A Genome-Wide Departure From the Standard Neutral Model in Natural Populations of Drosophila

Peter Andolfatto*,[1] and Molly Przeworski†,[1]

Institute for Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom and †Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois 60637

Manuscript received July 24, 1999
Accepted for publication May 25, 2000

ABSTRACT

We analyze nucleotide polymorphism data for a large number of loci in areas of normal to high recombination in *Drosophila melanogaster* and *D. simulans* (24 and 16 loci, respectively). We find a genome-wide, systematic departure from the neutral expectation for a panmictic population at equilibrium in natural populations of both species. The distribution of sequence-based estimates of $2Nc$ across loci is inconsistent with the assumptions of the standard neutral theory, given the observed levels of nucleotide diversity and accepted values for recombination and mutation rates. Under these assumptions, most estimates of $2Nc$ are severalfold too low; in other words, both species exhibit greater intralocus linkage disequilibrium than expected. Variation in recombination or mutation rates is not sufficient to account for the excess of linkage disequilibrium. While an equilibrium island model does not seem to account for the data, more complicated forms of population structure may. A proper test of alternative demographic models will require loci to be sampled in a more consistent fashion.

THE standard assumptions of the neutral theory of molecular variation (Kimura 1983) are that the vast majority of mutations are neutral and that genes are sampled from a panmictic population at equilibrium. Under this model, the population recombination parameter, $C = 4Nc$, and the population mutation parameter, $\theta = 4N\mu$, can be estimated from sequence polymorphism data (where N is the effective population size of the species, c the rate of recombination per base pair per generation, and μ the rate of mutation per base pair per generation). The ratio $\hat{C}/\hat{\theta}$ is then an estimate of the recombination rate between adjacent bases, scaled to the mutation rate per base pair. Estimates of these same parameters, c and μ, can be obtained from genetic and physical map data and from nucleotide divergence between closely related species, respectively. This allows for a direct comparison between the two methods of estimation (Hudson 1987). If the standard neutral model is correct, the two methods should yield similar results. This prediction is our point of departure.

Several authors have pointed out that estimates of C/θ for specific loci are different from those expected from independent estimates of c and μ (Hudson 1987; Leicht *et al.* 1995; Eanes *et al.* 1996; Hey and Wakeley 1997; Hasson *et al.* 1998; Andolfatto and Kreitman 2000). Low estimated values of C/θ have been interpreted as reflecting strong linkage disequilibrium (*e.g.*, Schaeffer and Miller 1993; Kirby and Stephan 1996). Higher than expected levels of linkage disequilibrium have also been noted for areas of normal recombination using a different approach (Miyashita *et al.* 1993; Miyashita and Langley 1994).

Excess linkage disequilibrium is often interpreted as reflecting the action of natural selection at the locus (*e.g.*, Miyashita *et al.* 1993; Schaeffer and Miller 1993; Kirby and Stephan 1996). However, this pattern can also result from demographic departures from model assumptions, *e.g.*, population structure (Li and Nei 1974; Ohta 1982). A characterization of background levels of linkage disequilibrium may help distinguish between these two explanations. Demography is a force that affects the entire genome, while natural selection has a local effect that is not expected to be the same for unlinked regions. Thus, a locus shaped by natural selection is expected to show a pattern of polymorphism that differs from most other loci. Here, we summarize the results for 24 independent loci in *Drosophila melanogaster* and 16 in *D. simulans*. We find that $\hat{C}/\hat{\theta}$ values are systematically too small to be consistent with the theoretical predictions of the standard neutral theory, given reasonable estimates of c and μ. In other words, levels of linkage disequilibrium (as measured by $\hat{C}$) are almost always significantly higher than expected.

METHODS

We use polymorphism data sets for regions of normal to high recombination ($>5 \times 10^{-9}$ per base pair per

Corresponding author: Molly Przeworski, Statistics Department, University of Oxford, 1 South Parks Rd., Oxford OX1 3TG, United Kingdom. E-mail: molly@stats.ox.ac.uk

[1]These authors contributed equally to this work.

generation) that have more than three segregating sites (24 genes in *D. melanogaster* and 16 genes in *D. simulans*). We include biallelic single nucleotide polymorphisms but not mutations that overlap deletions as they represent incomplete information. The sequences for most loci can be obtained from GenBank at http://www.ncbi.nlm.nih.gov/Entrez/. The data sets analyzed are available upon request to P. Andolfatto. Polymorphism data sets used in this study are the following: *Acp26A* (Aguadé *et al.* 1992; Tsaur *et al.* 1998), *Acp70A* (Cicera and Aguadé 1997), *Adh* (Kreitman 1983; Sumner 1991; S. C. Tsaur, unpublished results), *Boss* (Ayala and Hartl 1993), *Dpp* (Richter *et al.* 1997), *E(eve)* (Ludwig and Kreitman 1995), *Est-6* (Cooke and Oakeshott 1989; Karotam *et al.* 1993; Hasson and Eanes 1996), *G6pd* (Eanes *et al.* 1993), *Gld* (Hamblin and Aquadro 1996, 1997), *Hsp83* (Hasson and Eanes 1996), *In(3L)P* (Wesley and Eanes 1994), *In(2L)t* (Andolfatto *et al.* 1999; Andolfatto and Kreitman 2000), *Mlc1* (Leicht *et al.* 1995), *Pgd* (Begun and Aquadro 1994), *prune* (Simmons *et al.* 1994), *period* (Kliman and Hey 1993), *Ref(2)P* (Wayne *et al.* 1996), *Rh3* (Ayala *et al.* 1993), *runt* (Labate *et al.* 1999), *Sod* (Hudson *et al.* 1997), *Top2* (Palopoli and Wu 1996) and *Tpi* (Hasson *et al.* 1998), *white* (Kirby and Stephan 1995), *Vermilion* (Begun and Aquadro 1995), and *zeste* and *Yp2* (Hey and Kliman 1993). *Cec-C* (Clark and Wang 1997) and *Amy-d* (Inomata *et al.* 1995) were chosen as representative genes from their respective clusters. *In(2L)t* and *In(3L)P* are polymorphic inversions in *D. melanogaster*. Here, the labels refer to sequences spanning the breakpoint sites on standard chromosomes (proximal and distal breakpoint, respectively). For *D. simulans*, *In(2L)t* refers to the homologue of the *D. melanogaster In(2L)t* proximal breakpoint region. *Adh-Fast* haplotypes were excluded from the analysis, as were inverted alleles from *Adh*, *Dpp*, *Est-6*, *Hsp83*, *In(2L)t*, and *In(3L)P* samples. For *Sod*, we used region *2021*, which is ~12 kb upstream of the *Sod* coding region (Hudson *et al.* 1997).

Recombination estimates: For each locus, laboratory estimates of the regional rate of crossing over, c, are obtained as follows: for every chromosomal arm, polynomial curves were fitted to plots of cumulative genetic distance as a function of cumulative physical distance. The derivative of the polynomial at a given physical map position is taken to be the recombination rate (Ashburner 1989, pp. 453–457; True *et al.* 1996; Comeron *et al.* 1999). Laboratory estimates are not available for the second chromosome of *D. simulans*; we use the rates for the homologous region of *D. melanogaster* as surrogates for *Adh*, *E(eve)*, *In(2L)t*, and *Top2*.

To obtain an estimate of C, we need estimates of c and N (we refer to this estimate of C as C_{map}). Since males do not recombine in *D. melanogaster* and *D. simulans* (Ashburner 1989, p. 476), the population parameter C is $(1/2)4Nc = 2Nc$ for autosomal genes and $(2/3)3Nc = 2Nc$ for X-linked genes. The recombination rate c is taken to be the crossing-over rate estimated from laboratory crosses (see DISCUSSION). To estimate N, we equate the observed θ_w at silent and noncoding sites of each locus with $4N\mu$ (or $3N\mu$ for X-linked loci). Under the standard neutral model, θ_w (Watterson 1975) is an unbiased estimate of the population mutation rate, θ. Our estimate of the mutation rate $\hat{\mu}$ is obtained from $\hat{d}$, the estimated rate of divergence per year (which depends on $\hat{T}$, the estimated time to the common ancestor of the *melanogaster* and *obscura* species groups), and $\hat{g}$, the estimated number of generations per year. If $\hat{d} = 3 \times 10^{-8}|\hat{T} = 30$ million years (my) and $\hat{g} = 10$, our estimate of μ is 3×10^{-9}/bp/generation. (We discuss the validity of these assumptions in detail in the DISCUSSION.) Given these estimates for d, g, and T, the average $\hat{N}$ across loci is roughly 10^6 for *D. melanogaster* and 2×10^6 for *D. simulans*.

Hudson's (1987) estimator of C (henceforth referred to as C_{hud}) was calculated using a modification of a program kindly provided by J. Wakeley. Ideally, one would like to use all the information in the data. However, full-likelihood approaches (*e.g.*, Griffiths and Marjoram 1996) are not computationally feasible for high levels of recombination (Wall 2000). C_{hud} was chosen among many estimators of C because, under the standard neutral model, it can easily be related to the amount of linkage disequilibrium in the sample. Specifically, C_{hud} is a moment estimator obtained from the relationship expected between the sample variance of the number of pairwise differences (a measure of linkage disequilibrium) and the population recombination rate (Hudson 1987). In our implementation, any value of $C_{\text{hud}} \geq 10{,}000$ is taken to be 10,000. Similarly, if there is more association between sites than expected under the standard neutral model with no recombination, C_{hud} is set to zero. Note that C_{hud} is much more dependent on model assumptions than is our estimate of c and can only be interpreted as an estimate of $4Nc$ under a very restricted set of models.

Coalescent simulations of the standard neutral model: Coalescent simulations (Hudson 1990) of a panmictic population were run using modifications of programs kindly provided by R. Hudson and J. Wall. The simulations assume a Wright-Fisher population at equilibrium from which samples are drawn randomly (referred to as the "standard neutral model" or SNM). Mutations are neutral; each new mutation occurs at a previously unmutated site. Parameters for the simulations are the sample size, the sequence length in base pairs, and C_{map}. Ten thousand replicates are run for each parameter set.

Although we refer to our model as the "standard neutral model," standard coalescent simulations use the parameter θ and treat S as a random variable. Here, we generate genealogies and then place the observed number of mutations (at both synonymous and nonsynonymous sites), S, on the tree (Hudson 1993). We take

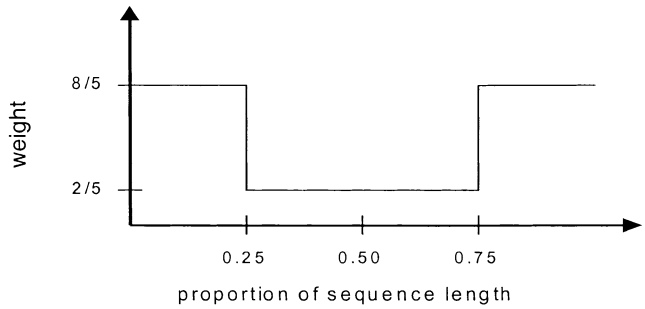

FIGURE 1.—A model of heterogeneity in selective constraints. The simulated sequence is divided into three parts, spanning one-fourth of the length, one-half, and one-fourth respectively. There is a fourfold difference in mutation rates between the middle segment of the sequence and the two end segments (see METHODS).

this approach because the number of segregating sites is observable, while we have little information about the population parameter θ (HUDSON 1993).

Note that S refers to the number of segregating sites, not the number of mutations, so we are effectively ignoring multiple hits. This choice is conservative for our purposes, since multiple hits will tend to decrease linkage disequilibrium. There are nine data sets in *D. melanogaster* with visible multiple hits (at most three sites) and seven in *D. simulans* (*Adh* has six sites with more than two alleles; the six other data sets have fewer than three). Results are virtually unchanged if we consider S to be the number of inferred mutations instead of the number of segregating sites (results not shown).

For each locus, given the observed number of segregating sites and our estimate of C, we ask what proportion P of simulated runs have a value of C_{hud} smaller than or equal to the observed value. That is, for locus i, $P_i = \Pr(\text{simulated } C_{\text{hud}} \leq \text{observed } C_{\text{hud}} \mid \text{SNM}, C_{\text{map}})$. If loci are independent, $C = C_{\text{map}}$, and the standard neutral model is accurate, the distribution of P values across loci should be uniformly distributed between 0 and 1. We test for a departure from uniformity by using the fact that, for n data sets, $-\Sigma_{i=0}^{n} 2 \ln(P_i)$ should be χ^2 distributed with $2n$ d.f. (Fisher 1954, as cited in SOKAL and ROHLF 1995). The independence of loci is a crucial assumption of our multilocus analysis. We verified that it is valid by running 1000 simulations with two loci separated by $C = 80$ (the sample size was 20, and intragenic recombination was $C = 10$). No correlation was detected between loci (results not shown).

Heterogeneity in selective constraints: We ran coalescent simulations to test the effect of a nonuniform distribution of mutations on estimates of C_{hud}. We model this spatial heterogeneity as variation in mutation rate. Coalescent simulations are run with the same parameters as for the panmictic, uniform mutation case. The sequence is divided into three parts, spanning one-fourth of the length of the sequence, one-half, and one-fourth, respectively (Figure 1). This case is meant to represent an exon flanked by two introns. Since on average one site out of four is silent in a coding region, the mutation rate in "introns" is taken to be fourfold higher than in the middle half. This model assumes that introns and silent sites have similar levels of constraint (*cf.* MORIYAMA and POWELL 1996; LI 1997). In practice, at the end of each simulated run, the total sequence is divided into segments with the same genealogy and the same mutation rate. Given a constant rate of mutation across base pairs, the probability that a mutation will be placed in segment i is given by

$$\frac{L_i T_i}{\Sigma_{j=1}^{n} L_j T_j},$$

where n is the number of segments, L_i is the length of segment i, and T_i is the total branch length of the genealogy for segment i. To add stepwise variation in mutation rates, we weight these probabilities by the "relative mutation rate" for each interval (y-axis in Figure 1).

Population subdivision: We also ran coalescent simulations for a symmetric island model. Since geographic subdivision increases the extent of linkage disequilibrium (LI and NEI 1974; OHTA 1982), it is not obvious that the loci can be treated as independent (NEI and MARUYAMA 1975; ROBERTSON 1975). To verify the assumption of independence, we simulated two loci separated by $C = 80$, with an intragenic rate of recombination of $C = 10$, $4Nm = 0.1$, and a sample size of 20; whether sampling from one or both demes, no correlations were detected (results not shown).

If we assume a symmetric two-deme model, observed values of F_{ST} (WRIGHT 1951) can be used to estimate the amount of gene flow between populations (HUDSON *et al.* 1992b). Estimates vary from roughly $4Nm = 0.5$ to 2 for *D. melanogaster* and 0.75 to 1.5 for *D. simulans*, depending on which loci are used (m is the number of migrants per deme per generation; IRVIN *et al.* 1998). We run coalescent simulations with $4Nm = 0.2$ to 1. Samples are either drawn equally (or close to equally) from both demes or only from one deme. Each parameter set is run 10,000 times. To gain insight as to how the variance in outcomes depends on the number of islands, we also run a five-island model with $4Nm = 1$ and $4Nm = 3$, sampling from only one deme.

To evaluate the fit of a symmetric island model, we use both C_{hud} and B' (WALL 1999). The statistic B' was developed as a one-tailed test of the standard neutral model (WALL 1999). B' is the number of pairs of adjacent segregating sites that partition the sample in the same way (WALL 1999). A partition of the sample consists of two disjoint subsets, whose union is the sample; each subset is composed of individuals with the same allele at that site. B' can be thought of as a measure of linkage disequilibrium among adjacent segregating sites. The expectation of B' should be higher under a geographic subdivision model than under a panmictic

one for the same level of recombination (Wall 1999). Given the true recombination rate, B' is the most powerful existing test for rejecting the panmictic neutral model when simulations are run for a two-island model and sampling is from a single locality (Wall 1999). The probabilities reported for B' correspond to the proportion of runs with a value greater than or equal to the observed B', *i.e.*, $P(B') = \Pr(\text{simulated } B' \geq \text{observed } B' \mid \text{SNM}, C = C_{\text{map}})$. A low value of $P(B')$ indicates high levels of linkage disequilibrium. Simulations for B' were run with the total number of (inferred) mutations (which in all cases equaled the sum of the number of segregating sites and the number of multiple hits). Sites with multiple hits were treated as multiple mutations with missing information. For each site, there are only three ways the missing information can be filled in, depending on which of the three alleles is ancestral. This leads to a range of values for B', from which the most conservative one is taken.

In some cases, we wish to demonstrate that there is too *little* linkage disequilibrium in the data; this corresponds to the other tail of the B' statistic or $\Pr(\text{simulated } B' \leq \text{observed } B' \mid \text{SNM}, C = C_{\text{map}})$. Note that this probability is not $1 - \Pr(\text{simulated } B' \geq \text{observed } B')$ since B' is discrete. Using the number of inferred mutations is no longer conservative, so we rerun the simulations with the number of segregating sites to calculate this tail. We do not examine the other tail of C_{hud} (*i.e.*, large values) because, for small samples, C_{hud} is expected to be much larger than the true mean (Hudson 1987). Thus, there is little power to detect an unusually large value of C_{hud}.

RESULTS

C_{hud}/θ_w is consistently too low: Figures 2 and 3 present two estimates of the number of recombination events per mutation for each locus. The first (represented by squares) is based on direct laboratory measurements of the crossing-over rate. This estimate is $c/2\mu$ if the gene is autosomal and $2c/3\mu$ if it is X-linked, where c is the recombination rate per base pair. C_{hud}/θ_w (shown with circles) is estimated from sequence polymorphism data. θ_w is based on the number of (silent and noncoding) segregating sites, not on the number of mutations; *i.e.*, multiple hits are ignored. (This is conservative for our purposes since it leads to a smaller value of θ_w than if the estimate were based on the number of mutations.) C_{hud} is a measure of linkage disequilibrium (see methods). Since both C and θ are scalar multiples of the effective population size under standard neutral assumptions, dividing C_{hud} by θ_w should make the ratio independent of the effective population size under the null model. This is of use because we do not have an estimate of the effective population size that is independent of genetic diversity levels. Scaling C_{hud} to θ_w could also be important if background selection is reducing the effective population size for some loci (Charlesworth *et al.* 1995). As can be seen by eye in Figures 2 and 3, C_{hud}/θ_w is almost always smaller than $c/2\mu$ (or $2c/3\mu$ if X-linked).

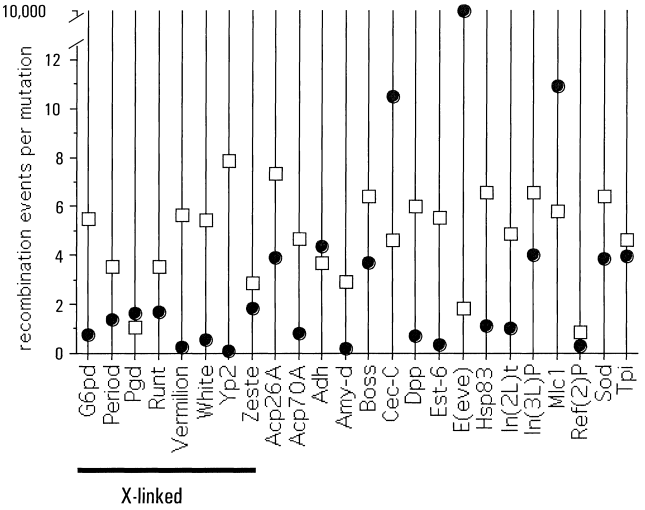

Figure 2.—Two estimates of the number of recombination events per mutation in *D. melanogaster*. Circles represent C_{hud}/θ_w. C_{hud} is a measure of linkage disequilibrium and an estimate of the population recombination parameter under the standard neutral model. θ_w is based on the number of segregating sites at silent sites and noncoding DNA. Squares represent $c/2\mu$ if the locus is autosomal and $2c/3\mu$ if it is X-linked, where c is a laboratory-based estimate of the rate of crossing over per base pair, and μ is an estimate of the mutation rate per base pair per generation (see methods).

One reason for using C_{hud} is that the median is known to be above the true mean (Hudson 1987)—in contrast to, *e.g.*, Hey and Wakeley's estimator of the recombina-

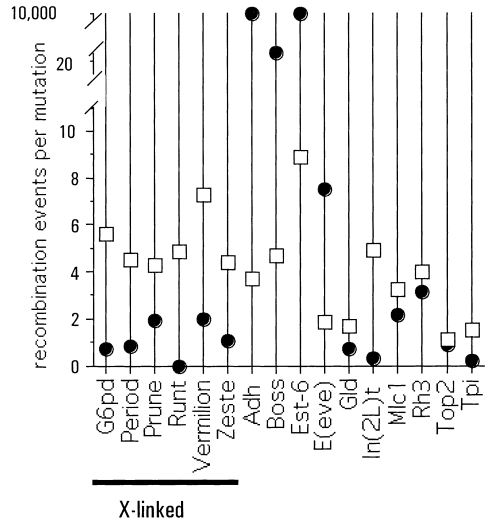

Figure 3.—Two estimates of the number of recombination events per mutation in *D. simulans*. Symbols as in Figure 2. Laboratory-based recombination rates (c) for *Adh*, *E(eve)*, *In(2L)t*, and *Top2* are taken from *D. melanogaster* (see methods).

TABLE 1

Probability of the observed C_{hud} value or less for 24 genes in *D. melanogaster*

Locus	CYT	n	L	S	C_{map}	C_{hud}	P
Acp26A	2L	49	1347	60	118.5	78.1	0.1010
Acp70A	3L	9	1147	34	64.5	11.7	0.0101
Adh	2L	9	2488	41	110.5	95.7	0.3028
Amy-d	2R	9	1485	41	52.6	10.0	0.0133
Boss	3R	5	1565	14	120.1	91.3	0.2422
Cec-C	3R	13	350	26	19.4	119.8	0.9452
Dpp	2L	19	1737	44	125.1	9.1	0.0000
Est-6	3L	9	1682	29	111.7	9.5	0.0010
E(eve)	2R	5	2083	18	46.7	10000	0.8697
G6pd	X	32	1709	29	84.8	17.5	0.0039
Hsp83	3L	8	1385	11	109.1	19.6	0.0391
In(2L)t	2L	35	1095	55	64.4	16.2	0.0091
In(3L)P	3L	9	1367	22	107.7	32.6	0.0464
Mlc1	3R	16	903	21	63.2	96.2	0.5396
Period	X	6	1874	29	59.6	54.2	0.2970
Pgd	X	13	4772	17	46.3	9.7	0.0200
Ref2P	2L	10	2758	38	29.2	7.6	0.0376
Runt	X	11	1931	42	62.0	55.5	0.3059
Sod(-2021)	3L	15	1193	48	91.9	56.8	0.1282
Tpi	3R	25	1074	37	59.1	78.7	0.5571
Vermilion	X	71	2080	111	105.7	10.7	0.0000
White	X	15	5996	82	293.8	18.9	0.0000
Yp2	X	6	1114	11	78.9	1.6	0.0020
Zeste	X	6	999	7	25.8	16.3	0.1861

For each locus in *D. melanogaster*, we list the chromosomal location (CYT), the sample size (n), the length in base pairs (L), the total number of segregating sites (S), the laboratory-based estimate of the population recombination rate for the locus (C_{map}), and the value of C_{hud} for the locus. Recall that our estimates of $2Nc$ do not include the contribution of gene conversion to the overall rate of exchange. P is the proportion of 10,000 runs where the estimate of C_{hud} is the observed value or less in our coalescent simulations (see METHODS).

tion rate γ (HEY and WAKELEY 1997; WALL 2000). Under the standard neutral model and taking estimates of c and μ to be the true rates, we expect that the median of C_{hud}/θ_w will be $>c/2\mu$ (or $2c/3\mu$ for X-linked loci). This was confirmed by simulation (results not shown). Since the loci are independent, we can use a signs test with probability one-half that C_{hud}/θ_w is above $c/2\mu$ (or $2c/3\mu$ for X-linked loci). Such a test is highly conservative, yet significant for both species ($P = 0.0032$ for *D. melanogaster* and $P = 0.019$ for *D. simulans*, one-tailed). Laboratory estimates of the crossing-over rate are not available for chromosome 2 of *D. simulans*. If the rates from *D. melanogaster* are used as surrogates, an additional four data sets can be considered (for 16 genes, $P = 0.038$, one-tailed); the true rates in *D. simulans* are probably higher (see TRUE *et al.* 1996).

***P* values are not uniformly distributed:** Of interest is not only the direction of the discrepancy between C_{hud}/θ_w and $c/2\mu$ but also the magnitude of the difference. To quantify this, we ran coalescent simulations under the assumption that $C = C_{map}$. Tabulated in Tables 1 and 2 for each locus are the proportion P of 10,000 simulated data sets with a C_{hud} value smaller than or equal to the observed one. Recall that under the null model, the distribution of P values across loci should be uniform; this is exceedingly unlikely: $P < 10^{-15}$ and $P < 10^{-9}$ for *D. melanogaster* and *D. simulans*, respectively. We might expect an excess of high P values since many of our assumptions are conservative. Instead we observe too many low P values (Figure 4): for 13 of the 24 *D. melanogaster* loci, $P(C_{hud}) < 0.05$. Similarly, in *D. simulans*, 7 out of 16 loci have P values below 0.05. We conclude that either C_{hud} is too small, *i.e.*, there is too much linkage disequilibrium given our assumed recombination rate, or θ is too large, *i.e.*, there is too much diversity given our assumed mutation rate.

DISCUSSION

Recombination rates: Our results rely on the assumption that the laboratory rates of crossing over, which are interpolated from measurements over large distances, are not overestimates. The correlation between diversity levels and recombination rates (BEGUN and AQUADRO 1992; CHARLESWORTH 1996) suggests that the relative rates in different regions are well characterized, but the absolute rates could be in error. Multiple lines of evidence suggest that they are not: first, four direct intragenic measurements of recombination in *D. melano-*

TABLE 2

Probability of the observed C_{hud} value or less for 16 genes in *D. simulans*

Locus	CYT	n	L	S	C_{map}	C_{hud}	P
Adh	2L	5	2503	57	222.3	10000	0.8580
Boss	3R	5	1655	40	185.4	1674.4	0.7169
Est-6	3L	4	1754	72	374.0	10000	1.0000
E(eve)	2R	6	2083	26	93.3	87.5	0.3269
Gld	3R	11	1551	26	64.5	25.9	0.0668
G6pd	X	12	1710	16	173.7	14.2	0.0026
In(2L)t	2L	11	1102	78	129.6	10.0	0.0000
Mlc1	3R	8	903	15	70.4	17.6	0.0460
Period	X	6	1878	54	153.2	56.1	0.0830
Prune	X	3	1416	17	109.9	41.8	0.1360
Rh3	3R	5	1130	30	108.0	184.6	0.4648
Runt	X	11	1931	20	169.2	0.0	0.0000
Top2	2L	5	556	22	15.0	11.6	0.2292
Tpi	3R	9	805	19	29.6	8.3	0.0176
Vermilion	X	21	1495	55	196.7	46.35	0.0000
Zeste	X	6	999	18	79.1	24.8	0.0766

Symbols are defined in Table 1. *D. melanogaster* recombination rates are used as surrogates for loci on chromosome 2.

gaster estimate rates that are equal to or above those inferred from larger distances (see references in Lindsley and Zimm 1992 for *rudimentary*, *white*, *lozenge*, and *rosy*). As an illustration, for *white*, the intralocus estimate of recombination is the same as the regional estimate of crossing over (2.5×10^{-8}/bp/generation), yet C_{hud}/θ_w is 18-fold smaller than predicted in the sample from *D. melanogaster*. Second, differences between laboratory and natural conditions such as age and temperature appear to have only minor effects on the rate of recombination (Ashburner 1989, pp. 460–461). Genetic background seems to have small effects as well: recombination rates estimated from F_1 progeny of laboratory strains and wild lines of *D. melanogaster* (Brooks and Marks 1986) are in close agreement with rates estimated from laboratory strains. To explain our observations, laboratory rates would have to be systematic overestimates of the true rate, rather than randomly erroneous, since for 40 loci, C_{hud} is almost always lower than expected. Yet the laboratory estimates ignore the contribution of gene conversion, which, on the scale of a gene, should be on the order of crossing over (Andolfatto and Norborg 1998). Thus, if anything, the laboratory measurements of crossing over are more likely to be twofold *under*estimates of the total rate of genetic exchange on an intralocus scale.

A concern in *Drosophila melanogaster* is the presence of high-frequency chromosomal inversions in natural populations (Lemeunier and Aulard 1992), since heterozygotes experience reduced recombination. The ensuing reduction in the population recombination rate is at most twofold (if inversions are at 50% frequency and with complete inhibition of recombination in heterozygotes); this is not sufficient to account for the skew in Figure 4a. The presence of inversions in the samples themselves could be expected to generate linkage disequilibrium. With the exception of *Boss* and *Mlc1*, inverted chromosomes were excluded when the locus is close to a breakpoint of a common inversion. The existence of a high-frequency inversion may still affect the extent of association between sites on standard haplotypes close to the breakpoints (this is analogous to sampling from one deme in a subdivided population). Farther from the inversion breakpoints, gene conversion between chromosomal rearrangements is likely to be high enough (Chovnick 1973; Andolfatto *et al.* 1999) to homogenize differences between arrangement classes for genes. In support of this, a test for subdivision (Hudson *et al.* 1992a) between inverted and standard chromosomes for *Est-6*, located in the middle of *In(3L)P*, was not significant ($P = 0.5871$). Most salient to our observation, inversions are rare and at low frequency in *D. simulans* and on the X chromosome of *D. melanogaster* (Lemeunier and Aulard 1992), which also have significantly low values of C_{hud} (for the eight loci on the X chromosome in *D. melanogaster*, a uniformity test yields $P < 10^{-9}$). Note finally that autosomal inversion heterozygosity can also increase rates of recombination on the X chromosome, which would make laboratory estimates of c for X-linked loci *under*estimates of recombination rates in natural populations (Schultz and Redfield 1951; Sniegowski *et al.* 1994).

Mutation rates: A second assumption is that the neutral mutation rate per base pair per generation is $\hat{\mu} = 3 \times 10^{-9}$. This assumption enters into the signs test directly and as a means to estimate N for our simulations (see methods). In what follows, we review what is known about the mutation rate in Drosophila and discuss the sensitivity of our results to this parameter.

Synonymous divergence estimates vary across loci

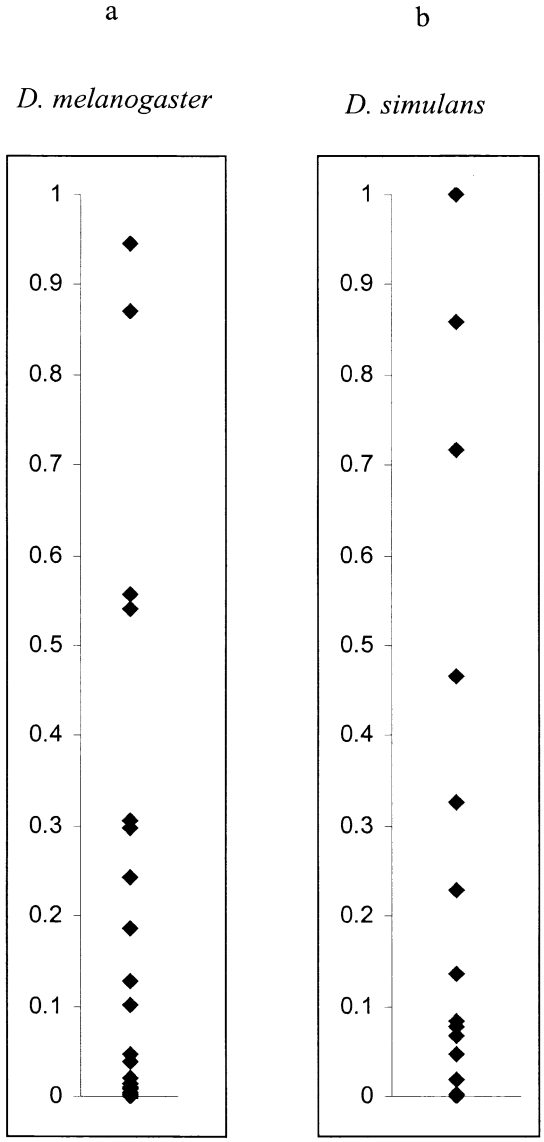

FIGURE 4.—(a) The distribution of *P* values for *D. melanogaster* (Table 1). *P* values represent the proportion of runs with a value of C_{hud} below or equal to the observed value (see METHODS). If the null model is correct, the *P* values should be uniformly distributed. (b) The distribution of *P* values for 16 loci in *D. simulans* (Table 2).

from 0.37×10^{-8}/bp/year to 2.98×10^{-8} (with an average of 1.56×10^{-8}; LI 1997, p. 191). These estimates rely on $\hat{T} = 30$ my for the time to the common ancestor of the *melanogaster* and *obscura* species groups (THROCKMORTON 1975). While $\hat{T}$ could be in error (*e.g.*, RUSSO *et al.* 1995 suggest $\hat{T} = 25$ mya), the substitution rates estimated for *D. melanogaster* are consistent with several independent estimates from Hawaiian Drosophila (ROWAN and HUNT 1991; KAMBYSELLIS *et al.* 1995; FLEISCHER *et al.* 1998) as well as with the results of mutation accumulation experiments (which do not rely on the same assumptions about the substitution process or the time to divergence). On the basis of the pooled results of three spontaneous mutation rate experiments, HARADA *et al.* (1993) estimated the allozyme band-morph mutation rate to be 7.5×10^{-7} per generation for an average protein size of 400 amino acids. Assuming that roughly one-third of changes are detected (KEIGHTLEY and EYRE-WALKER 1999), this rate corresponds to 2×10^{-9}/bp/generation (with an upper 95% confidence limit of 4.75×10^{-9}).

Comparing per-generation rates to per-year rates is complicated by the fact that we do not know the annual number of generations in the wild. In the laboratory, the generation time is ~14.5 days (*i.e.*, 25 generations a year) at 20° (ASHBURNER 1989, pp. 193–194). Generation time at 20° can be as long as 23 days in some species of Drosophila (~16 generations a year) although it is generally shorter for temperatures between 20° and 30°. Both species considered here, which are now cosmopolitan, are thought to have spent most of their evolutionary history in warm climates (LACHAISE *et al.* 1988). In summary, the plausible range of mutation rates per generation varies from 0.6×10^{-9} (the average synonymous substitution estimate with 26 generations a year) to 4.75×10^{-9}. There is a genome-wide excess of linkage disequilibrium for any mutation rate within this range (results not shown).

Variation in mutation rates: If our mutation rates and laboratory estimates of recombination are roughly correct, then C_{hud} is unexpectedly low; *i.e.*, there is a genome-wide excess of linkage disequilibrium. Several departures from the standard neutral model assumptions could potentially generate an excess of linkage disequilibrium. In our simulations, mutations are placed uniformly along the sequence while in actual data sets, mutations are clustered, presumably because of heterogeneity in selective constraints. Thus, the average distance between segregating sites may be smaller in actual data than in simulated data, perhaps resulting in more linkage disequilibrium.

To test this possibility, we ran coalescent simulations with the same parameters as before (see METHODS) but with a simple model of variation in mutation rates (see Figure 1). The results for *D. melanogaster* are presented in Table 3. To verify that our model produces at least as much spatial heterogeneity as observed in the data, we tabulated the longest distance (D_{max}) between any two consecutive polymorphisms in the actual data set (*cf.*, GOSS and LEWONTIN 1996). For an equal number of segregating sites, a greater D_{max} indicates more spatial heterogeneity of mutations. We kept track of the proportion of the simulated runs with D_{max} greater than or equal to the observed D_{max} when rates vary fourfold. As indicated in column three of Table 3, the actual data show much less heterogeneity than does our model. Yet a fourfold variation in mutation rates has very little effect on our results: for *D. melanogaster*, for instance, a uniformity test on the distribution of $P(C_{hud})$ still yields $P < 10^{-10}$. While actual data is not heterogeneous in exactly the same way as predicted by our model, these simulations (and simulations with an eightfold variation in

mutation rates; results not shown) suggest that even strong heterogeneity does not alter our qualitative conclusions. We can therefore rule out variation in mutation rates as being an important factor.

Variation in recombination rates: C_{hud} does not use information about the distance between segregating sites, only the extent of association among them. Small-scale variation in recombination rate should look similar to variation in mutation rates, with the history of sites within a locus more or less correlated than expected given the physical distance between them. In fact, for an infinite sites model, variation in recombination rates can be implemented in the same way as is done for variation in mutation rates (see METHODS). The simulated sequence can be thought of as a sequence in genetic rather than physical distance. Variation in recombination rates affects the translation of the genetic distance into physical distance by a factor similar to the one used to model variation in mutation rates. For example, if a particular interval has a low rate of recombination, then the number of mutations placed on it will be larger than expected given the (genetic) distance. So if recombination rates vary to the same extent as modeled for mutation rates, they should have only minor effects on C_{hud}.

On a larger scale, if the Drosophila genome were a collection of few hotspots and many coldspots for recombination, rates averaged over larger distances could yield systematic overestimates of the mean rate. Since we observe few data sets where C_{hud} is above the laboratory rate, recombination rates at these hotspots would have to be several orders of magnitude above those observed at most loci. Whether this is plausible is unclear. Recombination hotspots of this magnitude have been reported in fungi, humans, and mice (LICHTEN and GOLDMAN 1995; JEFFREYS *et al.* 1999) as well as in maize (*e.g.*, DOONER and MARTINEZ-FEREZ 1997) but, to our knowledge, not in Drosophila.

Departures from the demographic assumptions of the standard neutral model: Departures from the demographic assumptions of a panmictic equilibrium can also generate an excess of linkage disequilibrium (and possibly decrease C_{hud}). If there is population structure, for example, the effective population recombination rate should be decreased relative to panmixia since haplotypes in different subpopulations will not have a chance to recombine as often. In the data analyzed here, sampling schemes vary greatly across loci, with a third sampled entirely outside of Africa. If samples are partitioned by sampling location, population-specific estimates of C_{hud} are sometimes higher and sometimes lower than are estimates based on total samples, so there is no clear effect of pooling different populations (results not shown). However, few samples include a large number of sequences from multiple populations, so this approach is not very informative.

Instead, we try to assess the fit of the data to simple demographic models. Several researchers have argued that populations sampled at present are the result of the recent admixture of previously subdivided populations (*e.g.*, RICHTER *et al.* 1997; HASSON *et al.* 1998). This claim usually stems from the observation of two or more highly diverged haplotypes (*e.g.*, HUDSON *et al.* 1997; LABATE *et al.* 1999). Very recent admixture (*i.e.*, several hundred years ago) should look similar to sampling two demes at the present. This suggests an island model where samples are drawn from both subpopulations. But it is also possible that most or all samples are drawn from only one historical deme. *D. melanogaster* and *D. simulans* are thought to have an African origin (HALE and SINGH 1987; LACHAISE *et al.* 1988) but the number of demes is unknown. Most sampled localities may have been founded from one historical deme.

We evaluate the fit of a symmetric two-island model

TABLE 3

The effect of variation in mutation rates on the probability of C_{hud} (*D. melanogaster*)

Locus	Simulated > observed	Uniform	Step function
Acp26A	0.8217	0.1010	0.2416
Acp70A	0.8498	0.0101	0.0238
Adhs	0.3015	0.3028	0.3389
Amy-d	0.9752	0.0133	0.0234
Boss	0.5919	0.2422	0.2630
Cec-C	0.9928	0.9452	0.8603
Dpp	0.4836	0.0000	0.0010
Est6	0.7862	0.0010	0.0060
E(eve)	0.9821	0.8697	1.0000
G6pd	0.9535	0.0039	0.0403
Hsp83	0.3254	0.0391	0.1128
In2(L)t	0.9020	0.0091	0.0164
In3(L)P	0.7219	0.0464	0.0918
Mlc1	0.5169	0.5396	0.5169
Period	0.9308	0.2970	0.3265
Pgd	0.9425	0.0200	0.0663
Ref2p	0.2669	0.0376	0.0433
Runt	0.7438	0.3059	0.3359
Sod(-2021)	0.9847	0.1282	0.1999
Tpi	0.9122	0.5571	0.5349
Vermilion	0.9861	0.0000	0.0000
White	0.9150	0.0000	0.0000
Yp2	0.8345	0.0020	0.0191
Zeste	0.3212	0.1861	0.6647

To measure the extent of spatial heterogeneity in the distribution of segregating sites, we tabulated the length of the longest number of base pairs between any two adjacent polymorphisms in the sample D_{max}. For an equal number of segregating sites, greater D_{max} indicates more heterogeneity. Listed in the second column is the proportion of the simulated runs with a $D_{max} \geq$ the observed D_{max} when rates vary according to Figure 1. For each gene in *D. melanogaster*, we list the proportion of 10,000 runs with a simulated value of C_{hud} below or equal to the observed one. Column three lists the results when each site is equally likely to mutate (P values in Table 1), column four when mutation rates are variable (see Figure 1).

TABLE 4

A test of a simple model of population structure for *D. melanogaster* and *D. simulans*

Statistic	Model		
	Panmixia	$4Nm = 1$	$4Nm = 0.7$
	D. melanogaster		
C_{hud}	10^{-15}	0.006	0.054
B'	10^{-5}	0.684	0.956[a]
	D. simulans		
C_{hud}	10^{-9}	0.004	—[b]
B'	0.19[c]	0.996	—[b]

We consider a symmetric two-island model; all individuals are sampled from one deme. $4Nm$ ($3Nm$ for X-linked loci) is the population migration parameter and corresponds to the number of migrants per deme per generation. C_{hud} and B' are measures of linkage disequilibrium (see MATERIALS AND METHODS). $P(C_{hud})$ and $P(B')$ are one-tailed. Each cell lists the probability U that the distributions of P values for the statistic are uniformly distributed, as they should be under the null model. For C_{hud} and B', low U values correspond to excess linkage disequilibrium in the actual data relative to simulations.

[a] For $4Nm = 0.7$, there is significantly too *little* linkage disequilibrium in the actual data as measured by B'. (Note that this cannot be inferred from considering the value 1-U since B' is discrete.)

[b] Simulations were not performed for *D. simulans* with this migration rate.

[c] Even though this test does not reveal this feature, the distribution of B' is unusual when $4Nm = 1$ (values of B' are lower in actual data than in simulated data; *i.e.*, there is significantly less linkage disequilibrium in actual data than expected).

by considering changes in the distribution of $P(C_{hud})$ and $P(B')$ as the migration parameter $4Nm$ varies from 0.2 to 1. These values of $4Nm$ are lower than what is suggested by most F_{ST} values. For all parameter values we tried, sampling from both demes was a worse fit to the data than sampling entirely from one (results not shown). Thus, we present results only for the latter case. As expected, as the migration rates decrease, there is more linkage disequilibrium in the simulated data. As a result, low values of $P(C_{hud})$ and $P(B')$ become more likely. In Table 4, we report the probability U that the distributions of $P(C_{hud})$ and $P(B')$ are uniform (as they should be if the null model is correct). In both species, many C_{hud} values are still too low for $4Nm > 1$ ($U = 0.006$ for *D. melanogaster* and 0.004 for *D. simulans*). For $4Nm \leq 1$, there are too many high $P(B')$ values, *i.e.*, too *little* linkage disequilibrium in the data relative to the predictions of the model. Many of the assumptions made in this analysis are no longer conservative when considering the other tail (*i.e.*, too little linkage disequilibrium). However, this analysis does suggest that a simple island model will only account for the low values of C_{hud} if migration rates are lower than suggested by levels of population differentiation and that the model can explain some aspects of linkage disequilibrium (C_{hud}) but not others (B'). When a summary of the frequency spectrum (TAJIMA 1989) is considered along with aspects of linkage disequilibrium, an island model is found to be an even worse fit (results not shown). Under a five-island model where only one island is sampled, the qualitative results are the same (results not shown).

It may not be surprising that an equilibrium island model is an inadequate demographic model, as the history of the species is likely to have been much more complex. In particular, it has been argued that non-African populations have experienced a population bottleneck, *e.g.*, with the European colonization of the Americas (DAVID and CAPY 1988; BEGUN and AQUADRO 1993, 1994, 1995). A variant on this model further assumes that populations were subdivided in Africa and that non-African populations are the result of recurrent founder events from different historical demes (HAMBLIN and VEUILLE 1999). Either version predicts a reduction in variation outside of Africa relative to African populations. For *D. melanogaster*, the data in support of this hypothesis are equivocal: loci on the X show a reduction in variability in American populations relative to those of Zimbabwe (BEGUN and AQUADRO 1993), while loci on the autosomes are equally likely to be more or less variable outside of Africa (P. ANDOLFATTO, unpublished results). For *D. simulans*, there does appear to be a reduction in variability outside Africa (IRVIN *et al.* 1998), although population structure in Africa may be a confounding factor (P. ANDOLFATTO, unpublished results).

Interestingly, for loci in *D. melanogaster*, C_{hud} values appear to be higher when estimated on the basis of African populations alone than when estimated from non-African populations (such a comparison is possible only for *Acp26a*, *Adh*, *In(2L)t*, and *Vermilion*). This is not true of the two suitable loci in *D. simulans* (*G6pd* and *Vermilion*). These findings might suggest that African *D. melanogaster* populations are closer to linkage equilibrium than are non-African ones. However, samples from Africa tend to be smaller than those from outside Africa, and we expect the median C_{hud} value to be larger for smaller samples, even in the absence of population structure (HUDSON 1987). A proper test of this demographic model will require more consistently sampled data and a large number of sequences for multiple populations.

Natural selection: An alternative to a demographic explanation is that natural selection has acted on or near many of the loci. (These two alternatives are not mutually exclusive.) While a demographic departure from model assumptions may be a more parsimonious explanation for a systematic trend, it remains possible that natural selection is pervasive and commonly leads to increased linkage disequilibrium. In addition, it should

be kept in mind that many of the loci analyzed here were collected because of prior evidence for selection.

One mode of selection thought to be prevalent in the Drosophila genome are "selective sweeps" in which a rare advantageous allele is rapidly fixed in the population (Maynard Smith and Haigh 1974). What effect selective sweeps would have on C_{hud} is unknown. Along with a skew in the frequency spectrum, a prediction of this model is a reduction in variability near the selected site (Braverman *et al.* 1995). However, the loci with a significant skew in the frequency spectrum (Tajima 1989) do not have noticeably lower levels of variation than other genes (significance is assessed with simulations that condition on the laboratory-based estimate of *C*). Of the five such loci in *D. melanogaster* (*Dpp*, *Hsp83*, *Ref2p*, *Tpi*, and *Vermilion*) and three in *D. simulans* (*Period*, *Runt*, and *Tpi*), two and two, respectively, have θ_w above the median of loci (results not shown). The scenario just outlined assumes that a variant is selected while rare and swept to fixation in a panmictic population. A variation on this model is transient selection, where a variant is only swept to intermediate frequency (Hudson *et al.* 1994). Other types of selection that could generate linkage disequilibrium include local adaptation (*e.g.*, Stephan *et al.* 1998) and "traffic models" where multiple, linked beneficial mutations are selected for simultaneously (Kirby and Stephan 1996). These models may not lead to a detectable reduction in variability; none has yet been modeled explicitly (but see Gillespie 1997).

Epistatic interactions between nearby sites can also generate considerable linkage disequilibrium. In this type of model, recombinant haplotypes are selected against, reducing the effective frequency of crossing over. As an example, *Adh* protein production in *D. melanogaster* has been shown to be determined by epistatic interactions among multiple polymorphisms (Stam and Laurie 1996). Compensatory interactions between sites involved in the maintenance of secondary pre-mRNA structure might also generate linkage disequilibrium (*e.g.*, Kirby *et al.* 1995). Intralocus epistasis would have to be pervasive to account for the genome-wide departure reported here.

In conclusion, unless it can be demonstrated that laboratory-based estimates of the crossing-over rate are systematic overestimates of the recombination rate, any theory for patterns of variability in the genome will have to accommodate a genome-wide excess of linkage disequilibrium (as measured by C_{hud}). If this excess is best explained by a demographic departure from the standard neutral model, this has important implications for both parameter estimation (such as estimates of θ) and for inferences from patterns of polymorphism at a particular locus. An alternative interpretation is that the mutation rates are much higher than indicated by levels of divergence at silent sites and noncoding DNA for *any* gene. This would point to a different but equally serious problem with the neutral theory.

We thank B. Charlesworth, M. Foote, M. Hamblin, R. Hudson, M. Kreitman, and J. Wall for helpful discussions and comments on an earlier version. This manuscript was significantly improved by comments from A. Clark, C. Langley, and an anonymous reviewer. R. Hudson, J. Wakeley, and J. Wall provided computer programs, S.-C. Tsaur provided unpublished *Adh* data, and J. Comeron and J. True provided genetic and physical map data.

LITERATURE CITED

Aguadé, M., N. Miyashita and C. H. Langley, 1992 Polymorphism and divergence in the *Mst26A* male accessory gland gene region in Drosophila. Genetics **132:** 755–770.

Andolfatto, P., and M. Kreitman, 2000 Molecular variation at the *In(2L)t* proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans*. Genetics **154:** 1681–1691.

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

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.

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

Ayala, F. J., and D. L. Hartl, 1993 Molecular drift of the bride of sevenless (*boss*) gene in Drosophila. Mol. Biol. Evol. **10:** 1030–1040.

Ayala, F. J., B. S. Chang and D. L. Hartl, 1993 Molecular evolution of the *Rh3* gene in Drosophila. Genetica **92:** 23–32.

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

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

Begun, D. J., and C. F. Aquadro, 1994 Evolutionary inferences from DNA variation at the *6-Phosphogluconate Dehydrogenase* locus in natural populations of *Drosophila*: selection and geographic differentiation. Genetics **136:** 155–171.

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

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.

Brooks, L. D., and R. W. Marks, 1986 The organization of genetic variation for recombination in *Drosophila melanogaster*. Genetics **114:** 525–547.

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

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

Chovnick, A., 1973 Gene conversion and transfer of genetic information within the inverted region of inversion heterozygotes. Genetics **75:** 123–131.

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

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

Cooke, P. H., and J. G. Oakeshott, 1989 Amino acid polymorphisms for *Esterase 6* in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **86:** 1426.

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

Dooner, H. K., and I. M. Martinez-Ferez, 1997 Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell **9:** 1633–1646.

Eanes, W. F., M. Kirchner and J. Yoon, 1993 Evidence for adaptive evolution of the G6pd gene in *Drosophila melanogaster* and *D. simulans*. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

Eanes, W. F., M. Kirchner, J. Yoon, C. Biermann, I. N. Wang *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the G6pd locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027–1041.

Fleischer, R. C., C. E. McIntosh and C. L. Tarr, 1998 Evolution on a volcanic conveyor belt: using phylogeographic reconstructions and K-Ar-based ages of the Hawaiian Islands to estimate molecular evolutionary rates. Mol. Ecol. **7:** 533–545.

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

Goss, P. J. E., and R. C. Lewontin, 1996 Detecting heterogeneity of substitution along DNA and protein sequences. Genetics **143:** 589–602.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3:** 479–502.

Hale, L. R., and R. S. Singh, 1987 Mitochondrial DNA variation and genetic structure in populations of *Drosophila melanogaster*. Mol. Biol. Evol. **4:** 622–637.

Hamblin, M. T., and C. F. Aquadro, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with background selection. Mol. Biol. Evol. **13:** 1133–1140.

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.

Harada, K., S. I. Kusakabe, T. Yamazaki and T. Mukai, 1993 Spontaneous mutation rates in null and band-morph mutations of enzyme loci in *Drosophila melanogaster*. Jap. J. Genet. **68:** 605–616.

Hasson, E., and W. F. Eanes, 1996 Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. Genetics **144:** 1565–1575.

Hasson, E., I. N. Wang, L. W. Zeng, M. Kreitman and W. Eanes, 1998 Nucleotide variation in the Triose Phosphate Isomerase (*Tpi*) locus of *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **15:** 756–769.

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 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., 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. 23–26 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Japan Sci. Soc., Tokyo.

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

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

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.

Hudson, R. R., A. G. Saez and F. J. Ayala, 1997 DNA variation at the *Sod* locus of *Drosophila melanogaster*: an unfolding story of natural selection. Proc. Natl. Acad. Sci. USA **94:** 7725–7729.

Inomata, N., H. Shibata, E. Okuyama and T. Yamazaki, 1995 Evolutionary relationships and sequence variation of *alpha-Amylase* variants encoded by duplicated genes in the *Amy* locus of *Drosophila melanogaster*. Genetics **141:** 237–244.

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.

Jeffreys, A. J., R. Barber, P. Bois, J. Buard, Y. E. Dubrova *et al.*, 1999 Human minisatellites, repeat DNA instability and meiotic recombination. Electrophoresis **20:** 1665–1675.

Kambysellis, M. P., E. M. Craddock, F. Piano, M. Parisi and J. Cohen, 1995 Pattern of ecological shifts in the diversification of Hawaiian Drosophila inferred from a molecular phylogeny. Curr. Biol. **5:** 1129–1139.

Karotam, J., A. C. Delves and J. G. Oakeshott, 1993 Conservation and change in structural and 5′ flanking sequences of *esterase 6* in sibling Drosophila species. Genetica **88:** 11–28.

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

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

Kirby, D. A., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.

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.

Kirby, D. A., S. V. Muse and W. Stephan, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

Kliman, R. M., and J. Hey, 1993 DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

Kreitman, M., 1983 Nucleotide polymorphism at the *Alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature **304:** 412–417.

Labate, J. A., C. H. Biermann and W. F. Eanes, 1999 Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **16:** 724–731.

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.

Leicht, B. G., S. V. Muse, M. Hanczyc and A. G. Clark, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139:** 299–308.

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., 1997 *Molecular Evolution*. Sinauer Press, Sunderland, MA.

Li, W. H., and M. Nei, 1974 Stable linkage disequilibrium without epistasis in subdivided populations. Theor. Popul. Biol. **6:** 173–183.

Lichten, M., and A. S. H. Goldman, 1995 Meiotic recombination hotspots. Annu. Rev. Genet. **29:** 423–444.

Lindsley, D. L., and G. G. Zimm, 1992 *The Genome of Drosophila melanogaster*. Academic Press, San Diego.

Ludwig, M. Z., and M. Kreitman, 1995 Evolutionary dynamics of the enhancer regions of *even-skipped* in Drosophila. Mol. Biol. Evol. **12:** 1002–1011.

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

Miyashita, N. T., and C. H. Langley, 1994 Restriction map polymorphism in the *forked* and *vermilion* regions of *Drosophila melanogaster*. Jap. J. Genet. **69:** 297–305.

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

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

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

Ohta, T., 1982 Linkage disequilibrium with the island model. Genetics **101:** 139–155.

Palopoli, M. F., and C.-I. Wu, 1996 Rapid evolution of a coadapted gene complex: evidence from the *Segregation Distorter* (*SD*) system of meiotic drive in *Drosophila melanogaster*. Genetics **143:** 1675–1688.

Richter, B., M. Long, R. C. Lewontin and E. Nitasaka, 1997 Nucleotide variation and conservation of the *dpp* locus, a gene controlling early development in *Drosophila*. Genetics **145:** 311–323.

Robertson, A., 1975 Remarks on the Lewontin-Krakauer test. Genetics **80:** 396.

Rowan, R. G., and J. A. Hunt, 1991 Rates of DNA change and phylogeny from the DNA sequences of the *Alcohol Dehydrogenase* gene for five closely related species of Hawaiian Drosophila. Mol. Biol. Evol. **8:** 49–70.

Russo, C. A., N. Takezaki and M. Nei, 1995 Molecular phylogeny and divergence times of Drosophilid species. Mol. Biol. Evol. **12:** 391.

Schaeffer, S. W., and E. L. Miller, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

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

Simmons, G. M., W. Kwok, P. Matulonis and T. Venkatesh, 1994 Polymorphism and divergence at the prune locus in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **11:** 666–671.

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.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*. Freeman, New York.

Stam, L. F., and C. C. Laurie, 1996 Molecular dissection of a major gene effect on a quantitative trait: the level of *Alcohol dehydrogenase* expression in *Drosophila melanogaster*. Genetics **144:** 1559–1564.

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

Sumner, C. J., 1991 Nucleotide polymorphism in the *Alcohol Dehydrogenase* Duplicate locus of *Drosophila simulans*: implications for the neutral theory. Undergraduate thesis, Princeton University.

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

Throckmorton, L. H., 1975 The phylogeny, ecology and geography of Drosophila, pp. 421–469 in *Handbook of Genetics*, Vol. III, edited by R. King. Plenum Press, New York.

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.

Tsaur, S. C., C. T. Ting and C. I. Wu, 1998 Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of Drosophila: II. Divergence versus polymorphism. Mol. Biol. Evol. **15:** 1040–1046.

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

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

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

Wayne, M. L., D. Contamine and M. Kreitman, 1996 Molecular population genetics of *ref(2)P*, a locus which confers viral resistance in Drosophila. Mol. Biol. Evol. **13:** 191–199.

Wesley, C. S., and W. F. Eanes, 1994 Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91:** 3132–3136.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: A. G. Clark