Nucleic Acids Research, 1998, Vol. 26, No. 7 **1597–1604**

# *Drosophila* clipper/CPSF 30K is a post-transcriptionally regulated nuclear protein that binds RNA containing GC clusters

**Chunyang Bai and Peter P. Tolias***

Public Health Research Institute, 455 First Avenue, New York, NY 10016, USA

Received January 5, 1998; Revised and Accepted February 11, 1998

DDBJ/EMBL/GenBank accession no. AF033201

## ABSTRACT

**An essential component of the mammalian pre-mRNA 3′-end processing machinery is a multimeric protein complex known as cleavage and polyadenylation specificity factor (CPSF). The *Drosophila melanogaster* gene, *clipper* (*clp*), encodes a homolog of the CPSF 30K subunit. We have shown previously that CLP possesses N-terminal endoribonucleolytic activity and that the relative expression of its mRNA fluctuates during fly development. In the present study, we report that CLP's C-terminus, containing two CCHC zinc knuckles, confers a binding preference for RNAs that contain G- and/or C-rich clusters. We also show, for the first time, that a member of the highly conserved CPSF 30K family is a nuclear and developmentally regulated protein. Though *clp* transcripts are detectable throughout embryogenesis, CLP protein is not present. We demonstrate that post-transcriptional regulation of *clp* mRNA in the embryo occurs by a process that does not involve poly(A) tail length shortening. Thus, a key component of the pre-mRNA 3′-end processing machinery is subject to post-transcriptional regulation during development. These results support the existence of a distinct mechanism controlling eukaryotic gene expression through the regulated processing of pre-mRNAs in the nucleus.**

## INTRODUCTION

In eukaryotes, transcription of specific genes and formation of mature mRNAs are processes that occur in the nucleus. Prior to nuclear export, mRNAs require several post-transcriptional modifications, which are tightly coupled. These include 5′-end capping, intron splicing and 3′-end processing. The precise mechanism that couples these post-transcriptional events is not known. In spite of the fact that they are coupled, all three processes are subject to independent regulation.

Recognition and formation of mature mRNA 3′-ends involves endonucleolytic cleavage of the pre-mRNA followed by synthesis of a poly(A) tail (1,2). In mammals, at least two signals are required for pre-mRNA 3′-end processing: a nearly invariant AAUAAA sequence, referred to as the polyadenylation signal, and a much less conserved downstream GU-rich (or U-rich) element. The protein machinery that recognizes these sites performs a coupled two-step processing reaction (1,2). First, an endonucleolytic cleavage at the pre-mRNA cleavage site, usually located ~15 nucleotides (nt) downstream from the AAUAAA sequence, generates upstream (5′) and downstream (3′) cleavage products. This is followed by synthesis of a poly(A) tail which is added to the upstream cleavage product, and degradation of the downstream cleavage product. Five mammalian factors are required for the cleavage reaction (1,2). These include cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), two cleavage factors (CF I and CF II) and poly(A) polymerase (PAP). Efficient polyadenylation of the upstream cleavage product requires CPSF and PAP, as well as nuclear poly(A) binding protein (PAB II).

CPSF has been purified from HeLa cells and calf thymus and shown to consists of four subunits of 160, 100, 73 and 30 kDa that are required for both cleavage and polyadenylation (3–5). A recombinant version of the 160 kDa subunit can bind a pre-mRNA that contains AAUAAA, though with lower specificity compared with intact CPSF (6). UV-crosslinking experiments with CPSF fractions have indicated that proteins of 160 and 30 kDa interact with pre-mRNAs containing an AAUAAA element (7). Since only the 30 kDa component contains known RNA-binding motifs (5), it is thought that CPSF 30K cooperates with CPSF 160K to allow efficient binding of pre-mRNA targets.

CstF recognizes the downstream GU-rich (or U-rich) element and is only required for the cleavage reaction (1,2). It consists of a heterotrimeric protein complex composed of 77, 64 and 50 kDa subunits (8). CstF 64K directly binds the GU-rich (or U-rich) downstream element (9). The binding is apparently mediated by a RNP-type RNA binding motif (10), which resides in the N-terminal portion of the protein. This motif can sufficiently recognize RNA substrates that resemble the GU-rich (or U-rich) downstream element, thought to be one of the pre-mRNA 3′-end processing signals (11). CstF 77K is believed to physically bridge the 64 and 50K subunits (12) and also interacts specifically with CPSF (6). The 50K subunit is also required for CstF activity but its function is unknown.

CF I consists of three proteins of 68, 59 and 25 kDa and has been shown to bind preferentially RNAs that contain 3′-end processing signals (13). It can also stabilize CPSF–RNA complexes, presumably by interacting with CPSF. CF II has not been purified.

PAP is required to increase the efficiency of the cleavage reaction by interacting with CPSF 160K and stabilizing the RNA–CPSF complex (6). Since PAP binds RNA non-specifically

*To whom correspondence should be addressed. Tel: +1 212 578 0815; Fax: +1 212 578 0804; Email: tolias@phri.nyu.edu

with low affinity, the CPSF interaction also acts to tether PAP to the RNA. This is important because PAP is also required after cleavage for the polyadenylation reaction. CPSF and PAP initiate polyadenylation in a slow and distributive manner. The addition of at least 10 adenosine residues by these two factors provides a binding template for PAB II (14). When all three factors are engaged, this promotes the rapid and processive polymerization of the poly(A) tail.

Processing of pre-mRNA 3′-ends in yeast is clearly related to that described above for mammals. Differences include the recognition of more degenerate and redundant RNA target sequences and the generation of shorter poly(A) tails. In spite of these differences, the overall mechanism still employs a two step cleavage and polyadenylation reaction. Nine out of 14 known yeast genes whose products function in 3′-end processing are conserved and function in the same process in mammals (2).

Of all the factors required for efficient 3′-end processing of mammalian pre-mRNAs, CPSF plays the most diverse role. CPSF is a multifunctional protein complex that is required for both the cleavage and polyadenylation reactions. A bovine homolog of the CPSF 30K subunit was recently cloned and reported to contain five C-terminal putative CCCH zinc finger motifs and one C-terminal CCHC zinc knuckle (5). This protein displays significant animo acid sequence identity to *Drosophila* CLP, an endoribonuclease that cleaves RNA hairpins (15). We have shown that CLP's endonucleolytic activity resides in a region containing five copies of a CCCH zinc finger motif (15). Here, we demonstrate that CLP's C-terminus, containing two CCHC zinc knuckles, confers a binding preference for RNAs that contain G- and/or C-rich clusters. We also generated an affinity-purified polyclonal anti-CLP antibody and used it to reveal that CLP is a nuclear protein. However, CLP protein is not present in the embryo, even though *clp* transcripts are detectable throughout embryogenesis. This embryonic post-transcriptional regulation of *clp* mRNA occurs by a process that does not involve poly(A) tail length shortening. Our results suggest that post-transcriptional control of *clp* represents a unique mechanism by which many downstream genes can be governed through the regulation of an important component of the pre-mRNA 3′-end processing machinery.

## MATERIALS AND METHODS

### Preparation of an affinity-purified anti-CLP polyclonal antibody

Purification of the C-terminal portion of a glutathione-S-transferase (GST)–CLP fusion protein was performed as described previously (15). This fusion protein, which contains the C-terminal 112 amino acids of CLP (i.e. residues 184–296), was injected into rats by Pocono Rabbit Farms (Canadensis, PA). Polyclonal antibodies directed against GST were removed by passing the antiserum through a column that contained GST protein covalently attached to Affi-Gel 10 beads (Bio-Rad). The flow-through was loaded onto a second column containing the GST–CLP fusion protein covalently attached to Affi-Gel 10 beads. After several washes, anti-CLP polyclonal antibodies were eluted as described by Lasko and Ashburner (16). Western blots were treated with anti-CLP diluted 1:100 followed by a 1:1000 dilution of alkaline phosphatase-goat anti-rat IgG (Zymed Laboratories, CA) which served as the secondary antibody. CLP protein was visualized with the BCIP/NBT substrate kit (Zymed Laboratories, CA).

### Whole mount antibody staining

Immunolocalization of CLP protein in ovaries, embryos and in dissected third instar larvae was performed as described previously (17,18). The affinity-purified polyclonal anti-CLP antibody was used at a 1:100 dilution and the rabbit anti-rat secondary antibody was diluted 1:500. Histochemical reactions were performed using the DAB substrate kit (Vector Laboratories), as recommended by the manufacturer. Stained ovaries, embryos and larvae were mounted in Polyaquamount (Polysciences).

### Polyadenylation assay

Ovaries were dissected in 1× EBR (130 mM NaCl, 4.7 mM KCl, 1.9 mM $CaCl_2$, 10 mM HEPES, pH 6.9) or PBST (1× PBS with 0.2% Tween-20) from female flies fed wet yeast paste for 2–4 days after eclosion. Embryos were collected at 1 h intervals from well conditioned female flies at 25°C, aged appropriately and stored at –70°C. RNA samples were prepared as described by Bai and Tolias (15). Total RNA from egg chambers and embryos was treated with RQ1 RNase-free DNase (Promega) for 15 min at 37°C, phenol–chloroform extracted, ethanol precipitated and resuspended in 20 µl of $H_2O$. Approximately 1 µg of total RNA from each sample was subjected to the polyadenylation assay (19). Typically, 1 µl of the reverse transcription product was amplified with 25 pmol of the T-anchor primer (19) and a primer (5′-GTGTAGTCCA-GAGGTCGTAG-3′) directed between nucleotides 1079 and 1098 of the *clp* gene (15). The reaction mixture also contained 5 µCi of [α-$^{32}$P]dATP (New England Nuclear). Polymerase chain reaction (PCR) amplification was then performed as follows: 93°C for 5 min; 30 cycles at 93°C for 30 s; 62°C for 1 min; 72°C for 1 min and a final extension time of 7 min at 72°C. The expected size of the smallest possible PCR product was 197 nt (i.e. 167 nt at the 3′ end of *clp* plus 30 nt from the T-anchor primer). Products were separated from mineral oil, extracted with phenol chloroform and passed through a Sephadex G-50 column. Samples were then resolved by electrophoresis on a 5% non-denaturing polyacrylamide gel.

### Identification and purification of a mouse CLP homolog

A *Mus musculus* homolog of CLP was initially identified from a library of expressed sequence tags (EST) by the program BLAST. The cDNA clone was obtained from Research Genetics, Inc. (Huntsville, AL) as I.M.A.G.E. consortium clone ID 439437 (20), and was sequenced from both strands. The sequence has been deposited under DDBJ/EMBL/GenBank accession no. AF033201. Mu-CLP was then subcloned between the *Eco*RI and *Not*I sites of the pGEX1λT expression vector (Pharmacia Biotech). Affinity purification of GST-Mu-CLP was performed as described by Smith and Johnson (21). Cloning and purification of full length CLP has been described previously (15).

### Selection–amplification of RNA substrates (SELEX)

The RNA SELEX technique was performed as described in Brown and Gold (22) with the following modifications: 50 pmol of synthetic DNA template containing 30 nt of random sequence (5′-GCCGGATCCGGGCCTCATGTCGAA[**30N**]TGAGCGTTT-ATTCTGAGCTCCC-3′) was amplified by PCR using RNA SELEX 3′ primer (5′-GCCGGATCCGGGCCTCATGTCGAA-3′) and RNA SELEX 5′ T7 primer (5′-CCGAAGCTTAATACGAC-TCACTATAGGGAGCTCAGAATAAACGCTCAA-3′). PCR

amplification was carried out for 30 cycles under the following conditions: 93°C for 30 s; 55°C for 20 s and 72°C for 1 min. Products were phenol–chloroform extracted and precipitated with ethanol. Approximately 1 µg of each PCR product was transcribed with T7 RNA polymerase in the presence of RNasin RNase inhibitor (Promega). After 2 h of transcription, 2 U of RQ1 RNase-free DNase (Promega) was added for 15 minutes at 37°C. The transcription product was extracted with phenol–chloroform, passed through a Sephadex G-50 column and precipitated with ethanol. RNA transcripts were further separated on 2% low melting point agarose gels, purified with β-agarase (New England Biolabs), and resuspended in 100 µl of binding buffer containing 20 mM Tris–HCl pH 7.5, 250 mM of NaCl, 0.1 mM EDTA.

RNA was incubated in 100 µl of binding buffer with 100 nM of GST fusion protein on agarose beads at room temperature for 30 min. Agarose beads were then separated by centrifugation and washed for 30 min at room temperature with 200 µl of binding buffer. This was repeated three times. Bound RNA was eluted with 100 µl of 15 mM reduced glutathione in binding buffer. The eluted RNA was phenol–chloroform extracted and precipitated with 2.5 vol of ethanol in the presence of 0.3 M NaOAc (pH 5.2) and 20 µg of glycogen. The pellet was washed with ethanol, dried, and resuspended in $H_2O$. RNAs were reversed transcribed with SuperScript II RNase $H^-$ Reverse Transcriptase (Gibco BRL), as recommended by the manufacturer. One µl of the reverse transcription product was subjected to PCR amplification at the following conditions: 94°C for 30 s, 65°C for 30 s and 72°C for 30 s. PCR products were again extracted with phenol chloroform and precipitated with ethanol. Beginning with the second SELEX cycle, RNA transcripts were pretreated with GST–agarose beads for 30 min at room temperature before incubating with the GST fusion protein agarose beads. For SELEX experiments using Mu-CLP and a full length version of CLP, the following binding buffer was used: 20 mM Tris–HCl, 250 mM NaCl, 0.5 mM $ZnSO_4$.

To subclone the PCR amplification products, we passed them through a Sephadex G-50 column, digested them with *Bam*HI and *Hin*dIII and cloned them into pBluescript I KS+ vector (Stratagene). Sequencing was performed using the T7 Sequenase sequencing kit (Amersham Life Science) and KS primer (Stratagene).

## RESULTS

### *clp* encodes a developmentally regulated nuclear protein

Our earlier work showed that *clp* transcripts are not uniformly expressed during development (15). These studies used a developmental RNase protection assay as well as whole mount RNA *in situ* hybridization to highlight the accumulation and distribution of *clp* transcripts throughout the *Drosophila* life cycle. Though *clp* mRNAs were detected at every stage of development, the relative expression of these transcripts varied considerably.

To examine whether the *clp* encoded protein and RNA expression patterns coincide, we generated and affinity-purified a polyclonal antibody directed against the C-terminal end of CLP protein and used it for whole mount immunolocalization studies. Prior to initiating these experiments, we performed western blot analysis and confirmed that the isolated antibody specifically cross-reacted against CLP protein (data not shown). Since immunolocalization of CPSF 30K or any related protein had yet to be performed, these studies addressed the expectation that CLP is a nuclear antigen, given its putative involvement in nuclear

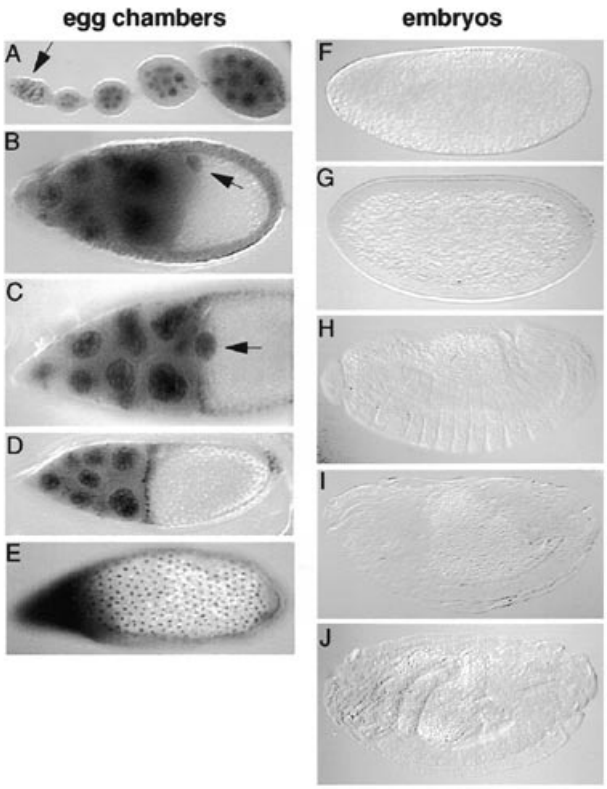


**Figure 1.** Expression of CLP protein in egg chambers and embryos. Egg chambers (**A**–**E**) and embryos (**F**–**J**) are oriented with anterior to the left and dorsal facing up. Expression is revealed by immunohistochemimcal staining using an affinity-purified anti-CLP polyclonal antibody. Nuclear localization of CLP protein is detected throughout oogenesis in the nurse cells (A–D), the oocyte (B and C) and the somatically derived follicle cells (E). The arrow in (A) points to the germarium. Arrows in (B) and (C) point to the oocyte nucleus. CLP protein cannot be detected throughout embryogenesis, even though maternally derived *clp* transcripts are present (15).

pre-mRNA 3′-end processing. The results presented in Figures 1 and 2, examining the distribution of CLP protein in egg chambers and larvae, confirm the predicted nuclear localization. During oogenesis, CLP protein was detected in the germarium (Fig. 1A) and in the nuclei of nurse cells (Fig. 1A–D). The nurse cell expression persisted until stage 12 (23) when these cells degenerated and emptied their contents into the oocyte. Throughout oogenesis, CLP protein was also detected in the oocyte nucleus (Fig. 1B and C) as well as in the nuclei of the somatically derived follicle cells (Fig. 1E). We conclude that during oogenesis, CLP protein is distributed in the nuclei of egg chambers, consistent with expression of its mRNA in the cytoplasm (15).

Though *clp* transcripts are present throughout embryogenesis (15), we could not detect CLP protein during any period of embryonic development using the whole mount antibody staining technique (Fig. 1F–J). This result was also confirmed by the absence of CLP protein in embryonic extracts as assayed by western blot analysis (data not shown). We also examined the distribution of CLP protein in third instar larvae where we have previously reported a zygotically derived burst of *clp* transcription (15). Figure 2 displays an abundant nuclear localization of CLP protein in virtually all larval organs and discs. The only obvious exceptions were the salivary glands (Fig. 2E–H), which initially

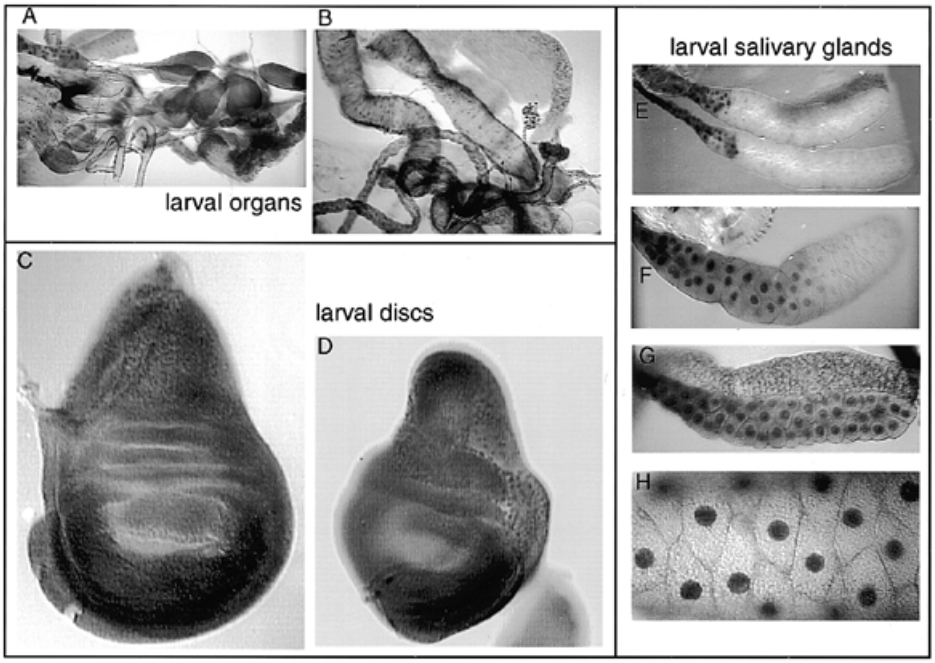


**Figure 2.** Expression of CLP protein in third instar larvae. CLP is localized in the nuclei of virtually all larval organs (**A** and **B**) and discs (**C** and **D**). The only exception is the salivary glands (**E–H**) where CLP is initially restricted to cells at the anterior end of the gland (E) and gradually progresses in the posterior direction (F) until it is expressed throughout the entire gland (G).

expressed CLP protein in a restricted pattern confined to cells at the anterior end of the gland (Fig. 2E). Later, this spatially restricted distribution gradually expands in the posterior direction (Fig. 2F) until CLP protein is present throughout the entire gland (Fig. 2G). The significance of this regulated wave of CLP expression in salivary glands is not known. Nevertheless, these immunolocalization studies have shown CLP to be a nuclear protein whose temporal and spatial distribution are developmentally regulated. However, the major difference between the distribution of the *clp* encoded protein and the previously reported RNA expression pattern is that they do not coincide during embryogenesis.

## Embryonic post-transcriptional regulation occurs without altering the poly(A) tail length of *clp* transcripts

The results presented above demonstrate that CLP protein is absent during embryonic development. This was unexpected considering that maternally derived *clp* transcripts are provided to the oocyte during oogenesis and persist in the embryo throughout most of embryogenesis (15). The absence of CLP protein during this period shows that *clp* is post-transcriptionally regulated during embryogenesis. A possible mechanism by which this regulation can be achieved is by changing the length of the poly(A) tail. Alterations in poly(A) tail length after fertilization have been shown to affect the translational initiation efficiency of other *Drosophila* genes (24,25). Thus, we compared the poly(A) tail length of ovarian versus embryonic *clp* mRNA using a PCR-based polyadenylation assay (19). Total RNAs from ovaries and several early embryonic stages were collected and subjected to reverse transcription with the T-anchor primer (19). This technique generates cDNAs that contain a poly(A) track equivalent to the length of the poly(A) tail on the corresponding mRNA. The cDNA was then amplified with a *clp* specific primer and the T-anchor primer.

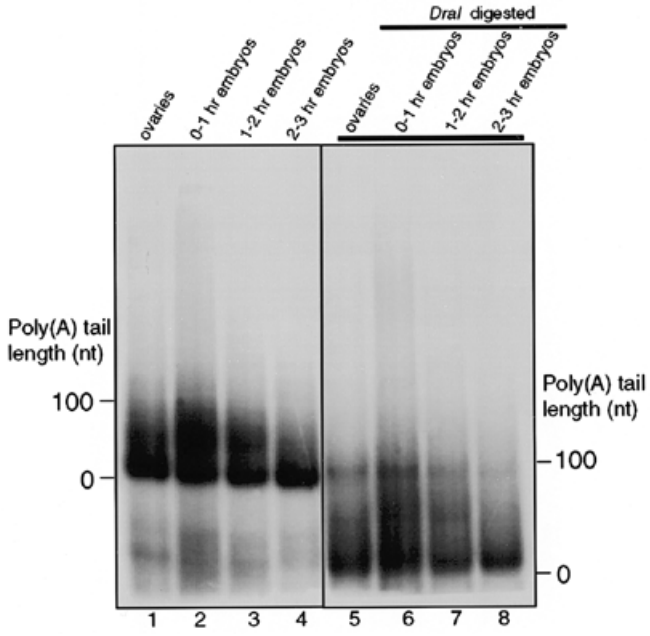


**Figure 3.** The length of the poly(A) tail in *clp* transcripts remains unaltered in early embryos. Total RNA from Oregon-R adult ovaries or the indicated age of early embryos was purified and subjected to the PCR based polyadenylation test (19). PCR amplification was performed using a *clp* specific primer (directed near the 3′ end) and the T-anchor primer (lanes 1–4). To confirm that these 3′ end products correspond to *clp* transcripts, the same reactions were also digested with *Dra*I (lanes 5–8) which cleaves 62 bp downstream from the *clp* primer. The 62 bp 5′ fragment has been run off the gel.

The results of this experiment are displayed in Figure 3. In egg chambers, where CLP protein is abundantly expressed, *clp*

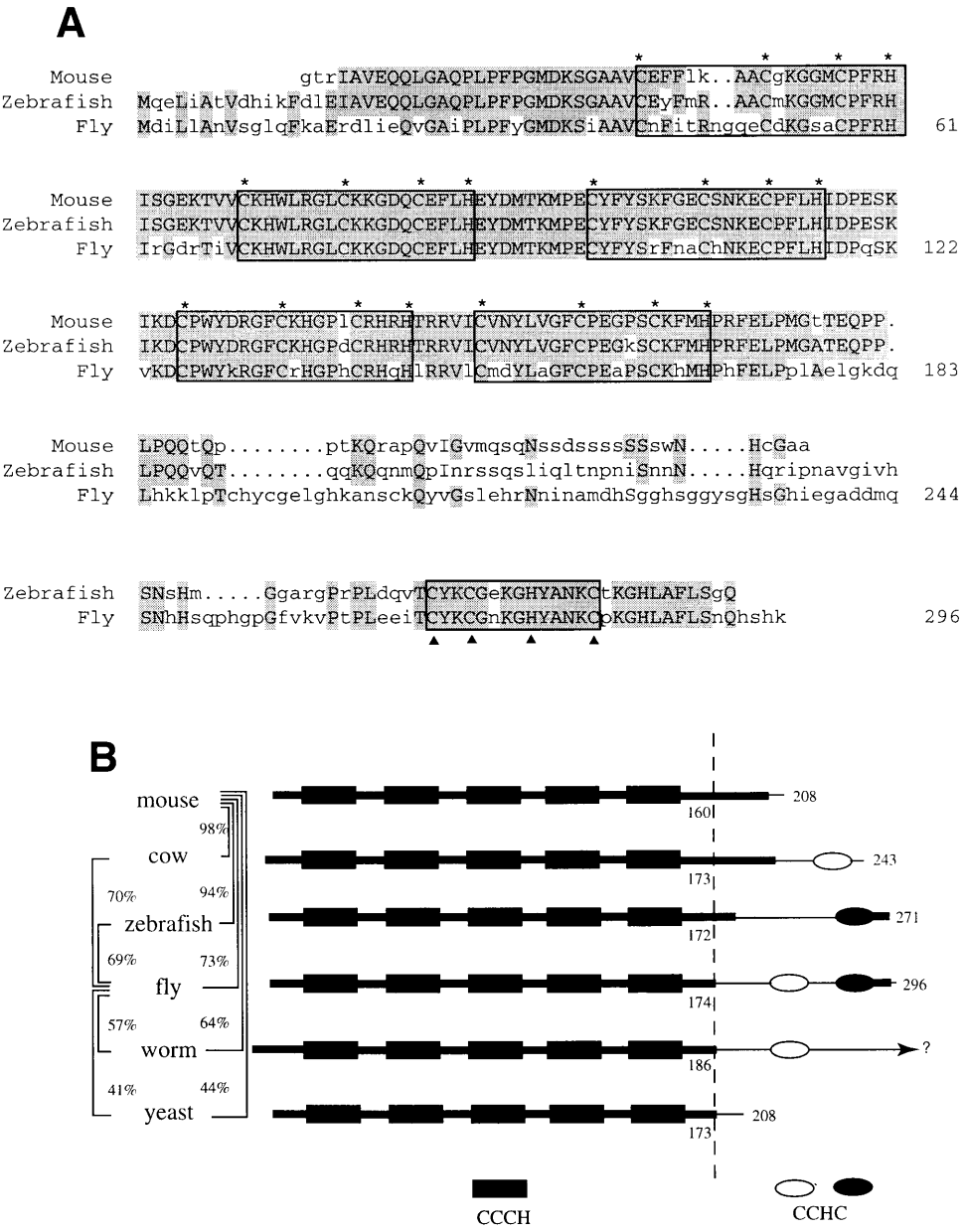


**Figure 4.** CLP/CPSF 30K homologs constitute a highly conserved family of eukaryotic proteins. (**A**) Protein alignment of CLP/CPSF 30K homologs from mouse (*M.musculus*), zebrafish (*Danio rerio*) and fly (*D.melanogaster*). The cDNA sequence of Mu-CLP has been deposited at DDBJ/EMBL/GenBank under accession no. AF033201. Alignments were performed using sequence data provided by DDBJ/EMBL/GenBank (sequences were retrieved using the program BLAST). Upper case shaded letters denote identical residues present in at least one more protein. Dots (.), indicate gaps in protein sequence introduced to achieve optimum alignment. The CCCH and CCHC zinc finger consensus sequences are boxed and indicated. Other DDBJ/EMBL/GenBank accession nos include U26549 (CLP from *D.melanogaster*) and U70479 (NAR from *D.rerio*). (**B**) Identity relationships among CLP/CPSF 30K family members. Amino acid identities between any two proteins are expressed as the percentage indicated on the right. These identities have been calculated for the N-terminal portion of each protein up to the amino acid position indicated by the dashed line. The total number of amino acids present in each protein is indicated on the right. Thicker horizontal lines and black filled regions in the schematic represent regions that are highly conserved. Other DDBJ/EMBL/GenBank accession nos include Z68297 (F11A10.3 from the worm *Caenorhabditis elegans*), U32445 (P8283.17 from the yeast *Saccharomyces cerevisiae*) and U96448 (bovine CPSF 30K from the cow *Bos taurus*).

mRNAs contain a poly(A) tail ~100 nt in length (Fig. 3, lane 1). During early embryogenesis (i.e. 0–1 h, 1–2 h and 2–3 h), we did not detect any significant changes in the length of the poly(A) tail in maternally provided *clp* transcripts (Fig. 3, lanes 2–4). To confirm that the bands on the gel were bona fide PCR products representing *clp* transcripts, they were digested at a unique restriction enzyme site predicted to shorten the products by 62 bp. The results shown in Figure 3, lanes 5–8, confirmed this prediction. Since the length of the poly(A) tail in *clp* transcripts does not substantially change between oogenesis and early embryogenesis, we conclude that the absence of CLP protein in the embryo does not involve translational repression of maternally provided *clp* mRNA through poly(A) tail length shortening.

**Table 1.** SELEX experiments using full length CLP versus a mouse homolog, did not reveal any obvious binding preference for specific RNA sequences, secondary structures or clusters of particular nucleotides

| Drosophila CLP | Mouse CLP Homolog |
|---|---|
| GGTAACAGG | TCC |
| TATGCCTATG | TAG |
| GTTCCTCGCTT | TGGCT |
| TGTTTCGTTGA | TTTTGA |
| CGTTCACCCGCAA | GGGCGC |
| TGGTTGCCGTGTAC | CGGTCA |
| TTGTTAGTGCCAAA | GTTGGTAT |
| CCTCGTACTCTTGCCTC | GGTTGTCTT |
| CCTTAATTCAATGCCTC | CGTGCCTGAC |
| TCTTGTCGCACCCTCTACAGCTTT | TCGCACACGCA |
| GATAGGTGTAGTGGGTGCACGTGC | GACTAATCGGCTCGCAC |
| GCCCAACGGCTGTTTCCCTCGTTAGGTA | TGATGTCGGCCGTTGCAC |
| CACGCCAGTTGTATTCACATAGCTAGATA | GATTGTCTCGTCGCATGCTGG |
| GATCTGGTTAGTGCCCTGGCAACATATACA | GGTTGTCTTGCACTAGGCGCTTGC |
| TGAACTCGTCTAATAGTGCGTAAAGAGTCC | CGTCTCCTGCGCTTGGCCCCTCCGC |
| GGAATAGTCATTGCATGGCCAGTGGTGCCC | GTTGCTCGACGATGAGTTGATTTGGT |
| TTACCACGCACCCACAAAAACATTGTCCCA | TCGTACGTGCAGAATGGTGTCACCGCCAC |
| ATTCAATCGTCATCAGTCGACACTTTCCCT | TTTACATGATGGATGTTAGGTGCCCGCGC |
| TTTGCTCGTCATTTTCTTGGTTACCTGCCA | TTTTCCCCTGTTTCCTTCAACGCAAGCGGT |
| CTCGCCCCACATAATTGTTTCTGCTCTGCG | CCAGGGAGTGCACCAGCCACATTTGCCATG |

Most of the sequences (selected at cycle eight) contain substantial deletions which reflect the ribonucleolytic activity associated with each full length protein.

## Sequence conservation and activity of CLP/CPSF 30K homologs

CLP is characterized by five CCCH zinc fingers. The CCCH motif is the rarest among the zinc finger classes. We have shown previously that the ribonuclease activity of CLP resides in a region containing five CCCH zinc fingers (15). CLP also contains two CCHC zinc knuckles at it's C-terminus. CCHC zinc knuckles are more common and are often present in proteins that bind exclusively to single stranded DNA or RNA (26). Mutagenesis experiments have demonstrated that zinc knuckles can play a critical role in RNA recognition (27).

To examine whether the activities and binding properties of *Drosophila* CLP were conserved in other species, we identified a mouse homolog of CLP as an expressed sequence tag (EST) from a mouse embryonic cDNA library. The mouse homolog was sequenced and aligned to *Drosophila* CLP and the zebrafish homolog [encoded by *no arches*(*nar*); 28] as shown in Figure 4A. We also compared the overall motif organization and calculated the relative amino acid sequence conservation among CLP homologs from three additional species (Fig. 4B). These comparisons revealed a remarkable degree of sequence conservation among all known CLP/CPSF 30K homologs. For example, between animo acids 17 and 205, bovine CPSF 30K is 99% identical to Mu-CLP. This region also contains the five CCCH zinc fingers which are highly conserved in all six CLP homologs as are the amino acid sequences that separate them. *Drosophila* CLP is 73% identical to Mu-CLP and 70% identical to bovine CPSF 30K over its N-terminal 174 amino acids. The fact that the N-terminal five CCCH zinc fingers are highly conserved from yeast to mammals suggests that they encode the main function of this protein family. In contrast, the less conserved C-terminus which contains one or two CCHC knuckles in four out of six proteins may play a supportive role.

Using a number of DNA and RNA polymers coupled to a solid matrix, Barabino *et al.* (5) have shown that bovine CPSF 30K binds to poly(U) and poly(G) and that deletion of the zinc knuckle motif decreases this activity. To examine whether *Drosophila* CLP and Mu-CLP possess similar activities, we employed a more stringent strategy which modified a PCR-based selection–amplification (SELEX) procedure originally described by Tuerk and Gold (29). After eight cycles of selection, we did not detect a binding preference for specific RNA sequences or secondary structures (Table 1). However, a large portion of the selected products contained substantial internal deletions which reflect an endoribonuclease activity associated with each protein.

Since CLP and Mu-CLP generated deletions in RNA SELEX experiments, we examined whether the C-terminal 112 amino acids of *Drosophila* CLP (i.e. residues 184–296), which contain two CCHC zinc knuckles, confer a preference for particular RNA sequences. We have established previously that this portion of the protein does not encode an endoribonuclease activity (15). In these experiments, we performed eight cycles of SELEX and monitored the progress after four and eight cycles. After subcloning and sequencing 24 clones from cycle four, we did not detect any binding preferences for specific nucleotide sequences or particular RNA secondary structures. However, 18 out of 24 additional clones sequenced after cycle eight contained a strong nucleotide bias for G and/or C, which averaged 70% and ranged from 60 to 80% (Table 2). In addition, 21 out of 24 sequences contained one or more clusters of five or more continuous GC, poly(G) or poly(C) tracks. It thus appears that the C-terminal portion of *Drosophila* CLP, which contains two zinc knuckles, confers a binding preference for RNA sequences that contain G and/or C clusters. This is reminiscent of the binding preference for poly(U) and poly(G) conferred by the zinc knuckle motif of bovine CPSF 30K (5).

## DISCUSSION

Our immunolocalization studies indicate that *clp* encodes a nuclear protein. This is consistent with its role as a component of the nuclear pre-mRNA 3′-end processing machinery. However, we were surprised to discover that CLP protein is not present in the *Drosophila* embryo, even though maternally derived *clp* transcripts are easily detected. These results demonstrate that *clp* is post-transcriptionally regulated during embryogenesis. Since alternative pathways of pre-mRNA processing have not been identified, our results suggest that either the CPSF 30K subunit is not required during all temporal periods or that an unknown protein may substitute its function during embryonic development.

There are several post-transcriptional mechanisms by which expression of *clp* can be controlled. The possibility that maternally derived *clp* mRNA may be modified in embryos to affect its stability can be excluded since *clp* transcripts are detectable throughout embryogenesis (15). A more likely mechanism by which *clp* may be post-transcriptionally controlled may involve alteration of its translational initiation efficiency. For example, it has been shown that maternally derived *bicoid* transcripts undergo cytoplasmic poly(A) elongation at their 3′ end, an event that correlates with their translation (24). A similar mechanism may operate to repress the translation of *clp* mRNA but through poly(A) tail length shortening. However, this possibility can be excluded since our results demonstrate that the poly(A) tail length of *clp* mRNA does not substantially change between oogenesis and early embryogenesis. Another possibility is that translation of *clp* mRNA is regulated by a *trans*-acting repressor protein(s) that functions in early embryos. Alternatively, it is possible that embryos may lack a translational activator that is present in ovaries and larval tissues. In any event, it seems likely that sequences within *clp* mRNA such as the 5′- or 3′-UTRs may provide the *cis*-acting sequences that are bound by translational regulatory proteins. An example of this type of regulatory strategy occurring in early embryos involves the translational repression of *hunchback* (*hb*) which requires two nanos response elements in the 3′-end of *hb* mRNA and two *trans*-acting factors: pumilio and an unknown 55 kDa protein (30).

A bovine homolog of CLP was recently identified as the 30K subunit of CPSF, a multimeric protein complex that recognizes the polyadenylation signal AAUAAA and is involved in both 3′ end cleavage and subsequent polyadenylation of pre-mRNAs (5). Though it has been shown that recombinant CPSF 160K can preferentially bind to pre-mRNAs containing AAUAAA (6), the affinity of this interaction was much weaker compared with purified CPSF. In addition, UV crosslinking experiments with HeLa nuclear extracts or partially purified CPSF, have shown that the 160K and 30K subunits crosslink to a substrate that contains an AAUAAA polyadenylation signal (31,32). Therefore, it is possible that the 30K subunit may enhance the ability of CPSF 160K to bind to the AAUAAA polyadenylation signal.

Since CLP contains two zinc knuckles at the C-terminus, a motif that has been implicated in RNA binding, we suspected that this portion of the protein may confer RNA recognition. This hypothesis was supported by experiments demonstrating that deletion of the zinc knuckle motif dramatically decreases the ability of bovine CPSF 30K to bind poly(U) and poly(G) polymers *in vitro* (5). Using a modified RNA SELEX procedure, we have shown that the C-terminal end of CLP, which contains two zinc knuckles, preferentially recognizes sequences that contain G- and/or C-rich clusters. Sequences such as these may be important for the formation of stable secondary structures that may be recognized by CLP *in vivo*. If these G- and/or C-rich clusters reside near the AAUAAA element, they may support CPSF-mediated recognition of the polyadenylation signal or RNA cleavage.

Immunodepletion of either HeLa nuclear extracts or a partially purified CPSF fraction with antiserum directed against CPSF 30K can significantly reduce cleavage and eliminate detectable polyadenylation as measured by *in vitro* assays (5). A temperature-sensitive allele that lacks the C-terminal 55 amino acids of a yeast CLP/CPSF 30K homolog but retains four of the five N-terminal CCCH zinc finger motifs has also provided some functional insight. Under permissive conditions, yeast extracts prepared from this temperature-sensitive allele can still cleave a pre-mRNA substrate, but cannot polyadenylate the upstream cleavage product. The uncoupling of the cleavage and polyadenylation reactions demonstrates that the cleavage activity of the yeast protein resides in a region containing four copies of the N-terminal CCCH zinc finger motif. This is consistent with our demonstration that the ribonuclease activity of CLP resides in the N-terminal portion of the protein that contains five CCCH zinc fingers (15). It is also supported by our finding that the Mu-CLP homolog also contains a similar ribonuclease activity, as demonstrated by the deleted substrates that were generated using the SELEX assay. We suggest that the ribonuclease activity of CLP/CPSF 30K family members is conserved and resides on the same N-terminal portion of the protein.

To date, our studies have illustrated CLP to be nuclear protein with a endoribonuclease activity that it is developmentally regulated both at the transcriptional and post-transcriptional level. Currently, *clp* mutant alleles from *Drosophila* are not available; their identification will be the focus of our future work. However, *clp* mutant alleles from two other organsisms have been generated. These studies have shown *clp* homologs from yeast and zebrafish to be encoded by essential genes (5,28). How can multicellular organisms regulate gene expression during development through an apparently essential component of the polyadenylation machinery which is thought to operate in all cells? One possibility is that in the absence of the 30K subunit, CPSF may process only a subset of nuclear pre-mRNAs. Under this scenario, when the 30K subunit reassociates with the other three CPSF components, this may modulate the specificity of CPSF to process another distinct subset of pre-mRNA targets. In support of this hypothesis, it was reported previously that some CPSF preparations do not contain detectable levels of the 30K subunit (6). This is further substantiated by the recent finding that CPSF molecules that lack the 30K subunit are first recruited to the transcription preinitiation complex by TFIID and then transferred through RNA polymerase II to nascent RNA transcripts (33). These observations have prompted our formulation of a model explaining the functional role of CLP/CPSF 30K homologs. Our model proposes that CPSF may function in the absence of the 30K subunit to process a select group of target pre-mRNAs. However, we speculate that when present, the 30K subunit can join the other three CPSF components and cooperate with the 160K subunit to alter target specificity so that a different set of pre-mRNAs targets are recognized and processed. Finally, the activity of CPSF 30K may be further modified post-translationally, thus promoting the recognition and processing of additional pre-mRNAs.

Evidence is accumulating to support that expression of specific sets of genes can be regulated during development by controlling pre-mRNA 3′-end processing. In zebrafish, the *nar* gene encodes a CLP/CPSF 30K homolog that is essential for zebrafish

**Table 2.** A C-terminal portion of *Drosophila* CLP, that contains two CCHC zinc knuckles, confers a binding bias for G- and/or C-rich RNA sequences

| Selected oligonucleotides | G+C |
|---|---|
| CGCCTGGTCCCAACCGAGGACCGT**CGCCGG** | 76% |
| CCCAGGTCACACGCCACT**GCCCCGCCCC**TG | 77% |
| AA**CCGCC**TCCGT**CGCCC**ACT**CGCCGCG**TG | 79% |
| CTT**GCGGCGC**A**GGGGGCC**T**GGCGGCG**ACTG | 80% |
| GAGTGATAAACCT**GCGCCGCCGCGCC**A | 67% |
| GGCCTTGGCT**GCCGC**T**GCGCCGGC**AGCTG | 79% |
| GCGAAGCAACCGGAGATA**CCCCGCG**TGTGT | 63% |
| GCATATAGTGTTGATCCGTT**GGGCG**TTACT | 50% |
| GCATATAGTGTTGATCCGTT**GGGCG**TTACT | 50% |
| CAGCAGAGTGTAGCA**CCGGCGGCC**AGTCCT | 67% |
| TTGGGTGGAGCGTAA**CGCCCGGGG**ACGA | 67% |
| CTAACAGT**GGGGCGGC**A**GGGGG**TCGGAAC | 69% |
| CAAACGAGAACAATCT**GCGGGG**AGGCAACGA | 57% |
| CTGCGCTA**CGCGGGGGGG**T | 79% |
| ATCGT**CGGGGG**A**GCGCC**AGTGTGTTCGAAA | 60% |
| **GGGGC**TTGCGTT**GCGCG**ACACCACCCGTTC | 67% |
| GACTGAGCT**GCGGCGC**ACGCGTGAATACTT | 67% |
| TATTCGCTA**GGGGCCG**ATACTGGGCAACCC | 60% |
| CTGATGATA**CGGGG**ATGCAACTAAGTTTGG | 47% |
| AGGT**GGCGGGGCC**AAATGCTGACGTGCCCA | 67% |
| ATTT**GGCCG**TTTCCCTTAATGTCACCTTGT | 43% |
| GGGTGCTCTCGGTGCGCACGCAT | 69% |
| GCTCCTCTCCTTGACCTCCGGTGGTACCGC | 67% |
| TGCCTTAGCTCATGACGGTGA | 52% |

Twenty four sequences selected after SELEX cycle eight are displayed. Clusters of three or four GC, poly(G) or poly(C) tracks are underlined, whereas clusters of five or more are indicated by bold text. The G + C content of each sequence is also shown.

development. Mutations in *nar* generate specific head phenotypes which include defects in the development of the eyes and pharyngeal arches (28). Given that the zebrafish genome is larger and more redundant compared with *Drosophila*, it is likely that *clp* is also an essential gene whose loss of function gives rise to specific developmental defects. Additional support for the notion that the 3′-end processing machinery can be developmentally regulated comes from the discovery that the gene encoding CstF 77K is a homolog of the *Drosophila suppressor of forked* [*su(f)*] gene, which encodes a modifier of gene expression (12). Mutations in *su(f)* can enhance or suppress the phenotypes generated by transposable element insertions in specific genes by altering the polyadenylation of their transcripts. Thus, fluctuations in the amount or activity of CstF 77K can change the polyadenylation efficiency of certain genes. A third independent example of controlling gene expression through pre-mRNA 3′-end processing involves CstF 64K and its role in regulating the transition from membrane-bound to secreted forms of IgM during B cell differentiation. Undifferentiated B cells produce large amounts of mRNA encoding the membrane-bound form of IgM heavy chain (μm), while differentiated B cells mostly produce secreted forms of IgM heavy chain (μs). The two different IgM heavy chains result from alternative use of polyadenylation sites at the 3′-end of IgM mRNAs. CstF 64K is the limiting factor responsible for the switch from membrane bound to secreted form of IgM during B cell differentiation (34). Another way by which pre-mRNA 3′-end processing can modulate gene expression is through the C-terminal end of PAP which contains a 25 kDa region rich in serine and threonine that can be phosphorylated to alter its activity (35). Finally, 3′-end processing can also control gene expression through PAB II which determines the length of the poly(A) tail. Since poly(A) tail length determines both mRNA stability and translational initiation efficiency, this is also a popular strategy employed for the post-transcriptional control of gene expression in the cytoplasm. In their totality, these observations support the hypothesis that the expression of specific genes can be regulated during development by modulating the components that mediate 3′-end processing of their pre-mRNAs in the nucleus.

## ACKNOWLEDGEMENTS

This work was supported by the Public Health Research Institute and by National Science Foundation grant IBN-9418722 awarded to P.P.T.

## REFERENCES

1 Keller, W. (1995) *Cell* **81**, 829–832.
2 Keller, W. and Minvielle-Sebastia, L.A. (1997) *Curr. Opin. Cell Biol.* **9**, 329–333.
3 Bienroth, S., Wahle, E., Suter-Crazzolara, C. and Keller, W. (1991) *J. Biol. Chem.* **266**, 19768–19776.
4 Murthy, K.G.K. and Manley, J.L. (1992) *J. Biol. Chem.* **267**, 14804–14811.
5 Barabino, S.M.L., Hubner, W., Jenny, A., Minvielle-Sebastia, L. and Keller, W. (1997) *Genes Dev.* **11**, 1703–1716.
6 Murthy, K.G.K. and Manley, J.L. (1995) *Genes Dev.* **9**, 2672–2683.
7 Keller, W., Bienroth, S., Lang, K.L. and Christofori, G. (1991) *EMBO J.* **10**, 4241–4249.
8 Takagaki, Y., Manley, J.L., MacDonald, C.C., Wilusz, J. and Shenk, T. (1990) *Genes Dev.* **4**, 2112–2120.
9 MacDonald, C.C., Wilusz, J. and Shenk, T. (1994) *J. Mol. Biol.* **14**, 6647–6654.
10 Takagaki, Y., MacDonald, C.C., Shenk, T. and Manley, J.L. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1403–1407.
11 Takagaki, Y. and Manley, J.L. (1997) *Mol. Cell. Biol.* **17**, 3907–3914.
12 Takagaki, Y. and Manley, J.L. (1994) *Nature* **372**, 471–474.
13 Ruegsegger, U., Beyer, K. and Keller, W. (1996) *J. Biol. Chem.* **271**, 6107–6113.
14 Wahle, E. (1991) *Cell* **66**, 759–768.
15 Bai, C. and Tolias, P.P. (1996) *Mol. Cell. Biol.* **16**, 6661–6667.
16 Lasko, P.F. and Ashburner, M. (1988) *Nature* **335**, 611–617.
17 Xue, F. and Cooley, L. (1993) *Cell* **72**, 681–693.
18 Stroumbakis, N.D., Li, Z. and Tolias, P.P. (1996) *Mol. Cell. Biol.* **16**, 192–201.
19 Salles, F.J. and Strickland, S. (1995) *PCR Methods Appl.* **4**, 317–321.
20 Lennon, G., Auffray, C., Polymeropoulos, M. and Soares, M.B. (1996) *Genomics* **33**, 151–152.
21 Smith, D.B. and Johnson, K.S. (1988) *Gene* **67**, 31–40.
22 Brown, D. and Gold, L. (1995) *Biochemistry* **34**, 14765–14774.
23 King, R.C. (1970) *Ovarian Development in Drosophila melanogaster.* Academic Press. New York, NY.
24 Salles, F.J., Lieberfarb, M.E., Wreden, C., Gergen, J.P. and Strickland, S. (1994) *Science* **266**, 1996–1999.
25 Lieberfarb, M.E., Chu, T., Wreden, C., Theurkauf, W., Gergen, J.P. and Strickland, S. (1996) *Development* **122**, 579–588.
26 Rajavashisth, T.B., Taylor, A.K. Andalibi, A., Svenson, K.L. and Lusis, A.J. (1989) *Science* **245**, 640–643.
27 South, T.L., Blake, P.R., Sowder, R.C., III, Arthur, L.O., Henderson, L.E. and Summers, M.F. (1990) *Biochemistry* **29**, 7786–7789.
28 Gaiano, N., Amsterdam, A., Kawakami, K., Allende, M., Becker, T. and Hopkins, N. (1996) *Nature* **383**, 829–832.
29 Tuerk, C. and Gold, L. (1990) *Science* **249**, 505–510.
30 Murata, Y. and Wharton, R.P. (1995) *Cell* **80**, 747–756.
31 Gilmartin, G.M. and Nevins, J.R. (1991) *Mol. Cell. Biol.* **11**, 2432–2438.
32 Jenny, A., Hauri, H.P. and Keller, W. (1994) *Mol. Cell. Biol.* **14**, 8183–8190.
33 Dantonel, J.C., Murthy, K.G., Manley, J.L. and Tora, L. (1997) *Nature* **389**, 399–402.
34 Takagaki, Y., Seipelt, R.L., Peterson, M.L. and Manley, J.L. (1996) *Cell* **87**, 941–952.
35 Colgan, D.F., Murthy, K.G.K., Prives, C. and Manley, J.L. (1996) *Nature* **384**, 282–285.