Cis-Elements Affecting Disease-Associated Repeat Sequences

by

Katharine Anne Hagerman

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Molecular Genetics
University of Toronto

© Copyright by Katharine Anne Hagerman 2009
Abstract

The expansion of repetitive sequences leads to more than 40 neurological, neurodegenerative and neuromuscular diseases. These diseases are frequently characterized by ongoing DNA repeat instability upon transmission, worsening of disease severity and decreasing age of onset with each successive generation. The mechanism of repeat instability and contribution of repeat instability to disease pathogenesis are unknown. My thesis examines the contribution of cis-elements – sequences around and within repeats as well as surrounding epigenetic environments – to repeat instability, and discusses their possible contribution to repeat diseases.

Here I identify the first cis-element regulating repeat instability, a DNA binding site for a trans factor protein, CTCF. Loss of CTCF binding at the spinocerebellar ataxia type 7 disease locus induces somatic and germline instability in an age- and tissue-specific manner in mice. CTCF protects against instability in an epigenetic manner, and may function at other disease loci also possessing CTCF binding sites near the repeat.

Given that CTCF flanks many repeat loci and is often situated between a replication origin and disease-associated repeat, I assess the role of CTCF on replication and instability at the myotonic dystrophy repeat locus. Templates with CTCF binding sites reduce overall replication efficiency in primate cells that may be partly due to replication fork stalling.
Mutating CTCF binding sites can alter the stability of the repeat depending on the distance from
the origin of replication to the repeat.

Finally I examine chromatinization of (ATTCT)_n repeats from the spinocerebellar ataxia
type 10 locus. These repeats induce very strong nucleosome formation, and at physiological
conditions form even more strongly on (ATTCT)_n repeats with interruptions that are also found
in some patients.

These data contribute to the understanding of repeat instability in the causation of many
diseases, and suggest that the presence of cis-elements at repeat-associated disease loci alter the
behaviour of repeats.
ACKNOWLEDGMENTS

I would like to take this opportunity to thank my supervisor, Dr. Christopher Pearson, for giving me the opportunity to do research in such an exciting field. For providing just the right amount of encouragement after failed experiments, to the many rounds of edits and practice presentations throughout my degree. I thank my committee members Dr. James Ellis and Dr. Stephen Meyn, for their guidance and support over the years. To the Pearson Lab and Program of Genetics and Genome Biology, I thank you for critiquing my findings at lab meetings, and sharing your enthusiasm for our field of research. You taught me the value of scientific collaboration, and provided the perfect environment for me to succeed.

I would like to thank Dr. Johanna Rommens, whose enthusiasm for genetics and science as a whole inspired me to pursue my goal of obtaining a PhD way back in high school. She opened her lab to young, inexperienced, yet eager budding scientists, and invited me to join her lab as a co-operative education high school student. She unknowingly solidified my curiosity in the field of human genetics, and provided me with an exciting introduction to lab research. I truly appreciate her open door policy over the last ten years, and willingness to discuss my work or write a letter of recommendation at a moment’s notice.

I’d like to thank my friends for providing me with both support for, and escape from graduate school when appropriate. To Bizjak, for being some of the only people that could really understand the difficulties of graduate student life, and providing me with more fun than I thought was possible on a Thursday evening. To my roommates, to Steph and to my Sauble Beach friends, I am so thankful to have met each and every one of you. You have helped me accept both the successes and failures that have come throughout graduate school, and provided outside perspectives on my accomplishments.

To all the teachers and professors from Islington J.M.S, Etobicoke Collegiate Institute, University of Western Ontario and University of Toronto, thank you for giving me the education and discipline to get me where I am today in my academic career.

Finally to my family, the Hagerman and Adamson clans, thank you for providing me with both the nature and nurture to get through. To my grandparents, Grace and Frank, Lois and Bert, I wish you were here to see me finally graduate. I know you would be proud, especially those who were former University of Toronto graduates more than 70 years ago. Your support in
every new endeavor I undertook, and attendance at my life’s milestones will not be forgotten. Thank you to my parents, Marilyn and Michael, whose interests in science and constant support in my life have made my transition from naïve student to budding scientist more smooth. Your late night homework help and lessons in the value of networking were truly appreciated, even if it did not seem like it at the time. To my sister, Sarah, who taught me very early on that the greatest tool I had in my arsenal was my brain, and took me under her wing through my life. My family’s support has been constant, and I appreciate it more than you know.
# TABLE OF CONTENTS

Abstract ............................................................................................................................................. ii

Acknowledgments .............................................................................................................................. iv

Table of Contents .............................................................................................................................. vi

List of Tables ...................................................................................................................................... ix

List of Figures .................................................................................................................................... x

List of Abbreviations ........................................................................................................................ xii

**Chapter 1** – Introduction to *cis*-elements affecting disease-associated repeat sequences ......... 1

1.1 – Repeat instability ...................................................................................................................... 2
  1.1.1 – Diseases ............................................................................................................................ 2
  1.1.2 – Somatic instability ........................................................................................................... 2
  1.1.3 – Germline instability and parental bias ............................................................................. 5

1.2 – *Cis*-elements ....................................................................................................................... 6
  1.2.1 – A definition ....................................................................................................................... 6
  1.2.2 – *Cis*-elements within repeats ......................................................................................... 7
  1.2.3 – *Cis*-elements surrounding repeats ............................................................................... 9
  1.2.4 – CpG methylation ............................................................................................................ 11
  1.2.5 – Chromatin environment ................................................................................................ 12

1.3 – CTCF background ................................................................................................................. 14
  1.3.1 – CTCF is a DNA binding protein ...................................................................................... 14
  1.3.2 – Classical function of CTCF binding .............................................................................. 15
  1.3.3 – Paralogue of CTCF, BORIS ............................................................................................ 16
  1.3.4 – CTCF at disease-associated DNA repeat loci ................................................................. 16

1.4 – DNA replication and repeat instability .................................................................................. 17
  1.4.1 – Evidence for replication-mediated instability ................................................................. 17
  1.4.2 – Replication fork stalling ................................................................................................ 20
  1.4.3 – Origins of replication in model systems ......................................................................... 21
  1.4.4 – Replication origins mapped to repeat disease loci ......................................................... 23
  1.4.5 – Origin placement models for replication-induced instability ....................................... 24

1.5 – Transcription and repeat instability ..................................................................................... 27

1.6 – Recombination and repeat instability .................................................................................... 30

1.7 – Repair and repeat instability ............................................................................................... 25
1.8 – Pathogenesis and repeat diseases ................................................................. 31
  1.8.1 – Myotonic dystrophy ........................................................................... 31
  1.8.2 – Spinocerebellar ataxia type 7 ............................................................... 33
  1.8.3 – Spinocerebellar ataxia type 10 ................................................................. 34

1.9 – Thesis outline ......................................................................................... 35

**Chapter 2** – CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner:
  A novel basis for mutational hot spot determination ........................................ 37

2.1 – Abstract ....................................................................................................... 38

2.2 – Introduction ............................................................................................... 38

2.3 – Materials and methods ........................................................................... 40
  2.3.1 – Generation of SCA7-CTCF-I-mut transgenic mice ................................. 40
  2.3.2 – Electrophoretic mobility shift assays ..................................................... 40
  2.3.3 – DNase I footprinting and methylation interference analysis .................. 41
  2.3.4 – DNA methylation sequencing ................................................................. 41
  2.3.5 – Chromatin immunoprecipitation ............................................................. 41
  2.3.6 – Repeat instability analysis ..................................................................... 42

2.4 – Results ....................................................................................................... 42
  2.4.1 – CTCF binds to the SCA7 locus in vitro and in vivo ................................. 42
  2.4.2 – SCA7-CTCF-I-mut mice display increased germ line instability ............. 46
  2.4.3 – SCA7-CTCF-I-mut mice display increased somatic instability .............. 47
  2.4.4 – CpG methylation of the CTCF binding site induces instability .......... 51

2.5 – Discussion ............................................................................................... 60

**Chapter 3** – CTCF induces replication fork stalling at the myotonic dystrophy sequence .... 69

3.1 – Abstract ..................................................................................................... 70

3.2 – Introduction .............................................................................................. 70

3.3 – Materials and methods .......................................................................... 72
  3.3.1 – Design of constructs ........................................................................... 72
  3.3.2 – Cell culture ........................................................................................ 75
  3.3.3 – Replication efficiency .......................................................................... 75
  3.3.4 – 2D gel analysis of replication intermediates ......................................... 76
  3.3.5 – Instability analysis .............................................................................. 76

3.4 – Results ..................................................................................................... 77
  3.4.1 – Generation of DM1 replication templates ............................................. 77
List of Tables

Chapter 1 – Introduction to cis-elements affecting disease-associated repeat sequences
Table 1.1 – Summary of repeat diseases and cis-elements ..........................................................3

Chapter 2 – CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner:
A novel basis for mutational hot spot determination
Table 2.1 – Repeat sizes of cortex DNA: (CAG)\textsubscript{n} tract length – Small-pool PCR.............56

Chapter 4 – The (ATTCT)\textsubscript{n} repeats of spinocerebellar ataxia type 10 display strong nucleosome
assembly which is enhanced by repeat interruptions
Table 4.1 – Competitive nucleosome reconstitution .................................................................99
LIST OF FIGURES

Chapter 1 – Introduction to cis-elements affecting disease-associated repeat sequences
Figure 1.1 – Replication fork progression through repeats can affect repeat stability .....................18
Figure 1.2 – Stability of Trinucleotide Repeats by Individual Product (STRIP) Analysis ..................22
Figure 1.3 – Origin placement models for replication-induced instability .................................25

Chapter 2 – CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner:
A novel basis for mutational hot spot determination
Figure 2.1 – Sequence of the SCA7 CTCF region .......................................................................43
Figure 2.2 – Analysis and mutagenesis of the SCA7-CTCF-I binding site ....................................44
Figure 2.3 – SCA7-CTCF-I-mut mice display increased germ line instability .................................48
Figure 2.4 – Mutation of SCA7-CTCF-I site also abrogates binding by BORIS ...............................50
Figure 2.5 – SCA7-CTCF-I-mut mice display increased somatic instability .....................................52
Figure 2.6 – Increased somatic instability in both SCA7-CTCF-I-mut transgenic lines ....................54
Figure 2.7 – Comparable somatic mosaicism in both SCA7-CTCF-I-mut transgenic lines ............55
Figure 2.8 – Epigenetic regulation of CTCF binding modulates instability at the SCA7 locus ..........................................................................................................................57
Figure 2.9 – Methylation of SCA7-CTCF-I-wt probe fragment for gel shift analysis .....................59
Figure 2.10 – Amplicon for bisulfite sequencing for epigenotype determination .........................61
Figure 2.11 – Epigenotype data for bisulfite sequencing analysis of the CTCF-I binding site region in SCA7-CTCF-I-wt transgenic liver ..................................................62
Figure 2.12 – Model for CTCF regulation of (CAG)_n repeat instability ........................................65
Figure 2.13 – Methylation of the SCA7-CTCF-I site abrogates binding of BORIS as well as CTCF ..................................................................................................................67

Chapter 3 – CTCF induces replication fork stalling at the myotonic dystrophy locus
Figure 3.1 – CTCF binding sites and (CTG)_n repeats from the endogenous DM1 region reduce replication efficiency of primate replication templates ............................................73
Figure 3.2 – CTCF binding sites and possibly (CTG)_n repeats from the endogenous DM1 region stall the replication fork .............................................................80
Figure 3.3 – CTCF binding sites and possibly (CTG)_n repeats from the endogenous DM1 region stall the replication fork (replicate experiments) .................................82
Figure 3.4 – The instability bias of each template used for 2D gel analysis was also analysed by the STRIP assay ..............................................................84

Chapter 4 – The (ATTCT)_n repeats of spinocerebellar ataxia type 10 display strong nucleosome assembly which is enhanced by repeat interruptions
Figure 4.1 – SCA10 genomic fragments .........................................................................................93
Figure 4.2 – Pure and interrupted SCA10 (ATTCT)_n repeats form strong nucleosomes with hypoacetylated, hyperacetylated, and recombinant histones by salt dilution nucleosome reconstitution .................................................................96
Figure 4.3 – Interrupted SCA10 (ATTCT)_n repeats form stronger nucleosomes than pure repeats with hyperacetylated histones by NAP-1 reconstitution ............................100
Figure 4.4 – Pure and interrupted SCA10 (ATTCT)$_n$ repeats form well positioned nucleosomes .................................................................102

**Chapter 5** – Summary and future studies
Figure 5.1 – Summary of findings and hypotheses.................................................................108
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BORIS</td>
<td>Brother of regulator of imprinted sites</td>
</tr>
<tr>
<td>CTCF</td>
<td>CCCTC binding factor</td>
</tr>
<tr>
<td>CTCFL</td>
<td>CTCF-like (or BORIS)</td>
</tr>
<tr>
<td>DM1</td>
<td>Myotonic dystrophy type 1</td>
</tr>
<tr>
<td>DM2</td>
<td>Myotonic dystrophy type 2</td>
</tr>
<tr>
<td>DMD</td>
<td>Differentially methylated domain</td>
</tr>
<tr>
<td>DRPLA</td>
<td>Dentatorubropallidoluysian atrophy</td>
</tr>
<tr>
<td>DUE</td>
<td>DNA unwinding element</td>
</tr>
<tr>
<td>FRAXA</td>
<td>Fragile X syndrome</td>
</tr>
<tr>
<td>FRDA</td>
<td>Friedreich’s ataxia</td>
</tr>
<tr>
<td>FSHD</td>
<td>Facioscapulohumeral muscular dystrophy</td>
</tr>
<tr>
<td>HD</td>
<td>Huntington’s disease</td>
</tr>
<tr>
<td>NAP-1</td>
<td>Nucleosome assembly protein-1</td>
</tr>
<tr>
<td>NI</td>
<td>Nuclear inclusion</td>
</tr>
<tr>
<td>polyQ</td>
<td>Polyglutamine</td>
</tr>
<tr>
<td>RSF</td>
<td>Remodelling and spacing factor</td>
</tr>
<tr>
<td>SBMA</td>
<td>Spinal and bulbar muscular atrophy</td>
</tr>
<tr>
<td>SCA</td>
<td>Spinocerebellar ataxia type</td>
</tr>
<tr>
<td>STRIP</td>
<td>Stability of trinucleotide repeats by individual product</td>
</tr>
<tr>
<td>TNR</td>
<td>Trinucleotide repeat</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
</tbody>
</table>
CHAPTER 1 - INTRODUCTION TO CIS-ELEMENTS AFFECTING DISEASE-ASSOCIATED REPEAT SEQUENCES
1.1 Repeat instability

1.1.1 Diseases

The expansion of repetitive sequences leads to more than 40 neurological, neurodegenerative and neuromuscular diseases (Pearson et al., 2005). Expansions leading to disease have been found in a number of sequences, the most common being trinucleotide repeats (TNRs) such as (CTG)$_n$ and (CGG)$_n$ that are associated with diseases such as myotonic dystrophy type 1 (DM1), spinocerebellar ataxia (SCA) types 1, 2, 3, 6, 7, 8, 12 and 17, Huntington’s disease (HD), fragile X syndrome (FRAXA) and many others. Larger repeat units such as tetranucleotide expansions and pentanucleotide expansions lead to DM2 and SCA10, respectively. Mutation of these repetitive sequences is a dynamic process, whereby the repeat often continues to mutate after reaching and exceeding a certain threshold size of around 35 to 40 repeat units in the case of TNR diseases (Pearson et al., 2005). The dynamic mutations can occur within an individual with some tissues being stable and others continuing to mutate, and upon transmission the repeats usually expand further. Genetic anticipation is common to many repeat diseases. Anticipation is the worsening of disease severity, increased repeat instability and decreased age of onset that occurs upon transmission of the disease-associated repeat to the next generation. Expansions can be found in a number of genic locations, such as the 5’ untranslated region (UTR) in FRAXA, exons in HD and SCA7, introns in DM2 and SCA10, and 3’ UTR in DM1 (Table 1.1) (Pearson et al., 2005). Depending on where in the gene the expansion occurs, it can lead to a number of pathogenic outcomes. Dysregulation of gene expression, splicing defects, protein loss of function, toxic polyglutamine-containing proteins, and gain of function mRNA are a few examples. Though much has been elucidated in these pathogenic pathways, the factors responsible for allowing and even driving the repeat to expand past the threshold size and continue to expand in somatic and germline cells are unknown.

1.1.2 Somatic instability

Repeat expansion diseases vary in the amount of somatic instability seen in different tissues. Possible causes for this variability include cell type-specific trans-factors, DNA metabolic events at the repeat, and cis-elements. Examples of these include chromatin packaging, replication activity, recombination, repair, transcription, and regulation of protein binding around the repeats. The complexity of the interplay between these systems makes understanding somatic
Table 1.1: Summary of repeat diseases and cis-elements

<table>
<thead>
<tr>
<th>Disease</th>
<th>Repeat and Gene</th>
<th>Repeat Size</th>
<th>Symptoms</th>
<th>Interruptions</th>
<th>CTCF sites*</th>
<th>Pathogenic mechanism</th>
<th>Origin Location</th>
<th>Nucleosome formation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM1</td>
<td>CTG DMPK</td>
<td>wt-5 to 37</td>
<td>Myotonia, cataracts, cardiac conduction defects</td>
<td>Rare CCG in wt allele</td>
<td>✓ ✓</td>
<td>RNA gain of function</td>
<td>5' and 3' of CTG</td>
<td>Very strong</td>
</tr>
<tr>
<td>SCA7</td>
<td>CAG ATAXIN-7</td>
<td>wt-7 to 34</td>
<td>Movement disorder and retinopathy</td>
<td>None</td>
<td>✓ ✓</td>
<td>PolyQ gain of function</td>
<td>3' of CAG</td>
<td>Very strong</td>
</tr>
<tr>
<td>SCA10</td>
<td>ATTCT ATAXIN-10</td>
<td>wt-10 to 29</td>
<td>Movement disorder, variable seizure incidence</td>
<td>ATGCT, ATTCTAT, ATTTCT, ATATTCT</td>
<td>✓ ✓</td>
<td>RNA gain of function</td>
<td>Centered on ATTCT</td>
<td>Unknown</td>
</tr>
<tr>
<td>HD</td>
<td>CAG HTT</td>
<td>wt-10 to 34</td>
<td>Chorea, movements, dementia</td>
<td>None</td>
<td>✓ ✓</td>
<td>PolyQ gain of function</td>
<td>5' of CAG</td>
<td>Very strong</td>
</tr>
<tr>
<td>FRAXA</td>
<td>CGG FMR1</td>
<td>wt-6 to 52</td>
<td>Mental retardation</td>
<td>AGG</td>
<td>✓ ✓</td>
<td>Protein loss of function</td>
<td>5' of CGG</td>
<td>Excludes nucleosomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>premutant-59 to 230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mutant-230 to 2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* bolded checkmarks denote confirmed binding, non-bolded are suggested by genome-wide studies
instability very complicated. Depending on the tissue, somatic instability may occur early in embryogenesis or throughout the patient’s life, and can occur to the same degree such that all the repeats expand to a similar size or to varying degrees leading to extensive length heterogeneity within the same tissues. DM1 shows a developmental window for instability during the 13 to 16 week fetal stage, corresponding to the transition from differentiation in the first trimester of fetal development when the tissue repeats are stable, to the rapid fetal growth in the second trimester when somatic instability occurs (Jansen et al., 1994; Wohrle et al., 1995; Martorell et al., 1997). DM1 muscle repeats are heterogeneous and expand over time, but plateau in adulthood (Zatz et al., 1995). In contrast, DM1 blood has little variability within each patient, but expands synchronously throughout the entire life of the patient (Martorell et al., 1998). Within a single DM1 patient, there can often be a difference of more than (CTG)_{4000} repeats in average size between blood and muscle, equating to between 2- and 13-fold size differences (Thornton et al., 1994). DM1 repeats in brain regions vary as well; the cerebral cortex has large expansions of the repeat above the blood repeat size whereas the cerebellar cortex has only small expansions (Ishii et al., 1996). Much less is known about somatic instability in SCA7, and has been ascertained mostly from mouse models. In SCA7 mice somatic instability is seen in the liver and brain, and at very low levels in most other tissues (Libby et al., 2003). Somatic instability in SCA10 is found in blood leukocytes, lymphoblastoid cells and buccal cells, but has not been assayed in any organs (Matsuura et al., 2004). In the same study, blood leukocytes had stable (ATTCT)_{n} repeats within a SCA10 patient over a five year period. Unfortunately, this data was obtained from only 5 patients, some of which were related. Therefore it is difficult to predict whether repeat stability in leukocytes would be expected in other patients because of the small sample size and similar genetic background. It is possible that different cis-elements and inherited trans-factors in unrelated individuals could alter instability. Overall it seems that somatic instability occurs to varying degrees in different tissues. Given that a similar cellular environment is expected in the same tissue of different repeat disease patients, the variability in repeat instability likely stems from either specific molecular pathways affected by mutation of the gene and/or cis-elements near the expansion. Cis-elements – defined in this thesis as variables around and within the disease-associated repeats – such as origins of replication, protein binding sites, methylation, and chromatin environment, have all been implicated in driving repeat instability.
Whether somatic instability contributes to patient symptoms and the progressive nature of the diseases remains a contested argument. The simplest arguments are based on the fact that anticipation in repeat diseases leads to worsening severity of symptoms with larger repeats. Given the variable severity of disease symptoms in an individual, it is possible that two individuals with the same transmitted repeat size could have varying amounts of instability in tissues, leading to symptoms of differing severity. Many pathogenic pathways believed to contribute to disease such as DNA structure formation (Pearson and Sinden, 1996; Pearson et al., 1998b), chromatinization of repeats (Wang and Griffith, 1995) and cleavage of polyglutamine proteins (Graham et al., 2006; Young et al., 2007) have stronger effects with increased repeat size over the threshold. Therefore it is not just the presence of a repeat over the size threshold that determines disease outcome, but also possibly how far past the threshold the repeat expands.

Most arguments for a contribution of ongoing repeat instability to disease are structured around observations in DM1 patients since instability is greatest in the two most affected tissues, brain and muscle. However an argument against an involvement of somatic instability comes from observations in HD and SCA1 brain. Patients with these diseases share similar TNR instability patterns including different brain regions, yet SCA1 has affected Purkinje cells of the cerebellum (Chong et al., 1995) and HD has degeneration of spiny neurons of the striatum (Strong et al., 1993). For now the general consensus in the field is that somatic instability may increase the progression of DM1 in some tissues, but is undetermined in other repeat diseases.

### 1.1.3 Germline instability and parental bias

One hallmark of repeat diseases is the involvement of germline mutations, which often lead to a high frequency of expansions upon transmission. Interestingly, instability varies depending on the parent of origin, suggesting active instability processes specific to spermatogenesis and oogenesis. Variability can arise from differences in the timing and frequency of replication events, availability of repair and recombination proteins, gene expression profiles and epigenetic modifications in the region of interest (reviewed in (Pearson, 2003)). Parental expansion biases differ depending on the disease and sometimes on the initial repeat size of the transmitting parent. DM1 has a paternal expansion bias in the premutation range of (CTG)$_{34-90}$, both paternal and maternal expansion bias in transmission of (CTG)$_{200-600}$, and maternal bias for expansions (CTG)$_{>600}$. In fact, expansion into the congenital form of DM1 resulting from repeats (CTG)$_{>1000}$
is almost exclusively transmitted from the maternal side. Other diseases such as FRAXA and Friedreich’s ataxia (FRDA) have only a maternal bias for expansion, and FRAXA full mutations are almost exclusively transmitted maternally. SCA7 and SCA10 have a paternal expansion bias (David, 1998; Gouw et al., 1998; Matsuura et al., 2004), with juvenile cases of SCA7 being exclusively paternal transmission. Even larger expansions are seen in SCA7 patient sperm, however these expansions are underrepresented in offspring suggesting lethality in embryos or dysfunction of large expansion-containing sperm (Monckton et al., 1999). One exception to the expected traditional definition of anticipation was found in SCA10. Two related SCA10 patients were also observed with anticipation in the form of instability and increased disease severity, however the transmitted alleles were unexpectedly contractions instead of expansions (Matsuura et al., 2004). A paternal contraction bias is found in SCA8, FRAXA and FRDA. Interestingly, full mutation FRAXA males carry only unmethylated premutation alleles (Reyniers et al., 1993). The degree of instability in germ cells can also increase over time. The paternal bias for expansion in HD mice arises through instability in mitotic diploid germ cells throughout life, and increased expansions are seen in older mice (Mangiarni et al., 1997; Zhang et al., 2002; Savouret et al., 2003). Male germline instability is not increased over time in DM1 patients (Martorell et al., 2004), suggesting either differing mouse and human mechanisms of germline instability or differences between the HD and DM1 diseases. Another interesting germline instability event is the rare reversal mutation when a normal allele length is transmitted from a clinically affected parent on the formerly expansion-containing chromosome (Brunner et al., 1993). These rare events are usually the result of contractions, though some instances of exchange of genetic material from the unaffected repeat allele onto the affected chromosome by gene conversion have been reported (Shelbourne et al., 1992).

1.2 Cis-elements

1.2.1 A definition

For the purpose of this thesis, I define “cis-elements” as variables found within the disease-associated DNA repeat region of the chromosome that affect repeat instability in cis. This includes known DNA elements that have a specific function or DNA motifs that bind known factors, epigenetic modifications such as CpG methylation or histone modifications, and structural features of chromatin such as nucleosome positioning or phasing on the DNA. Cis-
elements that affect repeat instability can be both internal within the primary repeat sequence, as well as external to the repeat in surrounding sequences. Internal cis-elements include the nucleotide sequence of the repeat, the size of the repeat tract, and the purity of the sequence. External cis-elements are in flanking sequences including origins of replication, protein binding sites (or motifs), CpG methylation and histone modifications, or chromatin structure. Given the large variation in instabilities, phenotypes, and even sequences causing repeat diseases, it is clear that the complexity of repeat diseases does not stem from a simple mechanism.

1.2.2 Cis-elements within repeats

The internal cis-elements of repeat tract length, purity and primary sequence all contribute to repeat instability. As mentioned previously, there is a relatively consistent repeat threshold of 35 to 40 TNR units associated with a switch from stable to unstable sequences in TNR diseases. As the repeat expands further above this threshold, it becomes increasingly unstable and eventually causes disease. The consistency of this threshold in spinal and bulbar muscular atrophy (SBMA), HD, DRPLA, SCA1, SCA2, SCA7, DM1, FRDA and FRAXE (threshold is 34 +/- 5 repeats; (Pearson et al., 2005)), all of which are TNR diseases, suggests that an inherent feature of the repeat sequence itself gains unstable properties. This is supported by studies showing repeat sequences form secondary DNA structures such as hairpins and loop-outs at repeat sizes above the threshold (Pearson and Sinden, 1996; Pearson and Sinden, 1998). It is hypothesized that above the threshold, there is formation of slipped-DNA structures, and through many metabolic processes slipped intermediate DNA structures with unequal amounts of each repeat strand cause similar structures. Though these structures have not been shown in vivo, there are in vitro studies demonstrating improper resolution of these structures by DNA repair proteins leads to instability (Panigrahi et al., 2005). Therefore repeat length alone is a cis-element that affects repeat instability.

Another internal cis-element affecting repeat instability is the actual expanded nucleotide sequence. There are large differences in the behaviour of different simple repeat sequences, such as the ability to form higher order structure like hairpins. (CTG)$_n$ repeats are more likely to form hairpins than (CAG)$_n$ repeats, and (CAG)$_n$ repeats are more likely to be found in single-stranded loop-outs than hairpins (Pearson and Sinden, 1996). Similarly in DM2, the (CAGG)$_n$ forms structured hairpins whereas the (CCTG)$_n$ does not form paired structures (Dere et al., 2004).
Recently it was shown that the \((CCTG)\_n\) repeat and nearby \((TG)\_n\) repeat in DM2 form Z-DNA and may protect against the formation of slipped-structures associated with mutagenesis (Edwards et al., 2009). \((CGG)\_n\) and \((CCG)\_n\) repeats from FRAXA both form hairpins (Pearson and Sinden, 1996), and \((ATTCT)\_n\) repeats form unpaired regions due to high AT-content with collapsed structures over the threshold repeat size (Potaman et al., 2003). \((CGG)\_n\) repeats can behave differently than \((CAG)\_n\) and \((CTG)\_n\) because \((CGG)\_n\) repeats can be methylated, which in turn can affect many downstream processes. \((CNG)\_n\) sequences are the most likely to form hairpins because they have two of three bases paired, whereas \((GAA)\_n\) repeats and non-TNRs associated with disease do not have as much pairwise binding. Since each sequence has a different propensity to form ordered structures, it is hypothesized this leads to varying levels of instability when the cell tries to resolve the structures by a number of methods.

The final internal \textit{cis}-element affecting repeat instability is purity of the repeat tract. Interruptions in repeat sequences are found in more than half of all repeat diseases, including DM1, DM2, SCA1, SCA2, SCA3, SCA8, SCA10, SCA17, FRAXA, facioscapulohumeral muscular dystrophy (FSHD) and FRDA (Cleary and Pearson, 2003). In some instances like DM1 (Table 1.1), the interruptions appear to be very rare and only documented in one individual (Leeflang and Arnheim, 1995), whereas in others such as SCA1, SCA2 and FRAXA, interruptions occur in most unaffected people with a short, genetically stable tract length (Cleary and Pearson, 2003). Loss of \((AGG)\_n\) interruptions in the \((CGG)\_n\) repeat tract is directly attributed to destabilization of the repeat in FRAXA, and even a single interruption can significantly stabilize the repeat (Eichler et al., 1994). A similar effect is seen in SCA1 where interrupted \((CAG)_{39}\) repeats are stable but pure \((CAG)_{40}\) repeats are unstable (Chong et al., 1995). However not all interrupted repeats are stable, such as in SCA3, SCA8, SCA10 and SCA17, where expanded unstable repeats still maintain interruptions. The mechanism of destabilization by loss of interruptions in TNR diseases is hypothesized to be based on the structure-forming properties of TNR sequences. Pure repeats form more slipped DNA that are believed to contribute to repeat instability, and interrupted repeats have fewer slip-outs (Pearson et al., 1998a). When interrupted, this structure is destabilized and sometimes lost. In a cell these stable structures from pure repeats would attract mismatch repair proteins, however the known mismatch repair proteins are not equipped to deal with slip-outs greater than 13 nucleotides (Genschel et al., 1998; Wilson et al., 1999) and instead induce incorrect repair and instability.
Another structural alteration that occurs with interrupted repeats is altered nucleosome formation or assembly. Pure \((CAG)_n\) repeats associated with SCA1 assemble nucleosomes strongly, however interruptions reduce the nucleosome formation to a level closer to random sequences (Mulvihill et al., 2005). Conversely, \((CGG)_n\) repeats associated with FRAXA strongly exclude nucleosome assembly, and interruption patterns found in patients increase the nucleosome assembly closer to random sequences. Therefore, interruptions can have an effect at the DNA level both by altering secondary structure of the DNA to reduce hairpin formation (Pearson et al., 1998a) as well as by altering the nucleosome formation to make it behave more like randomized sequences. Both effects could lead downstream to stabilization of the repeat.

\subsection{Cis-elements surrounding repeats}
External \textit{cis}-elements have been posited for many years to affect repeat instability. Before disease-causing repeat expansions were identified, common genetic backgrounds were found in disease gene regions. Haplotype analysis of DM1 (Imbert et al., 1993), DM2 (Bachinski et al., 2003), HD (Warby et al., 2009) and SCA3 (Martins et al., 2008) continues to support the involvement of \textit{cis}-sequences in repeat instability. Expanded DM1 and DM2 repeats are mostly from one haplotype and are in linkage disequilibrium, indicative of flanking sequences contributing to instability. Similarly two haplotypes at the SCA3 expansion were found, one associated with further expansions and another with contractions (Martins et al., 2008). Thus sequence context within which repeats are found likely contributes to repeat instability.

Further evidence for flanking \textit{cis}-elements affecting repeat instability comes from knock-in mouse models of TNR disease. The first studies inserted disease-associated repeats into several genomic regions, and resulted in large variation in repeat instability depending on the insertion site (Zhang et al., 2002). In order to more closely mimic the instability pattern seen in patients, mouse models were constructed with longer human sequences including more endogenous flanking sequence (Gourdon et al., 1997b; Seznec et al., 2000; Libby et al., 2003). This instability recapitulation with flanking sequences suggested instability depended on both the size of the repeat and the flanking sequences acting in \textit{cis}. Interestingly, many mouse models still lack the same degree of instability seen with identically sized repeats found in premutation and full mutation DM1 patients, therefore increased complexity still remains to be obtained.
The importance of understanding cis-elements surrounding disease-associated repeats extends beyond attempting to design mouse models that recapitulate human disease, as comparison between repeat-associated diseases in humans also demonstrates cis-element involvement. Each repeat disease has a distinct pattern of instability, and often is not recapitulated in other repeat diseases with the same expansion sequence. For example, the repeat diseases SBMA and SCA7 both occur with expansions greater than ~(CAG)\textsubscript{37} in their respective genes. However SBMA is relatively stable upon transmission with gains of less than (CAG)\textsubscript{2} (Biancalana et al., 1992) and SCA7 has an average gain of (CAG)\textsubscript{12} repeats each generation (Monckton et al., 1999; Lebre and Brice, 2003) and can be found with gains of several hundred. The large variation in repeat instability between diseases suggests that surrounding endogenous sequences play a factor in governing instability.

In order to understand the contribution of cis-elements to repeat instability, the La Spada and Pearson labs collaborated to analyse SCA7 mouse models containing (CAG)\textsubscript{92} repeats in either genomic DNA sequence context of the human ATAXIN-7 locus with exons 3 and 4 as well as intronic sequences, or in entire human cDNA context, and observed drastically different repeat instability patterns. Not only did genomic SCA7 mouse repeats behave in a similar manner to human instability with a propensity to expand upon transmission, but also the cDNA mouse line contrasted this with more stable repeats upon transmission (Libby et al., 2003). The instability differences between mice with genomic DNA versus cDNA insertions suggested there might be elements within intronic sequences driving instability. These observations were made in several different mouse lines with different integration sites, therefore the changes in instability were not due to the surrounding mouse sequences at the integration site but were due to the inherent instability of the human sequence inserted into mice. Mouse lines were then derived with a deletion of sequences 3' of exon 3, leading to significantly fewer length changes and a smaller range of repeat sizes. In the genomic fragment that was removed, there were a number of sequences bearing similarity to replication origins, as well as a previously identified CCCTC binding factor (CTCF) binding site (Filippova et al., 2001). Surprisingly, replication origins have been found at repeat diseases in SCA7, HD, SBMA, FRAXA and SCA10 (Nenguke et al., 2003; Brylawski et al., 2006; Liu et al., 2007), and CTCF binding sites have also been found in DM1, HD FRAXA, SCA2, SCA7, FSHD and dentatorubropallidoluysian atrophy (DRPLA) (Table 1.1) (Filippova et al., 2001; Ladd et al., 2007; Ottaviani et al., 2009). These
findings suggest origins of replication and CTCF binding sites near TNRs may affect instability at the SCA7 locus and possibly several other disease-associated TNR loci.

1.2.4 CpG methylation

DNA methylation and repeat instability are connected, from basic bacterial and mammalian model systems to affected patients with TNR diseases. Bacterial model systems show that CpG methylation of constructs containing repeat sequences such as (TC)$_n$, (CA)$_n$, (GC)$_n$, (CGG)$_n$, and (CTG)$_n$, have a stabilizing effect (Nichol and Pearson, 2002). Clearly not all of these repeats are methylated in the repeat tract, as the SssI methylase protein used in the experiments only methylates CpG sites. Therefore stabilization of the repeat occurred by methylation of flanking sequences around the repeat, protecting against contraction events that are normally seen in bacteria. Methylation of DNA changes the structure-forming properties, presumably because it increases the melting temperature of the DNA (Zacharias, 1993) causing alterations in the formation of cruciforms, Z-DNA, and triplex DNA (Hodges-Garcia and Hagerman, 1992). Methylation can also reduce the formation of nucleosomes on (CGG)$_n$ sequences (Wang and Griffith, 1996a), and may stabilize the repeat to allow methyl-directed strand-specific DNA mismatch repair to reduce structure formation (Wohrle et al., 1995; Smith and Crocitto, 1999). Genome-wide demethylation by treatment with 5-aza-deoxycytidine leads to destabilization, with large contractions in CHO cells and expansions in DM1 patient cells (Gorbunova et al., 2004). The mechanism by which methylation induces repeat instability is unknown.

Methylation is associated with gene silencing, imprinting, X-chromosome inactivation, repression of viral genomes, and silencing of repetitive elements. Gene silencing by methylation is suggested to occur through several mechanisms including blocking of the transcriptional machinery, recruitment of methylation sensitive co-repressors such as histone deacetylases and methyl-CpG-binding proteins that alter chromatin, and preventing some DNA binding proteins that activate or maintain activation of a nearby gene (Klose and Bird, 2006). These methylation-based silencing mechanisms are important during cell-specific activity such as differentiation, and the methylome varies depending on the tissue and developmental period (Shiota et al., 2002; Mathews et al., 2009). Early stages of embryogenesis have coordinated demethylation, including erasure of the paternal imprint in mouse zygotic cells four hours after fertilization and erasure and reprogramming of germ cells at embryonic day 11.5 in mice (reviewed in (Latham et al.,
The pattern of methylation in newly programmed zygotic cells is important to TNR instability and disease as methylation has been implicated in governing instability and activating pathogenic pathways (Verkerk et al., 1991; Nichol Edamura et al., 2005).

Methylation is an important variable in FRAXA; the general population has unmethylated, small, stable (CGG)_n repeats in the FMR1 expressed gene, whereas most FRAXA patients have methylated, silenced genes. There are some examples of apparently clinically unaffected “high functioning” males with expanded repeats, yet varying degrees of methylation and expression of FMR1 (Hagerman et al., 1994). The level of methylation inversely correlates with expression, and is a major determinant of clinical status.

The stability of methylated versus unmethylated FRAXA alleles suggests a connection between instability and methylation. The FRAXA patient repeats are methylated and stable, whereas the “high functioning” males have partially unmethylated, unstable, contraction-prone (CGG)_n repeat (Reyniers et al., 1993; Wohrle et al., 1998). There are also some methylation mosaic FRAXA patients whose methylation status is correlated with somatic contractions (Salat et al., 2000), and a clinically unaffected male with 40% of his full mutation FMR1 genes being unmethylated and premutation sized repeats in sperm (Rousseau et al., 1994). Therefore it appears that FRAXA instability and methylation are intricately linked.

Instability in FRAXA affected embryos is believed to occur at the same embryonic stage in which methylation reprogramming occurs. Whether the expanded repeat is methylated after this stage of reprogramming would be the major determinant of FRAXA disease outcome.

Beyond methylation of (CGG)_n repeats in FRAXA, the repeats surrounding the (CTG)_n repeat of congenital DM1 are also methylated (Steinbach et al., 1998). In fact many disease-causing TNRs are found within large CpG islands (Gourdon et al., 1997a; Brock et al., 1999), and may thus be regulated in part by methylation.

### 1.2.5 Chromatin environment

Nucleosomes are histone octamers bound to DNA. The formation or assembly of nucleosomes is strongly altered by nucleotide sequence of the DNA (Table 1.1). Strong nucleosome assembly occurs at (CTG)_n repeats (Wang et al., 1994; Wang and Griffith, 1995), (GGA)_n repeats (Yoshimura et al., 2002), (A/T)nNN(G/C)nNN repeats (Shrader and Crothers, 1989), (TATAAAACGCC)_n repeats (Widlund et al., 1999), and (TTGA)_n repeats (Cao et al., 1998),
whereas (CGG)$_n$ (Wang et al., 1996), and poly(dA) (Shimizu et al., 2000; Suter et al., 2000) exclude nucleosomes. Telomeres composed of (TTAGGG)$_n$ repeats also exclude nucleosomes (Cacchione et al., 1997), however exclusion only occurs in shortened telomeres less than 7 kb in length (Tommerup et al., 1994). At TNR disease loci, this altered nucleosome assembly is seen \textit{in vivo} with increased protection of DM1 expanded (CTG)$_n$ repeats from DNase I and loss of an adjacent wild type DNase I hypersensitive site (Otten and Tapscott, 1995), as well as nucleosome unwinding and fragile site formation hypothesized to be caused by exclusion of nucleosomes at (CGG)$_n$ repeats in FRAXA, FRAXE, FRAXF, FRA16A and FRA11B (Wang et al., 1996; Lukusa and Fryns, 2008). The FRAXA locus specifically has methylation at the 5' end of the \textit{FMR1} gene and TNR expansion leads to histone deacetylation and chromatin remodelling (Coffee et al., 1999). Nucleosome formation and exclusion can affect metabolic processes such as nucleotide excision repair (reviews by (Lieber, 1999; Green and Almouzni, 2002)), V(D)J recombination (Lieber et al., 1987; Roth and Roth, 2000) and initiation of replication origins (review in (DePamphilis, 2000)). Just as these metabolic processes differ between tissues and developmental stages, so too does nucleosome formation (Felsenfeld, 1996; Gribnau et al., 2000). One can imagine the interplay between chromatinization and gene activation and how it changes throughout a developing higher order organism, with altered nucleosome formation and increased acetylation at the most actively transcribed regions of the genome.

The connection between nucleosome formation and instability of repeats is less clear. The threshold of TNR instability may be connected to nucleosome formation due to similar DNA tract length sizes for the threshold at around 35 to 40 units (105-120 bp) and the size of DNA wrapped around a single nucleosome (146 bp). As previously mentioned, interruptions in repeat sequences are found at several repeat disease loci, and nucleosome formation is altered to more closely resemble random sequences than the pure repeat in (CAG)$_n$ and (CGG)$_n$ repeats (Mulvihill et al., 2005). Interruptions in FRAXA and SCA1 repeat tracts have a stabilizing effect, possibly due to the altered chromatin and DNA secondary structure formation. Chromatinization affects many metabolic processes such as replication, recombination and repair that are also implicated in repeat instability. The complex interplay between methylation, chromatinization and structure formation, along with varied efficiencies of metabolic processes likely contribute to repeat instability.
1.3 CTCF background

1.3.1 CTCF is a DNA binding protein

One cis-element at many disease repeat loci that is gaining attention as a possible regulator of repeat instability is CTCF binding sites. The ubiquitous CTCF protein is a highly versatile 11 zinc-finger protein involved in a wide range of cellular roles including transcriptional control, and nuclear architecture (Dunn and Davie, 2003; Barski et al., 2007). CTCF binds to DNA in a sequence-specific manner using its zinc-finger domains, and has a consensus DNA binding site at approximately 95% of its approximately 20,000 putative binding sites genome-wide as identified by ChIP-on-chip experiments (Kim et al., 2007; Jothi et al., 2008). CTCF is capable of binding to a variety of locations throughout the genome by using different combinations of zinc-fingers to recognize different sites (Filippova et al., 1996). Generally CTCF binds to GC-rich regions (Filippova et al., 1996) especially between differentially regulated gene-rich regions (Barski et al., 2007), though it has been found at AT-rich sequences such as the chicken lysozyme gene (Burcin et al., 1997). Binding is sometimes epigenetically regulated at methylation-sensitive sites such as the differentially methylated domain (DMD) of the IGF2/H19 locus and at the DM1 locus. The number of critical DNA base pairs required for CTCF binding varies between binding sites, likely due to the different zinc-fingers used to bind to each site. For example, mutation of a single base pair on the X-chromosome at a region associated with X-chromosome inactivation choice can either completely ablate CTCF binding with a C(-43)A mutation, or increase the strength of binding with a C(-43)G mutation (Pugacheva et al., 2005). Loss of binding to other sites such as the DM1 locus requires alterations of 7 bp to the site 5' of the (CTG)$_n$ repeat and 10 bp to the site 3' of the repeat, and the CTCF protein protects a 50 bp region when bound (Filippova et al., 2001).

CTCF is a highly conserved protein in metazoans, found from Drosophila and lamprey to humans. The most conserved protein region is the 11 zinc-finger DNA binding domain, which has a 99.5% sequence identity between human and other vertebrate orthologues (Hore et al., 2008). The chicken and Drosophila orthologues, for example, are able to bind the same CTCF target sequences, but can have different effects due to some variation between orthologues in the N- and C-terminal domains which have 90.1% and 80.7% identity in vertebrate orthologues, respectively (Hore et al., 2008). Even between the zinc-fingers of a given CTCF molecule there is a sequence consensus, where DNA base recognition residues lie at the tip of each zinc-finger,
and are 100% conserved from *Drosophila* to humans (Moon et al., 2005). In fact the amino acids responsible for DNA-protein interactions are 100% conserved across vertebrates (Hore et al., 2008). The highly conserved nature of CTCF suggests it is a very important protein for many organisms in the animal kingdom.

### 1.3.2 Classical function of CTCF binding

The classical function of CTCF is as an insulator, and is most thoroughly characterized at the *H19/IGF2* locus. CTCF binds in an epigenetic-dependent, parent-specific manner in the DMD. Here CTCF is believed to protect the *H19* DMD from *de novo* methylation during oocyte growth (Engel et al., 2006). When CTCF binds to the DMD of the maternal allele, it insulates downstream enhancers from inducing transcription of the *IGF2* locus of the maternal allele and instead enhances transcription of *H19*, whereas methylation of the paternal allele prevents CTCF binding, and the enhancers promote *IGF2* transcription. Perturbation of imprinting and altered CTCF binding at this region is associated with Beckwith-Wiedemann syndrome when expression of *H19* is lost from the maternal allele, and *IGF2* is expressed from both parental alleles. CTCF is one of the only known human insulator proteins, though others such as suppressor of Hairy wing protein binding at the gypsy locus have been found in less related organisms such as *Drosophila* (Gdula et al., 1996). More recently approximately 20,000 putative CTCF binding sites were mapped throughout the genome by ChIP on chip experiments, and were mapped predominantly to regions that divide actively transcribed genes from inactive genes (Kim et al., 2007; Jothi et al., 2008). CTCF is thus believed to function in genome-wide insulation and establishment of chromatin domains. Though CTCF binding between tissues of an individual can differ (Hikichi et al., 2003), most CTCF binding sites are invariant across tissues (Heintzman et al., 2009). Though binding may not differ, it is possible that different variants of CTCF could be bound in a tissue-specific and development-specific manner (Delgado et al., 1999). Different isoforms of CTCF are found in subcellular regions (Torrano et al., 2006) and tissues (Klenova et al., 1993), and CTCF has varied post translational modifications such as sumoylation (MacPherson et al., 2009), phosphorylation (Delgado et al., 1999) and poly(ADP-ribosyl)ation, the latter of which has been specifically associated with insulation at the *H19/IGF2* locus (Yu et al., 2004).
1.3.3 Paralogue of CTCF, BORIS

The 11 zinc-finger DNA binding domain of CTCF is also found in another highly similar protein, initially named for this similarity as Brother Of Regulator of Imprinted Sites, or BORIS (also known as CTCF-L). Human BORIS shares a 74% sequence identity with CTCF orthologues in the zinc-finger domain, but only 14% and 10%, respectively, in the N- and C-terminal domains (Hore et al., 2008). Both proteins bind the same sequences in vitro, however expression of the two genes is generally believed to be mutually exclusive in vivo (Loukinov et al., 2002). For the most part BORIS is only expressed in humans in the testis at one stage of spermatogenesis, whereas CTCF is expressed at all other stages and in all other tissues (Loukinov et al., 2002). More recently it was shown that BORIS is expressed in some tissues of platypus and bearded dragon, and led to the suggestion that BORIS evolved from an ancient gene duplication event of CTCF, and became more specialized with functions in imprinting in early therians (Hore et al., 2008). At this time, CTCF and BORIS expression became antagonistic, such that binding at sites like the IGF2/H19 locus is competitive when both proteins are expressed (Hore et al., 2008). The only time BORIS expression is found in humans outside of testis is in cancerous cells. Simultaneous expression of BORIS and CTCF is found in many types of cancers such as sporadic and invasive breast cancer, lung cancer, neuroblastoma, prostate cancer, and colon cancer (Hong et al., 2005; Vatolin et al., 2005), leading to the categorization of BORIS as a cancer-testes gene. CTCF knockouts are lethal in mice (Filippova, 2008), and CTCF gene deletions have only been found in cancerous human cells (Filippova et al., 1998).

1.3.4 CTCF at disease-associated DNA repeat loci

CTCF binding sites have been mapped to several disease-associated repeat loci such as DM1, HD, FRAXA, SCA2, SCA7, FSHD and DRPLA (Filippova et al., 2001; Ladd et al., 2007; Ottaviani et al., 2009), and were suggested at SCA10, SBMA, FRAXE and DM2 by genome-wide mapping studies (Barski et al., 2007; Cleary et al., 2009). Often CTCF binding sites are found on either side of the repeat, and are nearby other interesting cis-elements such as DNase I hypersensitive sites and origins of replication (Cleary et al., 2009). CTCF binding at a disease-associated repeat was first characterized at the DM1 locus by gel mobility shift assays of DNA fragments flanking and within the DM1 (CTG)n repeat (Filippova et al., 2001). Contact residues
responsible for CTCF binding were identified, and were used to design mutations to abrogate CTCF binding. In this same study the CTCF binding sites were shown to confer insulator function, along with the repeat. The function of these CTCF binding sites at other repeat loci is still unknown. However, it has been suggested that loss of CTCF binding in congenital DM1 patients by hypermethylation may contribute to the increased severity of the disease as well as the phenotypes unique to the congenital form (Filippova et al., 2001). Furthermore loss of a number of sequence motifs, including a CTCF binding site, in a SCA7 mouse model was associated with altered repeat instability (Libby et al., 2003). Therefore CTCF is associated with a number of repeat disease loci (Table 1.1), and may be conserved at these sites in order to protect against repeat instability and pathogenic effects.

1.4 DNA replication and repeat instability

1.4.1 Evidence for replication-mediated instability

The cause of repeat instability is hypothesized to be a combination of several malfunctioning metabolic processes including DNA replication. Supporting evidence for replication-induced repeat instability comes from the requirement for proliferation in several models of repeat instability (Cleary et al., 2002; Yang et al., 2003; Freudenreich and Lahiri, 2004), and ongoing instability in proliferating patient cells (Martorell et al., 1998). Also, a large proportion of repeat size variation occurs at the post-zygotic stage in development (Wohrle et al., 1993; Martorell et al., 1997) where rapid cell division occurs with unregulated initiation of replication (Aladjem and Fanning, 2004). Drug treatment of DM1 patient cells with ongoing instability using a replication initiation inhibitor had no effect, however, drugs that inhibit lagging strand synthesis or both leading and lagging strand synthesis caused a significant increase in expansion of only the disease-associated DM1 repeat (Yang et al., 2003). This suggests lagging strand synthesis and the coordination of leading and lagging strand synthesis are important for faithful replication of repeat sequences (Figure 1.1A).

The lagging strand template spends more time in the unstable single-stranded conformation, possibly allowing for increased structure formation time leading to instability. Disruption of Okazaki fragment processing by Rad27 in yeast increases expansion frequency (Kokoska et al., 1998; Schweitzer and Livingston, 1998; Yang and Freudenreich, 2007), though
Figure 1.1 – Replication fork progression through repeats can affect repeat stability (Cleary and Pearson, 2005). (A) Replication fork, including proteins involved in DNA synthesis, nucleosome architecture and demonstration of coordinated leading and lagging strand synthesis. Replication-induced repeat instability is suggested to occur by a number of means, including loss of coordination of leading and lagging strand synthesis, and stalling of the replication fork by DNA binding proteins. (B) Analysis of replication intermediates by 2D gel electrophoresis. Position of various replication fork structures shown along the y-arc and double y-arc. Red arrow denotes hypothetical accumulation of replication fork intermediates on y-arc caused by fork stalling. Graphic design of this figure almost exclusively performed by John Cleary, and included here with permission.
this has not been recapitulated in human cells or TNR model mice (van den Broek et al., 2006; Moe et al., 2008). Okazaki processing is implicated in lagging strand synthesis instability because Okazaki initiation is excluded in some repeats such as (CAG)$_n$ and (CGG)$_n$, and preferred in (CTG)$_n$ and (CCG)$_n$, suggesting that initiation of lagging strand synthesis on (CAG)$_n$ and (CGG)$_n$ repeats would be avoided leading to uncoupling of leading and lagging strand synthesis, and longer stretches between Okazaki initiation zones (Hay et al., 1984). These longer stretches would be single stranded for more time, allowing for increased structure formation and mutagenesis. Increased Okazaki initiations in preferred sequences could also contribute to instability due to exacerbated Okazaki processing errors in repeats. Okazaki fragments form on the lagging strand every 135-145 bp (Anderson and DePamphilis, 1979); a distance similar to the 105-120 bp (35-40 repeat units) of TNRs around the threshold of instability (Pearson et al., 2005). This size similarity may suggest that the threshold is determined by the ability of a cell to successfully replicate repeats on the lagging strand. Altogether the evidence suggests replication, especially lagging strand errors, contribute to repeat instability in dividing cells.

1.4.2 Replication fork stalling

Replication fork stalling is another mechanism hypothesized to contribute to repeat instability. Stalling of the replication fork can occur when the nucleotide pool is depleted, when the DNA template strand incurs damage, or when proteins are bound to the DNA (reviewed in (Pohlhaus and Kreuzer, 2006)). Detection of replication fork stalling is typically observed by 2D gel electrophoresis of replication intermediates (Figure 1.1B). This technique detects increased abundance of specific replication fork structures at various positions along a fragment of DNA. For example, a DNA binding protein in yeast stalls the replication fork as a part of a replication termination system (Brewer and Fangman, 1988). Replication forks also stall at naturally occurring DNA sequences, such as TNRs.

Stalling of the replication fork by TNRs has been shown in vitro with bacterial polymerases at (CTG)$_n$ and (CGG)$_n$ repeats (Kang et al., 1995b), and in vivo in E. coli (Samadashwily et al., 1997). Polymerase stalling at (CGG)$_n$ repeats was abolished with (AGG)$_n$ interruptions, often found in the general population. Stalling at (GAA)$_n$ repeats has also been demonstrated in yeast (Krasilnikova and Mirkin, 2004b). Stalling of the replication fork is direction-dependent, such that stalling occurs more predominantly with (CGG)$_n$ versus (CCG)$_n$,
and (CTG)$_n$ versus (CAG)$_n$ on the lagging strand (Samadashwily et al., 1997). The strength of the stall corresponds with the propensity for the sequence to form hairpins, as (CGG)$_n$>(CCG)$_n$>(CTG)$_n$>(CAG)$_n$ (Gacy et al., 1995; Petruska et al., 1996; Zheng et al., 1996). Replication fork stalling at TNRs in primates was recently demonstrated at (CGG)$_n$ repeats in COS-1 primate cells (Voineagu et al., 2009). Resolution of replication fork stalling by removal of the blockage is believed to be achieved through replication fork reversal, double strand break repair, or recombination (Pearson and Sinden, 1996; Haber, 1999; Michel, 2000; Saintigny et al., 2001). All of these processes might contribute to formation of mutagenic secondary structures and repeat instability. Though replication fork stalling is seen at some unstable sequences, instability and stalling have not yet been directly observed simultaneously.

### 1.4.3 Origins of replication in model systems

Several model systems have been designed to test the role of replication in repeat instability. Bacterial modelling showed expansions occur when (CAG)$_n$ is lagging strand template, whereas contractions occur when (CTG)$_n$ strand is lagging (Kang et al., 1995a). When constructs containing either (CTG)$_n$, (CGG)$_n$, (GTA)$_n$, (GAT)$_n$, (GTG)$_n$, or (TTA)$_n$ repeats were pooled and replicated in bacteria, most instability events were contractions. Of all the total rare expansion events that were observed, the (CTG)$_n$-containing construct contributed the most events, with 88% of all expansions observed, and (GTC)$_n$ and (CGG)$_n$ represented the remaining expanded repeats (Ohshima et al., 1996). In a yeast model system, contraction also occurred when the (CTG)$_n$ strand was the lagging strand template, however the repeat was mostly stable with (CAG)$_n$ as lagging (Freudenreich et al., 1997). A yeast model system designed to select for expansions also showed a bias for instability with (CTG)$_n$ on the lagging strand with an expansion rate of $\sim 1 \times 10^{-5}$ versus $\sim 1 \times 10^{-8}$ with (CAG)$_n$ (Miret et al., 1998). The rate of expansion was very low and the largest expansion was 38 repeats, unlike the high rates of expansion and rare contractions in humans and jumps of several hundred to thousands of repeats. Though both bacterial and yeast systems find a role for replication in repeat instability, clearly the contraction bias and lower rate of instability suggest these systems are not adequate models for human repeat instability.

The Pearson lab developed an experimental model for repeat instability in replicating COS-1 primate cells in order to more adequately model human replication forks (Figure 1.2)
Figure 1.2 – Stability of Trinucleotide Repeats by Individual Product (STRIP) Analysis (Cleary et al., 2002). Summary of the STRIP assay, including several ways to analyse instability in newly replicated plasmids. Black plasmids are input DNAs created through dam+ bacterial preparation; grey plasmids have been replicated in mammalian cells. Dam+ bacteria methylate both strands on adenine residues of GATC sites, a modification sensitive to DpnI digestion. This methylation is not maintained in mammalian cells. DpnI digestion cuts input DNA replicated in bacteria that have not subsequently been replicated by mammalian cells. Graphic design of this figure almost exclusively performed by John Cleary, and included here with permission.
(Cleary et al., 2002). These cells use all the same replication proteins as humans, save for the addition of T-antigen as a helicase. We found that replication direction through the (CTG)$_n$/(CAG)$_n$ repeat, repeat length and location of the SV40 origin relative to the repeat tract all affected instability. The STRIP assay (Stability of Trinucleotide Repeats by Individual Products) was designed to assess the repeat size alterations that occur after primate replication (Figure 1.2), with a variety of constructs containing either (CTG)$_{17}$ or (CTG)$_{79}$ repeats with the SV40 origin located upstream or downstream of the repeat at varying distances. The (CTG)$_{17}$ and (CTG)$_{30}$ constructs were continuously stable, whereas the (CTG)$_{79}$ constructs had significantly different instability patterns depending on the placement of the origin. Constructs with origins ~100 bp from the repeat had predominantly expansions using (CAG)$_n$ as lagging strand template and contractions using (CTG)$_n$. Origins further away from the repeat caused stabilization with (CTG)$_n$ as lagging strand template, whereas the (CAG)$_n$ constructs switched to predominantly contractions at 234 bp and 536 bp from the origin, and both contractions and expansions at 667 bp from the origin. This primate model was the first to demonstrate a bias for expansions, suggesting it was a strong model in which to test other variables of human replication-based instability.

The STRIP method has since been used to analyse instability of (GAA)$_n$ repeats associated with FRDA and (CCTG)$_n$ repeats of DM2, and expansion bias was also seen in both as well as direction-dependent instability with (CCTG)$_n$ repeats (Dere et al., 2004; Rindler et al., 2006). The Pearson lab also analysed the influence of pre-methylating templates with (TC)$_n$, (CA)$_n$, (GC)$_n$, (CGG)$_n$, and (CTG)$_n$ repeats, and found methylation in the SV40 replication system stabilizes the repeat (Nichol and Pearson, 2002). Given repeat instability was observed in this replication-based system, our lab analysed the overall efficiency of replication of the repeat-containing templates compared to non-repeat containing templates. (CGG)$_n$-containing templates had reduced replication efficiency relative to controls, but replication fork stalling assessed by 2D gel electrophoresis did not reveal accumulation of stalled replication forks at the TNR in primate cells (Nichol Edamura et al., 2005). Overall this system is a good model for primate replication, allowing systematic measurement of instability and replication fork analysis.

### 1.4.4 Replication origins mapped to repeat disease loci

With mounting evidence for a role of replication in repeat instability and a clear dependence of
replication direction for expansions, the natural origins in wild type and patient cells were mapped (Table 1.1). Through competitive PCR of nascent DNA, the abundance of newly replicated DNA at different positions along a region of interest can be quantified. This technique was used to locate the SCA7 and HD putative origins downstream and upstream of the disease-associated (CAG) \textsubscript{n} repeat, respectively, in wild type cells and in a rare homozygous HD patient (Nenguke et al., 2003). The SBMA peak was similarly found in wild type cells, however the origin was found directly centered at the (CAG) \textsubscript{n} repeat. The direction of replication through the disease-associated repeat was thus not determined in SBMA, but the lagging strand template is (CAG) \textsubscript{n} for HD and (CTG) \textsubscript{n} for SCA7, with increased instability associated with (CAG) \textsubscript{n} in primate cells (Cleary et al., 2002). Mapping of origin activity around the SCA10 locus revealed little to no activity in wild type cells, but a strong origin was activated in patient cells (Liu et al., 2007), supporting the idea that the (ATTCT) \textsubscript{n} repeat can act as a DNA unwinding element that gains strength with repeat expansions. The FRAXA and FRAXE origins were mapped adjacent to the (CGG) \textsubscript{n} repeat (Brylawski et al., 2006; Chastain et al., 2006; Gray et al., 2007), and the FRAXA region is known to replicate late in the cell cycle and even later when expanded (Torchia et al., 1994). These origins are slightly upstream of the (CGG) \textsubscript{n} repeat and use the (CGG) \textsubscript{n} strand as lagging strand template. (CGG) \textsubscript{n} repeats are more likely to form structures than (CCG) \textsubscript{n} repeats (Mitas, 1997; Yu et al., 1997; Moore et al., 1999; Mirkin and Smirnova, 2002; Lenzmeier and Freudenreich, 2003; Mirkin, 2006), and during lagging strand replication in primate cells are more prone to contractions than (CCG) \textsubscript{n} repeats (Nichol Edamura et al., 2005). Altered replication of expanded (CCG) \textsubscript{n} repeats could be due to the methylation and silencing found in FRAXA patients at the FMR1 repeat, as it has been previously shown that origin usage is dependent upon sequence motifs, nuclear organization, chromatin structure, methylation, transcription, and nucleotide pool (Leffak and James, 1989; Kitsberg et al., 1993; Aladjem and Fanning, 2004). These factors change between patients, tissues, and developmental stages, therefore there may be differences in origin usage that affect instability in different ways.

1.4.5 Origin placement models for replication-induced instability

Currently there are three main models to explain how movement of replication origins affects repeat instability; the origin-switch, origin-shift, and fork-shift models (Figure 1.3) (Cleary et al., 2002; Mirkin and Smirnova, 2002; Cleary and Pearson, 2005). Each of these models suggests
Figure 1.3 – Origin placement models for replication-induced instability (Cleary et al., 2002; Mirkin and Smirnova, 2002; Cleary and Pearson, 2005). (A) The origin-switch model suggests that moving the origin of replication to the opposite side of the repeat induces repeat instability. (B) The origin-shift model suggests that moving the origin closer to or further from the repeat induces instability. (C) The fork-shift model suggests that cis-elements (for example a DNA binding protein such as CTCF) cause a shift in the Okazaki initiation zone (OIZ) position without altering the origin placement, inducing repeat instability. Graphic design of this figure almost exclusively performed by John Cleary, and included here with permission.
that some of the varied instability observed between patients, between tissues, and between developmental stages could be explained by changes in the positioning of the origin. The origin-switch model (Figure 1.3A) postulates that by moving an origin of replication from an upstream position relative to the repeat to a downstream position, or vice versa, instability will be altered. Lagging strand template choice is believed to be the culprit in this variation, as certain sequences such as \((\text{CTG})_n\) repeats are more likely to form stable secondary DNA structures than \((\text{CAG})_n\) repeats that impede replication fork progression, especially on the lagging strand (Pearson and Sinden, 1996; Pearson et al., 1998b). Origin-switch-induced instability was demonstrated in primate cells that had predominantly expansions or contractions depending on the direction of replication through the \((\text{CTG})_n\) repeat (Cleary et al., 2002).

The origin-shift model (Figure 1.3B) postulates that movement of the origin without altering the direction of replication relative to the repeat alters repeat instability. Okazaki initiation zones are the culprit in this variation, as they are excluded in some repeat sequences, possibly forcing initiation sites to be further apart (Hay et al., 1984). These longer distances between initiations would leave larger stretches of single stranded DNA for longer periods of time, allowing for more structure formation and mutagenesis. Origin shifts would need to be close to the repeat because as the origin moves more than 350 bp from the repeat, the placement of Okazaki initiation zones becomes less stringently defined and would not have as predictable instability consequences. Our lab showed that origin shifts directly affect repeat instability in primate cells, as a shift of ~130 bp caused \((\text{CTG})_n\) repeats to go from predominantly expansion events to predominantly contraction events (Cleary et al., 2002). This proof of principle bolsters the hypothesis that cellular differences in the placement of origins could contribute to repeat instability.

The fork-shift model (Figure 1.3C) postulates that \textit{cis}-elements and \textit{cis}-events between the origin and repeat will affect the placement of Okazaki initiation zones to alter instability, independent of the origin location. Examples of altered \textit{cis}-elements and \textit{cis}-events include changes in the intervening sequence, protein binding for cellular processes such as DNA repair, recombination, and transcription, or epigenetic events such as altered methylation, protein binding, and chromatinization. Methylation has been shown to stabilize repeats (Nichol and Pearson, 2002), even when only flanking sequences and not the actual repeat are methylated. The fork-shift model has otherwise not been tested. Even when the origin is further than 350 bp
from the repeat, it is possible that a cis-element may stall the replication fork and sync the Okazaki initiation zones downstream once the stall has been resolved. This would cause directed instability that would not normally be seen when the origin is far away from the repeat.

1.5 Transcription and repeat instability

Not all instances of repeat instability can be explained by replication processes. Some non-dividing neural cells also exhibit repeat instability (Fortune et al., 2000; Pearson et al., 2005). Post-mitotic neural cells show ongoing expansions in HD mouse models, and human HD striatum samples sorted for neural cells suggest ongoing instability occurs in human patients as well (Gonitel et al., 2008). Interestingly the expanded repeats contribute a higher proportion of RNA molecules in these mouse HD striatal cells. Instability in post-mitotic cells suggests other non-replication-based metabolic processes that occur at the repeat such as transcription and repair may be involved. All disease-associated repeats are found within genes and for the most part are transcribed. In FRAXA full mutations, most patients have stable repeats and do not have transcription of the expanded repeat due to CpG methylation and transcriptional silencing.

Like replication, transcription induces unwinding of the DNA and makes it single stranded. This single strandedness would similarly allow for mutagenic intermediates to occur. TNR sequences are also prone to forming structures in RNA such as hairpins and tetraplexes (Michalowski et al., 1999; Tian et al., 2000; Sobczak et al., 2003; Khateb et al., 2007), and are suggested to be toxic on their own, and when bound to proteins that are then sequestered away from their normal function in the cell. Similar to the replication process, bacterial and yeast transcription induce mostly large contractions in a repeat size- and direction-dependent manner (Bowater et al., 1997; Freudenreich et al., 1997; Parniewski et al., 1999; Schumacher et al., 2001; Mochmann and Wells, 2004). Transcription-induced instability was observed in human cells through a contraction-based selection system, where transcription induction causes a 15-fold increase in contractions to a frequency of around $10^{-6}$ (Lin et al., 2006). In these cells neither expansions nor contractions were observed without selection, suggesting the contribution of transcription to repeat instability is relatively low. Knockdown of MSH3 and XPA in this system significantly reduced the contractions, suggesting repair processes on large loops and transcription-coupled repair are involved in the observed transcription-induced contractions.
Some models have had success with connecting transcription to instability such as the mouse HD model where three of four transgenic mouse lines had unstable \((CAG)_n\) repeats and high levels of transcription versus one line that had stable repeats and low levels of transcription (Mangiarini et al., 1997). A *Drosophila* model of SCA3 \((CAG)_n\) repeat instability found that transcription enhanced expansions (Jung and Bonini, 2007). In SBMA the tissues with greatest instability have the highest transcription level (Tanaka et al., 1999). However a DM1 mouse model showed somatic instability did not correlate with transcription levels (Lia et al., 1998). Though other DM1 mouse models have increased somatic instability of the repeat and increased transcription of the *DMPK* gene as the mice age, there is not a direct correlation between the two events across tissues (Seznec et al., 2000; Guiraud-Dogan et al., 2007). Overall transcription is clearly connected to instability in some models, and may contribute to disease-associated repeat instability in patients in varying degrees depending on the disease and cell type.

The role of transcription in repeat instability has become even more complex with the recent identification of bidirectional transcription throughout the genome (Katayama et al., 2005) as well as at TNR loci (Cho et al., 2005; Moseley et al., 2006; Ladd et al., 2007). Unpublished bidirectional transcription models discussed in a recent review suggest that transcription from either upstream or downstream of a repeat induce instability, but the bidirectional convergent transcription produces much more instability than both directions combined (Lin et al., 2008). This suggests collision of transcription machinery can induce instability, and is supported by studies showing collision of the transcriptional machinery, in this case with the replication machinery, induces instability (Liu and Alberts, 1995). Pausing of the transcriptional machinery also occurs at \((CNG)_n\) repeats (Parsons et al., 1998). Therefore slowing, stalling or collision of the transcriptional machinery may correlate with instability in some systems, however a direct connection between transcription and instability has yet to be convincingly demonstrated.

### 1.6 Recombination and repeat instability

Recombination is another metabolic process implicated in repeat instability. Rare instances of chromosomes containing repeats in the pathogenic range of DM1 and FRAXA being transmitted to offspring with a wild type repeat size have been attributed to both contraction and recombination (O’Hoy et al., 1993; van den Ouweland et al., 1994; Losekoot et al., 1997). It is hypothesized that ~1% of all premutation carriers of FRAXA will have recombination-induced
reverted alleles with a wild type size repeat, and therefore screening for the disease-associated haplotype is not sufficient for diagnosis (Brown et al., 1996). One might expect that replication defects could also be blamed for the change from expanded to wild type repeat sizes, however, several TNRs encoding polyalanine proteins were shown to expand due to repeat duplication events because the same repeat interruption pattern is seen in tandem (Goodman et al., 1997; Brais et al., 1998; Nakamoto et al., 2002). Interruptions would also be expected to reduce replication-based instability mechanisms, therefore expansion of these TNRs coding for polyalanine proteins are likely recombination-based. Occurrence of de novo losses of D4Z4 repeats causing FSHD are also believed to be caused by recombinogenic mutations in half of all de novo patients; these patients incurred the mutation after fertilization leading to two predominant repeat sizes throughout somatic tissues (van der Maarel et al., 2000; Lemmers et al., 2004). Therefore recombination is sometimes responsible for unstable transmission of repeats in several diseases.

Examples of recombinogenic loss of repeats are relatively rare. The detection of recombination events is not an easy process, and therefore may not be detected in most cases unless an unexpected loss or gain of repeats alerts researchers to the event. In order to more directly address the role of recombination in repeat instability, studies in bacteria and yeast have focussed on mutating proteins involved in the metabolic process to measure instability. In many knockout systems, no role of recombination in repeat instability was found (Miret et al., 1997; Freudenreich et al., 1998; Savouret et al., 2003). However recently a study of yeast knockouts of proteins responsible for recombination showed that srs2Δ yeast cells undergo active instability with frequent expansions and deletions, and deletion of RAD52 or RAD51 abolishes all expansions and most deletions (Kerrest et al., 2009). A similar loss of instability was seen on the sgs1Δ background as well. The authors hypothesize that replication fork stalling occurs at repeats and requires fork reversal, that loss of SRS2 or SGS1 allows damage to accumulate, and that homologous recombination is required to resolve the damage but triggers repeat rearrangements and instability. They also suggest previous models failed to find this connection due to smaller repeats below the threshold and the use of haploid yeast cells. Though there is likely a role of recombination and repair at stalled replication forks that lead to ongoing contractions in yeast and E. coli (Jakupciak and Wells, 1999; Richard et al., 2000; Hashem et al.,
2004; Kerrest et al., 2009), in humans it appears that recombination is involved only in rare instances of gene conversion and instability of repeats coding for polyalanine proteins.

1.7 Repair and repeat instability

The presence of instability in non-replicating cells such as neural tissues that can sometimes continue over a lifetime (Fortune et al., 2000) suggests that non-replicative metabolic processes are involved in instability. TNR mouse models with various repair gene mutations have altered instability, further strengthening the role of repair in repeat instability (Gomes-Pereira et al., 2001; Savouret et al., 2003; Spiro and McMurray, 2003; Wheeler et al., 2003). Overall the TNR mouse models show that expansions require Msh2, Msh3 and Pms2 for instability in the germline and somatic tissues regardless of replicative status (Manley et al., 1999; van Den Broek et al., 2002; Savouret et al., 2003; Wheeler et al., 2003; Savouret et al., 2004). Another disease with repeat instability is hereditary non-polyposis colorectal cancer. This form of cancer has genome-wide instability of microsatellites due to mutations in mismatch repair proteins, mostly in MLH1 and MSH2 (Quehenberger et al., 2005), some with MSH6 (Goodfellow et al., 2003) and rare mutations in PMS2 (Worthley et al., 2005). Some of these patients have minor repeat instability at DM1-, HD- and FRAXA-associated TNRs (Kramer et al., 1996; Goellner et al., 1997), even when the repeat is below the known threshold on TNR instability. Clearly repair processes are involved in repeat instability.

The wild type function of mismatch repair systems is to recognize mismatched bases and small insertion/deletion loops, excise a region around the mismatch on the newly replicated strand, polymerize across the excised region and ligate the DNA. MutSα is composed of MSH2 and MSH6, and is responsible for recognizing both base-base mismatches and insertion/deletion loops of 1 to 2 bases (Jiricny, 2006), whereas MutSβ is composed of MSH2 and MSH3, and is responsible for recognizing insertion/deletion loops up to 13 bases (Acharya et al., 1996; Wilson and Lieber, 1999). These mismatches occur regularly in healthy cells as a result from errors in replication, transcription, repair of double stranded and single stranded breaks, improper annealing of double stranded DNA and at DNA damage sites. The cell usually detects the mistakes with proofreading proteins, but if the mistake is not fixed before the next round of replication, it may induce a permanent mutation. It is therefore counterintuitive that proteins that have evolved for detection and repair of mismatches are implicated in inducing mutation in TNR
diseases. However large repeats are believed to have a propensity to form slipped structures (Pearson and Sinden, 1996; Sinden et al., 2002; Cleary and Pearson, 2003; Mirkin, 2007), which attract mismatch repair proteins in order to attempt to remove the structure. The Pearson lab has shown that slipped intermediate structures (double stranded DNA with one strand containing 50 TNRs annealed to another containing 30 TNRs) are either correctly repaired to restore the parental repeat size by removal of sequence from the nicked strand, escape repair so that the slip-out is not removed and both repeat sizes are maintained, or repaired in an error-prone manner resulting in incomplete excision and removal of the slip-out to yield heteroduplexed DNA of different sizes (Panigrahi et al., 2005). This process does not require MSH2, MSH3 or MLH1, and thus it is believed that mismatch repair processes may be involved in formation of slipped intermediate, but not in processing.

Much focus has been on recognition of proposed slipped structures by mismatch repair, however other repair processes can contribute to repeat instability including nucleotide excision repair (Kim et al., 1997; Lin and Wilson, 2007), base excision repair (Rada et al., 2002; Kovtun et al., 2007), double strand break repair (Freudenreich et al., 1998; Jankowski et al., 2000; Richard et al., 2000) and transcription-coupled repair (Nouspikel, 2002; Lin et al., 2008). The interplay between these processes and other metabolic processes in the cell makes the elucidation of factors involved in repeat instability very complex. Similar to replication, transcription and recombination, repair can also contribute to repeat instability in a tissue-specific and genome context specific manner. Cis-elements around repeats can affect repair by influencing the structure and stability of the substrates (Fazakerley et al., 1986; Werntges et al., 1986; Fagan et al., 1996), and chromatin structure also alters efficiency of repair (Li et al., 1999). Even similar lengths of \((\text{CTG})_n/(\text{CAG})_n\) TNR DNA from SCA1 and DM1 form biophysically distinct slipped-DNA (Pearson et al., 1997; Pearson et al., 1998a). Therefore endogenous cis-elements around repeats such as protein binding sites and chromatinization can also alter structure formation and repair.

1.8 Pathogenesis of repeat diseases

1.8.1 Myotonic dystrophy

Myotonic dystrophy (DM1) is believed to be caused by pathogenic repeat-containing RNA from the 3' UTR of \(DMPK\). The resulting toxic RNA has a gain of function; a mechanism also seen in
DM2 and FXTAS, and likely in other repeat diseases with expansion in UTRs. DM1 and DM2 have similar clinical symptoms with skeletal muscle wasting, myotonia, cataracts, cardiac conduction defects and insulin resistance, though DM1 has more distal muscle defects and DM2 more proximal. The tight overlap of symptoms supports the RNA gain of function of toxic (CUG)$_n$ and (CCUG)$_n$ RNA as the main contributor to disease symptoms over gene-specific effects of $DMPK$, $ZNF9$ and surrounding genes. $DMPK$ homozygous knockout mice only have mild cardiac conduction and skeletal muscle defects (Berul et al., 2000), and no loss of function mutations have been found in DM1 patients. An HSA$^{LR}$ mouse model overexpressing only the expanded CUG250 RNA in skeletal muscle recapitulates the myotonia symptoms seen in DM patients (Mankodi et al., 2000). Another mouse model, DM300, has a large region of the human $DMPK$ gene and an expanded (CTG)$_{300}$ repeat, and also has myotonia and skeletal muscle abnormalities (Seznec et al., 2000; Seznec et al., 2001; Guiraud-Dogan et al., 2007). This model has more accurate gene expression, splicing patterns, and levels of CUG RNA.

One of the downstream effects of expanded CUG and CCUG transcripts is formation of punctate nuclear RNA foci and co-localization of MBNL protein with the foci (Davis et al., 1997; Mankodi et al., 2001). Sequestration of MBNL at these foci then prevents it from acting in RNA splicing regulatory pathways and leads to misregulation (altered RNA level, splicing or decay) of around 175 transcripts (Osborne et al., 2009). Around 24% of these transcripts are involved in missplicing of Clcn1, leading to defective chloride channel protein that causes myotonia in DM1. Another main splicing defect occurs in transcripts for the insulin receptor and leads to insulin resistance in patients (Savkur et al., 2001). Both splicing defects code for proteins that fail to perform the adult function; CLCN-1 protein is only expressed to 10% of normal levels and IR has the non-muscle splice form in patient muscles (Moxley et al., 1978; Charlet et al., 2002). Of the main splicing defects in DM1, all are defects in regulation that turn off adult splicing patterns and turn on fetal splicing, as opposed to most other spliceopathies where mutation creates new splicing variants (reviewed in (Ranum and Cooper, 2006)).

Expansion of (CTG)$_n$ repeats leads to increased MBNL binding to (CUG)$_n$ repeats and hairpin formation (Napierala and Krzyzosiak, 1997; Miller et al., 2000), therefore increased severity of DM1 may correlate with the amount of MBNL the larger repeats are capable of binding. Another (CUG)$_n$ binding protein, CUG-BP, is overexpressed in DM1 patients (Dansithong et al., 2005), disrupts splicing (Ho et al., 2005), and is believed to antagonize
The toxic DM1 RNA also binds CUGBP, however this binding occurs at the base of the hairpin and is not proportional to the (CUG)$_n$ repeat size (Michalowski et al., 1999). Therefore it is likely not involved in a sequestration mechanism relating to DM1 pathogenesis.

With such a large number of transcripts being misregulated to varying degrees by toxic RNA of different lengths, it is not surprising that DM1 patients have a large spectrum of seemingly unrelated symptoms ranging from typical muscle defects, to premature balding, hypersomnia, gastrointestinal defects and testicular failure, with varying degrees of severity.

1.8.2 Spinocerebellar ataxia type 7
SCA7 is believed to be caused by pathogenic polyglutamine-containing ATAXIN-7 protein. The resulting expanded polyglutamine (polyQ) causes aggregation of the protein in nuclear inclusions (NI) (Garden et al., 2002) with protein chaperone and ubiquitin proteosome components (Holmberg et al., 1998; Takahashi et al., 2002). The resulting protein has a gain of function; a mechanism also seen in the polyQ diseases HD, SBMA, DRPLA, SCA1, SCA2, SCA3, and SCA17. SCA7 is differentiated from other the cerebellar-based movement disorders of SCA diseases by its characteristic retinopathy. There are no Ataxin-7 knockout mouse lines, however there are SCA7 mouse lines with expanded repeats showing motor coordination defects and retinopathy (Yvert et al., 2000; La Spada et al., 2001). Comparisons of gene expression patterns between the polyQ diseases HD, DRPLA, SBMA and SCA7 show that there is overlap between the pathways regulated by expanded polyQ protein expression (Luthi-Carter et al., 2002). Similarly a comparison of HD and SCA7 mouse model expression profiles shows that the shared retinal phenotype results from similar pathogenic pathways, and that neuronal differentiation and maintenance control defects are caused by expanded polyQ proteins independent of the protein context (Abou-Sleymane et al., 2006). Therefore polyQ expansions cause a toxic protein gain of function found in many TNR diseases, leading to similar pathogenic damage of neuronal tissues and death.

The discovery of NI formation in polyQ disease has led to controversial hypotheses about disease pathogenesis. Until recently, many believed that NIs were not only a marker of polyQ disease but also the pathogenic molecule. However studies show the disease phenotype in humans without visible inclusion formation in the most affected brain cells (Kuemmerle et al., 1999). Therefore the current hypothesis is that NI formation may be a cytoprotective way the
cell deals with the toxic proteins, but that the presence of expanded polyQ proteins without inclusions is enough to cause disease. The expanded polyQ protein has a longer half life, and is translocated from the cytoplasm to nucleus in an age-dependent manner (Yvert et al., 2001). The polyQ must exceed the threshold of 35 repeats, upon which it adopts a beta-sheet pleated structure (Perutz et al., 1994; Poirier et al., 2005). The misfolded proteins become more stable, aggregate, and induce a major stress response in the cell (Zander et al., 2001). It is likely not a coincidence that the threshold of repeat instability is around 35 TNRs and toxic protein stability over 35 polyQs. As the disease severity increases with larger repeats, so too would the stability of the protein, and strength of the aggregation by hydrogen bonds between pleated beta-sheets. The ATAXIN-7 protein has caspase cleavage sites that are reminiscent of those seen in polyQ proteins of HD, DRPLA, SCA3 and SBMA, and is truncated in a similar manner to form toxic protein fragments and aggregates (Young et al., 2007). The differences between polyQ diseases may arise from variation in tissue-specific expression patterns of the polyQ protein and abundance of other proteins involved in the degradation and clearance of the protein, as well as post-translational modifications of both.

1.8.3 Spinocerebellar ataxia type 10

SCA10 is hypothesized to be caused by several pathogenic processes including a toxic gain of function of the repeat-containing (AUUCU)ₙ RNA spliced out of intron 9 of ATAXIN-10 (also known as E46L) (Lin and Ashizawa, 2003). A general gain or loss of ATAXIN-10 protein function has been mostly ruled out as patients have the same transcription levels and splicing as unaffected individuals (Wakamiya et al., 2006). A different theory suggests the expansion induces altered binding of neuronal proteins to the DNA, whereas a third suggests altered chromatin structure may cause disregulation of nearby genes (Lin and Ashizawa, 2003). The toxic RNA gain of function theory is supported by studies showing that overexpression of the (ATTCT)ₙ repeat in cell culture induces accumulation of (AUUCU)ₙ-containing RNA foci (Matsuura et al., 2000; Marz et al., 2004; Lin and Ashizawa, 2005); an effect also seen in well established RNA gain of function diseases such as DM1, DM2, FXTAS and HD-like 2. SCA10 is differentiated from the other cerebellar-based movement disorders of SCA diseases by its characteristic seizures and anticipation, though even the preponderance for seizures is not always found with the SCA10 expansion (Grewal et al., 1998; Matsuura et al., 1999; Grewal et al., 2002;
Matsuura et al., 2006). The protein function is also unknown, though it is known to be expressed in brain, testis, kidney, heart, skeletal muscle and is essential for survival in mouse (Matsuura et al., 2000; Marz et al., 2004; Wakamiya et al., 2006). One clue about the disease mechanism is that the preponderance for seizures was seen in patients with interrupted (ATTCT)ₙ repeats, whereas patients with pure repeats did not have seizures (Matsuura et al., 2006). Future studies will need to focus on whether RNA foci are found in patient samples including brain, whether these foci sequester other proteins in a similar manner to other RNA gain of function mechanisms, and whether there is bidirectional transcription of the repeat. The molecular implications of repeat interruptions would also shed light on the contribution of repeats to disease phenotype. These clues may help categorize the pathogenic mechanism of SCA10 with other known repeat diseases. However given the extreme limitations in availability of SCA10 families and patient samples, testing these hypotheses is difficult, leaving the pathogenic mechanism of SCA10 largely unknown.

1.9 Thesis Outline
This thesis tests several hypotheses relating to the effect of cis-elements around and within unstable repeat sequences associated with human disease and the possible molecular mechanisms with which these effects occur. Many cis-elements such as protein binding sites, replication origins around the repeats, and chromatinization of the repeat have been postulated to contribute to human repeat diseases, but have largely remained untested to date. Using primate, mouse, and in vitro models, I have addressed the contribution of the following cis-elements to repeat disease:

1) Through mouse models of SCA7 repeat instability, I analysed the effect of CTCF binding site mutation and methylation on germline and somatic repeat instability over time. I found that binding site mutation induces hyper-instability in an age-dependent manner, seen in expansions and deletions upon transmission and expansions in some tissue types. Methylation also induced similar patterns of instability as mutation to CTCF binding sites.

2) Through primate replication models of DM1, I analysed the effect of CTCF binding sites on replication efficiency and replication fork progression through replication templates containing DM1 repeats and CTCF binding sites. I found that CTCF binding sites reduce replication efficiency in a replication-dependent manner, and induce replication fork stalling.
3) Through \textit{in vitro} nucleosome reconstitution assays, I analysed the effect of SCA10-associated pure and interrupted pentanucleotide sequences on nucleosome formation with variously acetylated histones. I found that (ATTCT)$_n$ repeats are the strongest known nucleosome positioning element, above (CTG)$_n$ repeats, and that interrupted (ATTCT)$_n$ repeats form nucleosomes more strongly than pure repeats with hyperacetylated histones.

Altogether my results suggest that \textit{cis}-elements around and within the repeat can directly affect repeat instability, and processes such as chromatinization and replication indirectly affect instability as well. I discuss implications of my findings on SCA7, DM1 and SCA10 including that CTCF may protect against replication-induced repeat instability at many repeat disease loci, and that chromatinization differences might lead to origin changes and phenotype variability in SCA10 patients.
Co-authorship statement: This work draws upon a collaborative project between the Pearson and La Spada labs. The La Spada lab established the mouse lines, confirmed CTCF binding and performed methylation analysis. The Pearson lab provided all measures of instability. I specifically performed somatic and transmission instability measures, and collaborated with Rachel Lau and Michelle Axford on small-pool PCR analysis of germline and tissue-specific instability. I made figures 2.3, 2.5, 2.6, 2.7 and Table 2.1, and contributed to figures 2.8 and 2.12. Together with Dr. Christopher Pearson, John Cleary and Dr. Al La Spada, we wrote the manuscript that was published in PLoS Genetics (2008) 4: e1000257.
2.1 Abstract

At least 25 inherited disorders in humans result from microsatellite repeat expansion. Dramatic variation in repeat instability occurs at different disease loci and between different tissues; however, cis-elements and trans-factors regulating the instability process remain undefined. Genomic fragments from the human spinocerebellar ataxia type 7 (SCA7) locus, containing a highly unstable (CAG)\textsubscript{92} tract, were previously introduced into mice to localize cis-acting “instability elements,” and revealed that genomic context is required for repeat instability. The critical instability-inducing region contained binding sites for CTCF – a regulatory factor implicated in genomic imprinting, chromatin remodelling, and DNA conformation change. To evaluate the role of CTCF in repeat instability, we derived transgenic mice carrying SCA7 genomic fragments with CTCF binding-site mutations. We found that CTCF binding-site mutation promotes trinucleotide repeat instability both in the germ line and in somatic tissues, and that CpG methylation of CTCF binding sites can further destabilize trinucleotide repeat expansions. As CTCF binding sites are associated with a number of highly unstable repeat loci, our findings suggest a novel basis for demarcation and regulation of mutational hot spots and implicate CTCF in the modulation of genetic repeat instability.

2.2 Introduction

Trinucleotide repeat (TNR) expansion is the cause of at least 25 inherited neurological disorders, including Huntington's disease (HD), fragile X syndrome, and myotonic dystrophy (DM1) (Pearson et al., 2005). One intriguing aspect of TNR disorders is ‘anticipation’ – a phenomenon whereby increased disease severity and decreased age-of-onset are observed as the mutation is transmitted through a pedigree (Harper et al., 1992). In spinocerebellar ataxia type 7 (SCA7), for example, disease onset in children, who inherit the expanded repeat, averages 20 years earlier than in the affected parent (Gouw et al., 1998). The basis of the profound anticipation in SCA7 stems from a significant tendency to undergo large repeat expansions upon parent-to-child transmission (Monckton et al., 1999). Other similarly-sized, disease-linked (CTG)\textsubscript{n}/(CAG)\textsubscript{n} repeat tracts do not exhibit strong anticipation, and are much more stable upon intergenerational transmission, as occurs at the spinobulbar muscular atrophy (SBMA) disease locus (La Spada et al., 1992). Drastic differences in the stability of (CTG)\textsubscript{n}/(CAG)\textsubscript{n} repeats, depending upon the locus at which they reside, strongly support the existence of cis-acting DNA elements that
modulate repeat instability at certain loci. Furthermore, dramatic variation in (CAG)$_n$ tract instability in tissues from an individual patient, together with disparities in the timing, pattern, and tissue-selectivity of somatic instability between (CTG)$_n$/(CAG)$_n$ disorders, indicates a role for epigenetic modification in DNA instability (Thornton et al., 1994; Hashida et al., 1997; La Spada, 1997; Ansved et al., 1998; Pearson et al., 2005). While the existence of cis-elements regulating disease-associated instability is widely accepted, the identities of cis-elements that define the mutability of any repeat are still unknown. Proposed cis-elements and activities that regulate repeat instability include: the sequence of the repeat tract, the length and purity of the repeat tract, flanking DNA sequences, surrounding epigenetic environment, replication origin determinants, trans-factor binding sites, and transcriptional activity (Sinden, 2001; Jung and Bonini, 2007; Mirkin, 2007). Such cis-elements may enhance or protect against (CAG)$_n$ tract instability.

To identify cis-elements responsible for (CAG)$_n$ expansion at the SCA7 locus, we previously introduced SCA7 (CTG)$_{92}$ repeat expansions into mice, either on 13.5 kb ATAXIN-7 genomic fragments or on ATAXIN-7 cDNAs. Comparison of (CAG)$_n$ repeat length change revealed that ATAXIN-7 genomic context drives repeat instability with an obvious bias toward expansion, while SCA7 (CAG)$_{92}$ repeats introduced on ATAXIN-7 cDNAs were stable (Libby et al., 2003). To localize the cis-acting elements responsible for this instability tendency, we derived lines of transgenic mice based upon the original 13.5 kb ATAXIN-7 genomic fragment, deleting a large region (~8.3 kb) of human sequence beyond the 3′ end of the (CAG)$_{92}$ tract (α-SCA7-92R construct). As deletion of the 3′ region in the α-SCA7-92R transgenic mice significantly stabilized the (CAG)$_{92}$ tract (Libby et al., 2003), we hypothesized that cis-elements within this 3′ region modify repeat instability at the SCA7 locus. To identify cis-acting instability elements at the SCA7 locus and the trans-acting proteins that regulate them, we evaluated the critical genomic region 3′ to the (CAG)$_{92}$ repeat for sequences that might regulate genetic instability. In the case of SCA7 and a number of other highly unstable (CAG)$_n$/(CTG)$_n$ repeat loci, including HD, DM1, SCA2, and dentatorubral-pallidoluysian atrophy, binding sites for a protein known as CTCF (i.e. the “CCCTC binding factor”) have been found (Filippova et al., 2001). CTCF is an evolutionarily conserved zinc-finger DNA binding protein with activity in chromatin insulation, transcriptional regulation, and genomic imprinting (Lobanenkov et al., 1990; Ohlsson et al., 2001). As CTCF affects higher order chromatin structure (Ling et al.,

39
we wondered if CTCF binding at the SCA7 locus might regulate \((\text{CAG})_n\) repeat instability. To test this hypothesis, we derived SCA7 genomic fragment transgenic mice with CTCF binding site mutations, and found that impaired CTCF binding yielded increases in both intergenerational and somatic instability at the SCA7 locus. Detection of increased somatic instability in association with hypermethylation of the CTCF binding site indicated a role for epigenetic regulation of SCA7 \((\text{CAG})_n\) repeat stability. Our results identify CTCF as an important modifier of repeat instability in SCA7, and suggest that CTCF binding may influence repeat instability at other tandem repeat expansion disease loci.

2.3 Materials and methods

2.3.1 Generation of SCA7-CTCF-I-mut transgenic mice

To derive the SCA7-CTCF-I-mut transgenic construct, we synthesized a PCR primer with randomly mutated nucleotides introduced at the CTCF-I contact sites for recombineering into the RL-SCA7-92R (SCA7-CTCF-I-wt) construct (Libby et al., 2003), and then confirmed loss of CTCF binding by the mutated fragment by electrophoretic mobility shift assay (protocol provided below). Using a standard recombineering approach (Lee et al., 2001), we PCR-generated a SCA7-CTCF-I targeting cassette containing a Chloramphenicol resistance gene and \(\text{ClaI}\) restriction site flanked by SCA7-CTCF-I region sequences with the following primer set: hSCA7-wt-CAM-F, 5'-tccccccctgcccccctctctgtctagtttaagggacaccaataactgc-3' and hSCA7-mut-CAM-R, 5'-catctctgcccctctctctctctgtctagtttaagggacaccaataactgc-3'. After recombineering the SCA7-CTCF-I-mut targeting cassette into the SCA7-CTCF genomic fragment carried on a plasmid, selection, and PCR screening, we deleted the Chloramphenicol gene by \(\text{ClaI}\) digestion and ligation. We verified the sequence of the SCA7-CTCF-I-mut construct prior to linearization with \(\text{SalI}\) – \(\text{SpeI}\) digestion, gel purification, and microinjection into C57BL/6J×C3H/HeJ oocytes. Transgene-positive founders were backcrossed onto the C57BL/6J background for more than 12 generations to yield incipient congenic mice before repeat instability analysis commenced. All experiments and animal care were performed in accordance with the University of Washington IACUC guidelines.

2.3.2 Electrophoretic mobility shift assays

We amplified a 161 bp DNA fragment (SCA7-CTCF-I) from the SCA7 locus with primers (5'-
ctccccctcaccctcctcgagac-3' & 5'-gtgacgcacctcagcagcgg-3') labelled at their 5' ends by $\gamma$-$^{32}$P-ATP. We gel-purified the 5' end-labelled fragment, and used it for electrophoretic mobility shift assays, with \textit{in vitro} translated proteins, as previously described (Filippova et al., 2001). We synthesized the CTCF 11 zinc-finger (ZF) DNA binding domain, full length CTCF and full length CTCFL/BORIS proteins using the pCITE-11ZF, pCITE-7.1, and pCITE-BORIS expression constructs (Filippova et al., 2001; Loukinov et al., 2002; Vatolin et al., 2005), with the TnT reticulocyte lysate coupled \textit{in vitro} transcription-translation system (Promega). For “super-shifts”, we used an anti-CTCF antibody (Upstate Biotechnology) or anti-BORIS antibody (Loukinov et al., 2002; Vatolin et al., 2005). We methylated the end-labelled SCA7-CTCF-I fragment with SsSI methyl-transferase (New England Biolabs) in the presence of 0.8 mM S-adenosylmethionine. We confirmed the methylation status by restriction enzyme digestion with NruI, and used unmethylated fragment as a control (Filippova et al., 2001).

2.3.3 \textbf{DNase I footprinting and methylation interference analysis}

We PCR-amplified the SCA7-CTCF-I fragment and labelled it at the 5' end on either the coding or anti-sense strand, incubated the purified probes with CTCF and then partially digested them with DNase I, or partially methylated them at guanine residues with dimethyl sulfate, and then incubated them with CTCF. Details of these protocols, as well as our methods for isolation and analysis of free probe DNA fragments on sequencing gels, have been described (Filippova et al., 2001).

2.3.4 \textbf{DNA methylation sequencing}

Bisulfite treatment of tissue DNAs was done as previously described (Laird et al., 2004), and PCR primers spanning the SCA7-CTCF-I region were designed so that they excluded CpG dinucleotides within the binding region. PCR products were then cloned into a Topo TA vector and sequenced. Sequencing of positive control samples, treated with SssI to methylate all cytosines in CpG dinucleotides, were included in every run, and revealed lack of C to T conversion at all CpG dinucleotides in all control samples analysed.

2.3.5 \textbf{Chromatin immunoprecipitation}

We prepared tissues, cross-linked proteins to DNA, and processed tissue samples essentially as
we have done previously (Chen et al., 2004). However, we doubled the length of the sonication step, and, prior to immunoprecipitation, we fractionated supernatant DNAs on agarose gels to gauge the extent of shearing. After confirming that the bulk of sheared DNAs migrated in the 500-1,000 bp range, we performed immunoprecipitation with an anti-CTCF antibody (Upstate Biotechnology), as described (Filippova et al., 2001). DNAs were isolated and then subjected to real-time qPCR analysis with different SCA7 genomic region primer and probe sets (available upon request) on an ABI-7700 sequence detection system. For each CTCF ChIP sample, we normalized SCA7 locus occupancy to a control region of the Myc locus lacking CTCF binding sites (Filippova et al., 2001). All primer and probe sequence sets are available upon request.

2.3.6 Repeat instability analysis
We PCR-amplified the SCA7 (CAG)_n repeat from genomic DNA samples in the presence of 0.1µCi of α-32P-ATP, and resolved the radiolabelled PCR products on 1.8% agarose gels (Libby et al., 2003). For small-pool PCR, dilution of genomic DNAs, yielding 1-5 genome equivalents, was performed prior to amplification and sizing (Monckton et al., 1999). In all experiments, ≥3 mice/genotype or 3 samples/time point were analysed. Primer sequences are available on request.

2.4 Results
2.4.1 CTCF binds to the SCA7 locus in vitro and in vivo
At the SCA7 locus, there are two CTCF binding sites that flank the (CAG)_n repeat tract; the CTCF-I binding site is located 3’ to the (CAG)_n repeat (Figure 2.1), within the critical region deleted from the SCA7 genomic fragment in the α-SCA7-92R mice (Figure 2.2A). As CTCF binding sites are associated with highly unstable repeat loci (Filippova et al., 2001), and CTCF binding can alter chromatin structure and DNA conformation (Ling et al., 2006; Filippova, 2008), we hypothesized that CTCF binding might be involved in SCA7 repeat instability. To test this hypothesis, we decided to compare SCA7 (CAG)_n repeat instability in mice carrying either the wild-type CTCF binding site or a mutant CTCF binding site that would be incapable of binding CTCF. To define the CTCF binding sites, we performed electrophoretic mobility shift assays to confirm that CTCF protein specifically binds to the putative CTCF-I binding site, and we found that both the CTCF DNase I binding domain truncated protein and full-length CTCF protein bind to the SCA7 repeat locus 3’ region (Figure 2.2B). When we mapped the CTCF-I
Figure 2.1 – Sequence of the SCA7 CTCF region (from Libby et al. (Libby et al., 2008)). Primary sequence for the 3' end of intron 2, all of exon 3, and the 5' end of intron 3 are shown. Intron sequence is lowercase; exon sequence is uppercase. CTCF binding sites are shown in blue. Note that the CTCF-I binding site is located in intron 3, while the CTCF-II binding site encompasses intron 2 - exon 3 boundary. Start site of translation is underlined in blue, and (CAG)$_n$ repeat is shown in red. Mapped contact regions from methylation interference and DNase I footprinting analysis are indicated by filled circles, and DNase I hypersensitive bases are marked by arrows (see Figure 2.2C). The primer sequences for generation of the probe fragment for all electrophoretic mobility shift assays are underlined in black.
Figure 2.2 – Analysis and mutagenesis of the SCA7-CTCF-I binding site (from Libby et al. (Libby et al., 2008)). (A) SCA7 genomic fragments used for transgenesis. Upper: SCA7-CTCF-I-wt; Middle: α-SCA7 3’ genomic deletion; Bottom: SCA7-CTCF-I-mut. Core CCCTC sequences are underlined, and sequence alterations in the SCA7-CTCF-I-mut transgenic construct are shown in gray. (B) Electrophoretic mobility shift assays with SCA7-CTCF-I-wt and -mut probe fragments were performed with probe only, empty lysate (no protein), full-length CTCF protein with pre-immune anti-CTCF sera (CTCF+pI), CTCF protein with anti-CTCF sera (CTCF+α-CTCF), or the 11 zinc-finger DNA binding domain region of CTCF. Arrows indicate shifted CTCF-DNA complexes. Addition of CTCF-DM1 probe as cold competitor prevented CTCF-DNA complex formation for SCA7-CTCF-I-wt fragment, while non-specific cold competitor did not (data not shown). (C) Methylation interference (Me I) and DNase I footprinting (DNase) on SCA7-CTCF-I fragment. Left and right panels correspond to the 5’-end labelled coding and antisense strands respectively. B, CTCF-bound DNA; F, free DNA; long bars, CTCF-protected from DNase I; arrows, DNase I hypersensitive bases induced by CTCF binding; filled circles, contact guanine nucleotides essential for sequence recognition by CTCF. See panel ‘A’ for precise location of sites. (D) ChIP on cerebellar lysates from SCA7-CTCF-I-wt and -mut mice (n = 3/genotype). Significantly decreased occupancy at the CTCF-I site was detected with the 3’ amplicon (primer set B) in SCA7-CTCF-I-mut mice (p = 0.02, one-way ANOVA), as this amplicon is not in close proximity to the 5’ CTCF-II site. No differences in CTCF occupancy between SCA7-CTCF-I-wt and -mut mice were detected with primer set A (or other adjacent primer sets; data not shown) due to the close proximity of the two CTCF binding sites. Results are normalized to SCA7-CTCF-I-wt. Error bars are standard deviation.
contact regions at the SCA7 repeat locus by methylation interference and DNA footprinting, we
defined a region that is protected from DNase I treatment upon CTCF binding and subject to
altered CTCF binding upon methylation treatment (Figure 2.2C). We then introduced point
mutations at 11 nucleotides within this 3' CTCF-I binding site, including eight contact
nucleotides contained within the footprinted region (Figure 2.2C; Figure 2.2A, bottom). After
confirming that CTCF binding was abrogated by these point mutations in electrophoretic
mobility shift assays (Figure 2.2B), we derived a RL-SCA7 94R 13.5 kb genomic fragment
construct, that was identical to our original RL-SCA7 92R genomic fragment construct
(Libby et al., 2003), except for: i) the presence of a mutant CTCF-I binding site, and ii) a minor repeat size
increase to (CAG)$_{94}$ repeats. The RL-SCA7 94R CTCF-I-mutant construct was microinjected,
and two independent lines of RL-SCA7 94R CTCF-I mutant transgenic mice were generated
(hereafter referred to as the $SCA7$-$CTCF-I$-mut line mice – to distinguish them from the original
RL-SCA7-92R transgenic mice with an intact CTCF-I binding site, hereafter referred to as the
$SCA7$-$CTCF-I$-wt line mice).

To assess in vivo occupancy of the CTCF-I binding site in $SCA7$-$CTCF-I$-wt and $SCA7$-$CTCF-I$-mut mice, we performed chromatin immunoprecipitation (ChIP) assays. To distinguish
between the two CTCF binding sites, separated by a distance of 562 bp, we used two primer sets,
including one extending 3' to the (CAG)$_{94}$ repeat. Quantitative PCR amplification with a primer
set (‘A’) within ~800 bp of the CTCF-I and CTCF-II sites yielded comparable CTCF occupancy
in $SCA7$-$CTCF-I$-wt and -mut mice. As most sheared DNA fragments isolated by ChIP exceed 1
kb, intact CTCF-II sites and the primer set ‘A’ amplicon will be present in sheared DNA
fragments isolated by ChIP from $SCA7$-$CTCF-I$-wt and -mut mice, accounting for comparable
CTCF occupancy with primer set A. However, a significant reduction in CTCF occupancy at the
CTCF-I site was observed in the $SCA7$-$CTCF-I$-mut mice for primer set ‘B’, which is closer to
the CTCF-I binding site (at a distance of ~700 bp) than the CTCF-II binding site (at a distance of
~1,200 bp, thereby exceeding the size of most sheared DNA fragments isolated by ChIP) (Figure
2.2D; p = 0.02, one-way ANOVA). Thus, ChIP analysis indicated that in vivo CTCF-I
occupancy is significantly diminished in the cerebellum of $SCA7$-$CTCF-I$-mut mice.

2.4.2 $SCA7$-$CTCF-I$-mut mice display increased germ line instability

We assessed intergenerational repeat length instability in 3 month-old $SCA7$-$CTCF-I$-wt and
SCA7-CTCF-I-mut mice by PCR amplification of the (CAG)_n repeat from tail DNAs, and found that mutation of the CTCF-I site destabilized the (CAG)_n repeat during intergenerational transmission (p = 0.002, Mann-Whitney two-tailed test) (Figure 2.3A). Increased intergenerational instability in the SCA7-CTCF-I-mut mice was reflected by a broader range of repeat length change, as mean expansion and deletion sizes were greater for SCA7-CTCF-I-mut mice in comparison to SCA7-CTCF-I-wt mice (+4.4 (CAG)_n repeats/−4.7 (CAG)_n vs. +2.6 (CAG)_n/−2.0 (CAG)_n). Analysis of repeat length instability between the two SCA7-CTCF-I-mut lines revealed similar intergenerational repeat instability (p = 0.93, chi-square), and there was no difference in expansion bias between the two lines (p = 0.25, chi-square). Thus, the SCA7-CTCF-I-mut mice did not show integration site effects, suggesting that increased instability in the two lineages results from altered CTCF binding. We then assessed germ line repeat instability by small-pool PCR of individual alleles in sperm DNAs from mice at age 2 months and 16 months (Figure 2.3B and C). As the mice aged, the (CAG)_n repeat in SCA7-CTCF-I-mut mice became increasingly unstable (p = 0.009, Mann-Whitney two-tailed test), as mean expansion and deletion sizes were significantly greater for 16 month-old SCA7-CTCF-I-mut mice in comparison to SCA7-CTCF-I-wt mice (+24.3 (CAG)_n/+15.5 (CAG)_n vs. +9.2 (CAG)_n/−1.0 (CAG)_n). Increasing (CAG)_n repeat instability with aging in SCA7-CTCF-I-mut mice suggests a role for CTCF in DNA instability during spermatogenesis, or for the male germ line-restricted CTCF-like parologue (CTCFL), also known as brother of the regulator of imprinted sites, or ‘BORIS’ (Loukinov et al., 2002). A potential role for CTCFL/BORIS in male germ line instability in the SCA7-CTCF-I-mut mice is plausible, as mutation of the SCA7-CTCF-I site also prevented binding of CTCFL/BORIS in electrophoretic mobility shift assays (Figure 2.4).

2.4.3 SCA7-CTCF-I-mut mice display increased somatic instability

Another intriguing feature of repeat instability is variation in repeat size within and between the tissues of an individual organism. This tissue-specific instability, or “somatic mosaicism”, occurs in human patients with repeat diseases, and in mouse models of repeat instability and disease (La Spada, 1997; Pearson et al., 2005; Mirkin, 2007). While shown to be age-dependent, the mechanistic basis of inter-tissue variation, which even occurs in postmitotic neurons (Gonitel et al., 2008), is unknown. To determine if somatic (CAG)_n mosaicism at the SCA7 locus...
Figure 2.3

A

![Histogram showing percentage of total alleles counted with repeat size change](image)

B

![Image showing SCA7-CTCF-I-wt and SCA7-CTCF-I-mut repeat size changes](image)

C

<table>
<thead>
<tr>
<th>2 months</th>
<th>16 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Histogram showing percentage of molecules observed with repeat size change" /></td>
<td><img src="image" alt="Histogram showing percentage of molecules observed with repeat size change" /></td>
</tr>
</tbody>
</table>
Figure 2.3 – *SCA7-CTCF-I-mut* mice display increased germ line instability (from Libby et al. (Libby et al., 2008)).  (A) Comparison of (CAG)$_n$ repeat instability in parent-offspring transmission for *SCA7-CTCF-I* mice. Repeat lengths are plotted as % of total alleles scored for 53 *SCA7-CTCF-I*-wt and 95 *SCA7-CTCF-I*-mut mice. The repeat size range in the *SCA7-CTCF-I*-mut mice was significantly different from the distribution of repeat alleles in the *SCA7-CTCF-I*-wt mice ($p = 0.002$; Mann-Whitney two-tailed test). (B) Small-pool PCR of sperm DNAs in 16 month-old SCA7 transgenic mice. *SCA7-CTCF-I*-wt mice typically exhibited small repeat length changes, while *SCA7-CTCF-I*-mut mice displayed pronounced instability. (C) Compilation of small-pool PCR data. At 2 months of age, only modest instability was noted. At 16 months of age, *SCA7-CTCF-I*-wt mice displayed moderate instability, but *SCA7-CTCF-I*-mut mice exhibited significantly greater instability ($p = 0.009$; Mann-Whitney two-tailed test).
Figure 2.4 – Mutation of *SCA7-CTCF-I* site also abrogates binding by BORIS (from Libby *et al.* (*Libby et al.*, 2008)). Electrophoretic mobility shift assays with *SCA7-CTCF-I*-wt and -mut probe fragments were performed with probe only, the 11 zinc-finger DNA binding domain region of CTCF, full-length CTCF protein, full-length BORIS protein, BORIS protein with anti-BORIS sera (BORIS+α-BORIS), or BORIS with pre-immune anti-BORIS sera (BORIS+pl). Arrows indicate shifted CTCF-DNA complexes, shifted BORIS-DNA complexes, and super-shifted BORIS-DNA complexes. Addition of CTCF-DM1 probe as cold competitor prevented CTCF-DNA and BORIS-DNA complex formation for the *SCA7-CTCF-I*-wt fragment, while non-specific cold competitor did not (data not shown).
involves CTCF binding, we surveyed repeat instability in various tissues from \textit{SCA7-CTCF-I}-wt and \textit{SCA7-CTCF-I}-mut mice. At two months of age, the SCA7 (CAG)\textsubscript{n} repeat was remarkably stable in all analysed tissues (Figure 2.5A). However, by \~10 months of age, \textit{SCA7-CTCF-I}-wt and \textit{SCA7-CTCF-I}-mut mice displayed large (CAG)\textsubscript{n} repeat expansions in the cortex and liver (Figure 2.5B). The liver also exhibited a bimodal distribution of repeat size (i.e. two populations of cells with distinct tract lengths) (Figure 2.5B). The most pronounced somatic instability differences existed in the kidney, with large expansions for \textit{SCA7-CTCF-I}-mut mice, but stable repeats in the \textit{SCA7-CTCF-I}-wt mice (Figure 2.5B). This pattern of increased kidney and liver repeat instability was present in both \textit{SCA7-CTCF-I}-mut transgenic lines (Figure 2.5B; Figure 2.6). Indeed, comparable somatic instability was also detected in both \textit{SCA7-CTCF-I}-mut transgenic lines at five months of age (Figure 2.7). When we closely examined repeat instability in the cortex by small-pool PCR, we observed significantly different repeat sizes (p = 8.6×10\textsuperscript{-5}, Mann-Whitney), with a range of 39 to 152 (CAG)\textsubscript{n} repeats in \textit{SCA7-CTCF-I}-wt mice and 26 to 245 (CAG)\textsubscript{n} repeats in \textit{SCA7-CTCF-I}-mut mice (Figure 2.5C; Table 2.1). The increased somatic instability occurred in both \textit{SCA7-CTCF-I}-mut transgenic lines, as an expansion bias was apparent in both lineages upon small-pool PCR analysis (Figure 2.5D; Table 2.1). These findings suggest that CTCF binding stabilizes the SCA7 (CAG)\textsubscript{94} repeat in certain tissues. Thus, as noted for the germ line and documented for two independent lines of \textit{SCA7-CTCF-I}-mut transgenic mice, SCA7 somatic (CAG)\textsubscript{n} instability is dependent upon age and the presence of intact CTCF binding sites.

2.4.4 \textit{CpG} methylation of the CTCF binding site induces instability

CTCF binding can be regulated by CpG methylation, as methylation at CTCF recognition sites abrogates binding (Ohlsson et al., 2001). This finding was confirmed for unmethylated and methylated versions of the SCA7 CTCF-I recognition site (Figure 2.8A; Figure 2.9). Highly variable levels of instability have been documented in the kidneys of transgenic repeat instability mouse models (Gomes-Pereira et al., 2001; van Den Broek et al., 2002), although the reasons for pronounced instability in this tissue are unknown. Interestingly, one mouse with a wild-type CTCF-I binding site (\textit{SCA7-CTCF-I}-wt\textsuperscript{*}) displayed marked (CAG)\textsubscript{n} repeat instability in its kidney DNA (Figure 2.8B), paralleling the considerable instability observed in the \textit{SCA7-CTCF-I}-mut mice (Figure 2.5B). Bisulfite sequencing of kidney DNA from this \textit{SCA7-CTCF-I}-wt\textsuperscript{*}
Figure 2.5

A

SCA7-CTCF-I-wt (2 Months)

(CAG)n
Heart Kidney Cerebellum Cortex Brainstem Liver
182 128 95

B

SCA7-CTCF-I-wt (10.5 Months)

(CAG)n
Heart Kidney Cerebellum Cortex Brainstem Liver
195 162 128 95

C

SCA7-CTCF-I-wt (10.5 Months)

(CAG)n
Heart Kidney Cerebellum Cortex Brainstem Liver
195 162 128 95

D

SCA7-CTCF-I-wt (10.5 months) SCA7-CTCF-I-mut (8 months) SCA7-CTCF-I-mut (12 months)

(CAG)n Repeat Size Change (CAG)n Repeat Size Change (CAG)n Repeat Size Change
Figure 2.5 – *SCA7-CTCF-I*-mut mice display increased somatic instability (from Libby et al. (Libby et al., 2008)). (A) At 2 months of age, the SCA7 (CAG)_n repeat is stable in the *SCA7-CTCF-I*-wt line and in both *SCA7-CTCF-I*-mut lines. (B) With advancing age, tissue-specific instability is seen in *SCA7-CTCF-I*-wt mice; however, this tissue-specific instability is much more pronounced in *SCA7-CTCF-I*-mut mice. Results for individuals from the two different *SCA7-CTCF-I*-mut mice are shown here. (C) To permit quantification of somatic instability, we performed small-pool PCR on tissue DNA samples from *SCA7-CTCF-I*-wt and *SCA7-CTCF-I*-mut mice. As shown here for cortex, *SCA7-CTCF-I*-mut mice displayed significantly greater instability than *SCA7-CTCF-I*-wt mice (p = 8.6×10^{-5}, Mann-Whitney two-tailed test). See Table 2.1 for a compiled list of repeat alleles. (D) Histogram of repeat length variation in the cortex of *SCA7-CTCF-I*-wt and *SCA7-CTCF-I*-mut mice. *SCA7-CTCF-I*-mut mice exhibit significantly greater instability than *SCA7-CTCF-I*-wt mice, and this expansion tendency exceeds that of *SCA7-CTCF-I*-wt mice, even when 2.5 months younger (p = 0.0003, Mann-Whitney two-tailed test). With advancing age, the expansion bias between the *SCA7-CTCF-I*-mut and -wt mice becomes more pronounced (p<0.0001, Mann-Whitney two-tailed test). Results for individuals from the two different *SCA7-CTCF-I*-mut mice are shown here.
Figure 2.6 – Increased somatic instability in both SCA7-CTCF-I-mut transgenic lines (from Libby et al. (Libby et al., 2008)). Here, we see representative results for PCR analysis of somatic repeat instability for aged mice from each of the two SCA7-CTCF-I-mut transgenic lines analysed in this study. Note that comparable patterns of increased somatic mosaicism are observed in each lineage.
Figure 2.7 – Comparable somatic mosaicism in both SCA7-CTCF-I-mut transgenic lines (from Libby et al. (Libby et al., 2008)). Here, we see representative images for PCR analysis of somatic repeat instability for 5 month-old individuals from each of the two SCA7-CTCF-I-mut transgenic lines analysed in this study. Note that comparable patterns of increased somatic mosaicism are again observed at this earlier point.
Table 2.1 – Repeat sizes of cortex DNA: (CAG)$_n$ tract length – Small-pool PCR (from Libby et al. (Libby et al., 2008)).

<table>
<thead>
<tr>
<th>SCA7-CTCF-I-wt (10.5 months)</th>
<th>SCA7-CTCF-I-mut (2 months)</th>
<th>SCA7-CTCF-I-mut (8 months) *</th>
<th>SCA7-CTCF-I-mut (12 months) *</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>98</td>
<td>91</td>
<td>103</td>
</tr>
<tr>
<td>78</td>
<td>98</td>
<td>91</td>
<td>103</td>
</tr>
<tr>
<td>83</td>
<td>101</td>
<td>91</td>
<td>110</td>
</tr>
<tr>
<td>83</td>
<td>101</td>
<td>91</td>
<td>130</td>
</tr>
<tr>
<td>83</td>
<td>104</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td>86</td>
<td>107</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>107</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>111</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>114</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>114</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>117</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>89</td>
<td>117</td>
<td>96</td>
<td>99</td>
</tr>
<tr>
<td>92</td>
<td>117</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>92</td>
<td>117</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>92</td>
<td>121</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>92</td>
<td>121</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>95</td>
<td>121</td>
<td>98</td>
<td>103</td>
</tr>
<tr>
<td>95</td>
<td>124</td>
<td>98</td>
<td>103</td>
</tr>
<tr>
<td>95</td>
<td>128</td>
<td>98</td>
<td>103</td>
</tr>
<tr>
<td>95</td>
<td>136</td>
<td>98</td>
<td>103</td>
</tr>
<tr>
<td>95</td>
<td>136</td>
<td>99</td>
<td>103</td>
</tr>
<tr>
<td>95</td>
<td>136</td>
<td>99</td>
<td>103</td>
</tr>
<tr>
<td>95</td>
<td>139</td>
<td>99</td>
<td>104</td>
</tr>
<tr>
<td>95</td>
<td>144</td>
<td>99</td>
<td>104</td>
</tr>
<tr>
<td>95</td>
<td>152</td>
<td>99</td>
<td>104</td>
</tr>
<tr>
<td>95</td>
<td>101</td>
<td>104</td>
<td>176</td>
</tr>
<tr>
<td>95</td>
<td>101</td>
<td>104</td>
<td>107</td>
</tr>
<tr>
<td>95</td>
<td>103</td>
<td>107</td>
<td>107</td>
</tr>
<tr>
<td>95</td>
<td>103</td>
<td>107</td>
<td>107</td>
</tr>
<tr>
<td>98</td>
<td>103</td>
<td>107</td>
<td>107</td>
</tr>
</tbody>
</table>

*Results for individuals from the two different SCA7-CTCF-I-mut lines are shown here.
Figure 2.8
Figure 2.8 – Epigenetic regulation of CTCF binding modulates instability at the SCA7 locus (from Libby et al. (Libby et al., 2008)). (A) CpG methylation prevents binding of CTCF to SCA7-CTCF-I site. Electrophoretic mobility shift assays with unmethylated (control) or methylated SCA7-CTCF-I fragments, using CTCF with no antisera (CTCF), CTCF with anti-CTCF antisera (CTCF+α-CTCF), or CTCF with pre-immune sera (CTCF+pI). Arrow indicates CTCF-bound probe. (B) Prominent somatic instability in kidney DNA (black arrowheads) from a SCA7-CTCF-I-wt mouse with CTCF-I site methylation (SCA7-CTCF-I-wt*) contrasts with somatic stability in SCA7-CTCF-I-wt mice with unmethylated CTCF-I sites. Note that SCA7-CTCF-I-wt lines display bimodal (CAG)_n repeat alleles. Prominent somatic instability is apparent in kidney DNA (gray arrowhead) from a SCA7-CTCF-I-mut mouse. All mice were 6 months of age. (C) Kidney DNAs from the SCA7-CTCF-I-wt* mouse are highly methylated. Circles, CpG dinucleotides; open circles, unmethylated; filled circles; methylated. Box highlights core CTCF binding site contact residue, based upon footprinting analysis. Diagrammed epigenotypes summarize results for five SCA7-CTCF-I-wt mice, eight SCA7-CTCF-I-mut mice, and the SCA7-CTCF-I-wt* mouse, and were consistent for at least 75% of all sequenced clones (n = 10–12/sample). (D) Liver DNAs from control SCA7-CTCF-I-wt mice are methylated. Bisulfite sequencing of the SCA7-CTCF-I region was performed upon liver DNAs from three SCA7-CTCF-I-wt mice at one year of age (n = 17 clones/mouse), and CpG methylation determined for the 13 CpG dinucleotides in the SCA7-CTCF-I region. A number of CpG dinucleotides, including the CpG-4 CTCF contact site, exhibit moderate to high levels of methylation.
Figure 2.9 – Methylation of SCA7-CTCF-I-wt probe fragment for gel shift analysis (from Libby et al. (Libby et al., 2008)). SssI was used to methylate cytosine residues in CpG dinucleotides in the SCA7-CTCF-I-wt probe fragment. Digestion of control (unmethylated) and SssI-methylated probe fragments with the methylation-sensitive restriction enzyme NruI revealed complete methylation of SssI-treated SCA7-CTCF-I-wt probe fragment.
mouse revealed high levels of CpG methylation at the wild-type CTCF-I binding site, including the central CTCF contact site (Figure 2.10); whereas methylation was not observed in kidney DNAs from 14 other SCA7-CTCF-I-wt mice that displayed only modest levels of (CAG)\textsubscript{n} instability (Figure 2.8C). The high levels of (CAG)\textsubscript{n} instability and the CpG methylation in this mouse were restricted to the kidney, as the cerebellum and tail DNAs of the same mouse, which showed limited (CAG)\textsubscript{n} instability (Figure 2.8B), were completely unmethylated (Figure 2.8C). This finding suggests a direct link between methylation status of the CTCF binding site and (CAG)\textsubscript{n} repeat instability. Of all the tissues analysed from SCA7-CTCF-I-wt mice, liver exhibits the greatest amount of somatic mosaicism, with the largest repeat expansions (Figure 2.5B). We hypothesized that the high levels of (CAG)\textsubscript{n} repeat instability in the liver of SCA7-CTCF-I-wt mice might result from methylation of the CTCF-I binding site. To address this question, we performed bisulfite sequencing analysis of liver DNAs from SCA7-CTCF-I-wt mice, and documented moderately high levels of methylation at the CTCF-I binding site (Figure 2.8; Figure 2.11). These results indicate a correlation between CpG methylation and (CAG)\textsubscript{n} repeat instability. Thus, in SCA7 transgenic mice, decreased CTCF binding, either by CpG methylation or mutagenesis of the CTCF-I binding site, enhanced (CAG)\textsubscript{n} repeat instability.

2.5 Discussion
We have identified a CTCF binding site as the first cis-element regulating (CAG)\textsubscript{n} tract instability at a disease locus. Furthermore, binding of the trans-factor CTCF to this cis-element influences (CAG)\textsubscript{n} instability, and this interaction is epigenetically regulated. At the SCA7 locus and four other (CTG)\textsubscript{n}/(CAG)\textsubscript{n} repeat loci known to display pronounced anticipation, functional CTCF binding sites occur immediately adjacent to the repeats, and CTCF binding can affect DNA structure and chromatin packaging at such loci, and elsewhere (Filippova et al., 2001; Libby et al., 2003; Cho et al., 2005; Navarro et al., 2006; Splinter et al., 2006). Although interplays between GC-content, CpG islands, epigenetic modification, chromatin structure, repeat length, and unusual DNA conformation has long been postulated to underlie TNR instability (Gourdon et al., 1997a; Brock et al., 1999; Nichol and Pearson, 2002; Mirkin, 2007), the mechanistic basis of this process is ill-defined. CTCF insulator and genomic imprinting functions are subject to epigenetic regulation, as methylation status is a key determinant of CTCF action at certain “differentially methylated domains” and methylation changes at CTCF
Figure 2.10 – Amplicon for bisulfite sequencing for epigenotype determination (from Libby et al. (Libby et al., 2008)). PCR amplification of bisulfite-converted genomic DNA for the fragment shown here was performed to derive CpG methylation status at the SCA7-CTCF-I binding site in murine tissues. Intron sequence is lowercase; exon sequence is uppercase. The SCA7-CTCF-I binding site is shown in blue. The thirteen CpG dinucleotides included in the epigenotyping are shown, and the dyad with filled circles corresponds to a critical CTCF contact site, based upon footprinting analysis (see Figure 2.2C).
Figure 2.11

SCA7-CTCF-I-wt (mouse 1)

SCA7-CTCF-I-wt (mouse 2)

SCA7-CTCF-I-wt (mouse 3)
Figure 2.11 – Epigenotype data for bisulfite sequencing analysis of the CTCF-I binding site region in SCA7-CTCF-I-wt transgenic liver (from Libby et al. (Libby et al., 2008)). Results of bisulfite sequencing analysis for liver DNAs obtained from three SCA7-CTCF-I-wt transgenic mice reveal moderate to high levels of CpG methylation in this tissue, especially when compared to the completely unmethylated status of CpG dinucleotides observed in all tail DNAs and kidney DNAs, with one exception.
binding sites are linked to oncogenic transformation (Ohlsson et al., 2001; Filippova, 2008). At the SCA7 locus, methylation status of the CTCF-I binding site may be similarly important for its ability to tamp down repeat instability, as hypermethylation of the CTCF-I site was associated with a dramatic enhancement of somatic instability in the SCA7 genomic fragment transgenic mouse model. Thus, inability to bind CTCF at sites adjacent to (CAG)_n tracts, because of binding site mutation or CpG methylation in the case of the SCA7-CTCF-I site, can promote further expansion of disease-length (CAG)_n repeat alleles (Figure 2.12).

In both human patients and transgenic mice with expanded repeat tracts, the repeat displays high levels of instability. The flanking sequence has been thought to contain elements that may protect or enhance repeat instability. Our results show that CTCF binding is a stabilizing force at the SCA7 repeat locus, suppressing expansion of the (CAG)_{94} repeat in the germ line and soma. Interestingly, deletion of ~8.3 kb of 3’ genomic sequence in our previous SCA7 transgenic mouse, including the CTCF-I site, stabilized the repeat (Libby et al., 2003). The (CAG)_{92} stabilization, arising from the ~8.3 kb 3’ genomic fragment deletion, suggests the existence of positive cis-regulators that were “driving” (CAG)_n instability. One such element could be a replication initiation site that was mapped within the genomic region 3’ to the CTCF-I binding site at the SCA7 locus (Nenguke et al., 2003). Hence, the ~8.3 kb 3’ deletion could grossly alter the chromatin organization of the adjacent repeat, and would likely ablate replication origin activity, stabilizing the (CAG)_{92} repeat tract. However, this ~8.3 kb genomic region likely also contained negative cis-regulators of (CAG)_n repeat instability, whose dampening effects would not be apparent due to the coincident loss of instability drivers. Our results indicate that CTCF binding negatively regulates expanded (CAG)_n repeat instability at the SCA7 locus. CTCF regulation of repeat instability potential is consistent with its many roles in modulating DNA structure. CTCF can mediate long-range chromatin interactions and can co-localize physically distant genomic regions into discrete sub-nuclear domains (Ling et al., 2006; Filippova, 2008). CTCF is contained within chromatin insulators, found between silenced genes and transcriptionally active genes, and binding sites occur at transition zones between X-inactivation regions and genes that escape from X-inactivation (Filippova et al., 2005). CTCF has been implicated in genomic imprinting, although recent studies indicate that such transcription insulator events may involve the coordinated action of CTCF with cohesin (Parelho et al., 2008; Stedman et al., 2008; Wendt et al., 2008). CTCF binding at the DM1 locus.
Figure 2.12 – Model for CTCF regulation of (CAG)$_n$ repeat instability (from Libby et al. (Libby et al., 2008)). Non-expanded TNR is stable, as CTCF is bound to adjacent site. Upon repeat expansion, chromatin environment and DNA structure of repeat region is altered, permitting instability. Loss of CTCF binding at adjacent CTCF binding site, either by CpG methylation or CTCF binding site mutation, further promotes repeat instability.
sequesters repeat-driven heterochromatin formation to the immediate repeat region, while repeat expansion-induced loss of CTCF binding may permit spreading of heterochromatin to adjacent genes, accounting for the mental retardation phenotype in congenital DM1 (Cho et al., 2005). As DNA structural conformation and transcription activity are two highly intertwined processes that appear fundamental to the instability of expanded tandem repeats (Jung and Bonini, 2007; Mirkin, 2007), CTCF appears a likely candidate for modulation of trinucleotide repeat instability.

At the SCA7 locus, a pronounced tendency for repeat expansion has been associated with transmission through the male germ line (David et al., 1997; Gouw et al., 1998; Monckton et al., 1999). Although we have hypothesized that CTCF is principally responsible for modulating SCA7 (CAG)$_n$ repeat instability both in the germ line and in the soma, we considered a possible role for the related CTCF-like factor BORIS. BORIS and CTCF share identical 11 zinc-finger domains for DNA binding (Loukinov et al., 2002); hence, both CTCF and BORIS can bind to the CTCF binding sites at the SCA7 locus. Upon mutation or methylation of the CTCF binding site 3′ to the SCA7 (CAG)$_{94}$ repeat, neither CTCF nor BORIS can bind (Figure 2.2C; Figure 2.8A; Figure 2.13). As BORIS can bind to the H19 differentially methylated domain even when it is methylated (Nguyen et al., 2008), our results suggest that the methylation dependence of BORIS binding is locus specific. BORIS and CTCF expression patterns overlap very little, if at all, and in the male germ line, BORIS appears restricted to primary spermatocytes, while CTCF occurs almost exclusively in post-meiotic cells, such as round spermatids (Loukinov et al., 2002). Interestingly, neither BORIS nor CTCF could be detected by immunostaining proliferating spermatogonia. In human HD patients and transgenic mouse models of (CTG)$_n$/(CAG)$_n$ instability, large repeat expansions have been documented in spermatogonia, but not in post-meiotic spermatids or spermatozoa (Zhang et al., 2002; Savouret et al., 2003; Yoon et al., 2003; Savouret et al., 2004). Thus, absence or low levels of BORIS or CTCF in spermatogonia – the cells in which the largest and most frequent repeat expansions occur – may contribute to the paternal parent-of-origin expansion bias common to most (CTG)$_n$/(CAG)$_n$ repeat diseases. In spermatocytes, BORIS may stabilize expanded (CAG)$_n$ repeats, just as CTCF binding appears to promote repeat stability in somatic tissues. Thus, in the SCA7-CTCF-I-mut mice, abrogated binding of BORIS may contribute to increased repeat instability and expansion bias in the male germ line.
Figure 2.13 – Methylation of the *SCA7-CTCF-I* site abrogates binding of BORIS as well as CTCF (from Libby et al. (Libby et al., 2008)). Gel retardation assays with unmethylated or *Sss*I-methylated *SCA7-CTCF-I*-wt probe fragments were performed with probe only, the 11 zinc-finger DNA binding domain region of CTCF, CTCF with pre-immune anti-CTCF sera (CTCF+pI), CTCF protein with anti-CTCF sera (CTCF+α-CTCF), BORIS with pre-immune anti-BORIS sera (BORIS+pI), or BORIS protein with anti-BORIS sera (BORIS+α-BORIS). Arrows indicate shifted CTCF-DNA complexes and shifted BORIS-DNA complexes. Methylation of the *SCA7-CTCF-I* probe fragment abrogates all binding. Success of *Sss*I methylation was confirmed by *Nru*I restriction digestion (see Figure 2.9).
Our findings suggest that CTCF is a *trans*-acting factor that specifically interacts in a methylation-dependent manner with the adjacent *cis*-environment to prevent hyper-expansion of disease length (CAG)$_n$ repeats. In a *Drosophila* model of polyglutamine repeat disease, expression of the mutant gene product modulated repeat instability by altering transcription and repair pathways (Jung and Bonini, 2007). Similarly, uninterrupted repeat sequences, and in particular, runs of CG-rich TNRs, can affect replication machinery, DNA repair pathways, and nucleosome positioning, though in *cis*, by altering the structure and conformation of the DNA regions within which they reside (Wang and Griffith, 1995; Mirkin, 2005). Association of adjacent CTCF binding sites with repeat loci is a common feature of unstable microsatellite repeats (Filippova et al., 2001). We propose that acquisition of CTCF binding sites at mutational hot spots represents an evolutionary strategy for insulating against mutation-inducing *cis*-sequences (Benzer, 1961), and our findings indicate that CTCF binding site utilization at a mutational hot spot is subject to epigenetic regulation. We thus envision a predominant role for CTCF in modulating genetic instability at DNA regions containing variably-sized repeats, unstable sequence motifs, or other repetitive sequence elements.
CHAPTER 3 - CTCF INDUCES REPLICATION FORK STALLING AT THE MYOTONIC DYSTROPHY SEQUENCE

Co-authorship statement: I performed the entirety of this chapter, including conception of the ideas, execution of experiments, statistical analysis and figure preparation. John Cleary helped with the instability analysis of pDMK constructs and graphical design of figures 3.2 and 3.3.
3.1 Abstract

Trinucleotide repeat (TNR) expansions are associated with at least 25 neurological, neuromuscular and neurodegenerative disorders. Myotonic dystrophy (DM1) is caused by a TNR expansion in the 3’ UTR of the DMPK gene. The expanded repeat is very unstable upon transmission, leading predominantly to expansions, worsening phenotypes and decreasing age of onset with each new generation. Contributors to TNR instability include cis-elements around the repeat such as replication origins and CTCF binding sites. Depending on the position of replication origins relative to the TNR, repeats can become more or less stable and have predominantly expansions or deletions. The chromosomal origin of replication has been mapped to the DM1 locus as well as other disease-associated TNR loci. CTCF binding sites are present between the origin and the repeat at the DM1 locus, and are proximal to other disease repeat loci. CTCF was shown to protect against hyper-instability of spinocerebellar ataxia type 7-associated TNRs in an epigenetic manner. To determine the mechanistic effect of CTCF at TNRs, we designed a primate model to test for replication efficiency across a fragment of the DM1 region including two CTCF binding sites and a (CTG)\textsubscript{83} repeat. CTCF binding strongly reduced replication efficiency in a replication direction-dependent manner. Loss of CTCF binding by binding site mutation increased replication efficiency by up to 10-fold. Analysis of replication intermediates showed that CTCF binding sites but not their mutant forms, cause replication fork stalling. CTCF binding sites affected TNR instability whereby loss of one or both binding sites altered the pattern of instability. Given that CTCF binding sites are found around many TNRs, the CTCF-induced replication fork stalling effect may also occur at many disease-associated TNRs and contribute to instability and disease.

3.2 Introduction

Trinucleotide repeat (TNR) expansion is the cause of at least 25 inherited neurological disorders, including myotonic dystrophy (DM1) and spinocerebellar ataxia type 7 (SCA7) (Pearson et al., 2005). Many TNR diseases exhibit anticipation with worsening disease phenotypes, decreased age of onset, and increased degree of instability upon transmission of the expanded repeat to the next generation. Moreover, somatic expansion through the patient’s life is thought to contribute to disease progression. TNR instability occurs in proliferative and non-proliferative tissues, and has been attributed to metabolic processes such as DNA replication and repair (Pearson et al.,
Strong evidence for replication-mediated instability comes from model systems (Cleary et al., 2002; Yang et al., 2003; Freudenreich and Lahiri, 2004) and DM1 patients (Wohrle et al., 1995; Zatz et al., 1995; Martorell et al., 1998). For example, one study showed ongoing instability in a human DM1 patient cell line that was further destabilized by replication fork inhibitors (Yang et al., 2003). Other studies showed that TNRs cause replication fork stalling in vitro (Pelletier et al., 2003; Krasilnikova and Mirkin, 2004a; Krasilnikova and Mirkin, 2004b; Voineagu et al., 2009). These studies suggest that instability can occur due to stalled replication forks near the repeat allowing for mutagenic DNA structures or slippage of the template DNA during Okazaki initiation and processing (Cleary et al., 2002). Support for replication-induced TNR instability comes from dramatic changes in repeat instability observed when an origin of replication is placed upstream versus downstream of TNRs as well as at varying distances to the repeat in vitro (Kang et al., 1995a; Freudenreich et al., 1997; Samadashwily et al., 1997; Freudenreich et al., 1998; Miret et al., 1998; Cleary et al., 2002). Clearly a portion of TNR instability can be caused by replication.

Another proposed contributor to repeat instability is cis-elements around TNRs. Some cis-elements and cis-events proposed to regulate TNR instability are the sequence of the repeat tract, the length and purity of the repeat tract, flanking DNA sequences, surrounding epigenetic environment, replication origin determinants, trans-factor binding sites, and transcriptional activity (Sinden, 2001; Cleary and Pearson, 2003; Jung and Bonini, 2007; Mirkin, 2007). Chromosomal origins of replication have been reported at disease-associated repeat loci for DM1, SCA10, and FRAXA (Nenguke et al., 2003; Brylawski et al., 2006; Chastain et al., 2006; Gray et al., 2007), though the effect on repeat instability was not directly examined. The region containing the expanded repeat of FRAXA is replicated late (Hansen et al., 1993; Subramanian et al., 1996), suggesting there may be epigenetic differences between expanded and wild-type repeat regions affecting replication around repeats. Another common element around repeat loci is the CTCF binding site, previously found at SCA2, SCA7, HD, dentatorubral-pallidoluysian atrophy, facioscapulohumeral dystrophy, FRAXA and DM1 (Filippova et al., 2001; Ladd et al., 2007; Ottaviani et al., 2009). CTCF is an evolutionarily conserved zinc-finger DNA binding protein that can affect higher order chromatin structure and establish chromatin domains (Ling et al., 2006; Filippova, 2008; Cuddapah et al., 2009). CTCF has activity in chromatin insulation, transcriptional regulation, genomic imprinting and replication timing (Lobanenkov et al., 1990;
Ohlsson et al., 2001; Bergstrom et al., 2007). CTCF binding sites are situated between the replication origins and the unstable repeats. However, the role of CTCF binding and replication progression has not been established.

We previously showed a connection between CTCF and repeat instability in a SCA7 mouse model by mutating the CTCF binding site, causing hyper-mutation of the (CAG)_{94} repeat (Libby et al., 2008). The mechanism by which CTCF-dependent instability occurs remains unknown. We and others have proposed a fork-shift model (Cleary et al., 2002; Cleary and Pearson, 2005; Mirkin, 2005) whereby sequences or epigenetic modifications around or within TNRs affect instability through mechanisms involving replication fork progression and Okazaki initiation. Studies of FRAXA (CGG)_{n} repeats, DM1 (CTG)_{n} repeats, FRDA (GAA)_{n} repeats and DM2 (CCTG)_{n} repeats clearly support the replication origin-shift model by moving the origin relative to the repeat by increments of Okazaki fragment lengths to induce varying levels of instability (Cleary et al., 2002; Dere et al., 2004; Nichol Edamura et al., 2005; Rindler et al., 2006). Given that the DM1 locus is sensitive to origin-shifts in vitro, it is possible that the locus may also be sensitive to fork-shifts caused by endogenous cis-elements such as CTCF binding sites around the repeat in DM1 patients causing disease-associated TNR instability. Towards determining the mechanistic role of CTCF binding sites at TNRs, we established a primate model system to test CTCF-dependent replication efficiency, replication fork progression and instability at DM1 sequence. Replication efficiency was greatly reduced in constructs with CTCF binding sites, which was attributed to replication fork stalling at the binding sites around the repeat. Furthermore, mutation of one or both CTCF binding sites changed the degree of replication stalling and the pattern of instability, supporting the fork-shift model of repeat instability.

3.3 Materials and methods

3.3.1 Design of constructs

Replication templates were constructed using the pCRScript-Amp plasmid (Stratagene) similar to previously published (Cleary et al., 2002) (Figure 3.1A). The DM1 region (OMIM: 605377) was cloned from patient fetal fibroblasts as previously described (Yang et al., 2003) using the primers 5'-CTGCCAGTT CACAACCGGTCCGAG-3' and 5'-AAGCAAATTTCCGAGCTAA GCAGGC-3'. A 219 bp fragment containing the SV40 origin of replication (viral nucleotide
Figure 3.1

A) Diagram showing the integration of selected DNA fragments.

B) Table showing the hybridization of different DNA constructs (pDM₁ to pDM₄) with their respective efficiencies.

C) Additional table showing the hybridization of different DNA constructs (pDM₁ to pDM₄) with their respective efficiencies.
Figure 3.1 - CTCF binding sites and (CTG)$_{83}$ repeats from the DM1 region reduce replication efficiency of primate replicated templates. (A) Design of replication templates with the SV40 origin of replication and DM1 sequence from human patient fibroblasts. pDM constructs have (CTG)$_{83}$ (grey box), CTCF binding sites (black box), bacterial elements (f1 ori, pUC and bla gene), additional kanamycin sequence added to pDMK constructs for 2D gel experiments (Kan fragment) and the SV40 origin of replication (blue square) either upstream of the (CTG)$_{83}$ repeat (SV40 ori 1, for constructs pDM$_{1-4}$ and pDMK$_{1-4}$) or downstream of the (CTG)$_{83}$ repeat (SV40 ori 2, for constructs pDM$_{5-8}$). The location of the hybridization probe for replication efficiency (RE) and 2D gel electrophoresis (2D) is illustrated. (B) Replication efficiency was measured following co-transfection with a control pSV40 plasmid lacking DM1 sequence, DpnI digestion to remove unreplicated parental template and linearized with AlwNI. Starting templates (S) and replicated material (R) were probed with a $^{32}$P-XbaI fragment of pCRScript-Amp to determine replication efficiency. The replication efficiency was calculated by comparing the ratio of pSV40 to pDM plasmid before primate replication to after primate replication (see Materials and methods) and displayed below the replicated material (%). Values represent the average of four experimental replicates with standard deviation shown in brackets below the replication efficiency value. (C) Replication efficiency of pre-methylated constructs by SsSI methylase treatment. Values represent the average of two to three experimental replicates with standard deviation shown in brackets below the replication efficiency value.
position 5,210/5,211, GenBank accession number NC_001669) (Hay and DePamphilis, 1982) was cloned into either the \textit{XhoI} or \textit{SapI} site. CTCF sites 1 and 2 found in the DM1 patient sequence were mutated by Site-Directed Mutagenesis (Stratagene) using substitutions known to abrogate CTCF binding (Filippova et al., 2001).

\textbf{3.3.2 Cell culture}

COS-1 cells were grown and transfected as described (Cleary et al., 2002) using Lipotaxi transfection reagent (Stratagene). Episomal DNA isolations were performed as described (Hirt, 1967; Cleary et al., 2002) 48 hours post-transfection for replication efficiency and STRIP analysis, and 24 hours post-transfection for 2D gel electrophoresis. Episomal DNA was purified with phenol and chloroform extractions, precipitated and resuspended in 10 mM Tris-HCl (pH 7.6) and 1 mM EDTA.

\textbf{3.3.3 Replication efficiency}

Replication efficiency was assessed as described (Nichol Edamura et al., 2005). Replication efficiency was calculated by using equal amounts of pDM and an identical template that lacks DM1 sequence or a TNR (pSV40). COS-1 cells were transfected with equal amounts of both plasmids. Replicated intermediates were linearized with AlwN\textit{I} and treated with \textit{DpnI} to remove unreplicated starting plasmid DNA. Replication products were resolved on a 1% agarose gel, transferred to Biodyne B membrane (Pall Corporation), and probed with the \textsuperscript{32}P-\textit{XbaI} fragment of the SV40 origin. An aliquot of the starting plasmid mixture was linearized with AlwN\textit{I} and analysed on the same gel. The densitometric ratio of pDM to pSV40 in starting material was compared to the ratio after primate replication (Phosphorimager scan of Southern blot followed by analysis with ImageQuant software -Molecular Dynamics) and used to determine percent replication efficiency. CpG methylation of replication templates for replication efficiency studies was performed by \textit{SssI} methylase treatment and successful methylation was confirmed by the DNA being resistant to digestion with the methyl-sensitive restriction enzyme \textit{AciI}. Replication efficiency analysis was performed in four replicates on unmethylated templates and two to three replicates for premethylated templates. Standard deviation between replicates was calculated for each template.
3.3.4 2D gel analysis of replication intermediates

Replication fork progression was assessed by 2D gel analysis (Samadashwily et al., 1997). To facilitate resolution of fork progression an additional ~600 bp $NaeI/SspI$ fragment of the Kanamycin gene was cloned into the $EcoRV$ site between the SV40 origin and the DM1 sequence (Figure 3.1A). Replication templates were transfected into primate cells and replicated for 24 hours. Episomal DNA was isolated, digested, and electrophoresed as described (Samadashwily et al., 1997). Episomal DNA digested with $XmnI$ and $AvrII$ were run in the first dimension for 24 h at 1 V/cm on a 0.4% agarose gel. Lanes were then cut out above the completely replicated band up to the well, embedded in 1% agarose gel with 0.5 mg/ml ethidium bromide, and run in the second dimension for 8 h at 5 V/cm at 4°C, with buffer recirculation (Cordeiro-Stone et al., 1997; Krasilnikova and Mirkin, 2004a). Gels were transferred as described previously (Nichol and Pearson, 2002) and were probed with the $^{32}P$-labeled $AlwNI/XmnI$ fragment of pPCR-Script Amp (Stratagene) away from the DM1 sequence.

3.3.5 Instability analysis

Repeat instability was assessed after primate replication by the Stability of Trinucleotide Repeats by Individual Products (STRIP) assay (Cleary et al., 2002). Episomal DNAs were digested with $DpnI$ (New England Biolabs) to eliminate unreplicated starting plasmids. DNAs were transformed into XL1-Blue MR bacteria (Stratagene), chosen for its low level of background instability. Individual colonies, representing single products of primate replication, were grown and plasmids isolated by Wizard SV Prep (Promega). Repeat-containing fragments were released by $PstI$ and $SacII$ digestion and resolved on 4% polyacrylamide gels, and repeat tract lengths scored relative to starting plasmid repeats and molecular weight marker. Corrected frequencies of expansion and deletion events were calculated by subtracting the background repeat heterogeneity found in plasmids transformed directly into bacteria without primate replication as described (Cleary et al., 2002). Repeats were categorized as ‘less than 83 repeats’, ‘83 repeats’ and ‘greater than 83 repeats’, and statistically assessed by $\chi^2$ analysis.
3.4 Results

3.4.1 Generation of DM1 replication templates
To assess the effect of CTCF binding sites at the DM1 locus on replication and repeat instability, replication templates were constructed with an SV40 origin of replication either upstream (pDM1-4) or downstream (pDM5-8) of repeat-containing sequence cloned from the DM1 locus of patient fibroblasts (Figure 3.1A) (Yang et al., 2003). This fragment contained (CTG)$_{83}$ repeats with DM1 flanking sequences (140 bp and 220 bp 5' and 3' of the (CTG)$_{83}$ repeat, respectively), including both CTCF binding sites (Filippova et al., 2001). Using mutations known to ablate CTCF binding, each CTCF binding site was mutated by site-directed mutagenesis in various permutations (Figure 3.1A) (Filippova et al., 2001). These templates replicate in transfected primate cells expressing T-antigen, and use the same replication proteins as human cells; replication initiates at the origin of bidirectional replication within the SV40 sequence, and the location of the origin determines the direction of fork progression through the repeat.

3.4.2 CTCF reduces replication efficiency
Efficiency of replication through DM1 sequence was determined by co-transfection of DM1 plasmids (pDM1-8) with a control plasmid (pSV40) containing the same SV40 origin but lacking any DM1 sequence. When transfected into COS1 cells where CTCF protein levels are not limiting (Vostrov et al., 2002), these templates are compacted into chromatin prior to replication (Cereghini and Yaniv, 1984). By comparing the abundance of each template before and after primate replication, the efficiency of pDM replication in primate cells can be calculated. Cells were transfected with relatively equimolar concentrations of pDM and pSV40 control plasmid (Figure 3.1B), yet after primate replication the pDM plasmids were less abundant than pSV40 overall (Figure 3.1B) reflecting a poor efficiency of replication. The replication efficiency of each template varied depending on both the direction of replication through the repeat as well as CTCF binding site status. Replication initiation 3' of the (CTG)$_{83}$ repeat (pDM5-8) yielded replication efficiencies of around 50% relative to the pSV40 control (Figure 3.1B, lower panel), regardless of CTCF binding site status. However, replication initiation 5' of the repeat (pDM1-4) led to large variation from 7% efficiency in pDM1 with both CTCF binding sites intact, to 73% efficiency in pDM4 with both sites mutated (Figure 3.1B, upper panel). The difference between both the wild type (pDM1) and site II mutation (pDM3) compared to double mutant (pDM4) in
this orientation (66% and 56% lower, respectively) suggests that site I has the largest effect on replication efficiency. This would suggest that CTCF binding may reduce replication efficiency through the DM1 locus in a direction-dependant manner.

3.4.3 Methylation alters replication efficiency
Given that CpG methylation abrogates CTCF binding at the DM1 locus (Filippova et al., 2001), I tested the effect of loss of CTCF binding by methylation as opposed to DNA binding site mutation. There were two possible methylation-specific effects that could alter replication efficiency. First, our lab previously showed that replication efficiency was generally reduced on all repeat-containing templates upon premethylation, regardless of whether there were CpG methylatable sites within the repeat tract (Nichol Edamura et al., 2005). Second, I expected that methylation of CTCF binding site-containing replication templates would increase replication efficiency to resemble templates with mutant binding sites. However, methylation of plasmid DNA is not actively maintained after transfection and is passively lost with replication (Nichol and Pearson, 2002; Nichol Edamura et al., 2005). Therefore a methylated CTCF binding site would be passively maintained after replication, diluting replication efficiency effects caused by loss of CTCF binding.

Plasmids containing CTCF binding sites intact (pDM1,5) or mutated (pDM4,8) with origins 5’ or 3’ of the repeat were methylated at CpG sites throughout with SssI methylase. Plasmids were co-transfected with premethylated pSV40 control plasmid, and assessed for replication efficiency. The plasmids with both sites intact (pDM1,5) had the lowest premethylated replication efficiency, whereas plasmids with the double mutation (pDM4,8) had increases in efficiency of +14% and +11%, when the origin was 5’ or 3’ of the repeat, respectively (Figure 3.1C). Thus both methylation-specific effects predicted to alter replication efficiency may be involved. The templates containing methylated intact CTCF binding sites (pDM1,5) had reduced replication efficiency relative to templates with methylated mutant sites, and yet the mutant templates had lower replication efficiencies than unmethylated templates. However the replication direction effect seen in unmethylated templates (Figure 3.1B) was not maintained. In summary, methylation alters replication efficiency in a complex CTCF-dependent and independent manner.
3.4.4 CTCF and possibly (CTG)\textsubscript{n} repeats stall the replication fork

To more directly assess the role of CTCF binding in the observed replication efficiency reduction of pDM\textsubscript{1-4} plasmids, I considered the possibility that bound CTCF mediated a stall of replication fork progression. This was done by 2D gel electrophoresis of replication intermediates which locates the region where stalling of the replication fork might occur. Replication templates (pDMK\textsubscript{1-4}) were redesigned for optimal observation of replication intermediates (Figure 3.1A). Based on this plasmid design, the position of stalled replication intermediates with different secondary structures was predicted (Figure 3.2A). Replication fork stalling was seen in all plasmids including those lacking CTCF binding sites, suggesting the (CTG)\textsubscript{83} repeat may cause part of the stalling seen along the y-arc (Figure 3.2B). CTCF binding site-dependent stalling was visible in constructs with intact binding sites (pDMK\textsubscript{1,2,3}) along the y-arc at regions mapping to the CTCF binding sites as well as the (CTG)\textsubscript{83} tract. The slight increase in thickness of the y-arc in pDMK\textsubscript{1} relative to pDMK\textsubscript{4} is suggestive of increased fork stalling. However, the strong reduction in replication efficiency of pDMK\textsubscript{1} relative to pDMK\textsubscript{4} (Figure 3.1B) is not likely fully accounted for by the slight difference in fork stalling by 2D gel analysis, and may be caused by a reduction in replication initiation that cannot be quantified by this experiment. Typically the stall along the entire y-arc was reduced in all replicates in constructs containing singly mutated binding sites and a more discrete stall emerged, as well as residual stalling remaining after both sites were mutated (Figure 3.3). However, replication fork stalling analysis was not performed with constructs lacking repeats, therefore stalling cannot be directly attributed to the expanded repeats and is only conclusively evident for CTCF binding sites. Accumulation of replication intermediates was also seen in the region attributed to double replication forks and regressed replication forks, with the largest accumulation seen with the construct lacking CTCF binding site II. Therefore CTCF binding sites around (CTG)\textsubscript{n} repeats of the DM1 locus cause replication fork stalling and account for reduced replication efficiency in the region.

3.4.5 CTCF binding alters repeat instability depending on origin distance to the repeat

Trinucleotide repeat instability has long been attributed to replication defects, amongst other DNA metabolic processes (Pearson et al., 2005). Slippage of DNA either on the template or nascent strand causes deletions and expansions. Replication fork stalling, such as that seen at the CTCF binding sites, may increase opportunities for slippage events to occur. Conversely,
Figure 3.2

A) 

B) 

pDMK_1  pDMK_2  pDMK_3  pDMK_4

Short Exposure

Long Exposure
Figure 3.2 – CTCF binding sites and possibly (CTG)$_{83}$ repeats from the DM1 region stall the replication fork. (A) Schematic representation of predicted replication intermediates by 2D gel electrophoresis of pDMK constructs. The relative locations of various replication intermediates such as replication forks along the y-arc, and double forks and/or regressed forks are shown, as well as the positioning of predicted stalling by CTCF and (CTG)$_{83}$ repeats along the y-arc. (B) 2D gel analysis of replication fork intermediates of pDMK constructs. The expected location of stalls within the repeat tract (bracket), flanking CTCF binding site (black arrow) along the y-arc are illustrated as well as the predominant regressed/double y-arc stall (white arrow).
Figure 3.3 – CTCF binding sites and possibly (CTG)$_{83}$ repeats from the DM1 region stall the replication fork (replicate experiments). Experimental replicates of 2D gel analysis demonstrate consistent stalling by CTCF binding sites, and residual stalling remaining in pDMK$_4$ constructs after mutation of binding sites suggesting pausing at the repeat. Third replicate of pDMK$_3$ shows pausing, though the y-arc is not clear.
replication fork stalling by CTCF may allow the replication machinery to slow as it approaches the TNR in order to replicate the repeat at high fidelity. To test the connection between the observed replication fork stalling and TNR instability, primate replicated templates were assessed for repeat instability by the Stability of Trinucleotide Repeats by Individual Products (STRIP) assay (Cleary et al., 2002). Similar to the replication efficiency assay, pDM and pDMK plasmids were transfected into primate cells and allowed to replicate. Episomal DNAs were isolated and parental plasmids eliminated by DpnI digestion to leave only newly primate replicated plasmids. These DNAs were assessed by the STRIP assay where individual colonies represent individual products of replication. The templates (Figure 3.1A) were used in the STRIP assay in order to assess the connection between CTCF-dependent replication deficiency and instability. Overall the eight plasmids (pDM1-8) did not exhibit significant repeat instability upon primate replication, and therefore did not exhibit CTCF-dependent effects (data not shown).

Replication-induced repeat instability is often directly correlated with the distance of the repeat to the origin of replication (Cleary et al., 2002). Moving the origin even a few base pairs can have large effects, shifting instability from expansion bias to deletion bias. Considering this sensitivity to origin distance, the constructs with a longer distance to the repeat that were optimized for 2D gel studies (pDMK1-4) were also assessed for repeat instability by STRIP. The four primate replicated constructs had instability that was significantly different from background bacterial instability (Figure 3.4). Furthermore, the pattern of instability differed in each of the 4 constructs with various combinations of CTCF binding site mutations. These replication templates demonstrate replication fork stalling and instability, supporting a link between replication fork progression through CTCF binding sites and (CTG)83 repeats and DM1 TNR instability.

3.5 Discussion
Several metabolic processes induce TNR instability including DNA replication, exemplified in patient cells by proliferation-dependent instability (Wohrle et al., 1995; Zatz et al., 1995; Martorell et al., 1998; Yang et al., 2003) and model systems by replication-dependent instability (Cleary et al., 2002; Yang et al., 2003; Freudenreich and Lahiri, 2004). Positioning of replication origins relative to TNR tracts drives instability depending upon the choice of leading
Figure 3.4

A)

- **pDMK\(_1\)** (Deletion bias)
  - Percent molecules observed:
    - Background: 33/100, 58/145, 63/100, 67/145, 14/100, 20/146
    - Transfected: 33/100, 58/145, 63/100, 67/145, 14/100, 20/146
    - *P* = 1.5x10\(^{-9}\)

- **pDMK\(_2\)** (Expand and deletion bias)
  - Percent molecules observed:
    - Background: 21/70, 65/185, 42/70, 63/147
    - Transfected: 21/70, 65/185, 42/70, 63/147
    - *P* = 4.0x10\(^{-9}\)

- **pDMK\(_3\)** (Expansion bias)
  - Percent molecules observed:
    - Background: 26/71, 69/183, 38/71, 69/183
    - Transfected: 26/71, 69/183, 38/71, 69/183
    - *P* = 8.2x10\(^{-8}\)

- **pDMK\(_4\)** (Deletion bias)
  - Percent molecules observed:
    - Background: 21/72, 63/207, 42/70, 63/207, 16/79, 36/207
    - Transfected: 21/72, 63/207, 42/70, 63/207, 16/79, 36/207
    - *P* = 1.0x10\(^{-8}\)

B)

- **pDMK\(_1\)**
- **pDMK\(_2\)**
- **pDMK\(_3\)**
- **pDMK\(_4\)**

Legend:
- **Background**
- **Transfected**

Instability Bias Range:
- -15 to +15
Figure 3.4 – The instability bias of each template used for 2D gel analysis was also analysed by the STRIP assay. (A) The bacterial preparation of parental template molecules contained a distribution of repeat tract lengths, each of which was determined by direct bacterial transformation. Lengths of individual molecules were then assessed and classified into three categories: ‘less than 83 repeats’, ‘83 repeats’ and ‘greater than 83 repeats’ (open bars). After primate replication, the length distribution of the replicated material for each template was similarly determined (solid bars). For each length category, the number of molecules observed out of the total number of molecules analysed is indicated inside the bars. After primate replication, templates for which there was no significant change ($\chi^2$-test, $P > 0.05$) in the length distribution were classed as stable. Indicated in the figure are the $\chi^2$ statistics for each template (2 degrees of freedom). (B) Summary of the instability bias (corrected frequencies of expansion or deletion) generated by primate replication were calculated by subtracting the background repeat length heterogeneities in the parental plasmid preparation.
and lagging strand template and distance of the origin to the repeat in several model systems (Mirkin and Smirnova, 2002; Dere et al., 2004; Cleary and Pearson, 2005; Rindler et al., 2006). Chromosomal origins of replication have been located near several disease-associated TNRs (Nenguke et al., 2003; Brylawski et al., 2006; Chastain et al., 2006; Gray et al., 2007), which may contribute to the large scale instability seen within somatic and germ cells of patients. Cis-elements other than origins can affect instability such as CTCF protein binding sites and methylation (Libby et al., 2008), possibly through replication-based mechanisms such as replication fork shifting or stalling (Cleary and Pearson, 2005).

In this study we show a direct link between the cis-elements found around numerous disease-associated TNR loci and altered replication. CTCF binding sites found surrounding the DM1 repeat strongly reduced replication efficiency in primate cells. The effect was replication direction-dependent, whereby the greatest reduction in efficiency was seen with origins placed upstream of the (CTG)$_{83}$ repeat. This effect may be caused by several factors including the use of (CAG)$_{83}$ versus (CTG)$_{83}$ as lagging strand template, as well as possible differences in the interaction between the oncoming replication fork and the bound CTCF molecule. Given that CTCF protein specifically binds to many DNA sequences and uses different combinations of zinc-fingers at different sites (Ohlsson et al., 2001; Kim et al., 2007), it is possible that the direction from which the replication fork approaches the binding site may affect their interaction. Furthermore, there were large differences in the replication efficiency with the origin upstream of the (CTG)$_{83}$ repeat depending on which CTCF binding site was mutated, suggesting the binding at each site may be inherently different and thus have varied effects on replication. Previous analysis of CTCF binding at the DM1 locus showed stronger binding of CTCF to site I, and the majority of insulator function contributed from this site (Filippova et al., 2001). A CTCF-mediated chromatin structure might explain differences in replication efficiencies between pDM$_2$ and pDM$_3$. The methylated replication efficiency in pDM$_{1,5}$ constructs may represent a combination of decreased replication efficiency due to overall methylation of the plasmid as well as increased replication efficiency due to loss of CTCF binding by methylation. After just one round of replication the methylation pattern may not be consistently maintained, further distorting any obvious CTCF effects.

Replication efficiency varied in a CTCF-dependent manner, therefore I assessed replication fork stalling at the CTCF binding sites and (CTG)$_{83}$ repeat. Stalling of the replication
fork at (CTG)ₙ/(CAG)ₙ repeats has previously been shown in bacteria (Samadashwily et al., 1997), marginally in yeast (Pelletier et al., 2003) and at (CGG)ₙ/(CCG)ₙ repeats in the same primate cells (Voineagu et al., 2009), however stalling at CTCF binding sites has never been reported as these templates did not contain CTCF binding sites. Accumulation of replication intermediates at a position indicative of double y-arcs and/or regressed replication forks in all constructs supports the previous finding that (CTG)ₙ/(CAG)ₙ repeats can form regressed replication fork structures (Fouche et al., 2006), and now the differing abundance of these intermediates between constructs suggests that CTCF binding also affects their formation. Stalling at the CTCF binding site and possibly at the (CTG)₈₃ repeat may allow the bidirectional replication fork progression to catch up from the other side of the plasmid and form double forks if given enough time. CTCF binding sites occur throughout the genome, and the coincidence at many disease-associated TNRs may have a biological role. It is possible that CTCF binding sites around repeats are positioned in order to slow the replication fork to ensure faithful replication of these structure-prone sequences. On the other hand, stalling and regression of the replication fork at these regions may allow more time for mutagenic intermediates to occur since single stranded regions of the replication fork would be more prone to structure formation.

In our final experiment we analysed instability of the CTCF-containing constructs and found that each binding site alters the stability in the same pDMK constructs used for replication stalling assessment. These results contrast STRIP analysis of pDM constructs with an origin closer to the repeat, showing no significant changes in constructs used for replication efficiency. This suggests that instability can vary greatly with small changes in the distance between the origin and repeat that lead to Okazaki initiation changes as previously described in the origin shift model (Cleary and Pearson, 2005). Interestingly, the construct with both sites mutated pDMK₄ had a deletion bias just like a previously published construct lacking flanking sequence (Cleary et al., 2002), though the distance between the origins differed. Clearly the presence of CTCF binding sites affects repeat instability, though the manner by which this occurs remains unclear. CTCF is found as a dimer and can mediate chromatin looping (Pant et al., 2004; Li et al., 2008), raising the possibility that there is interaction between the two binding sites. CTCF also establishes chromatin domains and acts as an insulator (Ohlsson et al., 2001; Dunn and Davie, 2003; Cuddapah et al., 2009), two functions that may contribute to the effect seen in this study. The CTCF binding site-dependent instability effects taken together with the observed
stalling of the replication fork suggest that alterations in replication fork progression lead to instability in this model.

In summary we demonstrate replication fork stalling at CTCF binding sites and \((\text{CTG})_{83}\) repeats as well as instability of the DM1 locus in primate cells. Though CTCF binding sites are not mutated in patients, it is possible that epigenetic changes in different patient tissues may prevent binding of CTCF in this region. Methylation of CTCF binding sites has already been observed in some tissues with unstable repeats in SCA7 transgenic mice (Libby et al., 2008), and in congenital DM1 patients with very large repeat tracts (Steinbach et al., 1998). It is possible that the role of CTCF at both the SCA7 and DM1 loci is to protect against instability, and our results suggest this may occur by slowing of the replication fork in order to faithfully replicate the sequence at high fidelity.
CHAPTER 4 – THE (ATTCT)$_n$ REPEATS OF SPINOCEREBELLAR ATAXIA TYPE 10 DISPLAY STRONG NUCLEOSOME ASSEMBLY WHICH IS ENHANCED BY REPEAT INTERRUPTIONS

Co-authorship statement: This work draws upon a collaborative project between the Pearson and Wang labs. The Pearson and Wang labs conceived of the experimental ideas. Kerrie Nichol Edamura (Pearson lab) prepared the repeat-containing DNA, and Haihe Ruan (Wang lab) performed the nucleosome reconstitution assay. I analysed and interpreted the data, designed the figures and wrote the manuscript that was published in Gene (2009) 434:29-34.
4.1 Abstract

Nucleosome packaging influences many aspects of DNA metabolism such as replication, repair and transcription, and via this link likely has further downstream effects on genome stability. The instability and expansion of repetitive sequences is associated with at least 42 human diseases, yet the molecular conditions contributing to repeat instability have remained largely undetermined. Previously we showed strong nucleosome formation on (CAG)$_n$ repeats associated with spinocerebellar ataxia type 1 and very weak formation on (CGG)$_n$ repeats associated with fragile X syndrome, and that interruption of these repeat tracts made the DNA behave more like random sequences. In this study we determined nucleosome formation on pure and interrupted (ATTCT)$_n$ pentanucleotides associated with spinocerebellar ataxia type 10 (SCA10). We report strong nucleosome formation on (ATTCT)$_n$ repeats, like (CAG)$_n$ tracts. Surprisingly, in contrast to the effect of interruptions on other repeat sequences, interruptions in the expanded (ATTCT)$_{71}$ tracts further strengthened assembly with hyperacetylated histones under physiological conditions with NAP-1. These differences may contribute to phenotypic variation seen between families having pure and interrupted SCA10 repeats, as well as the overall genetic instability at the SCA10 locus.

4.2 Introduction

The expansion of repeat tracts is associated with at least 42 human diseases (Cleary and Pearson, 2003; Pearson et al., 2005). Much attention has been paid to trinucleotide repeat (TNR) diseases such as myotonic dystrophy (DM1) and fragile X syndrome (FRAXA) as at least 14 diseases are caused by the instability of these repeats. However, other non-TNR diseases have been discovered such as myoclonus epilepsy type 1, myotonic dystrophy type 2 (DM2) and spinocerebellar ataxia type 10 (SCA10). SCA10 is caused by a large (ATTCT)$_n$ expansion in intron 9 of the ATAXIN-10 (ATXN10) gene on chromosome 22q13, and is the only known pentanucleotide expansion to cause human disease (Matsuura et al., 2000). SCA10 is characterized by progressive ataxia, seizures and anticipation (Grewal et al., 1998; Matsuura et al., 1999; Rasmussen et al., 2001; Grewal et al., 2002). Unaffected individuals have between 10 and 29 (ATTCT)$_n$ repeats, whereas affected individuals typically have between 800 and 4500 repeats (Matsuura et al., 2000; Matsuura et al., 2004). As has been suggested for other repeat diseases such as DM1 and SCA7, large variations in repeat instability may be associated with a
number of cis- and trans-factors at the ATXN10 locus (Pearson et al., 2005). Recent evidence has implicated the binding of the chromatin insulator protein CTCF adjacent to the SCA7 (CAG)$_n$ tract as a regulator of repeat instability (Libby et al., 2008).

Many cis-elements have been implicated in contributing to repeat instability such as the length and purity of the repeat tract, sequences flanking the repeat, and the surrounding epigenetic environment including DNA methylation and chromatin structure (Cleary and Pearson, 2003). Matsuura et al. examined the SCA10 repeat tract purity in detail, and found that the (ATTCT)$_n$ repeat had (ATGCT)$_n$ interruptions in the 5’ end and (ATTCTAT)$_n$ interruptions at the 3’ end, and revealed interruptions in 71% of alleles sized 17-29 in the penultimate repeat (Matsuura et al., 2006). This suggested that interruptions may explain the complex relationship between repeat length and disease severity. Association of pure and interrupted repeats with varied disease phenotypes also supported the belief that repeat purity may be a disease modifier. Interruptions have also been found in disease-associated genes of spinocerebellar ataxia types 1, 2, 3 and 8 (SCA1, SCA2, SCA3 and SCA8), FRAXA, DM2 and Friedreich’s ataxia (FRDA) (Cleary and Pearson, 2003), though the connections between interruptions, repeat instability and disease severity are not consistent from disease to disease. Generally it is believed that interruptions decrease repeat instability, such as found in FRAXA (Eichler et al., 1994; Kunst and Warren, 1994; Zhong et al., 1995). The SCA10 repeat is distinct from other disease-associated unstable repeats such as (CAG)$_n$ and (CGG)$_n$ repeats, in that it is A+T-rich, hence it is unknown if the general features of other repeats can be extended to it. Importantly, considerable differences exist even amongst the TNRs. Understanding the differences in behaviour of pure and interrupted DNA repeats is necessary in order to determine the role of DNA sequence purity in repeat diseases.

The molecular mechanism for the pathogenesis of SCA10 is unknown. Some suggest that the expanded (AUUCU)$_n$ mRNA products may act as toxic-RNA species, while others suggest neuronal protein binding to DNA is altered by the expansion (Handa et al., 2005; Lin and Ashizawa, 2005; Wakamiya et al., 2006; Waragai et al., 2006). Other data suggest altered chromatinization of the expanded repeat may alter expression of adjacent genes, as occurs in facioscapulohumeral dystrophy (FSHD) and FRDA (Gabellini et al., 2002; Saveliev et al., 2003). Recent data suggest that the SCA10 repeat can act as a DNA unwinding element (DUE) (Potaman et al., 2003) which has been shown to be a critical element for the initiation of DNA
replication (Kowalski et al., 1988). Given the SCA10 expansion causes the region to become highly A+T-rich this may induce DNA unwinding, which may affect gene expression over long distances as well as DNA replication at the ATXN10 locus (Liu et al., 2007). The (ATGCT)_n and (ATTCTAT)_n interruptions found in the SCA10 repeat would decrease and increase the A+T-content of the (ATTCT)_n tract, respectively, which might affect the propensity to form slipped-DNA structures, as occurs for interrupted (CAG)_{44} and (CGG)_{32} repeats (Pearson et al., 1998a), and its ability to readily unwind. The interruptions would also place the repeat out of phase with the DNA helical pitch thereby changing the face of DNA and possibly affect protein-DNA interactions. These features could affect nucleosome formation over the (ATTCT)_n repeats. Thus, interruptions in the (ATTCT)_n repeat may affect the overall DNA structure, and further influence chromatinization of the SCA10 region. However, there is no information about the ability of the (ATTCT)_n repeat to assemble into nucleosomes, or the effect of sequence interruptions on the assembly.

We previously reported altered nucleosome formation on TNRs depending on the repeat sequence, repeat purity, and the acetylation status of the histones (Mulvihill et al., 2005). We found that (CAG)_{49} and (CGG)_{53} repeats formed hypoacetylated nucleosomes more strongly and weakly, respectively, than random sequences. Naturally occurring interruption patterns found in patients caused the nucleosome formation to change, and behave more like random sequence than pure repeats. Here we examined the ability of the (ATTCT)_n repeat to enhance nucleosome formation, and the effect of (ATTTTCT) and (ATATTCT) repeat interruptions found in SCA10 patients and histone acetylation status on nucleosome assembly.

4.3 Materials and methods

4.3.1 DNAs and proteins

Genomic clones named (ATTCT)_{73}p and (ATTCT)_{71}i were obtained by amplifying regions of disease-associated (ATTCT)_n repeats from SCA10 patients. Repeat primed-PCR amplified the large repeat and determined that the (ATTCT)_{71}i plasmid had an interruption pattern of 39-A-1-B-1-B-1-A-11-B-5-B-7 where the numbers represent the number of uninterrupted (ATTCT)_n repeats, and an “A” represents an (ATTTTCT)_1 interruption and a “B” represents an (ATATTCT)_1 interruption (Figure 4.1) (Matsuura et al., 2006). Repeat containing fragments were isolated from the plasmids by EcoRI digestion, with the 104 bp of the endogenous ATXN10
Figure 4.1 – SCA10 genomic fragments (taken from Hagerman et al. (Hagerman et al., 2009)). The fragments used in nucleosome reconstitution were derived from clones of SCA10 patient cells (Matsuura et al., 2006). In the schematic, the (ATTCT)$_n$ repeats are indicated by hollow circles and interruptions indicated by filled circles. The interruption sequence is indicated by an ‘A’ for (ATTTTCT)$_1$ and ‘B’ for (ATATTCT)$_1$. The cloned repeats were isolated by EcoRI digestion, and 104 bp of endogenous ATXN10 flanking sequence is included 5′ of the repeat along with 9 bp and 34 bp of plasmid DNA upstream and downstream of the repeat, respectively.
locus 5' of the repeat as well as 9 bp and 34 bp of plasmid DNA 5' and 3' of the repeat, respectively. In total, the (ATTCT)73p fragment was 512 bp and (ATTCT)71i was 514 bp. A 488 bp pUC19 fragment was obtained by PCR amplification of pUC19 using primers from nt 137 to 161 and from nt 609 to 624. DNA fragments were purified by 4% polyacrylamide gel electrophoresis, and labelled with T4 polynucleotide kinase (New England Biolabs) in the presence of [γ-³²P ] ATP (Amersham Biosciences).

Hypoacetylated histone octamers were isolated from HeLa cells (Orphanides, 1998). Also, HeLa cells were treated with 10 mM sodium butyrate for 24 h to generate hyperacetylated histone octamers. Hyperacetylation of core histone proteins was about 33-fold higher compared to hypoacetylated octamers, as quantified by Western blot analysis using an antibody to acetylated H3 (Upstate Biotechnology). Recombinant histone octamers were purified from *Escherichia coli* which was expressing full-length *Xenopus laevis* histones (Luger et al., 1999). All three types of histones were tested to be fully active in nucleosome reconstitution (Trojer et al., 2007). Recombinant *Drosophila* nucleosome assembly protein-1 (NAP-1) was expressed and purified from baculovirus-infected SF9 cells (Pavri et al., 2006). Remodelling and spacing factor (RSF) was purified from the HeLa nuclear pellet fraction (Loyola et al., 2001).

### 4.3.2 Competitive nucleosome reconstitution

Competitive nucleosome reconstitution by a salt dilution method was carried out as previously described (Hsu and Wang, 2002; Trojer et al., 2007). Briefly, 50 ng of radioactively labelled DNA was mixed with 10 µg of unlabelled calf thymus DNA (Invitrogen) and 2.5 µg of histone octamers in a solution containing 2 M NaCl. NaCl was incrementally reduced to a final concentration of 0.1 M. The assembly mixtures were then electrophoresed directly on 4% polyacrylamide gels to separate free DNA from the nucleosome-assembled DNA. Results were visualized by autoradiography. The amount of DNA in each band was determined by PhosphorImager (Molecular Dynamics) scanning. A "no histones" assembly reaction was performed as a control, in which the reaction mixture did not contain histone octamers.

Competitive nucleosome reconstitution using the NAP-1 histone chaperone protein to more closely mimic physiological conditions was also carried out to examine the ability of the (ATTCT)_n repeats to form nucleosomes. 50 ng of radioactively labelled DNA fragment was mixed with 10 µg of unlabelled calf thymus DNA, 2.5 µg of histone octamers, and 5 µg of
recombinant NAP-1 in buffer (10 mM Hepes, pH 7.6, 50 mM KCl, 5 mM MgCl₂, 0.5 mM EGTA, 0.1 mM EDTA, and 10% glycerol), and incubated at 30°C for 4 h. The assembly mixtures were then electrophoresed on 4% polyacrylamide gels. Results were also visualized by autoradiography. The amount of DNA in each band was determined by PhosphorImager (Molecular Dynamics) scanning.

4.3.3 Assembly and analysis of nucleosome arrays
Assembly reactions were performed as described (Loyola et al., 2001) with 1 µg of plasmid DNA, 0.9 µg of hypoacetylated histones, 0.25 µg of RSF, 3.1 µg of NAP-1, 50 µg of bovine serum albumin (New England Biolabs), 3 mM ATP, 30 mM phosphocreatine, 0.2 µg of phosphocreatine kinase, 10 mM MgCl₂, 50 mM KCl, 10 mM Hepes, pH 7.6, 0.2 mM EDTA, and 5% glycerol at 30°C for 16 h (Loyola et al., 2001). Reaction mixtures were then partially digested with 0.0008 or 0.0025 U of micrococcal nuclease (Sigma). Digestion patterns were analysed by Southern blot probing with a (ATTCT)₅ oligo (Integrated DNA Technologies, Inc.).

4.4 Results
4.4.1 (ATTCT)ₙ repeat is a strong nucleosome positioning element
Towards determining the affinity of SCA10 (ATTCT)ₙ repeats to form nucleosomes, we isolated (ATTCT)ₙ repeats from cloned plasmids containing patient DNA sequences (Matsuura et al., 2006). The pure repeat clone (ATTCT)₇₃p had (ATTCT)₇₃ repeats whereas the interrupted clone (ATTCT)₇₁i had (ATTCT)₇₁ interrupted repeats (Figure 4.1). The patterns of interruptions were those present in described SCA10 families (Matsuura et al., 2006).

Using hyperacetylated, hypoacetylated, or recombinant histone octamers, nucleosomes were assembled upon these repeat-containing DNA fragments as well as a size-matched pUC control DNA used as a baseline of nucleosome assembly, according to the previously described competitive nucleosome reconstitution assay by salt dilution (Mulvihill et al., 2005). Following gel electrophoresis, the nucleosome-associated radiolabelled DNAs appeared as a retarded band above the protein-free radiolabelled DNA. Nucleosome assembly was quantified as the ratio of the nucleosome-assembled DNA to free DNA and normalized to the unit ratio generated from the pUC nucleosome assembly. Overall, both pure and interrupted (ATTCT)ₙ repeats formed nucleosomes significantly more efficiently than non-repetitive control pUC DNA (Figure 4.2,
Figure 4.2

A

<table>
<thead>
<tr>
<th></th>
<th>No Histones</th>
<th>Hypoacetylated</th>
<th>Hyperacetylated</th>
<th>Recombinant</th>
</tr>
</thead>
<tbody>
<tr>
<td>pUC (ATTCT)7/3p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC (ATTCT)7/11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC (ATTCT)7/3p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC (ATTCT)7/11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B

Hypoacetylated

Complex DNA/Free DNA

C

Hyperacetylated

Complex DNA/Free DNA

D

Recombinant

Complex DNA/Free DNA
Figure 4.2 – Pure and interrupted SCA10 (ATTCT)_n repeats form strong nucleosomes with hypoacetylated, hyperacetylated, and recombinant histones by salt dilution nucleosome reconstitution (taken from Hagerman et al. (Hagerman et al., 2009)). (A) Autoradiograms of a nucleosome reconstitution experiment. Lanes 1-3 are DNAs assembled without histones, lanes 4-6 are assembled with hypoacetylated histones, lanes 7-9 with hyperacetylated histones, and lanes 10-12 with recombinant histones. (B), (C), and (D) Graphic representations of nucleosome assembly efficiency performed with hypoacetylated, hyperacetylated, and recombinant histones, respectively. The assembling efficiency is represented by the ratio of DNA in complex over free DNA for each sample. The assembly of pUC DNA with all histones was assigned a value of 1. Each experiment was performed at least 3 times.
Table 4.1). Using hypoacetylated histone octamers, pure and interrupted repeats formed nucleosomes 7.5- and 8.1-fold respectively, more efficiently than the pUC DNA (p=0.001) (Figures 4.2A and B). The efficiencies of nucleosome assembly with hyperacetylated histones were 9.9- and 12.2-fold higher on pure and interrupted repeats, respectively, as compared to the pUC19 control fragment (p=0.03 and p=0.01). Finally, using recombinant histone octamers, the pure and interrupted repeats assembled nucleosomes more strongly than the control fragment by as much as 19.3- and 20.1-fold, respectively (p=0.02 and p=0.01). However, no significant difference in nucleosome assembly between the pure and interrupted repeats was found by this method. Overall (ATTCT)$_n$ repeats have the tendency to assemble nucleosomes stronger than random non-repetitive DNA, regardless of histone type.

4.4.2 Repeat interruptions increase hyperacetylated nucleosome formation on the (ATTCT)$_n$ repeat under physiological conditions with NAP-1

In the salt dilution method above, large variability between experimental replicates was observed. During salt dilution, (ATTCT)$_n$ repeats may undergo structural changes that could affect the stability of nucleosome formation, and cause the variability. This could mask the differences in nucleosome assembly, between pure and interrupted (ATTCT)$_n$ repeats. To further elucidate possible differences in nucleosome assembly of pure and interrupted (ATTCT)$_n$ repeats, we performed competitive nucleosome reconstitution under physiological conditions using NAP-1. The NAP-1 has been shown to chaperone histones to DNA, and prevent non-specific aggregation of histones (Fujii-Nakata et al., 1992; Ito et al., 1996; Gemmen et al., 2005). Using the same pure and interrupted (ATTCT)$_n$ repeats, nucleosome assembly was performed with the addition of NAP-1. Hyperacetylated histones showed a greater than 3-fold increase in nucleosome assembly on interrupted (ATTCT)$_n$ repeats over pure repeats (p=0.03) (Figure 4.3). There was no significant difference in nucleosome assembly between pure and interrupted (ATTCT)$_n$ repeats with hypoacetylated and recombinant histones. Therefore under more physiological conditions, interrupted (ATTCT)$_{71}$ repeats display a stronger assembly of hyperacetylated nucleosomes than pure (ATTCT)$_{73}$ repeats.

4.4.3 Both pure and interrupted repeats form evenly-distributed nucleosome arrays

Given that both pure and interrupted (ATTCT)$_n$ repeats display a strong ability for nucleosome
Table 4.1 Competitive nucleosome reconstitution (taken from Hagerman et al. (Hagerman et al., 2009)).

<table>
<thead>
<tr>
<th>DNA</th>
<th>COMPLEX/FREE</th>
<th>FREE ENERGY (cal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Salt dilution method</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypoacetylated histones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{73}p</td>
<td>7.5±3.2</td>
<td>-1098±387</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{71}i</td>
<td>8.1±3.8</td>
<td>-1133±351</td>
</tr>
<tr>
<td>hyperacetylated histones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{73}p</td>
<td>9.9±4.4</td>
<td>-1305±322</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{71}i</td>
<td>12.2±4.7</td>
<td>-1446±266</td>
</tr>
<tr>
<td>recombinant histones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{73}p</td>
<td>19.3±9.8</td>
<td>-1761±304</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{71}i</td>
<td>20.1±8.5</td>
<td>-1793±265</td>
</tr>
<tr>
<td><strong>NAP-1 method</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{73}p</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{71}i with hypoacetylated histones</td>
<td>1.5±0.9</td>
<td>-118±124</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{71}i with hyperacetylated histones</td>
<td>3.3±1.2</td>
<td>-688±209</td>
</tr>
<tr>
<td>(ATTCT)\textsubscript{71}i with recombinant histones</td>
<td>1.3±0.3</td>
<td>-160±127</td>
</tr>
</tbody>
</table>

* The ratio of DNA bound by histones to free DNA for each sample was determined by measuring the amount of DNA in each radioactive band by a phosphorimager. The ratios for the sample DNA were normalized against that of the pUC19 DNA fragment. The ratio for the pUC19 DNA fragment was assigned a value of 1. The free energy of nucleosome assembly was calculated according to the equation \( E = RT\ln K - RT\ln Q \) where \( E \) represents the free energy of assembly of the repeat-containing fragment, \( K \) is the ratio of DNA in complex to free DNA for pUC19 fragment, and \( Q \) is the ratio of DNA in complex to free DNA for the repeat-containing fragment (Shrader and Crothers, 1989; Wang and Griffith, 1996b). The free energy for the pUC19 DNA was defined as zero.

**A similar calculation was carried out as described for the salt dilution method. However, the ratios for the (ATTCT)\textsubscript{73}p fragment assembled with various histones were assigned a value of 1, and the free energy for the (ATTCT)\textsubscript{73}p fragment was defined as zero.
Figure 4.3 – Interrupted SCA10 (ATTCT)$_n$ repeats form stronger nucleosomes than pure repeats with hyperacetylated histones by NAP-1 reconstitution (taken from Hagerman et al. (Hagerman et al., 2009)). (A) Autoradiogram of nucleosome reconstitution experiments with NAP-1. (B) Graphic representation of nucleosome assembly efficiency with NAP-1. The assembling efficiency is represented by the ratio of DNA in complex over free DNA for each sample. Nucleosome reconstitution was performed with either hypoacetylated, hyperacetylated or recombinant histones (as labelled). The assembly of (ATTCT)$_{73}$p with histones was assigned a value of 1. Each experiment was performed 3 times.
assembly, we next investigated whether both repeats formed well positioned nucleosomes in a nucleosome array setting. Nucleosome arrays were assembled onto plasmids with pure or interrupted (ATTCT) \(_n\) repeats, using NAP-1 and the RSF protein. The RSF protein spaces nucleosomes and remolds chromatin (Loyola et al., 2001). The pUC19 control plasmid was analysed in parallel as well. All three plasmids were assembled with hypoacetylated histones, and the reaction mixtures were partially digested with micrococcal nuclease to examine the periodicity of nucleosome formation. In ethidium bromide stained gels (Figure 4.4A), 180 bp nucleosomal ladders were observed in all three samples, indicating that all three plasmids form regularly spaced nucleosome arrays. In both pure and interrupted (ATTCT) \(_n\) repeats, the nucleosomal ladders were further confirmed by Southern blot with a (ATTCT) \(_5\) probe (Figure 4.4). This suggests that phased nucleosomes, rather than altered periodicities or aggregates, are formed over the (ATTCT) \(_n\) repeat regardless of the presence or absence of interruptions.

### 4.5 Discussion

In this study we examined the ability of the (ATTCT) \(_n\) repeats to assemble nucleosomes, and the effects of disease-associated (ATTCT) \(_n\) repeat purity and histone acetylation on the assembly. We found that SCA10 (ATTCT) \(_n\) repeats are strong nucleosome assembling sequences compared to random sequences, with stronger formation on interrupted repeats with hyperacetylated histones at physiological conditions (with NAP-1), as well as formation of well-positioned nucleosomes, not aggregates, on the repeats. The differences observed in this study with over (ATTCT) \(_{70}\) repeats may be considerably amplified in the genomic context where the SCA10 disease-causing expansions are typically more than 800 and up to 4500 (ATTCT) \(_n\) repeats (Matsuura et al., 2004). Such increased binding has been observed for the large (CTG) \(_n\) tracts of DM1 (Wang et al., 1994; Wang and Griffith, 1995). The increased nucleosome assembly on the SCA10 repeats was similar to that on the (CAG) \(_n\) repeats of the SCA1 and DM1 locus (Table 4.1, (Wang et al., 1994; Wang and Griffith, 1995; Mulvihill et al., 2005)), and therefore the assembly on the interrupted SCA10 repeat could be even greater.

Nucleosome formation on repeats has been studied for many years, with focus on diseases caused by TNR expansions (Wang, 2007). (CAG) \(_n\) repeats from diseases such as DM1 and SCA1 are the strongest nucleosome positioning elements (Wang and Griffith, 1995) whereas (CGG) \(_n\) repeats associated with FRAXA preferentially exclude nucleosomes (Godde et al., 1996;
Figure 4.4 – Pure and interrupted SCA10 (ATTCT)$_n$ repeats form well positioned nucleosomes (taken from Hagerman et al. (Hagerman et al., 2009)). (A) Ethidium bromide stained gel of DNA assembled on nucleosomes with RSF. (B) Southern blot of DNA assembled on nucleosomes with RSF and hybridized with a (ATTCT)$_5$ probe.
Wang et al., 1996; Wang and Griffith, 1996a). The exclusion of nucleosome formation at (CGG)$_n$ repeats may be associated with both fragile site formation (Wang et al., 1996) and altered transcription dependent upon repeat size (Tassone et al., 2000; Tassone et al., 2007). Understanding the in vitro behaviour of the (ATTCT)$_n$ repeat may similarly lead to an understanding of its behaviour in cells and downstream metabolic and pathogenic mechanisms. SCA10 repeats are known to form unwound DNA and the high A+T-content is attributed to its serving as a DUE (Kowalski et al., 1988) and facilitating the in vitro activation of replication origins (Kowalski et al., 1988; Potaman et al., 2003). In vivo, there is a natural origin of replication in the ATXN10 gene just 5’ of the (ATTCT)$_n$ repeat which produces 5- to 10-fold more nascent DNA in SCA10 patient cells than controls (Liu et al., 2007). It is noteworthy that Kemp et al. found that alteration of chromatin packaging can dramatically affect the pattern of origin activation at various replication origins (Kemp et al., 2005). Thus, the strong nucleosome formation at (ATTCT)$_n$ sequences, and differences between pure and interrupted repeats may affect origin activity, transcription and DNA instability. The replication machinery must unwind and separate duplex DNA strands from each other and from histones, and in the process, replication fork progression may be stalled or slowed by the presence of hyper stable nucleosomes. This could be further exacerbated by the presence of interruptions. In addition, initiation of replication and transcription is controlled, at least in part, by the degree of local unwinding of nucleosomal DNA (Kowalski et al., 1988; Potaman et al., 2003). This unwinding can be regulated by histone acetylation; increased acetylation results in a more loosely wound structure with reduced thermal melting temperatures (Yau et al., 1982; Thomsen et al., 1991) allowing access of replication or transcription factors. Furthermore, acetylation of histones can effectively induce negative supercoiling (Norton et al., 1989) which may enhance DNA unwinding (Kowalski et al., 1988; Potaman et al., 2003). The DNA unwinding activity of the SCA10 repeat, coupled with our finding of its preferential binding to acetylated histones, along with their association with replication origins (Aggarwal and Calvi, 2004; Zhou et al., 2005), and the cell cycle regulation of histone hyperacetylation (Bradbury, 1992), prompts us to propose that histone acetylation at flexible repeat tracts may participate in replication initiation (Liu et al., 2007).

Repeat interruptions occur in various human diseases (DM2, FRAXA, FRDA, SCA1, SCA2, SCA3, SCA8, and SCA10), and only rarely detected in other diseases like DM1.
(Leeflang and Arnheim, 1995; Musova et al., 2009). Unlike the reduced nucleosome formation caused by interruptions in (CAG)$_n$ repeats, we show increased formation on interrupted (ATTCT)$_n$ repeats with hyperacetylated histones under physiological conditions. This difference may be reflected upon in the stability of the locus, since interrupted TNRs are typically more stable (Eichler et al., 1994) whereas interrupted SCA10 pentanucleotides found in patients are unstable (Matsuura et al., 2006). The strong nucleosome assembly by the (ATTCT)$_n$ repeats is likely due to the abundance of flexible TA dinucleotides, which increase bending of the repeats to wrap DNA around histones (Shrader and Crothers, 1989; Lowary and Widom, 1998; Packer et al., 2000; Tolstorukov et al., 2007). The interruption pattern described in this paper increases the number of flexible dinucleotides in the repeat tract, and may thus further increase the DNA bending ability, making nucleosome formation more favourable. Also, hyperacetylated histones are often associated with transcriptionally active genes; preferential formation of hyperacetylated nucleosomes onto the interrupted repeat may prime the interrupted alleles into a different transcription pattern from the pure alleles. Interestingly, the patient families that these pure and interrupted sequences were derived from have different disease symptoms, with the pure repeats found in family 1 that has no seizures and the interrupted repeats found in family 2 that has seizures (Matsuura et al., 2006). Most untranslated repeat disease pathogenic mechanisms are centralized around RNA expression, toxic RNAs (Ranum and Cooper, 2006), DNA repair (Slean et al., 2008) and gene chromatinization (Cho et al., 2005), therefore nucleosome formation on the repeats can be important in understanding the disease mechanism. The increased formation of nucleosomes on interrupted versus pure (ATTCT)$_n$ repeats could add to our understanding of SCA10 pathogenesis and instability.
CHAPTER 5 – SUMMARY AND FUTURE STUDIES
5.1 Summary

This thesis describes a culmination of experiments addressing the role of *cis*-elements in disease associated repeat instability. Since the discovery of TNR diseases in the early 1990’s (Fu et al., 1991; La Spada et al., 1991; Brook et al., 1992; Fu et al., 1992), great effort has been put forth to understand the mechanism of disease-associated repeat instability. The unique mechanism of genetic anticipation, with continual mutations and worsening of the disease, contradicts conventional Mendelian assumptions of inheritance of unchanged parental alleles as well as the static nature of mutations. Repeat sequences behave differently depending on the genomic context, therefore it is believed that *cis*-elements contribute to ongoing instability and possibly disease. *Cis*-elements may also variably contribute to instability in a tissue- and development-specific manner. For the purpose of this thesis, I defined *cis*-elements as sequences and epigenetic modifications around or within a disease-associated repeat. I focussed on the internal *cis*-element of repeat purity, and external *cis*-elements of origins of replication, chromatinization, methylation and CTCF binding sites. My experiments questioned the involvement of these *cis*-elements in the behaviour of the repeat, in order to ultimately understand the mechanism of repeat instability and disease.

In Chapter 2, I described the assessment of SCA7 mouse lines with intact and mutated CTCF binding sites from the human *ATAXIN-7* locus. These mice had an expanded (CAG)$_{94}$ repeat tract, greater than the known threshold causing disease. Mice with a mutated CTCF binding site transmitted more unstable repeats to the next generation, and had significantly larger expansions and contractions than mice with a wild type CTCF binding site. This finding was mirrored in the germline as sperm contained more unstable alleles, and the effect was amplified as the mice aged. Loss of CTCF binding also induced somatic repeat instability; kidney and some brain regions had increased instability which was also increased in older mice. Methylation of the CTCF binding site was found in wild type mice in the liver, suggesting that the cause of the high level of repeat instability in this tissue was also loss of CTCF binding; a finding further supported by the assessment of a mouse with an uncharacteristically high degree of repeat instability with unexpected methylation in the kidney. Overall this work supported a role for CTCF in disease-associated SCA7 repeat instability, was the first demonstration of an endogenous *cis*-element contributing to instability, and was the first demonstration of epigenetic regulation of repeat instability.
In Chapter 3, I described the effects of CTCF on replication of the DM1 locus. Through a primate model of replication, I determined that CTCF reduces the efficiency of replication of plasmids in a direction dependent manner, and then showed this effect was caused by replication fork stalling at CTCF binding sites. This was the first evidence of replication fork stalling at CTCF binding sites, and preliminary instability assessment suggested that this stalling may contribute to repeat instability at the disease locus. The function of CTCF binding sites around many repeats associated with disease was previously unknown, except for having insulator activity at the DM1 locus (Filippova et al., 2001). During the course of this project, an origin of replication was mapped to the SCA7 locus, positioning CTCF binding between the origin and the repeat (Nenguke et al., 2003). My results suggested that CTCF binding sites may be localized at repeat loci in order to slow the replication fork before it reaches the repeat. Many of the repeat loci, including SCA7 and DM1, have both CTCF binding sites and origins of replication. The CTCF binding site often lies between the origin and the repeat. Coupled with my results, this suggests that the positioning of origins of replication and CTCF binding sites around repeats may be a general phenomenon and have a functional role in reducing instability driven by replication.

In Chapter 4 I described nucleosome formation on (ATTCT)_n repeats of the SCA10 locus. I found that (ATTCT)_n repeats promote strong nucleosome formation, whether the repeat was pure or interrupted. Quantification of this effect revealed that (ATTCT)_n repeats formed nucleosomes more strongly than (CTG)_n repeats (Figure 5.1A), which were previously labelled the “strongest known natural nucleosome positioning elements” (Wang and Griffith, 1995). Using protocols that more closely mimicked physiological conditions the interrupted repeats formed moderately, yet significantly, stronger nucleosomes with hyperacetylated histones. A more open chromatin conformation at (ATTCT)_n may explain origin activation in SCA10 patient chromosomes (Liu et al., 2007). Altered nucleosome formation on pure versus interrupted expansions may explain why the families with pure and interrupted repeats have a different propensity for seizures, though other modifiers outside the repeat may also be involved.

5.2 Future studies
Overall my findings support a role for cis-elements in altering the characteristics and behaviour of repeats, however many questions remain.
Figure 5.1 – Summary of findings and hypotheses. (A) Comparison of nucleosome assembly on repeat sequences demonstrates that (ATTCT)$_n$ repeats are now the strongest known nucleosome positioning elements, stronger than (CAG)$_n$ repeats that were previously named the strongest. Interruptions of the pentanucleotide (iATTCT) slightly increase nucleosome formation. (B) The replication fork shift theory (Cleary and Pearson, 2005); a hypothesis that the Okazaki initiation zone is shifted by cis-elements between the origin of replication (ori) and the repeat. This model could explain how CTCF induces repeat instability at the myotonic dystrophy locus, and is a focus of future studies.
5.2.1  What is the molecular consequence of CTCF binding at the DM1 locus?

The effect of CTCF binding on the replication fork and repeat instability supports the fork-shift model for repeat instability (Cleary and Pearson, 2005; Mirkin, 2005). The fork shift model postulates that cis-elements such as CTCF binding between the origin and repeat will affect the placement of Okazaki initiation zones to alter instability (Figure 5.1B). To test this hypothesis, the sequence of RNA-primed Okazaki fragments from the DM1 plasmids can be determined by previously established methods (Hay et al., 1984; Gerbi and Bielinsky, 1997). If CTCF binding alters Okazaki initiation zones, then the templates with wild type CTCF binding sites will have different initiations than plasmids with mutated CTCF binding sites.

The mechanism by which CTCF stalls the replication fork is unknown. It is possible that structure formation of the repeats and CTCF create a physical barrier. CTCF is known to induce looping (Ling et al., 2006; Splinter et al., 2006), therefore the presence of CTCF binding sites flanking disease-associated repeats may also induce looping. In order to determine whether CTCF affects structure or loop formation, electron microscopy studies using my DM1 templates containing wild type and mutant CTCF binding sites incubated with purified CTCF protein will show whether structures form in vitro. Currently it is not possible to visualize structure formation in vivo.

5.2.2  Does CTCF protect the SCA7 repeat from other cis-elements driving instability?

In Chapter 2 I suggested that CTCF protects the SCA7 repeat from other cis-elements driving repeat instability. I proposed that the origin of replication localized 3' to the repeat and a CTCF binding site (Nenguke et al., 2003) may be the element responsible for driving instability and that CTCF may partially protect the repeat from instability driven by the origin. Future studies should map the origin in the SCA7 mice to determine whether it is active in the mice, and whether replication through the CTCF binding site is affected. The origin is surrounded by 5 kb of endogenous human ATAXIN-7 sequence upstream, and 8 kb downstream, increasing the likelihood that even within the mouse context, the human origin will still initiate replication. It is possible that CTCF binding may slow replication fork progression through this region before the fork reaches the disease-associated repeat, and slowing would be lost in the mice with a mutated binding site. Origin mapping experiments could demonstrate this by showing a paucity of newly replicated sequences in the region between the CTCF binding site and the repeat. Alternatively,
2D gel electrophoresis studies of *in vitro* progression of the replication fork through the SCA7 CTCF repeat would also support this theory.

5.2.3 *Does CTCF have universal functions at repeat disease loci?*

The frequent concurrent localization of CTCF binding sites to disease-associated repeats, taken together with my findings that CTCF affects replication and instability, suggest that the positioning of CTCF binding sites between repeats and origins of replication has a functional role in repeat disease and is not by chance. I have shown that CTCF binding sites induce replication fork stalling, and the absence alters instability in a manner that likely depends on other *cis*-elements nearby. Designing a DM1 mouse model with various permutations of mutated CTCF binding sites similar to the SCA7 mouse model would begin to directly address whether CTCF has universal functions at repeats. SCA7 and DM1 have the highest degree of somatic and germline instability, and should therefore be the two main diseases of focus in order to readily observe large changes in instability. It will be important to study other TNR diseases with origins and CTCF binding sites, but alterations in instability may be less dramatic and more difficult to detect. Replication efficiency and replication fork stalling studies could be executed with replication templates containing disease-associated repeat sequences and flanking CTCF binding sites such as those found at HD, FRAXA, SCA2, SCA7, FSHD, DRPLA, SCA10, SBMA, FRAXE and DM2 sequences. If replication differences between functional and mutant binding sites were observed, STRIP analysis could be performed on these constructs to determine whether the observed replication defect also leads to instability.

Expansion of repeats in human disease has not been found in homologous genes in other organisms, and repeat expansion has been identified as the causative mutation of a condition outside humans only twice (Lohi et al., 2005; Sureshkumar et al., 2009). It is possible that sequences surrounding repeats evolved to create CTCF binding sites in humans in order to prevent instability. Conversely, it is possible that CTCF binding sites are found in homologous genes but were not originally selected to protect against repeat instability. CTCF binding sites are often, if not always, found in UTRs and introns, and are therefore able to evolve more quickly without mutating the associated gene. I propose experimental comparisons of disease-associated homologous genes by computational techniques to look for sequence conservation at CTCF binding sites and compare the binding sites in other less complex organisms to both the
human sequence and the published consensus sequence (Kim et al., 2007). ChIP on chip studies of CTCF pulldowns in other organisms using an array of genomic sequence around and within homologues will also help determine if CTCF binding at these loci came before or after the emergence of repeat loci associated with disease. If CTCF binding sites arose before the wild type repetitive sequences, then CTCF may have evolved a function to protect against instability after the repeats arose. Otherwise if CTCF binding sites arose after repetitive sequences in genes, then CTCF binding sites may have evolved in these regions to counteract the possible mutagenic effects of the repeat. These studies will help determine how CTCF binding sites arose at so many disease loci.

5.2.4 Does CTCF-related instability affect phenotype?
An important question remaining in the field of repeat-associated diseases is whether somatic instability directly contributes to disease phenotype. In order to address this question, the SCA7 mouse can be used as a model to assess whether the instability induced by loss of CTCF binding leads to phenotypes found in SCA7 patients such as ataxic movement and retinopathy. Bidirectional transcription and translation need to be assessed, as well as characterization of the mouse phenotype. Preliminary studies by the La Spada lab suggest mutation of the CTCF binding site affects transcription and possibly phenotype (Pineda et al., 2006). It is possible that CTCF binding alterations contribute to disease phenotype in an instability-independent manner. If the SCA7 mouse phenotype is different between the wild type and mutant CTCF binding site lines, CTCF binding and/or methylation status at the binding site in patient tissues should be determined.

5.2.5 What is the molecular consequence of interruptions in SCA10 repeats?
SCA10 is the least understood repeat disease I have discussed in this thesis. Little is known about the gene function, and the pathogenic mechanism. The current belief is that expansion of the (ATTCT)$_n$ repeat acts as a DNA unwinding element to induce an origin of replication near the expanded allele (Potaman et al., 2003; Liu et al., 2007). In order to test this, SCA10 patient cell lines with smaller and larger repeats should be assessed for origin activation to determine whether the origin activation is proportional to the repeat size. The increased nucleosome
formation on interrupted repeats may also show increased origin activity, therefore measuring the firing of the origin in cell lines from the pure and interrupted repeat families.

Nucleosome structure is also associated with transcription, with a more open hyperacetylated nucleosome associated with increased gene activity. To test whether nucleosome formation differences between pure and interrupted repeats affects expression, the levels of both sense and possible antisense transcripts and acetylation status should be measured in several tissues of the two families from which the pure and interrupted constructs were developed (Matsuura et al., 2006). If repeat purity contributes to variation in phenotype through transcription or replication processes, then differences in origin activity or transcription may be seen.

5.3 Concluding comments
The work and ideas presented in this thesis suggested that cis-elements surrounding and within repeat sequences contribute to the molecular attributes of the repeat such as nucleosome formation, and mutability of the repeat. I identified the first cis-element, CTCF binding sites, regulating repeat instability. I suggested molecular mechanisms by which CTCF exerts its effects, and linked this element to many human diseases. The mutability induced by this cis-element may directly lead to increased severity of disease through anticipation in SCA7, and supports hypotheses that differential instability in tissues may contribute to tissue-specific phenotypes by regulation of cis-elements and trans-factors.
REFERENCES


Burcin, M., Arnold, R., Lutz, M., Kaiser, B., Runge, D., Lottspeich, F., Filippova, G.N., Lobanenkov, V.V. and Renkawitz, R. Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. Mol Cell Biol 17 (1997), pp. 1281-8.


Cleary, J.D., Hagerman, K.A., López Castel, A., Tomé, S., Foiry, L., Paradis, I., Panigrahi, G.B., Dorschner, M.O., Sroka, H., Chitayat, D., Stamatoyannopoulos, J.A., Drouin, R., Gourdon,


system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* 73 (2001), pp. 56-65.


drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. *Hum Mol Genet* **12** (2003), pp. 41-50.


Rasmussen, A., Matsuura, T., Ruano, L., Yescas, P., Ochoa, A., Ashizawa, T. and Alonso, E.


