Development of an Allele-Aware Method to Study the Nuclear Organization of the FSHD Locus

Kyle T. Siebenthall

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Barbara Trask, Chair

Stephen Tapscott

Evan Eichler

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Development of an Allele-Aware Method to Study the Nuclear Organization of the FSHD Locus

Kyle T. Siebenthall

Chair of the Supervisory Committee:
Professor Barbara Trask
Department of Genome Sciences, University of Washington, and
Division of Human Biology, Fred Hutchinson Cancer Research Center

The human genome carries out its functions, including the establishment and maintenance of tissue-specific gene expression programs, within the confines of the cell nucleus. The genome is not randomly organized within this space: its three-dimensional arrangement plays a role in executing its functions. One hallmark of nuclear organization is the compartmentalization of active and inactive genomic regions, with little contact between regions of different states. In many different cell types, the 4q subtelomere is located at the nuclear periphery, a repressive region adjacent to the nuclear lamina. Deletions in a tandem repeat array at this locus lead to the muscular dystrophy FSHD through an epigenetic mechanism that results in inappropriate expression of *DUX4*, a transcription factor encoded in each repeat unit that is normally active in the germline. *DUX4* is toxic to muscle cells, yet its high abundance in only a small fraction of muscle cells from patients hints that there might be multiple determinants of its expression. My thesis work addressed the hypothesis that nuclear organization is one of these determinants. Since the FSHD locus is duplicated in the genome, I developed an allele-aware method (4C-seq) that I used to interrogate the "nuclear neighborhood" of specific copies of the locus in primary muscle cells from control individuals and individuals with FSHD. 4C-seq uses a polymorphic "bait" to capture regions of the genome ("prey") that are in close physical proximity to the bait in a population of nuclei. By characterizing prey fragments captured by a bait within the FSHD locus, I found that the locus normally contacts lamin-associated regions with low transcriptional output within its own chromosome territory, and associates with putative insulator sequences (bound by the CTCF protein) and centromeres of other chromosomes. In FSHD cells, I found that the locus carrying a pathogenic deletion contacts regions within its own chromosome that have lower lamin-association and higher transcriptional output than the regions contacted in control cells. My results suggest that an altered nuclear neighborhood might play a role in the mis-expression of *DUX4* in FSHD.

TABLE OF CONTENTS

TABLE OF CONTENTS, CONTINUED

LIST OF FIGURES

LIST OF FIGURES, CONTINUED

LIST OF TABLES

Chapter 1: The nucleus in 3D

1.1 Nuclear organization, chromatin structure and genome regulation

The nucleus is the command center of a cell, where instructions encoded in DNA direct the activities that make a heart cell beat, help a brain cell send signals, allow an intestinal cell to process nutrients, and charge an immune cell to attack foreign invaders. The genome is the general in charge of a cell's construction and operation; not only must it direct what and how to build, it has to receive reports from the field, determine a course of action, and send out envoys to enforce a decision. A myriad of proteins live inside the nucleus to help the genome carry out its function, and even to help it fit inside its home: nearly two meters of linear DNA must be compressed into a structure that is typically only six thousandths of a millimeter in diameter.

The human genome contains approximately 20,000 protein-coding genes (in addition to thousands of other genes which do not encode proteins) that orchestrate the production of over hundreds of different cell types that make up our bodies[1]. However, not every gene is expressed in every cell: different cell types are defined by the subset of genes they express. This subset consists of a core set of genes involved in carrying out basic cellular functions such as DNA replication and mRNA translation that are required in every cell, along with a set of tissue-specific genes that carry out functions unique to a given cell type. It follows, then, that the purpose of organizing the genome within the nucleus could be to maintain the expression of the genes that define a given cell type, while at the same time repressing genes that define alternative cell types.

The strong interrelationship between structure and function is a basic tenet of biology, and can be seen in how the shape of an enzyme's active site determines what substrates it can modify or in how the structure of a finch's beak allows it to specialize in a particular food source. The nucleus is no exception to this rule. But while the physical structure of the organelle itself is well-described, and the identities of many of the proteins that help build it are known, the three-dimensional arrangement of the genome within the interphase nucleus— and how or whether this arrangement dictates its function— is difficult to probe, so has remained a mystery until quite recently. **Figure 1.1** depicts some of the basic principles of nuclear organization, which are described below[2-6].

The nucleus is bounded by a double-membrane that is traversed by pores that allow selective import and export. Underneath the inner membrane lies a meshwork of filamentous proteins called the nuclear lamina, which serves as a scaffold for various proteins that interact with DNA and link the nucleus to the cell's cytoskeleton. DNA in the nucleus is packaged into chromatin, a

pearl-necklace-like complex in which the DNA is wrapped tightly around histone proteins to form nucleosomes, which regulate access of other proteins to the DNA. The genome shares the interior of the nucleus with a number of proteinacious structures, called nuclear bodies. The most prominent of these is the nucleolus, the site where the genes encoding ribosomes coalesce and are regulated. Other bodies help carry out the basic functions of the genome, including gene expression, splicing and DNA replication.

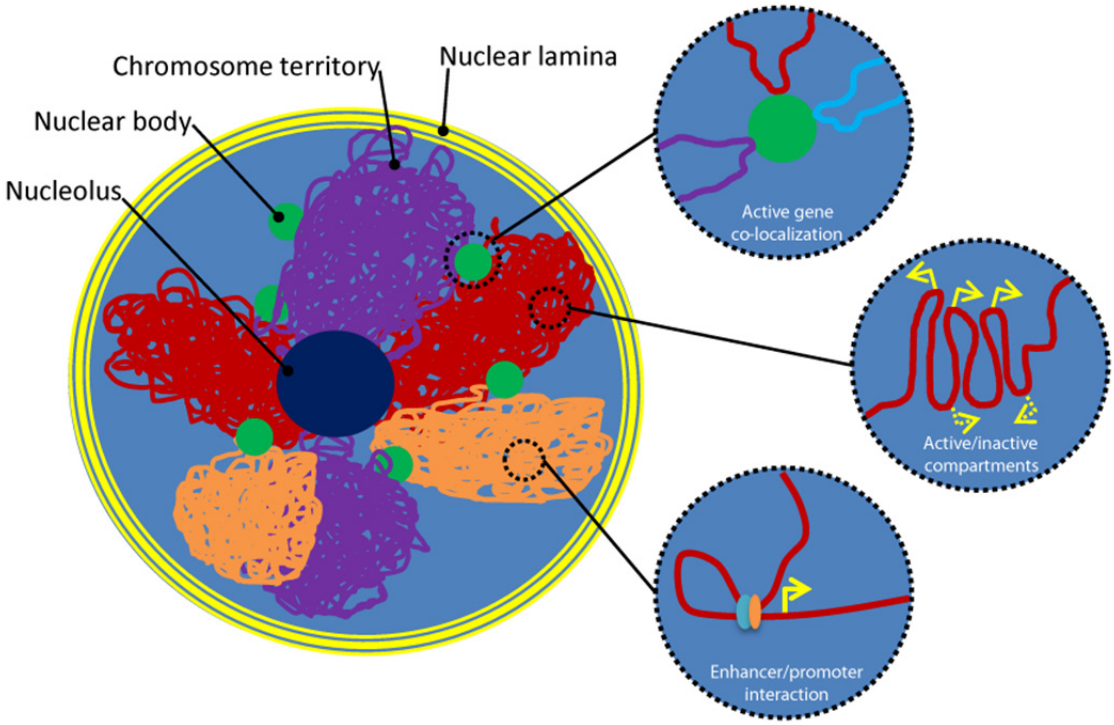

Figure 1.1 Overview of nuclear organization
Cartoon depiction of the basic features of nuclear organization that are described in the main text. Chromosome territories (colored lines), share the nuclear space with nuclear bodies (green circles), including the nucleolus (dark blue circle). Circles to the right depict local aspects of chromatin organization, including the co-localization of active genes at transcription factories, the compartmentalization of active and inactive loci, and looping of enhancers to contact promoters.

During division of a human cell, 46 chromosomes must be replicated, tightly compacted, and then divided evenly between two daughter cells. After this division, the nuclear membrane reforms around the chromosomes, which then de-condense to fill the nuclear volume. Largely as a consequence of loosening up within this space, de-condensed chromosomes occupy sub-volumes of the interphase nucleus, termed chromosome territories (CTs). CTs are not randomly organized within the nucleus, but have preferred radial positions[7-9]**.** While this size-correlated radial position could be a consequence of the mitotic forces acting differentially on small versus large

chromosomes[7,8,10], size isn't everything, as radial positions are also correlated with the gene density of chromosomes[7]. For example, chromosomes 18 and 19 have a similar length, but differ in the number of genes they encode. Gene-poor chromosome 18 is typically located closer to the nuclear periphery than is gene-rich chromosome 19 (Ref. 7).

On a coarse level, individual loci within a CT appear to be arranged in relation to each other as if they reside on a random polymer[10,11]. However, on closer examination, the activity states of loci appear to dictate their associations with one another and positions in nuclear space. This activity-correlated positioning of loci with respect to each other, their CTs, and other nuclear landmarks also changes in response to stimuli or differentiation[12-15]. Active genes have been found outside the apparent borders of their CT and appear to be transcribed in bursts when they associate with "transcription factories", concentrated foci of active RNA Polymerase II[12,16]. In contrast, silenced genes often reside at the nuclear periphery, adjacent to the nuclear lamina[6,17,18]. These genes may be silenced completely, never to be used again in a given lineage, or they can move away from the lamina during further differentiation[18]. On an even smaller scale, the chromatin fiber itself can fold in ways that bring regulatory sequences that may be far apart in linear sequence into close physical proximity to control gene expression.

A closer look at beta-globin— the poster child for studies of gene regulation and, recently, nuclear and chromatin organization— illustrates many of the principles I have just discussed. The beta-globin protein combines with alpha-globin to make hemoglobin, the oxygen-carrying molecule of red blood cells. Regulation of alpha- and beta-globin production is finely controlled: imbalances between the two protein levels lead to disease[19]. The beta-globin locus consists of a string of duplicated genes encoding beta subunits used at different times during development, and a locus control region (LCR) containing regulatory elements that control the properly timed expression of those genes[19]. In erythroid progenitor cells, the locus is silent and localized at the nuclear periphery. During differentiation to red blood cells, the locus begins to express one of the beta-globin genes and moves to the nuclear interior, where it associates with a transcription factory[20]. The beta-globin gene can still be expressed when it is located at the nuclear periphery, but its LCR-dependent movement into the nuclear interior is required for high-level expression[20,21]. At the chromatin level, the LCR contains an enhancer element that is brought into contact with the promoter of the activated beta-globin gene while looping out the intervening DNA, a configuration that results in high-level expression of the gene in a transcription factory[4,22]. Expression is enhanced so much that 80% of beta-globin alleles in a population of progenitors are engaged in productive transcription at a

given time[16]. Many of the lessons learned about the beta-globin locus have been applicable at other loci, even though the globin locus is developmentally regulated and has a complex structure not shared by all gene loci.

The residence of the expressed beta-globin gene in a transcription factory presents another emerging principle of nuclear organization: the co-localization of active genes. In erythroid nuclei, the beta-globin locus can be found juxtaposed with other active genes, including those that are coordinately regulated with it to carry out hemoglobin biosynthesis. This co-localization can include genes that are far away on the same chromosome, or even on different chromosomes[16,23]. Since the number of transcription factories present in a nucleus is lower than the number of active genes, it is not entirely surprising that active genes can share the same factory[16]. The presence of co-regulated genes in the same factory has been proposed to be beneficial, since it creates a "micro-environment" enriched in the specific transcription factors and general transcriptional machinery necessary for the high-level expression of those genes[2]. But what of inactive genes? Do these cluster as well? Here again, studies of beta-globin in a tissue where it is not expressed (brain) find that the locus adopts a different configuration than that found in erythroid cells and associates with other inactive genes on the same chromosome[24].

I have painted a picture of the interphase nucleus as an organized and dynamic organelle where the genome is functionally compartmentalized. At the chromatin level, chemical modifications of histones segment the genome into regions that are actively expressed, poised for future expression, or repressed altogether. Enhancer elements bind tissue-specific transcription factors to promote the expression of their target genes, looping over large distances to contact the genes' promoters and bringing them into contact with transcription factories. Insulator elements, bound by the CTCF protein, stand guard between different chromatin domains, blocking the spread of repressive histone modifications and preventing enhancers from contacting the wrong genes. Active genes come into close quarters with other active genes (located on the same or different chromosomes), be they tissue-specific or broadly expressed. At the same time, inactive genes have little contact with active genes, a separation that can be achieved by parking inactive loci at the nuclear periphery replete in repressive proteins. Genes can change their position during activation, either as the result of signals sent into the nucleus by the rest of the cell, or during the ordered process of differentiation that forms mature cell types. Since different cell types express different genes, and gene activation is reflected at the level of nuclear organization, the arrangement of the genome necessarily differs between cell types[2,3]. It also differs between organisms; for example, in

some *Drosophila* cell types, chromosomes are arranged in the Rabl configuration, with their centromeres clustered together and chromosome arms pointed away[25]. Finally, since nuclear organization is dynamic, and likely has a stochastic, random component, it is not stereotyped: no two nuclei have exactly the same organization. Rather, functional constraints on chromosome territories and active processes such as gene transcription serve to spatially organize the genome. In a population of cells, a given gene may be actively expressed in only a subset of nuclei, and the chromosome territory in which it resides will not be in the same location in every nucleus.

How did this image of nuclear organization emerge? In the next section I detail the early techniques that were used to discern details of chromosome structure and organization in the interphase nucleus. Following that, in Section 1.3, I describe a wave of relatively new techniques that have enabled the study of nuclear adjacencies between regions of the genome in a high-throughput manner; these studies have both broadened the conclusions generated by early techniques and provided new details about how two meters of DNA operates in a crowded nucleus to influence the fates of cells.

1.2 Early methods for studying nuclear organization (strengths, limitations and insight)

Biologists examining cells under early microscopes could readily observe the condensation of the chromosomes that occurs during the cell cycle and replicated chromosomes being pulled apart and segregated into two daughter cells. Yet when the cells entered interphase, the period between divisions, chromosomes de-condensed and adopted a new and mostly unobservable configuration. The chromosome-territory theory of genome organization emerged at the turn of the 20th century from observations of cell divisions in the early embryos of the horse roundworm, but was dismissed when electron microcopy studies failed to distinguish individual chromosomes in interphase nuclei[9]. More concrete evidence arose from an early experiment that directing focused radiation at small areas of an interphase nucleus, let the cell continue to metaphase, and examined the condensed chromosomes for signs of damage[9]. If the de-condensed chromosomes had been mingled together randomly, like noodles in a bowl of Vietnamese soup, then the radiation focused on a small spot would have cut through the DNA at sites on different chromosomes, and the subsequent repairs would have yielded a preponderance of inter-chromosomal rearrangements (i.e., translocations). If, on the other hand, chromosomes occupied discrete sub-volumes of the nucleoplasm, a given cell would most likely receive damage at sites on the same chromosome, and rearrangements would predominantly be intra-chromosomal. The latter result was found.

The advent of fluorescently-labeled DNA probes (fluorescence *in situ* hybridization, FISH) allowed the direct observation of vivid CTs in interphase nuclei treated to preserve their three-dimensional structure[9]. FISH experiments involve the preparation of labeled probes complementary to a region of interest (be it an individual locus or an entire chromosome) and the measurement of distances between the resulting fluorescent signal and nuclear landmarks, or other labeled genomic regions. In this manner, one can examine the positioning of a locus with respect to its CT or to a nuclear body, the radial positions of CTs, or the folding of a chromosome segment using alternatively colored probes. The characterization of nuclear proteins can be carried out using fluorescently-labeled antibodies (immunohistochemistry, IHC), a technique that can be combined with FISH to describe many aspects of nuclear structure and organization.

To date, FISH, in its many modified forms, has been the primary tool for dissecting nuclear organization. Much of the large-scale details I gave above about the beta-globin locus were determined using FISH, including its positioning relative to its own CT and to the nuclear lamina[20]. Immuno-FISH was used to characterize the co-localization of the locus with other genes at transcription factories, and RNA-FISH provided a means of identifying the proportion of active globin alleles in a population of nuclei[16,23]. Descriptions of the polymer-like behavior of chromosomes achieved with FISH made it possible to detect chromosomal aberrations (such as translocations) in interphase nuclei, whereas previously such aberrations could only be defined by examining condensed metaphase chromosomes[26].

FISH and IHC are subject to the inherent limitations of light microscopy, which cannot tell the difference between two signals that are < 100 nm apart (although advances in super-resolution fluorescence microscopy have pushed this resolution down to 10-100 nm[27]). They are also "snapshot" assays, displaying a static image of nuclear organization because they involve the fixation of cells, and are low-throughput due to of limitations in the number of distinct colors (and thus loci) that can be discerned at one time. Three-dimensional analyses with these methods are also extremely time-consuming. However, both methods have the advantage of illuminating the structural configuration of a single nucleus.

1.3 3C-based methods to study nuclear organization (strengths, limitations and insight)

The resolution barrier of FISH experiments can be broken using molecular techniques that assay the proximity of genomic sequences in nuclear space; one such technique is Chromosome Conformation Capture (3C). 3C measures how frequently two loci are in close physical proximity in a

population of nuclei at a single time-point[28,29]. It does so by fixing proteins within cells with formaldehyde to preserve the arrangement of the genome, digesting the DNA with a restriction enzyme and ligating together the ends of fragments held together in crosslinked complexes. This process creates a library of chimeric DNA molecules representing sequences that may be far apart in linear sequence (even on different chromosomes), but were close enough in the nucleus to be crosslinked together at the time of fixation. This juxtaposition of loci is closer than is resolvable by FISH. The library is then interrogated using PCR primers designed to amplify across the ligation junction, with one primer in a "bait" fragment and the other in a "prey" fragment it has captured.

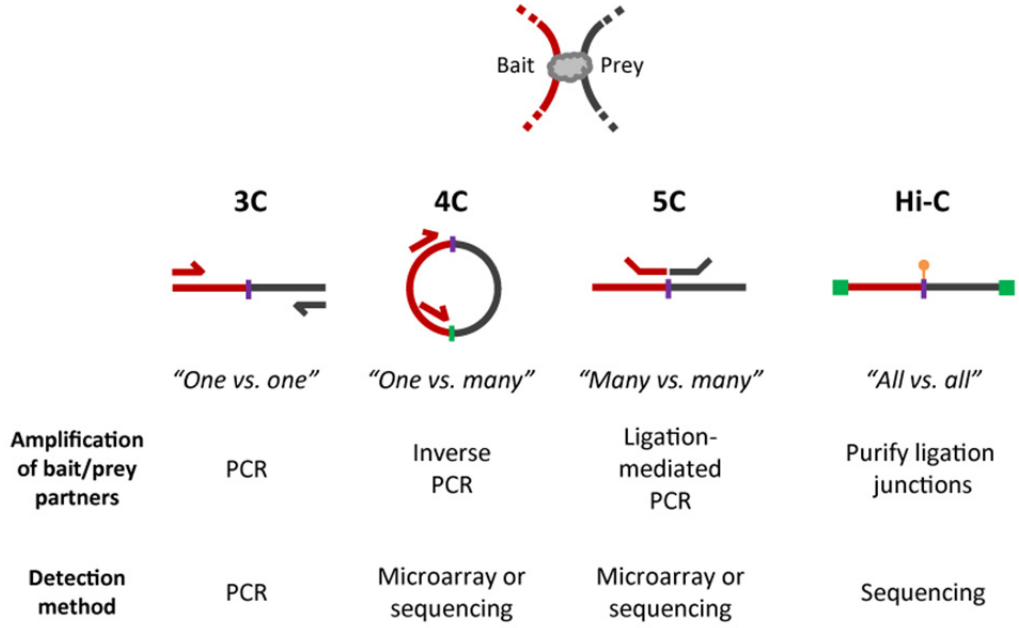

Figure 1.2 Comparison of 3C-based methods
3C and its derivative methods differ in the number of bait/prey interactions that can be identified, as well as how they amplify bait/prey ligation junctions and detect the amplified prey fragments, as described in the main text. "Sequencing" refers to high-throughput sequencing for 5C and Hi-C; for 4C either high-throughput or traditional Sanger sequencing can be used.

Because 3C is a low-throughput assay and requires a prior hypothesis about which loci might be interacting with one another, it has spawned many adaptations that are able to detect many more interactions in a single experiment[30] (**Figure 1.2**). Circular Chromosome Conformation Capture (4C, also known as "3C-on-chip"), which processes the 3C library to produce closed circles, uses inverse PCR primers in the bait fragment to amplify all prey captured by it, and identifies prey by hybridizing them to microarrays[24,31]. 3C carbon copy (5C) uses ligation-mediated PCR to amplify and then sequence hundreds of primer pairs that recognize different restriction fragments[32]. Hi-C purifies the ligation junctions of the 3C library and uses paired-end, high-throughput sequencing to

identify the fragments on both sides of the junction[33,34]. Other techniques combine the 3C philosophy with chromatin immunoprecipitation (ChIP, a technique that identifies genomic binding sites for proteins of interest) to answer questions about what sequences are brought into contact by proteins of interest. ChIA-PET and "enhanced 4C" start with ChIP using an antibody against the protein of interest and then process the DNA bound to the protein to produce ligation products between fragments that may be from different regions of the genome[23,35].

3C was initially developed in yeast, but has since (along with its derivative techniques) been used widely to study chromosomal interactions in cells from mammals and *Drosophila*, including dissections of the long-range interactions of the beta-globin locus discussed in Section 1.1. The use of 3C-based techniques has improved our understanding of how the genome folds and moves within the nucleus to carry out its functions, giving added dimensions to the linear maps of genes and regulatory elements that have been produced by so many genomic studies. Two themes emerge: functional elements interact with one another to control gene expression on a local scale, while at a larger scale, the genome is compartmentalized into active and inactive regions.

Genes that are vital to the specification of cell types have complex regulatory inputs that govern the timing and place of their expression. These inputs converge at enhancer elements, like the one described in the beta-globin LCR, which are often located at great distances (hundreds or even over a million bases) from the promoters they control[36]. Thanks to 3C-based methods, it is now widely accepted that these enhancers physically contact the promoters they control, looping out the intervening sequence[30,36,37]. This phenomenon has helped identify candidate genes for various human diseases by determining that sequence variants implicated through genome-wide association studies can lie in regulatory elements that control the expression of far-away genes[37,38]. CTCF (a zinc-finger DNA-binding protein) is one protein that plays a vital architectural role in mediating regulatory element interactions[35,39-41]. CTCF binds to insulator sequences, which function to demarcate active and repressed chromatin domains and block enhancer-promoter communication[40]. Insulators can associate with each other or be tethered to the nuclear lamina. CTCF also helps mediate enhancer-promoter contacts[40].

The active and inactive partitioning of the genome has been a theme of nuclear organization for decades, harkening back to electron microscopy studies of interphase nuclei that described condensed regions of "heterochromatin" that were distinct from de-condensed regions of "euchromatin". These distinctions were refined by chromatin studies that mapped histone modifications associated with active and inactive expression states across the genome to define the

euchromatic and heterochromatic regions. 4C and Hi-C studies have now shown that this partitioning is reflected in the contacts made by specific regions of the genome, with promiscuous interactions between active regions and more limited contact among inactive regions, and little contact between active and inactive regions[24,33,34]. The original Hi-C study[33] confirmed FISH results about the organization of the genome into chromosome territories (since sites on the same chromosome interact much more frequently than with sites on other chromosomes) and the properties of the chromatin fiber (interactions between two loci decay with increasing distance between them). Because so many sites were surveyed in this study, investigators using Hi-C were also able to describe a general segregation of active/inactive regions that is superimposed upon a simple polymer-like model for organization within chromosome territories.

Although 3C-based techniques have added to our understanding of interphase nuclear organization, they are limited to painting average pictures of genome structure because they sample interactions in a population of cells. The interaction frequencies determined by these methods do correlate with physical distance as determined by FISH, but interactions typically occur in only 10% of nuclei[24,30]. As with FISH and IHC, 3C-based methods are also "snapshot" assays that do not reveal the dynamics of nuclear organization, which requires live-cell imaging. One further shortcoming is that 3C-based assays are not inherently allele-specific, as the interactions of both bait alleles are averaged over the entire population of nuclei used to generate ligation libraries.

1.4 FSHD: an opportunity to examine nuclear organization in a disease state

Facioscapulohumeral dystrophy (FSHD), the third most common muscular dystrophy, provides an opportunity to examine the potential involvement of altered nuclear organization in a disease state. FSHD affects muscles of the face, shoulder and upper arm, with a typical onset in the second decade of life[42]. Despite being subject of some of the earliest genetic linkage studies, a firm candidate gene involved in the disease was not uncovered until very recently[42-45]. The genetic lesion used to molecularly diagnose the disease is a deletion in the D4Z4 macrosatellite repeat array in the subtelomere of the long arm of chromosome 4 (at 4q35): healthy individuals carry 11-100 copies of the repeat unit, but deletions that make the array smaller than 11 units lead to FSHD (**Figure 1.3**). Subtelomeres contain large blocks of duplicated sequence that are shared between non-homologous chromosomes, and the FSHD locus is no exception: the locus, extending from an inverted and truncated copy of D4Z4 to the telomere, is duplicated at the 10q subtelomere, and D4Z4 arrays are present on the short arms of all the acrocentric chromosomes (and adjacent to the

centromeric heterochromatin of chromosome 1)[46-48]. However, D4Z4 deletions are only pathogenic on certain haplotypes of the 4q subtelomere, defined by structural variants just distal of the repeat array and a short sequence polymorphism (SSLP) four kilobases proximal of the array[49].

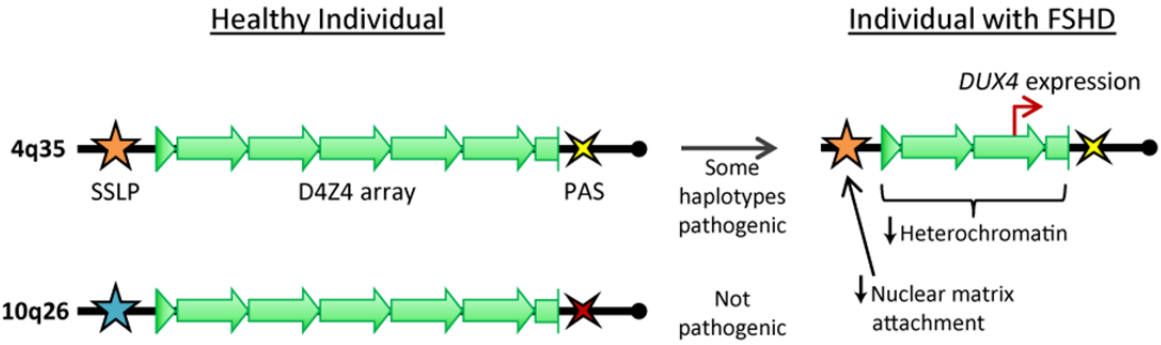

Figure 1.3 Model for FSHD etiology
The FSHD locus is found in the 4q subtelomeric region and is duplicated in the 10q subtelomere. Sequence variants (SSLP, short sequence length polymorphism; PAS, polyadenylation signal) define haplotypes of the locus and influence pathogenicity. Contractions of the D4Z4 macrosatllite array on certain sequence backgrounds lead to FSHD as described in the main text, due to altered chromatin structure that allows the *DUX4* gene within D4Z4 to be expressed.

Each D4Z4 unit contains a copy of a gene called *DUX4*, a double-homeobox transcription factor[43]. Array deletions remove integral numbers of D4Z4 units, leaving the *DUX4* reading frame intact. *DUX4* is referred to as a "retrogene", since its ancestral, intron-containing copy was retrotransposed to produce *DUX4*. Since *DUX4* expression was undetectable for many years, it was branded a "pseudogene" and ignored as a candidate for FSHD pathogenesis. During this time, various models were proposed to explain the etiology of the disease[50]. Did D4Z4 deletions affect the expression of other genes nearby, either through spreading of chromatin modifications or a looping mechanism? Was the binding of some protein factor to D4Z4 disturbed when there were too few repeat units? Studies found that cultured muscle cells from patients were overly sensitive to oxidative stress and defective in myogenesis[51] (the process of muscle cell differentiation), but still no clear candidate gene emerged. One group reported a distance-dependent increase in expression of 4q35 genes in FSHD, but these results could not be confirmed by others[43,52].

The idea that epigenetic mechanisms were at play, initially suggested by the determination that pathogenic D4Z4 deletions left at least one intact copy of *DUX4,* gained further traction when researchers described a loss of DNA methylation in 4q D4Z4 in FSHD patients[53-55]. This study also found a loss of methylation in 4q and 10q D4Z4 units in rare FSHD cases that lacked D4Z4 deletions but had the same phenotype as individuals with deletions (this form of the disease is now called FSHD2). The notion of a possible epigenetic cause led to numerous questions, including: Does

nuclear organization play a role in FSHD? Is the 4q subtelomere localized in a euchromatic or heterochromatic compartment in the nucleus, and does this position change in FSHD? In order to address these questions, two groups used FISH to study the position of 4q35 in the interphase nucleus and reported similar results[56,57]. Tam *et al.* found that the 4q subtelomere frequently localizes within the heterochromatic compartment near the nuclear periphery and adjacent to the nucleolus, with a preference for the periphery, while the centromere and 4p subtelomere had an internal localization. The positioning of the FSHD locus at the nuclear periphery was seen in multiple cell types, including lymphoblasts, fibroblasts, myoblasts and even a somatic cell hybrid line containing human chromosome 4. Both groups used 4q35 and D4Z4 probes, with the dimmer D4Z4 signal inferred to identify the chromosome carrying a pathogenic deletion, to compare the positions of homologous copies of the FSHD locus in muscle cell nuclei. They found no significant difference in position between the deleted and non-deleted alleles; both remained highly peripheral.

The epigenetic angle to FSHD pathogenesis continued to gather steam when it was found that D4Z4 arrays in FSHD patients with and without D4Z4 deletions shared a loss of repressive histone marks[58]. Combined with the description of the SSLP as an enhancer-blocking insulator that dissociated from the nuclear matrix in FSHD cells[59,60], these results suggested that D4Z4 deletions changed the local chromatin structure at 4q35, even if the locus was not grossly mis-localized in FSHD cells. Did de-repression of the D4Z4 array lead to expression of the *DUX4* retrogene? If so, why is it so hard to detect? Dedicated researchers soon uncovered evidence for *DUX4* expression, but only with two rounds of PCR, indicating that transcripts were very rare[61,62].

Strong evidence for the involvement of *DUX4* in the etiology of FSHD came with the identification of sequence variants distal to the D4Z4 array that altered a non-canonical poly-adenylation site (PAS). These sequence variants explained why D4Z4 deletions are only pathogenic in sequence contexts that contain a “permissive” PAS that allows stabilization of the *DUX4* transcript from the final repeat unit of the contracted D4Z4 array[44]. All FSHD1 patients carry this permissive PAS on the same chromosome as the deleted D4Z4 array[44,45]. Furthermore, all FSHD2 individuals (who don’t have D4Z4 deletions, but have de-repressed repeat arrays) carry at least one chromosome 4 with this permissive PAS, providing additional support for the involvement of *DUX4* in FSHD. *DUX4* appears to be normally expressed in the germline and its forced expression in cultured muscle cells activates germline genes (suggesting a possible adaptive immune response in patient muscle) and leads to apoptosis[62,63].

1.5 Bringing a high-throughput study of nuclear organization to bear on FSHD

I was introduced to the FSHD enigma when the role of DUX4 was still unknown. With the FISH studies showing no giant change in 4q35 localization in FSHD, but the chromatin studies indicating that local changes occur as a result of the D4Z4 deletion, I saw an opportunity to re-examine nuclear organization in FSHD using a new technique. I hypothesized that D4Z4 deletions altered the organization of the FSHD locus, reducing its association with inactive regions of the genome and increasing its association with active regions. I chose to use 4C to test this hypothesis, but to do so successfully I had to adapt the technique to be allele-aware, so that I could compare the organization of a single copy of the FSHD locus between control and FSHD cells. In Chapter 2, I discuss the development of 4C-seq, an allele-aware adaptation of 4C, and in Chapter 3 I show its application to FSHD. In Chapter 4, I summarize my findings and place them in the context of the picture of the nucleus I have developed above, ending with a discussion of the advantages and limitations of 4C-seq and the future directions for studying nuclear organization in a high-throughput manner.

Chapter 2: Development of 4C-seq

This chapter details my development of 4C-seq, an allele-aware version of the 4C assay that uses paired-end, high-throughput sequencing to identify prey fragments captured by a bait fragment and can distinguish alleles of the bait fragment. I first discuss my experimental design to study the nuclear organization in the FSHD locus in an allele-aware manner, and then I detail optimization of 4C conditions and preparation of material for high-throughput sequencing.

2.1 Outline of 4C assay design to study organization of the FSHD locus

I chose the 4C approach to examine the nuclear organization of the FSHD locus, because 4C is the most flexible of the 3C-based assays. 4C allows one to identify the collection of "prey" fragments captured by a "bait" locus without a prior hypothesis about the identity of the prey fragments. My hypothesis that pathogenic D4Z4 array contractions alter the organization of the FSHD locus in three-dimensional nuclear space predicts a change in the population of prey fragments captured by the deleted FSHD locus in FSHD muscle cells when compared to the population captured by the non-deleted locus in control cells. However, the FSHD locus on chromosome 4 is duplicated on chromosome 10, meaning that a bait fragment within the locus is present in four copies in a 4C library. In order to compare the prey fragments captured by a single locus between control and FSHD cells, I needed a bait fragment that could not only distinguish the 4q FSHD locus from the 10q copies, but also the 4q homologues from each other. I also needed a way to distinguish copies of that bait fragment.

Fortunately, the short sequence length polymorphism (SSLP) that defines haplotypes of the FSHD locus presents a solution to this problem. Other 4C studies have examined single-copy loci and ignored the sequence of the bait fragment, leaving them unable to resolve allele-specific interactions[24,31,64,65]. In my case, inclusion of the SSLP sequence in the bait fragment offers a means to resolve the prey fragments being captured by each copy of the FSHD locus, provided that cell lines with distinct SSLP genotypes are used and the sequence of each bait fragment is determined along with that of its captured prey fragment.

My 4C strategy is straightforward on its face: obtain myoblast cell lines from control and FSHD individuals with distinguishable SSLP genotypes, make 4C libraries from them, interrogate the libraries for prey fragments captured by the SSLP, and compare the prey fragment populations captured by a deleted and non-deleted FSHD locus (**Figure 2.1**). However, I overcame significant

challenges along the way to accomplishing this end-goal, from the setup of the 4C assay itself, to the means of exploiting SSLP sequence differences, to the preparation of material for high-throughput sequencing and analysis of the subsequent deluge of data. The development of my experimental approach followed an evolutionary rather than linear path, but for the sake of clarity I present it below in a linear manner.

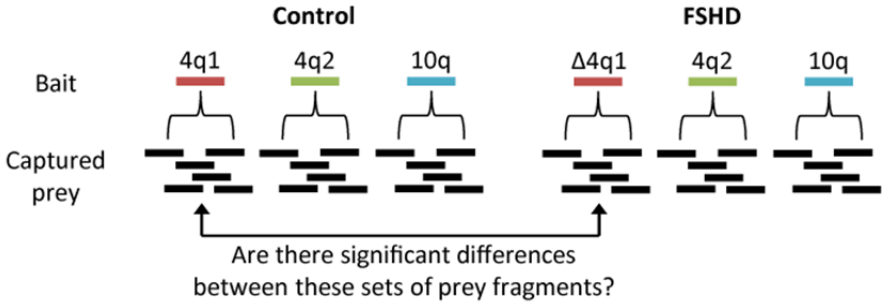

Figure 2.1 Outline of experimental approach
By performing 4C and using a bait fragment (colored lines) with polymorphisms that distinguish copies of the FSHD locus from one another, I endeavored to test for differences in the prey fragments (black lines) that were captured by the FSHD locus with and without a D4Z4 deletion.

The 4C assay exists in a handful of variations developed by independent laboratories[24,31,66]. After choosing a particular 4C protocol[24], I optimized the assay conditions for my experimental setup (Section 2.2) and tested ways to utilize the SSLP to make 4C an allele/paralog-specific assay (Section 2.3). During library preparation and testing, I used Sanger sequencing of cloned SSLP/prey PCR products from numerous 4C libraries to confirm their expected structure and the ability of the SSLP to resolve bait fragment haplotypes (Section 2.4). I then developed a means of sequencing the inverse PCR products that are the output of the 4C assay using the Illumina platform, allowing me to deeply sequence the prey fragments and to genotype the bait fragments that had captured them (Section 2.5).

2.2 Optimization of 4C conditions

4C determines the adjacencies of regions of the genome by crosslinking chromatin within the nuclei of a population of cells using formaldehyde. These crosslinks hold the genomic DNA in its original configuration while it is cut into pieces using restriction enzymes. Subsequent steps serve to join together fragments that are held to one another by crosslinked proteins, thus creating chimeric DNA molecules of sequences that may be far apart in linear genomic sequence—on separate chromosomes, even— but were physically close in three-dimensional nuclear space (**Figure 2.2A**).

Isolated nuclei from the crosslinked cells are digested with a primary restriction enzyme and solubilized to release the DNA/protein complexes, which are then treated with DNA ligase under very dilute conditions to favor intra-complex ligation. (This dilution step is necessary to reduce false-positives in the assay, since ligation between fragments contained in different complexes would make it appear as if they were crosslinked together in the original population of nuclei.) Crosslinks are then reversed, and a secondary restriction digestion is performed to reduce the size of the ligated DNA molecules. A second ligation step creates closed circles of DNA, which after purification constitute the 4C library.

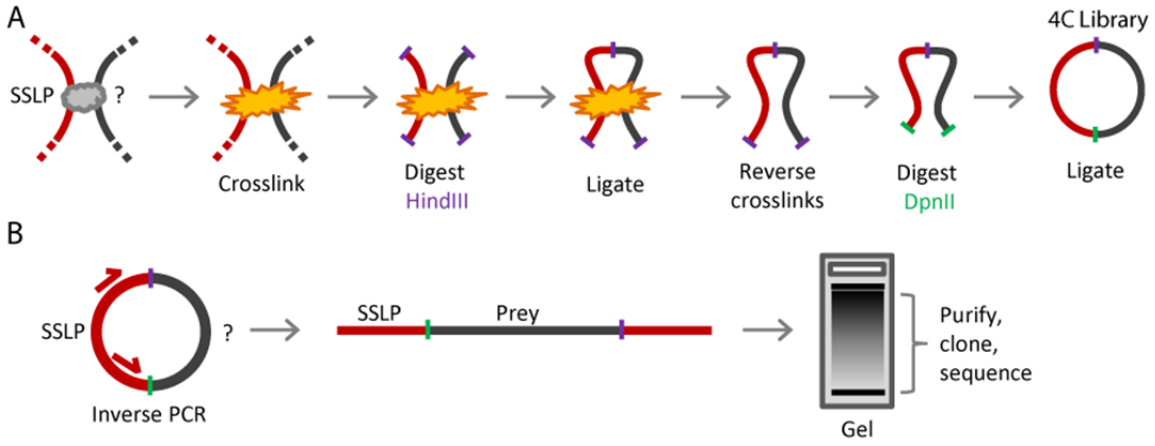

Figure 2.2 4C assay details
(A) Outline of steps for creating a 4C library. The assay creates circular DNA molecules composed of restriction fragments that were in close physical proximity in three-dimensional nuclear space. The library contains a population of such molecules, which can be interrogated to determine which regions of the genome were captured by a fragment of interest and thus were near it in at least one member of the population of crosslinked nuclei. **(B)** Interrogation of the 4C library. Inverse primers recognizing a fragment of interest ("bait", in red, the SSLP in my experiment) amplify all "prey" (black) fragments captured by it. These PCR products are then gel-purified and sequenced to discover the identity of the prey fragments.

The 4C library is interrogated by selecting a particular restriction fragment as bait (**Figure 2.2B**). PCR using inverse primers located in the bait fragment (which face away from each other in the linear genome) amplifies all of the prey fragments that were captured by the bait in the assay. When these PCR products are resolved on an agarose gel and stained with ethidium bromide, they typically form a smear bounded by two prominent bands composed of the most frequent ligation partners: the bait's immediate neighbors. In my experiment, incomplete cutting of the HindIII site on the 5' end of the bait would lead to the "capture" of the SSLP's 5' HindIII-DpnII neighbor at the end of 4C library creation (**Figure 2.3**). And if the ligation step after HindIII digestion leads the full HindIII bait fragment to self-ligate, this would lead to the "capture" of the SSLP's 3' HindIII-DpnII neighbor

(**Figure 2.3**). I reduced the presence of these two bait-neighboring fragments in my sequencing libraries by processing only the fragments lying between them on the gel. Cloning and sequencing of products in this "smear" identifies regions of the genome that were close enough in nuclear space to be crosslinked in the same complex with the bait fragment.

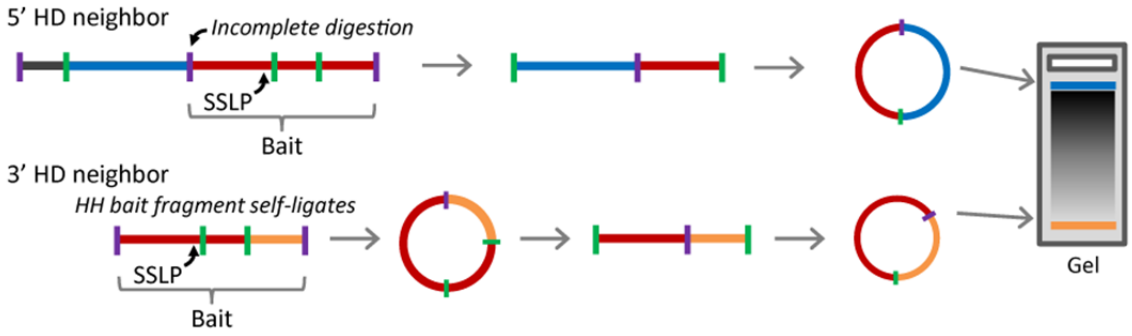

Figure 2.3 Genesis of prominent inverse PCR products
Two fragments typically predominate the prey fragments amplified from the 4C library by inverse PCR, forming prominent bands when the products are run on a gel. In the case of my bait fragment, the larger of the two bands comes from the bait's 5' HindIII-DpnII neighbor (blue line), which can remain attached to the bait if the HindIII site (purple vertical line) between them is incompletely digested. When the HindIII-HindIII bait fragment self-ligates, the DpnII digestion (green vertical lines) and subsequent ligation result in the "capture" of the bait's 3' HindIII-DpnII neighbor (orange line).

Selection of restriction enzymes is one of the most important aspects in the design of 3C-based assays. Some restriction enzymes do not function in the detergents (SDS and Triton) used in the assays, and those that do must be evaluated to determine how large they make the bait fragment, since ligation efficiency can be influenced by fragment size[29]. I decided to use HindIII and DpnII as my primary and secondary restriction enzymes, respectively. HindIII is one of the most commonly used enzymes in 3C-based assays, and together with DpnII, it creates a bait fragment that places the ligation junction as close to the SSLP sequence as possible. HindIII has a 6-bp recognition sequence (AAGCTT), cutting the genome into 4-kb fragments, on average, while DpnII has a 4-bp recognition sequence (GATC).

I also considered a modification of the 4C approach that uses only a single restriction enzyme, but found that it did not produce the same complex mixture of prey fragments as the "double-cutter" method (data not shown), which led me to adopt the latter approach. After testing the "double-cutter" method on fibroblast cell lines, I switched to myoblast lines. Inverse PCR products from my first pair of control and FSHD 4C libraries in myoblasts displayed a large band in the control sample that was absent in the FSHD sample (data not shown); however, this result did not replicate in the next set of 4C libraries (data not shown). Alterations in fixation conditions and

the differentiation state of the cells also did not replicate this result (data not shown). These efforts led me to settle on a single set of conditions for the assay that were used for all subsequent 4C libraries (see protocol in Appendix A).

2.3 Achieving bait-haplotype specificity

My ultimate goal with the 4C assay was to be able to compare the prey fragments captured by specific copies of a bait fragment in the FSHD locus, which exists in four copies in the genome. I exploited the SSLP that defines haplotypes of the locus for this purpose and considered two approaches for achieving haplotype-specificity:

- **Haplotype-specific primers:** I first tried to design haplotype-specific inverse PCR primers that would recognize specific SSLP genotypes and only amplify the prey fragments captured by those SSLPs. The 3' end of the SSLP sequence contains an 8-nucleotide insertion/deletion that differs between some 4q and 10q SSLPs, so I designed primers that terminated in either 4q- or 10q-specific sequence. I tested the haplotype specificity of these primers by cloning individual SSLPs into the TOPO-TA vector to mimic the circular molecules created by the 4C assay and performing PCRs with a common second primer. None of the primers I tested was entirely specific for a single SSLP: one greatly favored the 4q SSLP; one greatly favored the 10q SSLP; and one recognized both SSLPs, with a preference for the 4q sequence (**Figure 2.4**). In addition, given the SSLP genotypes of the cell lines I used for my study (see **Table 3.1**), the 4q-favoring primer would amplify prey captured by both 4q bait haplotypes in control cells but only one 4q bait haplotype in FSHD cells, complicating my desired comparison between cell types. Therefore, I developed an alternative approach.

- **Paired-end sequencing:** An alternative approach was to use inverse PCR primers that would amplify bait and prey sequences together and then sequence both ends of the product using paired-end sequencing, yielding one read giving the haplotype identity of an SSLP and a paired read giving the identity of the prey fragment that the SSLP had captured. This approach would give information on prey captured by all copies of the bait fragment, but would still allow me to compare prey captured by specific copies of the FSHD locus.

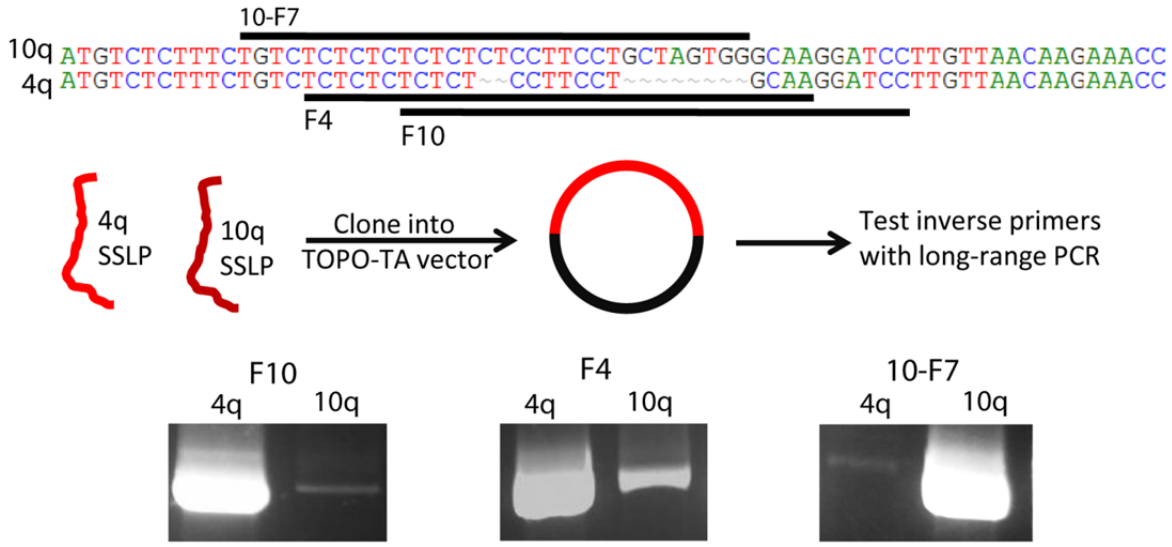

Figure 2.4 Test of haplotype-specific primers
A portion of the SSLP sequence that differs between 4q and 10q haplotypes is shown (top), along with the regions recognized by putative haplotype-specific primers (black bars). These primers were tested on SSLP sequences cloned into a vector to create circular molecules. Representative gels are displayed at the bottom of the figure.

I adopted the paired-end sequencing approach. At the beginning of this project, Illumina sequencing reactions produced only very short reads, typically 36 bp long. By the time I was ready to perform high-throughput sequencing on my 4C libraries, the technology had progressed to the point where high-quality 100-bp reads were being produced, which was long enough to sequence the entire SSLP. In addition, obtaining information on prey fragments captured by all copies of the SSLP provided an internal control to my experiment, as my hypothesis predicted no changes in the prey captured by the SSLP on the chromosomes without the deleted D4Z4 array in FSHD cells.

2.4 Sanger sequencing of inverse PCR products from 4C libraries

During the course of optimizing 4C assay conditions, I cloned and Sanger-sequenced clones of inverse PCR products from the 4C libraries I generated to verify their expected structure. I confirmed that the PCR products contained the SSLP bait fragment connected to prey fragments by a DpnII ligation junction on one side, and a HindIII ligation junction on the other side (**Figure 2.2B** and **Figure 2.5A**). However, some products contained two prey fragments (one DpnII-DpnII [DD] and one HindIII-DpnII [HD]) from different chromosomes (**Figure 2.5B**). The most likely explanation for these products is that DD fragments could ligate between two HD fragments in the formation of

circular DNA molecules during the second ligation step of the 4C assay (**Figure 2.2A**); in these cases, the HD fragment is presumed to be the "true" prey fragment.

Figure 2.5 Sanger sequencing of inverse PCR products
(A) Expected structure of inverse PCR products as confirmed by Sanger sequencing. Purple vertical line, HindIII site; green vertical line, DpnII site; F and R, inverse PCR primers. **(B)** Some products contained DpnII-DpnII fragments (red asterisk) between the SSLP and the expected HindIII-DpnII prey fragment.

The SSLP genotypes of the individuals whose primary myoblasts I used in this study were previously determined (R.J.L.F Lemmers, personal communication). By genotyping the SSLPs of my manually sequenced clones, I confirmed that I could detect bait-prey ligation products from all three SSLP haplotypes in each 4C library (4q SSLPs are distinct in the cell lines I used, but the 10q homologues cannot be distinguished from each other). On average, in a set of analyzed clones from four 4C libraries, 69% of SSLPs matched the expected sequences perfectly, while 28% contained one error and 3% contained two errors. SSLPs with errors typically lost a CA or CT unit from the repeat tracts that make up this complex sequence (which most likely arose due to DNA polymerase slippage during PCR amplification of the 4C library), but were otherwise identical to the reference sequences.

Despite the low depth of sampling in the 4C libraries that were subjected to manual sequencing, a handful of prey fragments were observed in multiple libraries (7/223 sequenced prey in an analysis of clones from eleven 4C libraries). This finding suggested that the 4C assay was reproducibly capturing certain contacts between the SSLP and various regions of the genome, which encouraged me to prepare my samples for high-throughput sequencing.

2.5 Preparation of 4C material for high-throughput sequencing

In order to measure the complexity of the SSLP-captured prey in my 4C libraries and test a method of sequencing-sample preparation, I carried out a pilot high-throughput sequencing run. I used the same 4C library to prepare two samples that differed in the amount of template used in the inverse PCR reaction (25 ng and 100 ng), and pooled two PCRs for each sample. I was only interested in the identities of the prey fragments in this pilot analysis. Because sequencing the ends of the PCR products would have only yielded bait sequences (see **Figure 2.2B**), I blocked their ends by adding non-templated ddNTPs, so that sequencing adapters could not be added to them. I then

gel-purified the products to minimize the presence of the bait's neighbors in the prey-fragment pool (see **Figure 2.3**) and fragmented the purified DNA using a Covaris instrument. The fragments were then prepared for single-end Illumina sequencing by Molly Weaver, a member of John Stamatoyannopoulos's laboratory. The resulting 36-bp reads were mapped to the reference human genome using their lab's automated pipeline. I subsequently filtered these reads and determined to which restriction fragments they mapped in the genome (data not shown).

My analysis of these pilot data confirmed and extended my observations from manually-sequenced inverse PCR products. Approximately 40% of fragments < 1,400 bp and identified with ≥ 4 reads were HD, while approximately 60% were DD. However, of the sequenced fragments shared between the two samples, 64% were HD, highlighting the artifactual nature of DD prey fragments. Put another way, 6% of DD fragments were shared between the samples, compared to 23% of HD fragments. As with the manually sequenced prey fragments, the prey in this dataset were found on every chromosome, but were enriched on the bait-containing chromosomes (data not shown).

I was surprised by the low overlap of HD fragments between the two samples, since the samples were derived from the same 4C library. As expected, fewer fragments were identified in the sample with the lower amount of inverse PCR template (904 in "25 ng" vs. 1,397 in "100 ng"). Almost half of the "25 ng" fragments were found in the "100 ng" dataset, but 69% of the "100 ng" fragments were novel. This observation suggested that I had not sampled the 4C library deeply enough, so I decided on two improvements for future sequencing-sample preparation. First, to increase the number of prey fragments I could observe, I would increase the number of inverse PCR reactions that I pooled together before sequencing library preparation. Second, because I worried that the inverse PCR efficiency might be hampered by competition between annealing of the PCR primers and re-annealing of the denatured 4C library constituents to reform circular molecules, I would linearize the 4C library before carrying out inverse PCR.

The pilot run also highlighted an advantage of paired-end sequencing. Since I only did single-end sequencing, I could not definitively tell which prey fragments were true prey, in the sense that they had been attached to a bait fragment and were not just free-floating restriction fragments in the 4C library that had been sequenced (although the inverse PCR step should bias against such fragments making up a substantial fraction of the DNA that was sequenced). However, with paired-end sequencing, "real" prey reads would be paired with a reads from the bait fragment.

I adapted my paired-end sequencing protocol from a method developed for assembling large contigs from short Illumina sequence reads with the help of Joe Hiatt and Jay Shendure, that

method's authors[67]. My goal was to place the adapters necessary for flowcell attachment and sequencing-primer recognition onto the ends of the inverse PCR products from the 4C library such that the first sequencing read would yield the SSLP sequence, and the second read would yield the prey fragment sequence. I describe the general approach here, with a graphical depiction in **Figure 2.6**, while the details about reagents can be found in Chapter 5 Section 3.

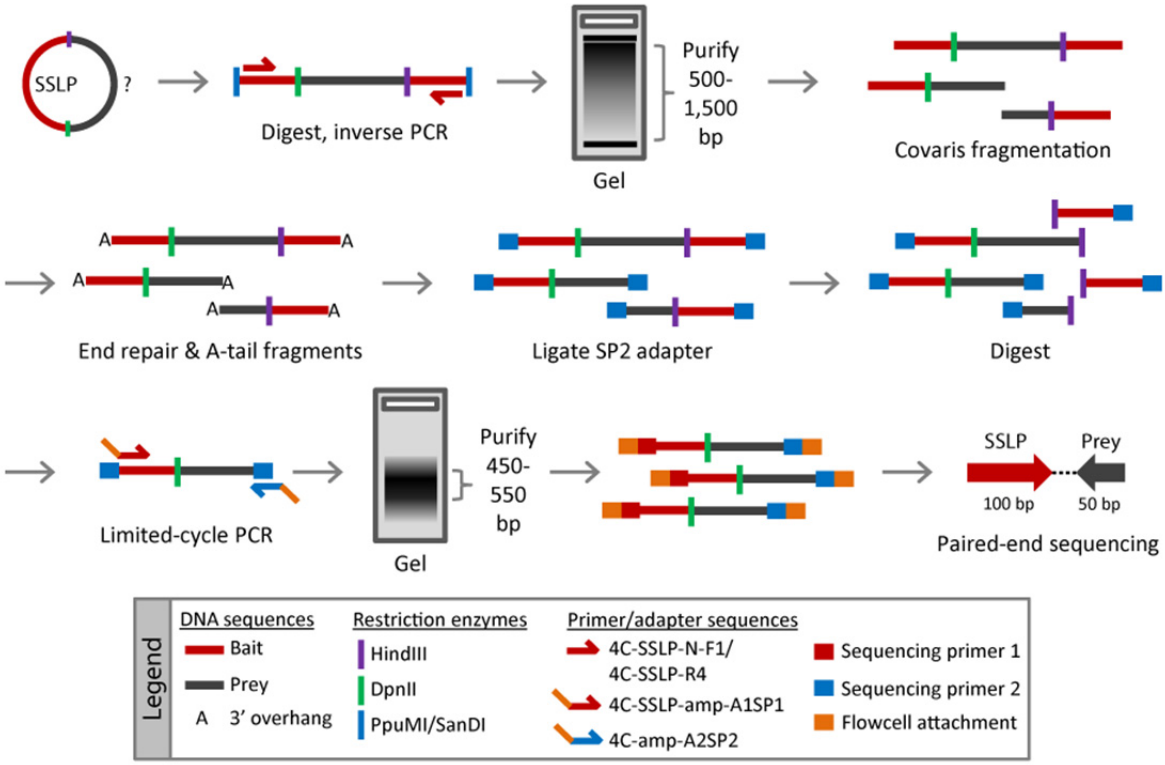

Figure 2.6 Preparing samples for high-throughput sequencing
Inverse PCR products from the 4C library are prepared for high-throughput sequencing by fragmentation, ligation of adapters, and a final PCR step to add a site for the first read's sequencing primer and attachment sequence for the flowcell, as described in the main text. Sequences of primer names listed in the legend can be found in **Table 5.1**.

First, I linearized the 4C library using two restriction enzymes with rare recognition sequences (PpuMI and SanDI) that cut inside the bait fragment. Approximately 10% of HD prey fragments in the reference genome contain these recognition sequences and could be lost, but I accepted this tradeoff for a step that could benefit the efficiency of the inverse PCR, as described above. I carried out and pooled six inverse PCRs per sample and gel-purified the products to minimize the presence of the two most prominent prey fragments (the bait's immediate neighbors). Some of the purified PCR products were larger than the upper size limit of DNA molecules that can be sequenced on an Illumina flowcell. To capture information from these large products, I

fragmented the gel-purified DNA with a Covaris instrument, which uses focused acoustics to randomly shear DNA. This fragmentation meant that the prey read would come from a random location within each prey fragment. I then end-repaired and A-tailed these fragments and ligated the sequencing primer 2 (SP2) adapter to both ends.

In initial tests of this protocol, I found that many of the end products had not been fragmented before SP2 adapter ligation. I wanted to prevent these products from being sequenced, since they would produce prey reads that consist of bait sequence (red line after the HindIII site in **Figure 2.2B**). Therefore, I added a HindIII digestion step after adapter ligation so that un-fragmented PCR products would not participate in sequencing-library construction, since they lacked the sequence for the SP2 primer to recognize during the final PCR step of library construction.

I used PCR to complete the addition of sequencing-primer and flowcell-attachment sequences to the ends of the fragments. The forward primer in this reaction is the first read's sequencing primer, and it starts the read at the beginning of the SSLP sequence; the reverse primer recognizes the SP2 adapter sequence. Both primers contained tails of the attachment sequences that secure the molecules to the flowcell surface. The PCR was carried out for only 18 cycles, since over-amplification would increase the number of reads with identical start sites in the genome, which I intended to filter out during data analysis. I gel-purified the PCR products to select a narrow, 100-bp size range, which is needed for efficient sequencing using the Illumina chemistry.

The paired-end sequencing runs were configured to give 100-bp SSLP reads, and 50-bp prey reads. As I analyzed the data from each flowcell, I optimized my sample preparation protocol to improve read depth and the fraction of informative reads (those remaining after my filters) that came out of subsequent sequencing reactions. Changes to the protocol included introducing the HindIII digestion step (described above) to reduce the number of prey reads mapping to the bait fragment and changing the position of this digestion step within the protocol; increasing the length of the gel run after the inverse PCR to make cutting between the prominent bands easier; and using improved Illumina sequencing chemistry and instrumentation to increase the overall number of reads. The first flowcell sequenced produced an average of ~6 million aligned prey reads per lane, of which 66% were near the bait, 3.7% matched the SSLP sequence, and 10% were used after filtering (see **Table 3.3**). These numbers improved to ~57 million aligned reads, 40% near the bait, 1% matching SSLP sequence, and 17% used after filtering for the final flowcell.

2.6 Conclusion

In this chapter, I have described my 4C experimental design based on use of the SSLP of the FSHD locus as a bait fragment. After testing modified versions of the 4C approach on fibroblasts and myoblasts, I settled on a single method and set of conditions and developed and optimized a protocol to interrogate these libraries using paired-end, high-throughput sequencing. In the next chapter, I detail the results obtained using these protocols to generate and characterize control and FSHD 4C libraries.

Chapter 3: Application of 4C-seq to FSHD

In this chapter, I apply the 4C-seq method developed in Chapter 2 to characterize the nuclear organization of the FSHD locus, with the goal of addressing the hypothesis that D4Z4 contractions alter the organization of this locus, reducing its contact with inactive regions of the genome while increasing its contact with active regions. First, I detail the filters I used to determine the prey fragments captured by the SSLP in my 4C libraries. I then characterize these prey fragments before and after assigning them to bait haplotypes. I describe the chromosomal distribution of the prey fragments and analyze the features of prominent contacts on "bait" chromosomes (4 and 10, where the bait itself resides). On "non-bait" chromosomes, I characterize prey fragments that are repeatedly captured across multiple 4C libraries. Finally, I identify differences between prey found in control and FSHD libraries.

Table 3.1 Primary myoblast cell lines & FSHD locus haplotypes
Haplotypes were determined by Southern Blot (A/B polymorphism) and STR analysis (SSLP) by R. Lemmers (personal communication). Sizes of SSLPs are indicated in subscripted numbers. In the FSHD1 lines, the 4q1 allele carries the pathogenic D4Z4 array contraction (Δ). The notations in parentheses for cell lines and haplotypes are simplified names used in the text to refer to each.

Cell line		SSLP-Defined Haplotype	
		4q	10q
Control	MB-NR-209 (C1)	$4qA_{161}$ (4q1) / $4qB_{163}$ (4q2)	$10qA_{166}$ (10q) / $10qA_{166}$ (10q)
	MB-NR-135 (C2)		
FSHD1	MB-FSHD-197 (F1)	$\Delta 4qA_{161}$ (4q1) / $4qB_{168}$ (4q3)	
	MB-FSHD-219 (F2)		
FSHD2	MB-FSHD-200 (F3)	$4qA_{161}$ (4q1) / $4qB_{168}$ (4q3)	

3.1 Preparation of 4C libraries from control & FSHD myoblast cell lines

I used primary myoblast cell lines from control, FSHD1, and FSHD2 individuals (**Table 3.1**) to generate 4C libraries in duplicate from each cell line as detailed in Chapters 2 & 5. These cell lines were all heterozygous for the 4q SSLP and homozygous for the $10qA_{166}$ SSLP; therefore, I could distinguish chromosome 4 homologues from each other and from chromosome 10, but could not distinguish chromosome 10 homologues from each other. Hereafter, I refer to these cell lines, the 4C libraries generated from them, and their SSLP genotypes using simplified nomenclature (see parenthetical notations in **Table 3.1**). Cell lines are denoted "C" for control or "F" for FSHD and

numbered, with numerals after the period indicating the replicate 4C libraries; SSLP genotypes are simplified to 4q1, 4q2, 4q3 and 10q.

3.2 High-throughput sequencing of 4C libraries

Each 4C library was interrogated using inverse PCR primers within the HindIII-DpnII fragment encompassing the SSLP in order to amplify prey fragments captured by it in the 4C assay. These PCR products were prepared for high-throughput Illumina sequencing as detailed in Chapter 2 Section 5 and Chapter 5 Section 3. Overall, I sequenced 27 samples on four flowcells; 26 samples produced usable data (**Table 3.2**). Some samples were sequenced twice on the same flowcell in order to assess technical reproducibility; in the case of flowcell #4, two samples were sequenced in two lanes each to increase sequence depth.

Table 3.2 Flowcell summary
X, sample sequenced; o, failed sample

4C Library	Flowcell #			
	1	2	3	4
C1.1	X o	X X		
C1.2	X	X		
C2.1	X		X	X X
C2.2		X	X	X
F1.1	X	X	X	
F1.2	X	X		
F2.1	X		X	X X
F2.2		X	X	X
F3.1			X	
F3.2			X	

Development of the 4C-seq assay (as detailed in Chapter 2) was an evolutionary process, during which I refined my method while also collecting data for my study. The unique nature of my Illumina sequencing runs, which generating paired-end reads of unequal length, necessitated filling the entire flowcell with my samples during each sequencing run (seven samples per run) for maximal cost effectiveness. After each run, I adjusted my protocol for sequencing-sample preparation to increase the number of informative prey reads (that is, reads passing the filters described below), and used the refined technique to prepare samples for the next flowcell. The first two flowcells were sequenced using a GAIIx machine, while the last two flowcells were sequenced using a HiSeq-2000 machine. The HiSeq-2000 machine uses different flowcells and sequencing chemistry than the GAIIx machine, further increasing read numbers.

Taking all of my sequencing runs into account, I obtained over 407 million prey reads and their corresponding bait reads, representing the original circular ligation products produced in the 4C assay. My next challenge was to determine which restriction fragments in the genome had been captured as prey (later, in Section 3.4, I will describe how I used the bait reads to determine which haplotype of the bait had captured each prey fragment). I started by using the Burrows-Wheeler Aligner (BWA) algorithm[68] to align prey reads that passed Illumina's default quality filters to the

reference human genome (details in Chapter 5 Section 5). I then used five filters to remove uninformative or spurious reads. Refer to **Table 3.3** for detailed statistics on read numbers for each sequenced sample, and to **Figure 3.1** for a graphical representation of four of these filters.

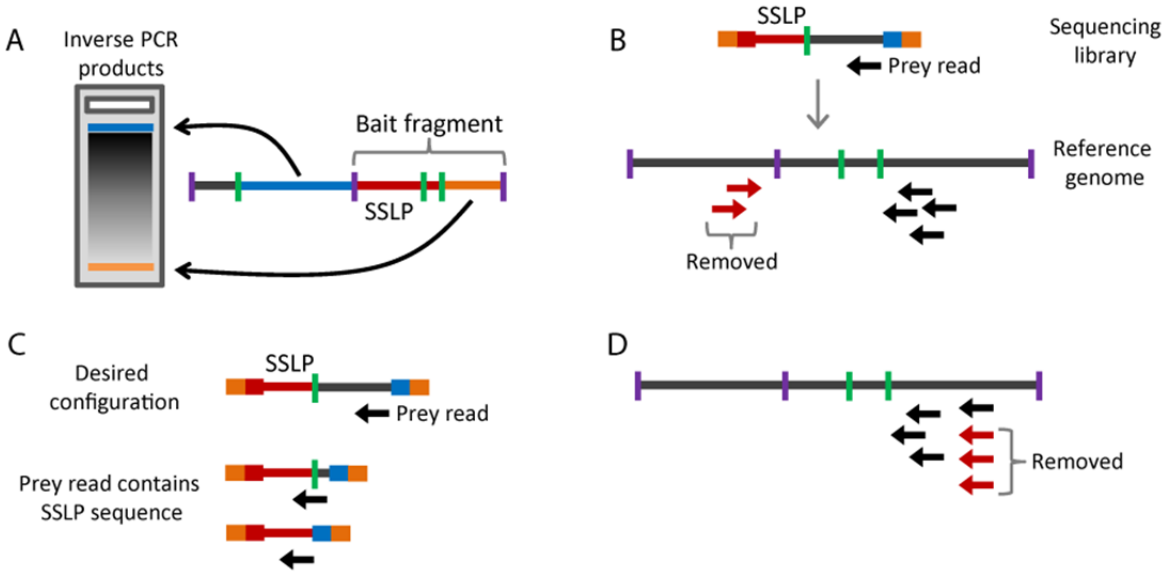

Figure 3.1 Prey read filters

Graphical depiction of prey reads that were removed from the analysis. **(A)** The two HindIII-DpnII fragments flanking the SSLP are its most frequent ligation partners in the 4C assay, forming the two most prominent bands when inverse PCR products are run on a gel (green vertical lines, DpnII sites; purple vertical lines, HindIII sites). Although I selected against these fragments by purifying the products between them on the gel, they still constitute a large proportion of prey reads. **(B)** The sequencing library should only produce prey reads pointing toward DpnII sites. Consequently, reads (in red) pointing toward HindIII sites without an intervening DpnII site in the reference genome were discarded. **(C)** Molecules with the SP2 adapter (blue) placed very close to the ligation junction between the SSLP and prey (or even within the SSLP) produce reads that match SSLP sequence, but can still be aligned to simple repeats elsewhere in the genome, so are filtered out. **(D)** Reads with identical start sites are most likely generated by the final PCR amplification step of sequencing library preparation, so one representative read is kept at each redundant position.

Table 3.3 Summary of prey read filtering

Prey read statistics are presented for each sample sequenced, along with averages for each flowcell. The "Aligned reads" column gives total uniquely-mapped reads; "% Near Bait" gives percentage of reads mapped to the bait fragment or its two flanking fragments (these reads were removed). The next columns give the percentage of reads not near the bait that had low alignment scores ("Map Q < 15"), pointed toward HindIII sites ("Wong Dir."), matched SSLP sequence ("SSLP Match"), or had a redundant start positions ("Red."). Reads with these attributes were removed, giving the "Final Filtered" counts (% values here indicate percentage of total aligned reads kept after filtering).

Flowcell	Lane	Aligned Reads	% Near Bait	% of Reads Not Near Bait With:					
				Map Q < 15	Wrong Dir.	SSLP Match	Red.	Final Filtered Reads	
1	1	4,476,881	68%	36%	0.8%	4.5%	36%	385,446	9%
	3	7,239,357	70%	32%	0.8%	4.2%	50%	370,502	5%
	4	6,885,130	75%	38%	0.8%	6.2%	33%	468,481	7%
	6	7,286,781	68%	28%	0.6%	3.3%	59%	271,135	4%
	7	8,366,630	60%	26%	1.3%	1.9%	35%	1,266,461	15%
	8	2,476,058	53%	24%	1.3%	1.8%	25%	574,662	23%
	avg	6,121,806	66%	31%	0.9%	3.7%	40%	556,115	10%
2	1	10,418,160	61%	14%	0.8%	1.3%	56%	1,184,198	11%
	2	7,559,417	63%	14%	0.8%	1.2%	56%	838,054	11%
	3	13,781,886	70%	15%	0.6%	1.3%	65%	831,050	6%
	4	14,133,093	65%	14%	0.6%	1.3%	67%	867,924	6%
	6	14,190,712	65%	16%	0.7%	1.3%	68%	781,122	6%
	7	11,240,982	49%	15%	1.6%	1.4%	50%	1,913,602	17%
	8	12,456,138	41%	13%	0.8%	1.0%	70%	1,214,373	10%
	avg	11,968,627	59%	14%	0.8%	1.3%	62%	1,090,046	10%
3	1	13,674,782	39%	20%	1.8%	1.6%	24%	4,285,778	31%
	2	17,061,377	38%	19%	1.8%	1.7%	33%	4,668,482	27%
	3	16,090,232	41%	20%	1.6%	1.6%	29%	4,471,646	28%
	4	12,105,519	40%	19%	1.3%	1.7%	37%	2,949,026	24%
	6	19,213,156	55%	21%	0.7%	2.1%	55%	1,767,852	9%
	7	25,583,413	44%	21%	1.3%	1.6%	44%	4,590,426	18%
	8	22,506,968	42%	20%	1.7%	1.6%	38%	4,989,198	22%
	avg	18,033,635	43%	20%	1.5%	1.7%	37%	3,960,344	23%
4	1	54,325,468	40%	20%	2%	1%	45%	10,413,909	19%
	2	49,011,585	40%	20%	2%	1%	43%	9,893,272	20%
	3	51,296,120	39%	20%	2%	1%	49%	9,013,275	18%
	4	52,626,843	39%	19%	2%	1%	49%	9,172,231	17%
	6	62,538,207	41%	19%	2%	2%	48%	10,694,496	17%
	7	76,094,557	40%	19%	1%	2%	64%	6,545,509	9%
	avg	57,648,797	40%	20%	2%	1%	50%	9,288,782	17%

The five prey read filters were:

- **Near bait:** The most prevalent fragments ligated to the SSLP are its immediate 5' and 3' neighbors (**Figure 3.1A**); however, I did not consider these to be true prey fragments, because their "capture" is most likely due to self-ligation of the HindIII-HindIII bait fragment during the first part of the 4C assay (3' neighbor), and incomplete digestion of the 5' HindIII site of the bait (5' neighbor) (see **Figure 2.3**). These fragments are connected to the bait so frequently that they are the most prominent bands seen when the inverse PCR products are run on a gel. While I tried to minimize the presence of these products in my sequencing libraries using a size-selection step after inverse PCR, reads mapping to these near-bait fragments still accounted for up to 66% of aligned prey reads (**Table 3.3**).

- **Alignment quality:** BWA produces an alignment quality score for each read that takes into account both the alignment and the quality of the sequence read. Based on advice from Zizhen Yao, a computational biologist with extensive experience working with high-throughput sequencing data, I chose a score cutoff of 15 and removed reads with a score below this value.

- **Wrong direction:** All prey reads should point towards DpnII sites due to the way I prepared the inverse PCR products from the 4C libraries for high-throughput sequencing (**Figure 3.1B**). Therefore, I added a directionality filter to remove reads that pointed toward HindIII sites without an intervening DpnII site. Reads pointing toward a HindIII site might in fact point toward a DpnII site due to sequence differences between the reference genome and the genomes of the individuals from whom my myoblast lines were derived. Despite this possibility, I filtered out reads pointing toward HindIII sites; doing so removed up to 2% of prey reads (**Table 3.3**). Sequencing each individual's genome might have revealed such polymorphic DpnII sites and rescued some of these prey reads, but such an effort was beyond the scope of my project and not worthwhile, given the small loss of prey reads from this filter.

- **Read matches SSLP:** I observed many positions in the genome to which prey reads aligned across all samples. Upon manual inspection, I found that sequences at these

locations contained CA and CT repeat tracts, similar to the SSLP sequence, leading me to believe that prey reads mapped to these locations were in fact derived from the bait fragment. These reads most likely resulted from Covaris fragmentation that led to placement of Illumina adapters very close to the bait/prey ligation junction, thereby allowing the prey sequencing read to extend into the bait (**Figure 3.1C**). Thus, I instituted a filter that scored all aligned prey reads for the presence of SSLP-like sequence and removed those reads that were highly similar to the SSLP (i.e., their scores were higher than the 99th percentile of scores from randomly-generated sequence reads; details in Chapter 5 Section 5).

- **Redundant position:** Due to a final PCR-amplification step in the Illumina sequencing protocol (see **Figure 2.6**), the same molecule of input DNA can be sequenced multiple times, producing reads with identical start positions in the reference genome (**Figure 3.1D**). These redundant reads are uninformative for further analyses, and all but one representative read were therefore removed from my datasets.

All reads that passed these filters were used for further analyses, regardless of whether their paired SSLP reads were successfully genotyped. In order to obtain the deepest possible coverage of prey fragments for each of my 4C libraries, I combined all reads for each library across all flowcells and assigned the reads to restriction fragments (details in Chapter 5 Section 7). Potential prey fragments are bounded by HindIII and DpnII restriction sites, the restriction enzymes used for 4C library construction. On average, 8% of HindIII-DpnII fragments in the reference genome contained at least one prey read (range 3-15%). As with my initial read data, I applied filters to these fragments that reflected parameters of 4C library and sequencing-sample preparation and endeavored to remove potentially artifactual fragments (depicted graphically in **Figure 3.2**). Detailed numbers for the assignment of prey reads to fragments are given in **Table 3.4**. Overall, I retained an average of 58% (range 34-76%) of fragments with any read coverage. I did not take relative read depth of fragments into account in subsequent analyses (except as a cutoff for assigning fragments to bait haplotypes, in Section 3.4), because the two rounds of PCR that take place on a bait/prey ligation product in the 4C library before sequencing on the flowcell could bias read counts.

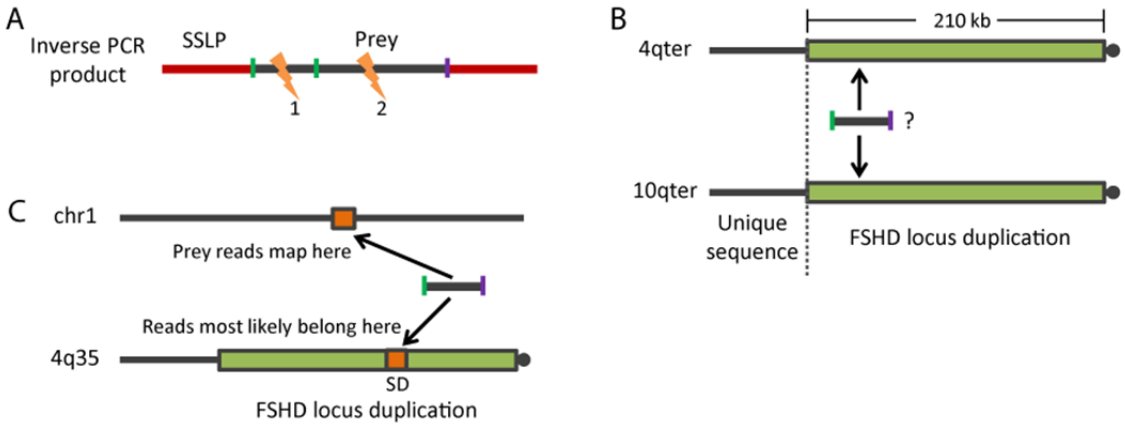

Figure 3.2 Prey fragment filters

Graphical depiction of prey fragments that were removed from the analysis. **(A)** Shearing of inverse PCR products at site 1 produces prey reads that map to DpnII fragments in the genome, while shearing at site 2 produces prey reads that map to HindIII-DpnII fragments, which are considered true prey fragments. **(B)** The sequence similarity between duplicated copies of the FSHD locus means that reads from prey fragments in the region could map to the wrong copy, making it hard to determine the true chromosomal origin of these prey fragments. **(C)** As in panel B, reads in segmental duplications on non-bait chromosomes with copies in the FSHD locus could be incorrectly mapped.

Table 3.4 Summary of prey fragment filtering

Prey fragment statistics are presented for each sample, starting with total number of filtered reads used to identify prey fragments. HindIII/DpnII (HD) restriction fragments containing reads were filtered to remove fragments that had only one read ("1 Read") and that were larger than 1,500 bp ("Size"), within the FSHD locus duplication boundaries ("NB"), or in a segmental duplication on a non-bait chromosome with a copy in the FSHD locus ("In SD").

4C Library	# Prey Reads	HD Frags >=1 Read	# of Frags Removed By Filters:				Final # Frags
			1 Read	Size	NB	In SD	
C1.1	2,407,698	74,151	18,781	332	34	22	54,982
C1.2	1,238,426	37,547	19,622	147	28	10	17,740
C2.1	25,859,420	200,854	70,844	2,015	35	68	127,892
C2.2	17,079,744	216,721	65,818	1,715	34	59	149,095
F1.1	3,067,383	85,596	41,610	284	35	26	43,641
F1.2	1,052,257	29,706	19,484	99	32	15	10,076
F2.1	23,428,650	215,262	80,940	1,765	35	68	132,454
F2.2	10,708,908	137,244	72,418	1,530	35	62	63,199
F3.1	4,590,426	67,944	19,235	513	34	29	48,133
F3.2	4,989,198	91,166	21,363	632	34	26	69,111

The prey fragment filters were:

- **Fragment type:** Prey reads mapped to both HindIII-DpnII (HD) and DpnII-DpnII (DD) restriction fragments (1-5% of all DD fragments) in the reference genome. As discussed above for the “wrong direction” read filter, it is possible that polymorphisms in the

genomes of my myoblast cell lines create restriction sites not in the reference genome, meaning that some DD fragments could actually be HD fragments. However, as discussed in Chapter 2 Section 4, I frequently observed DD fragments between the SSLP and an HD prey fragment in manually-sequenced inverse PCR products (**Figure 3.2A**). These DD fragments were often from a different chromosome than the HD fragments, and were likely to have ligated between two HD fragments during the second ligation step of the 4C assay. Consequently, I only considered HD fragments to be "true" prey fragments.

- **Read depth:** I considered a fragment to have been captured when it contained at least two prey reads, since a single read mapped to a fragment could be aligned incorrectly due to repetitive elements, while multiple reads mapped to the same fragment (and different portions of it) give greater confidence that the fragment had truly been captured in the 4C assay. An average of 40% (range 23-66%) of HD fragments containing prey reads were covered by only one read, and were removed by this filter (**Table 3.4**).

- **Fragment size:** During sequencing-sample preparation, I size-selected the inverse PCR products to minimize the presence of prey immediately adjacent to the bait (see "near bait" read filter, above, and **Figure 3.1A**). The upper limit of this selection was 1,500 bp, so I computationally removed fragments larger than this size. This filter removed 1-2% of fragments with at least two reads (**Table 3.4**).

- **Near bait:** The FSHD locus encompasses the terminal subtelomeric sequence of chromosome 4q, which is duplicated in the terminal subtelomere of 10q. Since these duplicated sequences are > 98% identical[47,69], prey reads from fragments in these regions could align to the wrong copy of the locus (**Figure 3.2B**). Thus, I removed the 28-35 prey-read-containing HD fragments located in the terminal ~210 kb of 4qter and 10qter (this region extends from the proximal boundary of the duplication to the end of the reference sequence; **Table 3.4**). I also removed prey contained within segmental duplications of these regions that were located on non-bait chromosomes (average 39 fragments, range 10-68; **Table 3.4**), reasoning that they could originate from reads mapped to the wrong place in the reference genome.

My filtered prey-fragment dataset contained over 415,000 unique fragments captured by the SSLP bait in the combined set of all ten 4C libraries. Libraries that had been sequenced only during the early phases of my protocol development for sequencing-sample preparation were exhausted in the process, resulting in their having a smaller number of identified prey fragments than libraries sequenced after the protocol was optimized (**Table 3.4**). The libraries sequenced on the final flowcell (C2.1, C2.2, F2.1, F2.2) have the highest coverage (both in terms of read depth and number of prey fragments). To answer some questions, I utilized prey fragments only from my high-coverage libraries, and for other questions I used prey from all libraries. For some analyses, I considered these captured prey fragments in aggregate, without assigning them to a bait haplotype (Sections 3.3 & 3.6). In other cases, I compared sets of bait-haplotype-assigned prey to each other (Sections 3.5-3.7).

3.3 Characterization of prey fragments before haplotype assignment

Prey fragments were distributed on every chromosome, but were highly enriched on the bait chromosomes (**Figure 3.3**). This enrichment was expected, because a locus has a higher probability of contacting sequences tethered to it within its own chromosome territory than sequences on other chromosomes.

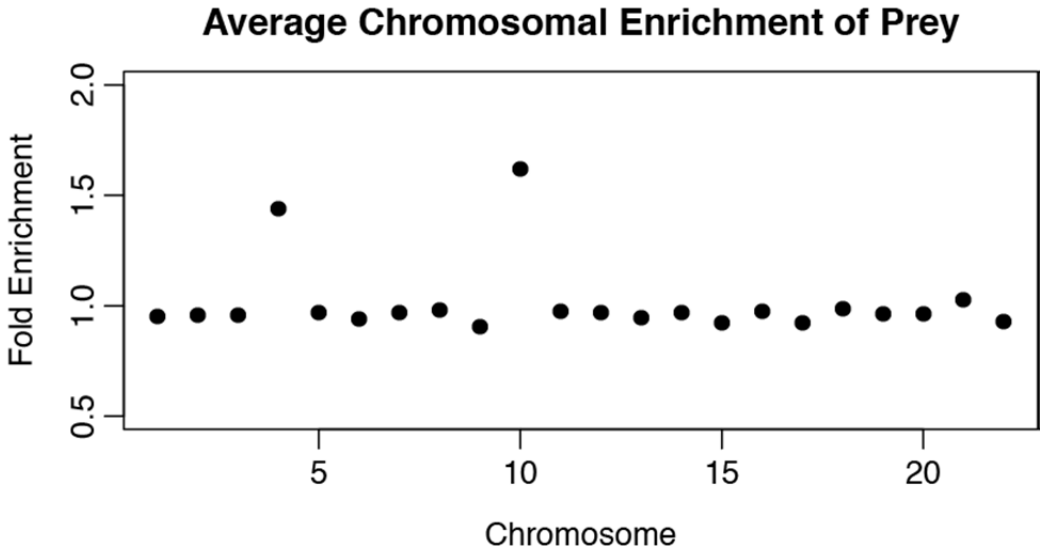

Figure 3.3 Prey fragments are enriched only on bait chromosomes
Fold enrichment of prey was determined by dividing the number observed on each chromosome by the number of total possible prey fragments on that chromosome and was averaged over all ten 4C libraries. The sex chromosomes were excluded, because one library was derived from a female cell line, while the rest were derived from male cell lines. The bait chromosomes (4 and 10) display a high enrichment of prey fragments, while no other chromosome displays enrichment.

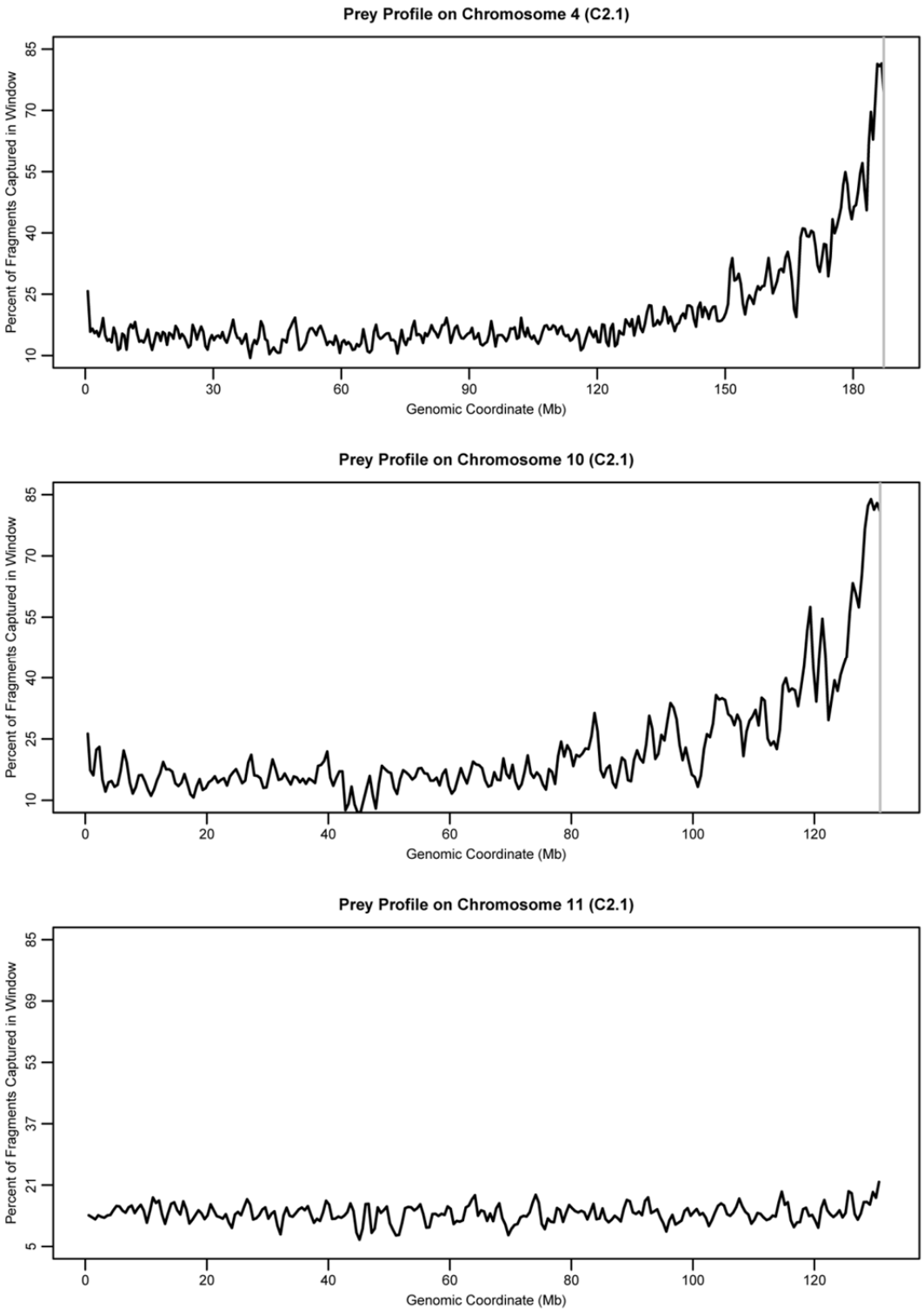

Figure 3.4 Sliding window analysis of prey distribution

The locations of all prey fragments identified in library C2.1 on the bait chromosomes (4 & 10) and chromosome 11 are shown using 1-Mb windows slid in 500-kb steps along the chromosomes. In each window, the number of observed prey is divided by the number of possible prey fragments, giving the percentage of fragments captured (y-axis). The location of the bait fragment is indicated with a grey vertical line.

For a localized perspective on captured prey fragments, I counted the number of observed prey in 1-Mb windows slid in 500-kb steps along the genome and divided each count by the number of potential prey fragments in each window. This analysis revealed that contacts made by the SSLP are enriched only on the same arm of the bait chromosomes as the SSLP, with the density of prey fragments decaying steeply with increasing distance from the bait (**Figure 3.4**). The "peaks" and "valleys" within these profiles represent regions of significantly increased and decreased contact frequency by the SSLP, respectively. Valleys cannot simply be explained by the inability to map reads to certain places in the genome, since some valleys in one profile are peaks in another profile (data not shown). In Section 3.7, I discuss the analysis of gene expression and lamin-associated domain overlap in these peaks and valleys in the context of control-versus-FSHD comparisons.

Windows on the bait chromosomes have, on average, 21% ± 12% (chromosome 4) and 24% ± 13% (chromosome 10) of their fragments captured in the three most deeply-sequenced 4C samples, compared to 15% ± 3% for windows on non-bait chromosomes (mean of all windows ± standard deviation, averaged across C2.1, C2.2, and F2.1). The most prominent contacts are made within the terminal 40 Mb of chromosome 4q and the terminal 30 Mb of 10q; beyond these points, contacts fall to a level found on non-bait chromosomes. The values for windows on the small arms of the bait chromosomes (4p, 16% ± 3%, 10p, 17% ± 3%; average ± SD) were comparable to those on non-bait chromosomes. Sliding-window profiles from low-coverage libraries also displayed prominent contacts near the bait, but overall were less sharply-defined than the profiles of high-coverage libraries (**Figure 3.5**).

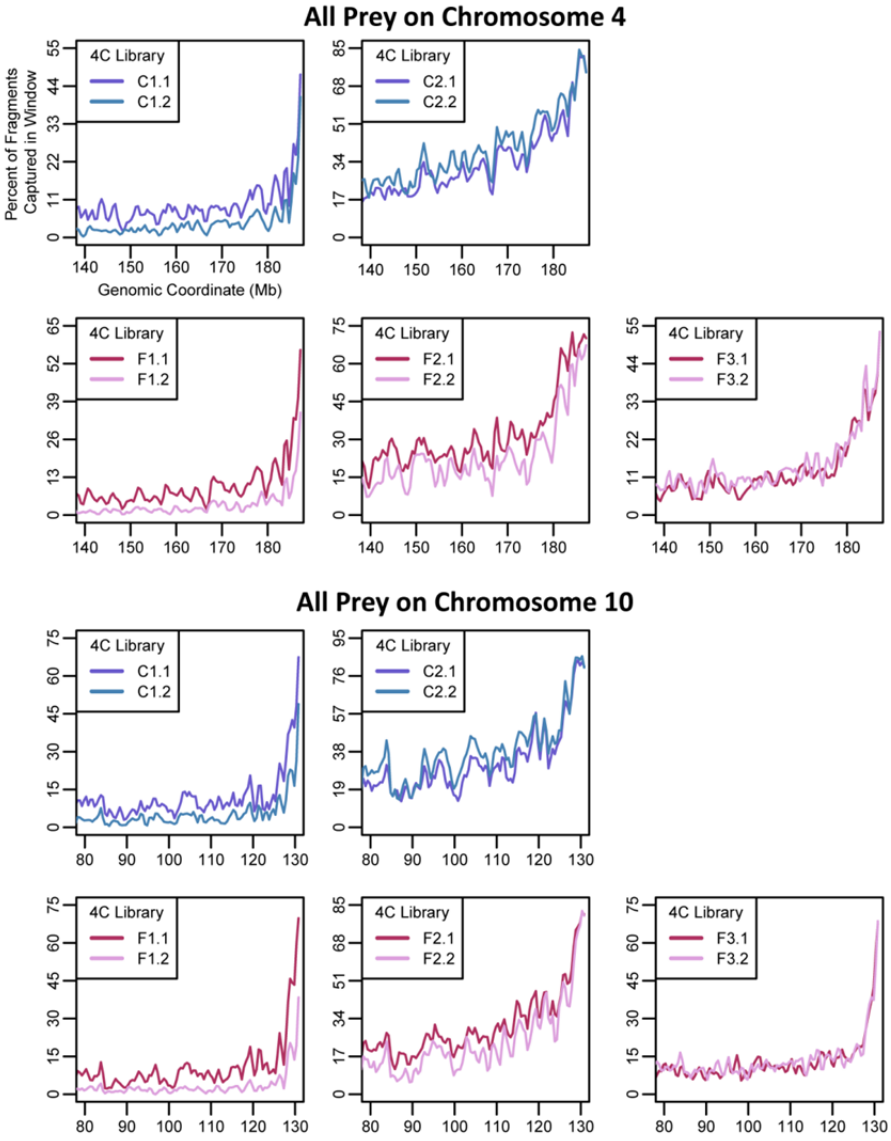

Figure 3.5 Sliding window profiles on bait chromosomes by sample
Sliding-window analysis was performed as described in the Figure 3.4 legend. Prey profiles on the bait chromosomes for replicate 4C libraries are plotted here, with axis labels displayed only for the top-left graph.

3.4 Assignment of prey to bait haplotypes

My bait fragment in the FSHD locus exists in four copies in the genome (two on the homologues of chromosome 4 and two on chromosome 10), and I wished to distinguish the sets of prey fragments captured by each copy of the bait. The SSLP within the bait fragment defines haplotypes of the FSHD locus[49], and I will thus refer to each distinct copy of the bait fragment as a "bait haplotype". The SSLP genotypes of the myoblast cell lines used for this study gave me the

potential to distinguish the chromosome 4 homologues from each other and from chromosome 10, but not chromosome 10 homologues from each other (**Table 3.1**).

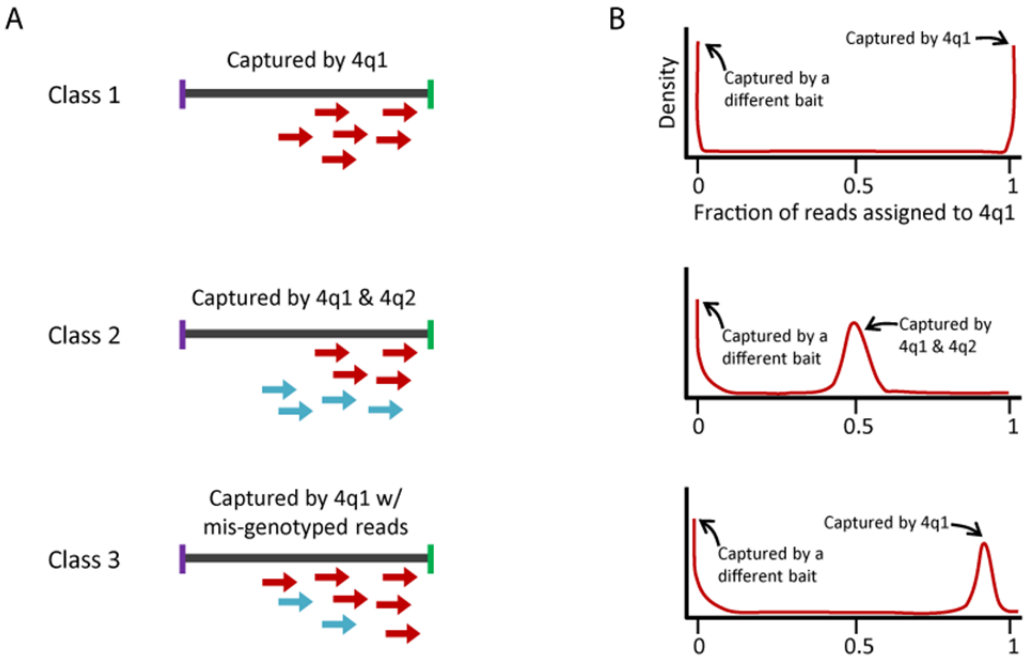

Figure 3.6 Expectations for assignment of prey to bait haplotypes
(A) Representations of a prey fragment captured by only one bait (case 1), evenly by two baits (case 2) or by one bait with mis-genotyped reads. Reads assigned to the 4q1 and 4q2 haplotypes are colored red and blue, respectively. **(B)** Representations of the distributions of the fraction of genotyped reads assigned to 4q1 for a collection of prey fragments. Prey captured by a single bait (as in case 1) result in a profile with peaks at 1 and at 0 (upper panel). Prey captured by two baits (case 2) result in a profile with peaks at 0.5 and at 0 (middle panel). Prey captured by one bait, but containing reads incorrectly assigned to another bait haplotype (case 3), result in a profile with a peak shifted to the left of 1 and a peak at 0.

Each prey fragment contains reads that are paired with a bait read, whose SSLP genotype reveals the identity of the bait fragment haplotype(s) that captured that prey fragment. I refer to prey reads as "genotyped" if their paired bait read was successfully assigned to an SSLP genotype (details in Chapter 5 Section 6). To determine which bait haplotype captured a prey fragment, I considered the proportions of genotyped reads assigned to each bait haplotype within that fragment. For example, a fragment captured only by the 4q1 bait would contain reads only of that genotype (class 1, **Figure 3.6A**). If that fragment was captured by 4q1 and 4q2 baits independently, it would contain reads of both those genotypes in roughly equal proportions, assuming similar sequencing depth (class 2, **Figure 3.6A**). If the fragment was captured by the 4q1 bait, but some reads were mistakenly assigned to the wrong genotype, the proportion of 4q1 reads would be less that 100% (class 3, **Figure 3.6A**).

Now imagine a collection of prey fragments and consider just those captured by the 4q1 bait. If these 4q1-prey fall in class 1, a plot of the distribution of 4q1 genotype proportions for *all* prey would have a peak at 1, representing those prey captured by 4q1, and a peak at 0, representing those prey captured by other bait haplotypes (**Figure 3.6B**). The areas under these two peaks would be proportional to the numbers of prey fragments captured by 4q1 and the other bait haplotypes, respectively. If, instead, the 4q1-prey fall in class 2, the 4q1 genotype distribution would have peaks at 0.5 and 0 (**Figure 3.6B**). Finally, if the 4q1-prey fall in class 3, the 4q1 genotype distribution would have peaks shifted to the left of 1 and at 0 (**Figure 3.6B**). These genotype distributions can be used to assign prey fragments to bait haplotypes by choosing cutoff values, whereby fragments containing greater than or equal to that proportion of genotyped reads for a particular bait haplotype are classified as being "captured" by that haplotype.

In order to assign prey fragments in each library to bait haplotypes, I first required a prey fragment to have at least 10 genotyped reads. I calculated the proportion of genotyped reads assigned to each bait haplotype for each prey fragment and plotted the distributions of these values by bait haplotype, as discussed above (**Figure 3.7**). These plots indicate that many prey are highly enriched for reads matched with a single bait haplotype; a small minority have 100% of their reads matching a single haplotype. The shift of the right-most peaks away from 1 indicates that these prey likely contain some incorrectly genotyped reads, as discussed above. In addition, small peaks at 0.5 indicate the presence of prey captured by two bait haplotypes. Some of the genotype distributions were tri-modal, where only bi-modal distributions were expected. For a full discussion of the explanations for this phenomenon, see Chapter 5 Section 8.

Despite the shifted positions of peaks in the genotype distributions, it was clear that many prey contained a majority of genotyped reads from a single bait haplotype. In order to decrease the likelihood of incorrectly assigning haplotypes, I chose stringent cutoffs placed at the lower sides of the upper peaks in the genotype distributions (displayed as vertical lines in **Figure 3.7** and listed in Chapter 5 Section 8). The stringency of these cutoffs resulted in each prey fragment being assigned to only one bait, or to none at all.

On average, I assigned 58% of prey fragments to a bait haplotype (**Table 3.5**), with higher-coverage libraries yielding more assigned fragments than lower-coverage libraries (since I required prey to have at least 10 genotyped reads to be assigned to a bait). If there were no allelic bias in my assay and bait assignment, I would expect 25% of prey fragments to be assigned to each of the 4q haplotypes and 50% to the 10q haplotype. My datasets are close to this expectation, with roughly

25% of prey assigned to 4q1 in each 4C library, though 4q2/3 counts are slightly lower and 10q counts slightly higher than expected (**Table 3.5** and **Figure 3.8**). Reassuringly, very few fragments were assigned to haplotypes not present in the samples, i.e., to 4q3 in control and 4q2 in FSHD.

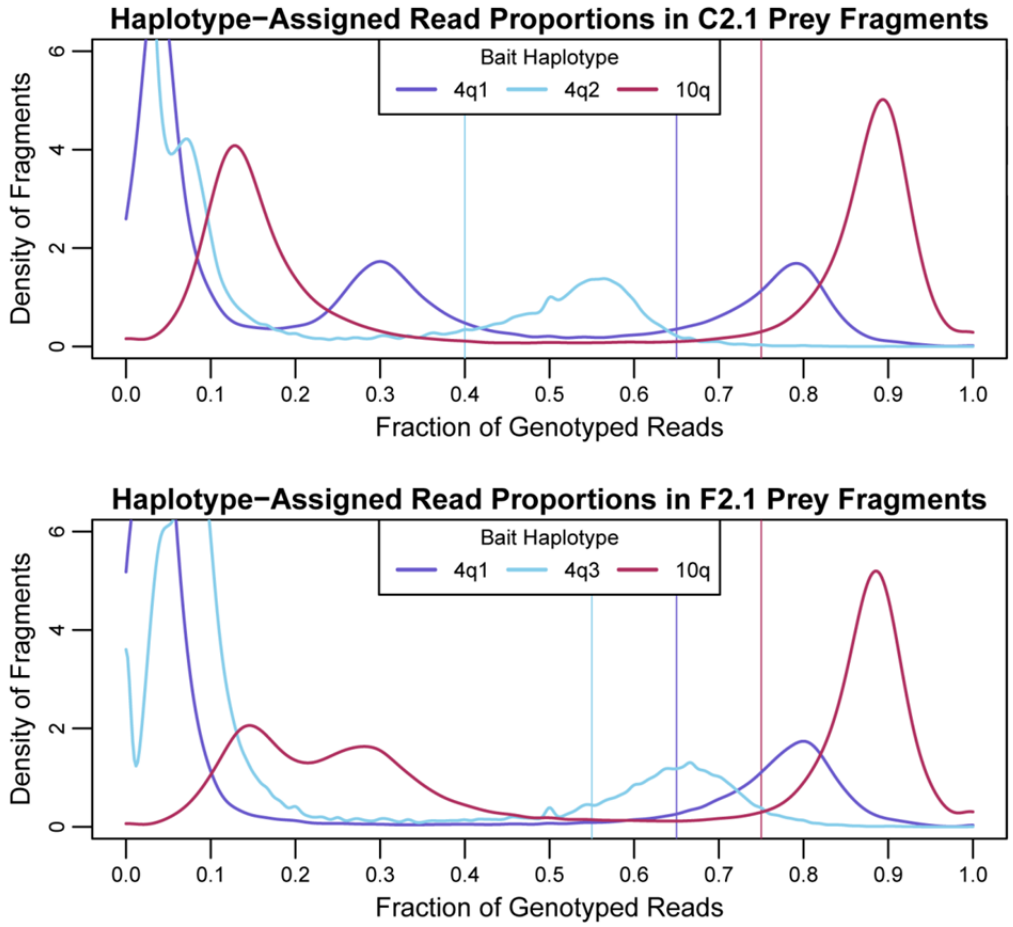

Figure 3.7 Proportions of haplotype-assigned reads in all prey
The proportion of all genotyped prey reads assigned to each bait haplotype was calculated for all prey fragments, and the distribution of those values is plotted here as a density graph for each bait haplotype in libraries C2.1 and F2.1. Cutoffs used for assigning prey to bait haplotypes are depicted as vertical lines, in the same color as the haplotype for which they were used.

Table 3.5 Summary of haplotype-assigned prey fragments
The percent of each sample's filtered prey fragments that were assigned to a bait haplotype is indicated in the second column, followed by the number of fragments assigned to each haplotype.

4C Library	Pct. Assigned	4q1	4q2	4q3	10q
C1.1	51%	7,042	5,608	5	15,497
C1.2	43%	1,837	1,418	1	4,291
C2.1	71%	21,124	21,116	8	48,922
C2.2	62%	22,160	21,076	35	48,724
F1.1	50%	5,255	70	4,248	12,264
F1.2	62%	1,610	3	1,194	3,447
F2.1	68%	22,109	1	18,784	49,080
F2.2	67%	10,305	6	10,192	21,752
F3.1	57%	9,090	2	9,185	9,101
F3.2	53%	12,441	3	11,945	12,222

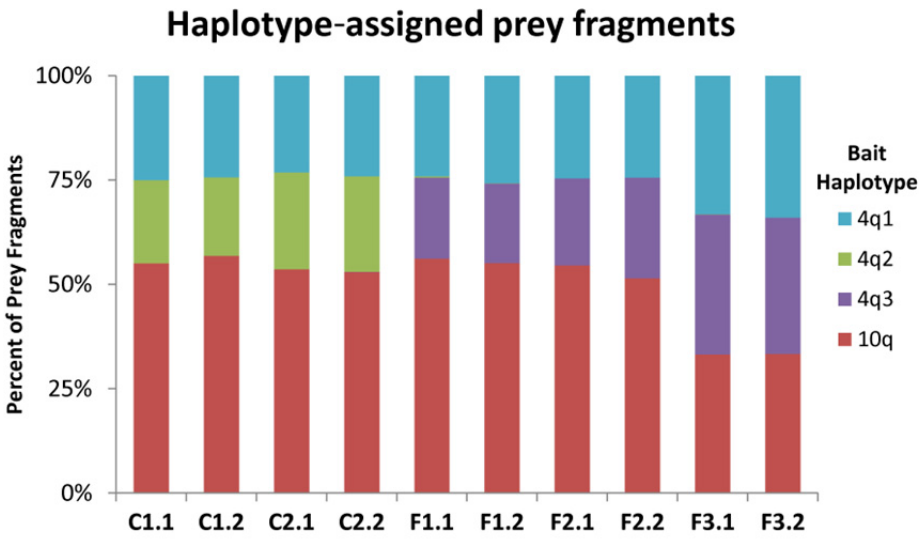

Figure 3.8 Proportions of haplotype-assigned prey fragments
The proportion of prey fragments assigned to each bait haplotype is plotted for each 4C library.

3.5 Characterization of haplotype-assigned prey

The bait-haplotype-assigned prey fragments are enriched on the bait chromosomes (**Figure 3.9**) as had been the case for all prey fragments (**Figure 3.1**). 4q1-, 4q2-, and 4q3-prey are enriched on chromosome 4 but not on chromosome 10, or any other chromosome. Conversely, fragments assigned to the 10q bait are enriched on chromosome 10, but not on any other chromosome. This enrichment is also observed at a local level in a sliding-window analysis, with haplotype-assigned prey profiles showing great enrichment in *cis* near their bait (i.e. 4q1-assigned prey are enriched near the bait at 4qter, but 10q-assigned prey are not greatly enriched in the same region, and vice versa; **Figure 3.10**).

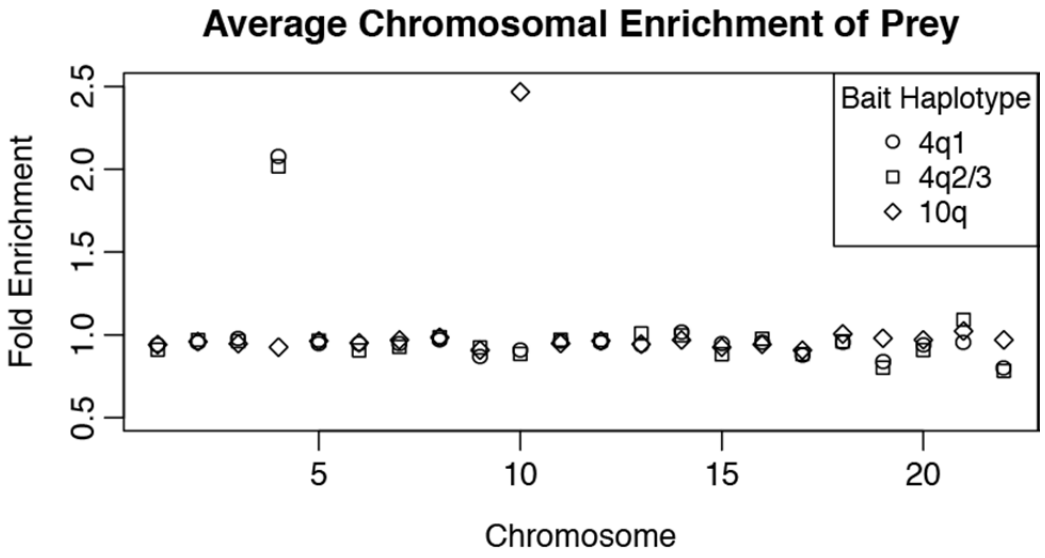

Figure 3.9 Haplotype-assigned prey are enriched on bait chromosomes
Fold enrichment of prey was determined by dividing the number observed on each chromosome by the number of total possible prey fragments on that chromosome and was averaged over all ten 4C libraries. The sex chromosomes were excluded because one library was derived from a female cell line, while the rest were derived from male cell lines. Prey assigned to each bait haplotype are depicted using different symbols, as indicated in the legend inside the figure. 4q1-, 4q2- and 4q3-assigned prey are enriched on chromosome 4 but not on chromosome 10. Conversely, 10q-assigned prey are enriched on chromosome 10 but not on chromosome 4.

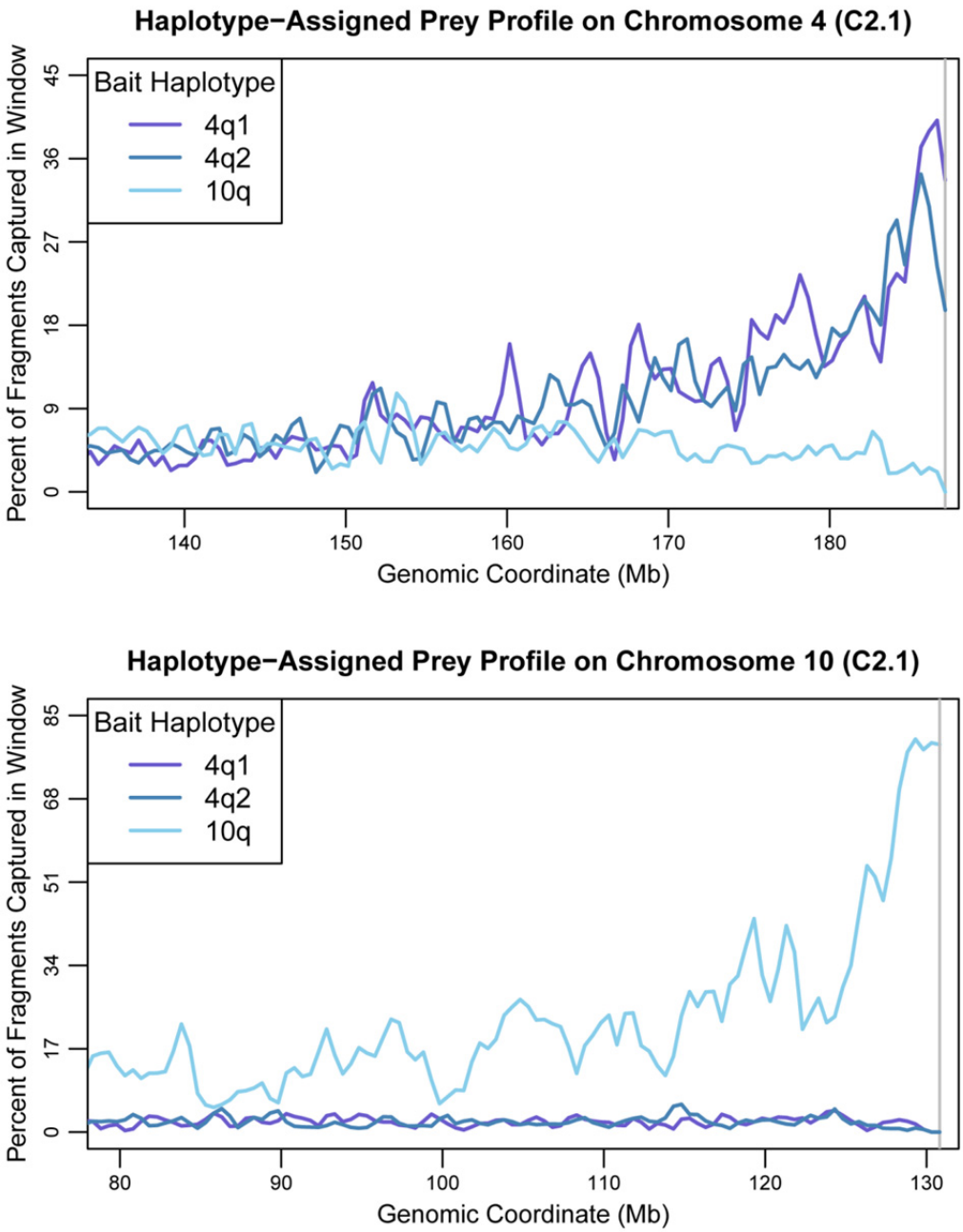

Figure 3.10 Sliding window analysis of haplotype-assigned prey distribution
The percentage of possible prey fragments captured by each bait haplotype is plotted in 1 Mb windows slid 500 kb along the bait chromosomes for library C2.1. 4q1- and 4q2-assigned prey are enriched near the bait on chromosome 4, but are not enriched along chromosome 10. Conversely, 10q-assigned prey are enriched near the bait on chromosome 10 but not on chromosome 4.

3.6 Characterization of prey found in multiple samples

Does each 4C library represent a unique collection of prey fragments captured by the SSLP, or are some fragments present in more than one library? If the latter is true, what might drive the repeated sampling of the same prey fragments, given that I capture ~10% of possible HD fragments in the genome in a given 4C library?

In order to compare prey-fragment sets across samples, I represented HD prey as HH fragments. Prey start as HH fragments in my 4C assay, and either end of an HH fragment can be captured by a bait fragment. As a result, comparing HD fragments between two samples can underestimate the number of shared prey fragments (**Figure 3.11**). For example, 17% of HD prey fragments from library C2.1 are also found in library C2.2, but when prey are represented as HH fragments, this percentage increases to 25%. Overall, 43% of HH prey fragments were found in two or more 4C libraries. The proportion of prey located on the bait chromosomes increased from 13% for all prey to 82% for those fragments found in all ten libraries. For comparison, the two bait chromosomes represent only roughly 12% of all possible HH fragments in the genome.

Figure 3.11 Representation of HindIII-DpnII prey fragments as HindIII fragments
Comparison of HindIII-DpnII fragments between samples can underestimate the number of shared fragments if two HD fragments originate from the same parent HindIII fragment, as depicted. Purple vertical lines represent HindIII sites, green vertical lines DpnII sites.

The presence of so many multi-library prey on the bait chromosomes can be explained by the chromosome territory effect and the prey's proximity to the SSLP. Fragments on the same chromosome as the bait are more likely to be captured by it, since they are confined to the same nuclear sub-volume occupied by that chromosome. Likewise, since so many prey fragments are captured near the bait (> 50% of prey on bait chromosomes are within 50 Mb of the bait fragment), these near-bait fragments have a high probability of being observed in multiple 4C libraries. Accordingly, sets of prey found in multiple libraries are located progressively closer to the bait as I increase the number of libraries considered (data not shown).

But, what phenomena could produce the multi-library prey on non-bait chromosomes? The SSLP has been previously characterized as an enhancer-blocking insulator[60]. Given that insulators in vertebrate genomes exert their function via the CTCF protein, and CTCF has been shown to mediate

chromosomal interactions in the nucleus[40,70], I examined whether CTCF binding might explain the recurring SSLP-prey interactions on non-bait chromosomes that I observed in my dataset.

I used CTCF binding-site locations in primary human myoblasts identified by the ENCODE Consortium[71] to calculate the distance between each prey fragment and the nearest CTCF site. Considering the entire collection of prey fragments on non-bait chromosomes, I find that prey captured in at least six 4C libraries are enriched at and near CTCF sites when compared to sets of GC-matched restriction fragments sampled from the same chromosomes (**Figure 3.12**, details in Chapter 5 Section 9.1).

I also found enrichment for proximity to CTCF sites in prey found in multiple libraries when I considered only the subset of prey assigned to bait haplotypes. While a given prey fragment could be found in multiple libraries, it was not assigned to a bait haplotype in every dataset. Therefore, my definitions of multi-library prey in the context of bait haplotypes accommodated this missing information. I classified a prey fragment as being captured by both 4q and 10q baits if it was assigned to a 4q haplotype in some of my ten libraries, and a 10q haplotype in others. I classified a prey fragment as being captured only by the 4q bait if it was assigned to 4q in any library, but not to 10q in any library (and *vice versa* for 10q assignment). Using these classifications, 10q-captured prey found in at least seven libraries and 4q+10q-captured prey found in at least six libraries appear to be enriched near CTCF sites (**Figures 3.13** and **3.14**).

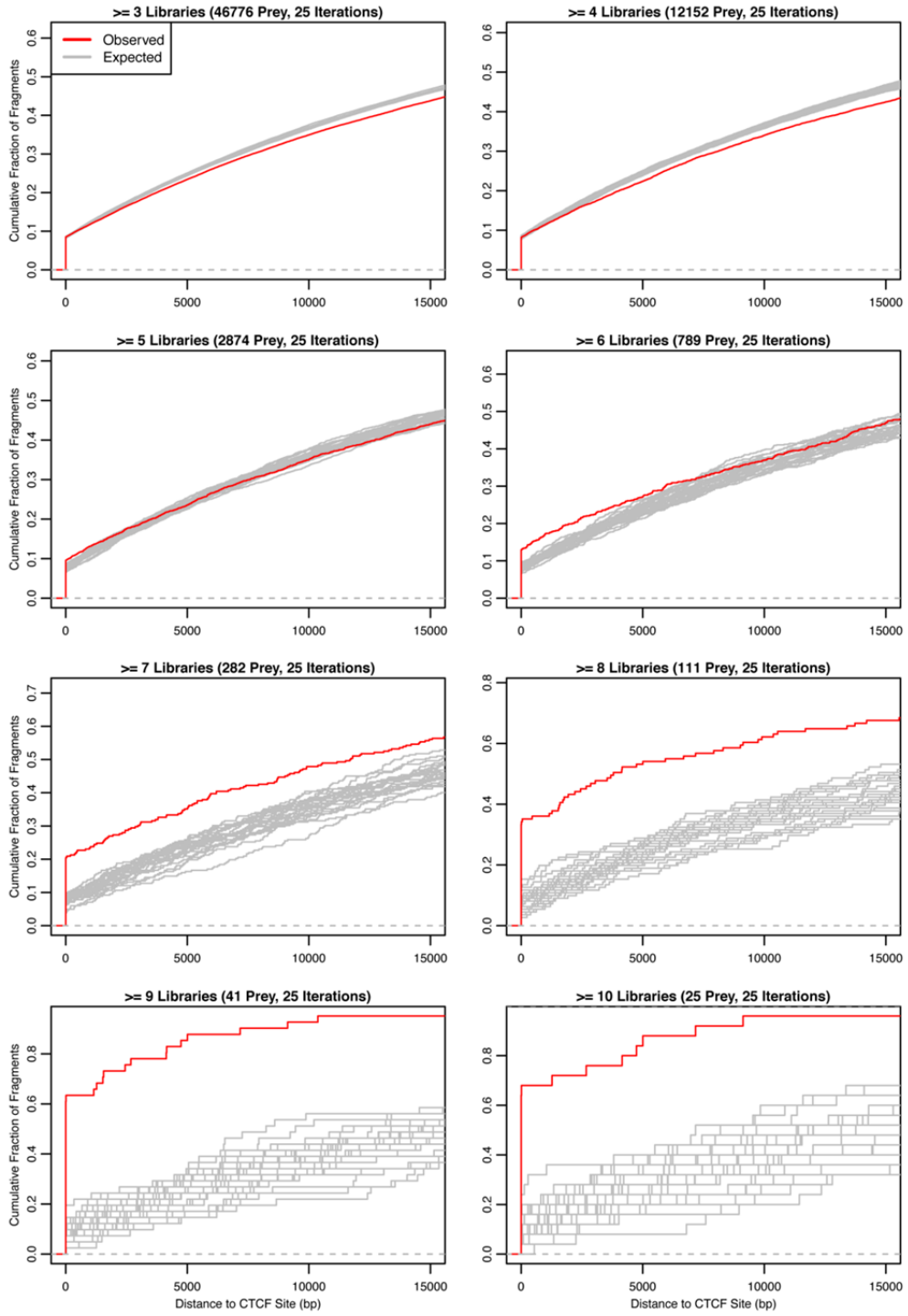

Figure 3.12 Prey in multiple libraries are enriched at and near CTCF sites on non-bait chromosomes
Cumulative density plots are shown depicting the distances between sets of prey fragments and CTCF sites. All prey fragments from non-bait chromosomes (not 4 and not 10) were used in this analysis. The distance to the nearest CTCF site is shown on the x-axis; fragments directly overlapping a CTCF site are given a distance of zero. The cutoff for the number of libraries in which a fragment is found is given in the title of each graph along with the number of fragments meeting that cutoff. The cumulative density for the observed fragments is plotted in red. The distributions of values for 25 sets of random fragments (described in main text) are plotted in grey.

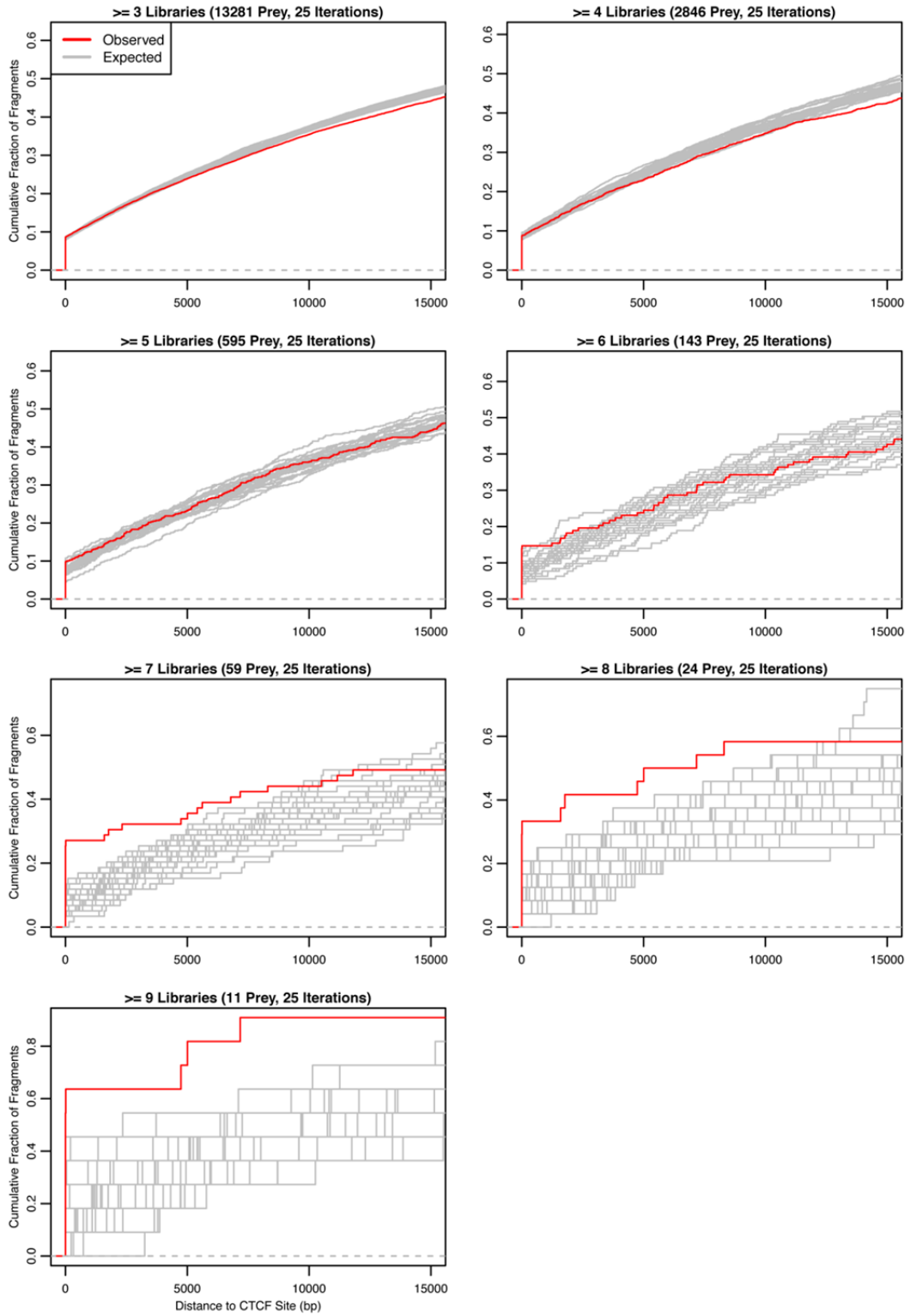

Figure 3.13 10q-assigned prey in multiple libraries are enriched at and near CTCF sites on non-bait chromosomes

See legend of Figure 3.12 for a description of the graph components and main text for a description of how 10q-assigned prey were determined.

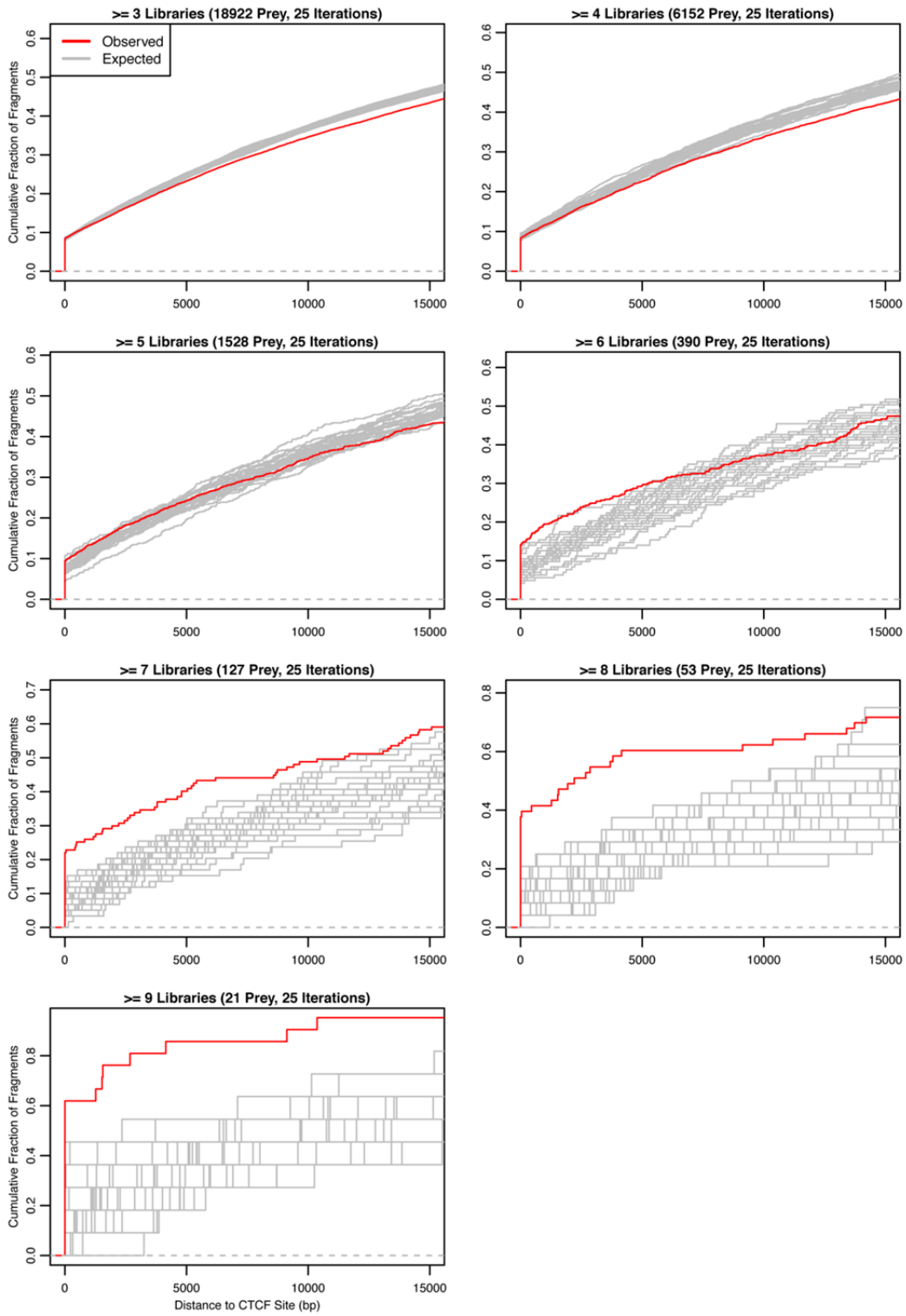

Figure 3.14 4q+10q-assigned prey in multiple libraries are enriched at and near CTCF sites on non-bait chromosomes

See legend of Figure 3.12 for a description of the graph components and main text for a description of how 4q+10q-assigned prey were determined.

During manual examination of the locations of prey (regardless of haplotype assignment) found in multiple libraries, I noticed that many appeared to be close to centromeres. I examined this trend quantitatively and found that the proportion of prey near centromeres on non-bait chromosomes increases as I consider prey found in progressively more 4C libraries, reaching 58% for prey captured in all ten libraries (**Figure 3.15**). Although CTCF sites are present at centromeres, the CTCF and centromere associations of these prey collections appear to be independent of one another, and not due to the presence of prey in segmental duplications (**Figure 3.16** and **Table 3.6**). Additionally, the presence of certain prey in multiple libraries does not appear to be due to a general enrichment for repetitive elements (**Figure 3.17**).

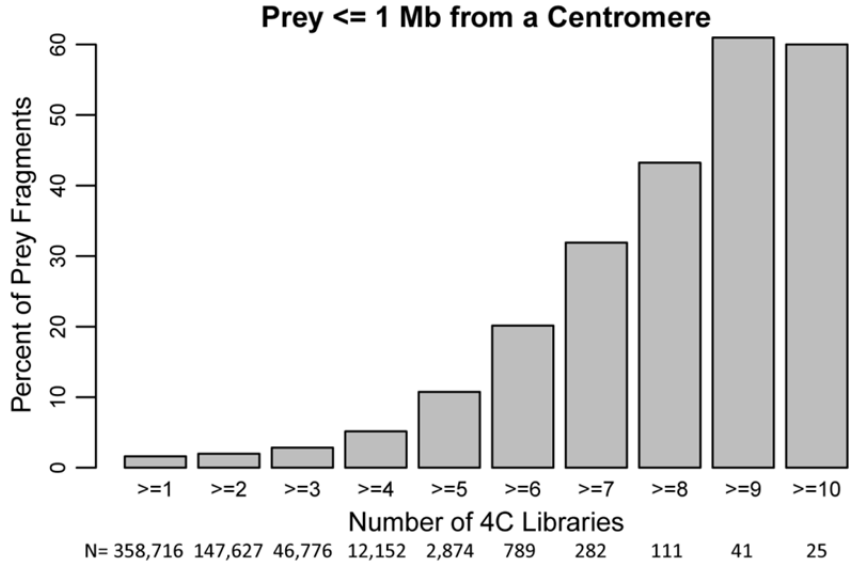

Figure 3.15 Prey in multiple libraries are enriched at centromeres of non-bait chromosomes
The percentage of prey fragments ≤ 1 Mb from a centromere on non-bait chromosomes (i.e., not chromosomes 4 or 10) increases as I consider the sets of prey fragments found in progressively more 4C libraries.

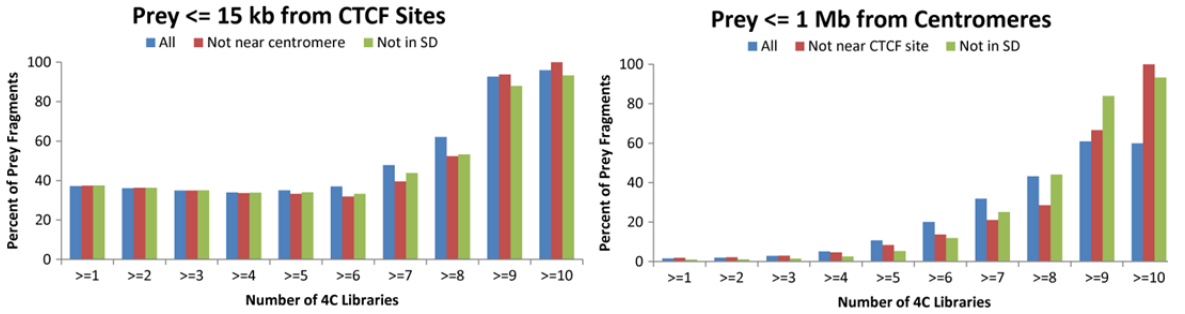

Figure 3.16 Enrichment of multi-library prey at CTCF sites and centromeres is not due to segmental duplications

For both plots, the y-axis gives the percentage of fragments in each set of prey fragments (defined by number of libraries, x-axis) meeting the given condition. As sets of prey fragments found in progressively more 4C libraries are considered, the percentage of fragments ≤ 15 kb from a CTCF site increases (blue bars, left panel). This trend is observed even when prey near centromeres (red bars) or in segmental duplications (green bars) are removed from the analysis. As sets of prey fragments found in progressively more 4C libraries are considered, the percent ≤ 1 Mb from a centromere also increases (blue bars, right panel). This trend remains when prey ≤ 15 kb from a CTCF site (red bars) or in segmental duplications (green bars) are removed from the analysis.

Table 3.6 Sample sizes of prey fragment sets used for Figure 3.16

For the plots shown in **Figure 3.16**, this table gives the total number of prey fragments used in each series, from which the percent meeting the given condition (CTCF site or centromere proximity) was calculated. Column names match series names from **Figure 3.16**.

# Lib	Total Number of Prey in Set					
	CTCF site proximity			Centromere proximity		
	All	Not near cen.	Not in SD	All	Not near CTCF	Not in SD
>=1	358716	352928	341482	358716	225041	341482
>=2	147627	144695	140297	147627	94219	140297
>=3	46776	45444	44059	46776	30436	44059
>=4	12152	11526	11172	12152	8020	11172
>=5	2874	2565	2506	2874	1866	2506
>=6	789	630	622	789	497	622
>=7	282	192	203	282	147	203
>=8	111	63	77	111	42	77
>=9	41	16	25	41	3	25
>=10	25	10	15	25	1	15

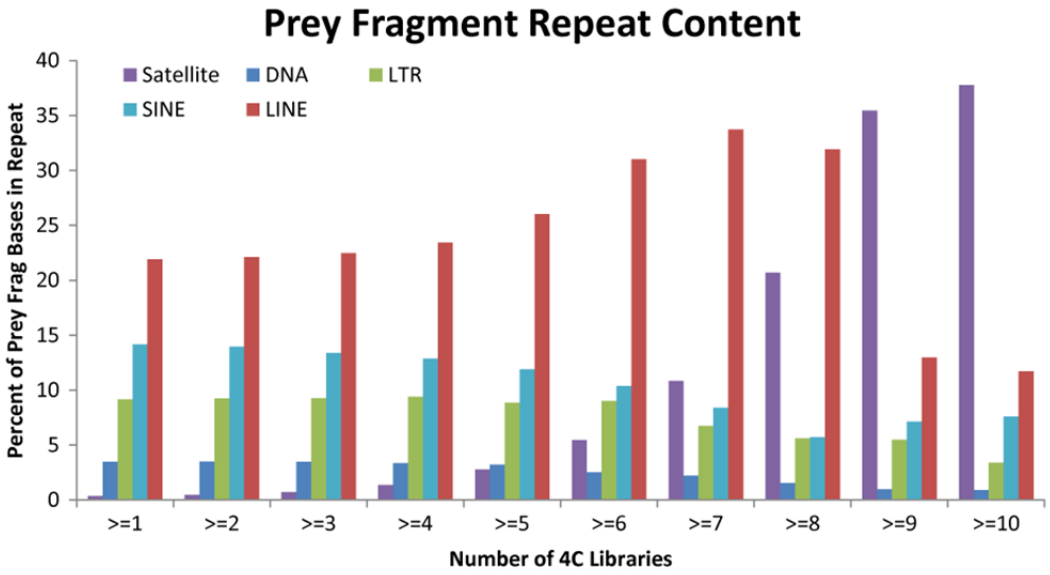

Figure 3.17 Presence of prey in multiple 4C libraries is not due to general repeat enrichment
The repeat content of each set of prey fragments found in progressively more 4C libraries was determined and expressed as a percentage of all bases in each prey collection covered by a given repetitive-element family. These percentages do not dramatically increase for prey fragments found in many libraries, except for satellite repeats (see **Figure 3.15** for centromere association), indicating that repetitive elements do not in general drive the presence of prey fragments in multiple libraries.

3.7 Differences between control and FSHD prey profiles

My hypothesis about altered nuclear organization in FSHD predicts that the 4q1 SSLP would associate with inactive regions of the genome in control cells and active regions in FSHD cells. I used the peaks and valleys of my 1-Mb sliding-window profiles in the terminal 50 Mb of chromosomes 4 and 10 to test this question, since they represent regions that are significantly enriched and depleted for SSLP-captured prey, respectively, and because the bait chromosomes contain the highest density of captured prey. I systematically selected 1-Mb windows for peaks and valleys that were significantly above and below a sliding median defined for each haplotype-assigned profile from C2.1, C2.2, F2.1 and F2.2 (**Figure 3.18**; see Chapter 5 Section 10.1 for details). This selection process yielded ~10 peaks and ~10 valleys for each profile.

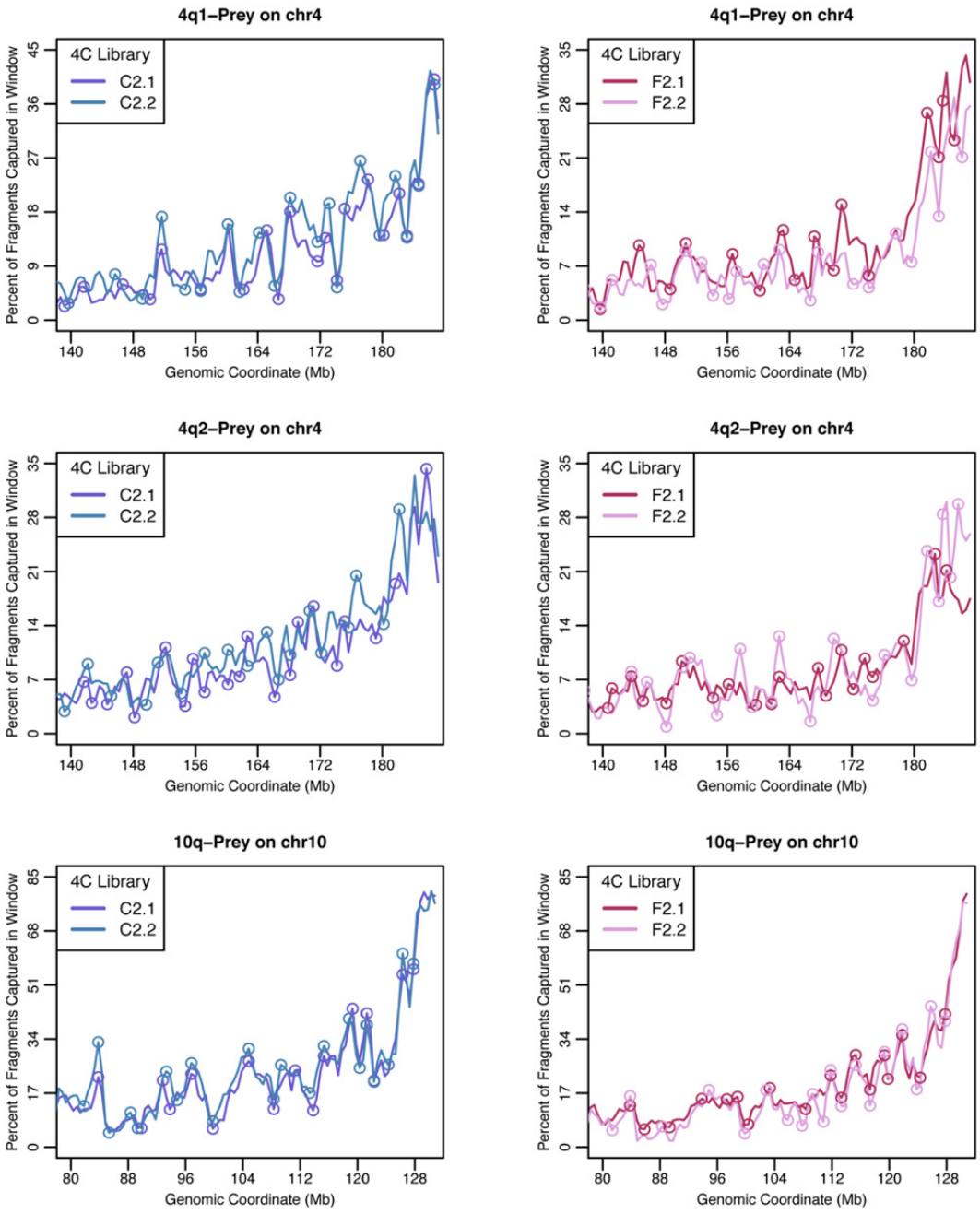

Figure 3.18 Selected peaks and valleys for gene expression and LAD analysis
Shown are 1-Mb sliding window plots for haplotype-assigned prey on chromosomes 4 and 10. 4C libraries are indicated in the inset legends. Windows used for peak and valley analysis are indicated by circles, in the same colors as the samples from which they were selected.

Overall, I find that peaks have a lower transcriptional output than valleys. Since I created my 4C libraries from primary myoblasts, I obtained RNA-seq data from the same cell type generated by the ENCODE Consortium[72]. In each peak and valley window, I summed expression values for all genes to give a measure of the region's total transcriptional output, and made boxplots of the distributions of these values (**Figure 3.19**, details in Chapter 5 Section 10.1). I then used a bootstrap version of the Kolmogorov-Smirnov (KS) test to determine whether two distributions were significantly different for various comparisons. For the 4q1-prey profile in control samples, peaks have a significantly lower transcriptional output than valleys ($p < 0.001$). For the 4q1-profile in the FSHD samples, this distinction is lost: there is no significant difference in the transcriptional output of peaks and valleys. However, the 4q1-peaks in FSHD have a significantly higher transcriptional output than control 4q1-peaks ($p < 0.01$). The difference in transcriptional output is also significant for FSHD 10q peaks vs. valleys ($p < 0.01$).

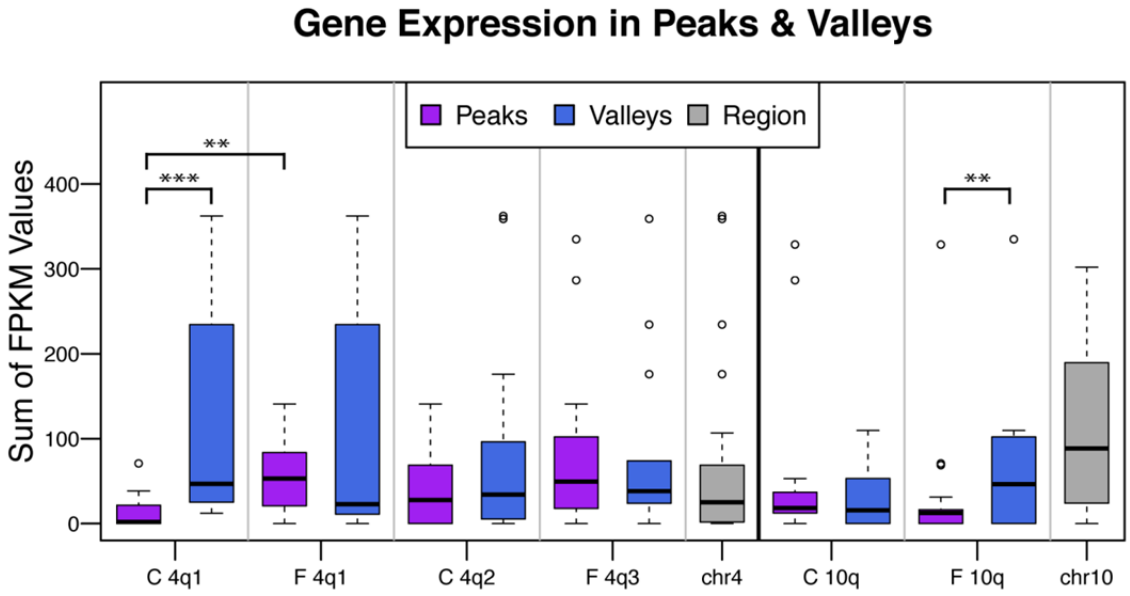

Figure 3.19 Gene expression within peaks and valleys
Boxplots for gene expression of peak and valley datasets. The x-axis lists the bait haplotype under consideration; "C" indicates controls, "F" indicates FSHD; 4q1, 4q2 and 4q3 peaks and valleys are located on chromosome 4, while 10q peaks and valleys are located on chromosome 10. Grey boxplots show gene expression for all windows in the terminal 50 Mb of chromosomes 4 and 10. Asterisks indicate significant differences (bootstrap KS test): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Inactive regions of the genome (including the FSHD locus) tend to localize at the nuclear periphery, adjacent to the nuclear lamina[6,18,73]. In order to test whether the FSHD locus preferentially associates with other lamina-adjacent regions of the genome, I compared the overlap of peak and valley windows with experimentally-defined lamin-associated domains (LADs) mapped

in fibroblasts[73], using the same approach as above. I found that peaks had higher LAD overlap than valleys (**Figure 3.20**). I observed significantly higher LAD overlap in peaks than in valleys of the 4q1-prey profiles in both control and FSHD samples ($p < 0.001$ and $p < 0.01$, respectively), but this comparison was not significant for other bait haplotypes. As with gene expression, I found a significant difference between control and FSHD 4q1 peaks. In this case, control peaks had higher LAD overlap than FSHD peaks ($p < 0.05$).

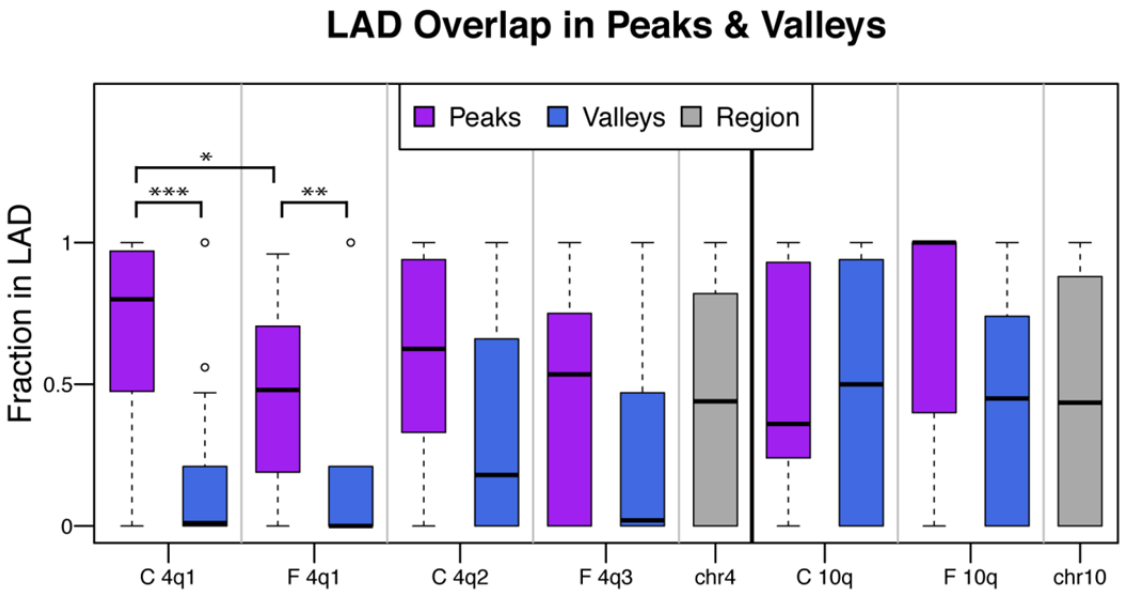

Figure 3.20 LAD overlap within peaks and valleys
Boxplots for LAD overlap of peak and valley datasets. See **Figure 3.19** legend for description. Asterisks indicate significant differences (bootstrap KS test): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In order to identify, in an unbiased manner, specific regions of the genome that might be differentially captured by the SSLP in control and FSHD samples, I performed a sliding-window analysis on the three most deeply-sequenced 4C libraries (C2.1, C2.2 and F2.1) comparing the density of prey captured by each bait haplotype. Briefly, a window was considered significant when the difference in the 'percent of fragments captured' in that window between the FSHD sample and both control samples exceeded the 99th percentile of differences scored between the two control samples for all windows on the same chromosome (see Chapter 5 Section 10.2 for details). This stringent analysis identified sixteen 500-kb windows where the 4q1-assigned prey counts differed significantly between control and FSHD libraries, five windows for 4q2/4q3-assigned prey and nine windows for 10q-assigned prey (**Table 3.7**). Sixteen of these 30 windows were located on the bait chromosomes within 40 Mb of the bait, and in the majority of these windows there was a larger normalized prey fragment count in control than in FSHD libraries (**Figure 3.21**). These control-versus-

FSHD differences predominantly occurred in LADs and were more extensive for 4q1-prey on chromosome 4 than 10q-prey on chromosome 10. The comparison of 4q2-prey on chromosome 4 did not identify any significant differences between control and FSHD prey counts.

Table 3.7 Regions of significant difference between control & FSHD prey counts

Chromosomal locations of 500-kb windows with significant differences in control & FSHD haplotype-assigned prey counts. "Ratio C/F" indicates the average "%Frags Captured in Window" of C2.1 & C2.2 divided by the value for F2.1. Orange shading indicates windows where FSHD > control; blue indicates control > FSHD. Validation values express the ratio of normalized 4q3 counts divided by 4q1 counts in the libraries listed, as discussed in the main text.

Bait	Chr	Start	End	% Frags Captured in Window			Ratio	Validation	
				C2.1	C2.2	F2.1	C/F	F2.1	F2.2
4q1	1	4,140,622	4,640,621	7%	8%	0%	-	-	-
	3	10,357,141	10,857,140	1%	2%	7%	0.2	0.1	0.0
	4	163,544,277	164,044,276	18%	20%	4%	4.8	2.2	0.7
		171,294,277	171,794,276	14%	18%	7%	2.5	1.3	0.6
		172,794,277	173,294,276	13%	13%	6%	2.4	1.3	2.9
		178,544,277	179,044,276	20%	17%	9%	2.1	1.0	0.6
		179,544,277	180,044,276	19%	19%	9%	2.0	1.2	1.5
		180,044,277	180,544,276	19%	24%	9%	2.5	1.4	0.7
		181,044,277	181,544,276	20%	20%	11%	1.9	1.1	0.8
		181,794,277	182,294,276	22%	19%	10%	2.1	1.9	1.8
		186,044,277	186,544,276	9%	11%	21%	0.5	1.2	1.1
		187,044,277	187,544,276	27%	28%	38%	0.7	0.7	1.3
	5	1,455,261	1,955,260	7%	6%	0%	-	-	0.6
	12	112,691,896	113,191,895	2%	1%	8%	0.2	0.5	3.0
	14	96,289,541	96,789,540	1%	0%	6%	0.1	0.2	-
	16	27,744,754	28,244,753	1%	0%	7%	0.1	0.2	4.0
4q2/3	2	64,942,374	65,442,373	1%	1%	6%	0.1		
	3	42,607,141	43,107,140	1%	3%	7%	0.3		
	6	97,555,068	98,055,067	3%	3%	9%	0.3		
	11	13,239,517	13,739,516	1%	1%	5%	0.2		
	15	89,021,393	89,521,392	2%	2%	9%	0.2		
10q	1	48,040,622	48,540,621	2%	2%	3%	0.2		
	10	107,724,748	108,224,747	3%	4%	2%	2.7		
		130,674,748	131,174,747	2%	3%	3%	1.4		
		130,174,748	130,674,747	0%	1%	1%	1.3		
		131,924,748	132,424,747	1%	0%	1%	1.2		
		132,924,748	133,474,747	0%	2%	0%	1.3		
		134,774,748	135,274,747	0%	0%	0%	0.9		
	12	133,341,896	133,841,895	6%	4%	5%	3.6		
	18	65,417,249	65,917,248	3%	2%	0%	7.0		

The D4Z4 deletion carried on the 4q1 haplotype in the FSHD cell lines would not be expected to alter the organization of the 4q3 haplotype of the FSHD locus in these cells. Thus, as a means of validating windows where I found control-versus-FSHD differences in 4q1-assigned prey density, I compared the normalized counts of the 4q1- and 4q3-assigned prey in F2.1. I expected

that in windows where control 4q1-prey outnumbered F2.1 4q1-prey, the normalized F2.1 4q3-prey count would also outnumber F2.1 4q1-prey. Indeed, in nine of the ten windows on chromosome 4, the difference in normalized counts went in the same direction as the control vs. FSHD comparison (**Table 3.6**). For three of those nine windows, this result also held in the F2.2 library.

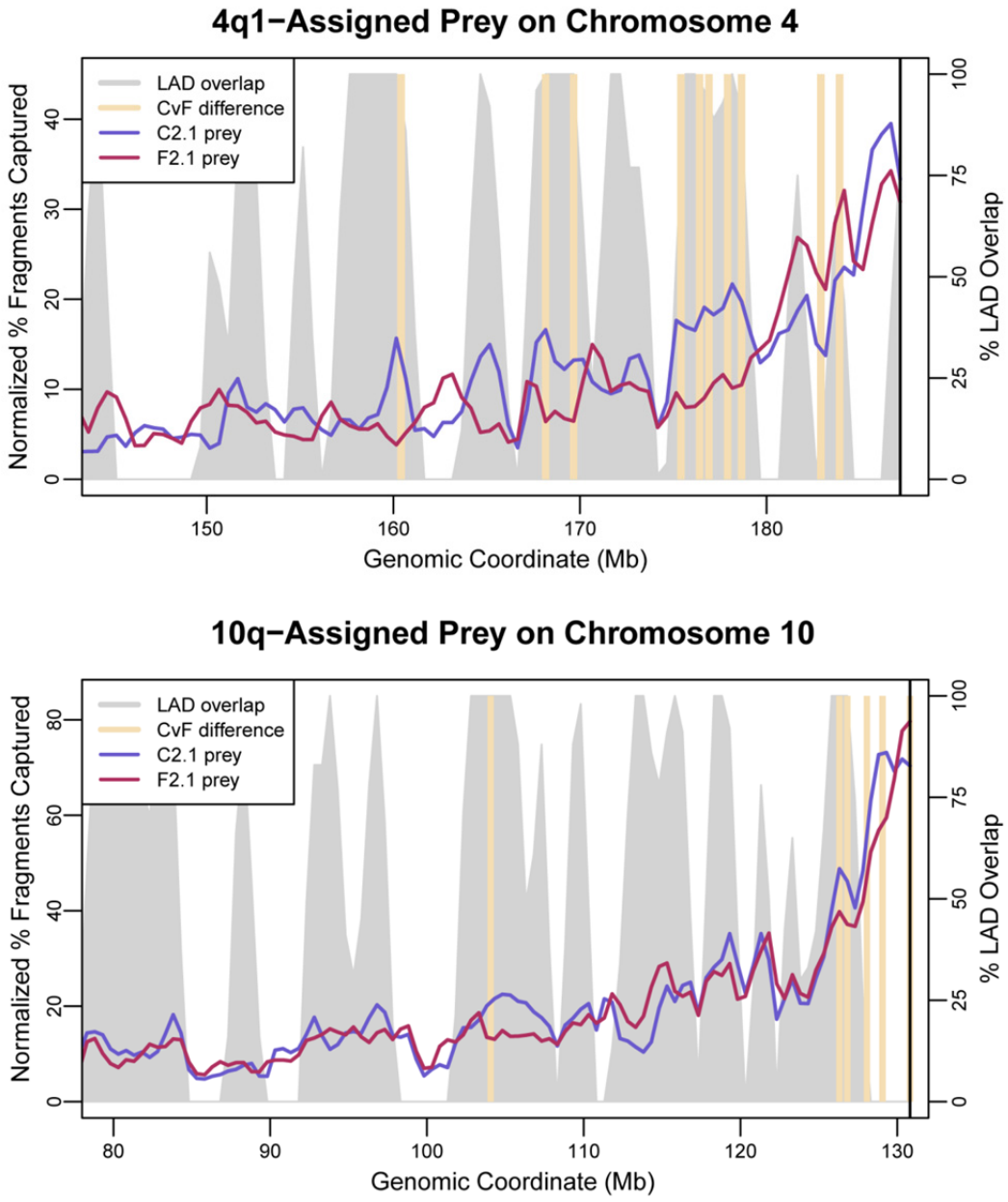

Figure 3.21 Regions of significant difference between control & FSHD prey counts on bait chromosomes
Sliding-window plots were generated as previously described. The haplotype-assigned prey profiles are shown for one control and one FSHD sample as indicated in the plot legend. LAD overlap in the same windows is plotted in grey (right-hand y-axis). Windows identified as having a significant difference in prey counts between control and FSHD are indicated with beige rectangles.

3.8 Conclusion

I have shown that the SSLP preferentially interacts with regions in its own chromosome territory. The most prominent of these interactions occur in the terminal 30-40 Mb of the bait chromosomes and coincide with areas of low gene expression and high lamin-associated domain overlap (when compared to regions depleted of interactions). Notably, these associations appear to be altered in FSHD. In addition, I find 16 regions on chromosomes 4 and 10 where control and FSHD prey counts differ significantly. In the next chapter I discuss the implications of these findings in the light of broader aspects of nuclear organization and with regard to the mechanism of FSHD pathogenesis.

Chapter 4: Discussion

In this chapter, I discuss the implications of my 4C-seq results for understanding the nuclear organization of the FSHD locus and whether it is altered by pathogenic D4Z4 deletions. I then propose a model for how the position of the FSHD locus within the nucleus might play a role in the inappropriate expression of *DUX4* in FSHD muscle cells. In Section 4.2, I consider the lessons I have learned from my thesis work about the advantages and limitations of 3C-based techniques for studying nuclear organization and suggest improvements for my 4C-seq approach. Finally, in Section 4.3, I give my perspective on future high-throughput studies of nuclear organization.

4.1 Implications of findings

Deletions within the D4Z4 macrosatellite repeat array in the 4q subtelomere cause FSHD, the third most common inherited muscular dystrophy. When I began my thesis research, the pathogenic mechanism for FSHD was unknown, but recent work has led to a model whereby the deletion leads to the de-repression of *DUX4* in muscle cells[44]. This gene, located within each D4Z4 repeat unit, encodes a transcription factor normally expressed in the germline[62,63]. DUX4 induces apoptosis in muscle cells in culture and activates cancer testis antigen genes that might induce an immune response in affected muscles *in vivo*[63]. Why D4Z4 deletion leads to de-repression of *DUX4* is still an unsolved mystery.

Before *DUX4* was identified as a prime candidate gene involved in FSHD pathogenesis, it was thought that D4Z4 functioned as a non-coding regulatory element, affecting the expression of proximal genes either through changes in nuclear organization or packaging of the 4q subtelomere[50,56]. Two FISH studies found no gross mis-localization of the FSHD locus away from the nuclear periphery in FSHD muscle cells[56,57]. However, other groups described local chromatin changes at 4q35, including a loss of heterochromatic marks and detachment from the nuclear matrix[58,59], hinting that D4Z4 deletions lead to more subtle changes in the nuclear organization of the FSHD locus that could not be detected by the earlier FISH experiments. Now, with a candidate gene identified, it remains reasonable to hypothesize that changes in nuclear organization contribute to the de-repression of *DUX4* to cause FSHD.

This thesis details my efforts to describe the normal nuclear organization of the FSHD locus and test whether that organization is altered by pathogenic D4Z4 deletions. To accomplish these two goals, I developed an allele-aware 4C assay, which I call 4C-seq. This assay allows one to

determine which regions of the genome ("prey") are in close physical proximity to a locus of interest ("bait") in three-dimensional nuclear space in a population of cells. By choosing a polymorphic bait fragment (the SSLP) and using paired-end, high-throughput sequencing, I was able to deeply sample the prey fragments captured by specific copies of the FSHD locus. I was especially interested in achieving this specificity because initial 3C studies of the FSHD locus were unable to distinguish duplicated copies of the locus, making their results challenging to interpret[74,75]. I generated ten 4C libraries from control and FSHD primary myoblast cell lines (see **Table 3.1**), which I then interrogated while optimizing my high-throughput sequencing protocol. I was able to use all of my data to describe the features of the prey fragments captured by the SSLP, despite the fact that my 4C libraries were not all sampled to an equal extent.

When I conceived this experiment, I imagined that I would use the 4C assay to paint a coarse picture of the "nuclear neighborhood" of the FSHD locus. In the same way a man can describe his physical location in a city based on the landmarks he sees from where he is standing, the prey fragments captured by the SSLP could describe the neighborhood of the FSHD locus based on its contacts with the rest of the genome. I then imagined that I would detect large changes in the neighborhood of the FSHD locus carrying a D4Z4 deletion by finding that the SSLP of this allele captured a different set of prey sequences in FSHD cells than the SSLP of a non-deleted FSHD locus in control cells. Thus, I would show that D4Z4 deletions change the regions of the genome with which the FSHD locus associates in the interphase nucleus.

I did not find such stark differences between these two prey sets, with the deleted FSHD locus contacting a completely different set of genomic regions than the non-deleted locus. Instead, prey fragments from both control and FSHD libraries were heavily enriched on chromosomes 4 and 10 ("bait chromosomes"), which contain the SSLP, but not on other ("non-bait") chromosomes (**Figure 3.3**). On the bait chromosomes, one might imagine a couple of extreme scenarios: that the bait locus makes a very limited number of interactions (which might differ between deleted and non-deleted alleles) or that the number of captured simply decays monotonically with increasing distance from the bait (expected if DNA is organized as a freely draining random polymer[10,11,76]). Instead, using a sliding-window analysis (Chapter 3 Section 3), I found a landscape of prominent interactions (peaks) whose height decreased with increasing distance from the bait (see **Figure 3.4**). The most prominent peaks extended for 40-50 Mb from the bait on both chromosomes 4 and 10. In analyzing these peaks on the bait chromosomes, as well as prey fragments on non-bait chromosomes that were repeatedly captured across multiple 4C libraries, my results provide a

picture of the nuclear neighborhood of the FSHD locus, and suggest that D4Z4 deletions might indeed shift the neighborhood of the locus within its own chromosome territory.

My results are consistent with the view that the FSHD locus normally associates with the inactive nuclear compartment. The FSHD locus is tethered to the nuclear lamina, a repressive region of the nucleus, and is normally packaged as heterochromatin, suggesting multiple layers of silencing[56,58]. As I discussed in Chapter 1, active and inactive regions of the genome are compartmentalized within the interphase nucleus: active regions contact other active regions on the same or different chromosomes, whereas inactive regions contact other inactive regions, primarily on the same chromosome[15,33,34]. Regions of the genome located at the nuclear periphery (lamin-associated domains, LADs) have been previously shown to contain lowly-expressed genes and to interact with one another[17,77]. In control cells, I found that the 4q1 SSLP contacts regions in the terminal 50 Mb of chromosome 4 that have low transcriptional output and high LAD overlap (**Figure 3.19** and **Figure 3.20**), suggesting that the FSHD locus associates with lowly-expressed or inactive regions within its own chromosome territory that are also positioned at the nuclear periphery. Although the LADs I used were defined in fibroblasts, I expect a substantial fraction of them remain as LADs in myoblasts, since a study of four mouse cell types found that LADs overlap by 73-87% between cell types[18].

The SSLP has been shown to possess the properties of an enhancer-blocking insulator sequence, which are typically conferred by the CTCF protein that binds to such sequences[40]. While it remains to be determined if CTCF directly binds the SSLP, I found that the SSLP associates with CTCF sites on non-bait chromosomes (**Figure 3.12**), which suggests that it may contact other insulator elements in the genome. In addition, I found that the SSLP associates with centromeres on non-bait chromosomes, which are known heterochromatic regions of the genome[78]. My analysis of CTCF sites considered all of my 4C samples in aggregate as well as all copies of the SSLP, suggesting that the CTCF association is a general property of the SSLP. It is formally possible that prey fragments captured by another bait in my libraries, analyzed in a similar fashion, would also show an association with CTCF sites, but my results are consistent with studies showing that insulator sequences interact with one another in the nucleus[35,77].

In contrast to its arrangement in control cells, the FSHD locus appears to associate with more active regions of chromosome 4 in FSHD cells. I found that the 4q1 SSLP in FSHD cells (which identifies the deleted locus) contacts regions in the terminal portion of chromosome 4 with higher transcriptional output and lower LAD overlap than the regions contacted by the 4q1 SSLP in control

cells (**Figure 3.19** and **Figure 3.20**). I did not find significant control vs. FSHD differences in gene expression or LAD overlap for the regions contacted by the 4q2 SSLP on chromosome 4 or the 10q SSLP on chromosome 10. Furthermore, my analysis of prey profiles to find regions of significant difference between control and FSHD prey counts found more extensive differences in 4q1-assigned prey on chromosome 4 than 10q-assigned prey on chromosome 10, and no significant differences in 4q2/3-assigned prey on chromosome 4 (**Table 3.7**). The 4q1-prey profiles on chromosome 4 showed large, lamin-associated regions where the control prey counts were higher than FSHD prey counts (**Figure 3.21**). Taken together, the altered contacts of the 4q1 SSLP on chromosome 4 suggest that D4Z4 deletions do not lead to gross changes in the nuclear position of 4q35, but rather *decrease* the association of the FSHD locus with the inactive nuclear compartment and *increase* its association with the active compartment.

What role might this shift into association with active regions play in FSHD pathogenesis? It is clear that *DUX4* is now the prime candidate gene involved in the development of FSHD through its inappropriate expression in muscle cells. However, expression of this gene is rare: it is difficult to detect in patient biopsies, and it is highly expressed at the RNA and protein level in only 0.1% of FSHD muscle cells in culture[62,63]. This low penetrance raises the possibility that there are multiple determinants of its expression.

I propose the following model for how nuclear organization could contribute to the end-point of *DUX4* expression (**Figure 4.1**). Epigenetic phenomena can be imagined as layers of control that either increase or decrease the probability of a gene being expressed. The FSHD locus is normally associated with markers of heterochromatin, including DNA methylation and the histone modifications H3K9me3 and H3K27me3, and is frequently located at the nuclear periphery, where it is tethered to the nuclear lamina[53,56-58]. The first condition that must be met for *DUX4* expression is the presence of a permissive polyadenylation signal distal to the D4Z4 array, which occurs on certain haplotypes of the FSHD locus on chromosome 4. Next, contractions of the D4Z4 array lead to a loss of heterochromatic marks at the locus. While this altered chromatin state does not cause 4q35 as a whole to move away from the nuclear periphery, it alters the organization of the FSHD locus to bring it into contact with regions of chromosome 4 that have higher gene expression than the regions with which it would normally associate. Gene transcription is not impossible at the nuclear periphery, but it is inefficient, and high-level transcription has been shown to occur at punctate sites (transcription factories) within the nuclear interior[2]. Thus, the loss of heterochromatin along with placement of the deleted FSHD locus near regions of active transcription might provide for the

stochastic activation of the *DUX4* gene. In this way, FSHD can be seen as resulting from a failure of multiple mechanisms for silencing transcription from the D4Z4 array.

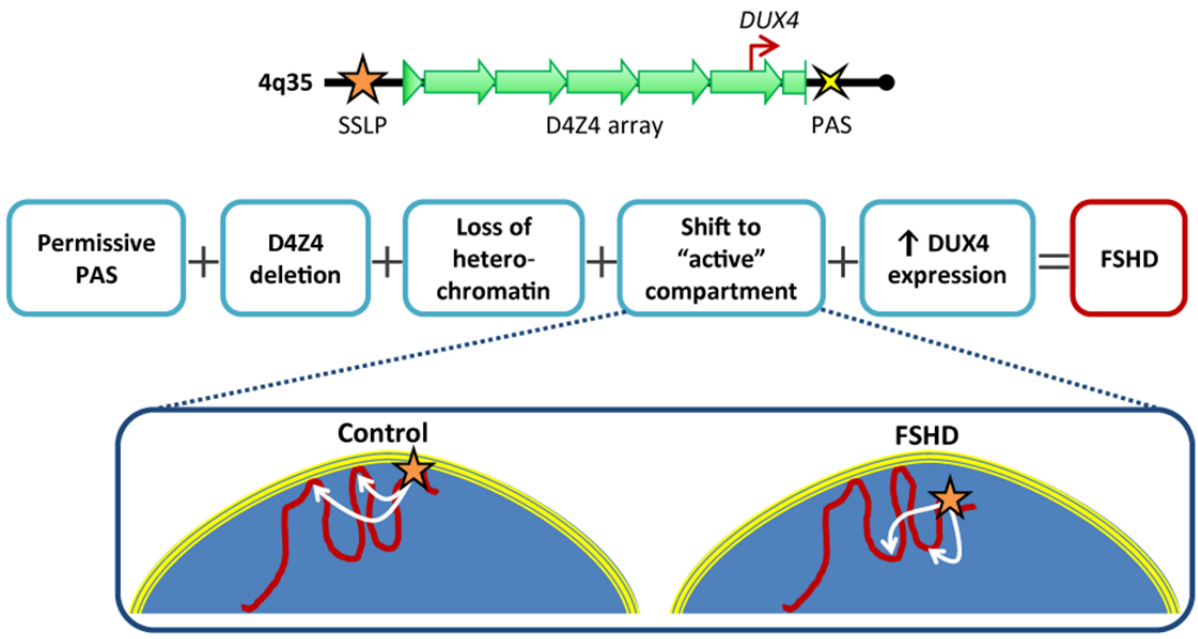

Figure 4.1 Model for the involvement of nuclear organization in FSHD
Model for the contribution of genetic and epigenetic factors that result in *DUX4* mis-expression and lead to FSHD. This thesis adds a fourth component to the model, the shift of the FSHD locus from associating with "inactive" regions of the genome to "active" regions.

Individuals with FSHD2 do not have D4Z4 array contractions, yet they share a loss of heterochromatic marks on a permissive chromosome 4 copy of the FSHD locus with FSHD1 individuals, pointing to a common disease mechanism[55,58]. However, in these individuals, all other copies of the locus *also* display this loss of heterochromatic marks, so the manner in which these marks are lost differs between FSHD1 and FSHD2. This phenomenon further highlights the epigenetic nature of FSHD pathogenesis, and it remains to be seen whether the *all* copies of the FSHD locus relocate to the active nuclear compartment in FSHD2 individuals. One of the muscle cell lines I used in my study was isolated from an FSHD2 individual, but the 4C libraries I generated from it were not sequence deeply enough to address this question.

The results I present in this thesis warrant further study of the nuclear organization of the FSHD locus. Because of the uneven coverage among my ten 4C libraries, I was only able to study contacts of the SSLP on the bait chromosomes in my most highly-sequenced libraries from a control cell line and an FSHD1 cell line. Sequencing additional 4C libraries from independent cell lines to a high depth would help confirm my findings about the genomic regions contacted by the SSLP and

the differences I observed between 4q1-assigned prey in control and FSHD cells. Doing so was beyond the scope of my thesis work, as many of my 4C libraries were exhausted while developing the 4C-seq technique. In particular, it would be important to examine the organization of the FSHD locus in FSHD2 cells, to test if its neighborhood changes when the locus is de-repressed in the absence of a D4Z4 deletion; and in cells with a D4Z4 deletion on a non-permissive chromosome 4, to help unravel whether the shifted neighborhood is a cause or consequence of *DUX4* expression.

4.2 Advantages, challenges and thoughts for improvement of 4C-seq

The application of 4C-seq provides a novel means of examining nuclear organization in an allele-aware manner: by selecting a bait fragment containing sequence polymorphisms and using paired-end high-throughput sequencing, this approach allows the researcher to deeply sample the prey fragments captured by different alleles of the bait.

Previous 4C studies used microarrays or low-throughput Sanger sequencing to identify prey fragments, which limits the prey that can be identified (this is a problem more for the latter approach than the former); in addition, these studies did not distinguish bait alleles[24,31,65,79]. A recent 4C study used a polymorphic restriction enzyme site to detect specific bait alleles along with high-throughput sequencing of captured prey fragments[80]. 4C-seq presents an advantage over this approach by giving information on the interactions of all bait alleles simultaneously, in the same sample of cells.

I overcame many challenges in the course of developing 4C-seq, including sample selection, the complexity of the 4C assay, preparation of Illumina sequencing libraries and analysis of high-throughput sequencing data. I discuss these challenges below in the context of suggestions for improving the 4C-seq to simplify the detection of prey captured by a bait of interest. Specifically, these improvements would increase the number of informative reads coming out of 4C-seq (i.e., reduce the number of reads removed by filtering) and allow one to use read depth to inform data analysis (I did not fully exploit read depth, because of concerns that the multiple PCR steps I used would bias read counts).

- **Use simple bait polymorphisms.** I exploited the SSLP as a vantage point for uncovering the nuclear neighborhood of the FSHD locus because this sequence distinguishes haplotypes of the locus on both chromosomes 4 and 10. However, realizing this potential required the procurement of appropriate cell lines, in

addition to modification of the 4C assay to determine the sequences of both the bait and prey fragments. Fortunately, by the time I was ready to perform 4C with muscle cells, a collection of primary myoblast cell lines from control individuals and individuals with FSHD was available at the University of Rochester, and the SSLP genotypes of these individuals had been determined. I was thus able to select cell lines from individuals who were heterozygous at their 4q SSLP, allowing the comparison of the interactions of single copies of the FSHD locus between control and FSHD cells (**Table 3.1**). The SSLP is a complex repeat sequence, which complicated my assignment of prey fragments to bait haplotypes because of errors introduced by DNA polymerase during PCR (discussed in Chapter 3 Section 4 and Chapter 5 Section 8). Consequently, the use of one or multiple heterozygous SNPs within a bait fragment would simplify the resolution of bait alleles.

- **Simplify detection of bait/prey ligation partners to increase the proportion of informative reads and allow for their counting.** In the 4C assay, closed-circle formation between bait and prey fragments allows other restriction fragments to ligate between the "true" bait/prey partners. In my experiment, this process resulted in many prey reads mapping to DpnII-DpnII fragments, instead of the true HindIII-DpnII fragments (**Figure 2.5** and **Figure 3.2**). This phenomenon decreases the number of informative reads coming out of the 4C-seq assay, and it can be ameliorated by purifying the bait/prey ligation junction directly after the first ligation step, as is done in the e4C and Hi-C methods[23,33]. Going after these ligation junctions directly has the additional benefit of removing the inverse PCR step used in the 4C assay, which would then allow one to use read counts to more confidently determine how frequently a particular prey fragment was ligated to the bait. In my thesis work, I simply used the presence of prey reads mapping within a given restriction fragment to determine that the fragment had been captured by the SSLP, and I did not use read depth to draw conclusions about the frequency of that capture.

- **Couple interaction detection with protein detection**. One way to simplify the functional characterization of bait/prey interactions is to couple chromatin immunoprecipitation with 3C-based methods[23,65]. This coupling allows the researcher to ask which interactions are mediated by a protein of interest, such as RNA polymerase or CTCF. An experiment using this design could be used to confirm my finding that the SSLP associates with CTCF sites by directly assaying CTCF-mediated interactions in myoblasts.

4.3 Perspective on the future of high-throughput nuclear organization studies

High-throughput sequencing has reinvigorated and accelerated many areas of genomics, from whole-genome sequencing to the genome-wide mapping of transcription factor binding sites and chromatin modifications. This technology has also brought the study of nuclear organization into the genomics era through its marriage with 3C-based methods and the development of new techniques that interrogate the contribution of physical proximity in the nucleus to various phenomena, from transcription to the formation of translocations[81-83]. Such efforts have increased the awareness that the genome's arrangement within nuclear space plays a role in carrying out its functions and provided evidence that this arrangement may impact diseases processes such as the development of cancer[2,84].

The cost of high-throughput sequencing continues to fall as the number (and length) of reads generated from each sequencing run continues to rise[85]. One group has already taken advantage of this fact to sequence Hi-C libraries to a ~20-fold increased depth over the original Hi-C paper, allowing the analysis of interactions at the sub-100-kb scale[86] (the original Hi-C paper analyzed the genome at a resolution of 1 Mb). This increase in resolution uncovered "topological domains" of the genome that are remarkably stable between cell types and conserved through evolution; these domains were not observed in the original Hi-C data[33].

Yet increased sequencing depth does not, by itself, resolve one of the major limitations of 3C-based studies: they report an *average* picture of the configuration of the genome in the nucleus because of their requirement for millions of cells as input. (FISH studies, on the other hand, can describe this configuration in single nuclei.) Single-cell sequencing would be required to overcome this limitation, allowing, for example, the preparation and sequencing of Hi-C libraries from multiple individual cells. While such an experiment would give a snapshot of nuclear organization at one time-point, it would provide a direct measure of the number of cells in which a given interaction is

occurring at that time-point and shed light on which interactions occur simultaneously in the same nucleus (both can currently only be assayed using FISH). Single-cell sequencing would also be highly advantageous for testing my model for the involvement of nuclear organization in FSHD: 4C-seq performed on single cells that had been sorted for *DUX4* expression would provide a clearer picture of whether the association of the deleted FSHD locus with active regions of the genome is a cause or consequence of *DUX4* expression. Development of single-cell 4C-seq would be necessary here instead of performing 4C-seq on a pool of flow-sorted *DUX4* expressing cells; since *DUX4* is only expressed in 1/1000 cells in culture[62], 10^{10} cells would need to be sorted to obtain the 10^7 cells necessary for a single 4C-seq experiment as performed in my thesis work. This is a prohibitive number of cells to grow in culture, especially for adherent cells such as myoblasts that must be kept at a certain density to prevent their differentiation.

The central question of the nuclear organization field is whether the structure of the genome in the nucleus determines gene expression, or is determined by gene expression. This question remains incompletely answered. 3C-based techniques now provide an invaluable tool for high-throughput studies that can, with additional technological advances, ultimately answer this question. To do so, I believe such studies must move beyond the simple mapping of (average) overall nuclear organization and comparison of those maps between cell types to perturbation experiments where organization is altered on a large scale and changes in gene expression are tested. Technological advances that reduce the number of cells required as input for 3C-based studies would help such experiments, so that, for example, RNAi-mediated knockdown of lamin A/C to disrupt the structure of the nuclear lamina could be performed in a small population of cells to see how it alters nuclear organization.

Nuclear organization studies of the differentiation process will also be enlightening, since they might reveal alterations in genome structure that precede changes in gene expression. One such experiment that has been proposed takes advantage of the ability of the master transcription factor MYOD to convert non-muscle cells into myoblasts[87]. Comparing the organization of the genome in fibroblasts before and after forced *MYOD* expression to the organization of the genome in myoblasts would test whether this transcription factor induces a “myogenic” organization before the muscle gene expression program is activated.

Ultimately, research in this field will unravel the mysteries of how the genome operates in three-dimensions (and over time, during the differentiation process). Elucidation of the rules that govern nuclear organization and the role this organization plays in establishing and maintaining

gene expression programs will (and have already begun to) provide a framework for understanding gene regulation beyond the primary DNA sequence. This framework will prove invaluable for understanding diseases such as FSHD and cancer, where there might not always be simple mutations in gene coding sequences that lead to the mis-regulation of gene expression and result in disease phenotypes.

Chapter 5: Materials and Methods

5.1 Cell culture

Primary human myoblasts were obtained from Rabi Tawil the University of Rochester Medical Center and grown in F-10 medium supplemented with 20% fetal bovine serum, 1% penicillin-streptomycin, 10 ng/ml bFGF and 1 µM dexamethasone.

5.2 Reagents

PCR primers were synthesized by IDT (Coralville, Iowa) and purified with standard desalting unless otherwise noted. Lyophilized stocks were resuspended in molecular grade H_2O to a concentration of 100 µM. Sources of other reagents used are identified in subsequent sections.

Table 5.1 Primer and adapter sequences
"*" indicates a phosphorothioate bond, "/5Phos/" indicates a 5' phosphate group, "#" indicates HPLC purification.

Name	Sequence (5' -> 3')	Use
4C-SSLP-N-F1	GCAGGGTGCGGTCATGTGTGAT	Inverse PCR
4C-SSLP-R4	GGGGAAGCAAGGAAGAGTGT	
SP2-AD	CTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T	Prey read adapter #
SP2-AD_rc	/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG	
4C-SSLP-amp-A1SP1	AATGATACGGCGACCACCGAGATCTACACACACATAAGGTGGAGTTCTGGTTTCAGC	Attachment & sequencing (SSLP read)
4C-amp-A2SP2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	Attachment & sequencing (prey read)
4C-SSLP-SP1	CGAGATCTACACACACATAAGGTGGAGTTCTGGTTTCAGC	SSLP read sequencing

5.3 Generation of 4C-seq libraries

The 4C assay was carried out as described in[24], with slight modifications; a detailed protocol can be found in Appendix A. 4C libraries were then linearized by digestion with SanDI and PpuMI to increase efficiency of the subsequent PCR step (rationale explained in Chapter 2 Section 5). Inverse primers within the HindIII-DpnII bait fragment encompassing the SSLP (4C-SSLP-N-F1 & 4C-SSLP-R4, **Table 5.1**) and the Roche High-Fidelity PCR system were used to amplify the SSLP sequence and prey fragment that had been ligated to the bait in the population of molecules generated by the 4C assay.

Inverse-PCR products were prepared for high-throughput sequencing as detailed below. All purification steps were carried out using either the QIAquick Gel Extraction Kit (QIAGEN 28704) or the QIAquick PCR Purification Kit (QIAGEN 28104) using standard columns (elution volumes ≥ 30 µl) or MinElute columns (elution volumes < 30 µl). DNA was eluted using Buffer EB from these kits. All

DNA quantitations were carried out using the Quant-iT dsDNA High Sensitivity Kit (Invitrogen Q-33120).

Fragmented, A-tailed DNA molecules ready for adapter ligation were prepared as follows. Six 25-µl PCRs (250 µM dNTPs, 1X High-Fidelity buffer, 2 mM $MgCl_2$, 300 nM primers, 75 ng linearized 4C library, 1.3 U High-Fidelity Polymerase) were pooled and run on a 1.5% agarose gel containing ethidium bromide (the equivalent of 2 PCR reactions/lane) at 80 V for 55 min. The cycling conditions for this PCR step were: 94°C 2min; 30 cycles of 94°C 15sec, 63°C 1min, 68°C 3min; 68°C 7min. PCR products of approximately 500-1,500 bp were excised from the gel under UV illumination and purified. DNA from each column was eluted in 30 µl EB, and the three eluates were pooled together and brought to 120 µl total volume with buffer EB. DNA was sheared using a Covaris E220 instrument (200 cycles per burst, duty cycle 5, intensity 4, 110 s), purified and eluted in 30 µl EB. Fragments were then end-repaired in a 50-µl reaction containing 1X NEBuffer 2, 100 µM dNTPs and 2 µl enzyme mix (Quick Blunting Kit, New England Biolabs E1201S) for 30 min at 24°C followed by heat inactivation for 10 min at 70°C. End-repaired fragments were A-tailed by adding 2.2 µl 5 mM dATP and 3.3 µl Klenow exo- (New England Biolabs M0212S) to each reaction and incubating for 30 min at 37°C. Reactions were purified and eluted in 16 µl EB, and quantitated.

Read 2 adapters were then ligated to the A-tailed DNA. A stock of 50 µM SP2 adapter was prepared by mixing 10 µl 100 µM 4C-SP2-AD with 10 µl 100 µM 4C-SP2-AD_rc (**Table 5.1**), heating to 95°C for 5 sec in a thermal cycler and gradually cooling over a period of at least 60 min to room temperature. Ligation was carried out in a 60-µl reaction with 1X Quick Ligation Buffer and 6 µl Quick T4 DNA Ligase (1:20 DNA:adapter; New England Biolabs M2200S) incubated for 15 min at 20°C. DNA was purified and eluted in 30 µl EB. To ensure that PCR products that had not been fragmented could not be sequenced, DNA was digested with 50 U HindIII (New England Biolabs R0104S) in a 50-µl reaction with 1X NEBuffer 2 incubated 3 h at 37°C followed by 20 min at 65°C. DNA was purified, eluted in 30 µl EB and quantitated.

A final PCR step was used to introduce attachment sequences for the Illumina flowcell to the ends of the DNA fragments and to amplify the template for sequencing (see **Table 5.1** for primer sequences). Three 50-µl PCR reactions were carried out (1X HF Buffer, 0.4 mM dNTPs, 2 mM $MgSO_4$, 1 µM 4C-SSLP-amp-A1SP1, 1 µM 4C-amp-A2SP2, 25 ng DNA, 1 U polymerase) using the Platinum Pfx system (Invitrogen 11708-013). Cycling conditions were: 94°C 2min; 18 cycles of 94°C 15sec, 65°C 30sec, 68°C 45sec; 68°C 5min. Each PCR reaction was split and purified using two QIAquick columns

eluted with 30 µl EB. The PCR reactions were split among four lanes of a 1.5% gel containing ethidium bromide, which was then run at 75 V for 80 min. PCR products of approximately 450-550 bp were excised from the gel under UV illumination, purified with a single column and eluted in 18 µl EB.

To verify the integrity of the adapter sequences and correct orientation of the product (read 1=SSLP, read2=prey), 2 µl of the sample was cloned into the pCR4-TOPO vector according to the manufacturer's instructions (Invitrogen K4575-J10), and approximately 12 clones were sequenced in standard Sanger sequencing reactions run on an Applied Biosystems 3730xl DNA Analyzer.

5.4 High-throughput sequencing

Paired-end sequencing was performed on Illumina GAIIx and HiSeq-2000 instruments according to the manufacturer's protocol for the particular reagent versions and machines used. Flowcells #1-2 (see **Table 3.2**) were run on the GAIIx machine with cluster generation kit version 2 and sequencing reagent kit version 4 for that machine. Flowcell #3 was run on the HiSeq machine with cluster generation kit version 2 and TruSeq SBS reagents for that machine. Flowcell #4 was run on the HiSeq machine with cluster generation kit version 3 and TruSeq SBS reagent kit version 3 for that machine. The first (SSLP/bait) read was sequenced from a custom forward primer (4C-SSLP-SP1, **Table 5.1**) for 100 cycles, and the second (prey) read was sequenced from the standard Illumina reverse read primer for 50 cycles.

5.5 Short read alignment and filtering

Prey reads that cross the ligation junction of two restriction fragments (joined by a DpnII site) will not align to the reference genome, so bases 3' of DpnII sites were trimmed from the reads. Trimmed prey reads of at least 20 nt were combined with the reads that did not contain DpnII sites, and all were aligned to the standard chromosomes (i.e., chr1 … chrY) of the reference human genome (NCBI Build 37.1/hg19) using BWA with default parameters, which allowed up to two mismatches in the seed region and kept only uniquely-mapping reads[68].

I analyzed sequence reads using custom R scripts (version 2.13.1) that relied upon open-source R packages for high-throughput sequence analysis from Bioconductor (www.bioconductor.org). Reads were filtered to remove those that aligned with low quality (mapq < 15), pointed toward HindIII sites, matched SSLP sequence or mapped to the two HindIII/DpnII fragments flanking the bait fragment (the rationale for each filter is discussed in Chapter 3 Section

2). I then removed redundant prey reads (those with identical start sites in the reference genome because this redundancy was expected to arise during sequencing rather than at prey capture), thereby retaining a representative read for a given position (whose paired SSLP read was assigned a genotype, if possible).

I determined whether prey reads matched the SSLP by aligning them to forward and reverse SSLP sequences, and determining whether these alignment scores exceeded the 99th percentile of alignment scores generated from 100,000 randomly-generated sequences (with the same mono-nucleotide and length distribution as the prey reads).

5.6 Genotyping of SSLP reads

I confirmed the expected SSLP genotypes in my cell lines in the course of sequencing clones of inverse PCR products from my 4C libraries generated during method development (Chapter 2 Section 4) and clones of Illumina sequencing libraries. Although the exact expected sequences were often observed, I also obtained sequences that differed from my expectation in the number of CAs or CTs, contained sequence errors, or both. DNA polymerase can slip during the copying of repeat tracts, and I frequently observed fewer CAs or CTs than expected. However, when I considered the SSLP sequence in its entirety, I was still able to assign genotypes to most of these non-standard SSLPs.

Regular expressions are programming tools used for pattern matching of strings of characters that easily incorporate ambiguity. In order to genotype the SSLP reads from my thigh-throughput data, I used regular expressions for each expected allele that were tolerant of slippage of the CA and CT repeat tracts as well as sequencing errors within the SSLP (**Table 5.2**). Reads that matched a given regular expression were assigned to that genotype, while reads that matched zero or more than one regular expression were classified as “unknown” genotype.

Table 5.2 Classification schemes for genotyping SSLP reads

Top row indicates generic SSLP sequence (M=C or A), other rows indicate schemes used for each SSLP genotype. "A" schemes match the reference SSLP sequences, while "B" and "C" schemes take into account sequencing errors and DNA polymerase slippage. By way of example, 4q1 scheme "A" reads: 8 or 9 CA's, followed by AA, followed by 9 or 10 CA's, followed by G, followed by any 27 letters, followed by 5 or 6 CT's, followed by any 6 letters, followed by TGCAAG.

SSLP	Scheme	CAn	MA	CAn	G/T	...	CTn	...	8nt
4q1	A	CA(8,9)	AA	CA(9,10)	G	*(27)	CT(5,6)	*(6)	TGCAAG
	B	*(16,18)	AA	*(18,20)	G	*(27)	*(10,12)	*(6)	TGCAAG
	C	*(16,18)	AA	*(18,20)	G	*(27)	CT(5,6)		
4q2	A	CA(8,9)	AA	CA(11)	G	*(27)	CT(5,6)	*(6)	TGCAAG
	B	*(16,18)	AA	*(22)	G	*(27)	*(10,12)	*(6)	TGCAAG
	C	*(16,18)	AA	*(22)	G	*(27)	CT(5,6)		
4q3	A	CA(17,18)			T	*(27)	CT(6,7)	*(6)	TGCTAG
	B	*(16)	CA	*(16,18)	T	*(27)	*(12,14)	*(6)	TGCTAG
	C	*(16)	CA	*(16,18)	T	*(27)	CT(6,7)		
10q	A	CA(7,8)	AA	CA(7,8)	T	*(27)	CT(6,7)	*(6)	TGCTAG
	B	*(14,16)	AA	*(14,16)	T	*(27)	*(12,14)	*(6)	TGCTAG
	C	*(14,16)	AA	*(14,16)	T	*(27)	CT(6,7)		

5.7 Assignment of prey reads to restriction fragments

I used the coordinates of HindIII and DpnII sites in the reference genome to generate an *in silico* list of HindIII-DpnII, HindIII-HindIII, and DpnII-DpnII restriction fragments. Prey reads that overlapped a restriction fragment by at least 20 bp were assigned to that fragment. I then filtered this fragment set to keep HindIII-DpnII fragments that contained at least two reads, were ≤ 1,500 bp in length (as expected from the gel-based size-selection step), were not located within the duplicated FSHD locus on chr4 or chr10 (i.e., > 190,944,164 on chr4, > 135,436,071 on chr10, hg19 coordinates), and were not within a segmental duplication with copy (copies) that mapped within the FSHD locus duplication (the rationale for each filter is discussed in Chapter 3 Section 2). The latter two filters effectively excluded ~200 kb of the terminal 40 Mb of the q arms of these chromosomes from further analysis, or ~72 captured fragments on average (0.15% of the total) from the analysis (see **Table 3.4**).

5.8 Assignment of prey fragments to bait haplotypes

Prey reads whose paired SSLP read was successfully assigned a genotype are referred to as "genotyped". For each prey fragment containing at least ten genotyped reads, I calculated the proportion of reads assigned to each bait haplotype and plotted the distributions of these values (see **Figure 3.7**). Although it is possible that the same prey fragment could be captured by more than

one bait haplotype in the population of nuclei sampled, these plots indicate that many prey are highly enriched for reads matched with the same bait haplotype (see Chapter 3 Section 4 and figures therein for further discussion). I used these distributions to choose cutoff values for each bait haplotype within each dataset, whereby fragments containing greater than or equal to that proportion of genotyped reads for a particular bait haplotype are classified as being "captured" by that haplotype. I chose stringent cutoffs that selected the right-most peaks of these distributions: 0.65 for 4q1 in all libraries; 0.4 for 4q2 and 4q3 in control libraries; 0.55 for 4q2 and 4q3 in FSHD libraries; and 0.75 for 10q in all libraries.

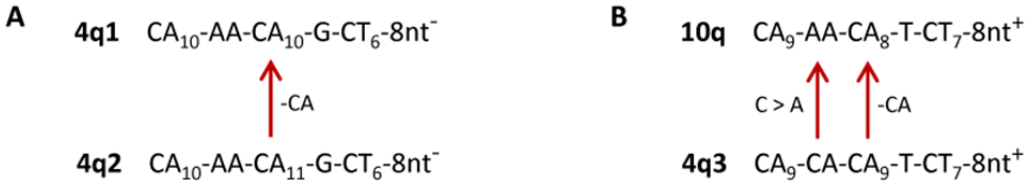

Figure 5.1 Polymerase and sequencing errors lead to SSLP misclassification
The sequences of the SSLP genotypes found in my 4C libraries are shown in simplified form, with numerical subscripts indicating the number of repeats of a given dinucleotide and the plus or minus after "8nt" referring to the presence or absence of an eight-nucleotide insertion. Red arrows demonstrate sequence changes necessary to convert one SSLP into another. **(A)** The 4q2 SSLP is highly similar to 4q1, differing by a single CA repeat. **(B)** Loss of a CA repeat and conversion of a C to an A in the proper position along the SSLP sequence makes 4q3 identical to 10q.

Some of the genotype distributions were tri-modal, where only bi-modal distributions were expected (see **Figure 3.6**). For example, in C2.1, the 4q1 profile has peaks centered at 0.8, 0.3 and 0.05, yet the same profile in F2.1 lacks the peak at 0.3. The presence or absence of this third peak can be explained by the distinctness of the two 4q SSLPs in a given library: the 4q1 and 4q2 SSLPs are very similar, differing by only a single CA, whereas the 4q3 SSLP is much more distinct from 4q1. If a 4q2 SSLP loses a CA repeat, it looks identical to a 4q1 SSLP (**Figure 5.1A**); if this occurs in the reads within a prey fragment that was truly captured by the 4q2 bait, the fragment's 4q2 read proportion would be reduced, but still be greater than the false 4q1 read proportion created by mistaken identity. These false 4q1 reads create the peak at 0.3 in the C2.1 4q1 profile. Since errors in the 4q3 SSLP sequence cannot make it identical to the 4q1 sequence, the 4q1 profile in F2.1 lacks the peak at 0.3. Consequently, the right-most peak in the 4q3 profile in F2.1 is centered around 0.68, whereas the corresponding peak in the C2.1 4q2 profile is centered around 0.55, since the 4q2-captured prey are "contaminated" with false 4q1 reads (though it is possible that some of these prey were also captured by 4q1).

The 10q genotype distribution also shows a difference between control and FSHD libraries, with a peak at 0.3 in F2.1 that is not present in C2.1, which can again be explained by the similarity between certain SSLP sequences: while 4q3 is very distinct from 4q1, 4q3 is similar to the 10q SSLP (though not as similar as 4q1 is to 4q2). If the 4q3 SSLP loses a CA, and PCR or sequencing error changes a CA to an AA in the right position, it looks identical to a 10q SSLP (**Figure 5.1B**). As with false 4q1 reads contaminating 4q2-captured prey described above, false 10q reads would thus contaminate 4q3-captured prey in FSHD libraries, which could explain both the peak at 0.3 in the F2.1 10q profile and why the right-most F2.1 4q3 peak is not as close to 1 as is the right-most F2.1 4q1 peak.

5.9 Analysis of prey captured in multiple libraries

Prey fragments are initially captured as HindIII fragments in my 4C assay, but are subsequently reduced to HindIII-DpnII fragments. Since it is possible to capture both ends of a HindIII fragment (I only view the capture from one side of the bait fragment; see **Figure 3.13**), I used HindIII instead of HindIII-DpnII fragments when comparing prey across multiple libraries. Indeed, this increases the proportion of prey fragments found in more than one 4C library from 30% to 43%.

5.9.1 Association with CTCF sites

For each HindIII prey fragment, I calculated the distance to the nearest CTCF site, defined by the ENCODE Consortium using ChIP-seq in primary human myoblasts[88]. Fragments directly overlapping a CTCF site were given a distance of zero. I then compared the observed distribution of distances to an expected distribution generated by 25 equally-sized random samples of HindIII fragments from the reference genome. Randomly-selected fragments were matched to the GC percentage and chromosomal origin of each observed fragment set. When the cumulative density graph of the observed distances exceeded the graph of the expected distances, I considered those observed prey to be enriched at and near CTCF sites. I focused on distances ≤ 15 kb, which is approximately the span of four HindIII fragments (average size 3.4 kb).

5.9.2 Association with centromeres

I used assembly gaps from the "gaps" track of the UCSC Genome Browser (http://genome.ucsc.edu) that were defined as "centromere" to determine the coordinates of each chromosome's centromere. For each prey fragment, I calculated the distance to the nearest border

of the centromere (i.e., distances for fragments on the p arm are measured to the p-arm side of the centromeric assembly gap) and determined which fragments were ≤ 1 Mb from that gap. I calculated this proportion of prey on non-bait chromosomes found "near" centromeres for different prey-fragment sets; the sets required the prey fragments be found in increasing numbers of libraries.

5.9.3 Repeat content

Short sequence reads derived from highly-similar copies of repetitive elements could create artifactual prey fragments that appear to be present in multiple 4C libraries. In order to assess this issue, I used the "RepeatMasker" track from the UCSC Genome Browser (http://genome.ucsc.edu), which contains the coordinates and identities of repetitive elements in the reference genome. For a given set of prey fragments, I calculated the fraction of bases in the collection that were in repetitive elements, classified by repeat family.

5.10 Sliding window analyses

Since examination of individual prey fragment locations along the genome is an overly-granular means of data analysis, especially for comparing different samples, I binned prey fragments using sliding windows. I arbitrarily selected 1 Mb (500-kb slide) and 500 kb (250-kb slide) as window sizes to examine and calculated coordinates starting from the q-arm end of each chromosome, adjusting windows as necessary to take assembly gaps into account so that each window covered the same number of sequenced bases in the reference assembly. I assigned my prey sets and the reference restriction fragments to windows, placing fragments that overlapped neighboring windows into the window with which they overlapped the most. In order to control for window-to-window differences in total fragment number, I normalized across windows by dividing prey counts by the number of "possible" restriction fragments in each window. This yields a "percent of fragments captured" value for each window, which is plotted at that window's midpoint in my graphs.

5.10.1 Assessment of gene expression and lamin-associated domain (LAD) overlap in peaks & valleys

I used my four most deeply-sequenced libraries (C2.1, C2.2, F2.1 and F2.2) to assess gene expression and LAD overlap within peaks and valleys of their haplotype-assigned, 1-Mb sliding-window profiles. I focused on the terminal 50 Mb of chromosomes 4 and 10, the I found the most

prominent SSLP interactions in these regions. In order to systematically select peaks and valleys, I first calculated a sliding median (M, 15 windows wide) of the "percent of prey captured" (PPC) values for a given prey profile. Then, for each window in that profile, I calculated a value N = PPC – M. Windows above the sliding median (peaks) have N > 1, windows below the sliding median (valleys) have N < 1. I then calculated the median, 25th and 75th percentiles for all N values of the windows in the last 50 Mb of chromosomes 4 and 10 and chose peaks that were higher than the 75th percentile and valleys that were lower than the 25th percentile. To be selected as a peak, a window had to have at least one adjacent window above the median. If multiple adjacent windows (which overlap, since the 1 Mb windows were slid by 500 kb) exceeded the 75th percentile cutoff, I chose the window with the highest N value as the peak. These same criteria were used to pick valley windows below the 25th percentile cutoff.

I examined total gene expression in peak and valley windows using expression values that had been calculated by the ENCODE Consortium using RNA-seq in primary myoblasts[72]. For each gene, I used the "fragments per kilobase per million reads sequenced" (FPKM) value to indicate expression level; this metric attempts to normalize for both size differences between genes and depth of sequencing. I summed the FPKM values of all genes overlapping each of my peak and valley windows to give a single measure of transcriptional output for each window. I then compared the distributions of these values between sets of peaks and valleys, either combined (control peaks added to FSHD peaks, etc.), or separately (control peaks vs. control valleys). Using the average FPKM of genes in each window gave similar results.

I also calculated the fraction of each peak and valley window that overlapped with lamin-associated domains that had been mapped in fibroblasts using the Dam-ID technique[17]. The domain coordinates were obtained from the "NKI Nuclear Lamina Associated Domains (LaminB1 DamID)" track of the UCSC Genome Browser.

With both gene expression and LAD data, I tested for statistically significant differences between various sets of windows (e.g., control 4q1 peaks vs. control 4q1 valleys) using a bootstrapped version of the Kolmogorov-Smirnov test in the "Matching" package of R. This version of the test works when data distributions are not entirely continuous, and is able to handle ties between the two sets being compared.

5.10.2 Identification of significant control vs. FSHD differences

I identified regions where the prey profiles differed significantly between control and FSHD using the three datasets with the largest prey coverage: C2.1, C2.2 and F2.1. This analysis was performed for the prey profiles associated with each bait haplotype in these three 4C libraries. In order to adjust for differences in the number of prey fragments assigned to the bait haplotypes between datasets, in a given comparison, the lowest prey fragment count on each chromosome was used to subset larger datasets so that all sets had the same prey count on each chromosome. For example, 4q1-assigned prey on chromosome 4 numbered 3,159 in C2.1, 3,617 in C2.2 and 3,073 in F2.1, so 3,073 prey were randomly selected from the C2.1 and C2.2 datasets. These prey were then assigned to 1-Mb and 500-kb sliding windows, as described at the beginning of Section 5.9.

For each window, I calculated a "difference score" between the fraction of fragments captured in two samples as the absolute value of the difference between a pair of prey profiles. This yielded three difference scores: C_1C_2 (|C2.1-C2.2|), C_1F (|C2.1-F2.1|) and C_2F (|C2.2-F2.1|). For each chromosome, windows where C_1F *and* C_2F exceeded the 99^{th} percentile value of C_1C_2 were classified as significant. That is to say, the control vs. FSHD difference in a window exceeded the control vs. control differences anywhere on the chromosome. In this way, I found regions of the genome where the 4q1, 4q2/3 or 10q SSLP bait captured prey differently in control vs. FSHD muscle cells. I was conservative in calling significant windows, requiring them to pass my significance test at both 1-Mb and 500-kb resolutions.

5.10.3 Comparison of FSHD prey counts in regions with significant control vs. FSHD differences

If D4Z4 deletions truly alter the nuclear organization of the FSHD locus, changes in the prey captured by the SSLP from the deleted chromosome 4 in FSHD cell lines should not be detected in the prey captured by the other chromosome 4 SSLP. Thus, I reasoned that in those windows where I found a control vs. FSHD difference in 4q1-prey counts, there should be a roughly equivalent difference between 4q1- and 4q3-prey counts in the FSHD samples. So, in a window where the control 4q1 count is higher than the FSHD 4q1 count, the 4q3 count should be higher than the 4q1 count in the FSHD sample.

I normalized the FSHD prey counts in my significant 500-kb windows to adjust for differences in the number of prey assigned to each bait by dividing the count in a given window by the number of prey per thousand prey captured on that chromosome. For example, on chromosome 4 in F2.1 I assigned 3,073 prey to the 4q1 bait and 2,577 to the 4q3 bait. Thus, I divided 4q1-assigned

prey counts in each significant chromosome 4 window by 3.073 and 4q3 counts by 2.577 to be able to compare the two counts to each other.

References

1. GENCODE genes. <http://www.gencodegenes.org/stats.html>.
2. Rajapakse I, Groudine M. On emerging nuclear order. J Cell Biol 2011;192(5):711-21.
3. Misteli T. Concepts in nuclear architecture. Bioessays 2005;27(5):477-87.
4. Misteli T. Beyond the sequence: cellular organization of genome function. Cell 2007;128(4):787-800.
5. Dechat T, Pfleghaar K, Sengupta K, Shimi T, Shumaker DK, Solimando L, Goldman RD. Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. Genes Dev 2008;22(7):832-53.
6. Kind J, van Steensel B. Genome-nuclear lamina interactions and gene regulation. Curr Opin Cell Biol 2010;22(3):320-5.
7. Pederson T. The spatial organization of the genome in mammalian cells. Curr Opin Genet Dev 2004;14(2):203-9.
8. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR and others. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol 2005;3(5):e157.
9. Cremer T, Cremer M. Chromosome territories. Cold Spring Harb Perspect Biol 2010;2(3):a003889.
10. Yokota H, van den Engh G, Hearst JE, Sachs RK, Trask BJ. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. J Cell Biol 1995;130(6):1239-49.
11. Yokota H, Singer MJ, van den Engh GJ, Trask BJ. Regional differences in the compaction of chromatin in human G0/G1 interphase nuclei. Chromosome Res 1997;5(3):157-66.
12. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature 2007;447(7143):413-7.
13. Takizawa T, Meaburn KJ, Misteli T. The meaning of gene positioning. Cell 2008;135(1):9-13.
14. Kosak ST, Skok JA, Medina KL, Riblet R, Le Beau MM, Fisher AG, Singh H. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science 2002;296(5565):158-62.
15. Shopland LS, Lynch CR, Peterson KA, Thornton K, Kepper N, Hase J, Stein S, Vincent S, Molloy KR, Kreth G and others. Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. J Cell Biol 2006;174(1):27-38.
16. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W and others. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet 2004;36(10):1065-71.
17. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W and others. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 2008;453(7197):948-51.
18. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Gräf S, Flicek P, Kerkhoven RM, van Lohuizen M and others. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell 2010;38(4):603-13.
19. Higgs DR, Engel JD, Stamatoyannopoulos G. Thalassaemia. Lancet 2012;379(9813):373-83.
20. Ragoczy T, Bender MA, Telling A, Byron R, Groudine M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. Genes Dev 2006;20(11):1447-57.

21. Bender MA, Ragoczy T, Lee J, Byron R, Telling A, Dean A, Groudine M. The hypersensitive sites of the murine beta-globin locus control region act independently to affect nuclear localization and transcriptional elongation. Blood 2012;119(16):3820-7.
22. Schoenfelder S, Clay I, Fraser P. The transcriptional interactome: gene expression in 3D. Curr Opin Genet Dev 2010;20(2):127-33.
23. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS and others. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet 2010;42(1):53-61.
24. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 2006;38(11):1348-54.
25. Marshall WF, Dernburg AF, Harmon B, Agard DA, Sedat JW. Specific interactions of chromatin with the nuclear envelope: positional determination within the nucleus in Drosophila melanogaster. Mol Biol Cell 1996;7(5):825-42.
26. Trask BJ, Allen S, Massa H, Fertitta A, Sachs R, van den Engh G, Wu M. Studies of metaphase and interphase chromosomes using fluorescence in situ hybridization. Cold Spring Harb Symp Quant Biol 1993;58:767-75.
27. Schermelleh L, Carlton PM, Haase S, Shao L, Winoto L, Kner P, Burke B, Cardoso MC, Agard DA, Gustafsson MG and others. Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy. Science 2008;320(5881):1332-6.
28. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science 2002;295(5558):1306-11.
29. Miele A, Dekker J. Mapping cis- and trans- chromatin interaction networks using chromosome conformation capture (3C). Methods Mol Biol 2009;464:105-21.
30. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. Genes Dev 2012;26(1):11-24.
31. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U and others. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 2006;38(11):1341-7.
32. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C and others. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 2006;16(10):1299-309.
33. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO and others. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326(5950):289-93.
34. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol 2012;30(1):90-8.
35. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F and others. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 2011;43(7):630-8.
36. Bulger M, Groudine M. Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev Biol 2010;339(2):250-7.
37. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J and others. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 2012;148(1-2):84-98.

38. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, He HH, Brown M, Liu XS, Davis M and others. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc Natl Acad Sci U S A 2010;107(21):9742-6.
39. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev 2006;20(17):2349-54.
40. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell 2009;137(7):1194-211.
41. Hou C, Dale R, Dean A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci U S A 2010;107(8):3651-6.
42. Tawil R, Van Der Maarel SM. Facioscapulohumeral muscular dystrophy. Muscle Nerve 2006;34(1):1-15.
43. van der Maarel SM, Frants RR, Padberg GW. Facioscapulohumeral muscular dystrophy. Biochim Biophys Acta 2007;1772(2):186-94.
44. Lemmers RJ, van der Vliet PJ, Klooster R, Sacconi S, Camaño P, Dauwerse JG, Snider L, Straasheijm KR, van Ommen GJ, Padberg GW and others. A unifying genetic model for facioscapulohumeral muscular dystrophy. Science 2010;329(5999):1650-3.
45. van der Maarel SM, Tawil R, Tapscott SJ. Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. Trends Mol Med 2011;17(5):252-8.
46. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 2005;437(7055):94-100.
47. van Geel M, Dickson MC, Beck AF, Bolland DJ, Frants RR, van der Maarel SM, de Jong PJ, Hewitt JE. Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. Genomics 2002;79(2):210-7.
48. Lyle R, Wright TJ, Clark LN, Hewitt JE. The FSHD-associated repeat, D4Z4, is a member of a dispersed family of homeobox-containing repeats, subsets of which are clustered on the short arms of the acrocentric chromosomes. Genomics 1995;28(3):389-97.
49. Lemmers RJ, Wohlgemuth M, van der Gaag KJ, van der Vliet PJ, van Teijlingen CM, de Knijff P, Padberg GW, Frants RR, van der Maarel SM. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. Am J Hum Genet 2007;81(5):884-94.
50. van der Maarel SM, Frants RR. The D4Z4 repeat-mediated pathogenesis of facioscapulohumeral muscular dystrophy. Am J Hum Genet 2005;76(3):375-86.
51. Winokur ST, Chen YW, Masny PS, Martin JH, Ehmsen JT, Tapscott SJ, van der Maarel SM, Hayashi Y, Flanigan KM. Expression profiling of FSHD muscle supports a defect in specific stages of myogenic differentiation. Hum Mol Genet 2003;12(22):2895-907.
52. Gabellini D, Green MR, Tupler R. Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell 2002;110(3):339-48.
53. van Overveld PG, Enthoven L, Ricci E, Rossi M, Felicetti L, Jeanpierre M, Winokur ST, Frants RR, Padberg GW, van der Maarel SM. Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. Ann Neurol 2005;58(4):569-76.
54. de Greef JC, Wohlgemuth M, Chan OA, Hansson KB, Smeets D, Frants RR, Weemaes CM, Padberg GW, van der Maarel SM. Hypomethylation is restricted to the D4Z4 repeat array in phenotypic FSHD. Neurology 2007;69(10):1018-26.
55. de Greef JC, Lemmers RJ, van Engelen BG, Sacconi S, Venance SL, Frants RR, Tawil R, van der Maarel SM. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. Hum Mutat 2009;30(10):1449-59.

56. Tam R, Smith KP, Lawrence JB. The 4q subtelomere harboring the FSHD locus is specifically anchored with peripheral heterochromatin unlike most human telomeres. J Cell Biol 2004;167(2):269-79.
57. Masny PS, Bengtsson U, Chung SA, Martin JH, van Engelen B, van der Maarel SM, Winokur ST. Localization of 4q35.2 to the nuclear periphery: is FSHD a nuclear envelope disease? Hum Mol Genet 2004;13(17):1857-71.
58. Zeng W, de Greef JC, Chen YY, Chien R, Kong X, Gregson HC, Winokur ST, Pyle A, Robertson KD, Schmiesing JA and others. Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). PLoS Genet 2009;5(7):e1000559.
59. Petrov A, Pirozhkova I, Carnac G, Laoudj D, Lipinski M, Vassetzky YS. Chromatin loop domain organization within the 4q35 locus in facioscapulohumeral dystrophy patients versus normal human myoblasts. Proc Natl Acad Sci U S A 2006;103(18):6982-7.
60. Petrov A, Allinne J, Pirozhkova I, Laoudj D, Lipinski M, Vassetzky YS. A nuclear matrix attachment site in the 4q35 locus has an enhancer-blocking activity in vivo: implications for the facio-scapulo-humeral dystrophy. Genome Res 2008;18(1):39-45.
61. Snider L, Asawachaicharn A, Tyler AE, Geng LN, Petek LM, Maves L, Miller DG, Lemmers RJ, Winokur ST, Tawil R and others. RNA transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: new candidates for the pathophysiology of facioscapulohumeral dystrophy. Hum Mol Genet 2009;18(13):2414-30.
62. Snider L, Geng LN, Lemmers RJ, Kyba M, Ware CB, Nelson AM, Tawil R, Filippova GN, van der Maarel SM, Tapscott SJ and others. Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. PLoS Genet 2010;6(10):e1001181.
63. Geng LN, Yao Z, Snider L, Fong AP, Cech JN, Young JM, van der Maarel SM, Ruzzo WL, Gentleman RC, Tawil R and others. DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. Dev Cell 2012;22(1):38-51.
64. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B. Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet 2011;7(3):e1001343.
65. Ren L, Wang Y, Shi M, Wang X, Yang Z, Zhao Z. CTCF mediates the cell-type specific spatial organization of the Kcnq5 locus and the local gene regulation. PLoS One 2012;7(2):e31416.
66. Simonis M, Kooren J, de Laat W. An evaluation of 3C-based methods to capture DNA interactions. Nat Methods 2007;4(11):895-901.
67. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods 2010;7(2):119-22.
68. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754-60.
69. UCSC Genome Browser. <http://genome.ucsc.edu/>.
70. Ohlsson R, Lobanenkov V, Klenova E. Does CTCF mediate between nuclear organization and gene expression? Bioessays 2010;32(1):37-50.
71. UCSC ENCODE data. <http://genome.ucsc.edu/ENCODE/>.
72. CalTech RNA-seq data. <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCaltechRnaSeq>.
73. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W and others. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 2008;453(7197):948-51.

74. Bodega B, Ramirez GD, Grasser F, Cheli S, Brunelli S, Mora M, Meneveri R, Marozzi A, Mueller S, Battaglioli E and others. Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation. BMC Biol 2009;7:41.
75. Pirozhkova I, Petrov A, Dmitriev P, Laoudj D, Lipinski M, Vassetzky Y. A functional role for 4qA/B in the structural rearrangement of the 4q35 region and in the regulation of FRG1 and ANT1 in facioscapulohumeral dystrophy. PLoS One 2008;3(10):e3389.
76. Sachs RK, van den Engh G, Trask B, Yokota H, Hearst JE. A random-walk/giant-loop model for interphase chromosomes. Proc Natl Acad Sci U S A 1995;92(7):2710-4.
77. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 2011;43(11):1059-65.
78. Lanctôt C, Cheutin T, Cremer M, Cavalli G, Cremer T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat Rev Genet 2007;8(2):104-15.
79. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B. Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet 2011;7(3):e1001343.
80. Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, Zhu Y, Kaaij LJ, van Ijcken W, Gribnau J, Heard E and others. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev 2011;25(13):1371-83.
81. De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat Biotechnol 2011;29(12):1103-8.
82. Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 2012;148(5):908-21.
83. Fudenberg G, Getz G, Meyerson M, Mirny LA. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat Biotechnol 2011;29(12):1109-13.
84. Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, Macdonald TY, Tripodi J, Bunting K, Najfeld V and others. Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci U S A 2012.
85. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet 2004;5(5):335-44.
86. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485(7398):376-80.
87. Rajapakse I, Scalzo D, Tapscott SJ, Kosak ST, Groudine M. Networking the nucleus. Mol Syst Biol 2010;6:395.
88. ENCODE Broad Histone data. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/

Appendix A: 4C protocol

Protocol for producing two 4C libraries simultaneously. Cells can be crosslinked and frozen before being used for library production. Use molecular-grade water (MG-H_20).

Day 1a

Solutions

9.5 ml 1% X-link Solution: 950 µl FBS + 7,925 µl PBS + 625 µl 16% Formaldehyde

10% FBS/PBS: 50 µl FBS + 450 µl PBS

5 ml Lysis Buffer (on ice): 50 µl 1M Tris HCl pH 7.5 (10 mM) + 50 µl 1M NaCl (10 mM) +
10 µl NP-40 (0.2%) + 4,890 µl dH_2O + 1/2 protease inhibitor tablet (Roche)

1. Wash cells 2x w/ PBS, trypsinize and count
2. Spin down cells (add formaldehyde to x-link soln.), resuspend in 500 µl 10% FBS/PBS
3. Set aside 500 µl X-link Soln., add 500 µl re-suspended cells to X-link tube through 40 µm strainer, then pipette 500 µl X-link Soln. through strainer
4. Rock X-link tubes for 10' @ RT
5. Place tubes on ice, add 2 x 712 µl ice-cold 1M glycine, mix
6. Spin 8' 225 g @ 4°C, discard super
7. Wash twice w/ 5 ml ice-cold PBS (add protease inhibitor tablet to lysis buffer & dissolve)
8. Resuspend in 5 ml ice-cold lysis buffer in dounce homogenizer, incubate on ice 10' w/ pipetting, then dounce 10x w/ pestle B
9. Spin 5' 400 g @ 4°C, discard super
10. Wash with 500 µl PBS, spin again, pipet off remaining liquid
11. Snap freeze in dry ice and EtOH mix, store at -80 °C

Day 1b

Solutions

1 ml 1X REact2: 100 µl 10X + 900 µl MG-H_2O

2 x 1.2X REact2 @ 37°C: 60 µl 10X + 7.5 µl 20% SDS (0.3%) + 440 µl MG-H_2O

1. Remove pelleted nuclei from -80 °C freezer and thaw for 10' @ RT
2. Wash in 500 µl 1X REact2 buffer, equalize # of cells, spin 5' 400g @ RT
3. Resuspend in 500 µl 1.2 REact2 containing 0.3% SDS (pre-warmed @ 37°C)
4. Incubate 1 hr 900 rpm @ 37°C in Eppendorf Thermomixer (Paddison Lab)
5. Add 50 µl 20% Triton (2%)
6. Incubate 1 hr 900 rpm @ 37°C in Thermomixer
7. Add 40 µl 10 U/µl HindIII (400 U)
8. Incubate overnight 900 rpm @ 37°C in Thermomixer (turn on 65°C waterbath)

Day 2

Solutions

1.41 ml 10x ligation buffer: 197.4 µl MG-H_2O + 930.6 µl 1M Tris-HCl pH 7.5 (660 mM)
+ 70.5 µl 1M DTT (50 mM) + 70.5 µl 1M $MgCl_2$ (50 mM)
+ 141 µl 100mM ATP (10 mM)

1 mg/ml RNase A: 4 µl 10 mg/ml RNase + 36 µl MG-H_2O

9. Take 2 x 20 µl digested control aliquots in 2 ml microtubes
10. Add 40 µl 20% SDS (1.6%), incubate 30' @ 65°C with manual shaking
11. Place samples @ 4°C until qPCR digestion quantification is complete

Day 2, cont.

Process control aliquots

I. Add 500 µl PK Buffer and 10 µl 2 mg/ml PK (20 µg) to control aliquots
II. Incubate 30' @ 65°C in waterbath, then equilibrate @ 37°C (make RNase dilution)
III. Add 1 µl 1 mg/ml RNase A, incubate 2 h @ 37°C in waterbath
IV. Place 250 µl in new 2 ml microtube, add 1,250 µl buffer PBI to each, mix
V. Using 1 QIAquick spin column/aliquot, transfer 800 µl at a time, apply vacuum
VI. Add 750 µl buffer PE to columns, incubate 5', apply vacuum
VII. Spin columns in collection tubes 2' 13,000 rpm
VIII. Put columns in new microtubes, add 30 µl buffer EB, incubate 3', spin 2' 13,000 rpm
IX. Run qPCR digestion quantification using 2 µl digested control / rxn

12. Add 704.4 µl 10x LB to 5,420.6 µl MG-H_2O, 375 µl 20% Triton (1%), add sample, mix
13. Incubate 1 hr @ 37°C with gentle shaking in oven
14. Add 100 U Roche T4 DNA ligase, incubate 4 hr @ 16°C in waterbath then 30' @ RT
15. Add 150 µl 2 mg/ml PK (300 µg), incubate overnight @ 65°C in waterbath

Day 3

Solutions

1.5 ml 10mM Tris 7.5: 15 µl 1M Tris pH 7.5 + 1,485 µl MG-H_2O

2 x 1X Digest Bffr. (2 ml screw-cap tubes): 60 µl DpnII buffer + 395 µl MG-H_2O

16. Add 30 µl 10 mg/ml RNase, incubate 30' @ 37°C in wb, transfer to phase-lock tubes
17. Add 7 ml PCI, mix and spin 15' 2,200g @ RT, transfer super to 50 ml tube
18. Add 7 ml MG-H_2O + 1.5 ml 2M NaOAc pH 5.6 + 35 ml EtOH
19. Incubate 1 hr @ -80°C
20. Spin 45' 2,200g @ 4°C, discard super
21. Add 10 ml 70% EtOH
22. Spin 15' 2,200g @ 4°C, discard super
23. Dry pellet, resuspend in 150 µl 10 mM Tris pH 7.5, dissolve @ 37°C
24. Add 145 µl sample to 1X Digest Buffer, add 1.2 µl 50 U/µl DpnII
25. Incubate overnight @ 37°C (turn on 65°C waterbath)

Day 4

Solutions

2,820 µl 10x ligation buffer: 394.8 µl MG-H_2O + 1,861.2 µl 1M Tris-HCl pH 7.5 (660 mM)
+ 141 µl 1M DTT (50 mM) + 141 µl 1M $MgCl_2$ (50 mM) + 282 µl 100mM ATP (10 mM)

26. Incubate 20' @ 65°C in waterbath
27. Add 600 µl PCI, mix and spin 5' 13,000 rpm
28. Transfer aqueous phase to 2ml microtube, add 1/10 NaOAc and 2 vol. EtOH, mix
29. Incubate 30' @ -80°C, spin 20' 13,000 rpm, discard super
30. Add 1 ml 70% EtOH, spin 5' 13,000 rpm, discard super
31. Air dry pellet, resuspend in 100 µl MG-H_2O, take 5 µl aliquot for gel
32. Transfer to 50 ml tube, add 12.46 ml MG-H_2O + 1.4 ml fresh 10X ligation buffer + 200 U T4 DNA ligase
33. Incubate 4 hr @ 16°C in waterbath, then 30' @ RT
34. Transfer to phase-lock tubes, add 14 ml PCI, mix
35. Spin 15' 2,200g @ RT
36. Split aqueous phase into two 50 ml tubes (7.25 ml each), add 7.25 ml dH_2O + 14.5 µl glycogen + 1.45 ml 2M NaOAc + 35 ml EtOH
37. Incubate overnight @ -80°C

Day 5

38. Spin 45’ 2,200g @ 4 °C, discard super
39. Add 15 ml 70% EtOH
40. Spin 15’ 2,200g @ 4°C, discard super
41. Dry pellets, resuspend each in 75 µl 10mM Tris pH 7.5, dissolve @ 37°C
42. Add 375 µl Buffer PBI to each sample, mix and transfer 150 µl to each of 3 QIAquick spin columns on vacuum manifold, apply vacuum
43. Add 750 µl Buffer PE, incubate 5’, apply vacuum
44. Spin columns in collection tubes 2’ 13,000 rpm
45. Elute each column with 30 µl Buffer EB, incubate 3’, spin 2’ 13,000 rpm; combine identical samples by spinning into one microtube

VITA

Kyle Siebenthall was born and raised in Cloverdale, California. His passion for biology developed in high school, which changed his career trajectory from computers to the life sciences. He attended Cornell University as a Pauline & Irving Tanner Dean's Scholar in the College of Arts & Sciences. He graduated *cum laude* from Cornell in 2006 with a BA in Molecular & Cell Biology. Not content to stay in any one place for too long, he spent his junior year of college abroad at the University of Edinburgh in Scotland.

Kyle's strong desire to conduct biological research led to his joining the laboratory of Dr. Andrew Clark at Cornell during his freshman year. He spent a summer in the lab as a Howard Hughes Scholar after his sophomore year, and completed his honors thesis on metabolic variation in *Drosophila melanogaster* during his senior year. During the summer after his year in Edinburgh, he worked on sequence variation in *Plasmodium falciparum* in the laboratory of Dr. Manolis Dermitzakis at the Wellcome Trust Sanger Institute in Cambridge, England. Kyle's efforts at undergraduate research resulted in authorship on two publications.

In 2012, Kyle earned his PhD in Genome Sciences from the University of Washington after completing his thesis research in the laboratory of Dr. Barbara Trask at the Fred Hutchinson Cancer Research Center.

Kyle's career goal is to be a college professor, so that he can both conduct biological research (in order to inspire himself) and teach courses (in order to inspire others).