Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8398–8402, August 1995
Genetics

Complete sequence of the bithorax complex of *Drosophila*

(Ultrabithorax/abdominal-A/Abdominal-B/glucose transporter-like/DNA analysis)

Christopher H. Martin*†, Carol A. Mayeda*†, Cheryl A. Davis*†, Cheryl L. Ericsson*†, John D. Knafels‡, David R. Mathog‡, Susan E. Celniker‡, Edward B. Lewis‡, and Michael J. Palazzolo*†§

*Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720; †Drosophila Genome Center and §Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720; and ‡Division of Biology, California Institute of Technology, Pasadena, CA 91125

Contributed by Edward B. Lewis, June 19, 1995

ABSTRACT **The bithorax complex (BX-C) of *Drosophila*, one of two complexes that act as master regulators of the body plan of the fly, is included within a sequence of 338,234 bp (SEQ89E). This paper presents the strategy used in sequencing SEQ89E and an analysis of its open reading frames. The BX-C sequence (BXCALL) contains 314,895 bp obtained by deletion of putative genes that are located at each end of SEQ89E and appear to be functionally unrelated to the BX-C. Only 1.4% of BXCALL codes for the three homeodomain-containing proteins of the complex. Principal findings include a putative ABD-A protein (ABD-AII) larger than a previously known ABD-A protein and a putative glucose transporter-like gene (1521 bp) located at or near the bithoraxoid (*bxd*), infra-abdominal-2 (*iab-2*) boundary on the opposite strand relative to that of the homeobox-containing genes.**

The bithorax complex (BX-C) of *Drosophila* is a set of master control genes that play a major role in determining the body plan of the fruit fly (1). Historically, the genetic studies of the original homeotic mutants of the BX-C, isolated principally by Bridges, Stern, Schultz, and Hollander (see ref. 2), were undertaken to test the cytologically based hypothesis of Bridges that the *Drosophila* genome contains naturally occurring tandem gene duplications (3). It was not until the discovery of the homeobox (HOX) (4, 5) that the hypothesis received strong molecular support. The high degree of conservation of these genes rapidly led to the identification of additional proteins of the two complexes and to the discovery of the genes in vertebrates as well as other invertebrates. The next stage was the identification of the various transcription units of the BX-C: those producing protein-coding transcripts—Ultrabithorax (*Ubx*) (6), abdominal-A (*abd-A*) (7), and Abdominal-B (*Abd-B*) (8–10)—and those producing noncoding transcripts—bithoraxoid (*bxd*) (11) and infra-abdominal-4 (*iab-4*) (12, 13).

Genetic studies have identified cis-regulatory regions of the BX-C that are involved in controlling the development of specific organs and body structures; 12 such regions are now known (14–16). Thus, the anterobithorax (*abx*), bithorax (*bx*), and postbithorax (*pbx*) regions function in the wild type to promote development of the third thoracic segment into a haltere-bearing instead of a wing-bearing segment. Nine regions from *bxd* to *iab-9*, inclusive, determine the pattern of differentiation of abdominal segments A1–9, respectively.

In all vertebrates and many invertebrates thus far studied, the BX-C and the related Antennapedia complex (ANT-C) (17) form a single tightly linked cluster, termed the homeotic complex (HOM-C) (18). In vertebrates the HOM-C is present in four partially redundant copies (19, 20), a situation that suggests it plays a vital and indispensable role in the development of these organisms as well. The HOM-C must have arisen before invertebrates and vertebrates diverged from a common ancestor over 500 million years ago.

With the availability of the complete sequence of the BX-C, as herein reported,¶ the stage is now set to integrate the varied types of genetic, biochemical, and developmental studies of the BX-C with the actual sequence. This paper presents an overview of the strategy used in obtaining the BX-C sequence, as well as an analysis primarily of the coding portion of that sequence. An analysis of the noncoding regions of the BX-C sequence is the subject of the accompanying paper (53).

MATERIALS AND METHODS

Directed Sequencing Strategy. The BX-C is located in the 89E region of the salivary-gland chromosomes. A sequence, SEQ89E, of 338,234 bp was generated from a set of six partially overlapping P1 bacteriophage clones (Fig. 1). The P1 library was made from an isogenic, multiple mutant strain (*y*; *cn bw sp*) (21). The DNA of these clones was sheared and the resultant fragments, averaging 3 kb in size, were subcloned. The subclones were ordered by a PCR-based screening method (22). The $\gamma\delta$ transposon was inserted into a minimal set of them to generate primer sites (23). The sites of transposon insertion were mapped by PCR and had an average spacing of 300–400 bp. Both strands were sequenced using automated fluorescent DNA sequencing technology (Applied Biosystems Dye Terminator Cycle Sequencing). Areas of ambiguity between the strands were resolved by resequencing using custom oligonucleotides as primers (24). Details of the sequencing strategy will be reported elsewhere.

ORF Analysis. SEQ89E has been analyzed for potential protein-coding regions. ORFs of >100 aa were identified by the DNA Inspector IIe ORF/peptide analysis program (Textco, West Lebanon, NH) and their codon preference was determined with the CODONPREFERENCE program (25). A total of 60 ORFs showing good *Drosophila* codon usage and an additional 15 with low codon preference but exceptional size (>200 aa) were compared with the sequences in public data bases by use of the FASTA (26) and BLASTX (27) programs for protein similarities. To search the entire sequence for putative genes we also used GENEFINDER (28) employing *Drosophila* codon usage files (Frank Eeckman, personal communication).

RESULTS

The assembled sequence generates a restriction map of *Eco*RI sites consistent with the previously determined physical map of these sites (14, 15), given that they were done for different strains.

A correlation of the molecular map of SEQ89E and the genetic map of the BX-C is shown in Fig. 1 *A* and *B*. Boundaries of 11 of the cis-regulatory regions of the complex are as yet very approx-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ORF, open reading frame; LDL, low density lipoprotein.

¶The SEQ89 sequence reported in this paper has been deposited in the GenBank data base (accession no. U31961).

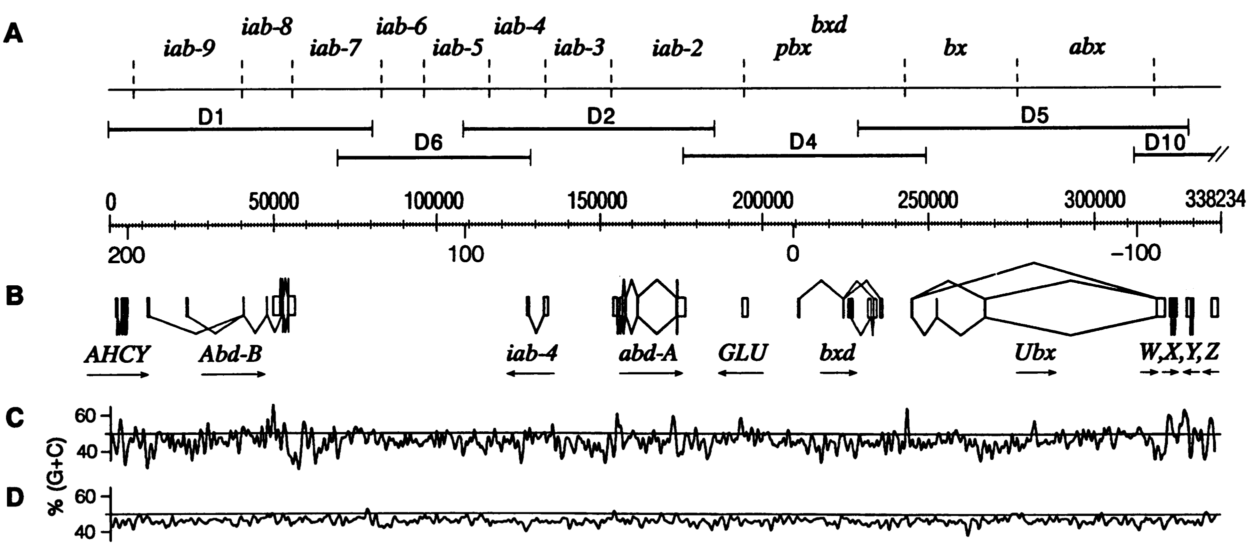

FIG. 1. Correlation of molecular and genetic maps of the BX-C. (*A*) Boundaries of cis-regulatory regions, delineated by hatched vertical lines, are only approximately known. Size and positions of P1 bacteriophage BX-C inserts (D1–D10) are shown by horizontal lines. Numbers below the line refer to the map positions in kilobases. The bases numbered 1 and 313,000 correspond approximately to +204 kb and −100 kb, respectively, on the BX-C walk. (*B*) The transcription map is based upon intron–exon mapping, cDNA, and computer analysis (see Table 1). Transcription units include homeobox-containing *Abd-B, abd-A*, and *Ubx*; the noncoding *bxd* and *iab-4*; and the open reading frames (ORFs) and predicted candidate genes. Arrows indicate the direction of transcription. The predicted genes labeled W, X, Y, and Z refer to two LDL receptor "a" repeats, serine protease-like, chaperonin-containing t-complex protein γ subunit-like, and no-on transient A-like, respectively (see Table 1). (*C*) G+C composition of SEQ89E was calculated for 1-kb windows, spaced every 100 bp. (*D*) G+C composition for a third-order Markov chain generated-random sequence of SEQ89E was calculated as in *C*.

imate and based principally on breakpoints of chromosomal rearrangements which were detected by their suppression of transvection (29). The sequence is numbered starting at the distal or *Abd-B* end of the complex in accordance with the direction of transcription. The lengths of the three homeobox-containing transcription units *Abd-B*, *abd-A*, and *Ubx* are 44.6 kb (from 11,204 to 55,760), 22.4 kb (from 153,391 to 175,816), and 78.1 kb (from 242,867 to 320,921), respectively. We have compared the sequence of the homeodomain-containing protein-coding regions of these three transcription units in BXCALL with the corresponding published sequences of those regions. Several base-pair changes were detected, presumably owing to strain variations. In the *Abd-B* unit (16) of 493 codons there are eight third-base substitutions; in the *abd-A* unit (7) of 330 codons, none; and in the *Ubx* unit (6) of 389 codons there are two changes, a third-base substitution and a first-base A → T substitution at position 244,057 resulting in a conservative amino acid change of asparagine to tyrosine.

The base composition of SEQ89E consists of 29.19% A, 29.11% T, 20.67% G, and 21.03% C. A plot of the G+C content of the complex (Fig. 1*C*) confirms the well-known higher A+T content of invertebrate cis-regulatory regions. Sustained peaks high in G+C correlate well with known or likely protein-coding regions. An exception is a single peak mapping to the third intron of the *Ubx* transcription unit.

Novel ORFs Within the BX-C. SEQ89E was searched for ORFs that have (*i*) codon usage consistent with that of other *Drosophila* genes and (*ii*) some degree of similarity to DNA sequences in the GenBank data base (Table 1). Perhaps the most remarkable ORF is located at or near the boundary between *bxd* and *iab-2*, but on the opposite strand (Fig. 1). It predicts a gene that lacks introns in its coding region (positions 195,089–193,569) and encodes a putative protein of 507 aa that starts with a methionine. It is 23.9% identical and 38.7% similar to a human glucose transporter type 3 (GLUT3) protein (Table 1 and Fig. 2*A*). These percentages are significant by the normalized alignment score (39).

A second finding is an ORF that is capable of encoding a putative second ABD-A protein and is located directly contiguous with the third exon of the *abd-A* transcription unit. It is 260 aa longer at the amino-terminal end than the previously described ABD-A protein of 330 aa (7) (Fig. 3).

Two other putative candidate genes could be of interest, although they have poor *Drosophila* codon usage. The first is an ORF of 426 aa that maps to the third intron of *Ubx*, where its position correlates, as already mentioned, with a high G+C peak; however it lacks a homolog in the public data base. Interestingly, the 3′ end of a nonpolyadenylylated transcript of 4.7 kb, expressed early in embryogenesis, apparently maps close to or within this peak (40). The second is candidate gene 5 that GENEFINDER identifies, albeit with a low score (26.22). It maps near or at the boundary of the *iab-5* and *iab-6* cis-regulatory regions and extends from bp 98,375 to bp 103,713. This putative gene contains an estimated nine exons and encodes a putative protein of 425 aa that lacks a homolog in the public data bases. It is not an essential gene. Thus the homozygote for a deficiency *Df(3)iab-4,5*^{DB} that deletes *iab-4*, *iab-5*, and part of *iab-6* (15, 16), and therefore this gene, survives as an adult and expresses only the expected homeotic transformations.

ORFs That Define the Limits of the BX-C. Fortunately SEQ89E includes several ORFs at the proximal and distal ends of the sequence that code for putative proteins (Table 1) not likely to be related to the BX-C. At the distal (telomeric) end is an ORF, commencing at bp 1258, which is a homolog of the human *S*-adenosylhomocysteine hydrolase (*AHCY*) gene. This *Drosophila AHCY*-like sequence falls within a known transcription unit, pH200, for which a cDNA has been identified (8). A comparison of the *AHCY*-like amino acid sequence with its human cognate finds 51.5% identity and 64.3% similarity over 423 aa (Table 1 and Fig. 2*B*). A high degree of amino acid similarity, 75%, exists across a region of the protein that contains the putative NAD$^+$-binding domain. Starting at bp 5991 and terminating at bp 7535 is an ORF (designated ORF1) of 515 aa that shows 30.9% identity over 97 aa with a human α-actinin protein of 888 aa. There are no more ORFs likely to code for proteins between ORF1 and the transcription start site for the most distal gene of the BX-C, *Abd-B*. The 5′ end of the complex therefore is assumed to commence after bp 7535.

Table 1. Amino acid comparisons of predicted genes

Gene	Start	Stop	No. of exons	No. of aa	No. of aa aligned[a]	Identity %	Similarity,[b] %	*Z* value[c]
Distal to BX-C								
S-adenosyl homocysteine hydrolase-like (*AHCY*-like)	1,258	5,002	7[d]	504	423	51.5	64.3	145
ORF1 (α-actinin-like)	5,991	7,535	1	515	97	30.9	47.4	14
Within BX-C								
Abd-B	23,173	55,760	9[e]	493 (I), 272 (II)	ND	ND	ND	ND
Candidate gene 5	103,713	98,375	9[f]	425	—	—	—	—
abd-A I	153,391	175,816	7[g]	330 (I)	ND	ND	ND	ND
New *abd-A II*			5[h]	590 (II)	ND	ND	ND	ND
Glucose transporter-like	195,089	193,569	1	507	507	23.9	38.7	33
Ubx	242,871	320,921	4[i]	380 (IA), 389 (IB), 363 (IIA), 372 (IIB), 346 (IVA)	ND	ND	ND	ND
Proximal to BX-C								
Low density lipoprotein (LDL)	322,431	322,835	1	135	100	35.0	42.0	15
receptor-like repeats	322,894	323,400	1	169	75	41.3	48.0	14
Serine protease-like[j]	323,432	324,602	3[k]	366	430	23.0	31.6	10
ORF2	326,067	327,074	1	336	—	—	—	—
Chaperonin-containing t-complex protein-1 γ subunit-like (*CCT*γ-like)	329,395	327,553	4[l]	532	531	69.5	80.4	122
no-on transient A-like	336,999	335,074	1	642	642	64.6	73.2	110

[a]Homologs to the predicted genes are human *AHCY* (30), *GLUT3* (31), *LDLRL* (32), and *CCT*γ (33); *Drosophila* snake serine protease (34) and no-on transient A (35); and chicken α-actinin (36). There are no known homologues to candidate gene 5 and ORF2. Splicing variants of *Abd-B* (refs. 8–10; M. Kuziora, personal communication); *abd-A* (ref. 7; S. Sakonju, personal communication), and *Ubx* (6, 37) are based on cDNA and computer analysis.

[b]The following amino acid similarities were used: F/Y, A/G, I/L/M/V, N/S/T/Q, R/K, D/E.

[c]Significance of the homology is given by the *Z* value (38). *Z* > 10 is significant.

[d]*AHCY* exons: bp 1258–1618, 1742–1802, 3078–3147, 3298–3964, 4036–4354, 4418–4661, and 4731–5002.

[e]*Abd-B* exons: bp 11,204–11,682, 23,173–23,469, 40,142–40,188, 47,337–47,514, 49,199–51,070, 51,828–52,039, 52,124–52,325, 53,590–53,804, and 53,862–55,760. Not shown is a detected, but not sequenced, *Abd-B* exon (10).

[f]Candidate gene 5 exons: 103,713–103,603, 103,550–103,489, 101,964–101,913, 101,005–100,824, 100,748–100,643, 100,602–100,535, 98,870–98,789, and 98,746–98,375.

[g]*abd-A* exons: bp 153,391–154,516, 155,540–155,698, 155,800–156,302, 157,130–157,386, 160,969–161,016, 173,083–173,306, and 173,377–175,816.

[h]Putative *abd-A* exon: bp 155,434–156,302.

[i]*Ubx* exons: bp 242,871–244,600, 251,992–252,042, 266,685–266,735, and 318,224–320,921.

[j]Although the homology for the putative serine protease is low, the MOTIFS program identified consensus sequences for the histidine and serine active sites (ITAAHC and GCSGG, respectively) (25).

[k]Serine protease exons: bp 323,432–323,767, 323,805–324,041, and 324,078–324,602.

[l]CCT γ exons: bp 329,395–329,362, 329,303–329,242, 329,118–328,658, and 328,591–327,553.

At the proximal (centrometic) end of the complex the *Ubx* transcription unit terminates at bp 320,921. The BX-C is assumed to end before bp 322,431, the start of the first of two ORFs that GENEFINDER links together. These ORFs show similarity to the class A cysteine-rich motif (41). This motif of ≈40 aa has a patterned spacing of cysteine residues and a conserved acidic region. It is found in a group of human proteins including the LDL receptor, LDL receptor-like pro-

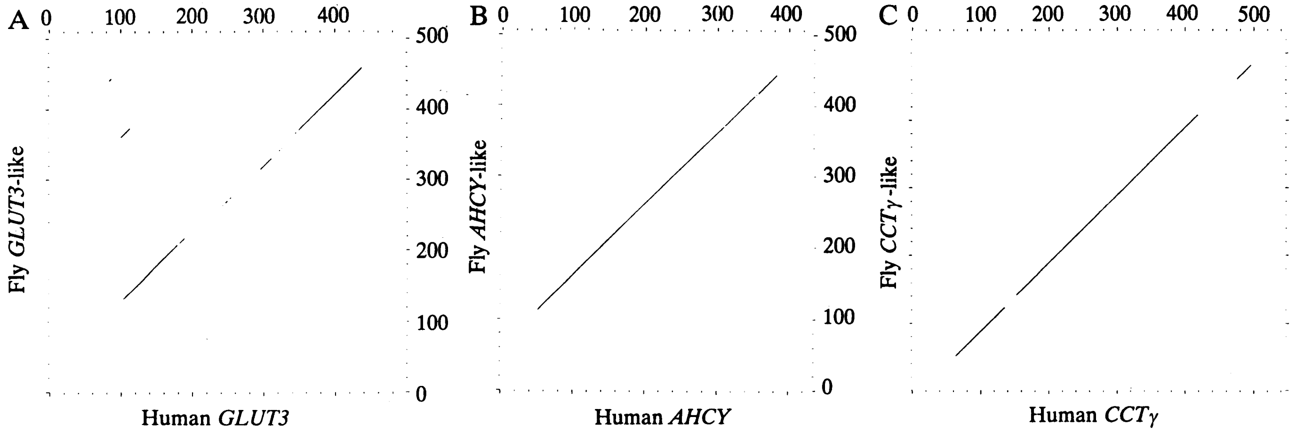

FIG. 2. Dot plot comparisons of fly and human genes in SEQ89E. (*A*) Human glucose transporter 3 (*GLUT3*) vs. fly glucose transporter-3-like (*GLUT3*-like), stringency = 40% amino acid identity. (*B*) Human *S*-adenosylhomocysteine hydrolase (*AHCY*) vs. fly *S*-adenosylhomocysteine hydrolase-like (*AHCY*-like); stringency = 50% identity. (*C*) Human chaperonin-containing t-complex polypeptide-1 γ subunit (*CCT*γ) vs. fly chaperonin-containing t-complex polypeptide-1 γ subunit-like (CCTγ-like), stringency = 70% identity. The comparison window is 100 residues for all three plots.

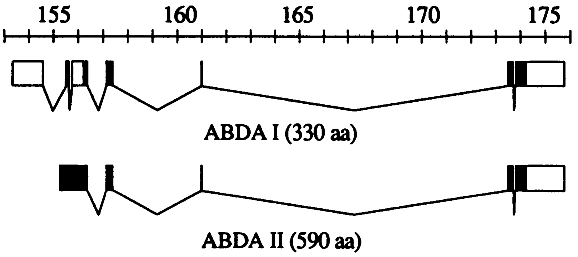

FIG. 3. Known and predicted ABD-A proteins. Numbers below the line refer to the map positions in kilobases (see Fig. 1). The transcription unit encoding ABD-AI is predicted from a previously sequenced cDNA (15) and from unpublished S1 nuclease and primer extension data (M. Lamka and S. Sakonju, personal communication). The translation start site of ABD-AII is predicted from our own sequence data, the transcription start site is not known. Filled boxes indicate the coding portions of the transcript, and open boxes indicate untranslated sequences.

tein (LRP) (32), and components of the terminal complement proteins (C9, C8α, C8β, and C7) (41). This group of proteins has the motif repeated from 7 to 31 times. Each of the two ORFs has the motif repeated twice.

The next most proximal gene (bp 323,432–324,602) is composed of three exons and encodes a predicted protein of 291 aa that is homologous to the trypsin-family serine proteases (42). A comparison with the *Drosophila* snake-, easter-, and alpha-encoded serine proteases (34) reveals homologies of ≈23% identity and ≈33% similarity. Although these identities are low, the predicted protein does contain two motifs, the histidine and serine active-site signatures (ITAAHC and GDSGG, respectively) that define all proteases (42).

Starting at bp 326,067 there is an ORF (ORF2) (Table 1) of 336 aa with good *Drosophila* codon usage but with no known homologue. Four hundred seventy-nine base pairs from this ORF, located on the opposite strand, is a gene identified by GENEFINDER (123.12) that encodes a homolog of the mammalian chaperonin-containing t-complex polypeptide-1 γ subunit (*CCT*γ) (33). A comparison of the deduced amino acid sequence (531 aa) with the mammalian cognate (556 aa) reveals 69.5% identity and 80% similarity (Fig. 2*C*). Two additional putative genes, located even more proximally, are also transcribed on the opposite strand. The first encodes a predicted protein of 642 aa from a single ORF. Comparison of this ORF with the public data bases reveals 64.6% identity and 73.2% similarity with no-on transient A (35). The last ORF in the sequence is the first coding exon of fasciclin I (43).

A sequence that represents just the BX-C itself has now been generated, eliminating from SEQ89E the presumably unrelated genes at each end of the sequence. The resultant sequence, BXCALL, has 314,895 bp and is assumed to approximate closely the BX-C.

DISCUSSION

The need to obtain the entire sequence of the BX-C has long been pressing, especially in view of the importance of understanding how the master control genes that make up the complex are themselves regulated and why there is a strong tendency for the genes of this complex and those of the related ANT-C to stay tightly linked, with their order along the chromosome being colinear with their order of expression along the body axis (1).

The pioneering work of Hogness *et al.* (40) showed that the protein-coding region of the *Ubx* transcription unit is only a small portion of the total unit—namely, 1167 bp for its longest variant. The *abd-A* and *Abd-B* protein-coding regions are also a relatively small fraction of the transcription units. In the case of the ABD-A proteins the putative ABD-AII identified here is the longer one and is encoded by 1770 bp. The larger of the two ABD-B proteins is ABD-BI and is encoded by 1479 bp. Thus 4416 bp out of 314,895 bp, or 1.4% of BXCALL, encode the homeodomain-containing proteins.

Transcripts That Do Not Code for Protein. In the *bxd* region, a complex set of transcripts ranging in size from 1.1 to 1.3 kb is detected between 3 and 6 hr of embryogenesis. They have multiple stop codons and therefore fail to code for protein (11). Similar noncoding transcripts come from the opposite strand of the *iab-4* region (12). Evidence for the possible importance of such transcripts comes from studies of the spatial distribution of RNAs encoded by the *bxd* region (44) and by the *iab-3* to *iab-7* cis-regulatory regions (13). We have not been able to predict probable candidates for such noncoding transcripts, since the high degree of degeneracy of splice-site junction sequences (45) results in vast numbers of such sites in BXCALL and, hence, excessive numbers of possible transcripts. A possible role for the *bxd* transcripts as structural RNAs that serve to facilitate the association of cis-regulatory regions has been proposed by one of us (46). The existence of a gene that generates noncoding polyadenylylated RNAs, as in the case of *bxd* and *iab-4*, has been reported in *Schizosaccharomyces pombe*. The gene, *sme2*, generates functional structural RNAs called meiRNA that are required for meiosis I when complexed with an RNA-binding protein, Mei2 (47). Other examples of noncoding RNAs are *XIST*, detected from the inactive human X chromosome (48); *H19*, an imprinted RNA derived from human chromosome 7 (49); and omega-n, an RNA derived from the third chromosome of *Drosophila* (50).

ORF Analysis. *glut-l.* The most surprising finding is an ORF encoding a glucose transporter-like (*glut-l*) gene located at the distal end of the *bxd* region, in the opposite strand relative to that of *Ubx, abd-A*, and *Abd-B*. Prior to this, no genes other than the HOX genes of the BX-C had been identified by genetic or molecular analysis. K. McCall and W. Bender (personal communication) have generated a deletion of ≈10 kb by P-factor excision that removes *glut-l*. When homozygous this deletion survives to adulthood and is phenotypically wild type except for a Uab phenotype in which the first abdominal (A1) segment is transformed toward the A2 segment. They find that the heterozygote, as well as the homozygote, has a Uab phenotype and that both genotypes are associated with misexpression of the ABD-A protein in A1. Presumably there is a negative cis-regulatory element (or elements) in the deleted region that is required for ABD-A expression, but it is unlikely that loss of *glut-l* from only one homolog would cause a dominant gain-of-function homeotic phenotype. Precedent for such genomic organization comes from the ANT-C, in which such genes as fushi tarazu, amalgam, zerknüllt, and still others are unrelated to, yet interspersed among, the HOX genes of that complex.

glut-l has no introns, at least in the coding portion of the gene. Although intronless genes are rare in *Drosophila*, a family of such genes encoding glutathione transferase (51) is known. The human cognate (*GLUT3*) and the human *HOXC* cluster are located in chromosome 12 but in different arms.

In mammalian cells, multiple glucose transporter protein isoforms are encoded by a family of glucose transporter genes (31). Our identification of a glucose transporter-like gene in *Drosophila* is consistent with the demonstration of a glucose transport function in *Drosophila* Kc cells (52). Since adults lacking *glut-l* survive we speculate that, as in the mammalian case, this putative gene may belong to a family of *Drosophila* glucose transporter genes and hence be at least partially dispensable.

abd-A. Support for an additional ABD-A protein comes from two studies. The first is an analysis by B. Appel and S. Sakonju (personal communication) of a Western blot that identifies two ABD-A proteins, 36 and 70 kDa, the larger of which presumably corresponds to our ABD-AII protein. The second study is based on an analysis of an *abd-A* mutant, *abd-A*^{MX2}. Although this mutant has a deletion (≈3 kb) for the known *abd-A* transcription start site, the homozygous mutant embryos still express an ABD-A protein in the epidermis and central nervous system (7). The distal end of this deletion maps at least 200 bp from the initiation codon of the putative *abd-A* ORF, and we speculate that the expression seen in *abd-A*^{MX2} mutants is due to the ABD-AII protein.

It is surprising how much of the control of development of the thorax and abdomen of the fly is accomplished by only three homeobox-containing genes. Evidently much of that control rests in the noncoding regions that occur either in the introns of those genes or in the ≈100-kb region between *abd-A* and *Abd-B* and the 67-kb region between *Ubx* and *abd-A*. An unknown fraction of the sequence may involve additional noncoding RNAs of the *bxd* and *iab-4* types, but whether such RNAs play a regulatory role is uncertain. An appreciable fraction of the sequence is expected to contain DNA sequence motifs. These motifs may play a variety of regulatory roles and are discussed in the accompanying paper (53).

We thank the sequencing group of the Berkeley *Drosophila* Genome Project, including Michelle Lee, Vivi Truong, Rebecca Zhang, Melvin Li, Kathryn Gunning, Annie Chiang, Marnel Bondoc, Kala Kowtha, Migdad Machrus, and Tom Cloutier, for their superb technical assistance. We thank Joe Jaklevic and his Lawrence Berkeley Laboratory Human Genome Center Automation group; Frank Eeckman, Ed Theil, and the Lawrence Berkeley Laboratory Human Genome Center Informatics group; Bill Kimmerly and the *Drosophila* Physical Mapping Group; and Suzanna Lewis and the *Drosophila* Genome Center Informatics group for their contributions to the sequencing effort. Analysis of the sequence was done by the California Institute of Technology group. We thank Mary Raney, Victor Hsu, Gretl Hornung, Mallory Zhang, and John Hubenschmidt for technical assistance. We thank Welcome Bender, Mike Kuziora, Michele Lamka, Kim McCall, and Shige Sakonju for communicating their unpublished results. We thank Welcome Bender, Howard Lipshitz, and Joanne Topol for critical reading of the manuscript. The sequencing was done by the Berkeley *Drosophila* Genome Project, which is a consortium of the *Drosophila* Genome Center [supported by grants from the National Center for Human Genome Research (P50-HG00750 to Gerald Rubin and HG00837-03 to M.J.P. and C.H.M.)] and the Human Genome Center of the Lawrence Berkeley Laboratory, Mohan Narla, Acting Director (supported by the U.S. Department of Energy). M.J.P. is a Lucille P. Markey Scholar and his effort was funded in part by the Lucille P. Markey Charitable Trust. The analysis was supported by research grants to E.B.L. from the National Institutes of Health (HD06331 and HD30727) and the March of Dimes.

1. Lewis, E. B. (1978) *Nature (London)* **276,** 565–570.
2. Lindsley, D. L. & Zimm, G. G. (1992) *The Genome of Drosophila melanogaster* (Academic, San Diego).
3. Lewis, E. B. (1951) *Cold Spring Harbor Symp. Quant. Biol.* **16,** 159–174.
4. McGinnis, W., Levine, M., Hafen, E., Kuroiwa, A. & Gehring, W. J. (1984) *Nature (London)* **308,** 428–433.
5. Scott, M. P. & Weiner, A. J. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4115–4119.
6. Kornfeld, K., Saint, R. B., Beachy, P. A., Harte, P. J., Peattie, D. A. & Hogness, D. S. (1989) *Genes Dev.* **3,** 243–258.
7. Karch, F., Bender, W. & Weiffenbach, B. (1990) *Genes Dev.* **4,** 1573–1587.
8. DeLorenzi, M., Ali, N., Saari, G., Henry, C., Wilcox, M. & Bienz, M. (1988) *EMBO J.* **7,** 3223–3231.
9. Celniker, S. E., Keelan, D. J. & Lewis, E. B. (1989) *Genes Dev.* **3,** 1425–1437.
10. Zavortink, M. & Sakonju, S. (1989) *Genes Dev.* **3,** 1969–1981.
11. Lipshitz, H. D., Peattie, D. A. & Hogness, D. S. (1987) *Genes Dev.* **1,** 307–322.
12. Cumberledge, S., Zaratzian, A. & Sakonju, S. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 3259–3263.
13. Sanchez-Herrero, E. & Akam, M. (1989) *Development (Cambridge, U.K.)* **107,** 321–329.
14. Bender, W., Akam, M., Karch, F. A., Beachy, P. A., Peifer, M., Spierer, P., Lewis, E. B. & Hogness, D. S. (1983) *Science* **221,** 23–29.
15. Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M. & Lewis, E. B. (1985) *Cell* **43,** 81–96.
16. Celniker, S. E., Sharma, S., Keelan, D. & Lewis, E. B. (1990) *EMBO J.* **9,** 4277–4286.
17. Kaufman, T. C., Lewis, R. & Wakimoto, B. (1980) *Genetics* **94,** 115–133.
18. Beeman, R. W. (1987) *Nature (London)* **327,** 247–249.
19. Krumlauf, R., Holland, P. W. H., McVey, J. H. & Horgan, B. L. M. (1987) *Development (Cambridge, U.K.)* **99,** 603–617.
20. Boncinelli, E., Simeone, A., La Volpe, A., Faiella, A., Fidanza, V., Acampora, D. & Scotto, L. (1988) *Hum. Reprod.* **3,** 880–886.
21. Hartl, D. L., Nurminsky, D. I., Jones, R. W. & Lozovskaya, E. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6824–6829.
22. Yoshida, K., Strathmann, M. P., Mayeda, C. A., Martin, C. H. & Palazzolo, M. J. (1993) *Nucleic Acids Res.* **21,** 3553–3562.
23. Hamilton, B. A., Mayeda, C. A., Simon, M. I., Meyerowitz, E. M. & Palazzolo, M. J. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 1247–1250.
24. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
25. Genetics Computer Group (1994) *Program Manual for the Wisconsin Package, Version 8* (Genetics Computer Group, Madison, WI).
26. Pearson, W. R. (1990) *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences* (Academic, San Diego).
27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
28. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., *et al.* (1994) *Nature (London)* **368,** 32–38.
29. Lewis, E. B. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50,** 155–164.
30. Coulter-Karis, D. E. & Hershfield, M. S. (1989) *Ann. Hum. Genet.* **53,** 169–175.
31. Gould, G. W. & Holman, G. D. (1993) *Biochem. J.* **295,** 329–341.
32. Herz, J. J., Hamann, U., Rogne, S., Myklebost, O., Gausepohl, H. & Stanley, K. K. (1988) *EMBO J.* **7,** 4119–4127.
33. Kubota, H., Hynes, G., Carne, A., Ashworth, A. & Willison, K. (1994) *Curr. Biol.* **4,** 89–99.
34. Smith, C., Giordano, H. & DeLotto, R. (1994) *Genetics* **136,** 1355–1365.
35. Jones, K. R. & Rubin, G. M. (1990) *Neuron* **4,** 711–723.
36. Baron, M. D., Davison, M. D., Jones, P. & Critchley, D. R. (1987) *J. Biol. Chem.* **262,** 17623–17629.
37. O'Connor, M. B., Binari, R., Perkins, L. A. & Bender, W. (1988) *EMBO J.* **7,** 435–445.
38. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
39. Doolittle, R. F. (1986) *Of URFS and ORFS* (University Science Books, Mill Valley, CA).
40. Hogness, D. S., Lipshitz, H. D., Beachy, P. A., Peattie, D. A., Saint, R. B., Goldschmidt-Clermont, M., Harte, P. J., Gavis, E. R. & Helfant, S. L. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50,** 181–194.
41. Stanley, K. K., Page, M., Campbell, A. K. & Luzio, J. P. (1986) *J. Mol. Immunol.* **23,** 451–458.
42. Brenner, S. (1988) *Nature (London)* **334,** 528–530.
43. Zinn, K., Mcallister, L. & Goodman, C. S. (1988) *Cell* **53,** 577–587.
44. Akam, M. E., Martinez-Arias, A., Wenzierl, R. & Wilde, C. D. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50,** 195–200.
45. Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Field, C. (1992) *Nucleic Acids Res.* **20,** 4255–4262.
46. Mathog, D. (1990) *Genetics* **125,** 371–382.
47. Watanabe, Y. & Yamamoto, M. (1994) *Cell* **78,** 487–498.
48. Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J. & Willard, H. F. (1992) *Cell* **71,** 527–542.
49. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. (1990) *Mol. Cell. Biol.* **10,** 28–36.
50. Hogan, N. C., Slot, F., Traverse, K. L., Garbe, J. C., Bendena, W. G. & Pardue, M. (1995) *Genetics* **139,** 1611–1621.
51. Toung, Y. S., Hsieh, T. & Tu, C. D. (1993) *J. Biol. Chem.* **268,** 9737–9746.
52. Wang, M. & Wang, C. (1993) *Fed. Eur. Biochem. Soc.* **317,** 241–244.
53. Lewis, E. B., Knafels, J. D., Mathog, D. R. & Celniker, S. E. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8403–8407.