MOLECULAR AND CELLULAR BIOLOGY, Dec. 1985, p. 3600–3609
0270-7306/85/123600-10$02.00/0

Vol. 5, No. 12

Electron Microscopic Heteroduplex Mapping Identifies Regions of the engrailed Locus That Are Conserved between *Drosophila melanogaster* and *Drosophila virilis*

JUDITH A. KASSIS, MEI LIE WONG, AND PATRICK H. O'FARRELL*

Department of Biochemistry & Biophysics, University of California, San Francisco, California 94143

Received 17 June 1985/Accepted 16 September 1985

Physical localization of mutations in the engrailed (*en*) gene suggested that at least 70 kilobases (kb) of genomic sequences contribute to the normal function of this gene. Molecular characterization has suggested that *en* function is encoded in a small, 4.5-kb primary transcript. To identify functional regions within the 70 kb of the *en* locus of *D. melanogaster*, we identified sequences conserved in the *D. virilis* genome (estimated divergence time, 60 million years). Based on homology to *D. melanogaster*, we isolated *en* DNA from a *D. virilis* genomic library. Electron microscopic heteroduplex analysis indicated that in 70 kb there is 20 kb of conserved DNA in 33 different regions dispersed throughout the *en* locus, including two which encode parts of the major embryonic transcript. The conserved regions are in the same linear order and are spaced by similar lengths of nonconserved sequences in the *D. virilis* and *D. melanogaster* DNAs. What functional constraints have enforced conservation of sequences throughout the entire 70 kb and protected the region from divergence of size and arrangement? Our working hypothesis is that sequences necessary for the complex spatial and temporal pattern of *en* expression are dispersed throughout the 70-kb *en* locus and that selection for proper regulation restricts evolutionary divergence.

Genetic and molecular analyses of developmental genes in *Drosophila melanogaster* have given us new ways to study the mechanisms involved in pattern formation in early embryos. For example, mutations in the fushi tarazu or engrailed (*en*) genes lead to severe perturbations of the segmental pattern that is established during the first 10 h of *Drosophila* development (13, 23, 30). Both of these genes are expressed in spatially localized patterns, generating a series of stripes along the anterior-posterior axis of the embryo (7, 8, 14). The patterns of expression precede morphological segmentation and anticipate those elements of the segmental pattern that are defective in the mutants. Extensive genetic analysis suggests that the product of the *en* locus acts in the posterior compartment of each segment to define the posterior cells as a distinct population (13, 16, 22). In the absence of *en* function, cells of posterior compartments acquire characteristics of anterior cells and mix with anterior cells of the same or adjacent segment (22), leading to erosion of segmental divisions in *en* lethal embryos (13, 23).

Molecular cloning and characterization of *en* mutant alleles has indicated that a large region of genomic DNA, 70 kilobases (kb), contributes to *en* function (15). Despite the large size of the genetic unit, the major embryonic transcript is derived from only 4.5 kb of genomic sequences (26), and no transcript spanning the 70 kb has yet been identified (B. Drees, P. H. O'Farrell, and T. Kornberg, manuscript in preparation). A cDNA corresponding to the major embryonic transcript has been sequenced and contains a highly conserved sequence, the homeobox, that is common to a number of developmental genes in *D. melanogaster* (7, 19, 20, 26, 28). The embryonic pattern of *en* gene expression is complex but orderly (14; S. DiNardo, J. Kuner, J. Theis, and P. H. O'Farrell, Cell, in press). Briefly, *en* product first accumulates in a single stripe just posterior to the cephalic furrow at the onset of gastrulation (DiNardo et al., in press). During the following 15 min, *en* product accumulates in the progenitors of the posterior compartments in each segment of the embryo, resulting in a zebra-striped pattern. In addition to its embryonic pattern of expression, the *en* gene plays a crucial role during metamorphosis. Although the details of the pattern of larval expression have yet to be established, product accumulation is evident in the posterior compartments in the imaginal disks (14; S. DiNardo, unpublished data). It is our working hypothesis that the major embryonic transcript encodes *en* function and elements in the rest of the 70 kb are involved in regulation of the complex temporal and spatial expression of this product.

The proposition that distant flanking sequences play a regulatory role is supported by genetic evidence suggesting that the *en* locus is not a simple complementation unit (13, 15). In addition, molecular characterization shows that the locus can be divided into two discrete regions, a 40-kb region that, when broken by chromosomal rearrangements, yields an embryonic lethal *en* phenotype and a 20-kb region that, when interrupted by rearrangements, produces nonlethal *en* mutants with adult morphological defects (15).

How can we identify the interesting regions within the 70 kb which act as the *en* locus? Numerous studies have shown that functionally important sequences are conserved throughout evolution (6, 25, 32). Although several studies have examined the conservation of the homeobox sequence (4, 15, 17), we addressed two different issues. Like *en*, a number of developmental loci are extremely large (1, 3, 15, 29). We would like to know whether these loci are large because of the presence of nonfunctional spacer DNA or whether functional elements are dispersed throughout. In addition, we were interested in identifying those sequences, beyond the homeobox, which are likely to be important for proper *en* expression and function. To these ends, we identified the regions of the *en* locus conserved between two *Drosophila* species which are separated by a sufficient evolutionary period (~60 million years; 2) such that regions

* Corresponding author.

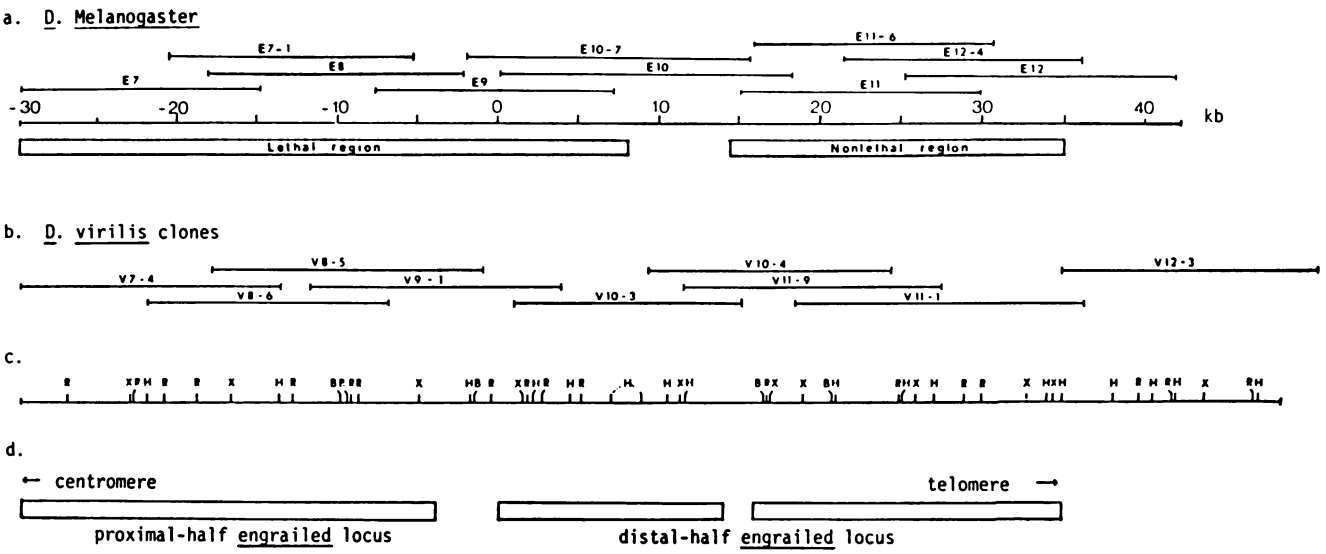

FIG. 1. *D. melanogaster en* clones. Also shown are the regions defined by breakpoint mutations to be involved in lethal and nonlethal *en* function (from reference 15). (b) *D. virilis en* clones. The total length of this DNA is approximately 10 kb (~15%) longer than the homologous *D. melanogaster* DNA. Because of this, proximal clones approximately line up with their analogs, whereas distal ones do not. (c) Restriction map of *D. virilis* clones. The restriction map was generated by single and double digests of phage DNA. There is no obvious correspondence of restriction sites between *D. melanogaster* and *D. virilis* (data not shown). H, *Hin*dIII; B, *Bam*HI; R, *Eco*RI; X, *Xho*I. (d) Location of *D. melanogaster* regions examined by heteroduplex analysis. Blank spaces indicate a gap in the electron microscopic heteroduplex data.

lacking functional constraints will show essentially no sequence relatedness.

MATERIALS AND METHODS

Recombinant DNA libraries. The λ vector 1059 was used to clone *Sau*3A partial digests of *D. virilis* (strain 1801.1) DNA as described by Karn et al. (11). This library was made by D. Swerdlow. An amplified library of Charon 30 clones carrying inserts from wild-type (Oregon R) *D. melanogaster* was obtained from Ron Blackman (Ph.D. thesis, Harvard University, Boston, Mass., 1983) for isolation of clones labeled E7-1, E10-7, E11-6, and E12-2. Other *D. melanogaster* clones used were from Kuner et al. (15).

Screening the *D. virilis* library. Standard methods were used for all procedures (18). Subclones of three adjacent *Eco*RI fragments from *D. melanogaster* DNA were nick translated and hybridized overnight at reduced stringency (20% formamide, 2× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate] at 42°C) to three replica lifts from the *D. virilis* library. Washing was in 1× SSC–0.1% sodium dodecyl sulfate at 50°C for 1 h. These conditions are theoretically 39°C below the T_m of *D. melanogaster* DNA. Clones which hybridized to more than one fragment were picked for further study. Restriction analysis and Southern blot analysis at reduced stringency with probes from *D. melanogaster en* DNA indicated the arrangement of the homologous fragments.

For the limited chromosome walk, two restriction fragments from near the appropriate ends were labeled by nick translation and hybridized to replica filters at regular stringency (50% formamide, 5× SSC; wash with 0.1× SSC, 0.1% sodium dodecyl sulfate). Clones picked were checked for overlap with adjacent clones by both restriction analysis and cross-hybridization.

Electron microscopy. Electron microscopic heteroduplexing was as described by Koller and Delius (12). Phage DNA was denatured in 0.1 N NaOH, neutralized, and renatured for 30 min at 25°C in 50% formamide–0.2 M NaCl. Hyperphase solution was 30% formamide–0.02 M NaCl. Hypophase was 0.005% octylglucoside in water. DNA was picked up on Parlodion-coated grids and platinum shadowed. DNA was measured with an Apple IIe graphics program. Standards were measured on the same grid and often on the same field (single stranded, a bacteriophage M13 derivative [7.5 kb]; double stranded, a plasmid, pJMC110 [4.755 kb], and phage arms).

RESULTS

Isolation of the *en* region from *D. virilis*. To isolate clones representing the entire region of the *D. virilis* genome that corresponded to the 70-kb *en* region of *D. melanogaster*, we relied partly on homology to *D. melanogaster* clones and partly on a limited chromosome walk in the *D. virilis* library. Figure 1a shows a map of the *D. melanogaster en* region including the positions of the lethal and nonlethal regions as defined by analysis of breakpoint mutations (15). Early analysis with different fragments from the *D. melanogaster en* walk as hybridization probes to Southern transfers of DNA from various fly species indicated that a 4.7-kb *Eco*RI fragment from the middle of E9 was highly conserved in evolution (data not shown); later it was found that this fragment encodes an exon of an *en* transcript (26). Using this fragment as a probe at reduced stringency, we isolated two overlapping clones from a *D. virilis* genomic DNA library in lambda phage 1059 (D. Swerdlow, unpublished data; Fig. 1b, V8-5 and V8-6). To ensure that these clones represented the homologous region of *D. virilis* rather than just coincidental homology, we used additional Southern analyses to show that other fragments of the V8-5 and V8-6 phage hybridized to adjacent fragments from the *D. melanogaster en* region and that homologous regions were arranged colinearly.

To isolate another section of the *D. virilis en* region, we screened the *D. virilis* library at reduced stringency with fragments from *D. melanogaster* phage E11. Two overlapping clones were obtained, V11-1 and V11-9. Each of the

phage hybridized to a number of *Eco*RI fragments from *D. melanogaster en* phage E10 and E11, and Southern analysis indicated they were similarly arranged.

To isolate *D. virilis* clones that might connect the two *en* regions we had isolated, we used a chromosome walking procedure. A step was taken from V8-5 toward the proximal end and from V11-9 toward the distal end. The newly isolated V9-1, V10-4, and V10-3 were shown to overlap with V8-5 and V11-9 by both restriction analysis and cross-hybridization of clones.

To extend the cloned region, we again relied on sequence homology and screened the *D. virilis* library at reduced stringency with probes from *D. melanogaster* phage E7 and E12 to isolate V7-4 and V12-3, respectively. The clones obtained were shown to overlap with the *D. virilis* walk both by restriction mapping and by hybridization to the adjacent step in the walk. These phage completed the *D. virilis* walk, which included a region of 80 kb homologous to the 70-kb *D. melanogaster en* region. A restriction map of the *D. virilis en*-related DNA clones is shown in Fig. 1c.

Heteroduplex mapping of the conserved regions of the *en* locus. It became apparent during the isolation of the homologous clones from *D. virilis* that the *en* locus was highly conserved in both length and arrangement. Without exception, at reduced stringency every *D. melanogaster en* fragment tested hybridized to a fragment from the *D. virilis en* locus. Tests with a few fragments at higher stringency indicated that fragments varied in the length or accuracy of homology. To localize the conserved regions responsible for hybridization, we hybridized analogous phage from *D. virilis* and *D. melanogaster* and analyzed the heteroduplexes by electron microscopy.

Before presenting the results, we will comment on several technical features of the heteroduplex analysis. Heteroduplexes were made under standard conditions (50% formamide, 0.2 M salt at 25°C) and spread under low-stringency conditions (30% formamide, 0.02 M salt). Conditions used for spreading were close to isodenaturing, that is, the denaturing conditions of the spreading solution were similar to those of the hypophase in an attempt to ensure that the molecules were at equilibrium (5). However, some degree of heterogeneity in heteroduplexes did occur, that is, a region would appear double stranded in one molecule and single stranded in the next on the same grid. This result has been seen by many other investigators and may indicate that the conditions used were very close to the T_m of the duplex (5). In addition, particular regions that were sometimes seen as long duplexes at other times were interrupted by unpaired regions. The stability of the duplex depends on many factors: the degree of mismatch, the arrangement of the mismatches within the area, the AT richness of an area, and the length of the conserved area. For these reasons, the degree of sequence conservation cannot be unambiguously defined without sequence data, but here we assume the generalization that duplexed regions are more conserved than single-stranded regions. For simplicity, we sometimes refer to the more consistently duplexed regions as more stable.

To present the data, we show example heteroduplexes and have tabulated the number of molecules measured for each region, the percent of the cases in which a particular duplex region appeared paired or partially paired, and the total number of molecules examined for a particular region (Table 1). We have numbered the regions consecutively and denote unpaired regions with ss (single stranded) and paired regions with a d (duplexed). We have summarized all the data in an aggregate schematic heteroduplex (Fig. 2), which also in-

TABLE 1. Summary of heteroduplex analysis

Genomic area	Length (SE) in kb[a]	No. measured	% Duplexed (% partially duplexed)	Total no. examined
1ss	1.52 (0.25)	13		
	1.56 (0.4)	13		
2d	0.81 (0.15)	12	73 (7)	15
3d[b]	1.24 (0.35)	10	7 (60)	15
4ss	1.57 (0.38)	10		
	1.36 (0.39)	10		
5d	1.01 (0.22)	9	42 (33)	12
6ss	0.83 (0.37)	8		
	0.76 (0.26)	8		
7d	0.5 (0.23)	9	75 (0)	12
8ss	1.19 (0.26)	6		
	1.05 (0.22)	6		
9d	0.91 (0.27)	12	53 (11)	19
10ss	2.38 (0.22)	9		
	2.61 (0.17)	9		
11d	0.71 (0.16)	10	63 (15)	13
12ss	2.45 (0.31)	9		
	2.44 (0.40)	9		
13d	0.69 (0.20)	19	62 (17)	24
14ss	0.55 (0.38)	15		
	0.53 (0.27)	15		
15d	0.41 (0.18)	15	79 (0)	19
16ss	0.61 (0.14)	15		
	0.29 (0.17)	15		
17d	1.47 (0.26)	12	78 (7)	14
18ss[c]	0.66 (0.23)	11		
	0.67 (0.22)	11	0 (36)	11
19d	0.38 (0.13)	11	78 (0)	14
20ss	1.35 (0.2)	7		
	1.05 (0.41)	7		
21d	0.52 (0.19)	7	58 (0)	12
22ss	0.76 (0.1)	5		
	0.85 (0.24)	5		
23d	0.19 (0.11)	6	50 (0)	12
24ss	1.11 (0.33)	7		
	0.96 (0.35)	7		
25d	0.3 (0.15)	7	58 (0)	12
26ss	1.69 (0.1)	4		
	1.12 (0.12)	4		
27d	0.28 (0.1)	4	36 (0)	13

Continued on following page

TABLE 1—*Continued*

Genomic area	Length (SE) in kb[a]	No. measured	% Duplexed (% partially duplexed)	Total no. examined
28ss[d]	1.80 (0.49)	4		
	2.97 (0.12)	4		
29ss[d]	8.11 (0.88)	5		
	0.20 (0.08)	5		
30d	0.71 (0.17)	5	38 (0)	13
31ss	0.75 (0.18)	6		
	0.27 (0.24)	6		
32d	0.27 (0.12)	12	92 (0)	13
33d	0.79 (0.18)	12		
	0.88 (0.24)	12		
34d	0.37 (0.16)	12	85 (0)	14
35ss	0.72 (0.21)	19		
	0.67 (0.24)	19		
36d	0.76 (0.18)	21	41 (45)	24
37ss	0.76 (0.23)	20		
	1.13 (0.25)	20		
38d	0.91 (0.14)	14	58 (0)	24
39ss	0.54 (0.22)	10		
	0.4 (0.12)	10		
40d	0.33 (0.14)	11	41 (4)	24
41ss	0.49 (0.16)	7		
	0.26 (0.24)	7		
42d	0.42 (0.14)	10	50 (0)	20
43ss	0.32	2		
	0.32	2		
44d	0.32	2	22 (0)	9
45ss	0.56 (0.1)	3		
	0.78 (0.06)	3		
46d	0.44 (0.07)	4	44 (0)	9
47ss	0.29 (0.09)	5		
	0.32 (0.09)	5		
48d	0.69 (0.1)	6	45 (9)	11
49ss	0.80 (0.17)	11		
	0.76 (0.18)	11		
50d	1.06 (0.16)	10	77 (0)	13
51ss[d]	7.6 (0.62)	10		
	0.73 (0.21)	10		
52ss[d]	8.06 (1.74)	7		
	0.88 (0.30)	7		
53d	0.95 (0.21)	9	44 (55)	9
54ss	1.37 (0.15)	6		
	1.17 (0.14)	6		
55d	1.10 (0.27)	6	25 (50)	8
56ss	2.18 (0.52)	8		
	1.67 (0.36)	8		
57d	0.81 (0.17)	8	75 (25)	8
58ss	1.8	2		
	1.3	2		
59d	0.35 (0.11)	12	66 (0)	18
60ss	3.06 (0.4)	10		
	0.45 (0.24)	10		
61d	0.13 (0.07)	10	90 (0)	11
62ss	4.0 (0.72)	8		
	5.52 (1.36)	8		
63d	0.1 (0.05)	10	100 (0)	10
64ss	1.44 (0.29)	12		
	1.96 (0.44)	12		
65d	0.8 (0.13)	15	47 (53)	15
66ss	1.3 (0.25)	8		
	1.65 (0.17)	8		
67d	0.27 (0.05)	8	72 (0)	11
68ss	0.49 (0.09)	6		
	0.62 (0.12)	6		
69d	0.15 (0.12)	9	81 (0)	11
70ss[d]	1.09 (0.14)	9		
	6.62 (0.66)	8		

[a] Length was calculated by comparison with single-stranded and double-stranded control DNAs spread on the same grids.

[b] Region 3d, though not separated from region 2d by a single-stranded section, is distinct from it because it has an unusually high frequency of formation of a partial duplex (that is, a duplex interrupted by very small bubbles).

[c] A short region of homology (~0.1 kb) sometimes split this region into two bubbles (Fig. 2). Because it was so short, the region was scored as a unit.

[d] These single-stranded regions were at the end of a phage pair, where no overlapping phage were examined.

cludes the position of the breakpoint mutations and the major embryonic transcript. We will discuss the data in two sections, the centromere-proximal half and the centromere-distal half (Fig. 1d). It should be noted that very AT-rich or very short regions of homology may not have been detected in these studies.

Conserved DNA segments in the centromere-proximal half of the *en* locus. We define the proximal side of the *en* region by the positions of two mutations, en^{SF52} and en^{SF37}, that have breakpoints that are 15 kb 3′ to the characterized structural gene. Conservation of the region extending through the position of the en^{SF52} mutation and the major embryonic transcript was analyzed by forming heteroduplexes between the *D. melanogaster* clones E7 and E7-1 and the corresponding *D. virilis* clones V7-4, V8-6, and

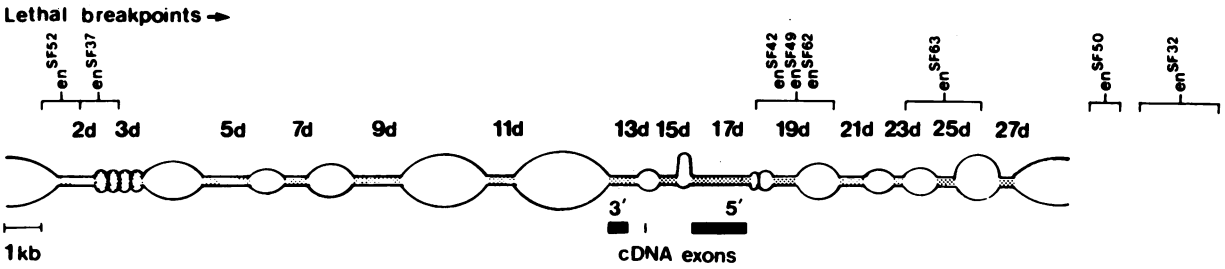

FIG. 2. Schematic representation of electron microscopic heteroduplex analysis showing the location of the conserved regions of the *engrailed* locus. Double-stranded, conserved regions are shown as shaded areas and are numbered. Single-stranded, nonconserved regions are shown as two curved lines; strands are shown to be the same length except where known to differ in size, in which case the longer strand is shown on top. Length of the shortest nonconserved strand was used to space the conserved areas. Also shown are the positions of the breakpoint mutations (from reference 15) and the exons of the major embryonic transcript (from reference 26).

V8-5. Figure 3 shows the approximate overlap of the phages hybridized and four example heteroduplexes. Table 1 (1ss through 28ss) is a compilation of data from at least 10 molecules of each heteroduplex shown in Fig. 3. Beside the regions corresponding to the major embryonic transcript (areas 17d and 13d), there are 12 other conserved areas in this region ranging in size from 0.1 to 1 kb. Some of these are as stable as the regions corresponding to the major embryonic transcript (Table 1, areas 2d, 7d, 11d, 15d, and 19d), whereas others are relatively unstable (especially areas 23d and 27d). It is interesting that a region within the first intron of the major embryonic transcript is conserved (area 15d). This conservation has been confirmed by sequence analysis and contains some AT-rich regions (manuscript in preparation). This may explain why this area was occasionally unpaired in the heteroduplexes (4 of 19 unpaired). In this region of the *en* walk, almost all of the nonconserved regions contained single strands of similar size; only areas 16ss and 26ss contained single strands that were significantly different, and these differences were minor (0.3 and 0.6 kb). This may indicate that, in this region of the chromosome, there is selection against large insertions or deletions even where the sequence itself is not conserved.

An example of the heterogeneity seen in our experiments is shown in Fig. 3C and D. Two different heteroduplexes of V8-5 versus E7-1 are shown. In Fig. 3D, the region near the short arm of the phage is totally unpaired, whereas in Fig. 3C there are four short areas of pairing (corresponding to areas 21d, 23d, 25d, and 27d) in this region. In contrast, the unpaired region (marked by the arrowhead) in the heteroduplex in Fig. 3C is partially paired in Fig. 3D. This is because conserved area 15d is unpaired in the molecule depicted in Fig. 3C.

The next ~4 kb of the *en* walk was not included in the heteroduplex analysis, because all analogous phage clones contained inserts in the opposite orientation. Alignment of the *D. virilis* and *D. melanogaster* clones by Southern analysis suggested that homology in this region is comparable to that in other regions, with one exception. A 3.6-kb *Eco*RI fragment from *D. melanogaster* hybridized to *D. virilis* DNA, which spanned about 5 kb. One 0.8-kb and one 0.2-kb fragment from *D. virilis* (which were not adjacent) did not hybridize at all to *D. melanogaster* DNA. By criteria described below, this extra DNA is not repetitive.

Conserved DNA segments in the distal half of the *en* locus, including the nonlethal region. Figure 4 shows heteroduplexes from phage covering the distal half of the *en* locus. Though it is entirely upstream of the characterized transcription unit, half of the localized *en* mutations map in this region (Fig. 2). For the first part of this region, *D. melanogaster* clones E10 and E10-7 were hybridized to *D. virilis* clones V10-3 and V10-4 (Fig. 4A and B; Table 1, areas 29ss through 51ss). This region of the *en* locus is densely packed with conserved DNA; 53% of the DNA appears to be conserved. Again, there is very little difference in the lengths of the single-stranded DNAs in the nonconserved areas. Following this, we have about a 1.5-kb gap in our data and then the most distal part of the *en* locus, areas 52ss through 70ss, where all the nonlethal *en* mutations map (Fig. 2). Although the pattern of conservation covered by the first pair of phage hybridized, E11-6 versus V11-1, was similar to what was seen in the more proximal region of the *en* locus, starting with 59d the pattern of conservation is strikingly different (Table 1, 59d to 70ss; Fig. 2 and 4D). There is only about 1.8 kb of conserved DNA extending over 11 kb (16%). In addition to the divergence of sequences in this region, we found that corresponding nonconserved regions differ greatly in length. For example, in area 60ss the one species has 3 kb more than the other. This extra DNA is in the *D. virilis* genome and appears to be due to a repeated element (see below). In addition, single-stranded DNAs in the large, unpaired region (area 62ss) differed by about 1.5 kb (standard error was large in this region, but Southern analysis confirmed this size difference). Hybridizing *D. melanogaster* fragments to Southern blots of the *D. virilis* clones indicated that the extra 1.5 kb was in the *D. virilis* genome. By criteria described below, there was apparently no substantial amount of repetitive DNA in this area.

Repeated DNA sequences within *D. virilis* and *D. melanogaster en* loci. Size differences in nonconserved regions of the DNA could be due to insertion of mobile elements in one genome or the other. Previous work showed that there were no insertion elements in the *D. melanogaster en* region of the chromosome (15). We were interested to see whether this was also the case in the *D. virilis en* locus.

An easy way to identify repetitive sequences within a stretch of cloned DNA is to probe a Southern blot of the clones with nick-translated genomic DNA. A repeated sequence will represent a greater portion of the probe than will a single-copy sequence; hence, repeated sequences will appear as darker bands on autoradiography. When this experiment was done on the clones shown in Fig. 1a from the *D. melanogaster en* locus, two bands were predominant. One of these bands is known to contain the repetitive sequence opa (31), which is present in the coding region of the major embryonic transcript (26). The other is within phage E10. When the same experiment was done on *D. virilis en* clones, a number of bands, including one known to contain an opa element (J. A. Kassis, D. Wright, S. Poole, and P. H. O'Farrell, manuscript in preparation) gave a weak

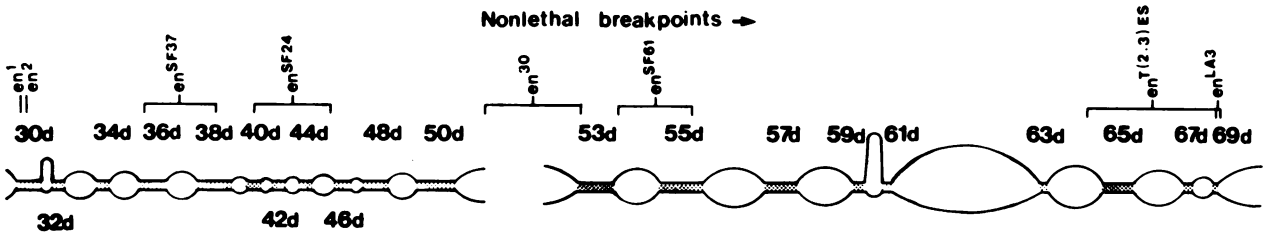

FIG. 2—*Continued*

signal. In addition, two bands gave stronger signals. One of those bands corresponded to DNA present in area 60ss; the other was a 6.5-kb band which contained areas 53d through 57d, where no large differences in the sizes of the nonconserved areas were observed. Hybridization of a nick-translated *D. virilis* fragment which contained the DNA from area 60ss to *Hin*dIII- and *Eco*RI-digested *D. virilis* and *D. melanogaster* DNAs showed that this repeated sequence is not present in *D. melanogaster* DNA and is moderately repeated in the *D. virilis* genome (data not shown). This suggests that it may be an insertion element (27).

DISCUSSION

In this study, we were interested in distinguishing domains of sequence whose evolutionary divergence is constrained from those in which it is not constrained. This requires that the two sequences compared have diverged sufficiently such that nonfunctional sequences have lost homology. *D. virilis* diverged from *D. melanogaster* 60 million years ago (2). This should be a sufficient evolutionary distance to distinguish between functional and nonfunctional sequences, since nonfunctional bases (such as intron sequences and the degenerate wobble positions in amino acid codons) evolve at a rate of about 1% per million years (9, 25; Blackman, Ph.D. thesis). In hybridization studies, 81.4% of the *D. virilis* genome was too divergent to hybridize at all with *D. melanogaster* under nonstringent conditions (33). A sequencing study of the gene encoding the 82-kd heat shock protein demonstrated that, although protein-coding sequences are very highly conserved at the amino acid level, there are many silent changes in the coding sequence and that the intron sequences are totally diverged between the two species (Blackman, Ph.D. thesis). Therefore, we felt that a thorough comparison of the *en* region between these two species would help point to the functional regions of this large locus.

The use of homology to identify a gene in a distantly related species can be complicated by cross-hybridization to related genes or to pseudogenes. If we are to interpret our results in terms of functional constraints on sequence divergence, it is essential that the sequences compared are functional homologs. We are convinced that the 80 kb of *D. virilis* DNA that we have cloned represents the functional homolog of the *D. melanogaster* engrailed (*en*) locus because of the pattern and extent of sequence homology. Homology is spread in a colinear fashion over the entire 80 kb of DNA examined, and no other areas of equivalent homology were observed.

Our analysis shows that there are numerous blocks of conserved sequences spread throughout the *en* locus. In addition to detecting the areas of homology, heteroduplex analysis reveals their arrangement. The data suggest that there may be functional constraints on the order and spacing of the blocks of conserved sequences. The *D. virilis* and *D. melanogaster* sequences are colinear. With few exceptions, regions that were not conserved in terms of sequence were conserved in terms of length. To assess the significance of the conservation of these parameters we need to know the rates of change in size and sequence arrangement during evolution. If these parameters change much more slowly than nucleotide sequence, then the conservation of these parameters between *D. virilis* and *D. melanogaster* may not be particularly significant. Unfortunately, there are few data with which we can compare our study. Nonetheless, there are at least indications that sizes can diverge rapidly. First, in the globin gene complexes of vertebrates, the total size of the complex and the number and spacing of the genes in the complexes have diverged substantially over evolutionary times, comparable to the difference separating *D. virilis* and *D. melanogaster* (10). Second, the description of numerous insertional polymorphisms in *D. melanogaster* indicates that large size changes can occur rapidly (27). Finally, Meyerowitz and Martin (21) have recently demonstrated that DNA in the 68C glue gene cluster evolves at a remarkably rapid rate. In addition to sequence divergence, there are large changes in the size and organization of this region of the chromosome after very short evolutionary periods (5 to 15 million years). These studies lead us to believe that there might be functional constraints on the spacing of the conserved *en* sequences or the size of the locus.

Based on the phenotypes and complementation behavior of *en* mutations, we have proposed that the extended sequences surrounding the identified *en* coding unit are involved in the complex spatial and temporal pattern of *en* expression (15, 24). Briefly, many of the *en* point mutations die with fused segments late in embryogenesis, whereas lethal breakpoint mutants die at a similar stage with nearly normal segments (13; T. Kornberg, personal communication). This suggests that the lethal breakpoint mutants produce an activity which is not present in some of the point mutants. Nonlethal breakpoint mutations cause a variety of adult cuticular defects dependent on the position of the break. Thus, we suggest that, although the major embryonic transcript may contain all the information necessary to produce normal segments, there must be other functional regions of the *en* locus which either regulate the expression of the identified *en* protein or modify its behavior in some way.

Our present analysis suggests that a very large amount of DNA in the *en* locus is functional. Within this large locus, 20 kb of DNA was sufficiently conserved to form duplexes, and these conserved sequences were divided into 33 distinguishable areas. The function of only two of these areas is known; these correspond to known coding regions. Although the present analysis cannot tell us the function of the remaining

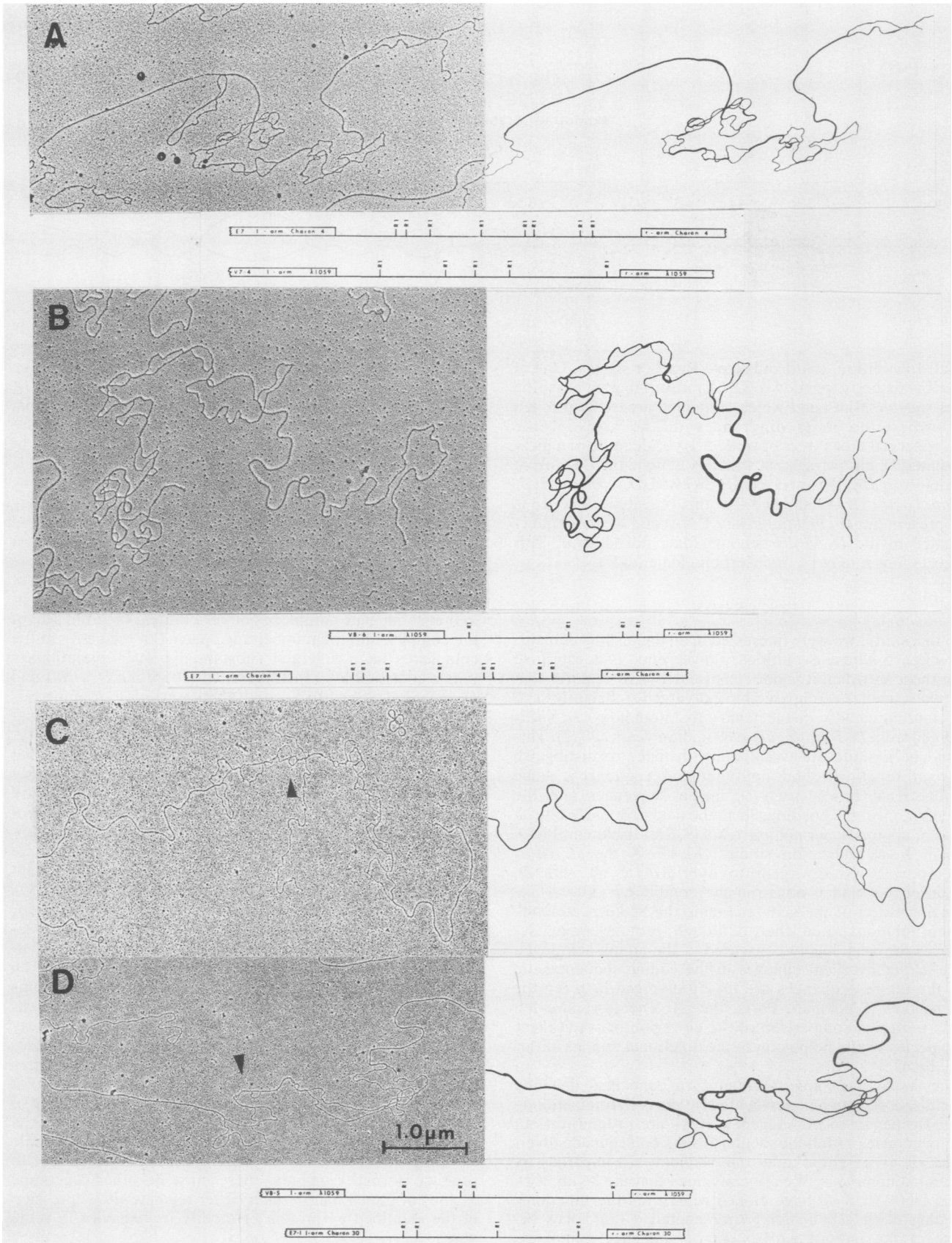

FIG. 3. Heteroduplexes from the proximal half of the *en* locus. Phage are lined up showing approximate overlap, and example heteroduplexes are shown. Data from many such molecules are tabulated in Table 1 (areas 1ss to 28ss). (A) V7-4 hybridized to *D. melanogaster* clone E7. Right arm of phage is on the right and in this case does not hybridize at the end (Charon 4 and λ 1059 are not homologous in this region). Circular DNA is single-stranded marker. Areas 1ss through 13d are covered by this phage pair. (B) V8-6 versus E7. Right arm of phage is on the left. Long arm is broken in one phage. Areas 9d through 16ss are covered by this phage pair. (C and D) *D. virilis* phage V8-5 hybridized to *D. melanogaster* phage E7-1. Right arms of phage in this case are totally homologous and duplexed and are on the right. Arrowhead points to a duplex area (15d) in D that is seen as a bubble in C. Note also that the area near the right arm is completely open in D, whereas in C there are regions that are duplexed. Areas 13d through 28ss are covered by these phage.

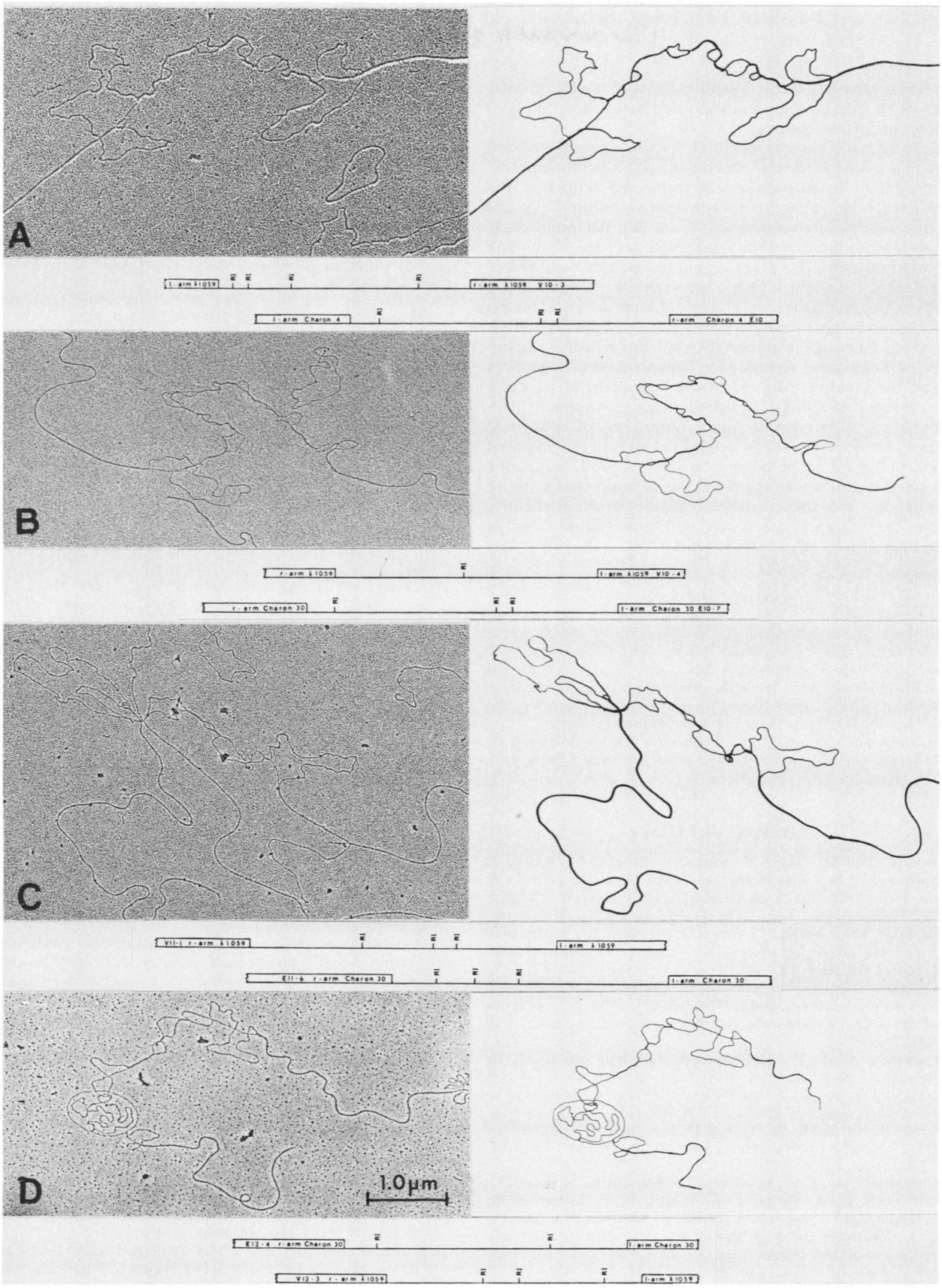

FIG. 4. Heteroduplexes in the distal half of the *en* locus. As in Fig. 3, phage overlap and example heteroduplexes are shown. Data from many molecules are tabulated in Table 1 (29ss to 70ss). (A) V10-3 versus E10. Right arm is on the left. Circular DNA is double-stranded DNA marker. Areas 29ss through 42d are covered by this phage pair. (B) V10-4 versus E10-7. Right arm is on the right. Areas 35ss through 51ss are covered by this phage pair. (C) V11-1 versus E11-6. Right arm is on the right. Areas 52ss through 59d are covered by this phage pair. (D) V12-3 versus E12-2. Right arm is on the left. Areas 59d through 70ss are covered by this phage pair.

conserved areas, it does dramatically confirm the genetic data, which suggest that the locus is complex, and points to regions for further studies.

Several developmental loci of *D. melanogaster* have been examined molecularly and share some interesting features. Many contain an extraordinarily conserved sequence, the homeobox, which suggests that they are evolutionarily and perhaps functionally related (7, 19, 26, 28). Although many studies have emphasized the conservation of the homeobox in evolution, there are other common features of developmental genes. In particular, antennapedia, ultrabithorax, scute, and engrailed are all physically large loci and are all expressed in complex manners (1, 3, 24, 29). Perhaps there is something functionally important and similar in the organization of these units. To study the conservation of sequences beyond the highly conserved homeobox element, we compared functionally homologous sequences. Although it remains to be seen whether the arrangement of conserved elements is similar in other developmental loci or in an *en* homolog after a much longer evolutionary period, our results indicate that extended regions of the genome contribute to *en* function. We suggest that large, dispersed regulatory regions may characterize genes with very complex spatial and temporal patterns of expression.

ACKNOWLEDGMENTS

We thank David Swerdlow for construction of the *D. virilis* genomic library and Joy Silen and David Swerdlow for isolation of some of the phage clones used in this study. Steve DiNardo, Jim Theis, Harald Biessman, Claude Desplan, Elisabeth Sher, and Steve Poole all provided helpful comments on the manuscript. We also thank Judy Piccini for her help in preparing the manuscript.

This work was supported by a National Science Foundation grant (P.H.O.) and a postdoctoral training grant (J.A.K.) (from institutional training grant 5 T32 CA09270).

LITERATURE CITED

1. **Bender, W., M. Akam, F. Karch, P. Beachy, M. Peifer, P. Spierer, E. Lewis, and D. Hogness.** 1983. Molecular genetics of the *bithorax* complex in *Drosophila melanogaster*. Science **221**:23–29.
2. **Beverley, S. M., and A. C. Wilson.** 1984. Molecular evolution in *Drosophila* and higher Diptera. II. A time scale for fly evolution. J. Mol. Evol. **21**:1–13.
3. **Campuzano, S., L. Carramolino, C. Cabrera, M. Ruiz-Gomez, R. Villares, A. Bornat, and J. Modolell.** 1985. Molecular genetics of the *achaete-scute* gene complex of *D. melanogaster*. Cell **40**:327–338.
4. **Carrasco, A. E., W. McGinnis, W. Gehring, and E. M. De Robertis.** 1984. Cloning of an X. laevis gene expressed during early embryogenesis coding for a peptide region homologous to *Drosophila* homeotic genes. Cell **37**:409–414.
5. **Davis, R. W., and R. W. Hyman.** 1971. A study in evolution: the DNA base sequence homology between coliphages T7 and T3. J. Mol. Biol. **62**:287–301.
6. **Efstratiadis, A., J. W. Posakony, T. Maniatis, R. M. Lawn, C. O'Connell, R. A. Spritz, J. K. DeRiel, B. G. Forget, S. M. Weissman, J. L. Slightom, A. Blechl, O. Smithies, F. E. Baralle, C. C. Shoulder, and N. J. Proudfoot.** 1980. The structure and evolution of the human β-globin gene family. Cell **21**:653–668.
7. **Fjose, A., W. J. McGinnis, and W. J. Gehring.** 1985. Isolation of a homoeo box-containing gene from the *engrailed* region of *Drosophila* and the spatial distribution of its transcripts. Nature (London) **313**:284–289.
8. **Hafen, E., A. Kuroiwa, and W. J. Gehring.** 1984. Spatial distribution of transcripts from the segmentation gene *fushi tarazu* during *Drosophila* embryonic development. Cell **37**:833–841.
9. **Hayashida, H., and T. Miyata.** 1983. Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. Proc. Natl. Acad. Sci. USA **80**:2671–2675.
10. **Jeffreys, A. J.** 1981. Recent studies of gene evolution using recombinant DNA. p. 2–29. *In* R. Williamson (ed.), Genetic engineering 2. Academic Press, Inc., New York.
11. **Karn, J., S. Brenner, L. Barnett, and G. Cesareni.** 1980. Novel bacteriophage λ cloning vector. Proc. Natl. Acad. Sci. USA **77**:5172–5176.
12. **Koller, B., and H. Delius.** 1984. Intervening sequences in chloroplast genomes. Cell **36**:613–622.
13. **Kornberg, T.** 1981. *engrailed*: a gene controlling compartment and segment formation in *Drosophila*. Proc. Natl. Acad. Sci. USA **78**:1095–1099.
14. **Kornberg, T., I. Siden, P. O'Farrell, and M. Simon.** 1985. The *engrailed* locus of *Drosophila*: *in situ* localization of transcripts reveals compartment-specific expression. Cell **40**:45–53.
15. **Kuner, J. M., M. Nakanishi, Z. Ali, B. Drees, E. Gustavson, J. Theis, L. Kauvar, T. Kornberg, and P. H. O'Farrell.** 1985. Molecular cloning of *engrailed*: a gene involved in the development of pattern in *Drosophila melanogaster*. Cell **42**:309–315.
16. **Lawrence, P. A., and G. Morata.** 1976. Compartments in the wing of *Drosophila*: a study of the *engrailed* gene. Dev. Biol. **50**:321–337.
17. **Levine, M., G. M. Rubin, and R. Tjian.** 1984. Human DNA sequences homologous to a protein coding region conserved between homeotic genes of *Drosophila*. Cell **38**:667–673.
18. **Maniatis, T., E. R. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
19. **McGinnis, W., R. L. Garber, J. Wirz, A. Kuroiwa, and W. J. Gehring.** 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. Cell **37**:403–408.
20. **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 (London) **308**:428–433.
21. **Meyerowitz, E. M., and C. H. Martin.** 1984. Adjacent chromosomal regions can evolve at very different rates: evolution of the *Drosophila* 68C glue gene cluster. J. Mol. Evol. **20**:251–254.
22. **Morata, G., and P. A. Lawrence.** 1975. Control of compartment development by the *engrailed* gene in *Drosophila*. Nature (London) **255**:614–617.
23. **Nusslein-Volhard, C., and E. Wieschaus.** 1980. Mutations affecting segment number and polarity in *Drosophila*. Nature (London) **287**:795–801.
24. **O'Farrell, P. H., C. Desplan, S. DiNardo, J. Kassis, J. Kuner, E. Lim, E. Sher, J. Theis, and D. Wright.** 1985. Molecular analysis of the involvement of the *Drosophila engrailed* gene in embryonic pattern formation. UCLA Symp. Mol. Cell. Biol. **31**:489–521.
25. **Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson.** 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555–566.
26. **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.
27. **Rubin, G. M.** 1983. Dispersed repetitive DNAs in *Drosophila*. p. 329–361. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
28. **Scott, M., and A. 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. USA **81**:4115–4119.
29. **Scott, M., A. Weiner, T. Hazelrigg, B. Polisky, V. Pirrotta, F. Scalenghe, and T. Kaufman.** 1983. The molecular organization of the *Antennapedia* locus of *Drosophila*. Cell **35**:763–776.
30. **Wakimoto, B. T., and T. C. Kaufman.** 1981. Analysis of larval

segmentation in lethal genotypes associated with *Antennapedia* gene complex in *Drosophila melanogaster*. Dev. Biol. **81**:51–64.

31. **Wharton, K. A., B. Yedvobnick, V. G. Finnerty, and S. Artavanis-Tsakonas.** 1985. *opa*: a novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in *D. melanogaster*. Cell **40**:55–62.
32. **Wilson, A. C., S. S. Carlson, and T. J. White.** 1977. Biochemical evolution. Annu. Rev. Biochem. **46**:573–639.
33. **Zwiebel, L. J., V. H. Cohn, D. R. Wright, and G. P. Moore.** 1982. Evolution of single-copy DNA and the ADH gene in seven drosophilids. J. Mol. Evol. **19**:62–71.