J Mol Evol (1989) 29:3–19

Journal of Molecular Evolution

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 (L1Oc) 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. L1Oc 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 L1Oc 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 L1Oc repeat has internal direct repeats of 78 bp in the 3′ untranslated region, whereas other L1Oc repeats have only one copy of this sequence. The longest repeat sequenced, L1Oc5, is 6.5 kb long, and genomic blot-hybridization data using probes from the 5′ end of L1Oc5 indicate that a full length L1Oc repeat is about 7.5 kb long, extending about 1 kb 5′ to the sequenced region. The L1Oc5 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 L1Oc 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 (Baltimore 1985). 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-

Offprint requests to: R.C. Hardison
* *Present address:* Fred Hutchinson Cancer Research Center, Seattle, Washington 98104, USA
† *Present address:* Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, Illinois 60201, USA

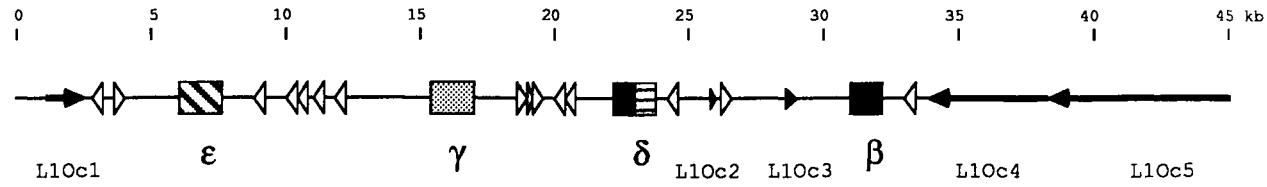

Fig. 1. Repetitive DNA in the rabbit β-like globin gene cluster. The β-like globin genes ε, γ, δ, and β are shown as boxes along the 45-kb segment of cloned DNA (Lacy et al. 1979). Transcription of the active genes is from left to right. The location and orientation of L1 repeats are shown by the filled arrows. The L1 repeats are named L1Oc1–L1Oc5 (Demers et al. 1986). The location and orientation of C repeats, a rabbit SINE, are shown by the open arrows.

spersed repeats, or LINEs, that are greater than 6000 bp long (Singer 1982). Although no precise sequence specificity has been observed at the insertion sites, SINEs and LINEs do have a regional preference for integration in the human genome, as shown by the enrichment of different chromosome bands for either LINEs or SINEs (Korenberg and Rykowski 1988).

Although several different sequences have been dispersed as SINEs in mammals (reviewed in Weiner et al. 1986), only one sequence element, called L1, has been found to be dispersed as a LINE in mammals (reviewed in Singer and Skowronski 1985). The L1 sequence has been identified in a wide variety of species including primates (Lerman et al. 1983), mice (Brown and Dover 1981; Fanning 1982), rats (Econonmou-Pachnis et al. 1985; Soares et al. 1985; D'Ambrosio et al. 1986), dogs (Katzir et al. 1985), cats (Fanning and Singer 1987), and rabbits (Demers et al. 1986). Genomic blot-hybridization analysis indicates that the L1 sequence is present in all mammalian species at a frequency of about 10^4–10^5 copies per haploid genome (Burton et al. 1986).

Although the parent genes of SINEs are transcribed by RNA polymerase III, the L1 repeats appear to be derived from an RNA polymerase II transcript. The parent gene of L1 is proposed to be a protein-coding gene (reviewed in Singer and Skowronski 1985). Long open reading frames (ORFs) are found in the L1 sequences (Manuelidis 1982; Martin et al. 1984; Potter 1984), and sequenced members from the mouse genome have two overlapping ORFs of 1137 bp (ORF-1) and 3900 bp (ORF-2) (Loeb et al. 1986; Shehee et al. 1987). The ORF-2 regions of primate and rabbit L1 are 65% similar, but the similarity ends abruptly at a conserved stop codon (Demers et al. 1986).

In previous studies on the L1 repeats from rabbits (L1Oc, for LINE 1 from *Oryctolagus cuniculus*), the B, E, and D repeats identified by Shen and Maniatis (1980) were shown to be parts of the L1Oc repeat. The sequence of one truncated L1 repeat and part of another repeat were presented as a composite sequence, and the ORF (corresponding to ORF-2) and 3′ untranslated region were identified (Demers et al. 1986). 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 L1Oc is related to type II cytoskeletal keratin.

Materials and Methods

Subcloning and Sequencing of L1Oc Repeats. The sequenced members of the L1Oc 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 L1Oc). 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,

```
5 GGCCGCACCCATCTCAAGCCTCCAAGGCTCCTCCAACAGCAGGCAGTCCACTTAACATGGACACAGTATAAAAAAAAAAAAAAAAGAAAAAAAAAAAACG  100

5 CACAGTGACACAAGAAGAATTAACTATGCCGAGTAACAAACACAGAAATAGAGGGAGCAAGATCAACGATGACACTATGATGCCTCCAAATAAGCAAAAC  200

5 ACCCCAAGCCAAGAGTATGAAGATGATGAGATAGAAGAAATGCAAGATACGGATTTCAAAAAATTTATGATAAGAACATTTAGAAGTTTTCAAAAGCAAA  300

5 TCCTTGAACTACAGAAATCCTTAATGGACAAGATTGAAAATCTCTCTCGTGAAAATGAAATTTTAAGGAAGAGTCAAAATGAAACTCAGAAACTAGTAGA  400

5 ACAGGAAAGTGTAATAGTGAAGAGAAATCAAAATGAAATGAAGAGCTCAATAGATCAAATGGCAAACACATTAGAAAGCCTTAAAAACAGAATGGGTGAA  500

5 GCAGAAGACAGAATATTGGACTTAGAAGACAGAGCACAGGAAAGTATACAGTCAAACCAAAGAAAAGAAGAGGAAATTAGAAATCTAAAAAATATTGTTG  600

5 GGAATCTACAGGATACTATTAAAAAAACCAACATTCGAGTTCTAGGAGTTCCTGAAGGCATGGAGAGAGAGAAAGGATTGGAAGGCCTTTTTAGTGAGAT  700

5 ACTAGCAGAGAACTTTCCAGGTTTGGAGAAGGACAGAGATATCCTAGTACAGGAAGCTCATAGAACCCCCAATAAACATGACCAAAAGAGATCCTCACAC  800

5 GACACGTGGTAATTAAACTTACCACAGTGAAACATAAAGAAAAGATCCTAAAATGTGCAAGAGAGAAACATCAGATTACTCTCAGAGGATCTCCAATCAG  900

5 ACTCACAGCAGACTTCTCATCAGAAACCCTACAAGCTAGGAGGGAATGGCGAGACATAGCACAGGTGCTAAGAGAGAAAAATTGCCAGCCCAGAATATTA 1000

5 TATCCTGCCAAGCTCTCATTTGTGAATGAAGGTGAAATAAAGACCTTTCATAGCAAACAGAAATTGAAAGACTTTGTGGCCACTTGTCCGGCCCTGCAAA 1100

5 AGATACTTAAAGATGTGCTACACTCAGAAACACAGAAACACGGCCATCAATATGAAAGAAGGGAAAGGAAGAACACCTACCAGTAAAAGAGCATGGGAAG 1200

5 CTCAAAGCATATACTAGAAAATATTTCCGGGAAAATGGCAGGGCAAAGTCACTACGTATCAATTGTCACATTGAACATTAATGGTCTGAATTCTTCAGTT 1300
4                              GGGAAAATGGCAGGGCAAAATTAC-ACTTATCAATAGTCACACTGAACGTTAATGGCCTGAACTGTCCAGTT

5 AAAAGACACCGTTTGGATGACTGGCTCACAGAACACAACCCAACTATTTGTTGCCTACAAGAAACACATCTCTCTAACAAAGAGGCATGCAGACTGAAAG 1400
4 AAAAGACATAGATTGGCTGATTGGGTTAAGGAACAAAACCCATCTATTTGCTGCTTACA-GAAACACATCTTTCCAACAAAGATGCATCCAGACTGAATG

5 TGAAAGGTTGGAAAAAGATATTCCATGCCAACAGAAACCAAAAAA-AGCAGGTGTAGCCATATTAATATCAGACAAAATAAACTTTAATACAAAAACTGT 1499
4 TGAAAGGCTGGAGAAAGATATTCCATGCCAACAGAAATGAAAAAAGAGCAGGCATAACCATCTTAATATCAGACAAAATAAACTTTAGCACAAAAACTGT

5 TAAGAGAGACAAAGAGGGACACTATATAATGATTAAGGGTTCAATTCAACAGGAAGA-TGTAACTATTATAAATGTATATGCACCTAATTACAGGGCACC 1598
4 TAAGAGAGACAAAGAGGGGCACTACATAATGATTAAGGGATGAATTCAACAGAAAAAATATAACGATTATCAATGTATATGCACCTAATTACAGGGCACC

5 GGTCTATTTAAAAGATATGTTAAGGGACTTAAAGGGAGACTTAGATTCCAATACAATAGTACTGGGGGACTTCAATACTCCACTCTCAGAAATAGACAGA 1698
4 GGTTTATTTAAAAGATTTGTTAAGAGAGTTAAAGGGAGACTTAGACTCCAATACAATAGTACTGGGGGACTTCAATACTCCACTCTCAGAAATAGACAGA

5 TCATCCGGACAGAAGATCAACAAGGAAACAGCAGATTTAATTGACACTATTGCCCAAATGGATCTAACAGATATCTACAGAACTTTCAACCCTACATCTA 1798
4 TCA-CAGGACAGAAGACTA-CA-GGAAACAGTACATTCAAAGGATACTATAGCCCAGATGGATCTGACACATATCTACAGAACTTTTCATCCTGCACTTA

5 CAGACTTCACATTCTTCTCAGCAGCGCATGGGACCTTCTCTAGGATTGATCACATACTAGGCCATAAAGCAAGTCTCAGCAAATTTAAAAGAATTAGAAT 1898
4 AAGAATTTACATTCTTCTCAGCAGTACATGGAACCTACTCTAAGATTAACCACATACTAGGCCATAAAGCAAGTCTCAGCACATTCAAAAGAATTAGAAT

5 CATACCATGCAGCTTCTCAGACCACAGTGGGATGAAGCTGGAAATTAGCAACTCAGGAAACCCAAGAAAGTATGCAAACACATGGAGACTGAACAACATG 1998
4 CATATGATGCAGCTTCTCAGACCATAATAGAATGAAGTCGGAAATTAGCAACTCAGGAATCCCTACAGCATATGCAAACACATGGAGAGTGAACAACATG

5 CTCCTGAATGAACACTGGGTCATTCAAGAAATCAAAAGAGAAATCAAAAACTTTCTGGAAGTAAATGAAGACAACAACACAACATATCAAAACTTATGGG 2098
4 CTCCTGAATGAACACTAGGTCATCAAAGAAATCAAAAGAGAAATCAAAAACTTTCTGGAAGTAAATGAGGATAACAGCACAACATACCAAAATGTATGAG

5 ATACAGCAAAAGCAGTATTGAGAGGCAAATTTATAGCAATAGGTGCCTATATCAAGAAATTGGAAAGGCACCAAATAAATGAGCTTTCAGTGCACCTCAA 2198
4 ATGCAGCGAAAGCAGTGTTAAGAGAAAAGTTTATATCAATAGGTGCCTACATCAAGAAATTGGAAAGGCACCAAATAGATGAGCTTTCAAATCACCTCAA

5 GGACCTAGAAAAACTGCAGCAAACCAAACCCAAATCTAGTAGGAGAAGAGAAATAATTAAAACCAGAGAAGAAATTAACAGGATTGAATCAAAAAAAAAA 2298
4 GGATCTAGAAAAACTGCAGCAAAGCAGACCCAAATCTAGTAGGAGAAGAGAAATAATTAAAATCAGAGAAGAAATCAACAGGATTGAATCCAAAAAAA--

5 AAAACATTACAAAAAATCAGCCAAGCGAGAAGCTGGTTTTTTGAAAAAATAAACAAAATTGACACCCCATTGGCCCAACTAACTAAAAAAAGAAGAGAAA 2398
4 -------TTACCAAAAATCAGCCAAATGGGGAGCTGGTTTTTTGAAAAAATAAACAAAATTGATACCGCATTAGCTCAACTAACTAAAAGAAGAAGAGA--

5 AGACCCAAATCAATAAAATCAGAGATGAAAAAGTAAACGTAACAACAGACACCACAGAAATAAAAAGAATCATCAGAAATTACTACAAGGACCTGTATGC 2498
  ----CCAAATCAATAAAATCAGAGATGAAATAGGAAATGTAACAGCAGACACCACAGAAATGAAAAGAATCATCAGAAATTACTACAAGGAC-TGTATGC

5 CAGCAAACAGGAAAACCTATCAGAAATGGATAGATTCCTGGACACATGCAATCTACCAAAATTGAACCATGAAGACATCGAAAACCTAAATAGACCCATA 2598
4 CAGCAAACAGGGAAATCTATCAGAAATGCATAGATTCCTGGACACCTGCAACCTACCTACATTGAACCAGGAAGACATCGAAAGCCTAAACAAACCCATA

5 ACTGAAACAGAAATTGAAACAGTAATAAAGGCCCTCCCAACAAAGAAAAGCCCAGGACCAGATGGATTCACTGCTGAATTCTACCAGACATTTAAAGAAG 2698
4 ACTGAGGCAGAAATTGAAACAGTAATAAAGGCCCTCCCAACAAAGAAAAGCCCAGGACCAGATGGATTCACTACTGAATTCTACCAGAAATTTAAAGAAG

5 AACTAATCCCATTATTCTCAAACTATTCAGAACAATCGAAAAAGAGGGAATCCTCCCAAATTCTTTCTATGAAGCCAGCATCACCTTAATCCCTAAGCCA 2798
4 AACTAACTCAATTCTTCTCAAACTATTCAGAACAATTGAAAAAGAGGGAATCCTCCCAAAATCTTTCTATGAATCCAGCATCACCTTAATTCCTAAGCCG

5 GAGAAAGATGCAGGACTGAAAGAAAATTACAGACCAATATCCCTGATGAACATAGATGCAAAAATCCTCAATAAAATTCTGGCCAATAGAATACAACAAC 2898
4 GAAAAAGATGCAGCATTGAAAGAAAATTACAGAACAATATACCTAATGAACATAGACTCAAAAATTCTCAATAAAATTCTGGCCAACGGAGTGCAACAAC

5 ACATCAGGAAAATCATCCACCCAGACCAAGTGGGATTCATCCCTGGTATGCAGGGATGGTTCAA-TGTTCGCAAATCAATCAATGTGATTCACCACATTA 2997
4 ATTTCAGAAAGATCATTCACCCAGACCAAGTGGGATATAACCCTGGTATGCAGGGATGGTTCAAGTGTTTGCAAAT---------GTGATACACCACATTA

5 ACAGACTGCAGAAGAAAAACCATATGGTTATCTCAATTGATGCAGAGAAAGCATTTGATAAAATTCAACACCCTTTCATGATGAAAACTCTAAGCAAATT 3097
4 ACAGACTGCAGAAGAAAAACCATATGATTATCTCAATAGATGCAGAGAAAGCATTCAATAAAACACAAGACCCTTTCATGGTGAAAACTCTAAGTAAACT
```

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, L1Oc1, L1Oc3, and L1Oc2, were placed in the alignment by inspection. The numbers at the right are for L1Oc5, the prototypical rabbit L1. The flanking direct repeats of L1Oc1 and L1Oc3 are in bold letters. The internal direct repeats in L1Oc5 and L1Oc1 are in lower-case letters. The conserved stop codon at positions 5060–5062 and the RNA polymerase II polyadenylation signal at positions 6431–6436 are underlined. Continued on pages 6 and 7.

```
5 GGGTATAGAAGGAACATTCCTCAATATAATCAAAGCAATTTATAAAAAACCCACAGCCAGCATCCTATTGAATGGGGAAAAGTTGGAAGCATTTCCACTA 3197
4 GGGTATAGAAGGAATGTTCCTCAATACAATCAAAGCAATTTATGAAAAACCCACGACCAGCATCCTATTGAATGGGAAAAAGTTGGAAGCATTTCCACTG

5 AAATCTGGCACCAGGCAGGGATGCCCACTCTCACCACTGCTATTTAACATAGTTCTGGATGTTTTAGCCAGAGCCATCAGACAAGAAAAAGAAATCAAAG 3297
4 AGATCTGGTACCATACAGGGATATCCATTCTCACCACTGCTATTCAGTATATTTCTGGAGGCTTTAGCCAGAGCTGTTAGGCAAGAAAAAGAAATTGAAG

5 GAATACAAATCAAGAAGGAAGAAGTCAAACTATCCCTCTTTGCAGACGATATGATTCTGTACTTAGAGGATCCAAAGAACTCTACTAAGAGACTATTGGA 3397
4 GGATACAAG----------

5 ACTCATAGAGGAGTTTGGCAAAGTGGCAGGATATAAAATCAATGCACAAAAATCAACAGCCTTTGTATACACAAGCAATGCCATGGCTGAGAAAGAACTG 3497
4 ----------

5 CTAAGATCAATCCCATTCACAATAGCTACAAAAACAATCAAATACCTTGGAATAAACTTAACCAAGGACGTTAAAGATCTCTACGATGAAAATTACAAAA 3597
4 ----------

5 CCTTAAAGAAAGAAATAGAAGAGGATACCAAAAAATGGAAAAATCTTCCATGCTCATGGATTGGAAGATCAACATCATCAAAATGTCCATTCTCCCAAAA 3697
4 ----------

5 GCAATTTATAGATTCAATGCAATACCAATCAAGATACCAAAGACATTCTTCTATGATCTAGAAAAAATGATGCTGAAATTCATATGGAGGCACAAGAGAC 3797
4 ----------

5 CTCGAATAGCTAAAGCAATCTTGTACAACAAAAACAAAGCCGGAGGCATCACAATACCAGACTTCAGGACATACTACAGGGCAGTAGTTATCAAAACAGC 3897
4 ----------

5 ATGGTACTGGTACAGAAACAGATGGATAGACCAATGGAACAGAATTGAAACACCAGAAATCAATCCAAACATCTACAGCCAACTTTTATTTGATCAAGGA 3997
4 ----------AGCCTACTTATATTTGATCAGGAA

5 TCTAAAACTAATTCCTGGAGCAAGGACAGTCTATTCAATAAATGGTGTTGGGAAAACTGGATTTCCACGTGCAGAAGCATGAAGCAAGACCCCTACCTTA 4097
4 TCTAAAACCAATTCCTGGAGCAAGGACAGTCTATTCAATAAATGGTGCTGGGA--------TTTCCACGTGCTGAAGCATGAAGAAAGACCCCTACCTTA

5 CATCTCACACAAAAATCCACTCAACATGGATTAAAGACCTAAATCCACGACCTGACACCATTAAGTTATTAGAGAACATTGGAGAAACCCTTCAAGATAT 4197
4 CACCTTACACAAAAATCCACTCAACATGGATTAAAGACCTAAATCTATGACCCGACACCATGAAGTTATTAGAGAACATTGGAGAAACCCTGCAAGATAT

5 TGGCACAGGCAAAGAATTTCTGGAAAAGACCCGGGAGGCACAGGCAGTCAAAGCCAAAATCAACTATTGGGATTGCATCAAATTGAGAAGTTTCTGTACT 4297
4 TGGCACCG-CAAAGACTTCTTGGAAAAGACCCTGGAGGCACAGGCAGTCAAAGCCAAAATTAACTATTGAGATTACATCAAATTGAGAAGTT-CTGTACT

5 GCAAAAGAAACAGTCAGGAGAGTGAAGAGACAACCAACAGAATGGGAAAAAATATTTGCAAACTATGCAACAGATAAAGGGTTAATAACCAGAATCTACA 4397
4 GCAAAAGAAGCAGTCAGGA-AGTGAAGAGGCAACTGACAGAATGGGAAAAAATATTTGCAAACTAAGCAACAGATAAAGCATTAATAGCTAGAATCCACA

5 AAGAAATCAAGAAACTCCACAACATCAAAACAAACAACCCACTTAAGAGATGGACCAAGGACCTCAATAGACATTTTTCAAAAGAGGAAATCCAAATGGC 4497
4 AAGACATAAAGAAACTCCACAGCATCAAAACAAACAACCCACTTAAGAGATGGGCCAAGGACCTCACTTGCCATTTTTGAAAAGAGGAAATCCAAATGGC

5 CAACAGGCACATGAAAAAATGTTCAAGGTCACTAGCAATCAGGGAAATGCAAATCAAAACCACAATGAGGTTTCACCTCACCCCGGTTAGAATGGTTCAC 4597
4 CAACATCCACATGAAGAAATGTTCAAGATCGCTAGCAATCAGGGAAATGCAAATCAAAACCACAATGAGGTTTCACCTCACCCCGCTTAGAATGGCTCAC

5 ATGCAGAAATCTACCAACAACAGATGCTGGTGAGGATGTGGGGAAAAAGGGACACTAACCCACTGTTGGTGGGAATGCAAACTGGTCAAGCCCCTATGGA 4697
4 ATACAGAAATCTACCAACAATAGATGCTGGAAAGGATGTGGGGAGAAAGGGACACTAACCCACTGTTGGTGGGAATGCAAACTGGTTAAGCCCCTATGGA

5 AATCAGTCTGGAGATTCCTCAGAAACCTGAATATAACCCTACCGTTCGACCCAGCCATCCCACTCCTTGGAATTTACCCAAAGGAGTTTAAATTGATAAA 4797
4 AGTCAGTCTGGAGAT-CCTCAGAAACCTGAATATAACCCTACCATACAACCCAGCCATCCCGCTCCT-GGAATTTACCCAAAGGAAATTAAATTGGCAAA

5 GAAAAAAGCGGTCTGCACCCTAATGTTTGTTGCAGCACAATTCACAATAGCCAACACCTGGAACCAACCTAAATGCCCATCAATGGTAGACTGGATAAAG 4897
4 CAAACAAGCTGTCTGCACCTCAATGTTTATTGCGGCTCACTTCACAATAGCTAAGACCTGGAACCAACCCAAATGCCCATCAACAGTAGACTGGAGGAAG

5 AAATTATGGGATATGTATTCTTTAGAATACTATACC---GCAGTAAGAAACAACGAAATCCAGTCATTTGCAACAAAATGGAGGAATCTGGAACACATCA 4994
4 AAATTATGGGACATGTACTCTATAGAATACTATATAGAAGCAAAAAAAAACAATGAAATCCGGTCATTTGCAACAAAATGGAGGAATCTGGAACACATCA

5 TGCTGAGTGAAGTAAGCCAGTCCCAAAGGGACAAATACCATATGTTCTCCCTGATCGGTGACAACTGACTGAACACCAAAAAGGAAACCTCCTGAAGTGA 5094
4 TGCTGAGTGAAATAAGCCAGTCCCAAAGGGACAAATATCATATGTTCTCCCTGATCAGTGACAACTGACCCAGCACCAAAAAGGAAACCTGTCAAAGTGA
1                                                    ATACTCATTCTGAACACCAAAAAGGAAATCTGTTGAAGTGA

5 AATGGACACTATGAGAAATGGTGACTTGATCAGC-ATAGCCCTGACTGCTAATGGACAACTTAATACATTATCCCTCATAGTATTTTTTTTGTCTGTTCT 5193
4 AATGAACACTATGAGAAACAGTGACTTGAACAGCCCTTGTCCTGACTGTTGATGTACAATGTAATACTTTATCCCTTTTAGTATTTTTT--GGTTGTTCT
1 AATGGACACTATGAGAAACGGTGACTTGATCAGC-AGAGCCCTGACTGTTAATGAACAACTTAATACATTATCCCTCTTATTAGTTTTTT-GTCTGTTCT

5 ACTTAATATGACTGGTTTAATTCTGTAATTATCACACAGTTATTCTTAAGTGTTGAAAATTAACTGAAATGTGATCCCTGTTAAACATAAGAGTGGCAAT 5293
4 ACTTAATACTATTGGTTGAACTCTGTAATTAACACACAATTATTCTTAGGTGTTTAAATTTAACTGAAAAGTAATCCCTGTTAAATATAAGAGTGGGAAT
1 ACTTAATATGACTGGTTTAATTCTGTAATTTATACACAGTTATTCTTAAGTGTTGAAAATCAACTAAAATGTGATCCCTGTTAAACATAAGAGTGGGAAT

5 AAGAGAGGGAAGAGATGTATAATTTGGGACATGCTCAGGCTGACTTGCCCCAATTGGTAGAGTTGGAAACATACCAGGGGATTCCAATTCAATCCCATCA 5393
4 AAGAGAGGGAAGAGATGTGCAATTCGGGACATGCTCAAACTGACTTACCTCAAATGGTAGAGTTAGAAACAGACCAGGGGATTCGAATTCAATCCCATTG
1 AAGAGAGGGAAGAGATGTACAATTTGGGACATGCTCAAGCTGACTTGCCCCAAATGGTAGAGTTAGAAACATACCAGGGGATTCCAATTCAATCCCATCA

5 AGGTGGC-ATGTGCCAATGCCATCTCACTATTCCAAGTGATCAATTTCAGTTCACAATTGATCATAATGAAAGGACTAAGAGTCAAAGGGAGCACATAAA 5492
4 AGGTGGCCATGTTCCAATGCCATCTCACTAGTCCCAGTGATCAATTTCTGTTCACAATTGATCGTAATGATAGGCATAAGAGTCAGAGGGATCACATAAA
1 AGGTGGC-ATGTACCAATGCCATCTC-CTAGTCCAAGTGATCAATTTCACTTCACAATTGATCATAATGAAAGGACTAAGAGTCAAAGGGAGCACATAAA

5 CAAGTCTAGTATCTGCTAACACTAACCGATAGAATAAATAAAGGGGAGAGTGATCCAACATGGGAAGTGAGATACTCAGCAGACTCATAGAATGGCGGAG 5592
4 CAAGACTAGTGTCTGCTAATACTAACTGACAGAATTAAAAATGGAGAGAACATTCCAACATGGGAAATGGGATACACATCAGACTCATAGAATGGCACA-
1 CAAGTCTAGTACCTGCTAACACTAACCGATAGAATAAATAAAGGGGAGAGTGATCCAACATGGGAAGTGAGGTACTCAGCAGACTCATAGAATGGCAGA-
```

Fig. 2. Continued

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 ^{32}P by nick-translation (Rigby et al. 1977) of DNA fragments or recombinant plasmids from L1Oc5 or L1Oc4.

Determination of Copy Number. The copy number of L1Oc was determined by plaque hybridization. Regions of L1Oc5 were ^{32}P-labeled and used as probes against the rabbit genomic λ library (Benton and Davis 1977) using the same hybridization

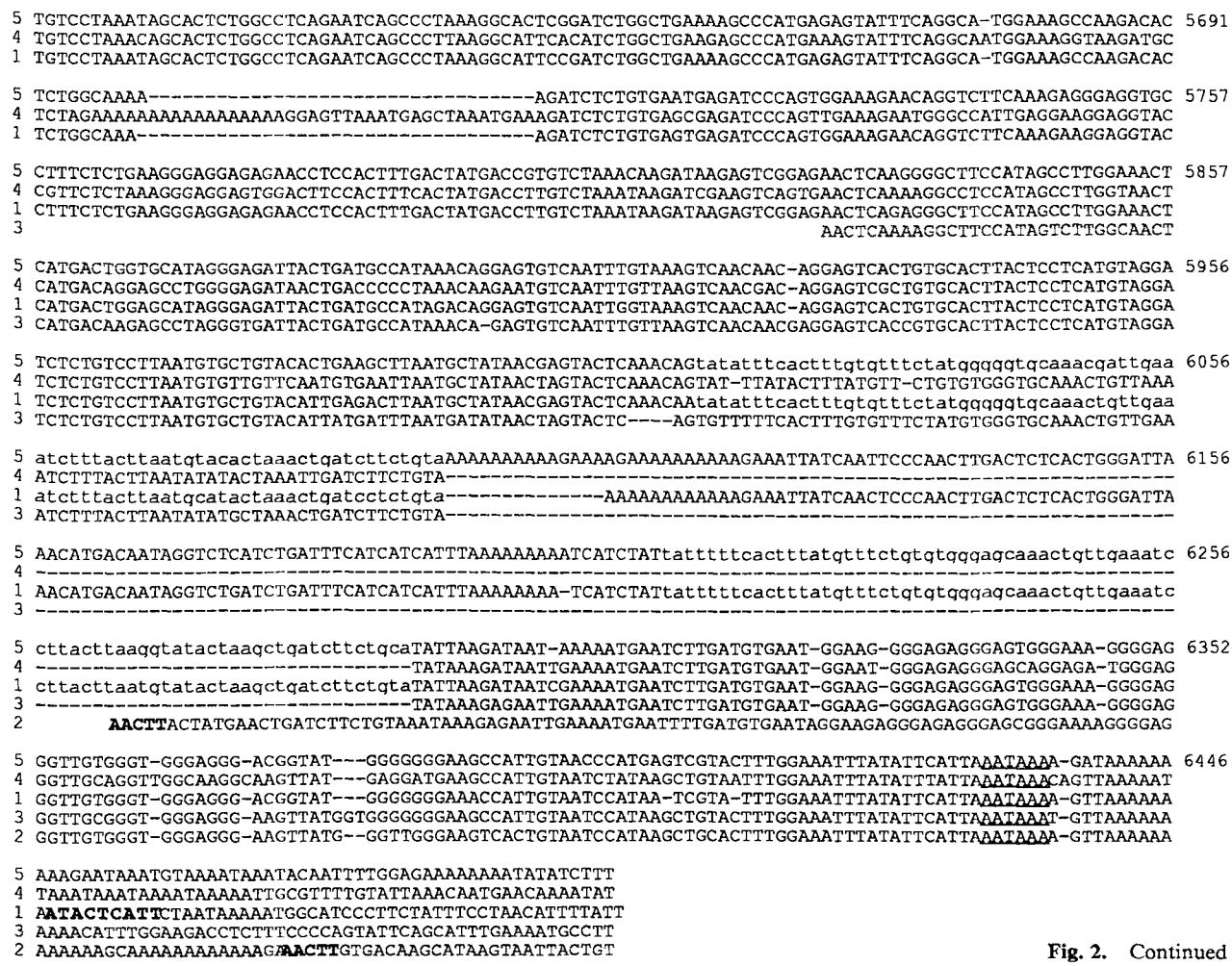

Fig. 2. Continued

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 $17 \times 10^3/3 \times 10^9$ 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, L1MdA2 (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 (Hattori et al. 1985). A consensus L1Hs sequence (Scott et al. 1987) was used in the analysis of ORF-1 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 (ψ β2) 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 L1Oc4 sequence (Fig. 2) in the sequence of the gene cluster (Margot et al. 1989).

The longest member of the rabbit L1 family in the β-like globin gene cluster is L1Oc5. The next longest member is L1Oc4; it has an internal deletion of 667 bp (Fig. 2, positions 3306–3973). This is clearly a deletion from L1Oc4 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

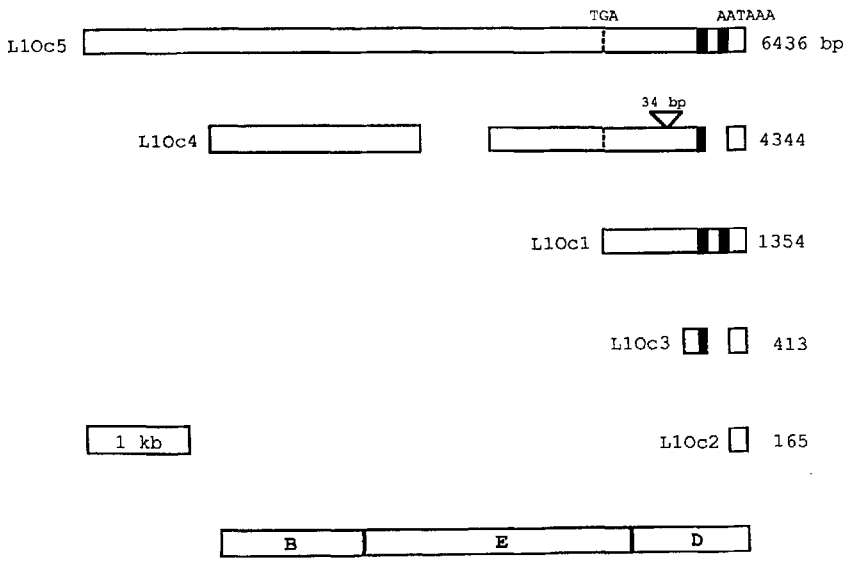

Fig. 3. Features of L1Oc revealed by the sequence alignment. The individual L1Oc repeats are shown as open boxes with the conserved termination codon, TGA, indicated by the dotted line and the polyadenylation signal, AATAAA, indicated at the 3′ (right) end. Gaps within an individual repeat are internal deletions. The 78-bp sequence that is present as a direct repeat in L1Oc5 and L1Oc1 is shown as a filled box. The other L1 members in the gene cluster have a single copy of the direct repeat. The positions of the B, E, and D repeats identified by Shen and Maniatis (1980) are shown at the bottom of the diagram.

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 L1Oc4 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 L1Oc5 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 $GTT(A)_7$ (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 $GTT(A)_7$ sequence at the target site.

The structural features revealed by the alignment and comparison of the L1 members from the rabbit β-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 β-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.

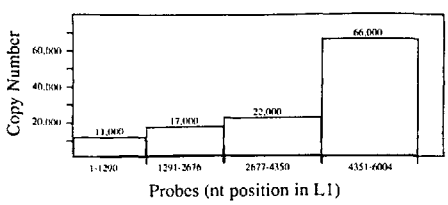

Fig. 4. Copy number of regions of L1Oc. The copy number per haploid genome is plotted as a function of the location of the probe from the L1 repeat. The location of the probe used for each region is given using the position numbers in Fig. 2.

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. The

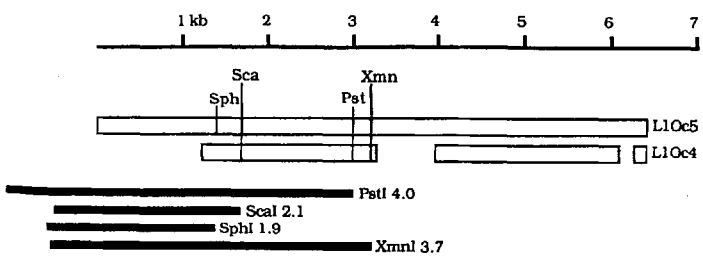

Fig. 5. Restriction map of 5′ end of L1Oc repeats. Partial restriction site maps of L1Oc5 and L1Oc4 are shown in the open boxes. The location of rabbit genomic DNA fragments (filled boxes) that hybridize to probes from L1Oc5 are shown below the restriction map; the fragments are labeled with the restriction enzyme and their size in kb.

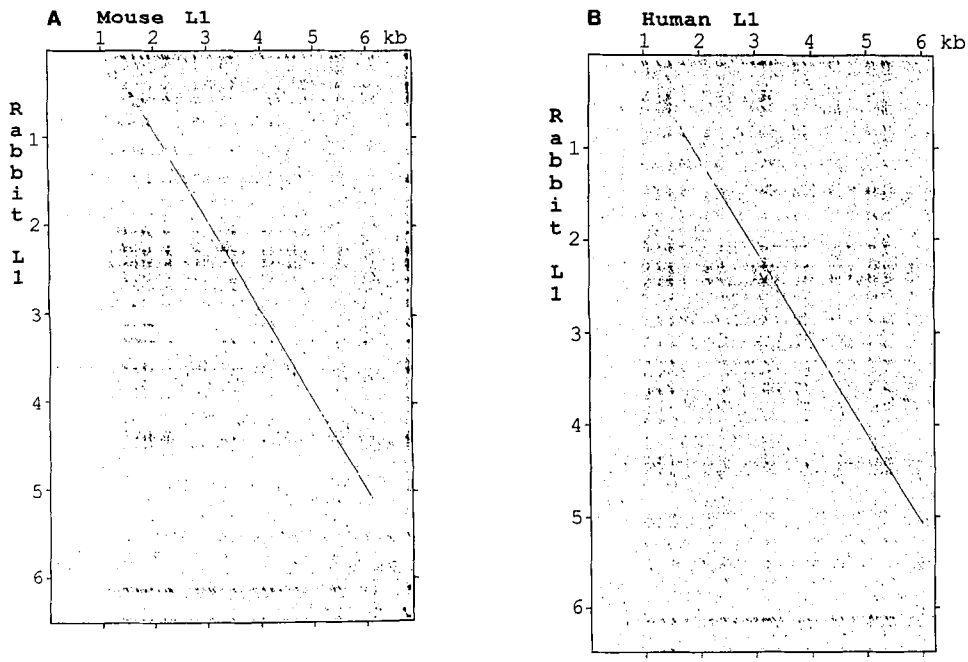

Fig. 6. Dot-plot comparisons of L1 repeats from rabbit, mouse, and human. The sequences of L1Oc5 (Fig. 2), L1MdA2 (Loeb et al. 1986), and L1Hs-TBG41 (Hattori et al. 1985) are compared using the graphical display of sequence matches generated by the program MATRIX (Zweig 1984). Segments that match at 23 out of 30 positions are shown by dots that form a diagonal. The comparison between rabbit and mouse L1 repeats is in part A, and the comparison between rabbit and human L1 repeats is in part B.

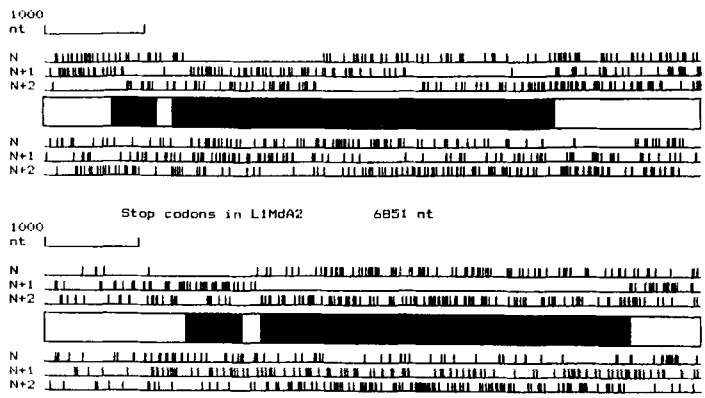

Fig. 7. Stop codons in rabbit and mouse L1 repeats. Stop codons in each reading frame are shown for L1 sequences from rabbit (top, L1Oc5), and mouse (bottom, L1MdA2). The positions of stop codons are indicated by vertical lines. The stop codon map for the L1 sequence from each species consists of the stop codons in the three reading frames of the top strand, followed by a diagram of the similar regions between species (indicated by the filled boxes—see Fig. 6), followed by the stop codons of the bottom strand.

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-

→

Fig. 8. Alignment of mammalian L1 sequences in the ORF-1 region and the beginning of ORF-2. The sequences of L1Oc5 and L1MdA2 were aligned using the program NUCALN, and the sequences of the rat L1Rn sequence (D'Ambrosio et al. 1986) and the consensus human L1Hs sequence (Scott et al. 1987) were added by inspection, using the results of dot-plot analyses (Fig. 6) and plots of stop codons (Fig. 7) as a guide. In positions where the consensus L1Hs is degenerate, the nucleotide in L1Hs-TBG41 (Hattori et al. 1985) was used. The names of the repeats are abbreviated Oc for L1Oc (rabbit), Md for L1Md (mouse), Rn for L1Rn (rat), and Hs for L1Hs (human). The nucleotide sequence is numbered beginning with the third nucleotide of the L1Oc5 sequence (Fig. 2), and the codons in the predicted translation frames for ORF-1 and ORF-2 of L1Oc are also numbered. The sequence of L1MdA2 begins at position 1648 of Loeb et al. (1986), L1Rn begins at position 1092 of D'Ambrosio et al. (1986), and L1Hs begins at position 876 of the L1 sequence in Scott et al. (1987). ORF-1 of L1Md, as defined by Loeb et al. (1986), begins at position 28, after the underlined TAA. The hyphens are gaps introduced to improve the alignment. All in-phase termination codons are underlined, and the termination

```
Oc CCG CAC CCA TCT CAA GCC TCC AAG GCT CCT CCA ACA GCA GGC AGT CCA CTT AAC ATG GAC   62
Md CCC TCC AGG TCT GCT CAT AGA GGC TAA CAG AGT CAC CTG AAG AAC AAG CTC TTA ACA GTG
Rn AAA CAG GTC TAC AGC ACT CCT GAC ACA CAG GCT TAT AGG ACA GTC TAG CCA CTG TCA GAA
Hs AAA CAG CAT CTG GAG TGG ACC TCC AGT AAA CTC CAA CAG ACC TGC AGC TGA GGG TCC TGA

Oc ACA GTA TAA AAA AAA AAA AAA AAG AAA AAA AAA AAA CGC ACA GTG ACA CAA GAA GAA TTA  122
Md ACA ACT AAA ACA GCT AGC TTC AGA GAT TAC CAG ATG GCG AAA GGC AAA CGT AAG AAT CCT
Rn ATA GCA GAA CAA AGT AAC ACT AGA GAT AAT CTG ATG GCG AAA GGC AAG CGC AGG AAC CCA
Hs CTG TTA GAA GGA AAA CTA ACA AAC AGA AAG GAC ATC CAC ACC AAA AAC CCA TCT GTA CAT

Oc ACT ATG CCG AGT AAC AAA CAC AGA AAT AGA GGG AGC AAG ATC AAC GAT GAC ACT ATG ATG  182
Md ACT AAC AGA AAT CAA GAC CAC TCA CCA TCA TCA GAA CGC AGC ACT CCC ACC CCA CCT AGT
Rn AGC AAC AGA AAC CAA GAC TAC ATG GCA CCA TCG GAG CCC AAT TCT CCC ATC AAA ACA AAC
Hs CAC CAT CAT CAA AGA CCA AAA GTA GAT AAA ACC ACA AAG ATG GGG AAA AAA CAG AGC AGA

Oc CCT CCA AAT AAG CAA AAC ACC CCA AGC CAA GAG TAT GAA GAT GAT GAG ATA GAA GAA ATG  242
Md CCT GGG CAC CCC AAC ACA ACC GAA AAT CTA GAC CCA GAT TTA AAA ACA TTT CTC ATG ATG
Rn ATG GAA TAT CCA AAC ACA CCA GAA AAG CAA GAT CTA GTT CCA AAA TCA TTT TTG ATC ATG
Hs AAA ACT GGA AAC TCT AAA AAT CAG AGT GCC TCT CCT CCT CCA AAG GAA CGC AGC TCC TCA

Oc Gln Asp Thr Asp Phe Lys Lys Phe Met Ile Arg Thr Phe Arg Ser Phe Gln Lys Gln Ile   20
Md Met Ile Glu Asp Ile Lys Lys Asp Phe His Lys Ser Leu Lys Asp Leu Gln Glu Ser Thr
Hs             Glu Gln Ser Trp Val Glu Asn Asp Phe Asp Glu Leu Arg Glu Glu Gly Phe
Oc CAA GAT ACG GAT TTC AAA AAA TTT ATG ATA AGA ACA TTT AGA AGT TTT CAA AAG CAA ATC  302
Md ATG ATA GAG GAC ATC AAG AAG GAC TTT CAT AAG TCA CTT AAA GAT TTA CAG GAG AGC ACT
Rn ATG CTG GAG GAC TTC AAG AAA GAC GTG AAG AAC TCC TTA GAG AAC AAG TAG AAG CCT ACA
Hs CCA GCA ATG GAA CAA AGC TGG GTG GAG AAT GAC TTT GAC GAG CTG AGA GAA GAA GGC TTC

Oc Leu Glu Leu Gln Lys Ser Leu Met Asp Lys Ile Glu Asn Leu Ser Arg Glu Asn Glu Ile   40
Md Ala Lys Glu Leu Gln Ala Leu Lys Glu Lys Gln Glu Asn Thr Ala Lys Gln Val Met Glu
Hs Arg Arg Ser Asn Tyr Ser Glu Leu Lys Glu Asp Val Gln Thr Lys Gly Lys Glu Val Lys
Oc CTT GAA CTA CAG AAA TCC TTA ATG GAC AAG ATT GAA AAT CTC TCT CGT GAA AAT GAA ATT  362
Md GCT AAA GAG TTA CAG GCT CTT AAA GAA AAG CAG GAA AAC ACA GCC AAA CAG GTG ATG GAA
Rn GAG AGG AAT CGC AAA AAT GCC TGA AAG AAT CGc aaa aat ccc tga aag aat tcc aag aaa 66 bp repeat
Hs AGA CGA TCA AAT TAC TCC GAG CTA AAG GAG GAC GTT CAA ACC AAA GGC AAA GAA GTT AAA

Oc Leu Arg Lys Ser Gln Asn Glu Thr Gln Lys Leu Val Glu Gln Glu Ser Val Ile Val Lys   60
Md Met Asn Lys Thr Ile Leu Glu Leu Lys Gly Glu Val Asp Thr Ile Lys Lys Thr Gln Ser
Hs Asn Phe Glu Lys Lys Leu Glu Glu Trp Ile Thr Arg Ile Thr Asn Thr Gln Lys Ser Leu
Oc TTA AGG AAG AGT CAA AAT GAA ACT CAG AAA CTA GTA GAA CAG GAA AGT GTA ATA GTG AAG  422
Md ATG AAC AAA ACC ATA CTA GAA CTA AAA GGG GAA GTA GAC ACA ATA AAG AAA ACC CAA AGC
Rn ACA ATC AAA CAG TTG AAG GAA TTA AAA ATG GAA ATA GAA GCA ATC AAA AAA GAA CAC ATG
Hs AAC TTT GAA AAA AAA TTA GAA GAA TGG ATA ACT AGA ATA ACC AAT ACA CAG AAG TCC TTA

Oc Arg Asn Gln Asn Glu Met Lys Ser Ser Ile Asp Gln Met Ala Asn Thr Leu Glu Ser Leu   80
Md Glu Ala Thr Leu Glu Ile Glu Thr Leu Gly Lys Arg Ser Gly Thr Ile Asp Ala Ser Ile
Hs Lys Asp Leu Met Glu Leu Lys Thr Lys Ala Arg Glu Leu Arg Asp Glu Cys Thr Ser Leu
Oc AGA AAT CAA AAT GAA ATG AAG AGC TCA ATA GAT CAA ATG GCA AAC ACA TTA GAA AGC CTT  482
Md GAG GCA ACG CTG GAG ATA GAA ACC CTA GGA AAG AGA TCT GGA ACC ATA GAT GCG AGC ATC
Rn GAA ACA ACC CTG GAT ATA GAA AAC CAA AAG AAG AGA CAA GGA GCT GTA GAT AAA AGC TTC
Hs AAG GAC CTG ATG GAG CTG AAA ACC AAG GCA CGA GAA CTA CGT GAC GAA TGC ACA AGC CTC

Oc Lys Asn Arg Met Gly Glu Ala Glu Asp Arg Ile Leu Asp Leu Glu Asp Arg Ala Gln Glu  100
Md Ser Asn Arg Ile Gln Glu Met Glu Glu Arg Ile Ser Gly Ala Glu Asp Ser Ile Glu Asn
Hs Ser Ser Arg Cys Asp Gln Leu Glu Glu Arg Val Ser Val Met Glu Asp Glu Met Asn Glu
Oc AAA AAC AGA ATG GGT GAA GCA GAA GAC AGA ATA TTG GAC TTA GAA GAC AGA GCA CAG GAA  542
Md AGC AAC AGA ATA CAA GAA ATG GAA GAG AGA ATC TCA GGT GCA GAA GAT TCC ATA GAG AAC
Rn ACC AAC AGA ATA CAA GAG ATG GAA GAG AGA ATC TCA GGA GCA GAA GAT TCC ATA GAA ATC
Hs AGT AGC CGA TGC GAT CAA CTG GAA GAA AGG GTA TCA GTG ATG GAA GAT GAA ATG AAT GAA

Oc Ser Ile Gln Ser Asn Gln Arg Lys Glu Glu Glu Ile Arg Asn Leu Lys Asn Ile Val Gly  120
Md Ile Asp Thr Thr Val Lys Glu Asn Thr Lys Cys Lys Arg Ile Leu Thr Gln Asn Ile Gln
Hs Met Lys Gln Glu Glu Lys Phe Arg Glu Lys Arg Ile Lys Arg Asn Glu Gln Ser Leu Gln
Oc AGT ATA CAG TCA AAC CAA AGA AAA GAA GAG GAA ATT AGA AAT CTA AAA AAT ATT GTT GGG  602
Md ATC GAC ACA ACA GTC AAA GAA AAT ACA AAA TGC AAA AGG ATC CTA ACT CAA AAC ATC CAG
Rn ATT GAC TCA ACT GTC AAA GAT AAT GTA AAG CGG AAA AAG CTA CTG GTC CAA AAC ATA CAG
Hs ATG AAG CAA GAA GAG AAG TTT AGA GAA AAA AGA ATA AAA AGA AAC GAA CAA AGC CTC CAA
```

codons proposed as the end of ORF-1 are in boldface. ATG codons proposed as the start point for ORF-1 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 L1Rn sequence are also underlined. The portion of the 66-bp tandem repeat in L1Rn that is included in the alignment is in lower-case letters. Continued on pages 12 and 13.

```
Oc Asn Leu Gln Asp Thr Ile Lys Lys Thr Asn Ile Arg Val Leu Gly Val Pro Glu Gly Met  140
Md Val Ile Gln Asp Thr Met Arg Arg Pro Asn Leu Arg Ile Ile Gly Ile Asp Glu Asn Glu
Hs Glu Ile Trp Glu Tyr Val Lys Arg Pro Asn Leu Arg Leu Ile Gly Val Pro Glu Ser Asp
Oc AAT CTA CAG GAT ACT ATT AAA AAA ACC AAC ATT CGA GTT CTA GGA GTT CCT GAA GGC ATG  662
Md GTA ATC CAG GAC ACA ATG AGA AGA CCA AAC CTA CGG ATA ATA GGA ATT GAT GAG AAT GAA
Rn GAA ATC CAG GAC TCA ATG AGA AGA TCA AAC CTA AGG ATA ATA GGT ATA GAA GAG AGT GAA
Hs GAA ATA TGG GAC TAT GTG AAA AGA CCA AAT CTA CGT CTA ATT GGT GTA CCT GAA AGT GAT

Oc Glu Arg Glu --- Lys Gly Leu Glu Gly Leu Phe Ser Glu Ile Leu Ala Glu Asn Phe Pro  160
Md Asp Phe Gln Leu Lys Gly Pro Ala Asn Ile Phe Asn Lys Ile Ile Glu Glu Asn Phe Pro
Hs Gly Glu Asn Gly Thr Lys Leu Glu Asn Thr Leu Gln Asp Ile Ile Gln Glu Asn Phe Pro
Oc GAG AGA GAG --- AAA GGA TTG GAA GGC CTT TTT AGT GAG ATA CTA GCA GAG AAC TTT CCA  719
Md GAT TTT CAA CTT AAA GGG CCA GCT AAT ATC TTC AAC AAA ATA ATA GAA GAA AAC TTC CCA
Rn GAC TCC CAG CTC AAA GGA CCA GTA AAT ATC TTC AAC AAA ACC ATA GAA GAA ANC TTC CCT
Hs GGG GAG AAT GGA ACC AAG TTG GAA AAC ACT CTG CAG GAT ATT ATC CAG GAG AAC TTC CCC

Oc Gly Leu Glu Lys Asp Arg Asp Ile Leu Val Gln Glu Ala His Arg Thr Pro Asn Lys His  180
Md Asn Ile Lys Lys Glu Met Pro Met Ile Ile Gln Glu Ala Tyr Arg Thr Pro Asn Arg Leu
Hs Asn Leu Ala Arg Gln Ala Asn Ile Gln Ile Gln Glu Ile Gln Arg Thr Pro Gln Arg Tyr
Oc GGT TTG GAG AAG GAC AGA GAT ATC CTA GTA CAG GAA GCT CAT AGA ACC CCC AAT AAA CAT  779
Md AAC ATA AAA AAA GAG ATG CCC ATG ATC AAT CAA GAA GCA TAC AGA ACT CCA AAT AGA CTG
Rn AAC CTA AAA AAA GAG ATA CCC ATA GAC ACA CAA GAA GCC TAC AGA ACT CCA AAT AGA TTG
Hs AAT CTA GCA AGG CAG GCC AAC ATT CAG ATT CAG GAA ATA CAG AGA ACG CCA CAA AGA TAC

Oc Asp Gln Lys Arg Ser Ser --- Arg His Val Val Ile Lys Leu Thr Thr Val Lys His Lys  200
Md Asp Gln Lys Arg Asn Ser Ser Arg His Ile Ile Ile Arg Thr Thr Asn Ala Leu Asn Lys
Hs Ser Ser Arg Arg Ala Thr Pro Arg His Ile Ile Val Arg Phe Thr Lys Val Glu Met Lys
Oc GAC CAA AAG AGA TCC TCA -CA CGA CAC GTG GTA ATT AAA CTT ACC ACA GTG AAA CAT AAA  838
Md GAC CAG AAA AGA AAT TCC TCC CGA CAC ATA ATA ATC AGA ACA ACA AAT GCA CTA AAT AAA
Rn GAC CAG AAA AGA AAC ACC TCC CGT CAC ATA ATT GTC AAA ACA CCA AAC GCA CAA AAT AAA
Hs TCC TCG AGA AGA GCA ACT CCA AGA CAC ATA ATT GTC AGA TTC ACC AAA GTT GAA ATG AAG

Oc Glu Lys Ile Leu Lys Cys Ala Arg Glu Lys His Gln Ile Thr Leu Arg Gly Ser Pro Ile  220
Md Asp Arg Ile Leu Lys Ala Val Arg Glu Lys Gly Gln Val Thr Tyr Lys Gly Arg Pro Ile
Hs Glu Lys Met Leu Arg Ala Ala Arg Glu Lys Gly Arg Val Thr His Lys Gly Lys Pro Ile
Oc GAA AAG ATC CTA AAA TGT GCA AGA GAG AAA CAT CAG ATT ACT CTC AGA GGA TCT CCA ATC  898
Md GAT AGA ATA TTA AAA GCA GTA AGG GAG AAA GGT CAA GTA ACA TAT AAA GGA AGG CCT ATC
Rn GAA AGA ATA TTA AAA ACA GTA AGG GAA AAA GGT CAA GTA ACA TAT AAA GGG AGA CCT ATC
Hs GAA AAA ATG TTA AGG GCA GCC AGA GAG AAA GGT CGG GTT ACC CAC AAA GGG AAG CCC ATC

Oc Arg Leu Thr Ala Asp Phe Ser Ser Glu Thr Leu Gln Ala Arg Arg Glu Trp Arg Asp Ile  240
Md Arg Ile Thr Pro Asp Phe Ser Pro Glu Thr Met Lys Ala Arg Arg Ala Trp Thr Asp Val
Hs Arg Leu Thr Ala Asp Leu Ser Ala Glu Thr Leu Gln Ala Arg Arg Glu Trp Gly Pro Ile
Oc AGA CTC ACA GCA GAC TTC TCA TCA GAA ACC CTA CAA GCT AGG AGG GAA TGG CGA GAC ATA  958
Md AGA ATT ACA CCA GAC TTT TCA CCA GAG ACT ATG AAA GCC AGA AGA GCC TGG ACA GAT GTT
Rn AGA ATC ACA CCA GAC TTC TCG CCA GAA ACT ATG AAG GCC AGA AGA TCC TGG ACT GAT GTT
Hs AGA CTA ACA GCT GAT CTC TCG GCA GAA ACT CTA CAA GCC AGA AGA GAG TGG GGG CCA ATA

Oc Ala Gln Val Leu Arg Glu Lys Asn Cys Gln Pro Arg Ile Leu Tyr Pro Ala Lys Leu Ser  260
Md Ile Gln Thr Leu Arg Glu His Lys Cys Gln Pro Arg Leu Leu Tyr Pro Ala Lys Leu Ser
Hs Phe Asn Ile Leu Lys Glu Lys Asn Phe Gln Pro Arg Ile Ser Tyr Pro Ala Lys Leu Ser
Oc GCA CAG GTG CTA AGA GAG AAA AAT TGC CAG CCC AGA ATA TTA TAT CCT GCC AAG CTC TCA 1018
Md ATA CAG ACA CTA AGA GAA CAC AAA TGC CAG CCC AGG CTA CTA TAC CCG GCC AAA CTC TCA
Rn ATA CAG ACC CTA AGA GAA CAC AAA TGC CAG CCC AGG TTA CTG TAT CCA GCA AAA CTC TCA
Hs TTC AAC ATT CTT AAA GAA AAG AAT TTT CAA CCC AGA ATT TCA TAT CCA GCC AAA CTA AGC

Oc Phe Val Asn Glu Gly Glu Ile Lys Thr Phe His Ser Lys Gln Lys Leu Lys Asp Phe Val  280
Md Ile Thr Ile Asp Gly Glu Thr Lys Val Phe His Asp Lys Thr Lys Phe Thr Gln Tyr Leu
Hs Phe Ile Ser Glu Gly Glu Ile Lys Tyr Phe Thr Asp Lys Gln Met Leu Arg Asp Phe Val
Oc TTT GTG AAT GAA GGT GAA ATA AAG ACC TTT CAT AGC AAA CAG AAA TTG AAA GAC TTT GTG 1078
Md ATT ACC ATA GAT GGA GAA ACC AAA GTA TTC CAC GAC AAA ACC AAG TTC ACA CAA TAT CTT
Rn ATT AAC ATT GAT GGA GAA ACC AAG ACA TTC CAT GAC AAA ACC AAA TTT ACA CAA TAT CTT
Hs TTC ATA AGT GAA GGA GAA ATA AAA TAC TTT ACA GAC AAG CAA ATG CTG AGA GAT TTT GTC
```

Fig. 8. Continued

```
Oc Ala Thr Cys Pro Ala Leu Gln Lys Ile Leu Lys Asp Val Leu His Ser Glu Thr Gln Lys  300
Md Ser Thr Asn Pro Ala Leu Gln Arg Ile Ile Thr Glu Lys Lys Gln Tyr Lys Asp Gly Asn
Hs Thr Thr Arg Pro Ala Leu Gln Glu Leu Leu Lys Glu Ala Leu Asn Met Glu Arg Asn Asn
Oc GCC ACT TGT CCG GCC CTG CAA AAG ATA CTT AAA GAT GTG CTA CAC TCA GAA ACA CAG AAA 1138
Md TCC ACG AAT CCA GCC CTT CAA AGG ATA ATA ACA GAA AAG AAA CAA TAC AAG GAC GGA AAT
Rn TCC ACA AAT CCA GCA CTA CAA AGG ATA ATA AAT GGT AAA GCC CAA CAT AAG GAG GCA AGC
Hs ACC ACC AGG CCT GCC CTA CAA GAG CTC CTG AAG GAA GCA CTA AAC ATG GAA AGG AAC AAC

Oc His Gly His Gln Tyr Glu Arg Arg Glu Arg Lys Asn Thr Tyr Gln
Md His Ala Leu Glu Gln Pro Arg Lys
Hs Arg Tyr Gln Pro Leu Gln Lys His Ala Lys Leu
Oc CAC GGC CAT CAA TAT GAA AGA AGG GAA AGG AAG AAC ACC TAC CAG TAA AAG AGC ATG GGA 1198
Md CAC GCC CTA GAA CAA CCA AGA AAG TAA --- --- --T CAT TCA ACA AAC CAA AAA GAA GAC
Rn TAT ACC CTA GAA GAA GCA AGA AAC TAA --- -TC GTC TTG GCA ACA AAA CAA AGA GAA TGA
Hs CGG TAC CAG CCA CTG CAA AAA CAT GCC AAA TTG TAA AGA CCA TCG AGG CTA GGA AGA AAC TGC

Oc                                                  Ala Gly Gln Ser His Tyr Val   7
Md                        Pro Thr Leu Thr Thr Lys Ile Lys Gly Ser Asn Asn Tyr Phe
Hs                                                  Thr Gly Ser Asn Ser His Ile
Oc AGC TCA AAG CAT ATA CTA GAA AAT ATT TCC GGG AAA ATG GCA GGG CAA AGT CAC TAC GTA 1258
Md AGC CAC AAG AAC AGA ATG CCA ACT CTA ACA ACA AAA ATA AAA GGG AGC AAC AAT TAC TTT
Rn AAG CAC ACA AAC ATA ACC TCA CAT CCA AAT ATG AAT ATA ACG GGA AGC AAT AAT CAC TAT
Hs ATC AAC TAA CGA GCA AAA TAA CCA GCT AAC ATC ATA ATG ACA GGA TCA AAT TCA CAC ATA

Oc Ser Ile Val Thr Leu Asn Ile Asn Gly Leu Asn Ser Ser Val Lys Arg His Arg Leu Asp  27
Md Ser Leu Ile Ser Leu Asn Ile Asn Gly Leu Asn Ser Pro Ile Lys Arg His Arg Leu Thr
Hs Thr Ile Leu Thr Leu Asn Val Asn Gly Leu Asn Ala Pro Ile Lys Arg His Arg Leu Ala
Oc TCA ATT GTC ACA TTG AAC ATT AAT GGT CTG AAT TCT TCA GTT AAA AGA CAC CGT TTG GAT 1318
Md TCC TTA ATA TCT CTT AAT ATC AAT GGA CTC AAT TCC CCA ATA AAA AGA CAT AGA CTA ACA
Rn TCC TTA ATA TCT CTC AAC ATC AAT GGC CTC AAC TCC CCA ATA AAA AGT CAT AGA TTA ACA
Hs ACA ATA TTA ACC TTA AAT GTA AAT GGG CTA AAT GCT CCA ATT AAA AGA CAC AGA CTG GCA

Oc Asp Trp Leu Thr Glu His Asn Pro Thr Ile Cys Cys Leu Gln Glu Thr His Leu Ser Asn  47
Md Asp Trp Leu His Lys Gln Asp Pro Thr Phe Cys Cys Leu Gln Glu Thr His Leu Arg Glu
Hs Asn Trp Ile Lys Ser Gln Asp Pro Ser Val Cys Cys Ile Gln Glu Thr His Leu Thr Cys
Oc GAC TGG CTC ACA GAA CAC AAC CCA ACT ATT TGT TGC CTA CAA GAA ACA CAT CTC TCT AAC 1378
Md GAC TGG CTA CAC AAA CAG GAC CCA ACA TTC TGC TGC TTA CAG GAA ACC CAT CTC AGG GAA
Rn AAC TGG ATA CAC AAC GAG GAC CCT GCA TTC TGC TGC CTA CAG GAA ACA CAC CTC AGA GAC
Hs AAT TGG ATA AAG AGT CAA GAC CCA TCA GTG TGC TGT ATT CAG GAA ACC CAT CTC ACG TGC

Oc Lys Glu Ala Cys Arg Leu Lys Val Lys Gly Trp Lys Lys Ile Phe His Ala Asn Arg Asn  67
Md Lys Asp Arg His Tyr Leu Arg Val Lys Gly Trp Lys Thr Ile Phe Gln Ala Asn Gly Leu
Hs Arg Asp Thr His Arg Leu Lys Ile Lys Gly Trp Arg Lys Ile Tyr Gln Ala Asn Gly Lys
Oc AAA GAG GCA TGC AGA CTG AAA GTG AAA GGT TGG AAA AAG ATA TTC CAT GCC AAC AGA AAC 1438
Md AAA GAC AGA CAC TAC CTC AGA GTG AAA GGC TGG AAA ACA ATT TTC CAA GCA AAT GGA CTG
Rn AAA GAC AGA CAC TAC CTC AGA GTG AAA GGC TGG AAA ACA AAT TTC CAA GCA AAT GGT CAG
Hs AGA GAC ACA CAT AGG CTC AAA ATA AAA GGA TGG AGG AAG ATC TAC CAA GCA AAT GGA AAA

Oc Gln Lys(Arg)Ala Gly Val Ala Ile Leu Ile Ser Asp Lys Ile Asn Phe Asn Thr Lys Thr  87
Md Lys Lys Gln Ala Gly Val Ala Ile Leu Ile Ser Asp Lys Ile Asp Phe Gln Pro Lys Val
Hs Gln Lys Lys Ala Gly Val Ala Ile Leu Val Ser Asp Lys Thr Asp Phe Lys Pro Thr Lys
Oc CAA AAA A-A GCA GGT GTA GCC ATA TTA ATA TCA GAC AAA ATA AAC TTT AAT ACA AAA ACT 1497
Md AAG AAA CAA GCT GGA GTA GCC ATT TTA ATA TCG GAT AAA ATC GAC TTC CAA CCC AAA GTT
Rn AAG AAG CAA GCT GGA GTA GCC ATT CTA ATA TCA AAT AAA ATC AAT TTC CAA CTA AAA GTC
Hs CAA AAA AAG GCA GGG GTT GCA ATC CTA GTC TCT GAT AAA ACA GAC TTT AAA CCA ACA AAG

Oc Val Lys Arg Asp Lys Glu Gly His Tyr Ile Met Ile Lys Gly Ser Ile Gln Gln Glu Asp  107
Md Ile Lys Lys Asp Lys Glu Gly His Phe Ile Leu Ile Lys Gly Lys Ile Leu Gln Glu Glu
Hs Ile Lys Arg Asp Lys Glu Gly His Tyr Ile Met Val Lys Gly Ser Ile Gln Gln Glu Glu
Oc GTT AAG AGA GAC AAA GAG GGA CAC TAT ATA ATG ATT AAG GGT TCA ATT CAA CAG GAA GAT 1557
Md ATC AAA AAA GAC AAG GAG GGA CAC TTC ATA CTC ATC AAA GGT AAA ATC CTC CAA GAG GAA
Rn ATC AAA AAA GAT AAG GAA GGA CAC TTC ATA TTC ATC AAA GGA AAA ATC CAC CAA GAT GAA
Hs ATC AAA AGA GAC AAA GAA GGC CAT TAC ATA ATG GTA AAG GGA TCT ATT CAA CAA GAA GAG
```

Fig. 8. Continued

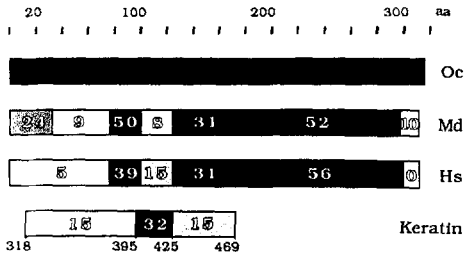

Fig. 9. Sequence similarities in the ORF-1 region. The L1Oc ORF-1 region is shown as a black box, numbered according to the codon positions in Fig. 8. The ORF-1 regions from L1Md and L1Hs are displayed as composite boxes. The darkness of the fill in each box is proportional to the extent of similarity of the L1Oc sequence. The percent identity of the encoded amino acids, compared to the L1Oc sequence, are given in the boxes. A box representing a portion of the type II cytoskeletal keratin sequence (Johnson et al. 1985) is aligned with the segment of the L1Oc sequence that matches it. The percent of amino acids identical to the L1Oc ORF-1 translated sequence is given in the boxes, and the amino acid positions in the keratin sequence are listed below the boxes. A gap penalty of -1 was assessed in calculating the percent identities.

larity corresponds to ORF-2 and the short region of similarity corresponds to the 3′ portion of ORF-1.

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 it 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 repeats from rabbit, mouse, rat, and humans. Figure 8 shows both the aligned nucleotide sequences and the predicted amino acid sequences. Sequences that match well between species are in reverse text, whereas sequences that do not match well are in plain text.

Inspection of the aligned L1 sequences allows a tentative identification of the start and stop sites of the ORFs. This analysis reveals that no overlap between reading frames is seen in rabbit and human L1 repeats. The end of ORF-1 in L1Md is the TAA at positions 1163–1165 (boldface in Fig. 8). The same sequence is found in the rat L1 sequence (L1Rn), and in-phase terminators are found nearby in L1Oc and L1Hs (boldface TAAs in Fig. 8). ORF-2 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 sequences 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 dotplots (Fig. 6). The sequence between the L1 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.

```
            10          20          30          40          50
L1  QDTDFKKFMI RTFRSFQKQI LELQKSLMDK IENLSRENEI LRKSQNETQK
     .    .      .:  ..  ..:  . .     :: .:.   .
KII KLDNLQQEID FLTALYQAEL SQMQTQISET NVILSMDNNR QFDLDSIIAE
           312         322         332         342         352

            60          70          80          90         100
L1  LVEQESVIVK RNQNEMKSSI DQMANTLESL KNRMGEAEDR ILDLEDRAQE
    .  :.  :.  .   :. :    .   . :.     : :... :    .:     .
KII VKAQNEDIAQ KSKAEAESLY QSKYEELQIT AGRHGDS-VR NSKIEISELN
           362         372         382         391         401

            110         120         130         140         150
L1  SIQSNQRKEE EIRNLKNIVG NLQDTIKKTN IRVLGVPEGM ERELKGLEGL
     .   ::     :: :.:  .. :::..:  ..  :   .           :  :: .
KII RV--IQRLRS EIDNVKKQIS NLQQSISDAE QRGENALKDA KNKLNDLEDA
           409         419         429         439         449

            160         170         180
L1  FSEILAENFP GLEKDRDILV QEAHRTPNKH
     .   :. .  : .:  . :. .
KII LQQA-KEDLT RLLRDYQELM NTKLALDLEI
           458         468         478
```

Fig. 10. Alignment of the amino acid sequences of the matching portions of ORF-1 from L1Oc and type II keratin. The sequence alignment generated by the FASTp program (Lipman and Pearson 1985) is shown starting at amino acid position 1 of ORF-1 from L1Oc5 (Fig. 8) and position 303 of the sequence of type II cytoskeletal keratin of humans (Johnson et al. 1985). The ORF-1 sequence of rabbit L1 is labeled L1, and the type II keratin sequence is labeled KII. Identical amino acids are indicated by colons, and similar amino acids are indicated by periods. The following groups of amino acids are considered similar: P, A, G, S, and T (neutral or weakly hydrophobic); Q, N, E, and D (acids and amides); H, K, and R (basic); L, I, V, and M (hydrophobic); F, Y, and W (aromatic); and C.

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 in the rabbit, mouse, and rat sequences at positions 240–242 of Fig. 8 (shown in boldface). An ATG is found three codons downstream in the human L1 sequence. Other 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 tentatively assigned as the start of ORF-1, and the codons in Fig. 8 are numbered starting here. This is 71 codons into ORF-1 as defined by Loeb et al. (1986). 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 L1Oc 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 L1Oc 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

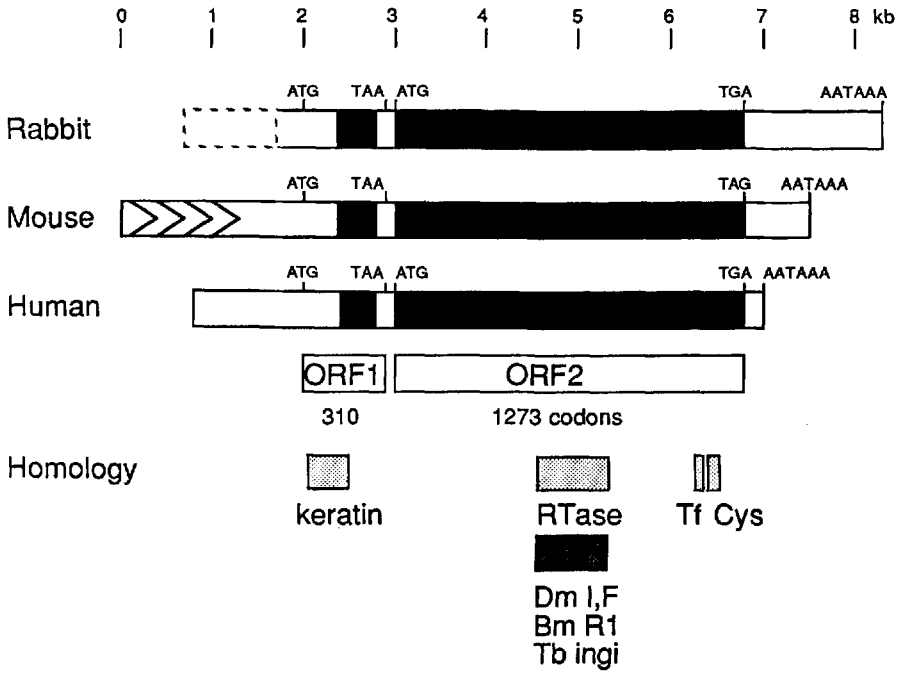

Fig. 11. Summary of regions of similarity in mammalian L1 sequences. Regions of similarity of L1s are represented by dark filled boxes, and nonconserved regions are shown as open boxes. The mouse sequence has a series of direct repeats (arrowheads) at the 5′ end. The position of the 5′ end of the rabbit L1 (dashed portion) is estimated from the genomic blot data presented in Fig. 5. The positions of ORF-1 and ORF-2 are shown below the diagrams of the L1 repeats, and regions that are similar to other proteins or repetitive elements are indicated in the lower part of the figure. Abbreviations are RTase, reverse transcriptase; Tf, transferrin; Cys, cysteine motif; DmI, I factor from *Drosophila melanogaster*; DmF, F element from *D. melanogaster*; BmR1, insertion sequence in some rRNA genes of *Bombyx mori*; Tb ingi, a repetitive element from *Trypanosoma brucei*.

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) is 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'Ambrosio 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 L1 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 L1Oc1 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 *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 AATAAA close to the 3′ end. The dicistronic structure proposed for L1Oc and L1Hs 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, Voliva CF, Martin SL, Edgell MH, Hutchison CA III (1986) Conservation throughout mammalia and extensive protein encoding capacity of the highly repeated DNA L1. J Mol Biol 187:291–304

Church GM, Gilbert W (1984) Genomic sequencing. Proc Natl Acad Sci USA 81:1991–1995

D'Ambrosio E, Waitzkin SD, Witney FR, Salemme 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) L1Oc: a long interspersed repetitive DNA element in the rabbit genome. Thesis, The Pennsylvania State University, University Park

Demers GW, Brech 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

Econonmou-Pachnis A, Lohse MA, Furano AV, Tsichlis PN (1985) Insertion of long interspersed repeated elements at the *Igh* (immunoglobulin heavy chain) and *Mlvi-2* (Moloney 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 *Kpn*I 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′-ϵ-γ-δ-β-3′ predates the mammalian radiation. Mol Biol Evol 1:390–410

Hardison RC, Butler ET, Lacy E, Maniatis T, Rosenthal N, Efstratiadis A (1979) The structure and transcription of four linked rabbit β-like globin genes. Cell 18:1285–1297

Hattori M, Hidaka S, 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

Hattori M, Kuhara S, Takenaka O, 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

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, Simoni 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-Moiyoi OK, Young JR (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

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) *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

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, Hutchison 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'Connel C, Quon D, Sim GK, Efstratiadis 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, Hutchison 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 *Kpn*I element. Proc Natl Acad Sci USA 81:1012–1016

Rigby PW, Deieckmann 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 AF, 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, Hutchison 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, Hutchison 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 retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661

Wilbur WJ, Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci USA 80:726–730

Xiong Y, Eickbush TH (1988) The site-specific ribosomal DNA-insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. Mol Cell Biol 8:114–123

Zweig SE (1984) Analysis of large nucleic acid dot matrices on small computers. Nucleic Acids Res 12:767–776

Received July 25, 1988/Revised December 9, 1988