Nucleotide Variation and Conservation at the *dpp* Locus, a Gene Controlling Early Development in Drosophila

Brent Richter, Manyuan Long, R. C. Lewontin and Eiji Nitasaka[1]

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received July 31, 1996
Accepted for publication October 25, 1996

ABSTRACT

A study of polymorphism and species divergence of the *dpp* gene of Drosophila has been made. Eighteen lines from a population of *D. melanogaster* were sequenced for 5200 bp of the Hin region of the gene, coding for the dpp polypeptide. A comparison was made with sequence from *D. simulans*. Ninety-six silent polymorphisms and three amino acid replacement polymorphisms were found. The overall silent polymorphism (0.0247) is low, but haplotype diversity (0.0066 for effectively silent sites and 0.0054 for all sites) is in the range found for enzyme loci. Amino acid variation is absent in the N-terminal signal peptide, the C-terminal TGF-β peptide and in the N-terminal half of the pro-protein region. At the nucleotide level there is strong conservation in the middle half of the large intron and in the 3′ untranslated sequence of the last exon. The 3′ untranslated conservation, which is perfect for 110 bp among all the divergent species, is unexplained. There is strong positive linkage disequilibrium among polymorphic sites, with stretches of apparent gene conversion among originally divergent sequences. The population apparently is a migration mixture of divergent clades.

THE determination of the nucleotide sequences of genes has been in the context of two quite divergent programs in biology. In population genetics the emphasis has been on the variation in nucleotide sequence within and between populations and species, as an extension of the general program of the study of variation in evolutionary genetics. Since the first population sequencing study in 1983 by KREITMAN, a number of gene loci coding for enzymatic amd structural polypeptides in Drosophila have been sequenced in population samples (see KREITMAN and WAYNE 1994 for a review). These studies all show that, with the exception of regions of virtually no recombination that have gone through a recent selective sweep-out of variation, there is considerable nucleotide polymorphism in flanking regions, introns, and "silent" changes in exons. Depending upon the gene and species, nucleotide polymorphism is between 3 and 9%. Amino acid polymorphism, on the other hand, may be completely absent as in *Adh* in *Drosophila pseudoobscura* (SCHAEFFER and MILLER 1993) or may be as high as 2% of codons as in *Est-5* of the same species (VEUILLE and KING 1995). These differences in amount of amino acid polymorphism, reflecting the same variability among loci as was originally seen in electrophoretic surveys, is presumably a consequence of very different physiological constraints on different enzymes. There is, in addition, evidence from haplotypic distributions in some cases for balanced polymorphism (KREITMAN and HUDSON 1991) or for other unspecified deviations from a model of simple neutral mutational variation (HUDSON *et al.* 1987) and for selective amino acid replacements between species (EANES *et al.* 1993). In general, population geneticists have concerned themselves with the explanation for patterns of standing *variation* within species, and of the *divergence* between species, while deemphasizing the importance of different levels of *constraint* on DNA and protein sequences.

In molecular developmental genetics the emphasis of sequence studies has been exactly the opposite of that in evolutionary genetics. Developmental geneticists have been concerned with finding constant motifs influencing development that transcend species and even phylum boundaries. The intensity of interest in *Hox* genes derives precisely from their apparent similarity or even identity over many organisms. The notion of sequencing "the genome" of Drosophila carries with it the implication that what is interesting about the genome can be learned from a single copy, so that if there is polymorphism one need not be concerned with it. Developmental geneticists are well aware of the phenotypic variation between individuals and species, but an explanation of that variation is not in their problematic. Variation is a source of disturbance and experimental noise, rather than the thing of central interest. There is, nevertheless, experimental information showing that nucleotide polymorphism at *Hox* genes is somehow relevant to phenotypic variation. So GIBSON and HOGNESS (1996) have shown that selection that increases the frequency of bithorax phenocopies in response to ether treatment, and that eventually produces a line that shows the bithorax

Corresponding author: R. C. Lewontin, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138.
E-mail: dick@mcz.harvard.edu

[1] *Present address:* Department of Biology, Kyushu University, Fukuoka, Japan.

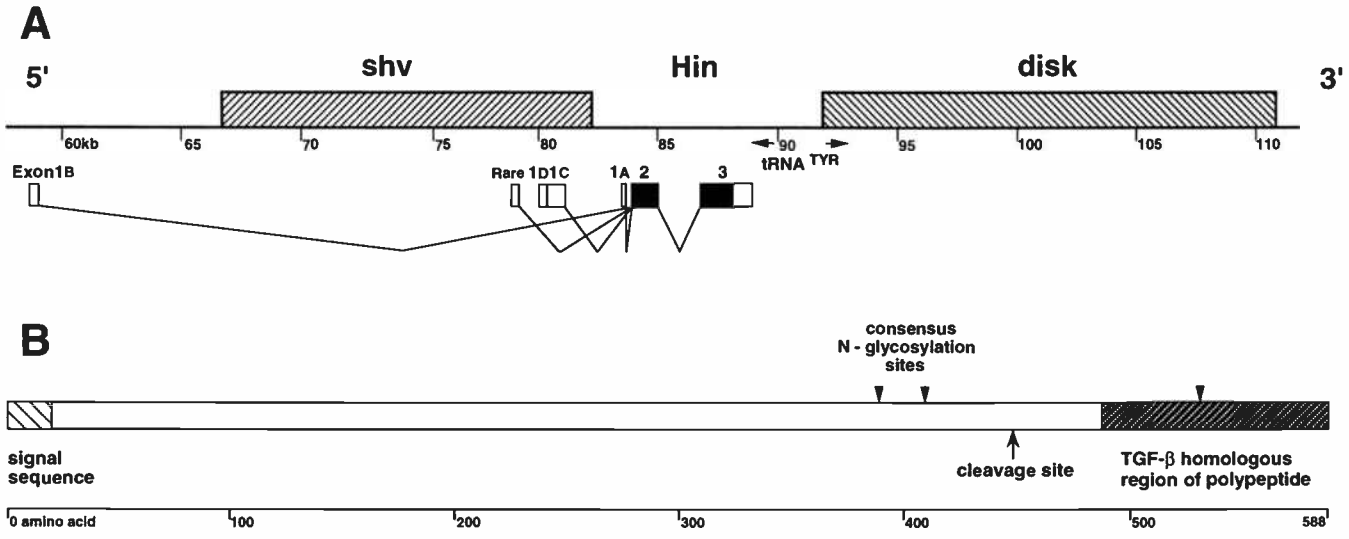

FIGURE 1.—(A) Organization of the *dpp* gene. The composition of the five known transcripts is given below the gene diagram, showing the alternate first exons and introns, and the common second exon, second intron and third exon. Filled-in portions of exons are coding. (B) Schematic diagram of the regions of the Hin protein. The arrows show the cleavage site and the positions of glycosylation sites (adapted from GELBART 1989).

phenotype in the absence of the treatment (Waddington's "genetic assimilation"), is accompanied by a change in the relative frequency of certain "wild-type" nucleotide polymorphisms at the *Ubx* locus. In addition, some population sequence studies have been done for parts of regulatory polypeptides, but these have often not encompassed the entire gene structure, but were carried out in the context of sampling the polymorphism from an available gene region (for example, HEY and KLIMAN 1993). An exception is the study by LEICHT *et al.* (1995) of the gene for myosin alkali light chain, which concentrates on regional constraints and function (see DISCUSSION below).

It is the purpose of the present paper to contribute to the bridge between developmental and population studies by reporting the results of a population genetic investigation of a gene, *dpp*, that is a major player in both the embryonic and imaginal development of Drosophila and that interacts with a variety of other loci concerned with early development.

THE DPP GENE COMPLEX AND POLYPEPTIDE

The *dpp* gene on chromosome *2* (location 2-4.0;22F1-2) of *D. melanogaster* occupies ~55 kb. Figure 1A, based on GELBART (1989) and ST. JOHNSTON *et al.* (1990), shows the main regions of the gene as differentiated by the effect of mutations. Mutations in the shv (short vein) region result in wing vein abnormalities; mutations in the Hin (haplo-insufficient) region are embryonic lethals; mutations in the disk region result in any of 15 different imaginal disc abnormalities. There are five known transcripts involving three exons, appearing in different tissues at different times. These transcripts differ in which one of five different 5′ exons are used, but all share the second and third exons from the Hin region that specify an open reading frame for the Hin protein. All these features are summarized in Figure 1. There are, in addition, two tyrosyl tRNA genes at the 5′ end of the disk region and several DNA binding sequences for the Ubx protein in the shv region.

The polypeptide encoded by the exons 2 and 3 from the Hin region is shown diagrammatically in Figure 1B. In *D. melanogaster* it has 588 amino acids (593 in *D. simulans*). About 40 amino acids at the N-terminus correspond to the consensus for the signal peptide of a secreted protein, and the 102-amino acid stretch at the C-terminus shows very extensive similarity to the TGF-β class of polypeptides in vertebrates that includes several BMPs (bone morphogenesis proteins), inhibins that suppress the release of follicle-stimulating proteins by the pituitary, and Mullerian duct-inhibiting substance. The amino acid similarity with BMP-2A is 75%. The middle 440-amino acid stretch contains a cleavage site at amino acid 455, shown by the arrow in Figure 1B, so that a smaller mature active protein containing the TGF-β fragment is created by cleavage. There is also evidence that the large N-terminal fragment continues to be physically associated with the TGF-β moiety after cleavage (see GELBART 1989, for a summary of features of the polypeptide). Activity of the Hin protein appears to be essential for dorso-ventral patterning of the embryo (IRISH and GELBART 1987).

Because of the importance of the Hin protein region, our first step in understanding the nucleotide variation at the *dpp* locus has been a study of the Hin region.

MATERIALS AND METHODS

Fly stocks: Eighteen isofemale lines of *D. melanogaster* caught from a natural population in New Jersey (courtesy of

TABLE 1

Polymorphism in *D. melanogaster*

Region	Length (bp)	Polymorphisms	Proportion of silent sites polymorphic
5′	295	5	0.017
Exon 1a	171	1	0.006
Intron 1	148	1	0.007
Exon 2-UT	14	0	—
Exon 2 coding	868	9 (3 amino acids)	0.026
Intron 2	1738	54	0.013
Exon 3 coding	899	14	0.062
3′ UT	1075	15	0.014
Total	5208	99	0.0247 (in 3883 effectively silent sites)

M. KREITMAN, University of Chicago) were isogenized for the second chromosome, using SM2, choosing one chromosome from each line. Four *D. simulans* lines (lines cyp, ndp, dsp and nbp), iso-chromosomal for the left arm of chromosome 2, were provided by J. COYNE, University of Chicago (originally assembled from independent laboratory sources worldwide). The remaining *D. simulans* lines are all iso-female and were obtained again from M. KREITMAN.

Genomic DNA preparation and cloning: Total genomic DNA was extracted from adult *D. melanogaster* flies according to standard procedure (AUSUBEL *et al.* 1987; ASHBURNER 1989) and cut with *Hin*dIII, which yields a 7.7-kb fragment including the entire Hin region. Libraries were constructed using (Lambda)DASH (Stratagene) according to manufacturer's directions. (Lambda)DNA was extracted using standard procedures (AUSUBEL *et al.* 1987) and cut again with *Hin*dIII. A fragment of 7.7-kb was subcloned into pEMBL 18+, with the resulting double-stranded plasmid DNA being used as a sequencing template.

PCR: The dpp region of interest was PCR amplified in three portions from single fly preps of *D. simulans* genomic DNA, purified and used directly in sequencing reactions. The general protocol involved double-stranded amplification, using rTth DNA polymerase (Perkin Elmer), with the following primers (nucleotide position corresponding to the sequence of *D. melanogaster*, NEWFELD *et al.* 1996): 11944-11963 and 14053-14073; 14002-14021 and 15646-15665; 15523-15542 and 17192-17211.

Sequencing: Templates were sequenced in both directions using 52 primers designed from a *D. melanogaster* sequence provided by W. GELBART (Harvard University). The procedure involved cycle sequencing using the Dye Terminator sequencing chemistry [Applied Biosystems (ABI)] according to the manufacturer's directions with slight modifications therof. Electrophoresis was carried out on an ABI 370A Automated DNA Sequencer. There were at least two independent coverages of each cloned sequence and three of each amplified sequence with a careful check for ambiguous positions and for short oligonucleotide sequences that are known to give repeatable automated sequencing errors. In a few cases positions were checked by manual sequencing.

Analysis: Nucleotide sequences were aligned manually in the Genetic Data Environment (GDE) (SMITH *et al.* 1994). Phylogenetic analyses were performed with the programs MacClade (MADDISON and MADDISON 1992), PAUP (SWOFFORD 1991) and MEGA (KUMAR *et al.* 1993). Divergence between species was analyzed using the SYNSUB algorithm (LEWONTIN 1989).

RESULTS AND ANALYSIS

***D. melanogaster* site polymorphism:** For all 18 genomes of *D. melanogaster* we determined 5.2 kb of sequence encompassing virtually the entire Hin region as defined by GELBART (1989). The sequence extends from a point 295 bp 5′ of Exon 1a, through a point 1075 bp downstream of the end of translation, 100 bp 3′ of the putative AATAAA cap signal. Thus, the sequence includes both exon 2 and exon 3 that code for the Hin polypeptide, the large common intron between them and one of the untranslated exons.

There were 99 polymorphisms including 15 insertion/deletion events in 5208 bp (counted by including all the insertions in the total). The distribution of these polymorphisms by DNA region is given in Table 1 and Figure 2. Of 23 polymorphisms in coding regions, three were replacements and 20 silent. There are 3883 effectively silent sites including coding and noncoding DNA, so the total proportion silent polymorphism is 96/3883 = 0.0247. The distribution of the allele frequencies, given as absolute number of copies of the rarer allele, are given in Table 2. A more detailed examination of the polymorphisms reveals a number of significant features of the pattern of variation and constraint.

1. There is a significantly higher silent polymorphism in coding (0.045) than in noncoding (0.022) regions ($\chi^2 = 7.78$; $P = 0.005$), and this is not simply a consequence of the high value in exon 3, as the two exon coding regions are not significantly different from each other ($\chi^2 = 2.3$; $P = 0.12$). There are, then, constraints on nucleotide variation in the noncoding regions of the transcript, including the intron and 3′ untranslated DNA, as compared with silent positions in the codons. A higher level of nucleotide polymorphism in coding than in noncoding regions is a common phenomenon in Drosophila (MORIYAMA and POWELL 1996), but the causes are diverse in different genes, including linkage with a balanced polymorphic site in the coding sequence and constraint on control regions, as well as a variety of observed conservations of 3′ regions that are unexplained.

2. The polymorphisms of nucleotides in untranslated regions are nonuniformly spread. Figure 2 shows the position of the nucleotide polymorphisms along the sequence. There is an apparent clustering of the 54

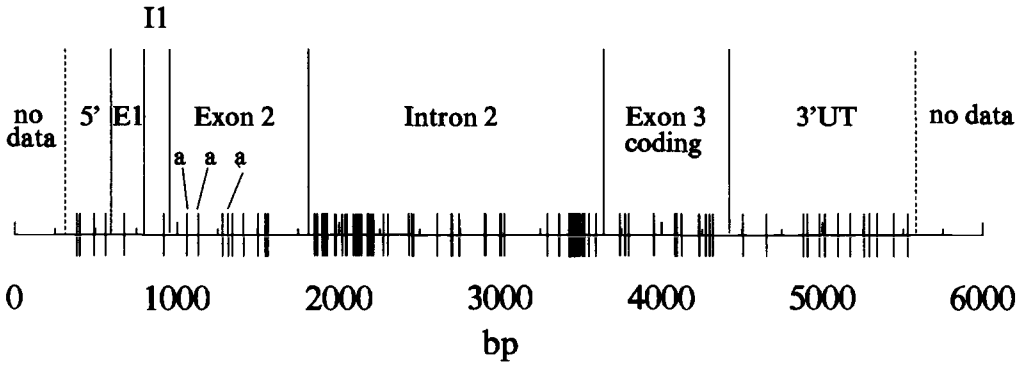

FIGURE 2.—Positions of the polymorphic sites in *D. melanogaster*. a denotes amino acid polymorphisms.

polymorphisms of the large intron at the ends as compared to the middle. When the spatial distribution of the polymorphisms is tested for randomness by the variance test of GOSS and LEWONTIN (1996), it is highly significantly clustered (observed variance of interval length = 0.00078; $P < 0.01$). There is no significant clustering within the exons.

3. Although it is not detected by a clustering test, there appears to be a region of low polymorphism in the middle of exon 3 around the boundary between the translated and untranslated portions. As will be shown below, from species comparisons, there is a 110-bp invariant region implying complete constraint located here.

4. The three replacement polymorphisms are at amino acids 59(V/G), 121(K/M) and 130(I/N), all toward the N-terminal end of the cleaved pro-protein fragment. Like most proteins, the empirical composition of the Hin polypeptide predicts that only 25% of random single base changes will be silent. Since there are 14 silent polymorphisms in the pro-protein fragment there should be 56 replacements expected, while only three are observed, giving an estimated selective constraint of 94.6%. While we might expect an absence of polymorphism in the signal peptide and TGF-β moieties, the source of amino acid conservation in the pro-protein fragment is not obvious. The interspecies comparison (see below) shows a striking parallel with the polymorphism result.

Comparison with *D. simulans* and other species: Only two *simulans* lines were completely characterized over most of the Hin region, and the results are given in Table 3. There were 37 polymorphisms in 4272 bp, of which 34 were silent out of a total of 2717.5 effectively silent sites (0.0125). While this is only half the proportion polymorphism observed in *D. melanogaster*, a lower value is expected because of the difference in sample size. A direct comparison can be made however, using the average pairwise silent difference between genomes (haplotype diversity). In *D. melanogaster*, excluding insertion/deletion polmorphisms, the average difference between pairs is 25.739 in 3883 effectively silent sites (0.0066), compared to 0.0125 in *D. simulans*, so the *simulans* sample is, if anything, more polymorphic than *melanogaster*. This may simply reflect the wider geographic origin of the *simulans* lines.

The three replacement polymorphisms observed in the two *simulans* genomes include an out-of-frame six-

TABLE 2

Frequency distribution of the absolute number of copies of the rare allele at the polymorphic sites in the 18 sequences of *D. melanogaster*

	No. of sites	
No. of copies	All sites	Biallelic sites
1	28	26
2	35	32
3	8	6
4	6	6
5	7	7
6	7	7
7	5	5
8	2	2
9	1	1
Total	99	92

TABLE 3

Distribution of *D. simulans* polymorphism by region in the two strains completely sequenced

Region	Length (bp)	Polymorphisms	Proportion silent site polymorphism (2 strains)
Exon 1a (partial)	70	0	0
Intron 1	148	1	0.0067
Exon 2 UT	14	0 (2)[a]	0
Exon 2 coding	883	4[b] (3)	0.0045
Intron 2	1777	27 (11)	0.0152
Exon 3 coding	899	2	0.0089
3' UT	481	3	0.0062
Total	4272		0.0121 (in 2717 effectively silent sites)

[a] 1 silent, 3 replacements.

[b] Figures in parentheses are additional polymorphisms in the eight partially sequenced strains.

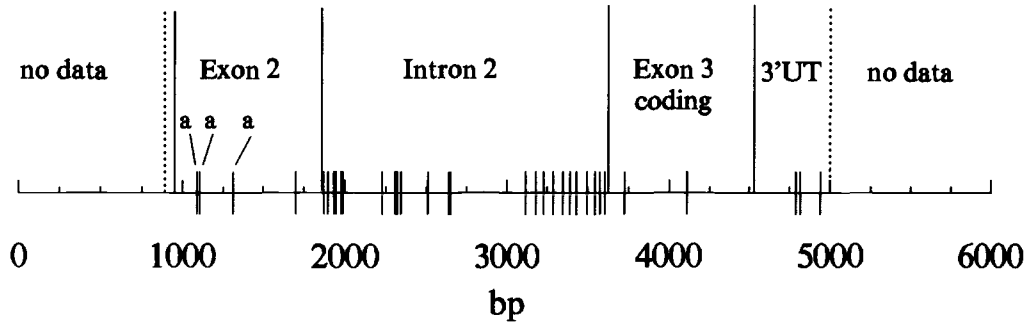

FIGURE 3.—Positions of the polymorphic sites in *D. simulans*. a denotes amino acid polymorphisms.

base insertion polymorphism in exon 2, ATCGGA, resulting in a polymorphic duplication, lacking in the *melanogaster* sequence, of a pair of amino acids, -Ser-Gly-, in a region of repeated -Ser-Gly-. Because such a polymorphism in polypeptide length has not been seen in previous studies on enzyme proteins, an additional eight lines of *simulans* were sequenced for 1231 bp including the end of intron 1a, all of exon 2, and 142 bp of intron 2. Four out of a total of 10 strains carried the insertion, so the polymorphism is in intermediate frequency. The sequencing of these additional eight lines revealed yet more *simulans* polymorphism as expected. Another insertion polymorphism, relative to *melanogaster*, of 9 bp in frame, AAC CAC AAC (-Asn-His-Asn-), was found in a region of repeated sequence 210 bp downstream from the first one. Moreover, there was an additional replacement polymorphism and 14 more silent polymorphisms detected in these additional lines.

The distribution of the *simulans* polymorphisms along the sequence, found in the two completely sequenced lines, is shown in Table 3 and Figure 3. The same pattern that appears in *melanogaster* is repeated (see Figure 2). There is a concentration of silent polymorphisms at the two ends of the common intron, with relative conservation in the middle ($P = 0.01$ by the Goss and Lewontin variance test.) The three replacement polymorphisms in the two completely sequenced lines, (as well as the two found in the eight partially sequenced lines) including the two insertion/deletions of two and three amino acids each are all in the region between amino acids 40 and 115, which is the N-terminal end of the cleaved pro-protein fragment. Finally, there is the same appearance of conservation in exon 3 around the end of translation, although, again, it is not statistically significant by the Goss and Lewontin variance test.

Finally, the same pattern appears when the fixed differences between *melanogaster* and *simulans* are examined. We include in fixed differences, in addition to sites differentially fixed, sites that are polymorphic in one species, but fixed for yet a different nucleotide in the other species. Table 4 and Figure 4 show the distribution by region of these differences. Of 76 fixed differences, 73 are synonymous, and in the coding region 10 out of 13 are synonymous. In Table 5 a comparison is made of the proportion of fixed and polymorphic differences between and within species. For this purpose only the polymorphisms observed in the original two *simulans* lines are included because the additional lines were sequenced only for exon 2 where we already expect a concentration of replacement polymorphisms. There is no difference in the ratio of silent to replacement substitutions between polymorphic and fixed substitutions (Fisher's Exact Test $P = 0.57$) so there is no evidence of a selective divergence in amino acids between the species (MCDONALD and KREITMAN 1991). Two out of three of the fixed amino acid differences between species are at the N-terminal end of the pro-protein. In fact, the fixed difference at amino acid 40, which is a histidine in *D. melanogaster*, is in the same codon that is polymorphic for glycine and glutamic acid in *simulans*. The third fixed difference, however, is in the middle of the pro-protein region. Again, the concentration of differences at the ends of the large intron is significant by the variance test ($P < 0.01$).

As with polymorphisms within species, there is a lack of fixed nucleotide difference in the region around the end of translation, but there are too few fixed differences to have any statistical power in the variance test. Merging the polymorphisms and fixed differences gives a total of 46 events in exon 3 and the observed variance in length of intervals between adjacent variable sites is 0.00051, which, although larger than expectation, is still not significant ($P = 0.15$). When this region is compared between more divergent species, however, a striking and unambiguous region of conservation is found. In the accompanying paper, NEWFELD *et al.* (1996) give the sequence of the Hin region from *D. melanogaster*,

TABLE 4

Fixed differences between *D. melanogaster* and *D. simulans*

Exon 1 a	0
Intron 1	1
Exon 2 coding	4 (2 amino acids)
Intron 2	59
Exon 3 coding	9 (1 amino acid)
3′ UT	3
Total	76

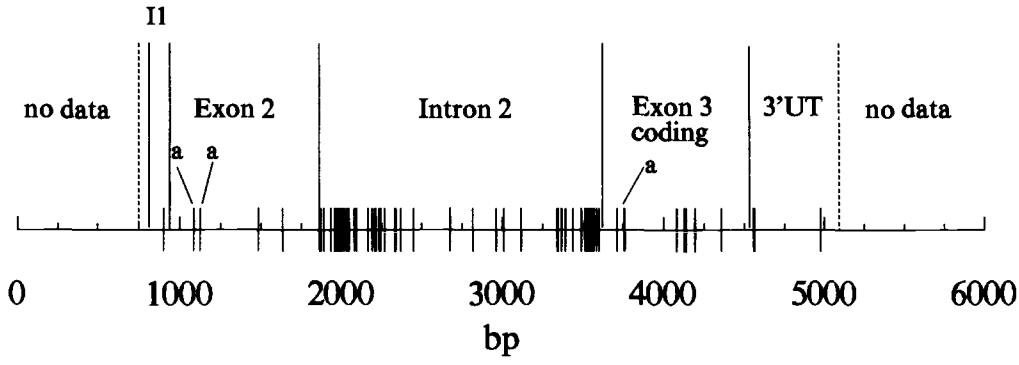

FIGURE 4.—Positions of fixed differences between *D. melanogaster* and *D. simulans*. a denotes amino acid replacements.

D. pseudoobscura and *D. virilis*. Figure 5 shows the aligned sequences for 300 bp beginning at the translation terminator. On a background of considerable differentiation among the species, there is a region of 110 bp of perfect conservation, starting ~100 bp downstream of the termination of translation. It is presumably this signal that is being detected, although not statistically significantly, by the variance test.

Haplotype analysis: The complete set of the 99 polymorphic positions of the 18 *D. melanogaster* haplotypes is shown in Figure 6. Dots denote agreement with the sequence of a standard laboratory strain used in species comparisons by NEWFELD *et al.* (1996). The letters X, Y and Z stand for duplication tracts of different lengths covering the same duplicated region. These are taken to be single alternative events. Table 2 gives the distribution of allele frequencies of the rarer allele at the 99 polymorphic positions. As is characteristic of molecular polymorphism data, most polymorphic sites are represented by variants present once or twice in the sample. The average haplotype diversity, including both silent and replacement sites and insertion events, is 0.0054.

It appears from Figure 6 that there are groups of similar haplotypes that differ little within groups, but have many differences from other groups. This is shown numerically in Table 6, giving the pairwise differences among haplotypes, and in Figure 7 giving the distribution of these pairwise differences. There are three clear modes, one at zero to three differences, one at 16–19 differences and one at 32–35 differences, indicating that there are clusters of related haplotypes and some subclustering within these. The structure of these clusters is shown in Figure 8, which is a phylogeny of the haplotypes using the PAUP neighbor-joining algorithm.

TABLE 5

Comparison of fixed and polymorphic synonymous and replacement substitutions in the ORF of the Hin protein for *D. melanogaster* and *D. simulans*

	Polymorphisms			
	melanogaster	*simulans*	Total	Fixed
Synonymous	20	3	23	10
Replacement	3	3	6	3

Branch lengths in the figure are proportional to divergence and bootstrap values are shown on the branches. Exactly the same phylogeny appears when the branch-and-bound algorithm is used.

An inspection of the haplotypes in Figure 6 shows that the differences that characterize the haplotypic clades are not scattered across the sequence, but that there tend to be linked blocks of sequence that separate haplotypes. The usual tests of linkage disequilibrium in two-by-two contingency tables are not applicable because so many of the polymorphic sites are represented by alleles present only once or twice in the sample, but it is possible to carry out a sign test on the direction of the linkages (see LEWONTIN 1995 for a discussion of the problem). Defining k as the absolute number of copies of the rarer allele at a biallelic site, and m as the absolute number of copies of the rarer allele at another

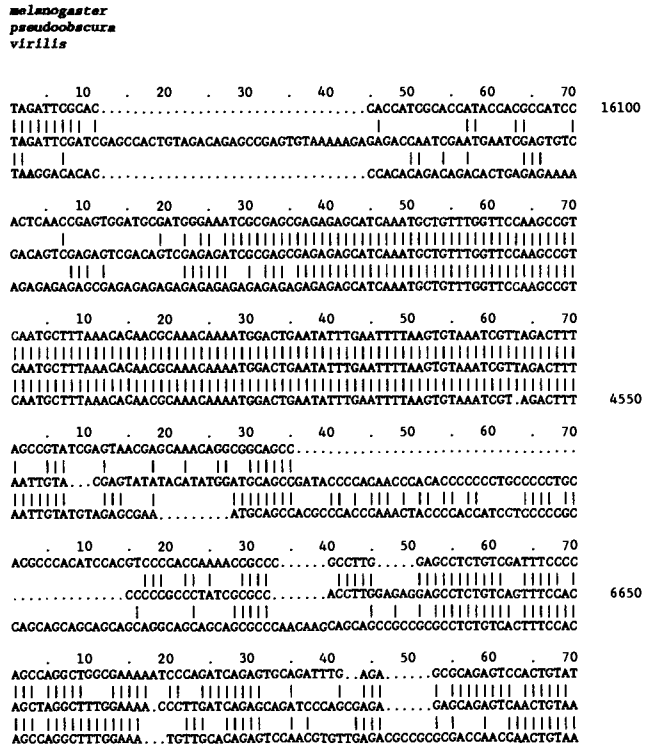

FIGURE 5.—Untranslated 3′ sequences from exon 3 of *D. melanogaster, D. doobscura* and *D. virilis* (from NEWFELD *et al.* 1996).

```
MK01  TXX.G..T.GA.....T.AAA.CATACC......ACCAGGA....X....C.....G.X..G.XX....C..CC..C.C....C.X..CX.......A.
MK11  ........TGA.........X..X......A...X..........X...T........X..G.XX....C....CT.........X...A....Y..X.
MK13  ..TT......A.TT.A....X..X..CXX.....ACCA.GA.C..X....C.......X..G.XY....C......C.C...T.....CX.CA.X..AA
MK14  ..Z.G...TGA.........X..X..........X..........X.......A....X..G.XX....C.........ATT...X..CXTC.CX..A.
MK15  .YT.G...TGA.........X..X..C.......ACCA..A....X............X....XY.........CT...ATT..AX...A....Y..X.
MK18  ..Z.G...TGA.........X..X..........X..........X.......A....X..G.XX....C.........ATT...X..CXTC.CX..A.
MK22  .YT.G.....AT.T......X..X..........X..........X....CG..GT..TTTGGGTATCGCC.........ATT.......A.....T.X.
MK23  ..Z.G...TGA.........X..X..........X..........X......C.....X..G.XX....C.........ATT...X..CXTC.CX..A.
MK26  .YT.G...TGA.........X..X..CXX.....X......GC..X....C...GT..TTTGGGTATCGCC.........ATT.....G.A.......X.
MK33  ..TT......A.TT.A....X..X..CXXT..A.Y.......CA.X....C.......X..G.XY....C......C.C...T.....CX.CA.X..A.
MK38  TXX.G..T.GA.....T.AAA.CATACC......ACCAGGA....X....C.......X..G.XX....C..CC..C.C....C.X..CX.......A.
MK42  ........TGA...T.....X..X..........X..........X............X....XY.........CT.T.ATT..AX...X......CX.
MK43  ....G.T.TGA.........X..X..........X.....X....X............X....XY.........CT...ATT..AX...X......CX.
MK47  ....GT....A......T..XT.X..C....A..ACCA.GA.C.CGA.A.C.......X..G.XT....CCGC...C.C...T..X..CX.C..X..X.
MK49  ......................X..X.........TX..........X............X....XY.........CT...ATT..AXT..X......CX.
MK54  ..Z.G...TGA.........X..X..........X..........X.......A....X..G.XX....C.........ATT...X..CXTC.CX..A.
MK92  ..T.G................X..X.........TX..........X............X....XY.........CT...ATT..AX.G.A....Y..X.
MK10  ....G...TGA.........X..X..........X..........X.A.........CX....XY.........CT...ATT..AXT..X......CX.
```

FIGURE 6.—Polymorphisms of the 18 haplotypes of *D. melanogaster*. · denote agreement with the standard *D. melanogaster* sequence given in NEWFELD *et al.* (1996). X, Y and Z are alternative duplications of different length in the same overlapping duplicated region.

site for which $m \geq k$, then pairs of sites can be classified into k, m classes. Defining a negative association between sites as one in which the common allele at one site is associated with the rare allele at the other site, the probability of an observed negative disequilibrium for a k, m class can be calculated under the null hypothesis of no true disequilibrium, and the observed proportion of negative associations can be compared to the expected by a goodness-of-fit test. For s sites there are $s - 1$ independent pairs that can be tested, one possible set, used here, being adjacent pairs down the sequence. The result for the 91 pairs involving biallelic sites is given in Table 7 for each k, m class. The expected total number of negative associations is 62.72, but only 50 were observed ($G = 9.46$; $P = 0.002$), so there is a significant excess of positive associations, that is, of rare alleles at different loci in coupling.

The linkage disequilibrium expected from mutation alone is a negative one since new rare alleles at a site should usually arise on the strand with the more common allele at another site. The excess of positive associations, together with the clade structure of the haplotypes, strongly suggests that the present population is a mixture of several divergent haplotypes that have subsequently recombined, but insufficiently to bring the population to linkage equilibrium. Moreover, the appearance of short, apparently recombined, tracts internal to the haplotypic sequences (see Figure 6) suggests that gene conversion rather than reciprocal crossing over has been involved. The observed linkage raises the possibility that the population may be segregating for second chromosome inversions. To check this, all strains were crossed to a standard line to look for inversion loops, but no inversions were detected with the

TABLE 6

Number of pairwise differences among the 18 lines of *D. melanogaster*

Line	M11	M13	M14	M15	M18	M22	M23	M26	M33	M38	M42	M43	M47	M49	M54	M92	M100
MK01	38	33	37	36	37	52	37	51	41	1	44	42	34	45	37	45	44
MK11		34	17	18	17	32	17	32	31	37	14	15	35	17	17	16	16
MK13			32	32	32	44	32	42	10	32	39	39	24	38	32	38	41
MK14				22	0	31	2	31	29	36	18	17	34	21	0	20	18
MK15					22	33	22	31	38	35	14	12	34	17	22	12	14
MK18						31	2	31	29	36	18	17	34	21	0	20	18
MK22							31	12	41	51	33	32	46	32	31	29	33
MK23								31	29	36	18	17	34	21	2	20	18
MK26									39	50	33	32	46	36	31	31	33
MK33										40	36	37	32	35	29	35	38
MK38											43	41	33	44	36	44	43
MK42												5	41	7	18	12	6
MK43													39	8	17	11	5
MK47														40	34	40	41
MK49															21	7	7
MK54																20	18
MK92																	12

Mean difference = 27.915
Variance of differences = 156.4598

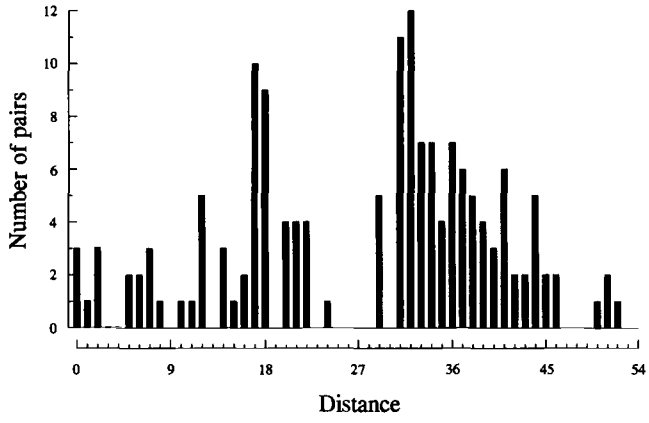

FIGURE 7.—Number of pairwise comparisons of haplotypes having different numbers of nucleotide differences.

exception of a single strain that carries an inversion for one-third the length of 3*R*, probably In(3R)P.

To aid in further analysis of the events leading to the haplotype structure, the following operations were performed on the phylogram of sequences produced by PAUP.

(1) Identical haplotypes were merged into a single representative, reducing the number of taxa from 18 to 15.

(2) Singleton characters, which are not phylogenetically informative, were removed from the haplotypes, reducing the number of characters from 99 to 59, and with the further reduction of distinguishable haplotypes from 15 to 14.

(3) The order of the haplotypes in the phylogram was rearranged, without changing the topology of the phylogram, to maximally order the hierarchy from the bottom to the top of the taxon list.

The result is shown in Figure 9. Both neighbor-joining and branch-and-bound algorithms produced the same maximum parsimony phylogeny. Asterisks at the top of the figure indicate sites that produce homoplasies in the phylogram. These homoplasies are the result either of mutations or of recombinations that are predominantly gene conversion events (BERTRAN *et al.* 1996). It is not possible to distinguish conversions from mutations unambiguously, but a conservative estimate of the number of conversion events can be obtained from homoplastic polymorphisms with the same phyletic structure. These are generally at multiple adjacent polymorphic sites, but may be interrupted by a single mutation in one haplotype, or by a second conversion within the first conversion tract. The location of these putative conversion events is shown at the top of the figure by underlines connecting adjacent homoplasies. In addition, there are some homoplasies whose topologies require more than one homoplastic event and these are indicated on the figure by numbers of independent events required. For example, the leftmost conversion tract involving two adjacent polymorphic sites requires a minimum of three independent conversions at the tips of the phylogeny plus one mutation. The long inferred conversion tract whose end points are the sixth and eleventh homoplastic sites is problematic, because it would imply a tract of at least 816 nucleotides, although conversion tracts of this length certainly occur (NASSIF and ENGELS 1993). The alternative is to invoke two independent short conversions, one at each end of the interval involving the same rare haplotype.

The 30 homoplasies can be accounted for by five conversion tracts and 17 isolated sites that may be either conversions or mutations. A striking feature of the data is the concentration of homoplasies in the 5′ half of the sequence, comprising 2700 bp. Four of the five clear conversion tracts appear in this region as do all of the sites with multiple independent homoplastic events. Taking only the multiple site putative conversion tracts,

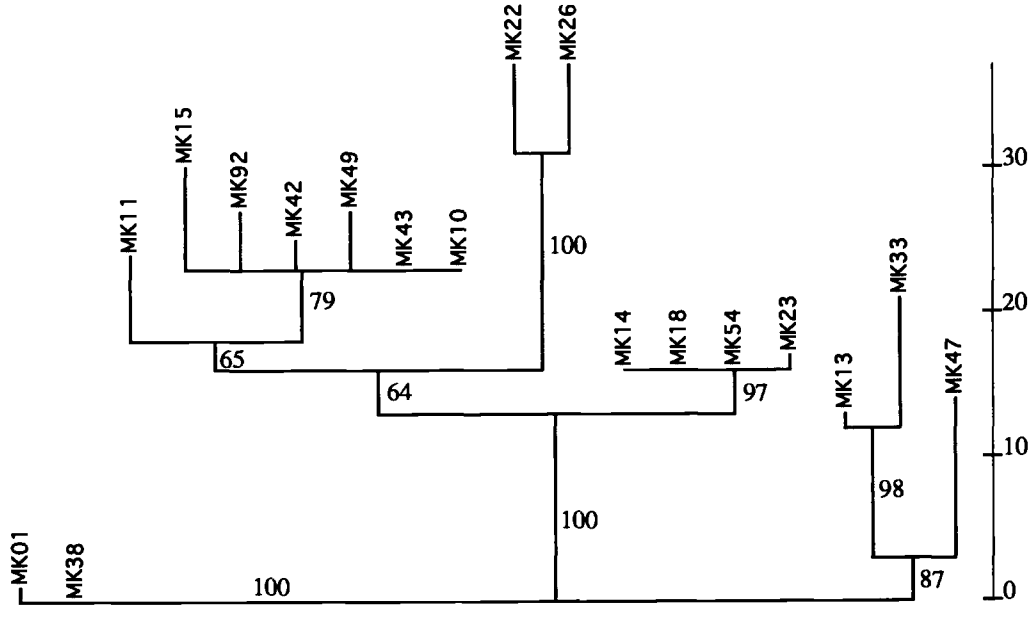

FIGURE 8.—Maximum parsimony tree showing phyletic relations among haplotypes with branch lengths shown proportional to number of character differences. Numbers on the branches are bootstrap values.

TABLE 7

Comparison of observed and expected numbers of negative linkage associations for pairs of biallelic sites falling in different *k*, *m* classes

k	*m*	NEG	TOT	P_{obs}	P_{exp}	*G*	*G* (corr.)
1	1	10	12	0.833	0.944	1.891	1.816
1	2	11	11	1.000	0.889	2.591	2.479
1	3	2	5	0.400	0.833	4.750	4.318
1	4	3	5	0.600	0.778	0.794	0.722
1	5	3	3	1.000	0.722	1.953	1.674
1	7	1	2	0.500	0.611	0.101	0.081
1	8	0	1	0.000	0.556	1.622	1.081
2	2	3	16	0.188	0.784	25.897	25.113
2	3	2	4	0.500	0.686	0.598	0.531
2	4	2	3	0.667	0.595	0.066	0.056
2	5	1	3	0.333	0.510	0.380	0.326
2	6	4	5	0.800	0.431	2.851	2.592
2	7	1	3	0.333	0.359	0.009	0.008
2	8	0	1	0.000	0.294	0.697	0.464
2	9	1	1	1.000	0.235	2.894	1.929
3	6	1	2	0.500	0.270	0.477	0.382
3	7	0	1	0.000	0.674	2.242	1.495
4	5	1	1	1.000	0.701	0.711	0.474
4	6	1	1	1.000	0.593	1.045	0.696
4	7	0	1	0.000	0.485	1.328	0.886
4	8	1	1	1.000	0.382	1.923	1.282
5	5	0	2	0.000	0.567	3.352	2.682
5	6	0	1	0.000	0.439	1.156	0.771
5	7	1	1	1.000	0.324	2.257	1.505
5	9	1	1	1.000	0.500	1.386	0.924
6	6	0	2	0.000	0.306	1.460	1.168
6	8	0	1	0.000	0.437	1.148	0.765
7	7	0	1	0.000	0.417	1.079	0.719

Total $G = 66.657$; d.f. = 28; $P < 0.001$. Total G with Williams correction = 56.937; d.f. = 28; $P < 0.001$. Observed number of negative pairs = 50. Expected number of negative pairs = 62.72. Goodness-of-fit $G = 9.46$; $P = 0.002$.

the observed mean length of the nine conversions involving the five distinguishable tracts is 169.3 bp including the very long tract of 816 bp, and 88.5 excluding this tract. These observed values do not take into account the fact that the end points of the actual conversions must be farther apart than observed and that many short tracts either leave no trace or involve only a single homoplastic site and so cannot be distinguished from mutation. While a method of estimation of the mean true lengths has been given by BETRAN *et al.* (E.

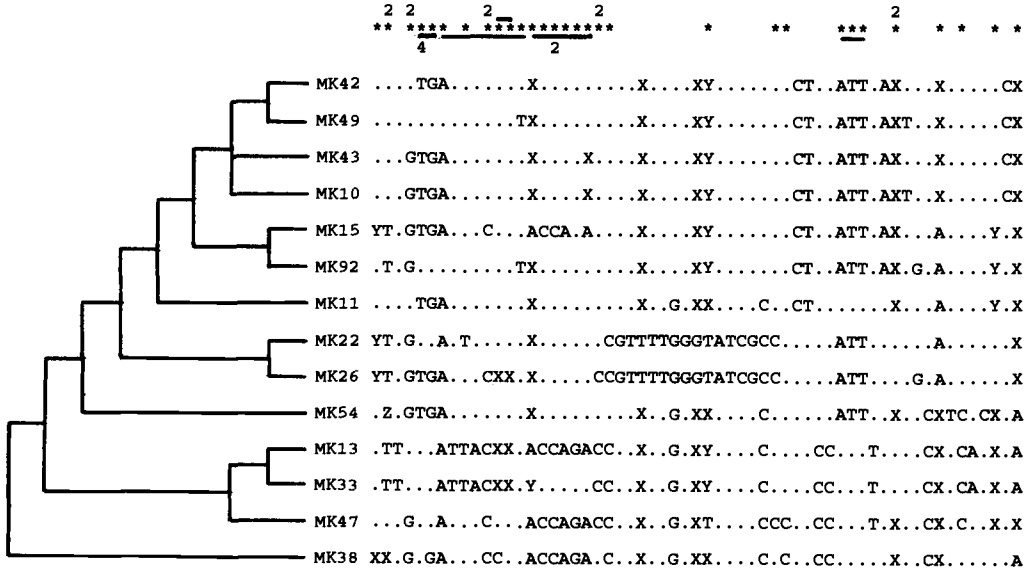

FIGURE 9.—Haplotypes arranged according to the phylogram. Homoplastic sites are indicated by * and putative conversion tracts by brackets at the top of the diagram. Numbers above the homoplasies indicate the number of independent homoplastic events at each site needed to explain the phylogram.

BETRAN, J. ROZAS, A. NAVARRO and A. BARBADILLA, unpublished results), this method is applicable only to data in which there are two *a priori* subpopulations of sequences being compared.

DISCUSSION

There are two kinds of questions that are addressed by our results. The first concerns the inferences that can be made about selection operating on the amino acid and nucleotide sequences in the evolution of the Hin region of the *dpp* locus. The other concerns the way in which migrational history and recombination mold the variation seen within a population.

The first most obvious and unsurprising result is that there are very strong overall constraints on substitutions, as there are for all genes ever studied by sequence comparisons in Drosophila. Nor are the constraints on the polypeptide for this gene of early development unusually strong, as compared, say, to an enzyme like Adh, which shows no amino acid polymorphism at all in a very large sample of *D. pseudoobscura* (SCHAEFFER and MILLER 1993).

Second, there is the question of whether there is evidence that the rate of evolution of amino acid sequences between species is any greater than would be expected from random fixation of allowable amino acid variation within species, as judged from polymorphism within *D. melanogaster*. It was already shown in Table 4 that the ratio of amino acid to synonymous divergence between *D. melanogaster* and *D. simulans* is what is expected from the polymorphism, using the test suggested by MCDONALD and KREITMAN (1991). These two species are quite close, however, and the inference can be supplemented by the much more distant comparisons with *D. pseudoobscura* and *D. virilis* from the data in the accompanying paper by NEWFELD *et al.* (1996). To make this comparison we have used the SYNSUB method of LEWONTIN (1989), which involves a codon by codon comparison to correct the observed divergence for multiple hits. The advantage of this method is that it makes no assumptions about equality of synonymous substitution rates among codons, and it also uses a parsimony method to infer unobserved intermediate amino acid changes. In making the comparisons, a region of 150 codons of the pro-protein region was omitted because no consistent alignment of the three species can be made in this region although short regions of homology are seen for pairs of species. The result of the corrected comparison of the alignable sequence is given in Table 8. The first column gives the synonymous and replacement polymorphisms we have observed in this paper (*D. melanogaster* and *D. simulans* combined from Table 6), while the second, third, and fourth columns give the fixed differences of *D. melanogaster* from *D. simulans*, *D. pseudoobscura* and *D. virilis*. Neither of the more distant species divergences from *D. melanogaster* is significantly different in the replacements/synonymous ratio from the ratio in the polymorphism data (*mel-pseudo:* $\chi^2 = 0.65$, $P = 0.4$; *mel-virilis:* $\chi^2 = 0.002$, $P = 0.94$), so on this evidence the divergence seems to be the result of random fixations. On the other hand, while neither differs significantly from the *melanogaster* polymorphism, they do differ significantly from each other ($\chi^2 = 8.8$; $P = 0.003$), so there may indeed be some selective divergence between *pseudoobscura* and *virilis*. The comparison of the synonymous/replacement ratio between these two distant species has a much larger sample size because of the large number of total differences, so the test is much more powerful than the comparisons of the divergences with the 29 polymorphisms. It is not simply a matter of power, however, because the observed value of the synonymous/replacement ratio in the total polymorphisms (3.83) is very close to the synonymous/replacement ratio in the divergence of *D. melanogaster* from *D. simulans* (3.33) and from *D. virilis* (4.25). Indeed, no other partition of the 29 polymorphisms would be as close to these species divergence ratios as the 23:6 actually observed. On the other hand, it is reasonable to suppose that a selective divergence occurred in the *D. pseudoobscura* divergence from *D. melanogaster*, which cannot be detected because of low power in the McDonald-Kreitman test, but which does show up in the *D. virilis-D. pseudoobscura* comparison.

Third, there is the question of whether the variation observed within *D. melanogaster* and *D. simulans* can be explained as purely neutral. Unfortunately, the most widely used statistical tests for selection when polymorphism and divergence data are available, the HKA test (HUDSON *et al.* 1987) and the TAJIMA (1989) test, make assumptions of stationarity, of homogeneity across the sequence of neutral mutation rate, and about linkage, that are strongly contravened in our data. The operating characteristics of the Tajima test have been explored by SIMONSEN *et al.* (1995) who showed that sample sizes below 20 had virtually no power to detect any deviations from the null model and that there is no difference in relative power to detect selective as opposed to breeding history deviations, so that nothing specific can be learned. As yet, no study of the operating characteristics of the HKA test has been published. Given the knowledge or absence of knowledge of the robustness of these tests to major deviations from their assumptions, it seems prudent not to use them in this case, since the result, no matter what it might be, would be uninterpretable. The McDonald-Kreitman test, on the other hand, does not depend on assumptions about stationarity of allelic frequency distribution or population structure, and is insensitive to assumptions about recombination and variation in mutation rates (MCDONALD and KREITMAN 1991). The test detects a change in the selective conditions of amino acid replacement in species divergence as compared to the selection operating on the intraspecies polymorphism. It is im-

TABLE 8

Polymorphisms in *D. melanogaster* and *D. simulans* combined, compared with the fixed differences between *D. melanogaster* and *D. pseudoobscura* or *D. virilis* corrected for multiple hits

	Polymorphism in *D. melanogaster* and *D. simulans*	Fixed differences from *D. melanogaster*		
		D. simulans	*D. pseudoobscura*	*D. virilis*
Synonymous	23	10	172	272
Replacements	6	3	74	64

portant to note, however, that the test itself cannot distinguish an adaptive divergence from a change in the fraction of effectively deleterious mutations because of changes in effective population size.

When we turn from overall amino acid polymorphism to regional variation along the protein, there are clearly different degrees of constraint. As might be expected, neither the signal peptide of ~40 amino acids at the N-terminus, nor the TGF-β fragment at the C-terminus show any polymorphism, nor any fixed differences between *D. melanogaster* and *D. simulans*, all amino acid variation being in the pro-protein fragment. The concentration of the pro-protein polymorphism to the N-terminal end of that fragment is correlated with the presence, in the C-terminal half, of a cleavage site and of glycosylation sites that would be involved in the association of the pro-protein with the TGF-β protein after cleavage. The lack of amino-acid polymorphism and of fixed differences in the signal peptide and TGF-β regions does not carry over completely to the more distant comparisons. Figure 10 shows the entire amino acid sequence for the four species discussed in this and the accompanying paper. There are five amino acid differences in the signal peptide region and 12 amino acid differences in the TGF-β region among the more distant species. There is also an extremely diverged region where no homology can be seen in the pro-protein between the cleavage site and the beginning of TGF-β, when the more distant species comparisons are made. The extreme divergence to the point of absence of any detectable homology in this region and at the previously noted N-terminal end of the pro-protein, coupled with a number of insertion/deletion changes in these regions, suggests some major genetic event, for example, horizontal transfer, rather than more regular substitution.

At the nucleotide level there are two evident conservations that require comment. The heterogeneity of polymorphism in the large common intron is a reflection of conservation in the middle 50% of the intron, since the polymorphism level of the intron ends (0.046) is same as the silent polymorphism level in the coding regions of the exons (0.048). This same conservation is seen in the species divergence from *D. simulans* and from the more distant species. This conservation is partially explained by the discovery of sites in this region that are involved in repression of *dpp* expression in the ventral part of the embryo (HUANG *et al.* 1993). This repression is mediated by the binding of the *dorsal* (*dl*) morphogen at 14 sites in the middle of the intron, each of which contains the consensus binding sequence GGGWWWWCC, where W stands for A or T. There are, however, two *melanogaster* polymorphisms, two *simulans* polymorphisms and two fixed differences that violate these rules. There are, moreover, very long stretches of conservation that do not contain these consensus sequences, but are characterized by many three- or four-base A repeats or T repeats separated by GC-rich stretches. No function for such structures is presently

FIGURE 10.—Aligned amino acid sequences of the Hin protein for four species of Drosophila.

known, but the conservation suggests that there must be one. There is evidence from constructs that the intron does contain enhancer sequences (see NEWFELD *et al.* 1996).

The perfect conservation of nucleotide sequence, even over the highly divergent species, in the 3′ untranslated region of exon 3 has no current definite explanation, although a variety of mechanisms of transcriptional control involving interaction between 5′ and 3′ sequences are known. Evidence of conservation of 3′ untranslated sequence is also found in other genes that have been studied for polymorphism in Drosophila, for example *Adh* (KREITMAN 1983), and Est-5B of *D. pseudoobscura* (VEUILLE and KING 1995), but the perfect conservation of 110 bp over highly divergent species has not been previously observed. We performed a Genbank search on this sequence, but found no matches, so if there are general causes of 3′ conservation there is no single motif common to them.

The second kind of outcome of the polymorphism study concerns the evidence that the *dpp* locus provides on population structure. The presence of a large excess of positive linkage disequilibrium, with clear contiguous stretches of intact sequence of different origins, points strongly to the origin of the present New Jersey population as a migration mixture of several (roughly three to five are suggested by the phylogeny in Figure 8) divergent lines in the recent, but not immediate, past. What that recent history is, however, is unclear. Presumably North American populations of *D. melanogaster* are ultimately derived from Africa, but there is no consistent evidence of a simple path of origin. So, BEGUN and AQUADRO (1995) found that the haplotypes at the *X*-chromosomal *vermillion* locus sampled from California were a small subset of Zimbabwean haplotypes, but the same authors (BEGUN and AQUADRO 1994) found that the haplotypes of the *X*-chromosomal *Pgd* from these populations showed completely unrelated evolutionary histories. A confirmation of this hypothesis of mixture comes from the observation of an excess of positive linkage disequilibria in the same population for the *Adh* gene, also on the second chromosome (BERRY and KREITMAN 1993; A. BERRY, personal communication).

Since the origin of our New Jersey population there has been time enough for some mutational divergence within haplotype clades as well as transfer of tracts of sequence, probably by gene conversion, between clades, but insufficient time to produce the kind of linkage equilibrium seen, for example, in *Xdh* in *D. pseudoobscura* (RILEY *et al.* 1989). A similar excess of positive linkages can be found in *Adh* in *D. pseudoobscura* (SCHAEFFER and MILLER 1993; LEWONTIN 1995). In judging the relative importance of gene conversion as opposed to reciprocal recombination one cannot use the presence or absence of reciprocal products of recombination in population data. Because there is no recombination in Drosophila males, and because only one product of each meiotic event in females appears in an egg nucleus, individual reciprocal recombinations do not leave reciprocal products in the population. On the other hand, the low rate of intracistronic recombination predicts that reciprocal recombination should leave a "signature" in the population of recombined segments that continue beyond the left or right hand boundaries of the sequence. Isolated intermediate segments could only arise from a second independent recombination involving a rare strand. The much greater likelihood is that isolated recombinant segments arise from gene conversion.

Our analysis and results are strikingly similar to the study by LEICHT *et al.* (1995) of the gene for the myosin alkali light chain, a molecule that may have a regulatory function in muscle. That study also found very strong conservation of "silent" sites both in introns and exons, and used a genealogical analysis, as we did, to locate regions of intracisronic recombination. Their most extraordinary observation was the *complete* conservation of all nucleotide sites in exons, which could not be explained by a recent selective sweep since there was nucleotide polymorphism in introns. Overall, our study and the accompanying one of NEWFELD *et al.* (1996) show that even for a gene that lies at the center of early morphogenesis, there can be considerable polymorphism within species and difference between species. While this variation is not uniform across regions, and there is strong evidence of considerable sequence constraint in some gene regions, there is no simple and unambiguous relation between the function of a region and its degree of constraint. The most constrained region in our study, the 110-bp invariant 3′ region is of completely unknown function. On the other hand, the repressor motifs that are part of the explanation of the low polymorphism in the middle of the intron are not, in fact, totally invariant. Nor is it clear why there is so little amino acid variation in most of the pro-region of the protein. The rationalization of patterns of polymorphism and invariance of genes, even genes of central importance in morphogenesis, is still an open question.

We are grateful to MARTIN KREITMAN for having provided lines of *D. melanogaster* and *D. simulans*. ANDREW BERRY, DAN WEINREICH and ANDY CLARK, acting as Editor, provided characteristically perceptive and useful criticisms, and two anonymous reviewers were extremely helpful. We also are grateful to ANDREW BERRY for providing original data from his study of *Adh*. We owe a special debt to WILLIAM GELBART for having introduced us to the *dpp* locus and for illuminating its complexities for us. Contributions of authors were as follows: design, R.C.L.; line extraction, R.C.L. and E.N.; libraries and cloning, E.N., B.R. and M.L.; sequencing, B.R. and M.L.; cytology, R.C.L.; data analysis and writing, B.R. and R.C.L. The research was carried out under National Institutes of Health grants GM-21179 and GM-29301.

LITERATURE CITED

ASHBURNER, M., 1989 *Drosophila: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEID-

MAN *et al.* (Editors), 1987 *Current Protocols in Molecular Biology.* Wiley Interscience, New York.

BEGUN, D. J., and C. F. AQUADRO, 1994 Evolutionary inferences from from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of Drosophila: selection and geographical differentiation. Genetics **136:** 155–171.

BEGUN, D. J., and C. F. AQUADRO, 1995 Molecular variation at the *vermillion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans.* Genetics **140:** 1019-1032.

BERRY, A., and M. KREITMAN, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134:** 869–893.

EANES, W. F., M. KIRCHNER and J. YOON, 1993 Evidence for adaptive evolution of the *G6pd* gene in *Drosophila melanogaster* and *D. simulans* lineages. Proc. Natl. Acad. USA **90:** 7475–7479.

GELBART, W. M., 1989 The *decapentaplegic* gene: a TGF-β homologue controlling pattern formation in Drosophila. Development **Supplement:** 65–74.

GIBSON, G., and D. S. HOGNESS, 1996 Effect of polymorphism in the Drosophila regulatory gene *Ultrabithorax* on homeotic stability. Science **271:** 200–203.

GOSS, P. J. E., and R. C. LEWONTIN, 1996 Detecting heterogeneity of substitution along DNA and protein sequences. Genetics **143:** 589–602.

HEY, J., and R. M. KLIMAN, 1993 Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* complex. Mol. Biol. Evol. **10:** 804–822.

HUANG, J.-D., D. H. SCHWYTER, J. M. SHIROKAWA and A. J. COUREY, 1993 The interplay bewteen multiple enhancer and silence elements defines the pattern of *decapentaplegic* expression. Genes Dev. **7:** 694–704.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

IRISH, V. F., and W. M. GELBART, 1987 The decapentaplegic gene is required for doral-ventral patterning of the Drosophila embryo. Genes Dev. **1:** 868–879.

KREITMAN, M., and M. L. WAYNE, 1994 Organization of genetic variation at the molecular level: lessons for Drosophila, pp. 157–183 in *Molecular Ecology and Evolution,* edited by B. SCHIERWATER, B. STREIT, G. P. WAGNER, and R. DESALLE. Birkhauser Verlag, Basel.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

KUMAR, S., K. TAMURA and M. NEI, 1993 *MEGA: Molecular Evolutionary Genetics Analysis, version 1.0.* The Pennsylvania State University, State College.

LEICHT, B. G., S. V. MUSE, M. HANCZYC and A. G. CLARK, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139:** 299–308.

LEWONTIN, R. C., 1989 Inferring the number of evolutionary events from DNA coding sequence differences. Mol. Biol. Evol. **6:** 15-32.

LEWONTIN, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.

MADDISON, W. P., and D. R. MADDISON, 1992 *MacClade: Analysis of Phylogeny and Character Evolution, version 3.0.* Sinauer Associates, Sunderland, MA.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation. Mol. Biol. Evol. **13:** 261–277.

NASSIF, N., and W. ENGELS, 1993 DNA homology requirments for mitotic gap repair in Drosophila. Proc. Natl. Acad. Sci. USA **90:** 1262–1266.

NEWFELD, S. J., R. W. PADGETT, B. RICHTER, S. D. FINLEY, M. DE CUEVA *et al.*, Molecular evolution at the *decapentaplegic* locus in Drosophila. Genetics **145:** 297–309.

RILEY, M. A., M. E. HALLAS and R. C. LEWONTIN, 1989 Distinguishing the forces controlling genetic variation at the *Xdh* locus in *Drosophla pseudoobscura.* Genetics **123:** 359–369.

ST. JOHNSTON, R. D., F. M. HOFFMAN, R. BLACKMAN, D. SEGAL, R. GRIMAILAN *et al.*, 1990 Molecular organization of the *decapentaplegic* gene in *Drosophila melanogaster.* Genes Dev. **4:** 1114–1127.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura.* Genetics **135:** 541–552.

SIMONSON, K. L., G. A. CHURCHILL and C. J. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

SMITH, S. W., R. OVERBEEK, C. R. WOESE, W. GILBERT and P. M. GILLEVET 1994 The genetic data environment, an expandable GUI for multiple sequence analysis. Comp. Appl. Biosci. **10:** 671–675.

SWOFFORD, D. L., 1991 *PAUP: Phylogenetic Analysis Using Parsimony, version 3.1.1.* Illinois Natural History Survey, Champaign.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

VEUILLE, M., and L. M. KING, 1995 Molecular basis of polymorphism at the *Esterase 5-B* locus in *Drosophila pseudoobscura.* Genetics **141:** 255–262.

Communicating editor: A. G. CLARK