

Recombination and the Frequency Spectrum in *Drosophila melanogaster* and *Drosophila simulans*

Molly Przeworski,* Jeffrey D. Wall,† and Peter Andolfatto‡

*Department of Statistics, Oxford University, Oxford, England; †2102 Biological Laboratories, Harvard University; and ‡Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh, Scotland

Most “tests of neutrality” assess whether particular data sets depart from the predictions of a standard neutral model with no recombination. For *Drosophila*, where nuclear polymorphism data routinely show evidence of genetic exchange, the assumption of no recombination is often unrealistic. In addition, while conservative, this assumption is made at the cost of a great loss in power. Perhaps as a result, tests of the frequency spectrum based on zero recombination suggest an adequate fit of *Drosophila* polymorphism data to the predictions of the standard neutral model. Here, we analyze the frequency spectrum of a large number of loci in *Drosophila melanogaster* and *D. simulans* using two summary statistics. We use an estimate of the population recombination rate based on a laboratory estimate of the rate of crossing over per physical length and an estimate of the species’ effective population size. In contrast to previous studies, we find that roughly half of the loci depart from the predictions of the standard neutral model. The extent of the departure depends on the exact recombination rate, but the global pattern that emerges is robust. Interestingly, these departures from neutral expectations are not unidirectional. The large variance in outcomes may be due to a complex demographic history and inconsistent sampling, or to the pervasive action of natural selection.

Introduction

In a survey of genes in three *Drosophila* species, Moriyama and Powell (1996) tested for departures from a model of a random-mating population of constant size with no selection. A test based on the frequency spectrum of alleles (Tajima 1989a) revealed very few departures from the predictions of this standard neutral model in both *Drosophila melanogaster* and *Drosophila simulans*. As is common in the application of “tests of neutrality,” the authors assumed zero recombination. Since there is clear evidence for recombination in the data (e.g., as detected by a four-gamete test; cf. Hudson and Kaplan 1985), this assumption is not realistic. In addition, while it is conservative, it is made at the cost of a great loss in power (Wall 1999).

An advantage of studies with many independent loci is that point estimates of the recombination rate can be used. Random error in the measurement of the recombination rate may lead to incorrect probabilities (P_i) for specific data sets. But with enough data, the distribution of P_i will reflect the extent to which there is a departure from null expectations. For *D. melanogaster* and *D. simulans*, dense genetic and physical maps provide independent estimates of the recombination rates for various gene regions (Ashburner 1989, pp. 453–457). This is one of the strengths of *Drosophila* as a model organism, in contrast to many less well studied species or to humans, for which genetic maps are limited by available pedigree information.

Here, we considered 24 loci in *D. melanogaster* and 15 loci in *D. simulans*, all in areas of normal to high recombination. We estimated the probability of two summaries of the frequency spectra given standard neu-

tral assumptions, an estimate of the species’ effective population size, and laboratory-estimated rates of crossing over per physical length. Laboratory-based estimates represented our best guess at the true crossing-over rate. With these parameters, we found that the frequency spectra of many loci departed from neutral expectations. This departure was not unidirectional: some loci showed an excess of rare frequency variants, while others showed an excess of common variants. To gain an understanding of the sensitivity of our conclusions, we also examined the distribution of P_i for a range of recombination values.

Materials and Methods

Polymorphism Data

Polymorphism data sets included all loci analyzed in Andolfatto and Przeworski (2000) for which the sample size was at least 4: *Acp26A* (Tsauro, Ting, and Wu 1998), *Acp70A* (Cicera and Aguadé 1997), *Adh* (Kreitman 1983; Summer 1991; S. C. Tsauro, personal communication), *Boss* (Ayala and Hartl 1993), *Dpp* (Richter et al. 1997), *E(eve)* (Ludwig and Kreitman 1995), *Est-6* (Cooke and Oakeshott 1989; Hasson and Eanes 1996), *G6pd* (Eanes, Kirchner, and Yoon 1993), *Gld* (Hamblin and Aquadro 1996, 1997), *Hsp83* (Hasson and Eanes 1996), *In(3L)P* (Wesley and Eanes 1994), *In(2L)t* (Andolfatto, Wall, and Kreitman 1999; Andolfatto and Kreitman 2000), *Mlc1* (Leicht et al. 1995), *Pgd* (Begun and Aquadro 1994), *Period* (Kliman and Hey 1993), *Ref(2)P* (Wayne, Contamine, and Kreitman 1996), *Rh3* (Ayala, Chang, and Hartl 1993), *Runt* (Labate, Biermann, and Eanes 1999), *Sod* (Hudson, Saez, and Ayala 1997), *Top2* (Palopoli and Wu 1996) and *Tpi* (Hasson et al. 1998), *White* (Kirby and Stephan 1996), *Yp2* and *Zeste* (Hey and Kliman 1993), and *Vermilion* (Begun and Aquadro 1995). *Amy-d* (Inomata et al. 1995) and *Cec-C* (Clark and Wang 1997) were chosen as representative genes from their respective clusters. *In(2L)t* and *In(3L)P* are polymorphic inversions in *D. melanogaster*.

Key words: *Drosophila*, frequency spectrum, recombination, neutral model.

Address for correspondence and reprints: Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3TG, United Kingdom. E-mail: molly@stats.ox.ac.uk.

Mol. Biol. Evol. 18(3):291–298. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Table 1
Probabilities Under the Standard Neutral Model for 24 Loci in *Drosophila melanogaster*

Locus	n^a	S^b	C_{map}^c	D^d	$P(D C=0)^e$	$P(D C=C_{\text{map}})$	$P(G_{\eta} C=C_{\text{map}})^f$
<i>Acp26A</i>	49	60	118.54	-0.343	0.7332	0.2190	0.2108
<i>Acp70A</i>	9	34	64.46	0.891	0.3754	0.0956	0.0076
<i>Adh</i>	9	41	110.47	-0.402	0.7264	0.2952	0.0293
<i>Amy-d</i>	9	41	52.57	-0.271	0.8360	0.5534	0.6218
<i>Boss</i>	5	14	127.43	0.304	0.7730	0.6530	0.0690
<i>Cec-C</i>	13	26	15.39	0.429	0.9752	1.0000	0.8264
<i>Dpp</i>	19	44	125.06	-0.889	0.3876	0.0238	0.0034
<i>Est-6</i>	9	29	111.68	-0.028	0.9687	0.9798	0.2194
<i>E(eye)</i>	5	18	46.66	-0.034	0.9268	0.5130	0.6836
<i>G6pd</i> [§]	32	29	84.77	-0.379	0.8092	0.4582	0.1189
<i>Hsp83</i>	8	11	109.14	-0.958	0.3886	0.0860	0.1920
<i>In(2L)t</i>	35	55	64.39	-0.306	0.6770	0.5726	0.0003
<i>In(3L)P</i>	9	22	107.72	1.232	0.1583	0.0059	0.0001
<i>Micl</i>	16	21	63.21	1.146	0.2856	0.0830	0.0049
<i>Period</i> [§]	6	29	59.59	-0.548	0.6964	0.2150	0.1389
<i>Pgd</i> [§]	13	17	46.29	0.778	0.3560	0.1710	0.0880
<i>Ref2P</i>	10	38	29.23	-1.118	0.2898	0.0294	0.0987
<i>Runt</i> [§]	11	42	61.99	0.175	0.8660	0.8932	0.6501
<i>Sod</i> (-2021)	15	48	91.86	0.788	0.3336	0.0442	0.0548
<i>Tpi</i>	25	37	55.07	-0.804	0.3916	0.5860	0.1172
<i>Vermilion</i> [§]	71	111	105.66	-0.539	0.5960	0.0788	0.0037
<i>White</i> [§]	15	82	293.80	-0.211	0.9070	0.4574	0.4711
<i>Yp2</i> [§]	6	11	78.87	1.150	0.2006	0.0409	0.0077
<i>Zeste</i> [§]	6	7	25.77	-1.011	0.4670	0.2084	0.3482

^a Sample size.^b Number of segregating sites.^c An estimate of the population recombination rate per site (see *Materials and Methods*).^d D is a summary of the frequency spectrum (Tajima 1989a).^e $P(D)$ is two-tailed (see *Materials and Methods*).^f $P(G_{\eta})$ is a measure of the departure of the entire frequency spectrum from the expectation (Fu 1996).[§] X-linked loci.

Here, the labels refer to sequences spanning the breakpoint sites on standard chromosomes (proximal and distal breakpoints, respectively). For *D. simulans*, *In(2L)t* refers to the homolog 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, Saez, and Ayala 1997). The data are available on request from P.A.

Four of the loci were entirely from noncoding regions (*In(2L)t*, *Dpp*, *In(3L)P*, and *Sod* (2021)). When a locus included a coding region, we combined all segregating sites (synonymous, nonsynonymous, and non-coding) in our analysis. An alternative approach would be to separate sites by type (e.g., synonymous/nonsynonymous). However, our goal was not to interpret particular data sets, but to highlight the importance of the population recombination rate for inferences from sequence polymorphism data. In addition, most loci had very few replacement polymorphisms; exceptions were *Acp26A* (with 27 such sites), *Amy-d* (10), *Est-6* (11), and *Ref(2)P* (7) in *D. melanogaster*, and *Est-6* (11) in *D. simulans*. Of these, only *Ref(2)P* showed a significant departure from the predictions of the standard neutral model (see table 1). The pattern for *Ref(2)P* was also significant if nonsynonymous polymorphisms were excluded (results not shown).

Summary Statistics

We use two summary statistics. The first was D , which compares two estimates of the neutral mutation parameter, θ (Tajima 1989a). D is negative when there is an excess of rare mutations, as would be expected under recent population growth (Tajima 1989b) or after a selective sweep (Braverman et al. 1995). It 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 1989b). Our second summary statistic, G_{η} , is based on the differences between the observed and the expected numbers of mutations in each frequency class (Fu 1996).

Coalescent Simulations

Simulations assumed a Wright-Fisher population at equilibrium, from which samples were drawn randomly. There was no selection, and every new mutation occurred at a previously unmutated site (Kimura 1969). In this implementation, we generated genealogies and then placed the observed number of mutations, S , on the tree (Hudson 1993). This approach departs from standard coalescent simulations that place mutations at a constant rate (of $\theta/2$) along each of the branches. We took this approach because we could observe the number of segregating sites while θ was unknown (cf. Hudson 1993). Simulations have shown that the type I errors of various

statistical tests under this “fixed S ” scheme are roughly equal to the nominal rejection probabilities (Kelly 1997; unpublished data).

Parameters of the simulations were, for each locus, the sample size, the number of base pairs, and the population recombination rate $C = C_{\text{map}}$ (see below). Although our simulations are somewhat unorthodox, we refer to our set of assumptions as the “standard neutral model” (SNM). To implement the coalescent with recombination, we used a modification of a program kindly provided by R. Hudson (see Hudson 1983, 1990). There is no gene conversion and the rate of crossing over is constant per base pair. Tracing the ancestral lineages of the sample backward in time, there are two possible genealogical events. Lineages can coalesce in the usual way, or they can split into two as a result of a crossing-over event. The result of a splitting event is that there are now two lineages, one to the left of the crossing-over point, one to the right. These two lineages are then followed back in time with the others and can themselves coalesce or split. At any given nucleotide site, there is a standard (no recombination) coalescent tree. However, recombination causes different sites to have trees that differ from each other. These trees are not independent of each other, since they share parts of their genealogical histories. For a more detailed explanation, see Hudson (1990).

Estimating Parameters

For the population recombination rate C_{map} , we took $C_{\text{map}} = 2N_e c$ where N_e is the effective population size of the species and c is the female rate of recombination per locus per generation. Since males do not recombine in *D. melanogaster* and *D. simulans* (Ashburner 1989, p. 476), the population parameter C was $(1/2)4N_e c = 2N_e c$ for autosomal genes and $(2/3)3N_e c = 2N_e c$ for X-linked genes (assuming that there was no sexual selection). As our estimate of c , we took the laboratory-based estimate of the rate of crossing over. Thus, our estimate did not include the contribution of gene conversion to the total rate of exchange. An estimate of the rate of crossing over was obtained as follows: for each chromosomal arm, polynomial curves were fit to plots of cumulative genetic distance as a function of cumulative physical distance. The derivative of the polynomial at a given physical map position was taken to be the crossing-over rate for the locus of interest (Ashburner 1989, pp. 453–457; True, Mercer, and Laurie 1996; Comeron, Kreitman, and Aguadé 1999). Laboratory estimates were not available for the second chromosome of *D. simulans*; we used the rates for the homologous region of *D. melanogaster* as surrogates for *Adh*, *E(eve)*, *In(2L)t*, and *Top2*. The true rates in *D. simulans* are likely to be higher (see True, Mercer, and Laurie 1996).

Under the standard neutral model, levels of diversity can be equated with $4N_e u$ (or $3N_e u$ if the locus is X-linked), where u is the mutation rate per generation. Given an estimate of u , this relationship can be used to solve for an estimate of N_e . Mutation rate estimates are

uncertain because they rely on numerous assumptions (e.g., if the rate is estimated from levels of divergence, assumptions about the constancy of the molecular clock, the number of generations per year, and the time to the split with the common ancestor). Plausible estimates of u vary 10-fold, roughly from 0.6×10^{-9} to 4.75×10^{-9} per base pair per generation (see Andolfatto and Przeworski [2000] for a review). A commonly used estimate is 1.6×10^{-9} (based on Sharp and Li [1989] assuming 10 generations per year). We took the value of $\hat{u} = 3 \times 10^{-9}$ per base pair per generation, an estimate based on a 30-MYA split for the obscure and melanogaster species groups, 10 generations per year, and the highest synonymous site divergence rate observed at any gene (Li 1997). For this \hat{u} , the average estimate of N_e across loci (weighting each locus equally) was roughly 10^6 for *D. melanogaster* and 2×10^6 for *D. simulans*. The use of lower mutation rates would lead to higher estimates of N_e and, therefore, to higher estimates of C .

Note that with this method, the estimate of N_e varies between loci. Under the standard neutral model, there is only one N_e for all loci, so variation in N_e is due to random chance. Under other models, such as background selection (Charlesworth, Morgan, and Charlesworth 1993), there is no longer a single N_e for the genome. Instead, different regions experience different effective population sizes (depending on the rates of recombination, the density of targets of selection, and so forth). Here, we included only loci in regions of normal to high recombination, where background selection is expected to have less of an effect (cf. Charlesworth 1996). We also verified that levels of diversity and c were not correlated (results not shown). Since there was no evidence for background selection in our data, and we were testing departures from the standard neutral model, we used the same value of N_e for all loci.

Estimating Probabilities

The statistic G_η was developed as a one-tailed test of the standard neutral model (Fu 1996). The probabilities reported for G_η are $P(G_\eta) = \Pr(\text{simulated } G_\eta \geq \text{observed } G_\eta | \text{SNM}, \hat{C})$, where \hat{C} is an estimate of C . The probabilities were estimated from 10^4 trials for each parameter set. D can be used to test for an excess of either rare or intermediate frequency alleles. In figures 1 and 2, we present $P(D) = \Pr(D \leq D_{\text{obs}} | \text{SNM}, \hat{C})$ for each locus, where D_{obs} is the value observed at a particular locus. Low P values indicate an excess of low-frequency variants. We used a one-tailed probability to demonstrate that departures from the standard neutral model were not unidirectional. Note that because D is discrete, $\Pr(D \geq D_{\text{obs}})$ is not exactly equal to $1 - P(D)$. In the tables, we report the results of a two-tailed significance test. We do so because we have no prior hypothesis about the direction of departure from standard neutral assumptions.

To combine the probabilities estimated for different loci, we used the fact that if all our assumptions held, the distribution of $P(G_\eta)$ and $P(D)$ should be uniform (Sokal and Rohlf 1995). The intuition behind this is as

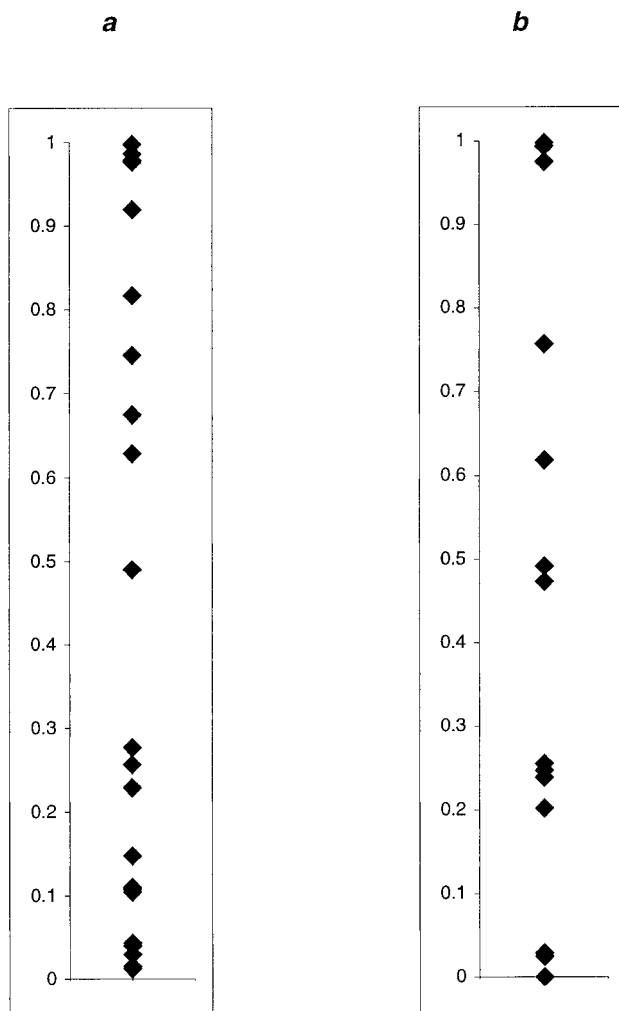


FIG. 1.—The distribution of $P(D) = \Pr(D \leq \text{observed } D | \text{SNM}, C = C_{\text{map}})$ under the standard neutral model for (a) *Drosophila melanogaster* and (b) *Drosophila simulans*.

follows: generate n D_i values ($1 \leq i \leq n$) under the null model, and order them. If the null model is true, then D_{obs} is just one more D value and is equally likely to lie between any two D_i values. Thus, the proportion of D_i values below D_{obs} , or $P(D)$, is uniformly distributed on $(0, 1)$. This is only roughly true for $P(D)$, since D is discrete (although it takes many values). If the P values are uniformly distributed, then for n data sets, $\sum_{i=0}^n -2\ln(P_i)$ should be χ^2 -distributed with $2n$ degrees of freedom (Fisher 1954, as cited in Sokal and Rohlf 1995).

Simulations for G_{η} excluded all sites with multiple mutations. Those for D were run with the 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 mutations at two sites with missing information. For each pair of sites, there were only three ways the missing information could be filled in, depending on which of the three alleles was ancestral. This led to a range of values for D , from which the most conservative one was taken. Similarly, in tables 1 and 2, we report the D val-

ues closest to 0 for the loci with multiple hits. There are nine data sets in *D. melanogaster* which have at most three sites with multiple hits: *Acp26A*, *Acp70A*, *Cec-C*, *E(eve)*, *In(2L)t*, *Mlc1*, *Runt*, *Tpi*, and *Vermilion*. Seven loci in *D. simulans* have more than two alleles at a site: *Adh* has six such sites, while *Boss*, *Est-6*, *Gld*, *In(2L)t*, *Rh3*, and *Vermilion* have fewer than three multiply hit sites.

Results

In tables 1 and 2, we list the estimated (two-tailed) probabilities $P(D)$ for loci in *D. melanogaster* and *D. simulans*, respectively. If we assumed no recombination, not one data set was significant at the 5% level. Similarly, $P(G_{\eta}) > 0.1$ for all loci (results not shown). In contrast, with recombination, several loci departed from the expectations of the standard neutral model (5/24 in *D. melanogaster* and 4/15 in *D. simulans*; see tables 1 and 2). G_{η} appeared to be most sensitive to the population recombination rate. A uniformity test on the distribution of the P values for G_{η} yielded $P < 10^{-10}$ and $P = 2 \times 10^{-4}$ for *D. melanogaster* and *D. simulans*, respectively. The probabilities that the (two-tailed) $P(D)$ values were uniformly distributed were 0.0006 for *D. melanogaster* and 0.0025 for *D. simulans*.

In spite of this widespread departure from standard neutral expectations, there was no clear trend (see fig. 1). For example, the mean and median of D were -0.014 and -0.122 for *D. melanogaster* and 0.104 and 0.022 for *D. simulans*, extremely close to their expectations under the standard neutral model. In *D. melanogaster*, two loci had significantly negative skews in their frequency spectra as measured by D (*Dpp* and *Ref2p*; two-tailed), while three had significantly positive ones (*In(3L)P*, *Sod*, and *Yp2*; two-tailed). In *D. simulans*, the numbers were two (*Period* and *Runt*) and two (*G6pd* and *In(2L)t*), respectively.

The P values for any particular locus were hard to interpret. The extent to which these data sets were unusual depended on the true population rate of recombination. To gauge the sensitivity of the results to the recombination rate, we estimated $P(G_{\eta})$ and $P(D)$ for $C = C_{\text{map}}/4$ to $2C_{\text{map}}$ (see fig. 2). Our best guess at the total recombination rate is $C > C_{\text{map}}$, since C_{map} does not include the effects of gene conversion. On the scale of several kilobases, the contribution of gene conversion to the total genetic exchange is expected to be on the same order as that of crossing over (Andolfatto and Nordborg 1998). For $C = 2C_{\text{map}}$, the distribution of $P(G_{\eta})$ was highly skewed: the probabilities that the distribution of $P(G_{\eta})$ across loci was uniform, as it should be if the null model is adequate, were $P < 10^{-14}$ for *D. melanogaster* and $P < 10^{-5}$ for *D. simulans* (for the distribution of the two-tailed $P(D)$, $P < 2 \times 10^{-5}$ and $P < 3 \times 10^{-4}$, respectively). Since the P values almost always decrease with increased recombination (see tables 1 and 2), the P values would be even lower if the true $C > 2C_{\text{map}}$.

For any particular locus, random errors in measurement could also yield a laboratory estimate above

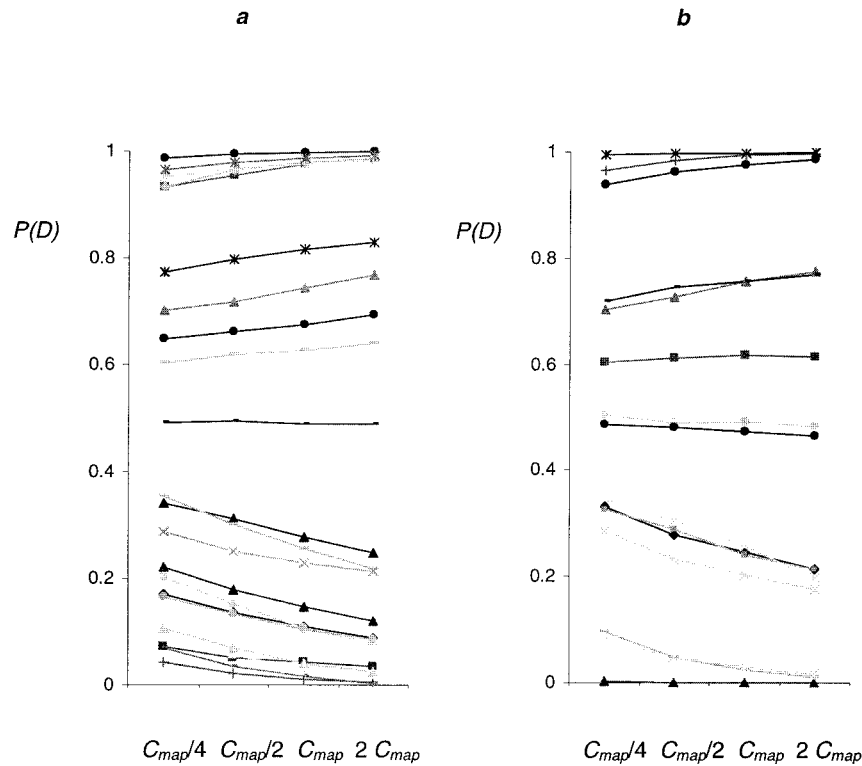


FIG. 2.—The sensitivity of $P(D)$ to the population recombination rate C . For each locus, we estimated $P(D) = \Pr(D \leq \text{observed } D | \text{SNM}, \hat{C})$ conditional on $\hat{C} = C_{\text{map}}/4, C_{\text{map}}/2, C_{\text{map}},$ and $2C_{\text{map}}$. Each line corresponds to a given locus. C_{map} was obtained from laboratory-based estimates of the rate of crossing over per physical length and an estimate of the effective population size (see *Materials and Methods*). Our best guess at the total rate of genetic exchange is $>C_{\text{map}}$ since C_{map} does not include gene conversion. *a*, *Drosophila melanogaster*. *b*, *Drosophila simulans*.

the true rate. The P values for that locus would then be lower than they would be for the true rate of recombination. For *D. melanogaster*, even if the true $C = C_{\text{map}}/4$, the distribution of $P(G_{\eta})$ would still be significantly different from uniform ($P = 0.005$). For $C = C_{\text{map}}/2$, the distribution of $P(G_{\eta})$ is still skewed for both *D. melanogaster* ($P < 10^{-6}$) and *D. simulans* ($P = 0.007$). Results for *D. melanogaster* appear to be less sensitive

to the exact value of C than those for *D. simulans*, perhaps because sample sizes tended to be larger.

Discussion

Effect of the Population Recombination Rate

With no recombination, there is only one genealogical tree at a locus, or one draw from the (highly sto-

Table 2
Probabilities Under the Standard Neutral Model for 15 Loci in *Drosophila simulans*

Locus	n	S	C_{map}	D	$P(D C = 0)$	$P(D C = C_{\text{map}})$	$P(G_{\eta} C = C_{\text{map}})$
<i>Adh</i>	5	57	222.27 ^a	0.232	1.0000	0.5434	0.5313
<i>Boss</i>	5	40	185.36	0.157	0.9176	1.0000	0.1278
<i>Est-6</i>	4	72	373.95	0.326	0.6636	0.7536	0.6825
<i>E(eve)</i>	6	26	93.32 ^a	-0.362	0.8120	0.4056	0.4985
<i>Gld</i>	11	26	64.52	1.016	0.1100	0.0044	0.0103
<i>G6pd</i> ^b	12	16	173.74	1.412	0.3346	0.0820	0.0060
<i>In(2L)t</i>	11	78	129.60 ^a	0.858	0.4186	0.0374	0.0035
<i>Mlc1</i>	8	15	70.43	0.348	0.6744	0.5246	0.3444
<i>Period</i> ^b	6	54	153.24	-0.592	0.7068	0.0476	0.1071
<i>Rh3</i>	5	30	108.03	0.104	0.9898	0.4788	0.6199
<i>Run1</i> ^b	11	20	169.16	-1.466	0.1208	0.0008	0.0181
<i>Top2</i>	5	22	15.01 ^a	-0.534	0.8544	0.5110	0.6933
<i>Tpi</i>	9	19	29.62	-1.084	0.3042	0.0578	0.4306
<i>Vermilion</i> ^b	21	55	196.74	0.000	1.0000	0.9842	0.0150
<i>Zeste</i> ^b	6	18	79.12	-0.066	1.0000	0.9456	0.8949

NOTE.—See footnotes to table 1 for abbreviations.

^a For loci on the second chromosome, there are no laboratory-based estimates of the rates of crossing over per physical length, so the *Drosophila melanogaster* rates of crossing over are taken as surrogates.

^b X-linked loci.

chastic) evolutionary process. With recombination, sites very close to each other are still likely to have similar genealogies, but sites farther apart will have quite different ones. As the recombination rate increases, so will the number of distinct trees at a locus (Hudson 1983). Summaries of the frequency spectrum at neutral sites (such as D and G_{η}) reflect aspects of the underlying trees. As the recombination rate increases, they reflect an average over more trees and thus will tend to take values closer to their expectation. In other words, their variance decreases as the recombination rate increases.

Wall (1999) demonstrated that a large number of statistical tests (based both on the frequency spectrum and on summaries of linkage disequilibrium) have very little power to detect population structure if simulations are run with $C = 0$ when the actual value of C is much higher. This is because the actual variances of the test statistics are smaller than the sample variances of the no-recombination simulations. However, it remained unknown if, in practice, many loci would show departures from the null model for $C > 0$. These results suggest that in *Drosophila*, population rates of recombination are high enough for alternative assumptions about C to have a major effect on one's conclusions. In particular, with our conservative estimate of the population rate of recombination, C_{map} , we found that the standard neutral model was a poor predictor of the frequency spectra observed at many loci.

Sensitivity to the Population Recombination Rate

The results reported here assume that the laboratory-based estimates of the rate of crossing over per physical length are not systematic overestimates and that the estimate of the mutation rate per generation is not a severe underestimate. These assumptions are reviewed in detail in Andolfatto and Przeworski (2000). Even if laboratory-based estimates of the rate of crossing over are not systematic overestimates, the population recombination rates may still be inaccurate because of random error in the measurement of c .

To gain an understanding of how sensitive our results were to the population recombination rate, we calculated $P(D)$ and $P(G_{\eta})$ for a variety of values of C , varying from $2C_{\text{map}}$ to eightfold less. We found that even if all the rates happened to be fourfold underestimates of our best guess at the recombination rate (e.g., because N_e for *D. melanogaster* was actually 250,000 and not 10^6), our qualitative conclusions would hold. More difficult to assess was the effect of random errors in c estimates. In particular, accidental overestimates of c may lead to P values that are lower than they should be. We cannot test this explicitly without an estimate of the sampling variance in laboratory-based estimates of the rate of crossing over. However, we can gain a qualitative understanding of how P values vary with C . P values are strikingly different for $C = 0$ and $C = C_{\text{map}}$ because $C_{\text{map}} \gg 0$. In contrast, twofold differences about $C = C_{\text{map}}$ estimates have little effect on $P(D)$ (see fig. 2). For example, all loci that showed a significant departure

from neutral expectations for $C = C_{\text{map}}$ also did so for $C = C_{\text{map}}/2$.

A central goal of this study was to demonstrate the large impact that different assumptions about C can potentially have on inferences from polymorphism data. However, our ultimate interest is not in methodology alone, but in determining how many loci actually depart from standard neutral assumptions, and for what reasons. To this end, we tried to make conservative assumptions in estimating N_e ; we also were conservative in running simulations without the additional contribution of gene conversion. Given the robustness of our results, it seems plausible that the picture of a widespread departure from neutrality is a realistic one. Below, we discuss possible explanations for this finding.

Possible Departures from the Standard Neutral Model

The demographic assumptions of panmixia and constant population size may be invalid for both species of *Drosophila*. Several authors have argued that one or both species have experienced long-term population structure (e.g., Richter et al. 1997; Hasson et al. 1998) or that non-African populations show the signature of recent founder events (e.g., David and Capy 1988; Begun and Aquadro 1993, 1994, 1995; Hamblin and Veuille 1999). For the data included in this analysis, sampling schemes varied greatly: while both species are thought to have an African origin (Hale and Singh 1987; Lachaise et al. 1988), roughly a third of the data sets contain only non-African samples. Within Africa, different localities are sampled in different studies. Inconsistent sampling and a complex demographic history may explain the large variance in observed frequency spectra. While in theory multilocus data can be used to assess the relative likelihood of alternative demographic models, this type of inference would be greatly facilitated by more consistently sampled data from multiple populations.

An alternative explanation for the large variance in frequency spectra is that distinct modes of selection (e.g., epistatic vs. directional) are operating on different loci, with varying effects on the frequency spectrum. The pervasive action of natural selection seems plausible in light of the fact that many of the samples were collected because of prior evidence for selection. If natural selection commonly shapes the pattern of variation at or near genes, the signature of population history may be obscured. If so, a systematic collection of noncoding regions may be useful for demographic inferences.

Weak Selection

In addition to replacement polymorphisms, silent sites appear to be under weak constraint, as evidenced by their rate of evolution (Akashi 1995); sites in introns and noncoding regions may also be evolving under weak selection (Li 1997). Thus, a more appropriate null model may be one where at least some fraction of the variants are weakly deleterious (Ohta 1973). Under a model of weak selection, the mean D value is expected to be negative (Tajima 1989a; Sawyer and Hartl 1992). In these

data, there is no evidence of a trend toward negative D values in either species: 14 out of 24 loci have negative D values in *D. melanogaster*; 6 out of 15 do in *D. simulans*. However, the power to detect departures may be low, depending on the selection coefficients and the population mutation and recombination rates (Akashi 1999; McVean and Charlesworth 2000; Tachida 2000). Modeling weak selection with recombination is not yet feasible with coalescent simulations. Up to now, infinite-sites coalescent simulations with weak selection have only been implemented for no recombination (Neuhäuser and Krone 1997; Przeworski, Charlesworth, and Wall 1999). Modeling the process forward in time would be extremely computationally intensive for the data analyzed here. It is unclear whether the overall picture would become more probable, as intermediate frequency variants would be less likely than under a strictly neutral model.

The use of laboratory-based estimates of the recombination rate reveals that many loci in *D. melanogaster* and *D. simulans* depart from standard neutral expectations. This picture is strikingly different from the one obtained under the assumption of no recombination. Because we have to rely on an indirect estimate of the population recombination rate, the extent of the departure from model predictions may be overestimated. However, what we wish to emphasize is how different the global picture is for alternative values of the nuisance parameter C . This difference warrants that more attention be given to assumptions about the population recombination rate when drawing inferences from sequence polymorphism data.

Acknowledgments

We thank Brian Charlesworth, Andy Clark, Dick Hudson, and Chuck Langley for comments on an earlier draft of this manuscript. A subset of the simulations were performed on a computer donated by Sun Microsystems Corporation to Richard R. Hudson and Martin Kreitman.

LITERATURE CITED

- AKASHI, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139**:1067–1076.
- . 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**:221–220.
- 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., and M. PRZEWSKI. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**:257–268.
- 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, New York.
- AYALA, F. J., B. S. CHANG, and D. L. HARTL. 1993. Molecular evolution of the Rh3 gene in *Drosophila*. *Genetica* **92**:23–32.
- 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.
- 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.
- . 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.
- . 1995. Molecular variation at the *vermillion* 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.
- 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.
- 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.
- 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.
- FU, Y. X. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**:557–570.
- 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* in consistent with background selection. *Mol. Biol. Evol.* **13**:1133–1140.
- . 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 M. VEUILLE. 1999. Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. *Genetics* **153**:305–317.

- 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.
- HUDSON, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**:183–201.
- . 1990. Gene genealogies and the coalescent process. Pp. 1–44 in D. FUTUYMA and J. ANTONOVICS, eds. *Surveys in evolutionary biology*. Vol. 7. Oxford University Press, Oxford, England.
- . 1993. The how and why of generating gene genealogies. Pp. 23–26 in N. TAKAHATA and A. G. CLARK, eds. *Mechanisms of molecular evolution*. Japan Science Society, Tokyo.
- HUDSON, R. R., and N. L. KAPLAN. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**:147–164.
- HUDSON, R. R., 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.
- KELLY, J. K. 1997. A test of neutrality based on interlocus associations. *Genetics* **146**:1197–1206.
- KIMURA, M. 1969. The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. USA* **63**:1181–1188.
- 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.
- 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. CARIU, J. R. DAVID, F. LEMEUNIER, L. TSACAS, and M. ASHBURNER. 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.
- LI, W. H. 1997. *Molecular evolution*. Sinauer, Sunderland, Mass.
- LUDWIG, M. Z., and M. KREITMAN. 1995. Evolutionary dynamics of the enhancer regions of *even-skipped* in *Drosophila*. *Mol. Biol. Evol.* **12**:1002–1011.
- 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.
- MORIYAMA, E. N., and J. R. POWELL. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**:261–277.
- NEUHAUSER, C., and S. KRONE. 1997. The genealogy of samples in models with selection. *Genetics* **145**:519–534.
- OHTA, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246**:96–98.
- 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.
- PRZEWSKI, M., B. CHARLESWORTH, and J. D. WALL. 1999. Genealogies and weak purifying selection. *Mol. Biol. Evol.* **16**:246–252.
- 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.
- SAWYER, S. A., and D. L. HARTL. 1992. Population genetics of polymorphism and divergence. *Genetics* **132**:1161–1176.
- SHARP, P. M., and W. H. LI. 1989. On the rate of DNA sequence evolution in *Drosophila*. *J. Mol. Evol.* **28**:398–402.
- SOKAL, R. R., and F. J. ROHLF. 1995. *Biometry*. Freeman, New York.
- SUMNER, C. J. 1991. Nucleotide polymorphism in the *Alcohol Dehydrogenase* Duplicate locus of *Drosophila simulans*: implications for the neutral theory. Undergraduate thesis, Princeton University, Princeton, N.J.
- TACHIDA, H. 2000. Molecular evolution in a multisite nearly neutral model. *J. Mol. Evol.* **50**:69–81.
- TAJIMA, F. 1989a. Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- . 1989b. The effect of change in population size on DNA polymorphism. *Genetics* **123**:597–601.
- 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.
- 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.

DAVID M. RAND, reviewing editor

Accepted October 26, 2000