The L1 Family of Long Interspersed Repetitive DNA in Rabbits: Sequence, Copy Number, Conserved Open Reading Frames, and Similarity to Keratin

G. William Demers,* Michael J. Matunis,† and Ross C. Hardison

Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Summary. The L1 family of long interspersed repetitive DNA in the rabbit genome (L10c) has been studied by determining the sequence of the five L1 repeats in the rabbit β-like globin gene cluster and by hybridization analysis of other L1 repeats in the genome. L10c repeats have a common 3' end that terminates in a poly A addition signal and an A-rich tract, but individual repeats have different 5' ends, indicating a polar truncation from the 5' end during their synthesis or propagation. As a result of the polar truncations, the 5' end of L10c is present in about 11,000 copies per haploid genome, whereas the 3' end is present in at least 66,000 copies per haploid genome. One type of L10c repeat has internal direct repeats of 78 bp in the 3' untranslated region, whereas other L10c repeats have only one copy of this sequence. The longest repeat sequenced, L10c5, is 6.5 kb long, and genomic blot-hybridization data using probes from the 5' end of L10c5 indicate that a full length L10c repeat is about 7.5 kb long, extending about 1 kb 5' to the sequenced region. The L10c5 sequence has long open reading frames (ORFs) that correspond to ORF-1 and ORF-2 described in the mouse L1 sequence. In contrast to the overlapping reading frames seen for mouse L1, ORF-1 and ORF-2 are in the same reading frame in rabbit and human L1s, resulting in a discistronic structure. The region between the likely stop codon for ORF-1 and the proposed start codon for ORF-2 is not conserved in interspecies comparisons, which is further evidence that this short region does not encode part of a protein. ORF-1 appears to be a hybrid of sequences, of which the 3' half is unique to and conserved in mammalian L1 repeats. The 5' half of ORF-1 is not conserved between mammalian L1 repeats, but this segment of L10c is related significantly to type II cytoskeletal keratin.

Key words: L1 — Long repetitive DNA — Rabbits — Genome evolution

Introduction

The repeated DNA sequences that are dispersed throughout eukaryotic genomes have been divided into two classes (reviewed by Weiner et al. 1986). Both classes appear to transpose by an RNA intermediate, and the insertion of either class of repeated DNA generates short flanking direct repeats at the target site—hallmarks of transposition first recognized in prokaryotes. One class of repeated DNA resembles retroviruses in that members of this class are flanked by long terminal repeats (LTRs). This class includes the yeast Ty-1 repeat, the Drosophila copia repeat, and the human THE1 repeat (Paulson et al. 1985). Another class of repeated sequences resembles processed pseudogenes and lacks long terminal repeats (LTRs). This second class of repeats has been termed retroposons (Rogers 1983), nonviral retroposons (Weiner et al. 1986), and non-LTR retrotransposons (Xiong and Eickbush 1988). In this paper, this second class of RNA-transposed repeats will be called retroposons. Two groups of retroposons have been identified based on their length: the short interspersed repeats, or SINEs, that are less than 500 bp long, and the long inter-
Analysis indicates that the L1 sequence is present in rats (Economou-Pachnis et al. 1985; Soares et al. 1985). The L1 sequence has been identified in a wide variety of species including primates (Lerman et al. 1986), eats (Fanning and Singer 1987), and rabbits (Economou-Pachnis et al. 1985). In this paper, the rabbit L1 repeats are characterized more thoroughly, and the similarities and differences of L1 sequences between species are explored further. Interspecies comparisons reinforce the conclusion that the L1 repeat has two ORFs that are conserved for their protein-coding capacity. However, the region between the two ORFs is not conserved among species, and this observation is used to indicate possible start and stop codons for the ORFs. ORF-1 encodes a composite protein, and the 5' half of ORF-1 from L10c is related to type II cytoskeletal keratin.

Materials and Methods

Subcloning and Sequencing of L10c Repeats. The sequenced members of the L10c family were from the rabbit β-like globin gene cluster isolated by Lacy et al. (1979). Interspersed repetitive DNA was identified by Shen and Maniatis (1980) by hybridization and heteroduplex mapping. The five L1 members (Demers et al. 1986) were sequenced by dideoxynucleotide chain termination reactions (Sanger et al. 1977) using subclones in M13 phages as templates (Messing 1983).

Analysis of DNA Sequences. Sequence matches were first identified by dot plots generated by the computer program MATRIX (Zweig 1984). This provides a graphical display of sequence similarity that plots matches (forward similarity) of 23 out of 30 bases. Similar sequences were then aligned by the computer program NUCALN (Wilbur and Lipman 1983) using the parameters K-tuple = 3, window size = 20, gap penalty = 7. The protein sequence databases at the Protein Identification Resource (National Biomedical Research Foundation) were searched using the FASTp program (Lipman and Pearson 1985). The statistical significance of the similarities found by FASTp were tested using the program RDF (National Biomedical Research Foundation); this program scrambles the target sequence (revealed by FASTp) into 20 shuffled sequences and computes the mean similarity score for the shuffled sequence with the test sequence (in this case, ORF-1 of L10c). The similarity score for the match between the true sequences is compared with the mean score for the shuffled sequences in terms of the number of standard deviations that separate them.

Genomic Blot-Hybridization. Rabbit genomic DNA was analyzed by Southern (1975) blot-hybridization using a modification of the hybridization procedure of Church and Gilbert (1984). Rabbit genomic DNA was digested by restriction enzymes and size fractionated on an 0.8% agarose gel before being transferred to a nylon filter (Nytran, Schleicher & Schuell). The hybridization solution was 0.5 M sodium phosphate, pH 7.2, and 5% sodium dodecyl sulfate. The blots were hybridized at 60°C overnight and then washed four times with 40 mM sodium phosphate, pH 7.2,
Fig. 2. Sequence alignment of the L1 repeats from the rabbit β-like globin gene cluster. The sequences of L1Oc5 and L1Oc4 were aligned by the program NUCALN (Wilbur and Lipman 1983). The other repeats, L1Oc3, L1Oc2, and L1Ocl, were placed in the alignment by inspection. The numbers at the right are for L1Oc5, the prototypical rabbit L1. The flanking direct repeats of L1Ocl and L1Oc3 are in bold letters. The internal direct repeats in L1Oc5 and L1Oc1 are in lower-case letters. The conserved stop codon and the RNA polymerase II polyadenylation signal are underlined. Continued on pages 6 and 7.
were labeled with np by nick-translation (Rigby et al. 1977) of DNA fragments or recombinant plasmids from L1 Oc5 or L1 Oc4. 

Determination of Copy Number. The copy number of L1 Oc5 was determined by plaque hybridization. Regions of L1 Oc5 were labeled with 32P by nick-translation (Rigby et al. 1977) of DNA fragments or recombinant plasmids from L1 Oc5 or L1 Oc4.

1 mM EDTA, and 1% sodium dodecyl sulfate. The wash solution was heated to 68°C before washing at room temperature. Probes were labeled with 33P by nick-translation (Rigby et al. 1977) of DNA fragments or recombinant plasmids from L1 Oc5 or L1 Oc4.
conditions as in the Southern blot analysis. The ratio of percentage of plaques that hybridized to the percentage of the rabbit genome in one λ clone gives the approximate copy number of the region. The average size of an insert in this λ library is 17 kb (Maniatis et al. 1978). Thus, the fraction of the rabbit genome per phage is 17 × 10^3 / 3 × 10^9 or 5.7 × 10^-4%. The fact that 96% of the phage in the library have rabbit DNA (Maniatis et al. 1978) was also taken into account.

Rodent and Human L1 Sequences. The mouse L1 sequence, L1Md2 (Loeb et al. 1986), and the rat L1 sequence, L1Rn or LINE3 (D’Ambrosio et al. 1986) are randomly isolated L1 members from their respective genomes. The human L1 sequence, L1Hs-TBG41, is located 3.3 kb 3’ to the human β-globin gene (Haftor et al. 1985). A consensus L1Hs sequence (Scott et al. 1987) was used in the analysis of ORF-I in Fig. 8.

Results

Comparisons among the Rabbit L1 Repeats in the β-like Globin Gene Cluster

The interspersion of repetitive sequences among the rabbit β-like globin genes is shown in Fig. 1. The genes ε and γ (formerly β4 and β3) are expressed in embryonic development (Rohrbaugh and Hardison 1983), δ (ψ2) is an inactive pseudogene (Lacy and Maniatis 1980), and β (β1) is expressed in fetal and adult life (Hardison et al. 1979; Rohrbaugh et al. 1985). The 5’ to 3’ orientations of the proposed RNA intermediates of the repetitive elements are indicated by the arrows in Fig. 1; the A-rich tracts are at the 3’ ends. The sequences of the five L1Oc repeats are presented in Fig. 2. L1Oc5 is adjacent to L1Oc4 (Fig. 1), so the last nucleotide in the L1Oc5 sequence is followed by the first nucleotide in the L1Oc5 sequence (Fig. 2, positions 3306-3973). This is the 5’ end of L1Oc5 because a similar sequence is present in L1Oc5 because a similar sequence is present in L1HS-TBG41 (Fig. 2). The 667 bp (Fig. 2, positions 3306-3973) is clearly a deletion from L1Oc5 and not an insertion in L1Oc5 because a similar sequence is present in both mouse and human L1s (Demers et al. 1986). L1Oc5 will be the prototypical rabbit L1 for further analysis because it is the longest and has no extensive internal deletions. The 5’ end of L1Oc5 is also the end of the cloned region of the rabbit β-like globin gene cluster (see Fig. 1). Only two of the
individual repeats, L1Oc4 and L1Oc5, contain sequences for the ORF region (Demers et al. 1986). The other three repeats contain part or all of the 3' untranslated region.

L1Oc5 and L1Oc1 have internal direct repeats of 78 bp in the 3' untranslated region. One copy of the repeat is at positions 6015–6092 and the other is at positions 6212–6289 (lower case letters in Fig. 2). L1Oc4 and L1Oc3 have only one copy of this 78-bp sequence, and they do not contain the sequence between the 78-bp direct repeat (present in L1Oc5 and L1Oc1). Thus, the class of L1Oc repeats containing one copy of the 78-bp sequence could be derived from the class containing two copies by a deletion between the two 78-bp sequences. Another example of a sequence rearrangement is the apparent insertion of 34 bp into L10e4 between positions 5701–5702 of L1Oc5.

Most members of the L1Oc family are flanked by short direct repeats. L1Oc1 and L1Oc2 are flanked by direct repeats of 9 bp and 5 bp, respectively (Fig. 2). The flanking direct repeats differ for the two individual L1 repeats, showing that they are not part of the L1 sequence. Such flanking direct repeats are often generated by insertion of transposable elements presumably by repair of a staggered break at the target site. The flanking direct repeats for L1Oc4 and L1Oc3 cannot be identified with the available data. The 5' end of L1Oc5 has not been cloned. Because L1Oc5 is juxtaposed to L1Oc4, it is possible that L1Oc5 may have inserted into L1Oc4, in which case the 5' end of L1Oc4 is also not available. The only other L1 member, L1Oc3, does not have obvious flanking direct repeats generated by a duplication of the target site. The sequence GTTAAAAAAA found just 3' to the polyadenylation site (positions 6438–6447) is also found upstream from L1Oc3 (Margot et al. 1989). However, because the sequence GTT(A), (or a slight variation of it) is also found in all of the other L1 sequences just 3' to the polyadenylation signal, it is likely not to have been generated by a target site duplication around L1Oc3. This terminal repetition could be generated by insertion of a circular form of L1 by homologous recombination into a GTT(A), sequence at the target site.

The structural features revealed by the alignment and comparison of the L1 members from the rabbit b-like globin gene cluster are summarized in Fig. 3. The B, E, and D repeats identified by Shen and Maniatis (1980) are also aligned with their position in the L1Oc sequence. The D repeat is confined to the 3' untranslated region, whereas the B repeat and most of the E repeat are from the ORF region. L1Oc1 begins immediately after the conserved translation stop codon. Figure 3 also illustrates the internal sequence rearrangements described above.

**Copy Number of Different Regions of L1Oc**

The diagram of L1Oc repeats in Fig. 3 shows that they are truncated at a variable distance from the 5' end of the longest elements. This truncation from the 5' ends is common in the whole population of L1 repeats, as demonstrated by using four regions of L1Oc5 as probes against the rabbit genomic DNA library in a plaque hybridization assay. By counting the number of plaques that hybridized to a given probe, the approximate copy number of each region of the L1Oc5 repeat was determined (see Materials and Methods). As shown in Fig. 4, the 5'-most region of L1Oc5 is represented about 11,000 times in the haploid genome of the rabbit, and regions of L1 located more 3' are found more frequently. The largest increase in copy number is seen in the region from positions 4351 to 6004 that includes the 3'
untranslated region; this region is represented at least 66,000 times. However, the relationship between the length of the repeat and the copy number is not linear; only a gradual decrease in copy number is observed as probes going from position 4350 to position 1 are used (Fig. 4). Therefore, many of the L1 repeats detected with the probe from the 5' end may be full length, indicating that up to 17% of the population of L1Oc repeats could be full length. This difference in copy number at the 5' and 3' ends of L1Oc repeats is also observed when uncloned genomic DNA is hybridized with the different L1Oc probes (data not shown). Thus, the lower copy number at the 5' end is not a result of underrepresentation in the cloned genomic library.

Approximate 5' End of Full-Length L1Oc Repeats

Because the 5' end of L1Oc5 is at the end of the cloned portion of the rabbit $\beta$-like globin gene cluster, it is likely that the nucleotide sequence obtained from L1Oc5 is not that of a full-length L1 repeat. Therefore, cloned subfragments of L1Oc5 were used as probes against Southern (1975) blots of rabbit genomic DNA to determine the average structure of full-length rabbit L1 repeats. Discrete genomic restriction fragments detected with L1Oc5 probes were mapped by two strategies. The portion of L1Oc contained within the genomic restriction fragment was determined by which probes from L1Oc5 hybridized to the fragment, and then the genomic restriction fragment was aligned with conserved restriction sites found in the cloned L1Oc DNA. This analysis is presented in detail in Demers (1987), and the portion relevant to the 5' end of L1Oc is summarized in Fig. 5.

The longest restriction fragment extending 5' to the cloned end of L1Oc5 is the $PstI$ 4.0-kb fragment that ends 1 kb 5' to the cloned region of L1Oc5 (Fig. 5). The $ScaI$ 2.1-kb, $SphI$ 1.9-kb, and $XmnI$ 3.7-kb genomic fragments all have 5' ends between the conserved $PstI$ site located outside L1Oc5 and the 5' end of L1Oc5 (Fig. 5). These data indicate that full-length L1Oc repeats will extend at least 1 kb further 5' than the sequenced portion of L1Oc5. Several clones from the rabbit genomic DNA library are currently being studied in order to determine the 5' end of L1Oc repeats.

Comparison of L1Oc with L1 Repeats from Mouse and Human

The sequence of the rabbit L1 repeat was compared with the sequences of the mouse and human L1 repeats by dot-plots and by sequence alignments. The dot-plot analyses in Fig. 6 show that the internal sequence of L1Oc is very similar to both L1Md (mouse) and L1Hs (human) over very long segments, whereas the 5' and 3' ends are not conserved between species. The internal region of sequence similarity of about 4.5 kb is divided into two parts, a short region of similarity of about 300 bp followed by a very long segment of similarity.

The long segments of internal similarity are in the portion of L1 that encodes open reading frames (ORFs). The ORFs found in the L1Oc5 sequence are shown in Fig. 7, along with a comparison of the ORFs from L1Md. The mouse L1MdA2 sequence contains two ORFs, one of 1137 nucleotides (top strand, N frame in Fig. 7, bottom panel) and one of 3900 nucleotides (top strand, N + 1 frame in Fig. 7), that overlap by 14 nucleotides (Loeb et al. 1986). Seven open reading blocks are in the rabbit L1Oc5 sequence in frames N, N + 1, and N + 2 (Fig. 7, top panel). The bar between the stop codon maps of each species shows the regions of similarity (Fig. 6) as filled boxes. It is apparent that the regions of L1 that are similar between species contain extensive ORFs, although the ORFs at the 5' end are not similar between species.

Rabbit L1 repeats have only two major ORFs. Although the data in Fig. 7 show that L1Oc5 has several ORFs, they are probably derived from longer reading frames in the ancestral L1 sequence.

Fig. 4. Copy number of regions of L1Oc. The copy number per haploid genome is plotted as a function of the location of the probe from the L1 repeat. The location of the probe used for each region is given using the position numbers in Fig. 2.

Fig. 5. Restriction map of 5' end of L1Oc repeats. Partial restriction site maps of L1Oc5 and L1Oc4 are shown in the open boxes. The location of rabbit genomic DNA fragments (filled boxes) that hybridize to probes from L1Oc5 are shown below the restriction map; the fragments are labeled with the restriction enzyme and their size in kb.
ORFs shown for L1Oc5 in Fig. 7 can be linked into two long ORFs by making substitutions found in L1Oc4, and by making insertions or deletions necessary to maintain the alignment of L1Oc5 with regions of similarity of L1s from mouse or human (Demers 1987). Examples of such insertions to maintain the alignment can be seen at positions 798 (ORF-1) and 1445 (ORF-2) of the L1Oc5 sequence in Fig. 8. By aligning the sequences of several human L1 repeats, Scott et al. (1987) recently concluded that L1Hs also contains two major ORFs. The diagram in Fig. 7 shows that the long region of simi-
The portion of the 66-bp tandem repeat in \( LIRn \) that is included in the alignment is in lowercase letters. Continued on all four species and that start ORFa in the \( LIRn \) sequence are also boldface. ATG codons proposed as the start point for ORFa and ORF-2 are in boldface, and in-phase ATGs close to the proposed beginning of ORF-1 in all four species and that start ORFa in the \( LIRn \) sequence are also underlined. The portion of the 66-bp tandem repeat in \( LIRn \) that is included in the alignment is in lower-case letters. Continued on pages 12 and 13.
Fig. 8. Continued
Analysis of ORF-1 of L1 Repeats

The two ORFs are overlapping in L1Md, and it is of interest to determine whether this feature is conserved in L1 repeats from other species. Also, ORF-1 appears to be a hybrid sequence because it is well conserved between species in the 3' half but is not well conserved in the 5' half. Therefore, the sequence of ORF-1 and the region between the ORFs were aligned for the L1 sequences before and after it are conserved, probably because this region is not conserved, whereas the sequence between the L1 ORFs lacks of conservation supports the proposed assignments for the start of ORF-2 in L1Oc and L1Hs. The mouse L1 sequence is ATA at positions 1235-1237; this same sequence is found in three sequenced members of the L1Md family (Shehee et al. 1987). Therefore, the overlap between ORFs in L1Md begins in a different reading frame at position 1149, and thus it overlaps with ORF-1 for 14 nucleotides. By aligning the sequences of the different L1s in the well-conserved ORF-2 region, it is apparent that an ATG is conserved in the rabbit and human L1 repeats at positions 1235-1237. An in-frame ATG two codons upstream was previously identified as the start of ORFb in the L1Rn sequence (D'Ambrosio et al. 1986) and an ATG is also in frame in the L1Md sequence seven codons upstream. One can propose that the TAA close to position 1163 is the end of ORF-1 and the ATG at positions 1235-1237 is the start of ORF-2 in rabbit and human L1 repeats. In an independent analysis of several individual L1Hs repeats, these same codons were assigned as the end of ORF-1 and the start of ORF-2 in the consensus L1Hs sequence (Scott et al. 1987). As shown in Fig. 8, ORF-2 is in the same reading frame as ORF-1 in the L1Oc and L1Hs sequences. Thus, the overlap in reading frames seen for L1Md is not observed in L1Oc and L1Hs. ORF-2 in L1Rn is in a different reading frame than ORF-1, but the L1Rn sequence does have an ATG proposed as the start of ORF-2. Thus, L1Rn has overlapping reading frames, but the sequence in the overlap may not be used to encode a protein.

The region between ORF-1 and ORF-2 is not conserved between mammalian species. The sequence between the TAA that ends ORF-1 and the ATG proposed to be the start of ORF-2 is in a region that is quite dissimilar between rabbit and mouse and between rabbit and human (plain text region between positions 1121 and 1240 in Fig. 8). This is the region of no similarity previously seen in dot-plots (Fig. 6). The sequence between the Ll ORFs is also not conserved in comparisons between the human and rodent sequences (Scott et al. 1987). Because this region is not conserved, whereas the sequences before and after it are conserved, probably for their capacity to encode a protein, it is unlikely that the inter-ORF region encodes a protein. This lack of conservation supports the proposed assignments for the start of ORF-2 in L1Oc and L1Hs. The mouse L1 sequence is ATA at positions 1235-1237; this same sequence is found in three sequenced members of the L1Md family (Shehee et al. 1987). Therefore, the overlap between reading frames 1 and 2 are conserved in mouse L1s, but the overlaps are not seen in the rabbit and human L1 sequences.

The ORF-1 sequence is a composite of conserved and nonconserved regions. As shown diagrammatically in Fig. 9, codons 79-294 are highly related between species in different mammalian orders, and a long segment from codons 171 through 294 shows a 52-56% amino acid identity in these comparisons. A short region from codons 97 to 122 is not conserved, nor are the last 14 codons in the sequence, but in general the C-terminal two-thirds of ORF-1 is conserved between orders. A search through the databanks at the Protein Identification Resource (National Biomedical Research Foundation) did not identify any known proteins (besides the L1 proteins) that are related to the C-terminal half of the ORF-1 sequence.
In contrast, the N-terminal portion of ORF-1 is not highly conserved between mammalian orders. This region shows almost no similarity between rabbit and human (sequence between nucleotide positions 3 and 476 in Fig. 8; Fig. 9), and the comparison between rabbit and mouse shows only a short segment of matching sequence at the 5' end (Figs. 8 and 9). The dissimilarity of the sequences makes it difficult to assign a start point to ORF-1. However, an ATG is found three codons downstream in the rabbit, mouse, and rat sequences at positions 240-242 of Fig. 8 (shown in boldface). An ATG is found in the rabbit, mouse, and rat sequences at positions 240-242 has been tentatively assigned as the start of ORF-1, and the ATG codons are either immediately adjacent (mouse and rat) or are 20 codons upstream (rabbit, underlined in Fig. 8). The ATG at positions 240-242 has been confirmed using the FASTp program (Lipman and Pearson 1985). Although the N-terminal half of ORF-1 differs among rabbits, mouse, and humans, it is similar between the two rodents, mouse and rat. This region surrounds a 66-bp tandemly repeated sequence in L1Rn (Soares et al. 1985; D’Ambrosio et al. 1986) and contains several in-frame stop codons in L1Rn (Fig. 8). It is possible that the coding function of this region has been lost in L1Rn.

The N-terminal half of ORF-1 from the rabbit L1 sequence is related to type II cytoskeletal keratin. Protein sequence databanks were searched using the FASTp program (Lipman and Pearson 1985), and a significant match was found with type II cytoskeletal keratin. The region of L10c ORF-1 that matches with keratin, along with the percent amino acid identity, is shown in Fig. 9, and the alignment with the human 67 kDa type II keratin (Johnson et al. 1985) is shown in Fig. 10. The sequences align over a 156-amino acid region, with an average of 20.5% identity. The segment between amino acid positions 95 and 126 of L1Oc ORF-1 is most similar to type II keratin; this segment contains identical amino acids at 32% of the positions.

The similarity between the N-terminal half of ORF-1 from L10c and type II cytoskeletal keratin is statistically significant. The sequence of the type II keratin was scrambled into 20 different sequences and aligned with the ORF-1 sequence to generate an average match score. The match score with the true keratin sequence is 13 standard deviations above the average match score with the scrambled sequences; a difference of 10 standard deviations in this test is an indicator of a significant evolutionary relationship (Lipman and Pearson 1985). Although statistical significance does not establish biological significance, it is helpful to compare this match with that of a part of ORF-2 with reverse transcriptases whose similarity has been cited as significant in the past (Hattori et al. 1986; Loeb et al. 1986). The alignment between the L1Md ORF-2 sequence and the sequence of reverse transcriptase from Moloney murine leukemia virus shows 17.5% amino acid identity, whereas the alignment between L1Oc ORF-1 and type II keratin shows 20.5% identity. It is apparent that ORF-1 of the rabbit L1 contains a region related in sequence to type II cytoskeletal keratin.

**Discussion**

The propagation of L1 repeats probably has occurred independently in different mammalian genomes. Although the L1 repeats from lagomorphs, rodents, and primates are similar in size and sequence organization, the 5' and 3' ends are distinctive (summarized in Fig. 11). Also, the L1 repeats...
are located in different positions in orthologous regions of chromosomes, specifically the β-like globin gene cluster of rabbits and humans (Margot et al. 1989) and mice (Shehee et al. 1989). Because the contemporary β-like globin gene clusters are descended from a preexisting gene cluster in the last common ancestor, the presence of L1 repeats at different positions in different species indicates that the L1 repeats have integrated independently into these gene clusters (and probably the whole genome) in each species.

It is noteworthy, therefore, that the structure of the population of L1 repeats is quite similar in several mammals. Most members of the L1 repeat family in rabbits (this paper), mouse (Voliva et al. 1983), and monkeys (Grimaldi et al. 1984) are truncated from the 5' end, resulting in a higher frequency in the genome of the 3' end of L1 (about 50,000 copies) than the 5' end (about 10,000 copies). This similarity in copy number suggests that the time of onset and the rate of propagation of L1 repeats is similar in the different species. The rabbit, mouse, and monkey L1 repeats also show a similar pattern for the increase in copy number in which the 5' regions increase gradually in copy number before a large increase in copy number at the very 3' end. This very large increase in copy number in the 3' region could indicate a strong stop for reverse transcriptase during the conversion of the L1 transcript to a DNA copy. Given this frequency of polar truncations of L1 in rabbits, humans, and mice, it is striking that most of the L1 repeats in rats are full length (D’Am-brosio et al. 1986). Some aspect of the mechanism for synthesis and propagation of the L1s is apparently different in rats, e.g., to allow more full length reverse transcripts or to select for these in the integration process.

Full length L1 transcripts have been observed in teratocarcinoma cells (Skowronski and Singer 1985). Given the assignments of start and stop codons proposed in this paper, then transcripts of the L1 repeat of rabbits and humans have the characteristics of a dicistronic RNA. Polycistronic mRNAs are common in bacteria, and a polycistronic arrangement of genes is found in the genomes of some RNA viruses that infect animals and plants, e.g., togaviruses, coronaviruses, and tobacco mosaic virus. In contrast, most mRNAs from eukaryotic cellular genes are monocistronic. Regardless of whether the ORFs are overlapping, as in L1Md, or are part of a dicistronic RNA, as in L1Oc and L1Hs, the structure of the L1 repeats resembles DNA copies of viral genomes more than conventional cellular transcription units. This suggests that the ancestor to L1 repeats in fact may be some type of animal virus rather than a normal cellular gene, as is often proposed (reviewed in Weiner et al. 1986). A viral ancestor with a wide host range would provide an explanation for the independent, and perhaps simultaneous, entry of the L1 element into different mammalian genomes.

The ORFs in the L1 repeat appear to encode hybrids of different types of proteins (Fig. 11). ORF-1 can be divided into two parts, the N-terminal por-
tion that is not well conserved between species and the C-terminal portion that is well conserved. In the rabbit L1 repeat, a sequence similar to keratin has been fused to the conserved C-terminal portion of ORF-1. Although ORF-2 is conserved in L1s from different orders of mammals it also seems to be a hybrid of sequences related to several proteins (Fig. 11). The middle portion of ORF-2 is related to reverse transcriptase (Hattori et al. 1986; Loeb et al. 1986). Different parts of the C-terminal region are related to transferrin (Hattori et al. 1986) and to nucleic acid binding proteins with the cysteine structural motif, such as the binding proteins derived from retroviral gag genes (Fanning and Singer 1987). The cysteine structural motif is related to the zinc fingers characterized in TFIIIA and other nucleic acid binding proteins (Fanning and Singer 1987). This pastiche of similarities suggests that the L1 element is a fusion of several different sequences, some of which are derived from cellular genes, possibly by a viral vector.

Another fusion event may account for the variation in sizes and sequences of the 3' untranslated regions of L1 repeats in different mammals. The 3' untranslated regions of orthologous globin genes in mammals have retained obvious sequence similarities over the course of eutherian evolution (e.g., Hardies et al. 1984; Hardison 1984), so it is puzzling that no sequence similarity is seen in the 3' untranslated region of L1Oc repeats in comparisons between mammals (Fig. 11). Perhaps the conserved coding region was fused to a different 3' untranslated sequence in each species. It is noteworthy that the 5' end of L1Oc begins immediately after the conserved termination codon that ends ORF-2, suggesting that the sequence corresponding to the 3' untranslated region of L1Oc may exist as a distinct repetitive element in the rabbit genome in addition to its presence in the L1 sequence. If so, this would be an additional factor in explaining the large increase in copy number of L1 repeats in this region. A similar situation has been observed in Drosophila melanogaster, in which a suffix, an element repeated about 300 times in the genome, is almost identical to the sequence of the 3' untranslated region (but not the coding region) of the F element that is present about 70 times in the genome (DiNocera and Casari 1987).

The mammalian L1 repeats show a clear similarity to the ingi repeat in the protozoan Trypanosoma brucei (Kimmel et al. 1987), the I factor of the I-R system of hybrid dysgenesis in D. melanogaster (Fawcett et al. 1986), F elements in D. melanogaster (DiNocera and Casari 1987), and the R1Bm (Xiong and Eickbush 1988) and R2Bm (Burke et al. 1987) insertion sequences in some rRNA genes of Bombyx mori (Fig. 11). The similarity has been recognized only in the region proposed to encode reverse transcriptase, and these sequences are more similar among themselves than to retroviral reverse transcriptases (DiNocera and Casari 1987; Xiong and Eickbush 1988). The mammalian L1s and these protozoan and insect repeats share other structural features, such as the absence of long terminal repeats, the presence of at least two ORFs (ORF-2 containing sequences similar to reverse transcriptase and either ORF-1 or ORF-2 encoding a cysteine motif), a length from 5 to 7.5 kb, and a 3' untranslated region with a sequence similar to AAUAAA close to the 3' end. The dicistronic structure proposed for L1Oc and LIHs may also be present in the I factor, the F element, and the R1Bm repeat (Fawcett et al. 1986; DiNocera and Casari 1987; Xiong and Eickbush 1988). Each type of repeated element also has some distinctive features, e.g., the specific insertion sites for R1Bm and R2Bm in the rRNA genes and the absence of A-rich tracts at the 3' ends of some of the insect repeats. However, at least parts of these repeats in mammals, insects, and a parasitic protozoan appear to be evolutionarily related. If this type of repeat is restricted to these groups of organisms, it may indicate that the genetic information was transferred among parasites, their mammalian hosts, and insect vectors (Kimmel et al. 1987). A viral progenitor, suggested by the dicistronic arrangement shown in this paper, would provide a means for the horizontal transmission of the L1 sequences.

Acknowledgments. We thank S. Davis for technical help. This work was supported by Public Health Service grants DK27635 and AM31961, a Research Career Development Award to R.C.H., DK01589, and a Pennsylvania State University Agricultural Experiment Station Hatch Grant.

References

Baltimore, D (1985) Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. Cell 40:481-482.
Benton WD, Davis RW (1977) Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196:180-182.
Brown SDM, Dover G (1981) Organization and evolutionary progress of a dispersed repetitive family of sequences in widely separated rodent genomes. J Mol Biol 150:441-466.
Burke WD, Calalang CC, Eickbush TH (1987) The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. Mol Cell Biol 7:2221-2230.
Burton FH, Loeb DD, Vellva CF, Martin SL, Edgell MH, Hutchinson CA III (1986) Conservation throughout mammalia and extensive protein encoding capacity of the highly repeated DNA LI. J Mol Biol 187:291-304.
Church GM, Gilbert W (1984) Genomic sequencing. Proc Natl Acad Sci USA 81:1991-1995.
D'Ambrosie O, Waitzkin D, Witney FR, Salemmie A, Furano AV (1986) Structure of the highly repeated, long interspersed...
DNA family (LINE or L1Rn) of the rat. Mol Cell Biol 6:411–424

Demers GW (1987) L10c: a long interspersed repetitive DNA element in the rabbit genome. Thesis, The Pennsylvania State University, University Park

Demers GW, Brenn K, Hardison RC (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

DiNocera PP, Casari G (1987) Related polypeptides are encoded by Drosophila F elements, I factors, and mammalian L1 repeats. Proc Natl Acad Sci USA 84:5843–5847

Economou-Pachnis A, Lohse MA, Furano AV, Tsichlis PN (1985) Insertion of long interspersed repeated elements at the IgH (immunoglobulin heavy chain) and Mlvi-2 (Mooney leukemia virus integration 2) loci of rats. Proc Natl Acad Sci USA 82:2857–2861

Fanning T (1982) Characterization of a highly repetitive family of DNA sequences in the mouse. Nucleic Acids Res 10:5003–5013

Fanning T, Singer MF (1987) The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res 15:2251–2260

Fawcett DH, Lister CK, Kellett E, Finnegan DJ (1986) Transposable elements controlling I-R hybrid dysgenesis in D. melanogaster are similar to mammalian LINEs. Cell 47:1007–1015

Grimaldi G, Skowronski J, Singer MF (1984) Defining the beginning and end of KpnI family segments. EMBO J 3:1753–1759

Hardies SC, Edgell MH, Hutchison CA III (1984) Evolution of the mammalian β-globin gene cluster. J Biol Chem 259:3748–3756

Hardison RC (1984) Comparison of the β-like globin gene families of rabbits and humans indicates that the gene cluster 5′-eryγb-β-3′ predates the mammalian radiation. Mol Biol Evol 1:390–410

Hardison RC, Butler ET, Lacy E, Maniatis T, Efstathiades A (1979) Linkage analysis and cDNA sequencing of four linked rabbit β-like globin genes. Cell 18:1285–1297

Hattori M, Hidaka S, Sakaki Y (1985) Sequence analysis of a KpnI family member near the 3′ end of human β-globin gene. Nucleic Acids Res 13:7813–7827

Hattori M, Kuhara S, Takenaka O, Sakaki Y (1986) LI family of long interspersed repetitive elements in mammalian genomes. Cell 53:391–400

Johnson LD, Idler WW, Zhou XM, Roop DR, Steinert PM (1985) Structure of a gene for the human epidermal 67-kDa keratin. Proc Natl Acad Sci USA 82:1896–1900

Katzir N, Rechavi G, Cohen JB, Unger T, Simon F, Segal S, Cohen D, Givol D (1985) “Retroposon” insertion into the cellular oncogene c-myC in canine transmissible venereal tumor. Proc Natl Acad Sci USA 82:1054–1058

Kimmel BE, Ole-Myoiyi OK, Young JR (1987) Ing i, 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

Korenberg JR, Rykowski MC (1988) Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. Cell 53:391–400

Lacy E, Maniatis T (1980) The nucleotide sequence of a rabbit β-globin pseudogene. Cell 21:545–553

Lacy E, Hardison RC, Quon D, Maniatis T (1979) Linkage arrangement of four rabbit β-like globin genes. Cell 18:1273–1283

Lerman MI, Thayer RE, Singer MF (1983) KpnI 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

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Loeb DD, Padgett RW, Hardies SC, Shehee WR, Comer MB, Edgell MH, Hutchinson CA (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

Maniatis T, Hardison RC, Lacy E, Lauer J, O’Connor C, Quon D, Sim GK, Efstathiades A (1978) The isolation of structural genes from libraries of eucaryotic DNA. Cell 15:687–701

Manuelidis L (1982) Nucleotide sequence definition of a major human repeated DNA, the HindIII 1.9 kb family. Nucleic Acids Res 10:3211–3219

Margot JB, Demers GW, Hardison RC (1989) Complete nucleotide sequence of the rabbit β-like globin gene cluster: analysis of intergenic sequences and comparison with human β-like globin gene cluster. J Mol Biol 205:15–40

Martin SL, Voliva CF, Burton FH, Edgell MH, Hutchinson CA 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

Messing J (1983) New M13 vectors for cloning. Methods Enzymol 101:20–78

Paulson KE, Deka N, Schmid CW, Misra R, Schindler CW, Rush MG, Kadyk L, Leinwand L (1985) A transposon-like element in human DNA. Nature (London) 316:359–361

Potter SS (1984) Rearranged sequence of a human KpnI element. Proc Natl Acad Sci USA 81:1012–1016

Rigby PW, Dejeckmann M, Rhodes C, Berg P (1977) Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J Mol Biol 113:237–243

Rogers J (1983) Retroposons defined. Nature (London) 301:460

Rohrbaugh ML, Hardison RC (1983) Analysis of rabbit β-like globin gene transcripts during development. J Mol Biol 164:395–417

Rohrbaugh ML, Johnson JE III, James MD, Hardison RC (1985) Transcription unit of rabbit β1-globin gene. Mol Cell Biol 5:147–160

Sanger FS, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Scott AE, Schmeckpeper BJ, Abdelrazik M, Comey CT, O’Hara B, Rossiter JP, Cooley T, Heath P, Smith KD, Margolet L (1987) Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics 1:113–125

Shehee WR, Chao SF, Loeb DD, Comer MB, Hutchinson CA, Edgell MH (1987) Determination of a functional ancestral sequence and definition of the 5′ end of A-type mouse L1 elements. J Mol Biol 196:757–767

Shehee WR, Loeb DD, Adey NB, Burton FH, Casavant NC, Cole P, Davies CJ, McGraw RA, Schnichman SA, Severynse DM, Voliva CF, Weyter FW, Wisely GB, Edgell MH, Hutchinson LA III (1989) The nucleotide sequence of the BALB/C mouse β-globin complex. J Mol Biol 205:41–62

Shen CKJ, Maniatis T (1980) The organization of repetitive sequences in a cluster of rabbit β-like globin genes. Cell 19:379–391

Singer MF (1982) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28:433–434

Singer MF, Skowronski J (1985) Making sense out of LINEs: long interspersed repeat sequences in mammalian genomes. Trends Biochem Sci 10:119–122
Skowronski J, Singer MF (1986) Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc Natl Acad Sci USA 82:6050–6054
Soares MB, Schon E, Efstratiadis A (1985) Rat LINE1: the origin and evolution of a family of long interspersed middle repetitive DNA elements. J Mol Evol 22:117–133
Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517
Voliva CF, Jahn CL, Comer MB, Hutchison CA III, Edgell MH (1983) The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res 11:8847–8859
Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retro...

Received July 25, 1988/Revised December 9, 1988