Nonrandom Distribution of Alu Elements in Genes of Various Functional Categories: Insight from Analysis of Human Chromosomes 21 and 22

Deepak Grover, Partha P. Majumder,† Chandrika B. Rao,* Samir K. Brahmachari,* and Mitali Mukerji**

*Functional Genomics Unit, Institute of Genomics and Integrative Biology (formerly Centre for Biochemical Technology), CSIR, Delhi, India; and †Anthropology and Human Genetics Unit, Indian Statistical Institute, Kolkata, India

The first draft of the human genome has revealed enormous variability in the global distribution of Alu repeat elements. There are regions such as the four homeobox gene clusters, which are nearly devoid of these repeats that contrast with repeat dense regions in other transcriptionally active regions of the genome. Our analysis of the completely sequenced chromosomes 21 and 22 revealed a striking bias in Alu distribution. These elements are more clustered in genes which are involved in metabolism, transport, and signaling processes. In contrast, they are significantly fewer in genes coding for information pathway components as well as structural proteins. This bias in Alu distribution is independent of the effect of Alu density of the flanking genomic region and is also not affected by the GC content of the gene and its upstream and downstream regions. The relative proportions of Alu subfamilies (Alu J, Alu S, and Alu Y) are not significantly different in genes with high Alu density belonging to the functional categories of transport, metabolism, and signaling. However, in the structural proteins and information genes, these proportions are lower than the other three categories. We suggest that Alu elements might be involved in regulatory mechanisms and are therefore differentially selected in primate genomes.

Introduction

The transposon-derived Alu elements, present exclusively in the primates, are the most abundant repeat elements in terms of copy number (1,090,000) and the second most abundant in terms of genome coverage (~15%) in the human genome (Lander et al. 2001). They belong to the SINE family of repeat elements and are predominantly present in the noncoding regions. The Alu repeats are divided into various subfamilies, namely Alu J (oldest), Alu S (intermediate age), and Alu Y (youngest) on the basis of their evolutionary age (Willard, Nguyen, and Schmid al. 1987; Britten et al. 1988; Jurka and Smith 1988; Labuda and Striker 1989). These subfamilies are further classified into sub-subfamilies based on their divergence from consensus sequence (Jurka and Milosavljevic 1991). A comparative analysis of genes across organisms has revealed that a number of homologous genes have accumulated Alus (Li et al. 1999). A minority of the Alus are still active and amplifying in the human genome (Deininger and Batzer 1999). Involvement of Alus in various functions and their association with various genetic disorders have been proposed in the course of studies carried out on disparate genes (Englander, Wolffe, and Howard 1993; Englander and Howard 1995; Norris et al. 1995; Babich et al. 1999; Deininger and Batzer 1999). Although there are indications about their role in evolving functional complexity and gene regulation (Kidwell and Lisch 1997; Hamdi et al. 2000), the basis of their retention and maintenance in 1 million copies in the human genome is still not clear. Accumulating evidence now shows that complex phenotypic traits observed in mammals are caused not only by genes and/or environment but also by heritable epigenetic modification of genes by retrotransposons (Whitelaw and Martin 2001). In an attempt to explore the functional role of the Alus at the genome-wide level, we have carried out an extensive analysis of the distribution of these repeats in the completely sequenced human chromosomes 21 and 22.

Key words: Alu, repeat distribution, chromosomes 21 and 22, functional classification, retrotransposons, gene regulation.

Methods

The nucleotide sequence, as well as information about the associated Alu repeats and genes of chromosome 22, was retrieved from the Web site http://www.sanger.ac.uk (version 2.4) and the same information for chromosome 21 was retrieved from http://hgp.gsc.riken.go.jp (Dunham et al. 1999; Hattori et al. 2000).

Detailed inspection of these chromosomes revealed a wide variability in the sizes of genes. Sizes ranged from as few as several hundred base pairs to as many as 0.8 million bp. To avoid inappropriate inferences about correlation arising from differences in the sizes of genes, the total Alu size (base pairs of an interval occupied by Alus) and total gene size (base pairs of an interval occupied by genes) were taken as measures of Alu and gene densities, rather than their numbers.

Correlation between Alu repeat and gene density was calculated for non-overlapping windows along the whole chromosome of sizes 100 kb, 200 kb, 500 kb, and 1,000 kb.

The density of Alu elements in each gene was expressed as a percentage, calculated using the expression, Alu percentage = [Alu size (bp)/Gene size (bp)] × 100. Because Alus are mostly present in the introns, there is a possibility that the differences in Alu density observed in the genes could be due to the small length of a gene or the absence of introns in it. Therefore, in a separate analysis, the exonic regions of the genes were excluded in the calculation of gene sizes.

The genes on chromosome 21 and 22 were classified into five functional classes: structural proteins, information storage and processing proteins, signaling pathways, metabolism proteins, and transport and binding proteins. The classification was based on information about function of the gene provided at Locus Link (http://www.ncbi.nlm.

Table 1
The Details of Chromosomes 21 and 22 in Terms of Size, Number of Genes, and Various Subfamilies of Alus

			No. of Alu Repeats				Alu Density		
Chromosome	No. of Genes	Size (Mb)	Alu J	Alu S	Alu Y	Total	Alu No.: Gene No.	Alu Covered Region in Genes (%)	Fraction of Alus in Gene (%)
21	285	33.8	2,741	6,992	1,880	12,341	43.3	10.7	34.5
22	814	34.6	5,270	13,506	3,073	21,993	27.0	17.2	42.5

nih.gov/LocusLink/), GeneCard (http://bioinfo.weizmann.ac.il/cards/), Gene Quiz Web server (http://www.sander.ebi.ac.uk/gqsrv/), Gene Ontology (http://www.geneontology.org), and the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/). Only those genes which were well characterized in terms of function and expression were considered (175 in chromosome 22, and 93 in chromosome 21, see Supplementary Material online).

Statistical tests of significance and relationship among different variables—e.g., Alu subfamily frequencies, Alu percentage, functional class, chromosome, and GC content—were carried out by the chi-square test, regression analysis, and analysis of variance (ANOVA).

Results

Chromosomes 21 and 22 differ substantially in both Alu density and gene density. The chromosomes are of similar size, but chromosome 22 has four times as many genes and twice as many Alu repeats (table 1). Interestingly, even though chromosome 22 has more Alu elements than chromosome 21, former seems to be less Alu dense based on its gene density (ratio of number of Alus and number of genes) than chromosome 21. However, the genes of chromosome 22 were found to be more enriched in Alu elements than those of chromosome 21 (table 1) as shown by higher values of both, the fraction of Alus in the genes, and Alu coverage in the genes.

The representation of Alu subfamilies has been found to be different in the genome. Alu S has the highest representation, followed by Alu J and Alu Y. We observed that the distribution of the Alu subfamilies within the genes is significantly different from their distribution in the intergenic region. This was true for both chromosome 21 (chi-square = 16.253, df = 2, $P = 0.0003$) and chromosome 22 (chi-square = 10.064, df = 2, $P = 0.0065$). Alu subfamily distribution in the intragenic region of the two chromosomes was not significantly different (chi-square = 4.932, df = 2, $P = 0.0849$), whereas the difference was highly significant for the intergenic region (chi-square = 31.91, df = 2, $P < 0.0001$).

In accordance with the previous observations (Chen et al. 2002), we observed a significant positive correlation ($P = 0.0001$) between Alu density and gene density in both the chromosomes at various window sizes ranging from 1,000 kb to 50 kb. However, the scatter plot of gene density versus Alu density (shown for the 200 kb window size) showed that this is not an all-or-none phenomenon (fig. 1). Some regions of high gene density are extremely Alu poor and vice versa.

To test whether there is a selective association of genes with Alus, we initially classified the genes of chromosomes 21 and 22 into five functional categories: structural proteins, information storage and processing proteins, metabolism proteins, signaling pathway proteins, and transport and binding proteins. We then calculated the Alu density in each functional category. Analysis of variance of Alu density between the functional categories revealed that genes coding for structural proteins and information storage and processing components were either devoid of Alu elements or were rarely associated with them. However, genes involved in metabolism and transport and in binding processes were extremely rich in Alus (F value = 14.294, df = 4, 266, $P < 0.0001$; fig 2).

It is possible that the differences in Alu density among the different functional categories of genes could be biased by intrinsic properties of the adjacent genomic sequence, such as GC content and Alu density. Therefore, we computed and analyzed the GC content and Alu percentage in the flanking 50 kb region (25 kb upstream + 25 kb downstream) with respect to various functional classes. Interestingly, regression analysis revealed that that GC content not only of the gene but of the downstream and upstream regions (in the order: $GC_{gene} > GC_{upstream} > GC_{downstream}$) also influenced Alu content of the gene (F ratio = 18.680, df = 3, 263, $P < 0.0001$). To identify whether any of the five variables (total GC content plus GC contents and Alu percentages in the 25 kb upstream and downstream regions) had any significant effect on Alu percentage, we carried out a stepwise regression analysis. The results showed that all five variables were statistically significant predictors of Alu percentage (F ratio = 47.197, df = 5, 266, $P < 0.0001$). We then regressed out the effects of these variables on Alu percentage and carried out ANOVA to test the equality of the adjusted mean values of Alu percentage among the functional categories. The F ratio (= 14.314, df = 4, 266) was highly significant ($P < 0.0001$), indicating that there are significant differences in Alu percentage among the functional categories even after adjusting for relevant correlates.

In the above exercise, Alu density was calculated by taking complete gene size (exons as well as introns) into account. Because Alu repeats are known to occur predominantly in introns, inclusion of exons for calculation of gene size may induce a bias in the analysis (particularly in the case of intron-less genes). To take this possibility into account, the analysis was repeated by calculating gene size as the sum of the sizes of its introns.

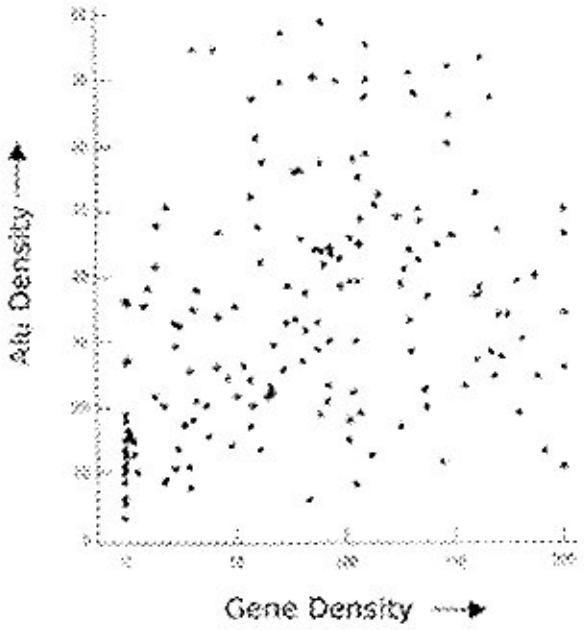

FIG. 1.—A scatter plot of Alu repeat density and gene density shown in a window size of 200 kb. The x-axis represents the cumulative length of genes (kb) in every 200 kb of sequence, and the y-axis represents the length occupied by Alu (kb) in the same 200-kb window.

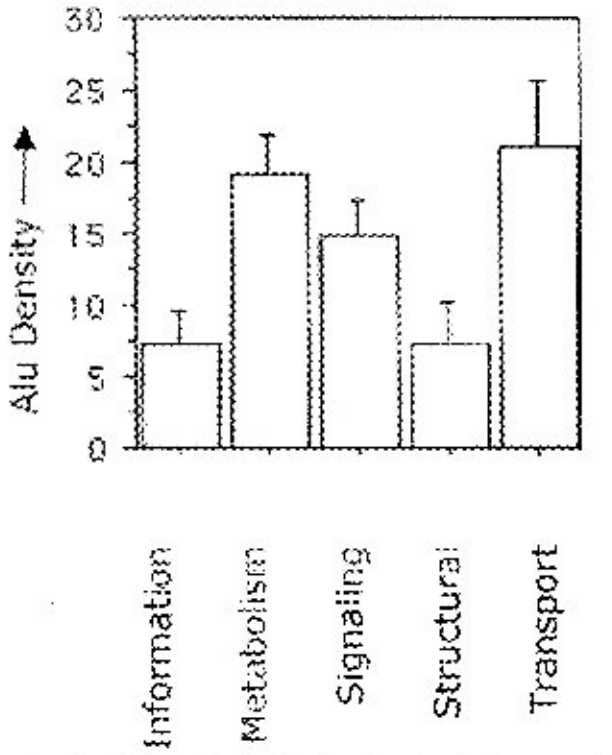

FIG. 2.—The mean Alu densities in various functional category in chromosomes 21 and 22 (error bars: 95% confidence interval).

Because regression analysis between Alu percentage and GC content of the intronic portions of genes revealed that GC content is not a statistically significant (F value = 0.274, df = 1, 262, $P > 0.6$) predictor of Alu percentage, ANOVA was carried out without regressing out the effect of GC content. Our results indicate that difference in mean Alu percentage values among different functional classes, even after excluding exons, is statistically significant (F value = 13.899, df = 4, 248, $P < 0.0001$).

We further determined whether there was a difference in the representation of the three Alu subfamilies in the different functional categories. For chromosome 22, there were no significant differences in the frequencies of S, J, and Y elements among the functional classes (chi-square = 6.28, df = 8, $P = 0.616$), but these differences were significant for chromosome 21 (chi-square = 22.7, df = 8, $P = 0.004$), which was also reflected in the pooled data (chi-square = 19.3, df = 8, $P = 0.013$). There was some difference in the frequencies of Alu S, J, and Y in the structural and information classes compared to the other two classes. When only three categories (signalling, transport, and metabolism) were considered, the chi-square value was not significant (chi-square = 6.19, df = 8, $P = 0.186$).

Discussion

Previous speculation about the predominance of Alu repeats in the actively transcribing regions of primate genomes (Schmid 1996) has recently been substantiated by analysis of the first draft of human genome (Lander et al. 2001). Higher Alu densities were observed in chromosomes with a greater number of genes and vice versa. These observations were made at the gross level of the chromosomes and indicated a general trend toward enrichment of Alus in gene-rich regions. In our analysis, scatter plots comparing Alu and gene densities indicate that although there is a positive correlation, this is not an all-or-none phenomenon. There are regions in the chromosome that are Alu rich but poorly represented in genes and vice versa. Another observation was that although chromosome 22 has more Alus and genes than chromosome 21, Alu density observed at the chromosomal level (Alu No.: Gene No.) is higher for chromosome 21 than for chromosome 22. This suggests that additional factors may govern Alu distribution in a chromosome and that gene density is not the only determinant of Alu density. Furthermore, it was observed that the genes on chromosome 22 are more Alu dense than those on chromosome 21 (table 1).

In an attempt to discern the properties that could influence Alu density in and around genes, we classified the genes from two chromosomes into five broad functional categories and then analyzed them with respect to Alu density. Surprisingly, we found a very biased distribution of Alu elements in these five functional categories. Alus were clustered in genes involved in metabolic pathways and signaling and transport processes, whereas they were poorly represented in genes coding for structural proteins and informational storage and processing components (fig. 2). Interestingly, the pattern of Alu distribution for each functional category was similar in the two chromosomes, despite a large difference in Alu and gene numbers between them.

Biased distribution of Alu in the human genome has been reported (Sainz et al. 1992) and ascribed to their

preference for GC-rich and gene-rich regions (Korenberg and Rykowski 1988; Pavlicek et al. 2001). It is possible that this bias is due to certain inherent differences in genomic architecture around the genes of various functional categories. However, we observed that although the GC content of the gene, as well as of the flanking sequence, influences Alu distribution considerably, it is the functional property of the gene which remains the dominant contributor toward Alu distribution as seen by ANOVA after regressing out the effect of GC content. This is in agreement with earlier observations that the distribution of young Alus in the human genome is not significantly influenced by GC content and transcriptional activity of the region (Arcot et al. 1995, 1996, 1998). In the earlier studies, it was concluded that the distribution of Alus was more or less random. We have demonstrated that this randomness is not observed if we classify the genes into various functional categories irrespective of GC content and Alu density of the surrounding genomic regions. Another suggested explanation for the nonrandom distribution is the abundance of sites that allow Alu insertion (Jurka, Klonowski, and Trifonov 1998) in certain genomic regions. If that were the case, one would observe a distribution of Alus that is a property of the genomic region, independent of genes and gene boundaries. However, we have observed that the bias in Alu distribution in genes was not influenced by Alu content of the flanking regions (see *Results*).

Based on these findings, we propose that Alus are nonrandomly distributed in the human genome and that the functional property of the gene seems to be the major factor contributing to the retention or exclusion of Alus within a gene. Given the increasing evidence of involvement of Alus in various regulatory functions (Oh et al. 2001; Hsieh et al. 2003; Le Goff et al. 2003), it is intuitively obvious that they might be negatively selected in structural genes as well as in the conserved information pathway genes. Because Alus are mostly present in the introns, it is also possible that absence of introns in the above categories could contribute to this bias. However, significant differences in Alu distribution across functional categories, even after excluding exonic sequences (thereby excluding genes without introns), further reinforced our hypothesis.

Our finding that the relative proportions of three Alu subfamilies are nearly same within the genes, but are significantly different outside the genes (chi-square test) indicates that there may be differential selection pressures operating on Alus within and outside genes. Furthermore, the relative proportions of these subfamilies for different functional categories were similar for chromosome 22 but somewhat different for chromosome 21, which was also reflected in the pooled data. In this analysis, two functional categories—informational and structural—were identified as outliers, and after removing these genes, the relative proportions became similar. This further corroborates our hypothesis of selection against insertion of these Alu elements in genes of structural and information functional classes. If Alus do play a role in gene regulation, it would be selectively disadvantageous—in fact cataclysmic—to have them in genes coding for structural proteins and information storage and processing components. This nonrandom distribution of Alu elements is in agreement with the analysis of the first draft of the human genome wherein homeobox gene clusters, which are extremely conserved across evolution, are found to be devoid of Alus or have low frequencies of them. The absence of these elements had been ascribed to the presence of large-scale cis-regulatory elements that cannot tolerate interruptions.

Alu elements harbor binding sites for various tissue-specific factors and hormone-responsive elements are involved in alternative splicing, can act as silencers as well as enhancers when present in 5′ untranslated regions (UTR) as well as 3′UTR, and also affect nucleosome positioning. Their role in differential gene regulation is exemplified by alternative splicing of the human epithelial sodium channel α gene (Oh et al. 2001) and the human β-amylase precursor protein, as well as by differential expression of genes like parathyroid hormone (PTH), the immunoglobulin E receptor, and the acetylcholine receptor (Hamdi et al. 2000) among many others. The higher physiological complexity in primates compared to lower organisms has been attributed to considerable amounts of change in the metabolic machinery as well as transport mechanisms (Hamdi et al. 2000). Therefore, it is possible that these elements may be positively selected in genes involved in metabolism, transport, and signaling processes because of a need for diverse regulatory functions in those genes. It is also possible that higher Alu density in regulated genes may result in a higher number of epigenotypes, as subtle epigenetic variations can be brought about by these elements in a number of ways. This hypothesis has been recently reinforced by the observation that SINEs are excluded from imprinted regions of human genome (Greally 2002). In this case, it has been proposed that methylation-induced silencing by these SINEs could lead to deleterious consequences in the imprinted loci, where inactivation of one allele is already established and expression is often essential for embryonic growth and survival (Greally 2002). The Alus could also contribute to the evolution of novel functions by serving to distribute functional and regulatable promoters (Ferrigno et al. 2001).

However, our study does not rule out the possibility of integration bias in genes of particular functional categories which could also lead to differences in Alu distribution. It has been reported in some studies that there are preferred sites of Alu integration in the genome (Daniels and Deininger 1985; Jurka and Klonowski 1996). Higher density of Alu repeats in genes of certain functional classes may therefore reflect the abundance of Alu integration sites in these genes. As more and more expression profiles become available, it will become possible to analyze the association of Alus with the function of genes.

In summary, our analysis of the Alu elements in chromosomes 21 and 22 clearly shows that there is a strong correlation between the functional class of the gene and Alu repeat maintenance. It remains to be seen whether this would be true for the entire genome.

Acknowledgments

We thank Debasis Dash for computational support and Adam Eyre-Walker for input into the manuscript. Fi-

nancial support was given to S.K.B. through the Department of Biotechnology (D.B.T.) of the Government of India Programme on Functional Genomics. The Council of Scientific and Industrial Research (CSIR) is duly acknowledged.

Literature Cited

Arcot, S. S., A. W. Adamson, J. E. Lamerdin, B. Kanagy, P. L. Deininger, A. V. Carrano, and M. A. Batzer. 1996. Alu fossil relics—distribution and insertion polymorphism. Genome Res. **6**:1084–1092.

Arcot, S. S., A. W. Adamson, G. W. Risch, J. LaFleur, M. B. Robichaux, J. E. Lamerdin, A. V. Carrano, and M. A. Batzer. 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. J. Mol. Biol. **281**:843–856.

Arcot, S. S., T. H. Shaikh, J. Kim, L. Bennett, M. Alegria-Hartman, D. O. Nelson, P. L. Deininger, and M. A. Batzer. 1995. Sequence diversity and chromosomal distribution of "young" Alu repeats. Gene **163**:273–278.

Babich, V., N. Aksenov, V. Alexeenko, S. L. Oei, G. Buchlow, and N. Tomilin. 1999. Association of some potential hormone response elements in human genes with the Alu family repeats. Gene **239**:341–349.

Britten, R. J., W. F. Baron, D. B. Stout, and E. H. Davidson. 1988. Sources and evolution of human Alu repeated sequences. Proc. Natl. Acad. Sci. USA **85**:4770–4774.

Chen, C., A. J. Gentles, J. Jurka, and S. Karlin. 2002. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. Proc. Natl. Acad. Sci. USA **99**: 2930–2935.

Daniels, G. R., and P. L. Deininger. 1985. Integration site preferences of the Alu family and similar repetitive DNA sequences. Nucleic Acids Res. **13**:8939–8954.

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

Dunham, I., N. Shimizu, B. A. Roe et al. (25 co-authors). 1999. The DNA sequence of human chromosome 22. Nature **402**: 489–495.

Englander, E. W., and B. H. Howard. 1995. Nucleosome positioning by human Alu elements in chromatin. J. Biol. Chem. **270**:10091–10096.

Englander, E. W., A. P. Wolffe, and B. H. Howard. 1993. Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. J. Biol. Chem. **268**:19565–19573.

Ferrigno, O., T. Virolle, Z. Djabari, J. P. Ortonne, R. J. White, and D. Aberdam. 2001. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. Nat. Genet. **28**:77–81.

Greally, J. M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc. Natl. Acad. Sci. USA **99**:327–332.

Hamdi, H. K., H. Nishio, J. Tavis, R. Zielinski, and A. Dugaiczyk. 2000. Alu-mediated phylogenetic novelties in gene regulation and development. J. Mol. Biol. **299**:931–939.

Hattori, M., A. Fujiyama, T. D. Taylor et al. (64 co-authors). 2000. The DNA sequence of human chromosome 21. Nature **405**:311–319.

Hsieh, S. Y., S. F. Liaw, S. N. Lee, P. S. Hsieh, K. H. Lin, C. M. Chu, and Y. F. Liaw. 2003. Aberrant caspase-activated DNase (CAD) transcripts in human hepatoma cells. Br. J. Cancer **88**:210–216.

Jurka, J., and P. Klonowski. 1996. Integration of retroposable elements in mammals: selection of target sites. J. Mol. Evol. **43**:685–689.

Jurka, J., P. Klonowski, and E. N. Trifonov. 1998. Mammalian retroposons integrate at kinkable DNA sites. J. Biomol. Struct. Dyn. **15**:717–721.

Jurka, J., and A. Milosavljevic. 1991. Reconstruction and analysis of human Alu genes. J. Mol. Evol. **32**:105–121.

Jurka, J., and T. Smith. 1988. A fundamental division in the Alu family of repeated sequences. Proc. Natl. Acad. Sci. USA **85**:4775–4778.

Kidwell, M. G., and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. USA **94**:7704–7711.

Korenberg, J. R., and M. C. Rykowski. 1988. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell **53**:391–400.

Labuda, D., and G. Striker. 1989. Sequence conservation in Alu evolution. Nucleic Acids Res. **17**:2477–2491.

Lander, E. S., L. M. Linton, B. Birren et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Le Goff, W., M. Guerin, M. J. Chapman, and J. Thillet. 2003. A CPF site and ALU repeat in the distal promoter region are implicated in regulation of human CETP gene expression. J. Lipid Res. **16**:16.

Li, L., T. Ohman, S. S. Deeb, and K. I. Fukuchi. 1999. Analysis of mouse intron 7 DNA sequence of the APP gene: comparison with the human homologue. DNA Seq. **10**: 219–228.

Norris, J., D. Fan, C. Aleman, J. R. Marks, P. A. Futreal, R. W. Wiseman, J. D. Iglehart, P. L. Deininger, and D. P. McDonnell. 1995. Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor–dependent transcriptional enhancers. J. Biol. Chem. **270**: 22777–22782.

Oh, Y. S., S. Lee, C. Won, and D. G. Warnock. 2001. An Alu cassette in the human epithelial sodium channel. Biochim. Biophys. Acta **1520**:94–98.

Pavlicek, A., K. Jabbari, J. Paces, V. Paces, J. V. Hejnar, and G. Bernardi. 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene **276**:39–45.

Sainz, J., L. Pevny, Y. Wu, C. R. Cantor, and C. L. Smith. 1992. Distribution of interspersed repeats (Alu and Kpn) on NotI restriction fragments of human chromosome 21. Proc. Natl. Acad. Sci. USA **89**:1080–1084.

Schmid, C. W. 1996. Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Prog. Nucleic Acid Res. Mol. Biol. **53**:283–319.

Whitelaw, E., and D. I. Martin. 2001. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. Nat. Genet. **27**:361–365.

Willard, C., H. T. Nguyen, and C. W. Schmid. 1987. Existence of at least three distinct Alu subfamilies. J. Mol. Evol. **26**:180–186.

Adam Eyre-Walker, Associate Editor

Accepted April 17, 2003