5-24-2006

Recently integrated Alu retrotransposons are essentially neutral residents of the human genome

Richard Cordaux
Center for BioModular Multi-Scale Systems

Jungnam Lee
Center for BioModular Multi-Scale Systems

Liv Dinoso
Center for BioModular Multi-Scale Systems

Mark A. Batzer
Center for BioModular Multi-Scale Systems

Recommended Citation

Cordaux, R., Lee, J., Dinoso, L., & Batzer, M. (2006). Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene, 373* (1-2), 138-144. https://doi.org/10.1016/j.gene.2006.01.020

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 373 (2006) 138 – 144

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Recently integrated *Alu* retrotransposons are essentially neutral residents of the human genome

Richard Cordaux, Jungnam Lee, Liv Dinoso, Mark A. Batzer *

Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA

Received 16 December 2005; received in revised form 18 January 2006; accepted 21 January 2006
Available online 9 March 2006
Received by W. Makalowski

Abstract

Alu elements represent the largest family of human mobile elements in copy number. A controversial issue with implications for both *Alu* biology and human genome evolution is whether selective pressures are affecting *Alu* elements on a large scale. To address this issue, we analyzed the genomic distribution of the three youngest known human *Alu* subfamilies (Ya5a2, Ya8 and Yb9) in conjunction with their insertion polymorphism status in the human population, since selection can only act on polymorphic elements. Our results indicate that: (i) polymorphic and fixed recently integrated *Alu* elements are found in genomic regions whose GC contents are statistically indistinguishable, and (ii) recently integrated *Alu* elements are inserted randomly, regardless of the GC content of the surrounding genomic DNA. These results provide strong evidence that recently integrated "young" *Alu* elements are not subject to positive or negative selection on a large scale. Therefore, young *Alu* elements can be regarded as essentially neutral residents of the human genome. These results also imply that selective processes specifically targeting *Alu* elements can be ruled out as explanations for the accumulation of *Alu* elements in GC-rich regions of the human genome.

Keywords: *Alu* elements; SINEs; Retrotransposons; Human; Recently integrated subfamilies; Neutral evolution

1. Introduction

Mobile elements constitute nearly half of the human genome (Lander et al., 2001). Among them are *Alu* retrotransposons, ~300-bp-long interspersed repeats which have been inserting in primate genomes for the past 65 million years (My) (Batzer and Deininger, 2002). They have reached over one million copies in the human genome, making them the largest family of human mobile elements by copy number (Lander et al., 2001; Batzer and Deininger, 2002). *Alu* elements have had a substantial impact on the architecture of the genome, and their ongoing expansion has resulted in various genetic disorders (Deininger and Batzer, 1999; Batzer and Deininger, 2002; Bailey et al., 2003; Deininger et al., 2003; Chen et al., 2005). The long-lasting presence of *Alu* elements along with their high copy number in the genome raises questions regarding the interactions of these elements with their primate host genomes during evolution. Do they possess a function that could lead to their massive positive selection? Are they simply neutral residents of the genome? Or, given that there is evidence of negative selection against at least some *Alu* insertions, does negative selection represent an important force acting on *Alu* elements?

Several potential functions have been proposed for *Alu* elements, such as stimulation of protein synthesis under stress conditions and regulation of gene expression (Schmid, 1998). These functions could drive the positive selection of *Alu* elements. On the other hand, it has been noted that "although there are numerous cases where individual *Alu* elements have had a positive impact on the human genome, it might be argued that none of them has been confirmed as a function" (Deininger and Batzer, 1999). Nevertheless, in the seminal publication of the human genome sequence, it was reported that *Alu* elements are not uniformly distributed across the human genome, as older *Alu* elements are preferentially found in GC-rich regions while younger *Alu* elements are slightly more abundant in AT-rich regions (Lander et al., 2001). This distribution shift was

Abbreviations: My, million years.

* Corresponding author. Tel.: +1 225 578 7102; fax: +1 225 578 7113.
E-mail address: mbatzer@lsu.edu (M.A. Batzer).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.01.020

interpreted as evidence of positive selection on the *Alu* family due to forces acting to maintain them in GC-rich, gene-rich regions (Lander et al., 2001). Subsequently, this conclusion was questioned based on inconsistencies with population genetics theory regarding the time scale on which the selective process would occur (Brookfield, 2001; Batzer and Deininger, 2002). However, a recent comparison of the human and chimpanzee genome sequences suggested that the *Alu* distribution shift might take place in a time window that is more compatible with population genetics predictions (Chimpanzee Sequencing and Analysis Consortium, 2005) and consequently with a possible selective process acting on *Alu* elements.

Because selection can only operate while *Alu* elements are polymorphic for insertion presence/absence in the human population, a positive selection-based model predicts that *Alu* elements fixed in the population (i.e. present in all individuals) should be preferentially found in GC-rich regions as a result of the completed selection process. By contrast, polymorphic elements are expected to show a more even distribution between AT-rich and GC-rich regions (more closely reflecting the initial insertion pattern of *Alu* elements) because the selection process is still incomplete. On the other hand, if *Alu* elements are not positively selected to accumulate in GC-rich regions of the genome, polymorphic and fixed elements are predicted to show similar genomic distributions. To test these hypotheses, we determined the insertion polymorphism status and genomic distribution for *Alu* elements belonging to the three youngest known human *Alu* subfamilies (Ya5a2, Ya8 and Yb9). An analogous approach has previously been used to test for selection on the base composition of isochores (Belle and Eyre-Walker, 2002). By contrast, here we focused exclusively on the youngest *Alu* subfamilies: (i) to ensure that any distribution shift between fixed and polymorphic elements could not be ascribed to non-selective, time-dependent processes that might take place in older subfamilies which possess lower levels of polymorphic elements, and (ii) to include in the analyses most known copies in each subfamily, thus avoiding any bias that could potentially arise from the analysis of very large *Alu* subfamilies for which subsets of elements would have to be selected.

2. Materials and methods

2.1. Identification of the youngest human Alu elements

The three youngest known human *Alu* subfamilies Ya5a2, Ya8 and Yb9 have been previously described but only subsets of these subfamilies were analyzed for polymorphism (Roy et al., 1999; Roy et al., 2000; Roy-Engel et al., 2001). To obtain a more exhaustive view of these subfamilies, we screened the May 2004 freeze of the human genome sequence available in the UCSC genome database (http://genome.cse.ucsc.edu/), using the BLAT program and the subfamily consensus sequences as queries. In addition, we performed searches with subfamily-specific oligonucleotide sequences (Ya5a2: AGAGATCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACTAAAAATACAAAAAAA; Ya8: CTACAAAAAATAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCT; Yb9: TAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTG).

2.2. PCR amplification

Alu elements were extracted along with 1000 bp of genomic sequence flanking each element on both sides. The RepeatMasker program (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) was then used to annotate all known repeat elements within the extracted flanking sequences. Oligonucleotide primers were designed in repeat-free portions (if any) of the flanking sequences of *Alu* loci, using the program Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi).

To determine the polymorphism status of the *Alu* elements in the human population, we genotyped them in a representative human population panel composed of DNA samples from 20 African Americans, 20 Europeans, 20 Asians (isolated from peripheral blood lymphocytes and available from previous studies in our lab) and 20 South Americans (HD17 and HD18 cell line panels purchased from the Coriell Institute for Medical Research). PCR amplification of each locus was performed in 25-μl reactions, using 280 nM of each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 2.0 mM $MgCl_2$, 10 mM Tris–HCl (pH 8.4), 2.5 units Taq DNA polymerase and 20 ng DNA. Reactions were subjected to an initial denaturation step of 95 °C for 5 min, followed by 35 cycles of 30 s at 95 °C, 30 s at optimal annealing temperature and 1 min at 72 °C, followed by a final extension step at 72 °C for 10 min. Resulting PCR products were separated on 2% agarose gels, stained with ethidium bromide and visualized using UV fluorescence. Detailed information on each locus including primer sequences, annealing temperature, PCR product sizes, chromosomal location and polymorphism status is published as Supplemental Table S1 and is also available in the Publications section of our website (http://batzerlab.lsu.edu).

2.3. Analyses of Alu subfamily sequence variation

The software NETWORK version 4.1 (http://www.fluxus-technology.com/sharenet.htm) (Bandelt et al., 1999) was used to calculate the proportion of elements within each subfamily that are identical to the subfamily consensus sequence and to determine the age of the subfamilies. We used an average mutation rate of 0.7965 ($231 \times 0.0015 + 50 \times 0.0015 \times 6$) mutations per *Alu* element per My (or one mutation per *Alu* element per 1.2555 My), based on the consensus sequence of the *Alu*Y subfamily which contains 231 non-CpG sites assumed to mutate at a neutral mutation rate of 0.0015 mutations per site per My (Miyamoto et al., 1987) and 50 CpG sites estimated to mutate 6 times faster than non-CpG sites (Xing et al., 2004).

We also conducted BLAST searches of the human genome reference sequence (http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html) using the Ya5 and Yb8 *Alu* consensus sequences (Carter et al., 2004; Otieno et al., 2004) as queries and the following parameters: expect threshold of $1e^{-100}$ and no filter. We identified a total of 155 autosomal exact matches, out of a

Table 1
Characteristics of the three youngest known human *Alu* subfamilies

Alu subfamily	Ya5a2	Ya8	Yb9
Copy number in human genome	46	36	57
Age±SD (million years)	0.63±0.14	1.36±0.26	1.78±0.24
Proportion of copies identical to consensus	65%	28%	26%

total of ~2850 Ya5 and Yb8 copies in human autosomes (Carter et al., 2004; Otieno et al., 2004).

2.4. Analyses of flanking sequence GC content

We extracted 10 kb of genomic sequence flanking each side of the 139 young *Alu* elements via the UCSC genome database. We then used the GeeCee program (http://bioweb.pasteur.fr/seqanal/interfaces/geecee.html) to calculate the percentage of GC nucleotides in all 20-kb sequence windows. Descriptive statistics and statistical tests were performed using an Excel spreadsheet.

3. Results

3.1. Polymorphism status of the youngest human Alu elements

Using subfamily consensus sequences and subfamily-specific oligonucleotides as queries, we identified a total 139 autosomal *Alu* elements belonging to the Ya5a2, Ya8 and Yb9 subfamilies in the human genome reference sequence (Lander et al., 2001). Our genome-wide subfamily copy numbers (Table 1) are similar to those reported in previous studies (Roy et al., 1999; Roy et al., 2000; Roy-Engel et al., 2001), suggesting that our detection strategy based on two different types of query sequences recovered most (if not all) Ya5a2, Ya8 and Yb9 *Alu* elements contained in the human genome reference sequence. A network analysis (Bandelt et al., 1999; Cordaux et al., 2004) of subfamily sequence variation confirmed that the three *Alu* subfamilies are very recent in origin, with estimated ages ranging from 0.6 to 1.8 My (Table 1). Another line of evidence further supporting the very recent origin of these *Alu* elements is that within each subfamily, a large proportion of the copies (26–65%, Table 1) are identical to their subfamily consensus sequences. By comparison, only ~5% of the autosomal copies of the older Ya5 and Yb8 subfamilies are identical to their subfamily consensus sequences, as demonstrated by BLAST searches of the human genome reference sequence (see Materials and methods section).

To determine the insertion polymorphism status in the human population for the *Alu* elements belonging to the Ya5a2, Ya8 and Yb9 *Alu* subfamilies, we genotyped all the elements for which PCR assays could be designed. Some elements have inserted within other repeated sequences and were therefore not amenable to PCR. However, we were able to determine the polymorphism status of the majority (74%) of these *Alu* elements. Each of the 103 *Alu* elements was genotyped in a panel of 160 human chromosomes, resulting in the recovery of 60 fixed and 43 polymorphic elements. The high proportion of polymorphic elements (42%) provides additional evidence for the recent origin of the Ya5a2, Ya8 and Yb9 *Alu* subfamilies (Roy et al., 1999; Roy et al., 2000; Roy-Engel et al., 2001).

3.2. Genomic distribution of the youngest human Alu elements

Next, we calculated the GC content of 20-kb windows of flanking genomic DNA centered on each *Alu* element, and we compared the genomic distribution of fixed and polymorphic recently integrated *Alu* elements. We found that the average GC contents of fixed and polymorphic elements within each subfamily are similar (Table 2), and *t*-tests showed that these data are not significantly different within any of the subfamilies (Ya5a2, $P=0.67$; Ya8, $P=0.08$; Yb9, $P=0.40$). This suggests that within each subfamily, *Alu* elements are found in similar genomic environments regardless of their polymorphism status. To test for heterogeneity in flanking GC content among *Alu* subfamilies, we performed *t*-tests using the elements regardless of their polymorphism status. We found that there is no significant difference among the different *Alu* subfamilies (Ya5a2 vs. Ya8, $P=0.64$; Ya5a2 vs. Yb9, $P=0.58$; Ya8 vs. Yb9, $P=0.33$). These results indicate that it is reasonable to pool the data from the different *Alu* subfamilies to increase the sample sizes of the categories to be compared, and thereby the statistical power of the tests. A *t*-test comparing the flanking GC contents of fixed and polymorphic *Alu* elements with the combined data from the three subfamilies also showed no significant difference ($P=0.83$). We emphasize here that this test leads to the same conclusions as the tests performed on individual *Alu* subfamilies, although it is based on 3–4 times more observations than the individual subfamily tests. Thus, the small sample sizes do not appear to have compromised the power of the tests.

To further compare the genomic GC content surrounding recently integrated *Alu* elements, we drew the frequency distribution of polymorphic and fixed elements according to GC content (Fig. 1). This analysis showed that the two categories of young *Alu* elements have very similar flanking GC content distributions and a χ^2 test comparing polymorphic and fixed elements in the different GC content classes depicted in Fig. 1 demonstrated that the two distributions are not significantly different ($P=0.99$). We calculated that in order to make this test significant at the 5% level, it would require

Table 2
Genomic distribution of the three youngest known human *Alu* subfamilies

	Number of elements	Average GC content±SD (20-kb window)	Proportion in ≥41% GC content
Fixed elements	60	39.6±4.1%	36.7%
Ya5a2	12	39.2±3.6%	33.3%
Ya8	18	38.5±2.6%	16.7%
Yb9	30	40.4±4.9%	50.0%
Polymorphic elements	43	39.8±3.6%	41.9%
Ya5a2	26	39.8±4.2%	42.3%
Ya8	8	40.6±3.2%	50.0%
Yb9	9	39.0±1.9%	33.3%

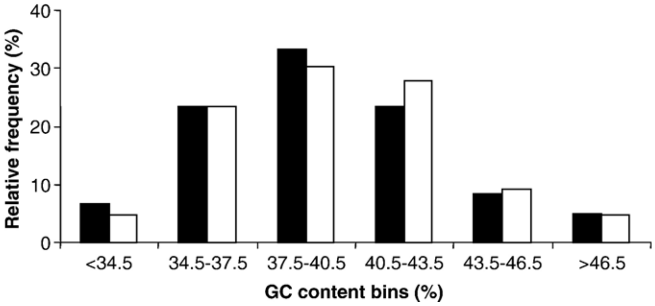

Fig. 1. Frequency distribution of fixed (black bars) and polymorphic (white bars) elements from the three youngest known human *Alu* subfamilies ($n=103$), according to the GC content of flanking genomic DNA (20-kb windows).

adding at least 30% more polymorphic elements all from the same most extreme GC content bins (i.e. <34.5% or >46.5%) (i.e. fixed elements cannot have been missed because they are present in all humans including the reference genome). It is highly unlikely that we would have missed such *Alu* elements because: (i) the flanking GC content of young *Alu* loci that were not amenable to PCR is not skewed towards extreme values (data not shown), and (ii) there is no reason to expect that polymorphic *Alu* loci not represented in the reference sequence would insert in GC content contexts that would be different from the reference sequence.

Our results indicated that, overall, 35–40% of young *Alu* elements are found in GC-rich regions of the genome, defined as regions with a GC content equal or higher than the 41% genome-wide average (Lander et al., 2001) (Table 2). Contrary to the expectations of a positive selection-based model in which fixed *Alu* elements would be accumulated in GC-rich regions as compared to polymorphic elements, we actually observed the opposite trend since ~37% of fixed elements are inserted in GC-rich regions vs. ~42% of polymorphic elements. However, the difference is not statistically significant (χ^2 test, $P=0.59$). Repeating this analysis for each subfamily separately revealed no significant difference between fixed and polymorphic elements (Ya5a2, $P=0.60$; Ya8, $P=0.08$; Yb9, $P=0.38$).

Our data also indicated that the majority (60–65%) of young *Alu* elements are found in AT-rich regions of the genome, defined as regions with a GC content lower than the 41% genome-wide average (Lander et al., 2001). The trend persists, although less clearly, if we take into account the fact that AT-rich regions represent 58% of the genome (Lander et al., 2001). These results suggesting a slight preferential insertion of the youngest *Alu* elements in AT-rich regions therefore concur with previously published results (Lander et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005; Hackenberg et al., 2005).

However, does the trend noted above significantly deviate from a random model of insertion? To address this question, we compared the number of *Alu* insertions expected in AT-rich and GC-rich regions under a random model of insertion to those observed for all 103 young *Alu* elements genotyped in this study. Because 58% of the genome consists of AT-rich regions (Lander et al., 2001), a random model of insertion predicts that 60 (103 × 58%) *Alu* elements will be in AT-rich regions and 43 (103 × 42%) will be in GC-rich regions. We observed that 63 *Alu* elements are inserted in AT-rich regions and 40 are in GC-rich regions. A χ^2 test indicated that the apparent insertional bias towards AT-rich regions does not significantly depart from a random model of insertion with respect to GC-richness of the surrounding genomic DNA ($P=0.55$). When the analysis was repeated with all 139 young *Alu* elements identified in this study and representing the nearly entire collection of existing Ya5a2, Ya8 and Yb9 elements, we found again that these subfamilies follow a random model of insertion with respect to GC content ($P=0.09$).

Having performed a number of statistical tests using the same dataset raises the issue of the use of corrections for multiple tests (such as the Bonferroni correction). We note that such corrections would result in lowering the significance threshold of each test to minimize significant tests arising by chance alone. As none of our tests was significant even without correcting for multiple testing, we conclude that the outcomes of the tests are even better supported.

4. Discussion

Our analyses of the youngest known human *Alu* subfamilies unambiguously show that the genomic distributions of polymorphic and fixed elements with respect to GC content are statistically indistinguishable. This conclusion is supported by a global analysis as well as separate analyses of three different subfamilies of recently integrated “young” *Alu* elements. This suggests that the patterns we observed cannot be attributed to bias from an atypical *Alu* subfamily altering the results, and that the relatively small sample sizes of the individual subfamilies did not compromise the power of the tests we employed. Actually, we identified a striking homogeneity in the genomic distribution of three different *Alu* subfamilies, although their copies have been produced by a number of independent source elements (Deininger et al., 1992; Cordaux et al., 2004). This strongly suggests that our results can be generalized to all recently integrated *Alu* elements, even though our analyses were restricted to only the young elements that are present in the human genome reference sequence and would have missed polymorphic elements that were absent from the reference sequence (Boissinot et al., 2004; Hedges et al., 2004).

4.1. Recently integrated Alu elements and selection

Because there is no difference in the genomic distribution of fixed and polymorphic elements with respect to GC content, our results provide evidence that young *Alu* subfamilies are not subject to large amounts of global or genome-wide positive selection in the human genome. This finding has important implications with respect to *Alu* evolution because it provides evidence that the successful expansion of *Alu* elements was not necessarily a result of their serving a function in their primate host genomes that would have been positively selected. Of course, it could be argued that selective pressures acting on *Alu* elements may have been high in the past, with strong positive selection operating only when the vast majority of *Alu* elements was produced ~40 My ago (Shen et al., 1991; Britten, 1994;

Kapitonov and Jurka, 1996). However, it is noteworthy that the burst of *Alu* amplification ~40 My ago was also accompanied by a burst in the formation of processed pseudogenes (Ohshima et al., 2003; Zhang et al., 2003; Marques et al., 2005). If *Alu* elements had been subject to genome-wide positive selection ~40 My ago, there would be no reason for processed pseudogenes to show a concomitant burst of formation. Since both *Alu* elements and processed pseudogenes are generated by the same retrotransposition mechanisms (Esnault et al., 2000; Dewannieux et al., 2003), it appears more likely that the burst of *Alu* amplification reflects a general burst of retrotransposition rather than positive selection. Consequently, there is currently no evidence to suggest that the amount of positive selection acting on *Alu* elements might have decreased in the recent past.

Interestingly, our results indicate that the insertion pattern of young *Alu* elements fits a random model of insertion with respect to GC content. This is consistent with an analysis of the distribution of recently integrated *Alu* elements inserted in human chromosome 19 (Arcot et al., 1998). By contrast, it a priori seems at odds with genome-wide analyses that suggested that young *Alu* elements are preferentially inserted in AT-rich regions (Lander et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005; Hackenberg et al., 2005), although we note that the statistical significance of this observation has not been tested in any of the studies. Even though our data are consistent with the previously reported apparent insertion bias towards AT-rich regions, we show that it is not statistically different from a random insertion model. This finding has important implications because it provides evidence against large levels of genome-wide negative selection acting on young *Alu* elements.

Indeed, the conditional fixation time for a deleterious allele with a selective disadvantage s is $(2/s)\ln(2N_e)$ generations on average (Graur and Li, 2000). As the effective population size (N_e) of the human species is ~10,000 individuals (Graur and Li, 2000 and references therein), a deleterious allele with $s=0.5\%$ that would reach fixation in the human population is expected to do so in only ~100,000 years on average (assuming a generation time of 25 years). Given that the probability of fixation of a deleterious allele is extremely small (Graur and Li, 2000), most deleterious alleles with $s \geq 0.5\%$ will be lost in much less than ~100,000 years. In other words, if *Alu* elements are subject to negative selection, they are expected to be lost very quickly after they inserted in the genome. Therefore, if young *Alu* elements were selected against on a large scale, the impact of negative selection would be expected to already be detectable in the 1–2-My-old *Alu* subfamilies analyzed in this study. Moreover, it would make sense to expect a significant *Alu* distribution shift towards AT-rich regions because *Alu* elements are presumably more likely to be deleterious in GC-rich, gene-rich regions of the genome, as suggested by the de novo *Alu* insertions involved in genetic disorders (Deininger and Batzer, 1999; Chen et al., 2005). Since we find that young *Alu* elements have inserted randomly in the genome with respect to GC content, we conclude that our data provide evidence against large-scale negative selection acting on *Alu* elements.

We note that the apparent slight preference of young *Alu* elements for AT-rich regions (Lander et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005; Hackenberg et al., 2005; this study) might be the outcome of negative selection against disease-causing *Alu* insertions in GC-rich regions. However, these deleterious elements seem not to be numerous enough to make the observed distribution differ significantly from a random model of insertion, suggesting that negative selection against *Alu* elements has a limited effect on their overall genomic distribution. This conclusion is also intuitive because it is difficult to envision how primate species could sustain the load of many deleterious *Alu* insertions constantly bombarding their genomes for tens of My.

4.2. Insight into the remarkable Alu genomic distribution shift

Our results also offer new insight into the discussion about forces potentially responsible for the remarkable shift in *Alu* genomic distribution towards GC-rich regions. Because polymorphic and fixed *Alu* elements from the youngest subfamilies have indistinguishable genomic distributions that are not skewed towards GC-rich regions of the genome, we conclude that the *Alu* distribution shift occurs *after* the fixation of *Alu* elements. This rules out the role of selection as the mechanism responsible for shaping the genome-wide distribution of *Alu* elements in the human genome (Lander et al., 2001). The ages of the three young *Alu* subfamilies we analyzed further indicate that the effect(s) of the mechanism(s) responsible for the distribution shift begin(s) to be detectable only after a subfamily has resided in the genome for at least ~2 My. This timing might even be extended to at least 3–4 My, based on comparisons using limited subsets of elements belonging to older *Alu* subfamilies (Arcot et al., 1998; Belle and Eyre-Walker, 2002). As an alternative explanation to positive selection of *Alu* elements in GC-rich regions, the differential loss or deletion of *Alu* elements from AT-rich regions could also produce the *Alu* distribution shift (Deininger and Batzer, 1999; Brookfield, 2001; Lander et al., 2001; Batzer and Deininger, 2002; Jurka et al., 2004). This scenario is considered plausible because deletions are presumably better tolerated in AT-rich, gene-poor regions than in GC-rich, gene-rich regions, and evidence has been presented showing that unequal homologous recombination among *Alu* elements can result in deletions (Hackenberg et al., 2005; van de Lagemaat et al., 2005). By contrast, a recent comparison of the human and chimpanzee genome sequences indicated that the number of human-specific *Alu*–*Alu* recombination-mediated deletions is not biased towards AT-rich regions (Chimpanzee Sequencing and Analysis Consortium, 2005). As such, these deletions would need to be much larger in AT-rich than GC-rich regions so that they could more frequently encompass *Alu* elements. Although such size difference (if any) alone is unlikely to account for the entire *Alu* distribution shift, it may be of interest to investigate this possibility further.

Two other non-selective processes have been suggested as causes for the distribution shift in *Alu* elements. For instance, it has been observed that the proportion of duplicated *Alu* elements is slightly higher in GC-rich than in AT-rich regions (although the statistical significance of the difference has not been tested), leading to the suggestion that segmental duplications may contribute to the accumulation of *Alu* elements in GC-rich regions (Jurka et al., 2004). Nevertheless, as only ~2% of *Alu* elements seem to have resulted from passive duplications (Jurka et al., 2004), it seems unlikely that this process substantially contributed to the *Alu* distribution shift. Alternatively, the *Alu* distribution shift could involve a change in insertion site preferences, although the reason for such a change is unclear. Moreover, a past strong insertional bias towards GC-rich regions would be difficult to reconcile with the fact that *Alu* elements use an L1 element encoded endonuclease for insertion with an AT-rich consensus motif (Jurka, 1997; Gentles et al., 2005).

In sum, the factors responsible for shaping the *Alu* genomic distribution remain largely undetermined. Nevertheless, our results allow narrowing down the spectrum of potentially contributing factors because they provide strong evidence against the involvement of a selective process specifically targeting *Alu* elements.

5. Conclusion

In this study, we have generated a nearly exhaustive collection of the youngest *Alu* elements inserted in the human genome reference sequence. We showed that: (i) the genomic distributions of polymorphic and fixed young *Alu* elements with respect to GC content are statistically indistinguishable, and (ii) young *Alu* elements follow a random pattern of insertion with respect to GC content. These findings provide strong evidence against massive positive or negative selection acting on human young *Alu* elements. This suggests that young *Alu* elements can be considered as neutral residents of the human genome. Nevertheless, it is undisputed that some particular *Alu* loci have had negative or positive impacts on the genome. Therefore, we conclude that young *Alu* elements are *essentially* neutral residents of the human genome.

Acknowledgements

We thank D. Grover, D.J. Hedges and S.W. Herke for constructive discussions. This research was supported by Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05, (2000-05)-01 and (2001-06)-02; National Institutes of Health GM 59290, National Science Foundation grants BCS-0218338 and EPS-0346411, and the State of Louisiana Board of Regents Support Fund.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2006.01.020.

References

Arcot, S.S., et al., 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. J. Mol. Biol. 281, 843–856.

Bailey, J.A., Liu, G., Eichler, E.E., 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. 73, 823–834.

Bandelt, H.J., Forster, P., Rohl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16, 37–48.

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

Belle, E.M., Eyre-Walker, A., 2002. A test of whether selection maintains isochores using sites polymorphic for Alu and L1 element insertions. Genetics 160, 815–817.

Boissinot, S., Entezam, A., Young, L., Munson, P.J., Furano, A.V., 2004. The insertional history of an active family of L1 retrotransposons in humans. Genome Res. 14, 1221–1231.

Britten, R.J., 1994. Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Proc. Natl. Acad. Sci. U. S. A. 91, 6148–6150.

Brookfield, J.F., 2001. Selection on Alu sequences? Curr. Biol. 11, R900–R901.

Carter, A.B., et al., 2004. Genome-wide analysis of the human Alu Yb-lineage. Hum. Genomics 1, 167–178.

Chen, J.M., Stenson, P.D., Cooper, D.N., Ferec, C., 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum. Genet. 117, 411–427.

Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.

Cordaux, R., Hedges, D.J., Batzer, M.A., 2004. Retrotransposition of Alu elements: how many sources? Trends Genet. 20, 464–467.

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

Deininger, P.L., Batzer, M.A., Hutchison 3rd, C.A., Edgell, M.H., 1992. Master genes in mammalian repetitive DNA amplification. Trends Genet. 8, 307–311.

Deininger, P.L., Moran, J.V., Batzer, M.A., Kazazian Jr., H.H., 2003. Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. 13, 651–658.

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

Esnault, C., Maestre, J., Heidmann, T., 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24, 363–367.

Gentles, A.J., Kohany, O., Jurka, J., 2005. Evolutionary diversity and potential recombinogenic role of integration targets of non-LTR retrotransposons. Mol. Biol. Evol. 22, 1983–1991.

Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution, 2 ed. Sinauer Associates, Sunderland.

Hackenberg, M., Bernaola-Galvan, P., Carpena, P., Oliver, J.L., 2005. The biased distribution of alus in human isochores might be driven by recombination. J. Mol. Evol. 60, 365–377.

Hedges, D.J., Callinan, P.A., Cordaux, R., Xing, J., Barnes, E., Batzer, M.A., 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. Genome Res. 14, 1068–1075.

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

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

Kapitonov, V., Jurka, J., 1996. The age of Alu subfamilies. J. Mol. Evol. 42, 59–65.

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

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., Kaessmann, H., 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3, e357.

Miyamoto, M.M., Slightom, J.L., Goodman, M., 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. Science 238, 369–373.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., Okada, N., 2003. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. 4, R74.

Otieno, A.C., et al., 2004. Analysis of the Human Alu Ya-lineage. J. Mol. Biol. 342, 109–118.

Roy-Engel, A.M., et al., 2001. Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159, 279–290.

Roy, A.M., et al., 1999. Recently integrated human Alu repeats: finding needles in the haystack. Genetica 107, 149–161.

Roy, A.M., et al., 2000. Potential gene conversion and source genes for recently integrated Alu elements. Genome Res. 10, 1485–1495.

Schmid, C.W., 1998. Does SINE evolution preclude Alu function? Nucleic Acids Res. 26, 4541–4550.

Shen, M.R., Batzer, M.A., Deininger, P.L., 1991. Evolution of the master Alu gene(s). J. Mol. Evol. 33, 311–320.

van de Lagemaat, L.N., Gagnier, L., Medstrand, P., Mager, D.L., 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res. 15, 1243–1249.

Xing, J., Hedges, D.J., Han, K., Wang, H., Cordaux, R., Batzer, M.A., 2004. Alu element mutation spectra: molecular clocks and the effect of DNA methylation. J. Mol. Biol. 344, 675–682.

Zhang, Z., Harrison, P.M., Liu, Y., Gerstein, M., 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 13, 2541–2558.