Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

GENOMICS 1, 113–125 (1987)

Origin of the Human L1 Elements: Proposed Progenitor Genes Deduced from a Consensus DNA Sequence

ALAN F. SCOTT,[1] BARBARA J. SCHMECKPEPER, MONA ABDELRAZIK, CATHERINE THEISEN COMEY,[2] BRUCE O'HARA,[2] JUDITH PRATT ROSSITER,[2] TIM COOLEY,[2] PETER HEATH, KIRBY D. SMITH, AND LOUISE MARGOLET

Division of Medical Genetics, Department of Medicine, Johns Hopkins University School of Medicine, Johns Hopkins Hospital, CMSC 1203, Baltimore, Maryland 21205

Received March 30, 1987; revised June 24, 1987

A consensus sequence for the human long interspersed repeated DNA element, L1*Hs* (LINE or *Kpn*I sequence), is presented. The sequence contains two open reading frames (ORFs) which are homologous to ORFs in corresponding regions of L1 elements in other species. The L1*Hs* ORFs are separated by a small evolutionarily nonconserved region. The 5′ end of the consensus contains frequent terminators in all three reading frames and has a relatively high GC content with numerous stretches of weak homology with *Alu*I repeats. The 5′ ORF extends for a minimum of 723 bp (241 codons). The 3′ ORF is 3843 bp (1281 codons) and predicts a protein of 149 kD which has regions of weak homology to the polymerase domain of various reverse transcriptases. The 3′ end of the consensus has a 208-bp nonconserved region followed by an adenine-rich end. The organization of the L1*Hs* consensus sequence resembles the structure of eukaryotic mRNAs except for the noncoding region between ORFs. However, due to base substitutions or truncation most elements appear incapable of producing mRNA that can be translated. Our observation that individual elements cluster into subfamilies on the basis of the presence or absence of blocks of sequence, or by the linkage of alternative bases at multiple positions, suggests that most L1 sequences were derived from a small number of structural genes. An estimate of the mammalian L1 substitution rate was derived and used to predict the age of individual human elements. From this it follows that the majority of human L1 sequences have been generated within the last 30 million years. The human elements studied here differ from each other, yet overall the L1Hs sequences demonstrate a pattern of species-specificity when compared to the L1 families of other mammals. Possible mechanisms that may account for the origin and evolution of the L1 family are discussed. These include pseudogene formation (retroposition), transposition, gene conversion, and RNA recombination.

INTRODUCTION

The genomes of man and other mammals contain a large fraction of repeated DNAs. The functions, if any, of these sequences have remained elusive. Repeated DNAs can generally be cataloged as either tandem or interspersed on the basis of their distribution. The interspersed repetitive elements can be further characterized according to length as either short (SINEs) or long (LINEs) (Singer and Skowronski, 1985). The most commonly described LINEs are the L1*sp* elements (where *sp* is the species designation). The human L1 elements have also been frequently referred to as *Kpn*I sequences. There are at least 10^4 copies of the L1 element in the human genome, although the number of complete 6-kb elements is much smaller due to truncation (Singer and Skowronski, 1985). L1 elements are present in all mammals that have been studied (including both placental and marsupial mammals; Burton *et al.*, 1986) and, based on their shared sequence homology in distantly related species, they most likely are derived from a common progenitor. Mammalian L1 repeats appear to evolve in concert such that the majority of elements within a species will be more like each other than they will be to the elements of another species. Consequently, Southern blots of restricted genomic DNA will often give species-specific patterns when hybridized with a LINE probe. However, when cloned elements from a particular species are compared, there are often significant differences between individual sequences.

Various investigators have identified open reading frames (ORFs) followed by an adenine-rich end in some L1 repeats from man and other species (Singer and Skowronski, 1985). L1-homologous transcripts have also been described in various cell types

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. J03034.

[1] To whom correspondence should be addressed.

[2] Predoctoral Training Program in Human Genetics.

0888-7543/87 $3.00

(Schmeckpeper *et al.*, 1984; Kole *et al.*, 1983; Shafit-Zagardo *et al.*, 1983) and a 6.5-kb polyadenylated RNA has been observed in teratocarcinoma cells (Skowronski and Singer, 1985). These observations have led to the description of LINEs as "retroposons" or "retrotranscripts" (Temin, 1985), suggesting that they may be derived by the integration of sequences involving either RNA or cDNA intermediates, as occur with retroviruses (Temin, 1985), and may occur in the formation of SINEs (Weiner *et al.*, 1986) and processed pseudogenes (Wagner, 1986). Likewise, similarities in size and organization and limited sequence homology with transposable elements from *Trypanosoma* (Kimmel *et al.*, 1987), *Drosophila* (Fawcett *et al.*, 1986), and yeast (Fink *et al.*, 1986) suggest that L1 elements may also be related to sequences such as these. If either model is correct, it would follow that a small number of functional genes have given rise to the majority of L1 sequences, which have subsequently diverged from the progenitors by the accumulation of mutations. If such changes have occurred randomly in these nonfunctional copies, then it should be possible to deduce the sequence of the original structural gene or genes from which they were derived. This can be done by determining shared bases in a sufficiently large number of L1 sequences. To that end we have assembled an L1*Hs* consensus sequence based on L1*Hs* sequence data published in the literature and generated in our laboratories. As described below, this analysis has yielded information about the origin of the human L1 elements and the organization of proposed progenitor genes.

MATERIALS AND METHODS

The human 3′ β-globin LINE, an apparently complete 6.2-kb sequence, has been described previously (Adams *et al.*, 1980). The b2 LINE, an autosomal element, was derived from a genomic clone, λHb2, which had been obtained from a λ-phage library of human DNA (Schmeckpeper *et al.*, 1981). The clones X63 and X65 were from a genomic library of AHA-11a DNA, a mouse–human hybrid cell line, containing an X as the only detectable human chromosome (Schmeckpeper *et al.*, 1979). This library was made by partial *Eco*RI digestion and ligation into Charon 4A λ-phage. λ-Phage X63 and X65 were isolated on the basis of hybridization of L1 subclones derived from λHb2. A gorilla L1 element was obtained from a similarly constructed library (Scott *et al.*, 1984) screened with a human β-globin cDNA probe. Restriction maps of the recombinant λ phages were prepared, and appropriate regions from these were subcloned into plasmids for further analysis. Because a majority of published sequences are homologous to the 3′ end of the element, we have accumulated new data from underrepresented portions of the 5′ end of the element.

Several sequences used to generate the consensus were derived from genomic fragments with particular restriction enzyme sites. Selection of sequences defined by particular restriction enzyme sites might bias the final consensus by selecting elements of a given type. In order to lessen this potential problem, the new elements analyzed for this paper were selected from genomic clones by hybridization criteria and because they were at least 6 kb long. The presence of particular restriction enzyme sites was not a criterion for selection. Further, many of the published elements were derived either by similar criteria or by the fact that they are adjacent to other genes or sequences. We observed no obvious clustering of elements into subgroups on the basis of how they were selected.

Fragments to be sequenced were ligated into either M13 phage (Messing and Vieira, 1982) or pEMBL plasmids (Dente *et al.*, 1983). Random length clones were prepared by a variety of methods (Dale *et al.*, 1985; Hong, 1982; Poncz *et al.*, 1983; Barnes and Bevan, 1983). Portions of sequence were also obtained using specific synthetic oligomers as primers. DNA sequencing was done by the enzymatic chain-termination method (Sanger *et al.*, 1977) using [^{35}S]dATP. Most samples were run on 6 or 8% acrylamide–urea gels (40 cm long, 0.2- to 0.4-mm wedge gels). Selected samples were also run on long (80 cm long, 0.2 mm thick) water-jacketed gels heated to 55°C (Garoff and Ansorge, 1981). Regions of sequence overlap between clones were located with the aid of homology identification programs (Queen and Korn, 1984), while data base searches were conducted using the programs of either Queen and Korn (1984) or Lipman and Pearson (1985).

The sequence of the 3′ β-globin LINE used in this analysis was determined in our laboratory. The identical element has been sequenced independently by Hattori *et al.* (1985). Five differences in 6239 nucleotides were noted between the two sequences which could not be resolved by reexamination of the sequence generated in our laboratory. The GenBank data base was searched for L1 homologous sequences that were aligned with respect to the β-globin element. Interspecies comparisons (Fig. 3) were made by aligning each pair of sequences, counting the number of matches per 60-nucleotide window, and then computing a moving average of three windows.

RESULTS

Figure 1 shows an overview of the organization of the human L1 element derived from the consensus sequence (Fig. 2). Figure 1 also indicates the relationship of the component sequences used in construction

of the consensus sequence. Each horizontal bar or group of bars per line represents DNA sequence from a separate element (as listed in Table 1). Some elements (e.g., sequences 5 and 6, Fig. 1 and Table 1) used in the consensus have regions that are scrambled relative to the 3′ β-globin element. We believe that the β-globin LINE lacks rearranged or atypical regions because its organization best accounts for the placement of blocks of sequence from the other elements. The consensus human L1 DNA sequence and predicted amino acid sequences of the two ORFs are shown in Fig. 2. Because of space contraints, the consensus is shown without the individual sequences used in its assembly (the complete data set is available upon request). At three positions in the consensus sequence there have been insertions or deletions of blocks of sequence in more than one element. The deleted regions are indicated in the consensus by dashed lines. At present, each element characterized at these positions has at least one block of sequence deleted relative to the consensus. Whether there are full-length elements corresponding to the consensus sequence remains to be shown. At a number of positions we observed alternative bases occurring with sufficient frequency that no particular base could be unambiguously chosen (Fig. 2).

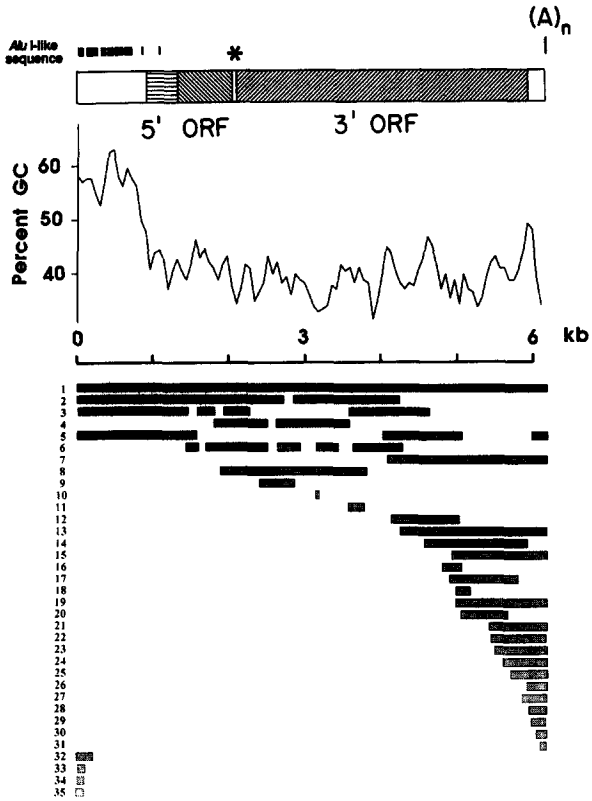
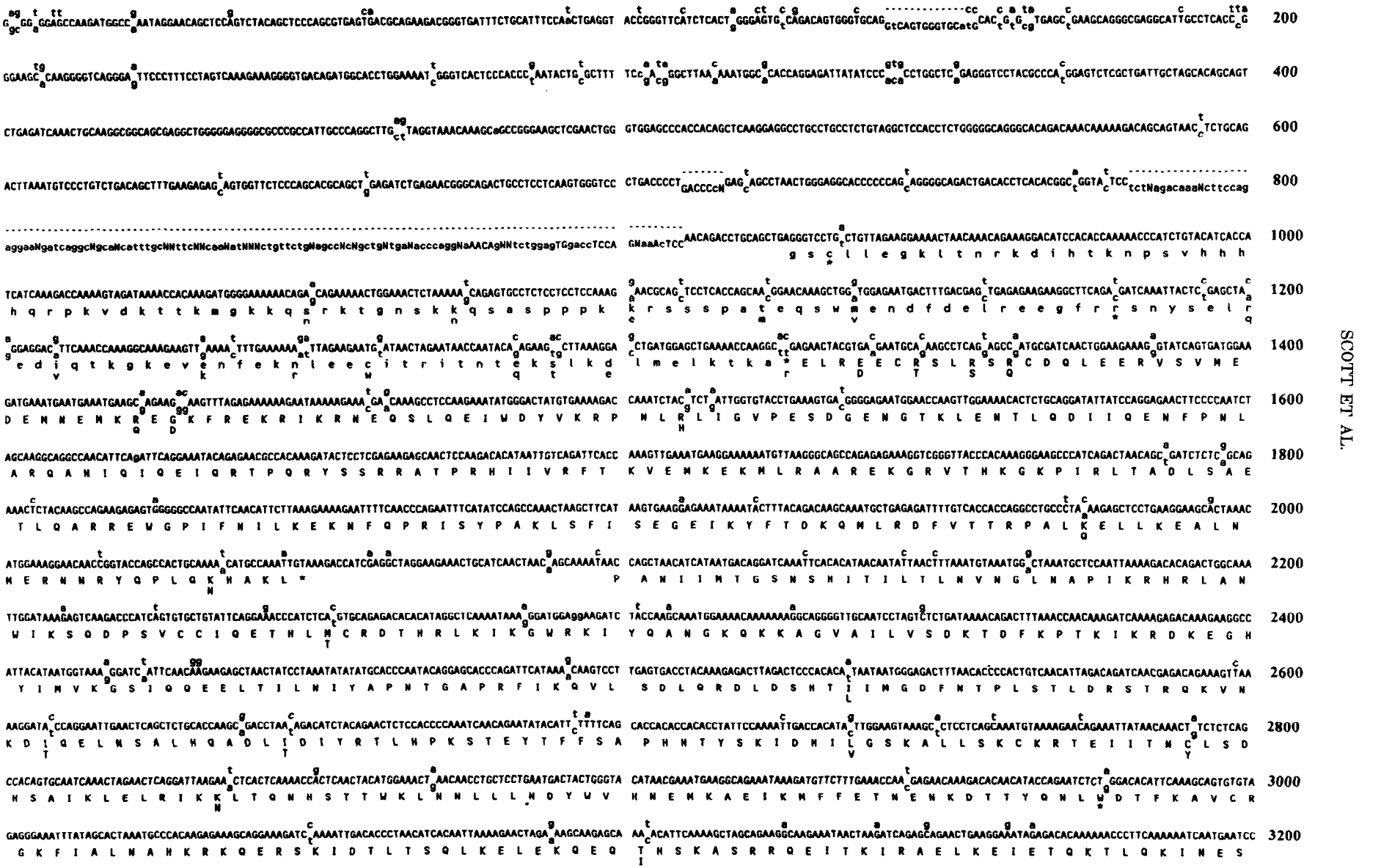


FIG. 1. A schematic diagram of the organization of the human L1 element is shown. Regions without long open reading frames are represented as open boxes. Boxes with diagonal lines correspond to conserved ORFs. A possible extension of the 5′ ORF is shown with horizontal lines. The asterisk denotes a small, poorly conserved region that coincides with sequence between the 5′ and 3′ ORFs in the human consensus. Regions with weak *Alu*I sequence homology are indicated at the 5′ end of the element. The GC content is plotted as a moving average of three 60-nucleotide blocks. Note that the highest GC content corresponds to the regions flanking the ORFs. Component sequences used in the generation of the L1*Hs* consensus are mapped relative to the consensus. The element 3′ of the β-globin gene was used as the archetype sequence. Black bars indicate elements sequenced in our laboratories. Shaded bars are published sequences. Gaps between bars along a line indicate either that the sequence was not obtained from this region or that these regions are missing from the element. The source for each sequence is presented in Table 1.

LINE Organization

As illustrated in Fig. 1 the human L1 consensus has a 5′ region of about 1 kb that is presumably noncoding. This portion of the consensus has a noticeably higher GC content (Fig. 1), a number of areas that show weak similarity to human *Alu*I sequences, and frequent termination codons in all three reading frames. The patches of homology to *Alu*I sequence, while small (14 to 32 bases with 60 to 82% homology averaging 72%), were confined largely to the noncoding region 5′ of the first ORF. This homology may reflect similar GC composition of the two classes of sequence rather than functional or evolutionary relatedness. Position 1327 has been designated as the beginning of the 5′ ORF because a terminator occurs immediately before this position in two of the four individual sequences used in the consensus. Other terminators which are present in at least 50% of the component elements occur at 1179–1181 and 933–935. We presume that these truly reflect the consensus, but sequence from additional elements will be necessary to confirm this. If these positions are not true terminators, then the 5′ ORF could be lengthened (although a longer 5′ ORF would extend into sequence that is relatively poorly conserved, as discussed below). Figure 2 shows the proposed amino acid sequence of the extended 5′ ORF in small letters. An in-frame terminator at position 924–926 occurs in all four elements used to deduce the consensus and presumably marks the most 5′ that the first ORF could be extended. This possible extension of the 5′ ORF is shown as a box with horizontal lines in Fig. 1. The reading frame beginning at 1327 is 241 codons long and would code for a protein of 218 amino acids (about 26 kDa) if the methionine encoded at position 1395–1397 is the amino terminus. A protein beginning at the methionine at nucleotides 1035–1037 in the extended 5′ ORF would be 338 amino acids, approximately 40 kDa. A data base search failed to identify significant homology between the protein predicted by the 5′ ORF and other protein sequences.

A small block of noncoding sequence (from 2049 to 2098; indicated by an asterisk in Fig. 1) was observed



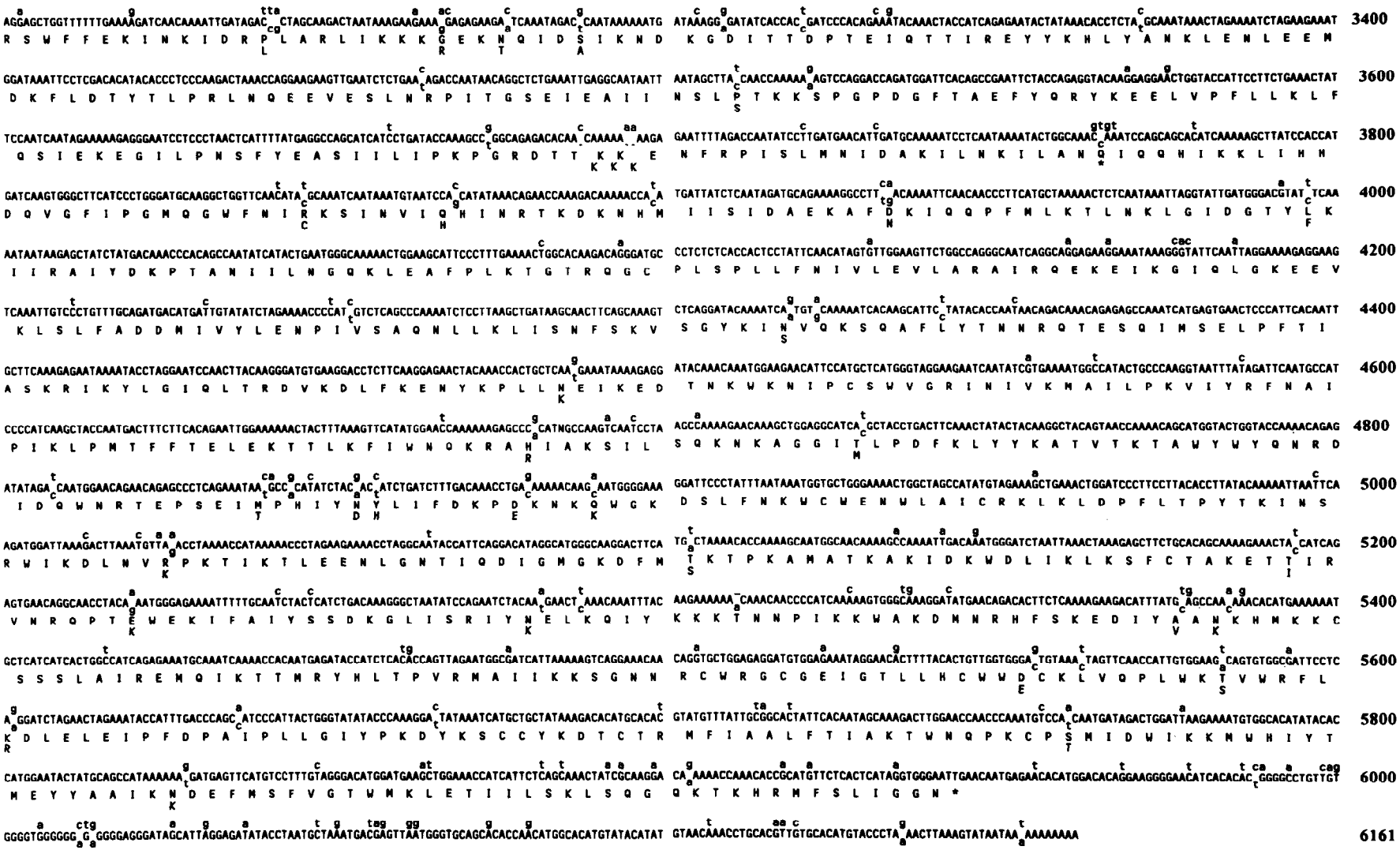


FIG. 2. The sequence of the human L1 consensus is shown. Upper- or lowercase lettering is used to indicate the relative frequency at which a particlar base occurs. A capital letter signifies that a base is present in 75% or more of the aligned sequences where fewer than six component sequences are available, or at two-thirds of positions when six or more component sequences are present. A lowercase letter indicates that the base occurred 65–74% of the time for fewer than six components, or 50% or better when multiple alternatives were seen (i.e., **t** = t, t, a or **t** = t, t, a, c) or more than six component sequences were present. If no clear choice was possible, the base is indicated by a N. Where two choices appeared with equal or nearly equal frequency (e.g., 2:2 or 2:3), both alternatives are indicated. Where a particular base is predominant but a second alternative is seen repeatedly, then a combination of upper- and lowercase is used (e.g., **c/T** = 8T, 3C, 1A). Where two bases are each displaced from the line, they occur with equal or nearly equal frequency. The amino acid sequence (using the single-letter code) for both open reading frames is indicated below the nucleic acid sequence. Small letters are used to indicate the possible extension of the 5′ ORF. The residue letter is shown directly below the first letter of the codon. Alternative residues are indicated only in cases where they are equally probable (i.e., residues predicted by codons with the less frequent nucleotides are not indicated). At some positions (especially in the 5′ ORF) several alternatives are possible, but for clarity only one is indicated. A terminator occurs as an alternative at position 3765–3767 in the 3′ ORF. A small deletion occurs as an alternative base in the 3′ ORF at position 5311 and, if true, would shift the reading frame and create a smaller protein. Two insertions also occur in the 3′ ORF at 3688 and 3695–3696. The alternative longer ORF would increase by one codon but would remain in frame.

TABLE 1

Sequences Used to Derive the Consensus

No.	Sequence name	Percentage homology with consensus	Description	Ref.
1	3′ β-globin LINE	97.1 (6026)	Genomic sequence	This paper and (19)
2	b2 LINE	94.5 (3730)	Genomic sequence	This paper
3	X63 LINE	97.1 (2925)	Genomic sequence from X chromosome	This paper
4	X65 LINE	96.8 (1633)	Genomic sequence from X chromosome	This paper
5	HUMRSKPNA	95.6 (3359)	Genomic sequence imbedded in alphoid repeat	(45)
6	ε-Globin LINE-1	72.7 (2223)	Genomic sequence adjacent to ε-globin gene	(6)
7	ε-Globin LINE-2	93.2 (2084)	Genomic sequence adjacent to ε-globin gene	(6)
8	HUMRSH3	96.8 (1905)	Genomic 1.9-kb *Hin*dIII restriction fragment	(38)
9	SHR4	98.2 (448)	Genomic sequence	(5)
10	HUMUG3PA	89.1 (46)	Genomic flanking sequence near U3.2 small nuclear RNA pseudogene	(3)
11	HUMRSKPA2	98.6 (210)	Genomic sequence	(41)
12	HUMRSKP08	97.3 (1756)	Fibroblast cDNA	(11)
13	HUMFIXG	96.8 (1914)	Genomic sequence flanking Factor IX gene	(65)
14	HUMMHDRB1	95.5 (1363)	Genomic sequence flanking MHC DR β locus	(33)
15	HUMRSKP04	96.7 (1239)	Fibroblast cDNA	(11)
16	HUMBLUR14	75.8 (264)	Genomic sequence near an *Alu*I repeat	(8)
17	HUMIFNAGS	89.3 (906)	Genomic sequence near α-interferon gene	(62)
18	HUMRSKP1	96.4 (197)	Genomic 1.8-kb *Kpn*I fragment	(42)
19	HUMRSKPE	95.7 (1184)	Genomic sequence	(58)
20	HUMIGHAD	94.5 (617)	Genomic sequence near an Ig heavy chain ε3 pseudogene	(61)
21	21-RING	97.1 (731)	Genomic sequence from chromosome 21	C. Wong *et al.*[a]
22	HUMRSKP03	94.6 (720)	Fibroblast cDNA	(11)
23	HUMRSKPT	92.4 (672)	Genomic sequence	(58)
24	HUMRSKP07	92.9 (562)	Genomic sequence	(11)
25	HUMRSKP84	83.4 (465)	Genomic sequence	(11)
26	HUMRSKP83	95.3 (254)	Genomic sequence	(11)
27	HUMPPD16	95.1 (307)	Genomic sequence	(8)
28	*Kpn*5′ε1	76.0 (229)	Genomic sequence identified 5′ of ε-globin gene	(35)
29	*Kpn*5′ε2	96.1 (154)	Genomic sequence identified 5′ of ε-globin gene	(35)
30	KT1	93.4 (137)	Genomic sequence near an HIV integration site	(43)
31	HUMRSKP2	98.8 (81)	Genomic sequence interrupting mitochondrial homologous DNA	(42)
32	PAC-32	92.8 (209)	Genomic sequence	J. W. Adams *et al.*[a]
33	HUMRSA16	92.2 (115)	Genomic sequence near *Alu*I repeat	(58)
34	HUMRSKPA1	100.0 (94)	Genomic sequence	(41)
35	HUMRSKA1	97.1 (103)	Genomic sequence	(41)
			Nonhuman L1 sequences	
36	AGMKPNRSA	91.7 (1745)	*Cercopithecus aethiops* genomic sequence	(34)
37	AGMRSKPNI	95.3 (832)	*Cercopithecus aethiops* genomic sequence interrupting alpha-satellite DNA	(60)
38	AGMKPNRSB	95.3 (595)	*Cercopithecus aethiops* genomic sequence	(34)
39	Rabbit L1[b]	69.3 (3945)	*Oryctolagus cuniculus* genomic sequence near β-globin gene	(9)
40	Rat L1[b,c]	64.7 (4560)	*Rattus norvegicus* genomic sequences	(57)
41	Mouse L1[b]	65.6 (4460)	*Mus domesticus* genomic sequence	(37)

Note. Numbers correspond to sequences diagrammed in Fig. 1 (except for nonhuman elements). *GenBank* filenames are given when available. The percentage homologies of individual sequences relative to the consensus used in Fig. 4 are given along with the length of homologous sequence in parentheses.

[a] Unpublished data.

[b] Nonhomologous 5′ and 3′ regions were excluded from this calculation.

[c] Consensus based on multiple sequences.

between the 5′ and 3′ ORFs. No obvious splice junctions were found bordering the noncoding region between ORFs. The consensus in the noncoding region is based on six sequences and it is unlikely that the separation of the two reading frames is a consequence of incomplete data. The 3′ ORF extends from position 2099 to 5942 for 1281 codons and has homology with various reverse transcriptases (Hattori *et al.*, 1986; Sakaki *et al.*, 1986). If the putative protein begins with the methionine encoded at position 2115–2117, then it

would be 1276 residues long. The 3′ ORF is followed by another nonconserved block of sequence that is 208 bp long with multiple terminators in all three reading frames. The L1 consensus terminates in a short stretch of adenines. A search of the L1*Hs* sequence in Fig. 2 failed to identify likely ribosome-binding sites near possible initiator methionines (Kozak, 1986). Further, no obvious promoter sequences were identified 5′ of either ORF.

In order to validate our observations about the human consensus, we compared published sequences from the mouse and rat to each other and to the human consensus (Fig. 3). Figure 3A shows a comparison of percentage homology of aligned mouse and rat L1 sequences calculated as a running average of three 60-nucleotide windows. A large region of conserved sequence is seen flanked by 5′ and 3′ ends which have markedly less homology. Furthermore, a short region of reduced homology (approximately 1979 to 2129) can be seen as indicated by the arrow. When the human consensus sequence is then aligned relative to the mouse–rat comparison in Fig. 3A, it can be seen that a good correspondence occurs between the predicted 5′ and 3′ noncoding regions from the L1*Hs* consensus and the regions of least homology between the two rodents. The sequence in L1*Hs* between the 5′ and 3′ ORFs aligns well with the area of reduced homology in the rodent comparison (asterisk in Fig. 3A). However, the region of reduced homology is longer than the gap between L1*Hs* ORFs because approximately the last 23 codons of the 5′ ORF and the first 10 codons of the 3′ ORF extend into nonconserved region. When either the mouse or the rat sequence is compared directly to the human sequence (Figs. 3B and 3C), the region between the human ORFs shows reduced homology in each case. The region corresponding to the 3′ ORF is relatively well conserved for both comparisons, since the sequence homology is as high as 65%. However, it appears that the length of the conserved sequence in the 5′ ORF is shorter in the human–rodent comparison than in the rat–mouse comparison. We also observed that this portion of the 5′ ORF gene product was less well conserved than the 3′ product, suggesting a possible species-specific function for this sequence. On the basis of these comparisons we are confident that the functional human L1 element (and by extension the rodent elements as well) must code for two separate proteins. Two ORFs have also been reported for the mouse L1 element, although in that species the 5′ reading frame extends (in a different register) into the 3′ ORF (Loeb *et al.*, 1986). The fact that the 5′ end of the element is poorly conserved, not only between the human and rodent elements but also between mouse and rat sequences, implies that this region does not contain large blocks of functionally important sequence (at least as measured by the criterion of evolutionary conservation).

Subfamily Analysis

Comparison of individual L1*Hs* elements to the consensus sequence (Fig. 2) indicates that subfamilies can be defined both on the basis of the association of alternative bases and by the presence or absence of blocks of sequence that occur at the same positions in different elements. Three such blocks (at L1*Hs* consensus positions 143–157, 710–716, and 782–909) can be used to define at least three subfamilies at the 5′ end as shown in Table 2. Subfamily 1 (as typified by the 3′ β-globin LINE) is characterized by a deletion from 710 to 716. Subfamily 2 (the X63 and *Kpn*-A elements) has a deletion from 782 to 909. In subfamily 3 (the b2 LINE) deletions occur from 143 to 157 and 710 to 716. PAC-32 LINE also has the subfamily 3 deletion from 143 to 157; however, since PAC-32 sequence does not extend into the region of the consensus where the other deletions are found, it is not yet possible to assign it to a subfamily. It could belong to subfamily 3 or represent a fourth subfamily. Other elements whose sequences include only a portion of this region may represent additional subfamilies.

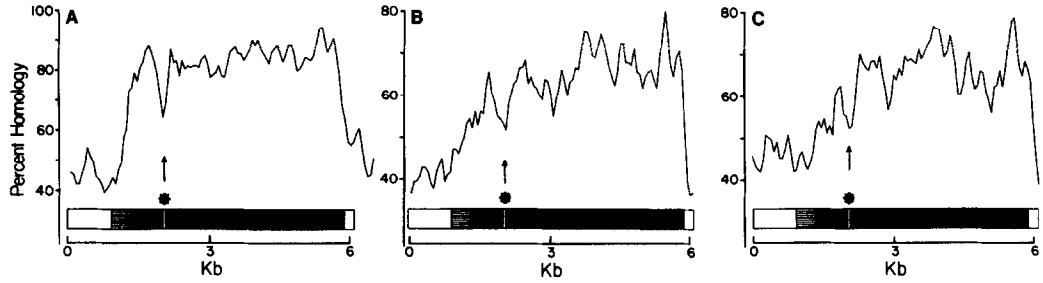


FIG. 3. Interspecies comparisons of L1 sequences are shown. DNA sequences were aligned and matched bases were counted in windows of 60 nucleotides. A moving average of three windows was calculated and plotted. Each of the plots is shown aligned with a schematic of the human element. Note that the scale for the rat–mouse comparison is changed to reflect the greater similarity between these species. Arrows above the asterisks point to the local homology minima corresponding to the evolutionarily nonconserved region between ORFs. (**A**) Mouse vs rat (elements 41 and 40, respectively; Table 1). (**B**) Human consensus vs rat. (**C**) Human consensus vs mouse.

TABLE 2

Variation at the 5′ End of L1*Hs* Defines Subfamilies

Subfamily type	L1*Hs* consensus sequence position 143–157	710–716	782–909	Example
1	+	−	+	3′ β-Globin LINE [1]
2	+	+	−	X63 LINE [3]
2	+	+	−	Kpn-A LINE [5]
3	−	−	+	b2 LINE [2]
?	−	?	?	PAC-32 LINE [32]

Note. Three major subfamilies at 5′ end of the consensus are defined by the presence (+) or absence (−) of sequence blocks (see Fig. 2). Numbers in brackets refer to sequences listed in Table 1.

However, it should be noted that all of the subfamilies can be accounted for by mixing blocks of sequence from as few as two different progenitor sequences. For example, 5′ subfamily 1 could result from a small conversion of a type 2 element by sequence from a type 3 element that includes the two blocks between 710 and 909. Additional evidence for subfamily groups can also be seen by the concordance of particular alternative bases at several positions at the 3′ end (Table 3). Several elements can be clearly assigned to either of two subfamilies, although it is important to note that evidence for mixing and the possibility of further subdivisions are suggested by the data. The existence of multiple subfamilies has also been inferred from restriction mapping studies of genomic DNA (Jubier-Maurin *et al.*, 1985; B. J. Schmeckpeper *et al.*, unpublished data). Nevertheless, the analysis presented here clearly suggests that the majority of L1 elements are derived from only a few classes of similar L1 progenitors, perhaps as few as two functional genes.

Age of L1Hs Elements Estimated by Evolutionary Analysis

Figure 4A illustrates the data presented in Table 1 on the homology between individual L1*Hs* sequences (Fig. 1) and the consensus sequence (Fig. 2). Most elements are less than 5% different from the consensus, although a few sequences are as much as 25% different. For this analysis a match to an alternative base in the consensus was accepted as perfect homology. Thus, differences among the subfamilies should not cause an overestimation of the percentage divergence of individual elements from either of the two proposed L1 progenitor sequences. During this analysis we also observed that the percentage difference by which any particular element differs from the consensus tends to be the same throughout the entire length of the element.

Figure 4B illustrates L1 sequence homology between the human consensus (Fig. 2) and sequence data from other mammals (Table 1). The first comparison is from sequence at the 5′ end of the orthologous L1 element adjacent to the gorilla β-globin gene. In 164 bp of homologous sequence there are only three differences. This value of about 2% change between man and gorilla, which diverged 8 to 10 million years (MY) ago, was also seen in other regions of the β-globin cluster (Scott *et al.*, 1984). A second comparison is available from the sequence of elements from the African green monkey *(Cercopithecus aethiops),* a cat-

TABLE 3

Variation at the 3′ End of L1*Hs* Elements Defines Two Major Subfamilies

Subfamily type	Example	L1*Hs* consensus sequence position 5105	5111	5202	5240	5282	5288	5307	5337	5348	5364	5411	5425	5502	5515	5523	5532	5566	5607	5609	5640	5688	5788	5838	5960
1	3′ β-Globin [1]	T	C	A	T	T	C	T	A	C	C	A	C	C	T	G	C	T	A	G	A	T	T	A	T
1	HUMRSKP04 [15]	g	C	A	T	T	C	C	A	C	C	A	C	C	T	G	C	T	T	G	A	C	T	G	T
1	HUMRSKP08 [12]				T	T	C	C	A	C	C	A	C	C	T	A	C	T	A	G	A	C	T	G	T
1	HUMFIXG [13]	T	C	G	T	T	C	C	A	C	C	A	C	C	T	G	C	T	A	G	A	C	T	G	T
2	ϵ-Globin 2 [7]	A	T	G	C	A	T	T	G	T	T	G	G	T	A	A	A	g	T	A	G	C	C	G	G
2	HUMIGHAD [20]	A	—	G	C	A	T	T	G	T	T	G	G	T	A	A	A								
2	HUMIFNAGS [17]	A	T	G	C	A	T	C	A	T	T	G	G	T	A	A	g	C	T	A	G				
2	HUMRSKPT [23]												G	T	A	G	A	C	T	A	G	T	C	A	G
2	HUMRSKP84 [25]																	C	T	A	G	T	C	A	G

Note. Two major subfamilies can be identified by associated bases from several elements at the 3′ end of the consensus from position 5103 to 6000. Most individual elements (Table 1) clearly fall into one of these subfamilies. Bases that do not correspond to either subfamily are shown in small letters. Discordant bases, which occur in a sequence from one subfamily but are most often seen in the other, are underlined.

tarhine primate which separated from the hominoids about 25 MY ago. Three partial monkey L1 sequences were compared to the human consensus. Overall homology with the human consensus is 95% for two and 92% for the third. If we assume that there are no, as yet uncharacterized, monkey elements more like the human consensus, then the 5% difference between monkey and human sequences can be used as an estimate of the percentage divergence that has occurred in 25 MY. The percentage similarity of rabbit, rat, and mouse sequences to the human consensus ranges from 64 to 69%. Since the mammalian radiation occurred about 80 MY ago, these species establish a third measure of the amount of sequence divergence with time. Altogether, the interspecies comparisons can be used to generate an approximation of the mammalian L1 substitution rate.

If the substitution rate calculated from the interspecies comparisons can be applied to elements within a species, then the age of particular L1*Hs* sequences can be determined. By such an analysis most are between 10 and 30 MY old. The validity of this analysis is dependent on several factors, the most important of which is how well the consensus reflects the proposed L1 structural genes. If there are more than the two L1 structural genes which we have inferred from the subfamily data, then the percentage homology will be greater and the predicted age of particular elements will be reduced. Nevertheless, a testable hypothesis follows from this model: that the estimated age of insertion of particular L1*Hs* elements can be independently determined by observing whether orthologous L1 sequences are present in the genomes of other primates. For example, of the five separate LINEs in the human β-globin gene cluster (open bars, Fig. 4A) two are apparently quite old. The two oldest elements (Table 1, Nos. 6 and 28) appear to have inserted near the ε-globin gene about 65 MY ago, while the most recent element (the 3′ β-globin LINE) may have integrated as recently as 25 MY ago. We know that the element present near the gorilla β-globin gene is the result of the same integration event as the human sequence, because its position relative to the gene is identical in both species and the first variable sequence block (position 143–155) is present in both elements. Although sequence data are not available, restriction maps and hybridization studies indicate that the 3′ β-globin LINE is also present in orangutans and chimpanzees but absent in the Old World monkeys (Fujita *et al.*, 1987). This is consistent with an age of insertion occurring approximately 25 MY ago which coincides with the estimated time of the hominoid–cattarhine divergence. Therefore, its integration may have occurred shortly after the divergence of the Old World monkey and hominoid lineages. By the same reasoning those elements least similar to the consensus, such as those near the ε-globin gene, might be present in primates with earlier divergence dates. Indeed, Rogan *et al.* (1987) have shown that at least one of the ε-globin LINEs is present in the Galago (a prosimian) which diverged from the higher primates at least 60 MY ago. Thus, the existing data, though limited, support the assumptions used to date the human elements.

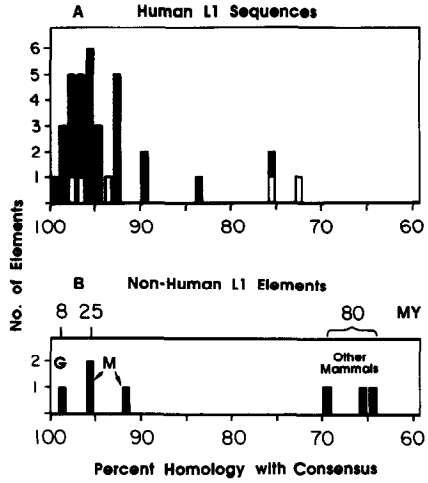


FIG. 4. (A) The percentage homology of individual human elements relative to the L1*Hs* consensus sequence (Table 1). Open bars, from left to right, represent the 3′ β-globin element, *Kpn*5′ε2, ε-globin LINE-2, *Kpn*5′ε1, and ε-globin LINE-1. (B) The percentage homology of nonhuman elements to the human consensus. Estimated divergence dates are indicated above the figure. The position of the gorilla β-globin element is indicated (G) and represents the percentage similarity between the gorilla and human orthologous sequences rather than the similarity between the gorilla element and the human consensus. The African green monkey (M) sequences (elements 36–38; Table 1) are plotted relative to the consensus and are shown at 25 MY. Other mammals include (from left to right) the rabbit, mouse, and rat (elements 39–41; Table 1) with a divergence date of 80 MY. The percentage homology for the nonprimate comparisons is calculated for sequence corresponding to a portion of the 3′ ORF from the rabbit and from the combined 5′ and 3′ ORFs of the mouse and rat. The less homologous flanking regions were excluded for this analysis.

DISCUSSION

Given that at least 1% of the human genome consists of L1 elements, how can we account for their origin and proliferation? The data presented here support the view that a relatively small number of similar L1 structural genes are the source of a much larger number of apparently nonfunctional elements which resemble processed pseudogenes. In the discussion below we make the assumption that transcription of an L1 structural gene produces a 6-kb mRNA containing two separate reading frames. However, transcription beginning at sites within the proposed structural genes is also possible. The observation that many elements appear to be truncated at the 5′ end could be accounted for by reintegration of

RNAs derived from multiple sites of initiation in the L1 structural genes or from various dispersed elements that remain transcriptionally active but do not produce a translatable mRNA. Indeed, C. T. Comey *et al.* (unpublished data) have recently shown that a portion of the 3′ β-globin LINE within the 3′ ORF has specific pol II promoter activity in an *in vitro* transcription system. Truncation of 5′ sequence is also consistent with a retroposon model in which L1 mRNAs are incompletely copied by a polymerase such as reverse transcriptase.

A retroposon model for the origin of the L1 family has often been proposed. In support of this model, several groups have analyzed the relationship between the putative L1 gene products and reverse transcriptases. As noted above, the amino acid sequence of the 3′ ORF predicts a protein very similar to that of Hattori *et al.* (1986) which contains regions of weak homology to the polymerase domain (Johnson *et al.*, 1986) of various reverse transcriptases (see Hattori *et al.* (1986) and Sakaki *et al.* (1986) for amino acid sequence alignments with these enzymes and the gene product of the 3′ ORF). Recently, Fanning and Singer (1987) constructed a mammalian amino acid consensus sequence for portions of the 3′ ORF by comparing the predicted L1 proteins from four different orders of mammals (cat, mouse, rabbit, and a combined human and monkey sequence). They concluded that the most highly conserved regions between species also corresponded to regions shown by Hattori *et al.* (1986) and Sakaki *et al.* (1986) to be homologous with various reverse transcriptases. Although these homologies are provocative, experimental evidence substantiating the retroposon model is needed. If the gene product of the 3′ ORF can be shown to have a polymerase activity, it would greatly strengthen the hypothesis that L1 elements originated by a retroposon mechanism and may be related to transposable elements or the general class of RNA viruses.

The origin of the L1*Hs* family can also be investigated by analysis of the subfamilies that we observed (Tables 2 and 3). As noted above, the study of individual elements suggests that the subfamilies can be accounted for by the mixing of blocks of sequences from as few as two structural genes. This could be the consequence of a process such as gene conversion involving DNA exchange between homologous sequences. Gene conversion in mammals often appears to be the result of homologous recombination between nearly identical, tandemly linked sequences, such as the γ-globin loci. Whenever conversion occurs at the level of the L1 structural genes, transcripts with mixed blocks of sequence would be generated. The reintegration of such transcripts would result in the accumulation of dispersed elements with the observed "patchwork" appearance. Another mechanism that would account for the apparent mixing of L1 sequences is one similar to that proposed for the recombination seen among certain RNA viruses (Keck *et al.*, 1987; Kirkegaard and Baltimore, 1986). For example, poliovirus, which contains a (+)-strand RNA genome, is copied by RNA polymerase to a (−)-strand intermediate. Evidence has accumulated that the RNA polymerase can switch to a second strand if it encounters a break during replication (Kirkegaard and Baltimore, 1986). If virus-infected cells contain two heterologous RNA (+)-strands, a hybrid (−)-strand can be formed. By analogy, if RNA recombination also occurs between L1 transcripts arising from the two proposed structural genes, then the subfamilies of these dispersed elements could be explained as retroposition events involving recombinant transcripts.

L1 elements occur in relatively high numbers in all mammals that have been studied (Burton *et al.*, 1986). They not only share homology between species, but they also show a pattern of species-specificity detectable by both DNA sequence analysis and the presence of discretely sized bands following Southern blotting with L1 probes. This pattern of within-species similarity is most obvious when distantly related species are compared, but it has also been observed when elements from different mouse species were analyzed (Jubier-Maurin *et al.*, 1985; Martin *et al.*, 1985) or when monkey and human elements were aligned (A. F. Scott, unpublished data). The relative similarity of the L1*Hs* elements used in constructing the consensus sequence (as manifested by the clustering of elements in Fig. 4A) is another measure of the within-species homogeneity of this repeated DNA family. At least three models can account for the species-specificity. The first would require the generation of the L1 repeated DNA family by the sudden amplification in different taxa of a small number of functional L1 elements, as proposed for the SINEs (Weiner *et al.*, 1986). If true, such amplifications would have had to occur recently in each species in order that individual elements would not have had sufficient time to diverge by the accumulation of random base substitutions. However, a difficulty in accounting for all intraspecies similarity of L1 sequences by this model is the necessary assumption that independent amplification events have occurred in each mammalian species. It is possible that such a predisposition to generate large numbers of dispersed elements is an intrinsic characteristic of the functional L1 genes. However, the observation (Fig. 4A) that some of the L1*Hs* elements are quite old argues against a single burst of amplification as the source of the L1*Hs* repeated DNA family.

A second model that would account for the within-species similarity of L1 sequences involves gene con-

version. These conversion events would have to affect a large fraction of the thousands of dispersed L1 elements in order to maintain the observed intraspecies homogeneity of this repeated DNA family. Such widespread conversion could occur by an RNA-mediated mechanism involving transcripts newly generated from the structural genes or by exchange of DNA sequences between various dispersed elements. In both yeast and fungi, conversion between dispersed tRNA genes occurs at relatively high frequency (Heyer *et al.*, 1986) and may involve RNA intermediates (Doolittle, 1985). In the case of the L1*Hs* family, conversion of dispersed elements is unlikely, because the alternative bases in Fig. 2 and Table 3 cluster into two groups with relatively little discordance. In addition, the percentage by which a given element differs from the consensus appears to be the same throughout the length of that element. This observation is opposite to the expectation that conversion of dispersed elements might produce LINEs with "new" sequence imbedded in otherwise "old" elements.

If neither conversion of dispersed elements nor recent bursts of amplification are adequate to account for within-species homogenization, how else can we account for this process? A third model that might explain the characteristics of the L1 repeats would require their continuous generation in each mammalian lineage. Indeed, evidence for new L1 integrations has been presented from studies of a canine *myc* gene (Katzir *et al.*, 1985) and from the rat *Mlvi-2* (Economou-Pachnis *et al.*, 1985), immunoglobin heavy chain (Economou-Pachnis *et al.*, 1985), and insulin (Lakshmikumaran *et al.*, 1985) loci. In each case allelic forms can be distinguished by the presence or absence of an associated element. If L1 elements are continuously produced, then we would expect an even distribution of sequences with a range of homology to the consensus instead of the clustering of "new" elements seen in Fig. 4A. The observed distribution could be accounted for either by a bias of ascertainment, so that older elements are less likely to be detected with probes most like the consensus, or by processes that remove DNA from the genome. As noted above, we detected no obvious bias of ascertainment of the elements in Table 1. We favor, therefore, the latter hypothesis. Support for this process can be found from several observations including differences in the lengths of intervening sequences and in the organization of nonfunctional DNA adjacent to orthologous genes in different species. Likewise, the fact that about 5% of mutant genes characterized in a variety of human genetic disorders are the result of deletions (Kazazian and Antonarakis, 1987) suggests that loss of large blocks of sequence is not uncommon. If removal of DNA from the genome occurs with moderate frequency, then as a consequence L1 sequences will be removed. Therefore, DNA loss coupled with new integration events and the decay of older L1 sequences by base substitution could produce the appearance of relatively recent L1 amplifications. Clearly, the mechanism or combination of mechanisms that accounts for the generation and evolution of these sequences remains to be fully understood.

The L1 consensus sequence suggests that transcripts of the L1 structural genes are distinct from typical processed eukaryotic mRNAs in that they have two ORFs separated by a noncoding region. The two reading frames and the reverse transcriptase homology of the 3′ ORF are characteristics shared with retroviruses, the *Ty* elements of yeast (Fink *et al.*, 1986), and the *I* factor transposable elements of *Drosophila* (Fawcett *et al.*, 1986). The possibility arises that all of these sequences are functional analogs and may have had a common evolutionary origin. If the 3′ ORF codes for a reverse transcriptase, then, as with retroviruses, that activity might enhance the copying and reintegration of L1 transcripts into the genome (Temin, 1985; Hattori *et al.*, 1986). The human L1 consensus sequence differs from that of typical retroviruses in that it lacks terminal repeats and has 5′ and 3′ nontranslated and nonconserved flanking regions together with an adenine-rich 3′ terminus. Although the origin of the L1 sequences is unclear at present, it is apparent that this class of DNA is very old. Perhaps the original L1 structural gene was derived from the integration of an RNA virus or transposable element prior to the mammalian radiation, and with selection the sequence was modified to become a conventional eukaryotic gene which continues to retain features of its origin. Assuming a retroposon mechanism for the generation of the high copy number for L1 elements, it follows that the structural genes may be transcribed in germ line cells. The demonstration of a large polyadenylated L1 RNA in certain teratocarcinoma cells (Skowronski and Singer, 1985) suggests that the structural genes may be expressed in ova. Because mammalian eggs are present in females at birth and can be maintained for long periods before fertilization, their DNA may represent excellent targets for invasion by sequences transcribed in these cells (Wagner, 1986). Although a functional role for L1 sequences remains unknown, it is not unreasonable to expect that their integration in or near genes might alter the expression of adjacent sequences and that they might serve as insertional mutagens. While single elements may have a small influence on neighboring genes, the cumulative effect of thousands of elements throughout the genome may be profound and, in part, account for the rapid rate of phenotypic change seen in mammalian evolution.

ACKNOWLEDGMENTS

We thank A. Nienhuis and J. Adams for providing the recombinant clones for the 3′ β-globin element and the sequence of PAC-32.

The unpublished sequence of a partial L1 element from chromosome 21 was provided by C. Wong, S. Antonarakis, S. Trusko, and H. Kazazian. Additional technical help was provided by J. Campbell. We thank M. Singer and T. Fanning for a helpful discussion of some of the ideas expressed in this manuscript and P. Rogan for a copy of his unpublished manuscript. This work was supported, in part, by NIH Grants GM28931 (A.F.S.) and HD17161 (K.D.S.) and Predoctoral Training Grant GM07814.

REFERENCES

1. ADAMS, J. W., KAUFMAN, R. E., KRETSCHMER, P. J., HARRISON, M., AND NIENHUIS, A. W. (1980). A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. *Nucleic Acids Res.* **8:** 6113–6128.
2. BARNES, W. M., AND BEVAN, M. (1983). Kilo-sequencing: An ordered strategy for rapid DNA sequence data acquisition. *Nucleic Acids Res.* **11:** 349–368.
3. BERNSTEIN, L. B., MOUNT, S. M., AND WEINER, A. M. (1983). Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. *Cell* **32:** 461–472.
4. BURTON, F. H., LOEB, D. D., VOLIVA, C. F., MARTIN, S. L., EDGELL, M. H., AND HUTCHISON, C. A. (1986). Conservation throughout mammalia and extensive protein-coding capacity of the highly repeated DNA long interspersed sequence one. *J. Mol. Biol.* **187:** 291–304.
5. CITRON, B. A., CHAUDARY, P. V., RAO, D. N., AND KAUFMAN, S. (1986). Evidence for transcription and potential translation of the human 1.9 kb *Hin*dIII repetitive element. *Nucleic Acids Res.* **14:** 3137–3142.
6. COLLINS, F. S., AND WEISSMAN, S. M. (1984). The molecular genetics of human hemoglobin. *Prog. Nucleic Acid Res. Mol. Biol.* **31:** 315–462.
7. DALE, R. M. K., MCCLURE, B. A., AND HOUCHINS, J. P. (1985). A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: Application to sequencing the corn mitochondrial 18 S rDNA. *Plasmid* **13:** 31–40.
8. DEININGER, P. L., JOLLY, D. J., RUBIN, C. M., FRIEDMANN, T., AND SCHMID, C. W. (1981). Base sequence studies of 300 nucleotides renatured repeated human DNA clones. *J. Mol. Biol.* **151:** 17–33.
9. DEMERS, G. W., BRECH, K., AND HARDISON, R. C. (1986). Long interspersed L1 repeats in rabbit DNA are homologous to L1 repeats of rodents and primates in an open-reading-frame region. *Mol. Biol. Evol.* **3:** 179–190.
10. DENTE, L., CESARENI, G., AND CORTESE, R. (1983). pEMBL: A new family of single stranded plasmids. *Nucleic Acids Res.* **11:** 1645–1655.
11. DIGIOVANNI, L., HAYNES, S. R., MISRA, R., AND JELINEK, W. R. (1983). *Kpn*I family of long-dispersed repeated DNA sequences of man: Evidence for entry into genomic DNA of DNA copies of poly(A)-terminated *Kpn*I RNAs. *Proc. Natl. Acad. Sci. USA* **80:** 6533–6537.
12. DOOLITTLE, W. F. (1985). RNA-mediated gene conversion. *Trends Genet.* **1:** 64–65.
13. ECONOMOU-PACHNIS, A., LOHSE, M. A., FURANO, A. V., AND TSICHLIS, P. N. (1985). Insertion of long interspersed repeated elements at the *Igh* and *Mlvi-2* (Moloney leukemia virus integration 2) loci of rats. *Proc. Natl. Acad. Sci. USA* **82:** 2857–2861.
14. FANNING, T., AND SINGER, M. (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucleic Acids Res.* **15:** 2251–2260.
15. FAWCETT, D. H., LISTER, C. K., KELLET, E., AND FINNEGAN, D. J. (1986). Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs. *Cell* **47:** 1007–1015.
16. FINK, G. R., BOEKE, J. D., AND GARFINKEL, D. J. (1986). The mechanism and consequences of retrotransposition. *Trends Genet.* **2:** 118–123.
17. FUJITA, A., HATTORI, M., TAKENAKA, O., AND SAKAKI, Y. (1987). The L1 family (*Kpn*I family) sequence near the 3′ end of human β-globin gene may have been derived from an active L1 sequence. *Nucleic Acids Res.* **15:** 4007–4020.
18. GAROFF, H., AND ANSORGE, W. (1981). Improvements of DNA sequencing gels. *Anal. Biochem.* **115:** 450–457.
19. HATTORI, M., HIDAKA, S., AND SAKAKI, Y. (1985). Sequence analysis of a *Kpn*I family member near the 3′ end of human β-globin gene. *Nucleic Acids Res.* **13:** 7813–7827.
20. HATTORI, M., KUHARA, S., TAKENAKA, O., AND SAKAKI, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. *Nature (London)* **321:** 625–628.
21. HEYER, W.-D., MUNZ, P., AMSTUTZ, H., AEBI, R., GYSLER, C., SCHUCHERT, P., SZANKASI, P., LEUPOLD, U., KOHLI, J., GAMULIN, V., AND SOLL, D. (1986). Inactivation of nonsense suppressor transfer RNA gene in *Schizosaccharomyces pombe:* Intergenic conversion and hot spots of mutation. *J. Mol. Biol.* **188:** 343–353.
22. HONG, G. F. (1982). A systematic DNA sequencing strategy. *J. Mol. Biol.* **158:** 539–549.
23. JOHNSON, M. S., MCCLURE, M. A., FENG, D. F., GRAY, J., AND DOOLITTLE, R. F. (1986). Computer analysis of retroviral *pol* genes: Assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. *Proc. Natl. Acad. Sci. USA* **83:** 7648–7652.
24. JUBIER-MAURIN, V., DOD, B. J., BELLIS, M. PIECHACZYK, M., AND ROIZES, G. (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.
25. KATZIR, N., RECHAVI, G., COHEN, J. B., UNGER, T., SIMONI, F., SEGAL, S., COHEN, D., AND GIVOL, D. (1985). "Retroposon" insertion into the cellular oncogene c-*myc* in canine transmissible venereal tumor. *Proc. Natl. Acad. Sci.* **82:** 1054–1058.
26. KAZAZIAN, H. H., AND ANTONARAKIS, S. E. (1987). The varieties of mutation. *In* "Progress in Medical Genetics" (A. Motulsky, C. Epstein, and B. Childs, Eds.), Vol. 7, in press.
27. KECK, J. G., STOHLMAN, S. A., SOE, L. H., MAKINO, S., AND LAI, M. M. (1987). Multiple recombination sites at the 5′-end of murine coronavirus RNA. *Virology* **156:** 331–341.
28. KIMMEL, B. E., OLE-MOIYOI, O. K., AND YOUNG, J. R. (1987). Ingi, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. *Mol. Cell. Biol.* **7:** 1465–1475.
29. KIRKEGAARD, K., AND BALTIMORE, D. (1986). The mechanism of RNA recombination in poliovirus. *Cell* **47:** 433–443.
30. KOLE, L. B., HAYNES, S. R., AND JELINEK, W. (1983). Discrete and heterogeneous high molecular weight RNAs complementary to a long dispersed repeat family (a possible transposon) of human DNA. *J. Mol. Biol.* **165:** 257–286.
31. KOZAK, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292.
32. LAKSHMIKUMARAN, M. S., D'AMBROSIO, E., LAIMINS, L. A., LIN, D. T., AND FURANO, A. V. (1985). Long interspersed repeated DNA (LINE) causes polymorphism at the rat insulin 1 locus. *Mol. Cell. Biol.* **5:** 2197–2203.

33. LARHAMMAR, D., SERVENIUS, B., RASK, L., AND PETERSON, P. A. (1985). Characterization of an HLA DR beta pseudogene. *Proc. Natl. Acad. Sci. USA* **82:** 1475–1479.

34. LERMAN, M. I., THAYER, R. E., AND SINGER, M. F. (1983). *Kpn*I family of long interspersed repeated DNA sequences in primates: Polymorphism of family members and evidence for transcription. *Proc. Natl. Acad. Sci. USA* **80:** 3966–3970.

35. LI, Q., POWERS, P. A., AND SMITHIES, O. (1985). Nucleotide sequence of 16-kilobase pairs of DNA 5′ to the human ϵ-globin gene. *J. Biol. Chem.* **260:** 14901–14910.

36. LIPMAN, D. J., AND PEARSON, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227:** 1435–1441.

37. LOEB, D. D., PADGETT, R. W., HARDIES, S. C., SHEHEE, W. R., COMER, M. B., EDGELL, M. H., AND HUTCHISON, C. A. (1986). The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retrotransposons. *Mol. Cell. Biol.* **6:** 168–182.

38. MANUELIDIS, L. (1982). Nucleotide sequence definition of a major human repeated DNA, the *Hin*dIII 1.9 kb family. *Nucleic Acids Res.* **10:** 3211–3219.

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

40. MESSING, J., AND VIEIRA, V. (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. *Gene* **19:** 269–277.

41. MIYAKE, T., MIGITA, K., AND SAKAKI, Y. (1983). Some *Kpn*I family members are associated with the *Alu* family in the human genome. *Nucleic Acids Res.* **11:** 6837–6846.

42. NOMIYAMA, H., TSUZUKI, T., WAKASUGI, S., FUKUDA, M., AND SHIMADA, K. (1984). Interruption of a human nuclear sequence homologous to mitochondrial DNA by a member of the *Kpn*I 1.8 kb family. *Nucleic Acids Res.* **12:** 5225–5234.

43. OKAMOTO, T., REITZ, M. S., CLARKE, M. F., JAGODZINSKI, L. L., AND WONG-STAAL, F. (1986). Activation of a novel *Kpn*I transcript by downstream integration of a human T-lymphotropic virus Type I provirus. *J. Biol. Chem.* **261:** 4615–4619.

44. PONCZ, M., SCHWARTZ, E., BALLANTINE, M., AND SURREY, S. (1983). Nucleotide sequence analysis of the delta beta-globin gene region in humans. *J. Biol. Chem.* **258:** 11599–11609.

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

46. QUEEN, C., AND KORN, L. J. (1984). A comprehensive sequence analysis program for the IBM personal computer. *Nucleic Acids Res.* **12:** 581–599.

47. ROGAN, P. K., PAN, J., AND WEISSMAN, S. M. (1987). L1 repeat elements in the human ϵ–γ-globin gene intergenic region: Sequence analysis and concerted evolution within this family. *Mol. Biol. Evol.* **4:** 327–342.

48. SAKAKI, Y., HATTORI, M., FUJITA, A., YOSHIOKA, K., KUHARA, S., AND TAKENAKA, O. (1986). The LINE-1 family of primates may encode a reverse transcriptase-like protein. *Cold Spring Harbor Symp. Quant. Biol.* **51:** 465–469.

49. SANGER, F., NICKLEN, S., AND COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74:** 5463–5467.

50. SCHMECKPEPER, B. J., SCOTT, A. F., AND SMITH, K. D. (1984). Transcripts homologous to a long repeated DNA element in the human genome. *J. Biol. Chem.* **259:** 1218–1225.

51. SCHMECKPEPER, B. J., SMITH, K. D., DORMAN, B. P., RUDDLE, F. H., AND TALBOT, C. C. (1979). Partial purification and characterization of DNA from the human X chromosome. *Proc. Natl. Acad. Sci. USA* **76:** 6525–6528.

52. SCHMECKPEPER, B. J., WILLARD, H. F., AND SMITH, K. D. (1981). Isolation and chracterization of cloned human DNA fragments carrying reiterated sequences common to both autosomes and the X chromosome. *Nucleic Acids Res.* **9:** 1853–1872.

53. SCOTT, A. F., HEATH, P., TRUSKO, S., BOYER, S. H., PRASS, W., GOODMAN, M., CZELUSNIAK, J., CHANG, L.-Y. E., AND SLIGHTOM, J. L. (1984). The sequence of the gorilla fetal globin genes: Evidence for multiple gene conversions in human evolution. *Mol. Biol. Evol.* **1:** 371–389.

54. SHAFIT-ZAGARDO, B., BROWN, F. L., ZAVODNY, P. J., AND MAIO, J. J. (1983). Transcription of the *Kpn*I families of long interspersed DNAs in human cells. *Nature (London)* **304:** 277–280.

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

56. SKOWRONSKI, J., AND SINGER, M. F. (1985). Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. *Proc. Natl. Acad. Sci. USA* **82:** 6050–6054.

57. SOARES, M. B., SCHON, E., AND EFSTRATIADIS, A. (1985). Rat LINE 1: The origin and evolution of a family of long interspersed middle repetitive DNA elements. *J. Mol. Evol.* **22:** 117–133.

58. SUN, L., PAULSON, K. E., SCHMID, C. W., KADYK, L., AND LEINWAND, L. (1984). Non-*Alu* family interspersed repeats in human DNA and their transcriptional activity. *Nucleic Acids Res.* **12:** 2669–2690.

59. TEMIN, H. M. (1985). Reverse transcription in the eukaryotic genome: Retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* **2:** 455–468.

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

61. UEDA, S., NAKAI, S., NISHIDA, Y., HISAJIMA, H., AND HONJO, T. (1982). Long terminal repeat-like elements flank a human immunoglobin epsilon pseudogene that lacks introns. *EMBO J.* **1:** 1539–1544.

62. ULLRICH, A., GRAY, A., GOEDDEL, D. V., AND DULL, T. J. (1982). Nucleotide sequence of a portion of human chromosome 9 containing a leukocyte interferon gene cluster. *J. Mol. Biol.* **156:** 467–486.

63. WAGNER, M. (1986). A consideration of the origin of processed pseudogenes. *Trends Genet.* **2:** 134–137.

64. WEINER, A. M., DEININGER, P. L., AND EFSTRATIADIS, A. (1986). Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55:** 631–661.

65. YOSHITAKE, S., SCHACH, B. G., FOSTER, D. C., DAVIE, E. W., AND KURACHI, K. (1985). Nucleotide sequence of the gene for human factor IX (anti-hemophilic factor B). *Biochemistry (USA)* **24:** 3736–3750.