Correlation between Exons and Dispersed Repetitive DNA Distribution on the Human Genome

Yoshiaki Hojo [12] hojo@jkk.hitachi.co.jp
Ikuo Uchiyama [2] uchiyama@ims.u-tokyo.ac.jp
Yataro Daigo [2] y-daigo@ims.u-tokyo.ac.jp
Yusuke Nakamura [2] yusuke@ims.u-tokyo.ac.jp
Toshihisa Takagi [2] takagi@ims.u-tokyo.ac.jp

[1] Hitachi, Ltd. Information Systems Group. Shinsuna Plaza 6-27, Shinsuna 1-Chome, Koto-ku, Tokyo 136, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

The human nuclear genome contains a large number of highly repeated DNA sequences. The *Alu* sequences are primate specific and are the most abundant family of repeated DNA sequences in the human genome. The human *Alu* sequence is approximate 300 bp long [2]. The *L1* sequence is a long interspersed nuclear element. *L1* is found in other mammals. Although their functions are not yet clear [4], some of them may affect gene functions or cause human diseases [3].

We have identified repeated DNA sequences from human genomic sequences in the region of 3p21.3-p22 and 9q32, both of which are more than 1M bp long. Our statistical analysis shows that the distributions of *Alu*s and exons have a weak positive correlation and those of *L1*s and exons have a weak negative correlation.

2 Method

Genomic sequence data of the human chromosome 3p21.3-p22 and 9q32 as well as cDNA sequences on these regions were obtained by Y.Daigo *et al.* (unpublished data). The lengths of sequences on the chromosome 3p21.3-p22 and 9q32 are 1.2M bp and 1.0M bp respectively. While the region 3p21.3-p22 contains 14 genes, the region 9q32 has only 3 genes.

Repetitive sequences were identified by the computer program CENSOR [1] with Repbase(Release 5.0). We divided each sequence into non-overlapping 100k bp segment and counted the exon, *Alu* and *L1*.

3 Result & Discussion

To characterize the exon, *Alu* and *L1* distributions, we compared their densities. The *Alu*s and exons were more abundant in the 3p21.3-p22 region whereas the *L1*s were more abundant in the 9q32 region. To test the significance of these differences, we applied a statistical analysis technique known as two-sample paired t-test. The significant difference ($p<0.05$) between the regions was observed in exon and *Alu* but not in *L1* (Table 1).

Next, densities of exon, *Alu* and *L1* in each 100k bp segment were plotted in Fig. 1. Due to extremely low gene content, no or little tendency was observed in the region 9q32. In 3p21.3-p22, positive correlation against exon density was observed for *Alu*($r = 0.42$) and negative correlation against exon density was observed for *L1* ($r = -0.42$). When the data in both regions were taken into account, the correlation coefficient between exon and *Alu* densities became 0.62 while that between exon and *L1* densities was unchanged ($r = -0.42$). However, when the segment size was enlarged up to 170k bp, the latter correlation coefficient also became a large negative value -0.7 (data not shown).

These results suggest that *Alu* elements but not *L1*s have a tendency to cluster into regions where the gene density is high. Although existence of direct relationships between exons and repetitive elements is not clear yet, the observed correlations might have an influence on gene function or gene expressions.

Table1:Comparison of exon, *Alu* and *L1* density

	exon	Alu	L1
3p21-22(1/k bp)	0.15	0.34	0.12
9q32(1/k bp)	0.02	0.16	0.16
Std.	0.12	0.13	0.11
t	3.28	4.78	-0.89
p	1.7×10^{-3}	4.5×10^{-5}	0.20

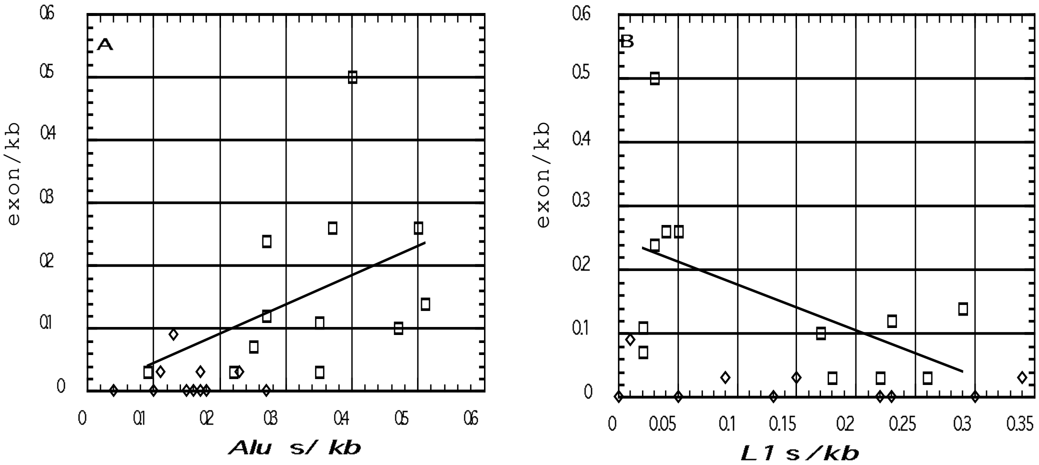

Figure 1: Correlation between exon density and *Alu* or *L1*
A, Correlation between *Alu* density and Exon density. Correlation coefficient is 0.62. B, Correlation between *L1* density and exon density. Correlation coefficient is -0.42. Squares are 3p21.3-p22. Diamonds are in 9q32.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Jurka, J., Klonowski, P., Dagman, V., Pelton, P., CENSOR - a program for identification and elimination of repetitive element from DNA sequence, *Comp. Chem.*, 20(1):119–122, 1996.

[2] Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S., Matsubara, K., Revision of consensus sequence of human Alu repeat – a review, *Gene*, 53:1–10, 1987.

[3] Mighell, A.J., Markham, A.F., Robinson, P.A., *Alu* sequences, *FEBS Letters.*, 417:1–5, 1997.

[4] Schmid, C.W., Dose SINE evolution preclude *Alu* function, *Nucl. Acids Res.*, 26(20):4541–4550, 1998.