

The Biased Distribution of Alus in Human Isochores Might Be Driven by Recombination

Michael Hackenberg,¹ Pedro Bernaola-Galván,² Pedro Carpena,² José L. Oliver¹

¹ Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain

² Departamento de Física Aplicada II, Universidad de Málaga, Málaga, Spain

Received: 28 June 2004 / Accepted: 1 October 2004 [Reviewing Editor: Dr. Jerzy Jurka]

Abstract. Alu retrotransposons do not show a homogeneous distribution over the human genome but have a higher density in GC-rich (H) than in AT-rich (L) isochores. However, since they preferentially insert into the L isochores, the question arises: What is the evolutionary mechanism that shifts the Alu density maximum from L to H isochores? To disclose the role played by each of the potential mechanisms involved in such biased distribution, we carried out a genome-wide analysis of the density of the Alus as a function of their evolutionary age, isochore membership, and intron vs. intergene location. Since Alus depend on the retrotransposase encoded by the LINE1 elements, we also studied the distribution of LINE1 to provide a complete evolutionary scenario. We consecutively check, and discard, the contributions of the Alu/LINE1 competition for retrotransposase, compositional matching pressure, and Alu overrepresentation in introns. In analyzing the role played by unequal recombination, we scan the genome for Alu trimers, a direct product of Alu–Alu recombination. Through computer simulations, we show that such trimers are much more frequent than expected, the observed/expected ratio being higher in L than in H isochores. This result, together with the known higher selective disadvantage of recombination products in H isochores, points to Alu–Alu recombination as the main agent provoking the density shift of Alus toward the GC-rich parts of the genome. Two independent pieces of evidence—the lower evolutionary divergence shown by recently in-

serted Alu subfamilies and the higher frequency of old stand-alone Alus in L isochores—support such a conclusion. Other evolutionary factors, such as population bottlenecks during primate speciation, may have accelerated the fast accumulation of Alus in GC-rich isochores.

Key words: Alu — LINE1 — Retrotransposons — Alu–Alu recombination — Selection — Isochores — Human genome

Introduction

A large part of the human genome (about 45%) is made up of mobile, repeated elements (Jurka 1995; Batzer and Deininger 2002; Deininger et al. 2003). The two most frequent interspersed repeats are Alu and LINE1 retrotransposons, with around 1,100,000 and 700,000 copies in the genome, respectively. Both mobilize (i.e., retrotranspose) via an RNA intermediate. Alus, relatively short elements (≈ 300 bp), are rich in GC and CpG dinucleotides, containing roughly one-third of all CpGs in the human genome. LINE1s are rather long, GC-poor elements (≈ 6500 bp), although in the genome they are often 5' truncated and their average length may reach less than 500 bp in GC-rich isochores (see Table 1). The LINE1 element is an autonomous retrotransposon that has two open reading frames, one of which codes for a retrotranscriptase/endonuclease (retrotranspos-

Table 1. Numbers, average lengths, GC ranges and mean %GC of the different isochore classes

Isochore class	Isochores				Alus			LINE1		
	<i>N</i>	Average length (bp)	%GC range	Mean %GC	<i>N</i>	%GC	Average length (bp)	<i>N</i>	%GC	Average length (bp)
L1	1,020	797,025	< 37.77	36.33	150,157	51.03	282	178,476	33.95	974
L2	1,634	495,487	37.77–41.32	39.51	258,670	51.16	283	163,087	34.51	887
H1	1,843	335,261	41.32–46.91	43.80	333,435	51.41	281	110,117	35.07	658
H2	724	260,731	46.91–51.23	48.86	120,123	51.64	281	26,574	36.58	529
H3	315	235,565	51.23–54.75	52.72	52,553	51.93	281	7,777	38.46	483
H4	200	163,698	≥54.75	57.32	16,322	52.53	279	2,542	41.41	435
Total	5,736				931,260			488,633		

Note. Only isochores longer than 50 kb were included. The numbers, lengths, and mean %GC contents of the Alus and LINE1s harbored by the different isochore classes are also shown. Note that in the definition of the isochores (GC ranges) the lower boundary is included and the upper one excluded.

ase). It has been recently demonstrated that LINE1 and Alus are amplified by the same enzymatic apparatus of active LINE1 elements (Dewannieux et al. 2003). Therefore, Alus should have the same insertion preference for AT-rich regions as LINE1. The binding site of the endonuclease is TT/AAAA, the slash indicating the cutting point. Indeed, the preference for AT-rich regions shown by young Alus has recently been demonstrated (Pavliček et al. 2001; IHGSC 2001).

While the density maximum of the LINE1 elements remains in AT-rich regions, the maximum for Alus is found in the H2 isochores (Bernardi 2001). Thus, if the insertion pattern is the same for Alus and LINE1s, the question arises: What evolutionary mechanism shifts the Alu density maximum from L to H isochores? This question has raised controversy in recent years, and various proposals have been advanced, most of which involve either positive or negative/purifying selection. Positive selection could act only if Alus had some identifiable function that would favor the organism. In recent years, various positive effects of Alus have been proposed (Chu et al. 1998; Schmid 1998; Deininger and Batzer 1999). For example, in many species, Alus are transcribed under conditions of stress, and the resulting RNAs specifically bind a particular protein kinase (PKR), blocking its ability to inhibit protein translation (Chu et al. 1998; Schmid 1998). Alu RNAs would thus promote protein translation under stress. Their location in open chromatin (which tends to correlate with GC content of the genomic region) appears to facilitate this task (Smit 1999). Alus can also insert into mature messenger RNAs via a splicing-mediated process termed exonization (Lev-Maor et al. 2003). Through exonization, intronic Alus can be converted into new coding exons. Indeed, about 5% of alternatively spliced internal exons in the human genome originate in Alu sequences (Sorek et al. 2002). An-

other possible function of Alus relates to gene expression levels. Alus are CpG rich, and by inserting near a gene this newly introduced CpG island may alter the expression pattern of the gene (Britten 1996). Since the genes are denser in GC-rich than in GC-poor regions, an equivalent distribution of the Alus may be positively selected for. However, generalized positive selection for Alus is problematic (Deininger and Batzer 1999)—that is, most of the attributed Alu functions refer to individual or few Alu repeats and not to the full set of repeats.

Consequently, most approaches to this problem involve negative selection. This mechanism could also provoke the Alu density shift if it were able to exclude or remove Alus more effectively from AT-rich than from GC-rich regions. Different negative influences of the Alus in AT-rich regions have been proposed. For example, the accumulation of Alus in GC-poor regions is counter-selected because it would severely change the local composition and/or the chromatin structure of these genomic regions, possibly affecting gene transcription (Rynditch et al. 1998). Bernardi's group also proposed a negative selection theory in which the major part of compositional matching is achieved by selective gain and loss of DNA, particularly repetitive DNA (Pavliček et al. 2001). This hypothesis is supported by comparative analyses of mammalian genomes (Pavliček et al. 2002; Paces et al. 2004).

A related mechanism is the compositional matching of repeats to the isochores harboring them. Compositional matching (or adjustment) to different genomic GC contents has been demonstrated in homologous genes and noncoding sequences of microorganisms and mitochondrial genomes (Jukes and Bushan 1986), as well as in genes moved between genomes (Oliver et al. 1990; Martínez-Zapater et al. 1993). Alu sequences can undergo compositional matching as well (Filipski et al. 1989)—base substi-

tutions decrease the average GC content of Alus located in AT-rich regions, whereas the Alus located in GC-rich genomic regions do not change their already high GC content. In this way, compositionally non-matching, recently inserted Alus could undergo stronger mutational pressure than the repeats adjusted to their host isochores. Since most of the elements found in the genome have adjusted their density maximum to regions where they have a highest compositional match, compositional matching can be considered responsible for the observed density distribution.

Another mechanism which appears to influence the distribution of these retrotransposons, and which does not require selection to act, involves Alu/LINE1 interactions. It has been argued (Gu et al. 2000) that Alus may switch their insertion preferences toward the GC-richer parts, thereby avoiding the competition for retrotransposase with the LINE1 elements in the L isochores.

A last mechanism that might be involved in the Alu density shift is the unequal homologous recombination among Alus repeats (Alu–Alu recombination), which is thought to be a recurrent process. Some authors (Deininger and Batzer 1999; Batzer and Deininger 2002) have argued that the selective disadvantage posed by Alu–Alu recombination is higher in H than in L isochores, given the positive correlation between gene density and GC content. The deletion rate of Alu elements should then be higher in L isochores. The evolutionary outcome of such a process would be a progressive decline in the relative density of Alus in L isochores, thus leading to the density shift.

However, by considering only average values or combining old with young elements, it is difficult to identify which of these is the main mechanism involved. Here, we present a new approach to this subject by analyzing Alu and LINE1 elements as a function of their evolutionary age, isochore membership, and intron vs. intergene location. This enables us to trace Alu evolution in the genome and in different genomic compartments, thereby disclosing the roles of the different mechanisms potentially involved in the Alu density shift.

Data and Methods

Chromosome Sequences and Gene Data

We used the human reference sequence (April 2003 freeze; UCSC version hgl5), based on NCBI Build 34 and produced by the International Human Genome Sequencing Consortium (IHGSC). The 24 chromosome sequences, in FASTA format, were downloaded from <ftp://genome.ucsc.edu/goldenpath/10april2003/big-Zips/chromFa.zip>. Gene data were derived from the GeneID database downloaded from the UCSC Genome Browser (<http://genome.cse.ucsc.edu/>).

Finding Isochores

In analyzing the relation of repeats with GC content, it is necessary to avoid the subjectivity implicit in choosing a window size to compute the surrounding GC for repeats, which would lead to unpredictable results (Bernaola-Galván et al. 1996; Li 2001; Oliver et al. 2001, 2002). Therefore, we first located isochores, thus obtaining GC values truly representative of the genomic environment in which the repeats are inserted.

Chromosome sequences were partitioned into fairly homogeneous genome regions (isochores) by using IsoFinder (Oliver et al. 2004; <http://bioinfo2.ugr.es/IsoF/isofinder.html>), an improved version of the segmentation algorithm described earlier (Bernaola-Galván et al. 1996; Oliver et al. 2001, 2002). Briefly, we move a sliding pointer from left to right along the DNA sequence. At each position of the pointer, we compute the mean G + C values to the left and to the right of the pointer. We then determine the position of the pointer for which the difference between left and right mean values (as measured by the t statistic) reaches its maximum. Next, we determine the statistical significance of this potential cutting point, after filtering out short-scale heterogeneities below 3 kb by applying a coarse-graining technique. Finally, the program checks whether this significance exceeds a probability threshold. If so, the sequence is cut at this point into two subsequences; otherwise, the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. This leads to partitioning of a DNA sequence into long homogeneous genome regions (LHGRs) with a well-defined mean GC level, each significantly different (at the 95% confidence) from the mean GC level of the adjacent regions. LHGRs may be assimilated into Bernardi's isochores (Oliver et al. 2001, 2002). The coordinates, sizes, and GC contents for all isochores identified in each human chromosome are available at our Web site (Online Resource on Isochore Mapping: <http://bioinfo2.ugr.es/isochores>) and also at the UCSC Genome Browser (<http://genome.cse.ucsc.edu/>).

We classified isochores into discrete compositional classes (Table 1) using the isochore abundances reported by Bernardi's group (Bernardi et al. 1985; Bernardi 2000). However, the classification given here differs slightly from that given by Bernardi, as we introduced a new human isochore (H4) by splitting the old H3 isochore into two new ones. The new H3 isochore corresponds to two-thirds of the old H3, while the new H4 is made up of the remaining one-third richest in GC of the old H3 (1.6% of the total DNA). We have chosen this classification because Alu density markedly decreases in this newly introduced H4 isochore, whereas gene density reaches its maximum. In this way, Alu and gene densities are positively correlated in GC-poor isochores but negatively correlated in the GC-richest isochore.

Scanning the Human Genome Sequence for Alus and LINE1s

To mask the repeats in the genome and align them with their respective family consensus, we used the RepeatMasker algorithm by Arian Smit (A.F.A. Smit and P. Green, unpublished data: <http://repeatmasker.genome.washington.edu>) and Rebase Update (Jurka 2000). We obtained the alignments between each element found in the genome (locus) and their family consensus by using the `-a` option in RepeatMasker. The proportion of nucleotide sites at which the two aligned sequences differ (p) was then transformed into evolutionary distance using the method of Tamura and Nei (1993). This method takes into account both the transition/transversion and the GC-content biases (Nei and Kumar 2000). Although we masked the whole genome, only repeats located in isochores longer than 50 kb were used here (see Table 1 for iso-

chore statistics). The total amount of DNA in the present analysis was thus limited to 2.58 Gb.

The repeats are often found fragmented in the genome. Based on the identification number (ID) provided by RepeatMasker, we joined together those fragments. As this ID number is assigned also to repeats that have potentially the same origin, we take into account the consensus coordinates to assemble only *true* fragments, and we used exclusively those *reassembled* repeats.

A total of 931,260 ($\approx 85\%$) Alus and 488,633 ($\approx 70\%$) LINE1 retrotransposons in 5736 human isochores longer than 50 kb were analyzed. Table 1 shows the number, average lengths, GC range, and mean %GC of the different isochore classes analyzed in this study. The numbers, lengths, and mean %GC contents of the Alus and LINE1s harbored by different isochore classes are also shown.

Perl Scripts

Perl scripts were developed to parse RepeatMasker output and compute the repeat GC content and evolutionary distance to the family consensus, as well as to perform all the subsequent data analyses. These scripts are available from the authors upon request.

Simulation of the Alu Insertion Process

To check the observed recombination products against the ones found in a random insertion process, we carried out a simulation of Alu insertion in the following way. (1) We used RepeatMasker to generate an exhaustive list of the Alus present in a particular genomic sequence we wished to simulate. (2) The Alus were ranked by age (evolutionary distance). (3) We first inserted the oldest Alu of the ranked list into a random position of a clean (Alu-free) random sequence, (4) We iterated step 3 for all the remaining Alus in the ranked list. By inserting the Alus in this ordered way, we generated the best approximation to the true insertion history. We also took into consideration the possibility of insertions into pre-existing Alus and the “blow-up” of the sequence produced by the Alu insertions. In this way, at each step of the simulation process the sequence generated increases its length, finally reaching the same length as the original genome sequence we wish to simulate. In computing the expected values of Alu trimers for a given genome sequence, we generated a hundred simulated sequences and then averaged the observed numbers of Alu trimers in each sequence.

Results and Discussion

The Density Shift: Alus Are Now Preferentially Located in H Isochores

To analyze the evolution of retrotransposon densities in the genome, we computed the *density ratios*—i.e., the densities in the H isochores divided by the densities in the L isochores (H^*/L , where L is the weighted sum of L1 + L2 isochores). This means, for example, that if the ratios are lower than 1, the absolute-density maximum is located in the L isochores.

Figure 1 shows the Alu and LINE1 density ratios as a function of age, estimated by evolutionary distance (see Data and Methods). Below evolutionary distances of around 0.025, the absolute Alu density maximum is located in the L isochores, given that all

the ratios are smaller than one. However, the density maximum shifts very fast toward the H isochores, which from this distance on register the maximum for Alus.

Note that the scale in Fig. 1 is not linear because CpG dinucleotides were included in the computations. When the fast CpG mutations were excluded, the density shift could not be resolved so clearly. We refer to this observable density shift at 0.025 as the “last” density shift, as it can be assumed that Alus during their evolution continually shifted their density.

Figure 1 (bottom) shows that the insertion pattern of the LINE1s is similar to that of Alus, but the AT-rich preference of LINE1 elements is more pronounced. This same observation was previously made by Pavliček et al. (2001), who offered two possible explanations: either the LINE1 elements are excluded from the GC-rich isochores or the Alus are removed from the AT-rich isochores. In this way, all happens as if some mechanism, acting after the insertion of these retrotransposons into the genome, were able to remove or exclude the Alus more effectively from the L than from the H isochores. What, then, is the nature of this mechanism(s)?

Factors Mediating the Initial Insertion Process: Alu/LINE1 Competition for Retrotransposase

The first mechanism that may change the Alu genomic distribution is competition for retrotransposase mediating the initial insertion process. This enzyme is encoded only by LINE1 but is used by both LINE1 and Alu repeats to retrotranscribe and reinsert into the genome. Gu et al. (2000) proposed that, since Alus use the retrotransposase encoded by LINE1, and since LINE1 prefers to insert into AT-rich regions, Alus could avoid competition for enzyme with LINE1 elements if they insert into GC-rich regions, which are replicated at different times from AT regions in the cell cycle. In this way, Alus could have shifted their density maximum to H isochores.

To test this hypothesis, we plotted the absolute densities of Alus and LINE1s as a function of isochore membership and evolutionary distance (Fig. 2). The LINE1s show a pronounced density minimum in the L1 and L2 isochores between evolutionary distances of 0.04 and 0.15 (Fig. 2, bottom), which appears concurrently with the maximum spread of the Alus (Fig. 2, top). Therefore, although the competition between the two elements is evident, the Alu insertion into L isochores does not appear to have been hampered. Instead, the massive Alu radiation has apparently somewhat impeded the LINE1 propagation, particularly in these isochores. Therefore, it appears that Alu/LINE1 competition for retrotrans-

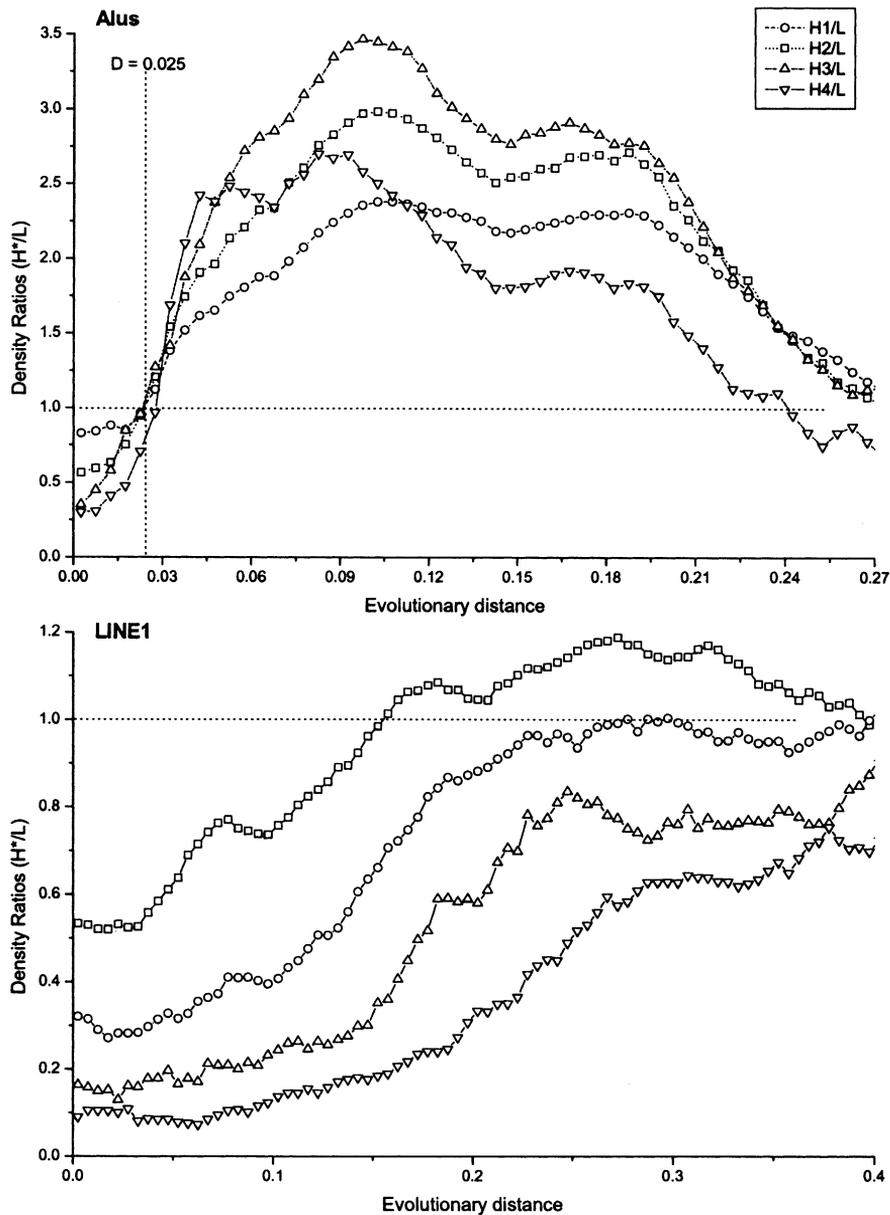


Fig. 1. Alu (top) and LINE1 (bottom) density ratios (H^*/L) as a function of age, where L is the weighted density sum of L1 and L2 isochores. The CpG mutations are included to compute the genetic distance, and therefore the time scale (abscissa) is not linear.

posase does not hinder Alu insertion and spread in the L isochores.

In their paper, Gu et al. (2000) assume that the competition between LINE1 and Alu RNAs for LINE1 ORF2 proteins takes place in the nucleus. This assumption is probably wrong. Current evidence (e.g., the LINE1 cis-preference) suggests that the interaction takes place on ribosomes (see, e.g., Boeke 1997), and consequently no competition occurs near the nuclear DNA.

The Density Shift of the Alus Occurred Long Before They Began to Match the Host Isochore Composition

Being GC rich, Alus have a higher compositional match in the H isochores, while the GC-poor LINE1s

match better to the L isochores. In general, most elements tend to fit their density maximum in regions where they compositionally match (Filipski et al. 1989). If so, compositional matching might be responsible for the density shift of the Alus.

To test this hypothesis, we analyzed the differences in GC content between the Alus and their family consensus as a function of age and isochore membership (Fig. 3). The GC-content decay of Alus is clearly appreciated in this plot. Up to evolutionary distances between 0.008 and 0.1, the slope is steeper, which corresponds to a higher rate of CpG mutations. Alus are rich in CpG dinucleotides that tend to mutate faster than the remaining sites. We call this range of evolutionary distances “the CpG domain.” During the CpG domain, no influence of the isochore on the composition of the Alus can be appreciated,

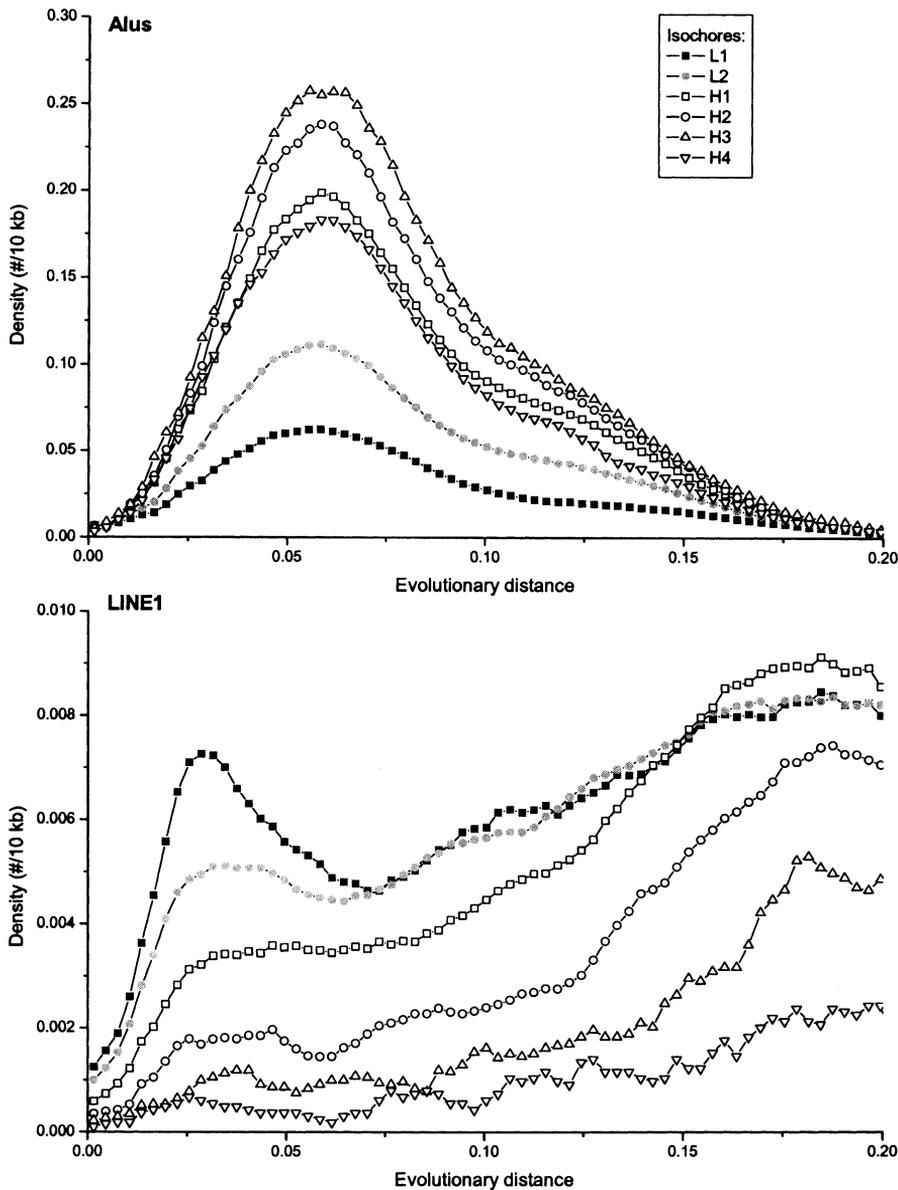


Fig. 2. Alu (**top**) and LINE1 (**bottom**) absolute densities as a function of isochore membership and evolutionary age. The LINE1 show a very pronounced density minimum in the L isochores, which happens concurrently with the massive spread of the Alus.

since the decay in Alu GC is dominated at this stage by fast mutations at the CpGs. After the CpG domain, the Alu GC decay is stronger in L than in H isochores, thus revealing a higher compositional matching pressure on Alus in L isochores.

By plotting the correlation between the isochore and the Alu GC content in different “age bins” (Fig. 4), the two stages in Alu evolution—the CpG domain and the compositional matching to the isochore—can be more clearly distinguished. During the CpG domain, the correlation remains low. Afterward, the correlation begins to increase and the effects of the compositional matching become appreciable. The constant increase in r^2 during this second stage indicates a time dependence of Alu compositional matching, a result that agrees with the regional mutation-pressure hypothesis (Wolfe et

al. 1989), which postulates that the mutation rate varies with chromosome regions. CpG levels of Alus have been previously shown to correlate with the GC levels of the long sequences in which they are located (Jabbari and Bernardi 1998). Our results also agree with previous observations that Alus accumulate point mutations at the rate expected for unselected DNA sequences, except for the special case of CpG dinucleotides (Jurka 1995; Schmid 1998).

As mentioned above, the density shift already takes place at evolutionary distances of 0.025. This signifies that long before Alus are influenced by the isochore on their base composition, Alu density shifts toward the H isochores. This observation rules out the matching mutational pressure as the agent of the density shift.

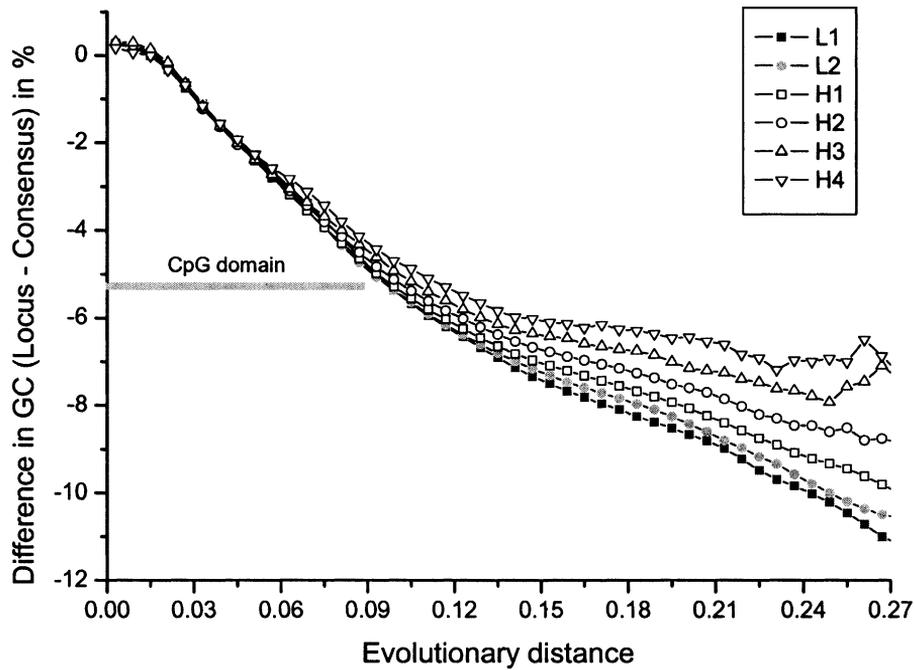


Fig. 3. Differences in GC content between the Alus and their family consensus. Two different slopes can be appreciated, corresponding to the CpG domain and the subsequent compositional matching of the Alus to the isochore.

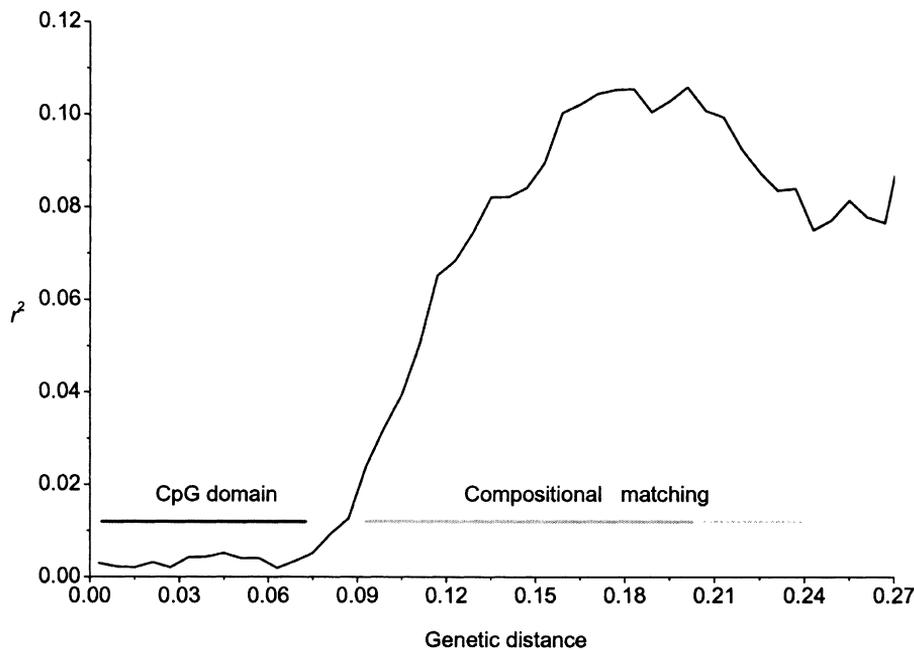


Fig. 4. Correlation between the isochore and the Alu GC content in different "age bins." The r^2 coefficient is plotted against the genetic distance.

Alu Overrepresentation in Introns Is Limited to L Isochores

Like Alus, the genes are also denser in GC-rich isochores (Zoubak et al. 1996; Bernardi 2000). By assuming a causal effect in this association, another hypothesis to explain the Alu density shift can be formulated: Alus would have followed the genes by concentrating in the GC-rich isochores, either through selective forces similar to those acting on genes (IHGSC 2001) or, more probably, through a hitchhiking effect. In fact, Alus are overrepresented in

introns and scarcer in intergenic sequences (Smit 1999). The higher gene (intron) density in H isochores could then explain the higher Alu density in GC-rich isochores.

To check this hypothesis we need to compare intron vs. intergenic *Densities* in the different isochores. The Alu excess in introns was measured as $R_{Ex} = [\rho_{IV}/(\rho_{IV} + \rho_{ig})] - 0.5$, where ρ_{IV} and ρ_{ig} are intron and intergenic Alu densities, respectively. The coefficient R_{Ex} is negative if intergenic densities are higher than intron densities and positive in the opposite case.

Table 2. Alu densities in introns and intergenic regions of the entire genome and the different isochores

	Introns		Intergenic regions		Intron excess
	<i>N</i>	Density	<i>N</i>	Density	
Entire genome	550,908	4.376	431968	3.12	0.084
Isochore					
L1	30,400	2.088	66164	1.752	0.044
L2	97,372	3.405	87347	2.917	0.039
H1	141,086	5.168	101627	5.897	-0.033
H2	57,537	6.257	32569	7.561	-0.047
H3	28,055	7.389	13568	8.977	-0.049
H4	8,641	5.289	4190	7.328	-0.081

Note. Intergenic regions were taken as sequence segments beginning 2 kb downward (3') from the stop codon of a gene and ending 2 kb upward (5') from the initiation codon of the next gene.

Table 2 shows the Alu excess in introns we found for the entire genome and for the different isochores. When the entire genome is considered, Alus are in fact overrepresented in introns, thus confirming the observation by Smit (1999). However, when R_{Ex} is computed per isochore, we got positive values in L, but negative values in H, isochores. Thus, Alu overrepresentation is limited to the introns of L isochores. Hence, the density shift to H isochores cannot be explained by the Alu concentration in the introns of these regions. When F5' and F3' flanking regions of genes, instead of introns, were used in the comparison with intergenic regions, the same conclusion was reached (not shown). The observation that in genes of GC-rich isochores the Alus are underrepresented compared to intergenic Alu copies rules out the hypothesis by IHGSC (2001) that Alu accumulate in genes and thus in GC-rich regions because they have some benefit for gene expression. Brookfield (2001) already suggested that the hypothesis of IHGSC (2001) is inconsistent with our knowledge of human population genetics.

Measuring Alu–Alu Recombination Activity in the Genome

Alu sequences can serve as substrates for either homologous or nonhomologous recombination events (Babcock et al. 2003; Deininger et al. 2003). The Alu repeat occurs approximately once every 3 kb in the human genome, and mispairing between such repeats, and a subsequent unequal crossing-over event between the mispaired repeats, may be a frequent cause of deletions and duplications (Fig. 5). In fact, Alu–Alu recombination has been proposed to be a possible mechanism driving the spatial, postinsertional Alu dynamics in the human genome (see, e.g., Lobachev et al. 2000; Brookfield 2001; Stenger 2001; Batzer and Deininger 2002; Medstrand et al. 2002; Deininger et al. 2003). However, an additional requisite for Alu–Alu recombination to provoke a net

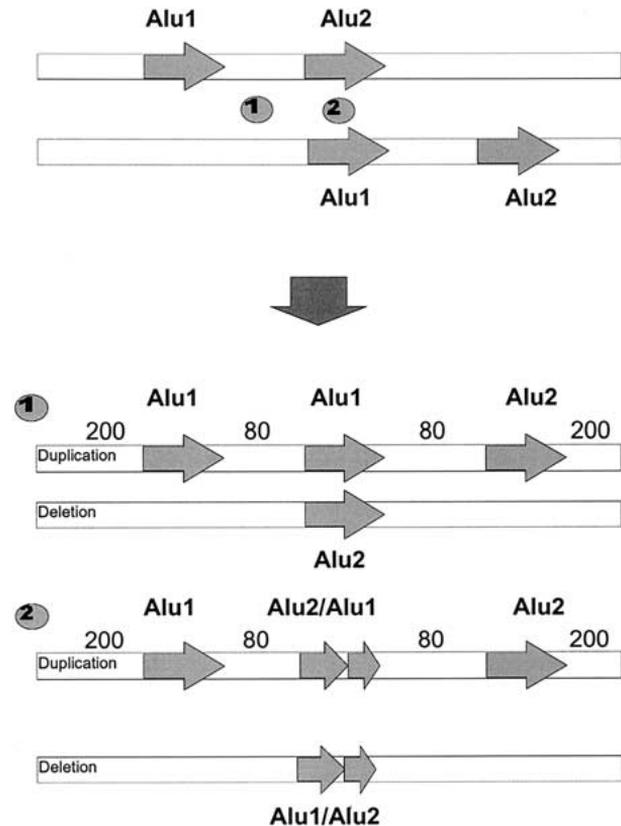


Fig. 5. Duplication products coming from the unequal homologous recombination between two directly oriented Alus. Two classes of Alu duplications are produced, depending on the specific site where the crossing-over occurs (1 or 2). Inner distances of 80 bp between the Alus involved and distances of 200 bp to the flanking Alus were used to scan the genome in search of these recombination products.

effect on Alu densities is that recombination products (deletion/duplication) must have a differential chance to survive in the distinct genomic regions.

Direct measurement of Alu–Alu recombination products in the genome sequence is difficult, as deletions usually do not leave identifiable footprints. Therefore, we estimated unequal recombination activity by counting only the duplication products. In

Table 3. Frequencies of Alu trimers (derived from Alu/Alu recombination) in the different isochore classes

Isochore	No. observed	No. expected	Obs/exp ratio
L1	124	1.84	67.39
L2	426	10.43	40.84
H1	804	33.34	24.12
H2	398	15.47	25.73
H3	217	8.40	25.83
H4	64	1.43	44.76
Total	2033	70.91	28.67

Note. The expected frequencies were derived from simulation experiments (see text). The comparisons between observed and expected frequencies were all statistically significant (chi-square test, $p < 0.001$).

the simplest case, if the crossing-over takes place between the two Alus, the duplication consists of three elements arranged in tandem (trimer), while if the recombination event occurs within the Alus, the product will be a tandem of four Alus (tetramer). We focus here on the frequencies of Alu trimers, as the tetramer numbers we found in the different isochores were too small (and noisy) to allow statistical tests (see below).

To track Alu duplications in the genome, we designed several Perl scripts to parse the RepeatMasker output. To locate these recombination products, we searched for three Alus in tandem separated by less of 80 bp and with no other Alu closer than 200 bp. Additionally, we imposed the condition that two consecutive Alus in a trimer must be of the same subfamily, due to the duplication. To filter out trimers originated by insertion of Alus into other pre-existing Alus, we used the ID number provided by RepeatMasker, afterward discarding trimers in which the two flanking Alus were actually the same element. In this way, we excluded most “nonrecombinogenic” trimers that did not originate through unequal Alu–Alu recombination.

However, another alternative source for Alu trimers is simply chance formation, in the “normal” Alu-insertion process over evolution. To control for this possibility, we performed computer simulations by randomly inserting Alus into artificial sequences (see Data and Methods). After generating 100 simulated sequences for each isochore class, we scanned them for Alu trimers, using the same setup as for the genome scan (see above). The frequencies of Alu trimers were then averaged for the 100 simulation runs for each isochore. We consider these values to be the expected chance frequencies.

Table 3 shows that the observed frequencies of Alu trimers in the genome were far higher than those expected to arise merely by chance. Furthermore, the observed/expected ratios were higher in L isochores,

thus pointing to a higher rate of unequal recombination, or a higher survival rate of recombination products, in these isochores. This suggests that the other recombination product (Alu deletion, which we cannot detect directly) should also be more frequent in L isochores.

The total number of trimers detected (2033) may appear too low to explain the dramatic shift in the Alu distribution. It should be taken into account, however, that we are detecting probably only one part of the recombination events. There may be a number of reasons for this underestimation. First, we imposed restrictive conditions on our genome scan (see above) to exclude most of the nonrecombinogenic Alu trimers. Second, the products of successive recombination events involving trimers cannot be detected with the setup we are using. Lastly, other recombination mechanisms, not quantified in this paper, may be also operating, in particular, those nonconservative ones preferentially producing deletions. For example, the intrachromatid single-strand annealing mechanism produces one genomic deletion and one episomal DNA that is lost before or during the next cell division.

Even so, the relatively higher number of recombination products we detected in L isochores is not enough to change the Alu densities in human isochores. It is necessary, furthermore, for each recombination product to have a differential survival probability. It seems that this would be the case. The analysis of human genetic disorders caused by Alu–Alu recombination has shown that deletions are by far more abundant than duplications, despite that even in-frame duplications/deletions can cause disease (Deininger and Batzer 1999; Kolomietz et al. 2002). Along the same line, experimental evidence for recombination in mammalian cells strengthens the idea of the increased frequency of deletions compared to duplications (Lambert et al. 1999). If this were the general rule, the higher level of unequal Alu–Alu recombination, together with the higher selective tolerance toward deletions in L isochores, could have promoted the preferential removal of Alus from GC-poor isochores and, therefore, their relative increase in the GC-rich genome regions (density shift).

If deletions by unequal recombination are more frequent in AT-rich DNA, all other repeats should also (though slowly) accumulate in GC-rich DNA. However, the opposite is true; with the exception of Alu and MIR SINES, all other repeats are more dense in AT-rich DNA. Nevertheless, when the density ratios were plotted against the evolutionary age of LINE1 elements (Fig. 1b), an accumulation trend of the older LINE1 elements in the GC-rich genome regions can be appreciated. A comparison between young and old LINE1s in Pavliček (2001) also suggested this trend.

Table 4. Average genetic distances (\pm SD), determined with the Tamura–Nei method, for the AluYa5, AluYb8, and AluY subfamilies and the AluJ subgroup in the different isochores

Isochore	AluYa5	AluYb8	AluY	AluJ
L1	0.010 \pm 0.012	0.016 \pm 0.023	0.039 \pm 0.023	0.143 \pm 0.042
L2	0.010 \pm 0.014	0.019 \pm 0.032	0.038 \pm 0.025	0.133 \pm 0.040
H1	0.015 \pm 0.030	0.029 \pm 0.051	0.038 \pm 0.023	0.129 \pm 0.042
H2	0.018 \pm 0.034	0.031 \pm 0.046	0.038 \pm 0.026	0.127 \pm 0.046
H3	0.015 \pm 0.015	0.041 \pm 0.049	0.039 \pm 0.032	0.129 \pm 0.053
H4	0.019 \pm 0.033	0.032 \pm 0.032	0.039 \pm 0.032	0.129 \pm 0.055

A Role for Alu–Alu Recombination: Two Independent Pieces of Evidence

The frequency of unequal homologous recombination between repeated Alu elements depends on both sequence similarity and physical distance on the chromosome between the two repeats involved. High sequence similarity and/or short physical distance should enhance recombination rates (Lobachev et al. 2000; Medstrand et al. 2002). On the contrary, a low level of sequence similarity or a large physical distance should lower the recombination rate. We took advantage of these two relationships and collected two sets of observations that support the implication of unequal recombination in the preferential Alu removal from L isochores. Below, these observations are described in detail.

Recently Inserted Alu Subfamilies Are Younger in L Than in H Isochores. New Alus are preferentially inserted in L isochores. If, likewise, the Alus were more frequently removed from these regions, a high turnover rate of Alu elements should be expected in these isochores. New inserted Alus are probably mutation-free copies of the master genes (Deininger et al. 1992). Therefore, due to the constant inflow of new, mutation-free Alus and the removal of other Alus, many of which have probably accumulated mutations, the compositional evolution of Alus in L isochores should become scaled down compared to those in H isochores. If so, we can formulate a first testable prediction: Alus should appear “younger” in L than in H isochores.

In addition, if the mechanism for Alu removal were Alu–Alu recombination, the rate of which directly depends on sequence similarity, we can add that the rejuvenation effect in L isochores should be stronger for members of the recently inserted subfamilies (more similar) than for the older ones (less similar). Summarizing both premises, we can formulate our hypothesis in the following way: Members of recently inserted Alu subfamilies will appear “younger” in L than in H isochores. This hypothesis may at first seem counterintuitive, since, given the Alu GC-richness, a faster rate of molecular evolution (compositional matching), and therefore faster aging, can

Table 5. Correlation between genetic distance (Tamura–Nei method) and isochore GC content in the AluYa5, AluYb8, and AluY subfamilies and the AluJ subgroup

	<i>N</i>	<i>r</i>	<i>p</i>	Slope
AluYa5	1836	0.14	$<10^{-6}$	35.32
AluYb8	1648	0.21	$<10^{-6}$	27.81
AluY	5610	−0.01	0.31	−7.46
AluJ	5689	−0.22	$<10^{-6}$	−64.29

be expected for Alus in L isochores, compared to those located on H isochores.

To discriminate between these two opposing effects, we computed the average evolutionary distances in Alu subfamilies of different ages (Table 4). We also plotted the evolutionary distances for each Alu against the isochore GC content, then computed the regression coefficients (Table 5). We observed a positive correlation between evolutionary distance and isochore GC content in young, recently inserted Alu subfamilies (AluYa5 and AluYb8, inserted 3–4 MYA [see Mighell et al. 1997]), but no correlation (AluY, inserted 20 MYA) or a negative correlation (AluJb, inserted 50–80 MYA) was found in the older ones. In this way, as predicted by our hypothesis, the rejuvenation effect in L isochores is limited to only recently inserted Alu subfamilies.

This age dependence of the rejuvenation process might point to recombination as the mechanism provoking the Alu removal from L isochores. It is known that the efficiency of recombination is directly related to the length of uninterrupted regions of nucleotide identity, with higher rates of recombination being associated with longer stretches of nucleotide identity (Waldman and Liskay 1988; Baker et al. 1996). In this way, Alu–Alu recombination should depend on the pairwise similarity between the sequences involved, with older Alu elements being less similar and less prone to recombine than younger, more similar Alu insertions (Batzer and Deininger 2002). Therefore, unlike any other evolutionary mechanism, recombination should act especially upon young, more similar elements, and its effects should decline sharply with growing sequence divergence or age of the elements.

In this way, the opposite correlations we observed between evolutionary distance and isochore GC content for Alu members of different ages (Tables 3 and 4) can be readily explained through the action of recombination. In recently inserted Alu subfamilies, such as AluYa5 or AluYb8, the average identity between repeats, and therefore the rate of Alu–Alu recombination, should be high, and a strong turnover of repeats should be expected. The consequence is that these Alus should appear younger in L than in H isochores, and a positive correlation between evolutionary distance and isochore GC content is then expected. Conversely, in the older Alu subfamilies, as AluJb, the average identity between repeats, and therefore the rate of Alu–Alu recombination should be low, and no turnover should take place. Therefore, Alu evolution would be driven only by compositional matching pressure, and a negative correlation between evolutionary distance and isochore GC content should be expected. In subfamilies of intermediate age, such as AluY, the lack of correlation we observed may be due to equilibrium between a moderate turnover rate and the increasing pressure of compositional matching.

In summary, the rejuvenation we found in recently inserted Alu subfamilies from L isochores may be due to the preferential Alu removal by Alu–Alu recombination. Therefore, unequal recombination could have contributed significantly to the Alu density shift in human isochores.

Old Stand-Alone Alus are More Frequent in L Isochores. Besides sequence similarity, unequal recombination requires a physical neighborhood on the chromosome between the two repeats involved. Closely spaced elements are more prone to provoke unequal recombination than the more distant ones. It is known that Alu elements are not uniformly distributed and tend to cluster with each other (Pavliček et al. 2001; Jurka et al. 2002, 2004). However, other retroposed elements exist outside the clusters. By measuring the densities of these “stand-alone” Alus, we found significant variations among different isochores, which may again suggest a role for unequal recombination in the preferential removal of Alus from L isochores.

For each Alu, we define a “nearest-neighbor distance” (NND) as the length (bp) separating this particular Alu element from the nearest one (in either the 5′ or the 3′ direction in the two DNA strands). We then considered “stand-alone” Alus as those showing NND > 2 kb (from this distance on, the Alus show a density maximum in L isochores). We found a total of 127,511 stand-alone Alus. Their densities in the different isochores are shown in Fig. 6. Stand-alone Alus are clearly abundant in L isochores and scarce in H isochores, a result consistent with the higher clus-

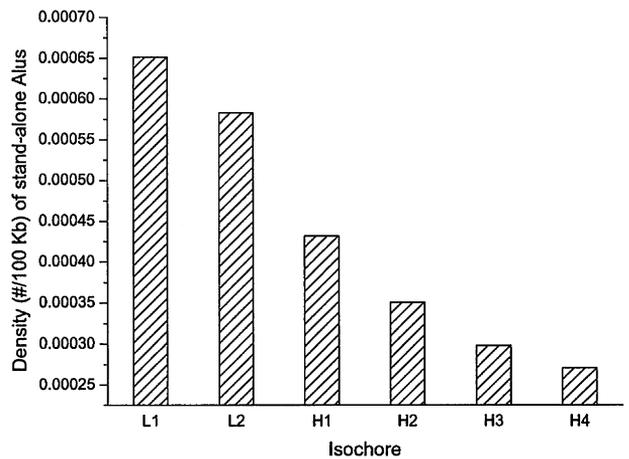


Fig. 6. Densities (number/100 kb) of stand-alone Alus (NND > 2 kb) in the different isochores. Most (81%) of these stand-alone Alu copies are from older AluS and AluJ families.

tering found by Jurka et al. (2004) in GC-rich regions.

Jurka et al. (2004) observed, furthermore, that recently retroposed elements are likely to be inserted outside the existing clusters, which should lead to an overrepresentation of young members among the stand-alone class. We confirmed such overrepresentation, as 19% of stand-alone Alus are members of the youngest group AluY (this group represents 10% of all Alu copies in the genome). However, the proposal of these authors that stand-alone elements appear to be rapidly eliminated from the genome was not supported by our genome scan. We observed that 81% of stand-alone Alu copies are from older AluS and AluJ families, thus indicating that many old elements are able to survive outside the clusters.

We interpret the high frequencies of stand-alone Alus in L isochores as a consequence of the preferential elimination of closely spaced Alus by unequal recombination. Only the Alus that are neighbors on the chromosome seem to take part in unequal recombination events, the isolated, stand-alone ones remaining untouched for long periods in the genome. In this way, the higher frequencies of old stand-alone Alus we detected in L isochores may reflect the higher activity of Alu–Alu recombination eliminating closely spaced Alus from these GC-poor isochores, thereby contributing to the density shift of Alus toward GC-rich isochores.

Conclusions

The genome-wide analysis of Alu densities we performed allowed us to rule out Alu/LINE1 competition for retrotransposase mediating the initial insertion process, compositional matching pressure,

and Alu overrepresentation in introns as mechanisms driving the Alu density shift in the human genome. However, the higher level of unequal recombination we detected in L isochores points to Alu–Alu recombination, coupled with the differential selective disadvantage of their recombination products, as the main agent provoking the Alu density shift in human isochores. Two independent pieces of evidence—the lower evolutionary divergence shown by recently inserted Alu subfamilies and the higher frequency of old stand-alone Alus in L isochores—support such a conclusion.

However, other recombination mechanisms as well as other evolutionary factors are probably involved. In particular, the contractions in population size or bottlenecks during primate speciation events (Hedges et al. 2004) may have been an important helper mechanism for Alu expansions. For example, the evolutionary distance of 0.025 at which the last density shift occurred (see Fig. 1) corresponds to 6.4 MY, which is also the date of the putative bottleneck during the human/chimpanzee split. It may be, therefore, that the Alu expansion around the speciation event was accelerated by such a population bottleneck, thus contributing to the fast accumulation of Alus in GC-rich isochores.

Acknowledgments. Helpful comments from A. Marín, J.P. Martínez-Camacho, M. Ruiz-Rejón, and two anonymous reviewers are greatly appreciated. We are also grateful to A. Smit for providing the RepeatMasker computer program. This work was supported by the Spanish Government (Grants BIO2002-04014-C03-01/02 to J.L.O. and P.B. and BFM2002-00183 to P.C. and P.B.) and Plan Andaluz de Investigación (CVI-162). M.H. acknowledges a predoctoral grant from the University of Granada (Spain). The help of David Nesbitt and Christopher Previti with the English version of the manuscript is also appreciated.

References

- Babcock M, Pavliček A, Spiteri E, Kashork CD, Isohikhes I, Shaffer LG, Jurka J, Morrow BE (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by *Alu*-mediated recombination events during evolution. *Genome Res* 13:2519–2532
- Baker MD, Read LR, Beatty BG, NG P (1996) Requirements for ectopic homologous recombination in mammalian somatic Cells. *Mol Cell Biol* 16:7122–7132
- Batzler MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3:1–10
- Bernaola-Galván P, Román-Roldán R, Oliver JL (1996) Compositional segmentation and long-range fractal correlation in DNA sequences. *Phys Rev E* 53:5181–5189
- Bernardi G, Olofson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958
- Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3–17
- Bernardi G (2001) Misunderstandings about isochores. Part 1. *Gene* 276:3–13
- Boeke JD (1997) LINEs and Alus—The polyA connection. *Nat Genet* 16:6–7
- Britten RJ (1996) DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci USA* 93:9374–9377
- Brookfield JF (2001) Selection on Alu sequences? *Curr Biol* 11:900–901
- Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW (1998) Potential Alu function: Regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol Cell Biol* 18:58–68
- Deininger PL, Batzer MA (1999) Alu repeats and human disease. *Mol Genet Metab* 67:183–193
- Deininger PL, Batzer MA, Hutchison CA, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet* 8:307–311
- Deininger PL, Moran TV, Batzer MA, Kazazian HH Jr (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
- Filipski J, Salinas J, Rodier F (1989) Chromosome localization-dependent compositional bias of point mutations in Alu repetitive sequences. *J Mol Biol* 206:563–566
- Gu Z, Wang H, Nekrutenko A, Li WL (2000) Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* 259:81–88
- Hedges DJ, Callinan PA, Cordaux R, Xing J, Barnes E, Batzer MA (2004) Differential Alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* 14:1068–1075
- IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Jabbari K, Bernardi G (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* 224:123–127
- Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. *J Mol Evol* 24:39–44
- Jurka J (1995) Origin and evolution of Alu repetitive elements. In: Maraia RJ (ed) *Impact of short interspersed elements (SINES) on the host genome*. Landes, Austin, TX, pp 25–41
- Jurka J (2000) Repbase Update, a database and an electronic journal of repetitive elements. *Trends Genet* 16:418–419
- Jurka J, Krnjajic M, Kapitonov VV, Stenger JE, Kohkany O (2002) Active Alu elements are passed primarily through paternal germlines. *Theor Popul Biol* 61:519–530
- Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV (2004) Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci USA* 101:1268–1272
- Kolomietz E, Meyn MS, Pandita A, Squire JA (2002) The role of *Alu* repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes Cancer* 35:97–112
- Lambert S, Saintigny Y, Delacote F, Amiot F, Chaput B, Lecomte M, Huck S, Bertrand P, Lopez BS (1999) Analysis of intrachromosomal homologous recombination in mammalian cell, using tandem repeat sequences. *Mutat Res* 433:159–168
- Lev-Maor G, Sorek R, Shomron N, Ast G (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* 300:1288–1291
- Li W (2001) Delineating relative homogeneous G + C domains in DNA sequences. *Gene* 276:57–72
- Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA (2000) Related inverted Alu repeats unstable in

- yeast are excluded from the human genome. *EMBO J* 19:3822–3833
- Martínez Zapater JM, Marín A, Oliver JL (1993) Evolution of base composition in T-DNA genes from *Agrobacterium*. *Mol Biol Evol* 10:437–448
- Medstrand P, van de Lagemaat LN, Mager DL (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* 12:1483–1495
- Mighell AJ, Markham AF, Robinson PA (1997) Alu sequences. *FEES Lett* 417:1–5
- Nei M, Kumar S (2000) *Molecular evolution and phylogenetics*. Oxford University Press, Oxford
- Oliver JL, Marín A, Martínez Zapater JM (1990) Chloroplast genes transferred to the nuclear plant genome have adjusted to nuclear base composition and codon usage. *Nucleic Acids Res* 18:65–73
- Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R (2001) Isochore chromosome maps of eukaryotic genomes. *Gene* 276:47–56
- Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, Hackenberg M, Bernaola-Galván P (2002) Isochore chromosome maps of the human genome. *Gene* 300:117–127
- Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P (2004) IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res* 32(Web Server Issue):W287–W292
- Paces J, Zika R, Paces V, Pavlíček A, Clay O, Bernardi G (2004) Representing GC variation along eukaryotic chromosomes. *Gene* 333:135–141
- Pavlíček A, Jabbari K, Paces J, Paces V, Henjar J, Bernardi G (2001) Similar integration but different stability of Alus and LINES in the human genome. *Gene* 276:39–45
- Pavlíček A, Clay O, Bernardi G (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? *FEES Lett* 523:252–253
- Rynditch A, Zoubak S, Tsyba L, Tryapitsina-Guley N, Bernardi G (1998) The regional integration of retroviral sequences into the mosaic genomes of mammals. *Gene* 222:1–16
- Schmid CW (1998) Does SINE evolution preclude Alu function? *Nucleic Acids Res* 26:4541–4550
- Smit AFA (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 9:657–663
- Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. *Genome Res* 12:1060–1067
- Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA (2001) Biased distribution of inverted and direct Alus in the human genome: Implications for insertion, exclusion, and genome stability. *Genome Res* 11:12–27
- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526
- Waldman AS, Liskay RM (1988) Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Mol Cell Biol* 8:5350–5357
- Wolfe K, Sharp P, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285
- Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. *Gene* 174:95–102