Conserved Arrangement of Nested Genes at the Drosophila *Gart* Locus

Steven Henikoff and Mohammad K. Eghtedarzadeh

Fred Hutchinson Cancer Research Center, Seattle, Washington 98104

Manuscript received July 13, 1987
Accepted August 26, 1987

ABSTRACT

The *Drosophila melanogaster Gart* gene encodes three enzymatic activities in the pathway for purine *de novo* synthesis. Alternative processing of the primary transcript leads to the synthesis of two overlapping polypeptides. The coding sequence for both polypeptides is interrupted by an intron that contains a functional cuticle protein gene encoded on the opposite DNA strand. Here we show that this nested organization also exists at the homologous locus of a distantly related species, *Drosophila pseudoobscura*. In both species, the intronic cuticle gene is expressed in wandering larvae and in prepupae. Remarkably, there are 24 different highly conserved noncoding segments within the intron containing the cuticle gene. These are found upstream of the transcriptional start, at the 3′ end, and even within the single intronic gene intron. Other introns in the purine gene, including the intron at which alternative processing occurs, show no such homologies. It seems likely that at least some of the conserved noncoding regions are involved in specifying the high level developmental expression of the cuticle gene. We discuss the possibility that shared *cis*-acting regulatory sites might enhance transcription of both genes and help explain their nested arrangement.

PERHAPS the most surprising discovery revealed by detailed molecular study of genes from higher eukaryotes was the existence of introns. Although baffling at first (BROKER *et al.* 1978), introns are now known to be functionally significant, as their removal by splicing sometimes allows for multiple protein products from a single gene by alternative processing (LEFF, ROSENFELD and EVANS 1986). In some cases, the existence of introns also appears to have allowed more rapid evolution of proteins by mechanisms such as exon shuffling, indicating an evolutionary role as well (SUDHOF *et al.* 1984). Recently, yet another role was found for an intron, when a *Drosophila melanogaster* intron was discovered to have a gene entirely within it (HENIKOFF *et al.* 1986a). This novel "nested" arrangement of the two genes has no obvious functional or evolutionary significance, since the two genes encode quite unrelated proteins on opposite DNA strands. The extent to which nested genes are common features of eukaryotic chromosomes is not known. Although no other cases of completely nested genes have been reported, more recently described examples of partially overlapping genes on opposite strands in Drosophila (SPENCER, GIETZ and HODGETTS 1986) and rodents (WILLIAMS and FRIED 1986; ADELMAN *et al.* 1987) suggest that the strict separation of genes along the eukaryotic chromosome is often violated. Whether there is any functional or evolutionary significance to these unconventional genetic arrangements is an open question.

This report concerns the organization of the Drosophila *Gart* locus, a gene encoding three purine pathway enzymatic activities on one strand and a cuticle protein on the other strand (HENIKOFF *et al.* 1986a). The cuticle gene lies entirely within the first intron of the purine pathway gene. This intron interrupts a portion of the protein-coding sequence that is homologous to its uninterrupted counterpart in yeast (HENIKOFF 1986). Since the intronic gene is a member of a diverged family of cuticle genes, it must have evolved elsewhere and been inserted into its present position. If this transposition were a recent event or the resulting arrangement detrimental, then distantly related Drosophila species might be expected to lack this gene within the purine gene intron. If, however, other Drosophila show the same nested organization, then this arrangement might be interpreted as having some functional or evolutionary significance.

Here we describe the organization, complete sequence, transcripts and proteins of the *Gart* locus from *Drosophila pseudoobscura*, a species thought to share a common ancestor with *D. melanogaster* about 45 million years ago (BEVERLEY and WILSON 1984). All organizational features of the gene, including alternative processing of the purine gene and the presence of a similarly expressed cuticle gene homologue within the first intron, are conserved. Our findings demonstrate that this first reported example of nested genes is evolutionarily stable and lead us to favor a functional basis for the arrangement.

MATERIALS AND METHODS

Genomic library screening: A *D. pseudoobscura* λ *Sau*3A genomic library constructed by C. LANGLEY using the strain

The sequence data presented in this article have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00606.

est-100 (of F. AYALA) and provided by D. CAVENER was used to infect *Escherichia coli* strain Q359. Infection, plating, plaque lifting, hybridization screening and plaque purification were carried out using standard techniques (MANIATIS, FRITSCH and SAMBROOK 1982). Probes were synthesized from single-stranded templates as described (HENIKOFF *et al.* 1986a).

***In situ* hybridization:** Slides were prepared from squashes of wandering third instar larvae and hybridized as described by SIMON *et al.* (1985). Nick-translated probes were prepared by a modification of the standard procedure (MANIATIS, FRITSCH and SAMBROOK 1982) except that biotinylated dUTP (ENZO Biochemical) was substituted for dUTP and enough deoxyribonuclease was added to give an average fragment size of about 150 base pairs (bp). As pointed out previously, the efficiency of hybridization *in situ* to polytene chromosomes is maximized when probes of this size are used (HENIKOFF and MESELSON 1977). Staining of biotinylated DNA using streptavidin and horseradish peroxidase was performed using "Bio-probe" reagents from ENZO Biochemical, following the procedure recommended by the manufacturer.

"Phagemid" subcloning and DNA sequencing: Purified λ DNA was digested with either *Eco*RI or *Sal*I and subcloned into either vector pEMBL18+ (DENTE, CESARENI and CORTESE 1983) or pVZ1. The latter is a derivative of "Bluescribe(+)" (Vector Cloning Systems) that contains an extended polylinker segment inserted into the *Eco*RI site. This segment adds *Not*I, *Nar*I, *Bbe*I, *Nsi*I, *Bst*XI and *Apa*I sites between the *Eco*RI site and the T7 promoter sequence of the Bluescribe(+) vector. This "phagemid" allows ColE1 plasmids to be replicated as single strands and secreted as filamentous phage-like particles in the presence of an M13 or F1 phage helper (DENTE, CESARENI and CORTESE 1983). The 4.1-kilobase (kb) *Eco*RI fragment derived from DpGar2 and the 6.3-kb *Sal*I fragment of DpGar1 (Figure 1) were used to construct deletion derivatives for DNA sequencing (HENIKOFF 1987) by the chain termination method (SANGER, NICKLEN and COULSON 1977). The short intervening region including overlaps was sequenced using a subcloned derivative of DpGar1 and synthetic oligonucleotide primers. The region to the 3′ side of the 6.3-kb *Sal*I fragment was sequenced using a *Kpn*I fragment derived from the right side of DpGar1 and synthetic oligonucleotide primers. Except for the most 3′ 200 bp, all sequence was verified by determining both strands.

Antibody preparation and Western blotting: A rabbit was injected with purified GARS-AIRS-GART polypeptide which had been isolated as described (HENIKOFF *et al.* 1986b) and mixed with Freund's adjuvant following standard methods. IgG was isolated from the resulting sera and used for affinity purification (GENTRY *et al.* 1983) by coupling *D. melanogaster* SL2 cell-line extract to a first column for removal of non-specific IgG and purified GARS-AIRS-GART protein to a second column for binding of specific antibody. Coupling to a solid support followed the procedure of BASSIRI, DVORAK and UTIGER (1979). Preparation of Drosophila extracts and polyacrylamide gel electrophoresis was carried out as described (HENIKOFF *et al.* 1986c). Western blotting, antibody probing and treatment with ^{125}I-labeled protein A followed standard techniques (TOWBIN, STAEHELIN and GORDON 1979) using nonfat dry milk as a blocking agent (JOHNSON *et al.* 1984).

RNA analysis: Isolation of poly(A)$^+$ RNA using guanidinium isothiocyanate was followed by oligo(dT)-cellulose fractionation as described by MANIATIS, FRITSCH and SAMBROOK (1982). Total cellular RNA was purified as described by AUFFREY and ROUGEON (1980). Northern blot analysis was carried out according to the procedure of THOMAS (1983).

Computer methods: Sequence data were entered directly into a computer file and verified using a sonic digitizer and editor software obtained from Riverside Scientific Enterprises (18332 57th Ave. N.E., Seattle, Washington 98155). Sequence display, restriction site searches, alignments, dot matrix analyses and homology searches were performed using the GENEPRO software package also obtained from Riverside.

RESULTS

Isolation of *D. pseudoobscura Gart*: A λ EMBL4 genomic library of *D. pseudoobscura* DNA was screened at normal hybridization stringency using two probes on duplicate filters. One probe corresponded to exon 2 and the other to exon 7 of the *D. melanogaster Gart* purine gene (HENIKOFF *et al.* 1986b). Five clones that appeared to be positive for both probes were chosen initially and restriction mapped. Four of the clones showed comigrating restriction fragments, some of which also hybridized to the *D. melanogaster Gart* purine and cuticle gene probes on Southern blots (data not shown). The most extensive clone, DpGar1, was chosen for subcloning and sequencing. None of these clones included the most 5′ portion of the *Gart* purine gene, so that a second set of three clones was isolated, screening with the most upstream *Eco*RI fragment of the DpGar1 insert. Restriction mapping and Southern blotting indicated that two of these clones (DpGar2,3) included the regions homologous to *D. melanogaster Gart* exon 1 (data not shown). The restriction maps for the inserts of clones DpGar1–3 are shown in Figure 1.

Hybridization *in situ* to *D. pseudoobscura* polytene chromosomes was performed using an homologous probe corresponding approximately to the upstream half of the gene. Figure 2 shows the typical labeling pattern. A single site of hybridization is seen at the proximal side of division 88 in the proximal portion of chromosome *4* (STOCKER and KASTRITSIS 1972). *D. pseudoobscura* chromosome *4* is thought to correspond to *D. melanogaster* chromosome arm *2L* (PATTERSON and STONE 1952). As *D. melanogaster Gart* is in the distal portion of 2L, at subdivision 27D (HENIKOFF *et al.* 1981; D. NASH personal communication; S. HENIKOFF, unpublished results), it is likely that these two species differ by at least one paracentric inversion on this chromosome arm. Many such inversions are known to have occurred during evolution of the subgenus *Sophophora*, particularly in the line leading to present-day *D. pseudoobscura* (PATTERSON and STONE 1952). We conclude that the clones we have isolated are true homologues of *D. melanogaster Gart* and are derived from a segment present at a single cytological location.

Sequence of *D. pseudoobscura Gart* and comparison to *D. melanogaster*: Overlapping segments from

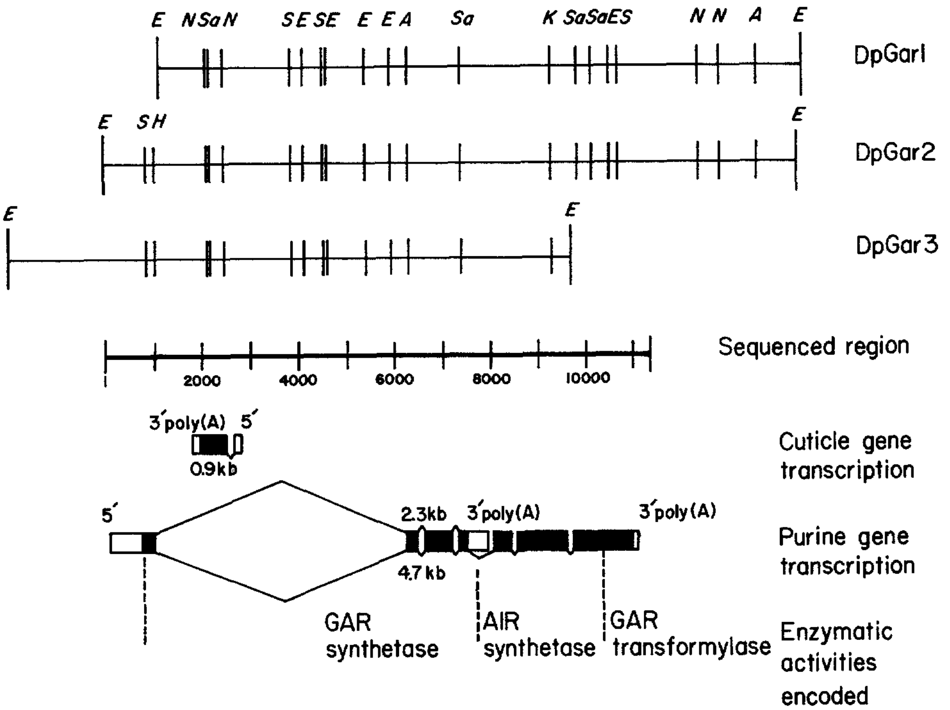
E NSaN SESE E EA Sa K SaSaES N N A E
DpGarl
E SH E
DpGar2
E E
DpGar3
Sequenced region
1 2000 4000 6000 8000 10000
3'poly(A) 5'
0.9kb
Cuticle gene transcription
5'
2.3kb 3'poly(A) 3'poly(A)
4.7kb
Purine gene transcription
GAR synthetase
AIR synthetase
GAR transformylase
Enzymatic activities encoded

FIGURE 1.—Restriction and transcriptional maps of the *D. pseudoobscura Gart* locus. The λ inserts used in the analysis (DpGarl-3) are shown above. The region that was sequenced and the transcriptional map is shown below. Protein-coding sequence is indicated by *filled boxes*. Restriction sites are: *Eco*RI (E), *Not*I (N), *Sac*I (Sa), *Sal*I (S), *Apa*I (A), *Kpn*I (K) and *Hin*dIII (H).

DpGar1 and DpGar2 were subcloned into "phagemid" vectors and used for DNA sequencing. The entire sequence of the purine gene coding strand, including 5′ and 3′ flanking material (11,392 bp), is shown in Figure 3. Nucleotide 1 corresponds to the first base of the genomic sequence determined, which is the first base of the DpGar2 insert.

The previously determined sequence of *D. melanogaster Gart* extended upstream of the transcriptional start site for only 51 bp (HENIKOFF *et al.* 1986a). We have now determined a total of 614 bp to the 5′ side (M. EGHTEDARZADEH, unpublished results), extending to an *Xba*I site mapped previously (HENIKOFF *et al.* 1986b). A comparison of this 9953-bp *D. melanogaster* sequence with the 11,392-bp *D. pseudoobscura* sequence is summarized in the form of a dot matrix plot (Fig. 4). A single dot is placed wherever 20 matches are found within a stretch of 25 bases. This degree of homology is highly unlikely to occur by chance; unrelated sequences of similar length generally show no dots (data not shown). Diagonal lines are produced where extensive homology exists between the two sequences being compared. Several very prominent diagonals are seen, corresponding to highly homologous regions. Only a single off-diagonal dot is seen. This dot corresponds to a portion of a diverged repeat described earlier (HENIKOFF, SLOAN and KELLY 1983; HENIKOFF 1986).

The previously determined transcriptional and translational (*filled boxes*) map of the *D. melanogaster* locus is shown along the y-axis (HENIKOFF *et al.* 1986a). Each coding region of the *D. melanogaster* locus has an homologous counterpart in *D. pseudoobscura*, both for the purine gene and the cuticle gene. This indicates a similar overall structure of *Gart* in the two species. Although the spacing between the homologous regions varies, they are all in the same order in the two species. When these homologous sequences are aligned, each of the experimentally determined intron-exon boundaries of *D. melanogaster* corresponds to the appropriate location of a consensus splice site in *D. pseudoobscura* (Table 1), for both the purine gene and the cuticle gene. These striking similarities indicate that the nested structure also exists in *D. pseudoobscura*.

Protein comparisons: Based on the unambiguous nucleotide sequence alignments in the coding regions, the amino acid (aa) sequences of the *Gart* locus proteins can be predicted. In *D. melanogaster* there are three such proteins: a 1353-aa polypeptide encoding GAR synthetase, AIR synthetase and GAR transformylase in that order from amino to carboxy terminus (GARS-AIRS-GART), a 434-aa polypeptide encoding GAR synthetase alone and a 184-aa predicted cuticle protein (HENIKOFF *et al.* 1986a,b). These correspond to predicted proteins that are respectively 1364 aa, 434 aa and 192 aa in *D. pseudoobscura* (Figure 5). The *D. pseudoobscura* GARS-AIRS-GART polypeptide is predicted to have an extra 2 aa that lie between the duplicated AIR synthetase domains, and an extra 9

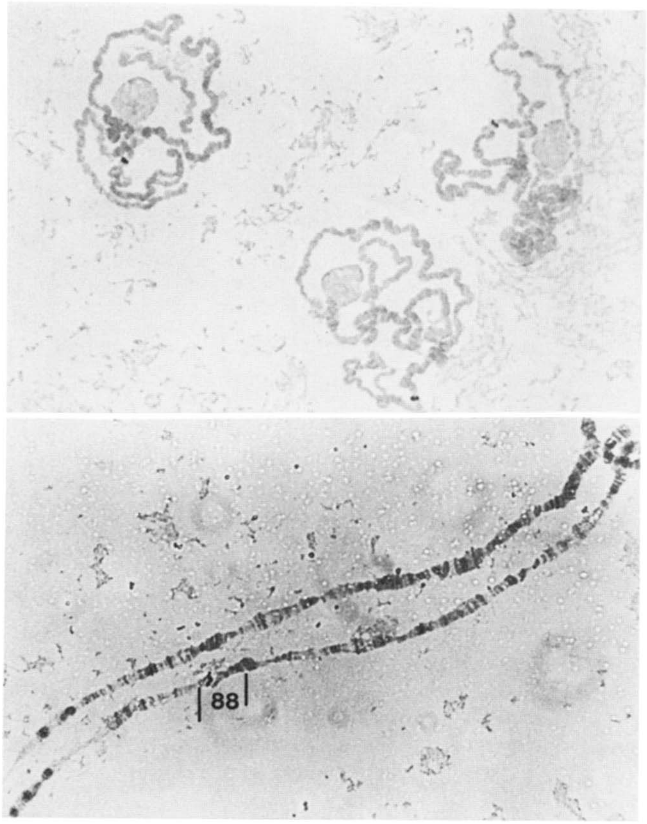

FIGURE 2.—*In situ* hybridization of *D. pseudoobscura Gart* DNA to homologous larval salivary gland polytene chromosomes. The 5.3-kb *Hin*dIII-*Apa*I fragment includes nearly all of the purine gene first intron with the cuticle gene. Three complete nuclei are shown, each with a single site of hybridization. Below is a stretched example of Chromosome *4*, showing labeling in the proximal portion of division 88.

aa at or near the poorly conserved COOH-terminal (Figure 5A). The predicted GAR synthetase polypeptide aligns throughout its length. The cuticle protein aligns for the first 4 aa comprising the coding portion of the first exon in each (Figure 5B). However, the remainder of the hydrophobic region that is presumed to be the signal peptide shows little homology and includes a 3-aa insertion in *D. pseudoobscura.* The remainder of the sequence aligns except for the last 15 aa of *D. melanogaster* and the last 20 aa of *D. pseudoobscura.* With these alignments, as shown in Figure 5, GARS-AIRS-GART is 86% homologous and the cuticle protein is 79% homologous overall between the two species.

Western blot analysis was carried out to test the predicted alignments. Crude extracts from adult *D. melanogaster* and *D. pseudoobscura* were electrophoresed on SDS polyacrylamide gels, transferred to nitrocellulose, and probed with affinity-purified antibody to purified *D. melanogaster* GARS-AIRS-GART. The *D. melanogaster* 145-kilodalton (kD) GARS-AIRS-GART polypeptide is seen in a purified preparation (Figure 6, GARS-AIRS-GART), in transformed cell-line overproducer extract (HGAR15), in 50-fold higher levels of untransformed cell-line extract (SL2) and in similarly high levels of *D. melanogaster* adult extract (*mel.*). The *D. pseudoobscura* extract (*pseudo.*) has a cross-reacting polypeptide that appears to be slightly larger than the 145-kD *D. melanogaster* GARS-AIRS-GART, consistent with its 11 aa (1.2 kD) larger predicted size. This slight difference is seen in adjacent lanes that have been alternately loaded with equal amounts of adult extract from the two species. This result provides direct evidence that the predicted *D. pseudoobscura* GARS-AIRS-GART polypeptide corresponds to the experimentally determined *D. melanogaster* polypeptide. The *D. melanogaster* 45-kD GARS polypeptide is detected by the antibody in an overproducer extract whether or not it is depleted in the 145-kD polypeptide (GARS, HGAR15). This confirms our previous detection of this protein in overproducer extracts by silver staining (HENIKOFF *et al.* 1986b). However, the GARS polypeptide is not detected in either *D. melanogaster* or *D. pseudoobscura* at levels that are above those of the more prominent background bands. This sensitivity problem is likely exacerbated by the high levels of comigrating proteins in this molecular weight range that could reduce accessibility of the smaller protein to the antibody.

Also electrophoresed in this experiment is adult extract from *ade*3^1, a purine auxotrophic strain with a single nucleotide substitution in the GAR transformylase domain which eliminates this single activity without affecting the other two (HENIKOFF *et al.* 1986c). The presence of normal amounts of cross-reacting material for GARS-AIRS-GART confirms our earlier conclusion concerning the lack of effect of this mutation on the presence of the proteins.

As we have been unable to raise specific antisera to the *D. melanogaster* cuticle protein, a comparable analysis is not possible. Thus far, the particular polypeptide encoded by this intronic cuticle gene has not been identified, perhaps because it is likely to be less abundant than other pupal cuticle proteins of this size class (HENIKOFF *et al.* 1986a; J. FRISTROM, personal communication). Nevertheless, the high degree of amino acid sequence homology, not much less than that of GARS-AIRS-GART, demonstrates that this gene must encode a conserved protein.

Silent substitutions: The degree of nucleotide substitution that has occurred since the latest common ancestor between these two species can be calculated by extrapolation from measurements on more closely related species (BRITTEN 1986) and estimates of the time since divergence (BEVERLEY and WILSON 1984). However, such a calculation makes untested assumptions and is uncertain in the absence of a dated fossil record for Drosophila. For our purposes, it is important to know if homologous regions indicate functional conservation, that is, whether these two species

```
CCGGATCTTTTTACGACTATTTGGTAGAATTTTATTATACCTGAAATATTGTTCCTATTAGGACGATGACCAGTGTGTACCTATCATAGCATTTATCTAATGCTCGCAGGAGAAACAGCT   120
GTTCTTTCCTCTCTCTCTATATTTTGCCTTTTATTTTTTTTTGAATTATTTTCATACGGAATAACAAATTTACGACTGATAAGAAGTGCAATTTGAGTTACAGTTACATTTACCCCCAGC   240
AGGGGTCTGGAAAGGCTATCGCCGACGTGTTTGTGTTATTTGTTTGAGCTAAAACAACAACAATTATTGCCGCGGCGAAAAATAAGTGCAATGCTGTGGATTGGGTCGCCTCCTAACGA   360
CGCGGCCCACGGAGGTTTTCTTATCGCGGTCACTGACGCGTCACAGTTGCTCTTACCTAGGCCTAGGGCCGCACACATAAAAAATATTTAATTCAAATTCAAGATTTCAGTTCAAAACGAG   480
GCGTCGTGCGGTGCGGTTTTACGGATCTGATCAGCCAAAAAAAAAAAAAAACAAATACATAAAACACGCAAAACCTGGCTCTCTTTCTCCCTCTCTCAAGCTCTCTCTCCATCGCGTCTA   600
TTAAACATACTGTGCTCTCACTCTCACTCTCTGGAAACGCTGCTGCTGTTGCTGCTGCCCGTTCAACAACTTGCAACACGTGGGCGAGAAAACGAAGAAAGAATTATAAATTCTTAAAAA   720
CGTGATTAGTTTTTGTTCTTATCGGCCAGTAAT
                                 ATGTCGCATAGCGTCTTGGTCATCGGCAGCGGCGGGCGGGAGCATGCCATCTGCTGGAAATTGTCGCAGTCGACGCTGGTGAAGCAG   840
ATCTATGCCCTGCCTGGCAGCTTTGGCATCCAACAGGTGGAAAAGTGCCGCAATTTGGATGCAAAAGTCTTGGATCCAAAGGATTTTGAA
                                                                                          GTGAGTTGAATGATTCCCCCAGCGACCTTG   960
GATTAACCTTCGCGAGTGACTTACAGTTACAGTGTGTGTGCCTGTGTGTGTGTGCACGCTATTTGGGGTGAAGCTTGAGCGTGGAATATTCGGCATTTACTTCGTTTTGGGTATTTTTTT  1080
ATTGCCGTCTGTGATACATTTTTATGGCTTCTTGAGAGATCTTTGAGATCGTAATGTATATTTTTATGATGCCCCAGAATGTTTTATGCTAAGCCGCCAATAAATCCGCAATCGGAACGA  1200
TCTTAAAGCTGATCATTCCATTGCACTGCGATCGCTGCTAATCGGTTTAACTTTAACCCTTGAGCGCGCCTTGGATTGCCATTGCAGCTGACGAAAATCGTGGCAAATATTGCAATTCGA  1320
GATTATTATTATATGTAGATCTGTGTAACGAGAGTCTGTCAATACGTTATCAATTGCAGGCTGGGATTTGCGTGCATTATTCGAGGTCTAGCAGCTTAACATTTTTGTAATGCATCCAAA  1440
TTGATTTTAATCATATTTGTGGCTGTCTTTTCTGTGTGGCTTAGTCGGCTCGTAATGTTGATAGATTAAGTGTGATTAGCAGAGAAAATGTATTCTACAATAATAAAGTATATTTGCTCG  1560
TTTAATGGGACTGTTGTCAGGGGACCAGGGTTTTGCTTGTGCAAAACGATCCTCATACGATCACGATTCCACCCCCCTATTGTCTACCCATGTCTAAACATAAATTGAAACTGTCTAGTT  1680
TAATTATCTAAATAAAGAATCAAATGCCAAAAGATTGTTCCATCGCGTTCCCCAGACATTGCCTCACATTAGCGCCTTCTTTTGCGACAGACAACAACAAAAATAAACACAAAATTGAGG  1800
CCTTTGTTCTTTGTTTTCGCATTATAGCTTATGTTTTGCTTAGCACTTCTATATTTATTGTTTTTGTGGCTTAATTAATTGTTAAGATTGAAATTCACAGATTTGCATGGTGTTCTATTG  1920
AGACCTGCCCCCAGCACCGATCAACGATCACCGATCATCGGATCA
                                             TCGGCGTCTCGAATCGTTTTGACGCTGTCTAAGAGGTCGCCGTGTTGGGGCATCCGATAGCGTGGGTGGATGGGT  2040
CAGATAGATGGTCTTTGGTGTGGAACTTGGCCGGGAACGAGGTGTCGTCTGGTCCTGGATGTTCCGGGTGTAGACATTAAAAGCGGCCGCATCGTGGGCCACTGTCTTGAGCTCGCCCGT  2160
ATAGTAATCCTTCACCTGGTACGGATGCGCGCGAATGTACTCCAGGGCGCGCAGGATGTAGTCGGGCACTTTGGGAATGTGATCACCCACGGGATGATAGCCGGTCTCGTCGGCCACATA  2280
GCTAACACTGATCACAGATCCTTCGGGGGAGGTGTAACTGGAGCCGCCCTGAACGGAGACGCCTCCGAGGCCCTCCTGAGTGGCCGAAATGCCATTGGATGTCTCGTAGGCATAGCGATA  2400
GTTACCATCCCGCTCCACCTGGAGATCATTCTGGAGTGTTTTCGTATTCCTATCCGAGTCGGGAATGTAGGCGGCCGCCCGAAACGCCAGCTGCTGCTGCATCACAGCCAGCACTCCAAA  2520
GAGACTCAT
         CTGTGGGGTAGAAGATTGATCGATTATAATCCGTTTTCGGGTGTACTGCCATTCTCGTTTCGCCTTTCGGTTTCTTAC
                                                                                       AAGCAAATGCAT
                                                                                                   GACGGGAGTGGAGTAGAGATC  2640
TCTGGGTTCTATCGTCTTTCGAGTATCAAGTGGAACTGTTTGTCTGCTTCTCGTACGCTTTCGTTTTTATATCCATCCGCTTGTCTGTCAGCTTGGATTTTTTTTTCGCAATATTTTCGC  2760
CTCGGAAACCCAGTTCCGAAACATCACCGACATGGCTCGTCATCATCATCGTTGCGGCGGCAGCGGGGCCGCTTGCCTTAGTCAGTCATAAATAAAATGATTTTATTTCTTATAAGTATT  2880
ATTTTATATAGGTATCAGCGATGTTATTTGTCGCCGCTGCAATTCCCAAAGTGTGGGATGGTAAAAATACATACATATACAATATTCGTGAATGGTTATGCCCTACTCGAAGGTTGGGGC  3000
ACTTACAGTTTCATATATAGTATATTGTGGCTGATCTTGAAATACTTACATGCTCAAGGAACTCGAAAATACCATAGCGATCATCGACACTGCGGGTGTCCTTAATTGGCTTCCATTGTG  3120
GAATATCAGAGGCAAAAGAAAGTTTTGTCTGAGGAAAATAAGGTTGATTTTATAGATCACATATTGTAGGGATTACTGGGGATAACTTCCACTTTATGGCTTCCTACTATAGAATATTGC  3240
GAAGGGAATCTGTTGTACGATACTTCGAGACCAAGAACAATACTACTTTCATCTCAGACTTGTAATAAAAATTGTACAGGAAATTGCATTCCAATTCAAGAAAATCTCCACCAATCAACA  3360
GCTCAAAGACCATCTATGGAACACCCCCCAGACAACCCATAATCTTGCTCAATTAATTCAACCACTCTCTTTGTTTTTTATAGTTTTTGTCTGCTAGTTATTATTAGAAATTGCATAGCA  3480
TCATAAATGATGAATGAAGTGTTATCTCTAAGCATATATTTTGATCATTAGACACAAACTTTATCAAAATTACGCGTACCTCCGATTATTTGATACGTCTTCTTTTTGTTGTTGTTTTG  3600
CTGGCTAATTCCCAAAATAAAAACCCGTACCGGTTCGACTAATTCACCAAACCGTATCTAATTTTGATTCCCTCCCCAAACTCTTATCCGCAAACAAATGTATCAGCATTCCTTGCTTCA  3720
ATACCCCGATTCCCATTCGCTGGCTCCCCCACTTAATTTGCATATAATACAGTAAATCATGAGCGCACTCTGTAAAGCATAAAGACGAAAACGTTCTAGCCAAATTTGGCAAGTAGTCGA  3840
CGCTCGCTCAGGTCAGAGATTCCGAACGATATATATAAAGGCAAAAGTCTTTCAAGACTTCTCTTACTCGGTTTATACGAAGTTGATTTGATGAAAAGCTGTGAAAAGTGAATGAAATCA  3960
GGAATTAGGAAGGCGCAGCCAACTTCTGTCCTTACAGGAATCAAGGGGGAAAACCAAAAACAGCTGATCAACTGTACAAATCTCTTCTGAAAGCAAAAAAAAAAGAAAACATACTGCAGA  4080
AAAGGAATCAAATGGAAAGTAAGAGAGATCTCTGTCCTATTCGAAAGATATAGACGCGGCGAAGCTGCTAGAACTGGCAACTTGCGTTCATTAATATCATGAATTCAGCTTCGTATAAAC  4200
ATAGTATATCTCTATCTGTCATAAGCCTTCATTTACGAGTATGTAGATGGTTAGATGGTGCTTGCCTTTCGCCTCATTCTTCTTGATCGCGCCTAACCTTCGTTACGGAAATTGAAACC  4320
TTGACCAGCTTTGCTCTTAATTTGGAAAGCCCAAATAGCTCTTAAAAGACGTCATATTTAGCAGCAAAATTGGAGAGCATGCGGCCGTGTGAACGTTCTTGAACTCTTAGGAAATCGGGA  4440
TGATTCTTTGCCCAAGTGCTAAACGGTCTGGAACGCAGACGATTGCAATTTCAACTCCGTGTTTCGTCGACTCTGCTGTTGTGGGGGCAAACCAACCGAAATTCTGCGTTGGCGCCGATA  4560
CGAATTAGAACTGAGGTAACGAAGCACCAGAGAACCGGACCAGAACAGCCCCGGATTACTGTGGTTGGTTGACATGGCTCTGACATAGCTGGCGCCTTAGTCAGTCGCATCGTTCGTTGA  4680
ATTCGCTGTTGATCCACACGCTGGCTCACGTTTGCCCGATCGAAAGACAATGCACCCGTACCGCCGGAGAGACCCAGTTTAATGGCAAAACAATTAGGAAGGCATTTAGTATTTTGGCGA  4800
CAAATTGGCGAACGGAACCGGTTGAGTACTTAGAGGAGTGTTATCTTCCATTTGCGGAATGTGGTTACAAAGAGGAGAGAGAGAGAGAATCCCCCGCCCAAAGAAGTTGCAGTGCGATCATG  4920
ACCCGGTTAGATGTTAGATGGGGGAAATCCAATTATGATCTCTGTTGATTGGGGATTGTTTTTTGATGCAAATTGTATTCAATGCCAATTCTGAAATTCTTAGAGAAGAGGTATTGATGG  5040
GGAGAGGGCAAAAGCAGGGATATGAGAATATTTAGATGGGATTTGGAGTTAGGAAAGGCTAAAGATAACGATAGGATACAAATTATAAGTAAAATGGAAGAGAAGTTTGGAAAGAAACAG  5160
TTTTATTTCTCATAAATTGCATTATTTTCCATACAATTTCCACATAAAGAATGGTTGTCATATTATATTTTAAATTTAAAAAACGCGGAACCTTCCTCTTAAATCAAGTAGATCTTTTCG  5280
TTACAATTTTATATGAAATTTTAAAAAGAAAGATGCAATTTTTTCATCAGTGTATCGTTTTATAAAGAACAAAGCCGATGCCAAATAAATAACTCGGTTTTGCGTTGAATTCAAATTCAA  5400
ACCGTACGGCTCTCCGATTATAAACATCACACCAAATATTTGTATCTTAATTAATTTTAAATTGCATGATTACTTAGCTTATGGAGACCTGTGTCGTCGCGGATAATGAATAACCCGAAT  5520
ATTTCAAATGGCTTGGAAAAAATATATAGATCGAAAATTTGCATGATTAAAATATCTGTGTATTTTTCAGGCCAGTGCTATATAATATTCTTTCTATAAGAAATTTTTTATATTTTAATG  5640
AAATATTATAATATATTCTTTCGTCTCTTGAGTTGTATTAAAATGCTACAAAAATGGAATACAAATTTACTAAATTTTAGTAGGATTTCAAAAAGTTTGGTTATCTATAATGACCAACAT  5760
TTTTGGGTTATGAGGTTTTTAAAATATGAACTAAAGATCGAACAGACCTATGGATACACTCAATAACATAGTTTCAGCGGTTCCCCAGATCATCATTAAATCTTTTGAATACGCTCCGAG  5880
GAGAAGATTGGGGGGAATTCTGAGGTTTAAGGACGTTCCGAAGGCAGGAGGGAATCGCGATTCTTTAGAACGTTTCGAAGATGGGAAGGAAGCAAGGGGAAAAGGCAGAGATTCTATCTT  6000
TGGCACAAGGACCACTTTCCGAGTTGGAAATACGGAAATATTAATGTTCTTCATTAAATATTATAAGTATAGAATAGACAGAATCTTATTTCTAGGAACATAAATGCGATTCCCAAAAAC  6120
TCAACATATTTATGGATGGTAATAGATCAATAATCATAAATCTATTCTAATCTTCAAATATTTCTTGCCCTGTTTTTCCTTCAG
                                                                                    GCCATTGCCAAATGGAGCAAGAAAAATGAAATCTCT  6240
CTAGTGGTCGTTGGACCCGAAGATCCCTTGGCCTTGGGTCTGGGCGATGTGCTGCAAAAGGAGGGCATCCCATGCTTTGGACCCGGCAAGCAGGGGGCCCAAATCGAGGCGGACAAGAAG  6360
TGGGCCAAGGACTTTATGCTGCGCCATGGAATACCAACGGCTCGATATGAGAGCTTCACGGATACGAACAAGGCCAAGGCGTTCATCAGGAG
                                                                                            GTGAGAAACGAATCTCTTGGAATGGAAA  6480
TTAGATTCGAATAACTGATGCCCTCTCTTCCAG
                                 TGCACCCTATCAGGCTCTGGTGGTGAAGGCCGCTGGTTTGGCTGCAGGCAAGGGTGTTGTGGTGGCCGCCAATGTGGATGAAGCCTG  6600
CCAAGCGGTGGATGAGATATTGGGAGACCTGAAATACGGACAAGCTGGGGCCACCCTTGTCGTAGAGGAGCTTCTCGAGGGCGAGGAAATATCGGTTCTAGCCTTCACCGATGGCAAGAG  6720
TGTTAGGGCCATGTTGCCTGCCCAGGATCACAAGCGTTTGGGCAATGGAGATACAGGACCAAACACCGGAGGAATGGGCGCCTACTGTCCCTGTCCTTTGATCAGTCAGCCCGCGTTGGA  6840
GCTGGTCCAGCGAGCCGTGCTGGAGAGAGCTGTTCAGGGTCTGATCAAGGAGCGCATCACCTATCAGGGTGTTCTCTATGCGGGACTCATGCTGACACGCGATGGACCCCGCGTCCTCGA  6960
GTTCAACTGTCGCTTTGGCGATCCCGAAACGCAGGTCATTCTGCCGCTCCTCGAAACTGATCTTTTCGAGGTGATGCAGGCCTGCTGCAGCGGTCAGCTGGATAGGCTGCCCCTCCAGTG  7080
GCGCAGCGGCGTGAGTGCTGTGGGTGTGGTCCTCGCAAGTGCTGGCTATCCAGAGACCTCCACGAAGGGTTGCCTCATCACTG
                                                                                   GTAAGGTTTCCTTGTGCTCTTTCCATTAAATCAATAT  7200
TAACTTGTTTCATTTTCTAG
                    GACTACCTGATGTCAATTCGCCCACTCAATTGATTTTCCACAGCGGTCTTTCTGTAAACAAACAAAAGGAGGCCCTCACGAATGGCGGACGTGTGCTGAT  7320
AGCCATTGCCTTGGATGCCAGCCTCAAGGAGGCGGCTGCAAAGGCCACAAAACTTGCAGGAACTATCACTTTTGCCGGCACAGGAGCTCAGTATCGCACGGACATTGCCCAGAAAGCATT  7440
TAAAAT
      GTAATTCAAAAGACGTTTACTCAATGGGTATATAAAGCCAATGAAGATAGGGAGATTTGGGGTCCTAAAAGATGGAGATTGAAGCTATTGATTGATAGAACTCTTACTAAAGA  7560
TTATTATGCTATAGATATAATTATTGATCCTCGAAAGATGAAATTTTGCCATACTGTAAAGTATGGAAATGGTTCCATTTTGTATTCAAATCTTTATCAGAAGTCCTATAAAGGGTGGAA  7680
ATTGTCAGTTTTTCATGCACAAGCATTTCAGGCATCGTTCGGATCCTCAAAAATTCAAATTAATCTCCTCTTAGCATCTAAAAACTCTTATTACTGTGAGATTTTGTTTGATCTGTCACC  7800
TTTTCGAAGTCTTGTTCGGAATCGTAACTTATTTTAGGTCAAAATTATAATCTTTCCGGACAAAATGGAGATTCTCACTAGGTTTCCCCTAGTACAACCTTCATCTTTATGCCTCTAATA  7920
AAGTATTTTCTACACATATAACTAATATATTACTTTTCTTTTCCACATCAG
                                                   AGCCATTGCCACGGCGCCAGGCCTCAGCTACAAAGACAGCGGCGTGGACATTGATGCCGGCGATGCTTT  8040
GGTCCAACGCATCAAGCCCTTGTCGCGTGGCACTCAAAGACCTGGGGTCCTTGGCGGTTTGGGTGGCTTTGGTGGTCTCTTTCGGCTGAAGGATCTCAGCTACAAGGAGCCCGTAATTGC  8160
AGAGGCCACGCAGGGTGTGGGCGCCAAGATCCAGTTGGCCCTGCAGAATGAATTGTATGAGAATATTGGCTACGATCTGTTTGCCATGTCTGCCAATGATCTGCTGGAACTAGGCGCCGA  8280
GCCGGTAGCTTTTCTGGATTACATAGCCTGTGGCAAGCTGCATGTGCCACTGGCCGCTCAGCTCGTGAAGGGAATGGCTGATGGCTGTCGGGATGCCAAGTGCGCATTAGTAG
                                                                                                                 GTGCGTA  8400
TCTAATCGAAGGAGAGGCAGTCTTTGAGGAAGATAATTAATAACTATATAATATATCCCATCCAG
                                                                 GTGGCGAAACAGCGGAGATGCCCTCGCTGTATGCCCCCGGGCAGCACGACATGGC  8520
CGGCTACTGTGTGGGCATCGTAGAGCAGGCTCGTGTCCTGCCACGCTTTGATCTGTACGAACCGGAGGATCTGCTCGTGGCCTGCCGTCCTCGGGCCTCCACTGTGCTGGCTTCAACGA  8640
GATCCTCACCCAGTTGGCCGCTTCGAAAGTGAATCTCAAGGAGTGCTCTCCTGTCGGCGGGGGCAAGCATGGCCTTAGCCTGGCCCAGGTCCTGGGCACGCCCACACGCCTGTACGTGCA  8760
ACAGTTGCTGCCCCATCTGCAGGCTGGCAACCAGATCAAAGCTGTGGCCCATGTCACGCATGGCCTGCTGCACGATGTCCAGCGCTTGCTGCCCGAAGGGTTCGAGGTCACCCTCGACTT  8880
TGGTGCAGTGCCCGTCCCGGATGTCTTTGGCTGGCTGGCGGGACAGCTGCAGCTGAGTGCCCAGACTCTGCTGGAGCGACATAACTGCGGTATTGGCATGGTGCTGGTTCTCCCACAGAG  9000
CAGTCTTCTGTGGCGCACAGCCCTGCCGGGGGCCAAGGTTCTGGGCGTGTTGAATCGTCAAGCGAAAGCCTCCGGCGGTGCTCCGCGCGTCAAAGTGCGAAATTTTGTGGAGCAACTGCA  9120
GAAGTTGGCTGCACCCTTTGGGGATTGGGGAATCGCAGCTGCCCGAGGAAGTCAAGGATGTGCCCTCATCGGAGTGAAGGCAACGACACGTGAAGAATGCTTTGAGAATGCCGTGGG  9240
ACGTCGCCTGACCCGCGTGCCCAATCATTACGTGGATCCCATACTCATCTTGGGTACCGATGGTGTGGGCACCAAGCTGAAGATTGCCCAGCAGACCCATCGGAATGCCAGCGTGGGCAT  9360
CGATCTGGTGGCCATGTGCGTCAACGATATACTCTGCAATGGCGCCGAACCATTTAGCTTTTCCAGCTATTATGCCTGCGGCAAATGGCAGGCCGCCTTGGCTGCGGAAGTCAACGCTGG  9480
CGTACAGGAAGGCGCCAGCCAGGCCAATAGTAGTTTTGTGG
                                         GTATGTTTTGGGTGGATCATTCTTTTGCTGATTCACATACTAATGCCATTTCCTCTCGATCACATAG
                                                                                                            CCTCCCATAGTG  9600
CTGCCCTACCACTCCTGTATGAGCCGCAGGTCTATGATCTGGCTGGCTTTGCTTTGGGCATAGCCGAACGCTCGGCCATCCTGCCACGCCTGGATGAGATTCAGCCCGGAGATGTGCTCA  9720
TTGGCTTGCCCTCATCGGGTGTCCACAGCAACGGATTCAGCCTCGTTCATGCGGTGCTCAAGCGTGCCGGCTTGGGCCTCAACGATCGGGCGCCATTCAGCGAAAAAACACTGGGGAGG  9840
AGCTCCTAGTGCCCACCAAGATCTATGTGAAGGCATTGTCCGCGCTGCTCTCTCGACCCAATCATGGGATCAAGGCGCTGGCGCACATCACTGGCGGTGGCCTGAGCGAGAATATACCCC  9960
GAGTGCTGCGCAAGGAATTGGCCGTTCGATTGGACGCCAACAAATACCCGCTGCCCCCAGTATTTGCCTGGCTGGCAGCAGCAGGAAACATCAGCTCCACAGAGCTGCAGAGGACCTACA 10080
ACTGTGGCTTGGGTCTGGTGTTGGTGGTGGAGCCACGGAAGTGGATGGCGTGCTCCGGGAGCTCCGTTATCCACAGCGCGCCTCTGTTGTGGGCGAAGTAGTGGCCCGAAAGGATCCGA 10200
AGAAACCCCAGGTGGTGGTCCAGAACTTTGAGGCATCGCTGACCAGGACCCAGAGAATGCTATCTCAACCACGGAAACGCGTTGCAGTGCTCATTTCAGGAAAAGGCAGCAATCTGCAGG 10320
CCCTCATCGATGCCATCCGTGACTCGGCTCAGGGCGTGTACGCAGAGATTGTCCTCGTCATCAGCAACAAGGCGGGTGTCTTGGGTCTGGAGAGGGCCGCCAAGGCGGCATACCCTCGA 10440
TGGTCATTTCGCACAAGGACTTCCCCAGTCGCGAGGTCTACGATGTTGAACTAACGCGCCATCTCAAGACAGCTCGCGTGGAATTCATCTGCCTGGCGGGCTTCATGCGCATCCTCAGTG 10560
TTCCCTTTGTACGCGAATGGCGTGGCCGTCTCATCAACATTCATCCCTCGCTGCTGCCTAAATTTCCGGGGCTACATGTCCAAAAGCAGGCTCTGGAGGCGGGTGAAACAGAGTCCGGGT 10680
GCACGGTTCACTATGTGGACGAGGGCGTCGACACGGGAGCCATAATCGTTCAGGCCGCTGTTCCCATACTGCCCGGGGATGATGAGGAGACGTTGACCCAGCGCATCCATTATGCCGAAC 10800
ATTGGCTTTTCCGCGTGCACTGGCGCTGTTGGCCAGCGGTGCCCTGCGTCGGGTGTCAGAGGTAAAGAAGGAGGCGCCGAAGGATATCAAGGATAGCCAG
                                                                                                     TGAAGTGTGAAGACATTAA 10920
AAACTTTAGTTTTTGTAAGTTTTTGACGCCCTCCAAGGGATTATTTTTATAGTGACAAACAATTTCTCTTTTGGTTAAGACTGGATTCCATGTTTTTTATAGATAACAAACCTATTCAAA 11040
ATATTTAAACGTATTCTCTACAACTAATATCACGGGGGATGCTTTCATATTCGATATCAATCACAAAAGACGGGTAATCAGGTCGAAAATATTGCTTTTGAAAGTGTTTTCCGAATGCT 11160
CATCCTTCGCTTTGTATTGGACTGCAGGACGCTTCTGATTACGTTTTAAGTCCGGCTCCTTAAACACCATGGCGGCCAGGGTCCAGGCGATAGCGAAAGAACTGATTCTTGAGTATGGGC 11280
ACAATCAGATTGTACTTATCAATCATCTTGTTCAGCTCCTGGACCTCCTCCTCGTGGTCTTGGGTGAACCTTAGCCAGGCGGCCTGCTCTTCGGCATTCTGTAGCGGCCAAT        11392
```

FIGURE 3.—Sequence of *D. pseudoobscura Gart.* Protein-coding regions are separated from noncoding regions by blank lines. Nucleotides that are within highly homologous regions (see Figure 9 and Table 4) and are identical between the two species are underlined.

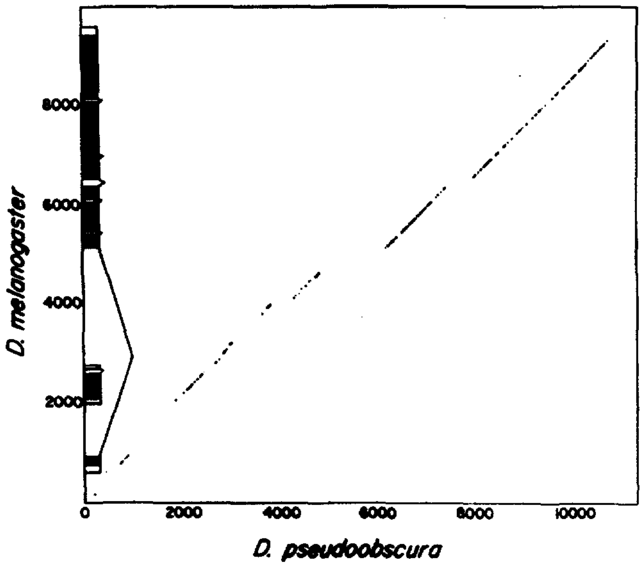

FIGURE 4.—Dot matrix comparison of *D. melanogaster* and *D. pseudoobscura Gart* loci. Each dot locates the position of 20 out of 25 nucleotides that are identical between the two species. The transcriptional and translational map of *D. melanogaster Gart* is indicated on the *y*-axis.

are sufficiently distant in time since divergence that mutation would have obliterated homologies at positions that are selectively neutral. As we have a total of more than 1000 aligned amino acids that are conserved at this locus, it should be possible to obtain a direct estimate of divergence at selectively neutral positions by examining third base codon differences.

Five amino acids are accounted for by four codons that differ only at the third position: threonine, proline, alanine, glycine and valine. For the aligned *Gart* locus proteins, these amino acids comprise 504 residues (Table 2). Of these, 284 (55%) differ at the third codon position. This compares with 75% differences expected for completely random codon choice. Because of a marked species bias in the choice of codons, however, a 75% level of difference is not expected to occur even for completely diverged sequences. Table 3 shows a codon usage comparison between the *D. pseudoobscura* and the *D. melanogaster Gart* locus polypeptides. The differences do not appear to be very significant between the two species. Overall codon usage conforms approximately to what has been seen for a large number of *D. melanogaster* genes (MARUYAMA *et al.* 1986). This allows us to superimpose the codon bias on our estimates of random codon choice, assuming only that the observed present-day bias has been in effect since the time of divergence. As Table 2 indicates, the expected number of differences for random choice of each of these five codons is 313 overall after this adjustment is made, not much higher than the observed value of 284. Therefore, about 90% of the neutral nucleotide positions in the protein coding regions are estimated to have mutated in one or the other lineage since their most recent common ancestor.

TABLE 1

Location of sequence landmarks

Description of site	Position in *D. pseudoobscura*	Position in *D. melanogaster*[a]
Purine gene transcriptional start	468	615
Purine gene translational start codon	754	775
Purine gene intron 1, 5′ side	931	952
Cuticle gene poly(A) addition	1831	1997
Cuticle gene 3′ AATAAA	1859	2025
Cuticle gene translational stop codon	1965	2116
Cuticle gene intron, 3′ side	2530	2673
Cuticle gene intron, 5′ side	2607	2743
Cuticle gene translational start codon	2619	2755
Cuticle gene transcriptional start	2681	2788
Cuticle gene TATA	2711	2818
Purine gene intron 1, 3′ side	6204	5093
Purine gene intron 2, 5′ side	6453	5342
Purine gene intron 2, 3′ side	6513	5395
Purine gene intron 3, 5′ side	7164	6046
Purine gene intron 3, 3′ side	7220	6101
Purine gene intron 4, 5′ side	7447	6328
GAR synthetase translational stop codon	7448	6329
Purine gene small transcript 3′ AATAAA	7917	6370
Purine gene intron 4, 3′ side	7971	6510
Purine gene intron 5, 5′ side	8394	6933
Purine gene intron 5, 3′ side	8465	6990
Purine gene intron 6, 5′ side	9522	8041
Purine gene intron 6, 3′ side	9588	8102
GARS-AIRS-GART translational stop codon	10902	9389
Purine gene 3′ TTTTTATA #1	10964	9532
Purine gene 3′ TTTTTATA #2	11013	9557
Purine gene 3′ AATAAA (AACAAA)	(11025)	9586

[a] Nucleotide #1 is defined as the first base of the nearest *Xba*I site on the 5′ side of the purine gene (HENIKOFF *et al.* 1986b).

Conservation of alternative processing: In *D. melanogaster*, alternative processing occurs with poly(A) addition after exon 4 to encode GAR synthetase and after exon 7 to encode GARS-AIRS-GART (Figure 1). Remarkably, however, there is no obvious resemblance between intron 4 of *D. melanogaster* and that of *D. pseudoobscura*. Figure 7 shows a dot matrix analysis in which each dot represents a match of 10 of 15 bp residues in order to detect relatively low levels of homology. Lack of homology for intron 4 is evident, since the density of dots in this section of the matrix is about the same as that for off-diagonal sections. Not only do we fail to detect homology between the two species, but also *D. pseudoobscura* intron 4 is about three times as large as that of *D. melanogaster*. Furthermore, the only obvious poly(A)-addition sequence, AATAAA, is located very near the

A.

```
MSHSVLVIGSGGREHAICWKLSQSTLVKQIYALPGSFGIQQVEKCRNLDAKVLDPKDFEAIAKWSKKNEISLVVVGPEDPLALGLGDVLQKEGIPCFGPGKQGAQIEADKKWAKDFMLRH   120
::: ::::::::::::::::::::  : :::::::: ::: :::::::::: :::::::::::::: : : ::::::::::::::::::::  ::::::::::::::::::::::::::::
MSHRVLVIGSGGREHAICWKLSQSPKVAQIYALPGSHGIQLVEKCRNLDAKTLDPKDFEAIAKWSKENQIALVVVGPEDPLALGLGDVLQSAGIPCFGPGKQGAQIEADKKWAKDFMLRH   120

GIPTARYESFTDTNKAKAFIRSAPYQALVVKAAGLAAGKGVVVAANVDEACQAVDEILGDLKYGQAGATLVVEELLEGEEISVLAFTDGKSVRAMLPAQDHKRLGNGDTGPNTGGMGAYC   240
::::::::::::: ::::::::::: ::::::::::::::::::::  ::::::::::::::::::::::::::::::::.:::::::::::::::::::::::::::::::::::::::
GIPTARYESFTDTEKAKAFIRSAPYPALVVKAAGLAAGKGVVVAANAKEACQAVDEILGDLKYGQAGATLVVEELLEGEEVSVLAFTDGKSVRAMLPAQDHKRLGNGDTGPNTGGMGAYC   240

PCPLISQPALELVQRAVLERAVQGLIKERITYQGVLYAGLMLTRDGPRVLEFNCRFGDPETQVILPLLETDLFEVMQACCSGQLDRLPLQWRSGVSAVGVVLASAGYPETSTKGCLITGL   360
:::::::::::::: ::::::::::::::: :::::::::::::::::::::::::::::::::::::: ::: :: ::::: ::  ::::: ::::::: :::::::::::::: : ::
PCPLISQPALELVQKAVLERAVQGLIKERINYQGVLYAGLMLTRDGPRVLEFNCRFGDPETQVILPLLESDLFDVMEACCSGKLDKIPLQWRNGVSAVGVILASAGYPETSTKGCIISGL   360

PDVNSPTQLIFHSGLSVNKQKEALTNGGRVLIAIALDASLKEAAAKATKLAGTITFAGTGAQYRTDIAQKAFKIAIATAPGLSYKDSGVDIDACDALVQRIKPLSRGTQRPGVLGGLGGF   420
:  : :::: ::::: :: :::::::::::::::::: :::::::::::::: : : : :::::::::::::::: :  :::::::::::::::::::::::::::::::::: ::::::
PAANTPTQLVFHSGLAVNAQKEALTNGGRVLIAIALDGSLKEAAAKATKLAGSISFSGSGAQYRTDIAQKAFKIASASTPGLSYKDSGVDIDACDALVQRIKPLSRGTQRPGVIGGLGGF   420

GGLFRLKDLSYKEPVIAEATQGVGAKIQLALQNELYENIGYDLFAMSANDLLELGAEPVAFLDYIACGKLHVPLAAQLVKGMADGCRDAKCALVGGETAEMPSLYAPGQHDMAGYCVGIV   600
::::::: : ::::::::::::::::: :::  : ::: ::::::  :::  ::: :::::::::::::::: :::::::::::::::::: ::::::::::::::::::::::::::::::
GGLFRLKELTYKEPVIAEATQGVGAKIHLALTHEFYENVGYDLFALAANDVLEVGAEPVAFLDYIACGKLQVPLAAQLVKGMADGCRDARCALVGGETAEMPSLYAPGQHDMAGYCVGIV   600

EQARVLPRFDLYEPEDLLVGLPSSGLHCAGFWEILTQLAASKVNLKECSPVGGGKHGLSLAQVLGTPTRLYVQQLLPHLQAGNQIKAVAHVTHGLLHDVQRLLPEGFEVTLDFGAVPVPD   720
:  : ::::::: : ::: :::::::::::::::::::::::::: : ::: ::  :: :: :: ::: ::::::::::: :  :: ::::::::: :  :::: ::: ::::::::::
EHSRILPRFDLYQPGDLLIGLPSSGLHCAGFWEILTQLAASKVNLRERSPVDGGDDGLTLAHVLATPTQLYVQQLLPHLQKGDEIKSVAHVTHGLLNDILRLLPDGFETTLDFGAVPVPK   720

VFGWLAGQLQLSAQTLLERHWCGIGMVLVLPQSSLLWRTALPGAKVLGVLNRQAKASGGAPRVKVRNFVEQLQKLAAPFGGLGESQLPEEVKDVPSSGVKATTREECFENAVGRRLTRVP   840
 :::::: : ::::: :::::::::::: ::::: :::: :::::::::: :  : ::    : :::::::: : : ::::::. ::::  :  : ::       :::::::: :::::: :
IFGWLAGKLKLSAQTILERHWCGIGMVLILPQSSQLWRTSLPGAKVLGVLQRRSKVSGSP--VQVRNFVEQLEKVASPFGGLGDRELPEELKKLPSNSDLSAPREECFENAAGRRLTRIP   838

NHYVDPILILGTDGVGTKLKIAQQTHRNASVGIDLVAMCVNDILCNGAEPFSFSSYYACGKWQAALAAEVNAGVQEGASQANSSFVASHSAALPLLYEPQVYDLAGFALGIAERSAILPR   960
 :: ::::::::::::::::::::: :: ::::::::::::::::::::: ::::::::: ::  ::  :  ::::::  ::::::  ::::::::::::::::::::::::::   :::
THYKDPILILGTDGVGTKLKIAQQTNRNTSVGIDLVAMCVNDILCNGAEPISFSSYYACGHWQEQLAKGVHSGVQEGARQANSSFIDSHSAALPLLYEPQVYDLAGFALGIAEHTGILPL   958

LDEIQPGDVLIGLPSSGVHSNGFSLVHAVLKRAGLGLNDRAPFSEKTLGEELLVPTKIYVKALSALLSRPNHGIKALAHITGGGLSENIPRVLRKELAVRLDANKYPLPPVFAWLAAAGN   1080
: :::::::::::::::::::::::::::::: :::: : :::: ::::::::::::::::::: ::::  :::::::::::::::::::::::: :::::::::  :::::::::::::
LAEIQPGDVLIGLPSSGVHSNGFSLVHAVLKRVGLGLHDKAPFSDKTLGEELLVPTKIYVKALSTLLSRGKHGIKALAHITGGGLSENIPRVLRKDLAVRLDANKFQLPPVFAWLAAAGN   1078

ISSTELQRTYNCGLGLVLVVGATEVDGVLRELRYPQRASVVGEVVARKDPKKPQVVVQNFEASLTRTQRMLSQPRKRVAVLISGKGSNLQALIDAIRDSAQGVYAEIVLVISNKAGVLGL   1200
::::::::::::::: ::::  :::  :: :::::::: ::::::::::::: :::::::::::.::: :::: :::::::::: :::::::::: :::::: :  ::::::: :::::
ISSTELQRTYNCGLGMVLVVAPTEVEDVLKELRYPQRAAVVGEVVARKDPKKSQVVVQNFEASLARTQKMLSQRRKRVAVLISGTGSNLQALIDATRDSAQGIHADVVLVISNKPGVLGL   1198

ERAAKAGIPSMVISHKDFPSREVYDVELTRHLKTARVEFICLAGFMRILSVPFVREWRGRLINIHPSLLPKFPGLHVQKQALEAGETESGCTVHYVDEGVDTGAIIVQAAVPILPGDDEE   1320
 :: :: ::: ::::::: :::::: :::: :: :::  :::::::: :: :::::::::: ::::::::  :::::::: :::::::::::::: ::::::: :::::::::::::::::::: :::
QRATQACIPSLVISHKDFASREVYDAELTRNLKAARVDLICLAGFMRVLSAPFVREWRGRLVNIHPSLLPKYPGLHVQKQALEAGEKESGCTVHFVDEGVDTGAIIVQAAVPILPDDDED   1318

TLTQRIHYAEHWAFPRALALLASGALRRVSEVKKEAPKDIKDSQ                                                                                1364
 :::::: ::::::::::: :  :      ::
SLTQRIHKAEHWAFPRALAMLVNGTALISPEVSSQ                                                                                         1353
```

B.

```
MHLLMSLFGVLAVMQQQLAVRAAAYIPDSDRNTKTLQNDLQVERDGNYRYAYETSNGISATQEGLGGVSVQGGSSYTSPEGSVISVSYVADETGYHPVGDHIPKVPDYILRALEYIRAHP   120
: ::      ::: :         ::::::::: :::::::::::: ::::::::::::: ::::::: :::::::::::: :::: ::::: :::::: ::: ::::::: ::::: ::
MYLLVMFIVALAVLQ---VQAGSSYIPDSDRNTRTLQNDLQVERDGKYRYAYETSNGISASQEGLGGVAVQGGSSYTSPEGEVISVNYVADEFGYHPVGAHIPQVPDYILRSLEYIRTHP   117

YQVKDYYTGELKTVAHDAAAFWVYTRWIQDQTTPRSRPSSTPKTIYLTHPPTLSDAPTRRPLRQRQNDSRRR                                                    192
:: ::::::::::: ::::::::::::::: : : :::: ::::::::::::       ::::::
YQIKDYYTGELKTVEHDAAAFWVYTRWIQDHTIPQSRPSTTPKTIYLTHPPTTTS----RPLRQRRALPTH                                                     184
```

FIGURE 5.—(A) Amino acid sequence alignments between the predicted *D. pseudoobscura* GARS-AIRS-GART (*top sequence*) and the *D. melanogaster* polypeptide. Identical residues are indicated with intervening colons. Alternative processing of the primary transcript leads to production of a GAR synthetase that is identical for residues 1–433 but has a methionine, rather than an isoleucine, at residue 434 in both species. (B) Similar alignments between the *Gart* cuticle proteins of the two species.

3′ splice junction of *D. pseudoobscura*, in contrast to the corresponding position of the functional *D. melanogaster* AATAAA which is closer to the 5′ splice site.

In spite of these striking differences, alternative processing does occur similarly in the two species. Figure 8 shows a Northern analysis of poly(A)$^+$ RNA from the two species run on the same gel and hybridized with appropriate species-specific probes. Although the 4.7-kb RNAs encoding GARS-AIRS-GART are about the same size in the two species, the transcript encoding GAR synthetase is substantially larger in *D. pseudoobscura* than in *D. melanogaster*, 2.3 kb compared to 1.7 kb. This difference can be accounted for primarily by the altered position of the AATAAA, 469 bp farther downstream from the 5′ splice site in *D. pseudoobscura*. The resulting polypeptides are predicted to be identical to one another for the two species at the residues affected by alternative processing: where the AATAAA is used, translation terminates following an AUG methionine codon, whereas where splicing occurs, the G of this codon forms the first base of the 5′ splice site, so that an isoleucine spanning the splice junction is encoded instead (HENIKOFF, SLOAN and KELLY 1983). In *D. melanogaster*, this isoleucine is encoded by AUU, whereas in *D. pseudoobscura* it is encoded by AUA. We conclude that GAR synthetase is encoded on the smaller transcript in *D. pseudoobscura* as it is in *D. melanogaster* (HENIKOFF *et al.* 1986b).

Highly conserved regions within intron 1 of the

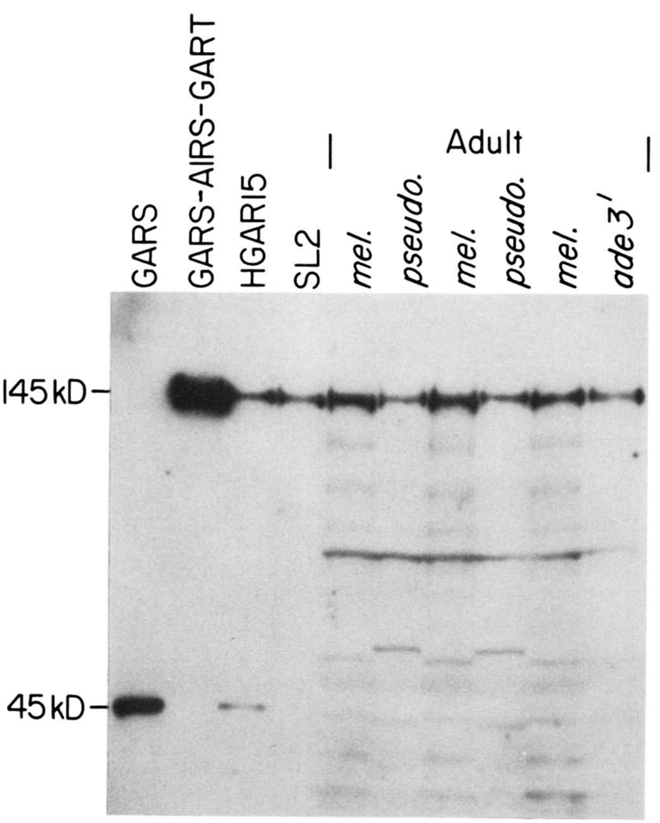

FIGURE 6.—Western blot analysis of Drosophila protein extracts probed with anti-GARS-AIRS-GART. GARS is HGAR15 tissue culture cell extract that has been depleted of the 145-kD GARS-AIRS-GART polypeptide by affinity chromatography on 10-formyl 5,8-deazafolate-Sepharose, and GARS-AIRS-GART is the resulting 145-kD protein, apparently homogeneously pure (HENIKOFF *et al.* 1986b). HGAR15 is a transformant of the *D. melanogaster* Schneider's line 2 (SL2) tissue culture cells that carries about 70 copies of the complete *D. melanogaster Gart* locus. Alternate lanes of *D. melanogaster* (*mel.*) and *D. pseudoobscura* (*pseudo.*) and the *ade*3[1] lane were each loaded with 1 mg protein extracted from adults.

TABLE 2

Third base differences for GARS-AIRS-GART and cuticle protein codons

Amino acid	Codon	Bias (%)[a]	No. matching residues	Expected No. differences[b]	Observed No. differences	Obs/Exp
Thr	ACU	9.6	51	27	32	1.19
	ACC	64.8				
	ACA	9.3				
	ACG	16.4				
Pro	CCU	10.1	76	50	47	0.94
	CCC	49.4				
	CCA	19.8				
	CCG	20.7				
Ala	GCU	19.1	132	81	74	0.92
	GCC	56.8				
	GCA	12.9				
	GCG	11.1				
Gly	GGU	22.1	138	89	75	0.84
	GGC	48.6				
	GGA	26.5				
	GGG	2.8				
Val	GUU	8.5	107	66	56	0.85
	GUC	33.3				
	GUA	6.8				
	GUG	51.3				
Totals			504	313	284	0.91

[a] Calculated from tabulations for *D. melanogaster* (MARUYAMA *et al.* 1986).

[b] Calculated from the 4 bias percentages using the formula: expected number of residues = No. matching residues × $\{1 - \Sigma[0.01\,(\text{percentage})^2]\}$.

TABLE 3

Codon usage of *D. pseudoobscura* and (*D. melanogaster*) *Gart* Loci[a]

	U	C	A	G	
U	phe 28 (19)	ser 8 (9)	tyr 23 (19)	cys 8 (4)	U
	phe 13 (22)	ser 14 (16)	tyr 22 (24)	cys 16 (19)	C
	leu 1 (6)	ser 4 (10)	och 1 (2)	uga 2 (1)	A
	leu 39 (42)	ser 22 (18)	amb 0 (0)	try 10 (10)	G
C	leu 10 (16)	pro 9 (12)	his 23 (18)	arg 16 (14)	U
	leu 43 (20)	pro 42 (34)	his 11 (22)	arg 31 (22)	C
	leu 10 (11)	pro 20 (14)	gln 15 (22)	arg 10 (14)	A
	leu 79 (83)	pro 14 (25)	gln 63 (51)	arg 13 (16)	G
A	ile 21 (24)	thr 9 (5)	asn 24 (21)	ser 16 (20)	U
	ile 38 (42)	thr 21 (30)	asn 20 (22)	ser 28 (24)	C
	ile 13 (15)	thr 18 (14)	lys 23 (16)	arg 5 (4)	A
	met 18 (16)	thr 22 (23)	lys 54 (62)	arg 7 (5)	G
G	val 20 (26)	ala 33 (41)	asp 54 (46)	gly 29 (32)	U
	val 36 (26)	ala 86 (69)	asp 17 (29)	gly 73 (61)	C
	val 11 (7)	ala 20 (25)	glu 31 (25)	gly 29 (42)	A
	val 67 (66)	ala 26 (21)	glu 55 (57)	gly 15 (9)	G

[a] Combined values for purine and cuticle genes.

purine gene: In addition to the protein-coding regions of the nested genes that are conserved between the two species, we detected extensive homologies in noncoding regions (Figure 4). All of these lay within the first intron of the *Gart* purine gene, except for a few homologies to the 5′ side. Figure 9 shows these noncoding homologies as open boxes in a linear alignment of the two sequences in this region. Regions were considered homologous if they matched precisely at greater than 10 residues or else at 20 of 25 residues. The length, degree of homology and position of each match is indicated in Table 4. Most of the homologous regions are within one of three clusters that lie upstream of the nested cuticle gene. Some of those within the cluster immediately upstream are surely involved in expression of this gene, as they include the start consensus sequence and the TATA box. The two other clusters lie approximately 1 kb and 2–3 kb upstream of the cuticle gene. In addition to these clusters, other homologies are detected at the 3′ side of the cuticle gene, including one of 82 bp that extends through both the poly(A) addition site and the AATAAA for this gene. Especially striking is the existence of an homologous region (20 of 21 identical nucleotides) within the single cuticle gene intron. This

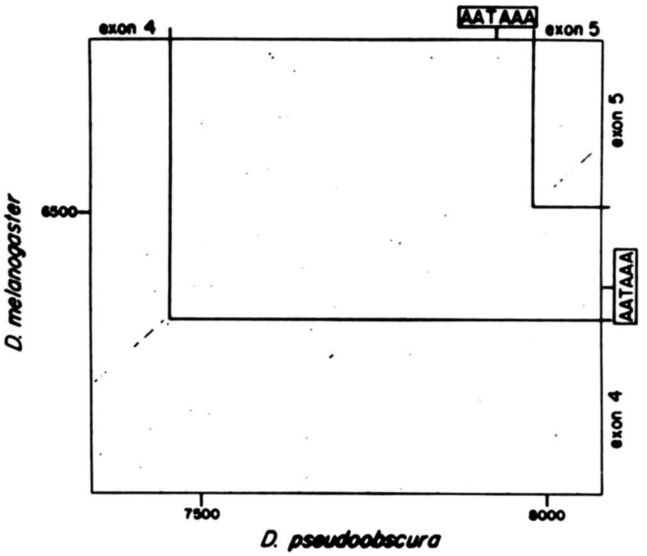

FIGURE 7.—Dot matrix comparison of the alternative processing region of *Gart* in the two species. The coordinates correspond to the landmarks listed in Table 1. Each dot locates the position of 10 out of 15 nucleotides that are identical between the two species. Except for protein-coding sequences, no detectable homologies are seen.

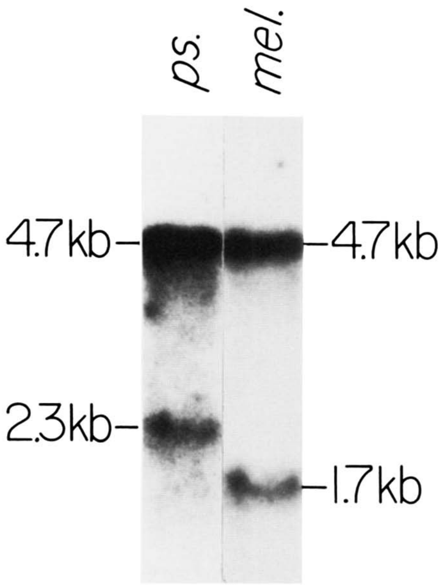

FIGURE 8.—Northern blot analysis of poly(A)$^+$ RNA using *Gart* purine gene probes. The lanes shown are from two halves of a single gel. The RNA was transferred to a filter which was then cut in half. Hybridization was carried out using homologous nick-translated probes. For the *D. pseudoobscura* lane, the probe corresponded to the 1069 bp *Apa*I-*Sac*I fragment spanning exons 2–4. For the *D. melanogaster* lane, the probe corresponded to an 1865-bp *Bss*HII-*Sma*I fragment spanning exons 1–4 but deleted for the entire intron 1 (data not shown).

particular sequence lies within introns of both genes on opposite strands.

These homologous segments range from 11 to 119 bp in length. Some are more highly homologous between species than are the protein-coding regions. For example, region 20 is 97% homologous with 83 of 86 identical nucleotides, whereas the conserved portion of the large cuticle gene exon is 80% homologous with 357 of 446 identical residues.

Overall, there are 24 homologous segments that average about 40 bp in length with greater than 90% average homology. All of these regions are in precisely the same order in the two species. However, the spacings of these homologies with respect to one another are quite variable. In some cases, nearly adjacent homologies in one species are separated in the other. For example, the distance between homologous segments 21 and 22 is 4 bp in *D. pseudoobscura* and 45 bp in *D. melanogaster*.

5′ and 3′ control regions of the purine gene: Outside of the first intron of the *Gart* purine pathway gene, there are only three regions of comparable homology, all of which lie upstream of the GARS-AIRS-GART coding region. One coincides with the experimentally determined transcriptional start site for the *D. melanogaster* gene (HENIKOFF *et al.* 1986a). Its counterpart in *D. pseudoobscura* is assumed to correspond to the transcription start in this species. Within this homologous stretch is the sequence TTCAGTT in *D. pseudoobscura*, a 6/7-bp match to the consensus cap site, ATCA(G/T)T(C/T), compiled for several other insect genes (HULTMARK, KLEMENZ and GEHRING 1986; PIRROTTA *et al.* 1987). Using this homology to locate the start of transcription in *D. pseudoobscura*, the expected length of 5′ untranslated leader would be 286 bp compared with 160 bp in *D. melanogaster*. This difference would be expected to contribute to the difference in size between the 1.7-kb *D. melanogaster* and the 2.3-kb *D. pseudoobscura* GAR synthetase transcripts (Figure 8). As was pointed out in the previous section, this *D. pseudoobscura* mRNA should be about 470 bp longer than *D. melanogaster* due to its more extensive 3′ untranslated region, so that this mRNA is expected to be about 600 bp longer overall, as observed. The other two regions of homology lie about 200 bp upstream of the transcriptional start site for *D. melanogaster Gart* and about 400 bp upstream of the corresponding *D. pseudoobscura* sequence. Although these sequences might be thought to be involved in *Gart* expression, the detection of another transcript in this region adjacent to the *D. melanogaster* gene (preliminary results) raises the possibility that these two homologous sequences are unrelated to *Gart*.

Also indicated in Figure 9 is a single region of alternating purines and pyrimidines found 40–81 bp upstream of the *D. melanogaster Gart* transcription start site (*hatched box*). It consists of 42 alternating residues with three exceptions. Such regions have been hypothesized to form Z-DNA and perhaps be involved in mediating processes such as gene expression and homologous recombination (NORDHEIM and RICH 1983; HAMADA *et al.* 1984; KMIEC and HOLLO-

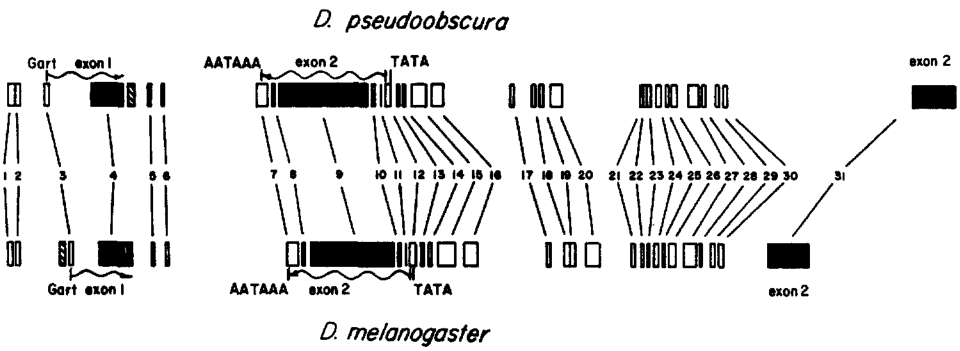

FIGURE 9.—Highly conserved regions within intron 1 of the purine gene. *Filled boxes* represent protein-coding regions. Transcription initiations are indicated by wavy lines. *Open boxes* are the homologous segments, nearly all of which appear in the dot matrix analysis shown in Figure 4. The precise locations of the individual homologies are shown in Table 4 and the sequences themselves are underlined in Figure 3. The *hatched box* in each map shows the location of the alternating purine-pyrimidine region.

TABLE 4

Highly conserved regions

No.[a]	Location	Position in		No. matches[c]/	No. insertions[d]	Percent
		D. pseudoobscura	*D. melanogaster*[b]	No. bases	(deletions)	conserved
1	5′ of purine gene	196–240	154–201	40/45	(1)	89
2	5′ of purine gene	241–282	216–252	32/41	2	78
3	purine gene 5′ end	451–486	597–633	26/31	0	84
4	*Gart* exon 1 coding	754–930	774–950	141/177	0	80
5	5′ side *Gart* intron 1	1100–1112	1112–1124	13/13	0	100
6	5′ side *Gart* intron 1	1180–1201	1183–1204	18/22	0	82
7	Cuticle gene 3′ end	1825–1906	1993–2070	72/82	3	88
8	Cuticle 3′ non-coding	1920–1931	2086–2097	11/12	0	92
9	Cuticle exon 2 coding[e]	1965–2529	2116–2672	357/446	0	80
10	Cuticle intron	2550–2570	2702–2722	20/21	0	95
11	Cuticle exon 1 coding	2608–2619	2744–2755	11/12	0	92
12	Cuticle 5′ end, TATA	2673–2721	2781–2828	45/48	0	94
13	5′ of cuticle gene	2757–2779	2853–2875	23/23	0	100
14	5′ of cuticle gene	2795–2809	2917–2931	14/15	0	93
15	5′ of cuticle gene	2835–2942	2971–3085	102/108	(2)	94
16	5′ of cuticle gene	2961–3042	3128–3212	73/82	1, (1)	89
17	Middle *Gart* intron 1	3457–3481	3621–3646	24/25	(1)	96
18	Middle *Gart* intron 1	3629–3663	3772–3806	31/35	0	89
19	Middle *Gart* intron 1	3683–3712	3812–3841	30/30	0	100
20	Middle *Gart* intron 1	3744–3829	3899–3984	83/86	0	97
21	3′ side *Gart* intron 1	4407–4424	4181–4198	18/18	0	100
22	3′ side *Gart* intron 1	4429–4443	4244–4258	15/15	0	100
23	3′ side *Gart* intron 1	4451–4478	4269–4296	26/28	0	93
24	3′ side *Gart* intron 1	4544–4582	4324–4361	34/39	1	87
25	3′ side *Gart* intron 1	4593–4603	4401–4411	11/11	0	100
26	3′ side *Gart* intron 1	4613–4665	4423–4473	46/53	(1)	87
27	3′ side *Gart* intron 1	4748–4866	4519–4632	108/119	2	91
28	3′ side *Gart* intron 1	4913–4932	4687–4706	16/18	0	89
29	3′ side *Gart* intron 1	4943–4962	4708–4728	19/20	(1)	95
30	3′ side *Gart* intron 1	4974–5019	4738–4783	33/46	0	84
31	*Gart* exon 2 (coding)	6205–6452	5094–5241	209/248	0	84

[a] Region numbers shown in Figure 9.
[b] Coordinates as in Table 1.
[c] Length of homologous region in *D. pseudoobscura*.
[d] Based on *D. pseudoobscura* coordinates.
[e] Only the conserved portion.

MAN 1986). A search of other Drosophila sequences for potential Z-DNA forming regions of similar length indicates that such regions are infrequent, occurring about once every 20 kb for regions that have been sequenced (data not shown). As no TATA box or other obvious promoter sequence (DYNAN and TJIAN 1985) is seen upstream of the purine gene, the possibility arises that this potential Z-DNA forming region is involved in transcription initiation. In *D. pseudoobscura*, a somewhat less extensive alternating purine/pyrimidine region also is detected. It is 28 bp in length with one exception. However, this region lies just to the 3′ side of *Gart* exon 1 (*hatched box*), rather than just upstream of transcription start, raising the possibility that these regions are not involved in gene expression. They might be involved in some other aspect of chromosome structure or function. A recent examination of regions of alternating purines and pyrimidines in Drosophila species by *in situ* hybridization indicates that such regions are conserved in their approximate distribution, consistent with their having some functional significance, perhaps in meiosis (PARDUE *et al.* 1987). In *Gart*, the imprecise correspondence of the two regions of alternating purines and pyrimidines in the two species might well reflect a function other than regulation of gene expression.

Examination of the 3′ region of *D. pseudoobscura Gart* shows no obvious poly(A) addition sequence, in contrast to the single AATAAA that is responsible for polyadenylation in *D. melanogaster*. Candidates include an ATTAAA which is found 12 bp after the stop codon and an AACAAA which is found 120 bp after the stop codon. There are two copies of the sequence TTTTTATA in this downstream region of the *D. pseudoobscura* gene (Table 1), apparently homologues of the two copies in the corresponding region of *D. melanogaster Gart* that were previously shown to function as transcription termination control sequences in yeast (HENIKOFF, KELLY and COHEN 1983; HENIKOFF and COHEN 1984). Subsequent work had shown similar sequences to be involved in transcription termination in vertebrate cells *in vivo* (SATO *et al.* 1986) and *in vitro* (BOHRMANN, YUEN and MOSS 1986). The very similar locations of the two TTTTTATA sequences in *D. melanogaster* and *D. pseudoobscura* further supports the suggestion that TTTTTATA functions in transcription termination of some higher eukaryotic genes as it does in yeast (HENIKOFF, KELLY and COHEN 1983; HENIKOFF and COHEN 1984).

Cuticle protein gene transcription: Previous Northern analysis had shown that the intronic gene was transcribed at high levels into a 0.9-kb poly(A)$^+$ RNA at the prepupal stage (HENIKOFF *et al.* 1986a). *In situ* hybridization further demonstrated that this mRNA was present specifically in the abdominal epidermal cells during the prepupal period. We had concluded that this gene is a pupal cuticle protein gene, since its temporal and spatial expression coincided precisely with the secretion of the abdominal pupal cuticle. We also reported that the 0.9-kb cuticle gene mRNA was present at lower levels in third instar larvae, and that cuticular preparations from this stage yielded the cDNAs that were used in the analysis. Since the RNA used in these experiments was isolated from mass cultures of larvae, and since third instar larval collections are generally contaminated with prepupae, we could not distinguish between the presence of the 0.9-kb mRNA in larvae and contamination by prepupae (HENIKOFF *et al.* 1986a). In the following analysis of RNA from late developmental stages, we used individually selected, rather than mass isolated larvae and prepupae in order to determine the expression of the cuticle gene. This allowed us to determine the developmental stage much more precisely, particularly since prepupae undergo rapid and obvious morphological changes. Both *D. pseudoobscura* and *D. melanogaster* individuals were collected in parallel so that possible species-specific differences could be detected. We separated feeding and wandering third instar larvae as well as early and later prepupae. Only the later prepupae secrete the pupal cuticle.

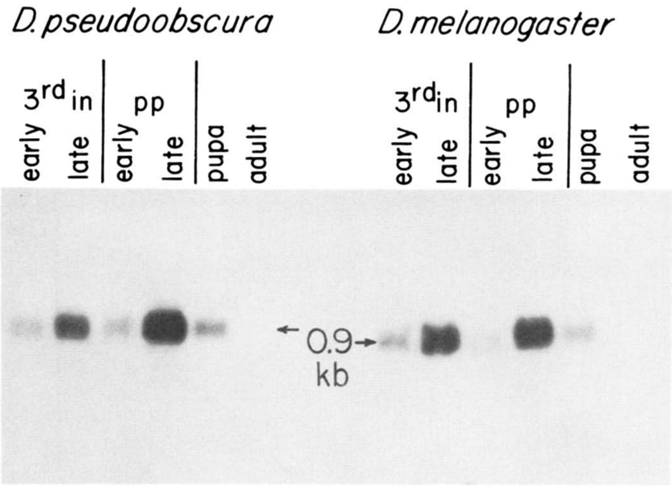

FIGURE 10.—Northern blot analysis of total RNA from late developmental stages. The lanes shown are from two halves of a single gel. The RNA was transferred to a filter which was then cut in half. Hybridization was carried out using homologous nick-translated probes. For the *D. pseudoobscura* lanes, the probe corresponded to the 1184-bp *Nsi*I fragment that includes the entire coding region of the cuticle gene and sequences downstream of it. For the *D. melanogaster* lanes, the probe corresponded to the 586-bp *Pvu*II-*Sac*I fragment containing most of the cuticle gene exon 2 (HENIKOFF *et al.* 1986a).

Figure 10 shows the resulting Northern analysis of total RNA from third instar larvae, pupae and adults, where the probe for each species is specific for the cuticle gene region. Both *D. pseudoobscura* and *D. melanogaster* have an RNA of about 0.9 kb in both wandering third instar larvae and late prepupae, with the *D. pseudoobscura* gene slightly larger. This difference in size is consistent with the mRNA in *D. pseudoobscura* being 55 bp longer, as predicted from the nucleotide sequence, assuming similar poly(A) tails. The presence of the cuticle gene mRNA during these two different developmental stages appears to be identical in the two species. As there are almost no detectable transcripts in the intervening early prepupal stage, we can conclude that mRNA accumulates during the third larval instar, is degraded prior to or during the prepupal stage and accumulates again at the time that the pupal cuticle is being secreted. A relatively low level of mRNA during the pupal stage, and absence of detectable transcript in adults is consistent with the degradation of this RNA when epidermal cells are histolyzed during metamorphosis.

DISCUSSION

The nested structure is conserved: Our molecular analysis of the *D. pseudoobscura Gart* locus has shown that this species has the same overall organization observed previously for *D. melanogaster Gart* (HENIKOFF *et al.* 1986a). Therefore, a particularly novel feature of this locus, the existence of a cuticle gene nested within the first intron, has been maintained in two lineages for at least 40–50 million years. Although considerable nucleotide evolution has occurred since the latest common ancestor, the gene is functional and shows similar developmental expression in both species. A high level of amino acid sequence homology further compels us to conclude that the cuticle gene has a similar function in the two species. This homology also confirms our previous assertion that the 0.9-kb poly(A)$^+$ RNA must encode a protein. Inability to identify this particular protein is likely due to the presence of several more abundant cuticle proteins of about the same size (CHIHARA, SILVERT and FRISTROM 1982; DOCTOR, FRISTROM and FRISTROM 1985).

Nucleotide sequence evolution at the locus: Alignment of homologous proteins encoded at the locus allowed us to estimate the extent of divergence at translationally silent nucleotide positions. When adjusted for codon bias, it was clear that the large majority of neutral positions have undergone changes. This confirms that the genes have been evolving separately for a length of time sufficient to generate sequence divergence. Thus, the homologies we detect must be significant.

It is somewhat surprising that no homologies were detected in intron 4, where alternative processing takes place in both species. Not only is the AATAAA in a different position with respect to the splice junctions, but also the *D. pseudoobscura* intron is three times the size of the *D. melanogaster* intron. This is the most sizable discrepancy thus far between introns in homologous positions of these two species for this gene and for others (BLACKMAN and MESELSON 1986; D. H. JOHNSON, personal communication). At least two explanations can be proposed for the striking dissimilarities between the functionally homologous introns where alternative processing occurs. The local sequence context or the relative efficiency of the splice sites and/or the polyadenylation site might lead to a balance of resulting mRNA products, so that no other specific sequences are necessary. Alternatively, such sequences might be necessary, but they are located outside of this intron, perhaps in neighboring exons where amino acid sequence conservation would hide any such homologies between the two species. This latter possibility is consistent with the demonstrated role of exonic sequences in splice-site selection *in vitro* (REED and MANIATIS 1986).

It is possible that there are species-specific differences in the relative amounts of the two purine gene mRNAs. This possibility has not been rigorously tested because of the coincidental electrophoretic migration of ribosomal RNA with the *D. pseudoobscura* 2.3-kb mRNA, but not with the corresponding *D. melanogaster* 1.7-kb mRNA. Because we have not been able to completely eliminate the ribosomal RNA from our poly(A)$^+$ preparations, and because this large amount of RNA in one position on a Northern blot appears to reduce the accessibility of the relatively rare 2.3-kb RNA, it is difficult to determine possible quantitative differences.

Conserved noncoding sequences: In contrast to the lack of sequence homology in the intron where alternative processing occurs, the intron containing the cuticle gene is particularly rich in highly homologous segments. Not counting the cuticle gene coding regions, there are 24 of these segments averaging about 40 bp in length. Their degree of homology exceeds that of the coding regions for both genes. These conserved regions are not connected by any consistent long open reading frame. Therefore, these segments are unlikely to be exons of yet another intronic protein-coding gene. The possibility that these segments are transcribed into noncoding conserved RNAs, such as has been observed for the Drosophila 93D heat shock puff locus (GARBE *et al.* 1986), seems unlikely as no other RNAs have been detected using intronic probes (M. EGHTEDARZADEH, unpublished results).

In addition to the 93D example, there have been other reports of extensive highly conserved noncoding intronic regions in Drosophila. A comparison between the *D. melanogaster* and *D. virilis engrailed* genes shows several short conserved intronic regions that resemble the ones reported here in that they do not appear to correspond to protein-coding regions (KASSIS *et al.* 1986). These authors suggest that the conserved segments represent *cis*-acting regulatory sequences and note that there are binding sites for the *engrailed* protein product similarly located in the first intron in both species. The first intron of the *Gld* locus also contains sequences that are homologous between distantly related Drosophila species (*D. melanogaster* and *D. pseudoobscura*) and show a similar pattern to the homologies reported here (KRASNEY *et al.* 1987).

The conserved intronic segments that we have found might represent *cis*-acting control sequences involved in regulation of one or both of the nested genes. Clearly some of these sequences are cuticle gene transcription control elements, as strong homologies are found around the cuticle gene TATA box and its polyadenylation site. Many of the others might correspond to tissue-specific and quantitative control elements found for other Drosophila genes (BOUROUIS and RICHARDS 1985; LEVIS, HAZELRIGG and RUBIN 1985; PIRROTTA, STELLER and BOZZETTI 1985; HIROMI, KUROIWA and GEHRING 1985; FISCHER and MANIATIS 1986).

It is possible that some of these conserved sequences within the purine gene intron are involved in purine gene regulation. We are unaware of Drosophila control elements known to lie within introns, although mapping of the *rosy* control element places it within a region that overlaps the first intron (LEE *et al.* 1987). In vertebrates, a few transcriptional enhancers are known to lie within introns (BANERJI, OLSEN and SCHAFFNER 1983; GILLIES *et al.* 1983; SLATER *et al.* 1985). One of these, the immunoglobulin heavy chain gene enhancer, is able to activate a transcriptional promoter that is 17.5 kb upstream (WANG and CALAME 1985). As all of the conserved regions that lie within the *Gart* purine gene first intron are no more than 5 kb downstream of the start of transcription, they are candidates for control elements of this gene.

The intronic cuticle gene is expressed at two different times during development: The use of individually staged Drosophila clearly demonstrates that late third instar larvae accumulate cuticle mRNA, which is degraded prior to reaccumulation in prepupae. This pattern of accumulation and degradation is found for both species. High level expression confined to the late larval period is characteristic of the small larval cuticle protein genes (SNYDER, HIRSH and DAVIDSON 1981; SNYDER *et al.* 1982). Similarly, high level prepupal expression is characteristic of the small pupal cuticle protein genes (CHIHARA, SILVERT and FRISTROM 1982). At least some of these low molecular weight larval and pupal cuticle proteins are known to be homologous to one another and to the *Gart* cuticle gene (HENIKOFF *et al.* 1986a; R. T. APPLE and J. W. FRISTROM, personal communication). It would thus appear that the protein product of the *Gart* intronic gene has a function that is appropriate for both larval and pupal cuticles, making it the most diverse known member of the gene family.

Possible significance of the nested arrangement: It has been suggested that the nested gene arrangement might have constrained developmental expression because of transcriptional interference (O'HARE, 1986). Nonetheless, the gene organization and complex pattern of expression have been conserved for at least 40–50 million years. We thus exclude the possibility that the nested arrangement is detrimental to function. However, this organization might have persisted simply because separation of the genes would require an exceedingly rare event. Accordingly, the gene arrangement might be of no consequence to the organism.

Another possibility is that the genes interact in a manner that is of some benefit. This raises the question of expression of the *Gart* purine gene at the times that the cuticle gene is active. We had previously presented evidence that low levels of cuticle gene mRNA are present in tissue culture cells that normally express the purine gene, suggesting that transcription can occur in a single cell on both strands (HENIKOFF *et al.* 1986a). Attempts at tissue localization of purine gene mRNA were unsuccessful because of low level transcription. Nevertheless, Northern blot analysis showed that purine gene transcripts were present at all times during development. The highest levels of purine gene mRNA accumulation were clearly seen during the third larval instar and prepupal periods (see Figure 2b of HENIKOFF *et al.* 1986a). These are precisely the two stages that we now know are times of intronic gene expression. We speculate that expression of the two genes might be coupled at these times. Perhaps some of the conserved segments are bidirectionally active enhancers shared by the two genes.

In conclusion, our comparison of the *Gart* locus between *D. melanogaster* and *D. pseudoobscura* shows that the organization of the nested genes is highly conserved. In each organism, the intronic gene is expressed as both a larval and a pupal cuticle protein gene and is surrounded and inhabited by conserved noncoding regions. These regions might be *cis*-acting regulatory regions such as transcriptional enhancers. Among the various possible explanations for the existence of nested genes, we argue that tolerance of a suboptimal arrangement is unlikely, given that the same complex expression has persisted over long evolutionary periods. Rather, we favor the possibility that the arrangement has some present-day function, perhaps by providing shared regulatory elements that enhance both cuticle gene and purine gene expression at larval and prepupal stages. This hypothesis can be tested by separation of the two genes from one another and from individual conserved sequences, followed by reintroduction into flies by transformation.

This work was supported by a grant from the National Institutes of Health (GM29009).

LITERATURE CITED

ADELMAN, J. P., C. T. BOND, J. DOUGLASS and E. HERBERT, 1987 Two mammalian genes transcribed from opposite strands of the same DNA locus. Science **235:** 1514–1517.

AUFFRAY, C., and F. ROUGEON, 1980 Purification of mouse immunoglobulin heavy-chain messenger RNAs from total myeloma tumor RNA. Eur. J. Biochem. **107:** 303–314.

BANERJI, J., L. OLSEN 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.

BASSIRI, R., J. DVORAK and R. D. UTIGER, 1979 Thyrotropin-releasing hormone. pp. 1–284. In: *Methods in Hormone Radioim-*

munoassay, Edited by B. M. Jaffe and G. R. Behrman. Academic Press, New York.

Beverley, S. M., and A. C. Wilson, 1984 Molecular evolution in *Drosophila* and the higher diptera II. A time scale for fly evolution. J. Mol. Evol. **21:** 1–13.

Blackman, R. K., and M. Meselson, 1986 Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila hsp82* gene. J. Mol. Biol. **188:** 499–515.

Bourouis, M., and G. Richards, 1985 Remote regulatory sequences of the Drosophila glue gene *sgs3* as revealed by P-element transformation. Cell **40:** 349–357.

Britten, R. J., 1986 Rates of DNA sequence evolution differ between taxonomic groups. Science **231:** 1393–1398.

Broker, T. R., L. T. Chow, A. R. Dunn, R. E. Gelinas, J. A. Hassell, D. F. Klessig, J. B. Lewis, R. J. Roberts and B. S. Zain, 1978 Adenovirus-2 messengers—an example of baroque molecular architecture. Cold Spring Harbor Symp. Quant. Biol. **42:** 531–553.

Chihara, C. J., D. J. Silvert and J. W. Fristrom, 1982 The cuticle proteins of *Drosophila melanogaster.* Dev. Biol. **89:** 379–388.

Dente, L., G. Cesareni and R. Cortese, 1983 pEMBL: a new family of single stranded plasmids. Nucleic Acids Res. **11:** 1645–1655.

Doctor, J., D. Fristrom and J. W. Fristrom, 1985 The pupal cuticle of *Drosophila*: biphasic synthesis of pupal cuticle proteins in vivo and in vitro in response to 20-hydroxyecdysone. J. Cell Biol. **101:** 189–200.

Dynan, W. S., and R. Tjian, 1985 Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature **316:** 774–778.

Fischer, J. A., and T. Maniatis, 1986 Regulatory elements involved in *Drosophila Adh* gene expression are conserved in divergent species and separate elements mediate expression in different tissues. EMBO J. **5:** 1275–1289.

Garbe, J. C., W. G. Bendena, M. Alfano and M. L. Pardue, 1986 A *Drosophila* heat shock locus with a rapidly diverging sequence but a conserved structure. J. Biol. Chem. **261:** 16889–16894.

Gentry, L. E., L. R. Rohrschneider, F. E. Casnellie and E. G. Krebs, 1983 Antibodies to a defined region of pp60^{sarc} neutralize the tyrosine specific kinase activity. J. Biol. Chem. **258:** 11219–11228.

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

Hamada, H., M. Seidman, B. H. Howard and C. M. Gorman, 1984 Enhanced gene expression by the poly(dT-dG) poly(dC-dA) sequence. Mol. Cell. Biol. **4:** 2622–2630.

Henikoff, S., 1986 The *Saccharomyces cerevisiae ADE5,7* protein is homologous to overlapping *Drosophila melanogaster Gart* polypeptides. J. Mol. Biol. **190:** 519–528.

Henikoff, S., 1987 Unidirectional digestion with Exonuclease III in DNA sequence analysis. Methods Enzymol. **155:** 156–165.

Henikoff, S., and E. H. Cohen, 1984 Sequences responsible for transcription termination on a gene segment in *Saccaromyces cerevisiae.* Mol. Cell. Biol. **4:** 1515–1520.

Henikoff, S., and M. Meselson, 1977 Transcription at two heat shock loci in Drosophila. Cell **12:** 441–451.

Henikoff, S., J. D. Kelly and E. H. Cohen, 1983 Transcription terminates in yeast distal to a control sequence. Cell **33:** 607–614.

Henikoff, S., J. S. Sloan and J. D. Kelly, 1983 A Drosophila metabolic gene transcript is alternatively processed. Cell **34:** 405–414.

Henikoff, S., K. Tatchell, B. D. Hall and K. A. Nasmyth, 1981 Isolation of a gene from *Drosophila* by complementation in yeast. Nature **289:** 33–37.

Henikoff, S., M. A. Keene, K. Fechtel and J. W. Fristrom, 1986a Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. Cell **44:** 33–42.

Henikoff, S., M. A. Keene, J. S. Sloan, J. Bleskan, R. Hards and D. Patterson, 1986b Multiple purine pathway enzyme activities are encoded at a single genetic locus in *Drosophila.* Proc. Natl. Acad. Sci. USA **83:** 720–724.

Henikoff, S., D. Nash, R. Hards, J. Bleskan, J. F. Woolford, F. Naguib and D. Patterson, 1986c Two *Drosophila melanogaster* mutations block successive steps of *de novo* purine synthesis. Proc. Natl. Acad. Sci. USA **83:** 3919–3923.

Hiromi, Y., A. Kuroiwa and W. J. Gehring, 1985 Control elements of the Drosophila segmentation gene *fushi tarazu.* Cell **43:** 603–613.

Hultmark, D., R. Klemenz and W. J. Gehring, 1986 Translational and transcriptional control elements in the untranslated leader of the heat-shock gene hsp22. Cell **44:** 429–438.

Johnson, D. A., J. W. Gautsch, J. R. Sportsman and J. H. Elder, 1984 Improved technique utilizing nonfat dry milk for analysis of proteins and nucleic acids transferred to nitrocellulose. Gene Anal. Techn. **1:** 3–8.

Kassis, J. A., S. J. Poole, D. K. Wright and P. H. O'Farrell, 1986 Sequence conservation in the protein coding and intron regions of the *engrailed* transcription unit. EMBO J. **5:** 3583–3589.

Kmiec, E. B., and W. K. Holloman, 1986 Homologous pairing of DNA molecules by Ustilago recl protein is promoted by sequences of Z-DNA. Cell **44:** 545–554.

Krasney, P. A., D. L. Cox, R. W. Whetten and D. R. Cavener, 1987 Nucleotide sequence comparison between the glucose dehydrogenase (*Gld*) genes of *Drosophila melanogaster* and *D. pseudoobscura.* Genetics **116:** s21.

Lee, S. S., D. Curtis, M. McCarron, C. Loce, M. Gray, W. Bender and A. Chovnick, 1987 Mutations affecting expression of the *rosy* locus in *Drosophila melanogaster.* Genetics **116:** 55–66.

Leff, S. E., M. G. Rosenfeld and R. M. Evans, 1986 Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annu. Rev. Biochem. **55:** 1091–1117.

Levis, R., T. Hazelrigg and G. M. Rubin, 1985 Separable *cis*-acting control elements for expression of the *white* gene of *Drosophila.* EMBO J. **4:** 3489–3499.

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

Maruyama, T., T. Gojobori, S. Aota and T. Ikemura, 1986 Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. **14:** r151–r197.

Nordheim, A., and A. Rich, 1983 The sequence $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ forms left-handed Z-DNA in negatively supercoiled plasmids. Proc. Natl. Acad. Sci. USA **80:** 1821–1825.

O'Hare, K., 1986 Genes within genes. Trends Genet. **2:** 33.

Pardue, M. L., K. Lowenhaupt, A. Rich and A. Nordheim, 1987 $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. EMBO J. **6:** 1781–1789.

Patterson, J. T., and W. S. Stone, 1952 *Evolution in the Genus Drosophila.* Macmillan, New York.

Pirrotta, V., H. Steller and M. P. Bozzetti, 1985 Multiple upstream regulatory elements control the expression of the *Drosophila white* gene. EMBO J. **4:** 3501–3508.

Pirrotta, V., E. Manet, E. Hardon, S. E. Bickel and M. Benson, 1987 Structure and sequence of the *Drosophila zeste* gene. EMBO J. **6:** 791–799.

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

ROHRMANN, G., L. YUEN and B. MOSS, 1986 Transcription of Vaccinia virus early genes by enzymes isolated from Vaccinia virions terminates downstream of a regulatory sequence. Cell **46:** 1029–1035.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SATO, K., R. ITO, K. H. BAEK and K. AGARWAL, 1986 A specific DNA sequence controls termination of transcription in the gastrin gene. Mol. Cell. Biol. **6:** 1032–1043.

SIMON, J. A., C. A. SUTTON, R. B. LOBEL, R. L. GLASER and J. T. LIS, 1985 Determinants of heat shock-induced chromosome puffing. Cell **40:** 805–817.

SLATER, E. P., O. RABENAU, M. KARIN, J. D. BAXTER and M. BEATO, 1985 Glucocorticoid receptor binding and activation of a heterologous promoter by dexamethasone by the first intron of the human growth hormone gene. Mol. Cell. Biol. **5:** 2984–2992.

SNYDER, M., J. HIRSH and N. DAVIDSON, 1981 The cuticle genes of Drosophila: a developmentally regulated gene cluster. Cell **25:** 165–177.

SNYDER, M., M. HUNKAPILLER, D. YUEN, D. SILVERT, J. FRISTROM and N. DAVIDSON, 1982 Cuticle protein genes of Drosophila: structure, organization, and evolution of four clustered genes. Cell **29:** 1027–1040.

SPENCER, C. A., R. D. GIETZ and R. B. HODGETTS, 1986 Overlapping transcription units in the dopa decarboxylase region of *Drosophila.* Nature **322:** 279–281.

STOCKER, A. J., and C. D. KASTRITSIS, 1972 Developmental studies in *Drosophila* III. The puffing patterns of the salivary gland chromosomes of *D. pseudoobscura.* Chromosoma **37:** 139–176.

SUDHOF, T. C., J. L. GOLDSTEIN, M. S. BROWN and D. W. RUSSELL, 1984 The LDL receptor gene: a mosaic of exons shared with different proteins. Science **228:** 815–822.

THOMAS, P., 1983 Hybridization of denatured RNA transferred or dotted to nitrocellulose paper. Methods Enzymol. **100:** 255–266.

TOWBIN, H., T. STAEHELIN and J. GORDON, 1979 Electrophoretic transfer of protein from polyacrylamide gels to nitrocellulose sheets: procedures and some applications. Proc. Natl. Acad. Sci. USA **76:** 4350–4354.

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.

WILLIAMS, T., and M. FRIED, 1986 A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. Nature **322:** 275–279.

Communicating editor: A. SPRADLING