Evidence for Selection at the *fused1* Locus of *Drosophila americana*

Jorge Vieira,* Bryant F. McAllister† and Brian Charlesworth*

**Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom and* †*Department of Biology, University of Texas, Arlington, Texas 76019-0498*

Manuscript received September 25, 2000
Accepted for publication February 2, 2001

ABSTRACT

We analyze genetic variation at *fused1*, a locus that is close to the centromere of the *X* chromosome-autosome (*X/4*) fusion in *Drosophila americana*. In contrast to other *X*-linked and autosomal genes, for which a lack of population subdivision in *D. americana* has been observed at the DNA level, we find strong haplotype structure associated with the alternative chromosomal arrangements. There are several derived fixed differences at *fused1* (including one amino acid replacement) between two haplotype classes of this locus. From these results, we obtain an estimate of an age of ~0.61 million years for the origin of the two haplotypes of the *fused1* gene. Haplotypes associated with the *X/4* fusion have less DNA sequence variation at *fused1* than haplotypes associated with the ancestral chromosome arrangement. The *X/4* haplotypes also exhibit clinal variation for the allele frequencies of the three most common amino acid replacement polymorphisms, but not for adjacent silent polymorphisms. These patterns of variation are best explained as a result of selection acting on amino acid substitutions, with geographic variation in selection pressures.

WHILE the maintenance of polymorphic chromosomal inversions in Drosophila has received much attention (Krimbas and Powell 1992; Powell 1997), there is only limited evidence on the nature of the evolutionary forces affecting other types of chromosomal arrangements, mainly because these are generally present only as fixed differences between species (Patterson and Stone 1952; Powell 1997). *Drosophila americana americana* and *D. americana texana* are two closely related subspecies of the virilis group of Drosophila (Throckmorton 1982). At the chromosomal level, *D. a. americana* is characterized by a derived fusion of the *X* and fourth chromosomes (Muller's elements A and B, respectively: Muller 1940), whereas *D. a. texana* retains the ancestral state, in which the *X* and fourth chromosomes segregate independently (Hughes 1939; Stalker 1940; Throckmorton 1982). Although previous studies suggested that the karyotype characteristic of *D. a. americana* is at a high frequency throughout the north central to northeastern United States, whereas the karyotype characteristic of *D. a. texana* replaces it abruptly in the south central to southeastern United States (Patterson and Stone 1952; Throckmorton 1982), recent data show that the *X/4* fusion is distributed through a very wide cline along a latitudinal gradient (B. F. McAllister, unpublished results). The two subspecies have also been found to be indistinguishable at the DNA level (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999; McAllister and McVean 2000). These observations suggest that there is considerable gene flow among populations distinguished by the different karyotypic forms of *D. americana*, and that the cline for the *X/4* fusion is maintained by a balance between gene flow and selection on the karyotypes themselves or on associated genes (Barton and Gale 1993).

If this is the case, some genetic differentiation could exist between the different chromosomal arrangements in regions where recombination between the two arrangements is restricted. In *D. melanogaster* the majority of laboratory-induced *X* autosome translocations that are viable and fertile are usually broken in the proximal *X* heterochromatin (Ashburner 1989, p. 566). If the breakpoint of the *X/4* fusion is also in the proximal *X* chromosome heterochromatin, the associated reduction in the amount of pericentric heterochromatin could cause suppression of recombination in the proximal euchromatin of the fusion *X* chromosome, as a result of its greater proximity to the centromere (Yamamoto and Miklos 1978). In addition, heterozygosity for the centric fusion may suppress crossing over between the centromere and proximal loci (Ashburner 1989, pp. 563–564). Although no significant differentiation was found between fusion and nonfusion fourth chromosomes at the *Adh* locus (McAllister and Charlesworth 1999), which is ~1 Mb from the centromeric heterochromatin on chromosome four, these considerations suggest that this might not be true for the base of the *X* chromosome.

In this context, *fused* (*fu*) is a suitable locus for study, since it is located on the *X* chromosome, ~1 Mb away

Corresponding author: Jorge Vieira, Departamento de Genética Molecular, Instituto de Biologia Molecular e Celular, Universidade do Porto, Rua do Campo Alegre 823, Porto 4150-180, Portugal.
E-mail: jbvieira@ibmc.up.pt

from centromeric heterochromatin (see MATERIALS AND METHODS). This gene encodes a serine-threonine kinase (PREAT *et al.* 1990) that has been implicated in the *hedgehog* signaling pathway (INGHAM 1993). This locus is duplicated in *D. americana* and *D. novamexicana*, with two paralogous loci *fu1* and *fu2* (VIEIRA and CHARLESWORTH 2000; J. VIEIRA, unpublished data). Here we show that there are several derived fixed differences at *fu1* (including one amino acid replacement) between the two chromosome arrangements and that several nucleotide variants are in strong linkage disequilibrium with the *X/4* fusion. The extent of divergence between allelic classes of *fu1* suggests an ancient origin of the *X/4* fusion. Haplotypes associated with the *X/4* fusion have less DNA sequence variation at *fu1* than haplotypes associated with the ancestral chromosome arrangement. Furthermore, the *X/4*-associated haplotypes show clinal variation in the allele frequencies of the three most common replacement polymorphisms. The possible causes of these patterns are discussed.

MATERIALS AND METHODS

Collection and analysis of chromosomes: Flies were collected at 10 localities representing a transect through the hybrid zone (MCALLISTER and CHARLESWORTH 1999; B. F. MCALLISTER, unpublished data). The sites, dates of the collections, and abbreviations for these populations are as follows: Niobrara, Nebraska (1997; NN97); Chicago, Illinois (1996; C96); Gary, Indiana (1996; G96); Howell Island, Missouri (1999; HI99); Puxico, Missouri (1999; PM99); Lake Ashbaugh, Arkansas (1999; LA99); Augusta, Arkansas (1999; AA99); Floodgate Park, Arkansas (1999; FP99); Monroe, Louisiana (1997; ML97); and Lone Star, Texas (1997; LP97). Identification of *X/4* chromosomes as fused or unfused was based upon linkage analysis of males and females (B. F. MCALLISTER, unpublished data). Wild-caught males were crossed with the multiply marked *D. virilis* strain V46 (see CHARLESWORTH *et al.* 1997), and restriction fragment length polymorphisms (RFLPs) on the fourth chromosome were identified as being sex linked *vs.* autosomal by examining at least six male and female F_1 progeny. Wild-caught females were crossed with strain V46 of *D. virilis*, and the F_1 male progeny were backcrossed individually to determine the pattern of inheritance (sex linked or autosomal) of the visible *cardinal* mutation on the fourth chromosome. DNA extractions of single wild-caught males, single F_1 males from wild-caught females, and single males from isofemale lines were used as template for amplification of regions for sequencing, and RFLP analyses at the *fu1* locus and were performed as described by MCALLISTER and CHARLESWORTH (1999).

***In situ* hybridization:** This technique was performed as described by VIEIRA *et al.* (1997a), using a 2.4-kb fragment of *fu1*. We have localized the *fu1* gene in *D. americana* and *D. novamexicana* to region 18C of the *X* chromosome, using the *D. novamexicana* photographic polytene chromosome map of VIEIRA *et al.* (1997b) for reference, since these three taxa are homosequential for this region of the *X* chromosome.

DNA sequencing and polymorphism analysis: DNA sequencing of both strands and analyses of DNA polymorphism were performed as described by VIEIRA and CHARLESWORTH (1999, 2000). The *fu1* DNA sequence GenBank accession nos. are AY014407–AY014454.

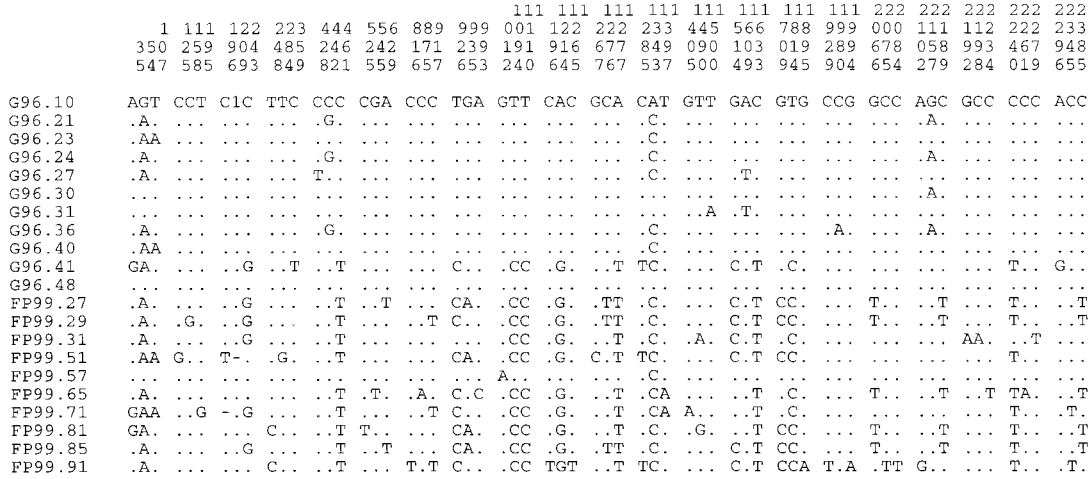
It is desirable to use an approach that allows us to directly determine *fu1* genomic DNA sequences from single males. We used a pair of primers (FUF and FU4IR; see VIEIRA and CHARLESWORTH 2000) that specifically support the amplification of a large fragment from *fu1* and performed a series of seminested PCRs using the primers listed in VIEIRA and CHARLESWORTH (2000) to determine the genomic DNA sequence of *fu1*. There are several fixed differences between *fu1* and *fu2*, one of which creates an additional restriction site for the enzyme *Cac*8I (VIEIRA and CHARLESWORTH 2000). The specificity of the primers FUF and FU4IR for amplifying *fu1* was evaluated by cloning the 2.4-kb amplification product. A set of 90 resulting clones was randomly chosen and digested with *Cac*8I. The same digestion pattern was present in all 90 clones, corresponding to the *fu1* locus, and thus confirming the specificity of the FUF and FU4IR primer pair for this locus.

RESULTS

The organization of DNA sequence variability at *fu1*: Genetic differentiation at the *fu1* locus may exist between the *X/4* fusion chromosome (characteristic of *D. a. americana*) and the unfused karyotype (characteristic of *D. a. texana*), since this gene is located in a region where recombination between the two arrangements is likely to be restricted (see Introduction). To examine this question, we initially studied two populations from the northern and southern regions of the range of *D. americana*, G96 and FP99, representative of *D. a americana* and *D. a. texana*, respectively (THROCKMORTON 1982). The frequencies of the fusion chromosomes in these populations have been estimated to be 96% (G96) and 12% (FP99) (MCALLISTER and CHARLESWORTH 1999; B. F. MCALLISTER, unpublished data). The sequences for a 2.4-kb region of *fu1* sampled from these populations are shown in Figure 1. The *fu1* region analyzed here is the same as in VIEIRA and CHARLESWORTH (2000) and includes most of the coding region of *fu1*, the four introns of this gene, and a small part of the 5′ flanking region.

Visual inspection reveals that each sample contains one sequence that is distinct from other sequences in its sample, but very similar to sequences in the other sample. The G96.41 sequence defines 15 additional segregating sites, representing 62.5% of the segregating sites in the G96 sample; the FP99.57 sequence defines 9 additional segregating sites, representing 16.4% of the segregating sites in the FP99 sample. However, only two additional segregating sites are defined by differences between the G96.41 sequence and FP99 sequences (excluding FP99.57), and one by the difference between the FP99.57 and G96 sequences (excluding G96.41). These two sequences are also distinguished by the variants at position 1633 that are very strongly associated with the nonfusion and fusion chromosomes (see below).

It is, therefore, very likely that sequence G96.41 is from a free *X* chromosome and that sequence FP99.57 is from a fusion *X* chromosome, given the direct evi-



```
                                                  111 111 111 111 111 111 111 111 222 222 222 222 222
                1 111 122 223 444 556 889 999 001 122 222 233 445 566 788 999 000 111 112 222 233
              350 259 904 485 246 242 171 239 191 916 677 849 090 103 019 289 678 058 993 467 948
              547 585 693 849 821 559 657 653 240 645 767 537 500 493 945 904 654 279 284 019 655

G96.10        AGT CCT C1C TTC CCC CGA CCC TGA GTT CAC GCA CAT GTT GAC GTG CCG GCC AGC GCC CCC ACC
G96.21        .A. ... ... ... .G. ... ... ... ... ... ... .C. ... ... ... ... ... .A. ... ... ...
G96.23        .AA ... ... ... ... ... ... ... ... ... ... .C. ... ... ... ... ... ... ... ... ...
G96.24        .A. ... ... ... .G. ... ... ... ... ... ... .C. ... ... ... ... ... .A. ... ... ...
G96.27        .A. ... ... ... T.. ... ... ... ... ... ... .C. ... .T. ... ... ... ... ... ... ...
G96.30        ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .A. ... ... ...
G96.31        ... ... ... ... ... ... ... ... ... ... ... ... ..A .T. ... ... ... ... ... ... ...
G96.36        .A. ... ... ... .G. ... ... ... ... ... ... .C. ... ... ... .A. ... .A. ... ... ...
G96.40        .AA ... ... ... ... ... ... ... ... ... ... .C. ... ... ... ... ... ... ... ... ...
G96.41        GA. ... ..G ..T ..T ... ... C.. .CC .G. ..T TC. ... C.T .C. ... ... ... ... T.. G..
G96.48        ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
FP99.27       .A. ... ..G ... ..T ..T ... CA. .CC .G. .TT .C. ... C.T CC. ... T.. ..T ... T.. ..T
FP99.29       .A. .G. ..G ... ..T ... ..T C.. .CC .G. .TT .C. ... C.T CC. ... T.. ..T ... T.. ..T
FP99.31       .A. ... ..G ... ..T ... ... ... .CC .G. ..T .C. .A. C.T .C. ... ... ... AA. ..T ...
FP99.51       .AA G.. T-. .G. ..T ... ... CA. .CC .G. C.T TC. ... C.T CC. ... ... ... ... T.. ...
FP99.57       ... ... ... ... ... ... ... ... A.. ... ... .C. ... ... ... ... ... ... ... ... ...
FP99.65       .A. ... ... ... ..T .T. .A. C.C .CC .G. ..T .CA ... ..T .C. ... T.. ..T ..T TA. ..T
FP99.71       GAA ..G -.G ... ..T ... ..T C.. .CC .G. ..T .CA A.. ..T .C. ... ... ... ... T.. .T.
FP99.81       GA. ... ... C.. ..T T.. ... CA. .CC .G. ..T .C. .G. ..T CC. ... T.. ..T ... T.. ..T
FP99.85       .A. ... ..G ... ..T ..T ... CA. .CC .G. .TT .C. ... C.T CC. ... T.. ..T ... T.. ..T
FP99.91       .A. ... ... C.. ..T ... T.T C.. .CC TGT ..T TC. ... C.T CCA T.A .TT G.. ... T.. .T.
```

FIGURE 1.—*D. americana* haplotypes in the G96 (Indiana) and FP99 (Arkansas) populations. Dots represent the same nucleotide as in the first sequence, and a dash represents a deletion.

dence for polymorphism of the fusion within these populations (MCALLISTER and CHARLESWORTH 1999; B. F. MCALLISTER, unpublished data), as well as the further evidence presented below. Unfortunately, the relevant strains are not available, thus preventing direct confirmation of this inference. When the G96 and FP99 samples are compared without including G96.41 and FP99.57, seven fixed differences are present between the samples in the 2.4-kb region analyzed, including one amino acid replacement [at position 1633; ACG (Thr)/ATG (Met), respectively]. In addition to the fixed differences between the two samples there are also 51 polymorphisms that are unique to either sample and a single polymorphism that is shared between them. The standard measure of population differentiation, F_{ST}, between G96 and FP99 (excluding G96.41 and FP99.57) is 0.57, as calculated by the method of HUDSON *et al.* (1992a). Highly significant differences between the two samples are detected by the HUDSON *et al.* (1992b) permutation test ($P < 0.001$).

Table 1 shows the estimated levels of nucleotide site diversity for the 2.4-kb region of the *fu1* locus in the G96 and FP99 samples (discarding the two sequences inferred to belong to the minority karyotypes for these populations, as described above). The level of synonymous DNA polymorphism for the FP99 sample (nonfusion karyotype) is ~0.02 per nucleotide site, which is similar to estimates for other genes surveyed in this species (including the *X* chromosomal locus *period*), suggesting an effective population size of $>10^6$ (HILTON and HEY 1996, 1997; MCALLISTER and CHARLESWORTH 1999; MCALLISTER and MCVEAN 2000). However, the estimated level of synonymous DNA polymorphism for the G96 sample (fusion karyotype) is only ~10% of that for FP99.

To determine whether this difference is statistically significant, we generated 10,000 pairs of independent gene trees (HUDSON 1990) with the same population size parameters and randomly distributed the total number of segregating sites between each pair of trees. The

TABLE 1

DNA sequence variation summary for the Gary and Floodgate Park populations

Sample		All (2401)	5′fl (58)	Nsyn (1609)	Syn (488)	Int (246)	Sil (792)
G96	*S*	9	1	5	3	0	4
	π	0.0014	0.0092	0.0011	0.0022	0	0.0021
	θ	0.0013	0.0061	0.0011	0.0022	0	0.0018
FP99	*S*	43	1	6	32	4	37
	π	0.0057	0.0067	0.0011	0.0194	0.0074	0.0152
	θ	0.0068	0.0063	0.0014	0.0241	0.0060	0.0178

Sample sizes are 10 and 9 for the Gary and Floodgate Park populations, respectively. *S* is the number of segregating sites; π (NEI 1987) is the average number of pairwise nucleotide differences per base pair; and θ is Watterson's estimator of $4N_e\mu$ (where N_e is the effective population size and μ the neutral mutation rate) based on the number of segregating sites (WATTERSON 1975) at nonsynonymous sites (nsyn), at synonymous sites (syn), at intron sites (int), at 5′ noncoding flanking sites (5′fl), or at silent sites (sil; 5′fl, syn, and int sites). The number of sites analyzed for each category is shown in parentheses.

frequency of pairs of trees that showed a difference in the numbers of segregating sites within trees as large or larger than that observed was estimated. Using both the total number of segregating sites or only silent segregating sites (synonymous sites, intron, and 5′ flanking region), this difference is significant ($P < 0.05$ and $P < 0.005$, respectively). This is a conservative test of the difference between the samples, since it assumes complete evolutionary independence between the two karyotypes and no recombination within the *fu1* gene. Other loci exhibit normal levels of variation in the G96 sample (McAllister and Charlesworth 1999; McAllister and McVean 2000), indicating that the low diversity at *fu1* is not caused by a recent bottleneck influencing nucleotide diversity in this population.

To examine the generality of this observation, DNA sequences of *D. americana* were also obtained for the HI99, LA99, NN97, and ML97 populations, for a shorter region that corresponds to the first 514 bp of the longer 2.4-kb region analyzed above. Variation at site 1633, which creates a *Cla*I RFLP marker, defines the two major haplotype classes at the *fu1* gene, with haplotypes having C at high frequency in the northern range of *D. americana* and haplotypes with T in the southern range. As shown in the next section, we can use this information to infer the karyotypes of randomly sampled flies with considerable confidence. On average, four to five individuals were sequenced for each putative karyotype and population analyzed. The haplotype structure of these populations is shown in Figure 2.

The estimated levels of nucleotide site polymorphism for the region analyzed here are summarized in Table 2. Although there are large variances associated with these diversity estimates, haplotypes with a C at 1633 are generally less variable than haplotypes with a T. This supports our results on the G96 (*D. a. americana*) and FP99 (*D. a. texana*) populations for a larger 2.4-kb region of the *fu1* gene. The lower variability among northern haplotypes, which are strongly associated with the *X/4* fusion (see below), is thus not limited to the G96 population.

Despite the evidence for considerable gene flow among *D. americana* populations (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999; McAllister and McVean 2000), analysis of the haplotype data in Figure 2 by the Hudson *et al.* (1992b) multiple permutation test (10,000 permutations) shows that there is significant population differentiation for chromosomes carrying the C variant at the *Cla*I 1633 site (and thus putatively the *X/4* fusion), but not for other chromosomes. The significant population differentiation seems to be mainly due to the presence of the haplotype with an amino acid replacement at position 442, for which G is at a high frequency in the northernmost populations (NN and G96), but is absent from the southernmost populations analyzed (HI and LA).

Patterns of geographic variation in karyotypes and DNA sequences: The results presented above demonstrate significant differentiation between two major haplotype classes at the *fu1* locus. The common haplotype class in samples from the northern range of *D. americana* has a lower level of genetic variability and a higher level of between-population differences in DNA sequences than the common haplotype class in the southern range of *D. americana*. The geographic distribution of the *fu1* haplotypes parallels the distribution of the alternative chromosomal arrangements in *D. americana*, so that restriction digestion patterns were used to determine the association between the *fu1* haplotypes and chromosomal arrangement.

We surveyed five nucleotide site polymorphisms that could be identified by the restriction enzymes *Cla*I, *Eae*I, and *Rsa*I in a set of 95 chromosomes from five samples

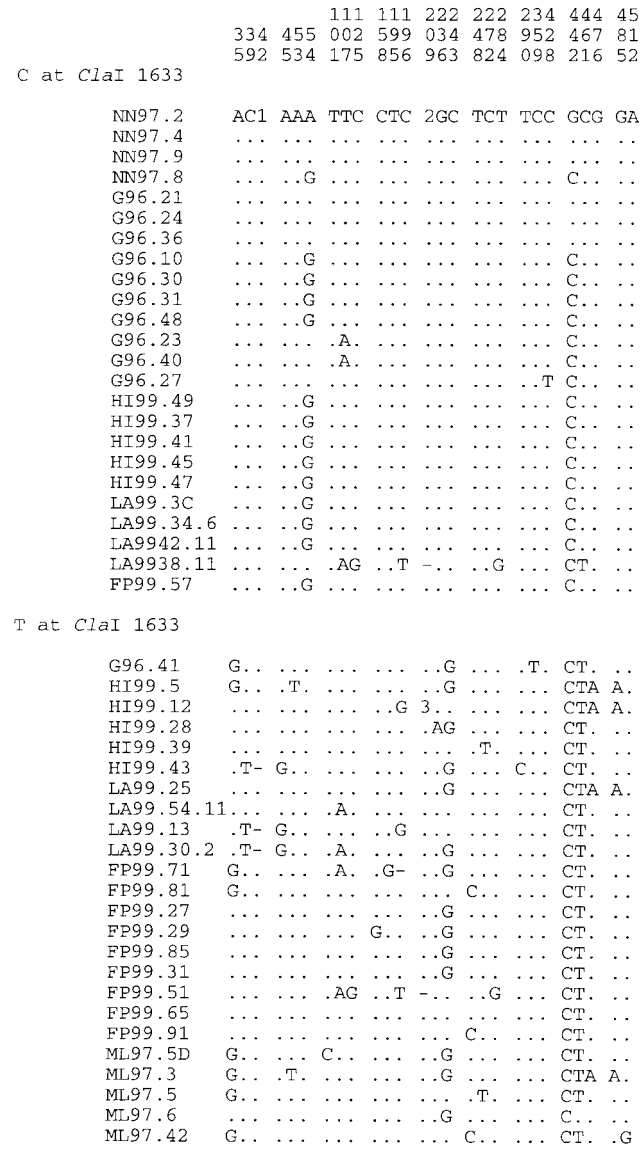


Figure 2.—*D. americana* haplotypes in the NN97 (Nebraska), G96 (Indiana), HI99 (Missouri), LA99 (Arkansas), FP99 (Arkansas), and ML97 (Louisiana) populations. Definitions are as in Figure 1.

TABLE 2

DNA sequence variation summary for chromosomes putatively carrying the *X/4* fusion (C at *Cla*I 1633 site) and unfused chromosomes from different *D. americana* populations

Sample	*N*	*Cla*I 1633		All (482)	Syn (86)	Nsyn (277)	5′ (47)	Int (72)	Sil (205)
NN97	4	C	*S*	2	0	1	1	0	1
			π	0.0021	0	0.0018	0.0106	0	0.0024
			θ	0.0023	0	0.0020	0.0116	0	0.0027
G96	10	C	*S*	4	2	1	1	0	3
			π	0.0032	0.0065	0.0017	0.0114	0	0.0053
			θ	0.0029	0.0082	0.0013	0.0075	0	0.0052
HI99	5	C	*S*	0	0	0	0	0	0
			π	0	0	0	0	0	0
			θ	0	0	0	0	0	0
LA99	4	C	*S*	6	4	0	1	1	6
			π	0.0063	0.0232	0	0.0106	0.0076	0.0151
			θ	0.0069	0.0254	0	0.0116	0.0083	0.0164
Average		C	π	0.0029	0.0074	0.0009	0.0082	0.0019	0.0057
			θ	0.0030	0.0084	0.0008	0.0077	0.0021	0.0061
HI99	5	T	*S*	12	4	0	4	4	12
			π	0.0113	0.0233	0	0.0372	0.0250	0.0269
			θ	0.0120	0.0223	0	0.0447	0.0267	0.0287
LA99	4	T	*S*	7	3	0	2	2	7
			π	0.0087	0.0194	0	0.0310	0.0162	0.0207
			θ	0.0080	0.0190	0	0.0254	0.0152	0.0190
FP99	9	T	*S*	8	3	1	1	3	7
			π	0.0055	0.0097	0.0008	0.0083	0.0180	0.0121
			θ	0.0062	0.0128	0.0013	0.0078	0.0170	0.0130
ML97	5	T	*S*	10	5	1	2	2	9
			π	0.0087	0.0233	0.0014	0.0170	0.0139	0.0186
			θ	0.0099	0.0280	0.0017	0.0204	0.0133	0.0211
Average		T	π	0.0086	0.0189	0.0006	0.0234	0.0183	0.0196
			θ	0.0090	0.0205	0.0008	0.0246	0.0181	0.0204

N is the sample size. Definitions are as in Tables 1 and 3.

representing a 500-km latitudinal transect exhibiting clinal variation for the *X/4* fusion (Table 3). For each of these 95 *X* chromosomes, identification of its status as fused or unfused to the fourth chromosome was performed by linkage analyses (see MATERIALS AND METHODS). The only variable site for which no significant association was found is site *Eae*I 107. The other four polymorphic sites surveyed exhibit significant associa-

TABLE 3

Associations between polymorphic markers at *fu1* and the status of the *X* chromosome of *D. americana*

Restriction enzyme	Site	Polymorphism	Fused	Unfused	*P*[a]
*Eae*I	107	A	5	12	0.098
		T	43	35	
*Rsa*I	1214	A	44	0	0.0001[b]
		G	4	47	
*Cla*I	1633	C	45	0	0.0001[b]
		T	3	47	
*Rsa*I	2157	A	7	0	0.020[b]
		G	41	47	
*Rsa*I	2187	C	45	32	0.003[b]
		T	3	15	

[a] 2×2 χ^2 test with continuity correction.
[b] Associations are significant after the sequential Bonferroni correction.

TABLE 4

Estimates of the *X/4* fusion frequency in *D. americana*

Sample	Latitude/longitude	Indirect estimate (site *Cla*I 1633)	*N*	Direct estimate[a]	*N*
NN97	42° N 40′, 98° W 2′	1	4	—	—
G96	41° N 33′, 87° W 22′	0.94 ± 0.06	16	0.96 ± 0.04	23
HI99	38° N 40′, 90° W 41′	0.72 ± 0.06	53	0.85 ± 0.06	39
PM99	36° N 58′, 90° W 9′	0.61 ± 0.08	38	0.55 ± 0.07	47
LA99	36° N 16′, 90° W 45′	0.42 ± 0.08	36	0.49 ± 0.08	39
AA99	35° N 17′, 91° W 23′	0.50 ± 0.13	16	0.44 ± 0.12	18
FP99	34° N 12′, 91° W 5′	0.16 ± 0.06	38	0.14 ± 0.05	44
ML97	32° N 30′, 92° W 2′	0	5	—	—

N is the sample size. The differences between the direct and indirect estimates of the frequency of the *X/4* fusion are not statistically different (χ^2 test).

[a] McAllister and Charlesworth (1999) and B. F. McAllister (unpublished results).

tions with the state of the centromere. The presence of C at the *Cla*I site 1633, A at the *Rsa*I site 1214, and A at the *Rsa*I site 2157 is always observed for *X/4* fusion chromosomes. The data in Table 3 show that the state of the *X* chromosome would have been erroneously deduced from the *Cla*I site 1633 in only 3 out of 95 chromosomes, *i.e.*, ~3% of the time.

The close association between the *X/4* fusion and the presence of the C variant at the *Cla*I site 1633 implies that it is possible to estimate the frequency of the former in a given population from the frequency of the latter and to examine the patterns of DNA polymorphism associated with each karyotype. For these purposes, a set of 208 chromosomes from single wild-caught males from several populations, and single males obtained from independent females, was surveyed for *Cla*I and eight other *fu1* polymorphic restriction sites (Tables 4 and 5; Figures 3 and 4). Most of these polymorphic sites were known to be present in the G96 sample, and this information was used when choosing the five restriction enzymes (*Eae*I, *Bbs*I, *Ava*I, *Rsa*I, and *Cla*I) used in this survey. This set of chromosomes partially overlaps the set of 95 chromosomes analyzed above, but the genetic information about their fusion status was not used in the following analyses.

No statistically significant differences from χ^2 tests are observed between direct and indirect estimates of fusion frequencies throughout the range of the *X/4* fusion cline (Table 4). There is, however, a 13% discrepancy between the two estimates for the HI99 sample; the indirect estimate is outside the 95% confidence limit for the direct estimate. The direct estimate of the frequency of the *X/4* fusion is significantly correlated both with latitude and longitude (stepwise regression: $N = 6$; $R^2 = 0.99$; $P < 0.001$), while the indirect estimate (based on the frequency of the variant at site 1633) is significantly correlated only with latitude ($N = 8$; $R^2 = 0.93$; $P < 0.001$). This difference may reflect the lack of complete association of site 1633 with karyotype.

The haplotypes for the 208 chromosomes are shown in Table 5, which is arranged so that the haplotypes of chromosomes inferred to be fusion or nonfusion from the state of their *Cla*I 1633 site are in the top and bottom sections, respectively. Figures 3 and 4 show the frequency of RFLP variants and RFLP haplotypes in several *D. americana* populations for individuals that were inferred to carry a *X/4* fusion or nonfusion chromosome, respectively. Only those sites for which the variant is represented more than once in the sample of chromosomes being considered are shown (see Table 5). Visual inspection of Figure 3 suggests that there is clinal geographic variation in nucleotide variant frequencies, not only for the *Ava*I site 442 but also for sites *Bbs*I 1609 and *Rsa*I 2157. These are the three most common replacement polymorphisms present in the G96 sample (see Figure 1).

Stepwise regression analyses reveal significant correlations between latitude and longitude and the variants *Ava*I 442 ($R^2 = 0.92$; $P < 0.005$), *Bbs*I 1609 ($R^2 = 0.88$; $P < 0.01$), *Rsa*I 2157 ($R^2 = 0.95$; $P < 0.001$), haplotype B ($R^2 = 0.90$; $P < 0.005$), and haplotype F ($R^2 = 0.88$; $P < 0.01$). Haplotypes B and F (see Table 5) include the RFLP variants *Ava*I 442, *Bbs*I 1609, and *Rsa*I 2157. In contrast, no obvious clines are found for silent sites or among nonfusion chromosomes.

In addition, for chromosomes carrying the *X/4* fusion, we calculated F_{ST} values between the group of northern populations (NN97, C96, G96, and HI99) and the group of southern populations (PM99, LA99, AA99, and FP99) for each of the polymorphic sites in Table 5 (F_{ST} values are 0.005, 0.146, 0.012, 0.021, 0.029, 0.133, and 0.021 for sites 107, 442, 1012, 1214, 1609, 2157, and 2187, respectively). The three highest F_{ST} values are for the replacement polymorphisms (sites 442, 1609, and 2157). This difference between silent and replacement variants is significant ($P < 0.05$; Mann-Whitney *U*-test), showing that differentiation betweeen the northern and southern populations is mainly due to the replacement variants.

TABLE 5

Haplotypes detected in a survey of 208 *D. americana* chromosomes

	Nucleotide site									Population										
Haplotype	1633	107	249	442[a]	1012	1214	1609[a]	2157[a]	2187	NN97	C96	G96	HI99	PM99	LA99	AA99	FP99	ML97	LP97	Total
A	C	T	T	C	G	A	A	G	C	0	0	5	24	14	10	4	5	0	0	62
B	C	T	T	G	G	A	A	A	C	3	1	4	5	0	0	1	0	0	0	14
C	C	T	T	C	A	A	A	G	C	0	0	1	4	1	3	2	1	0	0	12
D	C	A	T	C	G	A	A	G	C	0	0	2	2	3	1	1	0	0	0	9
E	C	T	T	C	G	A	A	A	C	0	0	1	2	2	1	0	0	0	0	6
F	C	T	T	C	G	A	T	G	C	1	0	2	1	1	0	0	0	0	0	5
G	C	T	T	G	G	A	A	G	T	0	0	0	0	1	0	0	0	0	0	1
H	C	T	T	C	G	G	A	G	C	0	0	0	0	1	0	0	0	0	0	1
									Subtotal:	4	1	15	38	23	15	8	6	0	0	110
I	T	T	T	C	G	G	A	G	C	0	0	1	5	7	12	4	10	4	0	43
J	T	T	T	C	G	G	A	G	T	0	0	0	6	2	2	1	12	0	0	23
K	T	A	T	C	G	G	A	G	C	0	0	0	4	2	4	1	4	0	1	16
L	T	A	T	C	G	G	A	G	T	0	0	0	0	2	2	0	2	0	0	6
M	T	T	C	C	G	G	A	G	C	0	0	0	0	0	0	2	2	1	0	5
N	T	T	C	C	G	G	A	G	T	0	0	0	0	1	0	0	2	0	0	3
O	T	T	T	G	G	G	A	G	T	0	0	0	0	0	1	0	0	0	0	1
P	T	T	T	C	G	G	T	G	C	0	0	0	0	1	0	0	0	0	0	1
									Subtotal:	0	0	1	15	15	21	8	32	5	1	98
									Total:	4	1	16	53	38	36	16	38	5	1	208

[a] Replacement polymorphisms.

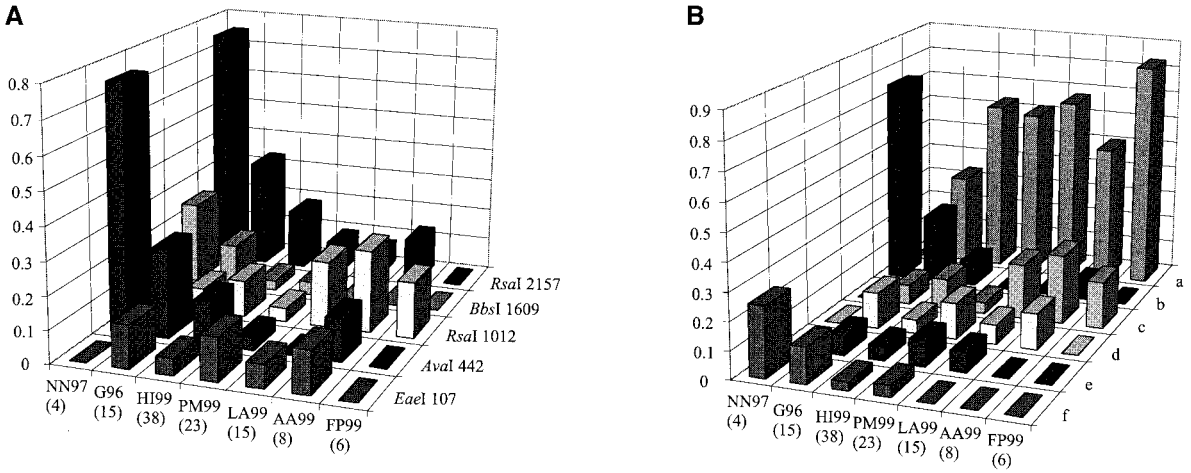


FIGURE 3.—Allele (A) and haplotype (B) frequency among chromosomes carrying the *X/4* fusion. Haplotype codes are as in Table 5.

Evidence for recombination: As noted in the Introduction, the location of *fu1* near the centromeric heterochromatin of the *X* chromosome implies that it may be located in a region of reduced crossing over, and that crossing over between *fu1* and the centromere may be suppressed in heterozygotes for fusion and nonfusion chromosomes (see the Introduction). In this section we summarize the evidence for recombination at *fu1* between unfused and fused chromosomes, within nonfusion chromosomes, and within chromosomes with the *X/4* fusion.

If we assume that the *X/4* fusion was derived through a single mutational event, some recombination (either gene conversion or crossing over) must have occurred between unfused and fused chromosomes, since there are at least two shared polymorphisms, *Eae*I 107 and *Rsa*I 2187 (see Table 3). Furthermore, Table 5 shows that haplotypes G, H, O, and P all have a variant that is present once in the sample (at positions 2187, 1214, 442, and 1609, respectively), but which is commoner among the alternative chromosome arrangement. It is likely, therefore, that these haplotypes are the result of crossing over or gene conversion events between nonfusion and fusion chromosomes.

From the FP99 sequence data on nonfusion chromosomes, a minimum of six recombination events can be inferred to have occurred within the *fu1* gene (HUDSON and KAPLAN 1985), and there are 64 out of 861 (7.4%) pairwise comparisons with all four possible gametic types present. Significant linkage disequilibrium ($P < 0.05$ from χ^2 tests) is detected only between 8 pairs of sites out of 120 pairwise comparisons. The rate of intragenic recombination, $C = (8N_e c)/3$ (where N_e is the effective population size for *X* chromosomal loci and c is the recombination frequency per nucleotide site in females), was estimated from the variance in the number of differences between pairs of DNA sequences (HUDSON 1987). We obtain $C = 0.04$, yielding $C/\theta = 1.7$, where the value of the θ estimator (WATTERSON 1975) of the scaled mutation parameter $4N_e u$ (where u is the mutation rate) is for synonymous sites only. There is also evidence for recombination in the other data sets (data not shown).

From the G96 sequence data on putative *X/4* fusion

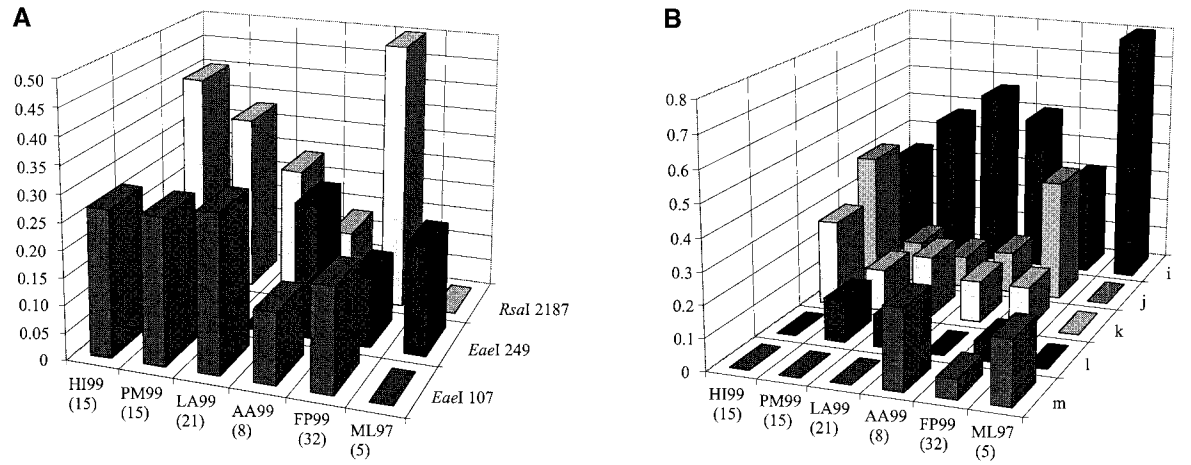


FIGURE 4.—Allele (A) and haplotype (B) frequency among unfused chromosomes. Haplotype codes are as in Table 5.

chromosomes (Figure 1), a minimum of one recombination event is inferred to have occurred (Hudson and Kaplan 1985), and each of the four possible gametic types is found for 4 out of 36 pairwise comparisons (11.1%). The data suggest that the recombination event was between two chromosomes carrying the *X/4* fusion, rather than between a nonfusion and a fusion chromosome, since none of the sequences have any of the many variants typically associated with nonfusion chromosomes. Significant (χ^2 test, $P < 0.05$) linkage disequilibrium is only detected between sites 54 and 1343 and 442 and 2157, out of 15 pairwise comparisons. None of these are significant after a Bonferroni correction, except for the association between sites 54 and 1343. From these data we estimate $C = 0.11$ between adjacent nucleotide sites (the value of C/θ is 50, where θ is for synonymous sites), but it should be noted that this estimate has a large sampling variance. There is only limited evidence for recombination in the other data sets (data not shown).

DISCUSSION

The results described above yield the following main conclusions, which require interpretation in terms of the evolutionary forces affecting the *X/4* fusion and the molecular variants at the *fu1* locus.

1. There are two divergent allelic classes at the *fu1* locus, and these are strongly associated with fusion and nonfusion *X* chromosomes.
2. The fusion chromosomes show much less variability at *fu1* than nonfusion chromosomes.
3. Within fusion chromosomes, there is clinal variation with respect to replacement but not silent site polymorphisms.

Testing for a selective sweep of the fusion chromosome: At first sight, the simplest interpretation of observations 1 and 2 is that the fusion chromosome originated as a single mutation, which rapidly increased in frequency, causing a loss of variability at the *fu1* locus among gametes carrying the fusion because of its close linkage to the centromere; *i.e.*, there has been a selective sweep of the fusion (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 1998). On this model, several neutral polymorphic sites linked to the target of selection must have been swept to a high frequency in the inferred fusion chromosomes in the G96 *D. a. americana* sample shown in Figure 1 and are not found in our relatively small sample of inferred nonfusion chromosomes from *D. a texana* (FP99), resulting in the apparent fixed differences between the fusion and nonfusion chromosomes. But the following argument suggests that this interpretation is unlikely to be correct.

We note first that the inferred derived states of the replacement site 1633, which shows a fixed difference between fusion and nonfusion chromosomes in these data, and all six synonymous fixed sites are associated with the fusion chromosome. This was deduced from comparisons with an outgroup sequence from *D. montana*, which shared a common ancestor with *D. americana* ~10 mya (Tominaga and Narise 1995; Nurminsky *et al.* 1996). The choice of *D. montana* was motivated by the fact that 5% of the shared polymorphisms between a pair of species are expected to be retained until 3.8 N_e generations after their separation (Clark 1997), which could lead to erroneous deductions of the ancestral state if a more closely related species were used.

We can ask if this asymmetry in the distribution of derived fixed variants between the fusion and nonfusion chromosomes can be accounted for by the selective sweep model. If we assume that the level of polymorphism of the ancestral population was the same as in the nonfusion chromosomes of the FP99 sample, and that the selective sweep involved a single randomly chosen *X* chromosome, we can estimate the expected number of variants captured by the fusion chromosomes and which are absent from the sample of nonfusion FP99 chromosomes (see the appendix). The estimated value of θ for all synonymous sites, based on the number of segregating sites in this sample, is 11.76 (Table 1); on this basis, only 1.18 apparent fixed differences are expected between fusion and nonfusion sequences in these samples (the 95% upper bound from the Poisson distribution is 3). Even if we use the highest true value of θ, which generates an estimated value as low as 11.76 with probability 5% (18.6, assuming independence between sites and the resulting Poisson distribution: Ewens 1979, p. 239), only 1.86 fixed differences between the two karyotypes are expected (with an upper 95% bound of 4 from the Poisson distribution). In either case, it is very unlikely that 6 differences would be observed ($P < 0.01$).

A single hitchhiking event cannot, therefore, explain all six derived synonymous fixed differences at *fu1* between the G96 (fusion) and FP99 (nonfusion) sequences. But the data might, in principle, be compatible with a single selective sweep involving the fusion chromosomes, followed by a period of neutral evolution that allows the accumulation of fixed differences, in addition to those associated with the spread of the fusion. The fixation of a small number of variants in association with the sweep would then remove the apparent paradox that all the fixed differences are associated with the fusion chromosomes, and none with the nonfusion chromosomes; for example, four fixations in one lineage and none in the other are not statistically different from the expectation of two fixations in each lineage.

This raises the question of whether other aspects of the data can be reconciled with this possibility. An estimate of the age of the *X/4* fusion from the numbers of fixed synonymous differences between the G96 fusion and FP99 nonfusion chromosomes can be obtained as follows. The above considerations suggest that, at most, three of the six differences are likely to have been associ-

ated with a selective sweep of the *X/4* fusion. A total of 488 synonymous sites were analyzed from the data in Figure 1, giving a synonymous site substitution frequency of at least 6.1×10^{-3} per site. Assuming a neutral mutation rate of 10^{-2}/site/million years (Aquadro *et al.* 1994; Vieira and Charlesworth 1999) and that three substitutions occurred within the fusion chromosomes after the sweep, we obtain an estimate of ~0.61 million years (with a lower 95% limit of ~0.27 million years) for the origin of this chromosome from the ancestral population of nonfused chromosomes. The putative sweep of the fusion is unlikely to be more recent than 0.27 mya.

We can now ask whether the observed low level of variation within the G96 fusion chromosomes is compatible with a selective sweep that occurred 0.27 mya. This can be done very simply by modifying a standard coalescent simulation (Hudson 1990), truncating it so that all surviving alleles coalesce into a single common ancestor at the estimated time of the sweep. The data discussed above suggest that a θ value of 0.02 per site is appropriate for *X*-linked synonymous sites in *D. americana*; this is twice the expected number of mutations per unit of coalescent time ($2N_e$ generations). Using the above mutation rate estimate of 2×10^{-8} per year, this yields an estimate of the coalescent time of 0.89 million years. The minimum time of the sweep of the *X/4* fusion estimated above (0.27 million years) is thus ~0.3 units of coalescent time. By placing mutations onto simulated trees truncated at this point, from a Poisson distribution with mean of 0.5θ in units of coalescent time, we can determine the expected distribution of the number of segregating sites. With the 488 synonymous sites surveyed at the *fu1* locus, the per locus mutation rate in units of coalescent time is 4.88. Three synonymous polymorphisms were observed among the 10 G96 fusion chromosomes. Among 10,000 replicate simulations with these parameters, only 2.6% have as few or fewer segregating sites as observed. Given the conservative assumptions involved in this calculation, this effectively rules out a single selective sweep of the *X/4* fusion as an explanation of the data, if it is assumed that the postsweep effective size of the fusion chromosomes is similar to that of the nonfusion chromosomes.

Other hypotheses: These results are, however, compatible with the possibility that the fusion chromosomes have persisted for a long time at a low effective population size, either because of a restricted geographical distribution or because subsequent hitchhiking events took place within the fusion chromosomes, due to the spread of alleles that were favored only in the genetic background or geographical location of the fusion chromosomes. Such hitchhiking events would not change the mean substitution rate of neutral alleles (Birky and Walsh 1988) and thus would not affect the above estimate of the age of the *X/4* fusion.

There are, in fact, features of the data that suggest that selection on amino acid variants has been operating at *fu1*, which may be relevant to the last hypothesis. In the first place, we note that, although most *fu1* variants are strongly associated with the state of the *X/4* centromere, the associations are not perfect (Table 3), and that (as described above) there is direct evidence from the data in Figure 2 for recombination between sequences derived from the two chromosome types. In particular, some fusion chromosomes carry variants derived from the nonfusion chromosomes. This weakens the case for a long period of purely neutral evolution with complete isolation between the two chromosomal arrangements.

In addition, all fusion chromosomes surveyed carry a mutation of methionine to a derived threonine at the *Cla*I site 1633 of *fu1*, whereas the unfused chromosomes mostly carry the ancestral state. It is possible that this amino acid replacement is advantageous in the "*americana*" background or in the ecological conditions prevailing in more northerly areas, so that a selective sweep associated with it may have contributed to the reduced variability at *fu1* among the fusion chromosomes. Selection maintaining this difference in amino acid sequence would reduce effective gene flow between arrangements and hence elevate divergence at linked silent sites (Charlesworth *et al.* 1997). This hypothesis can be tested by examining patterns of DNA sequence variability in the neighborhood of the *fu1* locus and by testing for any evidence of increasing levels of within-fusion chromosome variability and reduced silent site divergence between arrangements as the distance from *fu1* increases.

There is also a striking difference between the fusion and nonfusion chromosomes in the level of replacement *vs.* silent polymorphism. The results shown in Table 1 indicate 4 replacement and 5 silent polymorphisms within fusion chromosomes, and 6 replacement and 37 silent polymorphisms within nonfusion chromosomes; this pattern is significant at the 2% level on a 2×2 χ^2 test. While it is possible that this could be explained by reduced effective size of the fusion chromosomes, leading to relaxed selection against replacement mutations in the population carrying fusions, such a reduction would be expected to have a bigger effect on the ratio of replacement to silent changes between arrangements compared to the ratio for within-arrangement polymorphisms (Charlesworth 1994), contrary to what is observed (one replacement and six silent changes). The pattern is, therefore, indicative of selection maintaining the amino acid site variants.

Furthermore, among chromosomes with the *X/4* fusion, there are significant correlations between latitude and longitude and the frequencies of the three most common amino acid polymorphisms (at positions 442, 1609, and 2157), as well as for two haplotypes that include these (haplotypes B and F; see Table 5). All three replacement variants are derived and are likely to be younger than the *X/4* fusion since they are common only in chromosomes with the *X/4* fusion. In contrast,

there is no evidence for clinal patterns for silent variants within the fusion chromosomes, and the replacement variants show significantly higher divergence as measured by F_{ST}. This suggests that these apparent clines are the result of differential selection pressures in different parts of the species range, in combination with limited gene flow (Hedrick 2000, p. 301). If so, levels of silent site variability in *D. a. americana* could have been affected by the spread of gametes with these replacement polymorphisms. Among chromosomes with the *X/4* fusion, it is interesting to note that the southernmost population studied (LA99; Table 2) is the most variable at silent sites and that there is little variability in the other three northern populations. It is, therefore, not clear whether the low variability levels found mostly in the northernmost *americana* populations require any additional explanation. More detailed investigations of the geographic distribution of variants at *fu1* and its surrounding region should help to test these possibilities.

Conclusions: Overall, the *fu1* data are consistent with the hypothesis of at least one selective sweep in the *americana* lineage. There are two likely targets of selection for such a selective sweep. First, the presence of a wide frequency cline for the *X/4* fusion (B. F. McAllister, unpublished data) suggests that weak selection is maintaining it (Barton and Gale 1993) and raises the possibility that directional selection may have been responsible for the initial increase in frequency of the *X/4* fusion. But an ancillary hypothesis of a prolonged period of reduced effective population size for the fusion chromosomes is necessary to explain the *fu1* data fully on this basis. The *X/4* fusion is only likely to have influenced patterns of variation at *fu1* if the base of the *X* chromosome is a region with low levels of crossing over, at least for heterozygotes for fusion and nonfusion chromosomes. A direct estimate of this parameter is therefore essential to understand the forces that have shaped levels and patterns of variability at *fu1*. As an alternative, we suggest that a selectively favored replacement substitution at the *Cla*I site 1633 of *fu1* within fusion chromosomes may have occurred, influencing levels and patterns of variability at *fu1*. Whether this putative event could have also influenced the increase in frequency of the *X/4* fusion in the northerly areas depends on the level of crossing over between *fu1* and the centromere. Other evidence suggests the operation of local selection preserving replacement polymorphisms within the fusion chromosomes.

We thank B. Golding, C. P. Vieira, and an anonymous reviewer for helpful comments on this work. J.V. is supported by the Fundação para a Ciencia e Tecnologia (PRAXIS XXI/BPD/14120/97) and B.C. by the Royal Society. This work was partially supported by the Office of Research and Graduate Studies at the University of Texas at Arlington.

LITERATURE CITED

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and DNA polymorphism in Drosophila, pp. 46–56 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall: London.

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

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Barton, N. H., and K. S. Gale, 1993 Genetic analysis of hybrid zones, pp. 13–45 in *Hybrid Zones and the Evolutionary Process*, edited by R. G. Harrison. Oxford University Press, Oxford.

Birky, C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 6414–6418.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–227.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155–174.

Clark, A. G., 1997 Neutral behavior of shared polymorphism. Proc. Natl. Acad. Sci. USA **94:** 7730–7734.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Hedrick, P. W., 2000 *Genetics of Populations*. Jones & Bartlett, Boston.

Hilton, H., and J. Hey, 1996 DNA sequence variation at the *period* locus reveals the history of species and speciation events in the *Drosophila virilis* group. Genetics **144:** 1015–1025.

Hilton, H., and J. Hey, 1997 A multilocus view of speciation in the *Drosophila virilis* species group reveals complex histories and taxonomic conflicts. Genet. Res. **70:** 185–194.

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., M. Slatkin and W. P. Maddison, 1992a Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Hudson, R. R., D. D. Boos and N. L. Kaplan, 1992b A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

Hughes, R. D., 1939 An analysis of the chromosomes of the two sub-species *Drosophila virilis virilis* and *Drosophila virilis americana*. Genetics **24:** 811–834.

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

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitch-hiking effect" revisited. Genetics **123:** 887–899.

Krimbas, C. B., and J. R. Powell, 1992 *Drosophila Inversion Polymorphism*. CRC Press: Boca Raton, FL.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

McAllister, B. F., and B. Charlesworth, 1999 Reduced sequence variability on the Neo-*Y* chromosome of *Drosophila americana americana*. Genetics **153:** 221–233.

McAllister, B. F., and G. A. McVean, 2000 Neutral evolution of the sex-determining gene *transformer* in Drosophila. Genetics **154:** 1711–1720.

Muller, H. J., 1940 Bearings of the Drosophila work on systematics, pp. 185–268 in *New Systematics*, edited by J. Huxley. Clarendon Press, Oxford.

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

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.

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

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, Oxford.

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.

Stalker, H. D., 1940 Chromosome homologies in two sub-species of *Drosophila virilis*. Genetics **26:** 575–578.

Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism—analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Throckmorton, L. H., 1982 The virilis species group, pp. 227–296 in *The Genetics and Biology of Drosophila*, Vol. 3b, edited by M. Ashburner, H. L. Carson and J. N. Thompson, Jr. Academic Press, New York.

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. Ser. B **266:** 1905–1912.

Vieira, J., and B. Charlesworth, 2000 Evidence for selection at the *fused* locus of *Drosophila virilis*. Genetics **155:** 1701–1709.

Vieira, J., C. P. Vieira, D. L. Hartl and E. R. Lozovskaya, 1997a A framework physical map of *Drosophila virilis* based on P1 clones: applications in genome evolution. Chromosoma **106:** 99–107.

Vieira, J., C. P. Vieira, D. L. Hartl and E. R. Lozovskaya, 1997b Discordant rates of chromosome evolution in the *Drosophila virilis* species group. Genetics **147:** 223–230.

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

Yamamoto, M., and G. L. Miklos, 1978 Genetic studies on heterochromatin in *Drosophila melanogaster* and their implications for the functions of satellite DNA. Chromosoma **66:** 71–98.

Communicating editor: G. B. Golding

APPENDIX

There is probability x that a randomly chosen gamete from a population contains a variant that is present in the population at frequency x. Thus the probability that a site present at frequency x is absent from a sample of nonfusion chromosomes of size n, giving rise to an apparent fixed difference between the fusion and nonfusion chromosomes, is $(1 - x)^n$, in the absence of exchange at *fu1* between the two chromosomal arrangements. In a population at equilibrium under the infinite sites neutral model, the expected number of sites with a derived variant at frequency x is $\theta\, x^{-1}$ (Ewens 1979, p. 238), where θ is the product of $4N_e$ (N_e is the effective population size for *X* chromosomal loci) and the neutral mutation rate u for the region in question. Assuming independence between sites, as is justified by the low level of linkage disequilibrium detected in the FP99 sample (see results), if N is the size of the ancestral population, the expected number of variants captured by the fusion chromosome that are absent from the *D. a. texana* sample is

$$\theta \int_{2/3N}^{1-2/3N} (1 - x)^n dx \approx \frac{\theta}{(n + 1)} \quad \text{for large } N.$$