

Evolutionary selection against change in many *Alu* repeat sequences interspersed through primate genomes

(insertion/sequence conservation/gene regulation)

ROY J. BRITTEN

Division of Biology of the California Institute of Technology, Kerckhoff Marine Laboratory, 101 Dahlia Avenue, Corona Del Mar, CA 92625

Contributed by Roy J. Britten, February 18, 1994

ABSTRACT Mutations have been examined in the 1500 interspersed *Alu* repeats of human DNA that have been sequenced and are nearly full length. There is a set of particular changes at certain positions that rarely occur (termed suppressed changes) compared to the average of identical changes of identical nucleotides in the rest of the sequence. The suppressed changes occur in positions that are clustered together in what appear to be sites for protein binding. There is a good correlation of the suppression in different positions, and therefore the joint probability of absence of mutation at many pairs of such positions is significantly higher than that expected at random. The suppression of mutation appears to result from selection that is not due to requirements for *Alu* sequence replication. The implication is that hundreds of thousands of *Alu* sequences have sequence-dependent functions in the genome that are selectively important for primates. In a few known cases *Alu* inserts have been adapted to function in the regulation of gene transcription.

If *Alu* sequences have some positive significance (1), then the *Alu* inserts in primate genomes would have been under selection to preserve the valuable segments of their sequences. The *Alu* sequence is about 281 nucleotides long, is very high in G+C content, and usually has a poly(A) tail. The sequences are established by two principal processes: evolution of the source genes (2) that give rise to the sequences and mutation of the inserts *in situ* after insertion. The special features of the source genes that cause insertion of hundreds of thousands of copies are not known, but the source genes must be transcribed, and the mechanism of insertion is probably retroposition (3). The evolutionary changes in the source genes can be identified because large families of *Alu* repeats share diagnostic nucleotides at certain positions (2, 4). The relationships of the families of *Alu* sequences have been reexamined (5), including those judged to be recently inserted (6–9). It appears that several source genes are active at present or have been in the recent past, giving rise to several types of inserted *Alu* repeats matching each of the sources in sequence (7, 9, 10).

A central aspect of the *Alu* sequences is that almost all of the source gene sequence has been conserved through the history of the *Alu* sequences. The diagnostic positions are exceptions, since the nucleotides at these positions changed at some time in the past, giving rise to variant source genes. Some of the variants became predominant new sources of inserted copies, the changed nucleotides were maintained for extended periods of time, and the many copies formed recognizable families of *Alu* sequences. It is significant that the fully conserved positions of the source genes include most of the CpGs, even though these have changed rapidly after the copies were inserted. As a result of the conserva-

tion, there is a large group of positions for which we almost certainly know what the nucleotide was at the time of insertion. The existence of such positions has been previously shown (2), and the principal set of them (not including the CpGs, the diagnostic positions, and a few other positions) is examined in this work. Evidence is presented for their conservation. The 195 chosen positions are termed the CONSBI (conserved before insertion) positions and are shown as uppercase letters in Fig. 1.

Systematic and Stochastic Changes Identified by Comparing Two Randomly Chosen Sets of *Alu* Sequences

It is possible to recognize systematic processes as opposed to stochastic events that affect the nucleotides at specific positions by examining the relationship of mutations between subsets of the known sequences. A set of nearly full-length *Alu* sequences was divided into two equal randomly chosen sets (789 each), and all of the members of each set were compared with the consensus of recently inserted copies (Fig. 1). The divergences at each position were summed, and Fig. 2 is a graph of the fractional divergence of each position in one set plotted against the fractional divergence of the same position in the other set. The correlation of the points along the diagonal is expected, as many of these are diagnostic positions, and the differences from the modern consensus are shared by classes of *Alu* sequences. The highest points are positions that have recently changed to create the presently active source genes, and therefore almost all *Alu* inserts differ from the modern consensus sequence at these positions. The cluster of points at about 35–40% is mostly the C and G residues of CpGs, which evolve rapidly after insertion (2, 4, 15, 16).

The 195 points at the lower end of the diagonal (below 14%) are the positions that were CONSBI. This cluster of points exhibits correlation; that is, it is spread out along the diagonal. Part of the reason is that the C and G residues (that are not part of CpGs) have evolved about 1.5 times faster than the A and T residues, as shown on the bottom line of Table 1. The more rapid evolution of C and G residues is not specific for the *Alu* sequences, as shown by examining an alignment of ape and monkey sequences of the noncoding parts of the gamma globulin gene region (17) and some pseudogenes (18). The average substitution at G plus C residues was about 1.5 times the average substitution at A plus T residues, which agrees well with the ratio for the *Alu* sequences. Most of the spread of the lower points along the diagonal is due to the position-specific differences in mutation rate described in the next section.

Position-Specific Restriction on Observed Mutations

There are large differences between different positions in the most probable types of mutations that occur after insertion as

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CONSBI, conserved before insertion; pol, RNA polymerase.

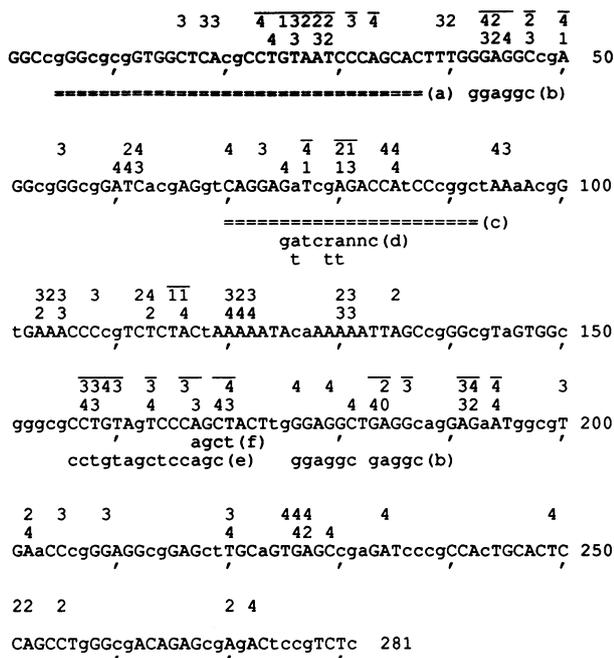


FIG. 1. Suppressed mutations in *Alu* insert sequences. The sequence is the consensus of a group of recent *Alu* inserts (6–9). The prime (') marks decades in the sequence. The CONSBI positions are in uppercase letters. The evidence for positions with suppressed mutations is discussed in the text. The upper line of numbers shows codes representing exceptional bias against certain types of changes: 1, observed changes (oc) <20% of expected mutations (em); 2, oc 20–30% of em; 3, oc 30–40% of em; 4, oc 40–50% of em. The lower line of numbers shows codes representing low total number of mutations: 0, oc <50% of em; 1, oc 50–57% of em; 2, oc 57–64% of em; 3, oc 64–71% of em; 4, oc 71–78% of em. The probability of occurrence of any of these degrees of suppression by chance is very low. Equivalent Monte Carlo models using Table 1 probabilities show only two or three "4s" in the whole sequence and no more-biased scores. Regions that appear clustered are overlined as are positions with a score of 1 or 0. The letters in parentheses mark regions that have been proposed in the literature as significant parts of the *Alu* sequence, possibly related to function. (a), enhancing element of "pol III promoter" (11); (b), "core sequence" of "pol III promoter" (12); (c), "directing element of pol III promoter" (11); (d), "consensus B box" of "pol III promoter" (13); (e), reducer element (end region) affecting pol II transcription (14); (f), *Alu* I site.

well as in the rates of change. Table 2 shows the distribution of the 780 (195 × 4) possible position-specific mutations of each type, expressed as the ratio of the number of occurrences to the average number of changes of that type. A small part of the spread of this distribution is of stochastic origin, and that is shown in the second column, which is the result of a Monte Carlo calculation. For this calculation, pseudo-random numbers were used to pick mutations in each of the positions of the modern consensus sequence making use of Table 1 to set the chance for each type of change (or no change). This was repeated 1750 times for each position, and the ratio of number of changes to expected number was calculated as shown in column 2 of Table 2. The probability that the distribution in column 1 occurred as a result of chance fluctuation is vanishingly small. There are 83 position-specific types of change that occur at less than half the expected (average) rate, while the Monte Carlo calculation indicates that only about three of these are due to chance.

The Pattern of Occurrence of Position-Specific Bias

The upper line of numbers in Fig. 1 shows the positions with unexpectedly low mutations of particular types. The degree

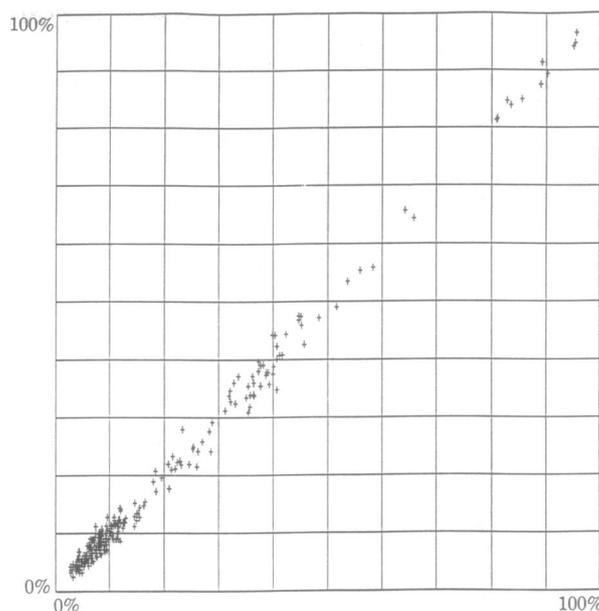


FIG. 2. Recognition of systematic as contrasted to stochastic processes in *Alu* sequence evolution. The known nearly full-length *Alu* sequences were randomly subdivided into two sets of 789 each. Each of the 281 positions was individually compared with the modern consensus sequence shown in Fig. 1, and the percentage of sequences that differed from it was calculated. The percent difference from the modern consensus in one set is graphed against that for the other set.

of suppression has been coded, and code 1 represents the most severe suppression. In addition to the suppression of specific types of change, many positions show reduced amounts of mutations of all kinds. The reduced total amount of mutation has also been encoded and is shown in the lower line of numbers above the sequence in Fig. 1, with code 0 representing the sole position with less than half the expected number of substitutions or deletions, based on the average shown in Table 1. The codes are described in the legend to Fig. 1.

The clustering of the positions showing suppression indicates that the reduced amounts of nucleotide change are significant and suggests that the clustered positions are parts of sites associated with potential function, such as "pol III promoters." There is no evidence that more than a small fraction of *Alu* sequences is transcribed. Thus the sequences in typical inserts, though homologous in sequence to true *Alu* "pol III promoters," may not be active promoters. For this

Table 1. Average occurrence in all CONSBI positions of specific substitutions

Nucleotide in <i>Alu</i> insert	Nucleotide in modern consensus			
	A	C	G	T
A	0.9431	0.0173	0.0539	0.0164
C	0.0099	0.9072	0.0133	0.0258
G	0.0237	0.0152	0.9108	0.0122
T	0.0121	0.0503	0.0123	0.9374
Del	0.0111	0.0099	0.0097	0.0081
Sum*	0.0568	0.0928	0.0892	0.0652

The values given are the expectations of replacement of the nucleotide in the modern consensus with the nucleotide in the *Alu* insert, on average. The data are for the 195 CONSBI positions from a set of ≈1750 full-length *Alu* sequences. Even the value for T to G transversions is based on 475 events. Del, deletion.

*Fraction of nucleotides in the inserts that were different from the nucleotide in the modern consensus.

Table 2. Occurrence of specific mutations at particular CONSBI positions of 1750 *Alu* inserts

Occurrence*	MC†	Ratio‡	Code(s)§
4	0	0.1775	1
1	0	0.1900	2
2	0	0.2150	2
6	0	0.2317	2
7	0	0.2643	2
5	0	0.2920	2 & 3
3	0	0.3233	3
7	0	0.3457	3
19	1	0.3805	3 & 4
8	1	0.4213	4
21	1	0.4733	
28	5	0.5214	
28	14	0.5779	
34	23	0.6441	
38	20	0.7074	
52	56	0.7706	
61	109	0.8597	
99	162	0.9638	
158	166	1.0291	
52	128	1.1600	
38	52	1.2837	
38	30	1.4179	
16	9	1.5681	
12	3	1.7392	
16	0	1.9231	
8	0	2.1287	
2	0	2.3250	
3	0	2.5600	
3	0	2.7733	
1	0	3.0800	
1	0	3.6100	
4	0	3.8575	
1	0	4.2400	
2	0	4.7400	
0	0	—	
2	0	5.8050	

To prepare this table, all of the mutations (differences from the modern consensus sequence) that occurred at each position were counted for the 1750 *Alu* sequences in the set. The ratio of the fraction for each type of mutation to the expected fraction from Table 1 was calculated. These ratios were placed in order, and the numbers in each of the classes of ratios were counted and are listed in column 1. For example, the four minimally occurring cases at the head of the column are position G to T at position 25, G deletion at position 81, T to G at position 115, and A to T at position 116. The average ratio for these four was 17.75% of expected as listed in column 3.

*The distribution of the ratio of occurrence to expectation for the four possible kinds of mutations at 195 CONSBI positions, totaling 780 entries.

†The distribution of the ratio of occurrence to average expectation for a Monte Carlo calculation (see text) as a measure of the stochastic spread of the distribution.

‡The average ratio of occurrence to expectation for the set of cases in column 1. For example, on the fourth line are six types of changes in specific positions that average about 23% of the expected occurrence.

§This column lists the code used in Table 3 to locate the positions with suppressed types of mutations. Code 1 represents the first four while code 2 represents the next 19, which average about 24% of the expected occurrence.

reason the phrase “pol III promoter” is placed in quotation marks where it refers to *Alu* sequences.

There have been several attempts to determine the most important sequences of *Alu* repeat sequence “pol III promoters” (11–13) with different methods of assay. Starting at position 23, 9 out of 11 positions are marked, and this falls in the center of a sequence that is claimed to be part of the “pol

III promoter” (11). This sequence was identified as the “enhancing element,” which increased transcription by 30-fold and is labeled (a) in Fig. 1. As shown in Fig. 1, the sequence labeled (e) has been identified as part of a “reducer element,” which means that it is involved in the reduction of pol II transcription, presumably by causing interference through increased pol III transcription (14). One short marked sequence starting at nucleotide 43 falls within a GGAGGC sequence, which Saegusa *et al.* (12) identify as the *Alu* core “required for RNA polymerase III promoter function.” This same region is also an active binding site for the LyF-1 transcription factor (19) with a consensus sequence of YYTGGGAGR.

Relationship of Different Positions Within a Site

One characteristic of a functional site is that a group of positions might act together and so be conserved together in a subset of functional *Alu* sequences. The result would be a correlation of the absence of changes between different positions. This section briefly describes the detection of such a correlation (details will be published elsewhere). In the analysis, the probability that a site is not mutated is written as Z_n , which is just the fraction of *Alu* sequences that match the modern consensus in that position. For a different site, it is Z_m . If the changes at the two sites are the result of independent events, the expectation for the number of pairs, of which neither is mutated, is $E_{nm} = Z_n Z_m L$, where L is the total number of *Alu* sequences examined at these positions. The observed number of cases in which both positions in a pair are unchanged is written as O_{mn} . $D(m, n) = O_{mn} - E_{mn}$ is the excess in the observed number of *Alu* sequences not changed in either of the two positions over that expected by chance and is a convenient measure of the correlation. The correlation between the 195 CONSBI positions was examined, and the results for one region are shown in Table 3, which is a small part of a 195×195 matrix. Listed in Table 3 is $D(m, n)$ for pairs where m is the position number given in the first column. The heading of each column is the distance between the two compared positions ($n - m$). For example, the number 16 listed at the intersection of the row numbered 35 and the column headed +1 equals $D(35, 36)$. In other words, the C at 35 and the A at 36 were both not mutated in 16 more *Alu* sequences than expected. This example is part of a cluster that stretches from positions 35 to 43, suggesting that the sequence CACTTTG is coordinately conserved in a subset of *Alu* sequences.

The $D(m, n)$ numbers, though small, are significantly larger than random fluctuations as the following controls show. For comparison, sequence positions out to 56 were included in Table 3 as a control, since they are positions that are not strongly correlated. $D(m, n)$ numbers in this region average 1.3. The larger value, 11 at (55, 56), is probably an actual correlation. This control region has no sets of larger signals such as those between positions 26 and 43, and thus there is a high probability that the stronger correlations between positions 26 and 43 are genuine. In an equivalent Monte Carlo run, 1500 sequences were generated that were randomly mutated from the modern consensus using the observed rate at each position. The resulting set of $D(m, n)$ varied primarily between +4 and -4, with many zeros.

Deletions do not occur at random locations in the *Alu* sequences, and thus deletions of several contiguous nucleotides artificially create correlations. To correct for this, the deleted regions were each replaced with a homologous piece of sequence taken from an *Alu* sequence of the same amount of divergence from the modern consensus. The effect of this correction was to remove many artificial correlations and reduce the background so that the clusters of significant correlations were more evident.

Table 3. Relationship of mutations among neighboring positions

Pos*	Mut†	nt	Position of second base‡									
			+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
22	129	C	9	4	5	4	1	2	-2	7	6	-2
23	135	C	5	2	10	7	5	4	3	5	0	3
24	73	T	2	1	6	1	0	4	10	-1	2	3
25	169	G	3	0	2	1	2	2	5	0	2	5
26	66	T	3	7	5	1	2	-1	3	0	2	3
27	97	A	7	8	4	8	1	8	4	4	4	3
28	61	A	9	3	9	3	7	0	2	8	9	10
29	63	T	7	7	1	8	1	3	4	3	6	5
30	132	C	13	4	5	4	5	7	6	3	3	0
31	140	C	7	13	6	13	5	6	10	3	7	7
32	178	C	9	8	1	11	4	4	-1	5	2	-1
33	80	A	12	8	5	5	7	0	1	2	0	0
34	193	G	8	2	8	4	6	1	3	0	1	0
35	156	C	16	10	8	4	1	8	7	4	3	5
36	79	A	9	14§	5	5	3	7	4	1	1	8
37	148	C	17	4	7	8	5	6	0	1	6	3
38	77	T	8	9	7	6	5	3	0	5	7	1
39	98	T	0	1	4	2	0	0	0	-1	2	5
40	97	T	11	6	0	1	3	0	2	1	1	0
41	108	G	7	3	0	6	2	2	0	0	0	4
42	114	G	12	4	4	2	4	2	3	4	2	7
43	99	G	10	1	10	3	4	5	0	8	4	0
44	49	A	0	2	6	1	2	0	0	0	0	0
45	97	G	4	0	2	0	1	3	0	0	0	3
46	117	G	3	2	0	0	5	4	0	0	1	1
47	93	C	1	2	2	2	1	0	0	0	1	3
50	41	A	4	4	4	5	0	0	0	0	0	0
51	102	G	2	2	7	2	3	1	3	0	0	0
52	138	G	0	4	2	0	3	1	0	0	0	0
55	122	G	11	3	0	5	0	0	1	2	0	0
56	130	G	6	-1	6	6	0	5	2	0	0	6

The entries in the body of the table are the observed number of cases in which both positions are not mutated, expressed as the excess over the expected number: $D(m, n) = O^{mn} - E^{mn}$ (see text).

*Position in the sequence of the first of two bases examined for correlation in the number of mutations.

†The total number of mutations (in the position listed in column 1) in a set of 1750 *Alu* inserts, with no corrections.

‡Position of the second base in the comparison, expressed as nucleotides to the right of the position listed in column 1.

§As an example the number 14 is the number of cases in which neither position 36 nor 38 is mutated, expressed as excess of what is expected.

Global Correlations in the *Alu* Sequence

Since the correlation method was effective in locating a site that is conserved in *Alu* inserts (Table 3), the method was used to search for other sites and to show that the absence of mutation in the different sites is correlated, suggesting joint function of distant sites. For this purpose the correlation between 5-nt blocks was examined, and the correlation numbers were calculated as the sum of all of the 25 individual nucleotide pair correlations between two 5-nt blocks. It turns out that hundreds of block correlations are probably significant so a strict criterion was used to select the strongest correlations. Only 5×5 block correlations >78 were included, and a total of 96 of these high-scoring correlations were observed. An equivalent Monte Carlo test showed only two accidental correlations scoring higher than 78. Forty-six of the observed 96 significant correlations were limited to the neighborhood of sites. The remaining 50 were significant correlations between distant sets of nucleotides. This suggests that the different conserved regions in the *Alu* inserts function in a coordinated fashion as might be expected if they cooperated together as a set of specific protein-binding sites.

Good correlations are seen for the "pol III promoter region." A central region beginning at position 101 is highly correlated both with near and far nucleotides and shows small amounts of substitution in Fig. 1. There is apparently no published evidence suggesting a function for this central "site-like region," which includes blocks centered between positions 101 and 107 (GAAACC) and some flanking sequences. This region is highly correlated with a region from 210 to 217 with a high score [sum of $D(m, n)$ ranging from 125 to 166]. It appears safe to conclude that many sites exhibit correlated suppression of mutations and probably are functional sites in many *Alu* inserts.

Lack of Lineages of *Alu* Inserts

The suppression of mutation in sequences that are part of "pol III promoters" indicates sequence-dependent selection and suggests that *Alu* sequence transcription is one of the processes responsible for the selection. It is easy to visualize selection favoring effective "pol III promoter" sites if there were lineages of *Alu* sequences giving rise to many more inserts, since selection would favor those with more efficient transcription, but this is not the case. The evidence shows that *Alu* repeats that are significantly different (at the CONSBI positions) from the source sequence are rarely replicated. The logic is that if the *Alu* sequences derive from one source, then any pair will differ from each other by just the sum of their differences from the source sequence, except for accidental identical mutations in both. Few such duplicate mutations are expected for modest divergence. The 1539 known nearly full-length *Alu* sequences were all compared with each other, and the divergence between them at CONSBI positions was determined (1,166,000 comparisons). A diagram was made (not shown) in which the sum of the divergence from the modern consensus of both members of a pair ($Da + Db$) was plotted against the divergence of the members of the pair from each other (Dab). Almost all comparisons (99%) fell near the diagonal, showing equality of Dab and $Da + Db$. There was no sign of events of copying of sequences that had significantly diverged from the modern consensus at CONSBI positions. At higher divergence from the source sequence, increasing numbers of pairs fall below the diagonal due to the increasing number of accidentally matching mutations in the two members of the pair that are being compared. The pattern seen is that expected if the *Alu* sequences derive from a common source (for the CONSBI positions).

The inserts that appeared to the left of the diagonal are candidates for other duplication mechanisms. A total of 86 pairs (with more than 5% divergence from the consensus) had less than one-third of the expected divergence between the members of the pair. *Alu* sequences in this set were all compared with each other using CLUSTAL V, and an attempt was made to form trees of relationship by the neighbor-joining method. No sets of lineage relationships were found. One group of eight inserts differed from each other by 0–5% divergence (CONSBI positions), and they each differed from the modern consensus by 6% of CONSBI positions. All of them were in the same locations in the 5' regions of the known HLA-DQ1 α gene set (20). These eight sequences did not form a lineage, since they were copied as a result of HLA gene duplications. Another group was found to occur in the growth hormone and somatotropin gene cluster, which contains 48 *Alu* inserts. The spacing of the similar pairs in this set shows that they are the result of past regional duplications. Thus an exhaustive search for self-replicating lineages of *Alu* inserts revealed none, and it is safe to say that they do not exist for inserts with more than 5% divergence (CONSBI positions) from the consensus. Thus the selection appears to have been

due to reduced survival or reproduction of primates as a result of certain mutations of interspersed *Alu* inserts.

Could a Role in Gene Regulation Account for the Apparent Selection?

The difficulty with this suggestion is of course that so many *Alu* sequences appear to be affected by selection. The data of Table 2 include four types of changes that occur at 17% of the expected amount and 19 that average 24% of what was expected for these mutations. To explain this using the minimum possible number of *Alu* sequences, no less than 75% of the known *Alu* sequences would have their mutations completely suppressed at these 23 positions. Most *Alu* sequences are not immediately adjacent to genes, and it is hard to visualize the weak control relationships acting at large distances that would be involved. However several genes are known to have large numbers of *Alu* sequences packed within the genes and in their neighborhood, and it is possible to consider a weak enhancer type role in these cases.

There are a number of published examples (20–28) suggesting an unexpected gene regulatory function for *Alu* inserts. In addition there is good evidence showing that pol III transcription of *Alu* inserts influences the effectiveness of nearby pol II promoters (29–33). There is a recently published example of an observed role of an *Alu* sequence in the CD8 gene (34). In searching for DNase I hypersensitive sites, an *Alu* repeat inserted into the last intron was detected that operates as part of an enhancer, which is apparently specific for T lymphocytes. Within the *Alu* sequence, four transcription factor binding sites were shown to be effective: two LyF-1 sites, bHLH, and GATA-3. This *Alu* sequence matches the modern consensus sequence with 12 differences. Five of the changes are at CpGs, which evolve rapidly and can probably be ignored. Of the remaining seven changes, four are within protein-binding sites that are important to the enhancer function. It is unlikely that by chance these four out of seven changes occurred in the 11% of the sequence occupied by the sites. Two of the changes are in the GATA-3 site and are necessary to its function. The authors interpreted this as probable positively selected change in the *Alu* sequence, suggesting that this sequence had adapted to function as an enhancer (34). Considering the ancient and central importance of the CD8 gene, it is a significant question what value the *Alu* insertion and subsequent mutations had for its function. The small degree of divergence of this particular *Alu* sequence from the modern consensus suggests that insertion was probably relatively recent (in the last 10–40 million years), and the selected changes at the binding sites probably occurred later. The putative role of *Alu* sequences is consistent with the proposal that repeated sequences could be a source of variation by moving into positions of significance to gene regulation (35–40), and there have been recent theoretical discussions of the possible role of *Alu* sequences (41–50).

Eric Davidson and Walter Fitch have improved the manuscript. The large set of *Alu* sequences was collected by Jerzy Jurka and supplied by file transfer protocol (FTP) from Barbara H. Rapp of the National Library of Medicine. The work was supported by grants from the National Institutes of Health.

- Wallace, M. R., Andersen, L. B., Saulino, A. M., Gregory, P. E., Glover, T. W. & Collins, F. S. (1991) *Nature (London)* **353**, 864–866.
- Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4770–4774.
- Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55**, 631–661.
- Jurka, J. & Smith, T. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4775–4778.
- Jurka, J. & Milosavljevic, A. (1991) *J. Mol. Evol.* **32**, 105–121.
- Matera, A. G., Hellmann, U. & Schmid, C. W. (1990) *Mol. Cell. Biol.* **10**, 5424–5432.
- Hutchinson, G. B., Andrew, S. E., McDonald, H., Goldberg, Y. P., Graham, R., Rommens, J. M. & Hayden, M. R. (1993) *Nucleic Acids Res.* **21**, 3379–3383.
- Leeflang, E. P., Liu, W.-M., Hashimoto, C., Choudary, P. V. & Schmid, C. W. (1992) *J. Mol. Evol.* **35**, 7–16.
- Jurka, J. (1993) *Nucleic Acids Res.* **21**, 2252.
- Matera, A. G., Hellmann, U., Hintz, M. F. & Schmid, C. W. (1990) *Nucleic Acids Res.* **18**, 6019–6023.
- Perez-Stable, C., Ayres, T. M. & Shen, C.-K. J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5291–5295.
- Saegusa, Y., Sato, M., Galli, I., Nakagawa, T., Ono, N., Iguchi-Ariga, S. M. M. & Ariga, H. (1993) *Biochim. Biophys. Acta* **1172**, 274–282.
- Murphy, M. H. & Baralle, F. E. (1983) *Nucleic Acids Res.* **11**, 7695–7700.
- Saffer, J. D. & Thurston, S. J. (1989) *Mol. Cell. Biol.* **9**, 355–364.
- Deininger, P. L., Batzer, M. A., Hutchison, C. A., III, & Edgell, M. H. (1992) *Trends Genet.* **8**, 307–311.
- Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. (1978) *Nature (London)* **274**, 775–780.
- Bailey, W. J., Hayasaka, K., Skinner, C. G., Kehoe, S., Sieu, L. C., Slightom, J. L. & Goodman, M. (1992) *Mol. Phylogenet. Evol.* **1**, 97–135.
- Blake, R. D., Hess, S. T. & Nicholson-Tuell, J. (1992) *J. Mol. Evol.* **34**, 189–200.
- Lo, K., Landau, N. R. & Smale, S. T. (1991) *Mol. Cell. Biol.* **11**, 5229–5243.
- Brini, A. T., Lee, G. M. & Kinet, J.-P. (1993) *J. Biol. Chem.* **268**, 1355–1361.
- Morzyczka-Wroblewska, E., Harwood, J. I., Smith, J. R. & Kagnoff, M. F. (1993) *Immunogenetics* **37**, 364–372.
- Wu, J., Grindlay, G. J., Bushel, P., Mendelsohn, L. & Allan, M. (1990) *Mol. Cell. Biol.* **10**, 1209–1216.
- Chang, D.-Y. & Maraia, R. J. (1993) *J. Biol. Chem.* **268**, 6423–6428.
- Tomilin, N. V., Bozhkov, V. M., Bradbury, E. M. & Schmid, C. W. (1992) *Nucleic Acids Res.* **20**, 2941–2945.
- Neznanov, N. S. & Oshima, R. G. (1993) *Mol. Cell. Biol.* **13**, 1815–1823.
- Kim, J.-H., Yu, C.-Y., Bailey, A., Hardison, R. & Shen, C.-K. J. (1989) *Nucleic Acids Res.* **17**, 5687–5700.
- Kato, N., Shimotohno, K., VanLeeuwen, D. & Cohen, M. (1990) *Mol. Cell. Biol.* **10**, 4401–4405.
- Sakamoto, K., Fordis, C. M., Corsico, C. D., Howard, T. H. & Howard, B. H. (1991) *J. Biol. Chem.* **266**, 3031–3038.
- Carlson, D. P. & Ross, J. (1983) *Cell* **34**, 857–864.
- Carlson, D. P. & Ross, J. (1986) *Mol. Cell. Biol.* **6**, 3278–3282.
- Chung, J., Sussman, D. J., Zeller, R. & Leder, P. (1987) *Cell* **51**, 1001–1008.
- Sussman, D. J., Chung, J. & Leder, P. (1991) *Nucleic Acids Res.* **19**, 5045–5052.
- Hull, M. W., Erickson, J., Johnston, M. & Engelke, D. R. (1994) *Mol. Cell. Biol.* **14**, 1266–1277.
- Hambor, J. E., Mennone, J., Coon, M. E., Hanke, J. H. & Kavathas, P. (1993) *Mol. Cell. Biol.* **13**, 7056–7070.
- Britten, R. J. & Davidson, E. H. (1969) *Science* **165**, 349–358.
- Britten, R. J. & Davidson, E. H. (1971) *Q. Rev. Biol.* **46**, 111–138.
- Davidson, E. H. & Britten, R. J. (1971) *J. Theor. Biol.* **32**, 123–130.
- Davidson, E. H. & Britten, R. J. (1973) *Q. Rev. Biol.* **48**, 565–613.
- Davidson, E. H. & Britten, R. J. (1979) *Science* **204**, 1052–1059.
- Britten, R. J. (1984) *Carlsberg Res. Commun.* **49**, 169–178.
- McDonald, J. F. (1990) *BioScience* **40**, 183–191.
- McDonald, J. F. (1993) *Curr. Opin. Genet. Dev.* **3**, 855–864.
- Okada, N. (1991) *Trends Ecol. Evol.* **6**, 358–361.
- von Sternberg, R. M., Novick, G. E., Gao, G.-P. & Herrera, R. J. (1992) *Genetica* **86**, 215–246.
- Shapiro, J. A. (1992) *Genetica* **86**, 99–111.
- King, C. C. (1992) *Genetica* **86**, 127–142.
- Brosius, J. (1991) *Science* **251**, 753.
- Schmid, C. & Maraia, R. (1992) *Curr. Opin. Genet. Dev.* **2**, 874–882.
- Howard, B. H. & Sakamoto, K. (1990) *New Biol.* **2**, 759–770.
- Vidal, F., Mougneau, E., Glaichenhaus, N., Vaigot, P., Darmon, M. & Cuzin, F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 208–212.