BIOINFORMATICS DISCOVERY NOTE

Vol. 20 no. 6 2004, pages 813–817
DOI: 10.1093/bioinformatics/bth005

Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition

Deepak Grover[1], Mitali Mukerji[1], Pankaj Bhatnagar[1], K. Kannan[2] and Samir K. Brahmachari[1,*]

[1]Functional Genomics Unit, Institute of Genomics and Integrative Biology (IGIB), CSIR, Mall Road, Delhi 110007, India and [2]School of Biotechnology, GGS Indraprastha University, Delhi 110 006, India

Received on August 31, 2003; accepted on October 31, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: Transposon-derived Alu repeats are exclusively associated with primate genomes. They have gained considerable importance in the recent times with evidence of their involvement in various aspects of gene regulation, e.g. alternative splicing, nucleosome positioning, CpG methylation, binding sites for transcription factors and hormone receptors, etc. The objective of this study is to investigate the factors that influence the distribution of Alu repeat elements in the human genome. Such analysis is expected to yield insights into various aspects of gene regulation in primates.

Results: Analysis of Alu repeat distribution for the human genome build 32 (released in January 2003) reveals that they occupy nearly one-tenth portion of the sequenced regions. Huge variations in Alu frequencies were seen across the genome with chromosome 19 being the most and chromosome Y being the least Alu dense chromosomes. The highlights of the analysis are as follows: (1) three-fourth of the total genes in the genome are associated with Alus. (2) Alu density is higher in genes as compared with intergenic regions in all the chromosomes except 19 and 22. (3) Alu density in human genome is highly correlated with GC content, gene density and intron density with GC content being major deterministic factor compared with other two. (4) Alu densities were correlated more with gene density than intron density indicating the insertion of Alus in untranslated regions of exons.

Contact: skb@igib.res.in

Supplementary information: Online supplementary data is available at the web page http://www.igib.res.in/manuscriptdata/aluanalysis.html

INTRODUCTION

Alu repeat elements belong to short interspersed nucleotide elements (SINE) family of repetetive sequences and are predominantly present in the non-coding regions of primate genomes. An Alu element consists of ∼282 bp conserved nucleotide sequence comprising of two tandem monomer units separated by a poly 'A' stretch. The monomers, homologous to 7SL RNA, are absolutely identical except for a 30 bp insertion in the right monomer (Jelinek *et al*., 1980; Ullu and Tschudi, 1984). The 3′ end of the Alu element has a long stretch of adenine residues, and is flanked by 4–10 bp of direct repeats at the site of insertion. These elements contain a pol III promoter that has been proposed to be involved in their retrotransposition mediated by RNA polymerase III (Deininger *et al*., 1992; Schmid and Maraia, 1992). These repeats are divided into different subfamilies according to their evolutionary age (Willard *et al*., 1987; Jurka and Smith, 1988; Britten *et al*., 1988; Labuda and Striker, 1989; Jurka and Milosavljevic, 1991). A majority of Alu repeats in human genome belong to old or intermediate subfamilies with relatively minimal representation of younger subfamilies (Schmid and Maraia, 1992; Deininger *et al*., 1992). However, the youngest Alus are considered to be biologically most active and most of the functions of Alus are attributed to this subfamily (Deininger *et al*., 1992). It has been shown earlier by cytogenetic studies that transcriptionally active regions of the genome (called R bands) were rich in Alu elements (Korenberg and Rykowski, 1988; Moyzis *et al*., 1989). Initial analysis of the first draft of human genome also revealed their association with gene and GC regions of the genome (Lander *et al*., 2001). It has also been reported that these elements are unevenly distributed and there is a positive correlation between exons and Alus in human chromosome 21 (Blinov *et al*., 2001). We had earlier shown that there is a difference in distribution of Alu repeats among genes of different functional categories in human chromosomes 21 and 22 (Grover *et al*., 2003). With the complete information about nucleotide sequence and genes now available publicly, we have attempted in this study to explore various factors that may drive the integration of Alu repeats in human genome.

*To whom correspondence should be addressed.

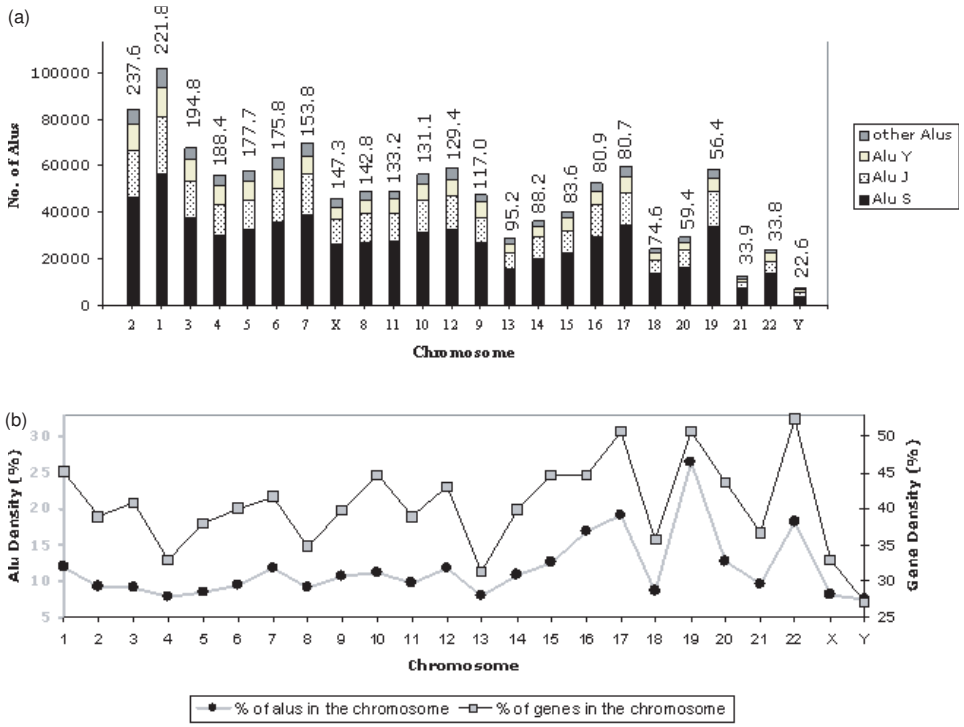

Fig. 1. (**a**) Number of Alu repeats in different chromosomes in human genome with vertical segments representing the numbers corresponding to each Alu subfamily. Chromosomes are arranged in descending order of their sizes (sequenced region) on *x*-axis with the size in Mb given at the top of each bar. (**b**) Variation in Alu and gene densities (percentage of region occupied in the chromosome) in human genome.

METHODS

The nucleotide sequence of human genome was downloaded from NCBI website (ftp://ftp.ncbi.nlm.nih.gov/genomes/h_sapiens; build no. 32). Information about genes, introns and exons was extracted from the available data using custom made PERL programs. Positions and subfamilies of Alu repeats in complete genome were identified using the program REPEATMASKER (http://repeatmasker.genome.washington.edu/) locally installed on compaq alpha sever ES40. Repeat numbers and density for each gene, intron, exon as well as chromosome was subsequently calculated using various PERL programs. Alu density in chromosomes and genes was expressed as Alu percentage, i.e. percentage of the gene/chromosomal region occupied by Alus. For correlation analysis, each chromosome was split into 100 kb intervals and Alu size, gene size, intron size and GC content was calculated individually for all these regions. Here, Alu, gene and intron sizes reflect the length occupied (in bp) by them in each 100 kb interval. Statistical analysis was performed using Statsview package (ver 4.0).

RESULTS

Alu distribution in whole genome

Alu repeats are present in 1 179 211 copies in the genome which together account for nearly 10.8% of the sequenced region of human genome (Supplementary material I). Interestingly, large sized chromosomes were not necessarily associated with more Alu repeats, as shown in Figure 1a where chromosomes have been arranged in descending order of lengths (bp). There are many chromosomes which have more Alus compared with other chromosomes with larger sizes. For example, chromosome 7 has more Alus than chromosomes 3, 4, 5 and 6; chromosome 12 has more Alu elements than chromosomes 8, 10, 11 and X; and chromosome 17 more than 8, 9, 10, 11, 12, 13, 14, 15, 16 and X. Despite all these variations, largest chromosome, i.e. chromosome 1 has highest number of Alus and chromosome Y, the smallest one, has least number of copies of these repeats. It is noteworthy, here, that although chromosome 1 is the largest human chromosome, chromosome 2 is larger than it in terms of region sequenced but contains less Alu repeats than the former.

Alu repeat density and association with genes

Alu densities were found to be highly variable among different chromosomes in human, with chromosome 19 being the most Alu rich chromosome having 26.3% of the region covered by these repeat elements. Other Alu dense chromosomes were chromosomes 17 (19.1%), 22 (18.2%) and 16 (16.8%). Chromosome Y, in addition to having the least number of Alu repeats, also has the least density (7.5%) closely followed

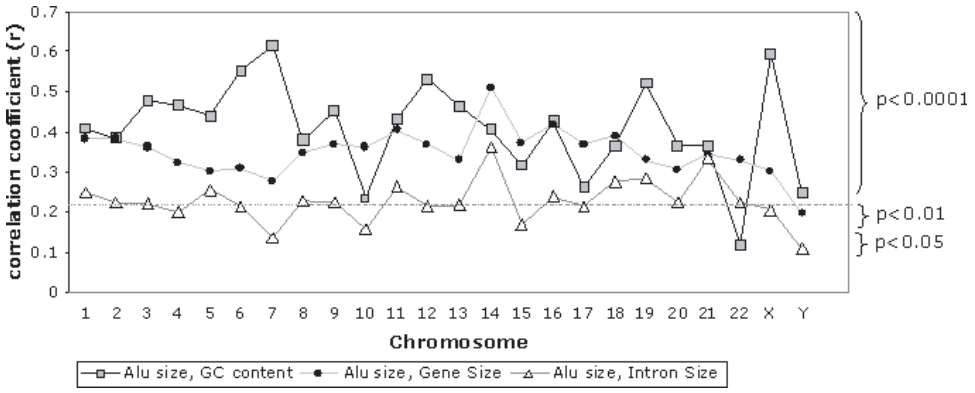

Fig. 2. Plot of correlation coefficients for various chromosomes in human genome. Correlation for Alu–intron, Alu–gene and Alu–GC is shown as separate curve and *P*-values are shown at right-hand side of the plot.

by chromosomes 4 (7.8%) and 13 (8.0%). Interestingly, the pattern of Alu densities across chromosomes was found highly similar to the pattern of gene densities (Fig. 1b). Statistical analysis revealed that Alu and gene densities are significantly correlated ($r = 0.83$, $P < 0.0001$). Chromosome 22 has the highest gene density in human genome (52.3%) and chromosome Y has least (27.1%).

Alus in intergenic and intragenic regions

Out of 27 963 genes so far identified in complete genome, Alu insertions were seen in nearly three-fourth of the genes. To analyze Alu distribution within genes, Alu percentage for each gene was calculated, which represents the Alu covered region within that gene. 8501 genes were found to have <10% Alu content, 6453 genes had Alus percentages ranging from 10 to 20%, 3534 genes had 20–30%, 1860 genes had 30–40% and 685 genes had >40% Alus whereas remaining 6930 genes were found without any Alu insertions. This clearly reflects that Alus are not uniformly distributed within genes (Supplementary Figure 4). A comparative analysis between intragenic and intergenic regions showed that Alu coverage is higher in genes (12.5%) than the intergenic regions (9.6%). The predominance of Alu densities in genes was observed in every chromosome except chromosomes 19 and 22 (Supplementary Figure 3). Within genes, Alus were mostly seen in intronic regions, occupying 12.8% of the intronic regions. As expected, they were rarely found in exons, with total Alu coverage of exons being 1.6%.

Distribution of Alu subfamilies

The most abundant Alu subfamily is the Alu S occupying 6.4% region of the genome and is followed by oldest Alu subfamily, Alu J (genomic coverage ∼2.5%). Members of younger Alu subfamily, known as Alu Y, are extremely less represented (genomic coverage ∼1.5%) compared with older subfamilies. A very small contribution (genomic coverage ∼0.4%) comes from Alu elements that do not belong to any of these subfamilies which include mainly truncated Alus and Alu monomers. At chromosomal level, there were significant differences in distribution of Alu repeat subfamilies ($\chi^2 = 2189.34$, d.f. $= 48$, $P < 0.0001$). Chromosome Y, the most Alu poor chromosome, was richer in members of Alu Y subfamily compared with other chromosomes whereas it has very low densities of Alu S and J, in fact, least density of Alu S in human genome. Similar trend was observed in chromosomes 13 and 9, with chromosome 13 having least density of Alu J subfamily (Supplementary material II). On the other hand, Chromosomes 8 and X were richer in Alu S and J subfamilies and contained very low densities of Alu Y, with chromosome X being the least Alu Y dense chromosome in the genome. Another interesting observation was that correlation of Alu J and Alu S with GC content was higher ($r = 0.38$, $P < 0.0001$) as compared with Alu Y ($r = 0.29$, $P < 0.0001$).

Correlation analysis

To elucidate the various factors that may affect the accumulation of Alu repeats in various regions of human genome, a correlation analysis was performed. It was found that Alu density is significantly positively correlated ($P < 0.0001$) with gene density, intron density and GC content (Fig. 2). Although the extent of correlation is different among different chromosomes, GC content seems to have highest association with Alu density overall, followed by gene density and intron density.

DISCUSSION

Nearly 45% portion of the human genome is occupied by transposon-derived repeat elements (Lander *et al.*, 2001). Earlier considered non-functional, they are now associated with various regulatory functions and are believed to interact with the whole genome to influence its evolution (Makalowski, 2000). To understand the mechanisms and means of such interactions, it is important to examine carefully

the positioning of these repeat elements. For this purpose, we have carried out extensive analysis of the distribution of Alu repeats, which are one of the most abundant and biologically important repeat elements in human. The analysis reveals many interesting features of these repeats which are as follows.

Alu repeats have a copy number of over one million in the human genome, which is much higher than previous estimates (Weiner, 2000; Deininger and Batzer, 1999). The frequency of their occurrence in the sequenced region, on average, comes out to be about one repeat element per 2.5 kb, which is also higher than previous estimates (Mighell *et al*., 1997). Moreover, we believe that the data reported here is an under-representation of the total Alu number as repeat rich regions are difficult to sequence completely. Alu repeats do not seem to be randomly spread in the genome, as indicated by unrelated Alu frequencies and chromosome sizes in many cases, particularly chromosomes 7, 12, 17 and 19. In case of random distribution, one would have expected a continuous decline in Alu numbers with decreasing chromosome sizes in Figure 1, but that is not the case.

The analysis showed that there is a clear difference between Alu densities across different chromosomes in human genome (Fig. 2). Chromosome 19 is the most Alu dense chromosome and chromosome Y has the least Alu density. Interestingly, Alu density variations were similar to gene density variations for all chromosomes, with correlation between two variables being highly significant. Alu occurrence was more frequent in intragenic regions as compared with intergenic regions ($P < 0.0001$). Thus, analysis of Alu distribution in genes elucidates two important points—(1) statistically significant correlation between Alu and gene densities across different human chromosomes indicates that gene density is a major driving factor for Alu accumulation within a chromosome and (2) a higher Alu density in intragenic regions indicates that these elements are preferred in genes.

Interestingly, the two chromosomes with highest Alu and gene densities in the human genome, namely 19 and 22, are the ones with higher intergenic Alu densities. It may be a consequence of gene underpredictions or other factors affecting Alu distribution and is a subject for future studies.

To explore the association of genes, introns and GC content with Alu repeats in genome, a correlation analysis was performed for equal sized intervals in human genome. Results of the analysis revealed that with minor differences from one chromosome to the other, overall Alu density is correlated with three variables in the order GC content > gene density > intron density. Higher correlation with genes than introns points towards the Alu insertions in untranslated region of exons.

The abundance of Alu subfamilies in human genome is in the order Alu S > Alu J > Alu Y, with significant variations in proportion of these subfamilies across chromosomes, e.g. chromosomes 9, 13 and Y contain more Alus belonging to young subfamilies in contrast to chromosomes 8 and X which are enriched in older Alus. These variations were not correlated with GC content, gene density or intron density. Higher correlation of older Alus with GC content than younger ones is in accordance with initial estimates (Lander *et al*., 2001). This indicates that Alus probably insert randomly in the genome followed by selection in specific regions that leads to non-random distribution finally.

To summarize, our analysis shows that there is significant association of Alu density in the human genome with gene density, intron density as well as GC content. However, a closer inspection shows that there are many regions inside the chromosomes where Alu distribution could not be explained by these factors. Non-uniform Alu distribution inside genes further strengthens this observation. Our previous work (Grover *et al*., 2003) as well as reports of other groups (Lander *et al*., 2001; Mighell *et al*., 1997) indicate that Alu distribution is influenced by specific functional properties of genomic regions. In future, extensive analysis of Alu repeats taking into account genomic composition as well as functional attributes would be helpful in elucidating the exact biological role for them.

REFERENCES

Blinov,V.M., Denisov,S.I., Saraev,D.V., Shvetsov,D.V., Uvarov,D.L., Oparina,N.Iu., Sandakhchiev,L.S. and Kiselev,L.L. (2001) Structural organization of the human genome: distribution of nucleotides, Alu-repeats and exons in chromosomes 21 and 22. *Mol. Biol. (Moskow)*, **6**, 1032–1038.

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

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

Deininger,P.L., Batzer,M.A., Hutchison,C.A., III and Edgell,M.H. (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet.*, **8**, 307–311.

Grover,D., Majumder,P.P., Rao,C.B., Brahmachari,S.K. and Mukerji,M. (2003) Nonrandom distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. *Mol. Biol. Evol.*, **20**, 1420–1424.

Jelinek,W.R., Toomey,T.P., Leinwand,L., Duncan,C.H., Biro,P.A., Choudary,P.V., Weissman,S.M., Rubin,C.M., Houck,C.M., Deininger,P.L. and Schmid,C.W. (1980) Ubiquitous, interspersed repeated sequences in mammalian genomes. *Proc. Natl Acad. Sci., USA*, **77**, 1398–1402.

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

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

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

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

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Devor,K., Doyle,M. and Fitzhugh,W. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Makalowski,W. (2000) Genomic scrap yard: how genomes utilize all that junk. *Gene*, **259**, 61–67.

Mighell,A.J., Markham,A.F. and Robinson,P.A. (1997) Alu sequences. *FEBS Lett.*, **417**, 1–5.

Moyzis,R.K., Torney,D.C., Meyne,J., Buckingham,J.M., Wu,J.R., Burks,C., Sirotkin,K.M. and Goad,W.B. (1989) The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics*, **4**, 273–289.

Schmid,C. and Maraia,R. (1992) Transcriptional regulation and transpositional selection of active SINE sequences. *Curr. Opin. Genet. Dev.*, **2**, 874–882.

Ullu,E. and Tschudi,C. (1984) Alu sequences are processed 7SL RNA genes. *Nature*, **312**, 171–172.

Weiner,A.M. (2000) Do all SINEs lead to LINEs? *Nat. Genet.*, **24**, 332–333.

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