A Comprehensive Analysis of The Transcriptome In Human Colorectal Cancer Cells Lacking DNA Methyltransferases

SYNOPSIS OF THE THESIS

Submitted in partial fulfillment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY

by

PAWAN KUMAR TIWARY

Under the supervision of
Dr. David Eric Symer MD PhD

Birla Institute of Technology and Sciences
Pilani (Rajasthan) India
2012

Completed at National Cancer Institute –Frederick, Frederick, Maryland, USA under BITS-NCI (NIH) Collaboration

Extensive studies have been conducted to elucidate the basis and impact of genetic abnormalities involved in carcinogenesis. Abnormalities such as point mutations, chromosomal rearrangements and other changes to DNA sequences have become well characterized over the past several decades. By contrast, epigenetic contributions to carcinogenesis have been appreciated only recently. In the past three decades, research efforts have established numerous connections between various stages of carcinogenesis and epigenetic aberrations.

While epigenetic processes are heritable, they are not encoded in primary DNA sequences in the genome. They include transcriptional gene silencing and post-transcriptional gene silencing. Their molecular mechanisms, including DNA methylation, histone modifications (including methylation, acetylation and phosphorylation) and certain RNA-mediated events, appear to be conserved throughout most eukaryotes. Epigenetic controls of the mammalian genome play fundamentally important roles in regulating gene expression, genomic stability, differentiation and development. While changes occur normally in development, disrupted controls can lead to cancers and other diseases. Two central questions persist in epigenetics: how are the intricate, nonrandom epigenetic marks established and maintained in the genome? And secondly, what are the consequences of changes in them, whether in normal development, in diseases, or in experimental models?

Chapter 1

In human cancers, both aberrant increases and decreases in DNA methylation at CpG dinucleotides located in different genomic regions have been reported to occur frequently. These involve genome-wide hypomethylation and focal hypermethylation. The latter often occurs aberrantly at CpG islands. These are dense patches of CpG dinucleotides that are associated with a large fraction of gene promoters. CpG islands are defined to be >200 nt long, to have >50% GC content, and to have an observed/ expected CpG ratio that is higher than expected at > 60%. They normally remain unmethylated in all tissue types throughout development. Aberrant hypermethylation at certain CpG islands has been

associated with transcriptional gene silencing of tumor suppressor genes in cancer formation; such genes include *RB*, *VHL*, *p16/INK4a* and *MLH*. Conversely, certain genomic compartments normally are methylated in most somatic tissues, but can become hypomethylated in cancers. These include repetitive elements such as retrotransposons, the inactive X-chromosome in females, pericentromeric sequences, embryonic genes, and imprinted genes. Normal monoallelic expression of imprinted genes depends on their parent of origin, but this can be disrupted in cancers through loss of imprinting. Several studies have shown that decreased genomic methylation (either by genetic or pharmacologic manipulation) reactivates genes that are aberrantly silenced in cancer cells. However, only a few studies have looked into the effects of such epigenetic manipulation on normally methylated and silenced genes.

A recent study by Weber et al. has shown that decreased genomic methylation of HCT116 cells, due either to pharmacologic or genetic manipulation, caused induction of illegitimate transcript from an intronic L1 anti-sense promoter (ASP) located in the proto-oncogene cMet, thus creating a fusion transcript L1-cMet. They showed that this L1-cMet transcript caused decreased expression of the cMet gene. These results demonstrate the effect of genomic hypomethylation on the expression of genes, even if the *bonafide* promoter of the gene is unmethylated.

DNA methylation is established in normal development by the essential, *de novo* DNA methyltransferases, *DNMT3a* and *DNMT3b*. Both are expressed at various levels in adult somatic tissues, suggesting that they continue to play functional roles in development. By contrast, *DNMT1* is the major maintenance methyltransferase, but also has some *de novo* activity. *DNMT1* is expressed ubiquitously, with high levels of expression in dividing cells. *Dnmt1* mouse knockouts show embryonic lethality, as they lose monoallelic expression of most imprinted genes, show inactivation of the active X-chromosome due to reactivation of *Xist*, express high levels of intra-cisternal particle A (IAP) retrotransposons, and exhibit genomic instability in mouse embryonicstem (ES) cells. More recently, a new

modification of cytosine has been identified– hydroxymethylcytosine; Tahiliani et al. demonstrated that the enzyme Tet1, an iron-dependent a-ketoglutarate dioxygenase, catalyzes the formation of 5hmC from 5meC. Additionally, they suggest that the 5hmC may be an intermediate in the conversion of 5meC to cytosine, thus identifying an enzyme that can potentially be involved in demethylating DNA. These forms of cytosine methylation are the most newly discovered types and are likely to confound and confuse measurements of methylcytosine, and also they appear likely to play significant biological roles including in active demethylation.

To gain insights into whether disrupted DNA methylation could result in chromosomal instabilityor reactivate transcriptionally repressed genes in a nearly diploid, cultured human cancer cell line, Rhee et al. knocked out certain exons of *DNMT1* in HCT116 cells by homologous recombination. Resulting single knockout derivatives are surprisingly viable, dividing at a slightly slower rate than parental cells. Despite a 95% decrease in methyltransferase activity, they exhibit only approximately 20% reductions in genome-wide methylation, with extensive demethylation at pericentromeric satellites but nearly normal methylation persisting at *p16* and short interspersed elements (SINE retrotransposons).

By contrast, subsequent knockout of *DNMT3B* in *DNMT1* single knockout cells, resulting in double knockout (DKO) derivative cells, resulted in 95% reduction in genome-wide methylation and nearly complete abolishment of their DNA methyltransferase activity. DKO cells survive in culture; however, they grow at a significantly slower rate than wild-type cells. DKO cells show significant hypomethylation at repeat sequences such as satellite repeats and *Alu* elements, a loss of imprinting at the *Igf2* imprinted locus and reactivation of *p16*. Establishment of DKO cells demonstrated that the original *DNMT1* knockout allele does indeed substantially affect normal expression and function of this DNA methyltransferase.

Egger et al. recently demonstrated that a truncated, hypomorphic DNMT1 protein residually is expressed in the knockout cells, due to previously undetected alternative

splicing. Thus the *DNMT1* knockouts have increased hemi-methylation at specific CpG sites. RNA interference (RNAi) directed against remaining *DNMT1* transcripts in the knockout cells caused further reductions in their genome-wide methylation and their viability, suggesting that *DNMT1* is essential for cell survival. More recently, a conditional knockout of all DNMT1 catalytic function has been developed in HCT116 cells, resulting in mitotic catastrophe. This result again verifies that DNMT1 is an essential cytosine methyltransferase.

Despite their residual problems of cumulative chromosomal instability and residual *DNMT1* activity, DKO cells have provided a useful *in vitro* system to study genome-wide effects of hypomethylation.

Other experimental investigations using DKO cells have included the work of Polyak et al., who used DKO cells' hypomethylated DNA to validate methylation sensitive digital karyotyping, a new method to map genomic DNA methylation patterns. More recently, elevated levels of certain microRNAs were identified in DKO cells compared to their parent HCT116 cells, demonstrating that DNA methylation represses expression of these particular miRNA precursors. Moreover, the hypomethylation caused by the hypomorphic *DNMT1* allele can be considered to be more stable and uniform than that caused by pharmacological agents which may be variably toxic, cause off-target effects, have heterogeneous uptake and metabolism in a population of cells, etc.

We undertook this project to study changes in the human transcriptome upon genome-wide hypomethylation.Here we constructed and analyzed long-tag Serial Analysis of Gene Expression (longSAGE)libraries prepared from parental HCT116 cells and their DKO derivatives, and also generated and analyzed these cell lines' transcriptomes using mouse exon microarrays. Long SAGE has several advantages over other methods:

1. transcripts can be quantified without prior knowledge of their sequence structure; results are quantitative, even for poorly expressed genes;
2. results are comparable between platforms and with many previously reported

SAGE reference libraries;

3. novel gene transcripts can be identified based on their tag sequences; and
4. Subtle sequence differences distinguishing transcripts may be specifically and sensitively identified by sequencing rather than by differential hybridization on microarrays.

In this study, as described in Chapter 1, we identified several classes of genes whose transcription is upregulated strongly in the context of genome-wide hypomethylation. In numerous cases, differential longSAGE results were verified by Northern blotting, qRT-PCR and exon microarray results, and a direct connection with promoter hypomethylation could be established. As described in Chapter 2, we also studied the effects of genome-wide hypomethylation on the transcription of transposable elements such as L1, Alu, HERV and SVA elements, which were not addressed by previous studies. Several previously unreported longSAGE tags were identified, suggesting expression of previously unreported transcripts or splice variants. We statistically compared the findings of our study (longSAGE and exon microarray) with various previously accomplished transcriptional profiling studies on a genome-wide scale, thereby highlighting the strengths and weaknesses of various profiling methods. We identified a potential "transcriptome signature" of genome-wide hypomethylation.

We extensively analyzed the transcriptomes of HCT116 and DKO2L cells by comparing Serial Analysis of Gene Expression (longSAGE) libraries both with previously published cDNA microarray data and with our own exon microarray assays of total RNA. This was done to test the hypothesis that disruption of methyltransferase activity, leading to profound decreases in genomic methylation, would result in pronounced differences in transcript levels, particularly of interspersed retrotransposons whose methylation status has been used as a surrogate for such genome-wide methylation changes. Our aim has been to identify differentially expressed genes and genetic loci in the two related cell lineages, and to study the methylation at those genes/loci to find any correlations between differential transcription and underlying methylation status.

We observed profound differences between the two cell lines' transcriptomes, as measured by longSAGE library tag counts. Indeed, hundreds of genes are differentially expressed in DKO2L cells compared with parental HCT116 cells, as reflected by a low correlation coefficient between them. The upregulated genes include interferon-inducible genes, cancer testis genes, several embryonic genes, HLA genes and metallothionein genes, while the downregulated genes include several ribosomal protein genes, RNA processing and RNA metabolism genes.We utilized Database for Annotation, Visualization and Integrated Discovery-2006 (DAVID-2006, NIH) bioinformatics resources to categorize the affected genes comprehensively into various pathways, and biological and molecular functions to facilitate our understanding of the biological meaning of these findings. In addition, we surveyed relationships to physical locations on cytobands and chromosomes. In general, those genes involved in negative regulation of biological and cellular processes, e.g. DNA damage response genes, are amongst the most highly upregulated genes in the DKO2L cells. This general finding could be due to the fact that DKO2L cells show a much slower growth rate and have a high level of genomic stability and DNA damage when compared to HCT116 cells. Genes related to biosynthesis, cellular physiology, RNA metabolism and processing, and translation are amongst the downregulated genes. This finding could be due to slower growth, lower protein synthesis and metabolism rates in DKO2L cells as compared to HCT116.

Two independent techniques for expression profiling, i.e. Northern blotting and qRT-PCR, corroborated the most highly upregulated genes that were identified initially by our longSAGE findings. In addition, exon microarrays corroborated a very large number of highly differentially expressed genes that had been identified by longSAGE. All tested genes showed significant upregulation in DKO2L cells when compared to HCT116. In addition, similar upregulation was observed in several independent DKO clones. In contrast to comparisons with previously published cDNA microarray data, which mostly missed the most upregulated genes identified here, these consistent results using a variety of techniques

indicate that genomic hypomethylation profoundly disrupts the human transcriptome in specific and reproducible ways.

Several studies have shown previously that epigenomic reactivation by genetic manipulation or drug treatment deregulates a large number of genes. However, the deregulation of many of the affected genes could be due to an indirect effect, i.e. mediated by factors in *trans* rather than by direct changes in promoter methylation in *cis*. To establish a correlation between transcriptome changes and promoter methylation changes, we carried out methylation analysis at the promoters of genes that are most upregulated in DKO2L cells. Bisulfite sequencing analysis was chosen for these methylation studies because it is highly quantitative, and provides high resolution analysis of several individual CpG sites simultaneously in one bisulfite PCR amplicon. As expected, comparative bisulfite sequencing of the promoters of highly upregulated genes generally showed an inverse correlation between changes in their methylation and changes in their expression. All genes analyzed by bisulfite sequencing showed heavy methylation of the promoter in the HCT116 cells that became significantly hypomethylated in DKO2L cells. This result suggests that a major portion of differential gene expression in DKO cells is attributable to hypomethylation of their promoters *in cis*. However, a few genes were upregulated in DKO2L cells did not appear to have such an inverse correlation with their promoter methylation status. This result suggests that while upregulation of most affected genes resulted directly from promoter demethylation, there could be some indirectly affected genes whose expression was changed due to some other transcriptional control factors such as histone tail modifications or the presence of a crucial transcription factor which may be directly or indirectly regulated by methylation. Also, it is possible that in certain cases, a predicted promoter region is not the actual promoter for a particular gene, and instead another cryptic promoter located elsewhere could affect the expression of such a gene. Nonetheless, methylation analyses of the highly upregulated genes suggest a usual pattern of negative transcriptional regulation by promoter methylation that fits well with the classical view of promoter methylation as a repressor of transcription.

We found that interferon-inducible genes are one of the most affected classes of genes, including the most highly upregulated gene, *IFI27*. In total, 15 genes of this class are significantly upregulated. Several of these genes have CpG islands comprising their promoters. This result corroborates several prior studies in a variety of cell types using microarrays, which documented activation of interferon-inducible genes in response to genome-wide hypomethylation caused by pharmacological treatments or by genetic disruptions. Intriguingly, several interferon alpha-inducible genes also can be activated by expression of double-stranded RNA. While recent work has demonstrated that miRNAs are induced in DKO cells, more studies are needed to investigate the possibility that double stranded or antisense RNAs also might be upregulated upon genomic hypomethylation.

Our results also corroborate previous findings that other classes of genes are upregulated in the context of genomic hypomethylation, including cancer testis (CT) genes, BORIS, embryonic genes, metallothionein genes clustered at chromosome 16q13, and MHC class I genes.

While the transcriptional repression of many tumor suppressor genes in cancers has been associated with localized hypermethylation at their promoters, *e.g. p16*, *Rb*, *MLH1*, *RASSF1*, *VHL*, etc., we did not detect significant upregulation of any of them. Their expression could be silenced persistently by the residual, truncated DNMT1 expressed in the DKO cells or by repressive histone modifications and chromatin condensation, and alternatively could be attributed to a lack of tissue-specific transcription factors required for their expression.Vatolin et al. suggested that sustained ectopic expression of BORIS can cause hypermethylation at several CTCF/BORIS-binding regulatory sequences at the promoters of various tumor suppressor genes. We observed a 31-fold upregulation of BORIS in DKO2L cells as compared to HCT116, suggesting that BORIS could play a role in persistent silencing of tumor suppressor genes in DKO2L cells.

Loss of imprinting has been observed in a wide range of cancers. DNA methylation is a major mechanism implicated in the maintenance of imprinting, implying that faulty

methylation in cells might cause loss of imprinting. Rhee et al. showed that imprinting of *IGF2* is disrupted in DKO cells, as that gene is biallelically expressed. However, we observed no effect on the expression of imprinted genes in DKO2L vs. HCT116 cells. One possible reason could be that their transcript levels are below the limit of detection in our longSAGE libraries, despite relatively deep sequencing.

We also observed that there could be effects of genomic hypomethylation on the expression of genes even if their *bona fide* promoter is unmethylated. Such effects could be due to the induction of alternate transcripts, fusion transcripts and/or other cryptic promoters which originate from intronic retrotransposons (sense or antisense promoters). One such example is the paradoxical downregulation of the *MET* proto-oncogene in DKO cells. This has been associated with the hypomethylation-dependent induction of an antisense fusion transcript initiated from an antisense promoter in the L1 retrotransposon located in the intronic region of this gene, thus giving rise to the fusion transcripts L1-ASP. The exact molecular mechanism by which this fusion transcript is linked to downregulated *MET* expression remains unclear.

Statistical comparisons between our longSAGE study and a previous microarray-based study of the same HCT116 cells and their derivatives unexpectedly revealed a relatively poor overall correlation ($r^2 = 0.1$). Some possible explanations for this striking discrepancy between the data sets include differences in the DKO clones or passage numbers used for RNA extractions, and/or fundamental differences in the sensitivity and specificity determined by different techniques and platforms used in the studies. Notably, we validated most of the highly upregulated longSAGE tags observed in DKO2L cells, using independent methods including exon microarray, Northern blotting and/or qRT-PCR, and verified that the most upregulated transcripts are similarly overexpressed in several, independently derived DKO clones.

Using publicly available transcriptome (SAGE) data and comparing our findings with previous studies of induced hypomethylation, we compiled a set of "signature tags"

which may well characterize differential gene expression in the context of genome-wide hypomethylation in human colorectal cells. Strikingly, this list does not include tags representing transposons, since despite extensive hypomethylation of those widespread elements, we did not observe substantial upregulation of them (Chapter 2). In our collection of signature tags, which includes interferon-inducible genes, cancer-testis genes, metallothionein gene cluster and MHC class I genes, most genes represented by tags had a corresponding CpG island at or near their promoters, and all are either not expressed or poorly expressed in normal colon tissue. Most of these genes have a testis-restricted expression pattern and significant numbers of these genes are present on X-chromosome and belong to the CT gene family. Together with extensive previous results, our longSAGE data suggest that upregulation of these normally or developmentally restricted classes of genes specifically could reflect genome-wide hypomethylation.

An approach to refine and improve this proposed transcriptome signature of genomic hypomethylation in cultured human colorectal cancer cells would be to re-introduce *DNMT1* and *DNMT3B* genes into DKO2L cells, to determine if expression of members of the transcriptome signature returns back to expression levels in the parental cells. Of course, this assumes that karotypic instability in the DKO cells does not preclude reestablishment of “wildtype” expression patterns. In future experiments, we will attempt to measure the transcriptomes comprehensively in additional clonal cell isolates of HCT116 lacking DNA methyltransferases accomplished either by genetic knockout or knockdown by RNA interference; to use even more comprehensive expression profiling platforms such as RNA-Seq; and/or to comparatively study the transcriptome in other hypomethylated cell lines derived from colorectal tumors or other tissues.

Chapter 2

Although previously considered as "junk" DNA, mammalian genomic transposable elements play many possible biological roles that recently have become more clearly recognized. The human and mouse genomes each contain an enormous number of

transposable elements, accounting for nearly 50% of genomic content overall, and even more according to some estimates. These are broadly divided into four classes, namely DNA transposons, Long interspersed elements (LINE), short interspersed elements (SINE) and long terminal repeat-containing (LTR) retrotransposons. Retrotransposons transpose via RNA intermediates. Most of these elements have accumulated mutations in their sequences and are therefore incapable of moving in the genome. However, active elements that are capable of mobilization also are present in the genome. LINE-1 (L1) retrotransposons are the most abundant and oldest, comprising approximately 17% of the human genome. These elements are the most active in mouse. *Alu* elements (SINEs) are most active in human and utilize L1 machinery for mobilization. There is a controversy over whether or not HERV-K elements have been mobile recently in the human genome, although certain classes of mouse ERVs remain very active. Rampant retrotransposition events could lead to genomic instability, insertion mutation and interference with transcription of adjoining genes.

L1 elements are an autonomous, mobile element abundantly found in mammalian genome. They are about 6 kb in length and are abundantly found in AT-rich, gene-poor regions of the chromosomes corresponding to the G-bands. X-chromosome has relatively high density of L1 elements (29%) as compared to total genome (17%). There are an estimated 450,000 L1 elements present in the human genome which could be classified into two families, most of the actives one belong to the Ta (Hs) family. A full-length L1 elements consists of a 5' UTR containing a internal promoter, two ORFs (ORF1 and ORF2) which are separated by 63 bp non-coding spacer region required for retrotransposition phenomenon and a 3' UTR ending with a polyadenylationsequence. ORF1 encodes a 40-kDa protein with RNA binding and nucleic acid chaperone activities in vitro. ORF2 encodes three distinct conserved domains, i.e. an N-terminal endonuclease domain, central reverse transcriptase domain and a C-terminal zinc knuckle-like domain. L1 is thought to move in the genome by a target-primed reverse transcription mechanism.

SINEs (Short Interspersed Elements)are the second most abundant retrotransposons comprising ~13% of the genomic content and are short (100-400 bp) in length. They have an internal RNA polymerase III promoter, do not encode for any protein and require L1 machinery *in trans* for their movement. *Alu* elements are the most numerous (~1,000,000 copies) and the only active family of this class, comprising 10% of the genome. They are approximately 300 bp in length and have a high density of CpG dinucleotides that are highly methylated in somatic tissues.

SVA elements are hominid-specific, non-autonomous, composite retrotransposons and are the youngest of all retrotransposon families. Their components are (in reverse order) SINE-R, VNTR and *Alu*. There are more than 2,500 SVA elements identified in human genome. They are enriched in G+C rich regions. SVA elements are classified into 6 sub-families (SVA-A to SVA-F). SVA elements have evolved recently, as demonstrated by their lack of high level of sequence divergence. Movement of SVA element is facilitated by L1 retrotransposons *in trans*. SVA elements are highly methylated in all somatic tissues of adult.

LTR (Long Terminal Repeat) retrotransposonsare autonomous retrotransposons comprising about 8.3% of the genome. These elements have long terminal repeats at both their 3' and 5' ends containing the required transcriptional regulatory sequences. In between their long terminal repeats, these elements often have *gag* and *pol* genes encoding protease, reverse transcriptase, RNaseH and integrase. The endogenous retrovirus-K (ERV-K) family of LTR class is an actively mobilized family and has about 8,000 copies in the mammalian genome.

Retrotransposition events in the mammalian genome can have several deleterious effects. L1 movement in the genome can promote unequal homologous recombination and/or insertion into genes, thus affecting normal transcription. During the retrotransposition process, two single stranded breaks that are created close to each other could act as a double stranded break, thereby increasing the chances of chromosomal breakage, deletion, translocation and illegitimate recombination. Once the retrotransposition

event has taken place, new insertions can cause various forms of transcriptional deregulation of the neighboring genes or transcription units depending upon their context and orientation. There are several documented cases of diseases caused by insertion of various actively mobilized classes of retrotransposons. L1 provides the necessary machinery for mobilization of other non-autonomous elements, *Alu* elements are known to be active and require L1 machinery to move. There are several documented cases of *Alu* insertions causing human diseases. SVA elements are one of the least studied retrotransposons; however, there are at least three-documented cases of SVA insertion-mediated human diseases. Further, L1s can also give rise to novel genes through shuffling by 3' or 5' transduction. Recent studies suggest that, although previously less emphasized, repetitive elements (retrotransposons) are commonly expressed in a highly tissue-specific manner (especially embryonic tissues) by utilizing their internal sense and/or antisense promoters. Retrotransposons close to the 5' end of a protein coding region may act as an alternate promoter that may express alternative mRNA and other non-coding RNAs, thus regulating the nearby genes and altering the transcriptome.

Given the deleterious potential consequences of retrotransposon movement, it is surprising to know that relatively low numbers of mutations and other harmful effects have been attributed to their movement to date. Does this suggest that the genome has some kind of defense system that checks the movement of these elements?

Most of the L1 elements in the genome are defective due to 5' truncations, point mutations and inversions, thereby leaving them incapable of moving. There are about 3,000-5,000 full length L1 elements residing in the human genome of which only about 80-100 are considered capable of actively moving in the genome. Bestor et al proposed that cytosine methylation may serve as a host genome-defense system that helps check the expression of these elements by silencing them through transcriptional gene silencing. L1s are generally silenced except in germ cells and during embryonic development. It is well known that endogenous L1 and other repetitive elements are highly methylated in somatic

cells, which is responsible for keeping these elements in a silent state. Any decrease in methylation at these transposable elements increases the risk of their transcription and movement in the genome. One of the initial publications showed using oligonucleotide microarray that genome-wide hypomethylation in mouse embryonic fibroblasts (with disrupted *Dnmt1*) caused increased expression of a particular L1 element (L1Md-Tf14). It was shown that in mouse germ cells, disruption of *Dnmt3L* prevents *de novo* methylation of non-LTR and LTR retrotransposons, thereby causing high expression of these elements in spermatogonia and spermatocytes. In a recent study, cancer-specific chimeric transcripts were isolated in cells where L1 retrotransposons were hypomethylated, leading to genomic instability and making them susceptible to cancer progression. Another study by Rangwala et al has shown that many L1 elements are expressed in human somatic cells, thus significantly contributing to the transcriptome.

RNA interference (RNAi) due to antisense promoter activity is also thought to play a modest role in regulating expression of human L1 retrotransposons. Further, it has been suggested that Miwi proteins, which interact with small RNAs called piRNAs, play a role in regulating expression of L1s; Mili mutant mouse testis shows expression of L1 and IAP elements. Interestingly, they also have decreased methylation at L1 elements. Additional cellular inhibitors involved in checking L1 expression are members of the APOBEC3 protein family, which appear to inhibit L1 movement without editing new integrant sequences.

There is evidence for DNA methylation playing a role in regulating the expression of HERVs. One study showed that treating Tera-1 cells with 5-azacytidine increased the expression of HERV-K(HML-2) Gag protein. Another study on Tera-1 cells supported CpG methylation as an important factor in silencing these elements. However, it was also suggested that CpG methylation is not the only factor needed for silencing the HERV promoter. In mice it was shown that disruption of *Dnmt1* causes increased IAP expression (one of the ERV LTR retrotransposon family). Oligonucleotide microarray analysis on Dnmt1-disrupted, p53-inactivated MEFs showed increased expression of IAP elements.

To our knowledge, no previous study has used genome-wide expression profiling either by microarray or sequencing based techniques to look directly at the effects of genome-wide hypomethylation on the expression of transposable elements. In this study, we measured gene expression profiles of HCT116 and DKO2L cells using longSAGE to compare the transcriptomes of HCT16 and DKO2L cells and to investigate the effects of genome-wide hypomethylation on the transcriptional regulation of L1 retrotransposons and other transposable elements such as *Alu*, HERV and SVA elements.

According to the "genome defense" model proposed by Bestor et al, CpG methylation is believed to be an important factor in silencing of transposable elements and repetitive sequence. In fact, an enormous portion of the human and mouse genomes (~ 45%) consists of transposable elements and most DNA methylation is focused on such elements. Activity of these transposable elements can lead to genomic instability, insertional mutagenesis and/or activation or inhibition of cancer-causing genes or oncogenes. For example, expression of IAP elements in mouse is kept under control by cytosine methylation and it was shown that in Dnmt1 hypomorphs ($Dnmt1^{chip/-}$), the centromeric repeats and IAP elements are hypomethylated and expressed.

One of the major focuses of our study has been to determine the effect of genomic demethylation in DKO2 cells on transcription of retrotransposons and endogenous retroviral elements. Surprisingly, our longSAGE results showed only a modest 3-fold increase in expression of human L1 retrotransposons in DKO2L cells. Bisulfite sequencing at the 5' UTR promoter region of L1 retrotransposons showed profoundly decreased methylation in DKO2L cells as compared to HCT116 suggesting that DNA methylation plays an important role in silencing of these parasitic elements. These observations fit the genome defense model proposed by Bestor et al. However despite profound hypomethylation there was only a modest increase in expression of L1, suggesting that DNA methylation might not be the only mechanism playing a role in regulating the expression or silencing these transposable elements. Other mechanisms such as histone tail

modifications and RNAi and cellular inhibitors including members of APOBEC family of proteins possibly could be involved in silencing of these elements. In addition, it is possible that transcription factors required for expression of these endogenous transposable elements are absent or limiting in somatic, human colorectal cancer cells.

Analysis of transposon expression is a complex undertaking, because of their highly repetitive nature genome-wide. For example, microarrays typically exclude probes for such Repeat Masker-identified sequences, because it would be impossible to identify which element(s) out of potentially thousands could give rise to transcripts. An additional complication is that the elements frequently are degenerate, due to nucleotide substitution, recombination events, etc. over time. Moreover, unlike single copy genes, thousands of repetitive elements could template transcripts, posing a challenge about normalization of transcript counts to template copies.

These problems are illustrated by L1 elements in the human genome, which have integrated over time as member of successful primate-specific or human-specific L1 subfamilies. Moreover, genomic L1 structures frequently are truncated from their 5' ends, so most templates for sense-strand transcripts would lack the L1-specific promoter in the 5' UTR, but still include 3' L1 sequences. L1 transcripts can undergo premature polyadenylation and termination, and alternative splicing. Recognizing that L1 genomic templates of many shapes, ages, sizes and numbers can give rise to complicated distribution of transcripts, we predicted and counted longSAGE tag frequencies in our libraries corresponding to every possible tag along the consensus "young" L1.3 sequence. This assumes transcripts' taggable 3' ends could occur anywhere along the L1 template in either orientation. We also recognize that active L1 variants might have different sequences at some of the tag positions. Given the significant numbers and complexity of such non-consensus tags, such tags are not analyzed further by our work here.

We studied the expression of sense-strand L1 tags in all publicly available long-SAGE libraries (http://www.ncbi.nlm.nih.gov/SAGE/). Our survey showed that most L1

tags are weakly expressed across various previously reported longSAGE libraries derived from many human tissues. However, the 3' most tag and the 3rd tag from the 5'end showed significant expression in those libraries. We also compared the expression of these predicted L1 tags in HCT116 and DKO2L libraries with the cumulative average expression of all the publicly available longSAGE libraries. Our HCT116 library did not show expression of any on these tags. However, our DKO2L library showed expression of 4 of predicted L1 tags which included the 3' tag and the 3rd tag from 5'end. The expression of these tags in DKO2L was slightly higher than the cumulative average expression in all the long-SAGE libraries previously sequenced. These comparisons suggest that various predicted L1 sense-orientation tags have higher expression in DKO2L libraries than other publicly available libraries.

We were interested in studying the effect of genome-wide hypomethylation on the expression of the other classes of transposons, including *Alu* elements. The comparison of expression profiles in HCT116 and DKO2L cells showed that there was no significant change in the expression level of *Alu* elements between the two cell lines. However, bisulfite sequencing across the entire *Alu* sequence, performed in bulk for genome-wide analysis, revealed significant decreases in cytosine methylation in the DKO2L cells as compared to HCT116 cells. These results showed once again that despite significant demethylation at *Alu* sequences, there is no significant change detected in the expression of these elements. Although *Alu* elements have high CpG densities and are highly methylated in somatic cells, decreases in DNA methylation may not be sufficient for their expression, unlike RNA polymerase II transcribed elements. Tissue specific-factors and/or lack of effect of CpG methylation on RNA polymerase III may play additional roles in this regard.

We were unable to detect the expression of any of the predicted HERV-K tags in both HCT116 and DKO2L libraries, suggesting either that there was no effect of genome-wide hypomethylation on the expression of these elements or that their expression was below the limit of detection of our long-SAGE libraries.

SVA elements are composite retrotransposons which are highly methylated in all somatic tissues, suggesting that DNA methylation might play an important role in regulating expression of these elements, again through transcriptional gene silencing. Upon comparing the expression of all the predicted SVA tags in HCT116 and DKO2L libraries, we found that genome-wide hypomethylation in DKO2L caused increases expression of SVA elements from 2-fold to 15-fold as compared to the HCT116 cells.

It should be noted that expression of other long, intergenic noncoding RNAs or retrotransposon-initiated fusion transcripts was not assayed by methods used in this study. In particular, antisense L1 promoters may initiate expression of many diverse fusion transcripts, but these transcripts' diverse 3' ends would not be uniquely or properly attributed by longSAGE tags to such promoters.

Collectively these results suggest that DNA methylation may play variable roles in the regulation of expression of various classes of retrotransposons. Different classes appear to show different levels of effects due to changes in methylation status. Thus DNA methylation is not the sole mechanism regulating expression of transposons in these cultured cells. There could be multiple overlapping regulatory mechanisms such as histone tail modifications, regulatory RNAs, and cellular inhibitors like APOBEC proteins, or limiting transcription factors affecting certain classes. The expression of a particular class of retrotransposons would depend on the interplay of these multiple regulatory mechanisms. Such regulatory mechanisms exert very tight governance over the expression of the repetitive elements, constituting a very large fraction of the genome, lest any rampant expression of the transposable elements could disrupt the transcriptome or lead to increased genomic instability and diseases.