ALLELE-SPECIFIC CHROMOSOME CONFORMATION AND ITS ASSOCIATION WITH ALLELIC EXPRESSION BIAS

Rex L. Williams, Jr.

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Genetics and Molecular Biology in the School of Medicine.

Chapel Hill
2013

Approved by:

Kerry S. Bloom
Rosann A. Farber
Terry S. Furey
Terry R. Magnuson
Brian D. Strahl
ABSTRACT

Rex L. Williams, Jr.: Allele-Specific Chromosome Conformation and its Association with Allelic Expression Bias
(Under the direction of Terry R. Magnuson)

It is frequently suggested in recent literature that long-range interactions, presumably between enhancer elements and promoters, are an important factor in mediating gene expression. Such “chromatin loops” are often described as regulatory models in which active or repressive loops are suggested to exist as predominate chromosome conformations that are tightly associated with expression states. These models are supported by numerous investigations of well-studied models such as the beta-globin locus and the Hox clusters. If correct, it follows from these models that differential chromosome conformation patterns, under various contexts, should be a widely observable property of loci that exhibit differential gene expression. To test predictions of this hypothesis, seven candidate genes exhibiting strong allele-specific expression bias were selected for analysis with allele-specific 4C-Seq assays. Conditions for chromosome conformation experiments were optimized for a trophoblast stem cell model that is ideal for interrogating allele-specific chromatin and expression regulation. Results from these experiments demonstrate that the profile of observable interactions may vary between copies. This is important because many genomics assays do not typically detect resolution to discriminate between alleles. The observations made here help to demonstrate that assumption of behavior and modification of loci with genome-wide assays may introduce errors related to inability to distinguish whether events are occurring on both alleles in similar proportions or even at all.
ACKNOWLEDGEMENTS

The work described in this dissertation would not have been possible without a substantial amount of assistance from various people along the way. The author would like to acknowledge, in general, the entire Magnuson Lab for regular feedback and encouragement during the development and execution of the experiments described in the following chapters. In addition to general input and assistance generously provided from all members of the Magnuson Lab, the participation of several people was substantial enough to warrant specific acknowledgements of their contributions.

Dr. Terry R. Magnuson was the principal investigator overseeing this project. In addition to providing funding to carry out the work described here, Dr. Magnuson has been a helpful advisor and long-time friend of the author. Without the advice and encouragement provided from him, it is likely that the author would never have entered a graduate program in the first place and, later, would likely have left the program without completing a doctoral degree.

Dr. Joshua W. Mugford was a Postdoctoral Fellow in the Magnuson Lab at the time that this work was performed. Dr. Mugford initially recruited the author to assist with chromosome conformation experiments in the Magnuson Lab upon learning of the likelihood that the author may leave his doctoral program without completing the degree. As such, Dr. Mugford was heavily involved in the optimization of conditions for performing the described techniques in trophoblast stem cells. Although the initial concept for proposed experiments did not come to fruition, Dr. Mugford’s intervention made the project discussed in the following chapters possible. Additionally, Dr. Mugford was the primary author of many of the Perl scripts and programs used in the mapping, alignment, and allelic assignment procedures described in Chapters 3 and 5. While many of these programs were occasionally modified and tweaked to meet requirements necessary for the specific analyses described here, the framework and design for much of the informatics pipeline leading up to statistical analysis with *fourSig* was primarily performed by Dr. Mugford.
Dr. Joshua Starmer was a Research Assistant Professor under the direction of Dr. Magnuson at the time this work was performed. Dr. Starmer’s initial work on observed allelic bias in expression and chromatin modification detections in the trophoblast stem cell models was the starting point from which the concept of testing allele-specific variations in chromosome conformation was founded. Additionally, Dr. Starmer’s heavy background in statistics and computational biology was crucial as he was the primary designer and developer of the fourSig analysis suite. Dr. Starmer has been heavily involved in the authorship of all manuscripts related to chromosome conformation studies in the Magnuson Lab to date and this work could absolutely not have been performed without his assistance. Furthermore, Dr. Starmer has had substantial involvement in both helping the author to learn how to develop software for informatics analysis and keeping him motivated to complete the work described here and in an associated manuscript.

Della Yee, a technician and the lab manager for the Magnuson Lab, was critical in the preparation of sample for the experiments described here. All culture of trophoblast stem cells was performed by Ms. Yee and she was critical in assisting the optimization of sample collection protocols and actually harvesting the cell pellets used for chromosome conformation experiments. This is laborious work that requires tremendous flexibility of scheduling and her efforts have been greatly appreciated.

Several minor contributions are worthy of specific acknowledgement as well. The described work is founded on initial work performed in the Magnuson Lab by Dr. J. Mauro Calabrese, a Postdoctoral Fellow at the time of this work. His input and guidance was crucial in facilitating effective navigation of the several datasets generated by his research projects.

Dr. Andrew Fedoriw, a Research Assistant Professor under the direction of Dr. Magnuson at the time this work was performed, has an extensive background in genomic imprinting and various molecular biology procedures. Specifically, Dr. Fedoriw was most helpful in offering feedback and suggestions related to allelic expression bias and optimization of conditions for fluorescent in situ hybridization.
Sequencing libraries were submitted to the High-Throughput Sequencing Facility at the University of North Carolina at Chapel Hill. For the specific projects described here and other unrelated chromosome conformation experiments, effective strategies for sequencing the hybrid libraries needed to be developed. The director of this facility, Dr. Piotr Mieczkowski, assisted by providing insightful feedback and suggestions leading to the adapted primer strategy described in Chapter 2.

Finally, it is important to acknowledge Dr. Lucy H. Williams, a former member of the Magnuson Lab and the author’s wife, for consistent encouragement during the course of the author’s graduate work. In addition to being a great contributor of ideas and intellectual assistance, during the course of this work she has had to tolerate numerous intrusions and inconveniences to both her professional work and personal life. Her unwavering support was absolutely critical to the successful completion of the author’s graduate work.
# TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................................ xii

LIST OF FIGURES ........................................................................................................................... xiii

LIST OF ILLUSTRATIONS ............................................................................................................ xv

LIST OF ABBREVIATIONS ............................................................................................................. xvi

LIST OF SYMBOLS ....................................................................................................................... xx

CHAPTER 1: INTRODUCTION ......................................................................................................... 1

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

Three-Dimensional Structure of the Genome ............................................................................ 2

Chromatin in Organization and Regulation ............................................................................ 3

Chromosome Territories and Relative Proximity .................................................................... 4

Chromosome Conformation Assays ........................................................................................ 6

Allele-Specific Gene Expression .............................................................................................. 7

Imprinting ................................................................................................................................. 8

Non-Imprinted Bias ................................................................................................................. 9

Concluding Remarks and Hypothesis ......................................................................................... 9

CHAPTER 2: EXPERIMENTAL DESIGN AND OPTIMIZATION .................................................. 12

Introduction ................................................................................................................................ 12
Future Directions ..................................................................................................................... 98
Possible Improvements ......................................................................................................... 99
Further Experimentation ....................................................................................................... 101
Concluding Remarks ............................................................................................................. 102

CHAPTER 5:  DETAILED METHODS ......................................................................................... 104

Tissue Culture and Sample Preparation ............................................................................. 104

4C-Seq Protocol .................................................................................................................. 104

Isolation of TS Cells ........................................................................................................... 105

Digestion of Chromatin .................................................................................................... 106

Proximity Ligation ............................................................................................................. 108

DNA Purification ............................................................................................................... 108

4C Restriction Digestion ................................................................................................... 111

4C Proximity Ligation and Linearization ........................................................................ 112

Library Amplification ......................................................................................................... 113

4C-Seq Data ....................................................................................................................... 115

Recipes for Solutions ....................................................................................................... 115

FISH Protocols .................................................................................................................. 116

Labeling of FISH Probes ................................................................................................. 116

Probe Precipitation ............................................................................................................ 117

Preparation of Coverslips ................................................................................................. 120
LIST OF TABLES

Table 2.1 Selected Candidates for Allelic 4C-Seq ................................................................. 18
Table 2.2 Primers for Allele-specific qRT-PCR Assays .......................................................... 25
Table 2.3 3C Libraries .............................................................................................................. 37
Table 2.4 4C Digestion Schemes for Candidate Loci ............................................................... 41
Table 2.5 Primers for Allele-specific 4C Assays ................................................................. 42
Table 2.6 4C-Seq Libraries .................................................................................................... 45
Table 3.1 Total Mapped Reads for 4C-Seq Libraries ............................................................... 69
Table 3.2 BACs, Fosmids, and Probes ................................................................................ 80
Table 3.3 Primers for Ibtk cDNA Probe ................................................................................. 82
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>F1 Trophoblast Stem Cells</td>
<td>13</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Process for Generating a 4C-Seq Library</td>
<td>19</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Validation of Allelic Expression Bias in Candidate Genes</td>
<td>28</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Optimization of Fixation Conditions for 3C in TS Cells</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Optimization of Lysis Conditions for 3C in TS Cells</td>
<td>33</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Optimization of Ligation Conditions for 3C in TS Cells</td>
<td>35</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>3C Libraries and Amplification Controls</td>
<td>38</td>
</tr>
<tr>
<td>Figure 2.8</td>
<td>4C-Libraries and Amplification Controls</td>
<td>44</td>
</tr>
<tr>
<td>Figure 2.9</td>
<td>Optimal Permeabilization Conditions for FISH in TS Cells</td>
<td>53</td>
</tr>
<tr>
<td>Figure 2.10</td>
<td>Optimal Fixation Conditions for FISH in TS Cells</td>
<td>55</td>
</tr>
<tr>
<td>Figure 2.11</td>
<td>Minimization of Size for FISH Probes in TS Cells</td>
<td>56</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Reads Distribution for Ibtk Allelic 4C-Seq Test Experiment</td>
<td>63</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Significant Interactions and Allelic Comparison for Ibtk 4C-Seq Test Experiment</td>
<td>64</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Low Proportion of Mappable Reads in Large Allelic 4C-Seq Experiments</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Proportion of Reads Lost to Primer-Dimer Contamination</td>
<td>68</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Mapped Reads Distributions for Allelic 4C-Seq Experiments</td>
<td>71</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Set Analysis of Ibtk 4C-Seq Libraries</td>
<td>73</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Validation of fourSig Prioritization Algorithm .............................................................. 74</td>
<td></td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>Allelic Comparison of Broad Ibtk Significant Contacts ............................................... 76</td>
<td></td>
</tr>
<tr>
<td>Figure 3.9</td>
<td>Allelic Bias in Contact Probability at a Putative Enhancer for Ibtk ............................. 77</td>
<td></td>
</tr>
<tr>
<td>Figure 3.10</td>
<td>Probe Design for Validation of Interactions by FISH .................................................. 79</td>
<td></td>
</tr>
<tr>
<td>Figure 3.11</td>
<td>Validation of Selected Interactions by FISH ............................................................... 83</td>
<td></td>
</tr>
<tr>
<td>Figure 3.12</td>
<td>Set Analysis of Airn 4C-Seq Libraries ........................................................................ 86</td>
<td></td>
</tr>
<tr>
<td>Figure 3.13</td>
<td>Allelic Comparison of Broad Airn Significant Contacts ............................................. 87</td>
<td></td>
</tr>
<tr>
<td>Figure 3.14</td>
<td>Set Analysis of Car2 and Bphl 4C-Seq Libraries ....................................................... 89</td>
<td></td>
</tr>
<tr>
<td>Figure 3.15</td>
<td>Allelic Comparison of Broad Significant Contacts for Car2 and Bphl ........................ 91</td>
<td></td>
</tr>
<tr>
<td>Figure 3.16</td>
<td>Locus View of Interaction Profiles for Car2 and Bphl ............................................... 92</td>
<td></td>
</tr>
</tbody>
</table>
LIST OF ILLUSTRATIONS

Illustration 2.1 Allelic 4C-Seq Candidate Suitability ............................................................... 16

Illustration 2.2 A Method for Determining Enriched Interactions ........................................... 48

Illustration 2.3 Prioritization of Enriched Contacts by Likelihood of Reproducibility .................. 50

Illustration 3.1 Strategy for Mapping 4C-Seq Reads ................................................................ 60
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C</td>
<td>Chromosome conformation capture</td>
</tr>
<tr>
<td>3-D</td>
<td>Three-dimensional</td>
</tr>
<tr>
<td>4C</td>
<td>Circular Chromosome Conformation Capture</td>
</tr>
<tr>
<td>4C-Seq</td>
<td>Circular Chromosome Conformation Capture with Sequencing</td>
</tr>
<tr>
<td>5C</td>
<td>Chromosome Conformation Capture Carbon Copy</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Chromosome</td>
</tr>
<tr>
<td>B6</td>
<td>C57BL/6 Mouse Strain (<em>M. m. domesticus</em>)</td>
</tr>
<tr>
<td>bp</td>
<td>Base Pairs</td>
</tr>
<tr>
<td>CAST</td>
<td>CAST/Eij Mouse Strain (<em>M. m. castaneous</em>)</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin Immunoprecipitation</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Chromatin Immunoprecipitation with Sequencing</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine-phosphate-Guanine</td>
</tr>
<tr>
<td>cm</td>
<td>Centimeter</td>
</tr>
<tr>
<td>CSK</td>
<td>Cytoskeletal Buffer</td>
</tr>
<tr>
<td>CTCF</td>
<td>CCCTC-binding Factor</td>
</tr>
<tr>
<td>Cy</td>
<td>Cyanine</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>DAPI</td>
<td>4',6-Diamidino-2-Phenylindole</td>
</tr>
<tr>
<td>dCTP</td>
<td>Deoxycytidine Triphosphate</td>
</tr>
<tr>
<td>dUTP</td>
<td>Deoxyuridine Triphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DNase</td>
<td>Deoxyribonuclease</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double-Stranded Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>FAIRE</td>
<td>Formaldehyde-Assisted Isolation of Regulatory Elements</td>
</tr>
<tr>
<td>FB</td>
<td>Fetal Brain</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal Bovine Serum</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescent <em>in situ</em> Hybridization</td>
</tr>
<tr>
<td>FL</td>
<td>Fetal Liver</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>Histone H3 Lysine 4 Monomethylation</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Histone H3 Lysine 4 Dimethylation</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Histone H3 Lysine 4 Trimethylation</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Histone H3 Lysine 27 Acetylation</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>Histone H4 Lysine 20 Monomethylation</td>
</tr>
<tr>
<td>ICE</td>
<td>Imprinted Control Element</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>LCR</td>
<td>Locus Control Region</td>
</tr>
<tr>
<td>LN$_2$</td>
<td>Liquid Nitrogen</td>
</tr>
<tr>
<td>M</td>
<td>Molar</td>
</tr>
<tr>
<td>m</td>
<td>Meter</td>
</tr>
<tr>
<td>mM</td>
<td>Millimolar</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
</tr>
<tr>
<td>mm9</td>
<td>Mouse Genome Sequencing Consortium Genome Assembly Version 37</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>ng</td>
<td>Nanogram</td>
</tr>
<tr>
<td>nm</td>
<td>Nanometer</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate Buffered Saline</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PFA</td>
<td>Paraformaldehyde</td>
</tr>
<tr>
<td>pol II</td>
<td>RNA Polymerase II</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative Polymerase Chain Reaction</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Quantitative Reverse Transcription Polymerase Chain Reaction</td>
</tr>
<tr>
<td>RCF</td>
<td>Relative Centrifugal Force</td>
</tr>
<tr>
<td>RFLP</td>
<td>Restriction Fragment Length Polymorphism</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RNase</td>
<td>Ribonuclease</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA Sequencing</td>
</tr>
<tr>
<td>RPM</td>
<td>Revolutions Per Minute</td>
</tr>
<tr>
<td>RPMI</td>
<td>Roswell Park Memorial Institute medium</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse Transcriptase</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse Transcription Polymerase Chain Reaction</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium Dodecyl Sulfate</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SSC</td>
<td>Saline-Sodium Citrate Buffer</td>
</tr>
<tr>
<td>TAD</td>
<td>Topologically Associated Domain</td>
</tr>
<tr>
<td>T4</td>
<td>Enterobacteria Phage T4</td>
</tr>
<tr>
<td>TE</td>
<td>Tris-EDTA Buffer</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TS</td>
<td>Trophoblast Stem</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Site</td>
</tr>
<tr>
<td>UCSC</td>
<td>The University of California, Santa Cruz</td>
</tr>
<tr>
<td>UNC</td>
<td>The University of North Carolina at Chapel Hill</td>
</tr>
<tr>
<td>µg</td>
<td>Microgram</td>
</tr>
<tr>
<td>µL</td>
<td>Microliter</td>
</tr>
<tr>
<td>µm</td>
<td>Micrometer</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS

3' Three Prime

5' Five Prime

C(t) Threshold Cycle

Gapdh Glyceraldehyde-3-phosphate (mouse gene)

Hbb-b1 Hemoglobin, beta adult major chain (mouse gene)

Ibtk Inhibitor of Bruton agammaglobulinemia tyrosine kinase (mouse gene)

NFx2 No Feeders in Culture for 2 Passages

W Window Size in fourSig program

X Significance threshold calculated by fourSig program

Δ Difference

°C Degrees Celsius
CHAPTER 1: INTRODUCTION

Introduction

Studies of how chromatin is organized have suggested that the three-dimensional (3-D) architecture of chromosomes, exhibited as both long- and short-range spatial interactions of genomic elements, may have a profound effect on gene expression. Variations of chromosome conformation capture (3C) have been used to demonstrate that specific loci, such as enhancers and promoters, can be separated by megabases (Mb) of linear sequence or, in some cases, located on different chromosomes yet are arranged in closer spatial proximity than intervening loci (Dekker et al. 2002; Dostie et al. 2006; Simonis et al. 2006; Zhao et al. 2006; Fullwood et al. 2009; Lieberman-Aiden et al. 2009). Such interactions have been associated with specific biological contexts including specificity of cell identity (Ferraiuolo et al. 2010) and response to extrinsic cues (Fullwood et al. 2009). These observations have led to the suggestion that higher-order chromatin conformation may be indicative of or directly involved in normal cellular processes. It has also been suggested that chromatin conformation may be altered in disease states, signifying its value as either a potential therapeutic target or as a biomarker (Crutchley et al. 2010). Therefore, understanding the dynamics of 3-D chromosomal architecture as it relates to gene expression is an important area of molecular biology whose exploration, to date, has been largely limited to a handful of model loci.

The proposed project and subsequently performed work described here will focus on exploring the role of higher-order chromatin conformation in genes that exhibit allelic expression bias. Parental expression bias, generally known as imprinting, has long been known to be important during embryonic development and misregulation of imprinted genes can lead to a variety of developmental disorders (Horstemke and Wagstaff 2008; Ideraabdullah, Vigneau, and Bartolomei 2008). Non-imprinted genes, however, have also been shown to sometimes exhibit allelic expression biases that may or may not be dependent on parental origin but continue to maintain allelically skewed
expression patterns in descendent daughter cells (Yan et al. 2002; Knight et al. 2003; Pandey et al. 2013). Previous work in the Magnuson Laboratory established the use of mouse trophoblast stem (TS) cells as a model system for studying various expression and epigenetic properties with allele specificity (Calabrese et al. 2012, see Chapter 2). Preliminary analysis of these datasets has revealed several non-imprinted genes that exhibit allelic biases in both their expression levels and in their occupancy of various chromatin modifications.

The joining of chromosome conformation experiments with the study of allelic bias in expression and chromatin landscapes represents an opportunity to apply relatively uncommon methodologies to an understudied biological phenomenon. The goal in the work described here was to use genes known to exhibit allelic expression bias as candidates to interrogate the allele-specific nature of chromosome conformations. The results (1) illuminate the degree to which chromosome conformation varies between copies of the same gene and (2) apply a basic interrogation of the hypothesis that chromosome conformation is functionally associated with control of gene expression.

Three-Dimensional Structure of the Genome

The nucleus is the compartment of the cell in which genetic material is stored, organized, and regulated for production of biomolecules, such as proteins and nucleic acids, necessary for the regular function of cellular processes. Genetic material, typically manifested as deoxyribonucleic and ribonucleic acids (DNA and RNA, respectively), contains inheritable units known as genes, which are generally considered stretches of inheritable genomic sequence that are associated with or contain some functional activity or information (Pearson 2006). In eukaryotes, genetic material exists as chromosomes, large structures comprised primarily of a long stretch of DNA coiled around histone octomers to create a series of nucleosomes. The mouse genome contains 40 chromosomes, which the most recent assembly (Dec. 2011 (GRCm38/mm10)) describes as having approximately 5.2 billion base pairs (bp) (Meyer et al. 2013). Assuming 0.34 nanometers (nm) per nucleotide (Watson and Crick 1953), one can calculate that, stretched out, the mouse genome contains approximately
1.77 meters (m) of DNA. When one considers that the average diameter of a mammalian nucleus is only around 6 micrometers (µm) (Alberts et al. 2002), that 1.77 m of DNA must be condensed by numerous packaging proteins, and that various regulatory factors must be free to access and dissociate from the DNA, it becomes clear that the nucleus is a very crowded environment and that efficient organization schemes must be in place for proper function to occur.

**Chromatin in Organization and Regulation**

The condensation of chromosomes varies with cell cycle, ranging from a highly compact state during mitosis to a more dynamic, though still occasionally highly condensed, state (i.e., chromatin) during the interphases (Rajapakse and Groudine 2011). In both states, the DNA polymer is wrapped around histone octomers, known as nucleosomes, which in turn are coiled or supercoiled into various secondary and tertiary structures depending on the phase of the cell cycle or current requirements for access. This ability to vary the condensed states of a chromosome is important for good organization and the malleability of this organization is necessary for functional regulation.

Genes are typically under tight control by regulatory mechanisms that ultimately determine the presence, absence, and abundance of most of the biomolecules within a cell. Consequently, the regulation of gene expression and replication requires that the nucleosomes be accessible to various regulatory proteins, RNAs, and other small molecules during all phases of the cell cycle (Beck et al. 2010). Two features central to the control of how tightly chromatin is compacted are histone modifications and DNA methylation. Histone octomers are comprised of 2 copies of the core histone proteins (H2A, H2B, H3, and H4), which can be covalently modified at many residues by various nuclear complexes (Chi, Allis, and Wang 2010). It has been proposed that the combinatorial modification of histones within a nucleosome may serve as a “code” for initiating or suppressing various nuclear processes, including gene expression and replication (Strahl and Allis 2000). Indeed, many studies over the past decade have provided substantial evidence that modifications, such as methylation and acetylation, of specific residues on histone tails influence nuclear processes by modulating nucleosome density and serving as docking sites to facilitate association of regulatory complexes with functional elements in the DNA (Voigt and Reinberg 2011). In addition to histone
modifications, methylation of the phosphate backbone between cytosines and guanines (CpG) has been well established as a mechanism for repressing functional regulation of DNA (Jones 2012). DNA methylation is traditionally associated with long-term silencing of genes that is maintained through cell divisions.

Beyond histone modification and DNA methylation, which lay the framework for association with regulatory factors, chromatin dynamics are influenced directly and indirectly by many other biomolecules. Chromatin remodeling factors may slide nucleosomes around to modify accessibility to DNA binding motifs (Hargreaves and Crabtree 2011). Non-coding RNAs, messenger RNAs (mRNA) that do not code for proteins, may block access to regulatory sites by hybridization or association with bound factors (Bierhoff, Postepska-Igielska, and Grummt 2013; Roberts, Morris, and Weinberg 2013). Additionally, various binding factors and structural proteins, such as the CCCTC-binding factor (CTCF) and the nuclear lamina, may play roles in compartmentalizing chromatin into spatial domains with differing propensities for active regulation and repression (Bickmore and van Steensel 2013). These mechanisms and others, many of which likely remain currently unknown, all work in combination to create a well-organized and highly dynamic system conducive to both packaging of genetic material and functional accessibility; a concert whose fine-tuning and plasticity is essential for healthy cellular function.

Chromosome Territories and Relative Proximity

Interphase chromosomes have long been observed to occupy “territories” that can be observed by fluorescent microscopy (Heard and Bickmore 2007; Cremer and Cremer 2010). Simply put, when nuclei are observed during interphase, fluorescent probes hybridizing to the same chromosome tend to occupy a specific region with minimal encroachment into the space of other chromosomes. Several studies have indicated that this organization could have regulatory consequences. For example, many experiments performed using fluorescent in situ hybridization (FISH) have consistently shown a tendency for actively expressed genes to localize to the exterior of their chromosome’s territory while repressed genes tend to be sequestered to the interior of the occupied space (Heard and Bickmore 2007; Morey, Kress, and Bickmore 2009). Such observations
have led to the hypothesis that genes whose expression are needed in high abundance tend to be moved to regions containing a higher local concentration of transcriptional machinery, called “transcription factories”, while genes not needed may be moved to the interior of the chromosome to reduce their potential for expression (Bartlett et al. 2006; Sexton et al. 2007; Papantonis et al. 2010). These transcription factories are known to occur for all three RNA polymerases and contain multiple, active loci. Recruitment into specialized transcription factories, which may contain tissue-specific transcription factors, may be a strategy for regulation of genes transcribed by RNA polymerase II (pol II). For example, active alleles of *Hbb-b1* and *Eraf* are found in nuclear foci enriched for RNA pol II in erythroid cells where they are highly transcribed (Osborne et al. 2004). Consistent with this idea, experiments providing an unbiased, genome-wide perspective of chromosomal interactions show that actively transcribed domains tend to associate with other actively transcribed domains, while inactive loci tend to cluster in an analogous manner (Lieberman-Aiden et al. 2009).

It remains unclear, however, whether this type of organization is directed by certain sequences or factors and whether it is a general feature of transcriptional regulation. It is likely that the formation of territories is simply a natural consequence of macromolecular crowding. Biomolecules are large and the nucleus, estimated at having a total density of more than 600 mg/mL (Handwerger, Cordero, and Gall 2005), is a very crowded place. Polymers under such confined constraints are known to become restricted to their own space, separating themselves from other polymers within the confined volume (Richter, Nessling, and Lichter 2008). This natural behavior may explain a number of observed nuclear phenomena, including association of active regions with other active regions (Lieberman-Aiden et al. 2009) and preferential associations with the interior or periphery of territories (Morey, Kress, and Bickmore 2009). Therefore, observations of how loci associate with one another in 3-D space provide potentially valuable information regarding possible regulatory effects on nuclear processes.
Chromosome Conformation Assays

3C is an experimental strategy that identifies the tendency of specific sequence elements in the genome to be co-localized in 3-D space, independent of their linear positioning along a chromosome. Initially developed by Job Dekker in 2002 to study general 3-D associations in the yeast genome (Dekker et al. 2002), the technique quickly became very popular as a tool to study potential effects of distal regulatory elements on gene expression. The genes in the mammalian beta-globin locus, long known to be regulated by a centralized locus control region (LCR), have been a target for extensive study by various techniques based on 3C. A wealth of data at this locus has demonstrated that, not only can genes be influenced by regulatory elements located more than 100 kilobases (kb) away, but that the presence of specific long-range interactions can be tightly associated with expression states indicative of cell identity (Tolhuis et al. 2002; Dostie et al. 2006; Baù et al. 2011). Furthermore, some studies have suggested that long-range interactions can occur on an even more dramatic scale. Interactions of regions separated by more than a Mb of sequence have been observed in association with V(D)J-recombination (Skok et al. 2007; Jhunjhunwala et al. 2008) and, although more controversial, several genes have been reported to regularly engage in significant interactions with regions on different chromosomes (Würtele and Chartrand 2006; Zhao et al. 2006; Robyr et al. 2011). These data, along with many other studies, have provided growing support for a hypothesis that spatial organization of chromosomes may influence expression regulation by positioning distant regulatory elements in close spatial proximity to the regulated gene.

Several methods for detecting these interactions have been used over the years, leading to different types of interaction information. All strategies based on 3C, however, employ the same basic library design: fixation, chromatin fragmentation by mechanical disruption or restriction digestion, proximity ligation of fragmented chromatin complexes, reversal of cross-links, and DNA purification. The key step in this process is the proximity ligation, which results in the ligation of DNA fragments that are tethered together by the binding of protein complexes. The newly ligated “hybrid sequences” are indicative of chromosomal regions that were in close proximity in the nucleus at the time of fixation. Basic 3C experiments employ the use of primers specific to individual restriction
fragments formed with the initial fragmentation (Dekker et al. 2002; Tolhuis et al. 2002). Primers are situated such that they will amplify the ligated hybrid sequences from a 3C library but not genomic DNA (gDNA, See Chapter 2). Consequently, this technique provides only information regarding the interaction of two pre-determined loci. Later, circular chromosome conformation capture (4C) was developed, which employs the use of re-digestion of the 3C library followed by circular ligation of the resulting fragments (Simonis et al. 2006; Zhao et al. 2006). Using microarray detection, amplification from primers in opposing orientation can be used for a specific fragment to gain information on unknown interacting loci with a desired locus. Additionally, in a variation known as chromosome conformation capture carbon copy (5C), microarrays and sequencing were used to detect libraries generated by amplifying conformation libraries with primers for several fragments within a 400 kb region in the human beta globin locus to gain information on all interactions within the (Dostie et al. 2006). Finally, though generally yielding a low resolution, the advent of high-throughput sequencing has allowed the development of 3C variations that can detect interactions occurring throughout the genome without regard to specific loci or regions (Lieberman-Aiden et al. 2009; Fullwood et al. 2009). In summary, the 3C assay was a substantial innovation that has led to the development of a plethora of tools for observing spatial associations of genomic loci within the nucleus.

**Allele-Specific Gene Expression**

Over the past 2 decades, high-throughput detection methods, such as microarrays and sequencing, have greatly facilitated the ability to observe the molecular landscape of the nucleus and the regulation of gene expression. In addition to isolating RNA transcripts, techniques such as chromatin immunoprecipitation (ChIP), formaldehyde-assisted isolation of regulatory elements (FAIRE), and DNA hypersensitivity assays have been used to generate libraries for high-throughput detection that provide copious amounts of data describing regulatory states and processes within the nucleus (Furey 2012). However, most studies using these techniques to date do not discriminate between alleles, allowing potentially interesting information about copy-specific differences in regulation and expression to go undiscovered. Although the advent and continuous improvement of
high-throughput sequencing offers an increasingly cost-effective platform for discriminating between alleles, the degree to which allelic expression bias occurs within the genome and how allelic differences in chromatin regulation are associated with them is still relatively unexplored. Considering the dynamic nature and diverse mechanisms involved in nuclear processes, understanding the degree to which different alleles behave differently is an important area of molecular biology.

Imprinting

It has long been known that maternal and paternal copies of the genome are not equivalent and that both are required for viability (Surani and Barton 1983; McGrath and Solter 1984; Surani, Barton, and Norris 1984). Allelic differences in gene expression and regulation have been documented and extensively studied at well-known loci and discriminatory expression patterns between alleles have most traditionally been associated with genomic imprinting, an epigenetic phenomenon in which either the paternal or maternal copy of a gene is designated to be silenced in the germline (Koerner and Barlow 2010).

In classical imprinting, several genes in the imprinted locus are under the control of a differentially methylated region known as the imprinted control element (ICE). The methylation state of the ICE typically instigates a mechanism that controls the expression state of all genes in the locus that are under imprinted control (Bartolomei 2009). Extensive study of the H19-Igf2 region has led to a detailed picture of how imprinted regulation can be implemented, which, at this locus, involves a pattern of allele-specific chromosome looping that sets up active and repressive domains in 3-D space (Bartolomei, Zemel, and Tilghman 1991; Murrell, Heeson, and Reik 2004; Zhao et al. 2006; Ling et al. 2006; Kurukuti et al. 2006; Engel et al. 2008; Nativio et al. 2009). Additionally, the Kcnq1 promoter, another widely studied imprinted locus, has been shown to preferentially interact with several nearby regulatory elements when active (Korostowski et al. 2011; Korostowski, Sedlak, and Engel 2012) while the repressed paternal copy of the locus may exhibit a preferential nucleolar association (Fedoriw et al. 2012). These observations help to underscore the potential involvement of spatial organization in establishing and maintaining allelic differences in expression regulation.
Non-Imprinted Bias

In addition to imprinting, in which the vast majority of expression comes from a single allele, genes with less extreme allelic biases and biases unrelated to parent-or-origin have been associated with disease states including schizophrenia, obesity, and colorectal cancer (Sutherland and Walley 2009). These forms of allele-specific expression have been reported to affect as much as 12% of detected transcripts in laboratory mouse strains (Keane et al. 2011). The existence of allele-specific expression bias implies the presence of differential regulatory schemes between copies; schemes that may be potentially related to spatial organization. Indeed, two recent studies have shown that chromosome conformation can be differentially associated with loci exhibiting allelic expression bias in both human colorectal cancer (Pittman et al. 2010) and lymphoblastoid cell lines (Cheung et al. 2010). Beyond potential involvement in imprinted regulation, these recent observations of widespread allelic expression bias in multiple cell types suggest that allele-specific 3-D architecture of chromosomes may be a general property of gene regulation throughout the genome. Further exploration of the association of higher-order conformation with expression patterns must be conducted in an allele-specific manner to verify whether this is the case.

Concluding Remarks and Hypothesis

Models of gene expression regulation have consistently evolved since their conceptual introduction over half a century ago (Yaniv 2011). In eukaryotic systems, it is generally accepted that expression of protein-coding genes occurs when the RNA pol II initiation complex, facilitated by the binding of specific proteins to regulatory elements such as enhancers, binds the promoter of a target gene and begins transcription (Ma 2011). Methods for controlling the selectivity of expression have been demonstrated by the study of epigenetics, which has illuminated the employment of histone modifications (Chi, Allis, and Wang 2010), DNA methylation (Jones 2012), and chromatin remodeling (Hargreaves and Crabtree 2011) as regulators of accessibility of genes to the transcriptional machinery. The roles of these influences highlight an increasingly complex system of expression
It is frequently suggested in recent literature that long-range interactions, presumably between enhancer elements and promoters, are an important factor in mediating gene expression (Felsenfeld and Dekker 2012). Such “chromatin loops” are often described as regulatory models in which active or repressive loops are suggested to exist as predominate chromosome conformations that are tightly associated with expression states (Krivega and Dean 2012). These models are supported by numerous investigations of well-studied models such as the beta-globin locus (Dostie et al. 2006; Baù et al. 2011) and the Hox clusters (Ferraiuolo et al. 2010; Lee et al. 2010; Wang et al. 2011). If correct, it follows from these models that differential chromosome conformation patterns, under various contexts, should be a widely observable property of loci that exhibit differential gene expression.

Outright deletion of a chromosomal region to test the validity of this phenomenon as a general property would create an uncertainty of affects due to possible changes in physical dynamics of the entire locus as well as the possibility for eliminating the effects of potentially unrelated protein-DNA interactions (Mirny 2011). Additionally, conformational differences between cell types exhibiting divergent expression of a target gene could be attributed to possible differences in nuclear organization between different cell types rather than related to the direct regulation of gene expression at target loci (Tolhuis et al. 2002; Simonis et al. 2006). Therefore, the identification of genes that exhibit regular and predictable allelic bias in expression and chromatin regulation within the same cell would offer an excellent opportunity to test the hypothesis that the 3-D spatial organization of a locus is a functionally associated property of differential gene expression. Furthermore, such a scenario would have the added benefit of being a well-controlled and minimally perturbed system. If correct, such a hypothesis would predict that if two copies of the same gene exhibit differential expression in a single cell type, then differences in long-range interactions should be observed between the copies of the two loci. It would also follow from this prediction that if the observed allele-specific differences in long-range interactions are functionally associated with the
regulation of gene expression, then the same interactions should be observed in association with the corresponding expression state whenever it is exhibited. The work proposed (Chapter 2) and described (Chapter 3) will aim to test these predictions.
CHAPTER 2: EXPERIMENTAL DESIGN AND OPTIMIZATION

Introduction

The 3-D context of spatial organization as it pertains to the regulation of gene expression was introduced in the previous chapter. Predictions of the proposed hypothesis may be tested by observing the existence of long-range interactions with allele-specific resolution in the context of allelically biased gene expression. If interactions between distant loci, presumably between gene promoters and distal regulatory elements, are a common and general phenomenon in expression regulation, then consistent and specific interactions should be present and detectable by 4C with sequencing (4C-Seq). Furthermore, when copies of the same gene exhibit differential expression patterns, it would be expected that such interactions should maintain strong correlations with the associated expression state. Simply put, when two copies of the same gene exhibit divergent expression states, any distal interactions that may be functionally associated with active expression should be observed at the active locus but largely absent from the repressed allele. Therefore, it was proposed that 4C-Seq be performed in an allele-specific manner using the promoters of genes showing a consistently strong bias of allelic expression as viewpoints.

TS cells derived from the F1 offspring of *M. m. castaneous* (CAST/EiJ, “CAST”) and *M. m. domesticus* (C57Bl/6, “B6”) mice were established in the Magnuson Lab as part of an earlier genomic study of allele-specific expression and epigenetic regulation (Calabrese et al. 2012). Stable lines were established from crosses in which CAST alleles were maternally contributed and B6 alleles were from paternal origin (Figure 2.1). Similarly, lines were established from the offspring of reciprocal crosses yielding paternally derived CAST alleles and maternal contribution from B6. Additionally, lines containing homozygous parental contribution (isotype lines) were established for both strains. Many genomic variations between these mouse strains have been discovered and reported from deep sequencing and annotation projects for the mouse genome (Keane et al. 2011).
Figure 2.1 F1 Trophoblast Stem Cells. Trophoblast Stem (TS) cells were derived by selectively growing trophoblast cells from blastocyst outgrowths. F1 lines were made by harvesting blastocysts from reciprocal matings of B6 and CAST strain mice. TS cell lines were also made from homozygous crossings (not depicted).
Among these, a large number of single nucleotide polymorphisms (SNPs) allow for effective discrimination of allelic origin by high-throughput sequencing. Therefore, as demonstrated by Calabrese et al. (2012), the TS cells established in the Magnuson Lab can serve as a model system for identifying allele-specific differences in genomic analyses.

In order to proceed with the plan to investigate allele-specific long-range interactions, several experimental conditions and tools needed to be worked out. With the F1 TS cells as a model system and the data from Calabrese et al. (2012) as a starting point, this chapter details the development of the experimental design for performing 4C-Seq at genes exhibiting allelic bias in gene expression.

**Strategy and Experimental Design**

4C is a locus-centric chromosome conformation technique; meaning that genome-wide interactions are captured with respect to a single locus known as the viewpoint. Therefore, it was necessary to focus the experimental design on a handful of candidates that were ideal for performing the assay. Additionally, the protocol for the 4C assay needed to be optimized to work with ideal lysis conditions determined for TS cells. Optimization of these conditions was a necessary first step for designing the experimental strategy.

**Candidate Selection**

Several candidate genes needed to be selected to serve as viewpoints for testing the role of long-range interactions in allele-specific expression bias by 4C-Seq. RNA-seq was previously performed in two TS cell lines derived from F1 crosses of CAST × B6 (CASTB6F1) and B6 × CAST (B6CASTF1) (Calabrese et al. 2012). Normalized allele-specific expression measurements (RPKM) were tested using a Fisher’s Exact test for deviation from equimolar detection of CAST and B6 transcripts to determine bias in expression. A false discovery rate (FDR) of 0.05 was used as a cutoff to score a gene as being allelically biased in expression, thereby allowing for the organization of genes into two groups: biased and unbiased.
Genes were then further categorized by identified allelic differences in chromatin regulation. Calabrese et al. (2012) also performed allele-specific DNase-seq, FAIRE-seq, and ChIP-seq experiments for H3K4me2, H3K27me3, H3K36me3, H4K20me1, pol II, and CTCF. Allelic biases in distribution were determined for these datasets in the same manner as was done for the RNA-seq data. Genes were categorized as having allelically biased chromatin signatures by assessing whether any biases in signal for these datasets were present within 2.5 kb of a transcription start site (TSS) of a gene. For this analysis, annotated TSS were taken from the UCSC Known Genes Table (Hsu et al. 2006). A list of candidate genes for interrogation by 4C-Seq was generated from the intersection of genes exhibiting biased expression and those with allelically biased chromatin signatures.

The list of candidate genes generated from the selection described above still contained several hundred genes, so two further filtrations were used to select final candidate loci for the 4C-Seq experiment. First, any gene whose promoter is to serve as a viewpoint must be suitable for performing allele-specific 4C. This requires several sequence features to be present in the initial 3C restriction fragment containing the TSS (Illustration 2.1). In order to determine allelic origin of the viewpoint, a known SNP must be present within the sequenced portion of the viewpoint fragment. This SNP may be associated with either the initial digestion site (3C site) or the secondary restriction site (4C site), which must be located between 140-700 bp of the 3C site (See “Generation of 4C Libraries”). Additionally, the identifying SNP should be as close as possible to an end of the targeted 4C viewpoint fragment. Since the intention was to use of 100 bp reads, and at least 20 bp were to be reserved for the 4C primer sequence, the requirement for SNP placement was set at being within a maximum of 60 bp of the end of the viewpoint fragment. This leaves at least 20 bp to be sequenced from the captured fragment, however shorter distances from SNP to viewpoint end are preferable to ensure maximum mappability and higher likelihood of detecting an informative SNP in the captured fragment. Further, the identifying SNP must also be flanked by an 8- or 6-base restriction site that lies between the ends of the targeted viewpoint fragment. This site may only cut once in the viewpoint fragment as it will be used to linearize the 4C product for amplification. These elements and their required constraints create sufficient spacing to allow for re-circularization of the 3C product,
Illustration 2.1 Allelic 4C-Seq Candidate Suitability. Several criteria are required for suitability of a locus to perform allele-specific 4C-seq assays. “3C” represents the restriction site used for initial fragmentation of the 3C library. “4C” represents the restriction site recognized by the enzyme used to generate the 4C library. “L” represents the motif recognized by the restriction enzyme used to linearize the circularized 4C intermediate.
linearization of the circularized DNA, and amplification of the captured fragment and known SNP.

One concern was that genes important for similar processes and roles may be subject to similar types of regulatory control. In specific, if two allelically biased genes were to be involved in the same pathway then it is possible that similar factors may be driving association with distal regulatory elements. This scenario could be envisioned as genes required for the same pathways being recruited to a transcription factory that is heavily involved with the regulation of the associated pathway. Since the goal was to test the existence of functional association with long-range interactions as a general property of gene expression, it was deemed preferable to avoid testing genes that may be under similar or related regulatory pressures. Therefore, the filtered candidate genes whose TSS were suitable for allele-specific 4C-Seq were subjected to gene ontology analysis using the AmiGO web application (Carbon et al. 2009) to separate them into subgroups based on similarity of biological process. To further streamline the selection of final candidate genes, it was also required that, regardless of statistical significance, any selected genes must consistently have meet a minimum allelic expression bias based on comparison of SNP containing reads from the RNA-seq data. Taken together, allele-specific 4C-compatible candidate genes were further selected on the requirements of having at least a 2-fold allelic expression bias in both F1 crosses and appearing in no more than two of the top hierarchical functions in the biological process ontologies determined by AmiGO. Seven final candidate genes were selected from the analyses to proceed with investigation by allele-specific 4C-Seq (Table 2.1).

**Allelic 4C-Seq**

4C libraries are generated from downstream processing of 3C templates and are used to identify genome-wide interactions with a fixed locus, referred to as the viewpoint (Figure 2.2). Initially, nuclei must be harvested from fixed TS cells. These nuclei are lysed to release chromatin, which is digested with a restriction enzyme to expose loose ends of DNA that are joined in a proximity ligation. The dilute nature of the ligation reaction increases the local proximity of loose ends belonging to fragments that are tethered together by fixed nuclear protein, thereby leading to a favorable
<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Expression-Bias*</th>
<th>Biological Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airn</td>
<td>17</td>
<td>-5.0397</td>
<td>IncRNA, imprinted &amp; antisense to Igf2r (Lyle et al. 2000; Hudson et al. 2010)</td>
</tr>
<tr>
<td>Igf2r</td>
<td>17</td>
<td>5.5100</td>
<td>Facilitates degradation of growth factors, imprinted (Barlow et al. 1991; Hudson et al. 2010; Suh et al. 2010)</td>
</tr>
<tr>
<td>Bphl</td>
<td>13</td>
<td>2.0419</td>
<td>Serine hydrolase, implicated in kidney cell proliferation and aging (Melk et al. 2005; Grigo et al. 2008)</td>
</tr>
<tr>
<td>Mtus1</td>
<td>8</td>
<td>-2.6087</td>
<td>Identified as possible tumor suppressor in several lines (Louis et al. 2010; Zuern et al. 2010; Ding et al. 2012; Zuern et al. 2012)</td>
</tr>
<tr>
<td>Tlr5</td>
<td>1</td>
<td>1.6215</td>
<td>Toll-like receptor, innate immunity, recognizes bacterial flagellin (Letran et al. 2011; Weile, Josefsen, and Buschard 2011)</td>
</tr>
<tr>
<td>Car2</td>
<td>3</td>
<td>2.4129</td>
<td>Carbonic anhydrase II, deficiency is associated with osteopetrosis (Margolis et al. 2008; Rajachar et al. 2009)</td>
</tr>
<tr>
<td>Ibtk</td>
<td>9</td>
<td>-2.5396</td>
<td>Tyrosine kinase inhibitor, involved in B-cell development and maturation (Liu et al. 2001; Janda et al. 2011)</td>
</tr>
</tbody>
</table>

Table 2.1 Selected Candidates for Allelic 4C-Seq. Seven candidate genes were selected as described. Expression Bias is displayed as a log2 transformation of average allele-specific read ratios from RNA-seq experiments. For Airn and Igf2r, read ratio is maternal:paternal. For all others, read ratio is B6:CAST.
Figure 2.2. Process for Generating a 4C-Seq Library. 3C libraries are prepared from fixed cells using a 6-base cutter (H). Ligation products are digested with a 4-base cutter (4), re-ligated, purified to form circular products, and linearized with a third enzyme (L). Primers designed against the desired viewpoint (blue) amplify the captured fragments (red). 4, L, and the primers are selected to capture a known SNP (star). Primers are adapted to contain a random barcode (green) followed by the desired sequencing adapter (yellow, purple) to prepare the final 4C-seq library.
probability of forming ligations between loci that were in close 3-D space in the nucleus over the formation of ligations between fragments that find each other randomly in solution. The resulting ligation products are purified by reversing formaldehyde crosslinks and protein digestion followed by several rounds of phenol and phenol-chloroform extraction to isolate the 3C templates from which the 4C libraries are made.

A significant challenge in the development of the protocol is the adaptation of the experimental conditions for the desired cell type and restriction enzymes used in the 3C portion of the process. Because steps are performed on fixed chromatin, there are no purifications and clean up steps prior to the final isolations of the 3C ligation products. This means that all solutions and reagents present in the nuclear lysis are still present during the digestion and that the restriction enzyme is not removed from solution prior to ligation. Each step in the 3C process uses quenching of ionic detergents with non-ionic detergents to prepare satisfactory conditions for the following enzymatic reaction. For example, sodium dodecyl sulfate (SDS) is an ionic detergent that is used to lyse the fixed nuclei. Since SDS is efficient at denaturing proteins, its presence in solution is not conducive to restriction digestion. Typically, DNA would be purified prior to digestion by restriction enzymes; however, since purification of the DNA would require the removal of the nuclear proteins that are essential to generating the desired 3C products, the SDS must be neutralized before fragmentation. The introduction of non-ionic detergents, such as Triton-X 100, in appropriate proportions is thought to generate micelles that can encapsulate the ionic detergents, thereby neutralizing unwanted denaturation of the enzyme. The presence of non-ionic detergents, however, is not exactly ideal for most enzymatic reactions. This denaturation and quenching step is also necessary after digestion to prevent unwanted cleavage of ligation products during the subsequent proximity ligation. Therefore, conditions must be optimized to find minimal concentrations of detergents to facilitate increased efficiency of enzymatic reactions (see “Generation of 3C Libraries”).

The generation of 4C libraries from 3C templates is very similar to the preparation of 3C libraries with the key difference being that all steps are performed with purified DNA in solutions with ideal conditions (Figure 2.2, middle). This makes the optimization of these steps much more straigh
forward. The isolated 3C libraries are first trimmed by a second restriction digestion to shorten the templates and ensure that linear products are present for a subsequent proximity ligation. The second proximity ligation gives rise to circular products containing shortened portions of the viewpoint and captured fragments which are linearized by digestion at a known site in the viewpoint. This linearization offers a 4C template that is amplified more efficiently than circular products. While these processing steps may generate many molecules, a significant proportion will consist of viewpoint sequence flanking an unknown sequence that was "captured" due to a genomic interaction at the time of fixation.

**Controls**

Chromosome conformation libraries are the result of several reactions and processes that are strung together prior to detection and analysis. Therefore, it is important to establish control experiments at several points throughout the process to ensure that the expected products are being produced (Dekker 2006). Before proceeding with proximity ligations, electrophoretic migration in 0.8% agarose of undigested sample (nuclear lysate for the 3C step and 3C library for the 4C step) is compared to a portion of the digested material to ensure efficient fragmentation. Additionally, each time a proximity ligation is performed, a portion of the digested sample is kept as a no ligase control. Efficient ligation can also be detected by tested by electrophoresis in comparison to the no ligase controls.

To ensure that the 3C ligation reaction is producing expected products, an amplification test can be performed using an amplicon designed against a ligation template. For this, a bacterial artificial chromosome (BAC) covering the Gapdh locus was used to generate 3C amplification control templates by digestion and ligation under ideal conditions. Unlike the proximity ligations in 3C and 4C, this ligation is performed in small volumes to generate an expected equimolar distribution of all possible ligation combinations of BAC fragments (Dekker 2006). Primers designed to amplify the junction of two adjacent fragments can be optimized against this template for testing the 3C libraries. The high local proximity of adjacent fragments in a nucleus leads to a near certainty of forming such a ligation event in 3C libraries generated from mammalian cells (van Berkum et al. 2010). Therefore, it
is expected that a correctly manufactured 3C library should contain such a product, whose detection in a 3C library and absence in a no ligase control can be used as an indicator of efficacious production.

The amplification control for proper production of a 4C library is similar; however, due to the locus-specific nature of the 4C template, an amplification control must be performed on an assay-by-assay basis. Since the 4C process places sequence from the known viewpoint on either side of an unknown fragment, no BAC control is necessary. The primer sequences designed for amplification of the 4C template can be optimized using the 4C libraries directly. Once the amplification scheme is efficiently optimized, simply testing for amplification in the no ligase samples and the source 3C library provides a useful indicator or successful library generation.

Allelic Design and Expectation

The presence of known SNPs between the CAST and B6 lineages is informative of the allelic source of origin and can be accurately determined with high-throughput sequencing. In contrast to a previous allele-specific 4C experiment that relied on restriction fragment length polymorphisms to identify allelic identity (Splinter et al. 2011), allelic identity in this analysis was determined based solely on the sequencing of SNPs. Therefore, digestion schemes and primer positions for 4C libraries were designed around known SNPs to exploit this indicator (Illustration 2.1). The presence of an informative SNP in the sequenced portion of the viewpoint ensures that allelic origin of the viewpoint will be known for each sequenced ligation event. Additionally, an observed polymorphism rate of approximately 1 in 300 nucleotides (Keane et al. 2011) allows for a reasonable expectation that SNPs should be detected frequently in captured fragments. To support this, a bioinformatics analysis performed in the Magnuson Lab showed that approximately 64% of all 3C fragments cut with HindIII could be accurately called B6 or CAST with 70 bp of sequence from each end the unknown fragment. Paired-end sequencing of 100 bp reads was used to increase the likelihood of such detections.
Primer Strategy

Primer sequences for the amplification of 4C libraries were designed from the viewpoint sequence such that each primer was close to either the 3C or 4C restriction end, facing 5’ → 3’ in the direction of the nearest restriction end, completely homologous to both strains, and positioned to amplify at least one SNP between CAST and B6 (Illustration 2.1). The orientation requirement ensures that amplification should not be possible from genomic DNA or 3C libraries, but only from digested and re-ligated 4C libraries. The strain blindness of the primer sequence is intended to ensure equal amplification efficiency between alleles from a single amplification event. Finally, the capture of a SNP in the viewpoint fragment is what allows for the discrimination of allelic origin of the viewpoint. Since these amplification products were to be sequenced using 2x100 bp on the Illumina HiSeq 2000, a double amplification strategy was designed to amplify the product while simultaneously adding sequencing adapters (Figure 2.2, bottom). Genomic primer sequences were appended to contain a portion of the desired sequencing adapter. Additionally, it was advised by the UNC High-Throughput Sequencing core facility that a random linker be introduced between the genomic sequence and the sequencing adapter. The presence of random bases at the start of the sequencing read facilitates an increased resolution of cluster identification on the Illumina platform, which should lead to a higher retention of reads in the final data. Additionally, this random linker creates a variable barcode on each amplified read that can be used to determine the potential presence of amplification bias in the library preparation (see Chapter 3). The product from this initial amplification is then size-selected to fit the needed specifications of the sequencer and amplified with primers containing sequence homologous to the 5’ portion of the adapter used in the previous amplification as well as the remainder of the sequencing adapter. Purification of this amplified product will yield a 4C-Seq library that is ready to be sequenced.

Intentions for Mapping and Analysis

For each candidate gene, 4C libraries will be generated and sequenced in duplicate for TS cells of reciprocal F1 crosses (i.e., CAST from paternal lineage and CAST from maternal lineage, 2 libraries for each cross). Sequencing libraries can be multiplexed on the Illumina platform using the
Illumina TruSeq indexes to determine library origin. The use of these indexes combined with the sequence of the viewpoint in the resulting read allows for an effective multiplexing strategy. Specifically, all 4C libraries from one cell line replicate can be sequenced on a single lane using the same TruSeq index. The same lane can also contain all 4C libraries from a reciprocal cell line replicate using a different TruSeq index. This allowed 14 experiments to be sequenced in a single lane by using the adapter index to identify the replicate to which a read belongs and the viewpoint sequence to identify the library to which a read belongs.

Sequenced reads were to be sorted into experiments based on their appropriate replicate, library, and allele. Subsequently, the portion of the read that did not map to the viewpoint (captured fragment) was mapped to reference genomes to determine position and, where SNPs are available, allelic origin. Significant interactions were identified using novel statistical methods loosely based on a previously reported algorithm (Splinter et al. 2011) and analyzed for reproducibility and association with expression state (see “fourSig: A Novel 4C-Seq Analysis Method”). Interactions that were reproducible amongst replicates and exclusive to a specific and consistent expression state in both cell lines were scored as having a potential functional association with the expression regulation of the candidate gene.

Validation of Allelic Expression Bias in Candidate Genes

Before proceeding with 4C-Seq experiments on the selected candidate genes, the allelic expression bias observed in the RNA-seq data for the selected candidate genes needed to be validated by a secondary method. To this end, allele-specific quantitative reverse transcription polymerase chain reaction (qRT-PCR) assays were designed for each of the candidate genes. In general, allele specificity was sought by designing amplicons to detect RT products using a single primer (typically the reverse primer) common to both alleles paired with a primer that aligns over at least one SNP (Table 2.2). The goal was to optimize amplicons capable of efficiently amplifying the correct allele while also suffering from a substantially unstable annealing to the primer pair’s
<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gapdhf1</td>
<td>TGGTCCCTACC CCCCAATGTGT</td>
<td>Forward primer</td>
</tr>
<tr>
<td>Gapdhr1</td>
<td>TGGAGGGAGATGCTAGTG</td>
<td>Reverse primer</td>
</tr>
<tr>
<td>lbtk_1fb</td>
<td>CTGTGCCGAGACTTGATATTTCC</td>
<td>Forward cDNA primer, B6-specific</td>
</tr>
<tr>
<td>lbtk_1fc</td>
<td>TCTGTGCCGAGACTTGATATTTCT</td>
<td>Forward cDNA primer, CAST-specific</td>
</tr>
<tr>
<td>lbtk_1r</td>
<td>GCCACGACAAGAACAACACTACAA</td>
<td>Reverse cDNA primer</td>
</tr>
<tr>
<td>Airn_e2s2BF</td>
<td>GCCATCACCTAACAGGGACATC</td>
<td>Forward cDNA primer, B6-specific</td>
</tr>
<tr>
<td>Airn_e2s2CF</td>
<td>GCCATCACCTAACAGGAAGAAGC</td>
<td>Forward cDNA primer, CAST-specific</td>
</tr>
<tr>
<td>Airn_e2s1R2</td>
<td>TCACATGCTGAGGCAAGAGC</td>
<td>Reverse cDNA primer</td>
</tr>
<tr>
<td>Igf2r_e45s1BF</td>
<td>CAGAGTTCAACAGCAACAGACCG</td>
<td>Forward cDNA primer, B6-specific</td>
</tr>
<tr>
<td>Igf2r_e45s1R2</td>
<td>CCATGCCACAGACCAGATAGC</td>
<td>Reverse cDNA primer for B6 amplicon</td>
</tr>
<tr>
<td>Igf2r_e14s1CF</td>
<td>TGCTAAGGCTGCAAGATCCCG</td>
<td>Forward cDNA primer, CAST-specific</td>
</tr>
<tr>
<td>Igf2r_e14s1R2</td>
<td>ACACAGCAAGAAGACGCTTTGG</td>
<td>Reverse cDNA primer for CAST amplicon</td>
</tr>
<tr>
<td>bphl_e2s1BF</td>
<td>CATTACCAGCGCGTGGGAGA</td>
<td>Forward cDNA primer, B6-specific</td>
</tr>
<tr>
<td>bphl_e2s1CF</td>
<td>CATTACCAGCGCGTGGGAAA</td>
<td>Forward cDNA primer, CAST-specific</td>
</tr>
<tr>
<td>bphl_e2s1R3</td>
<td>TCAGACTTGGAGCAAGGCAAT</td>
<td>Reverse cDNA primer</td>
</tr>
<tr>
<td>mtus1_e2s7BF</td>
<td>CTTTATTGGAAGTCTTTTGGGT</td>
<td>Forward cDNA primer, B6-specific</td>
</tr>
<tr>
<td>mtus1_e2s7CF</td>
<td>CTTTCATTACGAAGTCTTTTGGGT</td>
<td>Forward cDNA primer, CAST-specific</td>
</tr>
<tr>
<td>mtus1_e2s7R1</td>
<td>AGATGCCTAACGTGCAACCAG</td>
<td>Reverse cDNA primer</td>
</tr>
<tr>
<td>tlr5_e4s10BF</td>
<td>TTTGAAGAAAAGAAGACTTCTATTCCG</td>
<td>Forward cDNA primer, B6-specific</td>
</tr>
<tr>
<td>tlr5_e4s10CF</td>
<td>GCTTTGAAGAAAAGAAGACTTCTATTCCAG</td>
<td>Forward cDNA primer, CAST-specific</td>
</tr>
<tr>
<td>tlr5_e4s10R2</td>
<td>GCTTTGAAGAAAAGAAGACTTCTATTCCAG</td>
<td>Reverse cDNA primer</td>
</tr>
</tbody>
</table>

**Table 2.2 Primers for Allele-specific qRT-PCR Assays.** Primers used for allele-specific qRT-PCR assays are listed with descriptions for use.
unintended allele (Singer-Sam 1994). Two slightly different strategies were used for the design of these amplicons.

For the majority of the selected candidates, forward primers were designed to align over SNPs in the transcribed region of the gene. The TM Mismatch feature of the OligoAnalyzer tool from Integrated DNA Technologies, Inc. (http://www.idtdna.com/analyzer/applications/oligoanalyzer/) was used to estimate the least energetically favorable hybridization to cross-allele sequence for potential primers, thereby minimizing the number of primers needed to settle on an assay. Additionally, amplicons were designed to detect unspliced transcripts. Because amplification efficiency needed to be optimized for the correct allele in opposition to the incorrect allele for each assay, gDNA from allelic isotype mice was used as an ideal template to test the ability of the allele-specific assays to effectively discriminate between alleles.

Two of the candidate genes, *Ibtk* and *Car2*, had previously been studied by Dr. Joshua Starmer, a post-doc in the Magnuson Lab, as part of an earlier observation of allelic expression bias in hybrid TS cells. As a result, allele-specific qRT-PCR assays had already been designed for these two genes. To verify allelic expression bias, Dr. Starmer designed intron-skipping amplicons to detect allele-specific cDNA templates by placing a SNP at the most 3' nucleotide of the forward primers and pairing them with a common reverse primer. Because these assays were designed to detect cDNA made from processed transcripts, RT products from allelic isotype cells needed to be prepared for use in efficiency optimization and as positive and negative controls.

Expression can be measured in qPCR assays as a relative abundance calculation using the following equation:

$$\Delta C(t) = E \cdot (C(t)_{\text{Control}} - C(t)_{\text{Test}})$$

where C(t) is the threshold cycle determined in the qPCR experiment, "Test" refers to the experimental sample, "Control" refers to an untreated or baseline control, and "E" is the efficiency of the amplification reaction (Pfaffl 2001). Efficiency is commonly assumed to be 2 since efficient amplicons are expected to double the amplified product with each amplification cycle. However,
because the assays designed here are intended to detect allele-specific products by exploiting differences in efficiency of amplification between the favored and unintended alleles, it was deemed important to use accurately determined efficiencies for each amplification assay. Furthermore, it was also necessary to normalize the expression to a reference gene with no allelic bias for comparison of replicates. In the event that they are not very close to 2, efficiencies were needed for both the test and referenced amplification reactions to normalize ΔC(t) values using the Pfaffl equation (Pfaffl 2001):

$$\Delta \Delta C(t) = \Delta C(t)_{\text{Gene}} / \Delta C(t)_{\text{Reference}}$$

A qPCR assay lacking allele specificity for Gapdh was used as a reference to determine the normalized ratio. To determine the efficiencies for each assay, qPCR was performed in triplicate on 1:10 serial dilutions of optimal template. Average C(t)s were plotted against Log10-transformed mass of input DNA used to determine linear regression lines, from which efficiency was calculated using the following equation:

$$E = 10^{(-1/m)}$$

where “m” is the slope of the linear regression line (See Appendix). Efficiencies were calculated similarly for the cross-allelic templates (e.g., B6 assay using a CAST template) to ensure that a substantial loss of efficiency occurs. Such an absence of efficient amplification is necessary to ensure that transcript detected in the F1 TS cells is not due to unintended amplification of the unintended allele.

For the genes whose assays were designed to detect unprocessed transcript, allelic isotype gDNA was used as the control sample. For Ibtk and Car2, cDNA from allelic isotype TS cells was used as the control sample. C(t)s were determined in triplicate for cDNA from two different sample preparations of both reciprocally crossed TS cell lines, for cDNA from allelic isotype TS cell lines, for gDNA from allelic isotype mouse tissue, and for no RT controls for each cDNA preparation. Average C(t)s were used from each triplicate experiment to calculate ΔΔC(t) values and relative comparison of Gapdh-normalized expression was observed as a validation of the bias detected in the RNA-seq
Figure 2.3 Validation of Allelic Expression Bias in Candidate Genes. A. The proportion of allele-specific reads detected from RNA-seq experiments in two different TS cell lines show heavy expression biases for each candidate gene. In the figure legends, C/B refers to the CASTB6F1 cell line while B/C refers to the B6CASTF1 cell line. B. Expression biases of most candidates were confirmed by allele-specific qRT-PCR. Each cell line was assayed twice, in triplicate. The average normalized expression is displayed as relative values between the alleles of a cell line.
experiment (Figure 2.3). With the exception of Car2 and Mtu1, this analysis resulted in the faithful validation of allelic expression bias in each case.

While Car2 does show a bias in detected expression, the qRT-PCR analysis shown here identifies a bias in the opposite direction of what was originally observed in the RNA-seq experiments (Figure 2.3). This may be explained by the use of allelic isotype cDNA as the baseline control. All cDNA products were made from RT reactions using a starting amount of 2.5 nanograms (ng) RNA. Therefore, unlike the gDNA templates which should have near equivalent amounts of all templates in solution, it is possible that varying expression levels between samples could result in differing amounts of individual templates within the RNA solution. Previous experiments by Dr. Starmer in the Magnuson Lab have suggested that Car2 expression is expressed at different levels in the isotype lines and that the expression levels for each allele seen in the F1 TS cell lines are actually different than what is observed in a homozygous line (unpublished data). Therefore, while the isotype cDNAs serve as a good control to ensure discriminatory detection between the allelic assays, differences in template concentration may make them poor controls for comparing normalized relative expression. Indeed, the average C(t) values alone suggest that the B6 template reaches C(t) at more than 2 cycles earlier than the CAST template in each test (data not shown). This translates to more than a 4-fold enrichment of B6 template relative to CAST in the test samples, which is very similar to what was originally found by RNA-Seq. For Ibtk, the other gene that used an allelic isotype cDNA as the control sample, comparison of C(t)s for each allele was consistent with calculated ∆∆C(t) values and both were reflective of the strong CAST bias seen in the RNA-seq experiments (data not shown). With this data in hand, the decision was made to proceed with the 4C-Seq experiments for all genes rather than wait to design new allelic qRT-PCR assays for Car2 and Ibtk.

**Generation of 3C Libraries**

As described earlier (see “Allelic 4C-Seq”), reliable and efficient production of 3C libraries requires careful optimization of several steps to suit the targeted cell line. Many of the steps in the
protocol were optimized concurrently and most of the final conditions at each step were dependent upon results obtained in the following step. In other words, steps were not always optimized one at a time so some final conditions were settled on using conditions in the following step that may have changed later. Additionally, once the protocol is settled, the generation of 3C libraries must still be subjected to several controls along the way to ensure the quality of the library. Here the measures taken to arrive at the final optimized steps in the protocol for performing 3C in TS cells (see Chapter 5) and the generation of the libraries used in the 4C-Seq experiment are summarized.

Handling of TS cells

The first thing that required optimization was the isolation of tissue, which focused concurrently on two distinct factors. First, TS cells are typically grown on a feeder layer of fibroblasts, which are necessary to preserve maximally uniform morphology of TS cells in culture and is thought to help guard against variable differentiation amongst colonies (Quinn, Kunath, and Rossant 2006). Maintaining a high level of uniformity of a physiological indicator, such as morphology, is important for genomic assays because the result is a reflection of cell populations rather than individuals. However, the presence of two distinct cell types within the population is unacceptable because there is no way to distinguish the source of the detected DNA. TS cells can be grown off of feeders for a few passages before morphology becomes visibly compromised. It was ultimately settled that, after having harvested several batches of TS cells, growing them off of feeders for 2 passages (NFx2) offered the highest likelihood of preserving good uniform morphology while eliminating nearly all feeders from culture.

The second factor concerns the densely clumped colonies in which TS cells grow. Most published 3C protocols highly recommend isolation of single cells, so it was decided that colonies should be dispersed using trypsin and trituration to achieve single cells prior to fixation. Typically, TS cells at NFx2 were consistently brought to single cells by incubating at room temperature in 0.25% trypsin for approximately 4 minutes, followed by trituration with a glass pipette against the surface of the culture plate. Additionally, it was noticed that a short pre-incubation at room temperature in phosphate buffered solution (PBS) prior to trypsinization facilitates easier dispersal of colonies. Since
isolation occurred in batches, trypsin exposure was typically staggered with dispersed cultures being pooled and quenched in RPMI with 10% fetal bovine serum (FBS) on ice until all plates were processed. TS cells would then be fixed, quenched, flash-frozen in liquid nitrogen (LN₂), and stored at -80 °C until needed for 3C library preparations.

**Optimization of Fixation**

Nuclei are isolated by lysing cells with a non-ionic detergent (IGEPAL-CA630) in a dounce homogenizer. Chromatin is released from this fraction by nuclear lysis in a restriction digestion buffer with SDS, which should be kept to a minimum to reduce the amount of residual detergents in downstream steps. The optimization of this lysis, however, was also dependent upon fixation conditions. The published 3C protocols that used as a starting point called for 1% formaldehyde to fix cells (Lieberman-Aiden et al. 2009), however the ChIP-seq experiments in these TS cells that were used to screen candidate genes use 0.6% formaldehyde for fixation (Calabrese et al. 2012).

Conversations with Dr. J. Mauro Calabrese revealed that he had previously had problems with lysis by sonication, a harsher method than detergent-based lysis, with nuclei fixed with 1% formaldehyde. Dr. Calabrese had postulated that, because TS cells may be somewhat “tough”, higher fixation concentrations make for a difficult lysis. By reducing fixation to 0.6% formaldehyde, Dr. Calabrese was able to more efficiently lyse TS cells without compromising the preservation of protein-DNA interactions. Therefore, nuclei fixed with both 0.6% and 1% formaldehyde were used to test efficient isolation of chromatin.

Nuclei were resuspended with 0.5% SDS and incubated at 65 °C for 10 minutes with periodic agitation. At the time, nuclei were being lysed, quenched, and digested and the end product consistently contained a pelleted fragment that was found to still contain gDNA upon lysis. Therefore, this “pellet” was being exposed to a second lysis and digestion step, the result of which was intended to be reunited with the first digestion product prior to proximity ligation. Therefore, the amount of DNA released from both lysis steps was recorded. Samples were performed in duplicate alongside a gDNA preparation from TS cells lysed initially with a standard Proteinase K-based lysis buffer (Figure 2.4). Little difference in isolated DNA was found between the conditions, however a larger portion of
Figure 2.4 Optimization of Fixation Conditions for 3C in TS Cells. TS cells were tested for optimal DNA yield and restriction digestion after fixation with 0.6% and 1% formaldehyde. A. Similar DNA yields were obtained from multiple tests with both conditions. Both conditions resulted in a higher expected yield than what was seen with a Proteinase K-based lysis buffer. B. Both conditions resulted in similar fragmentation with HindIII digestion with a slight advantage apparent using the 1% fixation condition. Upper line represents 8 kb and lower line represents 2 kb.
Figure 2.5 Optimization of Lysis Conditions for 3C in TS Cells. TS cells were tested for optimal DNA yield and restriction digestion after fixation with 1% formaldehyde and SDS lysis followed by 10-fold quenching with Triton X-100. Lysis was performed as a single step as well as with a second lysis on recovered pelleted material following initial lysis. A. DNA yields from 5 million cell pellets were most consistent using 0.2% SDS (S) lysis followed by quenching with 2% Triton X-100 (T). B. All conditions resulted in similar fragmentation with HindIII digestion with the 0.2%S/2%T condition offering the best similarity to digested naked DNA. Upper line represents 8 kb and lower line represents 2 kb.
the isolated DNA was found to be digested into the expected range of products in the 1% formaldehyde solutions. This, combined with a fear of under-preserving protein-DNA interactions (Splinter, Grosveld, and de Laat 2004) resulted in the decision to proceed with a 1% formaldehyde fixation.

Optimization of Lysis

The minimization of SDS concentration in the lysis conditions is important to reduce the amount of Triton X-100 present in the digestion and ligation steps. Through several trial-and-error experiments, it was discovered that a 10-fold excess of Triton X-100 to SDS used provides an optimal quenching of ionic detergent so as to not interfere with HindIII digestion. To settle in on the final conditions of 0.2% SDS for lysis and 2% Triton X-100 for quenching, DNA collected from lysis and digestion using a gradient of 10-fold quenching conditions in single- and double-digestion experiments was analyzed for quantity and fragmentation (Figure 2.5). The 0.2% SDS/2% Triton X-100 conditions offered the best combination of consistent DNA yield and ideal fragmentation by HindIII as compared to an ideal digest on naked gDNA.

Optimization of Ligation

HindIII must be thoroughly inactivated prior to beginning the proximity ligation to ensure that residual cleavage activity does not re-digest ligated ends. Typically, any residual active enzyme is denatured by addition of SDS, which must be quenched prior to introduction of T4 DNA Ligase. To optimize the conditions for inactivating HindIII, two successive ligation experiments were performed using a gradient of Triton X-100 exposure (Figure 2.6). For the first experiment, digested chromatin samples prepared from a 0.2% SDS/2% Triton X-100 lysis were treated with or without SDS to a final concentration of 1.5% and incubated at 65 °C for exactly 30 minutes. Three of the samples exposed to SDS were then treated with Triton X-100 to final concentrations of 1%, 1.5%, and 2%. One SDS treated sample was kept free of Triton X-100 and T4 DNA Ligase to serve as a no ligase control. Of the samples not exposed to SDS, one was treated with 1% Triton X-100 to test the effects of the non-ionic detergent alone on the ligation reaction and one was not treated with Triton X-100 to test the -
Figure 2.6 Optimization of Ligation Conditions for 3C in TS Cells. Digested lysates from TS cells were tested for optimal quenching of SDS prior to proximity ligation. 0.2% SDS/2% Triton X-100 conditions were used for lysis of 1% formaldehyde fixed cells. A. HindIII was quenched at 65 °C for 30 minutes with and without 1.5% SDS. A Triton X-100 gradient was used to determine optimal ligation efficiency. HindIII lane is digested naked DNA and Lig Cntrls are BAC ligation controls. B. A similar experiment was performed using SDS to quench HindIII in all samples and a higher Triton X-100 gradient for quenching prior to ligation. 3C lane is a previously prepared 3C library using naked DNA. Upper line represents 8 kb and lower line represents 2 kb.
possibility of using only heat to inactivate HindIII. A clear improvement in the amount of ligated product could be seen with increased Triton X-100 treatment (Figure 2.6A). Although heat inactivation alone produced a tighter ligation product, the intensity of the product on the gel seemed to suggest that more DNA was ligated using the 2% Triton X-100 condition. Since increasing Triton X-100 concentration seemed to have a positive effect, a second experiment was performed using a concentration gradient of 2%, 2.5%, and 3% (Figure 2.6B). The 3% Triton condition appeared to generate the highest intensity ligation products while leaving the least amount of streaking in the digested sample range (2 – 8 kb). Therefore, inactivation of HindIII with approximately 1.5% SDS followed by proximity ligation in a solution containing 3% final concentration of Triton X-100 were selected as the working conditions for the ligation step.

The remainder of the 3C portion of the process involves several rounds of DNA extraction and precipitation aimed at cleaning up the 3C libraries and reducing the volume of solution. Several tips and tricks were found through experience to improve yield and facilitate work-flow along the way. Though not summarized in detail here, these specifics may be found in the detailed protocol (see Chapter 5)

Library Preparation

3C libraries were prepared by initial lysis of TS cells, quenching of SDS by TritonX-100, restriction digestion with HindIII and proximity ligation with T4 DNA Ligase (See Chapter 5). During the preparation, a portion of the digested product was withheld from the ligation reaction to serve as a control for effective production of the 3C libraries. Quality can be tested by comparing the electrophoretic migration of the re-ligated and non-ligated products in an agarose gel (not shown). More importantly, a high quality library should generate ligation products between adjacent fragments in favorable abundance for detection by PCR.

To test this property of the libraries, a ligation control was prepared by digesting a BAC mapping to the Gapdh locus (Table 3.2) and re-ligating it under ideal conditions. This product should create a template for amplifying the hybrid fragments from primers that are normally facing the same
<table>
<thead>
<tr>
<th>3C Library</th>
<th>Prepared</th>
<th>TS Line</th>
<th>Cross</th>
<th>Harvested</th>
<th>Passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C_6#2_A</td>
<td>3/5/2012</td>
<td>TS 6#2</td>
<td>CASTB6F1</td>
<td>2/20/2012</td>
<td>17</td>
</tr>
<tr>
<td>3C_6#2_B</td>
<td>5/12/2012</td>
<td>TS 6#2</td>
<td>CASTB6F1</td>
<td>2/20/2012</td>
<td>17</td>
</tr>
<tr>
<td>3C_6#2_C</td>
<td>5/14/2012</td>
<td>TS 6#2</td>
<td>CASTB6F1</td>
<td>4/3/2012</td>
<td>17</td>
</tr>
<tr>
<td>3C_10#1_C</td>
<td>5/19/2012</td>
<td>TS 10#1</td>
<td>B6CASTF1</td>
<td>4/20/2012</td>
<td>15</td>
</tr>
<tr>
<td>3C_10#1_D</td>
<td>5/21/2012</td>
<td>TS 10#1</td>
<td>B6CASTF1</td>
<td>3/3/2012</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2.3 3C Libraries. Large 3C Libraries used as source material for 4C-Seq experiments. "Prepared" is the date the library was prepared. "TS Line" refers to the name of the TS cell line in the Magnuson Lab repository. "Cross" describes the parental origin of the cell line. "Harvested" refers to the date of fixation and storage of cell pellets.
3C libraries were prepared and tested for quality by amplification of a hybrid fragment produced from adjacent fragments at the Gapdh locus. A BAC mapping to the Gapdh locus is used to prepare a ligation control. The test amplicon is only amplified when the BAC has been digested and re-ligated. Test amplifications were performed to verify the quality of (B) 3C_6#2_A, (C) 3C_6#2_B, (D) 3C_6#2_C, (E) 3C_10#1_C, and (F) 3C_10#1D prior to their use in producing 4C libraries.
orientation. An amplicon generating a hybrid template of approximately 186 bp was designed and tested on the BAC ligation control (Figure 2.7A). As an additional precaution, the resulting product was purified and digested with HindIII to ensure that the band is, in fact, the result of a ligation event. This amplicon was used as the primary control test to verify the quality of 3C libraries prior to their use in the 4C portion of the protocol (Figure 2.7B-F, Table 2.3).

**Generation of 4C Libraries**

The process for producing 4C templates is much more straightforward than making 3C libraries. The protocol involves similar steps as in 3C but, while these steps still need controls to ensure effective reactions, the reactions are generally performed under ideal conditions on relatively purified product. The absence of residual reagents, such as detergents, that creates a problem in the 3C process makes optimization of these steps largely unnecessary. The challenge in this portion, summarized below, is primarily related to finding optimal amplification conditions for each 4C assay.

*Digestion Schemes and Primer Design*

After successful quality control for the 3C libraries is performed, purified 3C libraries must be digested with a restriction enzyme that recognizes a 4-base motif (4-cutter, Illustration 2.1). This step provides two distinct benefits. First, a shorter DNA fragment ensures that circular ligation events are the result of a single capture event between one viewpoint fragment and one captured fragment (Figure 2.2, middle). One can envision a scenario in which several fragments bound by protein lead to a concatenation of more than two fragments during the proximity ligation in the 3C process. Without shortening of the 3C template, the circular ligation of such a fragment may lead to the amplification of two different captured fragments from the same viewpoint. Second, a long captured fragment creates several potential problems downstream. A 6-cutter, such as HindIII, produces fragments with an average sized of approximately 4 kb. Our tests on HindIII digestion of fixed chromatin show that the majority of DNA is fragmented between 2 – 8 kb. Circular ligation of a hybrid template consisting of just two fragments this size would contain a large distance of captured
sequence to amplify through during the 4C amplifications. This increases the probability that the
downstream linearization step will recognize sites in the unknown captured sequence, inhibiting
subsequent amplification. Additionally, any 4C products that are produced will need to be fragmented
to achieve an ideal size for cluster generation on the sequencer. Random fragmentation at this point
introduces the chance for multiple reads per capture event, thereby confounding any association
between detection frequency and contact probability. Trimming the size of 3C libraries prior to
circular ligation alleviates these concerns.

The ability to gain allele-specific resolution necessitates that the restriction enzymes used in
the 4C process are selected with respect to the positioning of known SNPs between B6 and CAST
strains in the viewpoint fragment. Possible digestion schemes were selected by searching for 4-
cutter sites near the end of the HindIII fragment containing the TSS for the desired gene positioned
such that a SNP was within 60 bp of either end of the resulting 4C viewpoint fragment and a 6- or 8-
cutter restriction site existed between the 4-cutter and the HindIII sites (Illustration 2.1). A maximum
distance of 60 bp between the end of the viewpoint fragment and associated SNP allows for
placement of up to 20bp for primer length while still ensuring a minimum of 20 bp sequence from the
unknown captured fragment in a 100 bp sequencing read. Additionally, all restriction sites needed to
be present in both genomes (i.e., no SNPs) to ensure that 4C products are produced from both
alleles.

The resulting schemes were then selected based on the percent of predicted resulting
fragments that would be mappable to the B6 and CAST genomes. Schemes that resulted in less
than 80% mappability from the predicted fragments were eliminated from potential use. Lastly, if 4C
templates can be made for several genes using the same digestion scheme, then the 4C libraries for
those genes can be made from the exact same template by simply changing the primers in the
amplification step. Therefore, an effort was made to limit the number of gene specific digestion
schemes required by prioritizing schemes that used similar enzymes between genes (Table 2.3).

Having selected digestion schemes to produce 4C libraries, genomic primers were designed
with opposite 5’ → 3’ orientation for each gene (Table 2.4). Primers were selected so that they do not
<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>3C Enzyme</th>
<th>4C Enzyme</th>
<th>Linearizer</th>
<th>4C Mappability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim</td>
<td>17</td>
<td>HindIII</td>
<td>Msel</td>
<td>Pvull</td>
<td>96%</td>
</tr>
<tr>
<td>Igf2r</td>
<td>17</td>
<td>HindIII</td>
<td>CviQI</td>
<td>Pvull</td>
<td>78%</td>
</tr>
<tr>
<td>Bphl</td>
<td>13</td>
<td>HindIII</td>
<td>DpnII</td>
<td>Dral</td>
<td>80%</td>
</tr>
<tr>
<td>Mtus1</td>
<td>8</td>
<td>HindIII</td>
<td>Msel</td>
<td>BglII</td>
<td>96%</td>
</tr>
<tr>
<td>Tlr5</td>
<td>1</td>
<td>HindIII</td>
<td>CviQI</td>
<td>BmrI</td>
<td>77%</td>
</tr>
<tr>
<td>Car2</td>
<td>3</td>
<td>HindIII</td>
<td>DpnII</td>
<td>Dral</td>
<td>80%</td>
</tr>
<tr>
<td>Ibtk</td>
<td>9</td>
<td>HindIII</td>
<td>Nlalll</td>
<td>Nsil</td>
<td>95%</td>
</tr>
</tbody>
</table>

Table 2.4 4C Digestion Schemes for Candidate Loci. 4C digestion schemes were designed such that an informative SNP may be sequenced from the viewpoint fragment. The mappability of expected detectable fragments was assessed by comparison to restriction motifs annotated in the mm9/NCBIv37 genome assembly.
<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
<th>Sequence (5'-&gt;3')</th>
<th>Ori</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>4C_Airn-H3_Tru</td>
<td>Airn-4C</td>
<td>[Adapter]-NNNAGAAAGAAAATAGGAGAGCAAGC</td>
<td>+</td>
<td>chr17:12941652-12941674</td>
</tr>
<tr>
<td>4C_Airn-MseI_Nex</td>
<td>Airn-4C</td>
<td>[Adapter]-NNNNGTGCTGGGAAGAGGATCAAGT</td>
<td>-</td>
<td>chr17:12941292-12941312</td>
</tr>
<tr>
<td>4C_Igf2r-CviQ_Nex</td>
<td>Igf2r-4C</td>
<td>[Adapter]-NNNNCCGGGTCACATGAGGCATC</td>
<td>+</td>
<td>chr17:12962652-12962669</td>
</tr>
<tr>
<td>4C_Igf2r-H31_Tru</td>
<td>Igf2r-4C</td>
<td>[Adapter]-NNNNTAGCCATTTGTCACTGAGTTGA</td>
<td>-</td>
<td>chr17:12961823-12961844</td>
</tr>
<tr>
<td>4C_Bphl-Dpn2_Tru</td>
<td>Bphl-4C</td>
<td>[Adapter]-NNNNAGGAGGAGGAGGAGCTCTTA</td>
<td>-</td>
<td>chr13:34132297-34132316</td>
</tr>
<tr>
<td>4C_Bphl-H3_Nex</td>
<td>Bphl-4C</td>
<td>[Adapter]-NNNNCCACACAGATGCACACAAGGA</td>
<td>+</td>
<td>chr13:34133006-34133024</td>
</tr>
<tr>
<td>4C_Mtus1-H3_Nex</td>
<td>Mtus1-4C</td>
<td>[Adapter]-NNNNCCCTCCACACACTTCACAGA</td>
<td>+</td>
<td>chr8:42227183-42227201</td>
</tr>
<tr>
<td>4C_Mtus1-MseI_Tru</td>
<td>Mtus1-4C</td>
<td>[Adapter]-NNNNCCCTCCACACACTTCACAGA</td>
<td>-</td>
<td>chr8:42226459-42226477</td>
</tr>
<tr>
<td>4C_Tlr5-CviQI_Tru</td>
<td>Tlr5-4C</td>
<td>[Adapter]-NNNNTCACCCTCGGTGCCAGGAGAGG</td>
<td>-</td>
<td>chr1:184894794-184894812</td>
</tr>
<tr>
<td>4C_Tlr5-H3_Nex</td>
<td>Tlr5-4C</td>
<td>[Adapter]-NNNNCTACCTCGGTGCCAGGAGAGG</td>
<td>+</td>
<td>chr1:184894896-184895014</td>
</tr>
<tr>
<td>4C_Car2-Dpn2_Tru</td>
<td>Car2-4C</td>
<td>[Adapter]-NNNNAGAATGAATGAGATGAGGTATG</td>
<td>-</td>
<td>chr3:14893894-148939200</td>
</tr>
<tr>
<td>4C_Car2-H3_Nex</td>
<td>Car2-4C</td>
<td>[Adapter]-NNNNTCCCAGAATACACTGGTCTGA</td>
<td>+</td>
<td>chr3:14894297-14894318</td>
</tr>
<tr>
<td>4C_Ibtk-Nla3_Nex</td>
<td>Ibtk-4C</td>
<td>[Adapter]-NNNNNTCTGATTCCTCTCTTCTTCTTCTT</td>
<td>-</td>
<td>chr9:85644465-85644486</td>
</tr>
<tr>
<td>4C_Ibtk-H3_Tru</td>
<td>Ibtk-4C</td>
<td>[Adapter]-NNNNNTCTGATTCCTCTCTTCTTCTTCTT</td>
<td>+</td>
<td>chr9:85645023-85645048</td>
</tr>
</tbody>
</table>

Table 2.5 Primers for Allele-specific 4C Assays. Genomic primers for the viewpoint of each candidate gene are displayed. Primers contain a random barcode (NNNN) and a portion of the desired sequencing adapter (proprietary). Nextera V1 Universal Adapter (Illumina) sequence was used for the side of the viewpoint containing the informative SNP and Illumina TruSeq Adapter sequences were used for the opposing primer.
align over SNPs, ensuring that both alleles can produce near equal proportions of 4C product.

Optimizations and Preparation

4C libraries were produced by digesting purified 3C products with secondary restriction enzymes and performing a second proximity ligation (Figure 2.2). The quality control steps used here are similar to those performed for the 3C libraries but vary slightly for each final 4C product. Each 4C library started with approximately 12.5 micrograms (µg) 3C library and efficient digestion was tested by observing electrophoretic migration in 0.8% agarose (Figure 2.8A). As with the 3C libraries, a portion of the digested 3C product was withheld from the subsequent proximity ligation to serve as a no ligase control in downstream steps. After circularization and linearization, amplification potential was assessed by performing a template dilution gradient test with the initial adapted primers (Figure 2.8B). This serves to ensure that an appropriate template exists and to gauge the approximate cycle numbers needed in the initial amplification step. Several amplification reactions are performed and pooled for a gel purification to isolate the 4C intermediate product (Figure 2.8C). This product serves as the template for the final amplification using the indexed adapter primers. A small portion of this may be run on a gel to verify the size prior to submission for sequencing (Figure 2.8D).

The challenge at this step is involved in identifying the most effective amplification conditions for each set of primers. For the genes assayed in this project, appropriate conditions for each 4C amplification were determined independently using several temperature gradients with varying concentrations of magnesium sulfate (not shown). 4 sequencing libraries, 2 from each cell line, were ultimately produced for each candidate gene (Table 2.6). A preliminary 4C library was produced for the *Ibtk* locus for a pilot experiment using the Illumina MiSeq sequencing platform (See Chapter 3).

fourSig: A Novel 4C-Seq Analysis Method

The creation of 4C libraries inevitably generates certain technical limitations that must be considered when analyzing 4C-Seq data. Among others, local proximity can have variable effects on the significance of an interaction. Because chromatin is essentially a polymer in confined space, the
Figure 2.8 4C Libraries and Amplification Controls. Quality control tests are depicted with representative results from the Ibtk_B 4C library are shown. 4C libraries were prepared from 3C libraries and tested for quality by (A) observing their fragmentation after secondary digestion and (B-D) observation throughout the amplification steps. B. Template dilution gradient for Ibtk_B 4C intermediate library C. Gel purification of pooled 4C intermediate product for Ibtk_B. D. Size check of small amount of final Ibtk_B 4C library. Invitrogen 1kb Plus ladder is used in panel A and Invitrogen 100bp ladder is used in panels B-D.
<table>
<thead>
<tr>
<th>4C Library</th>
<th>3C Source</th>
<th>Adapter Index</th>
<th>Platform</th>
<th>Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airn_A</td>
<td>3C_6#2_B</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Airn_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Airn_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Airn_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Igf2r_A</td>
<td>3C_6#2_B</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Igf2r_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Igf2r_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Igf2r_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Bphl_A</td>
<td>3C_6#2_B</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Bphl_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Bphl_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Bphl_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Mtus1_A</td>
<td>3C_6#2_A</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Mtus1_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Mtus1_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Mtus1_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Tlr5_A</td>
<td>3C_6#2_B</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Tlr5_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Tlr5_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Tlr5_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Car2_A</td>
<td>3C_6#2_B</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Car2_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Car2_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Car2_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Ibtk_Test</td>
<td>3C_6#2_A</td>
<td>TruSeq Index 4</td>
<td>MiSeq</td>
<td>eB3Ibtk</td>
</tr>
<tr>
<td>Ibtk_A</td>
<td>3C_6#2_A</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Ibtk_B</td>
<td>3C_6#2_C</td>
<td>TruSeq Index 4</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
<tr>
<td>Ibtk_C</td>
<td>3C_10#1_C</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt01</td>
</tr>
<tr>
<td>Ibtk_D</td>
<td>3C_10#1_D</td>
<td>TruSeq Index 6</td>
<td>HiSeq 2000</td>
<td>RWt02</td>
</tr>
</tbody>
</table>

Table 2.6 4C-Seq Libraries. 4C libraries for each candidate gene were prepared from separate 3C libraries to ensure the use of replicates. The same 3C library sources were used to prepare 4C libraries amongst sources. Multiplexing was facilitated by the use of indexed adapters to limit the project to 2 lanes on the HiSeq 2000.
probability of contact will generally be higher for loci that have a shorter linear distance (bp) separating them (Rippe 2001; Mirny 2011). This has been observed consistently in chromosome conformation experiments and has been substantiated as a general property with the identification of topologically associated domains (TADs) and the fractal configuration of chromosomes in the nucleus (Lieberman-Aiden et al. 2009; Dixon et al. 2012; Nora et al. 2012). It follows from these properties that in a 4C experiment, one will detect a higher level of background surrounding the viewpoint than at other regions of the chromosome. As such, it would be useful to have the capability of adjusting background thresholds across the lengths of chromosomes. Furthermore, it is expected that if two loci are interacting in a functional manner, then their neighboring sequences will also have a higher chance of interacting. Importantly, these expectations would not be found at interactions occurring by random chance.

At the time that this project was beginning, only one method for analyzing 4C-Seq data had been published (Splinter et al. 2011). This method, however, discards any quantitative information that can come from the sequenced reads to alleviate concerns over possible amplification bias that may occur during the library preparation. Instead, read counts are transformed to Boolean values, leading to 3C fragments being ruled simply as detected or not detected. This binarization of the data necessitates the use of a windowing strategy that pools at least 100 fragments into a single "interaction", thereby generating a lower-resolution analysis of the data. It was reasoned that if quantitative information from read counts could be used, however, more statistical power would be available to allow for the detection of much smaller domains as well as comparative analysis between alleles.

The presence of a variable barcode in the primers used to amplify the 4C libraries (Illustration 2.1) provides an indicator of the potential for amplification bias. Four random bases present on both the forward and reverse reads leads to more than 65,000 possible combinations that would be affixed to 4C products in the initial amplification round. Therefore, the degree to which reads containing identical barcodes contribute to the detection of a mappable event can be used as an indicator of how much excessive amplification of single capture events may be skewing the results. The ability to
account for possible amplification bias reduces the need to binarize 4C data as a precautionary step. To utilize the full range of quantitative information captured by 4C-Seq, the Magnuson Lab developed fourSig, a conceptually simple, yet powerful statistical method for analyzing 4C-Seq data (Williams et al.).

**Methodology**

Prior to statistical analysis with fourSig, a variety of information, including the size of the 4C primers, locations of the 3C, 4C, and linearization enzymes, and the size range of sequenced molecules is used to determine which 3C fragments can be identified by 4C. Fragments that are too large for the sequencer, too small to be aligned, or would not have been amplified by PCR are flagged such that they can be eliminated from the analysis (see Chapter 3). This allows fourSig to accurately model the random distributions used for threshold calculation (described below). The schematic in Illustration 2.2 describes the method for determining whether a signal is significant relative to background. Reads are first associated with the appropriate mappable 3C fragments (Illustration 2.2A). To determine the significance of a contact, a sliding window analysis is used to avoid introducing arbitrary window boundaries. The size of the windows, W, can be altered depending on the desired resolution of the contact map. Smaller window sizes provide higher resolution results, which are useful for investigating the details of specific interactions, while larger window sizes are better for showing general trends and identifying highly reproducible interactions.

For each chromosome, the total number of reads in each window is tallied to generate an observed distribution (Illustration 2.2B, Step 1). Then, a randomized distribution is generated by shuffling the observed reads amongst mappable 3C fragments in a manner that is independent of proximity to the viewpoint. The total number of shuffled reads per window is tallied (Illustration 2.2B, Step 2). The randomized data are used to calculate X, the minimum number of reads required for a window to be significant by satisfying a desired FDR (Illustration 2.2B, Step 3). The final cutoff value for significance of any given chromosome is calculated from a minimum of 1000 permutations, resulting in a distribution of Xs that can be used to define a minimum read count to qualify a window as significantly enriched (Illustration 2.2B, Steps 4 & 5). The permutations empirically derive a
Illustration 2.2  A Method for Determining Enriched Interactions. A. The sequence of captured fragments is mapped to the genome and the number of reads mapping to each 3C fragment is used as the observed distribution. B. (Step 1) Sliding windows of a desired number of 3C fragments, W, are demarcated and the total reads in each window are determined for the observed data. (Step 2) The reads on the chromosome are distributed randomly amongst 3C fragments and a new reads per window is calculated. (Step 3) A cutoff, X, is calculated for a desired FDR using the shuffled reads per window data. (Step 4) X is calculated from at least 1000 random shuffles to generate a histogram of possible cutoffs for the desired FDR. (Step 5) The final threshold for calling a significantly enriched window in the observed dataset is set at the 95th percentile for calculated Xs. In the depicted example, a desired FDR of < 0.01 leads to a final threshold of 40 reads per window for calling significant interactions in the 4C data.
distribution for X, which is used to choose a threshold with high confidence. Using the example demonstrated in Illustration 2.2, 99% of the Xs from shuffled reads were greater than 10, so it is likely that a cutoff as low as 10 reads in a window would be an underestimate of the true threshold for FDR $< 0.01$. Similarly, 50% of the shuffles generated Xs were greater than 30, so it is likely that 30 would also be an underestimate. However, only 5% of the shuffles generated cutoffs greater than 40, so it is reasonable to conclude that this threshold does not underestimate a FDR $< 0.01$.

*fourSig* employs two notable features to deal with common technical issues in the interpretation of chromosome conformation data. Due to the increased local contact probability associated with interacting regions and the disproportionate number of reads detected near the viewpoint, a single threshold calculated from the observed distribution of an entire chromosome could be set too high to detect very long-range interactions. While such a threshold may lead to more conservative significance calls, it is reasonable to desire the implementation of variable, position-dependent thresholds. To this end, *fourSig* allows the user to mask the reads for any specific region prior to determining a significance threshold. This is especially useful for the region surrounding the viewpoint because reads mapping to this fragment can sometimes be orders of magnitude larger than those observed at more distal locations. Using this feature, one can mask out as large or as small an area as desired, thereby allowing for limitations of distal background influence on threshold calculation.

Another important feature of *fourSig* is that it assigns a priority score for significant interactions, based on the distribution of reads within the window, which correlates with reproducibility (shown in Chapter 3). Because of the expected increase in local contact probability accompanying any given interaction that is not due to random chance, the likelihood that a contact may be consistently detected across replicates can be estimated by the shape of the curve generated by plotting the reads within a window. When an interaction is captured between two loci with significant frequency, it is expected that there will be a large number of reads for the fragment harboring the point of interaction (Illustration 2.3A). Due to various random events, such as incomplete digestions and increased local proximity, reads mapping to neighboring fragments should also be detected at a
Illustration 2.3 Prioritization of Enriched Contacts By Likelihood of Reproducibility. A. A model of a conformation capture event is shown (left) with a symbol legend (right). A frequently occurring interaction is expected to yield a distribution of reads mapping to locally available fragments and centering on the point of highest probability of contact (center). B. Peaks are assigned a priority level corresponding to expected distributions by reassessing the capacity of a window to exceed the significance threshold upon transformation of the read counts for the most abundant fragment (dashed line). Windows are categorized as Broad (Category 1) if the window is still significant when discounting the number of reads in the most abundant fragment, Intermediate (Category 2) if the threshold can be exceeded when the most abundant fragment is replaced with the average number of reads for the adjacent fragments, and Narrow (Category 3) if the significance threshold can only be exceeded when all data are included.
lower level. In the event of a consistent and frequent interaction, this behavior should give rise to a broader distribution of reads in the window because more opportunities for capture are available at the time of fixation.

Taking these expectations into consideration, additional criteria are added to the results after assessing the significance of a given window. For any given significant interactions, `fourSig` manipulates the fragment with the most reads mapped to it and re-evaluates whether the interaction still reaches our threshold of significance (Illustration 2.3B). A peak is described as being “Broad” (Category 1) if the fragment containing the highest number of reads can be removed and the remaining read count from all other fragments still exceeds the significance threshold. These types of peaks are the most likely to be reproduced (see Chapter 3). If this test fails, the value for the fragment with the highest read count is reduced to the average value of the two adjacent fragments. If the interaction can still exceed the significance threshold, we designate it as “Intermediate” (Category 2). Finally, if the interaction is only significant with the observed read counts, then it is designated as “Narrow” (Category 3).

**Optimization of FISH Methods**

Since an interaction that is involved in the regulation of gene expression should be present whenever the expression state is manifested, it was reasoned that detection of identical interactions across multiple replicates in which the associated expression state exists should provide a high-confidence indicator of functionally associated interactions (Sanyal et al. 2012). Nonetheless, the ability to validate the results using a non-redundant assay is highly desirable to ensure confidence that the 4C assays, and the analysis thereof, are identifying true and regular phenomena. Many previous studies have used FISH as a validation technique to confirm interactions observed in chromosome conformation experiments (Lieberman-Aiden et al. 2009; Splinter et al. 2011). Although FISH generally produces a much lower resolution analysis of relative proximity than can be extracted from conformation assays, general trends may be observed when validating interactions between loci.
positioned several hundred kb or a few Mb apart from each other. Previous FISH experiments in the Magnuson Lab have primarily used BACs, which can span several 100 kb, to prepare probes (Kalantry et al. 2009; Williams et al. 2011; Calabrese et al. 2012). However, successful validation of interacting loci would need to accurately pinpoint FISH probes such that the intervening distance is substantially larger than the areas canvased by the probes. Therefore, in addition to working out conditions for preparing tissue for analysis by FISH, it was deemed necessary to test the size limitations to reliable detection of DNA FISH probes.

**Permeabilization**

Standard protocols for FISH in the Magnuson Lab called for the permeabilization of cells using a cytoskeletal (CSK) extraction buffer with Triton X-100 (CSK-T) (Kalantry et al. 2009; Williams et al. 2011; Calabrese et al. 2012). CSK buffers are used during the permeabilization step to help remove soluble proteins that may contribute to higher background after hybridization. Consequently, CSK buffers have been used to remove substantial amounts of nuclear protein from tissues in structural studies of the nuclear matrix (Nickerson et al. 1997). Furthermore, permeabilization with CSK-T is typically performed prior to fixation. The potential for removing proteins that may be involved in genomic interactions prior to fixation led to the concern that these methods may not be suitable for the validation of interactions detected by 4C with FISH. Therefore, conditions were optimized to substitute phosphate-buffered saline (PBS) for CSK and to perform the permeabilization after fixing the cells. This should ensure that, similar to the harvest of cells for 3C libraries, the cells used for FISH are fixed as early as possible and that all proteins within the nuclear structure remain present for validation of detected interactions.

A previously prepared and tested DNA FISH probe for the Acbc7 locus (Williams et al. 2011) was used to test several permeabilization conditions on TS cells (Figure 2.9). Although at lower concentrations of Triton X-100 the PBS-based permeabilization was not as effective as CSK-T, post-fix permeabilization with 0.5% Triton X-100 in PBS (0.5% PBS-T) was equally as effective at producing consistent DNA FISH signal. Because no clear difference was observed between 1 minute and 30 second permeabilization steps (Figure 2.9D & E), the 1 minute incubation condition was
Figure 2.9 Optimal Permeabilization Conditions for FISH in TS Cells. TS cells were fixed in multiple conditions to determine optimal permeabilization conditions in the absence of CSK extraction buffer. A BAC probe for the Acbc7 locus was used for DNA FISH optimization and cells were permeabilized at room temperature with Triton X-100 in either a CSK buffer (A-C) or PBS (D-F). For this experiment, TS cells were fixed with 4% paraformaldehyde for 5 minutes on ice. CSK-T permeabilization (A-C) was performed prior to fixation while PBS-T permeabilization (D-F) occurred after fixation. All samples were treated with RNase A prior to hybridization. Images were taken at 100x magnification.
selected for ease of work-flow.

**Fixation**

Earlier protocols typically used 4% paraformaldehyde (PFA) fixation for up to 10 minutes (Kalantry et al. 2009; Williams et al. 2011; Calabrese et al. 2012). Because FISH was to be used here as a validation of 4C interactions, which are determined from tissue fixed for 10 minutes in 1% formaldehyde (See 3C Fixation Optimization), approximating similar fixation conditions was deemed important. To this end, observations were made of the consistency of DNA FISH signal obtained from various fixation conditions (Figure 2.10) Using a previously tested probe at the Ddx3x locus (Williams et al. 2011), TS cells were fixed and permeablized with the optimally determined conditions (see FISH Permeablization). DNA FISH signal was observed at all tested conditions but did begin to appear less consistent when fixed with 1% PFA for 10 minutes (Figure 2.10C). To improve consistency while still approaching a closer fixation condition to the 3C sample collection, 2% PFA fixation for 10 minutes was selected as the optimal condition.

**Minimization of Probe Size**

BACs are typically used to generate probes for DNA FISH experiments. BACs, however, tend to span several hundred kb in sequence. Anticipating that the resulting 4C experiments would yield interaction data at a much higher resolution than 150-300 kb, it was necessary to learn the lower size limit of reliable detection of DNA FISH probes. Fosmids are smaller than BACs; typically around 40 kb in length. To test the lower limit, a fosmid previously used as a DNA probe for the MeCP2 locus (Williams et al. 2011) was used in combination from probes consisting of pooled 5 kb amplification products. The 40 kb fosmid probe and 10 kb amplicon probe were tested individually and in co-hybridization to observe the consistency of visible foci. The fosmid probe (Figure 2.11A) was always robust and clean while the 10 kb probe (Figure 2.11B) displayed very high background and was difficult to distinguish real signals. Co-localization of the probes was tested to confirm the fidelity of foci resulting from the 10 kb probe (Figure 2.11C). True signal from the 10 kb probe, as evidenced by co-localization, was very infrequent. This eliminated the potential for using 10 kb probes in the absence of much more sophisticated optimizations. Provided that interactions selected
Figure 2.10 Optimal Fixation Conditions for FISH in TS Cells. TS cells were fixed on ice for (A) 5 minutes in 4% PFA, (B) 10 minutes in 2% PFA, and (C) 10 minutes in 1% PFA to best approximate conditions similar to fixation for 3C libraries without compromising reliability of hybridization. A BAC probe for the Ddx3x locus was used for DNA FISH optimization. TS cells were permeablized with 0.5% Triton X-100 in PBS after fixation. All samples were treated with RNase A prior to hybridization. Images were taken at 100x magnification.
Figure 2.11 Minimization of Size for FISH Probes in TS Cells. DNA FISH tests were performed using (A,C) a 40kb fosmid probe (white) for the MeCP2 locus and (B,C) pooled probes (red) from 5kb amplified fragments in the same region. (C) Co-localization of the two probes was assessed to determine if 10kb of sequence was sufficient for reliable detection of DNA FISH. 488nm channel (green) was imaged to test for auto-fluorescence.
for validations are far enough apart for the 40 kb range to be distinguishable, fosmid-derived probes would be acceptable for validation.

**Concluding Remarks**

In this chapter, a strategy was proposed for using allele-specific expression as a model to test the general availability and occurrence of long-range genomic interactions as regulatory loops. Additionally, preliminary tests and optimizations required to perform the proposed analysis were performed and are described. The successful development of these tools and methods sets the stage for carrying out the proposed work, the results of which are described in Chapter 3.
CHAPTER 3: RESULTS AND ANALYSIS

Introduction

In the previous chapter, a strategy was proposed to use 4C-Seq in an allele-specific manner to test the occurrence of long-range regulatory loops as a general principle of gene expression regulation. In short, the use of several replicates of TS cells from reciprocally crossed F1 embryos would be used to identify consistently occurring long-range interactions with high association to gene activity and address the ubiquity of their role in gene expression. Therefore, it was expected that the analyses resulting from this strategy would identify the existence and probable locations of functionally associated allele-specific long-range interactions. Many optimization tests and experiments, described in Chapter 2, were run to ensure that effective processes and tools were available to perform the proposed work. Several 4C-Seq experiments were carried out in accordance with the described strategy. This chapter describes the results of those experiments, subsequent analysis, and validation of the results.

Mapping Strategy and Sequencing Results

The creation of 4C libraries inevitably generates certain technical limitations that must be considered when analyzing 4C-Seq data. For one, the absolute resolution of the assay will be limited by the frequency of the restriction site for the enzyme used to create the initial 3C library. Additionally, the choice of enzymes used can result in fragments that are too large to sequence, too short to map to the genome, or will be cut in such a way that prevents PCR amplification. Therefore, it is necessary to perform a substantial amount of pre-processing and quality control on the raw data prior to analysis by fourSig.
Mapping and Processing Sequencing Reads

Each read is a hybrid of known (viewpoint fragment) and unknown (captured fragment) sequences (Figure 2.2, bottom). Since 4C libraries for several genes were pooled into a single sequencing library, the first step involves sorting the reads into groups based on their viewpoint identity and replicate origin (Illustration 3.1A). The FASTQ files returned from the High-Throughput Sequencing facility at the University of North Carolina at Chapel Hill (UNC) were separated by TruSeq index, so replicate origin was already separated in the raw data. First, each read was searched for the presence of the HindIII site that is supposed to be at the junction between the hybrid reads. This site serves as a point to split the read into viewpoint and captured fragments. Next, the viewpoint portion was searched for the sequence of the genomic primers used for amplification of the 4C libraries. This sequence provides the information needed to determine the 4C library to which the read belongs. Reads in which expected primers could not be found or have unexpected pairings were separated as having unidentified origins. Finally, the viewpoint fragment was searched for the presence of the expected SNP between B6 and CAST strains so that reads from each 4C library could be separated into B6 and CAST viewpoints.

The mm9 genome assembly of *M. m. musculus* for the B6 genome (Meyer et al. 2011) was used as a reference genome for mapping. A CAST genome sequence was generated by replacing the appropriate nucleotides in mm9 with reported SNPs for *M. m. castaneus* (Keane et al. 2011). After identification of the source allele and separation of the known sequence from the unknown captured sequences, the unknown portion of the reads were mapped to both the B6 and CAST genomes using the Bowtie algorithm (version 0.12.7) (Langmead et al. 2009) (Illustration 3.1A). 2 mismatches were allowed and only sequences that mapped uniquely to the genome were retained (settings \(-n 2 \-l 100 \-m 1 \-best \-strata\). When mismatches were found, the positions of the mismatches were checked against the sequence of both strains to determine whether the mismatch was due to an informative SNP (Illustration 3.1B). If so, allelic identity of the captured read was assigned.

Successfully mapped reads were then paired, where possible, to increase the likelihood of
Illustration 3.1 Strategy for Mapping 4C-Seq Reads. A. Reads from the FASTQ files are split at the recognition site for the 3C Enzyme. The viewpoint sequence is searched against the sequence for the amplification primers, supplied in a separate input file, to identify the locus and allelic origin of each fragment. Successfully sorted reads are mapped against genome sequences for both strains using the Bowtie algorithm. B. Allelic mapping is performed by searching mismatches in the mapped sequence from the aligned bowtie files against a supplied input file containing SNP annotations between the two mouse strains. Allelically and non-allelically mapped forward and reverse reads are paired in separate analyses. The result is two sets of mapped reads files, one with allelically mapped capture sequence and one with non-allelically mapped capture sequence, with auxiliary files for each set containing the sorting lists internal to the program.
detecting an informative SNP in the captured sequence. Reads were discarded in the event that the unknown portion of both reads did not map to identical loci. Ultimately, two sets of mapped reads files were generated for each 4C library; one in which allelic identity was not determined for the captured fragment (“non-allelic”) and one in which only reads with allele-specific SNPs were present in the captured fragment (“allelic”). Allelic origin of the viewpoint fragment, however, was always known in both sets of mapped reads. The “allelic” mapped reads files tend to be limited in data volume relative to the “non-allelic” files because captured sequence does not always contain an informative SNP. To maximize the volume of data used to call significant interactions, “non-allelic” reads files were adjusted by removing known allelic reads that are inappropriate for the specific mapped reads file (e.g., reads containing no ambiguities and SNPs for the CAST allele were removed from “non-allelic” files aligned to B6). Additionally, reads in which the unknown portion maps to the viewpoint were removed as these are indistinguishable from possible self-ligation events. The final result is 4 “allelic” and “non-allelic” mapped reads files for each allele (B6 to B6, B6 to CAST, CAST to B6, and CAST to CAST). The adjusted “non-allelic” mapped reads files were used to generate input files for fourSig to maximize data available for analysis of cis interactions while the “allelic” files were used again later to assign probable allele identity to the captured loci (see Chapter 3).

Because fourSig uses read counts to determine significance threshold, distinguishing PCR amplification bias from true multiple detections can be a source of concern. With the primer design strategy described in Chapter 2, this distinction can be made due to the inclusion of random barcodes incorporated between sequencing adapter sequences and 4C genomic primer sequences (Figure 2.2, bottom). Repeated amplification of single contacts will generate reads containing identical barcodes on both ends of the read, creating an indicator of the degree to which amplification bias may be affecting the detection of a mapped 3C fragment. To make this assessment, the barcodes in each of the 4C-Seq datasets were analyzed after sorting the mapped reads by allelic origin. Specifically, 3C fragments in which a single barcode could account for more than 15% of the reads mapped to a single fragment were selected for further analysis. Interactions in which this occurred were then re-evaluated to test if the interaction still reached the threshold when all reads containing identical barcodes were flattened to a single count. In each of the 4C experiments, this analysis found few, if
any, fragments containing a disproportionate number of reads with identical barcodes. No event was ever detected in which the interaction no longer reached the significance threshold once repetitive barcodes were flattened (data not shown).

Allelic 4C-Seq Test Experiment at Ibtk Locus

The sequencing of 28 4C-Seq libraries is an expensive and resource intensive prospect. To ensure that the methods developed and described in Chapter 2 were effective, a test run for a single 4C library was desired. The totality of the 4C-Seq experiments were intended to be sequenced using 2 x 100 bp on the Illumina HiSeq 2000 platform, which is capable of producing up to 6 billion paired-end reads on a single paired-end run (Illumina, Inc. 2011). Since this volume of data is not necessary or cost effective for a single, non-multiplexed run, 2 x 150 bp sequencing on the MiSeq platform was chosen for the test library. This platform, capable of up to 30 million paired-end reads (Illumina, Inc. 2013), produces a much smaller output but is quicker and more cost effective for a single experiment. Therefore, as a pilot experiment, a 4C library was generated from the TSS of the Ibtk locus in the CASTB6F1 TS cell line and was sequenced using the Illumina MiSeq platform. Approximately 2 million raw paired reads were returned from the MiSeq run with B6 and CAST having a near equal representation of sequenced viewpoints (Figure 3.1A). The reverse reads, however, saw a substantial loss in read quality and ability to detect the viewpoint sequence. For reads that were successfully mapped using the allelic and non-allelic analyses described above, similar balance in allele distribution was seen (Figure 3.1B).

This pilot library, though limiting in sequence depth, was analyzed by fourSig and several long-range interactions were identified as significant above background. fourSig was run unmasked using a window size of 5 fragments to calculate a significance threshold from the top 5\textsuperscript{th} percentile of 1000 calculations at FDR < 0.01. For visualization and comparative analysis, significant windows were transformed to 3C fragments within the windows and reads per fragment were plotted along the length of Chromosome 9 (Figure 3.2A). Despite the lack of replicates in the pilot experiment, a clear difference in interaction profile was apparent between alleles. 91 fragments within cis-interacting windows were found in common between both Ibtk alleles while 453 and 262 unique fragments
Figure 3.1 Reads Distribution for Ibtk Allelic 4C-Seq Test Experiment. A. Approximately 2 million paired reads were obtained from the pilot experiment on the MiSeq. Forward and reverse reads distributions are plotted as a percent of total reads. B. Allelic distributions for reads mapped by both allelic (AS) and non-allelic (NAS) analyses are plotted.
Figure 3.2 Significant Interactions and Allelic Comparison for Ibtk 4C-Seq Test Experiment. A. Reads per 3C fragment within significant windows are plotted for cis interactions. Green rectangle represents the location of the viewpoint. B. Allelic comparison of fragments within significant windows reveals some overlap between alleles.
within cis-interacting windows were specific to B6 and CAST alleles, respectively (Figure 3.2B). Common interactions would indicate similarity in the conformation of both B6 and CAST chromosomes and the unique interactions may hold possible sources for variation in allele regulation. While not sufficient in depth to draw effective conclusions, the resulting data demonstrated that the optimized experimental conditions and the fourSig analysis pipeline were sufficient for performing the proposed investigation.

**Full Allelic 4C-Seq Experiment on All Candidates**

Having successfully sequenced a test 4C-Seq library and extracted good results with the developed analysis pipeline, libraries for the complete experiment were generated and pooled for 2 x 100 bp sequencing using the Illumina HiSeq 2000 platform at the UNC High-Throughput Sequencing facility (Table 2.6). Two lanes were sequenced, each containing 4C libraries for each of the 7 selected TSS from CASTB6F1 (TruSeq Index 4) and B6CASTF1 (TruSeq Index 6) TS cells. Initially, just as was done with the MiSeq ibtk run, the FASTQ data files were analyzed using the FASTX-Toolkit (Hannon Laboratory) to assess the quality of the sequence. Much like the pilot library, Phred quality scores suggested a significant loss of quality in the reverse reads, relative to forward reads, which worsened early on in the read length (data not shown). While, on average, quality was reduced for reverse reads, acceptable quality still existed in sufficient quantity to assist in the detection of informative SNPs for mapped reads.

Next, the processing and mapping pipeline (described above) was applied to each of the FASTQ files. A significant setback was immediately noticed as the large majority of the reads failed to successfully align (Figure 3.3). For the forward reads, the majority of reads failed to contain a HindIII recognition site indicating a ligation junction (No Enzyme). Similar to the test run, the expected viewpoint sequence could not be associated with the majority of the reverse reads (No Primer). Other technical problems, including ambiguous base calls at SNP positions (Ambiguous SNP), too many mismatches (Primer Mismatch), sequences too small to align and pairs containing viewpoints from two different genes (Other), made up less than 10% of the reads in reverse primers and a negligible proportion in the forward reads.
Figure 3.3 Low Proportion of Mappable Reads in Large Allelic 4C-Seq Experiments.
A. The 4 HiSeq 2000 libraries returned an average of 66 million raw paired-end reads per library. For each library, the proportion of raw reads grouped by the mapping pipeline is plotted. The aligned category makes up the proportion of reads that were successfully mapped.
Having lost such a large proportion of the data in all sequencing runs, an attempt was initially made to identify what went wrong and if anything could be done to salvage some of the reads removed from the mapped data in the analysis pipeline. The thought was that, perhaps, an unforeseen circumstance may have caused the software to sort otherwise useful data into the non-useful groups. If a reason could be found that explained the discard of most of the data, perhaps this could be corrected and recovered prior to analysis by fourSig. Efforts were focused on the forward reads because the general read quality of the reverse reads was low. Additionally, since the majority of the data in all FASTQ files was placed into the “No Enzyme” category, which is sorted after viewpoint sequence has been successfully associated with the read, troubleshooting efforts were focused there.

Forward reads from an Airn No Enzyme file were visually inspected to identify patterns. It was immediately noticed that each read began with the 4 bp random barcode and the start of the forward genomic primer for the 4C viewpoint (Figure 3.4A). Complete sequence for the genomic primer was not present in most of the observed reads and the length of expected sequence varied amongst reads. Further analysis revealed that the remainder of the reads frequently contained sequence matching the reverse-complement of the TruSeq Indexed adapter found on the opposing primer. The similarity of sequence found between a portion of the genomic primer and the reverse-complement of the adapter on the opposing primer may indicate the formation of undesirable primer dimer in the amplification of the 4C library. A Perl script was written to search each read in the files for sequence matching the reverse-complement of the opposing primer and the majority of the reads could be accounted for by unintended amplification of primer dimer (Figure 3.4B). Reads in which no alignment to the opposing primer was found were aligned to the mouse genome using the BLAT tool (Kent 2002) and most forward reads mapped to the genome but at various loci different from the viewpoint. Similar analysis was performed on the other candidate genes and, with the exception of Igf2r, a similar problem was found in most of the No Enzyme files (data not shown). In summary, the majority of the loss to the No Enzyme designation can be accounted for by unintended amplification of primer dimer in the 4C libraries (Figure 3.4C), making them unsalvageable for further analysis.
Figure 3.4 Proportion of Reads Lost to Primer-Dimer Contamination. A. Visual analysis of reads in the No Enzyme category for *Aim* data revealed consistent detection of four random bases followed by a portion of the genomic primer and reverse complement of the opposing primer. An example model for a forward read is depicted. B. Proportion of No Enzyme category that can be accounted for by amplification of opposing primers for *Aim* is plotted. Random Location refers to reads not aligning to primer sequence that can be aligned to other areas of the mouse genome. Unknown refers to reads that could not be aligned to either the opposing primer or the mouse genome. C. The proportion of raw reads grouped by the mapping pipeline is plotted with the No Enzyme category replaced to reflect the results of primer dimer analysis.
<table>
<thead>
<tr>
<th>4C Library</th>
<th>Airn</th>
<th>Igf2r</th>
<th>Ibtk</th>
<th>Car2</th>
<th>Bphl</th>
<th>Mtus1</th>
<th>Tlr5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASTB6F1 #1</td>
<td>1.57E+06</td>
<td>0.00E+00</td>
<td>4.87E+06</td>
<td>2.06E+06</td>
<td>2.25E+06</td>
<td>8.92E+05</td>
<td>1.71E+04</td>
</tr>
<tr>
<td>CASTB6F1 #2</td>
<td>2.13E+04</td>
<td>1.00E+00</td>
<td>6.66E+06</td>
<td>8.04E+06</td>
<td>5.45E+06</td>
<td>2.25E+04</td>
<td>2.87E+05</td>
</tr>
<tr>
<td>B6CASTF1 #1</td>
<td>1.17E+05</td>
<td>0.00E+00</td>
<td>1.02E+05</td>
<td>2.09E+06</td>
<td>1.44E+06</td>
<td>2.28E+04</td>
<td>9.54E+04</td>
</tr>
<tr>
<td>B6CASTF1 #2</td>
<td>3.23E+05</td>
<td>1.70E+01</td>
<td>1.41E+05</td>
<td>8.69E+06</td>
<td>3.93E+06</td>
<td>1.58E+04</td>
<td>1.81E+06</td>
</tr>
</tbody>
</table>

Table 3.1 Total Mapped Reads for 4C-Seq Libraries. Total numbers of mapped reads for each library are displayed in the table. Libraries with more than half a million mapped reads are bolded. Proportions mapping to B6 or CAST depicted in Figure 3.5 are relative to the numbers here.
Having resolved to move forward with the analysis of aligned data, the distributions of mapped reads for each experiment were observed to identify useful datasets for replicate analysis by *fourSig*. Many of the libraries returned fewer than half a million mappable reads, limiting the number of replicates available for each gene (Table 3.1). Additionally, of the genes for which more than one library with substantial depth existed, only *Ibtk* had reads mapped in the expected equal proportion between alleles (Figure 3.5). Skewed distributions in favor of the B6 allele for *Bphl* and *Car2* imply some sort of preferential loss of alignability for CAST reads. Since this could be problematic for comparative analysis between alleles, replicate analysis was focused only on the *Ibtk* datasets using the two CASTB6F1 sequence libraries and the data from the test library (above) as a triplicate dataset.

**Set Analysis and Results for Allelic 4C-Seq at *Ibtk***

The availability of the pilot experiment for *Ibtk* offered the opportunity to perform a thorough replicate analysis of at least one allele-specific 4C-Seq experiment. Since the allelic expression bias for *Ibtk* is consistently strain-specific, rather than dependent on parent-of-origin, the lack of libraries from the B6CASTF1 TS cell line was not a strong concern. Therefore, 4C-Seq for the *Ibtk* TSS was performed in triplicate with the pilot library serving as one replicate (*Ibtk*-Rep 1) and the two CASTB6F1 libraries run on HiSeq 2000 as the other two replicates (*Ibtk*-Rep 2 and *Ibtk*-Rep 3).

Each replicate dataset was independently analyzed with *fourSig* to identify and prioritize windows representing significantly enriched interactions. Since the proposed hypothesis was that allele-specific contacts may harbor putative distal regulatory elements and planned validation experiments consisted of FISH, window sizes needed to be kept small to achieve high-resolution results. Therefore, an unmasked *fourSig* analysis was run on *Ibtk* 4C datasets using a window size of 5 fragments to calculate a significance threshold from the top 5th percentile of 1000 calculations at FDR < 0.01 (Illustration 2.2). Thresholds were assessed and enriched windows called independently for each chromosome. However, since the purpose of this investigation was to assess
Figure 3.5  Mapped Reads Distributions for Allelic 4C-Seq Experiments. The number of mapped reads for each 4C-seq library is plotted as a proportion of identified allelic viewpoints. CASTB6F1 cell lines are labelled as C/B and B6CASTF1 cell lines are labeled as B/C.
the potential for functional association in the regulatory state of a gene, and because evidence for trans-associated regulatory elements is limited and not generally considered as probable or common as with cis-association (Miele and Dekker 2008), analysis was preferentially limited to cis interactions. Therefore, for the *Ibtk* analysis, trans interactions mapping to the opposing Chromosome 9 to the analyzed viewpoint were excluded.

The allelic origin of all interactions was already known due to presence of an informative SNP in the viewpoint sequence. Since adjusted “non-allelic” mapped reads files were used in the *fourSig* analysis, however, allelic identity of capture sequences remain somewhat ambiguous. After enriched interactions were determined, allelic calls for the significant windows were made based on whether the number of SNPs in SNP-containing reads comprising the enriched window is significantly higher than what would be expected by random chance (Calabrese et al. 2012; Mugford et al.; Williams et al.). Interactions mapping to Chromosome 9 that lacked informative SNPs were assumed to be in cis. Extrapolation of the average proportion of trans calls in each library would suggest that less than 1% of the interactions may be erroneously called cis, making this a reasonably safe assumption (data not shown).

The positions of significant windows were cross-referenced with fragment read data to transform the interaction data into congruent coordinates, based on 3C fragments, between replicate libraries for set analysis. To focus analysis on the locations of precise overlap between replicates, individual 3C fragments found within the enriched cis interaction windows were compared among the three replicates for each viewpoint allele. The intersection of these libraries is defined as the set of consistent interacting fragments detected for each viewpoint (Figure 3.6). When generating intersection lists of enriched fragments, the highest interaction prioritization score (Broad > Intermediate > Narrow) was retained as the window’s assignment. The utility of the peak prioritization algorithm can be demonstrated through the distribution of Category calls amongst the replicate sets. The overwhelming majority of the fragments that were consistently detected amongst replicates tend to belong to Broad (Category 1) interactions (Figure 3.7A). Furthermore, in each library a substantially larger portion of fragments belonging to Broad interactions were present in the
Figure 3.6 Set Analysis of *Ibtk*-Seq Libraries. A-B. 3C Fragments found in enriched windows have similar proportions of overlap between replicate 4C-Seq experiments. A similar number of enriched fragments were found in the intersection of replicates for both the B6 (A) and CAST (B) alleles.
Figure 3.7 Validation of *fourSig* Prioritization Algorithm. Peak priority classifications for different sets of interacting fragments are displayed. Fragments designated as Broad are in purple, Intermediate in green and Narrow in red. A. Priority Categories for replicated enriched fragments are shown as a proportion of the total intersection. Proportion of interactions assigned a classification in each replicate are averaged for B6 and CAST alleles for (B) fragments found in the replicate intersection and (C) fragments unique to a replicate.
intersection of the three datasets (Figure 3.7B). In contrast, the majority fragments belonging to Narrow interactions were unique to that particular replicate (Figure 3.7C). These data demonstrate the ability of this feature to predict the likelihood that an interaction may be consistent among replicates.

Fragments from Broad interactions that were represented in all three 4C-Seq datasets were selected for further analysis to ensure that the focus is on high confidence contacts. To visualize the results from the experiments, reads from each dataset were summed per 3C fragment and plotted along Chromosome 9 (Figure 3.8A). The majority of reads from repeated interactions were located within 3 Mb of the viewpoint. Additionally, the majority of reads found within these contacts were shared between alleles (~66%), especially in the region upstream of the \( \text{Ibtk} \) TSS (Figure 3.8). As the distance from the viewpoint increased, allele-specific differences in the location of contacts became more apparent. In fact, at a distance of 3 Mb, interactions approach near allelic exclusivity. Despite sharing a largely similar spatial organization, there were a substantial number of allele-specific contacts that are associated with one allele versus the other.

Closer visual analysis of the allelic interaction data revealed that, in some instances, strong differences in read counts can indicate a preferential allelic contact even when the 3C fragment in question is found in interactions called for both alleles. Notably, a 3C fragment located within the \( \text{Ibtk} \) locus was identified containing substantially more reads interacting with the CAST allele versus the B6 allele (Figure 3.9). Further examination of previous chromatin data (Calabrese et al. 2012) at this fragment revealed a combination of genomic features associated with active enhancer elements, namely H3K27ac, H3K4me1, and DNaseI hypersensitivity (Heintzman et al. 2007; Creyghton et al. 2010; Ernst et al. 2011). This analysis demonstrated that the CAST 3C fragment was enriched for these marks relative to the B6 fragment. While this fragment is located only about 50 kb downstream of the TSS for \( \text{Ibtk} \), the difference in read counts between the alleles is striking. Also of interest is that a previous study of the human \( \text{Ibtk} \) locus identified the existence of a TSS for an alternative transcript for \( \text{Ibtk} \) located in approximately the same relative location (Spatuzza et al. 2008). This raises the possibility that, in addition to acting as a TSS for an alternative transcript, modification of this site may
Figure 3.8 Allelic Comparison of Broad Ibtk Significant Contacts. A. Distributions of detected reads (black = B6 and red = Cast) for replicated, Broad (Category 1) interacting fragments were plotted along Chromosome 9 (upper panel). The region surrounding the Ibtk locus (blue box) is expanded (lower panel). B. Intersection of Broad cis interactions between the B6 and CAST alleles. C. A proportional breakdown of the allelically unique and common interaction sets by distance from the Ibtk TSS.
Figure 3.9 Allelic Bias in Contact Probability at a Putative Enhancer for \textit{Ibtk}. UCSC Genome Browser screenshot of 4C interaction data and selected chromatin data at the \textit{Ibtk} locus is shown. Exons 24 and 25 of \textit{Ibtk} are highlighted with arrows. The 3C fragment containing the \textit{Ibtk} TSS is indicated by a dashed brown rectangle. A putative intragenic enhancer aligning with a previously reported regulatory element is enclosed by a solid black rectangle. For ChIP-seq tracks, green boxes indicate biallelic enrichment, blue boxes indicate CAST-specific enrichment, and gray boxes indicate insufficient SNP detection to make an allele-specific call.
act as an enhancer whose contact with the promoter is necessary for activating transcription of the gene. The precedent of this locus for having regulatory function combined with the presence of active histone marks and the detection of allele-specificity in the 4C-Seq analysis lends credibility to the ability of this strategy and analysis by fourSig to identify interactions with putative distal regulatory elements.

**Validation of *Ibtk* Contacts by FISH**

In order to test the validity and allelic nature of the interactions identified by the 4C experiment, three interactions were selected to test by FISH. DNA-FISH probes were made to measure two allele-specific *cis* interactions, one for B6 and one for CAST, and one interaction common to both alleles (Figure 3.10A). Since the majority of allele-specific contacts were found within 1-3 Mb of the TSS, interactions for verification were chosen from loci within this space. To ensure that the FISH test was reflective of the high resolution and precision with which fourSig calls interactions, loci for validation of allele-specific contacts were selected so that the called interactions were not isolated by large linear distances from interactions found on the opposing allele. Fosmids containing inserts that map to the desired loci were ordered from the BACPAC Resource Center. Correct identity of the insert for each fosmid was verified by restriction fragment length polymorphism (RFLP) fingerprinting analysis with HindIII and EcoRI. Successfully validated fosmids were labeled using the BioPrime DNA Labeling System from Invitrogen and used to probe DNA at the sites of predicted interactions and the region immediately upstream of the *Ibtk* gene (Table 3.2). Alexa Fluor 488-dUTP (Molecular Probes) was used to label the *Ibtk* TSS while Cy5-dCTP (GE Healthcare) was typically used to label fosmids for the interacting loci.

One caveat to using FISH to validate discovered interactions is the need to replicate allele-specific resolution. The 4C-Seq assays described here use differences in SNPs between the B6 and CAST strains to infer allelic origin of interactions. It is highly unlikely, however, that a mismatch of a single nucleotide would be enough to preferentially disrupt hybridization of a FISH probe to allow...
Figure 3.10  Probe Design for Validation of Interactions by FISH.  A. Fosmid probes were selected against a B6-specific interaction (2), a CAST-specific interaction (3), and a region with interactions common to both alleles (4). A fosmid probe covering the Ibtk TSS but upstream of the transcribed region was also selected to detect the gene locus (1). B. To detect Ibtk RNA, a mass-weighted cDNA probe was prepared by pooling purified RT-PCR product from six intron-spanning amplicons. The pooled probe contains just over 1 kb of total DNA sequence.
<table>
<thead>
<tr>
<th>Probe</th>
<th>Type</th>
<th>Source</th>
<th>Location</th>
<th>Label</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibtk RNA Probe</td>
<td>cDNA</td>
<td>RT-PCR</td>
<td>chr9:85590729-85637104</td>
<td>Cy3</td>
<td>4C Validation</td>
</tr>
<tr>
<td>Ibtk Probe 1</td>
<td>Fosmid</td>
<td>WI1-1164K11</td>
<td>chr9:85642568-85680931</td>
<td>AF-488</td>
<td>4C Validation</td>
</tr>
<tr>
<td>Ibtk Probe 2</td>
<td>Fosmid</td>
<td>WI1-1905A07</td>
<td>chr9:87194252-87232416</td>
<td>Cy5</td>
<td>4C Validation</td>
</tr>
<tr>
<td>Ibtk Probe 3</td>
<td>Fosmid</td>
<td>WI1-1702H10</td>
<td>chr9:88221104-88263244</td>
<td>Cy5</td>
<td>4C Validation</td>
</tr>
<tr>
<td>Ibtk Probe 4</td>
<td>Fosmid</td>
<td>WI1-0698D21</td>
<td>chr9:87954089-87992722</td>
<td>Cy5</td>
<td>4C Validation</td>
</tr>
<tr>
<td>Abcb7</td>
<td>BAC</td>
<td>RP24-274B9</td>
<td>chrX:101481493-101602397</td>
<td>Cy3</td>
<td>Optimization</td>
</tr>
<tr>
<td>Ddx3x</td>
<td>BAC</td>
<td>RP23-337M21</td>
<td>chrX:12750900-12985935</td>
<td>Cy3</td>
<td>Optimization</td>
</tr>
<tr>
<td>MeCP-40kb</td>
<td>Fosmid</td>
<td>WI1-0894A05</td>
<td>chrX:71271660-71315314</td>
<td>Cy5</td>
<td>Optimization</td>
</tr>
<tr>
<td>MeCP-10kb</td>
<td>Pooled PCR</td>
<td>Fosmid PCR</td>
<td>chrX:71279048-71288815</td>
<td>Cy3</td>
<td>Optimization</td>
</tr>
<tr>
<td>3C Ligation Control</td>
<td>BAC</td>
<td>RP23-356F10</td>
<td>chr6:125044436-125275813</td>
<td>None</td>
<td>3C Control</td>
</tr>
</tbody>
</table>

Table 3.1 BACs, Fosmids, and Probes. Materials used for FISH Probes and a 3C Ligation control are described.
single copy detection. Therefore, since \( Ibtk \) exhibits very strong bias of expression for the CAST allele, it was reasoned that monoallelic detection of \( Ibtk \) transcript with RNA FISH co-localized with a DNA FISH probe to the for the \( Ibtk \) TSS could be a suitable indicator of allele identity. A cDNA probe to detect the \( Ibtk \) transcript was prepared from a mass-weighted pool of intron-skipping RT-PCR products (Figure 3.10B, Table 3.3) and labeled with Cy3-dCTP (GE Healthcare) using the BioPrime kit. The combined sequence of all the amplicons used in this pool span an area of approximately 1 kb. Previous experiments to determine the limiting range of sequence coverage needed to reliably detect a FISH signal suggest that not enough area is covered by this probe to detect DNA (see Chapter 2). Since transcribed product, however, would likely be present in multiple copies, such a small coverage should be suitable for detecting RNA. Indeed, tests performed with and without RNase treatment prior to hybridization demonstrated that, while untreated samples contain signal from the cDNA probe in the majority of observed colonies, signal from this probe is never detected in TS cells that have been treated with RNase (data not shown). Additionally, when the cDNA probe was co-hybridized with a fosmid probe against the \( Ibtk \) TSS, monoallelic expression was observed in 92.8% of scored nuclei, confirming that the use of RNA-FISH is an acceptable indicator for allelic discrimination (data not shown).

For each experiment, DNA-FISH probes for the \( Ibtk \) TSS and the assayed interaction were co-hybridized with the \( Ibtk \) cDNA probe. Z-stack images were taken in grayscale at 63X using a Zeiss AxioImager M2 equipped with an AxioCam MRm camera (Carl Zeiss). Merged Z-stacks were then deconvolved using the Iterative algorithm in the AxioVision software package (Carl Zeiss MicroImaging GmbH 2011). In order to be scored for measurement, nuclei were required to have clear boundaries in the DAPI stain, biallelic signal from both DNA probes, and a clear RNA signal that overlaps with the \( Ibtk \) locus (Figure 3.11A & B). Distance measurements were taken in 3 dimensions from the approximate centers of DNA signals using ZEN 2011 (Carl Zeiss Microscopy GmbH 2012) in at least 50 scored nuclei for each experiment. For the allele-specific interactions, the measured distance between \( Ibtk \) and the interacting locus was consistently smaller for the expected allele (\( p \)-value < 0.001). The repressed allele was consistently closer in space to the B6-specific interaction, whereas the opposite was found for the CAST-specific interaction (Figure 3.11C & D, respectively).
<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
<th>Sequence (5’-&gt;3’)</th>
<th>Ori</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>IbtkExonOligo_1F2</td>
<td>Forward primer in Exon 2</td>
<td>GTGTTGGACTGGCTGATTGA</td>
<td>-</td>
<td>chr:85637085-85637104</td>
</tr>
<tr>
<td>IbtkExonOligo_1R3</td>
<td>Reverse primer in Exon 3</td>
<td>TGAGTTGGCCTGTCTTCTCAT</td>
<td>+</td>
<td>chr:85631095-85631114</td>
</tr>
<tr>
<td>IbtkExonOligo_2F1</td>
<td>Forward primer in Exon 4</td>
<td>CGGAAGCCAGAATAGCAAAC</td>
<td>-</td>
<td>chr:85628635-85628654</td>
</tr>
<tr>
<td>IbtkExonOligo_2R1</td>
<td>Reverse primer in Exon 5</td>
<td>CGGGACAGAAACACCGAAT</td>
<td>+</td>
<td>chr:85626310-85626329</td>
</tr>
<tr>
<td>IbtkExonOligo_3F1</td>
<td>Forward primer in Exon 6</td>
<td>CAGGCTTTTGGAAGGACTGT</td>
<td>-</td>
<td>chr:85624837-85624856</td>
</tr>
<tr>
<td>IbtkExonOligo_3R2</td>
<td>Reverse primer in Exon 7</td>
<td>GAGTCCACAGCACCCTATGGA</td>
<td>+</td>
<td>chr:85622325-85622344</td>
</tr>
<tr>
<td>IbtkExonOligo_4F3</td>
<td>Forward primer in Exon 9</td>
<td>GGTCTTGTGCTCTGGAGGTTG</td>
<td>-</td>
<td>chr:85620302-85620322</td>
</tr>
<tr>
<td>IbtkExonOligo_4R1</td>
<td>Reverse primer in Exon 10</td>
<td>ATTGACACCGGGGAGGATGA</td>
<td>+</td>
<td>chr:85617761-85617780</td>
</tr>
<tr>
<td>IbtkExonOligo_5F1</td>
<td>Forward primer in Exon 16</td>
<td>AGTGGTTGTGCTGGCTGTA</td>
<td>-</td>
<td>chr:85611214-85611233</td>
</tr>
<tr>
<td>IbtkExonOligo_5R2</td>
<td>Reverse primer in Exon 17</td>
<td>TCGCAGCTTTAAGGCTTAA</td>
<td>+</td>
<td>chr:85609003-85609023</td>
</tr>
<tr>
<td>IbtkExonOligo_8F2</td>
<td>Forward primer in Exon 24</td>
<td>CAATTCAGGAACGAACCGA</td>
<td>-</td>
<td>chr:85596674-85596693</td>
</tr>
<tr>
<td>IbtkExonOligo_8R1</td>
<td>Reverse primer in Exon 25</td>
<td>GCAACATTAGGGCTGGAGA</td>
<td>+</td>
<td>chr:85590729-85590748</td>
</tr>
</tbody>
</table>

Table 3.3 Primers for Ibtk cDNA Probe. Primers designed to amplify cDNA made from spliced transcripts are described.
Figure 3.11 Validation of Selected Interactions by FISH. A-B. Representative FISH images of (A) a common interaction and (B) a B6-specific interaction are displayed. Active allele (arrowhead) was determined by co-localization of RNA signal (red) with the *Ibtk* TSS (green). Distance was measured from the centers of the DNA probes (green, white). Nucleus is counterstained with DAPI (blue). C-E. Percent of scored nuclei are plotted against distance between the probed locus and the *Ibtk* TSS. Measured loci represent interactions identified by 4C to be specific to the (C) B6 allele (n=56), (D) CAST allele (n=51), or (E) common to both (n=51). Measurements of active alleles are traced in red, while repressed alleles are traced in black. All distances are in µM. Significance of trends was determined using an exact binomial test.
No discernible difference in the measured distances between loci was detected for the interaction common to both alleles (p-value > 0.5, Figure 3.11E). These results validate the interactions identified by the 4C-Seq experiments and fourSig analysis.

Analysis of Usable 4C-Seq Libraries

Upon further review, some of the 4C libraries that had been previously dismissed for technical reasons appeared to have potential for analysis with fourSig. The libraries for Airn had initially been dismissed due to low numbers in three of the replicates (Table 3.1). However, if the Ibtk pilot experiment from the MiSeq platform is used as a benchmark for the number of reads needed for analysis with fourSig, this assessment may need to be re-visited. Although fewer total interactions were called in the test experiment than in the HiSeq 2000 runs, fourSig was able to call a set of enriched interactions with similar level of overlap to each of the two larger libraries as the two larger experiments exhibited between each other with approximately 690 thousand mapped reads (Figure 3.6). The second replicate from the B6CASTF1 TS cell line for Airn returned approximately half this number of mapped reads, which may be enough to perform fourSig analysis to identify enriched interactions that can be compared with the CASTB6F1 experiment that returned 1.5 million mapped reads (Table 3.1). Additionally, since allelic analysis with fourSig is performed on each allele as independent datasets, the uneven distribution of CAST versus B6 reads seen in the Bphl and Car2 datasets may not be a substantial barrier to comparative analysis. Both of these libraries had a substantial overrepresentation of mapped reads originating from the B6 allele; however the number of total mapped reads from all four replicates of each was high enough to ensure that the CAST allele contained several hundred thousand mapped reads for each replicate (Table 3.1, Figure 3.5). Therefore, allelic 4C analysis was performed for all four replicates for Bphl and Car2 as well as on the CASTB6F1 #1 and B6CASTF1 #2 replicates.
Allelic 4C-Seq at Airn TSS

Each allele from the first replicate from the CASTB6F1 line (large Airn) and the second replicate from the B6CASTF1 line (small Airn) was independently analyzed with fourSig to identify and prioritize windows representing significantly enriched interactions. The same conditions used for the Ibtk analyses (see “Set Analysis and Results for Allelic 4C-Seq at Ibtk”) were used for the Airn datasets. Again, analysis was focused only on the cis interactions and cis was assumed for interactions mapping to Chromosome 17 in the absence of allelic data to indicate otherwise. After transposition of the enriched windows to underlying 3C fragments with mapped reads, the interacting fragments from the two replicates were compared to determine exclusivity of interactions related to both allelic influence (B6 vs CAST, Figure 3.12A & B, respectively) and parental influence (Maternal vs Paternal, Figure 3.12C & D, respectively).

Unlike the other candidate genes, the existence of a true imprinted regulation at this locus raises the possibility that interactions may be influenced by either strain (allelic) effects or parental effects. Therefore, if the imprinted regulation is the predominant driver of intrachromosomal interactions at the locus, one would expect that the parental comparison may yield more similarity between replicates than the allelic. Based on the percent of the total data between replicates (union) that is made up by the intersection between them, the allelic comparisons result in a slightly higher similarity than the parental; however, neither comparison is substantially higher in similarity than the other (Figure 3.12E). These data may suggest that either both allelic and parental origins exert similarly proportioned but different influences over interactions, or that neither association is very important in explaining the observed interactions.

Analysis of the Broad replicated interactions from both allelic and parental associations revealed that each grouping results in the same number of 3C fragments found in common between the sets (Figure 3.13A). To determine whether these intersections between allelic and parental comparisons were similar, each parental set was compared to the allelic sets. For both maternal and paternal sets, the 32 interacting fragments shared by B6 and CAST alleles were also found in the parental groupings (Figure 3.13B). Additionally, comparison of these intersections and set
Figure 3.12 Set Analysis of Airn 4C-Seq Libraries. A-D. The intersection of 3C fragments found in enriched windows is shown for (A) B6 and (B) CAST alleles, as well as for (C) maternal and (D) paternal origin. CAST alleles are represented by red areas and B6 alleles by black areas. Interactions from the B6CASTF1 (B/C) line are while CASTB6F1 (C/B) appear as diagonals. E. The similarity of libraries with respect to alleles is displayed as the percent of the union of the compared sets occupied by the intersection.
Figure 3.13 Allelic Comparison of Broad Airn Significant Contacts. A. Intersection of replicated Broad 3C fragments grouped by allelic (left) and parental (right) origin. B. Overlap of replicated Broad 3C fragments from maternal (left) or paternal (right) origin with allelically grouped interactions. C. Comparison of parental and allelic intersections and true parental unique interactions for Maternal (top) and Paternal (bottom) contacts. D. Distributions of detected reads for replicated Broad interacting fragments plotted along a portion of Chromosome 17 (upper panel). The region surrounding the Airn locus (blue box) is expanded (lower panel). Arrows indicate direction of gene expression.
differences to the total paternal and maternal sets reveals that the 32 fragments shared by B6 and CAST alleles are, in fact, identical to the 32 3C fragments shared between the parental sets (Figure 3.13C). These data suggest that the set of interactions that are truly functionally associated with observed imprinted effects, and not influenced by allelic origin, is a rather small subset of the replicated parental interactions.

To visualize the results from the Airn 4C-Seq experiments, reads from each dataset were summed per 3C fragment and all replicated interactions were plotted along Chromosome 17 (Figure 3.13D). Likely owing to the limited size of the B6CASTF1 dataset, the entirety of the replicated interactions were located between the 5 and 22 Mb positions on the chromosome. Paternal and maternal interactions are highly similar at very proximal distances to the TSS of Airn, but some differences are noted at greater than 100 kb away from the viewpoint. Furthermore, though very preliminary at the time of this writing, some of the observed differences in detection frequency, as indicated by read numbers, may align with sharp differences in some histone modifications observed at this locus (personal communication of unpublished data).

Allelic 4C-Seq at TSS for Car2 and Bphl

For the Car2 and Bphl experiments, the problem was related to the uneven distribution of mapped reads between allelic viewpoints (Figure 3.5), rather than the number of mapped reads returned (Table 3.1). Therefore, since each allele is analyzed independently, analysis was performed under the assumption that location of enriched interactions may be detected, even if relative frequency could not be compared due to the unequal distribution of reads. Each allele from all replicates of both candidate genes was independently analyzed with fourSig to identify and prioritize windows representing significantly enriched interactions. The same fourSig conditions used for the analysis of Ibtk and Airn were again used for these experiments (see “Set Analysis and Results for Allelic 4C-Seq at Ibtk”). Analysis was again focused on cis interactions with the assumption that interactions mapping to Chromosome 3 for Car2 and Chromosome 13 for Bphl and lacking allelic
Figure 3.14 Set Analysis of Car2 and Bphl 4C-Seq Libraries. Allelic library intersections of enriched 3C fragments are depicted for (A,B) Car2 and (C,D) Bphl 4C-Seq replicates. E. Comparison of allelic proportions of replicated enriched fragments to allelic proportions of mapped reads is displayed.
information were most likely *cis* interactions.

Enriched windows for each allelic replicate were transposed to underlying 3C fragments and set analysis was performed to collect the set of interacting fragments replicated in all experiments (Figure 3.14A-D). A sharp difference in the number of replicated interactions between alleles for both candidate genes was immediately noticed. Comparison of the proportion of total replicated interactions made up by each allele to the allelic proportion of reads mapped to each viewpoint revealed that, for both loci, the proportions of enriched interactions mirrored closely the inequalities in reads mapped to each viewpoint (Figure 3.14E). Therefore, although the identification of enriched interactions is performed independently for each allele, it appears that it is still necessary to have similarly proportioned detection of viewpoints for each allele to ensure that the scope of interactions detected for one allele is not highly underrepresented relative to the other.

Despite the incongruence of allelic interactions and likely underrepresentation of CAST interactions for both genes, allelic analysis of the Broad replicated interactions was performed for the sake of thoroughness. For *Car2*, a little less than half of the interactions enriched on the CAST allele were found in common with B6, while almost all of the CAST interactions for *Bphl* were also detected on the B6 allele (Figure 3.15A & B, respectively). Visualization of the results was performed as described for *Ibtk* and *Airm*; however, it must be noted that the unequal representation of reads mapping to each allele makes relative comparison of interaction frequency by read number a highly inappropriate comparison. Interestingly, very few B6 interactions were plotted far from the viewpoint region for *Car2*, while the CAST allele is more widely distributed (Figure 3.15C). It is possible that the threshold used for significance is too high for distal interactions in the B6 *Car2*, which may be improved by using a masked analysis in *fourSig*. The positioning of the interactions for *Bphl* is more reminiscent of what was seen for *Ibtk* and *Airm*, however, and observation of a 6 Mb region centered around the viewpoint suggests a large number of B6 interactions amidst a scarcity of CAST interactions (Figure 3.15D).

An exploded view of the regions within a few hundred kb of the viewpoints was used to observe the potential presence of correlation with marks of transcriptional activity as was previously
Figure 3.15 Allelic Comparison of Broad Significant Contacts for Car2 and Bphl. A-B. The intersection of reproducible Broad cis interactions between B6 and CAST alleles is shown for (A) Car2 and (B) Bphl viewpoints. C-D. Distributions of detected reads for replicated Broad interacting fragments plotted for (C) Car2 and (D) Bphl (upper panel). The region surrounding the Bphl locus (blue box) is expanded (lower panel). Arrows indicate direction of gene expression.
Figure 3.16  Locus View of Interaction Profiles for Car2 and Bphl. A UCSC Genome Browser screenshot of 4C interaction data and selected chromatin data at the (A) Car2 and (B) Bphl loci is shown. The 3C fragments containing the TSS for the candidate genes are indicated by a dashed brown rectangle. For ChIP-seq tracks, green boxes indicate biallelic enrichment, blue boxes indicate CAST-specific enrichment, and gray boxes indicate insufficient SNP detection to make an allele-specific call.
described for *Ibtk* (Figures 3.16 and 3.9, respectively). For both *Car2* and *Bphl*, CAST reads in the area are limited to almost exclusively the fragments immediately adjacent to the viewpoints. A periodicity of detected B6 interactions is observed and, for both genes, seems to correspond well with enrichments of H3K27ac and H3K4me1 modifications. Also of note is the general absence of CAST enrichment for these modifications, which further strengthens potential association of these marks with the observed interactions. Although limited confidence should be placed in the analysis of these data, due to the likely underrepresentation of CAST interactions, these data would suggest a higher level of promiscuity of local interactions in association with the actively transcribed allele relative to the repressed.

**Concluding Remarks**

The results of the analysis of *Ibtk* definitely suggest that the *fourSig* method described in Chapter 2 is highly effective at producing very high resolution analysis of 4C-Seq data. Additionally, the inclusion of an algorithm to prioritize interactions to reflect their likelihood of replicate detection is an important improvement to existing methodologies in the field. Underrepresentation of interactions seen in the CAST alleles for *Bphl* and *Car2*, however, do underscore the requirement of a substantial amount of sequence data to gain sufficient enrichment calls. Furthermore, for performing allelic-specific analysis, it is clear, from the *Airn* example, that multiple replicates of similar size and, from the *Bphl* and *Car2* analyses, that equal proportions of reads mapped to allelic viewpoints are critical attributes for high confidence results.

The large amount of data lost to primer dimer in the large experiments is troubling (Figure 3.4). The entire scope of the project would have likely been better positioned to shed light on the proposed hypothesis of the potential for allelic interactions to be functionally associated with observed allelic biases in expression had so much of what was sequenced not been plagued by this problem. In retrospect, this problem may be alleviated by using primers containing only genomic sequence for a single amplification round followed by standard ligation of sequencing adapters from Illumina kits.
While this may increase the cost of the library preparation due to the purchase of proprietary materials, it is highly likely that the amount of data lost due to amplification inefficiencies would be greatly reduced. Furthermore, the amount of mappable sequence used for the \textit{Ibtk} analysis relative to the volume of sequence data returned from a single lane on the HiSeq 2000 platform may suggest that an even larger diversity of multiplexing than what was attempted here could be used with such a preparation process.

Limited replicates notwithstanding, the results of the analyses described here appear to suggest that, while the majority of detected interactions are likely random in origin, a substantial number of detectable contacts occur exclusively on one copy or the other. That these contacts are not restricted solely to very distal regions (Figure 3.8) does support the possibility that regulatory elements may engage in copy-specific interactions in the event of allelic bias in expression. Indeed, the identification of the putative enhancer downstream of \textit{Ibtk} (Figure 3.9) further underscores that this method, in combination with other analyses, can be effective at finding distal regulatory elements. Therefore, while definitive statement cannot be made regarding the proposed hypothesis that long-range interactions, if functionally important, should be highly associated with observed expression states, the data presented here certainly make the case that copy-specific resolution is important in the analysis of chromosome conformation.
CHAPTER 4: DISCUSSION OF RESULTS AND CONCLUSIONS

Introduction

The observation of allele-specific associations of long-range interactions described in Chapter 3 was made possible by the preliminary work on condition optimization and experimental design described in Chapter 2. The overarching goal was to test the hypothesis that long-range interactions may be functionally associated with observed expression effects as a general property of gene regulation. While not conclusive in its own right in making definitive conclusions regarding the proposed hypothesis, the results of the *Ibtk* experiment do support the expected prediction that allelically biased genes may harbor interactions that are similarly specific to one allele over the other, at least in the single example. Lessons learned and possible paths forward are discussed in this chapter.

Analysis and Validation Methods

Currently, only a few algorithms exist to analyze 4C-Seq data. As a part of the work presented here, a new method for analyzing 4C-Seq data, *fourSig*, was developed. *fourSig* is similar to other recently published methods, such as *r3Cseq* (Thongjuea et al. 2013) and *4Cseqpipe* (van de Werken et al. 2012), in that it uses the full range of quantitative information derived from read depth to determine significance thresholds for 4C-Seq data. Previous 4C-Seq analyses have typically discarded this information due to concerns over possible PCR amplification bias during library preparation (Splinter et al. 2011; Holwerda et al. 2013). Instead, these studies employ a transformation of the read counts to Boolean values, leading to 3C fragments being ruled simply as detected or not detected. This binarization of the data necessitates the use of window sizes of at least 100 fragments, which generates a lower resolution analysis of the data. As a result these
methods would likely not be able to detect fine differences in interactions profiles, such as what is seen with the putative enhancer at the *Ibtk* locus (Figure 3.9). The incorporation of read counts into the threshold calculation allows the use of much smaller windows and improves the resolution available for 4C-Seq experiments. Therefore, higher resolution analysis methods such as *fourSig*, *r3Cseq*, and *4Cseqpipe* represent a substantial improvement over previous limitations.

Significance threshold can be determined in two different ways when using permutated data. All permutations can be performed up front and the average cumulative dataset can then be used to identify a single cutoff for the FDR. This strategy was initially used to determine cutoffs from 4C data at various loci on the inactive X chromosome (Splinter et al. 2011) and served as a starting point for analysis of 4C experiments in the Magnuson Lab. Tests of this strategy on preliminary datasets resulted in widely variable thresholds amongst separate analyses on the same datasets (data not shown). This was also observed in attempts to run the analysis from Splinter et al (2011) on the datasets described in that study. Therefore, a second strategy, involving the calculation of a threshold for the desired FDR after each permutation and selecting the final threshold from the distribution of cutoffs, became the focus of the *fourSig* method. This method was found to be more robust to extreme values that may occur with rare frequency amongst the permutations and resulted in more consistent threshold calculations when tested on preliminary datasets.

One notable feature of the *fourSig* method is the prioritization algorithm (Illustration 2.3). However, while the interaction prioritization feature may offer predictive value in terms of the likelihood that enriched interactions may be reproduced, it must be noted that interactions classified in all 3 categories were, in fact, consistently reproduced. The analysis of the intersection of replicates for *Ibtk* showed that more than 5% of the consistently detected 3C fragments were found within interactions designated as Narrow or Intermediate (Figure 3.7). This could occur, for example, in the event that the dominant 3C fragment in a window is very large. Such a scenario would leave little linear distance to either side of the primary point of contact to result in the capture of neighboring fragments. That said, due to the frequency of fragmentation that occurs when using a restriction enzyme that recognizes a 6 base motif, such as HindIII, the occurrence of such an event is largely
unlikely owing to the requirement of the rare very large fragment being involved in a specific interaction with the selected viewpoint.

To validate the interactions determined by fourSig to be enriched, FISH probes were designed to measure the relative proximities of selected allelically interacting loci for the Ibtk viewpoint. Measurements between FISH signals have frequently been used in the literature as first-pin validation methods for chromosome conformation studies (Simonis et al. 2006; Simonis, Kooren, and de Laat 2007; Lieberman-Aiden et al. 2009; Splinter et al. 2011). However, one shortcoming of this strategy is that resolution of standard microscopy methods is typically limited to around 0.2 µm. Since DNA in the nucleus exists as compacted chromatin, distances of less than 3 Mb do not always lead to large physical distances whose separations are obvious. Therefore, the limitation of resolution for FISH assays can make differential measurements between copies difficult due to the natural proximity of loci within this range. Therefore, the ability to consistently validate allele-specific interactions at this range by FISH is very significant and further speaks to the precision with which peaks can be called by fourSig.

Difference and Similarity of Contact Profile Between Alleles

Allele-specific bias in expression implies the existence of differential regulatory schemes between copies. The primary culprit behind such a scheme would seem most likely to be some sort of allelic variation in sequence at the binding site of a regulatory factor responsible for facilitating the appropriate expression program. Two of the candidate genes tested here, Ibtk and Car2, were previously studied with respect to the allelic expression bias in the Magnuson Lab by Dr. Joshua Starmer (unpublished work). As a part of Dr. Starmer’s analysis, these genes, along with several others, were examined in their promoter proximal regions for evidence of SNPs and genomic rearrangements between the B6 and CAST alleles to determine the regulatory source of the allelic expression effects. In almost every case, Dr. Starmer was unable to identify sequence differences that may account for the observed differences in expression. Interest in the possibility for influence
affected from distal loci was, in part, born out of the desire to further understand what mechanisms may be involved in exerting differential regulation of expression between copies.

While the results presented here, most notably for \textit{Ibtk}, do not prove that functional regulatory elements lie within the described interactions, the consistency at which allele-specific interactions were detected supports the notion that functional association of conformation with differential expression state is a possible source of regulatory control. The results for the \textit{Ibtk} locus indicate that the majority of enriched interactions detected within 1 Mb of the viewpoint are shared between alleles (Figure 3.8). This is consistent with numerous 4C studies that report substantially increased contacts with loci at shorter linear distances from the viewpoint (Splinter, Grosveld, and de Laat 2004; Simonis et al. 2006; Zhao et al. 2006; Holwerda et al. 2013). However, interactions that were found to be different between alleles were also mostly found no further than 3 Mb from the viewpoint, with only a very small portion of these interacting fragments occurring at very large distances from the TSS of \textit{Ibtk}.

Similarities and differences amongst allele groupings may prove to play a meaningful role in using this type of analysis to parse putative distal regulatory elements related to imprinted mechanisms. The comparison of groupings as allelic versus parental was able to narrow detected interactions down to a very small set when accounting for the potential of allelic influence over interactions found to be in common between parental origins (Figure 3.13). It is unfortunate that so much data were lost from this set, as well as from others, as a more thorough analysis of the variations in similarity and differences amongst allelic interactions may have provided insights into the prevalence of copy-specific interaction profiles as a general phenomenon.

\textbf{Future Directions}

This work may lay ground work for further development and experimentation to be performed on understanding the allele-specific variations in chromosome conformation. The experimental conditions and analysis methods described here, while optimized and reproducible, have several
shortcomings that may be improved with further development. Additional experiments may be helpful in improving the understanding of the functional properties, if any, of long-range interactions on observed expression effects.

Possible Improvements

There are several aspects of the work presented here that could potentially be improved with further experimentation and optimization. Chief amongst these is the volume of data loss that occurred in the filtration and mapping of the sequenced reads. As described in Chapter 3, the overwhelming majority of this loss was due to the presence of undesirable primer dimer contamination in the sequenced libraries (Figure 3.4). Being that amplified libraries are size selected and gel purified prior to submission for sequencing, it is not clear how a primer dimer product that would seem to be smaller than the lower limit of the size selection (approximately 200 bp, Figure 2.8) would have come to represent a substantial portion of the sequenced library. Nonetheless, the presence of reverse complement to the opposing sequence adapter as the major contributor to unaligned sequence is clear. Therefore, the design future 4C-Seq experiments would be wise to exclude the adapted primer strategy described here (See Chapter 2). While admirable in its attempt to reduce cost associated with proprietary adapter kits from Illumina, this step appears to have been directly responsible for a large amount of data loss in multiple genes. Such a deviation would also likely further remove concern of amplification bias because only one amplification step would be involved and barcodes can still be applied to the genomic primers.

Additionally, the need for more refined downstream analysis methods will always be present. Lacking from the work presented here is an efficient and streamlined strategy for performing meta-analyses with other genome wide data to loci highlighted by multiple properties indicating a higher likelihood of probably functional activity. The correlations made with existing ChIP-seq data here are highly subjective as they are the result of visual inspection of track files on the UCSC Genome Browser (Meyer et al. 2013). An attempt at automated correlation of interacting loci with epigenetic modifications was made in a separate and concurrent 4C-Seq project in the Magnuson Lab focusing on interaction patterns related to X Chromosome Inactivation (Mugford et al.). The assayed loci in
this project suffered some of the data loss problems seen here, but to a much less extent. Therefore, a more thorough analysis with several loci was possible; however, correlation with specific marks in a manner that would indicate clear associations with expected mechanisms was limited. While it is possible that clear association of modifications and interactions in these experiments is simply not present in a manner to indicate putative functionality, it is also possible that further refinement and sophistication of analysis methods may improve the scope of conclusions that can be made by such comparisons.

Finally, another comparison that is limited in the work presented here is comparison of relative interaction frequency between alleles. Much of the comparison of allelic associations of observed interactions focuses on exclusivity of interactions to one allele or another. However, transcription and gene expression are rarely all or nothing events. Recent evidence points to the likelihood that low level transcription of most genes may occur throughout the genome at loci that are not completely silenced (Adelman and Lis 2012). Similarly, none of the candidate genes described here displays complete exclusivity of expression to one allele or the other. In fact, low level detection of the repressed allele was a requirement of the candidate gene screen to ensure that observed bias was not a function of detection bias due to unforeseen technical error. Therefore, if any interactions truly have functional involvement in the observed expression patterns, it is very likely that interactions may be observable on both alleles but in a biased manner reflective of the observed expression bias. As such, an analysis focused on exclusivity of detection, as is primarily performed here, would be blind to relative differences in interaction frequency between alleles. Indeed, the putative enhancer for *Ibtk* referred to in Figure 3.9 is detected on both alleles, but contains a very strong read count bias in favor of the expressed (CAST) allele. An attempt to determine significance of bias was made involving an array of t-tests for allelic interactions at each 3C fragment (not shown), however the triplicate data for *Ibtk* proved too limited in scope to produce strong results. It is possible that improving the data volume and replicate numbers and adapting more sophisticated statistical methods, such as the edgeR suite (Robinson, McCarthy, and Smyth 2010) may prove more useful in detecting relative differences between allelic datasets.
Further Experimentation

The work presented here was designed to test the hypothesis that long-range interactions between loci are functionally associated with gene expression as a general property of regulation. Since the work is primarily observational, further inquisition may prove useful in determining the nature of such associations, especially within the context of preferential association between copies. Therefore, if such an investigation were to be pursued, three specific projects would be advisable.

First, the lack of depth in replicates due to the primer dimer problem compromises the ability of these data to provide information on the degree to which such associations might be a general property of regulation. It is necessary to study several candidates exhibiting the desired expression properties; therefore the work here should be repeated, and possibly expanded, using standard ligation of sequencing adapters according to the manufacturer’s instructions.

Secondly, this analysis has focused only on genes in which clear allelic biases in expression are consistently observable. As a comparison, allele-specific interaction profiles should be obtained using 4C-Seq on a similar size of candidate genes that exhibit equal distribution of expression from both alleles. The nature of chromosome conformation dictates that loci close to each other along the linear arrangement of the chromosome should have a higher contact probability relative to more distal loci due to natural physical properties of polymers in confined spaces (Richter, Nessling, and Lichter 2008; Mirny 2011). Therefore, one would expect that if long-range interactions are functionally associated with gene expression, genes whose expression patterns are similar between alleles should exhibit a higher proportion of similarity in allelic interaction profiles than genes whose allelic copies exhibit an imbalance of expression product. Furthermore, observation of relative association, or absence thereof, of interactions with other genomic features and modifications may be able to identify differing trends between interactions of unbiased genes compared to biased genes which may help to illuminate properties associated with functionally involved interactions.

Finally, perturbation experiments may be necessary to truly demonstrate functional involvement of an interaction with the observed expression effects. Care must be taken in the design of such experiments. For starters, targeted mutagenesis experiments are time consuming and costly.
Therefore, it would be prudent to ensure that substantial analysis has gone into the identification of specific interactions thought to be very likely to have functional involvement with the observed expression effect. Additionally, while the obvious experiment would be to delete the interacting locus and observe effects on expression or association between loci, such a strategy may be a mistake due to oversimplification. If one concedes that a long-range interaction is functionally involved in driving an expression state, it must be considered that complete removal of a region may have unintended and indirect consequences that may erroneously be interpreted as loss of expression or repression due to deletion of an enhancer or silencing element, respectively. Recent experiments have provided strong and convincing evidence that interphase chromosomes are packaged into a fractal globule configuration (Lieberman-Aiden et al. 2009; Mirny 2011) which may lead to the observation that loci tend to interact primarily within topologically associated domains (Dixon et al. 2012; Nora et al. 2012). Therefore, the removal of a chunk of DNA from a chromosome could, theoretically, change the physical constraints with which such partitioning occurs. Such an effect may lead to the erroneous conclusion that a regulatory element occurs within the deleted region when the observed affect may be due to a more comprehensive change in the available loci within the local environment. A correctly designed experiment, rather, should aim to replace the loci targeted for ablation with a similarly sized sequence not identified to interact with the viewpoint region.

**Concluding Remarks**

In addition to an added layer of complexity in the regulation of gene expression, diverse roles of the potential functionality of chromatin conformation provide a context of involvement for many genomic aberrations thought to be associated with many disease states. The unintentional introduction of SNPs, genomic rearrangements, and copy number variations could all potentially lead to changes in how chromatin is able to fold and arrange itself. The likelihood that disease states associated with such aberrations are interdependent upon a variety of complex functions have led to the suggestion that chromatin conformation states, either by virtue of being the result and/or cause of inappropriate behavior, could potentially serve as biomarkers indicative of disease phenotypes.
(Crutchley et al. 2010). Additionally, although speculative, improvements in medical technology and intracellular manipulation could lead to the possibility that, if properly understood, chromatin conformation states may one day be able to serve as therapeutic targets as well. Taken together, further investigation into how higher-order chromatin conformation is associated with genetic regulation is an important area of molecular biology.

Results from these experiments demonstrate that the profile of observable interactions may vary between copies. This is important because many genomics assays do not typically detect resolution to discriminate between alleles. The observations made here help to demonstrate that assumption of behavior and modification of loci with genome-wide assays may introduce errors related to inability to distinguish whether events are occurring on both alleles in similar proportions or even at all. Furthermore, a conceptually simple and relatively easy to use analytical program was developed and is now available for use by the scientific community. This is significant as interest in chromosome conformation is increasing and few tools exist to interpret data from these experiments. While the results of the 4C-Seq data alone are not particularly profound in terms of the biology of the candidate genes selected, the precedent for performing allele-specific analysis is established.
CHAPTER 5: DETAILED METHODS

Introduction

A substantial portion of the work described here was focused on the development and careful optimization of various techniques. As such, description of precisely how the procedures were performed is necessary. In this chapter, the procedures established are either described in detail or reference to where more precise details can be found is made.

Tissue Culture and Sample Preparation

TS cells were cultured as previously described (Quinn, Kunath, and Rossant 2006; Calabrese et al. 2012). TS cells were removed from feeder cells and grown independently for 2 passages on gelatin-coated plates prior to sample collection.

4C-Seq Protocol

The initial foundation for the 4C-Seq protocol was adapted from a collection of published methods. The generation of 3C libraries was based primarily on detailed descriptions of the identical portions of the Hi-C protocol established by the Job Dekker lab at the University of Massachusetts Medical School (Lieberman-Aiden et al. 2009; van Berkum et al. 2010). The remainder of the 4C protocol was adapted from descriptions of the 4C methods developed by the Wouter de Laat lab, currently at the Hubrecht Institute (Simonis et al. 2006; Splinter et al. 2011). Experimental conditions were specifically optimized for ideal use in the TS cell model and amplification and multiplexing strategies were independently designed in the Magnuson Lab (see Chapter 2). TS cells are prepared
in batches of 10 cm² dishes and collected once they reach ideal confluence at NFx2. Each 3C library is developed using approximately 30 million cells for starting material, which provides enough DNA to produce approximately 6-8 4C libraries. Helpful tips and suggestions learned from trial and error are described.

Isolation of TS Cells

1. Prior to sample collection, prepare fresh 2.5M glycine and filter it through a 0.22 µm syringe filter.

2. Use 2 mL of 0.25% Trypsin per 10 cm² plate of NFx2 TS cells. Trypsinization needs to occur quickly and dispersion to near single cell suspension is ideal.

   TIP: Allow the cells to sit in 1X PBS at room temperature for approximately 3 minutes prior to adding trypsin. This facilitates quicker and more thorough dispersion. Stagger the trypsinization in groups of no more than 3 plates at a time to avoid falling behind on incubation times. Additionally, use a pasteur pipette to triturate the cells against the surface of the plate.

3. To quench the trypsin, transfer the cells from all plates to a tube containing 6 mL of RPMI with 10% FBS. Use 3 mL of RPMI with 10% FBS to collectively rinse all 3 plates and transfer this rinse to the same tube as the previous cells. Keep the tube of dispersed cells on ice until all plates have been collected.

   TIP: Disperse the cells into the conical tube by pressing the pasteur pipette flat against the bottom of the tube.

4. The final collection should contain no more than 15 mL per tube. Centrifuge the cells at step 4 on a clinical centrifuge for approximately 4 minutes. Remove the supernatant and resuspend the cells in 2 mL of RPMI with 10% FBS and place on ice. Once all plates have been trypsinized and collected, consolidate all 2 mL aliquots into a single tube. Pellet the consolidated suspension and resuspend in 11.2 mL of RPMI with 10% FBS. The final volume should be approximately 11.25 mL.
5. Remove 50 µL from the suspension and dilute it 1:10. This will be used for counting the cells and verifying that the suspension contains mostly single cells.

6. Add 312 microliters (µL) of 37% formaldehyde for a final concentration of 1%. Fix the cells for 10 minutes at room temperature (RT) on a nutator. Count the cells using a hemocytometer during this incubation step.

7. Quench the fixation with 625 µL of the freshly made 2.5 molar (M) glycine for a final concentration of 129 millimolar (mM). Incubate at RT for 5 minutes on the nutator, then on ice for 15 minutes.

8. Pellet the cells at 480 relative centrifugal force (RCF) for 10 minutes and remove the supernatant. Resuspend the cells in sterile filtered 1X PBS for a concentration of 30 million cells per mL and distribute in aliquots of the desired cell number, preferably 30 million cells per aliquot. Pellet this suspension at 480 RCF for 5 minutes at 4 °C and remove the supernatant. Cells should be flash frozen with LN₂ and stored at -80 °C.

**Digestion of Chromatin**

1. Prepare 1.4 mL of Cell Lysis Buffer (Below) and 7 mL of 1X NEB Buffer 2. Pre-chill the lysis buffer and 1.7 mL of Buffer 2 on ice for at least 10 minutes. The remainder of Buffer 2 can remain at RT. Additionally, pre-cool a 2 mL dounce homogenizer with pestle B on ice as well. During the same time, thaw a 30 million cell pellet on ice.

2. Once the pellet is thawed, resuspend it in 660 µL of the chilled lysis buffer. Transfer this suspension to the dounce homogenizer and incubate on ice for 15 minutes.

3. Taking care to not allow foaming, lyse the cells by quickly douncing 10 times with pestle B. Let the suspension cool on ice for 1 minute, then dounce 10 more times. Move the suspension to a clean 2 mL microcentrifuge tube. Use an additional 660 µL of lysis buffer to rinse the homogenizer and pestle; then add this to the 2 mL tube. Pellet the nuclear fraction at 5000 revolutions per minute (RPM) in a microcentrifuge for 5 minutes at RT.
4. Remove the supernatant and wash the cells twice with 500 µL ice-cold 1X NEB Buffer 2. Pellet the nuclear fraction at 5000 RPM for 5 minutes between each wash. Ensure that the pellet is completely resuspended after each step.

5. During the wash spins, add 100 µL of 10% SDS to 4.25 mL of 1X NEB Buffer 2. Distribute 348 µL of this mixture to 12-1.5 mL microcentrifuge tubes and keep them at RT.

6. After the second wash, completely resuspend the pellet in 600 µL ice-cold 1X NEB Buffer 2. Split this sample amongst the 12 microcentrifuge tubes. 52 µL should result in 2.5 million cell aliquots in each tube. Use the pipette to gently disperse the solution to avoid foaming. These solutions should be at a final concentration of 0.2% SDS and a volume of approximately 400 µL.

7. Incubate the tubes at 37 °C for 30 minutes on a rocker, then transfer them to a 65 °C water bath and incubate for exactly 10 minutes. During this incubation, agitate the tubes by hand every 2-3 minutes. Immediately place the tubes on ice upon completion of the incubation. After about 1-2 minutes, tap spin the tubes at 4 °C to collect any condensation.

   TIP: Parafilm the microcentrifuge tubes to guard against sample loss and contamination.

8. Add 44.44 µL of 20% Triton X-100 to quench the SDS. This should lead to a final concentration of approximately 2% Triton X-100. Mix the solutions well by triturating with the pipette but be careful to avoid foaming. Incubate the solutions at RT for at least 10 minutes.

9. During the SDS quenching, prepare the Restriction Digest Master Mix (below). Do NOT use BSA from NEB. Make a 10 mg/mL solution of BSA in water and sterilize it with a syringe filter.

10. After quenching the SDS, distribute 51.56 µL of the Restriction Digest Master Mix to each tube. Add 400 U of HindIII. Mix well by trituration with a pipette but take care to avoid foaming. Incubate the digestions overnight at 37 °C in a vortex turning at 950 RPM.
**Proximity Ligation**

1. Inactivate the restriction enzyme by adding 39.66 µL of 20% SDS to each sample for a final concentration of 1.47% SDS. Mix well without foaming.

2. Place the tubes in a 65 °C water bath for exactly 30 minutes and immediately place them on ice upon completion of the incubation period. During the incubation period, prepare the Ligation Master Mix and store it on ice. BSA from NEB may be used in this solution. Place 6-15 mL conical tubes on ice to chill. Label 5 of them as “3C” and the sixth as “No Ligase”.

3. Pool all digested chromatin samples into the Ligation Master Mix and mix well by gentle inversion. Distribute 8.8 mL of the ligation solution to the “3C” conical tubes. Place the remainder of the solution in the “No Ligase” tube. Add 2.5 µL of 5U/µL T4 DNA Ligase (Invitrogen) to each “3C” tube. Do NOT add ligase to the “No Ligase” tube. Mix well by gentle inversion and incubate all tubes at 16 °C for 4 hours.

4. After the ligation incubation, add 50 µL of 10 mg/mL proteinase K to all tubes. Parafilm each tube and incubate tubes overnight at 65 °C.

**DNA Purification**

This portion of the protocol uses Phase Lock conical tubes to help improve yield from the large extractions. If Phase Lock tubes are not used, it is highly recommended to perform back extractions of the remaining organic layers. A significant amount of sample will be lost near the interphase of the extraction, so dilution of this with extra aqueous solution and subsequent extraction is necessary to improve yield.

1. First thing in the morning, add 50 µL of 10 mg/mL proteinase K to each tube and incubate at 65 °C for an additional 2 hours.

2. After at least 2 hours, cool the tubes to room temperature and move all samples into 50 mL Phase Lock conical tubes. Use 6 mL ddH₂O to successively rinse each “3C” tube and distribute
1.2 mL of this to each of the Phase Lock tubes. Similarly, rinse the “No Ligase” tube with 6 mL of ddH₂O and transfer to the corresponding Phase Lock tube. Add 10 mL of phenol (pH 8.0) to each Phase Lock tube and mix well by inversion. Do NOT vortex Phase Lock tubes.

3. Parafilm each tube and separate the phases by spinning at 480 RCF for 10 minutes at RT. Carefully transfer as much of the aqueous phase as possible to a new 50 mL Phase Lock conical tube.

TIP: Approximately 30-50 µL of aqueous phase will remain on the sides of the tubes after pouring off. A quick spin can be used to collect it so that it can be transferred by pipette.

4. Perform a second extraction by adding 10 mL phenol:chloroform:isoamyl alcohol (1:1:24, PC:IA) to each sample and mix well by inversion. Do NOT vortex Phase Lock tubes. Parafilm each tube and separate the phases by spinning at 480 RCF for 10 minutes at RT.

5. Transfer aqueous phases to new 50 mL conical tubes. Bring all volumes to 13 mL by adding 3 mL of ddH₂O. Precipitate samples by adding 1.3 mL of 3M NaOAC, pH 5.2 and 32.5 mL of ice-cold 100% ethanol to each tube. Place tubes at -80 °C for at least 1 hour or at -20 °C overnight.

6. If tubes are frozen, allow them to thaw but do not bring up to RT. Spin the tubes at 10,000 RCF for 20 minutes at 4 °C. Discard the supernatant and dissolve pellets in 450 µL of ddH₂O and transfer each to a new 1.5 mL microcentrifuge tube. Add 500 µL PC:IA to each tube, vortex and spin at top speed for 5 minutes at room temperature. Transfer the aqueous phase to a new tube and repeat the PC:IA extraction.

7. Transfer the aqueous phases to new 2 mL microcentrifuge tubes. Add 50 µL of 3M NaOAc, pH 5.2, and 1.25 mL of ice-cold ethanol. Mix the tubes by inversion and place tubes at -80 °C for 1 hour or at -20 °C overnight.

8. For each PC:IA extraction, pool the organic phase into a 15 mL conical tube. Use 1 mL ddH₂O to rinse each microcentrifuge tube and transfer it to the pooled organic phases. Add an extra 1 mL
of ddH$_2$O to the pooled organic phases and extract the aqueous phase. Place aqueous phases into a new 15mL tube and repeat the back extraction in the pooled organic phases with 2 mL ddH$_2$O. Pool all aqueous phases from each set of back extraction, which should be approximately 4 mL, and precipitate with 400 µL 3M NaOAc, pH 5.2 and 10 mL 100 % ethanol. Mix the tubes by inversion and place tubes at -80 °C for 1 hour or at -20 °C overnight.

9. If tubes are frozen, allow them to thaw but do not bring to RT. Spin microcentrifuge tubes at top speed for 30 minutes at 4 °C. Spin back extraction tubes in a large centrifuge at 10,000 RCF for 20 minutes at 4 °C. Rinse the pellets with ice-cold 70% ethanol and vortex. Pool back extracted pellets into a microcentrifuge tube and spin samples at top speed for 10 minutes at RT. Repeat the 70% ethanol wash until the pellet sizes stop decreasing. This will usually be at least 3 total washes.

10. Remove the supernatants and allow the pellets to air dry. Be careful not to over dry the pellet. Dissolve each pellet in 50 µL TE and combine all 3C samples in to one tube. Rinse the 3C tubes successively with 50 µL of TE to remove any remaining 3C sample and add this rinse to the pool. This volume should be approximately 300 µL for the 3C sample. Add 10 µL and 60 µL of 10 mg/mL RNase A (Sigma-Aldrich) to the No Ligase and 3C tubes, respectively. Incubate at 37 °C for 1-2 hours.

11. To remove the RNase A, add 340 µL and 40 µL of ddH$_2$O to the No Ligase and 3C tubes, respectively. Extract the DNA by addition of 400 µL PC:IA to each tube. Vortex and spin at top speed for 5 minutes. Move the aqueous phases to new 1.5 mL microcentrifuge tubes and ethanol precipitate by adding 40 µL of 3M NaOAc, pH 5.2 and 1 mL of ice-cold ethanol.

12. Perform 2 back extractions of the remaining organic phases from the most recent extraction. Vortex, spin and precipitate this sample as described for the primary extractions in step 11. Place all tubes at -80 °C for at least 1 hour or at -20 °C overnight.
13. If frozen, allow the tubes to thaw but do not bring them to RT. Spin all tubes at top speed for 30 minutes at 4 °C. Remove the supernatants and combine the pellets from the primary 3C and back-extracted pellets. Wash all pellets with ice-cold 70% ethanol. Vortex and spin tubes at top speed for 10 minutes at RT.

14. Resuspend the 3C, No Ligase, and back extraction pellets in 50 µL of TE and place at 37 °C for 20 minutes. After 20 minutes, combine the first and second back extractions (100 µL total) and add 15 µL TE to the empty tube previously containing the second back extraction. Incubate for another 20 minutes and pool the back extractions with the 3C pellet (150 µL total) and add 15 µL TE to the empty tube previously containing the first back extraction. Incubate for 20 more minutes and consolidate the back extraction rinse with the 3C libraries.

15. Determine the quantity of recovered DNA using the Quant-iT assay for the Qubit fluorometer (Invitrogen).

**4C Restriction Digestion**

Prior to proceeding with 4C library generation, the quality of the 3C libraries must be verified using the gel migration assessments and amplification of the expected ligation control for the Gapdh locus (see Chapter 2). Going further, 4C assays will be variable for each viewpoint. Below, DpnII and DraI are used as an example in the protocols described.

1. Setup a 125 µL digestion reaction using 12.5 µg of 3C DNA by adding 12.5 µL 10X NEB DpnII Buffer, 1.25 10 mg/mL BSA (NEB), and 25 U DpnII. Incubate the reaction at 37 °C for at least 4 hours.

2. After digestion, inactivate DpnII by incubation at 65 °C for 30 minutes. Remove 25 µL (approximately 25 µg) to keep as a 4C No Ligase control.

TIP: If 80 °C incubation is needed to inactivate the selected restriction enzyme, bring the solution to 10 mM EDTA to reduce denaturation of the DNA fragments. For the volume described here, 2.55 µL of 0.5M EDTA, pH8.0 should be appropriate.
3. Perform an assessment of the digestion efficiency by comparing gel migration of the undigested 3C library to 500 ng of the 4C No Ligase control.

4C Proximity Ligation and Linearization

1. For each 4C reaction, prepare a 4 mL proximity ligation reaction from 400 µL 10X T4 DNA Ligase Buffer (Invitrogen), 100 µL digested 3C library (approximately 10 µg), 3498.75 µL ddH2O, and 1.25 µL T4 DNA Ligase (Invitrogen). This mixture should be approximately 2.5 ng/µL DNA to facilitate circular ligations. Mix the reaction well with a P1000 pipetman and incubate the reactions in a 16 °C water bath for 4 hours.

2. To each ligation reaction, add 4 mL of PC:IA and vortex for 2 minutes. Separate the phases by centrifugation at 1500 RCF for 10 minutes at RT. Carefully transfer as much aqueous phase as possible to a new 50 mL conical tube (not Phase Lock).

3. Perform 2 back extractions of the organic phase with 4 mL ddH2O each. Consolidate the aqueous phases into the 50 mL conical tube holding the initial aqueous phase.

   TIP: This step achieves a substantial improvement in yield. If this step is omitted a large amount of sample loss should be expected.

4. Precipitate 4C libraries by adding 1.2 mL of 3M NaOAc, pH5.2 and 30 mL of ice-cold 100% ethanol to each tube. Place the tubes at -80 °C for 1 hour or -20 °C overnight.

5. If frozen, allow the mixtures to thaw. Spin the tubes at 10,000 RCF for 20 minutes at 4 °C. Carefully pour off the supernatant and rinse the pellet, which is likely fragmented, with 10 mL of ice-cold 70% ethanol. Vortex for 30 seconds, then spin the tubes again at 10,000RCF for 15 minutes at 4 °C. Pour off the supernatant and air-dry the pellet.

6. Resuspend the pellet in 200 µL ddH2O. This product is the circularized intermediate. When ready, prepare a linearization reaction by adding 30 µL 10X NEB buffer 4, 69 µL ddH2O, and 1 µL Dral (NEB). Incubate the reactions for at least 4 hours at 37 °C.
TIP: Modify the conditions as necessary for desired restriction enzymes. Be sure to check the manufacturer’s indications regarding star activity.

7. If possible, heat-inactivate the restriction enzyme. Otherwise, perform one PC:IA extraction with back extractions as described above. Either way, purify the DNA by ethanol precipitation to remove the reaction solutions. Determine the concentration using a Qubit® Fluorometer (Invitrogen) and adjust an aliquot to either 100 ng/µL or 50 ng/µL. Ideally, more than 1.5 µg of 4C template is necessary to ensure sufficient quantity for downstream reactions.

Library Amplification

If adapted primers are used, two steps of amplifications must be optimized. The following instructions assume the use of adapted primers as described in Chapter 2.

1. For each 4C assay, determine the optimal amplification conditions (annealing temperature and MgSO₄ concentration) for the genomic-adapted primers using the Pfx amplification system from Invitrogen. Use 35 cycles of 20 second denaturation, 20 second annealing, and 30 second 68 °C primer extension steps.

2. Once the ideal conditions are found, perform a test with a gradient of increasing 4C template. There should be PCR product ranging up to 2 kb, but two prominent bands may be visible between 300-500 bp. Be sure to use the 4C No Ligase, source 3C library, and 3C No Ligase samples as controls to test for inappropriate amplification.

3. Generate the 4C intermediate product by performing 20 cycle amplifications with the previously determined conditions. Prepare at least 5 reactions using 200 ng 4C template each.

4. Pool the replicate reactions from each 4C primer pair and performed an ethanol precipitation by adding 12.5 µL 3M NaOAc, pH 5.2, and 312.5 µL ice-cold 100% ethanol. Place the precipitations at -80 °C for at least 1 hour or at -20 °C overnight.
5. Pellet and wash the samples with 70% ethanol. Resuspend the pellet in 25 µL of TE and load the entire sample on a 2% low-range agarose gel (Bio-Rad). Purify the product that runs from 150-650 bp, but be careful not to capture the primer band. Perform the gel extraction using twice the manufacturer recommended volumes (6 volumes QG and 2 volumes isopropanol for the Qiagen Gel Purification Kit). Elute the columns with 2 rounds of 25 µL ddH₂O.

TIP: Spin the column several times to insure complete removal of ethanol. Do not trust that the number of spins in the manufacturer’s instructions is sufficient.

6. Using the SsoFast qPCR system (Bio-Rad), determine the linear range of amplification for the outside adapter primers on the prepared 4C intermediate products. Use 35 cycles of 20 second denaturation, 20 second 65 °C annealing, and 30 second 72 °C primer extension steps.

7. Prepare the final amplification reactions using the Phusion Polymerase system from NEB. Use the optimal number of cycles determined in the previous step for 20 second denaturation, 30 second 65 °C annealing, and 30 second 72 °C primer extension steps. Perform at least 3-50 µL PCR reactions and combine them after amplification.

8. Clean the pooled PCR product with an equal volume of AMPure XP beads (Beckman Coulter). Mix the solutions well and place them on a rotator for 5 minutes at RT. Reclaim the beads with a DynaMag (Life Technologies) and remove the supernatant.

9. Without removing the tubes from the magnet, add 250 µL freshly prepared 70% ethanol and let the beads incubate for 1 minute at RT. Remove the ethanol and repeat the wash. Pellet the beads at 1,000 RPM for 10 seconds and place the tube back on the magnet. Remove any residual ethanol with a P10 pipetman.

10. Dry the pellet at 37 °C for 2 minutes, then elute the sample by adding 26 µL TE and mix the beads well. Let the elution stand at RT for 5 minutes. Reclaim the beads and transfer the solution to a new tube. Test the size of the final 4C library by running 2 µL on a 2% agarose gel. The observed library will likely be very faint.
11. Determine the concentration using a Qubit® Fluorometer. Estimate the molarity of the solution by assuming 660 kiloDaltons per bp molecular weight. Estimate the average size of the library based on the gel image. Adjust an aliquot to 15 nM. This preparation is ready for sequencing using the Illumina HiSeq and MiSeq platforms.

4C-Seq Data

The raw data from the 4C-Seq experiments can be accessed from the GEO database under the accession number GSE50907. Processed data files are on file in the Magnuson Lab.

Recipes for Solutions

The following recipes are for solutions used in the 4C-Seq protocol described in this section. They should each be made fresh on the day of use.

Cell Lysis Buffer

This recipe is for lysis of cytoplasmic compartments to isolate insoluble nuclear fractions.

- 10 mM Tris-HCl, pH 8.0
- 10 mM NaCl
- 0.2% Igepal CA-630
- 1:10 Sigma Protease inhibitors.

Restriction Digest Master Mix

This recipe prepares enough for 13 digestion reactions. For a single reaction, divide all values by 13.

- 65 µL 10 mg/mL BSA (from scratch)
- 140.4 µL 10X NEB Buffer 2
- 464.88 µL ddH₂O

**Ligation Master Mix**

This recipe prepares enough for 6 ligation reactions. For a single reaction, divide all values by 6.

- 7.2 mL 20% Triton X-100
- 4.8 mL 10X NEB Ligation Buffer
- 480 µL 10 mg/mL BSA (NEB)
- 29.034 mL ddH₂O

**FISH Protocols**

The protocol describe here has been modified from a protocol that has circulated through the Magnuson Lab. The original protocol was provided courtesy of Susanna Miynarczyk-Evans of the Barbara Panning Lab. Modifications reflect the result of optimizations performed for performing RNA/DNA FISH in TS cells. Incubations are typically performed in a 6 well tissue culture plate, unless indicated otherwise.

**Labeling of FISH Probes**

FISH probes are labeled using reagents from the BioPrime™ Labeling Kit (Invitrogen).

1. Dilute 50-100 ng DNA to 19 µL with TE in a 1.5 mL microcentrifuge tube. Place on ice and add 20 µL 2.5X Random Primers Solution.

TIP: It may be a good idea to wrap the lids with parafilm. If you do this, however, the parafilm will be sticky when trying to remove.
2. Place mixture at 100 °C for 5 minutes. Immediately upon completion, place mixture on ice to cool. Allow mixture to cool for at least 1 minute.

3. A dNTP mix will need to be made for specific use depending on the labeled nucleotide being used. The mix should be 2mM for each dNTP that is not the same as the labeled nucleotide. The dNTP mix should include 1 mM of the unlabeled version of the labeled nucleotide desired. It is best to make this solution as a stock and use it when needed. See the following examples:

4. Add 5 µL of homemade dNTP solution to the cooled, primer-annealed probe template, 5µL fluorescent conjugated nucleotide (e.g., Cy3-dCTP), and 1µL Klenow Fragment.

5. Incubate the mixture at 37 °C for at least 2 hours. This step can go overnight if desired.

6. After labeling, add 5 µL of the Stop Buffer. Purify the probe using a G50 Sephadex Column (GE Healthcare). Follow the manufacturer’s instructions for purification.

7. Once purified, add 10 µL 20 mg/mL tRNA.

8. Add 3M NaOAc to a final concentration of 300mM. This should be approximately 7.23uL if volumes have not deviated from these instructions.

9. Add 2.5 volumes of 100% ethanol (approximately 200 µL). Mix the solution thoroughly, but gently. Allow the DNA to precipitate for at least 5 minutes at RT.

10. Spin the preparation at top speed in a microcentrifuge for 20 minutes at 4 °C.

11. Resuspend the pellet in 360 µL RNase-free ddH₂O. Then add 40 µL 3M NaOAc and 1 mL 100% ethanol. Mix the solution thoroughly, but gently. Store the labeled probe at -20 °C.

Probe Precipitation

If using two different probes (labeled with different fluorophores), then you can either use the two probes precipitated separately in a final hybridization solution or precipitate them together. To increase the activity of a probe, the starting amount (100 µL) may be increased, however it is
important to proportionately increase the amount of tRNA, Salmon Sperm DNA, and, most importantly, mouse COT-1 DNA in the subsequent steps. For example, if a combined volume of 200 µL of two probes is used for precipitation, tRNA and DNA additives must be doubled from the amounts described below. The final hybridization solution volume, however, should remain 100 µL per probe input.

1. Aliquot 100uL of stored probe. This is approximately 71.4% ethanol.

2. Add 300 µg of tRNA. This should be 15uL of 20 mg/mL stock (Invitrogen).

3. Add 15 µg of mouse COT-1 DNA. This should be 15uL of 1 mg/mL stock (Invitrogen).

4. Add 150 µg of sheared Salmon Sperm DNA. This should be 15 µL of 10 mg/mL stock (Invitrogen)

5. Add 10 µL 3M NaOAc, 45 µL RNase-free ddH₂O, and 250 µL 100% ethanol. Spin at top speed in a microcentrifuge for 20 minutes at RT.

6. Wash the precipitate with 70% ethanol. Vortex until the precipitate releases from the microcentrifuge tube, then spin at top speed for 2 minutes at RT.

7. Wash the precipitate with 100% ethanol. Vortex until the precipitate releases from the microcentrifuge tube, then spin at top speed for 2 minutes at RT.

8. Dry the precipitate thoroughly. There must be NO trace of ethanol remaining. Use a speed vacuum microcentrifuge without heat for approximately 5 minutes. Check that there is no residue. If residue persists, spin longer.

9. Resuspend the dried precipitate in 50 µL of 100% deionized formamide.

TIP: Formamide needs to be fresh (no older than 1 month at 4°C). It helps to use the pipette tip to crush the pellet after adding formamide. If you do this, be sure all visible pellet is dissolved before discarding pipette tip. Sometimes crushed pellet can get stuck in the tip.
10. Denature the resuspended probe for 10 minutes at 80-100 °C. After 10 minutes, move probe to ice immediately.

   **TIP:** It may be a good idea to wrap the lids with parafilm. If you do this, however, the parafilm will be sticky when trying to remove.

11. Make fresh 2X Hybridization Solution.

   **TIP:** 50% Dextran Sulfate is very viscous. It would be wise to make this solution prior to this step. Pay attention to the approximate volume in the pipette tip when pipetting water, SSC, and BSA. When it is time to pipette the Dextran Sulfate, break it into two parts and use approximately twice the volume setting for what is needed (assuming a P200 pipetman is used). Keep an eye on the volume level in the pipette tip and remove from stock when enough is withdrawn. The double volume setting in a P200 pipetman works well as an approximation.

12. Allow the probe to cool on ice for at least 1 minute. Quick spin at room temperature to consolidate probe to the bottom of the microcentrifuge tube.

13. Add 50 µL 2X Hybridization Solution. Use the pipette to mix the final solution thoroughly.

   **TIP:** It helps to stir the solution with the pipette tip while triturating.

14. Pre-anneal the probe at 37 °C for 1-1.5 hours. Store probe at -20 °C.

15. Using 8-10 µL of probe per 22 mm x 22 mm coverslip. Stretch parafilm out over a glass plate. Pipette the desired amount of probe onto the parafilm. Then lay the coverslip (cells down) onto the probe bubble. Use forceps to press the coverslip and ensure that no visible bubbles between parafilm and coverslip are present.

   **TIP:** If a probe has high specific activity, sometimes it can be used at up to 4 fold dilution. You will need to test activity by serial dilutions on coverslips. Make sure final hybridization solution is prepared as 1X Hybridization Solution.
Preparation of Coverslips

1. Acquire NFx1 TS cells cultured on coverslips. The cells should not be pre-plated prior to seeding on coverslips.

2. Rinse the coverslips twice with 1X PBS. Fix with 2% PFA in 1X PBS for 10 minutes at RT.

3. Permeabilize the cells. Rinse the PFA out with 1X PBS, then treat with ice-cold 1X PBS for 30 seconds, ice-cold 0.5% PBS-T (0.5% Triton X-100 in 1X PBS) for 1 minute, then ice-cold 1X PBS for 30 seconds. If desired, whole slips may be treated with RNase at this point. Use 20 µL of 100 µg/mL RNase A per coverslip and incubate at 37 °C for 10 minutes.

4. Rinse the cells twice with 1X PBS and store them at -20 °C in 85% ethanol. Coverslips may be stored for several months.

Hybridization

1. Make sure that the probes to be used have been pre-annealed. If this has not happened, pre-anneal the probe for 1 hour at 37 °C. Prepare a heat block at 80 °C at least 1 hour prior to use.

2. Etch a coverslip in quarters and use forceps to facilitate breaking into pieces. Use approximately 1 quarter of a coverslip per experiment.

3. Dehydrate the coverslip portion using an ethanol gradient of 85%, 95%, and 100% with 2 min RT incubations for each. Allow the coverslips to air dry.

4. While coverslips are drying, place a glass slide (3 coverslip quarters per slide) onto the heat block to allow it to come up to temperature.

5. Prepare 100 µL of fresh Denaturation Solution for each coverslip. Place 100 µL of Denaturation Solution per coverslip (max 3 per slide) onto the heated glass slide and invert the coverslip (cell side down) onto the solution.

6. Incubate the coverslips at 80 °C for 15 minutes.
TIP: Denaturation may need to be optimized for different probes.

7. During denaturation, pre-chill 2X SSC in an appropriate number of wells on ice. 2X SSC needs to be ice-cold by the time denaturation is complete.

8. Upon completion of denaturation, immediately invert each coverslip into one of the pre-chilled wells of 2X SSC and incubate on ice for 5 minutes. Place the coverslips in the well so that they are cell side up and use forceps to push them to the bottom to ensure complete submersion in the rinse buffer.

9. After at least 5 minutes, replace the 2X SSC with pre-chilled 70% ethanol and incubate on ice for at least 5 minutes or until the probes are finished pre-annealing.

10. Move the plate off of the ice and dehydrate the coverslips with an ethanol gradient of 85%, 95%, and 100% for 2 min each at RT. Allow the coverslips to air-dry.

11. Prepare a hybridization chamber in a slide box by lining the base with paper towels and soaking them with Wash Solution 1. Prepare a platform by stretching parafilm over a glass plate for western blots and place it in the hybridization chamber. Using a pencil, label the parts of the plate according to which hybridization will be placed there. Pipette 10μL pre-annealed probe in hybridization solution per 22x22 mm coverslip onto the prepared plate in the desired location. Invert the dried coverslip onto the probe (cell side down) and use forceps to ensure that the coverslip is completely covered by the probe solution. Close the hybridization chamber and place at 37 °C overnight.

TIP: If RNA and DNA are being probed in the same experiment, pre-hybridize with the RNA probe for at least 3 hours at 37 °C. After this pre-hybridization step, lift the coverslips and begin the overnight hybridization with the desired DNA probes. Annealing temperature, in both the probe pre-annealing step and in the overnight hybridization, can make a difference on hybridization efficiency. 37C is the standard expected temperature, but this can be altered if necessary.
12. Prepare wash solutions and warm them at 37 °C overnight. The stringency of the wash will increase with decreasing SSC concentration. The protocol for washing may need to be optimized for each probe, but begin with the following method: 3 washes with Wash Solution 1, 3 washes with 1X SSC, 2 washes with 0.5X SSC. Plan for 500uL per wash per coverslip (i.e., 1.5mL Wash 1, 1.5mL Wash 2, 1 mL Wash 3 per coverslip).

13. Upon completion of hybridization, wash the coverslips 3 times with Wash Solution 1, 3 times with 1X SSC, and 2 times with 0.5X SSC. All Wash solutions should be pre-warmed and each wash should be incubated at 39 °C for 5 minutes.

TIP: Temperature and timing is very important in this step, so determine a method for staggering workload prior to beginning the washes. Prior to use, add DAPI at an appropriate concentration to the 1X SSC to be used in the second wash step.

14. During the wash steps, bring an aliquot of Prolong Gold w/o DAPI (Molecular Probes) to room temperature. Approximately 8uL will be needed for mounting each quarter-slip.

15. After wash steps are complete, mount the coverslips with the Prolong Gold mounting media. Make sure that no bubbles are present between the slide and the coverslip. Place the slides in a dark place and allow them to cure overnight at RT. After overnight cure, store slides in a slide box at 4 °C.

Analysis Methods

Z-stack images were taken in grayscale at 63X using a Zeiss AxioImager M2 equipped with an AxioCam MRm camera (Carl Zeiss). Merged Z-stacks were deconvolved using the Iterative algorithm in the AxioVision software package (Carl Zeiss). Distances were measured in 3 dimensions from the approximate centers of DNA-FISH signals using the ZEN 2011 black edition software package (Carl Zeiss). Recorded measurements were used to produce cumulative distance plots with R. To produce representative images, z-projections for each color channel were produced and
exported with AxioVision, then merged and pseudo-colored with Adobe Photoshop CS5.1 (Adobe Systems).

Recipes for Solutions

The following recipes are for solutions used in the FISH protocols described in this section.

Denaturation Solution

- 70% Formamide
- 2X SSC

Hybridization Chamber Solution

- 50% Formamide
- 2X SSC

Homemade dNTP Solution for Labeling

- If the probe is to be labeled with a Cy3-dCTP, the homemade dNTP mix should consist of 2mM dATP, 2mM dGTP, 2mM dTTP, and 1mM dCTP.
- If the probe is to be labeled with a FITC-dUTP, the homemade dNTP mix should consist of 2mM dATP, 2mM dGTP, 2mM dCTP, and 1mM dTTP

2X Hybridization Solution

- 1 part RNase-free ddH₂O (Sigma)
- 1 part 20X SSC
- 1 part 10 mg/mL BSA (NEB)
- 2 parts 50% Dextran Sulfate
Allele-Specific qRT-PCR

In this section, the protocols and analysis methods used to verify allelically biased expression of the candidate genes are described. qPCR was performed using the SsoFast™ EvaGreen® SuperMix from Bio-Rad on a Bio-Rad CFX96 Real-Time PCR Detection System. RNA was harvested from TS cells using the TRizol Reagent and the manufacturer’s procedures (Ambion), treated with DNase I, and purified using the protocols described for the RNeasy® Mini Kit (Qiagen). RNA preps were quantitated using a Nanodrop 8000 (ThermoFisher) and cDNA samples were prepared from 250 ng RNA using SuperScript® III Reverse Transcriptase (Invitrogen).

Amplifications were performed over 40 cycles using a 10 second denaturation step and a 30 second annealing step with no primer extension. Melt curves were performed by running a 0.5 °C increment in 5 second steps from 65-95 °C. Annealing temperatures for each gene were as follows: Airn, Igf2r, Ibtk, Car2, and Bphl at 68 °C, Mtus1 at 66 °C, Tlr5 at 65 °C.

Informatics Pipeline

The informatics analysis was performed using a series of custom scripts and programs. As best as possible, programs were arranged in a series of shell scripts to be successively run on the Rocks Linux cluster at UNC. Below is an overview of the informatics pipeline used. The fourSig suite is available through SourceForge at http://sourceforge.net/projects/foursig/. The other programs and scripts referenced may be made available by request from the Magnuson Lab.

Alignment

FASTQ files are first decompressed using the gzip utility and analyzed for quality using the FASTX toolkit (Hannon Laboratory). Decompressed FASTQ files may then be aligned using the 4CSeqSplitAndAlign.pl program written in Perl. Aligned reads are then paired in both allelic and non-allelic fashions with the 4CFullPair.pl program written in Perl. Both of these programs perform their work as described in Chapter 3 (Illustration 3.1).
Statistical Analysis

At this point, FASTQ files are typically recompressed using the `gzip` utility. Mapped reads files are organized in to the tab files that serve as inputs for analysis with the programs in the `fourSig` suite as described in Chapters 2 and 3. This analysis is performed independently for each gene replicate using a script written in R. Output files for the statistical enrichment windows are generated and typically plotted with basic XY plots using graphical utilities in R. Allelic assignments are then made based on the probability that allelic reads are sufficiently enriched within a window to indicate a bias over equal allelic detection using the 4CAssign.pl and 4CAssignFast.pl programs written in Perl. These files produce output files describing enriched windows based on their likelihood to be enriched for SNPs in one allele.

Allelic Analysis

After `fourSig` analysis is performed on each replicate, set analysis may be performed as described in Chapter 3. The order of operations is typically carried out using separate scripts written in R for each `fourSig` analysis described. Several custom functions were written to facilitate this analysis in a program called SetAnalysisFunctions.r. Additionally, the Euler diagrams displayed in Chapter 3 are drawn using the VennDiagram package for R (Chen and Boutros 2011). Each area displayed within the diagrams is assembled using the subset command in R and resulting lists are printed in bedGraph format for visualization in the UCSC Genome Browser. The interallelic analyses are performed in R after completion of the replicate analysis using many of the same functions. General forms of the R scripts were produced, however heavy customization is required for each independent usage.
APPENDIX: AMPLICON EFFICIENCY

### Airn qRT-PCR Efficiency Curves

- **B6 Primer w/ B6 DNA**
- **B6 Primers w/ Cast DNA**
- **Cast Primers w/ Cast DNA**
- **Cast Primers w/ B6 gDNA**

- Linear (B6 Primer w/ B6 DNA)
- Linear (B6 Primers w/ Cast DNA)
- Linear (Cast Primers w/ Cast DNA)
- Linear (Cast Primers w/ B6 gDNA)

### Igf2r qRT-PCR Efficiency Curves

- **B6 Primer w/ B6 DNA**
- **B6 Primers w/ Cast DNA**
- **Cast Primers w/ Cast DNA**
- **Cast Primers w/ B6 gDNA**

- Linear (B6 Primer w/ B6 DNA)
- Linear (B6 Primers w/ Cast DNA)
- Linear (Cast Primers w/ Cast DNA)
- Linear (Cast Primers w/ B6 gDNA)
Ibtk qRT-PCR Efficiency Curves

- $y = -3.2967x + 25.74$, $R^2 = 0.992$
- $y = -2.962x + 26.647$, $R^2 = 0.9931$

Car2 qRT-PCR Efficiency Curves

- $y = -3.3793x + 25.559$, $R^2 = 0.9984$
- $y = -3.177x + 21.729$, $R^2 = 0.9982$
Mtus1 qRT-PCR Efficiency Curves

- **B6 Primer w/ B6 DNA**
  - $y = -3.1677x + 36.687$
  - $R^2 = 0.9912$
- **B6 Primers w/ Cast DNA**
  - $y = -1.497x + 39.132$
  - $R^2 = 0.6515$
- **Cast Primers w/ Cast DNA**
  - $y = -3.2673x + 35.869$
  - $R^2 = 0.9993$
- **Cast Primers w/ B6 gDNA**
  - $y = -1.6273x + 40.933$
  - $R^2 = 0.9003$

Log10 Conversion of input DNA (pg)

Average C(t)

Tlr5 qRT-PCR Efficiency Curves

- **B6 Primers w/ B6 DNA**
  - $y = 3.358x + 19.139$
  - $R^2 = 0.0506$
  - $y = 3.0027x + 36.86$
  - $R^2 = 0.9749$
- **B6 Primers w/ Cast DNA**
  - $y = -0.0195x + 37.064$
  - $R^2 = 0.0001$
- **Cast Primers w/ Cast DNA**
  - $y = -3.2047x + 37.339$
  - $R^2 = 0.9995$
- **Cast Primers w/ B6 gDNA**
  - $y = -3.1677x + 36.887$
  - $R^2 = 0.9912$

Log10 Conversion of input DNA (pg)

Average C(t)
Bphl qRT-PCR Efficiency Curves

- Linear (B6 Primers w/ B6 DNA)
- Linear (B6 Primers w/ Cast DNA)
- Linear (Cast Primers w/ Cast DNA)
- Linear (Cast Primers w/ B6 gDNA)

Data Points:

- y = -3.4x + 37.362, R² = 0.997
- y = -2.5183x + 36.548, R² = 0.9909
- y = -2.14x + 36.243, R² = 0.9713
- y = -3.165x + 36.116, R² = 0.9999
REFERENCES


Korostowski L, Sedlak N, Engel N. 2012. The Kcnq1ot1 long non-coding RNA affects chromatin conformation and expression of Kcnq1, but does not regulate its imprinting in the developing heart. PLoS Genet. 8:e1002956.


Mugford JW, Starmer J, Williams RL, Calabrese JM, Mieczkowski P, Yee D, Magnuson T. Escape from X chromosome inactivation occurs within topologically associated domains. Prep.


