Evidence for Selection at the *fused* Locus of *Drosophila virilis*

Jorge Vieira and Brian Charlesworth

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Manuscript received August 12, 1999
Accepted for publication April 7, 2000

ABSTRACT

The genomic DNA sequence of a 2.4-kb region of the X-linked developmental gene *fused* was determined in 15 *Drosophila virilis* strains. One common replacement polymorphism is observed, where a negatively charged aspartic amino acid is replaced by the noncharged amino acid alanine. This replacement variant is located within the serine/threonine kinase domain of the *fused* gene and is present in ~50% of the sequences in our sample. Significant linkage disequilibrium is detected around this replacement site, although the *fused* gene is located in a region of the *D. virilis* X chromosome that seems to experience normal levels of recombination. In a 600-bp region around the replacement site, all eight alanine sequences are identical; of the six aspartic acid sequences, three are also identical. The occurrence of little or no variation within the aspartic acid and alanine haplotypes, coupled with the presence of several differences between them, is very unlikely under the usual equilibrium neutral model. Our results suggest that the *fused* alanine haplotypes have recently increased in frequency in the *D. virilis* population.

DATA on the level and patterns of within-population nucleotide polymorphisms can give information about the action of natural selection on DNA and protein sequences. This action is revealed as deviations from the patterns expected if the variation is selectively neutral (Kimura 1983; Ohta 1992; Kreitman and Akashi 1995). Recently, we surveyed the level of nucleotide variability in a 500-bp region of the *fused* (*fu*) gene of *Drosophila virilis* and suggested that this gene is under selection (Vieira and Charlesworth 1999). The *fu* gene is a segment-polarity gene that encodes a putative serine-threonine kinase (Preat *et al.* 1990), which has been implicated in the *hedgehog* signaling pathway (Ingham 1993). This gene is required maternally for correct patterning in the posterior part of each embryonic metamere, but is also necessary later in development, since zygotic *fu* mutations lead to anomalies of adult cuticular structures and tumorous ovaries (Preat *et al.* 1990). The *fu* gene is composed of two domains, an N-terminal kinase domain and a C-terminal domain that may differentially regulate the *fu* catalytic domain according to cell position within the parasegment (Thérond *et al.* 1996). *fu* is X-linked, and its complete genomic DNA sequence has been previously determined both in *D. melanogaster* (Preat *et al.* 1990) and *D. virilis* (Blanchet-Tournier *et al.* 1995).

To obtain further evidence for selection acting on this gene, we have here analyzed a 2.4-kb region of *fu* in 15 *D. virilis* strains and one *D. lummei* strain; the region includes most of the coding region of *fu*, the four introns of this gene, and a small part of the 5′ flanking region. The level and distribution of variability along the gene suggest that one of the alleles has recently increased in frequency within the *D. virilis* population.

Corresponding author: Jorge Vieira, Institute of Cell, Animal and Population Biology, University of Edinburgh, Ashworth Laboratories, King's Bldgs., W. Mains Rd., Edinburgh EH9 3JT, United Kingdom. E-mail: j.vieira@ed.ac.uk

MATERIALS AND METHODS

The *D. virilis* strains used in this work are all the group A strains listed in Vieira and Charlesworth (1999). We also used strain w158 from Mishima, Japan, which was kindly provided by S. Hayashi. All these strains are from the eastern Paleoarctic and Oriental realms and are thought to represent a sample from a large central population (Vieira and Charlesworth 1999). The remaining seven strains (group B) listed in Vieira and Charlesworth (1999) were not used because they are of very diverse origin (California, Mexico, Argentina, Chile, and Holland), and so it is unclear whether they can be legitimately considered together with the strains from the eastern Paleoarctic and Oriental realms (Vieira and Charlesworth 1999). The *D. lummei* strain is from Kemi, Finland and was kindly provided by J. Aspi.

A 2440-bp DNA fragment, which includes a small region of the 5′ noncoding flanking region of *fu*, most of its coding region, and the four introns of this gene (Figure 1), was amplified using the primers FUF and FU4IR (Table 1) from a single male from each strain. Genomic DNA extraction was performed as described in Vieira and Charlesworth (1999). Standard PCR amplification conditions were 30 cycles of denaturation at 94° for 30 sec, primer annealing for 30 sec at 52°, and primer extension at 72° for 3 min. The PCR products were cloned into the pCR 2.1 vector, using the TA cloning kit (Invitrogen, San Diego). DNA sequencing of both strands was performed with a model 377 DNA sequencing system (Applied Biosystems, Foster City, CA) with the ABI PRISM dye termination cycle-sequencing kit (Perkin-Elmer, Norwalk, CT). The oligonucleotides used in the sequencing reactions are listed in Table 1. Because of nucleotide misincorporations that may have occurred during the amplification process, every nucleotide singleton (*i.e.*, a variant that is found in just one

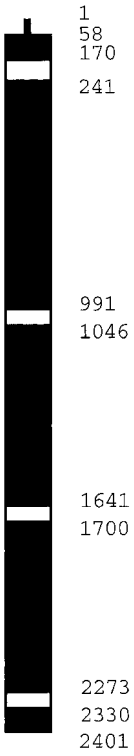

Figure 1.—Schematic organization of the region of the *fu* gene of *D. virilis* analyzed. Solid bars are exons, open bars are introns, and the black line represents the 5′ noncoding sequence. Numbers refer to exon and intron boundaries and are relative to the start of the region analyzed.

sequence in our sample) was checked by sequencing that particular region of the gene directly from the genomic DNA of a single male. DNA sequences are deposited in the GenBank database (*D. virilis* accession nos. AF239851–AF239865 and *D. lummei* accession no. AF239866).

The numbers of synonymous, nonsynonynous, intron, and 5′ flanking region differences between pairs of sequences were calculated using the DnaSP software (Rozas and Rozas 1997). The tests for heterogeneity across a region in the ratio of polymorphism to divergence were calculated using the DNA slider software (McDonald 1998).

TABLE 1

PCR primers used in this study

Forward primers	5′ TGTAAAACGACGGCCAGT 3′	M13F
	5′ AGCGTGTTGTCTGTCATTT 3′	FU457
	5′ TACAGGCAGCACAAACTTC 3′	FU937
	5′ GGGACAGAGCATTGACATA 3′	FU1516
	5′ CCGAACAGCATTTGGTAGC 3′	FU1965
	5′ CACACTGCGGCTTATTGAA 3′	FUF
Reverse primers	5′ ACATTTTGCGGCTTTAGAT 3′	FU523
	5′ GCCACATCGCTCAACTTTA 3′	FU1104
	5′ CTGCTGTGTCTTCCTATCC 3′	FU1594
	5′ CCGCTACCAAATGCTGTTC 3′	FU1967
	5′ GCTCGGCGTTCTCGGGTAG 3′	FU4IR
	5′ GTCGTGACTGGGAAAAC 3′	M13R

	111	111	111	112	222	222	222
	111	**222**	245	667	789	012	444	668	990	000	**011**	123
	339	**013**	316	393	923	515	038	697	255	678	**912**	794
Strain	**238**	**541**	458	776	034	395	179	712	065	598	**954**	241
47	**ATG**	**ATT**	GCG	GTA	CAT	ATG	GTC	ATA	GTC	**GTC**	**CCC**	CAC
31	**...**	**...**	...	.AC	.GC	...	...	GC.	A..	**.CT**	**GTT**	TG.
42	**...**	**..G**	...	...	.G.	...	...	...	...	**.CT**	**GTT**	TG.
skt	**...**	**...**	...	..C	.GC	...	...	GC.	A..	**.CT**	**GTT**	TG.
158	**..A**	**TCG**	CTT	...	GG.	TC.	C..	GC.	A..	**.CT**	**GTT**	TG.
12	**..A**	**TCG**	.T.	C..	GG.	.C.	C..	..T	..T	**...**	**...**	.G.
22	**..A**	**TCG**	...	...	GG.	.CA	...	...	.A.	**ACT**	**GTT**	TG.
9	**CCA**	**TCG**	...	...	GG.	.C.	...	...	...	**.CT**	**GTT**	TGT
s2	**CCA**	**TCG**	...	...	.G.	...	.C.	GC.	A..	**.CT**	**GTT**	...
s9	**CCA**	**TCG**	...	...	.G.	...	.C.	GC.	A..	**.CT**	**GTT**	...
11	**CCA**	**TCG**	...	...	.G.	...	.C.	GC.	A..	**.CT**	**GTT**	...
13	**CCA**	**TCG**	...	...	.G.	.C.	...	...	A..	**.CT**	**GTT**	...
33	**CCA**	**TCG**	...	.A.	.G.	...	.C.	GC.	A..	**.CT**	**GTT**	...
15	**CCA**	**TCG**	...	...	.G.	.C.	..T	...	...	**.CT**	**GTT**	TGT
sbb	**CCA**	**TCG**	...	...	.G.	.C.	..T	...	...	**.CT**	**GTT**	TGT

Figure 2.—*D. virilis* haplotypes. Represented in boldface are the two regions of significant linkage disequilibrium that coincide with two highest polymorphism peaks and the two most common replacement polymorphisms in our sample (the aspartic acid/alanine replacement polymorphism at position 132 and the serine/threonine replacement polymorphism at position 2099; see text for details).

RESULTS

Levels of nucleotide variability: The nucleotide polymorphisms found in our sample of 15 *D. virilis* alleles are shown in Figure 2, and the estimated level of DNA polymorphism is summarized in Table 2. There are 36 segregating sites within the sequenced region of the *D. virilis fu* gene, of which only 5 are replacement sites (3 of which are singletons). The replacement site at position 132 is located in the N-terminal serine/threonine kinase domain of *fu* and is present in ~50% of the sequences. In this case, a negatively charged aspartic amino acid is replaced by the noncharged amino acid alanine. The alanine haplotypes are not confined to any particular geographic region. The 13 alanine haplotypes so far analyzed are from different localities in Japan, Russia, Georgia, Caucasus, England, California, and Argentina, while the 18 aspartic acid haplotypes are from different localities in Japan, China, Georgia, Malta, England, Holland, California, Mexico, and Chile (Vieira and Charlesworth 1999; J. Vieira, unpublished results). This replacement polymorphism does not correspond to a change in any of the nine invariant or five almost invariant residues among all kinases that directly participate in adenosine triphosphate binding and phosphotransfer (Hanks *et al.* 1988).

The replacement site at position 2099 (serine/threonine) is the second most frequent replacement polymorphism in our sample, and it is located in the C-terminal domain of *fu* that may differentially regulate the *fu* catalytic domain (Blanchet-Tournier *et al.* 1995; Thérond *et al.* 1996). This rare variant is observed in only two haplotypes, from China and Japan, with aspartic acid at position 132.

Not surprisingly, the level of nucleotide site diversity, π (Nei 1987), at the *D. virilis fu* nonsynonymous sites is 17 times lower than that estimated for synonymous

TABLE 2

DNA sequence variation summary

	All	5′fl	Nsyn	Syn	Int
S	36	0	5	23	8
π	0.0046 ± 0.0025	0	0.0007 ± 0.0007	0.0135 ± 0.0080	0.0135 ± 0.0063
θ	0.0046 ± 0.0018	0	0.0010 ± 0.0005	0.0145 ± 0.0059	0.0081 ± 0.0040

Fifteen DNA sequences were analyzed. *S* is the number of segregating sites, π (Nei 1987) is the average number of differences per base pair, and θ is Watterson's estimator based on the number of segregating sites (Watterson 1975) at nonsynonymous sites (Nsyn), at synonymous sites (Syn), at intron sites (Int), or at 5′ noncoding flanking sites (5′fl). The standard deviations of π and θ due to stochastic factors, including sampling variance, were calculated according to Nei (1987; pp. 254–258) and Tajima (1993; pp. 37–59) under the assumption of no recombination (Nei 1987). In total, 2401 sites were analyzed (58 noncoding 5′ flanking sites and 1609 nonsynonymous, 488 synonymous, and 246 intron sites, respectively). Note that these values must be corrected to be compared with π values from autosomal genes.

and intron sites. The synonymous site and intron site π values are similar to the average level of intron variation for a sample of six *D. virilis* X-linked genes (1.36%; Vieira and Charlesworth 1999; note that these values must be corrected to be compared with π values from autosomal genes). A similar level of nucleotide variation at synonymous and intron positions is not the general pattern for *D. virilis*. The average level of variation at synonymous sites has been estimated to be approximately half of the level of variation at intron sites (Vieira and Charlesworth 1999). Although large variances are attached to these estimates, this finding suggests that the third codon positions of the *fu* gene are largely unconstrained, *i.e.*, this gene should show low codon bias. It should be noted that for intron sites, our estimate of π is larger than our estimate of θ (Watterson's estimator for $4N_e\mu$, in which μ is the neutral mutation rate and N_e is the effective population size; Watterson 1975), but the difference is not statistically significant according to Tajima's *D* test statistic (Tajima 1989).

An inverse measure of codon bias, the effective number of codons (ENC; Wright 1990), is larger for *fu* (49.64) than for the *D. virilis* average (45.5; McVean and Vieira 1999), indicating that *fu* does not have strong codon usage preferences. In fact, this gene is within the group of the 25% least biased genes in *D. virilis* [data not shown; only the 50 genes listed in McVean and Vieira (1999) were considered].

Recombination parameters and linkage disequilibrium: The *fu* gene is located in a region of the X chromosome that seems to experience normal levels of recombination (Vieira and Charlesworth 1999), and recombination can be detected in our sample of 15 chromosomes (a minimum number of four recombination events were detected between sites 231 and 697, 697 and 736, 1401 and 1667, and 1920 and 2172; Hudson and Kaplan 1985). Estimates of the level of recombination between adjacent sites based on synonymous and intron site variability ($C = 4N_ec$, where c is the recombination frequency per nucleotide site) and of the ratio C/θ (where θ is 0.013) are shown in Table 3. Hudson's estimator of C (Hudson 1987) is based on the variance of the number of base pair differences between DNA sequences, and Hey and Wakeley's estimator (Hey and Wakeley 1997) is based on the number of pairs of sites with incongruent genealogical histories. These two estimators are biased in opposite directions and therefore can be treated as rough bounds for the estimate of the recombination rate. It should be noted that the theory underlying these estimators includes several important assumptions, which if not verified, make unclear the interpretation of the estimate that is obtained.

In our data the nonneutral causes of the observed significant linkage disequilibrium between several pairs of sites (see below), together with an excess of intermediate frequency polymorphisms at intron sites (although not statistically significant), could in principle have an impact on the estimates of *C*. In general, as long as $\pi \approx \theta$, any nonneutral process that increases the variance of the number of base pair differences between DNA sequences will decrease the value of Hudson's estimator of *C*; similarly, any process that reduces incongruency will decrease the value of Hey and Wakeley's estimator. Using computer simulations of the coalescent process with recombination, under the usual neutral scenario

TABLE 3

Recombination rate estimates

Method	Estimated level of recombination between adjacent sites ($4N_ec$)	C/θ
Direct estimate[a]	0.176	13.54
Gamma estimator[b]	0.0096	0.74
C estimator[c]	0.0138	1.06

[a] Vieira and Charlesworth (1999).
[b] Hey and Wakeley (1997).
[c] Hudson (1987).

(Hudson 1990), the probability of obtaining a value as low as the population genetic estimate is <1 in 500, on the hypothesis that the direct estimate is the true value. Therefore, the observed discrepancy between Hudson's and Hey and Wakeley's estimators ($c = 2.5 \times 10^{-9}$ per base pair) and a direct estimate obtained from the comparison of the physical and linkage map ($c = 4.4 \times 10^{-8}$ per base pair; Vieira and Charlesworth 1999) is not wholly unexpected, in light of the evidence for selection (see below). This discrepancy cannot be attributed to the effect of polymorphic inversions that are known to influence observed rates of recombination (Ashburner 1989), because cytological analysis of >4000 *D. virilis* chromosomes from natural populations has shown a consistent chromosome pattern with no aberrations (Hsu 1952). We use Hudson's and Hey and Wakeley's estimators of C as a minimum value for a population genetic estimate of recombination.

Although recombination was detected in our sample of 15 sequences, significant linkage disequilibrium was detected between several pairs of sites (Figure 3), especially in the regions 132–231 and 2079–2124 [significant by Fisher's exact tests at $P < 0.05$, without Bonferroni correction for multiple tests; with the sequential Bonferroni correction for multiple tests (Rice 1989), significant linkage disequilibrium is observed only between sites 132–133 and 1667–1691], but not between the two regions.

Evidence for selection: Balancing selection acting on or near a region can give rise to an apparent pattern of locally reduced recombination, because theory (Strobeck 1983; Hudson and Kaplan 1988; Kaplan *et al.* 1988; Nordborg 1997) suggests that there should be strong linkage disequilibria and a polymorphism peak near a site under balancing selection. In our sample, the two highest synonymous plus intron polymorphism peaks are located around the aspartic acid/alanine replacement site at position 132 and around the serine/threonine replacement site at position 2099, and both coincide with regions of significant linkage disequilibrium (Figure 4). The latter is located in the C-terminal domain of *fu* that may differentially regulate the *fu* catalytic domain (Blanchet-Tournier *et al.* 1995; Thérond *et al.* 1996); variation at this site is observed in only two haplotypes with aspartic acid at position 131–133 (Figure 2). The significant linkage disequilibrium in this region is due to five consecutive synonymous polymorphisms in a 45-bp region that are not shared between the serine and the threonine sequences (in boldface in Figure 2). However, because of the small sample size for the threonine sequences, it is not possible to evaluate whether selection is maintaining this amino acid replacement polymorphism.

The replacement site at position 132 is located in the N-terminal serine/threonine kinase domain of *fu* and is present in ~50% of the sequences. In this case, a negatively charged aspartic amino acid is replaced by the noncharged amino acid alanine. Sequences with the aspartic residue have twice as much variability at synonymous and intron sites and three times as many segregating sites as sequences with alanine (Table 4). Furthermore, of the 25 polymorphisms that are not singletons, 8 are shared between the aspartic acid and alanine haplotypes, 14 are polymorphic only within the aspartic acid sequences, and 3 are polymorphic only within the alanine sequences. Therefore, it is possible that the alanine sequences may have risen in frequency only recently and acquired most of their variability through recombination with the aspartic acid sequences rather than through mutation. If the alanine haplotypes rose in frequency only recently, then the average level of divergence (0.0055) between the aspartic acid and alanine sequences should be similar to the average level of polymorphism for the oldest haplotypes, in this case the aspartic acid haplotypes, as is observed (the value is 0.0056; Table 5).

Because the aspartic acid/alanine replacement is common, it is possible in principle to evaluate if selection is maintaining this replacement polymorphism. However, the power of tests for detecting selection in regions of normal recombination is low when test statistics that assume no recombination are used (Wall 1999). Although the true rate of recombination to which the *fu* gene is exposed is unclear, there is at least some recombination. Therefore, we have included recombination in four statistical tests, using the methods described in detail in Filatov and Charlesworth (1999). When the entire 2.4-kb region is used, Kelly's test (Kelly 1997), which examines regions for excess of linkage disequilibrium compared with that expected under neutrality, is only significant ($P < 0.05$) if the level of recombination between adjacent sites ($C = 4N_e c$) is assumed to be >0.07. This level of recombination is higher than that estimated using both Hudson's estimator and Hey and Wakeley's estimator, but lower than the direct estimate (Table 3). Therefore, on the basis of this test statistic only, the inference of selection acting on this gene would be only tentative. However, two other haplotype tests, the B and Q tests (Wall 1999), based on the proportion of pairs of adjacent segregating sites that are congruent, *i.e.*, have consistent genealogies, are significant ($P < 0.05$) if $C > 0.006$ and 0.008, respectively. Furthermore, Hudson *et al.*'s (1994) haplotype test, which is based on the probability of occurrence of a subset of sequences with S segregating sites in a set of sequences with S_{total} segregating sites, is also significant ($P < 0.05$) if $C > 0.006$. The latter three values are all smaller than any of the recombination estimates (Table 3), and therefore these test statistics support the hypothesis that selection is acting on the *fu* gene.

The strongest deviations from the expected neutral pattern should be detected near the putative site under selection. If only the first 600 bp of sequence are analyzed (this is where the common replacement variant

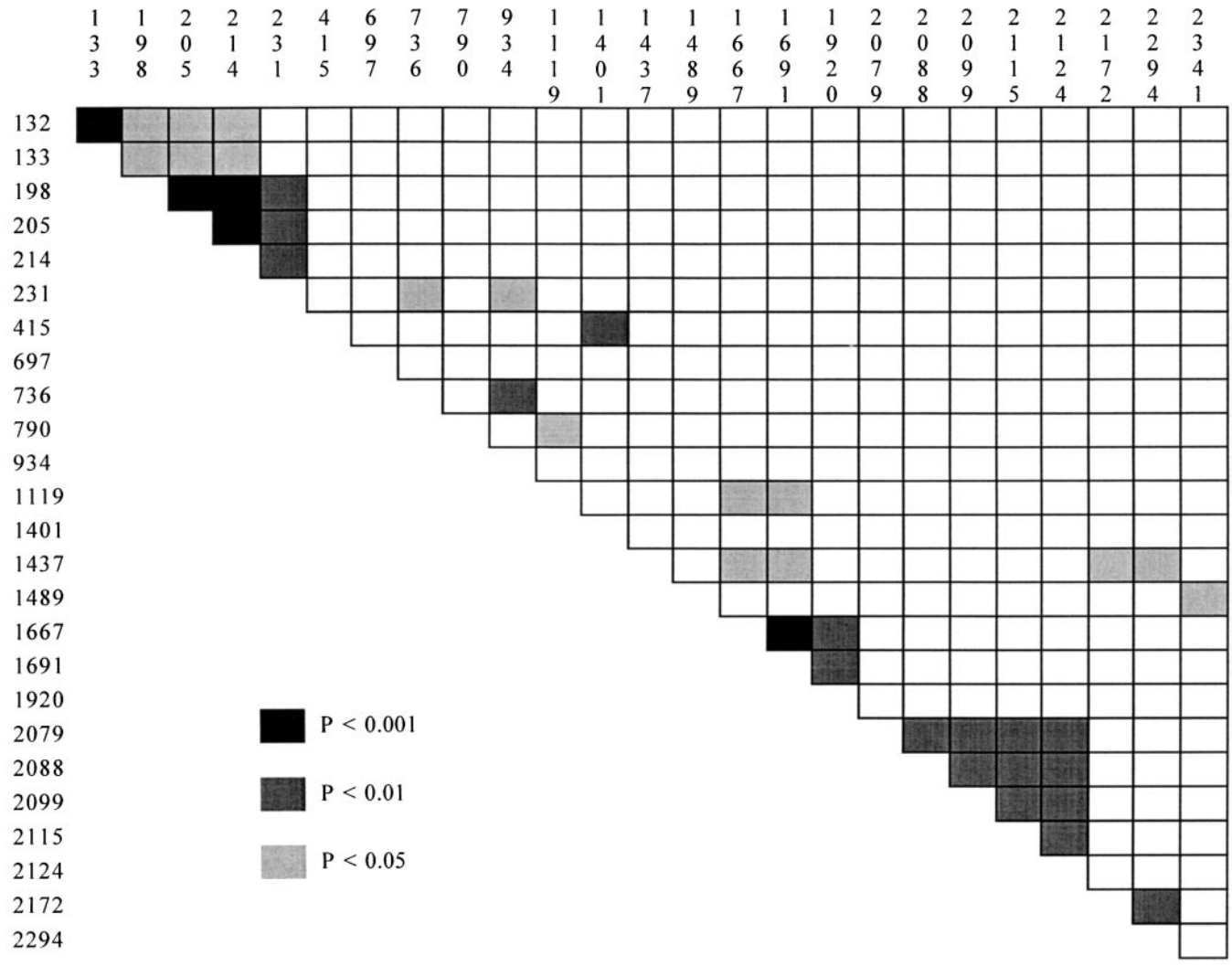

Figure 3.—Linkage disequilibrium among the 26 informative sites in our sample of 15 *fu* sequences. Linkage disequilibrium, significant by Fisher's exact test, without correction for multiple tests, is represented by different shades of gray and black.

is located), the aspartic acid sequences and the alanine sequences do not share any of the five polymorphic sites found in this region. Furthermore, all eight alanine sequences are identical, but of the six aspartic acid sequences, only three are identical. The occurrence of little or no variation within groups of sequences, cou-

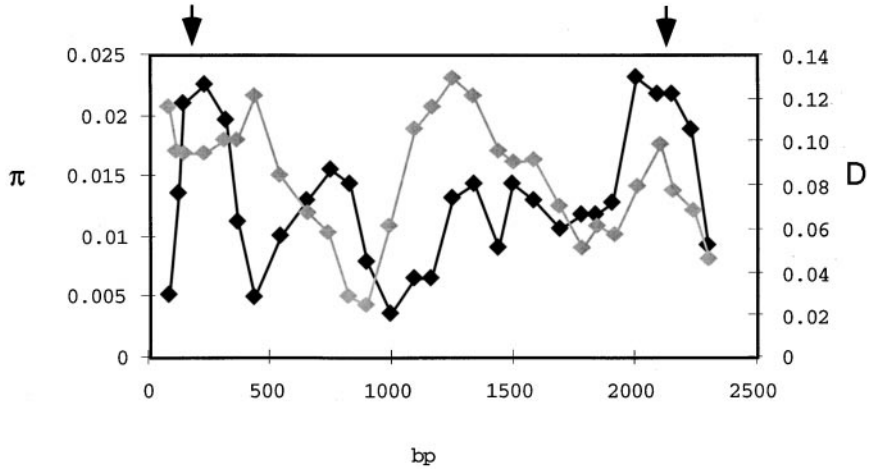

Figure 4.—*D. virilis* synonymous and intron variability (π; black line) and synonymous and intron divergence between *D. lummei* and *D. virilis* (*D*; gray line) along the *fu* gene. The window is 100 synonymous plus intron sites wide; the increment is 25 synonymous plus intron sites. Each diamond represents a sampling point. The arrows indicate the location of the two replacement polymorphisms that are associated with a peak of linkage disequilibrium.

TABLE 4

DNA sequence variation summary for the aspartic acid and alanine sequences

Class	*N*		All	5′fl	Nsyn	Syn	Int
Aspartic acid	7	*S*	31	0	4	19	8
		π	0.0056 ± 0.0031	0	0.0008 ± 0.0007	0.0164 ± 0.0095	0.0163 ± 0.0085
		θ	0.0053 ± 0.0025	0	0.0010 ± 0.0006	0.0159 ± 0.0078	0.0107 ± 0.0059
Alanine	8	*S*	11	0	0	8	3
		π	0.0022 ± 0.0011	0	0	0.0075 ± 0.0040	0.0068 ± 0.0036
		θ	0.0018 ± 0.0009	0	0	0.0063 ± 0.0034	0.0038 ± 0.0026
D			0.0055	0	0.0011	0.0154	0.0128

Definitions as in Table 2. *N* is the number of sequences analyzed. *D* is the uncorrected average number of nucleotide substitutions per site between the aspartic acid and alanine sequences (Nei 1987).

pled with the presence of several differences between them, is very unlikely under the usual equilibrium neutral model, even if no recombination is assumed [$P = 0$ using the Hudson *et al.* (1994) haplotype test; this analysis is equivalent to that of Vieira and Charlesworth (1999)].

Divergence between *D. virilis* and *D. lummei*: We have also determined the DNA sequence of the *fu* gene of one *D. lummei* strain (a close relative of *D. virilis*; Tominaga and Narise 1995; Nurminsky *et al.* 1996), which has aspartic acid at position 132. In Table 5, the average divergence between *D. lummei* and *D. virilis* is presented for the full region analyzed, at the 5′ noncoding flanking sequence, at intron sites, synonymous sites, and nonsynonymous sites. Synonymous sites are diverging as fast as 5′ noncoding flanking sites and almost twice as fast as intron sites. Because divergence is more sensitive than polymorphism to differences in the selection coefficients (Kimura 1983, p. 43), these results suggest that intron sites may be more constrained than synonymous sites, despite the level of polymorphism at synonymous sites and intron sites being similar (Table 2). The possibility that introns are more constrained only in the lineage leading to *D. lummei* is not supported, since in this case, the ratio of synonymous to intron fixed differences between species should be greater than the ratio of synonymous to intron polymorphisms within *D. virilis* (McDonald and Kreitman 1991), in contrast to what is observed. However, it should be noted that the power of the McDonald-Kreitman test is limited with the small number of differences observed here. Introns also diverge less than synonymous sites when orthologous genes of *D. melanogaster* and *D. simulans* are compared (Bauer and Aquadro 1997).

Furthermore, by comparing the level of divergence between *D. lummei* and *D. virilis* to the level of polymorphism in *D. virilis*, it is possible to determine whether the polymorphism peaks observed in the latter species could be attributed to regions of unusually low constraint. These regions seem not to be diverging more than surrounding regions that are less polymorphic and therefore these regions do not seem to represent regions of unusually low constraint (Figure 4). However, we have failed to show that there is significant heterogeneity in the distribution of polymorphic sites relative to fixed differences between *D. lummei* and *D. virilis* using the test statistics described in McDonald (1996, 1998), but the power of these is unknown for our parameters.

DISCUSSION

Overall, the statistical tests presented above suggest that the lack of variability within the haplotypes with the codon for alanine (GCC) at nucleotides 131–133 relative to those with the codon for aspartic acid (GAT) is inconsistent with an equilibrium neutral model and suggest a history of selection at the *fu* locus. There are three main such possibilities that can be imagined. The first is that balancing selection has been maintaining the amino acid polymorphism at this position (or at a closely linked site) for much longer than the standard neutral coalescence time; this is expected to produce a

TABLE 5

Divergence between *D. lummei* and *D. virilis* sequences

	All	5′ fl	Nsyn	Syn	Int
D	0.0307 ± 0.0035	0.1207 ± 0.0427	0.0049 ± 0.0017	0.0934 ± 0.0131	0.0539 ± 0.0141

D is the uncorrected average number of nucleotide substitutions per site between the *D. lummei* and *D. virilis* sequences (Nei 1987). In total, 2401 sites were analyzed (58 noncoding 5′ flanking sites and 1609 nonsynonymous, 488 synonymous, and 246 intron sites, respectively). The sampling standard deviation is given according to Kimura and Ohta (1972).

window of enhanced variability in the neighborhood of the target of selection, as observed here (Strobeck 1983; Hudson and Kaplan 1988; Kaplan *et al.* 1988). But such a process does not lead to a deficiency of expected polymorphism levels within haplotypes carrying one of the two selectively maintained alleles, although there may be some probability that such a deficit is observed in a given sample by virtue of the high stochasticity of structured coalescent processes (Wakeley 1996). The second possibility is that the alanine haplotype has originated recently (in comparison with the timescale of the coalescent process, $2N_e$) and has only recently reached its current frequency. The third possibility is that the polymorphism has been preserved for a long period of time, but that the alanine haplotype was at a low frequency until very recently, when it then rose to its present frequency. The last two models both predict a lower expected diversity for the alanine haplotype than for the aspartic acid haplotype.

The fit of the data to these models was investigated by simulations of a structured coalescent model, in which there are two alleles (A_1 and A_2) that are assumed to be the target of selection. The aspartic acid to alanine substitution is equated with the mutation from A_1 to A_2. Allele A_2 originated at time T in the past, measured in units of $2N_e$ generations, where N_e is the effective size of the population. If T is $\gg 1$, the two haplotypes are old; if $T \ll 1$, A_2 originated recently. The population frequencies of A_1 and A_2 at the time of sampling are p and q; if q is small, A_2 must have been maintained at a low frequency since its time of origination and has only recently increased to its current value of $\sim$0.5.

Variation at a set of neutral sites linked with recombination frequency r to the selected site was modeled by tracing the ancestry of an initial set of genes with n_1 copies that were A_1 in state and n_2 copies that were A_2. No recombination within the set of neutral sites was permitted; this is a conservative assumption in view of the nature of the hypotheses being tested (see below) and greatly simplifies the calculations. Up to time T, the rate per unit coalescence time at which a neutral site that is currently associated with allele A_1 is derived from A_2 is $R_{12} = qR$, where $R = 2N_e r$; the complementary rate for a neutral site associated with A_2 is $R_{21} = pR$ (Hudson and Kaplan 1988; Nordborg 1997). The corresponding rates of coalescence for neutral sites within the two allelic classes are p and q. At time T, all A_2 gene copies are assumed to coalesce instantaneously into a single copy; this forms a single panmictic population together with the A_1 alleles, which then follows the standard coalescent process.

Using the standard assumption that the possible events involved follow competing exponential distributions with the above rate constants (Hudson and Kaplan 1988; Hudson 1990), a single replicate simulation generates a gene tree connecting the alleles initially present in the sample. The number of segregating sites corresponding to the observed sample is then distributed randomly over the different branches of the tree. There are four mutually exclusive categories of such sites: (i) sites that segregate only within allele class A_1, (ii) sites that segregate only within allele class A_2, (iii) sites that segregate in both allele classes (shared sites), and (iv) sites that are fixed between alleles A_1 and A_2. Repeated simulations provide an estimate of the frequencies of these categories for assumed values of the parameters q, T, and R, which define the nature of the model.

Given the fact that this approach assumes no recombination between the neutral sites in question, it is only applicable to a group of closely linked sites. We have therefore applied it to the block of five segregating sites that are closest to the sites 132–133, which are involved in the aspartic acid/alanine substitution. This block spans sites 198–231 and shows no evidence for recombination in the set of 15 sampled *fu* alleles. From Figure 2, it is evident that the vector of the numbers of segregating sites in categories (i) to (iv), respectively, is (5, 0, 0, 0). An overall measure of the goodness of fit of the simulations to the data is thus given by the frequency of replicates in which all five segregating sites are found only among the A_1 alleles.

Figure 5 shows the results of the simulations. A gives the measure of the goodness of fit for the case when A_1 and A_2 have been maintained at equal frequencies ($q = 0.5$), as a function of R for various values of T. Since we have no *a priori* evidence as to whether the aspartic acid or alanine variants are ancestral, the probability values generated by the simulations have been multiplied by two. It is evident that only the cases with a relatively recent origin of the alanine haplotype ($T = 0.1$ or 0.01) have probabilities $\geqslant$1%. It is not possible to distinguish clearly between these two T values, although $T = 0.01$ consistently generates higher probabilities, nor to assign a value to R with much confidence, although the extreme values in Table 3 are both consistent with the data. The hypothesis that the two alleles have been maintained by selection at intermediate frequencies for a period that is as long or longer than the coalescence time is decisively rejected, however.

We have recently shown that the *fu* gene is duplicated in three closely related species of the virilis group (*D. americana*, *D. texana*, and *D. novamexicana*) but, by direct sequencing, there is no evidence for the presence of this duplication in any other species of the virilis group (J. Vieira, unpublished results). Of the nine *D. virilis* polymorphic sites observed in the first 600 bp, including the aspartic acid/alanine replacement polymorphism, seven have been observed either as fixed differences between the two duplicated *fu* genes or as polymorphisms in a small sample of *D. americana* and *D. texana* sequences (J. Vieira, unpublished results). Therefore, most of the polymorphisms observed in *D. virilis* in this region are ancient and must predate the divergence

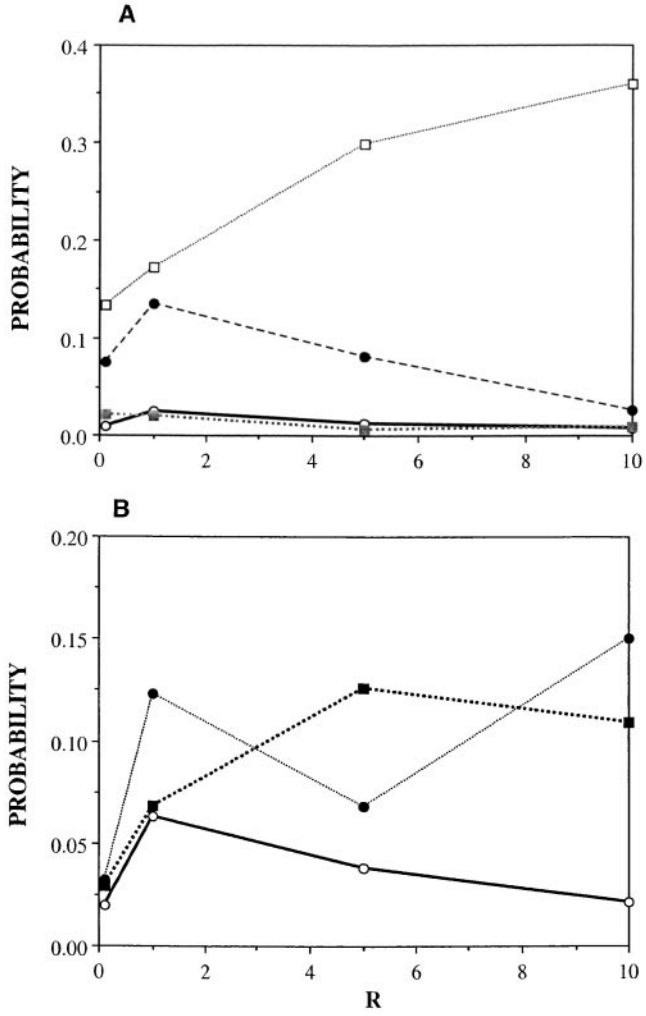

Figure 5.—(A) The frequency for each parameter set with which all five segregating sites are found only among one of the two alleles, among 100,000 replicate simulations. In each case, $p = q = 0.5$; open squares, $T = 0.01$; solid circles, $T = 0.1$; solid squares, $T = 1$; and open circles, $T = 10$. Seven copies of allele A_1 and eight copies of A_2 are assumed to be sampled. (B) The frequency with which all five segregating sites are found only among A_1, for the case when $T = 10$. Open circles, $q = 0.1$; solid squares, $q = 0.01$; and solid circles, $q = 0.001$. Further details are explained in the text.

between *D. americana*/*D. texana*/*D. novamexicana* and *D. virilis*. This observation is clearly incompatible with a relatively recent origin of the *D. virilis* alanine haplotype as required by the above model.

B displays the results of simulations where allele A_2 had a much lower frequency than A_1 up to the time of sampling, consistent with its lower level of variability, assuming $T = 10$. Since prior information is being used, it is legitimate to use the one-tailed probabilities generated by the simulations in this case. Overall, a value of q of 0.001 or 0.01 is consistent with the data at the 5% probability level for all R values considered; $q = 0.1$ is consistent with the data only for R in the neighborhood of one. Clearly, the alanine haplotype must have been kept at low frequency in the lineage leading to *D. virilis*. Because neither the aspartic acid nor the alanine haplotypes are confined to any particular geographic region, this implies that the alanine mutation has recently increased in frequency throughout the *D. virilis* populations. Consistent with this view is the observation that there is no polymorphism in this region in a worldwide *D. virilis* sample of 13 alanine haplotypes (Vieira and Charlesworth 1999; J. Vieira, unpublished results).

Because it seems that the two *fu* gene copies of *D. americana* and *D. texana* may be distinguished by the presence or absence of an aspartic acid/alanine at the same position where the common *D. virilis* replacement polymorphism is found, although the sample size is small (J. Vieira, unpublished results), it is possible that there may be an advantage in having both *fu* haplotypes. We therefore speculate that this may not be a selective sweep caught in midstream, but rather a balanced polymorphism that has experienced a shift in frequency. Our data fit the general observation that old balanced polymorphisms with intermediate allele frequencies seem to be rare in Drosophila. In the past few years, the pattern and level of intraspecific variation at the nucleotide level has been analyzed in detail for seven *D. melanogaster* loci with common allozyme polymorphisms (*Adh*, *6Pgd*, *G6pd*, *Gpdh*, *Sod*, *Est-6*, and *Tpi*; Hasson *et al.* 1998 and references therein). Hasson *et al.* (1998) have recently reviewed the data on these electrophoretic variants and concluded that *Tpi*, *Sod*, *6Pgd*, and one of the two polymorphisms at *G6pd* appear to be associated with recently derived mutations that have reached substantial frequencies in parts of the *D. melanogaster* distributional range, an observation compatible with region-specific directional selection. In the case of polymorphisms at *G6pd*, *Gpdh*, and *Adh*, some elevated differentiation between alleles is observed, but only in the case of *Adh* has the neutral model been rejected in a direction consistent with balancing selection. Our data on *fu*, together with the lack of evidence for selection at five other X-linked loci of *D. virilis* (Vieira and Charlesworth 1999), suggest that this pattern may extend to other classes of genes than enzyme loci and to species other than *D. melanogaster*.

We thank D. Filatov for providing us with his computer program for performing the statistical tests with recombination, M. Przeworski for performing the recombination simulations in relation to Table 3, and D. Charlesworth and C. P. Vieira for helpful comments on the work. J.V. is supported by the Fundação para a Ciencia e Tecnologia (PRAXIS XXI/BPD/14120/97). B.C. is supported by the Royal Society.

LITERATURE CITED

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Blanchet-Tournier, M. F., H. Tricoire, D. Busson and C. Lamour-Isnard, 1995 The segment-polarity gene *fused* is highly conserved in Drosophila. Gene **161:** 157–162.

Bauer, V. L., and C. F. Aquadro, 1997 Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **14:** 1252–1257.

Filatov, D. A., and D. Charlesworth, 1999 DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia *PgiC* locus. Genetics **153:** 1423–1434.

Hanks, S. K., A. M. Quinn and T. Hunter, 1988 The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science **241:** 42–52.

Hasson, E., I. N. Wang, L. W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the triosephosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **15:** 756–769.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Hsu, T. C., 1952 Chromosomal variation and evolution in the virilis group of Drosophila. Univ. Texas Publ. **5204:** 35–72.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

Ingham, P. W., 1993 Localized *hedgehog* activity controls spatial limits of *wingless* transcription in the Drosophila embryo. Nature **366:** 560–562.

Kaplan, N. L., T. Darden and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, United Kingdom.

Kimura, M., and T. Ohta, 1972 On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2:** 87–90.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26:** 403–422.

McDonald, J. H., 1996 Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **13:** 253–260.

McDonald, J. H., 1998 Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377–384.

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

McVean, G. A. T., and J. Vieira, 1999 The evolution of codon preferences in Drosophila: a maximum likelihood approach to parameter estimation and hypothesis testing. J. Mol. Evol. **49:** 63–75.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

Nurminsky, D. I., E. N. Moriyama, E. R. Lozovskaya and D. L. Hartl, 1996 Molecular phylogeny and genome evolution in the *Drosophila virilis* species group: duplications of the *alcohol dehydrogenase* gene. Mol. Biol. Evol. **13:** 132–149.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Preat, T., P. Thérond, C. Lamour-Isnard, B. Limbourg-Bouchon, H. Tricoire *et al.*, 1990 A putative serine/threonine protein kinase encoded by the segment-polarity *fused* gene of Drosophila. Nature **347:** 87–89.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

Rozas, J., and R. Rozas, 1997 DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

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

Tajima, F., 1993 Measurement of DNA polymorphism, pp. 37–60 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

Thérond, P., G. Alves, B. Limbourg-Bouchon, H. Tricoire, E. Guillemet *et al.*, 1996 Functional domains of *fused*, a serine-threonine kinase required for signaling in Drosophila. Genetics **142:** 1181–1198.

Tominaga, H., and S. Narise, 1995 Sequence evolution of the *Gpdh* gene in the *Drosophila virilis* species group. Genetica **96:** 293–302.

Vieira, J., and B. Charlesworth, 1999 X chromosome DNA variation in *Drosophila virilis*. Proc. R. Soc. Lond. B Biol. Sci. **266:** 1905–1912.

Wakeley, J., 1996 The variance of pairwise differences in two populations with migration. Theor. Popul. Biol. **49:** 39–57.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–275.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Communicating editor: A. G. Clark