ELSEVIER

Gene 276 (2001) 39–45

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Similar integration but different stability of Alus and LINEs in the human genome

Adam Pavlíček[a,b], Kamel Jabbari[b], Jan Pačes[a,c], Václav Pačes[a,c], Jiří Hejnar[a], Giorgio Bernardi[b,d,*]

[a]*Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, Prague, CZ-16637, Czech Republic*
[b]*Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France*
[c]*Center for Integrated Genomics, Flemingovo 2, Prague, CZ-16637, Czech Republic*
[d]*Laboratorio di Evoluzione Molecolare, Stazione Zoologica, Villa Comunale, 80121 Naples, Italy*

Received 11 April 2001; received in revised form 7 May 2001; accepted 26 July 2001
Received by C.W. Schmid

Abstract

Alus and LINEs (LINE1) are widespread classes of repeats that are very unevenly distributed in the human genome. The majority of GC-poor LINEs reside in the GC-poor isochores whereas GC-rich Alus are mostly present in GC-rich isochores. The discovery that LINES and Alus share similar target site duplication and a common AT-rich insertion site specificity raised the question as to why these two families of repeats show such a different distribution in the genome. This problem was investigated here by studying the isochore distributions of subfamilies of LINES and Alus characterized by different degrees of divergence from the consensus sequences, and of Alus, LINEs and pseudogenes located on chromosomes 21 and 22. Young Alus are more frequent in the GC-poor part of the genome than old Alus. This suggests that the gradual accumulation of Alus in GC-rich isochores has occurred because of their higher stability in compositionally matching chromosomal regions. Densities of Alus and LINEs increase and decrease, respectively, with increasing GC levels, except for the telomeric regions of the analyzed chromosomes. In addition to LINEs, processed pseudogenes are also more frequent in GC-poor isochores. Finally, the present results on Alu and LINE stability/exclusion predict significant losses of Alu DNA from the GC-poor isochores during evolution, a phenomenon apparently due to negative selection against sequences that differ from the isochore composition.

Keywords: Repeat; Retrotransposon; GC content

1. Introduction

The human genome is a mosaic of long, compositionally homogeneous DNA stretches, the isochores (Macaya et al., 1976), that belong to two GC-poor families, L1 and L2, and three GC-rich families, H1, H2 and H3. These families represent about 30, 33, 24, 7.5 and 4–5% of the human genome, respectively, and are characterized by increasing densities of coding sequences (see Bernardi, 2000, for a recent review). Compositional fractionation of human DNA not only allowed us to characterize the genome in terms of isochore families and gene densities, but also to investigate the distribution of interspersed repeated sequences. Reassociation kinetics provided the first hint that the distribution of intermediate repetitive sequences was different in DNA fractions from different isochore families (Soriano et al., 1981). Hybridization of appropriate probes on compositional DNA fractions showed that the GC-poor LINE1 and the GC-rich Alu families are predominantly located in GC-poor and GC-rich isochores, respectively (Meunier-Rotival et al., 1982; Soriano et al., 1983; Zerial et al., 1986). These results were later confirmed by assessments based on sequences from data banks (Smit, 1996, 1999; Jabbari and Bernardi, 1998).

Alus are short (~300 bp) GC-rich, non-autonomous elements, which derived from 7SL RNA about 80 million years ago; however, most Alu insertions occurred during the past 65 million years (Kapitonov and Jurka, 1996). They make up 10% of the human genome (Smit, 1996). Full-length LINEs are long (6–8 kb) GC-poor sequences encoding an RNA binding protein and a reverse transcriptase/

Abbreviations: Δ, divergence from the family consensus sequence

* Corresponding author. Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121, Naples, Italy. Tel.: +39-081-5833300; fax: +39-081-2455807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/01/$ - see front matter
PII: S0378-1119(01)00645-X

endonuclease. LINE1 elements represent the most abundant group of LINEs and correspond to 15% of the human genome (Smit, 1996).

Reverse transcriptase/endonuclease encoded by LINE1 elements is believed to be involved in Alu transposition because of similar target site duplication, common TT|AAAA insertion site specificity (Feng et al., 1996; Jurka, 1997; Cost and Boeke, 1998), and expression in the male germ line (Schmid, 1998; Branciforte and Martin, 1994). The location of the majority of Alus in GC-rich isochores in spite of their AT-rich insertion sites (which are more frequent in GC-poor isochores), together with suggestions that young Alus are distributed independently of the GC level (Smit, 1999; Arcot et al., 1998), raises a question concerning the integration and stability of Alus and LINEs.

In the present work we found that young Alus and LINEs are more frequent than old Alus and LINEs in the GC-poor isochores. We also found that processed pseudogenes on chromosomes 21 and 22 are more frequent in the GC-poor isochores. Processed pseudogenes are copies of their functional counterparts, characterized by a lack of introns and promoters, acquisition of 3′ poly(A) tails and the presence of target site duplications of varying length (Vanin, 1985; Weiner et al., 1986). Their insertion motifs strongly resemble the TT|AAAA hexanucleotide that is typical for LINE and Alu insertion (Jurka, 1997). The ability of the LINE insertion machinery to form processed pseudogenes was recently demonstrated experimentally in an *in vitro* assay (Esnault et al., 2000).

2. Materials and methods

We extracted 2751 contigs of finished sequences representing 21.70% of the genome (651,023,445 bp) from the Human Genome Sequencing resources at NCBI (http://www.ncbi.nlm.nih.gov/genome/seq/HsHome.shtml; Jang et al., 1999). The analysis of the distribution of Alu and LINE retroposons was performed with RepeatMasker (A.F.A. Smit and P. Green, unpublished data) version 09/20/2000 containing default libraries. Retrotransposon densities were calculated in 100 kb long, non-overlapping fragments. The age of Alus and L1 elements was estimated on the basis of their divergence from the consensus sequence of the corresponding family with a simple correction for multiple substitutions (Jukes and Cantor, 1969).

To compare the GC level of insertion regions, we calculated the mean GC level of 50 kb on both 5′ and 3′ flanks of the insertion site. Means of the GC level around insertion sites were compared by the Mann–Whitney *U*-test. To estimate the age of Alu families, we calculated the average distance, Δ, from the consensus for each family. For the youngest families, AluYa5, AluYb8 and AluY, we calculated the density in the 100 kb long fragments; a similar analysis was also performed for the other Alu families (data not shown). The AluYa8 family was not studied because of the small number, 47, of identified elements. The Alu density around each Alu insertion in the database was analyzed within 20 kb flanking sequences (10 kb on each side). The same was done for LINEs.

We also analyzed Alu and L1 distributions on human chromosomes 22 and 21 (Dunham et al., 1999; Hattori et al., 2000) using a 100 kb sliding window and plotted the locations of integration of the AluYa5 family and the LINE family L1PA2. The data from chromosome 22 (Dunham et al., 1999) were used for the extraction of pseudogenes. The annotations were scanned and all pseudogenes containing intron(s) or having an 'immunoglobulin variable region' in their annotation were excluded, the latter being known to be a result of tandem duplications (Dunham et al., 1999). Using this approach we obtained a dataset containing 114 intronless pseudogenes. For chromosome 21 (Hattori et al., 2000), annotations on the pseudogene intron/exon structure are not available; we scanned, therefore, the annotations for processed pseudogenes and found ten sequences. We then partitioned chromosome sequences into 100 kb long, non-overlapping segments and calculated the GC level and the number of pseudogenes embedded in them.

3. Results and discussion

3.1. Distribution of young Alus

The Alu and LINE distributions in different isochore families and their general genomic distributions in the analyzed dataset are shown in Table 1. As expected from previous results (Soriano et al., 1981, 1983; Meunier-Rotival et al., 1982; Zerial et al., 1986; Smit, 1996, 1999; Jabbari and Bernardi, 1998), Alus are more frequent in GC-rich isochores, whereas LINE1 density decreases with increasing

Table 1
Alu and LINE distribution in the human genome[a]

Δ	L1	L2	H1	H2	H3
All Alus	5.2	8.1	13.8	18.7	17.2
$\Delta < 2\%$	0.077	0.081	0.068	0.046	0.033
$\Delta = 2$–4%	0.062	0.052	0.069	0.073	0.069
$\Delta = 4$–6%	0.2	0.22	0.36	0.47	0.56
$\Delta > 6\%$	4.9	7.7	13.3	18.1	16.5
All LINEs	22.5	18.3	11.3	7.1	4.4
$\Delta < 2\%$	0.1008	0.0818	0.0410	0.0103	0.0011
$\Delta = 2$–4%	1.41	1.27	0.67	0.17	0.012
$\Delta = 4$–6%	1.1	1.2	0.48	0.11	0.031
$\Delta > 6\%$	19.9	15.8	10.1	6.8	4.3

[a] Isochore family intervals are: less than 37%, 37–41%, 41–46%, 46–52%, and more than 52% GC. Repeat densities were calculated for 10 kb long, non-overlapping sections using RepeatMasker. Alus and LINEs are divided into four categories corresponding to less than 2%, 2–4%, 4–6% and more than 6% divergence from the family consensus sequences (Δ). The general Alu and LINE distributions are also shown.

GC. Given the insertion similarities and (most probably) the same mechanism of propagation, the difference in the genomic distribution of Alus and LINEs (LINE1) raised the question of whether the current distribution of these elements is the result of some specific insertion, or of their stability in specific genomic regions.

The dataset was subdivided into four categories according to the divergence, Δ, from the family consensus sequence (see Table 1). This showed that very young Alus with $\Delta < 2\%$ (1563 of the 280,809 Alus analyzed) were found to be preferentially located in GC-poor isochores. Alus with Δ values of 2–4% show a pattern in which density is already slightly higher in GC-rich isochores, and Alus with $\Delta > 4\%$ are characterized by a bias toward the GC-rich part of the genome. Alus with $\Delta > 6\%$ show the typical distribution of all Alus with a maximum density in isochore family H2. The GC level of 100 kb insertion regions of Alus with $\Delta < 2\%$ is significantly lower compared to Alus with $\Delta > 6\%$ (P-value < 0.001; Mann–Whitney U-test). For young LINEs ($\Delta < 4$), the preference for GC-poor isochores is characterized by about 100-fold higher densities in L1 than in H3 isochores, whereas this ratio is only 5 for old LINEs.

Alus are CpG-rich and contain about 30% of all CpGs in the genome (Hellmann-Blumberg et al., 1993). Since a high rate of mutation of CpGs could affect age estimation based on the divergence from the consensus, we also used Alu classification (Batzer et al., 1996; Jurka, 2000) based on diagnostic sites as another criterion. The youngest family AluYa5 (Deininger and Slagel, 1988) has a genomic distribution similar to the least divergent Alus (Fig. 1a). The slightly older family AluYb8 is already more dense in GC-rich isochores (Fig. 1b), and AluY (Fig. 1c) and the other older families (members of the AluS and AluJ groups; data not shown) are biased toward GC-rich isochores. Remarkably, when young elements from the AluY and AluYb8 families with $\Delta < 2\%$ were considered, their distribution showed a pattern similar to that of AluY5a. In other words, families AluY and AluYb8 comprise relatively young elements with a distribution similar to those of AluY5a, but the majority of older Alus mask this pattern.

3.2. Chromosome 21 and 22 analysis

We also calculated the Alu and LINE1 distributions on human chromosomes 22 and 21 (Dunham et al., 1999; Hattori et al., 2000). The densities of Alu and LINE1 covary with the GC level, LINEs being found in the GC-poor regions and Alus being present in regions with GC enrichment. Insertions of the young Alu family AluYa5 seemed to be independent of the general Alu distribution on chromosomes 21 and 22 (Fig. 2a,b). The distribution of the LINE family L1PA2 is strongly correlated with regions of low GC content. To test this point, the Alu (or LINE) density around each Alu (or LINE) in the database was analyzed within 20 kb flanking sequences (Tables 2 and 3). We found that the density of Alus surrounding each Alu insertion is higher for old families. Similar conclusions hold for LINEs, since older LINEs tend to be preferentially located within LINE clusters. We conclude that both Alus and LINEs can integrate both inside and outside clusters of older elements and that there are chromosomal regions that are more favorable for their accumulation (GC-poor for LINEs and GC-rich for Alus). Remarkable exceptions are the GC-rich telomeric regions of the chromosomes, which are relatively free of Alus and rich in LINEs. This is presumably related to distinct properties of telomeric regions (Saccone et al., 1992; Collins et al., 1996).

The processed pseudogene distribution on both chromosomes was also studied. The majority of pseudogenes on chromosome 22 are intronless and some of them are flanked by direct repeats of variable length, suggesting that they correspond to processed pseudogenes (Dunham et al., 1999). They are more frequent in the GC-poor part of chromosome 22 (Fig. 3), in contrast to the gene distribution (Dunham et al., 1999). On chromosome 21, nine out of ten pseudogenes annotated as processed are localized in L1 isochores and only one in H1 isochores. Analysis of flanking repeats and experimental work indicated that processed pseudogenes are the products of LINE activity (Jurka, 1997; Esnault et al., 2000). Thus, integration into GC-poor regions seems to be a general feature of LINE1-mediated transpositions.

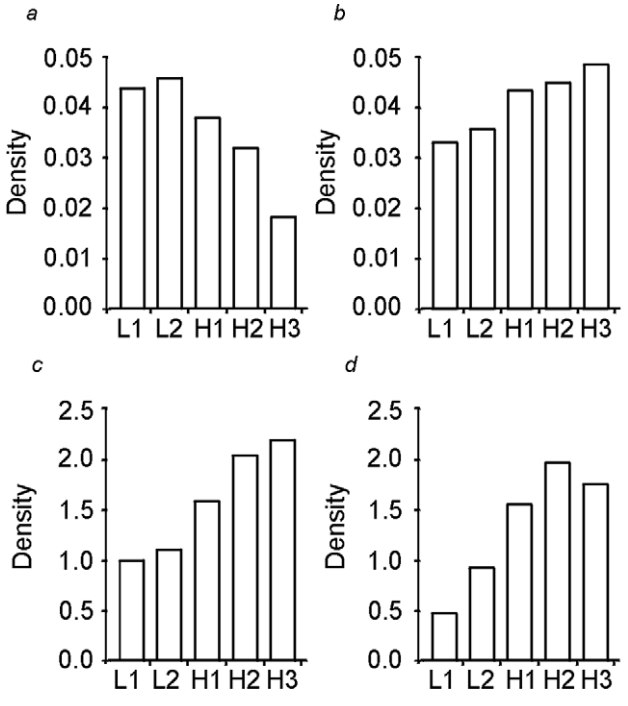

Fig. 1. (a) Genome distribution of the AluYa5 family (mean $\Delta = 2.26\%$). The density was calculated in 100 kb long, non-overlapping fragments. Isochore family intervals are: less than 37%, 37–41%, 41–46%, 46–52%, and more than 52% GC. (b) The distribution of the AluYb8 family (mean $\Delta = 5.23\%$). (c) Distribution of the AluY family (mean $\Delta = 7.15\%$). (d) Distribution of the AluJo family (mean $\Delta = 16.8\%$).

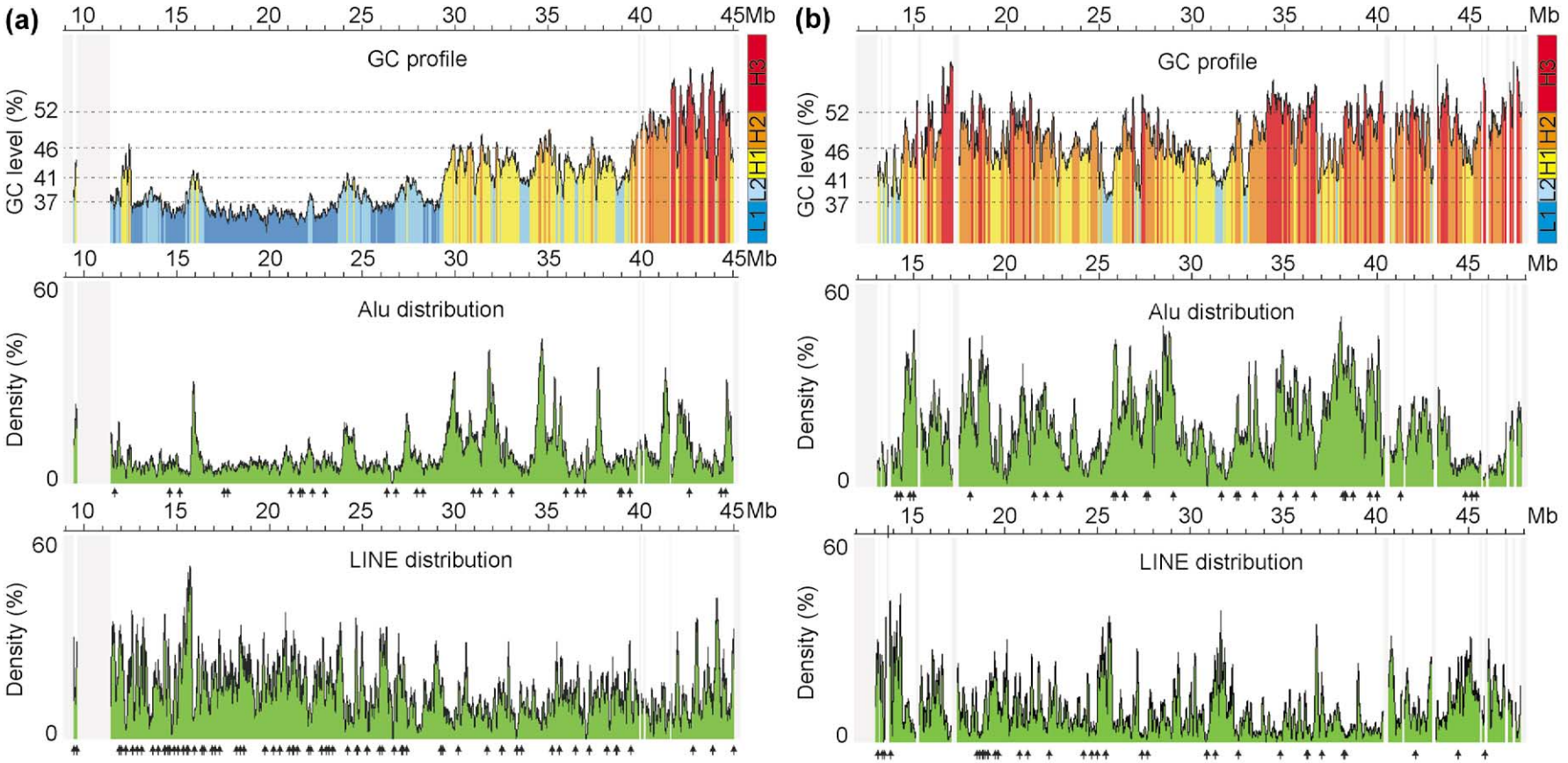

Fig. 2. Distribution of Alus and LINEs on chromosomes 21 (a), 22 (b) and correlation with the GC level. The GC level and the distributions are plotted using a 100 kb sliding window with a 10 kb step. The scale on the top shows positions in Mb on the chromosomes as provided by Human Genome Resources at NCBI; gray breaks represent gaps in the sequences. First, the GC profile was plotted; colors correspond to the GC range of five isochore families as defined in Fig. 1. The second plot shows the Alu density along the chromosome; arrows indicate insertions of the AluYa5 family. The bottom plot corresponds to LINE1 density and insertions of the young L1PA2 family.

Table 2
Alu density in 20 kb sequences flanking Alu insertions[a]

Family	L1	L2	H1	H2	H3
Alu density	5.2	8.1	13.8	18.7	17.2
AluYa5 ($\Delta = 2.25$)	6	9.2	16.7	29	24.8
AluSc ($\Delta = 9.9$)	6.8	12.3	22	30.8	27.9
AluSq ($\Delta = 10.8$)	7.2	13.2	22.3	30.2	27.4
AluJo ($\Delta = 16.8$)	8.2	15.2	23.8	31.7	29.8

[a] For different Alu families we calculated the mean Alu density within 20 kb flanking sequences (10 kb on both 5′ and 3′). In the second row are Alu densities in corresponding genomic compartments; they represent the expected values if Alus integrate and accumulate independently of other Alus. For each Alu family the mean divergence from the consensus (Δ) is shown.

3.3. Implications for repeat stability and genome organization

While young Alus and members of the AluYa5 family are more frequent in the GC-poor part of the genome, the GC-poor preference of young LINEs is even stronger. One possible explanation could be that the insertion pattern of LINEs is more similar to the distribution of young Alus but that LINEs are excluded from the GC-rich part of the genome. This explanation could be accepted if LINEs showed a trend toward accumulation in the GC-poor part of the genome. Instead, young LINEs are more frequent in L1 compared to H3 isochores than old LINEs. This corresponds to a relatively higher exclusion of LINEs from the GC-poor isochores. A second possibility is that the young LINE distribution reflects the Alu and LINE insertions and consequently that a significant part of Alus in the GC-poor isochores has already been lost. This could happen if the Alu excision process were fast enough to exclude new copies before their fixation in the population, which would make their detection impossible. Table 1 and Fig. 2 suggest such fast dynamics. Moreover, if we compare AluYa5 insertions with L1A2 insertions on both chromosomes, we can see differences between these two patterns. The LINE family L1A2 distribution is negatively correlated with the isochore GC level, while the correlation for the AluYa5 family is less clear. An objection to the second explanation could be the different sizes of Alus and LINEs, and possible constraints for insertion of the latter into the GC-rich isochores. Still,

Table 3
LINE1 density in 20 kb sequences flanking LINE1 insertions[a]

Family	L1	L2	H1	H2	H3
L1 density	22.6	19.3	11.3	7.1	4.4
L1PA ($\Delta = 5.1$)	26.7	26.9	17.7	8.9	6.6
L1PA7 ($\Delta = 7.9$)	29.2	29.2	20	10.2	7.2
L1PB1 ($\Delta = 14$)	38	33.1	20.3	9.6	6.3
L1M1 ($\Delta = 18.4$)	40.6	39.3	27.4	16.9	5.9

[a] The same procedure was used for LINES as was used for Alus (Table 2).

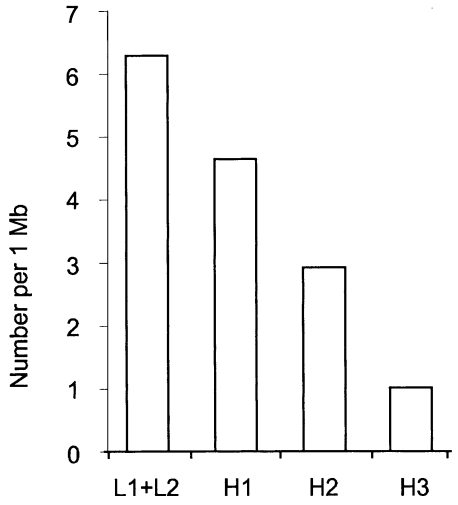

Fig. 3. Isochore distribution of processed pseudogenes on chromosome 22.

LINEs are often 5′ truncated (Voliva et al., 1983) and the mean size of L1PA2 elements is only 1.9 kb instead of 6 kb. Additional support for the second hypothesis comes from the analysis of the processed pseudogene distribution. Since pseudogenes are GC-rich (mean 50.2% GC) and short (mean 941 bp), they resemble Alus more than LINEs. Their preference for GC-poor regions is strong and very similar to that of young LINEs.

It has been suggested that Alus are preferentially fixed by positive selection in GC-rich DNA (Smit, 1999; International Human Genome Sequencing Consortium, 2001). This preference was interpreted as being linked to the hypomethylation of Alus in the male germ line and its suggested function in sperm chromatin assembly (Schmid, 1998). The positive selection and advantageous influence of Alus in the human genome might be a mechanism of Alu accumulation, but this mechanism can not apply to the majority of Alus, which are almost completely methylated in somatic cells (Jabbari and Bernardi, 1998), and more stable in the GC-

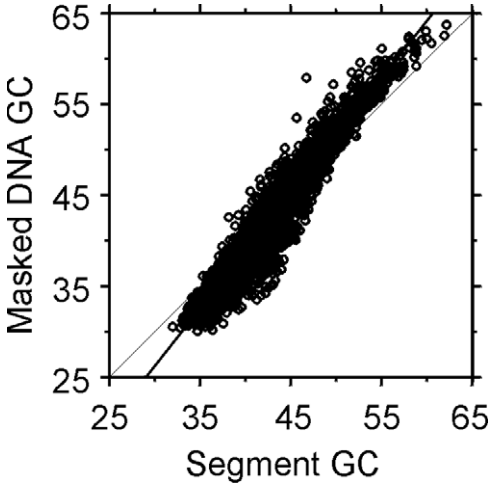

Fig. 4. Correlation between the GC level of 100 kb long segments before and after exclusion of all repeats.

rich part of the genome. In addition, the correlation of Alu and gene density and the proposed Alu involvement in gene regulation (International Human Genome Sequencing Consortium, 2001) are not obvious. Indeed, Alus are most frequent in the H2 isochore family, where genes are less abundant than in H3 isochores (Zoubak et al., 1996; International Human Genome Sequencing Consortium, 2001). Moreover, the most gene-rich and GC-rich telomeric regions of chromosomes 21 and 22 are relatively free of Alus (see Fig. 2a,b). An alternative explanation to positive or negative selection could be the different recombination frequencies for Alus and LINEs. The process could be more effective for short, more frequent Alus. Alternatively, loss of intervening DNA (including other repeats) in Alu-Alu recombination could be responsible for the apparent slow accumulation of LINEs in the GC-rich isochores.

Instead of a highly problematic positive selection, we propose here a mechanism based on the much more common negative selection, in which the accumulation of Alus in the GC-poor isochores is avoided because it would severely change the local composition of these genomic compartments and possibly affect, as a consequence, gene transcription. In other words, we propose that the current pattern of Alu and LINE distribution is, as already suggested (Rynditch et al., 1998), mainly the result of genomic stability, and that a significant part of Alu DNA in the GC-poor part of the genome has been lost. The mechanism of this exclusion is unclear. Alu excision seems to be very rare in the genome (see Edwards and Gibbs, 1992 for the only report), but the presence of a precise cellular mechanism capable of completely removing all traces of previously existing elements cannot be ruled out. The elimination could also be the result of different fixation probabilities and/or recombination frequencies. The latter would be a very efficient exclusion mechanism, as recently shown for direct and inverted Alu copies (Lobachev et al., 2000; Stenger et al., 2001).

Independent support for compositional stability of polyA-retrotransposons comes from other LINE classes. LINE2 and CR1-like elements are more GC-rich than LINE1 (50.6 and 46.6% GC, based on the consensus of LINE2 and CR1-like elements, respectively) and, in contrast to LINE1, are most frequent in the moderately GC-rich part of the genome, and less abundant in L1 isochores (LINE2 has the highest density in H2 isochores, whereas CR1-like elements are the most frequent in L2; unpublished data). MIRs, GC-rich tRNA-derived SINEs, are most frequent in the GC-rich H2 isochores (Smit, 1996, 1999; Matassi et al., 1998). Based on these additional observations, LINE and SINE isochore distributions depend on the repeats' GC level rather than on the repeat length.

An important question is whether isochores are just the result of a specific accumulation of repeats instead of the factor driving the accumulation. From the 100 kb long genomic segments, we excluded all repeats detected by RepeatMasker and calculated the difference between the GC level before and after exclusion of repeats, but the GC level of the segments did not change significantly (Fig. 4). In conclusion, the GC level of large genomic segments is not significantly influenced by the GC level of repeats. In fact, this possibility was already ruled out by experiments showing that GC variation was less than 1% on denatured and reassociated DNA over a 4 log cot range (Soriano et al., 1981).

From the above picture we can conclude that there is a strong selection for maintaining the GC level of long genomic stretches and that new insertions of compositionally non-matching retrotransposons are heavily counter-selected in the human genome. The proposed mechanism of exclusion by recombination could be more efficient for highly repetitive sequences; long distance recombination events are probably negatively selected, and the low copy repeats are thus more likely to keep their original insertion pattern in contrast to very abundant Alus. The strong GC-poor preference of pseudogenes gives support to this hypothesis.

Finally, it should be mentioned that some preferential integration of Alu sequences in the GC-rich isochores may also contribute to the present results. Indeed, the ratio of Alu densities in L1 and H2 is lower than the ratio of insertion sites in these two isochore families (as calculated statistically or actually measured). This preference, which clearly does not apply to LINEs, may have to do with the compositional match between Alus and GC-rich isochores.

Acknowledgements

We thank Oliver Clay and Giorgio Matassi for critical reading of the manuscript. This work was supported by grant No. 204/01/0632 of the Grant Agency of the Czech Republic to J.H. A.P. is supported by a PhD fellowship of the French Government program "Doctorat en cotutelle".

References

Arcot, S.S., et al., 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. J. Mol. Biol. 281, 834–856.

Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zietkiewicz, E., Zuckerkandl, E., 1996. Standardized nomenclature for Alu repeats. J. Mol. Evol. 42, 3–6.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Branciforte, D., Martin, S.L., 1994. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. Mol. Cell. Biol. 14, 2584–2592.

Collins, A., Frezal, J., Teague, J., Morton, N.E., 1996. A metric map of humans: 23,500 loci in 850 bands. Proc. Natl. Acad. Sci. USA 93, 14771–14775.

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

Deininger, P.L., Slagel, V.K., 1988. Recently amplified Alu family members share a common parental Alu sequence. Mol. Cell. Biol. 8, 4566–4569.

Dunham, I., et al., 1999. The DNA sequence of human chromosome 22. Nature 402, 489–495.

Edwards, M.C., Gibbs, R.A., 1992. A human dimorphism resulting from loss of an Alu. Genomics 14, 590–597.

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

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

Hattori, M., et al., 2000. The DNA sequence of human chromosome 21. Nature 405, 311–319.

Hellmann-Blumberg, U., McCarthy Hintz, M.F., Gatewood, J.M., Schmid, C.W., 1993. Developmental differences in methylation of human Alu repeats. Mol. Cell. Biol. 13, 4523–4530.

International Human Genome Sequencing Consortium, 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.

Jang, W., Chen, H.C., Sicotte, H., Schuler, G.D., 1999. Making effective use of human genomic sequence data. Trends Genet. 15, 284–286.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

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

Jurka, J., 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 9, 418–420.

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

Lobachev, K.S., Stenger, J.E., Kozyreva, O.G., Jurka, J., Gordenin, D.A., Resnick, M.A., 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J. 19, 3822–3830.

Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol. 108, 237–254.

Matassi, G., Labuda, D., Bernardi, G., 1998. Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. FEBS Lett. 439, 63–65.

Meunier-Rotival, M., Soriano, P., Cuni, G., Strauss, F., Bernardi, G., 1982. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc. Natl. Acad. Sci. USA 79, 355–359.

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.

Saccone, S., De Sario, A., Della Valle, G., Bernardi, G., 1992. The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA 89, 4913–4917.

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

Smit, A.F., 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743–748.

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

Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes. 2. Reassociation kinetics. Eur. J. Biochem. 115, 235–239.

Soriano, P., Meunier-Rotival, M., Bernardi, G., 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. USA 80, 1816–1820.

Stenger, J.E., Lobachev, K.S., Gordenin, D.A., Darden, T.A., Jurka, J., Resnick, M.A., 2001. Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res. 11, 12–27.

Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. 19, 253–272.

Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison, C.A., Edgell, M.H., 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res. 11, 8847–8850.

Weiner, A.M., Deininger, P.L., Esftradiatis, A., 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. 55, 631–661.

Zerial, M., Salinas, J., Filipski, J., Bernardi, G., 1986. Gene distribution and nucleotide sequence organization in the human genome. Eur. J. Biochem. 160, 479–485.

Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.