An Analysis of Replacement and Synonymous Changes in the Rodent L1 Repeat Family[1]

Stephen C. Hardies,[2] Sandra L. Martin,[3] Charles F. Voliva,[4] Clyde A. Hutchison III, and Marshall H. Edgell

Department of Microbiology and Immunology, University of North Carolina

L1 is a family of long interspersed repetitive sequences in mammals that includes the *Bam*HI family in rodents and the *Kpn*I family in primates. Previous studies have shown that L1 repeats contain a long open reading frame and that the family evolves in concert. Working with 32 rodent elements for which DNA sequence is available, we used the distribution of replacement and synonymous changes to determine which L1 lineages had been expressing their reading frame. The evidence obtained is consistent with there having been a small number of L1 genes that have been expressing a functional protein. Much of the concerted evolution in L1 is accounted for by the tendency of these functioning L1 genes to continually create nonfunctional pseudogenes by reinsertion into the genome of sequences derived from their transcripts. The gain of new pseudogenes is balanced by the loss of old pseudogenes with a half-life of 2 Myr. Therefore, most of the observed L1 repeats are at a dead end with respect to either the expression of the L1 protein or the potential to elaborate further copies of themselves. However, the turnover of L1 pseudogenes is sufficient to constitute a vast flux of sequences into and then out of the flanking regions of all cellular genes. If the presence of flanking L1 pseudogenes affects the expression of other genes in even a subtle fashion, this process should represent a major source of genetic variation. A second level of concerted evolution occurs within the functional L1 sequences in a pattern that did not meet our expectations for selfish DNA. Also, in spite of the marked suppression of replacement relative to synonymous changes in functioning L1 genes, they evolve at an overall rate accelerated to the level of their own pseudogenes.

Introduction

L1 or LINES-1 is a family of mammalian long interspersed repetitive sequences that includes the *Kpn*I family of humans and the *Bam*HI family in mice (Singer 1982*a*, 1982*b*; Singer et al. 1983; Singer and Skowronski 1985; Burton et al. 1986). L1 is characterized as a 6–7-kb-long sequence, one end of which is missing from most copies (Fanning 1983; Voliva et al. 1983; for review, see Rogers 1985; Singer and Skowronski 1985). Many elements have poly(A) tracts at one end and are surrounded by short direct repeats. These features have been taken to suggest that most L1 sequences

[1] Key words: concerted evolution, molecular drive, L1 repeat family, replacement and synonymous changes.

Address for correspondence and reprints: Dr. Stephen C. Hardies, Department of Biochemistry, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78284.

[2] Current address: Department of Biochemistry, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78284.

[3] Current address: Synergen, 1885 33d Street, Boulder, Colorado 80301.

[4] Current address: Department of Genetics, University of California, Berkeley, California 94720.

Mol. Biol. Evol. 3(2):109–125. 1986.

0737-4038/86/0302-3205$02.00

are insertions derived from a reverse-transcribed L1 transcript (for review, see Rogers 1985).

A protein-encoding function for the L1 family has been inferred in several ways. Sequenced L1 members often have open segments much greater than the 20-codon average expected for a random sequence (Manuelidis 1982; Thayer and Singer 1983; Martin et al. 1984; Potter 1984). Jagadeeswaran et al. (1982) noted that an individual *Kpn*I sequence had a triplet periodicity like that of coding sequences. However individual L1 members usually carry defects in the consensus reading frame. Martin et al. (1984) compared a 300-bp region between mouse and primate and showed a conserved frame with a higher proportion of changes at synonymous (*S*) positions than at replacement (*R*) positions. Similar results are found over at least a 1400-bp region (Burton et al. 1986). This marked perturbation of the ratio of *R* to *S* changes (*R*/*S*) is taken as strong confirmation that the long open reading frames have acted as the template for a protein, at least historically. So, although some L1 members carry defects, their ancestors descended under selective pressure to express a protein. This paper extends that result to recently diverged rodent L1 sequences, showing that functional L1 sequences have persisted to within at least the past few million years, if not until the present.

The L1 families undergo concerted evolution (Dover 1982; Martin et al. 1985). The analysis of Martin et al. (1985) was based on 32 300-bp sequences from three different species of mice. A tree was derived representing the history of sequence exchanges leading to these elements. In this paper we have further examined the concerted spread of changes into the population of repeats by localizing each change on the tree. Each change was cataloged according to whether it was an *R* or *S* change. This allowed us to determine which parts of the concerted evolution process were subject to conservative selection for the reading frame. The results of that analysis support the model that most L1 sequences are processed pseudogenes derived from the mRNA of functional L1 genes. We present an observation on the mechanism of the concerted evolution among the L1 pseudogenes that distinguishes between continuous dispersal of new pseudogenes and gene conversion. We show that there are multiple functional L1 genes and that they also undergo concerted evolution among themselves. The mechanism of concerted evolution among functional L1 genes appears to differ from that affecting L1 pseudogenes.

The theory of molecular drive (Dover 1982; Ohta and Dover 1983, 1984) is based partly on evaluation of concerted evolution of the rodent L1 families, under the name MIF-1. One conclusion of this work was that concerted evolution is accelerated in the L1 family because some members are more efficient donors in sequence exchanges than others. By a donor we mean any sequence that makes a second copy of itself. During gene conversion, the sequence that ends up in two copies will be considered the donor. During any duplicative process, regardless of the mechanism, the original template will be considered the donor.

The accelerating effect of preferred donors was illustrated by the elaboration of a model for concerted evolution by gene conversion in which some of the sequence elements had a conversion bias (Dover 1982). As noted by Dover, other mechanisms of concerted evolution will be similarly accelerated if a bias is built into the mechanism. For example, in the processed-pseudogene model for concerted evolution of L1, functional L1 genes would be preferred over pseudogenes to be the template for the creation of new pseudogenes. This is because functional genes would presumably be more likely to be transcribed.

The results of this study support this mechanism as a major contributor to the acceleration of concerted evolution of L1 pseudogenes. That is, we find that L1 pseudogenes are preferentially created from functional genes. Conversely, we found no evidence for the further elaboration of any pseudogenes into multiple copies.

We also report evidence for preferred sequence donors during concerted evolution of the functional L1 genes among themselves. Since the mechanism of concerted evolution among the functional L1 genes is not established, the reason why some are preferred sequence donors cannot presently be determined. We coin the term "molecular drivers" to refer to these special sequences. Base changes that arise in the molecular drivers will preferentially spread through the repeat family as described by the theory of molecular drive (Dover 1982).

Methods

The sequences and the tree relating them were taken from Martin et al. (1985). A donor is defined as being any ancestral sequence at a node. A lineage of sequences that includes multiple donors is defined as representing a molecular driver.

Each base change was positioned on the tree by the method of Fitch (1977) with assistance from the program MPN (Czelusniak et al. 1982). Because of the light mutation density in these data, there is little ambiguity arising from parallel and back mutations. There are 30 parallel or back mutations detected out of a total of 174 changes mapped on the tree. Of these, 15 are such well-separated parallel mutations that their resolution is unambiguous (e.g., a variant occurring in only one sequence from *Mus domesticus* and one sequence from *M. platythrix*). The few base changes that were not discretely localized were distributed evenly over their possible positions according to the method of Fitch (1971). The same general results were obtained when these changes were excluded from the analysis. A program named RSBRANCH was written in FORTRAN and run on an IBM mainframe to perform the above bookkeeping and to classify changes as *R* or *S*.

The absolute time scale was calibrated by assuming that the *M. domesticus/M. platythrix* split occurred 11 Myr ago (V. Sarich, personal communication) and that the node separating the *platythrix* repeats from those of the other two rodents represents that split (Martin et al. 1985). Nodes were timed by their relative position on the driver lineage. The results can be read from figure 2. (Fig. 1 is not to scale.) Figure 2 was scaled as follows: Branch lengths for the thickened lineages were taken from figure 1. These lineages include the molecular drivers plus the two direct descendants from each penultimate driver node. The latter are not formally part of the driver lineages; however, they are included because they provide the only way to estimate the height of the lowest driver nodes. An average-node-height tree was then constructed from these selected lineages; that is, each node was assigned a height equal to the sum of (1) one-half of the sum of the heights of its two descendants and (2) the length of the branches leading to these descendants. In this way, the position of each node representing the joining of two drivers was fixed and scaled on figure 2. The nondriver branches were then joined as in the following example: P21 joins its driver 2 changes above a node fixed by the above procedure and 3.9 changes below another fixed node (from fig. 1). It was scaled between these two nodes on figure 2 according to these proportions.

The topology-independent categorization (see Results) of changes as private or nonprivate was done with a program (TREELESS) written in FORTRAN for execution on an IBM mainframe. A private change occurs when a given nucleotide appears in only a single sequence at a particular position. A nonprivate change occurs at a given position if there are multiple nucleotides each shared by more than one sequence. Mutations to private bases were classified as being of the R or S type on the assumption that no other changes had occurred in the codon. If more than one nonprivate base appeared at the same position as a private base and if the nonprivate bases did not all give the same R or S change, the change was distributed among the two types in the proportion of the nonprivate bases that gave rise to those two types. Changes from one nonprivate base to another were proportioned between R and S types according to the average of all possible combinations of such changes per codon weighted by the relative frequency of the respective bases.

The R/S value expected for a pseudogene was calculated with the aid of a program (*RS*) written in FORTRAN for execution on an IBM mainframe. This program divided the potential changes to an L1 sequence into R and S sites and subdivided those classes into transitions and transversions. An unselected value of $R/S = 3.2$ was then found after correction for the fourfold excess of transitions over transversions observed in these sequences.

Martin et al. (1984) calculated an S/R ratio for a mouse/primate comparison that indicated selection on the reading frame. We recalculated this value in a form that would be directly comparable to the observed R/S values in this paper. R and S differences were subdivided into transitions and transversions and individually corrected for parallel and back mutations as described by Brown et al. (1982). In order to handle codons with multiple changes we used the procedure of Miyata and Yasunaga (1981) for averaging over the possible intermediate codons. Then, instead of using the averaging procedure of Brown et al. we back-calculated the number of changes in each category by multiplying the corrected divergence by the number of sites. Summing transitions and transversions we found R and S values that estimate the actual number of changes that occurred before the obscuring effect of parallel and back mutation. An SE that includes a contribution from the multiple-hit correction was calculated according to the method of Tajima and Nei (1984). The primate/rodent R/S ratio was calculated using the last 312 bp of the *M. domesticus* reading frame and the latest primate consensus sequence available from Singer and Skowronski (1985). Actual values were as follows: total R differences, 54.3; total S differences, 48.7; corrected R changes, 60.3; corrected S changes, 77.6; corrected R/S, 0.8; and SE, 0.3. A program for performing this calculation (RSTVS) was written in FORTRAN for execution on an IBM mainframe.

For R/S measurements on various groups of rodent sequences, SEs for the number of R (or the number of S) were taken to be the square root of R (or S) in accordance with the Poisson distribution. SEs for R/S values were calculated according to $(R/S)\sqrt{(1/R + 1/S)}$. This equation is derived from $\Delta(R/S)/(R/S) = \sqrt{(\Delta R/R)^2 + (\Delta S/S)^2}$.

A value of ~1 Myr under selection before inactivation was calculated by fitting points from figure 3 to the equation $R/S = [V_{ar}(x) + V_{pr}(t - x)]/[(V_{as}(x) + V_{ps}(t - x)]$, where V_{ar}, and V_{pr} are rates of R for active and pseudogenes, V_{as}, and V_{ps} are rates of S changes for active and pseudogenes, x is time under selection, and t is the average age of members in each category. The equation was simplified using $V_{ar} = 0.8 V_{as}$ and $V_{pr} = 3.2 V_{ps}$ (see above) and $V_{ar} + V_{as} = V_{pr} + V_{ps}$ (found empirically for these data; see Results). Solving for x gives $x = [t(3.2 - R/S)]/[1.33(1 + R/S)]$. The middle two points in figure 3B were used because greatest accuracy should be obtained where the branches have a nearly equal mixture of selected and unselected descent. Actual values obtained were $t = 2.35$ Myr, $R/S = 1.3$, and $x = 1.46$ Myr for the second point; and $t = 3.97$ Myr, $R/S = 2.7$, and $x = 0.4$ Myr for the third point. Taken collectively, the data from these two points were $t = 3.16$ Myr, $R/S = 1.94$, and $x = 1.0$ Myr. The estimated range covered by one SE was 0.4–1.9 Myr.

Results

Replacement/Synonymous Analysis

In a functional coding sequence, *R* changes are more likely to be detrimental—and therefore selected against—than are *S* changes. As a consequence, the *R/S* ratio, as measured in a molecular phylogenetic reconstruction, discriminates between functional genes and pseudogenes. An example of the use of this method can be seen in the analysis of β-globin sequences (Czelusniak et al. 1982). Functional β-globin genes on average showed an *R/S* ratio of ~1.0, whereas the average for β-globin pseudogenes was 2.2. Pseudogenes are expected to have 2.5–3.0 times as many *R* as *S* changes because there are more potential sites for *R* changes. The analysis of the mouse β-globin pseudogene, βh2, further illustrates the sensitivity of this method. Its *R/S* value of 1.8 is intermediate between that of a functional gene and that of a pseudogene. Subsequent analysis (Phillips et al. 1984) showed that βh2 had a history that included some time as a functional gene and some time as a pseudogene. A systematic approach to timing the inactivation of pseudogenes based on this principle has been published (Miyata and Yasunaga 1981).

Throughout this paper the term "functional," when applied to L1 sequences, refers to expression of the reading frame as evaluated by *R/S* values indicating selection. It does not refer to the capacity to transpose or to participate in other genetic or biological activities.

L1 Repeats Include Both Genes and Pseudogenes

The distribution of *R* and *S* changes in the tree relating 32 rodent L1 sequences is shown in figure 1. There is a low frequency (average of 4%; Martin et al. 1985) of base changes among these sequences; therefore, we expect a low incidence of parallel and back mutation. So the tree in figure 1 should be a good estimate of the historical sequence of base changes. The overall *R/S* ratio for these sequences is 1.6 ± 0.3. This is intermediate between the value expected for pseudogenes (3.2; see Methods) and that observed for the divergence of functional L1 sequences between rodents and primates (0.8; see Methods). So the lines of descent in figure 1 include a mixture of sequences under selection and pseudogenes.

The Molecular Drivers are Real Genes

Dover's (1982) treatment of molecular drive in this family predicts that a subset of the elements should be the predominant sequence donors regardless of the mechanism that supports its concerted evolution. On an evolutionary tree, any sequence above a node is a donor, because it has given rise to a second copy of itself. We find that a small number of lineages can be chosen that include all of the nodes (thick branches, fig. 1). Each of these lineages defines the descent of a sequence that is frequently a donor. The small number of such sequences observed is in agreement with Dover's analysis. Consideration of these lineages in isolation yields an *R/S* value of 0.9 ± 0.3, showing that they are functional genes under selection to make a protein. We refer to these sequences as molecular drivers. They are expected to drive their sequences into the genome in such a way that the long-term evolution of the entire

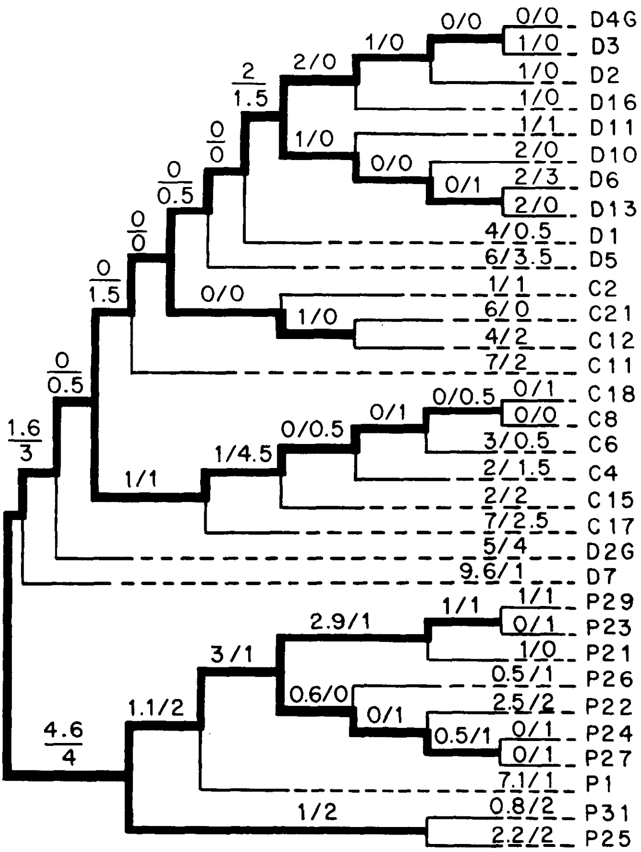

FIG. 1.—Replacement/synonymous changes incurred during the descent of rodent L1 sequences. The tree, taken from Martin et al. (1985), was derived by the method of maximum parsimony and has a total of 174 nucleotide substitutions. Sequences labeled D originate from Balb/c mice (*Mus domesticus*), sequences labeled C from *M. caroli*, and sequences labeled P from *M. platythrix*. D4G and D2G correspond to L1Md-4 and L1Md-2 of Voliva et al. (1984). Thick lines are molecular drivers. Dashed lines indicate the average period of descent during which sequences were estimated to reside in pseudogenes. Thin lines indicate the average period during which sequences were estimated to reside in functional L1 genes. The tree is not drawn to scale.

family will reflect the short-term behavior of the drivers. This expectation is fulfilled in this case, in that the *R*/*S* value between rodents and primates matches that of the rodent drivers considered in isolation.

When constructing a scaled tree (fig. 2), we had to recognize the possibility that the molecular drivers and the other sequences might evolve at different rates. This tree was scaled based on driver lineages only (see Methods). In this way the estimate of the time when the different sequences arise from the driver would not be altered if, for example, the nondrivers evolve at a faster rate.

Other Functional L1 Genes Are Also Present

The average *R*/*S* value among the branches that remain after the molecular drivers (those that end in a contemporary sequence) are factored out is 2.2 ± 0.4, still indicating

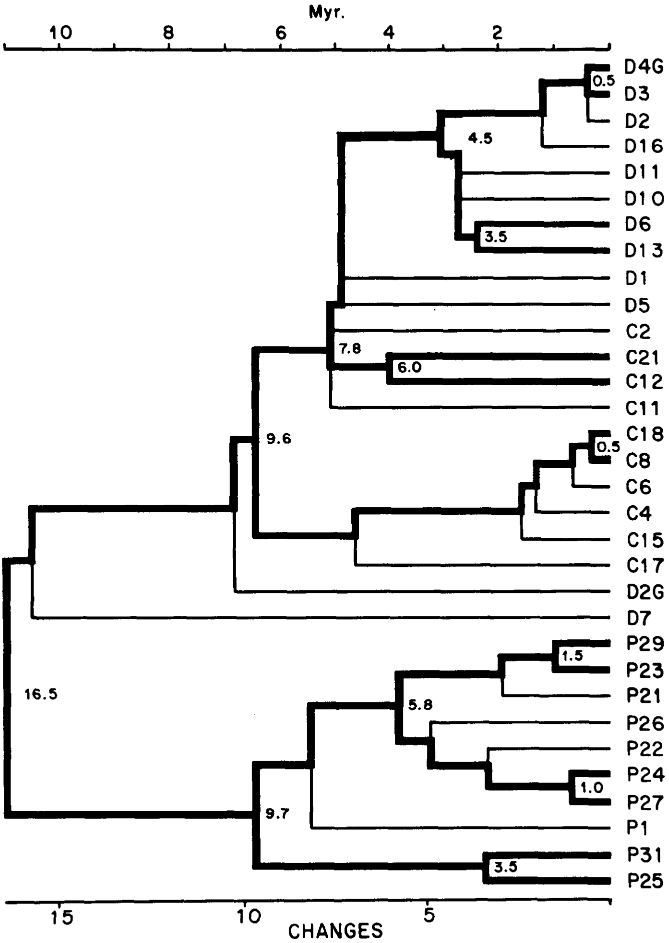

FIG. 2.—Scaled tree based on descent of molecular drivers only. An average-node-height tree consisting of only the thick lines was first constructed (see Methods). The corresponding node heights are marked on the figure in substitutions. That tree was scaled considering the *Mus platythrix/M. domesticus* split to be at 11 Myr. The other branches were then positioned according to their relative position on the driver lineages.

a mixture of functional genes and pseudogenes. When we categorized those branches by length, we discovered that short branches of one or two mutations showed the same *R/S* value as the molecular drivers and that longer branches showed an increasing *R/S*, until the pseudogene value was approached (fig. 3A). Furthermore, when the age of each branch was calculated from its position within the tree (fig. 2) rather than from its own length, the same result was obtained (fig. 3B). Thus we conclude that, on average, a nondonor sequence spends enough time under continued selective pressure to accumulate one or two mutations before being inactivated to become a pseudogene (see thin solid and dashed lines, fig. 1). Using the middle two points in figure 3B and an adaptation of the method of Miyata and Yasunaga (1981), we calculated that the average period of time spent under selection is ~1 Myr (see Methods). Note that this value is an average and that the actual behavior of any individual sequence cannot currently be measured.

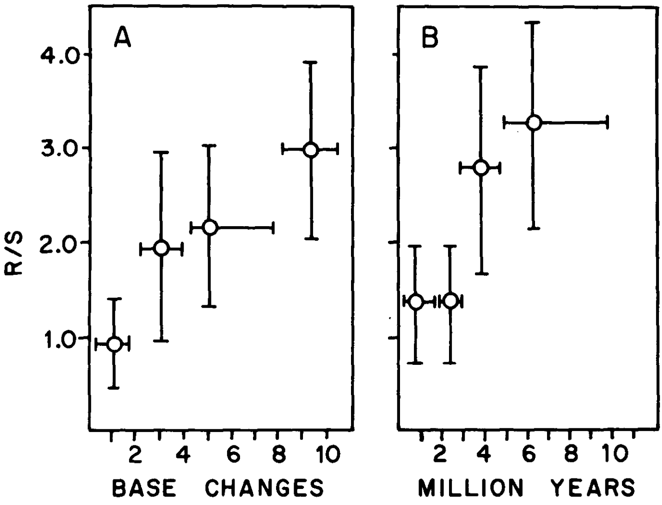

FIG. 3.—Replacement/synonymous ratios of branches by age categories. In panel A, age was estimated from the branch length; in panel B, from the position of the node on the driver sequences as in fig. 2. SEs are derived as in Methods. Horizontal error bars represent the range of values that were arbitrarily grouped. Sequences were grouped into a category containing the lowest 14, then into three categories containing six sequences each.

Each L1 Gene Generates Multiple Pseudogenes

At this point it is important to consider the effect of mobile sequences on the structure of the tree. When a sequence moves, it takes the divergence of its parent with it and then accumulates further divergence at the new location. This should have no effect at all on the topology of the reconstructed tree. Each branch should still represent the history of a sequence, although the sequence may not have spent its entire past in the place where it was found. However, if the character of the sequence is different in the new location, there may be an abrupt shift in its rate of evolution in the middle of a branch.

In order to reconcile the indication of function on each branch with the frequent truncation of L1 sequences, we interpret the inactivation occurring on each branch to represent movement of the sequence to a new genomic location. That is, ~1 Myr after derivation of its sequence from the molecular driver, a functional gene generated a truncated processed pseudogene at another location, which was subsequently sampled. This means that there must be multiple functional L1 genes that are sequence donors for the creation of pseudogenes. The difference between the above donors that are not molecular drivers and the donors that are molecular drivers is in the nature of the progeny that they produce. Donors that are not molecular drivers produce progeny that have no further capacity to donate copies of themselves and therefore are at a dead end with respect to the process of concerted evolution. If the number of sequences sampled approaches the number of these secondary donors, then two progeny from the same donor may be sampled and a node defining that ancestor will appear on the tree. As these events recede into the past, they leave no lasting effect on the evolution of the family. On the other hand, the molecular drivers are genes that preferentially donate in exchanges among functional L1 genes. This guarantees that their sequences will dominate the family of functional genes—and subsequently the pseudogenes.

Thus the molecular drivers emerge as masters over a two-tiered system of sequence exchanges. The first tier is concerted evolution among functional L1 genes, the second is the production of pseudogenes. From this position, the behavior of the molecular drivers determines the major trends to be found in the tree—and the long-term tempo and mode of evolution of the entire L1 family.

Depending on the number of pseudogenes that are derived from each functional gene, there will be an amplification of sequences that are separated from the molecular driver by >1 Myr. Such an amplification is observed in the frequency distribution of repeats categorized by age (fig. 4). We see that repeats of between 1 and 2 Myr (containing one change from the molecular driver) are the most prevalent, while older repeats drop off in frequency, indicating the operation of a clearance mechanism (see below). However, the very youngest repeats (those with no differences from the molecular driver) are vastly underrepresented. The observed amplification occurs at ~1 Myr, corresponding to the average time of inactivation.

A Control against Computationally Introduced Artifacts

We then devised a method to show that these results will reflect the true behavior of the L1 family even if the tree from which they were calculated should turn out to be incorrect. Any reasonable tree, regardless of how it was inferred, will have the vast majority of private changes (variants found in one sequence only) on the nondonor lineages. Similarly, all changes that produced a base shared by more than one sequence (traditionally called phylogenetically informative changes) will mostly be placed on the donor lineages of any reasonable tree. This is because the alternative would require

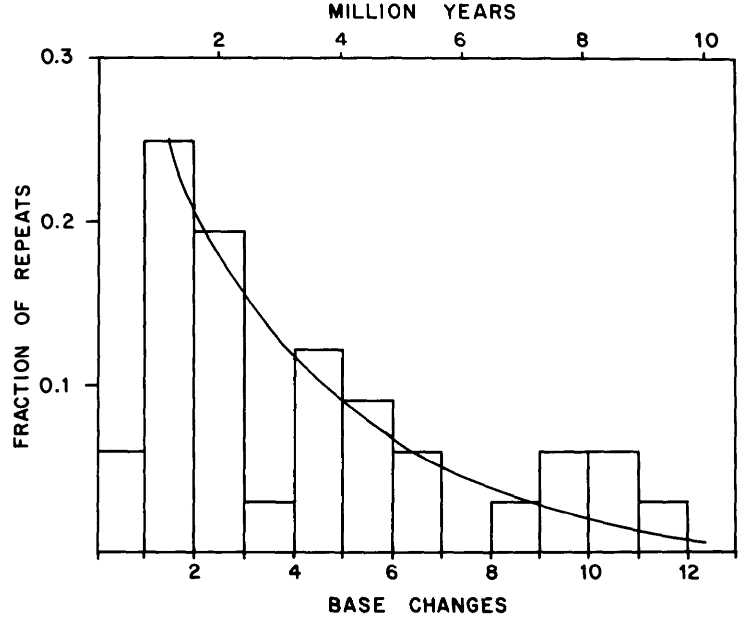

FIG. 4.—Frequency distribution of branches categorized by age. Age was estimated directly from the branch length using the rate obtained by Martin et al. (1985) of 4.1 × 10^{-9} changes/site-year. The curve illustrates the expected distribution for the steady-state case of one-half of the repeats being replaced by new repeats every 2 Myr.

such positions to be changed twice (i.e., would require a high incidence of parallel and back mutation). However, it is unlikely that many positions would be changed twice in a data set in which a large majority of positions are not changed at all.

We then tabulated the *R*/*S* values for these two classes of base changes (see table 1), finding them to be 3.4 for private changes and 1.0 for the other changes. The private base changes were grouped according to total number of private changes per sequence and plotted against the group *R*/*S* value (not shown) to derive a result similar to that in figure 3A. Similarly, the frequency distribution of private changes per sequence (not shown) mirrors the result in figure 4. So, our key observations are supported by a method that does not depend on any particular tree structure. The tree describing the true history must lead to similar results unless it is extremely unparsimonious.

A Model of the L1 Family

Our complete model for the history and population structure of the L1 family (fig. 5) specifies that all of the function is confined to a subset of presumably fully intact repeats. Concerted evolution within this subset is preferentially driven by a small number of its members, here referred to as molecular drivers. The functional L1 genes derive their sequences on an average of once every 1 Myr from the molecular drivers and frequently donate truncated copies to the genome, copies that thereafter evolve as pseudogenes.

The number of functional L1 genes apparently sampled in our tree is greater than 10 per species of rodent. Otherwise we would have sampled multiple truncated copies from the same parent, causing the nodes to appear at the time of inactivation

Table 1
Categorization of Private Changes in L1 Tree as Replacements (*R*) or Synonymous (*S*) Changes

	Type of Change			Type of Change			Type of Change	
Sequence	*R*	*S*	Sequence	*R*	*S*	Sequence	*R*	*S*
P25	1	2	D7	8	1	C12	3	0
P31	0	1	D2G	4	2	C21	4	0
P1	7	1	D5	6	2	C11	5.2	1.8
P27	0	1	D1	2.9	0.1	C2	1	0
P24	0	1	D13	2	0	C17	7	1
P22	1	1	D6	1	2	C15	2	0
P26	0	1	D10	2	0	C4	1	0
P21	1	0	D11	1	0	C6	3	0
P23	0	1	D16	0	0	C8	0	0
P29	1	0	D2	0	0	C18	0	0
			D3	0	0			
			D4G	0	0			

Sums	*R*	*S*	*R*/*S*
All private changes	64.1	18.9	3.4 ± 0.9
All nonprivate changes	25.6	26.4	1.0 ± 0.3

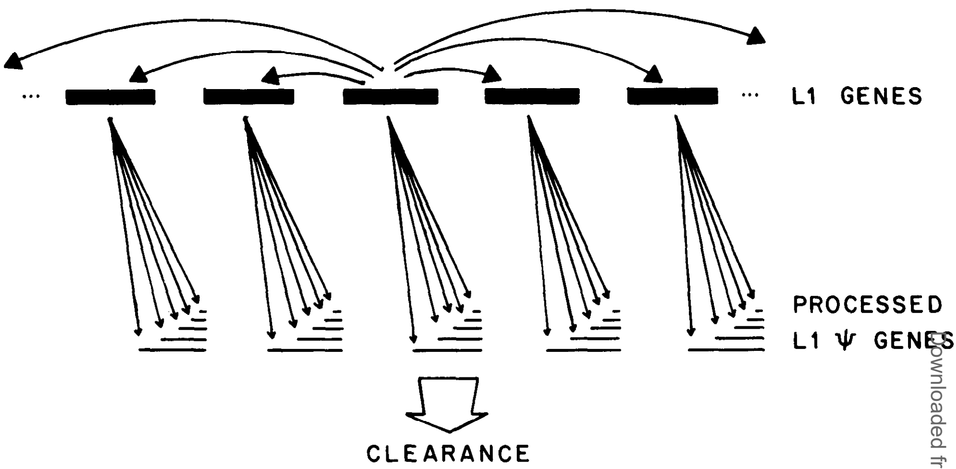

FIG. 5.—Model for the history of the L1 gene family. The upper part of the figure illustrates a family of functional L1 genes undergoing concerted evolution with a preferential donor. The lower part illustrates the associated family of processed L1 pseudogenes being distributed throughout the genome and subsequently being removed.

rather than before. There are some cases in our sample of sequences where this may have happened, e.g., D1 and D5. However, since there are no unambiguous cases, we suspect that even repeats like D1 and D5, which have identical parents, are probably paralogous. Since the sampling process involves a long period of time and the entire population of each species, models can be built with fewer functional L1 genes in any given mouse genome. In fact, a model can be built in which all functional L1 genes reside in a virus and never in a mouse genome. Assuming that all functional elements are intact over the entire 6–7 kb, their numbers would have to be small enough to be consistent with the number of observed intact elements, i.e., on the order of 10,000 (M. B. Comer, personal communication). Thus the possible range for the copy number of functional L1 genes per genome is rather large at this time.

The concerted evolution of the L1 family can be seen as the interplay of several continuous processes operating in a steady state. For the sake of the following illustration only, we will consider the sequence exchanges among the functional L1 genes as arising from a biased gene conversion mechanism, in the manner of Dover (1982). Then the limit of 1 Myr of functional divergence away from the molecular driver would simply represent the mean time required for any member of the functional L1 gene family to evolve away from the molecular driver before it was converted back to the driver sequence. All functional members of the L1 family would throw defective copies of themselves out into the genome at a constant rate. The amplification process, then, would not be a sudden burst but a smooth accumulation of copies throughout the period between successive conversions of the parent sequence.

Alternatively, consider a concerted evolution process for functional L1 genes in which they are continually created from the driver by a dispersal process similar to that which creates the pseudogenes. Then the 1-Myr limit for the divergence of functional genes from the driver would have to represent the average time until inactivation of functional genes. During this time they would be creating truncated pseudogenes, and at the time of inactivation they might persist as full-length pseudogenes themselves.

However, the loss of the capacity to make truncated copies would have to coincide with the loss of function.

This latter model uses similar mechanisms to explain the concerted evolution of the functional genes and the concerted evolution of the pseudogenes. However, to fit the data, some mechanism would have to cause the molecular drivers to repeatedly emerge from among the functional genes as the preferred sequence donor for future generations. This would have to be done without making the molecular drivers preferred direct donors of pseudogenes, or else the other functional genes would not have shown up in our analysis as intermediates in the flow of sequence information from driver to pseudogenes.

The Rate of Evolution of L1 Genes Is Unusually High

The divergence of mouse from primate L1 genes has been observed to be unusually high, with *S* sites changing faster than is normally expected even for pseudogenes (F. H. Burton, personal communication). For example, in the primate/rodent comparison we found 70% (48.7/69.7) of *S* sites to differ before correction for parallel and back mutation. By comparison, the large intervening sequences of the β-globins diverged by only 45% between these same species (Hardies et al. 1984). This trend is also apparent among the sequences examined in this paper (table 2). On average, from each node the subsequent length of the driver branch is similar to that of the pseudogene branch. Though there are fewer *R* changes to the drivers as a result of selection, there is a balancing increase in *S* changes. Therefore, on inactivation the rate of variation within the L1 pseudogenes accelerates at *R* sites but decelerates at *S* sites. Thus the rate of change of the L1 molecular drivers, before selection against unacceptable replacements, is high, even by comparison to that of their own pseudogenes. Furthermore, since the rate of evolution of the pseudogenes (4.1×10^{-9} changes/site-year; Martin et al. 1985) is typical of that for pseudogenes (4.9 ± 10^{-9} changes/site-year; Li et al. 1981), we can say that it is the molecular drivers that are behaving oddly.

The Rate of L1 Pseudogene Turnover

The rate calculation of Martin et al. (1985) was revised upward for the pseudogenes by using a frequency distribution of only the pseudogenes (fig. 4). The curve, which describes the relative number of repeats of different ages, decays by half approximately every 2 Myr. Consistent with the conclusion of Martin et al. (1985), we conclude that

Table 2
Total Length of Driver versus Nondriver Lineages

	Lengths of					
	Driver			Nondriver		
Driver/Nondriver Pair[a]	*R*	*S*	Total	*R*	*S*	Total
D4G/D16	1	0	1	1	0	1
D13/C11	5	3	8	7	2	9
C8/D7	3.6	11	14.6	9.6	1	10.6
P27/P26	0.5	3	3.5	0.5	1	1.5
P23/P1	6.9	4	10.9	7.1	1	8.1
Total	17	21	38 ± 6.2	25.2	5	30.2 ± 5.5

[a] Pairs were selected to avoid redundancy in the driver lineages.

one-half of the pseudogenes are lost and replaced by new copies every 2 Myr. So a clearance mechanism operates at about the same rate as the pseudogene generation mechanism, keeping the total copy number about the same. Also, we note a possible excess of older branches, which may represent sequences that have escaped the clearance mechanism. These conclusions about a clearance mechanism are based on the assumption that the rate of insertion is roughly constant with respect to time. Alternatively, one could explain the curve in figure 4 by proposing that the insertion process had been rapidly accelerating during the last few million years.

Martin et al. (1985) calculated a 3.3-Myr half-life for turnover of L1 sequences. Here we have resolved that figure into 1 Myr for sequence information to pass from the driver through functional genes into the pseudogene population and 2 Myr for the pseudogenes to turn over.

Discussion

To summarize our findings (see fig. 5), the L1 repeats consist of a set of functional L1 genes and a large associated family of pseudogenes. The term "functional" refers to the expression of the reading frame as evaluated by the suppression of *R* changes, not to the capacity to transpose or to participate in other genetic events. The pseudogenes undergo concerted evolution by continually being dispersed and turned over. A separate genetic mechanism causes concerted evolution among the functional L1 genes.

L1 sequences that generate additional copies of themselves are referred to as "donors" in this paper. Donor capacity and the capacity to function as a gene were evaluated separately with no assumption as to their relation. We found that donor capacity was correlated with evidence of functional-gene expression. We find no support for the idea that unexpressed L1 repeats ever propagate more copies of themselves. Our results are derived from averages of a number of repeats and so do not exclude other mechanisms operating at low frequency. However, the mechanisms that we identified are the ones that shape the overall dynamics of concerted evolution in the rodent L1 family.

Our results are consistent with the model suggesting that L1 repeats are mainly reinserted cDNAs that were reverse transcribed from RNA. This model explains the correlation between gene expression and donor capacity, since an mRNA would be intermediate for both. However, there may be other models that explain the correlation between expression and donor capacity.

The L1 Gene Family

All that we know about the functional L1 genes has been deduced from the reconstruction of their sequence from the sequences of the L1 pseudogenes. It is possible that none of the sequences actually characterized were functional L1 genes. Furthermore, it may be difficult to recognize a functional L1 gene even if it were cloned and sequenced. Much of what we know about the L1 repeats is really characteristics of the pseudogenes. For example, while the pseudogenes are interspersed among other genes, the functional L1 genes might just as well be clustered. The functional genes might have introns—and hence not conform to the consensus pseudogene restriction map. They might have homology with one another beyond the traditional boundaries of the repeat—and conceivably could contain structures not found in the pseudogenes, such as long terminal repeats.

A pseudogene generated from an L1 gene on a different chromosome will become separated from its parent during subsequent breeding cycles. In this way, the collection of pseudogenes in an extant animal will represent all of the L1 alleles in the previous interbreeding population. Consequently, each of our results contains an influence from population genetics as well as from intragenomic exchanges. We have not attempted to resolve these influences in this presentation.

There are several circumstances proposed under which selfish DNA can persist without benefiting the host organism (Doolittle and Sapienza 1980; Orgel and Crick 1980). We point out that the pattern of evolution in L1 is not consistent with at least the most prevalent paradigm for evolution of selfish DNA. In this case, a competent selfish sequence would have to constantly amplify itself at a sufficient rate to balance the loss of members resulting from inactivating mutations. Since the properties of the entire L1 family depend on the sequence of the molecular drivers, it is the rate of duplication and loss of molecular drivers that are relevant to the above process. In contrast to the high copy number and rapid amplification of L1 sequences in general, there are only one or two molecular drivers per species, and during most periods of time there are no duplications to produce new molecular drivers. Yet between such duplications there is a reduced (*R*/*S*) ratio for molecular drivers, suggesting that a form of selection based on something other than amplification has occurred.

Concerted Evolution among the L1 Genes

Concerted evolution proceeds among the functional L1 genes with a bias for two to three of the members per species to preferentially donate their sequences to the others. We have discussed how either gene conversion or a transposition mechanism would fit with our results (see Results). The mechanism of the concerted evolution among the functional L1 genes or the reason some donors should be preferred is unknown. The mechanism of concerted evolution among the functional L1 genes cannot be the same as that between the functional L1 genes and their pseudogenes. If it were, then the genes that drive the concerted evolution of the functional family would also directly produce most of the pseudogenes. Then the other members of the L1 gene family would not have shown up as intermediates in the flow of sequence information from the molecular drivers out into the pseudogene family.

Concerted Evolution of the L1 Pseudogenes

We considered a variation on the mechanism of concerted evolution of the L1 pseudogenes in which, after their original dispersal, they are gene converted many times by the functional L1 genes. In that case, the dynamics of their concerted evolution would predominantly be controlled by gene conversion rather than by the dispersal process. Jubier-Maurin et al. (1985) have argued for a gene conversion mechanism based on an analysis that included the sequences discussed in this paper. Gene conversion would remove an old copy for each new copy added and therefore naturally explain the overall balance of pseudogene production and clearance indicated by our analysis.

However, we conducted a test that favored the dispersal-only model over the dispersal-conversion model. It is based on the expectation that the small direct repeats of the flanking DNA generated during the initial insertion would not be affected by subsequent conversions. Thus we timed the true age of the insertion by the divergence of these small direct repeats. We averaged the data for multiple paired repeats to increase the accuracy of the measurement. Five sequenced repeats from the mouse β-globin cluster (Voliva et al. 1984) contain no changes out of 114 total base pairs of direct repeat. When we use the rate of sequence divergence in a neutral sequence derived by Li et al. (1981) (4.9×10^{-9} changes/site-year), we see that this translates to an average age of <2 Myr ($4.9 \times 10^{-9} \times 114$ sites = 1 change/1.8 Myr). It appears that the insertion-conversion model can only be salvaged if a conversion mechanism can be found that rejuvenates the direct repeats. A stronger test of this possibility will be made when it is known how often the L1 pseudogenes are in the same locations in closely related species. Meanwhile, we think that it is a reasonable conclusion that the predominant flow of sequence from the L1 genes into the genome at large is by way of new insertions.

Possible Mechanistic Explanations for the L1 Pseudogenes

The high rate of both accrual and elimination of L1 pseudogenes requires some explanation. A purely mechanistic explanation for the accrual might be that the L1 gene happens to be expressed heavily in germ cells. Why they should be lost at such a high rate is more problematic. Half of all L1 pseudogenes are deleted every few million years, yet nonessential DNA in general is not thought to disappear that fast. A mechanistic explanation might be that the short direct repeats created on insertion are sufficient to specifically guide subsequent deletion. Short direct repeats in this size range are known to guide large deletions in *E. coli* (Albertini et al. 1982). However, we would like to propose an alternative to a mechanistic explanation for the behavior of the L1 family, an alternative based on selection.

Potential Evolutionary Effects of the L1 Pseudogene Family

The possibility that repetitive sequences are involved in gene regulation has long been recognized (Davidson and Britten 1979, and references therein). We emphasize that the rate of insertion and deletion implicit in our results represents a large flux of genetic change. For example, the mouse β-globin gene cluster currently has eight L1 elements (Voliva et al. 1985). It is 65 kb long and has been separated from primates for ~100 Myr. Of 10^5 L1 copies, half will insert and half be deleted per 2 Myr, and 10^5 events/2 Myr × 100 Myr × 65×10^3 bp target size/3×10^9 bp/genome = 108 events in the rodent globin cluster since the separation from primates. Our times and rates are calibrated against a single speciation event (see Methods) and may subsequently be scaled up or down should better information become available. But the number of L1 insertions and deletions is still clearly going to be significant. If L1 insertions have even subtle effects on the expression of neighboring genes, then this process will be a major source of genetic variation. Such effects could be entirely passive, by being restricted to changing the spacing of active elements within the flanking sequences. Alternatively, the L1 genes may have acquired a transcriptional enhancer in the transcription unit, thus causing an active effect of their pseudogenes on neighboring genes. In either way, the high rate of insertion and deletion could reflect selective pressures rather than some unusual mechanism.

Such a process would bear on an important question in evolutionary biology. As suggested by Wilson et al. (1977), regulatory changes are a better candidate than structural changes to explain the vast morphological diversity arising during mammalian evolution. The proposed process fits the requirements perfectly. It would provide a continuous and plentiful source of subtle variation. The induced changes could also be qualitative in the same sense that rearranged flanking sequence can cause hereditary persistence of fetal hemoglobin (for review, see Weatherall and Clegg 1982). We emphasize that the proposed effects need not be large, since evolution proceeds by a series of small changes.

Finally, no other protein-encoding family known has the strikingly peculiar properties associated with L1. This generates for us the expectation that these unusual features are a harbinger of unusual function.

Acknowledgments

S. L. Martin is a fellow of the Jane Collin Childs Memorial Fund for Medical Research. This research was supported by Public Health Service grants AI08998 and GM21313 from the National Institutes of Health.

LITERATURE CITED

ALBERTINI, A. M., M. HOFER, M. P. CALOS, and J. H. MILLER. 1982. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell **29**:319–328.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

BURTON, F. H., D. D. LOEB, C. F. VOLIVA, S. L. MARTIN, M. H. EDGELL, and C. A. HUTCHISON III. 1986. Conservation throughout Mammalia and extensive protein-encoding capacity of the highly repeated DNA Long Interspersed Repeat One. J. Mol. Biol. (accepted).

CHENG, S. M., and C. L. SCHILDKRAUT. 1980. A family of moderately repetitive sequences in mouse DNA. Nucleic Acids Res. **8**:4075–4090.

CZELUSNIAK, J., M. GOODMAN, D. HEWETT-EMMETT, M. L. WEISS, P. J. VENTA, and R. E. TASHIAN. 1982. Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature **298**:297–300.

DAVISON, E. H., and R. J. BRITTEN. 1979. Regulation of gene expression: possible role of repetitive sequences. Science **204**:1052–1059.

DOOLITTLE, W. F., and C. SAPIENZA. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601–603.

DOVER, G. A. 1982. Molecular drive: a cohesive mode of species evolution. Nature **299**:111–117.

FANNING, T. G. 1983. Size and structure of the highly repetitive BAM HI element in mice. Nucleic Acids Res. **11**:5073–5091.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

———. 1977. On the problem of discovering the most parsimonious tree. Am. Nat. **111**:223–257.

HARDIES, S. C., M. H. EDGELL, and C. A. HUTCHISON III. 1984. Evolution of the mammalian beta-globin gene cluster. J. Biol. Chem. **259**:3748–3756.

JAGADEESWARAN, P., J. PAN, B. G. FORGET, and S. M. WEISSMAN. 1982. Sequences of non-alpha-globin genes in man. Cold Spring Harbor Symp. Quant. Biochem. **47**:1081–1082.

JUBIER-MAURIN, V., B. J. DOD, M. BELLIS, M. PIECHACZYK, and G. ROIZES. 1985. Comparative study of the L1 family in the genus *Mus:* possible role of retroposition and conversion events in its concerted evolution. J. Mol. Biol. **184**:547–564.

LI, W.-H., T. GOJOBORI, and M. NEI. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237–239.

MANUELIDIS, L. 1982. Nucleotide sequence definition of a major human repeated DNA, the Hind III 1.9 kb family. Nucleic Acids Res. **10**:3211–3219.

MARTIN, S. L., C. F. VOLIVA, F. H. BURTON, M. H. EDGELL, and C. A. HUTCHISON III. 1984.

A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein. Proc. Natl. Acad. Sci. USA **81**:2308–2312.

MARTIN, S. L., C. F. VOLIVA, S. C. HARDIES, M. H. EDGELL, and C. A. HUTCHISON III. 1985. Tempo and mode of concerted evolution in the L1 repeat family of mice. Mol. Biol. Evol. **2**:127–140.

MIYATA, T., and T. YASUNAGA. 1981. Rapidly evolving mouse alpha-globin-related pseudo gene and its evolutionary history. Proc. Natl. Acad. Sci. USA **78**:450–453.

OHTA, T., and G. A. DOVER. 1983. Population genetics of multigene families that are dispersed into two or more chromosomes. Proc. Natl. Acad. Sci. USA **80**:4079–4083.

———. 1984. The cohesive population genetics of molecular drive. Genetics **108**:501–521.

ORGEL, L. E., and F. H. C. CRICK. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

PHILLIPS, S. J., S. C. HARDIES, C. L. JAHN, M. H. EDGELL, and C. A. HUTCHISON III. 1984. The complete nucleotide sequence of a β-globin-like structure, βh2, from the [Hbb]^d mouse BALB/c. J. Biol. Chem. **259**:7947–7954.

POTTER, S. S. 1984. Rearranged sequences of a human *Kpn*I element. Proc. Natl. Acad. Sci. USA **81**:1012–1016.

ROGERS, J. H. 1985. The origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.

SINGER, M. F. 1982*a*. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. **76**: 67–112.

———. 1982*b*. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28**:433–434.

SINGER, M. F., and J. SKOWRONSKI. 1985. Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. Trends Biochem. Sci. **10**:119–122.

SINGER, M. F., R. E. THAYER, G. GRIMALDI, M. I. LERMAN, and T. G. FANNING. 1983. Homology between the *Kpn*I primate and *Bam*HI (MIF-1) rodent families of long interspersed repeated sequences. Nucleic Acids Res. **11**:5739–5745.

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

THAYER, R. E., and M. F. SINGER. 1983. Interruption of an alpha-satellite array by a short member of the *Kpn*I family of interspersed, highly repeated monkey DNA sequences. Mol. Cell. Biol. **3**:967–973.

VOLIVA, C. F., C. L. JAHN, M. B. COMER, C. A. HUTCHISON III, and M. H. EDGELL. 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res. **11**:8847–8859.

VOLIVA, C. F., S. L. MARTIN, C. A. HUTCHISON III, and M. H. EDGELL. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178**: 795–813.

WEATHERALL, D. J., and J. B. CLEGG. 1982. Thalassemia revisited. Cell **29**:7–9.

WILSON, A. C., S. S. CARLSON, and T. J. WHITE. 1977. Biochemical evolution. Annu. Rev. Biochem. **46**:573–639.

WALTER M. FITCH, reviewing editor

Received May 17, 1985; revision received November 27, 1985.