DOI: 10.1534/genetics.104.033456

DNA Sequence Polymorphism and Divergence at the *erect wing* and *suppressor of sable* Loci of *Drosophila melanogaster* and *D. simulans*

John M. Braverman,*,†,‡,[1] Brian P. Lazzaro,[§,‡] Montserrat Aguadé[†] and Charles H. Langley[‡]

**Department of Biology, Georgetown University, Washington, DC 20057-1229, [§]Department of Entomology, Cornell University, Ithaca, New York 14853, [†]Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08071 Barcelona, Spain and [‡]Center for Population Biology and Section of Evolution and Ecology, University of California, Davis, California 95616*

Manuscript received August 2, 2004
Accepted for publication April 5, 2005

ABSTRACT

Several evolutionary models of linked selection (*e.g.*, genetic hitchhiking, background selection, and random environment) predict a reduction in polymorphism relative to divergence in genomic regions where the rate of crossing over per physical distance is restricted. We tested this prediction near the telomere of the *Drosophila melanogaster* and *D. simulans X* chromosome at two loci, *erect wing* (*ewg*) and *suppressor of sable* [*su(s)*]. Consistent with this prediction, polymorphism is reduced at both loci, while divergence is normal. The reduction is greater at *ewg*, the more distal of the two regions. Two models can be discriminated by comparing the observed site frequency spectra with those predicted by the models. The hitchhiking model predicts a skew toward rare variants in a sample, while the spectra under the background-selection model are similar to those of the neutral model of molecular evolution. Statistical tests of the fit to the predictions of these models require many sampled alleles and segregating sites. Thus we used SSCP and stratified DNA sequencing to cover a large number of randomly sampled alleles (~50) from each of three populations. The result is a clear trend toward negative values of Tajima's *D*, indicating an excess of rare variants at *ewg*, the more distal of the two loci. One fixed difference among the populations and high F_{ST} values indicate strong population subdivision among the three populations at *ewg*. These results indicate genetic hitchhiking at *ewg*, in particular, geographically localized hitchhiking events within Africa. The reduction of polymorphism at *su(s)* combined with the excess of high-frequency variants in *D. simulans* is inconsistent with the hitchhiking and background-selection models.

SEVERAL evolutionary models of linked selection have been proposed to explain the patterns of DNA sequence variation observed in natural populations. Genetic hitchhiking is a model of strong directional selection in which the fixation of favorable variants removes linked neutral variation (Maynard Smith and Haigh 1974). This hitchhiking effect is expected to be strongest in genomic regions where crossing over is restricted per physical distance (Kaplan *et al.* 1989). The background-selection model also predicts a reduction in polymorphism that is due to what essentially amounts to a decrease in effective population size, caused by selection's removal of linked deleterious mutants (Charlesworth *et al.* 1993). Neither model predicts a reduction in interspecific divergence. A chief difference between the models is whether a skew toward rare polymorphisms is expected; the hitchhiking model predicts such a skew (Aguadé *et al.* 1989; Braverman *et al.* 1995), while such a skew is not expected in a practically sized sample of sequences under background selection (Hudson and Kaplan 1994; Charlesworth *et al.* 1995). The pseudo-hitchhiking model also yields reduced polymorphism and a skew in the frequency spectrum in regions of restricted recombination (Gillespie 2000). Finally, random-environment models involving linked selection can also produce reduction in polymorphism (relative to divergence) and a skew toward rare variants (Gillespie 1997). All these models of linked selection predict that the effect(s) on selectively neutral polymorphism will be most apparent in regions of the lowest crossing over.

The distal tip of the *X* chromosome of *Drosophila melanogaster* (and its close relatives) offers an excellent opportunity to test models of linked selection, since the rate of crossing over per physical distance decreases to zero at the gene-rich euchromatic region at the telomere. For example, Aguadé *et al.* (1989) found a reduction in polymorphism using RFLP in the *yellow-achaete-scute* (*y-ac-sc*) region of *D. melanogaster*. Begun and Aquadro (1991) and Martín-Campos *et al.* (1992) studied *y-ac* using greater sample sizes from additional geographic locations and extended the investigation to

[1]*Corresponding author:* Department of Biology, Georgetown University, 3700 O St. NW, Washington, DC 20057-1229.
E-mail: jmb24@georgetown.edu

the sister species *D. simulans*. All three studies found a reduction in polymorphism in both species and an excess of rare variants. When the site frequency spectrum was quantified with Tajima's *D* (Tajima 1989), observed values were negative, indicating a skew toward rare variants, although not always significantly so. Divergence data from *D. melanogaster* and *D. simulans* permitted a test of the neutral prediction that polymorphism and divergence are correlated; the levels of divergence observed were normal, thus ruling out a reduction in the neutral mutation rate or the exclusive action of genetic drift as an explanation for the data from these regions. Hence genetic hitchhiking appeared to explain the data from these studies.

Additional work on the *X* telomere extended the surveys to samples from Africa (Begun and Aquadro 1993, 1995b). Polymorphism was reduced in the telomeric genes *y* and *ac*. The levels of polymorphism were higher in Africa than on other continents, and population subdivision between African and non-African populations was detected. These results supported the theory that *D. melanogaster* originated in sub-Saharan Africa and migrated to Europe and North America (David and Capy 1988; Lachaise *et al.* 1988). *D. simulans* is thought to have a similar history. Thus demographic phenomena and/or local adaptation affect genetic variation in *D. melanogaster*, not unlike what was already known in *D. ananassae* (Stephan and Mitchell 1992). Yet sample sizes were generally limited and Tajima's *D* was not statistically different from zero, raising questions about statistical power and the applicability of the hitchhiking model.

More recent surveys of genes near the *X* chromosome's telomere consider regions with intermediate levels of crossing over and larger samples sizes. The studies of Aguadé *et al.* (1994) and Langley *et al.* (2000) investigated two loci, *suppressor of sable* [*su(s)*] and *suppressor of white apricot* [*su(w^a)*], which are (centromere) proximal to *y-ac-sc*. Crossing over is still reduced at these loci, but less so than at *y-ac-sc*. These authors found that the hitchhiking model could explain their data, according to the reduction in polymorphism, and a general trend of the skew in the site frequency spectrum toward rare variants, but again Tajima's *D* was not always significantly negative. In the North American sample, *D* was large and positive. Simulation analysis of the data found a better fit between that data and the hitchhiking model than between that data and the background-selection model, but neither model fit well. Further work is needed to examine these questions in a genomic region with even lower recombination using the same or similar samples. In such regions of extremely low crossing over, the impacts of both the hitchhiking and the background-selection models should be greater. The expected further reduction in polymorphism also means fewer segregating sites per base pair with which to evaluate the frequency spectrum, which thus requires greater survey effort.

One of the goals of the present study is to increase the statistical power of the tests for neutrality, such as Tajima's *D*, by using large sample sizes. We surveyed ~50 lines per population to find additional variation, especially rare variations. An additional reason for our generous sample sizes is to make informative comparisons among different Drosophila populations. We sampled from three continents, Africa, Europe, and North America.

Another goal of this article is to use interspecific divergence to gain insight into the evolutionary forces at work. Thus we surveyed both *D. melanogaster* and its sister species *D. simulans*. A normal level of divergence, for example, would rule out a low local neutral mutation rate and/or mutagenic recombination in regions of normal crossing over. In addition, we can test the generality of the phenomena by comparing data from the same genes experiencing similar but not identical genetic and population conditions in more than one species. Although the rate of crossing over per physical distance is restricted at the telomere of both species, crossing over in *D. simulans* is thought to increase faster when moving away from the tip (Sturtevant *et al.* 1929). Also, the effective population size may differ between these two species. The greater heterozygosity, greater codon bias, and fewer nonsynonymous polymorphisms observed in *D. simulans* has been interpreted as evidence that *D. simulans* has a larger population size than *D. melanogaster* (Aquadro 1992; Moriyama and Powell 1996; Irvin *et al.* 1998).

We surveyed two genes located near the telomeres of *D. melanogaster* and *D. simulans*. The gene *erect wing* (*ewg*) codes for a transcription factor and is located at polytene chromosome band position 1A1 (Koushika *et al.* 2000; Drysdale *et al.* 2005), distal to *yellow*. In this first region, excluding insertion-deletions (indels), we surveyed 3166 bp in *D. melanogaster* and 3193 bp in *D. simulans*. The gene *su(s)* encodes an RNA-binding protein and is located at position 1B13 (Geyer *et al.* 1991; Voelker *et al.* 1991; Drysdale *et al.* 2005). In this second region, excluding indels, we surveyed 2832 bp in *D. simulans*. The two loci are separated by ~360 kb. Our *D. simulans su(s)* data complement a previously published survey of the *su(s)* region in *D. melanogaster* (Langley *et al.* 2000).

Our results can be summarized as follows. First, the *ewg* region has an extreme reduction in polymorphism and a negative Tajima's *D* in both *D. melanogaster* and *D. simulans*, which is consistent with the hitchhiking model. Second, the pattern of variation across populations of *D. melanogaster* could be the result of geographically localized hitchhiking events, similar to what has been found in *D. ananassae* (Stephan and Mitchell 1992; Stephan *et al.* 1998; Baines *et al.* 2004) and in other regions of the *D. melanogaster X* telomere (Begun and Aquadro 1993). Third, variation at *su(s)* is reduced in *D. simulans*, but Tajima's *D* is positive; neither the

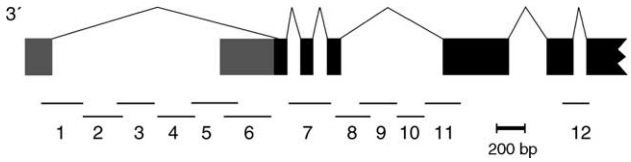

FIGURE 1.—The SSCP fragments of *ewg*, shown as small horizontal lines below the gene. Only part of the entire gene is shown, and it is oriented with the 3′-end on the left in contrast to the standard orientation to illustrate the fragment positions relative to the physical location of the *su(s)* fragments. The solid boxes are exons, the shaded boxes are alternatively spliced exons, and the thin lines connecting the solid boxes are the introns. The scale is indicated with a bar 200 nucleotides long.

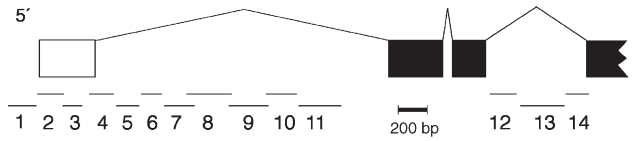

FIGURE 2.—The SSCP fragments of *su(s)*. These are similar but not identical to fragments in LANGLEY *et al.* (2000). The open box is the 5′-UTR. See the Figure 1 legend for more information.

hitchhiking model nor the background-selection model can explain results at that gene in that species.

MATERIALS AND METHODS

Samples: *D. melanogaster* flies were obtained from the following sites: North America (Raleigh, NC; same collection and extraction as for MIYASHITA *et al.* 1993), Europe [14 from the Canary Islands, Spain, 17 from Groningen, Holland, and 21 from Requena, Spain; same collection and extraction as MARTÍN-CAMPOS *et al.* 1992 (see their Figure 1)], and Africa (collected in September 1990 in the Sengwa Wildlife Preserve, Zimbabwe; same collection as BEGUN and AQUADRO 1993). The following collections of *D. simulans* were studied: North America (25 collected in September 1995 from the Noble Apple Orchard, Paradise, CA, and 25 collected in July 1995 from the Wolfskill Orchard, Winters, CA, and extracted in 1995 in the laboratory of M. Aguadé using the attached-*X* strain kindly provided by J. Coyne); Europe (collected in 1993 in Montblanc, Spain, by M. Aguadé and extracted in her laboratory using the attached-*X* strain); and Africa (collected about 1993 in Harare, Zimbabwe, and extracted using the attached-*X* strain in the laboratory of C. H. Langley). We refer to these samples by their continent of origin.

The same samples were used for both the *ewg* and *su(s)* studies. The study of *su(s)* in *D. melanogaster* was reported by AGUADÉ *et al.* (1994) for North America and by LANGLEY *et al.* (2000) for Europe and Africa. Line numbers in the figures in those publications are the same as those in supplementary Tables S1–S9 at http://www.genetics.org/supplemental/. The following lines were not represented in all three studies. For the *D. melanogaster* sample from Africa, lines 51, 52, and 53 were present only in the *ewg* study. For the *D. melanogaster* sample from Europe, line 46 was absent from the *su(s)* study. For the *D. melanogaster* sample from North America, line 13 was absent from the *ewg* study while lines 51 and 52 were absent from the *su(s)* study. For the *D. simulans* study of Europe, line 10 was absent in the *su(s)* study. The sample sizes are presented in Table 1.

SSCP and sequencing: The single-strand conformation polymorphism (SSCP) protocol of AGUADÉ *et al.* (1994) was used to bin sequence fragments (ranging in size from 136 to 345 bp) into allelic classes. The protocol of AGUADÉ *et al.* (1994) was modified in that the fragments were labeled with ^{33}P instead of being silver stained. The locations of the fragments are depicted in Figures 1 and 2. Representative alleles of each SSCP class were sequenced to identify underlying nucleotide polymorphisms. DNA sequencing was carried out on an ABI 377 automated sequencer using standard protocols.

Data analysis: We report $\hat{\pi}$, the average number of pairwise differences per nucleotide. When direct sequencing revealed polymorphism undetected by SSCP, the procedures of AGUADÉ *et al.* (1994) were followed to estimate $\hat{\pi}^*$, the average number of pairwise differences per nucleotide, which incorporates an estimate of the amount of hidden variation. The 95% confidence intervals associated with $\hat{\pi}$ and $\hat{\pi}^*$ were calculated by bootstrapping over alleles for 1000 replications. Calculations of the HKA test (HUDSON *et al.* 1987) and Tajima's *D* (TAJIMA 1989) assumed that sequences within SSCP classes were identical to the sequenced subsample. DnaSP 4.0 (ROZAS *et al.* 2003) was used for the HKA test, the calculation of F_{ST} and the permutation test (HUDSON *et al.* 1992a,b; HUDSON 2000), R_M (HUDSON and KAPLAN 1985), and the estimation of the number of silent sites (NEI and GOJOBORI 1986).

Gene regions: We annotated our *ewg* data from both *D. melanogaster* and *D. simulans* according to GenBank entry no. AE003417, which was prepared as part of the *D. melanogaster* genome annotation release 3.1 (CELNIKER *et al.* 2002). The *D. melanogaster ewg* study included introns (2200 bp, excluding polymorphic indels) and exons (966 bp). Also excluding polymorphic indels, the *D. simulans* survey covered 2252 bp of noncoding DNA (introns) and 941 bp of exons. The total number of silent sites (noncoding + synonymous coding; NEI and GOJOBORI 1986) studied was 2587.16 in *D. melanogaster* and 2472.16 in *D. simulans*. For *D. simulans su(s)*, we followed the GenBank entry no. M57889 (*D. melanogaster*) for our annotation of this gene, and accordingly 2832 noncoding bp were surveyed (excluding polymorphic gaps); this includes introns, a 5′-untranslated sequence, and a 5′-flanking sequence.

Computer simulations: First, neutral coalescent simulations (HUDSON 1990) were used to estimate confidence intervals for Tajima's *D*. We also ran these simulations (10,000 iterations) to estimate exact *P*-values for the observed Tajima's *D*'s. Second, the recurrent simulation method of BRAVERMAN *et al.* (1995) was used to assess the probability of obtaining the observed *D* values (D_o) or greater under a model of recurrent, strong directional selection at linked sites. That probability is labeled Prob $\{D \geq D_o|\text{H.H.}\}$, where H.H. stands for hitchhiking. Next we followed the logic that background selection can be modeled by a neutral coalescent simulation in which the effective population size is appropriately reduced (CHARLESWORTH 1996; STEPHAN *et al.* 1998; LANGLEY *et al.* 2000). Those simulations were used to calculate either Prob $\{D \geq D_o|\text{N.T. and } D_o > 0\}$ or Prob $\{D < D_o|\text{N.T. and } D_o < 0\}$, where N.T. stands for neutral theory. These simulations are conditioned on the observed number of segregating sites, and thus population size is not a factor.

The hitchhiking simulations require calibration. A rate of hitchhiking was chosen to produce, on average, the observed reduction in $\hat{\pi}$ from a value from a region of normal crossing over. It is important to choose a value matching the population source and the type of sequence (*e.g.*, silent sites). For *D. melanogaster*, we set the level of normal variation to be 0.023 in Africa and 0.0081 in North America and Europe. These numbers

were obtained from the DNA sequencing study of *vermilion* (Begun and Aquadro 1995a).

For *D. simulans*, we set the normal level of polymorphism to be 0.0347 for Africa, 0.0279 for Europe, and 0.0288 for North America. These were calculated from *vermilion* from corresponding populations (Begun and Aquadro 1995a; Hamblin and Veuille 1999). In some cases, their data were reanalyzed to obtain estimates of $\hat{\pi}$ for silent sites.

RESULTS

Polymorphism: The results of the SSCP and sequencing study of *ewg* and *su(s)* are presented in supplementary Tables S1–S9 at http://www.genetics.org/supplemental/. A total of 34 variable sites were found in the *ewg* region of *D. melanogaster*. Of these, 15 were indel polymorphisms of 1–34 bp long. The 17 variable sites found in *D. simulans ewg* include 5 indels, each 1 bp long. The *D. simulans su(s)* region was found to have 19 variable sites, including 7 insertion-deletion polymorphisms ranging from 1 to 8 bp.

Hierarchical DNA sequencing of a subset of the SSCP fragments identified the variants; the results are presented in part (b) of supplementary Tables S1–S9 at http://www.genetics.org/supplemental/. In a few cases, sequencing identified variants within SSCP classes. For *D. melanogaster ewg*, sequencing found two single-nucleotide polymorphisms in one fragment not detected by SSCP (in the exon of fragment 6 in the sample from North America). *D. simulans ewg* had eight instances (three different single-nucleotide variants among three fragments, only within Africa) of hidden variation. In *D. simulans su(s)*, there were five cases of the same hidden variant in fragment 9 in the African and North American samples.

A statistical analysis of polymorphism found by the survey of *ewg* and *su(s)* is located in Table 1. The two regions have different levels of polymorphism, with the values of $\hat{\pi}$ and $\hat{\pi}^*$ for *ewg* consistently lower than those for *su(s)*. The *D. melanogaster ewg* variation in the African sample, for example, was less than one-sixth that for *su(s)* (Langley *et al.* 2000). According to coalescent simulations, the probability of obtaining the observed number of segregating sites in the African *ewg* under the neutral model and no intralocus recombination, assuming it has the same value of $3N\mu$ as $\hat{\pi}$ for *su(s)*, is <0.001 (Hudson 1990). In *D. simulans*, variation at *ewg* is less than half that at *su(s)*. While values differ enough that the bootstrap 95% confidence intervals do not overlap in this comparison, the *ewg* and *su(s)* regions do not have significantly different estimates of $3N\mu$ according to neutral simulations. The other populations also were compared with simulations but power was too low to reveal differences.

Comparing across species, two different trends emerge (Table 1). Within Africa, the level of polymorphism ($\hat{\pi}$ and $\hat{\pi}^*$) is higher in *D. simulans* than in *D. melanogaster* at both *ewg* (0.00079 *vs.* 0.00035) and *su(s)* (0.00219 *vs.* 0.00182), although the confidence intervals overlap for the *su(s)* comparison. The same trend presents itself for *su(s)* of Europe and North America. The opposite trend appears in *ewg* of Europe and North America. A simulation analysis was conducted to test for a difference in the levels of genetic hitchhiking in the two species (see *Simulation analysis* below), but none was detected.

Compared to other *X*-linked loci from regions with normal levels of crossing over, *ewg* and *su(s)* have less variation. For this comparison, Langley *et al.* (2000) used averages of $\hat{\pi}$ values from the *white* and *vermilion* regions, studied in the same populations with RFLP (Miyashita and Langley 1988; Begun and Aquadro 1993). The averages for Africa and North America are 0.007 and 0.004, respectively. These numbers are well above all the values observed in this study. For example, the African *white-vermilion* average $\hat{\pi}$ is 21 times greater than the African *ewg* average $\hat{\pi}$.

More recent data from DNA sequencing studies are available for such comparisons against genes from *X*-linked regions of normal levels of crossing over. The *vermilion* locus, for example, was studied using DNA sequencing in a number of populations and in both *D. melanogaster* and *D. simulans* (Begun and Aquadro 1995a; Hamblin and Veuille 1999). Their data are an appropriate baseline for comparison because they did not reject the neutral model according to the HKA (Hudson *et al.* 1987) and Tajima (Tajima 1989) tests for most of the cases. The study of *D. simulans vermilion* by Hamblin and Veuille (1999) focused on a region of the gene with the highest level of polymorphism. So its $\hat{\pi}$ may not represent average levels in African and European populations. For comparison with our North American sample, we use the Begun and Aquadro (1995a) *vermilion* data from North Carolina. All of these data were reanalyzed to give values of $\hat{\pi}$ for silent sites (noncoding and synonymous sites combined).

Comparison of $\hat{\pi}$ from *vermilion* with $\hat{\pi}$ and $\hat{\pi}^*$ *ewg* and *su(s)* within *D. melanogaster* show remarkable reductions in variation. For example, the African sample (Table 1) exhibits a 65-fold reduction in polymorphism at *ewg* compared to *vermilion* ($\hat{\pi}^* = 0.00035$ at *ewg vs.* $\hat{\pi} = 0.0023$ for the *vermilion* silent sites). *D. melanogaster su(s)* polymorphism is also reduced (*e.g.*, >12-fold in the African sample; see Langley *et al.* 2000).

In *D. simulans, ewg* also has much less variation than *vermilion*. We recalculated the statistics for silent sites using the data collected by Hamblin and Veuille (1999). The value of $\hat{\pi}$ for *vermilion* from Africa (in Zimbabwe, but a different collection date), for example, is 0.035 for *vermilion*, but $\hat{\pi}^*$ is only 0.00079 for *ewg* (Table 1). The *su(s)* locus has a $\hat{\pi}^*$ of only 0.00219 in the African sample. Again, this is a major decrease in variation (>40-fold).

Divergence: The interspecific divergences between *D. melanogaster* and *D. simulans* at *ewg* and *su(s)* for all

TABLE 1

Polymorphism statistics for *ewg* and *su(s)*

Species	Population	Locus	n	$\hat{\pi}$ (nt)	Bootstrap	S		Tajima's D			Simulated P-values	
						nt	Indel	nt	Indel	Pooled	N.T.	H.H.
D. melanogaster	Africa	*ewg*	53	0.00035	0.00017–0.00053	10 (7)	7 (4)	−1.67*	−1.98**	−2.04**	0.0228	0.2126
D. melanogaster	Europe	*ewg*	52	0.00007	0.00001–0.00013	4 (4)	6 (4)	−1.87**	−1.88**	−2.21***	0.0093	0.5177
D. melanogaster	North America	*ewg*	51	*0.00057	0.00050–0.00071	4 (1)	6 (3)	−0.47 (NS)	−0.61 (NS)	−0.65 (NS)	0.3825	0.0537
D. melanogaster	Africa	*su(s)*	50	*0.00182	0.00170–0.00205	41 (18)	7 (2)	−1.28 (NS)	−0.47 (NS)	−1.19 (NS)	0.0823	0.0583
D. melanogaster	Europe	*su(s)*	51	*0.00035	0.0002–0.00044	8 (3)	6 (3)	−1.54*	−1.02 (NS)	−1.50*	0.0377	0.2340
D. melanogaster	North America	*su(s)*	50	*0.00102	0.00093–0.00109	10 (1)	3 (1)	+1.31 (NS)	−0.14 (NS)	+1.01 (NS)	0.0812	0.0012
D. simulans	Africa	*ewg*	51	*0.00079	0.0007–0.00087	9 (6)	5 (1)	−1.23 (NS)	−0.10 (NS)	−1.30 (NS)	0.1007	0.1088
D. simulans	Europe	*ewg*	44	0.00000	NA	0 (0)	0 (0)	NA	NA	NA	NA	NA
D. simulans	North America	*ewg*	50	0.00002	0.00000–0.00005	1 (1)	0 (0)	−1.10 (NS)	NA	−1.10 (NS)	NA	NA
D. simulans	Africa	*su(s)*	51	*0.00219	0.00201–0.00235	12 (2)	2 (1)	+0.93 (NS)	+0.93 (NS)	+0.80 (NS)	0.8592	0.0006
D. simulans	Europe	*su(s)*	43	0.00084	0.00052–0.00106	7 (0)	6 (4)	+1.28 (NS)	+1.28 (NS)	−1.57*	0.0821	0.0009
D. simulans	North America	*su(s)*	50	*0.00118	0.00097–0.00131	7 (0)	4 (0)	+1.96*	+1.96*	+2.44**	0.0258	0.0002

Values in the $\hat{\pi}$ column either are calculated directly as the average number of pairwise differences per site or are $\hat{\pi}$* (preceded by *; AGUADÉ *et al.* 1994). Values were calculated with the nucleotide (nt) site polymorphisms and not indels. The 95% confidence interval was calculated with a bootstrap of the polymorphic sites and 10,000 replicates. S is the number of segregating sites, either nt or indel; the number of singletons is indicated in parentheses. The asterisks after the values of Tajima's D indicate the following significance levels: $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$. The exact P-values of Tajima's D for nt polymorphisms are found in the first simulation results column. The last two columns on the right side of the table contain probabilities obtained from computer simulation of evolutionary models. H.H. indicates hitchhiking, specifically, Prob $\{D \geq D_o|\text{H.H.}\}$. N.T. indicates neutral theory and can be interpreted as background selection; the simulations yielded Prob $\{D \geq D_o|\text{N.T. and } D_o > 0\}$ or Prob $\{D < D_o|\text{N.T. and } D_o < 0\}$. D_o is the observed Tajima's D. See MATERIALS AND METHODS and RESULTS for more information.

TABLE 2

Divergence at *ewg* and *su(s)*

	No. of sites compared	No. of differences	Jukes-Cantor distance
ewg			
Total	3139	169	0.056
Silent	2409.58	150	0.101
Nonsynonymous	721.42	19	0.029
su(s)			
Silent	2794	318	0.123

"Silent" includes both synonymous and noncoding sites.

sites studied by SSCP are 0.056 and 0.123, respectively (Table 2). When considering only silent sites, *ewg* divergence is 0.101. These are similar to the average value, 0.061, reported for noncoding regions by Moriyama and Powell (1996). At *vermilion*, silent divergence is 0.185 (Begun and Aquadro 1995a). The level of divergence at *y-ac-sc* ranged between 0.0695 and 0.0558, depending on the type of data (Martín-Campos *et al.* 1992). The average of Jukes-Cantor divergences reported by Begun and Whitley (2000) for 21 *X*-linked loci in regions of normal crossing over is 0.112. Our divergence estimates for *ewg* and *su(s)* are comparable to these other values.

Polymorphism and divergence: We applied the HKA test (Hudson *et al.* 1987) to test the null hypothesis that the level of polymorphism is proportional to divergence (data not shown). The ideal reference locus matches the sequence type (here, silent) and the population source. These criteria are met in *vermilion* (Begun and Aquadro 1995a; Hamblin and Veuille 1999), except that a European sampling source was not available for *vermilion* from *D. melanogaster*, so that population sample was tested against North American data. The HKA using *ewg* and *su(s)* individually against *vermilion* was either highly significant ($P < 0.01$) or very highly significant ($P < 0.001$). The 5′-flanking region of *Adh* was also used (Kreitman and Hudson 1991), although the sample is a combination of 11 sequences from many global locales; the results were again always highly significant or very highly significant. Finally, we conducted the test comparing *ewg* and *su(s)*, the two loci from this study. None of those tests was significant. These results (and the normal level of divergence) can be interpreted as strong evidence that the level of polymorphism is reduced at the *ewg* and *su(s)* loci. This reduction of polymorphism is not consistent with the neutral model of molecular evolution.

Frequency spectrum: We used Tajima's *D* (Tajima 1989) to assess the deviations from a neutral expectation of the frequency spectrum of segregating sites. The results are presented in Table 1. The *ewg* region exhibits a number of significantly negative values of *D*, indicating a skew toward rare variants. For example, Africa *ewg* has a significant value (−1.67; $P = 0.0228$), even though a few variants in the African sample have intermediate frequencies (*e.g.*, site 29,790; supplementary Table S1 at http://www.genetics.org/supplemental/). Meanwhile, for the same sampled chromosomes from Africa, *su(s)* and *su(w^a)* have negative but not significant values of Tajima's *D*: −1.28 and −1.04, respectively (Langley *et al.* 2000).

Regarding the European *D. melanogaster* sample, the values of Tajima's *D* are also negative. They are significant in the case of Europe for both *ewg* and *su(s)*. For the same collection, *su(w^a)* also exhibits a negative but not significant Tajima's *D*. The North American *D*'s are negative for *ewg* and *su(w^a)* but not *su(s)*.

Turning to the results for *D. simulans*, Tajima's *D* for the *ewg* region has negative but not significant values (Table 1) in the African sample for both single-nucleotide and indel variation. Only one single-nucleotide variant was found in North America. The lack of polymorphism in the European sample precluded this analysis. The *su(s)* region of *D. simulans*, in contrast, did not have negative values at all, except for the indel variation; the North American sample actually had a significant positive value (+1.96; $P = 0.0258$). Likewise, the North American *D. melanogaster su(s)* had a large positive value. The European *D. simulans* sample has a large positive but not quite significant *D* at *su(s)*.

Simulation analysis: Simulations are a useful method for distinguishing the hitchhiking and background-selection models. They can provide probabilities of observing particular data sets under each model, which can then be compared.

For *D. melanogaster*, the simulation results (Table 1) can be interpreted as follows. First, Tajima's *D* from the *ewg* African and European samples can be explained better by the hitchhiking model than by the background-selection model. This is evident in the negative and significant values of Tajima's *D*'s observed (−1.67 and −1.86). In particular, the hitchhiking simulations showed relatively large *P*-values (0.2126 and 0.5177), while the background-selection (neutral) model is significantly inconsistent with the observed data ($P = 0.0228$ and 0.0093).

Second, the background-selection model seems to explain the value of Tajima's *D* (−0.47) observed in the *D. melanogaster ewg* North American sample better than the hitchhiking model (Table 1). The background-selection *P*-value is 0.3825 while the hitchhiking *P*-value is only 0.0537.

Third, for *su(s)* from *D. melanogaster*, we repeated the simulations presented in Figure 3 of Langley *et al.* (2000) (Table 1). Again, the hitchhiking model explains the observed *D* (−1.54; $P = 0.0377$) in Europe better than the background-selection model does. However, because we used different data (see materials and methods) to calibrate the hitchhiking model, the results are different in the case of *su(s)* for Africa. The

new values of $\hat{\pi}$ from *vermilion* are much larger than the values used by Langley *et al.* (2000). Thus the rate of recurrent hitchhiking required to achieve the observed relative reduction in $\hat{\pi}$ is larger, and the simulated values of Tajima's *D* are smaller. Therefore, the observed value of Tajima's *D* (−1.28), while negative, occurs less often in the hitchhiking simulation runs. However, $P = 0.0583$, so the hitchhiking model is still not rejected. Meanwhile, the background-selection model has $P = 0.0823$, which is also not a significant rejection. Thus, both the hitchhiking model and the background-selection model are marginally consistent with the data, although neither produces a very good fit.

Fourth, the value of *D* from the *D. melanogaster su(s)* from Europe (−1.54), as suggested by Langley *et al.* (2000), is explained better by hitchhiking, even with the new parameters (Table 1). The hitchhiking *P*-value is 0.2340, while background selection is significantly rejected by the data ($P = 0.0377$).

Fifth, the value of Tajima's *D* observed in the *D. melanogaster su(s)* sample from North America is explained better by the background-selection model. Because this sample has a large positive value of Tajima's *D*, we estimated the Prob $\{D \geq D_o|\text{N.T.}\}$ instead of Prob $\{D < D_o|\text{N.T.}\}$ (Table 1). The results indicate where the observed value falls in the upper half of the simulated distribution under the neutral or background-selection models. The simulations show that this value could be accounted for by the background-selection model, but it is not very likely ($P = 0.0812$). Just as the background-selection model is not likely to produce strongly negative values of Tajima's *D*, it is not likely to produce large positive values. This positive *D* is also inconsistent with the hitchhiking model ($P = 0.0012$).

Turning to the *D. simulans* results, Tajima's *D* at *ewg* from Africa has a negative value (−1.23; Table 1), but neither the background-selection model nor the hitchhiking model is rejected under these simulations. The power to discriminate among models is reduced in this case due to the small number of segregating sites. Low polymorphism precludes these analyses entirely in North American and European *D. simulans ewg* samples.

The remaining three cases are from *D. simulans su(s)* (Table 1). All three had positive Tajima's *D*s. The first case, *su(s)* from Africa, is explained better by the background-selection model ($P = 0.8592$). The final two cases had large positive values of Tajima's *D*. Their associated *P*-values, interpreted as Prob $\{D \geq D_o\}$ for both models, are very small. Consequently, neither the background-selection model nor the hitchhiking model is able to explain these cases very well.

Hitchhiking simulations were used to test for a difference in the rate of hitchhiking in the two species. In the case of the *ewg* sample from Africa, it appears that the rate of hitchhiking is greater in *D. melanogaster* than in *D. simulans*, since $\hat{\pi}$ is smaller in the former species. We used the same rates of hitchhiking used above for *D. melanogaster ewg* with the *D. simulans* sample size and number of segregating sites and asked how often the observed reduction, or a smaller one, in the total size of the coalescent tree was obtained. The size of the coalescent tree is proportional to the amount of variation and an indicator of the strength of hitchhiking (Braverman *et al.* 1995). If, all other things being equal, the rate of hitchhiking were significantly greater in *D. melanogaster*, then the distribution of the relative reduction would be well beneath the observed relative reduction in *D. simulans*. The simulation results did not detect evidence of such a difference ($P = 0.8586$). The converse simulations (using *D. simulans* rate and *D. melanogaster* parameters) also did not detect a significant difference ($P = 0.6516$).

Population subdivision: There is one fixed difference among the populations: a nonsynonymous change at site 28,218, fixed in the African *D. melanogaster ewg* sample as GCG (Ala) and as GGG (Gly) elsewhere (Table 3). There is one nearly fixed difference at site 27,501. All lines except one (no. 50, which has a T) in the African sample have an A. The non-African samples also have a T at this site. The *D. simulans* sequence at these two sites is the same as in the non-African populations, suggesting an African origin of the mutation subsequent to the species' colonization of the other locations.

Across species and genes, African populations stand out with the highest level of polymorphism (Table 1). The polymorphic sites in non-African *D. melanogaster ewg* populations are not a subset of those found in Africa. The only exception is one indel polymorphism, at which the rarer form is found only twice in the African sample, once in the European sample, and four times in the North American sample. Similarly, the polymorphisms at *D. simulans ewg* in Africa are not found in the non-African populations, as the latter have nearly no polymorphism.

For *D. simulans su(s)*, the variation is evenly distributed across the three population samples. Of 14 nonunique segregating sites, 8 segregate in all three populations, many at high frequencies. Three are polymorphic only in the African sample. Three tightly linked indels segregate only in the European and North American samples. Thus the European and North American variation cannot be said to be a subset of the African variation.

To measure the level of differentiation among the three populations, we calculated F_{ST} according to Hudson *et al.* (1992b) and applied the permutation test to various subdivision statistics (Hudson *et al.* 1992a; Hudson 2000). First, as a preliminary step, we calculated F_{ST} for *ewg* for comparisons of the three different locales from which the European *D. melanogaster* flies were collected; the values were very low and not statistically significantly different from zero subdivision. Thus we pooled these three groups. Second, we applied the same procedure to the two groups of *D. simulans* North American lines using data from *su(s)* (*ewg* had nearly no variation in

TABLE 3

The haplotypes of *D. melanogaster ewg* for nonunique single-nucleotide polymorphism

Haplotype	27,501 near-fixed	28,218 fixed nonsynonymous	28,430 polymorphism	29,790 polymorphism	*N*
Africa, no. 1	A	C	G	C	42
Africa, no. 2	A	C	T	T	6
Africa, no. 3	A	C	G	T	5
Europe, North America, and *D. simulans* (ancestral)	T	G	G	T	

The indel polymorphism at 29,138 was excluded (see text). Ancestral states were inferred on the basis of aligned sites in the non-African populations and in *D. simulans*.

these populations with which to detect any subdivision). The value was also not significant. This also justifies pooling these two locales.

Second, the estimates of F_{ST} for comparisons between Africa and North America and between Africa and Europe are reported in Table 4. All the comparisons exhibit statistically significant subdivision. The estimates for *D. melanogaster* range from 0.153 to 0.811. The subdivision at *ewg* is greater than at *su(s)*. Subdivision is also present in *D. simulans*, with F_{ST} values ranging from 0.100 to 0.312 (Table 4). They are slightly higher at *ewg* than at *su(s)*.

Linkage disequilibrium: We conducted Fisher's exact test on all pairs of polymorphisms present in at least two lines (*i.e.*, excluding unique polymorphisms) to test for nonrandom associations. Each population was treated separately. Table 5 summarizes the percentage of formally significant ($P < 0.05$) linkage disequilibria among polymorphic sites. Table 5 also presents R_M, the inferred minimum number of recombination events in a sample (Hudson and Kaplan 1985), and average r^2, the squared correlation coefficient. In most cases, $R_M > 0$, evidence for occasional recombination in the history of these sampled alleles in both species. At the same time, the proportions of "significant" tests and the average r^2 indicate substantial linkage disequilibrium. Interlocus linkage disequilibrium estimates (average r^2) in *D. melanogaster* in general are of the same order of magnitude as intralocus estimates, except in the North American sample, where both the *ewg* and *su(s)* intralocus values are higher than the interlocus values and, for *su(s)*, in the European sample. In *D. simulans*, the sample from Africa was the only one with enough polymorphic sites in both genes for this analysis. The average r^2 is the greatest within *su(s)* and is an order of magnitude lower in *ewg* and between the two loci. Average r^2 is even higher within *D. simulans su(s)* from North America and Europe.

DISCUSSION

The data and analysis presented here consider the telomere-proximal region of low crossing over per physical length where the impact of linked selection is most apparent. The evidence for skewed frequency spectra at *ewg* in the African populations of both *D. melanogaster* and *D. simulans* points toward strong positive selection shaping neutral (and more mildly selected) variation at the tip of the *X* chromosome in these two species. The remainder of this discussion considers other forces that may have shaped our data from *ewg* and *su(s)*. We also compare our data to previous publications.

Background selection: Several lines of reasoning argue against the background-selection model as an explanation for the data at *ewg* and *su(s)*. First, Hudson and Kaplan (1995) note that extremely high rates of deleterious mutation are required to obtain the large reductions observed at genes such as those at the telomere. Second, background selection cannot account for significant negative values of Tajima's *D* observed in practical sample sizes (Hudson and Kaplan 1994; Charlesworth *et al.* 1995). Our data include several cases of significantly negative Tajima's *D*'s. The case [*D. simulans su(s)* of North America] of Tajima's *D* that is large and significantly positive also does not fit the background-selection model. The large nonsignificant values of Tajima's *D* (*D. simulans* of Europe and *D. melanogaster* of North America) are not easily explained by the background-selection model according to our simulation analysis (Table 1). Third, Kim and Stephan (2000) compared the two models and found that in general the hitchhiking model better explains polymorphism in regions of very restricted crossing over.

Recombination: Another issue raised by our results is the unexpected evidence for recombination in our sample, indicated by $R_M > 0$. It seems unlikely that crossing over is responsible for these nonzero values of R_M because observed crossing over is very low in this region. In addition, crossing over should reduce the hitchhiking effect, yet polymorphism is in fact low. Another process, gene conversion, could result in $R_M > 0$, which we observed in both genes and in both species (Table 5). Because the population genetic consequence of unbiased gene conversion is effectively short-range double exchange, its impact on linkage disequilibrium is qualitatively different from that of crossing over

TABLE 4

Population structure at *ewg* and *su(s)*

Species	Locus	Populations	F_{ST}
D. melanogaster	*ewg*	Africa *vs.* Europe	0.811
D. melanogaster	*ewg*	Africa *vs.* North America	0.688
D. melanogaster	*su(s)*	Africa *vs.* Europe	0.153
D. melanogaster	*su(s)*	Africa *vs.* North America	0.245
D. simulans	*ewg*	Africa *vs.* Europe	0.312
D. simulans	*ewg*	Africa *vs.* North America	0.308
D. simulans	*su(s)*	Africa *vs.* Europe	0.100
D. simulans	*su(s)*	Africa *vs.* North America	0.256

F_{ST} was calculated according to Hudson *et al.* (1992b, Equation 3) using both the single-nucleotide and insertion-deletion data. The permutation test of the null no-subdivision model (1000 iterations) with all the statistics of Hudson *et al.* (1992a) applied to the new data is very highly significant ($P < 0.001$), except for *D. simulans su(s)* Africa *vs.* Europe ($P < 0.05$). The permutation test analysis of Hudson's S_{nn} (2000) is always very highly significant ($P < 0.001$). We did not adjust *P*-values for multiple tests. The values from *D. melanogaster su(s)* are from Langley *et al.* (2000) and are included as a reference without statistical tests.

(Andolfatto and Nordborg 1998; Frisse *et al.* 2001). For pairs of polymorphic sites less than a gene conversion-track length apart, gene conversion augments the decay of linkage disequilibrium with distance. In contrast, for pairs of polymorphic sites that are more widely separated, gene conversion reduces nonrandom association at a distance-independent rate. For example, Langley *et al.* (2000) noted a lack of long-distance linkage disequilibria and the presence of short-distance disequilibria on the scale of gene conversion, and thus they interpreted the inferred recombination in their samples as gene conversion, not crossover, events.

Before considering any linkage disequilibrium in our data, it is important to note that not much power is available to discern patterns. Not only is there low variation, but also, when there is a skew toward rare variants, the number of nonsingleton sites available for LD analysis is even fewer. Hence it is best to focus on the African sample, which has the highest amount of variation in these regions, and because the African population is probably closest to equilibrium. Two observations from the African *D. melanogaster* data are relevant. First, the average r^2 is of the same order of magnitude within both *su(s)* and *ewg* (0.083 and 0.035, respectively; Table 5), as well as between the loci in the intergenic comparisons [0.034 between *ewg* and *su(s)*]. Thus we did not detect a decrease in the magnitude of linkage disequilibrium over large genomic distances. Second, the proportion of intralocus comparisons with nominally significant linkage disequilibria (17.99% at *su(s)* and 6.67% at *ewg*; Table 5) is not greater at the more distal *ewg* despite the clear reduction in the level of polymorphism. While there is clear evidence of recombination in the history of the sampled alleles at both *ewg* and *su(s)*, the lack of any correlation with distance is consistent with gene conversion being the dominant form of recombination in this genomic region.

In *D. simulans*, the pattern of linkage disequilibrium is difficult to interpret. In African *D. simulans*, the order of magnitude of the r^2 is almost three times higher in *su(s)* than in *ewg* (Table 5). This difference between intralocus average r^2 and proportion of statistically significant associations may be ascribed to the strong skew in the frequency spectrum at *ewg*. On the other hand, the lack of significant interlocus associations between sites in *ewg* and *su(s)* suggests that the crossing over does contribute to recombination in this genomic region in *D. simulans*.

Little is known about the rate of gene conversion. Whether the few polymorphisms in these regions are those building up after a massive selective sweep or the equilibrium variation under background selection, the appearance of clear recombinants indicates that recombination (probably gene conversion) occurs at a rate comparable to (or larger than) that of neutral mutation. As new neutral mutations accumulate, they are recombined. A gene conversion rate of, for example, 10^{-8}/bp and a neutral mutation rate of 10^{-9} may be sufficient to accommodate the observations.

Our data are similar but not identical to those from surveys of DNA sequence polymorphism on the fourth chromosome that found long-distance disequilibria as well as evidence for some form of recombination on respective regions of the *D. melanogaster* fourth chromosome (Jensen *et al.* 2002; Wang *et al.* 2002). Wang *et al.* (2002) found Tajima's *D* to be −0.9745 ($P = 0.1739$) for all regions pooled, and Jensen *et al.* (2002) found Tajima's *D* to be +0.47 in *D. melanogaster* and −0.68 in *D. simulans* for single-nucleotide variation at the *ankyrin* gene. To contrast, we had large positive values of Tajima's *D*. They also found two haplotypes present over long distances. Thus their results do not immediately offer insight into our data.

Random-environment models: Linked selection models such as those studied by Gillespie (1997) might explain some of our results. He investigated random-environment-selection models and observed negative values of Tajima's *D* when selection reduces polymorphism at linked neutral sites. However, relevant sample properties of this statistic and/or appropriate parameter estimates under these models with which to conduct a statistical test on our data are not available.

Levels of polymorphism: A number of studies have measured polymorphism at other telomeric genes in the *D. melanogaster X* chromosome. A comparison of our *ewg* and *su(s)* data to previous results follows. The *yellow* (*y*) gene (and its proximal neighbors *ac* and *sc*), for example, is important because it is located between *ewg* and *su(s)*. As crossing over increases from *ewg* to *su(s)*, it would be interesting to see how polymorphism is

TABLE 5

Recombination and linkage disequilibrium

Species	Population	Quantity	*ewg*	*su(s)*	*ewg-su(s)*
D. melanogaster	Africa	% significant	6.67	17.99	6.5
		Average r^2	0.035	0.083	0.034
		R_M	1	6	1
D. melanogaster	Europe	% significant	0	25	0
		Average r^2	0.003	0.136	0.007
		R_M	0	2	0
D. melanogaster	N. America	% significant	60	38.18	0
		Average r^2	0.283	0.209	0.028
		R_M	2	3	1
D. simulans	Africa	% significant	23.81	25.45	0
		Average r^2	0.082	0.223	0.0153
		R_M	2	3	
D. simulans	Europe	% significant	NA	72.22	NA
		Average r^2	NA	0.446	NA
		R_M	NA	2	
D. simulans	N. America	% significant	NA	86.67	NA
		Average r^2	NA	0.449	NA
		R_M	NA	2	

"% significant" indicates the percentage of *formally* significant pairs; *i.e.*, *P*-values were not adjusted for multiple tests.

affected. For Zimbabwe collections of the *X*-linked *yellow* and *ac*, the values of $\hat{\pi}$ were estimated as 0.0017 and 0.0012 using RFLP data (Begun and Aquadro 1993). A DNA sequencing study of *yellow* from fly collections from Africa (Zimbabwe) estimated $\hat{\pi}$ as 0.0003 (Andolfatto and Przeworski 2001), and an expansion of that survey's sample size ($n = 49$) in more base pairs (2017 bp) yields a $\hat{\pi}$ of 0.000658 (recalculated from data reported by Andolfatto and Wall 2003). Meanwhile, the value of $\hat{\pi}$ reported for *ewg* is 0.00035. This number and its upper bootstrap confidence limit are lower than the last value reported for *yellow*. A $\hat{\pi}$ value of 0.00182 for the *su(s) D. melanogaster* Africa population (Zimbabwe) lies above the *yellow* numbers (Langley *et al.* 2000). Thus the levels of polymorphism at these three loci in the African populations are consistent with their relative distances from the telomere and presumed relative rates of crossing over.

In *D. simulans*, there are only three published studies of DNA sequence variation near the telomere of the *X* chromosome. Martín-Campos *et al.* (1992) found no variation at *y-ac* in a sample of 103 non-African samples. Begun and Aquadro (1991) found very low variation in non-African samples ($\hat{\pi} = 0.0001$ at the same genes in a North American population; $n = 36$). Sheldahl *et al.* (2003) surveyed variation among five lines of *D. simulans* at the same regions mentioned above for *D. melanogaster*. They found an average $\hat{\pi} = 0.00116$ for silent variation over two lines from Africa, two from North America, and one from the Seychelles Islands. While the species average is higher in *D. simulans* than in *D. melanogaster* for regions of normal crossing over (Moriyama and Powell 1996; Andolfatto 2001), these three studies and *ewg* and *su(s)* exhibit more reduced variation in *D. simulans* than in *D. melanogaster* at the *X* telomere.

In the region from the telomere to *ewg* where there is presumably even less crossing over, Sheldahl *et al.* (2003) also surveyed three regions. In the African (Zimbabwe) collection ($n = 4$), the values of $\hat{\pi}$ (silent) were 0, 0, and 0.00272, moving from the most distal to the most proximal. The trend stops just shy of the value reported in Table 1 for *ewg* Africa (Zimbabwe); thus these data from Sheldahl *et al.* (2003) are consistent with those from larger samples.

Demography: Our quantification of population structure (Table 4) can be compared to F_{ST} values from *D. melanogaster su(w^a)*, which had an Africa-Europe F_{ST} of 0.291 and an Africa-North America F_{ST} of 0.343 (Langley *et al.* 2000). The values at *su(s)* and *su(w^a)* are comparable to F_{ST} values for *X*-linked regions of normal crossing over, which have been reported for Africa-North America *D. melanogaster* (*e.g.*, on the basis of RFLP data: *white*, 0.28; *vermilion*, 0.32; *G6pd*, 0.30; *Pgd*, 0.25; Begun and Aquadro 1993). On the basis of DNA sequence data, *vermilion* has F_{ST} values of 0.370 for Africa *vs.* North

America (Begun and Aquadro 1995a). The values at *su(s)* and *su(w^a)* are slightly lower than other values of F_{ST} for regions of reduced recombination (Begun and Aquadro 1993). For example, Begun and Aquadro (1993) estimated F_{ST} as 0.56 for *yellow* and 0.54 for *ac*. Charlesworth (1998) showed that estimates of F_{ST} may be inflated when using low levels of polymorphism, which was the case for *yellow* and *ac*, so there may be no real difference in F_{ST} between the different regions. To contrast, *ewg* has an enormous value of F_{ST} (0.811), which was calculated using a larger number of polymorphic sites than those for *yellow* and *ac*, although the values of $\hat{\pi}$ at *ewg* are lower. The large geographic differentiation at *ewg* reflects the fixed difference and near-fixed difference (Table 3), and it is consistent with a geographically localized hitchhiking event(s). A single parameterization of a model of geographic differentiation by genetic drift and migration would not simultaneously account for this observation and data from the rest of the genome. Hitchhiking associated with strong selection, genomically localized to the *X* telomere and geographically differentiated, is proposed as an *ad hoc* explanation here but quantitatively documented elsewhere (*e.g.*, Baines *et al.* 2004).

Irvin *et al.* (1998) studied population substructure in *D. simulans* using microsatellites and found a much lower level of substructure than that found in *D. melanogaster*, similar to the trends seen in our data (Table 4). These authors interpreted this trend as the result of a much less severe bottleneck in *D. simulans* than what occurred in *D. melanogaster* and/or a more recent colonization of non-African locales by *D. simulans*.

We now consider whether demographic forces can explain our results for the African sample of *D. melanogaster*. The significantly negative Tajima's *D* in the African sample (-1.67, $P = 0.0228$; Table 1) could be the result of bottleneck or expansion. For example, Glinka *et al.* (2003) interpret their data as evidence of population expansion rather than hitchhiking. They studied many *X*-linked loci from the same population (Zimbabwe) and found many significantly negative Tajima's *D*'s yet no significant HKA test results and only a weak correlation between recombination and polymorphism. However, our study contrasts to theirs in several ways, leading to a different conclusion. First, Glinka *et al.* (2003) studied genes from regions of normal crossing over, while the two genes in the present study are from regions of highly restricted crossing over. Glinka *et al.* (2003) treat regions of reduced crossing over as exceptions, while to further understand such regions is precisely the goal of our study. Second, we observe that the amount of polymorphism at *ewg* is lower than that at *su(s)*, which does suggest a correlation between crossing over and polymorphism. A bottleneck or expansion alone could not explain this correlation. Third, our HKA test results are positive, indicating an extreme reduction in polymorphism, in contrast to those of Glinka *et al.* (2003). Fourth, we observed a fixed difference at one site (28,218) and a near-fixed difference at another site (27,501), and the ancestral forms of these differences occur only in samples collected outside Africa. It is unknown whether the first site is itself the target of selection, but the difference at this site is nonsynonymous, making it a more likely target than the remaining synonymous and noncoding sites. Beyond Glinka *et al.* (2003), Andolfatto and Przeworski (2001) studied many genes from an African sample and concluded that hitchhiking is a better explanation than demographic explanations for that data. Innan and Stephan (2003) applied a different method to the same data and also found hitchhiking to be the dominant force, although they were not considering demographic explanations.

Regarding demography and selection in the other cases of significant Tajima's *D*'s from non-African populations in this study, namely, European *D. melanogaster ewg* and *su(s)*, there is also reason to believe that hitchhiking played a role. In *D. melanogaster*, the large F_{ST} values and the greater variability in the African sample support a historical migration from Africa and subsequent restricted migration. This would indicate a demographic influence on non-African polymorphism. However, both Glinka *et al.* (2003) and Orengo and Aguadé (2004) found evidence of selection in European populations. Orengo and Aguadé (2004) point out that this is an expected process during colonization of new environments.

Our results for *D. simulans* included negative but not significant Tajima's *D*'s for Africa. Again, we view those results as consistent with a study by Quesada *et al.* (2003), who surveyed a different African sample, measuring variation in regions with normal to high levels of crossing over and also finding evidence for hitchhiking in *D. simulans*. For non-African populations, Wall *et al.* (2002) reanalyzed the North American *D. simulans* polymorphism data from Begun and Whitley (2000), and the patterns observed were found to be explainable by a simple bottleneck. However, their model fits the data only if the ancestral *X*:autosome effective population sizes ratio is low and if the bottleneck is strong and recent. The authors did not know how reasonable those conditions were (Wall *et al.* 2002). Further, those interpretations are from smaller sample sizes and genomic regions of normal crossing over per physical length and so may not be applicable to our data.

Conclusion: The excess of rare variants at *ewg*, the more distal of the two loci, and high F_{ST} values indicate strong population subdivision among the three populations at *ewg*. These results indicate genetic hitchhiking at *ewg* and perhaps geographically localized hitchhiking events within Africa. The reduction of polymorphism at *su(s)* combined with the excess of high-frequency variants in *D. simulans* is inconsistent with the hitchhiking and background-selection models. Although the *D. simulans su(s)* data are difficult to explain, our data

from *ewg* can be explained by hitchhiking in the telomeric region of the *X* chromosomes of both *D. melanogaster* and *D. simulans*. While this mechanism may reasonably be extrapolated to other telomeres (and perhaps centromere-proximal euchromatic sequences), the extremely reduced crossing over in these regions and the unique functional aspects of telomeres (*e.g.*, telomere capping and the telomere's role in mitotic and meiotic segregation) restrict generalization to the entire genome.

We are grateful to the following people for their assistance: Kalpana White for providing a genomic DNA sequence of the *erect wing* locus; Chip Aquadro for providing Zimbabwe *D. melanogaster* and *D. simulans* flies; Michael Turelli for collecting and providing the California *D. simulans* flies; and Amanda Frank, Debbie Davis, Kristy Martinez, Marc Crepeau, Erja Heikkinen, and members of the Aguadé and Langley laboratories. We thank two anonymous reviewers for improving the manuscript with their comments. Funding for this research came from National Science Foundation (NSF) grant DEB 95-09548 to C.H.L. and grants PB97-0918 and BMC2001-2909 from Comisión Interdepartamental de Ciencia y Tecnología, Spain, and 2001SGR-101 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Catalonia, Spain, to M.A. J.M.B. was supported by postdoctoral fellowships from the NSF/Sloan Foundation and the Ministerio de Educación y Ciencia, Spain.

LITERATURE CITED

Aguadé, M., N. Miyashita and C. H. Langley, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

Aguadé, M., W. Meyers, A. D. Long and C. H. Langley, 1994 Single-strand conformation polymorphism analysis coupled with stratified DNA sequencing reveals reduced sequence variation in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster X* chromosome. Proc. Natl. Acad. Sci. USA **91:** 4658–4662.

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279–290.

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657–665.

Andolfatto, P., and J. D. Wall, 2003 Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. Genetics **165:** 1289–1305.

Aquadro, C. F., 1992 Why is the genome variable? Insights from Drosophila. Trends Genet. **8:** 355–362.

Baines, J. F., A. Das, S. Mousset and W. Stephan, 2004 The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. Genetics **168:** 1987–1998.

Begun, D. J., and C. F. Aquadro, 1991 Molecular population genetics of the distal portion of the *X* chromosome in Drosophila: evidence for genetic hitchhiking of the *yellow-achaete* region. Genetics **129:** 1147–1158.

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548–550.

Begun, D. J., and C. F. Aquadro, 1995a Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **140:** 1019–1032.

Begun, D. J., and C. F. Aquadro, 1995b Evolution at the tip and base of the *X* chromosome in an African population of *Drosophila melanogaster*. Mol. Biol. Evol. **12:** 382–390.

Begun, D. J., and P. Whitley, 2000 Reduced *X*-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97:** 5960–5965.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Celniker, S. E., D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern *et al.*, 2002 Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol. **3:** research0079.0071–0079.0014.

Charlesworth, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res. **68:** 131–149.

Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15:** 538–543.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106–111.

Drysdale, R. A., M. A. Crosby, and The Flybase Consortium, 2005 FlyBase: genes and gene models. Nucleic Acids Res. **33:** D390–D395 (http://flybase.org/).

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Geyer, P. K., A. J. Chien, V. G. Corces and M. M. Green, 1991 Mutations in the *su(s)* gene affect RNA processing in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **88:** 7116–7120.

Gillespie, J. H., 1997 Junk ain't what junk does: neutral alleles in a selected context. Gene **205:** 291–299.

Gillespie, J. H., 2000 Genetic drift in an infinite population: the pseudohitchhiking model. Genetics **155:** 909–919.

Glinka, S., L. Ometto, S. Mousset, W. Stephan and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165:** 1269–1278.

Hamblin, M. T., and M. Veuille, 1999 Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics **153:** 305–317.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Rev. Evol. Biol. **7:** 1–44.

Hudson, R. R., 2000 A new statistic for detecting genetic differentiation. Genetics **155:** 2011–2014.

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. H., and N. L. Kaplan, 1994 Gene trees with background selection, pp. 140–153 in *Non-neutral Evolution: Theories and Data*, edited by G. B. Golding. Chapman & Hall, New York.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

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

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

Innan, H., and W. Stephan, 2003 Distinguishing the hitchhiking and background selection models. Genetics **165:** 2307–2312.

Irvin, S. D., K. A. Wetterstrand, C. M. Hutter and C. F. Aquadro, 1998 Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in New World populations. Genetics **150:** 777–790.

Jensen, M. A., B. Charlesworth and M. Kreitman, 2002 Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics **160:** 493–507.

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

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415–1427.

Koushika, S. P., M. Soller and K. White, 2000 The neuron-

enriched splicing pattern of Drosophila *erect wing* is dependent on the presence of ELAV protein. Mol. Cell. Biol. **20:** 1836–1845.

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

Lachaise, D., M. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup, pp. 159–225 in *Evolutionary Biology*, edited by M. K. Hecht, B. Wallace and G. T. Prance. Plenum, New York.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen and J. M. Braverman, 2000 Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster X* chromosome. Genetics **156:** 1837–1852.

Martín-Campos, J. M., J. M. Comeron, N. Miyashita and M. Aguadé, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics **130:** 805–816.

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

Miyashita, N., and C. H. Langley, 1988 Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

Miyashita, N. T., M. Aguadé and C. H. Langley, 1993 Linkage disequilibrium in the *white* locus region of Drosophila melanogaster. Genet. Res. **62:** 101–109.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Orengo, D. J., and M. Aguadé, 2004 Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics **167:** 1759–1766.

Quesada, H., U. E. Ramirez, J. Rozas and M. Aguadé, 2003 Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics **165:** 895–900.

Rozas, J., J. C. Sanchez-Delbarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Sheldahl, L. A., D. M. Weinreich and D. M. Rand, 2003 Recombination, dominance and selection on amino acid polymorphism in the Drosophila genome: contrasting patterns on the *X* and fourth chromosomes. Genetics **165:** 1195–1208.

Stephan, W., and S. J. Mitchell, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics **132:** 1039–1045.

Stephan, W., L. Xing, D. A. Kirby and J. M. Braverman, 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc. Natl. Acad. Sci. USA **95:** 5649–5654.

Sturtevant, A. H., C. B. Bridges, T. H. Morgan, L. V. Morgan and J. C. Li, 1929 Contributions to the genetics of *Drosophila simulans* and *Drosophila melanogaster*. Carnegie Inst. Wash. **399:** 1–62.

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

Voelker, R. A., W. Gibson, J. P. Graves, J. F. Sterling, M. T. Eisenberg *et al.*, 1991 The Drosophila suppressor of sable gene encodes a polypeptide with regions similar to those of RNA-binding proteins. Mol. Cell. Biol. **11:** 894–905.

Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.

Wang, W., K. Thornton, A. Berry and M. Long, 2002 Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134–137.

Communicating editor: L. Harshman