MOLECULAR AND CELLULAR BIOLOGY, Jan. 1986, p. 168–182
0270-7306/86/010168-15$02.00/0

Vol. 6, No. 1

The Sequence of a Large L1Md Element Reveals a Tandemly Repeated 5′ End and Several Features Found in Retrotransposons

DANIEL D. LOEB, RICHARD W. PADGETT,† STEPHEN C. HARDIES,‡ W. RON SHEHEE, MARY B. COMER, MARSHALL H. EDGELL, AND CLYDE A. HUTCHISON III§*

Department of Microbiology and Immunology, Curriculum in Genetics, Program in Molecular Biology and Biotechnology, University of North Carolina, Chapel Hill, North Carolina 27514

Received 12 August 1985/Accepted 21 October 1985

The complete nucleotide sequence of a 6,851-base pair (bp) member of the L1Md repetitive family from a selected random isolate of the BALB/c mouse genome is reported here. Five kilobases of the element contains two overlapping reading frames of 1,137 and 3,900 bp. The entire 3,900-bp frame and the 3′ 600 bp of the 1,137-bp frame, when compared with a composite consensus primate L1 sequence, show a ratio of replacement to silent site differences characteristic of protein coding sequences. This more closely defines the protein coding capacity of this repetitive family, which was previously shown to possess a large open reading frame of undetermined extent. The relative organization of the 1,137- and 3,900-bp reading frames, which overlap by 14 bp, bears resemblance to protein-coding, mobile genetic elements. Homology can be found between the amino acid sequence of the 3,900-bp frame and selected domains of several reverse transcriptases. The 5′ ends of the two L1Md elements described in this report have multiple copies, 4 2/3 copies and 1 2/3 copy, of a 208-bp direct tandem repeat. The sequence of this 208-bp element differs from the sequence of a previously defined 5′ end for an L1Md element, indicating that there are at least two different 5′ end motifs for L1Md.

L1Md (formerly known as Bam HI [59] and MIF-1 [5]) is a major long interspersed repetitive element in the mouse genome (19, 66). The ubiquity of these sequences, in the mouse genome, can be appreciated by digesting genomic DNA with either *Eco*RI or *Bam*HI and electrophoresing through an agarose gel. After ethidium bromide staining, one sees a smear of DNA representing a continuum of differently sized DNA fragments. Superimposed on this smear are a 1.35-kilobase (kb) *Eco*RI (5, 10, 26) and 0.5-kb *Bam*HI and 4.0-kb *Bam*HI (42) discrete ethidium-staining bands. These bands represent portions of the consensus L1Md structure.

The size of L1Md has been estimated to be as large as 7 kb in length by restriction mapping (19, 66). A majority of the members of this family are not full-length copies of the consensus sequence. Most members are truncated at apparently random distances from a common 3′ end (19, 66). Therefore, extreme 5′ sequences of L1Md are represented less frequently (ca. 10,000 times in the genome) than extreme 3′ sequences of L1Md (85,000 times in the genome) (22; M. B. Comer, unpublished results). The 3′ end of individual L1Md elements contains an adenine-rich tail (19, 22, 66). This, coupled with the observation that individual L1Md elements are surrounded by small, less than 15-base pair (bp) direct repeats, suggests that individual L1Md elements are generated via an RNA intermediate which is subsequently dispersed to distant locations (67, 73). The L1Md gene family has been shown to undergo concerted evolution (5, 38).

The *Kpn*I sequence family of primates (or primate L1) shares many of the above-mentioned features of the L1Md family (for a review, see references 54, 55, 56). Primate L1 sequences have been shown to be evolutionarily related to the L1Md family by DNA hybridization and sequence analysis (36, 37, 57). More recently, it has been shown that sequences homologous to L1Md are present in a wide variety of mammals, suggesting that L1 is ancient and has been conserved through mammalian evolution (30, 71; F. H. Burton et al., J. Mol. Biol., in press).

Several investigators have noted an open reading frame (ORF) in both primate L1 and mouse L1 (35, 37, 46). Martin et al. (37) compared a 312-bp region of monkey and mouse L1 sequences, finding a ratio of silent versus replacement differences indicating that this portion of L1 has evolved under selection for protein function.

The present state of knowledge of L1Md, and L1 in general, leaves several issues unresolved. The crux of these issues concerns the function of the L1Md gene product. Correlating genotype and phenotype, which is difficult to do in mammalian genetic systems, is made even more difficult by the unique properties of L1Md. It is difficult to isolate a functional L1Md gene because of the copy number and homogeneity of the family. L1Md transcripts can be identified but appear to be heterogenous in size (18, 59) and are transcribed from both strands (28; M. B. Comer, unpublished results). Transcription studies of primate L1 indicates both heterogeneous-sized (32, 52) and homogenous-sized (32, 58) strand-specific RNAs can be found. So far no L1 protein products have been identified.

We have found the DNA sequence analysis approach fruitful in gaining insight into the function of the L1Md family (37, 38, 67) and hence, we decided to sequence a large L1Md element. The complete nucleotide sequence of the large L1Md element, L1Md-A2, has been useful in determining the genetic organization of the L1Md family and the extent of the ORF and in identifying potential regulatory sequences.

* Corresponding author.

† Present address: Department of Cellular and Developmental Biology, Harvard University, Cambridge, MA 02138.

‡ Present address: Department of Biochemistry, University of Texas Health Science Center at San Antonio, San Antonio, TX 78284.

§ Present address: Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC 27514.

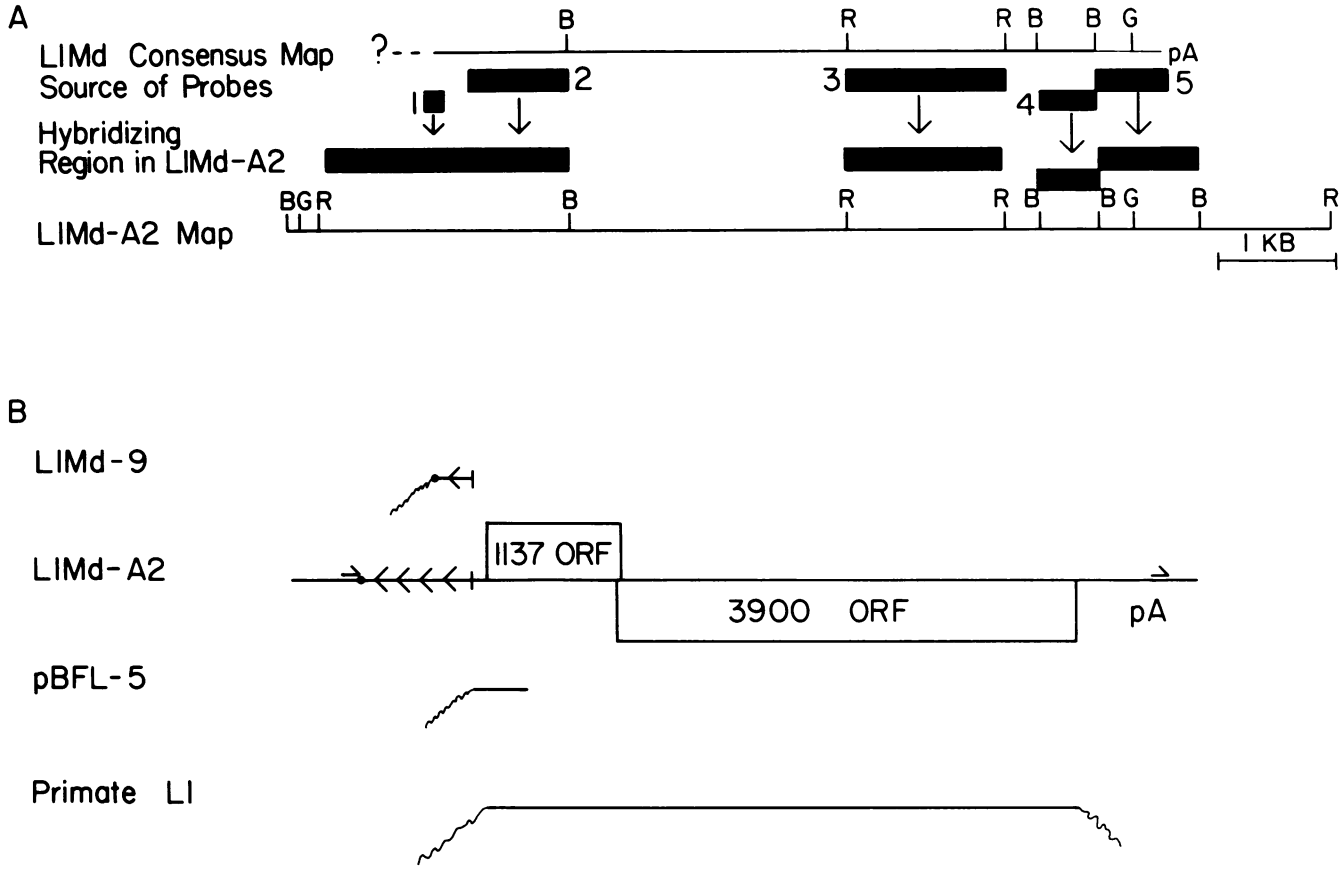


FIG. 1. (A) Restriction map of L1Md-A2 and comparison with an L1Md consensus map. Restriction enzyme sites are denoted as: B, *Bam*HI; R, *Eco*RI; G, *Bgl*I. The top line represents a genomic L1Md consensus map. pA indicates an A-rich 3′ tail. Dashed line and question mark indicate that the exact 5′ end is unknown. The second line indicates the source of the probes used in identifying L1Md-A2. Probes (see Materials and Methods) from left to right are: 1, 980-bp *Hpa*I-*Bam*HI fragment, which contains only 120 bp of L1Md sequence (the rest being single-copy genomic DNA); 2, 840-bp *Bam*HI fragment; 3, 1.35-bp *Eco*RI fragment; 4, 500-bp *Bam*HI fragment; and 5, an R-family fragment. The third line indicates which L1Md-A2 restriction fragments hybridized with individual probes. The fourth line represents the restriction map of L1Md-A2. (B) Map of sequence features of L1Md-A2 in comparison with other L1 sequences. The line labeled L1Md-A2 is a map of the L1Md-A2 sequence features. Boxed regions are the ORFs of 1,137 and 3,900 nucleotides. The 13-bp direct repeats surrounding L1Md-A2 are indicated by - - - - -. The 5′ tandemly repeating region is indicated by <——<——. Non-ORF sequences are represented by ———. The line labeled L1Md-9 indicates the position of the 5′ tandem repeats of L1Md-9 relative to L1Md-A2. The uneven wavy line represents the 5′ non-L1Md flank of L1Md-9. The main body of L1Md-9 is not represented in this diagram. The lines labeled pBFL-5 and Primate L1 indicate the homology relationships of the pBfl-5 sequence and the composite consensus primate L1 sequence, respectively, to the L1Md-A2 sequence. A straight horizontal line indicates homology with L1Md-A2, while an uneven wavy line indicates no homology with L1Md-A2.

MATERIALS AND METHODS

Identification and mapping of the lambda clone containing L1Md-A2. The lambda clone, CE102, was purified from a Charon 4A library made from BALB/c genomic DNA that was partially digested with *Eco*RI (a gift of Mark Davis). The isolation and characterization of clones carrying large L1Md elements will be described in detail elsewhere. Briefly, recombinant phage were plated at a density of 250 PFU/100-mm petri plate. Quadruplicate plaque lifts (3) were hybridized (29) with four separate probes representing the extreme left, right, and middle of L1Md. The 980-bp *Hpa*I-*Bam*HI fragment (the 3′ 980 bp of the *Bam*HI B fragment in Fig. 1 in reference 69) and the 840-bp *Bam*HI fragment (*Bam*HI fragment E in Fig. 1 in reference 69) represent the extreme left, or 5′ end, of L1Md and come from L1Md-9, a large element from the mouse beta-globin locus (8). The 1.35-kb *Eco*RI fragment (from 1.3Mm1 in reference 26) and the 500-bp *Bam*HI fragment (from L1Md-2 in reference 38) were the probes representing the middle and right end of L1Md, respectively. DNA was prepared from quadruply positive phage (34) and restriction mapped with the enzymes *Eco*RI, *Bam*HI, and *Bgl*I and the four above-mentioned probes plus an R-family probe (the 3′ end of L1Md-2 in reference 67). CE102 was chosen from 19 candidate clones.

Sequencing of the L1Md-A2 element. The 7.2-kb BglI fragment (10 μg) and the 0.84-kb *Bam*HI fragment (5 μg) (Fig. 1A) were individually isolated (16) from CE102 DNA. Random M13 subclones were generated individually for each fragment by sonicating the DNA preparation and cloning 300- to 600-bp fragments into the *Sma*I site of M13mp11 (1). Transformation of ligation products into the appropriate *Escherichia coli* host was performed by the procedure of Hanahan (24). Recombinant M13 phage growth, DNA preparation, and DNA chain-termination sequencing with [α-^{35}S]dATP as the isotope were carried out by the protocols of Bankier and Barrell (1) with slight modifications. Sequencing

reactions were performed in 96-well microtiter trays. Typically 10 clones (40 lanes) were electrophoresed through a buffer gradient sequencing gel (40 by 20 by 0.03 cm) (4). Each position in the presented sequence was determined an average of 8.3 times, for a total of over 64,000 nucleotides sequenced. The sequence was determined over 99% on both strands. The two single-stranded regions were determined multiple times, each from independent clones. Compression areas, especially in the 208-bp region (Fig. 2, position 577 to 1538), were resolved by substituting dITP for dGTP in the sequencing reactions. DBSYSTEM programs (61) were used to assemble the data.

Sequencing of the 5′ tandem repeats from L1Md-9. The cloning (69) and characterization (8) of L1Md-9 were previously described. The 5′ tandem repeats are within the 980-bp *Hpa*I-*Bam*HI and 840-bp *Bam*HI fragments described above in addition to a 200-bp *Bam*HI fragment (see Fig. 1 in reference 8 for map). These restriction fragments were cloned into M13mp10 and M13mp11 and sequenced as described above for L1Md-A2.

Mapping of the L1Md-A2 flanking sequence. Approximately 100 ng of phage DNA from sequencing preparations of clones representing the length of the 7.2-kb *Bgl*I fragment was spotted onto nitrocellulose filters and processed (3). Genomic DNA (200 ng) from *Mus domesticus* was radiolabeled with ^{32}P by nick translation (34) and hybridized to the above-mentioned filters. The BALB/c genome was probed with cloned DNA flanking the L1Md-A2 element as follows. Genomic DNA (5 μg) was restricted with either *Eco*RI or *Bam*HI and electrophoresed and blotted onto nitrocellulose (60). Insert DNA (200 ng) was individually prepared from M13 clones representing the 5′ (Fig. 2, position 1 to 550) and 3′ (Fig. 2, position 7450 to 7713) flanking regions of L1Md-A2. Each DNA was nick translated to a specific activity of 10^8 cpm/μg and hybridized to separate genomic blots.

Alignment with the consensus composite primate L1 sequence. Alignments between the L1Md-A2 sequence and the consensus composite primate L1 sequence (56) were generated with the ALIGN program of the Protein Identification Resource of the National Biomedical Research Foundation (12). Gapped positions were ignored when calculating percent homology. The significance of an alignment was determined with the statistics option of the ALIGN program. For the replacement-to-silent-site difference (R/S) ratio analysis (see below), the reading frame within the L1Md-A2 sequence was preserved during the alignment process. If a pad was introduced into the L1Md-A2 sequence by the ALIGN algorithm to align the mouse sequence with the primate sequence, additional pads were added to each sequence such that the total number of pads in a gap, in the L1Md-A2 sequence, was a multiple of three. Comparison of codons with pads was ignored during the R/S analysis.

R/S analysis. Replacement and silent changes between the primate and L1Md-A2 sequences were calculated as follows. The fraction of potential sites at which there were replacement or silent differences was individually determined for transitions and transversions and corrected for parallel and back mutations (7). To handle codons with multiple changes, we used the procedure of Miyata and Yasunaga (43) for averaging over the possible intermediate codons. Then, instead of the averaging procedure of Brown et al. (7), we back-calculated the number of changes in each category by multiplying the corrected divergence by the number of potential sites. Summing transitions and transversions, we found R and S values that estimate the actual number of changes that occurred before the obscuring effect of parallel and back mutation. A standard error which includes a contribution from the multiple hit correction was calculated (62). The R/S ratios expected for sequences evolving in the absence of selective pressure for expression of a reading frame were calculated as follows. Potential sites for replacement or silent changes were subdivided into transitions or transversions as described above. The two transition categories were multiplied by a factor representing the general excess of transitions over transversions. Transitions and transversions were summed for replacement and silent sites, respectively, and converted to an R/S ratio. With no correction for the excess of transitions over transversions, the R/S ratio tabulated for sequences not under selection for protein function was in the range of 2.9 to 3.5. Among the *Mus* L1 sequences of Martin et al. (38) there are twice as many transitions as transversions, or a fourfold excess, considering that each base pair is a potential site for two transversions but only one transition. Weighting the transitions with this factor of four reduces the above expectation values to a range of 2.2 to 2.5.

Phylogenetic tree construction. The most parsimonious tree relating the 208-bp 5′ sequences (Fig. 3) was derived as follows. Fourteen informative positions were tabulated in which more than one base was represented more than once each. By inspection we found that all but three of these were compatible with the tree shown in Fig. 3. Any rearrangement of the tree to make it compatible with each of the three discordant positions was found to introduce more parallel and back mutations than it saved. Thus, we are confident that the tree in Fig. 3 is most parsimonious. Since there is no information to support an assumption of a constant rate for the evolution of this sequence, the placement of the root on the longest branch should be considered arbitrary. Branch lengths were calculated by the method of Fitch (20).

Protein sequence database search. Version 4 (25 February 1985) of the Protein Identification Resource protein sequence library was obtained from W. R. Pearson. The program used to search the library is described by Lipmann and Pearson (33). The sequence library was searched for proteins containing any of the 64 permutations of the amino acid sequence Asp Asp X X X, where X is Ile, Leu, Met, or Val.

RESULTS

L1Md-A2 shares features with the canonical L1Md restriction map and with previously described L1Md sequence. The goal of this project was to sequence a long member of the L1Md family. Candidates were chosen from clones that hybridized to probes representing the extreme right, left, and middle of L1Md (Fig. 1A). To select an appropriate member, 19 long L1Md candidates were restriction enzyme mapped with *Eco*RI, *Bam*HI, and *Bgl*I. L1Md-A2 was chosen for sequencing because it contained six of six consensus *Eco*RI, *Bam*HI, and *Bgl*I restriction sites (Fig.1A). It contains the previously described 1.35-kb *Eco*RI fragment and the 0.5- and 4.0-kb *Bam*HI fragments in addition to the *Bgl*I site in the 3′ end (19). The mapping data also verified that this L1Md element is not rearranged or scrambled but canonical in organization (Fig. 1A). Because of these features, L1Md-A2 was chosen for sequencing.

The sequence of L1Md-A2 (Fig. 2) shows greater than 90% nucleotide homology to the previously described L1Md sequences from the MIF-1 (6, 67), *Bam*5 (38), and R-family regions (22, 67, 73). L1Md-A2 is flanked on the 3′ end by an adenine-rich region (Fig. 2, position 7,360 to 7,427), which is consistent with previously described L1Md and primate L1

5' tandem repeats — non-LIMd

```
AGGCTTATTGTGGCTTTCCCAGCTGCAGGCGGATTCTTAATCCCACACTCTTCGTTCACAGTGAAGCAGGGCTCTGCCTGCTTCCAAATTCCCCAGGAAGATCTTTCCTC   110

TCTTTTAGAGGCATCGACTGGAACCCAGCTAAAGGACAGGAGCTTCCCTCCACTTCCACTCTCAGATCAAAGGACCCTAGAGCGTTCCTAAGTGACTTTTGATGGTATAG   220

CCTGCCCAGGATTGGTGAAGTAGAGACCCCTAGATCAACCAGATTCTAGAAAGGGAGAATTCAGTTGAATTAAGTGTCTTTAGGACATTGAAGAGGGACAGAAAAACAAA   330

GCCTTATGCAATGGAGGGCAAGTTGGGTTGATCATTTGTGCCCTGACCACAGTGATCTTTGTCCTAGACCTCTACTTATATTTATACTTACATCACACATCATCTCATTC   440

TTATTCATAGAGAACAGTGGGAAGTTTCCAAATGTTGCCCTTCTGGACTTATCTCTTCAGTGAGTCTAGAAGTTCCCAACATAGAGTCCTGAGTTTTTGAGTCTGTATCC   550

AAGTTACAAATAATTTCTGGGGCCCGGATCTGGGGCACAAGTCCCTTCCGCTCGACTCGAGACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGC   660

AAGGGTCCACACGGGACTCCCCACGGGACCCTAAGACCTCTGGTGAGTGGATCACAGTGCCTACCCCAATCCAATCGCACGGAACTTGAGACTGCAGTACATAGGGAAGC   770

AGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTCGACTCGTGACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGTCCACAC   880

AGGACTCCCCGCGGGACCCTAAGACCTCTGGTGAGTGGATCACAGTGCCTGCCCCAATCCAATCGCGCGCGGAACTCGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGG   990

CCTGATCTGGGGCACAAGTCCCTTCCGCTCGACTCGAGACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGTCCACACGGGACTCCCCAC   1100

GGGACCCTAAGACCTCTGGTGAGTGGATCACAGTGCCTGCCCCAATCCAATCGCGCAGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGG   1210

CACAAGTCCCTTCCGCTCGACTCGAGACTCGAGCCCCGGGCTACCTTGACAGCAGAGTCTTGCCCAACACCCGCAAGGGCCCACACGGGACTCCCCACGGGACCCTAAGA   1320

CCTCTGGTGAGTGGAACACAGCGCCTACCCCAATCCAATCGCGTGGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCTTGATCTGGGGCACAAACCCCTT   1430

CCACTCCACTCGAGCCCCGGCTACCTTGCCAGCTGAGTCGCCTGACACCCGCAAGGGCCCACACAGGATTCCACACGTGATCCTAAGACCTCTAGTGAGTGGAACACAAC   1540

TTCTGCCAGGAGTCTGGTTCGAACACCAGATATCTGGGTACCTGCCTTGCAAGAAGAGAGGCTTGCCTGCAGAGAATACTCTGCCCACTGAAACTAAGGAGAGTGCTACCC   1650

                *** Gln Ser His Leu Lys Asn Lys Leu Leu Thr Val Thr Thr Lys Thr Ala Ser Phe Arg Asp Tyr Gln Met Ala Lys Gly Lys Arg Ly   29
TCCAGGTCTGCTCATAGAGGCTAACAGAGTCACCTGAAGAACGCAGCACTCCCACCCCACCTAGTCCTGGGCACCCCAACACAACCGAAAATCTAGACCCAGATTTAA   1760

s Asn Pro Thr Asn Arg Asn Gln Asp His Ser Pro Ser Ser Glu Arg Ser Thr Pro Thr Pro Ser Pro Gly His Pro Asn Thr Thr Glu Asn Leu Asp Pro Asp Leu L   66
GAATCCTACTAACAGAAATCAAGACCACTCACCATCATCAGAACGCAGCACTCCCACCCCACCTAGTCCTGGGCACCCCAACACAACCGAAAATCTAGACCCAGATTTAA   1870

ys Thr Phe Leu Met Met Met Ile Glu Asp Ile Lys Lys Asp Phe His Lys Ser Leu Lys Asp Leu Gln Glu Ser Thr Ala Lys Glu Leu Gln Ala Leu Lys Glu Lys Gln   102
AAACATTTCTCATGATGATGATAGAGGACATCAAGAAGGACTTTCATAAGTCACTTAAAGATTTACAGGAGAGCACTGCTAAAGAGTTACAGGCTCTTAAAGAAAAAGCAG   1980

Glu Asn Thr Ala Lys Gln Val Met Glu Met Asn Lys Thr Ile Leu Glu Lys Gly Glu Val Asp Thr Ile Lys Lys Thr Gln Ser Glu Ala Thr Leu Glu Ile Glu Th   139
GAAAACACAGCCAAACAGGTGATGGAAATGAACAAAACCATACTAGAACTAAAAGGGGAAGTAGACACAATAAAGAAAACCCAAAGCGAGGCAACGCTGGAGATAGAAAC   2090
```

1137 ORF

3900 ORF

```
r Leu Gly Lys Arg Ser Gly Thr Ile Asp Ala Ser Ile Ser Asn Arg Ile Gln Glu Met Glu Glu Arg Ile Ser Gly Ala Glu Asp Ser Ile Glu Asn Ile Asp Thr Thr V  176
CCTAGGAAGAGATCTGGAACCATAGATGCGAGCATCAGCAACAGAATACAAGAAATGGAAGAGAGAATCTCAGGTGCAGAAGATTCCATAGAGAACATCGACACAACAG  2200

al Lys Glu Asn Thr Lys Cys Lys Arg Ile Leu Thr Gln Asn Ile Gln Val Ile Gln Asp Thr Met Arg Arg Pro Asn Leu Arg Ile Ile Gly Ile Asp Glu Asn Glu Asp  212
TCAAAGAAAATACAAAATGCAAAAGGATCCTAACTCAAAACATCCAGGTAATCCAGGACACAATGAGAAGACCAAACCTACGGATAATAGGAATTGATGAGAATGAAGAT  2310

Phe Gln Leu Lys Gly Pro Ala Asn Ile Phe Asn Lys Ile Ile Glu Glu Asn Phe Pro Asn Ile Lys Lys Glu Met Pro Met Ile Ile Gln Glu Ala Tyr Arg Thr Pro As  249
TTTCAACTTAAAGGGCCAGCTAATATCTTCAACAAAATAATAGAAGAAAACTTCCCAAACATAAAAAAAGAGATGCCCATGATCATACAAGAAGCATACAGAACTCCAAA  2420

n Arg Leu Asp Gln Lys Arg Asn Ser Ser Arg His Ile Ile Ile Arg Thr Thr Asn Ala Leu Asn Lys Asp Arg Ile Leu Lys Ala Val Arg Glu Lys Gly Gln Val Thr T  286
TAGACTGGACCAGAAAAGAAATTCCTCCCGACACATAATAATCAGAACAAATGCACTAAATAAAGATAGAATATTAAAAGCAGTAAGGGAGAAAGGTCAAGTAACAT  2530

yr Lys Gly Arg Pro Ile Arg Ile Thr Pro Asp Phe Ser Pro Glu Thr Met Lys Ala Arg Arg Ala Trp Thr Asp Val Ile Gln Thr Leu Arg Glu His Lys Cys Gln Pro  322
ATAAAGGAAGGCCTATCAGAATTACACCAGACTTTCACCAGAGACTATGAAAGCCAGAAGAGCCTGGACAGATGTTATACAGACACTAAGAGAACACAAATGCCAGCCC  2640

Arg Leu Leu Tyr Pro Ala Lys Leu Ser Ile Thr Ile Asp Gly Glu Thr Lys Val Phe His Asp Lys Thr Lys Phe Thr Gln Tyr Leu Ser Thr Asn Pro Ala Leu Gln Ar  359
AGGCTACTATACCCGGCCAAACTCTCAATTACCATAGATGGAGAAACCAAAGTATTCCACGACAAAACCAAGTTCACACAATATCTTTCCACGAATCCAGCCCTTCAAAG  2750

g Ile Ile Thr Glu Lys Lys Gln Tyr Lys Asp Gly Asn His Ala Leu Glu Gln Pro Arg Lys ***  379
GATAATAACAGAAAAGAAACAATACAAGGACGGAAATCACGCCCTAGAACAACCAAGAAAGTAATCATTCAACAAACCAAAAAGAAGACAGCCACAAGAACAGAATGCCA  2860
                    *** Asn Asn Gln Glu Ser Asn His Ser Thr Asn Gln Lys Glu Asp Ser His Lys Asn Arg Met Pro  21

ACTCTAACAACAAAAATAAAAGGGAGCAACAATTACTTTCCTTAATATCTCTTAATGGACTCAATTCCCCAATAAAAAGACATAGACTAACAGACTGGCTACA  2970
Thr Leu Thr Thr Lys Ile Lys Gly Ser Asn Asn Tyr Phe Ser Leu Ile Ser Leu Asn Ile Asn Gly Leu Asn Ser Pro Ile Lys Arg His Arg Leu Thr Asp Trp Leu Hi  58

CAAACAGGACCCAACATTCTGCTGCTTACAGGAAACCCATCTCAGGGAAAAAGACAGACACTACCTCAGAGTGAAAGGCTGGAAAACAATTTTCCAAGCAAATGGACTGA  3080
s Lys Gln Asp Pro Thr Phe Cys Cys Leu Gln Glu Thr His Leu Arg Glu Lys Asp Arg His Tyr Leu Arg Val Lys Gly Trp Lys Thr Ile Phe Gln Ala Asn Gly Leu L  95

AGAAACAAGCTGGAGTAGCCATTTTAAATATCGGATAAAATCGACTTCCAACCCAAAGTTATCAAAAAAGACAAGGAGGGACACTTCATACTCATCAAAGGTAAAATCCTC  3190
ys Lys Gln Ala Gly Val Ala Ile Leu Ile Ser Asp Lys Ile Asp Phe Gln Pro Lys Val Ile Leu Lys Lys Asp Lys Glu Gly His Phe Ile Leu Ile Leu Lys Gly Lys Ile Leu  131

CAAGAGGAACTCTCAATTCTGAATATCTACGCACCAAATGCAAGGGCAGCCACATTCATTAGAGACACTTTAGTAAAGCTCAAAGCATACATTGCACCTCACACAATAAT  3300
Gln Glu Glu Leu Ser Ile Leu Asn Ile Tyr Ala Pro Asn Ala Arg Ala Ala Thr Phe Ile Arg Asp Thr Leu Val Lys Leu Lys Ala Tyr Ile Ala Pro His Thr Ile Il  168

AGTGGGAGACTTCAACACACCACTTCTTCAAAGGACAGATCGTGGAAACAGAAACTAAACAGGGACACAGTGAAACTAACAGAAGTTATGAACAAATGGACCTGACAG  3410
e Val Gly Asp Phe Asn Thr Pro Leu Ser Ser Lys Asp Arg Ser Trp Lys Gln Lys Leu Asn Arg Asp Thr Val Lys Leu Thr Glu Val Met Lys Gln Met Asp Leu Thr A  205

ATATCTACAGAACATTTTATCCTAAAACAAAAGGATATACCTTCTTCTCAGCACCTCACGGGACCTTCTCCAAAATTGACCATATAATTGGTCACAAAACAGGCCTCAAT  3520
sp Ile Tyr Arg Thr Phe Tyr Pro Lys Thr Lys Gly Tyr Thr Phe Phe Ser Ala Pro His Gly Thr Phe Ser Lys Ile Asp His Ile Ile Gly His Lys Thr Gly Leu Asn  241

AGATACAAAAATATTGAAATTGTCCCATGTATCCTATCAGACCACCACCATGGCCTAAGACTGATCTTCAATAACAACATAAATAATGGAAAGCCAACATTCACGTGGAAACT  3630
Arg Tyr Lys Asn Ile Glu Ile Val Pro Cys Ile Leu Ser Asp His His Gly Leu Arg Leu Ile Phe Asn Asn Asn Ile Asn Asn Gly Lys Pro Thr Phe Thr Trp Lys Le  278

GAATAACACTCTTCTCAATGATACCTTGGTCAAGGAAGGAATAAAGAAAGAAATTAAAGACTTTTTAGAGTTTAATGAAAATGAAGCCACAACGTACCCAAACCTATGGG  3740
u Asn Asn Thr Leu Leu Asn Asp Thr Leu Val Lys Glu Gly Ile Lys Lys Glu Ile Lys Asp Phe Leu Glu Phe Asn Glu Asn Glu Ala Thr Thr Tyr Pro Asn Leu Trp A  315

ACACAATGAAAGCATTTCTAAGAGGGAAACTCATAGCGCTGAGTGCCTCCAAGAAGAAACGGGAGACAGCACATACTAGCAGCTTGACACAACACATCTAAAAGCCCTAGAA  3850
sp Thr Met Lys Ala Phe Leu Arg Gly Lys Leu Ile Ala Leu Ser Ala Ser Lys Lys Lys Arg Glu Thr Ala His Thr Ser Ser Leu Thr Thr His Leu Lys Ala Leu Glu  351
```


AAAAAGGAAGCAAATTCACCCAAGAGGAGTAGACGGCAGGAAATAATCAAACTCAGGGGTGAAATCAACCAAGTGGAAACAAGAAGAACTATTCAAAGAATTAACCAAAC 3960
Lys Lys Glu Ala Asn Ser Pro Lys Arg Ser Arg Arg Gln Glu Ile Ile Lys Leu Arg Gly Glu Ile Asn Gln Val Glu Thr Arg Arg Thr Ile Gln Arg Ile Asn Gln Th 388
GAGGAGTTGGTTCTTTGAGAAAATCAACAAGATAGATAAACCCTTAGCTAGACTCACTAAAGGGCACAGGGACAAAATCCTAATTAACAAAATCAGAAATGAAAAGGGAG 4070
r Arg Ser Trp Phe Phe Glu Lys Ile Asn Lys Ile Asp Lys Pro Leu Ala Arg Leu Thr Lys Gly His Arg Asp Lys Ile Leu Ile Asn Lys Ile Arg Asn Glu Lys Gly A 425
ACATAACAACAGATCCTGAAGAAATCCAAAACACCATCAGATCCTTCTACAAAAGGCTATACTCAACAAAACTGGAAAACCTGGACGAAATGGACAAATTTCTGGACAGA 4180
sp Ile Thr Thr Asp Pro Glu Glu Ile Gln Asn Thr Ile Arg Ser Phe Tyr Lys Arg Leu Tyr Ser Thr Lys Leu Glu Asn Leu Asp Glu Met Asp Lys Phe Leu Asp Arg 461
TACCAGGTACCAAAGTTGAATCAGGATCAAGTTGACCATCTAAACAGTCCCATATCACCTAAAGAAATAGAAGCAGTTATTAATAGTCTCCCAACCAAAAAAAAGCCCAGG 4290
Tyr Gln Val Pro Lys Leu Asn Gln Asp Gln Val Asp His Leu Asn Ser Pro Ile Ser Pro Lys Glu Ile Glu Ala Val Ile Asn Ser Leu Pro Thr Lys Lys Ser Pro Gl 498
ACCAGATGGGTTTAGTGCAGAGTTCTATCAGACCTTCAAAGAAGATCTAATTCCAATTCTGCACAAACTATTTCACAAAATAGAAGTAGAAGGTACTCTACCCAACTCAT 4400
y Pro Asp Gly Phe Ser Ala Glu Phe Tyr Gln Thr Phe Lys Glu Asp Leu Ile Pro Ile Leu His Lys Leu Phe His Lys Ile Glu Val Glu Gly Thr Leu Pro Asn Ser P 535
TTTATGAAGCCACTATTACTCTGATACCTAAACCACAGAAAGATCCAACAAAGATAGAGAACTTCAGACCAATTTCTCTTATGAATATCGATGCAAAAATCCTCAATAAA 4510
he Tyr Glu Ala Thr Ile Thr Leu Ile Pro Lys Pro Gln Lys Asp Pro Thr Lys Ile Glu Asn Phe Arg Pro Ile Ser Leu Met Asn Ile Asp Ala Lys Ile Leu Asn Lys 571
ATTCTCGCTAACCGAATCCAAGAACACATTAAAGCAATCATCCATCCTGACCAAGTAGGTTTTATTCCAGGGATGCAGGGATGGTTTAATATACGAAAATCCATCAATGT 4620
Ile Leu Ala Asn Arg Ile Gln Glu His Ile Lys Ala Ile Ile His Pro Asp Gln Val Gly Phe Ile Pro Gly Met Gln Gly Trp Phe Asn Ile Arg Lys Ser Ile Asn Va 608
AATCCATTATATAAACAAACTCAAAGACAAAAACCACATGATCATCTCGTTAGATGCAGAAAAAGCATTTGACAAGATCCAACACCCATTCATGATAAAAGTTTTGGAAA 4730
l Ile His Tyr Ile Asn Lys Leu Lys Asp Lys Asn His Met Ile Ile Ser Leu Asp Ala Glu Lys Ala Phe Asp Lys Ile Gln His Pro Phe Met Ile Lys Val Leu Glu A 645
GATCAGGAATTCAAGGCCCATACCTAAACATGATAAAAGCAATCTACAGCAAACCAGTAGCCAACATCAAAGTAAATGGAGAGAAGCTGGAAGCAATCCCACTAAAATCA 4840
rg Ser Gly Ile Gln Gly Pro Tyr Leu Asn Met Ile Lys Ala Ile Tyr Ser Lys Pro Val Ala Asn Ile Lys Val Asn Gly Glu Lys Leu Glu Ala Ile Pro Leu Lys Ser 681
GGGACTAGACAAGGCTGCCCACTTTCTCCCTACCTTTTCAACATAGTACTTGAAGTATTAGCCAGAGCAATTCGACAACAAAAGGAGATCAAGGGGATACAAATTGGAAA 4950
Gly Thr Arg Gln Gly Cys Pro Leu Ser Pro Tyr Leu Phe Asn Ile Val Leu Glu Val Leu Ala Arg Ala Ile Arg Gln Gln Lys Glu Ile Lys Gly Ile Gln Ile Gly Ly 718
AGAGGAAGTCAAAATATCACTTCTTGCAGATGATATGATAGTATATATAAGTGACCCTAAAAATTCTACCAGAGAACTCCTAAACCTGATAAACAGCTTCGGTGAAGTAG 5060
s Glu Glu Val Lys Ile Ser Leu Leu Ala Asp Asp Met Ile Val Tyr Ile Ser Asp Pro Lys Asn Ser Thr Arg Glu Leu Leu Asn Leu Ile Asn Ser Phe Gly Glu Val V 755
TTGGATATAAAATAAACTCAAACAAGTCAATGGCCTTTCTCTATACAAAGAATAAACAGGCTGAGAAAGAAATTAGGGAAACAACACCCTTCTCAATAGTCACAAATAAT 5170
al Gly Tyr Lys Ile Asn Ser Asn Lys Ser Met Ala Phe Leu Tyr Thr Lys Asn Lys Gln Ala Glu Lys Glu Ile Arg Glu Thr Thr Pro Phe Ser Ile Val Thr Asn Asn 791
ATAAAATATCTTGGCGTGACTCTAACTAAGGAGGTGAAAGATCTGTATGATAAAAACTTCAAATCTCTGAAGAAAGAAATTAAAGAAGATCTCAGAAGATGGAAAGATCT 5280
Ile Lys Tyr Leu Gly Val Thr Leu Thr Lys Glu Val Lys Asp Leu Tyr Asp Lys Asn Phe Lys Ser Leu Lys Lys Glu Ile Lys Glu Asp Leu Arg Arg Trp Lys Asp Le 828
CCCATGCTCATGGATTGGCAGGATCAACATTGTAAAAATGGCTATCTTGCCAAAAGCAATCTACAGATTCAATGCAATCCCCATCAAAATTCCAACTCAATTCTTCAACG 5390
u Pro Cys Ser Trp Ile Gly Arg Ile Asn Ile Val Lys Met Ala Ile Leu Pro Lys Ala Ile Tyr Arg Phe Asn Ala Ile Pro Ile Lys Ile Pro Thr Gln Phe Phe Asn G 865
AATTGGAAGGAGCAATTTGCAAATTTGTCTGGAATAACAAAAAACCTAGGATAGCAAAAAGTCTTCTCAAGGATAAAAGAACTTCTGGCGGAATCACCATGCCAGACCTA 5500
lu Leu Glu Gly Ala Ile Cys Lys Phe Val Trp Asn Asn Lys Lys Pro Arg Ile Ala Lys Ser Leu Leu Lys Asp Lys Arg Thr Ser Gly Gly Ile Thr Met Pro Asp Leu 901
AAGCTTTACTACAGAGCAATTGTGATAAAAACTGCATGGTACTGGTATAGAGACAGACAAGTAGACCAATGGAATAGAATTGAAGATCCAGAAATGAACCCACACACCTA 5610
Lys Leu Tyr Tyr Arg Ala Ile Val Ile Lys Thr Ala Trp Tyr Trp Tyr Arg Asp Arg Gln Val Asp Gln Trp Asn Arg Ile Glu Asp Pro Glu Met Asn Pro His Thr Ty 938

3900 ORF


TGGTCACTTGATCTTCGACAAGGGAGCTAAAACCATCCAGTGGAAGAAAGACAGCATTTTCAACAATTGGTGCTGGCACAACTGGTTGTTATCGTGTAGAAGAATGCGAA 5720
r Gly His Leu Ile Phe Asp Lys Gly Ala Lys Thr Ile Gln Trp Lys Lys Asp Ser Ile Phe Asn Asn Trp Cys Trp His Asn Trp Leu Leu Ser Cys Arg Arg Met Arg I 975
TCGATCCATACTTATCTCCTTGTACTAAGGTCAAATCTAAGTGGATCAAGGAACTTCACATAAAACCAGAGACACTGAAACTTATAGAGGAGAAAGTGGGGAAAAGCCTT 5830
le Asp Pro Tyr Leu Ser Pro Cys Thr Lys Val Lys Ser Lys Trp Ile Lys Glu Leu His Ile Lys Pro Glu Thr Leu Lys Leu Ile Glu Glu Lys Val Gly Lys Ser Leu 1011
GAAGATATGGGCACAGGGGAAAAATTCCTGAACAGAACAGCAATGGCTTGTGCTGTAAGATCGAGAATCGACAAATGGGACCTAATGAAACTGCAAAGTTTCTGCAAGGC 5940
Glu Asp Met Gly Thr Gly Glu Lys Phe Leu Asn Arg Thr Ala Met Ala Cys Ala Val Arg Ser Arg Ile Asp Lys Trp Asp Leu Met Lys Leu Gln Ser Phe Cys Lys Al 1048
AAAAGACACTGTCAATAAGACAAAAAGACCACCAACAGATTGGGAAAGGATCTTTACCTATCCTAAATCAGATAGGGGACTAATATCCAACATATATAAAGAACTCAAGA 6050
a Lys Asp Thr Val Asn Lys Thr Lys Arg Pro Pro Thr Asp Trp Glu Arg Ile Phe Thr Tyr Pro Lys Ser Asp Arg Gly Leu Ile Ser Asn Ile Tyr Lys Glu Leu Lys L 1085
AGGTGGACTTCAGAAAATCAAATAACCCCATTAAAAAATGGGGCTCAGAACTGAACAAAGAATTCTCACCCGAGGAATACCGAATGGCAGAGAAGCACTTGAAAAAATGT 6160
ys Val Asp Phe Arg Lys Ser Asn Asn Pro Ile Lys Lys Trp Gly Ser Glu Leu Asn Lys Glu Phe Ser Pro Glu Glu Tyr Arg Met Ala Glu Lys His Leu Lys Lys Cys 1121
TCAACATCCTTAATCATCAGGGAAATGCAAATCAAAACAACCCTGAGATTCCACCTCACACCAGTCAGAATGGCTAAGATCAAAAATTCAGGTGACAGCAGATGCTGGCG 6270
Ser Thr Ser Leu Ile Ile Arg Glu Met Gln Ile Lys Thr Thr Leu Arg Phe His Leu Thr Pro Val Arg Met Ala Lys Ile Lys Asn Ser Gly Asp Ser Arg Cys Trp Ar 1158
AGGATGTGGAGAAAGAGGAACACTCCTCCATTGTTGGTGGGAGTGCAGGCTTGTACAACCACTCTGGAAATCAGTCTGGCGGTTCCTCAGAAAACTGGACATAGTACTAC 6380
g Gly Cys Gly Glu Arg Gly Thr Leu Leu His Cys Trp Trp Glu Cys Arg Leu Val Gln Pro Leu Trp Lys Ser Val Trp Arg Phe Leu Arg Lys Leu Asp Ile Val Leu P 1195
CGGAGGATCCAGCAATACCTCTCCTGGGCATATATCCAGAAGATGCCCCAACAGGTAAGAAGGACACATGCTCCACTATGTTCATAGCAGCCTTATTTATAATAGCCAGA 6490
ro Glu Asp Pro Ala Ile Pro Leu Leu Gly Ile Tyr Pro Glu Asp Ala Pro Thr Gly Lys Lys Asp Thr Cys Ser Thr Met Phe Ile Ala Ala Leu Phe Ile Ile Ala Arg 1231
AGCTGGAAAGAACCTAGATGCCCCTCAACAGAGGAATGGATACAGAAAATGTGGTACATCTACACAATGGAGTACTACTCAGCTATTAAAAAGAATGAATTTATGAAATT 6600
Ser Trp Lys Glu Pro Arg Cys Pro Ser Thr Glu Glu Trp Ile Gln Lys Met Trp Tyr Ile Tyr Thr Met Glu Tyr Tyr Ser Ala Ile Lys Lys Asn Glu Phe Met Lys Ph 1268
CCTAGCCAAATGGATGGACCTGGAGGGCATCATCCTGAGTGAGGTAACACATTCACAAAGAAACTCACACAATATGTATTCACTGATAAGTGGATATTAGCCCCAAACCT 6710
e Leu Ala Lys Trp Met Asp Leu Glu Gly Ile Ile Leu Ser Glu Val Thr His Ser Gln Arg Asn Ser His Asn Met Tyr Ser Leu Ile Ser Gly Tyr *** 1300
AGGATACCCAAGATATAAGATATAATTTGCTAAACACATGAAACTCAAGGAGAATGAAGACTGAAGTGTGGACACTATGCCCCTCCTTAGATTTGGGAACAAAACACCCA 6820
TGGAAGGAGTTACAGAGACGGAGTTTGGAGCTGAGATGAAAGGATGGACCATGTAGAGACTGCCATAGCCAGGGATCCACCCCATAATCAGCATCCAAACGCTGACACCA 6930
TTGCATACACTAGCAAGATTTTATTGAAAGGACGCAGATGTAGCTGTCTCTTGTGAGACTATGCCGGGGCCCAGCAAACACAGAAGTGGATGCTCACAGTCAGCTAATGG 7040
ATGGATCATAGGGCTCCCAATGGAGGAGCTAGAGAAAGTAGCCAAGGAGCTAAAGGGATCTGCAACCCTATAGGTGGAACAACATTATGAGCTAACCAGTACCCCGGAGC 7150
TCTTGACTCTAGCTGCATATATATCAAAAGATGGCCTAGTCGGCCATCACTGGAAAGAGAGGCCCATTGGACTTGCAAACTTTATATGCCCCAGTACAGGGGAATACCAG 7260
GGCCAAAAAGGGGGAGTGGGTGGGCAGGGGAGTGGGGGTGGGTGGATATGGGGGACTTTTGGTATAGCATTGGAAATGTAAATGAGTTAAATACCTAATAAAAAATGGAA 7370
3' untranslated

FIG. 2. Nucleotide and amino acid sequence of L1Md-A2. The number at the right end of a line is a running tally of the number of residues of that particular sequence whether they be nucleotides of L1Md-A2 or amino acids of the 1,137- or 3,900-bp ORF of L1Md-A2. Canonical L1Md restriction sites are at the following nucleotides: 2225, *Bam*HI; 4737, *Eco*RI; 6110, *Eco*RI; 6385, *Bam*HI; 6893, *Bam*HI; and 7184, *Bgl*I. The arrow above the13 bp at positions 564 to 576 and positions 7428 to 7440 indicates the 13-bp direct repeats surrounding L1Md-A2. The brackets beneath the nucleotide sequence starting at position 577 and ending at position 1538 delimit the 5′ tandem repeats. The 4-nucleotide-long arrows beneath the sequence starting at positions 717, 925, 1133, and 1341 indicate the 5′ end of the four full 5′ tandem repeats. The sequence from nucleotide 577 to 716 is the truncated 5′ tandem repeat. The labels on the right side of the figure divide the sequence into non-L1Md, 5′ tandem repeats, 1,137-bp ORF, 3,900-bp ORF, 3′ untranslated, and non-L1Md.

elements. Immediately flanking this adenine-rich region is a 13-bp sequence (Fig. 2, position 7428 to 7440) which is also found at the 5′ end of this L1Md element, 6,851 bp away (Fig. 2, position 564 to 576). These 13-bp sequences are most likely the direct repeats that are the result of the insertion mechanism. These direct repeats therefore determine the boundary, to the base, of the L1Md-A2 element. To confirm that these are the endpoints of the repetitive DNA in this clone, two hybridization experiments were performed to show that L1Md-A2 is surrounded by nonrepetitive DNA. Nick-translated mouse genomic DNA was hybridized to a set of 300- to 600-bp subclones representing the entire 7.2-kb *Bgl*I fragment (Fig. 1A). Clones that map to the left of the putative 5′ end, including one that comes to within 15 bp (Fig. 2, position 561), were not highly (<50 copies) repetitive. Clones that overlap the putative 5′ end, including one that overlaps by only 129 bp, were highly repetitive. Secondly, probes flanking the direct repeats hybridized to only two or three bands on a Southern blot of genomic restriction digests.

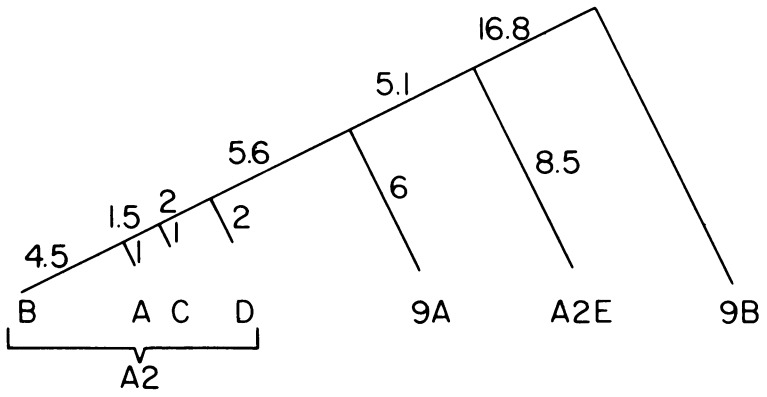
Sequence features. The L1Md-A2 element is 6,851 bp in length (Fig. 2, position 577 to 7427), excluding the 13-bp direct repeats. Two large ORFs of 1,137 (Fig. 2, position 1675 to 2811) and 3,900 (Fig. 2, position 2798 to 6697) bp are present, which overlap each other by 14 bp (Fig. 1B). These ORFs are each defined by the distance between two termination codons. The 3,900-bp ORF contains and extends 5′-ward the previously described L1Md ORF (37).

The 5′ end of the L1Md-A2 element (Fig. 2, position 577 to 1538) has 4 2/3 copies of a direct tandem repeat of 208-bp unit length. The 2/3 copy (Fig. 2, position 577 to 716) is the 5′-most member of this tandem repeating unit. Immediately adjacent, to the base, is the 13-bp sequence which is the direct repeat surrounding the element. Hybridization experiments indicate that this 208-bp sequence is a regular feature of many long L1Md elements (M. B. Comer, unpublished results). We have sequenced a homologous area (Fig. 1B) in a second L1Md element, L1Md-9, which is located in the mouse beta-globin locus (8). This L1Md element has approximately 1 2/3 copies (Fig. 3) of this same tandemly repeating motif. Each 208-bp tandem repeat, from L1Md-A2 and L1Md-9, has individual differences or polymorphisms. The homology between any two tandem units can range from 84 to 99%. A phylogenetic tree was constructed by maximum parsimony analysis of the individual tandem repeats from L1Md-A2 and L1Md-9 (Fig. 3). The distance between successive branch points on the tree is consistent with the model that individual units were added singularly to a tandem array as a function of time, instead of a single event creating a large tandem array. The decreasing branch length, as one goes from right to left on the tree, is consistent with the model that the age of the individual units decreases as one goes 3′ to 5′ in an individual L1Md element. These tandem repeats are unusual in that their G+C content is 62% (versus 40% for the rest of L1Md-A2) and that their dinucleotide CG content is high (5.7% versus 0.9% for the rest of L1Md-A2 and the genome).

This 208-bp tandem repeat, which is the 5′ end of the L1Md-A2 and L1Md-9 elements, shares no homology with a previously described 5′ end of L1Md (19). Analysis of unpublished sequence from the L1Md element in pBfl-5 (T. G. Fanning, personal communication) indicates that the L1Md-A2 element and the pBfl-5 L1Md element share extensive homology from the 3′ end, 5′-ward up to the base (Fig. 2, position 1539) where L1Md-A2 goes into the tandem repeating motif (Fig. 1B). The pBfl-5 L1Md element has a

```
1         '         '         '         '        50         '         '         '         '       100
                                                                GGATCTGGGGCACAAGTCCCTTCCGCTCGACT  L1MDA2 A
                                                                *
GTGCCTACCCCAATCCAATCGCACGGAACTTGAGACTGCAGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTCGACT  L1MDA2 B
      *         *         *
GTGCCTGCCCCAATCCAATCGCGCGGAACTCGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTCGACT  L1MDA2 C
                *
GTGCCTGCCCCAATCCAATCGCGCAGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTCGACT  L1MDA2 D
                *
GCGCCTACCCCAATCCAATCGCGTGGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCTTGATCTGGGGCACAAACCCCTTCCACTCCACT  L1MDA2 E
 *    *         *                                                  *                   **        *   *
GCGTCTGCCACAATCCAATCGAGCGGAACCTGAGACTGCAGGAAATAGGGAAGCAGACTACCCGGGCCTGACCCGGGCCACAAGTCCCTTCCACTCCACT  L1MD9 B
 * *    *         *       *        * * *         *                 * *   *                 *   *
                                                                CACAAGTCCCTCCCGCCCCTCT  L1MD9 A
                                                                        *    * **

          '         '         '         '       150         '         '         '         '       200         '
CGAG.ACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGTCCACACGGGACTCCCCACGGGACCCTAAGACCTCTGGTGAGTGGATCACA  L1MDA2 A
                                                  *         *                                 *
CGTG.ACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGTCCACACAGGACTCCCCGCGGGACCCTAAGACCTCTGGTGAGTGGATCACA  L1MDA2 B
  *                                               *                   *                 *
CGAG.ACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGTCCACACGGGACTCCCCACGGGACCCTAAGACCTCTGGTGAGTGGATCACA  L1MDA2 C
                                                  *         *                                 *
CGAG.ACTCGAGCCCCGGGCTACCTTGACAGCAGAGTCTTGCCCAACACCCGCAAGGGCCCACACGGGACTCCCCACGGGACCCTAAGACCTCTGGTGAGTGGAACACA  L1MDA2 D
                  *                               *
CGAG.........CCCCGGCTACCTTGCCAGCTGAGTC..GCCTGACACCCGCAAGGGCCCACACAGGATTCCACACGTGATCCTAAGACCTCTAGTGAGTGGAACACA  L1MDA2 E
            *         *         **                  *   *   *   *         *
CGAGCACTGGAG.........TGCCTTGCCATCGGAATC..GCCTGACACCCGCAAGAGCCCACACAGGATTCCCAACGGGATCCTAAGACCTCTAGTGAGTGGAACACA  L1MD9 B
      *         *       * *  *      **          *                 *   *   *         *
CGAGCA........CCGGGATACCTCGCCAGCGGAGTC..GACCAACACCCGCAAGGGCCCACACAGGATTCCCCACGGGATCCTAAGACCTCTGGTGAGTGGAACACA  L1MD9 A
            *     *     *       *                                 *                 *
```



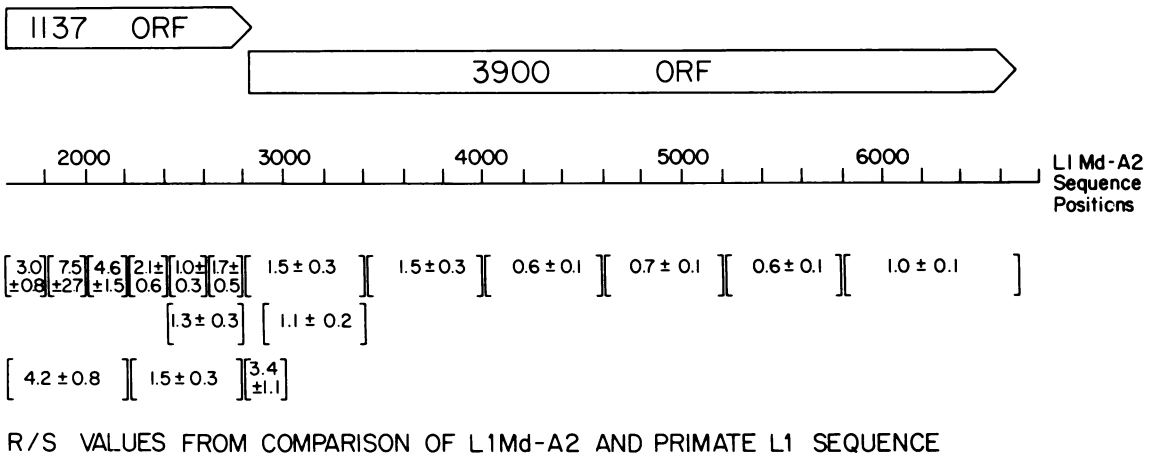


FIG. 4. R/S ratio between L1Md-A2 and the consensus primate L1 sequence. Labeled boxes representing the L1Md-A2 ORFs are drawn to scale. The line below the boxed ORFs indicates the corresponding L1Md-A2 sequence positions. The bracketed values below the sequence position line represents the R/S ratio between the two sequences for that region. The R/S ratios were corrected for multiple events as described in Materials and Methods. The R/S ratios expected for sequences evolving in the absence of selective pressure for expression of a reading frame are in the range of 2.2 to 2.5. An R/S ratio two standard error values below the unselected R/S ratio was taken to indicate selection for protein function, while an R/S ratio above this value was taken to indicate no selection for protein function. Initial sequence alignments were made as described in Materials and Methods.

different sequence at its 5′ end which is genomically repetitive as well as internally repetitive (T. G. Fanning, personal communication; R. W. Padgett, unpublished results).

Comparison with primate L1 sequence. A 6,000-bp composite consensus sequence for the primate L1 family has been compiled (56). The sequence was derived from clones from monkey and human L1 and represents the entire length of primate L1.

(i) Alignment of the 3,900-bp ORF region. The consensus primate L1 sequence was aligned with the L1Md-A2 sequence throughout the 3,900-bp ORF region (data not shown). The overall homology was 66%. The 3′ terminator of the 3,900-bp ORF (Fig. 2, position 6698 to 6700) is a conserved feature shared between primate and rodent L1. The sequence 3′ to the 3′ terminator of the 3,900-bp ORF (Fig. 2, position 6702 to 7427) cannot be aligned. The consensus primate L1 sequence, when aligned to the L1Md-A2 3,900-bp ORF, has three terminators creating frames of 400, 600, 1,300, and 1,600 bp long (data not shown). To test whether this ORF is evolving under selection for protein function, the ratio of amino acid replacement to amino acid silent (R/S) site differences between the L1Md-A2 3,900-bp ORF and the consensus primate L1 sequence was determined. An R/S ratio near 1.0 indicated selection, while an R/S ratio greater than 2.5 indicated no selection. The 3,900-bp ORF contains within it the sequence previously shown to have an R/S ratio characteristic of a protein-coding gene (37). Extending that analysis to the entire 3,900-bp ORF (Fig. 4), we find that the entire ORF, with the possible exception of the 5′-most 100 bp, has a R/S ratio near 1.0, characteristic of protein-coding sequence.

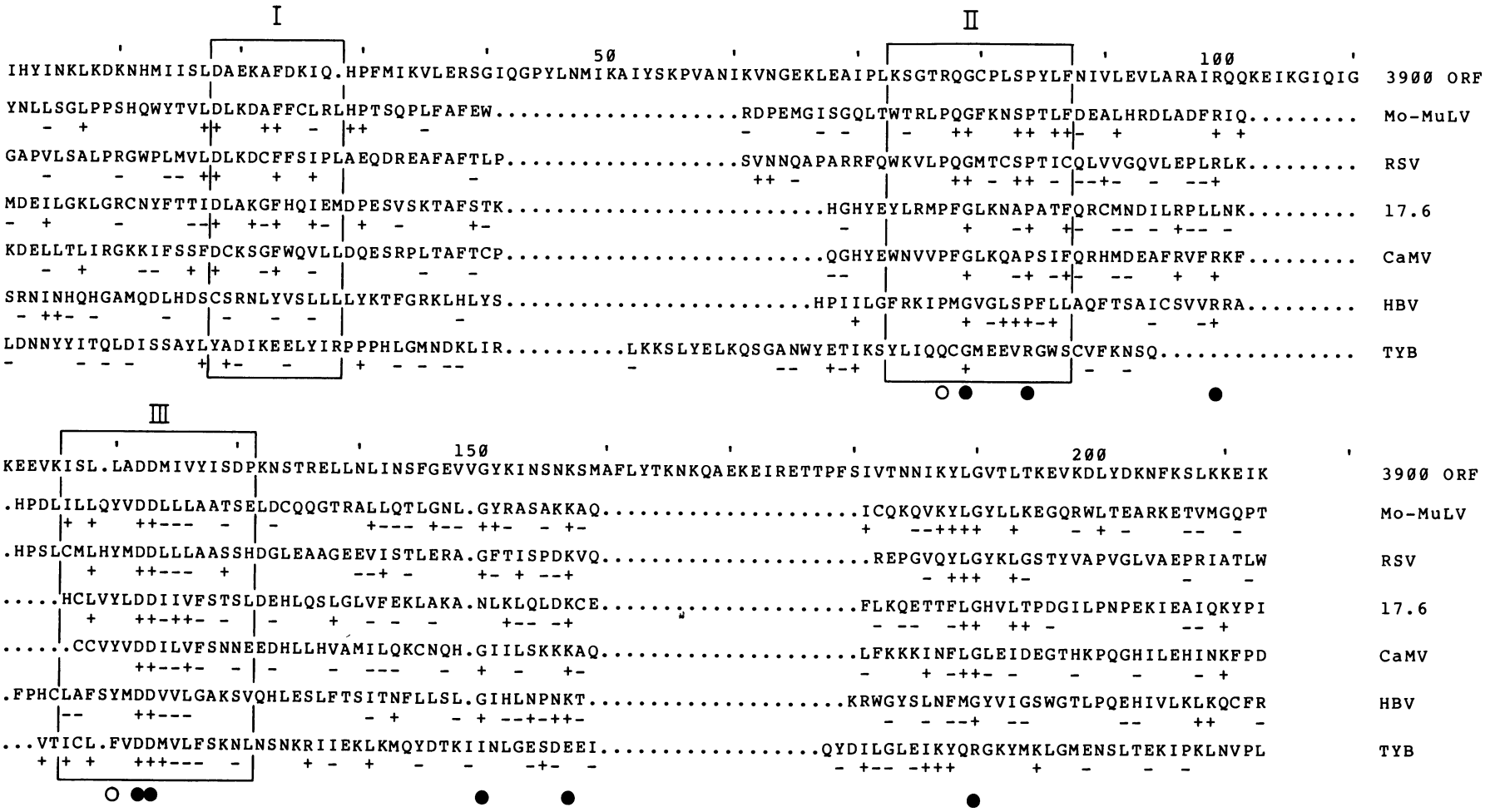
(ii) Alignment of the 1,137-bp ORF region. The 1,137-bp ORF, which is the 5′ proximal frame in L1Md-A2, has an overall homology of 55% with consensus primate L1 sequence. The homology between the two sequences is 61% in the 3′ half of this frame (Fig. 2, position 2261 to 2811) and 52% in the 5′ half (Fig. 2, position 1675 to 2260). The aligned primate sequence has four nonconserved terminators, with the longest ORF being 650 bp long. The R/S analysis (Fig. 4) indicates that the 3′ end of this frame is evolving under selection, while the 5′ end is not. The areas with the highest homology between the two sequences are evolving under selection, while the least homologous areas do not show evidence for such selection. The mouse 1,137-bp ORF does share a 3′-most terminator with the primate sequence just as the 3,900-bp ORF does.

We cannot detect significant homology between L1Md-A2 and the primate L1 sequence in the 208-bp tandemly repeating region (Fig. 2, position 577 to 1538) or in the 136 bp between the 208-bp region and the 1,137-bp ORF (Fig. 2, position 1539 to 1674).

DISCUSSION

L1Md-A2 has a total of over 5 kb of ORF, which is split into two frames of 1,137 and 3,900 bp. The size of the frames is such that it is unlikely they are open due to chance. To further test the significance of these ORFs, we aligned the L1Md-A2 sequence with a composite consensus primate L1 sequence and measured the R/S ratio in the ORFs (Fig. 4). This can be done readily within the 3,900-bp ORF region since the overall homology is 66%. The conclusion from this analysis is that essentially the entire 3,900-bp ORF region is

FIG. 3. Alignment of the 5′ direct tandem repeats from L1Md-A2 and L1Md-9. Individual unit repeats are designated at the right with a letter. The 5′-most repeat in an individual array is called A, with each succeeding 3′ unit receiving the next letter of the alphabet. Pads were added to the sequence to maximize homology. * indicates that the above nucleotide differs from the consensus nucleotide at that position. Below the sequence is a phylogenetic tree constructed by maximum parsimony analysis. The root of this tree was chosen such that branches descend to roughly the same point. Numbers on the tree indicate node heights and branch lengths in number of substitutions.


I
II
III
50
100
150
200
3900 ORF
Mo-MuLV
RSV
17.6
CaMV
HBV
TYB

evolving under selection for protein function. This means that this portion of the functional mouse and primate L1 genes is coding for a protein and that this protein most likely performs a similar function in the two species. Since homology has been shown in a wide cross-section of mammalian orders with probes from the 3,900-bp ORF region (F. H. Burton et al., in press), it is reasonable to speculate that L1 has, in general, an ORF throughout this region which generates a gene product of similar function throughout the mammalian orders.

The R/S ratio analysis between the L1Md-A2 sequence and the composite consensus primate L1 sequence in the 1,137-bp ORF region (Fig. 4) indicates that the 3′ half of the reading frame is evolving under selection while the 5′ half is not. One caveat concerning this apparent lack of selection on the 5′ half of the 1,137-bp ORF is that it is difficult to know if one is aligning highly diverged sequences with an ancestoral relationship or whether the alignment is inappropriate, since this region is highly diverged between the two species. From the R/S analysis, it appears that at least the 3′ half of the 1,137-bp ORF is evolving under selection for protein function. The existence of the 1,137-bp ORF region in L1Md-A2, coupled with the fact that L1Md-A2 has a canonical L1Md structure, suggests that the 1,137-bp ORF region is a feature of functional L1Md genes.

The 1,137- and 3,900-bp ORFs have the capacity to code for polypeptides of 43,789 and 151,839 daltons, respectively. The molecular weights of the polypeptides from the first possible initiation codon to termination codon are 41,227 and 149,590, respectively. One striking feature of both of these putative proteins is that lysine is the single most abundant amino acid and that these proteins are quite basic. This is suggestive of a DNA-binding protein. The high lysine content is a reflection of the high adenine content of the nucleotide sequence of both of the ORFs (37.5% overall).

Are the overlapping frames of L1Md-A2 a conserved feature of the L1Md family? Several additional L1Md elements were sequenced over the region of the overlapping reading frames, and none were found to have a frameshift relative to the L1Md-A2 sequence (R. W. Padgett and W. R. Shehee, unpublished results). If a sequence feature such as this is found in multiple L1Md elements, then it was apparently in a parental L1Md element. This means that L1Md elements which make copies or progeny have this arrangement. It has been shown through a nucleotide sequence analysis (S. C. Hardies et al., manuscript submitted for publication) that L1Md elements that produce progeny also evolve under selection for protein function. So, we think that the overlapping frame arrangement of L1Md-A2 is characteristic of an L1Md element competent to make more copies and produce protein.

Structural similarities to retroviruses and transposable elements. L1 has been proposed to disperse through the genome via an RNA intermediate (for a review, see reference 49). In comparison with other mobile genetic elements which have an RNA phase, a similarity in the overlapping organization and size of the L1Md-A2 ORFs to the *tya-tyb* genes of the yeast Ty1 element (11, 25), the ORF1-ORF2 genes of the *Drosophila* copia-like 17.6 element (50), and the *gag-pol* genes of several retroviruses, such as Rous sarcoma virus (51) is found. The arrangement of the 1,137-bp frame overlapping a 3,900-bp frame by 14 bp in L1Md-A2 is especially similar to the 1,332-bp *tya912* gene overlapping the 3,984-bp *tyb912* gene by 38 bp in Ty912.

A translational frame-shifting event producing a fusion precursor polypeptide has been shown to occur with Ty (11, 40) and has been proposed to occur with Rous sarcoma virus (51). This frame-shifting event is also thought to regulate the relative expression of the two reading frames. A similar mechanism could act on L1Md, suggesting that the 3,900-bp ORF would not have its own AUG translation start. This could explain why the first AUG codon in the 3,900-bp ORF shared between mouse and primate is 1,300 bp from the beginning of the 3,900-bp ORF (data not shown).

The above-mentioned mobile elements are thought to encode their own reverse transcriptases. This led us to examine L1Md-A2 and, particularly, the 3,900-bp ORF amino acid sequence for homology to known and proposed reverse transcriptase sequences. Figure 5 is an alignment indicating the homology we find between the 3,900-bp ORF sequence of L1Md-A2 and other known and proposed reverse transcriptase sequences from a wide variety of biological sources. Toh et al. (63) and Patarca and Haseltine (45) have reported regions of scattered homology between various known and proposed reverse transcriptases. Our alignment (Fig. 5) covers an area which contains boxes I, II, and III of Patarca and Haseltine (45). Overlapping this area is the region Toh et al. (Fig. 3 in reference 63) reported as conserved between various known and proposed reverse transcriptases. They report 10 invariant amino acids of 94 residues, 8 of which are present in the L1Md-A2 3,900-bp ORF. To test the significance of this homology, we searched the Protein Identification Resource protein sequence library for identity to the 64 permutations of the amino acid sequence Asp Asp X X X, where X is one of the four hydrophobic amino acids Ile, Leu, Met, or Val, a conserved sequence found in box III of Fig. 5. We found 31 proteins which contain a permutation of this sequence of 3,061. Of these 31 proteins, 23 are thought to interact with nucleic acids, with 20 of these thought to be nucleic acid polymerases. Of these 20 polymerases, 14 are known or proposed reverse transcriptases. The 17 proteins not thought to be reverse transcriptases were examined for reverse transcriptase homology outside of box III. None were found to have an extended reverse transcriptase homology outside of box III (Fig. 5). The degree of reverse transcriptase homology of the 3,900-bp ORF occurs very infrequently in known amino acid sequence and, therefore, appears significant. This evidence is suggestive that L1Md and, therefore, L1 may encode a protein with reverse transcriptase activity. This

FIG. 5. Amino acid sequence alignment of the L1Md-A2 3,900-bp ORF with known and proposed reverse transcriptases from six sources. 3900 ORF, L1Md-A2 3,900-bp ORF, residues 610 to 820; Mo-MuLV, Moloney murine leukemia virus residues 253 to 413 (53); RSV, Rous sarcoma virus, residues 90 to 250 (51); 17.6, *Drosophila melanogaster* 17.6 ORF 2, residues 284 to 434 (50); CMV, cauliflower mosaic virus, residues 323 to 472 (21); HBV, hepatitis B virus (*adr* strain), residues 465 to 620 (44); TYB, Ty912 *tyb912* gene, residues 910 to 1074 (11). All sequences are compared with the first sequence, 3900 ORF. An identity or a favored amino acid substitution between a sequence and 3900 ORF is indicated as a + or − respectively, directly beneath that residue. Favored substitutions are grouped as follows: A, S, T, P, and G; N, D, E, and Q; H, R, and K; M, L, I, and V; F, Y, and W (13). Boxed areas labeled I, II, and III are regions defined by Patarca and Haseltine (45). Positions with either an open or solid circle below them are the 10 invariant amino acid positions defined by Toh et al. (63): ●, residue conserved in the 3,900-bp ORF; ○, residue is not conserved in the 3,900-bp ORF.

concept is pleasantly consistent with the previous proposal that L1 is a mobile genetic element which moves via an RNA intermediate.

L1Md-A2 does not have long terminal repeats, which play an important role in the replication of retroviruses and transposable elements. Hence, it is interesting to speculate that the 208-bp tandem repeats in L1Md-A2 might play a functional role analogous to that of the long terminal repeats of retroviruses and transposable elements.

5′ end of L1Md-A2. The 208-bp tandemly repeating region described in L1Md-A2 is a newly observed feature of the L1Md family. This region, which is 5′ to the 1,137-bp ORF (Fig. 2, position 577 to 1672), is a reasonable area in which to look for potential regulatory sequences. This unusual structure shares some similarities with noncanonical *pol* II housekeeping promoters. It is G+C rich and has large amounts of the dinucleotide CG, which is also true for the promoter region of the housekeeping genes 3-hydroxyl-3-methylglutaryl (HMG) coenzyme A reductase (47), dihydrofolate reductase (74), adenine phosphoribosyltransferase (17), hypoxanthine phosphoribosyltransferase (41), and adenosine deaminase (64). Four 208-bp tandem repeats have a 5- of 6-bp match (Fig. 2, position 599 to 604 is the first example) with the sequence CCGCCC which is thought to be an important element in the HMG coenzyme A reductase promoter (47), the simian virus 40 promoter (for a review, see reference 68), and the herpes simplex thymidine kinase promoter (39). The four complete 208-bp tandem repeats have, in addition, a sequence complementary (Fig. 2, position 727 to 733 is the first example) to the simian virus 40 enhancer core sequence of TGGLLLG, where L is A or T (70). These sequence homologies, though provocative, are not conclusive. An interesting model of replication and transposition can be attributed to L1Md if the 208-bp region is a promoter. The model is as follows. Each individual 208-bp tandem repeat is a functional promoter. The transcription start site is within each unit, about 70 bp from the 5′ end of an individual unit. An element like L1Md-A2 would have four functional promoters and could give rise to four different-sized transcripts, each differing in size by one 208-bp tandem repeat. This model would predict that the transcriptional activity of an individual L1Md element would be a function of the number of tandem repeats present. An element such as L1Md-A2 could give rise via retrotransposition to progeny L1Mds with 3 2/3, 2 2/3, 1 2/3, and 2/3 copies of the 208-bp tandem repeat. This is consistent with the observation that individual L1Md elements have a variable number of the 208-bp tandem repeats. This would be a strategy of propagation with no net loss of sequence information, although with each transposition event at least one 208-bp unit is lost. The 208-bp tandem repeat region could then expand in size over time by unequal crossing over, thereby allowing continual rounds of transposition. This model would also predict that the truncated 208-bp repeat, which is a transcription start site, would always be 2/3 of a copy. The 5′-most 208-bp tandem repeat in L1Md-A2 and L1Md-9 are 140 and 130 bp, respectively. The HMG coenzyme A reductase promoter has five transcription start sites, four of which are clustered within a 30-bp region (47). A similar promoter mechanism could act on L1Md and explain why the 5′ ends of L1Md-A2 and L1Md-9 are within 10 bp of each other (Fig. 3).

L1Md can have different 5′ ends. Surprisingly, there are two quite different 5′ ends of L1Md, the sequence found on L1Md-A2 (A type) and the sequence found on the L1Md element in pBfl-5 (F tpye). Both of these ends are internally repetitive. The presence of at least two different 5′ ends could be an indication of different biological specificities for the L1Md elements, such as different temporal or tissue specificity of expression. The fact that both L1Md 5′ ends, which share no obvious sequence homology, are in tandem arrays suggests that there is an underlying reason for this tandem arrangement.

Is L1Md-A2 a full-length L1Md element? The issue of what is a full-length L1Md element cannot unequivocally be resolved with the present data. Restriction mapping data indicate that the longest L1Md members are approximately 7 kb. Though L1Md-A2, which is 6,851 bp in length, meets the above criterion for being full length, it is conceivable that the L1Md family contains an additional sequence 5′ to the 208-bp tandem repeat region.

Is L1Md-A2 a functional L1Md gene? Examination of sequence features of L1Md-A2 leaves the issue unresolved as to whether it is a functional gene. The majority of L1Md elements in the genome are truncated within ORF and, therefore, are pseudogenes. There are currently no criteria for determining whether there are long L1Md elements that contain the entire frame but are still defective in some other sense. We will present two arguments, one consistent with the idea that L1Md-A2 is a pseudogene and the other with L1Md-A2 being a functional gene. Experiments are in progress to test the biological activity of L1Md-A2.

Is L1Md-A2 a large processed pseudogene? A number of gene families (immunoglobulin lambda chain [27], epsilon chain [2], beta-tubulin [72], and dihydrofolate reductase [9], for example) have pseudogenes with structural features suggesting that they were generated from a functional gene via an mRNA intermediate. Typically these pseudogene sequences possess (i) an adenine-rich tail precisely where the poly(A) tail on the mature message would be; (ii) precise excision of introns, according to the rules of RNA splicing, leaving a large block of contiguous sequence which once coded for protein; and (iii) a 5′ sequence representing the 5′ untranslated flank of the mRNA. These processed pseudogenes are usually dispersed to a different chromosomal location than the functional gene and are usually flanked by short direct repeats. The L1Md-A2 structure is consistent with these features, although the overlapping ORFs and the tandemly repeating 5′ end of L1Md-A2 are two features previously not found in processed pseudogenes. It is conceivable that the functional L1Md genes are structurally different from the processed pseudogenes: e.g., they might contain introns, while the rest of the L1Md family, including L1Md-A2, are processed or truncated pseudogenes.

L1Md-A2 has features consistent with function. Examination of the L1Md-A2 sequence reveals no features inconsistent with function. The 208-bp tandem repeats of L1Md-A2 contain sequences homologous to known promoters and enhancers. The 1,137- and 3,900-bp ORF regions could code for protein. The ATG codon at position 1741 (Fig. 2) could act as a translation initiator. The apparent absence of introns within the region of the two ORFs is a feature found in mobile genetic elements which have an RNA intermediate. The 3′ region of L1Md-A2 (Fig. 2, position 7357 to 7362) contains a canonical polyadenylation sequence. Transcription and translation of overlapping reading frames such as we see in L1Md-A2 has precedent in retroviruses and transposons (11, 14, 40, 65).

If L1Md-A2 is a functional gene, two new attributes would be assigned to the L1Md family. First, since L1Md-A2 has partaken in movement, as inferred from the duplicated target site, this would mean that functional L1Md genes can move

and that L1Md is a bonafide transposable element. Second, since L1Md-A2 was randomly picked from a genomic library, with the only qualifications that it be large and canonical in organization, this would mean that a high proportion of the large L1Md elements would likely be functional. This would place the upper limit on functional genes at about 10,000.

This study increases the number of apparent structural and functional similarities between the L1Md repeat family of mouse and other, previously characterized, eucaryotic transposable elements. The overlapping reading frames are found in retroviruses as well as in transposable elements of yeast and *Drosophila*. Truncation at variable distances 5′ to an A-rich 3′ terminus is a feature shared with the F element of *Drosophila* (15). The 1723 element of *Xenopus* also contains a tandem array of approximately 200-bp repeats near one end, although unlike L1Md-A2, the actual ends of the element are inverted terminal repeats (31). The high copy number of L1 may set it apart from other well-characterized transposons but is a characteristic of some postulated retroposed elements such as the mammalian Alu family (48). Although none of these individual features is unprecedented, the particular constellation of traits found in L1Md is different from that of any of the other transposable elements mentioned above. The possible evolutionary relationship among these various elements poses an intriguing problem.

ACKNOWLEDGMENTS

We thank M. F. Singer for supplying the consensus primate L1 sequence, T. G. Fanning for communicating an unpublished sequence, and A. T. Bankier for supplying protocols and advice for sequencing.

This research was supported by Public Health Service research grants AI08998 and GM21313 from the National Institutes of Health.

LITERATURE CITED

1. **Bankier, A. T., and B. G. Barrell.** 1983. Shotgun DNA sequencing, p. 1–34. *In* R. A. Flavell (ed.), Techniques in nucleic acid biochemistry, vol. B5. Elsevier Scientific, Limerick, Ireland.
2. **Battey, J., E. E. Max, W. O. McBride, D. Swan, and P. Leder.** 1982. A processed human immunoglobulin ε gene has moved to chromosome 9. Proc. Natl. Acad. Sci. USA **79**:5956–5960.
3. **Benton, W. D., and R. W. Davis.** 1977. Screening λgt recombinant clones by hybridization to single plaques in situ. Science **196**:180–182.
4. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
5. **Brown, S. M. D., and G. Dover.** 1981. Organization and evolutionary progress of a dispersed repetitive family of sequences in widely separated rodent genomes. J. Mol. Biol. **150**:441–466.
6. **Brown, S. M. D., and M. Piechaczyk.** 1983. Insertion sequences and tandem repetitions as sources of variation in a dispersed repeat family. J. Mol. Biol. **165**:249–256.
7. **Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson.** 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.
8. **Burton, F. H., D. D. Loeb, S. F. Chao, C. A. Hutchison III, and M. H. Edgell.** 1985. Transposition of a long member of the L1 major interspersed DNA family into the mouse beta globin gene locus. Nucleic Acids Res. **13**:5071–5084.
9. **Chen, M.-J., T. Shimada, A. D. Moulton, M. Harrison, and A. W. Nienhuis.** 1982. Intronless human dihydrofolate reductase genes are derived from processed RNA molecules. Proc. Natl. Acad. Sci. USA **79**:7435–7439.
10. **Cheng, S.-M., and C. L. Schildkraut.** 1980. A family of moderately repetitive sequences in mouse DNA. Nucleic Acids Res. **8**:4075–4090.
11. **Clare, J., and P. Farabaugh.** 1985. Nucleotide sequence of a yeast Ty element: evidence for a novel mechanism of gene expression. Proc. Natl. Acad. Sci. USA **82**:2829–2833.
12. **Dayhoff, M. O.** 1978. Survey of new data and computer methods of analysis, p. 1–8. *In* M. O. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.
13. **Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt.** 1978. A model of evolutionary change in protein, p. 345–352. *In* M. O. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.
14. **Dickson, C., R. Eisenman, H. Fan, E. Hunter, and N. Teich.** 1982. Protein biosynthesis and assembly, p. 513–648. *In* R. Weiss, N. Teich, H. Varmus, and J. Coffin (ed.), Molecular biology of tumor viruses: RNA tumor viruses, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
15. **Di Nocera, P. P., M. E. Digan, and I. B. Dawid.** 1983. A family of oligo-adenylate-terminated transposable sequences in *Drosophila melanogaster*. J. Mol. Biol. **168**:715–727.
16. **Dretzen, G., M. Bellard, P. Sussone-Corsi, and P. Chambon.** 1981. A reliable method for recovery of DNA fragments from agarose and acrylamide gels. Anal. Biochem. **112**:295–298.
17. **Dush, M. K., J. M. Sikela, S. A. Khan, J. A. Tischfield, and P. J. Stambrook.** 1985. Nucleotide sequence and organization of the mouse adenine phosphoribosyltransferase gene: presence of a coding region common to animal and bacterial phosphoribosyltransferases that has a variable intron/exon arrangement. Proc. Natl. Acad. Sci. USA **82**:2731–2735.
18. **Fanning, T. G.** 1982. Characterization of a highly repetitive family of DNA sequences in the mouse. Nucleic Acids Res. **10**:5003–5013.
19. **Fanning, T. G.** 1983. Size and structure of the highly repetitive BAM HI element in mice. Nucleic Acids Res. **11**:5073–5091.
20. **Fitch, W. M.** 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.
21. **Gardner, R. C., A. J. Howarth, P. Hahn, M. Brown-Luedi, R. J. Shepherd, and J. Messing.** 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res. **9**:2871–2888.
22. **Gebhard, W., T. Meitinger, J. Hochtl, and H. G. Zachau.** 1982. A new family of interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. **157**:453–471.
23. **Grimaldi, G., J. Skowronski, and M. F. Singer.** 1984. Defining the beginning and end of *Kpn* I family segments. EMBO J. **3**:1753–1759.
24. **Hanahan, D.** 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. **166**:557–580.
25. **Hauber, J., P. Nelbock-Hochstetter, and H. Feldmann.** 1985. Nucleotide sequence and characteristics of a Ty element from yeast. Nucleic Acids Res. **13**:2745–2758.
26. **Heller, R., and N. Arnheim.** 1980. Structure and organization of the highly repeated and interspersed 1.3 kb Eco RI - Bgl II sequence family in mice. Nucleic Acids Res. **8**:5031–5042.
27. **Hollis, G. F., P. A. Hieter, O. W. McBride, D. Swan, and P. Leder.** 1982. Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature (London) **296**:321–325.
28. **Jackson, M., D. Heller, and L. Leinwand.** 1985. Transcriptional measurements of mouse repeated DNA sequences. Nucleic Acids Res. **13**:3389–3403.
29. **Jahn, C. L., C. A. Hutchison III, S. J. Phillips, S. Weaver, N. L. Haigwood, C. F. Voliva, and M. H. Edgell.** 1980. DNA sequence organization of the beta-globin complex in the BALB/c mouse. Cell **21**:159–168.
30. **Katzir, N., G. Rechavi, J. B. Cohen, T. Unger, F. Simoni, S. Segal, D. Cohen, and D. Givol.** 1985. "Retroposon" insertion into the cellular oncogene c-*myc* in canine transmissible veneral tumor. Proc. Natl. Acad. Sci. USA **82**:1054–1058.
31. **Kay, B. K., and I. G. Dawid.** 1983. The 1723 element: A long, homogeneous, highly repeated DNA unit interspersed in the genome of *Xenopus laevis*. J. Mol. Biol. **170**:583–596.
32. **Kole, L. B., S. R. Haynes, and W. R. Jelinek.** 1983. Discrete and

heterogeneous high molecular weight RNAs complementary to a long dispersed repeat family (a possible transposon) of human DNA. J. Mol. Biol. **165**:257–286.
33. **Lipmann, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.
34. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual, p. 1–545. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
35. **Manuelidis, L.** 1982. Nucleotide sequence definition of a major human repeated DNA, the Hind III 1.9 kb family. Nucleic Acids Res. **10**:3211–3219.
36. **Manuelidis, L., and P. A. Biro.** 1982. Genomic representation of the Hind II 1.9 kb repeated DNA. Nucleic Acids Res. **10**:3221–3239.
37. **Martin, S. L., C. F. Voliva, F. H. Burton, M. H. Edgell, and C. A. Hutchison III.** 1984. A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein. Proc. Natl. Acad. Sci. USA **81**:2308–2312.
38. **Martin, S. L., C. F. Voliva, S. C. Hardies, M. H. Edgell, and C. A. Hutchison III.** 1985. Tempo and mode of concerted evolution in the L1 repeat family of mice. Mol. Biol. Evol. **2**:127–140.
39. **McKnight, S. L., and R. Kingsbury.** 1982. Transcriptional control signals of a eucaryotic protein-coding gene. Science **217**:316–324.
40. **Mellor, J., S. M. Fulton, M. J. Dobson, W. Wilson, S. M. Kingsman, and A. J. Kingsman.** 1985. A retrovirus-like strategy for expression of a fusion protein encoded by yeast transposon Ty1. Nature (London) **313**:243–246.
41. **Melton, D. W., D. S. Konecki, J. Brennand, and C. T. Caskey.** 1984. Structure, expression, and mutation of the hypoxanthine phosphoribosyltransferase gene. Proc. Natl. Acad. Sci. USA **81**:2147–2151.
42. **Meunier-Rotival, M., P. Soriano, G. Cuny, F. Strauss, and G. Bernardi.** 1982. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc. Natl. Acad. Sci. USA **79**:355–359.
43. **Miyata, T., and T. Yasunaga.** 1981. Rapidly evolving mouse alpha-globin-related pseudo gene and its evolutionary history. Proc. Natl. Acad. Sci. USA **78**:450–453.
44. **Ono, Y., H. Onda, R. Sasada, K. Igarashi, Y. Sugino, and K. Nishioka.** 1983. The complete nucleotide sequences of the cloned hepatitis B virus DNA; subtype adr and adw. Nucleic Acids Res. **11**:1747–1757.
45. **Patarca, R., and W. A. Haseltine.** 1984. Sequence similarity among retroviruses—erratum. Nature (London) **309**:728.
46. **Potter, S. S.** 1984. Rearranged sequences of the human *Kpn* I element. Proc. Natl. Acad. Sci. USA **81**:1012–1016.
47. **Reynolds, G. A., S. K. Basu, T. F. Osborne, D. J. Chin, G. Gil, M. S. Brown, J. L. Goldstein, and K. L. Luskey.** 1984. HMG CoA reductase: a negatively regulated gene with unusual promoter and 5′ untranslated regions. Cell **38**:275–285.
48. **Rinehart, F. P., T. G. Ritch, P. L. Deininger, and C. W. Schmid.** 1981. Renaturation rate studies of a single family of interspersed repeated sequences in human deoxyribonucleic acid. Biochemistry **20**:3003–3010.
49. **Rogers, J. H.** 1985. Origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.
50. **Saigo, K., W. Kugimiya, Y. Matsuo, S. Inouye, K. Yoshioka, and S. Yuki.** 1984. Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature (London) **312**:659–661.
51. **Schwartz, D. E., R. Tizard, and W. Gilbert.** 1983. Nucleotide sequence of Rous sarcoma virus. Cell **32**:853–869.
52. **Shafit-Zagardo, B., F. L. Brown, P. J. Zavodny, and J. J. Maio.** 1983. Transcription of the *Kpn*I families of long interspersed DNAs in human cells. Nature (London) **304**:277–280.
53. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukaemia virus. Nature (London) **293**:543–548.
54. **Singer, M. F.** 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28**:433–434.
55. **Singer, M. F.** 1982. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. **76**:67–112.
56. **Singer, M. F., and J. Skowronski.** 1985. Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. Trends Biochem. Sci. **10**:119–122.
57. **Singer, M. F., R. E. Thayer, G. Grimaldi, M. I. Lerman, and T. G. Fanning.** 1983. Homology between the *Kpn* I primate and *Bam* HI (MIF-1) rodent families of long interspersed repeated sequences. Nucleic Acids Res. **11**:5739–5745.
58. **Skowronski, J., and M. F. Singer.** 1985. Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc. Natl. Acad. Sci. USA **82**:6050–6054.
59. **Soriano, P., M. Meunier-Rotival, and G. Bernardi.** 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. USA **80**:1816–1820.
60. **Southern E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
61. **Staden, R.** 1982. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10**:4731–4751.
62. **Tajima, F., and M. Nei.** 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.
63. **Toh, H., H. Hayashida, and T. Miyata.** 1983. Sequence homology between retroviral transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature (London) **305**:827–829.
64. **Valerio, D., M. G. C. Duyvesteyn, B. M. M. Dekker, G. Weeda, T. M. Berkvens, L. van der Voorn, H. van Ormondt, and A. J. van der Eb.** 1985. Adenosine deaminase: characterization and expression of a gene with a remarkable promoter. EMBO J. **4**:437–443.
65. **Varmus, H. E., and R. Swanstrom.** 1982. Replication of retroviruses, p. 369–512. *In* R. Weiss, N. Teich, H. Varmus, and J. Coffin, (ed.), Molecular biology of tumor viruses: RNA tumor viruses, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
66. **Voliva, C. F., C. L. Jahn, M. B. Comer, M. H. Edgell, and C. A. Hutchison III.** 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res. **11**:8847–8859.
67. **Voliva, C. F., S. L. Martin, C. A. Hutchison III, and M. H. Edgell.** 1984. The dispersal process associated with the L1 family of interspersed repetitive sequences. J. Mol. Biol. **178**:795–813.
68. **Wasylyk, B., and P. Chambon.** 1983. Potentiator effect of the SV40 72 bp repeat on initiation of transcription from heterologous promoter elements. Cold Spring Harbor Symp. Quant. Biol. **47**:921–934.
69. **Weaver, S., M. B. Comer, C. L. Jahn, C. A. Hutchison III, and M. H. Edgell.** 1981. The adult beta-globin genes of the "single" type mouse C57BL. Cell **24**:403–411.
70. **Weiher, H., M. Konig, and P. Gruss.** 1983. Multiple point mutations affecting the simian virus 40 enhancer. Science **219**:626–631.
71. **Witney, F. R., and A. V. Furano.** 1984. Highly repeated DNA families in the rat. J. Biol. Chem. **259**:10481–10492.
72. **Wilde, C. D., C. E. Crowther, M. G.-S. Cripe, and N. J. Cowan.** 1982. Evidence that a human beta-tubulin pseudogene is derived from its corresponding mRNA. Nature (London) **297**:83–84.
73. **Wilson, R., and U. Storb.** 1983. Association of two different repetitive DNA elements near immunoglobulin light chain genes. Nucleic Acids Res. **11**:1803–1816.
74. **Yang, J. K., J. N. Masters, and G. Attardi.** 1984. Human dihydrofolate reductase gene organization: extensive conservation of the (GC)-rich 5′ noncoding sequence and strong intron size divergence from homologous mammalian cells. J. Mol. Biol. **176**:169–187.