Investigation of the Epigenetic Landscape at Disease-Causing Polymorphic Repeat Loci

by

Rachel Mador-House

A thesis submitted in conformity with the requirements for the degree of Master's of Science
Molecular Genetics
University of Toronto

© Copyright by Rachel Mador-House 2014
Investigation of the Epigenetic Landscape at Disease-Causing Polymorphic Repeat Loci

Rachel Mador-House

Master's of Science

Molecular Genetics, University of Toronto, 2014

Abstract

There are over 40 genetic diseases caused by repeat locus-specific instability. The causes of repeat instability are poorly understood and a mechanism that explains their occurrence in full has yet to be identified. Numerous studies have shown epigenetic modifiers and cis-elements to play a role in repeat instability. Here, ENCODE datasets (histone modifications, DNA methylation, CTCF binding, DNase I hypersensitivity, and CpG sites) were analyzed at disease-causing repeat loci, control non-disease causing repeat loci, mono- and dinucleotide repeat loci, short interspersed nuclear element loci, long interspersed nuclear element loci, telomeric repeat loci, and classical satellite I, II, and III loci. Loci were assessed in hippocampus, skeletal muscle, liver and kidney. We identified epigenetic mark patterns at these loci. A striking symmetric distribution pattern centrally localized around specific tandem repeats showing epigenetic enrichment or depletion of marks was observed indicating a possible nucleosome positioning effect. Tissue-specific patterns were also identified.
Acknowledgments

I would like to thank my Master’s supervisor, Dr. Christopher Pearson, for the continued patience and compassion he has shown me over the last 3 years. I have learned so much, grown in diverse ways, and acquired many skills in your lab, for which I will always be grateful. I would also like to thank my committee members, Dr. Marc Meneghini and Dr. David Bazett-Jones, for the continued support and dedication to my current and future goals. I am also grateful to Dr. Mike Wilson and his amazing lab for taking me on as an “honorary” member. The help I received from the Wilson lab was critical in my ability to complete this degree.

Throughout my schooling I had the pleasure of working with truly inspiring individuals who I would like to take the time to acknowledge. From Mike Wilson’s lab I would like to thank Dr. Alejandra Rivera who took the time to be my teacher, and my friend. Also, Lina Antounians and I did many wet lab experiments together and lamented about ChIP-seq together. Jen Luo, thank you for sharing this experience with me, I found so much strength from our times together. Shirley Guan, thank you for making me laugh, hugging me when I cried, and looking at wedding dresses with me! Jodie Simard, thank you for the support, kindness, and empowerment you gave me, your caring nature makes you a wonderful friend and a fantastic mother. I also had the pleasure to work with Neha Thakkar, Dr. Gagan Panigrahi, Dr. Stephanie Tome, Dr. Lisa Yu, Monica Schmidt, Bilal Suboor and Dr. Stella Lanni.

Finally, I would like to thank my family. My brother, Nathan Alexander Mador-House is a continued inspiration for me in ever faucet of my life. He fills me with hope, courage, and excitement. My mother, Janet Louise Mador is the most self-sacrificing and loving person I know. I aspire to no better role model. And my fiancé, Patrick Arthur Cuaresma, who I shared many challenging and rewarding times with over the past years yet he continued to be supportive and positive through it all. I am very lucky to have such an amazing family and I am also lucky to have friends who helped, supported and celebrated with me during this process; Dorina Simeonov, Tsz Lung-Cheung, Nikol Paar, Marko Markovic, Arlie Cuaresma, Lara Deutch, and Robin freaking Tunley.
# Table of Contents

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

Acknowledgments ............................................................................................................................... iii

Table of Contents ............................................................................................................................... iv

List of Tables ....................................................................................................................................... vi

List of Figures ..................................................................................................................................... vii

1 Introduction .................................................................................................................................... 1

1.1 Repetitive units in the genome ..................................................................................................... 1

1.1.1 Tandem repeats ....................................................................................................................... 1

1.1.1.1 Evolutionary conservation of tandem repeats ................................................................. 2

1.1.2 Telomeric Repeats ................................................................................................................... 2

1.1.3 Mono- and dinucleotide repeats ............................................................................................ 3

1.1.4 Satellite I, II, III classical repeats ......................................................................................... 4

1.1.5 Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) ................................................................................................................................. 5

1.2 Epigenetic Marks .......................................................................................................................... 7

1.2.1 Histone modifications .............................................................................................................. 7

1.2.2 DNase I hypersensitivity ......................................................................................................... 8

1.2.3 CTCF binding ........................................................................................................................ 10

1.2.4 DNA CpG sites and DNA CpG methylation .......................................................................... 10

1.2.5 mRNA sequencing ................................................................................................................ 12

1.3 Disease-causing polymorphic repeats and instability ................................................................. 12

1.3.1 46 disease-causing microsatellite loci ................................................................................... 12

1.3.2 Mechanisms of instability ....................................................................................................... 13

1.3.3 Tissue specific instability ...................................................................................................... 16
List of Tables

Table 2.1. 46 disease-causing repeat loci locations ................................................................. 22

Table 2.2. Proportion of repeat sequences in 46 disease-causing repeat loci population and control population ........................................................................................................... 25

Table 2.3. Mono- and Dinucleotide repeat loci ................................................................. 28

Table 2.4. Datasets used for analysis ......................................................................................... 29
List of Figures

Figure 1.1. Repetitive elements in the human genome......................................................... 6
Figure 1.2. Epigenetic marks surrounding disease-causing repeat loci................................. 9
Figure 2.1. Hippocampus all epigenetic marks.................................................................. 32
Figure 2.2. Telomeric repeats, SINEs, LINEs and 46 disease-causing repeats in hippocampus.. 35
Figure 2.3. Mononucleotide, dinucleotide and 46 disease-causing repeat loci in hippocampus.. 37
Figure 2.4. Satellite I, II, III, and 46 disease-causing loci in hippocampus.......................... 40
Figure 2.5. Repeat composition analysis for disease-causing repeat loci and satellite I, II and III loci in hippocampus ......................................................................................... 42
Figure 2.6. Tissue specific evaluation of epigenetic marks at disease-causing and non-disease-causing repeat loci in hippocampus and skeletal muscle......................................................... 46
Figure 2.7. Disease-causing and non-disease-causing repeat loci enrichment for H3K9ac in hippocampus .................................................................................................................. 48
1 Introduction

1.1 Repetitive units in the genome

The absence of a correlation between the complexity of an organism and their DNA content was a mystery for early molecular geneticists. Unicellular have some of the largest genomes. The discovery of repetitive DNA within genomes began to unravel this dilemma. For example, repeats encompass 46% of the human genome and many genomes of simple organisms contain more repeats. There are a number of theories as to why some organisms have a larger portion of repetitive DNA such as providing an adaptive function, protecting against mutation, and aiding in chromatin structure. There are many varieties of repeats within the genome some with specific functions and others warrant further investigation.

1.1.1 Tandem repeats

Tandem repeats (TRs) are a category of repetitive DNA where each repeat unit is located next to the other (in tandem) (Fig. 1.1). TRs can be 1 to 9 nucleotides in length and are also known as microsatellites, simple sequence repeats (SSRs), or short tandem repeats (STRs). TRs that are 10 nucleotides or longer are called minisatellites and extremely long units of greater than 135 nucleotides constitute megasatellites (Fig 1.1).

TRs are unstable with mutation rates often 10 to 100,000 times higher than average mutation rates in other parts of the genome. Because of their variability, TRs are also known as VNTR (variable number of tandem repeat) elements. Mutations in TRs are due to repeat polymorphisms that occur when the number of the repeating units changes. The vast majority of these changes consist of the deletion or addition of a full repeat unit sequence.

TRs are often located within genes and regulatory regions. While some repeats are located within gene deserts, TRs are often found within coding and regulatory regions with approximately 17% of genes in the human genome containing repeats within their open reading frames\(^1\). The percentage of genes with repeats in open reading frames is similar in many other organisms\(^2\). Among repeats found in coding sequences, repeats with units that contain a multiple of 3 nucleotides, such as a tri- and hexanucleotide repeat, are by far the most common\(^3\).
1.1.1.1 Evolutionary conservation of tandem repeats

The representation of TRs is not uniform in all categories of genes. Analysis of TRs in the human genome shows that genes in certain categories of biological function are enriched for variable TRs\(^4\). Functional enrichment among genes with TRs has also been observed in yeast, where TRs tend to be found in genes involved in particular processes, with microsatellites being primarily found in regulatory genes such as those encoding transcription factors and minisatellites in cell-wall genes\(^2\).

Once believed to be neutral variations in junk DNA, TR polymorphisms are now known to be, in certain instances, linked to variation in function. An increasing body of evidence shows that the genomes of diverse organisms contain TR sequences that not only have specific biological functions, but by means of their intrinsic instabilities, may also confer faster rates of evolution of genes and their associated phenotypes\(^2,4\). For example, one of the earliest reported TR-mediated phase variation (the reversible, random, high frequency gain or loss of a phenotype) was discovered in surface genes in *Neisseria gonorrhoeae*\(^5\). In this bacterium, members of the P.II gene family contain a variable CTCTT repeat in the region encoding the member signal peptide. Changes in the number of repeat units cause frameshift mutations and lead to either correctly translated proteins (phenotype ON) or not correctly translated proteins (phenotype OFF). This ON/OFF switching of phenotype occurs during infection and is believed to create variants in the bacterial population capable of surviving in the host.

1.1.2 Telomeric Repeats

A telomere is a region of repetitive DNA at the end of each chromatid. It functions to protect the end of the chromosome from deterioration or from fusion with neighbouring chromosomes. As chromatin structures, telomeres undergo epigenetic regulation of their maintenance and function. Their DNA component is usually formed by repetitive TTAGGG sequences (Fig 1.1) and protects the chromosome ends from being mistaken for unrepaired chromosome breaks. The major part of telomeres are folded into nucleosomes\(^6,7\) which are regularly spaced, but show a 30-40 bp shorter periodicity than the bulk chromatin\(^8\). This shorter periodicity is caused by reduced stoichiometric ratios between the linker histone and the histone octamer\(^9\) and because telomeric DNAs are straight (due to the TTAGGG tandem repeat) they are out of phase with respect to the B-DNA periodicity\(^10\). It has been shown that telomeric nucleosomes occupy
isoenergetic positions having the periodicity of the telomere repeat which creates thermodynamic stability and a lack of rotationally phased positioning\textsuperscript{11,12}. Chromatin structure participates in the establishment of a telomeric capping complex in the nucleosomal context. Telomeres have been shown to exert effects on other chromatin domains both in cis- and trans-positions.

Telomeres are shortened during each round of chromosome replication. The enzymes that duplicate DNA cannot continue their duplication all the way to the end of a chromosome because synthesis of Okazaki fragments requires RNA primers attaching ahead on the lagging DNA strand\textsuperscript{13}. The truncation of telomeres during each round of replication is balanced by telomerase which is a ribonucleoprotein that adds DNA sequence repeats to the 3’ end of the DNA strand in the telomere. If the telomerase activity is not sufficient, then the cells will gradually lose telomeric sequences as a result of incomplete replication\textsuperscript{14}. As telomeres shorten it can cause cell senescence through the p53 and pRb pathways\textsuperscript{15}. With critically shortened telomeres, further cell proliferation can be achieved by inactivation of p53 and pRb pathways leading to gross chromosomal rearrangements, genome instability, and sometimes cancer.

Mammalian telomeres are transcribed from subtelomeric regions toward telomeres by RNA polymerase II and these TERRA transcripts (telomeric repeat-containing RNA) associate with the chromosome termini\textsuperscript{16}. This identification of telomeric RNA revealed the transcription of telomeres, although earlier reports indicated this possibility\textsuperscript{17,18}. It is believed that TERRA has the ability to epigenetically influence telomeres and telomeric chromatin.

**1.1.3 Mono- and dinucleotide repeats**

The most common microsatellite in the human genome is a dinucleotide repeat of CA, which occurs tens to thousands of times across the genome. The length and sequence of microsatellites can be variable and may actually consist of a mononucleotide repeat. Many of these small repetitive units have been shown to be unstable and, as such, are classified as microsatellite instability\textsuperscript{19}. The underlying mechanism of microsatellite instability is a defect in the DNA mismatch repair pathway. The MMR pathway corrects replication errors, such as mispaired nucleotides, as well as small insertions and deletions resulting from slippage of the polymerase during replication of microsatellites. MMR deficiency, therefore, leads to an accumulation of mutations in microsatellites and it is recognized by decreased or (less often) increased microsatellite lengths. This phenomenon is referred to as microsatellite instability and was first
described in Lynch syndrome patients because it was detected in over 90% of tumours in those patients\textsuperscript{19–22}. The mechanism behind the instability of these microsatellites in Lynch syndrome patients is a defective MMR pathway. Microsatellite instability is also found in a large proportion (15-25%) of sporadic colorectal and endometrial carcinomas\textsuperscript{23}. In 1998, the National Cancer Institute convened to determine which mono- and dinucleotide repeats represented the best indicators of instability, and therefore MMR activity. Stringent criteria were used by the group to recommend a panel of five candidate microsatellites, chosen as markers to determine microsatellite instability and therefore MMR activity; two monocucleoties and three dinucleotides. These markers were then used to evaluate tumours and classify tumour types as having high microsatellite instability or low microsatellite instability. Differing treatment methods have been shown to be more or less responsive to each tumour type.

\textbf{1.1.4 Satellite I, II, III classical repeats}

Representing about 5% of the total human genome, classical satellite DNA repeats I, II and III are highly repetitive sequences isolated from bulk DNA by CsSO4 or CsCl gradients\textsuperscript{24} due to their characteristic G+C content (Fig. 1.1). The genic regions of human chromosomes are flanked by large heterochromatic regions, often in arrays of diverged monomeric classical satellites or other tandemly repeated satellite DNA, and centromeres. Each classical DNA satellite is mainly composed of a single family of repeated sequences. Satellite I is made of alternating arrays of A and T rich sequences with satellite II and III containing different degrees of degenerate ATTCC tracts. Satellite II and III have been found to localize within pericentromeric regions of many chromosomes as well as heterochromatic regions intrachromosomally\textsuperscript{25}. There is no sequence relationship evident between the repeat elements of satellite I and those of satellite II and III. Satellite I, when digested by the restriction endonuclease \textit{RsaI}, breaks into fragments less than 150 base-pairs in length\textsuperscript{24}. Satellite II, when digested with \textit{Hin}fI, breaks into small fragments between 10-80 base pairs and satellite III, when digested by the same enzyme, breaks into a degenerate repeat with a single G+C-rich region at the 3’ end. Satellite III is present in a much smaller amount in the genome than satellite II. Satellites II and III have been shown via recent chromatin immunoprecipitation and sequencing (ChIP-seq) experiments to be enriched for stable nucleosome modifications H2A/H4R3me2s, H4K20me3, and/or H3K9me3 which are consistent with heterochromatic chromatin\textsuperscript{26}. Biochemical findings show that H4R3me2s provides a binding site for the DNA
methyltransferase (Dnmt3a) that methylates satellite II and III. Epigenetic variations for satellite sequences in the genome may be caused by alterations of heterochromatin-associated proteins.

1.1.5 Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)

Transposable elements make up approximately 33% of the human genome (Fig. 1.1). Two classes of transposable elements exist, transposons and retrotransposons. Transposons have inverted terminal repeats, encoding a transposase activity and the ability to move from one site to another. Retrotransposons move through an RNA intermediate encoded by reverse transcriptase. In this way, a copy of the original can be integrated into a new genomic location. Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are included in the retrotransponson category (Fig. 1.1). LINEs can be up to several kilobases (kb) in length and structurally contain an internal promoter for RNA polymerase II, a 5’ untranslated region, two open reading frames, and a 3’ terminal polyadenylation site. The open reading frames produce a RNA binding protein, a reverse transcriptase and DNA endonuclease.

LINEs are autonomous elements with the property of self-sufficiency for mobility whereas SINEs are nonautonomous relying on activity from autonomous retrotransposons and/or the host for their mobility. SINEs also have an internal promoter for RNA polymerase III and a 3’A-rich tract. The most prominent SINEs are the human Alu elements. Expression of these elements has been shown to lead to genetic instability; therefore it is important that they remain transcriptionally silenced. DNA methylation has been found to be a mechanism for SINE and LINE silencing.
Figure 1.1. Repetitive elements in the human genome

Repetitive elements can be classified in a variety of ways depending on sequence, repeat length, repeat homogeneity, and genomic location. Specific variations of repetitive elements can be classified as highly repetitive (such as satellite DNA) or moderately repetitive (such as tandem repeats and interspersed repeats) meaning the repetitive element is not always a pure sequence repeat. Tandem repeats can further be classified into minisatellites, microsatellites, or megasatellites depending on the size of the repeated unit. Telomeric repeats are an example of minisatellites where dinucleotide repeats are an example of microsatellites. Interspersed repeats can be further classified as RNA transposons or DNA transposons depending on the way the elements is able to move to a new genomic location. Endogenous retroviruses (ERVs), LINEs and SINEs are examples of RNA transposons which use an RNA intermediate to acquire a new genomic location.
1.2 Epigenetic Marks

1.2.1 Histone modifications

Nucleosomes are histone octamers bound to 1.7 turns (146 base pairs) of wrapped supercoiled DNA. The formation or assembly of nucleosomes is strongly altered by nucleotide sequences of the DNA. Strong nucleosome assembly occurs at CTG repeats, GGA repeats, (A/T)/NN(G/C)NN repeats, TATAAACGCC repeats, and (TTGA)n repeats. Repeats that have been shown to exclude nucleosomes include CGG repeats, and poly(dA) repeats. Telomeres composed of TTAGGG repeats also exclude nucleosomes. Exclusion, however, only occurs in shortened telomeres less than 7 kb in length. At trinucleotide repeat disease loci, this altered nucleosome assembly is seen in vivo with increased protection of myotonic dystrophy type 1 (DM1) expanded CTG repeats from DNaseI via the loss of an adjacent wild type DNaseI hypersensitive site (Fig 1.2). Nucleosome unwinding and fragile site formation is hypothesized to be caused by exclusion of nucleosomes at CGG repeats in fragile X syndrome (FRAXA), fragile X E syndrome (FRAXE), fragile X site F syndrome (FRAXF), fragile site 16q22 (FRA16A) and fragile site 11q23.3 (FRA11B). The FRAXA locus specifically has methylation at the 5' end of the FMR1 gene and trinucleotide repeat (TNR) expansion leads to histone deacetylation and chromatin remodelling. The threshold of TNR instability may be connected to nucleosome formation due to similar DNA tract length sizes for the threshold at around 35 – 40 units (105-120bp) and the size of DNA wrapped around a single nucleosome (146 bp). The relationship between nucleosome positioning and repeat instability has yet to be elucidated.

Another important aspect of the chromatin environment of repeat loci is histone modifications on the nucleosomes. Transcription of genetic information encoded in DNA is in part regulated by chemical modifications of histone proteins. Histones associate with DNA to form nucleosomes which bundle to form chromatin fibers. The nucleosome core is formed by two H2A-H2B dimers and an H3-H4 tetramer. H1 and H5 are known as linker histones. Histones are subject to post-translational modifications primarily on H3 and H4 long N-terminal tails which protrude from the nucleosome and can be covalently modified in several places by methylation, citrullination, acetylation, phosphorylation, SUMOylation, ubiquitination, and
ADP-ribosylation. Within the nucleosome, histones undergo posttranslational modifications that alter their interaction with DNA and nuclear proteins. Modifications of histones are related to different types of chromatin. Constitutively repressed chromatin is marked by the histone modification H3 lysine 9 trimethyl (H3K9me3) whereas facultatively repressed chromatin is marked by H3K27me3. Specific to promoters, H3K4me3 signals activity just after transcription start sites and H3K9ac follows a similar pattern of marking actively transcribed promoters. Actively transcribed gene bodies are correlated with the histone modification H3K36me3 and enhancer elements are marked by H3K4me1. A modification separating active from poised enhancers, and is a predictor of developmental state, is H3K27ac. Expanded DNA repeats have been shown to influence chromatin structure and it is suggested that expanded CTG alleles show a trend toward compacted chromatin. Chromatin packaging has been shown to be affected by repeat expansion and could lead to misregulation of mRNA transcripts.

1.2.2 DNase I hypersensitivity

Regions of the genome are compacted by nucleosomes in varying degrees. Early studies on chromatin used the accessibility of DNA to nucleases as a measure of chromatin compaction. DNase I hypersensitive sites are thought to correspond to nucleosome free regions. While the possibility of an unusual nucleosome DNA structure giving rise to hypersensitivity still exist, in general, these sites correspond to nucleosome free regions. Early on, DNase I hypersensitive sites were shown to occur at functional regions of the genome such as promoters. In budding yeast, DNase I hypersensitive site mapping of a 45 kb region (including 30 genes) showed that the promoters of almost all genes, active replication origins, and silencers showed hypersensitivity to nuclease digestion. New techniques now exist to analyze DNase I hypersensitive sites genome-wide through ChIP experiments followed by microarrays covering the whole genome. With this analysis, the theory that regulatory regions of the genome have lower nucleosome occupancy was reinforced. DNase I hypersensitive sites give investigators an understanding of the basic chromatin structure in an area. Changes in hypersensitive sites can also lead to suggestions about functional changes in chromatin structure. For example, the loss of a DNase I hypersensitive site adjacent to an expanded DM1 CTG repeat is suggestive of a chromatin structure change (Fig. 2.1). Although the direct link between repetitive elements and DNase I hypersensitive sites has yet to be established, this epigenetic mark can play a key role in our overall understanding of the chromatin structure and structural changes that may occur at specific repeats.
**Figure 1.2. Epigenetic marks surrounding disease-causing repeat loci**

There are specific epigenetic modifications found around some of the 46 disease-causing repeat loci. This is an example of the DM1 locus which shows many of the *cis*-elements found at other disease-causing repeat loci. The DM1 locus has two origins of replication, one upstream and one downstream of the (CTG)n repeat tract. There are also two CTCF binding sites flanking the repeat tract which is found within a large 3.5 kb CpG island. There is a DNase hypsensitive site that is lost when the (CTG)n expands over a certain threshold. There are two transcripts for the gene *DMPK* both a sense and antisense transcript, the antisense transcript coincides with the promoter region of the downstream *SIX5* gene.
1.2.3 CTCF binding

CCCTC-binding factor (CTCF) is a highly conserved, ubiquitously expressed nuclear protein with 11 zinc finger domains through which the protein is able to bind a diverse set of sequences and partners. CTCF participates in a number of cellular processes including transcriptional activator/repressor, gene regulation including promoter activation and repression, gene silencing, methylation-dependent chromatin insulation and genomic imprinting\(^49\). CTCF has also been shown to play a role in repeat instability. It is thought to provide an insulating property and one of its many functions is to insulate promoters from respective enhancer elements. There are CTCF binding sites flanking many of the disease-causing repeats\(^50\) (Fig. 2.1). Previous studies have attempted to elucidate the role CTCF may be playing in instability. One study used a human fragment of the ATXN7 locus containing an expanded CAG repeat, and showed an increase in repeat instability when the flanking CTCF binding sites were methylated or mutated\(^51\). Depletion of CTCF binding in the 5'UTR of the Frataxin gene has been shown to be coincident with both a decrease in sense transcriptions (which is known to contribute to the disease) and increase in antisense transcription\(^52\). A relationship between CTCF and the DM1 locus was first established when CTCF binding sites were identified flanking the CTG/CAG repeat\(^53\). In congenital myotonic dystrophy type 1 patients, two CTCF binding sites were shown to be methylation-sensitive insulators which were disrupted in patient tissues\(^53\). The insulator function of CTCF provides boundaries for chromatin regions, preventing the spreading of heterochromatic marks past a certain point. The disruption of CTCF binding could, therefore, lead to aberrant chromatin markings. CTCF sites were also shown to be involved in the regulation of asynchronous replication at the H19/IGF2 locus\(^54\). Given that many of the 46 disease-causing repeats are flanked by CTCF binding sites, potential binding effects could be a widespread phenomenon among this class of disease.

1.2.4 DNA CpG sites and DNA CpG methylation

In humans, DNA methylation occurs primarily, but not exclusively, at cytosine-guanine dinucleotide sites (CpG sites)\(^55,56\) in a tissue- and development-specific manner\(^57\). The covalent addition of a methyl group at the 5-carbon of the cytosine ring results in 5-methylcytosine and
occurs almost exclusively at CpG sites in somatic cells. Embryonic stem cells, however, show 5-methylcytosine at CpH sites (H = A/C/T) and have also been seen in neural development. Between 60-90% of all CpGs are methylated in mammals and all vertebrates show DNA methylation. Over half of all genes have CpG islands at their promoters and when methylated, the promoter becomes silenced. DNA methylation may affect the transcription of genes in two ways. First, the methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene, and second, methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins. These proteins then recruit additional proteins to the locus, such as histone deacetylases and other chromatin remodeling proteins that can modify histones and can lead to the compaction of DNA into heterochromatin. This link between DNA methylation and chromatin structure is very important. In particular, loss of methyl-CpG-binding protein 2 (MeCP2) has been implicated in Rett syndrome and methyl-CpG-binding domain protein 2 (MBD2) mediates the transcriptional silencing of hypermethylatable genes in cancer.

DNA methylation is essential for normal development and is associated with genomic imprinting, X-chromosome inactivation, chromatin structure and carcinogenesis. DNA methylation may stably alter the expression of genes in cells as they divide and differentiate from embryonic stem cells into specific tissues. The resulting change is normally permanent and unidirectional, preventing a cell from reverting to a stem cell or converting into a different cell type. This methylation is typically removed during zygote formation and re-established through successive cell divisions during development. However, recent studies show that hydroxylation of methyl groups occurs rather than complete removal of methyl groups in the zygote.

DNA methylation can be a heritable epigenetic change in gene expression leading to genomic imprinting. Genomic imprinting is an epigenetic phenomenon by which certain genes can be expressed in a parent-of-origin-specific manner. It may also ensure transposable elements remain epigenetically silenced throughout gametogenic reprogramming to maintain genome integrity. Genomic imprinting is an epigenetic process that can involve DNA methylation and histone modulation in order to achieve monoallelic gene expression without altering the genetic sequence. The heritability of methylation states and the methylation gain or loss suggests that DNA methylation is adapted for a specific cellular memory function during development.
Methylation patterns are specific depending on tissue and cell type, developmental time periods, race, disease, aging, and environmental factors\textsuperscript{73}.

DNA methylation suppresses the expression of endogenous retroviral genes and other harmful stretches of DNA including suppression of repetitive elements.

1.3 mRNA sequencing

RNA sequencing (RNA-seq) is a technology that uses the capabilities of next-generation sequencing to reveal quantifiable levels of RNA present from a genome at a given moment in time. This provides information about RNA transcripts in the cell, alternative gene splicing events, post-transcriptional changes, gene fusions, mutations/SNPs and changes in gene expression. RNA-seq can also provide insight into pathway-based analyses and a correlation between epigenetic marks and the impact on transcription. Although many repetitive elements within the genome are not actively transcribed, a select few tandem repeats are. Some of these tandem repeats have been shown to be polymorphic in nature and cause diseases. For these repeat elements, there is evidence to suggest that the repeat itself may be bidirectionally transcribed in specific cases (Fig 2.1). Therefore, mRNA-seq is important in providing information about repetitive elements in the genome.

1.4 Disease-causing polymorphic repeats and instability

1.4.1 46 disease-causing microsatellite loci

The human genome contains a large number of repetitive elements, including micro-, mini-, and megasatellites. Instability within a specific set of repetitive elements is an important and unique form of mutation linked to over 40 neurological, neurodegenerative and neuromuscular disorders\textsuperscript{74}. There are 46 known diseases associated with unstable repeat tracts. The first repeat diseases, spinobulbar muscular atrophy (SBMA), fragile X syndrome (FRAXA) and muscular dystrophy type 1 (DM1), were reported in 1991. Expansions leading to disease have been found in a number of sequences, the most common being trinucleotide repeats such as (CTG) and (CGG) that are associated with diseases such as DM1, spinocerebellar ataxia (SCA) types 1, 2, 3, 6, 7, 8, 12 and 17, Huntington's disease (HD), FRAXA and many others\textsuperscript{75}. Larger repeat units such as tetraneucleotide expansions and pentanucleotide expansions lead to DM2 and SCA10, respectively. Disease-causing repeats can occur within genes such as in HD, promoters, 5'UTR
as in FRAXA, introns such as in DM2, and 3'UTR such as in DM1\textsuperscript{75}. Depending on where in the gene the expansion occurs, it can lead to a number of pathogenic outcomes. Misregulation of gene expression, splicing defects, protein loss of function, toxic polyglutamine-containing proteins, and gain of function mRNA are a few examples\textsuperscript{75,76}. For most disease-causing repeats, unaffected individuals have less than 35 repetitive units and the repeats are stable throughout the lifetime and across generations. Only after a threshold number, do repeats become unstable; meaning the number of repeats expands or contracts within an individual and during transmission. The patterns of repeat instability differ between disease loci. The same repeat sequence can show different patterns of instability, for example, the CTG/CAG repeat of DM1 can expand to over 1000 repetitive units\textsuperscript{77} where the same CTG/CAG repeat unit of HD expands to only 100 repetitive units or less\textsuperscript{78}. Repeat length between diseases does not correlate to disease severity; therefore some other factor must be affecting disease severity. These dynamic mutations can expand further upon transmission and this 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. 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.

1.4.2 Mechanisms of instability

Many factors have been shown to play a part in repeat instability. One of these factors is repeat tract purity. Repeat tracts containing interruptions have been shown to have increased stability against expansions or contractions relative to pure tracts in both yeast and human cell models\textsuperscript{79,80}. Most times interruptions are associated with stability, but not always, and the mechanism behind interruptions increasing or decreasing instability is unknown.

Another factor influencing repeat instability is DNA replication. Repeat instability occurs in proliferative and non-proliferative tissues. However, DNA replication likely does play some part in instability (Fig 2.1). The Pearson lab has shown that replication direction affects the instability of trinucleotide repeats\textsuperscript{81} and other labs have shown that proteins such as flap endonuclease 1 (FEN1) and Werner syndrome ATP-dependent helicase (WRN) can have an effect on instability and are all involved in replication\textsuperscript{82–84}. FEN1 removes 5’ overhanging flaps
in DNA repair and processes the 5’ ends of Okazaki fragments in lagging strand DNA synthesis\textsuperscript{85}. WRN is a member of the RecQ helicase family with a 3’ to 5’ exonuclease activity and acts to unwind and separate double-stranded DNA\textsuperscript{86}. Replication origins may have an effect on instability and are often found around disease-causing repeats for example HD and SCA7 have replication origins downstream of the repeats, FRAXA has a replication origin in the promoter and SBMS has two origins flanking the repeat\textsuperscript{87–89}. In the Pearson lab it has been shown that there are two origins at the DM1 locus flanking the repeat\textsuperscript{90}.

Transcription has also been shown to be a factor influencing instability. All disease-causing repeats are transcribed and some are bidirectionally transcribed (Fig. 2.1). Activity of replication and transcription together probably plays a role in instability. It has been shown \textit{in vitro} that collision of replication and transcription machinery caused repeat instability\textsuperscript{91}. Terminally differentiated cells also show instability and transcription is likely a major contributor in those cells. It has been shown that bidirectional transcription results in 20-fold more CTG/CAG repeat instability when compared to single-direction transcription and leads to the production of R-loops\textsuperscript{92,93}. The Pearson lab has shown that R-loops, RNA-DNA hybrid structures, may additionally contribute to repeat instability\textsuperscript{93}. This shows that DNA structure may also influence repeat instability. Repeat containing tracts of DNA can create aberrant DNA structures that when coupled with transcription, lead to RNAPolII arrest and multiple attempted rounds of transcription-coupled repair\textsuperscript{94}. The repair process as well has been shown to lead to instability. Mismatch repair (MMR) and DNA binding proteins are \textit{trans}-acting factors that may contribute to repeat instability in tissues with little proliferation. MMR proteins show a strong effect on instability, for example MSH2 (binds to trinucleotide repeat DNA structures) and MSH3 are required for instability in some cases\textsuperscript{50,95,96}, but mostly shield from mutation. Instability is likely caused by error-prone or escaped repair of DNA structures\textsuperscript{97}.

CCCTC-binding factor (CTCF) has also been shown to play a role in repeat instability. This factor is a highly conserved, ubiquitously expressed nuclear protein with 11 zinc finger domains and participates in a number of cellular processes including transcriptional activation, repression, gene regulation, gene silencing, methylation-dependent chromatin insulation and genomic imprinting\textsuperscript{49}. There are CTCF binding sites flanking many of the disease-causing repeats\textsuperscript{50} (Fig. 2.1). Aberrant CTCF binding has been shown to have an effect on the repeat stability at the ATXN7 locus, DM1 locus, CDM locus, and H19/IGF2 locus\textsuperscript{51,53,54}. For example,
a depletion of CTCF and coincident heterochromatin formation involving the +1 nucleosome enrichment of H3K9me3 and recruitment of heterochromatin protein 1 was found in the 5′ UTR of the FXN gene in Friedreich ataxia (FRDA)\textsuperscript{52}. The expanded GAA triplet repeat results in deficiency of FXN gene transcription, which is reversed via administration of histone deacetylase inhibitors. This indicates that transcriptional silencing is at least partially due to an epigenetic abnormality. This abnormality was identified as FAST-1 (FXNAntisense Transcript – 1) which is a novel antisense transcript that overlaps the CTCF binding site and is expressed at higher levels in FRDA. The author concluded that CTCF depletion constituted an epigenetic switch that results in increased antisense transcription, heterochromatin formation and transcriptional deficiency in FRDA\textsuperscript{52}.

In humans, DNA methylation occurs primarily, but not exclusively, at CpG sites\textsuperscript{55,56} in a tissue- and development-specific manner\textsuperscript{57}. Although there is no general association amongst all repetitive elements and methylatable CpG sites, trinucleotide repeats that show the greatest instability are flanked by, or embedded in, CpG islands\textsuperscript{98,99} (Fig. 2.1). Furthermore, in vitro methylation of repetitive elements, including mono-, di-, tri-, penta-, and minisatellite sequences reduces instability for the majority of repeats, with a minority being destabilized or unaffected\textsuperscript{100}. The methylation status of flanking sequences, in addition to chromosomal context, can determine the differential susceptibility to instability for various repeat loci\textsuperscript{23,101,102}. In tumors, alterations of normal DNA methylation are associated with microsatellite instability and general genomic instability\textsuperscript{103}. It has been shown in Fragile X syndrome that large expansions become fully methylated resulting in relatively stable repeats after expansion into the diseased range\textsuperscript{104}. The disease FRDA has also shown aberrant hypermethylation upstream of the GAA repeat when compared to control\textsuperscript{105}. The DM1 repeat tract does not contain CpG methylatable sites\textsuperscript{106}, although recent results suggest non CpG-methylation may occur on CTG/CAG repeat tracts during early development\textsuperscript{60}. The DM1 repeat tract is, however, surrounded by a large CpG island, which becomes hypermethylated in the congenital form of the disease\textsuperscript{53}, and in a tissue-specific manner upstream of the repeat in adult and transgenic DM1 mouse tissues\textsuperscript{77}. The latter results were generated from methylation-restriction endonuclease studies, suggesting that more detailed investigation may reveal additional methylation changes and a connection with instability. Recently reported CCG sequence interruptions for DM1\textsuperscript{107} may introduce methylatable sites into the repeat tract. DNA methylation provides an insight into chromatin
structure and accessibility. This cis-element provides valuable information about interactions with trans-factors at or around repetitive sequences.

1.4.3 Tissue specific instability

Disease-causing repeats show a unique phenomenon of tissue-specific instability whereby within an individual, different tissue types are more prone to further expansions or contractions than others. This shows that mutations are dynamic within an individual. Repeat expansion diseases vary in the amount of somatic instability seen in different tissues. Possible causes for this variability include cell type-specific trans-acting 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 instability very complicated. Depending on the tissue, somatic instability may occur early in embryogenesis or through the patient's life and can occur to the same degree such that all the repeats expand to a similar size or to varying sizes leading to extensive length heterogeneity within the same tissue.

DM1 shows a developmental window for instability during the 13-16 week fetal stage, corresponding to the transition from differentiation in the first trimester of fetal development when the repeats are stable, to the rapid fetal growth in the second trimester when somatic instability occurs. DM1 muscle repeats are heterogeneous and expand over time, but plateau in adulthood. In contrast, DM1 blood has little variability within each patient, but expands synchronously through the entire life of the patient. 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. Lines of transgenic mice carrying either a SCA7 cDNA construct or a 13.5 kb SCA7 genomic fragment with 92 repeats have been created with the expanded repeat containing line reconstituting the diseased phenotype in humans. In these mice, somatic instability is seen in the liver and brain, and at very low levels in most other tissues. Somatic instability in SCA10 is found in blood leukocytes, lymphoblastoid cells and buccal cells but has not been assayed in any organs. The same tissue in two different diseases may show different levels of instability. Therefore, the variability in repeat instability likely stems from either specific molecular pathways affected by mutation and/or cis-elements near the expansion. Whether somatic instability contributes to patient symptoms and the progressive nature of the disease 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. Most arguments for a contribution of repeat instability to disease severity 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 the HD and SCA1 brain. Patients with these diseases share similar trinucleotide instability patterns including in different brain regions, yet SCA1 has affected Purkinje cells of the cerebellum and HD has degeneration of spiny neurons of the striatum. The role of somatic instability in disease progression may turn out to be disease specific, but for now remains to be defined.

1.5 Thesis goals

Dramatic repeat length variation evident at different disease-causing repeat loci can not be easily explained by repeat expansion alone. This raises the possibility that cis-elements and trans-acting factors may contribute to instability. Although the existence of cis-elements that alter instability have been widely-accepted, a non-diseased contextual reference using human tissues has not been established. With the advancement in research into cis-elements involved in repeat instability, a systematic investigation into disease-causing repeat loci within control human tissues would provide a baseline platform for further investigation. My thesis goals are to evaluate the epigenetic context at disease-causing repeat loci in multiple control tissue types. I also aim to evaluate the epigenetic context at other repetitive DNA elements within the genome.
2 The epigenetic landscape of disease-causing polymorphic repeat loci

2.1 Abstract

Repetitive sequences may comprise up to two thirds of the human genome. The dramatic variations in repeat sequences, biological functions, and phenotypic effects correlate to how diverse these repetitive units are in function and location. Some of the variation of repeats has begun to be elucidated and it is clear that chromatin structure plays a role in repeat maintenance, transposition, and disease-potential. Although a clear link between chromatin environment and repeat dynamics has yet to be established, a correlation has been shown time-and-again between epigenetic context and repeat maintenance and function. Here we take a broad perspective and evaluate different forms of repeats within the human genome for enrichment of specific epigenetic marks in human control hippocampus, skeletal muscle, liver and kidney. Publically available data from the ENCODE consortium was evaluated for histone modifications; histone H3 lysine 9 acetylation (H3K9ac), H3K4me1, H3K36me3, H3K4me3, H3K27ac, H3K27me3, and H3K9me3, CpG sites, DNA methylation, CTCF binding sites, DNase I hypersensitivity sites and mRNA-seq data for enrichment. The types of repeats that we focused our survey on were 46 known disease-causing microsatellite loci as well as 824 sequence and length matched microsatellite loci, 36 telomeric repeat loci located at telomeres or intrachromosomally, 22 mono- and dinucleotide polymorphic repeat loci, 8 short interspersed nuclear element (SINE) loci, 8 long interspersed nuclear element (LINE) loci, and 1236 loci for classical DNA satellites I, II, and III repeats. Analysis of sequenced regions by ChIP-seq processing using the Cistrome Galaxy platform and statistical analysis using SeqMINER revealed that several sequence variations actually show complex domain and/or higher order repeat organization.

Microsatellites causing 46 known neurodegenerative diseases showed a distinct epigenetic profile that was not recapitulated with any other repeat variety. Telomeric repeat loci, mono- and dinucleotide repeat loci, and SINE and LINE loci showed little enrichment of active histone modifications. Classical satellite I, II, and III repeat loci showed high levels of enrichment for active marks H3K4me3 and H3K27ac. I found that there are specific disease-causing repeat loci that show significant enrichment of certain histone modifications when compared to their
sequence, length and genetic location controls. I also found that some of these significantly enriched repeat loci show different enrichment levels within different tissues. The 46 disease-causing microsatellite repeat loci along with 824 sequences and location matched microsatellite loci were the only variety of repeats to show a possible nucleosome positioning effect surrounding the repeat unit. Further analysis of this positioning effect revealed it to be sequence specific with significantly more CGG and less CAG repeats in clusters showing a potential nucleosome positioning effect and less CGG and more CAG repeats in clusters without this epigenetic profile. My work shows that individual loci show tissue-specific effects, but no general tissue-specific effects could be seen when observing all loci for each repetitive element together.

2.2 Introduction

Repetitive sequences comprise approximately 46% of the human genome. The dramatic variation in repeat sequences, biological functions, and phenotypic effects correlates to how diverse these repetitive units are in function and location. Some of the variation of repeats has begun to be elucidated and it is clear that chromatin structure plays a role in repeat maintenance, transposition, and disease-potential. The exact interplay between the epigenetic environment and repetitive DNA has yet to be established, yet newly developed genome-wide analysis are beginning to provide insights into this complex relationship.

Repetitive sequences come in many forms and locations throughout the genome. Centromeric repeats constitute the largest tandem DNA family in the human genome called alpha satellite DNA, which has been extensively studied and has emerged as a paradigm for understanding the genomic organization of tandem DNA. Microsatellites and variable number of tandem repeats can be highly polymorphic, located in gene exons, introns, untranslated regions, and in non-genic areas, and are important for use as genetic marks. The evolutionary conservation of tandem repeats suggests a link to phenotypic advantage, which has been shown in specific cases. The chromatin context at specific repeats can change this phenotypic advantage, therefore indicating the importance of epigenetics when investigating repeat elements. For mono- and dinucleotide repeats, this polymorphic feature can be used as an indicator of general genomic instability. Another repetitive unit in the genome is telomeric repeats which can be found mainly at the ends of chromosomes but also rare instances of relatively short tracts are present at
intrachromosomal locations. SINEs and LINEs constitute transposable non-tandem interspersed repetitive units within the genome. Chromatin modifications and epigenetic context play a role in the maintenance and integrity of this variety of repeat. There are many varieties of repeats within the genome, all with specific characteristics, and all with a relationship to the chromatin environment. To better understand the majority of the human genome, it is important to investigate the epigenetic context (histone modifications, DNA methylation, protein binding, etc.) at a variety of repetitive units.

The development of genome-wide arrays for chromatin analysis has provided an efficient platform to begin exploring the epigenetic environment. Specifically, the ChIP-seq technique provides information on histone modifications and DNA/protein interactions. We can also infer indirect DNA/DNA interactions by information gathered about histone modifications and protein/DNA interactions. The combination of certain histone modifications and protein binding provides information about the probable state of chromatin at specific locations. The correlation between certain histone modifications and chromatin activity has been established and can be relied upon and an indicator of chromatin context. Active histone modifications include H3K9ac, a mark for actively transcribed promoters, H3K27ac, a mark for active and poised enhancers, H3K4me1, a mark of enhancers, H3K4me3, a mark for activity just after transcription start sites, and H3K36me3, a mark of actively transcribed gene bodies. Repressive histone modifications include H3K9me3, a mark of constitutively repressed genes and H3K27me3, a mark of facultatively repressed genes. Another indicator of chromatin environment is DNase I hypersensitivity. An active and open region will show sensitivity to the DNase I enzyme where a closed and compacted region will not show this sensitivity. The chromatin environment is also affected by protein binding. One of the key chromatin regulatory proteins is CTCF. The binding of this protein has been linked to transcriptional activation/repression, gene regulation, methylation-dependent chromatin insulation and genomic imprinting. DNA (cytosine) methylation is also a characteristic of compacted chromatin and the assessment of CpG methylation sites provides insight into the chromatin structure. Large CpG islands have been found around many disease-causing trinucleotide repeats. Although the direct relationship between repetitive sequences and G+C content is unknown, many repeats are either composed of cytosines and guanines or surrounded by areas of high G+C content. It has been suggested that this may be an evolutionarily conserved mechanism to protect repetitive elements through DNA
There are many factors that influence chromatin environment and can provide insight into possible mechanisms of repeat maintenance. The investigation into the chromatin context at repetitive units within the genome is crucial to our understanding of the maintenance and disease-causing potential of repeats, as well as to possibly provide insight into the natural functional role of these repeats. To determine the epigenetic status at specific repeat types within control human tissues, we have queried publically available ChIP-seq data and other epigenetic marks of interest.

2.3 Methods

*Locating Unstable Repeats and Tandem Repeat Controls*

I determined the location of 46 known tandem repeats with 45 disease phenotypes, relative to human genome build 19 (hg19; UCSC) and the 2009 human reference sequence (GRCh37; NCBI) (Table 2.1). The position of each disease gene was obtained through the UCSC genome browser and when multiple variants were encountered, the minimum and maximum positions were used. A region including the repeat and 5 kb upstream and 5 kb downstream was selected for each repeat tract. Although there may be interactions that reach farther than 5 kb, a window too broad would minimize repeat-centred distribution patterns. A manual search was carried out to locate each repeat and confirm its position in an exon, intron, or UTR region. To find sequence and location matched tandem repeats that are not associated with disease, the RepeatFinder program was used. Each sequence that causes disease was queried and included only if found in a genomic region which matched the disease repeat. Sequence proportions were kept the same between the 46 disease-causing population and the control population (Table 2.2). Also, sequence lengths were kept the same between the disease-causing population and the control population. Therefore, the control population was made of the same kinds of repeat sequences as the disease-causing population, with the same proportion breakdown of repeat type, the same length of repeats, and similar genomic positioning of repeats (intronic, exonic, 3’UTR, etc).
Table 2.1. 46 disease-causing repeat loci locations

There are 46 known disease-causing repeat loci. The repeat sequences, length and location may be similar between some loci. Most of these disease-causing loci show instability after the repeat has expanded past a certain threshold. This threshold is different for each loci.

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Gene</th>
<th>Location</th>
<th>Repeat</th>
<th>Disease length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding Repeats – Polyglutamine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBMA: Spinal and Bulbar Muscular Atrophy</td>
<td>AR</td>
<td>ChrX: 66765160 - 66765228</td>
<td>23 (CAG)*(CTG)</td>
<td>40 - 55</td>
</tr>
<tr>
<td>HD: Huntington’s disease</td>
<td>HTT</td>
<td>Chr4: 3076604 - 3076660</td>
<td>19 (CAG)*(CTG)</td>
<td>&gt; 35</td>
</tr>
<tr>
<td>DRPLA: Dentatorubral-pallidouysian atrophy</td>
<td>ATN1</td>
<td>Chr12: 7045942 - 7045936</td>
<td>15 (CAG)*(CTG)</td>
<td>49 – 88</td>
</tr>
<tr>
<td>SCA1: Spinocerebellar ataxia 1</td>
<td>ATXN1</td>
<td>Chr6: 16327867 - 16327952</td>
<td>29 (CAG)*(CTG)</td>
<td>39 – 81</td>
</tr>
<tr>
<td>SCA2: Spinocerebellar ataxia 2</td>
<td>ATXN2</td>
<td>Chr12: 112036755 - 112038905</td>
<td>23 (CAG)*(CTG)</td>
<td>&gt; 34</td>
</tr>
<tr>
<td>SCA3: Spinocerebellar ataxia 3</td>
<td>ATXN3</td>
<td>Chr14: 92537355 - 92537384</td>
<td>10 (CAG)*(CTG)</td>
<td>&gt; 55</td>
</tr>
<tr>
<td>SCA6: Spinocerebellar ataxia 6</td>
<td>CACN1A</td>
<td>Chr19: 13318683 - 13318712</td>
<td>13 (CAG)*(CTG)</td>
<td>20 – 29</td>
</tr>
<tr>
<td>SCA7: Spinocerebellar ataxia 7</td>
<td>ATXN7</td>
<td>Chr3: 63898342 - 63898391</td>
<td>10 (CAG)*(CTG)</td>
<td>37 – 306</td>
</tr>
<tr>
<td>SCA17: Spinocerebellar ataxia 17</td>
<td>TBP</td>
<td>Chr6: 170870996 - 170871110</td>
<td>38 (CAG)*(CTG)</td>
<td>47 – 63</td>
</tr>
<tr>
<td>KCNN3: no disease association</td>
<td>KCNN3</td>
<td>Chr1: 154842201 - 154842167</td>
<td>14 (CAG)*(CTG)</td>
<td>-</td>
</tr>
<tr>
<td>AIB-I: Increased prostate cancer risk</td>
<td>NCOA3</td>
<td>Chr20: 46279816 - 46279902</td>
<td>29 (CAG)*(CTG)</td>
<td>&lt; 29/&lt; 29</td>
</tr>
<tr>
<td>SEF2-1b: no confirmed disease association</td>
<td>SEF2-1b</td>
<td>Chr18: 53253387 - 53253458</td>
<td>24 (CAG)*(CTG)</td>
<td>-</td>
</tr>
<tr>
<td>ERDA1: no confirmed disease association</td>
<td>ERDA1/DIR1</td>
<td>Chr17: 49909030 - 49909092</td>
<td>21 (CAG)*(CTG)</td>
<td>-</td>
</tr>
<tr>
<td>Coding Repeats – Polyalanine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Clinical Syndrome</td>
<td>Chromosome Location</td>
<td>Repeat Type</td>
<td>Repeat Sequence</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>HOXD13</td>
<td>Synpolydactyly</td>
<td>Chr2: 176957893 - 176957827</td>
<td>(GCG)* (GCG)</td>
<td>5</td>
</tr>
<tr>
<td>OPMD</td>
<td>Oculopharyngeal Muscular Dystrophy</td>
<td>Chr14: 23790682 - 23790711</td>
<td>(GCG)* (CGC)</td>
<td>10</td>
</tr>
<tr>
<td>CBF1A1</td>
<td>Cleidocranial dysplasia</td>
<td>Chr6: 45390488 - 45390529</td>
<td>(GCG)* (CGC)</td>
<td>14</td>
</tr>
<tr>
<td>ZIC2</td>
<td>Holoprosencephaly</td>
<td>Chr13: 100634394 - 100634408</td>
<td>(GCG)* (CGC)</td>
<td>5</td>
</tr>
<tr>
<td>HOXA13</td>
<td>Hand-Foot-Genital Syndrome</td>
<td>Chr7: 27239298 - 27239586</td>
<td>(GCG)* (CGC)</td>
<td>18, 12, 14</td>
</tr>
<tr>
<td>FOXL2</td>
<td>Blepharophimosis/Ptosis/Epicanthus inversus syndrome type II</td>
<td>Chr3: 138664863 - 138664904</td>
<td>(GCG)* (CGC)</td>
<td>14</td>
</tr>
<tr>
<td>ARX</td>
<td>Infantile Spasm Syndrome</td>
<td>ChrX: 25031767 - 25031814</td>
<td>(GCG)* (CGC)</td>
<td>16</td>
</tr>
<tr>
<td>COMP</td>
<td>Multiple Skeletal dysplasias</td>
<td>Chr19: 18896845 - 18896860</td>
<td>(GAC)* (GTC)</td>
<td>5</td>
</tr>
<tr>
<td>C9orf72</td>
<td>no confirmed disease association</td>
<td>Chr9: 27573523 - 27573549</td>
<td>(GGGGGCCC)* (GGCCCC)</td>
<td>3</td>
</tr>
<tr>
<td>SCA36</td>
<td>Spinocerebellar ataxia 36</td>
<td>Chr20: 2633380 - 2633409</td>
<td>(GGCCTG)* (CAGGCC)</td>
<td>5</td>
</tr>
<tr>
<td>TCF4</td>
<td>no confirmed disease association</td>
<td>Chr18: 53253387 - 53253458</td>
<td>(CTG)* (CAG)</td>
<td>24</td>
</tr>
<tr>
<td>DMPK</td>
<td>Myotonic Dystrophy Type 1</td>
<td>Chr19: 46273463 - 46273522</td>
<td>(CTG)* (CAG)</td>
<td>21</td>
</tr>
<tr>
<td>FXN</td>
<td>Friedreich’s ataxia</td>
<td>Chr9: 71652203 - 71652220</td>
<td>(GAA)* (TTC)</td>
<td>6</td>
</tr>
<tr>
<td>ATXN8/ATXN8OS</td>
<td>Spinocerebellar ataxia 8</td>
<td>Chr13: 70713516 - 70713548</td>
<td>(CTG)* (CAG)</td>
<td>15</td>
</tr>
<tr>
<td>PPP2R2B</td>
<td>Spinocerebellar ataxia 12</td>
<td>Chr5: 146258292 - 146258321</td>
<td>(CAG)* (CTG)</td>
<td>10</td>
</tr>
<tr>
<td>TK2/BEAN</td>
<td>Spinocerebellar ataxia 31</td>
<td>Chr16: 66524299 - 66524300</td>
<td>(TGGAA)* (TTCCA)</td>
<td>0</td>
</tr>
<tr>
<td>JPH3</td>
<td>Huntington’s disease-like 2</td>
<td>Chr16: 87637890 - 87637935</td>
<td>(CAG)* (CTG)</td>
<td>15</td>
</tr>
<tr>
<td>MAB21L1</td>
<td>no confirmed disease association</td>
<td>Chr13: 36050434 - 36050490</td>
<td>(CAG)* (CTG)</td>
<td>19</td>
</tr>
<tr>
<td>ZNF9</td>
<td>Myotonic Dystrophy Type 2</td>
<td>Chr3: 128891420 - 128891499</td>
<td>(CCTG)* (CAGG)</td>
<td>22</td>
</tr>
<tr>
<td>ATXN1</td>
<td>Spinocerebellar</td>
<td>Chr22: 46191235 -</td>
<td>(AATCT)*</td>
<td>14</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Genes</td>
<td>Location</td>
<td>Repeat</td>
<td>Risk Factor</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>----------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>ataxia 10</td>
<td>EMP1</td>
<td>CSTB Chr21: 45196323 - 45196361</td>
<td>(AGATT)</td>
<td>10 ((CCCCGCCC CGCG))</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>IRDN</td>
<td>INS Chr11: 2182801 - 2183205</td>
<td>(ACAGGGGT (G/C)(T/C)GG G)</td>
<td>50 ((ACAGGGGT (G/C)(T/C)GG G))</td>
</tr>
<tr>
<td>Progressive Myoclonic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSHD: Facioscapulohumeral muscular dystrophy</td>
<td>FSHD1A</td>
<td>FSHMD Chr4: 190986409 - 191014477</td>
<td>(CGG)</td>
<td>11 D4Z4 (3.3 kb)</td>
</tr>
<tr>
<td>SMYD3: Cancer risk factor</td>
<td>SMYD3</td>
<td>SMYD3 Chr1: 246670629 - 246670639</td>
<td>(CCGCC)* (GGCGG)</td>
<td>3 ((CCGCC)* (GGCGG))</td>
</tr>
<tr>
<td>RELN: Risk of Autism</td>
<td>RELN</td>
<td>RELN Chr7: 103629805 - 103629829</td>
<td>(CGG)*(GCC)</td>
<td>8 ((CGG)*(GCC))</td>
</tr>
</tbody>
</table>

**Fragile Site-Associated Repeats – Folate-Sensitive Sites**

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genes</th>
<th>Location</th>
<th>Repeat</th>
<th>Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRAXA: Fragile X syndrome</td>
<td>FMR1/FMR4</td>
<td>ChrX: 146993569 - 146993652</td>
<td>(CGG)*(CCG)</td>
<td>20 ((CGG)*(CCG))</td>
</tr>
<tr>
<td>FRAXE: Fragile X syndrome</td>
<td>FMR2</td>
<td>ChrX: 147582234 - 147582128</td>
<td>(CCG)*(CCG)</td>
<td>15 ((CCG)*(CCG))</td>
</tr>
<tr>
<td>FRAXF: No confirmed disease association</td>
<td>FAM11A</td>
<td>ChrX: 148713420 - 148713437</td>
<td>(CGG)*(CCG)</td>
<td>6 ((CGG)*(CCG))</td>
</tr>
<tr>
<td>FRA10A: no confirmed disease association</td>
<td>FRA10AC1</td>
<td>Chr10: 95462355 - 95462228</td>
<td>(CCG)*(CCG)</td>
<td>8 ((CCG)*(CCG))</td>
</tr>
<tr>
<td>FRA11B: Jacobsen syndrome</td>
<td>CBL2</td>
<td>Chr11: 119077000 - 119077032</td>
<td>(CCG)*(CCG)</td>
<td>11 ((CCG)*(CCG))</td>
</tr>
</tbody>
</table>

**Fragile Site-Associated Repeats – Distamycin A – Inducible**

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genes</th>
<th>Location</th>
<th>Repeat</th>
<th>Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRA16B: no confirmed disease association</td>
<td>-</td>
<td>Chr16: 65357265 - 65357282</td>
<td>(33-bp AT-rich) VNTR</td>
<td>9 ((33-bp AT-rich) VNTR)</td>
</tr>
</tbody>
</table>

**Fragile Site-Associated Repeats – BrdU Inducible**

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genes</th>
<th>Location</th>
<th>Repeat</th>
<th>Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRA10B: no confirmed disease association</td>
<td>-</td>
<td>Chr10: 111823497 - 111823540</td>
<td>(~42-bp AT-rich) VNTR</td>
<td>22 ((~42-bp AT-rich) VNTR)</td>
</tr>
</tbody>
</table>
Table 2.2. Proportion of repeat sequences in 46 disease-causing repeat loci population and control population

Of the 46 disease-causing repeat loci, there are only 18 DNA sequences. The proportion of the total 46 that each sequence makes up is recapitulated in the 866 loci for repeats not known to cause disease. Some repetitive units did not have many none-disease loci within the genome to include in the none-disease loci population.

<table>
<thead>
<tr>
<th>Repeat Sequence</th>
<th>Number of disease-causing loci</th>
<th>Percent of 46 disease-causing repeat loci</th>
<th>Number of control loci</th>
<th>Percent of 866 control loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAGGGGTGTGGGG</td>
<td>1</td>
<td>2.13</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>AT</td>
<td>2</td>
<td>4.26</td>
<td>43</td>
<td>4.93</td>
</tr>
<tr>
<td>ATTCT</td>
<td>1</td>
<td>2.13</td>
<td>11</td>
<td>1.26</td>
</tr>
<tr>
<td>CAG</td>
<td>16</td>
<td>34.04</td>
<td>314</td>
<td>36</td>
</tr>
<tr>
<td>CCCCCCCCCGCG</td>
<td>1</td>
<td>2.13</td>
<td>4</td>
<td>0.46</td>
</tr>
<tr>
<td>CCG/GCC</td>
<td>2</td>
<td>4.26</td>
<td>40</td>
<td>4.59</td>
</tr>
<tr>
<td>CCGCC</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>CCTCATGGGTGGTGGCTG</td>
<td>1</td>
<td>2.13</td>
<td>2</td>
<td>0.23</td>
</tr>
<tr>
<td>CCTTG</td>
<td>1</td>
<td>2.13</td>
<td>9</td>
<td>1.03</td>
</tr>
<tr>
<td>CGG/GCG</td>
<td>12</td>
<td>25.53</td>
<td>255</td>
<td>29.24</td>
</tr>
<tr>
<td>CTG</td>
<td>2</td>
<td>4.26</td>
<td>39</td>
<td>4.47</td>
</tr>
<tr>
<td>D4Z4</td>
<td>1</td>
<td>2.13</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>GAA</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>GAC</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>GGCCTG</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>GGGGCC</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>TGGAA</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>ACGGCCC</td>
<td>1</td>
<td>2.13</td>
<td>21</td>
<td>2.41</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>866</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Locating Telomeric Repeats (intrachromosomal and telomeric), SINEs, LINEs, Satellites I, II and III**

I determined the location of 36 telomeric repeats intrachromosomally and at telomeres by querying the sequence TTAGGG in RepeatFinder for 5 or more repetitive units and confirming the genomic location also through RepeatFinder. The location of 8 SINEs and 8 LINEs were determined using the UCSC genome browser RepeatMasker annotation track randomly. A region including the repeat and 5 kb upstream and 5 kb downstream was selected for each repeat tract.

**Locating Mono- and Dinucleotide Repeats**

I determined the location of 22 mono- and dinucleotide loci by using previously published data and validating through the UCSC genome browser. The National Cancer Institute held a workshop in 1998 which outlined potential microsatellite instability markers. The five mono- and dinucleotide repeat loci chosen to be part of the microsatellite instability panel are included, as are the other candidates (Table 2.3). A region including the repeat and 5 kb upstream and 5 kb downstream was selected for each repeat tract.

**Relevant DNA and Chromatin State Markers**

Genomic context factors were selected because of their contribution to chromatin structure, their links to epigenetics and regulation or their general roles in chromatin maintenance. I used 24 genome-wide assays from the ENCODE project data repository, covering available DNA and chromatin state markers for human hippocampus, skeletal muscle, liver, and kidney tissues (Table 2.4). These tissues were derived from different individuals of different ages and genders and therefore may have variable sequence lengths at the same loci. The choice of these tissues was based on the availability of relevant markers and the implications to repeat disease relevant phenotypes. Analyses with these tissue types present repeats in their non-diseased and stable states.

I used several markers for euchromatin, heterochromatin and nucleosome occupancy, including DNaseI hypersensitivity assays. I used a large number of ChIP-Seq data sets including H3K27me3, H3K9me3, H3K4me1, H3K4me3, H3K27ac, H3K9ac, H3K36me3, and CTCF to probe histone modifications and protein binding. In addition, I analyzed datasets on DNA
methylation assays\textsuperscript{122} to examine methyl modifications at or near relevant locations. Furthermore, I analyzed RNA transcription by use of a mRNA-seq dataset\textsuperscript{122} to address the activity at relevant locations and directionality of transcription.

A range of markers for genomic context were selected for analysis and all the associated files were downloaded from the ENCODE repository. When multiple assays existed (different groups had performed the same assays), they were downloaded for analysis. Similarly, when multiple replicates were performed, they were all downloaded for analysis.

**Statistical Enrichment Data**

For each marker and each repeat, the centre of the region was identified and used for distance calculations. The program SeqMINER was used to create clusters, using K-means, and visualize enrichment through heatmaps including 5 kb upstream and 5 kb downstream of the repeat tract. K-means clustering is a method of vector quantization which aims to partition \( n \) observations in \( k \) clusters so that each observation belongs to the cluster with the nearest mean enrichment pattern\textsuperscript{123}. Clusters were analyzed for repeat sequence make-up, repeat type (disease-causing, non-disease-causing, mono- and dinucleotide, telomeric, SINE, LINE, satellites I, II, or III) by using Fisher’s exact test and 2 proportion z-score statistical tests in Rstudio.
Table 2.3. Mono- and Dinucleotide repeat loci

Some mono- and dinucleotide repeats show microsatellite instability. The sequence, location and length of these repeats can vary. The National Cancer Institute selected 5 of these repeat loci as markers of instability. The pool of 22 candidate markers was used in this analysis and is shown below.

<table>
<thead>
<tr>
<th>Repeat Name</th>
<th>Location</th>
<th>Repeat Type</th>
<th>Length of Repeat (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D17S791</td>
<td>chr17: 44856265-44856598</td>
<td>Dinucleotide</td>
<td>334</td>
</tr>
<tr>
<td>D2S123</td>
<td>chr2: 51288378-51288657</td>
<td>Dinucleotide</td>
<td>208</td>
</tr>
<tr>
<td>D17S250</td>
<td>chr17: 37152091-37152341</td>
<td>Dinucleotide</td>
<td>251</td>
</tr>
<tr>
<td>D5S346</td>
<td>chr5: 112213624-112213748</td>
<td>Dinucleotide</td>
<td>125</td>
</tr>
<tr>
<td>BAT-25</td>
<td>chr4: 55598212-55598237</td>
<td>Mononucleotide</td>
<td>26</td>
</tr>
<tr>
<td>BAT-26</td>
<td>chr2: 47641560-47641587</td>
<td>Mononucleotide</td>
<td>28</td>
</tr>
<tr>
<td>D20S100</td>
<td>chr20: 54314053-54314305</td>
<td>Dinucleotide</td>
<td>253</td>
</tr>
<tr>
<td>D7S519</td>
<td>chr7: 46115995-46116332</td>
<td>Dinucleotide</td>
<td>338</td>
</tr>
<tr>
<td>D17S787</td>
<td>chr17: 53281885-53282241</td>
<td>Dinucleotide</td>
<td>357</td>
</tr>
<tr>
<td>D13S153</td>
<td>chr13: 48890734-48891008</td>
<td>Dinucleotide</td>
<td>275</td>
</tr>
<tr>
<td>D18S69</td>
<td>chr18: 53842399-53842730</td>
<td>Dinucleotide</td>
<td>332</td>
</tr>
<tr>
<td>D8S87</td>
<td>chr8: 35300280-35300489</td>
<td>Dinucleotide</td>
<td>210</td>
</tr>
<tr>
<td>D5S107</td>
<td>chr5: 83599568-83599763</td>
<td>Dinucleotide</td>
<td>196</td>
</tr>
<tr>
<td>D17S588</td>
<td>chr17: 48215496-48215654</td>
<td>Dinucleotide</td>
<td>159</td>
</tr>
<tr>
<td>D13S175</td>
<td>chr13: 20848380-20848734</td>
<td>Dinucleotide</td>
<td>355</td>
</tr>
<tr>
<td>D10S197</td>
<td>chr10: 26526881-26527153</td>
<td>Dinucleotide</td>
<td>273</td>
</tr>
<tr>
<td>D3S1029</td>
<td>chr3: 44135857-44136026</td>
<td>Dinucleotide</td>
<td>170</td>
</tr>
<tr>
<td>D18S64</td>
<td>chr18: 57426031-57426378</td>
<td>Dinucleotide</td>
<td>348</td>
</tr>
<tr>
<td>D18S61</td>
<td>chr18: 67436057-67436316</td>
<td>Dinucleotide</td>
<td>260</td>
</tr>
<tr>
<td>D18S58</td>
<td>chr18: 72044456-72044713</td>
<td>Dinucleotide</td>
<td>258</td>
</tr>
<tr>
<td>D18S55</td>
<td>chr18: 61873501-61873648</td>
<td>Dinucleotide</td>
<td>148</td>
</tr>
<tr>
<td>ACTC1</td>
<td>chr15: 35083509-35083555</td>
<td>Dinucleotide</td>
<td>47</td>
</tr>
</tbody>
</table>
Table 2.4. Datasets used for analysis

Publically available datasets were used from the ENCODE consortium. Wiggle and BED files were uploaded for analysis for four tissues. All available ChIP-seq experiments were uploaded for each donor for each tissue type. When multiple replicates were available, all were uploaded and used for analysis.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Donor ID</th>
<th>Gender</th>
<th>Datasets used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hippocampus</td>
<td>112 (Broad Institute)</td>
<td>Female</td>
<td>H3K4me1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K9ac</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K27me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K9me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RRBS</td>
</tr>
<tr>
<td>150 (Broad Institute)</td>
<td>Male</td>
<td></td>
<td>H3K4me1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K9me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K27me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K27ac</td>
</tr>
<tr>
<td>149 (Broad Institute)</td>
<td>Male</td>
<td></td>
<td>H3K27ac</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K9me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K4me1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>mRNA-seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RRBS</td>
</tr>
<tr>
<td>Skeletal Muscle</td>
<td>19 (Broad Institute)</td>
<td>?</td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K9ac</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K9me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K4me1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K27me3</td>
</tr>
<tr>
<td>62 (Broad Institute)</td>
<td>Female</td>
<td></td>
<td>H3K27ac</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K4me1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H3K4me3</td>
</tr>
<tr>
<td>Sample Type</td>
<td>Tissue</td>
<td>Gender</td>
<td>H3K36me3</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>HAc cells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSMM cells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kidney</td>
<td>27</td>
<td>Male</td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td>153</td>
<td>Male</td>
<td>H3K36me3</td>
</tr>
<tr>
<td>Liver</td>
<td>4</td>
<td>?</td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Female</td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Male</td>
<td>H3K27me3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4 Results

2.4.1 There is a symmetric distribution pattern centrally localized around specific tandem repeats

There are 46 disease-causing repeat loci made up of 18 different repeat sequences (Table 2.1). As repeat sequences have been shown to have an influence on chromatin structure and DNA conformation, I hypothesized that the unexpanded disease-causing repeat loci might show specific epigenetic patterns. To test this hypothesis, I compared disease-causing repeat loci to non-disease causing repeat loci of related and varying sequences within the genome.

I assessed the epigenetic status of 46 disease-causing repeat loci and 824 non-disease causing repeat loci with similar locations, sequences, and lengths using epigenetic marks including H3K9ac, H3K4me1, H3K36me3, H3K4me3, H3K27ac, mRNA-seq, H3K27me3, H3K9me3, DNaseHS, CTCF, individual CpG sites, and RRBS in the hippocampus (Fig. 2.1, heatmap). The enrichment of these epigenetic marks was observed in a 10 kb window centrally localized at the repeat. The enrichment data for all epigenetic marks at the 870 loci were clustered using K-means into 3 clusters. Within the three clusters there are distinct enrichment patterns. Cluster 1 shows enrichment of H3K36me3 and mRNA-seq data within the 10 kb window. Cluster 2 shows low levels of enrichment for all epigenetic marks. Cluster 3 shows a distinct enrichment pattern localized to the repeats. In this cluster there is a symmetric distribution of enrichment surrounding the repeats for H3K9ac, H3K4me1, H3K27ac, and H3K4me3. For these histone modifications, there is enrichment approximately 1 kb upstream and downstream of the centrally located repeat and a decrease of enrichment signal directly at the repeat itself (Fig. 2.1, graph). Other epigenetic marks, including DNaseHS sites, CpG sites, and CpG methylation show enrichment directly at the repeats and spanning about 1 kb (Fig. 2.1, graph). When all factors are taken into account, there is no significant separation between known disease-causing repeat loci and other repeat loci not known to cause disease within each cluster.
Figure 2.1. Hippocampus all epigenetic marks

46 disease-causing repeat loci, as well as 824 location, sequence, and repeat length control loci that are not known to cause disease were analyzed for enrichment of specific epigenetic marks. The enrichment of these epigenetic marks was observed in a 10 kb window centrally localized at the repeat. Enrichment data was analyzed for epigenetic marks including H3K9ac, H3K4me1, H3K36me3, H3K4me3, H3K27ac, H3K27me3, H3K9me3, mRNA-seq, DNase I hypersensitivity sites (DNaseHS), CTCF binding, CpG sites, as well as CpG methylation via reduced rate bisulfite sequences (RRBS). Enrichment data for all epigenetic marks at the 870 repeat loci were clustered using K-means in seqMINER into 3 clusters. The heatmap is a visual representation of the clusters showing the 870 repeat loci aligned vertically and each epigenetic factor’s enrichment aligned horizontally. Within the three clusters there are distinct enrichment patterns. Cluster 1 shows enrichment of H3K36me3 and mRNA-seq data within the 10 kb window which is an indication of an actively transcribed gene body. Cluster 2 shows low levels of enrichment for all epigenetic marks. Cluster 3 shows a distinct enrichment pattern localized to the repeats. In this cluster we see a symmetric distribution of enrichment centred on the repeats for H3K9ac, H3K4me1, H3K27ac, and H3K4me3. For these epigenetic marks, there is enrichment approximately 1 kb upstream and downstream of the centrally located repeat and a decrease of enrichment signal directly at the repeat itself. Other epigenetic marks, including DNaseHS sites, CpG sites and CpG methylation show enrichment directly at the repeats and spanning about 1 kb. The graph is a visual representation of the enrichment profile of the RRBS, CpG, H3K4me3, H3K4me1, and H3K9ac marks. The profile shows that CpG sites and methylation of those sites are enriched directly at the repeats, spanning about 1 kb at the centre of the symmetrically distributed pattern seen for the histone modifications.
2.4.2 Telomeric, SINE, and LINE repeat loci do not show symmetric distribution pattern centrally localized around the repeat

To determine if disease-causing repeat loci show an epigenetic pattern distinct from other repeat types found in the genome I analyzed the enrichment pattern of specific epigenetic marks at the 46 disease-causing repeat loci combined with 36 telomeric repeat loci, 8 short interspersed nuclear element (SINE) loci, and 8 long interspersed nuclear element (LINE) loci (Fig. 2.2). Each variety of repeat has a different function and therefore epigenetic status. To determine if disease-causing repeats show any unique epigenetic pattern, I have compared them to other repeats within the genome with known function and well studied behaviors. Two histone modifications, H3K4me3 and H3K27ac were analyzed (both modifications show a cluster with a symmetric distribution centred upon the repeats and were available for multiple tissue types) for enrichment at these loci and clustering revealed two distinct clusters. Cluster 1 shows low levels of enrichment throughout the 10 kb window, in contrast, Cluster 2 shows a high level of enrichment centrally located at the repeat. Cluster 2 contains only disease-causing repeat loci and no telomeric repeat loci, SINE loci, or LINE loci. Therefore, the symmetric distribution pattern is not seen for telomeric, SINE, or LINE repeat loci.

2.4.3 Mono- and dinucleotide repeats do not show symmetric distribution pattern centrally localized around the repeat

Mono- and dinucleotide repeats are another well characterized form of repetitive DNA in the genome. Microsatellite instability in specific mono- and dinucleotide repeats have been found in hereditary nonpolyposis colorectal cancer (HNPCC) tumours and in a proportion of nonhereditary colorectal tumours. The polymorphic nature of these repetitive elements is similar to the 46 disease-causing repeat loci. To determine if disease-causing repeat loci show an epigenetic pattern distinct from another repetitive element which shows instability, I analyzed the enrichment pattern of histone modifications H3K4me3 and H3K27ac in hippocampus at 22 mono- and dinucleotide repeat loci and 46 disease-causing repeat loci (Fig. 2.3). Cluster 1 shows low levels of enrichment throughout the 10 kb window and contains loci for mononucleotide, dinucleotide, and disease-causing repeat loci. Cluster 2 shows a high level of enrichment centrally located at the repeat. Cluster 2 contains only disease-causing repeat loci and no mono- or dinucleotide repeat loci. Therefore, the symmetric distribution pattern is not seen for mono- and dinucleotide repeat loci.
46 disease-causing repeat loci as well as 36 telomeric repeat loci (intrachromosomal and telomeric), 8 SINE loci, and 8 LINE loci were analyzed for epigenetic mark enrichment within a 10 kb window centrally localized on the repeat element. Two histone modifications, H3K4me3 and H3K27ac were analyzed for enrichment at these loci. When more than one dataset was available, all were included in these analyses. K-means clustering analysis revealed two distinct clusters. The heatmap is a visual representation of the clustering and shows Cluster 1 has low levels of enrichment within the 10 kb window, where Cluster 2 shows a high level of enrichment centrally located at the repeat. Cluster 2 contains only disease-causing repeat loci and no telomeric repeat loci, SINE loci, or LINE loci.
Figure 2.3. Mononucleotide, dinucleotide and 46 disease-causing repeat loci in hippocampus

46 disease-causing repeat loci as well as 22 mono- and dinucleotide repeats were analyzed for epigenetic mark enrichment within a 10 kb window centrally localized on the repeat element. Two histone modifications, H3K4me3 and H3K27ac were analyzed for enrichment at these loci. K-means clustering revealed 2 distinct clusters. The heatmap is a visual representation of the clustering. Cluster 1 shows low levels of enrichment and contains all of the mono- and dinucleotide repeat loci. Cluster 2 shows high levels of enrichment centrally located at the repeats and contains only disease-causing loci. Disease-causing repeat loci were found in both clusters but mono- and dinucleotide repeat loci were only found in Cluster 1. Therefore, Cluster 2 was composed entirely of disease-causing repeat loci.
2.4.4 Satellite I, II, and III repeat loci do not show symmetric distribution pattern centrally localized around the repeat

Satellite DNA is another well characterized form of repetitive DNA which refers to repetitions of a short DNA sequence that show a different density from genomic DNA when separated on a CsCl density gradient due to its G+C content\(^{124}\). I analyzed the 46 disease-causing repeat loci and 1236 loci with satellite I, II, or III repeats for histone modifications H3K4me3 and H3K27ac and found four distinct clusters (Fig 2.4). Cluster 1 shows low levels of enrichment and Cluster 2 shows high levels of enrichment. Cluster 3 shows a distinct enrichment pattern with low levels of enrichment far upstream of the repeat and a gradual increase of enrichment towards the downstream side of the repeats. Cluster 4 shows high levels of diffuse enrichment. There is a difference in repeat type within each cluster. Only Clusters 1 and 2 contained disease-causing repeat loci. Therefore, Cluster 3 and 4 were composed entirely of satellites I, II, and III loci. Repeat composition analysis of the clusters revealed that the proportion of disease-causing repeat loci, compared to satellites I, II, and III loci, within Cluster 1 and 2 is statistically different than expected (\(p=1.08 \times 10^{-9}\) and \(p=0.011\) respectively) (Fig 2.5). This indicates that disease-causing repeat loci when compared to satellites I, II, and III loci are mainly found in the lowly or highly enriched clusters. Satellites I, II, and III loci are found in all enrichment distribution patterns but are exclusive to Cluster 3 showing upstream enrichment and Cluster 4 showing high levels of diffuse enrichment. Therefore, the disease-causing repeat loci clustered separately than the satellite I, II, and III repeat loci.
Figure 2.4. Satellite I, II, III, and 46 disease-causing loci in hippocampus

46 disease-causing repeat loci as well as 1236 loci with satellite I, II, or III repeats were analyzed for epigenetic mark enrichment within a 10 kb window centrally localized on the repeat element. Two histone modifications, H3K4me3 and H3K27ac were analyzed for enrichment at these loci. K-means clustering revealed 4 distinct clusters. The heatmap is a visual representation of the clustering. Cluster 1 shows low levels of enrichment and contains 37 of the 46 disease-causing repeat loci (80.4%). Cluster 2 shows high levels of enrichment and contains the remaining 5 disease-causing repeat loci. Cluster 3 shows a distinct enrichment pattern with low levels of enrichment far upstream of the repeat and a gradual increase of enrichment towards the downstream side of the repeat. Cluster 4 shows high levels of diffuse enrichment. Disease-causing repeat loci were only found in two of the clusters with over 80% being found in Cluster 1.
Figure 2.5. Repeat composition analysis for disease-causing repeat loci and satellite I, II and III loci in hippocampus

This is a graphical representation of the proportion of disease-causing repeat loci or satellite I, II, or III loci within each cluster from Fig. 2.4. Cluster 1 contains over 80% of the disease-causing repeat loci which is statistically more than would be expected by chance (p=1.08 x 10^{-9}). In Cluster 2 there are also statistically more disease-causing repeat loci then expected by chance (p=0.011). There are no disease-causing repeat loci contained in Cluster 3 or 4.
Cluster 1 Cluster 2 Cluster 3 Cluster 4

Disease Control

Total Repeats in Control or Disease causing groups

p=1.08x10^-9  p=0.011
2.4.5 There is no tissue-specific difference seen in enrichment patterns when all loci are considered concurrently

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. While shown to be age- and disease- dependent, the mechanistic basis of inter-tissue variation, which even occurs in postmitotic neurons, is unknown. To determine if disease-causing repeat loci show epigenetic differences between tissues I assessed enrichment patterns for epigenetic marks in both hippocampus and skeletal muscle for the 46 disease-causing repeat loci and 824 location, sequence, and length matched control repeat loci (Fig. 2.6). There were three datasets available for both tissues sets including H3K4me3, H3K27me3, and mRNA-seq. Enrichment patterns for H3K4me3, H3K27me3 and mRNA-seq in hippocampus and skeletal muscle revealed four distinct clusters. Cluster 1 shows high enrichment of mRNA-seq data in both hippocampus and skeletal muscle. Cluster 2 shows high levels of enrichment of H3K27me3 in both tissues. Cluster 3 shows low and disperse levels of enrichment for all epigenetic marks. Cluster 4 shows high levels of enrichment of H3K4me3 in a symmetric pattern directly surrounding the repeats and spanning about 1 kb on either side with a decrease of enrichment directly at the repeat units. This pattern is seen in both hippocampus and skeletal muscle. Thus, there is no tissue specific epigenetic pattern evident.

2.4.6 There is significant repeat sequence segregation in clusters showing low levels of enrichment and a symmetric distribution

The genomic landscape is determined by many factors including DNA sequence. It has been established that some repetitive sequences preferentially assemble or exclude nucleosomes. The epigenetic environment of disease-causing repeat sequences in control tissues has not yet been evaluated. To determine if the epigenetic environment is affected by repeat sequences known to cause disease I analyzed the enrichment of histone modification H3K9ac in hippocampus for 46 disease-causing repeat loci and 824 matched control repeat loci, matched for genic location, sequence, and length. With this comparison I found three distinct clusters (Fig. 2.7). Cluster 1 is lowly enriched for H3K9ac within the 10 kb window. Cluster 2 shows higher enrichment for H3K9ac and a dispersed pattern across the 10 kb window. Cluster 3 shows a symmetric distribution of enrichment with two areas of high enrichment about 1 kb on either side of the
repeat with a decrease of enrichment directly at the repeat. Cluster analysis for repeat sequence composition of each cluster showed significant differences between repeat sequence compositions between clusters. The proportion of 2 of the 18 sequences was used in this analysis, CGG and CAG repeats. To statistically analyze clusters I performed Fisher’s Exact Test which requires groups to contain 10 or more samples. Consistently throughout my cluster analysis, CGG and CAG repeat containing clusters had more than 10 loci in each cluster. Therefore, these are good sequence candidates to perform statistical analysis on. Also, each repeat is known to have the opposite affect on nucleosome assemble, so where CAG is known to be a strong nucleosome positioning sequence, CGG is known not to have a strong nucleosome positioning effect. Cluster 1 shows significantly less CGG and more CAG repeats then by chance \( (p=1.6 \times 10^{-9}) \). Cluster 3 showed significantly more CGG and less CAG repeats then by chance \( (p=2.8 \times 10^{-16}) \). This pattern of significantly less CGG and more CAG repeats in the lowly enriched cluster and more CGG with less CAG repeats in the highly enriched cluster was also found in hippocampus for enrichment of H3K27me3, H3K27ac, H3K4me1, H3K4me3, H3K36me3, RRBS, CpG and DNaseHS and in skeletal muscle for enrichment of H3K27me3, H3K27ac, H3K4me1, H3K36me3, and H3K4me3.
Figure 2.6. Tissue specific evaluation of epigenetic marks at disease-causing and non-disease-causing repeat loci in hippocampus and skeletal muscle

46 disease-causing repeat loci and 824 location, sequence, and length matched repeat loci not known to cause disease were analyzed for enrichment of H3K4me3, H3K27me3 and mRNA-seq in both hippocampus and skeletal muscle. K-means clustering analysis revealed 4 distinct clusters. The heatmap is a visual representation of the enrichment patterns in each cluster over a 10 kb window centrally located on the DNA repeat. Cluster 1 shows high enrichment mRNA-seq data in both hippocampus and skeletal muscle. Cluster 2 shows high levels of enrichment of H3K27me3 in both tissues. Cluster 3 shows low and disperse levels of enrichment for all epigenetic marks. Cluster 4 shows high levels of enrichment of H3K4me3 in a symmetric pattern directly surrounding the repeat and spanning about 1 kb on either side with a decrease of enrichment directly at the repeat unit. This pattern can be seen in both hippocampus and skeletal muscle.
Figure 2.7. Disease-causing and non-disease-causing repeat loci enrichment for H3K9ac in hippocampus

46 disease-causing repeat loci and 824 location, sequence, and length matched repeat loci not known to cause disease were analyzed for enrichment of H3K9ac in hippocampus tissue. K-means clustering analysis revealed three distinct clusters. The heatmap is a visual representation of these clusters. Cluster 1 is lowly enriched for H3K9ac within a 10 kb window centrally localized on the repeat. Cluster 2 shows higher enrichment for H3K9ac and a dispersed pattern. Cluster 3 shows a symmetric distribution of enrichment with two areas of high enrichment about 1 kb on either side of the repeat with a decrease of enrichment directly at the repeat. The graph is a visual representation of the distribution of sequences within each cluster. Each of the 18 repeat sequences is represented as a different colour. The first lane shows the distribution of sequences within the 824 non-disease-causing repeat loci as a proportion of the total. For Cluster 1 there is a different distribution of repeat sequences making up the total loci for that cluster. The repeat sequences within each cluster were analyzed and Fisher’s Exact Test was sued to compare the proportion of CGG or CAG repeat loci in each cluster to the total population of non-disease-causing repeats. There are significantly less CGG and more CAG repeat in Cluster 1 than expected by chance ($p = 1.6 \times 10^{-9}$). Cluster 3 showed significantly more CGG and less CAG repeats than expected by chance ($p = 2.8 \times 10^{-16}$). This pattern of significantly less CGG and more CAG repeats in the lowly enriched cluster and more CGG with less CAG repeats in the highly enriched cluster was also found in hippocampus for enrichment of H3K27me3, H3K27ac, H3K4me1, H3K4me3, H3K36me3, RRBS, CpG, and DNaseHS and in skeletal muscle for H3K27me3, H3K27ac, H3K4me1, H3K36me3, and H3K4me3.
2.5 Conclusions and Discussion

I have characterized the epigenetic chromatin context at different sequences of repetitive units in the genome. It has long been established that different repeated sequences within the genome have specific biological functions. A recent genome-wide analysis identified that 18.8% of all human genes contain at least one highly pure short tandem repeat in their upstream regulatory region. The function of the DNA sequence will be a factor in determining the epigenetic environment surrounding it. It has been suggested that local chromatin environment around a repeat is important in determining instability. I have found that repeat sequences that are known to cause 46 neurological and neuromuscular diseases display a specific and unique chromatin pattern (Fig 2.1). The epigenetic pattern shows a symmetric distribution of enrichment where there are high levels of enrichment upstream and downstream of the repeat spanning about 1 kb and a decrease of enrichment directly at the repeat. This pattern of enrichment for histone modifications, as well as other epigenetic marks, is not replicated with other types of repeats found within the human genome. When compared to telomeric repeats either at terminal chromosomal locations or interchromosomally, the epigenetic pattern was not recapitulated (Fig 2.2). Indeed, the disease-causing repeat loci all clustered within Cluster 2 in which there were no telomeric repeat loci. A similar enrichment pattern was seen when the enrichment for the same histone modifications were analyzed at mononucleotide, dinucleotide and disease-causing repeat loci (Fig 2.3). When compared to mono- and dinucleotide repeats, the disease-causing repeat loci clustered separately. Cluster 2 was composed of all disease-causing repeat loci and no mono- or dinucleotide repeat loci. This cluster distinction between mono- and dinucleotides repeat loci and disease-causing repeat loci is interesting because all of these forms of repeats have the potential to become unstable. Another variety of repeat elements within the human genome are satellites I, II, and III. These classical and well characterized repeats are closely related pentanucleotide ATTCC sequences isolated from bulk DNA by CsSO4 gradients and represent about 5% of the total human genome. Epigenetic analysis of these repeats showed a distinct chromatin pattern different from the pattern seen for the disease-causing repeat sequences (Fig 2.4). Satellite I, II, and III loci showed higher levels of enrichment in a more diffuse pattern for the two active histone modifications H3K4me3 and H3K27ac in the hippocampus.
Telomeres are the tips of chromosomes and consist of protein complexes and noncoding DNA comprising hundreds to thousands of tandem TTAGGG hexanucleotide repeats. Typically, telomeres shorten with aging (a primary risk factor for neurodegenerative disease). The major part of telomeres is folded into nucleosomes, and the chromatin environment at telomeres has long been investigated. Several studies have found heterochromatin-specific histone modifications associated with telomeric and subtelomeric chromatin. In Figure 2.2, all telomeric repeat loci fall within Cluster 1 which shows low levels of enrichment for H3K4me3 and H3K27ac in human hippocampus. When repressive histone modification H3K9me3 and H3K27me3 were analyzed, they showed low levels of enrichment in all clusters. The same pattern for all histone modifications was seen when analyzed in skeletal muscle, kidney and liver tissues. Telomeric repeats located at telomeres or within chromosomes do not show the same distinct nucleosome positioning effect as the 46 neurodegenerative disease-causing repeat sequences even though telomere shortening is also known to be correlated to neurodegenerative disease.

Short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) are another class of repetitive units within the genome. SINEs and LINEs are classes of retrotransposons able to move through the genome via an RNA intermediate and integrate into a new genomic location. DNA methylation plays a key role in the regulation of gene expression overall, including keeping transposable elements transcriptionally silent. In Figure 2.2, all SINE and LINE loci fall within Cluster 1 which shows low levels of enrichment for active histone modifications H3K4me3 and H3K27ac. SINE and LINE repeat loci do not show the same distinct nucleosome positioning effect as the 46 neurodegenerative disease-causing repeat sequences.

DNaseHS sites are an effective marker for DNA accessibility, and are often used to probe the chromatin packaging state. In Figure 2.1, the DNaseHS tract shows high levels of enrichment directly at the repeat unit in loci within Cluster 3. This enrichment pattern is evidence that the repeat is kept in an open accessible state. Flanking the repeats in Cluster 3 were high levels of enrichment of H3K9ac, H3K4me1, H3K4me3, and H3K27ac directly upstream and downstream of the repeat units. Enrichment pattern showing high enrichment directly upstream and downstream of the repeats and then a decrease of enrichment at the repeat itself was considered symmetric. The symmetric nature of this distribution of histone modifications may be due to the
repeats having an effect on nucleosome positioning effectively creating a nucleosome free area directly at the repeats and positioning a nucleosome directly upstream and downstream of the repeats. In hippocampus, histone modifications H3K9ac, H3K27ac and H3K4me3 showed this pattern of symmetric peaks centrally localized around the repeat. There were 13 disease-causing repeats that were consistently in the cluster showing symmetric peak distribution for the three histone modifications. They are C9orf72, CBFA1, EMP1, FRA10a, FRA11b, FRAXA, HD, H-Ras, MAB21l1, OPMD, SCA12, SCA2, and SCA7. In skeletal muscle, histone modifications H3K27ac and H3K4me3 showed the symmetric peak pattern around the repeats. There were 10 disease-causing repeats that showed the symmetric distribution in both histone modifications. They are C9orf72, DRPLA, FRA10a, FRA11b, FRAXA, HD, H-Ras, OPMD, SCA2, and SCA7.

DNA methylation plays a role in chromatin structure, gene expression and maintaining genome stability in repetitive DNA\textsuperscript{132}. DNA methylation is strongly associated with dense chromatin and the methylation of the CpG rich regions at the Fragile X loci alters the repeat stability in mouse\textsuperscript{100}. Further, DNA methyltransferase inhibitors destabilize the myotonic dystrophy repeat locus in humans\textsuperscript{133}. This implicates DNA methylation within or near a repeat as a suppressor of instability. In Figure 2.1 DNA methylation was measured via RRBS of CpG sites. Cluster 3 was the only cluster to show enrichment of CpG methylation directly on the repeat sequences of the loci within that cluster. In Cluster 3 at the repeat sequences there is concurrent enrichment for both DNaseHS sites and CpG methylation. Abberant DNA methylation directly upstream and downstream of the CTG/CAG repeat, was found in tissues from patients with myotonic dystrophy type 1\textsuperscript{75}. There is evidence that chromatin may be remodeled (via nucleosome positioning and histone modifications) to allow DNA methyltransferases access to targets\textsuperscript{132}. If chromatin is not properly re-modeled, then DNA methylation may not occur and unmethylated repeats may become unstable.

CTCF is an 11 zinc-figure DNA binding protein that plays a role in genomic imprinting, chromatin localization, DNA insulation and transcriptional regulation. It affects higher order chromatin structure and has been theorized to play a role in repeat instability\textsuperscript{51}. At the SCA7/ATXN7 locus and four other CAG/CTG repeat loci, CTCF binding has been show to have an influence on instability\textsuperscript{51}. Hypermethylation of CTCF binding sites have been shown to contribute to instability, particularly due to the increase in antisense transcription from loss of the CTCF insulator function\textsuperscript{51}. CTCF analysis at 870 repeat loci showed no discernable pattern,
possibly because of the large 10 kb observational window at each locus. Further investigation is necessary to determine the status of CTCF binding at these loci in control tissues.

Nucleosome placement is a highly regulated process for genome organization. Previous studies have shown specific repetitive sequences to either preferentially assemble or exclude nucleosomes. Strong nucleosome assembly occurs at CTG repeats\(^4\), GAA repeats\(^5\), (A/T),NN(G/C)\(_2\)NN repeats\(^6\), TATAAACGCC repeats\(^7\), and (TTGA)\(_n\) repeats\(^8\). Repeats that have been shown to exclude nucleosomes include CGG repeats\(^9\), and poly(dA) repeats\(^10\). Telomeres are composed of TTAGGG repeats. In telomeres shorter than 7 kb there is exclusion of nucleosomes\(^11\).\(^7\). Investigation of repeat sequences through individual cluster analysis revealed that there is a significant sequence-specific effect on enrichment of specific histone modifications. Figure 2.7 shows the proportion of each repeat sequence in the three clusters. There are significantly more CAG repeats and less CGG repeats than expected in the lowly enriched Cluster 1 and the opposite is true of the highly enriched Cluster 3. This pattern of a significantly lower proportion of CGG to CAG repeat sequences in the lowly enriched cluster and higher proportion of CGG to CAG repeats in the highly enriched cluster was also seen for histone modifications H3K27me3, H3K27ac, H3K4me1, H3K36me3, and H3K4me3 in skeletal muscle. In hippocampus this sequence specific effect was seen for enrichment of H3K27me3, H3K27ac, RRBS, CpG, H3K4me1, H3K36me3, H3K4me3, and DNaseHS in HAc cells. Telomeric repeat analysis as well as analysis of Satellites I, II, and III did not show specific nucleosome placement or significant repeat sequence effects when comparing clusters. These results support previous work establishing that repetitive DNA shows a sequence specific effect for nucleosome positioning. H3K36me3 showed the same pattern of significantly more CGG repeats and less CAG repeats in the enriched cluster but did not show the same pattern for H3K9ac, H3K27me3, H3K27ac, RRBS, CpG, DNaseHS, CTCF, H3K4me1, and H3K4me3. Histone modification H3K9me3 did not show this sequence specific pattern. Further analysis of sequences within clusters showed that repeat sequences TGGAA, ATTCT, and AT were consistently in the lowly enriched clusters (Figure 2.7). This was not due to a lack of nucleosome positioning because other histone modifications that I investigated showed high levels of enrichment (Figure 2.7). Some repeat sequences were found in the enriched or non-enriched clusters regardless of histone modification. This is interesting because simultaneous enrichment for histone modifications indicating activity and repression provide information about function
and epigenetic context. For example, concurrent enrichment of H3K4me1 and H3K27me3 is indicative of a poised enhancer where either histone modification on its own indicates another chromatin state. Although my analysis did not find any clusters with simultaneous histone enrichment indicating poised enhancers, I did consistently find simultaneous lack of enrichment for specific repeat sequences TGGAA, ATTCT, and AT.

Altered tandem repeat length in or near core promoters can change the local nucleosome positioning and is likely to hinder transcription factor binding and therefore affect rates of transcription and hence gene expression\(^{134,135}\). A recent study has shown that a polymorphic GA-repeat in the human SOX5 gene promoter can affect gene expression, with the longer allele resulting in a 2.7-fold increase in activity\(^{136}\). The authors report this as first evidence of a functional short tandem repeat in a human gene core promoter. The distribution of short tandem repeats around promoters has been proposed to have functional significance\(^{121}\). It has been established that as many as 10-20% of eukaryotic genes and promoters contain an unstable repeat tract\(^1\). Mutations in these repeats can have phenotypic consequences, either causing disease or conferring better fitness and evolutionary advantage. The polymorphic nature of these tandem repeats has been shown to be a conserved mechanism in many organisms. Figure 2.1 shows the epigenetic context at 870 tandem repeats, a section of which show histone modification enrichment congruent with active promoters (Cluster 3). These repeat loci may show more mutability within the repeat. The mechanism for repeat instability and specificity of certain repeats to cause disease and others with the same sequence to not cause disease remains unknown. However, we provide a human context to 870 tandem repeat loci and show the epigenetic environment in control tissues.

Many of the disease-causing repeats show tissue specific instability where certain tissues show greater rates of expansion compared to others. The mechanism behind tissue-specific instability is unknown. Because instability can occur in post-mitotic cells, such as neurons, epigenetic factors are thought to play a key role in the expansion of non-dividing cells. We investigated the epigenetic environment at 46 disease-causing repeat loci in hippocampus, skeletal muscle, kidney and liver to evaluate the non-expanded status. In Figure 2.1 we saw the overall chromatin environment in hippocampus. When skeletal muscle, kidney and liver were investigated in the same way, the same clusters were formed with the same enrichment patterns. Further investigation found some specific differences between tissues. For example, when
enrichment of H3K27me3 was investigated high levels of enrichment were found in skeletal muscle for HDL2, ZIC2, FOXL2, HOXD13, and HOXA13 loci. In hippocampus for the same histone modification the loci with enrichment were C9orf72, FRDA, FRAXA, ZIC2, FOXL2, HOXD13, and HOXA13. Therefore, the HDL2 locus only showed H3K27me3 enrichment in skeletal muscle, where the C9orf72, FRDA and FRAXA loci only showed H3K27me3 enrichment in hippocampus. The disease-causing repeat loci that were found to be enriched for H3K27me3 are expressed mainly in the brain with only a small amount of muscle involvement in the phenotype. This is interesting because H3K27me3 is a marker of facultatively repressed genes. It would be interesting to assess the activity and repression of these disease-causing loci in patient tissues and observe if a change in epigenetic status is correlated to the disease.

In the hippocampus, the HOXD13 locus was found to be enriched for H3K9me3, but not in skeletal muscle, kidney or liver. This is interesting because H3K9me3 is a repressive mark of constitutive heterochromatin. HOXD13 is also the only disease-causing repeat loci found to have enrichment for H3K9me3.

There were also tissue specific differences when analyzing histone modification H3K9ac. When observing the enrichment for H3K9ac, a mark of actively transcribed promoters, in skeletal muscle, it seems as though there was poor ChIP-seq coverage. However, patterns could be seen forming between hippocampus, liver and kidney. When looking at hippocampus alone (Fig. 2.7), we see a symmetric distribution of the most enriched cluster, Cluster 3. This cluster contains many disease-causing repeat loci, but we see a loss of some of these loci when all four tissue types are considered together. The hippocampus specific disease-causing loci that show a symmetric distribution pattern are DM1, SEF21b, FRA11b, SCA2, SCA7, OPMD, FRA10a, KCNN3, HD, C9orf72, and FRAXA. The loss of these disease-causing loci enrichment is countered by a gain of ZIC2, DRPLA, SCA1, and SBMA, in the symmetric distribution cluster, when other tissues are considered.

There are tissue-specific differences between disease-causing repeat loci when the histone modification H3K27ac, a mark of active enhancers, was observed (Fig. 2.1). High levels of enrichment were found spreading upstream and downstream of the disease-causing repeats DM1, SEF21b and DM2 in skeletal muscle, but for the same cluster pattern in hippocampus the disease-causing repeats were DM1, Zic2, and DRPLA. CTCF binding has been shown to act as a
chromatin looping modifier and can act to bring enhancers and promoters closer together. It is therefore interesting to note that DM1 showed enrichment for CTCF close to the repeat, as did FRA11b and FRA10a (Fig. 2.1). It has been previously shown that altered methylation of the CTCF binding sites can have an effect on instability making CTCF binding and H3K27ac enrichment important marks for further investigation. There was also a localized enrichment of H3K27ac surrounding a group of disease-causing loci which contained the CTCF enriched FRA11b and FRA10a loci. Another observation about the histone modification H3K27ac is that there is an additional cluster showing high levels of enrichment but extremely localized to the repeats. This cluster is observed in hippocampus and skeletal muscle but with a different repeat loci composition. The hippocampus specific loci are SEF21b, MAB21l1, SCA12, EMP1, CBFA1, KCNN3 and the skeletal muscle specific loci is DRPLA. When the two tissues are analyzed together, the differences become apparent. There seems to be more global H3K27ac levels in skeletal muscle tissue, but more specific enrichment surrounding the repeats in hippocampus. These differences were not statistically significant, but provide evidence that this modification may play a part in the epigenetic environment leading to tissue specific instability. It is interesting to note that although most repeat-causing diseases are neurodegenerative, the vast majority of disease-causing loci show a lack of active histone modification within the hippocampus. It would be beneficial to access the amount of enrichment of active histone modifications in diseased tissues.

Overall, there were tissue-specific findings when all histone modifications showing the symmetric distribution around the repeats were considered. In hippocampus there were three histone modifications which showed a cluster with a symmetric distribution around the repeat including H3K9ac, H3K27ac, and H3K4me3. There were disease-causing repeat loci consistently in the symmetrically distributed cluster for all these histone modifications; C9orf72, CBFA1, EPM1, FRA10a, FRA11b, FRAXA, HD, H-Ras, MAB21l1, OPMD, SCA12, SCA2, and SCA7. In skeletal muscle there were two histone modifications that showed a cluster with a symmetric distribution surrounding the repeats; H3K27ac, and H3K4me3. There were disease-causing repeat loci consistently in the symmetrically distributed cluster for these two histone modifications; C9orf72, DRPLA, FRA10a, FRA11b, FRAXA, HD, H-Ras, OPMD, SCA2, and SCA7. Therefore, there are tissue-specific differences when each individual disease-causing
repeat loci is considered independently. The tissue-specific differences are also apparent between histone modifications when analyzed independently.

The CpG data were comparable regardless of tissues as they are displaying the amount of CpG sites within the genome. Interestingly the methylation data was also the same between tissues, meaning that the same disease-causing loci were enriched for methylated CpG sites within each tissue.

Another mark of activity is histone modification H3K4me1, which marks transcriptionally active enhancers. Many of the disease-causing repeat loci showed enrichment for H3K4me1 and the pattern of enrichment was maintained between tissues (Fig. 2.1). When analyzed individually though, it was seen that enrichment for H3K4me1 in the hippocampus showed a significant finding, that in the cluster showing downstream enrichment, there were more disease-causing repeat loci than expected. Although the enrichment pattern was similar with skeletal muscle, the statistically significant repeat distribution was not recapitulated.

When other repeat variations were analyzed, no tissue-specific effect was observed. Telomeric, SINE, LINE, and classical satellite loci showed no tissue-specific effect. Therefore, only the repeat classification known to cause disease showed some tissue-specific epigenetic marks.
3 Conclusions

3.1 The epigenetic environment at repetitive elements within the human genome

There are a variety of repetitive elements, which make up a large portion of the human genome (Fig. 1.1). The classification of repetitive elements takes into account sequence, purity, location, length, and function of each repeat. Each of these factors has been shown to correlate to specific epigenetic factors such as histone modifications, DNase I hypersensitivity sites, protein binding, DNA methylation, and RNA transcription. As technology has advanced and made genome-wide data available, it is possible to further investigate the differences and similarities between repetitive elements using these genome-wide epigenetic marks. Here, we provide a human tissue control context to evaluate the epigenetic environment at many different varieties of repeats. We have been able to establish the chromatin context at telomeric repeats, tandem repeats, mono- and dinucleotide repeats, and satellite I, II, and III repeats containing centromeric repeats. For each of these repeat variations we evaluated specific epigenetic marks in human hippocampus, skeletal muscle, kidney, and liver tissues. To evaluate the activity at a specific loci containing a repetitive element we observed the enrichment for active histone modifications including H3K9ac, H3K27ac, H3K4me1, H3K4me3, and H3K36me3 as well as DNase I hypersensitivity sites, CTCF binding, and mRNA abundance. To evaluate the level of repression at a locus containing a repetitive element we observed the enrichment of repressive histone modifications including H3K9me3 and H3K27me3 as well as DNA CpG methylation. The link between repetitive elements and the epigenetic environment has yet to be established, but here we provide a base-line for further researchers to utilize when investigating disease, protein binding, polymorphic ability, and chromatin structure. Further investigation is warranted into the symmetric histone modification enrichment distribution seen at clusters of repeat loci. This symmetric pattern is mainly seen for microsatellite loci and could be due to repeat sequences, repeat length, flanking sequence or another epigenetic factor. Experimental assays using in vitro methods could elucidate which of the factors is most involved in the symmetric enrichment pattern. Probably, within each 10 kb genomic window selected for analysis, there are other repeat sequences. This is interesting because the symmetric enrichment distribution was not
observed in other areas of the 10 kb window for all loci. That distribution pattern seems to be localized upon those specific repeats.

Further investigation is also necessary to determine if there is a correlation between loci showing the symmetric histone modification enrichment distribution and specific gene functions. Previous studies in yeast have found that repeat sequences in promoters are mainly found in cell wall and transcription factor genes\(^2\). A simple computer analysis could determine if there are gene groups that constitute a significant proportion of the loci showing the symmetric peak distribution.

Further investigation into a possible function for the symmetric distribution around specific repeats should be conducted. It would be interesting to determine if there are expression differences for clusters showing the symmetric distribution pattern or if there are any other functional implications for this epigenetic pattern.

It is also possible that the symmetric distribution pattern around the repeat is caused by a large protein complex associating with the repeats and preventing nucleosome positioning. Although the DNase hypersensitivity and CpG methylation data seem to indicate the DNA is free from both nucleosomes and proteins, there are methylation insensitive proteins which could show such effects. Further investigation is warranted to distinguish between the possibility of a protein complex binding to the repeats or the nucleosome positioning of the unbound repeats.

3.2 Disease-associated repeats

There are over 40 human genetic diseases caused by repeat locus-specific instability. These polymorphic repeats are characterized by ongoing somatic expansions as the individual ages. Different levels of expansions arise in specific tissues of the same individual, where expansions can be greater in affected tissues. The causes of repeat instability are poorly understood and a mechanism that explains their occurrence in full has yet to be identified. Numerous studies have shown that epigenetic modifications and \textit{cis}-elements play a role in repeat instability. In this study, publicly available epigenetic mark datasets (histone modifications, DNA methylation, CTCF binding, DNase I hypersensitivity, CpG sites and DNA methylation) were analyzed at genomic locations of disease-causing repeat loci and sequence, length and location matched repeat loci that are not known to cause disease. We assessed these in various tissues including hippocampus, skeletal muscle, liver and kidney. Specifically, we identified patterns of epigenetic
marks at repeat loci. Histone modification H3K27me3 showed significantly (p=0.0035) more
disease-causing repeat loci with enrichment than non-disease causing repeat loci in both
hippocampus and skeletal muscle tissues. Tissue-specific patterns were also identified where
histone modification H3K27ac showed a significant (p=0.03) difference between disease-causing
repeat loci and controls within hippocampus but not in skeletal muscle. These patterns are
priority marks of unstable repeats and make ideal candidate elements for analysis in diseased
tissues. This work provides future investigators with knowledge of the epigenetic state on the
disease-causing loci in control tissues. Further, we have provided evidence to show which
epigenetic factors are most efficiently investigated for specific diseases, saving potential
researchers time and finances.
References


