OPEN ACCESS
CellPress

Cell Reports
Report

Evolution of Alu Elements toward Enhancers

Ming Su,[1,2,3] Dali Han,[1,2,3] Jerome Boyd-Kirkup,[1] Xiaoming Yu,[1,2] and Jing-Dong J. Han[1,*]
[1]Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China
[2]Center of Molecular Systems Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Lincui East Road, Beijing 100101, China
[3]These authors contributed equally to this work
*Correspondence: jdhan@picb.ac.cn
http://dx.doi.org/10.1016/j.celrep.2014.03.011

SUMMARY

The human genome contains approximately one million Alu repetitive elements comprising 10% of the genome, yet their functions are not well understood. Here, we show that Alu elements resemble enhancers. Alu elements are bound by two well-phased nucleosomes that contain histones bearing marks of active chromatin, and they show tissue-specific enrichment for the enhancer mark H3K4me1. A proportion of Alu elements were experimentally validated as bona fide active enhancers with an in vitro reporter assay. In addition, Hi-C data indicate that Alus show long-range interactions with gene promoters. We also find that Alus are generally more conserved when located in the proximal upstream region of genes. Their similarity to enhancers becomes more prominent with their age in the human genome, following a clear evolutionary continuum reminiscent of the evolutionary pattern of proto-genes. Therefore, we conclude that some Alu elements can function as enhancers and propose that many more may be proto-enhancers that serve as a repertoire for the de novo birth of enhancers.

INTRODUCTION

Transcriptional regulation is critical for the organization and coordination of cellular functions and organismal development (Vaquerizas et al., 2009). Elaborate gene expression regulation is a principle requirement for organismal complexity in higher animals and may depend on noncoding regions (Levine and Tjian, 2003). A large fraction of noncoding regions are repetitive sequences that are derived from transposable elements. In the human genome, 45% of sequences can be recognized as derived from transposons, among which L1 and Alu are the most numerous with >0.5 and 1.1 million copies each, comprising 17% and 10% of the human genome, respectively (Batzer and Deininger, 2002; Cordaux and Batzer, 2009).

Alu elements are preferentially distributed in gene-rich regions and contain one-third of the total CpG dinucleotides in the human genome (Batzer and Deininger, 2002), as well as many putative transcription factor (TF) binding sites, which may increase their likelihood to either enhance or repress gene expression (Polak and Domany, 2006). Although Alu insertions can mutate functional units, these features have been suggested to reflect a distinct advantageous contribution of Alu elements to the transcriptional landscape of the human genome (Batzer and Deininger, 2002; Cordaux and Batzer, 2009). However, the exact contribution of Alu to transcriptional regulation is unclear, and it remains unknown whether there exists a general role for Alu elements in gene regulation.

Here, we carried out genome-wide analyses of genomic distribution, evolutionary conservation, histone positioning, and epigenetic profiles to examine the characteristics of Alu elements. We found that the epigenetic profiles of Alu in multiple tissues and cell lines resemble those of putative transcription enhancers. Using recently published chromatin-interaction maps, we also found that Alu elements preferentially interact with nearby promoters. Intriguingly, these enhancer-like characteristics of Alu evolve with the age of the Alu elements in a clear evolutionary continuum, thus supporting Alus as proto-enhancers in the genome.

RESULTS

The Genomic Distribution of Alu in Comparison to Other Transposable Elements

We compared Alu with three other control retrotransposons, Mammalian apparent LTR-retrotransposons (MaLRs), which have similar lengths to Alus, mammalian interspersed repeats (MIR), which belong to the same ancient family of short nuclear interspersed elements (SINEs) as Alu (Smit and Riggs, 1995) and LINE-1 (long interspersed nuclear element 1, abbreviated as L1), which uses a similar retrotransposition mechanism as Alu and provides both L1 and Alu with the enzymes needed for retrotransposition (Cordaux and Batzer, 2009).

Although repetitive elements of all four families locate abundantly at gene proximal regions, Alus are 33.6%–89.8% more enriched in these regions than other types of transposons (Figure S1A). Thus, 60% of Alus are within gene proximal regions, compared to 30%–40% of L1, MaLR, and MIR (Figure S1B).

Evolutionary Conservation of Gene Proximal Alu

The functional importance of genomic DNA sequences can often be reflected by their evolutionary conservation. By comparing orthologous Alu elements in human and chimpanzee (based on

412,612 out of 1,180,972 human Alu elements that have chimpanzee orthologs, Supplemental Experimental Procedures), we found that Alu conservation in noncoding regions (reflected by a decreased substitution rate) increases progressively with proximity (but not close proximity) to transcription start sites (TSSs) and reaches its maximum just outside of the TSSs. By contrast, Alu elements are less conserved within the gene body and reach their lowest conservation level at transcription termination sites (TTS) (Figure S1C). These findings suggest that Alus immediately upstream of TSS are the most likely to be functional. Such patterns were not observed for MIR and MaLR (Figure S1C). L1 conservation also increases when proximal to TSS but reaches the maximum at a farther distance (20–30 kb) to TSS compared to that of Alu (~10 kb) (Figure S1C).

Although L1 and Alu tend to accumulate in AT-rich and GC-rich regions, respectively (Figure S1D), flanking region GC bias cannot explain the significant conservation of Alus at the gene vicinity (proximal and distal regions) because MIRs are also enriched in GC-rich regions but are not significantly conserved at the gene vicinity. In fact, gene proximal or distal regions with MIR insertions are significantly more GC enriched than those with Alu insertions (Figure S1E). Moreover, the gene proximal regions with MaLR insertions have a similar GC content as those with Alus, yet MaLRs display no conservation at these regions (Figure S1E). These results further indicate that the relative conservation of Alu at the gene vicinity is a feature of Alu that cannot be attributed merely to the GC bias of its flanking region.

Alu Has a Characteristic Histone Modification Pattern

We used deep-sequencing data to compare various epigenetic features for the four different repetitive elements with known functional elements, such as DNase hypersensitive sites (DHSs), enhancers, and RefSeq genes. In order to avoid any mapping bias, whenever possible, we used two different mapping strategies (redundant multiple mapping and unique mapping), along with background controls to verify the authenticity of a pattern or feature (Experimental Procedures). Applying both strategies to the nucleosome mapping (MNase-seq) and TF chromatin immunoprecipitation sequencing (ChIP-seq) data, we observed two well-phased nucleosomes on the Alu elements, and peaks of ChIP-seq tag distribution for some TFs (Figures S1F–S1K).

We then examined the histone modification ChIP-seq data from $CD4^+$ T cells (Barski et al., 2007; Emera and Wagner, 2012) (Figure 1A). By normalizing the number of ChIP-seq tags for a histone modification against the number of nucleosome mapping MNase-seq tags (Experimental Procedures), we found that Alu tends to possess H3K4me1, H3K4me2, H3K27me1, H3K36me3, and other modifications associated with open chromatin and enhancers (Barski et al., 2007; Göndör and Ohlsson, 2009) but lacks the active enhancer mark H3K27ac (Creyghton et al., 2010; Rada-Iglesias et al., 2011). In contrast, the modifications associated with gene repression, such as H3K9me2 and H3K9me3, are preferentially excluded from Alu but enriched on MaLR or L1, which also preferentially exclude gene expression activating marks (Figure 1A). Another heterochromatin mark, H4K20me3, was not enriched on any of these transposons. MIR showed almost no preferential association with specific histone modifications other than a strong exclusion of H4K20me3 (Figure 1A). Even though Alu elements are enriched near genes, and bear a high level of H3K36me3 as do transcribed genes, some of the active marks on Alu do not seem to be simply a consequence of the histone modification spreading from nearby genes, as there are obvious differences in H3K27ac, H3K36ac, and H3K79me1 marks between Alu elements and RefSeq genes (Figure 1A). Enrichment patterns were similarly observed, and therefore independent of, whether 0, 1, or 2 mismatches were allowed in the alignment, whether total tag counts or uniquely mapped tag counts were used, or whether tag counts were normalized against the nucleosome or immunoglobulin (Ig) G (Figure S1L).

To see whether such enrichment is dependent on genomic position, we separated the Alu elements according to their genomic positions and found that no matter where these Alus are in the genome (intergenic, gene proximal, gene distal, or intragenic), compared with their environment (flanking regions), H3K4me1 is always enriched (Figures 1B and 2A), arguing that this enrichment is independent of their locations in the genome and cannot be attributed to the influence of nearby genes or genomic distribution. Moreover, compared with their respective flanking regions, Alus are also enriched for H3K27me1, H3K27me2, and H3K36me3 and excluded of H3K9me3 whether they are localized near to genes or further away (Figure 1B), suggesting that these epigenetic signatures of Alu are not the result of their insertion environment. In contrast, only H3K9me3 and H4K20me3 show environment-independent L1-specific enrichment; e.g., they are enriched on L1 even when L1 elements are located in introns (Figure S1M). Although MIRs and MaLRs are also enriched for H3K4me1 when they are in gene proximal regions, unlike Alus, they are not enriched for H3K4me1 when they are located in intergenic regions, and the enrichment levels are always similar to, or less than, their genomic environments (flanking regions) (Figure S1M). This shows that H3K4me1 enrichment on MIRs and MaLRs is dependent on genomic distribution.

By hierarchical clustering of the histone modification profiles, Alu elements, but not the other three transposons, cluster together with RefSeq genes, putative enhancers (Heintzman et al., 2009), and DNase I hypersensitive sites (DHSs) (Crawford et al., 2006) (Figure 1A). Although we removed TSS −10 and +5 kb regions from DHS, the DHS regions may still contain not only enhancers but also some promoters, as shown by the high level of H3K4me3, a modification exclusively localized to promoters. H3K4me3 is excluded on all four transposons, consistent with their underrepresentation at promoter regions. Therefore, only Alus among the four tested transposons have the canonical enhancer signature of high H3K4me1 and low H3K4me3 (Figures 1C and 1D) (Heintzman et al., 2009). Despite high H3K4me1 level, Alus lack the active mark H3K27ac. This pattern is most reminiscent of poised enhancers, defined as inactive enhancers with low H3K27ac but high H3K4me1 levels that can acquire H3K27ac and turn active upon differentiation or other external cues (Creyghton et al., 2010; Rada-Iglesias et al., 2011).

Alus Display Tissue-Specific Enrichment for H3K4me1

In addition to epigenomic data from $CD4^+$ T cells, we also analyzed histone modifications in H1 and IMR90 cells, using

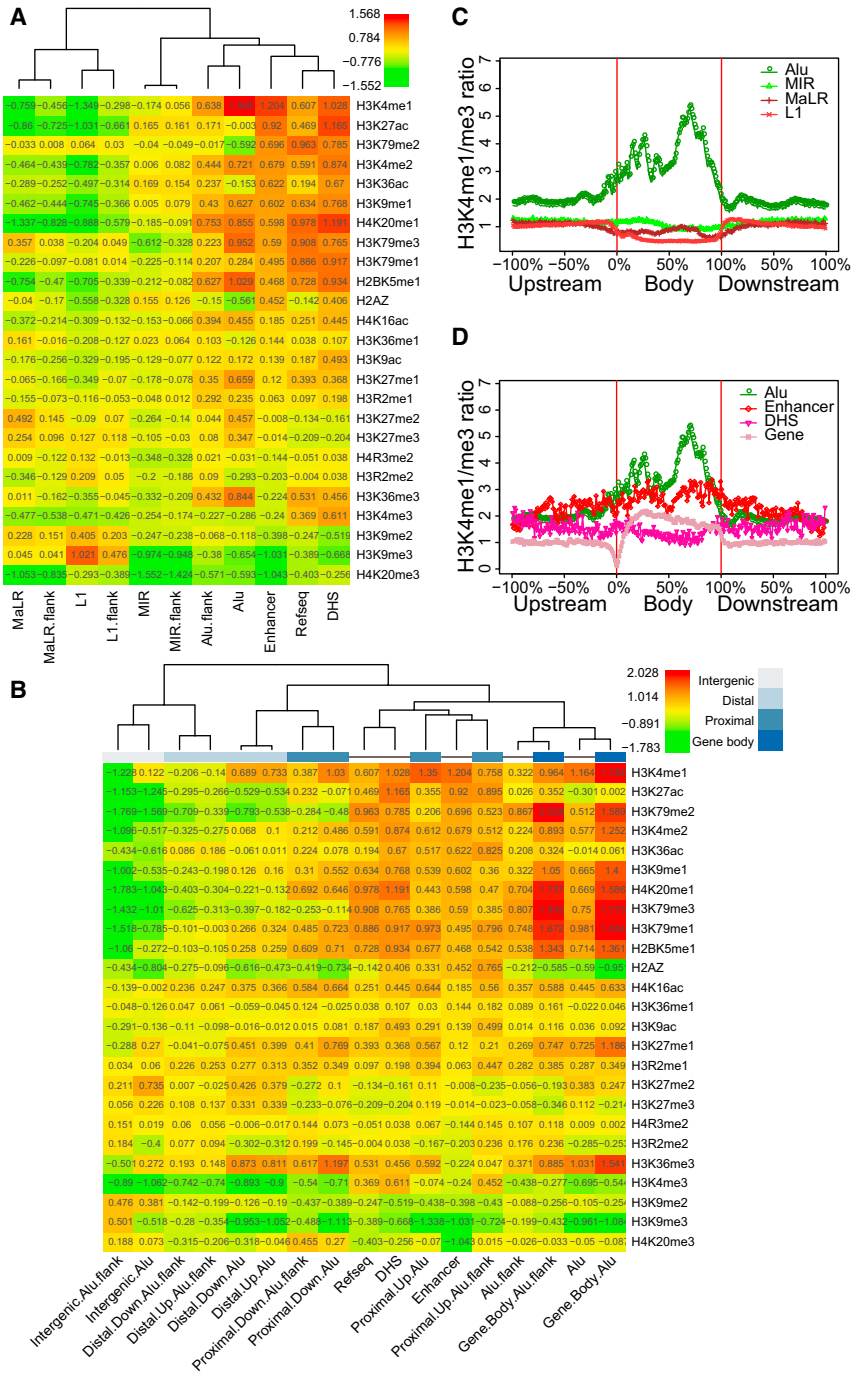

Figure 1. Alu Has Distinctive Active Histone Modifications

(A) Profiles of histone modification preferences on Alu, L1, MIR, MaLR, RefSeq genes, putative enhancers, and DNase hypersensitive sites (DHS) in CD4$^+$ T cells. Preference scores were calculated as described in the Experimental Procedures; a positive or negative number indicates relatively high or low levels of modification on each type of genomic elements, compared to the total-reads normalized nucleosome-mapping reads. Profiles here are based on multiple-reads mapping with at most two mismatches. Those based on 0 or 1 mismatch, unique-reads mapping, or compared to IgG controls are shown in Figure S2G.

(B) Profiles of histone modification preferences on Alus grouped by their distances to the nearest annotated genes. Gene proximal regions are sequences within 10 kb upstream of TSSs or downstream of TTSs. Distal regions are sequences within 10–100 kb regions up- or downstream of the transcription units. Intergenic regions are >100 kb from a nearest gene.

(C and D) Profiles of H3K4me1/H3K4me3 ratios across the body, upstream, and downstream (extending 100% of the transposon length from the boundary of the transposon) for transposons (B) and annotated elements (C). Red vertical lines mark the boundary of respective elements. The tag counts of the H3K4me1 and H3K4me3 ChIP-seq were linearly scaled to the same total counts.

See also Figure S1.

the NIH Roadmap Epigenomics Mapping Consortium (NREMC) data (Hawkins et al., 2010). Generally, in both cell types, Alus are enriched for active modifications (Figures S2A and S2B). Similar to CD4$^+$ T cells, Alus are also enriched for H3K4me1 signals in IMR90 cells although to a lesser extent (Figure S2A), but, in H1 ES cells, Alus show no enrichment of H3K4me1 (Figure S2B).

To further investigate the tissue specificity of H3K4me1 enrichment, we obtained additional data from the NREMC project (Table S1) and calculated the H3K4me1 enrichment on Alus located in different genomic regions in 20 different cell lines and tissues. H3K4me1 is highly enriched on Alus in HeLa cells, GM12878 cells, CD3$^+$ and CD4$^+$ blood cells, mesenchymal stem cells (MSCs), thymus and muscle cells, and moderately enriched in CD19$^+$ and CD56$^+$ blood cells, adrenal gland, and IMR90 cells (Figures 2A and S2C). In contrast, for CD14$^+$ blood cells, cells of the small and large intestine, trophoblasts, H1, H9 embryonic stem cells (ESCs), and neural progenitor cells (NPCs), Alus have weak or no enrichment for H3K4me1, and, in mesendoderm cells, Alus, in fact, show a depletion of H3K4me1 (Figures 2A and S2C). Such tissue-specific enrichment for H3K4me1 is characteristic of enhancers, which are generally tissue specific, but contain common epigenetic signatures.

Some Alus Display Enrichment for Active Enhancer Marks

To identify how many enhancers in a human cell could be derived from Alus, we first used the HOMER software (Heinz et al., 2010) to determine peaks of two enhancer marks H3K4me1 and

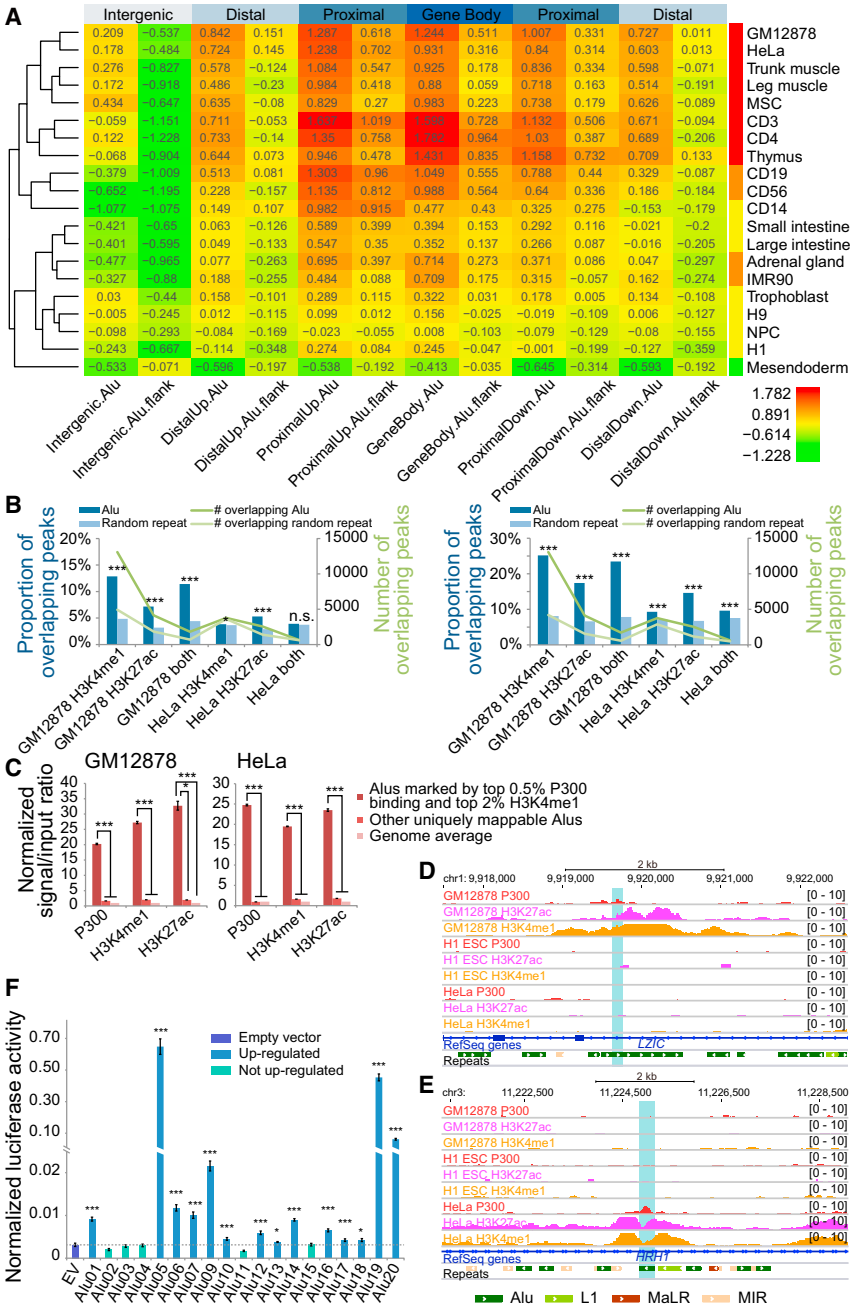

Figure 2. Alus Show Tissue-Specific Enrichment for H3K4me1

(A) Enrichment level of H3K4me1 on Alus at different genomic positions and their flanking regions, in various cell lines and tissues. Tissues are clustered based on the H3K4me1 enrichment profile in each tissue (one profile is one row). Red, orange, and yellow indicate high, moderate, and low enrichment, respectively, whereas green indicates depletion.

(B) Proportions of H3K4me1, H3K27ac, or their overlapped peaks that intersect with Alus (with >50% overlap) (left panel) and proportions of peaks that cannot be mapped to the mouse genome and intersect with Alus (right panel). Peaks were called by the HOMER software and mapped to the mouse genome by liftOver. The numbers of overlapping peaks are also shown. Also shown for comparison are the overlap percentages for the same number of random fragments with the same lengths as Alus from the all repeats background. * and ***, proportion test p values <0.05 and 0.0001, respectively; n.s., not significant. "All repeats" were acquired from UCSC genome browser repeat masker track, from which any family containing "RNA" in its name was excluded.

(C) Alus marked by top 0.5% P300 binding and top 2% H3K4me1 level also have significantly higher level of active enhancer mark H3K27ac than all other uniquely mappable Alus and the genome average. * and *** indicate t test p values <0.005 and 0.001, respectively.

(D and E) Examples of Alus bound by the active enhancer binding protein P300 in GM12878 cells (D) and HeLa cells (E). Alus marked by top 0.5% P300 binding and top 2% H3K4me1 are indicated by the cyan blocks. Alus, L1s and MIRs are marked with different colors in the repeats track.

(F) Experimental validation of the enhancer activity of the uniquely mappable Alus with the top 20 P300 binding and top 2% H3K4me1. Nineteen of the 20 Alus were successfully cloned into the luciferase reporter vector. Determined by the firefly luciferase versus the renilla luciferase activity, 14 showed upregulated luminescence compared to minimal promoter (empty vector), with fold changes ranging from 1.21 to 207.67, based on three biological replicates for each measurement. * and *** indicate one-tailed t test p values <0.05 and 0.01, respectively.

See also Figure S2.

H3K27ac and then defined intersections between peaks and Alus if 50% of the length of a peak overlapped with an Alu, with BEDTools (Quinlan and Hall, 2010). In GM12878 cells, among a total of 101664 H3K4me1 peaks, 12.85% show >50% overlap with Alus; among a total of 57,770 H3K27ac peaks, 7.16% show >50% overlap with Alus; among a total of 15,809 overlapping H3K4me1-H3K27ac peaks (with 50% length overlap), 11.42% show >50% overlap with Alus (Figure 2B left panel, also including results for HeLa cells). These percentages are higher than the overlap percentages for the same number of random fragments with the same lengths as Alus from the all repeats background (proportion test p values = 0, 2.09×10^{-204}, and 7.52×10^{-118} for H3K4me1, H3K27ac, and overlapping H3K4me1-H3K27ac peaks in GM12787 cells, 0.039, 1.19×10^{-78}, and 0.388 in HeLa cells) (Figure 2B left panel). Here, we focused on sharp H3K4me1 and H3K27ac peaks to avoid false-positives because the long regions that have continuous signals (as shown by NREMC) may accidentally overlap with one or more Alus simply because they are close to genes (Figure S2D). Thus one NREMC peak may cover many small sharp peaks (as shown by our analysis), resulting in the much larger number of enhancer mark peaks detected by our

analysis compared with that by NREMC (see Figure S2D for an example).

Although Alus are primate specific, many enhancers may have evolved before Alus appeared in the genome. We therefore determined whether Alus contribute a more significant fraction of enhancers that are not present in the mouse genome (Supplemental Experimental Procedures). Indeed, 25.15%, 17.42%, and 23.44% of the 51894 H3K4me1, 23687 H3K27ac, and 7379 overlapping H3K4me1-H3K27ac peaks, respectively, which cannot be mapped to the mouse genome, show >50% overlap with Alus (Figure 2B, right panel, together with results for HeLa cells). These are more than twice the fractions within all peaks (Figure 2B, left panel). Compared with the overlap to all peaks, these percentages are even higher than the overlap percentages for the same number of random fragments with the same lengths as Alus from the all repeats background (proportion test p values = 0, 8.64×10^{-286}, 1.11×10^{-149} for H3K4me1, H3K27ac and overlapping H3K4me1-H3K27ac peaks in GM12787 cells, and 2.29×10^{-29}, 1.62×10^{-124}, 2.41×10^{-5} in HeLa cells) (Figure 2B, right panel). We further examined the functions of nearby genes (up- and downstream 100 kb) for the overlapping H3K4me1-H3K27ac peaks using GREAT software (McLean et al., 2010). It would appear that these peak regions often regulate immunity and inflammation-related functions and pathways (Figure S2E), which is consistent with the enrichment of H3K4me1 on Alus in immune cells (Figure 2A).

Next, we asked whether some Alu elements were bound by active enhancer marks and the binding protein P300 in specific tissues. Indeed, by screening for both P300 binding and H3K4me1 level on uniquely mappable Alus (Supplemental Experimental Procedures), there are a total of 3,093 and 774 Alus having >5-fold enrichment for P300 and H3K4me1 over input and >5-fold enrichment for P300/input ratio over flanking regions in GM12878 and HeLa cells, respectively (Table S2). For example, the 34 Alus in GM12878 and 102 Alus in HeLa with the top 0.5% P300 enrichment and top 2% H3K4me1, also have a significantly higher level of H3K27ac compared with other Alus and the genome average (Figure 2C). The levels of enrichment of these enhancer or active enhancer marks are significantly higher than the all repeats background (Figure S2F, proportion test p value = 0). The enrichment level for enhancer binding acetyltransferases, such as P300 or GCN5, are also higher than the all repeats background, except for P300 in HeLa cells, which is lower than background (Figure S2F).

Two examples of Alus with specific P300, H3K27ac, and H3K4me1 peaks in one tissue but not in other tissues are shown in Figures 2D and 2E. Because P300, H3K27ac, and H3K4me1 peaks do not coexist on other Alus and control repeats in the same region, the high active modification levels are unlikely to be conferred by genomic background. The nearest genes to the Alus with top 0.5% P300 enrichment and top 2% H3K4me1 tend to show higher expression levels in their respective tissues (Figures S2G and S2H) and are enriched for some tissue-specific functions (Figures S2I–S2J, p value <0.05, fold enrichment >2). A general trend of decreasing nearest gene expression can be also observed with decreasing level of P300 binding and H3K4me1 signals on Alus (Figure S2K).

Among the Alus with the top 20 P300 enrichment and top 2% H3K4me1, we successfully cloned 19 into the luciferase reporter vector (Supplemental Experimental Procedures). Among the 19, 14 showed upregulated luminescence compared to minimal promoter (empty vector), with fold changes ranging from 1.21 to 207.67 (Figure 2F). This unequivocally demonstrated the capability and authenticity of these Alu elements to act as active enhancers.

Alu-Interacting DNAs Are Preferentially Located in Gene Promoters

Apostolou and Thanos (2008) have shown that three DNA fragments containing Alu elements on the same or different chromosome promote the activity of an NF-κB targeted promoter in *trans,* through NF-κB binding sites in these Alu elements. This has been suggested to recruit the promoter into enhanceomes (Göndör and Ohlsson, 2009). To examine whether Alu elements are generally involved in such TF-mediated DNA-DNA interactions, we used the Hi-C intra- and interchromatin interactions map (Lieberman-Aiden et al., 2009).

When normalized against the background tag distribution at the same distance from the TSS (Experimental Procedures), the DNA fragments interacting with the Alu elements in these maps (Experimental Procedures, Figure 3A) predominately localize to the promoter regions, with their highest density around the TSS (Figure 3B, the unnormalized and the background distributions are shown in Figure S3A). Their preference toward TSSs is even stronger than putative enhancers, but weaker than DHS (which may also contain promoter regions as shown above) (Figure 3B). In contrast, DNAs interacting with MaLR and L1 are slightly excluded around TSSs, whereas those interacting with MIR show no preference or exclusion (Figure 3C).

Furthermore, when Alu elements were grouped based on their distances to annotated genes, and even when normalized against the background tag distribution at the same distance to the TSS (Experimental Procedures), the DNA fragments interacting with gene proximal Alu are more concentrated around TSSs, whereas those interacting with distal Alu are moderately enriched at TSSs, and those interacting with intergenic Alu showed no apparent enrichment in promoter regions (Figure 3D, the unnormalized and the background distributions are shown in Figure S3B). These results also indicate that proximal Alu elements are indeed more likely to interact with promoters.

Again, the GC content bias of flanking regions cannot explain these long-distance interactions, as MIRs or MaLRs, which are more or similarly biased toward high GC sites, did not display such long-distance interactions (Figure 3C). Sequencing bias toward high GC tags also cannot explain the unique preferential long-distance interaction of Alu with TSSs. If we used reads mapping to the repeat flanking regions, where both MIR and MaLR have a higher GC content than the Alu flanking regions, we observed similar profiles as those that mapped directly to the repeats (Figure S3C). Finally, similar profiles were also found using a different Hi-C data set derived from human embryonic stem cells (hESCs) and human fibroblast IMR90 cells (Dixon et al., 2012) (Figures S3D and S3E).

In addition to the enhancer-like features of Alus, the variety of binding sites for tissue-specific or general transcription

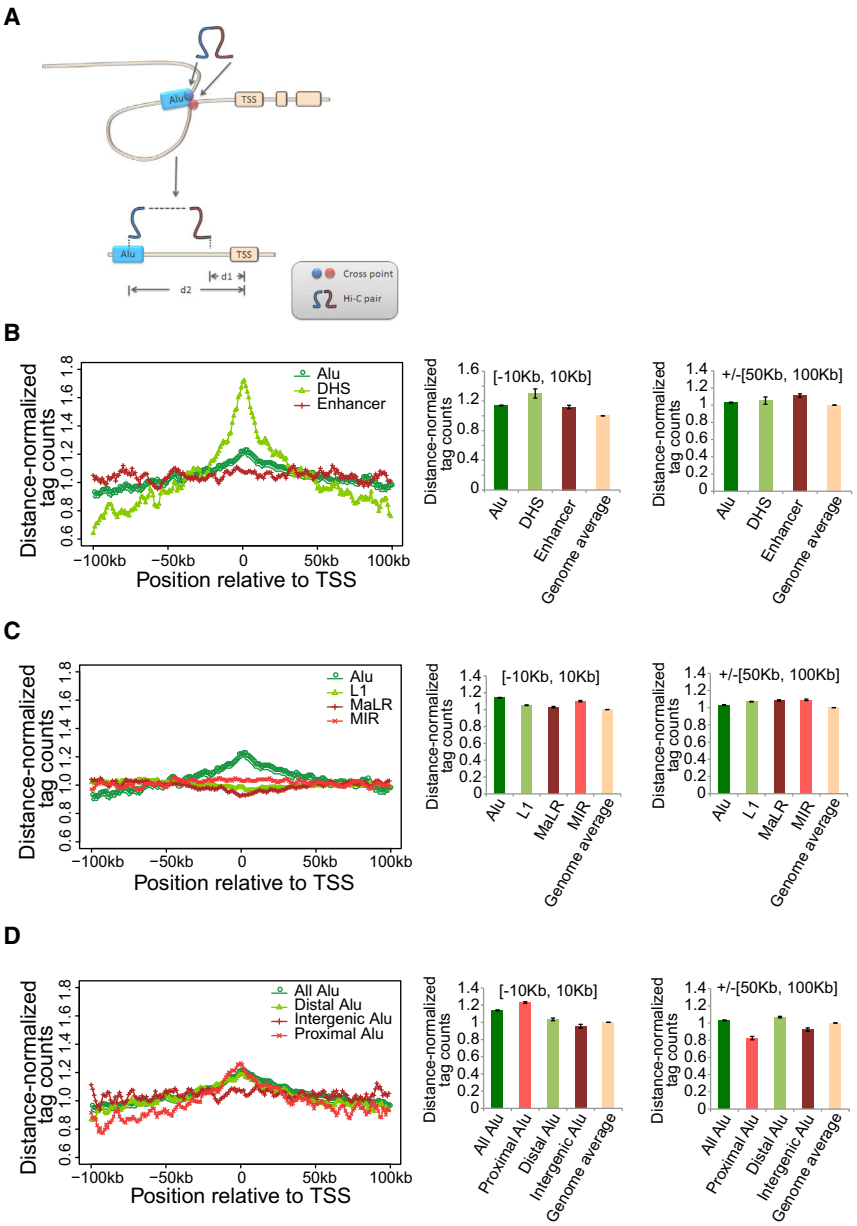

Figure 3. Alu-Interacting DNAs Show a Preference toward TSS

(A) Distance measurement and distance normalization strategy. In order to determine the above-background distribution of the "to TSS" distance of the tags interacting with each type of repeats or functional elements (d1), the background distribution of the elements' own "to TSS" distance (d2) was controlled by randomly sampling tags within the same 1 kb length interval (d2, Experimental Procedures).

(B and C) Distance-normalized distribution of DNA fragments interacting with Alu against their distances to TSS, compared with those interacting with annotated elements (B) or other transposons (C). The distance-normalized tag counts of interacting DNAs (A and Experimental Procedures) are plotted for each moving window of 5 kb at a step size of 1 kb within ±100 kb of TSS. The unnormalized and respective background distributions are shown in Figure S3A.

(D) Distance-normalized distribution of proximal, distal, or intergenic Alu-interacting DNA fragments against the fragments' distances to TSS. The interacting DNA fragments were mapped against the closest TSS. Gene proximal Alus are within 10 kb upstream of a TSS or downstream of a TTS. Distal Alus are within 10–100 kb regions up- or downstream of the transcription units. Intergenic Alus are >100 kb from the nearest gene.

In (B)–(D), histograms show statistics of distance-normalized tag counts within 10 or 50–100 kb region up- or downstream of TSS (Experimental Procedures). All comparisons are statistically significant (t test p value $<10 \times 10^{-10}$). Error bars indicate standard variances. The unnormalized and respective background distributions are shown in Figure S3B.

See also Figure S3.

regulators on Alus (Polak and Domany, 2006) is especially concordant with the concept of "enhanceosomes" (Farnham, 2009; Göndör and Ohlsson, 2009), which are specialized loci in the nucleus that organize active chromatin domains. As constituents of the enhanceosome often interact with each other, we also examined whether Alus tend to interact with other Alus. Using the Hi-C interactome maps, we indeed found that Alu preferentially interacts with Alu, whereas other repetitive elements do not show such preferences (Figure S3F).

Evolutionary Continuum of Alu toward Enhancers

With the genomic and epigenomic features of Alu elements resembling enhancers, what might be the real functions of Alu? Recently, it was found that the majority of newly emerged genes have evolved from intergenic sequences via proto-genes, which are translated non-genic sequences subject to evolutionary selection for their de novo functions (Carvunis et al., 2012). There is also a well-known concept called "exaptation," where transposons acquire new functions to facilitate their host's way of life (Bejerano et al., 2006). Anecdotal transposon elements have been reported to evolve into enhancers in human and other species (Bejerano et al., 2006; Emera and Wagner, 2012; Lynch et al., 2011; Santangelo et al., 2007). There are many reports that functional enhancers have evolved from Alus. For example, a DNA fragment containing an Alu element enhances transcription of the liver-specific HPR promoter in Hepatoma cell lines (Oliviero and Monaci, 1988); another DNA element, which may possess transcription enhancer-like properties for respiratory chain genes, is located within an Alu sequence (Liu and Bradner, 1993); several slow-evolving Alus are involved in regulating APOA5 gene expression (Ruiz-Narváez and Campos, 2008); and three Alu-containing genetic loci can be bound by the transcription

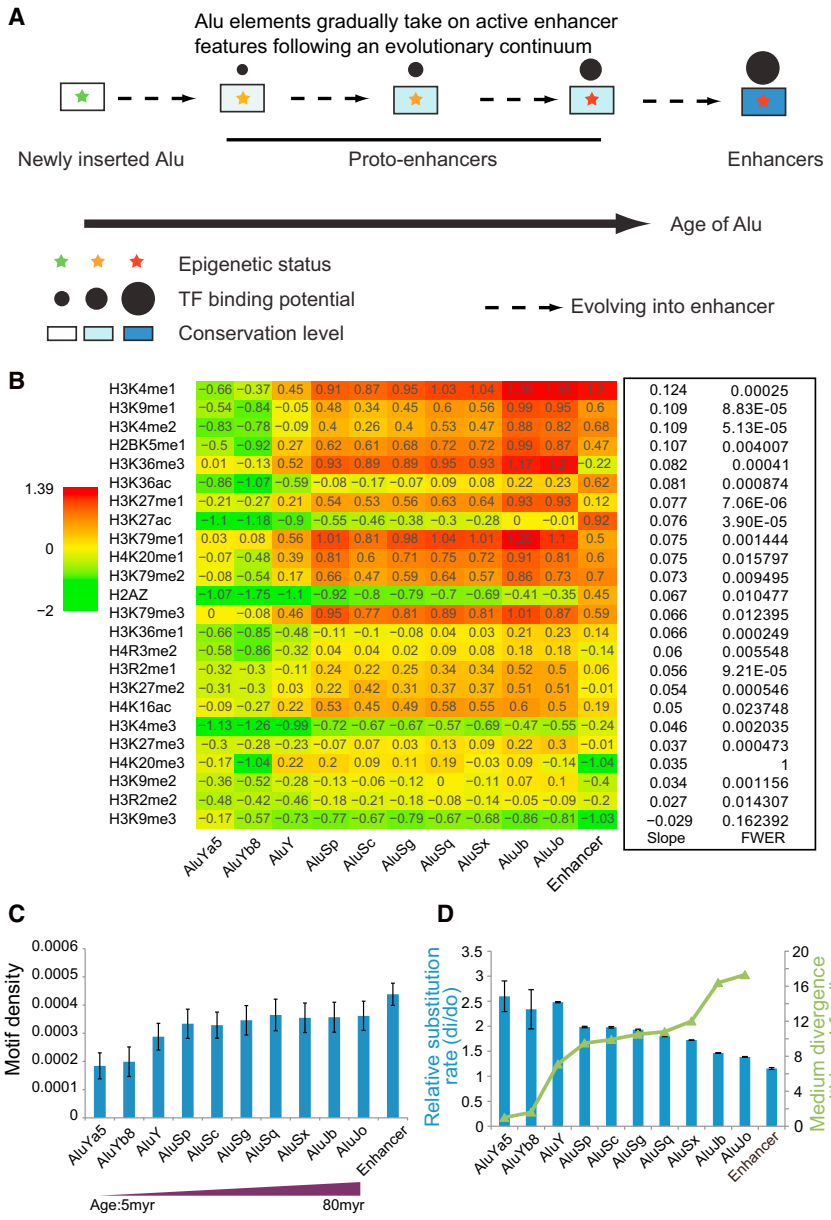

Figure 4. Evolutionary Continuum of Structural and Functional Features of Alu Elements

(A) The proposed model of Alu elements as proto-enhancers that follow an evolutionary continuum after insertion into the genome, and then evolve into functional enhancers. Conservation level refers to the conservation between human and chimpanzee. Darker color represents a higher level of conservation. Epigenetic status refers to the status (green, inactive; yellow, medium; red, active) of the histone modification on an Alu element. The size of circles represents the possibility of transcriptional regulator binding to Alu elements.

(B) Histone modification preferences on Alu shift toward an enhancer-like signature with increasing evolution age. Profiles here are based on unique-reads mapping and compared to nucleosome controls. Random mapping is not applicable here because of common sequences within the subfamilies. Linear regression is used to evaluate the correlation between the relative enrichment level and age (represented by sequence divergence). Histone modification signals were normalized against nucleosome signals from MNase-seq. The regression slopes and Bonferroni-adjusted p values are listed on the right side of the heat plot.

(C) The average and standard deviations of motif densities for all 216 well-known motifs on each Alu subfamily (Experimental Procedures).

(D) Substitution rate between human and chimpanzee decreases with the age of the Alu. The left y axis represents the mutation rate between two species; the right y axis represents the medium sequence divergence across different elements within each human Alu subfamily.

See also Figure S4 and Tables S3 and S4.

factor NF-κB and trigger enhanceosome assembly and activation of transcription (Apostolou and Thanos, 2008). Therefore, similar to proto-genes, we propose that Alu elements, with their enhancer-like sequence and epigenetic features, are proto-enhancers, defined as sequences that are prone to evolve into functional enhancers, given the right conditions and further evolutionary selection (Figure 4A). Such a model would predict (1) the genomic, epigenetic, and functional features of the Alu elements should follow an evolutionary continuum with evolutionary time, and/or the age of Alu in the genome; (2) because new enhancers will have gradually emerged from proto-enhancers, older elements should have a higher proportion of functional sites; and (3) because they gain functional advantages, older elements should be more conserved between closely related species, such as human and chimpanzee.

To test these predictions, we sorted the ten subfamilies of Alu elements according to their ages in the human genome (Giordano et al., 2007) and compared their average histone modifications, CpG content, distance to nearby TSSs, enrichment for regulatory motifs, nearby gene expression, and mutation rates between human and chimpanzee.

Relative to H3, most of the active histone modifications show a gradual enrichment over time, in particular, the enhancer mark H3K4me1, which is enriched with age from −0.66 $\log_2$(fold) in AluYa5 to 1.39 in AluJo. Even the active enhancer mark H3K27ac becomes less excluded on the Alu elements over time. In contrast, H3K9me3, a heterochromatin marker, showed a slight trend for gradual exclusion on the Alu elements over time (Figure 4B). Meanwhile, as genomic background controls, the flanking sequences of the Alu subfamilies do not show such strong enrichment of enhancer-like histone marks, or a histone modification continuum with age (Figure S4A).

A gain of enhancer signatures may also be accompanied by a gain of TF binding motifs over time. Indeed, we observed a gradually increased sequence match to TF binding motifs with an increase of evolutionary age across Alu subfamilies (Figure 4C; motif matching scores for different TFs and the top ten TFs with the best match scores are listed in Table S3 and S4). In addition, the CpG content showed a dramatic decrease over the evolutionary age of the Alu subfamilies, with the old Alu subfamilies reaching a level extremely close to that of enhancers (Figure S4B). Interestingly, and consistently, the TF motifs are on average less CpG-enriched than Alu elements (motif: 0.01246756, Alu: 0.02218925; t test p value $<2.2 \times 10^{-16}$).

As a consequence of gaining enhancer signatures and TF motifs, nearby gene expression could also be affected. We only observed significantly lower gene expression levels surrounding the most recently inserted Alu subfamilies, which may reflect an immediate effect as nearby gene expression is disrupted by Alu insertion into functional elements (Figure S4C). With time, the Alus retained in the genome become less disruptive to gene expression, or more tolerated. However, because both transcriptional activators and repressors can bind to the proto-enhancers in a tissue-specific manner, they may not necessarily display progressively increased expression over time. Consistently, the genes next to known enhancers also do not show higher expression levels than the genomic average (Figure S4C).

The de novo functions gained by Alu element insertion could provide advantages over the course of natural selection and lead to higher conservation levels with time. We tested this prediction by comparing the substitution rates of Alu elements between human and chimpanzee. Relative to substitution rates in their flanking regions, older Alus do show progressively lower substitution rates, and thus a higher conservation level. These are in stark contrast with the sequence divergence across different Alu elements within each subfamily over time (Figure 4D). Furthermore, consistent with the higher conservation levels observed for Alu elements that are closer to an upstream TSS (Figure S1C), the genomic distribution of Alu also follows a trend of moving closer to TSSs with age, from an average of 180 kb for the youngest to nearly 80 kb for the oldest, whereas putative enhancers are generally close to TSS (Figure S4D).

Taken together, our results show that Alus of different ages can be placed in an evolutionary continuum of both structural and functional features, implicating them as a repertoire for future enhancers that could evolve through natural selection.

DISCUSSION

In this study, we found that Alus are significantly more conserved at gene proximal regions. Alu elements in the human, and conceivably other primate genomes, have distinctive active epigenetic characteristics that are similar to nonactive enhancers. Alu elements are also similar to enhancers in that they are enriched for enhancer-like histone modifications in a tissue-specific manner, and preferentially engage in long-distance interactions with gene promoters and with themselves.

These observations are consistent with our hypothesis that the sequence and epigenetic features of Alu make them proto-enhancers. With partial enhancer-like features, given the right conditions and further evolutionary selection, some of these proto-enhancers will evolve into functional enhancers (Figure 4A). Such a model is supported by the observed evolutionary continuum (based on evolutionary age) of both the epigenetic and functional features displayed by Alu elements. The distance of Alu elements to TSSs resembles those of "shadow enhancers," which locate 10–20 kb away from TSSs (Hong et al., 2008) and often function redundantly to the main gene proximal enhancers to ensure the precision and robustness of their regulation. Such functional redundancy is necessary for proto-enhancers, allowing for trial and error during evolutionary selection.

Other repeats that we did not examine may also have similar enhancer-like features and serve as proto-enhancers. Similarly, not all enhancers are derived from Alus as many enhancers appeared in the genome before Alu. We observed H3K36me3, in a similar way to H3K4me1, shows genomic position-independent enrichment on Alus. Because H3K36me3 is a transcription elongation mark, this enrichment may reflect transcription activity on some Alu elements, which is not impossible given that many enhancers generate eRNA (enhancer transcribed/derived RNA). In addition to Alu, we also identified a genomic region that is significantly more conserved for L1 (Figure S1C), which is well known to be important for higher-order genomic organization, in particular, tethering of the chromatin to the nuclear envelope through nuclear lamina (Meuleman et al., 2013). However, the conserved positions for L1 are much farther away from TSS than for Alus, consistent with them having different functions and being under different evolutionary pressures. This further indicates that it is not the genomic position per se that contributes to the conservation; otherwise, one would expect to see the same genomic region significantly conserved for all types of repeats (including L1 and Alu).

The increasing sequence divergence of different Alu elements, along with decreasing substitution rate of an Alu element within the same subfamily over time, indicates divergence of mutational direction across different Alu elements (e.g., different Alu elements may harbor different TF motifs), hence functional divergence among different Alu elements and functional selection.

Although gene duplication contributes to the generation of new genes, de novo gene birth from proto-genes in the noncoding genomic regions also contributes significantly toward the evolution of new genes (Carvunis et al., 2012; Li et al., 2010; Wu et al., 2011). Similarly, although sequence duplication and transcriptional network rewiring are important mechanisms during the evolution of new enhancers (Li and Johnson, 2010; Tuch et al., 2008), the de novo birth of new enhancers from the enormous number of proto-enhancers harbored by Alu may play a critical role in shaping the transcriptome and regulatory network of the primate genomes.

EXPERIMENTAL PROCEDURES

Data Sets

Human RefSeq genes and the RepeatMasker track were obtained from http://genome.ucsc.edu in March 2006; all other data sets are summarized in the Supplemental Experimental Procedures.

Mapping High-Throughput Sequencing Tags to Repetitive Elements

For MNase-seq and ChIP-seq, unmapped tags in FASTQ format were first filtered to remove low complexity tags with a dusty-score >20 using the

DUST algorithm in the "ShortRead" R package. SOAP2 (Li et al., 2009) was then used to align filtered tags to hg18 genome sequences. For redundant all reads mapping, multiple mapped tags were reported randomly to a matching coordinate. For unique-reads mapping, all multiple mapped tags were discarded.

Preferences of Histone Modifications on Transposon Elements

For a specific histone modification *mod*, its $preference_{mod}$ on a certain family of transposon elements, whose total copy number is *N*, was calculated as

$$Preference_{mod} = \log_2\left[\left(\sum_1^N Reads_{i,mod} \Big/ TotalReads_{mod}\right) \Big/ \times \left(\sum_1^N Reads_{i,bg} \Big/ TotalReads_{bg}\right)\right],$$

where $Reads_{i,mod}$ stands for ChIP-seq tag counts for the histone modification *mod* on the transposable element *i*. Nucleosome mapping MNase-seq, histone H3 ChIP-seq, or IgG control experiment was used for the background distribution of histones, indicated by *bg*. For IMR90 and H1 data from NREMC, no MNase-seq data were available, so input data were used as background.

Analyzing Long-Range Interacting DNAs with Hi-C-Seq DNA Interactome Maps

We only used Hi-C tag pairs that mapped to the same chromosome and removed tag pairs that were <20 kb apart. We selected the tag pairs, requiring one end of the tag pairs to be uniquely mapped to Alu, MIR, MaLR, L1, putative enhancers, or DHS and then examined the distance (d1, Figure 3A) from their interacting tag (the other end of the pair) to the nearest TSS to evaluate their preference for gene promoters. To test the statistical significance of differences in each Hi-C-seq analysis, we repeated each analysis 100 times, each time using only 10% of randomly selected sequence tags and then calculated two-tailed t test p values between a group and the genome background and for all other pairwise comparisons. See the Supplemental Informationfor normalization procedure.

Testing the Evolutionary Continuum of Alu Elements

Ten subfamilies of Alu elements with different estimated age were selected using information in RepeatMasker. The order of a subfamilies' evolutionary age was based on the sequence divergence obtained from Giordano et al. (2007). Sequence and functional features were calculated for all subfamilies to examine the evolutionary continuum. Homer (Heinz et al., 2010) was used to scan the motifs of all 216 well-known transcription factors on all Alu subfamilies. For each motif, *Z* score transformation across the ten subfamilies was used to compare the motif density among the subfamilies.

Further details of the experimental procedures are described in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.03.011.

AUTHOR CONTRIBUTIONS

J.-D.J.H. conceived and designed the study. M.S. and D.H. performed the computational analysis. M.S. and X.Y. performed the experiments. J.-D.J.H., M.S., D.H., J.B.-K., and X.Y. interpreted the data and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Nicholas Baker (AECOM) for insightful and critical reading of the manuscript and Drs. Micheal Levine (UC Berkeley), Keji Zhao (NIH), and Haipeng Li for discussions and invaluable suggestions. This work was funded by grants from the National Natural Science Foundation of China (NSFC) (grants #31210103916, #91019019, and #31150110469), the Chinese Ministry of Science and Technology (grant #2011CB504206), and the Chinese Academy of Sciences (CAS) (grants #KSCX2-EW-R-02 and KSCX2-EW-J-15) and stem cell leading project XDA01010303 to J.-D.J.H. J.B.-K. holds a CAS Fellowship (#2011Y1SB05).

Received: March 16, 2013
Revised: January 29, 2014
Accepted: March 5, 2014
Published: April 3, 2014

REFERENCES

Apostolou, E., and Thanos, D. (2008). Virus Infection Induces NF-kappaB-dependent interchromosomal associations mediating monoallelic IFN-beta gene expression. Cell *134*, 85–96.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

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

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature *441*, 87–90.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D., et al. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. *16*, 123–131.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931–21936.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Emera, D., and Wagner, G.P. (2012). Transformation of a transposon into a derived prolactin promoter with function during human pregnancy. Proc. Natl. Acad. Sci. USA *109*, 11246–11251.

Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nat. Rev. Genet. *10*, 605–616.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., and Warburton, P.E. (2007). Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput. Biol. *3*, e137.

Göndör, A., and Ohlsson, R. (2009). Chromosome crosstalk in three dimensions. Nature *461*, 212–217.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell *6*, 479–491.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hong, J.W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147–151.

Li, H., and Johnson, A.D. (2010). Evolution of transcription networks—lessons from yeasts. Curr. Biol. *20*, R746–R753.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.

Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., and Wang, W. (2010). A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. *20*, 408–420.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Liu, A.Y., and Bradner, R.C. (1993). Elevated expression of the human mitochondrial hinge protein gene in cancer. Cancer Res. *53*, 2460–2465.

Lynch, V.J., Leclerc, R.D., May, G., and Wagner, G.P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat. Genet. *43*, 1154–1159.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Oliviero, S., and Monaci, P. (1988). RNA polymerase III promoter elements enhance transcription of RNA polymerase II genes. Nucleic Acids Res. *16*, 1285–1293.

Polak, P., and Domany, E. (2006). Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics *7*, 133.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Ruiz-Narváez, E.A., and Campos, H. (2008). Evolutionary rate heterogeneity of Alu repeats upstream of the APOA5 gene: do they regulate APOA5 expression? J. Hum. Genet. *53*, 247–253.

Santangelo, A.M., de Souza, F.S., Franchini, L.F., Bumaschny, V.F., Low, M.J., and Rubinstein, M. (2007). Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet. *3*, 1813–1826.

Smit, A.F., and Riggs, A.D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. *23*, 98–102.

Tuch, B.B., Li, H., and Johnson, A.D. (2008). Evolution of eukaryotic transcription circuits. Science *319*, 1797–1799.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Wu, D.D., Irwin, D.M., and Zhang, Y.P. (2011). De novo origin of human protein-coding genes. PLoS Genet. *7*, e1002379.