, not altered protein toxicity. In addition, a large (N = 48) group of HD patients from Crete with expanded but stable *HD* loci had a median age at onset 15-20 years later than expected (Tzagournissakis *et al.*, 1995). Significantly, the CAG repeat tract in these patients is uninterrupted, coding for a pure polyglutamine tract, further implicating somatic expansion, not polyglutamine toxicity, as the major modifier of disease progression (Kartsaki *et al.*, 2006). The finding that the repeat tract does not contain interruptions suggests that elements, other than repeat type, length and purity modify repeat stability.

Our model of instability-mediated disease pathogenesis is further supported by a recent computational study which predicted that repeat expansion rate in somatic tissue determines both age at onset and the rate of disease progression (Kaplan *et al.*, 2007). Employing mathematical modeling and computer simulations, it was shown that the more rapid disease progression observed in juvenile cases and the similar age at onset but more rapid disease progression observed in individuals homozygous for HD expansions could be accurately represented by a somatic-expansion rate model, but not by a cumulative polyglutamine toxicity model (Kaplan *et al.*, 2007).

Despite a plethora of data implicating *cis*-elements as potential modifiers of expanded CAG•CTG repeat stability, none have yet been identified *in vivo* other than flanking GC content. Flanking GC content may affect repeat stability by modifying the ability of the MMR machinery to process small mismatched loop-outs within the repeat tract. Modification of normal MMR by flanking GC content may be achieved directly through its effect on DNA melting potential. Indirectly, GC content may affect MMR by altering the ability of flanking sequences to form secondary structures, providing CpG sites for methylation leading to transcriptional changes or by alteration of chromatin state surrounding the repeat tract.

Our model provides a simple instability-mediated rationalization of the inter-locus toxicity differences observed between polyQ disorders and re-emphasizes

the importance of somatic mosaicism as both a powerful marker of disease progression and a possible site of therapeutic intervention.

3.4 Materials and methods

All genomic DNA analyses used the NCBI 36 (November 2005) assembly of the human genome, obtained from the Ensembl web server (url: http://www.ensembl.org/index.html). The accession numbers of the mRNA sequences employed for each disorder were; NM_001007026 (ATN1), NM_000332 (ATXN1), NM_002973 (ATXN2), NM_000333 (ATXN7), NM_004993 (ATXN3), NM_000044 (AR), and NM_002111 (HD). The accession numbers of the protein sequences employed were; NP_001007027 (ATN1), NP_000323.2 (ATXN1), NP_002964.2 (ATXN2), NP_000324 (ATXN7), NP_004984 (ATXN3), NP_000035 (AR), and NP_002102 (HD). Age at onset data for each locus, which was collated from published studies, was kindly supplied by Jim Gusella and Marcie MacDonald (Gusella and MacDonald, 2000). The data consists of age at onset and repeat number measurements for over 2,400 affected individuals, with over 100 data points for each disorder. Protein scales were obtained from the ExPASy proteomics server (url: http://www.expasy.ch/).

All GC content analyses were performed with custom written software implemented in the Perl programming language. STRIDE, DSSP, and STR secondary structure predictions were performed via the SAM server (url: http://www.soe.ucsc.edu/research/compbio/sam.html). SPSS (version 13) was used for statistical analyses.

4. Investigation of *cis*-acting modifiers of DM1 locus expandability in cell culture models

4.1 Introduction

Disease-associated expanded trinucleotide repeats are highly unstable in both germline and somatic tissue (Gomes-Pereira and Monckton, 2006). The mechanism underlying this expansion-biased instability is not precisely known. However, studies of murine models deficient for various components of the mismatch repair system have implicated a role for cell division independent mismatch repair in expanded repeat instability (Gomes-Pereira *et al.*, 2004; Manley *et al.*, 1999; Pearson *et al.*, 1997). Characterising repeat dynamics in individuals carrying expanded alleles and in animal and cell culture models of expanded repeat instability is crucial to determining the factors that mediate repeat expansion over time. Mounting evidence suggests somatic mosaicism is likely to be a major modifier of age at onset of symptoms and disease progression (Chapters 1 & 3), further emphasising the importance of understanding the process of expanded repeat instability.

Several observations suggest that *cis*-acting factors modify expanded repeat instability (Brock *et al.*, 1999; Fortune *et al.*, 2000; Frontali *et al.*, 1999; Libby *et al.*, 2003). The study of affected individuals has found that the sequence of the repeating triplet, the overall length of the repeat tract and its purity are major modifiers of repeat instability. Furthermore, when normalised for repeat length, expanded repeats of the same sequence at different loci exhibit markedly different levels of instability, suggesting that genomic location affects repeat stability (Brock *et al.*, 1999). In addition it was shown that locus expandability was significantly correlated with the GC content of the sequences directly flanking the expanded repeat loci (Brock *et al.*, 1999). Several murine models of expanded repeat instability have reported significantly differing levels of repeat instability between mouse lines carrying identical transgenes, further implicating genomic position as a modifier of repeat stability (Fortune *et al.*, 2000). A mouse model of SCA7 found that increasing the amount of endogenous human genomic sequence flanking an expanded repeat appeared to dramatically increase its instability upon integration into the mouse genome, suggesting the

presence of modifiers of repeat stability within the flanking sequence (Libby *et al.*, 2003).

The finding that site of transgene integration dramatically affects repeat stability has frustrated attempts to study *cis*-acting modifiers of repeat stability in animal models as separation of the effects of transgene-specific *cis*-elements from the effects of insertion site *cis*-elements is complicated. The generation of sufficient numbers of transgenic mouse lines containing identical, randomly integrated, and unstable expanded repeat tracts to allow for identification of true transgene-specific *cis*-elements is unfeasible. However, the generation of large numbers of mammalian cell lines carrying stably integrated transgenic repeats is feasible and thus, may offer a means to study and identify *cis*-acting modifiers of repeat stability.

However, once integrated into a host genome, expanded repeats may undergo repeat-stability modifying epigenetic alterations of sequences proximal to and within the transgenic repeat. Transcriptional silencing of transgenes integrated into mammalian genomes over time has been widely reported (Elgin and Grewal, 2003; Robertson et al., 1995). This phenomenon, called chromosomal position effect (CPE), results from the propagation of transcriptionally repressive condensed chromatin along the region of the host chromosome containing the integrated transgene (West *et al.*, 2002). Chromatin condensation of a transgenic repeat may affect its stability directly by altering its ability to form instability-mediating secondary structures, or indirectly by silencing its transcription. The relationship between expanded repeat stability and both its chromatin state and expression levels is unclear. However, an association between occurrence of transgene expression, and a presumptively open chromatin state, and presence of repeat instability has been indicated in several studies (Chapter 1).

Although chromatin condensation is self-propagating, vertebrate chromosomes contain distinct domains of condensed and open chromatin, suggesting the existence of elements that halt the spreading of chromatin condensation. Only one such vertebrate 'insulator' element has been well characterised, the chicken β-globin HS4 insulator (Chung *et al.*, 1997). Flanking a transgene with

two copies of this insulator element has been found to protect against CPE silencing in many model systems (Potts *et al.*, 2000; Recillas-Targa *et al.*, 2002; West *et al.*, 2002). How insulator elements function as barriers against the propagation of condensed chromatin is not fully understood. Although the chicken β-globin HS4 insulator element contains a binding site for CTCF (CCCTC-binding factor), a zinc-finger protein involved in regulation of gene expression domains by blocking non-specific enhancer-promoter interactions, CTCF does not seem to play a role in chromatin barrier activity (Recillas-Targa *et al.*, 2002). The most convincing model of insulator barrier activity proposes that proteins recruited to the insulator element by upstream transcription factor (USF) proteins, mediate the acetylation and methylation of specific histone H3 and H4 sites of proximal nucleosomes, preventing the propagation of histone modifications associated with chromatin condensation (West and Fraser, 2005; West et al., 2004).

Here, human and mouse cell lines containing stably integrated expanded repeat sequences are constructed, and their utility as models of expanded repeat instability is investigated. In addition, the ability of mammalian insulator elements to protect integrated repeats from chromosomal position effects on repeat stability is also studied.

4.2 Results

4.2.1 The effect of flanking insulator elements on expanded repeat instability

4.2.1.1 A HeLa cell model of expanded repeat stability

Studies of murine and cell culture models of expanded repeat stability have found that unstable transgenic repeats tend to be transcribed, though not all transcribed repeats are unstable. This association may reflect a requirement for expanded repeats to be located in open chromatin fibers, usually associated with areas of active transcription, in order to exhibit instability. Thus, we propose using flanking insulator elements to protect transgenic repeats from chromatin condensation and the possible consequential stabilising of the repeat. However, as the precise nature of the relationship between chromatin state and repeat stability is unclear, and much about the functioning of insulator elements is unknown, though counter-intuitive it is possible that insulator elements may have undesired effects on the stability of expanded repeats, such as preventing instability, rendering the use of insulators ineffective.

In order to assess the effect of proximal insulator elements on repeat dynamics, two constructs, one consisting of an expanded repeat flanked by insulator elements (Insulator +ve, INDI), another containing the same repeat lacking flanking insulator elements (Insulator -ve, ND), were constructed (Figure 4.1B, C). The repeat-containing portion of each construct consists of the *Dmt* transgene, cloned from *Dmt*-D mouse DNA. *Dmt*-D mice exhibit high levels of length-dependent expansion-biased instability, as do cell lines derived from tissues of *Dmt*-D transgenic mice (Chapter 1). The *Dmt* transgene consists of a CTG•CAG repeat from the human *DM1* locus flanked by 113 bp and 593 bp of endogenous human flanking sequence 5' and 3' of the repeat, respectively (Figure 4.1A). Sequencing of the constructs generated here found five interruptions in the 3'-end of the repeat tract, resulting in a repeat with the following configuration:

$(CTG)_{112}(CGG)(CTG)_6(CGG)(CTG)_3(CGG)(CTG)_6(CGG)(CTG)_6(CGG)(CTG)_8$.

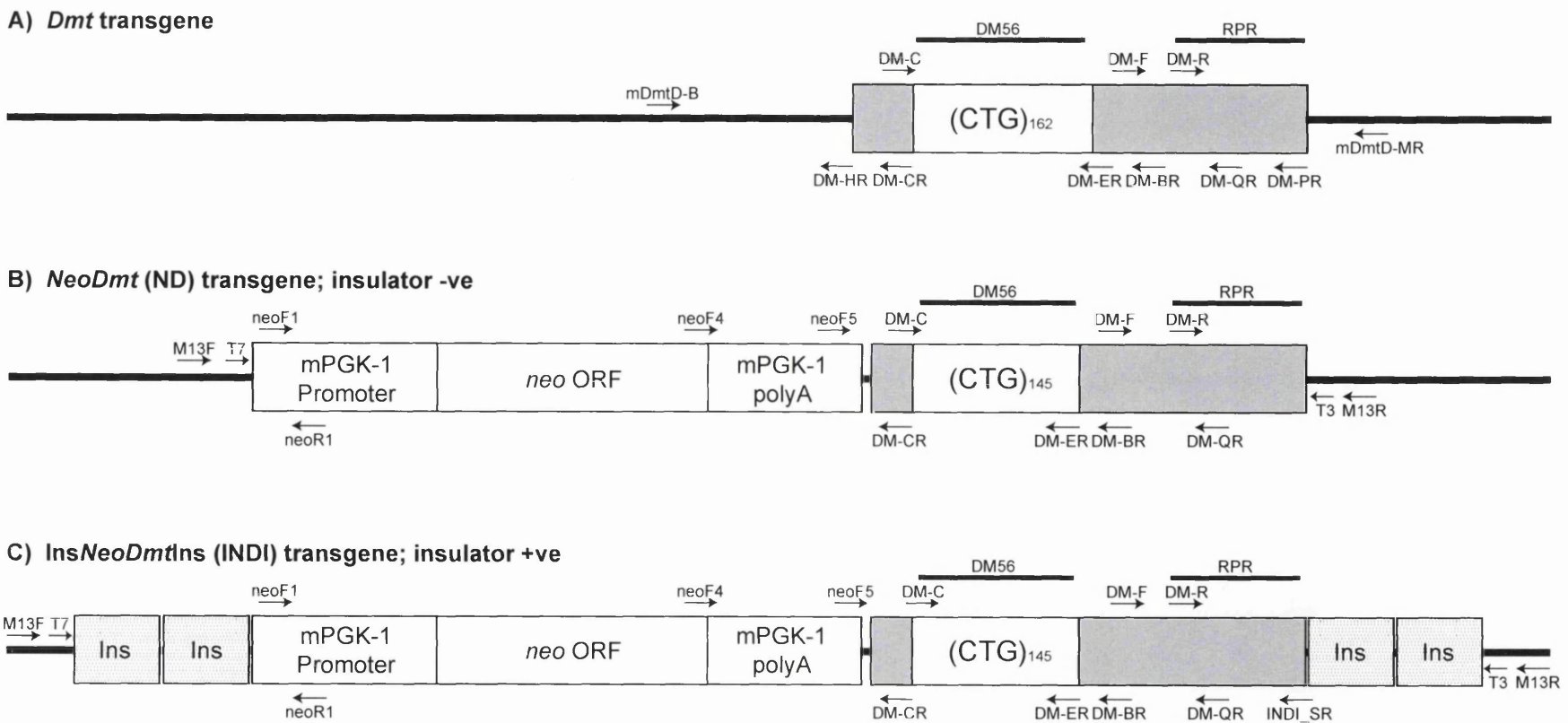


Figure 4.1. Schematic representation of DM1 transgenes. Primer binding sites are shown as arrows. Horizontal black bars indicate the region of binding of the probes DM56 and RPR. mPGK-1 = mouse phosphoglycerate kinase-1, Ins = insulator element (250 bp core of the chicken ß-globin locus hypersensitive site 4). Two copies of the chicken ß-globin insulator were isolated from pNI-CD (A. West, University of Glasgow) by digestion with *Kpn*I and cloned into the *Kpn*I site of pBluescript SK+ (Stratagene) to give pB-2CD. The neomycin expression cassette was isolated from pPNT by digestion with *Xho*I and *Bam*HI and subsequently cloned between the *Xho*I and *Bam*HI sites of pB-2CD, to yield pB-2CD-*Neo*. The *Dmt* transgene was amplified from tail-tip DNA of a *Dmt*-D mouse using 1:1 *Pfu*:*Taq* DNA polymerase and cloned into the pGEM-T-Easy vector (Promega) to give pT-*Dmt*. The *Dmt* transgene was isolated by digestion of pT-*Dmt* with *Not*I and cloned into the *Not*I site of pB-2CD-*neo*, to give pB-2CD-*Neo-Dmt*. pB-*NeoDmt* (ND), was obtained by removal of the insulator elements from pB-2CD-*Neo-Dmt* by digestion with *Kpn*I and religation of the plasmid. Finally, pB-2CD-*Neo-Dmt*-2CD (INDI) was generated by isolation of two copies the chicken ß-globin insulator from pB-2CD by digestion with *Ale*I, and their subsequent cloing into the *Ale*I site of pB-2CD-*Neo-Dmt*.

To allow for efficient selection of transfected cells, a *neomycin* cassette was cloned immediately 5' of the *Dmt* transgene. As the *neo* cassette was cloned in same orientation as the *Dmt* transgene, read-through of the CTG repeat may occur. Finally, the insulator positive construct was flanked by two copies of the chicken β-globin locus HS4 insulator at the 5' end of the *neomycin* cassette and the 3' end of the *Dmt* fragment.

HeLa cells were chosen as the cell culture system in which to model expanded repeat instability as, unlike many cancer cell lines, they possess a well characterised and functional mismatch repair system. Stably transfected, single HeLa cell clones were generated by cationic lipid transfection with linearised vector DNA followed by positive selection with G418 for 7 - 10 days. Between 7 - 10 days post-transfection, single cell colonies (10 - 50 cells) were identified under the light microscope, picked using the glass-ring technique (Chapter 2), and transferred to individual wells of a 24-well plate. Less than five single-cell clones were picked from any given culture dish, and care was taken not to pick colonies close together, as pairs of colonies often arise from cell clusters spalling-off from a parent colony. Single cell clone lines transfectant for the insulator -ve construct (N = 11) or insulator +ve construct (N = 12) were then maintained in culture for 50 population doublings (~50 days). Mouse kidney cell lines derived from *Dmt*-D mice carrying a mutant allele of ~CTG_{175} repeats, showed dramatic expansion-biased instability (+ 30 - 40 CTG) at the mutant locus over 50 populations in culture (Gomes-Pereira *et al.*, 2001).

To investigate changes in repeat length, the transgenic repeat was amplified (using primers neoF5 and DM-BR) from each clone and visualised by hybridisation with a CTG-containing probe (DM56) (Figure 4.2). Four insulator +ve clones possessed repeats which differed markedly in size from the repeat tract contained in the transfectant DNA. However, several observations suggest that the repeat length changes observed are not a consequence of expansion-biased instability as reported previously for the *Dmt* transgene. Firstly, as two clones have expanded alleles, and two have contracted alleles, no expansion bias is evident. Secondly, the size of the contracted repeats is similar to contractions observed during cloning of the transgene in bacteria (data not shown), indicating that the transfectant DNA may have contained a heterogeneous population of

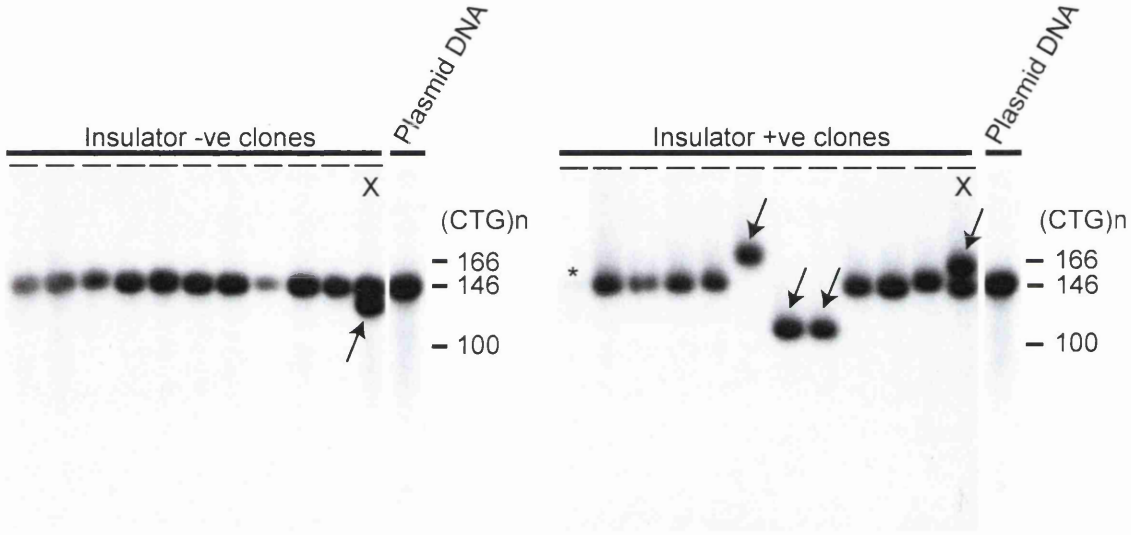


Figure 4.2. Instability of transgenic $(CTG)_{145}$ repeat in HeLa cells over 50 population doublings. Each lane represents DNA from an independent, stably transfected clone. The transgenic repeat tract was amplified with transgene-specific primers neoF5 & DM-BR. The products were blotted and hybridised to a CTG repeat containing probe (DM56). The size in repeats is indicated to the right of each autoradiograph. PCR of insulator negative plasmid DNA shows repeat size at time of transfection. Arrows indicate alleles exhibiting changes in repeat length. Asterisks indicates presence of a faint band.

constructs possessing different repeat lengths. This suggestion is supported by the observation that the two contracted alleles are apparently identical in size, suggesting that the deletion occurred pre-transfection. Alternatively, the two clones containing contracted alleles may not be independent clone lines, but have derived from the same single cell clone line. Analysis of the mutant allele in *Dmt*-D kidney cell lines revealed a dramatic increase in the average allele size and a broadening of the range of repeat sizes over time. These changes result in a both an increase in band size and smearing upon PCR amplification (Gomes-Pereira *et al.*, 2001). Here, the bands corresponding to the mutant expanded alleles are not more diffuse or 'smeary' than the shorter non-mutant alleles, suggesting that a single expansion or rearrangement event has occurred, as opposed to the numerous small expansions typical of instability at an expanded locus. One insulator -ve and one insulator +ve clone (both denoted as 'X') appear to have two alleles, suggesting that these clones contain multiple copies of the transgene. With the exception of clone 'X', no obvious repeat instability was observed in the insulator -ve clones.

As no credible, expansion-biased instability was observed in any clones, it was not possible to draw conclusions regarding the effect of insulators elements on repeat stability from this experiment.

The levels of somatic mosaicism observed in DM1 are highly tissue specific, suggesting a modifying influence of *trans*-acting factors on repeat stability. Furthermore, cultured cells derived from various tissues of mice containing the *Dmt* transgene recapitulate the tissue specificity of expanded *DM1* locus instability. Although HeLa cells contain a competent mismatch repair system, is possible that HeLa cells do not contain other *trans*-acting factors required to affect instability of expanded CAG•CTG repeats. Alternatively, as levels of repeat instability are repeat length dependent, it is possible the repeat tract employed here is too short to exhibit instability. Finally, the apparent stability of the mutant alleles may be due to chromosomal position effects in every clone.

4.2.1.2 *A mouse kidney cell model of expanded repeat stability*

In order to rule out cell-type specific effects it was decided to repeat the experiment using a cell line proficient for expanded CTG•CAG repeat instability. *Dmt*-D mouse kidney cells were previously generated from mice carrying the *Dmt* transgene and have been shown to exhibit dramatic, length-dependent, expansion-biased instability (Fortune et al., 2000; Gomes-Pereira et al., 2001; Gomes-Pereira and Monckton, 2004). Therefore, *Dmt*-D kidney cells possess the necessary *trans*-acting factors required to effect instability at expanded CTG•CAG loci. In addition, the expanded unstable repeat already present in this cell line (henceforth referred to as the 'primary transgene' or 'primary repeat') could act as an internal control for instability.

The *Dmt*-D cell line employed here, D2761Kc2, cloned by limiting dilution of the original D2763K cell line, was supplied by Dr. Mario Gomes-Pereira (Gomes-Pereira and Monckton, 2004). Single cell clone lines transfectant for the insulator -ve transgene (N = 20) or insulator +ve transgene (N = 15) were generated as before, and maintained in culture for 60 population doublings (~ 55 days). For clarity, henceforth these transgenes will be referred to as the 'secondary transgene' or 'secondary repeat'. To investigate changes in repeat length, the transgenic repeats were amplified (using primers DM-C and DM-BR) from genomic DNA prepared 30 and 60 population doublings post-transfection from each clone and visualised by hybridisation with a CTG-containing probe (DM56) (Figures 4.3 & 4.4).

As observed for stably transfected HeLa clones, no obvious instability was observed in the secondary repeats in any cell line. In contrast, the primary repeat showed dramatic expansion-biased instability, with a mean expansion size of 28 repeats. Furthermore, the increased smearing of the primary repeat band over time, indicates a broadening of its repeat size range, typical of a repeat undergoing expansion via numerous small size changes.

Interestingly, five single cell clones carrying an expanded yet stable primary transgene were identified (Figure 4.5). The presence of a population of stable expanded alleles, from which these clones are assumed to have derived, is

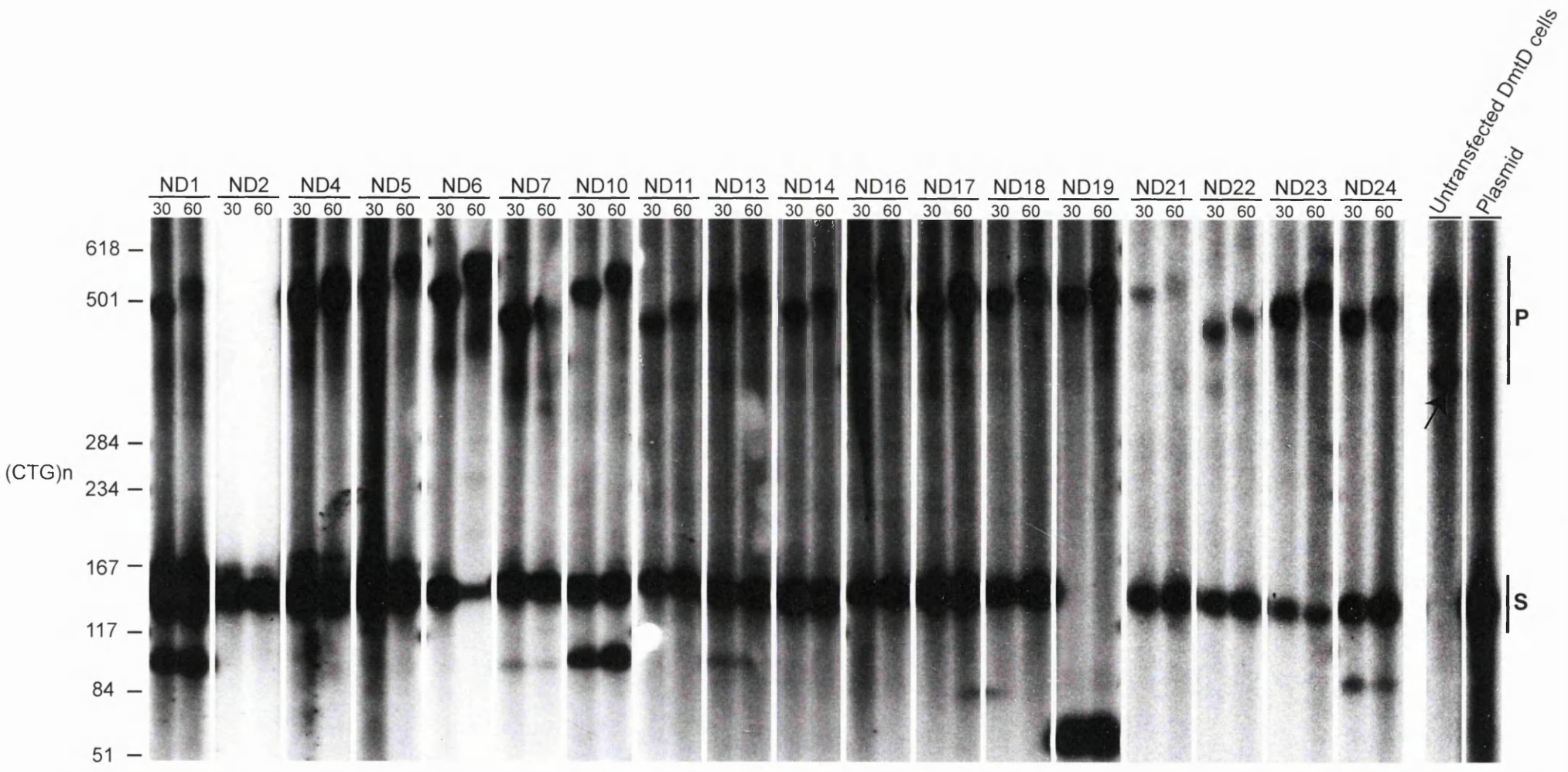


Figure 4.3. Analysis of transgenic repeat lengths in *Dmt*-D kidney cells stably transfected with the insulator negative transgene, ND. DNA was prepared from cells 30 and 60 population doublings post transfection. Both transgenic repeat tracts were amplified in the same reaction with the primers DM-C & DM-BR. The products were blotted and hybridised to a CTG containing probe (DM56). Size standards, in repeats, are indicated to the left of the figure. Amplified transfectant vector DNA and amplified untransfected *Dmt*-D kidney cell DNA are also shown. The arrow indicates a population of stable expanded alleles in *Dmt*-D kidney cells. Vertical bars on the right of the figure indicate the repeat size range of the primary (**P**) and secondary (**S**) transgenes.

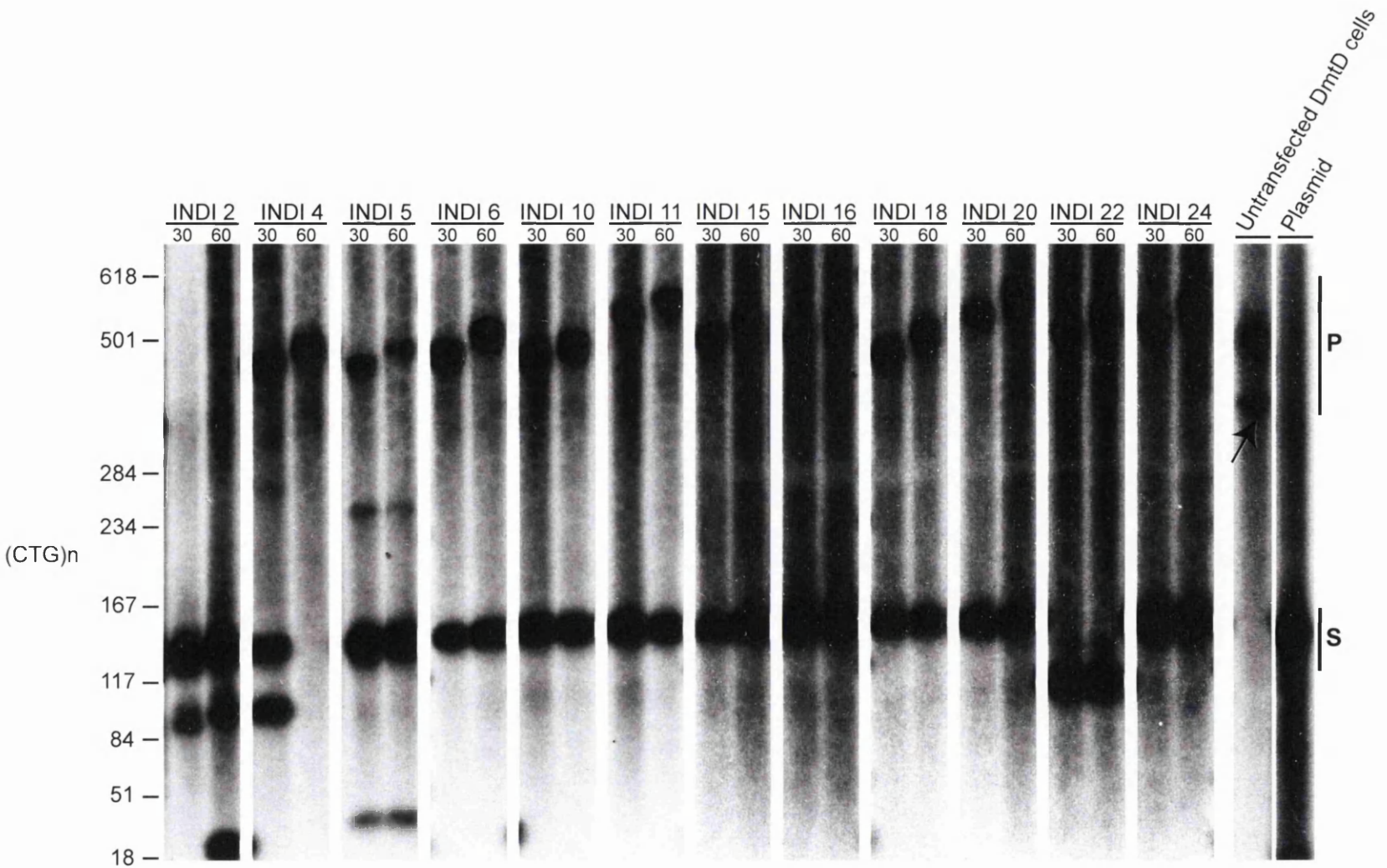


Figure 4.4. Analysis of transgenic repeat lengths in *Dmt*-D kidney cells stably transfected with the insulator positive transgene, INDI. DNA was prepared from cells 30 and 60 population doublings post transfection. Both transgenic repeat tracts were amplified in the same reaction with the primers DM-C & DM-BR. The products were blotted and hybridised to a CTG containing probe (DM56). Size standards, in repeats, are indicated to the left of the figure. Amplified transfectant vector DNA and amplified untransfected *Dmt*-D kidney cell DNA are also shown. The arrow indicates a population of stable expanded alleles in *Dmt*-D kidney cells. Vertical bars on the right of the figure indicate the repeat size range of the primary (**P**) and secondary (**S**) transgenes.

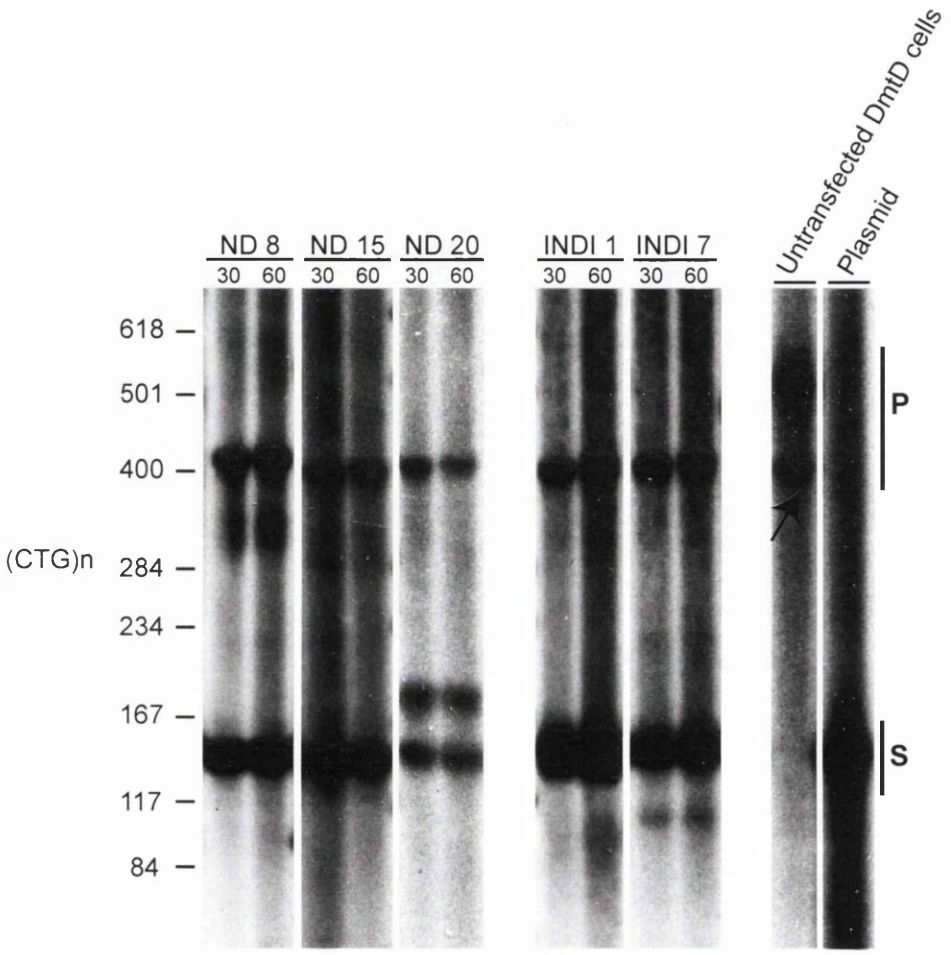


Figure 4.5. A subset of *Dmt*-D kidney cells contain an expanded stable allele. DNA was prepared from cells 30 and 60 population doublings post transfection. Both transgenic repeat tracts were amplified in the same reaction with the primers DM-C & DM-BR. The products were blotted and hybridised to a CTG containing probe (DM56). Size standards, in repeats, are indicated to the left of the figure. Amplified transfectant plasmid DNA and amplified untransfected *Dmt*-D kidney cell DNA is also shown. The arrow indicates a population of stable expanded alleles in *Dmt*-D kidney cells. Vertical bars on the right of the figure indicate the repeat size range of the primary (**P**) and secondary (**S**) transgenes.

evident within the untransfected *Dmt*-D cells (Figure 4.5). Thus, approximately 15% of expanded alleles within the *Dmt*-D kidney cells are stable.

As the secondary transgenic repeat used here is markedly shorter than the expanded primary repeat present in the *Dmt*-D kidney cells, and as expansion rate is length dependent, mutational events would be expected to occur less frequently at the transgenic allele over a given period of time. Thus, the time course over which the experiment was run may have been insufficient to observe expansion-biased instability at the transgenic loci. Five insulator -ve and five insulator +ve clones containing single copies of each transgene were maintained in culture for a further 40 population doublings, and repeat size analysed as before (Figures 4.6 & 4.7). Clones containing single copies of each transgene were identified by PCR amplification of clone DNA with primer pairs designed to amplify across tandemly integrated transgenes, followed by visualisation of PCR products by blotting and hybridisation with radioactive probes. Yet again, no increases in repeat length or obvious broadening in band size was observed for the secondary repeats; in contrast to the continued expansion-biased instability exhibited by the primary transgenic repeat.

4.2.1.3 Small-pool PCR analysis of repeat length variation in stably transfected *Dmt*-D kidney cell lines

The PCR-hybridisation method employed to analyse repeat length changes in the previous sections lacks the sensitivity to detect small (1-2 repeats) or rare changes in repeat length. In order to determine whether such repeat length changes have occurred, samples were analysed by small-pool PCR.

SP-PCR amplification of genomic DNA isolated 100 population doublings post-transfection from both insulator negative and insulator positive cell lines revealed the presence of small of repeat length changes (+/- 1 to 3 repeats) at secondary transgenic loci (Figure 4.8). Some larger (-10 to -30 repeats) contractions were also evident at the secondary transgenic loci (Figure 4.8 & 4.9). The levels of repeat length variation at the secondary repeat loci appeared to increase over time, indicating that the observed repeat length changes accumulated during culture, and did not occur pre-transfection (Figure 4.9).

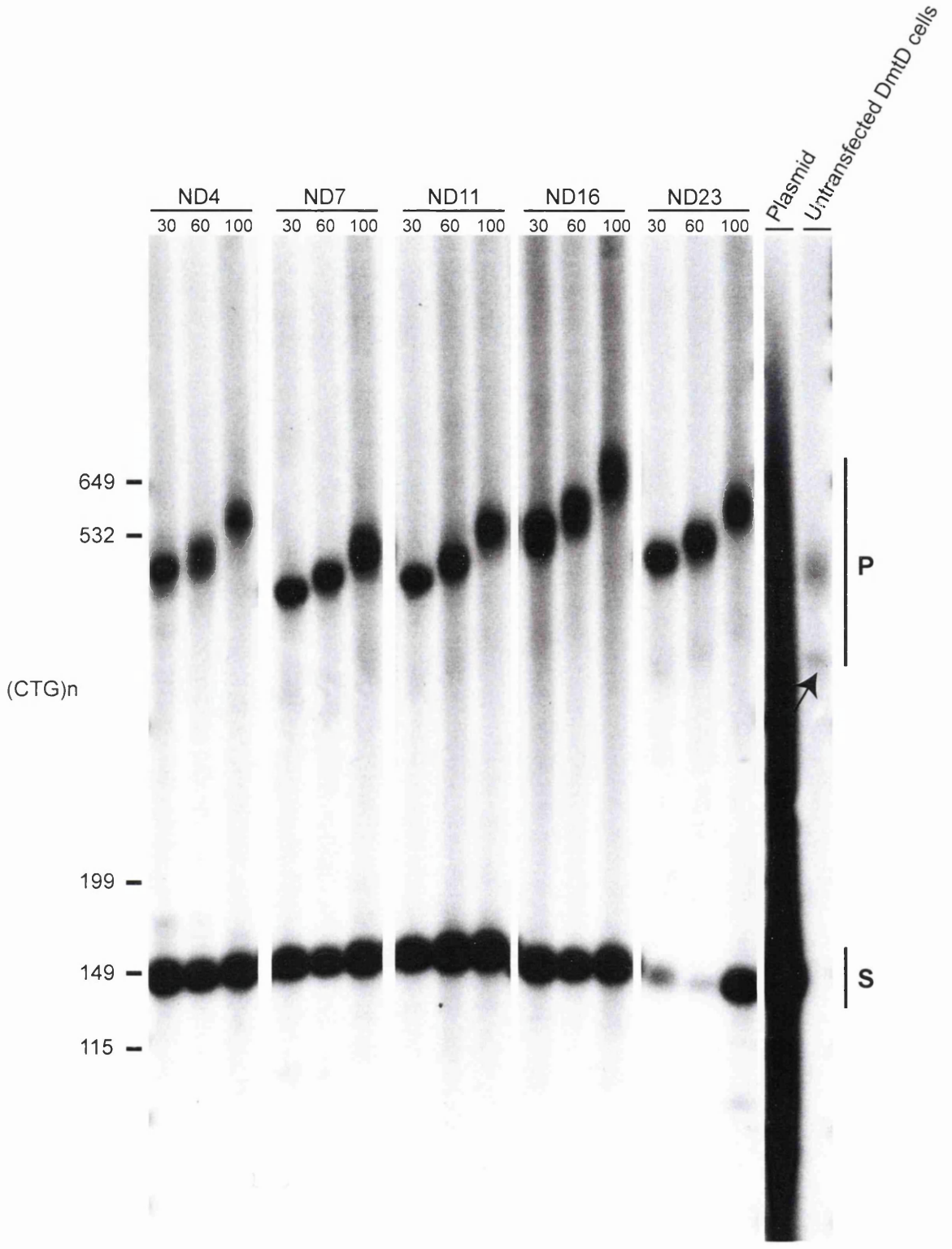


Figure 4.6. Analysis of transgenic repeat lengths in *Dmt*-D kidney cells stably transfected with the insulator negative transgene, ND. DNA was prepared fom cells 30, 60 and 100 population doublings post transfection. Both repeat tracts were amplified in the same reaction with the primers DM-C & DM-ER.The products were blotted and hybridised to a CTG containing probe (DM56). Size standards, in repeats, are indicated to the left of the figure. Amplified transfectant plasmid DNA and amplified untransfected *Dmt*-D kidney cell DNA is also shown. The arrow indicates a population of stable expanded alleles in *Dmt*-D kidney cells. Vertical bars on the right of the figure indicate the repeat size range of the primary (**P**) and secondary (**S**) transgenes.

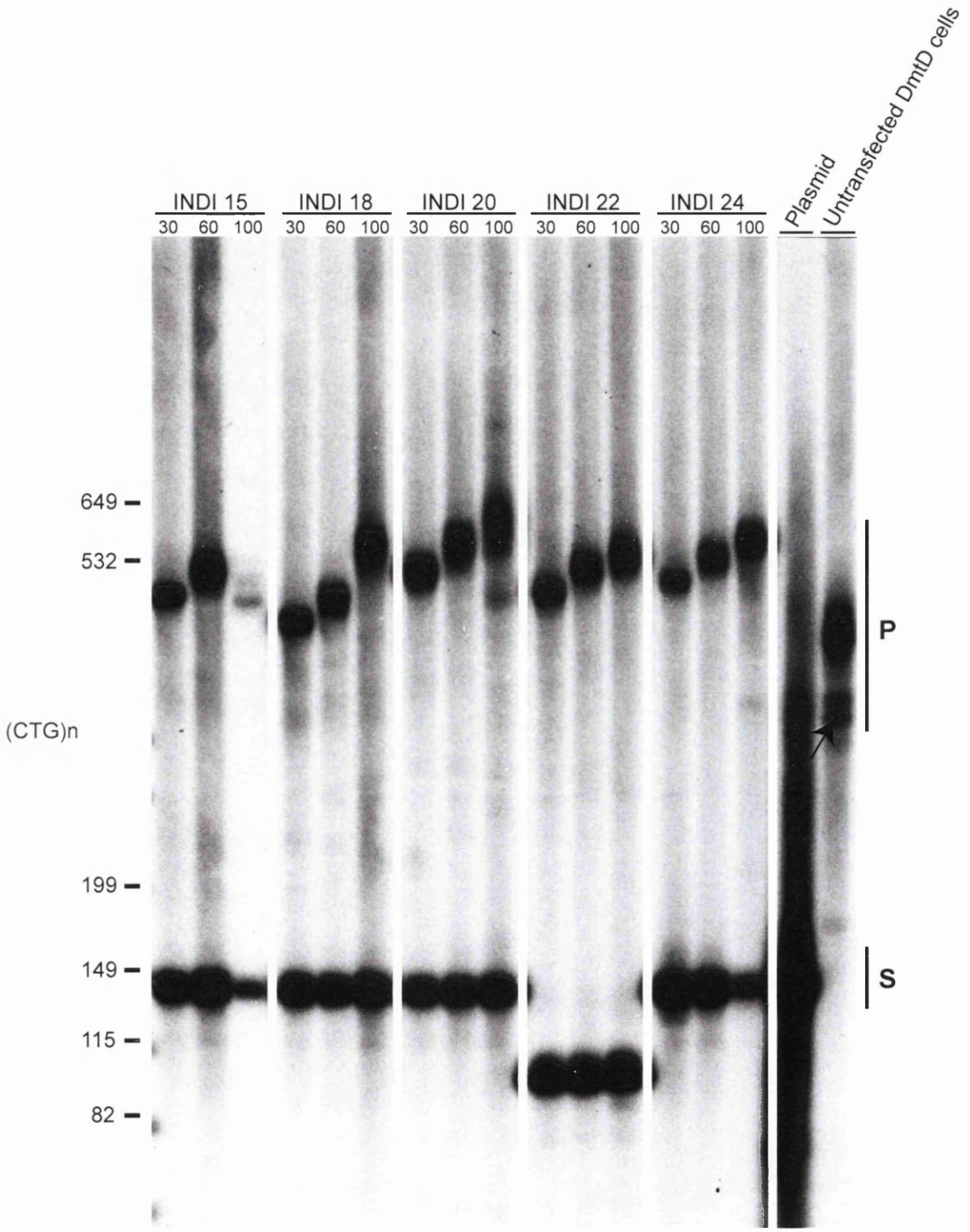


Figure 4.7. Analysis of transgenic repeat lengths in *Dmt*-D kidney cells stably transfected with the insulator positive transgene, INDI. DNA was prepared fom cells 30, 60 and 100 population doublings post transfection. Both repeat tracts were amplified in the same reaction with the primers DM-C & DM-ER.The products were blotted and hybridised to a CTG containing probe (DM56). Size standards, in repeats, are indicated to the left of the figure. Amplified transfectant plasmid DNA and amplified untransfected *Dmt*-D kidney cell DNA is also shown. The arrow indicates a population of stable expanded alleles in *Dmt*-D kidney cells. Vertical bars on the right of the figure indicate the repeat size range of the primary (**P**) and secondary (**S**) transgenes.

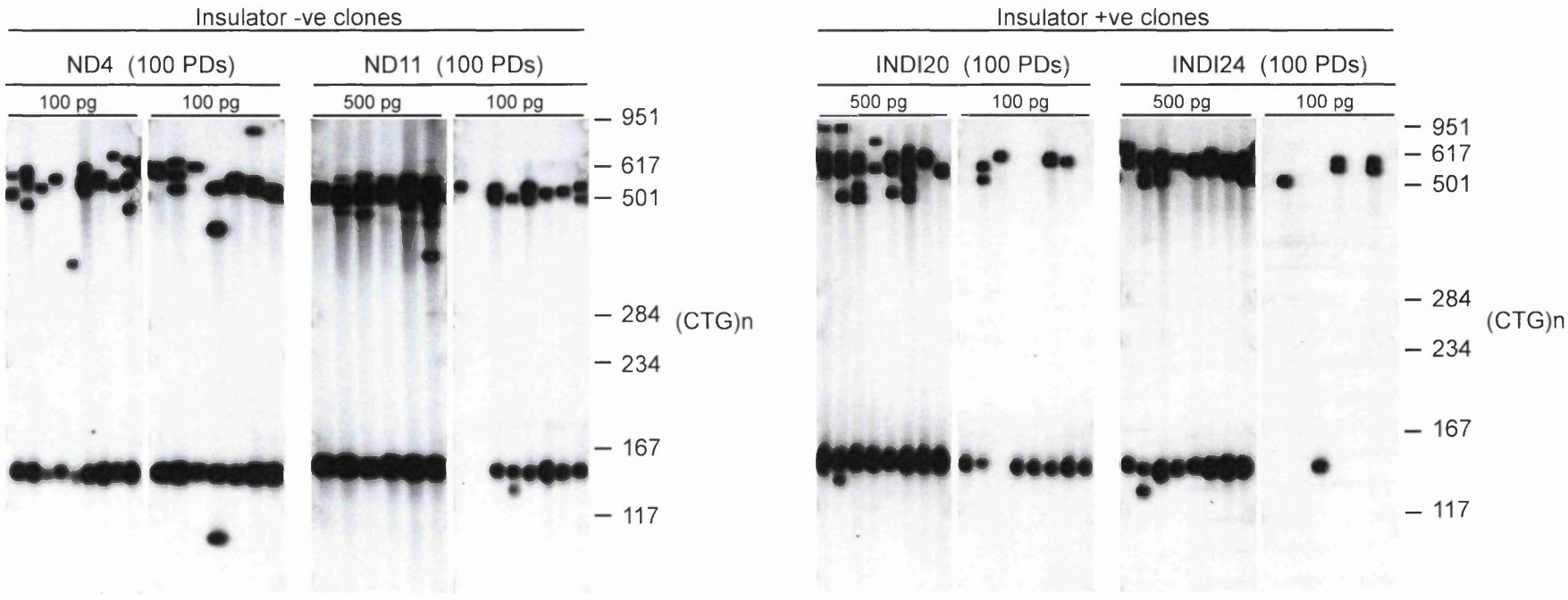


Figure 4.8. Small-pool PCR analysis of primary and secondary transgenic repeat tracts in DmtD kidney cells transfected with an insulator negative (ND) or insulator positive (INDI) secondary transgene. DNA was prepared 100 population doublings post-transfection. Size standards, in repeats, are indicated the right of the figure. Both transgenic repeats were amplified in the same reaction with the primers DM-C & DM-BR. The amount of genomic template DNA used in each reaction (each lane) is indicated at the top of each panel.

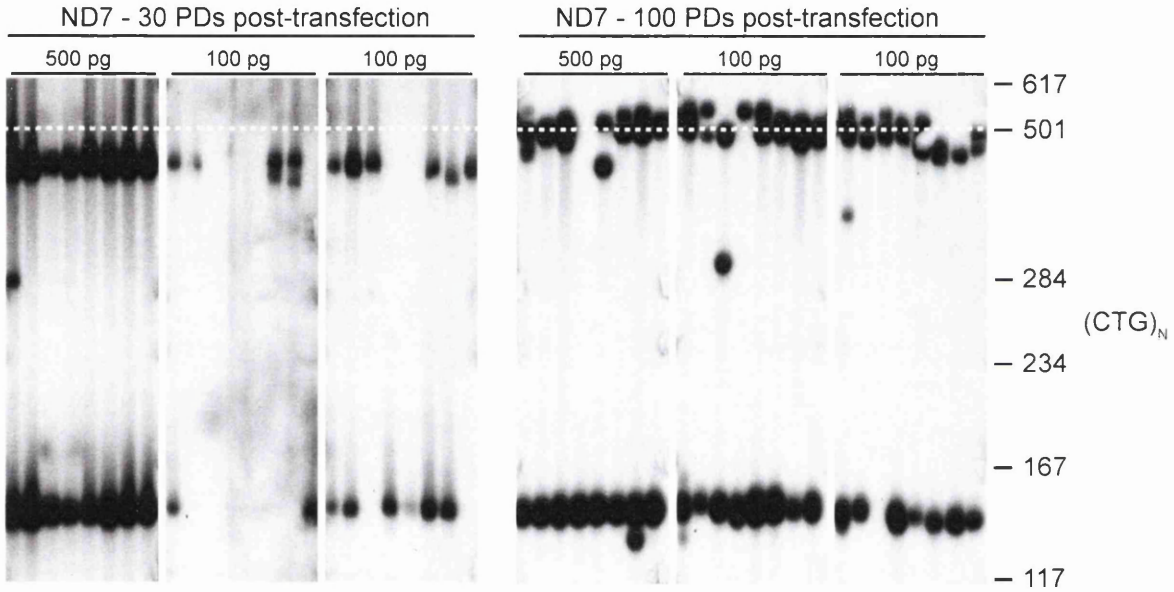


Figure 4.9. Small-pool PCR analysis of primary and secondary transgenic repeat tracts in DmtD kidney cells containing an insulator negative (ND) secondary transgene. DNA was prepared 30 and 100 population doublings post-transfection. Size standards, in repeats, are indicated the right of the figure. Both transgenic repeats were amplified in the same reacion with the primers DM-C & DM-BR. The amount of genomic template DNA used in each reaction (each lane) is indicated at the top of each panel. The white dashed line is shown to facilitate comparison of band size between blots. PD = Population Doubling.

However, in contrast to the obvious expansion-biased instability observed at the primary repeat loci, no gross expansion bias was evident at the secondary loci (Figure 4.9).

In order to more accurately quantify the difference in repeat length variation between the primary and secondary transgenes, alleles were sized from single molecule amplifications using molecular imaging software (Kodak MSI, v 4.0.5). As the degree of length variation in the secondary transgene was very low, the number of molecules amplified was predicted using Poisson analysis. The mean repeat length of ~40 and ~25 alleles was determined for the primary and secondary transgenes at each time-point, respectively. Quantitative analysis of repeat length variation in three insulator negative cell lines found no expansion-bias in repeat instability at the secondary loci; one cell line showing a significant decrease in repeat length during culture (ND11, Mann-Whitney U = 59, $P < 0.01$). In contrast, the primary repeat tracts showed dramatic expansion biased instability (Figure 4.10).

As outlined previously, repeat instability is length dependent, longer repeats exhibiting higher levels of repeat length variation. Therefore, the secondary transgenic repeat employed here may possess too few repeats to effect observable/quantifiable levels of instability over the time-scale of this experiment. To address this issue, a repeat length normalised measure of primary repeat expansion was determined using the following equation:

$$E_c = \frac{\mathbf{mT2 - mT1}}{\mathbf{mT1 - 35}}$$

where E_c = coefficient of expansion,

mT2 = median repeat length at time two

mT1 = median repeat length at time one

Thus, for every repeat in the primary repeat tract above the stability threshold (35 repeats), an expansion of 0.12 repeats was observed between 30 and 100 populations doubling in culture. Applying this coefficient of expansion to the secondary repeat, a median expansion of 14 repeats (42 bp) would be expected

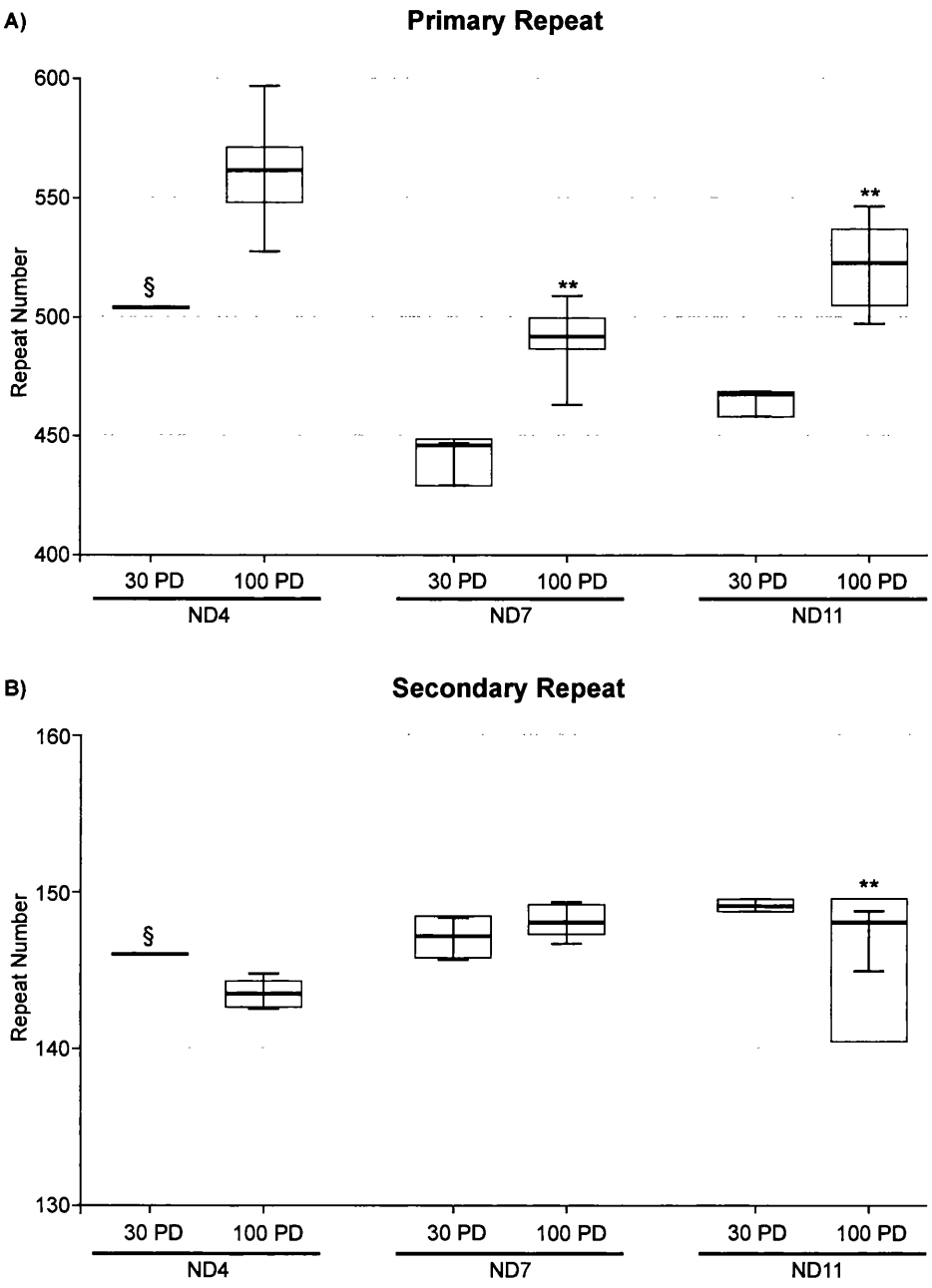


Figure 4.10. Change of transgenic repeat length in insulator negative cell lines as determined by SP-PCR. A) Repeat length of primary transgene at 30 population doublings and 100 population doublings post-transfection. **B)** Repeat length of secondary transgene at 30 population doublings and 100 population doublings post-transfection. Boxplots show median (solid bar), 95% confidence limits of the median (box), and the 25-75 % inter-quartile range (T-bars). The Mann-Whitney U-Test was used to determine if repeat sizes observed at the two time-points were significantly different (** $P < 0.01$). § indicates median repeat size was estimated from a single observation from a standard PCR-hybrisation (not SP-PCR) analysis. Note: panels A and B have different scales

over the course of this experiment, well within the detection limits of the analytical techniques employed here, suggesting that the repeat length alone does not explain the observed stability of the secondary transgene. In addition, the coefficient of variation (the standard deviation of a sample expressed as a percentage of the mean) of primary repeat length is approximately 2.5 times greater than the corresponding value for secondary repeat length (*N* = 3; *CV* primary rpt = 13.6%; *CV* secondary rpt = 5.3%). This indicates that repeat length variation, irrespective of direction of change (expansions or contractions), is also reduced at the secondary repeat loci.

The observation that two transgenes containing an expanded CTG•CAG repeat flanked by the same human DNA sequences present in the same cell show such contrasting stability profiles is striking. Interestingly, such cells represent a potent model system for identifying *cis*-acting modifiers of expanded repeat stability by facilitating comparison of a stable and unstable expanded repeat tract within the same cell.

4.2.2 Methylation state of expanded transgenic repeats

4.2.2.1 Methylation state of transgenic repeats in mouse Dmt-D kidney cells

The finding that expanded CTG•CAG transgenic repeats with similar flanking sequences show different levels of stability in the same cell line suggest the involvement of *cis*-acting modifiers of instability. As instability was not observed in 35 independent single cell *Dmt*-D clones, it is unlikely that simple site-of-integration position effects, such as flanking sequence composition or genic context, are the cause of the observed stability. However, it is possible that the secondary transgenes induce epigenetic changes in *cis*, which are not conducive to repeat instability. Evidence from affected individuals and murine models of expanded CTG•CAG repeat disorders have implicated DNA methylation of expanded repeats and their flanking sequences as a potential modifier of repeat stability (Chapter 1).

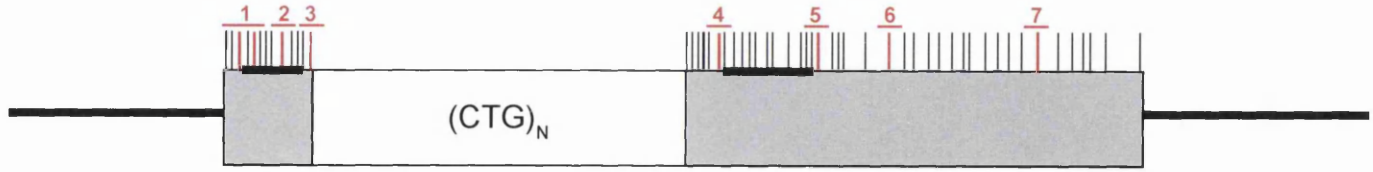
The methylation status of the primary and secondary transgenes was determined by methylation-sensitive restriction digest/PCR (MS-PCR) analysis. In brief, genomic DNA samples were digested with a methylation-sensitive restriction endonuclease with a recognition site within the transgenic sequence. Such restriction enzymes fail to cut their corresponding restriction site if it contains methylated CpG dinucleotides. Thus, subsequent PCR across the region containing the restriction site of interest will succeed only if the site was methylated and thus the template DNA sequence uncut and intact. An internal control sequence, either lacking or containing the site of interest (dependent on the observed methylation status of the test site) was also amplified in each reaction (Figure 4.11). A total of 8 CpG sites in the *Dmt* transgene were analysed, representing 14% of the 55 CpG sites present in the *Dmt* transgene (Figure 4.11). The primer combinations used to assay each site are given in Table 4.1.

Table 4.1. *Dmt* restriction sites and associated primer combinations used in methylation-sensitive restriction digest/PCR assays.

	Primary transgene		**Insulator –ve transgene**		**Insulator +ve transgene**	
CpG site	Test primers	Control primers	Test primers	Control primers	Test Primers	Control primers
1. *ApaI*	mDmtD-B DM-CR	neoF1 neoR1	neoF4 DM-CR	neoF1 neoR1	neoF4 DM-CR	neoF1 neoR1
2. *NaeI*	mDmtD-B DM-CR	neoF1 neoR1	neoF4 DM-CR	neoF1 neoR1	neoF4 DM-CR	neoF1 neoR1
3. *HpaII*	DM-C DM-BR	mP2-1 mP2-5	DM-C DM-BR	mP2-1 mP2-5	DM-C DM-BR	mP2-1 mP2-5
4. *HpaII*	DM-C DM-BR	mP2-1 mP2-5	DM-C DM-BR	mP2-1 mP2-5	DM-C DM-BR	mP2-1 mP2-5
5. *AvaI*	DM-F mDmtD-MR	neoF1 neoR1	DM-F T3	neoF1 neoR1	DM-F INDI_SR	neoF1 neoR1
6. *NaeI*	DM-F mDmtD-MR	neoF1 neoR1	DM-F T3	neoF1 neoR1	DM-F INDI_SR	neoF1 neoR1
7. *NruI*	DM-F mDmtD-MR	neoF1 neoR1	DM-F T3	neoF1 neoR1	DM-F INDI_SR	neoF1 neoR1

The MS-PCR analysis of both the primary and secondary transgenes was applied to genomic DNA isolated 30 population doublings post-transfection from both insulator negative (N = 12) and insulator positive (N = 15) cell lines. The primary and secondary transgenes were found to have significantly different levels of

A) CpG sites in *Dmt* transgene



B) Assay of *Nru*I site in insulator negative clones

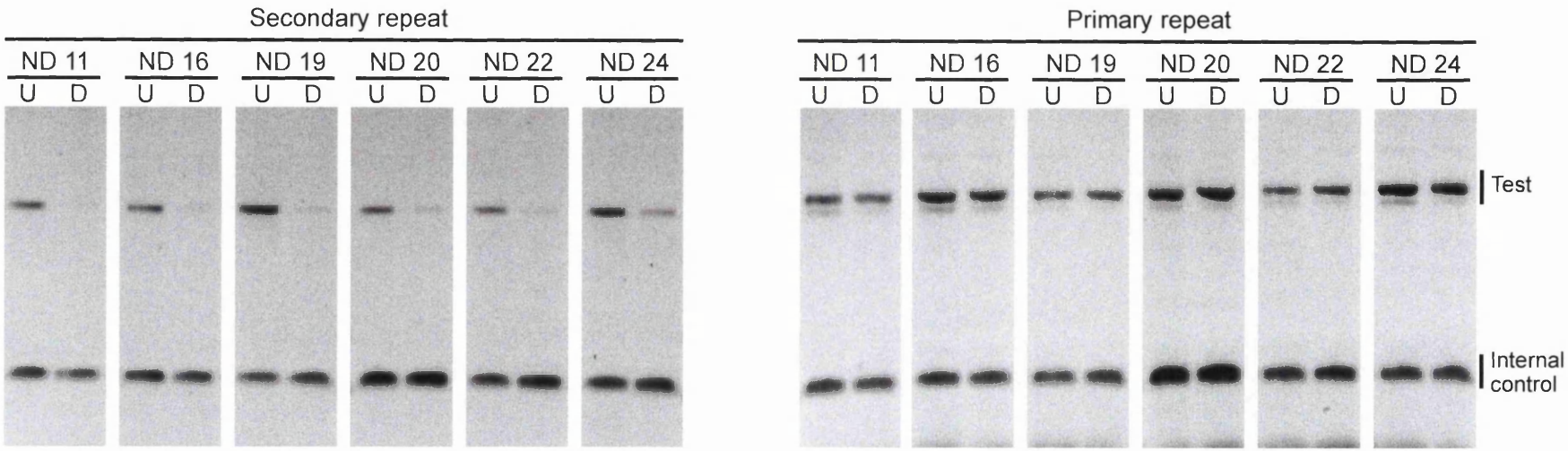


Figure 4.11. Methylation-sensitive restriction digest/PCR assay of transgenic repeats. A) Position of all CpG sites in the *Dmt* transgene. Red lines indicate CpG sites assayed here by methylation-sensitive restriction digest/PCR. Numbers indicate restriction enzyme site tested: 1 = *Apa*I, 2 = *Nae*I, 3 = *Hpa*II, 4 = *Hpa*II, 5 = *Ava*I, 6 = *Nae*I, 7 = *Nru*I. CTCF binding sites are shown as black bars **B)** Methylation-sensitive restriction digest/PCR of 3' *Nru*I site. Genomic DNA was digested overnight with *Nru*I. 70 ng of digested DNA was amplified with either the primers DM-F & mDmtD-MR (primary repeat) or DM-F & T3 (secondary repeat) and the primers neoF1 & neoR1 (internal undigested control) by 28 cycles of PCR. Failure to amplify the PCR product after digestion indicates an unmethylated, and thus digested CpG site (secondary transgene). U = undigested genomic DNA; D = digested genomic DNA.

CpG methylation on the sequences flanking the expanded repeats (Fishers Exact test; P = 0.03). The DNA flanking the primary repeat was highly methylated in all 27 clones (Figure 4.12). In contrast, the DNA flanking both secondary repeats was completely unmethylated in all 27 clones (Figure 4.12). The methylation profile of the stable primary repeats was identical to that of the unstable primary alleles.

4.2.2.2 Generation of cells lines transgenic for methylated expanded repeats

The finding that the primary and secondary repeats are associated with hyper-methylated and hypo-methylated flanking sequences, respectively, suggests that the methylation status of sequences flanking DNA repeats may be a modifier of expanded repeat stability. In order to determine the effect of flanking sequence methylation on expanded repeat stability, three insulator negative constructs were generated, each containing different levels of flanking CpG methylation.

As hyper-methylation of the *neo* promoter sequence would likely silence expression of the *neo* gene, reducing the likelihood of positive selection of transfected cells, a semi-methylated construct in which only the repeat containing portion of the construct was methylated was generated (M.CpG ND)(Figure 4.13A). In brief, the plasmid pB-ND, containing the insulator negative construct (ND) was methylated with the methylase M.CpG (M.*Sss*I) and digested with the restriction enodnucleases *Ale*I and *Nco*I. The digested fragments were resolved on an agarose gel, from which the *Dmt* containing fragment (M.*Dmt*) was purified. The M.*Dmt* fragment was then ligated to an umethylated fragment containing the *neo* portion of the ND construct. Ligation products were resolved on an agarose gel and the fragment corresponding to the ligated semi-methylated construct was purified. Integrity and methylation status of the M.CpG ND construct was assayed by digestion with methylation sensitive restriction endonucleases. Two partially methylated constructs were generated by methylation of the entire ND construct (including the *neo* cassette) with either M.*Hha*I or M.*Hpa*II methylase; methyltransferases which methylate the CpG dinucleotides, contained in either *Hha*I or *Hpa*II restriction sites, respectively (Figure 4.13 B & C). As the *neo* promoter contains just 5 *Hpa*II sites

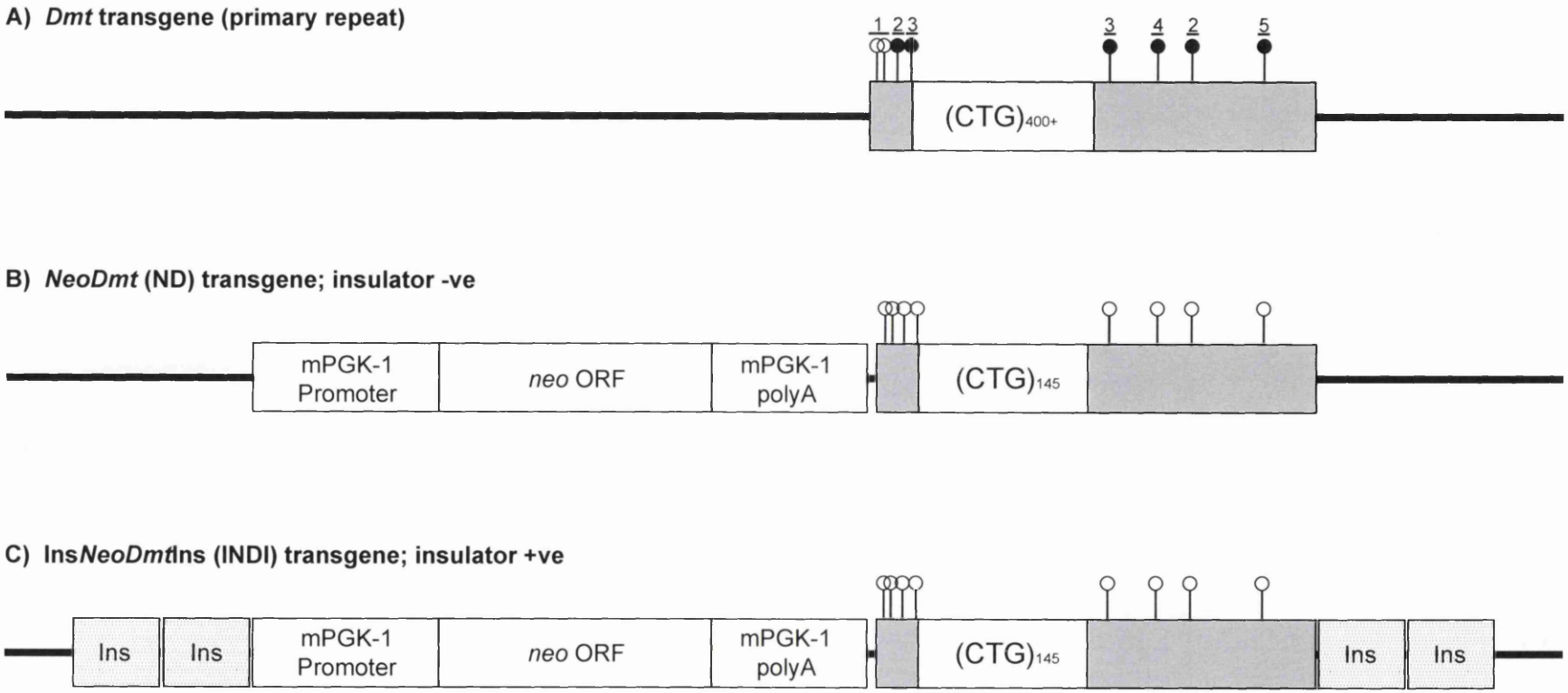


Figure 4.12. Methylation pattern of DmtD transgenic repeats. Empty circle indicates an unmethylated CpG dinucleotide, solid circle indicates a methylated CpG dinucleotide. Numbers indicate restriction enzyme site tested: 1 = *Apa*I, 2 = *Nae*I, 3 = *Hpa*II, 4 = *Ava*I, and 5 = *Nru*I. mPGK-1 = mouse phosphoglycerate kinase-1, Ins = insulator element (250 bp core of the chicken ß-globin locus hypersensitive site 4)

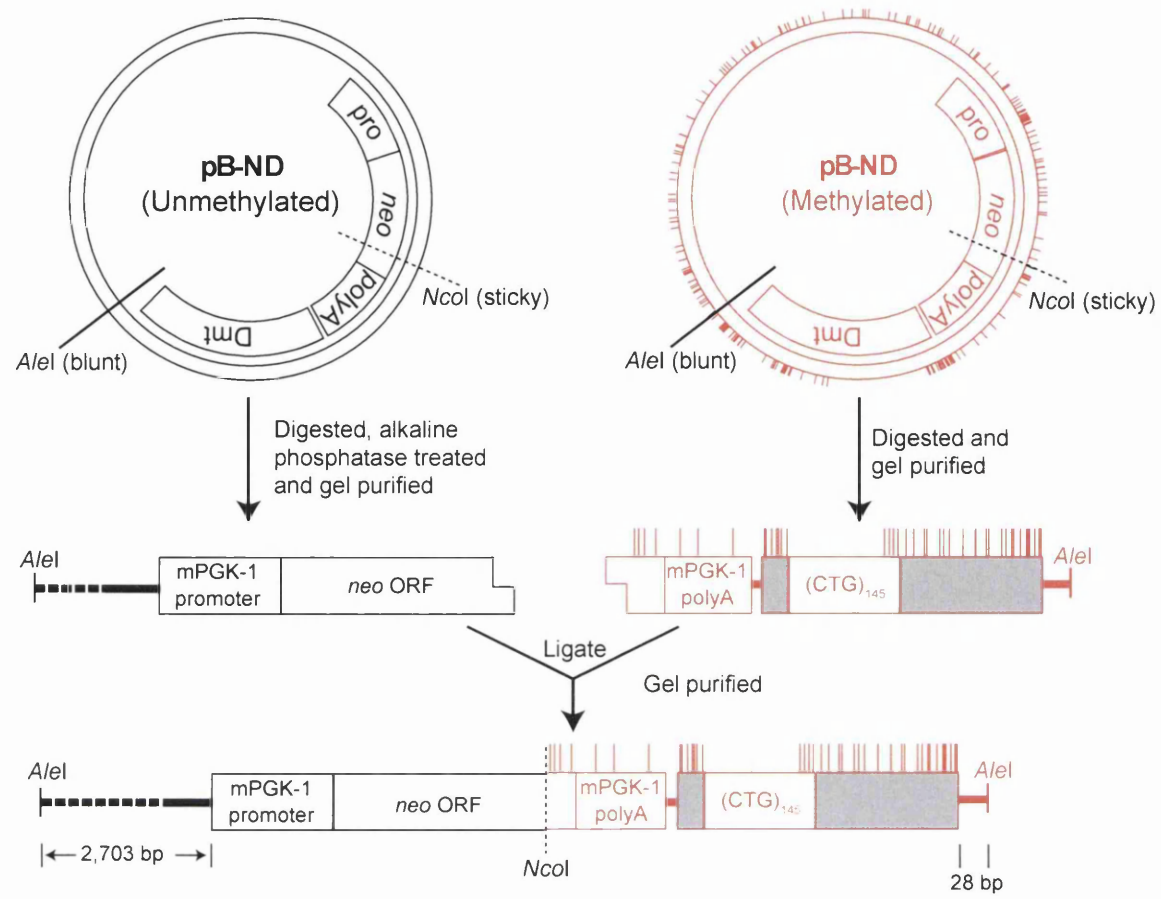


Figure 4.13. Generation of methylated insulator negative constructs. A) Generation of M.Chimera ND from a methylated and unmethylated pB-ND vectors (see text for description of cloning strategy). Methylated pB-ND plasmid is shown in red. **B)** Methylated *Hpa*II sites in M.*Hpa*II ND (see text for description of cloning strategy). **C)** Methylated *Hha*I sites in M.*Hha*I ND (see text for description of cloning strategy). Vertical bars indicate position of methylated CpG sites.

and 6 *Hha*I sites, promoter activity may not be fully repressed by methylation at these sites, allowing for *neo* expression and positive selection. As DNA methylation patterns are not maintained in bacteria, all cloning steps were performed without the use of bacteria.

Dmt-D kidney cells, stably transfected with either the M.*HhaI* transgene (N = 5), M.*HpaII* transgene (N = 7) or M.CpG ND transgene (N = 9) were generated and maintained in culture for 40 population doublings (~ 38 days). Routine analysis of transgene integrity in each cell line by PCR amplification assays identified several lines in which large deletions in which the 3' end of the transgene incorporating some or all of the *Dmt* sequence had been deleted. Clones in which some or the entire *Dmt* portion of the transgene was missing were excluded from the study.

MS-PCR of the remaining clones revealed that the M.CpG ND transgenes had retained their hyper-methylation in culture (N = 5), and the M.*Hha*I transgenes retained methylation at half of their *Hha*I sites (N = 3). However, the M.*Hpa*II transgenes appeared to have lost all CpG methylation (N = 5) (Figure 4.14).

To investigate changes in repeat length, the transgenic repeats were amplified (using primers DM-C and DM-BR) from genomic DNA prepared 20 and 40 population doublings post transfection from each clone and visualised by hybridisation with a CTG-containing probe (DM56). Once again, no obvious repeat length variation was observed in the secondary repeat tracts over time, whereas expansion-biased instability was evident in most primary repeats (Figure 4.15). In addition, SP-PCR analysis failed to reveal repeat length variation in the secondary repeat tracts of M.CpG ND (N = 2), M.*Hha*I (N = 2), or M.*Hpa*II (N = 1) cell lines (Figure 4.16).

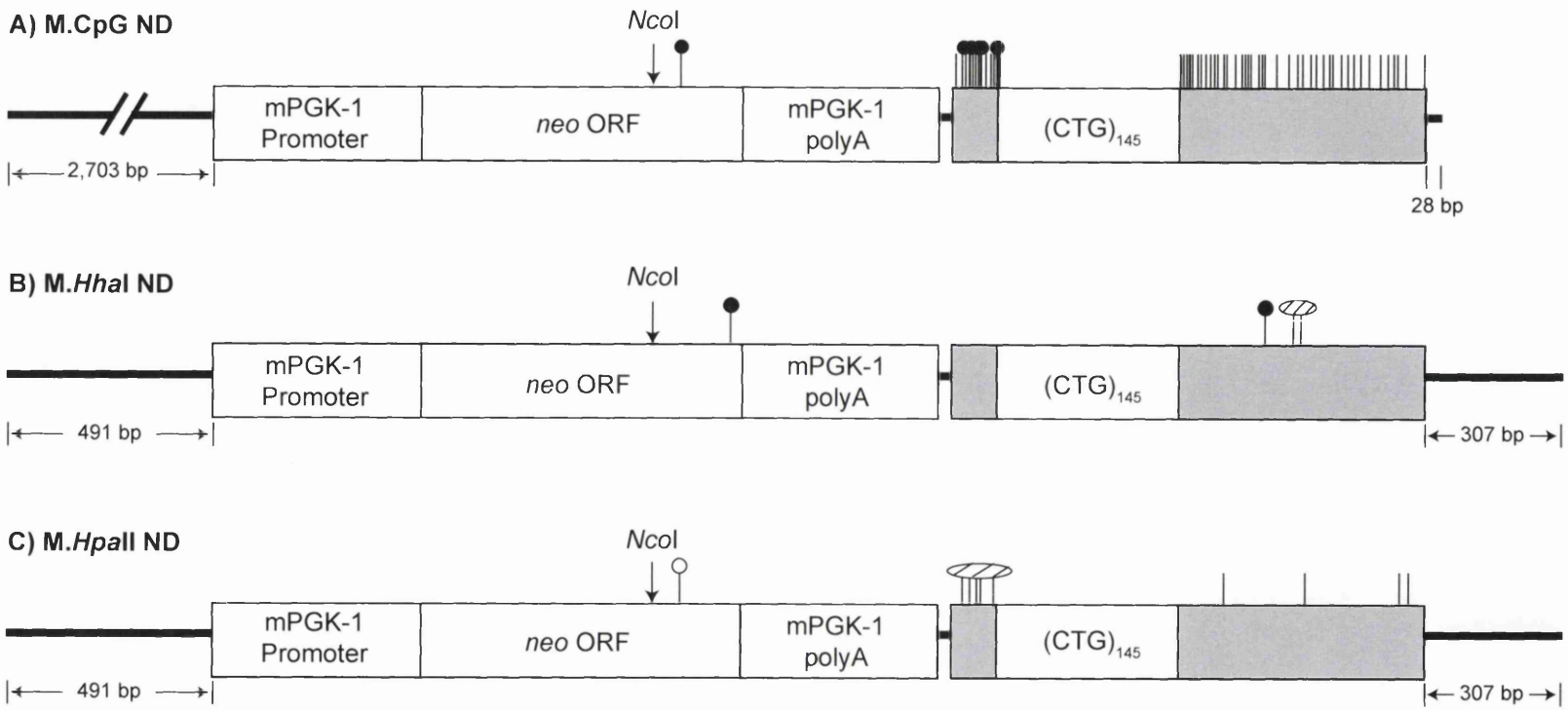


Figure 4.14. Methylation state of methylated transgenes post transfection. Vertical lines indicate the position of CpG sites in the *Dmt* transgene methylated pre-transfection. Solid circles indicate methylated CpG dinucleotides, crossed ovals indicate multiple sites assayed simultaneously in which at least one site is unmethylated. Arrow indicates location of *Nco*I site. mPGK-1 = mouse phosphoglycerate kinase-1.

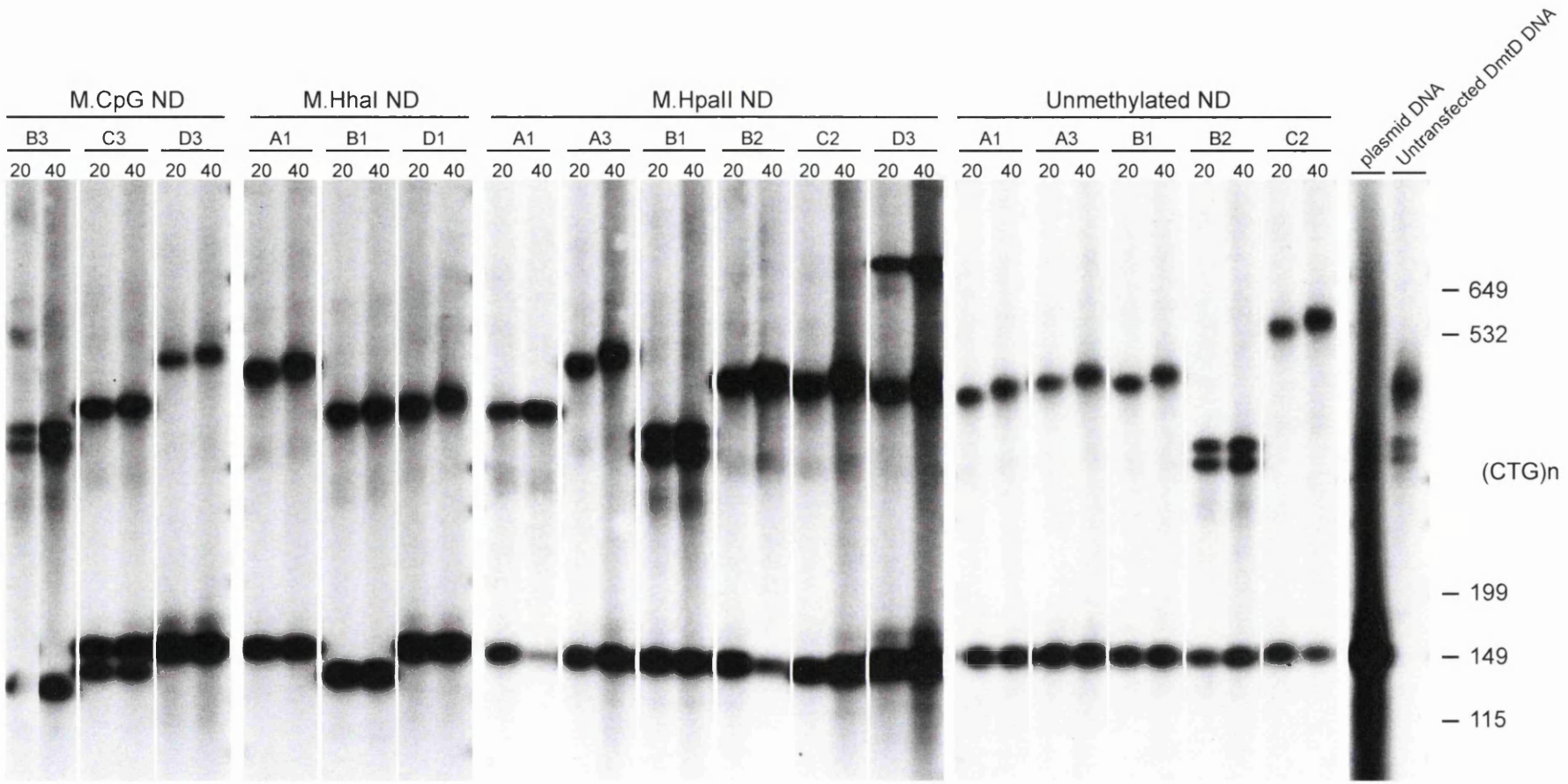


Figure 4.15 Analysis of methylated secondary repeats in DmtD kidney cells. DNA was prepared from cells 20 and 40 population doublings post-transfction. Both transgenic repeats were amplified in the same reaction with the primer DM-C & DM-ER. The products were blotted and hybridised to a repeat-containing probe (DM56). Size standards, in repeats, are indicated to the right of the figure. Amplified transfectant plasmid DNA and amplified untransfected DmtD kidney cell DNA are also shown.

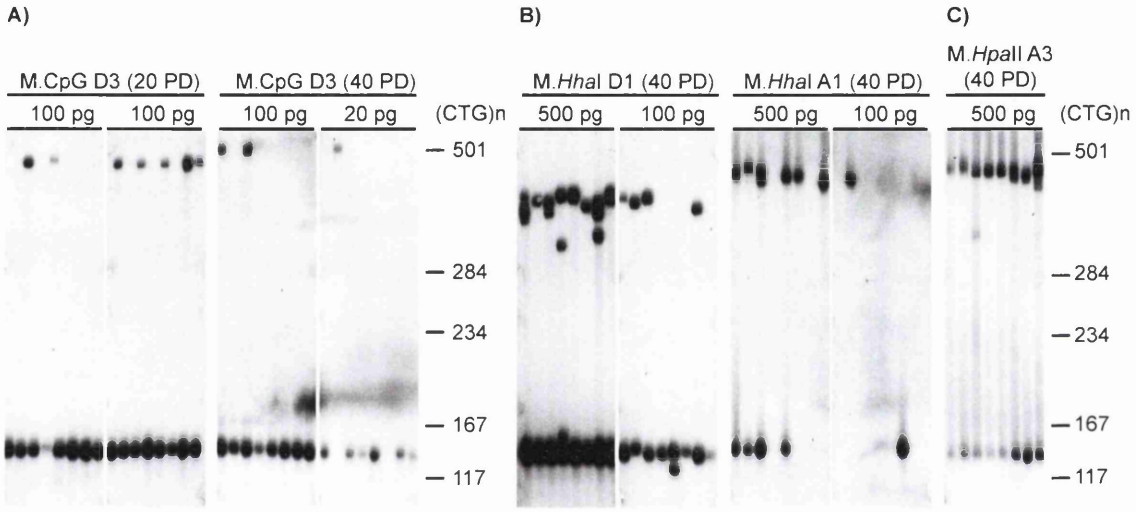


Figure 4.16. SP-PCR analysis of methylated transgenic repeats in DmtD cell lines. A) Repeat length variation is not evident in the hyper-methylated secondary transgene. **B)** Low levels of repeat variation are observed in the secondary transgene methylated only at *Hha*I sites (N = 2). **C)** Repeat length variation is not evident in the secondary transgene methylated only at *Hpa*II sites. Band sizes, in repeats, are shown on the right of the blots. The amount of genomic template DNA used in each reaction (each lane) is indicated at the top of each panel. PD = population doubling.

4.2.3 Semi-quantitative RT-PCR of expanded transgenic repeats

A positive association between repeat instability and repeat transcription has been suggested in transgenic models of expanded CTG•CAG disorders. Significantly, the *Dmt*-D transgene is expressed at low levels in kidney cells of the *Dmt*-D mouse model. Thus, differing expression levels of the primary and secondary transgenes may account for the contrasting stability profiles observed between the transgenic repeats.

In order to investigate transgene expression levels, total RNA was prepared from *Dmt*-D cell lines transgenic for the insulator negative (N =8) or insulator positive (N = 7) secondary transgenes. RNA concentration and purity was assayed by spectrophotometry; a ratio of sample absorbance at 260 nm and 280 nm of 2-2.2 indicating a clean RNA sample. RNA integrity (degree of degradation) was assayed by gel electrophoresis (Figure 4.17A). Contaminating genomic DNA was removed by digestion with DNaseI. cDNA was synthesised from RNA (2500 ng) using Superscript II reverse transcriptase (RT) primed with random hexamers. Samples were tested for the presence of contaminant genomic DNA by PCR amplification of the GAPDH gene from RT negative controls of each sample (Figure 4.17B).

As the primary repeat exhibits dramatic repeat length variation at 100 population doublings post-transfection, RT-PCR across the repeat tract would result in a broad smear when resolved on a gel, rendering comparison of expression levels with the stable secondary transgene problematic. Thus, the 3' flank of the primary transgene, insulator negative transgene, and insulator positive transgene was amplified using the primer pairs DM-F & mDmtD-MR, DM-F & T3, and DM-F & INDI_SR, respectively. Products were visualised by hybridisation with the probe RPR and quantified by densitometry, correcting for both input cDNA concentration and PCR efficiency (Figure 4.17C & 4.17D).

The analyses revealed that the primary transgene was expressed in all cell lines tested (N = 15). In addition, expression levels of the stable (N = 4) and unstable (N = 11) primary transgenes were not significantly different (Mann-Whitney U = 28; P = 0.54) (Figure 4.18A). Expression of the secondary transgene was observed

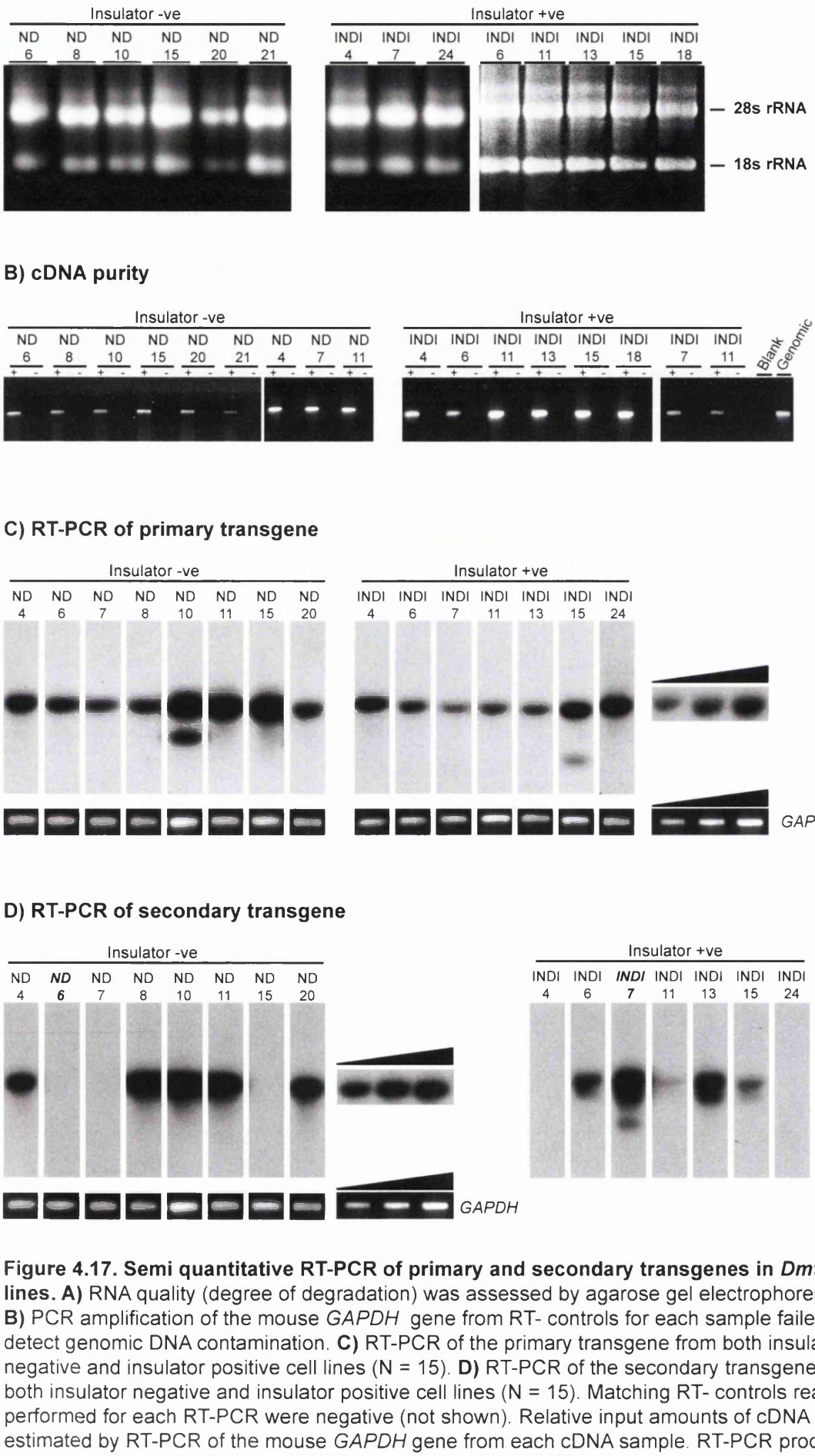


Figure 4.17. Semi quantitative RT-PCR of primary and secondary transgenes in *Dmt*D cell lines. A) RNA quality (degree of degradation) was assessed by agarose gel electrophoresis. **B)** PCR amplification of the mouse *GAPDH* gene from RT- controls for each sample failed to detect genomic DNA contamination. **C)** RT-PCR of the primary transgene from both insulator negative and insulator positive cell lines (N = 15). **D)** RT-PCR of the secondary transgene from both insulator negative and insulator positive cell lines (N = 15). Matching RT- controls reactions performed for each RT-PCR were negative (not shown). Relative input amounts of cDNA were estimated by RT-PCR of the mouse *GAPDH* gene from each cDNA sample. RT-PCR products were visualised by blotting and hybridisation to the probe RPR.

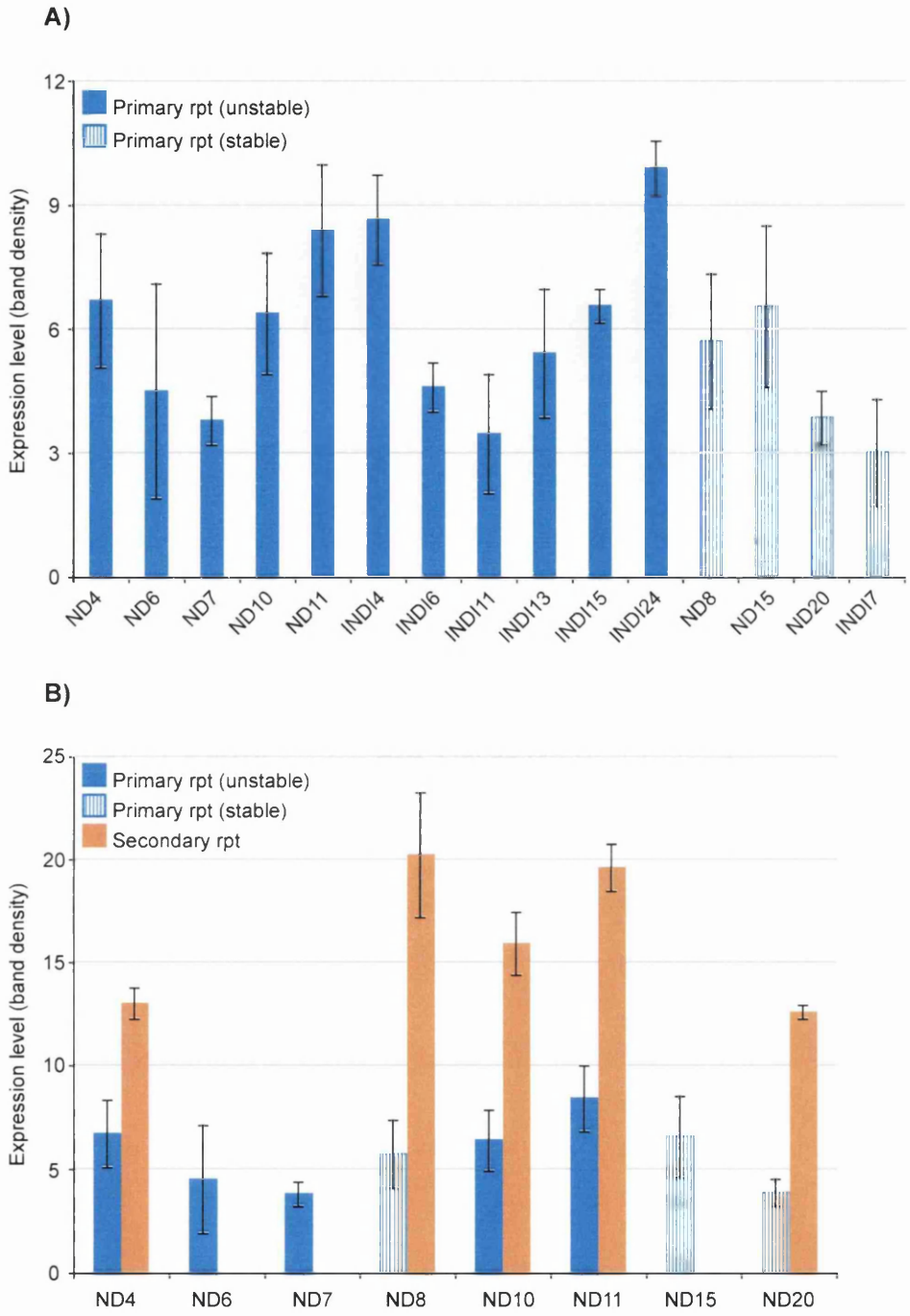


Figure 4.18 Semi-quantitative RT-PCR analysis of primary and secondary transgenes. **A)** Relative expresion levels of stable (N = 4) and unstable (N = 13) primary transgenes in *Dmt*D cell lines. **B)** Relative expresion levels of primary and secondary transgenes in nine insulator negative *Dmt*D cell lines. Expression levels are relative, and were estimated by densiometric analysis of band density. All expression measurements have been corrected for input cDNA template concentration and relative PCR efficiency.

in five of the eight insulator negative clones and four of the seven insulator positive clones. The insulator negative transgene was expressed at a significantly higher level than the primary transgene (Mann-Whitney U = 0, P = 0.02) (Figure 4.18B). RT-PCR of the insulator positive transgene was not of sufficient quality to allow for quantification.

These results exclude gross differences in repeat expression level as a major modifier of the contrasting repeat length variation observed between the unstable primary transgenes and either the stable primary transgenes or the secondary transgenes.

4.2.4 A model system for the study of *cis*-acting modifiers of expanded repeat stability

As outlined above (Section 4.1), chromosomal position effects complicate the study of *cis*-acting modifiers of expanded repeat instability. Thus, a model system whereby transgenic repeats could be faithfully, and repeatedly directed to the same genomic location, thereby normalising for insertion-site effects, is desirable. Such a model system is outlined in Figure 4.19.

In brief, the model system utilises the technique of recombinase-mediated cassette exchange (RMCE), which allows for repeated targeting of replacement cassettes to the site of integration of the targeting cassette using the Cre-Lox system (Figure 4.19). Replacement of the targeting cassette can be efficiently selected for by gain of resistance to gancyclovir due to loss of the *HyTK* gene. Moreover, by flanking the targeting cassette with insulators elements, the replacement cassettes would be predicted to be maintained in an open chromatin state, which is potentially instability promoting.

Once constructed, the model system would allow for comparison of the effects of flanking sequence on instability both within and between disorders, independent of the effects of site of integration. For example, the length of endogenous flanking sequence surrounding an expanded repeat locus could be progressively reduced in order to identify the location of *cis*-acting elements. In

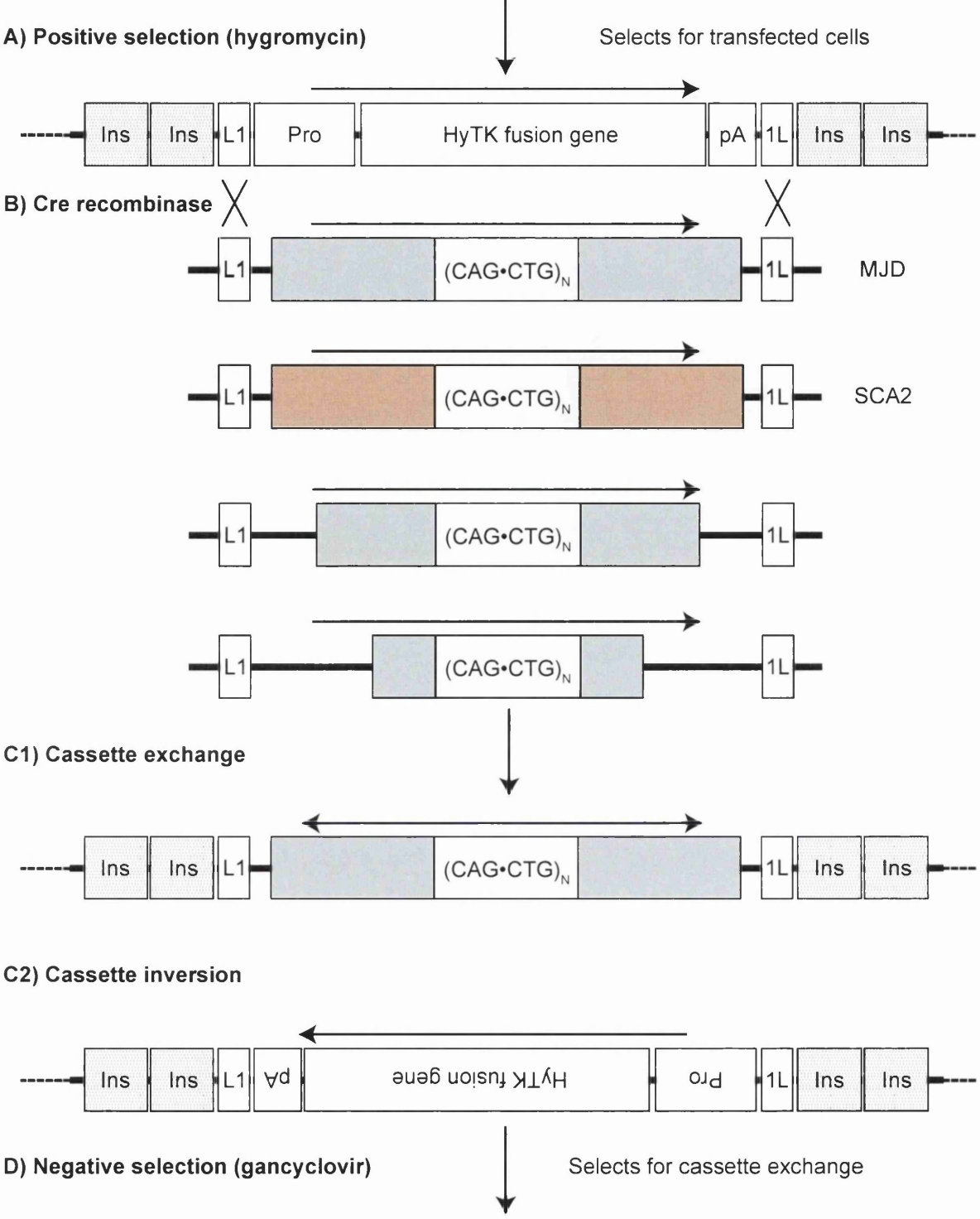


Figure 4.19. Recombinase-mediated cassette exchange of expanded repeats. The targeting cassette consists of a hygromycin phosphotransferase gene fused to a thymidine kinase gene (HyTK) which allows for positive selection with hygromycin B and negative selection with gancyclovir, flanked by inverted Lox sites (L1 and 1L), and two pairs of insulator elements from the chicken ß-globin locus (Ins). **A)** Hygromycin B is used to select for cells that are stably transfected with the targetting cassette. **B)** Clones containing the targeting cassette are co-transfected with a CRE expressing plasmid and a replacement cassette containing the expanded repeat and its flanking sequences. **C1)** CRE-mediated intra-molecular recombination between Lox sites results in exchange of the targetting cassette with the replacement cassette, whereas **C2)** inter-molecular recombination between Lox sites leads to inversion of the targeting cassette. **D)** Negative selection with gancyclovir selects for cassette exchange. Pro = human cytomegalovirus promoter, pA = simian SV40 polyA sequence. Grey boxes represent repeat flanking sequence, solid black bars represent vector backbone, dashed black bars represent genomic DNA, and horizontal arrows represent transgene orientation.

addition, the inherent instability of different expanded loci could be compared in the same genomic context (Figure 4.19 B).

Targeting cassettes with and without flanking insulator elements were constructed using conventional cloning techniques. Inversion of the *HyTK* gene was observed upon incubation of targeting cassette plasmid DNA with Cre recombinase (data not shown), indicating that the Cre-Lox component of the system was functional. Subsequently, three single-cell derived HeLa clones containing a single copy of the targeting cassette were generated.

Unfortunately, due to time constraints, we were unable to further develop and utilise this novel system.

4.3 Discussion

The characterisation of *cis*-acting modifiers of expanded repeat stability in transgenic animal and cell culture models has been frustrated by site-of-integration effects on repeat stability; the genomic location of a transgenic repeat tract markedly influences its potential to exhibit instability. Here, a transgene targeting system, which allowed for direction of transgenic repeats to the same genomic location and protection against changes in chromatin state of the transgenic locus over time was proposed (Figure 4.19). Such a system would allow for the study of *cis*-acting modifiers of expanded repeat stability independent of site-of-integration effects. It was proposed that by flanking the transgenic repeats with insulator elements, which prevent the propagation of condensed chromatin through a transgene, would maintain the transgenic DNA in an open-chromatin formation and allow for study of the transgenic repeats over long periods of time in culture unaffected by changes in chromatin state. Here, the feasibility of this system was tested by the analysis of both human and mouse cell lines transgenic for expanded repeats either with or without flanking insulator elements.

Analysis of repeat length variation in HeLa cell lines stably transfected with an expanded $(CTG)_{112}$ repeat failed to reveal gross instability in either the presence or absence of flanking insulator elements. As expanded CTG•CAG repeat instability is tissue-specific, it is possible that HeLa cells, which are derived from cervical cancer tissue, do not possess the relative *trans*-acting modifiers, such as a competent mismatch repair (MMR) system, necessary to affect instability. However, unlike many cancer cell lines, HeLa cells do not show widespread microsatellite instability, repair mispairs and hetero-duplexes efficiently, and are commonly used as a repair-proficient controls in studies of the MMR system (Panigrahi *et al.*, 2005; Thomas *et al.*, 1991; Vo *et al.*, 2005).

Interestingly, significant length variation of the transgenic repeat was not observed in mouse *Dmt*-D kidney cells stably transfected with the same transgenes (secondary repeats). Small-pool PCR (SP-PCR) analysis did identify small repeat length variants of the secondary repeat. However, these events

were rare and did not exhibit an expansion bias. The *Dmt*-D cell lines were proficient for expanded repeat instability, as evidenced by the continued expansion of the original *Dmt*-D repeat (primary repeat) in the cell lines created here, suggesting that *cis*-acting factors mediate the dramatic difference in stability observed between the primary and secondary transgenic repeats in each cell line. In addition, the identification of a population of expanded, but stable primary repeats suggested that those *cis*-elements may not be purely sequence based, but may be epigenetic and vary between cells within a given tissue.

Several disease-associated expanded repeats are associated with CpG islands (Brock *et al.*, 1999), and previous studies have found expansion of the *DM1* locus to be associated with changes in the methylation status and chromatin state of its flanking sequences (Cho *et al.*, 2005; Filippova *et al.*, 2001; Steinbach *et al.*, 1998). Methylation-sensitive restriction digest/PCR (MS-PCR) experiments were used to determine the methylation status of both the expanding primary repeat and stable secondary repeat in the stably transfected *Dmt*-D cell lines. The unstable primary repeat was hyper-methylated, whereas the stable secondary repeats were completely unmethylated, suggesting a positive association between CpG methylation of repeat flanking sequence and instability. These data are in agreement with similar analyses of genomic DNA from affected individuals (Steinbach *et al.*, 1998). However, the expanded stable primary repeats identified in five *Dmt*-D cell lines were also hyper-methylated, excluding a simple binary association between presence of flanking CpG methylation and expanded repeat instability. The analyses carried out here were restricted to sequences directly proximal to the repeats and do not rule out differences in methylation pattern at sequences > 500 bp from the repeat tract. Furthermore, none of the assayed CpG dinucleotides are located within the 5' CTCF site or 3' CTCF site located with the flanking sequences of the *Dmt* repeat tract. Methylation of these CTCF sites has been implicated with ablation of CTCF-binding, alteration of local chromatin state, and abnormal antisense transcription of the expanded repeat (Cho *et al.*, 2005; Filippova *et al.*, 2001).

If hyper-methylation of expanded repeat flanking sequences is required to affect instability, the use of neomycin to select for cell lines positive for the transgene may inadvertently select for stable, unmethylated repeats as the *neo* resistance gene will have been silenced in cell lines carrying hyper-methylated transgenes. However, *Dmt*-D cell lines stably transfected with semi-methylated constructs in which the repeat-containing *Dmt* portion of the transgene was methylated, but the neomycin cassette was unmethylated did not reveal gross repeat length variation in culture upon analysis by SP-PCR. These data also argue against a simple association between flanking CpG methylation and instability, but do not exclude a role for wider methylation context in mediating repeat stability. Due to degradation of the transgene during integration, relatively few cell lines containing complete transgenes were generated, limiting the power of this experiment. Moreover, as the relationship between DNA methylation and chromatin state is complex and not yet fully characterised, it is possible that the methylated secondary repeats do not possess the same chromatin structure as the unstable primary repeats.

Data from mouse models of expanded CAG•CTG disorders suggest that repeats that are not transcribed are stable, whereas transcribed repeats can be stable or unstable (Fortune *et al.*, 2000; Mangiarini *et al.*, 1997). However, the nature of the relationship is unclear as expression levels do not correlate with levels of instability (Fortune, 2001; Jung and Bonini, 2007; Mangiarini et al., 1997). It is possible that a common epigenetic feature, such as DNA methylation or chromatin state, facilitates both repeat transcription and instability. Semi-quantitative RT-PCR (sqRT-PCR) analysis revealed that both the stable and unstable primary alleles were expressed at similar levels. In addition, both the insulator positive and insulator negative secondary repeats were expressed in the majority of cell lines, and were expressed at a higher level than the primary repeats in the same cell line. Although these data exclude absence of transcription as the principal cause of the stability observed in the secondary repeats and the stable primary repeats the data do not completely exclude a role for transcription in the repeat stability profiles observed. Previous studies have identified bi-directional transcription at the *DM1* locus in human DM1 cell lines and in Dmt-D mice (Cho *et al.*, 2005; Fortune, 2001). It has been proposed that methylation-mediated ablation of CTCF-binding allows antisense-

transcription through the DM1 repeat tract leading to changes in chromatin state and subsequent destabilisation of the repeat (Cho *et al.*, 2005). It is possible that such bi-directional transcription is absent at the stable primary and secondary repeat loci.

As the Dmt_{112} transgene employed here contains 50 fewer repeats than the original Dmt_{162} transgene used to generate the *Dmt*-D mice, it is possible that the Dmt_{112} repeat is simply too short to exhibit expansion-biased instability. However, a mouse model of Huntington disease transgenic for a 1.9 kb fragment of the *HD* gene containing a $(CAG)_{113}$ expansion exhibited dramatic repeat length variation in kidney at just 10 weeks (Mangiarini *et al.*, 1997). Significantly, mice transgenic for a $(CTG)_{55}$ repeat surrounded by 45 kb of human *DM1* locus DNA exhibited dramatic, expansion-biased repeat length variation in kidney at 16 weeks, with 50 % of alleles having a repeat length differing from the original $(CTG)_{55}$ repeats (Lia *et al.*, 1998). Similarly, a knock-in mouse model of DM1, in which a $(CTG)_{84}$ repeat was inserted into the cognate position of the mouse DMPK gene showed dramatic instability in somatic tissues, particularly kidney, in which an average increase in repeat size of 13 repeats was observed in 6 months (van den Broek *et al.*, 2002). In addition, the isolation of *Dmt*-D cell lines with expanded (~ CTG_{400}) yet stable primary *Dmt* repeat tracts suggest that repeat length alone does not determine stability in the cell lines generated here.

Unlike the original Dmt_{162} transgene used to generate the *Dmt*-D mice, the Dmt_{112} transgene employed here was cloned immediately 3' of a neomycin cassette required for positive selection of clones carrying the transgene. The presence of the neomycin cassette could affect the observed stability of the Dmt_{112} transgene in two ways. First, as outlined above selection for neomycin resistance may inadvertently select for stable transgenes, by secondary selection for areas of open chromatin, methylation status of the transgene, or expression level, all of which may affect repeat stability. Secondly, the sequence of the neomycin cassette itself may contain, as of yet unidentified, repeat-stabilising elements. Indeed, the terminal 500 bp of the *neo* cassette, which directly precedes the Dmt transgene, has a low GC content (39% GC). Given the significant positive correlation between flanking GC content and expanded repeat instability reported previously (Brock *et al.*, 1999)(Chapter 3), the low GC

content of the *neo* cassette may have exerted a stabilising effect on the repeat sequence.

The contrasting instability of the expanding primary Dmt_{162} repeat and stable secondary Dmt_{112} repeat in the same cell lines, suggests the possible involvement of developmental triggers of instability. That is, DNA modifications which occur during development may be required to render an expanded repeat unstable. Indeed, the inhibition of DNA methyltransferases with 5-aza-deoxycytidine (5-aza-CdR) dramatically destabilised expanded $(CTG)_{80}$ and $(CTG)_{150}$ repeats in fibroblast lines from human DM1 patients (Gorbunova *et al.*, 2004). However, the repeat length changes observed were very rare, and occurred in unusually large jumps and no mechanistic explanation of how global demethylation might affect expanded repeat stability was proposed.

In summary, HeLa cells and mouse *Dmt*-D kidney cells failed to exhibit instability in a stably integrated transgene containing a CTG_{112} repeat tract. The *Dmt*-D kidney cell lines were proficient for expanded CTG•CAG instability, as evidenced by the continued instability of an expanded CTG•CAG tract already present in this cell line. The differences in repeat length variation between the unstable and stable transgenic repeats in each cell line did not reflect differences in either DNA methylation state or the transcription level of the transgene. These data further emphasise the importance of identifying the *cis*-acting modifiers of expanded repeat instability.

5. *Cis*-acting modifiers of CAG•CTG microsatellite mutability

5.1 Introduction

Microsatellites are short, tandemly repeated DNA sequences, where the repeating unit is one to six base pairs in length and are thought to comprise between 2 - 4% of the human genome (Leclercq *et al.*, 2007). Eukaryotic genomes are significantly enriched for microsatellite sequences (Bell and Jurka, 1997; Dieringer and Schlotterer, 2003). For example, although the sequence $(CAG)_5$ would be predicted to occur once by chance in the human genome, over 2,000 such sequences are present (Figure 5.1). This overabundance of microsatellite sequences is seen for all microsatellite motifs (Dieringer and Schlotterer, 2003; Ellegren, 2004). Despite decades of exhaustive study, it remains unclear whether this overabundance reflects some unknown function of microsatellites, or is simply a benign consequence of erroneous DNA replication.

Microsatellites are highly polymorphic. Microsatellite variation can occur as either a change in the sequence of the microsatellite (point mutations) or as changes in the length of the microsatellite (expansions or contractions) (Brohede and Ellegren, 1999; Kruglyak et al., 1998). The variability of microsatellite allele lengths, coupled with their ease of detection by PCR, rendered polymorphic microsatellites highly informative markers for use in genome mapping, DNA profiling, phylogenetic analyses and linkage analysis (Hagelberg et al., 1991; Kong et al., 2002; Tamaki and Jeffreys, 2005). However, despite the widespread use of microsatellite based markers, the mutational processes by which microsatellites change in length is not fully understood. It is widely assumed that repeat-length changes occur by replication slippage, followed by failure of mismatch repair (MMR) to correct the misalignment (Chapter 1) (Ellegren, 2004 ; Kelkar et al., 2008; Richards and Sutherland, 1994; Schlotterer and Tautz, 1992). A role for MMR in stabilisation of microsatellite sequences is further supported by the observation that widespread microsatellite instability, common to many human tumours, is usually associated with deficiencies in the MMR system (Woerner *et al.*, 2006). Whereas a role for non-reciprocal strand exchange (gene

conversion) has been implicated as a mutational mechanism of minisatellite sequences (Jeffreys *et al.*, 1994), its role in microsatellite evolution is unclear.

Similar to the disease-associated expanded simple repeats, the mutability of a microsatellite is positively correlated with its number of repeating units (Brandstrom and Ellegren, 2008; Ellegren, 2000; Webster et al., 2002). This appears to be true for all repeat motifs, however for a given number of repeat units different repeat motifs can display very different levels of mutability (Brandstrom and Ellegren, 2008; Ellegren, 2000; Webster et al., 2002). It is widely assumed that the association between the length of a microsatellite and its mutability reflects a greater potential for replication slippage in longer repeat tracts (Ellegren, 2004). However, the observation that the distribution of microsatellite repeat lengths in a given genome is stationary and that microsatellites appear to have an upper length limit suggest that other mechanisms are modifying microsatellite mutability (Ellegren, 2000; Li *et al.*, 2002). It is has been suggested that longer repeats may show a greater tendency to contract than shorter microsatellites resulting in a stable distribution of microsatellite allele lengths (Ellegren, 2000). In addition, mathematical modelling of microsatellite evolution has suggested that an equilibrium between replication-mediated expansion of microsatellites and microsatellite decay/shortening by accumulation of point mutations, best explains the microsatellite distributions observed in eukaryotic genomes (Ellegren, 2004; Kruglyak *et al.*, 1998).

As for the expanded trinucleotide repeat loci, the observation that microsatellites of similar length and motif can show significantly different levels of mutability (Ellegren, 2004), suggested the presence of *cis*-acting modifiers of microsatellite mutation rate. Tumors of hereditary non-polyposis colorectal cancer (HNPCC), deficient for various components of the mismatch repair machinery, show genome-wide microsatellite instability (MSI) (de la Chapelle and Peltomaki, 1995; Dietmaier et al., 1997). However, despite the obvious *trans*-factor determination of HNPCC MSI, the observation that loci similar in sequence, location and length can display markedly different levels of instability suggested a contribution of flanking sequence to misalignment mediated repeat

instability (Dietmaier *et al.*, 1997; Richards *et al.*, 1996). As of yet, no *cis*-acting modifiers of microsatellite mutability have been identified.

A negative correlation between flanking GC content and allelic diversity of microsatellites was reported in alligators (Glenn *et al.*, 1996). However, only 14 di-nucleotide loci were analysed in that study and a subsequent study in *Drosophila* failed to find an association between mutability and flanking GC content (Bachtrog *et al.*, 2000). Moreover, a genome-wide study of human and chimpanzee microsatellite evolution, found no correlation between isochore type and microsatellite mutability (Kelkar *et al.*, 2008). However, as of yet, no study has examined the relationship between flanking GC content and microsatellite mutability for a specific repeat type on a genome-wide dataset.

We hypothesised that, although the mechanisms underlying microsatellite length variation and dynamic mutation of expanded repeats are likely to be different, flanking GC content may be a modifier of both processes. A significant positive association between flanking GC content and instability has been reported for the expanded repeat disorders (Brock *et al.*, 1999) (Chapter 3). As dynamic mutation of disease-associated expanded repeats requires a functional MMR system we propose that a higher flanking GC may act to recruit or promote the activity of components of the MMR acting on the expanded repeat. In contrast, as microsatellite instability is caused by failure to repair misalignments due to a defective MMR system, loci with a lower flanking GC content may recruit or promote the activity of the MMR machinery less effectively than microsatellites flanked by sequences with a high GC content. If this assumption were correct, a negative correlation between microsatellite mutability and flanking GC content would be predicted.

Here, in an attempt to identify *cis*-acting modifiers of CAG•CTG microsatellite variability, the DNA sequence flanking CAG•CTG microsatellites in the human and chimpanzee genomes was analyzed, and sequence characteristics correlated with microsatellite mutability.

5.2 Results

5.2.1 Definition and identification of all CAG•CTG microsatellites in the human genome

Despite the inherent simplicity of short tandem repeats, their definition and subsequent identification in genomic sequences is complex. Consequently, marked differences in how microsatellite sequences are defined exist between published studies of microsatellite mutability, and such differences are likely to have had a major effect on the observed experimental outcomes (Leclercq *et al.*, 2007). Most notably, the minimum length (repeat number) of a microsatellite and the degree of imperfection (number of interruptions) permitted in a microsatellite varies between studies. Analyses of microsatellite mutability have generally excluded short (< 4 repeats) microsatellite sequences on the grounds that tandem repeat length mutations do not occur at very short microsatellites (Brandstrom and Ellegren, 2008; Kelkar et al., 2008; Webster et al., 2002). However, little evidence exists to support this assertion, and that which does is based on small samples (Rose and Falush, 1998). Moreover, recent studies have reported apparent misalignment events at microsatellites consisting of as few as two repeats (Brandstrom and Ellegren, 2007). Thus, it was decided to include microsatellites of all lengths (≥ 2 repeat units) in the analyses presented here. Imperfect (interrupted) repeats are difficult to define. Although some evidence suggests that microsatellite evolution may be influenced by the presence of directly proximal microsatellite sequences (Almeida and Penha-Goncalves, 2004), no objective classification system exists to differentiate between a true interrupted microsatellite and two individual, but proximal microsatellites. Moreover, existing repeat detection algorithms vary greatly in their ability to identify interrupted repeat microsatellite sequences (Kelkar *et al.*, 2008; Leclercq *et al.*, 2007). Thus, it was decided to exclude imperfect microsatellites from these analyses.

All triplet motif microsatellite sequences in the human genome (assembly: NCBI 36, October 2005) were identified using the Tandem Repeat Finder (TRF) program (Benson, 1999). To allow detection of short (< 4 repeat units) microsatellites, the minimum alignment score parameter of the TRF program

was reduced from the default setting of 50 to 10 (TRF parameter settings: Match = 2, MisMatch = 7, In-Del = 7, pM = 75, pI = 20, MinScore = 10, MaxPeriod = 3) (Benson, 1999; Leclercq *et al.*, 2007). CAG•CTG repeats belong to the AGC-motif class of microsatellites, which consists of all microsatellites with repeating units of the sequence AGC in all three reading frames and their reverse complements (ACG, CGT, CAG, CTG, CGA, TCG). All non-AGC microsatellites were removed. Microsatellites were allowed to vary in length by fractions of the repeating unit. For example, the microsatellite 'CAG CAG CAG CAG CAG C' was defined as having a length of 5.3 repeats; the terminal cytosine nucleotide being included in the sequence of the microsatellite, not the flanking sequence. It was rationalised that such partial repeat units should be included in the microsatellite sequence as they are equally likely to be involved in the formation of misalignment-promoting non B-DNA structures as nucleotides internal to the microsatellite.

Finally, microsatellites less than 10 bp apart were removed as repeats in close proximity may not evolve independently (Almeida and Penha-Goncalves, 2004; Kelkar et al., 2008) and the presence of repetitive sequences immediately proximal to microsatellites will render identification of truly orthologous sequences difficult (Webster *et al.*, 2002). Employing this approach 791,649 perfect AGC-motif microsatellite sequences were identified in the human genome (Figure 5.1). As the human genome sequence is single copy, the identified microsatellites represent one allele at each locus. As expected, the number of microsatellite loci decreased with increasing repeat length, the majority (99.4%) of loci consisting of short (≤ 4 repeats) microsatellites.

5.2.2 CAG•CTG microsatellite length and flanking GC content in the human genome

Inappropriate mismatch repair (MMR) by a competent MMR system has been suggested as a major modifier of expanded repeat instability (Gomes-Pereira *et al.*, 2004), whereas mutations in MMR genes result in widespread microsatellite instability in many cancers (Woerner *et al.*, 2006). As expanded trinucleotide instability is positively correlated with flanking DNA GC content, it has been suggested that flanking DNA GC content or other *cis*-elements may modify

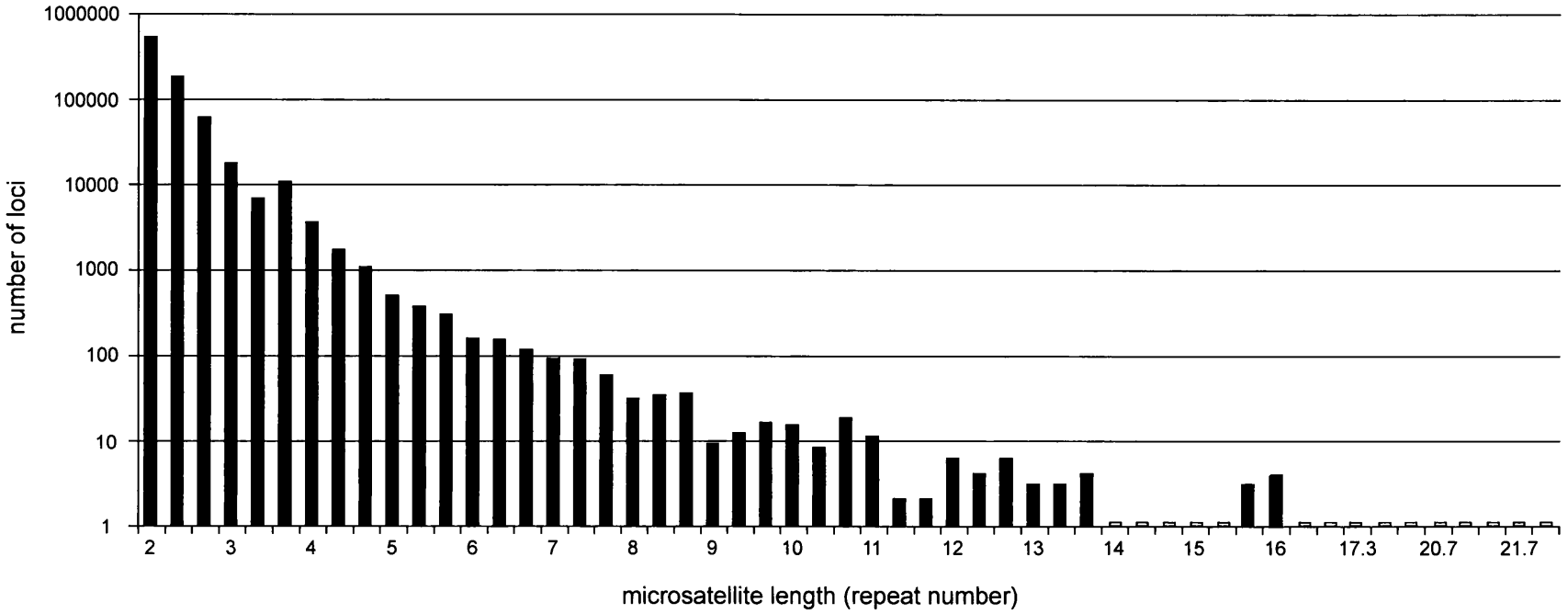


Figure 5.1. AGC-motif microsatellite length distribution in the human genome. Small empty bars represent a single repeat ($N = 1$).

expandability through effects on MMR (Chapter 3) (Brock *et al.*, 1999). Thus, it was hypothesised that flanking GC content may also modify microsatellite mutability through effects on MMR. As the microsatellite sequences identified in the human genome represent a single snapshot in a continuum of mutation events, it is not possible to measure the actual mutation rates of individual loci from these data. Whereas short microsatellites may have arisen from chance point mutations, longer microsatellites are more likely to have arisen from misalignment mutation events; microsatellite length can be employed as an estimator of the true mutation rate. Thus, the relationship between microsatellite length (repeat number) and flanking GC content in the human genome was investigated.

Analysing all identified AGC-motif microsatellite loci (N = 791,649), a highly significant rank correlation was found between flanking GC content and microsatellite length (Table 5.1). This correlation was greatest for sequences proximal to the repeat (≤ 1000 bp). To facilitate visualisation of the data, the mean GC content for each repeat length was determined and plotted against repeat length (Figure 5.2). Visual inspection of the data suggested that whereas the length of 'short' (≤ 7 repeats) microsatellites increased with increasing flanking GC content, no such relationship existed for 'long' (> 7 repeats) microsatellites (Figure 5.2). Indeed, analysis of short microsatellites found a significant rank correlation between microsatellite length and both flanking GC content and mean flanking GC content for each repeat length (Table 5.1 & Figure 5.3A). Conversely, no such correlation was found between long microsatellites and either flanking GC content or mean flanking GC content for each repeat length (Table 5.1 & Figure 5.3B).

As repeat length mutations in microsatellites associated with coding and non-coding DNA are most likely subject to different selective pressure, the dataset was divided into exonic and non-exonic microsatellites. Using all predicted and manually annotated exon coordinates from the Ensembl genome server (Dec, 2006), microsatellites were classified as exonic if all or part of the microsatellite sequence was located within an Ensembl exon. Approximately 6% (43,700) of all AGC microsatellites were identified as exonic. Exonic ACG-motif microsatellites (mean repeat number = 2.31 repeats) were significantly longer than non-exonic

Table 5.1. Rank correlation (Spearman's *rho*) of flanking GC content with AGC microsatellite length

		10 bp flanking sequence		50 bp flanking sequence		100 bp flanking sequence		1000bp flanking sequence		10000bp flanking sequence	
	N	*rho*	*P*	*rho*	*P*	*rho*	*P*	*rho*	*P*	*rho*	*P*
All AGC	791,649	0.06	$< 1x10^{-6}$	0.08	$< 1x10^{-6}$	0.09	$< 1x10^{-6}$	0.08	$< 1x10^{-6}$	0.05	$< 1x10^{-6}$
≤ 7 repeats	791,266	0.08	$< 1x10^{-6}$	0.1	$< 1x10^{-6}$	0.1	$< 1x10^{-6}$	0.09	$< 1x10^{-6}$	0.06	$< 1x10^{-6}$
> 7 repeats	383	0.11	0.038	0.06	0.2	0.08	0.1	0.08	0.1	0.04	0.5

Table 5.2. Rank correlation (Spearman's *rho*) of flanking GC content with exonic AGC microsatellite length

		10 bp flanking sequence		50 bp flanking sequence		100 bp flanking sequence		1000bp flanking sequence		10000bp flanking sequence	
	N	*rho*	*P*	*rho*	*P*	*rho*	*P*	*rho*	*P*	*rho*	*P*
All exonic	43,700	0.14	$< 1x10^{-6}$	0.16	$< 1x10^{-6}$	0.14	$< 1x10^{-6}$	0.12	$< 1x10^{-6}$	0.07	$< 1x10^{-6}$
≤ 7 repeats	43,624	0.14	$< 1x10^{-6}$	0.16	$< 1x10^{-6}$	0.14	$< 1x10^{-6}$	0.12	$< 1x10^{-6}$	0.07	$< 1x10^{-6}$
> 7 repeats	76	-0.02	0.86	-0.03	0.80	-0.02	0.87	0.10	0.40	0.10	0.4

Table 5.3. Rank correlation (Spearman's *rho*) of flanking GC content with non-exonic AGC microsatellite length

		10 bp flanking sequence		50 bp flanking sequence		100 bp flanking sequence		1000bp flanking sequence		10000bp flanking sequence	
	N	*rho*	*P*	*rho*	*P*	*rho*	*P*	*rho*	*P*	*rho*	*P*
All non-exonic	747,949	0.06	$< 1x10^{-6}$	0.08	$< 1x10^{-6}$	0.09	$< 1x10^{-6}$	0.06	$< 1x10^{-6}$	0.05	$< 1x10^{-6}$
≤ 7 repeats	747,642	0.06	$< 1x10^{-6}$	0.08	$< 1x10^{-6}$	0.09	$< 1x10^{-6}$	0.08	$< 1x10^{-6}$	0.05	$< 1x10^{-6}$
> 7 repeats	307	0.09	0.11	0.02	0.67	0.05	0.37	0.04	0.4	0.05	0.4

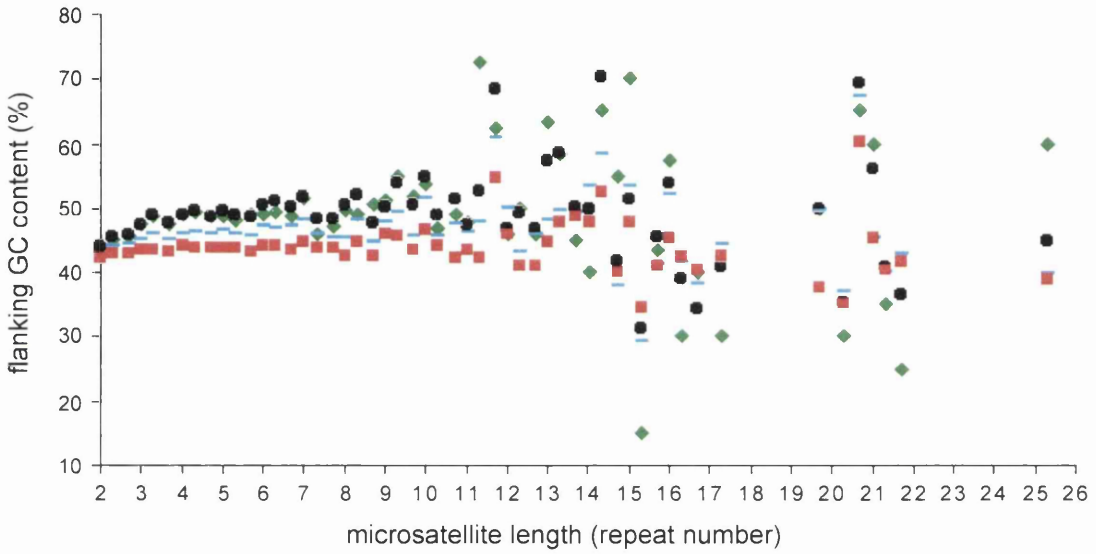
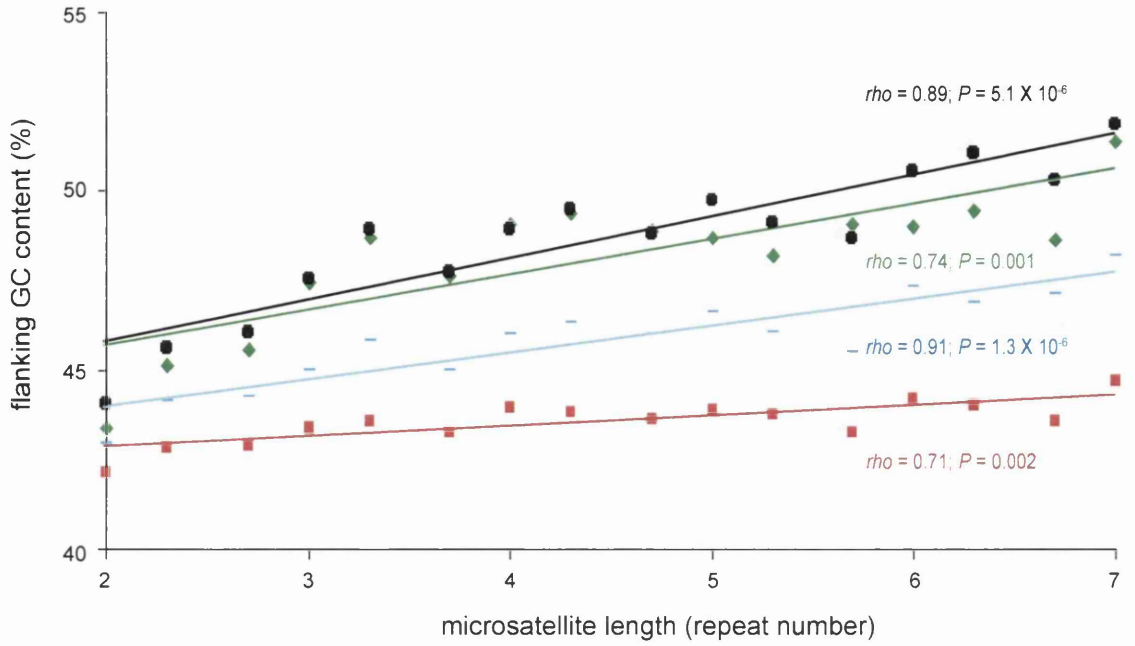


Figure 5.2 The relationship between AGC-motif microsatellite length and flanking GC content. Each data point represents the mean flanking sequence GC content (%) of a given microsatellite length class, grouped by repeat number. The GC content for 10 bp (green diamond), 100 bp (black circle), 1000 bp (blue line) and 10,000 bp (red square) of flanking sequence is shown.

A) Short microsatellites (≤ 7 repeats)



B) Long microsatellites (> 7 repeats)

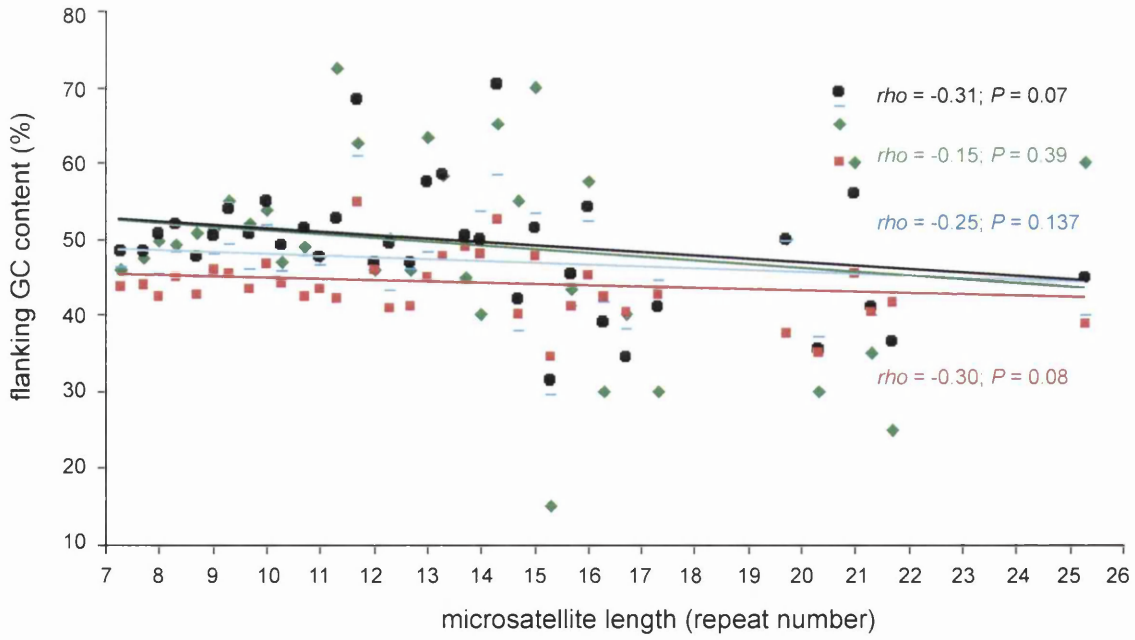


Figure 5.3 The relationship between AGC-motif microsatellite length and flanking GC content. The plots illustrate the relationship between microsatellite length and flanking GC content for **A**) microsatellites with seven or fewer repeat units and, **B**) microsatellites with greater than seven repeat units. Each data point represents the mean flanking sequence GC content (%) of a given microsatellite length class, grouped by repeat number. The GC content for 10 bp (green diamond), 100 bp (black circle), 1000 bp (blue line) and 10,000 bp (red square) of flanking sequence is shown. The rank correlation coefficient (Spearman's *rho*) of mean flanking GC content with microsatellite length, and corresponding trend line are also shown

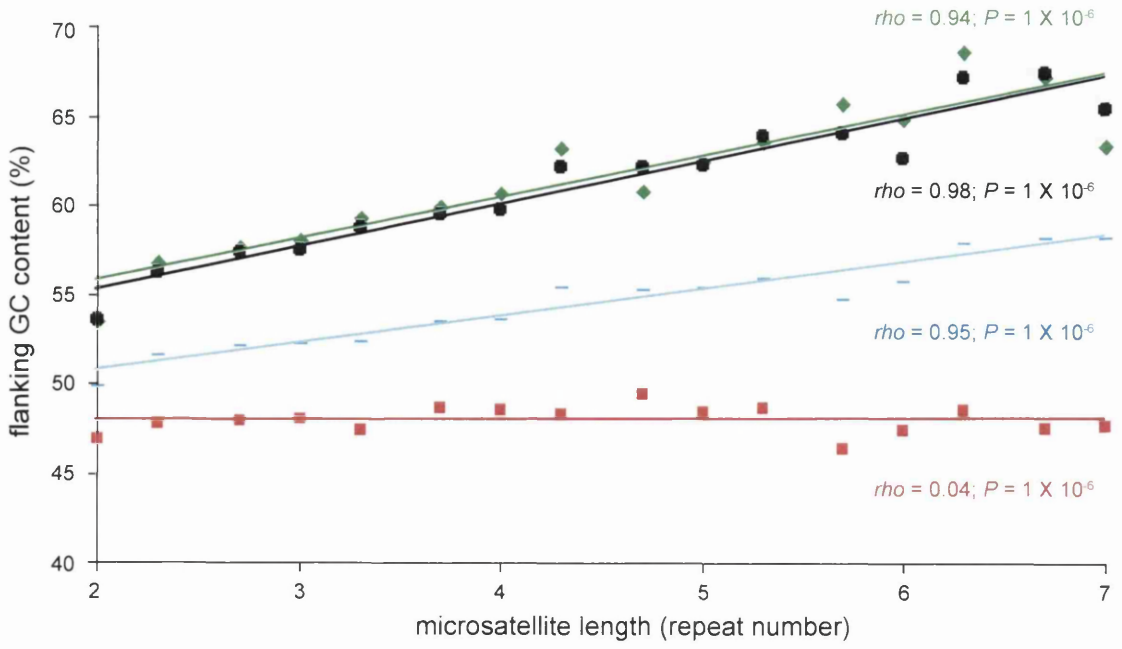
microsatellites (mean repeat number = 2.2 repeats) (Mann-Whitney U = 21284239, $P < 6 \times 10^{-5}$), possibly reflecting the high frequency of long polyglutamine-encoding CAG repeats in human genes. In addition, the relationship between repeat length and flanking GC content was stronger for exonic repeats than non-exonic repeats (Tables 5.2, 5.3 and Figure 5.4). Although the increased length of exonic repeats is most likely due to the higher GC content of coding regions, this does not readily explain the highly significant correlation of exonic microsatellite length with flanking GC content.

5.2.3 *Cis*-acting modifiers of misalignment microsatellite mutability

Microsatellite mutability increases with repeat number (Brandstrom and Ellegren, 2008; Dieringer and Schlotterer, 2003; Kelkar et al., 2008; Rose and Falush, 1998). However, the repeat length of a given microsatellite is a crude estimator of the mutational process which delivered a locus to its present length. That is, a given microsatellite may have arrived at its present length by the addition or deletion of whole repeat units as a result of misalignment mutation, or by base substitution in the microsatellite or its flanking sequences. As the MMR system is implicated in both instability of expanded disease-associated CAG•CTG repeats and both base substitution and misalignment mutations at polymorphic microsatellite loci, we sought to identify AGC microsatellite loci at which misalignment events had occurred. Once identified, the presence of *cis*-elements common to both type of loci could be investigated. Microsatellite loci at which misalignment mutation events have occurred can be detected by comparing the repeat number at orthologous microsatellite loci in closely related species, such as the human and chimpanzee (Figure 5.5A).

To identify orthologous microsatellites in humans and chimpanzees, the 5' and 3' flanking sequences (100 bp) of all human AGC-motif microsatellites were aligned to the chimpanzee genome (assembly: CHIMP2.1, Mar 2006) using the BLASTn program (version blastall 2.2.18) (Figure 5.5 B). All BLAST searches were carried out on a computing cluster of 30 linux servers. In order to identify truly orthologous, perfect ACG-motif microsatellite pairs, at which misalignment

A) Short (≤ 7 repeats) exonic microsatellites



B) Short (≤ 7 repeats) non-exonic microsatellites

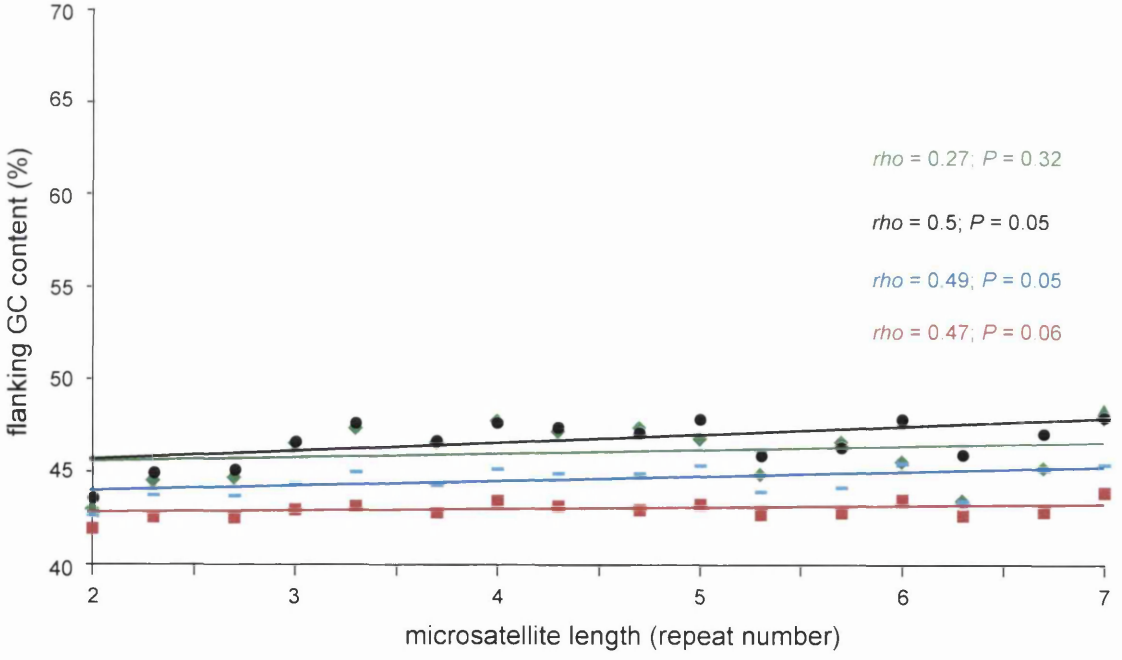
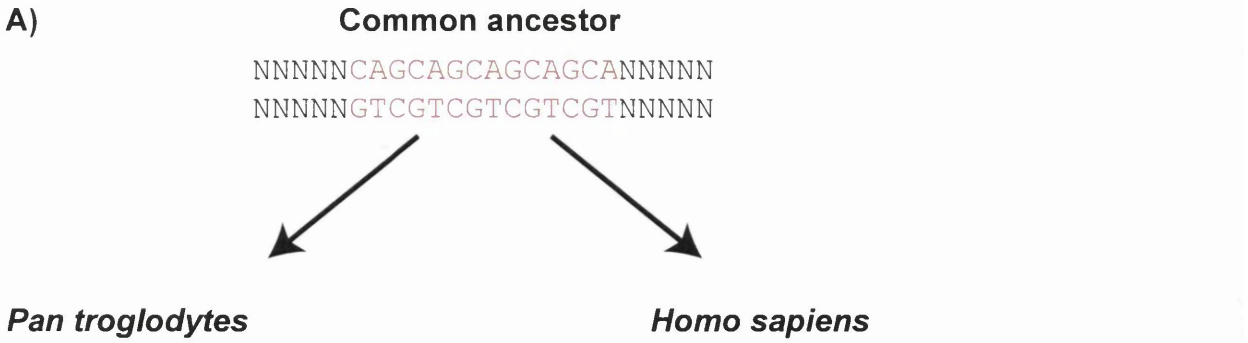


Figure 5.4 The relationship between AGC-motif microsatellite length and flanking GC content. The plots illustrate the relationship between microsatellite length and flanking GC content for **A**) microsatellites located within exons and, **B**) microsatellites located outside exons. Each data point represents the mean flanking sequence GC content (%) of a given microsatellite length class, grouped by repeat number. The GC content for 10 bp (green diamond), 100 bp (black circle), 1000 bp (blue line) and 10,000 bp (red square) of flanking sequence is shown. The rank correlation coefficient (Spearman's *rho*) of mean flanking GC content with microsatellite length, and corresponding trend line are also shown.



(i) No repeat number change in either species

NNNNNCAGCAGCAGCAGCANNNNN	NNNNNCAGCAGCAGCAGCANNNNN
NNNNNGTCGTCGTCGTCGTNNNNN	NNNNNGTCGTCGTCGTCGTNNNNN

(ii) Detectable repeat number change in one or both species

NNNNNCAGCAGCANNNNN	NNNNNCAGCAGCAGCAGCAGCANNNNN
NNNNNGTCGTCGTNNNNN	NNNNNGTCGTCGTCGTCGTCGTNNNNN

(iii) Undetectable repeat number change in both species

NNNNNCAGCAGCANNNNN	NNNNNCAGCAGCANNNNN
NNNNNGTCGTCGTNNNNN	NNNNNGTCGTCGTNNNNN

B)

Homo sapiens GTATCGATCGTCGATCAGCAGCAGCAGTTGACAGCTGGACTG

BLAST unique human flanking sequences against the chimpanzee genome

(i) Repeat number change by mis-alignment repair mutation within the repeat tract

```
Homo sapiens      GTATCGATCGTCGATCAGCAG---CAGCAGTTGACAGCTGGACTG
                  |||||..|||||.|||||                |||||||||.||||||
Pan troglodytes   GTATTTATCGTCGATCAGCAGCAGCAGCAGTTGACAGCTCGACTG
```

(ii) Repeat number change by point muation of flanking sequence

```
Homo sapiens      GTATCGATCGTCGATCAGCAGCAGCAGTTGACAGCTGGACTG
                  ||||.|||||||||||                ..||||||||||||||
Pan troglodytes   GTATAGATCGTCGATCAGCAGCAGCAGCAGACAGCTGGACTG
```

Figure 5.5 Identification of orthologous microsatellite loci in the human and chimpanzee at which mis-alignment repair mutation events have occurred. A) Variation of microsatellite repeat number can be identified by comparing repeat number at orthologous human-chimpanzee loci. **B)** By aligning the unique sequences flanking a human microsatellite to their orthologous chimpanzee sequences, (i) changes in repeat number resulting from the addition or deletion of whole repeat units can be distinguished from (ii) changes in repeat length resulting from point mutations in the flanking sequence. See text for a detailed description of the algorithm used.

mutation events had occurred a series of filters were applied the output obtained from the BLAST searches:

1) BLAST hits with an alignment length less than 100 bp or a percentage identity less than 95% were removed
2) Microsatellites mapping to non-homologous chimpanzee chromosomes were removed
3) Microsatellites mapping to more than one location on a homologous chimpanzee chromosome were removed
4) All loci for which the orthologous chimpanzee microsatellite was imperfect (interrupted) were removed
5) Finally, all loci which did not differ in length by whole repeat units (i.e. multiples of 3 bp) were removed

A total of 209,818 orthologous, perfect, AGC-motif microsatellite pairs were identified representing approximately 25 % of all human AGC microsatellites. In order to ascertain whether the set of orthologous microsatellites generated was a representative sample of all human AGC microsatellites, and free of significant sampling bias, the entire human AGC data set and the orthologous AGC dataset were compared. Repeat size did not significantly differ between the datasets ($N_{ALL\ HUMAN}$ = 791,649, $N_{ORTHOLOGOUS}$ = 209,818, Mann-Whitney U = 49618853, P = 0.296). In addition, the frequency distribution of repeat lengths for all human AGC-motif microsatellites and orthologous microsatellites did not differ significantly (Kolmogorov-Smirnov Z = 0.824, P = 0.51) (Figure 5.6).

As coding regions of the human and chimpanzee genomes are more highly conserved than intergenic regions, the flanking sequences of exonic microsatellites are more likely to be highly conserved and consequently more readily identified by the sequence comparison-based detection algorithm employed here, resulting in slight enrichment for exonic orthologous microsatellites. Indeed, the orthologous dataset was significantly enriched for exonic microsatellites (exonic loci/ total loci = 0.075) compared to the entire human AGC dataset (exonic loci/ total loci = 0.055) (χ^2 = 1142; $P << 0.00001$).

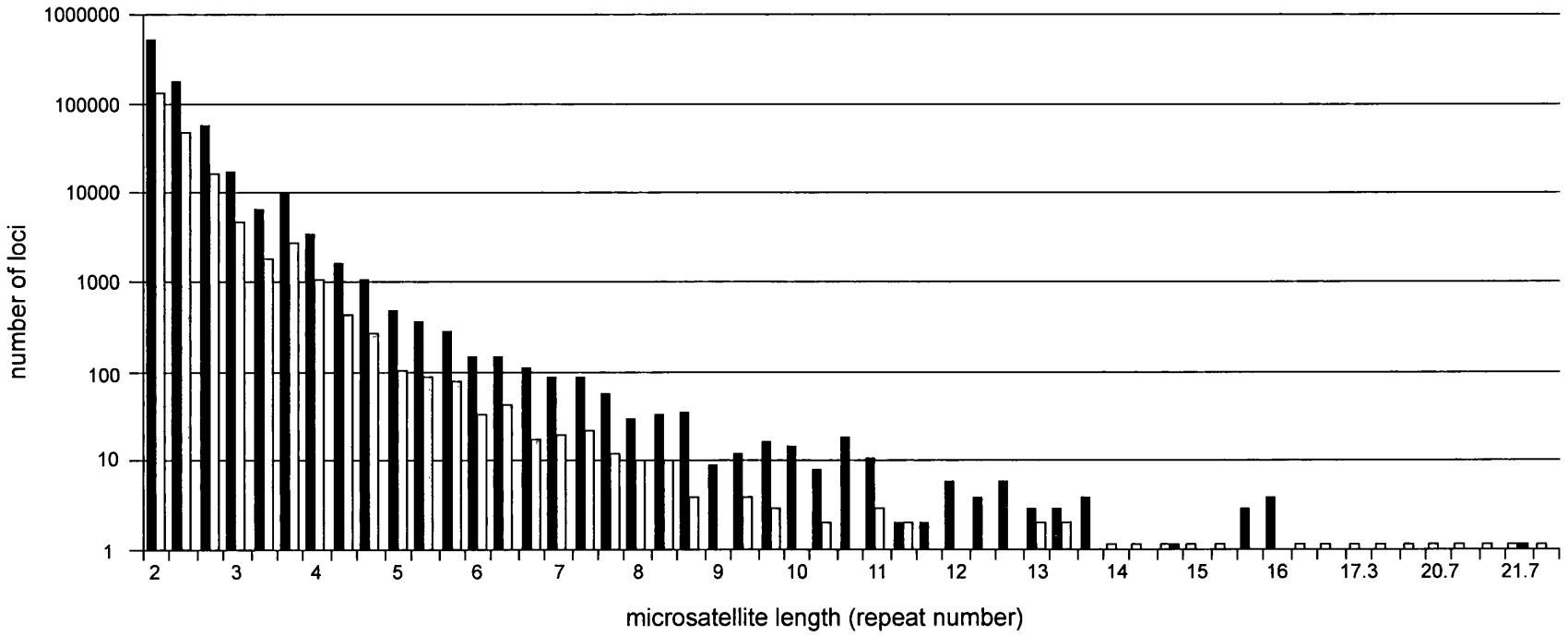
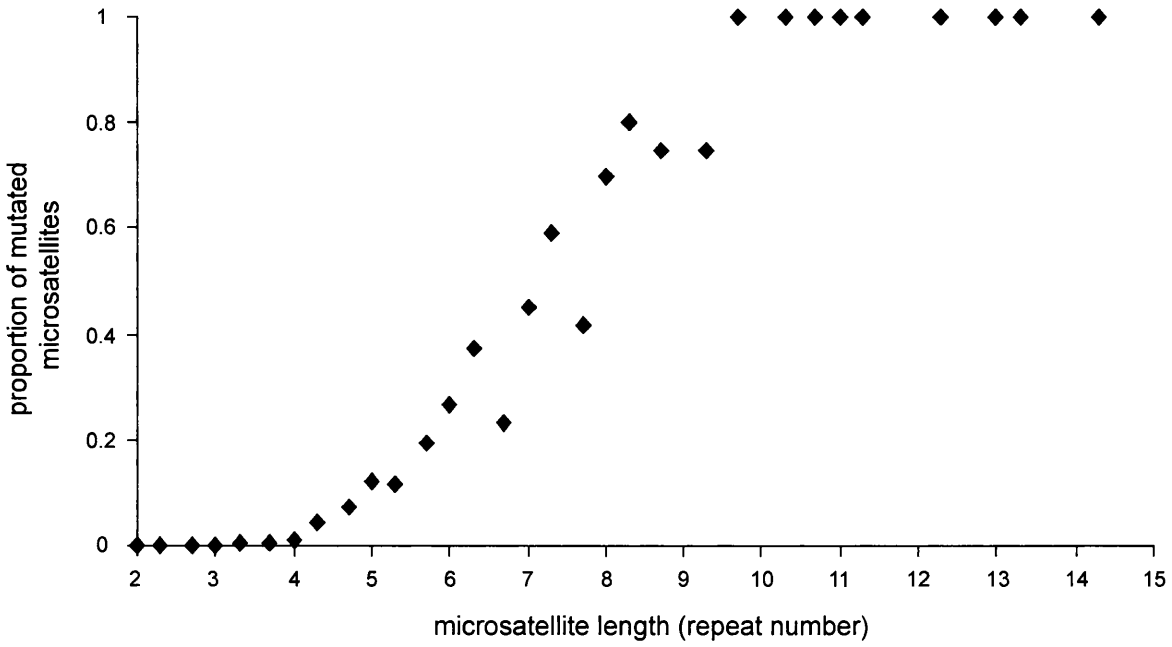


Figure 5.6. Comparison of length distributions of all human AGC-motif microsatellites and all human AGC-motif microsatellites with a detectable chimp orthologue. All human AGC microsatellite length distribution is shown in black. All orthologous human AGC microsatellite length distribution is shown in grey. Small empty bars represent a single repeat ($N = 1$).

Of the 209,818 orthologous loci identified, 324 loci differed by at least one repeat unit (approximately 0.2 % of all orthologous loci). Mutated loci were significantly longer than unchanged loci ($N_{UNCHANGED}$ = 209,494, $N_{MUTATED}$ = 324, Mann-Whitney U = 42010, $P < 0.00006$), in agreement with previous observations that misalignment events occur more frequently at microsatellite loci consisting of four or more repeats. However, although the proportion of mutated loci increased with increasing repeat number (Figure 5.7A), 43% of all mutated microsatellites were observed at loci with fewer than four repeats. Interestingly, the proportion of mutated non-integer microsatellites (e.g. 2.3 repeats, 2.7 repeats) was higher than the proportion of mutated integer microsatellites (e.g. 2 repeats) for each repeat class (Figure 5.7B). This is consistent with our prediction that misalignment events can occur in any reading frame of a microsatellite sequence, and consequently partial repeat sequences at the ends of microsatellites may be involved in misalignment events. Thus, the sequence '**CAG CAG C**' should be defined as having 2.3 CAG repeat units as misalignment may occur in the first two CAG units or last two ACG units (in bold). The proportion of exonic microsatellites at mutated orthologous loci was not significantly different from the proportion of exonic microsatellites found at unchanged orthologous loci (χ^2 = 1.014; P = 0.314), suggesting that genic location is not a major modifier of mutability.

Analysis of the human sequences flanking the orthologous microsatellite loci revealed that mutated loci had a significantly higher flanking GC content than unchanged loci (Table 5.4). This finding suggests a role for flanking DNA GC content in microsatellite misalignment mutation. However, the flanking GC content of mutated and unchanged orthologous loci with the same repeat number did not differ significantly (data not shown). Moreover, analysis of loci which differed in length between humans and chimpanzees failed to find a significant correlation between flanking GC content and the magnitude of inter-species microsatellite length change (Figure 5.8). This finding suggests, that once mutable, flanking sequence GC content may not be a modifier of mutability.

A) Longer microsatellites are more mutable



B) Mutability increases with non-integer repeat units

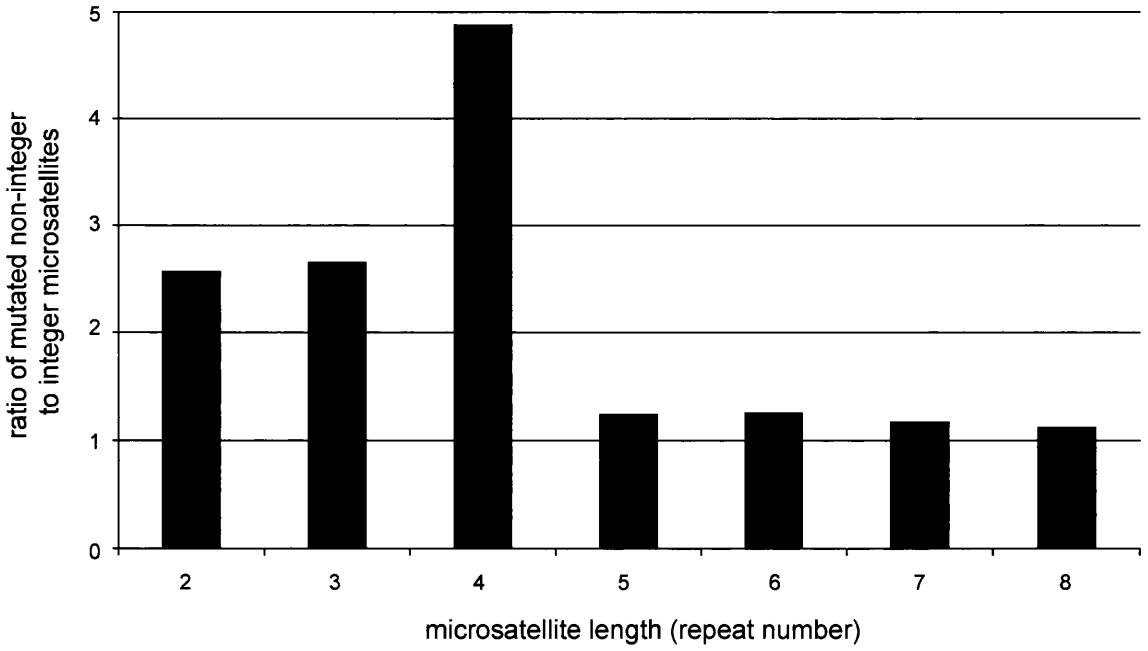
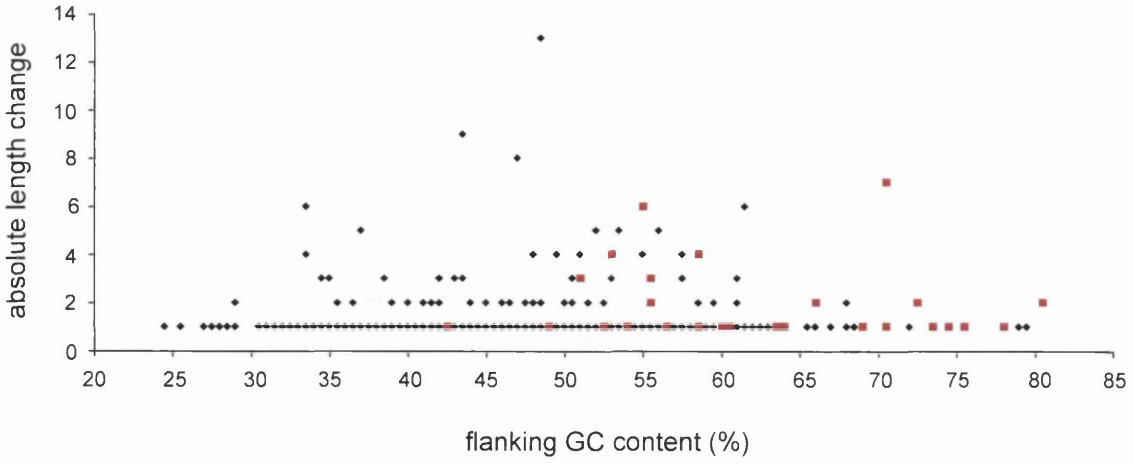


Figure 5.7 The frequency of length changes between orthologous human-chimpanzee microsatellites is dependent on microsatellite length (repeat number). A) The proportion of mutated orthologous microsatellites increases with microsatellite repeat number **B**) The ratio of mutated non-integer microsatellites to integer microsatellites (e.g proportion of mutated 2.3 + 2.7 microsatellites/proportion of mutated 2.0 microsatellites) for short repeats.

A)



B)

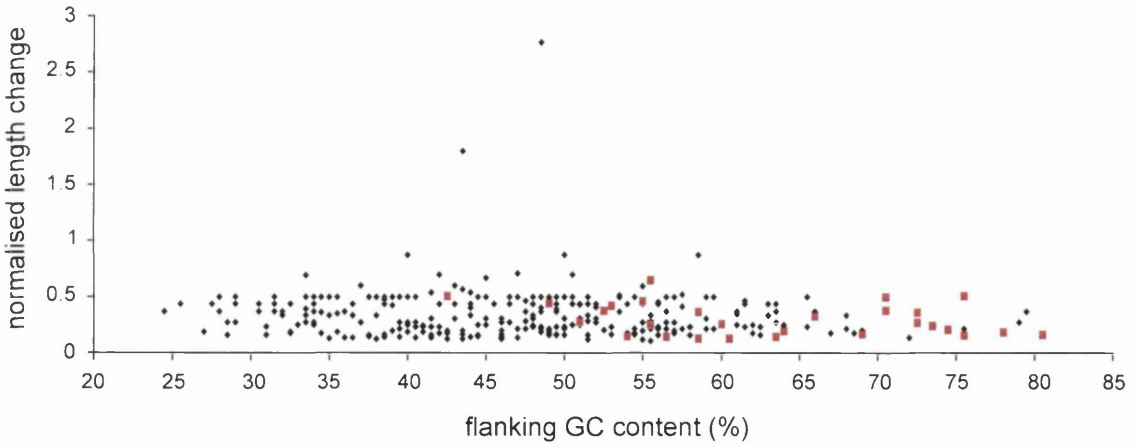


Figure 5.8. The magnitude of length changes observed between orthologous chimpanzee -human microsatellites is not correlated with flanking sequence GC content. A) Absolute microsatellite length change (|human repeat number - chimp repeat number|), and **B**) normalised microsatellite length change (absolute length change/ human repeat number) do not correlate with the GC content of the microsatellite flanking sequences (100 bp) for all (black diamonds) or exonic (red squares) orthologous microsatellites.

Table 5.4. Mutated orthologous loci have higher flanking GC content than unchanged orthologous loci.

	GC content (100 bp flanking sequence)	GC content (1,000 bp flanking sequence)	GC content (10,000 bp flanking sequence)	N
Unchanged	45.67	44.36	43.47	209,494
Mutated	48.25	45.61	43.72	324
MW U	29,562,768.5	30,529,860	32,700,176	
P	<0.000006	0.00194	0.271	

By necessity, human microsatellite sequences were defined as having a minimum of two repeating units. However, many orthologous chimpanzee loci were found to have one 'repeat' unit. As human microsatellites were used to locate their chimpanzee orthologues, but not vice versa, similar one unit 'microsatellites' would not be detected in the human genome using the methodology employed here. In order to investigate the nature of this ascertainment bias, and determine its effect on our analysis, the distribution of repeat length changes observed for human microsatellites of various lengths was determined (Figure 5.9). As predicted, chimpanzee loci were significantly shorter than their human orthologues for short microsatellites (Figure 5.9A). A more balanced normal distribution of repeat length differences was observed for medium length repeats (Figure 5.9B), suggesting that sampling bias resulting from failure to identify one 'repeat' human orthologues of chimp microsatellites with four or more repeats did not affect this dataset. However, long (≥ 7 repeats) microsatellites tended to be typically shorter in chimpanzee than human (Figure 5.9C). Although microsatellite mutability increases with repeat number, arbitrarily long microsatellites are not observed in eukaryotic genomes. It has been suggested that this microsatellite length limit reflects a tendency for longer repeats to undergo more contractions than expansions (Ellegren, 2000). Thus, as has been well documented, when a set of long microsatellites from one species are compared with a sister species, the loci in the sister species are typically shorter (Brandstrom and Ellegren, 2008). Thus, the apparent mutability of short and long microsatellite sequences is affected by two different ascertainment biases. However, exclusion of short and/or long microsatellites from our analysis failed to reveal a significant correlation between the

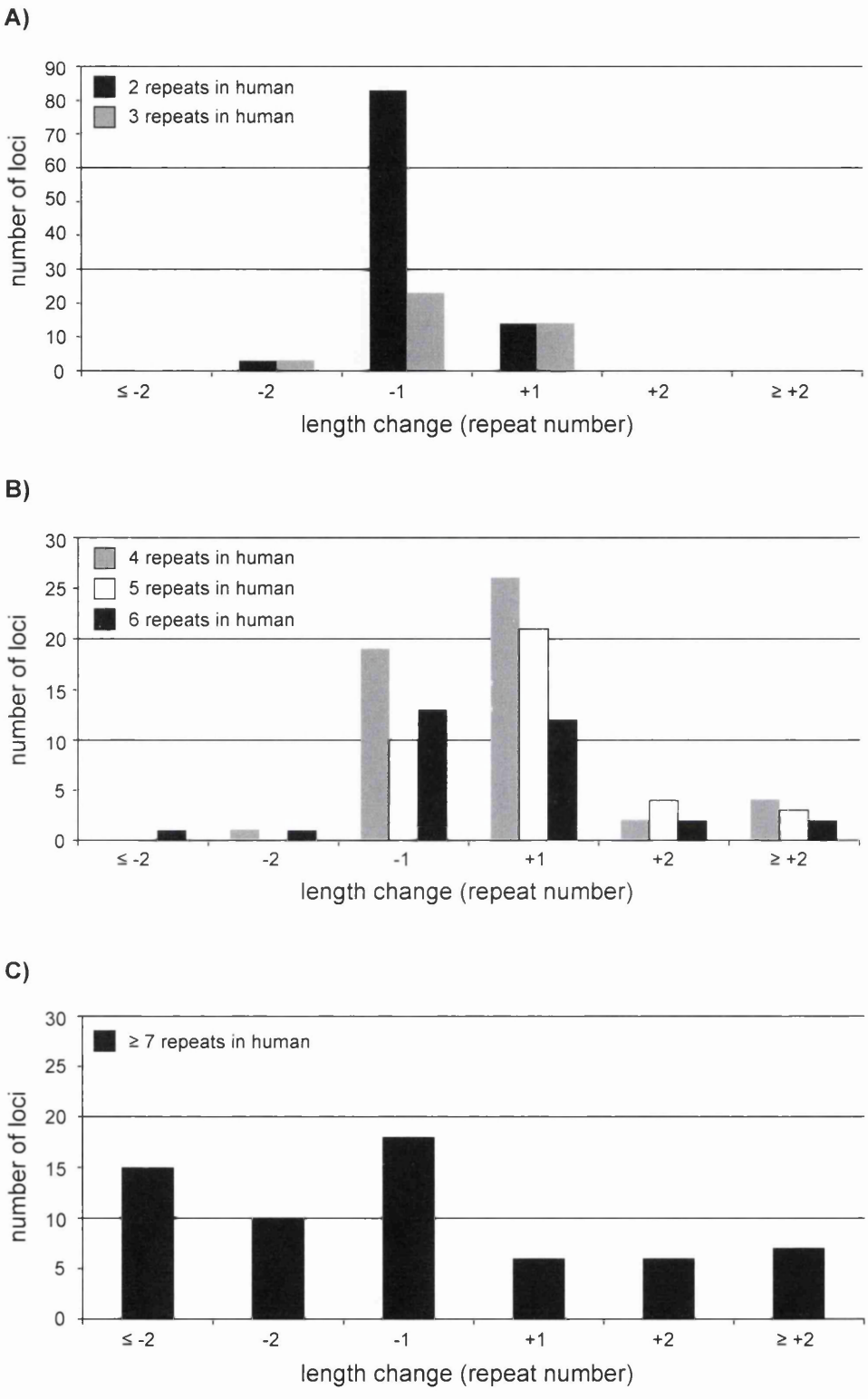


Figure 5.9. Distribution of length changes observed at orthologous microsatellite loci is dependent on human microsatellite length. Shown are the length differences (chimp repeat number - human repeat number) observed at mutated orthologous microsatellites. Distributions for A) short, B) medium length, and C) long microsatellites are shown. Repeats were grouped by integer repeat length (e.g. 2 repeats = 2.0, 2.3, and 2.7 repeat loci) to increase sample size.

magnitude of length change at mutated othologous loci and flanking GC content (data not shown).

5.2.4 Flanking GC content and CAG•CTG heterozygozity

Genome-wide, inter-species studies of orthologous microsatellite mutability allow for the analysis of large numbers of loci simultaneously. However, for any particular locus, only two alleles, one in each species is observed. As a consequence, such studies will tend to underestimate the true variability of loci. Whereas conventional genotyping analyses of microsatellite heterozygozity within a population are typically restricted to small sample sets and longer alleles, they provide more accurate measures of the true variability of individual microsatellite loci. Thus, we performed a meta-analysis of published data of exonic CAG•CTG microsatellite heterozygozity.

The repeat length of 51 exonic CAG•CTG microsatellites with a minimum repeat number of six was reported for 160 individuals in a Polish population (Rozanska *et al.*, 2007). Here, consistent with our genome-wide analyses, no significant correlation between the reported heterozygozity and flanking GC content of the loci was found (Figure 5.10).

Moreover, the heterozygozity of normal length range polyglutamine loci shows a significant negative correlation with the instability of the repeat when expanded (Figure 5.11). This observation suggests that the mechanism underlying expanded repeat instability differs from that underlying the normal mutability of these loci.

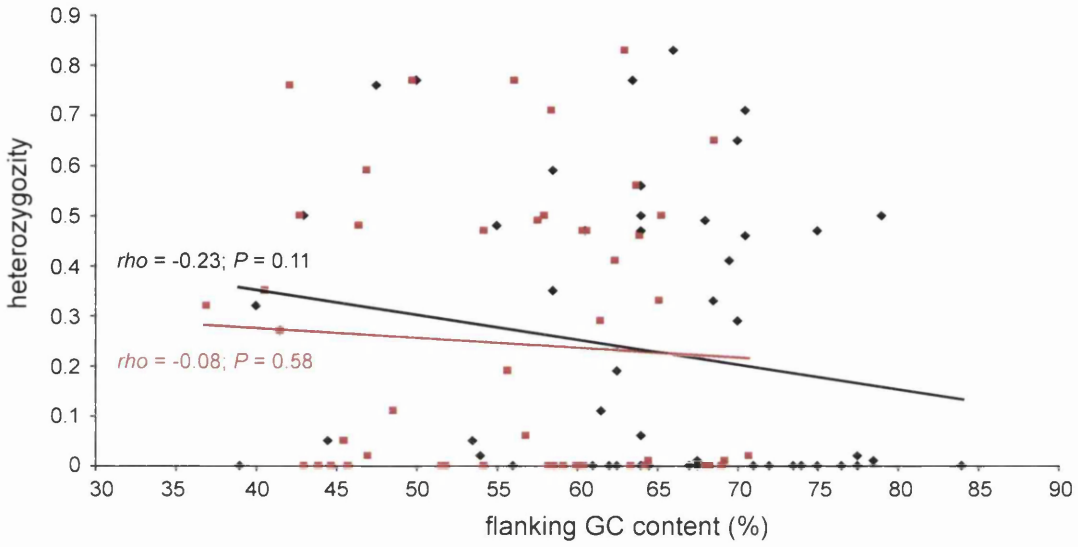


Figure 5.10 Heterozygozity of exonic CAG•CTG repeats does not correlate with flanking GC content. The plot illustrates the relationship between microsatellite heterozygozity and flanking GC content for 51 exonic microsatellites with a minimum length of six repeats. The GC content of 100 bp (black diamond) and 1000 bp (red square) of flanking sequence is shown. The rank correlation coefficient (Spearman's *rho*) of flanking GC content with heterozygozity is also shown.

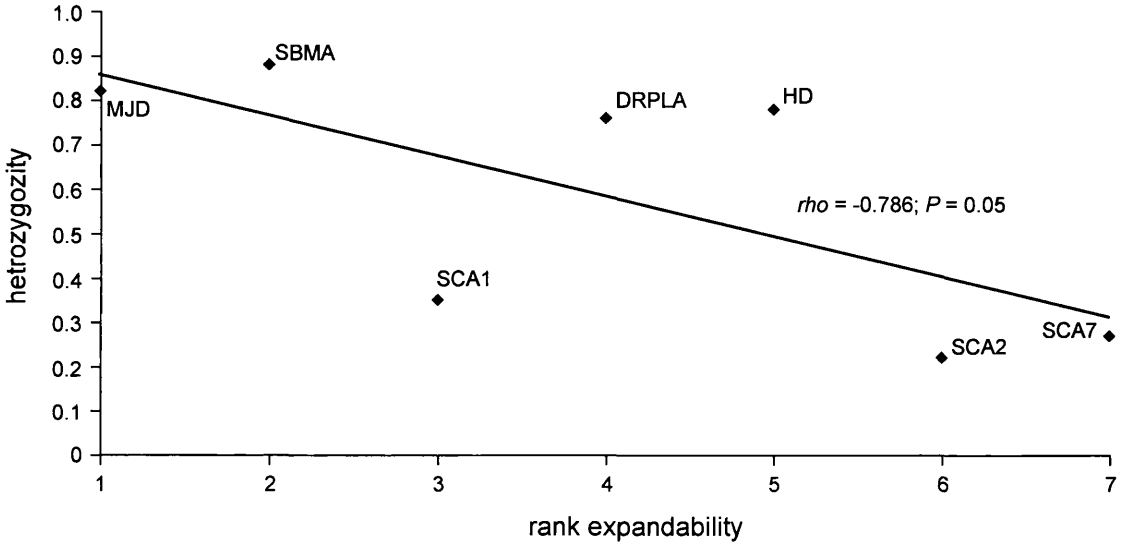


Figure 5.11. Heterozygozity and expandability of disease associated polglutamine loci are negatively correlated.

5.3 Discussion

The expandability of disease associated unstable CAG•CTG repeats is positively correlated with flanking GC content (Chapter 3) (Brock *et al.*, 1999). Competent mismatch repair (MMR) is required to effect instability at these expanded loci (Gomes-Pereira *et al.*, 2004; Manley *et al.*, 1999; van den Broek *et al.*, 2002). We have suggested that these observations may reflect a biological relationship; flanking GC content may promote the recruitment or action of MMR at expanded loci. MMR is also involved in the variability of normal length microsatellites. However, unlike expanded repeat instability, microsatellite instability is increased by deficiencies in the MMR machinery. Thus, we hypothesised that microsatellites flanked by GC-poor sequences may be more mutable than microsatellites flanked by GC-rich sequences, reflecting less efficient MMR recruitment to, or activity at, microsatellites flanked by low GC content sequences. To investigate the effect of flanking sequence composition on microsatellite instability we carried out the first genome-wide study of flanking sequence effects on AGC-motif microsatellite mutability in humans.

We found a positive association between short (≤ 7 repeats) microsatellite length and flanking GC content. Consistent with our hypothesis, a negative correlation between flanking sequence GC content and microsatellite length was found for long (> 7 repeats) microsatellites. However, this correlation was not statistically significant. As exonic sequences tend to have a higher GC content than non-exonic sequences, longer ACG-motif microsatellites would be expected to arise by chance in exonic regions. Indeed, we found that exonic repeats were significantly longer than non-exonic microsatellites. Strikingly, the correlation between short (< 7 repeats) microsatellite length and flanking GC content was even more significant when microsatellites located within exonic sequences were analysed separately. This finding is particularly interesting given the exonic nature of the expanded CAG•CTG repeat sequences (Chapter 3). A correlation between short microsatellite length and flanking GC content was expected, as the probability of forming a GC rich repeat by random chance, such as a AGC-motif microsatellite, increases with the GC content of the host sequence. For example, whereas a single point mutation in the sequence 'CGCCAGCAGCGG'

would result in lengthening of the CAG repeat tract at least two such mutations would be required to result in a similar repeat expansion in the AT rich sequence, 'AATCAGCAGAAT'.

However, chance association of nucleotides would predict that approximately one $(AGC)_5$ sequence would occur in the human genome. However, over 2,000 such $(AGC)_5$ sequences are observed. It seems unlikely that point mutations alone would result in such an overrepresentation of these sequences, unless base substitution in the flanking sequences was biased towards extending microsatellite length. Indeed, it has been suggested that the nucleotides directly flanking microsatellites show mutational biases (Vowles and Amos, 2004), however more recent work suggests no such biases are present (Webster and Hagberg, 2007). Thus, it is likely that processes such as misalignment mutation are also acting at short microsatellite loci. The lack of correlation between flanking GC content and repeat number for longer repeats, suggests that that local sequence composition is not a major modifier of mutability at these loci.

The observed repeat length of a microsatellite offers little insight into the series of mutational events that resulted in its present state. In order to identify loci at which misalignment mutation events have occurred, we sought to quantify repeat length variation at orthologous human and chimpanzee microsatellite loci. Pre-genomic era studies of human-chimpanzee microsatellite mutability were based on amplification of microsatellites in both species using primers designed for microsatellite marker detection in humans. As human microsatellite markers are typically long and highly polymorphic, and as a maximum length threshold exists for microsatellites (Ellegren, 2004), the orthologous loci in chimpanzee were usually found to be shorter than in humans. This 'ascertainment bias' frustrated early studies of microsatellite mutability based on inter-species comparisons. Publication of a high quality sequence of the chimpanzee genome facilitated the direct observation of microsatellite mutability on a genome-wide scale by comparison of orthologous microsatellites in the human and chimpanzee genomes. Whereas such inter-genome studies have furthered knowledge of the factors internal to microsatellites that affect their instability such as repeat number, motif size, and motif composition, few have analysed the effects of proximal flanking sequence on mutability.

Moreover, previous studies have tended to analyse the mutability of microsatellites grouped by motif size (all dinucleotide microsatellites, all trinucleotide microsatellites, etc.). As within each motif-size class, motif composition is known to have a significant effect on mutability, such studies may fail to detect the effect of flanking sequences which are specific to repeats of a particular motif-composition (Kelkar *et al.*, 2008). Moreover, most genome-scale studies exclude short microsatellite sequences, and are thus susceptible to the effect of sampling bias. Ostensibly, exclusion of short microsatellites is due to the widely held view that misalignment events are very rare at these loci, despite scant evidence to support this assumption. However, as approximately 99% of all microsatellite sequences are composed of fewer than 4 repeats, their exclusion most likely reflects attempts to reduce the computational magnitude of the analysis.

We identified 209,000 orthologous ACG-microsatellites in humans and chimpanzees, of which 324 were found to have changed by at least one whole repeat unit. As orthologous loci were identified by aligning the 5' and 3' flanking sequences of human microsatellites with the chimpanzee genome, the observed changes were not due to point mutations of either the microsatellite or its flanking sequences. Thus, we concluded that the observed changes were due to misalignment mutation events. Consistent with previous studies, it was found that microsatellite mutability increased with repeat number. However, a large proportion (43%) of mutated loci consisted of fewer than four repeats, suggesting that misalignment mutation is a major source of repeat number variation, even at very short loci. Significantly, we found that the frequency of length changes at mutated orthologous loci increased with increasing non-integer repeat units (mutability: 2.7 repeats > 2.3 repeats > 2.0 repeats), suggesting that such partial repeat units are an integral part of microsatellites as a biological unit.

Analysing flanking sequence GC content, it was found that mutated orthologous loci had a significantly higher flanking GC content than unchanged orthologous loci (Table 5.4). Moreover, consistent with our intra-human analysis, no correlation was found between the magnitude of length change observed at orthologous loci and flanking GC content. Taken together, the data from both

the intra-human and inter-species analysis suggest the following model of mutability for AGC-motif microsatellites. Sequence composition dependent, point mutation mediated expansion of short microsatellites delivers loci to a length at which sequence composition independent, misalignment mutation becomes the dominant mutational process (Figure 5.12). Thus a given microsatellite sequence is most likely the product of both point mutation and misalignment mutation repeat length variations. Our proposed model of AGC-motif microsatellite mutability suggests that GC content-dependent point mutations delivers short repeats to a length at which a GC content independent process of misalignment mutation dominates (Figure 5.12). As microsatellites flanked by sequences with higher GC contents are more likely to reach this point of mutational 'change-over', it would be predicted that microsatellites undergoing misalignment mutation would have a higher flanking GC content than those at which misalignment mutation has not taken place, as was found to be the case.

Despite the large sample size afforded by inter-genome analysis, such studies are subject to inherent biases, and biases resulting from the method of orthologous microsatellite detection employed. First, as only two alleles, one human and one chimpanzee are sampled, the direction and number of changes that have occurred cannot be inferred from the observed difference in repeat number between species. Moreover, where orthologous microsatellite pairs have independently mutated to the same repeat number, no change will be observed between species. Thus, inter-genome comparisons of microsatellite repeat number inherently underestimate the true levels of genome-wide mutability. Employing the genomic sequence of a third species as an out-group could serve to reduce, but not eliminate this bias. Secondly, as we have employed the flanking sequences of microsatellites identified in the human genome (focal genome) to locate those in the chimp genome (sister genome), but not vice versa, microsatellite loci of less than two repeats will not be identified in the human genome resulting in a distortion of repeat lengths observed for short microsatellites (Figure 5.9A). This sampling bias can be overcome by performing the reciprocal analysis employing the chimpanzee genome as the focal genome.

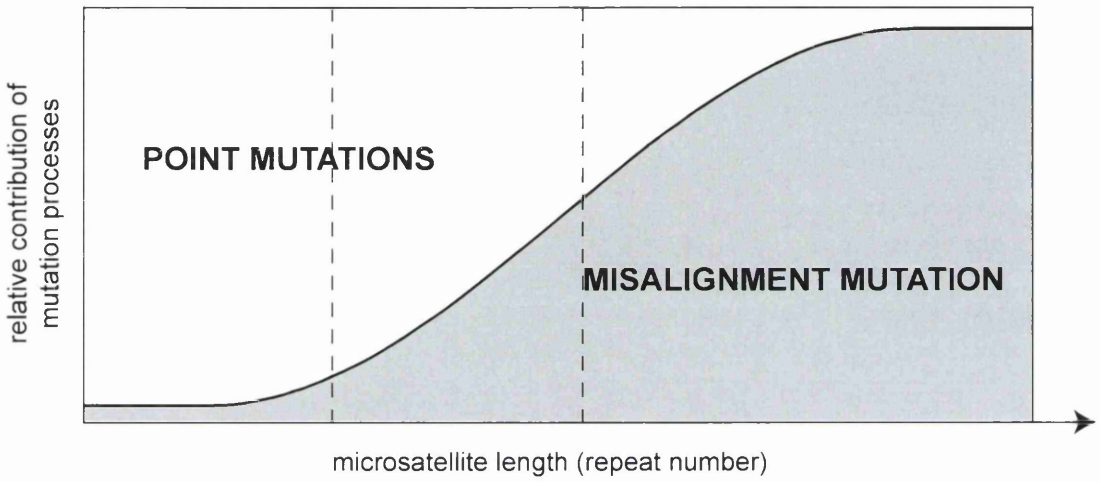


Figure 5.12. Relative contribution of mutational processes acting on AGC-motif microsatellites is length and flanking sequence dependent. Vertical dashed lines indicate a region where a complex combination of point mutations and misalignment mutations are mediating changes in repeat number

Consistent with our inter-genome analysis no correlation between the flanking GC content of exonic microsatellites and their heterozygozity was revealed. Interestingly, we noted that the heterozygozity of the normal length range disease-associated polyglutamine-encoding CAG repeats was significantly negatively correlated with their expandability in the disease associated length range. This striking observation suggests that the process of expanded repeat stability is not simply an extension of the process (presumably misalignment mutation) underlying the mutability of the disease loci in their normal length range.

6. Final discussion, main conclusions and future perspectives

Nearly two decades have passed since the expansion of a simple tandem repeat was identified as the mutation underlying fragile X syndrome (Kremer *et al.*, 1991). In the intervening years, a further 16 disorders have been linked to the expansion of usually benign short tandem repeats (Figure 1.1) (Gomes-Pereira and Monckton, 2006). The term 'dynamic mutation' was applied to the process of repeat expansion to distinguish this atypical mutational process from more canonical genetic mutations (Richards and Sutherland, 1992). The repeat tracts at these loci are typically small and polymorphic in the general population. However, upon expansion the repeat tracts become pathogenic and hypermutable, exhibiting expansion-biased, tissue-specific instability. As repeat pathogenicity increases with length, longer repeats result in an earlier age of disease onset and increased severity of symptoms. Intergenerationally, dynamic expansion of disease-associated expanded repeats results in successive generations being affected more severely, and at an earlier age; a phenomenon termed 'anticipation'. Given the typically milder symptoms and later disease onset in first generation carriers of an expanded repeat mutation, the development of early, more severe symptoms in the final generation can often be the point at which the disorder is first identified in a family, followed by testing and diagnosis in the wider family. Thus, the process of dynamic mutation has a direct and devastating human cost on entire families, for which there is currently no effective therapy for prevention or cure.

As dynamic mutation is causative, and common to all these disorders, it offers a unique site for therapeutic intervention. Unfortunately, in spite of intensive study over the past decade, our knowledge of the mechanism underlying dynamic mutation is limited.

6.1 Main conclusions

6.1.1 Somatic mosaicism is a mediator of disease progression

Expansion of trinucleotide repeats in the germline is responsible for the phenomenon of anticipation, whereas length- and age-dependent expansion in somatic tissues is believed to be responsible for the tissue-specificity and progressive nature of the expanded repeat disorders (Fortune et al., 2000; Gomes-Pereira and Monckton, 2006; Shelbourne et al., 2007). As repeat toxicity increases with length, it is intuitive to assume that the rate at which repeats expand in affected tissues will be a major modifier of age at onset and progression of symptoms. Thus, assuming an equal inherited repeat length, disorders with faster expanding repeat tracts would be predicted to have an earlier age at onset than loci expanding more slowly. Indeed, the polyglutamine disorders show marked differences in the repeat number (termed 'locus toxicity') required to effect a similar age at onset of symptoms (Figure 3.1). This inter-locus toxicity difference has been widely assumed to reflect protein context effects on the cytotoxicity of the expanded polyglutamine tract, although no evidence to support this assumption has been reported. However, we proposed that the inter-locus difference simply reflects differences in the rate at which somatic expansion delivers repeats to their cytotoxic state at each locus.

In order to test this hypothesis we analysed published data of locus expandability (Brock *et al.*, 1999) and the relationship between repeat number and age at onset for seven polyglutamine disorders (Gusella and MacDonald, 2000). Importantly, the repeat lengths quoted in both studies are data from affected individuals. Consistent with our hypothesis, we revealed a significant negative correlation between locus expandability and locus toxicity. That is, faster expanding loci such as SCA2 required fewer repeats to effect a given age at onset when compared to a more slowly expanding locus, such as MJD (Figure 6.1). Our somatic expansion rate model of inter-locus toxicity is also consistent with several clinical cases in which individuals with expanded but stable repeats tend to have a later age of onset and milder symptoms than individuals with an unstable repeat of similar length (Chapter 3). Our model proposes that disease

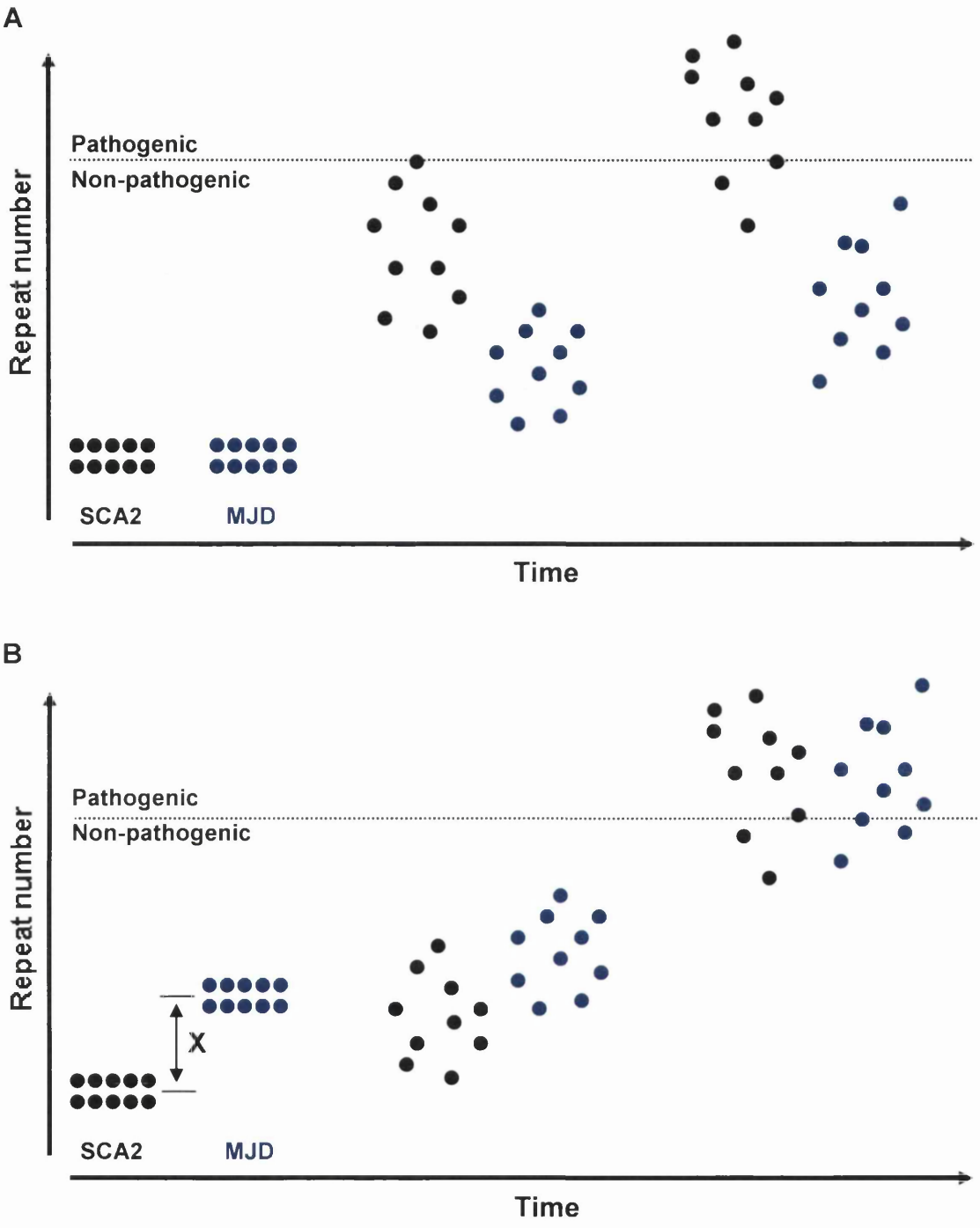


Figure 6.1. The effect of somatic expansion rate on age at onset and inter-locus toxicity. The change in repeat length in a generalised somatic tissue over time in SCA2 and MJD is shown. Solid circles represent cells containing an expanded CAG repeat in the same somatic tissue of a SCA2 affected individual (**black circles**) and a MJD affected individual (blue circles). **A)** If the same repeat length is inherited in SCA2 and MJD, the faster expansion of the repeat in SCA2 alleles delivers the repeat to is pathogenic length before MJD resulting in onset of symptoms in SCA2 before MJD. **B**) To affect a similar age at onset in SCA2 and MJD, the MJD affected individual must inherit a longer allele than the SCA2 affected individual. This length difference (**X**), corresponds to the inter-locus toxicity difference between MJD and SCA2

progression is governed by the rate at which somatic expansion delivers repeats to their pathogenic state, not just inherited repeat length. As such, this model predicts that age at onset of an individual homozygous for an expansion would be determined by the time taken for their longer allele to reach pathogenic length. That is, as the longer allele would take a shorter time to reach its cytotoxic state it would determine age at onset. Thus, an individual homozygous for an expanded allele would be expected to have an age at onset similar to that of an individual heterozygous for the longer expansion. However, as the second expanded allele would reach its pathogenic length not long thereafter in the homozygote, disease would progress more rapidly than for a heterozygote with a single pathogenic allele. Indeed, individuals homozygous for expanded alleles at the *SCA1* (Goldfarb *et al.*, 1996), *SCA2* (Sanpei *et al.*, 1996) and *HD* (Durr *et al.*, 1999) loci have a similar age at onset as length-matched heterozygotes. Significantly, a study of eight individuals homozygous for HD found that not only did age at onset occur within the range expected for heterozygotes with an expanded allele of similar length, but also that the rate of disease progression was significantly accelerated in the homozygous individuals (Squitieri *et al.*, 2003), consistent with a role for somatic expansion in disease progression.

Interestingly, during the course of our research an independent study employing computer simulations and mathematical modeling predicted that length-dependent somatic expansion underlies age at onset and disease progression in trinucleotide diseases (Kaplan *et al.*, 2007). Our results, based on data from affected individuals are consistent with this model.

Here, we have reported for the first time, a correlation between locus expandability and locus toxicity, implicating somatic mosaicism as a modifier of disease progression. Our finding re-emphasizes the importance of the role played by somatic expansion in disease progression and highlights the importance of understanding the mechanism of dynamic mutation. Of course, we acknowledge that correlation does not mean causation. However, we argue that our data provides a simple and intuitive explanation of the inter-locus toxicity differences observed between the polyglutamine disorders based on real data from affected individuals.

From a therapeutic perspective, the process of somatic expansion offers a potential site for intervention. Indeed, the data (Chapter 3) suggests that were the *SCA2* tract as stable as the *MJD* repeat, the majority of SCA2 individuals with less than 60 repeats would not develop symptoms. Moreover, whereas therapies targeted to the downstream affects of expanded polyglutamine proteins in affected tissues are likely to be disease-specific, therapies directed to somatic expansion are more likely to be generally applicable to all expanded CAG•CTG repeat disorders. Indeed, although admittedly limited in scope, preliminary screens for chemical modifiers of expanded repeat instability have been promising (reviewed in (Gomes-Pereira and Monckton, 2006). For example, in cell models of myotonic dystrophy the chemicals caffeine, aspirin, and 5-aza-2-deoxycytidine have been found to increase expansion rate, decrease expansion rate and increase the frequency of large deletions, respectively (Gomes-Pereira and Monckton, 2004; Gorbunova et al., 2004). These experiments provide proof of principle that somatic expansion can be modified by chemical agents. As we gain a better understanding of the mechanism of dynamic mutation more rational design of therapeutic agents, possibly targeted to the mismatch repair system may be designed.

6.1.2 Flanking GC content is a *cis*-acting modifier of expanded CAG•CTG repeat instability and genome-wide CAG•CTG microsatellite instability

Several observations have indirectly suggested a role for *cis*-acting modifiers of expanded repeat stability. Identifying these *cis*-elements may provide insights into the mechanism of dynamic mutation and provide potential sites for therapeutic intervention. Building on previous work (Brock *et al.*, 1999), we detailed a significant positive correlation between proximal (≤ 500 bp) flanking DNA GC content and locus expandability, and locus toxicity for seven polyglutamine disorders (Chapter 3). We hypothesized that this correlation reflected a true biological relationship, and resulted from an influence of flanking GC content on expanded repeat instability. In further agreement with this hypothesis, when we expanded this analysis to include the non-coding loci *DM1* and *ERDA1* the association between flanking GC content and expandability became even more significant.

As the polyglutamine encoding CAG repeats are located in exons, the differences in GC content at the DNA level may have reflected a requirement for particular amino acid sequences flanking the polyQ repeat in the mature protein; possibly consistent with the protein context mediated model of inter-locus toxicity. However, our analysis failed to reveal a similarly significant correlation between flanking GC content and locus toxicity at the level of the mRNA. Similarly, the physiochemical properties of the primary sequence flanking the polyglutamine tract did not correlate with toxicity, arguing against a protein context mediated model of inter-locus toxicity. Thus, we proposed that disease progression is modified by rate of somatic expansion, which is in turn, modified by flanking GC content.

How might flanking GC content modify dynamic mutation? Flanking sequence composition may affect the formation of instability mediating secondary structures or affect their downstream processing by the mismatch repair machinery. The formation of non B-DNA slipped strand structures is an assumption of the inappropriate mismatch repair model of repeat expansion (Chapter 1, Figure 1.3). It has been shown *in vitro* that flanking sequence can affect the propensity of CTG repeats to form S-DNA structures (Pearson *et al.*, 1998a). Flanking GC content may affect the formation of such structures by altering the size and lifespan of single stranded DNA bubbles formed during transcription or replication. Similarly, chromatin remodeling or nucleosome repositioning may result in DNA conformations favorable for secondary structure formation. Indeed, CTG repeats with as few as six repeats have been found to act as strong positioning signals for nucleosomes *in vitro* (Godde and Wolffe, 1996). Flanking GC content may affect the positioning or remodeling of such repeat-anchored nucleosomes. Similarly, a higher flanking GC content may facilitate the formation of single-stranded DNA directly by promoting spontaneous DNA melting (DNA breathing) (Dornberger *et al.*, 1999). Alternatively, once secondary structures formed the high flanking GC content may act to stabilize them. Current evidence suggests that the majority of contacts between the MutS heterodimers and DNA are with the DNA backbone rendering binding relatively sequence independent. However, recognition of the mismatched DNA requires the introduction of large conformational changes in the DNA which may be affected by the local sequence context (Kunkel and Erie,

2005). As MMR is required for expanded repeat instability, it is possible that flanking GC content promotes/facilitates recruitment or binding of MMR components to slipped strand structures within the repeat. It is also possible that the higher order chromatin structure of the DNA may affect its accessibility to or the effectiveness of the MMR machinery.

The mutability of normal (not hypermutable) microsatellites is thought to be mediated by replication slippage followed by failure to recognise or repair the misalignment by the MMR system (Ellegren, 2004). The involvement of MMR in microsatellite evolution is further suggested by the widespread microsatellite instability observed in tumours of HPNCC patients deficient for various components of the MMR system. Although microsatellite mutation over an evolutionary timescale and the hypermutation of an expanded repeat locus over the lifetime of an individual are most likely mediated by different mechanisms, flanking GC content may modify the involvement of MMR in both.

As microsatellite mutability increases with length, we analysed the relationship between flanking GC content and ACG-motif microsatellite length in the human genome. We revealed a striking correlation between flanking GC content and microsatellite length for short microsatellites (≤ 7 repeats), whereas the length of long microsatellites was not correlated with GC content. In order to identify microsatellites at which misalignment mutation events had occurred, we identified orthologous microsatellite loci in the human and chimpanzee at which the microsatellites differed in length by one or more repeat units. Two important, novel observations were made. First, although microsatellite mutability increased with increasing repeat number, many misalignment mutation events were observed at loci with fewer than four repeat units. Previous studies of microsatellite mutability have excluded microsatellites with fewer than four repeats on the false assumption that misalignment mutation does not occur at such loci. Thus, misalignment mutation of short microsatellites is a major source of variation in mammalian genomes. Secondly, mutability increased with increasing non-integer repeat number. That is, the magnitude of length changes observed at orthologous loci such as 'CAGCAGCAGC' (3.3 repeats; 10 bp) was greater than those observed at loci such as 'CAGCAGCAG' (3 repeats; 9 bp). This finding suggests that the conventional definition of microsatellite

sequences in terms of whole repeat numbers is incorrect, and that these fractional repeat units are an integral part of microsatellites as they exist and change in cells. Thus, nucleotides which have been typically excluded from the microsatellite sequence and classed as flanking sequence in previous studies of microsatellite mutability, affect the mutation rate of the repetitive sequence. Mutated microsatellites had a higher flanking GC content than non-mutated microsatellites suggesting an effect of flanking sequence on microsatellite mutability. However, no significant correlation between the magnitude of length changes observed between orthologous microsatellite loci and flanking GC content was found.

6.1.3 *Cis*-acting modifiers of expanded repeat instability

In order to directly assess the role of flanking sequence elements as modifiers of expanded repeat instability, we proposed constructing a mammalian cell culture model which would facilitate the targeting of different expanded repeats and their endogenous flanking sequences to the same genomic location, allowing for comparison of repeat dynamics independent of genomic position effects (Figure 3.19). To assess the feasibility of this system, we investigated the ability of mammalian cell lines to model expanded repeat instability.

Single cell-derived HeLa clones stably transfected with a CTG_{143} repeat from the DM1 locus, failed to exhibit expansion-biased instability in culture. As HeLa cells have a well characterised and functional MMR system, this result suggested that whereas a functional MMR system maybe necessary for expanded repeat instability, competent MMR is not sufficient for instability. Indeed, the observation that levels of somatic instability are tissue specific in affected individuals suggests the involvement of other tissue specific *trans*-acting factors. It is possible that these factors are absent in HeLa cells. However, similar results were obtained when single-cell *Dmt*-D clones were stably transfected with the same transgenic repeat. As the *Dmt*-D cells already contained a similar, unstable transgenic repeat the absence of *trans*-acting modifiers of instability could not explain the stability observed in our transgene. Thus, the presence of all the *trans*-acting factors required for expanded repeat instability is not always sufficient for instability to occur at an expanded repeat. This remarkable finding

emphasises the dominant effect *cis*-acting modifiers can have on repeat stability, and highlights the difficulty and unpredictability of generating mammalian models of expanded repeat stability.

However, the presence of two expanded repeats with identical immediate flanking sequences exhibiting markedly different levels of stability in the same cell line represented an ideal model system in which to study the affect of *cis*-elements on repeat stability. CpG methylation of the sequences flanking expanded repeats has been suggested as a potential modifier or marker of repeat expansion (Filippova *et al.*, 2001; Steinbach *et al.*, 1998). Employing restriction-digest-PCR assays we revealed that whereas the flanking sequences of the unstable repeat were hyper-methylated, the sequence flanking the stable repeat were completely unmethylated. However, subsequent methylation of the stable repeat tract failed to induce instability, suggesting that methylation of the repeat tract and its immediate flanking sequences is not sufficient for instability. Taken together these data argue against a simple link between flanking sequence methylation and instability. However, they do not exclude methylation as a potential modifier of instability. It is possible that methylation during embryonic development triggers or facilitates instability by effects in *cis*, such as changes in chromatin organisation or the recruitment of other epigenetic marks such as histone acetylation. If so, the generation of dynamic mutation models in adult mammalian cells may be very challenging. In addition, no association was found between repeat transcription and repeat instability, although all unstable repeats were transcribed. This is consistent with the general observation that transcription of expanded repeats is necessary, but not sufficient for instability.

6.2 Future directions

Further research is required to characterise the effect of *cis*-acting modifiers on expanded repeat instability. Our results suggest that flanking sequence composition, flanking sequence methylation and repeat transcription alone are not sufficient to facilitate instability. In order to further validate these results, the experiments presented here should be repeated using a longer (> 200 CTG) repeat tract, to rule out the unlikely possibility that the transgenic repeat employed here was not sufficiently long enough to exhibit instability. In addition, the use of a selectable marker such as *HyTK* flanked by lox sites would allow for excision of the selectable marker after generation of the stably transfected clones, eliminating the possibility of elements within the sequence of the selectable marker exerting a stabilising influence over the repeat tract.

Assuming our results are correct, several inter-related avenues of research seem a logical progression from the experiments presented here. Although both the degree of CpG methylation and expression level of a gene can reflect chromatin state, neither are direct measures of chromatin state. Moreover, although we revealed that flanking a repeat with insulator elements, which would be predicted to maintain the repeat in an open chromatin formation after integration into the host genome, had no affect on stability, we did not directly assess the chromatin state of the transgenes. DNaseI sensitivity assays could be used to elucidate differences in the chromatin state of the stable and unstable transgenic repeats. Moreover, as CTCF-binding to sites directly flanking the DM1 repeat has been suggested to lead to changes in chromatin state and repeat stability (Cho *et al.*, 2005; Filippova *et al.*, 2001), it would be interesting to assay for CTCF binding to the CTCF sites in the unstable and stable transgenes using chromatin immunoprecipitation (ChIP) assays. Although we have found no association between repeat stability and repeat transcription, it is possible that bi-directional transcription of expanded repeats is a modifier of stability. Indeed, bi-directional transcription has been found at the SCA8 (Moseley *et al.*, 2006) and DM1 loci (Cho *et al.*, 2005). Moreover, it has been suggested that anti-sense transcripts at the DM1 locus result in heterochromatin formation through a poorly understood process which results in the conversion of the anti-sense

transcripts into small interfering RNA (siRNA), which subsequently target chromatin modifying factors to the DM1 locus (Cho *et al.*, 2005). Interestingly, it has been shown that the *Dmt* transgene is bi-directionally transcribed in *Dmt*-D transgenic mice (Fortune, 2001).

We have provided evidence that somatic mosaicism may be a major modifier of disease progression in the expanded trinucleotide disorders, and that a somatic expansion rate model of disease progression can fully explain the striking inter-locus toxicity differences observed between the polyglutamine disorders. In addition, we have shown *cis*-elements can act as powerful modifiers of expanded repeat instability, and that a long, methylated, transcribed repeat is not sufficient for instability in mammalian cells. Thus, further investigation of the mechanism of dynamic mutation and the nature and function of its *cis*-acting modifiers is required.

References

Almeida, P., and C. Penha-Goncalves. 2004. Long perfect dinucleotide repeats are typical of vertebrates, show motif preferences and size convergence. *Mol Biol Evol.* 21:1226-33.

Al-Ramahi, I., Y.C. Lam, H.K. Chen, B. de Gouyon, M. Zhang, A.M. Perez, J. Branco, M. de Haro, C. Patterson, H.Y. Zoghbi, and J. Botas. 2006. CHIP protects from the neurotoxicity of expanded and wild-type ataxin-1 and promotes their ubiquitination and degradation. *J Biol Chem.*

Anvret, M., G. Ahlberg, U. Grandell, B. Hedberg, K. Johnson, and L. Edstrom. 1993. Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum Mol Genet.* 2:1397-400.

Arrasate, M., S. Mitra, E.S. Schweitzer, M.R. Segal, and S. Finkbeiner. 2004. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature.* 431:805-10.

Ashizawa, T., J.R. Dubel, and Y. Harati. 1993. Somatic instability of CTG repeat in myotonic dystrophy. *Neurology.* 43:2674-8.

Bachtrog, D., M. Agis, M. Imhof, and C. Schlotterer. 2000. Microsatellite variability differs between dinucleotide repeat motifs-evidence from Drosophila melanogaster. *Mol Biol Evol.* 17:1277-85.

Barbeau, A., M. Roy, L. Cunha, A.N. de Vincente, R.N. Rosenberg, W.L. Nyhan, P.L. MacLeod, G. Chazot, L.B. Langston, D.M. Dawson, and et al. 1984. The natural history of Machado-Joseph disease. An analysis of 138 personally examined cases. *Can J Neurol Sci.* 11:510-25.

Barcelo, J.M., M.S. Mahadevan, C. Tsilfidis, A.E. MacKenzie, and R.G. Korneluk. 1993. Intergenerational stability of the myotonic dystrophy protomutation. *Hum Mol Genet.* 2:705-9.

Bell, A.C., A.G. West, and G. Felsenfeld. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell.* 98:387-96.

Bell, G.I., and J. Jurka. 1997. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *J Mol Evol.* 44:414-21.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B.* 57:289 -300.

Bennett, E.J., T.A. Shaler, B. Woodman, K.Y. Ryu, T.S. Zaitseva, C.H. Becker, G.P. Bates, H. Schulman, and R.R. Kopito. 2007. Global changes to the ubiquitin system in Huntington's disease. *Nature.* 448:704-8.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573-80.

Benton, C.S., R. de Silva, S.L. Rutledge, S. Bohlega, T. Ashizawa, and H.Y. Zoghbi. 1998. Molecular and clinical studies in SCA-7 define a broad clinical spectrum and the infantile phenotype. *Neurology.* 51:1081-6.

Bernstein, B.E., A. Meissner, and E.S. Lander. 2007. The mammalian epigenome. *Cell.* 128:669-81.

Berry-Kravis, E., F. Lewin, J. Wuu, M. Leehey, R. Hagerman, P. Hagerman, and C.G. Goetz. 2003. Tremor and ataxia in fragile X premutation carriers: blinded videotape study. *Ann Neurol.* 53:616-23.

Bhaskaran, R., and P.K. Ponnuswamy. 1984. Dynamics of amino acid residues in globular proteins. *Int J Pept Protein Res.* 24:180-91.

Boyd, A.C., H. Davidson, B. Stevenson, G. McLachlan, H. Davidson-Smith, and D.J. Porteous. 1999. pSURF-2, a modified BAC vector for selective YAC cloning and functional analysis. *Biotechniques.* 27:164-70, 172, 175.

Brandstrom, M., and H. Ellegren. 2007. The genomic landscape of short insertion and deletion polymorphisms in the chicken (Gallus gallus) Genome: a high frequency of deletions in tandem duplicates. *Genetics.* 176:1691-701.

Brandstrom, M., and H. Ellegren. 2008. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Res.*

Brock, G.J., N.H. Anderson, and D.G. Monckton. 1999. Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. *Hum Mol Genet.* 8:1061-7.

Brohede, J., and H. Ellegren. 1999. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proc Biol Sci.* 266:825-33.

Brook, J.D., M.E. McCurrach, H.G. Harley, A.J. Buckler, D. Church, H. Aburatani, K. Hunter, V.P. Stanton, J.P. Thirion, T. Hudson, and et al. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell.* 69:385.

Brunner, H.G., H.T. Bruggenwirth, W. Nillesen, G. Jansen, B.C. Hamel, R.L. Hoppe, C.E. de Die, C.J. Howeler, B.A. van Oost, B. Wieringa, and et al. 1993. Influence of sex of the transmitting parent as well as of parental allele size on the CTG expansion in myotonic dystrophy (DM). *Am J Hum Genet.* 53:1016-23.

Calabresi, V., S. Guida, A. Servadio, and C. Jodice. 2001. Phenotypic effects of expanded ataxin-1 polyglutamines with interruptions in vitro. *Brain Res Bull.* 56:337-42.

Campuzano, V., L. Montermini, Y. Lutz, L. Cova, C. Hindelang, S. Jiralerspong, Y. Trottier, S.J. Kish, B. Faucheux, P. Trouillas, F.J. Authier, A. Durr, J.L. Mandel, A. Vescovi, M. Pandolfo, and M. Koenig. 1997. Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum Mol Genet.* 6:1771-80.

Campuzano, V., L. Montermini, M.D. Molto, L. Pianese, M. Cossee, F. Cavalcanti, E. Monros, F. Rodius, F. Duclos, A. Monticelli, F. Zara, J. Canizares, H. Koutnikova, S.I. Bidichandani, C. Gellera, A. Brice, P. Trouillas, G. De Michele, A. Filla, R. De Frutos, F. Palau, P.I. Patel, S. Di Donato, J.L. Mandel, S. Cocozza, M. Koenig, and M. Pandolfo. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science.* 271:1423-7.

Chai, Y., S.S. Berke, R.E. Cohen, and H.L. Paulson. 2004. Poly-ubiquitin binding by the polyglutamine disease protein ataxin-3 links its normal function to protein surveillance pathways. *J Biol Chem.* 279:3605-11.

Chai, Y., S.L. Koppenhafer, N.M. Bonini, and H.L. Paulson. 1999. Analysis of the role of heat shock protein (Hsp) molecular chaperones in polyglutamine disease. *J Neurosci.* 19:10338-47.
Chai, Y., L. Wu, J.D. Griffin, and H.L. Paulson. 2001. The role of protein composition in specifying nuclear inclusion formation in polyglutamine disease. *J Biol Chem.* 276:44889-97.
Chen, X., S.V. Mariappan, P. Catasti, R. Ratliff, R.K. Moyzis, A. Laayoun, S.S. Smith, E.M. Bradbury, and G. Gupta. 1995. Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc Natl Acad Sci U S A.* 92:5199-203.
Cho, D.H., C.P. Thienes, S.E. Mahoney, E. Analau, G.N. Filippova, and S.J. Tapscott. 2005. Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol Cell.* 20:483-9.
Chong, S.S., A.E. McCall, J. Cota, S.H. Subramony, H.T. Orr, M.R. Hughes, and H.Y. Zoghbi. 1995. Gametic and somatic tissue-specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. *Nat Genet.* 10:344-50.
Chou, P.Y., and G.D. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol.* 47:45-148.
Choudhry, S., M. Mukerji, A.K. Srivastava, S. Jain, and S.K. Brahmachari. 2001. CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. *Hum Mol Genet.* 10:2437-46.
Chung, J.H., A.C. Bell, and G. Felsenfeld. 1997. Characterization of the chicken beta-globin insulator. *Proc Natl Acad Sci U S A.* 94:575-80.
Chung, M.Y., L.P. Ranum, L.A. Duvick, A. Servadio, H.Y. Zoghbi, and H.T. Orr. 1993. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat Genet.* 5:254-8.
Cleary, J.D., and C.E. Pearson. 2005. Replication fork dynamics and dynamic mutations: the fork-shift model of repeat instability. *Trends Genet.* 21:272-80.
Cummings, C.J., E. Reinstein, Y. Sun, B. Antalffy, Y. Jiang, A. Ciechanover, H.T. Orr, A.L. Beaudet, and H.Y. Zoghbi. 1999. Mutation of the E6-AP ubiquitin ligase reduces nuclear inclusion frequency while accelerating polyglutamine-induced pathology in SCA1 mice. *Neuron.* 24:879-92.
Cummings, C.J., and H.Y. Zoghbi. 2000a. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet.* 9:909-16.
Cummings, C.J., and H.Y. Zoghbi. 2000b. Trinucleotide repeats: mechanisms and pathophysiology. *Annu Rev Genomics Hum Genet.* 1:281-328.
David, G., A. Durr, G. Stevanin, G. Cancel, N. Abbas, A. Benomar, S. Belal, A.S. Lebre, M. Abada-Bendib, D. Grid, M. Holmberg, M. Yahyaoui, F. Hentati, T. Chkili, Y. Agid, and A. Brice. 1998. Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). *Hum Mol Genet.* 7:165-70.
Day, J.W., R. Roelofs, B. Leroy, I. Pech, K. Benzow, and L.P. Ranum. 1999. Clinical and genetic characteristics of a five-generation family with a novel form of myotonic dystrophy (DM2). *Neuromuscul Disord.* 9:19-27.
de Chiara, C., R.P. Menon, F. Dal Piaz, L. Calder, and A. Pastore. 2005. Polyglutamine is not all: the functional role of the AXH domain in the ataxin-1 protein. *J Mol Biol.* 354:883-93.
de la Chapelle, A., and P. Peltomaki. 1995. Genetics of hereditary colon cancer. *Annu Rev Genet.* 29:329-48.
Debacker, K., and R.F. Kooy. 2007. Fragile sites and human disease. *Hum Mol Genet.* 16 Spec No. 2:R150-8.
Deleage, G., and B. Roux. 1987. An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* 1:289-94.
Devys, D., V. Biancalana, F. Rousseau, J. Boue, J.L. Mandel, and I. Oberle. 1992. Analysis of full fragile X mutations in fetal tissues and monozygotic twins indicate that abnormal methylation and somatic heterogeneity are established early in development. *Am J Med Genet.* 43:208-16.
Dieringer, D., and C. Schlotterer. 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.* 13:2242-51.
Dietmaier, W., S. Wallinger, T. Bocker, F. Kullmann, R. Fishel, and J. Ruschoff. 1997. Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression. *Cancer Res.* 57:4749-56.
Dornberger, U., M. Leijon, and H. Fritzsche. 1999. High base pair opening rates in tracts of GC base pairs. *J Biol Chem.* 274:6957-62.
Duennwald, M.L., S. Jagadish, P.J. Muchowski, and S. Lindquist. 2006. Flanking sequences profoundly alter polyglutamine toxicity in yeast. *Proc Natl Acad Sci U S A.* 103:11045-50.
Dunah, A.W., H. Jeong, A. Griffin, Y.M. Kim, D.G. Standaert, S.M. Hersch, M.M. Mouradian, A.B. Young, N. Tanese, and D. Krainc. 2002. Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. *Science.* 296:2238-43.
Durr, A., V. Hahn-Barma, A. Brice, C. Pecheux, C. Dode, and J. Feingold. 1999. Homozygosity in Huntington's disease. *J Med Genet.* 36:172-3.
Eichler, E.E., J.J. Holden, B.W. Popovich, A.L. Reiss, K. Snow, S.N. Thibodeau, C.S. Richards, P.A. Ward, and D.L. Nelson. 1994. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat Genet.* 8:88-94.
Eisenberg, D., E. Schwarz, M. Komaromy, and R. Wall. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol.* 179:125-42.
Elgin, S.C., and S.I. Grewal. 2003. Heterochromatin: silence is golden. *Curr Biol.* 13:R895-8.
Ellegren, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet.* 24:400-2.
Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 5:435-45.
Filippova, G.N., C.P. Thienes, B.H. Penn, D.H. Cho, Y.J. Hu, J.M. Moore, T.R. Klesert, V.V. Lobanenkov, and S.J. Tapscott. 2001. CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat Genet.* 28:335-43.
Foiry, L., L. Dong, C. Savouret, L. Hubert, H.T. Riele, C. Junien, and G. Gourdon. 2006. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Hum Genet.* 119:520-6.

Fortune, M.T. 2001. Developmental timing and the role of *cis* and *trans* acting modifiers on CTG repeat instability in murine models. *PhD Thesis*.
Fortune, M.T., C. Vassilopoulos, M.I. Coolbaugh, M.J. Siciliano, and D.G. Monckton. 2000. Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability. *Hum Mol Genet.* 9:439-45.
Freudenreich, C.H., S.M. Kantrow, and V.A. Zakian. 1998. Expansion and length-dependent fragility of CTG repeats in yeast. *Science.* 279:853-6.
Frishman, D., and P. Argos. 1995. Knowledge-based protein secondary structure assignment. *Proteins.* 23:566-79.
Frontali, M. 2001. Spinocerebellar ataxia type 6: channelopathy or glutamine repeat disorder? *Brain Res Bull.* 56:227-31.
Frontali, M., A. Novelletto, G. Annesi, and C. Jodice. 1999. CAG repeat instability, cryptic sequence variation and pathogeneticity: evidence from different loci. *Philos Trans R Soc Lond B Biol Sci.* 354:1089-94.
Gacy, A.M., G. Goellner, N. Juranic, S. Macura, and C.T. McMurray. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell.* 81:533-40.
Gatchel, J.R., and H.Y. Zoghbi. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet.* 6:743-55.
Genschel, J., S.J. Littman, J.T. Drummond, and P. Modrich. 1998. Isolation of MutSbeta from human cells and comparison of the mismatch repair specificities of MutSbeta and MutSalpha. *J Biol Chem.* 273:19895-901.
Geschwind, D.H., S. Perlman, C.P. Figueroa, L.J. Treiman, and S.M. Pulst. 1997. The prevalence and wide clinical spectrum of the spinocerebellar ataxia type 2 trinucleotide repeat in patients with autosomal dominant cerebellar ataxia. *Am J Hum Genet.* 60:842-50.
Glenn, T.C., W. Stephan, H.C. Dessauer, and M.J. Braun. 1996. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. *Mol Biol Evol.* 13:1151-4.
Godde, J.S., and A.P. Wolffe. 1996. Nucleosome assembly on CTG triplet repeats. *J Biol Chem.* 271:15222-9.
Goldfarb, L.G., O. Vasconcelos, F.A. Platonov, A. Lunkes, V. Kipnis, S. Kononova, T. Chabrashvili, V.A. Vladimirtsev, V.P. Alexeev, and D.C. Gajdusek. 1996. Unstable triplet repeat and phenotypic variability of spinocerebellar ataxia type 1. *Ann Neurol.* 39:500-6.
Gomes-Pereira, M. 2002. Genetic and Environmental Modifiers of Somatic Trinucleotide Repeat Dynamics. *PhD Thesis*.
Gomes-Pereira, M., M.T. Fortune, L. Ingram, J.P. McAbney, and D.G. Monckton. 2004. Pms2 is a genetic enhancer of trinucleotide CAG.CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Hum Mol Genet.* 13:1815-25.
Gomes-Pereira, M., M.T. Fortune, and D.G. Monckton. 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. *Hum Mol Genet.* 10:845-54.
Gomes-Pereira, M., and D.G. Monckton. 2004. Chemically induced increases and decreases in the rate of expansion of a CAG*CTG triplet repeat. *Nucleic Acids Res.* 32:2865-72.
Gomes-Pereira, M., and D.G. Monckton. 2006. Chemical modifiers of unstable expanded simple sequence repeats: what goes up, could come down. *Mutat Res.* 598:15-34.
Gorbunova, V., A. Seluanov, D. Mittelman, and J.H. Wilson. 2004. Genome-wide demethylation destabilizes CTG.CAG trinucleotide repeats in mammalian cells. *Hum Mol Genet.* 13:2979-89.
Gordenin, D.A., T.A. Kunkel, and M.A. Resnick. 1997. Repeat expansion--all in a flap? *Nat Genet.* 16:116-8.
Gourdon, G., F. Radvanyi, A.S. Lia, C. Duros, M. Blanche, M. Abitbol, C. Junien, and H. Hofmann-Radvanyi. 1997. Moderate intergenerational and somatic instability of a 55-CTG repeat in transgenic mice. *Nat Genet.* 15:190-2.
Gouw, L.G., M.A. Castaneda, C.K. McKenna, K.B. Digre, S.M. Pulst, S. Perlman, M.S. Lee, C. Gomez, K. Fischbeck, D. Gagnon, E. Storey, T. Bird, F.R. Jeri, and L.J. Ptacek. 1998. Analysis of the dynamic mutation in the SCA7 gene shows marked parental effects on CAG repeat transmission. *Hum Mol Genet.* 7:525-32.
Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science.* 185:862-4.
Gusella, J.F., and M.E. MacDonald. 2000. Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. *Nat Rev Neurosci.* 1:109-15.
Hagelberg, E., I.C. Gray, and A.J. Jeffreys. 1991. Identification of the skeletal remains of a murder victim by DNA analysis. *Nature.* 352:427-9.
Hagerman, P.J., and R.J. Hagerman. 2004. Fragile X-associated tremor/ataxia syndrome (FXTAS). *Ment Retard Dev Disabil Res Rev.* 10:25-30.
Heinig, M., and D. Frishman. 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* 32:W500-2.
Holmes, S.E., E.O. Hearn, C.A. Ross, and R.L. Margolis. 2001. SCA12: an unusual mutation leads to an unusual spinocerebellar ataxia. *Brain Res Bull.* 56:397-403.
Holmes, S.E., E.E. O'Hearn, M.G. McInnis, D.A. Gorelick-Feldman, J.J. Kleiderlein, C. Callahan, N.G. Kwak, R.G. Ingersoll-Ashworth, M. Sherr, A.J. Sumner, A.H. Sharp, U. Ananth, W.K. Seltzer, M.A. Boss, A.M. Vieria-Saecker, J.T. Epplen, O. Riess, C.A. Ross, and R.L. Margolis. 1999. Expansion of a novel CAG trinucleotide repeat in the 5' region of PPP2R2B is associated with SCA12. *Nat Genet.* 23:391-2.
Hornstra, I.K., D.L. Nelson, S.T. Warren, and T.P. Yang. 1993. High resolution methylation analysis of the FMR1 gene trinucleotide repeat region in fragile X syndrome. *Hum Mol Genet.* 2:1659-65.
Igarashi, S., Y. Takiyama, G. Cancel, E.A. Rogaeva, H. Sasaki, A. Wakisaka, Y.X. Zhou, H. Takano, K. Endo, K. Sanpei, M. Oyake, H. Tanaka, G. Stevanin, N. Abbas, A. Durr, E.I. Rogaev, R. Sherrington, T. Tsuda, M. Ikeda, E. Cassa, M. Nishizawa, A. Benomar, J. Julien, J. Weissenbach, G.X. Wang, Y. Agid, P.H. St George-Hyslop, A. Brice, and S. Tsuji. 1996. Intergenerational instability of the CAG repeat of the gene for Machado-Joseph disease (MJD1) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat. *Hum Mol Genet.* 5:923-32.

Ikeda, Y., R.S. Daughters, and L.P. Ranum. 2007. Bidirectional expression of the SCA8 expansion mutation: One mutation, two genes. *Cerebellum*:1-9.

Jacquemont, S., R.J. Hagerman, M. Leehey, J. Grigsby, L. Zhang, J.A. Brunberg, C. Greco, V. Des Portes, T. Jardini, R. Levine, E. Berry-Kravis, W.T. Brown, S. Schaeffer, J. Kissel, F. Tassone, and P.J. Hagerman. 2003. Fragile X premutation tremor/ataxia syndrome: molecular, clinical, and neuroimaging correlates. *Am J Hum Genet.* 72:869-78.

Jansen, G., P.J. Groenen, D. Bachner, P.H. Jap, M. Coerwinkel, F. Oerlemans, W. van den Broek, B. Gohlsch, D. Pette, J.J. Plomp, P.C. Molenaar, M.G. Nederhoff, C.J. van Echteld, M. Dekker, A. Berns, H. Hameister, and B. Wieringa. 1996. Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. *Nat Genet.* 13:316-24.

Jansen, G., P. Willems, M. Coerwinkel, W. Nillesen, H. Smeets, L. Vits, C. Howeler, H. Brunner, and B. Wieringa. 1994. Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in (CTG)n repeat variation and selection against extreme expansion in sperm. *American Journal Of Human Genetics.* 54:575-585.

Jeffreys, A.J., K. Tamaki, A. MacLeod, D.G. Monckton, D.L. Neil, and J.A. Armour. 1994. Complex gene conversion events in germline mutation at human minisatellites. *Nat Genet.* 6:136-45.

Jin, P., and S.T. Warren. 2000. Understanding the molecular basis of fragile X syndrome. *Hum Mol Genet.* 9:901-8.

Jiricny, J. 2006. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol.* 7:335-46.

Jung, J., and N. Bonini. 2007. CREB-binding protein modulates repeat instability in a Drosophila model for polyQ disease. *Science.* 315:1857-9.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577-637.

Kanadia, R.N., K.A. Johnstone, A. Mankodi, C. Lungu, C.A. Thornton, D. Esson, A.M. Timmers, W.W. Hauswirth, and M.S. Swanson. 2003. A muscleblind knockout model for myotonic dystrophy. *Science.* 302:1978-80.

Kaplan, S., S. Itzkovitz, and E. Shapiro. 2007. A Universal Mechanism Ties Genotype to Phenotype in Trinucleotide Diseases. *PLoS Comput Biol.* 3:e235.

Karchin, R., M. Cline, Y. Mandel-Gutfreund, and K. Karplus. 2003. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins.* 51:504-14.

Kartsaki, E., C. Spanaki, M. Tzagournissakis, A. Petsakou, N. Moschonas, M. Macdonald, and A. Plaitakis. 2006. Late-onset and typical Huntington disease families from Crete have distinct genetic origins. *Int J Mol Med.* 17:335-46.

Kelkar, Y.D., S. Tyekucheva, F. Chiaromonte, and K.D. Makova. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* 18:30-8.

Kennedy, L., E. Evans, C.M. Chen, L. Craven, P.J. Detloff, M. Ennis, and P.F. Shelbourne. 2003. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum Mol Genet.* 12:3359-67.

Kennedy, L., and P.F. Shelbourne. 2000. Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Hum Mol Genet.* 9:2539-44.

Klesert, T.R., D.H. Cho, J.I. Clark, J. Maylie, J. Adelman, L. Snider, E.C. Yuen, P. Soriano, and S.J. Tapscott. 2000. Mice deficient in Six5 develop cataracts: implications for myotonic dystrophy. *Nat Genet.* 25:105-9.

Knight, S.J., A.V. Flannery, M.C. Hirst, L. Campbell, Z. Christodoulou, S.R. Phelps, J. Pointon, H.R. Middleton-Price, A. Barnicoat, M.E. Pembrey, and et al. 1993. Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell.* 74:127-34.

Knight, S.J., M.A. Voelckel, M.C. Hirst, A.V. Flannery, A. Moncla, and K.E. Davies. 1994. Triplet repeat expansion at the FRAXE locus and X-linked mild mental handicap. *Am J Hum Genet.* 55:81-6.

Kong, A., D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, and K. Stefansson. 2002. A high-resolution recombination map of the human genome. *Nat Genet.* 31:241-7.

Koob, M.D., M.L. Moseley, L.J. Schut, K.A. Benzow, T.D. Bird, J.W. Day, and L.P. Ranum. 1999. An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nat Genet.* 21:379-84.

Kovtun, I.V., Y. Liu, M. Bjoras, A. Klungland, S.H. Wilson, and C.T. McMurray. 2007. OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. *Nature.* 447:447-52.

Kremer, E.J., M. Pritchard, M. Lynch, S. Yu, K. Holman, E. Baker, S.T. Warren, D. Schlessinger, G.R. Sutherland, and R.I. Richards. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science.* 252:1711-4.

Kruglyak, S., R.T. Durrett, M.D. Schug, and C.F. Aquadro. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A.* 95:10774-8.

Kunkel, T.A., and D.A. Erie. 2005. DNA mismatch repair. *Annu Rev Biochem.* 74:681-710.

Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 157:105-32.

La Spada, A.R., and J.P. Taylor. 2003. Polyglutamines placed into context. *Neuron.* 38:681-4.

Lavedan, C., H. Hofmann-Radvanyi, P. Shelbourne, J.-P. Rabes, C. Duros, D. Savoy, I. Dehaupas, S. Luce, K. Johnson, and C. Junien. 1993. Myotonic dystrophy: size and sex dependent dynamics of CTG meiotic instability, and somatic mosaicism. *American Journal Of Human Genetics.* 52:875-883.

Leclercq, S., E. Rivals, and P. Jarne. 2007. Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics.* 8:125.

Li, G.M. 2008. Mechanisms and functions of DNA mismatch repair. *Cell Res.* 18:85-98.

Li, H., S.H. Li, H. Johnston, P.F. Shelbourne, and X.J. Li. 2000. Amino-terminal fragments of mutant huntingtin show selective accumulation in striatal neurons and synaptic toxicity. *Nat Genet.* 25:385-9.

Li, Y.C., A.B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol.* 11:2453-65.

Lia, A.S., H. Seznec, H. Hofmann-Radvanyi, F. Radvanyi, C. Duros, C. Saquet, M. Blanche, C. Junien, and G. Gourdon. 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. *Hum Mol Genet.* 7:1285-91.

Libby, R.T., D.G. Monckton, Y.H. Fu, R.A. Martinez, J.P. McAbney, R. Lau, D.D. Einum, K. Nichol, C.B. Ware, L.J. Ptacek, C.E. Pearson, and A.R. La Spada. 2003. Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. *Hum Mol Genet.* 12:41-50.

Lin, B., J.M. Rommens, R.K. Graham, M. Kalchman, H. MacDonald, J. Nasir, A. Delaney, Y.P. Goldberg, and M.R. Hayden. 1993. Differential 3' polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. *Hum Mol Genet.* 2:1541-5.

Lin, X., J.W. Miller, A. Mankodi, R.N. Kanadia, Y. Yuan, R.T. Moxley, M.S. Swanson, and C.A. Thornton. 2006. Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. *Hum Mol Genet.* 15:2087-97.

Liquori, C.L., K. Ricker, M.L. Moseley, J.F. Jacobsen, W. Kress, S.L. Naylor, J.W. Day, and L.P. Ranum. 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science.* 293:864-7.

Maciel, P., I. Lopes-Cendes, S. Kish, J. Sequeiros, and G.A. Rouleau. 1997. Mosaicism of the CAG repeat in CNS tissue in relation to age at death in spinocerebellar ataxia type 1 and Machado-Joseph disease patients. *Am J Hum Genet.* 60:993-6.

Mahadevan, M., C. Tsilfidis, L. Sabourin, G. Shutler, C. Amemiya, G. Jansen, C. Neville, M. Narang, J. Barcelo, K. O'Hoy, and et al. 1992. Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science.* 255:1253-5.

Mangiarini, L., K. Sathasivam, A. Mahal, R. Mott, M. Seller, and G.P. Bates. 1997. Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nat Genet.* 15:197-200.

Mankodi, A., E. Logigian, L. Callahan, C. McClain, R. White, D. Henderson, M. Krym, and C.A. Thornton. 2000. Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. *Science.* 289:1769-73.

Mankodi, A., P. Teng-Umnuay, M. Krym, D. Henderson, M. Swanson, and C.A. Thornton. 2003. Ribonuclear inclusions in skeletal muscle in myotonic dystrophy types 1 and 2. *Ann Neurol.* 54:760-8.

Manley, K., T.L. Shirley, L. Flaherty, and A. Messer. 1999. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet.* 23:471-3.

Manto, M.U. 2005. The wide spectrum of spinocerebellar ataxias (SCAs). *Cerebellum.* 4:2-6.

Marsh, J.L., H. Walker, H. Theisen, Y.Z. Zhu, T. Fielder, J. Purcell, and L.M. Thompson. 2000. Expanded polyglutamine peptides alone are intrinsically cytotoxic and cause neurodegeneration in Drosophila. *Hum Mol Genet.* 9:13-25.

Martin, B.M. 1994. Tissue Culture Techniques: An Introduction.

Martorell, L., K. Johnson, C.A. Boucher, and M. Baiget. 1997. Somatic instability of the myotonic dystrophy (CTG)n repeat during human fetal development. *Hum Mol Genet.* 6:877-80.

Martorell, L., D.G. Monckton, J. Gamez, and M. Baiget. 2000. Complex patterns of male germline instability and somatic mosaicism in myotonic dystrophy type 1. *Eur J Hum Genet.* 8:423-30.

Martorell, L., D.G. Monckton, J. Gamez, K.J. Johnson, I. Gich, A.L. de Munain, and M. Baiget. 1998. Progression of somatic CTG repeat length heterogeneity in the blood cells of myotonic dystrophy patients. *Hum Mol Genet.* 7:307-12.

Maruyama, H., S. Nakamura, Z. Matsuyama, T. Sakai, M. Doyu, G. Sobue, M. Seto, M. Tsujihata, T. Oh-i, T. Nishio, and et al. 1995. Molecular features of the CAG repeats and clinical manifestation of Machado-Joseph disease. *Hum Mol Genet.* 4:807-12.

Matilla, A., E.D. Roberson, S. Banfi, J. Morales, D.L. Armstrong, E.N. Burright, H.T. Orr, J.D. Sweatt, H.Y. Zoghbi, and M.M. Matzuk. 1998. Mice lacking ataxin-1 display learning deficits and decreased hippocampal paired-pulse facilitation. *J Neurosci.* 18:5508-16.

Matsuyama, Z., Y. Izumi, M. Kameyama, H. Kawakami, and S. Nakamura. 1999. The effect of CAT trinucleotide interruptions on the age at onset of spinocerebellar ataxia type 1 (SCA1). *J Med Genet.* 36:546-8.

Mendlewicz, J., D. Souery, J. Del-Favero, I. Massat, K. Lindblad, C. Engstrom, D. Van den Bossche, R. Adolfsson, M. Schalling, and C. Van Broeckhoven. 2004. Expanded RED products and loci containing CAG/CTG repeats on chromosome 17 (ERDA1) and chromosome 18 (CTG18.1) in trans-generational pairs with bipolar affective disorder. *Am J Med Genet B Neuropsychiatr Genet.* 128:71-5.

Michalik, A., and C. Van Broeckhoven. 2003. Pathogenesis of polyglutamine disorders: aggregation revisited. *Hum Mol Genet.* 12 Spec No 2:R173-86.

Michlewski, G., and W.J. Krzyzosiak. 2004. Molecular architecture of CAG repeats in human disease related transcripts. *J Mol Biol.* 340:665-79.

Monckton, D.G., M.I. Coolbaugh, K.T. Ashizawa, M.J. Siciliano, and C.T. Caskey. 1997. Hypermutable myotonic dystrophy CTG repeats in transgenic mice. *Nat Genet.* 15:193-6.

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

Moseley, M.L., T. Zu, Y. Ikeda, W. Gao, A.K. Mosemiller, R.S. Daughters, G. Chen, M.R. Weatherspoon, H.B. Clark, T.J. Ebner, J.W. Day, and L.P. Ranum. 2006. Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nat Genet.* 38:758-69.

Mulvihill, D.J., K. Nichol Edamura, K.A. Hagerman, C.E. Pearson, and Y.H. Wang. 2005. Effect of CAT or AGG interruptions and CpG methylation on nucleosome assembly upon trinucleotide repeats on spinocerebellar ataxia, type 1 and fragile X syndrome. *J Biol Chem.* 280:4498-503.

Nucifora, F.C., Jr., M. Sasaki, M.F. Peters, H. Huang, J.K. Cooper, M. Yamada, H. Takahashi, S. Tsuji, J. Troncoso, V.L. Dawson, T.M. Dawson, and C.A. Ross. 2001. Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity. *Science*. 291:2423-8.

Ordway, J.M., S. Tallaksen-Greene, C.A. Gutekunst, E.M. Bernstein, J.A. Cearley, H.W. Wiener, L.S.t. Dure, R. Lindsey, S.M. Hersch, R.S. Jope, R.L. Albin, and P.J. Detloff. 1997. Ectopically expressed CAG repeats cause intranuclear inclusions and a progressive late onset neurological phenotype in the mouse. *Cell*. 91:753-63.

Otten, A.D., and S.J. Tapscott. 1995. Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. *Proc Natl Acad Sci U S A*. 92:5465-9.

Palombo, F., P. Gallinari, I. Iaccarino, T. Lettieri, M. Hughes, A. D'Arrigo, O. Truong, J.J. Hsuan, and J. Jiricny. 1995. GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. *Science*. 268:1912-4.

Panigrahi, G.B., R. Lau, S.E. Montgomery, M.R. Leonard, and C.E. Pearson. 2005. Slipped (CTG)*(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. *Nat Struct Mol Biol*. 12:654-62.

Pearson, C.E., E.E. Eichler, D. Lorenzetti, S.F. Kramer, H.Y. Zoghbi, D.L. Nelson, and R.R. Sinden. 1998a. Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry*. 37:2701-8.

Pearson, C.E., A. Ewel, S. Acharya, R.A. Fishel, and R.R. Sinden. 1997. Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. *Hum Mol Genet*. 6:1117-23.

Pearson, C.E., and R.R. Sinden. 1996. Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*. 35:5041-53.

Pearson, C.E., M. Tam, Y.H. Wang, S.E. Montgomery, A.C. Dar, J.D. Cleary, and K. Nichol. 2002. Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. *Nucleic Acids Res*. 30:4534-47.

Pearson, C.E., Y.H. Wang, J.D. Griffith, and R.R. Sinden. 1998b. Structural analysis of slipped-strand DNA (S-DNA) formed in (CTG)n. (CAG)n repeats from the myotonic dystrophy locus. *Nucleic Acids Res*. 26:816-23.

Perez, M.K., H.L. Paulson, and R.N. Pittman. 1999. Ataxin-3 with an altered conformation that exposes the polyglutamine domain is associated with the nuclear matrix. *Hum Mol Genet*. 8:2377-85.

Perutz, M.F., T. Johnson, M. Suzuki, and J.T. Finch. 1994. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc Natl Acad Sci U S A*. 91:5355-8.

Potts, W., D. Tucker, H. Wood, and C. Martin. 2000. Chicken beta-globin 5'HS4 insulators function to reduce variability in transgenic founder mice. *Biochem Biophys Res Commun*. 273:1015-8.

Quan, F., J. Janas, and B.W. Popovich. 1995. A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1. *Hum Mol Genet*. 4:2411-3.

Ranum, L.P., P.F. Rasmussen, K.A. Benzow, M.D. Koob, and J.W. Day. 1998. Genetic mapping of a second myotonic dystrophy locus. *Nat Genet*. 19:196-8.

Recillas-Targa, F., M.J. Pikaart, B. Burgess-Beusse, A.C. Bell, M.D. Litt, A.G. West, M. Gaszner, and G. Felsenfeld. 2002. Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*. 99:6883-8.

Reddy, P.H., M. Williams, V. Charles, L. Garrett, L. Pike-Buchanan, W.O. Whetsell, Jr., G. Miller, and D.A. Tagle. 1998. Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA. *Nat Genet*. 20:198-202.

Richards, R.I., J. Crawford, K. Narahara, M. Mangelsdorf, K. Friend, A. Staples, M. Denton, S. Easteal, T.A. Hori, I. Kondo, T. Jenkins, A. Goldman, V. Panich, E. Ferakova, and G.R. Sutherland. 1996. Dynamic mutation loci: allele distributions in different populations. *Ann Hum Genet*. 60 (Pt 5):391-400.

Richards, R.I., and G.R. Sutherland. 1992. Heritable unstable DNA sequences. *Nat Genet*. 1:7-9.

Richards, R.I., and G.R. Sutherland. 1994. Simple repeat DNA is not replicated simply. *Nat Genet*. 6:114-6.

Riley, B.E., and H.T. Orr. 2006. Polyglutamine neurodegenerative diseases and regulation of transcription: assembling the puzzle. *Genes Dev*. 20:2183-92.

Robertson, G., D. Garrick, W. Wu, M. Kearns, D. Martin, and E. Whitelaw. 1995. Position-dependent variegation of globin transgene expression in mice. *Proc Natl Acad Sci U S A*. 92:5371-5.

Roig, M., P.R. Balliu, C. Navarro, R. Brugera, and M. Losada. 1994. Presentation, clinical course, and outcome of the congenital form of myotonic dystrophy. *Pediatr Neurol*. 11:208-13.

Rose, O., and D. Falush. 1998. A threshold size for microsatellite expansion. *Mol Biol Evol*. 15:613-5.

Rozanska, M., K. Sobczak, A. Jasinska, M. Napierala, D. Kaczynska, A. Czerny, M. Koziel, P. Kozlowski, M. Olejniczak, and W.J. Krzyzosiak. 2007. CAG and CTG repeat polymorphism in exons of human genes shows distinct features at the expandable loci. *Hum Mutat*. 28:451-8.

Ruggiero, B.L., and M.D. Topal. 2004. Triplet repeat expansion generated by DNA slippage is suppressed by human flap endonuclease 1. *J Biol Chem*. 279:23088-97.

Sanpei, K., H. Takano, S. Igarashi, T. Sato, M. Oyake, H. Sasaki, A. Wakisaka, K. Tashiro, Y. Ishida, T. Ikeuchi, R. Koide, M. Saito, A. Sato, T. Tanaka, S. Hanyu, Y. Takiyama, M. Nishizawa, N. Shimizu, Y. Nomura, M. Segawa, K. Iwabuchi, I. Eguchi, H. Tanaka, H. Takahashi, and S. Tsuji. 1996. Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nat Genet*. 14:277-84.

Sarkar, P.S., B. Appukuttan, J. Han, Y. Ito, C. Ai, W. Tsai, Y. Chai, J.T. Stout, and S. Reddy. 2000. Heterozygous loss of Six5 in mice is sufficient to cause ocular cataracts. *Nat Genet*. 25:110-4.

Sato, T., M. Oyake, K. Nakamura, K. Nakao, Y. Fukusima, O. Onodera, S. Igarashi, H. Takano, K. Kikugawa, Y. Ishida, T. Shimohata, R. Koide, T. Ikeuchi, H. Tanaka, N. Futamura, R. Matsumura, T. Takayanagi, F. Tanaka, G. Sobue, O. Komure, M. Takahashi, A. Sano, Y. Ichikawa, J. Goto, I. Kanazawa, and et al. 1999. Transgenic mice harboring a full-length human mutant DRPLA gene exhibit age-dependent intergenerational and somatic instabilities of CAG repeats comparable with those in DRPLA patients. *Hum Mol Genet*. 8:99-106.

Savouret, C., E. Brisson, J. Essers, R. Kanaar, A. Pastink, H. te Riele, C. Junien, and G. Gourdon. 2003. CTG repeat instability and size variation timing in DNA repair-deficient mice. *Embo J.* 22:2264-73.

Savouret, C., C. Garcia-Cordier, J. Megret, H. te Riele, C. Junien, and G. Gourdon. 2004. MSH2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. *Mol Cell Biol.* 24:629-37.

Schlotterer, C., and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* 20:211-5.

Schols, L., I. Bauer, C. Zuhlke, T. Schulte, C. Kolmel, K. Burk, H. Topka, P. Bauer, H. Przuntek, and O. Riess. 2003. Do CTG expansions at the SCA8 locus cause ataxia? *Ann Neurol.* 54:110-5.

Seznec, H., O. Agbulut, N. Sergeant, C. Savouret, A. Ghestem, N. Tabti, J.C. Willer, L. Ourth, C. Duros, E. Brisson, C. Fouquet, G. Butler-Browne, A. Delacourte, C. Junien, and G. Gourdon. 2001. Mice transgenic for the human myotonic dystrophy region with expanded CTG repeats display muscular and brain abnormalities. *Hum Mol Genet.* 10:2717-26.

Seznec, H., A.S. Lia-Baldini, C. Duros, C. Fouquet, C. Lacroix, H. Hofmann-Radvanyi, C. Junien, and G. Gourdon. 2000. Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. *Hum Mol Genet.* 9:1185-94.

Shaw, J.A., T. Walsh, S.A. Chappell, N. Carey, K. Johnson, and R.A. Walker. 1996. Microsatellite instability in early sporadic breast cancer. *Br J Cancer.* 73:1393-7.

Shelbourne, P.F., C. Keller-McGandy, W.L. Bi, S.R. Yoon, L. Dubeau, N.J. Veitch, J.P. Vonsattel, N.S. Wexler, N. Arnheim, and S.J. Augood. 2007. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. *Hum Mol Genet.* 16:1133-42.

Shelbourne, P.F., N. Killeen, R.F. Hevner, H.M. Johnston, L. Tecott, M. Lewandoski, M. Ennis, L. Ramirez, Z. Li, C. Iannicola, D.R. Littman, and R.M. Myers. 1999. A Huntington's disease CAG expansion at the murine Hdh locus is unstable and associated with behavioural abnormalities in mice. *Hum Mol Genet.* 8:763-74.

Squitieri, F., L. Frati, A. Ciarmiello, S. Lastoria, and O. Quarrell. 2006. Juvenile Huntington's disease: does a dosage-effect pathogenic mechanism differ from the classical adult disease? *Mech Ageing Dev.* 127:208-12.

Squitieri, F., C. Gellera, M. Cannella, C. Mariotti, G. Cislaghi, D.C. Rubinsztein, E.W. Almqvist, D. Turner, A.C. Bachoud-Levi, S.A. Simpson, M. Delatycki, V. Maglione, M.R. Hayden, and S.D. Donato. 2003. Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course. *Brain.* 126:946-55.

Steinbach, P., D. Glaser, W. Vogel, M. Wolf, and S. Schwemmle. 1998. The DMPK gene of severely affected myotonic dystrophy patients is hypermethylated proximal to the largely expanded CTG repeat. *Am J Hum Genet.* 62:278-85.

Takai, D., and P.A. Jones. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A.* 99:3740-5.

Tamaki, K., and A.J. Jeffreys. 2005. Human tandem repeat sequences in forensic DNA typing. *Leg Med (Tokyo).* 7:244-50.

Tanaka, F., M.F. Reeves, Y. Ito, M. Matsumoto, M. Li, S. Miwa, A. Inukai, M. Yamamoto, M. Doyu, M. Yoshida, Y. Hashizume, S. Terao, T. Mitsuma, and G. Sobue. 1999. Tissue-specific somatic mosaicism in spinal and bulbar muscular atrophy is dependent on CAG-repeat length and androgen receptor--gene expression level. *Am J Hum Genet.* 65:966-73.

Telenius, H., B. Kremer, Y.P. Goldberg, J. Theilmann, S.E. Andrew, J. Zeisler, S. Adam, C. Greenberg, E.J. Ives, L.A. Clarke, and et al. 1994. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nat Genet.* 6:409-14.

Thomas, D.C., J.D. Roberts, and T.A. Kunkel. 1991. Heteroduplex repair in extracts of human HeLa cells. *J Biol Chem.* 266:3744-51.

Thornton, C.A., K. Johnson, and R.T. Moxley, 3rd. 1994. Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Ann Neurol.* 35:104-7.

Timchenko, N.A., Z.J. Cai, A.L. Welm, S. Reddy, T. Ashizawa, and L.T. Timchenko. 2001. RNA CUG repeats sequester CUGBP1 and alter protein levels and activity of CUGBP1. *J Biol Chem.* 276:7820-6.

Tomiuk, J., L. Bachmann, C. Bauer, A. Rolfs, L. Schols, C. Roos, H. Zischler, M.M. Schuler, S. Bruntner, O. Riess, and P. Bauer. 2007. Repeat expansion in spinocerebellar ataxia type 17 alleles of the TATA-box binding protein gene: an evolutionary approach. *Eur J Hum Genet.* 15:81-7.

Tsilfidis, C., A.E. MacKenzie, G. Mettler, J. Barcelo, and R.G. Korneluk. 1992. Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy. *Nat Genet.* 1:192-5.

Tybulewicz, V.L., C.E. Crawford, P.K. Jackson, R.T. Bronson, and R.C. Mulligan. 1991. Neonatal lethality and lymphopenia in mice with a homozygous disruption of the c-abl proto-oncogene. *Cell.* 65:1153-63.

Tzagournissakis, M., C.O. Fesdjian, P. Shashidharan, and A. Plaitakis. 1995. Stability of the Huntington disease (CAG)n repeat in a late onset form occuring on the Island of Crete. *Hum Mol Genet.* 4:2239-43.

Ueno, S., K. Kondoh, Y. Kotani, O. Komure, S. Kuno, J. Kawai, F. Hazama, and A. Sano. 1995. Somatic mosaicism of CAG repeat in dentatorubral-pallidoluysian atrophy (DRPLA). *Hum Mol Genet.* 4:663-6.

van den Broek, W.J., M.R. Nelen, G.W. van der Heijden, D.G. Wansink, and B. Wieringa. 2006. Fen1 does not control somatic hypermutability of the (CTG)(n)*(CAG)(n) repeat in a knock-in mouse model for DM1. *FEBS Lett.* 580:5208-14.

van den Broek, W.J., M.R. Nelen, D.G. Wansink, M.M. Coerwinkel, H. te Riele, P.J. Groenen, and B. Wieringa. 2002. Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum Mol Genet.* 11:191-8.

Vo, A.T., F. Zhu, X. Wu, F. Yuan, Y. Gao, L. Gu, G.M. Li, T.H. Lee, and C. Her. 2005. hMRE11 deficiency leads to microsatellite instability and defective DNA mismatch repair. *EMBO Rep.* 6:438-44.

Vowles, E.J., and W. Amos. 2004. Evidence for widespread convergent evolution around human microsatellites. *PLoS Biol.* 2:E199.

Wang, Y.H. 2007. Chromatin structure of repeating CTG/CAG and CGG/CCG sequences in human disease. *Front Biosci.* 12:4731-41.
Wang, Y.H., R. Gellibolian, M. Shimizu, R.D. Wells, and J. Griffith. 1996. Long CCG triplet repeat blocks exclude nucleosomes: a possible mechanism for the nature of fragile sites in chromosomes. *J Mol Biol.* 263:511-6.
Wang, Y.H., and J. Griffith. 1995. Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics.* 25:570-3.
Webster, M.T., and J. Hagberg. 2007. Is there evidence for convergent evolution around human microsatellites? *Mol Biol Evol.* 24:1097-100.
Webster, M.T., N.G. Smith, and H. Ellegren. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci U S A.* 99:8748-53.
West, A.G., and P. Fraser. 2005. Remote control of gene transcription. *Hum Mol Genet.* 14 Spec No 1:R101-11.
West, A.G., M. Gaszner, and G. Felsenfeld. 2002. Insulators: many functions, many mechanisms. *Genes Dev.* 16:271-88.
West, A.G., S. Huang, M. Gaszner, M.D. Litt, and G. Felsenfeld. 2004. Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol Cell.* 16:453-63.
Wheeler, T.M., J.D. Lueck, M.S. Swanson, R.T. Dirksen, and C.A. Thornton. 2007. Correction of ClC-1 splicing eliminates chloride channelopathy and myotonia in mouse models of myotonic dystrophy. *J Clin Invest.* 117:3952-7.
Wheeler, T.M., and C.A. Thornton. 2007. Myotonic dystrophy: RNA-mediated muscle disease. *Curr Opin Neurol.* 20:572-6.
Wheeler, V.C., L.A. Lebel, V. Vrbanac, A. Teed, H. te Riele, and M.E. MacDonald. 2003. Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. *Hum Mol Genet.* 12:273-81.
Wieringa, B. 1994. Myotonic dystrophy reviewed: back to the future? *Hum Mol Genet.* 3:1-7.
Woerner, S.M., M. Kloor, M. von Knebel Doeberitz, and J.F. Gebert. 2006. Microsatellite instability in the development of DNA mismatch repair deficient tumors. *Cancer Biomark.* 2:69-86.
Wohrle, D., I. Kennerknecht, M. Wolf, H. Enders, S. Schwemmle, and P. Steinbach. 1995. Heterogeneity of DM kinase repeat expansion in different fetal tissues and further expansion during cell proliferation in vitro: evidence for a casual involvement of methyl-directed DNA mismatch repair in triplet repeat stability. *Hum Mol Genet.* 4:1147-53.
Wong, L.J., T. Ashizawa, D.G. Monckton, C.T. Caskey, and C.S. Richards. 1995. Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *Am J Hum Genet.* 56:114-22.
Yu, S., M. Pritchard, E. Kremer, M. Lynch, J. Nancarrow, E. Baker, K. Holman, J. Mulley, S. Warren, D. Schlessinger, and A. Et. 1991. Fragile X genotype characterized by an unstable region of DNA. *Science.* 252:1179-1181.
Yu, Z.X., S.H. Li, H.P. Nguyen, and X.J. Li. 2002. Huntingtin inclusions do not deplete polyglutamine-containing transcription factors in HD mice. *Hum Mol Genet.* 11:905-14.
Zimmerman, J.M., N. Eliezer, and R. Simha. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol.* 21:170-201.