DOI: 10.1534/genetics.103.025494

Identification of a Locus Under Complex Positive Selection in *Drosophila simulans* by Haplotype Mapping and Composite-Likelihood Estimation

Colin D. Meiklejohn,*,[1] Yuseob Kim,[†,2] Daniel L. Hartl* and John Parsch[‡]

**Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, [†]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853 and [‡]Department of Biology II, Section of Evolutionary Biology, University of Munich (LMU), 80333 Munich, Germany*

Manuscript received December 10, 2003
Accepted for publication June 7, 2004

ABSTRACT

The recent action of positive selection is expected to influence patterns of intraspecific DNA sequence variation in chromosomal regions linked to the selected locus. These effects include decreased polymorphism, increased linkage disequilibrium, and an increased frequency of derived variants. These effects are all expected to dissipate with distance from the selected locus due to recombination. Therefore, in regions of high recombination, it should be possible to localize a target of selection to a relatively small interval. Previously described patterns of intraspecific variation in three tandemly arranged, testes-expressed genes (*janusA*, *janusB*, and *ocnus*) in *Drosophila simulans* included all three of these features. Here we expand the original sample and also survey nucleotide polymorphism at three neighboring loci. On the basis of recombination events between derived and ancestral alleles, we localize the target of selection to a 1.5-kb region surrounding *janusB*. A composite-likelihood-ratio test based on the spatial distribution and frequency of derived polymorphic variants corroborates this result and provides an estimate of the strength of selection. However, the data are difficult to reconcile with the simplest model of positive selection, whereas a new composite-likelihood method suggests that the data are better described by a model in which the selected allele has not yet gone to fixation.

THE recent action of positive selection is expected to leave a footprint on patterns of intraspecific DNA sequence variation. This footprint results from the effects that selection imposes on the genealogy of genomic segments linked to the selected variant and the spatial localization of these effects due to recombination. Because all sequences with complete linkage to the beneficial mutation must coalesce before the mutation event (going backward in time), these sequences will form a star-like tree of relationships with very short branch lengths relative to the neutral coalescent expectation (Kaplan *et al.* 1989). The star phylogeny results in a decrease in heterozygosity associated with the rapid increase in frequency of the beneficial mutation, which has been called genetic hitchhiking (Maynard Smith and Haigh 1974) or selective sweep. The extent of the region affected by a sweep is determined by the strength of selection and the local rate of recombination (Maynard Smith and Haigh 1974; Kaplan *et al.* 1988, 1989). As neutral mutations accumulate on a postsweep genealogy, most of those present in samples of reasonable size will be singletons or of low frequency, producing a skew in the site-frequency spectrum (Aguadé *et al.* 1989; Braverman *et al.* 1995). However, both the reduction in polymorphism and the site-frequency skew are expected to dissipate with distance along the chromosome from the selected site because of recombination. This dissipation should produce a valley of minimal heterozygosity in the region containing the selected site, with decay on either side of the valley to neutral levels of heterozygosity (Maynard Smith and Haigh 1974; Kim and Stephan 2002).

The combination of positive selection and recombination produces distinctive patterns at sites that are linked to the selected mutation, but distant enough that at least one sequence in the sample has undergone a recombination event between the neutral and selected sites. In such a region, the genealogy will resemble a star phylogeny connected by a long branch to other lineages with a more neutral-looking set of coalescent relationships (Fay and Wu 2000). This particular topology results in an excess of rare polymorphic variants due to the long branches between the swept lineages and the recombined lineages. Furthermore, because of the relatively long branch connecting the swept alleles with the most recent common ancestor of the sample, many of these polymorphisms will be in the derived state as determined by comparison to an outgroup sequence.

[1]*Corresponding author:* Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138. E-mail: cmeiklejohn@oeb.harvard.edu

[2]*Present address:* Department of Biology, University of Rochester, Rochester, NY 14627.

The excess of high-frequency-derived variants should therefore be diagnostic of a region with partial linkage to a swept site.

These genealogical relationships will also affect the haplotype structure among sampled chromosomes. The relatively long branches between swept and recombined alleles should produce an excess of linkage disequilibrium, due to the large number of polymorphic variants that are fixed between the haplotype groups (Przeworski 2002). Furthermore, when a recombination event occurs early in the sweep, then two or more chromosomes, each with its own distinctive set of polymorphisms, may rapidly move to high frequency, producing multiple haplotype groups with little or no variation within each, but an excess of variants fixed between them. As with the patterns of heterozygosity and polymorphism frequencies, this pattern of haplotype structure is also expected to decay with distance from the selected site, and the observation of such blocks of linkage disequilibrium has been proposed as a method for mapping the recent action of positive selection from genome sequence data (Sabeti *et al.* 2002).

A region located within polytene band 99D on chromosome arm 3R in *Drosophila simulans* was recently shown to contain a pattern of nucleotide polymorphism consistent with the recent action of positive selection by the criteria outlined above (Parsch *et al.* 2001a). This region contains three paralogous, testes-expressed genes, *janusA* (*janA*), *janusB* (*janB*), and *ocnus* (*ocn*; Yanicostas *et al.* 1989; Parsch *et al.* 2001b). A comparison of DNA sequences sampled from a worldwide collection of eight *D. simulans* lines revealed that this region contains low levels of DNA polymorphism and an excess of high-frequency-derived alleles. In addition, this region showed strong haplotype structure, with a high-frequency haplotype group containing very little heterozygosity and a low-frequency haplotype group with levels of variation that are more typical for *D. simulans*. The previous study revealed a break in the haplotype structure located between *janB* and *ocn*; however, recombination events that could define the proximal limit of the haplotype structure were not observed, leaving the extent of the selected region in question. To further characterize the geographic and chromosomal extent of this haplotype structure, we report here a survey of DNA sequence polymorphism in a worldwide sample of haplotypes from *D. simulans* at the *jan-ocn* region and in the three *serendipity* (*sry*) genes, *sry*δ, *sry*α, and *sry*β, which are located just proximal to *janA*. The results indicate distinct recombination events disrupting the haplotype structure on either side of *janB*, suggesting that the target of positive selection lies in or near this gene. Application of a likelihood-ratio test for a selective sweep also localizes the selected site to *janB*. Because the presence of low-frequency, ancestral alleles throughout the putative selected region is inconsistent with a single, *completed* sweep, we have elaborated previously developed methods for detecting selective sweeps (Kim and Stephan 2002) to allow discrimination between the hypotheses of complete and incomplete sweeps. The results indicate that a partial-sweep model fits the data significantly better than a completed sweep and suggest the historical or current action of more complex evolutionary forces, such as balancing or epistatic selection, in this region of the genome.

MATERIALS AND METHODS

Fly stocks: *D. simulans* lines were provided by P. Capy and Y. Tao. All lines were derived from a single female and maintained by full-sib matings for >50 generations. The majority of lines were initially collected by Rama Singh in 1983, and a number of others are described in Atlan *et al.* (1997). This worldwide collection of isofemale lines reflects the history of collection by researchers more than any aspect of *D. simulans* population structure. All flies were raised on standard cornmeal agar medium at 25°.

DNA sequencing: Genomic DNA was extracted from a single male of each line as described previously (Parsch *et al.* 2001b). The *janA-ocn* region was PCR amplified as a single 2.4-kb fragment from genomic DNA using primers and conditions described previously (Parsch *et al.* 2001a). Four primer pairs designed using the published *D. melanogaster* genome sequence (Adams *et al.* 2000) were used to amplify a ~700-bp fragment from *sry*α, a ~900-bp fragment from *sry*β, and two overlapping fragments totaling ~1200 bp from *sry*δ. PCR products were used as sequencing templates following purification with the QIAquick PCR purification kit (QIAGEN, Valencia, CA) or after treatment with the SAP/EXO reagent (United States Biochemical, Cleveland). Gene-specific internal primers and the original amplification primers were used for sequencing with the BigDye 2.0 cycle sequencing kit (Applied Biosystems, Foster City, CA) following the manufacturer's protocol. Sequences were run on an ABI 3100 automated sequencer, and each fragment was sequenced at least once on both strands. DNA sequences have been submitted to GenBank under the accession numbers AY663111–AY663284.

DNA polymorphism and haplotype analysis: Nucleotide sequence data were extracted in Sequencher 4.1 (Gene Codes, Ann Arbor, MI) and aligned in Clustal X (Thompson *et al.* 1997). Nucleotide polymorphism, haplotype and recombination statistics, and tests of neutrality were calculated using DnaSP 3.99 (Rozas *et al.* 2003) and SITES (Hey and Wakeley 1997). The probabilities associated with Tajima's *D* (Tajima 1989), haplotype diversity, and Fu and Li's *D* and *F* statistics (Fu and Li 1993) were calculated using DnaSP 3.99. Fay and Wu's *H* statistic (Fay and Wu 2000) and associated probabilities were calculated using a program available from J. Fay, assuming no recombination and a probability of back-mutation of 0.03 (using *D. melanogaster* as the outgroup) or 0.05 (using *D. yakuba* as the outgroup). The haplotype test of Hudson *et al.* (1994) was calculated using a program provided by J. Braverman. This test compares the observed configuration of segregating sites among haplotypes with that expected under neutrality. This comparison is done by partitioning the data into two groups of sequences and determining the probability of observing *i* sequences with *j* or fewer segregating sites. Partitions that maximized *i* and minimized *j* were chosen and corrected for the *a posteriori* choice of *i* and *j* by including the probability of all possible configurations more extreme than that observed. Probabilities of summary statistics and of configurations of the haplotype test were determined from 10,000 random coalescent simulations with no recombination, conditioned on the observed number of segregating sites. Heteroge-

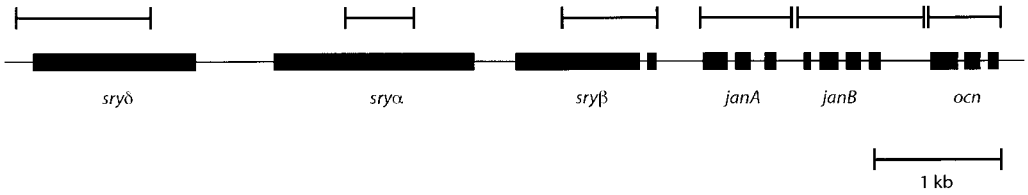

FIGURE 1.—Diagram of the portion of chromosomal band 99D studied here. Solid bars represent exons and intervening lines represent introns. Bars above represent sequenced regions.

neity in the ratio of polymorphism and divergence was assessed using the DNA Slider program (McDONALD 1998) with the published *D. melanogaster* sequence as an outgroup. Following the logic outlined by McDONALD (1996), 100 neutral simulations were run with the per-locus scaled population recombination parameter (R) set to 1, 2, 4, 8, 16, 32, and 64. The value(s) of R that maximized the probability of the data were then chosen, and the probability was recalculated with 10,000 simulations.

Composite-likelihood analysis: KIM and STEPHAN (2002) proposed a composite-likelihood method to detect the diagnostic features of a selective sweep from DNA sequence data. However, this method assumed that the population from which the sequences have been sampled is fixed for the putative beneficial mutation (*i.e.*, a complete sweep). Because we suspected that the pattern in the *jan-ocn* region was caused by the hitchhiking effect of a beneficial mutation that has not yet fixed (PARSCH *et al.* 2001a), we modified the method of KIM and STEPHAN (2002) to allow the frequency of the beneficial mutation at the time of sampling, β, to be less than one (*i.e.*, an incomplete sweep). The modified method proposes a new composite likelihood in which β is an additional parameter (APPENDIX).

For a given sample of DNA sequences, the maximum composite likelihood is obtained under three different models: the neutral model (L_0), the complete-sweep model (L_1), and the incomplete-sweep model (L_2). All three models assume that θ (the population mutation parameter $4N\mu$) and R_n (the scaled per-nucleotide recombination rate $4N\rho$) are known. L_1 is the maximum composite likelihood found by varying the location of the beneficial mutation, X, and the strength of selection, $2Ns = \alpha$, while setting $\beta = 1$. L_2 is obtained by allowing β to vary between 0 and 1, producing joint estimates of X, α, and β. These different hypotheses are tested using the likelihood ratios $LR_1 = \log(L_1/L_0)$ and $LR_2 = \log(L_2/L_1)$. To evaluate these likelihood ratios and examine the performance of the parameter estimation, maximum composite likelihoods were calculated for data sets simulated under various models. Simulations were conducted according to KIM and STEPHAN (2002) with the following modifications: if $\beta < 1.0$ in the selective sweep model, simulations were started with $n\beta$ sequences carrying the beneficial mutation and $n(1 - \beta)$ sequences carrying the ancestral allele. In other words, the ancestral recombination graph is constructed starting in the middle of the selective phase, with $k_B = n\beta$ and $k_b = n(1 - \beta)$ (the number of "B" and "b" edges, respectively; see KIM and STEPHAN 2002).

All simulations were conducted with the same structure as the sampled sequences in this study ($n = 36$ and six sequenced segments over a 7.8-kb region, as shown in Figure 1). For both simulation and composite-likelihood analyses, the sequence was divided into noncoding and coding regions, and the per-nucleotide mutation rate for each region was given as θ and 0.3θ, respectively, where θ is Watterson's estimator of the population mutation parameter calculated from the data (WATTERSON 1975). This approach is conservative for the likelihood-ratio test (KIM and STEPHAN 2002). Nonsynonymous and insertion/deletion polymorphisms were excluded from the analysis. Ancestral and derived alleles at polymorphic sites were identified by comparison with the published *D. melanogaster* sequence (ADAMS *et al.* 2000). The scaled per-nucleotide recombination parameter R_n was estimated as 0.065. This number was obtained from seven alleles of haplotype group II (PARSCH *et al.* 2001a) using the method of HUDSON (1987). We also estimated R_n using γ, a coalescent estimator of the population recombination rate (HEY and WAKELEY 1997), which gives values ranging from 0.009 (using all of the sequenced sites and all 36 lines) to 0.027 (using sites in the *janA-ocn* region with the same seven alleles as above). These are likely to be lower bounds on the true population recombination parameter, as γ under most circumstances underestimates the true R_n (WALL 2000). Estimates from the genetic map, using data from TRUE *et al.* (1996) and following the method of ANDOLFATTO and PRZEWORSKI (2000), give an R_n of 0.037 for interval 99D in *D. simulans*. We therefore included simulations and tests with $R_n = 0.005$, 0.03, and 0.1 to examine the effect of R_n on the composite-likelihood analysis, as these values represent a reasonable range of values for the true R_n in this region of the genome.

RESULTS

Reduced polymorphism and skewed site frequencies at 99D: We sequenced ~5 kb of a 7.5-kb region located within cytological band 99D on chromosome arm 3R from 36 lines of *D. simulans* (Figure 1). The configuration of segregating sites in this sample is shown in Figure 2. Polymorphism in this region is low relative to other loci that have been sequenced from chromosome arm 3R in *D. simulans* (BEGUN and WHITLEY 2000) and the other autosomes (MORIYAMA and POWELL 1996). Figure 3A depicts polymorphism scaled by divergence (π/K) calculated within a sliding window of 400 bp over the sampled region. Scaled polymorphism in this region is lowest at the 5′ end of *janB*. The ratio of polymorphism to divergence appears to be unusually heterogeneous across this region, as determined by a sliding window test that compares runs of consecutive polymorphic and fixed mutations with neutral coalescent simulations (McDONALD 1998). The number of runs of consecutive polymorphic or fixed sites across the 7.5 kb surveyed is significantly lower than expected under neutrality ($P = 0.0061$) and remains so ($P = 0.037$) following a Bonferroni correction for the multiple statistics that can be used to describe this heterogeneity (McDONALD 1998).

Departures from neutral genealogies will distort the spectrum of site frequencies found in DNA sequence data. Tajima's D-statistic (TAJIMA 1989) tests for deviations of this sort, with negative values indicating an excess of rare polymorphic variants, and positive values resulting from an excess of intermediate frequency vari-

ants. Tajima's *D* is significantly negative at *janB* and is negative, although not statistically significant, for five of the six genes and for all of the data combined (Table 1), indicating a deficiency of polymorphisms at intermediate frequency. Fu and Li have proposed tests of neutrality that compare the proportion of mutations found on the internal and external branches of a genealogy with the proportion expected under neutrality (Fu and Li 1993). Their *D*-statistic produces a significantly negative value at *janB* (Table 1), which results from the large number of singletons present in 10 of the 36 sequences (Figure 2). Figure 3B shows the values for Fu and Li's *D* calculated for a sliding window of 400 bp over the entire region. A localized segment with significantly negative values can be seen toward the 3′ end of *janB*, and similar results are obtained with Fu and Li's *F* (not shown). The remaining genes in this region show no significant departures from neutrality by these tests (Table 1). Thus, while the pattern of DNA sequence polymorphism and segregating site frequencies is unusual across the sampled region, these deviations from neutrality are strongest in the vicinity of *janB*.

Comparison with the *D. melanogaster* sequence reveals that, at a large fraction of the nonsingleton segregating polymorphic sites, the common allele is in the derived state (Figure 2). The presence of a high-frequency, derived haplotype with low levels of polymorphism was previously observed in the *janA-ocn* region in a subsample of the data presented here (Parsch *et al.* 2001a) and is a hallmark of a selective sweep (Fay and Wu 2000). When the *D. melanogaster* sequence is used as an outgroup, the *H*-test of Fay and Wu (2000) does not produce a value inconsistent with neutrality for any of the six genes or for the region as a whole (Table 1). This is due to the retention of relatively high amounts of genetic variation present in a minority of chromosomes (s18–s36 in Figure 2). However, if the more distantly related *D. yakuba* sequence is used to polarize the *D. simulans* polymorphisms as ancestral or derived, a significant *H*-statistic is obtained for all of the genes for which *D. yakuba* sequence data is available (Table 1). This is due to a number of sites where a rare polymorphism in *D. simulans* matches *D. yakuba* but not *D. melanogaster* (Figure 2, shaded sites). It is not clear whether this homoplasy can be attributed to multiple hits at these sites or to the retention of ancestral polymorphisms in the *D. simulans* and *D. melanogaster* lineages. Excluding all sites with conflicting outgroup information produces *H*-test results qualitatively similar to those obtained using the *D. melanogaster* sequence (not shown).

Composite-likelihood test of a selective sweep: To examine the ability of the new composite-likelihood method to detect partial sweeps, we applied the method to data sets simulated under an incomplete sweep model. Table 2 and Figure 4 show the results of parameter estimation for simulations where the frequency of the selected allele at the time of sampling was $\beta = 0.7$.

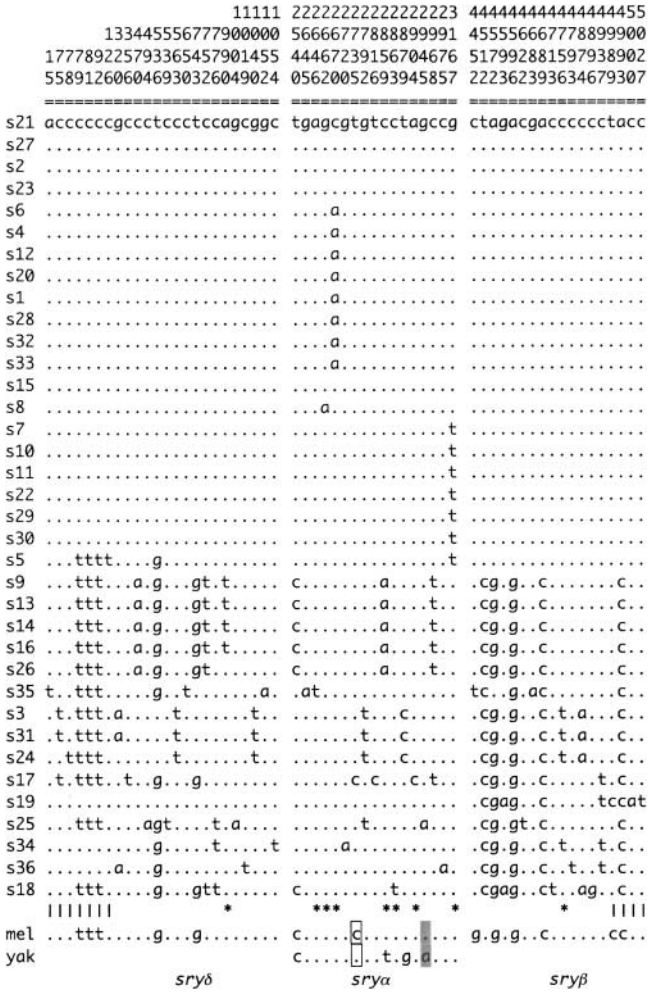

Figure 2.—Sequence data for six genes in region 99D. The *D. melanogaster* (mel) sequence is from Adams *et al.* (2000); the *D. yakuba* (yak) sequences are from Parsch *et al.* (2001b; *janA-ocn*) and Caccone *et al.* (1996; *sry*α). Asterisks below the sequences indicate nonsynonymous polymorphisms; vertical lines indicate noncoding polymorphisms. Boxed sites indicate that the rare *D. simulans* allele matches *D. melanogaster* and the common *D. simulans* allele matches *D. yakuba*; shaded sites indicate that the rare *D. simulans* allele matches *D. yakuba* and the common *D. simulans* allele matches *D. melanogaster*. Abbreviations for the location of origin of the *D. simulans* lines are: SA, South Africa; SM, St. Martin; JA, Japan; FR, France; TU, Tunisia; AU, Australia; HA, Haiti; US, United States; SE, Seychelles; PE, Peru; KE, Kenya; CO, Congo; PO, Polynesia; and ZI, Zimbabwe.

The method performs fairly well when estimating X and is more precise when $X = 6.3$ than when $X = 5$. This is due to the location of the gaps in the sequenced region (Figure 1). When a sweep occurs, it creates a valley of heterozygosity and skew in the site frequencies centered on the site of the beneficial mutation (Maynard Smith and Haigh 1974; Kim and Stephan 2002). The power to detect a sweep is dependent on the ability to observe sites that have experienced some recombination with the selected site, as these will produce the maximal skew in the site frequency spectrum. Detecting

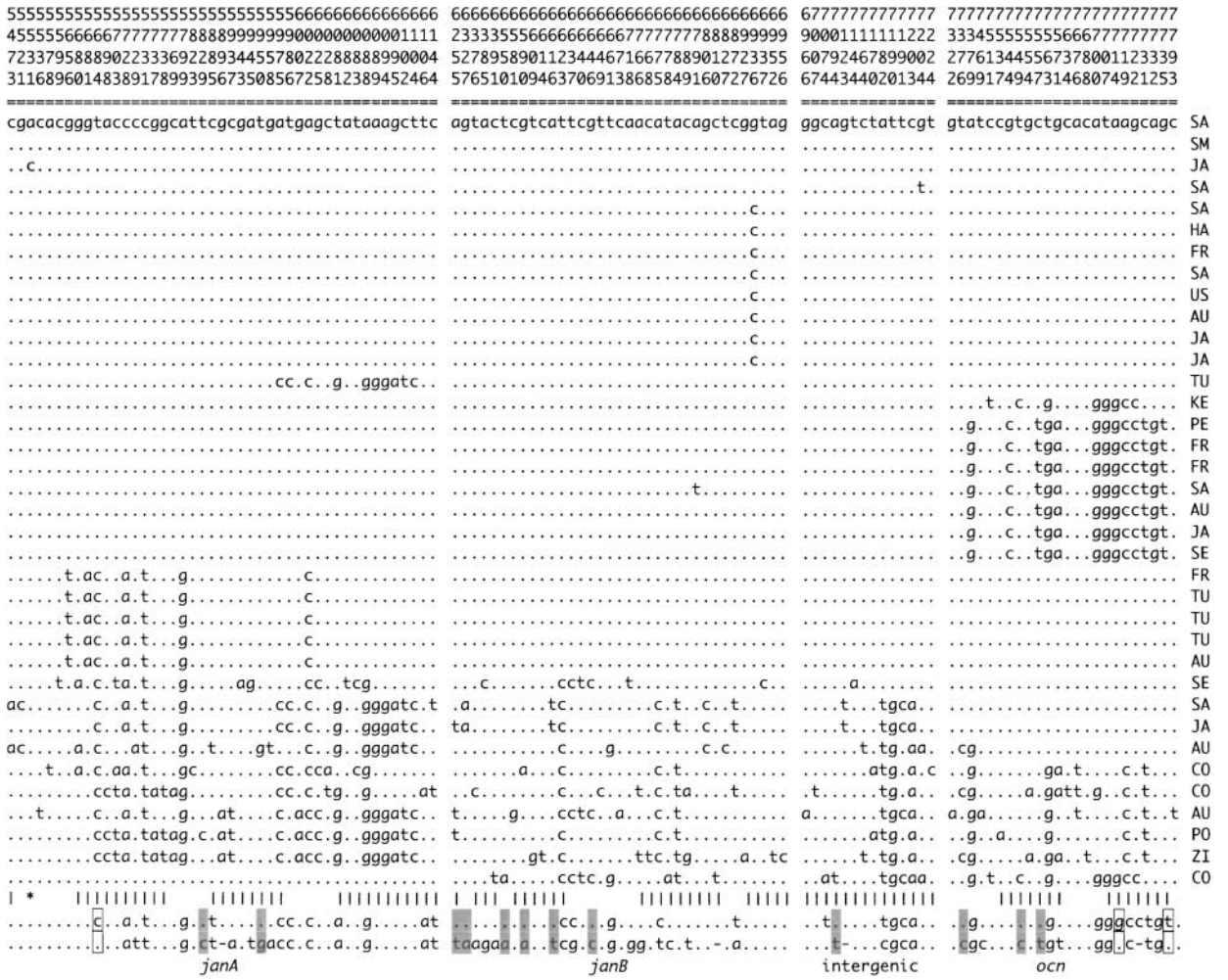

FIGURE 2.—*Continued.*

both "edges" of the valley of reduced heterozygosity should be more powerful than detecting only one edge. The sequenced regions from *sryB*, *janA*, *janB*, and *ocn* form an almost continuous stretch of sequences, with $X = 6.3$ near the center of this block (Figure 1). In contrast, $X = 5$ is located within the *sryB* sequence, and a simulated sweep located on $X = 5$ may contain an edge in a gap (*i.e.*, it may miss a recombination event).

However, the method produces poorer estimates of α and β. Mean estimates of α are upwardly biased due to a small fraction of simulated data sets that generate very large estimates, and the median $\hat{\alpha}$ seems to underestimate the true α (Figure 4B; Table 2). For many data sets, the maximization of L_2 leads to $\hat{\beta} = 1.0$, which therefore yields $LR_2 = 0$. From the joint distribution of $\hat{\alpha}$ and $\hat{\beta}$, it appears that a better fit is frequently given to a model of a complete sweep causing a reduction in variation over a small region, *i.e.*, an overestimate of β and an underestimate of α (Figure 4B). However, even excluding those simulations where $\hat{\beta} = 1.0$, $\hat{\beta}$ still overestimates β. Estimates of these parameters from data must therefore be interpreted cautiously.

We evaluated the power of this method to detect incomplete sweeps by comparing the distribution of LR_2 obtained from data sets simulated under the neutral, complete-sweep, and incomplete-sweep models (Table 3). To evaluate the ability of the method to distinguish partial sweeps from both neutral genealogies and complete-sweep simulations, null distributions of LR_2 were generated under complete sweeps as well as neutral equilibrium. In all cases, the cutoff for statistical significance was chosen to be the value of LR_2 generated under a null-distribution model (neutrality or a complete sweep) below which 95% of the null-distribution mass was contained. With $R_n = 0.065$ and across a range of values of α and X, the maximum 95% cutoff for LR_2 was <2.5 for neutral null distributions and <3.5 for completed-sweep null distributions (Table 3). Therefore, if the data produce an $LR_2 > 3.5$, we may reject either a neutral or a complete-sweep model in favor of an incomplete sweep. Using this threshold, the power to detect an incomplete sweep from simulated data (with $\beta = 0.7$) ranges from 0.013 to 0.855 depending the location and strength of selection (Table 3, cases 16–22) and increases with the

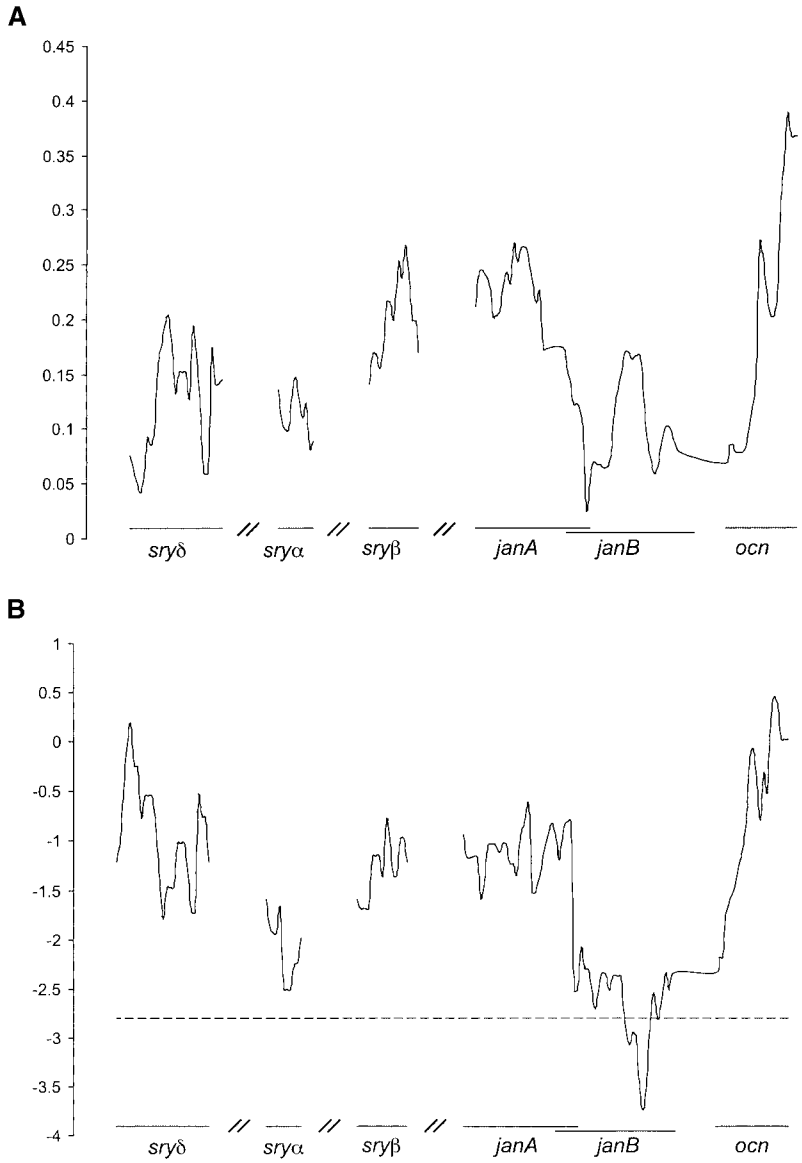

FIGURE 3.—Low polymorphism and excess of singletons at *janB*. Graphs were generated using DnaSP 3.99 (ROZAS *et al.* 2003) with a sliding window of 400 nucleotides and a step size of 25 nucleotides. (A) Average pairwise differences (TAJIMA 1983) divided by divergence. (B) Fu and Li's *D* (FU and LI 1993). The horizontal line indicates values of *D* that are significantly different from 0 at $P < 0.05$.

strength of selection. Inspection of Table 3 reveals that the power to detect a sweep (partial or complete) is greater for beneficial sites located at $X = 6.3$ than at $X = 5$. This is a result of the influence of the structure of the sequenced regions on the likelihood method described above.

Applying the complete-sweep composite-likelihood method to the data shown in Figure 2, and assuming $R_n = 0.065$, we obtain $LR_1 = 16.08$ ($\hat{\alpha} = 90.1$ and $\hat{X} = 7.01 \times 10^3$). Because the 99th percentile of LR_1 from the neutral simulations is 10.3, the neutral model is clearly rejected in favor of the complete sweep. However, the estimated strength of selection ($\hat{\alpha} = 90.1$) is too small to explain the unusual haplotype structure spanning the entire 7.8-kb region surveyed. A sweep is expected to influence variation only at linked neutral loci where the recombination fraction with the beneficial locus is $< s/2$ (MAYNARD SMITH and HAIGH 1974; STEPHAN *et al.* 1992). With $R_n = 0.065$, the region influenced by directional selection of this magnitude is not expected to extend beyond 1.4 kb in either direction from the selected site, and this expectation ranges from 1.2 to 1.6 kb over the values of R_n considered here. The complete linkage between sites as far apart as 4.6 kb (*i.e.*, sites 3167 and 7735) therefore suggests a stronger selective benefit associated with the favored site. This inference, as well as inconsistencies in the data with the predictions following a complete sweep, provide the impetus for a composite-likelihood-ratio (CLR) test of a partial sweep.

Applying the partial-sweep test to the data and again assuming $R_n = 0.065$, we obtain $LR_2 = 10.20$ ($\hat{\alpha} = 2.94 \times 10^4$, $\hat{X} = 7.20 \times 10^3$, and $\hat{\beta} = 0.60$). Because $LR_2 = 10.2 \gg Q_{0.95}$ derived from simulations of a complete sweep with a number of parameter values (Table 3), we can reject the complete sweep in favor of the incomplete-sweep model. In this case, $\hat{\alpha}$ is overestimated, as an incomplete sweep of this strength would completely

TABLE 1

Summary statistics for six genes in 99D

Gene	Sites[a]	π[a]	θ[a]	Tajima's D	Fu and Li's D	Fay and Wu's H[b]	hdiv	sub(i, j)
*sry*δ	1088 (317)	4.13 (13.1)	5.51 (17.6)	−0.86	−1.70	−0.64	0.687*	(21, 0)*
*sry*α	637 (155)	3.79 (8.16)	6.44 (15.6)	−1.36	−2.08	−1.29 (−1.64)	0.878	(21, 2)
*sry*β	740 (210)	5.01 (17.6)	5.87 (20.6)	−0.49	−1.23	−1.57	0.646*	(21, 0)*
janA	744 (410)	12.5 (21.7)	15.7 (27.0)	−0.60	−0.68	−5.49 (−9.37)	0.683**	(19, 0)**
janB	1027 (619)	6.67 (10.1)	13.2 (20.1)	−1.72*	−2.86**	−7.13 (−15.6*)	0.765**	(24, 1)**
ocn	573 (238)	10.8 (25.8)	10.2 (24.2)	0.23	−0.07	−4.72 (−6.78*)	0.633*	(21, 0)*
All	4809 (1949)	6.82 (15.0)	9.36 (20.8)	−0.97	−1.84	−20.90 (−31.74*)[c]	0.94*	(8, 0)*

π, average number of pairwise differences (Tajima 1983); θ, estimator of $4N\mu$ (Watterson 1975); hdiv, haplotype diversity (Nei 1987); sub(i, j), the most extreme subset of the sample, where i and j are the number of alleles and number of segregating sites in the subsample, respectively (Hudson *et al.* 1994). Significance levels were determined by 10,000 random coalescent simulations conditioned on the observed number of alleles and segregating sites and assuming no recombination. *janB* includes the region labeled "intergenic" in Figure 2. *$P < 0.05$; **$P < 0.01$.

[a] Values of π and θ were multiplied by 10^3; values in parentheses are for synonymous and noncoding sites only.

[b] Values in parentheses were determined using the *D. yakuba* sequence as an outgroup.

[c] *janA*, *janB*, and *ocn* only.

wipe out variation on affected chromosomes over a region >400 kb. This overestimation can likely be attributed to the ignorance of the correlation between polymorphic sites in the calculation of the composite likelihood. This likelihood is obtained by multiplying the probability of the observed frequency of derived alleles across sites, ignoring the correlation between polymorphic sites. An overestimation of α will then result from homogeneity in the frequency of derived alleles. If similar frequencies of derived alleles are present throughout the data, as seen in Figure 2, a model of a partial sweep driven by very strong selection is favored by the composite likelihood. In other words, the data in Figure 2 do not show a sufficient decay in the frequency spectrum skew in either direction to be compatible with a moderate α. This conclusion is supported by the profile of LR_2 as a function of X. Although there is a peak at $X = 7.10$ kb, the plot of LR_2 is almost flat over the entire region (the difference between multiple local optima is <0.5; data not shown). A much more limited footprint of a selective sweep is inferred from the changes of haplotype structure over this region (see below).

To examine the sensitivity of these results to our estimate of R_n, we repeated the analysis with $R_n = 0.005$, 0.03, and 0.1. Table 4 shows that the null distribution of LR_2 increases with decreasing R_n, and the same trend can be observed in the values of LR_2 obtained from the data. However, the empirical LR_2 is still $> Q_{0.95}$ for all values of R_n (Table 4). The profiles of LR_2 as a function of X for different recombination rates are also similar (data not shown). Therefore, the support for a partial-sweep model over a complete-sweep model and the inference of the location of the beneficial allele is largely insensitive to assumptions regarding R_n. The minimum estimated value of α (1.15×10^4) still appears to be greater than expected, on the basis of the observed extent of nucleotide variation.

While the composite-likelihood approach clearly supports the hypothesis of an incomplete sweep, it cannot identify which chromosomes carry the putative benefi-

TABLE 2

Composite-maximum-likelihood estimation of parameters applied to simulated data

α	X (kb)	$\hat{\alpha}$	$\hat{X}$ (kb)	$\hat{\beta}$
100	5	73 (21, 266)	5.02 (2.97, 6.81)	0.993 (0.766, 1)
500	5	381 (77, 1090)	5.11 (4.06, 6.26)	0.857 (0.694, 1)
2000	5	1754 (463, 5831)	5.24 (3.59, 6.89)	0.761 (0.666, 0.906)
100	6.3	82 (26, 257)	6.26 (4.86, 6.89)	0.918 (0.720, 1)
500	6.3	465 (117, 1190)	6.34 (5.71, 6.95)	0.796 (0.692, 0.983)
2000	6.3	1793 (318, 5960)	6.34 (5.03, 7.52)	0.758 (0.682, 0.853)
500[a]	5	447 (182, 853)	5.00 (4.71, 5.62)	—
500[a]	6.3	447 (119, 796)	6.30 (5.94, 6.73)	—
2000[a]	6.3	1359 (525, 2628)	6.29 (5.71, 6.99)	—

Median and (10%, 90%) values for each estimate are shown; $\beta = 0.7$. Results are based on 1000 simulations for each parameter set.

[a] Parameter values were estimated under the complete-sweep CLR method using subsets of chromosomes carrying a beneficial allele generated under incomplete-sweep simulations.

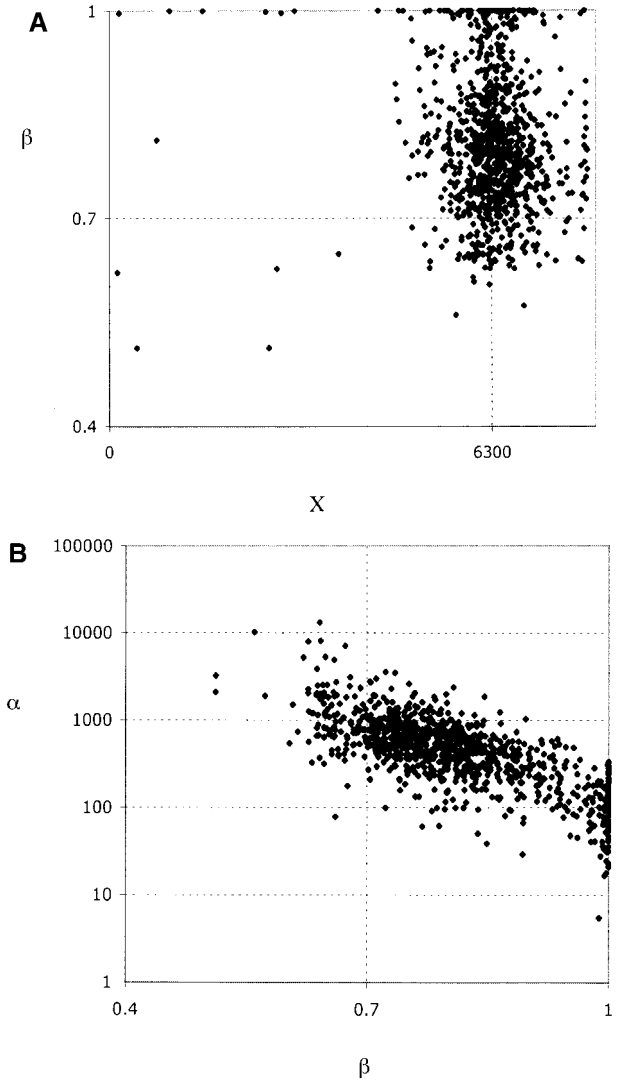

FIGURE 4.—Parameter estimation for data sets simulated under incomplete sweep with $\alpha = 500$, $X = 6300$, $\beta = 0.7$. (A) Joint distribution of $\hat{X}$ and $\hat{\beta}$. (B) Joint distribution of $\hat{\beta}$ and $\hat{\alpha}$.

cial mutation. Because the sequences carrying the beneficial mutations are descendants of a recent common ancestor, we may infer the haplotypes that have experienced the sweep by identifying a group of chromosomes within which diversity is greatly reduced. To identify this group, we selected a subsample of i chromosomes such that the estimated heterozygosity (π) within this group of chromosomes is minimized. This minimum is denoted $\pi_m(i)$. We then calculated the number of segregating sites for this subsample ($S(i)$) and graphed both $S(i)$ and $\pi_m(i)$ as a function of i (Figure 5). Figure 5 shows that $\pi_m(i)$ gradually decreases with decreasing i, as expected. On the other hand, $S(i)$ decreases rapidly and then reaches a plateau at $S(i) = 46$ from $i = 26$ to $i = 22$. This indicates that the 26 chromosomes chosen to minimize $\pi(i)$ contain a few sets of identical haplotypes, and this group of chromosomes is a good candidate for the one homogenized by the putative incomplete sweep. These 26 chromosomes are s21 to s26 in Figure 2, and we designate them as haplotype group I and the remaining 10 alleles as haplotype group II (Figure 2, s35 to s18). The second plateau in Figure 5 [$S(i) = 16$ from $i = 19$ to $i = 14$] contains the first 20 sequences in Figure 2, excluding s15. This subset is contained within the haplotype group I sequences, and so we designate it as haplotype group Ia. The haplotype group I sequences also maximize the linkage disequilibrium and the frequency of derived alleles (Figure 2), which are clear signatures of a selective sweep (Fay and Wu 2000; Kim and Stephan 2002; Przeworski 2002).

Given the wide confidence intervals associated with $\hat{\alpha}$ and $\hat{\beta}$ and the lack of resolution regarding the location of the selected site under the partial-sweep model, we reasoned that a complete-sweep CLR test using only those chromosomes likely to have been involved in the sweep might provide more accurate estimates of α and X. Theoretical work indicates that the frequency of a neutral allele conditional on its linkage to a beneficial mutation changes only slightly during the period when the frequency of the beneficial mutation increases from 0.5 to 1 (Stephan *et al.* 1992). In other words, the hitchhiking effect on neutral loci is mainly determined when the frequency of the beneficial mutation is low. The frequency of the putative beneficial mutation is likely to be >0.5, considering that the likelihood estimate of $\beta = 0.6$, and the frequencies of haplotype groups I and Ia are 0.72 and 0.53, respectively. This means that the skew in site frequencies among swept chromosomes under a partial sweep should be similar to that among all chromosomes following a complete sweep (see materials and methods) and justifies the application of a CLR test of a complete sweep to a subsample that is assumed to be in complete association with the beneficial mutation.

This reasoning is borne out by applying the complete-sweep CLR method to the subset of chromosomes carrying the beneficial allele in data sets simulated under an incomplete-sweep model (Table 2). Although the median $\hat{\alpha}$ sometimes more severely underestimates the true value, the median $\hat{X}$ is closer to the actual location of the selected site in all cases. In addition, the 10–90% range of estimated parameters is on average twofold narrower when using only the swept chromosomes than when using all of the sequences.

When the CLR test of a complete sweep is applied to the 19 haplotype group Ia sequences, a nonsignificant result is obtained ($R_n = 0.065$, $\hat{\theta}_W = 0.00196$; $LR_1 = 3.54$, $P < 0.057$). However, a significant CLR is obtained for the 26 haplotype group I sequences ($R_n = 0.065$, $\hat{\theta}_W = 0.0052$; $LR_1 = 16.6$, $P < 0.001$). We propose, therefore, that all of the haplotype group I chromosomes, rather than just haplotype group Ia, represent the best candidates for the partially swept haplotype. Under this model, the beneficial mutation is estimated

TABLE 3

Distributions of LR_2 obtained from simulations

Case	R_n	α	X (kb)	β	$Q_{0.05}$[a]	$Q_{0.2}$	$Q_{0.5}$	$Q_{0.8}$	$Q_{0.95}$
1	0.005	0[b]	—	—	0[c]	0.27	1.55	4.40	9.22
2	0.03	0[b]	—	—	0	0	0.26	1.32	3.76
3	0.065	0[b]	—	—	0	0	0.12	0.86	2.36
4	0.1	0[b]	—	—	0	0	0.06	0.71	1.79
5	0.065	50	2	1.0	0	0	0.08	0.66	1.82
6	0.065	100	2	1.0	0	0	0.07	0.72	1.97
7	0.065	500	2	1.0	0	0	0	0.34	1.45
8	0.065	1,000	2	1.0	0	0	0.01	0.81	2.73
9	0.065	50	7	1.0	0	0	0.17	1.39	3.28
10	0.065	100	7	1.0	0	0	0.09	0.99	2.86
11	0.005	500	7	1.0	0	0	0.62	2.40	5.67
12	0.03	500	7	1.0	0	0	0.01	0.92	4.11
13	0.065	500	7	1.0	0	0	0.01	0.61	3.09
14	0.1	500	7	1.0	0	0	0.01	0.39	2.51
15	0.065	1,000	7	1.0	0	0	0	0.29	2.71
16	0.065	100	5	0.7	0	0	0.11	0.74	1.99
17	0.065	500	5	0.7	0	0.01	0.73	3.18	7.00
18	0.065	2,000	5	0.7	0.01	1.59	5.90	11.2	16.7
19	0.065	100	6.3	0.7	0	0.01	0.63	2.45	4.89
20	0.065	500	6.3	0.7	0.02	0.99	4.21	8.07	12.6
21	0.065	1,000	6.3	0.7	0.29	2.52	6.53	11.4	16.5
22	0.065	2,000	6.3	0.7	1.68	4.63	9.10	13.8	19.0
23	0.065	10,000	5	1.0	0	0	0	0.36	2.78
24	0.065	10,000	7	1.0	0	0	0	0.99	6.12
25	0.065	14,000	6.3	1.0	0	0	0	0.31	2.05

All simulations were done with $\theta = 0.02$.
[a] Q_x is the xth percentile of LR_2 from 1000 simulations.
[b] Simulation under the neutral model.
[c] "0" means <0.01.

to be located near the 5′ end of *janB* (Figure 6; $\hat{X}$ = 6.39 kb) and its estimated strength is $\hat{\alpha} = 455$. Figure 6 shows that there are many local optima along the sequence. The difference between the highest and the second highest (located between *sry*α and *sry*β) peaks is ~2.4 CLR units. Therefore one may argue that the putative beneficial mutation is ~11 ($\approx e^{2.4}$) times more likely to be in or near the *janB* gene than between *sry*α and *sry*β.

Inferring the location of the selected region from haplotype structure: The inference of a selected site at *janB* is further supported by the pattern of linkage disequilibrium (LD), which the composite-likelihood analysis does not take into account. DNA polymorphism across this region is clearly grouped into distinct haplotypes. A number of alleles are identical (or nearly so) in their combination of polymorphic variants across the entire region (*e.g.*, s9, s13, s14, s16, and s26), and the

TABLE 4

Influence of R_n on composite-likelihood methods

	Complete-sweep model				Incomplete-sweep model				
R_n	$Q_{0.95}$[a]	LR_1	$\hat{\alpha}$	$\hat{X}$[c]	$Q_{0.95}$[b]	LR_2	$\hat{\alpha}$[c]	$\hat{X}$[c]	$\hat{\beta}$
0.005	9.22	14.88	5.81	6.45	5.67	11.4	14.3	7.07	0.60
0.030	3.76	14.69	47.2	7.02	4.11	11.62	11.5	7.09	0.59
0.065	2.36	16.08	90.1	7.01	3.09	10.20	29.4	7.20	0.60
0.100	1.79	16.14	129	7.02	2.51	10.13	29.3	7.09	0.60

[a] Neutral simulations.
[b] Complete-sweep simulations; $\alpha = 500$.
[c] Values were multiplied by 10^{-3}.

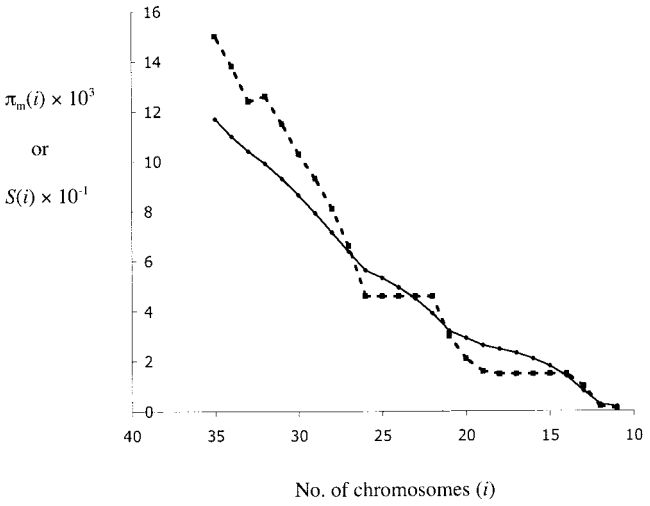

FIGURE 5.—Average number of pairwise differences (π, solid line) and number of segregating sites (S, dashed line) for subsets of chromosomes that minimize π ($\pi_m(i)$), graphed against the number of chromosomes in each subset. See text for details.

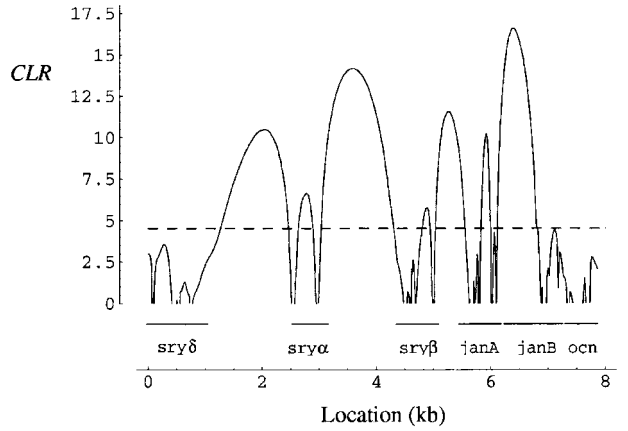

FIGURE 6.—The composite-likelihood ratio (CLR) as a function of the position of the putative beneficial mutation. Sequenced segments corresponding to six genes in this region are indicated by horizontal lines above the *x*-axis. The CLR was obtained from 26 chromosomes corresponding to haplotype group I. The dashed line represents the 95th percentile of CLR (4.52) determined by neutral simulations.

majority of the low-frequency and singleton variants are contained entirely within the 10 haplotype group II chromosomes. There is a significant reduction in haplotype diversity as compared with a neutral genealogy at all genes except *sry*α and across the region as a whole (Table 1). The same result is found with a haplotype test (HUDSON *et al.* 1994) that partitions the data into two groups and compares the number of segregating sites in each group with that expected under neutrality (Table 1). Consistent with the polymorphism data and the composite-likelihood results, the most extreme haplotype structure is also found at *janB*.

Because the duration of a sweep is expected to be very short, there should be limited opportunities for recombination events during the selective phase. Because recombination breakpoints on either side of the beneficial mutation should occur independently, observing a stretch of LD that extends across the site of the beneficial mutation is very improbable. Simulations confirm that an excess of LD is observed on both sides of a selected site but that the association is broken between the two sides (KIM and NIELSEN 2004). The extent of a genomic region affected by positive selection may also be inferred from the location of recombination events between derived (selected) and ancestral alleles. Derived polymorphisms are expected to be at their highest frequency near to the selected variant, and their frequency should decrease due to recombination as one moves away from the selected site in either direction (FAY and WU 2000; KIM and STEPHAN 2002). Note, however, that a region in complete linkage with a selected site that has recently gone to fixation should have *no* high-frequency-derived variants, as all polymorphism will be the result of new mutations.

Within subsets of sequences in haplotype group I, there are two stretches of ancestral polymorphic sites in complete LD (Figure 2). Alleles s9, s13, s14, s16, and s26 match the ancestral (*D. melanogaster*) haplotype at 15 sites ranging from position 79 in *sry*δ to position 6022 in *janA*, but then match the derived haplotype over the entire *janB-ocn* region. Similarly, alleles s7, s10, s11, s22, s29, and s30 match the derived haplotype at all but 2 sites over the *sry*δ-*janB* region, but then match the ancestral haplotype at 9 sites in the *ocn* gene. These two inferred recombination events both disrupt haplotype associations across *janB* and decrease the frequency of derived alleles on either side. The sequence of strain s5 is of particular interest in this context, as it shows evidence of recombination events on both sides of *janB* (*i.e.*, sites 536 and 7379 in Figure 2) that bring the derived sequence onto an ancestral haplotype both proximal and distal to *janB*. The localization of the maximal frequency of derived variants and the disruption of LD to the same region are further evidence for a partial, rather than a complete, selective sweep in this region. On the basis of the observed recombinants, the selected site is postulated to lie between position 6022 in *janA* and position 7379 in *ocn*, which is consistent with the estimates of $\hat{X} = 6390$, 7010, and 7200 from the partial- and complete-sweep likelihoods. It is noteworthy that, following recombination, a number of ancestral segments that became linked to the derived allele in the putative selected region appear to have increased in frequency themselves. For the proximally recombined alleles, the recombinants are in a frequency of 5/36; for the distally recombined alleles, the recombinants are in a frequency of 7/36. This could be the result of hitchhiking of ancestral segments that recombined early in the sweep onto a positively selected chromosome.

Estimation of the age of haplotype group I: The age of a recently derived haplotype group can be estimated on the basis of the number of new mutations that have occurred among the alleles since they last shared a common ancestor (Rozas *et al.* 2001). In the case of *janB*, there are 26 haplotype group I alleles with three new mutations (sites 6791, 6936, and 7204). Assuming a star phylogeny (as is expected following a selective sweep or population bottleneck), there are 26 branches of length t on which mutations can occur. If the number of mutations follows a Poisson distribution, then the expected number of mutations in the *janB* sample is $26t\mu$, where μ is the mutation rate per sequence per year. On the basis of the observed silent site divergence (synonymous and noncoding sites) at *janB* of 0.14 between *D. melanogaster* and *D. simulans*, and assuming a divergence time of 2.5 million years for these two species (Lachaise *et al.* 1988; Hey and Kliman 1993), μ is estimated to be 1.73×10^{-5}. The probability of observing three mutations is then

$$P(S = 3|t) = \frac{(26\mu t)^3}{3!} e^{-26\mu t}$$

and the maximum-likelihood estimate of t is 6667 years. Ninety-five percent confidence intervals for this age can be calculated by finding t_{max} and t_{min} such that $P(S \leq 3 \mid t_{max}) = 0.025$ and $P(S \geq 3 \mid t_{min}) = 0.025$. This produces the values $t_{min} = 1375$ and $t_{max} = 19{,}481$ years. The estimate of t should be taken as an approximate lower bound, as the 26 sequences in haplotype group I do not conform to the assumption of a star phylogeny.

An excess of nonsynonymous fixed differences: Recurrent positive selection on amino acid substitutions at a locus should result in an excessively low ratio of nonsynonymous intraspecific polymorphisms to nonsynonymous fixed interspecific differences, relative to the analogous ratio of synonymous substitutions (McDonald and Kreitman 1991). The number of each type of substitution for each gene and for the region as a whole is presented in Table 5, along with the neutrality index (Rand and Kann 1996), which is a ratio of the two ratios described above. A neutrality index less than one indicates a relative paucity of nonsynonymous polymorphisms. Five of the six genes in this region have an excess of fixed replacement substitutions between *D. simulans* and *D. melanogaster*, and this excess is statistically significant at *janB* and *sry*δ, as well as for the region as a whole. This result does not appear to be the result of an excess of unpreferred synonymous substitutions segregating within *D. simulans*, as there is no detectable difference in the frequency distribution of preferred to unpreferred *vs.* unpreferred to preferred changes (Akashi 1997; sites were pooled across all six loci, Mann-Whitney *U*-test, $z = 0.907$, $P > 0.05$). Using the *D. yakuba* sequence to polarize fixed differences to the *D. simulans* or *D. melanogaster* lineages results in nonsignificant McDonald-Kreitman tests for *janA*, *janB*, and *ocn* individually due to small sample sizes (not shown). However, pooling data across these three loci gives a significant excess of replacement fixations along the *D. simulans* lineage (Table 5). These results suggest that one (or more) of these genes has been a target of positive selection along the *D. simulans* lineage since its divergence from *D. melanogaster*.

TABLE 5

McDonald-Kreitman tests

Gene	D_S	P_S	D_N	P_N	N.I.	*P*-value
*sry*δ	15	16	8	1	0.12	0.021
*sry*α	18	10	3	7	4.2	0.060
*sry*β	12	13	4	1	0.23	0.176
janA	5	15	3	1	0.11	0.059
janB	11	11	7	0	0.00	0.026
ocn	6	10	2	0	0.00	0.183
All	67	75	31	10	0.29	0.001
janA-ocn[a]	8	36	6	1	0.04	<0.001

D_S, number of fixed synonymous substitutions; P_S, number of polymorphic synonymous substitutions; D_N, number of fixed nonsynonymous substitutions; P_N, number of polymorphic nonsynonymous substitutions; N.I., neutrality index (Rand and Kann 1996). *P*-values were calculated by a *G*-test except for *janB* and *ocn*, which were calculated by Fisher's exact test.

[a] Mutations were polarized to the *D. simulans* lineage using the *D. yakuba* sequence as an outgroup.

DISCUSSION

Theoretical studies have explored the effects of positive selection on heterozygosity (Maynard Smith and Haigh 1974; Kaplan *et al.* 1988, 1989), the distribution of segregating site frequencies (Braverman *et al.* 1995; Kim and Stephan 2002), the fraction of sites that are singletons (Fu and Li 1993), and the haplotype structure of linked neutral variation (Przeworski 2002; Wall *et al.* 2002). An excess of rare polymorphisms following a sweep is expected either due to the recovery of genetic variation on the star-like genealogy that results from the abrupt coalescence of all lineages during the sweep (Aguadé *et al.* 1989) or due to the persistence of ancestral mutations segregating on a limited number of recombined lineages some distance from the selected site (Fay and Wu 2000). In the data presented here, the rare variants are a combination of ancestral and derived polymorphisms retained in the haplotype group II sequences. Despite the presence of high-frequency-derived polymorphisms that are consistent with a sweep, nonsignificant values of Fay and Wu's *H* are obtained when *D. melanogaster* is used as an outgroup (Table 1). We infer this to be due to the presence of many low-frequency, derived variants among the haplotype group II alleles. These variants counteract the high-frequency, derived variants of haplotype group I, resulting in a nonsignificant value of *H*. Our interpretation is that this retention of a few haplotypes with

relatively high heterozygosity is the result of linkage to a beneficial mutation bringing one or a few alleles to high frequency, but not to fixation (*i.e.*, an incomplete sweep).

This interpretation is supported by the CLR analysis, which clearly indicates that the pattern of site frequencies across this region is better explained by a model of a complete sweep than by a neutral model and better explained by a model of a partial sweep than by one of a complete sweep. This result is robust to mis-estimation of the true recombination rate (Table 4). However, we are much less confident in our estimation of the parameters associated with this putative partial sweep. The parameter that is probably most accurately estimated is *X*, which shows the lowest bias and narrowest range in simulations (Table 2). *X* ranges from 6.39 to 7.20 kb across the various analyses, and this is consistent with the observed pattern of haplotype structure. Given the performance of the composite likelihood in estimating α, the sensitivity of these parameter estimates to R_n (Table 4), and the inconsistency between some of these estimates and the observed haplotype structure, it is difficult to say anything definitive about the strength of selection. Similarly, consistent overestimation of β and covariation between $\hat{\alpha}$ and $\hat{\beta}$ (Table 2, Figure 4) indicates that we cannot say much about the current frequency of the selected allele, except that β is most likely >0.5. Uncertainty regarding β also results from the discrepancy between the assumption of a single population used in the CLR analysis and the worldwide sample of lines used here.

As mentioned above, the presence of the haplotype group II chromosomes suggests either that the favored allele has not been fixed in a number of populations or that there is another explanation for the observed pattern of DNA sequence variation. It should be noted that the influence of a selective sweep on the site frequency spectrum is highly stochastic (for example, Figure 3 in Kim and Stephan 2002), and that this may be exacerbated by the nonrandom sampling of isofemale lines from various locations around the world. As a result, the gaps between the sequenced regions at the *sry* loci, or a locus outside of the sequenced region, could potentially harbor the selected site and show a pattern of nucleotide polymorphism more consistent with a completed sweep. However, none of these gaps is >1.5 kb, and the largest gap is the most proximal one (between *sry*δ and *sry*α). *sry*δ and *sry*α have the highest-frequency of haplotype group II sequences (40–50% for each). Given that a region more strongly affected would presumably be entirely (or almost so) haplotype group I sequences, this would require an extremely localized segment with this signature. A different study of nucleotide polymorphism in this region of the genome in *D. simulans* (Quesada *et al.* 2003) provides some evidence against these possibilities (see below). Another alternate explanation could be population structure—demographic history combined with restricted gene flow can produce large blocks of linkage disequilibrium and may be the cause of a genome-wide reduction from expected levels of recombination in Drosophila (Wall *et al.* 2002). However, three observations make it unlikely that demographic forces could have produced the pattern observed here. First, there is no obvious geographic pattern to the presence or frequency of the rarer haplotype group II sequences. Of the six geographic regions with more than two representative lines, at least three contain alleles of both haplotype groups. Second, a very similar pattern consisting of a single haplotype with low polymorphism has been observed at the *rp49* gene (which is immediately proximal to *sry*δ) in an independent sample of *D. simulans* (Rozas *et al.* 2001). In this study, a haplotype with very low heterozygosity (referred to here as haplotype group I for consistency) was observed at intermediate frequency within populations sampled from Spain and Mozambique. Estimation of the age of the haplotype group I alleles in these samples produced an estimate of ~6000 years (Rozas *et al.* 2001), which is consistent with our results. The persistence of haplotype group II alleles at similar frequencies in populations from two continents and in locations as geographically diverse as South Africa, Australia, and Japan is difficult to reconcile with a simple model of gene flow.

Third, there is strong evidence that this haplotype structure is restricted to this region of the genome. In a subsequent study, sequences from the Spanish and Mozambique *D. simulans* lines were sampled at intervals up to 35 kb away on either side of the *jan-ocn/rp49* region (Quesada *et al.* 2003). These data show convincingly that the reduction in polymorphism and unusual haplotype structure decay with increasing distance in either direction from *jan-ocn/rp49*. Furthermore, the observation of a reduction in haplotype structure from *sry*β to *sry*δ and beyond is similar to the one presented here and is additional evidence against a selected site located in a gap or outside the *sry-ocn* region. The observation of a valley of minimal heterozygosity and maximal haplotype structure, coupled with the retention of haplotype group II lineages (with normal levels of variation and linkage disequilibrium) throughout, led these authors to similarly conclude that a partial selective sweep was a plausible explanation for their observations (Quesada *et al.* 2003).

The CLR analysis indicates that the selected site is most likely somewhere in the vicinity of *janB*. However, there is not an obvious candidate for the selected mutation within this region. All of the polymorphic sites in *janA*, *janB*, and *ocn* are silent (with the exception of the singleton polymorphism in s2 at the 5′ end of *janA*), indicating that any selected site in this region must be regulatory in nature. One candidate region that might harbor important regulatory variants is the portion of the *janA* 3′-UTR that overlaps with the *janB* 5′-UTR and has been shown to be sufficient to regulate transcription

of *janB* and restrict translation to postmeiotic spermatids (Yanicostas and Lepesant 1990). However, none of the intermediate-frequency polymorphisms in the *janA* 3′-UTR (*e.g.*, those at positions 6081–6104 in Figure 2) lie in this region, and the most 5′ segregating site within the *janB* transcript that is fixed between the two haplotype groups is in the second exon (Figure 2, site 6623). If the selected site lies within the *janA* 3′-UTR, this would highlight the stretch of ancestral polymorphisms present in an otherwise haplotype group I chromosome, s15, which is likely the result of a gene conversion event. Such a gene conversion event between haplotype I and II sequences would be more likely to have occurred after the swept chromosomes reached an equilibrium frequency.

If this is indeed a partial selective sweep, what is maintaining the presence of the haplotype group II sequences? One possible explanation is that the sweep is ongoing, and that the new haplotype is destined for fixation. Assuming an effective population size for *D. simulans* of 2×10^6 and 10 generations per year, the transit time to fixation for an allele with a selective benefit $2Ns = 455$ could be between 10,000 and 20,000 years (Stephan *et al.* 1992), so catching this allele *in flagrante delicto* is not inconceivable and is consistent with the estimate of ~7000 years for the age of haplotype group I. However, if the sweep were ongoing, one might expect it to have gone to fixation in geographical regions closer to the origin of the selected mutation, rather than the consistently intermediate frequencies observed here across the world and in geographically disparate populations (Rozas *et al.* 2001).

An alternate explanation is that the fixation of the selected allele is inhibited by the presence of another beneficial mutation segregating nearby on a haplotype group II background, and both mutations must wait to fix until recombination can bring them together (Kirby and Stephan 1996). Given the high density of genes and the evidence for a history of positive selection in this region (Table 5), this might seem a plausible explanation. However, the patterns of variation at this locus and further away on chromosome arm 3R do not support this hypothesis. There is no evidence, for example, for a recently selected mutation in any of the three *sry* genes, or *rp49* (Rozas *et al.* 2001), as determined by haplotype structure and recombination breakpoints. Additionally, such a "traffic" model predicts that at some distance from *janB*, haplotype group II alleles would display a region of low polymorphism and strong haplotype structure in the neighborhood of the other selected site. None of the sequences sampled in either direction from *janB* show evidence for another recently derived haplotype that could be competing with haplotype group I (Quesada *et al.* 2003).

Finally, consider the possibility that the swept allele is not destined for fixation, but rather that it is nearing or has reached an equilibrium frequency determined by balancing selection at *janB* or epistatic selection between *janB* and an allele at another locus. The evidence for positive selection having acted on the genes in this study (Parsch *et al.* 2001b); and the expression of *janA*, *janB*, and *ocn* in testes (Yanicostas *et al.* 1989; Parsch *et al.* 2001b); is consistent with the observation of elevated rates of sequence evolution among genes associated with sex and reproduction in a wide range of taxa (Civetta and Singh 1998; Swanson and Vacquier 2002). A number of phenomena associated with male reproduction can also lead to stable polymorphisms, such as meiotic drive (Charlesworth and Hartl 1978), sperm competition (Clark *et al.* 1999), and sexually antagonistic pleiotropy (Rice 1984). Further functional analysis of the phenotypic differences between the genetic variants at these loci will be required to discriminate between these and other selective hypotheses.

The number of reports in the literature that invoke positive selection from DNA sequence data has become impressive of late (see Andolfatto and Przeworski 2001 and Przeworski 2002 for references). This alternative to the neutral (Kimura 1968) and mildly deleterious (Ohta and Kimura 1971) theories of molecular evolution is further supported by recent multilocus analyses suggesting that the fraction of interspecific substitutions driven by positive selection may be substantial (Fay *et al.* 2002; Smith and Eyre-Walker 2002), at least among Drosophila nuclear loci (Weinreich and Rand 2000; Bustamante *et al.* 2002). The large number of reported sweeps has prompted the reevaluation of the ability of current methods to discriminate positive selection from other forces shaping nucleotide variation; these studies suggest that migration and complex demographic history may explain part of the pattern (Przeworski 2002; Wall *et al.* 2002). However, more complex models of selection may also be necessary. While there are reports compatible with a recently completed selective sweep (*e.g.*, Nachman and Crowell 2000; Schlenke and Begun 2004), a significant number of studies that infer positive selection have found haplotype patterns suggestive of positive selection in conjunction with balancing or epistatic selection (Hudson *et al.* 1994; Kirby and Stephan 1996; Cirera and Aguadé 1997; Andolfatto *et al.* 1999; Benassi *et al.* 1999—but see Andolfatto *et al.* 1999 for possible reinterpretations of the statistical significance of some of the haplotype structures in these studies). This may indicate an important role for positive selection in shaping patterns of genetic divergence and also a significant contribution of epistatic and balancing selective forces to the maintainence of natural genetic variation (Lewontin 1974; Zapata *et al.* 2002). Systematic studies of nucleotide sequence data collected from natural populations without the biases associated with the studies of single loci will be instrumental in resolving this question.

We thank Sylvain Mousset for the estimate of the age of haplotype group I, Pierre Capy and Yun Tao for providing *D. simulans* stocks, and John Braverman for giving us computer programs. Comments from Rob Kulathinal, Daniel Weinreich, and Justin Blumenstiel greatly improved the article. Funding for this work was provided by National Institutes of Health grant GM60035 to D.L.H. and funds from the University of Munich to J.P. Y.K. was supported by funds from National Science Foundation grant DEB-0089487 to Rasmus Nielsen.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

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.

Akashi, H., 1997 Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205:** 269–278.

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

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., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

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

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

Benassi, V., F. Depaulis, G. K. Meghlaoui and M. Veuille, 1999 Partial sweeping of variation at the *Fbp2* locus in a West African population of *Drosophila melanogaster*. Mol. Biol. Evol. **16:** 347–353.

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.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in *Arabidopsis*. Nature **416:** 531–534.

Caccone, A., E. N. Moriyama, J. M. Gleason, L. Nigro and J. R. Powell, 1996 A molecular phylogeny for the *Drosophila melanogaster* subgroup and the problem of polymorphism data. Mol. Biol. Evol. **13:** 1224–1232.

Charlesworth, B., and D. L. Hartl, 1978 Population dynamics of the segregation distorter polymorphism of *Drosophila melanogaster*. Genetics **89:** 171–192.

Cirera, S., and M. Aguadé, 1997 Evolutionary history of the sex-peptide (*Acp70A*) gene region in *Drosophila melanogaster*. Genetics **147:** 189–197.

Civetta, A., and R. S. Singh, 1998 Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15:** 901–909.

Clark, A. G., D. J. Begun and T. Prout, 1999 Female X male interactions in *Drosophila* sperm competition. Science **283:** 217–220.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fay, J. C., G. L. Wyckoff and C.-I Wu, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature **415:** 1024–1026.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

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

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

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

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

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

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

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

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513–1524.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kimura, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–225.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

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

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

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

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

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

Nachman, M. W., and S. L. Crowell, 2000 Contrasting evolutionary histories of two introns of the Duchenne Muscular Dystrophy gene, *Dmd*, in humans. Genetics **155:** 1855–1864.

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

Ohta, T., and M. Kimura, 1971 On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. **1:** 18–25.

Parsch, J., C. D. Meiklejohn and D. L. Hartl, 2001a Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics **159:** 647–657.

Parsch, J., C. D. Meiklejohn, E. Hauschteck-Jungen, P. Hunziker and D. L. Hartl, 2001b Molecular evolution of the *ocnus* and *janus* genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **18:** 801–811.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.

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.

Rand, D. M., and L. M. Kann, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. **13:** 735–748.

Rice, W. R., 1984 Sex chromosomes and the evolution of sexual dimorphism. Evolution **38:** 735–742.

Rozas, J., M. Gullaud, G. Blandin and M. Aguadé, 2001 DNA variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics **158:** 1147–1155.

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

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Schlenke, T. A., and D. J. Begun, 2004 Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **101:** 1626–1631.

Smith, N. G. C., and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature **415:** 1022–1024.
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.
Swanson, W. J., and V. D. Vacquier, 2002 Reproductive protein evolution. Annu. Rev. Ecol. Syst. **33:** 161–179.
Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.
True, J. R., J. M. Mercer and C. C. Laurie, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics **142:** 507–523.
Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.
Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.
Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.
Weinreich, D. M., and D. M. Rand, 2000 Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. Genetics **156:** 385–399.
Yanicostas, C., and J.-A. Lepesant, 1990 Transcriptional and translational *cis*-regulatory sequences of the spermatocyte-specific *Drosophila janusB* gene are located in the 3′ exonic region of the overlapping *janusA* gene. Mol. Gen. Genet. **224:** 450–458.
Yanicostas, C., A. Vincent and J.-A. Lepesant, 1989 Transcriptional and posttranscriptional regulation contributes to the sex-regulated expression of two sequence-related genes at the janus locus of *Drosophila melanogaster*. Mol. Cell. Biol. **9:** 2526–2535.
Zapata, C., C. Nuñez and T. Velasco, 2002 Distribution of nonrandom associations between pairs of protein loci along the third chromosome of *Drosophila melanogaster*. Genetics **161:** 1539–1550.

Communicating editor: M. Veuille

APPENDIX

Assume that a beneficial mutation arises at a site located near a neutral locus. Let y be the frequency of alleles at the neutral locus whose ancestry traces back to the chromosome on which the beneficial mutation occurred (Gillespie 2000). At the moment that the beneficial mutation arises, $y = 1/2N$. With little recombination between the two loci, y increases along with the frequency of the beneficial allele ($0 < y \leq \beta$). When the beneficial allele is fixed, $y \sim \varepsilon^{r/s}$ (Kim and Stephan 2002), where r is the recombination rate between the selected and the neutral sites and s and ε are the selection coefficient and the frequency of the beneficial mutation at the beginning of the sweep, respectively. Therefore, y represents the fraction of the population that becomes identical by descent due to hitchhiking. With strong selection ($\alpha = 2Ns \gg 1$), the reduction of variation and the skew of the allele frequency distribution depend on the parameter y (Fay and Wu 2000; Gillespie 2000; Kim and Stephan 2002). Because of this, the hitchhiking effect on a neutral locus should be identical for a fixed value of y regardless of the final frequency of the beneficial mutation.

In the middle of the selective phase, y can be decomposed as $\beta y_1 + (1 - \beta) y_2$, where y_1 and y_2 are the frequency of the neutral allele originally linked to the first copy of the beneficial mutation, on chromosomes carrying the beneficial and the ancestral alleles, respectively. At t generations after the occurrence of the beneficial mutation, the expectation of y_1 is given approximately by

$$E[y_1]_t = 1 - r\int_0^t \frac{(1-\varepsilon)e^{-(s+r)\tau}}{\varepsilon + (1-\varepsilon)e^{-s\tau}}\,d\tau \qquad \text{(A1)}$$

(Stephan *et al.* 1992; Kim and Stephan 2002). Using the approximations given by Stephan *et al.* (1992), Equation A1 can be simplified to $(\varepsilon(1-\beta)/\beta(1-\varepsilon))^{r/s}$ if $\varepsilon \leq \beta < 0.5$ and $\varepsilon^{r/s}$ if $0.5 \leq \beta \leq 1$. A corresponding equation for y_2 can be solved to prove that y_2 is negligible compared to y_1 unless the recombination rate is high. We therefore obtain

$$y \approx \beta\varepsilon^{r/s} \qquad \text{(A2)}$$

[for $\varepsilon \leq \beta < 0.5$, $\beta(\varepsilon(1-\beta)/\beta(1-\varepsilon))^{r/s} \approx \beta\varepsilon^{r/s}$ if $r \ll s$]. Then, the frequency distribution of the derived allele at a neutral locus under the model of incomplete sweep is approximately

$$\phi(p) = \left(\frac{\theta}{p} - \frac{\theta}{1-y}\right) I_p(0, 1-y) + \frac{\theta}{1-y} I_p(y, 1), \qquad \text{(A3)}$$

where $I_p(a, b)$ is 1 if $a < p < b$ and 0 otherwise (Fay and Wu 2000; Kim and Stephan 2002). From this distribution the probability of observing k derived alleles at a site in a sample of n chromosomes is obtained, and this probability is used to calculate the composite likelihood under the model of an incomplete sweep. Because the deterministic change of the frequency of the beneficial allele from $1/(2N)$ to $1 - 1/(2N)$ is very different from the actual trajectory, which is influenced by genetic drift at the early stage (Barton 1998; R. Durrett, personal communication), choosing $\varepsilon = 1/(2N)$ underestimates this initial rate of increase. Barton (1998) showed that, conditional on fixation, the early increase in the frequency of a beneficial allele is accelerated by a factor $1/(2s)$ relative to the deterministic increase from $1/(2N)$. Then, the true trajectory might be approximated by a deterministic one that starts from $1/(4Ns)$. We therefore use $\varepsilon = 1/(4Ns) = 1/(2\alpha)$. As $y \approx \beta\varepsilon^{r/s} = \beta(2\alpha)^{-R/(2\alpha)}$, the composite likelihood under the selective-sweep model is now a function of scaled parameters α and R ($R = |X - m|R_n$, where X and m are the nucleotide positions of the selected and the neutral loci and R_n is the scaled recombination rate per nucleotide). Source code written in C for implementing this method is available from the authors upon request.