Joint Effects of Natural Selection and Recombination on Gene Flow Between *Drosophila ananassae* Populations

Ying Chen, Brenda J. Marsh[1] and Wolfgang Stephan

Department of Biology, University of Rochester, Rochester, New York 14627-0211

Manuscript received August 19, 1999
Accepted for publication March 23, 2000

ABSTRACT

We estimated DNA sequence variation in a 5.7-kb fragment of the *furrowed* (*fw*) gene region within and between four populations of *Drosophila ananassae*; *fw* is located in a chromosomal region of very low recombination. We analyzed gene flow between these four populations along a latitudinal transect on the Indian subcontinent: two populations from southern, subtropical areas (Hyderabad, India, and Sri Lanka) and two from more temperate zones in the north (Nepal and Burma). Furthermore, we compared the pattern of differentiation at *fw* with published data from *Om(1D)*, a gene located in a region of normal recombination. While differentiation at *Om(1D)* shows an isolation-by-distance effect, at *fw* the pattern of differentiation is quite different such that the frequencies of single nucleotide polymorphisms are homogenized over extended geographic regions (*i.e.*, among the two populations of the northern species range from Burma and Nepal as well as among the two southern populations from India and Sri Lanka), but strongly differentiated between the northern and southern populations. To examine these differences in the patterns of variation and differentiation between the *Om(1D)* and *fw* gene regions, we determine the critical values of our previously proposed test of the background selection hypothesis (henceforth called F_{ST} test). Using these results, we show that the pattern of differentiation at *fw* may be inconsistent with the background selection model. The data depart from this model in a direction that is compatible with the occurrence of recent selective sweeps in the northern as well as southern populations.

DURING the past decade studies of genetic variation in Drosophila have focused on the detection of natural selection at the DNA sequence level by comparing patterns of variation in gene regions of low and high recombination rates. Most of these studies have found that levels of average nucleotide diversity in low-recombination regions are reduced (Aguadé *et al.* 1989; Stephan and Langley 1989). In contrast, divergence between closely related species is not affected by recombination (Berry *et al.* 1991; Begun and Aquadro 1992). This lack of correlation between levels of variation and divergence is not consistent with a constant-mutation-rate neutral model (Kimura 1983), but can be explained by models invoking natural selection. In particular, two models have been suggested: (i) the hitchhiking model, which considers the effect of rare, strongly advantageous substitutions on linked, neutral polymorphisms (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992), and (ii) the background selection model, which assumes that the driving mutations are frequent and strongly deleterious (Charlesworth *et al.* 1993; Hudson and Kaplan 1995; Charlesworth 1996).

Properties of these selection models have been explored for panmictic populations, but little work has been done for substructured populations. In a substructured population, background selection against deleterious mutations has been shown to increase F_{ST}, a relative measure of differentiation between subpopulations, in chromosomal regions of low recombination because the effective size of local demes is reduced relative to that of high-recombination regions (Charlesworth *et al.* 1997). On the other hand, directional selection and genetic hitchhiking associated with the fixation of advantageous alleles may lead to greater homogeneity among populations if the selected allele causing the hitchhiking event in one deme migrates to other demes and causes a hitchhiking event in these demes as well. This scenario, however, is expected only in regions of zero or extremely low recombination. For larger recombination rates, different neutral variants may become linked to the selected allele in the population in which the advantageous mutation arose. In this situation limited migration of the selected allele may lead to temporarily increased differentiation at linked neutral loci between populations (Slatkin and Wiehe 1998). Yet another case is that of local adaptation, in which the selected allele causing the hitchhiking event is locally adapted. Here hitchhiking events are assumed to be

Corresponding author: Wolfgang Stephan, Department of Biology, University of Rochester, Rochester, NY 14627-0211.
E-mail: stephan@troi.cc.rochester.edu

[1] *Present address:* Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550.

restricted to single demes or parts of the species range and, as a consequence, may cause substantial genetic differences between populations (Stephan and Mitchell 1992; Begun and Aquadro 1993).

To test these ideas about the migration behavior of selected genes and to compare it to that of neutral genes, we utilize the facts that (1) in some well-characterized sexually reproducing species, such as Drosophila (Lindsley and Sandler 1977), recombination rates vary drastically along chromosomes and (2) the evolutionary dynamics of genes in regions of high recombination rates do not generally deviate from a (nearly) neutral model, whereas genes in regions of reduced recombination exhibit footprints of natural selection due to linkage to selected loci. As a species showing extensive population structure, we use *Drosophila ananassae* (Stephan *et al.* 1998). *D. ananassae* is a largely tropical species, which exists in many semi-isolated populations around the equator, particularly on mainland Southeast Asia and on the islands of the Pacific Ocean (Tobari 1993, chap. 3).

Here we report patterns of nucleotide variation in a 5.7-kb segment of *furrowed* (*fw*), a gene located in a region of very low recombination near the centromere on the right arm of the *X* chromosome of *D. ananassae* (Stephan and Mitchell 1992), and compare it with *Om(1D)*, a previously surveyed gene located in a region of intermediate recombination rates (Stephan *et al.* 1998). To explore the observed differences in the patterns of variation and differentiation between these two gene regions, we refine our recently proposed F_{ST} test of background selection (Stephan *et al.* 1998) and apply it to this new and much larger data set. As in our previous study, we estimate DNA sequence variation in four populations along a latitudinal transect on the Indian subcontinent: two from southern, subtropical areas (Hyderabad, India, and Sri Lanka), and two from more temperate zones in the north (Nepal and Burma).

MATERIALS AND METHODS

Strains and DNA preparation: A total of 45 *D. ananassae X* chromosome lines that originated from four different collections were used in this survey: 9 lines from Mandalay (Burma), 10 lines from Hyderabad (India), 12 lines from four localities near Kathmandu in Nepal (Bharatpur, Godawari, Hetauda, and Kathmandu), and 14 from two localities in Sri Lanka (Beruwala and Colombo). The lines and the isolation of the *X* chromosomes are described in Stephan *et al.* (1998). The *fw* sequence of *D. ananassae* (STD) was determined from three *Eco*RI subclones derived from a *ca*; *px* strain (Stephan and Mitchell 1992). The *D. pallidosa* strain was obtained from the National Drosophila Stock Center (Bowling Green, OH) and put through four generations of brother-sister mating (Stephan *et al.* 1998).

Sequencing of *fw* clones: The *fw* sequence (STD) of Figure 1 was determined using three *Eco*RI fragments (R1, R9, and R42) carried in pUC18 vectors and kindly provided by V. Corces. R1 is a 4.2-kb fragment of the *fw* gene region covering part of the 5′ untranslated region (UTR) and exons 1–9; R9 is a 1.2-kb fragment containing exon 12 as well as part of the 3′ UTR; and R42 is a 4.7-kb fragment encompassing the 3′ flanking region. The GenBank accession number of R1 is AF185289 and that of R9 and R42 (combined) is AF185290. The R11 fragment, located between the R1 and R9 fragments and containing exons 10 and 11, was not sequenced because of technical difficulties (it contained long stretches of repetitive DNA). Both DNA strands of all three *Eco*RI fragments were sequenced. Sequencing was performed on an ABI377 automated sequencer. The total length of the region sequenced was 10,025 bp and includes most of the transcriptional unit (except for the two exons in the R11 fragment) and a large fraction of the 3′ flanking region. The exon/intron structure was determined by comparison with the *D. melanogaster fw* sequence (Leshko-Lindsay and Corces 1997; see also GenBank accession no. U70770).

SSCP analysis and sequencing of the wild-caught *fw* alleles: Single-strand conformation polymorphism (SSCP) analysis and stratified sequencing were used to identify polymorphisms because of the greater efficiency of this method, relative to direct sequencing, in regions with low levels of variation. SSCP analysis was run as described in Stephan *et al.* (1998). A total of 34 pairs of primers were used to amplify overlapping segments of *fw* (numbers in parentheses are distances between primers in base pairs):

R1 fragment: 2151–2177 and 2352–2377 (227), 2317–2338 and 2543–2563 (247), 2512–2536 and 2681–2699 (188), 2659–2678 and 2859–2878 (220), 2806–2830 and 3038–3057 (252), 2927–2946 and 3173–3193 (267), 3142–3165 and 3410–3430 (289), 3385–3406 and 3638–3658 (274), 3603–3626 and 3781–3802 (200), 3719–3740 and 3950–3974 (256), and 3928–3949 and 4124–4157 (230).

R9 and R42 fragments: 6–34 and 248–266 (261), 215–235 and 480–495 (281), 441–460 and 590–607 (167), 480–495 and 719–741 (262), 590–607 and 814–835 (246), 719–741 and 934–957 (239), 814–835 and 1046–1066 (253), 934–957 and 1180–1200 (267), 1251–1285 and 1447–1467 (217), 1422–1444 and 1636–1655 (234), 1610–1629 and 1837–1860 (251), 1803–1828 and 2043–2065 (263), 1987–2009 and 2205–2224 (238), 2180–2202 and 2432–2452 (273), 2391–2414 and 2598–2617 (227), 2557–2584 and 2757–2788 (232), 2684–2703 and 2872–2891 (208), 2850–2871 and 3055–3079 (230), 3025–3050 and 3229–3246 (222), 3175–3201 and 3384–3407 (233), 3353–3373 and 3588–3609 (257), 3517–3535 and 3743–3765 (249), and 3685–3708 and 3925–3944 (260).

For each primer pair, all 45 wild-caught lines were scored by mobility class. Then the alleles were retested by grouping samples into mobility classes and electrophoresing similar classes side by side. Two randomly chosen lines were subsequently sequenced for each mobility class (unless only one variant existed). Both strands were sequenced. Sequencing did not reveal any undetected polymorphism; thus we proceeded assuming all variation was detected. Primers constructed for the SSCP analysis were used to sequence both *fw* strands of *D. pallidosa*.

***F*$_{ST}$ test of the background selection model:** The basic idea of this test of the background selection model in a substructured population was described in Stephan *et al.* (1998). For small samples, background selection generates genealogies that are approximately identical to those produced by a strict neutral model if the effective population size is adjusted such that the effects of recombination and background selection on the locus of interest are taken into account (Hudson and Kaplan 1995). The slight distortions of the allele frequency spectrum produced by background selection (Charlesworth *et al.* 1993; Fu 1997) are neglected because these can only be ob-

served in rather large samples (not used here). The effect of background selection on neutral variation in a substructured population can thus be analyzed by simulating a neutral coalescent in an appropriate model of population structure. We use the finite island model (Crow 1986, chap. 3.4), with background selection incorporated, as the general framework of our simulations (Stephan *et al.* 1998). In these simulations, the following parameters need to be specified: the per-locus nucleotide diversity θ_s, the migration rate M_s, the recombination rate R_s at the locus of interest, and the number of subpopulations k. The simulations were run using a modified version of the coalescent method of Hudson *et al.* (1992). The program with instructions can be downloaded from http://www.rochester.edu/college/BIO/StephanLab/YingChen.html.

The following modifications are introduced compared to the original version of the test. Instead of using Equation 1 of the previous article, the migration parameter of the locus being tested for the effect of selection, M_s, is now estimated from the data as

$$M_s = M_o \frac{\bar{\theta}_s}{\bar{\theta}_o} f_{so}, \qquad (1)$$

where M_o is the migration rate of the reference locus (located in a region of high or normal recombination), which is estimated for each pair of subpopulations of interest (rather than the whole set of subpopulations as in the previous version), $\bar{\theta}_{-s}$ and $\bar{\theta}_o$ are the arithmetic means of the per-site nucleotide diversities in the two subpopulations at the locus to be tested for the effects of selection and at the reference locus, respectively, and f_{so} is the ratio of the neutral mutation rate at the reference locus to that of the other locus. The factor f_{so} is newly introduced here to take into account that the neutral mutation rates (to be estimated from levels of divergence) may be different at the two loci. The new method of estimating M_o makes the test more conservative (see discussion). Furthermore, note that the subscript indicating the locus that may be under the effects of selection has been changed to s compared to the previous article.

Critical values of the F_{ST} test: We produced empirical distributions of the test statistic F_{ST} as a function of the migration rate M_s for specified values of the parameters θ_s, R_s, k, and the sample sizes of the two subpopulations. Using the abovementioned simulation program, 10,000 independent samples were generated for each parameter set. After these empirical distributions were obtained, the critical value at the 5% significance level was determined as the maximum value F^c_{ST} of F_{ST} such that the proportion of samples of the independently simulated samples with $F_{ST} \leq F^c_{ST}$ is equal to 5%. Examples of critical values for some parameter values are provided in Figure 2. To facilitate further analysis, the critical values obtained from simulation were fitted in interval [0, 1] by a function of the form

$$F^c_{ST}(M_s) = \frac{a}{M_s} + b + cM_s + dM_s^2 + e \ln(M_s), \qquad (2)$$

where the coefficients, a, b, c, d, and e were determined by least-squares analysis for each set of parameter values using Mathematica (Wolfram 1991, Chap. 3.8).

Achieved levels of significance: When performing the F_{ST} test, we have to substitute estimates for several parameters. Therefore, the achieved levels of significance with critical values computed as described above may not be close to the nominal levels of significance (see the F_s test in Fu 1997). To analyze the effect of parameter estimation on the achieved levels of the F_{ST} test, we followed Fu's (1996) suggestion and simulated an additional 1000 samples for a given set of parameter values, independently of those used to obtain the critical points. The simulations were done in parallel for the locus of interest (denoted by index s) and the reference locus. For each simulated sample (consisting of a pair of genealogies for each locus), we estimated the values of F_{ST} and of nucleotide diversity and then, using Equation 1, obtained an estimate of M_s, $\hat{M}_s$. Finally, the value of the statistic F_{ST} was compared to the achieved critical value obtained from Equation 2 by substituting $\hat{M}_s$ for M_s.

A full analysis of the properties of the F_{ST} test would require all parameters to be varied extensively. Such an analysis, however, is beyond the scope of this article. To reduce the number of parameters, we assumed that f_{so}, the ratio of the neutral mutation rate at the reference locus to that of the other locus, is accurately estimated from divergence data. The rest of the parameters were varied to some extent. Extensive simulations were done for parameter values close to those estimated from the data (see below).

RESULTS

DNA polymorphism at *fw*: Of the 10.0 kb of the *fw* clones sequenced, 5.7 kb were subjected to SSCP analysis and stratified DNA sequencing. These fragments comprise most of the 3′ half of the *fw* transcriptional unit and a large part of the 3′ flanking region. The complete polymorphism data (variations of the STD sequence) are shown in Figure 1. The estimates of nucleotide diversity (at silent + noncoding sites), $\hat{\pi}$ and $\hat{\theta}$, are extremely low for each population (Table 1), ~20–50-fold lower than in regions of normal rates of crossing over in *D. ananassae* (Stephan and Langley 1989; Stephan *et al.* 1998). However, they are comparable with the levels of nucleotide diversity found at *vermilion* (*v*), which is located in a chromosomal region of low recombination on the other side of the centromere on the *X* chromosome (Stephan *et al.* 1998). The values of Tajima's (1989) *D* statistic are negative in all four populations, reflecting the fact that single nucleotide polymorphisms are generally in low frequency (except for one polymorphism at coordinate 1070 in the Burma sample). Although the *D* values are consistently negative, they do not deviate significantly from zero.

Polymorphism and divergence: Average divergence (at silent + noncoding sites) between *D. ananassae* and its sibling species *D. pallidosa* is very low compared to *Om(1D)* and *v*, the only other gene regions for which divergence data between *D. ananassae* and a close relative are available. While divergence at *Om(1D)* and *v* has been estimated as 0.032 and 0.022, respectively, divergence of all *four* subpopulations from the *D. pallidosa* sequence at *fw* was only 0.0055. Hudson, Kreitman, and Aguadé (HKA) tests (Hudson *et al.* 1987) for each population, using *Om(1D)* as reference locus, revealed $\chi^2 = 5.18$ ($P < 0.025$) for Burma, 3.87 ($P < 0.05$) for Nepal, 1.89 ($P < 0.17$) for India, and 2.90 ($P < 0.09$) for Sri Lanka. Thus, there is a lack of correlation between levels of polymorphism and divergence in the two northern populations, rejecting a constant-rate, neutral model for these data. However, because of the

		R1b							R 9			R42					
		E		**E**	**E**	**I**			**F**								
		6		**7**	**9**	**9**			**3**								
		2	2	3	3	3	3	4			1	1	1	2	3	3	3
		5	6	1	6	9	9	0	6	9	0	5	9	2	3	4	9
		8	0	8	8	4	9	9	8	7	7	2	0	9	4	6	0
		9	6	1	1	5	0	5	8	0	0	6	7	5	1	6	7
	STD	T	T	T	T	$(GT)_2$	A_9	G	A	T	C	T	C	C	A_9	T_7	G
BUR	M68	.	.	.	.	.	A_{10}	.	.	.	T	.	.	.	.	.	.
	M79	.	.	.	.	.	A_{10}	.	.	.	.	.	.	.	.	.	.
	M89	.	.	.	.	.	A_{10}	.	.	.	.	.	T	.	.	.	.
	M90	.	.	.	.	.	A_{10}	.	.	.	.	.	.	.	.	.	.
	M91	.	.	.	.	.	A_{10}	.	.	.	T	.	.	.	.	.	.
	M92	.	.	.	.	.	A_{10}	.	.	.	.	.	.	.	.	.	.
	M97	.	.	.	.	.	A_{10}	.	.	.	T	.	.	.	.	.	.
	M117	.	.	.	.	.	A_{10}	.	.	.	T	.	.	.	.	.	.
	M119	.	.	.	.	.	A_{10}	.	.	.	.	.	.	.	.	.	A
NEP	B1	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	B2	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	B5	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	B7	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	B8	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	H5	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	H14	.	.	.	.	.	A_{11}	C	.	.	.	.	.	.	.	.	.
	H22	.	.	.	.	.	A_{11}	C	.	.	.	.	.	.	.	.	.
	H25	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	G8	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	G13	.	.	.	.	GT	A_{11}	.	.	.	T	.	.	.	.	.	.
	K8	.	.	G	C	.	.	.	G	.	.	.	.	T	A_8	T_8	.
IND	H11	.	A	G	C	.	.	.	G	.	.	A	.	T	.	T_8	.
	H15	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H19	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H23	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H26	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H36	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H50	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H62	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H76	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	H81	.	.	G	C	.	.	.	G	.	.	.	.	T	.	T_8	.
SRI	B2	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	B3	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	B4	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	B5	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	B6	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	B7	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	B8	.	.	G	C	.	.	.	G	C	.	.	.	T	.	T_8	.
	B11	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	C1	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	C2	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	C3	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	C5	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	C6	C	.	G	C	.	.	.	G	.	.	.	.	T	.	T_8	.
	C9	C	.	.	.	.	.	.	G	.	.	.	.	T	.	T_8	.
	PAL	.	.	G	C	.	A_{10}	.	G	C	.	.	.	.	.	T_8	.

Figure 1.—DNA polymorphisms of the *fw* gene region found in 45 lines of *D. ananassae*. The localities of the strains are indicated on the left. The nucleotides within the reference sequence STD are shown at the top. The numbers above the sequence represent the position numbers of each segregating site within the reference sequence. These coordinates refer to the *Eco*RI subclones used to determine the STD sequence: R1b, R9, and R42. Note that R1b covers most of the 3′ half of the *fw* gene, including exon 6 (E6), exon 7 (E7), exon 9 (E9), and intron 9 (I9), while the other clones cover the 3′ flanking region (F3). The homologous nucleotides within the *D. pallidosa* sequence (PAL) are at the bottom.

TABLE 1

Polymorphism at *furrowed*

	Total	Burma	Nepal	India	Sri Lanka
Sample size	45	9	12	10	14
Silent sites	4814				
Segregating sites	12	3	6	5	4
Singletons	5	2	4	2	2
Diversity $\hat{\theta}$		0.0002	0.0004	0.0004	0.0003
Diversity $\hat{\pi}$		0.0002	0.0002	0.0003	0.0002
Tajima's *D*		−0.36	−1.49	−0.68	−1.16
Divergence		0.0055	0.0055	0.0055	0.0055

The number of equivalent silent sites was calculated according to Rozas and Rozas (1997). Nucleotide diversity $\hat{\theta}$ was estimated according to Watterson (1975) and $\hat{\pi}$ according to Nei and Li (1979). The *D* value was obtained by Tajima's (1989) method. Divergence was estimated between each *D. ananassae* population and *D. pallidosa*.

relatively low rate of divergence between *D. ananassae* and *D. pallidosa* at *fw* and the lower levels of nucleotide diversity at *Om(1D)* in the two southern populations (see Stephan *et al.* 1998), the results of the HKA tests are only marginally significant for the populations from India and Sri Lanka.

Test of background selection model: Two alternative models have been proposed to explain the reduction of DNA sequence polymorphism in regions of low rates of crossing over and the lack of correlation between polymorphism and divergence: the hitchhiking model (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992) and the background selection model (Charlesworth *et al.* 1993; Hudson and Kaplan 1995; Charlesworth 1996). The first model assumes the hitchhiking of neutral (or nearly neutral) variants on chromosomes bearing rare, strongly selected, favorable mutations at closely linked loci that go rapidly to fixation. The second model involves the loss of neutral or nearly neutral variants as a result of steady elimination of linked deleterious mutations from the population. To test the background selection model, we used the method proposed by Stephan *et al.* (1998) with the modifications described in materials and methods. This method compares genetic differentiation at a neutral marker gene with the pattern of differentiation at a locus that is to be tested for the possible effects of selection.

The prediction that background selection is expected to increase differentiation between subpopulations at *fw* was tested by generating the probability density of F_{ST} for the finite island model with *k* demes, a given migration rate M_s, a given mutation rate per locus θ_s, and a given recombination rate per locus R_s. We chose a range of reasonable values for the unknown parameters *k* and R_s; θ_s and M_s were estimated from the data [see Equations 1 and 2 in Stephan *et al.* (1998)]. The correction factor f_{so} of Equation 1 was estimated as 5.82, reflecting the observation that divergence at *Om(1D)* is ~5.82 times higher than at *fw*. The results are presented in Table 2. For all combinations of *k* and R_s values used in the simulations, the northern (Burma, Nepal) and southern (India, Sri Lanka) subsamples produced the lowest probabilities ($P < 0.05$). This may suggest that the F_{ST} values observed for the two northern subpopulations as well as for the two southern subpopulations are not compatible with those predicted by a finite island model of population structure that incorporates background selection.

Statistical properties of the F_{ST} test: To interpret the *P* values in Table 2, we investigated the critical points of the test and the achieved levels of significance for a range of parameter values. Due to the multidimensionality of the parameter space, however, we had to concentrate on parameter values that are relevant for our data. Figure 2 shows the critical values of the F_{ST} test as a function of M_s for various values of the parameters *k* and R_s. The remaining parameter values are those used in the analysis of the two southern populations from India and Sri Lanka. The critical points as a function of M_s for other parameter sets look qualitatively very similar to those of Figure 2.

Next we examined the achieved levels of significance of the test using the critical values obtained for each parameter set. We found that parameter estimation had some influence on the test. For instance, the discrepancy between the achieved and nominal levels of significance increased with the number of subpopulations, *k*, and with nucleotide diversity at either locus. For the parameter sets examined, the critical points of the F_{ST} test (at 5% level) were the values corresponding approximately to the lower second to third percentile of its empirical distribution (similar to the F_s test; Fu 1997). Therefore, in Table 2, adjusted *P* values are presented for the two northern and the two southern populations that take this bias of the test into account. Some of the adjusted *P* values for the two northern populations are no longer <0.05.

TABLE 2

Probability of obtaining the observed or lower values of F_{ST} given the background selection model

		$k = 4$		$k = 10$		$k = 100$	
Population 1	Population 2	$R_s = 0$	$R_s = 0.1$	$R_s = 0$	$R_s = 0.1$	$R_s = 0$	$R_s = 0.1$
Burma	Nepal	0.026	0.026	0.025	0.025	0.027	0.011
		(0.040)	(0.048)	(0.081)	(0.069)	(0.096)	(0.047)
Burma	India	0.353	0.322	0.343	0.345	0.337	0.313
Burma	Sri Lanka	0.283	0.271	0.265	0.271	0.260	0.249
Nepal	India	0.452	0.414	0.443	0.458	0.424	0.400
Nepal	Sri Lanka	0.385	0.382	0.363	0.385	0.361	0.332
India	Sri Lanka	0.016	0.015	0.014	0.010	0.013	0.005
		(0.034)	(0.035)	(0.037)	(0.025)	(0.041)	(0.006)

Columns 3–8 contain the probabilities of obtaining the observed or lower values of F_{ST} for the given values of the parameters k and R_s and for the migration rate estimated from Equation 1. The values of F_{ST} are estimated using Equation 6 in Hudson *et al.* (1992). The numbers in parentheses below the probability values, P, for the northern and southern populations are adjusted probabilities that take the bias of the F_{ST} test into account (see text). They are defined as follows: Let $F_{ST}(P)$ be the point that corresponds to the tail probability P with regard to the empirical distribution of F_{ST} values. The adjusted probability is then the proportion of simulated samples with $F_{ST} \leq F_{ST}(P)$. The procedure to simulate these samples is described in materials and methods (*Achieved levels of significance*).

DISCUSSION

Variation within and between populations and divergence between species: The estimates of nucleotide diversity, $\hat{\pi}$ and $\hat{\theta}$, show extremely low levels of nucleotide polymorphism within each of the four populations from Burma, Nepal, India, and Sri Lanka. Nucleotide diversity at *fw* is lower than at *v*, which is located in a region of low recombination on the other side of the centromere of the *X* chromosome. This confirms earlier results from restriction map analyses and agrees with the hypothesis that recombination in the *fw* region is lower than at *v* (Stephan and Mitchell 1992). In contrast to the *vermilion* data (Stephan *et al.* 1998), there is no evidence for a recombination event in the *fw* data set, based on the four-gamete rule (Hudson and Kaplan 1985).

One haplotype (henceforth called M), including the lines K8 (Nepal), H11, H81 (both from India), B8, and C6 (both from Sri Lanka), is of particular interest. The four lines from India and Sri Lanka harbor all the variation detected in the southern populations, and K8 contains much of the variation found in the northern populations. Removing this haplotype, which is distinctly different from the rest of the sample and is thus apparently a recent immigrant from a different (unknown) source population, reduces the estimates of nucleotide diversity within the four populations even more; in fact, nucleotide diversity falls to zero within each of the two southern populations.

Another interesting feature of the data is the between-population distribution of variation. Again, removing haplotype M shows that the remaining seven segregating sites are organized in a very simple way: either they are in low frequency in the total sample (three polymorphisms) or they occur as fixed differences between the northern and southern populations (three). The only exception is the polymorphism at position 1070 of fragment R9. These results confirm an earlier restriction map survey of the *fw* region, in which two fixed differences between the populations from Burma and India were found although the estimates of variation within these populations were zero (Stephan and Mitchell 1992). While the northern and southern populations are strongly differentiated with respect to single nucleotide polymorphisms (but homogeneous within the northern as well as within the southern species range), it is noteworthy that the two northern populations from Burma and Nepal are differentiated with respect to two polymorphic microsatellites in intron 9.

Divergence between *D. ananassae* and *D. pallidosa* was relatively low at *fw*, a factor of ~4–6 lower than at *v* and *Om(1D)*, the only other genes that have been sequenced for these two species. Because of this small number of genes sequenced, it is difficult to evaluate the *fw* result. However, lower levels of divergence in regions of reduced recombination may not be unusual. Indeed, data from *D. melanogaster* suggest that the variance in levels of divergence may be quite large in regions of low recombination rates relative to that in regions of intermediate or high recombination (Begun and Aquadro 1992). Finally, given the fact that differentiation between the northern and southern populations is strong because of the presence of three nearly fixed differences, it should be noted that all four populations show the same average level of divergence from *D. pallidosa*. Although these two observations may appear to be inconsistent, they may simply be due to the fact that differentiation between populations was measured through-

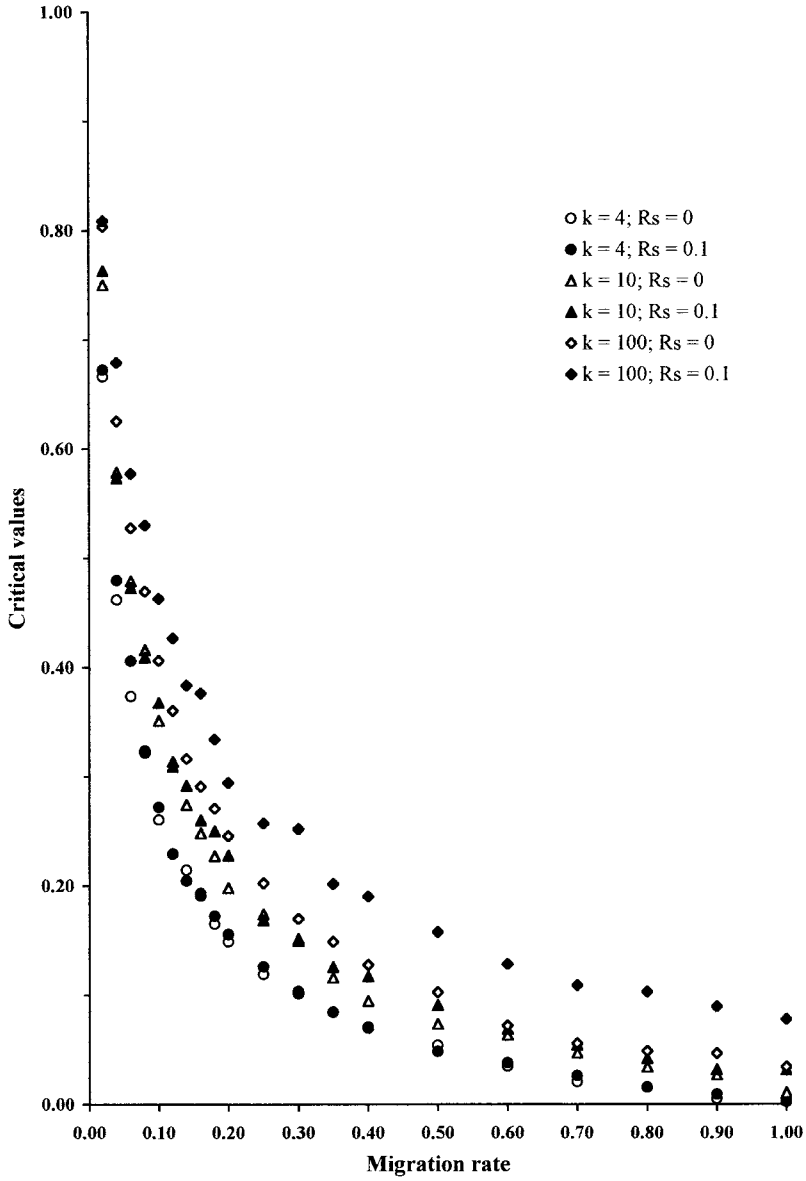

Figure 2.—Critical values of the F_{ST} test at the 5% level *vs.* migration rate M_s. The values of the parameters k and R_s are shown; the remaining parameter values are those pertaining to the two southern populations from India and Sri Lanka (*i.e.*, the sample sizes are 10 and 14, respectively, and the per-locus nucleotide diversity θ_s is 1.685).

out this study by F_{ST}, a relative measure of differentiation, whereas divergence between species was defined in absolute terms. Measuring differentiation between populations in absolute terms, as proposed by Charlesworth (1998), shows that the *fw* haplotypes of the northern and southern populations are much more similar to each other than either of them is to the *D. pallidosa* sequence.

Test of the background selection model: To explain our results, a basic model, such as the nearly neutral model (Ohta 1972), is not sufficient, although some aspects of the data (*e.g.*, the low levels of divergence at *fw*) are qualitatively in agreement with a hypothesis that invokes selective constraints due to slightly deleterious mutations. Adding more strongly selected mutations to this scenario, as postulated by the background selection model (Charlesworth *et al.* 1993), does not suffice either. It may explain the strong reduction of nucleotide diversity in regions of low recombination and also the relative insensitivity of the rates of divergence; *i.e.*, the sixfold lower level of divergence at *fw* relative to *Om(1D)* may be consistent with the background selection hypothesis. Furthermore, the background selection model may explain the strong differentiation between northern and southern populations. However, as our analysis seems to indicate, the background selection model cannot account for the observed homogeneity of single nucleotide polymorphism frequencies among the populations within the northern and southern species ranges.

There are two potential problems with our analyses. First, the F_{ST} test presumes that variation at the reference locus *Om(1D)* is (nearly) neutral, and the results of our test are based on the observation that migration rates between the four subpopulations at *Om(1D)* are low. As mentioned in Stephan *et al.* (1998), there is no evidence that the observed patterns of variation at *Om(1D)* deviate

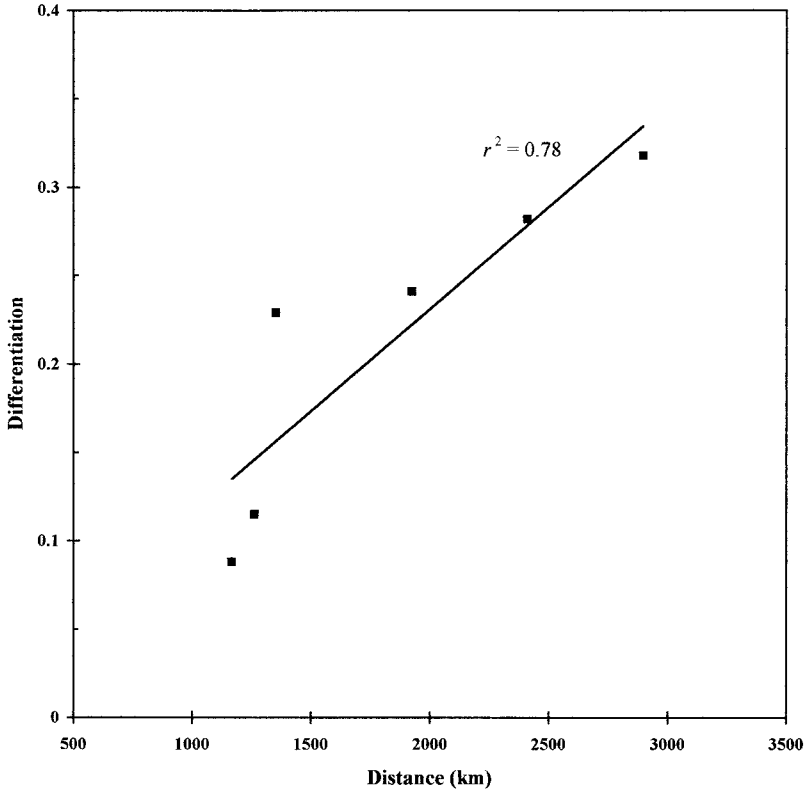

Figure 3.—Genetic differentiation at *Om(1D)* as a function of geographic distance. Genetic differentiation (measured by F_{ST}) is plotted against the geographic distance (in kilometers) between each pair of populations. The F_{ST} values were estimated using the method of Hudson *et al.* (1992). The distances between Mandalay (Burma), Kathmandu (Nepal), Hyderabad (India), and Colombo (Sri Lanka) were estimated by choosing the most direct route across land. The distances are ranked as follows: Hyderabad-Colombo < Mandalay-Kathmandu < Kathmandu-Hyderabad < Mandalay-Hyderabad < Kathmandu-Colombo < Mandalay-Colombo.

from neutrality. We have further addressed this question by measuring genetic differentiation at *Om(1D)* between all pairs of populations. As predicted by the neutral model, we found an "isolation-by-distance" effect (Wright 1943) in that the association between the geographic distance between populations and F_{ST} is statistically significant ($P < 0.025$, linear regression; see Figure 3). Furthermore, recent evidence from microsatellite studies (M. Schug, personal communication) suggests that the pattern of variation between the four subpopulations at *Om(1D)* is not locus specific. Our result that migration rates between these subpopulations are low at *Om(1D)* is rather in qualitative agreement with Schug's observation of strong differentiation at microsatellite loci.

Second, because the background selection model was incorporated into a specific model of population structure, the finite island model, the latter may be the problem instead of the background selection model itself. Is the finite island model a realistic model of migration for *D. ananassae* at the geographic scale used in this study? As mentioned above, we found an "isolation-by-distance" effect in that the association between the geographic distance between populations and F_{ST} is statistically significant (Figure 3). This suggests that the finite island model in which migration between each pair of populations is assumed to be identical may not be an adequate description of population subdivision for *D. ananassae.* At the same time, this result raises the question of how a more realistic, stepping-stone-type model in which migration rates between geographically more distant populations are lower would change our results.

We address this latter question for the two southern populations from Hyderabad, India, and Sri Lanka, but a similar argument applies to the subsample of the two northern populations, for which the pattern of differentiation may also be inconsistent with the background selection hypothesis (Table 2). Consider a coalescent process for a stepping-stone-type model with k demes (*i.e.*, populations that are assumed to be significantly differentiated among each other) with background selection incorporated such that the migration rates between populations are comparable (approximately equal) or smaller than the estimated rate of migration between the two populations from Hyderabad and Sri Lanka. This scenario is constructed based on the observation that the two southern populations under consideration (which are significantly differentiated; see Stephan *et al.* 1998) are separated by a smaller geographic distance than most other pairs of *D. ananassae* populations within the zoogeographic range of *D. ananassae.* (Note that this scenario includes all actual *D. ananassae* populations that are differentiated, not only those that have been sampled in this study.) Geography may imply that the migration rate between the Hyderabad and Sri Lanka populations is probably higher than the migration rates between most other pairs of populations within the zoogeographic range of *D. ananassae.* Thus, simulating a coalescent for this scenario under a background selection model would produce a distribution of F_{ST} values that is shifted to values larger than in the corresponding finite island model in which all migration rates are the same as between the populations from

Hyderabad and Sri Lanka. As a consequence, the probability of obtaining the observed F_{ST}, given the background selection model incorporated into a stepping-stone-type population structure, is likely to be lower than the corresponding value shown in Table 2 for the finite island model.

Alternative explanations: Our analyses indicate that, in addition to deleterious mutations, one may have to postulate the occasional occurrence of advantageous mutations. As with the *v* locus (Stephan *et al.* 1998), the observed pattern of differentiation at *fw* is consistent with recent selective sweeps, which homogenized single nucleotide polymorphism frequencies within the northern as well as within the southern populations. The sweep model seems also to be consistent with a large variance in between-species divergence rates in regions of low recombination (although no theoretical predictions are available at present). Furthermore, it appears that the fine-scale differentiation of two polymorphic microsatellite loci in intron 9 (between the Burma and Nepal populations) is in qualitative agreement with this explanation as length changes in microsatellites can occur more quickly than base substitutions (Schug *et al.* 1997).

Whether the data can be explained by a single-sweep model (Slatkin and Wiehe 1998) or whether independent sweeps have to be postulated (Stephan and Mitchell 1992) is difficult to decide. A single-sweep model seems to be more parsimonious than the other explanation. But for a single-sweep model to explain the data, at least two haplotypes of the *fw* locus must be linked to the advantageous allele. This would require that the (unknown) levels of nucleotide variation before the hitchhiking event are sufficiently high and that the sweep takes sufficiently long to spread through the northern and southern populations so that different haplotypes become associated with the advantageous mutation via recombination. The occurrence of (nearly) fixed differences between populations can be explained by both models. To distinguish between these models, low-frequency polymorphisms may be used. Most of the observed low-frequency polymorphisms, however, are probably due to recent mutations, and the variant C at position 970 of fragment R9 seems to be due to a recent migration event (discussed above).

Conclusions: Our results suggest that the pattern of variation between populations and thus the migration behavior of nuclear genes depends critically on at least two factors (other than geography): the recombination environment in which a gene is located and natural selection. These results are important as they shed new light on inferences of the migration history of populations (including humans) from molecular data.

We thank two reviewers for their excellent comments on a previous version of this article. Furthermore, we thank Seth Bordenstein and Alex Mohseni for their help with the SSCP analysis. This work was supported in part by National Science Foundation grant DEB-9896179 to W.S. and an Ernst Caspari fellowship to Y.C.

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.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. Nature **356:** 519–520.

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.

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

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, 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.

Crow, J. F., 1986 *Basic Concepts in Population, Quantitative, and Evolutionary Genetics.* Freeman, New York.

Fu, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

Fu, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

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

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

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

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

Leshko-Lindsay, L. A., and V. G. Corces, 1997 The role of selectins in Drosophila eye and bristle development. Development **124:** 169–180.

Lindsley, D. L., and L. Sandler, 1977 The genetic analysis of meiosis in female *Drosophila melanogaster*. Philos. Trans. R. Soc. Lond. B Biol. Sci. **277:** 295–312.

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

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

Ohta, T., 1972 Evolutionary rate of cistrons and DNA divergence. J. Mol. Evol. **1:** 150–157.

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

Schug, M. D., T. F. C. Mackay and C. F. Aquadro, 1997 Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nat. Genet. **15:** 99–102.

Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. **71:** 155–160.

Stephan, W., and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the *X* chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. Genetics **121:** 89–99.

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., 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.

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.

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

Tobari, Y. N., 1993 *Drosophila ananassae: Genetical and Biological Aspects.* Japan Scientific Societies Press, Tokyo.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wolfram, S., 1991 *Mathematica*: *A System for Doing Mathematics by Computer*. Addison-Wesley Publishing Co., Reading, MA.

Wright, S., 1943 Isolation by distance. Genetics **28:** 114–138.

Communicating editor: C.-I Wu