Novel transcripts from the Ultrabithorax domain of the bithorax complex.

H D Lipshitz, D A Peattie and D S Hogness

Genes Dev. 1987 1: 307-322
Access the most recent version at doi:10.1101/gad.1.3.307

Novel transcripts from the *Ultrabithorax* domain of the bithorax complex

Howard D. Lipshitz,[1] Debra A. Peattie,[2] and David S. Hogness

Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305 USA

We present a detailed analysis of the transcriptional products of the *bithoraxoid (bxd)* region of the *Ultrabithorax* domain in the bithorax complex of *Drosophila melanogaster*. This region is transcribed twice during development: between 3 and 6 hr of embryogenesis, a set of early transcripts, 1.1 to 1.3 kb in size, is synthesized; from the midthird larval instar through the adult stages, a late 0.8-kb transcript is synthesized. We have sequenced five cloned cDNAs representing early transcripts and three cDNAs representing the late transcript and have located their exons within the 40 kb of DNA comprising the *bxd* region. S1 nuclease protection and primer extension of both the early and late transcripts were used to further elucidate their structure. The early RNAs are produced by complex differential splicing of a series of exons derived from a 26-kb primary transcript. Curiously, these RNAs do not possess significant protein coding potential. The late *bxd* RNA comprises a single exon transcribed from an intronic region of the early transcription unit. This RNA, by contrast, possesses excellent coding potential and, if translated, would yield a 101-amino-acid polypeptide.

[*Key Words*: Homeotic genes; *Ultrabithorax* (*Ubx*); *bithoraxoid* (*bxd*); bithorax complex; *Drosophila* development; transcription]

Received February 6, 1987; revised version received and accepted March 4, 1987.

The bithorax complex (BX-C) of *Drosophila melanogaster* consists of a cluster of homeotic genes that transduce positional information into segmental identity for parasegments 5–13 of the developing larva and adult (posterior second thoracic segment through the eighth abdominal segment; Martinez-Arias and Lawrence 1985). The BX-C can be subdivided genetically into individual identity functions (Lewis 1978, 1981; Morata and Kerridge 1981; Casanova et al. 1985; Karch et al. 1985) that can be grouped into three complementation groups or functional domains: the *Ultrabithorax* (*Ubx*) domain, *abdominal-A* (*abd-A*) domain, and *Abdominal-B* (*Abd-B*) domain (Sanchez-Herrero et al. 1985; Tiong et al. 1985). The entire BX-C has been cloned and many of its mutations mapped molecularly (Bender et al. 1983; Karch et al. 1985). Each functional domain has been shown to contain a single homeo box (Scott and Weiner 1984; McGinnis et al. 1984; Beachy et al. 1985; Karch et al. 1985; Regulski et al. 1985).

The *Ubx* domain controls the identity of parasegments 5 and 6 (ps 5 and 6) and, to a lesser extent, that of ps 7–13, whose major identity functions derive from the *abd-A* and *Abd-B* domains. *Ubx* mutations inactivate all identity functions of the *Ubx* domain, subsets of which are inactivated by four other classes of recessive mutations within the domain; thus, *anterobithorax* (*abx*) and *bithorax* (*bx*) mutations define the ps 5 identity functions, while *postbithorax* (*pbx*) and *bithoraxoid* (*bxd*) mutations define the ps 6 identity functions, as well as the lesser ps 7–13 functions of the domain (for review, see Hogness et al. 1985).

Figure 1 shows that the *Ubx* domain is physically divisible into two regions: a 75-kb *Ubx* region defined by the sites of the *Ubx* mutations and containing the *abx* and *bx* mutations (not shown), and a 40-kb *bxd* region defined by the sites of the *bxd* mutations and containing DNA deleted by the two known *pbx* mutations. The *Ubx* region is coextensive with a long transcription unit (*Ubx* unit), whose primary transcript is differentially processed to yield at least five mRNAs (Beachy et al. 1985; Hogness et al. 1985; K. Kornfeld and D.S. Hogness, unpubl.). These mRNAs encode a family of *Ubx* proteins characterized by a variable internal region that joins constant amino- and carboxy-terminal regions, of which the latter contains the homeo domain encoded by the *Ubx* homeo box. Beachy et al. (1985) proposed that one or more of these *Ubx* proteins is required for the execution of each of the identity functions of the *Ubx* domain. Subsequent observations on the effects of *abx*, *bx*, *pbx*, and *bxd* mutations on the metameric distribution of this protein family are consistent with this proposal (Hogness et al. 1985; White and Wilcox 1985; Cabrera et al. 1985; S.L. Helfand and D.S. Hogness, in prep.).

Here, we focus on the *bxd* region and, more particu-

[1]Present address: Division of Biology 156-29, California Institute of Technology, Pasadena, California 91125 USA.
[2]Present address: Department of Tropical Public Health, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115 USA.

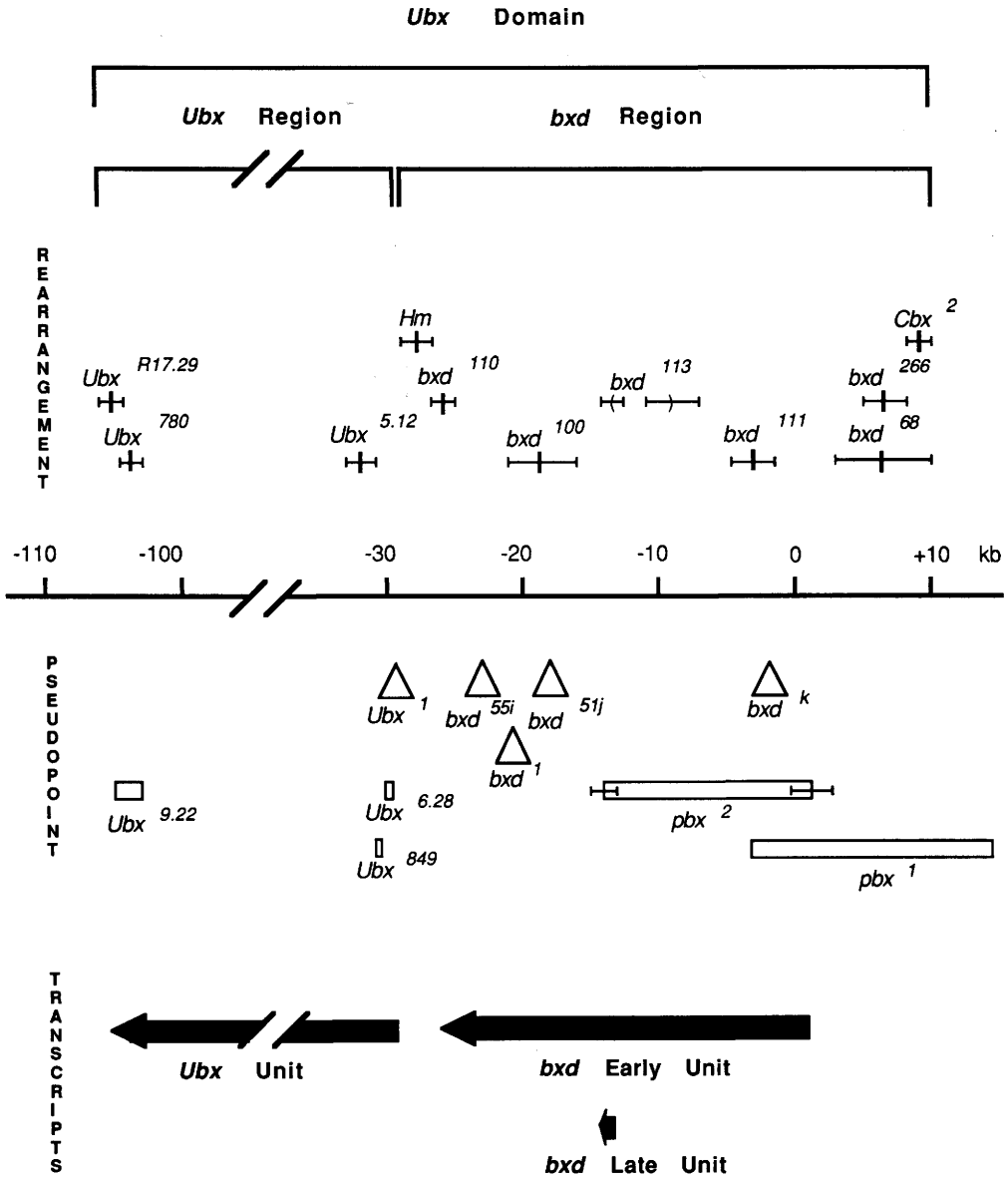

Figure 1. Molecular map of the *Ubx* domain with emphasis on the *bxd* region. The locations of rearrangement breakpoints are shown above the horizontal line which has the walk coordinates indicated in kilobases (Bender et al. 1983). The transposable element insertion sites (triangles) and deletions (open boxes) are shown below this line. The extents of uncertainty in map location are indicated by horizontal bars. All locations are as in Bender et al. (1983) with the following exceptions: Ubx^{1}, $Ubx^{6\text{-}28}$, Ubx^{849}, $Ubx^{9\text{-}22}$ (Akam et al. 1985); *Hm*, bxd^{266}, bxd^{68}, Cbx^{2} (Bender et al. 1985), and bxd^{110} (this study). The extent of the *Ubx* (Beachy et al. 1985), *bxd* early, and *bxd* late (this study) transcription units are indicated by the filled arrows below the molecular map.

larly, on its transcripts (Fig. 1). This region lies immediately upstream from the *Ubx* unit and appears to exert a *cis*-regulatory control over the quantitative and spatial expression of that unit. Curiously, the control elements within the *bxd* region, as defined by the phenotypes and locations of the *bxd* mutations, must be distributed over the entire 40 kb of the region (Hogness et al. 1985; Bender et al. 1985; S.L. Helfand and D.S. Hogness, in prep.). Furthermore, these *cis*-regulatory functions of the *bxd* region do not appear to be dependent upon its transcription (Hogness et al. 1985).

Transcription of the *bxd* region was detected by Northern blot hybridization experiments, which revealed that parts of the region are transcriptionally active both early (3–6 hr of embryogenesis) and late (midthird larval instar through the adult stage) in development (Hogness et al. 1985). These experiments and preliminary analysis of two early cloned cDNAs (3600 and 3601) at the level of Southern blot hybridization revealed that the early transcripts are polyadenylated, range in size from 1.1 to 1.3 kb, and derive, apparently by differential splicing, from a 27.5-kb region that we refer to as the *bxd* early unit. Late polyadenylated transcripts of about 0.8 kb were similarly detected with a genomic DNA probe from a region within the early unit —a region that we refer to as the *bxd* late unit (Fig. 1).

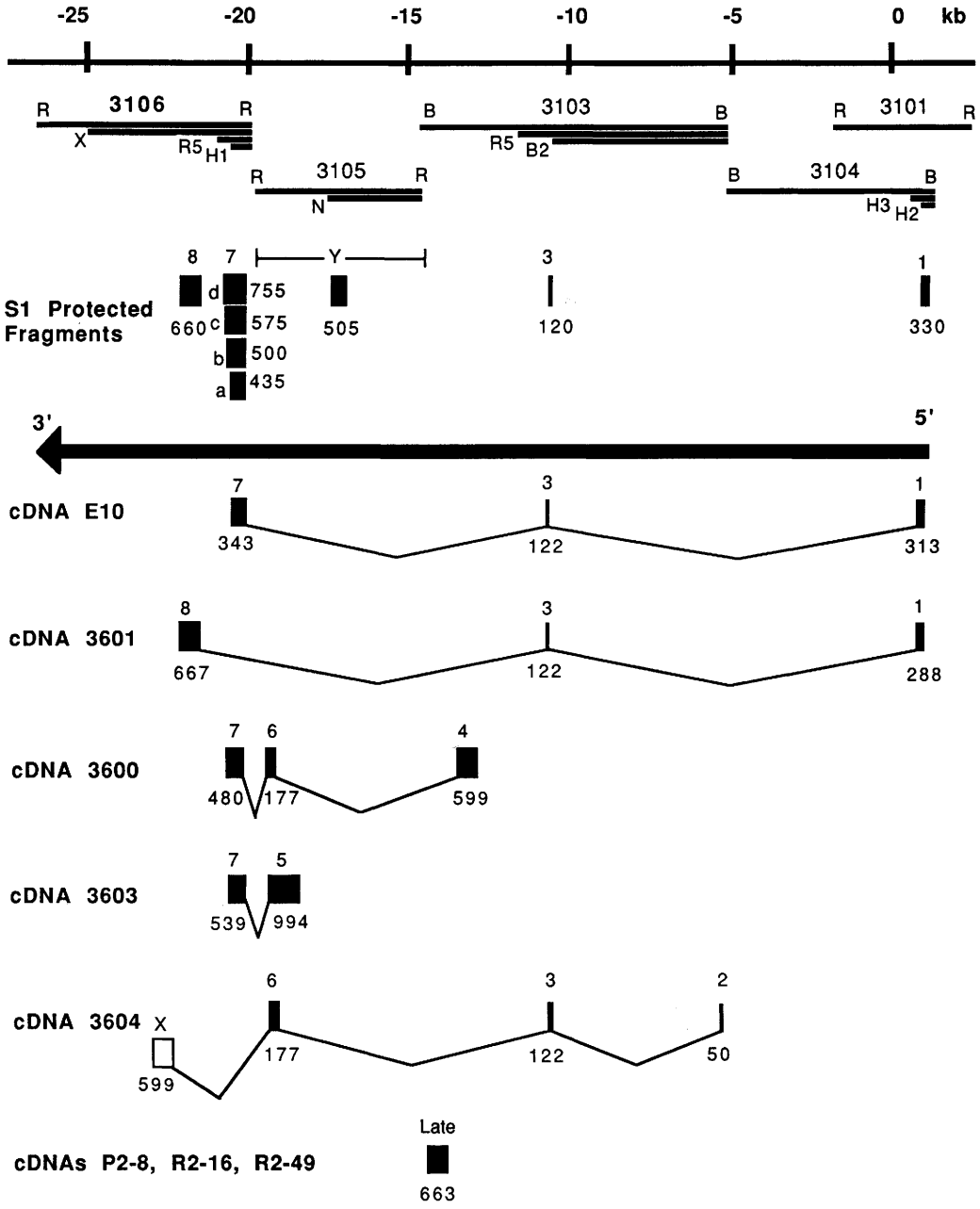

Figure 2. *bxd* early RNA S1 protected fragments and early and late cDNA exons localized on the molecular map of the *bxd* region. Map coordinates are as in Fig. 1. The exon numbers are shown above the filled boxes representing them. The sizes of the exons (in bp) are given below or to the side of the boxes. The filled arrow represents the extent of the early transcription unit. In the case of the cDNAs, shown below this arrow, all exons have been localized on the genomic DNA sequence (see Fig. 5) and sizes are given to the nucleotide. The location of the genomic clones and subclones used for S1 protection experiments against poly(A)$^+$ RNA are shown below the map coordinates at the top of the figure. The fragments protected are shown below these clones along with their calculated sizes. Their locations are based both on mapping with the deletion subclones used for S1 protection and on the assumption that they represent the exons (whose numbers are shown above the boxes) present in the cDNAs. Horizontal bars indicate the limits of resolution for the position of the Y fragment. The X exonic region of cDNA 3604 from the 42A region of chromosome 2 is shown as an open box. The E27 cDNA (not shown) maps entirely within genomic clone 3106, is 3 kb long and must represent a spliced RNA since it includes sequences derived from genomic fragments mapping greater than 3 kb apart: the *Eco*RI–*Xba*I fragment (−26.5 to −25 kb), the *Xba*I–*Eco*RV fragment, and the *Eco*RV–*Eco*RI fragment (−21 to −20 kb). Abbreviations: (B) *Bam*HI; (B2) *Bgl*II; (H1) *Hpa*I; (H2) *Hin*cII; (H3) *Hin*dIII; (N) *Nco*I; (R) *Eco*RI; (R5) *Eco*RV; (X) *Xba*I.

To define better the structure and possible functions of these *bxd* transcription units, we present here a detailed analysis of both the early and late transcripts. The results of S1 nuclease protection and primer extension studies, as well as the complete sequence of five early cDNAs, yield a complex picture of differentially spliced

early transcripts which, curiously, do not appear to possess protein coding potential. Similar analyses demonstrate that late *bxd* transcription produces a simple transcript that is initiated within an intronic region of the early transcription unit. This late RNA, in contrast to the early RNAs, is unspliced and contains an open reading frame (ORF) that potentially encodes a 101-amino-acid polypeptide.

Results

Complex splicing patterns of the early RNAs revealed by cDNA analyses

We have sequenced five early *bxd* cDNAs (cDNAs 3600, 3601, 3603, 3604, and E10). These cDNAs range in size from 778 bp (E10) to 1533 bp (3603). We have also determined the sequence of 35 kb of genomic DNA from the *bxd* region, which includes the early *bxd* transcription unit (D.A. Peattie, H.D. Lipshitz, L. Prestidge, and D.S. Hogness, in prep.). These analyses have made it possible to locate precisely all exons in the early cDNAs and hence to determine the exonic structure of the early *bxd* RNAs (Figs. 2 and 5). Eight exons (named exons 1–8), ranging in size from 50 bases (2) to 994 bases (5), are spliced differentially to produce the 1.1- to 1.3-kb early *bxd* RNAs. cDNA 3604 is unusual in that it includes sequences derived from chromosome 2 (see below).

A sixth early *bxd* cDNA (cDNA E27) has been characterized by Southern blot analysis (Fig. 2 legend). It includes sequences derived from the −25- to −26.5-kb region of the genomic map, placing the 3′ limit of *bxd* early transcription within 5 kb of the 5′ end of the *Ubx* transcription unit (Hogness et al. 1985). Together, these studies show that *bxd* early transcription extends over at least 26 kb of the 40 kb of *bxd* region DNA and that complex splicing patterns produce the *bxd* early RNAs.

Confirmation of complex early RNA structure by S1 nuclease protection and primer extension analyses

Because of the complexity of the splicing patterns revealed by the cDNA analysis, we have used S1 nuclease protection experiments to confirm that the cDNAs are representative of the early *bxd* RNAs as well as to reveal additional exons not represented in the six cDNAs. Single-stranded genomic DNA probes spanning the early transcription unit (genomic clones 3101, 3104, 3103, 3105, and 3106; Fig. 2) were used to protect poly(A)$^+$ RNA purified from 3- to 6-hr-old embryos. Given that *bxd* transcripts had previously been detected only with probes specific for leftward transcription (Hogness et al. 1985; Figs. 1 and 2), we used these same probes in the S1 protection experiments. Because these probes spanned large regions of DNA (5–10 kb), only abundantly represented exons could be detected. Eight exons were found (exons 1, 3, 7_a-7_d, 8 and Y; Fig. 2) ranging in size from 120 ± 10 bases (exon 3) to 755 ± 20 bases (exon 7_d). Figure 3a shows representative results obtained with the 3106 probe used for the detection of exons 7_a-7_d and 8. The difference in intensity among the protected fragments was reproducible and is a consequence of their different representation in the processed transcripts.

To map the exons at higher resolution, a series of deletions was generated unidirectionally from the 3′ end (relative to the direction of transcription) of each of the genomic clones. These deleted probes were then used in S1 protection experiments (Figs. 2 and 3). Seven of the eight exons defined by this analysis correspond with exons present in the cDNAs (exons 1, 3, 7_a-7_d, and 8; Fig. 2). The exception, exon Y, was not mapped accurately enough to determine whether or not it overlaps exons 5 and 6 (Fig. 2). Fragment Y is detected by 3105-derived probes, is present at roughly the same abundance as the other exons, and is of substantial length (505 ± 10 bases). However, the 3105 probe does not detect an RNA of discrete size on Northern blots of poly(A)$^+$ RNA, whether from the Oregon-R or Canton-S strains. One explanation for this result is that, although exon Y itself is abundant, it is present in a variety of RNAs that differ in size and each of which is relatively rare.

We determined that the 5′ end of exon 1 is the 5′ end of all *bxd* RNAs containing this exon by annealing a short DNA fragment from this region to the RNA and extending it with reverse transcriptase (data not shown). This 5′ end was mapped to the nucleotide on the genomic DNA sequence by electrophoresis of the S1-protected fragment and the primer extension product alongside a dideoxy-sequence ladder of genomic DNA from this region (data not shown, positions given in Fig. 5 below). Thus, E10 and 3601 are both nearly full-length cDNAs, respectively only 22 and 47 bp short of full length at their 5′ ends (Figs. 2 and 5).

To further establish the correspondence between the cDNAs and the structure of the early transcripts, four of the six cDNAS (cDNAS 3600, 3601, 3603, and 3604) were used to generate single-stranded DNA probes for S1 nuclease protection against the same 3- to 6-hr poly(A)$^+$ RNA preparation as that used in the experiments with genomic DNA probes. All *bxd*-homologous sequences in the cDNA probes represented exons, whereas as little as 1% of the *bxd*-homologous sequences in the genomic probes represented exons. Consequently, the signal-to-noise ratio was improved in the cDNA experiments.

Nineteen protected fragments were found. All of these fragments represent exons and exon-combinations found in the cDNAs (Fig. 4). In 15 cases, complete exons or exon-combinations were protected; in four cases molecular polymorphisms between the Oregon-R strain (from which the cDNAs were derived) and the Canton-S strain (from which the RNA was purified) led to base mismatches of sufficient extent (10 or more bases) that digestion occurred in these regions leading to fragments representing partial exons or exon-combinations (Fig. 4a,c and 5). The intensities of the fragments differed widely, again reflecting differences in the abundance of the exons or exon-combinations represented by them. Certain exons are found only in combination with others and so are not protected individually by the cDNA probes. For example, exon 8 is detected by the cDNA 3601 probe only in combination with exons 1 and 3, never alone; similarly, exon 1 is detected by 3601 only

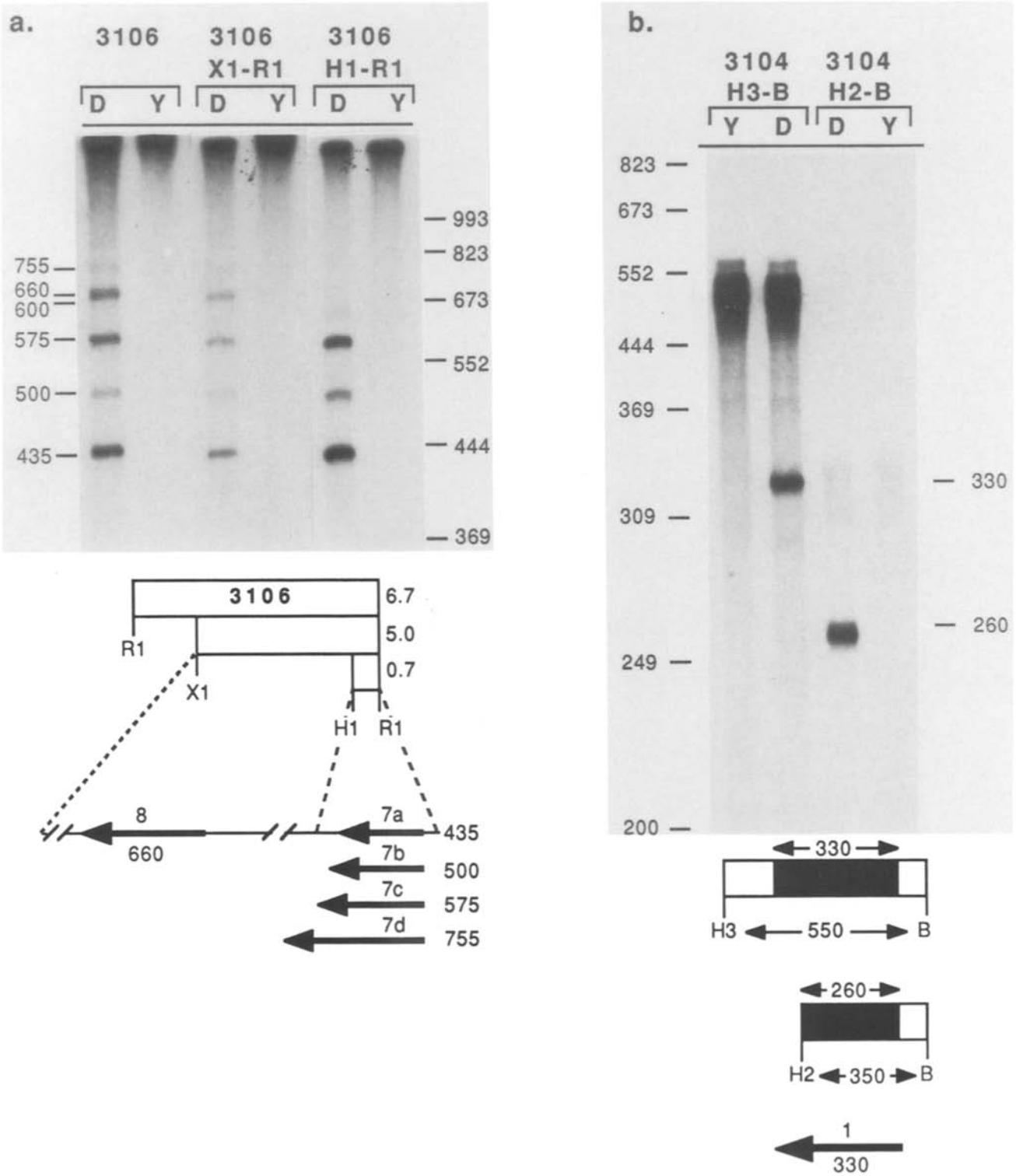

Figure 3. S1 nuclease protection analysis of the *bxd* early RNAs with genomic DNA probes. (*a*) S1 protection of the 3′ exons 7_a-7_d and 8 using single-stranded probes derived from the 3106 genomic clone. A total of 5 μg of RNA and 10^5 dpm of probe per lane were used. (D) *Drosophila* C-S poly(A)$^+$ RNA; (Y) yeast RNA (total). Marker sizes are shown to the right of the autoradiograph, calculated sizes of the protected fragments to the left. The results are interpreted schematically below the autoradiograph. Note that the *Hpa*I site is within exon 7_d; consequently a 600-base protected fragment is found in place of the 755-base fragment representing the entire 7_d exon. (*b*) S1 protection of exon 1 using single stranded M13 3104 probes: a 550-bp *Hin*dIII–*Bam*HI fragment (*left*) and 350-bp *Hin*cII–*Bam*HI subclone (*right*), each derived from the 5′ end of 3104 (Fig. 2). A total of 5 μg of RNA and 10^5 dpm of probe per lane were used. D and Y are as in *a* above. Marker sizes are shown to the left and calculated sizes of the protected fragments to the right of the autoradiograph. The 330-base exon could be localized to high resolution since the *Hin*cII site split it into a 260- and a 70- (data not shown) base fragment, as shown schematically below the autoradiograph.

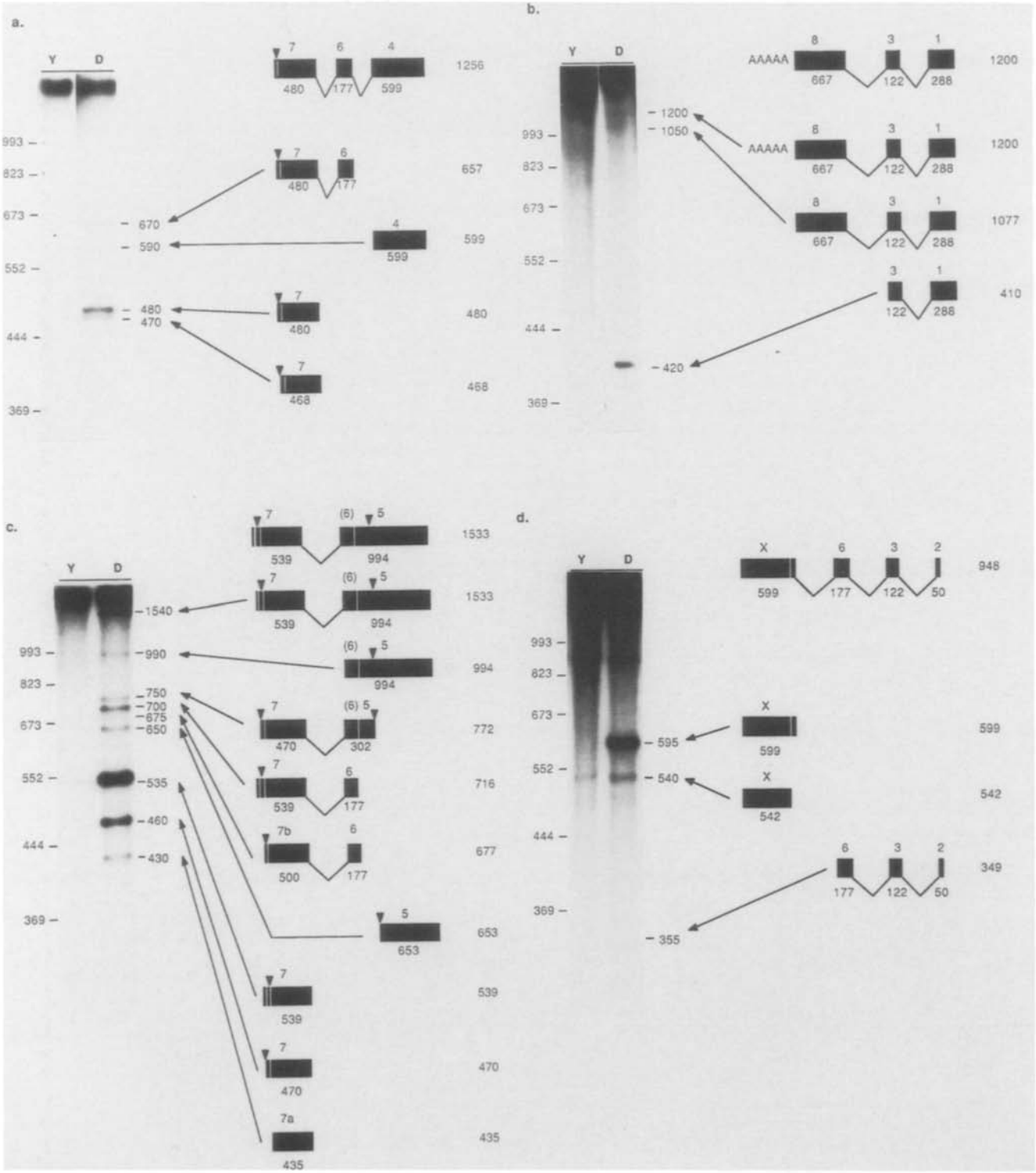

Figure 4. S1 nuclease protection analysis of the *bxd* early RNAs using cDNA probes. (*a*) 3600; (*b*) 3601; (*c*) 3603; (*d*) 3604. In each case single-stranded probes were used with 5 μg of RNA and 10^5 dpm probe per reaction. D and Y are as for Fig. 3. Marker sizes are indicated to the left of each autoradiograph. Two loadings were made of each reaction and different size standards were used to calculate the sizes of the protected fragments shown to the right of each autoradiograph. Only the second loading is shown in this figure. The hypothesized structures and exact sizes (determined by sequencing; see Fig. 5) of these structures are shown. The complete structure of each cDNA is shown above the hypothesized structures of its protected fragments. The inverted triangles above exons 5 and 7 indicate the locations of mismatches of 10 or more consecutive bases when the cDNA sequence is compared to that of the genomic DNA coding for the RNA used in these experiments (Fig. 5). Vertical white lines within exon 7 indicate the positions of alternative polyadenylation sites revealed by S1 experiments using genomic DNA probes (Figs. 2 and 3). The vertical white line in exon 5 represents the internal splice acceptor site of exon 6, and the vertical white line in exonic region X represents the location of a 58-nucleotide intron revealed by the sequence of, and S1 protection of, a genomic clone derived from the 42A region in chromosome 2 (data not shown; H.D. Lipshitz, unpubl.).

in combination with exon 3, never alone (Fig. 4b). This implies that exons 1 and 8 always splice to exon 3, but that exon 3 may splice to exons other than 1 and 8, which is indeed borne out by the sequence analysis of the cloned cDNAs (Fig. 2).

cDNA 3604 is unusual in that it includes sequences derived from region 42A of chromosome 2 (as revealed by in situ hybridization to polytene chromosomes) and detects an additional RNA of 2.3 kb on Northern blots. The junction between *bxd*-derived sequences and 42A-derived sequences occurs precisely at the splice donor junction of exon 6 (Figs. 2, 4d, and 5); thus cDNA 3604 may represent an in vivo example of *trans*-splicing (Konarska et al. 1985; Solnick 1985). We have isolated and sequenced a 42A-derived genomic clone containing the X exonic region and have found that the 542- and 57-base X exons (Fig. 4d) have adjacent (5′) invariant AG dinucleotides characteristic of splice acceptor sites. The preliminary results of S1 protection experiments using this genomic clone as probe are also compatible with the possibility that the fusion of 42A and *bxd* sequences in cDNA 3604 occurs precisely at the splice acceptor junction of the 57-base X exon (H.D. Lipshitz, unpubl.; data not shown).

In total, 26 of the 27 fragments protected by genomic or cDNA probes can be explained as protection of exons or exon combinations found in the cDNAs. These results confirm that the cDNAs represent actual spliced, polyadenylated transcripts.

The early RNAs may not be mRNAs

The complete sequence of the 11 exons present in the five sequenced cDNAs is shown in Figure 5 along with genomic sequences upstream of the exon 1 transcription initiation site and downstream of the exon 7 sequences present in the cDNAs. There is no obvious "TATA" consensus sequence (Goldberg 1979) at an appropriate distance (Breathnach and Chambon 1982) upstream of the exon 1 initiation site.

Analysis of the *bxd* cDNA sequences for ORFs and codon usage (using a table of codon preferences for *D. melanogaster* genes compiled by K. Burtis), indicated that there are only short ORFs in all three frames (Fig. 6) and that the codon usage is poor (data not shown). The longest AUG-initiated ORFs are: 17 codons for the *bxd*-derived portion of cDNA 3604, 38 codons for cDNA E10, 41 codons for cDNA 3600, 46 codons for cDNA 3601, 91 codons for cDNA 3603, and 102 codons for the 42A derived portion of cDNA 3604. This last ORF is contained completely within the non-*bxd* region-derived portion of the cDNA, has good codon usage, and may represent part of a translated ORF from the 42A region.

These data suggest that the early *bxd* transcripts may not be mRNAs. They may serve some other structural or catalytic function, or may instead represent nonfunctional processed RNAs (see Discussion).

The late RNA has a simple structure

In contrast to the early RNAs, the late *bxd* transcripts have a simple structure. On Northern blots, genomic clone 3103 (Fig. 2) hybridized to a 0.8-kb RNA band representing a late transcript produced from the midthird larval instar onwards (Hogness et al. 1985). We have carried out S1 nuclease protection experiments with poly(A)$^+$ pupal and adult RNA using single-stranded 3103 probes as well as probes made from unidirectional 3′ deletions of 3103 (Figs. 2 and 7). Both pupal and adult RNAs yield a single protected fragment 630 ± 20 bases long (Fig. 7). The size of this fragment relative to that of the polyadenylated transcript (0.63 vs. 0.8 kb) suggested that the late RNA comprises a single exon whose transcription initiation site lies within a intronic region of the early transcripts (Figs. 1 and 2).

The 5′ end of the 0.8-kb transcript was mapped more precisely by means of primer extension analysis. The primer used was a fragment of DNA derived from the R2-49 cDNA clone (see below), the 5′ end of which was 129 bases downstream of the 5′ end of the cDNA (Figs. 7 and 8). The difference in size between the extension product and the primer (110 ± 20 bases) confirmed that the late RNA consists of a single exon (Figs. 2, 7, and 8).

The late RNA may be a mRNA

Four cDNA libraries made from early pupae, late pupae, adult males, and adult females (L. Kauvar, unpubl.) were screened using the 3103 genomic clone (Fig. 2) as probe. Three cDNA clones were isolated, one (P2-8) from the early pupal library and two (R2-16 and R2-49) from the adult male library.

The sequences of the three late cDNA clones were determined and compared to the corresponding genomic sequence (Fig. 8). The cDNA sequence is coextensive with the genomic sequence, and consistent with the S1 and primer extension analyses. All three cDNAs have poly(A) tracts at their 3′ ends. The sizes [excluding poly(A) tracts] of the cDNAs are 598 bp (R2-16), 633 bp (P2-8), and 663 bp (R2-49). R2-49 is full length (see S1 protection and primer extension analyses above). Analysis of the cDNA sequences for ORFs with good codon usage for *D. melanogaster* (see above) revealed a 303-bp (101 codon) AUG-initiated ORF with excellent codon usage beginning at +116 and terminating at +418 (Fig. 8). Again there is little, if any, indication of a TATA sequence at the appropriate position upstream from the transcription initiation site.

Discussion

The early and late periods of *bxd* transcription stand opposed on almost every count. The early period is short, lasting only a few hours after its initiation during syncytial blastoderm; the late period long, lasting several days after initiation during third larval instar. The early transcription unit is long, occupying at least 26 kb; the late unit short, comprising 0.66 kb of DNA within an intron of the early unit. The early transcripts are complex, consisting of different combinations of 11 exons typically taken three at a time; the late transcripts are simple, formed from the primary transcript without splicing.

Exon 1 and 5' Upstream Sequences:

```
          -50                                                +1
Genomic     cgtgcatgttttaggatctccagggatcggcggctgatgttttgtaaacaCTCAGTATGAGTTCCAGTTCCCGGGTGCTTGGACGTTGCGGATCGGCTT
3601                                                          **  *                                          AG
         +50                                               +100
Genomic     AAAAAACGCATCGGAAGCGAGAATATTCTGTTTGCCAGTATTTCGAAATGAGTTAAAATGTGTGCAAAATATGATAAAAATTAATTAATTCGCCATAAAT
3601        CCGAT.................................C.A.....................................AT..------.........
        +150                                               +200
Genomic     GCGCAATCTAATTGGGCTGCTATCCATGAAAAAGTGTTAAGTGTAGTGTATCCTATGCAACGAAATGGGTCCATGTGTTAAAGATGTGCTTGCCAAGTAA
3601        .-...TA-----.....................................................................................
        +250                                               +300                                          +347
Genomic     AAAAAACCGAAAAGTAGAGGGGGTTAACAATAATTCAGCAAATGAATTGAATAAAGGATTCTACGGTAATATTTCTGCACATATTTTCAGAACAACATgt
3601        .....................................................................................
Genomic     aagtacct
3601
```

Exon 2:

```
                 +1                                                +50
Genomic      gcggaacaagACTCTATTCAGCGTAAATGGCAATTCGTAAAAAAGACATAAAGAACACAGgtatggccaa
3604              C.G.....................................................
```

Exon 3:

```
                 +1                                                +50
Genomic      tcaattacagAGCACTCACACAGCCAATAACCAGTCCGGCCTCCAGATTGCCACTCCAGGATACCAAATATTACACAACTCCAGAGGACTGTTCTCCAT
3601              .............................G.....................................................
3604              .....................................................................................
           +100                +122
Genomic    TTTGGGCACATTTAAAATTCAGAACGCGTTGAGgtgagtccac
3601       ..............................
3604       ..............................
```

Exon 4:

```
                 +1                                                +50
Genomic      acacccgccaGAATTCAGTTTTGCACCACAGTTAGCACTGATTTATGGGCAGTCGAATAGGGTTTTAAGGTGGGCTTTAAGCTAAGCTAGCAATTGGAA
3600              .....................................-................-.....................
           +100                                              +150
Genomic    TTAGTTTAAGTAGCCGGCGGGAAGTCAACGACCTGCGACCAGTCGGCAGTTCAAAGTCAAAAAGTCAATTATACGCGATACCTAAAAATACAACAGCGGG
3600       ..................C.................-...--.....................-.....................
           +200                                              +250
Genomic    TGGAGTGAAAATCTCGTTAAAAGCCAATAACAGACGCCATTAAATGTGCATGAAAATTAAATGAAAACGGCTTGATGATGAGGTTGCCTGCGGGAGGAGG
3600       .....................................................................................
           +300                                              +350
Genomic    AACTCGTAGCAGCCACGCGACCACTTGAACTTGGGCGGGAATATCAGTATTTGTGTGCAGTAATTCGAACTGACCCTCGTCGAGCATTCAGATCCAATTG
3600       ..........................................................-..........................
           +400                                              +450
Genomic    GAGGGCAAGTTCAATGGCATTGGGGATTCCAGGGGGAGTGGGGATCGAGGGCCGCAGCGAGATAAGGGAACTTCGATAAGTTTAATTTATTCGACCCCAC
3600       .....................................................................................
           +500                                              +550
Genomic    TGCCATCTGTTGCGATATAACCAACACTTATTTAATCAACGCGATAGGAATAGGCTCGTGCTGTAATTATAGATCGGATTAAAAACCATCGAACATGGAG
3600       .....................................................................................
           +600  +606
Genomic    GCTCAATTCAAAATGAGgtaaatgaaa
3600       ...........
```

Exons 5 and 6:

```
                 +1                                                +50
Genomic      acttgttgttTTTAATAAGTTTCTGGCTGTTGGCTGTTTGTTGTTGTAGCTCTGGCTGGCTTTCTTCCTTTGATGTTTCGATTAAGTTCGGAGTTCGCG
3603              AAC..C.....................................................................................
           +100                                              +150
Genomic    TTGCCGAACTGGTTGGTTAGGTCCCCTTTCCGTCCTCTCACCCCCTTCTGGGAACCATTCCCATCAGAGGGAGCATTTTGGCTTAAGTTTATGGCCATTT
3603       ............T.G.......--.............--.....................................................
           +200                                              +250
Genomic    CGTTCGTTTCACTGTCTTTGCTTTTTGGCGCTGCTCCTTGTTGGCTTCATTTTCTTGAGCATTTCACCCATTTGCCATTTACCTTTGCTTATTTGGGCTT
3603       ..................................T..................................................
           +300                                              +350
Genomic    GTGCTGGAAATTT-CTCACATCGCCGCTCTGACGCAACTCGGTTGCCGATTGGCC-GCTCCACTTCACTCCCCATCGTTCCGACCAACTCCCTATCAGCC
3603       ............T.....................-.............C.....................................
           +400                                              +450
Genomic    TATCCACCCACCGCTCGCAAATTTCGAATTT-ATTTTATTGTTATGATTTCAATTGTTTTCGCAACGTTAAACACTAGAGATTTCCAAAGCAGACCATGT
3603       ............................G.....................................-.....................
           +500                                              +550
Genomic    CGAGTTGTCGGTTAAATAATTTTCAATTTTTATTACCCTCCGCCAGAGAGGTTTTGCATTCAGCGCATACTGTTCCGAGTGACTCAGACCAAGGGGCTTT
3603       ..................................G.....................T.............-.....................CG
           +600                                              +650
Genomic    GAAAATAAGAAACCAAAGTTGGGATACCCAACTCTCGAAGGAACAATCCAAATCTATCTAGCATGCAAAAA----------GTT-------
3603       ............C.....................................................TTGTATGTAGCTAAAATTC...TTTTAAG
           +700                                              +750
Genomic    ----------GTGGTATTAATCATTGCCTCAAGTTGATGGATATGTGTTATTAGAAAAAC-TTGCACATTTTCTTCGAGAAATGTATTGATTC------G
3603       TAATAGCATA.............TC...-.....T............AA.A.....AT...A.A.....................AACCGAATCAAT.
           +800                                              +850
Genomic    GTTATTTGCTCGCATAATTATTATCGATCTTATTAGCATTCGCAGACATTTTGCCACAAACAAGTCACATTAGAACGCCCAATTTTCACGAGAACAATTA
3603       ..................................................T.......G.............A.........-.............
3600                                        ..............T.......G.............A.........-.............
3604                                        ....T.....................................-.............
```

```
                +900                                                  +950
Genomic     TGACCGTGGCACCGCCGAAGGTGATCCTAATAATTGCACCAATATGCTTCGGTAATCGTTTTGAGGCCAGACTCTCCCAGGACTCGAAGGCCGCCTTCCC
3603        .....-.....................................................................................................
3600        .....-.....................................................................................................
3604        .....-.....................................................................................................
                  +1004
Genomic     GAAACTACAAAAAAGgtatagattg
3603        .........................
3600        .........................
3604        .........................
```

Exon 7 and 3' Genomic Sequences:

```
                  +1                                                  +50
Genomic      tacactaaagCACATTATCACT-CCCATTGGACCGAGATGACTTCATTTTCGTGCCATATGCCTGTGCCGGTGCTGTACGTGTGGTTCCACCCGCTCCT
3600                  ............T.....................................................T...................
3603                  ............T.....................................................T...................
             +100                                                 +150
Genomic     CTTAAGTCGCTGCCATGATTTACTACAAGCAGATGATTTAATATAACCGCATATCTGGATATAGAAACACGCACACACCCGGCCAGAGTGCACTTCAGAT
3600        .....................................................................................................
3603        .....................................................................................................
             +200                                                 +250
Genomic     GCATTCGACTGTATTCGACTGCAGTGGAAACAGGAACCCGTTGAACTGGCCCACAAACAGAGGCAGACACCCCACACATGTTGTCAGTGAAGCGTCCGTT
3600        ........................C....................................................................................
3603        ........................C....................................................................................
             +300                                                 +350
Genomic     TTGGAGGAAAAACGGGACGACTGTTAAAGGGCCTCCATCTTGAAAACAGACGCCTCTAAGAGGCGCG--TCCGCAAGACGATAACTCAATTGAATATTTC
3600        ......................T.............................................GC--.....................................
3603        ......................T.............................................GC.....................................
             +400                                                 +450
Genomic     AACAAAAAAGCAATTGTAAATGGGTTAATACTTTTAACTATACTAATACTGTAACTAAAGGGTAATTTCAATATAATTATAGATTGTGATCG--------
3600        ........-.................................--------.....................CT....C..........ACAGATTGTC
3603        ........-.................................--------.....................CT....C..........ACAGATTGTC
             +500                                                 +550
Genomic     -GCTAGAATTAAAATATTTCCTTTTATTTATTAACATTATATTATGATCGAGAAAATATGTAATAAATAGTTTCTCCTGGGTTAAAATCAAAAAAATTTC
3600        A.
3603        A.....................................................
             +600                                                 +650
Genomic     ATATTACATTGTTTCGATTTTCTTAGCTATCAACCAAGTTAACTAGTGGAATAGTTGGCTTTGAAAACAAATAAAACTAAACAAAATAGAAGCCATCAAA
             +700                                                 +750            +767
Genomic     ATGAACTTCTGGCAGTGCGTAAATCTAGGCGTCAACGAGCTGCGACCTGACCGAAGGGCTCATCAAGATGCTGGCCAA
```

Exon 8:

```
                  +1                                                  +50
Genomic      ttccattcagCTATATTTTAGGCAAGATGTCCGAGCCATCCCCAGCACGATTCCGGTCATAATTCTTTTGGTTTGATGACCGCATTCCCACCTCCAAAA
3601                  .....................................................................................................
             +100                                                 +150
Genomic     CAAACGGCTAAGGCGCCGCGGTAAAAATTTGTTTGCGACTTGATGGCTTAATCCACCGCAAAGTGGCTCAGTGGTGCCTGGTATGGGCTCATTACGGGTC
3601        .....................................................................................................
             +200                                                 +250
Genomic     AGAAATCAATTGTGGTCCGAAATCGCTGTGAACCAACTGCTTATGCAAAGCATGCCATTCAAACAGAGCAAAACAATGGAGAGCGAAAGAGACCAGGCCA
3601        .....................................................................................................
             +300                                                 +350
Genomic     AAGAAATTCCTCGGACACAATGAAGACATTTGCTGCGAACCATACTACCTACTAATCCTGGCAATGAACCCTTGACTTTTGGCCGCAATTCATCAAACAA
3601        .........G.................T.....................................................................................
             +400                                                 +450
Genomic     TTGGCCGACCATTTGCATGCGGTGGTTGGCAGCTCTTCATTTTAAAACTCCCTTTCCCGTATCTGCATCCTAATCCGGAGGACTGAGGATTGCGGATTGA
3601        .....................................................................................................
             +500                                                 +550
Genomic     GGGCGACTAATTGAGCACTTTTGTTTGAGTGCCGTGCTGCGCGGATTTCGCCTACTTGTTAGTTGCTGCCCAATAACATCGCTCTCCTTCCTCATGCACT
3601        .....................................................................................................
             +600                                                 +650            +667
Genomic     GCGGAAAGGATATTTGATTTTAAATAAAGCCTCTCAATTAATTGTATCCTAAAAATATACGAAATTACAAAAAAAATAtatatattt
3601        ....................................C.....................................AAAAAAAAA
```

Figure 5. Complete sequence of the *bxd* early exons. Exonic sequences are in upper case; 5′-upstream sequences and intronic sequences are in lower case. Only sequences confirmed on both strands are shown; the genomic sequence is given above and the cDNA sequences are aligned below. Genomic sequences are from the Canton-S strain, while cDNA sequences are from the Oregon-R strain. Nucleotide differences are shown and gaps are indicated by dashes. Where the cDNA was not completely sequenced on both strands (E10), the relevant data are indicated by symbols: in exon 1, nucleotide 23 is underlined and represents the 5′ end of cDNA E10; in exon 7, nucleotide 344 is underlined and represents the 3′ end of cDNA E10. The absolutely conserved splice donor (gt) and acceptor (ag) nucleotides at the exon boundaries are underlined, including the internal acceptor of exon 6 (nucleotides 824 and 825 of exon 5). The location of the 5′ end of exon 1, as determined by both primer extension and S1 protection analysis, is indicated by asterisks, the major initiation site being accented. Nucleotide numbering represents the maximum length of each exon: In reality the gaps reduce these lengths (see Figs. 2 and 4). Mismatches of 10 or more nucleotides between the genomic and cDNA sequences (inverted triangles in Fig. 4) occur between nucleotides 661 and 699 of exon 5 and nucleotides 480 and 490 of exon 7. In exon 7, 3′ genomic sequences are appended and the putative polyadenylation sites (±10 nucleotides) based on S1 protection experiments (Figs. 2 and 3) are 439 = 7_a, 511 = exon 7_b, 587 = exon 7_c, and 767 = 7_d.

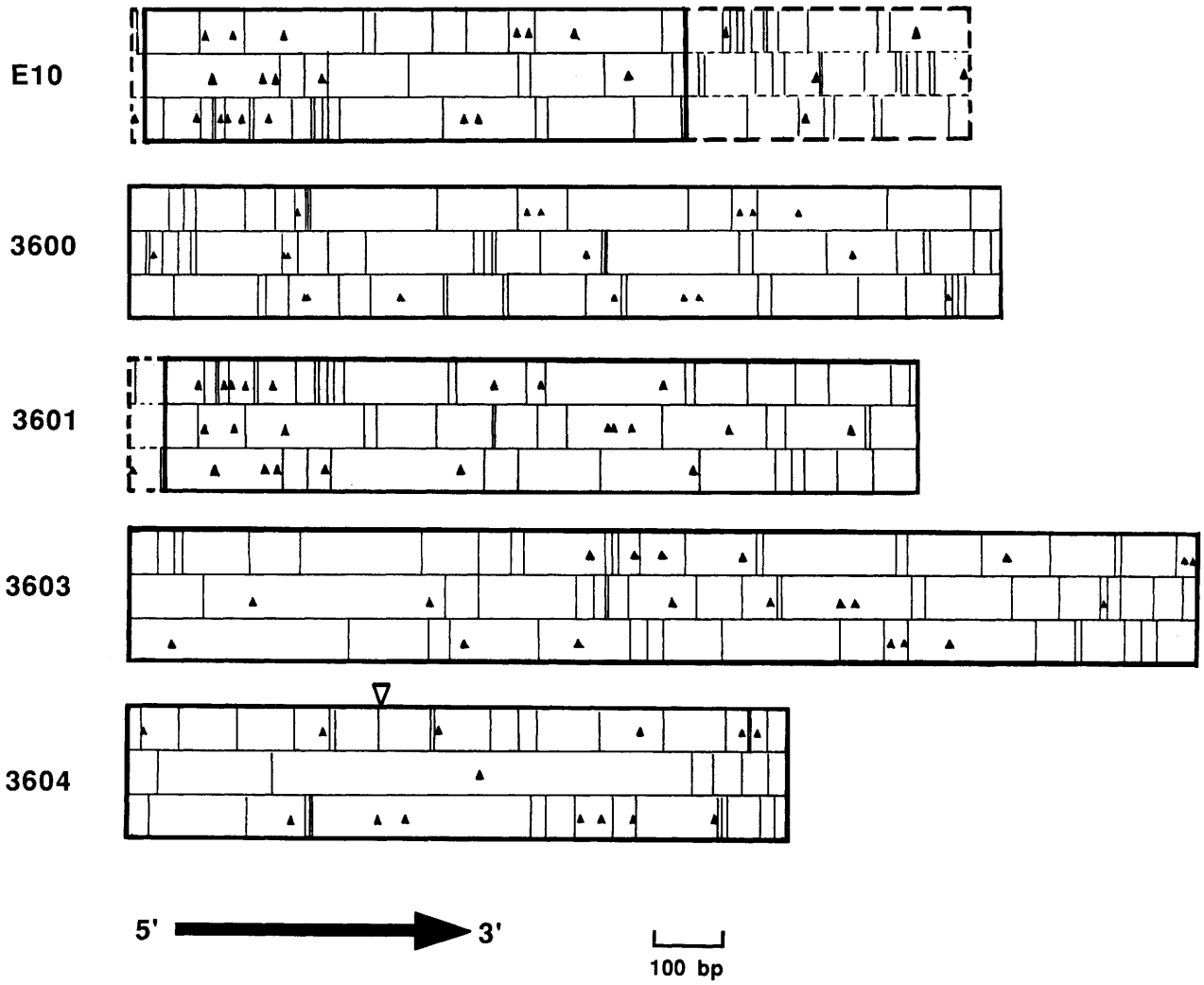

Figure 6. Open reading frames in the early cDNAs. Filled triangles represent initiation codons, vertical lines represent termination codons. The 5′ and 3′ sequences not present in the cDNAs have been appended to cDNAs E10 and 3601 (dotted lines) for completeness. To avoid redundancy, 3′ sequences were not appended to cDNAs 3600 and 3603. The unfilled inverted triangle above cDNA 3604 indicates the junction between *bxd* exon 6 and exonic sequences derived from the 42A region in chromosome 2, which lie downstream of this triangle. The only long ORF with good codon usage in any of the cDNAs occurs downstream of this junction.

The early RNAs do not possess significant coding potential; the late RNAs do.

These disparate properties of the early and late units suggest they have quite different functions. What might these functions be? Here we consider possible answers to this question, starting with the simplest unit.

Late unit functions

The transcribed sequence in the poly(A)$^+$ RNAs of the late unit can be divided into three regions: a 5′-proximal region of ~115 nucleotides that is devoid of AUG triplets, a central region consisting of an AUG-initiated ORF of 101 codons, and a 3′-proximal region of 217 or 245 nucleotides, depending on which of two polyadenylation sites is used (Fig. 8). Given the significant length of this ORF and the excellent fit of its codons to the codon usage frequencies for *D. melanogaster* structural genes, there is little doubt that these three regions correspond, respectively, to the 5′ noncoding, the coding, and the 3′ noncoding regions of mRNAs encoding a protein of 101 amino acids. The late unit therefore appears to define a structural gene that we shall refer to as the late *bxd* gene.

The larger question of whether the late gene is a homeotic gene whose appropriate temporal and spatial expression is required for specific metameric identity functions remains open and a focus for speculation. Given the location of this gene within the BX-C and its late expression, the most obvious speculation is that the adult ps6 identity functions require both the late protein and one or more of the *Ubx* proteins. According to this model, mutations within the *bxd* region could alter the adult ps6 identity either by interfering with the expression of the *Ubx* proteins via disruption of the appro-

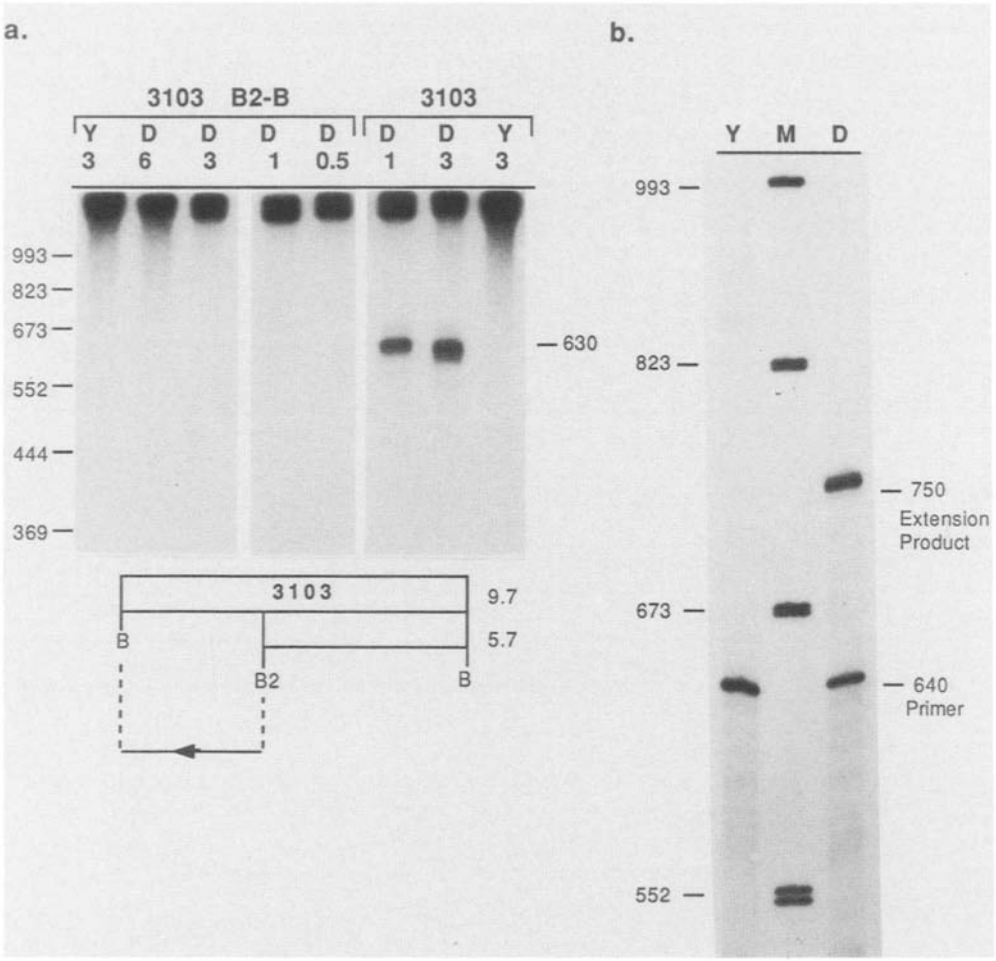

Figure 7. S1 nuclease protection and primer extension analysis of the late *bxd* RNAs. (*a*) S1 nuclease protection of a 630-base exon using single-stranded 3103 probes whose structure is schematized below the autoradiograph. Amounts of RNA and probe and abbreviations are as in Figs. 2 and 3. The number above each lane is the number of units of S1 nuclease ($\times 10^{-3}$) used. (*b*) Primer extension analysis of the 5′ end of the late RNAs. The primer was a single-stranded M13 probe in which the *D. melanogaster* sequences extended from the 3′ end of cDNA R2-49 to the *Xho*I site (nucleotide +130, see Fig. 8). Abbreviations are as in Figs. 2 and 3; (M) size standards, whose sizes are shown to the left of the autoradiograph. cDNA R2-49 appears full length, by this analysis.

priate *cis*-regulatory control of *Ubx* transcription, or by altering the structure or expression of the late protein. Mutations of the first kind should not be complemented by *Ubx* mutations and would be expected to alter the identity of ps6 in embryos and larvae, as well as in adults. By contrast, mutations of the second kind should be complemented by *Ubx* mutations and alter the adult, but not the embryonic and larval ps6 identities.

The *bxd* mutations appear to be of the first kind: they are not complemented by *Ubx* mutations; they alter ps6 identities in embryos, larvae, and adults (Lewis 1963, 1978, 1981, 1982; Bender et al. 1985); and they alter the quantitative and spatial expression of the *Ubx* proteins (Beachy et al. 1985; Hogness et al. 1985; S.L. Helfand and D.S. Hogness, in prep.). The *bxd* mutations would not, therefore, be expected to provide useful test systems for the above model. The two *pbx* deletions (Fig. 1) are more interesting in this regard as their ps6 phenotypes appear to be restricted to the adult (Lewis 1963, 1982 and pers. comm.). Most interesting is the pbx^2 deletion, since it is largely complemented by *Ubx* mutations (Lewis 1982), indicating it exhibits a lesser *cis*-regulatory defect than other mutations in the *bxd* region. Furthermore, the left end of the pbx^2 deletion has been mapped to a 1.5-kb restriction fragment that includes the late gene, and Northern analyses have failed to detect late mRNAs in pbx^2 homozygotes (H.D. Lipshitz, unpubl.). Part of the pbx^2 phenotype may therefore result from the lack of expression of the late gene—a possibility that can be tested by P-element-mediated insertion of a fully functional late gene.

Definition of the function of the late gene will, of course, also derive from the properties of the late protein, including its intracellular and metameric distributions. At present, data concerning this protein are limited to its deduced amino acid sequence, which exhibits no obviously unusual features and shows no significant homology to the sequences of proteins encoded by other

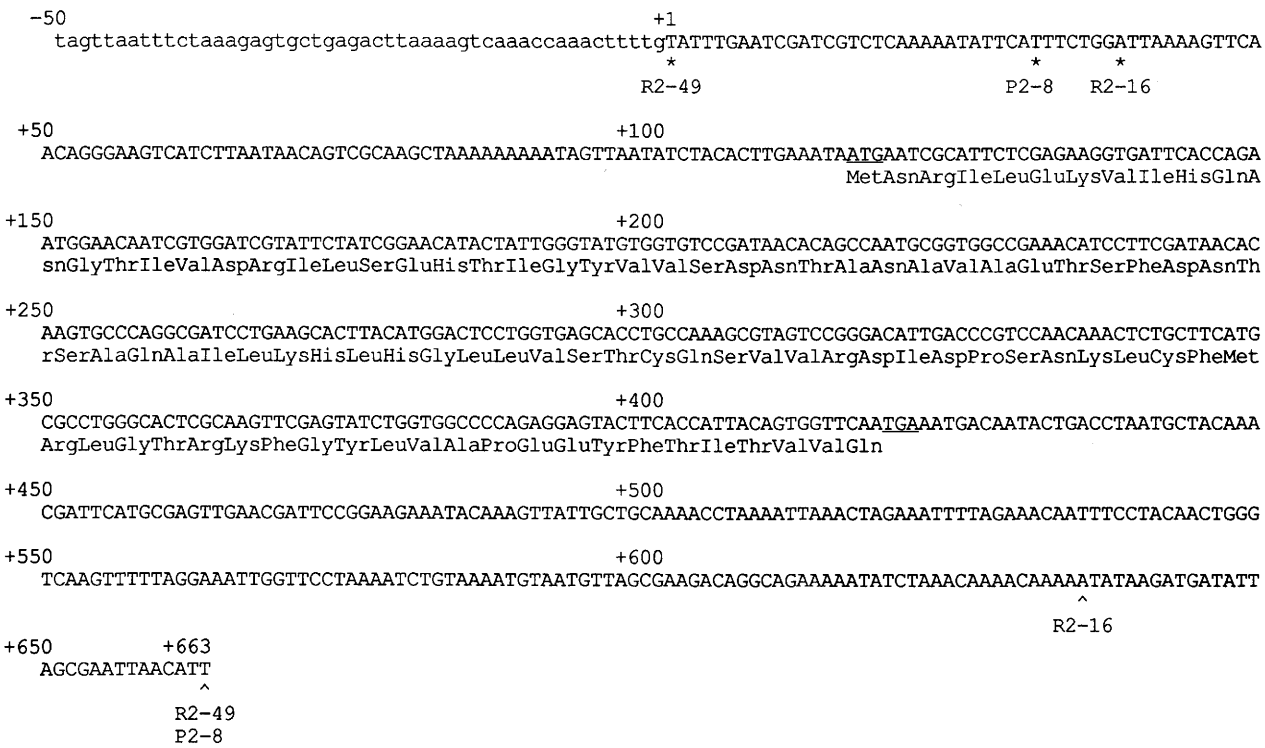

Figure 8. Complete sequence of the *bxd* late transcription unit. Transcribed sequences are shown in upper-case letters and upstream genomic sequence in lower-case letters. The 5′ and 3′ ends (* and ^, respectively) of the three cDNAs are indicated. The conceptual translation of the 101 codon ORF is shown below the DNA sequence. Only the sequence of the genomic DNA is shown, since the cDNAs were not sequenced completely on both strands.

homeotic genes, or of any proteins in the NIH sequence bank (SNBRF). We are currently attempting to isolate the late protein and raise antibodies against it in order to better define its function.

Early unit functions

We consider here three possibilities regarding the functions of the early *bxd* transcripts:

1. The complex set of processed transcripts from the early unit include mRNAs that are translated into polypeptides despite the short length and poor codon usage of their AUG-initiated ORFs. While we think this possibility unlikely, a precedent for poor codon usage can be found in the *dnaG* gene of *Escherichia coli*. In this case, it has been suggested that the poor codon usage provides a mechanism for the down-regulation of the *dnaG* primase relative to proteins encoded by other genes in the same operon (Smiley et al. 1982; Konigsberg and Godson 1983; Lupski et al. 1983).
2. The early transcripts are not mRNAs, but rather interact with the pre-mRNA or the mRNA of other genes to regulate processing or translation, respectively. Precedent for the involvement of intermolecular RNA–RNA interactions in the processing reactions derives from recent work implicating the small nuclear RNAs (snRNAs) in the splicing and polyadenylation reactions (Black et al. 1985; Krainer and Maniatis 1985; Moore and Sharp 1985), where they appear to interact with the pre-mRNA by a base-pairing mechanism (Zhuang and Weiner 1986). In cases such as the *Ubx* pre-mRNA, which can yield at least five different mRNAs by different splicing pathways (Beachy et al. 1985; Hogness et al. 1985; K. Kornfeld and D.S. Hogness, unpubl.), one can imagine that the early *bxd* RNAs might interact with the pre-mRNA to facilitate or inhibit one or more of the pathways and thereby favor the formation of a subset of the mRNAs.

Alternatively, one can imagine that the early *bxd* RNAs interact with the mRNAs to regulate their translation by a "hybrid-arrest" mechanism such as that proposed for certain prokaryotic genes whose transcripts exhibit sequence complementarity to the RNAs whose translation they regulate [*micF* (Mizuno et al. 1984); p*OUT* (Simons and Kleckner 1983); *NR1* (Womble et al. 1984)]. We have reported a sequence complementarity between the early *bxd* and the *Ubx* primary transcripts (Hogness et al. 1985), but these sequences reside in introns and their functional significance is untested.

It is difficult to evaluate the possibility of such *trans*-regulatory functions for the *bxd* early RNAs—particularly because possible *trans* functions of these RNAs cannot be inferred from the phenotypes of existing mutations in the *bxd* region, given the phenotypic dominance of their effects on the *cis*-regulatory functions of the *bxd* DNA (Hogness et al. 1985). We are attempting to resolve this difficulty by inactivating the *bxd* early RNAs with antisense RNAs produced in vivo by P-ele-

ment-mediated insertions, leaving unaltered the *bxd* DNA.

3. The third possibility is that early *bxd* transcription has no function, reflecting instead a change in state of the *bxd* chromatin resulting from the *cis*-regulation of *Ubx* transcription. A possible example of transcripts that exist only as functionless by-products of tissue-specific activation of a locus, is the I_μ transcripts produced from the $J–C_\mu$ intronic region of the μ heavy-chain immunoglobulin locus (Kemp et al. 1980; Alt et al. 1982; Nelson et al. 1983; Lennon and Perry 1985). In this case it has been proposed that the Ig-enhancer (Banerji et al. 1983; Gillies et al. 1983) activates a cryptic promoter in the $J–C_\mu$ intron, resulting in the production of spliced, polyadenylated, discrete-sized I_μ RNAs (Lennon and Perry 1985). Like the *bxd* early RNAs, the cryptic I_μ exon lacks a long ORF and is multiply closed in all three reading frames (Lennon and Perry 1985). No function for these transcripts has been demonstrated; it appears possible that the I_μ RNAs, rather than playing a role in the expression of the heavy-chain Ig locus (Lennon and Perry 1985; Yancopoulos and Alt 1985), instead represent by-products of the activity of the Ig enhancer, which has been reported to be capable of regulating promoters from a distance of as much as 17.5 kb (Wang and Calame 1985).

In the case of the *bxd* region, control elements regulating the *Ubx* promoter appear to be distributed as far as 40 kb upstream from that promoter (Hogness et al. 1985; S.L. Helfand and D.S. Hogness, in prep.). In analogy to the I_μ transcription, one can imagine that activation of one or more of these control elements by *trans*-acting regulatory proteins not only modulates transcription from the *Ubx* promoter but also activates a cryptic promoter located some 30 kb upstream from the *Ubx* promoter at the 5′ end of exon 1 in the *bxd* early unit (Figs. 1 and 2). It is also possible that other cryptic promoters within the unit can be activated. While we have demonstrated that the 5′ end of all early RNAs containing exon 1 is the 5′ end of exon 1, we have not shown that all of the processed early RNAs contain exon 1; thus, the 5′ ends of the RNAs represented by cDNAs 3604, 3600, and 3603 could be at the 5′ ends of exons 2, 4, and 5, respectively (Figs. 2, 4a,c,d), and could result from activation of additional cryptic promoters at these positions.

Why should the activation of cryptic *bxd* promoters be restricted to a short early embryonic period, when *Ubx* transcription is not similarly restricted? This question begets another: Is the regulation of *Ubx* transcription divisible into early and late periods that employ qualitatively different populations of regulatory proteins, and does the early period correspond to that for *bxd* early transcription? Two periods of this sort are indeed suggested by recent data on the effects of mutations in the segmentation and homeotic genes on *Ubx* transcription. [See Ingham and Martinez-Arias (1986) for a summary of such data.] During the first period, corresponding approximately to that for *bxd* early transcription, the spatial distribution of *Ubx* transcription appears to be determined by proteins encoded by segmentation genes that are transiently expressed during early embryogenesis. Subsequently, control of *Ubx* transcription appears to pass to other regulatory proteins, including those encoded by homeotic genes. One can therefore imagine that the activation of cryptic *bxd* promoters results from protein–DNA interactions that occur during the first, but not the second, of these two regulatory periods. Similar arguments can be made to explain the observation of Akam et al. (1985) that the spatial distribution of *bxd* early transcription does not completely overlap that of *Ubx* transcription.

It would clearly be of interest to determine whether those segmentation mutations that alter early *Ubx* transcription also alter *bxd* early transcription, and whether such alterations are correlated. We are also investigating the *cis*-regulatory activity of the *bxd* region in more detail by means of P-element-mediated cosmid transformation of constructs that fuse the *Ubx* promoter and the *bxd* region upstream from it to reporter genes such as *E. coli lacZ*.

Finally, we note that we have already presented arguments against a fourth possibility (Hogness et al. 1985): Namely, that although the *bxd* early transcripts are functionless, the *bxd* early transcription itself is required for appropriate *cis*-regulation of the *Ubx* unit. Thus, if the *bxd* early transcripts have no *trans* functions (i.e., possibilities 1 and 2 are not valid), we think *bxd* transcription has no function, *cis* or *trans*, and is simply an indicator of *Ubx* regulatory reactions. If this is the case, then the functional products of the *Ubx* domain would be limited to the proteins encoded by the *Ubx* unit and the late *bxd* gene, which at present yield a total of six different proteins.

Materials and methods

Strains

Wild-type *Drosophila melanogaster* Canton-Special (C-S) and Oregon-R (O-R) were maintained in large population cages. Plasmid vector pϕX was provided by R. Mulligan and is described by M. Goldschmidt-Clermont, R.B. Saint, and D.S. Hogness (in prep.). Phage cloning vectors M13mp18 and M13mp19 were provided by J. Messing and are described in Messing (1983). *E. coli* hosts were HB101 and DH1 for plasmids, JM101 for M13 phage, and K802 for λ phage (Maniatis et al. 1982).

Enzymes

ϕX174 replication enzymes were kindly provided by A. Kornberg and collaborators (Stanford), avian myeloblastosis virus (AMV) reverse transcriptase was purchased from Seikagaku, Inc., *E. coli* Klenow fragment from Bethesda Research Labs., and nuclease S1 and calf intestinal phosphatase from Boehringer-Mannheim Biochemicals. Restriction enzyme *Eco*RI was a gift from P. Modrich (Washington University, St. Louis). All other restriction enzymes were purchased from New England Biolabs or Boehringer-Mannheim Biochemicals.

DNA and RNA

Standard procedures were followed for the preparation of phage and plasmid DNA and for the construction of plasmid or M13 subclones (Davis et al. 1980; Maniatis et al. 1982; Messing 1983). Purification of RNA from whole organisms and the isolation of the poly(A)$^+$ RNA fraction were as described in Goldschmidt-Clermont et al. (in prep.).

Deletions within the pφX plasmid subclones were generated by digesting the insert DNA with an enzyme that did not cut the vector and with an enzyme with a unique site in the vector and no sites in the insert; the products were end-filled and religated. The site in the vector was chosen so that all deletions were unidirectional from the 3′ end of the subclone as measured with respect to the direction of transcription (see Fig. 2 for the sites within the inserts). The location of the *Bam*HI–*Hin*dIII fragment of genomic clone 3104 used for mapping exon 1 is also shown in Figure 2.

cDNA clones

cDNAs 3600, 3601, 3603, and 3604 were isolated from an embryonic cDNA library constructed by M. Goldschmidt-Clermont from poly(A)$^+$ RNA purified from 1.5- to 5-hr O-R embryos (M. Goldschmidt-Clermont et al., in prep.) and kindly provided by him. The isolation of cDNAs 3600 and 3601 is described by Goldschmidt-Clermont et al. The 3603 and 3604 cDNAs were isolated by the same strategy, using a nick-translated mixed probe from genomic clones 3101, 3104, 3103, 3105, and 3106. cDNAs E10 and E27 were isolated from a 3- to 12-hr embryonic O-R cDNA library (Poole et al. 1985) kindly provided by L. Kauvar (University of California, San Francisco), using a nick-translated probe specific for exon 7. A total of 2.5×10^5 recombinants were screened.

cDNAs representing the *bxd* late RNAs were isolated by screening libraries P (5.5- to 7.5-days, early pupal), Q (7- to 9-day, late pupal), R (adult male), and S (adult female) (a generous gift of L. Kauvar) using a nick-translated 3103 probe (Fig. 2). A total of 5×10^5 recombinants were screened from each library. Clone P2-8 was isolated from library P and clones R2-16 and R2-49 were isolated from library R.

Gel electrophoresis and blotting

Agarose gel electrophoresis of DNA and RNA samples was as described in Goldschmidt-Clermont et al. DNA was transferred to nitrocellulose, and RNA to ATP paper, nitrocellulose, or GeneScreen (New England Nuclear) nylon membrane. Electrophoresis of S1-protected probe and primer extension products was according to standard methods (Maniatis et al. 1975; Sharp et al. 1980).

Radioactive probes and hybridization

Nick-translated probe was prepared as described in Goldschmidt-Clermont et al. Procedures for the production of pφX single-stranded probes are also described in Goldschmidt-Clermont et al., as are the methods of filter hybridization. Probes for S1 nuclease protection and primer extension analysis are described below.

S1 nuclease protection and primer extension analysis

For S1 nuclease mapping, uniformly labeled single-stranded pφX or M13 extension products were synthesized. The production of the latter followed procedures described in detail by Burtis (1985) and is only summarized here. The universal sequencing primer (New England Biolabs) was annealed to single-stranded M13 subclone DNA followed by extension with radioactively labeled nucleotides. The product of the extension reaction was then digested at a unique site within or downstream of the insert fragment by an appropriate restriction enzyme, and the single-stranded probe was purified by electrophoresis through a preparative denaturing acrylamide/urea gel. M13 subclone-derived primers for extension by reverse transcriptase were prepared in the same way.

Hybridization of single-stranded probes/primers to RNA was carried out in the aqueous buffers described in Sharp et al. (1980) and Burtis (1985) at between 50 and 65°C for 3–6 hr; 5 μg of *Drosophila* poly(A)$^+$ RNA or 5 μg of yeast total RNA (control) were annealed to between 30 and 300 pg of probe/primer (between 10^5 and 10^6 dpm of probe/primer). Between 0.5×10^3 and 3×10^3 units of S1 nuclease (30-min units) were used per S1 protection reaction. Following digestion or extension, the products were electrophoresed through denaturing acrylamide (4–6%)/urea gels, dried under vacuum, and exposed to X-ray film. The sizes of the protected fragments or extension products were compared with standard markers or dideoxy sequencing reaction products run on the same gels.

Sequence analysis

This was carried out using standard dideoxy chain termination methods (Sanger et al. 1977; Biggin et al. 1983). cDNAs 3600, 3601, 3603, and 3604 were sequenced completely on both strands. cDNAs E10, P2-8, R2-16, and R2-49 were sequenced on one strand only, and the sequences compared with that of the completely sequenced cDNAs and to genomic sequence from the *bxd* region (D.A. Peattie, H.D. Lipshitz, L. Prestidge, and D.S. Hogness, in prep.). Computer analysis was carried out using BIONET sequence analysis programs and programs written by R. Staden (University of Cambridge, UK) run on a DEC VAX 11/780. FRAMESCAN (written by Staden) was run using an unpublished tabulation of codon usage frequencies for *D. melanogaster* compiled by K. Burtis and kindly provided by him.

Acknowledgments

We thank Michel Goldschmidt-Clermont for providing unpublished data, Michel Goldschmidt-Clermont and Larry Kauvar for kindly providing cDNA libraries, F. Sanger for providing laboratory space during the initial stages of DNA sequencing, Ken Burtis and Winship Herr for assistance with the DNA sequence analysis, Louise Prestidge for technical assistance, and Susanna Lewis, Jeremy Nathans, Carl Thummel, Mark Krasnow, Renato Paro, and Ken Burtis for critical comments. H.D.L. would like to thank Susanna Lewis for many stimulating discussions. H.D.L. and D.A.P. were supported by postdoctoral research fellowships from the Helen Hay Whitney Foundation. D.A.P. also received support from E.I. du Pont de Nemours and Co., Inc. This work was supported by grants from the National Institutes of Health to D.S.H. Computer resources used included the BIONET National Computer Resource for Molecular Biology, the computer facilities at the MRC Laboratory for Molecular Biology in Cambridge, UK, and the VAX 11/780 in the Department of Genetics, Stanford University.

References

Akam, M.E., A. Martinez-Arias, R. Weinzierl, and C.D. Wilde. 1985. Function and expression of *Ultrabithorax* in the *Dro-*

sophila embryo. *Cold Spring Harbor Symp. Quant. Biol.* **50:** 195–200.

Alt, F.W., N. Rosenberg, V. Enea, E. Siden, and D. Baltimore. 1982. Multiple immunoglobulin heavy-chain gene transcripts in Abelson murine leukemia virus-transformed lymphoid cell lines. *Mol. Cell. Biol.* **2:** 386–400.

Banerji, J., L. Olson, and W. Schaffner. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33:** 729–740.

Beachy, P.A., S.L. Helfand, and D.S. Hogness. 1985. Segmental distribution of bithorax complex proteins during *Drosophila* development. *Nature* **313:** 545–551.

Bender, W., M. Akam, F. Karch, P.A. Beachy, M. Peifer, P. Spierer, E.B. Lewis, and D.S. Hogness. 1983. Molecular genetics of the bithorax complex in *Drosophila melanogaster*. *Science* **221:** 23–29.

Bender, W., B. Weiffenbach, F. Karch, and M. Peifer. 1985. Domains of *cis*-interaction in the bithorax complex. *Cold Spring Harbor Symp. Quant. Biol.* **50:** 173–180.

Biggin, M.D., T.J. Gibson, and G.F. Hong. 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequencing. *Proc. Natl. Acad. Sci.* **80:** 3963–3965.

Black, D.L., B. Chabot, and J.A. Steitz. 1985. U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. *Cell* **42:** 737–750.

Breathnach, R. and P. Chambon. 1981. Organization and expression of eucaryotic split gene coding for proteins. *Annu. Rev. Biochem.* **50:** 349–383.

Burtis, K.C. 1985. "Isolation and characterization of an ecdysone inducible gene from *Drosophila melanogaster*." Ph.D. thesis, Department of Biochemistry, Stanford University.

Cabrera, C.V., J. Botas, and A. Garcia-Bellido. 1985. Distribution of *Ultrabithorax* proteins in mutants of *Drosophila* bithorax complex and its transregulatory genes. *Nature* **318:** 569–571.

Casanova, J., E. Sanchez-Herrero, and G. Morata. 1985. Prothoracic transformation and functional structure of the *Ultrabithorax* gene in *Drosophila*. *Cell* **39:** 663–669.

Davis, R.W., D. Botstein, and J.R. Roth. 1980. *Advanced Bacterial genetics*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Gillies, S.D., S.L. Morrison, V.T. Oi, and S. Tonegawa. 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33:** 717–728.

Goldberg, M.L. 1979. "Sequence analysis of *Drosophila* histone genes." Ph.D. Thesis, Department of Biochemistry, Stanford University.

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.

Ingham, P.W. and A. Martinez-Arias. 1986. The correct activation of *Antennapedia* and bithorax complex genes requires the *fushi tarazu* gene. *Nature* **324:** 592–597.

Karch, F., B. Weiffenbach, M. Peifer, W. Bender, I. Duncan, S. Celniker, M. Crosby, and E.B. Lewis. 1985. The abdominal region of the bithorax complex. *Cell* **43:** 81–96.

Kemp, D.J., A.W. Harris, S. Cory, and J.M. Adams. 1980. Expression of the immunoglobulin C_μ gene in mouse T and B lymphoid and myeloid cell lines. *Proc. Natl. Acad. Sci.* **77:** 2876–2880.

Konarska, M.M., R.A. Padgett, and P.A. Sharp. 1985. Trans splicing of mRNA precursors in vitro. *Cell* **42:** 165–171.

Konigsberg, W. and G.N. Godson. 1983. Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*. *Proc. Natl. Acad. Sci.* **80:** 687–691.

Krainer, A.R. and T. Maniatis. 1985. Multiple factors including the small nuclear ribonucleoproteins U1 and U2 are necessary for pre-mRNA splicing *in vitro*. *Cell* **42:** 725–736.

Lennon, G.C. and R.P. Perry. 1985. C_μ-containing transcripts initiate heterogeneously within the IgH enhancer region and contain a novel 5′-untranslatable exon. *Nature* **318:** 475–478.

Lewis, E.B. 1963. Genes and developmental pathways. *Am. Zool.* **3:** 33–56.

———. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276:** 565–570.

———. 1981. Developmental genetics of the bithorax complex in *Drosophila*. In: *Developmental biology using purified genes. ICN-UCLA Symp. Mol. Cell Biol.* **23:** 189–208.

———. 1982. Control of body segment differentiation in *Drosophila* by the bithorax gene complex. In: *Embryonic development. Part A: Genetic aspects* (ed. M. Burgher), pp. 269–288. A.R. Liss, New York.

Lupski, J.R., B.L. Smiley, and G.N. Godson. 1983. Regulation of the *rpsU-dnaG-rpoD* macromolecular synthesis operon and the initiation of DNA replication in *Escherichia coli* K-12. *Mol. Gen. Genet.* **189:** 48–57.

Maniatis, T., E.F. Fritsch, and J. Sambrook. 1982. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Maniatis, T., A. Jeffrey, and H. van deSande. 1975. Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry* **14:** 3787–3794.

Martinez-Arias, A. and P.A. Lawrence. 1985. Parasegments and compartments in the *Drosophila* embryo. *Nature* **313:** 639–642.

McGinnis, W., M.S. Levine, E. Hafen, A. Kuroiwa, and W.J. Gehring. 1984. A conserved DNA sequence in homoeotic genes of the *Drosophila antennapedia* and bithorax complexes. *Nature* **308:** 428–433.

Messing, J. 1983. New M13 vectors for cloning. *Methods Enzymol.* **101:** 28–78.

Mizuno, T., M.-Y. Chou, and M. Inouye. 1984. A unique mechanism regulating gene expression: Translational inhibition by a complementary RNA transcript (*mic*RNA). *Proc. Natl. Acad. Sci.* **81:** 1966–1970.

Moore, C.L. and P.A. Sharp. 1985. Accurate cleavage and polyadenylation of exogenous RNA substrate. *Cell* **41:** 845–855.

Morata, G. and S. Kerridge. 1981. Sequential functions of the bithorax complex of *Drosophila*. *Nature* **290:** 778–781.

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.

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.

Regulski, M., K. Harding, R. Kostriken, F. Karch, M. Levine, and W. McGinnis. 1985. Homeo box genes of the *Antennapedia* and bithorax complexes of *Drosophila*. *Cell* **43:** 71–80.

Sanchez-Herrero, E., I. Vernos, R. Marco, and G. Morata. 1985. Genetic organization of *Drosophila* bithorax complex. *Nature* **313:** 108–113.

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

Scott, M.P. and A.J. Weiner. 1984. Structural relationships among genes that control development: Sequence homology between the *Antennapedia*, *Ultrabithorax*, and *fushi tarazu* loci of *Drosophila*. *Proc. Natl. Acad. Sci.* **81:** 4115–4119.

Sharp, P.A., A.J. Berk, and S.M. Berget. 1980. Transcription maps of adenovirus. *Methods Enzymol.* **65:** 750–768.

Simons, R.W. and N. Kleckner. 1983. Translational control of IS10 transposition. *Cell* **34:** 683–691.

Smiley, B.L., J.R. Lupski, P.S. Svec, R. McMacken, and G.N. Godson. 1982. Sequences of the *Escherichia coli dnaG* primase gene and regulation of its expression. *Proc. Natl. Acad. Sci.* **79:** 4552–4554.

Solnick, D. 1985. Trans splicing of mRNA precursors. *Cell* **42:** 157–164.

Tiong, S.L., L.M. Bone, and T.R.S. Whittle. 1985. Recessive lethal mutations within the bithorax complex in *Drosophila melanogaster*. *Mol. Gen. Genet.* **200:** 335–346.

Wang, X-F. and K. Calame. 1985. The endogenous immunoglobulin heavy chain enhancer can activate tandem V_H promoters separated by a large distance. *Cell* **43:** 659–665.

White, R.A.H. and M. Wilcox. 1985. Regulation of the distribution of *Ultrabithorax* proteins in *Drosophila*. *Nature* **318:** 563–567.

Womble, D.D., X. Dong, R.P. Wu, V.A. Lucknow, A.F. Martinez, and R.H. Rownd. 1984. IncFII plasmid incompatibility product and its target are both RNA transcripts. *J. Bacteriol.* **160:** 28–35.

Yancopoulos, G.D. and F.W. Alt. 1985. Developmentally controlled and tissue-specific expression of unrearranged V_H gene segments. *Cell* **40:** 271–281.

Zhuang, Y. and A.M. Weiner. 1986. A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell* **46:** 827–835.