Role of Retrotransposons as a Major Source of Intra- and Inter-Individual Epigenetic Variations in the Mammalian Genome

Muhammad Badre Ekram

Louisiana State University and Agricultural and Mechanical College

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

Recommended Citation
Ekram, Muhammad Badre, "Role of Retrotransposons as a Major Source of Intra- and Inter-Individual Epigenetic Variations in the Mammalian Genome" (2013). LSU Doctoral Dissertations. 313.
http://digitalcommons.lsu.edu/gradschool_dissertations/313
ROLE OF RETROTRANSPOSONS AS A MAJOR SOURCE OF INTRA- AND INTER-INDIVIDUAL EPIGENETIC VARIATIONS IN THE MAMMALIAN GENOME

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

Muhammad Badre Ekram
B.Sc., University of Dhaka, 2008
May 2013
This dissertation is dedicated to my beloved parents
Mr. Badre Ebriz and Mrs. Akhtar Jahan
ACKNOWLEDGEMENTS

It is an absolute pleasure and honor for me to express my gratefulness to the individuals who has been the key to all of my achievements so far in my career. Definitely at the very beginning, I would like to sincerely thank and express my gratitude towards my mentor and advisor, Dr. Joomyeong Kim. I cannot imagine of a better mentor to have other than him during my graduate school years. When it comes to research, his constant support in the form of knowledgeable advice, attention, patience, and jollity keeps the lab in an environment to thrive in. Through his frequent and friendly conversations, Dr. Kim enlighten us on the nooks and crannies of the art of good science that help us tremendously to prepare for the career of a researcher. I would also like to sincerely thank my committee members, Dr. Mark Batzer, Dr. David Donze, Dr. Patrick DiMario, and Dr. Shisheng Li for all their generosity, support and time.

I thank my past and present colleagues Dr. Hana Kim, Dr. Keunsoo Kang, Dr. Michelle Thiaville, Dr. Jennifer Huang, Dr. Chris Faulk, Dr. Jeongdo Kim, Dr. Sungryul Yu, and Dr. Hongzhi He for all their support, encouragements, advice, and time over the last five years. It was very kind of them to teach me many of the techniques that I have used in this dissertation. At the same, I mention with sincerity the contributions in my research of Mohammad Faisal, Wesley Frey, KimNgoc Tran, Arundhati Bakshi, Pinithi Perera, An Ye, Corey Bretz, and Austin Wong. I would like to specially thank Dr. Scott Herke of the LSU Genomic Facility for his constant technical support and advice, without which the establishment of the HT-TREBS technique explained in chapter three would have been difficult.
Last but not least, I sincerely thank my family. I thank my parents for being the mentors of my life, for raising me, for providing me the best education that I could have had, and for being the relentless source of encouragement in life. It is perhaps the happiest and proudest time of my life that I am so close to fulfilling one of their most cherished dreams. I thank my lovely and beautiful wife for her perpetual support, inspirations, and advice. I thank her for being my constant companion through the thick and thin of life. Finally, I thank my brother, my sister, and their families for their endless support and encouragements.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS  iii  

LIST OF TABLES  vi  

LIST OF FIGURES  vii  

ABSTRACT  viii  

CHAPTER ONE: BACKGROUND  1  

CHAPTER TWO: RETROTRANSPOSONS AS A MAJOR SOURCE OF EPIGENETIC VARIATIONS IN THE MAMMALIAN GENOME  15  

CHAPTER THREE: HIGH-THROUGHPUT TARGETED REPEAT ELEMENT BISULFITE SEQUENCING (HT-TREBS)  51  

CHAPTER FOUR: SUMMARY  80  

APPENDIX: LICENSE TO PUBLISH  88  

VITA  89
# LIST OF TABLES

2.1 Candidate metastable retrotransposons tested in the COBRA of 1-week-old littermates…………………………………………………………………………………………………………………………… 24

3.1 Coverage and sequencing efficiency of the twelve samples sequenced by the HT-REBS technique……………………………………………………………………………………………………………… 56
LIST OF FIGURES

1.1 Structure and recombination of an IAP locus................................................................. 9

2.1 Scheme of the bioinformatic approach for identification of potential epialleles from retrotransposons........................................................................................................... 18

2.2 Detailed bioinformatic analysis of the candidate and other relevant retrotransposons.. 20

2.3 Combined Bisulfite Restriction Analysis (COBRA) of randomly selected candidate metastable retrotransposons in five 1-week-old littermates........................................... 23

2.4 Bisulfite sequencing of five highly potential candidate metastable retrotransposons in the tail of 1-week-old littermates................................................................. 27

2.5 COBRA of the identified metastable retrotransposons in different tissues from 2- and 5-month-old mice......................................................................................... 30

2.6 COBRA of the identified metastable retrotransposons in germ cells and blastocyst…. 32

2.7 Model of progression of DNA methylation status of retrotransposons through different developmental and growth stages................................................................. 33

2.8 Verification and quantification of the ectopic transcripts with accompanying DNA methylation analysis................................................................. 35

3.1 Scheme and efficiency of the proposed High-throughput Repeat Element Bisulfite Sequencing (HT-TREBS) technique using mouse IAPLTR elements............................... 55

3.2 Depiction of the canonical methylation pattern of the IAPLTRs by ‘sprinkler plot’…. 59

3.3 Analyses of the hypomethylated IAPLTRs............................................................. 62

3.4 DNA methylation variations of the IAPLTR retrotransposons............................... 65
ABSTRACT

Transcription of retrotransposons is usually repressed by DNA methylation, but a few elements partially escape this repression mechanism. I used two approaches as my dissertation study to identify unmethylated intracisternal A-particles (IAPs) and also to understand why some of IAPs escape DNA methylation-mediated repression. Firstly, I sought to catalog the retrotransposons (LTRs and LINEs only) in the mouse genome known as ‘epialleles’ which escape the repression variably even in individuals with an identical genome sequence. Using bioinformatic approaches, 143 candidate epialleles were first identified based on their promoter activity and association with active histone modification marks. Detailed methylation analyses suggested that a subset of these elements showed variable levels of DNA methylation inter-individually, revealing their stochastic nature (metastability) of DNA methylation. The analyses also identified two opposite patterns of DNA methylation, progressive gaining versus losing, during development. qRT-PCR analyses demonstrated that the expression levels of these elements are indeed variable among the individual mice, suggesting functional consequences on their associated endogenous genes. Next, for a large scale individual element or loci-based methylation analyses of the candidate epiallele IAPLTRs and other repeat elements, we proposed and validated a novel high-throughput targeted repeat element bisulfite sequencing (HT-TREBS) technique. We obtained CpG methylation data of 5135 loci of five IAPLTR subtypes in three different tissues of four 1-week-old mouse littermates. An average of 52,000 CpG positions per sample with average sequencing depth of 117x was reported. More than 80% of the targeted subtype loci and less than 15% of the non-targeted subtype loci of IAPLTRs have been covered by this technique. Data analyses revealed that 2% of the IAPLTR loci have less than 80%
average CpG methylation with no genomic position preference and the majority of these loci are of the IAPLTR2 and IAPLTR2a subtypes. Further analyses also revealed extensive tissue-specific, individual-specific (epialleles), stochastic (random), and even sex-specific CpG methylation variations in the IAPLTR loci. Overall, our data confirmed the presence of many new retrotransposon-derived epialleles and other epigenetic variations. Our efficient and robust technique HT-TREBS also provided a lot of novel insight into the epigenome of the IAPLTR repeat elements.
CHAPTER ONE:
BACKGROUND

Epigenetics and the Epigenomes

The old and conventional theory that the genetic makeup is the ultimate fate determinant of a cell or an organism is no longer a valid one. Now it is well known that there are other much greater layers of information storage in the cells of most multicellular organisms in the form of DNA methylation, histone modifications and others. It is the difference in these epigenetic states (DNA methylation or histone modifications) that makes each cell type of the body different from the other. Just as the overall genetic makeup of an organism is known as the genome, the overall epigenetic state of an organism is called the ‘epigenome’. With the exception of the germ cells and the immune cells, every cell in the human body has the same genetic makeup or genome and yet the human body has over 200 different cell types each with their own epigenetic states. These unique epigenetic states of each cell type gives rise to tissue-specific or cell-specific expression of certain genes in a spatial and temporal manner and cause each cell to become what it is. Thus the epigenome of the human body is at least 200 times more diverse than its genome. The epigenetic modifications on the human and other mammalian genomes plays a critical role in many biological processes, including transcriptional control of development and tissue-specific genes, X chromosomal inactivation, genomic imprinting (Jaenisch and Bird, 2003), and repression of repeat elements. As part of the normal developmental process, the epigenetic modifications of the body undergo two rounds of resetting: once during the gametogenesis and the other during the early embryogenesis. The resetting during the gametogenesis erases the somatic modifications from the primordial germ cells and subsequently sets up the epigenetic
modification for the germ cells. During the other resetting at embryogenesis, the germ-cell specific modifications are altered to provide full developmental potential of the newly formed zygote (Reik, 2007). Since these epigenetic resetting processes are mediated through mechanisms that could be affected in a stochastic manner by many variables, such as enzyme activity, substrates’ concentration, and environmental exposures, these resetting processes could have several different outcomes. For instance, even if the sequence of the genome is identical in the case of human monozygotic twin individuals, these processes could derive several different versions of the epigenome, resulting in inter-individual epigenetic variations (American Association for Cancer Research Human Epigenome Task Force, 2008).

The Epigenome and Environmental Exposures
Apart from cellular differentiation, the other prime manifestation of the epigenetic process in an individual is phenotypic variations. In spite of having the same genomes, phenotypic variability has been often observed from the studies of human monozygotic twins (Fraga et al., 2005) and other model organisms with identical genetic backgrounds, such as inbred mouse strains. Such phenomena can now be greatly attributed to the differences in the epigenomes of those identical individuals. Now what might cause such difference in the epigenome? Recent studies have demonstrated that environment can intervene in the remodeling of the epigenome, especially during the early stage of development when the human genome goes through genome-wide epigenetic resetting (Gluckman et al., 2008). Such interventions in early development can sometimes cause individuals of monozygotic human twins to become two very different persons, physically and mentally. At the same time these interventions cause them to have a very different susceptibility to several adult-onset diseases, including cancer, diabetes, obesity, and mental
illness. Also unlike the DNA sequence, most epigenetic marks are reversible (Handel et al., 2010). Such relative plasticity of the epigenome means that differences might arise due to changes in conditions later in the life as well and not just in the early development. Environmental exposures are thus a critical topic in the study of any disease that is not solely genetic, or in any study involving human or animal behavior. Nonetheless, it is still somewhat unclear which genomic elements are epigenetically labile and thus contribute most significantly to variations in the human or mammalian epigenome in terms of environmental exposures.

**DNA Methylation and Histone Modification**

Epigenetics manifests its effect in the mammalian genome through four different suggested mechanisms: DNA methylation, histone modification, RNA interference, and chromatin structure. DNA methylation is perhaps the most stable of all four mechanisms and at the same time it also the most studied and biochemically well understood epigenetic mechanism (Feinberg, 2010). In mammals, DNA methylation involves the addition of a methyl group (-CH₃) to the C5 position of cytosine of DNA nucleotides. Typically it occurs within the context of a CpG (cytosine-phosphodiester bond-guanine) dinucleotide, but methylation of cytosine in other contexts like CHG or CHH (where H is equal to adenine, cytosine or thymine) is also prevalent in human embryonic stem cells.

DNA methylation is achieved by the transfer of a methyl group from S-adenosyl methionine (SAM) to a cytosine by a group of enzymes known as DNA methyltransferases (Bird, 2002). In mammals, DNA methylation is usually catalyzed by two classes of enzymes in two different situations: i) maintenance methylation, and ii) de novo methylation. DNA methyltransferase 1
(Dnmt1) is responsible for DNA methylation maintenance by methylating hemi-methylated DNA or DNA in which only one strand has been methylated, resulting from the semiconservative replication of DNA (Sharif et al., 2007). This enzyme is also necessary for maintaining the correct methylation pattern in imprinted regions (Hirasawa et al., 2008). DNA methylation in the de novo context is carried out by the two de novo methyltransferases: Dnmt3a and Dnmt3b, which establish the DNA methylation pattern in early development (Okano et al., 1999). A related non-catalytic protein, Dnmt3l assists the de novo methyltransferases by forming complexes with Dnmt3a and/or Dnmt3b and increasing their ability to bind to specific target DNA regions as well as stimulating their activity. Hypomethylation of transposable elements and interruption of spermatocyte development has been observed in disruptions of Dnmt3l (Bourc’his and Bestor, 2004).

DNA methylation plays a very significant role in stability at the cellular and organismal levels. Promoters of genes usually have regions of higher than expected CpG dinucleotides known as CpG islands and the transcriptional activity of those genes can be largely affected by the methylation of those CpGs in the CpG islands. One mechanism by which DNA methylation achieves such repression of genes is by blocking certain transcription factors from binding to promoter and other regulatory regions. An important role of DNA methylation in the formation of chromatin structure has also been suggested (Weber and Schubeler, 2007). DNA methylation also plays a crucial role in the maintenance of genomic imprinting, which is a process in which only one specified parental allele is expressed (Li et al., 1993). This parental allele specific expression of an imprinted gene is achieved by the silencing of the gene in the parental allele from which it is not expressed by means of DNA methylation. As an example, Peg3 is a
paternally expressed imprinted gene in which the promoter of the maternal Peg3 allele is heavily methylated and silenced while the promoter of the paternal allele is unmethylated and expressed (Kuroiwa et al., 1996; Relaix et al., 1998). Perhaps, the one of the most important roles of DNA methylation is the repression or silencing of repeat elements, especially transposable elements, and as such DNA methylation is assumed to be the host cell’s defense against such elements (Yoder et al., 1997). In spite of such repressive mechanisms, as we demonstrate in chapters two and three, evasion of DNA methylation by small number of retrotransposons have been observed in mice and probably in other mammals as well.

Some of the other vital tools in the repertoire of epigenetic mechanisms are histone modifications. Such modifications have been observed primarily in the tails of almost all the major histone proteins and their derivatives. Relatively a new concept, the hypothesis of the “histone code” was only coined in 2001 by Thomas Jenuwein, David Allis, and Bryan Turner (Turner 2000; Jenuwein and Allis, 2001). The hypothesis states that the transcription of genes is in part regulated by chemical modifications to the histone proteins. The modifications to those histone proteins can be of various types, and among the more important ones are methylation, acetylation, phosphorylation, ubiquitination, and crotonylation (Strahl and Allis, 2000; Tan et al. 2011). The position of the amino acid residues that are modified in these proteins are also widely variable.

So far, the acetylation and methylation are two of the most studied modifications. Acetylations of histone are carried out by histone acetyltransferases (HATs) and these create a relaxed structure of the chromatin resulting in gene expression. Deacetylation of histones is catalyzed by
Histone deacetylases (HDACs) and is associated with repression. Unlike acetylation, the regulatory fate of methylation is rather complex and dependent on both the position of the amino acid residue that is being modified and also the number of methyl groups that are being added or removed. For example, mono-, di-, and tri-methylation of histone H3 lysine 4 (H3K4me1, H3K4me2, and H3K4me3) are associated with enhancers, tissue-specific gene regulation, and active promoters respectively. Contrary to the rather activating signals of these modifications, tri-methylation of histone H3 lysine 9 (H3K9me3) and of histone H3 lysine 27 (H3K27me3) are repressive with the H3K9me3 more abundant in constitutive heterochromatin regions and H3K27me3 more prevalent in promoters of lineage-specific genes that are silenced by Polycomb proteins (Berger, 2007). Repeat elements including retrotransposons are usually marked by the repressive H3K9me3 marks but as shown in chapter two, there are instances where the presence of other histone modifications, even activating ones, is observed in a subset of retrotransposons.

The regulation of gene expression and repression of the repeat elements is mediated by an interplay of both DNA methylation and histone modifications. Both these epigenetic mechanisms have a considerable influence on the chromatin structure and gene function and quite often have been found to act in concert (Cedar and Bergman, 2009).

**Retrotransposons and their Repression**

More than approximately half of the human genome is composed of repeat elements of different types (Schmid and Deininger, 1975; Batzer and Deininger, 2004). In fact, in a recent study using an alternative strategy to find repeat sequences, it is suggested that repeat elements might make up about two-thirds of the genome (de Koning et al., 2011). A minor fraction of these repeat
elements are composed of simple, tandem, repeated motifs, such as microsatellites or minisatellites. The bulk of the repeats, or in other words almost 45% of the genome are made up of mobile elements known as transposons (Cordaux and Batzer, 2009). These repeat elements are so named because of their mechanism of transposition or ability to move within the genome. The discovery of the mobile elements (transposons) during her study on color mosaicism in maize in the middle of 1900s earned Barbara McClintock the Nobel Prize in Physiology or Medicine in 1983 (McClintock, 1950; McClintock, 1956). Based on the mechanism by which mobile elements transpose, they are divided into two groups: i) DNA transposons – which transpose by so the called “cut-and-paste” mechanism to new position in the genome, and ii) retrotransposons – which transpose by the “copy-and-paste” mechanism in which an RNA intermediate is transcribed from the element and later a new DNA is synthesized by reverse transcription from that RNA, to be integrated to the new site. Retrotransposons, which comprise about 42% of the human genome (Cordaux and Batzer, 2009), represent the bulk of all transposons. Based on their ability to encode the enzymatic machinery necessary for their own retrotransposition, the retrotransposons are further classified into autonomous retrotransposons (capable of expressing the enzymes) and non-autonomous retrotransposons (incapable of expressing the enzymes). Autonomous retrotransposons are grouped into long terminal repeat (LTR) and non-LTR elements. Examples of LTR elements are human endogenous retroviruses (HERVs) in humans or intracisternal A-particles (IAPs) in mice while an example of a non-LTR element is long interspersed elements 1 (LINE-1) in humans. In humans, non-autonomous retrotransposons are generally classified as either short interspersed elements (SINEs), which include the Alu element, or as SVAs, which is a composite element and is named after its three main components, SINE-R, VNTR and Alu.
Because of the ability of retrotransposons to be transcribed from their own inherent promoters, it is crucial for the cellular mechanisms to repress these elements. Such repression is necessary, firstly, to prevent the elements from creating copies of themselves in random positions in the genome and maintain genomic integrity. Secondly, their presence to a nearby gene can provide an alternative promoter for the gene from which a transcript of the gene might be produced even when the original promoter is silenced (Michaud et al., 1994). Such constitutive transcription of a gene which is otherwise expressed in a temporal and spatial manner is known as ectopic transcription. Thus the repression of retrotransposons and other repeat elements is essential for the maintenance of genomic integrity and other cellular processes. Such repression is primarily achieved by DNA methylation. The histone modification H3K9me3 also serves as an important mechanism for repression of the elements.

**Epigenetic Variability of Retrotransposons: Metastable Epialleles**

A small number of retrotransposons in mice, in particular, of the IAP LTR type (Figure 1.1), have been found to be susceptible to environmental factors. Even though the majority of retrotransposons are transcriptionally silent due to DNA methylation and histone modifications, a subset of these elements (in the range of few hundred) escapes from epigenetic repression and are currently active in the mouse genome. Two excellent examples of such elements are the IAP insertions in the mouse endogenous genes agouti and axin causing the mutant mouse phenotypes, viable yellow agouti (A\textsuperscript{vy}) and axin-fused kinky (Axin\textsuperscript{fused}) respectively. In previous studies with both the mutant types, it was observed that these retrotransposons show very unusual inter-individual variations in their DNA methylation levels in the absence of genetic variation
In both cases the resultant ectopic transcriptions disrupt the regular transcripts of the genes giving rise to the phenotypes like coat color change and obesity for the *agouti* locus and tail kinkedness for the *axin-fused* kinky locus.

![Diagram of IAP locus](image)

**Figure 1.1.** Structure and recombination of an IAP locus. (a) Structure of a typical full-length IAP locus with a total length of approximately 7.2 kilobase (kb). Two identical LTRs 5’ and 3’ border three open reading frames (ORFs) that encode the retroviral genes gag, prt (protease) and pol. (b) Creation of a solitary (or solo) LTR by inter-LTR recombination. (c) Solo LTR after the removal of the 3’ LTR and internal protein coding sequences.

Also the severities of the phenotypes are very much positively correlated to the degrees of loss of DNA methylation. Retrotransposons and their associated alleles with such distinctive nature are termed as ‘metastable epialleles’ in order to distinguish them from the traditional alleles (Rakyan et al., 2002). This variable nature of DNA methylation in these retrotransposons and the presence of huge number of retrotransposons in both the human and mouse genome posited the
probability that phenotypic variations in humans and mice, such as disease susceptibility, may be caused by epigenetic variations stemming from similar retrotransposons, even though only a handful of such retrotransposons have been known in the mouse genome and almost none have been so far reported in the human genome.

Chapter Two describes our attempts to characterize the retrotransposons as a major source of epigenetic variations in mammalian genomes. It explains a potentially novel way of identifying more of the epialleles by means of their association with the active histone modification mark H3K4me3 and the presence of nearby annotated mRNAs.

**Study of Epigenetic Variability of Retrotransposons by Histone Modification Analysis and DNA Methylation Analysis**

Next Generation Sequencing (NGS) technology or High-Throughput (HT) sequencing technology is creating a revolutionary phase in the world of DNA sequencing. The continuous and rapid improvements in these technologies since their advent in the early 2000s have enabled sequencing of as much as 600 Gigabase (Gb) worth of DNA sequences within several days a reality. Such massive amounts of data in such short period of time not only changed the way experiments are designed now but also necessitated the development of numerous new computational/bioinformatics tools that are capable of handling such large amounts of data.

The use of NGS technologies opened a new way for histone modification analysis by the use of the technique known as ChIP-Seq. By means of this procedure, it is possible to sequence millions of short DNA fragments obtained during chromatin immunoprecipitation (ChIP) using
an antibody that specifically binds the targeted histone protein modified at the desired position. The reads obtained by such experiments enable the researcher to identify regions in the genome which have high enrichment (regions where many reads map) of the targeted histone modification. Chapter Two describes the use of ChIP-Seq enrichment data for the activating signal H3K4me3 from four different mouse tissues/cells in order to find the retrotransposons that are near regions with abundant H3K4me3. One primary assumption of our experiment was that such retrotransposons are more likely to be metastable epialleles.

NGS technologies also paved the way to revolutionize DNA methylation analysis. One of the gold standard techniques for DNA methylation analysis is the bisulfite conversion method. In this method, chemical conversion with the bisulfite reagent changes any unmethylated cytosines to uracil, while leaving methylated cytosines intact (Clark et al., 1994). PCR amplification after this conversion converts any uracil to thymine in the amplified products. NGS sequencing of the limited step amplification products of the bisulfite converted DNA gives millions of reads, which after careful comparison to a reference bisulfite converted genome identifies the cytosines which were methylated and which were not. At the moment, several different high-throughput bisulfite techniques are available that allow genome-scale study of cytosine methylation status, such as the reduced representation bisulfite technique (RRBS) and methylC-Seq (Lister et al., 2008; Meissner et al., 2008). Unfortunately, no high-throughput technique was available that allowed an individual locus-based bisulfite sequencing of repeat elements. Chapter Three proposes and validates a novel technique named High-throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS) that allows single-base resolution methylation analysis of individual
loci of repeat elements. Our results also prove the efficiency and robustness of the technique for the purpose of genome-wide repeat element bisulfite sequencing.

References


CHAPTER TWO: RETROTRANSPOSONS AS A MAJOR SOURCE OF EPIGENETIC VARIATIONS IN THE MAMMALIAN GENOME

Introduction

Epigenetic modification on the mammalian genomes is part of the normal process of development, and also plays a critical role in many biological processes, including transcriptional control of development and tissue-specific genes, X chromosomal inactivation, and genomic imprinting (Jaenisch and Bird, 2003). Epigenetic modification is usually mediated through specific enzymatic reactions, and thus influenced by many variables, such as substrates’ concentration and enzyme activity. As a result, even if the sequence of the human genome is identical as in the case of a monozygotic twin, these processes could easily derive several different versions of the human epigenome, providing a source for inter-individual variations (American Association for Cancer Research Human Epigenome Task Force and European Union, Network of Excellence, Scientific Advisory Board, 2008). According to recent epidemiological studies, the human population also displays very unusual inter-individual variations in terms of its susceptibility to several adult-onset diseases, including cancer, diabetes, obesity, and mental illness (Feinberg et al., 2006; Whitelaw and Whitelaw, 2006; Esteller, 2007; Jirtle and Skinner, 2007). Yet, this phenotypic variability has often been observed even between human monozygotic twins and among other model organisms with isogenic backgrounds (Fraga et al., 2005; Kaminsky et al., 2009). These observations suggest that epigenetic, but not genetic, variations are likely responsible for a lot of the phenotypic differences observed in humans and other model organisms.

*Reprinted by permission of Epigenetics (Landes Bioscience)
It is currently unknown what genomic elements contribute most significantly to variations in the epigenome. However, according to the studies of two mutant mouse phenotypes, viable yellow agouti (A\(^{vy}\)) and axin-fused kinky (Axin\(^{fused}\)), retrotransposons of relatively recent origin tend to show inter-individual variations in their DNA methylation levels and also concurrent phenotypic variations (Whitelaw and Whitelaw, 2006; Esteller, 2007; Jirtle and Skinner, 2007). Interestingly, both mutant phenotypes are caused by intracisternal A-particle (IAP) insertions and subsequent ectopic expression of the two endogenous genes, agouti and axin genes, driven by the IAP elements’ long terminal repeat (LTR). In both loci, the IAP LTR partially escapes DNA methylation-mediated repression, and triggers the production of fusion transcripts of LTR and endogenous gene’s transcripts. These LTR-driven ectopic expressions subsequently cause disruptions of the cell type specific and time or stage specific transcription of the endogenous genes, resulting in coat color change and obesity for the agouti locus and tail kinkedness for the axin-fused kinky locus. The different levels of DNA methylation on the two LTRs also correlate well with the expression levels of fusion transcripts and associated phenotypes.

The majority of retrotransposons are usually repressed by DNA methylation, and the degrees of DNA methylation are complete and static in different somatic tissues. In contrast, the DNA methylation on the two LTRs of the agouti and axin loci is partial and metastable. These specific LTRs display a wide range of DNA methylation (30 to 90\%) within the mice of an isogenic genetic background. The DNA methylation on these LTRs is also easily affected by nutritional supplements and other environmental agents particularly during early embryogenesis (Wolff et al., 1998; Waterland and Jirtle, 2003; Waterland et al., 2006). Such distinctive nature of these retrotransposons and their associated alleles earned them the term ‘metastable epialleles’ in order
to distinguish them from traditional alleles (Rakyan et al., 2002). This variable and metastable nature of DNA methylation in these retrotransposons further predicts that phenotypic variations in mice, humans and other mammals may be caused by epigenetic variations stemming from similar retrotransposons. To test this prediction, we performed a series of bioinformatic searches and DNA methylation analyses on a subset of potential epialleles in mouse in this study. The results from the analyses of this data indicate a strong probability that the mammalian genomes contain a large number of uncharacterized metastable epialleles that are derived from retrotransposons, and further that these elements may be a major source to inter-individual variations in the mammalian epigenome.

Results

Identification of potential metastable epialleles using bioinformatic searches

A series of bioinformatic searches was performed to identify additional retrotransposons with similar epigenetic features as the two epialleles, the agouti and the axin-fused kinky loci, that have been mentioned earlier (Figure 2.1 & http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf). For our study, we excluded SINE elements due to the prediction that the majority of Pol III-driven SINEs may not be a good source for promoter elements for Pol II transcription even though it is possible that some opportunistic SINEs could also function as promoters for Pol II transcription (Lai et al., 2009). So here and onwards, we will be meaning only LINEs and LTRs when we mention retrotransposons. First, we looked for retrotransposon and mRNA pairs where the transcription start site (TSS) of the mRNA is located within 50 bp upstream and 250 bp downstream of the 5'-end of the retrotransposon by intersecting the entire retrotransposon set (1,767,082 elements
containing LINEs, and LTRs) with the transcriptome of the mouse (228,765 mRNAs). This search identified 5,679 retrotransposons that may initiate or influence the transcription of adjacent mRNAs. Second, this set of retrotransposons was further filtered using the following criteria. The size of a given retrotransposon needed to be equal to or greater than 50 bp in length and also the element needed to contain at least 3 CpG dinucleotides, which were necessary for DNA methylation analyses at the later stages. These criteria further trimmed down the initial set to a group of 1,982 retrotransposons (1,266 LTRs and 716 LINEs) with potential Pol II promoter activity.

**Figure 2.1.** Scheme of the bioinformatic approach for identification of potential epialleles from retrotransposons. The database of approximately 1.77 million retrotransposons and 0.2 million mRNAs (version mm9) were intersected by using custom scripts and applying a set of defined filters. The intersection identified 1,982 retrotransposon (1,266 LTRs and 716 LINEs) and mRNA pairs which were then further intersected with the combined histone 3 lysine 4 trimethylation (H3K4me3) ChIP-Seq peak data derived from four different tissues and cell lines: adult liver, ES cell, MEF cell and NP cell. This last intersection provided some 143 candidate
metastable retrotransposons that have at least one mRNA originating in close proximity of its 5’-end and also have an overlapping H3K4me3 signal on it.

As an independent approach for finding retrotransposons with promoter activity, we performed a series of database searches using genome-wide ChIP-Seq data, which were designed to survey active (H3K4me3) regions in different types of cells, including ES (Embryonic Stem), NP (Neural Precursor), and MEF (Mouse Embryonic Fibroblast) cells (Mikkelsen et al., 2007), as well as adult liver (Hoffman et al., 2010). According to this survey, the majority of LINEs and LTRs (551,827) are not associated with an active histone mark (H3K4me3) but rather marked by repressive histone modifications (H3K9me3 or H3K20me3) and their genomic locations are mostly away from genes’ promoter regions (Figure 2.2B). However, a small fraction of LINEs and LTRs (4,226) are modified with an active mark (H3K4me3), and a large fraction of this set of retrotransposons (32%) tends to be located in close proximity (within 10 Kb upstream) to the promoter regions of genes (Figure 2.2A and B). This retrotransposon set with H3K4me3 was further intersected with the initially identified set of 1,982 retrotransposons with potential promoter activity (elements in proximity to mRNA TSS where the TSS of the mRNA is within -50 to 250 bp of the 5’ end of the retrotransposon). This intersection identified a set of 143 retrotransposons which contained both potential promoter activity and an active histone mark (Figure 2.1) and thereby can be considered excellent candidates for metastable epialleles (MEs). This set contains even greater fraction of retrotransposons (66%) located in close proximity (within 10 Kb upstream) to the promoter regions of genes (Figure 2.2B). It is also interesting to note that many retrotransposons are associated with active histone marks mainly in two cell or
tissue types, ES cells and adult liver, which might be indicative of the greater transcription rates in liver and ES cells even though MEF and NP cells are not as fully differentiated as liver cells are (Figure 2.2C).

**Figure 2.2.** Detailed bioinformatic analysis of the candidate and other relevant retrotransposons. (A) Verification of the presence of the H3K4me3 ChIP-Seq peak in the vicinity of a randomly chosen retrotransposon and its associated mRNA pair from the pool of 143 candidate metastable epialleles. The retro/mRNA pair associated with the gene Eps8R1 have been shown here. The figure shows a snapshot from the UCSC Genome Browser in the mm8 version of mouse genome where the ectopic transcript Eps8R1 was still annotated with the same name as the endogenous transcript Eps8l1. The current mm9 version of the UCSC Genome Browser, which has the proper annotation, does not have the “Broad H3 ChIPseq” track that was used to show the H3K4me3 ChIP-Seq peaks. The red arrow indicates the retrotransposon-driven ectopic transcript, the blue arrow indicates the endogenous transcript of Eps8R1 (annotated as Eps8l1), the orange arrow indicates the retrotransposon driving the ectopic transcript and the red box surrounds the confirmed H3K4me3 ChIP-Seq peak. This ChIP-Seq result also shows another peak around the endogenous promoter region marked by the blue box. (B) Percentage of retrotransposons in the three different pools based on their relative positions to annotated genes. The pools are: i) filtered set of retrotransposons that does not intersect with either mRNA or
H3K4me3 ChIP-Seq peaks (551,827 LTRs and LINEs), ii) filtered set of retrotransposons that intersects with H3K4me3 ChIP-Seq peaks (4,226 LTRs and LINEs) and iii) filtered set of retrotransposons that intersects with both mRNA and H3K4me3 ChIP-Seq peaks (143 LTRs and LINEs). (C) Four-way Venn diagrams showing the number of retrotransposons (LTRs and LINEs only) obtained for each type of cell lines or tissues during the separate intersection of H3K4me3 ChIP-Seq peak data with the retrotransposons having nearby mRNAs (143 elements) and with the retrotransposons that do not have nearby mRNAs (4,083 elements).

The identified set of 143 retrotransposons are comprised of 99 LTRs and 44 LINEs, and ERVK and L1 subfamilies make up the major subfamilies for LTRs and LINEs, respectively, with 47 ERVKs and 24 L1s (http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf). It is rather surprising though, that this set of 143 retrotransposons does not contain any copy of the mouse ERV family, the Early Transposons (ETns), some of which have been demonstrated to show significantly variable DNA methylation (Reiss et al., 2010). These retrotransposons are distributed among all the chromosomes except none are present on the Y chromosome. The genomic features surrounding these 143 retrotransposons were further examined using the UCSC genome browser (Figure 2.2A). According to this manual inspection, each of the 143 retrotransposons can be assigned exclusively to one of four groups based on their position relative to genes: a) 89 out of the 143 retrotransposons express transcripts from the promoter regions (within 10 Kb upstream) of endogenous genes. Thirty out of these 89 retrotransposons appear to function as completely alternative promoters for adjacent endogenous genes (Figure 2.2A), ultimately generating fusion transcripts between the retrotransposons and endogenous genes as seen in the viable yellow agouti locus. The rest 59 of the retrotransposons in this group are located very close to the endogenous promoters (within -1 to 0 Kb of the TSS or within the first exon); b) 21 out of the 143 retrotransposons also function as promoters for Pol II transcription in more gene-poor non-genic regions producing transcripts without any obvious
open reading frames within their sequences in most cases (http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf). Thus, the majority of these transcripts are novel and likely non-coding RNA genes; c) 34 of the 143 retrotransposons express transcripts from the inside of other endogenous genes as seen in the axin-fused kinky locus (Rakyan et al., 2003) (http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf). Interestingly, 11 out of these 34 retrotransposons have their transcriptional direction opposite to that of the endogenous genes, and thus the transcription of these elements might interfere with that of the associated endogenous genes; d) only 1 of the 143 retrotransposons express transcripts that originate from the downstream of an endogenous gene and is transcribed in the opposite direction of that gene. A list of the 143 retrotransposons (candidate MEs) along with their associated mRNAs and endogenous genes has been provided at http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-table1.xls. Overall, the series of bioinformatic approaches described above were implemented to identify an initial set of 143 retrotransposons, which were investigated as potential candidates for metastable epialleles in the following sections.

DNA methylation analyses on the identified retrotransposons

We analyzed the DNA methylation levels of thirteen randomly chosen retrotransposons from the set of 143 retrotransposons or candidate MEs (Figure 2.3 and Table 2.1). We also included two retrotransposons (associated the genes Rgs9 and Cog6) that have potential promoter activity but do not have any H3K4me3 mark (Table 2.1). Three additional loci were selected as controls for analysis: a house keeping gene (Gapdh), an imprinted gene (Peg3), and one IAP element (located within the 6th intron of Cdk5rap1) that was previously shown to be a metastable
epiallele (Druker et al., 2004). The DNA methylation levels of these retrotransposons were analyzed using a set of mouse tail DNA derived from one litter of five 1-week-old pups with an isogenic background (C57BL/6J).

![Figure 2.3](image)

**Figure 2.3.** Combined Bisulfite Restriction Analysis (COBRA) of randomly selected candidate metastable retrotransposons in five 1-week-old littermates. Tail DNAs isolated from five 1-week-old littermates of the C57BL/6J strain were bisulfite-treated, PCR-amplified and digested with restriction enzymes. The relative methylation status of the CpG sites at the enzyme recognition sites for each PCR product were judged from the intensities or the relative proportions of the bands obtained after gel electrophoresis. The endogenous genes associated with the mRNA/retrotransposon pairs are indicated on the left and the restriction enzymes used are indicated on the right of the electrophoresis gel picture. The bands expected in case of methylated and unmethylated states of the PCR amplicon after the enzyme digestion are marked with M and U, respectively. The promoter region of Gapdh and the DMR of Peg3 were used as negative controls and the IAP-LTR element associated with Cdk5rap1 was used as a positive control. The mRNA/retrotransposon pairs associated with the endogenous genes Rgs9 and Cog6 are H3K4me3 mark-deficient.

The DNAs were first treated with the bisulfite conversion method (Clark et al., 1994) and subsequently used for PCR amplification, for which one of the primers was designed from within the retrotransposon itself while the other primer from the adjoining endogenous and non-repeat
region. The amplified products were digested with their appropriate restriction enzymes, the digestion or non-digestion of which indicate the methylated or unmethylated status of the original DNAs (Xiong and Laird, 1997) (COBRA, Combined Bisulfite Restriction Analysis).

In order to test the completeness of digestion of the restriction enzymes, all the enzymes used in this experiment and throughout the study were made to digest specific PCR products (amplified from non-bisulfite treated genomic DNA) that had the restriction recognition site of the enzyme concerned (http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf).

### Table 2.1. Candidate metastable retrotransposons tested in the COBRA of 1-week-old littermates

<table>
<thead>
<tr>
<th>Associated Endogenous UCSC Gene</th>
<th>Coordinates</th>
<th>Retrotransposon</th>
<th>Associated mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rgs9*</td>
<td>chr11: 109165513-109165759</td>
<td>IAPEY3_LTR ERVK LTR</td>
<td>AK166518</td>
</tr>
<tr>
<td>Cog6*</td>
<td>chr3: 52798218-52798593</td>
<td>IAPLTR2_Mm ERVK LTR</td>
<td>BC071236</td>
</tr>
<tr>
<td>Gimap9</td>
<td>chr6: 48625715-48626103</td>
<td>RLTR9F ERVK LTR</td>
<td>AK041336</td>
</tr>
<tr>
<td>AK145129</td>
<td>chr2: 118380501-118380977</td>
<td>IAPLTR2_Mm ERVK LTR</td>
<td>AK145129</td>
</tr>
<tr>
<td>AK011682</td>
<td>chr11: 113034330-113034803</td>
<td>IAPLTR2_Mm ERVK LTR</td>
<td>AK011682</td>
</tr>
<tr>
<td>Bcat2</td>
<td>chr7: 52826548-52826692</td>
<td>LTR65 ERV1 LTR</td>
<td>AK160159</td>
</tr>
<tr>
<td><strong>Table 2.1 continued</strong></td>
<td><strong>Associated Endogenous UCSC Gene</strong></td>
<td><strong>Coordinates</strong></td>
<td><strong>Retrotransposon Name</strong></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------</td>
<td>-----------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Eps8R1</strong></td>
<td>chr7: 4411952-4412304</td>
<td>IAPLTR1_Mm ERVK</td>
<td>LTR</td>
</tr>
<tr>
<td><strong>Ogdh</strong></td>
<td>chr11: 6192483-6192694</td>
<td>LTR31 ERV1 LTR</td>
<td>BC049104</td>
</tr>
<tr>
<td><strong>Ms13l2</strong></td>
<td>chr10: 55826590-55826754</td>
<td>RMER15 ERVL LTR</td>
<td>BC037487</td>
</tr>
<tr>
<td><strong>AK009289</strong></td>
<td>chr10: 95143475-95143681</td>
<td>RLTR18B ERVK LTR</td>
<td>AK009289</td>
</tr>
<tr>
<td><strong>Alkbh8</strong></td>
<td>chr9: 3334707-3335338</td>
<td>Lx9 L1 LINE</td>
<td>BC050863</td>
</tr>
<tr>
<td><strong>Gsta3</strong></td>
<td>chr1: 21234684-21235118</td>
<td>RodERV21-int ERV1</td>
<td>LTR</td>
</tr>
<tr>
<td><strong>Popdc2</strong></td>
<td>chr16: 38376981-38377439</td>
<td>MTC MaLR LTR</td>
<td>AK171056</td>
</tr>
<tr>
<td><strong>Repin1</strong></td>
<td>chr6: 48548858-48549049</td>
<td>L3 CR1 LINE</td>
<td>AK085289</td>
</tr>
<tr>
<td><strong>Spsb3</strong></td>
<td>chr17: 25024291-25024393</td>
<td>L2 L2 LINE</td>
<td>AF403038</td>
</tr>
</tbody>
</table>

* The retrotransposons associated with the genes *Rgs9* and *Cog6* are H3K4me3 mark-deficient ones and are not part of the candidate metastable retrotransposon list.
As shown in **Figure 2.3**, the promoter region of a housekeeping gene (*Gapdh*) was not digested by TaqI, indicating the unmethylation status of this gene. In contrast, the Differentially Methylated Region (DMR) of one imprinted gene (*Peg3*) showed about 50% digestion with *HphI* and thus 50% methylation, consistent with the fact that only one allele, the maternal allele, is methylated (Huang and Kim, 2009; Kim et al., 2009). As expected, the known metastable IAP (*Cdk5rap1*) showed a partial methylation pattern with inter-individual variation: Mouse #1 and 5 had much lower methylation levels than their littermates, as evident from the intensities of the bands expected for methylated and unmethylated products after digestion. Among the thirteen candidate MEs (10 LTRs and 3 LINEs) tested, which does not include the H3K4me3 mark-deficient *Rgs9* and *Cog6*, only one locus showed complete unmethylation (*Gimap9*) whereas the other twelve loci showed partial methylation patterns, mostly unmethylated, with potential inter-individual variations based on the detection of different levels of digestion with the same enzyme for a specific locus in some of the cases (*Eps8R1*, *AK145129*, *Msl3l2*, *Gsta3*, and *Repin1*). We also performed a second set of COBRA using the same DNA samples for some of the thirteen candidate MEs and also *Rgs9* depending upon the availability of a different enzyme site within the retrotransposons ([http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf](http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf)). As observed in the first set of COBRA, most of the loci showed unmethylation while the loci *Repin1* showed mostly methylation.

Since COBRA, however, is limited to the methylation status of a single CpG site, we further followed up this possibility through cloning and sequencing the PCR products of five of the loci (*Bcat2, Eps8R1, Ogdh, AK011682*, and *AK145129*) for more extensive and accurate DNA methylation analysis (**Figure 2.4**).
Figure 2.4. Bisulfite sequencing of five highly potential candidate metastable retrotransposons in the tail of 1-week-old littermates. The PCR products of the bisulfite-treated tail DNA of the five 1-week-old littermates were cloned and sequenced to check the methylation status of the CpG sites of the potential metastable retrotransposons associated with the genes Bcat2, Eps8R1, Ogdh, AK011682, and AK145129. The bisulfite sequencing result has been shown using a bubble chart. Each row represents a different clone and each column represents a different CpG site. Filled and open circles indicate methylated and unmethylated cytosines, respectively. The CpG sites are arranged in the 5’ to 3’ direction of the sequenced region while going from left to right in each clone. The numbers on the far left of the figure represent the mouse number corresponding to the COBRA (Figure 2.3). The red boxes encompass the CpG sites that are within the retrotransposon. In case of AK011682, all the CpG sites shown are within the retrotransposon. The percentage (without the parentheses) on the right of each bubble chart indicates the percentage methylation of the CpG sites within the retrotransposon itself and the percentage within the parentheses indicates the percentage methylation of the CpG sites within the retrotransposon and adjoining regions in that specific sequenced DNA sample. The CpG sites analyzed in the first set of restriction enzyme digestion are marked by brown inverted triangles while those analyzed in the second set are marked by blue inverted triangles.
The results confirmed variable levels of DNA methylation in the retrotransposons among the individual mice: 0 to 44% for Bcat2, 26 to 72% for Eps8R1, 0.3 to 53% for AK011682, and 19 to 46% for AK145129. Even though the retrotransposon associated with Ogdh shows no significant variable methylation by itself (0 to 6%), the retrotransposon along with the flanking regions does show some variation in DNA methylation: 15 to 26%. Analyses of the DNA methylation of the flanking regions of the retrotransposons show some distinct patterns for at least three of the loci. The flanking region upstream of the retrotransposon in the loci Bcat2 shows a methylation pattern independent of that of the retrotransposon itself while the methylation pattern of the downstream flanking region mimics more or less that of the retrotransposon. The downstream flanking region (the only sequenced flanking region) of the retrotransposon in the loci Eps8R1 shows methylation pattern that is similar to that of the retrotransposon itself. The methylation of the flanking regions of the retrotransposon in the Ogdh loci show some interesting patterns: the upstream flanking region has a similar methylation pattern as the retrotransposon itself while the downstream flanking region shows substantial and variable methylation with an abrupt boundary at the border of the retrotransposon. In order to evaluate the statistical significance of the methylation variations of the CpG sites between individual mice for each of the five loci, we performed the Mann-Whitney U-test (also known as the Wilcoxon rank-sum test) on the bisulfite sequences of two individuals at a time in all combinations using the free online tool QUMA (Kumaki et al., 2008). The p-values obtained from the tests showed significant to moderate methylation variations among the individuals for the loci Bcat2, Eps8R1, and AK011682, while the loci Ogdh and AK145129 did not show any statistically significant variations (http://jookimlab.lsu.edu-supplemental/2011EPI0251R2-Sup-Table.pdf). In contrast to these five retrotransposons with variable DNA methylation patterns, the two retrotransposons without any
H3K4me3 mark (associated with the genes Rgs9 or Cog6) showed almost complete methylation with very minute intensities for unmethylated fragments in COBRA (both sets in case of Rgs9) and this was confirmed by cloning and sequencing the PCR product of one of those loci, Cog6, which showed almost complete methylation for all five littermates (Figure 2.3 & http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-Sup.pdf). Overall, the preliminary results confirmed that 12 out of the 13 chosen candidate MEs are partially methylated with either equal or more amount of unmethylation. The results, particularly the bisulfite sequencing, also demonstrated that some of these retrotransposons indeed display inter-individual variations in their DNA methylation levels, further supporting the feasibility of our approach.

We also analyzed the DNA methylation levels of the five candidate MEs that were sequenced in more detail along with the two H3K4me3 mark-deficient retrotransposons using the DNA derived from different tissues of two adult mice (2- and 5-month-old, Figure 2.5). The two H3K4me3 mark-deficient LTRs (Rgs9 and Cog6) had almost complete methylation patterns in all the tissues tested for both ages, which is similar to the patterns seen in 1-week-old mice (Figure 2.3). Two candidate MEs (AK011682, and AK145129) had similar levels of partial methylation patterns throughout all different tissues at both ages, which is also similar to the patterns seen in the neonatal set.

However, the DNA methylation levels of the three remaining candidate MEs (Bcat2, Eps8R1, and Ogdh) showed quite different patterns of DNA methylation among the tissues analyzed. The candidate ME associated with the Bcat2 gene had very reduced methylation in lung and spleen at both ages in addition to the tail at 5-month-old stage, whereas partial methylation was observed
in the other tissues. The candidate ME associated with the *Ogdh* gene had very reduced methylation in liver (particularly the 5-month-old liver), lung, and kidney at both ages, whereas it contained partial methylation in brain, tail, and spleen.

**Figure 2.5.** COBRA of the identified metastable retrotransposons in different tissues from 2- and 5-month-old mice. DNA methylation analysis of the identified metastable retrotransposons during different growth stages was done by COBRA on DNA obtained from six different tissues of two individual mice at two different ages (2 and 5 months). The tissues were chosen from three germ layers: ectoderm (brain and tail), endoderm (liver and lung), and mesoderm (kidney and spleen). As in the previous COBRA, *Gapdh* and *Peg3* were used as negative controls whereas the IAP-LTR associated with *Cdk5rap1* was used as a positive control. The names of the tissues are indicated at the top of each column. The associated genes, the restriction enzymes used and the expected bands in case of methylation and unmethylation are mentioned as in Figure 2.3.

However, the candidate ME associated with the *Eps8R1* gene did not have any stark unmethylation like the other two loci but did show significant unmethylation in lung at both ages as well as in tail, kidney, and spleen at 5-month-old stage, whereas partial methylation was
observed in the other tissues. Overall, the four retrotransposons (*Rgs9, Cog6, AK011682, and AK145129*) analyzed seem to maintain similar patterns of DNA methylation in the adult tissues as was previously observed in the viable yellow *agouti* locus, but interestingly three of the retrotransposons (*Bcat2, Eps8R1, and Ogdh*) displayed tissue-specific, fluctuating DNA methylation patterns which might be reminiscent of tissue dependent methylation profiles observed in the axin-fused kinky locus (Waterland et al., 2006). The possible reason for the hypomethylation at some of the 5 month-old stage tissues could be due to a combination of or either of the two phenomena: a progressive loss of methylation during the aging process and the presence of inter-individual variations among the 2 month- and 5 month-old mice.

**Developmental profiling of DNA methylation patterns of selected retrotransposons**

The DNA methylation levels of the same set of retrotransposons were further analyzed using the sperm, oocyte, and blastocyst-stage embryos derived from the C57BL/6J mice (*Figure 2.6*). For this series of COBRA, we isolated the sperm, individually, from 4 male littermates, a pool of 60 mature oocytes from three female littermates, and finally a pool of 10 blastocyst-stage embryos after timed mating. The DNA purified from these samples was subsequently analyzed in a similar manner as described above. To monitor the purity of germ cells, we first analyzed the DMRs of a control set, two imprinted genes *H19* (maternally expressed) and *Peg3* (paternally expressed). As shown in *Figure 2.6*, the DMRs of *H19* and *Peg3* were completely unmethylated in the oocyte and sperm, respectively, confirming the purity of the germ cells. Thus, we analyzed the DNA methylation levels of the 7 retrotransposons using the DNA purified from germ cells and early-stage embryos.
Figure 2.6. COBRA of the identified metastable retrotransposons in germ cells and blastocyst. Analysis of DNA methylation status of the newly identified metastable retrotransposons at different developmental stages. COBRA was performed on DNA obtained from sperm of four male littermates, combined oocytes from three female littermates and blastocyst-stage embryos from mating pairs of the same litter. Gapdh, Peg3 (in sperm), and H19 (in oocyte) were used as negative controls whereas the IAP-LTR associated with Cdk5rap1 was used as positive control. The associated genes, the restriction enzymes used and the expected bands in case of methylation and unmethylation are mentioned as in Figure 2.3.

We summarized the results from this series of analyses (Figure 2.6) along with the results from the neonates and the tissues of adult stages (Figure 2.3 and Figure 2.5), which subsequently provide a developmental overview of DNA methylation patterns for the tested retrotransposons (Figure 2.7).
Figure 2.7. Model of progression of DNA methylation status of retrotransposons through different developmental and growth stages. Four different patterns of methylation change throughout the developmental and growth stages have been proposed for retrotransposons after detailed DNA methylation analysis for the candidate metastable retrotransposons (those associated with genes or mRNAs: Bcat2, Eps8R1, Ogdh, AK011682, and AK145129), the established metastable retrotransposon (Cdk5rap1), and for the retrotransposons that do not intersect with any H3K4me3 data (those associated with Rgs9 and Cog6). The methylation state for each representative colored pie-chart has been described in the figure.

According to the three result sets, the retrotransposons display four developmental patterns of DNA methylation. First, the two H3K4me3 mark-deficient LTRs (Rgs9 and Cog6) display complete methylation throughout development, consistent with the fact that these two LTRs are not associated with H3K4me3 mark in any cell types examined. Second, one candidate ME (AK011682) shows almost complete unmethylation throughout development, including germ cells and blastocysts (Figure 2.6). Third, three candidate MEs (Bcat2, Ogdh, AK145129) show a pattern of progressive gaining of DNA methylation although the candidate MEs associated with
AK145129 appears to be partially methylated in the sperm. The endogenous genes associated with this group (Bcat2 and Ogdh) are expressed ubiquitously and thus the unmethylated status of the associated candidate MEs in germ cells and blastocysts somewhat agrees with the ubiquitous expression pattern of the endogenous genes. Fourth, the two candidate MEs (Eps8R1 and Cdk5rap1) start with complete methylation in the germ cells and/or blastocysts, but progressively lose their DNA methylation levels in the tissues of neonatal and adult stages. The endogenous genes associated with this group are tissue- and stage-specific genes, and thus the complete methylation of the associated candidate MEs in early stages also reflects the promoter activity of the endogenous genes. Overall, it is prudent to note that the DNA methylation patterns of the latter two groups of retrotransposons are somewhat consistent with the expression patterns of the associated endogenous genes. Furthermore, according to the results from detailed sequencing (Figure 2.4), the retrotransposons with relatively greater levels of inter-individual variations belong to these two groups, the groups with progressive gaining and losing of methylation. This observation suggests that the fluctuating DNA methylation patterns embedded in the associated endogenous genes might be a contributing factor to the inter-individual variations observed in the associated retrotransposons.

**Ectopic and variable expression of the endogenous genes by the associated retrotransposons**

The three candidate MEs (Bcat2, Eps8R1, and Ogdh) were further analyzed to test whether they trigger ectopic expression of their associated endogenous genes and whether that ectopic expression is variable or not at the same time. For this series of analyses, three tissues (brain, kidney, and lung) from three 5-month-old littermates were used for isolating total RNA and subsequent qRT-PCR analyses (Figure 2.8).
Figure 2.8. Verification and quantification of the ectopic transcripts with accompanying DNA methylation analysis. (A) Quantitative RT-PCRs were carried out for both the retrotransposon-driven ectopic transcripts and the endogenous transcripts for the genes Bcat2, Eps8R1, and Ogdh in brain, kidney, and lung of three 5-month-old littermates. The bands shown are the qRT-PCR products from the brain of the three mice (#1, 2, and 3) with the internal control β-actin at the top followed by the ectopic and the endogenous transcripts of the three loci as mentioned on the left. The exon structure of each gene is shown at the top of the gel bands with the location of the primers used for amplifying the ectopic and the endogenous transcripts marked by the red and green arrows, respectively. The exon 1 of the ectopic and the endogenous transcripts are indicated by the orange and blue boxes, respectively. (B) The fold differences of the ectopic transcripts with respect to the control β-actin are shown for the gene Eps8R1 in the three different tissues and three different animals (Ec #1, Ec #2, and Ec #3) as described before. (C) The COBRA data (using the restriction enzyme HpyCH4IV) of the three different tissues and the bisulfite sequencing data of the brain Eps8R1 ectopic transcript in the three different animals. In the electrophoresis gel picture, the bands expected after digestion in case of methylated and unmethylated CpGs are indicated by M and U, respectively. The bubble charts depict the sequencing results and the percentages indicate the percentage methylation in the same way as described in Figure 2.4. The red boxes encompasses the CpG sites that are within the retrotransposon while the numbers (#1, 2, and 3) refers to corresponding individual animals in both the pictures.
All of the three candidate MEs indeed trigger transcription based on the detection of the PCR products that are derived from fusion transcripts of the LTRs and the associated endogenous genes (Figure 2.8A). The LTR-driven ectopic expressions were detected in all three tissues examined in at least two out of the three animals. The relative expression levels of the ectopic versus endogenous transcripts were further measured for Bcat2 and Ogdh since the endogenous expression of these two genes was detected in all three tissues. The expression levels of the ectopic expression were very minimal, with ranges of 0.18-7.0 % (Bcat2) and 0.61-4.3 % (Ogdh) of the respective endogenous genes based on their relative C\textsubscript{T} (threshold cycle) values. Given the very low expression levels, the LTR-driven ectopic expression is thought to have minimal impact on the function of the Bcat2 and Ogdh loci. In the case of Eps8R1, out of the three mice tested, the expression of the endogenous transcript was detected only in the kidney of Mouse #2 as well as in all three tissues of Mouse #3, and yet the LTR-driven expression was readily detectable in all three tissues from all three mice (Figure 2.8A). Thus, the LTR-driven ectopic expression might have a significant impact on the function of the Eps8R1 gene.

Given the observations described above, we further tested the presence of potential inter-individual variations associated with the Eps8R1 locus. For each tissue, the expression level of the LTR-driven transcript was first normalized with that of \( \beta\)-actin, and later compared between the normalized levels of the three individual mice. As shown in Figure 2.8B, the expression levels of the LTR-driven transcript were relatively higher in Mouse #1 and 3 than in Mouse #2 for all three tissues examined. We also analyzed the DNA methylation levels of the LTR using the DNA isolated from the three adult mice (Figure 2.8C). As depicted by the COBRA of all three tissues and the bisulfite sequencing of the brain DNA from all three animals, the DNA
methylation levels of the LTR were indeed much lower in Mouse #1 and 3 than Mouse #2 in all three tissues, which is consistent with the higher expression levels observed in Mouse #1 and 3. Such an observation might indicate that the methylation might have been set prior to the divergence of the three germ layers from which the three tissues were derived from. Any relation of the sex of the mouse with the methylation status can also be ruled out because Mouse #1 and 2 are males and Mouse #3 is female. Overall, the results described above clearly demonstrate that the identified retrotransposons have inter-individual variations in terms of their DNA methylation levels, and further that some of these variations in DNA methylation levels have concurrent outcomes on the transcription levels of the nearby endogenous genes.

**Discussion**

In the current study, we have identified a set of potential metastable epialleles (MEs) that are derived from retrotransposons using a series of bioinformatic approaches. We also analyzed a subset of these potential metastable epialleles with DNA methylation and expression analyses. According to the results, some of the analyzed candidates are most likely epialleles with functional consequences on the nearby genes. This further implicates that mammalian genomes may contain a large number of uncharacterized metastable epialleles. Also, given the sheer number of retrotransposons in the mammalian genome, these elements may be a major source of inter-individual epigenetic variations resulting in more subtle phenotypic or physiological variations.

The bioinformatic approach used in this study has been successful for finding many epialleles. Two criteria are thought to be instrumental for this success. First, we sought to identify
retrotransposons (LINEs and LTRs only) that are located very close to the 5’-end of mRNAs. This approach revealed that about 2.5% of the entire transcriptome (5,679 out of 228,765) is likely initiated or influenced by these retrotransposons (Figure 2.1). Initially, this estimate was somewhat surprising in that retrotransposons could affect such a large number of genes in the transcriptome, but this fact has been independently confirmed by another study (Faulkner et al., 2009). Thus, mammalian retrotransposons may have reshaped not only the landscape of genomes but also the functions of individual genes (Belancio et al., 2008; Goodier and Kazazian, 2008; Cohen et al., 2009; Cordaux and Batzer, 2009). Despite this surprising estimate, however, our follow-up DNA methylation analyses suggest that many in this group of retrotransposons might be completely methylated in normal tissues without any inter-individual variations (Rgs9 and Cog6 in Figure 2.3). This might reflect the fact that this group of retrotransposons likely triggers Pol II transcription only in very unusual cell populations, such as cancer cells. This observation is consistent with the fact that a large fraction of the transcriptome is indeed derived from these types of cells. Second, as a measure to correct this bias in the transcriptome, we decided to utilize active histone modification (H3K4me3) profiles derived from ES and other cells, which likely have a significant impact on the DNA methylation levels in normal somatic tissues. A large number of retrotransposons (4,226) are indeed found to be associated with active histone marks. Furthermore, a minor fraction (143) of this group was already proven to function as promoters for Pol II transcription based on their fused exon structures between the 5’-side exons derived from retrotransposons and the 3’-side exons from endogenous genes. Interestingly, some of the retrotransposons that have been tested in this small set of 143 turns out to be epialleles with variable levels of DNA methylation among littermates of an inbred mouse (Figure 2.3 and Figure 2.4). Surprisingly though, it was shown in a recent study that for the
viable yellow agouti (A\textsuperscript{vy}) mice, groups of yellow and pseudoagouti mice did not exhibit any variations of the H3K4me3 mark in the LTR of the A\textsuperscript{vy} metastable epiallele in spite of showing variable DNA methylation (Dolinoy et al., 2010). It is also important to note that some of the retrotransposons that have been proven to function as promoters for Pol II transcription might not have associated active histone marks. Nonetheless, combining the two criteria, filtering retrotransposons based on their close proximity to the 5’-ends of mRNA and their association with active histone marks, turns out to be a good method to predict epialleles with a high rate of success.

The mechanistic basis by which a subset of retrotransposons escapes DNA methylation is still poorly understood and furthermore the reasons why the levels of this DNA methylation are variable among different individuals are also unknown. However, given the large number of potential epialleles identified in this study, we surmise that these unusual epigenetic features may be closely associated with the three factors described below. First, the location of a given retrotransposon appears to be important based on the fact that a large number of epialleles are located closely to the promoter regions of endogenous genes. An opportunistic insertion into a promoter region likely protects the inserted retrotransposon from DNA methylation. Second, this opportunistic position may also be accountable for the variable levels of DNA methylation observed from epialleles. The vast majority of retrotransposons are destined to be repressed by DNA methylation in the mammalian genomes (Belancio et al., 2008; Goodier and Kazazian, 2008; Cohen et al., 2009; Cordaux and Batzer, 2009). This general repressive trend might be in direct conflict with the unusual but opportunistic position of promoter proximity, resulting in an arms race between DNA methylation versus protection from methylation. This opposing
pressure might be a reason for the variable levels of DNA methylation that has been observed in some of the retrotransposon-derived epialleles. Third, the variable levels of DNA methylation observed in retrotransposons may also be influenced by the transcriptional programs embedded within their associated endogenous genes. Among several candidate MEs (Figure 2.3 and Figure 2.4), the retrotransposons that are associated with tissue- and stage-specific genes (Eps8R1 and Cdk5rap1) tend to show greater degrees of variations in DNA methylation levels among individual mice. Dynamic transcriptional programs in these tissue- and stage-specific genes during development might provide the associated retrotransposons more opportunities to accumulate DNA methylation variations than the other type of genes with static transcriptional programs. Overall, the position of retrotransposons and the transcriptional patterns of the associated genes are thought to be major contributing factors to the metastable nature of DNA methylation associated with epialleles.

It is currently unknown to what extent epigenetic variations contribute to phenotypic variations that are often observed among human populations. Many recent studies, however, suggest that a large fraction of human genetic disorders, in particular disorders with low penetrance, might be accountable for by the epigenetic variations (Feinberg et al., 2006; Whitelaw and Whitelaw, 2006; Esteller, 2007; Jirtle and Skinner, 2007). This is particularly likely since traditional genetic studies have focused mainly on genetic variations but not on epigenetic variations. In a similar line, it is equally important to note that the candidate and the confirmed epialleles identified in the current study are derived from retrotransposons, which tend to be ignored by traditional genetic studies as junk DNA with no functional relevance to human health. Nevertheless, the results from the current study highlight the possibility that retrotransposons
may have a much greater impact than previously thought on the epigenetic setting of mammalian genomes. In this regard, it is prudent to note that about 4,000 retrotransposons have been identified from the mouse genome based on their unusual association with active histone marks (Figure 2.2). A good number of retrotransposons in this group might be epigenetically unstable and likely contribute to epigenetic variations in the mouse genome. We also predict that a similar number of retrotransposons may exist as epialleles in the human genome since all the mammalian genomes have similar genomic makeup in terms of repeat and gene content. It will be very interesting to test whether the majority of these elements are indeed epigenetically metastable. If that is the case, this group of retrotransposons should provide a large set of epigenetically useful DNA elements, some of which might already explain potential phenotypic variations based on their chromosomal linkage to endogenous genes with disease connections. At the same time, this group of retrotransposons might also serve as epigenetic biomarkers for the diagnosis and/or prognosis of adult on-set diseases, such as diabetes and cancers, since the epigenetic status of these DNA elements is prone to be affected by environmental exposure during development (Finer et al., 2011).

**Materials and Methods**

**Data set construction**

Mouse (mm9) mRNA dataset (track: ‘Mouse mRNAs’; table: ‘all_mrna’) and repeat element dataset (track: ‘RepeatMasker’; table: ‘rmsk’) were downloaded from Table Browser of the UCSC Genome Bioinformatics website (http://genome.ucsc.edu/cgi-bin/hgTables?org=Mouse&db=mm9). The mRNA dataset contains coordinates of the pre-spliced mRNA only. Genome-wide histone lysine 4 trimethylation (H3K4me3) ChIP-Seq
A dataset of adult mouse liver was downloaded from the website of Canada’s Michael Smith Genome Sciences Centre (http://www.bcgsc.ca/data/histone-modification/histone-modification-data) (Hoffman et al., 2010). The dataset was already available as ChIP-Seq peaks which were generated by the tool FindPeaks 2.0 (Fejes et al., 2008). The H3K4me3 ChIP-Seq datasets for the V6.5 embryonic stem (ES) cell, embryonic day 13.5 mouse embryonic fibroblast (MEF) cell and ES-derived neural precursor cell (NP) were downloaded from the GEO data repository (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12241) (Mikkelsen et al., 2007). Since the authors for these ChIP-Seq datasets supplied the peaks that were generated using a different tool than the one used for the liver ChIP-Seq dataset, we regenerated the peaks for these cell lines using the same Findpeaks tool but using a newer version (FindPeaks 4.0) that is available now. All the peak datasets were subsequently converted to the mm9 version from their original mm8 version using the Lift-Over tool of the Galaxy genome analysis website (http://main.g2.bx.psu.edu/). Stretches of 100 bp were then added to both ends of each peak coordinates of all the ChIP-Seq datasets.

**Dataset intersections and computational analyses**

In the very first step of this bioinformatic approach, the mRNA and the repeat element datasets were uploaded to a MySQL database in order to run a series of coordinate intersections and filtering steps to identify the LTRs and LINEs that intersect with the transcription start sites (TSS) of mRNAs. The identified intersection result must also fit into the following criteria: i) the TSS of the mRNA is within 250 bp downstream and 50 bp upstream of the 5’-end of the retrotransposon, ii) the retrotransposon must be at least 50 bp in length and iii) the
retrotransposon must have at least 3 CpG sites in its sequence. In the cases where multiple mRNAs transcribing in the same orientation could have been mapped to a particular retrotransposon, only the mRNA that had its TSS closest to the retrotransposon 5’-end and also gave the longest transcription product was kept. For the retrotransposons that had two mRNA transcribing in opposite orientation, only one of the mRNAs was retained after a manual inspection of the surrounding genomic features. A full schematic representation for the preparation of this retro/mRNA pair dataset has been shown in http://jookimlab.lsu.edu-supplemental/2011EPI0251R2-Sup.pdf. For the intersection with histone modification data, the retro/mRNA pair dataset and all the H3K4me3 ChIP-Seq peak datasets were uploaded to the Galaxy genome analysis website (http://main.g2.bx.psu.edu/) and intersected using the ‘Intersect’ tool of the ‘Operate on Genomic Intervals’ section of the website. For the analyses of the pool of retrotransposons in which the retrotransposons did not intersect with the mRNA TSS and/or histone modification datasets, the LTR and LINE pool was created prior to those intersections from our constructed MySQL repeat element database using the following criteria: i) the retrotransposon must be at least 50 bp in length and ii) the retrotransposon must have at least 3 CpG sites in its sequence. These retrotransposon datasets were also intersected with the histone modification datasets using Galaxy in the same way mentioned earlier. In the analyses of the positions of the retrotransposons relative to the UCSC Known Gene dataset, the retrotransposons were considered to be in the promoter regions if their 5’-ends were located within 10,000 bp upstream of the TSS of the gene and the 3’-end of the Exon 1 of the gene.
Isolation and purification of sperm, oocyte and blastocyst-stage embryos

Mature sperms were isolated from the epididymis of adult C57BL/6J mice by following the ‘swim up’ method (Bunch and Saling, 1991). Briefly, freshly dissected and chopped epididymis was incubated in sterile PBS at 37°C for 30 minutes to allow the sperms to swim up to the top of the container. Following the incubation, PBS from the upper part of the container was transferred to a new container and centrifuged at 600 rpm for 5 minutes. The sperm cells were again allowed to swim up by having an incubation period at 37°C for 30 minutes. Following this round of incubation, the PBS in the upper part of the container was extracted and a small volume of it was observed under the microscope to check for purity (less somatic cells). Successive rounds of centrifugation, incubation and transfer were done until sperms of high purity (devoid of almost any somatic cells) were obtained. Finally the sperm cells were concentrated by centrifugation and resuspended in PBS for DNA methylation analyses.

For the isolation of mature oocytes in the MII phase, we followed the widely used superovulation protocol (Eppig and Telfer, 1993; Hogan et al., 1994). In brief, female C57BL/6J mice were first injected subcutaneously with 5 IU of the Pregnant Mare’s Serum (PMS) (Cat. G4877, Sigma-Aldrich), and the same mice were injected again 48 hours after the initial injection with 5 IU of the human Chorionic Gonadotropin (hCG) hormone (Cat. C1063, Sigma-Aldrich). These mice were sacrificed 12 h after the second injection, and mature oocytes were isolated from the swollen ampullae of the oviducts. The isolated eggs were incubated in hyaluronidase solution (Cat. H3506, Sigma-Aldrich) for several minutes to separate them from the cumulus cells, and subsequently washed three additional times to remove potential somatic tissue contamination. Finally, the oocytes were used to isolate DNA.
The blastocyst-stage embryos were also isolated using the same ovulation protocol as above. After the second injection, however, the female mice were paired with male littermates (both were of C57BL/6J strain). The female mice were then sacrificed after 3 days following the breeding setup, and the blastocysts were isolated from their uterus by flushing with PBS (Hogan et al., 1994). The isolated blastocysts were further examined under the microscope to estimate their developmental stages and purity, and subsequently washed three additional times to remove potential maternal tissue contamination. The blastocysts were then used for DNA isolation.

**COBRA (COmbined Bisulfite Restriction Analysis) and bisulfite sequencing**

The methylation levels of the CpG sites of a specific locus were analyzed using two widely used DNA methylation analyses, COBRA (COmbined Bisulfite Restriction Analysis) and bisulfite sequencing analysis. For the purpose of isolation and purification of genomic DNA from the six different adult tissues (brain, tail, liver, lung, kidney, and spleen), sperm, oocyte, and blastocyst-stage embryos homozygous C57BL/6J mice obtained from the Jackson Laboratory (http://www.jax.org/) were used. All experiments were performed in accordance with the National Institutes of Health guidelines for care and use of animals. For each type of DNA sample, 2 μg of the genomic DNA was modified using the bisulfite conversion reaction according to the manufacturer’s protocol (EZ DNA methylation™ kit, Zymo Research). The converted DNA (1 μl) was used as template for PCR reactions (Maxime PCR Premix Kit, Intron Biotech) using specific primer sets that lacked any CpG dinucleotides in the original pre-converted sequence and were designed to specifically amplify the bisulfite-converted (C-to-T converted) DNAs. Information regarding the primer sequences and detailed PCR conditions for
each tested loci has been provided in http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-table2.xls. The amplified DNAs were further used for the COBRA and bisulfite sequencing analysis.

For the COBRA, the amplified DNAs were digested using appropriate restriction enzymes (New England Biolabs) according to the manufacturer’s protocol in order to measure the methylation levels of particular CpG sites of the specific loci. All the genomic coordinates and restriction enzymes used for the regions tested are available in http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-table2.xls. For the bisulfite sequencing analysis, the amplified DNAs were separated by agarose gel electrophoresis and then purified by MEGA-spin™ Agarose Gel Extraction Kit (Intron Biotech). After cloning the purified DNAs into pGEM-T Easy Vector Systems (Promega), the cloned DNAs were isolated by using the DNA-spin™ Plasmid DNA Purification Kit (Intron Biotech) and sequenced with the ABI 3130XL DNA Sequencer (AME Bioscience). The sequences were analyzed and aligned using the software BiQ Analyzer v2.00 which also ensured that 1) clones with identical methylation patterns were in fact different clones by comparing the positions of the unconverted C’s in the clones, 2) each bisulfite sequence has a sequence identity of at least 80% with the genomic sequence, and 3) the bisulfite conversion rate of each sequence exceeds a minimum of 90% (Bock et al., 2005).

RT-PCR and quantitative real time PCR

Total RNA was isolated from the brain, kidney and lung of three 5-month-old littersmates (Mouse #1 and 2: male; Mouse #3: female) using the Trizol (Invitrogen) reagent as per the
manufacturer’s protocol. The isolated RNA was reverse-transcribed using the M-MLV Reverse Transcriptase (Invitrogen) and other compatible reagents from the manufacturer. PCR amplifications of the cDNA were performed with specific gene primer sets [http://jookimlab.lsu.edu/supplemental/2011EPI0251R2-table2.xls] using the Maxime PCR premix kit (Intron Biotech). Quantitative real-time PCR was also performed with the iQ SYBR Green Supermix (Bio-Rad) using the iCycler iQTM multicolor real-time detection system (Bio-Rad). All qRT-PCR reactions were carried out for 40 cycles under the standard PCR conditions. \(\beta\)-actin was used as the internal control. The results were analyzed based on the threshold (C\(_t\)) value. Each reaction was checked for the presence of the desired PCR product by gel electrophoresis and any product for which the C\(_t\) value exceeded the value of 31 was discarded. The experiments were performed in triplicate (for each transcript) and for each replicate a \(\Delta C_t\) value was first calculated by subtracting the C\(_t\) value of that given replicate from the average C\(_t\) value of the internal control (\(\beta\)-actin). Next, the fold difference was determined for each replicate by raising 2 to the \(\Delta C_t\) powers and then the average of those three fold differences were used for that particular transcript. The standard deviations were calculated from the fold differences of the replicates.

References


CHAPTER THREE:
HIGH-THROUGHPUT TARGETED REPEAT ELEMENT BISULFITE SEQUENCING (HT-TREBS)

Introduction

The vital role played by repeat elements in the genome of mammals including humans is very understandable from the fact that approximately half of it is comprised of repeat elements of different types (Schmid and Deininger, 1975; Batzer and Deininger, 2004). The bulk of these elements are the mobile elements known as retrotransposons and DNA transposons, which make up 42% and 2-3% of the human genome, respectively (Lander et al., 2001). Contrary to the old notion that these elements are ‘junk DNA’, they have recently been found to have vital role in mammalian evolution, vertebrate immune system, and genomic stability among other functions (Jurka et al., 2007). Nonetheless, the ability of retrotransposons and DNA transposons to transpose to new locations, and the fact that some possess their own promoters (long interspersed elements (LINEs) and long terminal repeats (LTRs)), or their own poly-A tails (short interspersed elements (SINEs)), means the cellular machineries have to repress these elements to maintain genomic integrity and other cellular processes. Among the different epigenetic mechanisms applied by the host cells for repression of these elements, DNA methylation is perhaps the most important and the most stable (Yoder et al., 1997) mechanism, and is usually set during early embryogenesis and gametogenesis (Lees-Murdock et al., 2008). Highly elevated levels of transcripts originating from unmethylated copies of a type of LTR retrotransposon known as Intracisternal A-Particles (IAPs) was observed in Dnmt1−/− mouse embryos (lacking the enzyme for maintenance of DNA methylation). However, on many different occasions, such epigenetic
repression has been found to be incomplete or inconsistent under diseased conditions such as in cancer (Howard et al., 2007; Baba et al., 2010) or even under normal conditions (Dolinoy et al., 2007).

The aberrant DNA methylation repression of repeat elements especially those of retrotransposons can have profound genetic and epigenetic effects. Decreased methylation in the LTR of a mouse endogenous retrovirus (ERV) in the intron of a gene has been responsible for increased levels of its truncated transcripts (Li et al., 2012) even at a distance. Opposing to the previous notion that LINE-1 (L-1) transposition is confined to germ line only, some recent studies have found evidence of somatic jumping of L-1 elements and have suggested hypomethylation of these elements as one of the causes (Coufal et al., 2009; Belancio et al., 2010; Muotri et al., 2010). One of the most intriguing epigenetic effects of incomplete DNA methylation mediated repression might be the presence of inter-individual variations imparted by retrotransposons in the absence of genetic variation. Two prime examples of such retrotransposons are the IAP insertions in the mouse endogenous genes agouti and axin causing the mutant mouse phenotypes, viable yellow agouti (Ay) and axin-fused kinky (Axinfused) respectively (Whitelaw and Whitelaw, 2006; Jirtle and Skinner, 2007). In both cases, the IAP LTRs bring about the mutant phenotype by causing ectopic transcriptions of the endogenous genes which would otherwise happen in a cell type specific and time or stage specific manner. Interestingly, even in the absence of any sequence variations, individuals harboring the IAPs can have very different DNA methylations in the LTRs and as a consequence variable phenotypic effect depending on the level of DNA methylation.
The huge implications of the DNA methylation status of repeat elements, and more importantly of retrotransposons, in their activity in the genome implies that large scale study of their DNA methylation status is very crucial. Due to the difference in nature between the repeat element classes, methylation study of individual class of repeats rather than a global study might also be needed. Also, since members of most of the repeat element classes are distributed almost throughout the entire genome, a mechanism of interrogating the methylation status of a repeat element class by individual locus is also warranted. The unavailability of an established scheme for large-scale cytosine methylation analysis of individual repeat element loci of specific type led us to the development of our protocol named High-throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS). This technique is capable of providing genome-wide, single-base resolution, and highly enriched DNA methylation data of any subset of repeat elements in an organism. Using this protocol, we investigated the methylation status of individual loci of five subtypes of the IAPLTR class of repeat elements. The data from the study shows that a minor fraction of the IAPLTR elements are hypomethylated and have the potential to act as ectopic promoters. The analyses from this study also reveals extensive divergence of DNA methylation repression for a large number of loci with intra-individual, inter-individual, as well as sex-based variations. These data suggest that such large scale, individual loci-based DNA methylation studies might be appropriate for other groups of repeat elements as well.

Results

High-throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS) for single-base resolution methylation analysis of individual locus of repeat elements
HT-TREBS is based on adaptations of the two widely used high throughput bisulfite sequencing techniques: reduced representation bisulfite sequencing (RRBS) (Meissner et al., 2008) and methylC-Seq (Lister et al., 2008), coupled with a PCR step in which one of the primers is specific to the repeat element class that is being investigated. This allows the DNA methylation analysis of only a subtype of repeat element to be done without the requirement of a vast number of reads. Thus the proposed technique avoids redundancy, but at the same time ensures enough coverage of the sequences that have been targeted. Using the HT-TREBS technique in this study, we analyzed the cytosine methylation status of five of the eight IAPLTR subtypes (IAPLTR1, IAPLTR1a, IAPLTR2, IAPLTR2a, and IAPLTR2b) in brain, liver, and kidney (one representative tissue from each of the three germ layers) of four 1-week-old littermates (two females and two males) of the isogenic background C57BL/6N. For this purpose, genomic DNA isolated from these twelve samples were separately sonicated to up to a predetermined certain length, end-repaired, and then ligated to custom Ion Torrent ‘A’ adaptors in which all the Cs have been methylated (Figure 3.1A; Methods). After a round of size-selection to remove any excess adaptors and undesirable short fragments, the adaptor-ligated DNAs were treated with the bisulfite conversion method (Clark et al., 1994). The bisulfite-converted DNA samples were then subjected to a round of PCR where the forward primer was chosen from the ‘A’ adaptor region and the reverse primer was chosen from a region that has been found to be well conserved among the IAPLTR subtypes mentioned earlier. The reverse primer also had the sequence of the Ion Torrent ‘P1’ adaptor ligated to its 5’ primer end. The amplification cycle number of the PCRs were kept to the minimum possible so as to produce just enough product for sequencing. After a round of agarose gel size-selection, the twelve PCR products were individually sequenced in Ion Personal Genome Machine (PGM) Sequencer using Ion 318 Chips. The
generated sequence reads were mapped to custom bisulfite-converted IAPLTR reference genome (mm9) by the aligner Bowtie 2 (Langmead and Salzberg, 2012) and then filtered by custom scripts to retain only those reads that mapped to IAPLTR loci and also had at least 10 bases of flanking unique sequences to ensure the reads were sequenced from unique positions of the genome (Methods). These filtered reads from each sample were then separately analyzed for their CpG methylation status using BiQ Analyzer HT (Lutsik et al., 2011).

**Figure 3.1.** Scheme and efficiency of the proposed High-throughput Repeat Element Bisulfite Sequencing (HT-TREBS) technique using mouse IAPLTR elements. (A) Schematic of the HT-TREBS depicting the steps involved up to the run in the next-generation sequencer. The blue bars and the green bars in steps I-IV represent the IAPLTR sequences and the adjoining unique sequences respectively. The orange bars in steps II-IV represent the methylated Ion Torrent ‘A’ adaptor. The orange arrow in step III depicts the forward primer corresponding to the ‘A’ adaptor region while the blue arrow depicts the reverse primer corresponding to the chosen IAPLTR site. The reverse primer also has the Ion Torrent ‘P1’ adaptor sequence attached to it as depicted by the purple bar. Step IV shows that the PCR products generated from a single specific IAPLTR locus might have varying lengths because of the random lengths of the unique sequences (green bars) caused by sonication. (B) Efficiency of the HT-TREBS. The genome graphs of the twelve sequenced samples as visualized using the UCSC Genome Browser website. (C) Specificity of the HT-TREBS. Graph showing the percentage of the total number of different IAPLTR subtypes present in the genome that have been sequenced in all twelve samples. The dashed red line shows the level of 80%. In all the sequenced samples, most of the targeted IAPLTR
subtypes have been sequenced for over 80% while the non-targeted subtypes have been sequenced for less than 15% of their total number in the genome.

In each sample, the methylation data were then filtered using custom scripts to retain only those IAPLTR loci that had at least 3 CpG positions for which the sequencing depth was at least 15X (at least 15 sequence reads were available for that CpG position). In a final filter, we discarded any IAPLTR loci that have not been sequenced in all twelve samples and built a representative set of 5135 IAPLTR loci. All further analyses in this study have been performed based on this representative set of IAPLTR loci. The efficiency and the robustness of the HT-TREBS was evident after the analyses of the sequenced reads and the CpG methylation analyses of the twelve samples even before applying the previous mentioned filters (Figure 3.1B,C; Table 3.1).

Table 3.1. Coverage and sequencing efficiency of the twelve samples sequenced by the HTREBS technique

<table>
<thead>
<tr>
<th>Sequenced sample</th>
<th>Number of mapped reads with IAPLTR sequences and unique sequences *</th>
<th>Number of IAPLTR loci retained after filtering</th>
<th>Total number of CpG positions sequenced at depth of &gt;15x</th>
<th>Average number of CpG positions sequenced per IAPLTR loci</th>
<th>Average coverage of the CpG positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female #1 Brain</td>
<td>984,969</td>
<td>6,520</td>
<td>49,720</td>
<td>7</td>
<td>118x</td>
</tr>
<tr>
<td>Female #1 Liver</td>
<td>794,884</td>
<td>6,494</td>
<td>49,746</td>
<td>7</td>
<td>91x</td>
</tr>
<tr>
<td>Female #1 Kidney</td>
<td>897,810</td>
<td>6,965</td>
<td>53,106</td>
<td>7</td>
<td>98x</td>
</tr>
<tr>
<td>Female #2 Brain</td>
<td>2,218,127</td>
<td>7,046</td>
<td>54,758</td>
<td>7</td>
<td>239x</td>
</tr>
<tr>
<td>Female #2 Liver</td>
<td>719,895</td>
<td>6,468</td>
<td>49,816</td>
<td>7</td>
<td>89x</td>
</tr>
<tr>
<td>Female #2 Kidney</td>
<td>931,659</td>
<td>7,106</td>
<td>55,186</td>
<td>7</td>
<td>100x</td>
</tr>
</tbody>
</table>
Table 3.1. continued

<table>
<thead>
<tr>
<th>Sequenced sample</th>
<th>Number of mapped reads with IAPLTR sequences and unique sequences *</th>
<th>Number of IAPLTR loci retained after filtering</th>
<th>Total number of CpG positions sequenced at depth of &gt;15x</th>
<th>Average number of CpG positions sequenced per IAPLTR loci</th>
<th>Average coverage of the CpG positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male #1 Brain</td>
<td>720,344</td>
<td>6,928</td>
<td>53,718</td>
<td>7</td>
<td>82x</td>
</tr>
<tr>
<td>Male #1 Liver</td>
<td>968,858</td>
<td>6,752</td>
<td>52,549</td>
<td>7</td>
<td>114x</td>
</tr>
<tr>
<td>Male #1 Kidney</td>
<td>1,107,989</td>
<td>7,228</td>
<td>56,262</td>
<td>7</td>
<td>118x</td>
</tr>
<tr>
<td>Male #2 Brain</td>
<td>1,641,525</td>
<td>6,961</td>
<td>53,251</td>
<td>7</td>
<td>170x</td>
</tr>
<tr>
<td>Male #2 Liver</td>
<td>1,107,765</td>
<td>7,100</td>
<td>53,933</td>
<td>7</td>
<td>111x</td>
</tr>
<tr>
<td>Male #2 Kidney</td>
<td>656,240</td>
<td>6,331</td>
<td>42,782</td>
<td>6</td>
<td>67x</td>
</tr>
</tbody>
</table>

* Number of reads with IAPLTR sequences that mapped to the reference genome and had at least 10 bases of unique sequences

The distribution of the IAPLTR loci that have been sequenced in all the samples is effectively over the entire lengths of all the chromosomes, without any bias for any particular region of the chromosomes (as shown in the genome graphs of Figure 3.1B). The percentages of the total number (in the genome) of the five targeted IAPLTR subtypes that have been sequenced are more than 80% in all but 5 cases (only 5 IAPLTR subtypes in 4 samples) and not less than 65% in any case. The specificity of the technique is also depicted by the fact that the other three non-targeted IAPLTR subtypes (IAPLTR3, IAPLTR4, and IAPLTR4_I), even though sequenced, were actually represented by a much smaller fraction of their total number in the genome (less than 15% in all the cases). This proves that our proposed HT-TREBS can be a very effective
scheme for a highly targeted high-throughput bisulfite sequencing for any class of repeat elements.

Canonical DNA methylation patterns of the IAPLTR retrotransposons in individual samples
IAPLTR loci, like most of the repeat elements, have been known to have an overall fully methylated state in most cases but also have been known to escape the repressive CpG methylation on many occasions. We tried to understand whether such escapes or variations in CpG methylation are CpG position based or cell based (one group of cell having full methylation while the other having none) or a combination of both. For that, we analyzed the CpG methylation status of each of the IAPLTR loci in our representative set both in terms of their individual CpG position and also in terms of the individual reads or sequences that were generated (http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Fig.pdf). For any given sequenced sample, when we plotted the overall CpG methylation percentage of a particular locus against the standard deviation of CpG methylation percentage of individual reads of that locus, a distinct pattern was evident. Such a similar pattern was also apparent when the overall CpG methylation percentages of those loci were drawn against their respective standard deviations of CpG methylation percentage of individual CpG positions. Due to the visual similarities of such plots to sprinklers, we would like refer to these plots as ‘sprinkler plots’ from hereon. As evident from both the read-based (Figure 3.2A) and the CpG position-based sprinkler plots (Figure 3.2B) of the Female #1 brain sample and also from similar plots of the other samples (http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Fig.pdf), majority of the IAPLTR loci are positioned within the 80-100% methylation point which is expected from many previous observations.
Figure 3.2. Depiction of the canonical methylation pattern of the IAPLTRs by ‘sprinkler plot’. 
(A) The two-dimensional sprinkler plot for the Female #1 brain sample showing the relation of the overall CpG methylation percentage of each sequenced IALTR locus to the standard deviation of CpG methylation percentage of individual reads of that locus. Likewise, in (B) the relations of the overall CpG methylation percentages of the loci to their respective standard deviations of CpG methylation percentage of individual CpG positions (only the ones sequenced) is shown. (C) The three-dimensional sprinkler plot of the same sample showing the relations of the overall CpG methylation percentages of the sequenced IALTR loci to both their standard deviations of CpG methylation percentage of individual reads and standard deviations of CpG methylation percentage of individual CpG positions. In all these sprinkler plots (two- and three-dimensional), each dot represents a single IAPLTR locus. Four representative patterns of CpG methylations of the loci have been shown in bubble charts by taking sequencing data from the mentioned sample: (D) mostly unmethylated, (E) partially methylated with all or none methylation among the reads, (F) partially methylated with all or none methylation among the CpG positions, and (G) mostly methylated. In these bubble charts, each row represents a different read and each column represents a different CpG position. Filled and open circles
indicate methylated and unmethylated cytosines, respectively, whereas the blue circles indicate CpG positions which did not have any or sufficient sequence information. The CpG positions are arranged in the 5’ to 3’ direction of the sequenced region while going from left to right in each read. The percentages at the bottom of each bubble chart shows their respective overall CpG methylation percentage (M), standard deviation of CpG methylation percentage of individual reads (R), and standard deviation of CpG methylation percentage of individual CpG positions (C). The approximate positions of these representative loci have been indicated in (A) and (B) by the red, blue, orange, and green arrows for (D), (E), (F), and (G) respectively.

Even though both the read-based and the CpG position-based sprinkler plots have a similar pattern, the CpG position-based plots look more spread out than the read-based plots even near the 100% methylation point. This might suggest that in majority of the IAPLTR loci variation in methylations might be more in individual CpG positions than in the reads. It is also worth noting that quite a few of the IAPLTR loci are located near the 0-50% methylation with wide range of standard deviations in both the sprinkler plots. A merged three-dimensional view (Figure 3.2C) of the previously mentioned sprinkler plots shows a similar pattern but at the same time it also shows that in the partially methylated region (~20-80% methylation), a particular IAPLTR loci can have close or very different standard deviations (read-based and CpG position-based). As shown in Figures 3.2D,E,F, and G, the sprinkler plots can help identify four distinct situations of the IAPLTR loci methylation status based on their locations in the plots: i) mostly unmethylated, ii) partial methylation with individual reads having full or no methylation but individual CpG positions having the same overall methylation level, iii) partial methylation with individual CpG positions having full or no methylation but individual reads having the same overall methylation level, and iv) mostly methylated. Overall, these plots once again showed the effectiveness of the HT-TREBS scheme as the majority of the IAPLTR loci are found to be mostly methylated (>80%) but some loci have varying degrees of methylation as was suggested by some previous studies.
Analyses of the hypomethylated IAPLTR retrotransposons

In a genome-scale DNA methylation study of retrotransposons one definite interest would be an analysis of the loci that are found to be hypomethylated, but under the normal circumstances, would be expected to be fully methylated. From that perspective, for each IAPLTR locus in the representative set, we calculated the average CpG methylation percentage of that locus in all twelve sequenced samples and then binned them in 10% CpG methylation windows starting from 0% to 100% (Figure 3.3A). As evident from that analysis, 4019 (78%), 995 (18%), and 121 (2%) of the loci have methylations in the ranges 90-100%, 80-90%, and less than 80% respectively. Considering less than 80% methylation as hypomethylation, we can find that only a fraction (2%) of the IAPLTR loci are probably constitutively active and this is consistent with the suggestions in many other previous findings.

In an effort to examine how the hypomethylation status of the 121 loci mentioned previously correlate with different genomic and epigenomic features we performed a set of analysis on these hypomethylated loci based on their genomic positions, their subtypes (sequence similarities), and their proximity to different histone modification enrichment regions. Firstly, we analyzed the percentages of the hypomethylated (below 80%) and the mostly methylated (above 80%) IAPLTR loci that are within 10kb, 100kb, and 500kb upstream and downstream of the transcription start sites (TSS) of known genes (Figure 3.3B; Methods). Interestingly, the CpG methylation of the IAPLTR loci appears to be largely unaffected by the proximity to promoters of genes since at all three distances the percentages between the hypomethylated and the mostly methylated groups are almost comparable.
Figure 3.3. Analyses of the hypomethylated IAPLTRs. (A) A breakdown of the IAPLTRs in the 5135 representative set based on their average CpG methylation percentage in the twelve samples. The numbers on top of each bar indicates the count of loci in that category. (B) The percentage of the hypomethylated (<80%) and mostly methylated (>80%) loci that are within 10kb, 100kb, and 500kb of the promoters of known genes. The orange bars and the blue bars indicate hypomethylated and mostly methylated loci respectively. (C) Percentage representation of the different subtypes of IAPLTR among the hypomethylated, mostly methylated, and combined (5135 representative set) categories. Once again the orange bars and the blue bars indicate hypomethylated and mostly methylated loci respectively while the green bars represent the combined loci. (D) Genome location of the IAPLTR2 locus at chr6:37,414,145-37,414,607. The red, orange, blue, and green arrows indicate the IAPLTR2 locus, the L1 fragment, the annotated mRNA AK040515, and the unspliced EST BB211325 respectively. The orientation of both the IAPLTR and the L1 fragments is in the positive strand.
Secondly, we studied the composition of the five IAPLTR subtypes we targeted in this study among the hypomethylated, mostly methylated, and the representative set (all) of IAPLTR loci (Figure 3.3C). Even though the mostly methylated group is mirroring the percentages in the representative set, the hypomethylated group has an abundance of IAPLTR2 (51%) and IAPLTR2a (39%) subtype compared to the methylated group. Some other previous studies regarding the age of endogenous retroviruses (ERVs) and their DNA methylation which suggested that the old ERVs are comparatively more methylated than the young ones (Reiss et al., 2010; Faulk et al., 2013). Hence our finding might be contrary to these previous studies since the IAPLTR2 and IAPLTR2a subtypes are comparatively older than the IAPLTR1 and IAPLTR1a subtypes.

Thirdly, we tried to analyze the possible role of histone modifications on the methylation values of the hypomethylated loci. For this purpose, for the three different tissues we studied (brain, liver, and kidney) we intersected these IAPLTR loci with the ChIP-Seq enrichment regions (peaks) of a broad range of histone marks obtained from those tissues: i) histone 3 lysine 4 trimethylation (H3K4me3) – marks active promoters, ii) histone 3 lysine 4 monomethylation (H3K4me1) – associated with enhancers and downstream of transcription starts, iii) histone 3 lysine 36 trimethylation (H3K36me3) – marks regions of RNA Pol II elongation, iv) histone 3 lysine 27 trimethylation (H3K27me3) – marks promoters of lineage-specific genes that are silenced by Polycomb proteins, and lastly v) histone 3 lysine 27 acetylation (H3K27ac) – associated with transcriptional initiation and open chromatin structure. The ChIP-Seq data for these histone modifications were obtained from the Mouse Encode Project experiments for whole brain, liver, and kidney (Methods). Even though the intersection reveals that only 11 loci
overlaps with at least one ChIP-Seq peak in at least one tissue ([http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Table.pdf](http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Table.pdf)), one of the loci (IAPLTR2 - chr6:37414145-37414607) has been found to be associated with the active promoter mark H3K4me3 in all three tissues and with H3K4me1 mark in liver and kidney. Interestingly this particular loci is also the one with the least average methylation percentage (6.8%) in the entire representative set. A close look at the nearby genomic features of this loci reveals the presence of an annotated mRNA (AK040515) and unspliced EST (BB590286) just upstream of it suggesting that a possible transcriptional activity of the loci ([Figure 3.3D](http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Fig.pdf)). Even though an L1 fragment is also present near adjacent to the IAPLTR, the transcripts are less likely to be due to its promoter activity due to the longer distance of the TSS of the promoters from its 5’ end from than from the 5’ end of the IAPLTR. Moreover, the subtype of this particular locus is IAPLTR2, which has already been shown to be the dominant subtype for the hypomethylated IAPLTRs. Overall, our data indicate that the hypomethylated IAPLTR loci might be more influenced by their sequence (subtype) rather than their genomic positions and the presence of their potential transcriptional activity might be influenced by the presence active histone modification marks.

**Intra-individual, inter-individual and sex-based DNA methylation variations of the IAPLTR retrotransposons**

We tried to judge the overall variation that might be present in the twelve samples that were sequenced by constructing two clustered heatmaps: one using the overall CpG methylation values of the representative set of 5135 IAPLTR loci in the twelve samples ([http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Fig.pdf](http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Fig.pdf)) and the other using the overall CpG methylation values of the loci among that representative set that had a methylation value of less
than 80% in at least one of the samples (Figure 3.4A). The distance and the clustering of all but one brain sample relative to the other tissue samples in the dendrogram in both the heatmaps hint that a lot of IAPLTR loci might have quite different CpG methylation status in brain than in other tissues.

**Figure 3.4.** DNA methylation variations of the IAPLTR retrotransposons. (A) Heatmap of the 12 sequenced samples showing the difference in CpG methylation of a subset of the representative IAPLTR loci that had a less than 80% CpG methylation in at least one of the twelve samples. The dendrogram on top shows the clustering of the samples while that on the left shows the clustering of individual loci based on their CpG methylation difference. The color key on the top left depicts the colors representing each value of CpG methylation while the histogram in it shows the number of loci present in the heatmap at those respective methylation values. (B) Intra-individual variation (between tissues of the same individual) of CpG methylation. Matrix showing the numbers of IAPLTR loci that have been found to be statistically significantly different (Kruskal-Wallis test; p < 0.001) in their CpG methylation percentages between brain, liver, and kidney in different combinations in the four individuals (Female #1-2, Male #1-2) that have been sequenced. The numbers on the far right of each row indicate the total number of IAPLTR loci that have been found to be varying in that particular individual while the percentages in italics in each cell indicate the number of loci in the
respective cells as percentage to the total number of loci for that individual. (C) Four-way Venn diagram showing the number of overlapping and non-overlapping IAPLTR loci of the four individuals that have been found to be varying intra-individually. (D) Inter-individual variation (between individuals in the same tissue) of CpG methylation. Three-way Venn diagram showing the number of overlapping and non-overlapping IAPLTR loci of the three tissues that have been found to be varying inter-individually (Kruskal-Wallis test; p < 0.001). (E) Intersection between IAPLTR loci varying intra-individually and those varying inter-individually. (F) Number of loci of the four categories of CpG methylation variation as a percentage of the total number of loci in the representative IAPLTR set. (G) Number of loci in the three different tissues that have been found to be varying between individuals of opposite genders but not between individuals of the same gender.

One of the prime objectives of developing a scheme like HT-TREBS was to assess the methylation status of individual IAPLTR locus so that we can identify loci with varying degrees of DNA methylation. Also due to the significance of CpG methylation variation in just a few CpG positions in conditions like predisposition to diseases (Rakyan et al. 2011), inheritance (Kile et al. 2010) and others, we included all the IAPLTR loci in the variation analyses because a bulk of those loci have DNA methylation greater than 80% but are not fully methylated. To this end we analyzed the CpG methylation variations of individual IAPLTR loci in three different scenarios: i) intra-individual variation (variations among the tissues of the same individual), ii) inter-individual variation (variations among the individuals in the same tissue), and iii) sex-based variation (variations among individuals of the opposite sexes but not among the individuals of the same sex). In each case, the variations were statistically determined (Kruskal-Wallis test; p<0.001; Methods) based the on the CpG methylation percentages of individual IAPLTR loci only in the samples involved for that specific variation. For the intra-individual variations, a total of 2,357 IAPLTR loci have been identified that have varying CpG methylation in at least one individual (Figure 3.4B,C) with the number of loci varying in any particular individual ranging from 587 to 1,971. At the same time no particular pattern among the sexes was evident.
A breakdown of the numbers of loci that are shared between the tissues (brain, liver, and kidney) shows that, in all the individuals except Male #1, even though the majority of loci are simultaneously varying in all three tissues (41-73%), the number of loci varying in only liver and kidney is the least among all the two-tissue combinations. This is again is probably a hint that IAPLTR loci are having much different CpG methylation status in brain than in other tissues.

For the inter-individual variations, a total of 2,325 IAPLTR loci have been identified that have varying CpG methylation in at least one of the tissues among the four individuals (Figure 3.4D). Even in this case, brain showed much difference compared to the other tissues by having 2,181 loci that are varying among the individuals while liver and kidney having 1,093 and 955 loci respectively. An intersection of the loci that are varying intra-individually and inter-individually, showed that in total 2,521 loci out of the 5,135 representative loci are varying in some way or the other while the remaining 2,614 loci or ~51% of the total are apparently not varying (Figure 3.4E,F). Such intersection showed that not only a large proportion of the IAPLTR are varying in their CpG methylation but also a large number are varying both intra- and inter-individually.

This also helped us categories the set of representative loci into four groups in terms of CpG methylation variability: i) Tissue-only variation loci (loci varying intra-individually only), ii) Epialleles (loci varying inter-individually only) (Rakyan et al., 2002), iii) Stochastic loci (loci varying both intra- and inter-individually), and iv) Non-variant loci. A list of all the loci in these four CpG methylation variation groups along with the other loci that have been analyzed in this study have been provided in http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Material.pdf.

An analysis among the varying groups of IAPLTR loci also helped us identify 26 loci that have been found to be varying in their CpG methylation between the individuals of the opposite sex.
and not between individuals of the same sex (Figure 3.4G). Such a find indicates the presence of a yet unknown sex-based methylation variations of any repeat element class. It is also worth noting that all of these loci have been found to be varying (sex-based) exclusively in one tissue. Once again brain was different from liver and kidney by having much more number of loci (14 loci in brain as opposed to 6 each in liver and kidney).

Similar to the hypomethylated IAPLTR loci, we performed a detail analyses of the genomic position preference, sequence preference, and histone modification overlap of the loci in the different variation categories (data not shown) but no significant preference have been observed for any of those. Overall, the results from our method so far indicate the presence of a large number of IAPLTR loci that have variations in their CpG methylation in different contexts and also helped us identify those individually.

**Discussion**

In this study we developed and verified the efficiency of a high throughput technique of targeted bisulfite sequencing known as High-throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS), which is able to provide genome-scale, single-base resolution, cytosine methylation data of any repeat element class down to an individual locus. Using this technique in this study we obtained genome-wide CpG methylation data of five of the subtypes of IAPLTR in the brain, liver, and kidney of four 1-week-old C57BL/6N littermates (two females and two males). The data acquired indicate the presence of a small subset of these repeat elements that are significantly hypomethylated (average CpG methylation is less than 80% across the twelve samples). A large proportion of these hypomethylated loci also tends to be from a specific
subtype, thereby, indicating the possible role of sequence and age in the methylation status of IAPLTR. At the same time, analyses of methylation data of all the sequenced loci exhibit extensive CpG methylation variations among the IAPLTR subtypes in the form of inter-individual, intra-individual, and sex-based variations. These findings apart from providing some novel insights proves, for the first time on a genomic scale, some of the presumed nature of IAPLTRs such as hypomethylation and methylation variability (Mietz et al., 1990; Morgan et al., 1999), and thereby, validates the results of the proposed HT-TREBS.

The unique aspect of HT-TREBS is the targeted nature it achieves through the amplification step involving primers specific to only certain portions of the genome. As a result, HT-TREBS is ideal for the bisulfite sequencing of subset of similar repeat elements because of their abundance and sequence similarity. A careful selection of the primer for HT-TREBS can ensure it amplifies and sequences the targeted repeat elements from a large number of loci instead of just few and thereby avoid redundancy. Such specificity and non-redundancy is well indicated from the results of our twelve samples where, by sequencing only ~5-6 million reads per run, we have been able to sequence almost 80% of our targeted regions with an average of approximately 52,000 CpG positions in each sample and at an average sequencing depth of 117x per CpG positions. These numbers also suggests that HT-REBS can used to sequence only one sample in sequencers with relatively small throughput like Ion PGM or by means of multiplexing can be used to sequence many samples using ultra-high-throughput sequencers like the Illumina HiSeq systems. Its independence from the use of restriction enzymes like MspI or BglII as used in techniques like reduced representation bisulfite sequencing (RRBS), means that its library can be made for even CpG-poor regions of the genome. Such feature of HT-TREBS might be ideal for
repeat elements like LINE elements which are mostly found in A+T rich regions even though they have a lot CpGs of their own (Lander et al., 2001). The requirement of having unique sequences (non-targeted sequences) in the HT-TREB reads is critical for the success of being able to provide individual loci-based methylation information for repeat elements. At the same time such requirement makes HT-TREBS inefficient with smaller read lengths (<150 bases) since those reads will not have enough length of non-targeted sequences and enough length of targeted sequences with sufficient number of CpGs in it. Fortunately, with the recent advancements in high-throughput sequencing technologies longer read lengths are becoming very commonplace in most platforms.

In this study we have been able to identify the presence of a minor fraction (121 or ~2% of the total number) of the IAPLTR loci to have hypomethylated status. Also the fact that these loci have been identified as individual locus gives immense weight to such finding. Careful study of these elements, locus by locus, could lead to important revelations about their ectopic or regular transcriptional activity, their possible role in the evolution of nearby genes or may be their possible roles as enhancers. Some of these can even be investigated as potential biomarkers with role in predisposition to disease or environmental exposures. Contrary to some previous suggestions that older ERVs are more likely to be fully methylated (Reiss et al., 2010; Faulk et al., 2013), these hypomethylated loci are predominantly of the relatively older subtypes IAPLTR2 and IAPLTR2a. One possible hypothesis behind such finding could be that since these elements are relatively older and many of them are present as solo LTRs, they are not under the same repressive force like the younger IAPLTRs and yet they have not aged enough to lose their promoter activity.
The vast number of IAPLTR loci that have been found to have DNA methylation variability in this study gives an insight into the often speculated diverse epigenome of retrotransposons as opposed to the previous notion that they are mostly silenced by DNA methylation. In our study approximately 49% of the representative IAPLTR loci (2521 out of 5135) have been found to be variable in either one or more of the variability types suggested. The loci that are varying only inter-individually might be good candidates for being metastable epialleles (Rakyan et al., 2002) and the number of those loci found in this study (164) is in close agreement to that suggested in one of our previous studies (Ekram et al., 2012). A detailed maternal supplement (Waterland et al. 2006) and transgenerational methylation study (Rakyan et al., 2003) on these loci might be necessary to confirm their viability to be epialleles. The fact that most of the variable loci are stochastic (both intra- and inter-individually variable) might suggest that their variability is influenced by a number of different genetic and epigenetic factors and might be different from loci to loci. In our study we also found that the loci in brain have more variability than in the other tissues. This might be reminiscent of the fact that brain consists of a more diverse cell type population than other tissues. It could also be suggestive that IAPLTRs in mice behave in a similar fashion like the L1s in humans, which have been known to cause somatic mosaicism in brains by means of retrotransposition (Coufal et al., 2009; Muotri et al., 2010). Our study also found a group of 26 IAPLTR loci which are variable only between the genders. Such a finding for repeat elements might possibly be totally novel and call for more study to find out the possible mechanisms behind such variation.

Overall, our study through the use of the proposed HT-TREBS technique have been able to prove that a fraction of the IAPLTR retrotransposons are certainly hypomethylated and yet a larger
fraction of them might be variable in terms of their DNA methylation. Due to the large number of retrotransposons in the mammalian genome, finding such a diverse epigenomic state in just a subset of retrotransposons might inspire researchers to perform similar studies in other types of retrotransposons.

Methods

Library construction for High-throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS)

Genomic DNAs have been isolated and purified from the brain, liver, and kidney of four 1-week-old littermates (two females and two males) of the isogenic C57BL/6N background mice obtained from the Jackson Laboratory (http://www.jax.org/). Due to the young age of the littermates, the sex of each individual was determined by genotyping of the Sry gene. All experiments were performed in accordance with the National Institutes of Health guidelines for care and use of animals. For each of the twelve samples, 1µg of the purified genomic DNA was sonicated to fragments that peaked at 700bp (Bioruptor NGS, Diagenode). Since the average lengths of the IAPLTRs are 300-350bp and the ultimate desired sequences are expected to contain flanking non-IAPLTR sequences as well, the peak size of the fragments are empirically determined to be at 700bp. The fragmented DNA was immediately end-repaired using the NEBNext End Repair Module (New England BioLabs) and then ligated using T4 DNA ligase (New England BioLabs) to custom made double-stranded Ion Torrent ‘A’ adaptors in which all the Cs have been methylated (Integrated DNA Technologies). The choice of using the ‘A’ adaptor over the ‘P1’ adaptor as the methylated adaptor was made because trial sequencing runs with ‘A’ adaptor yielded much higher percentage of mappable reads with unique sequences than
those using the ‘P1’ adaptor (http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Table.pdf; http://jookimlab.lsu.edu/Diss-HT-TREBS-Supp-Fig.pdf). The adaptor-ligated DNA fragments were then size-selected to get rid of any excess adaptors as well as any fragment smaller than 300bp using the Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer’s protocol. The size-selected samples were then quantified using the Agilent 2100 Bioanalyzer for verification of size and quantity. Fragments smaller 300bp were removed since the ultimate sequence reads were expected to be approximately 250-300bp. The adaptor-ligated DNA library was then modified using the bisulfite conversion reaction according to the manufacturer’s protocol (EZ DNA methylation kit, Zymo Research). The bisulfite-converted ‘A’ adaptor-ligated library was then used as template for a round of PCR (Maxime PCR Premix Kit, Intron Biotech) in which the forward primer (CCATCTCATCCCTGCGTGTCTCCGACTCAG) was designed in the 5’ end of the ‘A’ adaptor region of the library and the reverse primer was designed in a region of the bisulfite-converted genome which has been found to be well conserved among the consensus sequences of the IAPLTR subtypes that have been analyzed in this study (IAPLTR1, IAPLTR1a, IAPLTR2, IAPLTR2a, and IAPLTR2b). Attached to its 5’ primer end, the reverse primer (CCACTACGCCTCCGCTTTTCCCTCTATGGGCGATCGGTGAT^CTCCCTAAATACTACA ACCCATC) also had the sequence of the regular Ion Torrent ‘P1’ adaptor. The break in the previous sequence shows the joining point of the IAPLTR-specific sequence and the P1 adaptor sequence. The amplification cycle number of the PCR was individually determined for each of the twelve samples so that the minimum possible cycles were used to produce just enough product for sequencing. The PCR product was then size selected for range of 250-300bp using agarose gel electrophoresis. The size-selected PCR product was then verified for its length and
quantity using the Agilent 2100 Bioanalyzer. Each of the twelve PCR products was then individually sequenced in the Ion Personal Genome Machine (PGM) Sequencer using Ion 318 Chips (Ion Torrent, Life Technologies). The amplification of the library with the primers that had the Ion Torrent adaptor sequences at the ends as mentioned earlier ensured the amplification products were directly compatible for use in the Ion PGM sequencer.

Mapping of the sequences and other computational analyses

The sequence reads generated from the twelve Ion PGM runs were individually mapped using the aligner Bowtie 2 (Langmead and Salzberg, 2012) to a curated reference genome made up of bisulfite-converted top (Original Top: OT) and bottom (Original Bottom: OB) strands of the IAPLTR sequences (along with the 350bp flanking sequences) and the complementary sequences to those strands (Complementary to Original Top: CTOT; Complementary to Original Bottom: CTOB). The original IAPLTR and their flanking sequences were obtained from the mm9 version of the mouse genome. The mapped reads were then filtered using custom Python scripts to extract only the sequences that mapped to the IAPLTR regions and also had at least 10 bases of the flanking non-IAPLTR sequences to ensure the reads were generated from a specific IAPLTR locus rather than multiple loci. The filtered reads from each sample were then separately analyzed using the BiQ Analyzer HT tool (Lutsik et al., 2011) that provided, for each IAPLTR locus in a sample, CpG methylation status, pattern of CpG methylation in individual reads and other relevant information regarding the quality of the reads. These methylation data were then put through the following filter using custom scripts: i) discard any IAPLTR loci that did not have at least 3 CpG positions for which the sequencing depth was at least 15X (at least 15 sequence reads were available for that CpG position), ii) discard any IAPLTR loci that have
not been sequenced in all twelve samples. Application of all these filters helped us build the representative set of 5135 IAPLTR loci. The CpG methylation values of each locus in all the samples were calculated using only the methylation data of the CpG positions that are only within the IAPLTR locus itself and not in the flanking sequences.

**Intersection with datasets and preparation of ChIP-Seq data**

The distance of the IAPLTR loci from the transcription start sites (TSSs) of promoters were calculated using the UCSC Known Gene (mm9) dataset downloaded from the Table Browser of the UCSC Genome Bioinformatics website (http://genome.ucsc.edu/cgibin/hgTables?org=Mouse&db=mm9). The histone modification ChIP-Seq data were downloaded from the Mouse Encode Project section of the UCSC Genome Bioinformatics website. All the ChIP-Seq data that were downloaded were produced at the Ludwig Institute for Cancer Research (LICR) under the Mouse ENCODE project. The dataset that were used are H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3 of whole brain, liver, and kidney. In all cases, the embryonic day 14.5 dataset was used whenever available rather than the adult dataset in order to maintain maximum consistency with the methylation data that was produced using 1-week-old mice. For each ChIP-Seq data, raw sequence files for both replicates along with the accompanying input data were obtained. The replicate sequence files were then individually mapped to the mouse mm9 genome (with repeats) using the Bowtie alignment tool (Langmead et al., 2009). Areas of enrichment (ChIP-Seq peaks) were then found in each replicates using the MACS 1.4 tool (Zhang et al., 2008). The ChIP-Seq peak files from the replicates were then combined and merged based on their genomic coordinates. The intersection between the ChIP-Seq peaks and the IAPLTR loci were done using
custom script which ignored any ChIP-Seq peak that was completely within the coordinates of an IAPLTR locus.

**Statistical analyses for measurement of CpG methylation variations**

The methylation variations among the representative IAPLTR loci were calculated using the following sample combinations: i) intra-individual variation (separately in each of the four individuals using the respective brain, liver, and kidney sample methylation data), and ii) inter-individual variation (separately for each of the three tissues using the methylation data of the four individual in the respective tissues). In each of the variations, for each specific loci, the CpG methylation values of the reads of the relevant samples were compared with each other to find variations by using the non-parametric Kruskal-Wallis test (p<0.001). Each positive test of variation was followed by a post hoc Mann-Whitney test with Bonferroni corrections to isolate the pair of samples that actually gave rise to the variation. The sex-based variations was found by considering the inter-individual variations in each tissue and isolating the loci that had statistically significant variations among individuals of the opposite sex but not between individuals of the same sex.

**References**


CHAPTER FOUR:
SUMMARY

Retrotransposons have been found in all mammalian genomes that have been analyzed or sequenced (Boeke and Stoye, 1997). Also, their importance in these genomes might be reflected by the fact that they form the bulk of these genomes: approximately 42% and 37% of the human and mouse genome respectively (International Human Genome Sequencing Consortium, 2001; International Mouse Genome Sequencing Consortium, 2002). Yet, they are known to exhibit a rather ‘parasitic’ nature by being able to move or copy themselves to new positions in the genome. They also carry their own promoters, driven either by RNA polymerase (pol) II or for RNA polymerase (pol) III, which can be potentially disruptive to regular gene transcription when present near to those regular genes. These features of retrotransposons have led to most of them becoming permanently epigenetically repressed by DNA methylation and histone modifications, mainly histone H3 lysine 9 tri-methylation (H3K9me3) for LINEs as well as LTRs and histone H3 lysine 27 tri-methylation (H3K27me3) for SINEs (Huda et al., 2010). Interestingly, previous data have revealed, such DNA methylation is far from complete in many retrotransposon loci. Also, retrotransposon loci have been found in areas of the genome that are enriched in histone modifications that are not necessarily repressive and sometimes are even activating. More interestingly, such inconsistencies in repression have often been observed with variations among individuals as is the case in the mutant phenotypes caused by the mouse metastable epiallele IAP elements in the genes agouti and axin (Whitelaw and Whitelaw 2006; Jirtle and Skinner 2007). Such findings clearly indicate that retrotransposons can be a major source of epigenetic variations in the mammalian genome. Unfortunately though, the search for the factors and the mechanisms underlying those epigenetics variations is far from complete. Also the lack of the
proper techniques for genome-scale DNA methylation analysis of repeat elements has been a stumbling block on the road to that search.

In this dissertation, I studied some of the epigenetic and genetic marks that might define the nature of metastable epialleles in the mouse genome and then developed and validated a novel high-throughput technique of repeat element bisulfite sequencing known as High-throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS). Search for the potential metastable epialleles in the mouse genome yielded 143 candidates among the vast number of LTR and LINE elements. Such number was very close to the one suggested by the study using HT-TREBS. The study also indicated the presence of a minor fraction of hypomethylated IAPLTRs as well as large scale methylation variations among the IAPLTRs intra-individually, inter-individually, and between the two genders.

In chapter two, we searched for the existence of additional retrotransposon-derived epialleles in the mouse genome. Using a series of bioinformatic approaches, 143 candidate epialleles were first identified from the mouse genome based on their promoter activity (proximity to the annotated mRNAs) and their association with the active histone modification mark H3K4me3 in a set of four tissues and cell lines. Detailed DNA methylation analyses on a subset of these elements using COBRA and regular bisulfite sequencing were performed from tail DNA of five 1-week-old mice littermates, in two representative tissues of each germ layer in a 2-month-old and 5-old-month-old mouse, in the germ cells and in blastocyst-stage embryos. Such analyses revealed that some of these elements showed variable levels of DNA methylation among the individual mice of the isogenic background C57BL/6J, revealing a stochastic nature
The dynamic nature of their DNA methylation patterns was confirmed since the analyses also identified two opposite patterns of DNA methylation, progressive gaining versus losing, during development. The expression levels of these elements were also found to be variable among the individual mice when analyzed by qRT-PCR. This also suggested probable functional consequences of the ectopic transcripts of these elements on their associated endogenous genes. All these data confirm the presence of a number of new retrotransposon-derived epialleles, and suggests the presence of more.

From this study we tried to find the possible mechanistic factors by which a subset of retrotransposons escapes DNA methylation and sometimes also in a variable manner among different individuals. However, given the large number of potential epialleles identified in this study, we surmise that these unusual epigenetic features may be closely associated with the three factors described below. Firstly, given that about 66% of the candidate epialleles are located in close proximity of the promoter of endogenous genes, the location of a given retrotransposon appears to be important. Secondly, the presence of conflicting histone modification signals is expected for the candidate epialleles in the promoter regions. Even though these elements are supposed to have the repressive H3K9me3 signal, they might also contain activating marks like H3K4me3 that are typical in promoter regions. Such opposing forces might explain the variable DNA methylation levels observed in the epialleles. Thirdly, the variable levels of DNA methylation of the metastable epialleles may also be influenced by the transcriptional properties of their associated endogenous genes. A lot of the candidate metastable epialleles are located near tissue- and stage-specific genes. As a result, the dynamic transcriptional programs in these genes during development might provide the associated retrotransposons more opportunities to
accumulate DNA methylation variations than the other retrotransposons associated to genes with static transcriptional programs. Overall, the nearness to promoters, the presence of conflicting histone marks, and the transcriptional pattern of the associated endogenous genes are assumed to be major contributing factors to the metastable nature of DNA methylation associated with epialleles.

In chapter three, we proposed and validated the results of a next generation bisulfite sequencing technique for repeat elements, which we termed as High-Throughput Targeted Repeat Element Bisulfite Sequencing (HT-TREBS). This technique can be used to interrogate the cytosine methylation status of individual loci of any group of repeats in a precisely targeted manner. Using HT-TREBS in this study, we obtained the CpG methylation data of 5135 loci or elements of five of the IAPLTR subtypes (IAPLTR1, IAPLTR1a, IAPLTR2, IAPLTR2a, and IAPLTR2b) in three different tissues (brain, liver, and kidney) of four 1-week-old mice littermates. The efficiency and the robustness of the designed technique was proven since the data reported the CpG methylation status of on average 52,000 CpG positions per sample at an average sequencing depth of 117x. The targeted nature of the technique was also well proven since more than 80% of the loci of the targeted subtypes and less than 15% of the non-targeted subtypes of IAPLTR have been covered by this technique. In each sample, the overall pattern of the CpG methylation percentage shows that even though most of the loci are between 80-100% methylated, a fair number of them show lower levels of methylation. An integrated analysis of all twelve samples revealed that approximately 2% of the IAPLTR loci have less than 80% average CpG methylation. These hypomethylated IAPLTR loci show no genomic position preference and the majority of these loci tend to be of the IAPLTR2 and IAPLTR2a subtype. In
an effort to find DNA methylation variations, different analyses were done on all the sequenced loci in different combinations of samples. The results revealed that the numbers of loci that vary in tissue-specific, individual-specific, and stochastic (both tissue-specific and individual-specific) manner are 196, 164, and 2521 respectively. A group of 26 loci were also found to be varying in a sex-specific (among the genders) manner. These results provided some unique perceptions into the epigenome of the IAPLTR repeat elements.

In this study, we have been able to demonstrate for the first time the single-base resolution CpG methylation status of a subset of repeat elements at level of individual locus using our HT-TREBS technique. Such analysis of the five subtypes of the IAPLTR repeat elements verified some of the previously assumed properties of the IAPLTRs: i) majority of the IAPLTRs are mostly methylated, ii) a fraction of the IAPLTRs escape the repressive methylation, and iii) a fraction of the IAPLTRs like the metastable epialleles exhibit variations in DNA methylation among individuals of the isogenic background (Mietz et al., 1990; Morgan et al., 1999). Such verifications have been done with the accompanying novel CpG methylation information of specific loci pertaining to those properties. These loci-specific information would be particularly useful in the future study of the hypomethylated and variable loci. Of particular importance is the finding that not only the mouse IAPLTR subtypes have loci that are variable in methylation as expected, their numbers and varieties are rather unexpected as well. In spite of the large number of loci having an average CpG methylation of greater than 80% in all the samples, subtle to major difference in their methylation values in each sample means that almost half of the 5135 loci analyzed show one or the other type of variation (intra-individual, inter-individual, or both). Also the presence of sex-based methylation variation on all chromosomes and not just the active
and inactive X chromosomes is a unique finding for mouse IAPLTR elements. Overall, these findings show the usefulness of obtaining the methylation information of individual locus of retrotransposons using the technique HT-TREBS.

An overview of the findings stated in chapters two and three indicates that mammalian retrotransposon families are in fact a versatile population of genetic elements. It contradicts some of the previous assumptions that the retrotransposons are fully methylated and marked by repressive histone marks with the exception of a few. Data in this dissertation suggests that those exceptions are more prevalent than previously thought. The studies in this dissertation show that as many 164 of the IAPLTR loci are potential metastable epiallele candidates. Also the numbers of variable loci of other variation types constitute almost half of the IAPLTR that have been studied. From these studies we can also hypothesize that the DNA methylation variability of IAPLTRs and other retrotransposons in general might be affected by the following factors:

i) Distance from nearby promoters of endogenous genes

ii) Spatial and temporal expression pattern of the nearby endogenous genes

iii) Types of nearby histone modification marks

It should be mentioned that not all these factors might simultaneously influence the variability of a particular locus since we the findings described in Chapter Three suggested that the hypomethylated IAPLTR loci do not have a preference for proximity to promoter and also some of the variation types apparently do not have preference for any particular histone modification. Therefore a locus by locus study of the IAPLTRs is necessary. Fortunately, the locus-based CpG methylation data that can be obtained from our proposed technique HT-TREBS would give a
more focused direction to such study with only a handful of specific loci to analyze instead of starting with all of them that are present in the genome.

Indeed the data in this dissertation strongly suggests that retrotransposons can be a major source of intra- and inter-individual epigenetic variations in the mammalian genome. In humans, any retrotransposon locus showing DNA methylation variations or hypomethylation might also be important as biomarkers of predisposition to diseases or as biomarkers of the diagnosis and prognosis of diseases. So considering the plethora of retrotransposons that are present in the human and other mammalian genomes, it is only fitting for the improvement of human health and for better understanding of the mammalian systems to carry out this kind of study in many other organisms including humans.

References


APPENDIX:
LICENSE TO PUBLISH

License to Publish

Journal/Title: Epigenetics
Author(s): Muhammad B. Elkan, Keonseong Kang
Manuscript #: 2011EPI0251R2

Title of the Contribution: Retrotransposons as a Major Source of Epigenetic Variations in the Mammalian Genome

To: Landes Bioscience (hereinafter "Landes")

1. In consideration of Landes agreeing to publish the Contribution, the Authors grant to Landes an exclusive license to:
   (a) Publish, reproduce, transmit, distribute and display the Contribution in all forms and formats whether now known or hereafter conceived or developed throughout the world,
   (b) Prepare translations and other derivative works based on the Contribution,
   (c) License any or all of the foregoing to others.

2. Copyright. Ownership of the copyright remains with the Authors, and provided that, when reproducing all or part of the Contribution, the Authors first acknowledge and reference publication in the Journal, the Authors retain the following non-exclusive rights:
   (a) Use of all or part of the Contribution (after publication in the Journal) in any book, article, thesis or dissertation written by the Authors without first requiring permission from Landes.
   (b) To reuse figures or tables created by the Authors and contained in the Contribution in the Authors' future works.
   (c) Make photocopies of all or part of the Contribution for use by the Authors and any academic institution where they work for the purpose of classroom teaching.
   (d) To post a copy of the Contribution as accepted for publication after peer review on the Author's personal website or institutional repository, or the Author's funding body's designated archive as soon as possible but no later than 12 months after publication in the Journal.

3. Author's Duties:
   (a) If the Contribution contains copyrighted material not owned by the Authors, then Authors shall obtain written permission from the copyright owner in a form acceptable to Landes, to include such copyrighted material in the Contribution and to reprint it in the Journal, and shall deliver such written permission to Landes together with delivery of the manuscript of the Contribution.
   (b) Landes may copyedit the Contribution as necessary in order to prepare it for publication in the Journal. The Author will proofread and correct the edited manuscript and shall return the proofs in a timely manner as directed. If proofs are not corrected in a timely manner, Landes may proceed with the publication of the Contribution as it deems appropriate.
   (c) The full text of the Contribution may not be posted on any social network, commercial website or self-archived on any website that is not the Author's personal website or the funding body's database prior to publication and for the first 12 months after publication unless the Open Access fee has been paid to Landes.
   (d) This license must be signed by the Authors and delivered to Landes as a condition for publication of the Contribution in the Journal.

4. The Authors warrant and represent that:
   (a) The Contribution has not been previously published, is not in the public domain and is the original work of the Authors.
   (b) The Authors are the sole owners of the copyright of the Contribution and have the full power and authority to assign the license to Landes. If the Contribution includes materials of others, the Authors have obtained permission of the original copyright owner to enable them to grant the rights contained herein.
   (c) The Contribution does not infringe the copyright or any other proprietary right of any party.
   (d) The Contribution contains no material that is defamatory, violates any right of privacy or is otherwise contrary to law.
   (e) All statements in the Contribution asserted as facts are based upon reasonable research for accuracy.
   (f) No formula or procedure contained in the Contribution would cause injury if used in accordance with the instructions and/or warnings contained in the Contribution.

5. Authors employed by the US Government:
   If the Author's contribution was performed as part of the Author's official duties as an employee of the US Government, the Publisher acknowledges that copyright may not be established in the United States.

6. Authors with funding from agencies with deposition requirements:
   (a) For Authors who must comply with NIH Public Access Policy, Landes will deposit an electronic version of the final, peer-reviewed manuscript upon acceptance for publication in NIH's PubMed Central archive to be publicly available no later than 12 months after publication by the Journal.
   (b) For Authors who must comply with UPMC funding agencies (including Wellcome Trust) Landes will deposit an electronic version of the final, peer-reviewed manuscript upon acceptance for publication in the UPMC archive to be available no later than 6 months after publication by the Journal.
   (c) For Authors who have paid the Open Access fee, please see the Open Access License Agreement.

Signed for and on behalf of the Authors: [Signature]
Date: 01-11-2013

Please print name: Muhammad B Elkan
VITA

Muhammad Badre Ekram is the youngest son of Engr. Badre Ebriz and Ms. Akhtar Jahan. He was born in 1984 and raised in Dhaka, Bangladesh. Muhammad graduated with a Bachelor of Science (with Honours) degree in Biochemistry & Molecular Biology from the University of Dhaka, Bangladesh in 2008. Subsequently he began his doctoral research under the guidance of Dr. Joomyeong Kim in the fall of 2008 in the Department of Biological Sciences at Louisiana State University. Mr. Ekram will graduate with the degree of Doctor of Philosophy in May 2013.