MOLECULAR AND CELLULAR BIOLOGY, Nov. 1984, p. 2518–2528
0270-7306/84/112518-11$02.00/0

Vol. 4, No. 11

Characterization of the Expressed Gene and Several Processed Pseudogenes for the Mouse Ribosomal Protein L30 Gene Family

LEANNE M. WIEDEMANN AND ROBERT P. PERRY*

Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111

Received 2 July 1984/Accepted 21 August 1984

Five cloned genes encoding the mouse ribosomal protein L30 were isolated from a recombinant DNA library and characterized by restriction mapping and nucleotide sequence analysis. Only one of these genes has introns and is expressed; the others are inactive processed pseudogenes. The expressed gene consists of five exons and four introns spanning 2,723 nucleotides. Transcripts of this gene are processed into the mature L30 mRNA by pathways that exhibit both constraints and flexibility with regard to the order of intron excision. The L30 mRNA which is 457 to 468 nucleotides in length excluding the polyadenylic acid tail, exhibits some microheterogeneity at its 3′ end and encodes a basic protein of 115 amino acids. The 5′ portion of the rpL30 gene has some novel features which are remarkably similar to the previously characterized mouse rpL32 gene. These include homologous sequences in the −60 to −340 region, the absence of a good TATA consensus sequence, and the presence of a palindromic pyrimidine sequence that spans the cap site.

A eucaryotic ribosome is composed of four RNA components and ca. 75 distinct structural proteins (56). The genes encoding these proteins belong to the so-called housekeeping category, i.e., their activity is required for the growth and maintenance of all cell types. Although the mechanisms that regulate and coordinate ribosomal protein (rp) gene expression in eucaryotes are still largely unknown, some initial observations have suggested that regulation can occur at both transcriptional (10, 40) and translational (12, 38, 40) levels. To improve our understanding of these regulatory mechanisms it is important to know whether the genes specifying different rp's have any common structural features. Accordingly, efforts are currently being made to identify and characterize rp genes in several eucaryotic species.

Our laboratory has been concerned with mammalian rp genes, particularly those of mice. Although previous studies demonstrated that mammalian genomes contain multiple copies of each rp sequence (8, 29, 30), more recent analyses (9, 18, 38; M. Wagner and R. P. Perry, manuscript in preparation) as well as data presented here indicate that most of these copies are processed pseudogenes which are not expressed. In this communication we describe the identification and sequence characterization of the expressed gene for rpL30. A comparison of this sequence with that of another mouse rp gene, rpL32, has revealed distinctive common features, some of which appear to be shared by other genes in the housekeeping category.

MATERIALS AND METHODS

Materials. Restriction enzymes and T4 ligase were purchased from New England Biolabs. T4 polymerase was obtained from Bethesda Research Laboratories; T4 polynucleotide kinase was obtained from P-L Biochemicals, Inc.; reverse transcriptase was from Life Sciences, Inc.; and S1 nuclease was from Miles Laboratories, Inc. [γ-^{32}P]ATP (2,000 Ci/mmol) was supplied by Amersham Corp., and [α-^{32}P]dATP (800 Ci/mmol) and [α-^{32}P]dCTP (800 Ci/mmol) were supplied by New England Nuclear Corp. RNA and DNA were isolated from cultured MPC 11 mouse myeloma cells (25). Methods for DNA isolation, RNA isolation, and blot hybridization have been described previously (34).

Screening and subcloning. A library constructed from BALB/c sperm DNA partially digested with *Alu*I and *Hae*III was obtained from Davis et al. (7) and screened with an L30 cDNA probe (29) by the methods described previously (30). Restriction enzyme maps were determined by analysis of single and double digests of the genomic clones and by a modification of the partial-digest mapping procedure described by Schibler et al. (44), in which *Cla*I was used in place of *Sma*I. Since there are *Cla*I sites in both arms of Charon 4A, we were able to map the genomic clones from both ends. Because *Cla*I is sensitive to methylation (27), it is necessary to propagate the recombinant phage in a dam$^-$ host.

Regions from the rpL30-1 phage containing the entire gene (3.2-kilobase [kb] *Pst*I-*Pvu*II fragment) and the 5′ adjacent fragment (4.5-kb *Hin*dIII-*Pst*I fragment) were subcloned. Portions of the 3.2-kb fragment were further subcloned to obtain the 5′ flank (0.2-kb *Pst*I-*Sac*I fragment), IVS 3 (0.5-kb *Hin*dIII-*Hin*dIII fragment), and 3′ flank (0.3-kb *Sac*I-*Pvu*II fragment) probes (see Fig. 6). Segments spanning the regions hybridizing to the L30 cDNA from rpL30-2 (1.75-kb *Xba*I-*Xba*I fragment), rpL30-3 (6.8-kb *Bam*HI-*Bam*HI and 1.3-kb *Hin*dIII-*Hin*dIII fragments), and rpL30-4 (2.4-kb *Sal*I-*Eco*RI fragment) were also subcloned for use in sequence analysis.

The fragments were subcloned into pBR322 or the series of pUC vectors described by Vieira and Messing (53) and propagated in HB101 or JM83, respectively. Plasmids were isolated by the alkaline-sodium dodecyl sulfate method (2). Fine-structure restriction mapping of the subclones was done by the partial-digest protocol of Smith and Birnstiel (47).

DNA sequencing. Sequencing reactions were done by the chemical cleavage method of Maxam and Gilbert (26), with formic acid substituted for piperidine formate in the G + A reaction (24). Fragments were end labeled with T4 polynucleotide kinase and strand separated (26) or restricted with a second enzyme, yielding a single labeled end. The chemical reaction products were routinely electrophoresed on three separate gels, a 20% (50-cm) and two 6% (80-cm) urea–poly-

* Corresponding author.

acrylamide (30:1) gels (26), yielding an average of 200 to 300 nucleotides of sequence per labeled end.

Primer extension protocol. A 65-base-pair *Hin*fI fragment was isolated from the third exon of rpL30-1 (see Fig. 4) and end labeled with T4 polynucleotide kinase and [γ-^{32}P]ATP (26). This primer (5×10^6 cpm) was annealed to 10 μg of polyadenylated [poly(A)$^+$] cytoplasmic RNA in 80% formamide–0.4 M NaCl–10 mM PIPES [piperazine-*N*,*N*′-bis(2-ethanesulfonic acid); pH 6.5] at 50°C (4) for 18 h. The hybridization reaction also contained 30 μg of nonpolyadenylated RNA. The poly(A)$^+$ RNA was purified by chromatography on oligodeoxythymidylic acid-cellulose to remove unhybridized primer (0.2% of the counts per minute eluted with the RNA) and extended by avian myeloblastosis virus reverse transcriptase (5). The extended product was electrophoresed in a 60-cm 6% urea–polyacrylamide sequencing gel together with an unrelated sequenced fragment as a size marker.

Determination of 3′ ends with S1 nuclease. A 385-base-pair *Hin*fI-*Bam*HI fragment was isolated from the 3.2-kb *Pst*I-*Pvu*II subclone of rpL30-1. This fragment contains most of exon 5, 310 nucleotides of 3′ flanking sequence, and a few nucleotides of the pUC 12 vector. The fragment was labeled with T4 polymerase and [α-^{32}P]dATP by an adaptation of the method of O'Farrell (37). Briefly, the fragment was incubated for 2 min at 37°C with 2.5 U of T4 polymerase in 33 mM Tris-acetate (pH 7.9)–66 mM potassium acetate–10 mM magnesium acetate–0.5 mM dithiothreitol in the absence of deoxynucleotide triphosphates. Then, 50 μCi of [α-^{32}P]dATP and unlabeled dCTP, dGTP, and TTP were added to a concentration of 0.1 μM, and the polymerization reaction was allowed to progress for 5 min at 37°C. Unlabeled dATP was added to a final concentration of 0.1 μM, and the reaction was incubated for an additional 5 min to ensure complete polymerization. The end-labeled fragment was strand separated on a 5% acrylamide–bis gel (60:1) as described previously (24). The anti-coding strand was identified by sequence analysis, and 13,000 cpm of labeled DNA was hybridized for 6 h at 45°C to 3 μg of poly(A)$^+$ RNA in 50% formamide–0.6 M NaCl–10 mM PIPES (pH 6.5) (17). The hybridization mixture was digested with 4,000 U of S1 nuclease (Miles) in 0.5 ml of 160 mM NaCl–30 mM sodium acetate–3 mM zinc acetate at 37°C for 30 min (45). The undigested material was ethanol precipitated, denatured in 80% formamide, and electrophoresed on a 6% sequencing gel along with the sequenced fragment as the size marker (52).

Analysis and display of sequence data. The computer programs used during sequence acquisition, compilation, and subsequent analysis are implemented in the Institute of Cancer Research SEQuence program package and are drawn and modified from many sources (see references 14, 16, 23, 50, 51, and 54). The programs for sequence display were provided by Bob Stodola, Institute for Cancer Research.

RESULTS

Isolation and characterization of the rpL30 genomic clones. The partial cDNA for the rpL30 mRNA hybridizes to at least 15 discrete fragments of different sizes and intensities in an *Eco*RI digest of BALB/c DNA (see Fig. 6; 30). To characterize these sequences in more detail, we isolated several genomic clones from a recombinant DNA library prepared from BALB/c sperm DNA. Five nonoverlapping segments containing sequences that hybridize to L30 cDNA were identified. These cloned segments, which account for six of the *Eco*RI fragments observed in the genomic DNA hybridization pattern, were characterized by fine-structure restriction mapping and were designated rpL30-1 to rpL30-5. To delineate the regions containing the L30 mRNA sequences and to locate repetitive sequences in the vicinity, we digested the cloned DNA with various restriction enzymes and analyzed the digests by the method of Southern (49) with ^{32}P-labeled L30 cDNA and ^{32}P-labeled total mouse DNA probes (Fig. 1). All of the genes were flanked by repetitive DNA sequences. For three of them, rpL30-2, rpL30-4, and rpL30-5, the restriction fragment containing the L30 mRNA sequences also contained some repetitive DNA. Sequence analysis, discussed below, indicated that some of this repetitive DNA included members of the B2 family (19).

Sequencing of the L30 cDNA and rpL30-1. Comparison of the restriction maps of the genomic clones with that of the cDNA suggested that rpL30-1 might represent an intron-containing gene: the region of rpL30-1 that hybridized to cDNA was interrupted by restriction sites not present in the cDNA (i.e., *Hin*dIII, *Eco*RI, and *Bam*HI), and moreover, a 0.56-kb *Hin*dIII fragment that did not hybridize to the cDNA mapped internal to fragments containing cDNA sequences (Fig. 1). Therefore, the appropriate regions of rpL30-1 were subcloned into plasmid vectors and sequenced with the cDNA according to the strategies shown in Fig. 1.

The sequence data (Fig. 2) confirmed that rpL30-1 does indeed contain introns. The cDNA sequence was located between nucleotides 443 and 2711 of rpL30-1 and found to be interrupted by two introns of 1,278 and 626 nucleotides. There is only one nucleotide difference between the cDNA sequence and the corresponding portions of the rpL30-1 gene: a neutral A → G substitution at position 517. Since the cDNA was derived from cultured L-cells, which originated from a mouse of the C3H strain, whereas the genomic clone was constructed from DNA of the BALB/c strain, we suspect that this substitution represents a strain polymorphism.

The cDNA sequence contains a single extended open reading frame of 108 codons. When the amino acid sequence predicted by these codons was compared with the previously determined amino acid sequence of 32 N-terminal residues of rat liver L30 (55), we observed that the mouse and the rat L30 proteins are identical from residues 7 to 32. Moreover, a search of upstream sequences of the rpL30-1 gene revealed a separate segment which could code for the missing six amino acids and the initiating methionine residue. Since appropriate splice junction sequences were also identified, the 3′ boundary of the N-terminal encoding exon was tentatively assigned (Fig. 2). The identification of the 5′ end of this gene required additional information as discussed below.

Sequence analysis of rpL30-2, rpL30-3, and rpL30-4. To investigate the structure of other members of the rpL30 gene family, we subcloned the appropriate regions of three additional genomic clones (rpL30-2, rpL30-3, and rpL30-4) and determined their sequences by the strategies shown in Fig. 1. All three of these genes bear the structural hallmarks of processed genes which probably arose from cellular mRNA intermediates. They lack introns, contain a polyadenylic acid-rich stretch 3′ to the cleavage-polyadenylation recognition signal (AATAAA), and are flanked by direct repeats or their remnants (Fig. 3). Although sequence analysis of rpL30-5 was not done, the relative location of the *Bgl*I and *Sac*I sites at the 3′ end of the gene is identical to that in the cDNA and L30-4 (Fig. 1), suggesting that it is also a processed gene. The sequence analysis also revealed the presence of repetitive elements belonging to the mouse B2

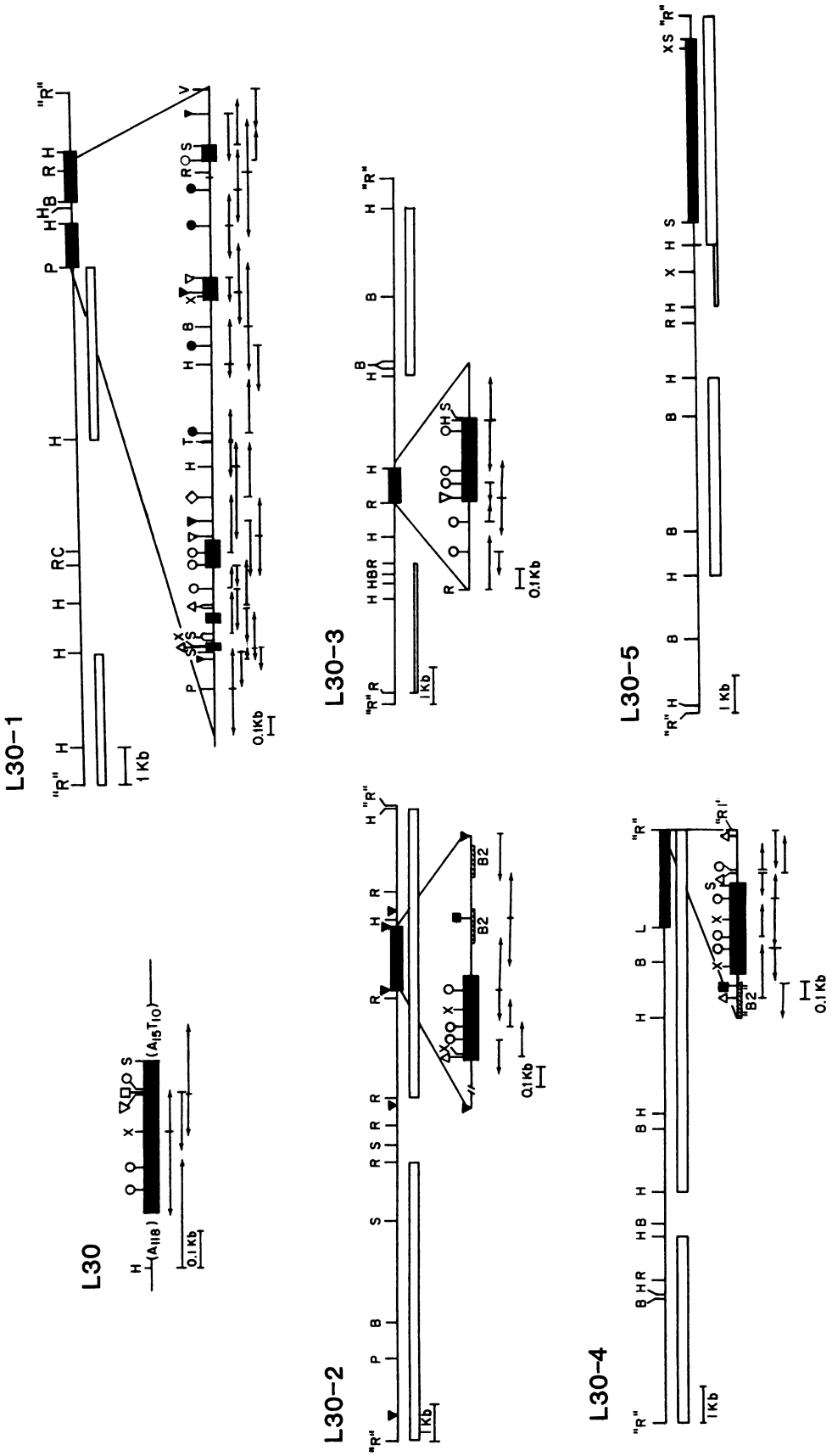


FIG. 1. Restriction maps and sequencing strategies of the rpL30 gene family. Filled bars on the lines mark cDNA hybridizing fragments. Open bars beneath the lines mark regions containing repetitive sequences; the thickness of the bar is proportional to the intensity of the hybridization signal with total mouse DNA. "R", Artificial *Eco*RI site derived from linkers used in the construction of the gene library; R, *Eco*RI; H, *Hin*dIII; B, *Bam*HI; C, *Cla*I; L, *Sal*I; ⍑, *Hin*fI; ⍒, *Bst*NI; S, *Sac*I; P, *Pst*I; ⍐, *Msp*I; ✳ *Bgl*I; ▼, *Bgl*II; ▼, *Xba*I; V, *Pvu*II; ▼, *Ava*II; ⍑, *Hph*I; T, *Taq*I; ◇, *Acc*I; ●, *Dde*I. Only the enzyme sites used for end labeling are shown. The location of B2-like sequences is shown by thin hatched boxes.

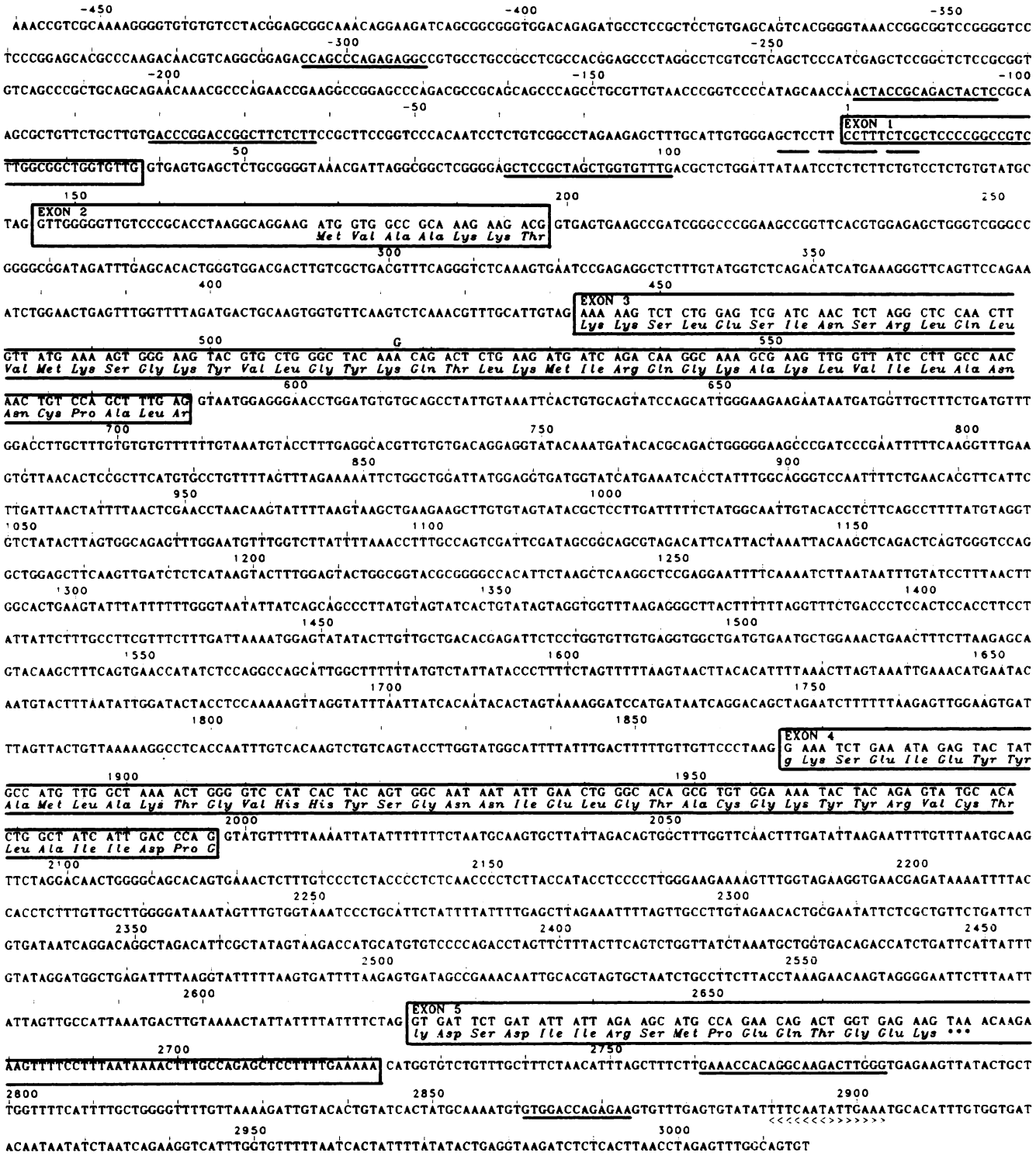


FIG. 2. Sequence of rpL30-1 gene. The cap site as determined by primer extension and homology to the processed genes is numbered 1. The exons are boxed, and the predicted protein product of the gene is shown. The cDNA sequence begins at position 443, terminates at position 2712, and is identical to the fusion of exons 3, 4, and 5 except for the A → G substitution at position 517. The true palindromic sequence at the cap site is underlined, and an inverted repeat in the 3′ flank is designated by <<< >>> symbols. Flanking and intron sequences with homology to another mouse rp gene are underlined.

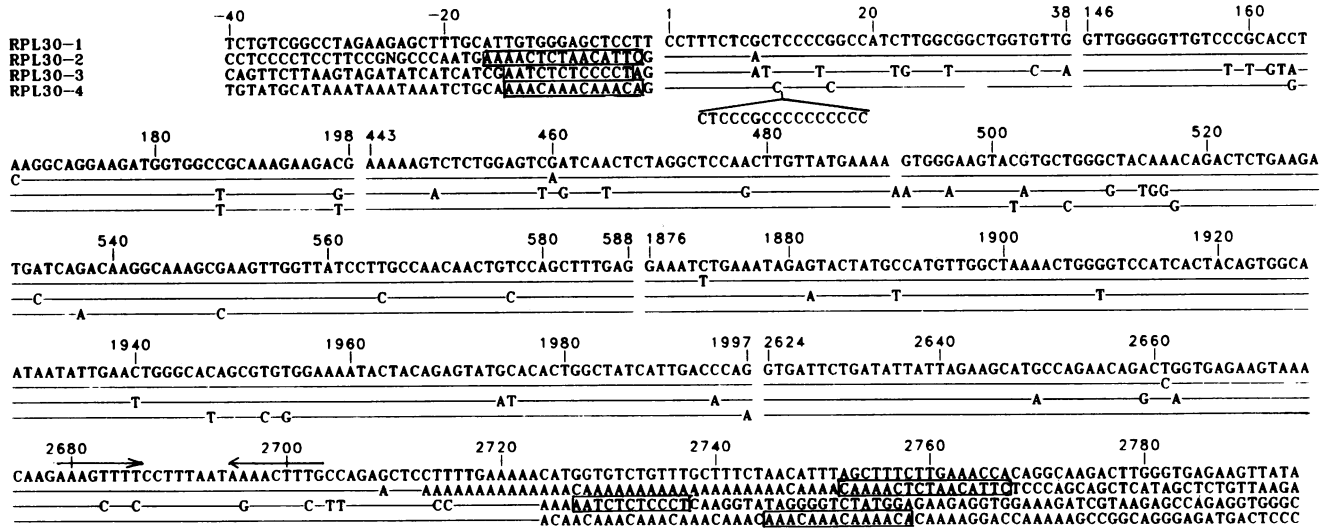


FIG. 3. Sequences of three processed genes and their flanks compared with a spliced version of the intron-containing gene rpL30-1 and its flanks. The sequences are numbered according to the intact rpL30-1 gene shown in Fig. 2. The lines indicate identity with the rpL30-1 sequence. Gaps are inserted to maintain optimal alignment. Exon junctions are separated by gaps and marked by the boundary positions. Terminal direct repeats are boxed. The 16-nucleotide insert in rpL30-4 is shown beneath the location of its insertion. rpL30-2 is 1.3% divergent at the nucleotide level; its encoded protein would differ from that of rpL30-1 by only one residue. rpL30-3 is 10.3% divergent and has a nucleotide inserted at position 491 which causes a frame shift; it could encode a 59-residue protein of which only 21 amino acids would be homologous to the authentic rpL30 protein. rpL30-4 is 3.8% divergent and would encode a protein with six amino acid replacements.

family (19) on the 3′ flank of rpL30-2 and the 5′ flank of rpL30-4 (sequence data to be published elsewhere, location shown in Fig. 1).

Since the position of the flanking direct repeats usually delineates the 5′ end of the mRNA sequence within 1 or 2 nucleotides (J. H. Rogers, Int. Rev. Cytol. Suppl., in press), we compared the sequences immediately downstream of these repeats in the three processed genes. Except for a 16-nucleotide insertion in rpL30-4, the sequences were very similar (Fig. 3), thus predicting a 5′ untranslated region of ca. 70 nucleotides. When this 70-nucleotide sequence was compared with the 5′ portion of the rpL30-1 sequence, homology was lost 32 nucleotides upstream of the AUG codon and regained 107 to 145 nucleotides further upstream. The homologous stretches were bounded by appropriate 5′ and 3′ splice junctions (32). The predicted spliced RNA from rpL30-1 is shown with flanking 5′ and 3′ sequences and compared with the three sequenced processed genes in Fig. 3. This analysis suggests that rpL30-1 has a 5′ untranslated exon of 38 base pairs and a cap site that is located in a pyrimidine stretch. It is noteworthy that there is no TATA box in the usual location 30 base pairs upstream of the predicted cap site.

Mapping of the 5′ end by primer extension. To determine whether the 5′ exon sequence predicted from the analysis of the processed genes is, in fact, present in mRNA, we used the method of primer extension. In this experiment, a 65-base-pair *Hin*fI fragment was isolated from the coding region of rpL30-1, hybridized to cytoplasmic poly(A)$^+$ mRNA, and used to reverse transcribe the mRNA sequence into a cDNA copy. The length of the extended product was used to identify the cap site (Fig. 4). The extension product was 183 ± 1 nucleotides long. This is precisely the length expected if the predicted 5′ exon were spliced onto the exon containing the AUG codon. No other major products were observed in this experiment.

Mapping of the 3′ end. Comparison of the 3′ ends of the processed genes (Fig. 3) showed variability in the polyadenylation site. Although the cleavage-polyadenylation site of the rpL30-2 gene agreed with that of the cDNA, the rpL30-3 and rpL30-4 genes contained an additional 11 nucleotides of homology with the rpL30-1 gene. To determine whether this 3′ variability among the processed genes is a reflection of microheterogeneity of the site of cleavage-polyadenylation, we examined the 3′ end of cytoplasmic poly(A)$^+$ RNA by S1 nuclease mapping (45).

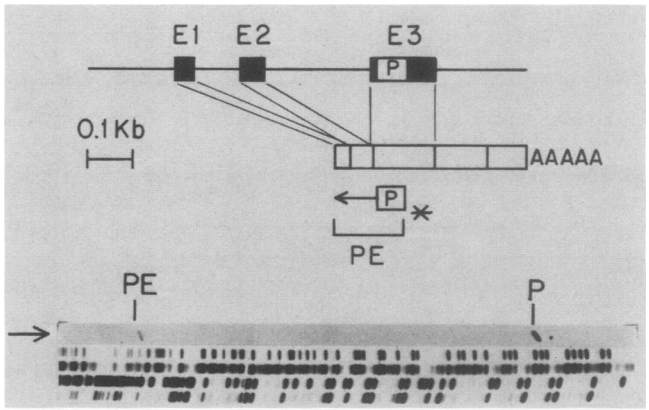


FIG. 4. Determination of the 5′ end by primer extension. A *Hin*fI fragment (P) from exon 3 of rpL30-1 (nucleotides 457 to 552) was isolated, end labeled, and used to prime cDNA synthesis from cytoplasmic poly(A)$^+$ mRNA as shown in the schematic diagram. The product (PE) and unextended primer were electrophoresed on a 6% acrylamide gel together with a size marker consisting of an unrelated DNA fragment chemically cleaved by the Maxam-Gilbert method.

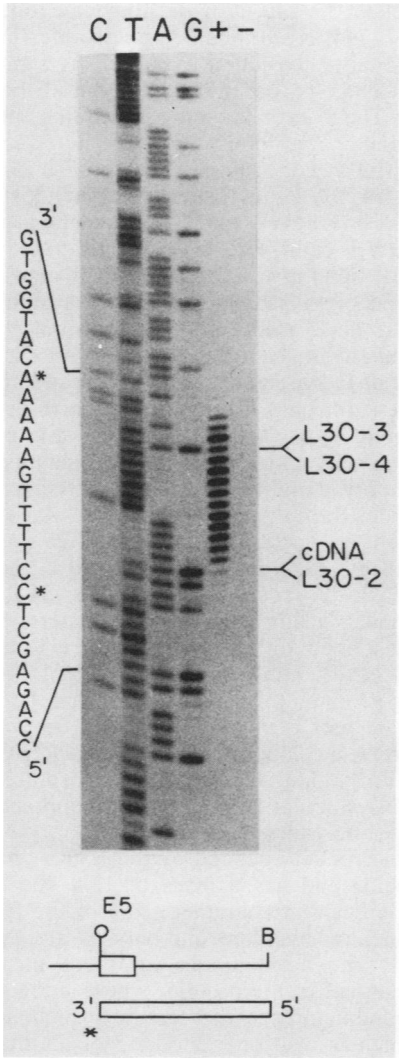


FIG. 5. Determination of the cleavage-polyadenylation sites of rpL30 mRNA. The 3'-end-labeled antisense strand of the *Hin*fI-*Bam*HI fragment containing the region from nucleotide 2627 in exon 5 to nucleotide 3015 in the 3' flank was hybridized to cytoplasmic poly(A)$^+$ RNA and digested with S1 nuclease. The protected DNA fragments were electrophoresed alongside the chemical-cleavage products of the antisense strand. The sequence of the sense strand is shown on the left with the nucleotides where polyadenylation occurred in the cDNA, and processed genes are marked with an asterisk. The corresponding bands in the S1 nuclease-treated product (+) are indicated on the right. The − lane is a control treated in the same manner as + except that poly(A)$^+$ mRNA was omitted from the hybridization reaction.

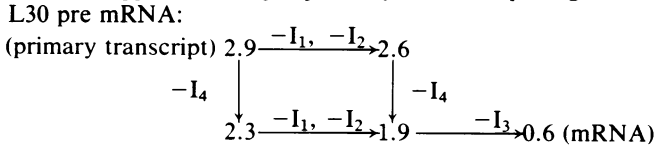
A 365-base-pair fragment, containing most of exon 5 and ca. 300 base pairs of 3' flanking DNA, was 3'-end labeled with T4 polymerase and strand separated. The minus strand was identified by sequence analysis, hybridized to the RNA, and treated with S1 nuclease. The protected fragments were electrophoresed on a 6% sequencing gel alongside the sequenced fragment (Fig. 5). The protected fragments showed an exceptionally high degree of heterogeneity, considerably greater than that normally associated with S1 nuclease nibbling. This suggests that the polyadenylation site can vary by as much as 14 nucleotides. When a correction of 1 nucleotide is made for the migration of a 3'-end-labeled S1-treated product relative to a chemically cleaved fragment (48), the two extremes (a C at 2712 and an A at 2723) observed in the comparison of cDNA and processed gene sequences are reflected in the array of protected fragments. Except for this array of 90 ± 7 nucleotide fragments, no other protected fragments were detected, indicating that there are no other functional polyadenylation sites farther downstream. These results also suggest that there are no other expressed L30 genes with substantially different 3' untranslated regions. The mRNAs from such genes would have protected, short (~50-nucleotide) fragments corresponding to the conserved coding region, and no such fragments were detected.

Unique expression of the rpL30-1 gene. Our collection of L30 genomic clones comprises only ca. 40% of the total rpL30 gene sequences that are present in the mouse genome. It is therefore important to determine whether rpL30-1 is actually an expressed gene and whether there might be other expressed intron-containing L30 genes which are not represented in our clone collection. Fortunately, these questions could be answered by examining the expression of the unique portions of the rpL30-1 gene, namely, its intron sequences.

Specific probes for the third and fourth introns, IVS 3 and IVS 4, respectively, were isolated and annealed with size-fractionated nuclear poly(A)$^+$ RNA from rapidly growing myeloma cells (Fig. 6). These probes, which were shown by a Southern blot analysis to be unique in the mouse genome (Fig. 6A), hybridized to a set of discrete RNA components of the same size as those revealed by the L30 cDNA probe (Fig. 6B). Identical results were obtained with samples of poly(A)$^+$ nuclear RNA from three other mouse myeloma lines (data not shown). Thus, the rpL30-1 gene is indeed expressed. Moreover, except for the mature size mRNA, there are no components that hybridize uniquely to the L30 cDNA. If another intron-containing rpL30 gene were expressed, one might expect to find additional cDNA-positive components that are not recognized by the unique IVS probes. Therefore, we conclude that rpL30-1 is likely to be the only intron-containing gene expressed in these cells.

To determine whether there are any additional transcribed sequences in close proximity to the rpL30-1 gene and to confirm the delineation of the 5' and 3' boundaries of the rpL30-1 gene, we isolated and subcloned the two flanking fragments (Fig. 6) and used them as hybridization probes against the nuclear poly(A)$^+$ RNA. Neither probe hybridized to any precursor or mature-size RNAs (Fig. 6B). This lack of hybridization was not due to a technical problem since both probes were capable of hybridizing to the correct-size fragments of *Eco*RI-digested genomic DNA (Fig. 6A). These results confirm that the sequences present in L30 mRNA are confined to a 2.75-kb region of the rpL30-1 gene and further indicate that there are no other stable poly(A)$^+$ RNAs produced from this region.

Processing pathway of the rpL30-1 transcript. The data (Fig. 6) suggest two major pathways for the splicing of the L30 pre mRNA:



According to this scheme, intron 4, which can be excised either before or after introns 1 and 2, is always removed before intron 3. It is apparent that there is not an intermediate for every possible combination of intron excisions,

although since we are looking at steady-state RNA, very short-lived intermediates might not be detectable. The 1.6-kb component is not easily related to any predicted intermediate that contains all of the exons. However, it could represent a cleaved but unspliced derivative of the 1.9-kb component which lacks exons 1, 2, and 3. Such processing products have been previously observed in studies of other pre-mRNAs (39) and may actually be normal splicing intermediates (33).

DISCUSSION

Genomic structure and expression of the L30-1 gene. The expressed gene for the mouse rp L30 contains five relatively short exons of 38, 53, 146, 131, and 100 nucleotides and four introns of 107, 244, 1,278, and 626 nucleotides in length. It has conventional splice junction consensus signals and an AATAAA cleavage-polyadenylation signal at the appropriate positions. The 5′ exon is totally noncoding and very rich (68%) in pyrimidine residues. We have noted that several other genes with 5′ noncoding exons are also in the housekeeping category, e.g., cytoskeletal actin (35) and cytochrome *c* (43). Moreover, many housekeeping genes have pyrimidine-rich 5′ ends, e.g., histones (46, 57), cytoskeletal actins (31, 35), tubulins (21, 22), and other mouse rp genes (9; Wagner and Perry, manuscript in preparation).

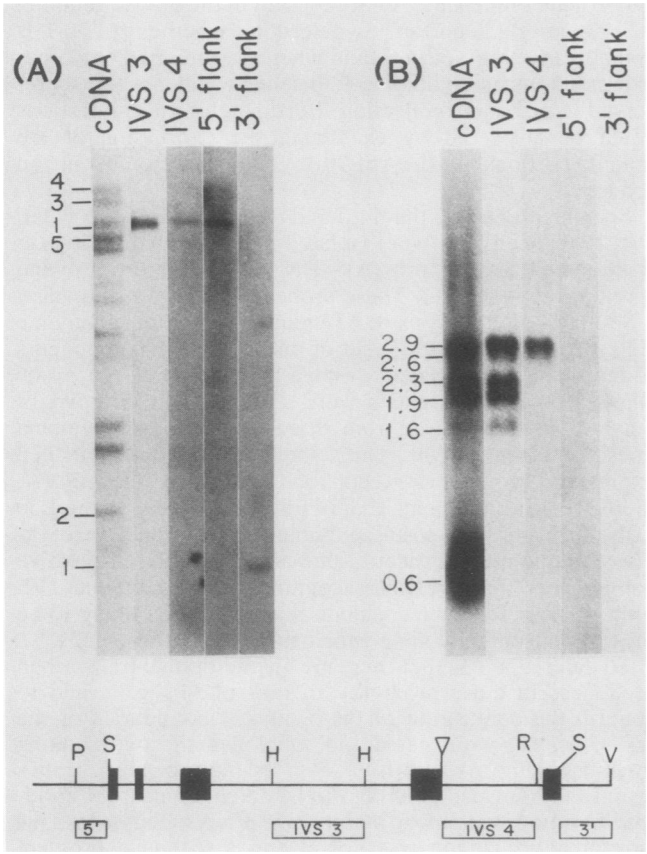


FIG. 6. Blot analyses of mouse genomic DNA and nuclear poly(A)$^+$ RNA hybridized with L30 cDNA and genomic probes. (A) *Eco*RI-digested DNA hybridized with five different probes as indicated. The genomic fragments used as probes are shown as open rectangles in the schematic diagram. Those for the 5′ flank, IVS 3, and 3′ flank were subcloned as described in the text. The IVS 4 probe was an isolated 0.6-kb *Bst*NI-*Eco*RI fragment. Restriction enzyme symbols are the same as in Fig. 1. The numbers on the left identify the five cloned rpL30 genes. The sizes of the *Eco*RI fragments corresponding to the cloned genes are: rpL30-1, 11.5 kb and 2.1 kb; rpL30-2, 2.7 kb; rpL30-3, 14.8 kb; rpL30-4, 19.0 kb; rpL30-5, 9.4 kb. The sizes of the other *Eco*RI fragments are 25.0, 8.8, 7.2, 6.5, 5.5, 5.3, 5.0, 3.7, 3.35, and 2.3 kb. (B) Hybridization of the same set of probes to poly(A)$^+$ nuclear RNA from MPC 11 cells. The relative sizes of the pre-mRNAs were estimated by comparison with a set of ribosomal RNA components run in an adjacent lane and stained with ethidium bromide. The absolute sizes in kb were obtained by normalization to the 2.9- and 0.6-kb values given by the sequence data (assuming ca. 150 to 200 nucleotides of polyadenylic acid).

The rpL30-1 gene produces a 2.9-kb primary transcript which is polyadenylated and spliced into mature L30 mRNA. The pathway of intron splicing is neither random nor completely ordered. Intron 3 is always the last to be removed, but intron 4 may be excised either before or after introns 1 and 2. The mRNA exhibits microhereogeneity at its 3′ terminus, the polyadenylic acid addition site varying from 14 to 28 nucleotides from the end of the AAUAAA signal. This 3′ heterogeneity is apparently a chronic feature of rpL30 expression since it must have been present in the ancestral animals in which the processed genes were created.

A very similar example of 3′ microheterogeneity was previously described for the yeast actin (11) and bovine prolactin (42) genes. These genes and the rpL30 gene have an interesting common property, namely that the AAUAAA signal could be part of a metastable ($\Delta G \cong -4$ kcal [ca. -16.8 kJ]) stem and loop structure in the RNA transcript (see below). Variation in the tendency to form this structure among the RNA population might lead to imprecision in the site of cleavage-polyadenylation.

The protein product encoded by the L30-1 gene contains 115 amino acids and has a mass of 12.8 kilodaltons. As expected, it is basic in character: 18% of the residues are arginine, lysine, and histidine, and only 7% are glutamic and aspartic acid. The basic amino acids are mostly in the amino-terminal half of the protein, whereas the acidic residues are predominantly in the carboxy-terminal half. This property, which is also characteristic of another recently elucidated mouse rp, rpL32, may be indicative of a ribosomal RNA binding function (9).

Processed members of the rpL30 gene family. Our analyses of four additional cloned L30 genes indicate that they are processed genes. Recent studies of other mouse ribosomal gene families, L18 (38a), L7 (18), L32 (9), and S16 (Wagner and Perry, manuscript in preparation), indicate that many members of these families are also processed genes and that only one or at most a few are functional, intron-containing genes.

The rpL30-3 gene, which has ca. 10% nucleotide sequence divergence from the rpL30-1 gene, has a single nucleotide insertion which causes a frameshift and unmasks an early termination codon that is not normally in phase. Thus this gene cannot produce an rp. The rpL30-2 and rpL30-4 genes, respectively, show 1.3 and 3.8% divergence from the rpL30-1 gene, although in this case the nucleotide changes do not result in a termination codon and the deletions or insertions fall outside of the coding sequence. However, it is unlikely that rpL30-2 is expressed. An mRNA from this gene would bear a U → C substitution at position 2661, causing mispairing and the probable generation of a 30-nucleotide fragment in the S1 nuclease protection experiment (Fig. 5); no such fragment was observed. Similarly, if rpL30-4 were expressed, the 16-nucleotide insert in its 5′ untranslated region should have produced a corresponding longer product in the

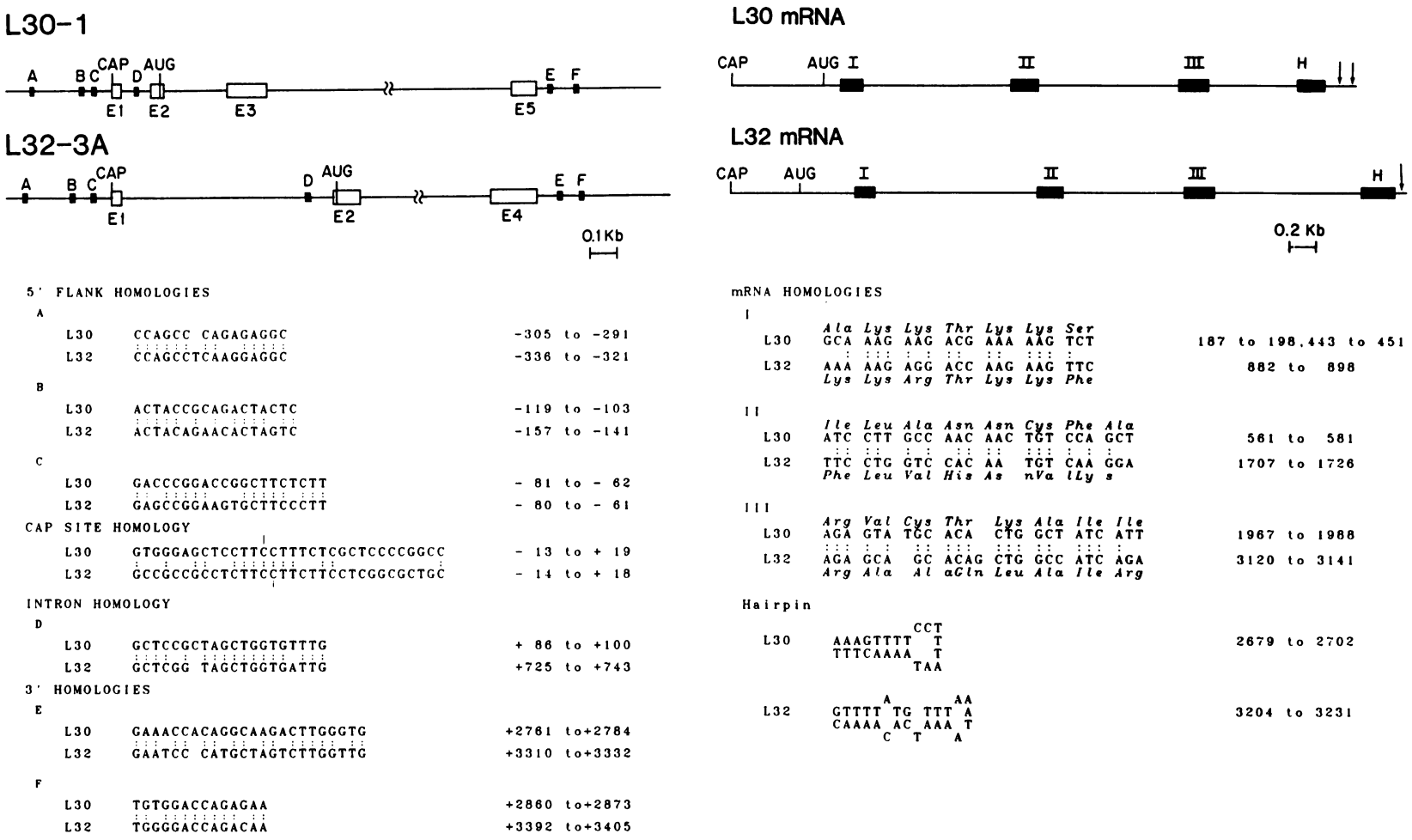


FIG. 7. Homologies identified by computer analysis of the expressed rpL30 and rpL32 genes. Left panels: The flanking and intron sequences of rpL30-1 and rpL32-3A were compared to identify locally homologous subsequences according to the computer algorithm of Goad and Kanehisa (13). Those homologies which were located in a similar position relative to the cap site, intron boundary, or polyadenylation site are designated A to F and the location is shown in the schematic diagram. Exons are shown as open boxes, and homologies are shown as filled boxes. Right panels: The sequences of the spliced mRNA products of the rpL30-1 and rpL32-3A genes were similarly compared. Homologous regions are designated I, II, and III. A potential stem and loop structure at the 3' end is designated H. The estimated ΔG values of the L30 and L32 H structures are -3.6 kcal (ca. -15 kJ) and -0.6 kcal (ca. -2.5 kJ), respectively, according to the rules tabulated by Salser (41). The vertical arrows show the polyadenylation sites.

primer extension experiment (Fig. 4), and no such product was observed. We conclude therefore, that the three sequenced L30 processed genes are in fact nonexpressed pseudogenes.

Since the recombinant DNA library was initially screened for L30 sequences under stringent hybridization and washing conditions (final wash, 0.015 M NaCl–0.0015 M sodium citrate at 68°C), we expect that we have isolated and characterized the more conserved processed members of the rpL30 gene family. Thus, it is likely that most, if not all, of the uncharacterized rpL30 genes are also pseudogenes. Studies of two other mouse rp gene families have reached similar conclusions with respect to the lack of expression of the processed genes (9; Wagner and Perry, manuscript in preparation).

Comparison of the structure of the expressed rpL30 and rpL32 genes. In searching for elements that govern the expression of rp genes one might predict at least two kinds of sequences with regulatory roles: (i) those which control transcription and RNA processing and (ii) those which might play a role in translational control or mRNA stability. Since the sequence of another expressed mouse rp gene (rpL32 [9]) was available, we compared it with the rpL30 sequence in the hope of finding homologies and structural similarities which might play a role in coordinating rp gene expression.

A comparative sequence analysis (Fig. 7) revealed several interesting similarities, particularly in the 5′ region. Both 5′ exons are noncoding, and both 5′ untranslated regions are very rich in pyrimidine residues. Both cap sites initiate with a cytosine that is located in an almost perfect palindromic pyrimidine stretch. In fact, the sequences surrounding the cap site form an interesting common motif. Neither gene has a canonical TATA box 25 to 30 nucleotides upstream of the cap site, which apparently helps define the position of polymerase II initiation (3). The appropriately located sequences most homologous to the TATA consensus are TAGAAGA for rpL30 and CATCATA for rpL32. The G in the third position of the rpL30 box is considered to be particularly unfavorable for TATA function (3). Although viral genes without TATA consensus sequences have been identified (1, 6, 15), only a few nuclear genes lacking these sequences have been found in higher eucaryotes (21, 28).

A computer program was used to compare sequences up to 500 base pairs upstream of the cap sites. Confining our search to homologies that are in close proximity (within 50 bases) of each other when the cap sites are aligned, we identified a number of homologies in addition to the cap site. Some of these homologous sequences, e.g., those designated A, B, and C (Fig. 7), are located such that they could have a role in regulating the transcription of the rp genes. One of the homologies (C) is ca. 80 base pairs upstream from the cap, a region that has been implicated in the transcriptional regulation of several other eucaryotic genes (3). The absence of a canonical TATA box and the presence of these common sequences may provide the rp genes with a distinctive promoter. Experiments are currently underway to test these ideas.

A similar comparison of the intronic and 3′ flanking regions revealed a common sequence in the 3′ portion of the first intron and two homologous stretches, E and F, in the 3′ flanks. Further experiments will be needed to determine the biological importance of these sequences. The sequence TACTAAC, present in introns 1 and 2 of the L32 gene (9) and shown to be required for proper splicing of yeast introns (20), is absent in the rpL30 gene.

If translational control is an important mode of regulation for the coordinate synthesis of ribosomal proteins, one might hope to find some sequence homology between mature rp mRNAs, possibly near the 5′ end. We have already noted the remarkably conserved cap site sequence. A further comparison of the mRNA sequence identified another homologous segment near the beginning of the translated region. This segment, designated I (Fig. 7), is similar in both nucleotide and amino acid sequence, even though it is interrupted by an intron in rpL30 and is uninterrupted in rpL32. Two other homologous segments (II and III) were found in the coding region. For these segments the nucleotide homology is very good, but the translation products are totally unrelated owing to a difference in reading frame. Conceivably, these homologous sequences could be recognized by a factor that enhances or blocks translation of the mRNAs. Finally, we noted that both of these genes contain a sequence at their 3′ end (H) which could form a stem-loop structure encompassing the AATAAA signal. However, the estimated stability of the L30 structure ($\Delta G \cong -3.6$ kcal [ca. -14.2 kJ]) is considerably greater than that of the L32 structure ($\Delta G \cong -0.6$ kcal [ca. -2.5 kJ]).

It is important to keep in mind that the homologies described above are in genes which are functionally related but which have no obvious evolutionary relationship. When regions of homology are identified in the cognate genes of two different organisms or in evolutionarily related genes in the same organism, one can argue either that the homologies are vestigial remnants of the ancestral sequence which has not yet diverged because insufficient time has elapsed or that the sequences have been preserved because they are functionally important. There is no such ambiguity for comparisons among rp genes, and therefore, any sequence homology that is not due to random occurrence may be credited to convergent evolution, presumably to fulfill a functional requirement.

ACKNOWLEDGMENTS

We are indebted to Dawn E. Kelley and Joni Brill for providing myeloma cell RNA and mouse liver DNA preparations. We also thank Kalin Dudov for stimulating discussions.

The research was supported by a grant from the National Science Foundation, by a Public Health Service grant from the National Institutes of Health, and by an appropriation from the Commonwealth of Pennsylvania.

LITERATURE CITED

1. **Baker, C. C., and E. B. Ziff.** 1981. Promoters and heterogeneous 5′ termini of the messenger RNAs of adenovirus serotype 2. J. Mol. Biol. **149**:189–221.
2. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7**:1513–1523.
3. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. **50**:349–383.
4. **Casey, J., and N. Davidson.** 1977. Rates of formation and thermal stabilities of RNA: DNA duplexes at high concentrations of formamide. Nucleic Acids Res. **4**:1539–1552.
5. **Caton, A. J., G. G. Brownlee, J. W. Yewdell, and W. Gerhard.** 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell **31**:417–427.
6. **Cattaneo, R., H. Will, G. Darai, E. Pfaff, and H. Schaller.** 1983. Signals regulating hepatitis B surface antigen transcription. EMBO J. **2**:511–514.

7. **Davis, M. M., K. Calame, P. W. Early, D. L. Livant, R. Joho, I. L. Weismann, and L. Hood.** 1980. An immunoglobulin heavy chain gene is formed by at least two recombination events. Nature (London) **283**:733–739.
8. **D'Eustachio, P., O. Meyuhas, F. Ruddle, and R. P. Perry.** 1981. Chromosomal distribution of ribosomal protein genes in the mouse. Cell **24**:307–312.
9. **Dudov, K. P., and R. P. Perry.** 1984. The gene family encoding the mouse ribosomal protein L32: characteristics of the uniquely expressed intron-containing gene and an unmutated processed gene. Cell **37**:457–468.
10. **Faliks, D., and O. Meyuhas.** 1982. Coordinate regulation of ribosomal protein mRNA level in regenerating rat liver. Study with corresponding mouse cloned cDNAs. Nucleic Acids Res. **10**:789–801.
11. **Gallwitz, D., F. Perrin, and R. Seidel.** 1981. The actin gene in yeast *Saccharomyces cerevisiae*: 5′ and 3′ end mapping flanking and putative regulatory sequences. Nucleic Acids Res. **9**:6339–6350.
12. **Geyer, P. K., O. Meyuhas, R. P. Perry, and L. F. Johnson.** 1982. Regulation of ribosomal protein mRNA content and translation in growth-stimulated mouse fibroblasts. Mol. Cell. Biol. **2**:685–693.
13. **Goad, W. B., and M. I. Kanehisa.** 1982. Pattern recognition in nucleic acids: a general method for finding local homologies and symmetries. Nucleic Acids Res. **10**:247–263.
14. **Isono, K.** 1982. Computer programs to analyse DNA and amino acid sequence data. Nucleic Acids Res. **10**:85–89.
15. **Hartzell, S. W., B. J. Byrne, and K. N. Subramanian.** 1984. Mapping of the late promoter of simian virus 40. Proc. Natl. Acad. Sci. U.S.A. **81**:23–27.
16. **Kanehisa, M. I.** 1982. Los Alamo sequence analysis package for nucleic acids and proteins. Nucleic Acids Res. **10**:183–196.
17. **Kelley, D. E., C. Coleclough, and R. P. Perry.** 1982. Functional significance and evolutionary development of the 5′-terminal regions of immunoglobulin variable-region genes. Cell **29**:681–689.
18. **Klein, A., and O. Meyuhas.** 1984. A multigene family of intron lacking and containing genes coding for mouse ribosomal protein L7. Nucleic Acids Res. **12**:3763–3776.
19. **Krayev, A. S., T. V. Markusheva, D. A. Kramerov, A. P. Ryskov, K. G. Skryabin, A. A. Bayev, and G. P. Georgiev.** 1982. Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. Nucleic Acids Res. **10**:7461–7475.
20. **Langford, C. J., and D. Gallwitz.** 1983. Evidence for an intron-contained sequence required for the splicing of yeast RNA polymerase II transcripts. Cell **33**:519–527.
21. **Lee, M. G.-S., S. A. Lewis, C. D. Wilde, and N. J. Cowan.** 1983.Evolutionary history of a multigene family: an expressed human β-tubulin gene and three processed pseudogenes. Cell **33**:477–487.
22. **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.
23. **Maizel, J. V., Jr., and R. P. Lenk.** 1981. Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc. Natl. Acad. Sci. U.S.A. **78**:7665–7669.
24. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual, p. 475–478. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
25. **Mather, E. L., and R. P. Perry.** 1981. Transcriptional regulation of immunoglobulin genes. Nucleic Acids Res. **9**:6855–6867.
26. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base specific chemical cleavages. Methods Enzymol. **65**:499–560.
27. **Mayer, H., R. Grosschedl, H. Schütte, and G. Hobom.** 1981. *Cla*I, a new restriction endonuclease from *Caryophanon latum* L. Nucleic Acids Res. **9**:4833–4845.
28. **Melton, D. W., D. S. Konecki, J. Brennand, and C. T. Caskey.** 1984. Structure, expression and mutation of the hypoxanthine phosphoribosyltransferase gene. Proc. Natl. Acad. Sci. U.S.A. **81**:2147–2151.
29. **Meyuhas, O., and R. P. Perry.** 1980. Construction and identification of cDNA clones for several mouse ribosomal proteins: application for the study of r-protein gene expression. Gene **10**:113–129.
30. **Monk, R. J., O. Meyuhas, and R. P. Perry.** 1981. Mammals have multiple genes for individual ribosomal proteins. Cell **24**:301–306.
31. **Moos, M., and D. Gallwitz.** 1983. Structure of two human β-actin-related processed genes one of which is located next to a simple repetitive sequence. EMBO J. **2**:757–761.
32. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10**:459–472.
33. **Mount, S. M., I. Pettersson, M. Hinterberger, A. Karmas, and J. A. Steitz.** 1983. The U1 small nuclear RNA-protein complex selectively binds a 5′ splice site *in vitro*. Cell **33**:509–518.
34. **Nelson, K. J., J. Haimovich, and R. P. Perry.** 1983. Characterization of productive and sterile transcripts from the immunoglobulin heavy chain locus: processing of μ_m and μ_s mRNA. Mol. Cell. Biol. **3**:1317–1332.
35. **Nudel, U., R. Zabut, M. Shani, S. Newman, Z. Levy, and D. Yaffe.** 1983. The nucleotide sequence of the rat cytoplasmic β-actin gene. Nucleic Acids Res. **11**:1759–1771.
36. **O'Farrell, P.** 1981. Replacement synthesis method for labeling DNA fragments. Bethesda Research Labs Focus **3**:1.
37. **Pearson, N. J., H. M. Fried, and J. R. Warner.** 1982. Yeast use translational control to compensate for extra copies of a ribosomal protein gene. Cell **29**:347–355.
38. **Peled-Yalif, E., I. Cohen-Binder, and O. Meyuhas.** 1984. Isolation and characterization of four ribosomal protein-L18 genes that appear to be processed pseudogenes. Gene **29**:157–166.
39. **Perry, R. P., D. E. Kelley, C. Coleclough, J. G. Seidman, P. Leder, S. Tonegawa, G. Matthyssens, and M. Weigert.** 1980. Transcription of mouse *k* chain genes: implications for allelic exclusion. Cell **77**:1937–1941.
40. **Pierandrei-Amaldi, P., N. Campioni, E. Beccari, I. Bozzoni, and F. Amaldi.** 1982. Expression of ribosomal protein genes in *Xenopus laevis* development. Cell **30**:163–171.
41. **Salser, W.** 1977. Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp. Quant. Biol. **42**:985–1002.
42. **Sasavage, N. L., M. Smith, S. Gillam, R. P. Woychik, and F. M. Rottman.** 1982. Variation in the polyadenylation site of bovine prolactin mRNA. Proc. Natl. Acad. Sci. U.S.A. **79**:223–227.
43. **Scarpulla, R. C., and R. Wu.** 1983. Nonallelic members of the cytochrome c multigene family of the rat may arise through different messenger RNAs. Cell **32**:473–482.
44. **Schibler, U., A.-C. Pittet, R. A. Young, O. Hagenbüchle, M. Tosi, S. Gellman, and P. K. Wellauer.** 1982. The mouse α-amylase multigene family: sequence organization of members expressed in the pancreas, salivary gland and liver. J. Mol. Biol. **155**:247–266.
45. **Sharp, P. A., A. J. Berk, and S. M. Berget.** 1980. Transcription maps of adenovirus. Methods Enzymol. **65**:750–768.
46. **Sittman, D. B., R. A. Graves, and W. F. Marzluff.** 1983. Structure of a cluster of mouse histone genes. Nucleic Acids Res. **11**:6679–6697.
47. **Smith, H. O., and M. L. Birnstiel.** 1976. A simple method for DNA restriction site mapping. Nucleic Acids Res. **3**:2387–2398.
48. **Sollner-Webb, B., and R. H. Reeder.** 1979. The nucleotide sequence of the initiation and termination sites for ribosomal RNA transcription in *X. laevis*. Cell **18**:485–499.
49. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
50. **Staden, R.** 1978. Sequence handling by computer. Nucleic Acids Res. **4**:4037–4057.
51. **Staden, R.** 1978. Further procedures for sequence analysis by computer. Nucleic Acids Res. **5**:1013–1015.
52. **Van Ness, B. G., M. Weigert, C. Coleclough, E. L. Mather, D. E. Kelley, and R. P. Perry.** 1981. Transcription of the unrearranged mouse C_κ locus: sequence of the initiation region and comparison of activity with a rearranged V_κ-$C_\kappa$gene. Cell **27**:593–602.
53. **Vieira, J., and J. Messing.** 1982. The pUC plasmids, an M13mp7 derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene **19**:259–268.

54. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. U.S.A. **80**:726–730.
55. **Wittmann-Liebold, B., A. W. Geissler, A. Lin, and I. G. Wool.** 1979. Sequence of the amino-terminal region of rat liver ribosomal proteins S4, S6, S8, L6, L7a, L18, L27, L30, L37, L37a, and L39. J. Supramol. Struct. **12**:425–433.
56. **Wool, I. G.** 1979. The structure and function of eukaryotic ribosomes. Annu. Rev. Biochem. **48**:719–754.
57. **Zhong, R., R. G. Roeder, and N. Heintz.** 1983. The primary structure and expression of four cloned human histone genes. Nucleic Acids Res. **11**:7409–7425.