MOLECULAR AND CELLULAR BIOLOGY, Jan. 1990, p. 316–323 Vol. 10, No. 1
0270-7306/90/010316-08$02.00/0

The *Drosophila Hrb98DE* Locus Encodes Four Protein Isoforms Homologous to the A1 Protein of Mammalian Heterogeneous Nuclear Ribonucleoprotein Complexes

SUSAN R. HAYNES,[1]* GOPA RAYCHAUDHURI,[2] AND ANN L. BEYER[2]

Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, Bethesda, Maryland 20892,[1] and Department of Microbiology, University of Virginia School of Medicine, Charlottesville, Virginia 22908[2]

Received 24 July 1989/Accepted 6 October 1989

The *Drosophila Hrb98DE* locus encodes proteins that are highly homologous to the mammalian A1 protein, a major component of heterogeneous nuclear ribonucleoprotein (RNP) particles. The *Hrb98DE* locus is transcribed throughout development, with the highest transcript levels found in ovaries, early embryos, and pupae. Eight different transcripts are produced by the use of combinations of alternative promoters, exons, and splice acceptor sites; the various species are not all equally abundant. The 3′-most exon is unusual in that it is completely noncoding. These transcripts can potentially generate four protein isoforms that differ in their N-terminal 16 to 21 amino acids but are identical in the remainder of the protein, including the RNP consensus motif domain and the glycine-rich domain characteristic of the mammalian A1 protein. We suggest that these sequence differences could affect the affinities of the proteins for RNA or other protein components of heterogeneous nuclear RNP complexes, leading to differences in function.

Transcripts synthesized by RNA polymerase II rapidly associate with a specific set of nuclear proteins to form heterogeneous nuclear ribonucleoprotein (hnRNP) complexes (for recent reviews, see references 12, 16, and 18). These RNA-protein complexes are the substrates for the processing of pre-mRNA and its transport to the cytoplasm. Studies of complexes isolated by sucrose gradient sedimentation (2), in vivo UV cross-linking (17, 19), and immunopurification (11, 35) have identified a characteristic set of proteins that are bound to the hnRNA. One of the major components of mammalian hnRNP complexes is the A1 protein, a basic, glycine-rich species. The sequence of the A1 protein has been determined by partial sequence analysis of protein purified from HeLa hnRNP particles (28, 41) and by sequencing rat and human cDNA clones (8, 14, 41); also, a human genomic clone for the locus has recently been isolated (3). The mammalian A1 protein has two domains. The N-terminal domain consists of two copies of an RNA-binding domain that is approximately 90 amino acids long. Each copy contains short, conserved peptide sequences, termed RNP consensus sequences 1 and 2, that have been found in many, but not all, RNA-binding proteins (1, 18, 38). Photochemical cross-linking experiments have demonstrated that certain phenylalanine residues in the RNP consensus sequences can be cross-linked to nucleic acid, and thus these residues may form part of a nucleic acid-binding pocket (32). The C-terminal domain is glycine rich and contains interspersed aromatic amino acids. This region is thought to be involved in both protein-protein and protein-RNA interactions (13).

Most of the data on hnRNP complexes have been derived from studies of mammalian tissue culture cells. Comparatively little is known about the structure and composition of hnRNP complexes in *Drosophila melanogaster* or other invertebrates. Risau et al. (40) identified proteins of 27, 37, 52, and 57 kilodaltons (kDa) as major components of *Drosophila* hnRNP complexes sedimenting between 50 and 150S. No further biochemical characterization of these proteins has been reported, and their relationship to the major mammalian proteins is unknown. Recently, we isolated a cDNA clone (p9) encoding a protein that, on the basis of sequence and structural comparisons, is probably a *Drosophila* homolog of the mammalian A1 hnRNP protein (23). The *Drosophila* protein encoded by clone p9 has three domains relative to the A1 protein: a short, N-terminal domain that has no counterpart in A1, a domain containing the RNP consensus sequences, and a C-terminal domain consisting of 43% glycine residues. The RNP consensus sequence domain has 58% sequence identity with the RNA-binding domains of the rat A1 protein. The glycine-rich regions of the two proteins are poorly conserved in exact sequence but similar in composition; in particular, the *Drosophila* protein also shows interspersion of aromatic amino acids. We now report further characterization of this *Drosophila* gene and its transcripts. The use of alternative exons and splice acceptor sites generates eight different transcripts. These transcripts encode four protein isoforms that differ in the short N-terminal domain. All isoforms contain the RNP consensus sequence domain and are thus likely to be components of *Drosophila* hnRNP complexes.

MATERIALS AND METHODS

Isolation and characterization of clones. The genomic clone was isolated from the *Drosophila* Canton S library of Maniatis et al. (31), and cDNA clones were derived from the following sources: p9, the Oregon R pupal library of Goldschmidt-Clermont (24); L3, the Oregon R 0- to 3-h embryonic library of Kauvar (37); ov12, the Canton S ovarian library prepared by L. Kalfayan and the Recombinant DNA Facility of the Laboratory for Reproductive Biology at the University of North Carolina, Chapel Hill. Sequence determination was done by the dideoxy method (42).

Characterization of transcripts. Preparation of RNA by guanidinium thiocyanate-phenol extraction, methylmercuric hydroxide gels, and hybridizations were performed as previously described (15). Reverse transcription reactions were

* Corresponding author.

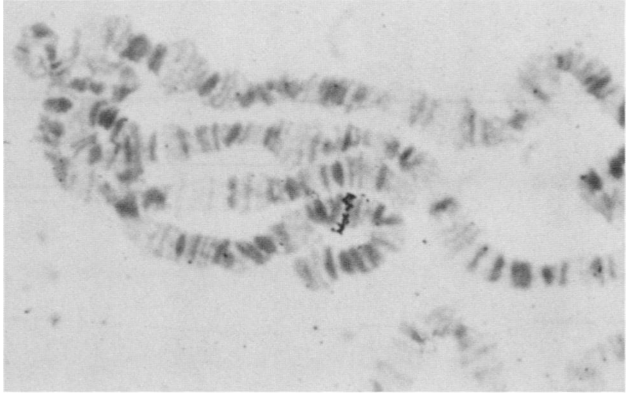

FIG. 1. Cytological localization. Polytene chromosomes from salivary glands were hybridized with a genomic clone corresponding to the p9 gene. A single site of hybridization at 98DE is seen.

done by annealing 1 μg of poly(A)$^+$ RNA with the appropriate primer in 20 mM Tris (pH 8.3)–0.1 mM EDTA–100 mM NaCl and then adding an equal volume of 2× reverse transcription (RT) buffer (80 mM Tris [pH 8.3], 80 mM KCl, 12 mM $MgCl_2$, 10 mM dithiothreitol, 200 μg of actinomycin D per ml, 400 μM deoxynucleoside triphosphates) and 10 U of avian myeloblastosis virus reverse transcriptase (Promega Biotec). S1 protection analyses were done as described previously (22). Polymerase chain reactions (PCR [29]) were done under conditions recommended by the supplier of the Taq polymerase (U.S. Biochemical Corp.), using 20 to 40 cycles of 1 min of denaturation at 94°C, 2 min of annealing at 55°C, and 2.5 min of extension at 70°C. The products were analyzed on gels containing 4% Nusieve (FMC BioProducts) and 1% standard agarose.

In situ hybridizations. Squashes of salivary glands of wild-type Oregon R *Drosophila* larvae were prepared and hybridized according to the method of Bingham et al. (4). The probe was a genomic clone containing the *Hrb98DE* locus labeled with [^{3}H]dTTP (ICN Pharmaceuticals Inc.).

RESULTS

Cytological localization. To determine the location of the gene corresponding to the p9 cDNA clone, a genomic clone containing the gene (see below) was hybridized to polytene chromosomes from salivary glands of wild-type *Drosophila* larvae. A single band of hybridization was seen on the right arm of the third chromosome at 98DE (Fig. 1). We have named this gene *Hrb98DE*, since, based on sequence homology to the rat A1 protein (14), it probably encodes an hnRNA-binding protein. A portion of the chromosome around 98DE has recently been cloned in connection with the study of the fork head (*fkh*) locus (45). The entry point for the chromosome walk was the E7Δ6 cDNA clone isolated by Knust et al. (27), using a probe that contains many GGN triplets (pen repeats [23]) on both strands; these repeats are also abundant in transcripts from the *Hrb98DE* gene. Comparison of genomic restriction maps, Southern hybridization experiments, and partial sequence analysis of the E7Δ6 cDNA clone (data not shown) demonstrated that this clone is derived from the *Hrb98DE* locus. Thus, *Hrb98DE* is located 15 kilobases (kb) distal of the *fkh* locus, in the region −4 to +2 kb on the chromosome walk (see Fig. 1 of reference 45).

Developmental expression and structure of *Hrb98DE* transcripts. To determine the sizes and pattern of expression of transcripts from the *Hrb98DE* locus, a probe from the p9 clone was hybridized to a blot containing poly(A)$^+$ RNA from different developmental stages. A single band of 2.3 kb was present at all stages, although its level varied (Fig. 2). The lower panel shows the same blot rehybridized with a probe from the ribosomal protein gene *rp49* (34), which provides an estimate of the relative amount of RNA loaded in each lane (46). The *Hrb98DE* transcript levels are high in ovaries and early embryos, and decline later in embryogenesis. (The relatively high level in the 4- to 12-h sample is due to overloading of that lane, as can be seen by comparison with the *rp49* transcript levels.) Transcript levels remain low during larval periods and then increase during pupation.

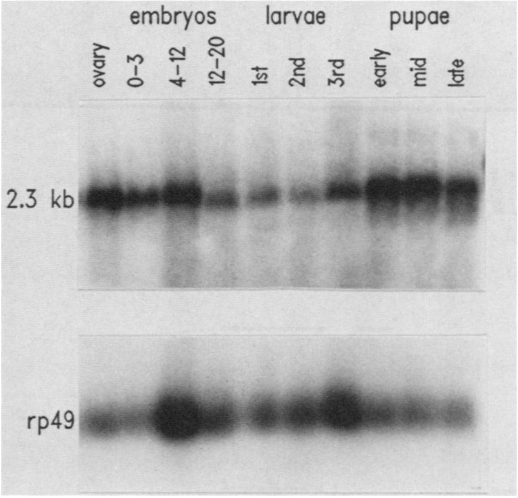

FIG. 2. Expression of *Hrb98DE* transcripts during development. Samples of 1 μg of poly(A)$^+$ RNA from the indicated stages (ovaries; 0- to 3-, 4- to 12-, or 12- to 20-h embryos; first-, second-, or third-instar larvae; and early, mid, or late pupae) were electrophoresed on a methylmercuric hydroxide gel and electrophoretically transferred to a nylon membrane. The blot was probed with a 550-nt fragment of the p9 clone (upper panel) which contains the RNP consensus sequences but not the glycine-encoding GGN repeats, then stripped, and reprobed (lower panel) with a fragment from the *rp49* locus (34). RNA markers were used as size standards.

The p9 clone insert is only 1.5 kb long and was presumed to be missing about 700 nucleotides (nt), depending on the length of the poly(A) tail, since the transcript seen on the RNA blot is 2.3 kb. Ovarian and embryonic cDNA libraries were screened to isolate longer clones; a genomic clone covering the locus was also isolated (see Materials and Methods for details of the libraries screened). Restriction analysis of the new cDNA clones indicated that they differed from the original p9 clone at the 5′ and/or 3′ ends. Two of the clones were sequenced completely, and several others were sequenced in the regions of divergence. Comparison of the nucleotide sequences (see below) indicated that these clones did indeed diverge significantly at the 5′ and 3′ ends. Since some of the cDNA clones were derived from the Canton S library and some were derived from the Oregon R libraries, it was possible that the sequence divergence was due to differences between *Drosophila* strains. Alternatively, the cDNA clones could simply represent differentially spliced transcripts. To resolve this question, splice junction sites were localized on the cDNA and genomic DNA. Initial localization was done by comparison of the restriction maps of the clones, and the exact boundaries were determined by sequencing the genomic DNA in the relevant regions. In

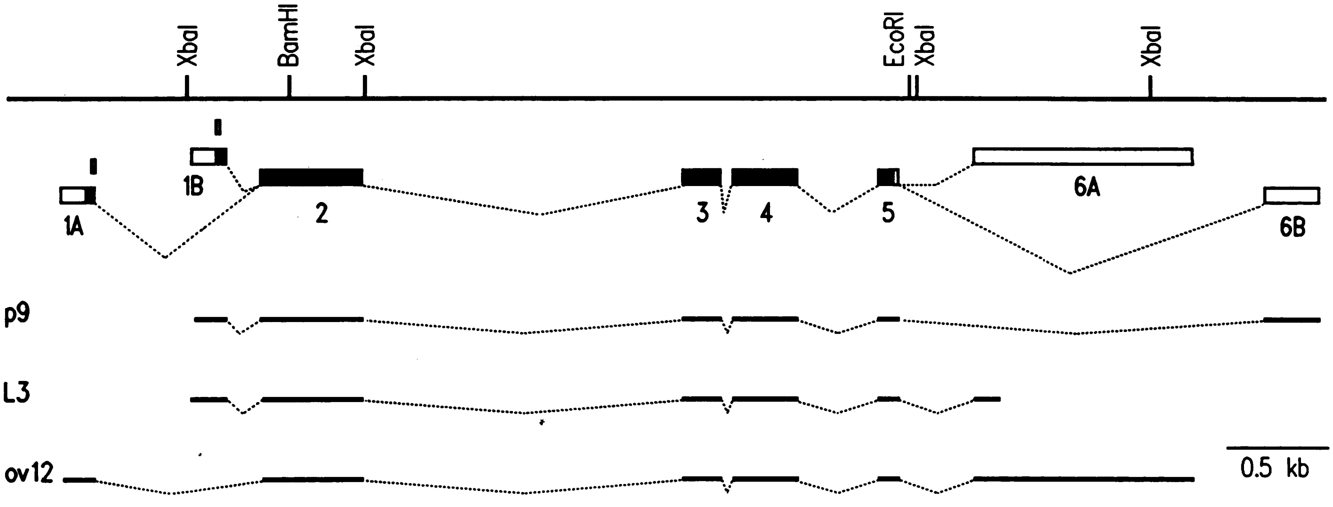

FIG. 3. Map of *Hrb98DE* genomic and cDNA clones. The top line shows a portion of a genomic clone isolated from the library of Maniatis et al. (31). Boxes below represent the exons of the *Hrb98DE* transcripts. Solid boxes represent the coding sequences. Transcription is from left (proximal) to right (distal). Small rectangles above exons 1A and 1B indicate the locations of the primers used for the experiment described for Fig. 5. The three cDNA clones that were completely sequenced are shown below. The L3 clone is truncated at its 3′ end.

cases in which the exons did not contain restriction enzyme sites useful for unambiguous localization on the genomic DNA, the PCR was used to map the exons. Oligonucleotide primers corresponding to the adjacent exons were used in reactions with either the cDNA clone or the genomic DNA clone. The increase in size of the band amplified from the genomic DNA compared with that from the cDNA corresponds to the length of the intron separating the two exons. The results of these analyses are shown in Fig. 3. All of the sequences in the cDNA clones could be located on the Canton S genomic DNA clone, indicating that the differences are due to the use of alternative exons and not to the presence of different exons in different strains. Each transcript contains six exons; exons 2 to 5 are common to all the species, but there is a choice of two 5′ and two 3′ exons. In addition, sequencing of the cDNA clones indicated that exon 2 has two alternative splice acceptor sites at its 5′ end that are 12 nt apart, leading to further variability in the structure of the transcripts. Because exons 1A and 1B are approximately the same size, transcripts containing either of these exons and exon 6A will all be about 2,250 nt and thus not separable on RNA blots. However, exon 6B is considerably shorter than 6A (314 versus 1,035 nt), and therefore it was surprising that only a single band was seen in Fig. 2. Hybridization of RNA blots with a probe specific for exon 6B showed that transcripts containing this exon are rare in comparison with those containing exon 6A and are not readily detectable at the exposure times used for the blot in Fig. 2 (data not shown). These transcripts are clearly present early in development and during pupation and may be present at other stages as well. They are about 1.6 kb in size, and thus cDNA clone p9 represents a nearly full-length clone and includes a complete coding region.

There are eight different transcripts that could be generated by combinations of alternative exons and splice acceptor sites, and we have isolated cDNA clones corresponding to four of these species. To determine whether the other species are also produced and whether production of the various species is developmentally regulated, the following experiments were done. Usage of exons 1A and 1B and the alternative splice acceptor sites was assayed by a combination of RT and PCR analyses (Fig. 4). Poly(A)$^+$ RNA from the stages of development shown in Fig. 2 or from males or ovariectomized females was copied by reverse transcriptase using a primer located 3′ of the alternative splice acceptor sites in exon 2 (the RT primer). The reaction products were divided in half, and each was incubated in a PCR reaction with the RT primer and a primer specific for exon 1A or 1B.

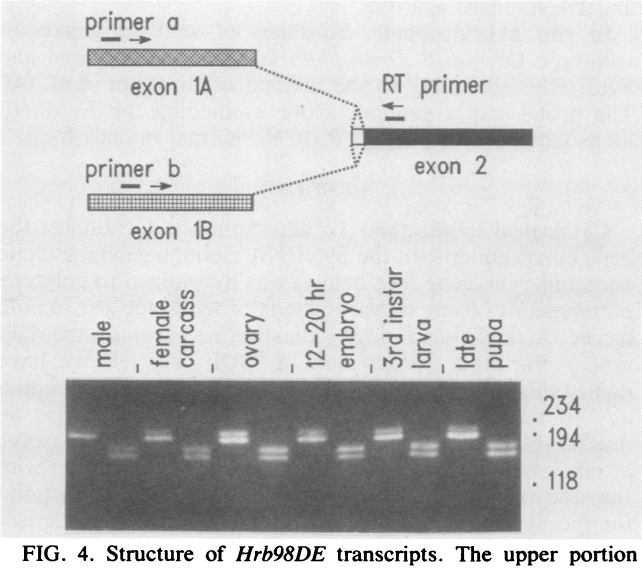

FIG. 4. Structure of *Hrb98DE* transcripts. The upper portion shows a schematic diagram of the experiment. cDNA was synthesized from 1 μg of poly(A)$^+$ RNA from different developmental stages using reverse transcriptase and a primer specific for exon 2 (RT primer). Each sample was split in half and incubated with the RT primer and either primer a or primer b, and a standard PCR reaction was performed. The amplified material was electrophoresed on a 4% Nusieve–1% standard agarose gel (lower panel). The positions of some of the bands from a *Hae*III digest of φX174 DNA are shown at right. The female carcass sample consisted of ovariectomized females. The analysis was done for all the developmental stages shown in Fig. 2, with similar results from all samples. Only representative ones are shown here.

The sizes of the expected products were 200 and 188 nt for exon 1A and 170 and 158 nt for exon 1B; some of the results are shown in the lower part of Fig. 4. All of the RNA samples tested generated bands of the expected sizes with each exon-specific primer, indicating that both exons 1A and 1B are used at all stages of development. However, usage of the alternative splice acceptor sites was not uniform. Transcripts containing exon 1B showed approximately equal usage of the two splice acceptor sites for exon 2, but those containing exon 1A preferentially spliced to the 3′-most acceptor site (compare the intensities of the 200- and 188-nt bands, particularly in the male and late pupae sample lanes). Thus the choice of the 5′ exon may play an important role in the relative utilization of the splice acceptor sites, as has been suggested for other systems (39).

Because transcripts containing exon 6A are much more abundant than those containing exon 6B, we assume that they contribute the majority of the material amplified in the PCR experiment of Fig. 4. Similar experiments using an exon 6B-specific primer were done to characterize the 6B-containing species present in pupal RNA (data not shown). The results showed that both exons 1A and 1B are used in conjunction with exon 6B. To examine the alternative splice acceptor sites used, a small portion of the amplified material was reamplified by using the primers described in Fig. 4. Both alternative splice sites are used with each 5′ exon. Taken together, these experiments demonstrate that all possible combinations of exons 1A, 1B, 6A, and 6B and the alternative splice acceptor sites of exon 2 are in fact utilized, although not at the same frequency, generating eight different transcripts from the *Hrb98DE* locus.

Analysis of the exon structures of the *Hrb98DE* transcripts suggested that they might be derived from two promoters but did not rule out the existence of a common leader exon further upstream. The transcription start sites for exons 1A and 1B were determined by primer extension and S1 protection analyses using exon-specific primers; the results are shown in Fig. 5 and suggest that *Hrb98DE* transcripts initiate from two promoter regions about 600 nt apart. For exon 1A, there is a single start site 163 nt from the 5′ end of the primer. (The locations of the primers used are shown in Fig. 3.) The situation is more complicated for exon 1B. Four major start sites are seen in the region 130 to 110 nt from the primer, and additional minor start sites are visible further upstream on longer exposure of the gel. Thus, heterogeneous 5′ termini were found for transcripts containing exon 1B. The same multiple initiation sites were detected by both S1 mapping and primer extension and with several nonoverlapping primers and different RNA samples (data not shown), rendering it unlikely that this result is an artifact. The region 30 nt upstream of the exon 1A initiation site does not contain a TATA sequence; there is a TTTAAT sequence 5′ of exon 1B, but if it functions as a TATA box it does not specify a unique start site. Other non-TATA elements have been found to specify initiation sites (44), and the promoters in the *Hrb98DE* gene may be of this type. The sequences surrounding the transcription start sites are shown in the lower part of Fig. 5. These sequences are similar to each other and are related to the *Drosophila* consensus sequence for transcription initiation derived by Hultmark et al. (25). In most of the sequences used to derive the consensus, the initiation site is the first A residue. However, there are several transcription initiation sites in which the first A of the consensus sequence is missing, and initiation occurs at the second A (7, 25). This is the situation in all of the start sites mapped for the *Hrb98DE* transcripts.

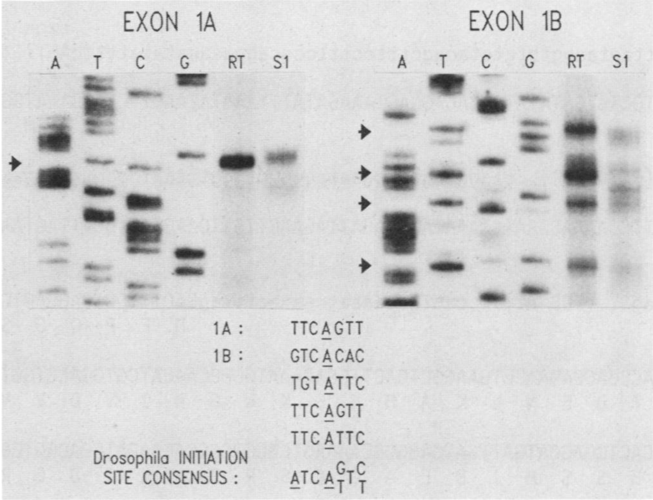

FIG. 5. Mapping transcription initiation sites. The transcription initiation sites for exons 1A and 1B were mapped by RT and S1 protection analyses, using the exon-specific primers indicated on the map in Fig. 3. The S1 probe, RT reaction, and sequencing reaction used the same primers. The sequence (leftmost four lanes of each panel) is that of the antisense strand. Arrows indicate major initiation sites. The sequences surrounding the initiation sites and the *Drosophila* consensus sequence are shown at the bottom of the figure.

Sequence and protein isoforms. The nucleotide sequences of the exons and portions of the introns near the splice junctions and the derived protein sequences are shown in Fig. 6. The major transcription initiation sites and a polyadenylation signal in exon 6B are underlined. Translation probably begins at the first ATG in exon 1A or 1B, both of which are in good contexts for *Drosophila* translation initiation codons (9). The termination codon (* in Fig. 6) is found near the end of exon 5. The final exon is completely noncoding, an unusual feature found in few genes (21), and therefore the choice of exon 6A or 6B does not influence the protein sequence. The use of alternative exon 1A or 1B and the alternative splice acceptor sites (Fig. 6, arrowheads) can result in the production of four protein isoforms that differ at their N termini. The N termini are relatively rich in asparagine, a characteristic shared with portions of the glycine-rich domain. The RNP consensus sequence domain is common to all isoforms, and thus all are likely to be components of hnRNP complexes. The isoforms are predicted to be approximately the same molecular mass (ranging from 38.5 to 39 kDa) but differ in calculated pI because of differences in the content of acidic amino acids. These structural differences in the isoforms may reflect functional distinctions between the proteins (see Discussion).

DISCUSSION

Transcripts from the *Drosophila Hrb98DE* locus encode proteins that are very homologous to the mammalian A1 hnRNP protein and therefore are likely to be components of *Drosophila* hnRNP complexes (23). Eight transcripts, which can generate four protein isoforms, are produced by the use of alternative exons and splice acceptor sites. Although some of the transcripts are relatively rare, transcripts encoding each of the protein isoforms are present at all stages of development. Unless production of the proteins is regulated at the level of translation, this suggests that the different

```
                                                exon 1A
ttagtaaggtctgtctacggctttcctttccacaggaaaaatatattttcAGTTTTAGGGAAGGGGTGCTACAGTGAGCGTCTTTCGTTCCCAGTGTCGTTATTTCTATAGTAT

TGCTGAGATATATATCAGAGCAGTAAAGATATTTAAATATAAGTTCTTCGAAATGGGTGGTCACGACAACTGGAACAATGGTCAAAATGAGGAGCAAGATgtaagtag......
                                                     M  G  G  H  D  N  W  N  N  G  Q  N  E  E  Q  D
                                                            exon 1B
(~410 nt)...ctaggtagtatcgaataggtgctgggtttaatcgggttgtagaaacggtcACACTGTATTCAGTTTTTTCATTCGTCTCGTGTGTCAGGGGACTTTCTCCAT

TTGCAAAAGCAAAAAAAATAGTTAGATTAGAAGTTGTTCAACTTTTGCATTAGTAAAATGGTGAACTCGAACCAGAACCAGAACGGCAACTCCAATGGCCATGATGATgtaagt
                                                          M  V  N  S  N  Q  N  Q  N  G  N  S  N  G  H  D  D
                                      ▼  exon 2   ▼
ag...(106 nt)...cattgataaaatcaataatttagGACTTTCCTCAGGACTCCATCACCGAGCCGGAGCATATGCGCAAGCTGTTCATCGGTGGCTTGGACTACCGTACC
                                        D  F  P  Q  D  S  I  T  E  P  E  H  M  R  K |L  F  I  G  G  L| D  Y  R  T

ACCGACGAGAACCTGAAGGCTCACTTCGAGAAATGGGGCAACATCGTGGACGTGGTGGTCATGAAGGATCCGCGCACAAAGCGTTCCCGCGGATTCGGATTCATCACCTATTCC
 T  D  E  N  L  K  A  H  F  E  K  W  G  N  I  V  D  V  V  V  M  K  D  P  R  T  K  R  S |R  G  F  G  F  I  T  Y| S

CACTCGAGCATGATTGATGAGGCGCAAAAGTCGCGTCCCCACAAGATCGACGGTCGAGTGGTGGAGCCCAAGCGCGCCGTTCCCCGTCAGGACATTGATTCCCCGAATGCCGGA
 H  S  S  M  I  D  E  A  Q  K  S  R  P  H  K  I  D  G  R  V  V  E  P  K  R  A  V  P  R  Q  D  I  D  S  P  N  A  G

GCTACCGTAAAGAAGCTCTTTGTTGGCGCCCTCAAGGACGACCATGATGAGCAGAGCATCCGCGACTACTTCCAGCACTTTGGCAACATCGTCGACATCAACATCGTCATCGAC
 A  T  V  K  K |L  F  V  G  A  L| K  D  D  H  D  E  Q  S  I  R  D  Y  F  Q  H  F  G  N  I  V  D  I  N  I  V  I  D
                                                                                                            exon 3
AAGGAGACTGGCAAGAAACGCGGATTCGCCTTCGTCGAGTTCGACGACTACGATCCCGTGGACAAAGTTGTGTgtaagtac...(~1.6 kb)...tctttattttgcagTGC
 K  E  T  G  K  K |R  G  F  A  F  V  E  F| D  D  Y  D  P  V  D  K  V  V                                         L

AAAAGCAGCATCAGCTGAATGGCAAAATGGTGGACGTAAAGAAGGCCTTGCCCAAGCAGAATGACCAACAGGGAGGTGGCGGAGGACGCGGTGGTCCGGGAGGTCGTGCCGGTG
Q  K  Q  H  Q  L  N  G  K  M  V  D  V  K  K  A  L  P  K  Q  N  D  Q  Q  G  G  G  G  G  R  G  G  P  G  G  R  A  G
                                                                                          exon 4
GAAACCGCGGAAACATGGGCGGTGGAAACTACGGCAACCAGAATGGTGGCGGCAACTGGgtaagaac...(41 nt)...taattggcattgcagAACAACGGTGGCAACAACT
G  N  R  G  N  M  G  G  G  N  Y  G  N  Q  N  G  G  G  N  W                                     N  N  G  G  N  N

GGGGCAACAACCGCGGGGGTAACGACAACTGGGGCAACAACAGCTTCGGTGGTGGCGGCGGCGGCGGTGGTGGCTATGGCGGTGGCAACAACAGCTGGGGCAATAACAATCCGT
W  G  N  N  R  G  G  N  D  N  W  G  N  N  S  F  G  G  G  G  G  G  G  G  Y  G  G  G  N  N  S  W  G  N  N  N  P

GGGACAATGGCAATGGAGGCGGCAACTTTGGAGGCGGCGGCAACAATTGGAACAATGGTGGCAATGATTTTGGAGGCTACCAGCAGAACTATGGCGGCGGTCCGCAGCGAGGTG
W  D  N  G  N  G  G  G  N  F  G  G  G  G  N  N  W  N  N  G  G  N  D  F  G  G  Y  Q  Q  N  Y  G  G  G  P  Q  R  G
                                                                                                    exon 5
GCGGCAACTTCAACAACAATCGCATGCAACCCTACCAAGGAGGTGGTGGATTCAAAGCAGgtgagcaa...(~370 nt)...tgggtttcttttagGCGGTGGCAATCAAGG
G  G  N  F  N  N  N  R  M  Q  P  Y  Q  G  G  G  G  F  K  A                                       G  G  G  N  Q  G

CAACTATGGCGGAAACAATCAGGGCTTCAATAACGGTGGCAACAACCGCAGATATTGAGAGAGATTCCGGACGAGgtgagaca...(~340 nt)....atcgacattttatag
  N  Y  G  G  N  N  Q  G  F  N  N  G  G  N  N  R  R  Y  *
exon 6A
GTGATAAAGTAAAAGTCGCTCAGTATAATCCCAAAATTCAAAGTCAAGTAAACGATACGCAATTCAATGATCAATATGATGAAGTATGAACGCACGTTCATAGTCCTAAGAAGT

TTTCCGAGCACCACAAACCAAACATTACAGATTGATTGATCAGGATCAGTGCAGTGCAGTGAAGTCGTAAGAAGTTTGGTTCAGTCTTGGATTAGAACAGAAAGAGAGATCAGT

ACCAAGCAAGACCAGACCAATACAGAGTGCAGTGCAGTGCAGAGCAGTGTGTGAGCTAAACAGAAGAAACACAGCAAGTGGCAGTACACGTACACAGAAGGTGACCAGGATTCC

ATTCAATTTGTCTTTGCAAGCGATGATAAATTGAATGTCAAGAGGCGCATCGAAAACTATCCAATACCCACAGAATTGTAATCTATAATTTGCAACCCATATGTAGCAGCCCAT

ATAAGTAACATTGTAATTAGATCCACAAAGCCGACGACAGGCCAAACATTTGTATCTCCACCACACGAGCAACCAACCAACCAACCTACTTAAAACAACCAGTCATAAACGAAT

CAGCTATGTAAAACTTCAAGGCGGCAGCCGAGAGATATGTTTCTTTGGTTGCAGACAGCTCTCCAGTCCATCGGCTGCCGGCAACTTTTGATAATCAGAGCAGAAAACACAAAA

TATATAAAAACATGTATTACATTTTACAACTTTACTAGGCTAAGTTCTCATAGAATTAAATCGAGTAGAGCGTGCGGAGCGACTTCTGCATCCATCCGCACGCTACCCTAGTCT

TAATAATTTTACTATTCAAGTTCTAAGCGAACCTGACCGTGTGTGGAGGAATCTAGATACTTAGATTTTTGAATAAGTGTCGACATGGATATAACCTCTTCCGCTGCTTAGATG

TAACCGATCAACAAACAACCTTTATTGTAATTTATACCTTTTAGCAAGCAGTCTGTCTGAGGAGTCAGCGCAATTATACTTAACAAAATTTACATTTTAATGTAGCTAAAAAAT
                             exon 6B
ACAAGATGCA...(~300 nt)...aatttctgttgccagTTGGCTGCGAAAAGTATGCGAAGCGCTGAAAGACTGCCGCCGCGCTACTGAATGCAGCCACGGAAGGTTCCCGC

GTTTGCTAGCAACTCCGGTCGCTGGGTTTCGGTCCCTCTGGGTAAGTAATCCTCCGCAGATCCCAAATCCGGAAAAAGTTGTAACTTCAGGTGGTCTGAATCGGCTCGTCGCAC

CTGAGGATGGAATGGATGTTGAACGGTTTCCCCTCTGCACCATCGTAAATATACGACTTTTATTTATTACTTCATTTTATTTTCATTATTACTATCTTTGAATAAAGAAATTAC

AGAATACCAAACA
```

protein forms are not stage specific, but it does not rule out the possibility of tissue-specific distribution. We do not know whether all of the possible protein isoforms are in fact produced, but preliminary evidence indicates that at least some of them are. Antibodies generated to a fusion protein containing the two RNP consensus sequence domains of the *Hrb98DE* protein recognize several *Drosophila* proteins in the predicted 38 to 40 kDa range. Preliminary experiments also indicate that these proteins are found in association with newly synthesized RNA (unpublished data).

The mammalian A1 protein is related to other hnRNP proteins (A2 and the B group proteins) antigenically and structurally (29, 30), but currently the complete sequence of only the A1 protein is available. Therefore, we do not know which of these proteins is most closely related to the *Hrb98DE* proteins. However, there are some striking similarities between the structures of the human A1 gene and the *Hrb98DE* gene. Comparison of the splicing patterns of the human A1 (3) and *Hrb98DE* transcripts revealed similar locations of some of the splice sites. In both cases, the first exon contained the translation initiation codon and a short stretch of coding sequence which is separated from the domain with the RNP consensus sequences by an intron. The final exon is completely noncoding and is preceded by a short exon containing the termination codon. Internally, the splices between *Hrb98DE* exons 2 and 3 and human A1 exons 4 and 5 occur at the same site relative to the protein sequence. The locations of the remaining splice sites differ. In the human A1 gene, an intron separates the RNP consensus sequence domain from the glycine-rich domain; in *Hrb98DE*, exon 3 contains the end of the RNP consensus sequence domain and the beginning of the glycine-rich domain. Also, in the *Drosophila* gene, the majority of the RNP consensus domain is found in a single exon, whereas in the human A1 gene it is split into four exons. In both genes, however, the RNP consensus sequences themselves are not interrupted by introns, in contrast to the situation for the mouse nucleolin (6) and human SS-B/La (10) RNA-binding protein genes.

Analysis of mammalian hnRNP proteins by two-dimensional gel electrophoresis reveals multiple protein spots of the same size as the A1 protein which may be related to each other by slight differences in charge or molecular weight (30, 35). It is not known whether some of these proteins represent different posttranslational modifications of a single protein or different primary sequences. In the case of the human A1 protein, cDNA sequencing has identified two transcripts encoding proteins that differ by two conservative amino acid substitutions (8). Consistent with this, the human A1 gene family has about 30 members, at least two of which are probably active genes (3). Also, there is some evidence for alternative splicing of one of the active genes; an additional exon encoding a glycine-rich sequence might be spliced into the transcript (3). In contrast, the *Hrb98DE* gene is probably a single-copy gene, and the diversity is generated by alternative splicing. Here the differences are confined to the extreme N-terminal 16 to 21 amino acids, which have no homology to the mammalian A1 protein. Examination of the sequences immediately upstream of exon 2 suggests a simple explanation for the use of two alternative splice acceptor sites. In general, such regions are characterized by a branchpoint consensus sequence ($\genfrac{}{}{0pt}{}{C}{T}T\genfrac{}{}{0pt}{}{A}{G}A\genfrac{}{}{0pt}{}{T}{C}$ for *Drosophila* genes [26]), which typically occurs 18 to 35 nt upstream of the splice acceptor site consensus sequence, $Y_{11}NYAG$ (33). Not all *Drosophila* splice acceptor sites are pyrimidine rich, but almost all lack AG dinucleotides between the branchpoint and the splice acceptor site AG (26, 43), as do mammalian introns. For the *Hrb98DE* transcripts, the 5'-most acceptor site of exon 2 is the first AG downstream of a sequence (TTGAT [dashed underline in Fig. 6]) that is a good fit to the branchpoint consensus. However, it is not preceded by a pyrimidine-rich region and is only 18 nt away from the branchpoint A residue, a distance which has been shown to be acceptable but suboptimal for splicing of simian virus 40 small t-antigen RNA (20). The 3'-most acceptor splice site is located at a more favorable distance from the branchpoint consensus (30 nt) and is preceded by a fairly pyrimidine-rich stretch, which however is interrupted by an AG dinucleotide (i.e., the alternative splice acceptor site). The splicing machinery is thus faced with two alternative and closely spaced splice acceptor sites, neither of which is optimal, but both of which are acceptable to the splicing apparatus and preserve the reading frame for the rest of the mRNA.

The *Hrb98DE* proteins are probably basic housekeeping components of the cellular machinery, yet transcripts from the *Hrb98DE* locus are not present at uniform levels during development. The transcripts in 0- to 3-h embryos presumably represent maternal RNA, since *Hrb98DE* transcripts are abundant in the ovary. The RNA levels are reduced in older embryos and larvae and then rise again late in development. Transcripts from the rat A1 gene have also been shown to vary in abundance. In this case, mRNA levels show a rapid increase shortly after stimulation of quiescent cells with growth factors or serum (36). We do not know whether the *Hrb98DE* transcripts respond similarly to mitogenic stimulation. However, if protein levels at each stage correspond to transcript levels, these results suggest that less protein may be required at certain stages. Alternatively, the maternal transcripts may generate enough protein to last until pupation, when higher transcript levels are required again.

The *Hrb98DE* protein isoforms could have different tissue distributions of have somewhat different functions in hnRNP complexes. Each isoform has a slightly different charge because of different numbers of acidic amino acids. This could produce differences in affinity for RNA or for other proteins, leading to differences in structure or function of the complex. Recently, the U1 RNA-binding domain of the 70K U1 small nuclear RNP protein has been found to include and

FIG. 6. Sequences of the *Hrb98DE* transcripts and proteins. Intron and upstream sequences are given in lowercase letters; exon sequences are in capital letters, and each exon is numbered (above the 5' end) according to the numbers in Fig. 3. The approximate lengths of the remaining intron sequences are shown. Transcription initiation sites and a polyadenylation signal in exon 6B are underlined; exon 6A does not have an AATAAA sequence, but there are sequences related to some of the variant polyadenylation signals that have been described previously (5). A possible branchpoint sequence upstream of exon 2 is indicated (-----). The positions of the alternative splice acceptor sites in exon 2 are indicated by arrowheads; clones L3 and ov12 use the downstream site, and clone p9 uses the upstream site. The previously reported p9 sequence (23) contains 69 nt at the 5' end that are not found in genomic or other cDNA clones. This sequence is found, in inverted orientation, in the middle of the p9 sequence, and its presence at the 5' end of the clone is probably a cloning artifact. There are eight single-base changes between clones derived from Canton S and Oregon R cDNA libraries, but none affects the protein sequence. RNP consensus sequences are boxed. The termination codon is indicated (*). The nucleotide sequence data reported in this paper have been submitted to the EMBL, GenBank, and DDBJ Nucleotide Sequence Databases and assigned the accession numbers M15766, M28870, M28871, and M28872.

extend beyond the RNP consensus sequence domain of that protein (38). Therefore, portions of the protein outside the RNP consensus sequences can potentially be important for specific binding. Determining whether this is also true for the *Hrb98DE* proteins will require the use of genetic, immunological, and biochemical approaches to characterize the structure and function of *Drosophila* hnRNP complexes.

ACKNOWLEDGMENTS

We thank Detlef Weigel for sending clones form the *fkh* region, Paul Adler for help with in situ hybridization techniques, Igor Dawid for support and encouragement, and Peter Good for comments on the manuscript.

This work was supported in part by Public Health Service grant GM39271 to A.L.B. A.L.B. is the recipient of an American Cancer Society Faculty Research Award.

LITERATURE CITED

1. **Adam, S. A., T. Nakagawa, M. S. Swanson, T. K. Woodruff, and G. Dreyfuss.** 1986. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. Mol. Cell. Biol. **6:**2932–2943.
2. **Beyer, A. L., M. E. Christensen, B. W. Walker, and W. M. LeStourgeon.** 1977. Identification and characterization of the packaging proteins of core 40S hnRNP particles. Cell **11:** 127–138.
3. **Biamonti, G., M. Buvoli, M. T. Bassi, C. Morandi, F. Cobianchi, and S. Riva.** 1989. Isolation of an active gene encoding human hnRNP protein A1. J. Mol. Biol. **207:**491–503.
4. **Bingham, P. M., R. Levis, and G. M. Rubin.** 1981. Cloning of DNA sequences from the white locus of *Drosophila melanogaster* by a novel and general method. Cell **25:**693–704.
5. **Birnstiel, M. L., M. Busslinger, and K. Strub.** 1985. Transcription termination and 3′ processing: the end is in site! Cell **41:**349–359.
6. **Bourbon, H.-M., B. Lapeyre, and F. Amalric.** 1988. Structure of the mouse nucleolin gene. J. Mol. Biol. **200:**627–638.
7. **Bowtell, D. D. L., M. A. Simon, and G. M. Rubin.** 1988. Nucleotide sequence and structure of the *sevenless* gene of *Drosophila melanogaster*. Genes Dev. **2:**620–634.
8. **Buvoli, M., G. Biamonti, P. Tsoulfas, M. T. Bassi, A. Ghetti, S. Riva, and C. Morandi.** 1988. cDNA cloning of human hnRNP protein A1 reveals the existence of multiple mRNA isoforms. Nucleic Acids Res. **16:**3751–3770.
9. **Cavener, D. R.** 1987. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. Nucleic Acids Res. **15:**1353–1361.
10. **Chambers, J. C., D. Kenan, B. J. Martin, and J. D. Keene.** 1988. Genomic structure and amino acid sequence domains of the human La autoantigen. J. Biol. Chem. **263:**18043–18051.
11. **Choi, Y. D., and G. Dreyfuss.** 1984. Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): a unique supramolecular assembly. Proc. Natl. Acad. Sci. USA **81:**7471–7475.
12. **Chung, S. Y., and J. Wooley.** 1986. Set of novel, conserved proteins fold pre-messenger RNA into ribonucleosomes. Proteins Struct. Funct. Genet. **1:**195–210.
13. **Cobianchi, F., R. L. Karpel, K. R. Williams, V. Notario, and S. H. Wilson.** 1988. Mammalian heterogeneous nuclear ribonucleoprotein complex protein A1. J. Biol. Chem. **263:**1063–1071.
14. **Cobianchi, F., D. N. SenGupta, B. Z. Zmudzka, and S. H. Wilson.** 1986. Structure of rodent helix-destabilizing protein revealed by cDNA cloning. J. Biol. Chem. **261:**3536–3543.
15. **Digan, M. E., S. R. Haynes, B. A. Mozer, I. B. Dawid, F. Forquignon, and M. Gans.** 1986. Genetic and molecular analysis of *fs(1)h*, a maternal effect homeotic gene in *Drosophila*. Dev. Biol. **114:**161–169.
16. **Dreyfuss, G.** 1986. Structure and function of nuclear and cytoplasmic ribonucleoprotein particles. Annu. Rev. Cell Biol. **2:** 459–498.
17. **Dreyfuss, G., Y. D. Choi, and S. Adam.** 1984. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. Mol. Cell. Biol. **4:**1104–1114.
18. **Dreyfuss, G., M. S. Swanson, and S. Piñol-Roma.** 1988. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. Trends Biochem. Sci. **13:**86–91.
19. **Economidis, I. V., and T. Pederson.** 1983. Structure of nuclear ribonucleoprotein: heterogeneous nuclear RNA is complexed with a major sextet of proteins in vivo. Proc. Natl. Acad. Sci. USA **80:**1599–1602.
20. **Fu, X.-Y., J. D. Colgan, and J. L. Manley.** 1988. Multiple *cis*-acting sequence elements are required for efficient splicing of simian virus 40 small-t antigen pre-mRNA. Mol. Cell. Biol. **8:**3582–3590.
21. **Hawkins, J. D.** 1988. A survey on intron and exon lengths. Nucleic Acids Res. **16:**9893–9908.
22. **Haynes, S. R., B. A. Mozer, N. Bhatia-Dey, and I. B. Dawid.** 1989. The *Drosophila fsh* locus, a maternal effect homeotic gene, encodes apparent membrane proteins. Dev. Biol. **134:** 246–257.
23. **Haynes, S. R., M. L. Rebbert, B. A. Mozer, F. Forquignon, and I. B. Dawid.** 1987. *pen* repeat sequences are GGN clusters and encode a glycine-rich domain in a *Drosophila* cDNA homologous to the rat helix destabilizing protein. Proc. Natl. Acad. Sci. USA **84:**1819–1823.
24. **Hogness, D. S., H. D. Lipshitz, P. A. Beachy, D. A. Peattie, R. B. Saint, M. Goldschmidt-Clermont, P. J. Harte, E. R. Gavis, and S. L. Helfand.** 1985. Regulation and products of the *Ubx* domain of the bithorax complex. Cold Spring Harbor Symp. Quant. Biol. **50:**181–194.
25. **Hultmark, D., R. Klemenz, and W. Gehring.** 1986. Translational and transcriptional control elements in the untranslated leader of the heat-shock gene *hsp-22*. Cell **44:**429–438.
26. **Keller, E. B., and W. A. Noon.** 1985. Intron splicing: a conserved internal signal in introns of *Drosophila* pre-mRNAs. Nucleic Acids Res. **13:**4971–4981.
27. **Knust, E., U. Dietrich, U. Tepass, K. A. Bremer, D. Weigel, H. Vässin, and J. A. Campos-Ortega.** 1987. EGF homologous sequences encoded in the genome of *Drosophila melanogaster*, and their relation to neurogenic genes. EMBO J. **6:**761–766.
28. **Kumar, A., K. R. Williams, and W. Szer.** 1986. Purification and domain structure of core hnRNP proteins A1 and A2 and their relationship to single-stranded DNA-binding proteins. J. Biol. Chem. **261:**11266–11273.
29. **Leser, G. P., J. Escara-Wilke, and T. E. Martin.** 1984. Monoclonal antibodies to heterogeneous nuclear RNA-protein complexes. J. Biol. Chem. **259:**1827–1833.
30. **Leser, G. P., and T. E. Martin.** 1987. Changes in heterogeneous nuclear RNP core polypeptide complements during the cell cycle. J. Cell Biol. **105:**2083–2094.
31. **Maniatis, T., R. C. Hardison, E. Lacy, J. Lauer, C. O'Connell, G. K. Sim, and E. Efstratiadis.** 1978. The isolation of structural genes from libraries of eucaryotic DNA. Cell **15:**687–701.
32. **Merrill, B. M., K. L. Stone, F. Cobianchi, S. H. Wilson, and K. R. Williams.** 1988. Phenylalanines that are conserved among several RNA-binding proteins form part of a nucleic acid-binding pocket in the A1 heterogeneous nuclear ribonucleoprotein. J. Biol. Chem. **263:**3307–3313.
33. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10:**459–472.
34. **O'Connell, P., and M. Rosbash.** 1984. Sequence, structure, and codon preference of the *Drosophila* ribosomal protein 49 gene. Nucleic Acids Res. **12:**5495–5513.
35. **Piñol-Roma, S., Y. D. Choi, M. Matunis, and G. Dreyfuss.** 1988. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. Genes Dev. **2:**215–227.
36. **Planck, S. R., M. D. Listerud, and S. D. Buckley.** 1988. Modulation of hnRNP A1 protein gene expression by epidermal growth factor in Rat-1 cells. Nucleic Acids Res. **16:**11663–11673.
37. **Poole, S. J., L. M. Kauvar, B. Drees, and T. Kornberg.** 1985. The *engrailed* locus of *Drosophila*: structural analysis of an embryonic transcript. Cell **40:**37–43.

38. **Query, C. C., R. C. Bentley, and J. D. Keene.** 1989. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell **57:**89–101.

39. **Reed, R., and T. Maniatis.** 1986. A role for exon sequences and splice-site proximity in splice-site selection. Cell **46:**681–690.

40. **Risau, W., P. Symmons, H. Saumweber, and M. Frasch.** 1983. Nonpackaging and packaging proteins of hnRNA in *Drosophila melanogaster*. Cell **33:**529–541.

41. **Riva, S., C. Morandi, P. Tsoulfas, M. Pandolfo, G. Biamonti, B. Merrill, K. R. Williams, G. Multhaup, K. Beyreuther, H. Werr, B. Henrich, and K. P. Schafer.** 1986. Mammalian single-stranded DNA binding protein UP1 is derived from the hnRNP core protein A1. EMBO J. **5:**2267–2273.

42. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.

43. **Shapiro, M. B., and P. Senapathy.** 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. **15:**7155–7174.

44. **Smale, S. T., and D. Baltimore.** 1989. The "initiator" as a transcription control element. Cell **57:**103–113.

45. **Weigel, D., G. Jürgens, F. Küttner, E. Seifert, and H. Jäckle.** 1989. The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the *Drosophila* embryo. Cell **57:**645–658.

46. **Zachar, Z., T.-B. Chou, and P. M. Bingham.** 1987. Evidence that a regulatory gene autoregulates splicing of its transcript. EMBO J. **6:**4105–4111.