Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 6148–6150, June 1994
Genetics

Evidence that most human *Alu* sequences were inserted in a process that ceased about 30 million years ago

(repeated sequences/evolution/sequence divergence/retroposition)

Roy J. Britten

Division of Biology, California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625

Contributed by Roy J. Britten, March 14, 1994

ABSTRACT **The primate *Alu* interspersed repeats can be subdivided into classes on the basis of shared nucleotides at a set of diagnostic positions. Each of the classes of *Alu* sequences is apparently the result of past retrotransposition of transcripts of highly conserved class-specific source genes that differed from each other at the diagnostic positions. The nucleotides at the majority of positions are identical among the source genes and therefore were identical among all of the *Alu* sequences at the time of their insertion. These CONSBI (conserved before insertion) positions are useful because the changes that have occurred after insertion are recognizable and the divergence resulting from nucleotide substitutions, insertions, and deletions is informative. The divergence of *Alu* sequences at the CONSBI positions is a measure of the time since a class was inserted. The greatest majority of *Alu* sequences are in one class (identified as class II), and it is particularly suitable for such examination, since nearly full-length sequences are now known for nearly a thousand members of this class. The average divergence of class II members indicates that the class has an average age of about 40 million years. The distribution in divergence of class II accurately fits a sum of two Poisson distributions. The implication is that class II *Alu* sequences were derived from two massive past events of insertion of many *Alu* sequences. In this model the younger subset of class II sequences (corresponding to about 300,000 copies in the genome) has an average divergence of 5% at CONSBI positions. The older set of class II sequences (corresponding to about 150,000 genomic copies) has a 9% average divergence. Based on the drift rate of primate DNA sequences, the events of insertion probably occurred 30–50 million years ago. The goodness of the fit to the Poisson distribution indicates that no significant number of members of class II have been inserted since 30 million years ago.**

Short interspersed sequences (SINES) occur in most eukaryotic DNA, and the primate *Alu* sequence is the most numerous known example. Most copies are about 281 nucleotides long with a terminal poly(A) tract of about 20 nucleotides. The inserted copies are probably the result of retrotransposition (1) of the transcript of a limited set of *Alu* sequences, which are referred to as the source genes (2, 3). The source genes have been highly conserved over the 100 million years or so of the existence of primate *Alu* sequences. However a few changes have occurred during the evolution and branching of the source gene lineages, and the changed nucleotides at these diagnostic positions permit identification of the families. Certain source genes have given rise to very many inserted copies, while others have not been so productive. Retrotransposition from the source genes has occurred during various periods in the past, with particular source genes being active for different and probably overlapping periods of time. These latter concepts are based on the amount of base substitution of the *Alu* sequences after their insertion. While selection has prevented most substitutions in the source genes, it appears that there is no strong selection against substitutions of the retroposed *Alu* inserts, and they drift freely at most positions (4). The large number of sequences of *Alu* inserts that have become available recently permits improved statistical analysis of the divergence of the families leading to a new assessment of the periods of activity of the different source genes.

Diagnostic Positions, Source Genes, and Families

The known diagnostic positions identify five classes (2), which can be further subdivided into eight families (3, 5). The names, sizes, and probable ages of the classes and families are shown in Table 1. The oldest class, I, includes only one identified family (J) but is very heterogenous and likely will be resolved into families. The next and largest class (II) has been split (5) into three families (Sx, Sp, Sq) based on a few diagnostic positions, but all appear to be of about the same age and can be discussed together. There are more recently inserted classes, III (Sc) and IV (Sb), and finally the very small class V group of sequences (Sb1) that includes those being currently inserted into the human genome (6). From these recently inserted *Alu* sequences (9), it is possible to derive a consensus sequence referred to here as the modern consensus (4).

The Conserved-Before-Insertion (CONSBI) Positions

There are 195 positions that are suitable for statistical analysis (4), not including the rapidly evolving CpG dinucleotides. Since the nucleotides in these positions are conserved in the source genes, they are known at the time of insertion of each *Alu* sequence (2). Thus, they are useful for the study of the evolution of the *Alu* sequences after insertion and are here identified as the CONSBI positions (4). At the time of insertion, the nucleotides in the CONSBI positions are the same as those of the modern consensus sequence. The divergence from the modern consensus sequence at the CONSBI positions was determined for the 955 known nearly full-length members of class II, and the distribution is shown in Fig. 1 (boxes). There is a large spread in divergence, and the half-maxima of the distribution are at 3% and 8% divergence, with a mode somewhat greater than 5%.

Poisson Fit to the Majority Class (II) Distribution

The distribution in Fig. 1 is too broad to be fit by a single Poisson distribution. Therefore, a least-squares fit was made to the sum of two Poisson distributions, and the result of each of the distributions is shown by the lower curves in Fig. 1,

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; SINES, short interspersed sequences.

Table 1. Events of insertion of *Alu* repeats in primate DNA

Class*	Fam.†	Quan.‡	Poisson fit§		Event time,¶	Number inserted‖	
			N	*a*	MYBP	in genome	
I	J	266	132	17.6	56	65,000	
			127	12.7	41	62,000	
II	Sx	708	415	10.1	32	203,000	437,000**
			242	16.5	53	119,000	
	Sp+Sq	247	175	10.0	32	86,000	
			60	17.8	57	29,000	
III	Sc	112	87	8.0	26	43,000	
			25	14.1	45	13,000	
IV	Sb	170	32	3.3	11	16,000	
			130	7.5	24	64,000	
V	Sb1	23††	2–3	1.7	(5)	about 1,000	
Total		1528				750,000‡‡	

*Classes of *Alu* inserts as previously defined (2), with class V added for the newly inserted sequences (6).

†Familial names of *Alu* repeats (2, 5).

‡Number of sequences of each class in the set of 1528 nearly full-length *Alu* inserts remaining after removal of nearly all duplicates and those including large segments of foreign sequence.

§A least-squares fit to the sum of two Poisson distributions was done, each having the form $(Ne^{-a}a^x)/x!$ where x is the number of mutations in the 195 CONSBI positions (see text) and a is the average.

¶The time since the multiplication event was calculated from a, the average number of mutations in the 195 positions when a rate of mutation of 0.16% per million years is assumed (7, 8). MYBP, million years before present.

‖The number of inserts estimated to be present in the genome calculated as $N \times 750{,}000/1528$.

**The number of all class II inserts, showing that they are in the majority.

††Most of the known *Alu* sequences of this class are derived from human genetic studies where they cause variation or were specifically screened on the basis of diagnostic positions (6). Thus, the number 23 is not comparable to the number of other classes of *Alu* sequences that were found by sequencing gene regions. Only 3 of the 23 known class V members appear to have been identified by sequencing gene regions, and these 3 form a statistical sample equivalent to that for the other *Alu* sequences. Class V (B1) *Alu* sequences are therefore rare. The average number of nucleotide differences from the modern consensus shown in column 5 is calculated for the total known set of 23 as is the time estimate in column 6. The number estimate in column 7 is based on the 3 discovered by gene region sequencing.

‡‡Arbitrary figure for the total number of *Alu* inserts in the human genome for which estimates range from 500,000 to a million.

while their sum is the upper curve. As shown in Table 1, the average divergence for the two components is about 10 and 17 substitutions per 195 positions or 5% and 8.5% divergence. Table 1 describes in more detail the analysis that was done. The collection of *Alu* sequences available from the National Library of Medicine was used, accepting the classification into families listed there. Files were set up for each class showing the number of members as a function of the number of nucleotide differences from the modern consensus at CONSBI positions, ignoring insertions, and counting deleted nucleotides with the same weight as substitutions. A least-squares fitting program was written that compared a file with the sum of two Poisson distributions. The four parameters (average divergence for two sets and the number of sequences for two sets) were varied stepwise in all combinations, moving to the best fit in each iteration, and the steps were reduced when a stable fit was reached. The final best fits are shown in Table 1. Class II was subdivided into two files Sx and (Sp+Sq) to examine the relative ages of these families, showing that they could not be distinguished on the basis of the distribution of degrees of divergence from the modern consensus. The sum of the quantities in column 4 does not quite agree with the total in column 3, since the least-squares fit numbers are shown uncorrected. The agreement is fairly close, indicating that there are not significant quantities of sequences outside the sets represented by the two Poisson distribution components in the least-squares solution.

To test whether there may have been more recent events, a third Poisson component was included in another least-squares analysis (not shown). If the third component was fixed at 2% divergence (four substitutions at CONSBI positions), the best fit was for an insignificant quantity of this component. If both the amount of divergence and quantity of the third component were free, then the best-fit solution included a small component with larger divergence as follows: 313 sequences with 4.6% divergence and 313 sequences with 7.2% divergence and 72 sequences with 13% divergence. It is known (4) that the rate of substitution varies widely between different positions, and it has been shown that, as expected, this does not affect the accurate fit to a Poisson distribution of the distribution of total divergence of the *Alu* sequences at CONSBI positions.

Times of the Major Events of Alu Insertion and Times of Divergence of Primate Lineages

In order to estimate the time of insertion of the *Alu* sequences, Table 2 lists the measurements of interspecies DNA sequence divergence among the primates calculated from total single-copy DNA measurements (7) and for drifting pseudogene sequences (7, 8). The rate of change has been estimated as 0.13% per million years (7), while the value from Table 2 is 0.16% per million years based on a date of 34 million years ago for the branch between the lineages leading to New World monkeys and that leading to Old World monkeys and apes (8). The time uncertainty is large. It is likely that most of the CONSBI positions of *Alu* sequences drift at the same rate (4) as the neutrally drifting positions of primate DNA in the same lineages. The divergence of the more recent of the two Poisson components (Table 1) is 5% at CONSBI positions, and thus it is likely to have occurred after the branch between the lineages of the New World and Old World monkeys or at about the same time. A reasonable model is that the insertion of class II became a major process about 50 million years ago, lasting to about 30 million years ago and then terminated. The DNA sequences of Old World monkeys differ from those of apes by 7.2%, and this is the sum of the changes on both lineages. Therefore, this branching event is listed at half of the total DNA divergence in Table 2 to show

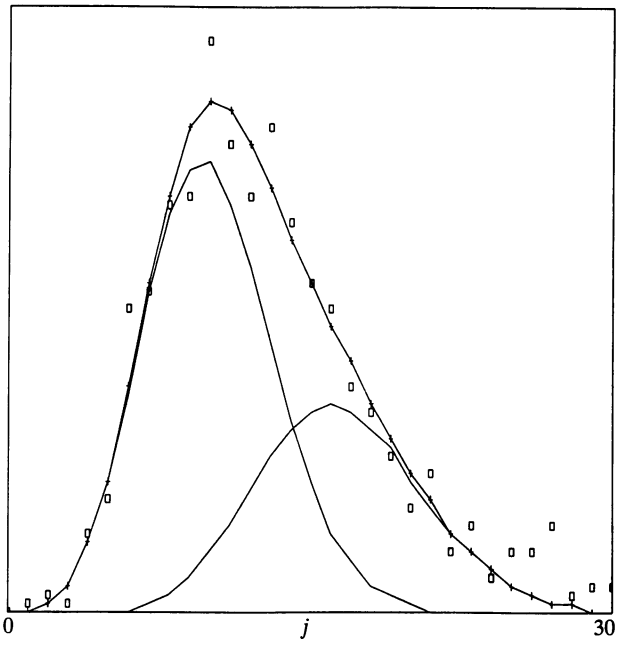

FIG. 1. Divergence of class II *Alu* sequences since insertion. The horizontal axis is the number of differences, *j*, from the modern consensus at the 195 CONSBI (see text) positions, and the vertical axis is the number of class II *Alu* sequences. Boxes show the observations for the 955 almost full-length sequences available at present. The two lower curves are Poisson distributions resulting from the least-squares fit for all of the class II *Alu* inserts. The left-hand lower curve corresponds to the sum of lines 3 and 5 in Table 2, and the right-hand curve corresponds to the sum of lines 4 and 6 in Table 2. The upper curve (+) is the sum of these two calculated components.

the probable drift that has occurred in each lineage since the time of branching. The single-copy DNA divergence between New World monkeys and apes or Old World monkeys is about 11.2% (range from 10% to 12.5%, listed as 5.6%). This number is not significantly different from the 5% estimate for the more recent of the major events of *Alu* insertion listed in Table 1. The earlier major event of insertion occurred well before this time, and thus many of the class II insertions occurred before the separation of the lineages leading to the New World and Old World monkeys (and apes). However, it is possible that this branching event occurred before the termination of the insertion of class II *Alu* sequences. Thus, the suggestion is that some class II *Alu* sequences in Old World monkey and ape DNA are absent from the homologous positions in the DNA of New World monkeys, but many are held in common. An examination of the flanking sequences of many *Alu* sequences will be required to investigate this issue, since some of the individual *Alu* sequences may have been lost in either lineage because of events of deletion by unequal crossover or other possible mechanisms.

Table 2. Drift of primate DNA sequences since branching

	% sequence difference/2		
	η globin* (sequence)	Single copy† (Tm reduction)	Branching time,‡ MYBP
Chimpanzee	0.8	0.8	5
Gorilla	0.9	1.2	7
Orangutan	1.7	1.8	10
Gibbon	2.5	2.4	16
Monkey			
Old World	3.6	3.5	22
New World	5.5	5.6	34

*One-half of the percentage difference in sequence between human DNA and the DNA of the species listed (8).

†One-half of the best estimates of percent sequence difference of total single-copy DNA based on melting temperature reduction of interspecies DNA hybrid duplexes.

‡Estimate of the time of branching calculated from the DNA divergences, based on fossil evidence of the Catarrhine/Platyrrhine branch 34 million years before present (MYBP) (see ref. 8).

Note Added in Proof. At a recent meeting (Workshop on Open Questions in Molecular Evolution, April 18–23, 1994, Guanacaste, Costa Rica) a number of papers (to be published soon in issues of the *Journal of Molecular Evolution*) presented evidence for a wide range of rates of synonymous substitution in different genes of mammals. It would appear, from the samples now available, that two genes picked at random from a single genome would be expected to differ from each other more than 2-fold in the rate of synonymous substitution. The observation that the rate of drift of a region flanking a coding sequence is correlated with the synonymous substitution rate of the coding region is germane to this work on *Alu* sequences. In other words, regions of the genomes exhibit high and low rates of base substitution, apparently due to drift, more or less, free of selection.

There is limited evidence applying to primate genomes during the last 40–50 million years, but it would be surprising if ape and monkey genome evolution did not also include such a range of rates. The implication is that the breadth of the upper curve in Fig. 1 could be due to *Alu* sequences sharing the different rates of substitution of the rate regions in which they occur. Thus, it is possible that the spread in degrees of divergence is entirely due to this phenomenon, and all of the class II *Alu* sequences might have been inserted in a single short period. Such an explanation would have simplicity, but the issue cannot be settled at this time.

I thank Eric Davidson for useful criticism and Jerzy Jurka and Barbara Rapp for the collection of *Alu* sequences. This work was supported by National Institutes of Health grants.

1. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55,** 631–661.
2. Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 4770–4774.
3. Jurka, J. & Smith, T. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 4775–4778.
4. Britten, R. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5992–5996.
5. Jurka, J. & Milosavljevic, A. (1991) *J. Mol. Evol.* **32,** 105–121.
6. Leeflang, E. P., Liu, W.-M., Hashimoto, C., Choudary, P. V. & Schmid, C. W. (1992) *J. Mol. Evol.* **35,** 7–16.
7. Britten, R. J. (1986) *Science* **231,** 1393–1398.
8. Bailey, W. J., Fitch, D. H. A., Tagle, D. A., Czelusniak, J., Slightom, J. L. & Goodman, M. (1991) *Mol. Biol. Evol.* **8,** 155–184.
9. Matera, A. G., Hellmann, U., Hintz, M. F. & Schmid, C. W. (1990) *Nucleic Acids Res.* **18,** 6019–6023.