2017

Epigenetic Profiling of Mammalian Retrotransposons

Arundhati Bakshi

Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Part of the Life Sciences Commons

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/4357

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
EPIGENETIC PROFILING OF MAMMALIAN RETROTRANSPOSONS

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy in The Department of Biological Sciences

by

Arundhati Bakshi
B.S., Louisiana State University, 2013
August 2017
ACKNOWLEDGEMENTS

I joined Dr. Joomyeong Kim’s lab as an undergraduate student, a “newbie” to the world of science. He discovered in me the potential to be a good scientist, and then gave me the opportunity in his lab to realize that potential. For that, I will always remain grateful to Dr. Kim, my boss and mentor for the past six years. He gave me the opportunity to learn from his knowledge and experience, and his constructive criticisms frequently helped me fine-tune my ideas. He also trained me to think of science both in publication units as well as the “big-picture,” which is an invaluable skill for a scientist. I am equally indebted to Dr. Hana Kim, who has been nothing short of a second mentor to me since I joined the lab, and has shaped most of what I know about science today. I did not just learn science techniques from her, but also gained valuable insights regarding the philosophy behind doing science, and all the dedication and hard work it requires. She has been exceedingly generous in sharing her knowledge and friendship with me, and I cannot thank her enough for all that she has done. My thanks also extends to all the other members of the Kim lab (past and present) – Dr. Michelle Thiaville, Dr. Muhammad Ekram, Mr. Mohammad Faisal, Dr. An Ye, Dr. B. Piniti Perera, Dr. Wesley Frey, Mr. Corey Bretz, Ms. Terri Cain and Mr. Carter Pesson. They have all shaped my “scientific-personality” in one way or another, and have added uniquely to my experiences at the Kim lab. I thank my dissertation committee members – Dr. Mark Batzer, Dr. David Donze, and Dr. Samithamby Jeyaseelan – for their scientific guidance, as well as for all the memorable non-scientific chats. I am equally grateful to Dr. Doug Gilman and Dr. Huangen Ding, my undergraduate thesis committee members, for their unrelenting support of my scientific ambitions. I extend a special thank-you to Dr. E. William Wischusen, who first noted in me the potential to be a good educator and mentor, and has continued to direct me towards honing that potential. I thank him
for always encouraging and supporting me in my continually evolving goals, and for providing me with opportunities so I could aim to achieve them. On that note, I would like to thank all of the undergraduate students I have worked with; they have taught me more about teaching and mentoring than I could have ever learned from courses, seminars or science education literature. I would like to make special mention of Dr. Elzbieta Cook and Dr. Carol Taylor, who have guided, supported and mentored me since I first began as an international student at LSU. They saw potential in me that I never knew I had, and I am immensely grateful to them for that. A huge thank-you to Dr. Scott Herke for his dedication towards the best execution of my project; Ms. Chimene Williams and Dr. Ann Jolissaint always having a ready answer to all my logistical problems; Mrs. Jasmine Baker, Ms. Terri Cain and Mr. Corey Bretz, for adding warmth, humor and genuine companionship to my years in graduate school, and often when I needed it most. This acknowledgement would be incomplete without a mention of my parents – they have constantly encouraged me to shoot for all my unconventional dreams, and have imbued me with the strength of character to make them come true. Finally, last but never the least, my most heartfelt thanks goes to my husband, Mr. Randy LeJeune, who has always been my personal pillar of strength. His support, encouragement, humor and companionship have been the very foundations on which I have built my hopes and dreams; words cannot express how grateful I am to have him in my life.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS........................................................................................................... ii

LIST OF TABLES....................................................................................................................... vi

LIST OF FIGURES...................................................................................................................... vii

ABSTRACT................................................................................................................................... viii

CHAPTER 1: BACKGROUND ........................................................................................................ 1
  The Jumping Genes ................................................................................................................ 1
  Epigenetics for “Host Defense” ............................................................................................ 8
  Research Objectives ............................................................................................................ 12
  References ............................................................................................................................ 15

CHAPTER 2: LOCUS-SPECIFIC DNA METHYLATION ANALYSIS OF RETROTRANSPOSONS IN ES, SOMATIC AND CANCER CELLS USING HIGH-THROUGHPUT TARGETED REPEAT ELEMENT BISULFITE SEQUENCING ......................................................................................................................... 22
  Introduction .......................................................................................................................... 22
  Experimental Design, Materials and Methods ...................................................................... 23
  References ............................................................................................................................ 29

CHAPTER 3: RETROTRANSPOSON-BASED PROFILING OF MAMMALIAN EPIGENOMES: DNA METHYLATION OF IAP LTRS IN EMBRYONIC STEM, SOMATIC AND CANCER CELLS .......................................................................................................................... 30
  Introduction .......................................................................................................................... 30
  Results ................................................................................................................................... 32
  Discussion ............................................................................................................................. 42
  Materials and Methods ....................................................................................................... 45
  References ............................................................................................................................ 48

CHAPTER 4: DNA METHYLATION VARIATION OF HUMAN-SPECIFIC ALU REPEATS ................................................................................................................................. 52
  Introduction .......................................................................................................................... 52
  Results ................................................................................................................................... 54
  Discussion ............................................................................................................................. 68
  Materials and Methods ....................................................................................................... 72
  References ............................................................................................................................ 76
CHAPTER 5: SUMMARY AND DISCUSSION .........................................................81
  Summary .........................................................................................81
  Retrotransposons as Drivers of Carcinogenesis .............................83
  Retrotransposons as Drivers of Mammalian Evolution ..................90
  Conclusions ......................................................................................93
  References .........................................................................................94

APPENDIX: LETTERS OF PERMISSION .....................................................98

VITA ........................................................................................................103
LIST OF TABLES

4.2: Characteristics of AluYa5/Yb8 elements located within 1-kb of the nearest transcription start site (TSS) and their associated gene ..................................................58
# LIST OF FIGURES

1.1: Distribution of major retrotransposons families in the human genome ..................................1
1.2: Structure of murine IAP LTR retrotransposons .................................................................3
1.3: IAP LTR elements can affect nearby gene expression .........................................................5
1.4: Structure and retrotransposition of Alu elements ..............................................................6
1.5: Alu elements can disrupt gene expression .................................................................7
2.1: Library preparation for HT-TREBS ..................................................................................24
2.2: Bioinformatics workflow for HT-TREBS ........................................................................27
3.1: HT-TREBS analyses of IAP LTRs in ES, somatic and Neuro2A cells .................................34
3.2: DNA methylation profiles of IAP LTRs in ES and Neuro2A cells compared to somatic cells ..............................................................................................................................36
3.3: DNA methylation pattern-based grouping of IAP LTRs ....................................................39
3.4: COBRA analyses of representative IAP LTR loci .............................................................41
4.1: HT-TREBS of AluYa5 and AluYb8 in human fibroblast cells ............................................55
4.2: Methylation variation of AluYa5 and Yb8 loci closely associated with a gene in normal and cancer samples ........................................................................................................60
4.3: Methylation variation of CEBPG-Alu in normal and cancer samples ...............................62
4.4: Inter-individual variation in DNA methylation of AluYa5 and Yb8 .................................64
4.5: Tumorigenesis-related variation in DNA methylation of AluYb8 .....................................66
5.1: Retrotransposons as drivers of carcinogenesis .................................................................84
5.2: Gene association analyses of IAP LTRs in LHL group ....................................................86
5.3: Expression level changes of Alu-associated genes in cancer tissues ...............................87
5.4: Conceptual diagram for Junk RNA-seq ...........................................................................89
5.5: Retrotransposons as drivers of mammalian evolution ...................................................93
ABSTRACT

Over evolutionary time, mammalian genomes have accumulated a large number of retrotransposons, making up about half of the genome in any given species. These retrotransposons are typically repressed by epigenetic mechanisms, one of the main ones being DNA methylation. It is well known that improper DNA methylation of retrotransposons can have unwanted consequences on nearby gene expression, and hypomethylation of retrotransposons has been frequently observed in various cancers. Nevertheless, it has been notoriously difficult to study retrotransposon loci individually due to the highly repetitive nature of their sequences. To address this issue, we have developed a novel protocol termed HT-TREBS (High-Throughput Targeted Repeat Element Bisulfite Sequencing), which is designed to survey the DNA methylation levels of a large number of interspersed repeat elements on an individual-locus basis. Here we have used this technique on two mammalian retrotransposon families: IAP LTRs in mouse, and the AluYa5 and AluYb8 subfamilies of Alu elements in humans. According to the results, the majority of retrotransposons (~95%) are heavily methylated in mammalian somatic cells. Moreover, only a fraction of loci appear to be sensitive to cell state in both species. Approximately 25% IAP LTRs showed hypomethylation in mouse ES cells, and ~50% in mouse cancer (Neuro2A) cells. In humans, however, the level of response to tumorigenesis in the breast was much more constrained, with only 1% AluYb8 elements being expected to show hypomethylation at an early stage. Interestingly, our results also revealed extensive (up to 10-fold) inter-individual variation in the level of DNA methylation of AluYa5 and AluYb8 elements in humans, similar to the variation previously noted regarding IAP LTRs in mice. Overall, these results highlight the dynamic nature of DNA methylation at retrotransposons, which further leads us to speculate its unique contribution to mammalian evolution and disease susceptibility by
allowing for epigenetic variation within one species. Furthermore, it also suggests the potential utility of some of these elements which are sensitive to cell state, but show less variability between individuals, to be used as epigenetic biomarkers for tracking disease progression.
CHAPTER ONE

BACKGROUND

The “Jumping Genes”

When Dr. Barbara McClintock first described “controlling elements” in maize [1], she was largely ignored by the scientific community. More than 30 years later, advances in molecular biology were finally able to support her visionary ideas, which led to her being awarded the

Figure 1.1. Distribution of major retrotransposon families in the human genome. Non-transposable elements comprise approximately half of the human genome, whereas the other half is composed of transposable elements. Human genomes harbor relatively few DNA transposons, and the bulk of transposable elements belong to the retrotransposon category. Two families of retrotransposons, L1 (LINE) and Alu (SINE), together make up about ~30% of the entire genome, with other retrotransposons contributing to ~15% of the genome. This figure has been adapted from Figure 1A in R. Cordaux, M.A. Batzer, The impact of retrotransposons on human genome evolution, Nature Reviews Genetics, 10 (2009) 691-703.
Nobel Prize. Today, whole-genome sequencing has confirmed that over half of any given mammalian genome consists of these “jumping genes” known as transposable elements (Figure 1.1) [2, 3].

There are two major classes of transposable elements: DNA transposons, which move by a “cut-and-paste” mechanism; and retrotransposons, which use an RNA intermediate to move by a “copy-and-paste” mechanism. The vast majority (~99%) of all transposable elements in the mammalian genome are the retrotransposons, which rely on the host genome for their transcription, and viral element-encoded reverse transcriptase and endonuclease for their transposition [4]. Retrotransposons could be largely divided into two groups: those that contain Long Terminal Repeats (LTR elements), and those that do not (non-LTR elements). LTR elements (~8% of the human genome) have typically originated from ancient retrovirus insertions that have lost their infectious capacity, but retain the viral transposition mechanisms which lead to the formation of the long terminal repeats that flank these elements. Often, LTR elements only exist as a single LTR, due to homologous recombination between the two LTRs resulting in the loss of the gene body of the retroelement. Such solo-LTRs vastly outnumber full-length elements in mammalian genomes [5, 6]. HERVs (Human Endogenous Retroviruses) in humans and IAP LTRs (Intracisternal A-Particle Long Terminal Repeats) in mice are two of the major mammalian LTR-type retrotransposon families. Of the non-LTR elements, there are two major groups: SINEs (Short INterspersed Elements), which are <500 bp and harbor no protein-coding genes; and LINEs (Long INterspersed Elements) which are >1 kb long and encode all proteins necessary for reverse transcription and re-integration into the genome. Based on the number of recently reported cases of de novo insertions in humans, it is believed that the non-LTR elements are currently actively retrotransposing in humans, and as such, pose a threat to...
genome stability [7-10]. Two such elements are the Alu (which is a SINE, showing one new insertion per ~20 births [11]) and L1 retrotransposons (LINE-1, with about 80-100 active loci out of the 500,000 copies [7]), that together make up nearly a third of the human genome (Figure 1.1) [8].

Though LTR elements pose little threat to the human genome in current evolutionary time, many are still active in other mammals, such as mice, and are responsible for a large number of insertional mutagenesis events due to their retrotransposition [12]. With about 10,000 copies, IAP LTRs (Intracisternal A-Particle Long Terminal Repeats) are one of the most abundant retrotransposons in mice (Figure 1.2). Originating from ancient retroviruses, full-length IAP elements (~7 kb long) contain 5’ and 3’ LTRs that flank the viral gag, prt and pol genes for reverse-transcription and re-integration within the genome. While the presence of these genes makes the IAP retroviral elements autonomous, i.e. they carry all necessary factors for their own transposition, murine IAP are not infectious due to the absence of the env (envelope) gene. However, full-length IAP elements can produce virus-like intracisternal A-particles (IAP),

![Diagram of murine IAP LTR retrotransposons](image)

**Figure 1.2.** Structure of murine IAP LTR retrotransposons. Full length IAP LTR elements carry their own the viral gag, prt, and pol genes for reverse-transcription and re-integration into the genome. The three viral genes are flanked by a 5’ and 3’ LTR that usually serve as the promoter and polyadenylation signal respectively.
visible within the cisternae of the endoplasmic reticulum, thus lending the nomenclature for these elements [13]. The IAP LTR harbors the viral 5’-U3-R-U5-3’ LTR structure [14], with transcription typically starting at the beginning of R in the 5’ LTR, where the “CAT” and “TATA” boxes serve as the RNA polymerase II promoter; the polyadenylation signal in the 3’ LTR is responsible for transcriptional termination. Occasionally, RNA polymerase II may read through the polyadenylation sequence into the neighboring genomic regions, and create fusion transcripts. With YY1 (Ying-Yang 1), SP1 (Specificity Protein 1), AP1 (Activator Protein 1) and other transcription factor-binding motifs on the LTR, fusion transcripts may also be created if the LTR acts an alternative promoter for a neighboring gene through its bidirectional promoter activity [13, 15-17]. Such fusion transcripts, triggered by the nearby LTR, often have disrupted open reading frames (ORF), resulting from an extended 5’ UTR (untranslated region), potentially leading to adverse effects on gene function. Therefore, mechanistically, IAP LTR retrotransposons harbor the ability to functionally disrupt the host genome in more ways than just by active retrotransposition (Figure 1.3).

While LTR elements comprise the main source of retrotransposition activity in rodents, certain subfamilies of the SINE element Alu are the major cause for concern in human genomes. Approximately 300 bp-long, these derivatives of the 7SL RNA gene (a component of the signal recognition particle) derived their name from the presence of AluI restriction enzyme recognition sites within their sequence [18, 19]. With a bipartite structure, linked by a middle A-rich region, and ending with a poly(A) tail (Figure 1.4A), Alu elements do not code for any ORF [11]. Hence, these elements are non-autonomous, meaning they require the retrotransposition machinery of other autonomous retrotransposons in order to retrotranspose [20]. Nonetheless, with over 1 million copies, Alu elements are the most successful retrotransposon in the human
genome, in terms of copy number, and comprise ~10% of its total mass [2]. Being primate-specific, these retrotransposons are estimated to have originated approximately 65 million years ago, before the radiation of primate species [21]. While members of the older Alu families (e.g. AluSx and J) are shared between humans, apes and monkeys, newer Alu elements (such as those belonging to the “young” AluY family) are shared only between humans and their closest ape ancestors. Interestingly, the AluY family has undergone a massive expansion specifically in humans, yielding various subfamilies (such as AluYa5, Yb8, Ya5a2, Ya8, Yb9, Yc1 and Yc2) which are human-specific. In fact, several members of these Alu subfamilies have integrated so recently in the human genome that not all humans share each individual insertion. Depending on the “young” Alu subfamily, the insertion polymorphism in humans can range from 20-80% [20]. Indeed, it is these “young” Alu which are currently active in humans that pose much of the threat to the genome through retrotransposition activity [9].
Active Alu elements are believed to retrotranspose using a mechanism primarily used by LINEs, called Target-Primed Reverse Transcription (TPRT) [20, 22]. Before TPRT can occur, Alu elements are first transcribed from their internal RNA polymerase III promoter (containing the A- and B-boxes). The transcript then extends through the length of the Alu element, and terminates at the poly(T) RNA polymerase III transcription termination signal located downstream of the Alu locus (Figure 1.4AB). Enzymatic machinery encoded by LINE elements cleave the bottom strand of the DNA after a poly(T) sequence, as shown by the carat (^): 5’-TTTT^AA-3’, which then anneals to the poly(A) sequence within the Alu RNA. The exposed 3’ hydroxyl group then allows the Alu RNA to be reverse-transcribed, and re-integrated into the genome following a second, staggered nick on the opposite strand, and filling in of the gaps, which in turn lead to the target site duplications (TSD) that flank the Alu elements [20, 23].
Using the TPRT mechanism, Alu retrotransposition can cause considerable harm to the human genome if it lands close to regulatory regions or within an active gene, and disrupts the expression level, ORF or proper splicing of the gene (Figure 1.5) [9]. Alu elements contain several putative transcription factor binding sites [24], such as those for YY1 [25], ER (Estrogen Receptor) [26], NF-κB (Nuclear Factor kappa-light-chain-enhancer of activated B cells) [27] and p53 (Tumor Protein 53) [28], and thus possess the ability to affect the expression of nearby RNA polymerase II transcribed genes by acting as an alternative promoter or enhancer (Figure 1.5A). Moreover, the poly(A) stretches of Alu elements are particular well-suited to mutating to the polyadenylation signal (AATAAA), thus engendering the potential to introduce alternative transcription termination sites for nearby genes (Figure 1.5CD). Therefore, Alu elements, like IAP LTRs described before, can continue to influence the host genome, even when not actively transposing.

Figure 1.5. Alu elements can disrupt gene expression. Alu elements (vertical black bars) near enhancer (A) or promoter (B) regions may affect the expression pattern of the genes; in the genic regions, they can disrupt the open reading frame (ORF; C) and/or interfere with proper splicing (D). This figure has been adapted from Figure 1A in P.L. Deininger, M.A. Batzer, Alu repeats and human disease, Molecular Genetics and Metabolism, 67 (1999) 183-193.

In fact, since Dr. McClintock’s first description of how mobile genetic elements can affect gene expression and thereby the phenotype, there have been numerous such examples observed in mammals, indicating that the potential for retrotransposons to affect mammalian genomes is not
just theoretical. Studies in mice involving IAP LTRs demonstrated how these elements can affect gene expression that result in outward phenotypes such as coat color variation (for the $A^{vy}$ allele; agouti viable yellow) and tail-kinkedness ($Axin^f$ locus; Axin-fused) [29, 30]. Similarly, a human-specific endogenous retrovirus, associated with the $PRODH$ (Proline Dehydrogenase 1) gene, has been shown to behave as an enhancer that increases the expression of $PRODH$ in the hippocampus of the brain [31]. As such, it may have contributed to the evolution of the central nervous system in humans. While there are such documented instances where retrotransposons have been co-opted into the host genomes [17, 32], many insertion events are associated with negative outcomes. In fact, ~0.3% of all human diseases could be attributed to the effect of transposable elements [33]. Specifically, ~25 L1 and ~60 Alu insertions have been implicated in various human diseases including multiple types of cancers [7]. As discussed in detail previously regarding IAP LTR and Alu elements, retrotransposons within or close to genes can interfere with endogenous promoter activity, splicing, transcription initiation/termination signals and can even cause deletions resulting from unequal recombination events. Such disruption of gene expression can be very deleterious, as seen in the cases of $BRCA1$ and 2 (breast cancer), $FVIII$ and $IX$ (hemophilia A and B), $NF1$ (neurofibromatosis) and many others [7, 9, 34]. Thus, given the potential for retrotransposons to result in so many disease conditions, it has been important for mammalian genomes to find ways to silence their activity.

**Epigenetics for “Host Defense”**

The term “epigenetics” was first coined by Conrad Waddington in 1942 to explain the role of the environment in embryonic development. Since then, the field has emerged to encompass much more, not only with regards to development, but also in the context of cellular differentiation and integrity. To reflect that, the term “epigenetics” is currently defined as all modifications which
result in the heritable change in gene expression without changing the underlying DNA sequence. However, even that definition falls short of acknowledging one of the major roles of epigenetic modifications in the mammalian genome: silencing retrotransposons [35].

The four major epigenetic mechanisms – DNA methylation, histone modifications, RNA interference and chromatin remodeling – are all involved, and function with extensive cross-talk between each other, to silence retrotransposition activity in mammalian genomes at various developmental stages [36, 37]. While DNA methylation is the major epigenetic mechanism repressing retrotransposons in differentiated cells, repressive histone modifications, such as H3K9me3 (trimethylation of lysine 9 on histone H3), act independently of it in mouse embryonic stem cells to silence retrotransposons such IAP LTRs [38-41]. Similarly, H3K9me3 has also been shown to play a major role in the silencing of Alu transcription from its endogenous RNA polymerase III promoter in HeLa cells [42]. Various studies have also explored the role of specific proteins, such as chromatin remodelers and histone methyltransferases, which mediate the cross-talk between DNA methylation, histone modifications and RNA interference pathways in order to keep the retroelements silenced [43-46]. Overall, these studies underscore the importance of epigenetic mechanisms as “host defense” systems in mammalian genomes.

Of the four epigenetic mechanisms, DNA methylation is perhaps the one that is best characterized in the context of retrotransposon silencing. DNA methylation in mammals is mainly observed on the cytosines that occur directly 5’ guanines (known as CpG sites), converting them to 5-methylcytosines [47]. As early as the 1990s, it was noted that most CpG sites in mammalian genomes lie within retroelements (with about a quarter in Alu elements [48]), and that for the most part, they are heavily methylated [49]. DNA methylation was correlated with transcriptional silencing through the activity of various transcription factors, such
as MeCP1 and MeCP2 (Methyl-CpG binding protein 1 and 2) [50, 51]. Around the same time, it was also first noted that certain retrotransposons could trigger de novo methylation [52] and that DNA methylation could be responsible for silencing retrotransposon activity [13, 53-56]. Based on this evidence, it was proposed that epigenetic mechanisms such as DNA methylation may have evolved for “host defense” against parasitic sequences [49]. Consistent with this hypothesis, phenotypic alterations in mice due to IAP LTR activity at the A\textsuperscript{\textregistered} and Axin\textsuperscript{\textalpha} loci were reversed upon methylation of the IAP LTRs [29, 30]. Similarly, the HERV near the PRODH locus was unable to influence the expression level of the gene in human brain when it was methylated [31].

DNA methylation in mammals is carried out by three main enzymes. DNMT1 (DNA Methyltransferase 1) is the canonical “maintenance” methylation enzyme, responsible for the methylation of hemimethylated DNA during replication. DNMT1 is recruited to hemimethylated CpG sites by another protein, UHRF1/NP95 (Ubiquitin-like with PHD and Ring Finger domains 1/Nuclear Protein 95), which in turn binds replication-associated histone modifications [57]. As the major DNMT expressed in somatic cells, the role of DNMT1 in retrotransposon silencing is highlighted by the fact the 5’ UTR of L1 elements are heavily methylated when their transposition is suppressed; conversely, hypomethylation is associated with mobility [58-60]. Maintenance of DNA methylation on retrotransposons by DNMT1 even in embryonic stem (ES) cells is crucial, as shown by the fact that embryos lacking proper DNMT1 function show high levels of IAP and L1 hypomethylation and increased IAP expression [56, 61, 62]. DNMT3A and 3B (DNA Methyltransferase 3A and 3B) are the canonical de novo methyltransferases [63], active only during early embryogenesis, and are responsible for restricting the evolutionary propagation of retrotransposons. An interacting partner of DNMT3A/3B, DNMT3L (DNA Methyltransferase 3-Like), is required for the proper methylation of retrotransposons by the de
*de novo* methyltransferases in the male germline; absence of this protein results in hypomethylation of both LTR as well as non-LTR retrotransposons [64-66]. Although they share interacting partners, the two *de novo* methyltransferases have somewhat non-overlapping functions in terms of retrotransposon methylation. SINE elements (like B1 in mice) are expected to be methylated by DNMT3A, whereas both DNMT3A and 3B work concurrently to methylate LINEs and LTR elements (like L1 and IAP) [37]. Interestingly, IAP elements were not hypomethylated in embryos lacking DNMT3A function, but they were found to be somewhat hypomethylated in those without DNMT3B activity [63]. Overall, these data highlight both the functional separation as well as the partial overlap between the two *de novo* methyltransferase enzymes.

Despite the long-held “two-step” model for establishment and maintenance, recent findings suggest that DNMT1 and DNMT3A/3B work together to establish both *de novo* as well as maintenance methylation, adding to the complexity of the mechanisms by which retrotransposons are methylated [67, 68]. ES cells lacking both DNMT3A/3B enzymes (the canonical *de novo* methyltransferases) nonetheless showed methylation and suppression of LTR elements (especially IAP elements), with only a functioning DNMT1 (the canonical “maintenance”) enzyme. Combined with previous data, these new data suggest that DNMT1 (in conjunction with UHRF1/NP95) has a larger role in both *de novo* as well as maintenance methylation of LTR elements, with some variation across families and subfamilies of the retroelements. On the other hand, deletion of only the canonical “maintenance” enzyme, DNMT1, resulted in retention of DNA methylation at LINE promoters but not LTR elements. This suggests that the canonical *de novo* methyltransferases (DNMT3A/3B) also have some maintenance activity at the LINEs. Therefore there is an intriguing division of labor between the DNMT1 and DNMT3A/3B in ES cells: DNMT1 appears to be more important for methylation at
LTR elements, whereas DNMT3A/3B are crucial for LINE methylation. Taken together, these studies establish that all three DNA methyltransferase enzymes are critical for the proper methylation and silencing of mammalian retrotransposons.

**Research Objectives**

Despite our current broader understanding of how DNA methylation is established and maintained at retrotransposons, a deeper understanding of the DNA methylation status at individual retrotransposon loci has remained limited. Yet the importance of such an understanding cannot be overstated in the light of various reports describing the hypomethylation of repeat elements in many human cancers. It is further highlighted by a recent meta-analysis which concluded that hypomethylation of retrotransposons is likely to be associated with a poor tumor prognosis [69]. However, high sequence degeneracy and short sequencing reads have prevented the locus-specific analysis of individual retrotransposons on a genome-wide scale. Hence, so far they have only been studied in bulk (using restriction digestion by methylation sensitive enzymes followed by Southern blot [70]) or a few loci at a time (using techniques like individual locus bisulfite-sequencing [71]). As a result, two important questions have remained unanswered:

1. How does the methylation level of individual retrotransposon loci compare among different cell states, such as somatic, pluripotent (ES) and cancer cells?

2. Is the hypomethylation observed in cancer cells occurring uniformly across all loci of a targeted retrotransposon, or is it a more locus-specific event?

Addressing these fundamental questions is crucial to gaining a better understanding of the epigenetic landscape of different cell types. Given the current interest in developing the
methylation level of retrotransposons as cancer biomarkers [69, 72], the answers to these questions would provide a significant gain in resolution that could better inform our search towards specific loci whose methylation level can help detect and track cancer progression.

In order to address these questions, we have used a newly developed technique called High-Throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS) [73, 74], described in detail in chapter two (in the context of the study described in chapter three). Briefly, this technique involves ligation of methylated adaptors compatible with a high-throughput sequencing platform (that is capable of ≥ 300 bp long reads) to sheared and end-repaired DNA. The adaptor-ligated DNA is then subjected to bisulfite conversion and amplification using adaptor- and retrotransposon subfamily-specific primers. The amplified DNA is then size-selected and sequenced on a high-throughput sequencing machine to generate millions of reads. These reads are mapped to a custom-prepared library of retrotransposons along with some flanking unique sequence. In order to obtain accurate locus-specific information, a predetermined number of bases must match both the retrotransposon as well as its unique sequence. The mapped reads are then processed to yield DNA methylation information for individual retrotransposon loci of the targeted subtype.

Using this technique, Ekram and Kim previously reported the various methylation patterns of a small fraction of IAP LTR loci which escape DNA methylation in a set of normal tissues [15, 73]. Based on this study, we hypothesized that the hypomethylation of retrotransposons in different cell states is likely to be a locus-specific event as well. We tested our hypothesis in two model systems, mouse IAP LTR (chapter three) and human Alu elements (chapter four), using the HT-TREBS technique. In both model systems, we have been successful in analyzing the
DNA methylation patterns of over 5000 individual loci of the targeted retrotransposon on a genome-wide scale.

Research Objectives I (chapter three):

**Objective IA.** Establish the level of DNA methylation at IAP LTR retrotransposons in murine ES and cancer cells, compared to somatic cells.

**Objective IB.** Determine whether IAP LTR elements are hypomethylated in a uniform or locus-specific manner in ES and cancer cell states.

The results from the IAP LTR methylation study revealed that ~25% loci escape DNA methylation in ES cells, a number that increases up to 50% in Neuro2A cancer cells, though <5% remain hypomethylated in somatic cells (Objective I). Moreover, the results suggested that not all IAP LTR loci tend to be equally sensitive to cell state in terms of their DNA methylation levels (Objective II). While about half of the loci tested remained stably methylated in all three cell types, the other half fluctuated in various different patterns with a majority of them being hypomethylated in Neuro2A cancer cells [75]. Overall, the data highlighted the locus-specific nature of DNA hypomethylation at IAP LTR elements.

Research Objectives II (chapter four):

**Objective IIA.** Establish the level of DNA methylation at two human-specific Alu retrotransposon subfamilies, AluYa5 and AluYb8, in normal somatic cells.

**Objective IIB.** Determine the fraction of AluYa5/Yb8 that would be expected to lose DNA methylation in response to a cancer cell state.

**Objective IIC.** Evaluate the extent to which the variation in DNA methylation could be attributed to tumorigenesis as compared to inter-individual variation.
According to the results, ~90% of the Alu repeats are highly methylated in human fibroblast cells; but the ~10% that escape DNA methylation are often located close to gene promoters and show high levels of variation in DNA methylation. The DNA methylation variation was characterized in the context of tumorigenesis and inter-individual differences, revealing that a small subset of Alu loci may respond early and specifically to tumorigenesis, albeit in the background of high levels of inter-individual variation [76]. Overall, these results corroborated those from the IAP LTR study (chapter three), highlighting the locus-specific nature of DNA hypomethylation at retrotransposons in various cell states.

Taken together, these results from mouse and human retrotransposons emphasize the non-uniform nature of DNA methylation in different cell states. While some retrotransposons remain highly methylated regardless of cell type, others are likely to show detectible levels of difference in DNA methylation even upon subtle changes. This information gives us the opportunity to seek specific retrotransposon loci which are uniquely hypomethylated in different types and stages of cancer. Techniques such as HT-TREBS make such a search feasible by allowing multiple individuals to get sequenced for multiple loci so that the inter-individual variation of certain loci could be taken into account. The specific elements that then show predictable changes only in response to cell state, but do not vary significantly among multiple individuals, can then be developed as excellent epigenetic biomarkers for diagnosing and tracking the progression of cancer.

References


CHAPTER TWO

LOCUS-SPECIFIC DNA METHYLATION ANALYSIS OF RETROTRANSPOSONS IN ES, SOMATIC AND CANCER CELLS USING HIGH-THROUGHPUT TARGETED REPEAT ELEMENT BISULFITE SEQUENCING

Introduction

DNA methylation is a major epigenetic mark associated with multiple aspects of retrotransposons within the mammalian genome. In order to study DNA methylation of a large number of retrotransposons on an individual-locus basis, we have developed a new protocol termed High-Throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS) [1]. We have used this technique to characterize the locus-specific patterns of DNA methylation of 4799 members of the mouse IAP LTR (Intracisternal A Particle Long Terminal Repeat) retrotransposon family in embryonic stem, somatic and Neuro2A cells [2], as described in chapter three. In this chapter, we describe in detail the sample preparation and bioinformatic analyses used for these studies. Briefly, we ligated methylated Ion Torrent “A” adaptors to sonicated genomic DNA, and amplified it after bisulfite-conversion using a primer scheme that was designed to select for five IAP LTR subfamilies. With the “P1” adaptor being incorporated into the amplification products through a primer extension, the resulting fragments were then sequenced on the Ion Torrent PGM machine. We have then used a custom-designed pipeline to analyze the locus-specific methylation levels of IAP LTR in the three cell states.

1 This chapter first appeared as A. Bakshi, M.B. Ekram, J. Kim, Locus-specific DNA methylation analysis of retrotransposons in ES, somatic and cancer cells using High-Throughput Targeted Repeat Element Bisulfite Sequencing, Genomics Data, 3 (2015) 87-89.
Experimental Design, Materials and Methods

Sample Preparation: DNA Isolation and Library Construction

Whole brain, liver and kidney from a 1-week-old male C57BL6/N mouse were lysed overnight in 10X w/v Tail Lysis Buffer (100 mM Tris-Cl [pH 8.5], 5 mM EDTA [pH 8.0], 200 mM NaCl, 0.2% w/v sodium dodecylsulfate [SDS]) and 0.01X v/v 20 mg/mL Proteinase K (Sigma-Aldrich). The tissue-lysis products were divided into 300 µL aliquots and stored at -20°C. One 300 µL aliquot was used to isolate DNA using phenol-chloroform-isoamyl alcohol followed by ethanol precipitation. A similar DNA isolation protocol, involving cell lysis followed by phenol-chloroform extraction and ethanol precipitation, was implemented on ES and Neuro2A cell extracts as well. The isolated DNA was resuspended in 50-100 µL 1X TE and its concentration quantified using the Nanodrop (Thermo-Scientific).

Approximately 1 µg of the isolated DNA was sonicated using the Bioruptor NGS (Diagenode) to obtain fragments which were approximately 700 bp in length (4 cycles, on/off cycle time: 15”/90”) (Figure 2.1AB). Next, these fragments were end-repaired using NEBNext® End Repair Module (New England BioLabs), and cleaned using the DNA Clean and Concentration Kit (Zymo Research) with 5X v/v Binding Buffer and eluted in 30 µL HPLC water. The end-repaired DNA was then incubated at 20°C for 2 hours with 50 pmols of custom-made methylated-C Ion Torrent “A” adaptors (Integrated DNA Technologies) and 800 units of T4 DNA Ligase (New England Biolabs) (Figure 2.1C). Unligated adaptors were removed using the DNA Clean and Concentration Kit (Zymo Research) with 5X v/v Binding Buffer and the resulting “A” adaptor-ligated fragments were eluted in 50 µL HPLC water. The adaptor-ligated DNA fragments were further size-selected to remove any excess adaptors and DNA fragments smaller than 300 bp in length using the Agencourt AMPure XP beads (Beckman Coulter) using
Figure 2.1. Library preparation for HT-TREBS. The isolated DNA (A) is subjected to sonication (B), to yield ~700 bp fragments, and end-repaired before methylated-C Ion Torrent “A” adaptor-ligation (C). Following one round of size selection (D), all fragments >300 bp are bisulfite treated (E) and PCR amplified. The cycle number for PCR was determined for each individual library to be the one that corresponds to the midpoint of the exponential portion of the amplification curve from qPCR (F). Finally, the amplified library was size selected for 250-300 bp insert size, quantified by Bioanalyzer and sequenced on the Ion Torrent PGM platform (G). The color codes within this figure is as follows: yellow bars indicate unique sequence, blue bars represent IAP LTRs, green and gray boxes indicate the Ion Torrent “A” and “P1” adaptors respectively and orange arrows represent the primers used for amplification.
DNA:bead ratio of 1:0.7 (100 µL DNA + 70 µL beads) (Figure 2.1D). This library was then quantified using Bioanalyzer (Agilent) and subsequently subjected to one round of bisulfite conversion (Figure 2.1E) using the EZ DNA Methylation Kit (Zymo Research). This bisulfite-converted library (1 µL) was then used to conduct quantitative real-time PCR using SYBR-Green (Life Technologies) to determine the appropriate number of cycles to amplify the bisulfite-converted library for subsequent size selection and sequencing. The cycle number for PCR amplification was individually determined for each library to be the one corresponding to the midpoint of the exponential portion of the amplification curve (Figure 2.1F). The forward primer used for PCR was complementary to the 5’ end of the “A” adaptor region (5’-CCATCTCATCCCTGCTGTCTCCGACTCAG-3’). The reverse primer (5’-CCACTACGCCTCCGCTTTTCTCTATGGGCA GTCGGTGA^CTCCCTATTAACTAACCCATC-3’) was designed to bind the 24-bp region within the U3 portion of the LTR which is conserved between five subtypes of IAP LTR (IAPLTR1, IAPLTR1a, IAPLTR2, IAPLTR2a, and IAPLTR2b), and is devoid of any CpG sites. The sequence at the 5’ end of the carat (^) corresponds to the “P1” adaptor which is a part of the amplification scheme used by Ion Torrent (Life Technologies) (Figure 2.1F). The amplified libraries were then size-selected using agarose gel extraction (MEGAquick-spin™ Total Fragment DNA Purification Kit, Intron Biotechnology) to have approximately 250-300 bp insert length flanked by the “A” and “P1” adaptors. This was achieved by excising the gel fragment corresponding to 330-380 bp in order to account for the ~80 bp combined-length of the two adaptors (Figure 2.1G). The size-selected library was then quantified using Bioanalyzer (Agilent Technologies). Approximately 25 µL of the size-selected library at 10 pM was used for the emulsion PCR and subsequent next-
generation sequencing using the Ion Torrent Personal Genome Machine (PGM) Sequencer and Ion 318 chips (Ion Torrent, Life Technologies).

**Bioinformatics Analyses**

We have implemented the following bioinformatics pipeline to process the raw sequence reads from the NGS platform (Figure 2.2). We have used Bowtie2 [3] for mapping and BiQ analyzer HT [4] for DNA methylation analyses. Several Unix-based command lines were used, along with custom-made Perl scripts, for the pipeline (Figure 2.2).

First, we used the following steps for the construction of a custom database (below and Supplemental Material 2.1)².

1) Download the sequence of 9282 IAP LTRs containing 330-bp LTR along with two 350-bp flanking regions, in fasta format, from the Table database at UCSC genome browser. Reverse complement half of the IAP LTRs in an opposite direction so that the entire set will have an identical forward direction with a custom Perl script.

2) Convert all the IAP LTRs into bisulfite-converted sequences with a custom Perl script.

3) Compile all the bisulfite-converted sequences into one large sequence with a Unix command line.

4) Construct a searchable index file with a large compiled sequence with bowtie2-build.

The execution of these steps will provide the following files: (i) a directory containing the 9282 files with the original sequences in a fasta format (Step 2) and (ii) a directory containing 6 indexable file (Step 5). These two directories will be used for the analyses that follow.

---

² All supplemental materials are available online with the original publication at journal’s website.
Figure 2.2. Bioinformatics workflow for HT-TREBS. All the major steps in the HT-TREBS data analysis have been shown in a stepwise fashion, including custom database construction, processing sequenced reads, processing the mapped reads and finally, methylation analysis using BiQ Analyzer HT [4]. Precise information regarding each step of the workflow may be found in Supplemental Material 2.1, along with the custom Perl scripts used to execute them. The box and text in red indicate the attainment of the main result (text) file which can then be imported in Excel. Filled boxes indicate the major phases of the data processing pipeline whereas the unfilled boxes indicate the steps performed for each phase. Unfilled lines and arrows indicate files from the respective steps which feed into the next major phase of the pipeline. Filled arrows are used elsewhere to indicate the dataflow.
Second, we used the following steps for the mapping of sequence reads and subsequent DNA methylation level analyses (below and Supplemental Material 2.1).

1) Remove raw sequence reads smaller than 40 bp in length.
2) Map the raw reads against the custom-made database using Bowtie2 [3].
3) Remove headerlines from the sam file derived from mapping with a Unix command line.
4) Sort each line of the sam file based on each locus of IAP LTR and generate a directory containing each IAP LTR file containing mapped raw reads with a custom Perl script.
5) Filter out raw reads from each IAP LTR file that are not qualified based on the insufficient overlap with the flanking unique region (greater than 10 bp in length) with a custom Perl script.
6) Execute BiQ Analyzer HT [4] with two directories, the directory containing all the sequences of IAP LTRs (Step 2 of database construction) and the directory containing all the mapped bisulfite sequence reads, with a custom Perl script.
7) Extract the necessary information (number of reads used, % methylation, standard deviation) from the output of BiQ Analyzer HT [4] with a custom Perl script.

The execution of these steps will finally derive one text file containing the DNA methylation level for the entire set of IAP LTRs, which can be imported into an excel file for further calculation and inspection (Figure 2.2). This series of bioinformatics analyses require multiple Unix command lines and custom Perl scripts. We have provided these scripts and command lines as Supplemental Material 2.1

\[\text{\textsuperscript{3}}\] The Perl scripts and Unix command lines used for the HT-TREBS data analysis are released under the GNU General Public License, and can be accessed from the publisher’s website.
References


CHAPTER THREE

RETROTRANSPOSON-BASED PROFILING OF MAMMALIAN EPIGENOMES: DNA METHYLATION OF IAP LTRS IN EMBRYONIC STEM, SOMATIC AND CANCER CELLS

Introduction

Mammalian genomes have accumulated a large number of retrotransposons during evolution, making up about half of the genome in any given species [1, 2]. These retrotransposons are usually repressed by two main mechanisms. In germ cells, DNA methylation is the major mechanism which represses the majority of retrotransposons [3, 4]. Later, in fertilized eggs, the germ cell-driven DNA methylation is removed, and subsequently, retrotransposons are temporarily repressed by various histone modifications [5-9]. During the implantation stage, however, these transient histone marks are again replaced by DNA methylation, a more stable and permanent modification. Thus, histone modifications are usually responsible for the temporary repression of retrotransposons in transient stem cell populations whereas DNA methylation is responsible for the more stable and permanent repression in further committed and differentiated cell populations [4, 10, 11].

It is well known that improper epigenetic regulation of retrotransposons can result in unwanted consequences due to their transposition [12] as well as their ability to influence adjacent gene activity by acting as alternative promoters [13] or disrupting the endogenous exon structure. The latter has been well demonstrated by the two cases of mouse epialleles, where the ectopic expression of two independent IAP LTRs (Intracisternal A Particle Long Terminal Repeats)

1 This chapter first appeared as A. Bakshi, J. Kim, Retrotransposon-based profiling of mammalian epigenomes: DNA methylation of IAP LTRS in embryonic stem, somatic and cancer cells, Genomics, 104 (2015) 538-544.
interfere with the transcription of two endogenous loci, $A^v$ (agouti viable-yellow) and $Axin^{fu}$ (Axin-fused). In the mouse, these ectopic expressions are responsible for visible phenotypic consequences, such as coat color variation for $A^v$ and tail kinkedness variations for $Axin^{fu}$ [14, 15]. Similar hypomethylation on retrotransposons is also seen in cancer genomes in humans [16, 17], which may result in functional consequences for cancer progression [18, 19]. However, at the same time, hypomethylation of retrotransposons also provides a unique opportunity for mammalian genome evolution through allowing their co-option into the host genome as regulatory elements [20]. A recent study suggests that endogenous retroviruses (ERVs) in the human genome may have played an important role in the evolution of the central nervous system by affecting gene expression in key areas of the brain [21]. Overall, DNA methylation at retrotransposons is not only of interest towards understanding complex human disorders with an epigenetic underpinning, including cancers [22, 23], but also for their role in shaping mammalian evolution.

Given the importance of epigenetic modifications on retrotransposons, much research is currently focused on understanding the extent to which they are repressed in normal as well as diseased cells and tissues. However, due to the paucity of methods allowing for deep sequencing coverage at repeat elements and flanking loci, it has thus far been difficult to perform a systematic analysis of the effect of DNA methylation at individual loci in a locus-specific manner. To address these issues, we have developed a new protocol termed HT-TREBS (High-Throughput Targeted Repeat Element Bisulfite Sequencing) [24], which is designed to survey the DNA methylation levels of a large number of interspersed repeat elements on an individual-locus basis. According to the results from a set of pilot experiments analyzing mouse IAP LTRs, the majority of this retrotransposon family is properly repressed in normal cells, but a very small
number of IAP LTRs (about 5%) escape this repression mechanism [24, 25]. In the current study, this protocol has been utilized to characterize the DNA methylation profile of this retrotransposon family in AB2.2 embryonic stem and Neuro2A cancer cells, and subsequently compared with the somatic cell data derived from Ekram and Kim [24]. The results revealed that the DNA methylation levels of a large fraction of IAP LTRs, though not all, are dynamically fluctuating between different cell types. Further, the methylation variation at a fraction of those loci are easily detectible by independent restriction enzyme-based techniques. Overall, these analyses demonstrate the feasibility of using HT-TREBS to find potential epigenetic biomarkers of cell state in multiple different cell types.

Results

HT-TREBS analyses of IAP LTRs in ES, somatic and Neuro2A cells

In the current study, we analyzed and compared the DNA methylation levels of the IAP LTR family using the following three cell types: embryonic stem (ES), somatic and Neuro2A cancer cells. Genomic DNA isolated from ES cells (AB2.2 cell, 129 origin, from Baylor College of Medicine) and a neuroblastoma cell line of strain A origin (Neuro2A) were treated according to an established protocol of HT-TREBS [24]. In brief, the genomic DNA was fractionated by sonication, ligated to an adaptor and modified with bisulfite conversion protocol [26]. It was then amplified with a PCR scheme designed to enrich IAP LTR-genomic regions. This was accomplished by designing one of the primers complementary to the 24-bp portion within the U3 region of the LTR that is conserved between five subtypes of IAP LTR (IAPLTR1, IAPLTR1a, IAPLTR2, IAPLTR2a, and IAPLTR2b), and is devoid of any CG dinucleotides. The region is located approximately 150 bp downstream of TGTTGGG (denoting the extreme 5’ end of the
LTR), and includes the CAT box. The amplified libraries were finally size-fractionated with agarose gel electrophoresis, and sequenced with a Next-Generation-Sequencing (NGS) protocol.

We have obtained 6.3 and 6.8 million raw sequence reads from ES and Neuro2A cells, respectively. Each set of bisulfite-converted sequence reads were mapped and processed to derive the methylation level of each IAP LTR locus as described below. In brief, individual raw sequence reads were mapped to a custom database containing the sequences of about 10,000 IAP LTR loci. In this database, each locus is represented with a 1,030-bp sequence covering the 330-bp LTR plus two 350-bp flanking regions. A set of sequence reads mapped to a given locus was subsequently used for calculating its methylation level. We have successfully obtained the methylation values for 5637 and 5575 loci of IAP LTRs for ES and Neuro2A cells respectively. In both sets, 100 mapped raw sequences were used on average to derive methylation level of each IAP LTR locus, covering an average of 7 CpG sites per locus.

Next, these two new data sets were compared with the somatic data set representing the average methylation values of IAP LTRs that were individually derived from the three organs (brain, liver and kidney) of a two-month-old C57BL/6N mouse (Figure 3.1) as well to the data sets derived from each of the organs representing the three tissue layers individually [24]. According to our analyses with the individual tissues (Supplemental Figure 3.1^2), liver and kidney showed overall more similarity in DNA methylation levels between each other than with the brain. However, on the whole, over 95% of all IAP LTRs were methylated at similar levels in all three organs, as observed by Ekram and Kim [24]. Therefore, in order to compare ES and Neuro2A cells to a representative DNA methylation value in normal somatic cells, all further analyses

---

^2 All supplemental materials are available online with the original publication at the journal’s website.
were performed with the average somatic data set. Finally, the three data sets (ES, somatic and Neuro2A) were compiled to derive 4799 IAP LTRs with a minimum of 15X coverage at each CpG site that were represented in all three cell types for further comparisons (Supplemental Data 3.1 and Figure 3.1).
Initial tabulation of these data sets revealed the following immediate conclusions. The overall DNA methylation levels of IAP LTRs are variable among the three samples: the DNA methylation levels in somatic cells are the highest, followed by ES and Neuro2A cells. This was indicated by the result of COBRA (Combined Bisulfite Restriction Analysis [27]), showing a substantial level of DNA hypomethylation in Neuro2A cells as compared to the much greater levels of DNA methylation in both ES and somatic cells (Figure 3.1A). Next, in order to quantify the level of DNA methylation difference, the set of 4799 loci was divided into two groups: hypermethylated (≥ 80% methylation level) and hypomethylated (< 80% methylation level). IAP LTRs methylated above 80% showed no major differences amongst them, whereas those under 80% methylation level showed some unique patterns of hypomethylation, as discussed previously by Ekram and Kim [24]. The same was largely true for thresholds under 80% (70-75%), however the number of loci in the hypomethylated category were too few for meaningful comparisons with the hypermethylated category under these conditions. Therefore, the empirically selected 80% threshold, which is consistent with that established by Ekram and Kim [24], ensured an adequate number of loci in each group while staying close to the base of the curve shown in Figure 3.1B. According to the results, about 96% of IAP LTRs (4587 loci) showed at least 80% methylation level in somatic cells whereas only 77% (3689 loci) and 51% (2465 loci) were hypermethylated in ES and Neuro2A cells respectively (Figure 3.1BC). In somatic cells, only 4% of the loci (212 IAP LTRs) belonged to the hypomethylated category, while 23% of IAP LTRs (1110 loci) and nearly half (2334 loci) showed less than 80% methylation levels in ES and Neuro2A cells respectively (Figure 3.1BC). Overall, these initial analyses demonstrate the severe hypomethylation of IAP LTR loci in Neuro2A cells which is in stark contrast with the much higher levels of methylation seen in ES and somatic cells.
Comparative analyses of DNA methylation profiles in ES and Neuro2A vs. somatic cells

The DNA methylation pattern of each individual IAP LTR locus in ES and Neuro2A cells was analyzed in comparison to somatic cells using a two-dimensional dot plot display (Figure 3.2). In a given plot, each dot represents one IAP LTR with two values: X- and Y-axis values representing DNA methylation levels derived from two comparing cell types. As shown in Figure 3.2A, the majority of dots (IAP LTRs) are located within the 80-100% methylation range in both ES and somatic cells. Yet, a substantial fraction of IAP LTRs are spread horizontally over the methylation ranges less than 80% in ES cells but still greater than 80% in somatic cells. In terms of actual number, this group is estimated to contain around 932 IAP LTRs based on the initial tabulation (Figure 3.1C). A small number of IAP LTRs also show even greater degrees of DNA methylation difference between ES and somatic cells based on their location in the upper

Figure 3.2. DNA methylation profiles of IAP LTRs in ES and Neuro2A cells compared to somatic cells. The methylation levels of IAP LTRs were compared between ES and somatic cells (A) and between somatic and Neuro2A cells (B). For both scatter plot analyses, each dot represents one IAP LTR, and the position of the dot (on the XY plane) indicates the methylation levels from the two cell types that are being compared. The dots spreading horizontally toward left on the first plot indicates that a substantial fraction of IAP LTRs is hypomethylated in ES cells but not in somatic cells. Similarly, a large number of dots spreading horizontally to the left from the right corner of the second graph represents the hypomethylation of half of IAP LTRs in Neuro2A cells, but their methylation levels are still greater than 80% in somatic cells.
left corner of the dot plot where the DNA methylation levels are less than 40% in ES but greater than 80% in somatic cells. Another set of IAP LTRs is located along the diagonal line connecting the lower left to upper right corner. These dots belong to the group of IAP LTRs showing hypomethylation regardless of cell type, whose methylation values, interestingly enough, are also known to be variable between different individuals [24, 25]. Overall, this dot plot analysis highlights the observation that a large portion of IAP LTRs are hypomethylated in ES cells but not in somatic cells (Figure 3.2A).

Comparing DNA methylation patterns of IAP LTRs in Neuro2A and somatic cells (Figure 3.2B) indicate that a much greater number of IAP LTRs, around 2152 elements, show different levels of methylation between the two cell types. A majority of these loci (2137 elements), represented by a large number of dots spreading horizontally to the left from the upper right corner in the dot plot, show greater than 80% methylation in somatic cells but less than 80% methylation in Neuro2A cells. It is interesting to note that all the individual IAP LTRs of this group have different levels of DNA methylation change: some have much greater levels, such as the ones in the top left corner which are nearly completely methylated in somatic cells but completely unmethylated in Neuro2A cells, whereas others have more modest levels of change. This individuality of DNA methylation change is even more contrasting in the case of the other remaining half of the IAP LTRs, which still show similar levels of DNA methylation between somatic and Neuro2A cells. Overall, this analysis indicates that about half, but not all, of the IAP LTRs are hypomethylated in Neuro2A compared to somatic cells. Given all the different degrees and the various patterns of changes, the observed methylation difference between the two cell types is thought to be locus-specific rather than uniform, which may be reflecting the fact that not all IAP LTRs are equally sensitive to cell type in terms of their DNA methylation.
DNA methylation pattern-based grouping of IAP LTRs

The various different methylation levels of each individual IAP LTR in the three different cell types were further studied by Venn diagram analysis using Venny [28]. In this analysis, each IAP LTR has one of the two states in a given cell type, either High (≥ 80%) or Low (< 80%) level of methylation based on the division established from Figure 3.1BC. Since each IAP LTR is present in three different cell types, ES, somatic and Neuro2A, eight different combinations of DNA methylation states are possible for any given IAP LTR (Figure 3.3). The first Venn diagram was constructed through comparing three sets of IAP LTRs that show less than 80% methylation level in all three cell types (Figure 3.3A). This, subsequently, derived seven groups of IAP LTRs, which are represented by the seven sections of the first Venn diagram. The three non-overlapping sections indicate the three groups with one Low state in the three different cell types (LHH, HLH, HHL in Figure 3.3C), whereas the three overlapping sections between two cell types (or circles) indicate the groups with two Low states (LLH, LHL, HLL). One remaining group is overlapped by three Low states (LLL), meaning that this group of IAP LTRs maintains Low methylation in all three cell types. One final group is missing in the first Venn diagram, but is found in the center section of the second Venn diagram that has been constructed with the three sets of IAP LTRs showing greater than 80% methylation level. The final group of IAP LTRs is characterized by greater than 80% methylation levels in all three different cell types (HHH) (Figure 3.3B).

The actual numbers of IAP LTRs constituting each of the eight groups are summarized in Figure 3.3C. According to this summary, there are three groups which are represented by very small numbers of IAP LTRs, which include LLH (11), HLL (30) and HLH (4). Of the remaining five groups, two of them, LLL (167) and HHH (2257), show constant DNA methylation and make up
Figure 3.3. DNA methylation pattern-based grouping of IAP LTRs. A given IAP LTR has one of two states, either Low (<80% methylation) or High (≥80% methylation) in each of the three cell types, resulting in eight total possible categories of DNA methylation patterns. The first Venn diagram was constructed through comparing the three sets of IAP LTRs showing less than 80% methylation level (Low) in the three cell types (A). The second Venn diagram was derived from comparing the three sets of IAP LTRs with their methylation levels being at least 80% (B). The resulting eight different categories of IAP LTRs are presented along with the actual numbers of loci in (C).

3.5 and 47% of the entire IAP LTR family, respectively. The other three groups with fluctuating methylation levels altogether make up the remaining half: the HHL group (29%, 1398 members), the LHL group (15%, 739 members) and the LHH group (4%, 193 members).

These IAP LTRs within these Venn diagram groups were further analyzed in order to ascertain any biological significance behind the various patterns of methylation observed for these IAP
LTRs. They were studied with regards to expressed sequence tags (ESTs), histone modifications (data not shown), gene association and genomic position preference along with the 20 kb flanking sequence in order to survey a representative portion of the mouse genome covered by the IAP LTRs (Supplemental Figure 3.2). No specific epigenetic mark distinguished any one of the groups from others. IAP LTRs within all groups were associated with 0-2 genes, albeit only distally, with most IAP LTRs being positioned within 50-500 kb of transcription start sites (TSS) (Supplemental Figure 3.2AB) as previously observed by Ekram and Kim [24]. There was no genomic location preference observed for any these groups of IAP LTRs with respect to their chromosomal location or distance to TSS (Supplemental Figure 3.2BC) which distinguished them from each other, after accounting for the large variation in sample sizes. In summary, according to this series of Venn diagram analyses, the DNA methylation levels of half of the IAP LTRs are static and constant whereas the remaining half fluctuates between different cell types without any gene association or chromosomal position preference.

**COBRA analyses of representative IAP LTR loci**

We employed COBRA as an independent method of assessing DNA methylation at some of the loci representing the various Venn diagram groups. It provided us with a rapid view of the differences between cell types through testing the methylation level associated with 1-3 CpG sites associated with the retrotransposon. Bisulfite-treated brain DNA was chosen as a representative organ for somatic cells.

One representative example of COBRA analysis of the locus IAPLTR1a_Mm-ERVK-LTR chrX: 39388013-39388339, associated distally with the gene Xiap, is shown along with the corresponding HT-TREBS data set (Figure 3.4A). The black triangles over the heatmaps indicate one of the two HpyC4IV restriction sites in the amplified portion of the IAP LTR which was
also sequenced by HT-TREBS. This particular CpG site was methylated in 133 reads out of a total of 191 reads in ES cells (70% methylation), 89 out of 137 in the brain (65% methylation) and 0 out of 187 reads (0% methylation) in Neuro2A cells. The stark difference in methylation states at this position between Neuro2A, and ES and somatic cells, is clearly demonstrated by the COBRA which shows an undigested DNA fragment for Neuro2A whereas complete digestion is observed for the ES and somatic cell samples.

![Figure 3.4](image)

**Figure 3.4. COBRA analyses of representative IAP LTR loci.** (A) Representative COBRA analysis shown along with heatmaps from HT-TREBS analysis. Red boxes indicate methylated CpG sites and blue boxes indicate unmethylated CpG sites in the heatmaps in the bottom panel, with a blank box indicating insufficient sequencing coverage. Black triangles indicate one of the restriction sites, which was also sequenced by HT-TREBS. (B) COBRA analysis from four randomly chosen loci representing the various Venn diagram groups. All loci in (A) and (B) are named after the gene they are (distally) associated with and the Venn diagram group they belong to. A full description of each locus may be found in Supplemental Data 3.2. The percent numbers shown under the images indicate the average methylation level of this locus in the indicated cell type. The expected fragment sizes following the restriction enzyme digestion as indicated are marked by “U” (unmethylated) or “M” (methylated) based on the CpG site methylation status. The no-enzyme control for each restriction digestion is indicated by a minus sign.

COBRA results of four other loci from various different groups are shown in Figure 3.4B. A full description of each of these loci, along with the gene they are associated with, may be found in Supplemental Data 3.2. It becomes apparent from these COBRA analyses that the difference in
methylation level between Neuro2A and other cell types at the tested CpG sites is greater at some IAP LTRs compared to others. Two of the loci presented in this panel (IAP LTRs associated with *Fos* and *Tcam1*) show severe hypomethylation in Neuro2A cells, such that the undigested DNA fragments are easily detectible by COBRA. In contrast, the IAP LTR located within the second intron of *Itpk1* shows marked hypermethylation in Neuro2A, but much lower levels of methylation in ES and somatic cells. In general, these results are consistent with the detailed DNA methylation data obtained from these loci through HT-TREBS. However, not all loci can be tested as easily using COBRA due to technical limitations, such as low sensitivity and resolution capacity, as seen in the case of the IAP LTR associated with *Aebp2*. In conclusion, this set of analyses demonstrate that the DNA methylation difference at certain loci, though not all, could be verified using an independent method, COBRA.

**Discussion**

In the current study, the DNA methylation patterns of the mouse retrotransposon family IAP LTR have been studied in three distinct cell types: embryonic stem (ES), somatic and a neuroblastoma cell line, Neuro2A (Figure 3.1). Previous studies on DNA methylation of retrotransposons have focused on analyzing them as an entire group, without regard for locus-specific variations in methylation patterns. Therefore, thus far, it was unclear whether all loci behaved similarly in different cell types or whether some were more susceptible to changes in DNA methylation levels based on cell type and stage. Using HT-TREBS [24], we have now been able to analyze retrotransposons for the first time in a locus-specific manner, which has revealed its non-uniform nature of hypomethylation in ES and Neuro2A cells (Figure 3.2). The results suggest that not all IAP LTRs are equally sensitive to cell types and stages in terms of their DNA methylation level. Certain IAP LTRs showed severe hypomethylation in ES and Neuro2A cells,
while others were stably methylated at either high or low levels (Figures 3.3-3.4). Interestingly, much greater levels of hypomethylation was observed in Neuro2A cells than in ES cells, which may be due to two reasons. First, it may be due to strain related differences between strain A, 129 and C57BL/6N or cell line related artifacts which are not relevant in vivo. Second, Neuro2A cell line is derived from a murine neuroblastoma and hence is likely to show some cancer related signatures. From this perspective, the severe DNA hypomethylation of Neuro2A cells is consistent with various other reports about genome-wide hypomethylation of retrotransposons in cancer cells [17, 18, 29-32]. According to the current study, however, it is apparent that not all retrotransposons are likely to behave in a similar fashion when a cell undergoes transformations such as in development or cancer. Only a subset of all loci are likely to undergo changes in DNA methylation in response to such stresses, with varying degrees of severity depending upon the locus.

The hypomethylation we observed in ES cells had also been previously predicted based on the propensity for temporary silencing marks in these transient cell populations, even on retrotransposons [5, 6]. In fact, a recent whole-genome bisulfite sequencing analysis by Stadler et al. [33] showed that repeat elements comprised approximately 34% of all “low-methylation regions” (55-75% methylation) and 7% of all “unmethylated regions” (< 55% methylation). The scope of this study, however, does not allow it to focus on the composition of the repeat elements analyzed within these under-methylated groups. Here we show with greater specificity and sequencing depth for IAP LTR retrotransposons that 25% of the loci tend to be hypomethylated (< 80% methylation level) in these 129-derived ES cells. Such levels of hypomethylation in ES cells, however, is not a sign of increased transcriptional activity since only 0.23% of all ESTs in a blastocyst are associated with IAP elements [34]. Furthermore, Dnmt1-deficient ES cells are
only found to overexpress IAP mRNA upon differentiation and not in their pluripotent state [11]. This apparent paradox is explained by recent studies which have shown that KAP1, SETDB1 and HP1 act in tandem to silence retrotransposons in a DNA methylation-independent manner through H3K9 trimethylation in ES cells [7-9]. These temporary, but strong, silencing histone marks sustain proper repression while still allowing a significant fraction of the retrotransposons to remain hypomethylated. A subset of these hypomethylated retrotransposons then have the opportunity to become co-opted into the host genome as regulatory elements. Instances of such adaptation of retrotransposons have been previously observed in ES cells [35, 36] as well as other tissues [21, 37] in both humans and mice. It is possible, then, that a small fraction of the 1110 hypomethylated IAP LTRs detected in this study may also function as regulatory elements for nearby genic regions, much as the stem cell-related promoters and enhancers originating from ERVs that were recently identified by Fort et al. [35].

In conclusion, through this study, we have been able to demonstrate the feasibility of using HT-TREBS as a Next-Generation-Sequencing based approach to reliably assess the DNA methylation levels of retrotransposons in multiple different cell types. With greater specificity for targeted repeat elements and deep sequencing coverage, HT-TREBS may be used for identifying those retrotransposons which may be important for genome evolution or be associated with disease states in humans as well as mice. For instance, our observations in Neuro2A cells (Figure 3.2), which are potentially cancer-like, allow us to predict that only a subset of retrotransposons are likely to suffer from hypomethylation in most cancers with some facing much greater levels of demethylation compared to other loci. Moreover, based on the heterogeneity of cancer, it is predicted that the loci facing hypomethylation will most likely be different based on the origin of the tumor and to different degrees depending on the locus. HT-
TREBS can be employed on a large-scale basis to test which retrotransposons suffer hypomethylation and to what levels in different types of cancers. Since certain loci are likely to be severely hypomethylated, such as the ones demonstrated in Figure 3.4, they can be easily and quickly detected by COBRA and used as cancer biomarkers. Furthermore, given that certain loci are hypomethylated at extreme levels in Neuro2A cells (0-5% methylation), it is possible that demethylation events at these IAP LTRs are additive over time. If one may indeed consider Neuro2A cells to be somewhat reflective of a neuroblastoma cell state, this could indicate that certain specific retrotransposable loci may start losing methylation very early during tumorigenesis events. Such loci, as biomarkers, could be useful for tracking cancer progression. The sensitivity of HT-TREBS also provides us with the opportunity to find the subset of retrotransposons which are potentially associated with tumor suppressors or oncogenes and can affect their expression patterns, since early demethylation at these loci could indicate their status as drivers of cancer progression. Overall, the various different patterns of DNA hypomethylation observed in ES and Neuro2A cells via HT-TREBS, in this study, introduces the possibility of identifying specific repeat elements which may serve as biomarkers for different cell states and be especially useful for tracking disease progression.

Materials and Methods

HT-TREBS analyses of DNA isolated from ES and Neuro2A cells

For the current study, the HT-TREBS protocol developed and employed by Ekram and Kim to characterize somatic tissues (brain, liver and kidney) from a two-month-old C57BL/6N mouse [24] was applied to ES (AB2.2 from Baylor College of Medicine) and Neuro2A cells. For each sample, 1 µg of the purified genomic DNA was fragmented with sonication (Bioruptor NGS, Diagenode) to generate a pool of DNA fragments with the peak size being around 700 bp in
length. The fragmented DNA was end-repaired using the NEBNext® End Repair Module (New England BioLabs), and ligated to custom-made Ion Torrent ‘A’ adaptors in which all the cytosines have been methylated (Integrated DNA Technologies). The adaptor-ligated DNA fragments were further size-selected to remove any excess adaptors and DNA fragments smaller than 300 bp in length using the Agencourt AMPure XP beads (Beckman Coulter). The adaptor-ligated DNA library was modified using the bisulfite conversion reaction according to the manufacturer’s protocol (EZ DNA Methylation™ kit, Zymo Research). The bisulfite-converted library was used as template for a round of PCR (Maxime PCR Premix Kit, Intron Biotech) using the following two primers: the forward primer (5’-CCATCTCATCCCTGCGTGTCTCCGACTCAG-3’) designed to bind to the 5’ end of the ‘A’ adaptor region and the reverse primer (5’-CCACTACGCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT^CTCCCTAATTAACTAC AACCCATC-3’) designed to bind to the 24-bp region that is well conserved among the IAP LTR subtypes (IAP LTR1, 1a, 2, 2a, and 2b). The sequence in the 5’-side of the reverse primer marked by caret (^) corresponds to the ‘P1’ adaptor, which is part of the amplification strategy used for the Ion Torrent NGS scheme (Ion Torrent, Life Technologies). The PCR product was finally size-selected for a range of 250-300 bp in length using agarose gel electrophoresis. Each of the two PCR products was then individually sequenced in the Ion Personal Genome Machine (PGM) Sequencer using Ion 318 Chips (Ion Torrent, Life Technologies). The sequence reads generated from the two Ion PGM runs were individually mapped using the aligner Bowtie2 [38] to a curated reference genome made up of bisulfite-converted IAP LTR sequences. The mapped reads were filtered through several custom Perl scripts to extract only the sequences covering the IAP LTR and flanking unique regions. The filtered reads from each sample were separately analyzed using the BiQAnalyzerHT tool [39]. The detailed information regarding Perl scripts
and bioinformatic pipelines are available upon request\(^3\). ES and Neuro2A datasets have been added to the NCBI’s Gene Expression Omnibus [40] data repository and can be viewed under the accession number GSE60007 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60007).

**COBRA analyses**

Approximately 500 ng of purified genomic DNA from AB2.2 ES cell, Neuro2A and C57BL/6N brain was treated using the EZ DNA Methylation™ kit according to the manufacturer’s protocol (Zymo Research). This bisulfite-treated DNA was then used for the COBRA analyses. Specifically, 1 µL (~ 20 ng) of the converted DNA was used for methylation-unbiased PCR (Maxime PCR Premix Kit, Intron Biotech) using bisulfite primers which lacked any CpG dinucleotides or cytosines (all cytosines were converted to thymines). Next, amplified DNA was digested using appropriate restriction enzymes which recognized at least one CpG site as part of their recognition sequence (New England BioLabs). All primers, PCR conditions and amplified region coordinates (according to the mm9 version of the mouse genome) may be found in Supplemental Data 3.2, along with the restriction enzymes used for COBRA.

\(^3\) A complete protocol for HT-TREBS library preparation and data analysis was later published as A. Bakshi, M.B. Ekram, J. Kim, Locus-specific DNA methylation analysis of retrotransposons in ES, somatic and cancer cells using High-Throughput Targeted Repeat Element Bisulfite Sequencing, Genomics Data, 3 (2015), and has been reproduced in this dissertation as “chapter two.”
References


CHAPTER FOUR

DNA METHYLATION VARIATION OF HUMAN-SPECIFIC ALU REPEATS

Introduction

Alu elements are one of the most successful primate retrotransposons, with over one million copies in the human genome [1]. They are SINEs (Short INterspersed Elements), derived from the 7SL RNA gene, that have a dimeric structure, linked by a middle A-rich region [2]. The 5’- ends of Alu elements house the A- and B-boxes that are the internal hallmarks of a RNA polymerase III (Pol III) promoter, and their propagation is thought to occur through transcription by Pol III [3]. The 3’ends of the Alu have a poly-A sequence which is crucial to their retrotransposition mechanism, called Target-Primed Reverse Transcription [4-7]. Currently, relatively few Alu copies are capable of retrotransposition and those elements belong largely to the young Alu (AluY) family and its derivatives [8, 9]. Some of these elements have undergone massive expansions specifically in humans, thereby comprising largely human-specific subfamilies, such as AluYa5 and AluYb8 [10]. These expansions have been driven by both “master” Alu copies and secondary “source” elements [11]. Further, some of Alu subfamilies are predicted to contain “stealth-driver” elements, which escape negative selection by mobilizing at only very low rates over long periods of time [12].

DNA methylation is the major epigenetic mechanism which represses all retrotransposons in the human genome, including Alu [13]. However, Alu elements may be more affected in terms of DNA methylation than other retrotransposons because of their relatively high CpG density [14].

---

1 This chapter first appeared as A. Bakshi, S.W. Herke, M.A. Batzer, J. Kim, DNA methylation variation of human-specific Alu repeats, Epigenetics, 11 (2016) 163-173.
Indeed, most Alu elements tend to be heavily methylated in somatic tissues, with some locus- and tissue-specific differences [15-19]. They also show a unique pattern of differential methylation in the male and female germ cells, compared to somatic cells [20]. Primate oocytes and human dysgerminoma (primary germ cell tumors usually occurring in the ovary) show high levels of DNA methylation, similar to somatic cells [20]. In contrast, many Alu elements, especially the young Alu (e.g. the AluYa5 subfamily), show distinct hypomethylation in sperm [16, 17, 21]. Consequently, Alu-specific RNAs have been observed in spermatozoa, indicating their transcriptional activity [17].

Hypomethylation of Alu elements is not just associated with sperm, but also with several disease conditions [5, 22, 23]. It is predicted that hypomethylation may lead to Alu retrotransposition that can disrupt gene expression; it may also allow for Alu-mediated recombination which is believed to contribute to about 0.3% of all human diseases [22]. BRCA1 and BRCA2 (two breast cancer susceptibility genes) represent the best-characterized cases of diseases caused by Alu insertions and Alu-mediated recombinations [24-26]. In addition to breast cancer [27], recent reports suggest extensive hypomethylation of Alu in several other types of cancers, such as gastric carcinoma [28], multiple myeloma [29], epithelial ovarian cancer [30], and lung adenocarcinoma [31]. Hypomethylation of Alu in all these types of cancers is expected to be associated with tumor progression, as a recent meta-analysis concluded that hypomethylation of multiple repetitive elements, including Alu, has significant negative effects on tumor prognosis [32].

Based on the propensity for Alu hypomethylation in many cancers, efforts are underway to develop them as epigenetic cancer biomarkers [33]. Thus far, these efforts have focused on Alu elements as an entire group, instead of studying the methylation levels of individual Alu loci.
This has been due mainly to the lack of methods that could provide reliable methylation levels of individual retrotransposon loci in a targeted fashion and on a genome-wide scale. Hence, it has remained largely unknown which particular Alu loci, and how many, show early changes in DNA methylation specifically in response to changes in cell state, such as during tumorigenesis.

In this study, we have used High-Throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS) [34, 35] to derive the methylation levels regarding >5000 elements belonging to two of the most active Alu subfamilies, AluYa5 and AluYb8, which have been implicated in many diseases, including breast cancer [22, 24]. Our data show that ~90% of AluYa5/Yb8 loci are highly methylated; however, the hypomethylated loci (~10%) are often located close to gene promoters and show high degrees of variation in DNA methylation. We have characterized this variation in the context of tumorigenesis in the breast and with regard to inter-individual differences. The results indicate that AluYa5/Yb8 loci proximal to promoter regions may respond to tumorigenic events, but this response occurs in the background of very high levels of inter-individual variation in DNA methylation. In fact, genome-wide, only ~1% of AluYb8 elements are expected to be early responders specifically to tumorigenesis, suggesting the potential use of specific Alu elements as epigenetic biomarkers for the early detection of cancer.

Results

HT-TREBS of AluYa5 and AluYb8 in human skin-derived fibroblast

Here, we have analyzed the DNA methylation of >5000 individual Alu elements from one of the most commonly used human cell lines, skin-derived fibroblast cells, using an established protocol for HT-TREBS [34, 35]. We enriched for two human-specific subfamilies of the young Alu group, AluYa5 and AluYb8, using a primer scheme that took advantage of their subfamily-specific diagnostic mutations (Figure 4.1A) [5]. We obtained a total of 3.9 million final library
sequencing reads, after filtering for read quality. Reads were then processed and mapped to a
custom database containing unidirectional, bisulfite-converted sequence for the individual loci
(and flanking sequence) belonging to the AluYa5 and AluYb8 subfamilies. Given the high
sequence similarity within all AluY subfamilies, we repeated the mapping process against

Figure 4.1. HT-TREBS of AluYa5 and AluYb8 in human fibroblast cells. (A) Primer scheme
depicting the enrichment of AluYa5 and AluYb8 sequences using HT-TREBS. Briefly,
methylated Ion Torrent “A” adaptors were ligated to sheared and size-selected genomic DNA,
before the sample was bisulfite-treated. Ion “A” primers and subfamily-specific primers (with
Ion “P1” adaptor on their 5’ ends) amplified fragments containing part of the Alu element (the
first monomer and middle A-rich region) along with some flanking genomic sequence. (B)
Methylation profile for AluYa5 and AluYb8 elements, binned in increments of 5%, for the
5238 loci sequenced. Loci were subdivided into three broad groups roughly corresponding to
the inflection points in the graph: Low (0-50%), Medium (50-75%) and High methylation (75-
100%). (C) Characterization of loci in terms of variation in methylation level (expressed in
standard deviation, SD) for the three methylation groups. Dashed horizontal lines represent the
thresholds used in the analysis (SD 0.15 and 0.25). Variation in DNA methylation was generally
observed to be lower for highly methylated elements compared with those in the medium and
low methylation groups. (D) Distance of AluYa5 and Yb8 elements to the nearest transcription
start site (TSS) based on methylation level. Alu elements with low levels of methylation
appeared more likely to be present close to gene promoters compared with those methylated at
>50% levels.
databases for elements in the AluYc, Yd, Yf and Yk groups and discarded reads which also mapped to those groups. In the end, ~63% of the mapped reads belonged to the human-specific AluYa5 and Yb8 subfamilies, yielding methylation levels regarding ~75% of all AluYa5 and Yb8 elements in the human genome (Supplemental Data 4.1\(^2\)). On average, for the 3102 AluYa5 and 2136 AluYb8 elements, ~25 reads were used to calculate the percent methylation at each locus, with a minimum of 10X coverage.

The results indicated that both AluYa5 and Yb8 elements tend to be highly methylated in the human epigenome (Figure 4.1B). For these analyses, the entire data set of 5238 loci was divided into three groups: Low (0-50% methylation); Medium (50-75% methylation); and, High (75-100% methylation). For human skin-derived fibroblast cells, high methylation levels were observed for nearly 90% of these human-specific Alu elements (2788 in Ya5; 1867 in Yb8). By contrast, there were only 30 loci (~0.5%) in the “Low” category and 533 loci (~10%) in the “Medium” category. The distribution of Ya5:Yb8 loci within each methylation category ranged from ~60:40 to ~55:45.

For these elements within the fibroblast cell line, we also assessed the level of variability in DNA methylation as calculated by the standard deviation in methylation levels for all reads mapped to each specific locus (Figure 4.1C). Variation appeared to be inversely correlated with the methylation categories, with 84% of the elements (3903) in the “High” category having the lowest levels of standard deviation (SD 0-0.15) while 40% of the elements (12 out of 30) in the “Low” category had SD values in the highest range (≥ 0.25). Loci in the “Medium” category

\(^2\) All supplemental materials are available online with the original publication at the journal’s website.
showed intermediate levels of variation, with 75% (413 elements) having SD values between 0.15 and 0.25.

To test whether the methylation level of the AluYa5 and Yb8 elements correlated with any positional bias, we mapped their distance to the nearest transcription start site (TSS) using the Genomic Regions Enrichment of Annotations Tool (GREAT; [36]) (Figure 4.1D). For AluYa5 and Yb8 elements methylated at high or medium levels, most were found 5-500 kb from the nearest TSS. In contrast, those methylated at low levels were often (though not exclusively) found within 5 kb upstream of the TSS. Thus, most AluYa5 and AluYb8 elements are not present near gene-rich regions; however, those with low levels of methylation may be more frequently associated with proximity to gene promoters.

Characteristics of AluYa5 and AluYb8 elements located near gene promoters

A search for the nearest gene with a TSS within 1 kb of the AluYa5/Yb8 elements in our data set detected 18 genes, with equal numbers associated with AluYa5 vs. AluYb8 elements. With respect to their association with Alu loci methylation categories, the 18 genes were distributed as follows: “Low”, 7 genes (~40%); “Medium”, 4 genes (~20%); and, “High”, 7 genes (~40%). This highlights the proximity bias of low-methylation AluYa5/Yb8 elements; over 20% of the 30 elements with low levels of methylation were located within 1 kb of the nearest transcription start site, whereas only 0.1% of loci in the high-methylation (7 out of 4655) and 0.7% of those in the medium-methylation group (4 out of 553) were present so close to a gene (Figure 4.1D, Table 4.1).

The data in Table 4.1 also suggest that most AluYa5 and Yb8 elements found proximal to gene promoters often bear active epigenetic signatures in a tissue-specific manner. For instance, in
Table 4.1. Characteristics of AluYa5/Yb8 elements located within 1-kb of the nearest transcription start site (TSS) and their associated gene. Loci were subdivided into three broad groups: Low (0-50%), Medium (50-75%) and High methylation (75-100%). Negative sign (-) indicates that the Alu locus is upstream of the transcription start site.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Alu sub-family</th>
<th>Methylation Level (%)</th>
<th>Group</th>
<th>Distance to nearest TSS (bp)</th>
<th>H3K27ac shore</th>
<th>Hypomethylated Tissue(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR10Q1</td>
<td>Ya5</td>
<td>82</td>
<td>High</td>
<td>-160</td>
<td>—</td>
<td>Sperm</td>
</tr>
<tr>
<td>OSBPL10</td>
<td>Ya5</td>
<td>17</td>
<td>Low</td>
<td>-160</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>TMSB4Y</td>
<td>Ya5</td>
<td>24</td>
<td>Low</td>
<td>-200</td>
<td>—</td>
<td>Various</td>
</tr>
<tr>
<td>†UBE2T</td>
<td>Ya5</td>
<td>8</td>
<td>Low</td>
<td>529</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>†CEBPγ</td>
<td>Ya5</td>
<td>83</td>
<td>High</td>
<td>-578</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>HIGD1B</td>
<td>Yb8</td>
<td>66</td>
<td>Medium</td>
<td>-603</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>†DHODH</td>
<td>Yb8</td>
<td>21</td>
<td>Low</td>
<td>634</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>RABEPK</td>
<td>Yb8</td>
<td>44</td>
<td>Low</td>
<td>-660</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>UGGT2</td>
<td>Ya5</td>
<td>17</td>
<td>Low</td>
<td>-698</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>HENMT1</td>
<td>Ya5</td>
<td>66</td>
<td>Medium</td>
<td>729</td>
<td>—</td>
<td>Various</td>
</tr>
<tr>
<td>INTS5</td>
<td>Yb8</td>
<td>97</td>
<td>High</td>
<td>-779</td>
<td>Yes</td>
<td>Sperm</td>
</tr>
<tr>
<td>HTR3E</td>
<td>Yb8</td>
<td>83</td>
<td>High</td>
<td>-783</td>
<td>Yes</td>
<td>Sperm</td>
</tr>
<tr>
<td>†MAP3K7 (TAK1)</td>
<td>Yb8</td>
<td>20</td>
<td>Low</td>
<td>798</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>AASDH</td>
<td>Yb8</td>
<td>66</td>
<td>Medium</td>
<td>830</td>
<td>Yes</td>
<td>Various</td>
</tr>
<tr>
<td>ADAD1</td>
<td>Ya5</td>
<td>88</td>
<td>High</td>
<td>-854</td>
<td>—</td>
<td>Sperm</td>
</tr>
<tr>
<td>IDNK</td>
<td>Yb8</td>
<td>84</td>
<td>High</td>
<td>-931</td>
<td>—</td>
<td>Sperm</td>
</tr>
<tr>
<td>†EDAR</td>
<td>Yb8</td>
<td>61</td>
<td>Medium</td>
<td>-936</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>EIF2B1</td>
<td>Ya5</td>
<td>89</td>
<td>High</td>
<td>952</td>
<td>—</td>
<td>Sperm</td>
</tr>
</tbody>
</table>

† Loci further characterized in this study
several tissues, nearly 60% (10/18) of these Alu elements were found along the shores of regions bearing histone marks indicative of enhancer regions [37-39], with especially strong H3K27ac (acetylation on lysine 27 of histone H3) signals, and somewhat weaker signals for H3K4me1 (monomethylation on lysine 4 of histone H3) for some of these elements (data not shown). Next, consistent with the results of our study, nearly all elements from the “Low” and “Medium” methylation groups were also found to be hypomethylated in various other normal and cancer cell lines [40-45]. Finally, nearly all AluYa5/Yb8 elements in the high methylation category in fibroblast cells (as well as most loci in the low and medium categories) were nonetheless hypomethylated in the sperm [21].

Gene-associated AluYa5/Yb8 elements show variable DNA methylation

Most of the genes located within 1 kb of the 18 Alu elements have previously been found to be misregulated in various different cancers (Supplemental Figure 4.1){sup}. Hypothesizing that such changes could affect nearby Alu elements, we assessed the methylation levels of Alu elements located near five genes (CEBPG, UBE2T, EDAR, DHODH and MAP3K7/TAK1) as well as the genes’ endogenous promoters. Two genes, CEBPG and UBE2T, were associated with elements from the AluYa5 subfamily; the other three genes were associated with elements from the AluYb8 subfamily. Using two independent approaches, Combined Bisulfite Restriction Assay (COBRA; [46]) and NGS-based bisulfite sequencing of the amplicons, we tested the methylation level of these five Alu loci in five samples: matched breast normal; matched breast tumor; breast cancer; lung normal; and, lung cancer.

---

3 This figure appears in chapter five as Figure 5.3.
The COBRA results showed that four out of the five loci showed significant variation in at least one of the three cancer samples (Figure 4.2). The CEBPG-Alu locus was the most affected, showing significant hypermethylation by 20% on average in the matched breast tumor, and hypomethylation by 16% on average in the unmatched breast cancer sample. This particular locus was also affected especially severely in the lung cancer sample, showing near-complete demethylation (Figure 4.2A). UBE2T-Alu and EDAR-Alu also seemed affected in the context of

![Figure 4.2](image-url)

**Figure 4.2.** Methylation variation of AluYa5 and Yb8 loci closely associated with a gene in normal and cancer samples. Alu elements are named after their associated gene. (A-E) Methylation levels are quantified through COBRA, followed by densitometry, in five tissues: matched breast normal (Br-mat-N); matched breast primary tumor (Br-mat-T); breast cancer (Br-C); normal lung (Lu-N); and lung cancer (Lu-C). Bands are labeled as either methylated (red “M”) or unmethylated (blue “U”), along with the appropriate restriction enzyme. * P < 0.05; ** P ≤ 0.01; *** P ≤ 0.001; **** P ≤ 0.0001. Error bars indicate standard error of mean (SEM). (F) Relative to the appropriate control, CEBPG-Alu, UBE2T-Alu, EDAR-Alu and DHODH-Alu showed significant difference in methylation in at least one test sample. Heatmap summarizes the results as hypermethylation (red), hypomethylation (green), and no significant change (grey).
breast tumorigenesis (Figure 4.2BC), with UBE2T-Alu showing almost no methylation in both the matched breast tumor as well as the unmatched breast cancer sample, and EDAR-Alu showing 25% less methylation only in the unmatched breast cancer. In addition to CEBPG-Alu, lung cancer samples were hypomethylated by 10% for EDAR-Alu and by 12% for DHODH-Alu (Figure 4.2CD). The fifth locus, MAP3K7-Alu, showed no significant difference between any of the normal and cancer samples (Figure 4.2E). Figure 4.2F summarizes the COBRA-derived variation in methylation in the cancer samples relative to their normal controls.

The methylation patterns of three out of the five loci, CEBPG-Alu, EDAR-Alu and DHODH-Alu, were further tested with next-generation-sequencing on the bisulfite-PCR products. Overall, the bisulfite-NGS confirmed the COBRA results, demonstrating hypermethylation of CEBPG-Alu in the matched breast tumor and hypomethylation in the lung cancer samples (Figure 4.3), as well as being concordant for EDAR-Alu and DHODH-Alu (Supplemental Figures 4.2-4.3). Detailed analysis of the sequencing pattern also showed that the methylation of the CpG sites associated with the A- and B-boxes of the elements showed no unique patterns and was similar to nearby CpG sites which did not correspond to any functional element (Figure 4.3B). Additionally, variation in Alu methylation was unrelated to the methylation status of the endogenous gene promoters, which were completely unmethylated in all samples, aside from some hypermethylation at the CEBPG promoter in lung cancer (Supplemental Figures 4.4-4.5).

In sum, based on the COBRA and bisulfite-NGS data, we concluded that locus-specific differences in methylation exist for AluYa5 and AluYb8 elements proximal to four cancer-associated genes in breast and lung cancer samples. The changes observed in the breast tumor relative to the matched normal can be attributed to tumorigenesis, suggesting that methylation levels of UBE2T-Alu and CEBPG-Alu are likely to respond to changes in cell state.
Figure 4.3. Methylation variation of CEBPG-Alu in normal and cancer samples. (A) Heatmaps show the methylation patterns of the CEBPG-Alu locus (after nested PCR) in the five samples described in Figure 4.2. Red and blue in the heatmap indicate methylated and unmethylated CpG sites, whereas white marks CpG sites with unknown methylation status. CpG sites within actual Alu elements are enclosed by black rectangles. The % methylation is shown for the entire amplicon (top) as well as for CpG sites within vs. flanking the AluYa5 element (bottom). Significant differences ($P < 0.0001$) are shown with a large asterisk. Arrowheads indicate the A-box (grey) and B-box (purple) in the Alu [top], or CpG sites analyzed in COBRA ($BstUI$, green; $HpyCH4III$, orange) [bottom]. (B) Methylation levels of a few pertinent CpG sites within the locus (marked by the arrowheads) in the five samples.
**Inter-individual variation in AluYa5 and AluYb8 elements**

Various molecular events during normal development, independent of those related to tumorigenesis, may potentially lead to inter-individual variation in Alu DNA methylation. To assess the background level of variation in DNA methylation at these Alu loci, we tested the DNA methylation levels in eight normal breast tissues by COBRA and bisulfite-NGS for CEBPG-Alu, DHODH-Alu, EDAR-Alu and UBE2T-Alu. Methylation levels were high enough for accurate quantitation in only two of the four loci, with the range in normal tissue being ~9-85% for CEBPG-Alu and ~9-80% for DHODH-Alu (Figure 4A). Further, NGS-based bisulfite sequencing supported the COBRA data by showing various different methylation patterns between the eight normal and eight tumor samples (Supplemental Figure 4.6A-C). Overall, the data suggest that AluYa5 and Yb8 loci close to gene promoters are likely to show high levels of inter-individual variation in DNA methylation.

To determine whether some of the variation could be tumor-related, we then compared the COBRA methylation data obtained from the normal breasts to those from forty breast tumor samples. For these analyses, we grouped all individual methylation counts into normal vs. tumor samples. With respect to overall methylation state, the pooled site data (BstUI and HpyCH4III) for CEBPG-Alu showed no significant differences between normal and tumor samples (Wilcoxon-Mann-Whitney test, Figure 4.4B). However, analyzed separately, significant hypomethylation in the tumor was observed at the HpyCH4III site (Figures 4.4CD). We also quantified the percentage of normal and tumor samples in five methylation-ranges: Low (0-20%), Medium-Low (20-40%), Medium (40-60%), Medium-High (60-80%) and High (80-100%). Here, regardless of whether the COBRA sites for CEBPG-Alu were analyzed separately or together, ~25% more tumor samples were methylated in the low range compared to normal.
Figure 4.4. Inter-individual variation in DNA methylation of AluYa5 and Yb8 elements. (A) Heatmap showing the average methylation level of CEBPG-Alu and DHODH-Alu in eight normal and forty breast tumor samples as determined by the COBRA assay: dark blue, 0-20% methylation (Low); blue, 20-40% (Medium-Low); green, 40-60% (Medium); gold, 60-80% (Medium-High); and red, 80-100% (High). (B-G) Variation in methylation levels for CEBPG-Alu in normal vs. tumor samples for CpG sites recognized by: both BstUI and HpyCH4III (B, E); BstUI only (C, F); and, HpyCH4III only (D, G). Panels B-D represent statistical analyses of all reads as pooled into normal vs. tumor, with significant hypomethylation only at the HpyCH4III site; Panels E-G depict the relative numbers of samples in each group as pooled into five levels of variation in methylation, with more tumor samples methylated at low levels at all tested CpG sites.
(Figure 4.4E-G). Altogether, this analysis indicated the presence of inter-individual variations at AluYa5 and AluYb8 loci located close to gene promoters, with the methylation level at one CpG site in CEBPG-Alu being significantly lower in the breast tumor sample set.

**Aberrant DNA methylation of AluYa5 and AluYb8 in response to tumorigenesis**

Despite the large inter-individual variation in DNA methylation documented through the eight normal samples, initial analysis of the matched breast pair suggested that at least two loci, CEBPG-Alu and UBE2T-Alu, responded specifically to tumorigenesis (Figure 4.2AB). To determine how many AluYa5 and Yb8 elements are likely to respond similarly genome-wide, we performed another HT-TREBS on the same matched breast normal and primary tumor. Reads for normal (6.3 million) and tumor (6.8 million) samples were processed as described for the fibroblast cells, yielding methylation values regarding 104 AluYa5 and 1776 AluYb8 elements with a minimum of 10X coverage in both the normal and tumor samples (Supplemental Data 4.2).

Initial analysis of the HT-TREBS results revealed a subtle trend towards hypomethylation of AluYa5/Yb8 elements in the breast tumor compared to normal (Figure 4.5A). For instance, in terms of the percentages of Alu elements, there was a disparity between the tumor and normal samples within the 85-100% methylation level, with ~10% of the elements in the tumor being shifted into the 65-85% methylation range. In fact, 36% of all loci with methylation levels of <85% were unique to the tumor sample, while 55% were common to both the normal and the tumor (inset in Figure 4.5A). Next, to determine the pattern of change for individual elements, the methylation level of each AluYa5/Yb8 locus in the tumor sample was plotted against that in the normal sample (Figure 4.5B). Most loci clustered between 80-100% methylation, with a
Figure 4.5. Tumorigenesis-related variation in DNA methylation of AluYb8 elements. HT-TREBS analysis of matched breast normal and primary tumor tissues showing tumorigenesis-related variations in DNA methylation of AluYb8 elements. (A) Methylation profile of the normal (orange) and tumor (blue) samples, based on the percent Alu elements belonging to methylation bins in increments of 5%. The venn-diagram (inset) shows the 4X higher number of hypomethylated loci (<85% methylation; dotted line) in the tumor (blue) compared to the normal tissue (orange). (B) Methylation level of individual Alu loci in the matched breast tumor vs. normal, with each dot representing one Alu locus. Most loci follow the y=x dotted line, indicating no major difference in methylation between the two tissues. (C) Variation in methylation (standard deviation) vs. methylation level of the 22 differentially methylated AluYb8 loci, indicating lower methylation and higher levels of variation in the tumor tissue (blue) compared to the normal tissue (orange). (D) Heatmaps showing methylation levels of two AluYb8 loci in normal and tumor tissues. (E) Summary of information with respect to the nearest gene associated with the 22 differentially methylated AluYb8 loci (green for hypomethylation and red for hypermethylation). Negative sign (-) indicates that the Alu element is upstream of the gene; “N/A” indicates no gene within 1000 kb of the Alu element. Gene regulation information is from the invasive breast cancer (BRCA) and Pan-Cancer (PANCAN) data sets from The Cancer Genome Atlas (TCGA); using Student’s t-test (P < 0.05), genes were classified as downregulated (blue) or upregulated (gold). Grey indicates no significant change.
minority of loci ranging between 20-80% methylation. Loci falling on the diagonal y=x line represent Alu elements showing no change in DNA methylation during tumorigenesis; deviations from the line represent loci showing differential methylation between the two samples. Loci were considered to be responding to tumorigenesis if they deviated by at least 20% in their methylation between normal and tumor samples; at this threshold, as few as ten reads were sufficient to establish statistical significance. Twenty-two loci (all AluYb8) were differentially methylated at >20% levels, with 19 loci (~86%) showing hypomethylation in the tumor (Supplemental Data 4.2). Overall, the HT-TREBS data on the matched breast pair revealed that the primary breast tumor has at least 4X more uniquely hypomethylated loci than its matched normal sample. Furthermore, about 1% of AluYb8 loci are expected to show at least 20% difference in methylation in response to tumorigenesis.

Close examination of the 22 AluYb8 loci showed that, along with a decrease in DNA methylation, these loci also displayed an increase in the variability of their methylation state (Figure 4.5C). As seen in the individual heatmaps, this change is the result of read-specific hypomethylation, either in the entire span or in only the 5’ end of the Alu element (Figure 4.5D). Assuming that each individual methylation read is derived from the DNA from different cells, this pattern may indicate the accumulation of cells in the primary tumor that are losing DNA methylation at these loci. To test whether these early changes in DNA methylation correlate with events in later stages of breast cancer, we searched The Cancer Genome Atlas (TCGA) for the gene nearest to the Alu elements and determined whether that gene showed any misexpression in the invasive breast cancer or Pan-Cancer (PANCAN; [47]) data sets from TCGA (Figure 4.5E). Of the 22 loci responding to tumorigenesis, 14 had their nearest gene misexpressed in both the invasive breast cancer as well as in the PANCAN set. Considering the two data sets combined,
the AluYb8 elements appeared to be associated with a significantly greater number of down-regulated genes than up-regulated genes ($P=0.02$, chi-square test). This gave rise to a slight correlation between hypomethylated loci in breast tumor, and downregulated genes in the invasive breast cancer stage. With regard to the BRCA data set, 8 out of 19 (42%) loci hypomethylated in the primary breast tumor were associated with a gene downregulated in breast cancer, and 4 (21%) loci were associated with an upregulated gene (Figure 4.5E). Unfortunately, however, the significance of this correlation could not be established due to the small sample number. Overall, detailed analysis of the 1% AluYb8 elements responding to breast tumorigenesis indicates that a fraction of the tumor cells may show complete unmethylation of the Alu, either in part or in its entirety. Further, such changes in DNA methylation of Alu elements may serve as early indicators of downstream aberrations in gene expression, characteristic in the later stages of breast cancer.

Discussion

Here, we have determined DNA methylation levels regarding 5238 individual Alu loci belonging to the AluYa5 and AluYb8 subfamilies. Consistent with previous results, we find that ~90% of these elements are highly (>75%) methylated (Figure 4.1B), presumably as a means of repressing expression by retrotransposons in the human genome. Nevertheless, ~10% of the elements are hypomethylated and tend to be located in regions of active chromatin (Figure 4.1D, Table 4.1), potentially allowing them to escape this repression mechanism. Interestingly, these hypomethylated loci exhibit high variation in their methylation status from cell to cell in a tissue sample (Figure 4.1C). Loci near gene promoters also tend to show high levels of inter-individual variation in DNA methylation (Figure 4.4A, Supplemental Figure 4.6).
The existence of Alu elements with variable levels of activity, secondary and “stealth” drivers, have been proposed in previous studies [11, 12]; yet, the molecular mechanisms underlying these models remain unclear. Given that DNA methylation is one of the major epigenetic mechanisms repressing Alu retrotransposition, we speculate that inter-individual variations in DNA methylation, as reported here for AluYa5 and AluYb8 elements near gene promoters (Figure 4.4, Supplemental Figure 4.6), might play an important role. After all, for “stealth” drivers and secondary elements to be successful, they must be repressed in most individuals in order to maintain low retrotranspositional activity within the overall population. Differential repression of Alu elements between individuals is likely to be at the epigenetic level because the epigenome is highly plastic and easily influenced by various environmental factors. Further, because only a few individuals in a population are expected to have very low levels of DNA methylation at an Alu locus, very few copies should be active at the population level, with others being methylated (and perhaps active) at various different levels. Thus, differential methylation of the human-specific Alu elements, as found in this study, may decouple them from selection pressures and allow their continued propagation according to the Stealth Model [12].

With respect to disease, hypomethylation at Alu elements has been associated with various cancer types, and it may have a bearing on tumor prognosis [32]; however, there is a low penetrance of cancer incidences which can be directly attributed to Alu activity. The high inter-individual variation in methylation of Alu elements proximal to promoter regions may be a contributing factor towards the low penetrance phenotype. Only a small fraction of the population is expected to harbor severely hypomethylated Alu loci, potentially increasing their chances of developing cancer. The underlying mechanism could be either Alu retrotransposition or its activity in cis. As discussed below, cis activity is especially interesting with regard to Alu
elements proximal to cancer-associated genes, such as CEBPG-Alu (Figure 4.4, Supplemental Figure 4.6). In addition to the evolutionary perspective, further investigations regarding inter-individual variation in Alu methylation may prove fruitful in relation to elucidating disease susceptibility factors.

A long-standing question regarding the role of Alu retrotransposons in cancer has been whether they can be drivers of tumorigenesis or if they are simply “passengers.” Answering this question involves understanding how many Alu elements are likely to affect nearby gene promoters in cis. Here, focusing on only two subfamilies of human-specific Alu elements, we report characteristics for 18 AluYa5 and Yb8 elements located within 1 kb of a gene promoter (Table 4.1). Many of these loci are both hypomethylated and are located near enhancer regions in various tissues, with some bearing active histone marks themselves. Moreover, nearly all of these loci are associated with genes that are frequently misregulated in cancer (Supplemental Figure 4.1). Two loci, UBE2T-Alu (chr1:202310424-202310734) and CEBPG-Alu (chr19:33863506-33863809) are particularly noteworthy for the following reasons:

1. Early changes in DNA methylation during tumorigenesis were displayed by the AluYa5 elements associated with both of these genes, as indicated by significant changes in a primary breast tumor relative to the matched normal. CEBPG-Alu was hypermethylated; by contrast, UBE2T-Alu was hypomethylated (Figures 4.2-4.3).

2. Both elements are closely associated with tumor-suppressor/oncogenes and are located within the active promoter regions of their respective genes. In addition, CEBPG-Alu shows some enhancer-related histone marks in some cell lines (e.g. HMEC, normal human mammary epithelial cells) [38, 39], suggesting potential functional exaptation of the locus.
3. Both elements have potential function as \textit{cis} elements, being associated with genes that are directly involved in tumorigenesis. \textit{UBE2T} is frequently overexpressed in breast cancer and is thought to aid in breast carcinogenesis by downregulating BRCA1 [48]. \textit{CEBPG}, by contrast, is mainly a lung cancer gene, with inter-individual differences in expression level of \textit{CEBPG} being associated with the risk of lung cancer [49, 50]. In this study, \textit{CEBPG-Alu} showed inter-individual differences in methylation in the breast tissue (Figure 4.4); assuming similar inter-individual variation exists at the Alu locus in the lung, it may correlate with inter-individual variation in \textit{CEBPG} expression and the risk of lung cancer.

In summary, based on their methylation profiles in normal and cancer samples, we have identified several Alu elements for further research into their potential roles as drivers of tumorigenesis.

Regardless of their role as “drivers” or “passengers,” Alu elements have the potential to serve as epigenetic cancer biomarkers. We performed HT-TREBS on a primary breast tumor and its matched normal sample, and found 22 elements (~1%) with signatures of being early responders to breast tumorigenesis. These elements all belong to the AluYb8 subfamily and they are all located far from promoter regions (Figure 4.5). Alu hypomethylation has been considered to be a later event in breast carcinogenesis, given that significant changes in overall Alu methylation were not seen until the later more invasive stages [27]. Thus, the percentage of ‘early responders’ may vary amongst tumor types; for example, we expect it to be much higher in lung cancer where Alu hypomethylation is considered to be an early event [31]. Here, even in an early stage of breast tumor and using very stringent criteria, deep-sequencing of 1800 individual AluYa5/Yb8 loci revealed that the primary tumor had four times more uniquely hypomethylated loci (with \( \leq 85\% \) methylation) than the normal (Figure 4.5A), and that 19 of these loci showed
significant hypomethylation in a read-specific manner only in the tumor (Figure 4.5D). The greater number of reads showing complete unmethylation at the 5' end of the Alu (which houses regulatory elements for Alu propagation) may indicate an accumulation of cells within the tumor sample which are completely losing DNA methylation at this locus.

These results demonstrate that HT-TREBS is capable of detecting when very few cells within a potentially cancerous tissue show loss of methylation at a retrotransposon. In fact, as few as ten reads were deemed sufficient to establish statistical significance when the difference in methylation was at least 20%. Further, according to TCGA, many of the genes found within 1000 kb of these 22 AluYb8 elements showed a significant change in their expression level in invasive breast cancer as well as in the PANCAN data set [47]. This suggests that Alu elements located far from genes, which show an early change in DNA methylation during tumorigenesis, may be good biomarkers for the early detection and risk-assessment of cancer. However, given their distance from the associated genes, it seems likely that these Alu elements are merely neutral bystanders (a.k.a., “passengers”) in the genome, and that their methylation levels are only indicative of a certain cell state. Taken together, these factors render HT-TREBS a potentially useful technique in detecting epigenetic biomarkers for the early detection of cancer through the deep-bisulfite-sequencing of individual loci of the targeted retrotransposon.

Materials and Methods

HT-TREBS analysis of AluYa5 and AluYb8

The established protocol for HT-TREBS [34, 35] was used for this study, with some modifications. Purified DNA was sonicated to a mode of ~700-bp fragments, end-repaired and ligated to custom-designed Ion Torrent “A” adaptors with methylated cytosines (Integrated DNA Technologies). Excess adaptors and DNA fragments <300-bp were then removed using
Agencourt AMPure XP beads (Beckman Coulter). The final product was subjected to a bisulfite conversion reaction using EZ DNA Methylation™ kit (Zymo Research) according to manufacturer’s protocol. The bisulfite-converted DNA was then amplified with a forward primer complementary to the Ion Torrent “A” adaptor and reverse primers specifically designed to select for AluYa5 (5’-CCACTACGCCCTCGCTCTCCCTATGTCGCGTGT^CA
AATAACTAAAAACTACAACCRCCCRRCCACT-3’) and AluYb8 (5’-CCACTACGCCTCCGC
TTCTCTCTATGTCGCGTGT^CTCTATCRCCCAAACCRAACTACTA-3’) using their diagnostic mutations [5]. The underlined sequence before the caret (^) belongs to the Ion Torrent “P1” adaptor which was included as a 5’-extension on the reverse primers; ‘R’ stands for either ‘A’ or ‘G’ (IUPAC DNA nomenclature). The PCR products from the separate “AluYa5” and “AluYb8” reactions were then size-selected to a range of 450-500 bp in length using the Egel Precast Agarose Electrophoresis System (Life Technologies), purified by Select-A-Size DNA columns (Zymo Research), quantified on the Bioanalyzer DNA-HS chip (Agilent Technologies) and combined in equimolar concentrations. The combined library was templated on the Ion™ OneTouch 2 (using the Ion PGM Hi-Q™ OT2 Kit) and then sequenced on the Ion PGM (using the Ion PGM Hi-Q Sequencing Kit and a 318-v2 chip). Sequencing reads were first filtered for quality and size by the Ion Torrent Suite software (v-4.4.3). All remaining reads >100-bp in length were then used for mapping using Bowtie2 [51]. Reference genomes were custom-prepared for the mapping, which consisted of ~20,000 Alu sequences from six AluY subfamilies (a, b, c, d, f, & k) plus 700-bp flanking regions, from Human Genome Build hg19, in which all non-CpG cytosines were converted to thymines and all CpG cytosines to ‘Y’ (IUPAC ambiguous base for C/T). The mapped reads were then processed using various Perl scripts to yield only those sequences that contained the Alu element and at least 10-bp of flanking genomic sequence.
The filtered reads were then processed through BiQAnalyzerHT [52] to derive the individual methylation levels of each Alu locus as defined by the following equation: \[
\left(\frac{\# \text{ methylated CpG sites from all reads}}{\# \text{ all CpG sites from all reads}}\right) \times 100.
\] Due to the nature of the HT-TREBS protocol, these values applied to \(~60\%\) of each AluYa5 and \(~85\%\) of each AluYb8 element; the remaining portions of each element were not part of the amplification products and thus were not sequenced. All HT-TREBS datasets have been added to the NCBI’s Gene Expression Omnibus (GEO) [53] data repository and can be viewed under the accession number GSE74420 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74420).

**COBRA (Combined Bisulfite Restriction Analysis)**

Purified DNA (~500 ng) was treated with sodium bisulfite according to the EZ DNA Methylation™ kit protocol (Zymo Research). The bisulfite-treated DNA (~1 µL; ~20 ng) was then amplified by PCR using primers (Supplemental Data 4.3) lacking any CpG-dinucleotides and in which all cytosines were converted to thymines. PCR products were digested using a restriction enzyme that recognizes at least one CpG site within the Alu sequence and none in the flanking sequence (New England BioLabs). The gel band densities of the restriction products were analyzed using the Quantity One software (BioRad) to determine the percent methylation of the locus. Data from at least two restriction enzyme sites were used for all AluYa5/Yb8 elements tested in this manner. All densitometric analyses were repeated using an independent software, ImageJ, to ensure consistency. Error bars in Figure 4.2 were derived from repeating the entire process at least three times, starting with bisulfite-conversion and ending with the densitometry. By contrast, for the data in Figure 4.4 & Supplemental Figure 4.6, the bisulfite-conversion process could not be repeated due to insufficient DNA stocks; thus, the error bars in Figure 4.4 were derived from repeating the process from PCR to COBRA-densitometry 2-3.
times for each locus. The boxplots in these figures were generated by combining the data from all the independent trials of all restriction enzyme sites tested for the eight normal and forty tumor samples.

Bisulfite-NGS

One set of PCR products used for the COBRA analyses was used for NGS-based bisulfite sequencing. This was accomplished as follows. Multiple PCR products of the same sample were combined into one barcoded library, with different barcodes for each sample. These libraries were then end-repaired and ligated to Ion Torrent “A” and “P1” adaptors lacking 5’-phosphate, using a novel scheme that uses T4 ligase and Bst 2.0 WarmStart® DNA Polymerase (New England BioLabs). Unligated adaptors were then removed by agarose gel extraction and further purification by Select-A-Size DNA columns (Zymo Research), before the libraries were quantified on the Bioanalyzer (Agilent Technologies). Next, each library, with its unique barcode, was combined for multiplexed next-generation sequencing using the Ion Torrent 318-v2 chip as described above for HT-TREBS. After processing for quality and size (Torrent Suite 4.4.3), reads were filtered first by barcode sequence and then by the primer sequence for each locus from each sample. To derive individual methylation heatmaps and levels, the processed reads were analyzed using BiQ Analyzer [54] against the unconverted sequence for each amplification product. Percent methylation was calculated using: \[\frac{(\# \text{ methylated CpG sites from all reads})}{(\# \text{ all CpG sites from all reads})}\] * 100. Separate methylation levels for only the Alu locus were calculated by applying the formula above to only the CpG sites that mapped within the Alu sequence.
Statistical analyses

COBRA-densitometry data in Figures 4.2-4.3 were analyzed by the Student’s t-test (P < 0.05),
given that all analyzed cohorts were normally distributed (mean = median). For HT-TREBS and
bisulfite-NGS data, at least one group in each case was not normally distributed (mean ≠
median); thus, these data were analyzed by the non-parametric Wilcoxon-Mann-Whitney test in
RStudio. For data derived from next-generation sequencing, significance was set at P < 0.001
due to the high depth of coverage at each locus. The boxplots in Figure 4.4 were also subjected
to the Wilcoxon-Mann-Whitney test; however, as these data were derived from COBRA
analyses, P < 0.05 was deemed statistically significant.

References

Dewar, M. Doyle, W. FitzHugh, e. al., Initial sequencing and analysis of the human genome,


Batzer, P.L. Deininger, Active Alu element "A-tails": size does matter, Genome Research, 12


[6] M. Dewannieux, T. Heidmann, Role of poly(A) tail length in Alu retrotransposition,


Active Alu retrotransposons in the human genome, Genome Research, 18 (2008) 1875-1883.


Simpson, B. Ren, Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer, Genome Research, 22 (2012) 246-258.


CHAPTER FIVE

SUMMARY AND DISCUSSION

Summary

Approximately 50% of mammalian genomes consist of retrotransposons (Figure 1.1) [1, 2]. With their ability to retrotranspose, affect nearby gene expression (Figure 1.3; 1.5) and cause unequal homologous recombination, they can be a potent source of disruption to the genome [3]. DNA methylation is one of the major epigenetic mechanisms used to silence retrotransposon activity [4-7], and hypomethylation of retrotransposons has been observed in various cancers [8]. However, the highly repetitive nature of the sequences has thus far prevented the analysis of DNA methylation at individual retrotransposon loci (chapter one). To address these concerns, we have developed a new technique called HT-TREBS (High-Throughput Targeted Repeat Element Bisulfite Sequencing), described in detail in chapter two (Figure 2.1-2.2) [9, 10]. In this study, we have used HT-TREBS on the mouse IAP LTR family (Figure 1.2; chapter three) and two subfamilies of Alu elements in humans, AluYa5 and AluYb8 (Figure 1.4; chapter four), to address two major questions. First, we asked how the DNA methylation level of individual retrotransposon loci compare among different cell states, such as somatic, pluripotent and cancer cells. Second, we assessed whether the hypomethylation that has been frequently observed of retrotransposons in certain cell states, such as cancer, is a uniform or a more locus-specific event.

Chapter three describes our results from the mouse IAP LTR study comparing the DNA methylation states of >5000 elements in three cell types: embryonic stem (ES), somatic and cancer cells. According to the results, IAP LTRs are most heavily methylated in somatic cells at ~95%, followed by ES cells (~75%), and show the lowest levels of methylation (~50%) in
Neuro2A cancer cells (Figure 3.1). However, the hypomethylation in ES and cancer cells were not uniform; *i.e.* not all retrotransposon loci lost 25-50% of their methylation level. The hypomethylation was locus-specific, with about half the loci showing no difference in methylation in the three cell types. Of the other half, a large majority were hypomethylated in Neuro2A and about one-third showed shared hypomethylation patterns in ES and Neuro2A cells (Figure 3.2-3.3).

Similar conclusions regarding the methylation of human-specific retrotransposons in normal and cancer cells were derived from our AluYa5/Yb8 study (chapter four), which involved the analysis of >5000 AluYa5 and AluYb8 elements. We found that ~90% elements were highly methylated in skin-derived fibroblast cells; but interestingly, the 10% loci that were hypomethylated were frequently located close to gene promoters (Figure 4.1; Table 4.1). While these loci showed a tendency for high levels of inter-individual variation (Figure 4.4), a small fraction also responded early to tumorigenesis in the breast (Figure 4.2-4.5). Genome-wide analysis in a primary breast tumor and its matched control revealed that ~1% AluYb8 are likely to respond early and specifically to tumorigenesis (Figure 4.5).

Taken together, the results described in chapters three and four highlight several intriguing aspects of retrotransposon DNA methylation. First, our reports were the first to provide actual values for DNA methylation levels of a large number of mouse and human retrotransposons in several different cell types. Using HT-TREBS, we demonstrated that a majority of mammalian retrotransposons are highly methylated in normal somatic cells. While the overall levels of DNA methylation appears to be lower in other cell types, such as pluripotent and cancer, this loss does not affect all retrotransposons equally. Thus, we have shown conclusively that the loss of DNA methylation at retrotransposons, which has been frequently observed in many different cancers
[8, 11-15], is a locus-specific event [16, 17]. Second, we have added to previous IAP LTR studies revealing intra-individual and inter-individual variation within mouse tissues [9, 18] with similar observations regarding human retrotransposons. According to our analysis of AluYa5 and Yb8 elements, human retrotransposons also show within-tissue variation in DNA methylation from an individual, as well as inter-individual variation in the same tissue type. These observations, regarding the level of epigenetic variation in mammals, being attributed to retrotransposons allow us to speculate about the unique contribution of retrotransposons at both the individual as well as the population levels.

**Retrotransposons as Drivers of Carcinogenesis**

At the individual level, we find certain retrotransposon loci showing loss of DNA methylation in only some cells of a tissue, but not others, leading to the intra-tissue variation (Figures 4.1C) observed by us as well as others [9, 17, 18]. On a methylation heatmap, they show a “read-specific” difference in methylation (Figure 4.5D), where some sequencing reads show more methylated CpG sites at the locus than others. Here, each read may represent one cell in which the locus was sequenced; thus suggesting, overall, that some loci may only escape DNA methylation in a small number of cells in a tissue. This may be significant if that particular retrotransposon is located next to a disease-related gene, say a tumor-suppressor or an oncogene. Though hypomethylated in only a small number of cells, this retrotransposon may be affecting the expression of the nearby cancer-related gene (Figure 5.1A). Though in most circumstances, the body’s defense mechanisms are likely able to counteract any untoward effects caused by the aberrant gene expression (Figure 5.1B), given the right set of environmental cues (such as the activation of other genes with oncogenic potential; Figure 5.1C), this retrotransposon could potentially become a driver of carcinogenesis (Figure 5.1D).
Retrotransposons as drivers of carcinogenesis. A hypomethylated retrotransposon upstream of an oncogene may act as an alternative promoter through the binding of transcription factors (TF), thereby promoting the ectopic activation of the oncogene (A). Assuming that the retrotransposon-driven misregulation of the oncogene only occurs in a small number of cells in a tissue (cells with blue DNA; B), tissue identity remains unaltered. However, the addition of another stressor, such as the downregulation of a tumor-suppressor gene in these cells (C), can then result in uncontrolled cell division (D), and eventually, tumorigenesis. Retrotransposons are shown as green boxes (A) with the methylation state encoded by the filled (methylated) or unfilled (unmethylated) circles (CpG sites). The average methylation level of the tissue state (B-D) is shown as percent methylation.
In both IAP LTR and Alu elements, we have noted preliminary indications that a fraction of the loci may indeed be able to influence gene expression, which in turn could have a functional impact on cellular processes. An initial survey of genes surrounding the LHL group of IAP LTRs (which showed shared hypomethylation patterns in ES and cancer cells; Figure 3.3) suggested enrichment for biological processes shared between both cell types, such as regulation of cell proliferation, cell death and cell adhesion [19, 20]. Continuous cell division and proliferation are common features of both pluripotent as well as carcinogenic cells, as is the regulation of cell adhesive properties, such as in EMT (Epithelial-Mesenchymal Transition) during development and cancer metastasis [21, 22]. There were also correlations to signaling pathways that are known to play a role in both development as well as tumorigenesis such as the Wnt signaling pathway. Other cancer-related biological processes and pathways, such as angiogenesis and p53 signaling pathway, were also enriched in this gene set. Interestingly, several genes were associated with multiple types of cancers, such as renal cell carcinoma and basal cell carcinoma (Figure 5.2; unpublished data). Similarly, in humans, we found most of the genes located within 1 kb of AluYa5/Yb8 elements to be misregulated in various different cancer types (Figure 5.3) [17, 23]. Many of these Alu loci are located close to the promoter or enhancer regions of the cancer-related genes (Table 4.1; Figure 5.3), thus strengthening the suggestion that some retrotransposons may be indeed be able to have a functional impact on tumorigenesis.

As retrotransposons frequently become “re-activated” in cancer cells, one of the ways in which the hypothesis posed above could be tested is through the detection of fusion transcripts emanating from the hypomethylated retrotransposons located near tumor-suppressor or oncogenes. In order to find these retrotransposons that may drive tumorigenesis, loci that show hypomethylation in a small number of cells in pre-neoplastic tissues would be the ones of the
Figure 5.2. Gene association analyses of IAP LTRs in LHL group. A set of 915 genes, out of a total of 928, located within the 20 kb region surrounding the 773 IAP LTRs hypomethylated in ES and cancer cells (but not somatic cells) was used for gene association and pathway analyses using EGAN [20]. Several processes that share functional significance in ES and cancer cells appear to be highly enriched in this gene set. Set-based enrichment statistics (P-values) are calculated by hypergeometric tests. Enriched pathways were identified using the following databases: Gene Ontology Process (green lines); PANTHER (violet lines); KEGG (blue lines); MeSH (thistle lines). Gene associations are shown as following: PubMed co-occurrence (pink lines); human protein-protein interactions (purple lines); proximity on the chromosome (grey lines).
greatest interest (Figure 5.1AB). With a few retrotransposon loci hypomethylated even in normal tissues, it should be possible to detect at least a few ectopic transcripts, through sensitive high-throughput RNA sequencing techniques. If the retrotransposon is indeed a cancer-driver, then we would expect the cells that express the ectopic fusion transcript originating from it to be disproportionately affected by other oncogenic events co-occurring in the cell, and begin uncontrolled cell division (Figure 5.1CD). As a tissue progresses from normal to carcinogenic, we predict that many more of the potentially cancer-driving ectopic transcripts would be

![Figure 5.3. Expression level changes of Alu-associated genes in cancer tissues.](image-url)

16 out of 18 genes located within 1-kb of AluYa5 and AluYb8 elements showed either upregulation (red) or downregulation (green) according to five databases in The Cancer Genome Atlas: BRCA (breast invasive carcinoma), LUSC (lung squamous cell carcinoma), LUAD (lung adenocarcinoma), Lung (lung cancer) and PANCAN (pan-cancer). This figure first appeared as Supplemental Figure 1 in A. Bakshi, S.W. Herke, M.A. Batzer, J. Kim, DNA methylation variation of human-specific Alu repeats, Epigenetics, 11 (2016) 163-173, and has been previously referenced in chapter four as Supplemental Figure 4.1.
detectible, from the over-representation of the cells with the cancer-driving retrotransposon (Figure 5.1D).

In order to detect these low-level transcripts, which may not be well represented in a traditional RNA-seq, we propose a novel high-throughput sequencing technique designed to target fusion transcripts of a particular retrotransposon subtype. Modified from a protocol initially used by Horie, *et al.* [24], we have dubbed this technique “Junk RNA-seq,” since it is aimed at detecting RNA derived from what were traditionally considered “junk” DNA. Based on CapSeq [25], we envision this protocol to initially use the RLM-RACE (RNA Ligase Mediated Rapid Amplification of cDNA Ends) [26] approach to enrich for capped transcripts (Figure 5.4AB), followed by the ligation of the RNA RACE-adaptor to their 5’ ends (Figure 5.4C). The RACE adaptor-ligated RNA is then subjected to double-stranded cDNA synthesis, where the random primers are attached to the Ion Torrent P1 adaptors (Figure 5.4D). Next, the cDNA is amplified with a primer that overlaps the junction between the RACE-adaptor and the targeted retrotransposon (left arrow in Figure 5.4E), with another primer complementary to the P1-adaptor. We predict that this amplification scheme could be the key to selecting for transcripts triggered by the retrotransposons. The primer that targets the RACE-adaptor/retrotransposon would also carry the Ion Torrent-A adaptor as a 5’-extension. Thus, the resulting PCR products are enriched for cDNA derived from retrotransposon-driven transcripts, flanked by the two Ion Torrent sequencing adaptors (Figure 5.4F). This library can then be size selected and sequenced on a suitable Ion Torrent sequencing platform (Figure 5.4G).

Our unpublished data regarding IAP LTR-driven fusion transcripts in Neuro2A cells suggest the utility of the proposed protocol. For example, we have detected a potentially novel fusion
Figure 5.4. Conceptual diagram for Junk RNA-seq. Junk RNA-seq is designed to enrich for and sequence fusion transcripts arising from retrotransposons (green boxes; coded same as in Figure 5.1) driving ectopic transcription of endogenous genes. These 5’-capped transcripts are then selected for using a series of enzymatic treatments (A, B), followed by the ligation of the RNA RACE-adaptor at the 5’-end of the transcripts (C), and conversion to double-stranded (ds) cDNA with random primers that carry the Ion Torrent “P1” adaptor as a 5’-extension (D). Retrotransposon-driven transcripts are selected for in the next step (E), where we propose to employ an amplification scheme that uses a primer overlapping the junction between the RACE-adaptor and retrotransposon sequence (left arrow; E), with the other binding to the “P1” adaptor (right arrow; E). Since the novelty of the proposed technique lies in this unique amplification scheme, this step is highlighted in red. With the Ion Torrent “A” adaptor attached to the primer complementary to the RACE-adaptor/retrotransposon sequence, the resulting PCR products yield a library ready to size-selected and sequenced on a suitable Ion Torrent sequencing machine.
transcript driven from an IAP element upstream of the Glipr1l1 (Glioma pathogenesis-related protein-like 1) gene. Glipr1l1 is a member of a p53-regulated gene cluster, related to the tumor-suppressor genes Glipr (Glioma pathogenesis-related protein) and Glipr1 (Glioma pathogenesis related protein 1) [27]. Initially identified in the context of brain cancer, both Glipr and Glipr1 are also regulated by p53, and both have been shown to induce cell death in prostate cancer [28, 29]. Based on amino acid-sequence similarity and shared regulation by p53, Glipr1l1 is predicted to also function as a tumor-suppressor, similar to Glipr and Glipr1 [27]. Since an upstream-originating transcript of this gene is likely to be missing the first ATG-containing exon (due to the typical lack of splice acceptor sites on the first exon), the resulting disrupted ORF from the IAP LTR-induced ectopic transcript has the potential to dilute the tumor-suppressor function of the protein. Therefore, it would be particularly interesting to validate this novel transcript arising from an IAP LTR upstream of the Glipr1l1 gene, and further explore whether its expression can interfere with the native tumor-suppressor function. If so, then this locus might yield the proof of principle that the cancer biology field has been seeking for decades regarding the potential role of retrotransposons as cancer drivers. Regardless, this locus highlights the importance of developing sensitive high-throughput techniques that would be capable of detecting such fusion transcripts, and potentially leading us to the elusive retrotransposon loci that may drive carcinogenesis.

**Retrotransposons as Drivers of Mammalian Evolution**

The targeted high-throughput techniques discussed in this dissertation are useful ways of studying retrotransposons not only in the context of tumorigenesis, but also mammalian evolution. The inter-individual variation observed for IAP LTRs in mice and Alu elements in humans point towards the role of retrotransposons as a major source of epigenetic variation at the
population level [9, 17, 18]. We propose that such retrotransposon-based epigenetic variation may be a key player in mammalian evolution.

Variability in gene expression within a population is the main tenet of natural selection. Mendelian genetics dictates that such phenotypic variation must arise from genetic variation. However, only about 0.1% of the human DNA shows variability between individuals. This level of homogeneity in the human genetic code is an anomaly even amongst its closest ancestors, with chimpanzees showing greater genetic variation than humans [30-32]. Therefore, the human genome needs an alternate source of variation within the population. Early studies in monozygotic twins first revealed that epigenetic variation amongst humans might play a role in differential gene expression [33, 34]. Today, it is being increasingly acknowledged that epigenetic variation in human populations may be an important source of variability in the population from an evolutionary perspective [35-38].

With a large number of CpG sites and more-or-less even distribution throughout the genome, interspersed repeat elements are great candidates to harbor the epigenetic variation across the mammalian genome. Considering humans in particular, Alu elements and L1 promoters harbor over 30% of all CpG sites, spread over the entire genome. Thus, it is conceivable that DNA methylation at these retroelements could be an important source of epigenetic variation between humans. Moreover, the inter-individual variation observed for Alu elements nearby gene promoters may be a significant indication towards the prevalence of retrotransposon-driven “metastable epialleles” in humans, much as the ones previously discovered with regard to IAP LTR in mice [18, 39]. “Metastable epialleles” are so named because they code for variable expression of the gene through variable levels of epigenetic modifications, without alterations to the underlying genetic code [40]. Various AluYa5/Yb8 elements were identified in chapter four
that lay close to regions marked by enhancer-related histone modifications (Table 4.1). We speculate that the variable DNA methylation at these elements may result in variable levels of expression of the endogenous genes, assuming that the Alu elements act as enhancers for those genes. Then in times of stress, selection pressure can act on the phenotypic variation resulting from these retrotransposon-based epialleles, thus driving evolution.

This hypothesis can, theoretically, be tested using HT-TREBS and Junk RNA-seq with a model comparing the level of DNA methylation variation at interspersed elements before and after the application of some environmental stress forcing natural selection. Performing HT-TREBS before the application of the stress will yield the basal level of inter-individual variation in DNA methylation, with the most meaningful loci being associated with genes relevant for the specific stress response within a relevant tissue. Comparing the DNA methylation values with Junk RNA-seq data from the same tissues will further select for retrotransposons likely to be affecting gene expression in the “pre-stress” tissue state. If the hypothesis regarding the retrotransposon-driven epialleles were correct, we would expect comparatively lesser variation between individuals of the population after the application of the stress, since only some of the epiallele-driven phenotypes would be selected for. An example of how this hypothesis may play out in a real world scenario is presented in Figure 5.5. If the methylation level of a particular retrotransposon results in a phenotype where an individual requires more or less amount of food (Figure 5.5AB), then the methylation level of this retrotransposon may come under selection pressure following a period of food shortage (Figure 5.5CD). With the possibility of such retrotransposon-based selection yielding only subtle epigenetic changes at the population level, techniques like HT-TREBS and Junk RNA-seq may prove imperative for our ability to detect any potential metastable epialleles in humans.
Conclusions

Taken together, chapters three and four in this dissertation provide a high-resolution map of retrotransposon DNA methylation in mammalian cells. Based on the locus-specific methylation analysis for IAP LTR in mice, and AluYa5 and AluYb8 elements in humans, we conclude that most retrotransposons are highly methylated in normal somatic cells, with some loci showing inter-individual variation in DNA methylation. Our data also revealed the non-uniform nature of DNA hypomethylation at retrotransposons in ES and cancer cells; only a fraction of loci showed DNA methylation changes in response to cell state in both mice and humans. Overall, these results reveal some intriguing aspects of retrotransposon DNA methylation, and open up novel avenues for further research. The HT-TREBS techniques (chapter two) gives us the opportunity

Figure 5.5. Retrotransposons as drivers of mammalian evolution. Variably methylated retrotransposons within a population may act as metastable epialleles, driving differential gene expression (A, B). In the presence of a selection pressure (C), the epiallele offering the selective advantage to the individual would be favored in the population (D).
to track individual retrotransposon methylation levels through various stages of cancer. Such information regarding the dynamics of each locus, and their potential influence on nearby genes, can lead us in the search for cancer-driving retrotransposons. Furthermore, the highly sensitive technique can help us develop better retrotransposon methylation-based biomarkers for assessing the susceptibility for cancer among different individuals, detecting tumors in earlier stages and potentially tracking them as they progress to determine the best time for medical intervention. Finally, the HT-TREBS technique, along with the preliminary information from this dissertation, may also be critical in the search for retrotransposon-based metastable epialleles in humans, thereby answering critical questions about human evolution.

References


**APPENDIX: LETTERS OF PERMISSION**

**NATURE PUBLISHING GROUP LICENSE**

**TERMS AND CONDITIONS**

<table>
<thead>
<tr>
<th>Service</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>License Number</td>
<td>4091551042433</td>
</tr>
<tr>
<td>License date</td>
<td>Apr 17, 2017</td>
</tr>
<tr>
<td>Licensed Content Publisher</td>
<td>Nature Publishing Group</td>
</tr>
<tr>
<td>Licensed Content Publication</td>
<td>Nature Reviews Genetics</td>
</tr>
<tr>
<td>Licensed Content Title</td>
<td>The impact of retrotransposons on human genome evolution</td>
</tr>
<tr>
<td>Licensed Content Author</td>
<td>Richard Cordaux and Mark A. Batzer</td>
</tr>
<tr>
<td>Licensed Content Date</td>
<td>Oct 1, 2009</td>
</tr>
<tr>
<td>Licensed Content Volume</td>
<td>10</td>
</tr>
<tr>
<td>Licensed Content Issue</td>
<td>10</td>
</tr>
<tr>
<td>Type of Use</td>
<td>reuse in a dissertation / thesis</td>
</tr>
<tr>
<td>Requestor type</td>
<td>academic/educational</td>
</tr>
<tr>
<td>Format</td>
<td>print and electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>figures/tables/illustrations</td>
</tr>
<tr>
<td>Number of figures/tables/illustrations</td>
<td>1</td>
</tr>
<tr>
<td>High-res required</td>
<td>no</td>
</tr>
<tr>
<td>Figures</td>
<td>Figure 1a</td>
</tr>
<tr>
<td>Author of this NPG article</td>
<td>no</td>
</tr>
<tr>
<td>Title of your thesis / dissertation</td>
<td>Epigenetic Profiling of Mammalian Retrotransposons</td>
</tr>
<tr>
<td>Expected completion date</td>
<td>Aug 2017</td>
</tr>
<tr>
<td>Estimated size (number of pages)</td>
<td>100</td>
</tr>
<tr>
<td>Requestor Location</td>
<td>Baton Rouge, LA 70808</td>
</tr>
<tr>
<td></td>
<td>United States</td>
</tr>
<tr>
<td></td>
<td>Attn: Arundhati Bakshi</td>
</tr>
<tr>
<td>Total</td>
<td>0.00 USD</td>
</tr>
</tbody>
</table>
ELSEVIER LICENSE
TERMS AND CONDITIONS

License Number: 4091430411240
License date: Apr 17, 2017
Licensed Content Publisher: Elsevier
Licensed Content Publication: Molecular Genetics and Metabolism
Licensed Content Title: Alu Repeats and Human Disease
Licensed Content Author: Prescott L. Deininger, Mark A. Batzer
Licensed Content Date: July 1999
Licensed Content Volume: 67
Licensed Content Issue: 3
Licensed Content Pages: 11
Type of Use: reuse in a thesis/dissertation
Portion: figures/tables/illustrations
Number of figures/tables/illustrations: 1
Format: both print and electronic
Are you the author of this Elsevier article? No
Will you be translating? No
Order reference number: FIG. 1A
Title of your thesis/dissertation: Epigenetic Profiling of Mammalian Retrotransposons
Expected completion date: Aug 2017
Estimated size (number of pages): 100
Elsevier VAT number: GB 494 6272 12
Arundhati Bakshi
4033 Burbank Dr #23
Requestor Location: Baton Rouge, LA 70808
United States
Attn: Arundhati Bakshi
Publisher Tax ID: 98-0397604
Total: 0.00 USD
Title: Locus-specific DNA methylation analysis of retrotransposons in ES, somatic and cancer cells using High-Throughput Targeted Repeat Element Bisulfite Sequencing

Author: Arundhati Bakshi, Muhammad B. Ekram, Joomyeong Kim

Publication: Genomics Data
Publisher: Elsevier
Date: March 2015
Copyright © 2014 The Authors. Published by Elsevier Inc.

Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND)
This article is published under the terms of the Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND).

For non-commercial purposes you may distribute and copy the article and include it in a collective work (such as an anthology), provided you do not alter or modify the article, without permission from Elsevier. The original work must always be appropriately credited.

Permission is not required for this non-commercial use.
ELSEVIER LICENSE

TERMS AND CONDITIONS

This Agreement between Arundhati Bakshi ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number: 4084251043816
License date: Apr 08, 2017
Licensed Content Publisher: Elsevier
Licensed Content Publication: Genomics
Licensed Content Title: Retrotransposon-based profiling of mammalian epigenomes: DNA methylation of IAP LTRs in embryonic stem, somatic and cancer cells
Licensed Content Author: Arundhati Bakshi, Joomyeong Kim
Licensed Content Date: December 2014
Licensed Content Volume: 104
Licensed Content Issue: 6
Licensed Content Pages: 7
Type of Use: reuse in a thesis/dissertation
Portion: full article
Format: both print and electronic
Are you the author of this Elsevier article?: Yes
Will you be translating?: No
Order reference number: 4084251043816
Title of your thesis/dissertation: EPIGENETIC PROFILING OF MAMMALIAN RETROTRANSPOSONS
Expected completion date: Aug 2017
Estimated size (number of pages): 100
Elsevier VAT number: GB 494 6272 12
Arundhati Bakshi
4033 Burbank Dr #23
Baton Rouge, LA 70808
United States
Attn: Arundhati Bakshi
Publisher Tax ID: 98-0397604
Title: DNA methylation variation of human-specific Alu repeats

Author: Arundhati Bakshi, Scott W. Herke, Mark A. Batzer, Joomyeong Kim

Publication: Epigenetics
Publisher: Taylor & Francis
Date: Feb 1, 2016

Copyright © 2016 Taylor & Francis

Thesis/Dissertation Reuse Request

Taylor & Francis is pleased to offer reuses of its content for a thesis or dissertation free of charge contingent on resubmission of permission request if work is published.
VITA

Arundhati “Runa” Bakshi, daughter of Mrs. Alokananda (née Sinha) and Mr. Amitava Bakshi, is a native of India. She completed high school there before moving to Baton Rouge to attend Louisiana State University. As an undergraduate student, she completed a thesis entitled “Characterization of Aebp2 mutant mice” in lab of Dr. Joomyeong Kim, for which she was awarded college honors. She completed her B.S. in Biological Sciences in 2013, graduating summa cum laude with a university medal. She then continued in Dr. Kim’s lab as a graduate student, while being involved in developing a Course-based Undergraduate Research Experience (CURE) for freshmen in Biological Sciences. She anticipates graduating with a Ph.D. in August 2017, and continuing her career in the biomedical sciences.