MOLECULAR AND CELLULAR BIOLOGY, July 1988, p. 2797–2803
0270-7306/88/072797-07$02.00/0

Vol. 8, No. 7

A Highly Conserved Mouse Gene with a Propensity To Form Pseudogenes in Mammals

DIANA L. HELLER,[1†] KATHERINE M. GIANOLA,[2] AND LESLIE A. LEINWAND[1,2*]

Departments of Genetics[1] and Microbiology and Immunology,[2] Albert Einstein College of Medicine, Bronx, New York 10461*

Received 5 October 1987/Accepted 4 April 1988

A mouse cDNA clone corresponding to an abundantly transcribed poly(A)$^+$ mRNA was found to be represented by 200 copies in mammalian genomes. To understand the origin and nature of this sequence family, we studied two genomic members and two cDNA clones from mouse liver. The DNA sequence of the coding strand of a full-length cDNA clone was shown to have an open reading frame capable of encoding a 25-kilodalton polypeptide that has not been previously described. In vitro transcription-translation experiments verified the presence of an open reading frame encoding a protein of the predicted size. Restriction analysis of genomic DNA and DNA sequence analysis of genomic clones indicated that many of the 200 members of this family represent processed pseudogenes, with one or a small number of active structural genes. The vast majority of the genomic copies are heterogeneous in length, truncated at their 5′ ends with respect to the mRNA, and do not appear to have intervening sequences. Two distinct genomic members of this family were sequenced and found to represent incomplete copies of the mRNA. Both are 5′ truncated at slightly different points with respect to the mRNA. Both pseudogenes have multiple base changes, insertions, and deletions relative to the mRNA, and one of them encodes the poly(A) tail of the mRNA. The expression of this gene family is highest in rapidly dividing cells such as early mouse embryos and testis, but was seen in all tissues tested. This gene shows extremely high sequence conservation, extending to chicken, amphibian, and nematode genomes. Surprisingly, the gene appears to exist in only one copy in these organisms.

The genomes of higher organisms are complex, consisting of many classes of sequences ranging from single-copy genes encoding proteins to the millionfold-reiterated simple satellite sequences of no known function. Other classes of sequence have reiteration frequencies that fall between these two extremes. Many eucaryotic genes exist as multigene families whose members show tissue-specific and developmentally regulated expression. Examples include myosin, actin, and tubulin genes (2–4, 17). Included in the members of many repeated genes are processed pseudogenes that appear to have arisen and been dispersed by transposition via an RNA intermediate (12, 26). Processed pseudogenes have no intervening sequences and have accumulated mutations, and the genomic sequences encode the poly(A) tail. Lack of promoter elements prevents the pseudogene from being expressed. Retroposition has also been proposed as the origin for the thousands of interspersed repeated DNA sequences that exist in mammalian genomes and are not known to encode any proteins (25). These include repetitive DNA sequence families such as *Alu* and L1 (11). These sequences have structural features consistent with RNA-mediated transposition (14, 22, 24). The presence of long open reading frames (ORFs) in several members of the L1 interspersed repeated DNA sequence family suggests that copies of this repeated sequence family were derived from a structural gene or a small number of genes whose transcripts were copied and inserted into the genome. In other words, interspersed repeated sequence families may be an example of a highly efficient generation of processed pseudogenes.

We previously described a highly conserved mouse repetitive DNA sequence family, LLRep3, which is present in 200 copies per haploid mouse and human genomes (9). In contrast to other moderately repeated DNA sequence families, which hybridize to a broad range of cellular RNAs, representative members of the LLRep3 sequence family hybridize to a single 1.7-kilobase (kb) polysomal poly(A)$^+$ RNA that is found at high abundance in rapidly growing mammalian cells (9). We present here the characterization of the transcript and the gene family. The transcript from this family contains an ORF capable of encoding a 25-kilodalton protein. Restriction endonuclease digestion and hybridization analysis of mouse genomic DNA revealed that most copies have the appearance of processed pseudogenes and are truncated at their 5′ ends with respect to the mRNA. The two genomic members of this family which were isolated and subjected to DNA sequence analysis have no introns, multiple stop codons, and are truncated at different points relative to the 5′ end of the cDNA. While the sequence family is represented by 200 members in mammals, it is a single-copy sequence in the chicken genome, in which it is highly transcribed. It is also a unique sequence in the genomes of *Xenopus laevis*, fish, and nematodes.

MATERIALS AND METHODS

Genomic and cDNA clones. LLRep3 cDNAs were isolated by screening two mouse liver cDNA libraries with a Chinese hamster ovary cDNA clone corresponding to the CHOB cDNA clone described by Harpold et al. (7, 8). One clone derived from a C57BL/6 pBR322 liver cDNA library, while the second clone derived from a C57BL/6 λgt10 liver cDNA library. The genomic clones λ8B1-2 and λ8B2-1 were isolated from a partially digested *Eco*RI Charon 4A BALB/c mouse genomic library (provided by L. Hood, California Institute of Technology).

Hybridizations. Hybridizations to genomic Southern blots or Northern (RNA) blots were done at 65°C in 5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–1×

* Corresponding author.
† Present address: Rockefeller University, New York, NY 10021.

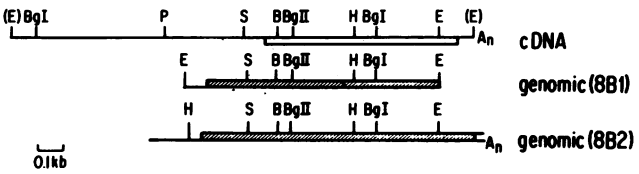

FIG. 1. Restriction map of LLRep3 full-length cDNA clone and two genomic clones. The open box in the cDNA clone is the ORF. The cross-hatched areas in the genomic clones correspond to the regions of overlap with the cDNA. Restriction sites are indicated. (E), Artificial *Eco*RI sites derived from linkers in the cDNA-cloning process. Abbreviations: E, *Eco*RI; H, *Hin*dIII; P, *Pst*I; S, *Sma*I; B, *Bam*HI; BgI, *Bgl*I, BgII, *Bgl*II.

Denhardt solution–10 mM $NaPO_4$ (pH 7.4)–10% dextran sulfate–50 μg of denatured salmon sperm DNA per ml for 12 to 16 h. Filters were prehybridized in the above buffer without dextran sulfate or radioactive probe for 2 to 6 h. After hybridization, filters were washed in 2× SSC–0.2% sodium dodecyl sulfate at 65°C for 2 to 4 h. DNA probes were radioactively labeled by the method of Rigby et al. (20) and included at a concentration of 2×10^5 to 5×10^5 cpm/ml.

RNA isolation. Total RNA was isolated from mouse embryos and tissues by the guanidine hydrochloride method (1). Nuclear and cytoplasmic RNA was extracted from tissue culture cells by lysis in Nonidet P-40 and phenol-chloroform extractions (16). Poly(A)$^+$ RNA was isolated by oligo(dT)-cellulose chromatography (Collaborative Research, Inc., Waltham, Mass.). RNA from chicken embryo fibroblasts and tissues was a gift from L.-H. Wang (Rockefeller University, New York., N.Y.).

In vitro transcription-translation. T7-generated transcripts were generated from the full-length cDNA clone whose insert had been cloned into pGEM1. The transcripts were capped (18) and used to direct protein synthesis in vitro in a wheat germ cell-free translation system (see reference 5). [^{35}S]methionine was used as a radioactive tracer. Translation products were separated on a sodium dodecyl sulfate-7% polyacrylamide gel, fluorographed (13), and exposed to X-ray film.

DNA sequence analysis. DNA fragments were subcloned into M13 bacteriophage (mp18 or mp19 or both), and DNA was sequenced by the chain termination method of Sanger et al. (21) or by the chemical base-specific reaction method of Maxam and Gilbert (15). The rapid deletion system (IBI Technologies, Inc.) was used to generate M13 clones for DNA sequencing.

RESULTS

Characterization of mouse liver cDNA clones corresponding to the LLRep3 repetitive DNA sequence family. The mouse LLRep3 family was originally described as a 200-copy interspersed repeated DNA sequence family in mouse and human genomes (9). We showed that there was a prominent 1.7-kb discrete poly(A)$^+$ transcript homologous to this family in hepatoma cell RNA (9). To determine the origin of this transcript and its relationship to the 200 genomic copies, we isolated mouse liver cDNA clones homologous to this family. Two cDNA clones isolated had inserts of 560 base pairs (bp) and 1.7 kbp. The 1.7-kb cloned insert represents a full-length cDNA copy. The restriction map of the full-length cDNA clone is shown in Fig. 1.

One (or more) members of the 200-copy sequence family can encode a protein. The presence of transcripts homologous to this family in polysomal poly(A)$^+$ RNA (8) in rodent cells suggests that the transcription product of this family encodes a protein. The coding strand was determined by hybridization of bacteriophage promoter-generated transcripts of the cDNA clone to mRNA (data not shown). To determine whether the sequence of this transcript contains an ORF, we subjected both cDNA clones to DNA sequence analysis. These clones were isolated from two different mouse liver cDNA libraries. Sequence analysis of both cDNAs revealed an ORF of 221 amino acids on the coding strand (Fig. 2). There is a poly(A) addition signal that is located 15 nucleotides from the poly(A) tail, and there is a short 18-nucleotide untranslated region at the 3′ end and a long 5′ untranslated sequence of 1.0 kb. To determine whether these sequences correspond to known genes or proteins, we searched Genbank and the Dayhoff library. The Genbank search revealed homology with the DNA sequence of rat and human α-tubulin. There is 46% homology between the mouse cDNA clone and human α-tubulin cDNA over a 1,674-nucleotide region (data not shown). There is, however, no homology at the amino acid level between the protein encoded by LLRep3 cDNA and human α-tubulin or any other known protein.

Sequence analysis of the two cDNA clones also addresses the tissue of the homogeneity of LLRep3 transcripts. If multiple members of this 200-copy family are transcribed, it might be expected that the transcripts would exhibit sequence microheterogeneity, especially in third base positions. The two cDNA clones are identical over their entire length of overlap (data not shown).

To verify the presence of the ORF predicted by the DNA sequence, a bacteriophage T7-generated transcript was synthesized and the RNA was capped and translated in vitro in a wheat germ cell-free translation system. The results are shown in Fig. 3. When compared with the extract to which no RNA had been added, the transcript of the LLRep3 family directed the synthesis of a protein migrating at about 25,000 daltons. This is consistent with the size predicted by the DNA sequence. There is also a prominent 10-kilodalton product seen in Fig. 3, lane 3. This is most likely due to premature termination or internal initiation. There are no other possible long ORFs in the sequence on the coding strand (data not shown), so the 25-kilodalton protein must derive from the sequence indicated in Fig. 1.

Genomic members of LLRep3 have features of pseudogenes. The presence of 200 genomic copies of this family and the single discrete RNA species homologous to this family led us to investigate the relationship of these 200 genomic members. It seemed unlikely that all 200 copies were functional genes, expressing a single mRNA species. The other possibility is that there could be a repeated sequence in the 5′ or 3′ untranslated region of this mRNA, while the protein-coding region is single copy. There are precedents for the latter possibility in several mRNAs which contain repetitive DNA sequences in their untranslated regions (e.g., the low-density lipoprotein receptor and mouse histocompatibility sequences) (23, 27). To distinguish between these two possibilities, the cDNA clone was digested into fragments that were radioactively labeled, and the various fragments were used as probes against genomic blots. We found that several different restriction fragments spanning the entire 1.1-kb *Pst*I-*Eco*RI fragment of the cDNA clone (Fig. 1) hybridized to approximately 200 genomic fragments (see Fig. 5 and 6, for example). These include the *Pst*I-*Sma*I and *Sma*I-*Bgl*II fragments. However, a 0.65-kb *Pst*I-*Eco*RI fragment derived from the 5′ end of the full-length cDNA clone

```
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTTCACATTGCCCTTTTTTATTT
GTAACAGGGGAGCATCCTCTTCCCATCAAGGTCTCGGCCTCACAGGCGTCCCTGTTGGC
ATAAGGTGTCATCTGCCTGCGTACGAGGACAACATGGCAGGGAGTGGCGCTCCTAACAG
TTCTATATCTCCACTGTCCAGCAGGAAATCTCTGCTCCTGGGAACCCTTCTTGGTAACT
CAACTTGAGCTCTCTGGTAGCCAGGTAGTCAACTCCTTGAACTTTGACTGTCCGTTTTG
TATTATCGTGGCTCGACGGGTGGCTCCCAGACGATGTCCCGGTAAAACTGACCAAGCTC
CCCTTTAACTGTTGGAGAATGCTGGTGTCATTCAAAAGCAGCATTAGTCCTCTCTGAGG
GTCATCCAGGGACAGGAGGCCGATGGTGACTTTGTCTGGGCCACTGAGCTGTAGGAGTC
CCATCATATTCATGGTCATCTTCAAGCCAGGTAACACAGAGAACAGCGTCTTCTTCCAC
TTGTATTCAGTGTTTTTTCTGCCTCGTGTCACCACCAGTCGTTCAGGGGTTGCATGGAC
CTGCAGGTGAGGCTTCTGGATGGTCACTTCGAGCCACGAGAACCCAGTATTCTCATACT
TACCAGTCACCTCCAGACCTAAGAAGGCCAGACCAGAGATTTCAAAGTCAATTCAGAGC
AATGGCCCAAACAAGGAATCTCCCCTATGTCCTTTCTGGAACGGGGTCCCCAGAAGAAT
CTTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTC
GGGAGAACACCAAATGGCGGATGACGCCGGTGCAGCGGGAGGGCCCGGAGGACCCGGGG
GCCCAGGATTAGGTGGCCGCGGCCGGCTTCCGCGGAGGATTCGGCAGCGGTCTTAGGGC
CCGTGGTCGTGGCCGAGGCCGTGGCCGTGGTCGAGGCCGCGGGGCTCGTGGAGGTAAA_
                                                            1
                                                           Met
CTGAAGACAAGGAGTGGATCCCCGTCACCAAGCTGGGCCGTCTGGTTAAGGAC     ATG
                                        10
Lys Ile Lys Ser Leu Glu Glu Ile Tyr Leu Phe Ser Leu Pro Ile
AAG ATC AAG TCC TTG GAG GAG ATC TAC CTG TTC TCC CTG CCC ATT
            20                                          30
Lys Glu Ser Glu Ile Ile Asp Phe Phe Leu Gly Ala Ser Leu Lys
AAG GAG TCT GAG ATC ATT GAT TTC TTC CTG GGT GCG TCC CTA AAG
                                        40
Asp Glu Val Leu Lys Ile Met Pro Val Gln Lys Gln Thr Arg Ala
GAT GAG GTT CTG AAA ATC ATG CCA GTG CAG AAG CAG ACT CGG GCT
            50                                          60
Gly Gln Arg Thr Arg Phe Lys Ala Phe Val Ala Ile Gly Asp Tyr
GGC CAG CGG ACC AGG TTC AAG GCT TTC GTC GCT ATT GGG GAC TAC
                                        70
Asn Gly His Val Gly Leu Gly Val Lys Cys Ser Lys Glu Val Ala
AAT GGT CAC GTT GGT CTT GGT GTT AAG TGC TCC AAG GAG GTT GCC
            80                                          90
Thr Ala Ile Arg Gly Ala Ile Ile Leu Ala Lys Leu Ser Ile Val
ACT GCC ATC CGA GGG GCC ATC ATC TTG GCC AAG CTT TCC ATC GTC
                                       100
Pro Val Arg Arg Gly Tyr Trp Gly Asn Lys Ile Gly Lys Pro His
CCT GTG CGG AGA GGC TAC TGG GGG AAC AAG ATT GGC AAG CCC CAC
           110                                         120
Thr Val Pro Cys Lys Val Thr Gly Arg Cys Gly Ser Val Leu Val
ACT GTT CCA TGC AAG GTG ACA GGC CGC TGT GGC TCT GTG CTG GTG
                                       130
Arg Leu Ile Pro Ala Pro Arg Gly Thr Gly Ile Val Ser Ala Pro
CGT CTC ATC CCT GCC CCC AGA GGC ACT GGC ATT GTC TCT GCT CCT
           140                                         150
Val Pro Lys Lys Leu Leu Met Met Ala Gly Ile Asp Asp Cys Tyr
GTG CCC AAG AAG CTC CTG ATG ATG GCC GGT ATA GAT GAC TGC TAC
                                       160
Thr Ser Ala Arg G y Cys Thr Ala Thr Leu Gly Asn Phe Ala Lys
ACT TCA GCC AGA GGC TGC ACT GCC ACC CTG GGC AAC TTT GCT AAG
           170                                         180
Ala Thr Phe Asp Ala Ile Ser Lys Thr Tyr Ser Tyr Leu Thr Pro
GCC ACC TTT GAT GCC ATC TCC AAG ACT TAC AGC TAC CTG ACC CCC
                                       190
Asp Leu Trp Lys Glu Thr Val Phe Thr Lys Ser Pro Tyr Gln Glu
GAC CTC TGG AAA GAG ACT GTC TTC ACC AAG TCT CCT TAT CAG GAA
           200                                         210
Phe Ser Asp His Leu Val Lys Thr His Thr Arg Val Ser Val Gln
TTC TCG GAT CAT CTT GTG AAA ACC CAC ACC AGA GTC TCT GTT CAG
                                   220 221
Arg Thr Gln Ala Pro Ala Val Ala Thr Thr OC
AGG ACC CAG GCT CCA GCT GTG GCT ACC ACA TAA GGGTTTTTATATGAGAAA

AATAAAAGAATTAAGTCTGCTGAAAAAAA
```

FIG. 2. DNA sequence and predicted protein sequence of LLRep3 full-length cDNA clone. The termination codon is marked OC for ochre. Amino acids encoded are presented above the DNA sequence.

hybridized to one prominent fragment in genomic digests (this will be discussed below; see Fig. 5B). This excludes the possibility that a repetitive element lies in the 3′ or 5′ untranslated region of the mRNA, since DNA fragments corresponding to the coding sequence hybridize to 200 copies.

To test the hypothesis that most of the 200 LLRep3 genomic members are pseudogenes created by retrotransposition and that only one or a small number are actual

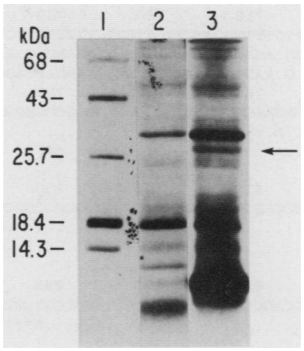

FIG. 3. In vitro translation of the LLRep3 transcript. The LLRep3 full-length cDNA clone was inserted into pGEM1, and T7 promoter-generated transcripts were produced. The capped transcript (0.5 μg) was translated in vitro in a wheat germ extract in the presence of [^{35}S]methionine. Products were electrophoresed on a sodium dodecyl sulfate-7% polyacrylamide gel and visualized by fluorography and exposure to X-ray film. Lane 1, Molecular size markers; lane 2, wheat germ extract with no exogenous RNA; lane 3, wheat germ extract with 0.5 μg of LLRep3 mRNA. The arrow indicates the protein product that is specific to lane 3.

structural genes, we studied the sequence organization of the genomic members. If most of the genomic members are processed pseudogenes, the genomic and cDNA sequences should be colinear and lack intervening sequences and the genomic sequences would have poly(A) tracts at the 3′ ends. It was shown earlier that the primary transcript of this gene was larger than the mRNA (3.6 kb compared with 1.7 kb) (7). Therefore, the gene(s) must have intervening sequences. Two different genomic copies of LLRep3 were isolated and subjected to DNA sequence analysis. The portion of one genomic clone which is homologous to the cDNA resides on a 0.94-kb *Eco*RI fragment, while that of the second genomic clone resides on two *Eco*RI fragments of 2.0 and 1.8 kb. The restriction maps of the genomic clones and their alignment with the cDNA sequence are shown in Fig. 1. The comparisons of the genomic and cDNA sequences are shown in Fig. 4. The cDNA and genomic sequences are colinear over 815 bp for λ8B1-2 and 990 bp for λ8B2-1. One genomic clone (λ8B1-2) does not contain the sequences of the 3′ end of the cDNA because of an *Eco*RI site in the genomic clone and cDNA. This genomic clone (derived from a partial *Eco*RI library) does not contain the adjacent *Eco*RI fragment. Both genomic sequences are truncated at their 5′ ends with respect to the full-length cDNA and would, therefore, represent incomplete reverse transcription products. Additionally, the two genomic sequences are truncated at slightly different places on the cDNA, suggesting that they derive from independent events. In addition, the DNA sequences flanking the LLRep3 homology are different in the two genomic clones. There are multiple base changes, insertions, and deletions in the genomic sequences relative to the cDNA. These data show that the genomic clones studied cannot encode the transcript represented by the cDNA clone. Also, there are no intervening sequences in these genomic segments. One of the clones, λ8B2-1, contains the very 3′ end of the mRNA and also contains a long A-rich region and poly(A) tail that coincides with the poly(A) tail of the mRNA. These features and the 5′ truncation of the sequences at different points are consistent with features of a retroposon.

Since it was impractical to sequence the remaining members of this family, an additional method was used to analyze

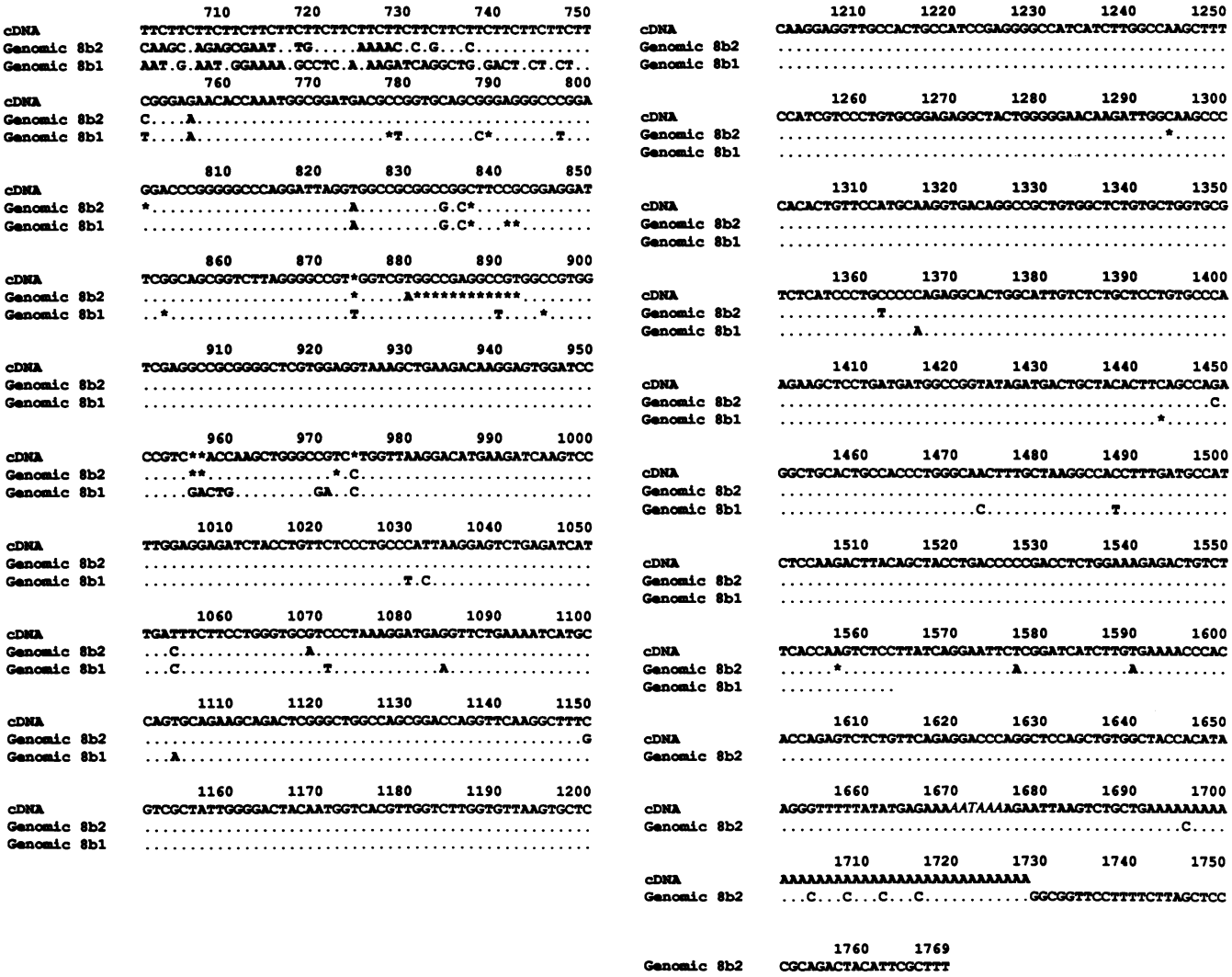

FIG. 4. Sequence comparison of cDNA and two genomic clones of the LLRep3 family in the region of their overlap. The numbering corresponds to residues of the full-length cDNA which is presented in Fig. 2. Dots (. . . . .) indicate sequence identity, and asterisks (*) indicate spaces inserted to allow alignment.

the remaining genomic copies. One method of determining how homologous different members of a repetitive family are is to determine whether they conserve restriction sites. If all or most LLRep3 genomic members are processed pseudogenes, then one would expect them to have the same structure as the cDNA. In other words, they would have no intervening sequences. We asked whether restriction sites that span the length of the cDNA are conserved among the genomic members of LLRep3. *Hin*dIII and *Eco*RI sites bound a 329-bp fragment (Fig. 1). When the LLRep3 cDNA was used to probe genomic DNA double cut with *Eco*RI and *Hin*dIII, 30% (as estimated by densitometry) of the dispersed hybridizing bands condensed to a 329-bp fragment that coincided with the *Eco*RI-*Hin*dIII fragment found in the two cDNA clones (Fig. 5A, lane 3). Similar results were obtained with *Hin*dIII-*Pvu*II (Fig. 5A, lane 5) and *Hin*dIII-*Bam*HI and *Dde-Pvu*II (data not shown). These digests span the entire ORF. These results show that at least 30% of the 200 genomic copies of LLRep3 are colinear with regard to the cDNA and are likely to be processed pseudogenes since there is no evidence of intervening sequences in these copies. However, not all of the bands cut to this single fragment. These sequences could represent older pseudogenes with increased accumulation of sequence changes.

As mentioned above, we observed that the two genomic copies of this family which we isolated were truncated at their 5′ ends with respect to the cDNA sequence. When the 5′ 650-bp *Eco*RI-*Pst*I fragment of the cDNA clone was used as a probe against *Eco*RI-digested genomic DNA, a simple pattern of hybridization was observed (Fig. 5B, lanes 1 and 2). This is in contrast to the pattern of hybridization when fragments 3′ to this were used as a probe (Fig. 5A, lanes 1 and 2). This result indicates that most if not all of the genomic copies are truncated at their 5′ ends with respect to the cDNA and cannot encode the transcript represented by the cDNA clone.

Growth-regulated expression of LLRep3 gene. LLRep3 transcripts were more abundant in cytoplasmic RNA compared with nuclear RNA, and no nuclear precursor was visible in steady-state RNA even on long exposures (Fig. 6A). This is consistent with the observation of Harpold et al. (8) regarding the short half-life of nuclear transcripts.

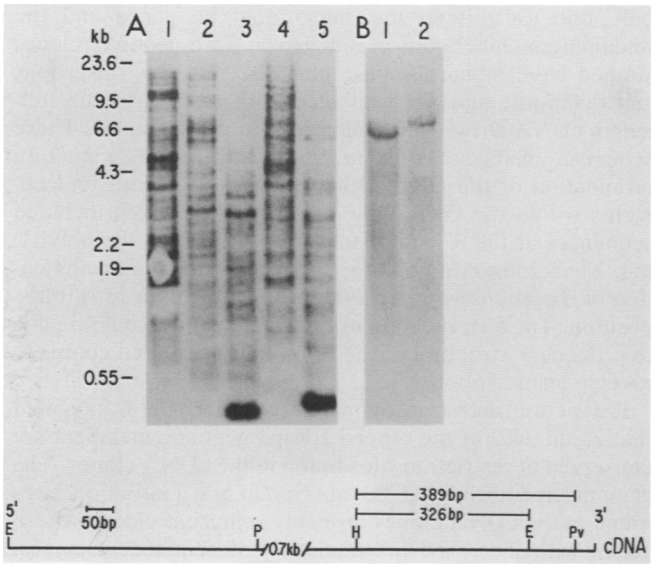

FIG. 5. Genomic analysis of LLRep3 sequences. (A) Mouse genomic DNA was digested with *Eco*RI (lane 1), *Hin*dIII (lane 2), *Eco*RI and *Hin*dIII (lane 3), *Pvu*II (lane 4), and *Hin*dIII and *Pvu*II (lane 5). The filter shown in panel A was probed with radiolabeled LLRep3 cDNA. (B) Mouse genomic DNA was digested with *Eco*RI (lane 1) or *Bam*HI (lane 2). The filter shown in panel B was probed with the 5' 0.65-kb *Eco*RI-*Pst*I fragment indicated in Fig. 1. E, *Eco*RI; P, *Pst*I; H, *Hin*dIII; Pv, *Pvu*II.

LLRep3 transcripts were only detectable in steady-state nuclear and cytoplasmic poly(A)$^+$ RNA and not in poly(A)$^-$ RNA (Fig. 6A). The LLRep3 transcripts are also found on polysomes in Chinese hamster ovary cells (7, 8). These characteristics combined with the ORF found in the sequence suggest that this transcript is translated into a protein.

We characterized the expression of this gene (or gene family) in steady-state RNA isolated from mouse embryo and adult tissues (brain, kidney, liver, and thymus). Figure 6B shows the patterns of hybridization to RNA from mouse embryos and adult mouse liver. Different amounts of the LLRep3 transcript accumulated in the various samples examined. LLRep3 transcripts were much less abundant in adult tissues than in embryonic tissues, and the abundance of the transcript decreased during development. Since it appeared that this sequence family has an unusually large number of pseudogenes, we wished to investigate the hypothesis that retrotransposition occurs much more frequently in genes that are expressed in the germ line (25). If this hypothesis is correct, one might expect that LLRep3 would be expressed at high levels in germ line tissue. LLRep3 is indeed expressed at a much higher level in testis than it is in other adult tissues (data not shown). This is consistent with the hypothesis, but does not prove it. For example, the correlation also holds that this gene is expressed in rapidly growing and dividing cell types.

Evolutionary conservation of LLRep3 in sequence but not in copy number. The above data suggest that only one or a small number of genes exist within the 200-copy family. We therefore investigated the gene number in other organisms and whether the repetitive nature of its sequence might have been a recent evolutionary event. LLRep3 is present in about 200 copies in human, rat, hamster, and mouse genomes (Fig. 7, lanes 1 to 4 and lane 9). In human genomes, there is also 1.7-kb poly(A)$^+$ RNA homologous to this family (Fig. 6C, lane 1). LLRep3 is extremely well conserved in lower organisms but not as a repetitive element. LLRep3 hybridized to a single *Pst*I restriction fragment in chicken DNA and to one to three *Pst*I fragments in the genomes of *Xenopus laevis*, goldfish, and *Caenorhabditis elegans* (Fig. 7, lanes 5 to 8). This pattern of hybridization is indicative of a single-copy structural gene. The presence of a single copy in chicken DNA prompted us to examine transcription of this sequence in chicken cells. If our theory is correct, then the single sequence in chicken represents a gene which gave rise to multiple sequences following some recent event. LLRep3 was found to be heavily transcribed in chicken poly(A)$^+$ RNA from a chicken embryo fibroblast cell line, chicken brain, and chicken kidney (Fig. 6C, lanes 2 to 4).

DISCUSSION

RNA-mediated transposition appears to be restricted to eucaryotes and could have been the origin of many interspersed repeated DNA and pseudogene sequence families. It has been shown that reiterated gene sequence families frequently consist of both genes and processed pseudogenes, the latter being derived from mRNA copies. Usually, there is

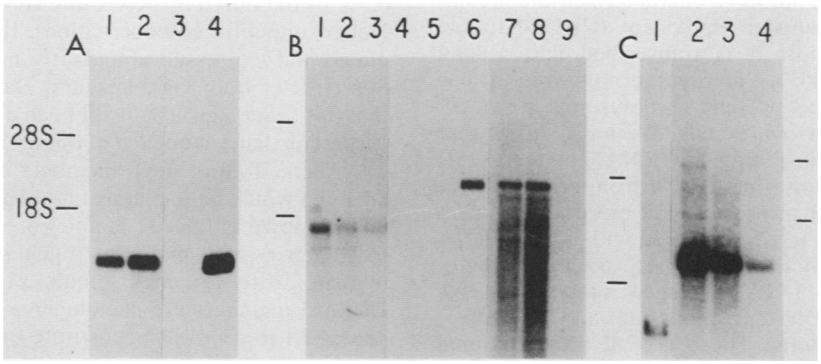

FIG. 6. Abundance levels of LLRep3 transcripts. (A) Northern hybridization of LLRep3 DNA probe to RNA from various sources. Lanes: 1, hepatoma nuclear RNA; 2, hepatoma cytoplasmic RNA; 3, hepatoma cytoplasmic poly(A)$^-$ RNA; 4, hepatoma cytoplasmic poly(A)$^+$ RNA. (B) Lanes 1, 4, and 7, 10-day-old mouse embryo; lanes 2, 5, and 8, 12-day-old mouse embryo; lanes 3, 6, and 9, adult mouse liver RNA. Lanes 1 to 3 were probed with radiolabeled LLRep3 cDNA. Lanes 4 to 6 were proved with albumin cDNA. Lanes 7 to 9 were probed with α-fetoprotein cDNA. (C) Poly(A)$^+$ RNA from HeLa cells (lane 1), chicken embryo fibroblasts (lane 2), chicken brain (lane 3), and chicken kidney (lane 4). Filters in panel C were probed with nick-translated LLRep3 cDNA.

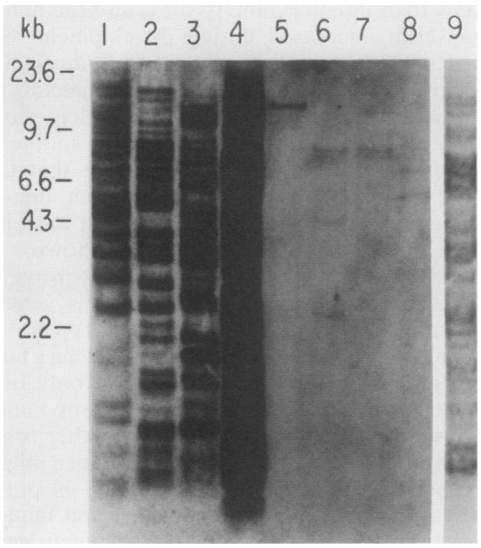

FIG. 7. Evolutionary conservation of LLRep3. *Pst*I-digested genomic DNA from human (lane 1), rat (lane 2), hamster (lane 3), mouse (lane 4), chicken (lane 5), *X. laevis* (lane 6), goldfish (lane 7), and *C. elegans* (lane 8). The filter was hybridized with radiolabeled LLRep3 cDNA. Lane 9 is a shorter exposure of lane 4. The filter was washed in 0.2× SSC at 55°C for 1 h.

a relatively small number of pseudogenes relative to the number of expressed genes (see reference 25). However, glyceraldehyde-3-phosphate dehydrogenase represents an example of a single functional gene and approximately 200 pseudogenes. In mouse DNA, there are 200 pseudogenes; in human DNA, there are perhaps 20 to 30, and in chicken DNA, there appears to be a single copy (19). Loeb et al. (14) proposed that the LINE-1 family of repeats which numbers approximately 5×10^4 copies is really a collection of pseudogenes with an as yet elusive member that is or was an active functional gene. Owing to the high copy number of the LINE family and its fortuitous transcription from adjacent structural genes (10), it is difficult to ascertain which members, if any, may still be active. The ability to look at sequence families such as LLRep3 with lower copy numbers may elucidate such events.

LLRep3 is unique compared with other repeated sequences studied in that it is transcribed into a discrete RNA species. Harpold et al. (7, 8) have characterized the metabolism of the Chinese hamster equivalent (CHOB) to the mouse LLRep3 transcript. It is transcribed into a long nuclear precursor of 3.6 kb with a rapid turnover rate. It has been demonstrated to be present in polysomal poly(A)$^+$ RNA in Chinese hamster ovary cells. We have shown that mouse transcripts of this family are RNA polymerase II products and that the transcripts derive from one strand of the repeat (10). LLRep3 transcripts are more abundant in steady-state cytoplasmic RNA compared with nuclear RNA and are found exclusively in the poly(A)$^+$ fraction of cytoplasmic RNA. Finally, DNA sequence analysis of two independently derived cDNAs demonstrates an ORF with 5′ and 3′ untranslated regions. These are all transcriptional characteristics of a structural gene.

As mentioned previously, the unusual property of a 200-copy genomic sequence family with a homologous discrete transcript suggests several possibilities. All genomic members could be transcribed into a single-size mRNA, and each LLRep3 member would represent a gene. Alternatively, only one (or a few) members could be genes and the remaining members are pseudogenes. Two genomic clones studied have numerous base changes, including insertions and deletions, and are truncated with respect to the full-length cDNA at two different points on the sequence. These sequence changes result in frame shifts and premature termination of the ORF. Other changes are silent replacements within the ORF. One genomic clone which included sequences at the 3′ end of the mRNA encodes the poly(A) tail. These observations are consistent with the accumulation of mutations within pseudogenes after their retrotransposition. The 5′ truncation of the pseudogenes could be due to secondary structure of the RNA that prevented complete reverse transcription.

Restriction digestion of mouse genomic DNA indicated that about 30% of the other LLRep3 genomic members are conserved at restriction sites found in the cDNA clones. The genomic members that do not retain the restriction sites found in the cDNA clones probably represent older pseudogenes with increased accumulation of sequence changes. One might hypothesize that these noncolinear sequences represent functional genes of this family which have intervening sequences. However, Fig. 5 suggests that there is most likely a single functional gene. Further analysis is required to unambiguously identify the gene. The strongest pieces of evidence that the genomic members of this family represent retroposons are that the two genomic clones studied here represent incomplete copies of the cDNA without intervening sequences and that one of them encodes the poly(A) tail.

Another piece of evidence that suggests that one member of LLRep3 is a functional protein-encoding gene is the conservation of LLRep3 in chicken DNA as what appears to be a single structural gene. Apparently, some event occurred during evolution that resulted in dispersal of these sequences through mammalian genomes. The single gene appears to be abundantly transcribed in the chicken as observed by hybridization of the mouse DNA probe to RNA from a chicken embryo fibroblast cell line, brain, and kidney. The chicken genome, in general, has fewer pseudogenes than seen in mouse and human genomes (25).

It will be very interesting to determine the identity of the LLRep3 gene product. The transcript appears to be ubiquitous and quite abundant. It has been estimated to be about 0.5% of the poly(A)$^+$ RNA in CHO and hepatoma cells (D. Heller, unpublished observation). It is strongly growth regulated and expressed abundantly in early mouse embryos. Searches of both Genbank and Dayhoff libraries revealed that this sequence has not yet been described. It is somewhat surprising that a sequence of this abundance has not yet been described. If all or most members of this family are genes, LLRep3 would be the largest multigene family in a mammal. Other known multigene families are only on the order of 10 to 40 genes (e.g., histone, 40 genes; actin, 16 to 20 genes; myosin, 10 to 15 genes; tubulin, 10 to 15 genes). If most LLRep3 members are pseudogenes with only one or a few genes, LLRep3 would constitute one of the largest known number of pseudogenes generated from a single gene.

We would like to know the identity, function, and expression of the putative LLRep3 protein that could be encoded by LLRep3. Since LLRep3 is expressed in many tissues and at very high levels in growing cells, it could be a housekeeping gene such as one encoding a cytoskeletal protein.

ACKNOWLEDGMENTS

We thank M. Rizzo for secretarial assistance and M. Okun and D. Shields for excellent assistance with the in vitro transcription and translation. We also thank Ken Krauter for his help with sequence alignments.

L.L. is an American Heart Association Established Investigator. This work was supported by Public Health Service grant GM 29090 to L.L. from the National Institutes of Health.

LITERATURE CITED

1. **Childs, G., R. Maxon, and L. Kedes.** 1979. A family of moderately repetitive sequences in mouse DNA. Nucleic Acids Res. **8**:4075–4090.
2. **Cleveland, D. W., M. A. Lopata, R. J. MacDonald, N. I. Cowan, W. J. Rutta, and M. W. Kirchner.** 1980. Number and evolutionary conservation of α- and β-tubulin and cytoplasmic β- and γ-actin genes using specific cDNA probes. Cell **20**:95–105.
3. **Engel, J. N., P. W. Gunning, and L. Kedes.** 1981. Isolation and characterization of human actin genes. Proc. Natl. Acad. Sci. USA **78**:4674–4678.
4. **Engel, J. N., P. W. Gunning, and L. Kedes.** 1982. The human genome contains multiple cytoplasmic actin genes. Mol. Cell. Biol. **2**:674–684.
5. **Erickson, A. H., and G. Blobel.** 1983. Cell-free translation of messenger RNA in a wheat germ system. Methods Enzymol. **96**:38–50.
6. **Grimaldi, G., and M. F. Singer.** 1983. Members of the KpnI family of long interspersed repeated sequences join and interrupt α-satellite in the monkey genome. Nucleic Acids Res. **11**: 321–333.
7. **Harpold, M. M., R. M. Evans, M. Salditt-Georgieff, and J. E. Darnell.** 1979. Production of mRNA in Chinese hamster cells: relationship of the rate of synthesis to the cytoplasmic concentration of nine specific mRNA sequences. Cell **17**:1025–1035.
8. **Harpold, M., M. C. Wilson, and J. E. Darnell, Jr.** 1981. Chinese hamster polyadenylated messenger ribonucleic acid: relationship to nonpolyadenylated sequences and relative concentration during messenger ribonucleic acid processing. Mol. Cell. Biol. **1**:188–198.
9. **Heller, D., M. Jackson, and L. Leinwand.** 1984. Organization and expression of non-Alu family interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. **173**:419–436.
10. **Jackson, M., D. Heller, and L. Leinwand.** 1985. Transcriptional measurements of mouse repeated DNA. Nucleic Acids Res. **13**: 3389–3403.
11. **Jelinek, W. R., and C. W. Schmid.** 1982. Repetitive sequences in eukaryotic DNA and their expression. Annu. Rev. Biochem. **51**: 813–814.
12. **Lemischka, I., and P. A. Sharp.** 1982. The sequences of an expressed rat α-tubulin gene and a pseudogene with an inserted repetitive element. Nature (London) **300**:330–335.
13. **Lingappa, V. R., A. Devillers-Thiery, and G. Blobel.** 1977. Nascent prehormones are intermediates in the biosynthesis of authentic bovine pituitary and growth hormones. Proc. Natl. Acad. Sci. USA **74**:2432–2436.
14. **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 several features present in other retrotransposons. Mol. Cell. Biol. **6**:168–182.
15. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
16. **Nevins, J. R.** 1980. Definition and mapping of adeno virus 2 nuclear transcription. Methods Enzymol. **65**:768–785.
17. **Nguyen, H., R. Gubits, R. Wydro, and B. Nadal-Ginard.** 1982. Sarcomeric myosin heavy chain is coded by a highly conserved multigene family. Proc. Natl. Acad. Sci. USA **79**:5230–5234.
18. **Perara, E., and V. Lingappa.** 1985. A former amino terminal signal sequence engineered to an internal location directs translocation of both flanking protein domains. J. Cell Biol. **101**: 2292–2301.
19. **Piechaczyk, M., J. M. Blanchard, S. Riaad-El Sabouty, C. Dani, L. Marty, and P. Jeanteur.** 1984. Unusual abundance of vertebrate 3-phosphate dehydrogenase pseudogenes. Nature (London) **312**:469–470.
20. **Rigby, P. W., M. Dieckmann, C. Rhodes, and P. Berg.** 1977. Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113**: 237–251.
21. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
22. **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.
23. **Steinmetz, M., J. Frelinger, D. Fisher, T. Hunkapiller, D. Pereira, S. Weissman, H. Uehara, S. S. Nathenson, and L. Hood.** 1981. Three cDNA clones encoding mouse transplantion antigens: homology to immunoglobulin genes. Cell **24**:125–134.
24. **Ullu, E., and C. Tschudi.** 1984. Alu sequences are processed 7SL RNA genes. Nature (London) **312**:171–172.
25. **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.
26. **Wilde, C. D., C. Crowlher, T. Corpe, M. Gwo-Shu Lee, and N. J. Cowan.** 1982. Evidence that a human β-tubulin pseudogene is derived from its corresponding mRNA. Nature (London) **292**:83–84.
27. **Yamamoto, T., C. G. Davis, M. S. Brown, W. J. Schneider, M. L. Casey, J. L. Goldstein, and D. W. Russell.** 1984. The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. Cell **39**:27–38.