Regions of Lower Crossing Over Harbor More Rare Variants in African Populations of *Drosophila melanogaster*

Peter Andolfatto* and Molly Przeworski†

*Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom and
†Department of Statistics, Oxford University, Oxford, OX1 3TG, United Kingdom

ABSTRACT

A correlation between diversity levels and rates of recombination is predicted both by models of positive selection, such as hitchhiking associated with the rapid fixation of advantageous mutations, and by models of purifying selection against strongly deleterious mutations (commonly referred to as "background selection"). With parameter values appropriate for Drosophila populations, only the first class of models predicts a marked skew in the frequency spectrum of linked neutral variants, relative to a neutral model. Here, we consider 29 loci scattered throughout the *Drosophila melanogaster* genome. We show that, in African populations, a summary of the frequency spectrum of polymorphic mutations is positively correlated with the meiotic rate of crossing over. This pattern is demonstrated to be unlikely under a model of background selection. Models of weakly deleterious selection are not expected to produce *both* the observed correlation and the extent to which nucleotide diversity is reduced in regions of low (but nonzero) recombination. Thus, of existing models, hitchhiking due to the recurrent fixation of advantageous variants is the most plausible explanation for the data.

IT has been known for a decade that levels of nucleotide diversity in *Drosophila melanogaster*, but not levels of divergence with its sibling species, are correlated with crossing over rates, a pattern inconsistent with a strictly neutral model of molecular evolution (Aguadé *et al.* 1989; Berry *et al.* 1991; Begun and Aquadro 1992; Aquadro *et al.* 1994). A similar pattern has emerged in a variety of organisms (Stephan 1994; Nachman 1997; Nachman *et al.* 1998; Stephan and Langley 1998; Przeworski *et al.* 2000). Explanations for the correlation include positive selection (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Gillespie 1997, 2000) as well as selection against strongly deleterious mutations (Charlesworth *et al.* 1993, 1995; Hudson and Kaplan 1995). These explanations assume that the sampled variation is neutral and reflects the action of selection at linked sites.

Particular theoretical attention has been paid to a model of recurrent but nonoverlapping episodes of positive selection, in which neutral variants linked to a strongly favored mutation are swept to fixation in the population. This so-called "hitchhiking" model predicts that loci in regions of low recombination will harbor lower levels of variation and more low frequency polymorphisms than loci in regions of normal recombination (Kaplan *et al.* 1989; Braverman *et al.* 1995; Gilles-

pie 2000). This occurs because the physical length of the hitchhiking region depends on the strength of selection relative to the recombination rate (Kaplan *et al.* 1989). If selection coefficients are similar across the genome, loci in regions of low recombination will be affected by more selective sweeps per unit time, and hence are more likely to be sampled shortly after a sweep. In the recovery phase, most polymorphisms will be young (*i.e.*, postsweep) and at low frequency.

An alternative to positive selection models is purifying selection against *strongly deleterious* mutations, hereafter referred to as "background selection" (Charlesworth *et al.* 1993, 1995; Hudson and Kaplan 1994, 1995). In this model, a neutral allele will persist in the population only if it finds itself on a deleterious mutation-free chromosome (or segment of chromosome), either when it first arises in the population or by recombination. If selection coefficients and deleterious mutation rates are the same in different regions, the rate of recombination will determine the extent of the reduction in neutral diversity (*i.e.*, the extent to which neutral alleles can escape from background selection).

While purifying selection undoubtedly occurs, uncertainty about key parameters, such as the distribution of selection coefficients and the deleterious mutation rate, renders unclear its importance in reducing levels of variability. Similar uncertainty exists for positive selection models. As a result, considerable debate has revolved about the relative importance of background selection and hitchhiking in shaping patterns of variability.

While positive selection models predict an excess of rare alleles at linked neutral sites relative to a model of

Corresponding author: Peter Andolfatto, Institute for Cell and Animal Population Biology, Ashworth Labs, Kings Bldgs., University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom.
E-mail: peter.andolfatto@ed.ac.uk

no selection (Braverman *et al.* 1995; Gillespie 2000), the background selection model does not, so long as the population is large and the deleterious mutation rate not extremely high (Hudson and Kaplan 1994; Charlesworth *et al.* 1995). In *D. melanogaster*, these two conditions are likely to be met (Li *et al.* 1999; McVean and Vieira 2001). Thus, polymorphism data from *D. melanogaster* provide a means to distinguish between models. Two previous surveys of loci in *D. melanogaster* did not detect a skew toward rare variants in regions with low rates of crossing over (Begun and Aquadro 1993; Charlesworth *et al.* 1995). These observations suggested that background selection might be a sufficient explanation for the correlation between diversity levels and recombination rates, *i.e.*, that there exists no unequivocal evidence for positive selection.

The effects of background selection and hitchhiking on the frequency spectrum of linked neutral sites are known for a random-mating population of constant size. Demographic departures from these model assumptions can alter the signature of selection (*e.g.*, see Nordborg 1997; Slatkin and Wiehe 1998; Hamblin and Di Rienzo 2000). Recent multilocus surveys suggest that patterns of polymorphism in *D. melanogaster* do not conform to the expectations of a panmictic population of constant size (Andolfatto and Przeworski 2000; Andolfatto 2001; Przeworski *et al.* 2001). In particular, the species probably has an African origin and has only recently become cosmopolitan (David and Capy 1988; Lachaise *et al.* 1988). There is also evidence that populations are geographically structured (*e.g.*, Hale and Singh 1991; Begun and Aquadro 1993). Given that African populations are thought to be ancestral, they are likely to have a more stable evolutionary history than non-African populations. As a result, they may be closer to the demographic assumptions of population genetic models; *i.e.*, the effective population size is likely to be large and may not have changed much in size over time. With these considerations in mind, we revisit the relationship between frequency spectra and crossing over rates by focusing, where possible, on samples from single African populations. Our goal is to establish whether background selection alone can account for patterns of nucleotide variation.

MATERIALS AND METHODS

Data collection and previously published data: We collected sequence polymorphism data for 10 loci distributed over a range of crossing over rates on the X chromosome. We concentrate on loci in regions with lower than average crossing over rates, which are underrepresented in published data. Primers used to PCR amplify and sequence these loci are listed in Table 1. Genomic DNA was prepared (Gentra Systems, Research Triangle Park, NC) from one male fly for each of 13 isofemale lines kindly provided by C.-I Wu. These lines are sampled from a Zimbabwe, Africa population (*cf.* Begun and Aquadro 1993). All 10 loci (Table 1) encode 600–700 bp of intron

sequence (except *zeste*, which is 3′ untranslated DNA) and were sampled from the same 13 individuals. PCR products were sequenced directly on both strands using a Big-Dye sequencing kit and run on an ABI377XL automated sequencer (Perkin-Elmer-Applied Biosystems, Norwalk, CT). Sequence contigs were managed with Sequencher 3.0 software (Gene Codes, Ann Arbor, MI) and aligned manually. Aligned sequences were deposited into GenBank under accession nos. AF345663–AF345791.

From published data, we include the 19 loci available for African populations with sample sizes greater than three. For the majority of these loci (12), population samples were drawn from a Zimbabwe population (as above) with the exceptions of *Fbp2*, *Su(H)*, and *Vha68*, where samples are from an Ivory Coast population; *Acp29AB*, from a Malawi population; and *Dras1*, *Dras2*, and *R*, from a number of African populations. *CecC* was chosen as a representative from a cluster of closely related genes (*cf.* Clark and Wang 1997).

For each locus, we include all polymorphisms at synonymous sites and in noncoding DNA (including both insertion-deletions and single nucleotide polymorphisms). Single nucleotide polymorphisms within deletions and overlapping deletions were excluded from analyses. Excluding insertion-deletion variation has no effect on the conclusions (results not shown). In situations where a nucleotide polymorphism occurs immediately adjacent to a deletion, and the two are in complete coupling, the two polymorphisms are treated as a single event. For these reasons, the numbers reported in tables may differ slightly from previously published values. DnaSP3.0 (Rozas and Rozas 1999) was used for a subset of polymorphism analyses.

Estimating rates of crossing over: For each locus, we estimate *r*, the sex-averaged local rate of meiotic crossing over per kilobase per generation, following the approach of Charlesworth (1996). In this method, each chromosomal arm is divided into several regions. Each region was assigned either a linear or a quadratic function relating genetic distance to physical distance. We modified this method to incorporate estimates of the DNA content of each cytological band (Heino *et al.* 1994). Thus, the rate of crossing over is expressed in units of centimorgans per million bases (cM/Mb) rather than centimorgans per band (cM/band). Since there is no crossing over in *D. melanogaster* males (Ashburner 1989), female rates of crossing over are multiplied by $\frac{1}{2}$ for autosomes and $\frac{2}{3}$ for X-linked loci.

Previous approaches have fit high-order polynomial curves to whole chromosome arms using standard map distances available in the databases (*e.g.*, Ashburner 1989, chapter 11; Comeron *et al.* 1999). These approaches produce estimates of *r* that are highly correlated with our estimates (*e.g.*, for our 29 loci, the correlation with estimates from Comeron *et al.* 1999 yields a Pearson correlation coefficient of $R = 0.91$). The main difference between approaches is the estimates of *r* in regions of very low crossing over, *i.e.*, telomeric and centromeric regions of *D. melanogaster*. Using Charlesworth's method, we recover *yellow-su(s)*, *yellow-su(wᵃ)*, and *yellow-white* map distance estimates ($\sim 10^{-4}$, 10^{-3}, and 0.015 M, respectively) that are in good agreement with published measurements (*cf.* Takano-Shimizu 1999; Langley *et al.* 2000).

For the *white* locus (cytological map 3C2, genetic map 1.5), the appropriate *r* estimate is less clear. Charlesworth (1996) chooses this locus to mark the transition from a region with a quadratic relation of genetic to physical distance (from the telomere) to one with a linear relation in the centromere-proximal direction. We take the *r* predicted by the telomere-proximal quadratic function (0.6 cM/Mb). Note, however, that genetic distance in the centromere-proximal direction from *white* appears to increase faster than quadratically. If,

TABLE 1

PCR and sequencing primers for new X-linked loci used in this study

Locus	Forward (5′ to 3′)	Reverse (5′ to 3′)
Yellow	GTTGCGAGGTTTTAGGACTGA	CAATGGTAACTGCCGTGCTAA
SNF1A	GTGGCCGTCAAGATACTCAAT	AAGGTTAGCAAGCAGGTTTAT
Broad	AAAAAGTTGGAGAGGGGTAGT	CAACAACAGCGAATCACATAA
Armadillo	TCCAACAATGCGACTCTCACT	AAGAGGGTTTCGCTCACACTG
Actinin	CAAAAATAGCACAAGGGAAAT	GAGGAGTGGGAGAATAGAAGC
Pgd	AATGTGCAAAGGGGAACGATA	TCCGAGTAAATCCACCTGAAA
Vinculin	TTGGGTCAGAGGGCTAAACTT	CTCCGCTCGTTTGTTTCTTCG
Zeste	CCCTCCTTGTTGTCGTTATTG	TTCTGCATCACTAATCTATCG
Period	CCAAAAAGTGAGACCATAAGT	GAATAAGCTGAAAATACCAAG
White	AGAGCAGCGGCGGATAAAATG	ATGGGTGGTGGGAGTGACAAG

instead, we assume an exponential increase in genetic distance from *yellow* (cytological map 1B1, genetic map 0) to *Notch* (cytological map 3C7–9, genetic map 3.0), we estimate an *r* for *white* that is ∼2.5-fold larger. The use of this *r* estimate for *white* has no effect on our conclusions (results not shown).

Summary of the frequency spectrum of mutations: To summarize the frequency spectrum, we use the statistic *D* (Tajima 1989c). *D* considers the (approximately normalized) difference between two estimates of the population mutation rate, θ: π, the mean number of pairwise differences in the sample, and $θ_W$, an estimate based on the number of segregating mutations in the sample (Watterson 1975). Under the standard neutral model, the expectation of *D* is roughly zero (Tajima 1989c). *D* is positive when there are too many intermediate frequency variants, such as in the presence of a balanced polymorphism, or under certain population subdivision models (Tajima 1989a). *D* is negative when there is an excess of rare mutations, as would be expected under recent population growth (Tajima 1989b), or under hitchhiking models (Braverman *et al.* 1995; Gillespie 2000). For each locus, Tajima's *D* was calculated on the basis of the total number of mutations rather than on the number of segregating sites. This approach is conservative since it tends to make the *D* values of loci in regions of high recombination (and greater diversity) more negative. This said, the number of multiply hit sites is very small.

We also report a second summary of the frequency spectrum, *D**, proposed by Fu and Li (1993). Similarly to *D*, *D** compares two estimates of θ: $θ_W$ and an estimate based on the number of mutations found only once in a sample of chromosomes (referred to as singletons). The expectation of *D** is approximately zero under the neutral model; it is negative when there is an excess of singletons in the sample (Fu and Li 1993).

Coalescent simulations: To perform weighted regressions (see results), we estimated the variance of *D* under the null model for each locus. To do so, we ran 10^4 coalescent simulations for a random-mating population of constant size with no selection (*cf.* Hudson 1990). Every new mutation occurs at a previously unmutated site. We implemented the coalescent for a fixed number of segregating sites (Hudson 1993); *i.e.*, we generated a genealogy, then placed the observed number of segregating sites on the tree. For each locus, we conditioned on the sample size the total number of base pairs sequenced and an estimate of the population recombination rate *C* = 4*Nr* (or 3*Nr* if X-linked, assuming no sex differences and equal sex ratios), where *r* is the sex-averaged rate of crossing over per base pair per generation. *N*, the diploid effective population size of the species, is taken to be 10^6 (Li *et al.* 1999;

Andolfatto and Przeworski 2000). The exact value of *N* is not important, since we are mainly interested in the *relative* variance of *D* for loci in regions with different crossing over rates. Note finally that weighted regression ignores error in the estimation of *r*.

To estimate the variance of *D* under a model of background selection, we ran simulations as before, but conditional on *C* = *f*4*Nr* (or *f*3*Nr* for X-linked loci), where the *f* value for each locus is the expected local reduction in effective population size due to background selection, as estimated according to Charlesworth (1996). We use the values of *f* that provided the best fit to the available restriction fragment length polymorphism data, when both transposable elements and point mutations are considered (Charlesworth 1996).

We also ran coalescent simulations to estimate the distribution of the Pearson correlation coefficient *R*. For a model of a single population of constant size, simulations are implemented as above. We generated *D* values for the 29 loci 10^4 times, and calculated *R* each time. We then tabulated the number of runs with simulated *R* greater than or equal to the observed one. We also estimate the distribution of *R* for a model of population growth and a model of population structure. We model population growth as a constant population size of 10^5 until 10^5 generations ago, followed by an exponential increase to a present size of 10^7 (*cf.* Marjoram and Donnelly 1994). Other aspects of the simulation are as above. Our model of geographic subdivision is a symmetric two-island model (*e.g.*, Tajima 1989a), where samples are drawn entirely from one deme. The number of individuals in each subpopulation, N_{pop}, is taken to be *N*/2. We considered a fairly extreme model of structure, where the number of migrants per island per generation is one. This value is lower than suggested by most F_{ST} values observed in *D. melanogaster* (Irvin *et al.* 1998). Simulations are run using modifications of programs kindly provided by R. R. Hudson.

RESULTS

Summaries of diversity and of the frequency spectrum of mutations are positively correlated with rates of crossing over: In Table 2, we summarize patterns of nucleotide variation at synonymous sites and noncoding DNA for 10 newly sequenced loci and 19 previously published data sets. To allow for comparisons between X and autosome diversity levels, a point of departure is to multiply X-linked diversities by ⁴⁄₃ (*e.g.*, Begun and

Aquadro 1992); this approach assumes equal sex ratios and no sexual selection. In Figure 1, we plot the relationship between the corrected π, the average number of pairwise differences per site, and the estimated rates of crossing over per physical distance. As expected (Begun and Aquadro 1992; Aquadro et al. 1994), nucleotide diversity is positively correlated with crossing over rates (Pearson's $R = 0.61$, $P = 0.001$; Spearman's $R = 0.66$, $P < 0.001$; all reported P values are two-tailed). It is unclear, however, whether this scaling is appropriate. Sexual selection in natural populations can cause the scaling of X to autosome diversities to be closer to unity (Caballero 1995). If diversities in Table 2 are left uncorrected, the correlation is stronger (Pearson's $R = 0.65$, $P < 0.001$; Spearman's $R = 0.71$, $P < 0.001$).

In Figure 2, we plot the values of Tajima's D and Fu and Li's D^* against rates of crossing over for the 29 loci in Table 2. Values of D are sharply negative in areas of low crossing over and increase with increasing rates of crossing over (Pearson's $R = 0.56$, $P = 0.002$; Spearman's $R = 0.54$, $P = 0.002$). The correlation between crossing over and D^* is even less likely to occur by chance (Pearson's $R = 0.67$, $P < 0.001$; Spearman's $R = 0.64$, $P < 0.001$). These patterns are expected under a model of hitchhiking associated with the rapid fixation of advantageous mutations (Braverman et al. 1995; Gillespie 2000) and under some random-environment-selection models (Gillespie 1997), but not under the background selection model (Hudson and Kaplan 1994; Charlesworth et al. 1995). A significant correlation is also obtained when the analysis is restricted to the 22 loci sampled only in Zimbabwe (for D, Pearson's $R = 0.53$, $P = 0.011$; Spearman's $R = 0.51$, $P = 0.016$; while for D^*, Pearson's $R = 0.66$, $P = 0.001$; Spearman's $R = 0.58$, $P = 0.004$).

One concern is that loci in areas of very low crossing over may not represent independent data points. However, considering these loci as a single data point has no effect on our conclusions. As an illustration, if we replace the D values for the nine X-linked loci with a crossing over rate of $\leq 5 \times 10^{-6}$/kb/generation by one value (either the average D or the total D), the correlation is still highly significant ($P < 0.01$).

Also relevant is the presence of inversion polymorphisms in African populations of *D. melanogaster* (Lemeunier and Aulard 1992). When heterozygous, inversions suppress recombination locally and can increase crossing over rates on other chromosomes (Schultz and Redfield 1951; Sniegowski et al. 1994). These features make our estimates of the crossing over rate less certain. In particular, because inversion polymorphisms are more common on autosomes than on the X chromosome, estimates of crossing over rates for autosomes could be systematically biased upward relative to those on the X. However, the correlation of D (or D^*) and crossing over rates is still significant if autosomal crossing over rates are reduced by one-half (for D, $P < 0.02$;

and for D^*, $P < 0.002$). The main effect of inverted chromosomes should be to introduce additional noise (Andolfatto et al. 2001). In fact, the correlation is improved if inverted chromosomes are excluded where possible [*In(2L)t*, *Fbp2*, *Su(H)*, and *Vha*, results not shown]. In this light, it is worth noting that the correlation between diversity and crossing over is stronger for X-linked loci (Pearson's $R = 0.90$, 15 loci), where there are no high frequency inversions, than for autosomal loci (Pearson's $R = 0.39$, 14 loci).

Sensitivity to demographic assumptions: Our null model assumes a random-mating population of constant size. Population structure will alter the distribution of D relative to panmixia, even if samples are drawn from a single deme (see Wall 1999). When samples are drawn from one island in a symmetric island model, the mean D is slightly increased, and there is more weight on both tails of the distribution of D (Wall 1999; M. Przeworski, unpublished results). To test the significance of the correlation between D and r in the presence of structure, we ran simulations of a symmetric two-island model for $4N_{pop}m = 1$, where N_{pop} is the number of individuals in each population and m is the fraction of migrants per generation. Sampling was entirely from one deme. We ran 10^4 trials to estimate the distribution R under this null model (see materials and methods). Only 9 of the 10^4 trials had $R \geq 0.56$ (the observed value). This model of population subdivision is a highly simplified one; nonetheless, the results suggest that our correlation is no less unusual in the presence of population structure.

Inequality of variance among samples and the background selection hypothesis: Standard correlation tests are not entirely appropriate, since under the null model the expected variance in D values increases with decreasing recombination (Hudson 1983; Wall 1999; Przeworski et al. 2001). In addition, samples from regions of low recombination tend to contain fewer segregating sites (see Table 2). With this in mind, we performed a weighted regression, weighting each observed D value by the reciprocal of the estimated variance of D for each locus (see materials and methods). When the inequality of variances among samples is taken into account in this way, $P = 0.01$ ($R = 0.47$).

Background selection against strongly deleterious alleles in large populations can be thought of as a simple reduction in effective population size (Hudson and Kaplan 1994, 1995), so should not result in a distortion of the frequency spectrum. However, by reducing C, background selection will further increase the variance of D. We therefore performed a second weighted regression. This time, the variance of D was estimated from simulations that condition on $C = f4Nr$ (or $f3Nr$ for X-linked loci), where f for each locus is the expected local reduction in effective population size due to background selection, as estimated by Charlesworth (1996). Under this model, the probability of observing

TABLE 2

Patterns of variability and local rates of crossing over for 29 loci in African populations of *D. melanogaster*

Gene region	n	S^a	Indels[b]	No. of silent sites[c]	π (%)	r (cM/Mb)	Tajima's D	Fu and Li's D^*	Reference or GenBank accession nos.
X chromosome									
Yellow (*y*)	13	1	0	518	0.03	0.04	−1.15	−1.37	This study
suppressor of sable [*su(s)*]	50	46	7	3213	0.20	0.07	−1.25	−1.78	Langley *et al.* (2000)
suppressor of white-apricot [*su(w^a)*]	50	50	11	1955	0.42	0.17	−0.92	−2.02	Langley *et al.* (2000)
SNF1A	13	9	1	527	0.65	0.24	0.73	0.51	This study
broad (*br*)	13	9	2	572	0.29	0.27	−1.71	−1.72	This study
Armadillo (*arm*)	13	3	1	570	0.33	0.36	−1.65	−2.03	This study
Actinin (*Actn*)	13	7	2	546	0.33	0.43	−0.78	−1.40	This study
Phospogluconate-dehydrogenase (*Pgd*)	13	14	2	540	0.58	0.47	−1.27	−1.56	This study
Vinculin (*Vinc*)	13	11	1	545	0.67	0.50	0.14	−0.06	This study
Zeste (*z*)	13	24	1	581	1.11	0.63	−0.73	−0.49	This study
Period (*per*)	13	19	1	579	0.75	0.67	−1.23	−1.25	This study
White (*w*)	13	7	0	393	0.31	0.69	−1.77	−1.95	This study
Vermilion (*v*)	11	79	2	1381	2.01	2.27	0.14	0.31	Begun and Aquadro (1995)
Glucose-6-phosphate dehydrogenase (*Zw*)	12	32	1	490	2.10	1.36	−0.14	0.08	U42765–67 U42738–49 U44721
Runt (*run*)	7	33	0	792	1.64	1.36	−0.22	0.08	AF077074–80
Chromosome 2									
Acp26Aab	10	10	0	526	0.72	2.35	0.33	0.34	AF053250–57 AF053270–71
Acp29AB	13	52	2	1183	1.46	2.35	0.15	0.01	AJ240539–51
Fat-body protein 2 (*Fbp2*)	10	67	6	1530	1.69	2.35	0.44	0.89	AF045787–96
In(2L)t-34A breakpoint	7	22	3	684	1.19	1.45	−0.53	−0.39	AF217947–49 AF275826–29
Vacuolar H^+-ATPase 68 (*Vha68*)	20	9	0	474	0.30	1.41	−1.52	−1.20	Depaulis *et al.* (2000)
Alcohol dehydrogenase (*Adh*)	10	33	2	1434	0.69	1.19	−0.76	−0.75	S.-C. Tsaur (unpublished results)
Suppressor of Hairless [*Su(H)*]	20	47	7	803	2.16	1.16	1.24	1.22	AF088255–74
Chromosome 3									
Roughened (*R*)	4	3	0	127	1.32	1.85	0.17	0.17	Gasperini and Gibson (1999)
Ras oncogene at 64B (*Dras2*)	4	13	0	387	1.85	1.85	0.11	0.11	Gasperini and Gibson (1999)
Glucose dehydrogenase (*Gld*)	10	1	0	385	0.05	0.15	−1.11	−1.24	Hamblin and Aquadro (1997)
Relish (*Rel*)	6	11	0	841	0.49	0.28	−0.86	−0.91	AF204284–89
Ras oncogene at 85D (*Dras1*)	4	7	1	1406	0.25	0.33	−0.81	−0.82	Gasperini and Gibson (1999)
Triose-phosphate isomerase (*Tpi*)	5	23	0	505	2.06	1.04	−0.43	−0.43	U60837, -39, -49, -59–60
Cecropin-C (*CecC*)	5	9	1	199	2.38	1.04	0.71	0.46	AF019014–18

For each locus, we list the sample size (*n*), the number of segregating mutations (*S*), the number of insertion-deletion mutations, the number of silent sites, the average pairwise difference at silent sites (π%, Tajima 1989c), the sex-averaged rate of crossing over in centimorgans per million bases per generation (*r*), the value of Tajima's *D* (Tajima 1989c), and the value of Fu and Li's *D** (Fu and Li 1993).

[a] Including single nucleotide substitutions and two-state insertion-deletion polymorphisms.

[b] Number of two-state insertion-deletion polymorphisms.

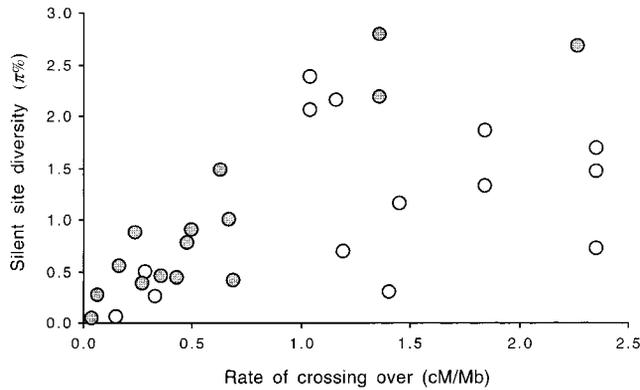[c] Including 5′ or 3′ untranslated, intron, and synonymous sites.

FIGURE 1.—Relationship between silent nucleotide diversity (π%) and the rate of crossing over for 29 loci sampled from African populations of *D. melanogaster*. Shaded circles represent X-linked loci. X-linked diversities (Table 2) have been multiplied by $\frac{4}{3}$ to make them comparable to autosomal diversities, assuming equal sex ratios and no sexual selection. There is a significant, positive correlation (Pearson's $R = 0.61$, $P = 0.001$, Spearman's $P < 0.001$, two-tailed).

such a strong positive correlation between D and r ($R = 0.48$) by chance is estimated to be $P < 0.01$. Thus, background selection *alone* is an unlikely explanation for the correlation.

Effects of varying sample size: We have shown that if we assume a random-mating population of constant size [where $E(D) \approx 0$], the correlation between D and crossing over rates is highly unlikely. Next, we consider whether the correlation between D and r might reflect negative D values at *all* loci. This is a concern because when the *population* D is not zero, the sample estimate of D is no longer unbiased. In particular, small samples will yield estimates closer to zero than will large ones (*cf.* Gillespie 2000). In our data, there is a marginally significant negative correlation between sample size (n) and crossing over rates (Pearson's $R = -0.344$, two-tailed $P = 0.07$). Suppose that population D values are all nonzero, for example, due to population expansion. Might the unequal sample sizes result in a correlation between diversity and crossing over rates? To test this possibility, we randomly subsampled 10 alleles for each of the two data sets with $n = 50$. For this version of the data, D values are -1.35 for *su(s)* and 0.55 for *su(w^a)* and all sample sizes are <20. The correlation between n and estimated crossing over rates has all but disappeared (Pearson's $R = -0.13$, two-tailed $P = 0.51$), yet there is still a significant correlation between D and crossing over rates (Pearson's $R = 0.48$, two-tailed $P < 0.01$). Thus, the correlation between D and r observed in these samples is likely to reflect an underlying correlation in population D values (rather than nonzero population D values at all loci).

Whether we even expect the negative correlation between n and r to produce a spurious correlation between D and r is unclear. While loci in areas of high
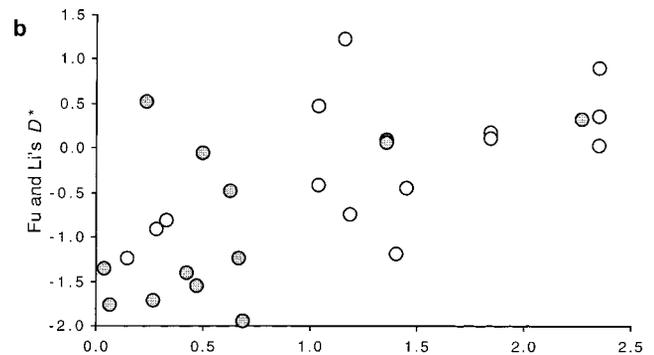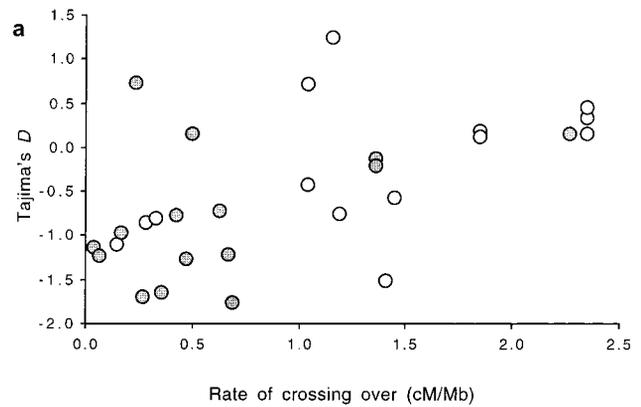

FIGURE 2.—Scatterplot of (a) Tajima's D and (b) Fu and Li's D* *vs.* the rate of crossing over for 29 loci sampled from African populations of *D. melanogaster*. Shaded circles represent X-linked loci. D *vs.* crossing over: Pearson's $R = 0.56$, $P = 0.002$, Spearman's $P = 0.002$, two-tailed. D* *vs.* crossing over: Pearson's $R = 0.67$, $P < 0.001$, Spearman's $P < 0.001$, two-tailed. The correlation with D is improved by excluding insertion-deletion variation (Pearson's $R = 0.59$).

crossing over tend to have smaller sample sizes, they also tend to have many more polymorphic sites. This follows from the strong correlation observed between diversity and rates of crossing over in our data (Figure 1). Also, the variance of D under the null model is much smaller for loci in regions of high recombination than it is for loci in regions of low recombination (Przeworski *et al.* 2001). To explore this further, we investigated whether a positive correlation is expected for a model of population growth. Specifically, we estimated the distribution of Pearson's R for the 29 loci for a model of constant population size ($N = 10^5$) followed by 10^5 generations of exponential growth to a present size of 10^7. This model is not intended as a realistic depiction of the population history of *D. melanogaster* but as an illustration. With these parameters, for $n = 10$ and 10 segregating sites, the mean D is ~ -0.80, while for $n = 50$ and 50 segregating sites, the mean D is ~ -1.76 (as estimated from 10,000 simulations). The average R value in 10^4 simulations of the 29 loci was $+0.08$ (*i.e.*, significantly greater than 0). However, the $\Pr(R \geq 0.56)$ was estimated to be 0.001; in other words, the correla-

tion that we observe is still highly unlikely. While this is only one model among many, it suggests that the effect of varying sample sizes on R is quite weak when other facets of the data are also taken into account.

DISCUSSION

Models of positive selection: The positive correlation between D and crossing over rates is unlikely under a simple model of background selection but is expected under simple recurrent hitchhiking models (Braverman *et al.* 1995; Gillespie 2000). As one illustration, the expected D value in Gillespie's (2000) pseudo-hitchhiking model is ~ -1.5 for regions of no recombination and samples of 20 drawn from "large" populations ($N \geq 10^4$). Thus, hitchhiking due to the rapid fixation of advantageous variants appears to be the best explanation for the data. Patterns of polymorphism may also be consistent with random-environment-selection models (Gillespie 1997), but their sampling properties have yet to be investigated explicitly.

Also of interest are models that incorporate both deleterious and favorable mutations, as both processes are likely to be occurring simultaneously (Hudson 1994). Such models are at their infancy. The sole study of the joint effects of hitchhiking and background (Kim and Stephan 2000) concludes that hitchhiking is likely to be the main force shaping patterns of variation in regions of low recombination, while background selection may dominate when recombination rates are higher. Thus, the correlation between D and r reported here may be consistent with the joint action of these forces.

Alternative deleterious evolution models: Next, we consider whether alternative models of deleterious evolution might also be consistent with the patterns reported in Figures 1 and 2. A large class of deleterious mutations may have selection coefficients smaller than values typically used in models of background selection (Fry *et al.* 1999; Keightley and Eyre-Walker 1999). Such selection intensities may be particularly relevant to transposable elements (*cf.* Charlesworth 1996) as well as to a subset of synonymous sites and amino acid variants (Akashi 1999). While selection against strongly deleterious mutations does not result in a skewed frequency spectrum at linked neutral sites (Hudson and Kaplan 1994), selection against weakly deleterious mutations can lead to an excess of rare neutral alleles in large samples (*e.g.*, for a sample size of 100 chromosomes, Charlesworth *et al.* 1995). Whether a skew would also be detectable in small samples (our median sample size is 13) is unknown. In addition, the parameters (selection coefficient and deleterious mutation rate) required to produce a skew in the frequency spectrum of linked neutral variants are unlikely to cause the drastic reduction of nucleotide diversity observed in regions of low recombination in *D. melanogaster* (Charles-

worth *et al.* 1995; Charlesworth 1996; Nordborg *et al.* 1996).

When purifying selection is weak, linked deleterious alleles may persist long enough in a population to co-segregate. Selection at one site can then reduce the efficacy of selection at linked sites (Hill and Robertson 1966; Felsenstein 1974), a phenomenon termed Hill-Robertson interference (HRI). Weak selection HRI models (*e.g.*, Li 1987; Comeron *et al.* 1999; McVean and Charlesworth 2000; Tachida 2000) can produce a negative skew in the frequency spectrum of linked neutral variants under some parameters (Tachida 2000). However, while HRI models produce a reduction in diversity levels, they cannot explain diversity reductions on the scale of those observed in areas of low (but nonzero) crossing over in *D. melanogaster* (Tachida 2000; G. McVean, personal communication).

In addition, the effect of weak selection HRI is probably weaker than has been estimated in simulations, as published models (*e.g.*, McVean and Charlesworth 2000) include crossing over but not gene conversion. Gene conversion may be the prevalent mechanism determining the strength of associations among segregating mutations in regions of very low crossing over in Drosophila (Langley *et al.* 2000), and thus the strength of HRI. Preliminary simulations suggest that the effect of HRI on the frequency spectrum and diversity levels at linked neutral variants is considerably weakened by the presence of gene conversion in regions of nonzero crossing over (J. Comeron, personal communication; G. McVean, personal communication).

The models described thus far have considered the effect of selection on linked neutral variation. However, sampled variation may not be neutral, but may itself be under very weak selection (*i.e.*, on the order of $1/N$). In this case, a negative D value is expected for *all* rates of recombination (McVean and Charlesworth 2000). In addition, only a minor reduction in variability is expected for regions of low recombination relative to regions of high recombination (*e.g.*, see Figure 3 of McVean and Charlesworth 2000).

In summary, existing models of deleterious evolution, weak or strong, are unlikely to account for *both* the reductions in diversity and the skew in D observed in small samples from regions of low crossing over. However, models with a distribution of selection coefficients have yet to be investigated. Recent work on the selective effects of deleterious mutations suggests a large class of deleterious mutations with minor effects in both Drosophila and *Caenorhabditis elegans* (Davies *et al.* 1999; Keightley and Eyre-Walker 1999). The challenge is to construct a more realistic model of deleterious evolution (including a distribution of selection coefficients, crossing over, and gene conversion) that is constrained by plausible estimates of the deleterious mutation rate (McVean and Vieira 2001), yet can account for the pattern reported here.

The potential importance of demographic factors: In this study, we focused on African samples of *D. melanogaster*. We did so because non-African populations are likely to have experienced a more complicated demographic history (David and Capy 1988; Lachaise *et al.* 1988). In this light, it is noteworthy that no significant correlation has been found between *D* and crossing over rates in the available data for non-African populations (*cf.* Aquadro *et al.* 1994; Charlesworth *et al.* 1995). The reasons for the discrepancy between African and non-African populations are not understood. One possibility is that a recent demographic perturbation (*e.g.*, founder effects) in the history of non-African populations (David and Capy 1988; Begun and Aquadro 1993, 1995) obscures the pattern of natural selection. Long-term population structure (*e.g.*, Slatkin and Wiehe 1998; Hamblin and Di Rienzo 2000) could also alter the signature of selection relative to the predictions of existing models.

Species-specific demographic considerations are likely to be of relevance for evolutionary inferences from other species as well. A large quantity of polymorphism data will soon be available for humans. Like Drosophila, anatomically modern humans may have an African origin and have only recently become cosmopolitan (Cann *et al.* 1987). Preliminary data suggest a correlation between diversity levels and crossing over rates (Nachman *et al.* 1998; Przeworski *et al.* 2000). By carefully considering the demographic history of populations, it may be possible to recover a similar pattern in the human data.

CONCLUSIONS

A decade of empirical work has been devoted to distinguishing between a simple model of background selection (Charlesworth *et al.* 1993; Hudson and Kaplan 1995) and a simple model of positive selection (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989). Here we have shown that both levels of diversity and a summary of the frequency spectrum of mutations, *D*, are positively correlated with rates of crossing over in African populations of *D. melanogaster*. This multilocus pattern is unlikely under the first model and is expected under the second. Together with an increasing number of examples of selective sweeps at individual loci in Drosophila and other organisms (*e.g.*, Schlötterer *et al.* 1997; Chen *et al.* 2000; Hamblin and Di Rienzo 2000; Nachman and Crowell 2000; Yi and Charlesworth 2000), our results suggest that positive Darwinian selection is a common force shaping patterns of variability in natural populations.

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.

Akashi, H., 1999 Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene **238:** 39–51.

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

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., F. Depaulis and A. Navarro, 2001 Inversion polymorphism and nucleotide variability in *Drosophila*. Genet. Res. **77:** 1–8.

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

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

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. 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.

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

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.

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.

Caballero, A., 1995 On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics **139:** 1007–1011.

Cann, R. L., M. Stoneking and A. C. Wilson, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

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

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

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

Chen, Y., B. J. Marsh and W. Stephan, 2000 Joint effects of natural selection and recombination on gene flow between *Drosophila ananassae* populations. Genetics **155:** 1185–1194.

Clark, A. G., and L. Wang, 1997 Molecular population genetics of Drosophila immune system genes. Genetics **147:** 713–724.

Comeron, J. M., M. Kreitman and M. Aguade, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

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

Davies, E. K., A. D. Peters and P. D. Keightley, 1999 High frequency of cryptic deleterious mutations in *Caenorhabditis elegans*. Science **285:** 1748–1751.

Depaulis, F., L. Brazier, S. Mousset, A. Turbe and M. Veuille, 2000 Selective sweep near the *In(2L)t* inversion breakpoint in an African population of *Drosophila melanogaster*. Genet. Res. **76:** 149–158.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

Fry, J. D., P. D. Keightley, S. L. Heinsohn and S. V. Nuzhdin, 1999 New estimates of the rates and effects of mildly deleterious mutation in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **96:** 574–579.

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

Gasperini, R., and G. Gibson, 1999 Absence of protein polymorphism in the *Ras* genes of *Drosphila melanogaster*. J. Mol. Evol. **49:** 583–590.

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

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

Hale, L. R., and R. S. Singh, 1991 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. IV. Mitochondrial DNA variation and the role of history *vs.* selection in the genetic structure of geographic populations. Genetics **129:** 103–117.

Hamblin, M. T., and C. F. Aquadro, 1997 Contrasting patterns of nucleotide sequence variation at the *Glucose dehydrogenase* (*Gld*) locus in different populations of *Drosophila melanogaster*. Genetics **145:** 1053–1062.

Hamblin, M. T., and A. Di Rienzo, 2000 Detection of the signature of natural selection in humans: evidence from the *Duffy* blood group locus. Am. J. Hum. Genet. **66:** 1669–1679.

Heino, T. I., A. O. Saura and V. Sorsa, 1994 Maps of the salivary gland chromosomes of *Drosophila melanogaster*. Dros. Inf. Serv. **73:** 621–738.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Japan Scientific Society, Tokyo.

Hudson, R. R., 1994 How can the low levels of DNA sequence variation in regions of the Drosophila genome with low recombination rates be explained? Proc. Natl. Acad. Sci. USA **91:** 6815–6818.

Hudson, R. R., and N. L. Kaplan, 1994 Gene trees with background selection, pp. 140–153 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, London.

Hudson, R. R., and N. L. Kaplan, 1995 The coalescent process and background selection. Philos. Trans. R. Soc. Lond. Ser. B **349:** 19–23.

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

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

Keightley, P. D., and A. Eyre-Walker, 1999 Terumi Mukai and the riddle of deleterious mutation rates. Genetics **153:** 515–523.

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

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

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

Lemeunier, F., and S. Aulard, 1992 Inversion polymorphism in *Drosophila melanogaster*, pp. 339–405 in *Drosophila Inversion Polymorphism*, edited by C. B. Krimbas and J. R. Powell. CRC Press, Boca Raton, FL.

Li, W. H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

Li, Y. J., Y. Satta and N. Takahata, 1999 Paleo-demography of the

Drosophila melanogaster subgroup: application of the maximum likelihood method. Genes Genet. Syst. **74:** 117–127.

Marjoram, P., and P. Donnelly, 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136:** 673–683.

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

McVean, G. A. T., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

McVean, G. A. T., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245–257.

Nachman, M. W., 1997 Patterns of DNA variability at X-linked loci *in Mus domesticus*. Genetics **147:** 1303–1316.

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.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133–1141.

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

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 The effect of recombination on background selection. Genet. Res. **67:** 159–174.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **18:** 291–298.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Schlötterer, C., C. Vogl and D. Tautz, 1997 Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics **146:** 309–320.

Schultz, J., and H. Redfield, 1951 Interchromosomal effects on crossing-over in Drosophila. Cold Spring Harbor Symp. Quant. Biol. **16:** 175–197.

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

Sniegowski, P. D., A. Pringle and K. A. Hughes, 1994 Effects of autosomal inversions on meiotic exchange in distal and proximal regions of the X-chromosome in a natural population of *Drosophila-melanogaster*. Genet. Res. **63:** 57–62.

Stephan, W., 1994 Effects of genetic recombination and population subdivision on nucleotide sequence variation in *Drosophila annanasae*, pp. 57–66 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, London.

Stephan, W., and C. H. Langley, 1998 DNA polymorphism in Lycopersicon and crossing-over per physical length. Genetics **150:** 1585–1593.

Tachida, H., 2000 Molecular evolution in a multisite nearly neutral mutation model. J. Mol. Evol. **50:** 69–81.

Tajima, F., 1989a DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics **123:** 229–240.

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

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

Takano-Shimizu, T., 1999 Local recombination and mutation effects on molecular evolution in Drosophila. Genetics **153:** 1285–1296.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **73:** 65–79.

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

Yi, S., and B. Charlesworth, 2000 A selective sweep associated with a recent gene transposition in *Drosophila miranda*. Genetics **156:** 1753–1763.

Communicating editor: N. Takahata