Published in final edited form as:

Gene. 2007 April 1; 390(1-2): 153–165.

Repetitive sequence environment distinguishes housekeeping genes

C. Daniel Eller¹, Moira Regelson^{1,3}, Barry Merriman¹, Stan Nelson¹, Steve Horvath^{1,2}, and York Marahrens^{1,*}

¹ *UCLA Department of Human Genetics David Geffen School of Medicine, Gonda Center, 695 E. Young Drive South, Los Angeles, California 90095-7088, USA*

² *UCLA Department of Biostatistics, School of Public Health, Box 951772, Los Angeles, California 90095-1772, USA*

Abstract

Housekeeping genes are expressed across a wide variety of tissues. Since repetitive sequences have been reported to influence the expression of individual genes, we employed a novel approach to determine whether housekeeping genes can be distinguished from tissue-specific genes their repetitive sequence context. We show that Alu elements are more highly concentrated around housekeeping genes while various longer (>400-bp) repetitive sequences ("repeats"), including Long Interspersed Nuclear Element 1 (LINE-1) elements, are excluded from these regions. We further show that isochore membership does not distinguish housekeeping genes from tissue-specific genes and that repetitive sequence environment distinguishes housekeeping genes from tissue-specific genes in every isochore. The distinct repetitive sequence environment, in combination with other previously published sequence properties of housekeeping genes, were used to develop a method of predicting housekeeping genes on the basis of DNA sequence alone. Using expression across tissue types as a measure of success, we demonstrate that repetitive sequence environment is by far the most important sequence feature identified to date for distinguishing housekeeping genes.

Keywords

random forest; Alu; SINE; LINE; repeat; tissue-specific genes; isochores

1. INTRODUCTION

Housekeeping genes perform the basic functions common to all dividing cells; they are widely expressed across tissues, and are associated with CpG islands (Bird, 1986). Housekeeping genes also typically have small introns (Eisenberg and Levanon, 2003) that lack repetitive sequences (Han et al., 2004). Housekeeping genes have been found to cluster together on the genome to some degree (Lercher et al., 2002), and to preferentially localize to GC-rich fractions of genomic DNA known as isochores on cesium sulfate gradients (Lercher et al., 2003).

* to whom correspondence should be addressed: York Marahrens, UCLA Department of Human Genetics, Gonda Center, Room 4554b, 695 Charles E. Young Drive, Los Angeles, CA 90095, USA, Phone: (310) 267-2466, Fax: (310) 794-5446, E-mail: ymarahrens@mednet.ucla.edu.

³Current address: Moira Regelson, Senior Research Scientist, Yahoo!, 3333 West Empire Ave, Burbank CA 91504, (818) 524-3549, moira@Yahoo-inc.com

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

We were interested in the relationship of housekeeping genes to repetitive sequences. Nearly half of the human genome consists of repetitive sequence, the majority of which is transposon-derived and widely considered to be "junk DNA." The major classes of repeats are LTR retrotransposons (7.9% of genome sequence), non-LTR retrotransposons (32.0%), DNA transposons (2.8%), satellite and satellite-related sequences (0.34%), low complexity repeats (0.54%), and simple sequence repeats (0.84%). The non-LTR retrotransposons consist primarily of the Long Interspersed Nuclear Element-1 (LINE-1, 15.6%) and the non-autonomous Alu element (10.1%). LINE-1 transposons encode the enzymatic activities required for both their own mobility and for the mobilization of Alu elements (Hagan and Rudin, 2002; Kajikawa and Okada, 2002; Dewannieux et al., 2003). Alu transposons differ from other SINEs in that they are not derived from tRNA genes, but rather from the 7SL RNA gene (Ullu and Tschudi, 1984; Quentin, 1994; Smit, 1996; Okada and Hamada, 1997; Terai et al., 1998; Lander et al., 2001) which encodes the RNA component of the signal recognition particle that mediates the translocation of nascent secretory and membrane proteins (Wild et al., 2004). Aside from favoring TT|AAAA target sequences (Feng et al., 1996; Jurka, 1997; Cost and Boeke, 1998), human Alu and LINE-1 elements have been reported to insert at random positions in the genome (Smit, 1999; Boissinot et al., 2001; Lander et al., 2001; Ovchinnikov et al., 2001; Gilbert et al., 2002; Myers et al., 2002; Symer et al., 2002; Szak et al., 2002; Jurka et al., 2004; Gilbert et al., 2005). However, there is some evidence for insertional hot spots (Cost and Boeke, 1998; Myers et al., 2002; Graham and Boissinot, 2006). A popular idea is that the non-random distribution of these repetitive sequences arises from their loss via purifying selection (Boissinot et al., 2001; Myers et al., 2002; Graham and Boissinot, 2006).

There are a number of reports of repetitive sequences influencing gene expression. For example, in fragile X patients expansion mutations of a tandem simple sequence repeat located in an intron of the *FMR1* gene result in the transcriptional silencing of the *FMR1* gene (Pieretti et al., 1991). Transposable elements in *Drosophila* and plants have been implicated in the transcriptional silencing of nearby genes by the spread of heterochromatin (Lippman et al., 2004; Sun et al., 2004) raising the possibility that transposons may also be capable of reducing expression if located near genes in humans. DNA methylation is an important feature of heterochromatin that silences gene expression (Stancheva, 2005). Tissue-specific differences in DNA methylation have been reported for various repetitive sequences including LINE-1 elements (Sano and Sager, 1982; Breznik et al., 1984; Nishioka, 1988; Mietz and Kuff, 1990; Allingham-Hawkins et al., 1996; Hassan et al., 2001; Chalitchagorn et al., 2004; Khodosevich et al., 2004) raising the expectation that repetitive sequences are more repressive to the expression of nearby genes in some tissues than in others. Here we sought to determine whether the repetitive sequence environments flanking housekeeping genes that are widely expressed and important across tissues are subject to unique constraints. We show that long (>400-bp) repeats including LINE-1 elements are excluded from the regions flanking housekeeping genes and that short repeats, in particular Alu elements, are particularly highly enriched around these genes. We demonstrate that repetitive sequence environment is by far the most important sequence feature identified to date for distinguishing housekeeping genes and speculate that Alu elements are advantageous for housekeeping genes.

2. METHODS

2.1 Assembly of Gene Lists

For housekeeping genes, we used a published list of 575 genes that were expressed in all available tissues above 200 standard Affymetrix average-difference units on an Affymetrix U95A microarray chip containing 12,600 probes from 47 different human tissues and cell lines (Eisenberg and Levanon, 2003). We assembled a list of tissue-specific genes by combining the published lists of two studies (Warrington et al., 2000; Hsiao et al., 2001). Genes from all lists

were identified by either Genbank RefSeq ID or Unigene ID and were converted to RefSeq ID via DAVID Tools (<http://apps1.niaid.nih.gov/david>) (Dennis et al., 2003). We then looked up each gene by its RefSeq ID in the UCSC Genome Browser (<http://genome.ucsc.edu>) and marked it as HK (housekeeping) or TS (tissue-specific), as appropriate. Genes with common transcription start or stop positions on the same chromosome were considered to be the same gene were treated identically. Seventeen genes appeared in both the housekeeping and tissue-specific lists and were therefore excluded from both groups. After all conversions, we had 586 autosomal housekeeping genes and 468 autosomal tissue-specific genes.

2.2 Sequence Characteristics

Human sequence information including repetitive sequences were obtained from the July 2003 assembly (hg16) UCSC annotation tracks in the *chrN_rmsk* tables (<http://hgdownload.cse.ucsc.edu/downloads.html#human>). For each gene, we initially defined a region of analysis extending from 100-kb upstream of the transcription start position (txStart) to 100-kb downstream of the transcription end position (txEnd). We excluded the transcribed gene regions from our analysis to avoid effects that can be attributed to displacement by the coding sequence or splicing elements, or to the interference of transcriptional elongation by repetitive sequences in introns (Han et al., 2004). Gaps in the DNA sequence were also omitted.

The total number of base pairs comprising each repeat by family was then calculated and divided by the total number of base pairs included in the region. These data were extracted from a copy of the UCSC Genome Browser Database running locally on MySQL (<http://hgdownload.cse.ucsc.edu/downloads.html#human>; <http://www.mysql.com>) (Regelson et al., 2006). Calculations were performed using the Perl scripting language (<http://www.perl.org>) and MySQL functions.

CpG islands were treated in the same manner as repetitive sequence composition: the total number of bases comprising CpG islands was divided by total bases in each region of analysis. CpG islands are defined by the UCSC Genome Browser as sequences that are at least 50 base pairs long and have at least 50% GC content. The extent of gene clustering was estimated by counting the number of transcription start positions found within each region of analysis. Sense and antisense orientation were determined relative to the gene being analyzed; repeats oriented in the same direction as the gene are designated "sense," while genes in the opposite orientation are designated "antisense."

The long-range distribution of each characteristic was measured in much the same way as above, except flanking regions extended 40-Mb in each direction, excluding the gene, rather than 100-kb. Also, rather than working with the entire flanking region at once, we divided the 80-Mb regions into 1-Mb segments and calculated the fraction of each segment comprising repeats, CpG islands and number of genes that have transcription start sites in that segment.

Isochores are defined as contiguous regions along a chromosome sharing a homogenous GC composition and are identified by a custom track of the UCSC Genome Browser provided by IsoFinder (<http://genome.ucsc.edu>) (Oliver et al., 2004). Genes were assigned to isochores if their transcription start positions fell within the isochore's boundaries.

2.3 Statistical Analysis

Statistical analysis was performed using the software R (<http://www.R-project.org>) (Team, 2003). Means for each group were compared using the Kruskal-Wallis test. Error bars indicate 1.96x standard error (95% confidence interval).

The long-range effect of each sequence characteristic was plotted using local regression (loess) curves, as implemented in the R *loess* function (below). These plots represent a moving average

and can be adjusted with respect to the smoothness of the curve. We used a smoothness parameter of 0.02 for plots presented in Figure 1b and 0.10 for the correlation plots in Figures 4a and c, as well as for Supplementary Figure 12, online.

2.4 Random Forest Classification

A random forest predictor is an ensemble of individual classification tree predictors (Breiman, 2001). For each observation, each individual tree votes for one class and the forest predicts the class that has the plurality of votes. The user specifies the number of randomly selected variables to be searched through for the best split at each node, using the Gini index (Breiman et al., 1984) as the splitting criterion. We used the random forest package in R, which also implements partial dependence plots (Liaw and Wiener, 2002).

The root node of each tree in the forest contains a bootstrap sample of roughly 2/3 of the original data as the training set. Observations not in the training set are referred to as out-of-bag observations. For a case in the original data, the outcome is predicted by plurality vote involving only those trees that did not contain the case in their corresponding bootstrap sample. By contrasting these out-of-bag predictions with the training set outcomes, one can arrive at an estimate of the prediction error rate. Our out-of-bag error rate was 23%, which compared favorably with the proportion of housekeeping genes used in the training set (55%). Thus the input variables contain predictive information of housekeeping status.

We estimated the specificity of our random forest classifier by dividing the number of incorrectly classified genes (*i.e.* tissue-specific genes classified as housekeeping genes) by the total number of genes classified with a stringent probability (>90%).

The random forest construction allows one to define several measures of variable importance. In this article, we used the "node purity"-based variable importance measure. For each variable, it measures the mean decrease in Gini index over all node splits that involve it. The absolute values of the variable importance measure have no meaning; instead, this measure is used to rank variables.

2.5 Affymetrix Microarray Atlases

Two independent microarray datasets were used. Gene Atlas #1 was by the authors (A.D., B.M. and S.N., submitted). Gene Atlas #2 was provided by the Genomics Institute of the Novartis Research Foundation (<http://symatlas.gnf.org>) (Su et al., 2004). Conversion from RefSeq gene ID to Affymetrix probe ID was performed using the DAVID website (<http://apps1.niaid.nih.gov/david>).

We considered a gene to be expressed in a tissue if its presence call p-value was below 0.05. Gene Atlas #1 presence call p-values were generated with the software MAS 5.0 (<http://www.affymetrix.com/products/software/specific/mas.affx>). Gene Atlas #2 presence call p-values were generated with the software DNA-Chip Analyzer (dChip) (<http://biosun1.harvard.edu/complab/dchip>).

3. RESULTS

3.1 Differences in flanking sequence composition between housekeeping and tissue-specific genes

We chose for our analysis 1000 randomly selected autosomal Reference sequence (RefSeq) genes, published lists of 583 housekeeping genes (Eisenberg and Levanon, 2003), and 468 tissue-specific genes (Warrington et al., 2000; Hsiao et al., 2001). For each gene, we initially considered a region extending 100-kb upstream from the transcription start and 100-kb

downstream from the transcription end, but excluding the transcribed gene region. We excluded the transcribed regions from our analysis to avoid effects attributable to displacement by the coding sequence or splicing elements, or to the interference of transcriptional elongation by repetitive sequences in introns (Han et al., 2004). For the 200-kb region flanking each gene, we identified the types and positions of repeats from the RepeatMasker output provided by the UCSC genome browser (<http://genome.ucsc.edu>) (Supplementary Table 1 online). We then obtained a value for each repeat type representing the percent of the 200-kb flanking sequence occupied by that repeat type. We also considered the percent CpG island sequence, the size of the each gene, and the number of neighboring genes (gene clustering) whose transcription start position fell within the 200-kb regions.

Housekeeping (HK) genes contained more genes in their 200-kb flanking regions and were flanked by significantly higher concentrations of CpG island sequence than tissue-specific (TS) genes or the random sample of genes (Fig. 1a), in accordance with previously published reports (Bird, 1986; Lercher et al., 2002). HK genes were also flanked by significantly higher concentrations of Alu elements (Fig. 1a) and this was consistent across all subtypes (Supplementary Fig. 6a online) except for four of the five youngest subtypes examined; the only one of the youngest Alu subtypes that was significantly more enriched around HK genes (Ya8) displayed a more modest enrichment than the older subtypes (Supplementary Fig. 6b online). In addition, HK genes were flanked by significantly lower concentrations of the repeats CR1, LINE-1, MaLR, MER1-type, AcHobo, DNA, ERV1, ERVL, LINE-2, Mariner, MER2-type, MIR and Tip100 elements (Fig. 1a and Supplementary Fig. 7 online). These effects were G independent of gene clustering (Supplementary Table 2 online) and also independent of the orientation of the repeat with respect to the gene (Supplementary Fig. 8 online). A trend was that shorter (<400-bp) transposons were significantly more enriched around housekeeping genes ($p < 1 \times 10^{-50}$) while longer (>400-bp) transposons were significantly more abundant around tissue-specific genes ($p = 1.6 \times 10^{-32}$). The correlation of short repeats appeared to be almost entirely due to the dominant Alu element. However, the correlation for long repeats was more significant than the corresponding correlations involving exclusively the most significant, and most abundant long repeat, LINE-1 ($p = 3.4 \times 10^{-31}$, Fig. 1a).

3.2 Comparison of housekeeping and tissue-specific genes among isochores

Isochores are defined as long stretches of a chromosome that exhibit a more or less homogenous GC composition. There are currently five commonly recognized isochores, identified by their relative GC content (low 1 or 2; high 1, 2 or 3) (<http://genome.ucsc.edu>) (Oliver et al., 2004). To determine whether isochore membership distinguishes HK from TS genes, we determined the isochore assignment to which each HK and TS gene belonged. Both HK and TS genes resided in the high-GC isochores (H1, H2 and H3) and in the low-GC isochores (L1 and L2) (Table 1) indicating that isochore membership did not distinguish HK from TS genes in agreement with previous reports (D'Onofrio, 2002). However, there were very few HK genes in isochore L1 (Fig. 2a). We then compared the repeat environment of HK and TS genes in each isochore. Although some repeats were significantly enriched or depleted according to the GC content of the isochore, (Supplementary Fig. 9 online), HK genes could be distinguished from TS genes by significant differences in repeat environment in each of the five isochores (Fig. 2). In several instances the expected trends were clearly apparent but, statistical significance was not achieved due to the small sample sizes that resulted from splitting the HK and TS gene samples into five groups, particularly when considering isochore L1. An oddity was the repeat CR1 that was most abundant in L1 and H3 where it distinguished HK from TS genes with opposite trends. Short (<400-bp) repeats were significantly more abundant around HK versus TS genes across all isochores except L1 while longer repeats were more abundant around TS genes in most isochores (Fig. 2b). We confirm that isochore membership influences repetitive sequence environment but fails to distinguish HK from TS genes (D'Onofrio,

2002). Repeat environment and not isochore membership therefore distinguishes HK from TS genes.

3.3 Sequence composition differences extend over megabase distances

To determine how far along the chromosome the flanking sequence differences between housekeeping and tissue-specific genes extended, we calculated the density of each repeat for 1-Mb intervals extending 40-Mb upstream and downstream from the three sets of genes. Alu elements were significantly more enriched around housekeeping genes than tissue-specific genes across an 18-Mb region (Fig. 1b, arrows). The regions over which repeat densities were reduced around housekeeping genes relative to tissue-specific genes ranged from 3-Mb (MER1-type) to 20-Mb (LINE-1) (Fig. 1b, arrows).

Visual inspection revealed that most of the trends that distinguish housekeeping genes are also evident around tissue-specific genes, but to a lesser extent. Nevertheless, most repeats displayed elevated or reduced densities over similar distances around tissue-specific genes as housekeeping genes (Fig. 1b). The exceptions were the CR1 and MER1-type elements, whose densities were reduced over 10-Mb regions around housekeeping genes but were not significantly reduced around tissue-specific genes (Fig. 1b).

3.4 Using repetitive sequence environment to identify housekeeping genes

The finding that repetitive sequence environment distinguishes housekeeping genes suggested that it may be possible to identify housekeeping genes based on DNA sequence features alone. We constructed a random forest classifier of housekeeping status using repeats, CpG islands, gene size and gene clustering as input variables. The out-of-bag accuracy ($1 - \text{error}$) was 77%, which compares favorably with the accuracy of 45% if all genes in the training set had been called tissue-specific. However, only a small proportion of the published set of housekeeping genes was successfully classified (Supplementary Table 3 online). We evaluated the importance of each characteristic in the random forest classification by removing them one at a time and recreating the classifier. Removal of CpG islands, gene size or clustering effects had no significant impact on the accuracy of the classifier, whereas removal of repeat information significantly increased the classifier's error rate (Fig. 3a). We therefore conclude that repeat environment is the most important DNA sequence feature we tested for predicting housekeeping genes. We also calculated the relative importance of each individual repeat in the classifier and found Alu to be the most important repeat, followed by LINE-1 (Fig. 3b).

3.5 Validation of the Random Forest prediction using microarray data

We applied the intact random forest classifier to more than 16,000 RefSeq genes and identified >800 genes with housekeeping gene prediction probabilities ("HK probability") greater than 80%. If these high-scoring genes are indeed housekeepers, they should be expressed across tissues. To test this, we calculated the proportion of tissues in which each gene is reported to be expressed in two independent Affymetrix microarray atlases. Genes scoring over 80% HK probability had markedly higher expression levels than those with lower scores (Fig. 4 and Supplementary Fig. 10 online). As expected, the published set of housekeeping genes was widely expressed across tissues while the tissue-specific genes were expressed in low numbers of tissues (Supplementary Fig. 11 online). Of the >800 candidate housekeeping genes, 489 were expressed in more than 80% of the tissues examined (Supplementary Table 4 online).

3.6 Alu concentration around a gene correlates with percent of tissues in which the gene is expressed

A gradual increase in the predicted probability of being a housekeeping gene coincided with a corresponding gradual increase in the number of tissues in which those genes were expressed

(Fig. 5a). This raised the possibility that wide gradients of one or more sequence features may correspond with the entire range of HK probabilities. To investigate this, we used partial dependence plots (Friedman, 2001) to track the relationship between the density of a sequence feature and its individual contribution to the overall HK probability. Modest increases in gene size, gene clustering and CpG island density were accompanied by sharp changes in contribution of these features to the HK probability (Fig. 5b and Supplementary Fig. 12 online), suggesting threshold values for these features. In contrast, a wide range of Alu concentration (0–30%) was required to achieve a 60% change in contribution to the HK probability (Fig. 5b). This suggested that Alu concentration may correspond to breadth of gene expression. A comparison of gene expression to Alu concentration revealed that Alu elements are indeed progressively more abundant in flanking regions as one moves from narrowly expressed genes to ubiquitously expressed genes (Fig. 5c).

4. DISCUSSION

4.1 Housekeeping genes are distinguished by a distinct repetitive sequence environment

Repetitive sequences have been shown to silence nearby genes via the spread of heterochromatin (Lippman et al., 2004; Sun et al., 2004). However, several longer repetitive sequences or tracts of repetitive sequences have been reported to show differences in an important heterochromatin property, DNA methylation (Sano and Sager, 1982; Breznik et al., 1984; Nishioka, 1988; Mietz and Kuff, 1990; Allingham-Hawkins et al., 1996; Hassan et al., 2001; Chalitchagorn et al., 2004; Khodosevich et al., 2004), suggesting that such influences should vary with tissue type. We therefore reasoned that housekeeping genes, which are utilized across all or nearly all tissues, ought to be subject to stronger constraints in the repetitive sequence environment of their flanking regions. We show that housekeeping genes indeed reside in a distinct repetitive sequence environment that distinguishes these genes more accurately than their previously identified sequence characteristics: association with CpG islands, gene clustering and the presence of small introns. Alu elements were enriched in large regions around housekeeping genes but peaked sharply near the genes themselves. Most other repetitive sequences, including LINE-1 elements, showed the inverse pattern: partial exclusion over large regions with a sharp drop in concentration in the immediate vicinity of the genes. Most repeats showed similar but significantly less pronounced trends around the less widely expressed tissue-specific genes. Interestingly, the most significant correlations were the enrichment of all short (<400-bp) transposons around HK genes and the exclusion of all long (>400-bp) transposons or repeat tracts from these regions.

Though isochore membership influenced repeat abundance, both HK and TS genes were distributed across isochores, and repeat environment distinguished HK and TS genes in every isochore. Isochore membership, therefore, failed to distinguish the two types of genes. Several recent studies have concluded that isochores, while excellent for broad descriptions of chromosomal properties, lack the resolution necessary to analyze subtle influences on gene expression (Häring and Kypr, 2001; Lander et al., 2001; Cohen et al., 2005). Alu enrichment was only 16% correlated with housekeeping genes, but even this modest correlation was highly significant. Clearly, there are one or more processes at work that cause housekeeping genes to acquire a distinctive repetitive sequence environment. Therefore, we find it prudent to consider all of the many quantifiable elements that contribute to a gene's sequence environment.

4.2 The possibility that non-random insertion patterns contribute to the distinct repeat environment of housekeeping genes

Why are Alu elements abundant around housekeeping genes while long repeats are scarce? One possibility is that Alu elements preferentially insert around housekeeping genes while longer transposons including LINE-1 elements preferentially insert elsewhere. The notion that

human transposons preferentially integrate in certain regions gains support from evidence for insertion hotspots. For example, it was reported that nine of 14 disease-causing LINE-1 insertions were restricted to only three genes (Ostertag and Kazazian, 2001). Alu elements might preferentially recognize the chromatin near housekeeping genes while LINE-1 elements and other long transposons avoid such chromatin. However, the notion of strikingly distinct insertion biases seems contrary to the finding that the LINE-1 transposition machinery mobilizes both LINE-1 and Alu elements (Jurka, 1997; Dewannieux et al., 2003) which consequently insert at the same consensus TT/AAAA sequence (Feng et al., 1996; Jurka, 1997; Cost and Boeke, 1998). Indeed, the factor IX gene has been shown to be targeted by one disease-causing LINE-1 insertion and two independent Alu insertions (Ostertag and Kazazian, 2001) suggesting that Alu and LINE-1 elements can share an insertion hotspot. The notion that Alu and LINE-1 elements have different insertion biases is also inconsistent with the finding that evolutionarily recent insertions of active Alu and LINE-1 subfamilies do not follow the non-random genomic distribution of the older elements (Feng et al., 1996; Ovchinnikov et al., 2001; Gilbert et al., 2002; Symer et al., 2002; Szak et al., 2002; Gilbert et al., 2005). Indeed, we show that all older (>2.4 myr) Alu subfamilies examined are significantly more abundant around housekeeping genes than tissue-specific genes while this trend is weaker or not evident among the youngest subfamilies. The insertion bias scenario would therefore also require all young Alu and LINE-1 subfamilies to have lost their contrasting regional insertion preferences.

4.3 Non-random repeat distributions via natural selection

Another explanation is that natural selection rather than insertion bias is responsible for the non-random repeat distribution. Selection may simply involve new insertions being lost when they are deleterious. In this case longer (>400-bp) repeats would be disadvantageous when located near housekeeping genes but not detrimental in gene poor regions. In contrast, Alu elements would not be deleterious near housekeeping genes, but disadvantageous when too abundant around tissue-specific genes and most detrimental when concentrated in gene-poor regions. This selection scenario allows us to avoid invoking a mechanism whereby the LINE-1 transposition machinery leads to Alu elements being inserted in different regions from LINE-1 elements. It would also account for why all of the youngest Alu and LINE-1 subfamilies are more randomly distributed than the older subfamilies as not enough time would have passed to select for the favored distributions (Gu et al., 2000; Pavlíček et al., 2001; Medstrand et al., 2002; Belle et al., 2005; Hackenberg et al., 2005).

How might high Alu concentrations be increasingly detrimental as one moves away from housekeeping genes? We show that this decrease in Alu concentration is accompanied by an increase in long repeats and repeat tracts, the most abundant of which are LINE-1 elements. Long repeats including LINE-1 elements are normally heavily DNA-methylated (Woodcock et al., 1988; Crowther et al., 1991; Woodcock et al., 1997), a feature of heterochromatin. DNA methylation (and heterochromatin in general) has been reported to suppress homologous recombination (Páldi et al., 1995; Maloisel and Rossignol, 1998; Schnable et al., 1998; Fu et al., 2002; Yao et al., 2002; Yamada et al., 2004; Myers et al., 2005) as well as transposition (Yoder et al., 1997; Walsh et al., 1998; Hirochika et al., 2000; Robertson, 2001; Bird, 2002; Kato et al., 2003). Accordingly, reports of deletions caused by homologous recombination between LINE-1 elements are rare (Segal et al., 1999) except in cancers (Florl and Schulz, 2003) where LINE-1 elements are frequently hypomethylated (Santourlidis et al., 1999; Takai et al., 2000; Ehrlich, 2002; Carnell and Goodman, 2003; Florl et al., 2004; Roman-Gomez et al., 2005). LINE-1 transposition has also been reported to be elevated in cancers (Schulz, 2006). If high concentrations of Alu elements cause nearby LINE-1 elements and other repeats to be hypomethylated (or lose other heterochromatic features), the expected result would be genome instability due to an increase in illegitimate homologous recombination and an increase in transposition.

Studies on a small number of Alu elements describe properties that, if widespread among Alu elements, suggest a general mechanism for the genome destabilization scenario. In an impressively thorough pair of studies (Thorey et al., 1993; Willoughby et al., 2000), one of three Alu elements was shown to protect transgenes against transcriptional repression by position effects and also by being integrated as tandem multicopy repeats (Garrick et al., 1998; Selker, 1999). This led to the suggestion that a subset of all Alu elements define transcriptionally permissive (euchromatic) domains (Willoughby et al., 2000). A substantial proportion of Alu elements are hypomethylated (Schmid, 1998; Yang et al., 2004). In one study, it was shown that seven out of 21 Alu elements examined displayed properties of euchromatin (H3-K4 methylation and histone acetylation) and furthermore were bound by a SNF2-containing chromatin remodeling complex (Hakimi et al., 2002). Treatment of cells with the DNA methyltransferase inhibitor 5-azacytidine resulted in additional Alu elements being bound by the chromatin remodeling complex (Hakimi et al., 2002). Unfortunately, it was not addressed whether the same Alu elements are consistently hypomethylated or whether Alu elements slowly alternate between heterochromatic (e.g., DNA methylated) and euchromatic (e.g., DNA hypomethylated) states as has been documented for a mouse IAP element near a gene (Whitelaw and Martin, 2001) and a proviral reporter (Lorincz et al., 2002). We favor alternation between heterochromatic and euchromatic Alu states in part because the same Alu element has been shown to be associated with both heterochromatic and euchromatic histone modifications (Kondo et al., 2004). An Alu element has also been shown to display variability in its histone modifications in mice (Martens et al., 2005).

If a significant proportion of all Alu elements are euchromatic, a consequence should be illegitimate recombination-induced genome instability that is limited by the short lengths of the repeats but possibly increased by the presence of a chi-like recombination sequence (Lupski, 2004). Indeed, there are numerous reports of disease-causing deletions resulting from recombination between Alu elements (Batzer and Deininger, 2002; Nishimura et al., 2005; Casarin et al., 2006; Has et al., 2006; Kozak et al., 2006; Li et al., 2006; Matejas et al., 2006; Nissen et al., 2006; Sen et al., 2006; Shabbeer et al., 2006; Uddin et al., 2006; Xie et al., 2006; Zhang et al., 2006). If a large proportion of Alu elements indeed fosters euchromatic domains as suggested by the aforementioned study (Willoughby et al., 2000), then flanking sequences may also be destabilized. Interestingly, high concentrations of Alu elements have been associated with disease-causing deletions whose breakpoints were not in the Alu elements themselves (Abrao et al., 2006; Abu-Safieh et al., 2006). LINE-1 elements located among high concentrations of such Alu elements might be rendered euchromatic. Since efficiency of illegitimate homologous recombination increases with length of homology (Waldman and Liskay, 1988; Baker et al., 1996), this would engender genome instability via illegitimate homologous recombination and possibly also via transposition if the element is suitably intact (Baker et al., 1996; Ostertag and Kazazian, 2001; Robertson, 2001). We would expect the genome instability engendered by euchromatic LINE-1 elements to result in these elements' being selected against.

4.4 Positive selection may enrich Alu elements near genes

Why are Alu elements more enriched around housekeeping genes and less so around other genes? One possibility is that nearby Alu elements increase gene expression levels under normal circumstances. There are several reports of variant Alu sequences functioning as tissue-specific enhancers for genes (for a partial list, see <http://zmbe.uni-muenster.de/expath/alltables.htm>) (Britten, 1996). However, Alu elements might also aid the expression of the more widely expressed housekeeping genes. The Alu element shown to protect transgenes (Thorey et al., 1993; Willoughby et al., 2000) was proposed to support open chromatin barriers that acted as barriers against the spread of heterochromatin. A mutational analysis implicated conserved elements in the internal Pol III

promotor in the protective function of this Alu (Thorey et al., 1993). Pol III promoters do not display tissue-specificity (Lander et al., 2001) and have been shown to also function as barriers that protect genes from the spread of heterochromatin in *S. cerevisiae* (Donze et al., 1999) and *S. pombe* (Noma et al., 2006; Scott et al., 2006). In *S. cerevisiae*, Pol III promotor barrier function has been shown to require histone acetyltransferases that are chromatin remodeling enzymes that foster the formation of euchromatin (Donze and Kamakaka, 2002). Other heterochromatin barriers are also believed to serve as entry sites for the recruitment of histone acetyltransferase or chromatin remodeling activities that disrupt the binding of silencing proteins to histones and thereby terminate the spread of heterochromatin structures (Grewal and Moazed, 2003). Pol III heterochromatin barriers in yeast have been shown to require subunits of the cohesin complex (Donze and Kamakaka, 2002) and the Alu elements reported to be bound by a SNF2h-containing chromatin remodeling complex also recruited cohesin subunits (Hakimi et al., 2002). Since transgenes tend to become partially or completely silenced over time by repressive chromatin (Pannell and Ellis, 2001; Iba et al., 2003; Kwaks and Otte, 2006; Lavigne and Gorecki, 2006), most chromosomal regions appear to be repressive to gene function. Such repressiveness would select for the appearance and retention of elements that facilitate the expression of housekeeping genes. We therefore speculate that Alu elements are not only tolerated by housekeeping genes due to the scarcity of long repeats, but that they are more abundant around the highly expressed housekeeping genes than other genes because they promote gene expression across tissues.

Another possible pressure driving the greater enrichment of Alu elements around housekeeping genes arises from the observation that in a number of species, stress has been shown to trigger Alu transcription resulting in Alu transcripts that bind the protein kinase PKR; this blocks the ability of PKR to inhibit protein translation (Chu et al., 1998; Schmid, 1998; Deininger and Batzer, 1999). The concentration of Alu elements in the euchromatic environment around genes may facilitate this. Since housekeeping genes provide a more consistent euchromatic environment than other genes, the positive selective pressure may be stronger for housekeeping genes.

4.8 Long repeats may be disadvantageous to nearby housekeeping genes

Regardless of whether Alu elements are beneficial or merely tolerated, the increased abundance of Alu elements around housekeeping genes stood in stark contrast to the scarcity of longer (>400-bp) repeats and repeat tracts (LINE-1 elements and various other repeats) in these same regions. Although, the lower abundance of the TT|AAAA target sequence near housekeeping genes may very well contribute to LINE-1 scarcity (Jurka, 1997; Cost and Boeke, 1998; Lander et al., 2001; Graham and Boissinot, 2006) despite their otherwise random insertion pattern (Smit, 1999; Boissinot et al., 2001; Lander et al., 2001; Ovchinnikov et al., 2001; Gilbert et al., 2002; Myers et al., 2002; Symer et al., 2002; Szak et al., 2002; Jurka et al., 2004; Gilbert et al., 2005), at least one additional explanation is needed since long repeats in general were scarce. One reason why long repeats might be selected against near housekeeping genes is that an abundance of these repeats might reduce gene expression via heterochromatin spread. Long transposons (Lyon, 1998; Marahrens, 1999; Bailey et al., 2000; Allen et al., 2003; Lippman et al., 2004; Sun et al., 2004) and long tracts of tandem repeats (Pieretti et al., 1991; Hansen et al., 1997; Saveliev et al., 2003) have been implicated in gene silencing via the spread of heterochromatin. Another reason why LINE-1 and other long transposons might be scarce around housekeeping genes is that the euchromatin could spread into the transposons and activate their internal promoters (Swergold, 1990; Minakami et al., 1992; Leib-Mösch and Seifarth, 1995; Speek, 2001; Athanikar et al., 2004). This could cause gene over-expression as has been reported for and IAP element insertion in the 5' upstream region of the mouse *Agouti* gene (Whitelaw and Martin, 2001) or down-regulate genes if the transcription proceeds through the gene in the antisense direction (Whitelaw and Martin, 2001). Promotor activation

in a transposon could also cause transposition if the element is suitably intact. Finally, the encroachment of euchromatin into long repeats would facilitate illegitimate homologous recombination between repeats.

Long repeats were much more abundant around tissue-specific genes than housekeeping genes. Tissue-specific genes are known to employ tissue-specific enhancers and LCRs to open the chromatin for transcription. Long repeat content might help prevent unwanted expression in all other tissues and might overwhelm the Alu elements near tissue-specific genes. However, in tissues where the gene is to be expressed, Alu elements might interact with tissue-specific promoter or enhancer elements to modulate gene expression, as has been reported for the *K18* gene (Rhodes and Oshima, 1998; Willoughby et al., 2000). Note that although short repeats (other than demethylated Alu elements) also display properties of heterochromatin, there is no evidence that heterochromatin can spread more than a few bases beyond the repeat. Indeed, in plants a polymorphic SINE element has been reported to serve as a nucleation center for DNA methylation but the methylation only spread a few bases beyond the SINE (Arnaud et al., 2000). Limited spread of methylation would provide an additional explanation for why Alu elements are tolerated the vicinity of widely expressed genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

WEB SITE REFERENCES

1. <http://apps1.niaid.nih.gov/david/>; DAVID Tools
2. <http://genome.ucsc.edu/>; UCSC Genome Browser
3. <http://hgdownload.cse.ucsc.edu/downloads.html#human>; UCSC Genome Browser downloads
4. <http://www.mysql.com/>; MySQL database software
5. <http://www.perl.org/>; Perl scripting language
6. <http://www.R-project.org/>; The R Project for Statistical Computing
7. <http://symatlas.gnf.org/>; Gene Atlas #2 (Microarray data)
8. <http://www.affymetrix.com/products/software/specific/mas.affx>; MAS (Affymetrix Microarray Analysis Software)
9. <http://biosun1.harvard.edu/complab/dchip/>; dChip (DNA chip Analyzer)
10. <http://zmbe.uni-muenster.de/expath/alltables.htm> (Tables of Retronuons)

Acknowledgements

C.D.E. was supported by a UCLA-IGERT bioinformatics traineeship (NSF DGE-9987641). M.R. was supported by a Tumor Cell Biology Fellowship (USHHS Institutional National Research Service Award #T32 CA09056). Y.M. was supported in part by National Institutes of Health Grants GM6100701 and HD041451-02.

References

- Abrao MG, Leite MV, Carvalho LR, Billerbeck AE, Nishi MY, Barbosa AS, Martin RM, Arnhold IJ, Mendonca BB. Combined pituitary hormone deficiency (CPHD) due to a complete PROP1 deletion. *Clin Endocrinol (Oxf)* 2006;65:294–300. [PubMed: 16918947]

- Abu-Safieh L, Vithana EN, Mantel I, Holder GE, Pelosini L, Bird AC, Bhattacharya SS. A large deletion in the adRP gene PRPF31: evidence that haploinsufficiency is the cause of disease. *Mol Vis* 2006;12:384–8. [PubMed: 16636657]
- Allen E, Horvath S, Tong F, Kraft P, Spiteri E, Riggs AD, Marahrens Y. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc Natl Acad Sci U S A* 2003;100:9940–5. [PubMed: 12909712]
- Allingham-Hawkins DJ, Brown CA, Babul R, Chitayat D, Krekewich K, Humphries T, Ray PN, Teshima IE. Tissue-specific methylation differences and cognitive function in fragile X premutation females. *Am J Med Genet* 1996;64:329–33. [PubMed: 8844075]
- Arnaud P, Goubely C, Pelissier T, Deragon JM. SINE retroposons can be used in vivo as nucleation centers for de novo methylation. *Mol Cell Biol* 2000;20:3434–41. [PubMed: 10779333]
- Athanikar JN, Badge RM, Moran JV. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* 2004;32:3846–55. [PubMed: 15272086]
- Bailey JA, Carrel L, Chakravarti A, Eichler EE. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A* 2000;97:6634–9. [PubMed: 10841562]
- Baker MD, Read LR, Beatty BG, Ng P. Requirements for ectopic homologous recombination in mammalian somatic cells. *Mol Cell Biol* 1996;16:7122–32. [PubMed: 8943368]
- Batzler MA, Deininger PL. Alu repeats and human genomic diversity. *Nat Rev Genet* 2002;3:370–9. [PubMed: 11988762]
- Belle EM, Webster MT, Eyre-Walker A. Why are young and old repetitive elements distributed differently in the human genome? *J Mol Evol* 2005;60:290–6. [PubMed: 15871040]
- Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;16:6–21. [PubMed: 11782440]
- Bird AP. CpG-rich islands and the function of DNA methylation. *Nature* 1986;321:209–13. [PubMed: 2423876]
- Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* 2001;18:926–35. [PubMed: 11371580]
- Breiman L. Random forests. *Machine Learning* 2001;45:5–32.
- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. *Classification and Regression Trees*. 1984.
- Breznik T, Traina-Dorge V, Gama-Sosa M, Gehrke CW, Ehrlich M, Medina D, Butel JS, Cohen JC. Mouse mammary tumor virus DNA methylation: tissue-specific variation. *Virology* 1984;136:69–77. [PubMed: 6330997]
- Britten RJ. DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A* 1996;93:9374–S7. [PubMed: 8790336]
- Carnell AN, Goodman JJ. The long (LINEs) and the short (SINEs) of it: altered methylation as a precursor to toxicity. *Toxicol Sci* 2003;75:229–35. [PubMed: 12773759]
- Casarin A, Martella M, Polli R, Leonardi E, Anesi L, Murgia A. Molecular characterization of large deletions in the von Hippel-Lindau (VHL) gene by quantitative real-time PCR: the hypothesis of an alu-mediated mechanism underlying VHL gene rearrangements. *Mol Diagn Ther* 2006;10:243–9. [PubMed: 16884328]
- Chalitchagorn K, Shuangshoti S, Hourpai N, Kongruttanachok N, Tangkijvanich P, Thong-ngam D, Voravud N, Sriuranpong V, Mutirangura A. Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene* 2004;23:8841–6. [PubMed: 15480421]
- Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW. Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol Cell Biol* 1998;18:58–68. [PubMed: 9418853]
- Cohen N, Dagan T, Stone L, Graur D. GC composition of the human genome: in search of isochores. *Mol Biol Evol* 2005;22:1260–72. [PubMed: 15728737]
- Cost GJ, Boeke JD. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* 1998;37:18081–93. [PubMed: 9922177]

- Crowther PJ, Doherty JP, Linsenmeyer ME, Williamson MR, Woodcock DM. Revised genomic consensus for the hypermethylated CpG island region of the human L1 transposon and integration sites of full length L1 elements from recombinant clones made using methylation-tolerant host strains. *Nucleic Acids Res* 1991;19:2395–401. [PubMed: 1710354]
- Deininger PL, Batzer MA. Alu repeats and human disease. *Mol Genet Metab* 1999;67:183–93. [PubMed: 10381326]
- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003;4:P3. [PubMed: 12734009]
- Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 2003;35:41–48. [PubMed: 12897783]
- D’Onofrio G. Expression patterns and gene distribution in the human genome. *Gene* 2002;300:155–60. [PubMed: 12468096]
- Donze D, Adams CR, Rine J, Kamakaka RT. The boundaries of the silenced HMR domain in *Saccharomyces cerevisiae*. *Genes Dev* 1999;13:698–708. [PubMed: 10090726]
- Donze D, Kamakaka RT. Braking the silence: how heterochromatic gene repression is stopped in its tracks. *Bioessays* 2002;24:344–9. [PubMed: 11948620]
- Ehrlich M. DNA hypomethylation, cancer, the immunodeficiency, centromeric region instability, facial anomalies syndrome and chromosomal rearrangements. *J Nutr* 2002;132:2424S–2429S. [PubMed: 12163705]
- Eisenberg E, Levanon EY. Human Housekeeping genes are compact. *Trends in Genetics* 2003;19:362–365. [PubMed: 12850439]
- Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 1996;87:905–16. [PubMed: 8945517]
- Flori AR, Schulz WA. Peculiar structure and location of 9p21 homozygous deletion breakpoints in human cancer cells. *Genes Chromosomes Cancer* 2003;37:141–8. [PubMed: 12696062]
- Flori AR, Steinhoff C, Muller M, Seifert HH, Hader C, Engers R, Ackermann R, Schulz WA. Coordinate hypermethylation at specific genes in prostate carcinoma precedes LINE-1 hypomethylation. *Br J Cancer* 2004;91:985–94. [PubMed: 15292941]
- Friedman J. Greedy function approximation: the gradient boosting machine. *Annals of Statistics* 2001;29:1189–1232.
- Fu H, Zheng Z, Dooner HK. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc Natl Acad Sci U S A* 2002;99:1082–7. [PubMed: 11792865]
- Garrick D, Fiering S, Martin DI, Whitelaw E. Repeat-induced gene silencing in mammals. *Nat Genet* 1998;18:56–9. [PubMed: 9425901]
- Gilbert N, Lutz S, Morrish TA, Moran JV. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* 2005;25:7780–95. [PubMed: 16107723]
- Gilbert N, Lutz-Prigge S, Moran JV. Genomic deletions created upon LINE-1 retrotransposition. *Cell* 2002;110:315–25. [PubMed: 12176319]
- Graham T, Boissinot S. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *J Biomed Biotechnol*. 2006;2006;(Art75327)
- Grewal SI, Moazed D. Heterochromatin and epigenetic control of gene expression. *Science* 2003;301:798–802. [PubMed: 12907790]
- Gu Z, Wang H, Nekrutenko A, Li WH. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* 2000;259:81–8. [PubMed: 11163965]
- Hackenberg M, Bernaola-Galvan P, Carpena P, Oliver JL. The biased distribution of Alus in human isochores might be driven by recombination. *J Mol Evol* 2005;60:365–77. [PubMed: 15871047]
- Hagan CR, Rudin CM. Mobile genetic element activation and genotoxic cancer therapy: potential clinical implications. *Am J Pharmacogenomics* 2002;2:25–35. [PubMed: 12083952]
- Hakimi MA, Bochar DA, Schmiesing JA, Dong Y, Barak OG, Speicher DW, Yokomori K, Shiekhatter R. A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature* 2002;418:994–998. [PubMed: 12198550]

- Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 2004;429:268–274. [PubMed: 15152245]
- Hansen RS, Canfield TK, Fjeld AD, Mumm S, Laird CD, Gartler SM. A variable domain of delayed replication in FRAXA fragile X chromosomes: X inactivation-like spread of late replication. *PNAS* 1997;94:4587–4592. [PubMed: 9114034]
- Häring D, Kypr J. No isochores in the human chromosomes 21 and 22? *Biochem Biophys Res Commun* 2001;280:567–73. [PubMed: 11162557]
- Has C, et al. Molecular basis of Kindler syndrome in Italy: novel and recurrent Alu/Alu recombination, splice site, nonsense, and frameshift mutations in the KIND1 gene. *J Invest Dermatol* 2006;126:1776–83. [PubMed: 16675959]
- Hassan KM, Norwood T, Gimelli G, Gartler SM, Hansen RS. Satellite 2 methylation patterns in normal and ICF syndrome cells and association of hypomethylation with advanced replication. *Hum Genet* 2001;109:452–62. [PubMed: 11702227]
- Hirochika H, Okamoto H, Kakutani T. Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. *Plant Cell* 2000;12:357–69. [PubMed: 10715322]
- Hsiao LL, et al. A compendium of gene expression in normal human tissues. *Physiol Genomics* 2001;7:97–104. [PubMed: 11773596]
- Iba H, Mizutani T, Ito T. SWI/SNF chromatin remodelling complex and retroviral gene silencing. *Rev Med Virol* 2003;13:99–110. [PubMed: 12627393]
- Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retrotransposons. *Proc Natl Acad Sci U S A* 1997;94:1872–7. [PubMed: 9050872]
- Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* 2004;101:1268–72. [PubMed: 14736919]
- Kajikawa M, Okada N. LINEs mobilize SINES in the eel through a shared 3' sequence. *Cell* 2002;111:433–44. [PubMed: 12419252]
- Kato M, Miura A, Bender J, Jacobsen SE, Kakutani T. Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. *Curr Biol* 2003;13:421–6. [PubMed: 12620192]
- Khodosevich KV, Lebedev Iu B, Sverdlov ED. [The tissue-specific methylation of human-specific endogenous retroviral long terminal repeats]. *Bioorg Khim* 2004;30:493–8. [PubMed: 15562970]
- Kondo Y, Shen L, Yan PS, Huang TH, Issa JP. Chromatin immunoprecipitation microarrays for identification of genes silenced by histone H3 lysine 9 methylation. *Proc Natl Acad Sci U S A* 2004;101:7398–403. [PubMed: 15123805]
- Kozak L, Hrabincova E, Kintr J, Horky O, Zapletalova P, Blahakova I, Mejstrik P, Prochazkova D. Identification and characterization of large deletions in the phenylalanine hydroxylase (PAH) gene by MLPA: Evidence for both homologous and non-homologous mechanisms of rearrangement. *Molecular Genetics and Metabolism*. 2006In Press, Corrected Proof
- Kwaks TH, Otte AP. Employing epigenetics to augment the expression of therapeutic proteins in mammalian cells. *Trends Biotechnol* 2006;24:137–42. [PubMed: 16460822]
- Lander ES, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921. [PubMed: 11237011]
- Lavigne MD, Gorecki DC. Emerging vectors and targeting methods for nonviral gene therapy. *Expert Opin Emerg Drugs* 2006;11:541–57. [PubMed: 16939390]
- Leib-Mösch C, Seifarth W. Evolution and biological significance of human retroelements. *Virus Genes* 1995;11:133–45. [PubMed: 8828141]
- Lercher MJ, Urrutia AO, Hurst L. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics* 2002;31:180–183. [PubMed: 11992122]
- Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. A unification of mosaic structures in the human genome. *Hum Mol Genet* 2003;12:2411–2415. [PubMed: 12915446]
- Li L, McVety S, Younan R, Liang P, Du Sart D, Gordon PH, Hutter P, Hogervorst FB, Chong G, Foulkes WD. Distinct patterns of germ-line deletions in MLH1 and MSH2: the implication of Alu repetitive element in the genetic etiology of Lynch syndrome (HNPCC). *Hum Mutat* 2006;27:388. [PubMed: 16541406]
- Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;2:18–22.

- Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R. Role of transposable elements in heterochromatin and epigenetic control. *Nature* 2004;430:471–6. [PubMed: 15269773]
- Lorincz MC, Schubeler D, Hutchinson SR, Dickerson DR, Groudine M. DNA methylation density influences the stability of an epigenetic imprint and Dnmt3a/b-independent de novo methylation. *Mol Cell Biol* 2002;22:7572–80. [PubMed: 12370304]
- Lupski JR. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biol* 2004;5:242. [PubMed: 15461806]
- Lyon MF. X-chromosome inactivation: a repeat hypothesis. *Cytogenet Cell Genet* 1998;80:133–137. [PubMed: 9678347]
- Maloisel L, Rossignol JL. Suppression of crossing-over by DNA methylation in *Ascombolus*. *Genes Dev* 1998;12:1381–9. [PubMed: 9573054]
- Marahrens Y. X-inactivation by chromosomal pairing events. *Genes Dev* 1999;13:2624–32. [PubMed: 10541548]
- Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J* 2005;24:800–12. [PubMed: 15678104]
- Matejas V, Huehne K, Thiel C, Sommer C, Jakubiczka S, Rautenstrauss B. Identification of Alu elements mediating a partial PMP22 deletion. *Neurogenetics* 2006;7:119–26. [PubMed: 16570190]
- Medstrand P, van de Lagemat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* 2002;12:1483–95. [PubMed: 12368240]
- Mietz JA, Kuff EL. Tissue and strain-specific patterns of endogenous proviral hypomethylation analyzed by two-dimensional gel electrophoresis. *Proc Natl Acad Sci U S A* 1990;87:2269–73. [PubMed: 2107549]
- Minakami R, Kurose K, Etoh K, Furuhashi Y, Hattori M, Sakaki Y. Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res* 1992;20:3139–45. [PubMed: 1320255]
- Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* 2002;71:312–26. [PubMed: 12070800]
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 2005;310:321–4. [PubMed: 16224025]
- Nishimura DY, Swiderski RE, Searby CC, Berg EM, Ferguson AL, Hennekam R, Merin S, Weleber RG, Biesecker LG, Stone EM, Sheffield VC. Comparative genomics and gene expression analysis identifies BBS9, a new Bardet-Biedl syndrome gene. *Am J Hum Genet* 2005;77:1021–33. [PubMed: 16380913]
- Nishioka Y. Tissue specific methylation of human Y chromosomal DNA sequences. *Tissue Cell* 1988;20:875–80. [PubMed: 3245036]
- Nissen PH, Damgaard D, Stenderup A, Nielsen GG, Larsen ML, Faergeman O. Genomic characterization of five deletions in the LDL receptor gene in Danish Familial Hypercholesterolemic subjects. *BMC Med Genet* 2006;7:55. [PubMed: 16796766]
- Noma K, Cam HP, Maraia RJ, Grewal SI. A role for TFIIC transcription factor complex in genome organization. *Cell* 2006;125:859–72. [PubMed: 16751097]
- Okada N, Hamada M. The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. *J Mol Evol* 1997;44(Suppl 1):S52–6. [PubMed: 9071012]
- Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P. IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res* 2004;32:W287–92. [PubMed: 15215396]
- Ostertag EM, Kazazian HH Jr. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* 2001;35:501–38. [PubMed: 11700292]
- Ovchinnikov I, Troxel AB, Swergold GD. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res* 2001;11:2050–8. [PubMed: 11731495]
- Pàldi A, Gyapay G, Jami J. Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies. *Curr Biol* 1995;5:1030–5. [PubMed: 8542279]

- Pannell D, Ellis J. Silencing of gene expression: implications for design of retrovirus vectors. *Rev Med Virol* 2001;11:205–17. [PubMed: 11479927]
- Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar J, Bernardi G. Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 2001;276:39–45. [PubMed: 11591470]
- Pieretti M, Zhang F, Fu YH, Warren S, Oostra BA, Caskey CT, Nelson DL. Absence of expression of the FMR-1 gene in Fragile X syndrome. *Cell* 1991;66:817–822. [PubMed: 1878973]
- Quentin Y. Emergence of master sequences in families of retrotransposons derived from 7sl RNA. *Genetica* 1994;93:203–215. [PubMed: 7529207]
- Regelson M, Eller CD, Horvath S, Marahrens Y. A link between repetitive sequences and gene replication time. *Cytogenet Genome Res* 2006;112:184–93. [PubMed: 16484771]
- Rhodes K, Oshima RG. A regulatory element of the human keratin 18 gene with AP-1-dependent promoter activity. *J Biol Chem* 1998;273:26534–42. [PubMed: 9756890]
- Robertson KD. DNA methylation, methyltransferases, and cancer. *Oncogene* 2001;20:3139–55. [PubMed: 11420731]
- Roman-Gomez J, Jimenez-Velasco A, Agirre X, Cervantes F, Sanchez J, Garate L, Barrios M, Castillejo JA, Navarro G, Colomer D, Prosper F, Heiniger A, Torres A. Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia. *Oncogene* 2005;24:7213–23. [PubMed: 16170379]
- Sano H, Sager R. Tissue specificity and clustering of methylated cytosines in bovine satellite I DNA. *Proc Natl Acad Sci U S A* 1982;79:3584–8. [PubMed: 6954504]
- Santourlidis S, Florl A, Ackermann R, Wirtz HC, Schulz WA. High frequency of alterations in DNA methylation in adenocarcinoma of the prostate. *Prostate* 1999;39:166–74. [PubMed: 10334105]
- Saveliev A, Everett C, Sharpe T, Webster Z, Festenstein R. DNA triplet repeats mediate heterochromatin-protein-1-sensitive variegated gene silencing. *Nature* 2003;422:909–13. [PubMed: 12712207]
- Schmid C. Does SINE evolution preclude Alu function? *Nucl Acids Res* 1998;26:4541–4550. [PubMed: 9753719]
- Schnable PS, Hsia AP, Nikolau BJ. Genetic recombination in plants. *Curr Opin Plant Biol* 1998;1:123–9. [PubMed: 10066571]
- Schulz WA. L1 retrotransposons in human cancers. *J Biomed Biotechnol* 2006;2006:83672. [PubMed: 16877821]
- Scott KC, Merrett SL, Willard HF. A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. *Curr Biol* 2006;16:119–29. [PubMed: 16431364]
- Segal Y, Peissel B, Renieri A, de Marchi M, Ballabio A, Pei Y, Zhou J. LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *Am J Hum Genet* 1999;64:62–9. [PubMed: 9915944]
- Selker EU. Gene silencing: repeats that count. *Cell* 1999;97:157–60. [PubMed: 10219236]
- Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P, Batzer MA. Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* 2006;79:41–53. [PubMed: 16773564]
- Shabbeer J, Yasuda M, Benson SD, Desnick RJ. Fabry disease: identification of 50 novel alpha-galactosidase A mutations causing the classic phenotype and three-dimensional structural analysis of 29 missense mutations. *Hum Genomics* 2006;2:297–309. [PubMed: 16595074]
- Smit AF. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development* 1996;6:743–748. [PubMed: 8994846]
- Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development* 1999;9:657–663. [PubMed: 10607616]
- Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol* 2001;21:1973–85. [PubMed: 11238933]
- Stancheva I. Caught in conspiracy: cooperation between DNA methylation and histone H3K9 methylation in the establishment and maintenance of heterochromatin. *Biochem Cell Biol* 2005;83:385–95. [PubMed: 15959564]

- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004;101:6062–6067. [PubMed: 15075390]
- Sun FL, Haynes K, Simpson CL, Lee SD, Collins L, Wuller J, Eissenberg JC, Elgin SC. cis-Acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. *Mol Cell Biol* 2004;24:8210–20. [PubMed: 15340080]
- Swergold GD. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* 1990;10:6718–29. [PubMed: 1701022]
- Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* 2002;110:327–38. [PubMed: 12176320]
- Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. *Genome Biol* 2002;3:research0052. [PubMed: 12372140]
- Takai D, Yagi Y, Habib N, Sugimura T, Ushijima T. Hypomethylation of LINE1 retrotransposon in human hepatocellular carcinomas, but not in surrounding liver cirrhosis. *Jpn J Clin Oncol* 2000;30:306–9. [PubMed: 11007163]
- Team, RDC. R Foundation for Statistical Computing. 2003. R: A language and environment for statistical computing.
- Terai Y, Takahashi K, Okada N. SINE Cousins: The 3'-End Tails of the Two Oldest and Distantly Related Families of SINEs are Descended from the 3' Ends of LINES with the Same Genealogical Origin. *Mol Biol Evol* 1998;15:1460–1471. [PubMed: 12572609]
- Thorey IS, Cecena G, Reynolds W, Oshima RG. Alu sequence involvement in transcriptional insulation of the keratin 18 gene in transgenic mice. *Mol Cell Biol* 1993;13:6742–51. [PubMed: 7692231]
- Uddin RK, Zhang Y, Siu VM, Fan YS, O'Reilly RL, Rao J, Singh SM. Breakpoint Associated with a novel 2.3 Mb deletion in the VCFS region of 22q11 and the role of Alu (SINE) in recurring microdeletions. *BMC Med Genet* 2006;7:18. [PubMed: 16512914]
- Ullu E, Tschudi C. Alu sequences are processed 7SL RNA genes. *Nature* 1984;312:171–172. [PubMed: 6209580]
- Waldman AS, Liskay RM. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Mol Cell Biol* 1988;8:5350–7. [PubMed: 2854196]
- Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* 1998;20:116–7. [PubMed: 9771701]
- Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2000;2:143–7. [PubMed: 11015593]
- Whitelaw E, Martin DI. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat Genet* 2001;27:361–5. [PubMed: 11279513]
- Wild K, Rosendal KR, Sinning I. A structural step into the SRP cycle. *Mol Microbiol* 2004;53:357–363. [PubMed: 15228518]
- Willoughby DA, Vilalta A, Oshima RG. An Alu element from the K18 gene confers position-independent expression in transgenic mice. *J Biol Chem* 2000;275:759–768. [PubMed: 10625605]
- Woodcock DM, Crowther PJ, Diver WP, Graham M, Bateman C, Baker DJ, Smith SS. RglB facilitated cloning of highly methylated eukaryotic DNA: the human L1 transposon, plant DNA, and DNA methylated in vitro with human DNA methyltransferase. *Nucleic Acids Res* 1988;16:4465–82. [PubMed: 2837736]
- Woodcock DM, Lawler CB, Linsenmeyer ME, Doherty JP, Warren WD. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J Biol Chem* 1997;272:7810–6. [PubMed: 9065445]
- Xie F, Wang X, Cooper DN, Chuzhanova N, Fang Y, Cai X, Wang Z, Wang H. A novel Alu-mediated 61-kb deletion of the von Willebrand factor (VWF) gene whose breakpoints co-locate with putative matrix attachment regions. *Blood Cells Mol Dis* 2006;36:385–91. [PubMed: 16690331]
- Yamada T, Mizuno K, Hirota K, Kon N, Wahls WP, Hartsuiker E, Murofushi H, Shibata T, Ohta K. Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot. *Embo J* 2004;23:1792–803. [PubMed: 14988732]

- Yang AS, Estecio MR, Doshi K, Kondo Y, Tajara EH, Issa JP. A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Res* 2004;32:e38. [PubMed: 14973332]
- Yao H, Zhou Q, Li J, Smith H, Yandau M, Nikolau BJ, Schnable PS. Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proc Natl Acad Sci U S A* 2002;99:6157–62. [PubMed: 11959909]
- Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 1997;13:335–40. [PubMed: 9260521]
- Zhang G, Fukao T, Sakurai S, Yamada K, Michael Gibson K, Kondo N. Identification of Alu-mediated, large deletion-spanning exons 2–4 in a patient with mitochondrial acetoacetyl-CoA thiolase deficiency. *Molecular Genetics and Metabolism*. 2006In Press, Corrected Proof

Abbreviations

HK	housekeeping gene
TS	tissue-specific gene
LINE-1	Long Interspersed Nuclear Element 1
SINE	Short Interspersed Nuclear Element
repeat	repetitive sequence

Figure 1.

Sequence environment of housekeeping genes and tissue-specific genes. **(a)** Sequence composition in 200-kb regions flanking housekeeping genes (HK), tissue-specific genes (TS) and a random sample of genes (RS). P-values were obtained using the Kruskal-Wallis test; error bars represent 95% confidence intervals. We also considered CpG island densities at various sized intervals around either the gene body or the start of transcription and found, in both cases, a 200-kb region to yield the most significant differences between the gene groups (data not shown). **(b)** Sequence properties for 1Mb intervals extending 40-Mb upstream and 40-Mb downstream around genes. The results were plotted using smoothed local regression (loess) curves and the regions over which significant differences between housekeeping genes

and tissue-specific genes extend were calculated at a 95% confidence interval (region between arrows). Horizontal bars mark the regions around housekeeping (black) or tissue-specific (gray) genes with significant enrichment or reduction for a sequence feature. The length of the each horizontal bar was calculated by determining where each loess curve exceeds one standard deviation from the average value of the regions 30–40 Mb away from each gene.

Figure 2. HK and TS genes are grouped according to isochore membership for comparison of their 200kb flanking regions. Numbers of HK and TS genes assigned to each isochore are given beneath the columns in **(a)**. **(a)** Repeats with statistically significant differences in concentration in at least one isochore. **(b)** Repeats grouped according to size. Short repeats are less than 400 bases in length and long repeats are greater than 400 bases. For **(a)** and **(b)**, p-values were obtained using the Kruskal-Wallis test; error bars represent 95% confidence intervals. Isochore membership is assigned according to boundaries published as a custom track to the UCSC Genome Browser (<http://genome.ucsc.edu>) (Oliver et al., 2004).

Figure 3. Housekeeping gene classification. **(a)** Changes in false positive error rate when various characteristics are removed. Error bars represent 95% confidence intervals. **(b)** Relative importance of individual repeats in classifying housekeeping genes, as generated by the *partial.plot* function in the random forest package in R.

Figure 4. Local regression (loess) curves relating HK probability scores of genes to average expression level across all tissues. Gene expression levels were obtained from Gene Atlas #1. Spearman's correlation coefficient and its associated p-value are reported.

Figure 5.

Genome-wide relationships between HK probability, breadth of gene expression across tissues, and repeat concentration in 200-kb regions flanking genes. **(a)** Relationship between the calculated probability of being housekeeping genes (HK probability) and the proportion of tissues in which the genes are expressed. Gene expression data were obtained from Gene Atlas #1. **(b)** Partial dependence plots depict the marginal effect of each characteristic on HK Probability as determined by the random forest classifier. Negative probability values indicate that the characteristic weakens the classifier. Rug marks denote deciles and can be seen in greater detail in Supplementary Fig. 12 online. **(c)** Relationship between breadth of gene

expression and concentration of Alu elements. For **(a)** and **(c)**, Spearman correlation coefficients and associated p-values are reported.