DNA Methylation Patterns of Specific L1 Loci on the Short Arm of Chromosome 21

Sarah Ann Tincher
Loyola University Chicago

Recommended Citation
http://ecommons.luc.edu/luc_theses/3153

This Thesis is brought to you for free and open access by the Theses and Dissertations at Loyola eCommons. It has been accepted for inclusion in Master's Theses by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License.
Copyright © 2015 Sarah Ann Tincher
Copyright by Sarah A. Tincher, 2016
All rights reserved
ACKNOWLEDGMENTS

I would first like to express my gratitude to my advisor, Dr. Jeffrey Doering, for his unwavering support, patience, and encouragement throughout my years at Loyola University Chicago. He taught me what it means to be a scientist and gave me the chance to develop the skills I would need for a career in research. I can say, without a doubt, that I would not be where I am without his mentorship. I would also like to thank the faculty, staff, and fellow students in the Department of Biology, particularly my committee members, Dr. Howard Laten and Dr. Stefan Kanzok, for their input and moral support throughout this entire process. I must also acknowledge the many individuals with whom I have had the pleasure to share a bench in the Doering lab: we laughed, we cried, and we high-fived. Always remember that “the world is our methylation oyster.”

Finally, there are very few words in the English dictionary that I can use to express how grateful I am to my family and close friends, particularly my parents and brothers. Their steadfast faith in me gave me the courage to pursue my dreams and the motivation to keep going even when it all seemed impossible.
To my parents, for always believing in me.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ iii

LIST OF TABLES .................................................................................................................... vii

LIST OF FIGURES .................................................................................................................. viii

ABSTRACT .............................................................................................................................. xi

CHAPTER ONE: INTRODUCTION ......................................................................................... 1
The Heterochromatic Portion of the Human Genome .......................................................... 1
An Overview of LINE-1 Retrotransposons ........................................................................... 2
DNA Methylation and The Regulation of L1s ...................................................................... 4
The Role of L1s in X Chromosome Inactivation ................................................................ 6
L1s and Cancer ..................................................................................................................... 7
Research Approach and Rationale ...................................................................................... 10

CHAPTER TWO: MATERIALS AND METHODS .................................................................. 12
DNA from Cell Lines and Tissues ....................................................................................... 12
Bisulfite Sequencing PCR (BSP) .......................................................................................... 12
Primer Design ...................................................................................................................... 12
Bisulfite Conversion ............................................................................................................ 13
Polymerase Chain Reaction (PCR) .................................................................................... 14
Cloning and Transformation ............................................................................................. 15
Sequence Analysis ............................................................................................................. 16

CHAPTER THREE: RESULTS .............................................................................................. 27
Analysis of Locus Specificity ............................................................................................... 27
Methylation Analysis in Hybrid Cell Lines and Normal Peripheral Leukocytes .............. 31
Locus 1 ............................................................................................................................... 31
Locus 2 ............................................................................................................................... 33
Locus 3 ............................................................................................................................... 35
Locus 4 ............................................................................................................................... 36
Long Arm Locus 1 ............................................................................................................. 37
Long Arm Locus 2 ............................................................................................................. 38
Future Analysis of HC21q ................................................................................................. 38
LRE3 ................................................................................................................................. 39
Methylation Comparisons among Loci .............................................................................. 41
Analysis of Critical CpGs ................................................................................................. 41
Methylation Analysis in Prostate Cell Lines .................................................................... 42
Locus 1 ............................................................................................................................... 42
Locus 2 ............................................................................................................................... 46
Long Arm Locus 1 ............................................................................................................. 48
LRE3 ................................................................................................................................. 49
LIST OF TABLES

Table 1. Subfamily and Positions of L1 Loci .................................................................17
Table 2. Primer Sequences for Each Locus .......................................................................18
Table 3. Primer Information for Outside PCRs .................................................................19
Table 4. Primer Information for Nested PCRs .................................................................20
Table 5. Total Percent Methylation of L1 loci .................................................................56
Table 6. Percent Methylation at Each Critical CpG Site .................................................77
Table 7. Summary of Critical CpG Site Methylation in Leukocytes .................................112
Table 8. Conclusions for HC21 L1s as Cancer Biomarkers .............................................112
# LIST OF FIGURES

Figure 1. The Structure of an L1 Element .................................................................11

Figure 2. Depiction of the Difference between Overall Percent Methylation and Methylation Patterns .................................................................11

Figure 3. Locus 1 PCR on Converted Human Control DNA .........................21

Figure 4. Locus 2 PCR on Converted Human Control DNA .........................21

Figure 5. Locus 3 PCR on Converted Human Control DNA .........................22

Figure 6. Locus 4 PCR on Converted Human Control DNA .........................22

Figure 7. Long Arm Locus 1 PCR on Converted Human Control DNA ..........23

Figure 8. Long Arm Locus 2 PCR on Converted Human Control DNA ..........23

Figure 9. LRE3 PCR on Converted Human Control DNA .............................24

Figure 10. Locus 1 Acrocentric Chromosome PCR Gel ..............................24

Figure 11. Locus 2 Acrocentric Chromosome PCR Gel ..................................25

Figure 12. Locus 3 Acrocentric Chromosome PCR Gel ..................................25

Figure 13. Locus 4 Acrocentric Chromosome PCR Gel ..................................26

Figure 14. Locations of L1 Loci 1-4 in an Alphoid Region on HC21p .............55

Figure 15. Locus 1 in WAV17 ..............................................................................57

Figure 16. Locus 1 in WAV17 Using Consensus Sequence of All L1 Loci .........57

Figure 17. Locus 1 in Leukocytes .....................................................................58

Figure 18. Locus 1 in Leukocytes Using Consensus Sequence of All L1 Loci ....58

Figure 19. Locus 2 in WAV17 ..............................................................................59

Figure 20. Locus 2 in WAV17 Using Consensus Sequence of All L1 Loci .........59
Figure 21. Locus 2 in Leukocytes .................................................................60
Figure 22. Locus 2 in Leukocytes Using Consensus Sequence of All L1 Loci ......60
Figure 23. Locus 3 in WAV17 .......................................................................61
Figure 24. Locus 3 in WAV17 Using Consensus Sequence of All L1 Loci ........62
Figure 25. Locus 3 in Leukocytes .................................................................63
Figure 26. Locus 3 in Leukocytes Using Consensus Sequence of All L1 Loci ......64
Figure 27. Locus 4 in WAV17 .......................................................................65
Figure 28. Locus 4 in WAV17 Using Consensus Sequence of All L1 Loci ........66
Figure 29. Locus 4 in Leukocytes .................................................................67
Figure 30. Locus 4 in Leukocytes Using Consensus Sequence of All L1 Loci ......68
Figure 31. Long Arm Locus 1 in WAV17 .......................................................69
Figure 32. Long Arm Locus 1 in WAV17 Using Consensus Sequence of All L1 Loci .......................................................................................................70
Figure 33. Long Arm Locus 1 in Leukocytes .................................................71
Figure 34. Long Arm Locus 1 in Leukocytes Using Consensus Sequence of All L1 Loci .......................................................................................................72
Figure 35. Long Arm Locus 2 in WAV17 .......................................................73
Figure 36. Long Arm Locus 2 in WAV17 Using Consensus Sequence of All L1 Loci .......................................................................................................73
Figure 37. Long Arm Locus 2 in Leukocytes .................................................74
Figure 38. Long Arm Locus 2 in Leukocytes Using Consensus Sequence of All L1 Loci .......................................................................................................74
Figure 39. LRE3 in HC2 .................................................................................75
Figure 40. LRE3 in Leukocytes .................................................................76
Figure 41. Locus 1 in RWPE-1 .......................................................................78
Figure 42. Locus 1 in RWPE-1 Using Consensus Sequence of All L1 Loci ......78
Figure 43. Locus 1 in LNCaP ..................................................................................79
Figure 44. Locus 1 in LNCaP Using Consensus Sequence of All L1 Loci ..........79
Figure 45. Locus 1 in DU145 ..................................................................................80
Figure 46. Locus 1 in DU145 Using Consensus Sequence of All L1 Loci ..........81
Figure 47. Locus 2 in RWPE-1 ..................................................................................82
Figure 48. Locus 2 in RWPE-1 Using Consensus Sequence of All L1 Loci ..........83
Figure 49. Locus 2 in LNCaP ..................................................................................84
Figure 50. Locus 2 in LNCaP Using Consensus Sequence of All L1 Loci ..........85
Figure 51. Locus 2 in DU145 ..................................................................................86
Figure 52. Locus 2 in DU145 Using Consensus Sequence of All L1 Loci ..........87
Figure 53. Long Arm Locus 1 in RWPE-1 .................................................................88
Figure 54. Long Arm Locus 1 in RWPE-1 Using Consensus Sequence of All L1 Loci ..................................................................................................................89
Figure 55. Long Arm Locus 1 in LNCaP ................................................................. 90
Figure 56. Long Arm Locus 1 in LNCaP Using Consensus Sequence of All L1 Loci ..................................................................................................................91
Figure 57. Long Arm Locus 1 in DU145 ................................................................. 92
Figure 58. Long Arm Locus 1 in DU145 Using Consensus Sequence of All L1 Loci ..................................................................................................................93
Figure 59. LRE3 in RWPE-1 .................................................................................... 94
Figure 60. LRE3 in LNCaP .................................................................................... 95
Figure 61. LRE3 in DU145 .................................................................................... 96
Figure 62. Mechanisms for RNAi-Mediated Chromatin Modifications .......... 113
Figure 63. Proposed Mechanism of Heterochromatin Formation on HC21p .... 114
ABSTRACT

While 10-15% of the human genome is composed of heterochromatic DNA, these regions are not included in the completed genome sequence. The short arm of chromosome 21 (HC21p) serves as a model for understanding the structure and function of heterochromatin. LINE-1 (L1) retrotransposons are underrepresented in heterochromatin, including HC21p, and have not been extensively studied in these regions. However, there are disproportionately more full length L1s on HC21p than in euchromatic genomic regions. Decreased DNA methylation in the promoters of these L1s on HC21p may allow them to facilitate heterochromatin formation, which would be analogous to their proposed role in X chromosome inactivation. Using bisulfite sequencing PCR on both chromosome-specific hybrid cell lines and leukocytes, I found that overall methylation of four specific L1s on HC21p was substantially lower than control L1 loci located in euchromatic regions. This trend was also present for four critical CpG sites in the L1 promoter that have been specifically implicated in the suppression of L1 expression. I propose a model wherein decreased L1 promoter methylation promotes heterochromatin formation by intrachromosomal interactions and transient bidirectional expression.

Additionally, due to evidence that some L1s become hypomethylated in cancer cells, a limited number of HC21p L1 loci were analyzed in a small number of prostate cancer cell lines to determine if changes in methylation at those loci
could be relevant to the development of a cancer biomarker. I found that the L1s on HC21p were generally less methylated in cancer versus normal cell lines and differed between loci at different tumor stages. This supports the hypothesis that changes in methylation of a variety of specific L1 loci may be sensitive biomarkers for both cancer detection and prognosis.
CHAPTER ONE
INTRODUCTION

The Heterochromatic Portion of the Human Genome

The Human Genome Project covered the euchromatic regions of the genome, but not the 10-15% of the genome consisting of heterochromatin. Heterochromatin is highly condensed and composed of clusters of tandemly repetitive DNA sequences. The tandem repeats, or satellite DNAs, exist as a number of discrete sequence families, each with its own characteristic repeat sequence (Lee et al., 1997). These satellite regions are very similar in sequence on the short arms of all the acrocentric chromosomes (13, 14, 15, 21, and 22), resulting in nonhomologous pairing during meiosis (deCapoa et al., 1973). This incorrect pairing of nonhomologous regions can lead to nondisjunction.

HC21p is an ideal model for studying the structure and function of heterochromatin (Doering et al., 1993). HC21 is the smallest chromosome in the genome and a large percentage of it is heterochromatin located on the short arm. The short arm of this chromosome (HC21p) has already been very well characterized, more so than any other acrocentric chromosome in the genome (Doering et al., 1993). It is a mosaic composed primarily of large clusters of tandem repeats that are interspersed with islands of low copy number repeats. The satellite families include alphoid (α), beta (β), satellite I and satellite DNA, and there are also many interspersed sequence families, including
retrotransposons (Doering et al., 1993).

**An Overview of LINE-1 Retrotransposons**

Transposable elements make up roughly 45% of the non-coding portion of the genome (Lander et al., 2001). These mobile elements are divided into four categories: long interspersed elements (LINEs), short interspersed elements (SINEs), long terminal repeat (LTR) retrotransposons, and DNA transposons. LINEs comprise almost half of the total transposons in the human genome (~21%). Thus, they are largely considered be the most successful of all the transposable elements (Lander et al., 2001; Boissinot et al., 2005). The persistence of LINE activity over the course of hundreds of millions of years of eukaryotic evolution is likely due to the adaptability of these elements to host pressures (Boissinot et al., 2005). As the host evolved to respond to LINE mobility, novel LINE families emerged in order to maintain retrotransposition capability (Boissinot et al., 2005). The dominant LINE family is LINE-1 (L1) which constitutes 17% of the total genome (Boissinot et al., 2005). L1s can be divided into subfamilies based on shared nucleotide differences, specifically in the 3’ UTR, that correlate with the age of the element (Smit et al., 1995; Boissinot et al., 2005). The majority of L1s in the human genome are ancient, dating back to before the divergence of the primate lineage (Smit et al., 1995). Over the past 25 million years, five families of primate-specific L1s have emerged: L1PA5 (oldest), L1PA4, L1PA3, L1PA2, and L1PA1 (youngest). The human-specific L1PA1 (L1PHs) is the predominantly active L1 in the genome and retrotransposition of elements belonging to this family have been implicated in numerous genetic disorders (Hancks and Kazazian, 2012).
A full-length L1 element is ~6 kb and consists of a 5'-untranslated region (5'-UTR) with both sense and antisense internal promoters, two ORFs (ORF1 and ORF2) that encode the proteins required for autonomous retrotransposition, a 3'-UTR, and a weak polyadenylation signal followed by a polyA tail (Figure 1). The 5'-UTR contains several transcription regulatory elements, including transcription factor (TF)-binding sites. The yin yang 1 (YY1) TF-binding site (+13 to +21) is known to be important for transcription initiation by correctly positioning the initiation protein complex (Kurose et al., 1995; Athanikar et al., 2004). There are also two sex-determining region Y (SRY) TF-binding sites (+472 to +477; +572 to +577) that are located near the antisense promoter and may be necessary for the regulation of transcription in the antisense direction (Tchénio et al., 2000). Runt-domain transcription factor 3 (RUNX3)-binding sites are present at two separate sites within the promoter (+83 to +101; +526 to +508) and are also considered to be important for transcription initiation (Yang et al., 2003). In addition, there are four methyl-CP2 (MeCP2)-responsive elements (+36, +101, +304, +481) that are known to be involved in the regulation of chromatin structure (Hata and Sakai, 1997). The protein encoded by ORF1 (ORF1p) is proposed to provide nucleic acid chaperone activities involved in the assembly of a ribonucleoprotein complex in the cytoplasm (Doucet et al., 2010). ORF2 encodes the ORF2p protein that contains an endonuclease and a reverse transcriptase domain, both of which are essential for completion of retrotransposition (Rhabari et al., 2015). The genes located within the L1 element are transcribed and the resulting mRNA, upon export to the cytoplasm, is translated to generate the proteins for retrotransposition via a mechanism called
target primed reverse transcription, or TPRT. Once the L1 mRNA that is complexed with ORF1p and ORF2p re-enters the nucleus, the endonuclease, which mainly targets AT-rich sequences, will make a nick at the integration site. The free 3’-OH created by this cut serves as a primer that the reverse transcriptase then uses to generate a DNA sequence complementary to the mRNA (cDNA) at the target site (Rhabari et al., 2015). The second strand of the copied L1 sequence is then polymerized to make a double-stranded DNA, flanked by target site duplications (TSDs). Most L1s are 5’-truncated because reverse transcriptase fails to generate a complete cDNA copy of the L1 mRNA at the site of integration. Thus, the majority of L1s in the genome lack the internal promoter and are transcriptionally inactive. Since L1s do not mobilize by a “cut-and-paste” mechanism, but rather by a “copy-and-paste” mechanism, these elements contributed significantly to the expansion of primate genomes (Lander et al., 2001).

DNA Methylation and the Regulation of L1s

The L1 internal sense promoter is located in the highly conserved, GC-rich 5’ UTR and contains a CpG island that serves as an epigenetic target for DNA methylation (Thayer et al., 1993). Indeed, 90% of methylated CpGs in the genome occur in retrotransposons, indicating the importance of epigenetic regulation of these elements (Goll and Bestor, 2005). The major consequences of L1 mobilization are genomic instability due to insertion events that induce double-strand breaks (DSBs) and structural rearrangements (Gasior et al., 2006; Gilbert et al., 2002). In response to transposon mobility, the host necessarily evolved mechanisms to suppress potentially detrimental mobilization events. The
detection of antisense transcripts originating from the antisense promoter (ASP) within the L1 5’ UTR has implicated the RNA interference (RNAi) machinery as a potential mechanism for repression (Chow et al., 2010; Yang and Kazazian, 2006). However, DNA methylation within L1 promoters is currently the most well-studied host defense mechanism to suppress L1 retrotransposition. While 5’ truncations do prevent the vast majority of L1s from mobilizing, DNA methylation functions to suppress the retrotransposition of the full-length L1s containing an intact promoter (Hata and Sakai, 1997). Four specific CpG sites located within the promoter region when methylated appear to be necessary and sufficient for repression of L1 transcription in vitro (Hata and Sakai, 1997).

DNA methylation is a heritable epigenetic modification that has been extensively studied since 1975, when it was originally proposed to be involved in gene regulation and cellular differentiation in eukaryotes (Holliday and Pugh, 1975; Riggs, 1975). The establishment as well as maintenance of DNA methylation is a complex process involving multiple components. The addition of a methyl group to the fifth carbon in the cytosine pyrimidine ring to form 5-methylcytosine (5-mC) is catalyzed by DNA methyltransferases (DNMTs). This epigenetic mechanism is not only necessary for silencing of repetitive elements, but also for the process of X chromosome inactivation, imprinting, and the regulation of tissue-specific gene expression (Pogribny and Beland, 2009). Local and global disruptions in DNA methylation patterns have been observed during the aging process as well as in cancer cells (Liu et al., 2003). Hypermethylation and hypomethylation indicate a deviation from the “normal” methylation state and the degree to which methylation is altered is often specific to the cell type.
(Dunn, 2003). While it is unlikely that one single mechanism is responsible for the establishment of these aberrant methylation states, altered functioning of DNMTs likely plays a large role (Li et al., 1992). Consistent with this hypothesis, experimentally reducing the expression of DNMTs or overriding its functionality with demethylating agents results in a significant reduction in total genomic DNA methylation (Gaudet et al., 2003).

The Role of L1s in X Chromosome Inactivation

The human X chromosome is enriched 2-fold in the density of L1s compared to the autosomes (Bailey et al., 2000). The majority of these overrepresented L1s belong to the evolutionarily younger subfamilies and are highly concentrated in the regions important for X chromosome inactivation (XCI): the X inactivation center and the X inactive specific transcript (XIST) locus. Thus, it has been proposed that these L1s are involved in the assembly of heterochromatic nuclear compartments during XCI, perhaps acting as “way stations” to ensure the spreading of heterochromatin (Lyon, 1998; Chow et al., 2010; Hansen, 2003). While most of the L1s expected to be involved in XCI are transcriptionally inactive, an unexpected finding in mouse embryonic stem cells was that some young, full-length L1s are expressed from the X chromosome that is undergoing inactivation (Chow et al., 2010). These expressed L1s may be facilitating the spread of silencing into certain regions of the X chromosome that are prone to escaping inactivation. Indeed, Hansen et al. (2003) have provided evidence that the expression of L1s on the inactive X may be due to the hypomethylation of the L1 promoters. These data emphasize the theory that mobile elements, including LINEs, have the capability to regulate genes. Another
study provided evidence for this hypothesis by demonstrating that methylation levels of intragenic L1s are potentially involved in the regulation of their host genes (Aporntewan et al., 2011). In cells where L1s were hypomethylated, genes containing L1s were more likely to be repressed than genes without L1s. This indicates that the methylation patterns of L1s may be important for controlling their surrounding chromatin structures as well as the expression of nearby genes.

Previous work has shown that while L1 elements are underrepresented in heterochromatin and on HC21p, there is a disproportionate number of full-length L1s on HC21p (Beris, 2003). In a given HC21-specific hybrid cell line, there are twenty times as many full-length L1s on the short arm as on the long arm. There is also substantial population polymorphism in the copy number of L1s on HC21p (Beris, 2003). Due to this skewed presence of full-length L1s on HC21p, it is possible that their presence in these regions may be involved in initiating and/or maintaining the highly condensed chromatin state as is seen on the X chromosome, possibly through DNA methylation dynamics. DNA methylation patterns of L1s in heterochromatic regions of the genome, such as the acrocentric chromosome short arms, have not previously been studied.

L1s and Cancer

It has been known for decades that global genomic alterations of DNA methylation is one of the most common characteristics of tumor cells (Kanai, 2010). Interspersed repetitive elements such as L1s are overwhelmingly hypomethylated in malignant cells and as hypomethylation increases, there is an association with advanced tumor grade and poorer prognosis (Kitkumthorn and Mutirangura, 2011). One consequence of hypomethylation may be the
reactivation of previously repressed L1s. In support of this hypothesis, previous studies have correlated the hypomethylation of L1s with an increase in their activity and subsequent de novo insertions (Kitkumthorn and Mutirangura, 2011; Iskow et al., 2010). In the study by Iskow et al. (2010), at least 30% of the lung tumors examined contained one or more de novo L1 insertions that were not seen in the adjacent normal tissue. The lung tumors that were permissive for L1 somatic retrotransposition were significantly less methylated than their matched tissues, although the degree of hypomethylation was not specified (Iskow et al., 2010). In another study, methylation decreased by >10% compared to normal matched tissues in 31% of prostate tumors (Schulz et al., 2002). This was further linked to chromosomal aberrations, such as loss or gain of specific chromosome arms (Schulz et al., 2002). An additional study reported significant decreases in L1 methylation in prostate adenocarcinomas compared to normal controls (35.7% vs. 46.1%) and this hypomethylation was associated with poorer prognostic parameters (Cho et al, 2007). In patients with colon cancer, L1s were significantly less methylated compared to matched normal tissue (61.4% vs. 71.8%), although this study did not find a significant correlation between L1 methylation and clinical features such as tumor grade (Ogino et al., 2008). Another study demonstrated L1s were hypomethylated in 74.5% of chronic myelogenous leukemia (CML) cell lines compared to the normal controls (Roman-Gomez et al., 2005). Furthermore, the degree of hypomethylation increased as the disease progressed and was also correlated with poorer prognosis and response to treatment (Roman-Gomez et al., 2005). The result of L1 hypomethylation observed in CML patients may be associated with dysfunctions such as
chromosomal aberrations, hypermethylation of tumor suppressor genes, and alternate transcription of oncogenes (Roman-Gomez et al., 2005).

The studies described emphasize the importance of global L1 methylation trends in cancer. However, other authors have proposed that if aberrant L1 methylation is driving tumorigenesis, it may not be at a global level (Pobsook et al., 2011; Phokaew et al., 2008; Nüsgen et al., 2015). In fact, the driving force may be alterations in methylation patterns in the promoters of specific L1s. This hypothesis formed after observation that the methylation patterns at specific L1 loci often varied between different tumor cells displaying the same global L1 percent methylation (Figure 2) (Pobsook et al., 2011). For instance, by looking at differences in methylation patterns at specific L1 loci, it was possible to distinguish oral cancer cells from normal cells (Pobsook et al., 2011). This same distinction could not be made upon comparison of genomewide L1 percent methylation (Pobsook et al., 2011). Furthermore, methylation patterns of individual DNA strands at a specific locus are not always identical (Phokaew et al., 2008). This observation indicates that L1 methylation is a dynamic process and genomic decreases in methylation can influence different chromosomes or cells independently (Phokaew et al., 2008). Hypomethylation of intragenic L1s is correlated with reduced mRNA levels of their host genes in cancer cells (Aporntewan et al., 2011). Thus, it has been suggested that individual L1 loci and their respective methylation patterns may play different roles in cancer depending on their position in the genome (Phokaew et al., 2008; Aporntewan et al., 2011). The cumulative implications of these studies indicate that the methylation status of specific L1 loci may serve as potential biomarkers for
cancer detection and prognosis.

**Research Approach and Rationale**

Since L1s and their methylation patterns have never before been investigated in heterochromatic regions of the genome outside of the inactive X chromosome, bisulfite sequencing PCR (BSP) was used to explore DNA methylation patterns in the promoter regions of four specific full length L1s on the heterochromatic HC21p for inter-loci comparison. For control experiments, the four specific L1 loci on HC21p were compared to two full-length L1 loci on the long arm of HC21 (HC21q). This allows for the determination of any differences in the patterns of methylation between heterochromatic and euchromatic regions. As a positive control a human-specific L1 element, LRE3, that has been determined to be retrotranspositionally active was analyzed with the expectation that it is hypermethylated in normal cells (Sassaman et al., 1997). Compared to the euchromatic control on HC21q, the HC21p L1s are hypomethylated in normal tissues, which supports my hypothesis that they may be influencing heterochromatin formation on HC21p.

Due to the evidence that some L1s become hypomethylated in cancer cells, the same HC21p L1 loci were analyzed in a small number of prostate cancer cell lines to determine if changes in methylation at those loci could be relevant to the development of a cancer biomarker. The results show that the L1s on HC21p are overall hypomethylated compared to the HC21q control loci, but the extent varies between loci. It will be necessary to study these L1s in other tissues, as well as include additional L1s on HC21p, in order to determine if these inter-loci differences can be used as a biomarker.
Figure 1. The Structure of an L1 Element. SP: sense promoter; ASP: antisense promoter; TSD: target site duplication; pA: polyadenylation signal.

Figure 2. Depiction of the Difference between Overall Percent Methylation and Methylation Patterns. Both cells have the same overall methylation level (50%), but different patterns of methylation. Closed circles represent methylated cytosines and open circles represent unmethylated cytosines. Adapted from Pobsook et al., 2011.
CHAPTER TWO

MATERIALS AND METHODS

DNA from Cell Lines and Tissues

DNA from the human chromosome 21 (HC21)-containing WAV17 hybrid cell line was obtained from Coriell Cell Repositories. DNAs from hybrid cell lines containing human HC13, HC14, HC15, and HC22 were also obtained from Coriell Cell Repositories and were used for acrocentric chromosome specificity analysis. Genomic DNA from normal human peripheral blood leukocytes was obtained from Biochain. DNAs from the prostate cell lines were generously provided by Dr. Jennifer Doll. RWPE-1 is a normal human prostate epithelial cell line immortalized with HPV and is p53- and Rb-negative. LNCaP is a human prostate androgen-sensitive cancer cell line isolated from a lymph node metastasis and is thus less aggressive. DU145 is a human androgen-insensitive prostate cancer cell line isolated from a brain metastasis and is thus more aggressive.

Bisulfite Sequencing PCR (BSP)

Primer Design

Primers were designed using the criteria described in Darst et al. (2010) from the indicated accession numbers (Table 1). L1 subfamilies were determined using the web-based RepeatMasker tool (http://repeatmasker.org). Each primer set included one primer in the unique sequence adjacent to L1 promoter and one
primer within the conserved L1 promoter. For most loci, a second nested PCR reaction was performed using 1 ul of a 1:50 or a 1:100 dilution of the first product with a primer set that was designed internal to the first primer set. The sequences for each primer are reported in Table 2. The number and location of CpG sites, as well as the length of the amplicons, are specific to each locus (Tables 3 and 4). The specificity of the primer sequences for the sequenced portion of the human genome reference was determined in silico using the NCBI web-based Primer-BLAST tool (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). In order to determine if the primers could successfully amplify bisulfite treated DNA, they were tested on bisulfite converted EpiTect PCR Human Control DNA (QIAGEN) before experimental use. PCR products from these reactions were run on a 1% agarose gel to verify the expected amplicon size (Figures 4-10). The specificity of primers in terms of the other acrocentric chromosomes was determined by PCR on the DNAs from the hybrid cell lines for the other acrocentric chromosomes (Figures 11-14). After cloning and transformation of individual acrocentric chromosome PCR products, DNA sequences obtained from the clones belonging to each acrocentric chromosome were aligned using Geneious version 7.1.5 (http://www.geneious.com, Kearse et al., 2012) and alleles were determined based on nucleotide differences.

Bisulfite Conversion

For each sample, 1 ug of input DNA was converted using the EpiTect Plus kit (QIAGEN) according to the manufacturer’s protocol. The online software QUantification tool for Methylation Analysis (QUMA; quma.cdb.riken.jp) was used to determine conversion efficiency with a lower limit of 95%. Due to initial
conversion efficiencies below 95%, a modification was made at the end of the bisulfite conversion thermal cycler conditions protocol (Table 3 in manufacturer’s handbook) to increase efficiency by adding a denaturation step at 95°C for 5 minutes and a final incubation step at 60°C for two hours.

**Polymerase Chain Reaction (PCR)**

PCR was performed using PCR SuperMix (Invitrogen), Pyromark PCR kit (QIAGEN), or EPIK Amplification Kit (Bioline). Annealing temperatures were calculated for each primer set using the Integrated DNA Technologies OligoAnalyzer tool (https://www.idtdna.com/calc/analyzer). For each primer set, a temperature gradient was designed around the calculated annealing temperature and used to determine an ideal annealing temperature for each template. In some cases, the ideal annealing temperature was outside of the calculated temperature gradient. The calculated annealing temperature gradients as well as the experimentally determined ideal temperatures are indicated in Tables 1 and 2.

The PCR SuperMix kit was used only for untreated samples. A total of 400 ng of input DNA was included in each reaction. Each reaction also contained 45 ul PCR SuperMix, 0.2 uM of each primer, and water to a total volume of 50 ul. The PCR cycling program begins with an initial activation step at 94°C for 2 minutes; then 35 cycles of 94°C for 15 seconds, annealing for 30 seconds, 72°C for 1 minute; and a final extension of 72°C for 10 minutes.

The PyroMark PCR kit was used for both untreated and treated samples. For untreated samples the amount of input DNA was 200 ng per reaction and for bisulfite-treated samples the amount of input DNA was 10 ng per reaction. Each
reaction also contained 12.5 ul PyroMark PCR Master Mix (2x), 0.2 uM of each primer, and water to a total volume of 25 ul. The PCR cycling program begins with an initial activation step at 95°C for 15 minutes; then 45 cycles of 94°C for 30 seconds, annealing for 30 seconds, 72°C for 30 seconds; and a final extension of 72°C for 10 minutes.

The EPIK Amplification kit was used only for bisulfite-treated samples. A total of 10 ng of input DNA was included in each reaction. Each reaction contained 25 ul EPIK Amplification Mix (2x), 0.4 uM of each primer, and water to a total volume of 50 ul. The PCR cycling program begins with an initial activation step at 95°C for 2 minutes; then 35 cycles of 95°C for 15 seconds, annealing for 15 seconds, 72°C for 30 seconds; and a final extension of 72°C for 10 minutes.

**Cloning and Transformation**

PCR products were cloned with the TOPO TA cloning kit (Invitrogen) using 4 ul of fresh PCR product, 1 ul of salt solution, and 1 ul of TOPO vector and the reaction mixtures were incubated at room temperature for 45 minutes. The transformation was performed using the entire 6 ul of the cloning reaction products mixed with either TOP10 competent cells (Invitrogen) or Zymo 10B Mix & Go Competent Cells (Zymo Research). For TOP10 competent cells (Invitrogen), the cells were then incubated on ice for 30 minutes, followed by heat-shock for 30 seconds at 42°C. After addition of 250 ul S.O.C. medium, the cells were shaken horizontally (200 rpm) at 37°C for one hour. After this incubation period, an additional 100 ul S.O.C. medium was added and 200 ul of cells was spread on each LB Ampicillin plate (2 plates per reaction). For Zymo 10B Mix & Go
Competent Cells (Zymo Research), the cells were then incubated on ice for 10 minutes and 50 ul of cells was spread on each LB Ampicillin plate (2 plates per reaction). A minimum of 8 colonies were picked and DNA was purified from the clones with the QIAprep Spin Mini Prep kit (QIAGEN). Sanger sequencing was performed on products by either ACGT, Inc. or Genewiz.

**Sequence Analysis**

Sequences were analyzed using Geneious version 7.1.5 (http://www.geneious.com, Kearse et al., 2012) and low quality sequences (HQ% < 90%) were excluded. For total genomic DNA, clones were isolated by allele. Analysis of methylation patterns was performed using the online software QUantification tool for Methylation Analysis (QUMA; quma.cdb.riken.jp) as described in Kumaki and Okano (2008). Clone sequences with less than 90% percent identity to the unconverted reference sequence and/or conversion efficiency less than 95% were excluded from further analysis. The strict CpG site check of bisulfite sequence for repetitive sequence analysis parameter was chosen in order to account for potential amplification of similar sequences originating from the other acrocentric chromosomes by taking into account polymorphisms at CpG sites. Without adding this parameter, a polymorphic CpG site is automatically considered to be unmethylated regardless of the underlying nucleotides. When using this parameter, if any of the sequences lacked an expected CpG site as determined from the reference sequence or contained an additional CpG site that was perhaps lost from the reference, the program removed that polymorphic CpG data point from the analysis.
### HC21p

<table>
<thead>
<tr>
<th>Locus</th>
<th>Family</th>
<th>BAC Accession Number</th>
<th>L1 Position within BAC</th>
<th>Amplicon Location within BAC</th>
<th>Chromosome Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L1PA3</td>
<td>AF254982</td>
<td>68891-74913*</td>
<td>74553-75314</td>
<td>chr21:10,766,787-10,772,809</td>
</tr>
<tr>
<td>2</td>
<td>L1PA4</td>
<td>AF254982</td>
<td>177341-183471*</td>
<td>182885-183568</td>
<td>chr21:10,875,237-10,881,367</td>
</tr>
<tr>
<td>3</td>
<td>L1PA3</td>
<td>AF254982</td>
<td>38755-44808*</td>
<td>44550-44874</td>
<td>chr21:10,736,651-10,742,704</td>
</tr>
<tr>
<td>4</td>
<td>L1PA3</td>
<td>CT476838</td>
<td>90359-96518*</td>
<td>95869-96695</td>
<td>chr21:7,280,238-7,286,391</td>
</tr>
</tbody>
</table>

### HC21q

<table>
<thead>
<tr>
<th>Locus</th>
<th>Family</th>
<th>BAC Accession Number</th>
<th>L1 Position within BAC</th>
<th>Amplicon Location within BAC</th>
<th>Chromosome Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long 1</td>
<td>L1PA2</td>
<td>AL163207</td>
<td>268670-274700</td>
<td>268604-269018</td>
<td>chr21:16,299,404-16,305,434</td>
</tr>
<tr>
<td>Long 2</td>
<td>L1PA2</td>
<td>AP001694</td>
<td>325412-331436</td>
<td>257200-257718</td>
<td>chr21:25,918,112-25,918,617</td>
</tr>
</tbody>
</table>

### Controls

<table>
<thead>
<tr>
<th>Locus</th>
<th>Family</th>
<th>BAC Accession Number</th>
<th>L1 Position within BAC</th>
<th>Amplicon Location within BAC</th>
<th>Chromosome Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRE3</td>
<td>L1Hs</td>
<td>AC217417</td>
<td>25167-31184</td>
<td>24946-25396</td>
<td>Not Present in Reference Sequence</td>
</tr>
</tbody>
</table>

Table 1. Subfamily and Positions of L1 Loci. The BACs from which the primers were designed as well as the positions of the full-length L1 and the amplicon are indicated. The chromosome position within the reference genome is listed. *L1 is in the reverse orientation relative to the plus (+) strand. Full-length L1 in hg38 reference sequence.
### Table 2. Primer Sequences for Each Locus

The name of the primer specific to the unconverted (_uncon) or converted (_con) genomic sequence as well as the primer sequence (5’ to 3’) is listed.

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
<th>Locus 3</th>
<th>Locus 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1PF9.Con</td>
<td>GTAGGATTTTTTGAGTTAGGTGTG</td>
<td>L1PF10.Uncon</td>
<td>CTACCTGATCGGATTCGTCAGCT</td>
</tr>
<tr>
<td>AP254.R5.Con</td>
<td>AAAACCCTTTATTTGGCTATGCCC</td>
<td>L1PF10.Con</td>
<td>CTTACTGATCGGATTCGTCAGCT</td>
</tr>
<tr>
<td>LIP.N1.Con</td>
<td>GAGCCAGGTGTGGGATATAGTCTCATGGTGCT</td>
<td>LIP.N3.Uncon</td>
<td>CTTACTGATCGGATTCGTCAGCT</td>
</tr>
<tr>
<td>AF254.N1.Uncon</td>
<td>CTCTTTGTGATGAATACATTCCTCACAC</td>
<td>LIP.N5.Uncon</td>
<td>CTCTTTGTGATGAATACATTCCTCACAC</td>
</tr>
<tr>
<td>L1PF13.Con</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>L1PF14.Uncon</td>
<td>CTCTTTGGAATGTGTGCAATTCTC</td>
</tr>
<tr>
<td>L1PF14q.Uncon</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>L1PF12.Uncon</td>
<td>TGTTCTGGGAGAACCACTGTTCTC</td>
</tr>
<tr>
<td>L1PF14q.Con</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>L1PF12.Con</td>
<td>TGTTCTGGGAGAACCACTGTTCTC</td>
</tr>
<tr>
<td>L1PF8q.Uncon</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>CT476.R1.Uncon</td>
<td>CTCTTTGGGAATGTTGCAATTCTC</td>
</tr>
<tr>
<td>L1PF8q.Con</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>CT476.R1.Con</td>
<td>CTCTTTGGGAATGTTGCAATTCTC</td>
</tr>
<tr>
<td>L1PF15.Uncon</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>CT476.N1.Uncon</td>
<td>CTCTTTGGGAATGTTGCAATTCTC</td>
</tr>
<tr>
<td>L1PF15.Con</td>
<td>CTGCTGCCCTTTTATTTGGCTATGCCC</td>
<td>CT476.N1.Con</td>
<td>CTCTTTGGGAATGTTGCAATTCTC</td>
</tr>
</tbody>
</table>

**Table 2.** Primer Sequences for Each Locus. The name of the primer specific to the unconverted (_uncon) or converted (_con) genomic sequence as well as the primer sequence (5’ to 3’) is listed.
Table 3. Primer Information for Outside PCRs. L1 loci and their corresponding primer names used for first round PCRs, as well as temperature gradients used to determine ideal annealing temperatures and the length of the resulting amplicons are indicated. The number of CpGs within each amplicon is indicated in last column and is specified by the number of CpGs within the L1 promoter/total number of CpGs in the amplicon. The amplicon length and number of CpGs for LRE3 reflects the expected sequence for which primers for designed and does not account for the heterogeneity observed in the data for this locus.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Outside Forward (Unconverted/Converted)</th>
<th>Outside Reverse (Unconverted/Converted)</th>
<th>Outside Gradient (°C) (Unconverted/Converted)</th>
<th>Ideal Temperature (°C) (Unconverted/Converted)</th>
<th>Outside Length (bps)</th>
<th># of CpGs within L1/Total CpGs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>HC21p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long arm 1</td>
<td>L1PF14q.Uncon/Con</td>
<td>AL163.R1q.Uncon/Con</td>
<td>58-63/44-49</td>
<td>63/48</td>
<td>415</td>
<td>13/14</td>
</tr>
<tr>
<td>Long arm 2</td>
<td>L1PF15q.Uncon/Con</td>
<td>AP001.R1q.Uncon/Con</td>
<td>54-59/50-55</td>
<td>59/53</td>
<td>519</td>
<td>16/18</td>
</tr>
<tr>
<td>LRE3</td>
<td>LRE3.F.Uncon/Con</td>
<td>LRE3.R.Uncon/Con</td>
<td>55-60/46-51</td>
<td>61/53</td>
<td>451</td>
<td>16/16</td>
</tr>
</tbody>
</table>
Table 4. Primer Information for Nested PCRs. L1 loci and the corresponding primer names used in the second round of PCRs, as well as temperature gradients used to determine ideal annealing temperatures and the length of the resulting amplicon. The length of the LRE3 amplicon reflects the expected sequence for which primers for designed and does not account for the heterogeneity observed in the data for this locus.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Nested Forward (Unconverted/Converted)</th>
<th>Nested Reverse (Unconverted/Converted)</th>
<th>Nested Gradient (°C) (Unconverted/Converted)</th>
<th>Ideal Temperature (°C) (Unconverted/Converted)</th>
<th>Nested Length (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L1P.N1_Uncon/Con</td>
<td>AF254.N1_Uncon/Con</td>
<td>57-62/54-59</td>
<td>57/54</td>
<td>681</td>
</tr>
<tr>
<td>2</td>
<td>AF254.N3_Uncon/Con</td>
<td>L1P.N2_Uncon/Con</td>
<td>62-67/53-58</td>
<td>63/56</td>
<td>649</td>
</tr>
<tr>
<td>3</td>
<td>L1P.N5_Uncon/Con</td>
<td>AF254.N4_Uncon/Con</td>
<td>56-61/46-51</td>
<td>60/57</td>
<td>298</td>
</tr>
<tr>
<td>4</td>
<td>L1P.N4_Uncon/Con</td>
<td>CT476.N1_Uncon/Con</td>
<td>54-59/46-51</td>
<td>60/48</td>
<td>655</td>
</tr>
<tr>
<td></td>
<td><strong>Long arm 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L1P.N6_Uncon/Con</td>
<td>AL163.N1_Uncon/Con</td>
<td>56-61/42-47</td>
<td>57/49</td>
<td>268</td>
</tr>
<tr>
<td></td>
<td><strong>Long arm 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L1P.N7_Uncon/Con</td>
<td>AP001.N1_Uncon/Con</td>
<td>53-58/48-53</td>
<td>58/55</td>
<td>506</td>
</tr>
<tr>
<td></td>
<td><strong>LRE3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LRE3.NeF_Uncon/Con</td>
<td>LRE3.NeR_Uncon/Con</td>
<td>59-64/46-51</td>
<td>66/57</td>
<td>361</td>
</tr>
</tbody>
</table>

**HC21p**
Figure 3. Locus 1 PCR on Converted Human Control DNA. The ideal primer annealing temperature for this locus was determined by a temperature gradient using the nested primers (Table 4).

Figure 4. Locus 2 PCR on Converted Human Control DNA. The ideal primer annealing temperature for this locus was determined by a temperature gradient using the nested primers (Table 4).
Figure 5. Locus 3 PCR on Converted Human Control DNA. The ideal primer annealing temperature for this locus was determined by a temperature gradient using the outside primers (Table 3).

Figure 6. Locus 4 PCR on Converted Human Control DNA. The ideal primer annealing temperature for this locus was determined by a temperature gradient using the nested primers (Table 4).
Figure 7. Long Arm Locus 1 PCR on Converted Human Control DNA. A first round of PCR for this locus was performed using the outside primers at one annealing temperature (Out; Table 3). A second round of PCR was performed using the nested primers with a temperature gradient (Nested; Table 4).

Figure 8. Long Arm Locus 2 PCR on Converted Human Control DNA. A first round of PCR for this locus was performed using the outside primers at one annealing temperature (Out; Table 3). A second round of PCR was performed using the nested primers with a temperature gradient in order to determine the ideal annealing temperature (Nested; Table 4).
Figure 9. LRE3 PCR on Converted Human Control DNA. The ideal primer annealing temperature for this locus was determined by a temperature gradient using the nested primers (Table 4). Multiple bands and smearing are present, which demonstrates the polymorphic locations of this locus within the human genome. The band with the expected amplicon size is indicated (arrow).

Figure 10. Locus 1 Acrocentric Chromosome PCR Gel. The specificity of the primers for this locus was determined by performing PCR reactions on the unconverted acrocentric chromosome hybrid cell lines. The nested primers (Table 4) were used in the reactions with 64°C as the annealing temperature.
Figure 11. Locus 2 Acrocentric Chromosome PCR Gel. The specificity of the primers for this locus was determined by performing PCR reactions on the unconverted acrocentric chromosome hybrid cell lines. The nested primers (Table 4) were used in the reactions with 64°C as the annealing temperature.

Figure 12. Locus 3 Acrocentric Chromosome PCR Gel. The specificity of the primers for this locus was determined by performing PCR reactions on the unconverted acrocentric chromosome hybrid cell lines. The outside primers (Table 3) were used in the reactions with 64°C as the annealing temperature.
Figure 13. Locus 4 Acrocentric Chromosome PCR Gel. The specificity of the primers for this locus was determined by performing PCR reactions on the unconverted acrocentric chromosome hybrid cell lines. The nested primers (Table 4) were used in the reactions with 58°C as the annealing temperature.
Analysis of Locus Specificity

In this study, bisulfite sequencing PCR (BSP) was employed to determine the methylation patterns within the promoter regions of the specific L1s. The L1s chosen for this analysis included four elements from HC21p, designated Loci 1, 2, 3, and 4; two elements from HC21q, designated Long Arm Loci 1 and 2; and a potentially active element designated LRE3 (Sassaman et al., 1997). Detailed information regarding each of these L1s is presented in Tables 1-4. The locations of the loci on HC21p are shown in Figure 14.

Primers were designed for these specific loci as described in Materials and Methods. In order to ensure robust amplification of a single amplicon, annealing temperatures were optimized for each individual converted primer set by first performing PCRs on a bisulfite converted control DNA sample. Once this step was completed, methylation data were collected from experimental cell lines. The first cell line used in this study was WAV17, which is a human chromosome 21 hybrid cell line that contains a single copy of HC21 (Lyle et al., 2007) and theoretically a single allele of each locus. This allows it to serve as a control for any variations in methylation that can occur within a single sample as well as sequencing errors that may be introduced during a PCR reaction. However, since this is a hybrid cell line and does not reflect HC21 in its natural environment, it
may not be a completely accurate representation of methylation in vivo. Methylation data for all loci were also collected from diploid DNA from normal human peripheral leukocytes. Additionally, some loci were analyzed in genomic DNA from a normal prostate cell line and two prostate cancer cell lines described in Materials and Methods.

An important consideration for this study was the high sequence similarity among the short arms of all of the acrocentric chromosomes. Furthermore, the acrocentric chromosome short arms have not yet been completely sequenced. Thus, data sets obtained from total genomic DNA can potentially contain alleles originating from sequences on any of the acrocentric chromosomes. To investigate this possibility, primer sets designed for the four L1 elements on HC21p were tested on unconverted hybrid cell line DNAs containing each of the other acrocentric chromosomes. Locus 1 was also found on HC13 (Figure 10), which is not surprising because out of all the acrocentric chromosomes, HC13 is the most similar to HC21 (Vissel & Choo, 1991). The sequences obtained from the PCRs performed on the acrocentric chromosome cell lines showed that for Locus 1 only a single nucleotide polymorphism (SNP) distinguished HC21 from HC13. In addition to being on HC13, Locus 2 was also found to be present on HC15 (Figure 11). Similar to the results seen for Locus 1, Locus 2 on HC13 could be distinguished from HC21 by only one SNP, whereas this locus on HC15 has only 95% identity to HC21. Furthermore, two alleles with 96.9% identity to each other were present in the sequence pool obtained from the HC15 hybrid cell line. Therefore, two separate L1 loci present on HC15 can potentially be amplified by this primer set. Locus 3 and Locus 4 were determined
to be present on all acrocentric chromosomes (Figures 12 and 13). Due to this complexity, the sequences of these loci were not determined for the other acrocentric chromosomes. Thus, data sets collected from genomic DNA were analyzed with the knowledge that alleles originating from either copy of any of the acrocentric chromosomes were potentially present in the sequence data set. This possibility must also be considered for Loci 1 and 2. While the hybrid cell line experiments predicted the presence of these loci on only some acrocentric chromosomes, this information still pertains only to what is seen on the one chromosome copy in the hybrid cell line. Therefore, it cannot be ruled out that additional loci are present in the various diploid tissue samples, which could not be detected in the hybrid cell lines. However, as described below, the number of alleles detected for any given locus appears to be limited.

Due to the complexity inherent in repetitive DNA analysis, it was necessary to develop a method that could effectively sort sequences after methylation analysis. The repetitive sequences version of the QUMA program was used in this study, which is an excellent tool for analyzing initial methylation data. (described in Materials and Methods). When a CpG present in the reference sequence is absent in a converted DNA sequence, this version of QUMA will ignore this polymorphism whereas the regular version of QUMA will interpret any polymorphism as an unmethylated CpG site. However, this program is not designed for allelic analysis and further steps must be taken using the QUMA output. In order to sort sequences into groups that potentially originated from the same locus, percent identity of converted sequences to the unconverted WAV17 reference sequence was first evaluated. Percent identity
alone is not sufficient to group sequences because loci originating from separate alleles of acrocentric chromosomes can be very similar in sequence. Furthermore, sequences lacking any CpG sites that are present in the reference give a strong indication that it originated from a separate allele. In most cases, a missing CpG site did not drastically alter the percent identity of the sequence to the reference. Sequences with similar CpG methylation pattern haplotypes may indicate that they originated from the same allele, so methylation pattern haplotypes were then used to further classify sequences into subgroups. Any clone displaying a different pattern of methylated and unmethylated CpGs across the sequenced region was considered a separate haplotype. However, methylation patterns at a single locus can vary within a single sample, so differences in methylation haplotypes do not necessarily indicate that the sequences originated from separate alleles. Therefore, the most robust analysis resulted from the combination of percent identity to the WAV17 reference sequence with methylation pattern haplotypes.

It must be noted that this method was performed manually and has limitations. For instance, bisulfite treatment results in the creation of extensive polypyrimidine tracts that causes DNA polymerase to stutter, which produces errors during PCR. These will be incorrectly counted as mismatches in the QUMA program, thereby lowering percent identity values and skewing the analysis. Additionally, bisulfite converted DNA templates are difficult to sequence by Sanger sequencing, resulting in lower quality sequencing reads compared to unconverted templates. To account for potential sorting errors, parameters were set using the variations in percent identity and number of
haplotypes observed within the WAV17 control samples. The WAV17 control data across all loci demonstrated that a locus within this single sample can have percent identity as low as 97% to the unconverted WAV17 reference sequence. Thus, converted sequences with percent identity ranging from 97%-100% were considered to have likely originated from HC21, although in some cases they would be indistinguishable from HC13. Sequences with less than 97% percent identity to the WAV17 reference sequence were considered to originate from other acrocentric chromosomes and could be further divided into subgroups based on their sequence similarity to each other. Furthermore, the WAV17 control data demonstrated that a single locus within one sample can have as many as 8 different methylation haplotypes. Some of these methylation patterns were very similar, varying at only one CpG position. Thus, converted sequences that were very similar to each other but containing different methylation pattern haplotypes could have potentially originated from the same allele. The small number of sequence alleles found for a given locus suggested that each locus has only a few highly similar copies.

**Methylation Analysis in Hybrid Cell Lines and Normal Peripheral Leukocytes**

**Locus 1**

As mentioned previously, Locus 1 was also present on HC13 where it differed from HC21 by only one SNP. At this position, HC21 was identified by a cytosine whereas HC13 contained a thymine. After bisulfite treatment, the cytosine present on HC21 will be a thymine. Thus, it was impossible to identify from which chromosome the sequences originated in converted genomic DNA samples. When all CpGs in this locus were included in the methylation analysis
of the converted HC21 control hybrid cell line WAV17, the overall percent methylation was 78.9% (Table 5). Due to the complete lack of methylation of CpGs in the adjacent sequence, overall percent methylation was 89.6% when only the CpGs within the L1 promoter were included (Table 5). All of the converted clones had greater than 99% percent identity to the WAV17 reference sequence, indicating that they may represent one allele. However, three different methylation haplotypes were present, demonstrating the heterogeneity that can occur within one sample (Figures 15 and 16). While Haplotypes 2 and 3 were methylated at all of the critical CpG sites within the L1 promoter, Haplotype 1 was unmethylated at two of these sites (+61 and +70). Haplotypes 1 and 2 were unmethylated at position +269, whereas Haplotype 3 was methylated at this site. Haplotype 2 contained one additional unmethylated CpG site at +285 that was not seen in the other methylation haplotypes.

Leukocytes had 74.4% overall methylation when the adjacent CpGs were included in the analysis (Table 5). Since the adjacent L1s were hypomethylated in this cell line, this percentage was 82.9% when only CpGs within the L1 promoter were included (Table 5). All of the converted sequences displayed high similarity to the reference sequence (>98%), although the CpG at position +171 in the sequence obtained from WAV17 was absent from all leukocyte clones (Figures 17 and 18). This does not necessarily indicate that all of the clones did not originate from HC21 as there may simply be a polymorphism at this position in the genomic leukocyte sample. All ten clones obtained from leukocytes were too similar to separate into alleles by only comparing percent identity to WAV17. Seven different methylation haplotypes and varying levels of methylation were
observed among the 10 clones (Figure 17). Four of these methylation haplotypes (1, 2, 3, and 5) each had only one representative clone sequence and overall lower percent methylation (<75%) than the rest of the clone sequences. The remaining methylation haplotypes were each represented twice and had overall higher levels of methylation (between 75% and 88%). Thus, while HC13 and HC21 are indistinguishable in terms of underlying sequence, it is possible that some or all of the alleles from each chromosome are present in the data. Overall, Locus 1 is similarly methylated in WAV17 and leukocytes.

**Locus 2**

Locus 2 was also present on HC13 in addition to HC15. As with Locus 1, HC13 was indistinguishable from HC21 in converted DNA samples due to a C/T SNP. This locus in WAV17 showed 56.7% overall methylation (Table 5). All of the clones had greater than 97% identity to the WAV17 reference. Eight different methylation haplotypes were present in the WAV17 group, six of which were represented only once (Figures 19 and 20). Furthermore, the total percent methylation of each clone had a large range of 16.7%-75%, again indicating the heterogeneity that can occur within a single sample.

Locus 2 in leukocytes had 73.4% methylation, which was higher than what was seen in WAV17 (Table 5). The sequences derived from leukocytes were split into two groups based on percent identity to the WAV17 reference, although there were several different methylation haplotypes present in the data set (Figure 21 and 22). Group A had lower sequence identity (~96%) to the WAV17 reference and was missing a CpG at position +341. The lower sequence identity that is observed in Group A cannot be explained by the heterogeneity that can
occur at a single locus. This indicates that sequences belonging to Group A likely do not originate from HC21. In order to investigate the possibility that Group A represented sequences on HC15, an additional QUMA analysis was performed using the consensus HC15 sequence derived from the initial acrocentric chromosome experiments as the reference. The converted clones had a much lower percent identity (~92%) to the HC15 reference sequence than would be expected if they originated from HC15. Three methylation haplotypes were present in the 4 clones belonging to Group A with a span of 63.6%-72.7% overall percent methylation (Figure 21). Thus, even though the percent identity of Group A to the reference sequence is lower than 97%, the even lower percent identity to HC15 indicates that the sequences in this group most likely originate from a highly similar allele on HC21, HC13, or even perhaps another acrocentric chromosome. Group B had an identity above 99%, which indicates that the sequences could have originated from either HC13 or HC21. Each of the 8 clones present in Group B displayed a different methylation haplotype and much of the heterogeneity between methylation haplotypes occurred at CpGs further downstream within the L1 promoter (Figure 21). It is possible that the sequences in Group B came from the same allele of Locus 2 and the variable methylation haplotypes are due to the heterogeneity that can occur at the same locus. However, it cannot be determined if the sequences originate from HC13, HC21, or both. Thus, based on the sequences in Group A as well as the several methylation haplotypes in Group B, Locus 2 is more heterogeneous than predicted perhaps indicating that it is also present on the other acrocentric chromosomes and in more than one copy.
Locus 3

Locus 3 was found to have 80.0% overall methylation in the WAV17 hybrid cell line; this value was 78.8% when the single adjacent CpG was removed from the analysis (Table 5). This sample was more homogeneous than was seen for the first two loci, with high percent identity to the WAV17 reference (>99.4%) and 9 out of the 11 clones displaying only one methylation haplotype (Haplotype 2) (Figures 23 and 24).

Locus 3 was less methylated in leukocytes with 46.2% total methylation and 42.6% total methylation when the adjacent CpG was excluded (Table 5). It was possible to delineate two groups of sequences in this data set based on percent identity to the WAV17 reference sequence (Figures 25 and 26). Group A had less than 93% identity to WAV17, indicating that these sequences likely originated from another acrocentric chromosome. In this group, three of the CpGs within the L1 promoter (+49, +58, and +70), as well as the CpG within the adjacent sequence were absent (Figure 25). Furthermore, there was only one methylation pattern haplotype in Group A (Haplotype 1), which had 33% total methylation (Figure 25). The similarity between sequences in Group A (>99%) and the presence of the single methylation haplotype indicates that they originated from the same allele. Group B had high percent identity to WAV17 (>99%) and had two distinct methylation haplotypes (Figure 25). Haplotype 2 displayed 30% total methylation while haplotype 3 displayed 80% total methylation. The percent identity to WAV17 suggests that Group B originates from HC21 and the large difference between the methylation haplotypes indicate that each haplotype may represent a different allele. Without reference sequences
for this locus on the other acrocentric chromosomes, it is not possible to
determine at this point if any of the sequences in this group represent this locus
on the other acrocentric chromosomes. For instance, Locus 3 on HC13 may be
present in the data set if its sequence is indistinguishable from the sequence on
HC21 as observed for Loci 1 and 2.

Locus 4

Locus 4 was 52.4% methylated in WAV17 (Table 5) and all of the
sequences had greater than 99.8% identity to the WAV17 reference. Six different
methylation haplotypes were present (Figures 27 and 28). Both Haplotypes 1 and
2 had 43.9% total methylation, while the remaining methylation haplotypes
displayed 57%-72% total methylation.

Locus 4 was 69.4% methylated in leukocytes, which was similar to
WAV17 (Table 5). The data were divided into two groups based on percent
identity. Group A consisted of two clones that had greater than 98% sequence
identity to the WAV17 reference, and one distinct methylation pattern haplotype
(Haplotype 1) was present in the two clones with 50% total methylation (Figures
29 and 30). Initially, these data indicate that this group may have originated from
HC21. However, three of the expected CpG sites for this locus were absent,
which implies that these sequences originated from another locus. Group B had
higher percent identity to WAV17 (~99%) and five distinct methylation pattern
haplotypes within this group of 11 clones were identified. Haplotypes 5 and 6
were more frequently represented than the others making up 5/11 clones and
3/11 clones, respectively (Figure 29). Haplotype 5 displayed 71.4% methylation,
while as Haplotype 6 was 85.7% methylated. Each of the remaining three
methylation haplotypes displayed 57% total methylation and each had only one representative clone. Group B likely originated from HC21, although it cannot be discounted that it was from the locus on another acrocentric chromosome.

Long Arm Locus 1

Long Arm Locus 1 was almost fully methylated across all clones in WAV17 with 99.2% methylation and had greater than 99.6% identity to the WAV17 reference sequence (Table 5). Three methylation haplotypes were present in the WAV17 sample; Haplotypes 1 and 2 each had only one representative clone, whereas Haplotype 3 made up the majority with 7 clones (Figures 31 and 32). While all of the clones displaying Haplotype 3 were completely methylated, the clone belonging to Haplotype 1 contained one unmethylated CpG at +35. Haplotype 2 is missing a CpG at +155, but this is likely a sequencing artifact as this sequence is otherwise identical to the WAV17 reference.

Long Arm Locus 1 had 97% total methylation in leukocytes which was similar to WAV17 (Table 5). Three methylation haplotypes were present (Figures 33 and 34). Haplotype 3 had greater than 99% identity to the WAV17 reference and was methylated at every CpG site. The two clones that represented Haplotype 2 also had high percent identity to the WAV17 reference (>99%) and were completely methylated at the CpGs within the L1 promoter. However, unlike Haplotype 3, the CpG in the region adjacent to the L1 was unmethylated. Haplotype 1 was represented only once and had lower percent identity to the WAV17 reference (96%), with the CpG at +160 absent from the sequence. This
methylation haplotype had 76.9% total methylation, which was much lower than what was seen for the other methylation haplotypes. Due to the high percent identity to WAV17 seen in Haplotypes 2 and 3, these sequences likely originated from either allele present on HC21. Long Arm Locus 1 is a unique sequence on HC21q, so it is unlikely that another locus would be amplified by this primer set. However, the variations in percent identity and methylation haplotype cannot be explained by heterogeneity at a single locus that can occur within a sample. Thus, the clone belonging to Haplotype 1 most likely originated from an unexpected locus.

*Long Arm Locus 2*

Long Arm Locus 2 was also highly methylated in WAV17 with 99.1% total methylation (Table 5). Two haplotypes were present in the WAV17 sample, both with greater than 99.6% sequence similarity to the WAV17 reference. Haplotype 1 differed from Haplotype 2 by only one unmethylated CpG at position +204 (Figures 35 and 36).

This locus had a 98.8% methylation in leukocytes (Table 5). The three methylation haplotypes within the leukocyte sample displayed high percent identity to the WAV17 reference (>99%). The majority of the sample consisted of Haplotype 3, which was completely methylated (Figure 37 and 38). Haplotypes 1 and 2 each had one representative clone and each contained only one unmethylated CpG within the L1 promoter region (Figure 37). All three haplotypes most likely originated from the expected locus on HC21q.

*Future Analysis of HC21q*

The ENCODE project data available on the UCSC genome browser was
used to determine if any data were collected in the regions of HC21q containing Long Arm Loci 1 and 2 and if they would correspond to the results for WAV17 and leukocytes. There was not any significant ENCODE methylation data in these regions, which reflects the general challenges related to repetitive sequence analyses.

Future work will analyze two additional full length L1 loci on HC21q that belong to the L1PA3 and L1PA4 families and so correspond to the L1 families analyzed on HC21p. Using RepeatMasker, full-length L1s in the hg38 reference genome were filtered to isolate candidates by the desired family as well as those located on HC21q. After applying these filters, one L1 element from each subfamily that met these criteria was chosen as a suitable candidate for future investigation (L1PA3, chr21:14166720-1417287; L1PA4, chr21:13131779-13137920). Similar to the observation for Long Arm Loci 1 and 2, there was not any informative ENCODE methylation data for either of these loci.

LRE3

The control locus chosen for this study was LRE3, which is a retrotransposition-competent L1Hs located on chromosome 2 that is polymorphic and not present in the reference genome (Brouha et al., 2002). The authors of the paper listed a HC2 BAC from which they designed primers but when using the Primer-BLAST tool on NCBI, the primers were not found on this BAC, but on a different HC2 BAC with an L1Hs at that position (Accession #AC217417). Thus, primers were designed from this BAC and it is unclear how the authors designed their primers. Amplification of LRE3 produced unexpected results. As shown in Figure 9, the primers designed for this locus produced
bands and smearing in the human control DNA sample that were larger than the expected amplicon size of 361 base pairs, indicating that the primer set was not specific. PCR on DNA from the HC2 hybrid cell line produced an unexpected amplicon larger than LRE3. However, PCR on DNA from leukocytes produced the expected amplicon of 361 bps with exact sequence identity to the L1Hs on the HC2 BAC. Furthermore, the unexpected sequence obtained from the HC2 hybrid cell line DNA has 9 CpG sites, which is different than the 14 CpG sites present in the expected LRE3 sequence in leukocytes. It is likely that the copy of HC2 in the hybrid cell line does not contain the LRE3 locus as seen in the reference genome whereas LRE3 is present in the leukocyte sample. By aligning the sequence obtained from the HC2 hybrid cell line DNA to the hg38 reference genome, it was determined that while the L1 is located on HC2, it is not an L1Hs and belongs to the L1PA2 subfamily.

LRE3 was 95.6% methylated in HC2 (Table 5) and had only two methylation haplotypes with greater than 99.5% sequence identity to the HC2 reference sequence. Haplotype 1 differed from Haplotype 2 by only a single unmethylated CpG site at position +19 (Figure 39).

LRE3 was 97.9% methylated in leukocytes and the sequences had greater than 98% sequence identity to the leukocyte reference sequence (Table 5). The high percent identity of the sequences to each other indicates that they all originate from the same locus, but could still be divided into four methylation haplotypes (Figure 40). All four methylation haplotypes had high levels of methylation. Haplotype 1 had the lowest percent methylation at 86% and Haplotype 2 and 3 had 92.9% methylation. Haplotype 4 made up the majority of
sequences and was completely methylated at every CpG site.

*Methylation Comparisons Among Loci*

Overall, the loci located on HC21p have substantially lower overall methylation levels (42.6%-89.6%) than the control loci on HC21q and LRE3 (95.6%-99.2%; Table 5). The methylation levels did differ between WAV17 and leukocytes for some loci on HC21p. However, there do not appear to be any obvious trends in the methylation differences between WAV17 and leukocytes. For instance, Loci 1 and 3 have lower levels of percent methylation in the leukocyte sample compared to WAV17, whereas Loci 2 and 4 have higher levels in leukocytes compared to WAV17. These patterns cannot be explained by the family to which each locus belongs since Loci 1, 2, and 4 all belong to the L1PA3 family. In leukocytes, the L1PA3 loci had higher levels of methylation than Locus 3, which is the only L1PA4 locus. However, this trend was not seen in WAV17 since Locus 1 methylation was more similar to Locus 3 than to the other L1PA3 loci. These differences could reflect the fact that in leukocytes the loci are in a different genomic context than in the hybrid cell line. The control loci did not vary greatly from each other, with values of methylation all greater than 97%. The “LRE3” sequence in HC2 had slightly lower levels at 95.6%, but this was a different amplicon than expected.

*Analysis of Critical CpGs*

There are four critical CpG sites in the L1 promoter known to be important for the regulation of L1 expression by methylation (Hata and Sakaki, 1997). For each locus, the average percent methylation of the critical CpG sites was used to compare the loci to each other and by cell line. The average critical
CpG site methylation in Loci 1, 3, and 4 was higher in WAV17 (92.6%, 100%, 100%, respectively) than in leukocytes (86.7%, 78.2%, 36.4%, respectively; Table 6). Locus 2 contains only one of the critical CpG sites (+58) and the percent methylation at this site was consistently low in WAV17 and leukocytes (70% and 83.3%, respectively; Table 6). The control loci were completely methylated at each of these CpG sites in both WAV17 and leukocytes (Table 6). It should be noted that not all CpG sites were present in each locus and at this point it is unclear whether the absence of one of these 4 critical CpG sites has a similar effect as if it is unmethylated. However, the study by Hata and Sakaki (1997) indicated that all four of these CpG sites must be methylated to inhibit transcription. This is the case for the control loci, but not the HC21p loci in leukocytes nor Loci 1 and 2 in WAV17. All four of the critical CpG sites were present in the LRE3 amplicon in leukocytes and were 100% methylated. The unexpected amplicon seen in the HC2 hybrid cell line contained three of the four critical CpG sites and these sites were also 100% methylated. Given that this sequence most likely belongs to an L1PA2 element as determined from the BLAT results, the data from HC2 supplement what was seen for Long Arm Loci 1 and 2, which are also from the L1PA2 subfamily.

**Methylation Analysis in Prostate Cell Lines**

**Locus 1**

Locus 1 in RWPE-1 had overall 35.6% methylation when all CpGs were included and 41.9% methylation when the two adjacent CpGs were removed from the analysis (Table 5). Two distinct groups of sequences were isolated based on two particular CpGs at positions +171 and +251: both of these CpGs were
absent in Group A but present in Group B (Figures 41 and 42). As mentioned above, position +171 was also absent from all clones in leukocytes, but this is the first cell line where a polymorphism was also present at the CpG at position +251. The percent identity of Group A to the WAV17 reference sequence ranged from 97.7-98.2%, which was slightly lower than the range of percent identity to WAV17 seen in Group B (98.1%-99.4%). Group A consisted of one single methylation haplotype that had 6.7% total methylation, while Group B contained 7 different methylation haplotypes with a range 47.1%-76.5% total methylation. The percent identity of both groups fell within the range set to account for variability at a single locus, which suggested that both groups potentially originated from HC21. However, when taking into account the decreased range of percent identity to WAV17, the CpG polymorphisms, and the differences in total percent methylation, it is possible that Group A originated from a different allele than Group B and perhaps even a different locus. It should also be noted that sequences from Group A and Group B were obtained from different batches of bisulfite-converted DNA from RWPE-1, so it is possible that variations between batches resulted in the biased amplification of separate loci.

Locus 1 in LNCaP had 16.8% total methylation when all CpGs were included and 22.0% methylation when the two adjacent CpGs were excluded from the analysis (Table 5). All clones had between 97.7%-98.2% sequence identity to the WAV17 reference sequence and the CpG at position +171 was absent. The sequences were too similar to each other to be separated into alleles based on percent identity alone. However, a second CpG site at position +251 allowed for the separation of the sequences into two groups due to its presence
(Group A) or absence (Group B) (Figures 43 and 44). Group A contained 3 different methylation haplotypes displaying various levels of percent methylation. Haplotype 1 had 75% total methylation; this was much higher than Haplotype 2, which had 12.5% total methylation and Haplotype 3, which was completely unmethylated. Thus, Haplotype 1 most likely represents a separate allele than Haplotypes 2 and 3. Group B also contained 3 different methylation haplotypes; Haplotype 4 had the highest level of methylation within this group with 26.7% total methylation. Haplotype 5 contained only one methylated CpG whereas Haplotype 6, which comprised half of the sequences obtained from LNCaP, was completely unmethylated. Thus, Haplotypes 5 and 6 may represent one allele, while Haplotype 4 represents another allele. The polymorphism at the CpG located at +251 was only observed in RWPE-1 and LNCaP. Group A in RWPE-1 and Group B in LNCaP lack this CpG and had high sequence similarity (>99%), indicating that they represent the same locus. All of the sequences in LNCaP fall within the indicated range of percent identity to WAV17 to assign them to a locus on HC21. The CpG site polymorphisms present in this sample do not rule out this possibility. In particular, it may be possible that the CpG at +171 in WAV17 is more likely to be absent at the HC21 locus in a genomic sample. The absence of the CpG at +251 in some sequences could potentially indicate another highly polymorphic site at the HC21 locus. It is also possible that these CpG sites are characteristic of a locus on another acrocentric chromosome, although this cannot be determined definitively due to the high sequence similarity among sequences.

Locus 1 in DU145 was 89.5% methylated within the L1 promoter and
81.2% methylated when adjacent CpGs were included in the analysis (Table 5). The sequences in this sample were highly similar to the WAV17 reference sequence (>97.7%) and could not be distinguished from each other based on percent identity. The CpG at site +171 was absent in all sequences. While the methylation status at the majority of CpGs across Locus 1 was consistent across all clones, variations in methylation at two particular CpG sites were used to separate the sequences into four different haplotypes (Figures 45 and 46). One of these CpGs was in the sequence adjacent to the L1 a position -301, while the other was within the L1 promoter at one of the critical CpG sites, +52. Haplotype 1, which comprised 6/15 sequences, was methylated at position -301 and unmethylated at position +52. Haplotype 4 had seven representative sequences and displayed the opposite pattern, with the CpG at -301 unmethylated and the CpG at +52 methylated. The clone belonging to Haplotype 3 was the only sequence with an unmethylated CpG at +231, but was otherwise identical to the methylation pattern of Haplotype 4, indicating that they most likely represent the same allele. Haplotype 2 was represented by one clone and was methylated at both positions. Based on sequence identity to the WAV17 reference and the lack of polymorphisms at the +171 and +251 CpG sites, all of the sequences may have originated from the same locus on HC21. The differences in methylation at the two specific CpG sites suggest that there are two alleles of this locus represented in the population. Of course it is still not possible to definitively reject the possibility that these sequences represent a locus on another acrocentric chromosome.

Overall, Locus 1 in RWPE-1 and LNCaP were less methylated than in
WAV17 and leukocytes. This locus was similarly methylated in DU145 compared to WAV17 and leukocytes.

Locus 2

Locus 2 had 43.9% total methylation in RWPE-1 (Table 5). Two groups of sequences were isolated based on percent identity to the WAV17 reference sequence (Figures 47 and 48). Group A had less than 97% sequence identity and six methylation haplotypes. The most obvious difference in Group A was variable presence of reference CpG sites. Haplotypes 1-4 lacked the CpG sites at +115 and +341. Haplotype 5 did not have the reference CpG sites at +160 and +341, while Haplotype 6 only lacked the reference CpG site at +160. The polymorphic presence of the CpG at position +341 was previously observed in the leukocyte sample. Group B had greater than 99% sequence identity to the WAV17 reference sequence and all of the CpG sites were present. Another six distinct methylation haplotypes were observed with low levels of methylation. Haplotype 7 was completely unmethylated and the rest of the haplotypes were 25%-50% methylated. This is in contrast to Group A, with most of the sequences having between 60-90% methylation (Haplotype 6 was an exception with 45% methylation). Based on these data, it is most probably that Group B originated from alleles of HC21. Sequences in Group A most likely originated from one or more of the other acrocentric chromosomes; Haplotypes 1-4 were most similar to each other and potentially came from an allele of the same chromosome. Haplotypes 5 and 6 may had different alterations in presence of CpG sites, indicating that it either came from the other allele of the first four methylation haplotypes or potentially another chromosome.
Locus 2 in LNCaP was almost entirely unmethylated with 0.7% total methylation (Table 5). Two groups emerged based on their percent identity to the WAV17 reference (Figures 49 and 50). Group A had less than 97% sequence identity and three different methylation haplotypes. Haplotype 1 differed from Haplotype 2 by only a single methylated CpG, but CpG sites at +160 and +341 were absent in both methylation haplotypes. Thus, it is likely that these two originated from the same allele. Group B had greater than 98% sequence identity to the WAV17 reference sequence, indicating that the sequences originated from HC13, HC21, or both. Haplotype 3 in Group A has less than 97% sequence identity to Haplotypes 4 and 5 in Group B, but all of these haplotypes have identical methylation patterns. Thus, even though Haplotype 3 contains all of the CpG sites in the reference sequence and has the same methylation pattern seen in Group B, it likely originated from another acrocentric chromosome.

Locus 2 in DU145 has 48.9% total methylation (Table 5) and was split into two groups based on percent identity to the WAV17 reference sequence: Group A had between 93%-96% sequence identity, whereas Group B had greater than 98% sequence identity (Figures 51 and 52). Furthermore, Group A had on average 55% total methylation which was higher than the 37% total methylation seen in Group B. There were a total of 9 methylation haplotypes in Group A, all of which did not have the reference CpG at position +341. Additionally, Haplotypes 7-9 lacked the reference CpG at position +115. Group B contained an additional 4 methylation haplotypes. The majority of the differences causing the distinction between methylation haplotypes occurred within the first 200 base pairs of the L1 promoter. From these data, Group B potentially originated from
alleles on HC21, while Group A contained sequences belonging to another acrocentric chromosome. While all of the methylation haplotypes observed in Group B could have originated from the same locus, the differences in Haplotypes 1-6 versus Haplotypes 7-9 indicates that they came from different alleles of the same locus or separate loci.

Methylation at Locus 2 in LNCaP was substantially less in LNCaP compared to all other cell lines. This locus was less methylated in RWPE-1 and DU145 compared to WAV17 and leukocytes, but the difference was not as extreme as in LNCaP.

Long Arm Locus 1

Long Arm Locus 1 in RWPE-1 was 93% methylated within the L1 promoter and 93.5% methylated when the adjacent CpG was included (Table 5). The percent identity to the WAV17 reference sequence was greater than 99% for all sequences. Five methylation haplotypes were isolated and had overall high levels of methylation (Figures 53 and 54). Haplotype 1, which had one representative clone, had 64% total methylation whereas the remaining haplotypes had greater than 85% total methylation. Haplotype 5, which contained 6 out the 11 clones, was completely methylated at every CpG site. It is possible that all sequences originated from the same locus on HC21 due to the high sequence identity to WAV17 observed across all clones. The differences in methylation haplotypes could be due to the potential heterogeneity occurring at a single locus or they could represent alleles of the same locus. This cannot be confirmed due to the high sequence identity across all sequences.

Long Arm Locus 1 in LNCaP was 94.9% methylated within the L1
promoter and 92.9% methylated when the adjacent CpG was included (Table 5). All of the sequences had greater than 99% identity to the WAV17 reference. Seven different methylation haplotypes were present in the population (Figures 55 and 56). The total percent methylation ranged from 78.6%-92.9% among Haplotypes 1-6, while Haplotype 7 was completely methylated. The larger number of methylation haplotypes may indicate that both alleles of this locus are represented in the sequences.

Long Arm Locus 1 in DU145 was completely methylated and all sequences had greater than 99% identity to the WAV17 reference (Table 5; Figures 57 and 58). Thus, these sequences likely originated from HC21, but it would be impossible to determine if they represent separate alleles of the same locus.

Overall methylation at Long Arm Locus 1 in RWPE-1 and LNCaP was slightly lower than in WAV17 and leukocytes. However, this locus in DU145 was similarly methylated as in WAV17 and leukocytes.

**LRE3**

The majority of sequences obtained from the unconverted RWPE-1 cell line were LRE3 and matched the leukocyte amplicon. However, three other amplicons of different sizes (307, 345, and 455 bps) were also obtained that were unexpected and not observed in the converted sample. The 307 bp amplicon aligned to HC15q26.1 with 98.1% sequence identity and is classified as an L1PA3. The 345 bp amplicon aligned to 7p12.2 with 99.0% sequence identity and is an L1Hs. The last unexpected amplicon with a length of 455 bp aligned to HC3p22.3 with 98.6% sequence identity and is likely an L1PA4. PCR of LRE3 in converted
RWPE-1 resulted in three different amplicons. Using the expected 361 bp LRE3 sequence as the unconverted input sequence in QUMA, the majority of the amplicons were determined to be LRE3 (Figure 59, Group B), while two sequences were not LRE3 and were also not present in the unconverted RWPE-1 data (Figure 59, Group A). Without a representative sequence in the unconverted samples, it is impossible to determine where they are located in the genome using bioinformatic tools. Thus, the primers designed for LRE3 amplify several other loci in the human genome, some of which are not necessarily observed in the unconverted sample. Haplotype 1 had only 94% sequence identity to LRE3 and was completely unmethylated. Furthermore, four of the CpG sites within the promoter were absent from this amplicon. Haplotype 2 was 66.7% methylated and had only 96% sequence identity to LRE3. Interestingly, this amplicon is the same length as the amplicon in the HC2 hybrid cell line (~411 bps), but the two sequences had less than 96% sequence identity to each other, indicating that they were not the same locus. The sequences in Group B had greater than 99.5% sequence identity to LRE3, indicating that they are in fact LRE3. This group was divided into five different methylation haplotypes. Haplotypes 5 and 6 made up the majority of Group B (7/16 and 6/16, respectively) and were 99.2% methylated. The remaining 3 clones were 85.6% (Haplotypes 3 and 4) and 100% (Haplotype 7) methylated. Thus, LRE3 was much more methylated in RWPE-1 than the two other loci observed in Group A.

The data obtained from LNCaP were much more complex than that seen in the previous samples. Seven different amplicons were present in the unconverted sequence pool with sizes ranges from 306-420 bps and none of these
amplicons were LRE3 nor the locus present in the HC2 hybrid cell line (data not shown). Analysis using the BLAT tool and UCSC genome browser allowed for preliminary mapping of 6 out of the 7 amplicons to the human reference genome, with greater than 97% sequence identity to the locations to which they were mapped. Preliminary data using the RepeatMasker tool embedded in the UCSC browser indicate that the six mapped amplicons are not L1Hs elements and are evolutionarily older (L1PA2-L1PA4). It was not possible to obtain an unconverted reference sequence specific to LNCaP due to the presence of numerous different amplicons, ranging in size from 306-420 base pairs, with less than 94% identity to each other. The analysis was further limited by the low frequency of each amplicon in the sample (only 1-4 clones). Thus, all of the converted sequences were compared to the 361 bp LRE3 reference sequence (Figure 60). Like the unconverted LNCaP sample, neither LRE3 nor the HC2 amplicon were present in the converted LNCaP sequence pool. The sequences obtained from the converted sample had on average 51.2% methylation, ranging from 0%-78.6% total methylation. Five different amplicon sizes were present and sequences were divided into groups of sequences with similar lengths, resulting in 5 different subgroups (Groups A-E, Figure 60). However, sequences that were the same length were not necessarily the same locus, as determined by both methylation haplotypes and percent sequence identity (Groups A, C, D, and E). For instance, Group A had less than 70% sequence identity to LRE3, but the sequences within the group only shared 96% identity to each other. The remaining groups had higher percent identity to LRE3 (> 94%), but the differences in amplicon lengths as well as polymorphisms at CpG sites indicate
that the sequences represent many different L1 loci. Thus, the clones for this sample were not further divided into separate methylation pattern haplotypes because of the complexity of the sequence data. The two clones in Group B are identical in sequence and contain the same CpG sites, indicating that they likely originate from the same locus (blue rectangle, Figure 60). The five clones in Group D are the same length, but only three of them originate from the same locus, based on high sequence identity to each other (>98%, red rectangle, Figure 60).

The sequences obtained from unconverted DU145 were also not LRE3 nor the amplicon on HC2. Two rounds of amplification were performed on separate occasions and interestingly, the first round produced a single amplicon that was 345 bps in length, while the second round resulted in five different amplicons, ranging in size from 344-356 bps. The first amplicon (345 bps) was also seen in LNCaP and was tentatively mapped to chromosome 3 using the BLAT tool in the UCSC genome browser. Two of the remaining 5 amplicons seen in the second round of PCR could also be confidently mapped to the reference genome on 2p24.2 (99.5% sequence identity, L1Hs subfamily) and 15q25.3 (99.2% sequence identity, L1PA2 subfamily), but the three amplicons are from unknown locations. Thus, these L1s may be polymorphic in the population and are not present in the reference genome. The sequences obtained from converted DU145 were similarly as complex as LNCaP and the LRE3 and HC2 amplicons were not present. Since the 345 bp amplicon was observed in the majority of sequences obtained from unconverted DU145, the initial consideration was to use this sequence as the input reference sequence. However, none of the converted sequences were the
345 bp amplicon (<95% sequence identity) and ranged in size from 343-411 bps. Thus, the 361 bp LRE3 sequence was again used as the input reference sequence to maintain consistency. The sequences were divided into two groups based on percent identity (Figure 61). Group A had a range of 94%-96% identity to LRE3, whereas Group B had 68%-71% identity. The total percent methylation was 58.4%, but this is not an informative number as the sequences spanned the whole spectrum from 0%-100% total methylation. The large ranges of amplicon length and percent identity of sequences to each other prevented further division into subgroups. However, it appears that at least 9 different L1s are present in the sample and each has more than one representative allele.

The lack of specificity of the primers designed for LRE3 as well as this locus’ polymorphic presence in the population indicates than an additional control L1 locus that is known to be actively retrotransposing must be included in this study. Recent papers identified and characterized several previously unknown “hot” L1s, some of which were hypothesized to be the most active in the human genome (Tubio et al., 2014; Nüsgen et al., 2015). These loci would be excellent candidates for use as a control in methylation studies and primers were provided by the authors for use in BSP analysis (Tubio et al., 2014; Nüsgen et al., 2015).

*Methylation Comparisons Among Loci*

Loci 1 and 2 in the prostate cell lines had substantially lower methylation levels in RWPE-1 and LNCaP than in WAV17 and leukocytes (Table 5). Locus 2 was also less methylated in DU145, but Locus 1 was similarly methylated in this cell line as in WAV17 and leukocytes. Long Arm Locus 1 was similarly highly
methylated in the prostate cell lines as in WAV17 and leukocytes and was much more methylated than the HC21p loci. The numerous amplicons obtained for LRE3 in the prostate cell lines make comparisons with the other loci difficult. The actual LRE3 locus was only detected in RWPE-1 and was highly methylated as it is in leukocytes. The amplicons in LNCaP and DU145 were on average less methylated than other cell lines, but there was a large range in each cell line.

**Analysis of Critical CpGs**

In the prostate cell lines, the critical CpG sites were much less methylated in the HC21p loci than in Long Arm Locus 1, which is similar to the results observed in WAV17 and leukocytes (Table 6). Locus 2 had the lowest average methylation of the critical CpG sites in the prostate cell line, ranging from 0%-35.3% (Table 6). The average methylation of the critical CpG sites was slightly higher for Locus 1, ranging from 25.0%-86.75 (Table 6). The CpG sites in Long Arm Locus 1 were completely methylated in LNCaP and DU145, but slightly less methylated in RWPE-1 (90.9% average methylation). For LRE3, the critical CpG average methylation ranged from 69.6%-94.3%. It is not possible to draw strong conclusions because of the large number of LRE3 amplicons present in the prostate cell line data and the low frequency of each amplicon in the samples.
Figure 14. Locations of L1 Loci 1-4 in an Alphoid Region on HC21p. Mp1-5 indicate the regions of alphoid sequence mapped to HC21p. Other clusters of repetitive sequences such as Satellites I and III are indicated. Hash marks represent a gap in the map.
<table>
<thead>
<tr>
<th>Locus 1 (Promoter/Total)</th>
<th>WAV17</th>
<th>Leukocytes</th>
<th>RWPE-1</th>
<th>LNCaP</th>
<th>DU145</th>
<th>HC2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>89.6%/78.9%</td>
<td>82.9%/74.4%</td>
<td>41.9%/35.6%</td>
<td>22.0%/16.8%</td>
<td>89.5%/81.2%</td>
<td>N/A</td>
</tr>
<tr>
<td>Locus 2</td>
<td>56.7%</td>
<td>73.4%</td>
<td>43.9%</td>
<td>0.7%</td>
<td>48.9%</td>
<td>N/A</td>
</tr>
<tr>
<td>Locus 3 (Promoter/Total)</td>
<td>78.8%/80.0%</td>
<td>42.6%/46.2%</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>N/A</td>
</tr>
<tr>
<td>Locus 4</td>
<td>52.4%</td>
<td>69.4%</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>N/A</td>
</tr>
<tr>
<td>Long Arm Locus 1</td>
<td>99.2%/99.1%</td>
<td>97%/98.1%</td>
<td>93.0%/93.5%</td>
<td>94.9%/92.9%</td>
<td>100%/100%</td>
<td>N/A</td>
</tr>
<tr>
<td>Long Arm Locus 2</td>
<td>99.0%/99.1%</td>
<td>98.8%/98.9%</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>N/A</td>
</tr>
<tr>
<td>LRE3</td>
<td>N/A</td>
<td>97.9%</td>
<td>87.4%</td>
<td>51.2%</td>
<td>58.4%</td>
<td>95.6%</td>
</tr>
</tbody>
</table>

Table 5. Total Percent Methylation of L1 Loci. The values of percent methylation within the amplicon for each given locus across all clones collected from each sample are indicated here. For the loci with amplicons containing one or more CpG dinucleotides in the sequence adjacent to the L1 element, the values are given as total percent methylation within the L1 promoter/total percent methylation within the entire amplicon (promoter/total). TBD: To be determined; N/A: Not applicable (the locus is not present on the chromosome in that hybrid cell line).
Figure 15. Locus 1 in WAV17. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 16. Locus 1 in WAV17 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 17. Locus 1 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 18. Locus 1 in Leukocytes Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 19. Locus 2 in WAV17. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 20. Locus 2 in WAV17 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 21. Locus 2 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 22. Locus 2 in Leukocytes Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 23. Locus 3 in WAV17. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 24. Locus 3 in WAV17 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 3 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 25. Locus 3 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 26. Locus 3 in Leukocytes Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 3 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 27. Locus 4 in WAV17. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 28. Locus 4 in WAV17 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 4 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 29. Locus 4 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 30. Locus 4 in Leukocytes Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 4 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 31. Long Arm Locus 1 in WAV17. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 32. Long Arm Locus 1 in WAV17 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 33. Long Arm Locus 1 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 34. Long Arm Locus 1 in Leukocytes Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 35. Long Arm Locus 2 in WAV17. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 36. Long Arm Locus 2 in WAV17 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 37. Long Arm Locus 2 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 38. Long Arm Locus 2 in Leukocytes Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 39. LRE3 in HC2. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 40. LRE3 in Leukocytes. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
<table>
<thead>
<tr>
<th>Critical CpG Site</th>
<th>WAV17</th>
<th>Leukocytes</th>
<th>RWPE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>52</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>Locus 1</td>
<td>100.0%</td>
<td>-</td>
<td>88.9%</td>
</tr>
<tr>
<td>Locus 2</td>
<td>-</td>
<td>70.0%</td>
<td>-</td>
</tr>
<tr>
<td>Locus 3</td>
<td>-</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Locus 4</td>
<td>-</td>
<td>100.0%</td>
<td>-</td>
</tr>
<tr>
<td>Long Arm Locus 1</td>
<td>-</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Long Arm Locus 2</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>LRE3</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Critical CpG Site</th>
<th>LNCaP</th>
<th>DU145</th>
<th>HC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>52</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>Locus 1</td>
<td>25.0%</td>
<td>-</td>
<td>25.0%</td>
</tr>
<tr>
<td>Locus 2</td>
<td>-</td>
<td>0.0%</td>
<td>-</td>
</tr>
<tr>
<td>Locus 3</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>Locus 4</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>Long Arm Locus 1</td>
<td>-</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Long Arm Locus 2</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td>LRE3</td>
<td>70.0%</td>
<td>64.3%</td>
<td>71.4%</td>
</tr>
</tbody>
</table>

Table 6. Percent Methylation at Each Critical CpG Site. The percent methylation across all clones at each critical CpG site (+52, +58, +61, +70) for each locus is indicated when present in the sequence. The average percent methylation for all four critical CpG sites is indicated in blue. TBD: to be determined. N/A: Not applicable.
Figure 41. Locus 1 in RWPE-1. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 42. Locus 1 in RWPE-1 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 43. Locus 1 in LNCaP. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.

Figure 44. Locus 1 in LNCaP Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 45. Locus 1 in DU145. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 46. Locus 1 in DU145 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 47. Locus 2 in RWPE-1. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 48. Locus 2 in RWPE-1 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 49. Locus 2 in LNCaP. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 50. Locus 2 in LNCaP Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 51. Locus 2 in DU145. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 52. Locus 2 in DU145 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Locus 2 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 53. Long Arm Locus 1 in RWPE-1. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 54. Long Arm Locus 1 in RWPE-1 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 55. Long Arm Locus 1 in LNCaP. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 56. Long Arm Locus 1 in LNCaP Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 57. Long Arm Locus 1 in DU145. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 58. Long Arm Locus 1 in DU145 Using Consensus Sequence of All L1 Loci. The individual amplicon sequences of the L1 loci in this study were aligned to create a consensus sequence that includes the 40 CpG sites for which data was collected across all loci. A schematic of the L1 promoter is shown above the methylation data and the nucleotide position of each CpG site is indicated. The 4 critical CpG sites are highlighted in red. The portion of the promoter sequence that is covered in the Long Arm Locus 1 amplicon is indicated in purple. The sense and antisense promoters (SP and ASP, respectively) are indicated by black arrows. The elements that are known to be important for the regulation of L1 transcription are indicated by colored lollipops. Open circles: unmethylated CpG; Closed circles: methylated CpG.
Figure 59. LRE3 in RWPE-1. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 60. LRE3 in LNCaP. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
Figure 61. LRE3 in DU145. The position of each CpG site is indicated on the top. Open circles: unmethylated CpG; Closed circles: methylated CpG. The summary of average methylation at each CpG site is shown at the bottom of the figure. Haplotype numbers are shown at the right of the diagram.
CHAPTER FOUR

DISCUSSION

DNA methylation is one of the major defense mechanisms that the host cell uses to suppress transposon mobilization and, in fact, the majority of methylated cytosines within the human genome of normal cells reside in transposons (Goll and Bestor, 2005). DNA methylation in the promoter regions of L1s in particular is an epigenetic mechanism that has long been known to play a critical role in suppressing retrotransposition (Hata and Sakaki, 1997). To date, studies investigating L1s focused only on those elements that are located in euchromatic regions. There has been no previous work regarding L1 methylation status in the constitutively heterochromatic regions of the genome such as the acrocentric chromosome short arms. In order to explore DNA methylation of L1 elements in heterochromatin, I chose the well-characterized HC21p as a model (Doering et al., 1993). Using the BSP method for methylation analysis, I designed primers for four full-length L1s on HC21p, as well as three euchromatic controls. I hypothesized that the L1s on HC21p are hypomethylated, which may facilitate heterochromatin formation as on the inactive X chromosome during XCI.

My results demonstrate that the L1s on HC21p are substantially hypomethylated compared to the euchromatic controls in leukocytes. In WAV17 and leukocytes, the total percent methylation of the loci on HC21p ranged from 42.6-89.6%, whereas the euchromatic HC21q controls had greater than 97%
methylation in both cell lines (Table 5). The four critical CpG sites known to be essential for L1 suppression by DNA methylation (Hata and Sakaki, 1997) were differentially methylated in the HC21p loci compared to the euchromatic control loci. One or more of these sites were more frequently absent from the HC21p loci and were substantially less methylated compared to the same sites in the euchromatic controls (Table 7).

I also sought to determine if DNA methylation of the L1s on HC21p can serve as potential cancer biomarkers. L1s are known to be hypomethylated in cancer cells on a genome-wide level (Kanai, 2010; Kitkumthorn and Mutirangura, 2011). However, very few studies have investigated the methylation status of individual L1 elements in the genome, and those that did only looked at L1s located in euchromatic regions (Pobsook et al., 2011; Phokaew et al., 2008; Singer et al., 2012). The general consensus of those studies was that specific L1s are hypomethylated, which corresponds to the global trend, but their methylation patterns varied significantly. To study cancer cell L1 methylation in heterochromatic genome regions, I performed BSP on two prostate cancer cell lines for two of the HC21p loci. Overall, my results show that the L1s on HC21p are hypomethylated in all prostate cell lines (0.7-89.5% total methylation) compared to the euchromatic controls (>93% total methylation) and leukocytes (Table 8). These data suggest that decreased methylation levels of the HC21p loci may be candidate biomarkers for cancer detection and prognosis.

Challenges and Limitations

Primer design for this study was difficult due to the fact that the L1 promoter region is very conserved across all L1 subfamilies. In order to design
primers that would amplify specific L1s, one of the primers must be within the unique sequence adjacent to the 5' end of the L1. However, the repetitiveness of the sequences surrounding the L1s on heterochromatic HC21p made it difficult to design a unique primer. The presence of multiple highly similar L1s on the acrocentric chromosome short arms is due to the high frequency of segmental duplications caused by regular non-homologous exchanges in these genomic regions (Choo, 1990; Ziccardi, 2012). Additional complexity for specific primer design was due to incomplete sequence data for the acrocentric chromosome short arms. As a result, there are no available bioinformatic tools to determine if the known HC21p L1 loci are present on the other acrocentric chromosomes. The only precaution I could take was to rule out non-specific amplification in the sequenced portions of the human genome. The Primer-BLAST tool provided by NCBI performs this function by using a BLAST and global alignment algorithm to screen primers against the RefSeq genome (Ye et al., 2012). If the results generated in this program confirmed that they had the potential to anneal elsewhere in genome, I redesigned the primers. While I was able to eliminate the possibility that my primers would anneal in the non-heterochromatic regions of the genome, I knew that they could still anneal to similar sequences at other locations on HC21p or any of the other acrocentric chromosome short arms.

To address this issue, I began by collecting methylation data for the HC21p loci in a hybrid cell line containing one copy of HC21. From these data, I was able to determine that natural variability does occur at a single copy of one locus both in the sequence, as well as the number of potential methylation pattern haplotypes. From these data, I assigned cut-off values that accounted for
both types of variability in order to characterize the origin of the sequences obtained from diploid cell lines. Overall, the primers designed for all of the L1 loci on HC21p did not amplify unique L1s, but actually a limited number of highly similar L1 loci located on more than one acrocentric chromosome short arm, and in potentially more than one copy on HC21p. Ultimately, sequences for each locus would need to be obtained from the other acrocentric chromosome hybrid cell lines in order to determine the origin of the undefined sequences.

Another challenge in this project involved the CpG sites themselves. While L1s are overall highly conserved, CpG sites in general are considered to be “hot spots” for spontaneous mutation (Robertson and Wolffe, 2000). This is due to the increased deamination of 5-methylcytosine compared to cytosine as well as the inefficient repair of deaminated cytosines (Robertson and Wolffe, 2000). The consequences of these events are either a CpG to TpG transition or the removal of the deaminated cytosine product (Robertson and Wolffe, 2000). Thus, even highly similar L1 sequences may differ in the number of CpG sites within the promoter. Fortunately, the QUMA program that was used for all of my methylation analyses provides an optional repetitive sequence parameter that takes into account polymorphisms in CpG sites (described in Materials and Methods). Without adding this parameter, a polymorphic CpG site is automatically considered to be unmethylated regardless of the underlying nucleotides. When using this parameter, if any of the sequences lacked an expected CpG site as determined from the reference sequence or contained an additional CpG site that was perhaps lost from the reference, the program removed that polymorphic CpG data point from the analysis.
Additional euchromatic control L1s will be necessary due to limitations of the control loci that I chose for this study. For instance, LRE3 proved to be highly complex in both genomic location and percent methylation, making it difficult to draw many conclusions from the data. The polymorphic nature of this locus decreases the likelihood that an attempt to re-design primers would increase specificity, so it will be necessary to include an additional known transcriptionally active or “hot” L1. Another “hot” L1Hs that is located on chromosome 22q12 was recently characterized and is a promising candidate for an additional control (Tubio et al., 2014). Furthermore, the L1 loci located on HC21q that were chosen for this study belong to the L1PA2 subfamily, whereas the HC21p loci were either L1PA3 or L1PA4. In order to perform a more direct comparison for the euchromatic controls, it would be useful to choose additional loci on HC21q that are of the same evolutionarily subfamilies as the HC21p loci.

**Mechanisms for L1-Mediated Heterochromatin Formation**

The idea that there were elements on the X chromosome acting as “way stations” or “boosters” to facilitate X chromosome inactivation (XCI) was initially proposed in 1989, but the specific identity of these components was unknown (Riggs, 1989). Close to a decade passed before L1s were proposed as agents promoting XCI based on Lyon’s observation that L1s were twice as dense on the X chromosome compared to any of the autosomes (Lyon, 1998; Bailey et al., 2000). Numerous studies have provided supporting evidence to Lyon’s proposal, identifying decreased methylation and transient bidirectional transcription of the L1s on the inactive X chromosome as major contributing factors to their role in XCI (discussed below).
In addition to the higher L1 density on the X chromosome compared to the autosomal chromosomes, the L1s are not evenly dispersed along the length of the X chromosome (Bailey et al., 2000). In L1-rich regions of the X chromosome, the L1s are less methylated on the inactive X than on the active chromosome and by cooperating with the Xist RNA, the nearby genes are readily pulled into the heterochromatic compartment during the early stages of XCI (Chow et al., 2010; Singer et al., 2012). However, genes in regions of the X chromosome that are depleted of L1s are more resistant to silencing and tend to be the genes that escape XCI (Bailey et al., 2000). When pieces of the X chromosome are translocated to autosomal chromosomes, heterochromatization occurs that resembles XCI and is accompanied by the spreading of DNA methylation into the autosomal regions at the translocation breakpoint (Cotton et al., 2014; Tannan et al., 2014; Peeters et al., 2014). The autosomal regions that are most susceptible to this process are rich in both L1Hs and L1PA elements, which suggests an evolutionarily conserved ability to facilitate heterochromatin formation that is not necessarily specific to the X chromosome sequences (Zhang et al., 2012; Cotton et al., 2014; Tannan et al., 2014). L1 rich regions in autosomal regions containing monoallelically expressed imprinted genes may also be facilitating heterochromatin formation of one allele (Allen et al., 2011).

The high sequence conservation of L1 elements suggests a potential mechanism wherein the elements participate in intrachromosomal pairing in order to alter chromatin structure (Hansen, 2003; Chow et al., 2010). A chromosomal region that has a high density of L1s can interact at many positions to promote the formation of loops that will allow the chromatin to condense,
whereas the regions depleted of repetitive sequences that escape XCI will be unable to participate in these extensive interactions (Hansen, 2003; Chow et al., 2010; Cotton et al., 2014). During the early stages of XCI, initial heterochromatin formation may temporarily prevent DNMTs from accessing the L1s, which would account for the observed decrease in L1 methylation on the inactive X chromosome (Hansen, 2003; Chow et al., 2010). Hypomethylated L1s are hypothesized to form interactions more easily than hypermethylated L1s, which supports a looping mechanism for facilitating heterochromatin formation during XCI (Hansen, 2003; Chow et al., 2010).

While global L1 transcription is generally suppressed in differentiating cells, it has been shown that a subset of L1s is transcribed from the inactive X chromosome to generate a population of small RNAs during the specific time period that XCI occurs (Chow et al., 2010). Furthermore, small sense and antisense RNAs (~19–21 nucleotides) that map to L1 promoters are also detected during XCI in mice, indicating that the L1s are transiently expressed (Chow et al., 2010). This unusual, localized L1 expression suggests a potential mechanism that requires the upregulation of L1 transcription to facilitate the spreading of heterochromatin into regions resistant to inactivation. The classic RNA interference (RNAi) pathway has already been implicated in the suppression of L1 retrotransposition at a post-transcriptional level by inducing mRNA degradation (Yang and Kazazian, 2006). A process similar to RNAi that instead acts on the chromatin level may explain the function of L1 transcription from the inactive X chromosome during XCI (Chow et al., 2010). RNAi-mediated chromatin modification mechanisms utilize small RNAs that interact with the
DNA from which they were transcribed in order to recruit chromatin modifying proteins (Holoch and Moazed, 2015). In this type of mechanism, dsRNAs are processed by proteins in a similar manner as in RNAi, but the resulting small RNAs are instead directed back to the nucleus to induce changes in chromatin structure in cis (Figure 62)(Holoch and Moazed, 2015). This process has been well-characterized in model organisms such as Arabidopsis thaliana and Drosophila melanogaster, but the details of the homologous pathways in mammalian cells have yet to be completely elucidated (Holoch and Moazed, 2015). One study using mice has described the presence of an RNAi-dependent DNA methylation pathway wherein small retrotransposon RNAs are able to direct DNMTs to genomic retrotransposon sequences in cis (Aravin et al., 2008).

A similar mechanism has been proposed that also utilizes small L1 RNAs derived from the ASP to direct chromatin modifications, but in an RNAi-independent manner. Transcription from the L1 ASP into sequences upstream of the L1 5’UTR will result in the generation of L1-chimeric transcripts containing both L1 and the adjacent sequence. A well-studied example is the L1-cMet fusion transcript that originates from the ASP of a L1PA2 element (Roman-Gomez et al., 2005; Weber et al., 2010). When transcription is initiated at the ASP of the L1 located within an intron of the cMet proto-oncogene, a fusion transcript is generated that contains both L1 and cMet sequence (Roman-Gomez et al., 2005). It has been suggested that this chimeric transcript may have a regulatory role as its presence is associated with decreased cMet expression (Roman-Gomez et al., 2005). A mechanism comparable to an RNAi-mediated chromatin modification pathway may be activated upon demethylation of the L1 promoter to permit ASP
transcription and the generation of the L1-cMet fusion transcript (Figure 62)(Weber et al., 2010). In this alternative mechanism, the single-stranded antisense transcript will then be processed to make a small AS RNA containing only the cMet sequence (Weber et al., 2010). Thus, this pathway is RNAi-independent in that it does not process dsRNA. While this type of mechanism has been well-characterized in model organisms, there is currently not any substantial data to support an RNAi-independent process specifically involving retrotransposons in mammals (Holoch and Moazed, 2015). However, Xist RNA that coats the inactive X chromosome during XCI has been suggested to promote an RNAi-independent process regulating chromatin modifications, indicating that the machinery necessary for this mechanism likely exists in mammals (Holoch and Moazed, 2015).

**L1 Methylation and Heterochromatin Formation on HC21p**

I propose a model that incorporates both intrachromosomal pairing and small RNA regulatory mechanisms that have been observed in other systems to support my hypothesis that the L1s dispersed along HC21p may be influencing heterochromatin formation (Figure 63). The observed hypomethylation within the promoters of the HC21p L1 loci may permit a localized euchromatic environment that would allow intrachromosomal pairing and loop formation (Figure 63). In contrast, the L1 loci on HC21q are highly methylated, which would inhibit pairing and allow the chromatin to remain open. Hypomethylation of the HC21p L1 loci can also potentially allow for sense and antisense transcription to promote the RNAi-mediated chromatin modification mechanism described above (Figure 62). While the HC21p L1 loci are evolutionarily older
and likely retrotransposition-incompetent, they may retain the ability to initiate transcription if the transcription factor binding sites and promoter sequences are relatively conserved, as in the case of the L1-cMet chimeric transcript produced from a L1PA2 (Cruickshanks and Tufarelli, 2009; Roman-Gomez et al., 2005; Yang and Kazazian, 2006). Even if there are mutations in the promoter regions that have weakened transcription processivity, this may not necessarily prevent the generation of short transcripts from both the SP and ASP (Yang and Kazazian, 2006). The L1 promoter sequence within the two transcripts can pair to form a partial dsRNA and the region of the dsRNA that overlaps is cleaved by an RNAi-type nuclease to form a small interfering RNA (siRNA) containing only L1 sequence (Figure 62). One strand of this siRNA associates with an unknown complex that will direct it back to the L1 locus on HC21p in cis, similar to what is observed during retrotransposition (Yang and Kazazian, 2006). This will lead to the recruitment of chromatin modifying enzymes such as DNA methyltransferases (DNMT) and histone deacetylases (HDAC) to promote heterochromatin formation (Figure 63). In addition, the full single-stranded antisense L1 RNA transcript can be processed to make a small antisense RNA containing only the adjacent HC21p sequence to promote the undefined RNAi-independent mechanism that was also described earlier (Figure 62). This presents another opportunity for recruitment of chromatin modifying machinery specifically to the regions surrounding the L1s on HC21p.

Additional observations may also support the potential for transcription from both promoters of the HC21p L1 loci. First, the four critical CpG sites that are known to be important for L1 transcription have decreased average
methylation in the HC21p L1 loci (36.4-86.7% in leukocytes) compared to the control loci where all four sites are 100% methylated (summarized in Table 7). The lower levels of methylation at these sites in the HC21p loci could potentially allow transcription to occur because it is known that methylation of these critical CpG sites is necessary for suppression of L1 transcription (Hata and Sakaki, 1997). While the role of the methylation of CpGs outside of gene promoters, in regions referred to as “CpG island shores,” has not been well characterized, they are known to affect the tissue-specific expression of nearby genes, even when located up to 2 kilobases upstream (Irizarry et al., 2009; Jones, 2012). The very low methylation at the two CpG sites in the region upstream of Locus 1 sites may also facilitate expression from the ASP into the adjacent region to regulate the nearby chromatin structure (Figures 15 and 17). This is in contrast to the CpG sites in the shores of both control loci, which were almost completely methylated (Figures 31, 33 35, and 37). The Locus 3 amplicon also contained a CpG site upstream of the L1 promoter but unlike Locus 1, it was highly methylated when present (Figures 23 and 25). However, I cannot draw any definitive conclusions about the potential for transcription from the ASP due to the limited number of CpG sites in the shores and the contrast between the levels of methylation of these sites in Loci 1 and 3. Only the HC21p Loci 2 and 4 amplicons span the actual ASP, which begins approximately 400 base pairs downstream of the beginning of the L1. Methylation of CpG sites in the ASP region of these two loci is relatively low at roughly 37% on average in leukocytes (Figure 21, CpG sites +425 and +492; Figure 29, CpG site +492). While I again cannot draw any major conclusions from these observations, my preliminary data is consistent with
expression from the ASP promoter.

Local changes in other repetitive sequences on the acrocentric chromosomes may also play a role in heterochromatization. Other work in our lab is focusing on different repetitive sequences that are characteristic of heterochromatin, such as NBL-2 as well as gamma and beta satellites. However, preliminary DNA methylation data obtained from these regions indicate that these repetitive elements are not hypomethylated in WAV17 and leukocytes. While much of this evidence points to a major role of L1s in heterochromatin formation, changes in DNA methylation in L1 promoters are unlikely to be the sole initiating factor in heterochromatization. Further work to analyze the histone modifications within both the L1 promoter and in the adjacent regions will help to better elucidate the role of the HC21p L1 loci in heterochromatin formation. Activating histone modifications within and immediately upstream of the L1 promoter, as well as the presence of small transcripts that map to the HC21p loci, would support my hypothesis.

In summary, the overall hypomethylation that I observed in the promoters of the HC21p L1 loci may facilitate intrachromosomal pairing and transcription from the L1 SP and ASP to induce an RNAi-mediated chromatin modification pathway. The hypomethylation of the critical CpG sites as well CpG sites in the adjacent genomic regions further support the possibility of regulation by transcription from the SP and ASP.

**Prostate Cancer**

Genomewide hypomethylation of L1s is a major characteristic of tumor cells (discussed in Chapter One). As described earlier, the most likely
consequences of this aberrant methylation in cancer cells are overall genome instability and tumor progression. However, it is still unknown if the consequences of hypomethylation are at a global, locus-specific level, or perhaps both. A locus-specific mechanism is supported by evidence that the methylation patterns of specific L1s can vary between different cells with identical global L1 methylation (Figure 2) (Pobsook et al., 2011; Phokaew et al., 2008; Nüsgen et al., 2015). In many cases, measuring global L1 methylation levels was not sensitive enough to detect any clinical differences between cell types in patients, such as malignancy or tumor grade (Pobsook et al., 2011; Nüsgen et al., 2015). Thus, a biomarker that detects methylation patterns of specific L1 loci may prove to be a sensitive detection and/or prognostic tool. My preliminary data support the hypothesis that the methylation of specific HC21p loci may be better candidates for a precise biomarker than genomewide L1 methylation by demonstrating that HC21p L1 promoter methylation varies between loci in cancer and can distinguish between cancer and normal cell lines.

Loci 1 and 2 were substantially less methylated in LNCaP when compared to normal peripheral leukocytes, whereas methylation of Long Arm Locus 1 was very high across all cell lines (summarized in Table 8). This indicates that the HC21p loci, but not Long Arm Locus 1, are potential candidates for diagnostic biomarkers. Additionally, this trend of hypomethylation was also observed at the four critical CpG sites, which supports the hypothesis that methylation analysis of specific CpG sites may be sufficient for creating a sensitive biomarker (Table 6). However, my data indicate that methylation of Loci 1 and 2 changed differently in response to tumor stage (Table 8). While Locus 2 was much less
methylated in both LNCaP and DU145 compared to leukocytes, methylation of Locus 1 did not change much between DU145 and leukocytes (Table 8). LNCaP is considered to be a less aggressive cell line than DU145, indicating that Locus 2 has the potential to serve as a detection tool throughout tumor progression, whereas Locus 1 methylation alone could not detect a more aggressive cancer (Russell and Kingsley, 2003).

My data also shows the importance of choosing relevant cell lines for methylation studies and analyzing the results in the context of their tissue origin. For instance, the purpose of including RWPE-1 in this study was to perform a direct comparison of a “normal” prostate cell line to the prostate tumor cell lines. I expected that L1s would be similarly methylated in RWPE-1 as in leukocytes, which was the other “normal” control cell line used in this study. However, the HC21p loci were less methylated in RWPE-1 compared to leukocytes. Since RWPE-1 is an immortalized cell line derived from normal prostate epithelial cells it is possible that the decreased methylation observed in RWPE-1 may reflect transformation-dependent hypomethylation. In support of this hypothesis, the transformation process is frequently observed to result in global hypomethylation (Grafodatskaya, 2009). However, an alternative theory would be that methylation may be tissue-specific, in that the L1s are normally less methylated in prostate cells than in leukocytes. Additional studies using DNA from other normal tissues will be required in order to make this distinction. Thus, it is necessary to characterize the methylation of many additional L1 loci in order to develop a highly sensitive cancer biomarker tool. In addition, the consistently high levels of methylation in Long Arm Locus 1 in the prostate
cancer cell lines suggests that the HC21q loci can serve as internal controls.

**Future Directions**

While the hypomethylation that I observed in the HC21p loci provides preliminary support for the role of L1s in heterochromatin formation, the conclusions that I can draw from my data remain limited. Additional methylation analyses for the other full-length loci on HC21p will be required to determine if L1 hypomethylation is a consistent trend for the acrocentric chromosome short arms. Going forward, I have designed primers for two other L1 loci on HC21p. As mentioned earlier, additional euchromatic controls on HC21q will be necessary in order to specifically enable a direct comparison by L1 subfamily. I chose to include a highly active L1 in my study as an additional euchromatic control because it should be highly methylated in normal cells. However, LRE3 proved to be an inadequate control due to its polymorphic presence in the human population, as well as the lack of specificity of my primer set. Thus, one or more additional active (“hot”) L1s will be chosen for methylation analyses. It will be useful to select additional “normal”, non-malignant cell lines and tissues in order to determine if the L1s on HC21p are consistently hypomethylated as in leukocytes. As already indicated, immortalized cell lines such as RWPE-1 may not be ideal to use for methylation comparisons to malignant cells, so DNA from primary cells should be collected. If methylation of specific L1s on HC21p is to be used as a cancer biomarker, additional cancer models, including a number of matched normal-cancer samples, must be carefully selected in order to determine if any differences in methylation are specific to tumor origin tissue and/or tumor grade.
Table 7. Summary of Critical CpG Site Methylation in Leukocytes. The percent methylation for each locus in leukocytes is shown with the average methylation across all four CpG sites shown on the right.

<table>
<thead>
<tr>
<th>Locus</th>
<th>+52</th>
<th>+58</th>
<th>+61</th>
<th>+70</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 1</td>
<td>70.0%</td>
<td>-</td>
<td>100.0%</td>
<td>90.0%</td>
<td>86.7%</td>
</tr>
<tr>
<td>Locus 2</td>
<td>-</td>
<td>83.3%</td>
<td>-</td>
<td>-</td>
<td>83.3%</td>
</tr>
<tr>
<td>Locus 3</td>
<td>-</td>
<td>100.0%</td>
<td>84.6%</td>
<td>50.0%</td>
<td>78.2%</td>
</tr>
<tr>
<td>Locus 4</td>
<td>-</td>
<td>36.4%</td>
<td>-</td>
<td>-</td>
<td>36.4%</td>
</tr>
<tr>
<td>Long 1</td>
<td>-</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Long 2</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Table 8. Conclusions for HC21 L1s as Cancer Biomarkers.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Leukocytes (normal)</th>
<th>LNCaP (somewhat aggressive)</th>
<th>DU145 (very aggressive)</th>
<th>Biomarker?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 1</td>
<td>82.9%</td>
<td>22.0%</td>
<td>89.5%</td>
<td>Detect early stages</td>
</tr>
<tr>
<td>Locus 2</td>
<td>73.4%</td>
<td>0.7%</td>
<td>48.9%</td>
<td>Detect throughout progression</td>
</tr>
<tr>
<td>Long 1</td>
<td>97.0%</td>
<td>94.9%</td>
<td>100%</td>
<td>Unlikely</td>
</tr>
</tbody>
</table>
Figure 62. Mechanisms for RNAi-Mediated Chromatin Modifications. The L1 is transcribed from its sense promoter (SP) and antisense promoter (ASP) within its 5’ untranslated region (5’UTR) to produce sense (S) and antisense (AS) transcripts, respectively. In a RNAi-mediated chromatin modification mechanism that is similar to RNA interference (RNAi) (blue box, right), the two complementary transcripts pair to form a partial dsRNA. The region of the dsRNA that overlaps is cleaved by an RNAi-type nuclease to form a small interfering RNA (siRNA) containing only L1 sequence. One strand of this siRNA associates with an unknown complex that will direct it back to the genomic L1 locus in cis. This will lead to the recruitment of chromatin modifying enzymes such as DNA methyltransferases (DNMT) and histone deacetylases (HDAC) to promote heterochromatin formation. Another potential variation of this mechanism (orange box, bottom left) involves the processing of the full AS RNA transcript to make a small AS RNA containing only the unique genomic sequence adjacent to the L1. This small AS RNA can then localize to the genomic location from which it was transcribed and recruit chromatin modifying enzymes. Since the sense transcript does not contain adjacent genomic sequence, dsRNA will not form at the unique genomic region of the ASP transcript. Thus, this process is likely RNAi-independent.
Figure 63. Proposed Mechanism of Heterochromatin Formation on HC21p. The hypomethylated L1s on HC21p pair to form loops and are transcribed to induce the RNAi-mediated chromatin modification pathway.
LITERATURE CITED


VITA

Sarah Tincher was born and raised in the Northwest suburbs of Chicago, Illinois. She received her Bachelor of Science in Biology in 2012 at Loyola University Chicago and continued her studies there to pursue a Master of Science in Biology. Currently, she is in her second year of a PhD program at University of Illinois at Chicago, where she investigates the role of nitric oxide in the regulation of epigenetic mechanisms in cancer.