The Distributions of "New" and "Old" Alu Sequences in the Human Genome: The Solution of a "Mystery"

Maria Costantini,[1] Fabio Auletta,[1] and Giorgio Bernardi[1,*†]

[1]Laboratory of Cellular and Developmental Biology, Stazione Zoologica Anton Dohrn, Naples, Italy

†Present address: Laboratory of Theoretical Biology and Bioinformatics, Biology Department, Rome 3 University, Rome, Italy.

***Corresponding author:** E-mail: gbernardi@uniroma3.it.

Associate editor: Norihiro Okada

Abstract

The distribution in the human genome of the largest family of mobile elements, the Alu sequences, has been investigated for the past 30 years, and the vast majority of Alu sequences were shown to have the highest density in GC-rich isochores. Ten years ago, it was discovered, however, that the small "youngest" (most recently transposed) Alu families had a strikingly different distribution compared with the "old" families. This raised the question as to how this change took place in evolution. We solved what was considered to be a "mystery" by 1) revisiting our previous results on the integration and stability of retroviral sequences, and 2) assessing the densities of acceptor sites TTTT/AA in isochore families. We could conclude 1) that the open state of chromatin structure plays a crucial role in allowing not only the initial integration of retroviral sequences but also that of the youngest Alu sequences, and 2) that the distribution of old Alus can be explained as due to Alu sequences being unstable in the GC-poor isochores but stable in the compositionally matching GC-rich isochores, again in line with what happens in the case of retroviral sequences.

Key words: Alu, genome, isochores, repeated sequences, SINEs.

Research article

Introduction

Over 40 years ago, the development of preparative ultracentrifugation in Cs_2SO_4 density gradients run in the presence of sequence-specific ligands (Corneo et al. 1968) allowed us to demonstrate that the "main band" DNAs of vertebrates are characterized by a long-range compositional heterogeneity (Filipski et al. 1973; Thiery et al. 1976). In fact, these genomes are mosaics of isochores (Macaya et al. 1976), megabase-size DNA sequences that are endowed with a fairly homogeneous base composition. Isochores are distributed in a small number of families that cover a broad compositional range. For example, in the human genome, a typical mammalian genome, the five isochore families (called L1, L2, H1, H2, and H3, in order of increasing GC), cover a 30–60% GC range (see Bernardi 2004 for a review; Costantini et al. 2006).

The physical separation of isochores belonging to different families allowed us to investigate the genome distribution of specific sequences, such as genes and repeated sequences. By the end of the '70s, hybridization of appropriate probes on isolated isochore families had already led us to localize some genes (Cuny et al. 1978; Cortadas et al. 1979), as well as some integrated retroviral sequences (Kettmann et al. 1979, 1980).

As far as the genome distribution of intermediate repetitive sequences was concerned, we found it to be different in different isochore families as investigated by reassociation kinetics (Soriano et al. 1981). Hybridization experiments showed that the two major classes of interspersed repeats, the GC-poor LINE-1 and the GC-rich Alu (SINE) sequences, had their highest densities in GC-poor and GC-rich human isochores, respectively (Meunier-Rotival et al. 1982; Soriano et al. 1983; Meunier-Rotival and Bernardi 1984; Zerial et al. 1986). This finding pointed not only to a strikingly non-random distribution of SINEs and LINEs, later confirmed by assessments based on genome sequences (Smit 1996, 1999; Jabbari and Bernardi 1998), but also to the existence of compositional correlations between the two classes of interspersed repeats and the isochore families in which they were located. This was, in fact, the first indication that interspersed repeated sequences were obeying genome organization rules, later called the "genomic code" (Bernardi 1990). Further work indicated that genes were also non-uniformly distributed among isochore families. In fact, gene density showed a marked increase from the GC-poorest L1 to the GC-richest H3 isochore families (Bernardi et al. 1985; Mouchiroud et al. 1991; Zoubak et al. 1996). Again compositional correlations were found to hold between coding sequences and extended flanking sequences (Bernardi et al. 1985; Costantini and Bernardi 2008).

When different families of human Alu sequences, on which we will concentrate here, were investigated, it was found that the density of the youngest (the most recently transposed) *Alu*Y sequences, *Alu*Ya5, was at least 2-fold higher in the GC-poorest compared with the GC-richest bins in the human genome (The International Human Genome Sequencing Consortium [IHGSC] 2001) and in L1 compared with H3 isochores of chromosomes 21 and 22 (Pavliček et al. 2001). In other words, the compositional tendencies shown by the youngest Alus appeared to be strikingly different from those shown by older Alu sequences. These results raised a question about the reason for the change of the Alu distribution over evolutionary time.

The IHGSC (2001) considered this question to be a "mystery," especially in view of the fact that LINEs families of

different ages were stably located in GC-poor regions of the genome, that the insertion sequence of Alu is TTTT/AA (where the slash indicates the point of cleavage), which one would expect to be more frequent in GC-poor DNA, and that insertion depends on the transposition machinery encoded by LINEs (the last point being irrelevant in our opinion). The IHGSC (2001) considered three possible explanations for the accumulation of Alus in GC-rich isochores: 1) Alus target GC-rich isochores for insertion; 2) Young Alus target GC-poor isochores, but there is a higher rate of random loss, elimination by negative selection, or more tolerance for their deletion; and 3) Alus are positively selected in GC-rich isochores. The conclusion was that "the first two possibilities seem unlikely because AT-rich DNA is gene-poor and tolerates the accumulation of other transposable elements. The third seems to be more feasible . . . but positive selection in GC-rich regions would imply that they benefit the organism" and "clearly, much additional work will be needed to prove or disprove the hypothesis that SINEs are genomic symbionts."

Immediately after the IHGSC (2001) presented this viewpoint, we critically discussed it and proposed that the evolutionary shift of Alus from GC-poor to GC-rich isochores was due to their higher stability in compositionally matching chromosomal regions (Bernardi 2001; Pavliček et al. 2001). This idea mainly arose from our work on the integration, stability, and expression of retroviral sequences (see Results and Discussion) and was further discussed in Bernardi (2004).

Here, we have reanalyzed this problem because 1) the gaps present in the draft sequence (IHGSC 2001) that mainly concerned GC-rich regions were filled in later in the human genome sequence that is now used (see Materials and Methods and Lander 2011); 2) the definition of Alu families is no longer based on divergence from the consensus but on actual sequences; and 3) the borders of isochore families are now more precisely defined (Costantini et al. 2006). The main point of this investigation was, however, that we assessed the density of Alu acceptor sites TTTT/AA in different isochore families taking advantage of recent work on short-sequence frequencies in isochore families (Costantini and Bernardi 2008), which allowed an appropriate comparison of Alu insertions and acceptor site frequencies. Incidentally, acceptor site frequencies were taken into consideration previously (Bernardi 2001; 2004; Pavliček et al. 2001), but a precise assessment was only done more recently (Costantini and Bernardi 2008).

Materials and Methods

The coordinates of repeat sequences were downloaded from the UCSC Web site (http://genome.ucsc.edu). Because release hg17, which was used for mapping isochores of the human genome (see Costantini et al. 2006), differs from the latest release hg19 only in that a small number of gaps were filled in, we localized Alu sequences on release hg19 (Kent et al. 2002; IHGSC 2004). According to Batzer et al. (1996) "In the universal nomenclature a logical (alphabetical) progression from the oldest (J) to intermediate (S) and young (Y) Alu sequences using capital letters is used to denote major subfamilies branches." In the present work, we took into consideration *AluJo* as representative of old families, *AluSq* and *AluSc* as intermediate famiiles, and *AluY*, *AluYa5*, and *AluYb8* as youngest families. *AluYa5/8* was also analyzed because of previously considered polymorphic *AluYa5* repeats (Arcot et al. 1997, 1998).

We calculated the density values for each Alu family as the number of sequences per megabase of isochore. Alu densities and Alu numbers were plotted against the mean GC values of each isochore family (36% for L1, 39% for L2, 43% for H1, 48.5% for H2, and 54.5% for H3).

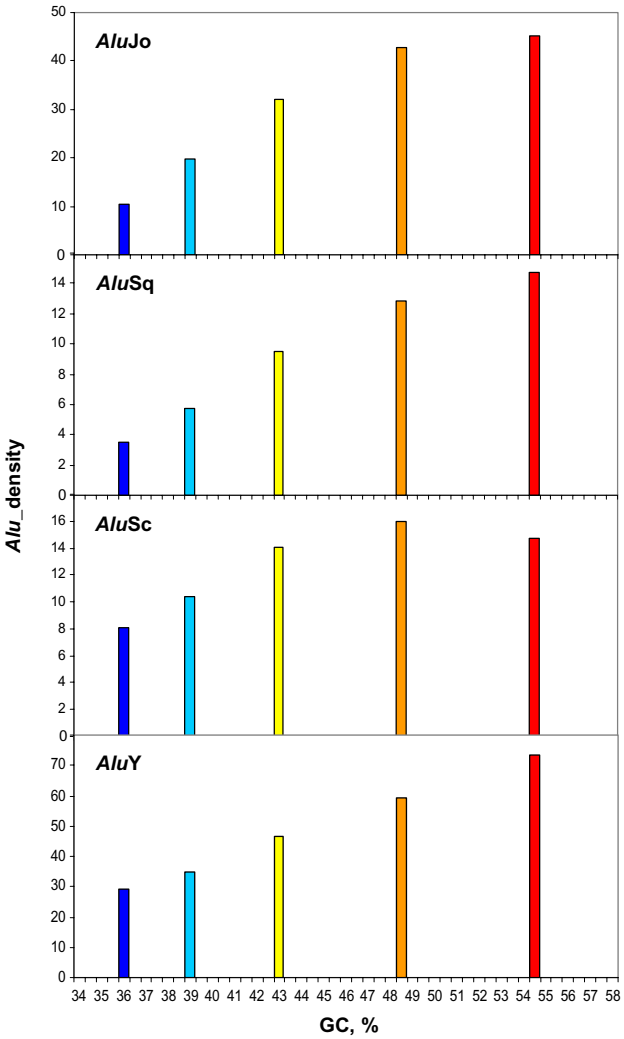

FIG. 1. Densities (counts per megabase) of *Alu*Jo, AluSq, *Alu*Sc, and *Alu*Y families in the five isochore families (L1, L2, H1, H2, and H3) of the human genome.

Results and Discussion

Figure 1 displays the densities of *Alu*Jo, *Alu*Sq, *Alu*Sc, and *Alu*Y sequences in the isochore families of the human genome. All density plots show an increase from isochore family L1 to isochore family H3, the ratio L1/H3 showing an increase in *Alu*Sc and *Alu*Y compared with *Alu*Jo and *Alu*Sq.

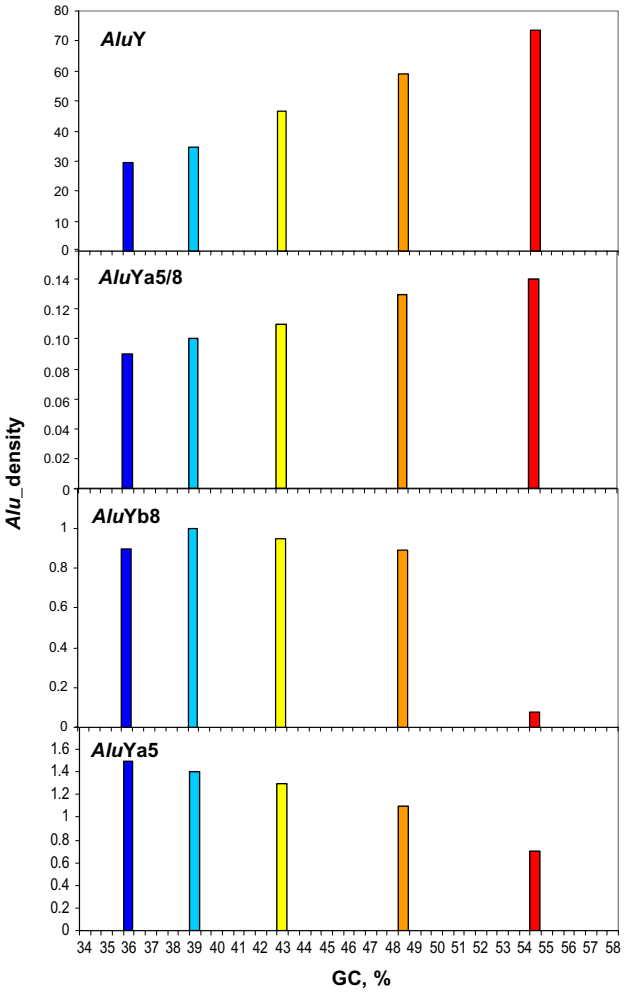

FIG. 2. Densities (counts per megabase) of *Alu*Y, *Alu*Ya5/8, *Alu*Yb8, and *Alu*Ya5 families in the five isochore families (L1, L2, H1, H2, and H3) in the human genome.

The distribution is different in the case of *Alu*Ya5/8, *Alu*Yb8, and *Alu*Ya5, which are compared with the *Alu*Y in figure 2. *Alu*Ya5/8 shows a regular increase in density in increasingly GC-richer isochore families, namely a distribution pattern rather similar to that of the AluY family. In *Alu*Yb8, the densities are similar in all isochore families (except for the H3 isochore family, which shows a still unexplained very low density). In *Alu*Ya5, densities show a 2-fold lower level in the GC-poorest compared with the GC-richest isochore family. Both densities and numbers of "old" and "young" Alu sequences are presented in table 1.

The results of figures 1 and 2 provide a picture of the distinction of old and young Alu sequences which is based on the sequence distribution in the genome, and this picture fits with that established on the basis of sequence divergence (Batzer et al., 1996; Batzer and Deininger, 2002). Indeed, the distribution of the two families that are most divergent from the consensus, *Alu*Jo and *Alu*Sq, shows a very steep increase in Alu sequence density with increasing isochore GC compared with *Alu*Sc and *Alu*Y, and the series *Alu*Ya5, *Alu*Yb8, *Alu*Ya5/8, and *Alu*Y seem to catch the transition between young and old Alus.

We will now discuss two questions. The first question is why the density of the youngest *Alu*Ya5 is higher in GC-poor compared with GC-rich isochores (see above). The second question is why there was a shift in the Alu distribution with evolutionary time. Both questions can be answered in the light of our previous results on the distribution of retroviral sequences in the isochores of the mammalian genome and of the assessment of acceptor site densities in isochore families.

Our earlier investigations on 45 complete retroviral genomes showed that they exhibit a bimodal compositional distribution, consisting of two different classes, GC poor (about 43% GC) and GC rich (about 53% GC; Zoubak et al. 1992; Tsyba et al. 2004; see also Bernardi 2004 for a review). Investigations on retrovirus integration were carried out on both GC-rich retroviruses, such as BLV (bovine leukemia; Kettmann et al. 1979, 1980), RSV (Rous sarcoma; Rynditch et al. 1991), HTLV-I (human T-cell leukemia; Zoubak et al. 1994), and MuLV (murine leukemia; Rynditch et al. 1998), and on GC-poor retroviruses, such as MMTV (mouse mammary tumor; Salinas et al. 1987) and HIV (Glukhova et al. 1999; Tsyba et al. 2004). The results so obtained showed that the initial integration into GC-rich isochores (which correspond to open chromatin regions in cells from adult organisms; Saccone et al., 2002; Di Filippo and Bernardi 2008; see fig. 3A) is favored not

Table 1. Number and Density Values of Alu Families per Megabase in the Five Isochore Families of the Human Genome.

	Number						Density				
Alu family	**L1**	**L2**	**H1**	**H2**	**H3**	**Total**	**L1**	**L2**	**H1**	**H2**	**H3**
Old											
AluJo	**6,237**	**20,566**	**25,140**	**15,076**	**3,039**	**70,058**	**10.30**	**19.80**	**32.20**	**42.80**	**45.20**
AluSc	**4,975**	**10,788**	**11,032**	**5,640**	**990**	**33,425**	**8.10**	**10.40**	**14.10**	**16.00**	**14.70**
AluSq	**2,177**	**5,966**	**7,494**	**4,499**	**991**	**21,127**	**3.50**	**5.70**	**9.50**	**12.80**	**14.70**
AluY	**18,079**	**36,002**	**36,329**	**20,885**	**4,935**	**116,230**	**29.40**	**34.60**	**46.50**	**59.20**	**73.40**
						240,840					
Young											
AluYa5	**914**	**1,438**	**994**	**385**	**50**	**3,781**	**1.50**	**1.40**	**1.30**	**1.10**	**0.70**
AluYb/8	**606**	**1,059**	**747**	**312**	**52**	**2,776**	**0.90**	**1.00**	**0.95**	**0.89**	**0.08**
AluYa5/8	**59**	**114**	**91**	**48**	**10**	**322**	**0.09**	**0.10**	**0.11**	**0.13**	**0.14**
						6,879					

NOTE.—Numeric values in italics correspond to the sum of the total number for "old" and "young" Alu families.

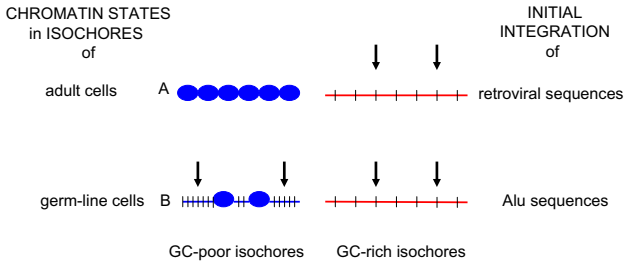

Fig. 3. A scheme of chromatin states (open/closed) in GC-poor and GC-rich isochores as present in adult cells (*A*) and germ cells (*B*) along with the initial integration of retroviral and Alu sequences. Ovals represent nucleosome and vertical bars represent acceptor sites.

only by the GC-rich but also by the GC-poor proviruses, a tendency which was corroborated by other studies (Elleder et al. 2002; Schroeder et al. 2002). This is a strong indication that the open chromatin of GC-rich isochores plays a major role in the initial proviral integration. Stability of integration and expression are, however, associated with an "isopycnic" localization, namely a localization within compositionally matching isochores of the host genome (see Rynditch et al. 1998 and Bernardi 2004 for reviews). In conclusion, the initial integration of both GC-rich and GC-poor retroviral sequences targets the open chromatin of GC-rich isochores, the chromatin of GC-poor isochores being typically closed in the cells of adult organisms. In contrast, their stability (and expression) of the integrated provirus is dictated by a compositional match with the host sequences, GC-poor, and GC-rich retroviral sequences being stable only in GC-poor and GC-rich isochores, respectively.

Going back now to the first question, contrary to a first impression based on the results of *Alu*Ya5 (see fig. 2), the acceptor site frequencies actually indicate a preference for integration into GC-rich isochores even of these youngest Alus. This surprising conclusion derives from an assessment of Alu acceptor sites in isochore families. We know that the frequencies of short sequences in the human genome (and in all genomes for that matter) are not random and,

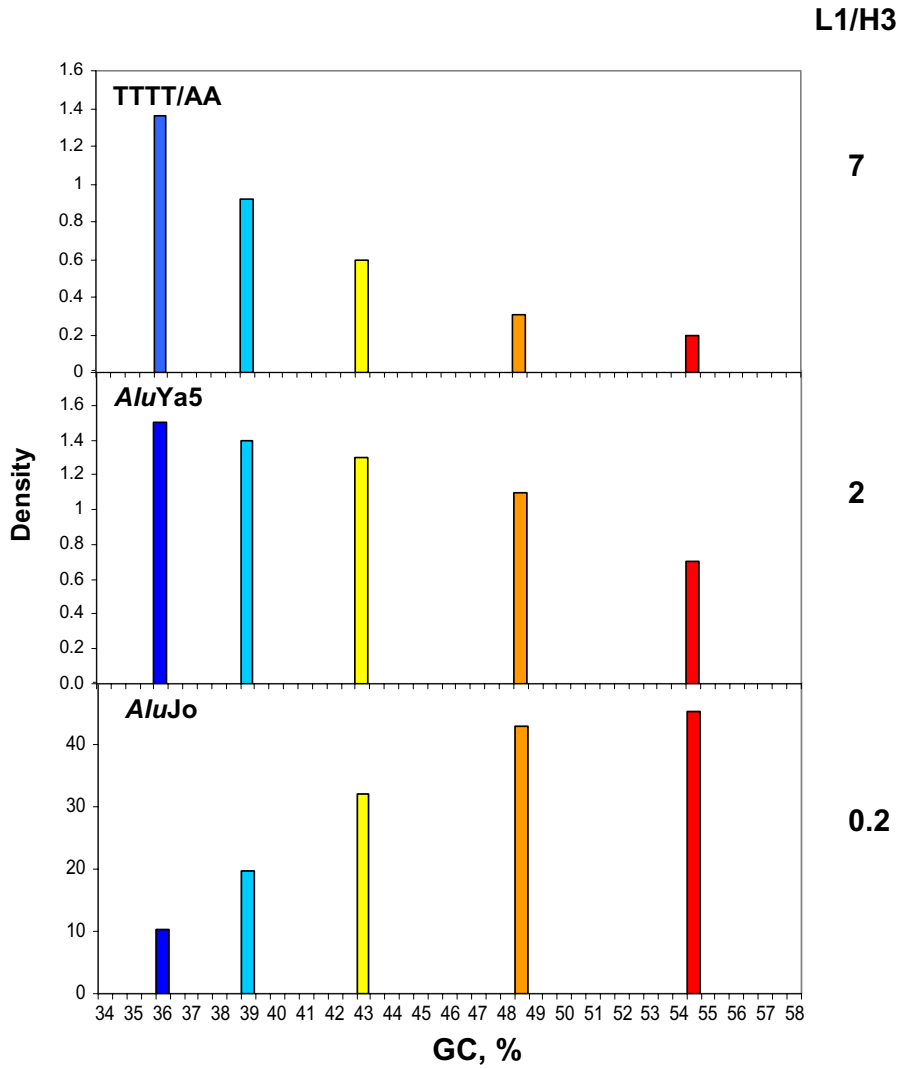

Fig. 4. Densities (counts per megabase) of Alu acceptor sites TTTT/AA ($\times 10^{-3}$) in isochore families are compared with the densities of representative "young" and "old" Alu sequences. The L1/H3 ratios of densities are also shown.

therefore, they cannot be calculated from base composition. However, frequencies of trinucleotides as assessed in different isochore families are now available (Costantini and Bernardi 2008). The frequencies of TTTT/AA could then be calculated as the product of the frequency of TTT by that of TAA for different isochore families (see fig. 4). The Alu acceptor sites turn out to be seven times more frequent in L1 than in H3. In other words, if the distribution of the youngest Alu sequences only depended upon the frequency of acceptor sites, one should find a distribution in which frequencies in the GC-poorest isochore families would be seven times higher than in the GC-richest ones, whereas only a 2-fold difference was found.

This difference can be understood if one takes into consideration the following points. 1) The retrotransposition process of Alu sequences takes place in the germ-line cells, more so in the male (Jurka 2004). 2) In those cells, open chromatin regions are present not only in GC-rich isochores (as in adult cells) but also, partially at least, in GC-poor isochores of the genome (fig. 3). In fact, it is known that GC-poor regions of the genome preferentially harbor genes that are primarily active during development (Hiratani et al. 2004; Kikuta et al. 2007; Ren et al., 2007; Navratilova and Becker 2009) and that those regions are shut off in the closed chromatin of the adult where the chromatin of GC-rich isochores typically remains open (Saccone et al. 2002; Di Filippo and Bernardi 2008; see fig. 3). 3) The acceptor sites of Alu sequences TTTT/AA have their highest levels in L1 isochores and their lowest ones in H3 isochores. These points provide the solution of the mystery since they indicate that the initial integration (as represented by the youngest Alu sequences) targets both the relatively rare acceptor sites of the open GC-rich isochores and the more frequent acceptor sites of the partially open GC-poor isochores of germ cells (see fig. 3B). This indicates, however, a preference of Alu integration in GC-rich isochores because of the 7-fold lower density of Alu acceptor sites (fig. 4) and the 9-fold lower amount (fig. 5, top panel) of H3 compared with L1 isochores.

The second question, concerning the shift from the initial to the final Alu distribution, can be answered by proposing, as done before (Bernardi 2001, 2004; Pavliček et al. 2001), that the "shift" in isochore localization of the old Alu sequences compared with young ones is due to the instability of the latter in compositionally non-matching isochores. There are, however, some additional factors that could contribute for the higher density of Alu sequences in GC-rich isochores. Indeed, 1) segmental duplications are more frequent in GC-rich regions, 2) Alu sequences favor such duplications, and 3) duplications increase Alu frequencies because the duplicated regions are Alu rich (Jurka et al. 2004). Moreover, a comparison of the genome of Craig Venter with the human reference genome (Costantini and Bernardi 2009) showed that both deletions and insertions are extremely scarce in L1 isochores and increase gradually in increasingly GC-rich isochore families. In fact, frequencies of deletions and

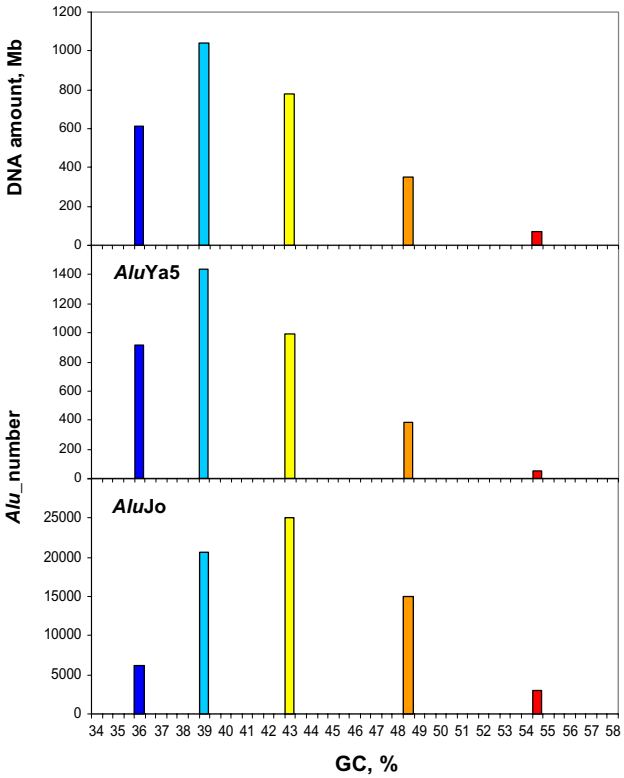

FIG. 5. Numbers of representative "young" and "old" Alu sequences are compared with each other and with DNA amounts (top panel) in the corresponding isochore families.

insertions largely parallel the densities of genes, of retroviral integrations, and of Alu sequences.

As a final point, it should be noted that if the actual numbers of Alu sequences are taken into consideration (see table 1 and supplementary figs. 1 and 2, Supplementary Material online) instead of their densities as done so far, the distribution among the isochore families is strongly influenced, as expected, by the amounts of DNA in different isochore families (fig. 5). Still, the different distributions of young and old Alus can be perceived.

At a more general level, it is obvious that the concept of "selfish DNA" (Doolittle and Sapienza 1980; Orgel and Crick 1980) as applied to sequences such as Alu sequences should be abandoned for good since Alu sequences can modulate gene expression (Cordaux and Batzer 2009; Shen et al. 2011), play a role in nucleosome formation (Tanaka et al. 2010; Bettecken et al. 2011), increase local recombination rates (Witherspoon et al. 2009), and contribute to the stabilization of the gene-rich isochores of the mammalian genome. The latter point is suggested by a comparison of two congeneric fishes living at 40 °C–42 °C and 15 °C–20 °C, respectively, which has shown an amplification of a GC-rich minisatellite in the gene-rich regions of the former (Bucciarelli et al. 2009).

To sum up, we dealt here with two basic subjects (see fig. 3). The first one is the integration of young Alu sequences in the genome. Although this can only occur at acceptor sites, the latter may be available, in open chromatin, or not

available, in closed chromatin. In adult cells, open chromatin is typically restricted to GC-rich isochores, and this explains why retroviral sequences, whether GC rich or GC poor, integrate preferentially in GC-rich isochores. In germ cells, however, open chromatin is not only present in GC-rich isochores but also to some extent in GC-poor isochores. Alu sequences can, therefore, integrate in both GC-poor isochores and GC-rich isochores. If one takes into account that L1 isochores are 9-fold more abundant and have a 7-fold higher density of acceptor sites, it is difficult to escape the conclusion that even the initial integration of Alu sequences has a preference for the GC-richest isochores compared with the GC-poorest ones. With evolutionary time, both retroviral sequences and Alu sequences are predominantly found in compositionally matching isochores where they fulfill a role, by being expressed in the case of retroviral sequences or by contributing to the modulation of gene expression, helping nucleosome formation, and providing thermodynamic stability. In other words, integrated sequences are unstable and are eventually lost in compositionally non-matching isochores. These conclusions rule out the explanations previously proposed (IHGSC 2001) and stress the crucial role of the chromatin state and the sequence context for the initial integration and the integration stability of both Alu and retroviral sequences.

Acknowledgments

The authors are very grateful to Oliver Clay for the critical reading of this paper. This work was supported by the author's institute, Stazione Zoologica Anton Dohrn. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Arcot SS, Adamson AW, Risch GW, LaFleur J, Robichaux MB. 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *J Mol Evol.* 281:843–856.

Arcot SS, De Angelis MM, Sherry ST, Adamson AW, Lamerdin JE, Deininger PL, Carrano AV, Batzer MA. 1997. Identification and characterization of two polymorphic Ya5 Alu repeats. *Mutat Res.* 382:5–11.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet.* 3:370–379.

Batzer MA, Deininger LD, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E. 1996. Standardized nomenclature for Alu repeats. *J Mol Evol.* 42:3–6.

Bernardi G. 1990. Le génome des vertébrés: organisation, fonction et evolution. *Biofutur* 94:43–46.

Bernardi G. 2001. Misunderstandings about isochores. Part 1. *Gene* 276:3–13.

Bernardi G. 2005. Structural and evolutionary genomics. Natural selection in genome evolution. Amsterdam: Elsevier.

Bernardi G, Olofsson 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.

Bettecken T, Frenkel ZM, Trifonov EN. 2011. Human nucleosomes: special role of CG dinucleotides and Alu-nucleosomes. *BMC Genomics*. 12:273.

Bucciarelli G, Di Filippo M, Costagliola D, Alvarez-Valin F, Bernardi G, Bernardi G. 2009. Environmental genomics: a tale of two fishes. *Mol Biol Evol.* 26:1235–1243.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet.* 10:691–703.

Corneo G, Ginelli E, Soave C, Bernardi G. 1968. Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. *Biochemistry* 7:4373–4379.

Cortadas J, Olofsson B, Meunier-Rotival M, Macaya G, Bernardi G. 1979. The DNA components of the chicken genome. *Eur J Biochem.* 99:179–186.

Costantini M, Bernardi G. 2008. Correlations between coding and contiguous non-coding sequences in isochore families from short sequence vertebrate genomes. *Gene* 410:241–248.

Costantini M, Bernardi G. 2009. Mapping insertions, deletions and SNPs on Venter's chromosomes. *PLoS One*. 4:e5972.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. *Genome Res.* 16:536–541.

Cuny G, Macaya G, Meunier-Rotival M, Soriano P, Bernardi G. 1978. Some properties of the major components of the mouse genome. In: Boyer WH, Nicosia S, editors. Genetic engineering. Amsterdam: Elsevier. p. 109–115.

Di Filippo M, Bernardi G. 2008. Mapping DNase I-hypersensitive sites on human isochores. *Gene* 419:62–65.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601–603.

Elleder D, Pavlicek A, Paces J, Hejnar J. 2002. Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence. *FEBS Lett.* 517:285–286.

Filipski J, Thiery JP, Bernardi G. 1973. An analysis of the bovine genome by $Cs_2SO_4^-Ag^+$ density gradient centrifugation. *J Mol Biol.* 80:177–197.

Glukhova LA, Zoubak SV, Rynditch AV, Miller GG, Titova IV, Vorobyeva N, Lazurkevitch ZV, Graphodatskii AS, Kushch AA, Bernardi G. 1999. Localization of HTLV-1 and HIV-1 proviral sequences in chromosomes of persistently infected cells. *Chromosome Res.* 7:177–183.

Hiratani I, Leskovar A, Gilbert DM. 2004. Differentiation-induced replication-timing are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proc Natl Acad Sci U S A.* 101:16861–16866.

International Human Genome Sequencing Consortium (IHGSC). 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

International Human Genome Sequencing Consortium (IHGSC). 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.

Jabbari K, Bernardi G. 1998. CpG doublets, CpoG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* 224:123–128.

Jurka J. 2004. Evolutionary impact of human Alu repetitive elements. *Curr Opin Genet Dev.* 14:603–608.

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

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler DWJ. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006.

Kettmann R, Cleuter Y, Mammerickx M, Meunier-Rotival M, Bernardi G, Burny A, Chantrenne H. 1980. Genomic integration of bovine leukemia provirus: comparison of persistent lymphocytosis with lymph node tumor form of enzootic bovine leukemia. *Proc Natl Acad Sci U S A.* 77:2577–2581.

Kettmann R, Meunier-Rotival M, Cortadas J, Cuny G, Ghysdael J, Mammerickx M, Burny A, Bernardi G. 1979. Integration of

bovine leukaemia virus DNA in the bovine genome. *Proc Natl Acad Sci U S A.* 76:4822–4826.
Kikuta H, Laplante M, Navratilova P, et al. (17 co-authors). 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* 17:545–555.
Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* 470:187–197.
Macaya G, Thiery JP, Bernardi G. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol.* 108:237–254.
Meunier-Rotival M, Bernardi G. 1984. The Bam repeats of the mouse genome belong in several superfamilies the longest of which is over 9 kb in size. *Nucleic Acids Res.* 12:1593–1608.
Meunier-Rotival M, Soriano P, Cuny 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 U S A.* A79:355–359.
Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G. 1991. The distribution of genes in the human genome. *Gene* 100:181–187.
Navratilova P, Becker TS. 2009. Genomic regulatory blocks in vertebrate and implication in human disease. *Brief Funct Genomic Proteomic.* 8:333–342.
Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604–607.
Pavlíček A, Jabbari K, Pačes J, Pačes V, Hejnar J, Bernardi G. 2001. Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 276:39–45.
Ren L, Gao G, Zhao D, Ding M, Luo J, Deng H. 2007. Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. *Genome Biol.* 8:R35.
Rynditch A, Kadi F, Geryk J, Zoubak S, Svoboda J, Bernardi G. 1991. The isopycnic, compartmentalized integration of Rous. *Gene* 106:165–172.
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, Federico C, Bernardi G. 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* 300:169–178.
Salinas J, Zerial M, Filipski J, Crépin M, Bernardi G. 1987. Non-random distribution of MMTV proviral sequences in the mouse genome. *Nucleic Acids Res.* 15:3009–3022.
Schroeder ARW, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* 110:165–172.
Shen S, Cai JJ, Jinag P, Kenkel EJ, Stroik M, Sato S, Davidson BL, Xing Y. 2011. Widespread establishment and regulatory impact of Alu exons in human genes. *Proc Natl Acad Sci U S A.* 108:2837–2842.
Smit AF. 1996. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev.* 6:743–748.
Smit AF. 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 non-uniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A.* 80:1816–1820.
Tanaka Y, Yamashita R, Suzuki Y, Nakai K. 2010. Effects of Alu elements on global nucleosome positioning in the human genome. *BMC Genomics.* 11:309–318.
Thiery JP, Macaya G, Bernardi G. 1976. An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol.* 108:219–235.
Tsyba L, Rynditch AV, Boeri E, Jabbari K, Bernardi G. 2004. Distribution of HIV-1 in the genomes of AIDS patients. *Cell Mol Life Sci.* 61:721–726.
Witherspoon DJ, Watkins WS, Zhang Y, Xing J, Tolpinrud WL, Hedges DJ, Batzer MA, Jorde LB. 2009. Alu repeats increase local recombination rates. *BMC Genomics.* 10:530–540.
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.
Zoubak S, Richardson J, Rynditch A, Hîllsberg P, Hafler D, Boeri E, Lever AML, Bernardi G. 1994. Regional specificity of HTLV-I proviral integration in the human genome. *Gene* 143:155–163.
Zoubak S, Rynditch A, Bernardi G. 1992. Compositional bimodality and evolution of retroviral genomes. *Gene* 119:207–213.