Published in final edited form as:

Gene. 2012 January 15; 492(1): 104–109. doi:10.1016/j.gene.2011.10.046.

Distribution of a marker of germline methylation differs between major families of transposon-derived repeats in the human genome

Martin I. Sigurdsson^{a,b}, Albert V. Smith^c, Hans T. Bjornsson^{d,e,*}, and Jon J. Jonsson^{a,b,**}

^aDepartment of Biochemistry and Molecular Biology, Faculty of Medicine, University of Iceland, IS-101, Iceland ^bDepartment of Genetics and Molecular Medicine, Landspítali-University Hospital, Reykjavik, IS-101, Iceland ^cIcelandic Heart Association, Kopavogur, IS-201, Iceland ^dMcKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, 21287 MD, United States ^eHarriet Lane Pediatric Residency Program, Baltimore, MD, United States

Abstract

A potential relationship between transposon-derived repeats (TDR) and human germline methylation is of biological importance since many genes are flanked by TDR and methylation could affect the expression of nearby genes. Furthermore, DNA methylation has been suggested as a global defense mechanism against genome instability threatened by TDR. We studied the correlation between the density of HapMap methyl-associated SNPs (mSNPs), a marker of germline methylation, and proportion of TDR.

After correcting for confounding variables, we found a negative correlation between proportion of *Alu* repeats and mSNP density for 125–1000 kb windows. Similar results were found for the most active subgroup of repeats. In contrast, a negative correlation between proportion of L1 repeats and mSNP density was found only in the larger 1000 kb windows.

Using methylation data on germ cells (sperm) from the Human Epigenome Project, we found a lower proportion of *Alu* repeats adjacent (3–15 kb) to hypermethylated amplicons. On the contrary, there was a higher proportion of L1 repeats in the 3–5 kb of sequence flanking hypermethylated amplicons but not in the 10–15 kb flanks.

Our data indicate a differential response to the major repeat families and that DNA methylation is unlikely to be a uniform global defense system against all TDR. It appears to play a role for the L1 subgroup, with sequences adjacent to L1 repeats methylated in response to their proximity. In contrast, sequences adjacent to *Alu* repeats appear to be hypomethylated, arguing against a role of methylation in germline defense against those elements.

Keywords

Repeats; *Alu*; L1; HapMap; Germline; Methylation

1. Introduction

About 50% of the human genome consists of repeat sequences, with the vast majority being transposon-derived repeats (TDR). The L1 subfamily of the LINE elements represents about 16.9% of the human genome (Lander et al., 2001) and is the major family capable of active retrotransposition (Dombroski et al., 1991). A second major active TDR is the *Alu* subfamily of SINE elements, representing about 10.6% of the human genome (Lander et al., 2001). The *Alu* repeats are thought to be dependent on the L1-coded ORF2p for retrotransposition (Dewannieux et al., 2003). Both *Alu* and L1 are preferentially inserted into a short AT-rich L1 endonuclease target site (Boissinot et al., 2004). Flanking sequences of insertion sites of the active polymorphic L1 Ta-1 TDR have a GC percentage lower than the genome average, albeit higher than fixed Ta-1 insertions (Boissinot et al., 2004). However a recent study found that the 20 kb regions flanking de novo L1 insertions had the same GC ratio as the genome average (Gasior et al., 2007). Fixed *Alu* repeats are preferentially located within GC-rich regions of the genome, in contrast to L1 repeats which are preferentially located in AT-rich regions (Lander et al., 2001). This might suggest a similar insertional preference and different post-insertional selection for *Alu* repeats compared to L1 repeats (Gasior et al., 2007).

The retrotransposition of active TDR imposes a threat to the host genome (Amariglio et al., 1993; Belancio et al., 2008; Callinan et al., 2006). Several examples of L1 and *Alu* insertions causing human diseases have been described (Miki, 1998). Estimations of the contribution from TDR insertions to the human mutational spectra vary greatly, but have been suggested to be 0.1% (Kazazian, 1998) and 0.2% of all human mutations (Callinan et al., 2006). This is most certainly an underestimation given the difficulty of PCR based methods in uncovering TDR insertions (Kazazian, 1998). Genomic stability is also threatened by TDR sequence homology (Hedges et al., 2007) with an estimated 0.17% of all human genetic disease resulting from unequal *Alu/Alu* recombination (Callinan et al., 2006).

The existence of TDRs is conflicting from an evolutionary standpoint. While the beneficiary effects of allowing TDR activity are less clear, genome stability is threatened by TDR, and unrestrained TDR activity eventually eliminates the fitness of the species (Brookfield, 1991). This calls for defense mechanisms capable of controlling TDR activity tightly (Zamudio et al., 2010). These include mechanisms based on enzymes editing TDR RNA such as the APOBEC3 family (Bogerd et al., 2006), some of which are *Alu* specific (Hulme et al., 2007), and the PIWI defense system that mediates defense against L1 and IAP repeats (Girard et al., 2006). A defense system against long terminal repeats utilizing histone methylation has also been recently discovered (Matsui et al., 2010).

Methylation of cytosine, a heritable DNA modification is targeted to the CpG dinucleotide in the human genome. Patterns of DNA methylation differ between healthy tissue and cancer (Feinberg et al., 1983) and are changed within individuals over time (Bjornsson et al., 2008). Deamination of methylated cytosine creates a thymine base that can escape the DNA repair machinery. This can cause a C→T transition and is thought to cause hypermutability of methylated CpGs.

The “Genome Host Defense” hypothesis proposes that DNA methylation might suppress TDR transcription and thereby minimize harmful TDR effects (Yoder et al., 1997). The hypothesis is supported by observations of frequent occurrence of the CpG dinucleotide within TDR in combination with generalized methylation of TDR in somatic tissues. Also, degradation of TDRs by mutations of methylated cytosines might render them inactive. A major argument against the “Genome Host Defense” hypothesis is hypomethylation and the apparent transcription of some TDR in the germline (Hata et al., 1997; Ostertag et al., 2002),

where the defense system should be most critical (Bird, 1997). Methylation analysis of *Alu* repeats based on Southern blotting suggests hypomethylation in sperm (Hellmann-Blumberg et al., 1993) and they are inherited in an unmethylated state from male germ cells (Rubin et al., 1994). Bioinformatic analysis also suggests a relative hypomethylation of *Alu* repeats compared to other TDR classes in the germline (Kim et al., 2007; Xie et al., 2009; Xing et al., 2004). L1 repeats, at least the subset cleaved by MspI/HpaII restriction enzymes, are however inherited in an unmethylated state from the female germline (Sanford et al., 1987). Other defense mechanisms have also been described more recently including a family of proteins (e.g. APOBEC3) inhibiting TDR retrotransposition (Belancio et al., 2008).

It has been suggested that both L1 (Liang et al., 2002) and *Alu* (Graff et al., 1997) repeats might induce de novo methylation of adjacent sequences (Turker, 2002), although the evidence is limited. The size range of this methylation spread is unknown, but can possibly be within the range of 1 kb, as described for the de novo methylation of recently inserted retroviral elements (Jähner et al., 1985). The methylation of TDR could therefore have profound effects on methylation of adjacent sequences.

The “Genome Host Defense” predicts a positive correlation between the proportion of TDR and adjacent germline methylation. The correlation should be evident even if the germline is overall hypomethylated compared to somatic tissues. TDR might also participate in the phenomenon of metastable epialleles, where epigenetic modifications of particular alleles in the genes occur and result in extensive interindividual variability in gene expression (Waterland et al., 2003). We therefore might expect TDR content to be correlated both with global methylation patterns and gene expression in certain circumstances.

Besides direct spread of methylation from TDR to adjacent sequences there are two alternative explanations for a relationship between TDR and methylation of adjacent sequence in the germline: Either the insertion site of TDR or the post-insertional fixation of TDR could be influenced by sequence methylation. None of these explanations are mutually exclusive. Because of the biological importance of these three different explanations, it is of interest to study whether a correlation exists between TDR and the methylation of adjacent sequence in the germline. Previously we argued that the density of CpG/TpG SNPs (methyl-associated SNPs, mSNPs) can be used as a surrogate marker of germline DNA methylation under the neutral theory of molecular evolution (Sigurdsson et al., 2009). Subsequently we used the density of 434,198 HapMap mSNPs to create a model of human regional germline methylation. After validation, we demonstrated a positive correlation between our methylation marker and regional recombination rate (Sigurdsson et al., 2009). Our germline methylation marker is well suited to study the possible correlation between TDR and methylation of adjacent sequence in the germline.

2. Materials and methods

2.1. Definitions

We defined methylation-associated SNP (mSNP) as any C/T polymorphism with an adjacent 3' guanine or G/A polymorphism with adjacent 5' cytosine (corresponding to a C/T polymorphism on the opposite strand). All mSNPs where the ancestral allele was either a T or A were removed, maximizing the likelihood of mSNP representing methylation. As previously described, we mapped 434,198 mSNPs from the HapMap dataset using this approach. We demonstrated that the SNP spectra of the HapMap database represented the suggested hypermutability of methylated cytosine, and that the ratio of observed versus expected mSNPs was inversely correlated with the density of CpG islands (Sigurdsson et al., 2009). These data were then added to a window-based data set of proportions of GC, CpGs, exons, recombination rate and repeat subfamilies.

2.2. HEP data set

We downloaded the most recent data release from the Human Epigenome Project (HEP) from the project website (<http://www.epigenome.org>, downloaded on 6/15/2008) (Eckhardt et al., 2006). Results include bisulfite sequencing results of 2524 amplicons chosen from chromosomes 6, 20 and 22 in 12 different tissues (Eckhardt et al., 2006). We wrote programs to extract sperm specific data which allowed us to average the methylation across amplicons. We then created two subgroups of amplicons based on methylation, a hypermethylation group (>80% methylation, n=1001) and a hypomethylation group (<20% methylation, n=701) as defined by Eckhardt et al. (2006). We then extracted 3 kb, 5 kb, 10 kb and 15 kb of sequence flanking the middle location of each amplicon from the NCBI Build 34 of the human genome. Subsequently, the sequences were submitted to a CENSOR server (Jurka et al., 2005) installed on a local cluster computer (Version 4.2). The server was set to normal sensitivity and compared the submitted data to a data set of human repeated sequences.

We tested the correlation between methylation of a single CpG and the average methylation of the adjacent 3–15 kb of sequence by downloading the entire dataset of a whole genome bisulfite sequencing of the H1 and IMR90 cell lines (Lister et al., 2009), and calculating the average methylation of the 3–15 kb of adjacent sequence to every hundredth CpG in the data set.

2.3. Programs

All programs were developed in the JAVA programming language using the Textpad editor (Helios Software Solutions, England) and rigorously tested prior to data analysis. Several scripts were developed for statistical analysis and graphics in the R statistical language. The source code of all major programs used is available at www.hi.is/~mis.

2.4. Statistical analysis

All genome-wide correlations were done using four different sizes (125 kb, 250 kb, 500 kb and 1000 kb) of non-overlapping windows. For each window the relevant sequence properties were counted as absolute number of bases per window to calculate mSNP density, proportion of repeat subfamily, GC ratio, CpG dinucleotide proportion, genes and exons density. These properties were extracted from the human sequence (NCBI build 35) and UCSC tables describing genome properties (Karolchik et al., 2004). Recombination rate was extracted from the HapMap phase II data set (International HapMap Consortium, 2007). Windows containing sequencing gaps were removed prior to analysis. All variables failed the Kolmogorov–Smirnov normality test. We therefore used Spearman ranked partial correlation to correlate mSNP density of windows and proportion of TDR subfamilies while correcting for the content of GC ratio, CpG dinucleotide, SNP density exons and recombination rate.

Prior to multiple linear regression analysis, the data was transformed to the natural logarithm. Linear regression was then done using proportion of the TDR subfamily as the response variable and mSNP density, SNP density, GC ratio, CpG dinucleotide proportion, recombination rate and exon/gene density as the predictor variables. Although linear models are appropriate and widely used to test for significance and directionality of a relationship between several variables, a non-linear contribution cannot be excluded. This would more likely affect the effect magnitude of the relationship rather than directionality.

To estimate *p*-values for the HEP data set analysis, we did permutation testing. The 1702 amplicons were randomly split 10,000 times into two groups (A and B, with $n_A = 1001$ and $n_B = 701$) using an in-house JAVA program. We then compared the observed difference in

TDR proportion between the hypermethylated and hypomethylated group with the distribution of mean differences of TDR proportion from the randomly created groups. A p -value was then estimated by counting number of randomized mean difference equal or greater than the observed value.

We adjusted the partial correlations using a Bonferroni correction. Since 25 correlations were done, a p -value less than $0.05/25=0.002$ was considered statistically significant. For the HEP data set, significance of measurements for 10 different TDR subfamilies was compared, therefore a p -value less than $0.05/10=0.005$ was considered statistically significant. We note, however, that certain features tested were inter-correlated (such as absolute proportion of SINE and *Alu* repeats), possibly increasing the likelihood of type II errors. All statistical analysis as well as preparation of figures was done using the R package version 2.8.1, using the `ks.test` (Kolmogorov–Smirnov test), `cor.test` (correlation), `pcor.test` (partial correlation) and `lm` (linear model) routines.

3. Results

3.1. Correlation between mSNPs and TDR subfamilies

We examined both absolute and partial correlation between the mSNP density and the proportion of TDR and TDR subfamilies while correcting for known confounding variables affecting either variable (Table 1). The variables affecting mSNP density and potentially TDR density are SNP density and CpG dinucleotide proportion (Sigurdsson et al., 2009). Similarly, GC ratio and gene density affect TDR density significantly (Lander et al., 2001). We also corrected for recombination rate as we have recently demonstrated a positive correlation between mSNP density and recombination rate (Sigurdsson et al., 2009), and recombination rate is correlated with TDR repeats density (Myers et al., 2005). Partial correlation was emphasized as it holds constant the effect on the confounding variables that vary greatly between the different TDR subfamilies (such as the GC ratio) and the mSNP marker (such as SNP density). This expands a prior analysis using a much smaller data set limited to sequence surrounding 385 genes (Bjornsson et al., 2006).

We found a statistically significant absolute and partial negative correlation between mSNP density and the proportion of TDR in all window sizes tested (125 kb, 250 kb, 500 kb, 1000 kb, Table 1). For the 125–500 kb window sizes this was mostly due to a strong negative correlation between mSNP density and proportion of *Alu* repeats (similar for *AluY*, *AluJ* and *AluS*). The partial correlation between mSNP density and proportion of L1 repeats was strong and negative in 1000 kb window size, but not significant in smaller window sizes of 125–500 kb (Table 1). The pattern for LTR was similar to the L1 repeats pattern in all window sizes. There was a positive correlation between proportion of simple repeats and mSNP density over all window sizes. Previously it is known that methylation is critical for the stability of simple repeats in somatic cells (Nichol et al., 2002). Our observation suggests that this might also apply in the germline. We defined active repeats as those about 11,000 *Alu*, L1 and SVA elements differentially present in the human and the chimpanzee genomes (Mills et al., 2006). This subset has a higher likelihood of including active elements as they have evidence of recent transposition. Interestingly, a negative absolute and partial correlation was found between our mSNP marker and this subset of repeats in window sizes of 125–500 kb (Table 1).

3.2. Multiple linear regression modeling of TDR subfamilies as a function of genetic and epigenetic variables

We pursued a multiple linear models of the two largest TDR families that include active transposons (L1 and *Alu*) in addition to the subset of active TDR. The predictor variables were known sequence features affecting proportions of TDR (density of exons, GC ratio,

recombination rate and proportion of repeats) in addition to mSNP density and known confounding variables (SNP density and CpG dinucleotide proportion) (Table 2). For *Alu* repeats there was a significant negative contribution from mSNP density to the model, with mSNP density the second strongest predictor (standardized β value) of proportion of *Alu* repeats in all window sizes (125–1000 kb) (Table 2, Supplemental Table S1). The CpG density was the strongest predictor of *Alu* proportion. The adjusted R^2 of our linear model for *Alu* repeats was 0.530–0.769 for different window sizes.

For the L1 repeats, there was a significant negative contribution from mSNP density to the model in 1000 kb windows only, where mSNP density was the third strongest predictor of L1 proportion (Table 2). Here, CpG density was also the strongest predictor of L1 proportion. For the other window sizes there was either a small or no contribution from mSNP density to the model (Supplemental Table S1). The adjusted R^2 of our linear model for L1 repeats was 0.384–0.607 for different window sizes.

There was a significant negative contribution from mSNP density to the model of active repeats, with mSNP being the third strongest predictor in the model in all window sizes tested (125–1000 kb) (Table 2, Supplemental Table S1). As in the other two models, CpG density was the strongest predictor of active repeats proportion. The adjusted R^2 of our linear model for active repeats was 0.483–0.730 for different window sizes.

The analysis was performed with non-transformed data and yielded similar results albeit slightly lower adjusted R^2 values. Furthermore, adjusting for gene density instead of exon density did not change the results (data not shown).

3.3. TDR analysis context of differentially methylated regions

Next we studied TDR subfamily proportions flanking differentially methylated sequences using an independent biological data set. The Human Epigenome Project (HEP) has released methylation information for 2524 amplicons in three chromosomes for 12 different tissues using bisulfite sequencing, the current gold standard for methylation analysis (Eckhardt et al., 2006). We used the HEP data from sperm as an indicator of the male germline methylation, comparing proportion of TDR in sequences flanking the hypermethylated (>80% average methylation) and hypomethylated group (<20% average methylation) of amplicons. Previously we have shown that the methylation of the several CpGs within each amplicon is similar (Sigurdsson et al., 2009). Furthermore, we found a positive correlation between the absolute methylation value of a single CpG and the mean methylation of the adjacent 3–15 kb of sequence using the whole genome bisulfite sequencing results of the H1 ($r=0.29$ – 0.34 based on size, $p<10^{-15}$) and IMR90 ($r=0.32$ – 0.36 based on size, $p<10^{-15}$) cell lines (Lister et al., 2009). We also found that the average methylation of the 3–15 kb of adjacent sequence was significantly higher for the CpGs with more than 80% methylation compared to CpGs with less than 20% methylation (e.g. 50% vs 12%, $p<10^{-15}$ for the 15 kb flank size in the H1 cell line). This indicates that the methylation at a given region can be used to infer the methylation of adjacent sequence.

Results for 5 kb flanks are shown in Table 3 (Supplemental Tables S2 and S3 contain the 3, 10 and 15 kb flanks). For all flank lengths, there was a significantly lower proportion of *Alu* repeats in the hypermethylated amplicon group compared to the hypomethylated amplicon group (Fig. 1a). In the two smaller flank lengths (3 kb and 5 kb) there was a higher proportion of L1 repeats in the hypermethylated amplicon group compared to the hypomethylated amplicon group (Table 3) (Fig. 1b) (Supplemental Table S2), but the trend was not significant in the larger flank sizes (Supplemental Table S3). The results for the less abundant LTR had a similar trend as L1 repeats in all flank sizes although they were non-significant after correcting for multiple hypothesis testing (Table 3) (Fig. 1c) (Supplemental

Tables S2 and S3). We found no difference in the proportion of DNA transposons between the hypermethylated and hypomethylated amplicon groups (Table 3) (Supplemental Tables S2 and S3).

4. Discussion

It is tempting to speculate on the several possible relationships between the major TDR subfamilies and germline methylation, although it should be stressed that a cause and effect can never be determined based on correlation data. Several scenarios can be envisioned. Perhaps methylation affects the likelihood of TDR insertion where the hypomethylated and more open chromatin might be more susceptible to TDR insertion. Methylation could also affect the likelihood of TDR fixation; hypomethylated regions include more transcribed areas and these might be more sensitive to TDR insertion. However, this could be subfamily-specific as the insertion of some TDR elements is potentially beneficial (Lev-Maor et al., 2008). Methylation of TDR also affects their transcription, at least for L1 repeats (Hata et al., 1997; Ostertag et al., 2002). Finally, there is some evidence that methylation of TDR spreads over neighboring regions of few kilobases (Jähner et al., 1985).

The dependence of *Alu* repeats on the L1 repeat machinery for retrotransposition (Dewannieux et al., 2003) suggests similar insertion preferences these subfamilies. This is supported by data suggesting that recent L1 repeats insert into sequences with similar GC ratios as *Alu* repeats (Gasior et al., 2007). We assume similar insertion preferences when discussing our two major observations : a) A significant negative correlation between regional *Alu* repeat levels and germline methylation for both proximal sequences (3–15 kb) close to genes as well as distant regions (125–1000 kb) and b) A correlation gradient between regional L1 repeat levels and markers of germline methylation; in the 3–5 kb proximal sequences there is a positive correlation (in male germline at least), in 10–15 kb proximal sequences and 125–500 kb regions there is no correlation and for the largest distant regions (1000 kb) there is a negative correlation. We note, however, that the difference between the 125–500 kb results and the 1000 kb result could be due to a higher signal-to noise ratio of the mSNP density measurement in the largest windows.

We propose that after TDR insertion, the two major families evolve differently. The L1 repeats and adjacent sequences could either be methylated after insertion or the repeats undergo differential selection resulting in their preferential retention in methylated regions. In our HEP data set, we see a sharp loss of L1 repeats flanking hypermethylated amplicons when we extend our flanks from 3–5 kb to 10–15 kb. This is more suggestive of a post-insertional methylation of L1 repeats and adjacent sequences that does not extend more than a few kilobases. The size range is similar to the one described for de novo retroviral insertions (Jähner et al., 1985). In contrast, differential selection would, however, be observed over 10–15 kb sequences or longer.

Observed differences in methylation of adjacent sequences between *Alu* and L1 repeats in the intermediate range (10–500 kb) might be due to differences in post-insertional selection. A negative selection against L1 insertion into gene-rich and therefore hypomethylated regions has previously been suggested to explain the observed location bias between L1 and *Alu* repeats despite a similar insertional preferences (Gasior et al., 2007). The consistent negative correlation between our germline methylation marker and *Alu* suggests positive selection for *Alu* repeats into relatively hypomethylated sequences that counters the negative effect of insertional mutation by known mechanisms such as exaptation i.e. exonization (Lev-Maor et al., 2008) and by influencing gene expression (Eller et al., 2007). This might even indicate importance of *Alu* repeats expression for the biology of the germline. Alternatively, *Alu* elements preferentially inserted into hypomethylated regions might be

relatively unaffected by purifying selection and therefore more prone to fixation within those regions.

The methylation of L1 repeats could be a subfamily-specific methylation defense mechanism decreasing transcription of L1 repeats. L1 repeats can retrotranspose in the germline (Ostertag et al., 2002) and site-specific methylation seems to be able to control their transcription (Hata et al., 1997). Of note, L1 repeats are inherited in a methylated state from male germ cells, but in an unmethylated state in female germ cells (Sanford et al., 1987). *Alu* repeats are however inherited in an unmethylated state from male germ cells, but a methylated state in female germ cells (Rubin et al., 1994). A proximal methylation pattern of sequences adjacent to *Alu* repeats could also exist, but methylation information for the female germline does not exist in the HEP data set. It is currently unknown to what degree these differences in methylation also reflect cells in the germline other than the matured germ cells.

Several notes of caution should be kept in mind. It is very difficult to determine the magnitude of the correlation between mSNP and TDR density, with the inherent imprecision of the mSNP density as the independent variable. Therefore, the biological correlation could be even stronger. Our analysis was focused on finding significance and directionality of the relationship between our markers of germline methylation and TDR density, rather than the absolute magnitude. We chose linear correlation, as it provides easy to interpret results on significance and directionality of this relationship and it provides means to appropriately correct for confounding factors. A non-linear contribution cannot be excluded, but this is more likely to affect the magnitude of the relationship. While interpreting the data we must keep in mind that the observations come from two largely unrelated data sets. While the mSNP data set has a genome-wide resolution and minimal bias, the HEP dataset is from a limited number of amplicons from three chromosomes. In addition, the amplicons of the HEP dataset are manually selected to represent a subset of the genome and are less likely to include TDRs (Eckhardt et al., 2006). However, it is not likely that hyper- and hypomethylated amplicons differ regarding any bias related to TDRs. It is also possible that a fraction of mSNPs does not represent methylation, and that correlation between the mSNP density and TDR density could be due to the correlation with a third variable. An example of direct relation to TDR would be the APOBEC3 defense system, causing increased cytosine to thymine transitions. However, this system should not show a preference for CpG dinucleotides. Finally while mSNP are correlated with TDR density, and we have previously observed a correlation between mSNP density and measurements of methylation, this does not necessarily imply a correlation between TDR density and methylation.

In summary, our data is consistent with a “Genome host defense thesis” with regards to L1 and possibly LTR elements in the male germline, as these elements are more common immediately close to hypermethylated sequences. In contrast our data does not support a role for DNA methylation in suppressing *Alu* repeats, the other major active TDR family. Similarly, DNA methylation patterns adjacent to active repeats were similar to the patterns adjacent to *Alu* repeats. Therefore it seems unlikely that DNA methylation serves as a uniform global defense mechanism against the various genome stability threats posed by different TDR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the University of Iceland Research Fund, the Icelandic Student Innovation Fund, the Landspítali University Hospital Science Fund and the Memorial Fund of Bergþóra Magnúsdóttir and Jakob Bjarnason.

Abbreviations

TDR	Transposon-derived repeats
mSNPs	methyl-associated SNPs

References

- Amariglio N, et al. Insertional mutagenesis by transposable elements in the mammalian genome. *Environ Mol Mutagen*. 1993; 21:212–218. [PubMed: 8385004]
- Belancio VP, et al. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res*. 2008; 18:343–358. [PubMed: 18256243]
- Bird A. Does DNA methylation control transposition of selfish elements in the germline? *Trends Genet*. 1997; 13:469–472. [PubMed: 9433135]
- Bjornsson HT, et al. Transposon-derived repeats in the human genome and 5-methylcytosine-associated mutations in adjacent genes. *Gene*. 2006; 370:43–50. [PubMed: 16446059]
- Bjornsson HT, et al. Intra-individual change over time in DNA methylation with familial clustering. *JAMA*. 2008; 299:2877–2883. [PubMed: 18577732]
- Bogerd HP, et al. APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. *Nucleic Acids Res*. 2006; 34:89–95. [PubMed: 16407327]
- Boissinot S, et al. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res*. 2004; 14:1221–1231. [PubMed: 15197167]
- Brookfield JF. Models of repression of transposition in P-M hybrid dysgenesis by P cytotype and by zygotically encoded repressor proteins. *Genetics*. 1991; 128:471–486. [PubMed: 1649073]
- Callinan PA, et al. Retrotransposable elements and human disease. *Genome Dyn*. 2006; 1:104–115. [PubMed: 18724056]
- Dewannieux M, et al. LINE-mediated retrotransposition of marked *Alu* sequences. *Nat Genet*. 2003; 35:41–48. [PubMed: 12897783]
- Dombroski BA, et al. Isolation of an active human transposable element. *Science*. 1991; 254:1805–1808. [PubMed: 1662412]
- Eckhardt F, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*. 2006; 38:1378–1385. [PubMed: 17072317]
- Eller CD, et al. Repetitive sequence environment distinguishes housekeeping genes. *Gene*. 2007; 390:153–165. [PubMed: 17141428]
- Feinberg AP, et al. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*. 1983; 301:89–92. [PubMed: 6185846]
- Gasior SL, et al. Characterization of pre-insertion loci of de novo L1 insertions. *Gene*. 2007; 390:190–198. [PubMed: 17067767]
- Girard A, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*. 2006; 442:199–202. [PubMed: 16751776]
- Graff JR, et al. Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation. *J Biol Chem*. 1997; 272:22322–22329. [PubMed: 9268383]
- Hata K, et al. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene*. 1997; 189:227–234. [PubMed: 9168132]
- Hedges DJ, et al. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res*. 2007; 616:46–59. [PubMed: 17157332]

- Hellmann-Blumberg U, et al. Developmental differences in methylation of human *Alu* repeats. *Mol Cell Biol*. 1993; 13:4523–4530. [PubMed: 8336699]
- Hulme AE, et al. Selective inhibition of *Alu* retrotransposition by APOBEC3G. *Gene*. 2007; 390:199–205. [PubMed: 17079095]
- International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–861. [PubMed: 17943122]
- Jähner D, et al. Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity. *Nature*. 1985; 315:594–597. [PubMed: 2989695]
- Jurka J, et al. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 2005; 110:462–467. [PubMed: 16093699]
- Karolchik D, et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res*. 2004; 32:D493–D496. [PubMed: 14681465]
- Kazazian HHJ. Mobile elements and disease. *Curr Opin Genet Dev*. 1998; 8:343–350. [PubMed: 9690999]
- Kim T, et al. Germline methylation patterns inferred from local nucleotide frequency of repetitive sequences in the human genome. *Mamm Genome*. 2007; 18:277–285. [PubMed: 17514347]
- Lander ES, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
- Lev-Maor G, et al. Intronic Alus influence alternative splicing. *PLoS Genet*. 2008; 4:e1000204. [PubMed: 18818740]
- Liang G, et al. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Mol Cell Biol*. 2002; 22:480–491. [PubMed: 11756544]
- Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009; 462:315–322. [PubMed: 19829295]
- Matsui T, et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature*. 2010; 464:927–931. [PubMed: 20164836]
- Miki Y. Retrotransposal integration of mobile genetic elements in human diseases. *J Hum Genet*. 1998; 43:77–84. [PubMed: 9621510]
- Mills RE, et al. Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet*. 2006; 78:671–679. [PubMed: 16532396]
- Myers S, et al. A fine-scale map of recombination rates and hotspots across the human genome. *Science*. 2005; 310:321–324. [PubMed: 16224025]
- Nichol K, et al. CpG methylation modifies the genetic stability of cloned repeat sequences. *Genome Res*. 2002; 12:1246–1256. [PubMed: 12176932]
- Ostertag EM, et al. A mouse model of human L1 retrotransposition. *Nat Genet*. 2002; 32:655–660. [PubMed: 12415270]
- Rubin CM, et al. *Alu* repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res*. 1994; 22:5121–5127. [PubMed: 7800508]
- Sanford JP, et al. Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse. *Genes Dev*. 1987; 1:1039–1046. [PubMed: 3428592]
- Sigurdsson MI, et al. HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res*. 2009; 19:581–589. [PubMed: 19158364]
- Turker MS. Gene silencing in mammalian cells and the spread of DNA methylation. *Oncogene*. 2002; 21:5388–5393. [PubMed: 12154401]
- Waterland RA, et al. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol*. 2003; 23:5293–5300. [PubMed: 12861015]
- Xie H, et al. SNP-based prediction of the human germ cell methylation landscape. *Genomics*. 2009; 93:434–440. [PubMed: 19442638]
- Xing J, et al. *Alu* element mutation spectra: molecular clocks and the effect of DNA methylation. *J Mol Biol*. 2004; 344:675–682. [PubMed: 15533437]

- Yoder JA, et al. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* 1997; 13:335–340. [PubMed: 9260521]
- Zamudio N, et al. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity.* 2010; 105:92–104. [PubMed: 20442734]

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10. 1016/j.gene.2011.10.046.

Fig. 1.

Observed difference (solid bar) of repeat proportion in hypermethylated and hypomethylated amplicons compared to 10,000 random assignments ($\Delta\%$) into the two groups using permutation. Repeat groups were of a) *Alu* b) L1 and c) LTR repeats within sequences flanking hypermethylated (>80% methylation) and hypomethylated (<20% methylation) amplicon groups. Using amplicons from the Human Epigenome Project sperm dataset with available methylation data, absolute proportion of TDR repeats in 5 kb flanking each amplicon was compared between the two groups. The observed difference (vertical line, -1.7%, 1.0% and 0.4% for *Alu*, L1 and LTR respectively) was compared against the distribution of $\Delta\%$ when the data was randomly assigned to two groups. The observed difference in *Alu*, L1 and LTR occurred 0, 4 and 514 times in 10,000 simulations ($p < 0.0001$, $p = 0.0004$, $p = 0.0514$) for L1, *Alu* and LTR families.

Genome-wide correlation coefficients between proportion of TDR and mSNP density for all window sizes, both absolute correlation (absolute) and partial correlation (partial) correcting for SNP density, recombination rate and proportion of GC, CpG dinucleotides and exons. Levels of statistical significance are marked by; $p < 10^{-50}$ (***), $p < 10^{-10}$ (***) and $p < 10^{-5}$ (**), with $p < 0.002$ (*) and $p < 0.002$ (*) being the lowest level of significant correlation given multiple testing.

Table 1

	125 kb		250 kb		500 kb		1000 kb	
	Absolute	Partial	Absolute	Partial	Absolute	Partial	Absolute	Partial
Total	-0.29***	-0.15***	-0.32***	-0.18***	-0.35***	-0.19***	-0.44***	-0.39***
LINE	-0.20***	-0.01	-0.23***	-0.01	-0.28***	-0.02	-0.43***	-0.18***
L1	-0.21***	-0.01	-0.25***	-0.01	-0.29***	-0.01	-0.45***	-0.17***
L2	0.09***	-0.02*	0.09***	-0.02	0.09***	-0.03	0.10**	0.03
SINE	0.02	-0.18***	0.04**	-0.21***	0.08**	-0.22***	0.20***	-0.18***
MIR	0.30***	0	0.32***	-0.01	0.34***	-0.01	0.37***	0.06
AluY	-0.16***	-0.14***	-0.16***	-0.16***	-0.12***	-0.16***	-0.03	-0.20***
AluJ	-0.06***	-0.12***	-0.04**	-0.15***	0	-0.15***	0.12**	-0.12**
AluS	-0.04**	-0.14***	-0.02	-0.17***	0.03	-0.17***	0.14***	-0.18***
LTR	0.02	-0.01	0	0	-0.02	0.03	-0.20***	-0.14***
DNA transposons	0.04**	0.01	0.05**	0.02	0.07**	0.01	0.08*	0.02
Simple repeats	0.18***	0.11***	0.22***	0.16***	0.26***	0.18***	0.25***	0.14***
Active repeats ^a	-0.10***	-0.15***	-0.08***	-0.18***	-0.03	-0.18***	0.09**	-0.21***

^a Active repeats subset comprise approximately 11,000 elements from the *Alu*, L1 and SVA TDR subfamilies that are differently present in the human and the chimpanzee genomes and thus have a greater likelihood of including active elements.

Multiple linear regression analysis of TDR subfamilies as a response to sequence and model feature predictors. Here genome-wide linear regression was performed with a window size of 1000 kb, results for other window sizes are presented in supplemental Table S1.

Table 2

	Alu repeats		L1 repeats		Active repeats	
	β^a	p-Value	β^a	p-Value	β^a	p-Value
mSNP density	-0.28	$<10^{-16}$	-0.24	$<10^{-16}$	-0.29	$<10^{-16}$
SNP density	0.11	$3 \cdot 10^{-10}$	-0.14	$3 \cdot 10^{-11}$	0.10	$8 \cdot 10^{-7}$
Exon density	0.13	$<10^{-16}$	-0.03	0.048	0.10	$2 \cdot 10^{-15}$
GC ratio	-0.17	$7 \cdot 10^{-10}$	-0.01	0.863	-0.29	$<10^{-16}$
CpG proportion	0.98	$<10^{-16}$	-0.32	$2 \cdot 10^{-15}$	1.19	$<10^{-16}$
Recombination Rate	-0.10	$3 \cdot 10^{-13}$	-0.13	$8 \cdot 10^{-13}$	-0.10	$1 \cdot 10^{-12}$
Repeats proportion ^b	-0.18	$<10^{-16}$	-0.30	$<10^{-16}$	-0.01	0.315
Model adjusted R ²		0.769		0.607		0.730

^aStandardized β values are shown. The β value is the number of standard deviations that the outcome variable will change as a result of one standard deviation change in the predictor variable.

^bProportion of repeats excluding the response TDR group for each model (L1, Alu, active repeats).

Table 3

Proportion of TDR subfamilies in the 5 kb regions flanking hypermethylated and hypomethylated amplicons. Using amplicons from the Human Epigenome Project sperm dataset, proportion of TDR was compared between the two groups (absolute numbers [bp] and percentages [%] of bases studied are given). The absolute % difference ($\Delta\%$) was calculated and compared against 10,000 permutations of the data to estimate a p-value. Statistically significant results given multiple testing ($p<0.005$) are typeset in bold.

Family	<u>Hypermethylation</u>		<u>Hypomethylation</u>		$\Delta\%$	p-Value
	bp	%	bp	%		
Total	1,860,310	18.6%	1,280,672	18.4%	0.2%	0.6862
LINE	567,614	5.7%	302,893	4.3%	1.3%	0.0001
L1	406,287	4.1%	212,281	3.0%	1.0%	0.0004
L2	136,953	1.4%	78,567	1.1%	0.2%	0.0253
SINE	708,465	7.1%	618,748	8.9%	-1.8%	<0.0001