MOLECULAR AND CELLULAR BIOLOGY, Apr. 1988, p. 1385–1397
0270-7306/88/041385-13$02.00/0

Vol. 8, No. 4

Unit-Length Line-1 Transcripts in Human Teratocarcinoma Cells

JACEK SKOWRONSKI,† THOMAS G. FANNING, AND MAXINE F. SINGER*
Laboratory of Biochemistry, National Cancer Institute, Bethesda, Maryland 20892

Received 19 August 1987/Accepted 29 December 1987

We have characterized the approximately 6.5-kilobase cytoplasmic poly(A)$^+$ Line-1 (L1) RNA present in a human teratocarcinoma cell line, NTera2D1, by primer extension and by analysis of cloned cDNAs. The bulk of the RNA begins (5′ end) at the residue previously identified as the 5′ terminus of the longest known primate genomic L1 elements, presumed to represent "unit" length. Several of the cDNA clones are close to 6 kilobase pairs, that is, close to full length. The partial sequences of 18 cDNA clones and full sequence of one (5,975 base pairs) indicate that many different genomic L1 elements contribute transcripts to the 6.5-kilobase cytoplasmic poly(A)$^+$ RNA in NTera2D1 cells because no 2 of the 19 cDNAs analyzed had identical sequences. The transcribed elements appear to represent a subset of the total genomic L1s, a subset that has a characteristic consensus sequence in the 3′ noncoding region and a high degree of sequence conservation throughout. Two open reading frames (ORFs) of 1,122 (ORF1) and 3,852 (ORF2) bases, flanked by about 800 and 200 bases of sequence at the 5′ and 3′ ends, respectively, can be identified in the cDNAs. Both ORFs are in the same frame, and they are separated by 33 bases bracketed by two conserved in-frame stop codons. ORF 2 is interrupted by at least one randomly positioned stop codon in the majority of the cDNAs. The data support proposals suggesting that the human L1 family includes one or more functional genes as well as an extraordinarily large number of pseudogenes whose ORFs are broken by stop codons. The cDNA structures suggest that both genes and pseudogenes are transcribed. At least one of the cDNAs (cD11), which was sequenced in its entirety, could, in principle, represent an mRNA for production of the ORF1 polypeptide. The similarity of mammalian L1s to several recently described invertebrate movable elements defines a new widely distributed class of elements which we term class II retrotransposons.

Line-1 (L1) is a family of long highly repeated DNA sequences dispersed in all mammalian genomes (9–12, 14, 22, 23, 55, 58, 65). In primates, the longest known family members are about 6 kilobase pairs (kbp), although many family members are truncated and internally rearranged (1, 22, 23, 31, 35, 39, 48). The structure of randomly selected genomic L1s from various primates is similar to that of processed pseudogenes, including the presence on one strand of long but broken open reading frames (ORFs), an A-rich 3′ terminus (on the strand containing the ORFs), the apparent lack of introns interrupting the ORFs, and variable-sized target site duplications (Fig. 1). The proteins predicted by the ORFs include regions with homology to reverse transcriptase and nucleic acid-binding proteins (19, 24, 36).

L1 elements in other mammals are similar to those in primates with regard to abundance and overall organization (3, 19, 20, 33, 36, 37, 50, 60, 63). Moreover, both the nucleotide sequence of the ORF region and the polypeptides predicted from the ORFs are homologous in all mammalian orders that have been analyzed. In contrast, the regions flanking the ORF region on both the 5′ and 3′ sides (called here the 5′-leader and 3′-trailer segments, respectively) are not conserved between orders.

These findings have led to two suggestions regarding L1 families: first, that in each species, one or more family members may be functional genes encoding one or more conserved polypeptides; and second, in analogy with processed pseudogenes, L1 units are amplified and dispersed by a mechanism that involves reverse transcription. Both of these suggestions imply that unit-length L1 transcripts should occur.

Nuclear RNA from various cell types and species contain sequences that anneal with L1 probes (11, 14, 29, 31, 33, 35, 52, 54, 56, 60, 67). The size of these RNA polymerase II transcripts (54) ranges from 0.5 to 14 kilobases (kb), and they contain sequences homologous to both strands of the L1 unit (56). In several human cell lines, the nuclear (or total) RNA contains more copies of the 3′ end of the L1 unit than of the 5′ end, as is true of the L1 elements in the genome (56, 60). Most nuclear transcripts are likely to represent L1 sequences that are included in unrelated primary transcripts, being neither polyadenylated nor transported to the cytoplasm in most of the cell types that have been screened (31, 52, 56).

The cytoplasm of the NTera2D1 line of human teratocarcinoma cells (2) contains an approximately 6.5-kb-long RNA selected by oligo(dT) and poly(U) (56). On Northern (RNA) blots, this RNA anneals with nonoverlapping probes that together encompass the entire 6 kbp of a unit-length human L1. Only one L1 strand is detected in this RNA, the strand containing the ORFs. The cytoplasmic 6.5-kb RNA is most abundant in NTera2D1 cells with an embryonal carcinoma phenotype, is barely detectable in sparsely growing cells, and is undetectable after the cells are induced to differentiate with retinoic acid. The properties of the RNA thus suggested that the NTera2D1 RNA might include functional mRNA. Recently, discrete 8-kb L1 transcripts were also detected in poly(A)$^+$ cytoplasmic RNA from murine lymphoid cells (17).

To test whether the properties of the NTera2D1 6.5-kb L1 RNA are consistent with it including functional mRNA or intermediates in transposition or both, we have now (i) carried out primer extension experiments to define the 5′ end of the RNA and (ii) characterized cDNA clones isolated from a library prepared with cytoplasmic poly(U)-selected RNA of NTera2D1 cells with the embryonal carcinoma phenotype. Preliminary accounts of some of these data have

* Corresponding author.
† Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

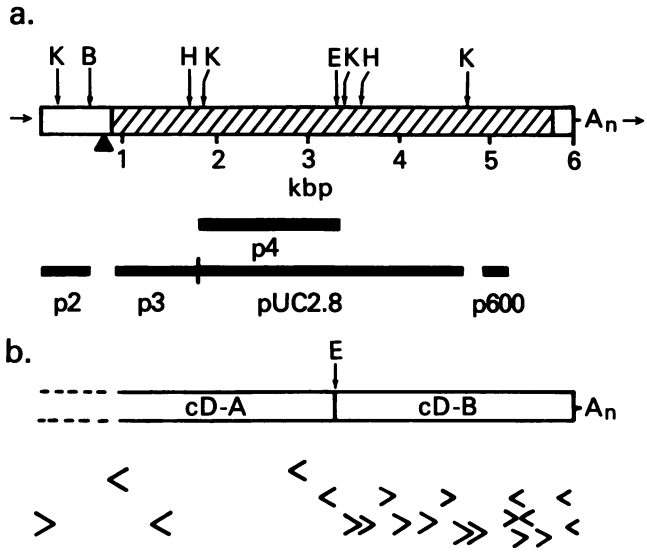


FIG. 1. Schematic diagrams of a (a) unit-length human genomic L1 unit and (b) showing the regions of the cDNAs subcloned in the A and B subclones. (a) The drawing is based on deduced consensus sequences (56; Sakaki, personal communication; Scott, personal communication). The slashed region contains extensive ORFs, but these are interrupted at different positions in various genomic cloned L1s. The slashed region also corresponds to the segment conserved among mammalian orders. A few typical restriction endonuclease sites are shown: E, *Eco*RI; K *Kpn*I; H, *Hin*dIII; B, *Bgl*II. Arrowhead indicates the position at which an additional 132 bp are inserted in about 50% of genomic L1 units (23). The A-rich region at the 3' end is marked A_n. Short arrows external to the two ends of the L1 represent the target site duplications that are found surrounding some L1s. The subcloned probes used in this work are shown below (see Materials and Methods). (b) Dashed lines indicate the variable lengths of the cDNAs at the 5' end. The positions and polarities of the oligonucleotides used to sequence the cD-Bs and small regions of some cD-As are indicated at the bottom. The sequence of cD11A was obtained as described in Materials and Methods.

been presented (57; M. F. Singer, J. Skowronski, T. G. Fanning, and S. Mongkolsuk, *in Eukaryotic Transposable Elements as Mutagenic Agents*, in press).

MATERIALS AND METHODS

All materials and methods were as previously described (56) unless indicated otherwise.

Preparation of RNA. NTera2D1 cells were grown to high density, at which time they begin to pile up and take on the embryonal carcinoma morphology. At various times during the further progression of the culture, cells were collected and RNA was prepared from the cytoplasmic fraction. The RNA was enriched for poly(A)$^+$ RNA by binding to and elution from poly(U)-Sepharose columns (Bethesda Research Laboratories, Gaithersburg, Md.). The several preparations were then analyzed by gel electrophoresis and blotting followed by hybridization with plasmid pUC2.8 (Fig. 1) which contains a cloned 2.8-kbp *Kpn*I fragment derived from the central portion of a typical primate L1 (34). As shown previously (56), the 6.5-kb transcript is visible against a faint smear of material that presumably represents contamination with nuclear RNA. This smear, however, can represent a substantial portion of the total RNA in the preparation. Therefore, the RNA preparation showing the highest ratio of the discrete 6.5-kb transcript to the total background hybridization was used to construct the cDNA library. In the sample used, about 50% of the hybridized radioactivity was contained in the discrete band.

Preparation of cDNA library. The duplex cDNA was synthesized as previously described (53). *Eco*RI linkers were attached to the double-stranded cDNA, which was then fractionated according to size by chromatography on 1-ml columns of Bio-Gel A-150m (27). The cDNA excluded from the column was ligated to *Eco*RI-digested λgt10 DNA (27) and packaged into phage particles in vitro (Gigapack; Vector Cloning Systems). The final library (λgt10.1) contained about 1.5×10^6 independent recombinants, and the mean size of the inserts was between 2 and 3 kbp. Quadruplicate plaque lifts were prepared from plates containing the complete library (4). Each of the four filters from each plate was then hybridized to one of four cloned probes representing nonoverlapping regions within a 6-kbp L1 unit (Fig. 1). Probes p2, p3, p4, and p600 have all been described previously (23, 35).

Subcloning. The cDNA inserts in λgt10 were cleaved with endonucleases *Eco*RI, and the resulting fragments were cloned into the pUC18 plasmid vector (43) by standard protocols.

Preparation of primers, probes, and markers for 5'-end analysis of RNA by primer extension. A restriction fragment derived from the 5' region of L1 (Fig. 1a) was used as a primer for mapping the 5' ends of L1 transcripts. Briefly, the 565-base-pair (bp)-long *Asp*718/*Bgl*II restriction fragment was isolated from cD11A (positions 65 to 630 in cD11 sequence in Fig. 4). (Note that *Asp*718 is an isoschizomer of *Kpn*I.) The ends were filled in with the large fragment of DNA polymerase I and [α-^{32}P]dATP (3,000 Ci/mmol) together with the remaining three deoxynucleoside triphosphates. The fragment was digested with *Bst*NI, and the 100-bp-long *Asp*718/*Bst*NI (3'-end-labeled) fragment (positions 65 to 163 in Fig. 4) was reisolated from an 8% polyacrylamide gel and used as a primer (see Fig. 2b) after its identity was confirmed by sequencing (38).

Plasmid p20/14AS, which was used for generating L1 transcripts to be used as a positive control, was constructed by subcloning (from cD14) the approximately 3.4-kb *Eco*RI fragment encompassing the 5' part of L1 (see Fig. 3) downstream of the T7 RNA polymerase promoter in the pIBI20 vector (International Biotechnologies, Inc., New Haven, Conn.). This orientation allows the synthesis of the L1 sense strand transcript by T7 RNA polymerase. T7 transcripts were generated by using as a template p20/14AS linearized with *Bam*HI, which cleaves the construct once within the polylinker 3' to the insert; conditions recommended by the supplier (International Biotechnologies, Inc.) were used.

An approximately 850-bp-long *Eco*RI fragment containing the junction between the 5' segment of a unit-length monkey genomic L1 sequence and the flanking low-copy-number sequence was isolated from pCaF2.5 (22) and used to provide markers for the primer extension experiments. This fragment was further cut with *Bst*NI. Its 5' ends were labeled with [γ-^{32}P]ATP and polynucleotide kinase. It was then cleaved with *Dra*I, and the *Bst*NI/*Dra*I subfragment containing 60 bp of the non-L1 sequence adjacent to 198 bp of the 5' end of L1 was isolated. These 198 bp correspond to positions 1 to 165 in the cD11 sequence extended at the 5' end by the 5'-most 32 bp of the longest genomic L1s. The structure of the *Bst*NI/*Dra*I fragment was confirmed by nucleotide sequence analysis (38).

Primer extension assays. Annealing of the primer was carried out as described before (44), using 1 to 10 μg of total, cytoplasmic, nuclear, or cytoplasmic poly(A)$^+$ RNA prepa-

rations and 5×10^4 cpm of labeled primer. Control T7 polymerase transcripts were annealed in the presence of 10 μg of tRNA as carrier. Extension reactions were carried out with Moloney murine leukemia virus reverse transcriptase in a final volume of 50 μl under the conditions recommended (Bethesda Research Laboratories), but with the omission of actinomycin D. Reactions were terminated and the products were processed and analyzed on 6% sequencing gels as described previously (44).

Sequence of cDNAs. A total of 19 L1 cDNAs were analyzed at least in part. All of these cDNA inserts were contained in λgt10 vectors and each had a single internal *Eco*RI site (Fig. 1b). Two additional *Eco*RI sites occur at the insert-vector junctions. The 5′ and 3′ portions of each cloned cDNA were designated A and B, respectively, and each was subcloned into the *Eco*RI site of pUC18 as indicated in Fig. 1b (e.g., cD11 was subcloned to yield cD11A and cD11B). Sequence data for the B subclones and selected regions of the A subclones were obtained by using the dideoxy sequencing method, denatured double-stranded plasmid DNA as template, and oligonucleotide primers (25). The sequencing reactions were primed either with commercially available primers (forward or reverse sequencing primers from P-L Biochemicals, Inc., Milwaukee, Wis.) which gave the sequences at the extremities of the subcloned segments or with oligonucleotides whose sequence was predicted from the previously compiled primate L1 sequence (55) or from already sequenced portions of the cDNAs. The custom synthesized oligonucleotides were kindly supplied by Michael Brownstein, National Institutes of Health. The oligonucleotides were 18-20 residues long and their locations and polarity are depicted on Figure 1b.

Sequence data in the 5′ regions of some cDNAs (e.g., cD11A) were obtained by using deletions constructed with *Exo*III and *Exo*VII. The 5′-proximal 3.4 kbp of cD11 (cD11A) was cloned into the phage vector M13mp18, and approximately 5 μg of replicative-form DNA was digested to completion with restriction enzymes *Bam*HI (insert proximal) and *Sph*I (insert distal). The DNA was then treated with *Exo*III and *Exo*VII as described previously (69). The DNA was treated with Klenow polymerase I and T4 DNA ligase, and samples of the ligated materials were transfected into *Escherichia coli* JM101. Clear plaques were selected and the size of the M13 recombinant DNAs was determined. The phage DNAs containing deletions of the appropriate sizes were used for DNA sequence determination by the dideoxy method (51).

RESULTS

Primer extension experiments on NTera2D1 cytoplasmic poly(A)$^+$ RNA. If L1 units are amplified and dispersed through the intermediary formation of RNA followed by reverse transcription, then the 5′ end of at least some of the RNA intermediates should coincide with the 5′ end of the longest genomic L1s. To define the 5′ end of the NTera2D1 cytoplasmic poly(A)$^+$ L1 RNA, we performed primer extension experiments (Fig. 2). A 100-bp-long *Asp*718/*Bst*NI L1 DNA fragment corresponding to residues 100 to 200 (Fig. 1a) and representing the complement of the RNA was labeled at the 3′ end and used to prime reverse transcriptase reactions on the RNA. A second fragment, encompassing the 5′ segment of a monkey unit-length genomic L1 sequence, was chosen to provide size markers after base specific chemical degradation. The 5′ terminus of the monkey L1 element was previously shown to be representative of primate L1 units (22). Both the primer and marker fragments have one end at the corresponding *Bst*NI sites located within L1 (Fig. 2b).

Comparison of the size marker ladder with the mobility of the major products of the primer extension assay, using poly(U)-selected NTera2D1 cytoplasmic RNA as a template, indicates that many if not most L1 transcripts in this RNA population are initiated within one residue from the position defined as the 5′ end of the primate L1s (residue 1 on Fig. 1a and 2b). The major band of the extension products shows a slight length heterogeneity when compared with the band obtained for the control L1 transcript generated with T7 phage RNA polymerase. This could reflect the complexity of the 6.5-kb L1 transcripts in NTera2D1 cells because the analysis of cDNAs indicates that many of them differ from one another by an occasional single-base-pair deletion or insertion (see below).

The results of similar assays performed on total NTera2D1 RNA show that L1 transcripts initiated at 5′ ends of unit-length L1s are enriched in the poly(U)-selected cytoplasmic RNA fraction (cf. Fig. 2c, lane 4, and Fig. 2a, lane R). The majority of extension products formed with total RNA map to sequences located downstream from the 5′ end of complete L1 units, although products that are longer are also detectable (though not visible in Fig. 2c, lane 4). Both of the latter classes may reflect either the presence of incorrectly initiated products or the turnover of nonpolyadenylated nuclear L1 transcripts. These products were even more abundant when nuclear L1 templates were compared with total RNA templates (Fig. 2c, lanes 6 and 7). Correctly initiated L1 transcripts were not detectable in total RNA isolated from 293 cells (not shown) or Hela cells (Fig. 2c, lane 5), while they were found in total RNA isolated from the JEG3 cell line (Fig. 2c, lane 3).

Preparation and screening of the cDNA library. Transcripts that anneal with L1 sequences are estimated to represent as much as 1% of the nuclear RNA in several primate cell lines, including NTera2D1, while the poly(U)-selected 6.5-kb-long cytoplasmic L1 transcripts in NTera2D1 cells are at least 20 times less abundant than the mRNA for β-actin. Thus, the 6.5-kb cytoplasmic RNA represents only a small proportion of the total L1 transcripts in the cells. Several precautions were therefore taken to maximize the representation in the library of cDNAs copied from the 6.5-kb transcript of interest. First, the RNA used for cDNA synthesis was isolated from the cytoplasm of cells in which the 6.5-kb RNA constituted not less than 50% of the cytoplasmic L1 RNA that could be detected with a probe from the central portion of a genomic L1 unit. Second, the conditions for cDNA synthesis were optimized for the synthesis of long cDNAs and, prior to cloning, the cDNAs were fractionated to enrich for molecules that were several kilobase pairs in length. Third, to ensure that each cDNA selected represented an independent recombinant, the original master plates of the cDNA library were screened rather than samples from a previously amplified library. This was particularly important because the sequence complexity of the 6.5-kb RNA was unknown, and therefore a representative sample of the cDNAs was expected to be informative. Fourth, the recombinant phage were screened on quadruplicate replicas of the master plates with four hybridization probes (p2, p3, pUC2.8, and p600) that were not overlapping and together spanned almost the entire 6 kbp of a typical L1 unit (Fig. 1). This procedure allowed the identification of cDNAs that include both the 5′ and 3′ ends of the L1 unit and permitted the elimination of clones that contained short or rearranged L1 sequences.

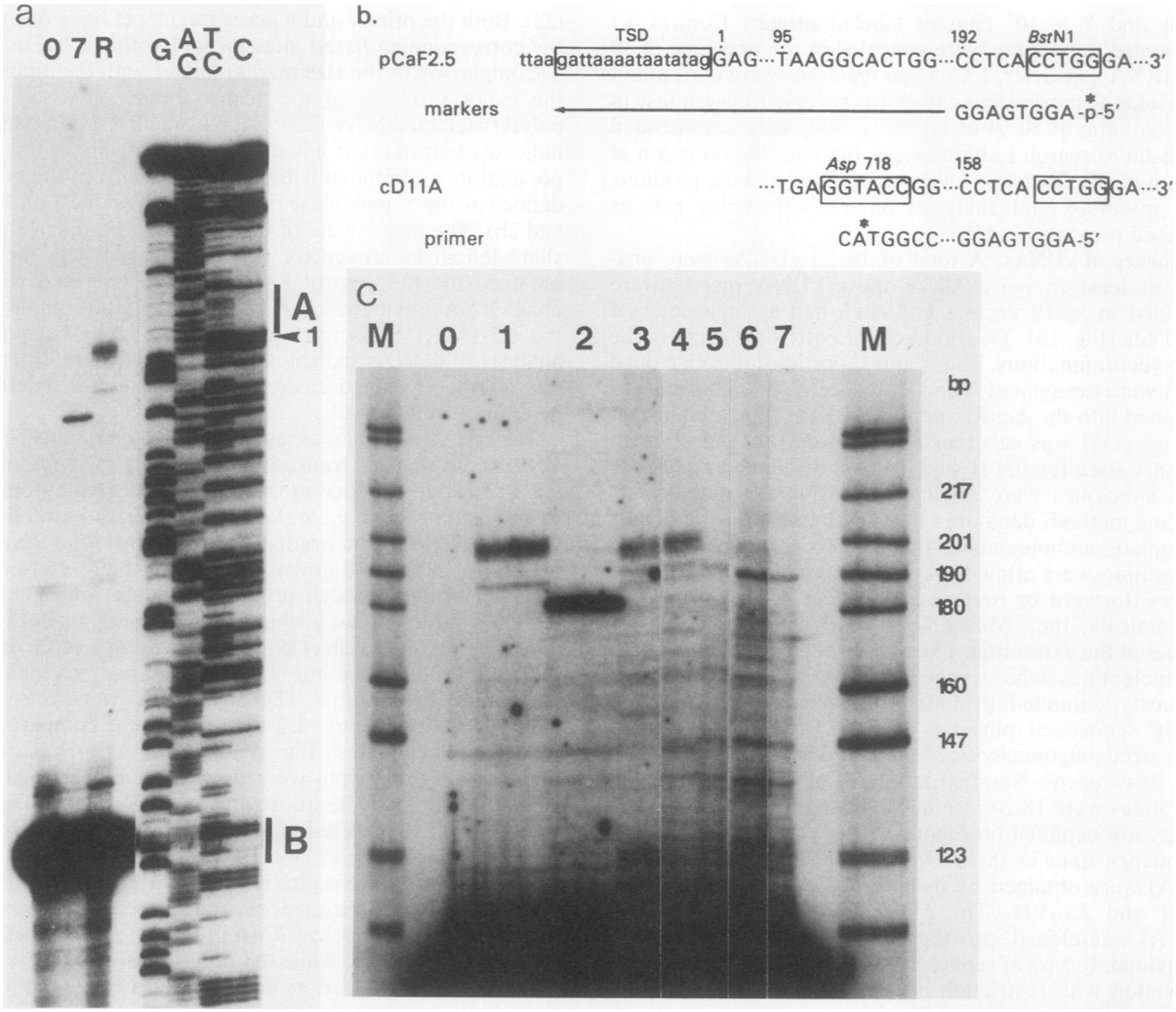


FIG. 2. Primer extension analysis of L1 transcripts. (a) Results of experiment designed to map 5′ ends of NTera2D1 cytoplasmic poly(A)$^+$ L1 transcripts. Reactions were primed with the *Asp*718/*Bst*NI restriction fragment isolated from cD11A plasmid. Templates were 10 μg of tRNA (lane 0), 10 ng of L1 transcript generated with T7 RNA polymerase from p20/14AS (as described in Materials and Methods) plus 10 μg of tRNA (lane 7), or 1 μg of cytoplasmic poly(U)-selected NTera2D1 RNA plus 10 μg of tRNA (lane R). The length of the extension product generated from the T7 polymerase L1 transcript is equal to that predicted from the structure of p20/14AS plasmid (not shown), indicating that no premature termination occurs on this template. The products of the base-specific chemical degradation of the marker fragment labeled at the 5′ end of the *Bst*NI site were separated on the same sequencing gel in lanes marked G, A/C, T/C, and C. The pCaF2.5 nucleotide sequence displayed in panel a is the complement of that presented in the top line of panel b, where the schematic alignment of the primers and markers used is shown in the context of the nucleotide sequences of the respective molecules. In panel a, the 5′-most nucleotide of the λF2 L1 element (from which pCaF2.5 is derived) that follows immediately after the 16-bp target site duplication is indicated by an arrowhead (residue 1) (22). In panel b, the site of radiolabel is indicated by an asterisk. The 5′ repeat of the 16-bp target site duplication (TSD) that flanks the L1 element in λF2, the consensus *Asp*718 (*Kpn*I) site present in cD11 but mutated in the L1 in λF2, and the *Bst*NI sites used for generation of the probes are indicated. (c) Results of the primer extension assay performed cn nuclear RNA templates isolated from various cell lines. Assays were performed as described above and in Materials and Methods. A 10-μg portion of total or 1 to 2 μg of nuclear RNA was used per assay. Lanes: M, pBR322 digested with *Msp*I to provide markers; 0, tRNA; 1, NTera2D1 cytoplasmic poly(A)$^+$ RNA; 2, 10 ng of L1 RNA made with T7 RNA polymerase as in lane 7 of panel a; 3, total RNA from JEG3 cells (56); 4, total RNA from NTera2D1; 5, total RNA from HeLa cells; 6 and 7, two independent preparations of nuclear RNA from NTera2D1 cells.

In the initial screening of the cDNA library, 15 phage annealed with all four of the probes and an additional 31 phage annealed with all but p2, which represents the most 5′ region of the L1 unit. A total of 153 other phage annealed with other combinations of the probes and were not characterized further. Of the 46 phage considered, 27 gave the same hybridization pattern with the L1 probes after plaque purification as they did originally, and these were analyzed further.

Characterization of the cDNA clones. The selected phage were digested with endonuclease *Eco*RI, and the products were analyzed by gel electrophoresis. Most of the inserts contained a single internal *Eco*RI site. Moreover, the majority of those with a single *Eco*RI site produced an approximately 2.5-kbp *Eco*RI fragment that annealed, in DNA-blotting experiments, to p600, the probe containing sequences from the 3′ end of the L1 unit. This is the result expected if the inserts extend to the 3′ end of an L1 unit and contain the *Eco*RI site typically found in genomic units (Fig. 1). The second fragment varied in length among the different phage and annealed with p4, which contains sequences derived from the 5′ portion of the long L1 unit. The total

length of the cDNA inserts varied from about 4 to about 6 kbp (57). The cDNA clones that did not contain a single internal *Eco*RI site were set aside, and a total of 20 cDNAs were analyzed further.

The *Eco*RI fragments produced by endonuclease digestion of the cDNAs were subcloned into the *Eco*RI site of pUC18. Subclones representing the 3′ and 5′ portions of the L1 unit were identified by annealing with appropriate probes. The subclones with the 3′ portions are designated with the letter B; thus, cD11B represents the subclone containing the 3′ portion of cD11. The subclones containing the 5′ portions are designated A, e.g., cD11A. All of the cD-B subclones were subjected to at least partial sequence analysis, as described in Materials and Methods. Although most of the sequence data represent single determinations on one strand, their reliability is good. This is because sequencing reactions with a particular oligonucleotide primer were carried out on multiple cDNAs simultaneously and the samples were run on single gels. The identity of the residues at most positions was evident, as were the changes from the consensus sequence. Overlapping sequences obtained with neighboring oligonucleotide primers as well as comparison with L1 genomic sequence (55; Y. Sakaki, personal communication) permitted the data to be assembled.

Figure 3 summarizes the structures of the cDNA inserts and the extent to which they were sequenced and points out several features. The sequence analyses show that all but two of the cDNAs arose from unique cloning events because they have unique junctions between the 3′ end of the L1 sequence and the vector. Two clones, cD14B and cD15B, were identical throughout, including the junction, and thus appear to represent independent isolations of the same clone; they are treated as a single clone, cD14. All of the cDNAs except cD28B contain only sequences previously shown to occur in typical genomic L1 units. In cD28, sequences that are not homologous to known L1 sequence follow the 3′ end of the L1 sequences (not shown on Fig. 3).

All of the cDNAs diverge from one another in one or more positions, indicating that they represent transcripts of different genomic L1s. This is apparent from the distribution of stop codons shown in Fig. 3 as well as the sequences at the 3′ ends (see below) and randomly distributed base changes not shown here (all sequence data are available upon request). Overall, the sequences between residues 3400 and 5800 diverge an average of 1.7% from a consensus of the cDNA sequences. This contrasts with an average divergence of about 7% for randomly selected genomic L1s compared with the genomic consensus (58).

Several of the cDNAs are nearly unit length, i.e., 6 kbp, as estimated from the size of *Eco*RI restriction fragments produced from the clones. All of the cDNAs have long ORFs. However, in the sequenced regions, these reading frames are interrupted by an occasional stop codon or single (or double)-base-pair deletion (Fig. 3). Some of the stop codons are randomly distributed, but the locations of several appear to be conserved. Thus, in every cDNA for which pertinent data were obtained, there is a TGA stop codon about 200 bp from the 3′ end (see below). This stop codon was used to define the beginning of the 204-bp-long 3′-trailer (noncoding) region. Two additional stop codons occur about 2 kbp from the 5′ end of the unit-length sequence in the three cDNAs (cD5, cD11, and cD14) for which pertinent data are available (Fig. 3). These same two stop codons are also conserved in human genomic L1 consensus sequences (Sakaki, personal communication; A. Scott et al., personal communication). The significance of these stop codons as

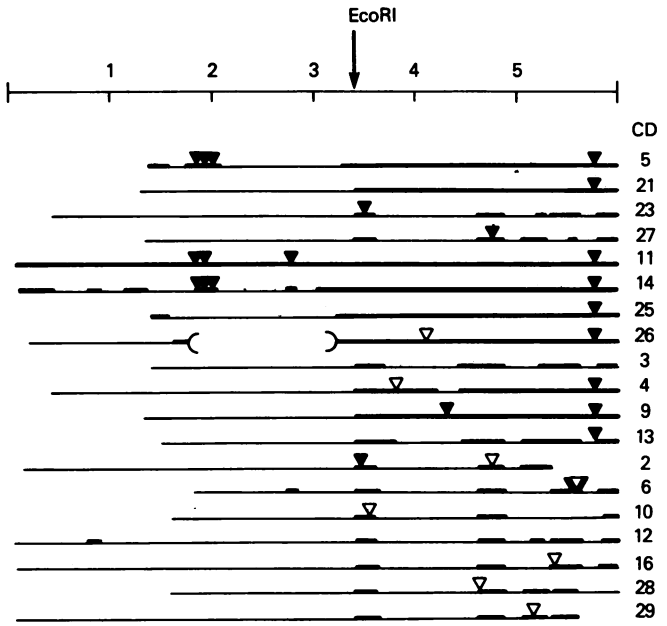


FIG. 3. Each horizontal line represents a different cDNA (CD), as numbered on the right. Thin lines show the length of the cDNA as estimated from restriction fragment lengths. Thick lines indicated the region of each cDNA that has been sequenced (sequence data are available upon request). Closed arrowheads mark the position of stop codons that are in the same frame as the long ORFs. Open arrowheads indicate base pair deletions that shift the frame so that a stop codon is encountered within a short distance. The parentheses in cD26 indicate a deletion of typical L1 sequences.

well as the other features of the cDNAs are most easily described in relation to the structure of cD11, the one cDNA sequenced in its entirety. One cDNA, cD26, is missing about 1.5 kbp compared with the others (Fig. 3). The sequence at the deletion site was not consistent with a standard intron splice junction.

Sequence of cD11. The sequences of the inserts of cD11A (3,395 bp) and cD11B (2,580 bp) were determined in their entirety. The two are joined to give the complete sequence of cD11 (5,975 bp) in Fig. 4. A comparison of the sequence of the 5′ end of cD11 with the primate 5′ consensus sequence indicates that the cDNA clone falls short of being a complete copy of a poly(A)$^+$ cytoplasmic NTera2D1 L1 RNA by 32 bp.

Altogether, we distinguish five regions in cD11, each of which will be discussed in turn: (i) a 5′ leader from residue 1 to the beginning of a long ORF (ORF1) at residue 768, (ii) ORF1 (residues 769 to 1889), (iii) the 39 bp (residues 1890 to 1928) between ORF1 and a second long ORF, (iv) ORF2 (residues 1929 to 5781); and (v) a 3′ trailer (residue 5781 to end). Several of the cDNAs (though not cD11) extend through the A-rich 3′-end; this 6th region of the L1 cDNAs will be described separately.

5′ Leader. Considering the strand that contains ORF1 and ORF2, there are no long ORFs in the 767 bp starting at residue 1 of cD11 (Fig. 4). Moreover, there is only one ATG codon (residues 575 to 577) in all three possible frames, and the coding region initiated by this ATG is stopped after seven codons. Notably, ATG codons occur at a normal frequency in all frames 3′ to the start of ORF1 and also occur in all three frames on the opposite strand in the 5′-leader region. Selective forces may have operated to suppress start codons in the 5′ leader during the evolutionary history of the DNA sequence encoding cD11. The other cDNAs are likely

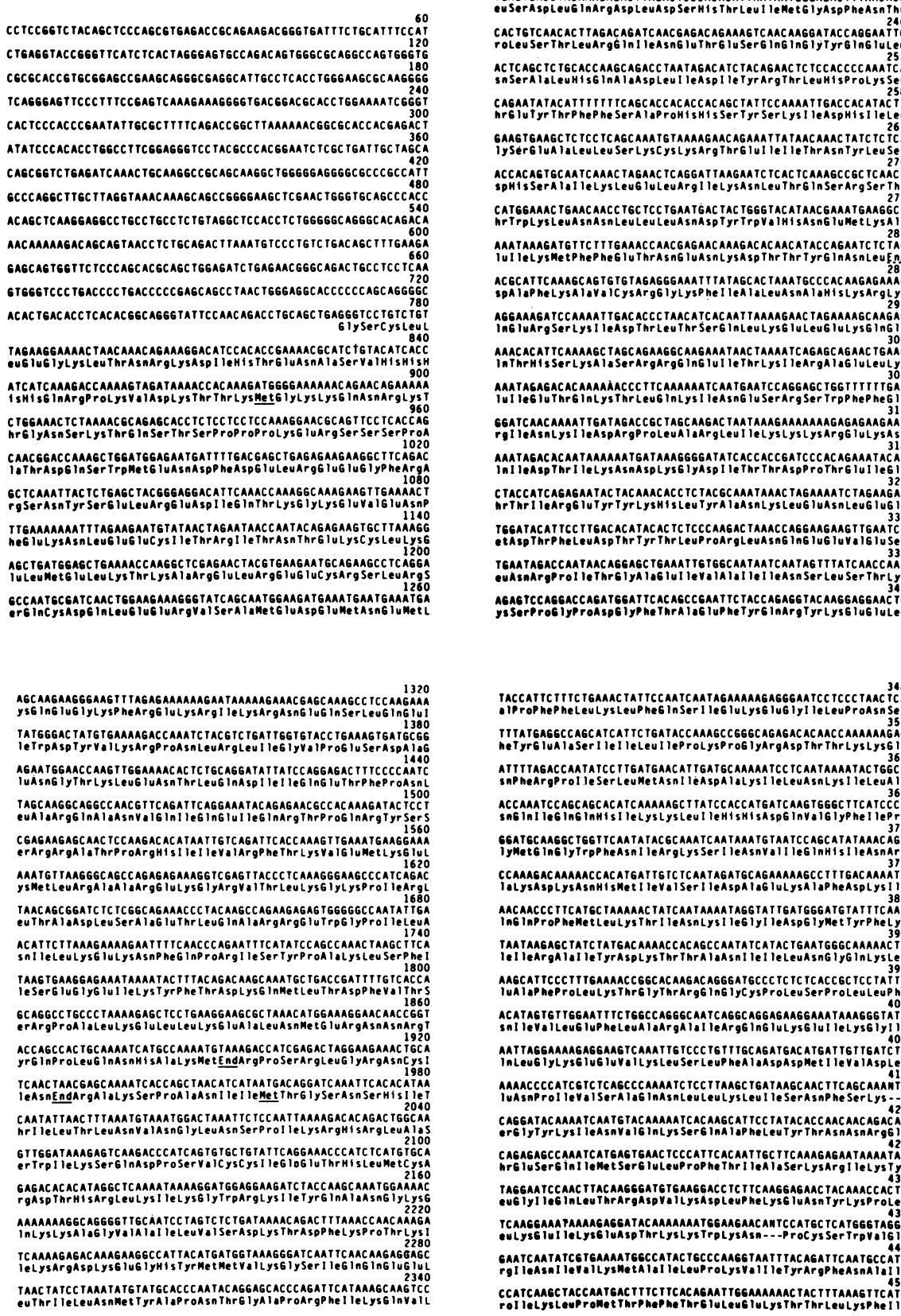
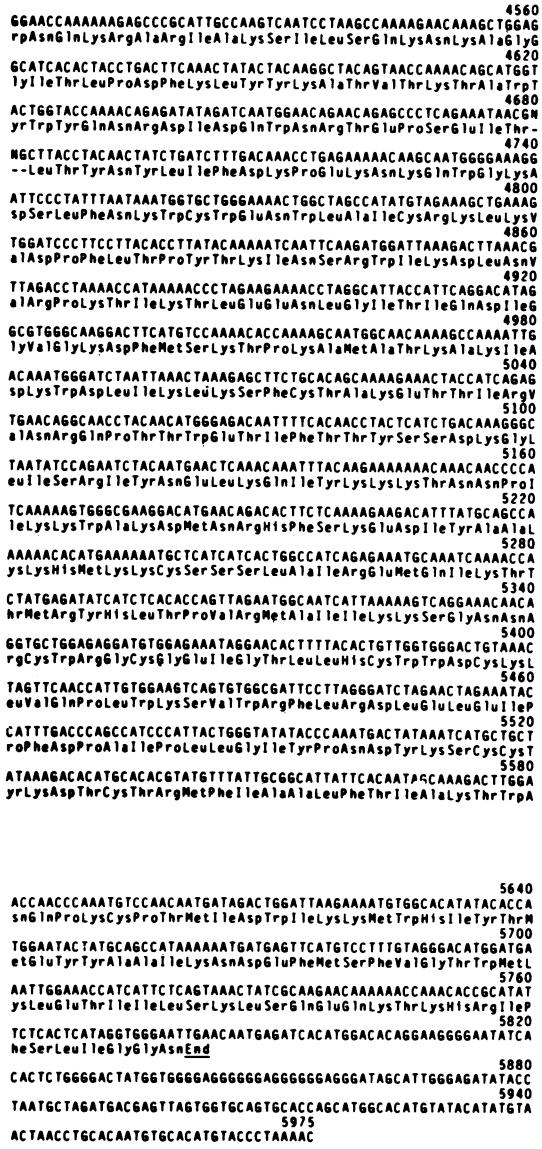


FIG. 4. Nucleotide sequence of cD11 and the predicted amino acid sequences of ORF1 and ORF2. N represents bases not identified; the amino acid is omitted where the codon contains N. The first methionines in ORF1 and ORF2 are underlined, as are the conserved stop codons (end). The nonconserved stop codon at residue 2817 is indicated by broken underlining. The displayed sequence requires three corrections. (i) Residues 28 and 29 (AC) should be replaced by three residues, CGA. (ii) The C at position 2412 should be deleted. (iii) A C should be added after residue 2450. The total number of residues, after correction, is 5976.



to have similar 5′ leaders because the cD11 sequence matches well the 370 bp of cD14 that were determined in this region and is similar to the genomic L1 consensus.

cD11 has a skewed distribution of CpG dinucleotides. There are 29 CpGs within the first 400 bp of cD11 and only 57 more in the remaining 5,575 bp. A similar clustering of CpGs is found near the 5′ end of the mouse L1Md-A2 clone (36) and near the 5′ end of some mammalian genes (6). As in the latter case, the CpGs may play a role in regulating L1 activity, perhaps correlated with the degree of cytosine methylation in the 5′ leader (16).

Hattori et al. (23) reported that about one-half of all unit-length human genomic L1s contain an insert of 132 bp in the 5′ leader. The sequence of cD11 shows that it does not contain this 132-bp insertion which would have been located between residues 750 and 751. Dot blots of cD2A, cD12A, cD14A, cD16A, and cD26A, all of which anneal to probe p2, did not hybridize to a probe containing the 132 bp (kindly supplied by A. Scott).

ORF1 of cD11. ORF1 of cD11 consists of 1,122 bases (residues 769 to 1899). The first ATG in ORF1 is at residue 876. From this ATG to the TAA stop codon at residues 1890 to 1892, ORF1 can encode a protein of 338 amino acids and approximately 40 kilodaltons (Table 1). The ATG at 876 occurs also in cD12A and cD14A, the only other cDNAs that were sequenced in this region, and it is the first ATG in this frame in the consensus sequence of human genomic L1s. The amino acid composition of the protein predicted by ORF1 suggests that it is somewhat basic (66 Arg + Lys and 55 Asp + Glu).

It is of interest to compare ORF1 with the corresponding ORF of the mouse L1, L1Md-A2 (36), although L1Md-A2 is a genomic clone and its ability to be transcribed is unknown. L1Md-A2 contains two ORFs that are related to those in cD11. A comparison of the mouse ORF1 (1,100 bp) and cD11 ORF1 (Table 1) indicates that (i) the nucleotide homology is about 53%, suggesting a lack of strong selection to conserve the coding sequence, and (ii) the two predicted proteins share only 35% homology at the amino acid level overall, although the homology is higher (53%) in the carboxy-terminal region of the protein than in the NH_2 terminal (not shown). A search of the protein data bank (March 1987) for homology between the cD11 ORF1 and known proteins revealed nothing of obvious significance.

Inter-ORF, the region between ORF1 and ORF2. ORF1 and ORF2 of cD11 are separated by 39 bp, including the TAA codon that ends ORF1 (residues 1890 to 1928 in cD11) (Fig. 4). This inter-ORF region includes one additional in-frame stop codon, 33 bp downstream of the first. Both stop codons are conserved in the two additional cDNAs sequenced in this region (cD5A and cD14A) as well as in the human genomic L1 consensus (Fig. 5).

ORF2 of cD11. The reading frame from residue 1929 to 5781 is in frame with ORF1. It is broken by a single in-frame TAG stop codon at position 2817. Two additional cDNA clones, cD14 and cD16, were sequenced through this region, and both were found to code for tryptophan (TGG) at this position. Thus, the TAG in cD11 is not a consensus stop signal and we refer to the entire region as ORF2 (3,852 bp).

The first ATG in ORF2 occurs at residue 1956. This ATG is conserved in cD5A and cD14A, the only other cDNAs sequenced in this region (Fig. 5). It is also conserved in the genomic L1 consensus. Upstream of the ATG, within ORF2, cD5A and cD14A each contain an additional in-frame stop codon, although the position of this extra stop is different in the two clones.

The 3,825 bp of ORF2 starting at the ATG and extending to the conserved stop codon at residue 5781 could encode a protein of 1,275 amino acids and about 150 kilodaltons, except for the nonconsensus stop codon at residue 2817 (Table 1). The amino acid composition of the predicted protein suggests that it is fairly basic (209 Arg + Lys and 135 Asp + Glu).

Overall, the human and mouse (L1Md-A2) ORF2 share 67% homology at the nucleotide level and 60% at the amino acid level (Table 1). This suggests that ORF2 in mammals is under more stringent selective pressure than ORF1. Several regions of the ORF2 proteins are conserved over 85% in both the human and the mouse (as well as in carnivore and lagomorph). Some of these exhibit homology to retroviral reverse transcriptase; they are found between residue 3504 and 4262 on the cD11 sequence (Fig. 4) and are labeled A through G by Fanning and Singer (19). Near the carboxy terminus of the ORF2 protein is another highly conserved region (labeled I by Fanning and Singer [19], residues 5342 to 5438 in cD11) that is homologous with the "Cys" motif of retroviral nucleic acid-binding proteins. In addition, a highly

conserved region (labeled H by Fanning and Singer [19], residues 5225 to 5315 in cD11) of the ORF2 protein has homology to transferrin (24).

3′-Trailer region of the cDNAs. The 3′-trailer sequences of several of the cDNAs are reported in Fig. 6. Sequence data were not obtained for cD28B and neither cD2B nor cD29B extended into this region. Most of the cDNAs that extend through the trailer have 204 bp in this segment. A consensus sequence for the 204 bp was deduced, and the percent divergence of each cDNA from the cDNA consensus was calculated (Fig. 6). The highest divergence is 11% (cD10B) but the majority of the cDNAs diverge <5% from the consensus. Figure 6 also shows a consensus sequence for the 3′-trailer region (205 bp) calculated from the sequence of 20 randomly selected human genomic L1s (57). The genomic L1s diverge much more from their own consensus (average of 13%) than do the cDNAs from the cDNA consensus. Moreover, the two consensus sequences differ from one another at 17 positions. (Note that a value of 16 is mentioned by Skowronski and Singer [57]. The correction results from recompilation of the consensus sequence.) Only two of the cDNAs, cD6B and cD10B, have sequences in the 3′-trailer region that match the genomic consensus better than they do the cDNA consensus. We conclude from these observations that a distinct subset of genomic L1s, subset T, is the main contributor to the 6.5-kb cytoplasmic RNA in the NTera2D1 cells. Four of the cDNAs, cD5B, cD21B, cD23B, and cD27B, differ from the cDNA consensus in the same four residues and thus represent a subset (subset Ta) of the cDNAs. These four cDNAs diverge from the subset Ta consensus only about 0.6%.

3′-Terminal A-rich region. Figure 7 shows the A-rich ends of those cDNAs that extend beyond the 204 bp of the 3′ trailer. The first base shown in Fig. 7 follows directly after the last in Fig. 6. The last base shown abuts the *Eco*RI site of the vector. Each of the cDNAs has a different sequence in this region. Of the seven sequences, six have at least one polyadenylation signal, AATAAA, the positions of which are variable. No other AATAAA sequences occur upstream of the A-rich region in the 3′ trailer.

The 6.5-kbp cytoplasmic transcripts used to generate the cDNA library were stably bound to poly(U)-Sepharose in 10% formamide (56) and thus are likely to have poly(A) tails (66). The tails do not appear in the cDNAs, presumably because of the cloning procedures. Two other cDNA clones for non-L1 transcripts that have been isolated from the same λgt10 library also lack long poly(A) tails (47; S. Detera-Wadleigh, personal communication).

TABLE 1. Properties of ORFs in cD11[a]

Property	ORF1	ORF2
Length		
Nucleotides	1,014	3,825
Codons	338	1,275
Polypeptide (kilodaltons)	39.8	149.5
% Homology to mouse L1		
Nucleotides	53	67
Amino acids	35	60
Amino acid composition[b]		
A	16	61
R	34	61
N	20	74
D	12	60
C	4	19
Q	21	58
E	43	75
G	11	46
H	2	27
I	18	129
L	29	115
K	32	148
M	11	31
F	10	51
P	15	48
S	21	71
T	21	94
W	3	29
Y	6	42
V	9	32

[a] In this table ORF1 and ORF2 are measured from the first consensus ATG codon to the first consensus stop codon.
[b] There were three unspecified amino acids in ORF2.

DISCUSSION

The initial characterization of the discrete cytoplasmic L1 transcript in NTera2D1 cells with the embryonal carcinoma phenotype suggested that the RNA is derived from unit-length L1 sequences. This finding was consistent with current speculations concerning (i) the potential of some L1 units to encode proteins and (ii) the mechanism of amplification and dispersal of L1s. The results reported here confirm and extend this conclusion. The primer extension experiments show that the 5′ end of the bulk of the transcripts corresponds to the nucleotide residue previously identified as the 5′ end of the longest genomic L1 sequences. The properties of the characterized cDNAs suggest that most of them represent the discrete 6.5-kb cytoplasmic poly(A)$^+$ RNA, which represented 50% of the L1 RNA in the sample used for cDNA synthesis. Only one, cD28, had non-L1 sequences joined to L1 sequence. Another, cD26, contains a deletion. About half of the cDNAs analyzed start within a few hundred base pairs of the 5′ end of the cytoplasmic RNA and at least two, cD11 and cD14, start within 32 bases of the 5′ end of the RNA. The data indicate that the bulk of the NTera2D1 6.5-kb cytoplasmic RNAs represent transcripts of entire L1 units and not transcripts of

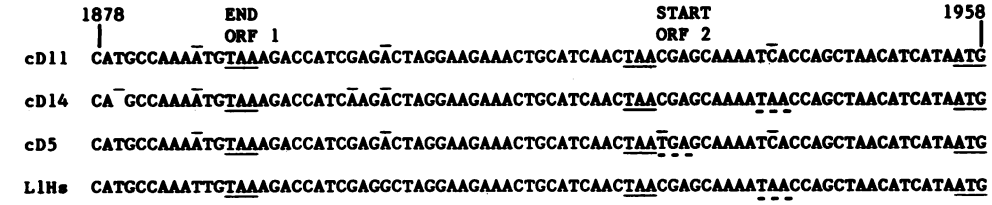


FIG. 5. Nucleotide sequences of cD5A, cD11A, cD14A, and the human genomic L1 consensus sequence (L1Hs) in the region of the stop codons separating ORF1 and ORF2. The strand with the ORF is shown. The conserved stop codons are underlined. Other in-frame stop codons in cD14, cD5, and L1Hs are indicated by broken underlining. Overlines indicate differences between the cDNAs and the genomic consensus sequence.

L1 units fortuitously inserted in unrelated transcription units.

We established the complete nucleotide sequence of one of the cDNAs, cD11. Besides the 5′ most 32 bp, this 5,976-bp cDNA lacks 30 bp of the 204-bp 3′-trailer sequence that precedes the A-rich region of genomic L1s and the other cDNAs. We also obtained partial sequences of 18 other cDNAs. No two of the cDNAs are identical, although they are very similar. All sequence data are consistent with there being the potential for two major ORFs in some L1 units. However, we do not know if both frames are completely open in any one of the cDNAs. The structure of cD11, an open ORF1 and a closed ORF2, is consistent with it representing an mRNA that could be translated to yield the ORF1 protein but not the ORF2 protein. If ORF2 is completely open in one or more of the RNAs, its ability to be translated involves interesting problems. The two L1 ORFs are in the same frame but are separated by a short region that includes two conserved in-frame stop codons. Two of the cDNAs are known to have a third in-frame stop codon in this region. In many mouse genomic L1s, including the full-length L1Md-A2, the two ORFs that are homologous to the human L1 ORFs are in different frames and overlap by five amino acids (36, 41). Similarly, retroviruses and retrotransposons in diverse organisms have ORFs that are in different overlapping frames or in the same frame but separated by an in-frame stop codon. The prototype of the first group in Rous sarcoma virus, in which *gag* and *pol* ORFs are in two frames and overlap by 19 amino acids (28). An example of the second group is Moloney murine leukemia virus, in which a single in-frame stop codon separates the *gag* and *pol* genes (70). Thus, the human L1 cDNAs have a unique arrangement more like that of Moloney murine leukemia than Rous sarcoma virus, while the mouse L1 has a Rous sarcoma virus type of arrangement.

Translation of Moloney murine leukemia virus RNA to yield the *gag-pol* fusion protein involves suppression of the dividing stop codon (70). A fusion protein might similarly be translated from ORF1 to ORF2 of a human L1, but this would required suppression of two or even three stop codons. Such an event might be restricted to a particular tissue at a particular developmental time. Additional models include the independent initiation of translation at an internal AUG in ORF2 possibly by the relaxed scanning model or by direct binding of a ribosome or reinitiation in ORF2 after termination at the end of ORF1 without prior dissociation of ribosomes (32). Recent experiments demonstrate that such reinitiations occur if the translation of an upstream ORF terminates within a short distance of the start codon of the downstream ORF (45, 46). An alternative possibility is splicing of a primary transcript to remove the inter-ORF region.

Our observation that the poly(A)$^+$ cytoplasmic 6.5-kb NTera2D1 RNA is a heterogeneous mixture of a substantial number of different transcripts indicates that a sizable percentage of the approximately 3×10^3 to 4×10^3 unit-length genomic L1s may be templates for the RNA. Moreover, these particular L1s (subset T) are more homogeneous in sequence than are the bulk of genomic L1s, many of which are also truncated. Subset T has a distinctive consensus sequence in the 3′-trailer region that distinguishes its members from most randomly selected genomic L1s. We term those units that have the 3′-trailer consensus of randomly selected genomic L1s are subset U. To account for the discrete cytoplasmic NTera2D1 RNAs, we propose that subset T, but not subset U, is associated with specific sequences that regulate transcription, the processing of primary transcripts, or transport to the cytoplasm. The members of subset T are notable for the high degree of sequence conservation. This could reflect a recent origin of the transcribed subset from a single "founder" sequence or a higher rate of homogenization than in the majority of genomic L1s or both.

The observation that the 5′ end of the NTera2D1 RNA coincides with the 5′ end of unit-length genomic L1s is important when considering models for transcriptional regulation of L1 and, in particular, the possible location of *cis* regulatory elements. One possibility is that members of subset T are all associated with common upstream regulatory sequences. For example, the subset T L1s might be clustered in a tandem array in a transcriptionally active genomic region much like the tandemly amplified loci containing functional U1 (5), dihydrofolate reductase, or CAD genes (59). The amplified unit could have included the L1 element as well as upstream promoters and other regulatory signals, if the amplification occurred at the DNA level. All of the units might then be kept relatively homogeneous by recombinational processes. However, the varying structures of the A-rich regions at the 3′ termini of the cDNAs are not consistent with this model.

Another explanation, which allows for subset T to be dispersed in the genome, is that the promoter and other regulatory sequences are contained within the unit-length L1 elements themselves, as proposed for the L1-like I element in *Drosophila melanogaster* (21). This model implies that subset T could include the products of recent rounds of transposition or gene conversion or both in which the template sequence was a functional L1 element. An interesting version of this second model has been proposed for mouse L1s (36).

Of the two subsets, A and F (36, 41), of mouse genomic L1 elements that have been described, the A family, which has tandemly repeated units of 208 bp beginning at the 5′-terminus, is the basis for the model. Accordingly, L1 units that possess at least one copy of the 5′-repeat unit, including reverse transcribed and reintegrated L1 units, will be capable of transcriptional activity because the repeats contain promoter elements that direct downstream transcription initiation. Neither the L1 cDNA clones, described here, nor primate L1 consensus sequences (Sakaki, personal communication; Scott et al., personal communications) have such repeat structures. A rat L1 (L1Rn) contains 65-bp repeat sequences in its 5′ portion, but these are about 1 kbp from the 5′ end (11). Thus, the model proposed for the control of L1Md transcription may not be applicable to primates and possibly not even to other rodents, at least based on current data.

Although the cDNAs are well conserved in the sequenced portions of the ORFs, and in the 3′ trailer, their A-rich regions are all different from one another and from the similarly unique A-rich regions of random genomic L1s. (The set of genomic L1 sequences analyzed are those described in Fig. 6.) Similarly, a truncated cDNA that was isolated from a library constructed with poly(A)$^+$ cytoplasmic RNA from human primary fibroblasts and ends at the usual L1 3′ end also has a unique A-rich structure between the AATAAA signal and a long (56-base) poly(A) tail (14). Only two features of the A-rich region are conserved in subsets T and U: the preponderance of A residues and a polyadenylation signal, AATAAA (although the position of the AATAAA unit varies). Although it has not been demonstrated that the NTera2D1 6.5-kb RNA is an RNA polymer-

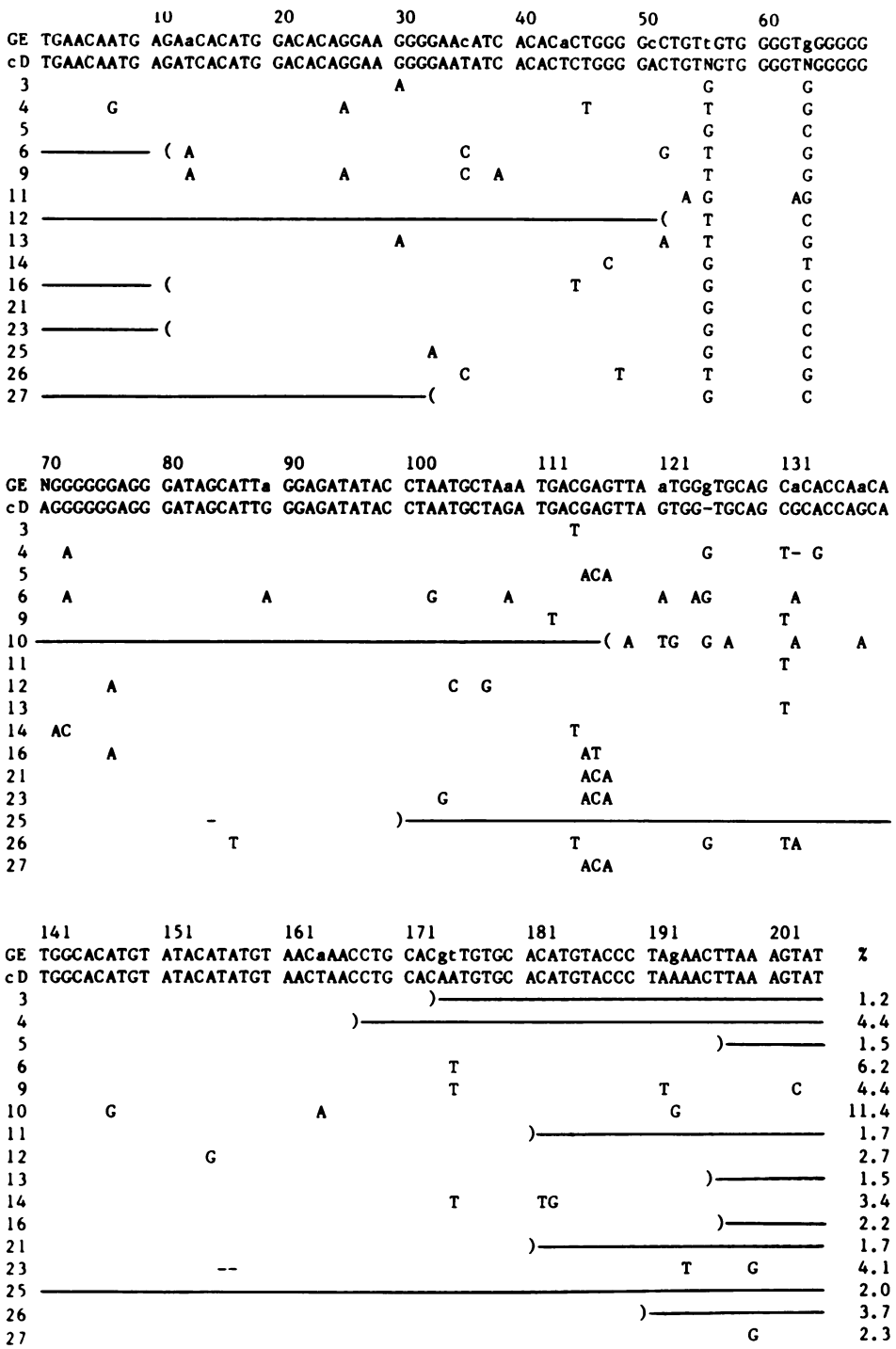


FIG. 6. cDNA, cDNA consensus (cD), and genomic consensus (GE) sequences in the 3'-trailer region of human L1 units. The top line is the consensus sequence derived from 20 randomly selected human genomic L1s. The identities and references for all of these sequences have been reported (57). The second line shows the consensus sequence deduced from the cDNAs described in this work; those cDNA sequences are shown below. Only those residues that differ from the cDNA consensus are given for the individual cDNA clones. Lowercase letters in the genomic consensus indicate divergence from the cDNA consensus. An N indicates no consensus base. A dash stands for a missing base. A line indicates that no sequence is available because either it was not determined or the cDNA clone lacks the region. Parentheses mark the beginning and end of the determined sequences. The numbers at the right of the last lines are the percent that each cDNA diverges from the cDNA consensus; residues for which there is no consensus (N) were ignored in the calculation.

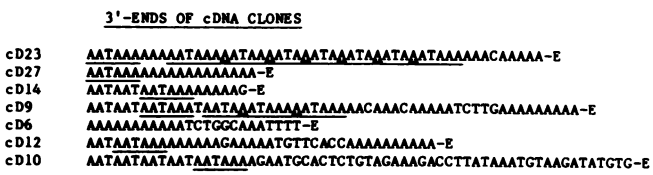


FIG. 7. A-rich region. Polyadenylation signals, AATAAA, are underlined; where several overlap, the initial A is underlined separately. The ends of the cDNA clones are shown such that the first residue follows immediately after the last residue in Fig. 2. E indicates the *Eco*RI site that joins the cDNA to the vector.

ase II product, the structure of the 3′ ends of most of the cDNAs is consistent with the possibility that subset T is transcribed by RNA polymerase II (7) from genomic units that also have heterogeneous A-rich termini.

The irregular and diverse A-rich segments in genomic L1s were pointed out earlier (49) and compared with the similar A-rich segments at the ends of retroposons such as *Alu* sequences. In many of these elements, the A-rich region is made up of short stretches of tandem repeats of the type $T(A)_n$ or $C(A)_n$. It was suggested that such sequences may arise during the insertion of the elements into new genomic locations, as a consequence of the transposition mechanism and preferred (AT-rich) target sites. If this suggestion is correct, then our observations indicate that the 3′ ends of both subset T and subset U L1s may have been generated by such processes. Note that cD9B and cD23B have the $T(A)_n$ motif (Fig. 7). However, it is also possible that the sharp boundary between conservation and divergence that occurs at the junction of the 3′-trailer and A-rich regions marks a restriction on a homogenization process such as gene conversion. Recurring borders for the extent of gene conversion have been noted in other systems, for example, the dispersed 5S rRNA genes of *Neurospora crassa* (40).

Current models suggest that most L1s are processed pseudogenes. Typically, processed pseudogenes do not have A-rich regions between the polyadenylation signal and the start of the poly(A) tail. Rather, they are homologous to their parental genes in this region. However, several families of processed pseudogenes such as those for human argininosuccinate synthetase (42), mouse λ-chain (26), and mouse ribosomal protein L30 (64) do diverge markedly from the corresponding genes and cDNAs and from one another in the region preceding the poly(A) tail. Some of these (L30 and mouse λ-chain) also display tandem repeats of $C(A)_n$ just prior to the 3′ copy of the target site duplication. Thus, the structures of L1s are, as far as is known, consistent with the suggestion that most genomic units including subset T are processed genes.

Several lines of evidence are consistent with earlier proposals suggesting that, in addition to the large number of pseudogenes, the L1 family contains one or more functional genes. Perhaps the most compelling is the conservation of the two ORFs in the L1s of various mammals. Overall, the ORF1 coding region is less well conserved between humans and mice than is the ORF2 coding region, although 80 amino acids (23% of the protein) near the carboxy terminus of ORF1 are 53% identical. It is possible that this 80-amino acid region represents a functional domain of the ORF1 protein that is much less tolerant of change than the rest of the protein. The proteins predicted by human and mouse ORF2s are 60% identical, overall. In some regions the homology is >80%, and these regions are also well conserved in lagomorphs and carnivores. The significance of these striking homologies is underscored by their similarity to polypeptide regions conserved in revere transcriptases and nucleic acid-binding proteins (19, 24, 36). These findings support the suggestion that some L1s encode proteins that can foster L1 amplification and dispersal through the intermediary formation of L1 RNA and, thus, the possibility that at least some L1s are transposable elements.

L1 elements are organizationally similar to the F (15; P. P. DiNocera and G. Casari, Proc. Natl. Acad. Sci. USA, in press), G (15), and I (21) transposable elements of *D. melanogaster*, the ingi element of *Trypanosoma brucei* (30), and the R2 element of *Bombyx mori* (8). These elements lack long terminal repeats, have A-rich 3′ termini, and contain long ORFs that encode polypeptides homologous to that encoded by ORF2 of cD11, including the regions of reverse transcriptase homology. The ORF2 homology between all of these elements and L1 is significantly greater than between any of the five elements and retroviral reverse transcriptases. Current data indicate that F, I, G, and ingi, like L1, contain several nonoverlapping ORFs; R2 has a single ORF. Together these elements constitute a recently recognized class of movable elements which we term class II retrotransposons to distinguish them from retrotransposons that have long terminal repeats (e.g., Ty, copia, and IAP). Although not proven, it is likely that class II is amplified and transposed through reverse transcription of an RNA intermediate. Thus, the L1 RNAs described in this report could include intermediates in transposition. Appropriate RNAs have yet to be identified in *Drosophila, Bombyx,* or *Trypanosoma* spp. These considerations indicate that at least four classes of elements that may encode and depend on reverse transcription for amplification are found in many organisms: retroviruses (61), class I retrotransposons (containing long terminal repeats) (61), class II retrotransposons (no long terminal repeats), and hepadnaviruses (18, 62).

ACKNOWLEDGMENTS

We are very grateful to Michael Brownstein for synthesizing the oligonucleotides. Our colleagues Ronald E. Thayer, Gisela Heidecker, and Roberto di Lauro provided help and advice. Y. Sakaki and Alan Scott kindly made available consensus sequences for human genomic L1 (L1Hs). P. P. DiNocera provided data prior to publication. The manuscript was ably prepared by Gail Gray.

LITERATURE CITED

1. **Adams, J. W., R. E. Kaufman, P. J. Kretschmer, M. Harrison, and A. W. Nienhuis.** 1980. A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. Nucleic Acids Res. **8**:6113–6128.
2. **Andrews, P. W., I. Damjanov, D. Simon, G. Banting, N. C. Dracopoli, and J. Fogh.** 1984. Pluripotent embryonal carcinoma clones derived from the human teratocarcinoma cell line Tera-2. Lab. Invest. **50**:147–162.
3. **Bennett, K. L., and N. D. Hastie.** 1984. Looking for relationships between the most repeated dispersed DNA sequences in the mouse: small R elements are found associated consistently with long MIF repeats. EMBO J. **3**:467–472.
4. **Benton, W. D., and R. W. Davis.** 1977. Screening λgt recombinant clones by hybridization to single plaques in situ. Science **196**:180–182.
5. **Bernstein, L. B., T. Manser, and A. M. Weiner.** 1985. Human U1 small nuclear RNA genes: extensive conservation of flanking sequences suggests cycles of gene amplification and transposition. Mol. Cell. Biol. **5**:2159–2171.
6. **Bird, A. P.** 1986. CpG-rich islands and the function of DNA methylation. Nature (London) **321**:209–213.

7. **Birnstiel, M. L., M. Busslinger, and K. Strub.** 1985. Transcription termination and 3′ processing: the end is in site! Cell **41:** 349–359.
8. **Burke, W. D., C. C. Calalang, and T. H. Eickbush.** 1987. The site-specific ribosomal insertion element type II of *Bombyx mori* (RsBm) contains the coding sequence for a reverse transcriptase-like enzyme. Mol. Cell. Biol. **7:**2221–2230.
9. **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.
10. **Burton, F. H., D. D. Loeb, C. F. Voliva, S. L. Martin, M. H. Edgell, and C. A. Hutchison III.** 1986. Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. J. Mol. Biol. **187:**291–304.
11. **D'Ambrosio, E., S. D. Waitzkin, R. R. Witney, A. Salemme, and A. V. Furano.** 1986. Structure of the highly repeated long interspersed DNA family (LINE or L1Rn) of the rat. Mol. Cell. Biol. **6:**411–424.
12. **Deininger, P. L., and G. R. Daniels.** 1986. The recent evolution of mammalian repetitive DNA elements. Trends Genet. **2:**76–80.
13. **Demers, G. W., K. Brech, and R. C. Hardison.** 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.
14. **DiGiovanni, L., S. R. Haynes, R. Misra, and W. R. Jelinek.** 1983. *Kpn*I family of long-dispersed repeated DNA sequences of man: evidence for entry into genomic DNA of DNA copies of polyA-terminated *Kpn*I RNAs. Proc. Natl. Acad. Sci. USA **80:** 6533–6537.
15. **DiNocera, P. P., M. E. Digan, and I. B. Dawid.** 1983. A family of oligoadenylate-terminated transposable sequences in *Drosophila melanogaster*. J. Mol. Biol. **168:**715–727.
16. **Doerfler, W.** 1983. DNA methylation and gene activity. Annu. Rev. Biochem. **52:**93–124.
17. **Dudley, J. P.** 1987. Discrete high molecular weight RNA transcribed from the long interspersed repetitive element L1Md. Nucleic Acids Res. **15:**2581–2592.
18. **Enders, G. H., D. Ganem, and H. Varmus.** 1985. Mapping the major transcripts of ground squirrel hepatitus virus: the presumptive template for reverse transcriptase is terminally redundant. Cell **42:**297–308.
19. **Fanning, T., and M. Singer.** 1987. The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retroviral proteins. Nucleic Acids Res. **15:**2251–2260.
20. **Fanning, T. G.** 1983. Size and structure of the highly repeated BAM H1 element in mice. Nucleic Acids Res. **11:**5073–5091.
21. **Fawcett, D. H., C. K. Lister, E. Kellett, and D. J. Finnegan.** 1986. Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs. Cell **47:** 1007–1015.
22. **Grimaldi, G., J. Skowronski, and M. F. Singer.** 1984. Defining the beginning and end of Kpn1 family segments. EMBO J. **3:** 1753–1759.
23. **Hattori, M., S. Hidaka, and Y. Sakaki.** 1985. Sequence analysis of a *Kpn*I family member near the 3′-end of human β-globin gene. Nucleic Acids Res. **13:**7813–7827.
24. **Hattori, M., S. Kuhara, O. Takenaka, and Y. Sakaki.** 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.
25. **Hattori, M., and Y. Sakaki.** 1986. Dideoxy sequencing method using denatured plasmid templates. Anal. Biochem. **152:**232–238.
26. **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.
27. **Huynh, T. V., R. A. Young, and R. W. Davis.** 1985. Constructing and screening cDNA libraries in λgt10 and λgt11, p. 49–78. *In* D. M. Glover (ed.), DNA cloning, vol. 1. IRL Press, Oxford.
28. **Jacks, T., and H. E. Varmus.** 1985. Expression of the Rous Sarcoma Virus *pol* gene by ribosomal frameshifting. Science **230:**1237–1242.
29. **Jackson, M., D. Heller, and L. Leinwand.** 1985. Transcriptional measurements of mouse repeated DNA sequences. Nucleic Acids Res. **13:**3389–3403.
30. **Kimmel, B., O. K. Ole-Moiyoi, and J. R. Young.** 1987. Ingi, a 5.2-kilobase pair 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.
31. **Kole, J. 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.
32. **Kozak, M.** 1986. Bifunctional messenger RNAs in eukaryotes. Cell **47:**481–483.
33. **Lakshmikumaran, M. S., E. D'Ambrosio, L. A. Laimins, D. T. Lin, and A. V. Furano.** 1985. Long interspersed repeated DNA (LINE) causes polymorphism at the insulin 1 locus. Mol. Cell. Biol. **5:**2197–2203.
34. **Lee, T. N. H., and M. F. Singer.** 1986. Analysis of LINE-1 family sequences on a single monkey chromosome. Nucleic Acids Res. **14:**3859–3870.
35. **Lerman, M. I., R. E. Thayer, and M. F. Singer.** 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.
36. **Loeb, D. D., R. W. Padgett, S. C. Hardies, W. R. Shehee, M. B. Comer, M. H. Edgell, and C. A. Hutchison III.** 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.
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. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65:**499–559.
39. **Miyake, T., K. Migita, and Y. Sakaki.** 1983. Some *Kpn*I family members are associated with the Alu family in the human genome. Nucleic Acids Res. **11:**6837–6846.
40. **Morzycka-Wroblewska, E., E. U. Selker, J. N. Stevens, and R. L. Metzenberg.** 1985. Concerted evolution of dispersed *Neurospora crassa* 5S RNA genes: pattern of sequence conservation between allelic and nonallelic genes. Mol. Cell. Biol. **5:** 46–51.
41. **Mottez, E., P. K. Rogan, and L. Manuelidis.** 1986. Conservation in the 5′-region of the long interspersed mouse L1 repeat: implications of comparative sequence analysis. Nucleic Acids Res. **14:**3119–3136.
42. **Nomiyama, H., K. Obaru, Y. Jinno, I. Matsuda, K. Shimada, and T. Miyata.** 1986. Amplification of human argininosuccinate synthetase pseudogenes. J. Mol. Biol. **192:**221–233.
43. **Norrander, J., T. Kempe, and J. Messing.** 1983. Construction of improved M13 vectors using oligonucleotide-directed mutagenesis. Gene **26:**101–106.
44. **Paterson, B. M., and J. D. Eldridge.** 1984. α-Cardiac actin is the major sarcomeric isoform expressed in embryonic avian skeletal muscle. Science **224:**1436–1438.
45. **Peabody, D. S., and P. Berg.** 1986. Termination-reinitiation occurs in the translation of mammalian cell mRNAs. Mol. Cell. Biol. **6:**2695–2703.
46. **Peabody, D. S., S. Subramani, and P. Berg.** 1986. Effect of upstream reading frames on translation efficiency in simian virus 40 recombinants. Mol. Cell. Biol. **6:**2704–2711.
47. **Persico, M. G., G. Viglietto, G. Martini, D. Toniolo, G. Paonessa, C. Moscatelli, R. Dono, T. Vulliamy, L. Luzzatto, and M. D'Urso.** 1986. Isolation of human glucose-6-phosphate dehydrogenase (G6PD) cDNA clones: primary structure of the protein and unusual 5′ non-coding region. Nucleic Acids Res. **14:**2511–2522.

48. **Potter, S. S.** 1984. Rearranged sequences of a human *Kpn*I element. Proc. Natl. Acad. Sci. USA **81**:1012–1016.
49. **Rogers, J. H.** 1985. The origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.
50. **Sakaki, Y., M. Hattori, A. Fujita, K. Yoshioka, S. Kuhara, and O. Takenaka.** 1986. The LINE-1 family of primates may encode a reverse transcriptase-like protein. Cold Spring Harbor Symp. Quant. Biol. **51**:465–469.
51. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
52. **Schmeckpeper, B. J., A. F. Scott, and K. D. Smith.** 1984. Transcripts homologous to a long repeated DNA element in the human genome. J. Biol. Chem. **259**:1218–1225.
53. **SenGupta, D. N., B. Z. Zmudzka, P. Kumar, F. Cobianchi, J. Skowronski, and S. H. Wilson.** 1986. Sequence of human DNA polymerase β mRNA obtained through cDNA cloning. Biochem. Biophys. Res. Commun. **136**:341–347.
54. **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.
55. **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.
56. **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.
57. **Skowronski, J., and M. F. Singer.** 1986. The abundant LINE-1 family of repeated DNA sequences in mammals: genes and pseudogenes. Cold Spring Harbor Symp. Quant. Biol. **51**:457–464.
58. **Soares, M. B., E. Schon, and A. Efstratiadis.** 1985. Rat Line-1: the origin and evolution of a family of long interspersed middle repetitive DNA elements. J. Mol. Evol. **22**:117–133.
59. **Stark, G. R., and G. M. Wahl.** 1984. Gene amplification. Annu. Rev. Biochem. **53**:447–491.
60. **Sun, L., K. E. Paulson, C. W. Schmid, L. Kadyk, and L. Leinwand.** 1984. Non-Alu family interspersed repeats in human DNA and their transcriptional activity. Nucleic Acids Res. **12**: 2669–2690.
61. **Temin, H. M.** 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons and retrotranscripts, a review. Mol. Biol. Evol. **2**:455–468.
62. **Tuttleman, J. S., C. Pourcel, and J. Summers.** 1986. Formation of the pool of covalently closed circular viral DNA in Hepadnavirus-infected cells. Cell **47**:451–460.
63. **Voliva, C. G., S. L. Martin, C. A. Hutchison III, and M. H. Edgell.** 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178**:795–813.
64. **Weidemann, L. M., and R. P. Perry.** 1984. Characterization of the expressed gene and several processed pseudogenes for the mouse ribosomal protein L30 gene family. Mol. Cell. Biol. **4**: 2518–2528.
65. **Weiner, A. M., P. L. Deininger, and A. Efstratiadis.** 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**:631–661.
66. **Wilson, M. C., S. G. Sawicki, P. A. White, and J. E. Darnell.** 1978. A correlation between the rate of poly(A) shortening and half-life of messenger RNA in adenovirus transformed cells. J. Mol. Biol. **126**:23–36.
67. **Witney, F. R., and A. V. Furano.** 1984. Highly repeated DNA families in the rat. J. Biol. Chem. **259**:10481–10492.
68. **Wolf, S. F., and B. R. Migeon.** 1985. Clusters of CpG dinucleotides implicated by nuclease hypersensitivity as control elements of housekeeping genes. Nature (London) **314**:467–469.
69. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUc19 vectors. Gene **33**:103–119.
70. **Yoshinaka, Y., I. Katoh, T. D. Copeland, and S. Oroszlan.** 1985. Murine leukemia virus protease is encoded by the *gag-pol* gene and is synthesized through suppression of an amber termination codon. Proc. Natl. Acad. Sci. USA **82**:1618–1622.