Alu and LINE-1 are Determinants for Repressive Mark Type at Genes and Degree of Gene Expression Level Variation

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Wuming Gong

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

York Marahrens

January 2013

Acknowledgements

I would most of all like to thank my parents for their unwavering support and the enthusiasm they show toward my career in science and thank York Marahrens for taking me on in his laboratory, for his excellent guidance, and for providing a first rate working and learning environment. I would also like to thank Naoko Shima and Anindya Bagchi, and their laboratory members for insightful and enjoyable discussions during weekly meetings. Furthermore, I would also like to acknowledge the members of Marahrens' laboratory.

Abstract

A fundamental unanswered question in biology is why the Alu transposable elements that constitute 11% of the human genome have amassed at varying levels in the vicinity of most genes and appear to have been selected against in non-gene regions while LINE-1 transposons that constitute 16% of the genome have accumulated around a minority of genes and are abundant in non-gene regions. Here we show that genes flanked by increasingly higher Alu concentrations display progressively less variation in expression level among humans and across human cell types while genes flanked by increasingly higher LINE-1 concentrations show progressively higher variation. Bayesian network modeling indicates that Alu elements promote H3K36me3 chromatin that fosters low gene expression variation while LINE-1 elements encourage H3K9me3 chromatin that procures high expression variation. Accordingly, genes in high Alu low LINE-1 environments produce H3K36me3 in their transcribed regions of at levels reflecting the levels of transcription while genes residing in progressively higher LINE-1 environments are increasingly likely to establish H3K9me3 in their transcribed regions, again at levels reflecting the levels of transcription. Further along these lines, silent genes occupied by repressive H3K36me3 across their regulatory regions reside primarily in high Alu low LINE-1 chromosomal environments, while occupation of silent genes by H3K9me3 is most prevalent in low Alu high LINE-1 environments, and genes silenced by H3K27me3 are concentrated in low Alu low LINE-1 environments. Genes residing in medium-high LINE-1 but extreme low Alu genes are exceptional as they tend to be occupied by an uncharacterized chromatin type. Finally, gene-distant regions are generally high LINE-1 low Alu, and here LINE-1 also promotes uncharacterized non-H3K9me3 chromatin that Alu opposes. Our findings indicate that local Alu and LINE-1 concentrations determine chromatin systems utilized by many genes.

Table of Contents

List of Figures

Introduction

Mobile elements constitute almost one half of the human genome with the most abundant being the LINE-1 transposable element (~900,000 copies or ~21% of the human genome) and the Alu transposon (1.4 million copies or ~ 11% of the human genome). The DNA sequences of mobile element insertions degrade over millenia and the number of recognized elements in the genome depends on the threshold at which one begins to recognize degraded copies as elements. An intact full length LINE-1 element is approximately 6-kb and encodes protein and signal sequences responsible for its own mobilization. An internal promotor drives the production of RNA copies of the LINE-1 that are then reverse transcribed and inserted into the genome by a LINE-1-encoded reverse transcriptase and endonuclease (*1*, *2*). The Alu element is a non-autonomous transposon that has been shown to hijack the LINE-1-encoded proteins for reverse transcription and insertion (*1*, *3-14*).

A fundamental unanswered question in biology is why Alu elements are concentrated near most genes. Transposition studies in cultured cells indicate that the LINE-1 machinery inserts newly synthesized elements at largely random positions in the genome (*1*, *2*). The relatively recent Alu and LINE-1 transposition events in human evolution similarly show little insertion bias (*1*, *3-14*) aside from somewhat favoring a few insertion hot spots (*10*, *15-17*), TT|AAAA target sequences(*5*, *15*, *18*), and A+T rich regions in general (*1*). Despite their insertion at largely random positions, the Alu and LINE-1 distributions in the human genome are strikingly non-random with Alu elements having concentrated to varying degrees around most genes and LINE-1 elements displaying a largely reciprocal distribution compared to Alu (*9*, *12*, *19-21*). Analysis of insertion patterns supports the idea that selection against LINE-1 retrotransposons from the vicinity of genes results primarily from their ability to mediate ectopic recombination (*22*). A popular idea is that the non-random distribution of Alu elements is the product of natural selection (*4*, *10*, *16*), but what drives this selection remains to be elucidated. Alu and LINE-1 concentrations correlate with magnitude and breadth of gene expression (*19*, *20*, *23*) and the transposons have been calculated to account for 78% of peak gene expression level, 76% of breadth of expression across tissues, and 66% of tissue specificity (*23*). The aforementioned studies suggest that Alu elements promote gene expression by an unknown mechanism, but the Alu accumulations around many genes do not conform to these correlations suggesting that additional selective forces are at work.

The accumulation of Alus near the promotors of genes is all the more puzzling because Alu elements are silenced by DNA methylation (*24*), a component of all known repressive chromatin structures in humans. Transposon repression limits gene disruption by novel Alu and LINE-1 insertions to about 0.3% of human genetic disease (*25*, *26*). Even LINE-1 elements in introns that do not disrupt splicing have been shown to reduce gene transcription (*27*). Transposon repression also increases genome instability brought on by incomplete or abortive transposition reaction or by ectopic double strand breaks by the LINE-1 endonuclease. However, Alu and LINE-1 transposons can also create deletions and translocations via illegitimate homologous recombination between copies and a large proportion of copy number variations in humans have LINE-1 and Alu elements at their breakpoints. A subset of Alu elements located near gene promotors has been reported to be hypomethylated. As long as the LINE-1 proteins are not expressed, hypomethylated Alu elements do not multiply (*28*-*44*).

Repetitive sequences amount for the majority of constitutive heterochromatin in the human genome. Interestingly, there are at least three distinct and apparently mutually exclusive forms of repressive chromatin that frequently are identified as H3K27me3, H3K9me3, and H3K36me3 chromatin. These can alternate along heterochromatic regions. H3K9me3 heterochromatin is found at centromeric repeats (*45*), pericentromeric regions (*46*), telomeric heterochromatin (*47*), and many endogenous retroviral (ERV) elements (*48*). H3K27me3 is the most common facultative heterochromatin (*49*) but also represses some ERV elements (*50*, *51*) including HIV (*52*). H3K36me3 is associated with pericentromeric heterochromatin (*53*) and with the transcribed regions of genes where it is thought to suppress spurious transcriptional initiation (*54*). Interestingly, H3K36me3 appears to also promote transcriptional elongation and, in addition, helps up-regulate genes on the X chromosome in male flies, suggesting that the role of H3K36me3 extends beyond merely suppressing transcriptional initiation (*54*). Whether the LINE-1 and Alu elements of the human genome are associated with H3K27me3, H3K9me3, or H3K36me3 marks has not been the focus of any studies.

Here we link Alu elements to H3K36me3 and gene expression variation in the human population. We show that Alu elements are associated with reduced gene expression variation and LINE-1 elements with increased variation between monozygotic twins, across the human population, and across tissue types. We show that Alu and LINE-1 elements are either in H3K36me3, H3K9me3, or H3K27me3 chromatin depending on what chromosomal domain they are in and show what promotes one of these modifications to prevail throughout the domain. Bayesian modeling indicates that H3K27me3 and H3K9me3 increase gene expression variation while H3K36me3 and H3K79m2 reduce variation. Finally, we develop gene expression variation maps of the human genome for numerous tissues and show that all this occurs within the context of a much larger megabase domain organization and that tissue-specific differences arise at this larger domain level.

Genes differentially expressed in monozygotic twin pairs and unrelated individuals reside in low-Alu high-LINE-1 regions

To explain the highly non-random distribution of Alu and LINE-1 elements in the genome (Figure 1A), we hypothesized that these elements may be playing a role in gene expression variation. To investigate this possibility, we first examined whether gene expression variation in humans could be associated with repetitive sequence environment. We considered a region extending 100-kb upstream from the transcription start and 100-kb downstream from the transcription end of each gene. We excluded the transcribed region from our analysis to avoid effects attributable to displacement by the coding sequence, splicing elements, or to the disruption of transcriptional elongation by repetitive elements in introns (*27*). We downloaded RepeatMasker file UCSC genome browser (hg19) and calculated the densities of Alu and LINE-1 on upstream and downstream 100-kb regions for each human gene. It has been shown that density of Alu, LINE-1 as well as other repetitive elements such as ERV1, from genes' upstream and downstream 100-kb have been successfully used to infer biological features, such as imprinted genes in mouse and human (*55*, *56*).

We sorted all genes into a 2-D matrix consisting of 21 %Alu categories (Y-axis) and 21 %LINE-1 categories (X-axis). This revealed that genes varied widely with respect to the abundance of Alu and LINE-1 in their 200-kb flanking regions and a loose reciprocal relationship with genes in low Alu gene environments being flanked by high LINE-1 concentrations and vice versa (Figure 1B).

We compared the gene expression profiles of lymphoblastoid cell lines (LCLs) derived from the blood of five healthy monozygotic (MZ) twin pairs (*57*). Gene expression of one individual (individual one), selected at random from a twin pair, was compared to the monozygotic sibling (individual two). Genes were split into three categories: ≥ two-fold higher expression in individual one compared to the monozygotic sibling, ≥ two-fold higher expression in individual two compared to the monozygotic sibling, and less than two-fold difference. For each of the five MZ twin pairs, >900 genes displayed ≥ two-fold lower expression in individual one compared to the monozygotic sibling, between 11,000 and 13,000 genes showed a less than two-fold difference (which we categorized as 'same'), and another >900 genes displayed ≥ two-fold higher expression in individual one. We then determined the average amount of Alu or LINE-1 sequence, expressed as percents, in the 200-kb flanking sequences of the 'lower', 'same' and 'higher' categories of genes. For all five monozygotic twin pairs, the 'lower' and 'higher' genes were flanked by significantly less Alu sequence and significantly more LINE-1 sequence than the 'same' genes (Figure 1C). For all five monozygotic twin pairs, the 'lower' and 'higher' genes were also flanked by significantly more ERVL sequence, and MalR sequence than the 'same' genes (data not shown).

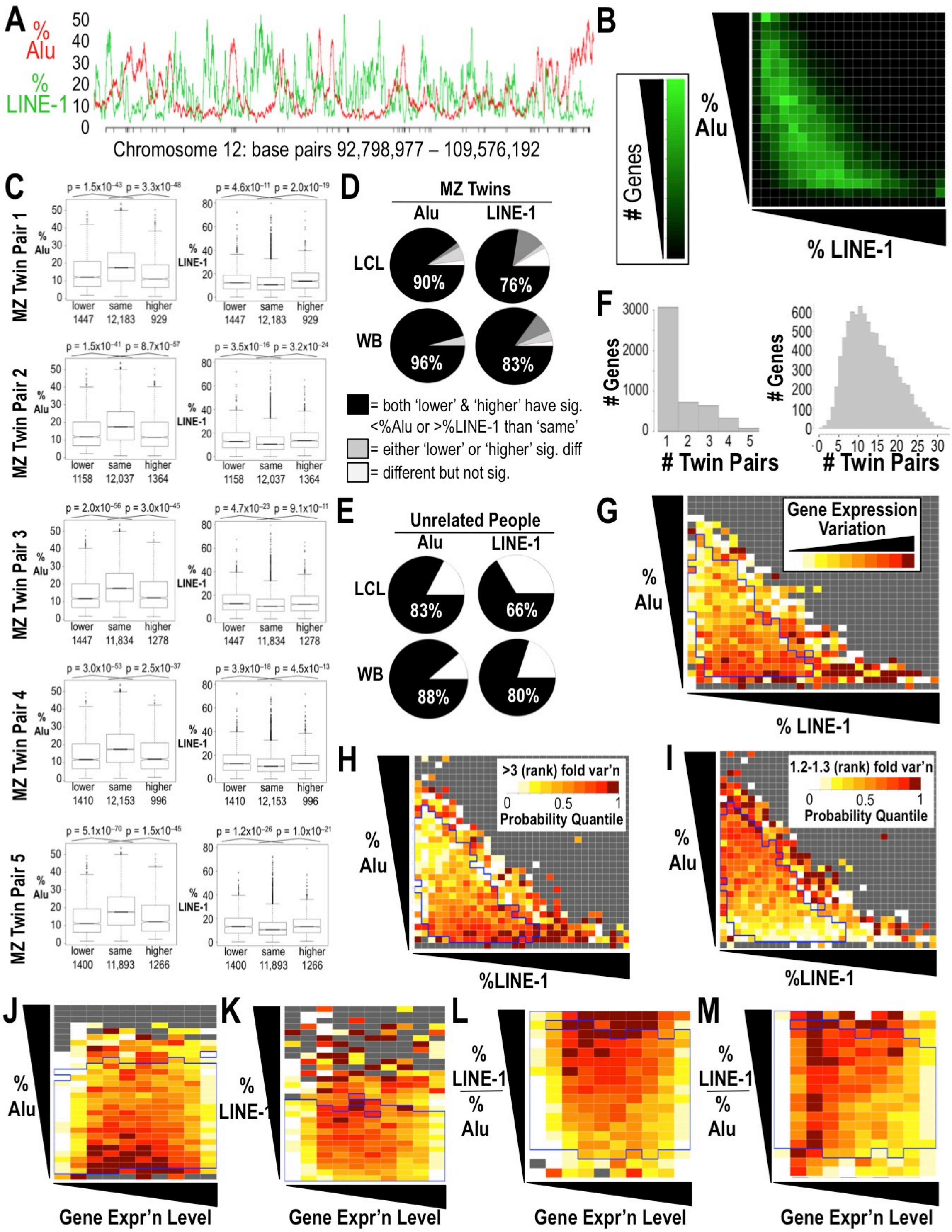

A
% Alu
% LINE-1
50
40
30
20
10
0
Chromosome 12: base pairs 92,798,977 – 109,576,192
B
# Genes
% Alu
% LINE-1
C
MZ Twin Pair 1
p = 1.5x10^-43 p = 3.3x10^-48
% Alu
lower 1447 same 12,183 higher 929
p = 4.6x10^-11 p = 2.0x10^-19
% LINE-1
lower 1447 same 12,183 higher 929
MZ Twin Pair 2
p = 1.5x10^-41 p = 8.7x10^-57
lower 1158 same 12,037 higher 1364
p = 3.5x10^-16 p = 3.2x10^-24
lower 1158 same 12,037 higher 1364
MZ Twin Pair 3
p = 2.0x10^-56 p = 3.0x10^-45
lower 1447 same 11,834 higher 1278
p = 4.7x10^-23 p = 9.1x10^-11
lower 1447 same 11,834 higher 1278
MZ Twin Pair 4
p = 3.0x10^-53 p = 2.5x10^-37
lower 1410 same 12,153 higher 996
p = 3.9x10^-18 p = 4.5x10^-13
lower 1410 same 12,153 higher 996
MZ Twin Pair 5
p = 5.1x10^-70 p = 1.5x10^-45
lower 1400 same 11,893 higher 1266
p = 1.2x10^-26 p = 1.0x10^-21
lower 1400 same 11,893 higher 1266
D
MZ Twins
Alu
LINE-1
LCL
90%
76%
WB
96%
83%
■ = both 'lower' & 'higher' have sig. <%Alu or >%LINE-1 than 'same'
▨ = either 'lower' or 'higher' sig. diff
□ = different but not sig.
E
Unrelated People
Alu
LINE-1
LCL
83%
66%
WB
88%
80%
F
# Genes
3000
2000
1000
0
1 2 3 4 5
# Twin Pairs
# Genes
600
500
400
300
200
100
0
0 5 10 15 20 25 30
# Twin Pairs
G
Gene Expression Variation
% Alu
% LINE-1
H
>3 (rank) fold var'n
0 0.5 1
Probability Quantile
% Alu
%LINE-1
I
1.2-1.3 (rank) fold var'n
0 0.5 1
Probability Quantile
% Alu
%LINE-1
J
% Alu
Gene Expr'n Level
K
% LINE-1
Gene Expr'n Level
L
% LINE-1 / % Alu
Gene Expr'n Level
M
% LINE-1 / % Alu
Gene Expr'n Level

Figure 1.

Gene expression variation in LCLs and whole blood correlates with abundance of Alu and LINE-1 sequence in 200-kb regions flanking the gene (100-kdb immediately upstream plus 100-kb downstream of the transcribed region). **A.** Nonrandom distribution of Alu and LINE-1 elements in a randomly selected chromosomal region. Alu and LINE-1 abundance was calculated as percents in of the total sequence rolling averages over 100-bp. Red, Alu. Green, LINE-1. Vertical black tricks, transcription start sites of genes. **B.** Heatmap showing the relative abundance of the genes of the genome across Alu and LINE-1 concentrations. Both axes run, in increments of 2%, from 0% to 40%, followed by a single Alu row or LINE-1 column representing >40%. **C.** Boxplots showing the abundance of Alu (left) and LINE-1 (right) sequence around genes that are either more than two-fold or less than two-fold differently expressed in LCLs from the two individuals of monozygotic twin pairs. Results from five monozygotic twin pairs (GEO accession number GSE7624) are shown. 'Lower', ≥ two-fold lower expression in individual A vs. monozygotic sibling 'B'. 'Same', <2-fold differently expressed. 'Higher', ≥2-fold higher expression in individual A vs. B. Numbers of genes falling into each category are shown below the corresponding labels. P values for the indicated comparisons were determined using the Wilcoxon Rank Sum Test. **D.** Pie charts showing the percentage of pairwise comparisons among 49 MZ twin pairs for lymphoblastoid cell lines (LCLs, top) and 47 MZ twin pairs for whole blood (WB, bottom) where both 'lower' and 'higher' expressed genes reside among significantly less Alu (left charts, black) or significantly more LINE-1 sequence (right charts, black) than 'same' genes. The GEO accession number for both the 47 and 49 pairs of samples is GSE33321. Comparisons were performed as in 'C' above, but do not include the five twin pairs from 'C'. If only one of two differential expression categories ('higher' or 'lower') is significantly different, the corresponding pie chart section is dark gray. If a ≥ two-fold differentially expressed gene is lower Alu or higher LINE-1 than 'same' genes but the difference is not statistically significant, they fall into the light gray section of the pie chart. **E.** Pairwise comparisons among unrelated individuals drawn from the 49 MZ twin LCL (upper) and 47 MZ twin WB (lower) sample sets performed as in 'D' above. Here the gray sections were consolidated with the white sections. **F.** Number of genes that are ≥2-fold differently expressed one MZ twin pair, two twin pairs, three twin pairs, from among the LCLs from the set of five twin pairs (left) and from the WB sets of 47 MZ twins (right). **G.** Map showing the level of expression variation across the genome with expressed genes sorted according to their %Alu and %LINE-1 environments (100kdb upstream + 100-kb downstream of genes). Coefficient of gene expression variation (*η*) was calculated by dividing the standard deviation of gene expression level by the mean expression level. Data for LCLs from 210 individuals from HapMap project is shown. The darker the color, the greater the expression level variation. The blue line encompasses squares that represent ≥20 genes. **H.** Heatmap showing the frequency of high gene expression variation across the range of %Alu / %LINE-1 environments for LCLs of 210 HapMap individuals. Frequency of gene expression variation expressed as probability quantile. **I.** Heatmap of low gene expression variation in the indicated %Alu / %LINE-1 environments in the aforementioned LCL profiles of 210 HapMap individuals. **J.** Color map of coefficient of gene expression variation (η) for expressed genes sorted according to %Alu (Y-axis) and gene expression level quantile (X-axis). **K.** Heatmap of coefficient of gene expression variation (η) for expressed genes sorted according to %LINE-1 (Y-axis) and gene expression level quantile (X-axis). **L.** Heatmap of coefficient of expression level variation (η) for expressed genes sorted according to the ratio of %LINE-1 divided by %Alu (Y-axis) and gene expression level quantile (X-axis). **M.** Heatmap of the mode of ran change spectrum for expressed genes sorted according to the ratio of %LINE-1 divided by %Alu (Y-axis) and gene expression level quantile (X-axis). For J-M, the darker the color, the larger the gene expression variation.

To test the generality of the finding that genes displaying ≥2 fold expression level differences reside in low-Alu high-LINE-1 environments across MZ twins, we examined the gene expression profiles of LCLs derived from 49 monozygotic twin pairs (*58*). The ≥2 fold differently expressed genes followed the low-Alu pattern in 89% (44/49) of MZ twin pairs and the high LINE-1 pattern in 76% (37/49) of MZ twin pairs (Figure 1D, upper). To determine whether the low-Alu high-LINE-1 trend is an artifact of the immortalization or transformation that occurs during the derivation of LCL lines, we repeated this analysis using the expression profiles obtained from whole blood (WB) from 47 MZ twin pairs (*58*). For the WB samples: 96% (45/47) MZ twins followed the low-Alu trend and 83% (39/47) of MZ twin pairs followed high-LINE-1 trend (Figure 1D, lower). LCL lines derivation may therefore be modestly degrading the low-Alu high-LINE-1 trend. The same trends were seen when gene expression profiles from unrelated individuals from these LCL or WB datasets were subjected to such pairwise comparisons (Figure 1E). Figure 1C-E had been generated after first performing global median normalization on the microarray data, which transforms all expression values to produce a constant median across samples. Essentially the same results were obtained if the data was not first normalized or if the data was first subjected to quantile normalization, which causes the overall distributions of expression profiles to be the same for all samples (data not shown).

Next, we examined whether the ≥2 fold differently expressed genes constitute a distinct subset of low-Au high-LINE-1 genes. Among the initial five MZ twin pairs examined, most genes displayed a ≥2 fold expression difference in only one of the five twin pairs (Figure 1F, left), In the larger twin pair set, the mode frequency for dislodging a gene from its typical expression level was 10/47 MZ twins (21%) in whole blood (Figure 1F, right) and 13/49 MZ twins (27%) for LCLs. Occasional ≥2 fold expression difference in expression, therefore, appeared to be a widespread property of genes in WB and LCLs.

Gene expression variation correlates with local abundance of Alu and LINE-1 sequence

We next used %Alu vs. %LINE-1 2D heatmaps to compare the incidence of expression level variation across the spectrum of Alu and LINE-1 environments. For these and subsequent analysis, we switched from employing 'fold' variation to either using coefficient of variation, or using rank change when comparing gene expression when coefficient of variation could not be appropriately used.

Coefficient of variation is defined as the standard deviation of gene expression level divided by the mean, which is a robust and unbiased estimate of the amount of gene expression variation (*59*, *60*).

Rank change in gene expression has been shown to be more reliable than fold change for microarray data comparisons due to it being less impaired by the technical inconsistencies of microarrays (*61*). Furthermore, rank change methods facilitate the comparison of gene expression variation across microarray data sets from different platforms. We adopted a simple and effect normalization method based on rank change (*62*). Specifically, for gene *i*, we first replaced its signal by its absolute rank (AR) among all *N* expressed genes, that is, the lowest and highest expressed genes received the lowest and highest rank, respectively. The absolute rank was then transformed into a relative one (r = AR * 100 / N), expressed on a 0-100 scale. The non-expressed genes are not ranked. In this way, the raw signals are expressed on the same scale and are directly comparable. The expression variation of gene *i* between expression profile A and B is then expressed as RC= $|r_A - r_B|$, which is the absolute value of the difference of relative rank change between profile A and B. It has been shown that rank change based methods are effective on measuring the gene expression variation and capturing the differentially expressed genes (*62*, *63*).

To characterize the full range of variation amplitudes for each gene, we divided the range of rank change variation into 20 equal increments (quantiles) and calculated the probability that fall into each quantile.

Specifically, once the RC was obtained for each of N expressed genes between any two expression profile, we split genes into 20 RC groups based on the RC quantiles (RCQ) (0.05, 0.1, ..., 0.95 and 1). The genes that fall into high RC quantile (e.g. RCQ_1) show high expression variation between two profiles, while the genes that fall into low RC quantile (e.g. $RCQ_{0.05}$) show low variation.

To characterize the variation of gene *i* among *M* profiles from a specific tissue or cell types, we performed pairwise comparison of any two different expression profiles, and count the times that this gene fall into any one of the RC quantiles ($n_{0.05}$, ..., n_1). The 'rank change spectrum' (RCS) for gene *i* (RCS_i) is defined as the probability that this gene fall into each RC quantile among all the pairwise comparisons:

$$RCS_{i,k} = \frac{n_{i,k}}{\sum_{h=1}^{20} n_{i,h}}$$

where k = 0.05, ..., 1. The rank change spectrum provides a measurement of gene expression variation among multiple profiles.

High-resolution analysis of relatively infrequent events requires larger populations, so we turned to the gene expression profiles of LCLs from 210 HapMap individuals(*64*, *65*), which include 45 unrelated people of Chinese from Beijing (CHB), 45 unrelated Japanese from Tokyo (JPT), 60 unrelated Utah residents with Northern and Western European ancestry (CEU), and 60 unrelated African people (YRI). The data were quantile normalized within replicates, followed by median normalization between individuals within each population. The probes which detection p value is >0.05 are considered to be non-expressed.

A heatmap for magnitude of coefficient of variation for expressed genes across the same ranges of %Alu and %LINE-1 environments as in Figure 1B revealed that gene expression variation was greatest for genes residing in low Alu high LINE-1 environments (Figure 1G). The same conclusion was reached when the probability of being top 5% most fluctuated genes was used in place of coefficient of variation (Figure 1H). In contract, the probability of being top 5% least fluctuated genes was frequent in high-Alu low-LINE-1 genes (Figure 1I). Finally, we looked at see if the association of high Alu low LINE-1 environments with high gene expression variation was consistent across gene expression levels. 2-D heatmaps for coefficient of variation among genes sorted according to Alu% and gene expression level (Figure 1J), LINE-1% and gene expression (Figure 1K), or the ratio of %LINE-1 to %Alu and expression (Figure 1L) revealed the greatest variation and strongest correlations with %Alu and %LINE-1 to be among the genes in the middle two-thirds of the expression level spectrum. Replacing coefficient of variation with the mode of rank change spectrum (RCS), that is, the most frequent rank change quantile that a gene likely fall into by comparing any two different expression profiles, produced a similar chart (Figure 1M).

The finding that gene expression variation correlates with abundance of Alu and LINE-1 sequence in blood and LCLs derived from blood raised the question of whether this represented a general trend across all cell types. We identified large numbers of gene expression profile datasets that include large numbers of samples from healthy people from GEO database (*66*). To present large amount of information in a manageable space, we converted 2-D plots to linear plots. To start things off, the high amplitude and low amplitude 2-D heatmaps for 210 HapMap LCLs in Figure 1G and H were combined by, dividing the probability of high-amplitude variation in each square of the 1G heatmap, by the probability of low-amplitude variation in the corresponding square of the 1H heatmap. The resulting 2-D plot was then duplicated and one copy was collapsed along the X-axis so only the Alu gradient is presented (Figure 2, left top row) and the other copy was collapsed along the Y-axis so only the LINE-1 gradient is presented (Figure 2, right top row). All additional datasets were presented in the same manner below the

HapMap LCL row, organized into Mesoderm (top), Ectoderm (middle), and Endoderm (bottom). In all datasets, gene expression variation formed a gradient that ran from more high amplitude to more low amplitude among the %Alu gradient (Figure 2, left). In most datasets, gene expression variation also formed a gradient that conformed to the %LINE-1 gradient (Figure 2, right). Datasets from several studies involving whole blood revealed the effect of differences in sample quality. Most of the deviations in the heatmap gradients can therefore probably be attributed to sample quality. We conclude that the correlation of gene expression variation with Alu environment is widespread across tissues and the correlation with LINE-1 may be widespread also.

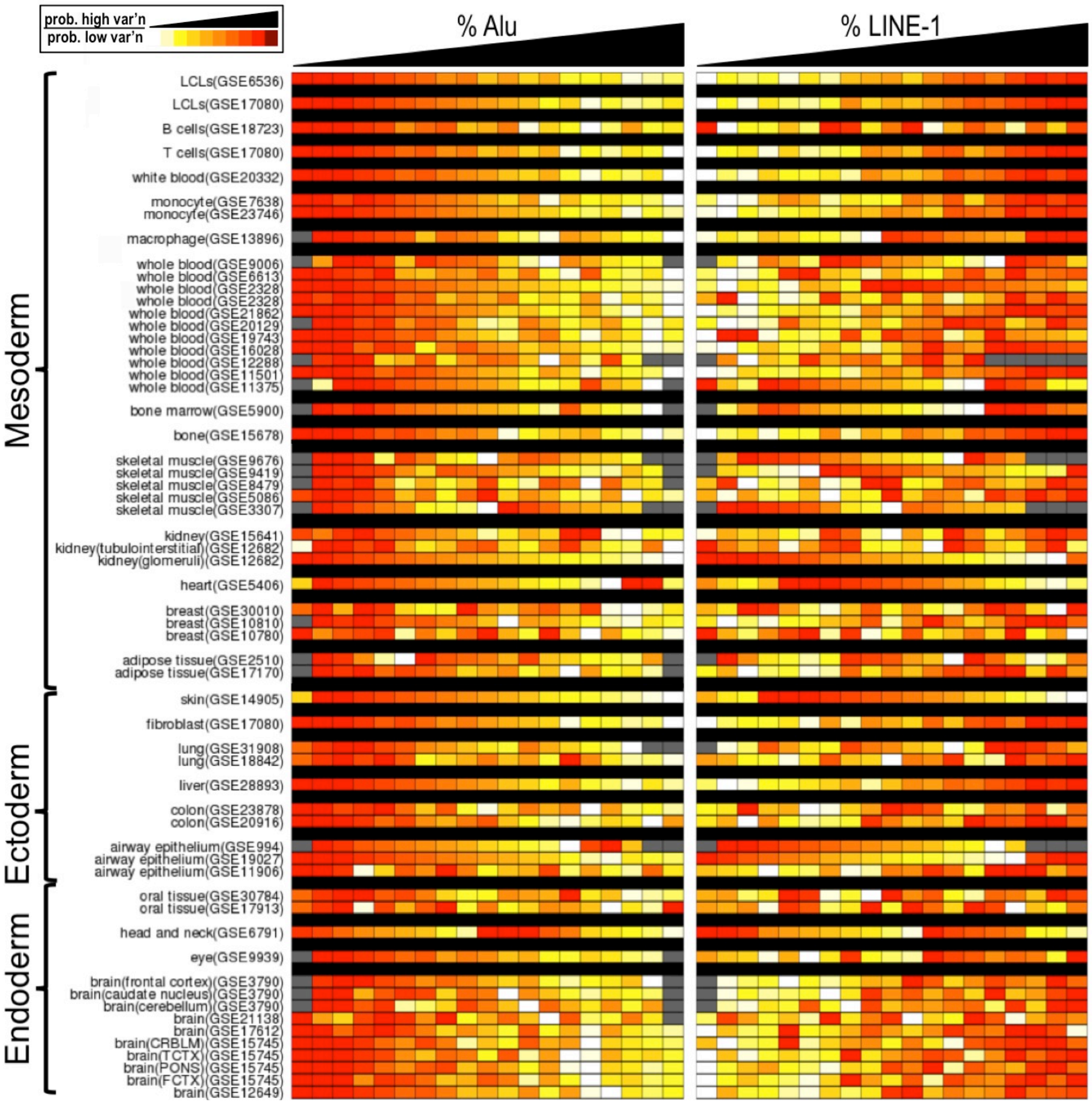


Figure 2

Gene expression variation correlates with Alu and LINE-1 sequence abundance across cell types and tissues. Healthy normal cell or tissue types and corresponding NCBI GEO accession for gene expression profiles are identified at the left of the corresponding heatmaps. The heatmap represents the probability of being top 5% highest fluctuated genes measured by rank change, divided by the probability of being top 5% lowest fluctuated genes for all genes in the indicated %Alu or %LINE-1 window. The Alu% gradient runs from 2% to 40% in 2% increments and the LINE-1 gradient runs from 1% to 20% in 1% increments.

Outiside of these ranges there tended to be few genes for the analysis. Grey boxes represent Alu% or LINE-1% windows that had too few genes for analysis (<200 genes).

The further dissect the pattern of variation, genes are clustered based on their rank change spectra derived from HapMap dataset. The distance between the rank change spectrum (RCS) of gene *i* and *j* was computed as (1 - c_{ij}) / 2, where c_{ij} is the Pearson's correlation coefficient between RCS_i and RCS_j. The genes were then clustered into four groups by using Partitioning Around Medoids (PAM) algorithm (*67*). The four clusters resembled the consolidation of four divisions of a gradient (Figure 3A). Each of the four clusters displayed significantly different Alu concentrations (Figure 3B) and significantly different LINE-1 concentrations (Figure 3C). Each of the four clusters furthermore displayed by a strikingly disparate median gene expression level with the gene cluster showing the least variation containing the highest average gene expression level (Figure 3D). We next examined the enriched Gene Ontology terms in biological process in each of the four clusters (*68*) and found that cluster displaying the lowest variation was dominated by genes involved in basic metabolism. As one moved to higher variation clusters, genes involved in development and in the function of multicellular organisms became increasingly prominent (Figure 3E).

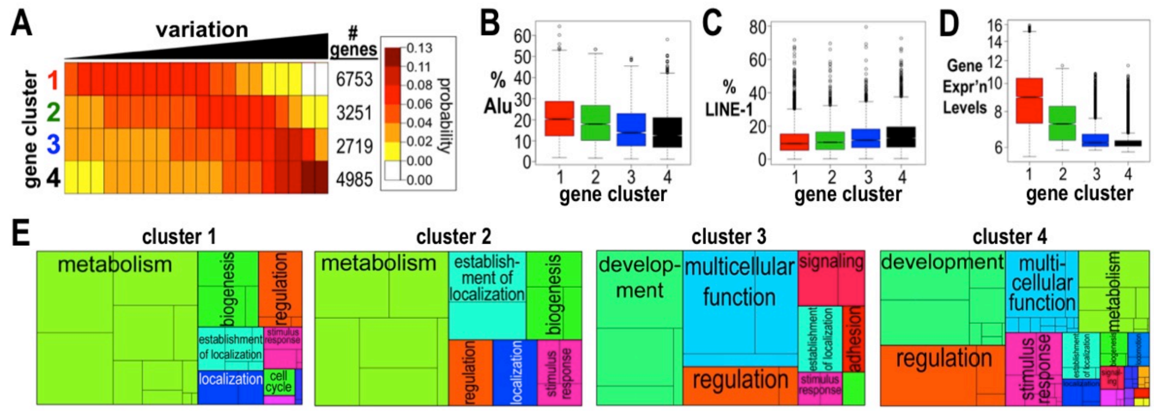


Figure 3

Rank change spectrums of genes. The rank variation of all gene expression in the HapMap dataset was divided into 20 equal increments (quantiles). A rank change spectrum (RCS) was then drawn up for each gene by calculating the probability of variation at each of the 20 quantiles. **A.** The expressed genes in the HapMap dataset clustered into four groups based on their RCS. The average probability of the genes of a given cluster displaying variation in each of the 20 quantiles is denoted using a heatmap with darker colors representing higher probabilities. **B.** Boxplots displaying the median and distribution of %Alu environments of the genes in each of the four gene clusters. %Alu was calculated from the 100-kb immediately upstream and 100-kb immediately downstream of the transcribed region of each gene. P values are clusters 1 vs. 2: 9.8e-19; 1 vs. 3: 3.08e-100; 1 vs. 4: 1.5e-221; 2 vs. 3: 7.8e-30; 2 vs. 4: 2.0e-72; 3 vs. 4: 2.2e-06. **C.** Boxplots displaying the median and distribution of %LINE-1 environments of the genes in each of the four gene clusters. P values are clusters 1 vs. 2: 1.0e-283; 1 vs. 3: 4.1e-105; 1 vs. 4: 2.8e-94; 2 vs. 3: 5.8e-56; 2 vs. 4: 3.8e-39; 3 vs. 4: 4.7e-07. **D.** Expression levels of the genes in four gene clusters. P values are clusters 1 vs. 2: 1.2e-297; 1 vs. 3: 0; 1 vs. 4: 0; 2 vs. 3: 1.1e-171; 2 vs. 4: 0; 3 vs. 4: 2.6e-13. All p values in parts B, C, and D were determined by two-sample Wilcoxon tests. **E.** Treemap of Gene Ontology Analysis of gene function in the four clusters. Each uniformly colored rectangle represents a group of genes with a shared classification according to the indicated function. Only the most prominent gene groups are labeled and subdivisions within rectangles bearing the same color represent functional subgroups.

The finding that gene variation profiles across human populations are associated with Alu and LINE-1 environments led us to wonder whether genes also displayed such trends across tissues. To address this, the rank change spectrum (RCS) of all expressed genes were compiled for the tissue datasets in Figure 2.

Specifically, for gene *i*, the 'RCS profile' in *T* tissues, $RCSP_i$, is defined as a *T* X 20 matrix where each row is the RCS_i in tissue *t*, where $t = 1, ..., T$. The RCS profile can therefore describes the variation of gene *i* in multiple tissues. If a gene is less fluctuated across multiple tissues, the RCS from different tissues should be similar with each other.

To quantitatively characterize genes' tissue-wise variation, we defined a tissue-wise variation score for each gene. Specifically, let K_i = cor($RCSP_i$,$RCSP_i$), which is the correlation matrix of rank change profile spectrum for gene *i* and K_0 is a *T* X *T* ideal consistency matrix that everywhere is one. The tissue-wise consistency score for gene *i*, A_i, is defined as:

$$A_i = 1 - \frac{\left\langle K_i, K_0 \right\rangle_F}{\sqrt{\left\langle K_i, K_i \right\rangle_F \left\langle K_0, K_0 \right\rangle_F}}$$

where $<.,.>_F$ is the Frobenius inner product (*69*). The score A can be seen as a distance score based on the cosine of the angle between K_i and K_0 in an appropriate space. In our situation, since both K_i and K_0 are always positive semidefinite, this score ranges between 0 and 1, from least to most fluctuated across tissues (Figure 4A).

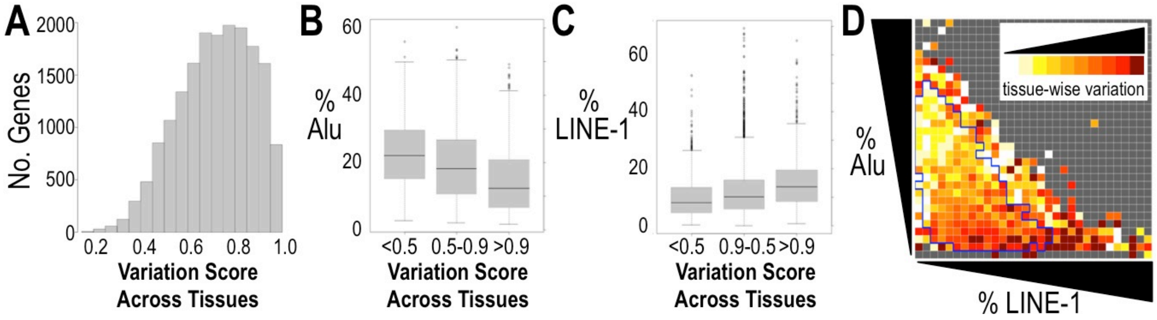


Figure 4

Gene variation across tissues correlated with %Alu and %LINE-1. A tissue-wise variation score was calculated for every gene. Only genes expressed in ≥70% of tissues were considered. All p values are determined by two-sample Wilcoxon tests. A. Distribution of genes across the tissue-wise variation spectrum that runs from 0 (no variation) to 1 (complete variation). B. %Alu in the 200-kb flanking regions of genes falling into three tissue-wise variation ranges (tissue-wise variation score smaller than 0.5, between 0.5 and 0.9, and greater than 0.9). P values are '<0.5' vs. '0.5-0.9': 2.5e-29; '<0.5' vs. '>0.9': 1.7e-113; '0.5-0.9' vs. '>0.9': 4.0e-87. **C.** %LINE-1 in the 200-kb flanking regions of genes falling into the indicated tissue-wise variation ranges. P values are '<0.5' vs. '0.5-0.9': 1.8e-23; '<0.5' vs. '>0.9': 1.6e-89; '0.5-0.9' vs. '>0.9': 7.1e-71. **D.** Color map of tissue-wise variation across the range of %Alu / %LINE-1 environments. %Alu and %LINE-1 increments are the same as in Figures 1H and 1I. The darker the color, the greater the tissue-wise variation.

The RCS profile of every gene expressed in ≥70% of tissues was compared across tissues. Increasing tissue-wise gene expression variation was strongly associated with progressively less %Alu in the 200-kb flanking region (Figure 4B) and progressively more %LINE-1 (Figure 4C). A heatmap of tissue-wise variation scores across a %Alu vs %LINE-1 spread revealed a strong progression toward higher variation in increasingly low Alu high LINE-1 gene group (Figure 4D).

We next asked whether the correlations of %Alu and %LINE-1 with gene expression variation display regional differences along chromosomes. The human genome was systematically scanned using a series of overlapping 10 million base pair windows that differed from one window to the next by 100-kb. In 10 million base pair windows containing ≥30 genes, correlation coefficients were calculated between coefficient of variation and %Alu, and between coefficient of variation and %LINE-1. The correlation coefficients were then assigned a color and mapped to the corresponding 100-kb location. A map correlating %Alu to gene expression variation in LCLs of four human populations (Caucasian, Chinese, Japanese and African) and four types of brain tissues (cerebellum, frontal cortex, pons, and temporal cortex) is shown in Figure 5A and corresponding map for %LINE-1 is shown in Figure 5B. The four LCL variation maps were highly similar to each other and the four brain maps were also quite similar to each other. However, the brain maps were remarkably different from the LCL maps. Although %Alu tended to be negatively correlated and %LINE-1 positively correlated with gene expression variation, the strength of these correlations varied tremendously throughout the genome. Surprisingly, the correlation were reversed from the usual trends in several regions with %Alu positively correlating with gene expression variation or %LINE-1 negatively correlating with expression variation or both. Strikingly, the strongest positively and negatively correlated regions tended to be tissue-specific suggesting that tissue-specific chromatin changes at the domain level alter the relationship of %Alu and %LINE-1 to gene expression variation. Reproducible %Alu and %LINE-1 variation correlation maps required high quality gene expression microarray data (not shown) and we consider RNA-seq data to be better suited for such maps.

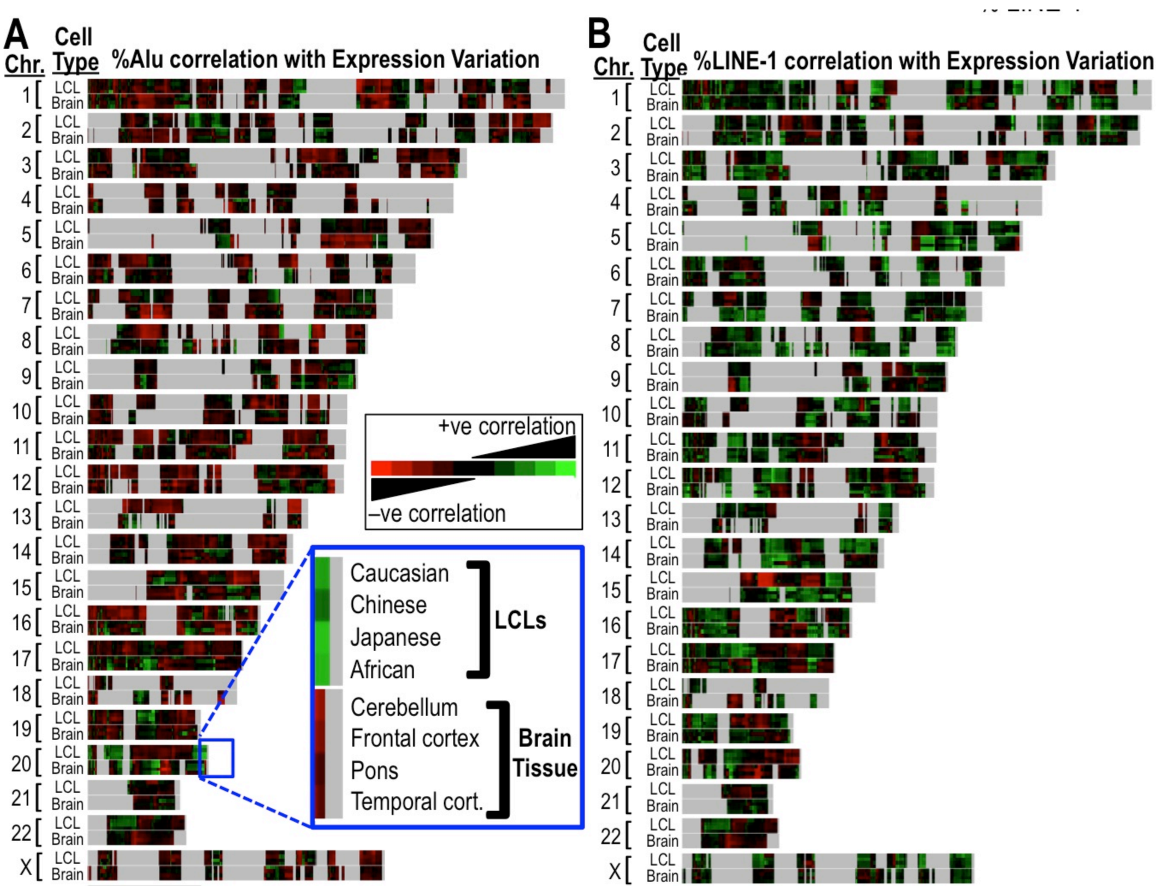


Figure 5

Correlation between 200-kb flanking Alu (panel A) and LINE-1 (panel B) concentrations with gene expression variation across the genome in LCLs and in brain tissue. Chromosomes are identified on the left. Each chromosome was systematically scanned using a series of overlapping 10 million base pair window that differed from one window to the next one by 200-kb. For each 10 million base pair window, correlation coefficients between gene expression variation and %Alu (panel A) or %LINE-1 (panel B) were calculated only if the window contained ≥30 genes. Correlation coefficients were assigned a color from a color map gradient running from bright green (strong positive correlation, correlation coefficient = 1) to black (no correlation = 0) to bright red (strong negative correlation, correlation coefficient = -1) and the color added to the corresponding 100-kb section. Regions representing windows with <30 genes are gray. Cell or tissue types, moving from the top to the bottom of each horizontal chromosome are: LCLs from 60 Caucasians (GSE6536), 45 Chinese (GSE6536), 45 Japanese (GSE6536), 60 Africans (GSE6536), 136 cerebellum samples (GSE15745), 146 frontal cortex samples (GSE15745), 145 pons samples (GSE15745), and 147 temporal cortex samples (GSE15745).

Finally, we used 210 HapMap individuals to examine the effect of common cis-SNPs and CNVs on the correlation between %Alu and %LINE-1 and gene expression variation. We wonder whether the correlation between gene expression variation and local abundance of Alu and LINE-1 still exists after excluding the variations that can be explained by common cis-SNPs and CNVs. The common SNPs and CNVs for each of 210 HapMap individual were obtained from HapMap public release #28 (*64*).

For each gene, we find the cis-SNPs that locate between 500 kb upstream and downstream of transcriptional start site. It has been shown that cis-SNPs have more reproducible association with expression QTL (*70*). The cis-SNPs that have more than 15% missing values in each group were removed, and we only considered cis-SNPs that have >5% minimum allele frequency (MAF) (*70*). A simple linear regression was performed on a cis-SNP and gene expression levels, and the cis-SNPs which raw p value is <0.00001 was considered as being significant (*70*). Similarly, to determine the significant CNVs, we also fit a linear model between expression value and each CNV. The CNV is codes as 0 (homozygous deletion), 1 (heterozygous deletion), 2 (wild type), 3 (heterozygous duplication) or 4 (homozygous duplication). The CNV which raw p value < 0.001 was considered to be significant.

For gene i, let y_i be the log-transformed and normalized expression values within a population group that has J individuals, n_i^{SNP} and n_i^{CNV} be the number of significantly associated cis-SNPs and CNVs, X_i^{SNP} (J X n_i^{SNP}) and X_i^{CNV} (J X n_i^{CNV}) be the explanatory matrix of cis-SNP and CNV. A linear model was used to model the relationship between expression levels and significant cis-SNPs and CNVs. We fit a linear model:

$$y_i = \mu_i + \beta_i^{SNP} X_i^{SNP} + \beta_i^{CNV} X_i^{CNV} + \varepsilon_i,$$

where μ_i is the intercept and ε_i is the i.i.d. noise term. Let $\hat{y}_i$ be the estimated expression levels by above linear model, and $y_{0i} = y_i - \hat{y}_i$ is therefore the residual expression that cannot be explained by SNP and CNV effects, and σ_{0i} be the standard deviation y_{0i}. We defined the coefficient of residual gene expression variation (GEV), $\eta_{0i} = \sigma_{0i} / \mu_i$, which is a robust and unbiased estimate of the amount of GEV that cannot be explained by SNP and CNV effects (*59*).

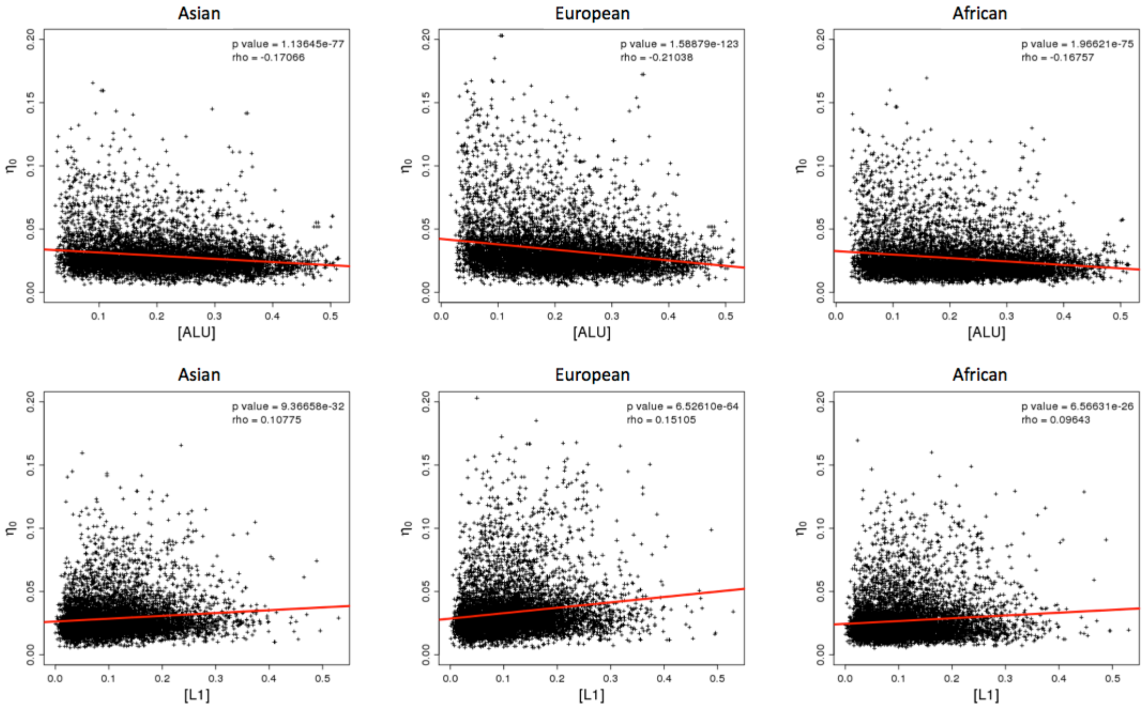


Figure 6.

Correlation between coefficient of residual gene expression variation and %Alu (top row), and %LINE-1 (bottom row) in Asian (left column), European (middle column) and African groups (right column).

Figure 6 suggested in three population groups (Asian, European and African), the coefficient of residual GEV (η_0) is significantly positively correlated with %LINE-1, and negatively correlated with %Alu (p value < 1e-16). This suggests that the strong association between gene expression variation and %Alu and %LINE-1 is not due to the effect of common SNPs and CNVs.

Taken together, our results show that, generally, genes with low Alu high LINE-1 environment show low gene expression variation, while the genes with high Alu low LINE-1 environment show high gene expression variation. Such associations do not only exist within individual tissues and cell types, they also exist between tissue and cell types. The relationship between gene expression variation and Alu/LINE-1 environment looks also independent of the effect of common SNPs and CNVs.

Characterizing the histone modifications of Alu and LINE-1

Histone modifications of mobile elements have been studied at the global level in mice where the majority of LINE-1 and Alu lacked all of the histone modifications queried (*71*). No genome-wide studies on the histone modifications of Alu and LINE-1 elements could be found for humans so it was first necessary to characterize the histone modifications of these elements. We used the histone modifications for LCLs (GM12878) from the ENCODE web site (http://genome.ucsc.edu/ENCODE/). The ENCODE data for LCLs (GM12878) is in the highest priority set in ENCODE project and thus have the most comprehensive histone modification data (*72*). For each type of histone modification for LCLs, we used the 'BroadPeak' file downloaded from ENCODE web site. The BroadPeak files contain enriched regions estimated by the peak-calling program. Whereas a chromosomal region is covered by BroadPeak file, the corresponding histone modification is considered to be present in this region (*72*).

W found that Alu and LINE-1 elements are associated with a variety of histone modification, the most common being H3K9me3, H3K27me3, and H3K36me3 (Figure 7A). H3K36me3 was the most common mark for Alu where it was more than twice as than at LINE-1 elements (Figure 7A). An exception was the 200-kb flanking regions of expressed genes where H3K36me3 was the most common histone modification for both Alu and LINE-1 (Figure 7B, left). The association of H3K36me3 with expressed genes therefore extends beyond the transcribed regions. H3K9me3 were more plentiful over LINE-1 than Alu and elevated at LINE-1 in the 200-kb flanking regions of expressed genes (Figure 7B, left). Maps of randomly selected 10-Mb chromosomal regions showed that the genome consists of alternating stretches of H3K9me3, H3K27me3, and H3K36me3 regions but many gene desert regions contained none of these three histone modifications (Figure 7C and 7D). Domain size ranged from one or a few nucleosomes to megabase regions. The promotors of expressed genes resided mostly in H3K36me3 domains while H3K9me3, H3K27me3, and H3K36me3 all contained unexpressed promotors (Figure 7C and 7D). Strikingly, the histone modification of Alu and LINE-1 elements corresponded almost perfectly with histone modification of the domain they resided in (Figure 7C). Both Alu and LINE-1 were dominated by H3K36me3 in high Alu low LINE-1 regions (Figure 7E, left), by H3K27me3 in low Alu low LINE-1 regions (Figure 7E, middle) and by H3K9me3 in low Alu high LINE-1 regions (Figure 7D, right). The exceptions were the medium-to-high LINE-1 regions with extremely low Alu (Figure 7D left): these were dominated by none of the three aforementioned histone modifications (Figure 7C and 7E). The association of H3K36me3 with high Alu concentrations and H3K9me3 with high LINE-1 densities suggested that Alu attracts H3K36me3 and LINE-1 attracts H3K9me3 and that local prevalence of Alu or LINE-1 can cause H3K36me3 or H3K9me3 to 'win' a domain. If neither Alu nor LINE-1 sequence is abundant, H3K27me3 may win.

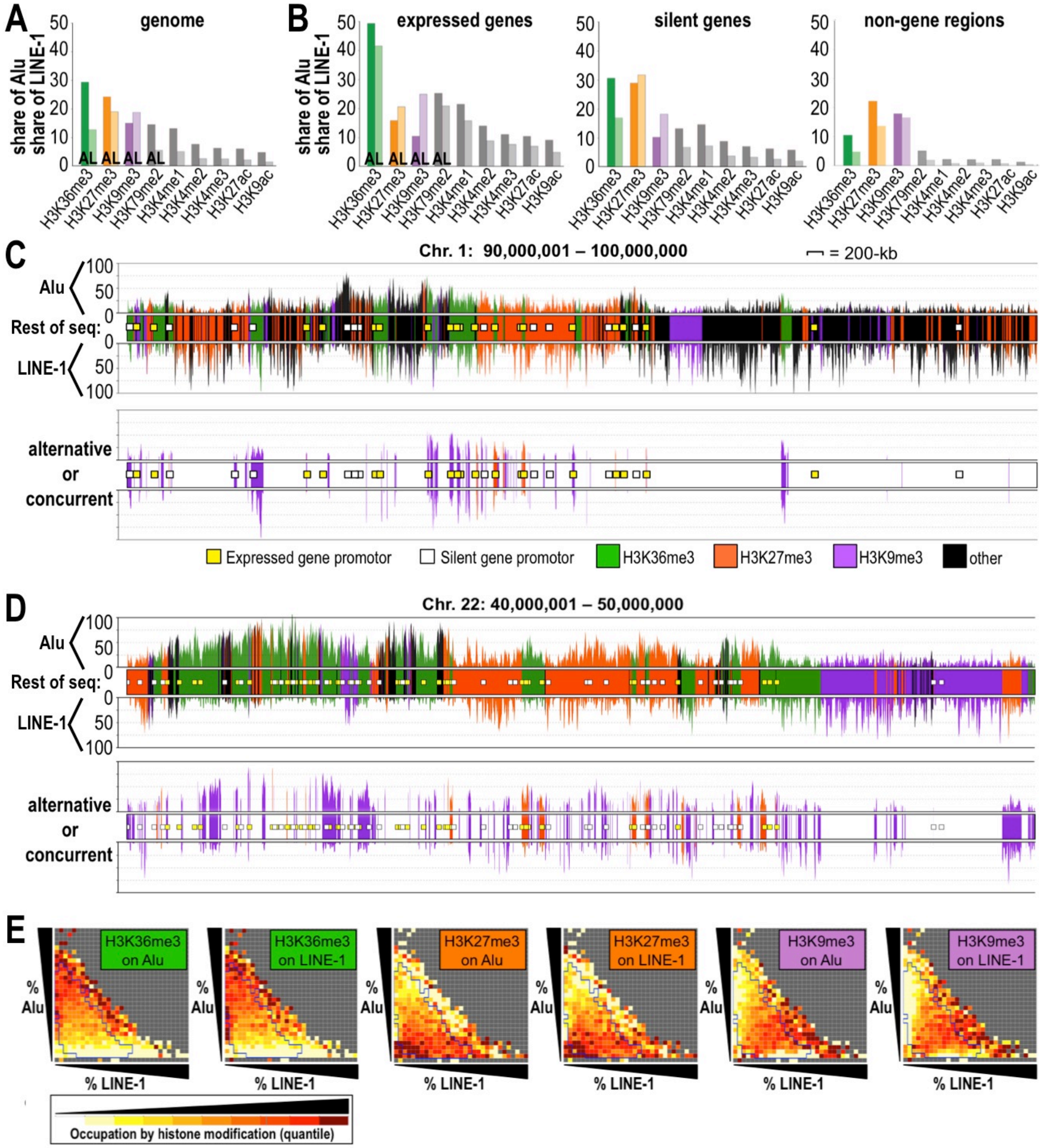


Figure 7

Histone modifications of Alu and LINE-1 elements in LCLs. A, B. Distribution of histone modifications for Alu (left bars) and LINE-1 (right bars) transposons in the human genome. Y-axes are the percent of the indicated transposon overlapped with the histone modification indicated on the X-axis. **A.** Histone modifications across the entire human genome. **B.** Histone modification in the 200-kb regions flanking the transcribed regions of expressed genes (left) and unexpressed genes (enter) and histone modifications in non-gene regions (right). Non-gene regions are defined as the subset of the genome that is >100-kb from any gene transcription start site. **C.** Map of H3K36me3 (green), H3K27me3 (orange), and H3K9me3 (purple) in a randomly selected chromosomal region (positions 90,000,001 - 100,000,000 of chromosome 1) using data from the ENCODE project (*73*). The upper map identifies these histone modifications chromosomal domains across the region and the lower map identifies second histone modifications in certain regions. It is not known whether two of the aforementioned histone modifications are simultaneously present in these regions or whether they represent alternative chromatin structures. Regions with all three histone modifications were extremely rare and therefore ignored. The center of each map identifies the histone modification in 3-kb rolling windows. Black regions contain none of these three histone modifications. The promotors of expressed genes are identified as yellow squares and the promotors of non-expressed genes are white squares. The bar graph aligned with and above the map of the chromosomal region displays the abundance of Alu sequence in rolling 3-kb windows. The downward facing graph below the map of the chromosome is a bar graph displaying the abundance of LINE-1 sequence in rolling 3-kb windows. The color of each bar in the Alu and LINE-1 graphs denote the histone modifications of the Alu or LINE-1 sequence in the window represented by the bar. Note that the histone modification of the Alu or LINE-1 sequence almost always matches the histone modification of the domain. The ChIP-seq data uses presence/absence calls and therefore the relative abundance of two histone modifications is not known. **D.** Map of h3K36me3, H3K27me3, and H3K9me3 in a randomly selected chromosomal region (positions 40,000,001 - 50,000,000 of chromosome 22) using data from the ENCODE project. This map is at the same scale and was derived in the same manner as the map in C. **E.** 2-D heatmaps showing the prevalence of the indicated histone modifications on Alu or LINE-1 elements in the 200-kb region flanking the transcribed regions of genes. The genes are sorted according to the %Alu and %LINE-1 in their 200-kb flanking regions using the same %Alu and %LINE-1 scales as in Figure 1H and I.

Alu/LINE-1 environment is a determinant for whether gene expression promotes H3K36me3 or H3K9me3 in the transcribed region.

Since gene expression influences gene expression variation, we examined the relationship between gene expression level and H3K36me3, H3K27me3 and H3K9me3 levels. Genes were sorted by expression level (X-axis) and by the %Alu or %LINE-1 in the 200-kb flanking region (Y-axis) and the differences in H3K36me3, H3K27me3 and H3K9me3 levels across these groups were mapped using quantiles. As one moved from lower to higher expressed genes, H3K36me3 modestly increased in the 200-kb regions flanking high Alu low LINE-1 genes (Figure 8A, left), while H3K27me3 decreased (Figure 8A, center). Surprisingly, H3K9me3 also increased in the 200-kb regions with increased expression level, but only at low Alu high LINE-1 genes (Figure 8A, right). The transcribed regions of genes followed the same trends as the 200-kb regions (Figure 8B), except that H3K36me3 showed a much stronger gradient and increased across all Alu and LINE-1 environments with the most striking increase among high LINE-1 genes (Figure 8B, right). High H3K36me3 levels in the transcribed regions suppress spurious transcription and our data suggested that H3K9me3 may also contribute to this role in high LINE-1 low Alu genes.

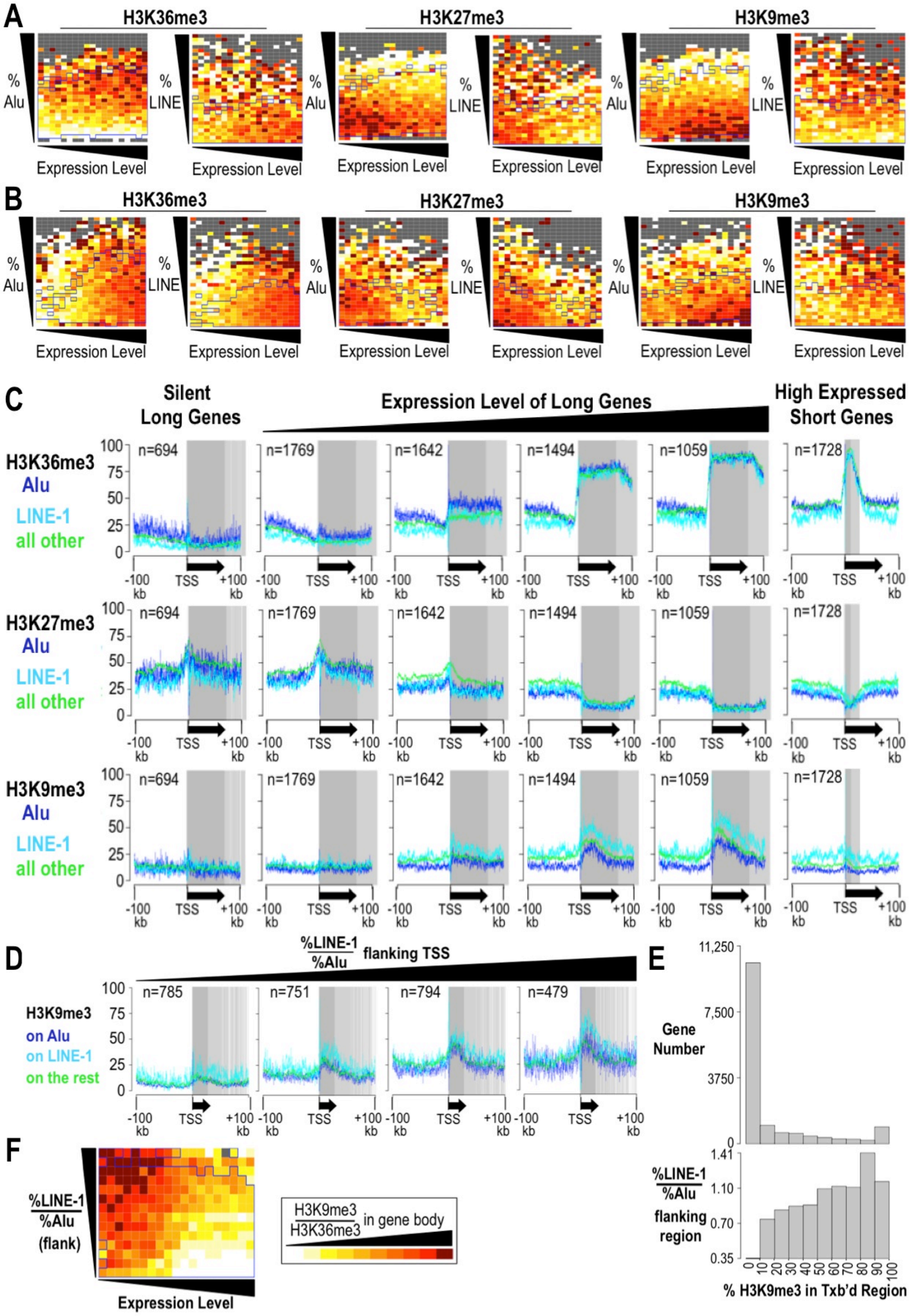

A
H3K36me3
H3K27me3
H3K9me3
% Alu
% LINE
Expression Level
B
H3K36me3
H3K27me3
H3K9me3
% Alu
% LINE
Expression Level
C
Silent Long Genes
Expression Level of Long Genes
High Expressed Short Genes
H3K36me3
Alu
LINE-1
all other
H3K27me3
Alu
LINE-1
all other
H3K9me3
Alu
LINE-1
all other
n=694
n=1769
n=1642
n=1494
n=1059
n=1728
100
75
50
25
0
-100 kb
TSS
+100 kb
D
%LINE-1 / %Alu flanking TSS
H3K9me3
on Alu
on LINE-1
on the rest
n=785
n=751
n=794
n=479
100
75
50
25
0
-100 kb
TSS
+100 kb
E
Gene Number
11,250
7,500
3750
0
%LINE-1 / %Alu flanking region
1.41
1.10
0.70
0.35
0
10
20
30
40
50
60
70
80
90
100
% H3K9me3 in Txb'd Region
F
%LINE-1 / %Alu (flank)
Expression Level
H3K9me3 / H3K36me3 in gene body

Figure 8

Relationship of H3K36me3, H3K27me3, and H3K9me3 to level of gene expression in different %Alu and %LINE-1 environment. **A.** 2-D heatmaps showing the prevalence of the indicated histone modifications in the 200-kb regions flanking the transcribed regions of genes. These genes are sorted according to the %Alu or %LINE-1 in their 200-kb flanking regions (Y-axis), and by gene expression level quantile (X-axis). **B.** 2-D heatmaps showing the prevalence of the indicated histone modifications in the transcribed regions of genes. The genes are sorted according to the %Alu or %LINE-1 in their 200-kb flanking regions on the Y-axis, and by gene expression level quantile on the X-axis. **C.** Levels of the indicated histone modifications across a region extending from 100-kb upstream of the TSS to 100-kb downstream of the TSS. Each graph represents a group of genes that is aligned at their TSSs and is transcribed from left to right. Histone modification data is reported for H3K36me3 (top row of graphs), H3K9me3 (middle row), and H3K27me3 (bottom row). The genes category represented by a graph is listed above the graphs and the number of genes represented by each graph is listed in the top left corner of the corresponding graph. Each graph represents either long (> 69707-kb) or short (<9730-kb) genes as indicated. The region transcribed by all of the genes in a graph is given a darker grey background and the region transcribed by only a subset of genes is lighter gray. **D.** Levels of H3K9me3 across a 200-kb region flanking the TSS in genes sorted according to %LINE-1 / %Alu in the 200-kb region flanking the TSS. Graphs are organized as in panel C above. **E.** Top panel: bar graph showing the number genes in the LCL line with the indicated level of H3K9me3 in their transcribed regions. Bottom panel: bar graph showing the average %LINE-1 / %Alu in the 200-kb regions flanking genes with the indicated level of H3K9me3 in their transcribed region. **F.** 2-D heatmap showing the distribution of the ratio of concentration of H3K9me3 to H3K36me3 on gene body. The genes are sorted according to the %LINE-1 / %Alu ratio in their 200-kb flanking regions (Y-axis), and by gene expression level quantile (X-axis).

To see how the changes in H3K36me3, H3K27me3 and H3K9me3 were distributed across gene regions and to see more quantitative data, we graphed the level of these histone modifications across the gene region spanning from 100-kb upstream of transcription start sites (TSS) to 100-kb downstream of the TSS. Genes were sorted according to the transcript length and expression level and each pooled set of genes aligned at their transcription start sites (TSSs) with transcription oriented from left to right. H3K36me3 levels increased in both the gene body and in the flanking regions as one moved from silent genes to increasingly higher expressed genes with H3K36me3 on Alu until expression was high; however, the rate of increase of H3K36me3 was much steeper in the transcribed region than in the flanking region (top row of Figure 8C). As previously reported (*53*), the level of H3K36me3 increased as one moved along the length of the transcribed region (top row of Figure 8C). As expected, H3K27me3 displayed the opposite trend, with the transcribed regions of high expressed genes displaying greater H3K27me3 suppression than the flanking region (middle row of Figure 8C). H3K9me3 increased in both the gene body and in the flanking regions as one moved from low to high expressed genes with H3K9me3 on LINE-1 sequence leading the way; as with H3K36me3, the transcribed region showed a greater H3K9me3 increase with increased expression level than the flanking regions (Figure 8C). In contrast to H3K36me3, H3K9me3 levels progressively decreased as one moved from the 5' end to the 3' end of the transcribed region. This raised the possibility that a subset of genes is enriched for H3K9me3 in the 5' end of the transcribed region that the H3K9me3 is progressively replaced by H3K36me3 as one moved to the 3' end of the transcribed region. H3K9me3 displayed a more modest increase with expression level compared to H3K36me3 and this was presumably due to the H3K9me3 increase being restricted to high LINE-1 low Alu genes (Figure 8A and 8B). Sorting longer high expressed genes according to %LINE-1 / %Alu ratio of the 200-kb flanking region revealed the highest H3K9me3 level in the transcribed region of the genes with the highest LINE-1-to-Alu ratio of their flanking regions (Figure 8D). Sorting genes according to the level of H3K9me3 in their transcribed regions revealed that ~950 expressed genes where >90%

H3K9me3 in their transcribed regions while in a few thousand additional genes had 10-90% of the transcribed regions occupied by H3K9me3 (Figure 8E, upper). The %LINE-1 / %Alu ratio of the 200-kb flanking regions of these sorted genes increased as one considered genes with greater H3K9me3 representation in the transcribed region (Figure 8E, lower). When both expressed and unexpressed genes were considered, gene number were roughly 1.5 fold higher but the relationship of %LINE-1/%Alu to H3K9me3 content in the transcribed region was nearly the same. Although the %LINE-1 / %Alu ratio of the 200-kb flanking regions of these sorted genes appeared to identify threshold of ~0.7 in the flanking region above witch H3K9me3 started to appear in the transcribed regions of genes (Figure 8E), closer examination revealed a more gradual progression that varied with gene expression level (Figure 8F).

Silent genes occupied by H3K36me3, H3K27me3, or H3K9me3 according to Alu/LINE-1 environments

The finding that gene expression promotes H3K36me3 or H3K9me3 depending on the LINE-1/Alu environment raised the possibility that gene expression was responsible for the presence of H3K36me3 or H3K9me3 in the flanking regions. To address this, we mapped the histone modifications associated with silent genes. This revealed essentially the same pattern as for expressed genes: H3K36me3 was associated with silent genes in residing in high Alu low LINE-1 regions, H3K27me3 was concentrated in the flanking regions of silent low Alu low LINE-1 genes, and H3K9me3 was most concentrated in the 200-kb flanking regions of medium Alu high LINE-1 genes (Figure 9A). This raised the possibility that the Alu/LINE-1 environment of a gene might be a determinant for the identity of the repressive mark occupying the promotors and TSSs of silent genes. A Venn diagram was used to show the distribution of repressive marks across a 3-kb region flanking the TSS of genes in LCL cells, extending from 2-kb upstream of the TSS to 1-kb downstream. This promotor/TSS region was occupied exclusively by H3K27me3 at 5141 silent genes, by H3K36me3 at 1208 silent genes, and only by H3K9me3 at 579 unexpressed genes (Figure 9B). Strikingly, an additional 1332 silent genes had 3-kb regions of associated with both H3K27me3 and H3K9me3. 2-D Alu vs. LINE-1 color maps revealed that the three repressive marks genes followed the same trends as in the 200-kb flanking regions: H3K36me3 occupied the 3-kb at the promotor and TSS of high Alu low LINE-1 genes, H3K27me3 was most concentrated in low Alu low LINE-1 genes, and the H3K9me3 repressive mark occupied the 3-kb regions in genes residing in mid-to-low Alu high LINE-1 regions. Occupancy of these marks in 500-bp regions flanking the TSS produced nearly identical maps. Finally, the 3-kb regions 5412 silent genes were occupied by none of these marks and these genes were distributed across essentially the entire Alu/LINE-1 spectrum (Figure 9D), with the highest concentrations residing in low Alu mid-LINE-1 regions where we had hypothesized a fourth not-yet identified repressive mark to reside.

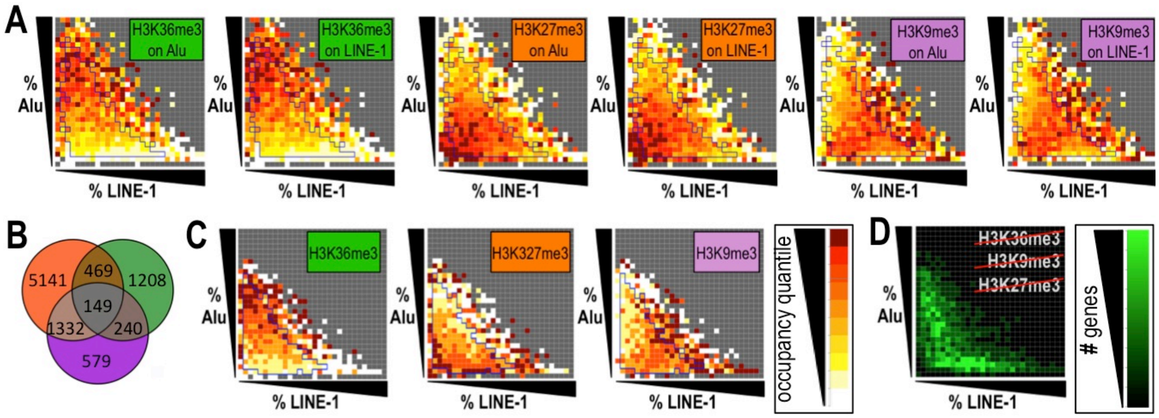


Figure 9

H3K36me3, H3K27me3, and H3K9me3 occupancy at silent genes. **A.** Heatmaps showing the distribution of H3K36me3, H3K27me3, and H3K9me3 on Alu or LINE-1 elements in the 200-kb regions flanking the transcription start sites of unexpressed genes. Silent genes were sorted according to Alu and LINE-1 abundance in the 200-kb flanking regions and, for each histone modification, ten colors representing ten quantile of occupancy a re used (same color scale as shown in panel C). **B.** Venn diagram showing the numbers of genes associated with H3K36me3 (green), H3K27me3 (orange), H3K9me3 (purple), or by more than one of these histone modifications at 3-kb regions extend from 2-kb upstream of the transcription start sites (TSSs) of genes to 1-kb downstream of the TSSs. **C.** Heatmaps showing the distribution of H3K36me3, H3K27me3, and H3K9me3 at 3-kb regions extending from 2-kb upstream of the TSSs of genes to 1-kb downstream of the TSSs. Genes were sorted according to Alu and LINE-1 abundance in the 200-kb flanking regions. **D.** Heatmap showing unexpressed genes whose 3-kb regions associated with none of the three histone modifications: H3K36me3, H3K27me3, or H3K9me3. These silent genes were sorted according to the %Alu and %LINE-1 in their 200-kb flanking regions.

Specificity of Alu for H3K36me3 and LINE-1 for H3K9me3

The association of high Alu low LINE-1 environments with H3K36me3 and low Alu high LINE-1 environments with H3K9me3 led us to hypothesize that Alu elements suppresses gene expression variation by promoting H3K36me3 and LINE-1 promotes variation by promoting H3K9me3.

As a first step towards testing this hypothesis, we derived restriction/exclusion maps at 1-kb resolution across 200-kb regions centered on the TSSs of genes to see if H3K36me3 preferentially mapped to Alu elements and H3K9me3 preferentially mapped to LINE-1 elements. Genes of interest were aligned at their TSSs and oriented so that they were all transcribed from left to right (Figure 10A, top). Large groups of genes were used to ensure that every 1-kb increment across the lined up 200-kb regions was represented by Alu sequence in a subset of genes and by LINE-1 sequence in another subset of genes (Figure 10A, top). For each of the 2000 100bp increments, we calculated the proportion of Alu and LINE-1 sequence that was associated with each histone modification and proportion of non-Alu and non-LINE-1 sequence that was associated with the same histone marks. The significance of restriction or exclusion was determined by Fisher's exact test with the alternative hypothesis 'greater' or 'less'. Each 1-kb increment was then assigned a color from a gradient (Figure 10A, bottom) that runs from bright green (histone modification strongly favors Alu or LINE-1 sequence, Fisher's exact test p value < 0.00001 under the alternative hypothesis 'greater') to black (no preference for transposon, Fisher's exact test p value > 0.05 under either alternative hypothesis) to bright red (histone modification strongly exclude from Alu or LINE-1, Fisher's exact test p value < 0.00001 under the alternative hypothesis 'less'). In the hypothetical 'Gene Group 1' occupying the left half of Figure 10A, H3K9me3 resides exclusively on LINE-1 elements (black rectangle) except near the TSS where these is no preference so the resulting heatmap for H3K9me3 is green except at the TSS where it is black. In the hypothetical 'Gene Group 2' in the right half of Figure 10A, there is much less H3K9me3 on Alu elements (yellow squares) than on non-Alu sequence throughout the 200-kb except directly the TSS so the Group 2 restriction/exclusion heatmap for Alu is red except at the TSS where it is black.

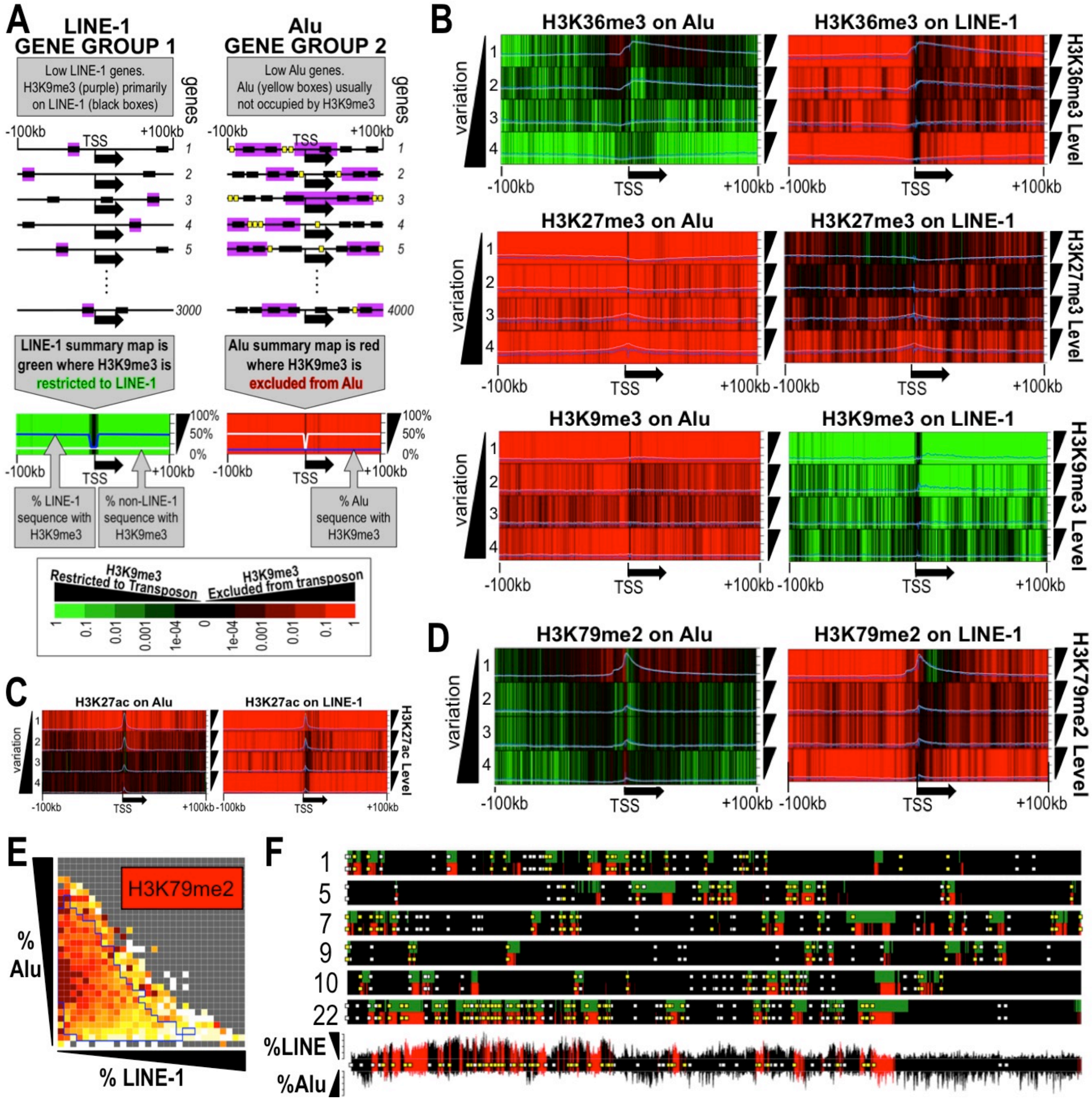


Figure 10

Heatmaps localizing the restriction/exclusion of histone modifications to/from Alu and LINE-1 elements across 200-kb regions flanking the TSSs of genes. **A.** Explanation of how the heatmaps in this figure were obtained. Each heatmap reports histone modification data from a group of genes that are aligned at their TSSs and are transcribed from left to right. Histone modification data is reported in 1-kb increments across a region extending from 100-kb upstream of the TSS to 100-kb downstream of the TSS. Each heatmap represents a large number of genes, and every 1-kb increment consequently is occupied by Alu (or LINE-1) sequence in a subset of genes and not occupied by Alu or LINE-1 sequence in the remaining genes. Each 1-kb increment derives its color from the gradient shown at the bottom. The color gradient (shown at bottom) runs from bright green (histone modification exclusively at transposon sequence) to black (no preference for transposon) to bright red (histone modification completely excluded from transposon). Superimposed on the heatmaps are line graphs charting the proportion of the transposon occupied by the histone modification (blue line) and the proportion of all remaining sequence occupied by the histone modification (white line). *Left example.* Gene group 1 consists of 3000 genes that have low numbers of LINE-1 elements (black boxes) in their 200-kb regions. H3K9me3 (purple) is sparse and occurs almost exclusively at LINE-1 sequence. The restriction of H3K9me3 to LINE-1 sequence in every 1-kb increment except very close to the TSS causes the graph to be green. Since ~50% of the LINE-1 sequence is occupied by H3K9me3, the superimposed blue line is at 50%. Since only ~5% of the non-LINE-1 sequence is occupied by H3K9me3, the superimposed white line is at 5%. *Right example.* Gene group 2 consists of 4000 genes with high numbers of LINE-1 elements and low numbers of Alu in their 200-kb regions. Although many of these Alu elements are associated with H3K9me3, the non-H3K9me3 regions are overwhelmingly Alu sequence except very close to the TSS where H3K9me3 is missing from both Alu and non-Alu sequence. The entire 200-kb except for the TSS is therefore bright red. **B.** Restriction/exclusion heatmap sets for H3K36me3 (top), H3K37me3 (middle) and H3K9me3 (bottom). Each histone modification has a restriction/exclusion heatmap set for Alu (left) and LINE-1 (right) and each set consists of consists of four heatmaps corresponding to the four gene expression level variation clusters (defined in Figure 3) stacked on top of each other in order of least variation (top) to greatest variation (bottom). Note that Alu shows specificity for H3K36me3 while LINE-1 shows specificity for H3K9me3. **C.** Restriction/exclusion heatmap sets for H3K27ac organized like the sets in panel *B*. **D.** Restriction/exclusion heatmap sets for H3K79me2 organized like the sets in panel *B*. **E.** 2-D color maps showing the prevalence of H3K79me2 in the 200-kb regions flanking the transcribed regions of genes sorted according to the %Alu and %LINE-1 in their 200-kb flanking regions. Occupancy color scale uses quantiles as in Figure 9C.

H3K36me3, H3K27me3, and H3K9me3 restriction/exclusion heatmaps were derived for each of the four variation gene clusters (Figure 3A) and the four maps were stacked on top of each other in order of lowest variation cluster map (top) to highest variation cluster map (bottom) with the Alu stack on the left and LINE-1 stack on the right (Figure 10B). As expected, H3K36me3 displayed restriction to Alu in all four variation clusters but this ran from modest restriction in the lowest variation genes to severe restriction in the highest variation genes (Figure 10B, top left). Since higher variation was associated with low Alu concentrations, severe restriction of H3K36me3 to Alu sequence was a feature of low Alu abundance. Line graphs showing the level of H3K36me3 on Alu (blue line) and h3K36me3 on non-Alu sequence (white line) were superimposed on the heatmaps; these revealed that as Alu concentrations increased, H3K36me3 increased on both Alu and non-Alu sequence but at a slightly higher rate on non-Alu sequence (Figure 10B, top left). Alu elements, therefore, showed visible specificity for H3K36me3 at low concentrations and the higher the Alu abundance the greater the overall level of H3K36me3 and the greater the spread of H3K36me3 beyond Alu elements into non-Alu sequence. A similar relationship was observed between H3K9me3 and LINE-1 where restriction of H3K9me3 to LINE-1 was observed across variation groups (Figure 10B, lower right), and the higher the LINE-1 concentrations (and the higher the variation group) the more abundance the overall H3K9me3 and the less H3K9me3 was limited to LINE-1 sequence (Figure 10B, lower right). LINE-1 therefore appeared to attract H3K9me3 and the higher the LINE-1 concentrations, the greater the spread of H3K9me3 beyond the confines of LINE-1 elements. The increased spread of H3K36me3 beyond Alu elements and H3K9me3 beyond LINE-1 elements at higher transposon concentrations slightly outstripped the increased occupancy of the histone modifications at the elements themselves resulting in less green and more black. Consistent with Alu specificity for H3K36me3 and LINE-1 specificity for H3K9me3, Alu excluded H3K27me3 and H3K9me3 at all concentrations and LINE-1 excluded H3K36me3 and H3K27me3 (Figure 10B).

Restriction / exclusion heatmaps were also derived for several additional histone modifications. H3K27ac displayed very similar patterns: very low levels of histone modifications throughout the 200-kb region except for a peak centered on the TSS. Both Alu and LINE-1 excluded these histone modifications except close to the TSS and with Alu showing weaker exclusion in higher variation groups (Figure 10C). H3K4me3, H3K4me2, and H3K9ac all had restriction / exclusion that were highly similar to H3K27ac (not shown). Unexpectedly, H3K79me2 was represented throughout the 200-kb regions, was more abundant in lower variation groups, and peaked at the TSSs (Figure 10D). H3K79me2 showed restriction to Alu except in the lowest variation group where H3K79m2 was most abundant and showed modest repulsion (Figure 8D). Mapping H3K79m2 to the 200-kb flanking regions of genes revealed a pattern similar to H3K36me3 but with the highest concentration perhaps a little lower on the Alu scale (Figure 10E). Mapping H3K79m2 to the six randomly selected 10-Mb regions that we previously considered for H3K36me3, H3K27me3, and H3K9me3 revealed a pattern that closely mirrored H3K36me3 but usually with smaller domains nested within the larger H3K36me3 domains (Figure 10F, top). Domains occupied by H3K79m2 included this histone modification on both Alu and LINE-1 elements (Figure 10F, bottom). H3K79m2 therefore appeared to be related to H3K36me3 and may play a role in gene expression variation.

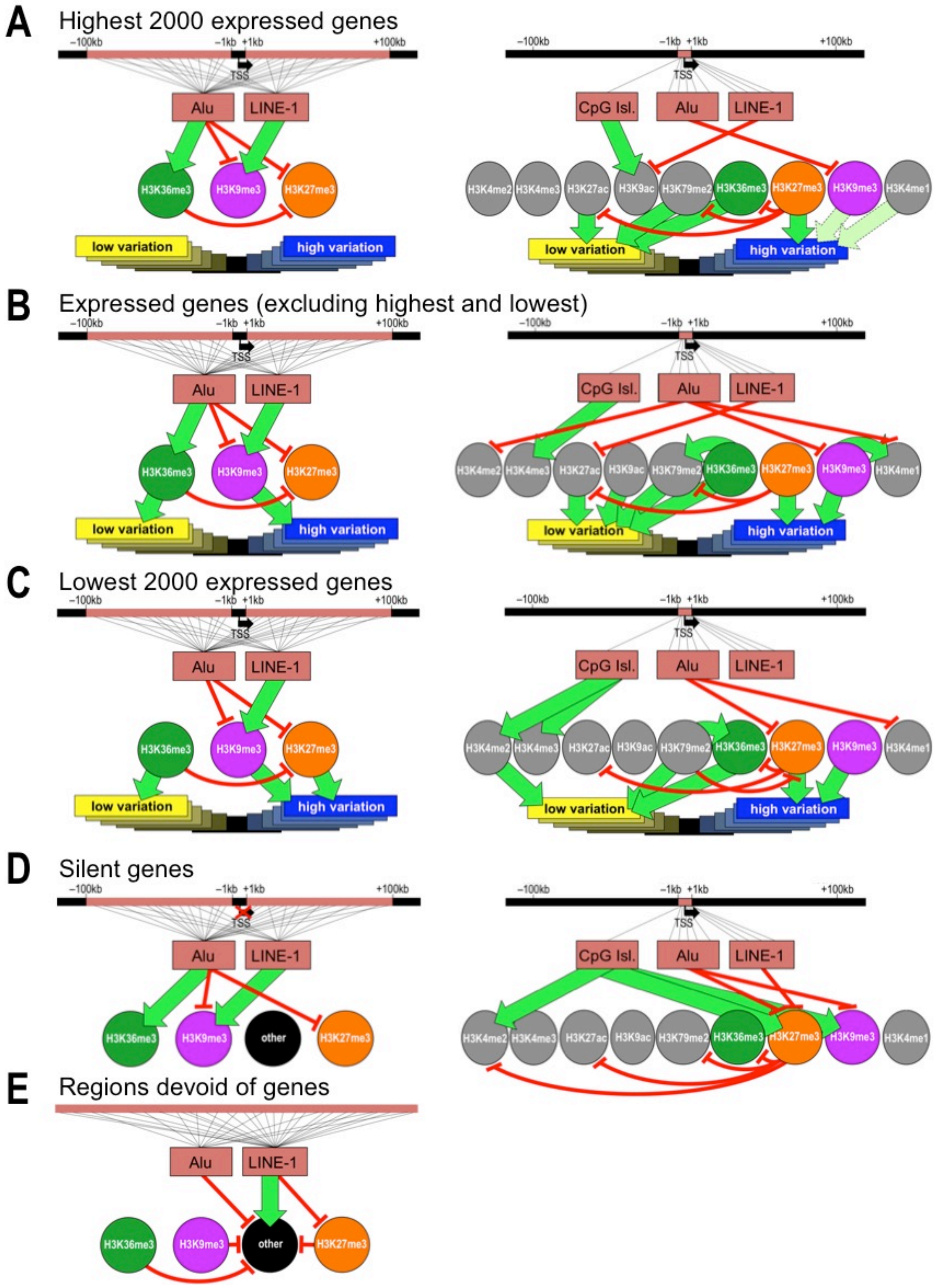
Bayesian network inferring the causality between DNA sequence, histone modifications and gene expression variation

The aforementioned data suggested that Alu may reduce gene expression variation via H3K36me3 and that LINE-1 may promote variation via H3K9me3. To determine whether this is the case, we turned to Bayesian network that are by far the most established and widely used method for inferring causality in large datasets.

We tried to build a model to infer the causal relationship between Alu / LINE-1 concentrations, CpG island, various histone modifications, and gene expression variation in a Bayesian network, where for each gene, Alu / LINE-1 concentrations within a chromosomal region, occurrence of histone modifications and gene expression variation (as coefficient of variation) are treated as an observable event.

We used the modeling strategy described by (*74*). Specifically, to ensure the robustness of the Bayesian network, genes were randomly partitioned into 10 non-overlapping groups, where each nine groups combination was used to train a Bayesian network (leave one out). All the undirected edges were removed from each inferred network, and only directed edges were kept. The common network in which the directed edges were agreed by at least seven of Bayesian networks derived from each of nine group combinations were chosen as the derived Bayesian network at this round. The randomly grouping was performed 100 times and at each time, we derived a common Bayesian network from the

LOO procedure. The final Bayesian network contained the directed edges that were agreed by at least 70 common Bayesian networks estimated by each LOO procedure. When building the Bayesian network, certain edges that do not have biological meaning were excluded, such as edges from gene expression variation to all other nodes, edges from histone modifications to Alu/LINE-1/CpG island, edges from Alu/LINE-1/CpG island directly to gene expression variation and edges within Alu/LINE-1/CpG island. We used Grow-Shrink (GS) method (*75*) from bnlearn package (http://www.bnlearn.com/) in R (http://www.r-project.org) for estimating the network structure.



A Highest 2000 expressed genes

B Expressed genes (excluding highest and lowest)

C Lowest 2000 expressed genes

D Silent genes

E Regions devoid of genes

Figure 11

Bayesian networks inferring causality between DNA sequence, histone modification, and gene expression variation. Bayesian networks were derived independently for the 2-kb regions flanking the TSSs of genes and for the remaining 198-kb sequence flanking the TSS. Bayesian models for **A.** high expressed genes, **B.** medium expressed genes, **C.** low expressed genes, **D.** transcriptionally silent genes, **E.** non-gene regions. Chromosomal maps showing the region whose DNA sequences were used to derive the networks are shown at the top. Alu, LINE-1 and CpG islands (reddish rectangles) influence the appearance of histone modifications (circles and ovals) some of which in turn promote low variation (yellow rectangles) or high variation (blue rectangles). Note that the 'Other' chromatin category was only included when deriving Bayesian networks for silent genes and non-gene regions.

Bayesian network modeling were applied to several size windows surrounding genes and we concluded that the data could most clearly be summarized we separately derived networks for the 2-kb regions flanking the TSS and for 198-kb region that flanking the TSS of genes and skips the 2-kb around the TSS. Moderately enlarging the window around the TSS did not produce different results (data not shown). Because the correlations of %Alu and %LINE-1 with gene expression variation appeared to be weaker for the highest and the lowest expressed genes (e.g. see Figure 1J-M), Bayesian networks were derived separately for these genes. Bayesian network derived from the main set of genes (all but the 2000 highest and 2000 lowest expressed genes) revealed that, in the 198-kb region flanking the 2-kb around the TSS, Alu promoted the appearance of H3K36me3 and inhibited the appearance of both H3K27me3 and H3K9me3 while LINE-1 promoted H3K9me3 (Figure 11B, left). H3K36me3, in turn, inhibited H3K27me3 and promoted low gene expression level variation while H3K9me3 promoted high gene expression variation (Figure 11B, left). At the highest expression genes, the same causality was seen, except that H3K36me3 and H3K9me3 failed to influence gene expression variation (Figure 11A, left). At the lowest expressed genes the scenario was also similar except that here Alu failed to measurably promote H3K36me3 and H3K27me3 promoted gene expression variation (Figure 11C, left). For all of the aforementioned gene groups, the presence of H3K79me2 in the 198-kb regions failed to influence gene expression variation and H3K79me2 was consequently ommited from these figures.

In the 2-kb flanking the TSS of genes, the situation was more complicated. Here, Alu failed to promote H3K36me3 and LINE-1 failed to promote H3K9me3 but both elements inhibited competing histone modifications from forming (Figure 11A-C). H3K36me3 consistently promoted low variation and both H3K27me3 and H3K9me3 consistently promoted high variation, although the role of H3K9me3 in highest expression genes was very weak. In the 2-kb regions, H3K79me2 very strongly promoted low gene expression variation across all gene expression levels and H3K36me3 promoted H3K79me2 in the main set of genes. Finally, histone acetylation (H3K27ac and H3K9ac) promoted low variation in the main set of genes.

In the 198-kb region around silent genes, Alu still promoted H3K36me3 and inhibited H3K9me3 and H3K27me3 and LINE-1 promoted H3K9me3 (Figure 9D). The only difference between 198-kb regions around silent genes and around the main set of genes was that H3K36me3 failed to inhibit H3K27me3 at silent genes (Figure 11D). However, in non-gene regions (defined as all regions > 100-kb from the TSS of any known gene), Alu failed to promote H3K36me3 and LINE-1 failed to promote H3K9me3 (Figure 11E). In these regions, which tend to be low Alu high LINE-1, LINE-1 promoted something that we labeled as 'non-of-the-above' or 'other', which is defined as the regions that do not have any of the known histone modifications in ENCODE dataset for LCLs. Other appears to be an as-yet uncharacterized chromatin structure. It is not known whether the uncharacterized chromatin in non-gene regions is the same or different from the uncharacterized chromatin that appears to be concentrated in the 200-kb flanking regions of extreme low Alu high LINE-1 genes (Figure 9D).

Discussion

This study presents three fundamental steps forward in our understanding of biology. First we show that the striking non-random distribution of Alu and LINE-1 transposon sequence, that constitutes approximately 30% of the human DNA, is responsible for modulating gene expression variation in humans. Alu elements reduce variation by promoting H3K36me3 and LINE-1 elements increase variation by promoting H3K9me3. Second, we show that Alu and LINE-1 abundance is a determinant for the identity of the repressive chromatin that occupies core promotors and transcription start sites of silent genes and a determinant for the identity of the repressive chromatin that occupies the gene bodies of expressed genes. Finally, this provides a framework for understanding why the genome is organized the way it is into alternating domains of H3K36me3, H3K27me3 and H3K9me3 chromatin and at least one additional type of chromatin whose histone modifications have not yet been characterized.

The finding that Alu and LINE-1 modulate gene expression variation provides an explanation for the observed distribution of Alu and LINE-1 elements in the vicinity of genes. Strikingly, the low variation genes surrounded by high concentration of Alu and low concentrations of LINE-1 sequence were dominated by genes involving in metabolism as determined by the gene function analysis while the largest groups of genes among the low Alu high LINE-1 highest variation genes were involved in development (Figure 3). It has previously been reported that high expressed housekeeping genes display low gene expression variation (*76*) and this is the first study to provide an explanation for why. The negative correlation of expression level variation and local Alu concentrations was evident in all 59 data sets examined that spanned several tissue types (Figure 2, left). However, the positive correlation of variation with LINE-1 concentrations was only seen in subset of datasets (with the size of the subset depending on the stringency of criteria: 30 / 59 datasets by our count). Inconsistencies between sample sets of the same cell or tissue type from different investigators led us to conclude that differences in data quality were making it impossible to discern tissue-specific differences (if any) for the LINE-1 correlation (Figure 2, right). For example, LINE-1 showed an overall positive correlation with variation among the 11 whole blood data sets but among these were data sets that completely failed to show this trend. The strength of these correlations could be mapped to sections of chromosomes suggesting domain controls (Figure 5E, F). However, these domain maps were even more sensitive to sample quality than broader LINE-1 correlations (not shown) and we cannot rule out that some of the tissue differences between LCLs and brain are due to different sample processing protocols. Standardized sample collection and storage practices and the use of RNA-seq rather than microarrays should permit tissue-specific differences to be examined.

Genome sequencing had revealed that humans were remarkably genetically uniform (*77*) and it is generally agreed that this is due to the surviving human population having been only a few thousand individuals at some time in the past. One hypothesis is that the human population was reduced to approximately 10,000 or even fewer individuals by the enormous Toba volcanic eruption approximately 70,000 years ago whose volcanic ash and droplets of sulfuric acid may have sufficiently obscured the sun to raise Earth's reflectivity of solar radiation resulting in an extended 'volcanic winter' (*78*). This is on the scale of what is thought to be the minimum population size for humans and other terrestrial vertebrates (*79*). We speculate that LINE-1 mediated increases in gene expression variation may increase variation in the absence of genetic variation and may reduce the size of the minimal viable population and thus may help protect a species against extinction.

Throughout this study, we determined the gene expression level and gene expression level variation from microarray data. We realized that the expression levels derived from RNA-seq method are generally more promising and consistent, and the estimate of gene expression variation should be therefore more accurate. Large-scale RNA-seq analysis that include >20 normal samples from the same tissue have also been done for certain cell types, such as lymphoblastoid cell lines (*80*). However, large-scale microarray datasets (>20 normal samples) are much more available and easy to obtain. For example, we investigated the correlation between gene expression variation with Alu and LINE-1 sequence abundance across 24 tissues from 55 microarray datasets. In comparison, the RNA-seq data for these normal tissues are mostly not available. We will extend our analysis to RNA-seq derived gene expression level and gene expression level variation as more of them becoming available for different tissues.

Even though microarray datasets for normal tissues are more available, the large-scale study for some tissues like heart or skin are still scarce. After searching the GEO database, only one large-scale microarray for normal sample (from >20 normal individual) were found for each of them, respectively. In heart (GSE5406) and skin (GSE14905), the Alu concentration is negatively correlated gene expression variation, which is the same as the general trend. However, LINE-1 seems also decreasing gene expression variation, which contradict the general trend that LINE-1 promotes gene expression variation. Because there is no other independent large-scale microarray datasets for these two tissues available for verifiation, there are at least two possible explanations: 1) in heart and skin, unlike most of other tissues, LINE-1 used a different tissue-specific mechanism to influence gene expression variation, or 2) more likely, such deviations are due to the difference in the sample quality or experimental procedure. Indeed, even for whole blood, while most of the datasets follow the general trend, that is, low-Alu high-LINE-1 promote gene expression variation, some of the datasets such as (GSE11375) shows a reversed trend for LINE-1. Nevertheless, following analysis by using the more accurate RNA-seq derived gene expression variation should make the general trend more conclusive.

In Figure 5 we found the regional difference of the correlation between %Alu and %LINE-1 with gene expression variation. While in most of the regions, the %Alu negatively and %LINE-1 positively correlated with gene expression variation, there exists quite a few regions that showed the reversed trend. One explanation is that the tissue-specific genes in these regions used a different way to generate gene expression variation, other than Alu/LINE-1 facilitated repressive histone modification. Or through unknown mechanism, Alu attracts H3K9me3 or LINE-1 attracts H3K36me3 to influence gene expression variation. Once the large-scale RNA-seq (for gene expression variation) and ChIP-seq data (for histone modifications) become available for different normal tissues, it is very interesting to see how the associations between Alu/LINE-1 and repressive marks change in these 'trend-reversed' region.

We found that after excluding the effect of common SNP and CNVs derived from HapMap project, the significant association between Alu/LINE-1 and unexplained gene expression variation still existed, suggesting that beside the genetic causes, epigenetic factors such as the LINE-1/Alu facilitated histone modifications also contribute to the observed gene expression variation. It should be noted a large number of rare SNPs and structural variation are not included in the HapMap data (*81*). Only the effect of cis-SNPs, which is 500kb upstream and downstream of any gene, has been considered, yet distant SNPs may also influence the gene expression variation through the trans-mechanism. Moreover, some non-linear or gene-level random effects between SNPs/CNVs and gene expression variation may not be explained by our simple linear model. There may be underestimation of the gene expresssion variation that can be explained by genetic causes. Nevertheless, our analysis showed an important way that

Alu and LINE-1 based chromatin change also lead to the variability in gene expression, which stochasticity has to be explained by both genetic and epigenetic factors (*60*).

Statistical methods have been widely used to model the relationship between chromatin marks obtained from high-throughput technique like ChIP-seq and functional components along the chromosomes (*82-93*). For example, Thurman et al. combined wavelet analysis and hidden Markov models to discover 53 active and 62 repressed functional domains by using the ENCODE data. They found that LINE-1 elements are enriched in repressive regions and Alu in active regions (*84*). Ernst et al. again used Hidden Markov model to identify 51 distinct chromatin states, in which LINE-1 repeats are enriched in a state with dominant H3K9me3, and Alu repeats are enriched around active enhancers and promoter regions of medium and high expressed genes (*86*). The Hidden Markov Model based methods have some limitations, such as handling missing data, requiring interpolation and smoothing to process regions where data is not available. The Bayesian network based method provides the additional flexibility of modeling complex relationships among variables, and has been successfully applied to histone modification data to discover the causal relationship between histone modifications and gene expression (*74*), and the chromatin structure (*92*). However, none of aforementioned analysis has explored the relationship between repetitive sequences, histone modifications and particularly, gene expression variation. In this analysis, for the first time we incorporated the repetitive sequence information in genes' flanking region into the Bayesian network model to study the causal relationship between repetitive sequences and histone marks. Our current model, however, only considered the concentrations of repetitive sequences and histone modifications in flanking or promoter regions surrounding each gene. We may extend our model to study such causal relationship in different genomic segmentation.

Our Bayesian network analysis indicate that Alu elements reduce variation by promoting H3K36me3 and LINE-1 elements increase variation by promoting H3K9me3. Restriction / exclusion maps furthermore indicated that Alu elements promote H3K79m2. Our Bayesian networks indicated that the effect of Alu elements on H3K79me2 is indirect: Alu elements promote H3K36me3 which in turn promotes H3K79me2 that then reduces gene expression variation even more strongly than H3K36me3. Very little has been known about the role of H3K79me2 in the genome and our finding illustrates how performing ChIP-seq followed by bioinformatic analysis can uncover important role for histone modification.

Our analysis provides a framework for understanding why the genome is organized the way it is into alternating domains of H3K36me3, H3K27me3 and H3K9me3 chromatin and at least one additional type of chromatin whose histone modifications have not yet been characterized. That the genome primarily consists of alternating regions of different types of repressive chromatin had already been known but why one type of chromatin prevails over another had not been known. In addition, it had not been known why some genes are silenced by H3K36me3, while other genes are silenced by H3K9me3 and most others by H3K27me3. Our study indicates that in the absence of Alu and LINE-1 sequence, the prevailing repressive mark is H3K27me3 which therefore may be considered the default chromatin type. It had already been appreciated that genes such as the Hox genes are devoid of LINE-1 and Alu elements and encased in H3K27me3 when not expressed but the generality of this had not been known. Our study indicates the high concentrations of Alu elements allow H3K27me3 to be replaced by H3K36me3 which is more conducive to high gene expression level (*54*) and prevents unwanted gene expression variation. Our study also indicates that high LINE-1 concentrations near genes promote H3K9me3 that increases gene expression variation. Elevated variation was seen among genes involved in development and multicellular processes we propose are the genes were diversity in expression level is adaptive at the population level. This diversity is at the expense of human complex disorders that arise in the absence of causative mutations, as in seen arising from the random gene expression differences among the monozygotic discordant for almost every known human complex disorder (*94*).

Bibliography

1. S. L. Gasior *et al.*, Characterization of pre-insertion loci of de novo L1 insertions, *Gene* **390**, 190–198 (2007).

2. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease, *Annu Rev Genomics Hum Genet* **12**, 187–215 (2011).

3. M. Dewannieux, C. Esnault, T. Heidmann, LINE-mediated retrotransposition of marked Alu sequences, *Nat. Genet.* **35**, 41–48 (2003).

4. S. Boissinot, A. Entezam, A. V. Furano, Selection against deleterious LINE-1-containing loci in the human lineage, *Mol. Biol. Evol.* **18**, 926–935 (2001).

5. J. Jurka, Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1872–1877 (1997).

6. N. Gilbert, S. Lutz, T. A. Morrish, J. V. Moran, Multiple fates of L1 retrotransposition intermediates in cultured human cells, *Mol. Cell. Biol.* **25**, 7780–7795 (2005).

7. G. Abrusan, J. Giordano, P. E. Warburton, Analysis of transposon interruptions suggests selection for L1 elements on the X chromosome, *PLoS Genet.* **4**, e1000172 (2008).

8. J. Jurka, O. Kohany, A. Pavlicek, V. V. Kapitonov, M. V. Jurka, Duplication, coclustering, and selection of human Alu retrotransposons, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1268–1272 (2004).

9. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome, *Nature* **409**, 860–921 (2001).

10. J. S. Myers *et al.*, A comprehensive analysis of recently integrated human Ta L1 elements, *Am. J. Hum. Genet.* **71**, 312–326 (2002).

11. I. Ovchinnikov, A. B. Troxel, G. D. Swergold, Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion, *Genome Res* **11**, 2050–2058 (2001).

12. A. F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).

13. D. E. Symer *et al.*, Human l1 retrotransposition is associated with genetic instability in vivo, *Cell* **110**, 327–338 (2002).

14. S. T. Szak *et al.*, Molecular archeology of L1 insertions in the human genome,

Genome Biol. **3**, research0052 (2002).

15. G. J. Cost, J. D. Boeke, Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure, *Biochemistry* **37**, 18081–18093 (1998).

16. T. Graham, S. Boissinot, The genomic distribution of l1 elements: the role of insertion bias and natural selection, *J. Biomed. Biotechnol.* **2006**, 75327 (2006).

17. E. M. Ostertag, H. H. J. Kazazian, Biology of mammalian L1 retrotransposons, *Annu. Rev. Genet.* **35**, 501–538 (2001).

18. Q. Feng, J. V. Moran, H. H. J. Kazazian, J. D. Boeke, Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition, *Cell* **87**, 905–916 (1996).

19. C. D. Eller *et al.*, Repetitive sequence environment distinguishes housekeeping genes, *Gene* **390**, 153–165 (2007).

20. T. M. Kim, Y. C. Jung, M. G. Rhyu, Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively, *J. Korean Med. Sci.* **19**, 783–792 (2004).

21. R. Versteeg *et al.*, The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes, *Genome Res* **13**, 1998–2004 (2003).

22. M. Song, S. Boissinot, Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination, *Gene* **390**, 206–213 (2007).

23. D. Jjingo, A. Huda, M. Gundapuneni, L. Mariño-Ramírez, I. K. Jordan, Effect of the transposable element environment of human genes on gene length and expression, *Genome Biol Evol* **3**, 259–271 (2011).

24. J. A. Yoder, C. P. Walsh, T. H. Bestor, Cytosine methylation and the ecology of intragenomic parasites, *Trends Genet.* **13**, 335–340 (1997).

25. P. L. Deininger, M. A. Batzer, Alu repeats and human disease, *Mol. Genet. Metab.* **67**, 183–193 (1999).

26. H. H. Kazazian, An estimated frequency of endogenous insertional mutations in humans, *Nat. Genet.* **22**, 130 (1999).

27. J. S. Han, S. T. Szak, J. D. Boeke, Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes, *Nature* **429**, 268–274 (2004).

28. M. A. Batzer, P. L. Deininger, Alu repeats and human genomic diversity, *Nat Rev Genet* **3**, 370–379 (2002).

29. B. Burwinkel, M. W. Kilimann, Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease, *J Mol Biol* **277**, 513–517 (1998).

30. A. Casarin *et al.*, 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* **10**, 243–249 (2006).

31. C. Has *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* **126**, 1776–1783 (2006).

32. L. Kozak *et al.*, Identification and characterization of large deletions in the phenylalanine hydroxylase (PAH) gene by MLPA: evidence for both homologous and non-homologous mechanisms of rearrangement, *Mol. Genet. Metab.* **89**, 300–309 (2006).

33. L. Li *et al.*, 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* **27**, 388 (2006).

34. V. Matejas *et al.*, Identification of Alu elements mediating a partial PMP22 deletion, *neurogenetics* **7**, 119–126 (2006).

35. D. Y. Nishimura *et al.*, Comparative genomics and gene expression analysis identifies BBS9, a new Bardet-Biedl syndrome gene, *Am. J. Hum. Genet.* **77**, 1021–1033 (2005).

36. P. H. Nissen *et al.*, Genomic characterization of five deletions in the LDL receptor gene in Danish Familial Hypercholesterolemic subjects, *BMC Med. Genet.* **7**, 55 (2006).

37. S. K. Sen *et al.*, Human genomic deletions mediated by recombination between Alu elements, *Am. J. Hum. Genet.* **79**, 41–53 (2006).

38. J. Shabbeer, M. Yasuda, S. D. Benson, R. J. Desnick, 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* **2**, 297–309 (2006).

39. R. K. Uddin *et al.*, 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.* **7**, 18 (2006).

40. F. Xie *et al.*, 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 **36**, 385–391 (2006).

41. G. Zhang *et al.*, Identification of Alu-mediated, large deletion-spanning exons 2-4 in a patient with mitochondrial acetoacetyl-CoA thiolase deficiency, *Mol. Genet. Metab.* **89**, 222–226 (2006).

42. H. H. J. Kazazian, J. L. Goodier, LINE drive. retrotransposition and genome instability, *Cell* **110**, 277–280 (2002).

43. S. L. Gasior, T. P. Wakeman, B. Xu, P. L. Deininger, The human LINE-1 retrotransposon creates DNA double-strand breaks, *J Mol Biol* **357**, 1383–1393 (2006).

44. V. P. Belancio, A. M. Roy-Engel, R. R. Pochampally, P. Deininger, Somatic expression of LINE-1 elements in human tissues, *Nucleic Acids Res* **38**, 3909–3922 (2010).

45. M. Guenatri, D. Bailly, C. Maison, G. Almouzni, Mouse centric and pericentric satellite repeats form distinct functional heterochromatin, *J Cell Biol* **166**, 493–505 (2004).

46. A. H. Peters *et al.*, Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability, *Cell* **107**, 323–337 (2001).

47. M. Garcia-Cao, R. O'Sullivan, A. H. Peters, T. Jenuwein, M. A. Blasco, Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases, *Nat. Genet.* **36**, 94–99 (2004).

48. T. Mikkelsen, M. Ku, D. Jaffe, B. Issac, E. al, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature* (2007).

49. P. Trojer, D. Reinberg, Facultative heterochromatin: is there a distinctive molecular signature? *Mol Cell* **28**, 1–13 (2007).

50. D. C. Leung, M. C. Lorincz, Silencing of endogenous retroviruses: when and why do histone marks predominate? *Trends Biochem Sci* **37**, 127–133 (2012).

51. M. Leeb *et al.*, Polycomb complexes act redundantly to repress genomic repeats and genes, *Genes Dev* **24**, 265–276 (2010).

52. J. Friedman *et al.*, Epigenetic silencing of HIV-1 by the histone H3 lysine 27 methyltransferase enhancer of Zeste 2, *J Virol* **85**, 9078–9089 (2011).

53. S. Chantalat *et al.*, Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin, *Genome Res* **21**, 1426–1437 (2011).

54. E. J. Wagner, P. B. Carpenter, Understanding the language of Lys36 methylation at

histone H3, *Nat Rev Mol Cell Biol* **13**, 115–126 (2012).

55. P. P. Luedi, A. J. Hartemink, R. L. Jirtle, Genome-wide prediction of imprinted murine genes, *Genome Res* **15**, 875–884 (2005).

56. P. P. Luedi *et al.*, Computational and experimental identification of novel human imprinted genes, *Genome Res* **17**, 1723–1730 (2007).

57. I. Helbig *et al.*, Gene expression analysis in absence epilepsy using a monozygotic twin design, *Epilepsia* **49**, 1546–1554 (2008).

58. J. E. Powell *et al.*, Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent, *Genome Res* **22**, 456–466 (2012).

59. J. Li, Y. Liu, T. Kim, R. Min, Z. Zhang, Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility, *PLoS Comput. Biol.* **6**, e1000910.

60. M. Kaern, T. C. Elston, W. J. Blake, J. J. Collins, Stochasticity in gene expression: from theories to phenotypes, *Nat Rev Genet* **6**, 451–464 (2005).

61. I. B. Jeffery, D. G. Higgins, A. C. Culhane, Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, *BMC Bioinformatics* **7**, 359 (2006).

62. D. E. Martin, P. Demougin, M. N. Hall, M. Bellis, Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data, *BMC Bioinformatics* **5**, 148 (2004).

63. R. Breitling, P. Armengaud, A. Amtmann, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS letters* (2004).

64. B. E. Stranger *et al.*, Relative impact of nucleotide and copy number variation on gene expression phenotypes, *Science* **315**, 848–853 (2007).

65. B. E. Stranger *et al.*, Population genomics of human gene expression, *Nat. Genet.* **39**, 1217–1224 (2007).

66. T. Barrett *et al.*, NCBI GEO: archive for functional genomics data sets--update, *Nucleic Acids Res* **41**, D991–5 (2013).

67. S. Theodoridis, K. Koutroumbas, Pattern recognition, *Academic* (1999).

68. A. Alexa, Gene set enrichment analysis with topGO, *Bioconductor Improvments* (2007).

69. M. Gönen, Multiple kernel learning algorithms, *Journal of Machine Learning Research* (2011).

70. R. S. Spielman *et al.*, Common genetic variants account for differences in gene expression among ethnic groups, *Nat. Genet.* **39**, 226–231 (2007).

71. J. H. A. Martens *et al.*, The profile of repeat-associated histone lysine methylation states in the mouse epigenome, *EMBO J.* **24**, 800–812 (2005).

72. ENCODE Project Consortium, A user's guide to the encyclopedia of DNA elements (ENCODE), *PLoS Biol.* **9**, e1001046 (2011).

73. J. R. Ecker *et al.*, Genomics: ENCODE explained, *Nature* **489**, 52–55 (2012).

74. H. Yu, S. Zhu, B. Zhou, H. Xue, J. D. Han, Inferring causal relationships among different histone modifications and gene expression, *Genome Res* **18**, 1314–1324 (2008).

75. D. Margaritis, Learning Bayesian network model structure from data, (2003).

76. A. Sharma *et al.*, Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays, *Physiol. Genomics* **21**, 117–123 (2005).

77. R. Dawkins, *The Accestor's Tale, A Pilgrimage to the Dawn of Life* (2004).

78. C. Oppenheimer, Limited global change due to the largest known Quaternary eruption, Toba≈ 74kyr BP? *Quaternary Science Reviews* (2002).

79. B. W. Brook, L. W. Traill, C. J. A. Bradshaw, Minimum viable population sizes and global extinction risk are unrelated, *Ecol. Lett.* **9**, 375–382 (2006).

80. S. B. Montgomery *et al.*, Transcriptome genetics using second generation sequencing in a Caucasian population, *Nature* **464**, 773–777 (2010).

81. R. M. Durbin, A map of human genome variation from population-scale sequencing, *Nature* **467**, 1061–1073 (2010).

82. N. Day, A. Hemmaplardh, R. E. Thurman, J. A. Stamatoyannopoulos, W. S. Noble, Unsupervised segmentation of continuous genomic data, *Bioinformatics* **23**, 1424–1426 (2007).

83. L. Jia *et al.*, Functional enhancers at the gene-poor 8q24 cancer-linked locus, *PLoS Genet.* **5**, e1000597 (2009).

84. R. E. Thurman, N. Day, W. S. Noble, J. A. Stamatoyannopoulos, Identification of higher-order functional domains in the human ENCODE regions, *Genome Res* **17**, 917–

927 (2007).

85. B. Schuettengruber *et al.*, Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos, *PLoS Biol.* **7**, e13 (2009).

86. J. Ernst, M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome, *Nat. Biotechnol.* **28**, 817–825 (2010).

87. C. Taslim, S. Lin, K. Huang, T. H.-M. Huang, Integrative genome-wide chromatin signature analysis using finite mixture models, *BMC Genomics* **13 Suppl 6**, S3 (2012).

88. L. Steiner *et al.*, A global genome segmentation method for exploration of epigenetic patterns, *PLoS One* **7**, e46811 (2012).

89. J. Wang, V. V. Lunyak, I. K. Jordan, Chromatin signature discovery via histone modification profile alignments, *Nucleic Acids Res* **40**, 10642–10656 (2012).

90. L. Teng, K. Tan, Finding combinatorial histone code by semi-supervised biclustering, *BMC Genomics* **13**, 301 (2012).

91. J. Cotney *et al.*, Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb, *Genome Res* **22**, 1069–1080 (2012).

92. M. M. Hoffman *et al.*, Unsupervised pattern discovery in human chromatin structure through genomic segmentation, *Nat Methods* **9**, 473–476 (2012).

93. J. Ernst *et al.*, Mapping and analysis of chromatin state dynamics in nine human cell types, *Nature* **473**, 43–49 (2011).

94. J. T. Bell, T. D. Spector, A twin approach to unraveling epigenetics, *Trends Genet.* **27**, 116–125 (2011).