Letter to the Editor

A Test of Whether Selection Maintains Isochores Using Sites Polymorphic for *Alu* and *L1* Element Insertions

Elise M. S. Belle and Adam Eyre-Walker[1]

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom

Manuscript received June 6, 2001
Accepted for publication October 22, 2001

ONE of the most striking features of the human genome is the large-scale variation in GC content that is found along chromosomes. This variation was discovered in 1973 (Filipski *et al.* 1973) and was originally thought to be organized as a series of "isochores," blocks of DNA of homogeneous base composition separated by borders of sharp transition (Bernardi 1995). While more recent studies of long sequences (Nekrutenko and Li 2000) and the draft human genome (International Human Genome Sequencing Consortium 2001) have suggested that the mammalian genome probably does not fit the isochore model exactly, it is still evident that base composition varies over scales as large as 10 Mb.

The origins of the large-scale variation in GC content remain controversial (Eyre-Walker and Hurst 2001). It has been suggested that it could be a consequence of (i) mutation bias (Filipski 1987; Suoeka 1988; Wolfe *et al.* 1989), (ii) natural selection (Bernardi and Bernardi 1986; Hughes and Yeager 1997; Eyre-Walker 1999), or (iii) biased gene conversion (Holmquist 1992; Eyre-Walker 1993). In this study we test whether natural selection is acting upon base composition by considering the distribution of SINE and LINE elements.

If selection is acting upon the base composition of isochores, then the distribution of newly integrated elements, and elements that are fixed within the population, should be different. For example, if selection favors high GC content in certain regions of the genome, then one might expect the GC content flanking newly inserted *L1* elements, which are still polymorphic in the population, to be greater than the GC content flanking young *L1* elements that are fixed within the species. This is because *L1* elements are GC poor; they will therefore be selected against if they integrate into the GC-rich parts of the genome.

[1]*Corresponding author:* School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom.
E-mail: a.c.eyre-walker@sussex.ac.uk

Although we have little direct information about the pattern of element integration, we can use sites that are polymorphic for SINE and LINE insertions to investigate whether selection on composition affects the distribution of elements, since sites that are polymorphic for element insertions will reflect the pattern of integration more closely than sites at which the element is fixed. We therefore tested whether the sequences flanking sites polymorphic for *Alu* and *L1* elements are different in composition from the sequences flanking elements fixed within the human population. Since the distribution of fixed elements may also be affected by the rate at which elements decay, we have restricted our comparisons between fixed and polymorphic elements to elements of the same subfamilies.

Mark Batzer, Prescott Deininger, and colleagues have investigated whether a number of *Alu* insertions from young *Alu* subfamilies are polymorphic in humans (Batzer *et al.* 1995, 1996; Arcot *et al.* 1996, 1998; Sherry *et al.* 1997; Roy *et al.* 1999; Roy-Engel *et al.* 2001). Using their primer sequences and a BLAST search of the draft human genome, we located the genomic position of 240 of their elements: 86 Ya5, 12 Ya8, 11 Yb8, 34 Yb9, 90 Yc1, and 7 Yc2; 76 of these sites were polymorphic for the element. Similarly we used the primer sequences or accession numbers given by Boissinot *et al.* (2000) to locate 41 young Ta *L1* repeats: 14 Ta-0 and 27 Ta-1; 18 of these sites were polymorphic. All the sites polymorphic for an element appear to be a consequence of insertion rather than deletion. From these data we extracted the 500 and 5000 bp each side of the repeat, or the site midway between the two primers, if the element was missing from the genomic sequence. All the data are published as supplementary information. For comparison we also extracted 200 randomly chosen sequences of 1 and 10 kb from the human genome.

Figure 1 shows the distribution of the GC content of the 10 kb of sequence flanking sites polymorphic or fixed for *Alu* or *L1* insertion, along with the distribution of randomly selected sequences of the same length;

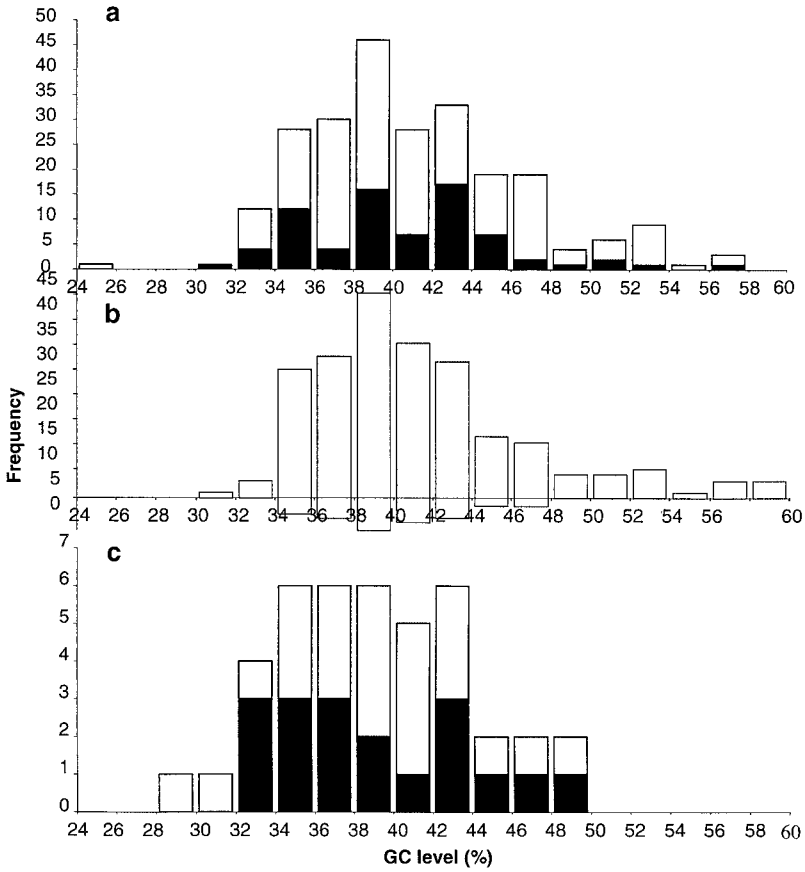

FIGURE 1.—The GC content distribution of the 10-kb flanking polymorphic (solid bars) and fixed (open bars) *Alu* (a) and *L1* (c) elements. (b) The GC content distribution for 200 randomly selected human sequences of 10 kb.

similar results were obtained with 1-kb flanking sequences. There is no significant difference between the mean GC content of the sequences flanking the polymorphic and fixed elements, for either *Alu* or *L1* (Table 1), which suggests that compositional selection is not affecting the distribution of either element. However, such an analysis could mask the action of selection, if selection is stabilizing and similar numbers of elements integrate into regions that are higher and lower than the GC content of the element; for example, if *L1* elements,

TABLE 1

Mean (and standard error) of GC content flanking fixed and polymorphic *Alu* and *L1* elements

	Fixed		Polymorphic	
Element	*n*	Mean (SE)	*n*	Mean (SE)
Alu Ya5	51	0.410 (0.008)	35	0.409 (0.009)
Alu Ya8	7	0.398 (0.014)	5	0.431 (0.024)
Alu Yb8	9	0.476 (0.014)	2	0.410 (0.049)
Alu Yb9	25	0.392 (0.011)	9	0.414 (0.014)
Alu Yc1	68	0.416 (0.006)	22	0.388 (0.008)
Alu Yc2	4	0.418 (0.033)	3	0.434 (0.004)
Alu all	164	0.413 (0.004)	76	0.406 (0.006)
Alu (freq. <0.2)	164	0.413 (0.004)	19	0.411 (0.010)
Alu $(x - y)^2$	164	0.040 (0.002)	76	0.037 (0.002)
Alu $(x - y)^2$ (freq. <0.2)	164	0.040 (0.002)	19	0.042 (0.005)
L1 Ta-0	12	0.382 (0.013)	2	0.385 (0.010)
L1 Ta-1	11	0.385 (0.015)	16	0.395 (0.012)
L1 all	23	0.383 (0.010)	18	0.394 (0.011)
L1 (freq. <0.2)	23	0.383 (0.010)	5	0.396 (0.025)
L1 $(x - y)^2$	23	0.0025 (0.0005)	18	0.0030 (0.0008)
L1 $(x - y)^2$ (freq. <0.2)	23	0.0025 (0.0005)	5	0.0038 (0.0019)

which in our sample have an average GC content of 39%, integrate at roughly the same frequency into DNA, which is either <39% or >39%, then selection will manifest itself as a difference in the variance of the GC content of sequences flanking sites with polymorphic and fixed elements. To investigate this we tested whether there was a significant difference in the statistic $(x - y)^2$ between sites with fixed and polymorphic elements, where x is the GC content flanking an element and y is GC content of the element itself. For *Alu* elements, we took the value of y from Gu (2000), since all the *Alu* sequences were of a similar length and sequence, but for *L1* elements we calculated the composition of each element individually. We tested for a difference using a Mann-Whitney test. No significant difference was found for either the *L1* or the *Alu* elements (Table 1).

Unfortunately, the methods that were used to detect sites polymorphic for *Alu* and *L1* insertions were biased toward the detection of common variants; the average frequency of the polymorphic *Alu* elements is 0.471, and for *L1* elements it is 0.394; both of these are high above neutral expectations (0.275 if 12 chromosomes were sampled, as there were for some of the *L1* elements, and 0.052 if 100 chromosomes were sampled, as there were for some *Alu* elements). The bias toward common variants will reduce the power of our test. We therefore tested whether the GC content flanking fixed elements was different from that flanking elements that were segregating at <20%. In none of the cases was there a significant difference (Table 1).

There is therefore no evidence that compositional selection acts upon the distribution of repetitive DNA in the human genome and, hence, no evidence that selection maintains isochores. However, the power of the analysis may be limited since many of the polymorphic elements are segregating at high frequencies, and the differences between the GC contents of the elements and the DNA into which they integrate is fairly small.

We thank two referees for comments and the Royal Society for funding.

LITERATURE CITED

Arcot, S. S., A. W. Adamson, J. E. Lamerdin, B. Kanagy, P. L. Deininger *et al.*, 1996 Alu fossil relics - distribution and insertion polymorphism. Genome Research **6:** 1084–1092.

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

Batzer, M. A., C. M. Rubin, U. Hellman-Blumberg, M. Alegria-Hartman, E. P. Leeflang *et al.*, 1995 Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J. Mol. Evol. **247:** 418–426.

Batzer, M. A., S. S. Arcot, J. W. Phinney, M. Alegria-Hartman, D. H. Kass *et al.*, 1996 Genetic variation of recent Alu insertions in human populations. J. Mol. Evol. **42:** 22–29.

Bernardi, G., 1995 The human genome: organization and evolutionary history. Annu. Rev. Genet. **29:** 445–476.

Bernardi, G., and G. Bernardi, 1986 Compositional constraints and genome evolution. J. Mol. Evol. **24:** 1–11.

Boissinot, S., P. Chevret and A. V. Furano, 2000 L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. **17:** 915–928.

Eyre-Walker, A., 1993 Recombination and mammalian genome evolution. Proc. R. Soc. Lond. Ser. B **252:** 237–243.

Eyre-Walker, A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 675–683.

Eyre-Walker, A., and L. D. Hurst, 2001 The evolution of isochores. Nat. Rev. Genet. **2:** 549–555.

Filipski, J., 1987 Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **217:** 184–186.

Filipski, J., J. P. Thiery and G. Bernardi, 1973 An analysis of the bovine genome by Cs2SO4-Ag+ density centrifugation. J. Mol. Biol. **80:** 177–197.

Gu, Z., H. Wang, A. Nekrujenko and W-H. Li, 2000 Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequences. Gene **259:** 81–88.

Holmquist, G. P., 1992 Chromosome bands, their chromatin flavors and their functional features. Am. J. Hum. Genet. **51:** 17–37.

Hughes, A. L., and M. Yeager, 1997 Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. **45:** 125–130.

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

Nekrutenko, A., and W.-H. Li, 2000 Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. **10:** 1986–1995.

Roy, A. M., M. L. Carroll, D. H. Kass, S. V. Nguyen, A.-H. Salem *et al.*, 1999 Recently integrated human Alu repeats: Finding needles in the haystack. Genetica **107:** 149–161.

Roy-Engel, A. M., M. L. Carroll, E. Vogel, R. K. Garber, S. V. Nguyen *et al.*, 2001 Alu insertion polymorphisms for the study of human genomic diversity. Genetics **159:** 279–290.

Sherry, S. T., H. C. Harpending, M. A. Batzer and M. Stoneking, 1997 Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics **147:** 1977–1982.

Suoeka, N., 1988 Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 2653–2657.

Wolfe, K. H., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Communicating editor: T. F. C. Mackay